BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Nanoinformatics for environmental health and biomedicine

Edited by Rong Liu and Yoram Cohen

Imprint

Beilstein Journal of Nanotechnology
www.bjnano.org
ISSN 2190-4286
Email: journals-support@beilstein-institut.de

The *Beilstein Journal of Nanotechnology* is published by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften.

Beilstein-Institut zur Förderung der Chemischen Wissenschaften
Trakehner Straße 7–9
60487 Frankfurt am Main
Germany
www.beilstein-institut.de

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Nanoinformatics for environmental health and biomedicine

Rong Liu*[1,2] and Yoram Cohen*[1,2,3]

Editorial

Open Access

Address:
[1]UC Center for Environmental Implications of Nanotechnology, University of California, Los Angeles, Los Angeles, California 90095, United States, [2]UCLA Institute of the Environment and Sustainability, Los Angeles, California 90095, United States, and [3]UCLA Chemical and Biomolecular Engineering Department, Los Angeles, California 90095, United States

Email:
Rong Liu* - rongliu@ucla.edu; Yoram Cohen* - yoram@ucla.edu

* Corresponding author

Keywords:
nanoinformatics

Beilstein J. Nanotechnol. **2015,** *6,* 2449–2451.
doi:10.3762/bjnano.6.253

Received: 28 October 2015
Accepted: 07 November 2015
Published: 21 December 2015

This article is part of the Thematic Series "Nanoinformatics for environmental health and biomedicine".

Editor-in-Chief: T. Schimmel

Nanotechnology has become a significant enabling technology for a wide array of industries being integrated across diverse areas such as medicine, electronics, biomaterials, and energy production. For example, nano-scaled systems have been designed and utilized for safe and effective targeted delivery of therapeutic agents, demonstrating the rapid advancements of nanotechnology in medical-treatment and diagnosis. At the same time, there is also mounting concern regarding the potential impact of nanotechnology on the environment and human health. As a result, there is a global drive to ensure that the development of beneficial nanotechnologies is accomplished in a responsible manner so as to avoid adverse impacts on environmental and human health.

In order to develop safe-by-design nanomaterials for their various intended applications, large amounts of data are being generated for better understanding and mapping the toxicology and pharmacology of nanomaterials. Nanomaterials data are typically sought regarding their physicochemical and structural properties, environmentally related properties, toxicity behavior, processing information, production levels, environmental releases, and more. Accordingly advanced informatics techniques are urgently required for the collection and curation, management (e.g., achieving and sharing), analysis and modeling of the large amount of data involved with nanotechnology processes and materials (i.e., "nano-data"). In order to address these requirements, nanoinformatics has emerged over the last decade as "The science and practice of determining which information is relevant to the nanoscale science and engineering community, and then developing and implementing effective mechanisms for collecting, validating, storing, sharing, analyzing, modeling, and applying that information." [1]. At present, nanoinformatics focuses primarily on: nano-data management and database development, nano-data curation, assessment of the value of information in nano-data, literature mining for nano-data collection and meta-analysis, data mining/machine learning of nano-data (e.g., development of quantitative structure–activity relationships (QSARs)), simulation of the fate and transport of nanomaterials, nano-bio interactions, and assessment of potential environmental and health risks associated with nanomaterials.

As an interdisciplinary field consisting mainly of nanotechnology and data science, nanoinformatics has significantly advanced over the last decade, playing an increasingly important role in research and development in nanomedicine and environmental health impact assessment of nanomaterials (often termed NanoEHS). In addition, efforts in nanoinformatics research have provided in a multitude of tools and resources that are being made available through nanoinformatics cyberinfrastructures and web platforms (e.g., nanoinfo.org [2] in the US and eNanoMapper [3] in the EU). However, much of the current research and advances in nanoinformatics are not documented in dedicated resources and, given the interdisciplinary nature of nanoinformatics, are dispersed throughout a wide range of sources and journals. As a consequence, researchers and practitioners in other fields of nanotechnology have been at a disadvantage not having easy access to the most recent resources and tools provided by the nanoinformatics research community. Accordingly, this Thematic Series is devoted to bring together the state-of-the-art in nanoinformatics with a particular focus on the latest related developments/applications for environmental health and biomedicine.

In this Thematic Series, recent advances in the development of databases are reported. These databases represent a collection of valuable data related to the physicochemical properties and bioactivity of nanomaterials. In one contribution, the latest version of caNanoLab is described along with a critical discussion of the challenges associated with database development for nanomaterials, as well as the needs for nano-data curation and sharing by the biomedical research community [4]. The latest development of the eNanoMapper database for nanomaterial safety information is summarized in another contribution [5], while a third contribution reports on the NanoE-Tox database that is concerned with the ecotoxicity of nanomaterials [6]. In addition, important improvements are reported for the Nanotechnology Consumer Products Inventory that progressively documents the marketing and distribution of nano-enabled products into the commercial marketplace [7].

The progress in nano-data curation is covered in two contributions. One describes the Nanomaterial Data Curation Initiative, a collaborative effort by the nanoinformatics research community for nano-data discovery and extraction, quality assessment, integration, and reuse [8]. Another contribution illustrates key concepts, and discusses current practices and challenges in the field of nano-data curation [9]. In order to facilitate nano-data discovery and extraction, a data collection framework was developed [10] through ISA-TAB-Nano (a set of standardized specifications for nano-data representation). Advances in automating nano-data discovery and extraction is the subject of two other contributions that report on using advanced literature/text mining techniques, such as natural language processing [11] and corpus-based automatic information extraction [12]. In addition, bibliometric and social network analysis is introduced and adopted in the field of nanoinformatics to identify collaboration networks and developmental patterns of nano-enabled drug delivery for brain cancer [13].

As an imported aspect of nanoinformatics, recent advances in data mining/machine learning of nano-data are also reported in this Thematic Series. In one study, the toxicity of ZnO nanoparticles to zebrafish (measured by mortality rate (%)) was correlated to two principal components calculated from nanoparticle size and surface properties using Kriging estimations [14]. Another contribution reports on the development of models to predict the cytotoxicity of PAMAM dendrimers using molecular descriptors [15]. Nanomaterials that have potential to cause disease (e.g., TiO_2 nanoparticles, carbon black, and carbon nanotubes) were also identified using biclustering of gene expression data and gene set enrichment analysis methods [16]. Various visual analytical approaches (e.g., bipartite graphs, log-ratio analysis, and multidimensional scaling) are demonstrated in another study for exploring the impact of manufactured nanoparticles (ZnO and TiO_2) on soil bacterial communities [17], which is an area of nanoinformatics that is only now receiving increased attention.

The present Thematic Series also presents a simulation tool for estimating the release and environmental distribution of nanomaterials, which provides critical information for the environmental impact assessment of nanomaterials [18]. Another contribution addresses the issue of nanomaterial risk assessment and proposes a decision analysis scheme for furthering nanoinformatics work [19]. This work considers an array of decision analysis techniques (e.g., multicriteria decision analysis, value of information, weight of evidence, and portfolio decision analysis) that are potentially capable of assessing and classifying the multitude of available nanomaterial data. Such an approach can serve as the basis for both establich a decision making process and future research priorities in the field.

This Thematic Series was made possible by the contribution of numerous authors to whom we owe our gratitude. We appreciate the time and effort of the numerous referees that helped shape this Thematic Series and we are also grateful for the unwavering support of the team at the Beilstein-Institut. We particularly acknowledge and commend the Beilstein Journal of Nanotechnology for its open access policy, which has provided a wonderful incentive for researchers and practitioners to contribute to this journal

while is freely available to all scientific and professional communities.

Rong Liu and Yoram Cohen

Los Angeles, October 2015

References

1. Nanoinformatics 2020 Roadmap. http://eprints.internano.org/id/eprint/607 (accessed Nov 5, 2015).
2. Nanoinformatics for Environmental Impact Assessment of ENMs. http://nanoinfo.org/ (accessed Nov 5, 2015).
3. eNanoMapper: Computational infrastructure for toxicology data management of engineered nanomaterials. http://www.enanomapper.net/ (accessed Nov 5, 2015).
4. Morris, S. A.; Gaheen, S.; Lijowski, M.; Heiskanen, M.; Klemm, J. *Beilstein J. Nanotechnol.* **2015,** *6,* 1580–1593. doi:10.3762/bjnano.6.161
5. Jeliazkova, N.; Chomenidis, C.; Doganis, P.; Fadeel, B.; Grafström, R.; Hardy, B.; Hastings, J.; Hegi, M.; Jeliazkov, V.; Kochev, N.; Kohonen, P.; Munteanu, C. R.; Sarimveis, H.; Smeets, B.; Sopasakis, P.; Tsiliki, G.; Vorgrimmler, D.; Willighagen, E. *Beilstein J. Nanotechnol.* **2015,** *6,* 1609–1634. doi:10.3762/bjnano.6.165
6. Juganson, K.; Ivask, A.; Blinova, I.; Mortimer, M.; Kahru, A. *Beilstein J. Nanotechnol.* **2015,** *6,* 1788–1804. doi:10.3762/bjnano.6.183
7. Vance, M. E.; Kuiken, T.; Vejerano, E. P.; McGinnis, S. P.; Hochella, M. F., Jr.; Rejeski, D.; Hull, M. S. *Beilstein J. Nanotechnol.* **2015,** *6,* 1769–1780. doi:10.3762/bjnano.6.181
8. Hendren, C. O.; Powers, C. M.; Hoover, M. D.; Harper, S. L. *Beilstein J. Nanotechnol.* **2015,** *6,* 1752–1762. doi:10.3762/bjnano.6.179
9. Powers, C. M.; Mills, K. A.; Morris, S. A.; Klaessig, F.; Gaheen, S.; Lewinski, N.; Hendren, C. O. *Beilstein J. Nanotechnol.* **2015,** *6,* 1860–1871. doi:10.3762/bjnano.6.189
10. Marchese Robinson, R. L.; Cronin, M. T. D.; Richarz, A.-N.; Rallo, R. *Beilstein J. Nanotechnol.* **2015,** *6,* 1978–1999. doi:10.3762/bjnano.6.202
11. Lewinski, N. A.; McInnes, B. T. *Beilstein J. Nanotechnol.* **2015,** *6,* 1439–1449. doi:10.3762/bjnano.6.149
12. Dieb, T. M.; Yoshioka, M.; Hara, S.; Newton, M. C. *Beilstein J. Nanotechnol.* **2015,** *6,* 1872–1882. doi:10.3762/bjnano.6.190
13. Huang, Y.; Ma, J.; Porter, A. L.; Kwon, S.; Zhu, D. *Beilstein J. Nanotechnol.* **2015,** *6,* 1666–1676. doi:10.3762/bjnano.6.169
14. Zhou, Z.; Son, J.; Harper, B.; Zhou, Z.; Harper, S. *Beilstein J. Nanotechnol.* **2015,** *6,* 1568–1579. doi:10.3762/bjnano.6.160
15. Jones, D. E.; Ghandehari, H.; Facelli, J. C. *Beilstein J. Nanotechnol.* **2015,** *6,* 1886–1896. doi:10.3762/bjnano.6.192
16. Williams, A.; Halappanavar, S. *Beilstein J. Nanotechnol.* **2015,** *6,* 2438–2448. doi:10.3762/bjnano.6.252
17. Liu, R.; Ge, Y.; Holden, P. A.; Cohen, Y. *Beilstein J. Nanotechnol.* **2015,** *6,* 1635–1651. doi:10.3762/bjnano.6.166
18. Liu, H. H.; Bilal, M.; Lazareva, A.; Keller, A.; Cohen, Y. *Beilstein J. Nanotechnol.* **2015,** *6,* 938–951. doi:10.3762/bjnano.6.97
19. Bates, M. E.; Larkin, S.; Keisler, J. M.; Linkov, I. *Beilstein J. Nanotechnol.* **2015,** *6,* 1594–1600. doi:10.3762/bjnano.6.162

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Simulation tool for assessing the release and environmental distribution of nanomaterials

Haoyang Haven Liu*[1,2], Muhammad Bilal[1], Anastasiya Lazareva[3], Arturo Keller[1,3] and Yoram Cohen*[1,2]

Full Research Paper

Open Access

Address:
[1]Center for the Environmental Implications of Nanotechnology, California NanoSystems Institute, University of California, Los Angeles, CA 90095, USA, [2]Chemical and Biomolecular Engineering Department, University of California, Los Angeles, Los Angeles, CA 90095, USA, and [3]Bren School of Environmental Science & Management, University of California, Santa Barbara, Santa Barbara, CA 91306, USA

Email:
Haoyang Haven Liu* - haven.liu@ucla.edu; Yoram Cohen* - yoram@ucla.edu

* Corresponding author

Keywords:
engineered nanomaterials; environmental exposure assessment; life cycle assessment; nanoinformatics; web-based simulation tool

Beilstein J. Nanotechnol. **2015,** *6,* 938–951.
doi:10.3762/bjnano.6.97

Received: 12 December 2014
Accepted: 23 March 2015
Published: 13 April 2015

This article is part of the Thematic Series "Nanoinformatics for environmental health and biomedicine".

Guest Editor: R. Liu

Abstract

An integrated simulation tool was developed for assessing the potential release and environmental distribution of nanomaterials (RedNano) based on a life cycle assessment approach and multimedia compartmental modeling coupled with mechanistic inter-media transport processes. The RedNano simulation tool and its web-based software implementation enables rapid "what-if?" scenario analysis, in order to assess the response of an environmental system to various release scenarios of engineered nanomaterials (ENMs). It also allows for the investigation of the impact of geographical and meteorological parameters on ENM distribution in the environment, comparison of the impact of ENM production and potential releases on different regions, and estimation of source release rates based on monitored ENM concentrations. Moreover, the RedNano simulation tool is suitable for research, academic, and regulatory purposes. Specifically, it has been used in environmental multimedia impact assessment courses at both the undergraduate and graduate levels. The RedNano simulation tool can also serve as a decision support tool to rapidly and critically assess the potential environmental implications of ENMs and thus ensure that nanotechnology is developed in a productive and environmentally responsible manner.

Introduction

Engineered nanomaterials (ENMs) are reported to be utilized in more than 1,000 commercial products owing to their unique size-related beneficial properties [1-4]. It is estimated that global ENM production levels will be in excess of 340,000 tons by 2016 [5]. Given the rapid growth of nanotechnology, it is critical to assess the potential impacts associated with ENMs

and thus to ensure that nanotechnology is developed in an environmentally compatible manner. In this regard, various environmental impact assessment (EIA) frameworks have been proposed [6], which all require knowledge of the potential environmental distribution of ENMs in addition to their potential toxicological effects. However, reported ENM source release rates, environmental monitoring data of ENM concentrations, as well as suitable ENM measurement techniques are presently scarce. Thus, computational models have been proposed as support tools to estimate ENM release rates [7,8] and potential environmental exposure concentrations [9-11].

It has been proposed that analysis of the multimedia environmental distribution and exposure concentrations of contaminants can be accomplished via a tiered approach [12]. A screening level assessment (tier-1 analysis) can be carried out based on multimedia compartmental models (MCMs) [12] to identify major exposure pathways and to monitor data gaps. In such analysis, the environmental entry, movement, and distribution of contaminants are described by a set of mathematical expressions. Specifically, MCMs require mechanistic quantification of intermedia transport rates (e.g., dry and wet deposition, sedimentation, dissolution) and rates of contaminant release to various environmental media. Typically, such a screening level analysis is expected to provide an order of magnitude (or better) assessment. Although MCMs have been developed to estimate non-steady-state (i.e., temporal dynamic) environmental concentrations of gaseous and dissolved chemical pollutants (e.g., Mend-Tox [13,14], CalTOX [15], TRIM.FaTE [16]), these are not directly applicable for ENMs. Unlike gaseous and dissolved chemical pollutants, for which interphase mass transport rates are governed by chemical potential (fugacity) driving forces that are constrained by thermodynamic equilibrium, the intermedia transport of ENMs is governed by physical transport processes of particulate matter. Therefore, a description of the environmental fate and transport of ENMs requires the particle size distribution (PSD) to be accounted for within the modeling framework, as well as the PSD dependence of the various transport processes. Higher tier analyses, which may include the use of detailed single medium models, can provide higher spatial resolution of the predicted ENM distribution for the studied region (in contrast to a regional average of ENM media concentrations). However, such an approach requires extensive site-specific geographical information and meteorological data for the target region (i.e., $\sigma\left(10^1\right)-\sigma\left(10^2\right)$ higher relative to the tier-1 approach [14]), and thus can be more complex and computationally demanding.

Irrespective of model complexity, an important factor in assessing the environmental multimedia distribution of ENMs is their release rates. In order to estimate ENMs release rates, life cycle inventory assessment (LCIA) based approaches have been developed to track the target ENM mass throughout its life cycle from production, through use, to final disposal and/or release into the environment. LCIA approaches are based on ENM production rates and empirical transfer coefficients that quantify the fraction of mass transferred between compartments (including technical compartments, such as waste processing facilities, as well as environmental compartments, such as air, water and soil) [7,8,17-19]. Although there are uncertainties in the LCIA approaches (primarily due to the inherent uncertainty in the estimated ENM production rates and intercompartmental transfer coefficients [7]), such methods are considered at present as being reasonably suitable for assessing potential ENM release rates [7,17]. There have also been attempts to extend LCIA-based methods to estimate the ENM media concentrations (e.g., via material flow analysis) [17-19] relying on empirically estimated media transfer coefficients under laboratory (i.e., not environmental) conditions. In the above methodology, estimated transport rates may violate constraints imposed by intermedia transport mechanisms [9]. A recently proposed approximate treatment for steady-state ENM multimedia concentrations was provided by SimpleBox4nano [11], which is yet to be validated against environmentally measured concentrations of particulate matter. This model considers a range of intermedia transport processes (including episodic events such as rain scavenging) as continuous processes, with constant rate coefficients throughout the simulation period. SimpleBox4nano also does not consider temporal variability of meteorological conditions or source releases, and processes such as wind resuspension, aerosolization, foliage washoff, and uptake by biological organisms are not included. It is stressed that SimpleBox4nano only considers the average particle size in each particle class (primary ENM (with size of 10 nm), ENM attached to colloids, and ENM attached to larger particles), while assuming an arbitrary value of 0.1 for both aggregation and attachment efficiencies [11]. As a consequence, the above approach does not account for the temporal dynamics of multimedia distribution and the strong dependence of ENM intermedia transport on the complete PSDs [9].

In earlier work, a multimedia environmental distribution of nanomaterials (MendNano) model was developed [9] based on a mechanistic description of various intermedia transport and reaction (including dissolution) processes, which considers the complete PSD of ENMs and ambient particulates. This study reported that dry and wet depositions (from air) are important intermedia transport pathways for ENM removal from the atmosphere and their input to the aquatic and terrestrial environments, the latter being particularly significant in the absence of direct ENM release to those compartments. Also, the dissolution of sparingly soluble ENMs in the water compartment can

be the dominant mechanism for removal of particulate ENMs from water. MendNano was also applied to the modeling of the environmental distribution of semi-volatile organics. These organics adsorb onto ambient particles [20,21] and thus their transport behavior is governed by the particle phase as is the case with ENMs [9,12]. Simulation results have demonstrated excellent agreement with environmental monitoring data to within a factor of 2 or better [9], which is an acceptable level for compartmental models [22-24].

Compartmental models can be used to provide a first-tier analysis for estimating the magnitudes of potential ENM exposure concentrations. However, in order to support timely decision analysis regarding the potential environmental impact of ENMs, it is imperative to make available integrated tools that enable rapid analysis. Accordingly, in the present work, an integrated simulation tool for estimating the potential release and the environmental distribution of nanomaterials (RedNano) was developed. This tool integrates MendNano [9] with a LCIA-based model for estimating ENM release rates [7,25]. RedNano is a simulation tool suitable for estimating the potential environmental ENM release and distribution, for performing multimedia scenario analysis, and for evaluating the significance of intermedia transport pathways. RedNano has been deployed as a web application and was developed as a modular system. Its structure and utility are demonstrated in the present study with a number of illustrative use cases.

Computational Modeling

Overview of RedNano simulation tool

RedNano consists of five main elements (Figure 1): (1) user interface for scenario design and results visualization, (2) MendNano, which is a fate and transport model for estimating environmental ENM concentrations, (3) lifecycle environmental assessment for release of nanomaterials (LearNano) model for estimating ENM release rates, (4) a parameter database, and (5) a repository for building a library of scenarios and simulation cases. The RedNano graphical user interface (GUI) provides guidance for scenario design and parameter specification; the latter may be obtained from an integrated parameter database, input manually, or calculated by various submodels. Based on the designed scenario, MendNano computes the multimedia mass distribution of ENMs given a release rate and/or initial concentration of the selected ENMs in one or more of the environmental compartments. Simulation results are then graphically represented via visualization modules as well as provided in standard numerical formats. Additionally, scenario input data as well as intermediary and final simulation results

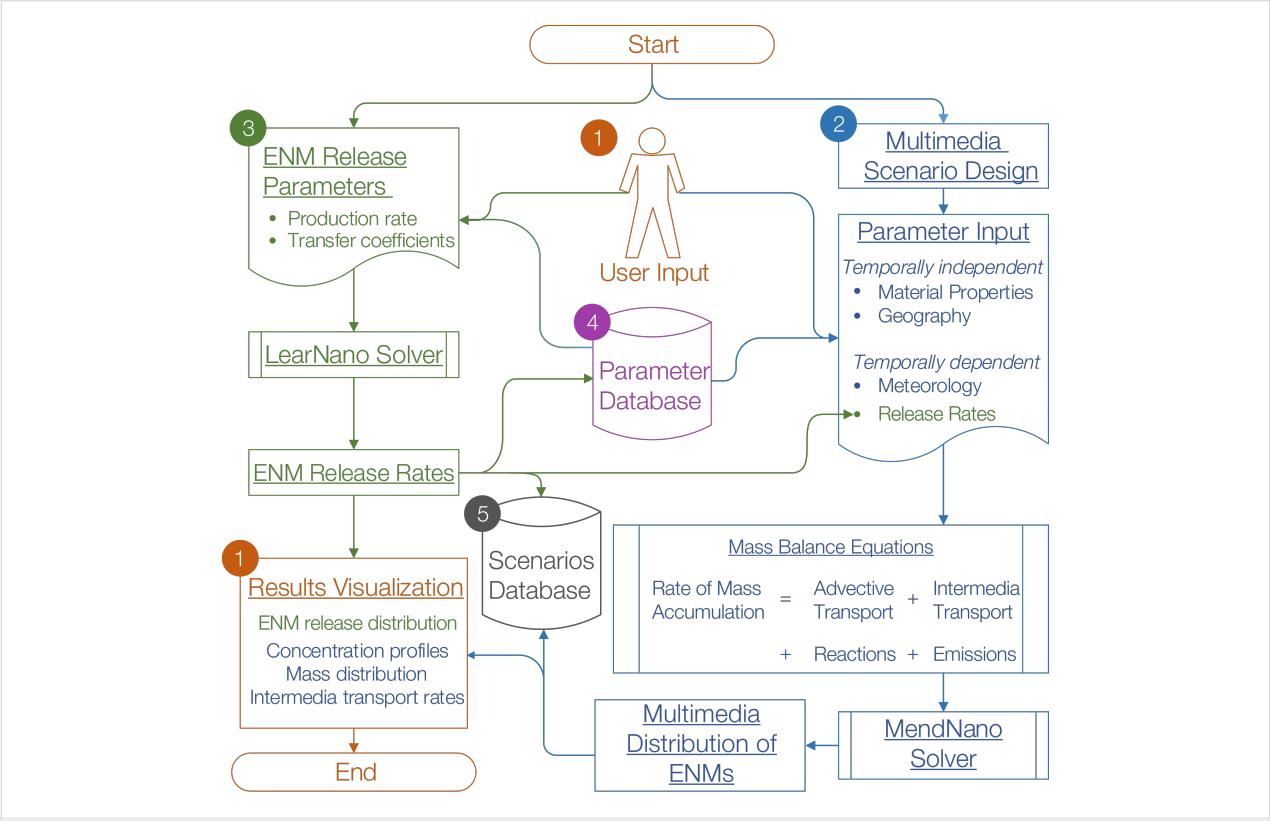


Figure 1: Overview of the release and environmental distribution of nanomaterials (RedNano) simulation tool and its components: (1) GUI, (2) MendNano, (3) LearNano, (4) parameter database, and (5) scenarios database.

are stored in the scenario database. The RedNano integrated simulation tool was designed as a client–server web application using a standard web development environment (i.e., HTML, PHP, JavaScript, MySQL).

MendNano

The theoretical basis describing the dynamic distribution of ENMs in the multimedia environment is provided in detail elsewhere [9]. Briefly, MendNano treats the multimedia environment as a set of well-mixed compartments (e.g., air, water, soil, sediment, biotas) linked via intermedia transport processes (ITP) meaning among compartments (e.g., dry/wet deposition, resuspension, sedimentation, dissolution) as listed in Figure 2. The resulting unsteady state, mass balance, ordinary differential equations (Supporting Information File 1, Equation S1) are then solved to obtain the mass of the ENMs in the various environmental compartments, and thus the temporal evolution of their mass distribution, concentration, and intermedia transport rate. Intermedia transport rates are specified by mechanistic transport processes, and are governed by geographical and meteorological parameters, as well as material properties. The compartmental modeling approach, which is generally suitable for regional assessments [26-28] of a minimum area of 1 km^2 [12], lends itself to screening level analysis. Spatial resolution, however, may be increased by using nested or subcompartments, as well as via hybrid approaches that integrate spatial and well-mixed compartments [14]. In addition, the simulation time should be greater than the longest convective residence time in the model compartments (e.g., hours to days for air and water, respectively [12]). MendNano accounts for the complete PSD of both ENMs and ambient particulates by discretizing the PSD into bins, and the association of ENMs with ambient particulates is described by an attachment factor [9]. The PSD of ambient particulates is typically taken to be self-preserving [29-33], but may be altered when there is significant removal (e.g., during precipitation events). The PSD of ENMs may also be altered in a given compartment as the result of intermedia transport processes such as dry and wet deposition from the atmosphere, gravitational settling in aqueous systems, as well as dissolution and reaction processes in air and water (Figure 2).

MendNano includes modules for: (a) mechanistic submodels for rates of intermedia transport processes [9,12], (b) dynamic compartmental mass balance equations consisting of a set of 50–204 (depending on the user-specified scenario) ordinary differential equations (ODEs), (c) event tracking (for episodic events, e.g., precipitation, wind resuspension), and (d) an ODE solver. The modular construction of MendNano allows for adding/upgrading compartments and transport submodels as new information becomes available (e.g., biological compartments and associated uptake mechanisms). The compartmental mass balance ODEs (Supporting Information File 1, Equation S1) are solved via the Adams–Bashforth–Moulton predictor–corrector method [34], with time steps dynamically selected to achieve the numerical solution error (in terms of compartmental ENM mass) set with 0.1% relative error tolerance (defined as percent change in two consecutive solutions). At each time step, the rates of advective (i.e., via air and water flow) and intermedia transport, reactions, and source release are computed based on the temporally varying parameters

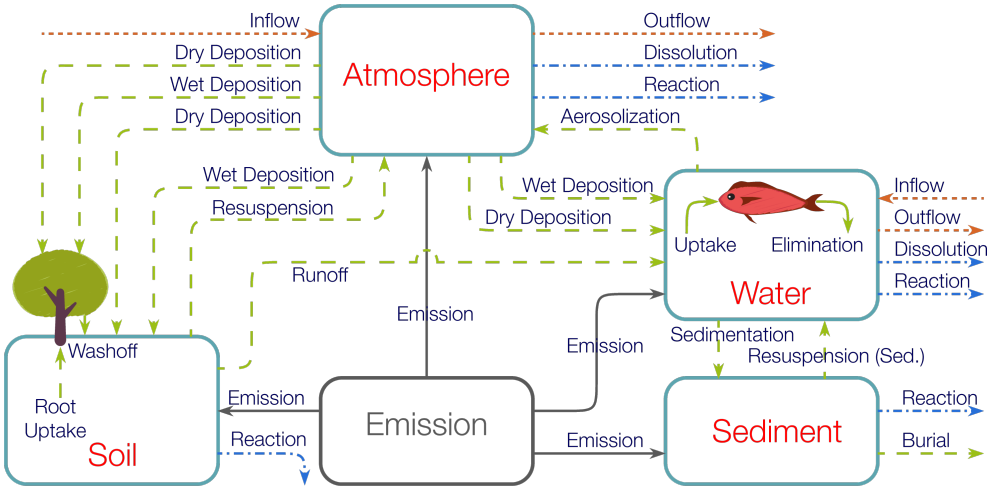


Figure 2: Transport processes in MendNano. Green dashed lines represent intermedia transport processes, blue dash-dot lines represent reactions (including dissolution) within the compartments that eliminate the ENM from particle phase, orange dotted lines represent advection (i.e., transport of ENMs via the flow of air and water) into and out of the given compartment, and gray solid lines represent emissions (i.e., ENM release events into the compartments).

(e.g., wind speed, temperature, biological organism mass, ENM release rates).

LearNano

Estimation of the ENM release rates can be accomplished by the LCIA modeling approach as described in detail elsewhere [7,17]. Briefly, in LCIA-based models, reported ENM mass production rates [5] are allocated to the various ENM applications (e.g., paints, cosmetics, electronics, catalysts), waste processing facilities (i.e., technical compartments), and eventually environmental compartments (Figure 3) [7,17]. Transfer coefficients, which are dependent on the ENM type, ENM application, and region under consideration [7,17], then serve to quantify the fraction of ENMs entering the "source" compartments that are subsequently transferred to the "target" compartment (Figure 3). Accordingly, a series of algebraic mass balance equations that describe ENM mass release rates related to the various environmental compartments [7,17] are incorporated in LearNano (Supporting Information File 1, Equations S2–S4).

Implementation of the LearNano model includes user guidance and visualization tools for data input and simulation results, a model solver, and a parameter database. The analysis scenario (i.e., a given combination of ENM, region, and application(s)) is constructed within the GUI, which also captures ENM production rates and the various transfer coefficients between adjoining compartments (both technical and environmental). ENM production rates and transfer coefficients can be obtained from a parameter database by specifying the ENM(s), application(s), and region(s) of interest (see section, Databases). The mass balance equations (Supporting Information File 1, Equations S1–S4) are then solved to determine the average ENM release rates to the environmental compartments (i.e., air, water, and soil). Mass "flows" of ENMs among the various compartments can be visualized using a dynamic and interactive Sankey diagram (Figure 4). Also, the global distribution of ENM release (to various environmental compartments) in different countries can be represented on a world map (Figure 5). It is noted that, while the present version of LearNano computes ENM release rates on a country level, estimates of regional ENM release rates may be obtained by scaling country level release rates on the basis of population, area, or economic indicators [7,17].

Graphical user interface (GUI)

The web-based GUI for RedNano enables building multimedia scenarios, initiating model execution, as well as visualization of simulation results. A multimedia scenario refers to the specification of a model environment (i.e., geographical region and its meteorology), the target ENM, and its release rate. A multimedia scenario is built by specifying or selecting the

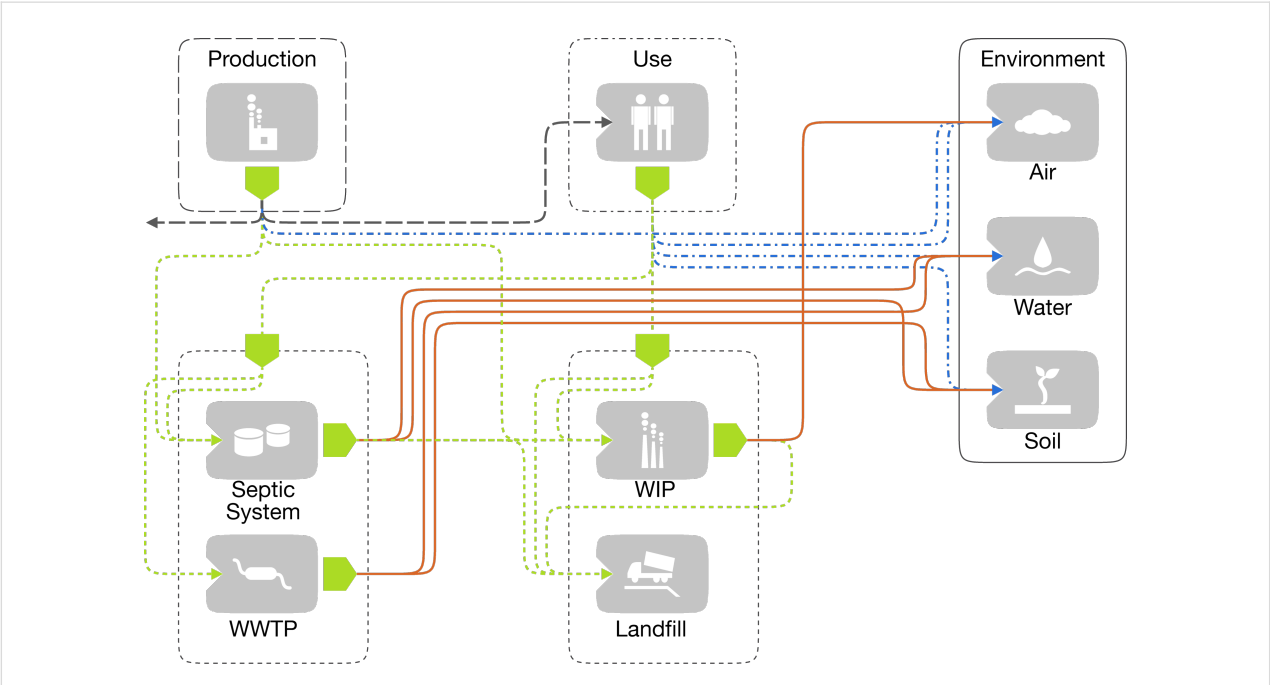


Figure 3: Lifecycle tracking of ENMs. The various lines represent the paths for which transfer coefficients quantify the portion of ENMs transferred from the source to the target compartments. Blue dash-dot lines represent direct release to environmental compartments from production and use, green dotted lines represent ENM transfer from production and use to waste processing facilities, orange solid lines represent indirect release to environmental compartments from waste processing facilities, and gray dashed lines represent import/export and ENM transfer from production to phase.

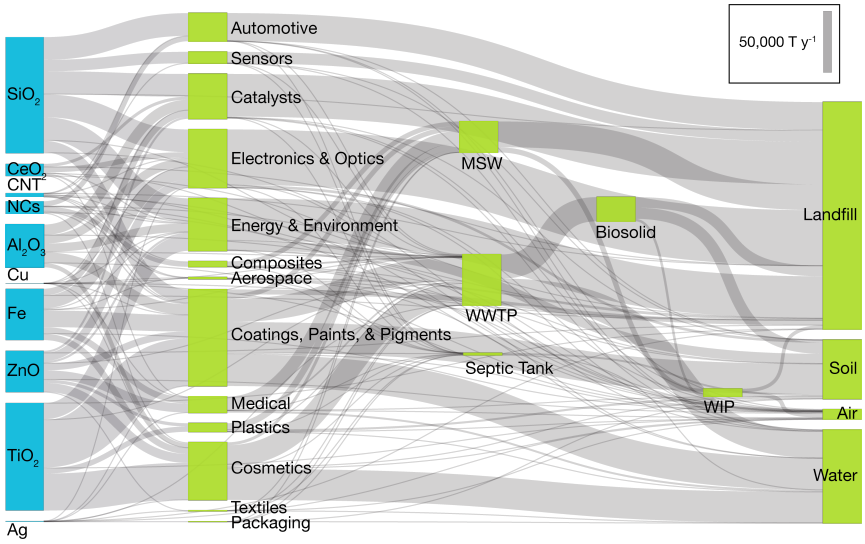


Figure 4: Sankey diagram depicting the flows of different ENMs from production and use, through technical compartments, to disposal and release to the environment. The vertical size of the bars and thickness of the links represent the magnitude of the ENM mass transfer rate.

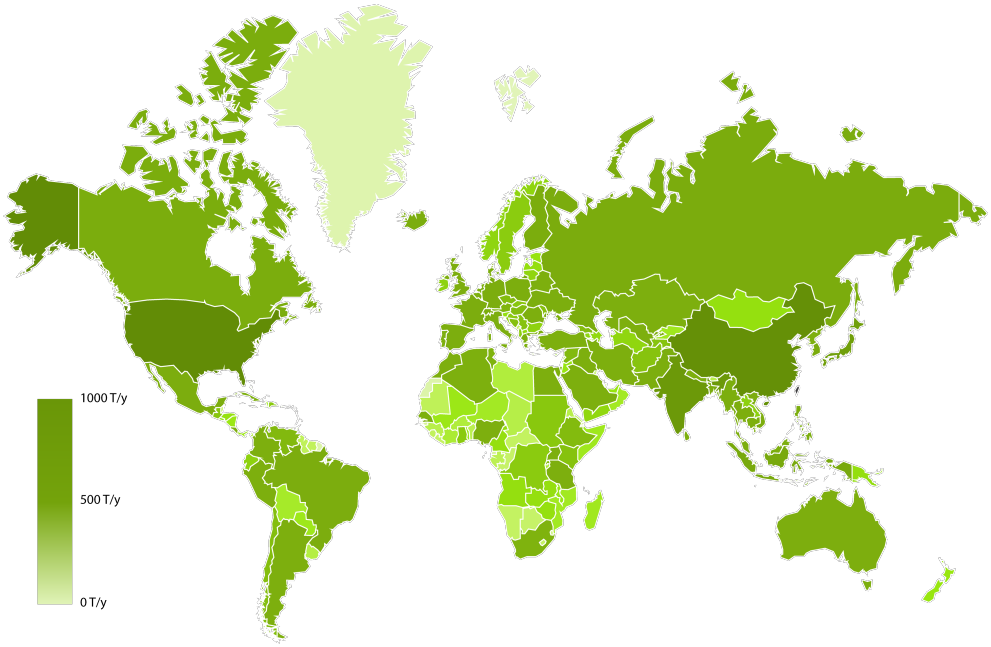


Figure 5: Example of the global distribution of the release rates of TiO_2 into water.

required parameters from modules that include: (a) geography, (b) meteorology, (c) material properties, and (d) source release (Figure 6).

Scenario design is initiated by selecting the environmental compartments (e.g., air, water, soil, sediment, vegetation canopy, biota) and ITPs (e.g., dry/wet deposition, resuspension, sedimentation, dissolution) of interest for the desired simulation period (typically ≈1 year) and the target ENM and its properties (Figure 6). Subsequently, submodels are selected for the specified ITPs (Figure 2) and the regional geographical and meteorological parameters are specified for the selected region (Figure 6). The values for these parameters may be obtained from the system's parameter database, or can be provided by

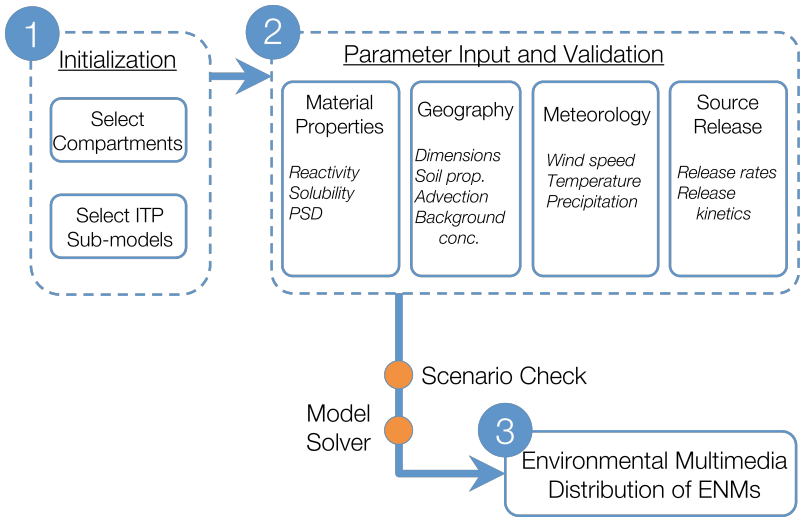


Figure 6: Workflow for assessing the environmental distribution of ENMs. ITP: intermedia transport processes, PSD: particle size distribution.

the user. ENM release rates to the various compartments are also required and these can be obtained from LearNano by selecting the target ENM, region, and applications of interest, or specified directly by the user (Figure 6). The temporal profile of the ENM release rate kinetics can be specified as constant or periodic sinusoidal (e.g., to mimic seasonal and diurnal variability).

The specification of the required parameter values is accomplished in a series of web pages (or views; Figure 7) within the GUI corresponding to the modules shown in (Figure 6). The parameter input is validated, prior to model execution, to ensure that the specified values are within a reasonable range and/or constraints (e.g., minimum regional area, maximum rainfall intensity). Additional simulation scenario validation is also conducted to ensure that scenarios are not ill-defined (e.g., simulation with neither source release nor initial compartmental concentration). Upon simulation scenario design completion, model execution is initiated (a unique Simulation ID is assigned for compilation of a scenario library). The results can then be visualized via a series of graphical representations. The dynamic multimedia ENM distributions can be represented as: (a) ENM temporal concentration (or mass) profiles in various compartments (Figure 8), (b) intermedia mass transport rates or fluxes, (c) ENM mass distribution (percent) among the various compartments, (d) ENM apportionment throughout the ambient particle size distribution (Figure 8), and (e) the magnitude of intermedia transport rates, as a fraction of the ENM release rates, that allows assessment of the relative significance of various intermedia transport processes (Supporting Information File 1, Figure S5). For example, in the illustration of Figure 8, ENM concentrations in air and water (left upper plot) rapidly reach pseudo-steady state, except during episodic rain events, in which a sharp decrease in ENM concentration in air is observed, followed by a rapid increase after the rain event. In contrast, ENM concentrations in soil and sediment continue to increase, since ENM removal rates from soil and sediment are significantly lower than the rate of ENM entering the soil and sediment. Given these considerations and that the ENM release rate to water was greater relative to air (Supporting Information File 1, Figure S5, Table S5), the majority of ENM mass accumulated in the sediment (right upper subplot). The ENM mass distribution in air among the particle size fractions of ambient aerosol is shown to follow the expected tri-modal distribution (lower subplot). It is noted that such information can be utilized to convert MendNano reported ENM mass concentrations to surface area concentration [35,36] given the knowledge of the primary particle size.

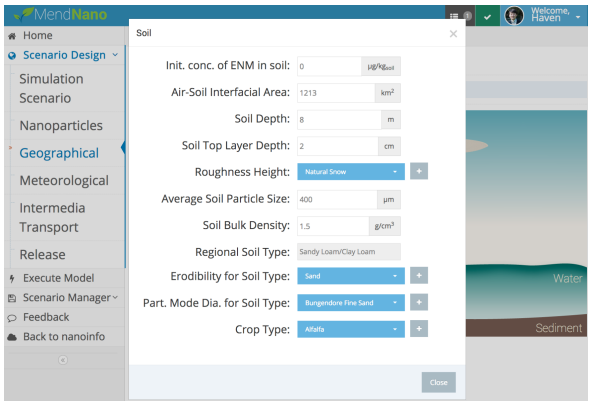


Figure 7: Examples of MendNano web-based graphical user interface for scenario building showing inputs of soil parameters.

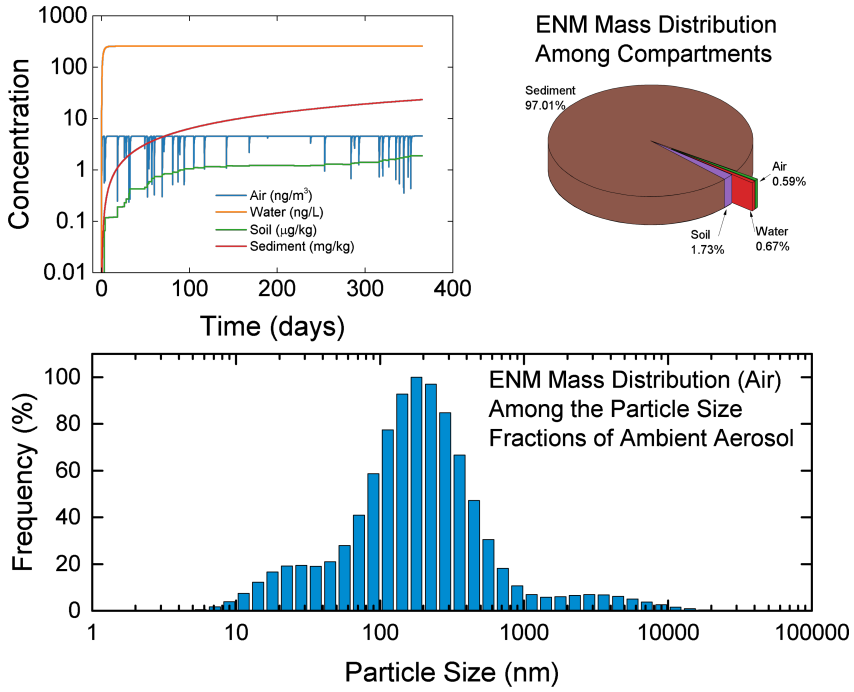


Figure 8: Examples of graphical representations of MendNano simulation results depicting concentration profiles and mass distributions of TiO_2 in the Los Angeles region among the various compartments and among the ambient particles in air. Release of TiO_2 in the above example is in air (5,000 kg yr^{-1}) and water (19,381 kg yr^{-1}).

Table 1: Parameters database.

Category	Subcategory	Property[a]
Material properties		PSD (ENM and aerosol)
Geographical parameters	Physical description	Interfacial Area (air–water, air–soil)
		Mixing height
		Water depth
		Water flow rate
		Average suspend solids diameter
		Sediment depth
		Soil depth
	Dry deposition to vegetation	Roughness factor
		Characteristic field length
		Crop vegetation factor
	Dry deposition to soil	Roughness height
	Wind resuspension of soil	Soil erodibility
Meteorological parameters		Monthly Temperature (air, water)
		Wind speed (monthly, annual average, max)
		Rainfall rate (monthly)
LearNano parameters		ENM Global production rate
		Transfer coefficients (ENM specific)
		Transfer coefficients (application specific)
		Transfer coefficients (region specific)

[a]Additional parameters, including those calculated internally by the model, are provided in Supporting Information File 1, Table S1.

Databases

The parameter database contains material properties, geographical, and meteorological parameter values (Table 1), which are compiled from various literature and database sources [31,37-39]. The parameter database also includes a library of ENM production rates and transfer coefficients corresponding to specific ENMs and applications, for different geographic regions (Table 1), compiled from various published studies [17], public databases [40], and market research [5], and estimated based on economic indicators [41]).

Use cases for assessing multimedia distribution of ENMs

The integrated RedNano simulation tool is suitable for a variety of assessments regarding the environmental distribution of ENMs and their fate and transport behavior. These assessments can be classified into use cases that include, but are not limited to, the following:

1. Environmental ENM concentrations and mass distribution based on a specified multimedia scenario;
2. Dynamic response of the environmental system to temporally varying ENM release rates;
3. Impact of specific intermedia transport processes on the temporal dynamics of ENM distribution in the environment;
4. Comparison of estimated environmental ENM concentrations in various regions;
5. Contribution by ENM applications (or use) to the overall ENM releases and exposure concentrations in the various environmental compartments;
6. Estimation of source release rates, based on matching of model estimates and reported environmental concentrations.

Results and Discussion

In order to demonstrate the above use cases, illustrative simulations were conducted to estimate the environmental distributions of TiO_2, CeO_2, SiO_2, and CNT in selected regions. The multimedia distribution of ENMs (use case #1) and the dynamic response of an environmental system to temporal variations of ENM release rate (use case #2) are illustrated for TiO_2 in Los Angeles. Due to a lack of transfer coefficients specific to Los Angeles, TiO_2 release rates for Los Angeles were estimated by scaling from US release rates on the basis of a population ratio. TiO_2 release rates to air and water were taken to follow a sinusoidal release function with a cycle period of 100 days, where the release rates fluctuated between 0 to 27.4 and 0 to 106.2 kg day^{-1}, for release into air and water, respectively, and were terminated thereafter. The results, as shown in Figure 9, indicate that TiO_2 concentrations in air and water fluctuate between 3.3–4.4 ng m^{-3} and 195–267 ng L^{-1}, respectively, representing an ≈15% deviation (in both media) above and below the time-averaged concentration in the respective compartments. Following cessation of source release into air and water (at t = 100 days), the TiO_2 concentration in both compartments decreased rapidly (Figure 9) to 90% of the levels just prior to the termination of the release in ≈1 day and ≈4 days, respectively. The TiO_2 concentrations continued to decrease until a pseudo-steady state was reached in air and water, within ≈4 and ≈38 days, respectively. Although ENM release into air and water ceased after 100 days, the ENM concentrations in these compartments did not vanish since ENMs in the soil (accumulated during the first 100 days) continued to be transported to air via soil–wind resuspension, and subsequently deposited to water via dry and wet deposition.

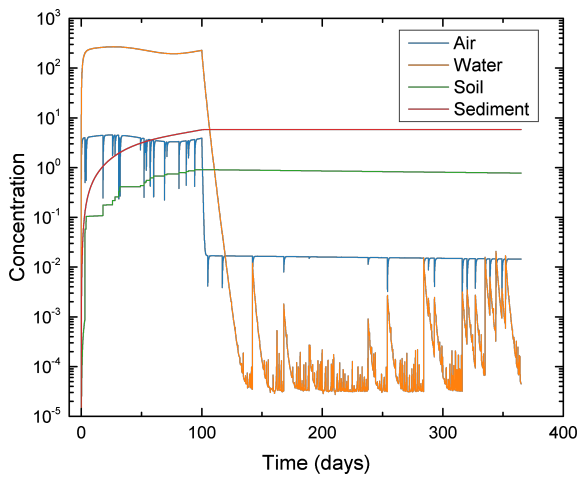


Figure 9: Effect of release scenario on temporal dynamics of TiO_2 media concentrations in Los Angeles. TiO_2 release rates to air and water were obtained from LearNano (Supporting Information File 1, Table S5). The ENM release rates (into air and water) followed a sinusoidal function for the first 100 days (cycle period of 100 days, amplitude of 13.7 and 53.1 kg/day, for releases to air and water, respectively), after which the source releases are terminated. Regional geographical parameters are reported in Supporting Information File 1, Table S4.

The impact of specific intermedia transport processes on the temporal dynamics of the ENM distribution in the environment (use case #3) is highlighted via a series of simulations for TiO_2 in Los Angeles focusing on intermedia transport via dry deposition, rain scavenging, and wind dilution (Supporting Information File 1, Figure S1). In these scenarios, the initial TiO_2 concentration in air is taken to be the steady state TiO_2 concentration reached after 1 year with all other compartments being initially free of TiO_2.

Dry deposition is a process in which particles (including ENMs) are collected onto terrestrial (e.g., soil, vegetative canopy) and

aquatic surfaces due to Brownian diffusion, impaction, and interception [42]. The intermedia transport rate due to dry deposition is a function of wind speed (among other parameters, e.g., surface roughness), which is typically reported to be 3.3 ± 0.95 m s^{-1} (1 standard deviation for 1996–2006) [43], with a maximum of ≈10 m s^{-1} in the Los Angeles region (LAX station). An increase in wind speed would lead to an increase in the rates of collection by impaction and interception [42], and thus an increase in the overall rate of dry deposition. The predicted temporal ENM concentration profiles in air and soil (Figure 10) reveal that the time to remove 90% of TiO_2 by dry deposition alone is ≈100–230 days for wind speed in the range of 2.7–10 m s^{-1}. Additionally, at the end of a 1 year simulation, 0.1–3.4% of the initial ENM mass in air remains in the air compartment for this wind speed range.

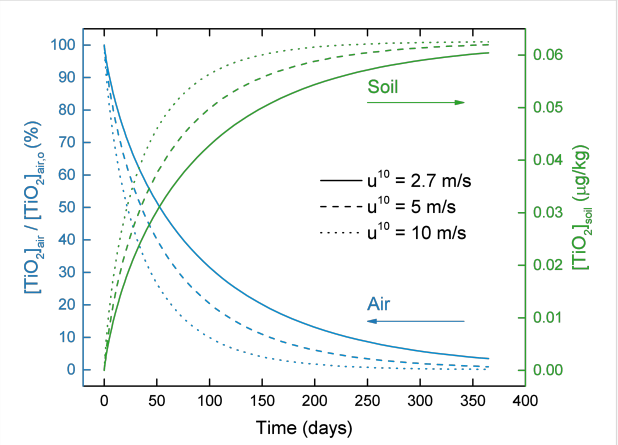


Figure 10: Effect of dry deposition on the reduction of TiO_2 concentrations in air and soil (postcessation of all ENM releases) in Los Angeles as a function of wind speed (range of 2.7–10 m s^{-1}). Regional geographical parameters are reported in Supporting Information File 1, Table S4.

Rain scavenging of particulate matter (including ENMs) by raindrops results in the removal of particulate matter from the atmosphere and its subsequent deposition onto terrestrial and aquatic surfaces. The ENM removal rate by rain scavenging is governed by rainfall intensity (typically in the range of 1–10 mm h^{-1} for light to moderate rain [44], and can exceed 50 mm h^{-1} for intense storms [45]). Rain scavenging can typically remove atmospheric particles at a faster rate relative to dry deposition. As illustrated in Figure 11, even with a mild rainfall intensity of 1–5 mm h^{-1}, 90% of TiO_2 can be removed in hours (i.e., ≈2–6 h, corresponding to a rainfall intensity of 5–1 mm h^{-1}), compared to many days for removal by dry deposition (Figure 10). Since rain scavenging is an episodic process (in contrast to the continuous dry deposition), the annually averaged ENM removal rate by rain scavenging is expected to be lower than the instantaneous removal rate during rainfall events as shown in Figure 11. Nonetheless, the averaged transport rate by rain scavenging can exceed that by dry deposition. For example, in Los Angeles, the estimated annually averaged TiO_2 removal by rain scavenging is a factor of ≈10 greater than by dry deposition (Supporting Information File 1, Figure S5), indicating that rain scavenging has a more significant impact on the environmental ENM distribution relative to dry deposition.

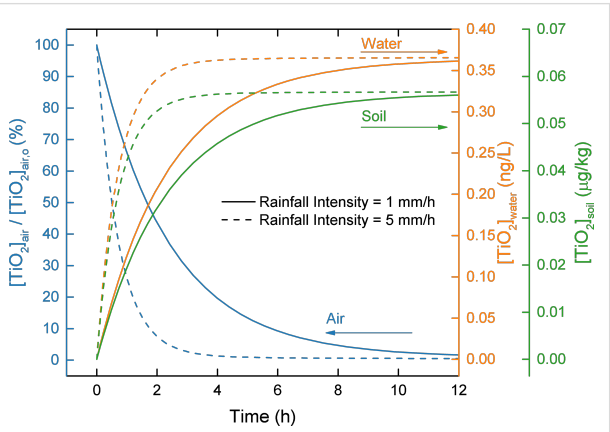


Figure 11: Effect of rain scavenging on TiO_2 concentration in air, water, and soil in Los Angeles as a function of rainfall intensity (1–5 mm h^{-1}). All ENM release rates are terminated at the start of a long rain event, which was taken to last for 12 h. Regional geographical parameters are reported in Supporting Information File 1, Table S4.

A comparative analysis of the potential environmental ENM concentrations in various countries (use case #4) is given using the example of CeO_2 ENMs, whereby release rates were estimated via LearNano for 12 selected countries. These countries were selected to represent the high ENM producing (and high emission) regions. The estimated CeO_2 release rates (high estimate) for the 12 countries span over the range of 7.2–486 T yr^{-1} for Chile and China (Figure 12). The high estimates for the release rates for the 12 countries are, on average, a factor of ≈12 greater than the low estimates, with the highest difference being by a factor of 86 (e.g., for release to water in Switzerland). The release rates into air, water, and soil represent, on average for the different countries, 10% (3–40%), 38% (33–46%), and 52% (24–60%) of the total release rates, respectively (Supporting Information File 1, Figure S2). The above analysis suggests that while some differences exist in apportionment of total release to various compartments between countries, the majority of ENM release events are into water, followed by soil and air. It should be noted that among the total ENM release to soil, only the direct release portion (≈79%, which excludes release from WWTP biosolids) may be considered to be distributed over the entire soil area in the region. The distinction between direct release to soil and that from WWTP biosolids is important. Although biosolids are applied to some

agricultural lands in the USA, the USEPA estimates that <1% of agricultural lands receive biosolids [46], which suggests that the application of biosolids to soil does not represent a wide spread release in the USA. Similarly, it has been reported that in Switzerland, biosolids are not applied to soil, and are instead processed in waste incineration plants [17].

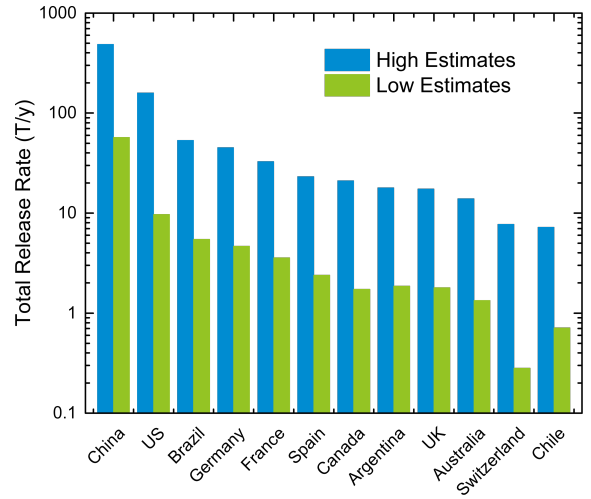


Figure 12: Estimated CeO_2 release rates for 12 selected countries.

The compartmental concentrations of CeO_2 for the 12 countries were estimated via MendNano using the release rate estimates shown in Figure 12, and country specific geographical and meteorological conditions (Supporting Information File 1, Table S3). The simulations were carried out assuming that only direct release to soil is regionally distributed. The predicted CeO_2 concentrations using the high release rates estimates are in the range of 0.0003–0.097 ng m^{-3}, 0.0058–2.7 ng L^{-1}, 0.0095–0.74 µg kg^{-1}, and 0.0054–0.25 mg kg^{-1} for air, water, soil, and sediment, respectively (Figure 13). Relative to these predictions, the CeO_2 concentrations predicted using the low release rates estimates are a factor of 5–1243 lower (Supporting Information File 1, Figure S3). Clearly, there is a large uncertainty in the estimated media concentrations due to uncertainties in ENM release estimates. Nonetheless, it is noted that the above predicted CeO_2 concentration range is significantly below concentrations typically used in experimental toxicity studies [47].

It is interesting to note that while the USA ranks second highest in terms of release rates (for all compartments), it ranks 7th (out of 12) in terms of CeO_2 concentration in air and soil, and 11th based on concentration in water and sediment. In contrast, while the UK and Switzerland rank 9th and 11th with respect to total release rates, respectively, they rank first (i.e., highest) in terms of the compartmental concentrations in air and water, respectively. Additionally, the environmental concentrations in the European countries are all significantly higher than that in the US (by a factor of 1.4–15), despite having total release rates that are lower than the USA (by a factor of 3.5–20). The apparent resulting discrepancy between release and environmental concentrations is attributed to differences in geography and meteorology. For example, Supporting Information File 1, Figure S4 shows that the release rate into air per unit area (combined soil and water) in Switzerland is a factor of 17 greater than in the US; similarly, release rates into water per unit area in the UK are a factor of 46 greater than in the US.

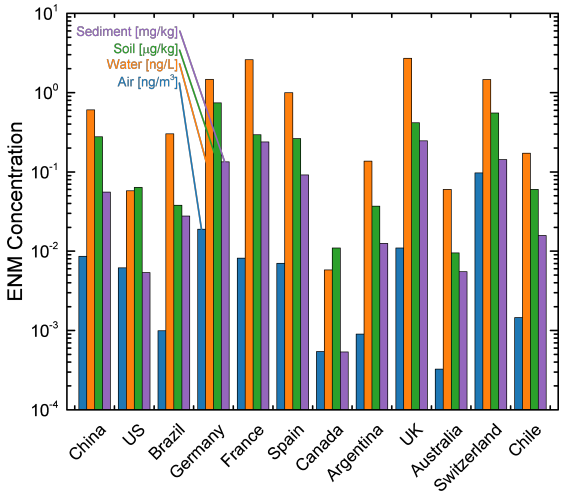


Figure 13: Predicted compartmental concentrations for CeO_2 in 12 selected countries at the end of a 1 year simulation for the ENM release rates reported in Figure 12**.** Regional geographical and meteorological parameters are reported in Supporting Information File 1, Table S4.

The contribution of ENM release rates by various ENM applications (or use) to the overall ENM release and exposure concentrations in the various environmental compartments (use case #5) is shown in the example of Figure 14 and Supporting Information File 1, Figure S6. For Los Angeles, the simulations were carried out for TiO_2 and SiO_2, which were selected since these are produced in the largest quantity [7], and CNT was included due to its diverse applications [7]. The TiO_2 release rates attributed to coating, paint, and pigment applications are the primary contributors of the release of this ENM into air (≈45%) and soil (≈77%). In water, TiO_2 release is associated with cosmetic applications, which represent the largest fraction at ≈53%, while those associated with coatings, paints, pigments represent ≈44%, with remainder due to energy applications (e.g., photovoltaics, energy storage [7]), environmental (e.g., remediation [7]), and plastic applications. These results are consistent with reported TiO_2 use in coatings, paints, and pig-

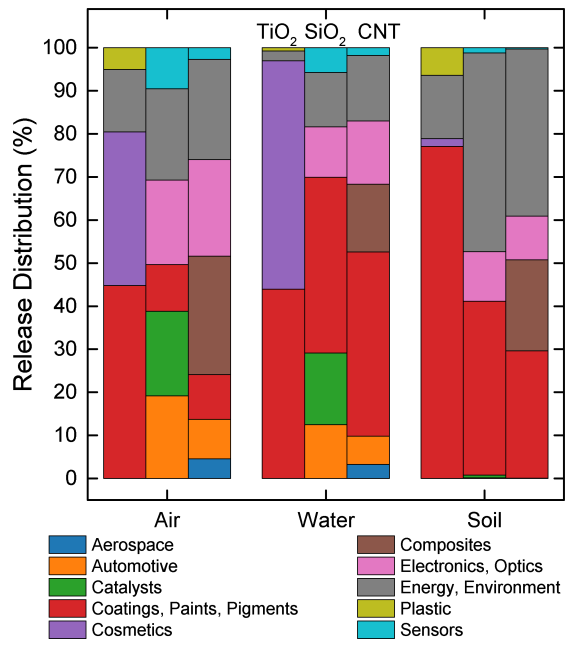


Figure 14: Apportionment of environmental release rates of selected ENMs to specific compartments in the Los Angeles region to different ENM applications.

ments and associated release into the environment due to weathering [48] and TiO_2 used in cosmetics is primarily released during washing into waste water [49]. The release of SiO_2 into air (Figure 14) associated with energy and environmental applications is the largest fraction (≈21%), while other applications (i.e., automotive, catalysis, coatings/paints/pigments, electronics/optics, and sensors) contribute less, but still a significant amount (9.5–19.6%). In contrast, the release of SiO_2 into soil is dominated by energy and environmental applications, and the group of coating, as well as paint and pigment applications (46% and 40%, respectively), while other applications collectively contribute less than 14% of the total SiO_2 release to soil. The most significant contribution to SiO_2 released into water is also associated with coating, paint, and pigment applications (≈41%). Finally, the largest contributions to the release of CNTs into air, water and soil are associated with composites (≈28%), coatings, paints and pigments (≈43%), and energy and environmental applications (≈40%), respectively.

The contributions of the various ENM applications to compartmental concentrations (Supporting Information File 1, Figure S6) are, as expected, typically qualitatively similar to their contributions to the ENM release rates shown in Figure 14. However, noticeable differences can be observed in some cases due to intermedia transport of these ENMs from soil to air. For example, an ENM associated with a given ENM application can be transported to the air compartment via soil–wind resuspension in larger portion relative to other applications. Thus, increased ENM concentration in air may occur for that application. Such a behavior can be expected when an ENM application contributes to the ENM release to soil in larger proportion relative to its contribution to ENM release to air. The above behavior is demonstrated in Supporting Information File 1, Figure S6 for TiO_2, for which the release associated with coatings, paints and pigments contributes ≈45% to the total TiO_2 release to air while contributing ≈77% of total TiO_2 release to soil (Figure 14). As a result, ≈54% of the TiO_2 mass concentration in air is attributed to releases associated with coatings, paints, and pigments. In contrast, when 36% of the total TiO_2 release to air is associated with cosmetics applications, and only 1.8% of total TiO_2 release to soil is associated with cosmetics, less than 28% of the TiO_2 mass concentration in air is related to this category of ENM application. Therefore, since wind resuspension from soil may be a significant transport pathway of ENMs into the air compartment, the apportionment of the total ENM release to soil associated with the various applications may have a notable impact on the contribution of ENM application to its concentrations in air.

The estimation of ENM release rates, based on reported environmental ENM concentrations (use case #6), can be accomplished as described in the example of simulations of CeO_2 environmental distribution in Newcastle (UK). In this example, the release rate of CeO_2 ENMs from fuel additives in Newcastle was estimated based on matching reported atmospheric concentrations before and after the introduction of the fuel additive with MendNano simulation results. Monitoring the results showed that following the introduction of Envirox (a CeO_2 ENM-based diesel fuel combustion catalyst) to a bus fleet in the Newcastle area, the ambient CeO_2 concentration increased by a factor of ≈4.2 (0.574 ng m^{-3}, from 0.145 to 0.612 ng m^{-3}) [50]. MendNano simulations carried out considering the geographical and meteorological scenario setup for the Newcastle region revealed that a CeO_2 release rate of 43.96 kg yr^{-1} would result in the reported increased CeO_2 concentration. The MendNano estimate of the CeO_2 release rate is consistent with the release rates estimated based on: (a) vehicle miles travelled (VMT) and (b) the diesel fuel consumption rate in the region of Northumberland, which is in proximity to Newcastle and of similar population (Supporting Information File 1, Estimation of CeO_2 Release Rates in Newcastle UK by VMT and Diesel Fuel Consumption). The estimated CeO_2 release rates for the above two cases are 21.48 and 44.82 kg yr^{-1}, respectively.

Applications and Merits

In summary, an integrated release and environmental distribution of nanomaterial (RedNano) simulation tool was developed

and implemented as a web-based application to enable rapid "what-if?" scenario analysis. The RedNano simulation tool is suitable for both research as well as educational purposes, and can be utilized in both undergraduate and graduate level courses for multimedia environmental assessment. It is envisioned that the present multimedia analysis platform can assist regulators, industry, and researchers to rapidly assess the potential environmental implications of ENMs that may be released into the environment.

Supporting Information

Supporting Information File 1

Additional equations and results regarding the model equations, intermedia transport factors, use cases, and parameters used for simulations carried out in the study.
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-97-S1.pdf]

Acknowledgements

This work was supported, in part, by the National Science Foundation and the Environmental Protection Agency under Cooperative Agreement Number DBI 0830117, the UCLA Water Technology Research Center and the California Department of Water Resources. Any opinions, findings, conclusions or recommendations expressed herein are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Environmental Protection Agency. This work has not been subjected to an EPA peer and policy review. The study was performed, in part, in a renovated collaboratory by the National Science Foundation under Grant No. 0963183, which is an award funded under the American Recovery and Reinvestment Act of 2009 (ARRA).

References

1. Guo, Z.; Tan, L. *Fundamentals and Applications of Nanomaterials,* 1st ed.; Artech House Publishers: London, United Kingdom, 2009.
2. Handy, R. D.; Owen, R.; Valsami-Jones, E. *Ecotoxicology* **2008,** *17,* 315–325. doi:10.1007/s10646-008-0206-0
3. Klaine, S. J.; Alvarez, P. J. J.; Batley, G. E.; Fernandes, T. F.; Handy, R. D.; Lyon, D. Y.; Mahendra, S.; McLaughlin, M. J.; Lead, J. R. *Environ. Toxicol. Chem.* **2008,** *27,* 1825–1851. doi:10.1897/08-090.1
4. Wiesner, M. R.; Lowry, G. V.; Alvarez, P.; Dionysiou, D.; Biswas, P. *Environ. Sci. Technol.* **2006,** *40,* 4336–4345. doi:10.1021/es062726m
5. Future Markets, Inc., *The global market for nanomaterials 2002-2006: production volumes, revenues, and end use markets*, Bristol, United Kingdom, 2012.
6. Cohen, Y.; Rallo, R.; Liu, R.; Liu, H. H. *Acc. Chem. Res.* **2013,** *46,* 802–812. doi:10.1021/ar300049e
7. Keller, A.; McFerran, S.; Lazareva, A.; Suh, S. *J. Nanopart. Res.* **2013,** *15,* No. 1692. doi:10.1007/s11051-013-1692-4
8. Sun, T. Y.; Gottschalk, F.; Hungerbuhler, K.; Nowack, B. *Environ. Pollut.* **2014,** *185,* 69–76. doi:10.1016/j.envpol.2013.10.004
9. Liu, H. H.; Cohen, Y. *Environ. Sci. Technol.* **2014,** *48,* 3281–3292. doi:10.1021/es405132z
10. Praetorius, A.; Scheringer, M.; Hungerbühler, K. *Environ. Sci. Technol.* **2012,** *46,* 6705–6713. doi:10.1021/es204530n
11. Meesters, J. A. J.; Koelmans, A. A.; Quik, J. T. K.; Hendriks, A. J.; van de Meentt, D. *Environ. Sci. Technol.* **2014,** *48,* 5726–5736. doi:10.1021/es500548h
12. Mackay, D. *Multimedia environmental models: the fugacity approach,* 2nd ed.; CRC Press: Boca Raton, FL, U.S.A., 2001. doi:10.1201/9781420032543
13. Cohen, Y.; Cooter, E. J. *Pract. Period. Hazard., Toxic, Radioact. Waste Manage.* **2002,** *6,* 87–101. doi:10.1061/(asce)1090-025x(2002)6:2(87)
14. Cohen, Y.; Cooter, E. J. *Pract. Period. Hazard., Toxic, Radioact. Waste Manage.* **2002,** *6,* 70–86. doi:10.1061/(asce)1090-025x(2002)6:2(70)
15. McKone, T. E. *CalTOX, a multimedia total exposure model for hazardous-waste sites; Part 1, Executive summary;* California Environmental Protection Agency: Sacramento, CA, USA, 1993. doi:10.2172/139702
16. Office of Air Quality Planning and Standards, USEPA, *Total Risk Integrated Methodology, TRIM.FaTE Technical Support Document, Volume I: Description of Modules*, USEPA: Research Triangle Park, NC, 2002.
17. Gottschalk, F.; Sonderer, T.; Scholz, R. W.; Nowack, B. *Environ. Sci. Technol.* **2009,** *43,* 9216–9222. doi:10.1021/es9015553
18. Gottschalk, F.; Scholz, R. W.; Nowack, B. *Environ. Model. Software* **2010,** *25,* 320–332. doi:10.1016/j.envsoft.2009.08.011
19. Gottschalk, F.; Sonderer, T.; Scholz, R. W.; Nowack, B. *Environ. Toxicol. Chem.* **2010,** *29,* 1036–1048. doi:10.1002/etc.135
20. Pirrone, N.; Keeler, G. J.; Holsen, T. M. *Environ. Sci. Technol.* **1995,** *29,* 2123–2132. doi:10.1021/es00008a035
21. Simcik, M. F.; Franz, T. P.; Zhang, H.; Eisenreich, S. J. *Environ. Sci. Technol.* **1998,** *32,* 251–257. doi:10.1021/es970557n
22. Ryan, P. A.; Cohen, Y. *Chemosphere* **1986,** *15,* 21–47. doi:10.1016/0045-6535(86)90577-1
23. Yaffe, D.; Cohen, Y.; Arey, J.; Grosovsky, A. J. *Risk Anal.* **2001,** *21,* 275–294. doi:10.1111/0272-4332.212111
24. Harrison, R. M.; Smith, D. J. T.; Luhana, L. *Environ. Sci. Technol.* **1996,** *30,* 825–832. doi:10.1021/es950252d
25. Lazareva, A.; Keller, A. A. *ACS Sustainable Chem. Eng.* **2014,** *2,* 1656–1665. doi:10.1021/sc500121w
26. Office of Air Quality Planning and Standards, USEPA *Total Risk Integrated Methodology. TRIM.Expo Technical Support Document. External Review Draft (EPA-453/D-99-001)*, USEPA: Research Triangle Park, NC, 2002.
27. ICF Consulting. *Air Toxics Risk Assessment Reference Library*, USEPA: Research Triangle Park, NC, 2004.
28. National Research Council. *Multimedia approaches to pollution control: a symposium proceedings*; National Academy Press: Washington, DC, USA, 1987.
29. Friedlander, S. K. *Smoke, dust, and haze: fundamentals of aerosol dynamics,* 2nd ed.; Oxford University Press: New York, NY, USA, 2000.
30. Friedlander, S. K.; Wang, C. S. *J. Colloid Interface Sci.* **1966,** *22,* 126. doi:10.1016/0021-9797(66)90073-7

31. Seinfeld, J. H.; Pandis, S. N. *Atmospheric chemistry and physics: from air pollution to climate change,* 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2006.
32. Farley, K. J.; Morel, F. M. M. *Environ. Sci. Technol.* **1986,** *20,* 187–195. doi:10.1021/es00144a014
33. Spicer, P. T.; Pratsinis, S. E. *AIChE J.* **1996,** *42,* 1612–1620. doi:10.1002/aic.690420612
34. Hairer, E.; Nørsett, S. P.; Wanner, G. *Solving ordinary differential equations;* Springer Verlag: Berlin, Germany, 1987.
35. Braakhuis, H. M.; Cassee, F. R.; Fokkens, P. H. B.; de la Fonteyne, L. J. J.; Oomen, A. G.; Krystek, P.; de Jong, W. H.; van Loveren, H.; Park, M. V. D. Z. *Nanotoxicology* **2015,** in press. doi:10.3109/17435390.2015.1012184
36. Cohen, J. M.; Teeguarden, J. G.; Demokritou, P. *Part. Fibre Toxicol.* **2014,** *11,* No. 20. doi:10.1186/1743-8977-11-20
37. The World Factbook. https://www.cia.gov/library/publications/the-world-factbook/ (accessed Dec 1, 2010).
38. National Climatic Data Center: Climate Data Online. http://www.ncdc.noaa.gov/cdo-web/ (accessed Dec 1, 2013).
39. Van de Water, R. B. Modeling the Transport and Fate of Volatile and Semi-volatile Organics in a Multimedia Environment. Masters Thesis, University of California, Los Angeles, 1995.
40. United Nations Statistics Division. Municipal waste treatment. http://unstats.un.org/unsd/environment/wastetreatment.htm (accessed Oct 8, 2013).
41. United Nations Development Programme. *Human Development Reports 2012. International Human Development Indicators: Inequality-adjusted HDI value;* UNDP: New York, NY, USA, 2012.
42. Giorgi, F. *J. Geophys. Res.: Atmos.* **1986,** *91,* 9794–9806. doi:10.1029/jd091id09p09794
43. Daily Observational Data: Global Summary of the Day (GSOD). http://www.climate.gov/daily-observational-data-global-summary-day-gsod-%E2%80%93-gis-data-locator (accessed Sept 1, 2013).
44. American Meteorological Society. “Rains” in Glossary of Meteorology. http://glossary.ametsoc.org/wiki/Rains (accessed Sept 1, 2013).
45. Met Office, United Kingdom. Fact sheet No. 3 - Water in the atmosphere, National Meteorological Library and Archive. http://www.metoffice.gov.uk/media/pdf/4/1/No._03_-_Water_in_the_Atmosphere.pdf (accessed Sept 1, 2013).
46. United States Environmental Protection Agency. Water: Sewage Sludge (Biosolids), Frequently Asked Questions. http://water.epa.gov/polwaste/wastewater/treatment/biosolids/genqa.cfm (accessed Oct 8, 2013).
47. Holden, P. A.; Klaessig, F.; Turco, R. F.; Priester, J. H.; Rico, C. M.; Avila-Arias, H.; Mortimer, M.; Pacpaco, K.; Gardea-Torresdey, J. L. *Environ. Sci. Technol.* **2014,** *48,* 10541–10551. doi:10.1021/es502440s
48. Kaegi, R.; Ulrich, A.; Sinnet, B.; Vonbank, R.; Wichser, A.; Zuleeg, S.; Simmler, H.; Brunner, S.; Vonmont, H.; Burkhardt, M.; Boller, M. *Environ. Pollut.* **2008,** *156,* 233–239. doi:10.1016/j.envpol.2008.08.004
49. Keller, A. A.; Vosti, W.; Wang, H.; Lazareva, A. *J. Nanopart. Res.* **2014,** *16,* 2489. doi:10.1007/s11051-014-2489-9
50. Park, B.; Donaldson, K.; Duffin, R.; Tran, L.; Kelly, F.; Mudway, I.; Morin, J.-P.; Guest, R.; Jenkinson, P.; Samaras, Z.; Giannouli, M.; Kouridis, H.; Martin, P. *Inhalation Toxicol.* **2008,** *20,* 547–566. doi:10.1080/08958370801915309

License and Terms

The definitive version of this article is the electronic one which can be found at: doi:10.3762/bjnano.6.97

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Using natural language processing techniques to inform research on nanotechnology

Nastassja A. Lewinski[1] and Bridget T. McInnes[*2]

Review

Open Access

Address:
[1]Department of Chemical and Life Science Engineering, Virginia Commonwealth University, Richmond, VA, USA and [2]Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA

Email:
Bridget T. McInnes[*] - btmcinnes@vcu.edu

* Corresponding author

Keywords:
data mining; informatics; name entity recognition; nano-informatics; nanoparticles; nanotechnology; nanotoxicity; natural language processing; text mining

Beilstein J. Nanotechnol. **2015,** *6,* 1439–1449.
doi:10.3762/bjnano.6.149

Received: 30 March 2015
Accepted: 11 June 2015
Published: 01 July 2015

This article is part of the Thematic Series "Nanoinformatics for environmental health and biomedicine".

Guest Editor: R. Liu

Abstract

Literature in the field of nanotechnology is exponentially increasing with more and more engineered nanomaterials being created, characterized, and tested for performance and safety. With the deluge of published data, there is a need for natural language processing approaches to semi-automate the cataloguing of engineered nanomaterials and their associated physico-chemical properties, performance, exposure scenarios, and biological effects. In this paper, we review the different informatics methods that have been applied to patent mining, nanomaterial/device characterization, nanomedicine, and environmental risk assessment. Nine natural language processing (NLP)-based tools were identified: NanoPort, NanoMapper, TechPerceptor, a Text Mining Framework, a Nanodevice Analyzer, a Clinical Trial Document Classifier, Nanotoxicity Searcher, NanoSifter, and NEIMiner. We conclude with recommendations for sharing NLP-related tools through online repositories to broaden participation in nanoinformatics.

Introduction

Nanotechnology may still be considered a relatively new field. However, its impact is already realized with engineered nanomaterials (ENMs) incorporated in over 1800 consumer products, included in over 100 clinical trials, and contained in 40 FDA approved nanomedicines [1-3]. At the onset of the U.S. National Nanotechnology Initiative, researchers spearheaded efforts to "get it right the first time" by studying the potential human health and environmental impacts of ENMs in parallel with ENMs discovery and development. However, the creation and establishment of data repositories as well as algorithms to automatically analyze the collected resources has lagged behind. As a consequence, unlike bioinformatic areas such as genomics or systems biology, nanoinformatics is still in its infancy.

Nanoinformatics is defined as "the science and practice of determining which information is relevant to the nanoscale science and engineering community, and then developing and

implementing effective mechanisms for collecting, validating, storing, sharing, analyzing, modeling, and applying that information" [4]. Applications of nanoinformatics include data integration and exchange (e.g., caNanoLab, GoodNanoGuide), nanoparticle characterization (e.g., caNanoLab, Nanomaterial Registry), domain ontologies (e.g., NanoParticle Ontology), terminologies and standards (e.g., ISA-TAB-Nano), data and text mining (e.g., NEIminer, TechPerceptor), and modeling/simulation (e.g., HDAT). Extracting information usually comes from two different sources: (1) literature to which natural language processing methods are applied, and (2) experimental data to which data modeling methods, such as those used in HDAT and NanoMiner, are applied [5,6]. Despite being a largely overlooked area of informatics, several reviews have been published that list the different databases and tools currently available [7-11]. In this review, we focus on the tools that utilize natural language processing.

Natural language processing (NLP) involves the use of computers to perform practical tasks involving written language, such as extracting and analyzing information from unstructured text. What separates NLP applications from other data processing systems is their use of knowledge about human language [12]. Many of the NLP applications utilize literature retrieved from databases. Information retrieval, document classification, and pattern matching methods are often utilized to ensure that the documents being analyzed by the NLP systems contain relevant engineered nanomaterials information [13,14].

In the nanoinformatics literature discussed in this review, there are several NLP methods and systems that were proposed to extract, classify, and understand ENM-related information within unstructured text. One of the most commonly explored NLP applications by nanoinformatics researchers was Entity Extraction, which is the task of identifying mentions of a specific entity within unstructured text. The entities explored by nanoinformatics researchers varied between very specific entities such as the particle diameter of a poly(amidoamine) dendrimer [15] to very broad such as any toxicological hazard of nanoparticles [16]. Within the literature, there was also a discussion of the prospective NLP tools and algorithms that may be useful to provide information about a set of nanotechnology related documents. For example, the development of a topic identification and summarization component was proposed for incorporation into the NanoPort system to provide researchers with an automatically generated abstract or listing of relevant information based on a document [13].

Terminologies and taxonomies are equally important when building many of the NLP-based algorithms. Information Retrieval and Entity Extraction can be guided by relevant ontologies. Thomas et al. developed the first NanoParticle Ontology (NPO) based on the Open Biomedical Ontologies (OBO) Foundry principles, which were set up to promote the standardization of ontologies and common controlled vocabularies for data integration [17,18]. Recently, the eNanoMapper project has developed an ontology that merges and extends existing ontologies, including the NPO [19]. Ontologies in other languages, such as Japanese and Russian, have also been developed [20,21]. In the following section, we describe our method for identifying the nanoinformatics literature discussed in this paper, and then review the different informatics methods that have been applied such as patent mining, nanomaterial/device characterization, nanomedicine, and environmental risk assessment.

Methods

This review was limited to the English language literature included in two databases, PubMED and Web of Science [22,23]. The searches were conducted on February 12, 2015. For the search term (nano* AND "natural language processing"), Web of Science retrieved 5 records (2 excluded) and PubMED retrieved 2 records (2 excluded). For the search term (nanoinformatic*) Web of Science retrieved 38 records (34 excluded) and PubMED retrieved 24 records (22 excluded). For the search term (nano* AND "text mining"), Web of Science retrieved 38 records (34 excluded) and PubMED retrieved 2 records (2 excluded).

The following exclusion criteria were applied to the retrieved records:

- Bioinformatics papers not specifically focused on nanotechnology were not included.
- Bibliometric approaches were not included.
- Non-text based approaches (such as QSAR or image analysis) were not included.
- NLP approach(es) not described in full detail were not included.

After excluding duplicates, an initial set of 7 papers was retrieved using the described Boolean searches. We then expanded our search to include the literature cited within these 7 papers as well as the literature citing these 7 papers as identified in PubMED and Web of Science. A final set of 14 papers were included for detailed review, and the results are presented in the following section.

Review

Patent mining

Three groups across the globe (USA, Japan, China) have developed independent, NLP-based patent text mining systems. NLP

is not the only approach to text mining and we refer the reader to a recent review by Abbas et al. on the state of the art in patent analysis [24].

NanoPort

NanoPort is a web portal that (1) automatically identifies nano-related documents (website articles, patent documents, and academic articles), and (2) supports the searching and analysis of the documents [13]. The portal contains a content analysis module that utilizes NLP technology in order to help the researcher to understand and analyze the documents returned by the search engine of the portal. The authors proposed to include (1) a document summarizer, (2) a document clusterer, (3) a topic mapper, and (4) a patent analyzer.

The proposed document summarizer automatically develops an abstract containing the important points of the document for the researcher. The authors propose using their previously developed Arizona Txttractor system, which was initially developed for web pages. The document clusterer groups the documents returned by the portal based on common topics identified within the document using the author's Arizona Noun Phraser (ANP). ANP identifies noun phrases in text and then ranks them based on their frequency. The highly frequent noun phrases are used as topics by the clusterer as well as to support visualization of the search results in the topic mapper. The proposed Patent analyzer supports the basic analysis, content map analysis and citation network analysis. The basic analysis contains traditional patent analysis information such as number of patents based on country, institution or technology field. The content map allows for the concepts from multiple patents to be viewed and analyzed over time. The patent citation network allows for the visualization of links between entities such as countries, institutions and technology fields providing a wider scope of the field for the researcher. NanoPort was hosted at http://www.nanoport.org but unfortunately is no longer available online.

NanoMapper

NanoMapper expands on the proposed patent analyzer within the NanoPort system [25]. The NanoMapper prototype provides search capability, visualization and analytical tools to analyze nanotechnology patents from the United States Patent and Trademark Office (USPTO), European Patent Office (EPO), Japan Patent Office (JPO), and grants from the U.S. National Science Foundation (NSF). It includes basic statistics, citation network analysis and content map analysis as described in the proposed NanoPort patent analyzer as well as publication trend analysis to compare trends of patents and grants. Similarly to NanoPort, the NSF-funded NanoMapper was hosted at http://nanomapper.eller.arizona.edu but is no longer available online.

TechPerceptor

TechPerceptor is a text mining tool to conduct patent analysis and generate a patent map based on a subject–action–object (SAO) approach [26-28]. Their training corpus consisted of 136 patents and was initially analyzed for trends in carbon nanotube synthesis methods [26,27]. More recently, the research group expanded the scope to include applications of carbon nanotubes such as incorporation in photovoltaic cells and prostate cancer therapeutics [28]. The patents, which spanned the years 1992 to 2009, were collected from E.U., Japan, Korea and U.S. patent databases with patents in Japanese and Korean translated using K2E-PAT or Google Translate. The group followed a four step procedure for both their SAO-based static and dynamic patent map construction: 1) collect patent data, 2) extract SAO structures using NLP, (3) generate a patent dissimilarity matrix, and (4) visualize as dynamic patent [26,27]. The patent maps were also automatically analyzed to identify areas of high or low activity, infringement and novelty, which were determined based on degrees of (dis)similarity to other patents [28].

Their static tool revealed 8 patent clusters with the most patents reporting arc-discharge and laser vaporization synthesis methods [26]. Chemical vapor deposition (CVD) methods were also mentioned as being invented frequently. Top patenting companies included NEC, Samsung and Sony. Their dynamic tool revealed a possible patent vacuum of using low temperature or microwave-based synthesis of single-walled carbon nanotubes [27]. Analyzing hot spots revealed changes in the type of synthesis method patented over time, with synthesis methods evolving from arc discharging in 1999–2000 to metal-catalyzed heat-treatment syntheses and CVD in 2003–2004, to arc discharge with purification control in 2005–2006, to plasma-enhanced and thermal CVD in 2007–2010. CVD is the dominant commercial synthesis approach and catalyzed CVD with fluidized bed has been used by Bayer to synthesize Baytubes [29]. Competitor analysis revealed overlap between Sony and an individual researcher, Young Sang Cho.

Text mining framework for Nano S&T

Junpeng et al. developed a patent text mining tool using NLP [14]. Patents were retrieved from Science Citation Index, Engineering Information Compendex, International Information Services for Physics and Engineering communities, and the Chinese Patent database. Text extraction was conducted, with fuzzy logic used to cleanse the data. Fuzzy matching techniques were used to identify and combine similar entities. List Process, Matrix Process, Factor Analysis, Technology Group Clustering, and Concept Hierarchy were used in the framework to analyze the database. Multi-dimensional scaling was employed with a path erasing algorithm. The data presented

focused on identifying leading countries, companies and inventors in the nanotechnology field. At the time of publication, the top three patenting institutions representing the top three patenting countries included the Naval Research Laboratory (USA), Cavendish Laboratory (UK), and Hitachi Ltd (Japan).

Nanomaterial/device characterization

Not all ENMs or nanodevices and their respective synthesis or fabrication methods are patented. In addition, the information provided in a patent can be limited compared to that included in a research article. Therefore systems that can automatically retrieve and annotate literature on ENMs/nanodevices can be valuable tools for accelerating the discovery/design, synthesis/fabrication and optimization of ENMs/nanodevices.

Nanodevice fabrication and characterization analyzer

Dieb et al. generated a tool to automatically collect literature relevant to nanodevice design and a tool to automatically annotate literature on nanodevices [30,31]. A training set, which consisted of two fully annotated papers with 129 sentences, was manually annotated by graduate students with the assistance of an annotation support tool, XConc Suite [32]. The terms included: source material (SMaterial), characteristic feature of material (SMChar), experiment parameter (ExP), value of the experiment parameter (ExPVal), evaluation parameter (EvP), value of the evaluation parameter (EvPVal), manufacturing method (MMethod), and final product (TArtifact).

Because terms can overlap with other terms, four tag groups were created where the terms within a group did not overlap. With these four tag groups, cascading style annotation could be applied [31]. To automate the annotation process, a biomedical entity extraction method using the supervised machine learning algorithm, support vector machines (SVM), was applied to their literature library. Supervised machine learning algorithms learn patterns and make predictions based on a set of training data. The training data for this system was generated by first parsing the text using a part-of-speech (POS) tagger, with tag category and boundary represented using the BIO format. The part-of-speech information, category, and context surrounding the term where used as features (or parameters) for the machine learning algorithm. For the source material, a publicly available chemical entity recognizer, OSCAR3-a5, was first used to parse the papers. However, since the precision (the percentage of correctly identified entities over all the entities identified by the system) of OSCAR-a5 was poor (0.59), the group developed a custom chemical entity recognizer called CNER, where they improved issues related to chemical symbol and acronym confusion. CNER had improved precision (0.92) with similar recall (0.97 compared to 0.99 for OSCAR-a5). Recall is the percentage of correctly identified entities over all the entities in the datatset. The authors also used a text chunk annotator based on the sequence labeling tool called YamCha (available at http://chasen.org/~taku/software/yamcha/) and a POS tagger called GPoSTTL (available at http://gposttl.sourceforge.net/).

The tool was further improved by applying a physical quantities list (based on the one listed on the website chemistry.about.com) to refine the extraction of two tags: evaluation parameter and experiment parameter [31]. However, their annotated library only expanded from two to five papers, and the group only used two papers to test their improved system. The group also further improved their CNER, renaming it SERB-CNER or syntactically enhanced rule-based chemical entity recognizer. SERB-CNER still focused on the Source Material tag. Here the POS tagger used was rb tagger. The machine learning system used was CRF++. This new system had recall improvements of 4–7% depending on which parameter was examined.

Nanomedicine

Through targeted and activatible delivery, nanomedicine has the potential to greatly improve drug efficacy while reducing side effects. Improved design can also address emerging challenges to disease treatment such as adaptive resistance. Despite the promise, few nanomedicines have successfully advanced from the bench to the clinic. For both developing and marketed nanomedicines, there still remain questions on the long-term safety. Two groups have developed NLP-based systems to annotate and classify nanomedicine articles or clinical trials.

Nanotoxicity Searcher

The Nanotoxicity Searcher is a tool to automatically annotate nanomedicine and nanotoxicology literature using pattern matching techniques [9,16,33]. The group used ABNER (available at http://pages.cs.wisc.edu/~bsettles/abner/), a biomedical named entity recognizer, to identify names of nanomaterials (NANO), potential routes of exposure (EXPO), target organs and/or organisms (TARGET), and types of toxicity/damage (TOXIC) [16,34]. ABNER contains the supervised machine learning algorithm linear-chain conditional random fields (CRFs) from Mallet (available at http://mallet.cs.umass.edu/), an open source freely available Java-based statistical natural language processing toolkit [35]. To create training data for the CRF, the authors manually annotated 300 sentences collected from 654 abstracts retrieved in PubMed after searching "nanoparticles/toxicity (MeSH major topic)". For example, the authors manually labeled the sentence

"The purpose of this study was to review published dose-response data on acute lung inflammation in rats after instillation of titanium dioxide particles or six types of carbon nanoparticles."

with the NANO, EXPO, TARGET and TOXIC mentions within the sentence

"*The purpose of this study was to review published dose-response data on acute <TARGET>* ***lung*** *</TARGET> <TOXIC>* ***inflammation*** *</TOXIC> in <TARGET>* ***rats*** *</TARGET> after <EXPO>* ***installation*** *</EXPO> of <NANO>* **titanium dioxide particles** </NANO> or six types of <NANO> **carbon nanoparticles** </NANO>)."

Features extracted from the context surrounding the mentions were used to train the CRF.

The performance of their NER software was measured based on three factors: precision, recall, and F-measure score. F-measure is the harmonic mean of precision and recall. The authors evaluated how well their system performed in identifying the entire entity string (entity-level) and partial matches (token-level). For each level, their results were reported to be greater than 0.85, with almost all factors examined at the token level greater than 0.9. The performance of the Nanotoxicity Searcher was also compared to a baseline method, which combines a dictionary-based approach with a term selection scheme. The dictionary was created manually from the same 300 sentences used to train the CRF plus terms identified from two ontologies, the Foundational Model of Anatomy (FMA) and the NanoParticle Ontology [36]. The results demonstrated that overall the CRF method obtained a significantly higher F-measure than the baseline.

NanoSifter

The NanoSifter, which focused on a specific type of ENM, is finer grained than the Nanotoxicity Searcher, which used four broad nano entities encompassing all types of ENMs [15]. NanoSifter was designed to identify quantitative data (i.e., numerical values for different characterization parameters) associated with a specific class of dendrimer, poly(amidoamine) (PAMAM), which shows promise for cancer treatment. PAMAM dendrimers are three-dimensional, highly-branched, polymeric ENMs synthesized by growing shells of branched molecules from a central core ethylenediamine molecule. Each doubling of the number of amine surface groups constitutes a new shell or generation.

The NanoSifter algorithm contains two steps. The first to identify possible mentions of the entities associated with PAMAM, and the second to associate the numeric values and dendrimer property terms. The entities associated with PAMAM were based on the NanoParticle Ontology and included: (1) hydrodynamic diameter, (2) particle diameter, (3) molecular weight, (4) zeta potential, (5) cytotoxicity, (6) IC_{50}, (7) cell viability, (8) encapsulation efficiency, (9) loading efficiency, and (10) transfection efficiency [17]. To identify mentions associated with PAMAM entities, the authors utilize the freely available open source NLP pipeline General Architecture for Text Engineering (GATE, https://gate.ac.uk/) and its IE module ANNIE (a Nearly-New Information Extraction System, https://gate.ac.uk/ie/annie.html) [37]. GATE, originally developed by the University of Sheffield, is a widely employed suite of Java tools developed for the processing unstructured text [37]. ANNIE is an information extraction module within GATE that contains a tokenizer, sentence splitter, part-of-speech tagger and named entity extractor. The named entity extractor of ANNIE is tailored to extract entities such as persons, organizations and dates, but the components are highly configurable and can be adapted to extract a variety of entities.

To create a training set for the entity extractor, two domain experts annotated 100 articles for the numeric values and dendrimer property terms using the Java Annotations Patterns Engine (JAPE) and integrating components from ANNIE. The training data was then utilized by ANNIE's IE module to identify mentions associated with PAMAM. The identified numerical values cannot be automatically assumed to associate with a PAMAM property. Therefore, to determine if the associated numeric values of the PAMAM entities were referring to the dendrimer property, the authors utilized a proximity metric. The proximity metric requires the mention of a PAMAM property to be within so many characters of the property term. This provides the system with context information used in the literature when referring to the entity. The authors selected a proximity distance metric threshold of 200 characters based on preliminary experiments using the training set. Too large of a proximity metric provides the system with too much information to accurately discriminate whether the word is an entity, which increases the false positive rate, whereas too little of a proximity metric does not provide the system with enough context information. Evaluating their results using precision, recall and F-measure metrics showed that their algorithm obtained a high accuracy and recall when identifying entities associated with the PAMAM properties. The performance of NanoSifter was based on comparison with annotations generated by researchers working in the Ghandehari lab at the University of Utah. Overall, NanoSifter demonstrated good recall (95–100% - 99%), poor precision (59–100% - 84%), a passing F-measure (73–100% - 91%).

Clinical trial document classifier

De la Iglesia et al. proposed a method to automatically classify clinical trial summaries as those testing nanotechnology products and those testing conventional drugs [38]. A benefit of this system is that it can automatically identify summaries of interest for further processing by more computationally intensive systems such as those discussed elsewhere in this review. Looking for just the term "nano" is not sufficient to determine if a summary contains nanotechnology products because many summaries do not explicitly state that they are testing nanotechnology products. For example, many nanotechnology products encapsulate insoluble or highly cytotoxic drugs within liposomal or micellar particles, which alters the kinetics of the drug in the body.

To develop their system, the group used the Natural Language Toolkit (NLTK, http://www.nltk.org/), a suite of freely available, open source, Python-based modules developed for processing unstructured text. They evaluated seven supervised machine learning algorithms implemented in the package: (1) multinomial naive Bayes classifier, (2) decision trees, (3) stochastic gradient descent (SGD) logistic regression, (4) L-1 regularized logistic regression, (5) L-2 regularized logistic regression, (6) linear support vector machine and (7) polynomial support vector machine. The authors explored four vector-based methods for representing the document each using a "bag-of-words" approach containing unigrams (single content words) and bigrams (sequence of two content words) as features (or parameters) for the machine learning algorithm. The first is a binary representation, where a zero or one is used to indicate the absence or presence of the feature in the summary. The second is a feature-based representation, which uses the number of times the feature occurred in the summary. The third is inverse-document frequency (IDF), which quantifies how discriminative a feature is based on the number of documents it occurred within. And lastly, the fourth is term frequency-inverse document frequency (TFIDF), which weights IDF based on how often the term occurs.

The authors trained their algorithm on 1000 clinical trial summaries from clinicaltrials.gov, where 500 were nanomedicine-focused (nano) and 500 were not involving any nanomedicines or nanodevices (non-nano). The author evaluated their system using the leave-one-out and 10-fold cross validation evaluation methodology and report the overall: (1) precision, (2) recall, (3) F-measure, (3) true-positive vs false-positive rates, (4) Mathews correlation coefficient (MCC) and (5) area under the curve (AUC). The MCC measures the quality of the nano/non-nano classification by the system and the AUC measures the discriminativeness of the classifier. The results show an F-measure greater than 0.85 regardless of the machine learning algorithm or feature representation. The overall results indicate that the context within the unigram and bigram features is able to discriminate between non-nano and nano clinical summaries.

The authors describe several advantages of automatically categorizing clinical trials investigating nano versus non-nano drugs. These include facilitating comparisons between clinical trials testing nano and non-nano drug formulations involving the same active ingredient (e.g., Doxil = pegylated liposome [nano] encapsulated *doxorubicin* compared to Adriamycin = *doxorubicin*). In addition, categorization could facilitate information retrieval by users interested in this distinction. In the consumer product arena, labeling consumer products containing ENMs has been discussed widely, and a similar NLP categorization tool tailored to consumer products could potentially facilitate the categorization of products containing nanomaterials or generated using nanotechnology-based processes from those not involving nanotechnology.

Environmental risk assessment

Environmental release and exposure to ENMs is already occurring, and it is the obligation of nanotechnology researchers to also consider the potential effects of commercialized ENMs on human health and environment. A wealth of data has been collected through large-scale centers, which in the U.S. include the Center for Biological and Environmental Nanotechnology (CBEN) and the two Centers for Environmental Implications of Nanotechnology (CEIN and CEINT). Surprisingly, only one group was found to describe the use of NLP techniques in a tool analyzing the environmental nanotechnology literature.

NEIMiner

The Nanomaterial Environmental Impact data Miner, or NEIMiner, is a web-based tool built using CMS and Drupal [39]. NEIMiner consists of four parts: 1) nanomaterial environmental impact (NEI) modeling framework – similar to Framework for Risk Analysis of Multi-Media Environmental Systems (FRAMES), 2) data integration, 3) data management and access, and 4) model building. This web-based tool is supported by the company's previously developed tool, ABMiner (available at http://discover.nci.nih.gov/abminer/). Three databases (ICON, caNanoLab, and NBI) were used as the data sources. Data extraction was performed using application programming interface (API) calling via web services and data scraping via parsing web pages. The model building component of NEIMiner utilizes machine learning algorithms from ABMiner, such as nearest neighbor algorithms, tree algorithms and support vector machines. This allows for the systematic evaluation of a variety of algorithms. The model building component also contains a meta-optimizer, which automatically iterates

through the algorithms in ABMiner that can be used to solve the input problem to determine which algorithm will provide the most optimal results. To demonstrate the applicability of the model building component, the authors developed a predictive model based on the Nanomaterial-Biological Interactions (NBI) knowledge base. The NBI includes data on the mortality, delayed development and morphological malformations of embryonic zebrafish due to the toxicity of various nanomaterials including metal nanoparticles, dendrimer, metal oxide and polymeric materials [40]. Java Applets were used to visualize the data in 3D histograms and scatterplots. NEIMiner was hosted at http://neiminer.i-a-i.com but is no longer accessible.

Conclusion

NLP perspective

Nine nanoinformatics systems utilizing NLP have been described in the literature. Table 1 shows the components of these systems from a NLP perspective. "NLP tasks" describes the applications discussed by the researchers when developing their system. "NLP subtasks" shows the underlying NLP components that were utilized within the systems. For example, NanoMapper, a patent analyzer developed by Li et al., utilized a part-of-speech (POS) tagger and parser within their system to automatically annotate the words in the document with their part-of-speech and extract the phrasal chunks from the sentences [25]. Similarly, the TechPerceptor system developed by Yoon et al. utilizes a stemmer in order to normalize words to their base form, and sentence similarity algorithms to compare how close the contextual content of one sentence is with another [26].

Many of the nanoinformatics systems were implemented using pre-existing NLP software packages. These NLP packages were developed to perform specific tasks, such as Abner, a biomedical named entity extractor, or more general NLP systems that provide various NLP tools such as Mallet and Natural Language Toolkit (NLTK) [34,35]. Utilizing and adapting these previously developed NLP tools allows for nanoinformatics researchers to build their automated systems without needing to develop low level NLP functionality. There were three main types of algorithms utilized by the systems: machine learning, pattern matching and clustering. The most common was machine learning algorithms such as Conditional Random Fields and Support Vector Machines (SVMs). These algorithms require manually annotated training data. For example, in building the Nanotoxicity Searcher, Garcia-Remesal et al. manually annotated documents for various nanoparticles and their toxicological hazards to train their entity extraction system [16]. In many cases, the annotation toolkit (if used) was not reported, but two annotation systems were mentioned in the articles reviewed: 1) GATE and 2) XConc Suite.

Lastly, although not specifically an NLP component, five groups incorporated visualization of the extracted information as part of their system. Visualization provides researchers with additional capabilities to explore and analyze the data.

Data perspective

Table 2 shows the components of the nanoinformatics systems from a data perspective. With the growing number of nanotechnology publications, more refined databases that automatically identify records (e.g., articles, patents, grants, clinical trials) relevant to specific ENMs or properties can greatly facilitate trend analyses. The amount of information gathered automatically differed widely between the systems reviewed. The Clinical Trial Document Classifier focused on differentiating between two variables, nanotechnology products and non-nanotechnology products [38]. The four patent mining systems (i.e., NanoPort, NanoMapper, TechPerceptor, and Text Mining Framework) primarily extracted publication information, which allowed for patents to be clustered by date, inventor, country, and institution. However, the TechPerceptor also extracted information on nanomaterial type and synthesis method [26]. Moving beyond bibliographic information, the Nanodevice Fabrication and Characterization Analyzer automatically extracted nanodevice physico-chemical characterization properties as well as the fabrication and evaluation parameters and their associated values [30]. Comparing the parameters that were extracted to the proposed minimum information for nanomaterials characterization, referred to as MINChar in the table, 64% of parameters were captured [41]. This system was trained using two annotated articles, and its application to a larger literature corpus has not been published. This may be due to future plans to integrate a system, similar to the patent analyzers, where the extracted data are associated with the citation information.

The amount of physico-chemical characterization data extracted by the systems analyzing literature for exposure and biological response data (i.e., Nanotoxicity Searcher, NanoSifter, and NEIMiner) varied greatly. Focused primarily on the toxicity endpoints, the Nanotoxicity Searcher extracted several biological response endpoints but only associated these effects with the ENMs' core composition [16]. The NanoSifter collected size, surface charge and molecular weight data beyond the core composition, which was fixed to PAMAM [15]. Incorporating almost 80% of the minimum characterization data, the NEIMiner appears to be the most comprehensive with regards to extraction of physico-chemical characterization properties.

When assessing the human health or environmental impact of ENMs, it is important to recognize that risk is a function of exposure and hazard. Without exposure, there is no risk. All

Table 1: Nanoinformatic system components from an NLP perspective.

		Nano Porter	Nano Mapper	Tech Perceptor	Text Mining Framework	Nano Device F & C	Nano Toxicity Searcher	Nano Sifter	Clinical Trial Doc. Class.	NEI Miner
machine learning algorithm	CRF					×	×			
	decision trees								×	×
	logistic regression								×	
	naive Bayes								×	
	nearest neighbor									×
	SVM					×			×	×
algorithm class	machine learning						×		×	×
	pattern matching							×		
	clustering	×	×	×	×					
visualization	visualization modules	×	×	×	×					×
taxonomy	FMA (in UMLS)						×			
	MeSH (in UMLS)						×			
	WordNet			×						
	NanoParticle Ontology						×			
NLP tools	GATE (NLP Toolkit)							×		
	Xconc Suite (annotator)					×				
	ABMiner (NLP Toolkit)									×
	Abner (NER)						×			
	YamCha (Parser)					×				
	GPoSSTTL (POS Tagger)					×				
	ANNIE (GATE module)							×		
	Mallet (NLP Toolkit)						×			
	NLTK (NLP Toolkit)								×	
NLP sub task	POS tagging	×	×	×		×				
	parsing	×	×			×				
	concept mapping			×						
	stemming			×						
	sentence similarity			×						
NLP task	document classification								×	
	document clustering	×								
	entity extraction					×	×	×		×
	information retrieval	×				×				
	patent analyzer	×	×	×	×					
	summarization	×								
	topic identification	×	×							

substances are potentially hazardous depending on the dose or concentration encountered. In addition, the biological response data of interest can be dependent upon the application. Nanomedicine applications are often evaluated using performance parameters, such as drug loading efficiency and efficacy, in addition to biological response, such as cytotoxicity or IC_{50}. Since efficacy and cytotoxicity are dependent upon the administered dose, concentration and exposure dose parameters are critical for the interpretation of this data. While text mining is useful, it is only the first step. Current nano-focused NLP systems are not sufficient to reveal relationships or connections between data. Close collaboration and communication between nanotoxicology and nanoinformatics researchers will provide interpretive context so that computer understandable patterns

Table 2: Nanoinformatic system components from a data perspective.

		MIN Char	Nano Porter	Nano Mapper	Tech Perceptor	Text Mining Framework	Nano Device F & C	Nano Toxicity Searcher	Nano Sifter	Clinical Trial Doc. Class.	NEI Miner
publication information	citation (e.g., author, journal, date)		×	×	×	×		×			×
	laboratory/ organization			×		×					
	location		×	×		×					
	content description		×	×	×						
	patent classification (e.g., US, EU)		×	×	×						
physico-chemical character-ization	particle diameter	×					×		×		×
	particle size distribution	×									×
	hydrodynamic diameter								×		
	agglomeration and/or aggregation	×									×
	shape	×					×				×
	core composition	×			×		×	×			×
	crystallinity/crystalline state	×					×				×
	surface area	×					×				
	surface charge/zeta potential	×					×		×		×
	surface chemistry	×					×				×
	purity	×					×				×
	stability	×									
	solubility	×									
	concentration (mass, number, SA)	×									×
	method of synthesis/preparation	×			×		×				×
	molecular weight								×		
exposure	exposure media										×
	exposure pathway/route							×			×
	exposure duration										×
	exposure dose										×
biological response	bioavailability/uptake										×
	biomagnification										×
	cell viability							×	×		
	cytotoxicity							×	×		×
	inflammatory response							×			
	genotoxicity							×			×
	EC_{50} (ppm)							×			
	IC_{50}							×	×		
	LC_{50} (ppm)							×			
	organ response							×			
	whole organism response							×			×

can be developed to enable future knowledge discovery from the literature.

Recommendations

There is a critical need to automatically extract and synthesize knowledge and trends from nanotechnology literature. New ENMs are continuously being discovered and NLP approaches can semi-automate the cataloguing of ENMs and their unique physico-chemical properties. As shown in this review, various NLP methods have been used for patent mining, nanomaterial/device characterization, nanomedicine, and environmental risk assessment. We believe these approaches can be expanded upon to automatically aggregate studies on the exposure and hazard of ENMs as well as link the physico-chemical properties to the measured effects. Towards this end, we conclude with the following recommendations:

- Add the NPO to the Unified Medical Language System (UMLS). → Impact: provide a nano-specific terminology source that can be used by pre-existing systems that currently utilize sources from the UMLS.
- Create a publicly available annotated corpus for nanotechnology. → Impact: develop new nanoinformatics tools; provide a benchmark dataset to compare nanoinformatic systems.
- Encourage authors to include more experimental details, such as the minimum characterization data, in their manuscripts. → Impact: increase experimental reproducibility and inter-study comparison.
- Encourage researchers to add nanoinformatics tools to freely available, online repositories, such as nanoHUB or NCIPhub. → Impact: Promote broader participation in the nanoinformatics field.

References

1. Woodrow Wilson International Center for Scholars. Project on Emerging Nanotechnologies, Consumer Product Inventory. http://www.nanotechproject.org/cpi (accessed March 9, 2015).
2. Venditto, V. J.; Szoka, F. C., Jr. *Adv. Drug Delivery Rev.* **2013,** *65,* 80–88. doi:10.1016/j.addr.2012.09.038
3. Schütz, C. A.; Juillerat-Jeanneret, L.; Mueller, H.; Lynch, I.; Riediker, M. *Nanomedicine (London, U. K.)* **2013,** *8,* 449–467. doi:10.2217/NNM.13.8
4. de la Iglesia, D.; Harper, S.; Hoover, M. D.; Klaessig, F.; Lippell, P.; Maddux, B.; Morse, J.; Nel, A.; Rajan, K.; Reznik-Zellen, R.; Tuominen, M. T. Nanoinformatics 2020 Roadmap. 2011; http://eprints.internano.org/id/eprint/607 (accessed March 9, 2015).
5. Liu, R.; Hassan, T.; Rallo, R.; Cohen, Y. *Comput. Sci. Discovery* **2013,** *6,* 014006. doi:10.1088/1749-4699/6/1/014006
6. Kong, L.; Tuomela, S.; Hahne, L.; Ahlfors, H.; Yli-Harja, O.; Fadeel, B.; Lahesmaa, R.; Autio, R. *PLoS One* **2013,** *8,* e68414. doi:10.1371/journal.pone.0068414
7. Maojo, V.; Martin-Sanchez, F.; Kulikowski, C.; Rodriguez-Paton, A.; Fritts, M. *Pediatr. Res.* **2010,** *67,* 481–489. doi:10.1203/PDR.0b013e3181d6245e
8. Thomas, D. G.; Klaessig, F.; Harper, S.; Fritts, M.; Hoover, M.; Gaheen, S.; Stokes, T.; Reznik-Zellen, R.; Freund, E.; Klemm, J.; Paik, D.; Baker, N. *Wiley Interdiscip. Rev.: Nanomed. Nanobiotechnol.* **2011,** *3,* 511–532. doi:10.1002/wnan.152
9. Maojo, V.; Fritts, M.; de la Iglesia, D.; Cachau, R. E.; Garcia-Remesal, M.; Mitchell, J. A.; Kulikowski, C. *Int. J. Nanomed.* **2012,** *7,* 3867–3890. doi:10.2147/IJN.S24582
10. de la Iglesia, D.; Cachau, R. E.; García-Remesal, M.; Maojo, V. *Comput. Sci. Discovery* **2013,** *6,* 014011. doi:10.1088/1749-4699/6/1/014011
11. Panneerselvam, S.; Choi, S. *Int. J. Mol. Sci.* **2014,** *15,* 7158–7182. doi:10.3390/ijms15057158
12. Jurafsky, D.; Martin, J. *Speech and language processing;* Pearson: Englewood Cliffs, NJ, U.S.A., 2014.
13. Chau, M.; Huang, Z.; Qin, J.; Zhou, Y.; Chen, H. *Decis. Support Syst.* **2006,** *42,* 1216–1238. doi:10.1016/j.dss.2006.01.004
14. Junpeng, Y.; Jin, H.; Donghua, Z.; Hailong, B.; Chunning, Y. A Text Mining Framework to Support Nano Science and Technology Management. In *Proceedings of the IMACS Multiconference on Computational Engineering in Systems Applications,* Beijing, China, Oct 4–6, 2006; pp 2086–2091. doi:10.1109/CESA.2006.4281982
15. Jones, D. E.; Igo, S.; Hurdle, J.; Facelli, J. C. *PLoS One* **2014,** *9,* No. e83932. doi:10.1371/journal.pone.0083932
16. García-Remesal, M.; García-Ruiz, A.; Pérez-Rey, D.; de la Iglesia, D.; Maojo, V. *BioMed Res. Int.* **2013,** No. 410294. doi:10.1155/2013/410294
17. Thomas, D. G.; Pappu, R. V.; Baker, N. A. *J. Biomed. Inf.* **2011,** *44,* 59–74. doi:10.1016/j.jbi.2010.03.001
18. Smith, B.; Ashburner, M.; Rosse, C.; Bard, J.; Bug, W.; Ceusters, W.; Goldberg, L.; Eilbeck, K.; Ireland, A.; Mungall, C.; The OBI Consortium; Leontis, N.; Rocca-Serra, P.; Ruttenberg, A.; Sansone, S.; Scheuermann, R.; Shah, N.; Whetzel, P.; Lewis, S. *Nat. Biotechnol.* **2007,** *25,* 1251–1255. doi:10.1038/nbt1346
19. Hastings, J.; Jeliazkova, N.; Owen, G.; Tsiliki, G.; Munteanu, C.; Steinbeck, C.; Willighagen, E. *J. Biomed. Semantics* **2015,** *6,* 10. doi:10.1186/s13326-015-0005-5
20. Nanotechnology Structured Knowledge Platform Nanostructure Indexes. http://mandala.t.u-tokyo.ac.jp/english/nanoindex.html (accessed March 4, 2015).
21. Ivanov, V. German-Russian School on Semantic Technologies, Towards an applied ontology of nanomaterials and nanotechnologies. 2012; http://www.workshop-misis.ru/documents/ros_germ/iva.pdf (accessed March 4, 2015).
22. PubMed. http://www.ncbi.nlm.nih.gov/pubmed (accessed Feb 12, 2015).
23. Web of Science. http://www.isiknowledge.com (accessed Feb 12, 2015).
24. Abbas, A.; Zhang, L.; Khan, S. U. *World Pat. Inf.* **2014,** *37,* 3–13. doi:10.1016/j.wpi.2013.12.006
25. Li, X.; Hu, D.; Dang, Y.; Chen, H.; Roco, M. C.; Larson, C. A.; Chan, J. *J. Nanopart. Res.* **2009,** *11,* 529–552. doi:10.1007/s11051-008-9491-z
26. Yoon, J.; Kim, K. *Scientometrics* **2011,** *88,* 213–228. doi:10.1007/s11192-011-0383-0
27. Yoon, J.; Park, H.; Kim, K. *Scientometrics* **2013,** *94,* 313–331. doi:10.1007/s11192-012-0830-6
28. Park, H.; Kim, K.; Choi, S.; Yoon, J. *Expert Syst. Appl.* **2013,** *40,* 2373–2390. doi:10.1016/j.eswa.2012.10.073

29. De Volder, M. F. L.; Tawfick, S. H.; Baughman, R. H.; Hart, A. J. *Science* **2013,** *339,* 535–539. doi:10.1126/science.1222453
30. Dieb, T.; Yoshioka, M.; Hara, S. Automatic Information Extraction of Experiments from Nanodevices Development Papers. In *2012 IIAI International Conference on Advanced Applied Informatics (IIAIAAI),* IIAI - International Conference on Advanced Applied Informatics, Fukuoka, Japan, Sept 20–22, 2012; 2012; pp 42–47. doi:10.1109/IIAI-AAI.2012.18
31. Dieb, T.; Yoshioka, M.; Hara, S.; Newton, M. In *Proceedings of the 4th International Workshop on Computational Terminology,* Dublin, Ireland, Aug 23, 2014; 2014; pp 77–85.
32. Kim, J.-D.; Ohta, T.; Tsujii, J.'i. *BMC Bioinf.* **2008,** *9,* 10. doi:10.1186/1471-2105-9-10
33. Chiesa, S.; Garcia-Remesal, M.; de la Calle, G.; de la Iglesia, D.; Bankauskaite, V.; Maojo, V. Building an Index of Nanomedical Resources: An Automatic Approach Based on Text Mining. In *Knowledge-Based Intelligent Information and Engineering Systems, 12th International Conference, KES 2008, Zagreb, Croatia, September 3-5, 2008, Proceedings, Part II;* Lovrek, I.; Howlett, R. J.; Jain, L. C., Eds.; Lecture Notes in Computer Science, Vol. 5178; Springer: Berlin, Germany, 2008; pp 50–57. doi:10.1007/978-3-540-85565-1_7
34. Settles, B. In *Proceedings of the COLING 2004 International Joint Workshop of Natural Languate Processing in Biomedicine and its Applications,* 2004; pp 104–107.
35. *MALLET: A Machine Learning for Language Toolkit;* McCallum, A., 2002, http://mallet.cs.umass.edu.
36. Thomas, D. G.; Pappu, R. V.; Baker, N. A. Ontologies for cancer nanotechnology research. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009,* EMBC 2009, Minneapolis, MN, U.S.A., Sept 3–6, 2009; 2009; pp 4158–4161. doi:10.1109/IEMBS.2009.5333941
37. Cunningham, H. *Comput. Humanit.* **2002,** *36,* 223–254. doi:10.1023/A:1014348124664
38. de la Iglesia, D.; García-Remesal, M.; Anguita, A.; Muñoz-Marmol, M.; Kulikowski, C.; Maojo, V. *PLoS One* **2014,** *9,* No. e110331. doi:10.1371/journal.pone.0110331
39. Tang, K.; Liu, X.; Harper, S.; Steevens, J. A.; Xu, R. *Int. J. Nanomed.* **2013,** *8,* 15–29. doi:10.2147/IJN.S40974
40. Liu, X.; Tang, K.; Harper, S.; Harper, B.; Steevens, J. A.; Xu, R. *Int. J. Nanomed.* **2013,** *8,* 31–43. doi:10.2147/IJN.S40742
41. Card, J.; Magnuson, B. *J. Food Sci.* **2009,** *74,* vi–vii. doi:10.1111/j.1750-3841.2009.01335.x

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Influence of surface chemical properties on the toxicity of engineered zinc oxide nanoparticles to embryonic zebrafish

Zitao Zhou[1], Jino Son[2], Bryan Harper[2], Zheng Zhou[1] and Stacey Harper[*1,2,3]

Full Research Paper

Open Access

Address:
[1]School of Chemical, Biological and Environmental Engineering, Oregon State University, Corvallis, Oregon, 97330, United States, [2]Department of Environmental and Molecular Toxicology, Oregon State University, Corvallis, Oregon, 97330, United States and [3]Oregon Nanoscience and Microtechnologies Institute, Eugene, Oregon, United States

Email:
Bryan Harper - bryan.harper@oregonstate.edu; Stacey Harper* - stacey.harper@oregonstate.edu

* Corresponding author

Keywords:
kriging estimation; modelling; nanomaterials; nanotechnology; toxicology

Beilstein J. Nanotechnol. **2015,** *6,* 1568–1579.
doi:10.3762/bjnano.6.160

Received: 04 April 2015
Accepted: 01 July 2015
Published: 20 July 2015

This article is part of the Thematic Series "Nanoinformatics for environmental health and biomedicine".

Guest Editor: R. Liu

Abstract

Zinc oxide nanoparticles (ZnO NPs) are widely used in a variety of products, thus understanding their health and environmental impacts is necessary to appropriately manage their risks. To keep pace with the rapid increase in products utilizing engineered ZnO NPs, rapid in silico toxicity test methods based on knowledge of comprehensive in vivo and in vitro toxic responses are beneficial in determining potential nanoparticle impacts. To achieve or enhance their desired function, chemical modifications are often performed on the NPs surface; however, the roles of these alterations play in determining the toxicity of ZnO NPs are still not well understood. As such, we investigated the toxicity of 17 diverse ZnO NPs varying in both size and surface chemistry to developing zebrafish (exposure concentrations ranging from 0.016 to 250 mg/L). Despite assessing a suite of 19 different developmental, behavioural and morphological endpoints in addition to mortality in this study, mortality was the most common endpoint observed for all of the ZnO NP types tested. ZnO NPs with surface chemical modification, regardless of the type, resulted in mortality at 24 hours post-fertilization (hpf) while uncoated particles did not induce significant mortality until 120 hpf. Using eight intrinsic chemical properties that relate to the outermost surface chemistry of the engineered ZnO nanoparticles, the highly dimensional toxicity data were converted to a 2-dimensional data set through principal component analysis (PCA). Euclidean distance was used to partition different NPs into several groups based on converted data (score) which were directly related to changes in the outermost surface chemistry. Kriging estimations were then used to develop a contour map based on mortality data as a response. This study illustrates how the intrinsic properties of NPs, including surface chemical modifications and capping agents, are useful to separate and identify ZnO NP toxicity to zebrafish (*Danio rerio*).

Introduction

Accelerated advancements in nanotechnology and nanoscience have found applications in a variety of scientific fields, leading to a rapid increase in the types of engineered nanoparticles on the market. In particular, zinc oxide nanoparticles (ZnO NPs) are the third highest production volume nanoparticles at roughly 550 tons per year [1]. Given their value as UV-protects [2], self-cleaning surfaces [3], sensors [4] and catalysts [5], it is expected that the use of engineered ZnO NPs will continue to increase with the increasing market demand. Such widespread use will also inevitably result in increased environmental release and a higher potential for human exposure [6]. As such, understanding which features of ZnO NPs increase their risks to humans and/or the environment is of paramount importance [7]. Despite this fact, very few studies to date have looked across a wide-range of engineered ZnO nanoparticle types to investigate how surface chemical modifications alter toxicity.

The toxicity of ZnO NPs to a wide range of species can be found elsewhere in literature from in vivo [8,9] to in vitro studies [10,11]. Bare ZnO NPs (lacking surface ligands) are known to cause delayed embryo hatching, developmental abnormalities [12] through dissolution and release of ionic zinc [13,14] as well as induction of DNA damage through generation of reactive oxidative species (ROS) [12,15]. ZnO NPs are often coated with a variety of capping agents or surface ligands with differing chemical properties to functionalize the surface and improve stability against agglomeration and dispersibility in a given medium [16]. These surface alterations have the potential to alter their toxicity as a result of differences in the release of Zn^{2+} ions and ROS production compared to bare ZnO NPs [17,18]. In addition, the behaviour of surface functionalized ZnO NPs may vary compared to non-functionalized (bare) ZnO NPs by altering stability and/or agglomeration, potentially altering bioavailability and toxicity to aquatic organisms [18-21]. While the dissolution kinetics and agglomeration state of the ZnO NPs is known to influence the toxicity of the materials, this study aimed to determine if specific intrinsic features could be used in lieu of empirical data on the material's behaviour.

Surface chemical ligands and capping agents are more closely related to the fate and effects of ZnO NPs than the core composition alone [18,19,22]. Thus, it is expected that surface chemical properties can be employed as descriptors to model the toxicity of various types of engineered ZnO NPs. The development of such relationships between a set of intrinsic properties of ligands and/or capping agents with their biological effects could serve as the basis of nanomaterial structure–activity relationships (nanoSARs) [23,24]. However, there is a limited understanding of how to link different nanoparticle surface chemistries directly to the fate and effects of ZnO NPs in organisms, and whether these properties can be used to develop predictive models useful in the development of safer engineered ZnO materials [7].

The main objective of this study were 1) to investigate whether the intrinsic properties of different capping agents or surface ligands of engineered ZnO NPs alter their toxicity and 2) to determine if these features can be used to model the developmental toxicity of ZnO nanoparticles to embryonic zebrafish (*Danio rerio*) (Figure 1). Zebrafish embryos were selected as vertebrate test species as their transparent tissues allow for easy visual assessment of multiple developmental malformations and

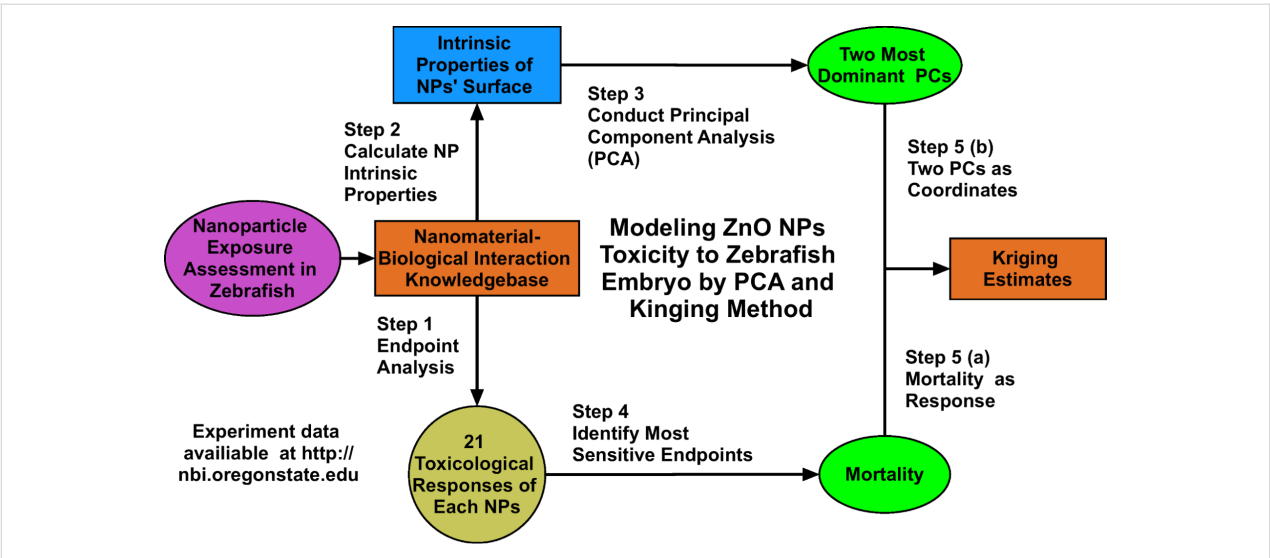


Figure 1: Data processing for model development.

their rapid development makes them ideal for studies of numerous types of NPs [25,26]. Due to the agglomeration of ZnO NPs in fishwater, the chorionic membrane can serve as a barrier to the direct interaction of NPs or dissolved oxygen with the developing embryo, thus we chose to remove this barrier in our study. The removal also allows for the visual analysis of the developing embryo, which can be hampered when the chorion is intact and coated with nanoparticles [25,27]. To achieve these objectives, we conducted zebrafish embryo toxicity testing for 17 different types and sizes of ZnO NPs with differing surface chemistries. Then, using bare and surface modified NP toxicity data and eight intrinsic chemical properties related to the outermost surface chemistry, we conducted principal component analysis (PCA) to extract descriptors useful as coordinates to develop a model of how surface chemistry impacts ZnO NP toxicity.

Selected surface features used in the PCA were those deemed likely to influence biological interactions with the NP surface. Size (SZ) was chosen as it has been reported by others to influence NP toxicity [11,28]. Hydrophobicity was selected as the Log P (partitioning coefficient) of NPs has been found to be related to toxic responses in other organisms [29]; however, since ZnO NPs can release zinc ions [30] and Log P is pH-independent [31], distribution coefficient (Log D) was also considered for both ionic and non-ionic forms. Polarizability was selected (PL) as a factor to describe the molecules electronic properties and its ability to change with external fields in biochemical reactions [32]. Polar surface area (PS) represents the area formed by the polar areas of the molecule and has been used to predict drug intestinal absorption in humans, thus it may be a useful predictor of other biological interactions [33]. Van der Waals (VDW) surface area calculated by VDW radius, is associated with the likelihood of NP agglomeration [34]. Solvent accessible surface area (SASA) can be used to estimate the protein-ligand binding free energy [35], and molar refractivity (RF) represents the energy required to polarize one mole of the substance and is associated with receptor binding affinity [36]. Dreiding energy (DE) will be used to predict the binding affinity of organic molecules with Zn and membrane proteins [37]. Although zeta potential is known to be crucial to biological response [38]; it's dependent on the environment in which it is measured and thus is not an intrinsic feature of the NP and thus was omitted from the model.

Following PCA, the ordinary kriging (OK) method was applied to estimate the pattern of variation of mortality in a given coordinate system. We hypothesized that surface chemical modifications would result in significant alterations in toxicity that would depend on the type of surface chemical modification performed.

Results

Estimation of intrinsic capping agent properties

The 17 ZnO NPs (Table 1) had 6 different surface chemistries including bare ZnO, oleic acid, octanoic acid, para-nitrobenzoic acid, cyclohexanecaboxylic acid and benzoic acid (Figure 2). The average primary particle sizes in this study ranged from 4 to 70 nm (Table 1). Table 2 provides the values

Table 1: Description of zinc oxide nanoparticles included in this study (17 in total).

NBI record	Particle descriptor	Manufacturer	Surface group	Size (nm)
nbi_085	ZnO + oleic acid	Voxtel	oleic acid	62
nbi_086	ZnO + oleic acid	Voxtel	oleic acid	26
nbi_087	ZnO	Sigma-Aldrich	—	62
nbi_088	ZnO	Voxtel	—	26
nbi_089	ZnO + octanoic acid	Voxtel	octanoic acid	62
nbi_090	ZnO + octanoic acid	Voxtel	octanoic acid	26
nbi_091	ZnO + para-nitrobenzoic acid	Voxtel	para-nitrobenzoic acid	62
nbi_092	ZnO + para-nitrobenzoic acid	Voxtel	para-nitrobenzoic acid	26
nbi_093	ZnO + cyclohexane carboxilic acid	Voxtel	cyclohexane carboxilic acid	62
nbi_094	ZnO + cyclohexane carboxilic acid	Voxtel	cyclohexane carboxilic acid	26
nbi_095	ZnO + benzoic acid	Voxtel	benzoic acid	62
nbi_096	ZnO + benzoic acid	Voxtel	benzoic acid	26
nbi_136	ZnO	Boise State University	—	14.6
nbi_137	ZnO	Boise State University	—	33.6
nbi_138	ZnO	Boise State University	—	4.5
nbi_139	ZnO	Boise State University	—	10.2
nbi_187	NanoGard ZnO (NGZ)	Alfa Aesar, NanoGard, Prod.#44898, lot#D28X017	—	70

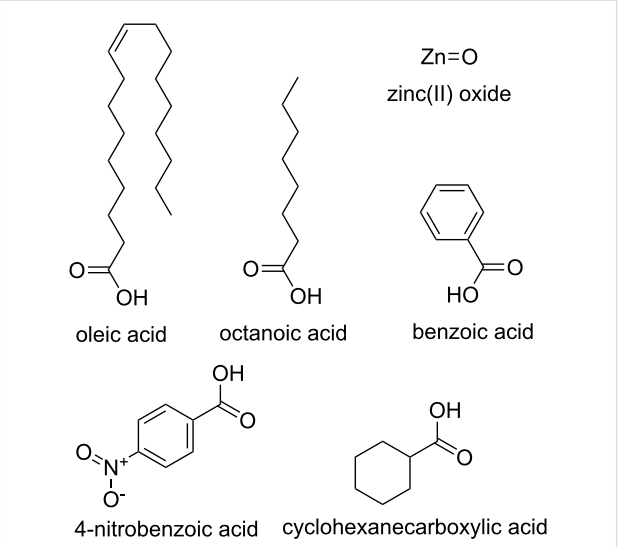


Figure 2: Chemical structures used to calculate the surface properties.

calculated for the intrinsic features of the 6 surface chemistries. The calculated distribution coefficient (Log D) had the least variance of all the parameters ranging from −1.22 to 5.62. Van der Waal surface area is the surface of the union of the spherical atomic surfaces defined by the van der Waals radius of each component atom in the molecule. Van der Waal surface area values for bare ZnO were 50.3 $Å^2$ and ranged from 173 to 560.40 $Å^2$ for other surface chemistries. These values had the highest variance in our estimations.

ZnO nanoparticle toxicity

Embryonic zebrafish mortality was concentration dependent and varied with different types of bare and surface engineered ZnO NPs as expected. Mortality for the bare and surface modified ZnO NPs as a function of exposure concentration is shown in Figure 3. Surface modified ZnO particles caused significant mortality at 24 hpf, in some cases at exposure concentrations as low as 0.08 mg/L; however, despite the exposures continuing until 120 hpf, no significant mortality or developmental problems were noted after 24 hpf (Figure 3A). Bare ZnO NPs showed similar results with 2 out of 7 displaying no visible signs of toxicity at the highest concentration tested (Figure 3B). In contrast to the surface engineered particles, the toxicity of bare particles occurred more frequently at 120 hpf (3 out of 7 materials, Supporting Information File 2). Bare NanoGard ZnO (NGZ) showed the highest 120 hpf mortality of all the tested particles (bare and surface modified) with 100% mortality (n = 24 embryos) at 50 mg/L. In addition, NGZ was the only ZnO particle tested (bare or surface modified) that resulted in any significant sublethal responses, eliciting swim bladder malformations at 10 mg/L and notochord malformations at the highest exposure concentration (see Supporting Information File 1). The results of the endpoint analysis using the Fisher's exact test for all tested NPs are provided in Supporting Information File 2. Detailed raw toxicity data for each individual exposure is also available online from the Nanomaterial-Biological Interactions knowledgebase (nbi.oregonstate.edu) [39].

Analysis of the 5 pairs of surface modified particles, with the same surface chemistries and differing average particle sizes, showed no clear trend related to the primary particle size (Figure 3A). Smaller oleic acid coated ZnO NPs (26 nm) caused significant mortality at the highest test concentration that did not occur for the larger (62 nm) oleic acid functionalized particles. In contrast, the larger octanoic acid coated ZnO NPs caused significant mortality at 0.4 mg/L while the smaller 26 nm particles did not induce toxicity until exposure concentrations reached 50 mg/L. Similarly, the ZnO NPs coated with cyclohexane carboxylic acid had a significantly different mortality rate between sizes, with the larger particles being more toxic than the smaller version (p = 0.009, 0.234 respectively).

Principal components analysis

By selecting the most dominant components to explain the majority of data variance, PCA effectively reduced the dimen-

Table 2: Intrinsic properties of different surface chemistries.

Intrinsic descriptor	Oleic acid	Octanoic acid	4-Nitrobenzoic acid	Cyclohexane carboxylic acid	Benzoic acid	Zinc oxide
Log D	5.62	0.53	−1.22	−0.43	−1.08	−0.20
Polarizability ($Å^3$)	34.5	16.1	15.8	13.4	13.2	1.00
Polar surface area ($Å^2$)	37.3	37.3	83.1	37.3	37.3	17.1
VDW surface area ($Å^2$)	560	283	211	221	173	50.3
Solvent-accessible surface area ($Å^2$)	689	403	330	260	284	156
Molar refractivity (cm^3/mol)	87.1	40.7	39.7	39.7	33.2	1.44
Dreiding energy (kcal/mol)	35.7	12.1	23.1	24.8	16.6	0.00

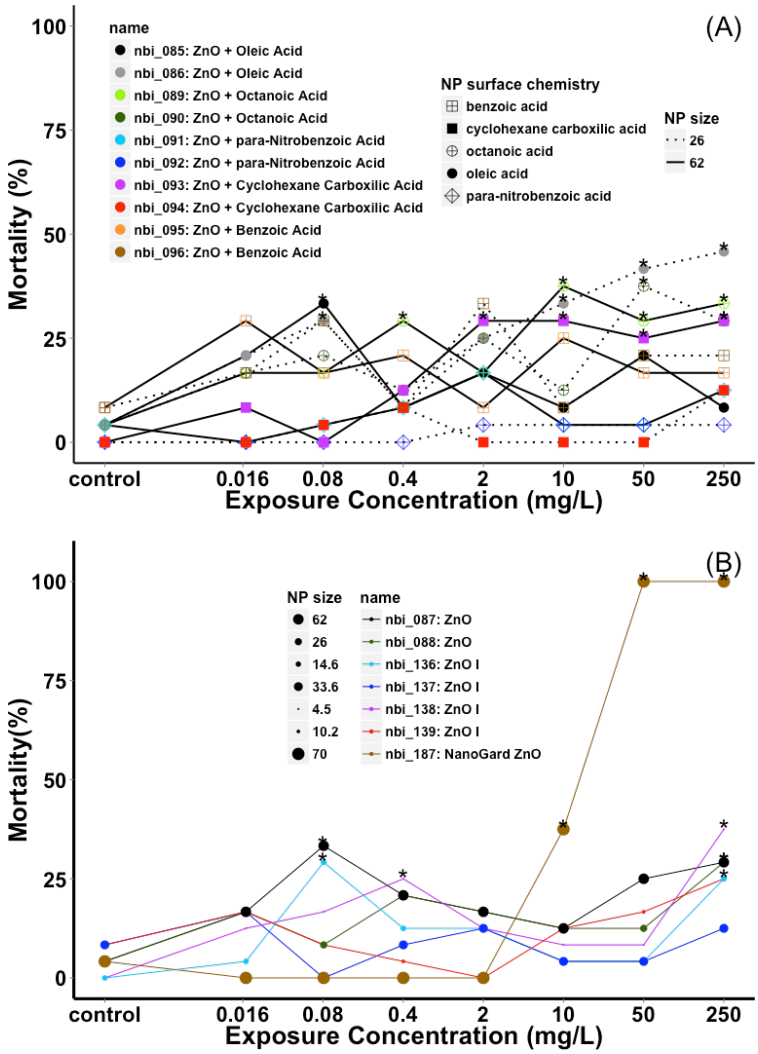


Figure 3: Zebrafish mortality at 120 hpf following exposure to: (A) ZnO NPs with and (B) without surface modification.

sions of the dataset with keeping most information. It eliminated the correlation between different independent variables by creating different linear combinations which are independent of each other [40]. PCA was conducted on the database that consists of 8 property descriptors: size (SZ), Log D, polarizability (PL), polar surface area (PS), van der Waals surface (VS), solvent-accessible surface area (SASA), molar refractivity (RF) and Dreiding energy (DE) with 10 surface modified and 7 bare ZnO NPs (17 ZnO NP datasets × 8 properties). Each individual NP exposure dataset is comprised of results from experiments conducted at 8 exposure concentrations, thus the final matrix of the database was comprised of 136 rows and 8 columns (17 materials × 8 concentrations × 8 surface chemical properties).

The first two principle components (PCs), whose standard deviations both were greater than 1, explained 87.3% of the total variance of the matrix. As the linear combinations (or weights) of these two PCs were calculated based on all of the input data, they represent all of the particle information. As such, these two PCs were determined to be appropriate to represent the variability in this dataset (Figure 4). These two PCs were selected as the new independent variables, reducing the independent variables' dimensions from 8 to 2.

Table 3 shows the 8 descriptors all have moderately similar weights in PC1, but Log D, PS and SZ have outstanding weights in PC2. The variable coefficients in the PC1 linear combination all have the same sign, suggesting these parameters have similar effects on the model. In contrast, the sign of the variable coefficients for SZ and PS in PC2 are opposite to the other parameters suggesting these variables help separate the particles. Graphing the PCA scores for PC1 versus PC2 allows for the use of Euclidean distance to identify clusters of

Table 3: Rotation of PCA (weighting of each property).

Property	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
SZ[a]	0.188	0.669	0.711	0.072	−0.077	−0.027	0.001	0.000
PS[b]	0.270	0.497	−0.610	0.454	−0.262	0.100	0.063	0.139
SASA[c]	0.404	−0.025	−0.002	0.173	0.844	0.196	−0.090	0.218
RF[d]	0.407	−0.058	−0.063	−0.205	−0.182	−0.320	−0.803	0.062
DE[e]	0.378	−0.001	−0.039	−0.634	−0.222	0.531	0.217	0.274
Log D[f]	0.292	−0.535	0.339	0.538	−0.359	0.142	0.069	0.266
VS[g]	0.410	−0.099	−0.015	0.053	−0.020	0.191	0.063	−0.882
PL[h]	0.408	−0.070	−0.051	−0.150	0.037	−0.714	0.536	0.072

[a]Size; [b]polar surface; [c]solvent-accessible surface area; [d]molar refractivity; [e]dreiding energy; [f]distribution coefficient; [g]van der Waals surface; [h]polarizability.

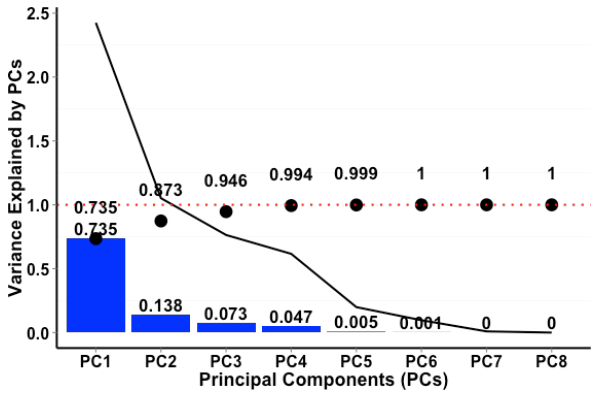


Figure 4: Individual variance for each of the principal components (PCs). Black dots represent the accumulated variance explained by each PC, while the solid line shows the Eigenvalue.

similar NPs with respect to their toxicity to embryonic zebrafish. As predicted, the various surface modifications to ZnO NPs resulted in distinct groupings based on these capping agent properties (Figure 5). When partitioned into three clusters, the plot shows a clear separation as: (Group 1) oleic acid; (Group 2) octanoic acid, para-nitrobenzoic acid, cyclohexane carboxylic acid and benzoic acid; (Group 3) bare ZnO with blank control responses (Figure 5). Similar analysis using either four or five clusters shows minor differences compared to the use of three clusters, namely the coated 26 nm NPs (except octanoic acid) separated out of Group 3 in the four cluster calculation and the blank control point separated out of Group 1 in the five clusters calculation in addition to 62 and 70 nm bare ZnO NP separating out of Group 3 (See Supporting Information File 3).

Estimation of toxicity by ordinary kriging method

By using the two most dominant PCs identified earlier as coordinates (XY-direction) and mortality data as the response (Z-direction), we calculated the kriging estimation of mortality. The ordinary kriging method, based on the spherical model, was used to model the mortality of zebrafish embryos at each of the different exposure concentrations for each of the 17 tested NPs. The resulting contour map for the highest exposure concentration (250 mg/L) is shown in Figure 6 and the contour maps for other exposure concentrations can be found in Supporting Information File 4. The coefficient of determination was calculated to determine how well the estimation fit the original data. Similar coefficients of determination were found at each concentration (0.702–0.778).

Discussion

ZnO NP toxicity to embryonic zebrafish

Of the numerous sub-lethal endpoints evaluated in our study, most of the significant toxicity resulting from exposure to ZnO NPs was associated with mortality, regardless of the type of surface chemistry found on the nanoparticle. Interestingly, when mortality occurred in the surface functionalized ZnO NPs, it was always within the first 16–18 hours of exposure (observed at the 24 hpf evaluation). Embryos surviving exposure to surface coated ZnO NPs after this initial period had almost 100% survival and no significant developmental abnormalities (see Supporting Information File 1 and Supporting Information File 5). In contrast, the bare ZnO particles resulted in mortality at both 24 and 120 hpf for some materials and a complete lack of toxicity in others. This result supports the hypothesis that outermost surface chemistry is a primary driver of biological interactions, even more than core composition. This finding has been supported in other studies investigating a wide range of NP types [27,41,42].

Given that dissolution and the resulting release of zinc ions and ROS are the primary cause of ZnO NP toxicity [8], it is possible that the lack of late-onset mortality in coated particles is the result of decreased dissolution of these particles [7,21]. It has been reported that the release of zinc ion from ZnO NPs coated

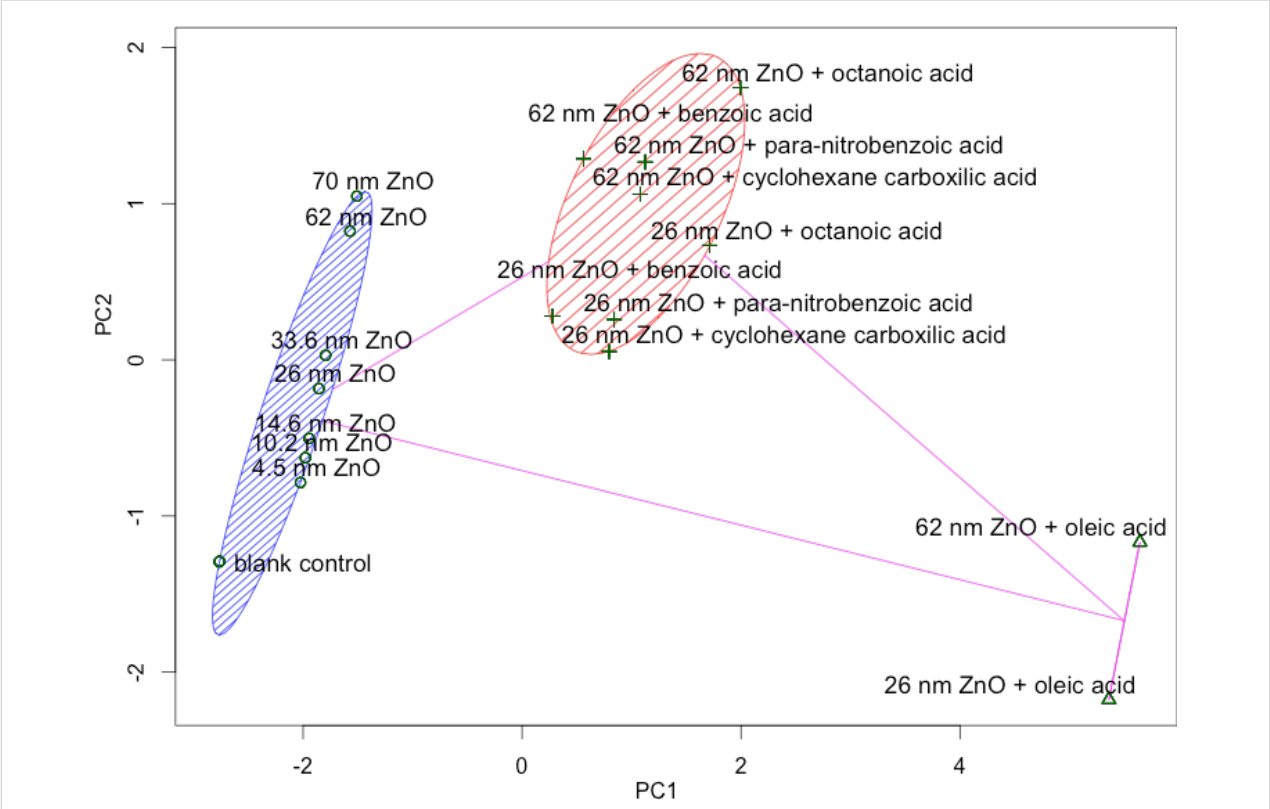


Figure 5: Clustering analysis based on Euclidian distance for ZnO NPs partitioned into 3 clusters. Shown on the left (blue hash marks) are the bare ZnO NPs with the blank control point. In the middle (tan hash marks) are ZnO NPs with 4 different surface chemistries and on the right are the oleic acid modified particles.

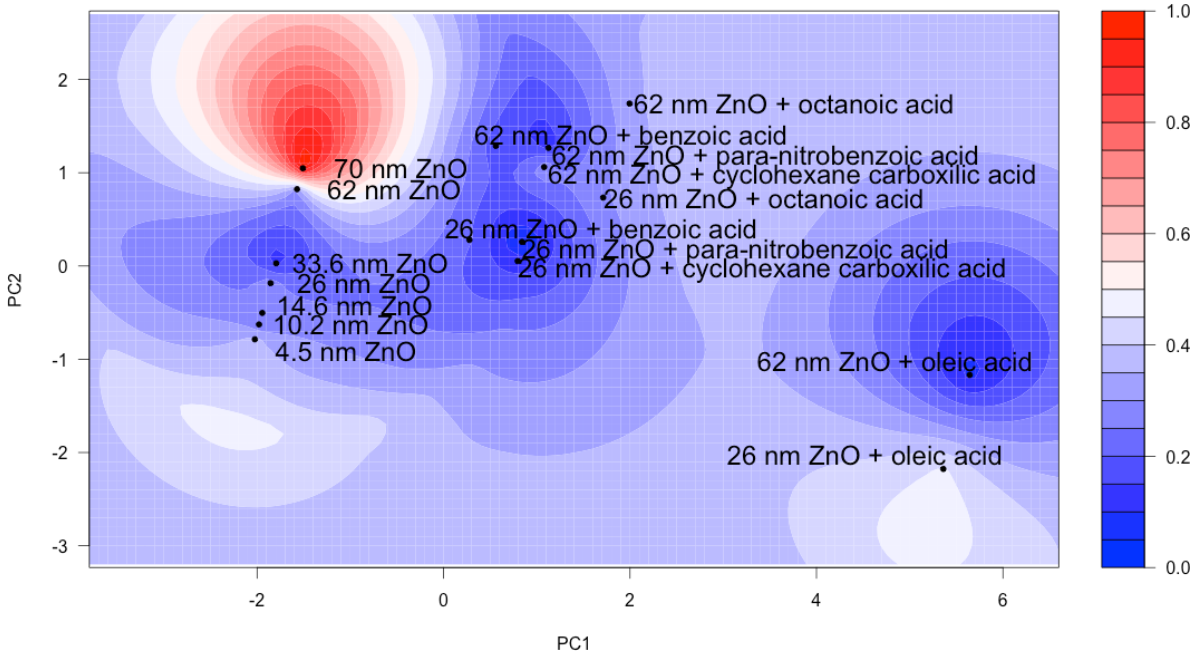


Figure 6: Kriging estimation contour map for embryonic zebrafish exposed to 250 mg/L of each type of zinc oxide nanoparticle using the first two surface chemistry-based principal components as the coordinates and 120 hpf total mortality as response. The coefficient of determination was found to be 0.702.

with organic molecule can be slower than uncoated ZnO NPs by up to 10 days, due to the protective effect of the surface coating [43]. The idea that coated particles were more benign overall is also supported by the most toxic response being noted for a bare particle (NGZ, Figure 3). In addition, the observed mortality at 24 hpf for some of the surface functionalized particles could have been due to either residual impurities or zinc ions, as any dissolved zinc would have remained in the exposure media due to the static nature of these experiments. The delayed mortality response in the bare ZnO particles could also relate to the onset of mouth-gaping behavior during fish development that led to increased uptake over the exposure period; however, this would likely have occurred with the coated particles as well unless this was specific to zinc ion uptake or direct impacts of generated ROS.

Only one ZnO NP (NGZ) caused any significant sublethal impacts in the developing fish with notochord malformations as well as significant malformations of the swim bladder. Despite NGZ being an uncoated ZnO NP, its unique toxicity relative to the other non-coated ZnO NPs suggests some other features, such as crystal morphology, may be contributing to the observed differential toxicity. It is known that ZnO NPs with sharper angles have been noted to contribute to lower viability in cell culture studies with A549 and HT29 cells [30]. Similar morphology effects on toxicity have been observed in studies of manganese oxide, where the sharp points and edges were found to generate more ROS than smooth surfaces [44]. We tested this hypothesis by comparing X-ray diffraction (XRD) results for NGZ relative to a representative sample of the other bare ZnO NPs (Sigma-Aldrich, 63 nm, NBI_0215) using a Bruker-AXS D8 Discover XRD instrument (Karlsruhe, Germany and Madison, WI). No differences in the lattice parameters were identified, thus other intrinsic factors must be contributing to the unique toxicity of this commercial ZnO NP (see Supporting Information File 6).

Since the size of the ZnO NP did not elicit any general trends in the toxic responses observed, it is likely that surface features of the particle impacting interactions with biological membranes may drive toxicity more than the size of the particle itself. NP agglomeration in aquatic environments often occurs and can be influenced by physicochemical properties of the particle surface and environmental factors affecting the zeta potential [27,45,46]. Therefore, it is possible that the agglomeration of the particles in the fishwater media could indirectly affect dissolution or interactions with the developing embryo. Previous studies have found that uncoated ZnO NPs form smaller aggregates on the surface of bacteria than are formed in suspension [47], and this type of surface aggregation cannot be ruled out as a contributing factor in our results. Previous studies with the freshwater crustacean *Daphnia magna* based on 30, 80–100 and 200 nm ZnO NPs found that toxicity was not dependent on the primary particle size [11]. This is similar to what we found for the bare ZnO NPs in our study which range from 4 to 70 nm.

Overall, the toxicity results suggest that surface features do impact ZnO NP toxicity. In addition, the evaluation or mortality at multiple time points during development is useful in modeling nanoparticle–biological interactions using zebrafish [45].

PCA

PCA combines as much information as possible to provide an overview of the known and unknown relationships between inherent NP features and developmental toxicity. The eight original intrinsic properties descriptors were correlated with each other based on similarities in value of PC1 weights, however more separation was gained using the weighting of PC2 (Table 3). The latent factor suggested by PC2 is the Log D, which plays a different role in the ZnO NPs toxicity compare to size and polar surface effects. The unique clustering of both sizes of oleic acid functionalized particles suggests the properties of this ligand are somewhat unique relative to the others, perhaps due to the long chain length (Figure 2) and high hydrophobicity of oleic acid (Table 2). Oleic acid coated ZnO NPs which have the highest hydrophobicity (Log D 5.62), showed the smaller size one was more toxic and separated from the remainder of the coated particles in the PCA. In contrast, the remaining surface functionalized particles all had much lower log D values (Table 2) and clustered together in our analysis. The Log D calculations can be affected by electrolyte concentration, however in our study this was too small (Cl^- 0.0174 mol/L and Na^+, K^+ 0.0165 mol/L) to affect its value relative to water, thus these inherent properties value are expected to reflect the true properties in fishwater. This suggests that future studies should continue to investigate surface features impacting the hydrophobicity of the particle as potential contributors to toxicity. However, this result depends on our assumption that the coating chemicals dominate the hydrophobicity of the metal oxide NP [22]. Even when surface chemistry is constant among ZnO NPs, differential particle morphology and variations in the suspension media will likely affect dissolution and alter the hydrophobicity in comparison to theoretical values of Log D [30].

Other intrinsic properties not considered, such as the proportional amount of ligand coverage on the surface of the nanoparticle, may improve model performance further. Unfortunately this level of detailed characterization of the surface chemistry is often unavailable from manufacturers and is cost- and time-

intensive to determine for a wide range of surface chemistries. Further refinement of the model could likely also be achieved by including more complex calculation of intrinsic values that are based on the actual ligand-nanoparticle structure rather than surface ligand structure alone (in the absence of consideration of bonding with the NP). In studies of multiple engineered nanoparticles, it is nearly impossible to set single variable control groups due to correlated descriptors and constraints in characterizing NPs in the experiment conditions. However, we have shown that PCA can be used as a valuable alternative method to estimate the relative effects of multiple inherent properties simultaneously to support the development of predictive models that will allow for the development of safer ZnO materials.

Based on the large differences in molecular properties between the organic surface coatings and the bare zinc oxide properties (Table 2), it was expected that each group would separate during clustering analysis, as was the case with this data (Figure 5). Identified clusters suggest that a set of appropriate intrinsic properties of surface chemistry can be used to partition NPs into different groups. The 17 ZnO NPs partitioned into clusters that were fairly easy to identify using only capping agent properties. However, with more complex surface structures, overlap between clusters might happen making determination of the cluster number the first concern. Although there are several algorithms to decide the cluster number, the lack of robust data sets such as this preclude a current understanding of which algorithm may be appropriate [48].

Kriging estimation

Based on the two most dominant PCs that explained 87.3% of the variance in the toxicity data, we performed the kriging estimation at each of the exposure concentrations. Interestingly, the exposure concentrations had little influence on the coefficients of determination with similar values being determined at each concentration (Figure 6, Supporting Information File 4). Kriging estimation further elucidated the impacts of NP size. Based on Figure 6, we can see that the largest bare particle (NGZ) also has the highest mortality (Figure 3B) and the cluster 2 surface modified 26 nm particles were predicted to have overall lower toxicity than the larger versions of the same particle. However, this trend does not hold for the oleic acid functionalized particles as the smaller particles are predicted to be higher in toxicity. Therefore, outermost surface chemistry continues to play a more important role in determining toxicity.

Conclusion

The observed toxic responses of developing zebrafish embryos to ZnO NP exposure varied with surface chemical modification and were only minimally impacted by particle size. Only NGZ, a bare ZnO NP, had relatively high toxicity, suggesting specific product features of bare ZnO NPs drive toxicity. This work has shown that large databases of similar NPs with varying surface features studied under identical experimental design protocols, are invaluable in the development of models of nanoparticle-biological interactions. We have shown that intrinsic features of NPs, particularly those encompassing the outermost surface chemistry, are useful in the classification and clustering of NP toxicity data. Our finding that hydrophobicity was the strongest determinant of toxicity of the many surface features we investigated will contribute to the development of predictive models of ZnO NP-biological interactions. We have found that PCA is a useful tool for reducing numerous surface molecular properties to fewer dimensions. Future development of highly accurate predictive models will depend on detailed information provided by in silico modeling and analysis of the outermost surface of the nanoparticle. Overall, identification of specific material features, such as outermost surface chemistry, that drive biological interactions appears feasible and models such as this should continue to be tested and refined to achieve safer design principles for the manufacture of ZnO NPs.

Experimental

Nanomaterials

The ZnO NPs with different capping agents and sizes were obtained from a variety of commercial and research laboratories (Table 1). More detailed characterization of the nanomaterials are also available on the open-source Nanomaterial-Biological Interactions Knowledgebase [39] provided by Oregon State University.

Estimation of surface chemical parameters

The eight surface chemical descriptors we utilized were size, hydrophobicity (Log D), polarizability, polar surface area, van der Waals surface area, solvent accessible surface area, molar refractivity and Dreiding energy (Table 2). Except for the primary particle sizes (which were provided by manufacturers), the seven other intrinsic properties of capping agents were calculated by software (Table 2). Log D is calculated using Advanced Chemistry Development (ACD/Labs) Software version 11.02. PL is retrieved from ChemSpider (Mar. 2014), which was predicted by ACD/Labs Percepta Platform - PhysChem Module. VDW surface (VS), PS, SASA, RF and DE were calculated in Marvin Beans (version 6.2.2, Cambridge, MA). All inherent chemical properties were calculated based on the pH used in zebrafish toxicity test.

Embryonic zebrafish assay

Wild-type 5D zebrafish (*Danio rerio*) embryos were obtained from group spawns of adult fish housed at the Sinnhuber

Aquatic Research Laboratory at Oregon State University (Corvallis, OR). All NP dilutions and exposures were conducted in fish water (FW). The FW was prepared with 0.26 g/L Instant Ocean salts (Aquatic Ecosystem, Apopka, FL) combined with approximately 0.01g $NaHCO_3$ pH buffer in reverse osmosis water (pH 7.0–7.4, conductivity 450–600 μS). Embryos were collected at 6 hours post-fertilization (hpf) and maintained at 27 °C under 14/10 light and dark cycle. Embryos were exposed individually in 96-well plates to 7 different concentrations (0.016 to 250 mg/L) of each type of ZnO NP suspended in FW. Prior to exposure, embryos were dechorionated at 6 hours post-fertilization (hpf) with pronase (Sigma-Aldrich) and then rinsed several times with FW [25]. The control groups are FW alone without NPs present. A total of 21 endpoints were observed during development at 24 and 120 hpf that included mortality as well as morphological, behavioral and developmental endpoints in sub-lethal exposures [49]. The 19 sub-lethal endpoints include developmental progression (DP), spontaneous movement (SP), notochord (N), yolk sac edema (Y), axis (A), eye (E), snout (Sn), jaw (J), otic (O), heart (H), brain (B), somite (So), pectoral fin (PF), caudal fin (CF), pigment (P), circulation (C), trunk (T), swim bladder (SB), and touch response (TR).

Statistical analysis

Due to the non-parametric nature of the data and the small sample size (<30 embryos for each exposure concentration), the Fisher's exact test (Sigma Plot v12.0, San Jose, CA) was used to analyze individual endpoints recorded at 24 and 120 hpf [50]. *P*-value was calculated based on two-tailed test and a $p \leq 0.05$ significance level was maintained for all analyses. Mortality data was compared between NPs with the same capping agent but different sizes using two-way analysis of variance (R, version 3.1.0, Vienna, Austria).

Principal component analysis (PCA) was conducted in R using the primary particle size and seven intrinsic properties of NPs' surface chemistry shown in Table 1 and Table 2, respectively. To include control groups (blank group) in the analysis, all of the intrinsic NP properties are set to 0 for the blank groups. The same intrinsic properties were used for all exposure concentrations (0.016 mg/L to 250 mg/L) for a given particle type. The normalization process was conducted on the dataset as a matrix in PCA, with the mean of normalized data equal to 0 and standard deviation equal to 1. Then 8 different linear combinations consisting of 8 independent variables and their coefficients (also called "rotation" in R) were generated as new vectors, called principal components (PCs). The converted value, called score (stored as "x" in R), was used to model the toxic responses. The ordinary kriging was conducted in R using the additional "Kriging" and "gstat" packages.

Supporting Information

Supporting Information File 1

Zebrafish malformation and behavioral data. The 19 sub-lethal endpoints are developmental progression (DP), spontaneous movement (SP), notochord (N), yolk sac edema (Y), axis (A), eye (E), snout (Sn), jaw (J), otic (O), heart (H), brain (B), somite (So), pectoral fin (PF), caudal fin (CF), pigment (P), circulation (C), trunk (T), swim bladder (SB), and touch response (TR).
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-160-S1.xlsx]

Supporting Information File 2

Fisher's exact test *p*-value. The 19 sub-lethal endpoints are developmental progression (DP), spontaneous movement (SP), notochord (N), yolk sac edema (Y), axis (A), eye (E), snout (Sn), jaw (J), otic (O), heart (H), brain (B), somite (So), pectoral fin (PF), caudal fin (CF), pigment (P), circulation (C), trunk (T), swim bladder (SB), and touch response (TR). Included are three mortality (M) endpoints at 24 and 120 hours post fertilization after the exposure to ZnO NP and the sum of two M.
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-160-S2.xlsx]

Supporting Information File 3

Cluster analysis of converted data using Euclidean distance to partition into A) 3, B) 4, C) 5, D) 6 clusters.
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-160-S3.png]

Supporting Information File 4

Kriging estimations of zebrafish mortality data at A) 0.016 ppm, B) 0.08 ppm, C) 0.4 ppm, D) 2 ppm, E) 10 ppm, F) 50 ppm.
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-160-S4.png]

Supporting Information File 5

Embryonic zebrafish mortality at 24 and 120 hours post fertilization after ZnO NP exposure.
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-160-S5.xlsx]

Supporting Information File 6

XRD analysis of three different ZnO NPs.
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-160-S6.png]

Acknowledgements

We thank the staff of Sinnhuber Aquatic Research Laboratory for providing the embryos used in these studies. These studies were supported with funding from #ES017552-01A2; #P30ES03850; #ES0166896-01; #FA8650-05-1-15041; #P30E5000210.

References

1. Piccinno, F.; Gottschalk, F.; Seeger, S.; Nowack, B. *J. Nanopart. Res.* **2012,** *14,* No. 1109. doi:10.1007/s11051-012-1109-9
2. Osterwalder, U.; Sohn, M.; Herzog, B. *Photodermatol., Photoimmunol. Photomed.* **2014,** *30,* 62–80. doi:10.1111/phpp.12112
3. Gao, D.; Chen, C.; Ma, J.; Duan, X.; Zhang, J. *Chem. Eng. J.* **2014,** *258,* 85–92. doi:10.1016/j.cej.2014.07.072
4. Khan, S. B.; Rahman, M. M.; Asiri, A. M.; Asif, S. A. B.; Al-Qarni, S. A. S.; Al-Sehemi, A. G.; Al-Sayari, S. A.; Al-Assiri, M. S. *Physica E* **2014,** *62,* 21–27. doi:10.1016/j.physe.2014.04.007
5. Assi, N.; Sharif, A. A. M.; Bakhtiari, H.; Naeini, Q. S. M. *Int. J. Nano Dimens.* **2014,** *5,* 145–154.
6. Dumont, E.; Johnson, A. C.; Keller, V. D. J.; Williams, R. J. *Environ. Pollut. (Oxford, U. K.)* **2015,** *196,* 341–349. doi:10.1016/j.envpol.2014.10.022
7. Ramasamy, M.; Das, M.; An, S. S. A.; Yi, D. K. *Int. J. Nanomed.* **2014,** *9,* 3707–3718. doi:10.2147/IJN.S65086
8. Buerki-Thurnherr, T.; Xiao, L.; Diener, L.; Arslan, O.; Hirsch, C.; Maeder-Althaus, X.; Grieder, K.; Wampfler, B.; Mathur, S.; Wick, P.; Krug, H. F. *Nanotoxicology* **2013,** *7,* 402–416. doi:10.3109/17435390.2012.666575
9. Ma, H.; Williams, P. L.; Diamond, S. A. *Environ. Pollut.* **2013,** *172,* 76–85. doi:10.1016/j.envpol.2012.08.011
10. Adam, N.; Schmitt, C.; Galceran, J.; Companys, E.; Vakurov, A.; Wallace, R.; Knapen, D.; Blust, R. *Nanotoxicology* **2014,** *8,* 709–717. doi:10.3109/17435390.2013.822594
11. Lopes, S.; Ribeiro, F.; Wojnarowicz, J.; Łojkowski, W.; Jurkschat, K.; Crossley, A.; Soares, A. M. V. M.; Loureiro, S. *Environ. Toxicol. Chem.* **2014,** *33,* 190–198. doi:10.1002/etc.2413
12. Zhao, X.; Wang, S.; Wu, Y.; You, H.; Lv, L. *Aquat. Toxicol.* **2013,** *136–137,* 49–59. doi:10.1016/j.aquatox.2013.03.019
13. Leung, Y. H.; Chan, C. M. N.; Ng, A. M. C.; Chan, H. T.; Chiang, M. W. L.; Djurišić, A. B.; Ng, Y. H.; Jim, W. Y.; Guo, M. Y.; Leung, F. C. C.; Chan, W. K.; Au, D. T. W. *Nanotechnology* **2012,** *23,* 475703. doi:10.1088/0957-4484/23/47/475703
14. Applerot, G.; Lipovsky, A.; Dror, R.; Perkas, N.; Nitzan, Y.; Lubart, R.; Gedanken, A. *Adv. Funct. Mater.* **2009,** *19,* 842–852. doi:10.1002/adfm.200801081
15. Bai, W.; Zhang, Z.; Tian, W.; He, X.; Ma, Y.; Zhao, Y.; Chai, Z. *J. Nanopart. Res.* **2010,** *12,* 1645–1654. doi:10.1007/s11051-009-9740-9
16. Meißner, T.; Oelschlägel, K.; Potthoff, A. *Int. Nano Lett.* **2014,** *4,* No. 116. doi:10.1007/s40089-014-0116-5
17. Tang, K.; Liu, X.; Harper, S. L.; Steevens, J. A.; Xu, R. *Int. J. Nanomed.* **2013,** *8* (Suppl. 1), 15–29. doi:10.2147/IJN.S40974
18. Punnoose, A.; Dodge, K.; Rasmussen, J. W.; Chess, J.; Wingett, D.; Anders, C. *ACS Sustainable Chem. Eng.* **2014,** *2,* 1666–1673. doi:10.1021/sc500140x
19. Tang, E.; Cheng, G.; Ma, X.; Pang, X.; Zhao, Q. *Appl. Surf. Sci.* **2006,** *252,* 5227–5232. doi:10.1016/j.apsusc.2005.08.004
20. Ramasamy, M.; Kim, Y. J.; Gao, H.; Yi, D. K.; An, J. H. *Mater. Res. Bull.* **2014,** *51,* 85–91. doi:10.1016/j.materresbull.2013.12.004
21. Merdzan, V.; Domingos, R. F.; Monteiro, C. E.; Hadioui, M.; Wilkinson, K. J. *Sci. Total Environ.* **2014,** *488–489,* 316–324. doi:10.1016/j.scitotenv.2014.04.094
22. Xiao, Y.; Wiesner, M. R. *J. Hazard. Mater.* **2012,** *215–216,* 146–151. doi:10.1016/j.jhazmat.2012.02.043
23. Fourches, D.; Muratov, E.; Tropsha, A. *J. Chem. Inf. Model.* **2010,** *50,* 1189–1204. doi:10.1021/ci100176x
24. Puzyn, T.; Rasulev, B.; Gajewicz, A.; Hu, X.; Dasari, T. P.; Michalkova, A.; Hwang, H.-M.; Toropov, A.; Leszczynska, D.; Leszczynski, J. *Nat. Nanotechnol.* **2011,** *6,* 175–178. doi:10.1038/nnano.2011.10
25. Usenko, C. Y.; Harper, S. L.; Tanguay, R. L. *Carbon* **2007,** *45,* 1891–1898. doi:10.1016/j.carbon.2007.04.021
26. Harper, S. L.; Carriere, J. L.; Miller, J. M.; Hutchison, J. E.; Maddux, B. L. S.; Tanguay, R. L. *ACS Nano* **2011,** *5,* 4688–4697. doi:10.1021/nn200546k
27. Bonventre, J. A.; Pryor, J. B.; Harper, B. J.; Harper, S. L. *J. Nanopart. Res.* **2014,** *16,* No. 2761. doi:10.1007/s11051-014-2761-z
28. Kim, K.-T.; Truong, L.; Wehmas, L.; Tanguay, R. L. *Nanotechnology* **2013,** *24,* 115101. doi:10.1088/0957-4484/24/11/115101
29. Moyano, D. F.; Goldsmith, M.; Solfiell, D. J.; Landesman-Milo, D.; Miranda, O. R.; Peer, D.; Rotello, V. M. *J. Am. Chem. Soc.* **2012,** *134,* 3965–3967. doi:10.1021/ja2108905
30. Mu, Q.; David, C. A.; Galceran, J.; Rey-Castro, C.; Krzemiński, L.; Wallace, R.; Bamiduro, F.; Milne, S. J.; Hondow, N. S.; Brydson, R.; Vizcay-Barrena, G.; Routledge, M. N.; Jeuken, L. J. C.; Brown, A. P. *Chem. Res. Toxicol.* **2014,** *27,* 558–567. doi:10.1021/tx4004243
31. LogP and logD calculations. https://docs.chemaxon.com/display/CALCPLUGS/LogP%20and%20logD%20calculations (accessed March 30, 2015). https://docs.chemaxon.com/display/CALCPLUGS/LogP and logD calculations.
32. Hansch, C.; Steinmetz, W. E.; Leo, A. J.; Mekapati, S. B.; Kurup, A.; Hoekman, D. *J. Chem. Inf. Comput. Sci.* **2003,** *43,* 120–125. doi:10.1021/ci020378b
33. Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. *Pharm. Res.* **1997,** *14,* 568–571. doi:10.1023/A:1012188625088
34. Nel, A. E.; Mädler, L.; Velegol, D.; Xia, T.; Hoek, E. M. V.; Somasundaran, P.; Klaessig, F.; Castranova, V.; Thompson, M. *Nat. Mater.* **2009,** *8,* 543–557. doi:10.1038/nmat2442
35. Wang, J.; Hou, T. *J. Chem. Inf. Model.* **2012,** *52,* 1199–1212. doi:10.1021/ci300064d
36. Crippen, G. M. *J. Comput. Chem.* **1999,** *20,* 1577–1585. doi:10.1002/(SICI)1096-987X(19991115)20:14<1577::AID-JCC11>3.0.CO;2-I
37. Mayo, S. L.; Olafson, B. D.; Goddard, W. A. *J. Phys. Chem.* **1990,** *94,* 8897–8909. doi:10.1021/j100389a010
38. Bhattacharjee, S.; Ershov, D.; Islam, M. A.; Kämpfer, A. M.; Maslowska, K. A.; van der Gucht, J.; Alink, G. M.; Marcelis, A. T. M.; Zuilhof, H.; Rietjens, I. M. C. M. *RSC Adv.* **2014,** *4,* 19321–19330. doi:10.1039/C3RA46869K
39. http://nbi.oregonstate.edu (accessed March 1, 2015).
40. Jolliffe, I. T. *Principal component analysis;* Springer Series in Statistics; Springer: Berlin, Germany, 1986; p 27. doi:10.1007/978-1-4757-1904-8
41. Walkey, C. D.; Olsen, J. B.; Guo, H.; Emili, A.; Chan, W. C. W. *J. Am. Chem. Soc.* **2012,** *134,* 2139–2147. doi:10.1021/ja2084338

42. Perreault, F.; Popovic, R.; Dewez, D. *Environ. Pollut. (Oxford, U. K.)* **2014,** *185,* 219–227. doi:10.1016/j.envpol.2013.10.027
43. Gelabert, A.; Sivry, Y.; Ferrari, R.; Akrout, A.; Cordier, L.; Nowak, S.; Menguy, N.; Benedetti, M. F. *Environ. Toxicol. Chem.* **2014,** *33,* 341–349. doi:10.1002/etc.2447
44. Gotić, M.; Jurkin, T.; Musić, S.; Unfried, K.; Sydlik, U.; Bauer-Šegvić, A. *J. Mol. Struct.* **2013,** *1044,* 248–254. doi:10.1016/j.molstruc.2012.09.083
45. Liu, X.; Tang, K.; Harper, S.; Harper, B.; Steevens, J. A.; Xu, R. *Int. J. Nanomed.* **2013,** *8* (Suppl. 1), 31–43. doi:10.2147/IJN.S40742
46. Hotze, E. M.; Phenrat, T.; Lowry, G. V. *J. Environ. Qual.* **2010,** *39,* 1909–1924. doi:10.2134/jeq2009.0462
47. Jiang, W.; Mashayekhi, H.; Xing, B. *Environ. Pollut.* **2009,** *157,* 1619–1625. doi:10.1016/j.envpol.2008.12.025
48. Mirkin, B. *Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery* **2011,** *1,* 252–260. doi:10.1002/widm.15
49. Truong, L.; Moody, I. S.; Stankus, D. P.; Nason, J. A.; Lonergan, M. C.; Tanguay, R. L. *Arch. Toxicol.* **2011,** *85,* 787–798. doi:10.1007/s00204-010-0627-4
50. Liao, Y.-Y.; Lee, T.-S.; Lin, Y.-M. *Radiology (Oak Brook, IL, U. S.)* **2006,** *239,* 300–301. doi:10.1148/radiol.2391051114

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Experiences in supporting the structured collection of cancer nanotechnology data using caNanoLab

Stephanie A. Morris*[1], Sharon Gaheen[2], Michal Lijowski[3], Mervi Heiskanen[4] and Juli Klemm[4]

Full Research Paper

Open Access

Address:
[1]Office of Cancer Nanotechnology Research, National Cancer Institute/NIH, 31 Center Drive, Bethesda, MD, 20892, USA, [2]Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research, 8560 Progress Drive, Frederick, MD, 21701, USA, [3]Essential Software, Inc., 9024 Mistwood Drive, Potomac, MD, 20854, USA and [4]Center for Biomedical Informatics and Information Technology, National Cancer Institute/NIH, 9609 Medical Center Drive, Rockville, MD, 20850, USA

Email:
Stephanie A. Morris* - morriss2@mail.nih.gov

* Corresponding author

Keywords:
caNanoLab; cancer research; databases; nanomaterials; nanomedicine

Beilstein J. Nanotechnol. **2015,** *6,* 1580–1593.
doi:10.3762/bjnano.6.161

Received: 16 April 2015
Accepted: 29 June 2015
Published: 21 July 2015

This article is part of the Thematic Series "Nanoinformatics for environmental health and biomedicine".

Guest Editor: R. Liu

Abstract

The cancer Nanotechnology Laboratory (caNanoLab) data portal is an online nanomaterial database that allows users to submit and retrieve information on well-characterized nanomaterials, including composition, in vitro and in vivo experimental characterizations, experimental protocols, and related publications. Initiated in 2006, caNanoLab serves as an established resource with an infrastructure supporting the structured collection of nanotechnology data to address the needs of the cancer biomedical and nanotechnology communities. The portal contains over 1,000 curated nanomaterial data records that are publicly accessible for review, comparison, and re-use, with the ultimate goal of accelerating the translation of nanotechnology-based cancer therapeutics, diagnostics, and imaging agents to the clinic. In this paper, we will discuss challenges associated with developing a nanomaterial database and recognized needs for nanotechnology data curation and sharing in the biomedical research community. We will also describe the latest version of caNanoLab, caNanoLab 2.0, which includes enhancements and new features to improve usability such as personalized views of data and enhanced search and navigation.

Introduction

The U.S. annual report to the nation on the state of cancer indicates a steady decline in overall mortality rates, with increases in incidence for many cancers [1]. Internationally, cancer incidence paints a more dramatic picture in which the number of new cases has increased from 12.7 million in 2008 to 14.1 million in 2012, with this number expected to rise even

further by an additional 75% in the next two decades [2]. Regardless of whether the focus is limited to the U.S. or considered internationally, the implied and actual burden of cancer is clear, calling for earlier detection and treatment modalities to alleviate this problem. Standard cancer therapeutics are often characterized by poor water solubility and rapid degradation leading to narrow therapeutic windows and doses limited by toxicity [3]. In turn, diagnostics are often hindered at the level of sensitivity, and time between testing and diagnosis. Opportunities for the potential to improve current cancer therapeutics and diagnostics are sorely needed. Nanotechnology provides tremendous opportunities in applications to medicine to make improvements in both these areas. At the nanoscale, the properties of materials yield unique chemical, physical, and biological features that make them advantageous drug delivery vehicles and imaging agents that can target tumor cells, while sparing healthy cells – thereby drastically reducing the toxicity of treatments [4]. Even more so, nanotechnology can be utilized to deliver newer drugs that in the absence of nanotechnology-based vehicle are undeliverable at effective doses [5].

Yet, major hurdles remain to be overcome before we can expect to see regular use of nanotechnology in the clinic that are inherent to new technologies at the clinical trial stage, such as the cost of development, and biological challenges that need to be addressed to ensure patient safety and efficacy. There are only five U.S. Food and Drug administration approved nanotechnology-based drugs – Doxil, DaunoXome, DepoCyt, Marqibo, and Abraxane – while many more are in clinical trials [6]. Similarly, there are a limited number of approved diagnostic devices and tests [7]. In other areas of research, especially genomics, the sharing of experimental data has been shown to be vital for the advancement of scientific discovery and translation [8,9]. Databases such as dbGaP have provided investigators access to hundreds of genomics studies, resulting in three times that number of publications and scientific advances in the genetic basis of disease [8]. Unlike genomics, nanotechnology data management systems, which are at relatively early stages of development, must consider the heterogeneity of nanomaterial data and varied needs based on application (e.g., research focus – environmental vs medical vs energy). Even within a given research area, multi-disciplinary contributions to the field further complicate the development of management systems that address the needs of different communities.

The task of creating relevant databases for nanotechnology risk assessment, manufacturing, characterizations, and literature data is being taken on globally by government, academic, and regulatory organizations. To date, there are approximately 38 databases at various stages of development from initial schema integration to storage of structured, accessible data [10]. However, obstacles still exist in accessing well-characterized datasets and computational tools for further analyses, validation, and guidance in the design optimization of nanomaterials. Further, the development and adoption of data standards to enable efficient data deposition into databases and sharing between laboratories and individual investigators is of great importance. Building the infrastructure for organized data management systems is seen as a potential avenue to overcome these challenges to technology development and clinical translation.

Here we discuss considerations for developing a user-friendly nanomaterial repository in biomedicine and sharing well annotated nanotechnology data. In particular, we describe the cancer Nanotechnology Laboratory (caNanoLab) data portal, a web-based database that allows users to submit and retrieve information on highly described nanomaterials used in biomedicine. We provide an overview of caNanoLab functionality and the release of caNanoLab 2.0, which contains new features and enhancements that address some of the barriers to data sharing described above and enable more efficient data submission and greater support for users.

Results and Discussion

caNanoLab 2.0 navigation, search, and submission

As we have previously reported, the caNanoLab project (https://cananolab.nci.nih.gov/) was initiated as a collaborative effort between the National Cancer Institute's (NCI) Office of Cancer Nanotechnology Research and Center for Biomedical Informatics and Information Technology to address the characterization requirements for federal regulatory review of nanomaterial-based investigational new drugs, diagnostic devices, and imaging agents [11,12]. caNanoLab was originally designed to capture information about the nanomaterial sample and its composition, associated in vitro characterizations, experimental protocols, and relevant publications. The ultimate goal being to accelerate the clinical use of cancer nanomedicines by providing efficacy and safety information to support the above mentioned review process for the use of these nanomaterial in human cancer clinical trials, one of the first step to clinical use. Moreover, caNanoLab was designed to enable the sharing of highly described and complete nanomaterial datasets that can then be re-used for downstream analyses and nanomaterial optimization. In the past decade since its launch, caNanoLab has been expanded to further address the needs of the biomedical research community by enabling the submission and retrieval of diverse nanomaterial types (Figure 1) and characterizations, including in vivo and ex vivo characterizations, to additionally support computational modeling and simulation of

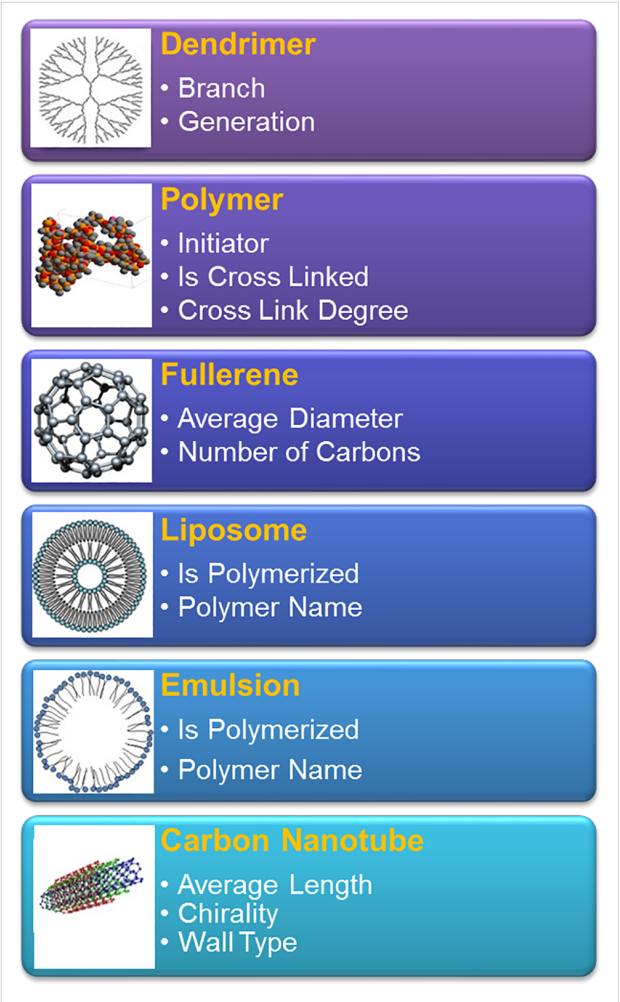


Figure 1: Example nanomaterial types supported by caNanoLab. Pictured is a subset of supported nanomaterials, along with a list of example high-level metadata specific for the listed particle types.

nanoparticle behavior. Standardized metadata are provided to aid these efforts.

caNanoLab navigation and search features

In support of data sharing, caNanoLab compliments other nanomaterial data resources [11] and provides facilities that enable the retrieval and submission of standardized nanomaterial data. Currently, more than 1,000 curated nanomaterial records are publicly accessible and can be queried directly from the caNanoLab homepage. Web usage statistics indicate the majority of users are from the U.S., but has grown to include users from several other countries such as Great Britain, Germany, China, the Netherlands, Spain, and Japan. In 2014, the number of unique portal visitors numbered over 3,000. Options for browsing curated protocols, samples, and publications are available on the homepage. In the caNanoLab 2.0 release, the homepage layout and interface were changed to improve navigation, including enhancements to the User Actions options, and access to commonly asked questions and answers. By selecting "Search Samples," users are taken to a screen from which nanomaterial samples can be queried by keyword, name, or nanomaterial feature. Each sample provides information on the nanomaterial developer, which is also provided as a search option (Sample Point of Contact), and listed in detail in the subsequent Sample Search Results screen (Figure 2).

By selecting "View" next to the sample of interest, users can analyze information about individual nanomaterial sample records such as composition, which includes standard metadata used to describe composition properties (Figure 3). Importantly, the "Navigation Tree" allows for viewing of other pertinent features of the selected nanomaterial such as general information about the developer (e.g., organization and role) and performed characterizations. Similarly, recommended metadata are provided for various characterization assay information such as assay type, experimental techniques, protocols, instruments, and experimental conditions to ultimately support comparison between nanomaterial studies (Figure 4). These metadata were derived from review of nanomaterial properties provided by NCI's Nanotechnology Characterization Laboratory (http://ncl.cancer.gov/), collaborations with the NanoParticle Ontology (NPO; http://www.nano-ontology.org/), and discussions with the research community.

In addition to sample searches, caNanoLab users can search for protocol and publication information by name or nanomaterial feature from the caNanoLab homepage or by using tabs at the top of a viewed nanomaterial sample record (Figure 3). Query results can be either printed or exported into spread-sheet based reports using options available on the results screen. In caNanoLab 2.0, a search for sample characterization and composition information using the associated publication's identifier has been implemented and returns a compiled sample information page (Figure 5). Users can search by either Digital Object Identifier (DOI) or PubMed ID. This feature is also available for publication vendors to interface online articles with corresponding caNanoLab data by leveraging the publication's DOI. By creating this interface, we hope to promote the discoverability and usage of data in caNanoLab.

caNanoLab submission

To submit information into caNanoLab, data submitters are guided through the process with the help of a workflow diagram containing active links (Figure 6) that directs users to web-based forms. Users request an account on the homepage and once credentials are provided, may login to submit protocols, samples, and publications. All data submissions are reviewed for completeness by an in-house curator, and require approval

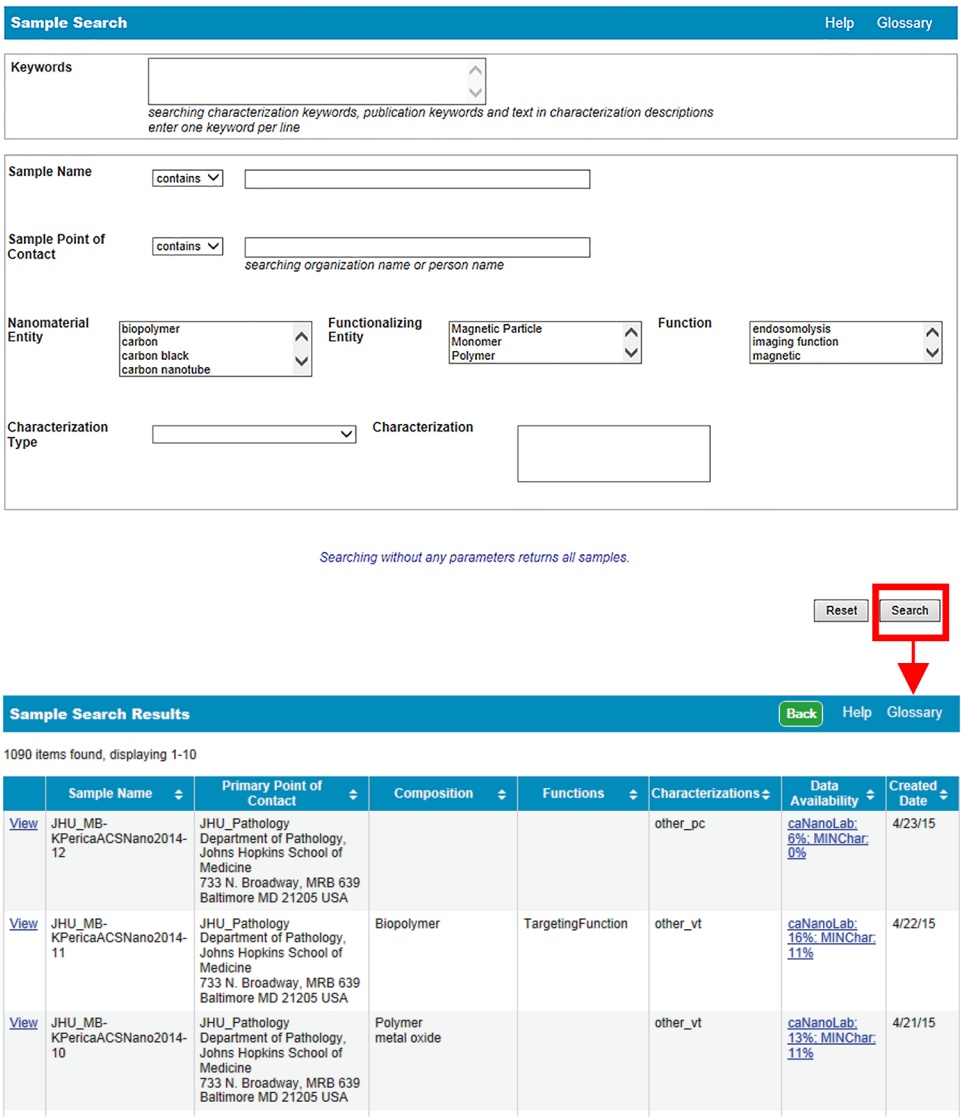


Figure 2: Sample search. Users can search for samples by keyword, name, point of contact, or feature. Following a search (red highlighted box and arrow), users are taken to a sample search results screen from which users can review the results and select sample records to view.

before being made publicly available on the caNanoLab website. To improve this process, caNanoLab 2.0 introduces a MyWorkspace feature as illustrated in Figure 7 to allow submitters to view and access their submitted data, and monitor submission status.

Nanotechnology protocols (Figure 8) for characterization, safety, radiolabeling, sample preparation, and other detailed procedures that might be part of an experiment can be entered into the portal. Protocols currently available are primarily for physico-chemical and in vitro characterizations, however, other protocol assays are strongly encouraged and welcomed, including video-recorded procedures. Submitters can specify protocol type from a drop-down list (e.g., in vitro assay, sample preparation, other) and protocol version if multiple variations or updates exist. Protocols can be submitted as files or URLs to videos or other protocol documents maintained externally. Once submitted, protocols can then be associated with characterization assays described for submitted samples.

In addition to protocols, caNanoLab supports the submission of sample composition and characterizations. For the purposes of caNanoLab, a sample is defined as a formulation of a base nanomaterial platform and any additional components that contribute to the function(s) of the nanomaterial. Submitters can enter nanomaterial composition information (Figure 9)

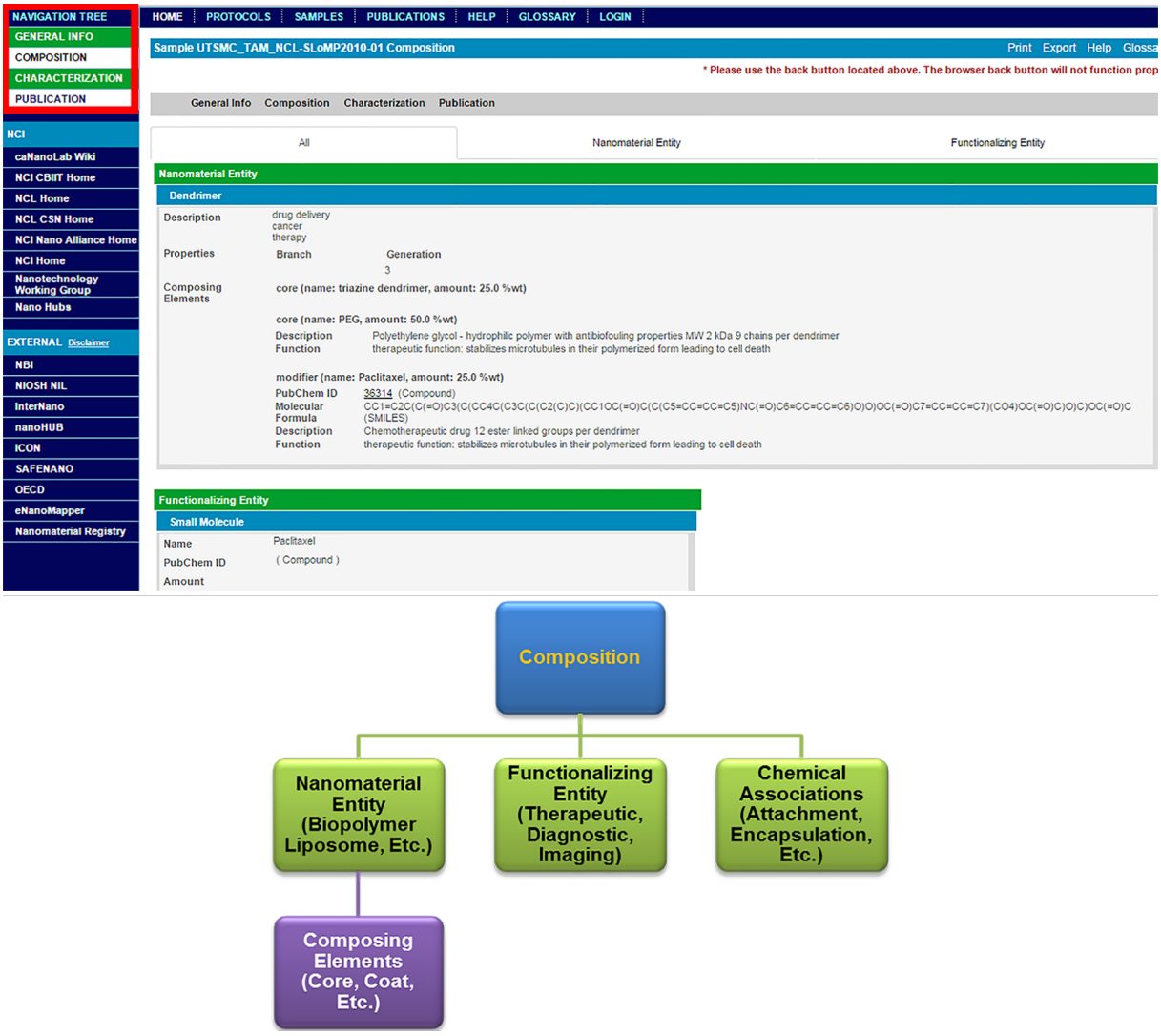


Figure 3: Example nanoparticle composition in caNanoLab of a triazine dendrimer with paclitaxel. Composition information captures properties inherent to the dendrimer (e.g., generation), as well as properties inherent to several particle types (e.g., chemical name, molecular formula). Bottom diagram highlights high level concepts and properties pertaining to composition.

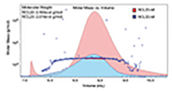
including: nanomaterial entities (e.g., dendrimer), functionalizing entities (e.g., small molecule), and chemical associations (e.g., covalent bond). This composition model supports the submission of complex particles (e.g., liposome encapsulated in a quantum dot) and supports the capture of properties unique to each particle type. Nanomaterial characterizations include physico-chemical, in vitro, and in vivo characterizations. When submitting characterizations, submitters can specify the protocol, instruments, and techniques used in the described characterization assay (Figure 10). Research findings information, including empirical data and experimental conditions, may also be uploaded as files and/or in a data matrix (Figure 11). Once a sample is successfully submitted to the database, either the submitter or curator can generate a data availability metrics table for the sample (Figure 12). Such a data availability metrics compares the submitted data to a checklist of data supported by caNanoLab and data recommended in the MinChar standard (https://characterizationmatters.wordpress.com/parameters/). The caNanoLab identified metadata illustrates information pertinent for nanomaterial composition and specific characterizations, while MinChar is suggested minimum metadata proposed by researchers and others involved in assessing nanomaterial safety to enable cross-comparison of nanomaterial data and data interpretation. Access to this table is available following a sample search on the sample search results screen (Figure 2).

Assay Type	molecular weight		
Point of Contact	DNT		
Characterization Date	N/A		
Protocol	N/A		
Design Description	N/A		
Experiment Configurations	**Technique**	**Instruments**	**Description**
	asymmetrical flow field-flow fractionation with multi-angle laser light scattering(AFFF-MALLS)		

Characterization Results — **Data and Conditions**

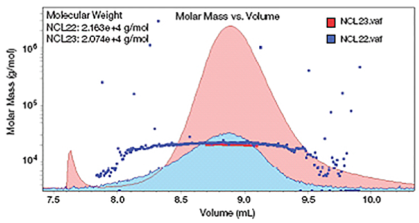
sample concentration (observed,mg/mL)	molecular weight (observed,kDa)	solvent media (observed)	PDI (observed)
2	20.74	PBS	1.078

Files



Molar mass versus elution time plot of NCL22 and NCL23 by AFFF-MALLS. Concentration of NCL22: 1 mg/mL in H O; concentration of NCL23: 2 mg/mL in PBS; Conditions: Injection volume: 100 uL; 10kDa regenerated cellulose membrane; 350 um channel thickness; 1 mL/min channel flow; 3 mL/min cross-flow. AFFF is an innovative separation method for an efficient separation and characterization of nanoparticles, polymers, and proteins that is both fast and gentle. When coupled with a MALLS system, the molar mass and rms radius can be obtained for the fractionated sample. The molar mass distribution plot shows that NCL22 and NCL23 have similar molar mass by using AFFF as separation method. The calculated molar mass of NCL22 and NCL23 was 21.63 kDa and 20.74 kDa, and the polydipersity index was 1.046 and 1.078, respectively (the molar mass of both NCL22 and NCL23 was determined by using the dn/dc value of NCL22, which was measured using an RI detector).

Analysis and Conclusion — NCL23 and NCL22 have a similar molar mass by using AFFF as a separation method. NCL23 is NCL22 with associated Magnevist.



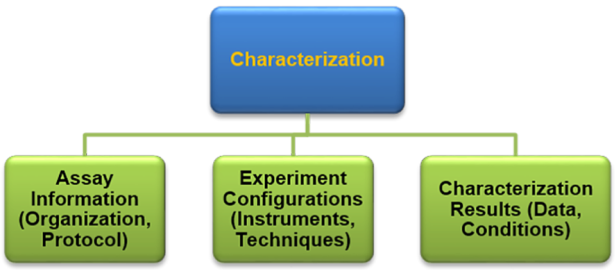


Figure 27. Molar mass versus elution time plot of NCL22 and NCL23 by AFFF-MALLS. Concentration of NCL22: 1 mg/mL in H_2O; concentration of NCL23: 2 mg/mL in PBS; Conditions: Injection volume: 100 µL; 10kDa regenerated cellulose membrane; 350 µm channel thickness; 1 mL/min channel flow; 3 mL/min cross-flow. AFFF is an innovative separation method for an efficient separation and characterization of nanoparticles, polymers and proteins that is both fast and gentle. When coupled with a MALLS system, the molar mass and rms radius can be obtained for the fractionated sample. The molar mass distribution plot shows that NCL22 and NCL23 have similar molar mass by using AFFF as a separation method. The calculated molar mass of NCL22 and NCL23 was 21.63 kDa and 20.74 kDa, and the polydispersity index was 1.046 and 1.078, respectively (the molar mass of both NCL22 and NCL23 was determined by using the dn/dc value of NCL22, which was measured using an RI detector).

Figure 4: Example Nanoparticle Characterization in caNanoLab of a Dendrimer. Characterization information captures information about the assay type and experimental conditions (e.g., technique, concentrations, and observed measurements). Pictured here is an example in which the molecular weights of two curated nanomaterials are compared by light scattering (top and bottom left). Bottom right diagram highlights parameters and factors specific to characterization assays.

caNanoLab also supports the submission of publications (Figure 13) and other reports. Through integration with PubMed, information about publications can be populated into caNanoLab simply by providing the PubMed ID. Previously submitted samples can be associated with a publication during the publication submission process (if samples were described in a published work), enabling the simultaneous retrieval of publication and sample information following a query.

Data submitters are allowed to make their data public or private, with the option to grant access to a limited number of users for varied levels of sharing. Submission instructions are provided in caNanoLab's online user manual, as well as through a video

Sample Information by Publication Back Help Glossary

Publication REF	Authors	Title	Sample Composition	Sample Characterization	Journal	Year	Vol(Iss)Pg
DOI Id: 10.1111/j.1751-1097.2007.00163.x	Rancan, F, Helmreich, M, Mölich, A, Jux, N, Hirsch, A, Röder , B, Böhm, F	Intracellular uptake and phototoxicity of 3(1),3(2)-didehydrophytochlorin-fullerene hexaadducts	Samples curated in caNanoLab: UC_HU_UEN-FRancanPhPh2007-01 UC_HU_UEN-FRancanPhPh2007-02 UC_HU_UEN-FRancanPhPh2007-03 UC_HU_UEN-FRancanPhPh2007-04	Samples curated in caNanoLab: UC_HU_UEN-FRancanPhPh2007-01 UC_HU_UEN-FRancanPhPh2007-02 UC_HU_UEN-FRancanPhPh2007-03 UC_HU_UEN-FRancanPhPh2007-04	Photochemistry and Photobiology	2007	83:1330-1338

The main challenge in searching for new photosensitizers is to improve their specificity for target cells to avoid toxicity towards normal cells. New modular drug delivery systems were proposed consisting of a multiplying unit with the property of carrying several drug moieties and an addressing unity with high selectivity for target cells. Following this concept, two new fullerene-bis-pyropheophorbide a derivatives were synthesized: a mono-(FP1) and a hexa-adduct (FHP1). The photophysical characterization of the compounds revealed significantly different parameters related to the number of addends at the fullerene core. In this study, the derivatives were tested with regard to their intracellular uptake and photosensitizing activity towards human leukemia T-lymphocytes (Jurkat cells) in comparison with the free sensitizer, pyropheophorbide a. The C(60)-hexa-adduct FHP1 resulted to have a significative phototoxic activity (58\% dead cell, after a dose of 400 mJ/cm(2), 688 nm) while the mono-adduct FP1 had a very low phototoxicity and only at higher light doses. The photosensitizing activity of the fullerene hexa-adduct, FHP1, resulted to be lower than that of pyropheophorbide a. The lesser intracellular concentration reached by the C(60)-hexa-adduct FHP1 is probably the reason for its lower phototoxicity with respect to pyropheophorbide a.

Figure 5: caNanoLab sample information by publication. List and active links to all curated data for a given publication are provided following a DOI-based search for samples by publication. This option is available under the Publication tab once the user has initiated a publication search.

tutorial that guides users through the caNanoLab 2.0 submission procedures. Both resources can be found on the caNanoLab FAQ webpage (https://wiki.nci.nih.gov/x/UKml), accessible through the caNanoLab homepage under the “How To” box. Assistance is also provided by the in-house curator.

Data integration and sharing

To optimize the design and utility of nanomaterials in biomedicine, researchers need to integrate and compare datasets generated by different research groups. However, the lack of availability and access to datasets stored across a variety of resources

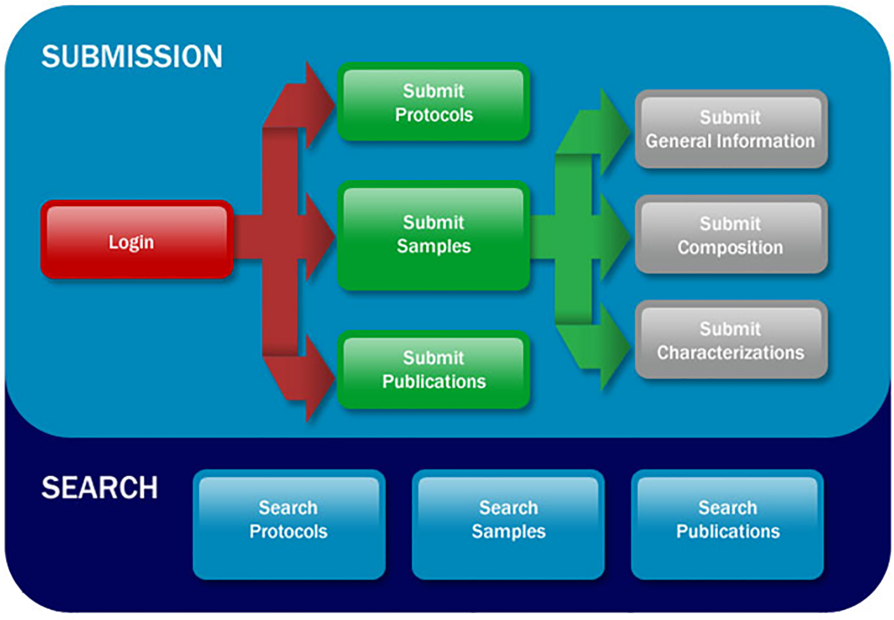
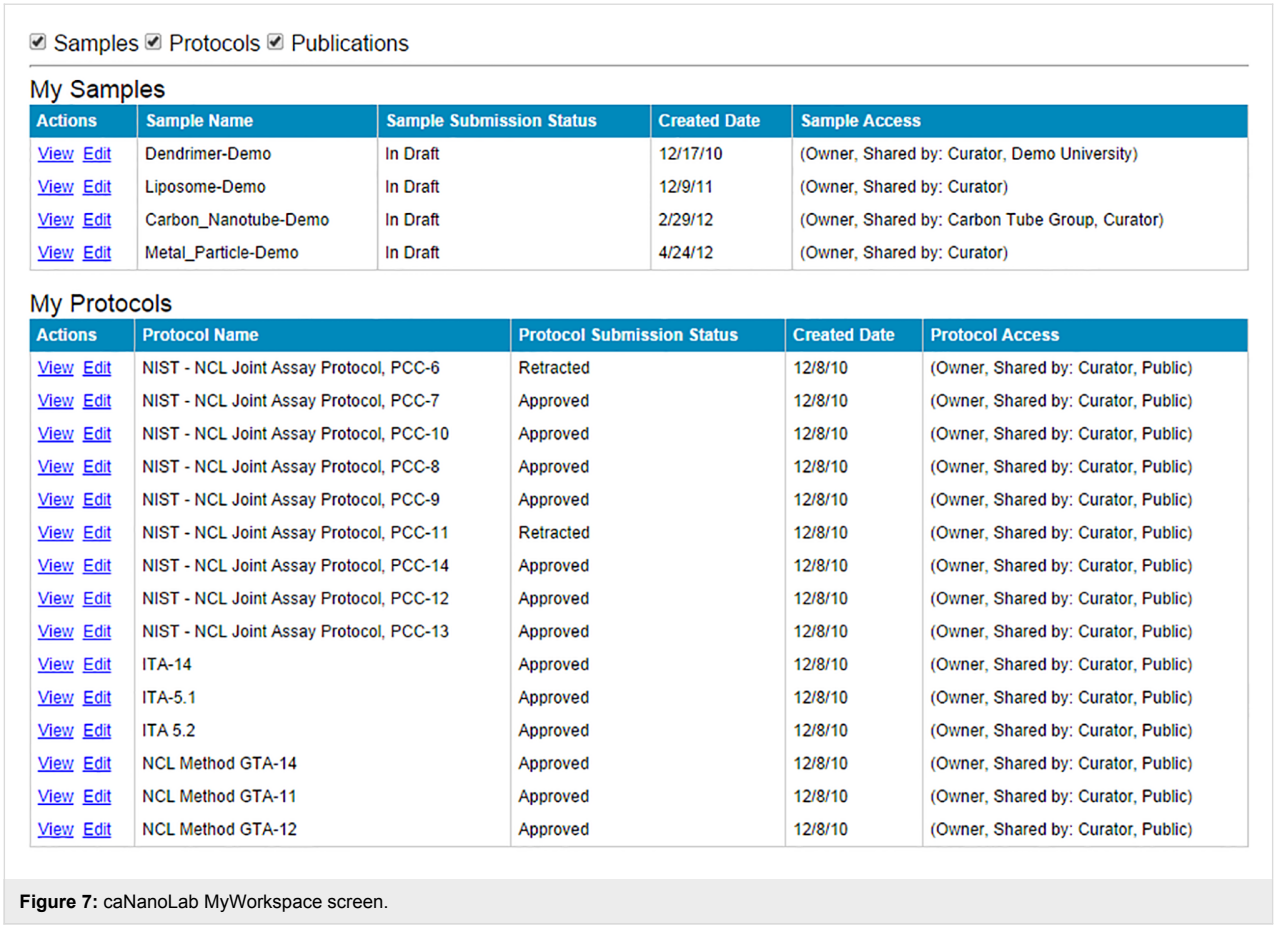


Figure 6: caNanoLab data submission and search workflow. A graphic available upon login that illustrates the functionality in caNanoLab. Both workflows provide active links for the indicated options. Reprinted with permission from [12]. Copyright 2014 IEEE.



☑ Samples ☑ Protocols ☑ Publications

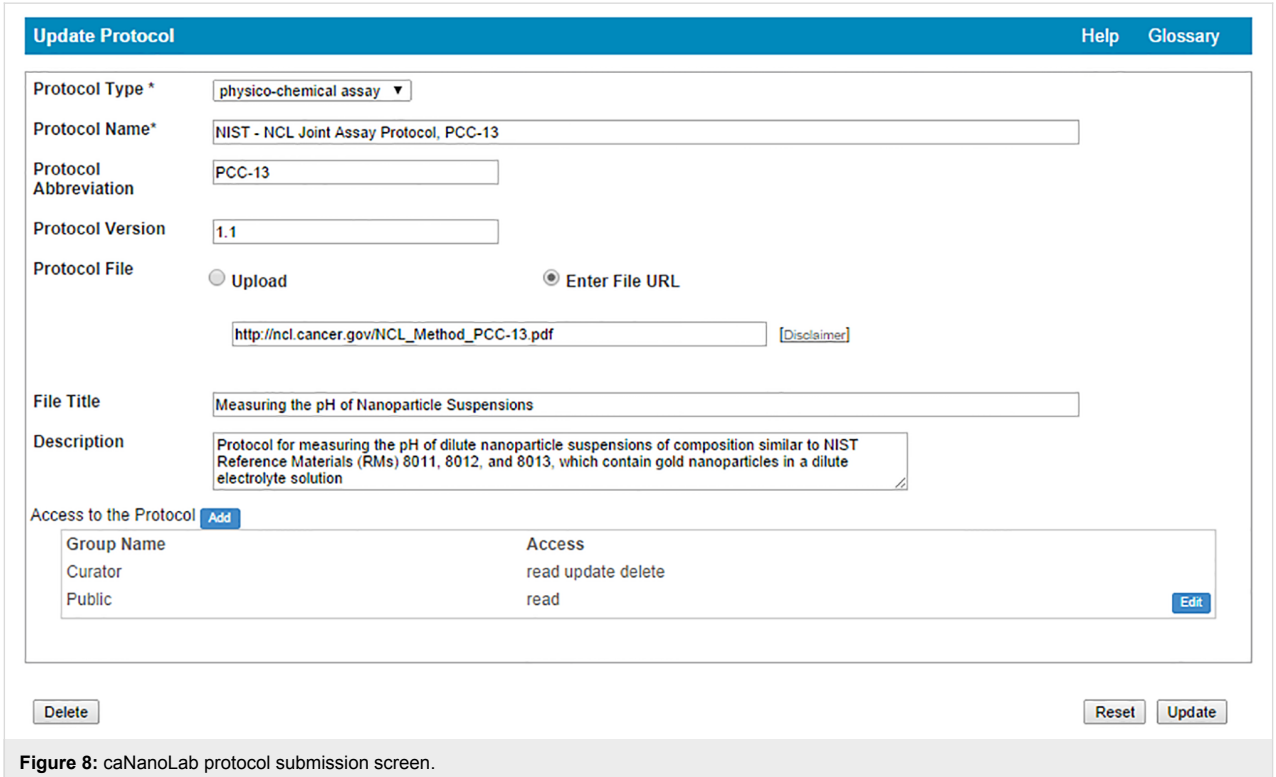
My Samples

Actions	Sample Name	Sample Submission Status	Created Date	Sample Access
View Edit	Dendrimer-Demo	In Draft	12/17/10	(Owner, Shared by: Curator, Demo University)
View Edit	Liposome-Demo	In Draft	12/9/11	(Owner, Shared by: Curator)
View Edit	Carbon_Nanotube-Demo	In Draft	2/29/12	(Owner, Shared by: Carbon Tube Group, Curator)
View Edit	Metal_Particle-Demo	In Draft	4/24/12	(Owner, Shared by: Curator)

My Protocols

Actions	Protocol Name	Protocol Submission Status	Created Date	Protocol Access
View Edit	NIST - NCL Joint Assay Protocol, PCC-6	Retracted	12/8/10	(Owner, Shared by: Curator, Public)
View Edit	NIST - NCL Joint Assay Protocol, PCC-7	Approved	12/8/10	(Owner, Shared by: Curator, Public)
View Edit	NIST - NCL Joint Assay Protocol, PCC-10	Approved	12/8/10	(Owner, Shared by: Curator, Public)
View Edit	NIST - NCL Joint Assay Protocol, PCC-8	Approved	12/8/10	(Owner, Shared by: Curator, Public)
View Edit	NIST - NCL Joint Assay Protocol, PCC-9	Approved	12/8/10	(Owner, Shared by: Curator, Public)
View Edit	NIST - NCL Joint Assay Protocol, PCC-11	Retracted	12/8/10	(Owner, Shared by: Curator, Public)
View Edit	NIST - NCL Joint Assay Protocol, PCC-14	Approved	12/8/10	(Owner, Shared by: Curator, Public)
View Edit	NIST - NCL Joint Assay Protocol, PCC-12	Approved	12/8/10	(Owner, Shared by: Curator, Public)
View Edit	NIST - NCL Joint Assay Protocol, PCC-13	Approved	12/8/10	(Owner, Shared by: Curator, Public)
View Edit	ITA-14	Approved	12/8/10	(Owner, Shared by: Curator, Public)
View Edit	ITA-5.1	Approved	12/8/10	(Owner, Shared by: Curator, Public)
View Edit	ITA 5.2	Approved	12/8/10	(Owner, Shared by: Curator, Public)
View Edit	NCL Method GTA-14	Approved	12/8/10	(Owner, Shared by: Curator, Public)
View Edit	NCL Method GTA-11	Approved	12/8/10	(Owner, Shared by: Curator, Public)
View Edit	NCL Method GTA-12	Approved	12/8/10	(Owner, Shared by: Curator, Public)

Figure 7: caNanoLab MyWorkspace screen.



Update Protocol — Help Glossary

Protocol Type * physico-chemical assay

Protocol Name* NIST - NCL Joint Assay Protocol, PCC-13

Protocol Abbreviation PCC-13

Protocol Version 1.1

Protocol File ○ Upload ◉ Enter File URL

http://ncl.cancer.gov/NCL_Method_PCC-13.pdf [Disclaimer]

File Title Measuring the pH of Nanoparticle Suspensions

Description Protocol for measuring the pH of dilute nanoparticle suspensions of composition similar to NIST Reference Materials (RMs) 8011, 8012, and 8013, which contain gold nanoparticles in a dilute electrolyte solution

Access to the Protocol Add

Group Name	Access	
Curator	read update delete	
Public	read	Edit

Delete Reset Update

Figure 8: caNanoLab protocol submission screen.

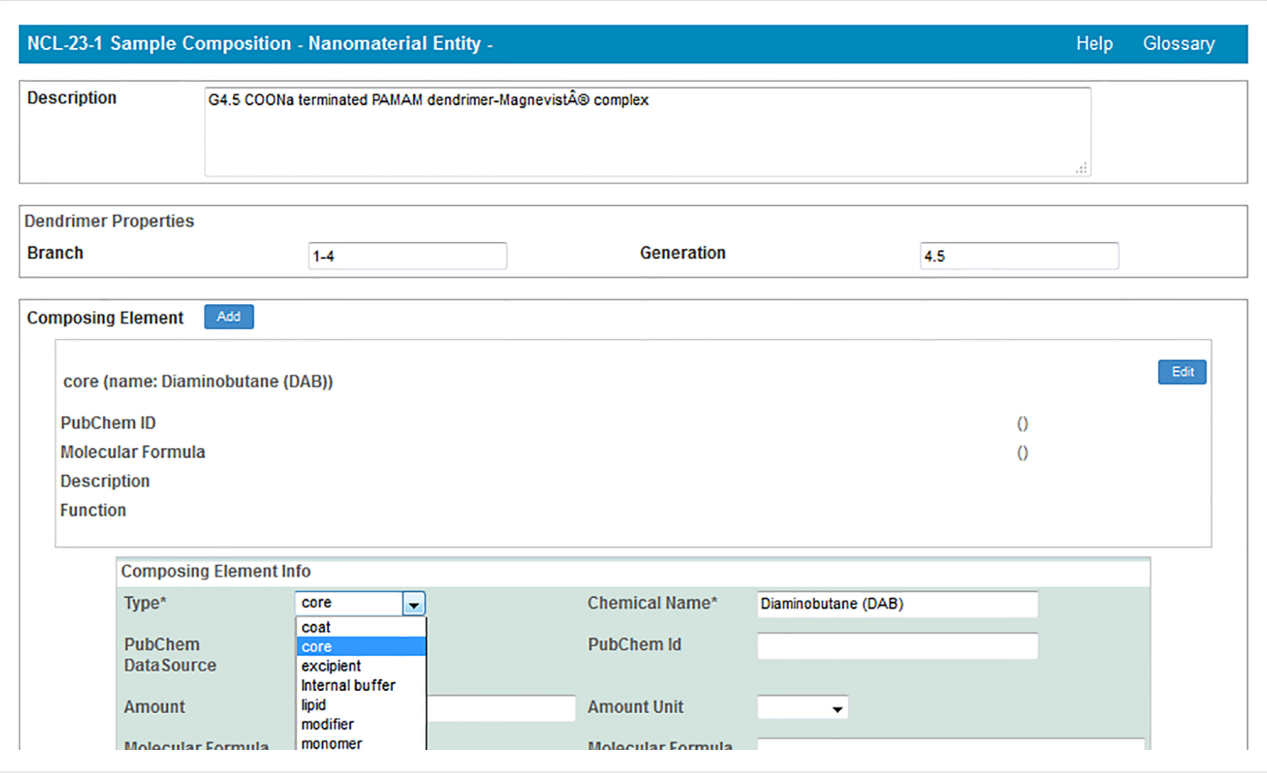


Figure 9: caNanoLab sample composition submission screen.

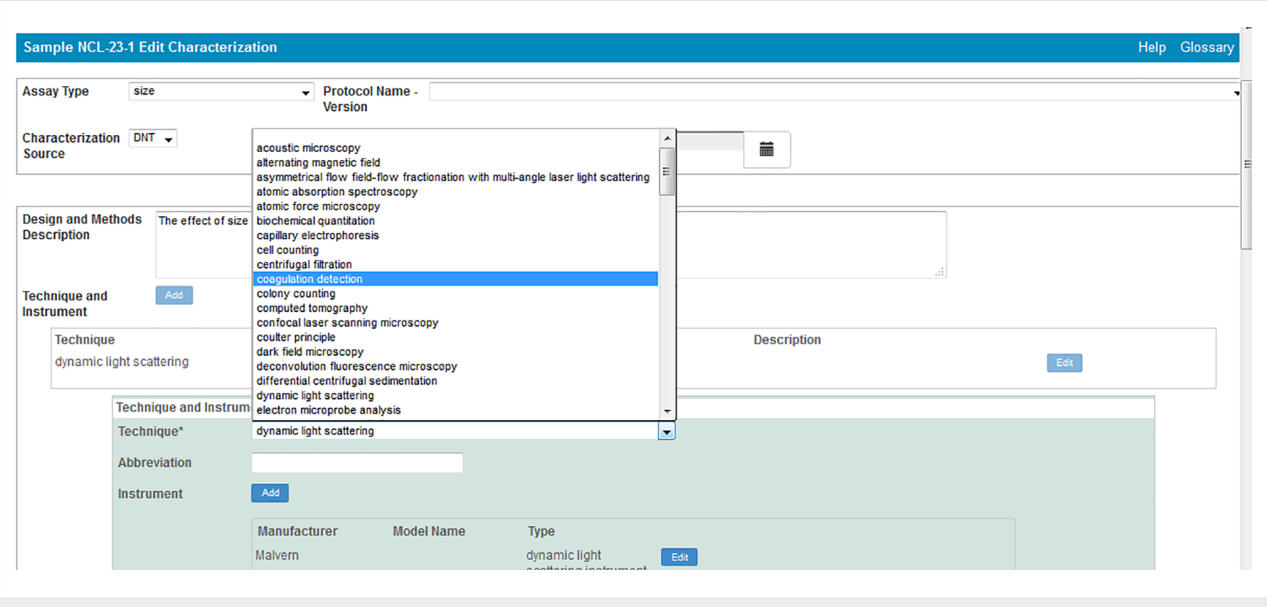


Figure 10: caNanoLab sample characterization submission screen – techniques and instruments.

with limited data exchange hinders this goal. The caNanoLab team strongly supports interoperability between databases, and engages in activities focused on the development of standards to enable data exchange. In particular, the design of the caNanoLab data model was informed by the NPO, which represents knowledge underlying the description, preparation, and characterization of nanomaterials in cancer nanotechnology research [13]. caNanoLab data model class names and attributes are maintained in the NCI cancer Data Standards Repository (https://cdebrowser.nci.nih.gov/CDEBrowser/), and definitions for caNanoLab concepts are maintained in the NCI Thesaurus (http://ncit.nci.nih.gov/). The caNanoLab team is

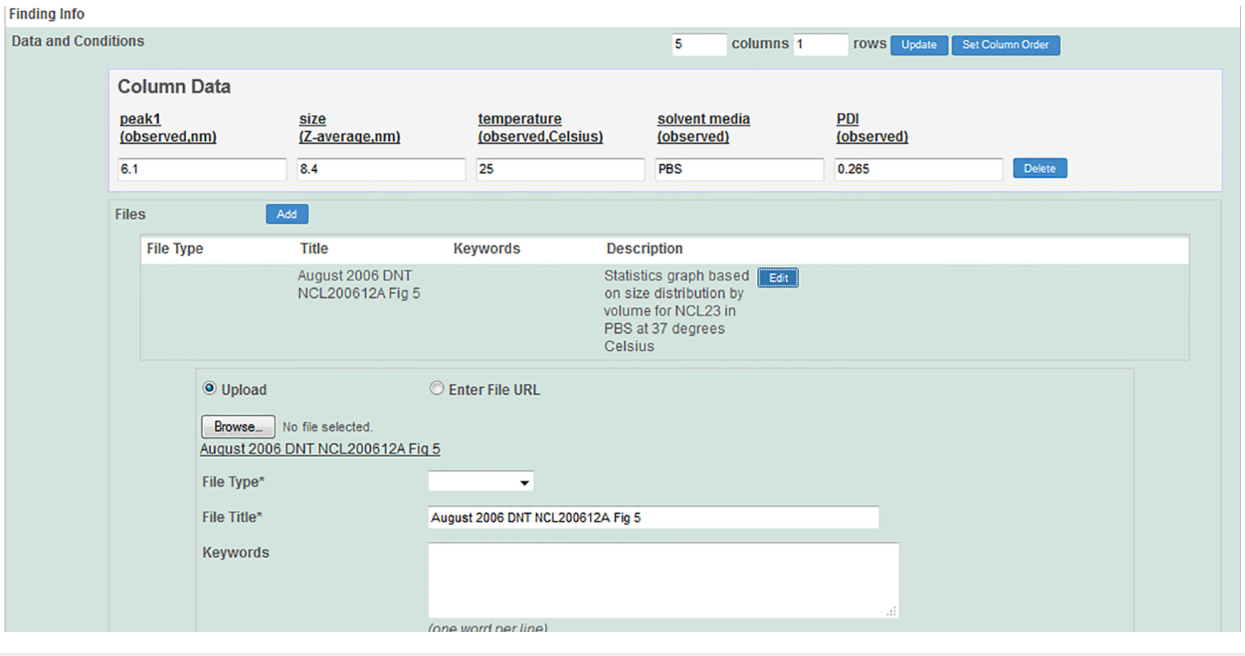


Figure 11: caNanoLab sample characterization submission screen – data and conditions.

also working with the ISA-TAB (http://isatab.sourceforge.net/) and nanotechnology communities to develop a specification that provides descriptive information applicable to nanotechnology using spreadsheet-based file formats – ISA-TAB-Nano [14]. Curated caNanoLab data are annotated by terms from Bioportal (http://bioportal.bioontology.org) and entered into ISA-TAB-Nano files that are available for download at https://wiki.nci.nih.gov/x/lgFwBg by individual users or other databases to enable data exchange.

In addition to the development and utilization of data exchange standards, another challenge to data sharing, as viewed by caNanoLab, has been access to investigator-derived data, and submission of these data by individual investigators. The majority of data submitted into caNanoLab are curated from published articles. The most challenging aspect of this process is acquiring additional information from the author. To address this challenge, many of the features in caNanoLab 2.0 to enhance navigation and enable personalized views of data were designed to improve individual investigator/user data submission. Further, the NCI Alliance for Nanotechnology in Cancer program (http://nano.cancer.gov), a network of extramural research centers and projects also supported by NCI's Office of Cancer Nanotechnology Research, now requires awardees to share data through appropriate publicly accessible databases such as caNanoLab, and has made nanomaterial data deposition a Term and Condition of award (see RFA-CA-14-013 (http://grants.nih.gov/grants/guide/rfa-files/RFA-CA-14-013.html); PAR-14-25 (http://grants.nih.gov/grants/guide/pa-files/PAR-14-285.html)). A nanomaterial data sharing coordinator must be named for each award and plans for data sharing must be included with each application submission. Information on how to incorporate the use of caNanoLab into a data sharing plan is available on the caNanoLab website to make this process easier. Although this is not yet a requirement for other nanomaterial-related funding opportunity announcements, NCI's Office of Cancer Nanotechnology Research hopes this will encourage data sharing and acceptance of nanomaterial data deposition as a standard practice similar to what has been observed for genomics data and currently instituted federal data sharing policies [8,15].

Addressing future needs of biomedical databases supporting nanotechnology

The genomics community expressed the need for standards and databases to house the extensive amount of data generated by gene expression and sequencing experiments, yielding such efforts as the development of the minimum information about a microarray experiment (MIAME) [16]. As a result, the MIAME guideline, and others, have been adopted by journals, databases, and researchers as an accepted format for annotating data – a requirement called for by these groups [17]. Similarly, in order for the nanoinformatics field to grow, the relevance of nanotechnology data and associated information must be emphasized by the community. In discussions amongst community members, primarily in consultation with journals, researchers acknowledged and agreed with the importance of implementing minimum characterization requirements and guidelines, but the

caNanoLab Availability Score: 36.0% (11 out of 30) MINChar Availability Score: 44.0% (4 out of 9)

caNanoLab	MINChar	Caltech-HHanBC2013-10
	agglomeration and/or aggregation	
	crystal structure/crystallinity	
General Sample Information		✓
Sample Composition	chemical composition	✓
nanomaterial entities		✓
functionalizing entities		✓
chemical associations		
attachment	surface chemistry	✓
encapsulation		
entrapment		
sample function		✓
Physico-Chemical Characterization		
surface		
surface area	surface area	
surface charge	surface charge	
zeta potential	surface charge	✓
molecular weight		
physical state		
purity	purity	
relaxivity		
shape	shape	
size	particle size/size distribution	✓
solubility		
In Vitro Characterization		
blood contact		
cytotoxicity		✓
enzyme induction		
immune cell function		
metabolic stability		
oxidative stress		
sterility		
targeting		✓
transfection		
In Vivo Characterization		
pharmacokinetics		
toxicology		
Publications		✓

Close

Figure 12: caNanoLab data availability metrics table. The first and middle columns list data supported and recommended by caNanoLab and the MinChar standard, respectively. The last column is a comparison of the data curated for the indicated sample to the caNanoLab and MinChar column lists. Data availability is provided for samples in Sample Search Results.

manner in which to identify these features were debated [18]. Different types of information are needed based on the purpose of the study, which may vary based on the nanotechnology application [19]. Considering these issues, caNanoLab and other nanomaterial databases require input and support from users including informatics experts, nanotechnologists, biologists, and clinicians to better understand their needs. Active outreach and collaborations are required to meet these goals, as well as sustained interest in the use of databases by the community, and increased data exchange between resources and researchers.

Enhancing data interoperability by collaborative development of data standards and best practices

The caNanoLab team is engaged in many activities to better serve the needs of the nanotechnology research community and increase adoption of caNanoLab and other nanomaterial resources. Activities range from engaging publication vendors to facilitate linkages between publications and nanotechnology databases (as described above), to working with other groups to develop data standards and guidelines for data submission and sharing. In particular, interoperability with other databases is seen as important both for NCI and the caNanoLab user community. To achieve this goal, the caNanoLab team actively works with other databases, community-based programs, and federal initiatives such as the National Cancer Informatics Program (NCIP) Nanotechnology Working Group (Nano WG) and the National Nanotechnology Initiative (NNI; http://www.nano.gov), to develop data standards and deposition guidelines. Accelerating the meaningful exchange of information across the nanotechnology community is a priority for the Nano WG. Consisting of researchers from academia, government, and industry, much of the group's focus has been on the collaborative development and dissemination of data standards. Key efforts in this area have included development and enhancement of the NPO and ISA-TAB-Nano. ISA-TAB-Nano is currently used by NCI, the NBI Knowledgebase (http://nbi.oregonstate.edu/), and the EU NanoSafety Cluster (http://www.nanosafetycluster.eu/) to enable interoperability between databases. Most recently, the Nano WG established a subgroup focused on developing guidelines for data curation, and is in the process of writing a series of consensus papers on curation workflows, data completeness and quality, curator responsibilities, metadata, and integration between datasets and databases, as an overview of current curation practices and recommendations (Nanomaterial Data Curation Initiative, https://nciphub.org/groups/nanotechnologydatacurationinterestgroup) [20,21].

In line with the goals of this subgroup, the journal Nature Nanotechnology recently published an editorial to announce their plans to participate in Nature's initiative to improve consistency and reporting of data in life sciences articles [22]. Starting in January 2015, the journal requires the submission of a checklist that ensures authors disclose all the information necessary for others to reproduce their work. This full disclosure includes the deposition of data into comprehensive public

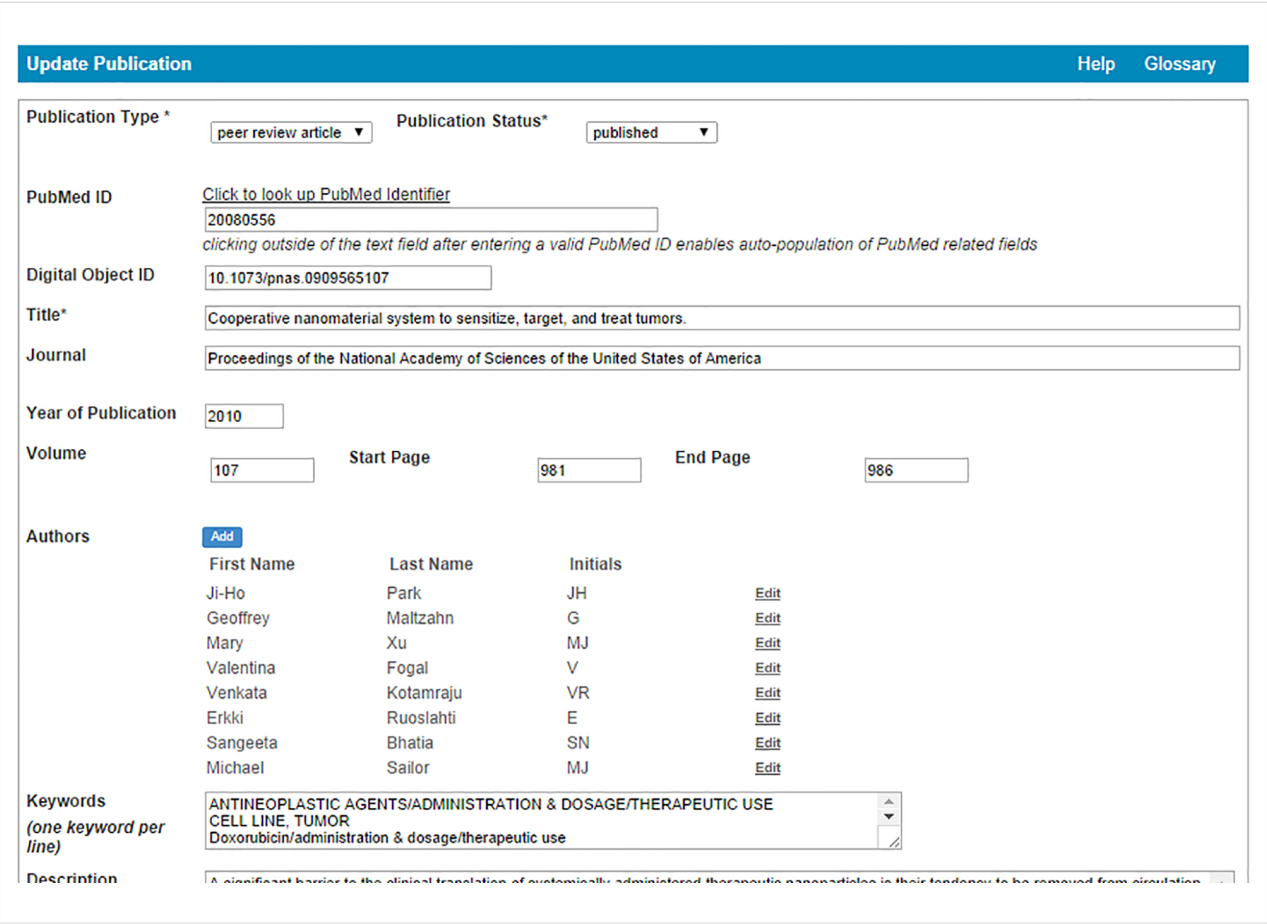


Figure 13: caNanoLab sample publication submission screen. Information for PubMed articles is auto-populated by leveraging PubMed's Application Programming Interface for information retrieval.

databases such as caNanoLab and the Nanomaterial Registry (https://www.nanomaterialregistry.org/). The journal expressed interest in working with communities to develop customized checklists appropriate for specific research fields to streamline data reporting and deposition during the manuscript submission process. As part of this effort, caNanoLab is listed as a recommended data repository for Scientific Data, a Nature journal that publishes descriptions of scientific datasets, and the caNanoLab team participates in the NCIP Nano WG's Nanomaterial Data Curation Initiative. Increased interactions between caNanoLab and journal publishers are also underway to facilitate the development of reporting guidelines in an effort to increase data deposition at the manuscript submission stage [12].

Federal members of the caNanoLab team participate in the NNI Signature Initiative on Nanotechnology Knowledge Infrastructure (NKI) – enabling national leadership in sustainable design [23]. The purpose of the NNI Signature Initiatives is to rapidly advance science and technology by coordinating the programmatic efforts of member federal agencies in areas identified to be of national importance such as nanotechnology data management. The NKI is focused on major thrust areas, including the creation of a data infrastructure to support data sharing, and management to enable novel nanotechnology-based innovations across disciplines. As such, the NKI works with varied groups to accomplish the initiative's goals of ultimately sustaining new innovation and knowledge discovery in the design and application of nanomaterials in science.

Conclusion

Access to detailed nanomaterial characterization data is seen as a prominent need to advance cancer nanomedicines to the clinical environment. To aid this process, caNanoLab will continue to evolve as a valuable resource to the biomedical nanotechnology community through portal enhancements and through integration with other community-identified resources. Plans are underway for a caNanoLab 2.1 release, which will include increased usability and performance enhancements, a Google-like search capability, advanced search and query features, pop-up instructions for data submission fields, and enhancements to the MyWorkspace feature. The caNanoLab 2.1 release will be available in late summer 2015. caNanoLab software is open

source and available for download from GitHub for local installation (https://github.com/NCIP/cananolab). This code is customizable, and code contributions back to the community via GitHub are strongly encouraged to support further development of caNanoLab. As part of the evolution of the portal, the caNanoLab team plans to maintain collaborations with other nanomaterial resources used by the community in support of nanomaterial data standards development, integration, and analysis. The future development of caNanoLab will be guided by community practices supporting data interoperability and exchange, such as the use of ISA-TAB-Nano and community developed common web services.

User Feedback

The caNanoLab team is interested in feedback from the user community on the new caNanoLab features and plans for future enhancements. A discussion forum was created to receive this feedback at https://nciphub.org/groups/cananolab_usability. The team is especially interested in the community's ideas for needed features, as well as data.

Acknowledgements

The caNanoLab project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

References

1. Kohler, B. A.; Sherman, R. L.; Howlader, N.; Jemal, A.; Ryerson, A. B.; Henry, K. A.; Boscoe, F. P.; Cronin, K. A.; Lake, A.; Noone, A.-M.; Henley, S. J.; Eheman, C. R.; Anderson, R. N.; Penberthy, L. *J. Natl. Cancer Inst.* **2015,** *107,* djv048. doi:10.1093/jnci/djv048
2. International Agency for Research on Cancer. *World Cancer Report 2014;* World Health Organization: Geneva, Switzerland, 2014.
3. Gabizon, A.; Bradbury, M.; Prabhakar, U.; Zamboni, W.; Libutti, S.; Grodzinski, P. *Lancet* **2014,** *384,* 2175. doi:10.1016/S0140-6736(14)61457-4
4. Langer, R.; Weissleder, R. *JAMA, J. Am. Med. Assoc.* **2015,** *313,* 135. doi:10.1001/jama.2014.16315
5. Wicki, A.; Witzigmann, D.; Balasubramanian, V.; Huwyler, J. *J. Controlled Release* **2015,** *200,* 138. doi:10.1016/j.jconrel.2014.12.030
6. Etheridge, M. L.; Campbell, S. A.; Erdman, A. G.; Haynes, C. L.; Wolf, S. M.; McCullough, J. *Nanomedicine: NBM* **2013,** *9,* 1. doi:10.1016/j.nano.2012.05.013
7. Neely, L. A.; Audeh, M.; Phung, N. A.; Min, M.; Suchocki, A.; Plourde, D.; Blanco, M.; Demas, V.; Skewis, L. R.; Anagnostou, T.; Coleman, J. J.; Wellman, P.; Mylonakis, E.; Lowery, T. J. *Sci. Transl. Med.* **2013,** *5,* 182RA54. doi:10.1126/scitranslmed.3005377
8. Paltoo, D. N.; Rodriguez, L. L.; Feolo, M.; Gillanders, E.; Ramos, E. M.; Rutter, J. L.; Sherry, S.; Wang, V. O.; Bailey, A.; Baker, R.; Caulder, M.; Harris, E. L.; Langlais, K.; Leeds, H.; Luetkemeier, E.; Paine, T.; Roomian, T.; Tryka, K.; Patterson, A.; Green, E. D.; National Institutes of Health Genomic Data Sharing Governance, C. *Nat. Genet.* **2014,** *46,* 934. doi:10.1038/ng.3062
9. Robinson, P. N. *Genome Med.* **2014,** *6,* 78. doi:10.1186/s13073-014-0078-2
10. Mustad, A. P.; Smeets, B.; Jeliazkova, N.; Jeliazkov, V.; Willighagen, E. L. Summary of the Spring 2014 NSC Database Survey. 2015; (accessed July 6, 2015). doi:10.6084/m9.figshare.1195888
11. Gaheen, S.; Hinkal, G. W.; Morris, S. A.; Lijowski, M.; Heiskanen, M.; Klemm, J. D. *Comput. Sci. Discovery* **2013,** *6,* 014010. doi:10.1088/1749-4699/6/1/014010
12. Morris, S. A.; Gaheen, S.; Lijkowski, M.; Heiskanen, M.; Klemm, J. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* Belfast, Ireland, Nov 2–5, 2014; Zheng, H.; Hu, X.; Berrar, D.; Wang, Y.; Dubitzky, W.; Hao, J.-K.; Cho, K.-H.; Gilbert, D., Eds.; IEEE: Piscataway, NJ, United States of America, 2014; pp 29–33. doi:10.1109/BIBM.2014.6999371
13. Thomas, D. G.; Pappu, R. V.; Baker, N. A. *J. Biomed. Inf.* **2011,** *44,* 59. doi:10.1016/j.jbi.2010.03.001
14. Thomas, D. G.; Gaheen, S.; Harper, S. L.; Fritts, M.; Klaessig, F.; Hahn-Dantona, E.; Paik, D.; Pan, S.; Stafford, G. A.; Freund, E. T.; Klemm, J. D.; Baker, N. A. *BMC Biotechnol.* **2013,** *13,* 2. doi:10.1186/1472-6750-13-2
15. Pham-Kanter, G.; Zinner, D. E.; Campbell, E. G. *PLoS One* **2014,** *9,* e108451. doi:10.1371/journal.pone.0108451
16. Brazma, A.; Hingamp, P.; Quackenbush, J.; Sherlock, G.; Spellman, P.; Stoeckert, C.; Aach, J.; Ansorge, W.; Ball, C. A.; Causton, H. C.; Gaasterland, T.; Glenisson, P.; Holstege, F. C.; Kim, I. F.; Markowitz, V.; Matese, J. C.; Parkinson, H.; Robinson, A.; Sarkans, U.; Schulze-Kremer, S.; Stewart, J.; Taylor, R.; Vilo, J.; Vingron, M. *Nat. Genet.* **2001,** *29,* 365. doi:10.1038/ng1201-365
17. *Nat. Genet.* **2006,** *38,* 1089. doi:10.1038/ng1006-1089
18. *Nat. Nanotechnol.* **2013,** *8,* 69. doi:10.1038/nnano.2013.19
19. Fadeel, B.; Savolainen, K. *Nat. Nanotechnol.* **2013,** *8,* 71. doi:10.1038/nnano.2013.2
20. Powers, C. M.; Mills, K.; Morris, S. A.; Klaessig, F.; Gaheen, S.; Lewinski, N.; Hendren, C. O. *Beilstein J. Nanotechnol.* **2015,** *6,* in press.
21. Marchese Robinson, R. L.; Cronin, M. T. D.; Richarz, A.; Rallo, R. *Beilstein J. Nanotechnol.* **2015,** *6,* in press.
22. *Nat. Nanotechnol.* **2014,** *9,* 949. doi:10.1038/nnano.2014.287
23. NSTC *Nanotechnology Knowledge Infrastructure: Enabling National Leadership in Sustainable Design*, National Nanotechnology Initiative, 2012.

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

How decision analysis can further nanoinformatics

Matthew E. Bates[1], Sabrina Larkin[2], Jeffrey M. Keisler[3] and Igor Linkov[*1]

Commentary

Open Access

Address:
[1]Environmental Laboratory, U.S. Army Engineer Research and Development Center, Concord, MA, USA, [2]Contractor to the Environmental Laboratory, U.S. Army Engineer Research and Development Center, Concord, MA, USA, and [3]Department of Management Science and Information Systems, College of Management, University of Massachusetts Boston, Boston, MA, USA

Email:
Igor Linkov[*] - Igor.Linkov@usace.army.mil

* Corresponding author

Keywords:
decision analysis; nanoinformatics; policy; portfolio analysis; risk assessment; value of information; weight of evidence

Beilstein J. Nanotechnol. **2015,** *6,* 1594–1600.
doi:10.3762/bjnano.6.162

Received: 18 April 2015
Accepted: 09 July 2015
Published: 22 July 2015

This article is part of the Thematic Series "Nanoinformatics for environmental health and biomedicine".

Guest Editor: R. Liu

Abstract

The increase in nanomaterial research has resulted in increased nanomaterial data. The next challenge is to meaningfully integrate and interpret these data for better and more efficient decisions. Due to the complex nature of nanomaterials, rapid changes in technology, and disunified testing and data publishing strategies, information regarding material properties is often illusive, uncertain, and/or of varying quality, which limits the ability of researchers and regulatory agencies to process and use the data. The vision of nanoinformatics is to address this problem by identifying the information necessary to support specific decisions (a top-down approach) and collecting and visualizing these relevant data (a bottom-up approach). Current nanoinformatics efforts, however, have yet to efficiently focus data acquisition efforts on the research most relevant for bridging specific nanomaterial data gaps. Collecting unnecessary data and visualizing irrelevant information are expensive activities that overwhelm decision makers. We propose that the decision analytic techniques of multicriteria decision analysis (MCDA), value of information (VOI), weight of evidence (WOE), and portfolio decision analysis (PDA) can bridge the gap from current data collection and visualization efforts to present information relevant to specific decision needs. Decision analytic and Bayesian models could be a natural extension of mechanistic and statistical models for nanoinformatics practitioners to master in solving complex nanotechnology challenges.

Introduction

Extensive nanomaterial research has yielded an increasing amount of nanomaterial data [1]. The nanomaterial data are currently so vast that it has become difficult to find data relevant to a specific need. However, a formal knowledge infrastructure, inclusive of current nanomaterial data, is essential to future developments in nanomaterial research [2]. Nanoinformatics is defined as (a) "the science and practice of determining which information is relevant to the nanoscale science and engineering community", and (b) "developing and implementing effective mechanisms for collecting, validating,

storing, sharing, analyzing, modeling, and applying that information" [3]. This definition implies the integration of top-down methods for assessing scientific community needs with bottom-up methods for data collection and management [4,5]. Such integration will enhance the reproducibility and distribution of data and the ability to transform the vast nanomaterial data into accessible, integrated information.

Two recent workshops sponsored by the National Nanotechnology Initiative [5] and the National Nanomanufacturing Network [6] were focused on assessing the state of nanomaterial risk management, nanoinformatics, determining gaps in the information and risk management technologies, and evaluating opportunities for improvement. These nanoinformatics workshops highlighted a number of resources that were already using nanoinformatics to aggregate and organize nanomaterial data [6]. The Nanoparticle Information Library (NIL) is a database from the National Institute for Occupational Safety and Health (NIOSH) that aggregates the physical characteristics of nanomaterials for industrial users, researchers, and health professionals to access and share [7]. The NanoHub offers a collaborative workspace for users to share research, identify possible opportunities to work with others, and to learn more about nanotechnology [8]. This includes the GoodNanoGuide, a resource that serves as a best practice exchange for nanomaterials in the workplace [9]. The Nanomaterial Registry archives nanomaterial data according to their properties and environmental and health implications, including their compliance scores [1]. These efforts all focus on developing resources that satisfy the bottom-up part of the nanoinformatics definition presented above. The top-down part, in which the appropriateness of information to a specific need is determined, is not addressed to the same extent in any of the aforementioned efforts. A few existing efforts implement parts of the envisioned top-down strategy but none have bridged the gap to link top-down analytics to the bottom-up data. Some of the closest existing efforts include the various hazard and control banding tools [10], as well as the SUN [11] and LICARA [12] projects of the European Union Seventh Framework Programme. The need for comprehensive top-down approaches was called for after the NNI workshop and decision analytic tools were specifically mentioned as a way of supplementing data intensive visualization methods for the goals of risk management [5,13,14].

For a successful nanoinformatics enterprise, top-down decision analytic tools and bottom-up data management methods need to be integrated. Decision analytic tools are able to bridge the gap between the data needed and the data available to make informed decisions about a new technology. Decision analysis typically formulates models for important decisions in order to identify which alternatives are most desirable given the available information and the preferences of the decision makers, thus incorporating the top-down (decision) perspective. In addition, once decision modeling structures are in place, it is possible to shift attention from selection of alternatives to understanding the data's support for those alternatives. In other words, decision modeling structures can be used to first synthesize information toward a decision focus and second to identify gaps and delve further in areas of need in order to establish which particular data would be most relevant to the decisions at hand. The ability of decision modeling to identify the relevance of existing data and to distill which areas of research would be most helpful are especially useful when large amounts of data are available and when the data are uncertain and ambiguous.

This paper discusses several decision analytic tools that hold promise for nanoinformatics. We describe the methodology and application of case studies. In particular, we review the use of multicriteria decision analysis (MCDA), value of information (VOI), weight of evidence (WOE), and portfolio decision analysis (PDA) from the perspective of nanoinformatics. We propose that this set of decision analytic methods should be explicitly developed as the next step to advance the nanoinformatics vision of efficiently guiding research and seamlessly identifying and synthesizing available information for decision making.

Discussion

Multicriteria decision analysis

Multicriteria decision analysis (MCDA) refers to a set of methods that are employed to rank decision alternatives from most to least preferred. To accomplish this, MCDA allows the user to break down complex problems into more manageable pieces, assess those pieces with respect to the relevant data for each alternative, and reassemble them to present an overall conclusion to decision makers [15]. The process of completing an MCDA can be divided into four steps: (1) identifying the problem, the stakeholders, and the criteria relevant to the decision; (2) extracting weights, thresholds, and other parameters to be inputs in the mathematical model, and assigning measurements for each alternative; (3) executing the model via software; and (4) evaluating the results of the model [16].

MCDA can be applied to nanoinformatics decisions, for example, to help users evaluate and choose a nanomaterial type, formulation, fabrication technique, supplier, coating, or risk management strategy for a new product. From a portfolio of alternatives, MCDA pinpoints those that are most worthy of further consideration based on an aggregated score across all selected evaluation criteria. Most nanomaterial hazard and control banding tools implicitly implement MCDA by using physiochemical property data to relate hazard scores to indi-

vidual criteria. The criteria are weighted by importance, and the sum of these weighted scores is used to derive an overall hazard score for a nanomaterial. In this way, MCDA-based tools can synthesize data in the context of material development decisions to identify materials with the highest overall hazard scores, typically omitted from use or selected for additional study. The MCDA structure can be used to loosely guide more detailed research and development, because the criteria most in need of further review can be compared in the decision model to find which has the greatest contribution to the overall hazard score [17].

In a case study by Tervonen et al. [18], an MCDA framework was applied for the classification of five nanomaterials: nC_{60}, multiwalled carbon nanotubes (MWCNTs), CdSe, silver nanoparticles (Ag NPs), and aluminum nanoparticles (Al NPs). The SMAA-Tri MCDA model was selected as it is well suited for the classification of nanomaterials with uncertain or unavailable physiochemical properties. Five extrinsic characteristics (agglomeration, reactivity, critical functional groups, particle size and contaminant dissociation) and three factors that are dependent on the characteristics listed above and that may influence hazards (bioavailability, bioaccumulation and toxic potential) were used to evaluate the selected nanomaterials [18].

Five alternative risk classifications were proposed for the materials: extreme risk, high risk, medium risk, low risk, and very low risk. The nanomaterials were sorted based on the probability of classification in a particular risk category, given complete information. CdSe was identified as the nanomaterial most likely to receive the highest hazard score, with a 98% chance of being categorized as "high risk." With these results in mind, the contribution of each criterion to the total score can be evaluated to see which of the eight factors might reasonably benefit from further investigation [18]. This method of determining relevant information with MCDA is a top-down approach. Decision analysis starts with the research objective and ends with decision making. Standard risk assessments, on the other hand, begin with data and end with risk measurements [4]. By starting with the goal of the research, the top-down approach is able to clarify the research needed to achieve the objective and to efficiently make an informed decision.

Beyond this, a series of next steps can be explored to expand the use of MCDA in nanoinformatics. Hazard and control banding tools can be tailored for each funding or regulatory agency's mission and goals, and additional tools can be developed to meet the needs of other common types of decisions. Furthermore, MCDA capabilities can be integrated into existing nanoinformatics platforms to let users develop their own top-down frameworks, which are linked to the bottom-up data, and to interactively explore evaluations of the best materials for a given design or product. Finally, MCDA can potentially address the need for rapid, real-time screening of nanomaterial hazards and the need for incorporating cost–benefit information alongside environment, health and safety data in a cost–benefit screening.

Value of information

Value of information (VOI) is a decision analytic concept characterizing the amount a decision maker would pay to acquire additional information that would improve the quality of a decision [19]. As such, it prioritizes research based on its decision relevance, which is the degree to which it is expected to reduce uncertainty regarding the best alternative. Decision relevance is context dependent but vastly more nuanced than approaches that only consider the magnitude of uncertainties in the unweighted and uncontextualized underlying data. Specifically, to calculate the VOI associated with a decision under uncertainty, (i) the best perceived alternative is selected with the benefit of some contemplated information; these outcomes will always be, on average, preferable or at least equal to those of the same decision where (ii) the best perceived alternative is selected in the absence of that information. The expected value of information is the maximum cost which would be spent to get that information while still leaving the decision maker indifferent between (i) and (ii).

The significance of new nanomaterial research and data for a decision maker is often initially unknown. Ideally, further studies would be prioritized such that research plans addressing the greatest amount of uncertainty, or eliminating the uncertainties the decision maker most wants to eliminate, are completed first. The VOI is able to quantify the benefits of this complex bundle of information for a particular decision making situation. In some cases, the VOI also locates a point at which enough information is known, that is, where the marginal returns to additional information diminish to less than the marginal cost of obtaining that information [19].

In a case study from Linkov et al., an MCDA framework evaluates four alternative technologies for single wall carbon nanotube synthesis and a VOI model prioritizes further research [20]. The MCDA process identified pertinent criteria: synthesis cost, material efficiency, energy consumption, life cycle environmental impacts, and risks to human health. A probability distribution of scores for each technology was specified for each criterion via author judgment and the literature. Monte Carlo simulations were used to normalize and aggregate individual criteria distributions into distributions of overall performance using criteria weights associated with preferences of different key stakeholders [20].

After developing result distributions that reflect current uncertainties, the study evaluated research that might best improve decision confidence. Monte Carlo simulations of possible research outcomes (to reduce uncertainty in the input data) and decision outcomes (resulting reduced uncertainty in the distributions of overall scores) were produced for each nanomaterial, showing the likelihood that each nanomaterial would rank first for each stakeholder under different research efforts. This revealed the VOI in terms of increase in the average score of the best alternative selected with the benefit of increasing manufacturing research, health research, both types of research, or neither. The VOI analysis showed that the biggest potential gain in decision confidence in that case would come from health research, which would substantially increase confidence in decisions for both regulators and environmental groups, but not for other stakeholders. In contrast, additional manufacturing research would not substantially improve decision confidence for any of the stakeholders [20]. Applied broadly, this type of analysis can provide a strong basis for identifying and promoting research relevant to future technology development.

A series of next steps can be explored for including VOI in nanoinformatics efforts. Databases can be expanded to include uncertainties for criteria other than hazards (e.g., cost or performance), providing a foundation in the data for the VOI. This is important because research activities that quantify or reduce uncertainty about environmental concerns, material costs, and other cost–benefit parameters are of great value to funding agencies and scientists. Like the suggestion for MCDA technology, VOI algorithms can be imbedded within existing nanoinformatics platforms and tied to the data, putting new capabilities into the hands of the user. Finally, VOI can potentially enable the continuous and immediate classification of uncertainties based on aggregated nanoinformatics data. In this way, the focus could be shifted towards those uncertainties that are relevant to technologies with high potential.

Weight of evidence

A major challenge in nanoinformatics is how to compare and harmonize the large volume of independently derived, possibly conflicting, and possibly incompatible data into a coherent argument. Weight of evidence (WOE) is a method of integrating and aggregating different and diverse types of evidence to draw a conclusion [21]. The WOE method can be used to fuse information such that discrepancies in data quality and gaps in evidence are considered [21]. WOE was first introduced in the form of a Bayesian model [22] that updates prior beliefs about a hypothesis to form posterior beliefs due to the introduction of new evidence. In this formulation, the Bayes factor is defined as the ratio of prior odds to posterior odds, and the WOE is the natural logarithm of the Bayes factor. More varied qualitative and quantitative applications of the WOE methodology have evolved since then [23].

On the basis of experience with WOE approaches, the National Research Council has recommended a shift towards defensible qualitative and quantitative methods. Quantitative Bayesian approaches and MCDA were both recommended as quantitative supplements and replacements for solely qualitative WOE practices. Thus, the Bayesian approach is able to account for uncertainty and varied sources and types of evidence, while the MCDA approach considers the quality of the evidence and its source as criteria [23]. As in the previous sections, information is first synthesized using the analytical tools, and from this, critical information for decisions or further nanomaterial research is identified.

A case study by Hristozov et al. used a quantitative WOE framework to evaluate the hazards associated with titanium dioxide nanoparticles. Three sets of criteria (physiochemical properties, toxicity, and data quality) were used to evaluate and calculate the hazard scores by means of MCDA. Uncertainties derived from expert judgment were considered in Monte Carlo simulations [24]. As with MCDA, once the WOE hazard score is determined, each contributor to the hazard score can be further reviewed to see which had the largest effect on the score and which might benefit from further research.

A series of next steps can also be explored for including WOE in nanoinformatics efforts. When data is added to nanoinformatics databases, additional quantitative and qualitative metrics (e.g., data statistical significance, precision, applicability, soundness, completeness, uncertainty and variability, degree of review) can be included to contextualize the weight that each data source should carry based on its relevance, quality, resolution, etc. WOE approaches can also be imbedded in nanoinformatics toolsets to help users clarify conflicting and uncertain evidence for early stage nanomaterial evaluations. WOE approaches can be implemented alongside or within hazard and control banding tools to allow differentiation between input data. In the future, continuous and immediate application of a standardized WOE approach with nanoinformatics data could provide a real-time and more accurate initial summary of nanomaterial hazards or other conclusions that can be drawn from the body of knowledge [24].

Portfolio decision analysis

Portfolio decision analysis (PDA) is similar in aim to the tools discussed earlier, but with one major distinction: instead of choosing one option from a set of choices, a subset of items (a portfolio) is selected [25]. The MCDA, VOI, and WOE methods are all appropriate for use with either single choice

decision analysis or portfolio decision analysis. Once a series of possible portfolios has been evaluated, the portfolios with the highest score at any given budget or level of resource availability can be further investigated. The nanomaterials that contribute most to the portfolio score will be identified, along with the qualities shared among the high scoring nanomaterials.

Bates et al. applied PDA to sets of nanomaterial hazard research efforts, in order to prioritize research portfolios at the national level. This PDA was an extension of a VOI approach evaluating multiple research topics for three emerging nanomaterials: multiwalled carbon nanotubes, silver nanoparticles, and titanium dioxide nanoparticles [26]. First, a preliminary screening tool (CB Nanotool 2.0 [17], an MCDA-based approach) was used to assign distributions of hazard scores for each characteristic of a chosen nanomaterial. These scores were summed across properties to assign a distribution of overall hazard scores for each material. Based on these total scores, the materials were probabilistically classified as high risk, moderate risk, and low risk.

From there, the VOI model estimated the improvement in hazard-identification accuracy for each unique research effort. Each research effort was assumed to reduce the uncertainty associated with a single parameter for a single nanomaterial. Research portfolios for each nanomaterial were defined as sets of research efforts addressing parameters for that material. Monte Carlo simulations were used to estimate the expected benefit of each research effort and portfolio, with the assumption that research undertaken on a material property would reveal a true hazard score prior to the decision, and otherwise, that score would only become known after material classification. For each realization of the simulation, the correct score and classification of the material are assumed to be the score and classification identified when all parameter values are known. The proportion of realizations for which a research portfolio is expected to lead to the correct classification and the degree to which it produces hazard scores matching the correct hazard scores can be tabulated. By comparing this performance to that of a baseline portfolio in which no research is done, it is possible to determine the average increase in value for each research portfolio. These calculations are properly performed at the portfolio level because the potential for any given effort to affect a material's classification and significantly reduce hazard uncertainty depends on the state of knowledge of other parameters for the material [26].

To better reflect the national decisions that are typical of funding agencies, the portfolios of research efforts were also aggregated across materials. Plotting each aggregated portfolio's increase in performance against its difficulty or cost revealed an efficient set of most desirable portfolios (those with a value higher than any others of similar cost) [26]. It is then simple to inspect any of these types of portfolios and observe what research on which nanomaterials and properties might be most attractive at different levels of overall investment.

A series of next steps can also be explored for including PDA in nanoinformatics efforts. Funding agencies, research institutions, corporations, and individual research teams can use nanoinformatics data with PDA techniques to help prioritize future research efforts. PDA algorithms can be tailored to work more seamlessly with existing and future MCDA, VOI, and WOE tools supporting decisions in nanotechnology. Finally, as with the other tools, PDA algorithms can be added to nanoinformatics tool sets to put greater top-down analytical power in the hands of the end user.

Conclusion

Recent discussions from the Nanotechnology Knowledge Infrastructure have heralded the creation of a communication portal for the various nanotechnology databases and tools. The tremendous amount of data that would be available via that portal would necessitate not only the bottom-up accumulation, sorting, and visualization of data, but the top-down identification of decision-relevant information. The four tools described here can accomplish both facets of that goal, and overall, provide capability to expand the reach of current nanoinformatics tools.

Part of this expansion should be accomplished through use of expert elicitations, which are often featured in decision analysis to supplement and connect hard data to the decision while leaving a transparent record of the way in which this connection is made. In the context of nanoinformatics, properly implemented human judgments can help users navigate and incorporate available information resources. Each of the applications described herein uses such judgments. The weights on criteria for a given stakeholder are nearly always subjectively assigned (although they use techniques that are transparent, maximize logical consistency, and minimize psychological biases). While some uncertainties involving the outcome of repetitive processes can be readily characterized on the basis of statistical data, it may be impossible or inadequate to do so in situations involving new or ambiguous factors. It is a philosophical point emphasized in decision analysis that in making choices, it is rational for decision makers to act consistently with what is implied by their beliefs in conjunction with the information they have.

The use and implementation of these decision analytic techniques are not without challenges [27]. These include involving

the right experts and stakeholders so that results will be credible, guarding against motivational and other biases in elicitation and dissemination [28], and communication of results in a way that they will be known, understood and trusted by the people who can use them [29]. In addition, the academic decision analysis community is often focused on the creation of new tools, and is less interested in their immediate application. Open advocacy and networking from the community could better relay the benefits of these approaches and techniques.

Thus expanded from information retrieval to decision support, nanoinformatics has the potential to improve the characterization of nanomaterials, the reproducibility of nanomaterial research, and the accessibility of data. Currently, nearly all nanoinformatics efforts are working from a bottom-up perspective to create databases and archives and to organize all of the available data instead of employing a top-down decision approach to identify relevant data. Without the incorporation of both top-down and bottom-up concepts, the full definition and scope of the nanoinformatics vision may not be realized. A range of decision analytic techniques, starting with MCDA, VOI, WOE, and PDA, as described here, can help to sort through and organize the vast nanomaterial data to inform both current choices and the prioritization of future nanomaterial research. These techniques focus the attention of researchers and policy makers toward what is most relevant to their decisions and provide consistent and transparent frameworks for integrating that information. In the future, we expect that both decision analytic techniques and Bayesian models will be used as extensions of standard mechanistic and statistical models to leverage and advance developments in nanoinformatics [21].

Acknowledgements

This work was funded by the nanotechnology and emerging material risk research programs of the US Army Engineer Research and Development Center. Publication of this material has been approved by authority of the Chief of the US Army Corps of Engineers.

References

1. Nanomaterial Registry. https://www.nanomaterialregistry.org/ (accessed March 13, 2015).
2. NSI: Nanotechnology Knowledge Infrastructure (NKI) – Enabling National Leadership in Sustainable Design. http://www.nano.gov/NSINKI (accessed March 8, 2015).
3. http://nanotechinformatics.org/nanoinformatics/index.php/Main_Page (accessed March 13, 2015).
4. Linkov, I.; Anklam, E.; Collier, Z. A.; DiMase, D.; Renn, O. *Environ. Syst. Decisions* **2014,** *34,* 134–137. doi:10.1007/s10669-014-9488-3
5. National Nanotechnology Initiative. Stakeholder Perspectives on Perception, Assessment, and Management of the Potential Risks of Nanotechnology. http://www.nano.gov/sites/default/files/pub_resource/2013_nni_r3_workshop_report.pdf (accessed March 15, 2015).
6. Morse, J. Nanoinformatics 2015: Enabling Successful Discovery and Applications. http://www.internano.org/content/view/948/251/ (accessed March 25, 2015).
7. National Institute for Occupational Safety and Health. Nanoparticle Information Library (NIL). http://nanoparticlelibrary.net/index.asp (accessed March 13, 2015).
8. https://nanohub.org/ (accessed March 13, 2015).
9. GoodNanoGuide. https://nanohub.org/groups/gng (accessed March 13, 2015).
10. Brouwer, D. H. *Ann. Occup. Hyg.* **2012,** *56,* 506–514. doi:10.1093/annhyg/mes039
11. SUN Project - Sustainable Nanotechnologies Project. http://www.sun-fp7.eu/ (accessed April 17, 2015).
12. EU NanoSafety Cluster. http://www.nanosafetycluster.eu/eu-nanosafety-cluster-projects/seventh-framework-programme-projects/licara.html (accessed April 17, 2015).
13. Isaacs, J. A.; Alpert, C. L.; Bates, M.; Bosso, C. J.; Eckelman, M. J.; Linkov, I.; Walker, W. C. *Environ. Syst. Decisions* **2015,** *35,* 24–28. doi:10.1007/s10669-015-9542-9
14. Subramanian, V.; Semenzin, E.; Hristozov, D.; Zondervan-van den Beuken, E.; Linkov, I.; Marcomini, A. *Environ. Syst. Decisions* **2015,** *35,* 29–41. doi:10.1007/s10669-015-9541-x
15. Belton, V.; Stewart, T. J. *Multiple Criteria Decision Analysis: An Integrated Approach;* Kluwer Academic Publishers: Norwell, MA, U.S.A., 2002. doi:10.1007/978-1-4615-1495-4
16. Linkov, I.; Satterstrom, F. K.; Monica, J. C., Jr.; Hansen, S. F.; Davis, T. A. *Nanotechnol. Law Bus.* **2009,** *6,* 203–220.
17. Zalk, D. M.; Paik, S. Y.; Swuste, P. *J. Nanopart. Res.* **2009,** *11,* 1685–1704. doi:10.1007/s11051-009-9678-y
18. Tervonen, T.; Linkov, I.; Figueira, J. R.; Stevens, J.; Chappell, M.; Merard, M. *J. Nanopart. Res.* **2009,** *11,* 757–766. doi:10.1007/s11051-008-9546-1
19. Raiffa, H.; Schlaifer, R. *Applied Statistical Decision Theory;* Clinton Press, Inc.: Boston, MA, U.S.A., 1961.
20. Linkov, I.; Bates, M. E.; Canis, L. J.; Seager, T. P.; Keisler, J. M. *Nat. Nanotechnol.* **2011,** *6,* 784–787. doi:10.1038/nnano.2011.163
21. Linkov, I.; Massey, I.; Keisler, J.; Rusyn, I.; Hartung, T. *ALTEX* **2015,** *32,* 3–8. doi:10.14573/altex.1412231
22. Good, I. J. *J. R. Stat. Soc.* **1960,** *22,* 319–331.
23. Statistica Help – Weight of Evidence (WOE) Introductory Overview. http://documentation.statsoft.com/STATISTICAHelp.aspx?path=WeightofEvidence/WeightofEvidenceWOEIntroductoryOverview (accessed March 8, 2015).
24. Hristozov, D. R.; Zabeo, A.; Foran, C.; Isigonis, P.; Critto, A.; Marcomini, A.; Linkov, I. *Nanotoxicology* **2014,** *8,* 72–87. doi:10.3109/17435390.2012.750695
25. Salo, A.; Keisler, J.; Morton, A. *Portfolio Decision Analysis: Improved Methods for Resource Allocation;* Springer: Berlin, Germany, 2011. doi:10.1007/978-1-4419-9943-6
26. Bates, M. E.; Keisler, J. M.; Zussblatt, N. P.; Plourde, K. J.; Wender, B. A.; Linkov, I. *Nat. Nanotechnol.*, in press.
27. Parnell, G. S.; Bresnick, T.; Tani, S. N.; Johnson, E. R. *Handbook of decision analysis;* John Wiley & Sons, Inc.: Hoboken, NJ, U.S.A., 2013. doi:10.1002/9781118515853

28. Hora, S. Eliciting probabilities from experts. In *Advances in Decision Analysis: From Foundations to Applications;* Edwards, W.; Miles, R. F., Jr.; von Winterfeldt, D., Eds.; Cambridge University Press: New York, NY, U.S.A., 2007; pp 129–153.
29. Keisler, J. M.; Noonan, P. S. *Decis. Anal.* **2012,** *9,* 274–292. doi:10.1287/deca.1120.0238

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

The eNanoMapper database for nanomaterial safety information

Nina Jeliazkova*[1], Charalampos Chomenidis[2], Philip Doganis[2], Bengt Fadeel[3], Roland Grafström[3], Barry Hardy[4], Janna Hastings[5], Markus Hegi[4], Vedrin Jeliazkov[1], Nikolay Kochev[1,6], Pekka Kohonen[3], Cristian R. Munteanu[7,8], Haralambos Sarimveis[2], Bart Smeets[7], Pantelis Sopasakis[2,9], Georgia Tsiliki[2], David Vorgrimmler[10] and Egon Willighagen[7]

Full Research Paper

Open Access

Address:
[1]Ideaconsult Ltd., Sofia, Bulgaria, [2]National Technical University of Athens, School of Chemical Engineering, Athens, Greece, [3]Karolinska Institutet, Stockholm, Sweden, [4]Douglas Connect GmbH, Zeiningen, Switzerland, [5]European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom, [6]Department of Analytical Chemistry and Computer Chemistry, University of Plovdiv, Plovdiv, Bulgaria, [7]Department of Bioinformatics, NUTRIM, Maastricht University, Maastricht, The Netherlands, [8]Computer Science Faculty, University of A Coruna, A Coruña, Spain, [9]IMT Institute for Advanced Studies Lucca, Lucca, Italy and [10]in silico toxicology Gmbh (IST), Basel, Switzerland

Email:
Nina Jeliazkova* - jeliazkova.nina@gmail.com

* Corresponding author

Keywords:
database; EU NanoSafety Cluster; nanoinformatics; nanomaterials; nanomaterials ontology; NanoQSAR; safety testing

Beilstein J. Nanotechnol. **2015,** *6,* 1609–1634.
doi:10.3762/bjnano.6.165

Received: 31 March 2015
Accepted: 03 July 2015
Published: 27 July 2015

This article is part of the Thematic Series "Nanoinformatics for environmental health and biomedicine".

Guest Editor: R. Liu

Abstract

Background: The NanoSafety Cluster, a cluster of projects funded by the European Commision, identified the need for a computational infrastructure for toxicological data management of engineered nanomaterials (ENMs). Ontologies, open standards, and interoperable designs were envisioned to empower a harmonized approach to European research in nanotechnology. This setting provides a number of opportunities and challenges in the representation of nanomaterials data and the integration of ENM information originating from diverse systems. Within this cluster, eNanoMapper works towards supporting the collaborative safety assessment for ENMs by creating a modular and extensible infrastructure for data sharing, data analysis, and building computational toxicology models for ENMs.

Results: The eNanoMapper database solution builds on the previous experience of the consortium partners in supporting diverse data through flexible data storage, open source components and web services. We have recently described the design of the

eNanoMapper prototype database along with a summary of challenges in the representation of ENM data and an extensive review of existing nano-related data models, databases, and nanomaterials-related entries in chemical and toxicogenomic databases. This paper continues with a focus on the database functionality exposed through its application programming interface (API), and its use in visualisation and modelling. Considering the preferred community practice of using spreadsheet templates, we developed a configurable spreadsheet parser facilitating user friendly data preparation and data upload. We further present a web application able to retrieve the experimental data via the API and analyze it with multiple data preprocessing and machine learning algorithms.

Conclusion: We demonstrate how the eNanoMapper database is used to import and publish online ENM and assay data from several data sources, how the "representational state transfer" (REST) API enables building user friendly interfaces and graphical summaries of the data, and how these resources facilitate the modelling of reproducible quantitative structure–activity relationships for nanomaterials (NanoQSAR).

Introduction

Nanotechnology is an increasingly dynamic area in materials science research and development, introducing novel materials with unique properties due to their size in the range of nanometres. A database and framework supporting nanomaterials safety has to comply with diverse requirements, set-up by the nanotechnology community. A number of challenges exist in the representation and integration of engineered nanomaterials (ENMs) data mainly due to the complexity of the data and the need to capture provenance.

Physicochemical identity

The eNanoMapper framework must capture the physical and chemical identity of ENMs, including the notion of mixtures and their particle size distributions, differences in the amount of surface modification, manufacturing conditions and batch effects. It must also capture the biological identities (e.g., toxicity pathways, effects of ENM coronas, modes of action), interactions (cell lines, assays) and a wide variety of measurements. A number of analytic techniques have been proposed and developed to characterise the physicochemical properties of nanomaterials, including the commonly used dynamic light scattering to measure the particle size distribution and zeta potentiometry to estimate the pH-dependent surface charge.

Biological identity

With the expanding insight into the factors determining toxicity, the list of measurable effects is growing increasingly long. The need for validated in vitro tests has been advocated since 2006 [1]. It is proposed to extend the list of endpoints for hazard identification to include cell uptake, cell viability, oxidative stress, inflammation, fibrosis, immunotoxicity, cardiovascular toxicity, ventilation rate, gill pathologies, mucus secretion and brain pathology. The EU guidance document lists the main known effects from experimental studies [2]. High-throughput omics data and kinetics [3] are becoming increasingly important in the assessment of nanomaterials, presenting challenges in both data management and analysis. A common requirement of all categories of users is to link the ENM entries with studies in which toxicology or biological interference of the nanomaterials have been studied, in addition to an accurate physicochemical characterisation.

Data input, data formats, provenance, visualisation

The framework should allow for the representation of data and facts compatible with regulatory expectations and (inter)national standards. This usually translates into a set of available study summaries (rarely raw data) for a given ENM. The inclusion of links to product databases could also be considered (e.g., whether the nanomaterial occurs in nature, whether it is emitted by cars or is present in certain food sources, as well as known therapies in which the nanomaterial is used). However, supporting raw data files (including microscopy images) is an important requirement in contexts other than regulatory, enabling the reproducibility of the data preprocessing and analysis. Links to the corresponding protocols and data sources should be added, where available. Clear visualisation of nanomaterials that goes beyond just structural formulae should be available, in order to make the data less abstract biologists with less knowledge about nanomaterials.

Support for data analysis

The modelling community presents a different requirement: the data analysis methods usually require a "spreadsheet" or matrix view of data for multiple ENMs. The experimental data in the public datasets are usually not in a form appropriate for modelling. Standardisation in these sources is specific to each database. Even in curated collections the preparation of data for modelling is not a straightforward exercise (e.g., the experimental values can be merged in many different ways into a matrix, depending on which experimental protocols and conditions are considered similar; also there could be multiple values due to replicates or similar experiments). The framework should

allow for the addition of information based on the outcomes of the predictive toxicology models, including the biological role of the ENM, clearance, accumulation, and pathway information (e.g., WikiPathways entries [4]).

Existing databases

Several databases exist that are relevant for ENM toxicity assessment. They list nanomaterials and a variety of their properties, or products containing nanomaterials: NanoMaterialRegistry (http://www.nanomaterialregistry.org/) [5], Nanoparticle Information Library NIL (http://nanoparticlelibrary.net/) [6], Nanomaterial-Biological Interactions Knowledgebase (http://nbi.oregonstate.edu/), caNanoLab (http://cananolab.nci.nih.gov/caNanoLab/) [7], InterNano (http://www.internano.org/), Nano-EHS Database Analysis Tool (http://icon.rice.edu/report.cfm), nanoHUB (nanohub.org/resources/databases/), NanoTechnology Characterisation Laboratory (http://ncl.cancer.gov/), EC JRC NanoHub (http://www.napira.eu/), the DaNa Knowledge Base (http://nanopartikel.info/) [8], and NanoWerks Nanomaterial Database (http://www.nanowerk.com/). The EU NanoSafety Cluster alone (http://www.nanosafetycluster.eu/) has many projects with database generating activities, such as NanoMiner [9]. An extensive review of existing nano-related data models, databases, and nanomaterials-related entries in chemical and toxicogenomic databases is presented in two recent publications [10,11]. Reviews of emerging databases and analysis tools in nanoinformatics have started to appear in the literature [12]. It becomes clear that nano-related data is relatively abundant, but also quite dispersed across many different sources. Combining data from various sources is hampered by the lack of programmatic access in most cases and the absence (or infrequent use) of suitable domain ontologies.

Experimental

The eNanoMapper prototype database (http://data.enanomapper.net/) is part of the computational infrastructure for toxicological data management of ENM, developed within the EU FP7 eNanoMapper project [13]. It provides support for upload, search and retrieval of nanomaterials and experimental data through a REST web services API (http://enanomapper.github.io/API/) and a web browser interface. It is implemented by a customized version of AMBIT web services [14]. The database has been populated with content provided by project partners. We have recently described the design of the eNanoMapper prototype database [10] along with a summary of ENM data representation challenges and comparison to existing data models used to describe nanomaterials and assay data. The focus of this paper is the database functionality exposed through an application programming interface (API), and the use of the API for visualisation and modelling. While starting from the chemical compound-centric OpenTox API, the eNanoMapper prototype database implements a REST API, allowing for the representation of chemical substances with complex composition, and experimental data associated with those substances. The NMs are considered a special case of substances, which is consistent with the ontology representations, ECHA guidelines and peer-reviewed publications as elaborated in the next section.

Chemical structures, substances, nanomaterials and measurements

The Nano Particle Ontology (NPO) defines a nanomaterial (NPO_199) as equivalent to a chemical substance (NPO_1973) that has as constituent a nano-object, nanoparticle, engineered nanomaterial, nanostructured material, or nanoparticle formulation. Chemical substances are classified as types of chemical entity (NPO_1972). The default approach for representation of chemical compounds in ISA-Tab [15] is an ontology entry, which typically points to a single chemical structure. This is insufficient for describing substances of complex composition such as nanomaterials, hence a material file was introduced to address this need in ISA-Tab-Nano [16]. The latest ISA-Tab-Nano 1.2 specification recommends using the material file only for material composition and nominal characteristics, and to describe the experimentally determined characteristics in regular ISA-Tab assay files. The definitions of the terms “substance” and “material” are discussed in [17], comparing ISO, REACH and general scientific definitions of the terms. The REACH definition of a substance encompasses all forms of substances and materials on the market, including nanomaterials; a substance may have complex composition. The paper [17] notes that the OECD Harmonized Templates (OHT) definition of “reference substances” is very similar to the definition of the term “reference material”. The same publication refers to the “test” and “measurement” terms as the fundamental concepts [17]. The OECD guideline defines the “test” or “test method” as the experimental system used to obtain the information about a substance. The term “assay” is considered a synonym. The term “testing” is defined as applying the test method. The endpoints recommended for testing of nanomaterials [18] by the OECD Working Party on Manufactured Nanomaterials (OECD WPMN) use the terms and categories from the OECD Harmonized Templates. The NPO distinguishes between the endpoint of measurement (e.g., particle size, NPO_1694) and the assay used to measure the endpoint (e.g., size assay, NPO_1912), where the details of the assay can be further specified (e.g., uses technique electron microscopy, NPO_1428). This structure is generally the same as the one supported by the OHT (e.g., in the OHT granulometry type of experiment several size-related endpoints can be defined, as well as the equipment used, the protocol and specific conditions). The CODATA UDS [19,20] requires specification of

how each particular property is measured. ISA-Tab-Nano also allows for defining the qualities measured and detailed protocol conditions and instruments. The level of detail in the OHT, CODATA UDS, ISA-Tab-Nano and available ontologies differ, which is due to their different focus. Mapping between terms defined in the different sources is an ongoing effort supported by the eNanoMapper ontology team and the EU NanoSafety Cluster database working group. In Supporting Information File 1, we provide a table of OECD WPMN recommended endpoints and their potential correspondence to UDS and ISA-Tab-Nano concepts.

To summarise, the most important data objects necessary to represent nanomaterials and NM characterisation are the substance with its composition, and a data object, able to represent a test method, its application to the substance under specific conditions and the measurements obtained as a result of this process. Therefore, the objects supported by the API are "substances" (as a superclass of nanomaterials), "protocols", "endpoints", "conditions", "protocol applications" and "measurements". A "protocol application" (a term borrowed from ISA-Tab) explicitly describes a single step of the experimental graph, namely the application of a particular protocol with its specific parameters to the source material and includes the corresponding results (be it a sample or data readouts). For the purposes of ENM database integration, the source material is always a chemical substance (ENM) with its composition and linkage, while the result is a set of measurements, each annotated with the relevant endpoints and experimental conditions. While we support importing files generated from IUCLID5 database and thus all OECD WPMN recommended endpoints, the list of endpoints in the database is not fixed, and arbitrary endpoints can be imported through spreadsheets and further annotated with ontology entries. The measurement can be specified by a value, range of values, error measure and units, or by a link to a raw data file (e.g., an image). This representation directly supports the OHT data model, and the notion of a set of measurements is very similar to the measurement group concept in the Bio Assay Ontology (BAO) [21], as well as encompassing the measurement value concept in the CODATA UDS. In order to support raw data, we decided to extend the measurement value beyond scalar values and include links to measurement artifacts, such as image and raw data files, similarly to the ISA-Tab approach. The ability to describe derived measurements, by linking measurement groups, as supported by BAO and implied in UDS, is currently being considered, especially in order to support the modelling activities in eNanoMapper. The data model is sufficiently flexible to represent scenarios like multiple endpoints readouts within a single experiment, dose response data as well as replicated measurements. Examples are shown in the visualisation section.

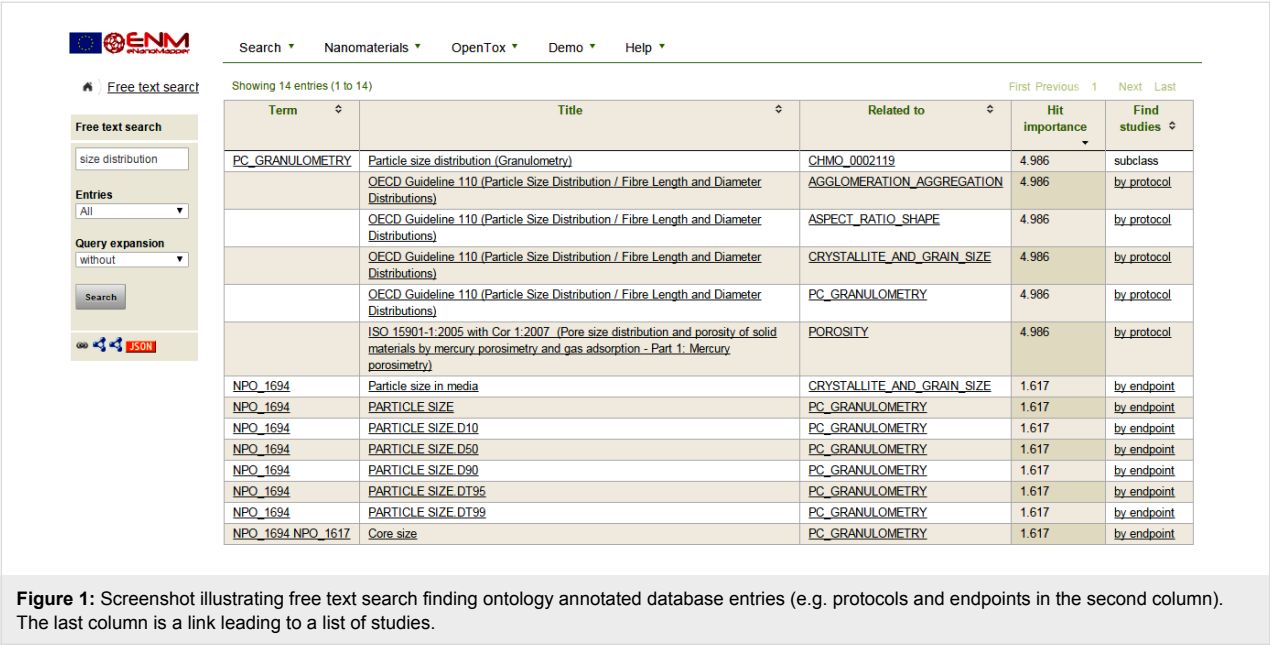
Ontology

The eNanoMapper strategy to adopt and extend ontologies in support of data integration has recently been described [22]. eNanoMapper supports ontology re-use, for example it re-uses the content of the NPO and BAO, through automated modular import of content subsets into an integrated whole. However, the scope of the ontology goes beyond any of the individually imported ontologies, encompassing the whole of the domain of nanomaterial safety assessment. The strategy of re-use of existing ontology content enables downstream annotated data in different repositories to be integrated wherever the same identifiers are used in annotation. The ontology is available at http://purl.enanomapper.net/onto/enanomapper.owl, from BioPortal at http://bioportal.bioontology.org/ontologies/ENM, and for download in full from the development repository on GitHub (https://github.com/enanomapper/ontologies). This section describes the strategy for application of the ontology to the annotation of the prototype eNanoMapper database content.

All data in the database is targeted for annotation with relevant ontology entries from the composite eNanoMapper ontology. Each entry in the ontology has a unique IRI (International Resource Identifier), for example "nanomaterial" (a class imported from the NPO) has the IRI http://purl.bioontology.org/ontology/npo#NPO_199. The IRI consists of an ontology namespace as prefix, followed by a unique identifier for the particular term. For brevity, throughout this manuscript we have referred simply to ontology identifiers (IDs) without the full IRI including the prefix. However, expansion from the short ID to the full IRI is a deterministic transformation. Classes are also associated with a unique label and a descriptive textual definition. The IRI, based on the same underlying Semantic Web technology as the eNanoMapper database prototype, offers a semantics-free stable identifier that is suitable for use in data annotation, as it is resistant to minor changes in the label and improvements in the definition of the class.

Examples of annotations that have already been included in the database are: "particle size distribution (granulometry)" annotated to the ID CHMO_0002119 in the Chemical Methods Ontology namespace, "aspect ratio" annotated to the ID NPO_1365 and "shape" to ID NPO_274 in the NPO namespace (Figure 1).

Annotations are selected from the available classes in the eNanoMapper ontology; a best match approach is used which aims to select the most specific class available for annotation. When no suitable class is present, a suitable class may be found in the broader BioPortal collection which is then targeted for inclusion in the eNanoMapper ontology. If no suitable class exists even within the full collection of ontologies in BioPortal,



Term	Title	Related to	Hit importance	Find studies
PC_GRANULOMETRY	Particle size distribution (Granulometry)	CHMO_0002119	4.986	subclass
	OECD Guideline 110 (Particle Size Distribution / Fibre Length and Diameter Distributions)	AGGLOMERATION_AGGREGATION	4.986	by protocol
	OECD Guideline 110 (Particle Size Distribution / Fibre Length and Diameter Distributions)	ASPECT_RATIO_SHAPE	4.986	by protocol
	OECD Guideline 110 (Particle Size Distribution / Fibre Length and Diameter Distributions)	CRYSTALLITE_AND_GRAIN_SIZE	4.986	by protocol
	OECD Guideline 110 (Particle Size Distribution / Fibre Length and Diameter Distributions)	PC_GRANULOMETRY	4.986	by protocol
	ISO 15901-1:2005 with Cor 1:2007 (Pore size distribution and porosity of solid materials by mercury porosimetry and gas adsorption - Part 1: Mercury porosimetry)	POROSITY	4.986	by protocol
NPO_1694	Particle size in media	CRYSTALLITE_AND_GRAIN_SIZE	1.617	by endpoint
NPO_1694	PARTICLE SIZE	PC_GRANULOMETRY	1.617	by endpoint
NPO_1694	PARTICLE SIZE.D10	PC_GRANULOMETRY	1.617	by endpoint
NPO_1694	PARTICLE SIZE.D50	PC_GRANULOMETRY	1.617	by endpoint
NPO_1694	PARTICLE SIZE.D90	PC_GRANULOMETRY	1.617	by endpoint
NPO_1694	PARTICLE SIZE.DT95	PC_GRANULOMETRY	1.617	by endpoint
NPO_1694	PARTICLE SIZE.DT99	PC_GRANULOMETRY	1.617	by endpoint
NPO_1694 NPO_1617	Core size	PC_GRANULOMETRY	1.617	by endpoint

Figure 1: Screenshot illustrating free text search finding ontology annotated database entries (e.g. protocols and endpoints in the second column). The last column is a link leading to a list of studies.

a request is issued for the class to be added in the eNanoMapper ontology manually. We formally document all such requests via our public GitHub issue tracker (https://github.com/enanomapper/ontologies/issues). Once the term has been included in the ontology it is released to the wider community and becomes available in tools such as BioPortal automatically.

The hierarchical classification structure of the ontology, together with the use of domain-specific relationships, is envisioned to enable intelligent searching, browsing and clustering tools to be developed in the future, as well as to enable templates to be implemented for database content entry compliant with Minimum Information guidelines.

Application programming interface (API)

The eNanoMapper architecture has been informed by the prior experience of several of the authors in designing and building the OpenTox predictive toxicology framework for chemicals [23] and their involvement in developing and supporting the ToxBank [24] data warehouse for the SEURAT-1 research cluster [25]. The framework design adopts the REpresentational State Transfer (REST) software architecture style, a common information model that supports ontology annotation, and an identity service and an access control based on OpenAM [26]. The REST architecture can be briefly summarized as being composed of a collection of information entities (resources), in which each entity can be retrieved by its address and supports a limited number of operations (e.g., read and write). The overall system architecture of eNanoMapper extends the OpenTox [23] and ToxBank [24] designs. Both consist of a set of web services that provide access to experimental protocols, raw and processed data, and data analysis tools. The web services do not need to be deployed on the same machine, but can also be distributed on independent servers. Communication through well-defined interfaces facilitates adding new services, such as services that support new data types or search functionality. The eNanoMapper API is documented online using the Swagger (http://swagger.io/) specification, accessible as interactive documentation at http://enanomapper.github.io/API/.

Substance resource

While the OpenTox framework is intentionally centred on chemical compounds, eNanoMapper uses an extension, allowing representation of chemical substances with a defined composition (Figure 2) and experimental data, associated with substances, rather than associated with chemical structures.

The substance resource supports assigning a nanomaterial type, a chemical composition with relevant concentration and constituents roles, and links to the OpenTox compound resources for specifying the chemical structure, where relevant. NMs are considered a special case of substances. Figure 3 shows the eNanoMapper prototype database user interface displaying the components of a gold nanoparticle with an organic coating. The visualisation is implemented as a JavaScript widget, which consumes the substance API.

The experimental data are assigned to a substance (e.g., nanoparticle) and a JSON (JavaScript Object Notation) representation of the data can be retrieved through a "/substance/{uuid}/study" API call. As an example, in Figure 4,

Figure 2: Top level substance API documentation. The “GET /substance” call is used to retrieve or search a list of NM, subject to multiple query parameters defining the NM search. The “POST /substance” call is used to upload NM and study data in supported formats. The “/substance/{uuid}” call is used to retrieve the substance specified by its unique identifier. Each substance is identified with an unique identifier, generated or specified on import in the form of UUID. The rest of the calls allow to retrieve the component of the NM, the study data and a summary of the available data for the NM, grouped by endpoints.

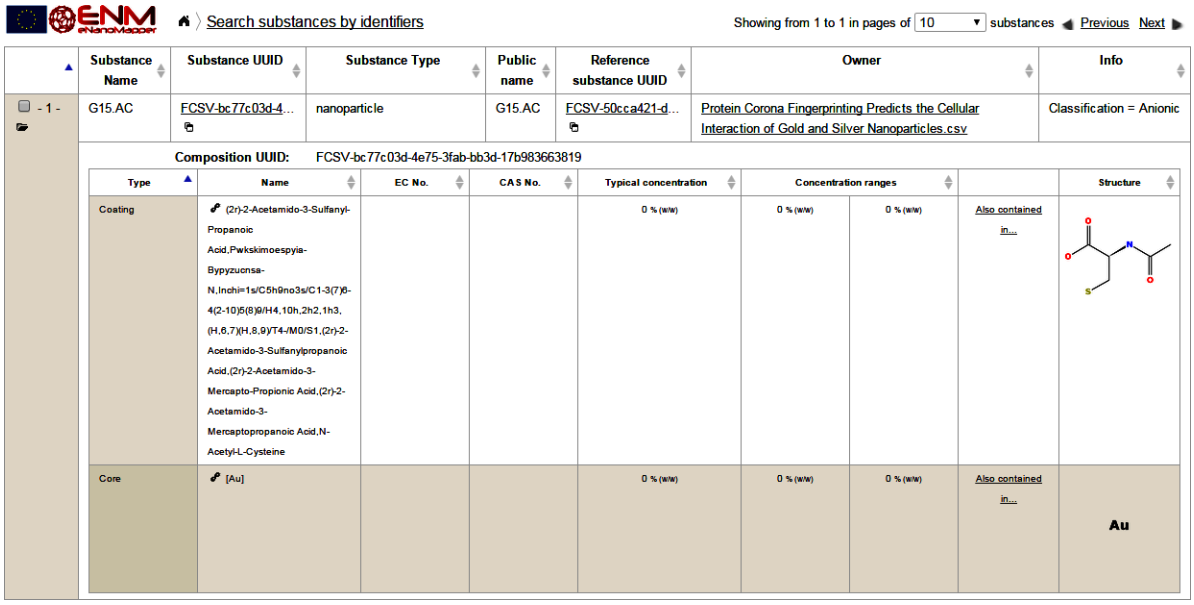


Figure 3: Screenshot showing a nanomaterial entry (a gold nanoparticle with the name G15.AC) and its components (a gold core and organic coating). The components can be retrieved through the “/substance/{uuid}/composition” API call and are linked to the OpenTox API compound resources, which allows for the execution of chemical structure based calculations and predictions. This NM entry is part of the the Protein Corona dataset described below and was imported via a spreadsheet (.csv) file. The “reference substance UUID” refers to the chemical structure, which is considered the main component (Au in this case). The “Owner” column typically refers to the NM manufacturer, or if such information is missing it refers to the data file used for import. The “Info” column may contain an arbitrary key-value data, typically referring to the NM identifiers in other systems.

we present an excerpt from the JSON serialisation of a cell viability assay for the NanoWiki [27] entry with identifier *NWKI-56d49cc3-4a76-354b-9a77-4b2ecb2dbef0*, retrieved from https://apps.ideaconsult.net/enanomapper/substance/NWKI-56d49cc3-4a76-354b-9a77-4b2ecb2dbef0/study.

Similarly to the nanoparticle composition shown in Figure 3, the visualisation of physico-chemical and biological data (Figure 5) is implemented as a JavaScript widget, consuming the substance API.

Search

The API offers access to a variety of searches by substance identifier, any combination of measurement endpoints, and/or chemical structure (Figure 6). The JSON serialisation is the same as above, screenshots of the currently implemented user interface are shown in the Results section.

Data import

The data model (Figure 7) allows for integration of content from a variety of sources, namely OHTs (IUCLID5 .i5z files or

```
{
  "study":[
    {
      "uuid":"NWKI-0271c1d6-e324-4eba-85eb-b060d5807faf",
      "owner":{
        "substance":{
          "uuid":"NWKI-56d49cc3-4a76-354b-9a77-4b2ecb2dbef0"
        },
        "company":{
          "uuid":"NWKI-9f4e86d0-c85d-3e83-8249-a856659087da",
          "name":"NanoWiki"
        }
      },
      "citation":{
        "title":"http://dx.doi.org/10.3109/17435390903276933",
        "year":"0",
        "owner":"Nanotoxicology"
      },
      "protocol":{
        "topcategory":"P-CHEM",
        "category":{
          "code":"PC_GRANULOMETRY_SECTION",
          "title":"4.5 Particle size distribution (Granulometry)"
        },
        "endpoint":"Particle Size"
      },
      "parameters":{
        "DISTRIBUTION_TYPE":null,
        "TESTMAT_FORM":null
      },
      "reliability":{
        "r_isRobustStudy":"false",
        "r_isUsedforClassification":"false",
        "r_isUsedforMSDS":"false",
        "r_purposeFlag":null,
        "r_value":null
      },
    },
}
```


Figure 4: Experimental data JSON example.

direct retrieval of information from IUCLID5 servers, http://iuclid.eu/); custom spreadsheet templates (e.g., Protein Corona CSV files or ModNanoTox Excel files), and custom formats, provided by partners (e.g., the NanoWiki RDF dump [27]). ISA-Tab [15] files are converted by compressing the chain of protocols into a single entry, yet retaining all the protocol parameters and recording the material as a substance and the rest of the factors as experimental conditions. The NanoWiki RDF dump is converted with a custom parser. The supported import formats are currently being extended to include ISA-Tab-Nano [16] and a large set of custom spreadsheet templates.

Taking into account the observation that the use of spreadsheet templates is the preferred approach for data entry by the majority of the EU NanoSafety Cluster projects, we developed a configurable spreadsheet parser facilitating user friendly data preparation and upload. The parser enables import of the data, stored in the supported set of spreadsheet templates, and accommodates different row-based, column-based or mixed organizations of the data. The parser configuration is defined in a separate JSON file, mapping the custom spreadsheet structure into the internal eNanoMapper storage components: “Substance”, “Protocol”, “Measurement”, “Parameters” and “Conditions”. The JSON configuration syntax includes a set of keywords,

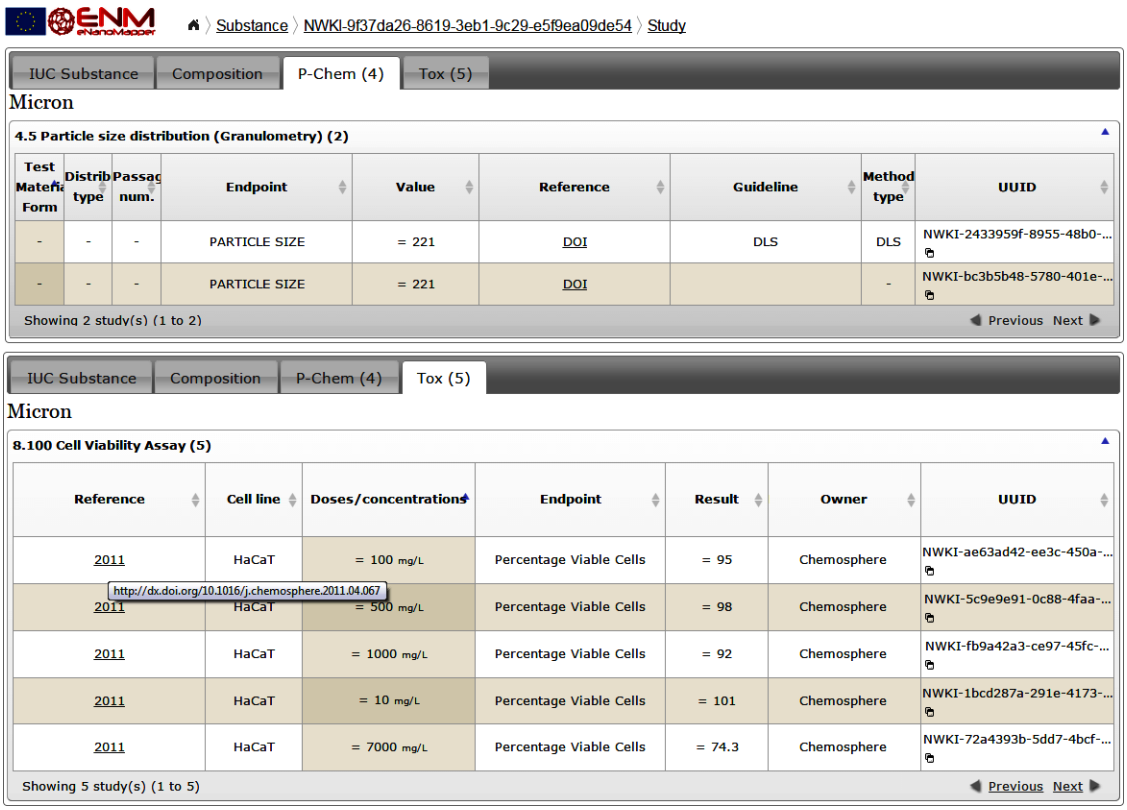


Figure 5: Physico-chemical and toxicity data from the NanoWiki data set.

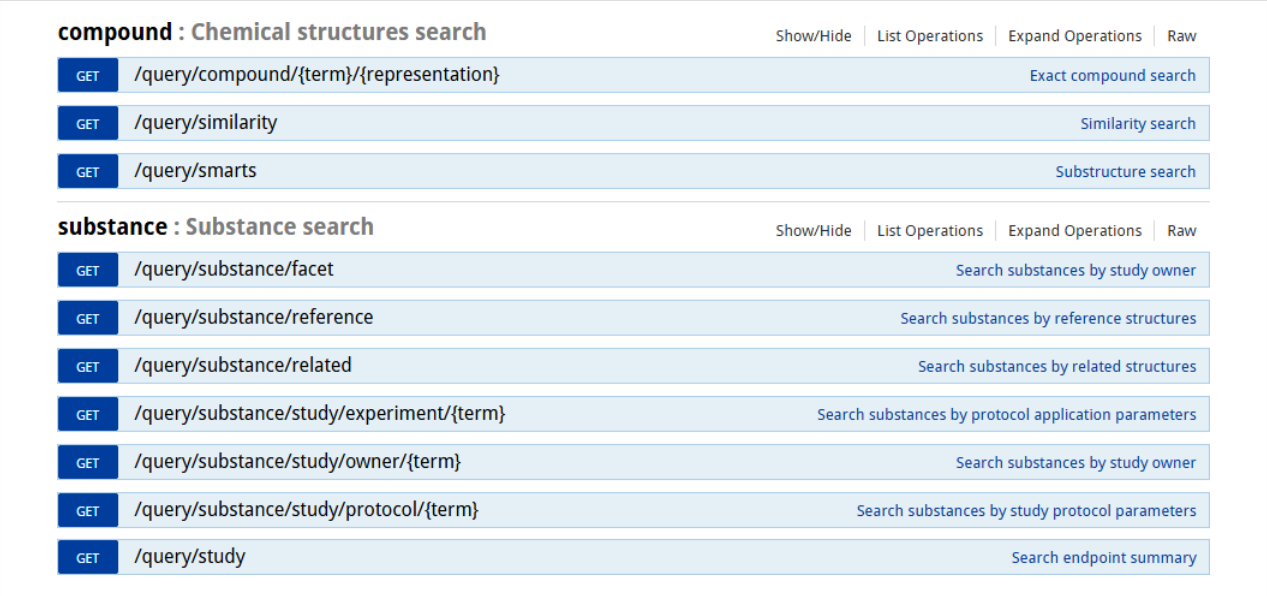


Figure 6: Compound, substance and study search API documentation.

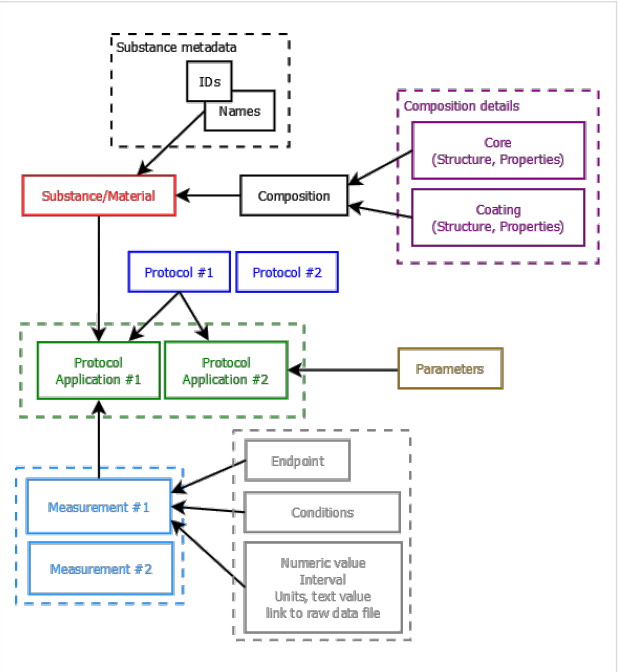


Figure 7: Outline of the data model: Substances are characterised by their "composition" and are identified by their names and IDs. The event of applying a test protocol to a substance/material is described by a "protocol application" entity. Each protocol application consists of a set of "measurements" for a defined "endpoint" under given "condition". The measurement result can be a numeric value with or without uncertainty specified, an interval, a string value, or a link to a raw data file (e.g., a microscopy image).

specifying different strategies for reading the data from one or several sheets, as well as allowing combination of the excel structures (sheets, rows, columns, blocks of cells and cells) into the eNanoMapper data model. The parser code, the JSON syntax, documentation and example files are available at https://github.com/enanomapper/nmdataparser/. The mapping enables a uniform approach towards import, storage and searching of the ENM physicochemical measurements and biological assay results. While the parser itself is open source, the configuration files may not be, thus not revealing the organisation of confidential data templates. The parser is currently being used to parse ModNanoTox templates and confidential templates from EU NanoSafety Cluster projects. Maps of the confidential spreadsheet templates are available on request, in compliance with the agreements between the corresponding projects. More formats will be supported as needed for indexing data from different sources. The development of ISA-Tab-Nano and RDF import and export tools is ongoing.

The data import is performed by HTTP POST to the substance resource (Figure 2), which translates to a regular web form for file upload (Figure 8). The two checkboxes control whether the

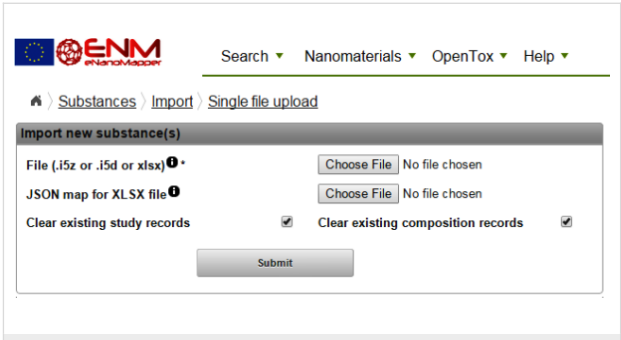


Figure 8: Data upload web page of the database system showing support for two file formats.

composition records and study records for the materials being imported will be cleared, if already in the database. Each material entry in the database is assigned a unique identifier in the form of a UUID. If the input file is *.i5z or *.i5d, the identifiers are the IUCLID5 generated UUIDs already present in these files (e.g., IUC5-5f313d1f-4129-499c-abbe-ac18642e2471). If the input file is a spreadsheet, the JSON configuration defines which field to be used as an identifier and uses the field itself or generates UUID from the specified field (e.g., FCSV-bc77c03d-4e75-3fab-bb3d-17b983663819 indicates the entry imported from CSV file). The parser may be configured to use a custom prefix on import, e.g., "NWKI-" for NanoWiki entries, generating UUID like "NWKI-71060af4-1613-35cf-95ee-2a039be0388a".

Datasets of substances (bundles)

A "bundle" (Figure 9) is a REST resource that groups a selected set of substances and a selected set of endpoints. This functionality was introduced to enable creating groups of diverse nanomaterials, to specify the endpoints of interest, which can vary from physicochemical to proteomics assays, and to enable retrieving all this data with a single REST call. A bundle may include the nanomaterials and assay data from a single investigation as well as serve as a container for a set of NMs and for data (typically representing different experiments) retrieved from the literature. The latter is currently difficult to achieve in ISA-Tab, as its purpose is to capture the experimental graph of a single investigation. The bundle API can be considered an extension of the original OpenTox compound-centric dataset concept to allow for datasets of nanomaterials. The experimental values may include replicates and range values and can be merged in many different ways into a matrix (Figure 10), depending on which experimental protocols and conditions are considered similar. The API in Figure 9 provides one of many possible ways of conversion into a matrix form through the "/bundle/{id}/matrix" call. The users can build external applications, retrieving the experimental data and applying custom conversion procedures, as does the Jaqpot Quattro application described in the "Modelling" section.

Results

The results include using the eNanoMapper database described above to import and publish online ENM and assay data from several sources; as well as the demonstration of how the REST API enables building a user friendly interface and graphical summaries of the data, and last but not least, facilitates reproducible Quantitative Structure Activity Relationship for nanomaterials (NanoQSAR) modelling.

The demonstration data provided by eNanoMapper partners – (i) NanoWiki, (ii) a literature dataset on protein coronas and (iii) the ModNanoTox project dataset – illustrates the capability of the associated REST API to support a variety of tests and endpoints, as recommended by the OECD WPMN.

NanoWiki

NanoWiki was originally developed as an internal knowledge base of the toxicity of, primarily, metal oxides at the Karolinska

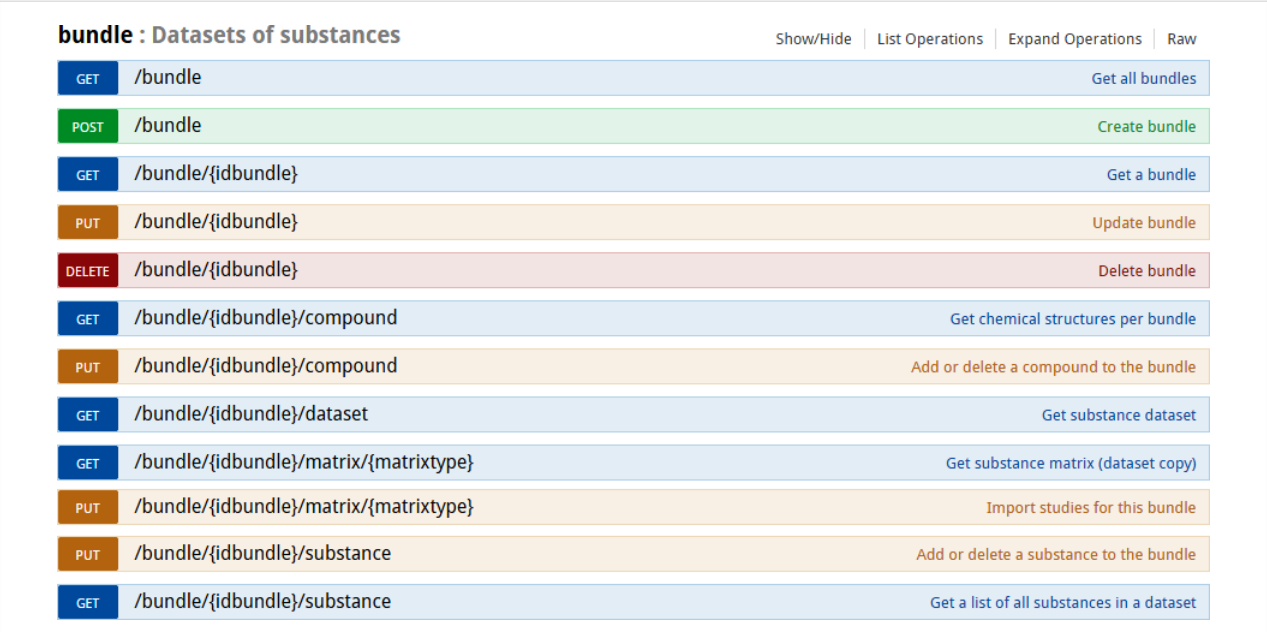


Figure 9: Bundle API documentation at http://enanomapper.github.io/API. A bundle is a REST resource, allowing one to retrieve all information about a selected set of NMs and endpoints by a singe REST call. The PUT calls allow one to select or deselect the NMs and the endpoints.

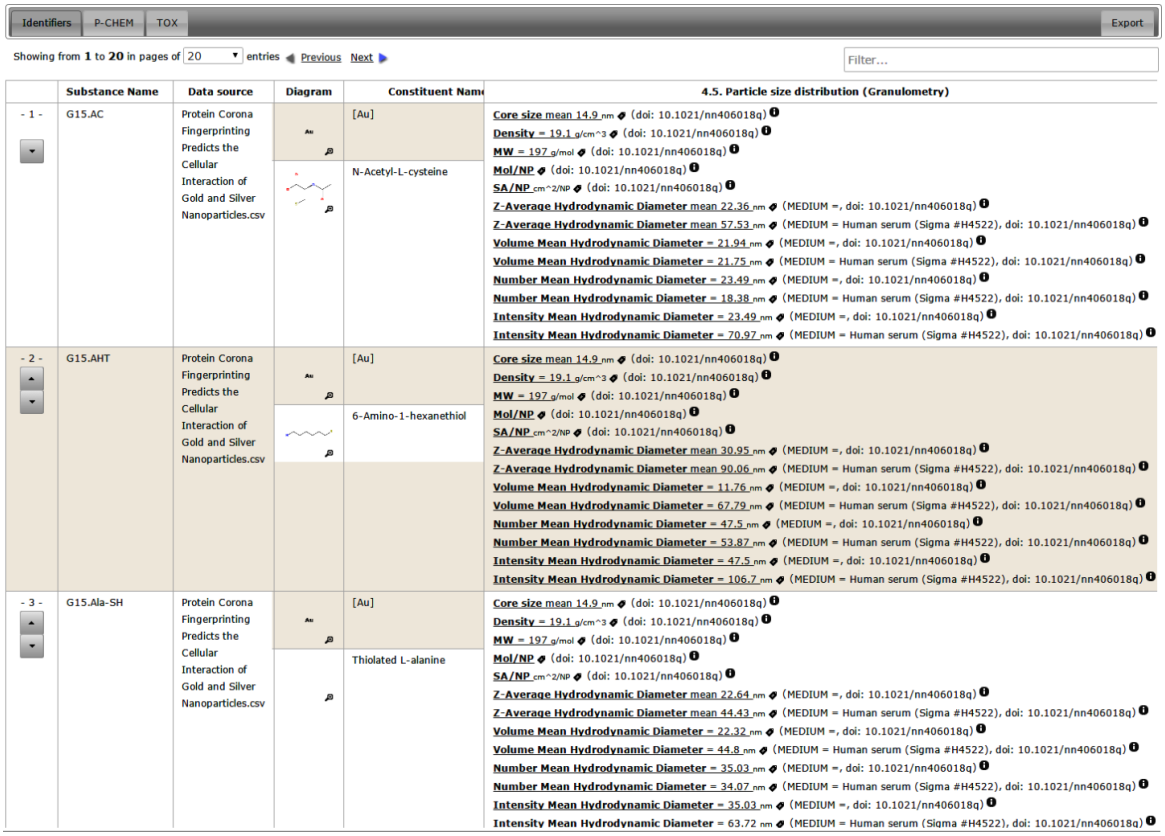


Figure 10: Screenshot of the bundle view with the Protein Corona data set. In addition to the Substance API, which allows one to retrieve study data for a single NM as in Figure 5, the bundle API provides efficient means to retrieve information about a set of NMs.

Institutet and Maastricht University. The database is developed as a wiki using the Semantic MediaWiki platform, running on a virtual machine using the VirtualBox software. The wiki contains physicochemical properties and toxicological data for more than three hundred nanomaterials: more than two hundred metal oxides, 80 carbon nanotubes, and a few metal and alloy particles. All nanomaterials originate from data in 34 papers, identified by Digital Object Identifier (DOI), from twenty scientific journals. Because the amount of physicochemical detail differs from one paper to another, each material is characterized with different measured characteristics. Each measurement may have a single value (median or average, though this is not always specified), a minimum and maximum value, or a single value and a standard deviation. Biological measurements are linked to assays (such as cytotoxicity, cell growth, cell viability, genotoxicity, and oxidative stress), endpoints measured on that assay (e.g., ROS concentration, GI50, percentage viable cells), and cell line information, though not consistently.

Importing the data into eNanoMapper takes advantage of NanoWiki using Semantic MediaWiki and its template framework: all data relevant to NanoQSAR can be retrieved from the wiki as RDF, in the form of a RDF/XML data dump [27] (in addition to the common MediaWiki XML and SQL dumps of the wiki content).

ModNanoTox

The ModNanoTox EU FP7 project (http://www.birmingham.ac.uk/generic/modnanotox/index.aspx) has produced a survey and selection of relevant physicochemical properties to use towards building a range of descriptors of engineered nanoparticles (mainly metal-based) and their potential toxicity. This dataset nicely demonstrates the complexity of the nanosafety domain. The ModNanoTox database provides physicochemical descriptors and toxic activities of nanoparticles from several studies. The database version from August 2013 includes 86 assays with more than 100 different endpoints affecting 45 species.

Unfortunately, only a few nanoparticles (usually fewer than three) have been tested for each endpoint. Physicochemical descriptors for the characterisation of nanoparticles are incom-

plete as well (about 75% missing values). The two most comprehensive species in the dataset are *Daphnia magna* (water flea) and *Danio rerio* (zebrafish), with 34 and 14 assays each. The best represented endpoint for Daphnia is "Mortality", and we were able to extract about forty "LC50" and sixty "% survival" data entries. In both cases the number of measured nanoparticle properties was very low. Most studies report only two to four different nanoparticle properties (descriptors) and the descriptor types are very inconsistent (overall 36 different descriptors, which results in very sparse matrices with a high number of missing values).

The ModNanoTox data import is currently being tested and is not yet available online. The ModNanoTox data set was provided as a MSExcel spreadsheet file. It consists of four sheets describing, respectively, (i) investigation study details, (ii) particle details and physicochemical properties, (iii) assay protocol description and (iv) assay measurement outcomes. The information in all sheets is organized as a sequence of dynamic blocks of data, each one containing a variable number of rows. The configurable spreadsheet parser described in the "Data Import" section supports the recognition of blocks and the synchronization between blocks within the four sheets. The next step is to divide the data in each block into groups and subgroups and match them across the sheets. This last operation is implemented by a dedicated command line application, built on top of the configurable data parser and allowing parsing of the entire ModNanoTox complex organisation into the internal eNanoMapper data model.

Protein Corona

The demonstration data set, extracted from [28], focuses on the biological identity of ENMs. The authors used the composition of the protein corona "fingerprint" to predict the cell association of a 105-member library of surface-modified gold nanoparticles (see Figure 3). 785 distinct serum proteins were identified by LC-MS/MS, from which 129 were suitable for relative quantification. The fingerprint of serum proteins was defined by the relative abundance of each protein on a nanoparticle formulation. The value of individual proteins within the serum protein fingerprint for predicting cell association was explored by the authors by developing a series of log-linear models that model the influence of the relative abundance of each adsorbed serum protein on net cell association. Among the factors in play in protein corona, biological interaction was chosen to be represented by cell association because of its relevance to biodistribution, inflammatory response potential, and in vivo toxicity. The eNanoMapper prototype described in this paper is able to capture this protein corona, and modelling approaches were extracted from these data for statistical analysis.

Data quality considerations

While there is a common agreement on the importance of data curation, there is no well established common understanding of how it should be performed. Approaches range from simple data cleaning to the entire spectrum of data-related activities including evaluation, on-going data management, and added value provisioning through analytic tools. The focus of this publication is on the data management system, allowing for a unified approach to storage and querying of NM related data. Using the data for modelling and being able to write the prediction results back is only one of the possible ways to add value. Future developments may include providing support for emerging paradigms such as Adverse Outcome Pathways [29], categorization strategies via decision trees [30] and principal components [31]. We intentionally do not discuss data evaluation and clean-up for the following reasons. Firstly, at present we are not aware of universally adopted criteria for evaluation of NM data, although there are a number of related activities in the EU NanoSafety Cluster projects and worldwide, as well as specific sets of rules implemented in existing databases such as the NanoMaterial Registry (https://www.nanomaterialregistry.org/about/WhatIsCuratedData.aspx). In regulatory toxicology the Klimisch codes [32] are the accepted approach, enforced in Europe by the relevant guidance [33] and the IUCLID database. They provide definitions and support for annotating the data records by relevance, reliability and adequacy. Some of the criteria necessarily overlap with rules defined elsewhere (availability of the raw data, adequate description of the study, protocols, parameters, purities/impurities and the origin of the test substances; proof of ability of the lab to do the study). Klimisch codes (or scores) define four reliability categories (1 = reliable without restrictions, 2 = reliable with restrictions, 3 = not reliable, 4 = not assignable), where score 1 or 2 can only be assigned if the data are generated through accepted standard methods (e.g., OECD guidelines or equivalent national or international standards) and according to Good Laboratory Practice (GLP). In practice, very few of the publicly available NM datasets can be assigned reliability code 1 or 2, due to the lack of standard or validated protocols, deviations, or just an absence of details. The criteria for experimental protocol validation are out of scope for this paper as well as for the eNanoMapper project. However, the database and import templates are designed to require that the test protocol be specified for every data entry. Secondly, as the goal is to support data originating from different sources and typically already having undergone some kind of evaluation and assigned relevant labels, the most appropriate way is to import the data as it is and keep the original quality labels. For example the OECD HT templates do include fields for Klimisch scores and the eNanoMapper database does store these scores, as is shown in the JSON serialization. The data generated or gath-

ered from the literature by EU NanoSafety Cluster projects have already been evaluated as part of these project activities, and we intend to keep this information, where it is available. Once the data are converted into the common data model, rules checking the presence or absence of raw data, protocols, deviations, and parameters can be applied automatically, which is a more efficient approach than checking these rules manually before import. The ontology annotation might help to overcome some of the challenges, such as different evaluation criteria and different terminology for the quality labels. In cases where automatic tools fail, working closely with data providers to improve the quality and gain common understanding of the data is necessary. This approach is also in line with the intention "not to exclude automatically the unreliable data from further considerations" [32] and that "there is unlikely to be a single out-of-the-box solution that can be applied to the problem of data curation. Instead, an approach that emphasizes engagement with researchers and dialogue around identifying or building the appropriate tools for a particular project is likely to be the most productive" [34].

Visualisation

User interface

The following screenshots illustrate the eNanoMapper prototype database user interface, as implemented by AMBIT web services [14], with the help of JavaScript widgets consuming the REST API. The screenshots in Figure 11 and Figure 12 illustrate the data model support and the visualisation of experimental data, consisting of a variety of endpoints, experimental conditions and multiple endpoints values. The origin of the data is the ECHA dissemination site [35], and the data were manually entered into a local IUCLID5 instance, exported into IUCLID5 .i5z file and imported into the database.

The API is tightly integrated with a chemical structure and chemical similarity search (implementation details previously published in [14,36,37]). Chemical similarity is a pivotal concept in cheminformatics, encompassing a variety of computational methods quantifying the extent to which two chemical structures resemble each other. Apart from the "intuitive notion" of chemical similarity typically acquired during chemistry education, the computational methods vary from structure-based (2D, 3D), descriptor- and field-based approaches [38]. Chemical similarity evaluation requires two components, namely a numerical representation of the chemical structure and a measure allowing for comparing two such representations. The representations derived from the molecular graph are by far the most common (e.g., hashed fingerprints and various flavours of substructure keys) and the Tanimoto coefficient is the most popular similarity measure. The chemical similarity values usually range from zero (no similarity) to one (identical structures). Similarity searching (along with chemical substructure searching) in chemical databases is considered standard functionality and is nowadays offered by all state-of-the-art chemical databases and cheminformatics tools [39].

The chemical similarity search in the eNanoMapper prototype database enables querying by a chemical structure of a NM

IUC Substance | Composition | Eco Tox (8) | Tox (13) | P-Chem (15) | Env Fate (1)

Multi-Walled Carbon Nanotubes (MWCNT), Synthetic Graphite In Tubular Shape

4.1 Appearance (1)

4.5 Particle size distribution (Granulometry) (10)

Distribution type	Endpoint	Value	Reference	Guideline	Owner
imaging technique	PARTICLE SIZE.D10	120 nm	2010	Transmission Electron Microscopy (TEM)	ECHA Dossier
	PARTICLE SIZE.D50	360 nm			
	PARTICLE SIZE.D90	980 nm			
	PARTICLE SIZE	ca. 380 nm			
volumetric distribution	PARTICLE SIZE.D10	331 µm	2010	Laser diffraction	ECHA Dossier
	PARTICLE SIZE.D50	569 µm			
	PARTICLE SIZE.D90	945 µm			
imaging technique	PARTICLE SIZE.D10	5 nm	2010	Transmission Electron Microscopy (TEM)	ECHA Dossier
	PARTICLE SIZE.D50	10 nm			
	PARTICLE SIZE.D90	18 nm			
volumetric distribution	PARTICLE SIZE.D10	597 µm	2010	Digital microscopy	ECHA Dossier
	PARTICLE SIZE.D50	842 µm			
	PARTICLE SIZE.D90	995 µm			
imaging technique	MASS MEDIAN DIAMETER	[3, 20) nm	2011	Transmission Electron Microscopy (TEM)	ECHA Dossier
	PARTICLE SIZE.D90	12.7 nm			
imaging technique	PARTICLE SIZE.Tube thickness	[5, 20] nm	2008	- Transmission Electron Microscopy (TEM) - EDXS Energy Dispersive X-Ray Spectroscopy (Z>=6)	ECHA Dossier

Figure 11: Physicochemical data for multi-walled carbon nanotubes. The screenshot illustrates the data model and UI support for size distribution (through percentiles D10, D50, D90), multiple endpoints per measurement (Mass median diameter and particle size), and multiple experiments using different protocols.

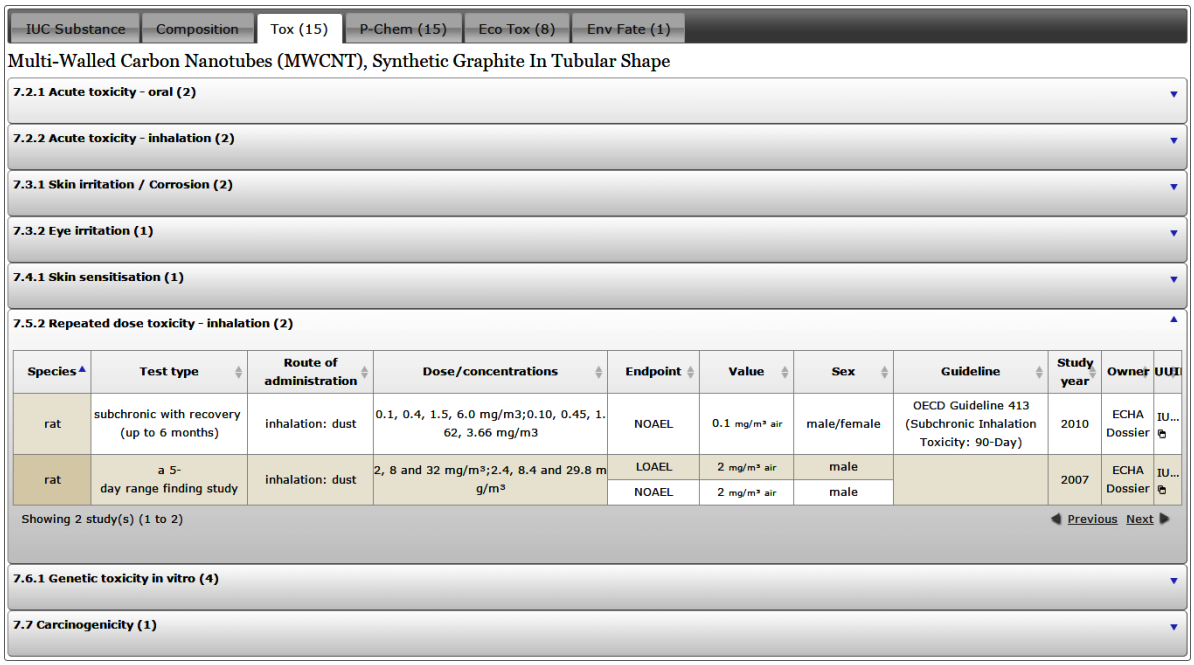


Figure 12: Toxicity data for multi-walled carbon nanotubes. The repeated dose toxicity (inhalation) is shown in the expanded row, illustrating support for multiple endpoints (LOAEL, NOAEL) and test types.

component and highlighting the results as a core, coating or functionalisation component (Figure 13). The reason for the wide adoption of the similarity approach is the assumption of the "similarity property principle" or "neighbourhood behaviour", namely that "similar compounds should have similar properties". This principle puts the chemical similarity at the core of methods and tools supporting property prediction, structure–activity relationship, chemical database screening, virtual screening in drug design, and diversity selection. The similarity assessment based on structure analogy is the basis of read across and chemical grouping. However, there is a common understanding that the most difficult part in read across is "rationalising the similarity". Violations of the "similarity property principle" exist due to a variety of reasons [38], and nowadays the existence of "activity cliffs" (small changes in the chemical structure leading to a drastic change in the biochemical activity) is well known. A recent review by Maggiora [40] outlines the methods used as well as the pros and cons of using the molecular similarity framework in medicinal chemistry. In the context of nanosafety assessment there is not yet a standardized approach for NM similarity, however a number of attempts for NM grouping and read across have been published recently [41,42].

Apart from enabling searching by well-defined chemical structures, the chemical similarity and substructure search enhances the data exploration capabilities of the system (e.g., finding nanoparticles with similar coatings). The data exploration is also supported by REST API calls retrieving data summaries (e.g., number of zeta potential entries) and endpoint prefix queries, allowing for building dashboards and supporting autocompletion fields. Therefore a suitable user interface can be built to allow data search without requiring a priori knowledge of the database content and field names (Figure 14). The search and results retrieval API can be used for many applications, one of which being NanoQSAR modelling. Future extensions, currently under development, include free text search with query expansion based on the eNanomapper ontology and annotated database entries, with an indication of the relevance of the hits. Visual summaries can be integrated in the eNanoMapper web interface, as well as used as widgets in external web sites as demonstrated in the following section.

JavaScript visual summaries

To further demonstrate the use of the eNanoMapper API for visualisation we have developed a series of example web pages (HTML) using the JavaScript d3.js library [43]. This library has been used for a wide variety of visualisations (as can be seen on their website), and here used to summarize some of the data in the database. To simplify the interaction with the eNanoMapper API a JavaScript client library, ambit.js, was written to allow asynchronous calls to the web service [44]. However, because the d3.js methods require the data to be provided in a specific JavaScript object, the JSON returned by the API has to be

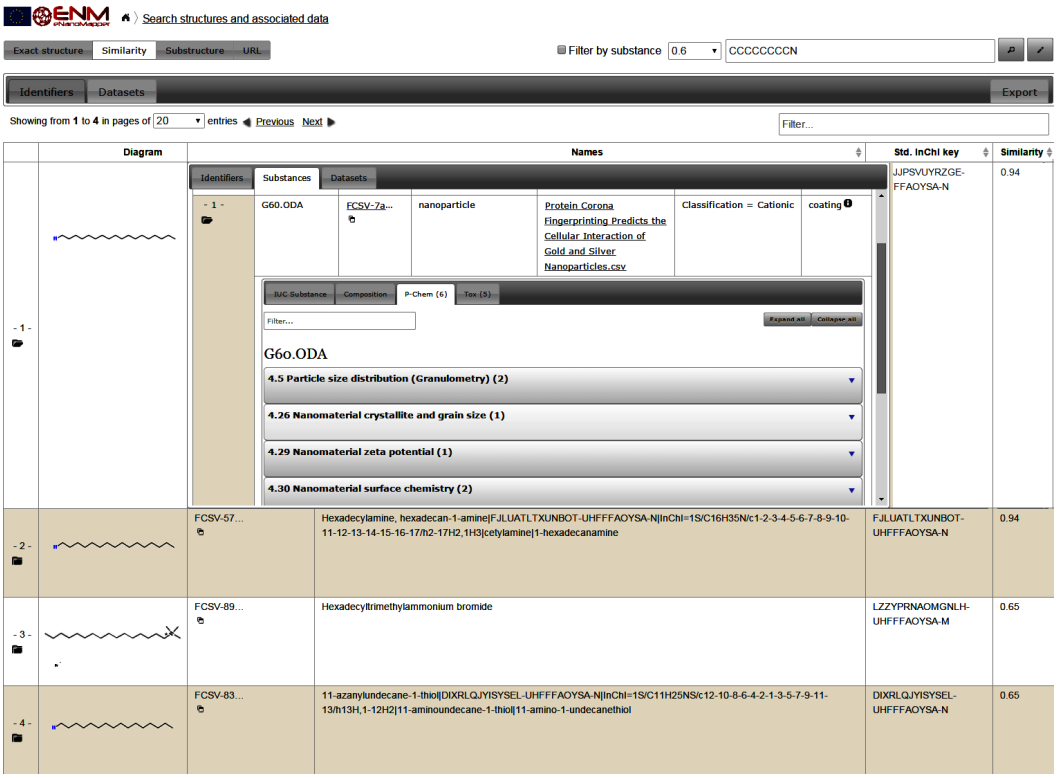


Figure 13: Screenshot showing the results of a chemical similarity query (octyl amine, SMILES CCCCCCCCN) with a similarity threshold Tanimoto coefficient = 0.6. The results include octadecylamine (similarity 0.94), hexadecylamine (similarity 0.94), hexadecyltrimethylammonium bromide (similarity 0.65), 11-amino-1-undecanethiol (similarity 0.65), all used as coating of silver and gold nanoparticles in the protein corona dataset. The first row shows expanded view with details of the NM.

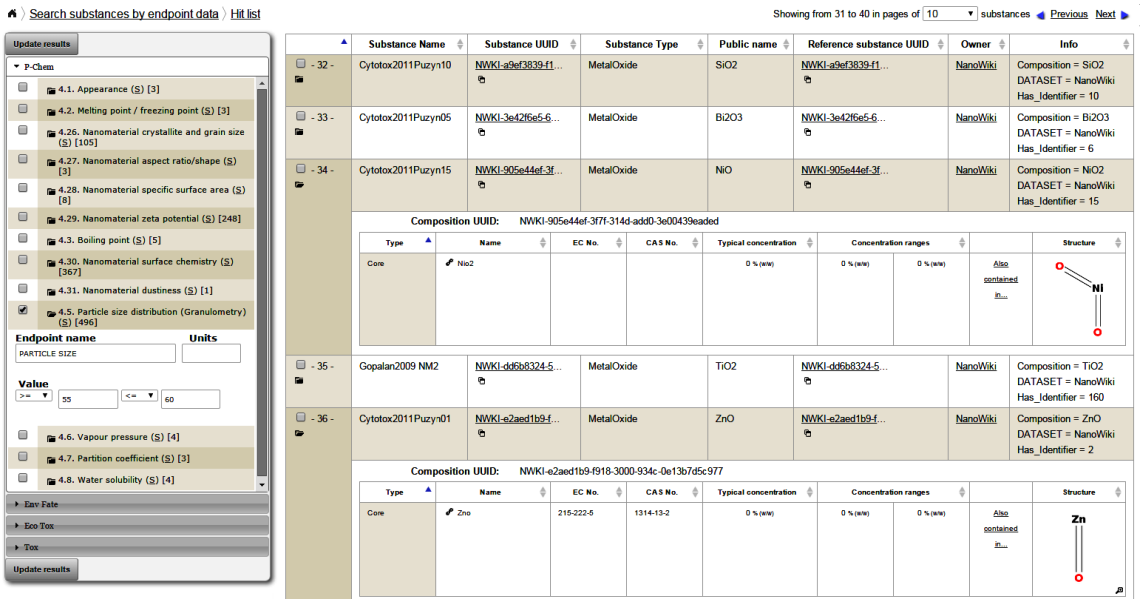


Figure 14: Screenshot showing query results in the NanoWiki data set for particle sizes between 50 and 60 nm. The widget at the left side represents an overview of all experimental data in the system, organized in four groups of physicochemical, environmental, ecotoxicological and toxicity sections. Each section lists available endpoints and the number of available data entries. The text boxes support auto-completion, i.e., the available values will be displayed and can be selected by either pressing an arrow-down button (to list all available values) or by entering the first letters of a possible value.

converted to a structure understood by the d3.js code. The sources of the examples presented here are available from the ambit.js project page at http://github.com/enanomapper/ambit.js/. The source code and documentation of the ambit.js library are available at the same location.

The first example shows a summary of the number of materials in the database, sorted by the dataset they originate from (NanoWiki, protein corona, and others), as shown in Figure 15. Here, a single API call was sufficient and the data needed for the pie chart were extracted from the JSON returned by this call. Because of the asynchronous nature of the client–server interaction, a callback function has to be defined. The combination of the callback function (the full implementation is left out for brevity but is available from the ambit.js repository as with Example 2) and the actual API call is done by the ambit.js code given in Figure 16.

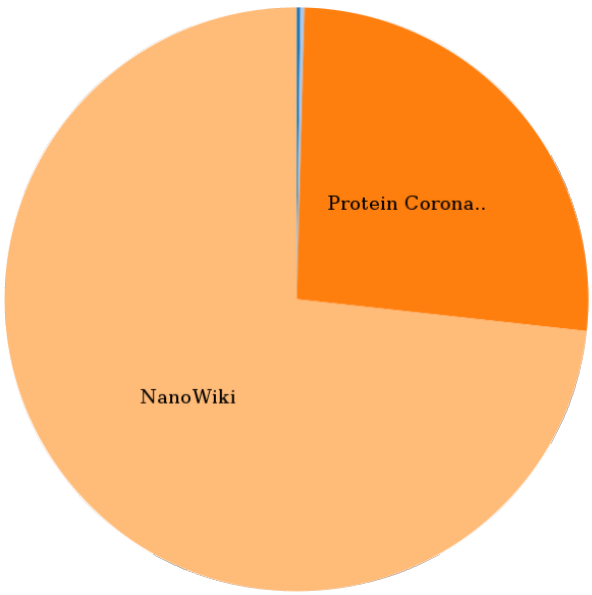


Figure 15: Pie chart created with d3.js and ambit.js in a web page showing that the NanoWiki and Protein Corona datasets contain the most nanomaterials in the database.

```
var callback = function(success, status, response){
    // here the data is extracted from the JSON in the
    // response variable, and visualized with d3.js
}
var substances = new Ambit.Substance(
   "https://apps.ideaconsult.net/enanomapper"
);
substances.list(callback);
```

Figure 16: API call in ambit.js code.

The second example shows a histogram of nanomaterial sizes (size reported, or average if a size range was given). Because the list of materials does not provide the size information, the callback function of the "Ambit.Substance.list()" call has to make a subsequent call for each material in the list. The example web page keeps track of the number of remaining calls to this second "Ambit.Substance.info()" API call in a second callback function which also aggregates the material sizes in a global variable. Therefore, the total number of API calls equals the number of materials plus one. When the second callback function notices that there are no further calls to be returned, it calls a plot function that takes the aggregated list of sizes and visualizes it with d3.js, resulting in Figure 17.

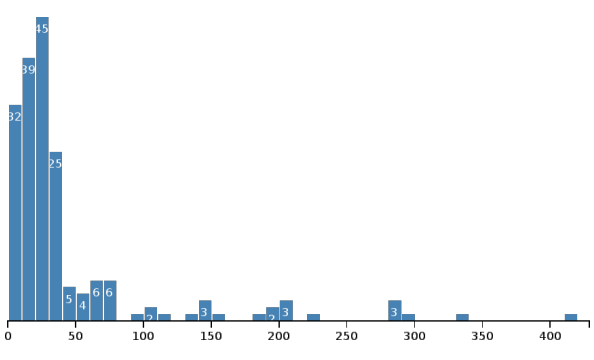


Figure 17: Histogram of nanomaterial sizes created with d3.js and ambit.js.

A variation of the second example shows a scatter plot of the zeta potential values against nanomaterial sizes. Here, the same approach is used and the bits of information are aggregated in a global variable. The results are shown in Figure 18. The red colour of the dots was chosen arbitrarily, but could reflect another feature, possibly the data sources as shown in the first example.

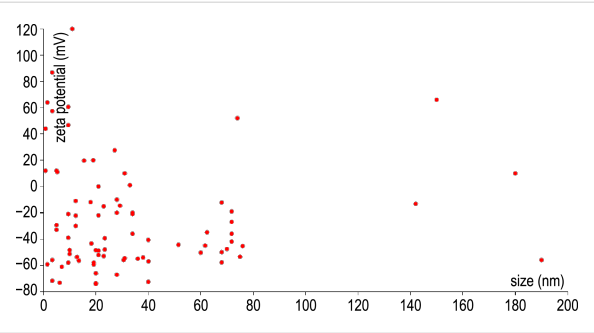


Figure 18: Scatter plot of nanomaterial zeta potentials against the nanomaterial sizes, also created with d3.js and ambit.js.

Modelling

The OpenTox API implementations contain all major statistical and machine learning (ML) algorithms required for the development of regression, classification or clustering models, as well

as cheminformatics algorithms, such as structure optimisation and descriptor calculation. A ML algorithm is made available as a web resource and a model is created by sending a HTTP POST to the algorithm URI, with specified dataset URI and modelling parameters, where relevant. The model is again a web resource, and another HTTP POST to the model URI can be used to launch prediction of a specified dataset of chemical structures or materials. However, the OpenTox algorithm and modelling API is centred on chemical structures, and requires clean datasets in a specific form. On the other hand, the eNanoMapper prototype database is explicitly designed to handle all peculiarities of experimental data, including replicates, range and error values. Therefore, a tool, converting the experimental data into a form suitable for modelling algorithms, is required.

This section describes the approach taken by eNanoMapper, namely the Jaqpot web application, the API documentation of which can be found at http://app.jaqpot.org:8080/jaqpot/swagger, providing one possible solution for this challenge. Jaqpot is a web application that currently supports data preprocessing, statistical, data mining and machine learning algorithms and methods for defining the applicability domain of a predictive model. A screenshot of the Jaqpot web services is presented in Figure 19. Jaqpot provides asynchronous execution of tasks submitted by users, authentication, authorisation and accounting mechanisms powered by OpenAM. It was originally developed during OpenTox [23] and is an open-source project, written in Java and licensed with the GNU GPL v3 licence. Jaqpot Quattro is an extension, developed within eNanoMapper and featuring improved efficiency and additional functionality. Jaqpot Quattro is part of the eNanoMapper framework and communicates with other web services in the framework via the common REST API described above. The source code is publicly available from https://github.com/KinkyDesign/JaqpotQuattro. The main features of Jaqpot Quattro are presented next.

Producing datasets from bundles

The Jaqpot algorithm services require input data in a standardized format in order to generate a predictive model and raw experimental data cannot be used directly for modelling purposes. The experimental data are, more often than not, heterogeneous by nature and properly structuring these is not a trivial task. To this end, a web service acting as a link between experimental data and data for modelling was introduced, which will be hereafter referred to as the "conjoiner service". This service performs the task of mapping the experimental data into a modelling-friendly format and producing standardized datasets as specified in the OpenTox API. One can initiate a conjoiner service operation by specifying a bundle URI. A bundle (see Figure 9) is an eNanoMapper resource that acts as an assortment of experimental effects, images and molecular structures, for nanomaterials, and the job of the conjoiner service is to combine all that disparate data into a dataset suitable to be fed to an algorithm service. Concerning experimental data, multiple individual measurements, interval-valued measurements (lower and upper values), or values accompa-

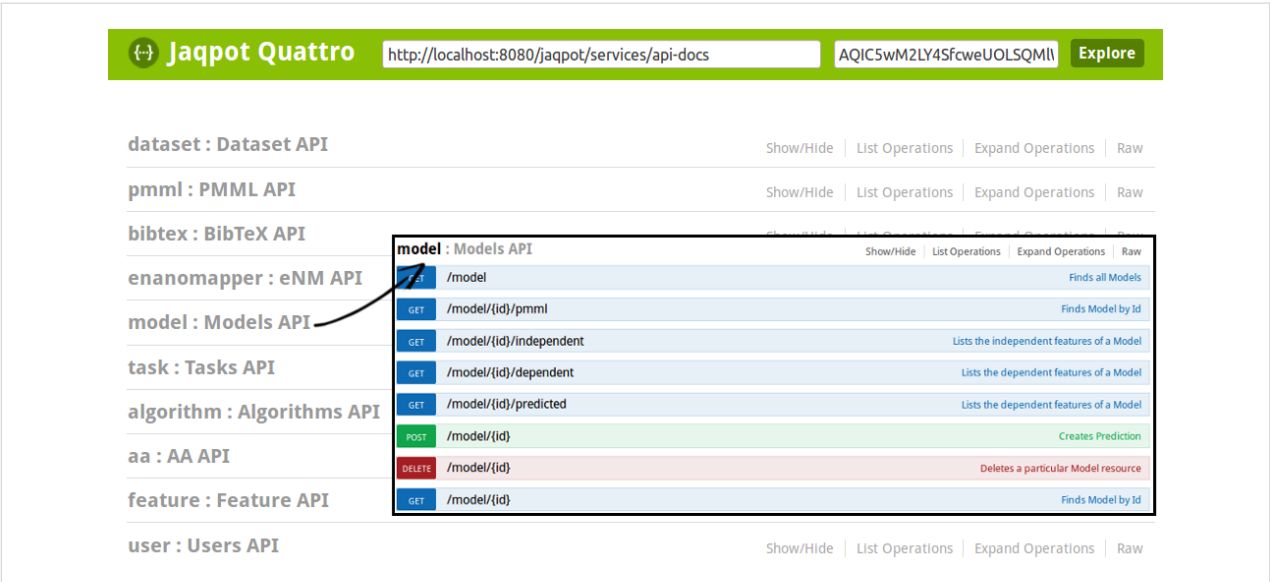


Figure 19: Screenshot of the Jaqpot Quattro modelling web services API, compatible with the eNanoMapper API. A list of REST endpoints is presented to the end user. These correspond to the main entities/resources of eNanoMapper: datasets, models, algorithms, BibTeX entities, asynchronous tasks and more. The user can click on any of these to get a list of the available operations related to each entity. In the inset of this figure we see the list of model-related operations. For more information consult the OpenTox Model API http://opentox.org/dev/apis/api-1.2/Model.

nied by a standard measurement error, may be included for the same endpoint in a bundle, and need to be aggregated into a single value. This is currently done by taking the average value of all experimental measurements having excluded outliers identified by a Dixon's q-test [45], but different aggregation procedures will be implemented in the future based on more elaborate outlier detection criteria and rejection/aggregation schemata [46,47]. The client will then be able to customise this procedure. The overall procedure is illustrated in Figure 20.

Preprocessing

Scaling, normalization and handling of missing values are important preprocessing steps for efficient model training, as most algorithms are sensitive to nonscaled data [48] such as SVM [49]. All these preprocessing steps are offered as options when a client calls a Jaqpot Quattro algorithm service. Furthermore, Jaqpot Quattro makes use of the Predictive Model Markup Language (PMML) file format that allows clients to define a "data dictionary" and a "transformations dictionary", by providing the URI of a PMML document [50,51]. The data dictionary selects a number of features out of the original dataset that will be provided as inputs to the modelling algorithm, while the transformation dictionary defines mathematical formulae to be applied on the selected features. The predictive model will be then trained using the transformed features as input.

PMML, which has been developed for enabling models to be portable across different computational platforms, is a well-adopted standard in the machine learning and QSAR community. PMML documents are essentially XML documents that contain all necessary information to reproduce a model including the definition of input parameters, targets (predicted properties), preprocessing steps (e.g., scaling, normalization, transformation of inputs), and the main model (e.g., MLR, SVM). The PMML format of the produced NanoQSAR models is also supported by Jaqpot Quattro algorithm services.

An example of a PMML document that selects two properties and applies subtraction, division and absolute value operations is given in Figure 21.

Notice that the "DataDictionary" block defines the required input features. The trained model, however, needs to transform these features into the *internal variables* "zp_ch", "zp_rel", "zp_synth_mag" and "zp_serum_mag" as specified in the "TransformationDictionary" of the PMML document.

API for dynamic algorithm integration

The Jaqpot Protocol of Data Interchange, in short JPDI, is a new feature of the Jaqpot Quattro web services that allows developers of machine learning algorithms to integrate their implementations in the framework. This integration requires

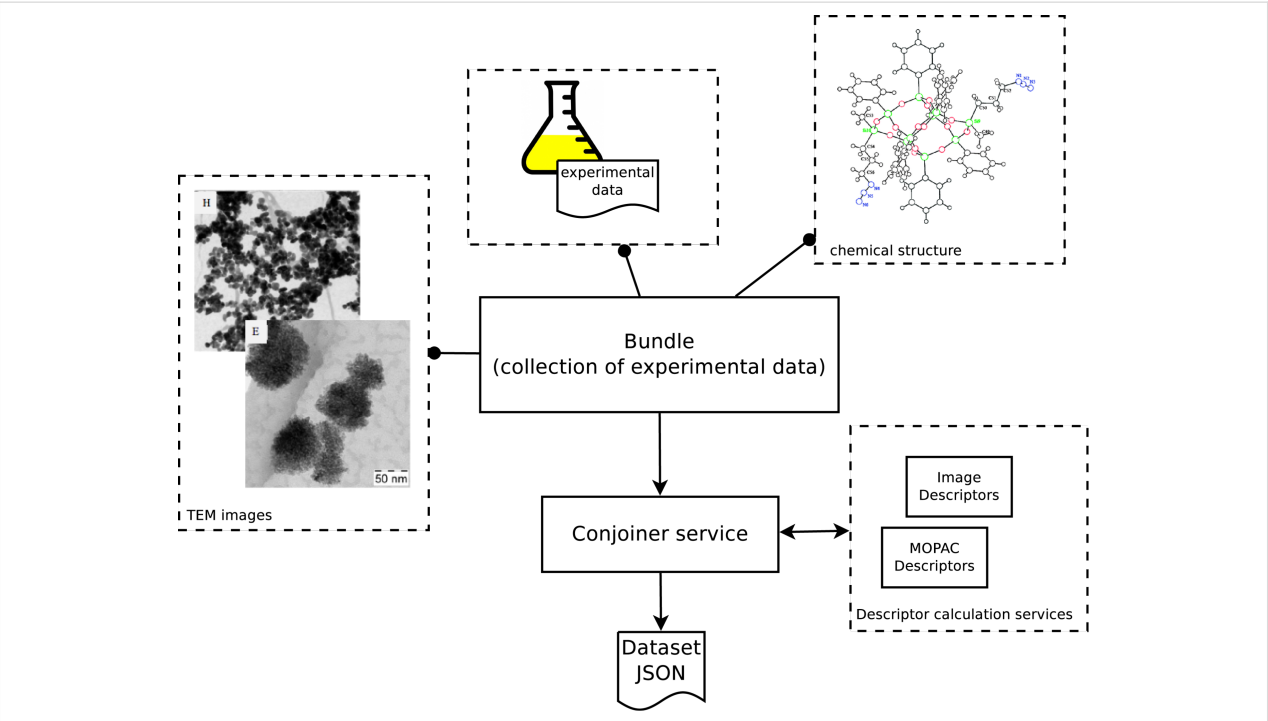


Figure 20: Conjoiner API: modelling-oriented information can be extracted from bundles of experimental data. Data as heterogeneous as chemical structures, raw experimental measurements, spectra and microscopy images can be combined by the conjoiner service to produce a dataset for modelling purposes.

```
<PMML version="4.0"
   xsi:schemaLocation="http://www.dmg.org/PMML-4_0
   http://www.dmg.org/v4-0/pmml-4-0.xsd"
   xmlns="http://www.dmg.org/PMML-4_0"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
<DataDictionary numberOfFields="2" >
<DataField name="property/1" optype="continuous" dataType="double" />
<DataField name="property/2" optype="continuous" dataType="double" />
</DataDictionary>
<TransformationDictionary>
   <DerivedField dataType="double" name="zp_ch" optype="categorical">
     <Apply function="-">
       <FieldRef field="property/1"/>
       <FieldRef field="property/2"/>
     </Apply>
   </DerivedField>
   <DerivedField dataType="double" name="zp_rel" optype="categorical">
     <Apply function="/">
       <FieldRef field="property/1"/>
       <FieldRef field="property/2"/>
     </Apply>
   </DerivedField>
   <DerivedField dataType="double" name="zp_synth_mag" optype="categorical">
     <Apply function="abs">
       <FieldRef field="property/1"/>
     </Apply>
   </DerivedField>
   <DerivedField dataType="double" name="zp_serum_mag" optype="categorical">
     <Apply function="abs">
       <FieldRef field="property/2"/>
     </Apply>
   </DerivedField>
</TransformationDictionary>
</PMML>
```

Figure 21: Example of a PMML document.

little engagement with intricate software development and allows algorithm developers to outsource their implementations and make them available to the nanomaterials design community through the eNanoMapper framework.

The communication between eNanoMapper services and third-party JPDI services is carried out by exchanging JSON documents that contain no more information than a modelling service needs to train a predictive model, calculate descriptors, perform a prediction, evaluate the domain of applicability of a model, or perform other tasks. This is well illustrated in Figure 22.

Once a developer (possibly third-party) has prepared a JPDI-compliant web service, they need to register it to the eNanoMapper framework and specify (i) the name of the algorithm, (ii) metadata for the algorithm, such as a description, tags, copyright notice, bibliographic references and any other metadata supported by the Dublin core ontology (http://dublincore.org/) and/or the OpenTox ontology [52], (iii) the URI of their implementation to be used as an endpoint for training, (iv) the corresponding URI for the prediction web service, (v) an ontological characterization of the algorithm according to the OpenTox Algorithms ontology (e.g., “ot:Regression” or “ot:Classification”, or “ot:Clustering” (http://www.opentox.org/

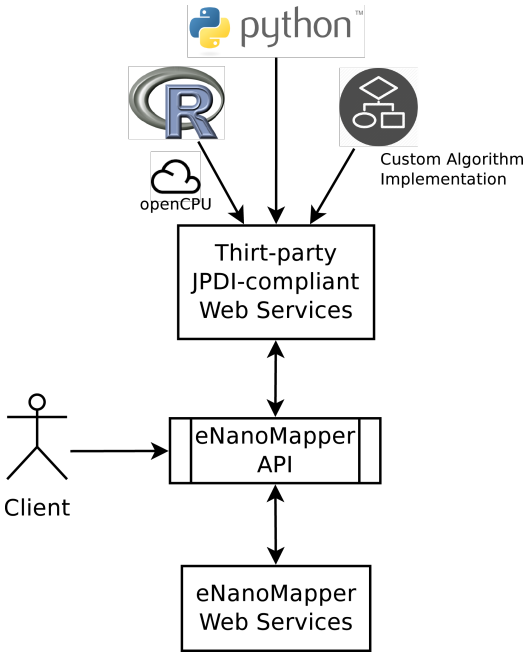


Figure 22: JPDI-compliant web services can be seamlessly incorporated into the eNanoMapper framework. The client communicates with eNanoMapper services through the eNanoMapper API while certain operations such as model training are delegated to JPDI-compliant services.

dev/apis/api-1.1/Algorithms), and (vi) a set of tuning parameter definitions, optional or mandatory, that the client may provide during training. The algorithm is then registered by POSTing a JSON document containing all this information to "/algorithm". Once registered, the algorithm acquires a URI, and is exposed as a web service, that can be consumed. Algorithms can be registered (POST), removed (DELETE) and modified (PATCH) using the Algorithm API presented in Figure 23, which extends the OpenTox Algorithm API (http://opentox.org/dev/apis/api-1.2/Algorithm).

A JPDI request for training is presented in Figure 24. This request is issued by an algorithm web service of eNanoMapper to a JPDI-compliant web service.

Notice the three most important components in a training request, which are the "dataset", the "prediction feature" and the "tuning parameters" of the algorithm. Once the model is trained, the JPDI service will return it to the caller in JSON format in which the actual model is encoded. Figure 25 gives an example:

```
{
 "dataset": {
   "dataEntry": [
     {
      "compound ": {
       "URI": "http://some.server.org/substance/1"
                    },
      "values": {
       "http://some.server.org/property/1": 0.268,
       "http://some.server.org/property/2": 0.667,
           ...
                 }
       },
     {
      "compound": {
       "URI": "http://some.server.org/substance/2"
                    },
      "values": {
       "http://some.server.org/property/1": 0.115,
       "http://some.server.org/property/2": 0.759,
           ...
                 }
                    },
           ...
     ]
    },
   "predictionFeature": "http://some.server.org/feature/1",
   "parameters": {
       "theta": 49,
       "mvh": 1.0
    }
}
```

Figure 24: A JPDI request for training.

Notice that the JPDI web service may select only some of the features of the initial dataset, which are defined in the PMML. Then, the JPDI service requires that a dataset containing these features be posted back to it, i.e., a JPDI service in order to perform predictions requires (i) the model it has previously produced and (ii) a dataset containing values for the features it has selected.

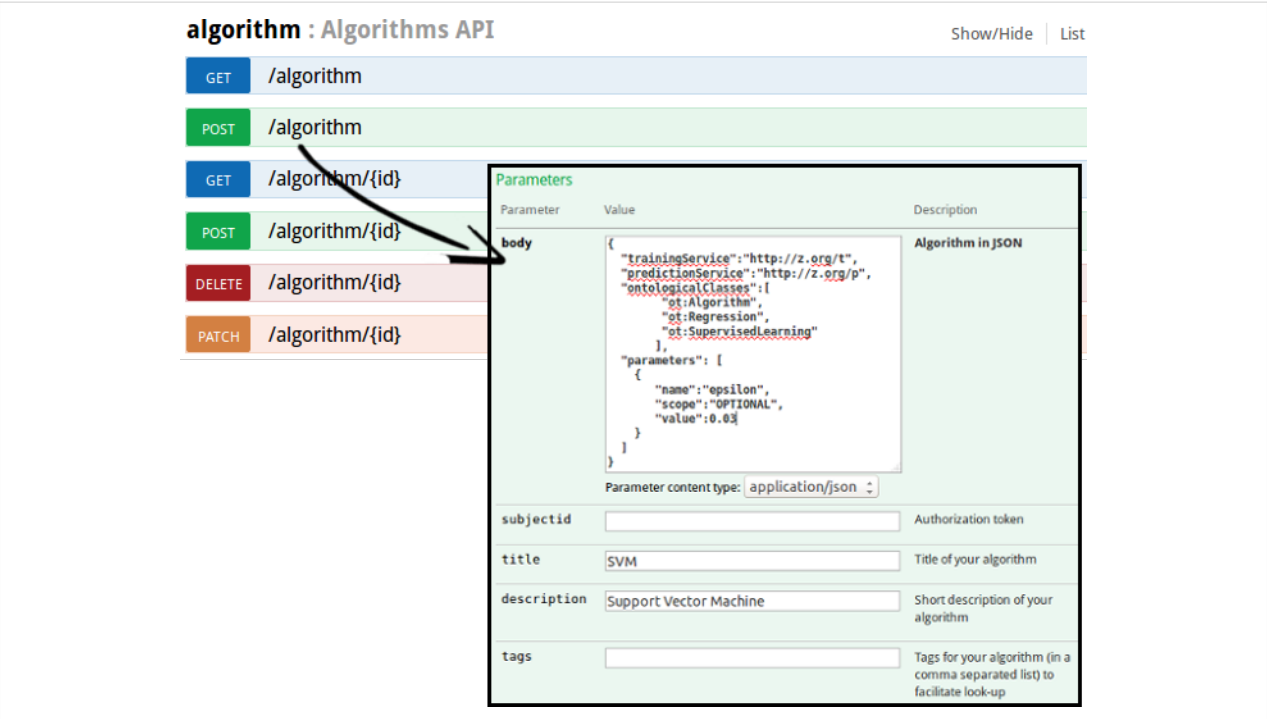


Figure 23: Algorithm API that allows to consume as well as register new algorithms (following the JPDI specification). Clients can use this API to (i) GET a list of all algorithms, (ii) register a new algorithm, (iii) GET the representation of an existing algorithm, (iv) Use an algorithm, (v) Delete an existing algorithm or (vi) use the HTTP method PATCH to modify an algorithm resource.

```
{
  "rawModel": "<Raw model (encoded)>",
  "pmmlModel": "<PMML-XML>",
  "additionalInfo" : "<Extra information the algorithm
       service needs saved with the model>",
  "independentFeatures": [
    "http://some.server.org/property/1",
    "http://some.server.org/property/2"
   ]
}
```

Figure 25: A model returned by JPDI service in JSON format.

Upon training, the model returned to the caller is stored as-is by the called service and will be returned back to the JPDI-compliant service when the client requests a prediction. This way, as already mentioned, the JPDI service providers do not need to maintain a database while the eNanoMapper services do not need to know how the third-party services perform computations.

Likewise, when Jaqpot Quattro needs to consume a JPDI web service to perform predictions, it POSTs to it a JSON document with (i) the input dataset containing substances and (ii) the model that was previously created by the JPDI service. An example of JSON prediction request is shown in Figure 26.

Integration with third party services

The JDPI protocol allows one to dynamically and seamlessly incorporate any custom algorithmic implementation into eNanoMapper and without any need for resource management (i.e., the algorithm providers do not need to maintain a database system). The protocol specifies the form of data exchange between eNanoMapper services and third party algorithm web service implementations. The eNanoMapper framework already provides wrappers for WEKA [53] and the R language [54]. Integration with R is made possible through the OpenCPU (https://www.opencpu.org/) system, which defines a HTTP API for embedded scientific computing based on R although this approach could easily be generalized to other computational back ends [55]. OpenCPU acts as a wrapper to R that is readily able to expose R functions as RESTful HTTP resources. The OpenCPU server takes advantage of multi-processing in the Apache2 web server to handle concurrency. This implementation uses forks of the R process to serve concurrent requests immediately with little performance overhead. By doing so it enables access to those functions on simple HTTP calls converting R from a stand-alone application to a web service. R (http://www.r-project.org/) has become the most popular language for computational statistics, visualization and data science, in both academia and industry [56]. One of the most important benefits for R users is cost-free, easy access to the frontline of methods in predictive modelling and statistics that are produced and are under continuous review from leading data science researchers [57]. In Bioinformatics, the Bioconductor R branch (http://www.bioconductor.org/), provides open source tools for high-throughput omic data analysis. Bioconductor users enjoy access to a wide array of statistical and graphical methods for genomic data analysis and makes it much easier to incorporate biological metadata in genomic data analysis, e.g., PubMed literature data (http://

```
{
 "dataset": {
  "datasetURI": ["http://some.server.org/dataset/1"],
  "dataEntry": [{
    "compound": {
      "URI": "http://some.server.org/substance/1"
                },
       "values": {
         "http://some.server.org/property/1": 0.268,
         "http://some.server.org/property/2": 0.667,
                    ...
                 }
                }, {
    "compound": {
      "URI": "http://some.server.org/substance/2"
                },
       "values": {
         "http://some.server.org/property/1": 0.115,
         "http://some.server.org/property/2": 0.759,
                    ...
                }
            },
            ...
        ],
  "rawModel": ["<Encoded raw model>"],
  "additionalInfo": ["<other info the JPDI algorithm service needs>"]
    }
}
```

Figure 26: An example of a JSON prediction request.

www.ncbi.nlm.nih.gov/pubmed), annotation data extracted from Entrez genes, etc. This is one of its important features, since users can easily gather all the relevant biological information and analyse their integrated findings or validate their results. We are planning on integration with other software packages, developed in Matlab (or Octave) and Python. Python is gaining considerable momentum for machine learning applications as various packages facilitate the analysis of data, development and validation of models, conduction of various statistical analyses and other tasks. Scikit-learn (http://scikit-learn.org/stable/), pyBrain (http://pybrain.org), and mlpy (http://mlpy.sourceforge.net/) are a few of the numerous machine learning packages for Python.

Algorithm Implementations

Currently, Jaqpot Quattro contains the following API-compliant algorithm services: two implementations of multiple linear regressions (MLR) (using R and Weka [53] functionalities), and implementations of the partial least squares (PLS) algorithm (based on Weka), the support vector machine method (using the LIBSVM library [58]) and the sub-clustering algorithm developed in-house for Radial Basis Function Neural Networks [59]. As an example, the R implementation of the MLR regression algorithm was applied on the corona dataset to generate a linear NanoQSAR model that relates net cell association of gold nanoparticles (the logarithm base 2 transformed values) to zeta potential after synthesis, zeta potential after serum exposure, and a number of transformation defined in the PMML file found at http://app.jaqpot.org:8080/jaqpot/services/pmml/corona-standard-transformations. The produced model, trained with the algorithm with ID "ocpu-lm" (located at http://app.jaqpot.org:8080/jaqpot/services/ocpu-lm) can be found under the following address: http://app.jaqpot.org:8080/jaqpot/services/model/corona-model. OCPU-LM is implemented in R (using OpenCPU) and exposed via the JPDI API as explained in the previous section. To access these resources the client needs to provide an authentication token as specified by the access control API. Alternatively, the end user can easily access it via the Jaqpot Swagger interface (http://app.jaqpot.org:8080/jaqpot/swagger) using an authorization token produced automatically.

Apart from experimental descriptors available through the database, datasets used for modelling may contain theoretical descriptors, which are calculated using services that were originally developed during the OpenTox project, but are now being updated and extended, such as CDK [60] and MOPAC [61]. The eNanoMapper MOPAC implementation (available at: https://apps.ideaconsult.net/enanomapper/algorithm/ambit2.mopac.MopacOriginalStructure) was used to calculate quantum-mechanical descriptors for metal oxides, including HOMO (highest occupied molecular orbital), LUMO (lowest unoccupied molecular orbital), band gap and ionization potential. Figure 27 shows the results for Sb_2O_3 (available at http://enanomapper.github.io/bjnano7250433 ; login as guest is required for access). Calculations are available in various formats, including CSV, JSON, CML and SDF.

The leverage method for defining the "applicability domain" (AD) of NanoQSAR models has also been implemented and offered as a service. According to the OECD definition, the "applicability domain of a (Q)SAR model is the response and chemical structure space in which the model makes predictions with a given reliability" [62,63]. Defining the chemical structure space for nanomaterials is not trivial, hence the descriptor-based approach is adopted. The AD is created by applying a POST at an instance of the AD web service. Then, the predictive model can be linked to the AD model in such a way that predictions are accompanied by an indicator that informs us whether the query compound is in or out of the AD of the model.

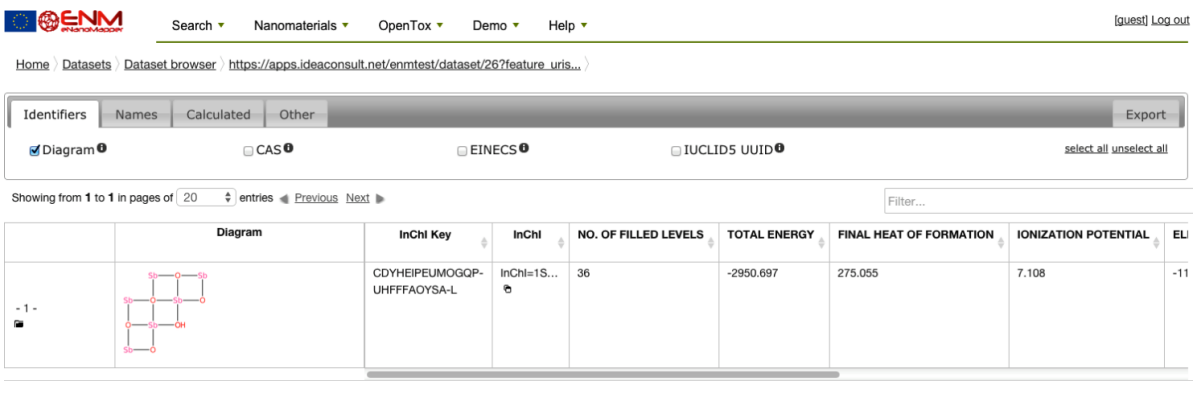


Figure 27: Screenshot of the descriptors calculated with quantum mechanics MOPAC web service.

Integration of modelling services in the framework

Dataset resources from any dataset service may be used by any modelling service which in turn will store the produced dataset of prediction on any dataset service. The eNanoMapper web services design assumes a distributed architecture in which data are not required to be stored or even indexed by a common system. Among services that implement the API, input data can come from any dataset service, be used by any modelling service, which in turn will submit the produced dataset with prediction results to any dataset service for storage. Linked-data principles are combined here with a REST-based design to enable this distribution of resources.

Discussion

The API with resources supporting substances, protocols and measurements is in line with recent publications in the domain and is able to support a variety of tests and endpoints, recommended by the OECD WPMN. The annotation with ontology entries is an ongoing collaboration between the eNanoMapper database and ontology teams and the EU NanoSafety Cluster. Data heterogeneity is a pervasive challenge within the nanosafety domain, with the complexity of the nanomaterials and their biological interactions being measured via multiple different types of assays and endpoints across a wide range of experimental technologies. While our prototype database and ontology already illustrate a range of these different measurements, the list of possible endpoints and characterisation properties is growing all the time as the science evolves, and our objective is ultimately to represent all relevant properties and endpoints in our ontology, which is currently growing through community feedback and as it is being used for annotations. Given the heterogeneity of the data being represented, a challenge of inconsistency may also emerge. Our platform is inspired by the OECD recommendations to define a minimum set of information that needs to be included as metadata in the case of each experiment type. Through templates, the fields that are required for different protocols can be customised.

The demonstration data provided by partners illustrates the capability of the API and the implementation to handle diverse information. It has been used for NanoQSAR modelling. Research is ongoing to extend the OpenTox algorithm and modelling APIs for nanomaterials, allowing these new models to be exposed with unique URIs suitable for reuse. The REST API with JSON serialisation is the current state of the art in web system development and data integration and enables building graphical summaries of the data, JavaScript widgets, custom user interfaces and programmatic interaction. The next steps include provision of RDF serialisation of the resources, support for multiple data formats on import and export, support for multiple search interfaces (including ones based on semantic technologies), and improvements of the data model, API and the implementation, based on the feedback and close collaboration with all eNanoMapper partners and EU NanoSafety Cluster working groups.

The eNanoMapper database discussed here is a design architecture that allows, in a first stage, for the import of experimental data and calculated descriptors by those who have measured or calculated them respectively, and in a second stage the use of data from the database for propagation or modelling. The eNanoMapper team from the beginning of the project paid attention to designing a system that would be strict in enforcing traceability of data, and in recording the details in its representation of nanomaterials and the specifics of how the data were generated (experimental conditions, methods). Users of the platform prototype feed it with data they have curated and know to be accurate. Any problematic uploads can be traced back to their source. In future work, metrics on the data such as the compliance level suggested by the Nanomaterial Registry (https://www.nanomaterialregistry.org/about/HowIsComplianceCalculated.aspx) could be introduced in order to progress the nanomaterial safety community towards a holistic approach to data quality that may be triggered from data storage, but this also needs to go back to the data origins, i.e., the specifics of experiments/measurements/calculations. Besides being accessible online at data.enanomapper.net, the system presented is an open source solution, which can be downloaded, installed and hosted by individual researchers or labs, and as such presents an open distributed platform for NM data management, rather than being restricted to use as a single database instance.

Data format conversions

Formatting experimental data as ISA-Tab files manually is very cumbersome and time consuming, even if using "semantically aware" tools, such as ISAcreator (https://github.com/ISA-tools/ISAcreator). Formatting data as ISA-Tab-Nano is even more challenging, as there is no publicly available validator of ISA-Tab-Nano, and the available examples at https://wiki.nci.nih.gov/display/ICR/ISA-TAB-Nano are more useful to convey the idea of the format, rather than to be considered the ultimate specification-compliant instances. Furthermore, while ISA-Tab validation relies heavily on XML assay templates, specifying the fields required by experiments with a defined endpoint and technology, the ISA-Tab-Nano wiki does not provide such templates, which makes it impossible to use existing ISA-Tab tools to generate ISA-Tab-Nano compliant files, even if ignoring the ENM-specific material files. Last but not least, the ISA-Tab specification only defines the metadata format and does not impose any restrictions on the actual data files. We consider two parallel roads towards improvement of

the status quo. First, enabling ISA-Tab-Nano support by the core ISA-Tab tools (ISAcreator), and second, an automatic generation of ISA-Tab archives, given the ubiquitous and convenient Excel templates as input. We have initiated work towards the first goal by extending a fork of the ISA-Tab core code to enable parsing of ISA-Tab-Nano files (https://github.com/enanomapper/ISAvalidator-ISAconverter-BIImanager). As this code is part of the ISAcreator application, it would potentially allow for loading and validating of ISA-Tab-Nano files through the core ISA-Tab tools. While ISA-Tab is designed to ensure that all experimental details are retained, the chemical compound or ENM is hidden in a step of the experimental graph, and such a data model is usually less convenient for preparing and querying the data and applying subsequent predictive modelling. Building on previous experience and taking into account the observation that the majority of EU NanoSafety Cluster projects prefer to prepare their experimental data using custom spreadsheet templates, the eNanoMapper team took an alternative, but pragmatic, approach by implementing support for a large set of custom spreadsheet templates for data preparation. We developed the configurable Excel parser described in the "Data import" section above. Being able to parse diverse spreadsheets, as well as other input formats (such as OHT) into the same internal data model and export the data from this data model into different formats allows us to provide format converters, in the same fashion as OpenBabel [64] (http://openbabel.org/) interconverts between chemical formats. Extending the tools to include ontology annotations and to be able to write the internal data model into ISA-Tab files will not only accomplish the second goal of automatically generating the files, but will also enable exporting query results from the database in a desired format.

Modelling

We are now developing a new R package that automates the creation of the best possible NanoQSAR regression model (validated using cross validation and external testing), by searching over many different regression algorithms and tuning the parameters for each algorithm. The suggested workflow automates the development of a reliable and well-validated NanoQSAR model or set of models by a simple call to an R function. The R package will be integrated within the eNanoMapper system using the JDPI and OpenCPU functionalities, described before.

Transmission electron microscopy (TEM) is a valuable technique for the characterization of nanomaterials. TEM image analysis yields number-based results, allows the extraction of size and shape-related attributes and characterization of surface topologies, and provides distinctions between the characterizations of primary particles and of aggregates/agglomerates. Based on TEM images, Gajewicz et. al. have proposed a set of image-derived descriptors for characterizing nanomaterials, such as volume, area, porosity and circularity [65]. These descriptors will be included in the set of descriptors to be computed by an image analysis tool that is under development in the context of the eNanoMapper project based on the standard and well accepted Fiji/ImageJ [66] open-source software, which was selected after an assessment of the most relevant software tools that are available and in use by the scientific community.

Integration of the facilities provided by R will allow for easy access to a wealth of additional algorithms and methods focusing on the analysis of omics data and utilization of useful information included in public ontologies such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [67]. Recent studies suggest that integrating multi-omics additional genomic knowledge can greatly assist towards fully understanding various phenotypes [68], as opposed to the conclusions drawn by focusing on only a single level of genomic data. Along these lines, we are working on an integration clustering analysis of the proteomics data included in protein corona datasets also incorporating information from the underlying relations in the data using the Gene Ontology [69]. For example, a hierarchical clustering algorithm is applied to NP proteomics data to build a hierarchy of protein clusters and compare them to those established by Gene Ontology; similarities between the two should reinforce any toxicology related outcome.

Technology

The REST API has become the most commonly used approach for web application development. Because of its simplicity and performance scalability it has replaced solutions such as the simple object access protocol (SOAP). The OpenTox project was in 2008 one of the first to define and implement a REST API in the cheminformatics and QSAR domains [14,23,70], but nowadays all the major chemical (and some material) databases provide access via REST. This applies to both data as well as computational functionality, including wrappers for popular software as R, science-as-a-service platforms, and high-performance computing, because the demand for interfacing via web services increases. REST is defined as a software architecture style designated for network-based applications, as the outcome of a thorough analysis of network architectures [71]. It is compliant with the successful architectural principles behind the World Wide Web that characterizes RESTful applications. Specifically, the principles were selected to ensure the distributed system will feature a set of particular properties: simplicity, scalability, performance, modifiability, visibility of communication, portability, reliability, and resistance to failure. The granularity of the REST resources is not fixed, but can be designed to fit particular application needs. A set of resources is

a resource itself, hence there are no efficiency limitations on the retrieval of large amounts of data. A potential challenge when processing large amounts of data is the output in textual format, however a resource representation can be compressed, and in principle the JSON output used is much more terse than other formats (e.g., RDF). REST allows for the provision of representations in multiple formats, hence formats suitable for representing sparse data can be utilized. A known limitation is that a REST API specifies how questions can be asked and therefore restricts the users in what they can ask, compared to being able to access the data directly via, e.g., SPARQL (SPARQL protocol and RDF query language) or SQL (structured query language). While eNanoMapper plans to enable SPARQL queries for NM data, this approach has its own drawbacks which often motivate the hiding of SPARQL behind a REST API. Despite the overwhelming use of REST with HTTP protocol and HTTP URIs, originally REST was a protocol-independent architecture and could be used outside of the HTTP context, which, in principle, allows for the adoption of binary protocols for effectiveness (such as Google protocol buffers, Apache Thrift, etc.). However binary protocols are much harder to use. We do not expect a solution other than HTTP to be required in the lifetime of the eNanoMapper project. Finally, it deliberately adopts the choice of a distributed database system, which follows the same idea as the World Wide Web, and is in accordance with the REST architecture of an ecosystem of distributed entities that interact and are made available independently from each other.

Conclusion

The eNanoMapper database builds on previous experience from the OpenTox and ToxBank projects in supporting diverse data through flexible data storage, semantic web technologies, open source components and web services. A number of opportunities and challenges exist in nanomaterials representation and integration of ENM information, originating from diverse systems. We adopted the concept of substances, allowing a more elaborate representation of ENMs, overcoming limitations of existing compound-based databases and integration solutions. We describe how an approach of adopting an ontology-supported data model, covering substances and measurements, provides a common ground for integration. The data sources supported include diverse formats (ISA-Tab, OECD harmonized templates, custom spreadsheet templates), as well as other formats via custom import scripts. Besides retaining the data provenance, the focus on measurements provides insights into how to reuse chemical structure database tools for nanomaterials characterization and safety.

The database is still under development within the eNanoMapper project. Future work includes support for high-throughput screening (HTS) data, further annotation with ontologies, and support for data from aforementioned third-party databases, such as PubChem and ArrayExpress. HTS and high-content analysis data are currently being generated in several of the projects within the EU NanoSafety Cluster, including the eNanoMapper partners. As these datasets become available, they will be able to serve in generating use cases for further development, refinement and proof-of-concept of the current state of the eNanoMapper database and ontology framework.

Nanomaterials synthesis until the final product stage may potentially involve several analyses, where go, no-go decisions are made from evaluating safety and other aspects of the materials. Ultimately, we envision that the eNanoMapper infrastructure should be directly applicable to such a safe-by-design principle, directly coupling the material and product development stages with safety analysis. The current prototype provides us with the means of comparing new nanomaterials to an expanding collection of reference data.

Supporting Information

Supporting Information File 1

OECD WPMN recommended endpoints and their potential correspondence to UDS and ISA-Tab-Nano concepts.
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-165-S1.pdf]

Acknowledgements

The eNanoMapper project is funded by the European Union's Seventh Framework Programme for research, technological development and demonstration (FP7-NMP-2013-SMALL-7) under grant agreement no. 604134.

References

1. Maynard, A. D.; Aitken, R. J.; Butz, T.; Colvin, V.; Donaldson, K.; Oberdörster, G.; Philbert, M. A.; Ryan, J.; Seaton, A.; Stone, V.; Tinkle, S. S.; Tran, L.; Walker, N. J.; Warheit, D. B. *Nature* **2006,** *444,* 267–269. doi:10.1038/444267a
2. European Commission. *Types and uses of nanomaterials, including safety aspects;* 2012.
3. Cárdenas, W. H. Z.; Mamani, J. B.; Sibov, T. T.; Caous, C. A.; Amaro, E., Jr.; Gamarra, L. F. *Int. J. Nanomed.* **2012,** *7,* 2699–2712. doi:10.2147/IJN.S30074
4. Kelder, T.; van Iersel, M. P.; Hanspers, K.; Kutmon, M.; Conklin, B. R.; Evelo, C. T.; Pico, A. R. *Nucleic Acids Res.* **2012,** *40,* D1301–D1307. doi:10.1093/nar/gkr1074
5. Mills, K. C.; Murry, D.; Guzan, K. A.; Ostraat, M. L. *J. Nanopart. Res.* **2014,** *16,* No. 2219. doi:10.1007/s11051-013-2219-8

6. Miller, A. L.; Hoover, M. D.; Mitchell, D. M.; Stapleton, B. P. *J. Occup. Environ. Hyg.* **2007,** *4,* D131–D134. doi:10.1080/15459620701683947
7. Gaheen, S.; Hinkal, G. W.; Morris, S. A.; Lijowski, M.; Heiskanen, M.; Klemm, J. D. *Comput. Sci. Discovery* **2013,** *6,* 014010. doi:10.1088/1749-4699/6/1/014010
8. Marquardt, C.; Kühnel, D.; Richter, V.; Krug, H. F.; Mathes, B.; Steinbach, C.; Nau, K. *J. Phys.: Conf. Ser.* **2013,** *429,* 012060. doi:10.1088/1742-6596/429/1/012060
9. Kong, L.; Tuomela, S.; Hahne, L.; Ahlfors, H.; Yli-Harja, O.; Fadeel, B.; Lahesmaa, R.; Autio, R. *PLoS One* **2013,** *8,* e68414. doi:10.1371/journal.pone.0068414
10. Jeliazkova, N.; Doganis, P.; Fadeel, B.; Grafstrom, R.; Hastings, J.; Jeliazkov, V.; Kohonen, P.; Munteanu, C. R.; Sarimveis, H.; Smeets, B.; Tsiliki, G.; Vorgrimmler, D.; Willighagen, E. The first eNanoMapper prototype: A substance database to support safe-by-design. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* IEEE: Piscataway, NJ, United States of America, 2014; pp 1–9. doi:10.1109/bibm.2014.6999367
11. Mustad, A. P.; Smeets, B.; Jeliazkova, N.; Jeliazkov, V.; Willighagen, E. *Summary of the Spring 2014 NSC Database Survey;* 2014. doi:10.6084/m9.figshare.1195888
12. Panneerselvam, S.; Choi, S. *Int. J. Mol. Sci.* **2014,** *15,* 7158–7182. doi:10.3390/ijms15057158
13. Lynch, I. Compendium of Projects in the European NanoSafety Cluster. *European NanoSafety Cluster;* 2014; pp 250 ff.
14. Jeliazkova, N.; Jeliazkov, V. *J. Cheminf.* **2011,** *3,* 18. doi:10.1186/1758-2946-3-18
15. Sansone, S.-A.; Rocca-Serra, P.; Field, D.; Maguire, E.; Taylor, C.; Hofmann, O.; Fang, H.; Neumann, S.; Tong, W.; Amaral-Zettler, L.; Begley, K.; Booth, T.; Bougueleret, L.; Burns, G.; Chapman, B.; Clark, T.; Coleman, L.-A.; Copeland, J.; Das, S.; de Daruvar, A.; de Matos, P.; Dix, I.; Edmunds, S.; Evelo, C. T.; Forster, M. J.; Gaudet, P.; Gilbert, J.; Goble, C.; Griffin, J. L.; Jacob, D.; Kleinjans, J.; Harland, L.; Haug, K.; Hermjakob, H.; Ho Sui, S. J.; Laederach, A.; Liang, S.; Marshall, S.; McGrath, A.; Merrill, E.; Reilly, D.; Roux, M.; Shamu, C. E.; Shang, C. A.; Steinbeck, C.; Trefethen, A.; Williams-Jones, B.; Wolstencroft, K.; Xenarios, I.; Hide, W. *Nat. Genet.* **2012,** *44,* 121–126. doi:10.1038/ng.1054
16. Thomas, D. G.; Gaheen, S.; Harper, S. L.; Fritts, M.; Klaessig, F.; Hahn-Dantona, E.; Paik, D.; Pan, S.; Stafford, G. A.; Freund, E. T.; Klemm, J. D.; Baker, N. A. *BMC Biotechnol.* **2013,** *13,* 2. doi:10.1186/1472-6750-13-2
17. Roebben, G.; Rasmussen, K.; Kestens, V.; Linsinger, T. P. J.; Rauscher, H.; Emons, H.; Stamm, H. *J. Nanopart. Res.* **2013,** *15,* 1455. doi:10.1007/s11051-013-1455-2
18. Series on the Safety of Manufactured Nanomaterials No. 27. List of manufactured nanomaterials and list of endpoints for phase one of the sponsorship programme for the testing of manufactured nanomaterials: revision, 2010.
19. Rumble, J.; Freiman, S.; Teague, C. The description of nanomaterials: A multi-disiciplinary uniform description system. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* IEEE: Piscataway, NJ, United States of America, 2014; pp 34–39. doi:10.1109/BIBM.2014.6999372
20. CODATA-VAMAS Working Group on the Description of Nanomaterials, Uniform Description System for Materials on the Nanoscale, 2015.
21. Abeyruwan, S.; Vempati, U. D.; Küçük-McGinty, H.; Visser, U.; Koleti, A.; Mir, A.; Sakurai, K.; Chung, C.; Bittker, J. A.; Clemons, P. A.; Brudz, S.; Siripala, A.; Morales, A. J.; Romacker, M.; Twomey, D.; Bureeva, S.; Lemmon, V.; Schürer, S. C. *J. Biomed. Semantics* **2014,** *5* (Suppl. 1), S5. doi:10.1186/2041-1480-5-S1-S5
22. Hastings, J.; Jeliazkova, N.; Owen, G.; Tsiliki, G.; Munteanu, C. R.; Steinbeck, C.; Willighagen, E. *J. Biomed. Semantics* **2015,** *6,* 10. doi:10.1186/s13326-015-0005-5
23. Hardy, B.; Douglas, N.; Helma, C.; Rautenberg, M.; Jeliazkova, N.; Jeliazkov, V.; Nikolova, I.; Benigni, R.; Tcheremenskaia, O.; Kramer, S.; Girschick, T.; Buchwald, F.; Wicker, J.; Karwath, A.; Gütlein, M.; Maunz, A.; Sarimveis, H.; Melagraki, G.; Afantitis, A.; Sopasakis, P.; Gallagher, D.; Poroikov, V.; Filimonov, D.; Zakharov, A.; Lagunin, A.; Gloriozova, T.; Novikov, S.; Skvortsova, N.; Druzhilovsky, D.; Chawla, S.; Ghosh, I.; Ray, S.; Patel, H.; Escher, S. *J. Cheminf.* **2010,** *2,* 7. doi:10.1186/1758-2946-2-7
24. Kohonen, P.; Benfenati, E.; Bower, D.; Ceder, R.; Crump, M.; Cross, K.; Grafström, R. C.; Healy, L.; Helma, C.; Jeliazkova, N.; Jeliazkov, V.; Maggioni, S.; Miller, S.; Myatt, G.; Rautenberg, M.; Stacey, G.; Willighagen, E.; Wiseman, J.; Hardy, B. *Mol. Inf.* **2013,** *32,* 47–63. doi:10.1002/minf.201200114
25. European Commission. *Towards the replacement of in vivo repeated dose systemic toxicity testing;* 2013; Vol. 3.
26. Thangasamy, I. *OpenAM;* Packt Publishing Ltd.: Birmingham, U.K., 2011; pp 276 ff.
27. Willighagen, E. NanoWiki (release 1), 2015. doi:10.6084/m9.figshare.1330208
28. Walkey, C. D.; Olsen, J. B.; Song, F.; Liu, R.; Guo, H.; Olsen, D. W. H.; Cohen, Y.; Emili, A.; Chan, W. C. W. *ACS Nano* **2014,** *8,* 2439–2455. doi:10.1021/nn406018q
29. Vinken, M. *Toxicology* **2013,** *312,* 158–165. doi:10.1016/j.tox.2013.08.011
30. Godwin, H.; Nameth, C.; Avery, D.; Bergeson, L. L.; Bernard, D.; Beryt, E.; Boyes, W.; Brown, S.; Clippinger, A. J.; Cohen, Y.; Doa, M.; Hendren, C. O.; Holden, P.; Houck, K.; Kane, A. B.; Klaessig, F.; Kodas, T.; Landsiedel, R.; Lynch, I.; Malloy, T.; Miller, M. B.; Muller, J.; Oberdorster, G.; Petersen, E. J.; Pleus, R. C.; Sayre, P.; Stone, V.; Sullivan, K. M.; Tentschert, J.; Wallis, P.; Nel, A. E. *ACS Nano* **2015,** *9,* 3409–3417. doi:10.1021/acsnano.5b00941
31. Lynch, I.; Weiss, C.; Valsami-Jones, E. *Nano Today* **2014,** *9,* 266–270. doi:10.1016/j.nantod.2014.05.001
32. Klimisch, H.-J.; Andreae, M.; Tillmann, U. *Regul. Toxicol. Pharmacol.* **1997,** *25,* 1–5. doi:10.1006/rtph.1996.1076
33. ECHA. Evaluation of available information. *Guidance on information requirements and chemical safety assessment;* 2011; pp 16 ff.
34. Jahnke, L.; Asher, A.; Keralis, S. D. C. *The Problem of Data;* Council on Library and Information Resources, 2012.
35. ECHA Dissemination site, Multi-Walled Carbon Nanotubes (MWCNT), synthetic graphite in tubular shape. http://apps.echa.europa.eu/registered/data/dossiers/DISS-b281d1a0-c6d8-5dcf-e044-00144f67d031/AGGR-64114fd2-a71c-46aa-ad6a-b519fe96dbc1_DISS-b281d1a0-c6d8-5dcf-e044-00144f67d031.html#AGGR-64114fd2-a71c-46aa-ad6a-b519fe96dbc1 (accessed June 6, 2015).
36. Jaworska, J.; Nikolova-Jeliazkova, N. *SAR QSAR Environ. Res.* **2007,** *18,* 195–207. doi:10.1080/10629360701306050
37. Jeliazkova, N.; Kochev, N. *Mol. Inf.* **2011,** *30,* 707–720. doi:10.1002/minf.201100028
38. Nikolova, N.; Jaworska, J. *QSAR Comb. Sci.* **2004,** *22,* 1006–1026. doi:10.1002/qsar.200330831

39. Willett, P.; Barnard, J.; Downs, G. *J. Chem. Inf. Model.* **1998,** *38,* 983–996. doi:10.1021/ci9800211
40. Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. *J. Med. Chem.* **2014,** *57,* 3186–3204. doi:10.1021/jm401411z
41. Arts, J. H. E.; Hadi, M.; Irfan, M.-A.; Keene, A. M.; Kreiling, R.; Lyon, D.; Maier, M.; Michel, K.; Petry, T.; Sauer, U. G.; Warheit, D.; Wiench, K.; Wohlleben, W.; Landsiedel, R. *Regul. Toxicol. Pharmacol.* **2015,** *71,* S1–S27. doi:10.1016/j.yrtph.2015.03.007
42. Gajewicz, A.; Cronin, M. T. D.; Rasulev, B.; Leszczynski, J.; Puzyn, T. *Nanotechnology* **2015,** *26,* 015701. doi:10.1088/0957-4484/26/1/015701
43. *D3: A JavaScript visualization library for HTML and SVG;* 2015, https://github.com/mbostock/d3/.
44. *ambit.js,* Release 0.0.2; E. Willighagen, 2015. doi:10.5281/zenodo.16517
45. Rorabacher, D. A. *Anal. Chem.* **1991,** *63,* 139–146. doi:10.1021/ac00002a010
46. Angiulli, F.; Pizzuti, C. Fast Outlier Detection in High Dimensional Spaces. In *Principles of Data Mining and Knowledge Discovery;* Elomaa, T.; Mannila, H.; Toivonen, H., Eds.; Lecture Notes in Computer Science, Vol. 2431; Springer: Berlin, Germany, 2002; pp 15–27.
47. Hautamaki, V.; Karkkainen, I.; Franti, P. Outlier Detection Using k-Nearest Neighbour Graph. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04),* IEEE Computer Society: Washington, DC, USA, 2004; pp 430–433.
48. Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. A Practical Guide to Support Vector Classification. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (accessed March 15, 2015).
49. Arora, M.; Bhambhu, L. *Int. J. Adv. Res. Comput. Sci. Software Eng.* **2014,** *4,* 271.
50. Guazzelli, A.; Zeller, M.; Lin, W.-C.; Williams, G. *The R Journal* **2009,** *1,* 60–65.
51. Pechter, R. *ACM SIGKDD Explorations Newsletter* **2009,** *11,* 19–25. doi:10.1145/1656274.1656279
52. Tcheremenskaia, O.; Benigni, R.; Nikolova, I.; Jeliazkova, N.; Escher, S. E.; Batke, M.; Baier, T.; Poroikov, V.; Lagunin, A.; Rautenberg, M.; Hardy, B. *J. Biomed. Semantics* **2012,** *3* (Suppl. 1), No. S7. doi:10.1186/2041-1480-3-S1-S7
53. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. *ACM SIGKDD Explorations Newsletter* **2009,** *11,* 10–18. doi:10.1145/1656274.1656278
54. R Core Team. *R: A language and environment for statistical computing;* R Foundation for Statistical Computing: Vienna, Austria, 2012.
55. Ooms, J. *arXiv* **2014,** No. 1406.4806.
56. Smith, D. R is hot. http://www.revolutionanalytics.com/whitepaper/r-hot (accessed March 31, 2015).
57. Kim, J.; Wang, G.; Bae, S. T. *Int. J. Semantic Comput.* **2014,** *08,* 99–117. doi:10.1142/S1793351X14500056
58. Chang, C.-C.; Lin, C.-J. *ACM Trans. Intell. Syst. Technol.* **2011,** *2,* No. 27. doi:10.1145/1961189.1961199
59. Sarimveis, H.; Alexandridis, A.; Bafas, G. *Neurocomputing* **2003,** *51,* 501–505. doi:10.1016/S0925-2312(03)00342-4
60. Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. *Curr. Pharm. Des.* **2006,** *12,* 2111–2120. doi:10.2174/138161206777585274
61. Stewart, J. J. P. *J. Comput.-Aided Mol. Des.* **1990,** *4,* 1–103. doi:10.1007/BF00128336
62. Netzeva, T. I.; Worth, A.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D.; Schultz, T.; Stanton, D. W.; van de Sandt, J. J. M.; Tong, W.; Veith, G.; Yang, C. *ATLA, Altern. Lab. Anim.* **2005,** *33,* 155–173.
63. Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. *ATLA, Altern. Lab. Anim.* **2005,** *33,* 445–459.
64. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminf.* **2011,** *3,* 33. doi:10.1186/1758-2946-3-33
65. Gajewicz, A.; Schaeublin, N.; Rasulev, B.; Hussain, S.; Leszczynska, D.; Puzyn, T.; Leszczynski, J. *Nanotoxicology* **2015,** *9,* 313–325. doi:10.3109/17435390.2014.930195
66. Schneider, C. A.; Rasband, W. S.; Eliceiri, K. W. *Nat. Methods* **2012,** *9,* 671–675. doi:10.1038/nmeth.2089
67. Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M.; Tanabe, M. *Nucleic Acids Res.* **2012,** *40,* D109–D114. doi:10.1093/nar/gkr988
68. Kim, D.; Joung, J.-G.; Sohn, K.-A.; Shin, H.; Park, Y. R.; Ritchie, M. D.; Kim, J. H. *J. Am. Med. Inf. Assoc.* **2015,** *22,* 109–120.
69. Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. *Nat. Genet.* **2000,** *25,* 25–29. doi:10.1038/75556
70. Jeliazkova, N. *Expert Opin. Drug Metab. Toxicol.* **2012,** *8,* 791–801. doi:10.1517/17425255.2012.685158
71. Fielding, R. T.; Taylor, R. N. Principled Design of the Modern Web Architecture. In *Proceedings of the 22Nd International Conference on Software Engineering,* ACM: New York, NY, U.S.A., 2000; pp 407–416.

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Analysis of soil bacteria susceptibility to manufactured nanoparticles via data visualization

Rong Liu*[1,2], Yuan Ge[1,3], Patricia A. Holden[1,3,4] and Yoram Cohen*[1,2,5]

Full Research Paper

Open Access

Address:
[1]Center for Environmental Implications of Nanotechnology, University of California, United States, [2]Institute of the Environment and Sustainability, University of California, Los Angeles, United States, [3]Earth Research Institute, University of California, Santa Barbara, United States, [4]Bren School of Environmental Science and Management, University of California, Santa Barbara, United States and [5]Chemical and Biomolecular Engineering Department, University of California, Los Angeles, United States

Email:
Rong Liu* - rongliu@ucla.edu; Yoram Cohen* - yoram@ucla.edu

* Corresponding author

Keywords:
environmental impact; manufactured nanoparticles; nanoinformatics; soil bacteria; visualization

Beilstein J. Nanotechnol. **2015,** *6,* 1635–1651.
doi:10.3762/bjnano.6.166

Received: 06 April 2015
Accepted: 02 July 2015
Published: 28 July 2015

This article is part of the Thematic Series "Nanoinformatics for environmental health and biomedicine".

Associate Editor: P. Ziemann

Abstract

The impact of ZnO and TiO_2 manufactured nanoparticles (MNPs) on soil bacterial communities for different exposure periods and MNP doses was explored via data visualization techniques. Interrelationships between MNP treatments and responses of bacterial taxa were illustrated by bipartite graphs, allowing fast identification of important soil bacterial taxa that are susceptible to MNPs. Contribution biplots with subcompositional coherence property were generated via log-ratio analysis (LRA), which jointly display the treatment distribution and the variance (contribution) of bacterial taxa. The LRA contribution biplots and nonmetric multi-dimensional scaling (NMDS) of the dataset, along with hierarchical clustering, demonstrated that high doses of ZnO and TiO_2 MNPs caused significant compositional changes in soil bacterial communities. The suitability of family level for MNP taxonomic impact assessment was demonstrated by both the LRA biplots and simplified NMDSs with quantification provided by the distance correlation between MNP impacts summarized at different taxonomic levels. The present study demonstrates that visual exploration could potentially assist in knowledge discovery and interpretation of data on soil bacterial communities exposed to MNPs and thus evaluate the potential for environmental impacts.

Introduction

Manufactured nanoparticles (MNPs) are now routinely used in numerous products and applications due to their novel functional properties that arise at the nanoscale [1,2]. However, as the applications of MNPs rapidly expand [2,3], there is an increased public concern regarding the potential environmental and health risks associated with MNPs [4-9] throughout their

lifecycle [10-14]. MNPs may be released to the environment as the result of a variety of human-related activities (air emissions and/or direct discharge to surface water, etc.), wherein they can move across environmental boundaries and are therefore likely to be found in most media [13,14]. The presence of MNPs in the environment could lead to exposures of ecological receptors to MNPs via multiple pathways [13]. Although there is lack of field monitoring data regarding environmental concentrations for most MNPs, various simulations [14,15] of multimedia environmental distributions of MNPs suggest that MNPs tend to accumulate in soil and sediment [16,17]. Various studies [18-22] have reported that MNPs could lead to adverse environmental impacts. For example, Ag and Pt MNPs may interfere with zebrafish embryo hatching [23]; ZnO MNPs may cause compositional changes in soil bacterial communities [18,19]; quantum dots (QDs) were linked to DNA damage of both freshwater mussels and gills [24]; and carbon nanotubes have been found to induce harmful effects to various organs (such as aquatic animals, bacteria, and plants) [25].

MNPs in soil can cause compositional changes to soil bacterial communities and thus may induce profound impacts on terrestrial ecosystems [16,26]. Soil microbial communities, as one of the most abundant and diverse groups of organisms on earth, perform many critical ecosystem functions (e.g., element cycling and waste decomposition) [27,28] and are important biotic indicators of soil health [29]. Therefore, information about MNP effects on soil microbial communities is critical for environmental impact assessment [13]. Recently, efforts [18,19,26,30,31] have been devoted to investigate the impacts of various MNPs on soil bacterial communities, resulting in large datasets of high dimensionality (e.g., over 10^5 soil DNA sequences extracted for a treatment) [18,19]. Therefore, advanced data exploration/visualization approaches are required to allow researchers to design subsequent confirmatory experiments and/or perform detailed statistical analyses. Graphical displays of multivariate (high-dimensional) ecological data can also facilitate data comparison and interpretation (e.g., acquainting variables of important roles/contributions and identifying similarity/distribution among samples) [32]. In addition, since bacterial community data are usually compositional (each sample is profiled by a set of non-negative values that add up to unity), it is important that their analyses are subcompositionally coherent (i.e., the relationship between two components (variables) should be the same and not dependent on the presence/absence of other components) [32].

Accordingly, in the present work, we report on a range of visual exploration approaches suitable for analysis of high content dataset for bacterial communities exposed to MNPs. Bipartite graphs [33-35] were established to illustrate interrelationships between MNPs and responses of bacterial taxa. Log-ratio analysis [32,36,37] that has subcompositional coherence property was utilized to generate biplots for joint displays of sample (treatment) separation/distribution and the contribution of bacterial taxa (i.e., the variances of bacterial taxa across all the treatments). In addition, the impacts of different MNPs were projected and explored via two-dimensional (2D) maps constructed by hierarchical clustering [32,38,39] and multidimensional scaling [32,40]. Also, a recently developed distance correlation [41] was employed to quantify the consistency between MNP impacts summarized at a range of taxonomic levels.

Materials and Methods

Data for soil bacterial communities exposed to MNPs

Visual exploration was conducted for a previously reported dataset of MNP impacts on soil bacterial communities [18]. The dataset contained 15 treatments (i.e., different MNP exposure tests) including TiO_2 and ZnO MNPs of primary size in the range of about 15–20 nm and about 20–30 nm [42], respectively. The soil bacteria were exposed to the above MNPs for 15 and 60 days at three different doses (0.5, 1.0, and 2.0 mg/g (soil) for TiO_2 MNPs and 0.05, 0.1, and 0.5 mg/g (soil) for ZnO MNPs) as well as 0, 15, and 60 day controls (without MNPs) [18]. Soil DNA sequences were recovered for the above 15 treatments (in quadruplicate). The recovered DNA sequences were clustered into 31,621 bacterial operational taxonomic units (OTUs) [18], with the number of DNA sequences clustered into the same OTU counted to quantify the impact of the 15 treatments on soil bacterial communities [18]. The OTUs were further summarized/assigned into a set of hierarchical taxa (i.e., genus (446), family (135), order (53), class (41), and phylum (19); the total number of taxa at each taxonomic level is given in the parentheses) [18]. For each taxonomic level (including OTU), the total counts of sequences assigned to a specific taxon represent its abundance, while the relative abundance of the taxon in the whole community was used as a measure of the impacts of the 15 treatments [18].

Exploration workflow

Visual exploration of the above soil bacterial community data [18] followed a workflow summarized in Figure 1. The analysis was conducted to identify significant MNP-bacterial taxon interrelationships and to assess the similarity of MNP impacts on soil bacterial communities. For each taxonomic level (from genus to phylum), bacterial taxa that are susceptible to MNP treatments were identified according to a threshold of interpercentile range. Interrelationships between the MNP treatments and the identified susceptible bacterial taxa were illustrated using bipartite graphs [33-35]. Biplots were generated by

log-ratio analysis [32,36,37] (of subcompositional coherence property) to jointly display the separation (distribution) of treatments and the contribution (variance) of bacterial taxa. Multidimensional scaling analysis [32,40] was conducted, along with hierarchical clustering, in order to illustrate the main underlying structure of the soil bacterial community dataset. In addition, distance correlation coefficients [41] were calculated to assess the consistency of MNP impacts summarized at different taxonomic levels.

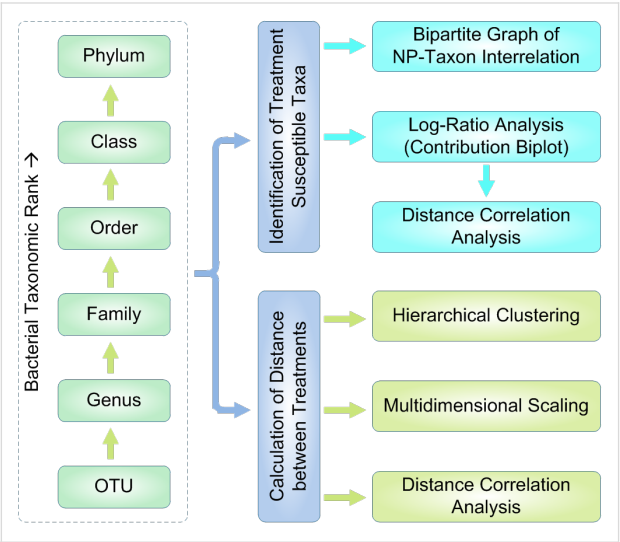


Figure 1: Workflow for visual data exploration of soil bacteria susceptible to MNP treatments.

MNP-Bacteria Interrelationships

The interrelationships between MNPs and the responses of bacterial taxa were explored using bipartite graphs [33-35]. It is noted that some bacterial taxa demonstrated only marginal variance across the 15 treatments (in quadruplicate), indicating their insusceptibility to the treatments. It is noted that the presence of treatment insusceptible bacterial taxa will complicate bipartite graphs without adding useful information. Therefore, in the present work, bacterial taxa which is in the 95th–5th percentile range in terms of relative abundance across all the 15 treatments (in quadruplicate) less than a prescribed threshold (e.g., 10/n, where n denotes the total number of bacterial taxa at a given taxonomic level) were discarded as being treatment insusceptible. The relative abundances of the remaining bacterial taxa that were considered as treatment susceptible were re-scaled to sum up to unity for each treatment. Bipartite graphs were then established based on the averaged relative abundance of bacterial taxa for each quadruplicated treatment. In an established bipartite graph, treatments and bacterial taxa were represented as nodes on opposite sides of the graph, with linkages between them indicating the bacterial taxa (and their relative abundance) identified for each treatment or vice versa.

Log-ratio analysis

Log-ratio analysis (LRA) [32,36,37] was conducted for the bacterial taxa that were identified as treatment susceptible in order to further explore and visualize the impact of TiO_2 and ZnO MNPs on the soil bacterial communities. In LRA, the relative abundances of bacterial taxa (i.e., compositional variables) were transformed to log-ratios to attain subcompositional coherence [32,36,37]. For example, given a dataset of four compositional variables (i.e., components) a, b, c, and d, a subcompositional dataset of a', b', and c' can be obtained by discarding component d (note that the subcompositional dataset is closed again, i.e., $a' = a / (a + b + c)$, $b' = b / (a + b + c)$, and $c' = c / (a + b + c)$ so that $a' + b' + c' = 1$). After log-transformation, the distance between the composition a' and b' is given by:

$$\sqrt{1/n\sum_{i=1}^{n}(\log(a_i')-\log(b_i'))^2} = \sqrt{1/n\sum_{i=1}^{n}(\log(a_i)-\log(b_i))^2} \quad (1)$$

where n denotes the total number of samples in the dataset. It is noted that the log-ratio distance between two components remains the same irrespective of the presence/absence of other components (i.e., subcompositional coherence).

In LRA, once a compositional data matrix $\boldsymbol{G}$ (e.g., relative abundance of bacterial taxa) is transformed into log-ratios, a double centered matrix (i.e., row and column sums are all equal to zero) is constructed as:

$$\boldsymbol{A} = \left(\boldsymbol{I} - \mathbf{1}\boldsymbol{r}^{\mathrm{T}}\right)\log(\boldsymbol{G})\left(\boldsymbol{I} - \mathbf{c}\mathbf{1}^{\mathrm{T}}\right)^{\mathrm{T}} \quad (2)$$

where $\boldsymbol{I}$ and $\mathbf{1}$ denotes identity matrix and vectors of ones of appropriate size, respectively. In addition, the two vectors $\boldsymbol{r}$ and $\boldsymbol{c}$ are the row and column sums of $\boldsymbol{G}$ relative to the grand total. The above double centered matrix is further weighted as follows:

$$\mathrm{S} = \boldsymbol{D}_r^{1/2}\boldsymbol{A}\boldsymbol{D}_c^{1/2} \quad (3)$$

where $\boldsymbol{D}_\mathrm{r}$ and $\boldsymbol{D}_\mathrm{c}$ are the diagonal matrices corresponding to vectors $\boldsymbol{r}$ and $\boldsymbol{c}$, respectively. Singular value decomposition (SVD) [43] of the weighted matrix produces:

$$\boldsymbol{S} = \boldsymbol{U}\,\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \quad (4)$$

From the above SVD, the following coordinate matrices can be obtained:

$$\begin{aligned}&\text{Contribution row coordinates}: \boldsymbol{U}\\&\text{Contribution column coordinates}: \boldsymbol{V}\\&\text{Standard row coordinates}: \boldsymbol{D}_r^{-1/2}\boldsymbol{U}\\&\text{Standard column coordinates}: \boldsymbol{D}_c^{-1/2}\boldsymbol{V}\\&\text{Principal row coordinates}: \boldsymbol{D}_r^{-1/2}\boldsymbol{U}\boldsymbol{\Sigma}\\&\text{Principal column coordinates}: \boldsymbol{D}_c^{-1/2}\boldsymbol{V}\boldsymbol{\Sigma}\end{aligned} \quad (5)$$

Based on the coordinates provided by LRA, various biplots can be constructed to represent treatments (samples) and bacterial taxa (variables) together. For example, principal row and standard column coordinates can be displayed (using the first two columns of the coordinate matrices) jointly as a row-principal biplot, while the combination of standard row and principal column coordinates yields a column-principal biplot. When there are many components (e.g., bacterial taxa) a convenient alternative is to derive a contribution biplot by combining standard row and contribution column coordinates or contribution row and standard column coordinates [36]. It is noted that LRA requires the compositional data matrix to be strictly positive. However, a few zeros could remain in the compositional data matrix even after the removal of the bacterial taxa that are identified as treatment insusceptible. In the present work, for a given taxonomic level, the remaining vanishing relative abundances of bacterial taxa was substituted by half of the smallest non-zero value in the complete data (before the removal of treatment insusceptible bacterial taxa) [36], followed by a rescaling step to close the data again (i.e., the relative abundance sums to unity for each treatment).

Multidimensional scaling analysis

Multidimensional scaling (MDS) analysis [32,40] was also conducted for the soil bacterial community dataset with the objective of representing the treatments in a two-dimensional (2D) map while maintaining (as closely as possible) the inter-treatment distance. Unlike LRA, MDS is not subcompositionally coherent [32,36,37] and thus was conducted with the complete dataset (i.e., no bacterial taxa removed) of each taxonomic level (from OTU, genus, …, to phylum). For a given taxonomic level, in order to conduct MDS, distances between treatments need to be calculated first based on their relative abundances. In the present work, Bray-Curtis dissimilarity (BCD), as the most widely used dissimilarity metric in ecological data analyses [32,44], was calculated to quantify the difference between the 15 treatments (in quadruplicate). For raw OTU counts, BCD between two treatments [32] was calculated by:

$$d_{ij} = \Sigma_k \,|\, n_{ik} - n_{jk} \,|\, / \Sigma_k \,|\, n_{ik} + n_{jk} \,| \quad (6)$$

in which n_{ik} and n_{jk} represent the k-th OTU count for treatment i and j, respectively. As the OTU counts were converted into relative abundances ($r_{ik} = n_{ik}/\Sigma_k n_{ik}$), the BCD reduces to the regular L_1 distance [32]:

$$d_{ij} = \Sigma_k \,|\, r_{ik} - r_{jk} \,|\, /2 \quad (7)$$

The above L_1 distance calculation resulted in a 60 × 60 matrix for each taxonomic level since quadruplicates were used for each of the 15 treatments.

Coordinates for plotting the treatments in 2D maps were derived from the L_1 distance matrices via MDS [32,40] (using the isoMDS function of R package MASS [45]). Since the L_1 distance is a non-Euclidean distance, the above MDS is referred to as nonmetric MDS (NMDS) [32,40]. The quality of the NMDSs was then quantified by the normalized sum of squared approximation errors known as *stress* [32,40]. In the NMDS established for each taxonomic level there were 60 points, corresponding to the 15 treatments (in quadruplicate). In order to avoid obscureness induced by treatment replicates, reduced NMDSs were developed by using the average-link as the metric to measure the distance between different treatments. The average-link between treatment S_i and S_j was calculated as:

$$d\left(S_i, S_j\right) = \Sigma_{x\in S_i}\Sigma_{y\in S_j} d\left(x,y\right) / \left|S_i\right|\left|S_j\right| \quad (8)$$

The developed NMDs were converted into biplots by adding vectors to represent bacterial taxa [32]. For a bacterial taxon, the relevant vector was obtained via linear regression of the relative abundance (quadruplicates averaged for the bacterial taxon) on the NMDS coordinates. The vector was formed by the regression coefficients of the NMDS coordinates which then served to indicate the direction the greatest ascent in the regression plane (i.e., gradient vector) [32].

In addition, hierarchical clustering [32,38,39] was carried out based on the L_1 distance matrices to identify treatments that induced similar impacts on the soil bacterial communities (i.e., the main underlying structure of the MNP soil bacterial community data). Hierarchical clustering successively merges together similar treatments or treatment groups until a single cluster is attained [38,39], providing a dendrogram of hierarchical similarity among the treatments. In the hierarchical clustering, average-link (defined as $\sum_{x\in C_i}\sum_{y\in C_j} d\left(x,y\right) / \left|C_i\right|\left|C_j\right|$ for two clusters C_i and C_j) was used as inter-cluster distance measure since it is robust to outliers [38,39]. An advantage of the hierar-

chical clustering based on the L_1 distance matrix is that $L_1 < 0.5$ represents a meaningful threshold to cut a dendrogram (hierarchical tree) into suitable meta-clusters, whereas a threshold above 0.5 will lead to clustering of treatments that are more dissimilar than similar [32].

Consistency analysis of MNP impact

A recently developed distance correlation [41] was used to assess the consistency of MNP impacts on soil bacterial communities summarized in different taxonomic levels. It is noted that each taxonomic level contained a range of taxa, representing a set of vectors where the number of components (i.e., dimensionality) could be much larger than the total treatments (e.g., there are 446 bacterial taxa in genus level and 31,624 in OTU levels). Therefore, conventional correlation analyses such as Pearson correlation [46] and canonical correlation [47] are not applicable for analyzing the consistency between different taxonomic levels. For the above problem, distance correlation is particularly suitable, which quantifies the similarity in treatment distance for different taxonomic levels. In distance correlation analysis [41], a new matrix $\boldsymbol{A}$ is first constructed from the distance matrix $\boldsymbol{a}$ that was calculated at taxonomic level T_A as $A_{ij} = a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}$, in which $\bar{a}_{i\cdot}$, $\bar{a}_{\cdot j}$ and $\bar{a}_{\cdot\cdot}$ are the means of the i-th row, j-column, and the entire matrix $\boldsymbol{a}$, respectively. Similarly, another matrix $\boldsymbol{B}$ can be derived from the distance matrix $\boldsymbol{b}$ calculated at taxonomic level T_B. The distance variances for taxonomic level T_A and T_B along with their distance covariance can be defined as:

$$\begin{aligned} \mathrm{V}^2(\boldsymbol{A}) &= \frac{1}{n^2}\sum_{i,j=1}^{n} A_{ij}^2, \\ \mathrm{V}^2(\boldsymbol{B}) &= \frac{1}{n^2}\sum_{i,j=1}^{n} B_{ij}^2, \\ \mathrm{V}^2(\boldsymbol{A},\boldsymbol{B}) &= \frac{1}{n^2}\sum_{i,j=1}^{n} A_{ij}B_{ij} \end{aligned} \tag{9}$$

where n identifies the dimensionality of matrix $\boldsymbol{A}$ and $\boldsymbol{B}$. Accordingly, the distance correlation between taxonomic level T_A and T_B is given by:

$$\mathrm{R} = \frac{\mathrm{V}(\boldsymbol{A},\boldsymbol{B})}{\sqrt{\mathrm{V}(\boldsymbol{A})\,\mathrm{V}(\boldsymbol{B})}} \tag{10}$$

An important property of the above distance correlation is that it becomes zero if and only if the random variables (e.g., different taxonomic levels) are statistically independent [41].

Results and Discussion

Bipartite graphs between MNP treatments and bacteria responses

For taxonomic levels from genus to phylum, soil bacterial taxa for which the range of 95th–5th percentile with respect to relative abundance (across all the quadruplicated treatments) was no less than $10/n$ (n denotes the total number of bacterial taxa at a given taxonomic level) were identified as treatment susceptible. Interrelationships between the 15 treatments and the responses (quantified as relative abundance) of bacterial taxa were illustrated as the bipartite graphs [33-35] established in Figure 2, Figure 3 and Figure 4, as well as Figure 5, Figure 6 and Figure 7. In the bipartite graphs (Figures 2–7), the relative abundances of the soil bacterial taxa identified as treatment susceptible were re-closed (i.e., rescaled such that the relative abundances sums up to unity for each treatment), and then averaged for the quadruplicate of each treatment. It is also noted that, for the genus level, the threshold of 95th–5th percentile range was increased to $50/n$ (where n = 446 denotes the total number of bacterial taxa at genus level) in order to avoid cluttering the bipartite graph.

In the bipartite graphs (Figures 2–7), soil bacterial taxa identified as treatment susceptible are denoted by the bars (nodes) on the right side, with the bar height proportional to their total relative abundance over the 15 treatments. For example, *Actinomycetales* is abundant in all the 15 treatments with an average relative abundance of 52% (Figure 2), while, for a specific treatment with ZnO MNPs at the dose of 0.1 mg/g (soil) and exposure time of 60 days, its relative abundance is 50% (Figure 4). Each taxon bar is further split into sub-bars representing its distribution (in terms of relative abundance) across the 15 treatments. The bars on the left side of the bipartite graphs (Figures 2–7) identify the 15 treatments with the bar height indicating the total relative abundance of the taxa identified for the treatments. In the present work, such total relative abundance was 100% for each treatment since the soil bacterial taxa identified as treatment susceptible were re-closed.

The established bipartite graphs can be useful for inspecting soil bacterial taxa that are susceptible to MNPs along with their relative abundance for each treatment. For example, the bipartite graph (Figure 2) for order level shows that only 14 of the 53 bacterial taxa were identified as treatment susceptible, based on the threshold of 95th–5th percentile range $\geq 10/n$ in relative abundance. It is also noted that relative abundances of the above order bacterial taxa vary significantly from 1% to 52%. Moreover, bipartite graphs (Figures 2–7) allow bidirectional exploration of the soil bacterial community data for detailed information about a specific treatment (i.e., bacterial taxon → treatment) or a taxon at different taxonomic levels (i.e., treatment →

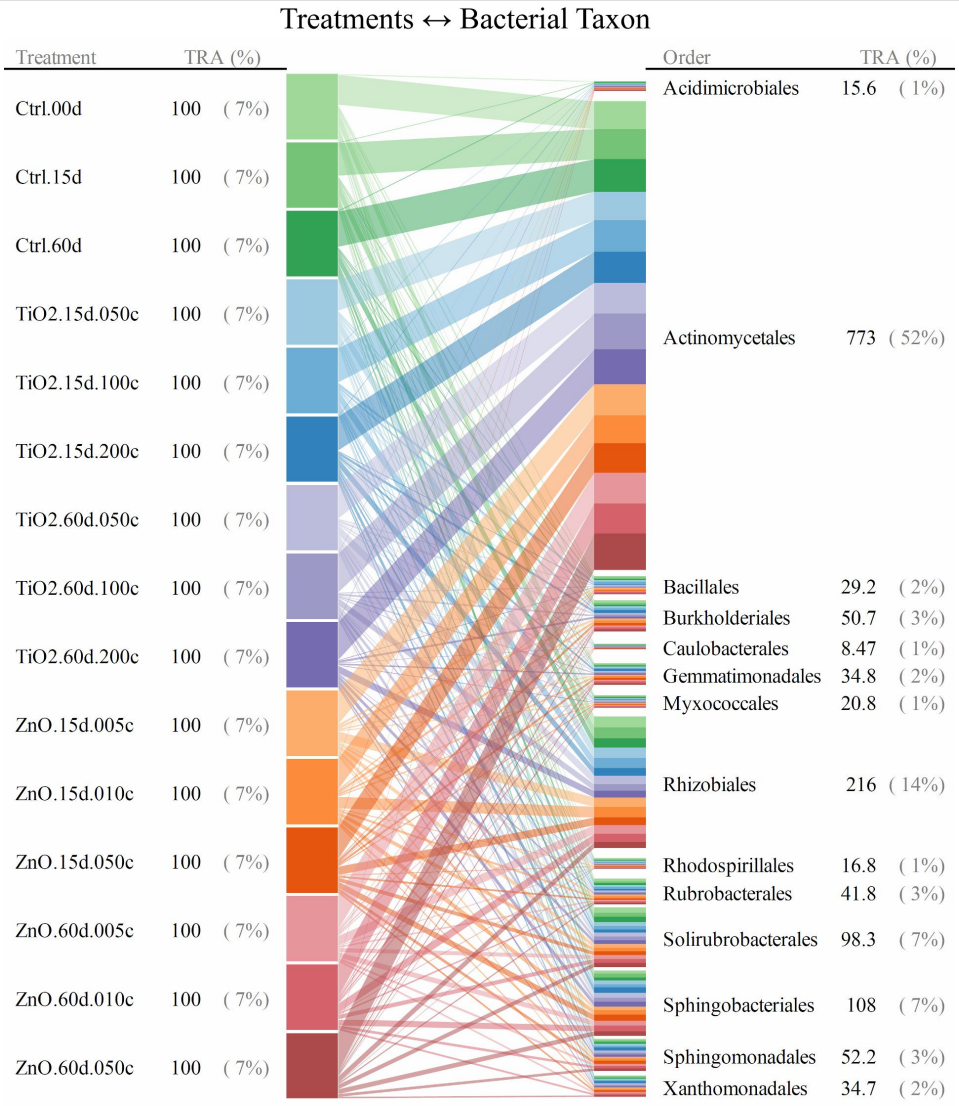


Figure 2: Bipartite graph for MNP-bacteria interrelationships at order level. Soil bacteria taxa identified for the above graph are for the 95th–5th percentile range ≥ 10/*n* (*n* denotes the total number of bacterial taxa at a given taxonomic level) in relative abundance. TRA denotes total relative abundance. The treatments are labelled as "--.##d.##c", where "--" identifies the treatment type (i.e., TiO_2 MNP, ZnO MNP, or control), "##d" denotes exposure time of ## day, and "##c" represents exposure dose of ##×10^{-2} mg/g (soil).

bacterial taxon). For example, in the direction of bacterial taxon → treatment, focusing the bipartite graph of order level on *Rhizobiales* (Figure 3) revealed that, compared to the controls, the exposure to high TiO_2 (2.0 mg/g (soil)) or ZnO (0.5 mg/g (soil)) MNP doses for 15 and 60 days reduced the relative abundance of *Rhizobiales* by up to 32% and 35%, respectively. Such relative abundance reductions of *Rhizobiales* indicate that the two MNPs at high dose could stress the *Rhizobiales*. Studies have reported that *Rhizobiales* is an important order taxon containing N_2-fixing bacteria that are able to symbiotically associate with legume roots to fix atmospheric N_2 into ammonium for plant growth [48]. One can also explore the effect of treatment on bacterial taxa (treatment → bacterial taxon). For example, the relative abundances of the 14 order taxa displayed in Figure 4 illustrates treatment with ZnO MNPs at the dose of 0.1 mg/g (soil) and exposure time of 60 days, showing that *Actinomycetales* and *Caulobacterales* are the bacterial taxa of the highest (49.7%) and lowest (0.5%) relative abundance, respectively. The above bidirectional exploration using bipartite graphs can be conducted along the taxonomic hierarchy (Figures 5–7) to identify informative MNP-bacteria interrelationships at different levels (e.g., drill down to genus level or roll up to phylum level).

Contribution biplots generated by log-ratio analyses

Results of the log-ratio analysis (LRA) [32,36,37] for the soil bacterial community dataset are illustrated in the contribution

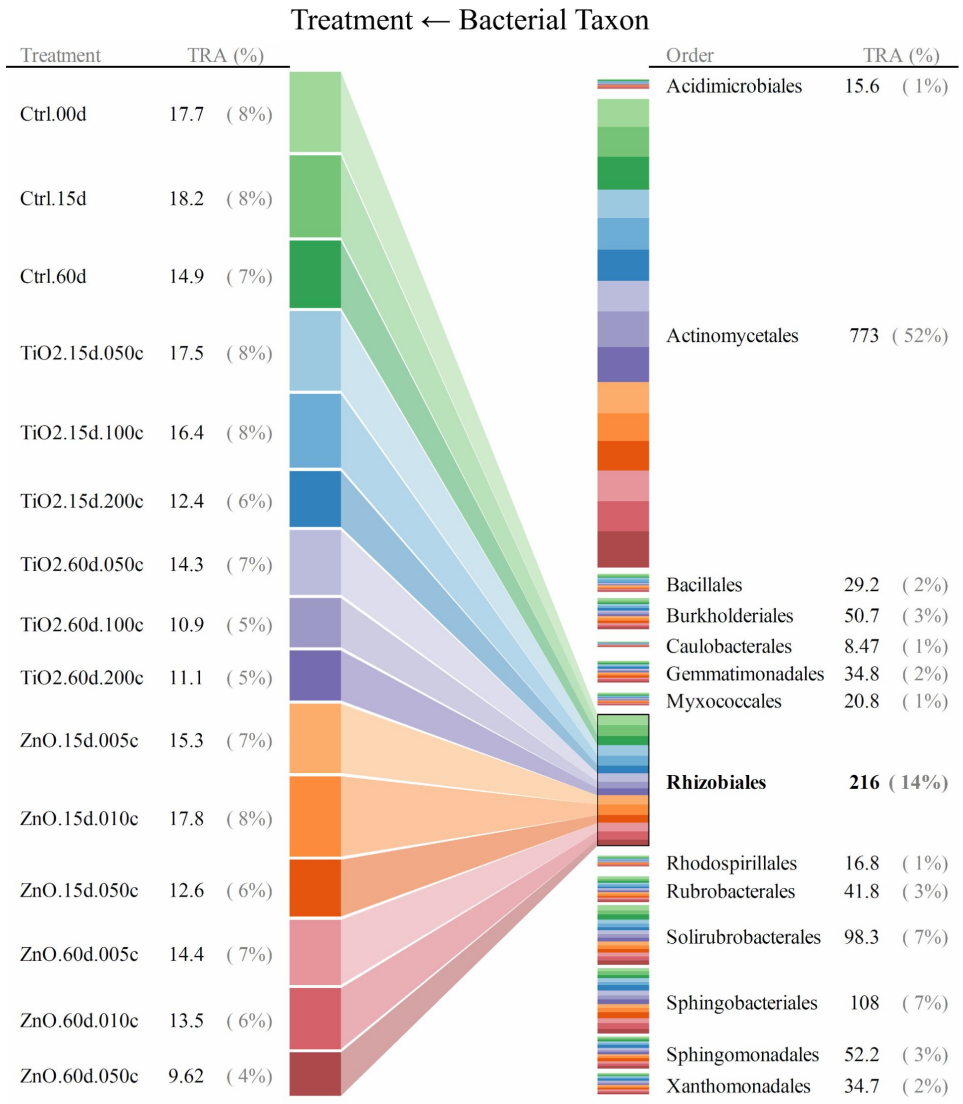


Figure 3: Bipartite graph for bacterial taxon → MNP treatment at order level. Soil bacteria taxa identified for the above graph are for the 95th–5th percentile range ≥ 10/*n* (*n* denotes the total number of bacterial taxa at a given taxonomic level) in relative abundance. TRA denotes total relative abundance. The treatments are labelled as in Figure 2.

biplots [32,36] given in Figure 8, which display treatments and bacterial taxa jointly in the same maps. In a contribution biplot (Figure 8), the treatments (samples) are displayed as scatter points using the first two principal row coordinates (i.e., dim1 and dim2) provided by LRA, while the bacterial taxa contributions (variables) were added as vectors (from the origin) scaled to fit into the same range of the principal row coordinates. The scatter plots maintain the distance between different treatments in the complete datasets to a reasonable approximation. The vectors, on the other hand, are indicative of both the contribution (variance across all the treatments) of the bacterial taxa (via vector length) and the correlations between them (via angles between the vectors).

The above configuration of biplots (Figure 8) that display bacterial taxa according to their contributions (variances) to the principal row coordinates allows a visual separation of determinant ones from the large number of bacterial taxa. The correlations between bacterial taxa can be readily inferred from the biplots (Figure 8) along with their contribution to treatment separation. For example, a number of bacterial taxa of significant contribution (vectors of large length) to treatment separation are outlined in each biplot (Figure 8). It is noted that, for order level, *Rhizobiales* is a primary bacterial taxon that separates TiO_2 and ZnO MNPs from the controls at the high dose. The biplot for order level (Figure 8d) demonstrate that the MNP treatments at high dose had lower relative abundances of *Rhizo-*

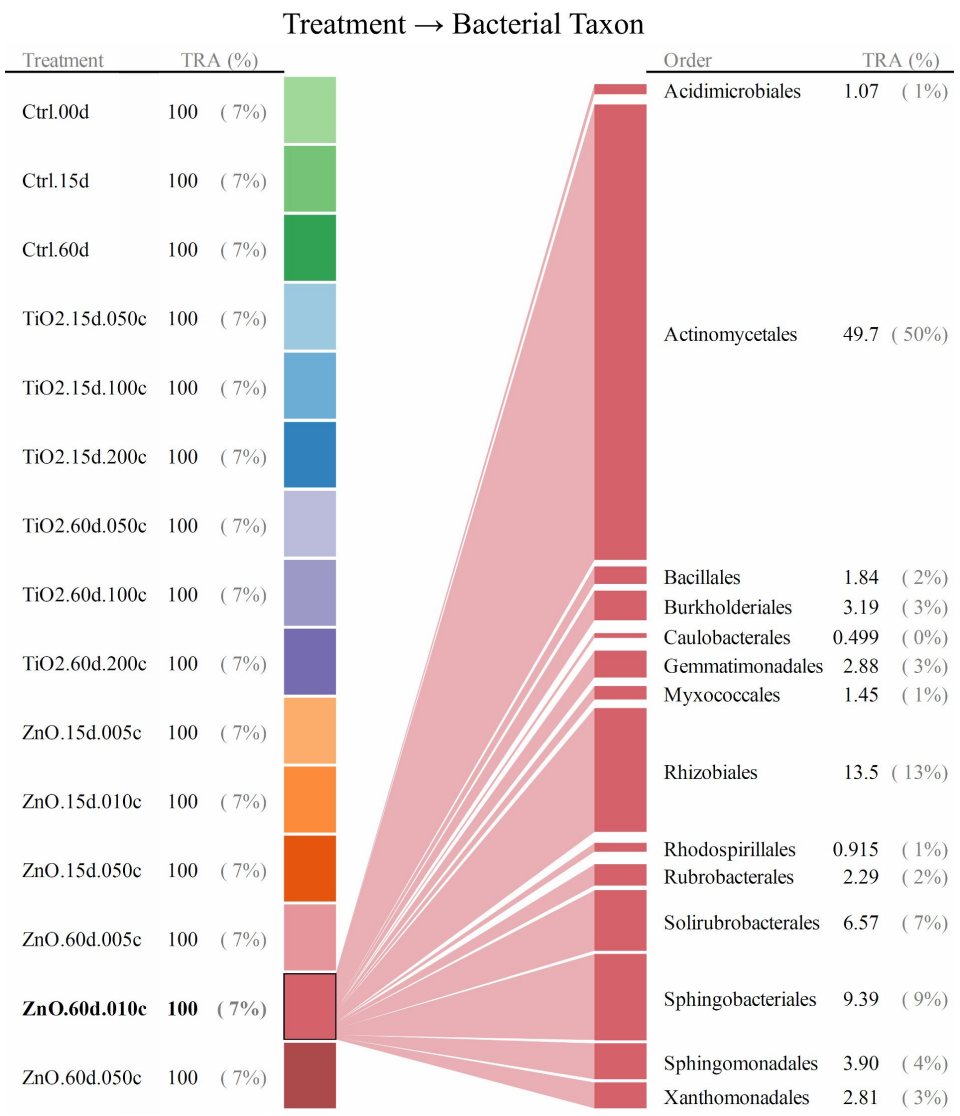


Figure 4: Bipartite graph for MNP treatment → bacterial taxon at order level. Soil bacteria taxa identified for the above graph are for the 95th–5th percentile range ≥ 10/*n* (*n* denotes the total number of bacterial taxa at a given taxonomic level) in relative abundance. TRA denotes total relative abundance. The treatments are labelled as in Figure 2.

biales compared to controls. The above observation (Figure 8d) is consistent with the bipartite exploration result of order level (Figure 3). Moreover, due to the subcompositional coherence property of LRA [32,36,37], the removal of some bacterial taxa will not change the correlations between the remaining bacterial taxa. For example, the biplot for phylum level remains essentially the same with (Figure 8f) or without (Figure 9) the *Gemmatimonadetes*.

The biplots given in Figure 8 also provide useful information regarding the main underlying structures in the soil bacterial community dataset. For example, the biplots for OTU, genus, and family levels (Figure 8a–c) demonstrate that there are two groups of MNP treatments (corresponding primarily to 15 days and 60 days exposure, respectively) separated from the controls. However, as the taxonomic hierarchy increases to order, class, and phylum levels (Figure 8d–f), the treatments are more dispersed (less separable). This indicates that the above taxonomic levels are too high to differentiate the impact of MNPs on soil bacterial communities. In other words, family, as the highest taxonomic level that maintains the main underlying structure of the soil bacterial community data, could be a suitable taxonomic level for MNP impact assessment. Indeed, the distance correlation (Figure 10a) calculated between log-transformed relative abundance of bacterial taxa at different taxonomic levels revealed that the six bacterial taxonomic levels can

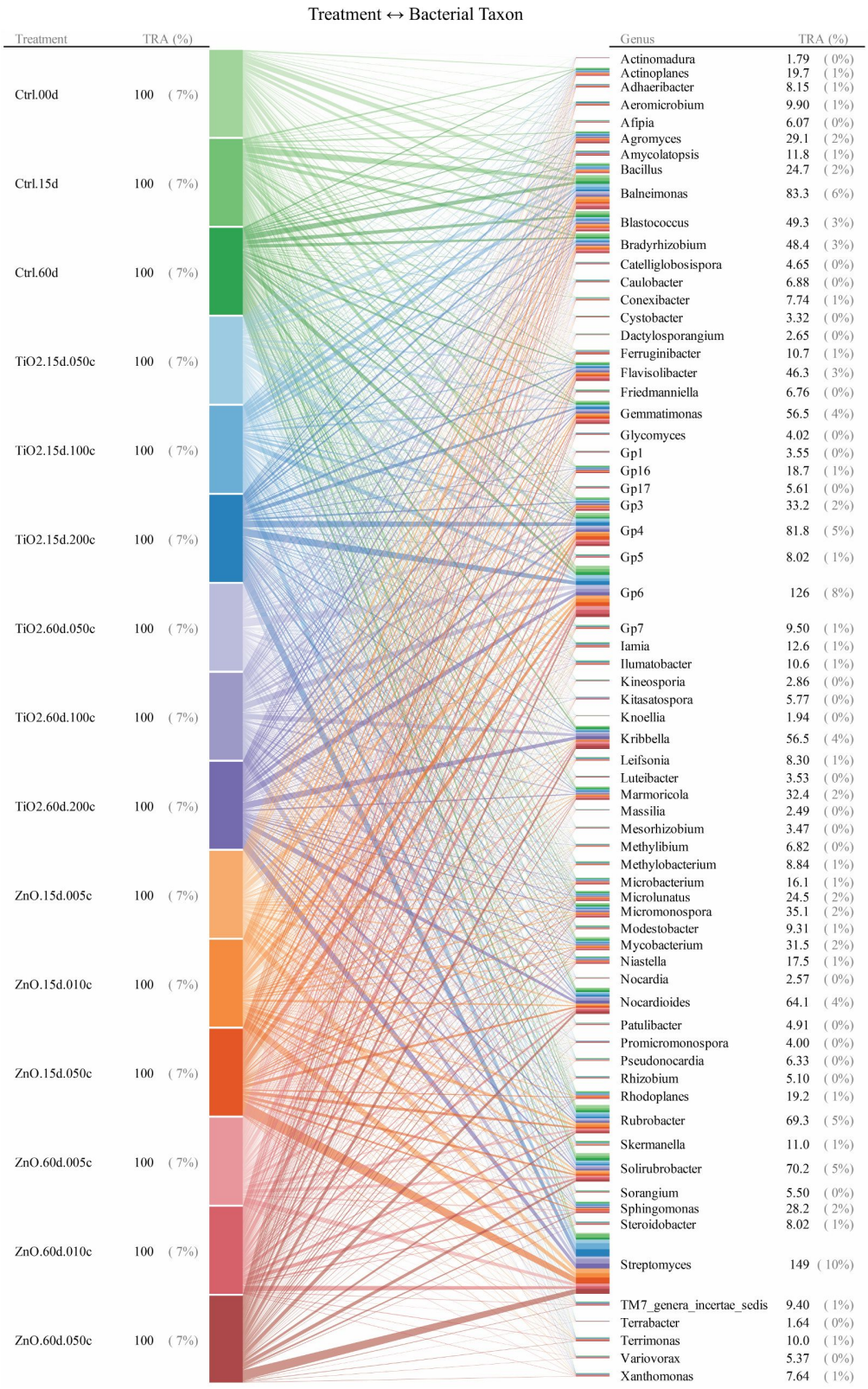


Figure 5: Bipartite graphs for MNP-bacteria interrelationships at genus levels. At genus level, soil bacteria taxa were identified according to an increased threshold of 95th–5th percentile range ≥ 50/*n* (*n* denotes the total number of bacterial taxa at genus level) to avoid cluttering the bipartite graph. TRA denotes total relative abundance. The treatments are labelled as in Figure 2.

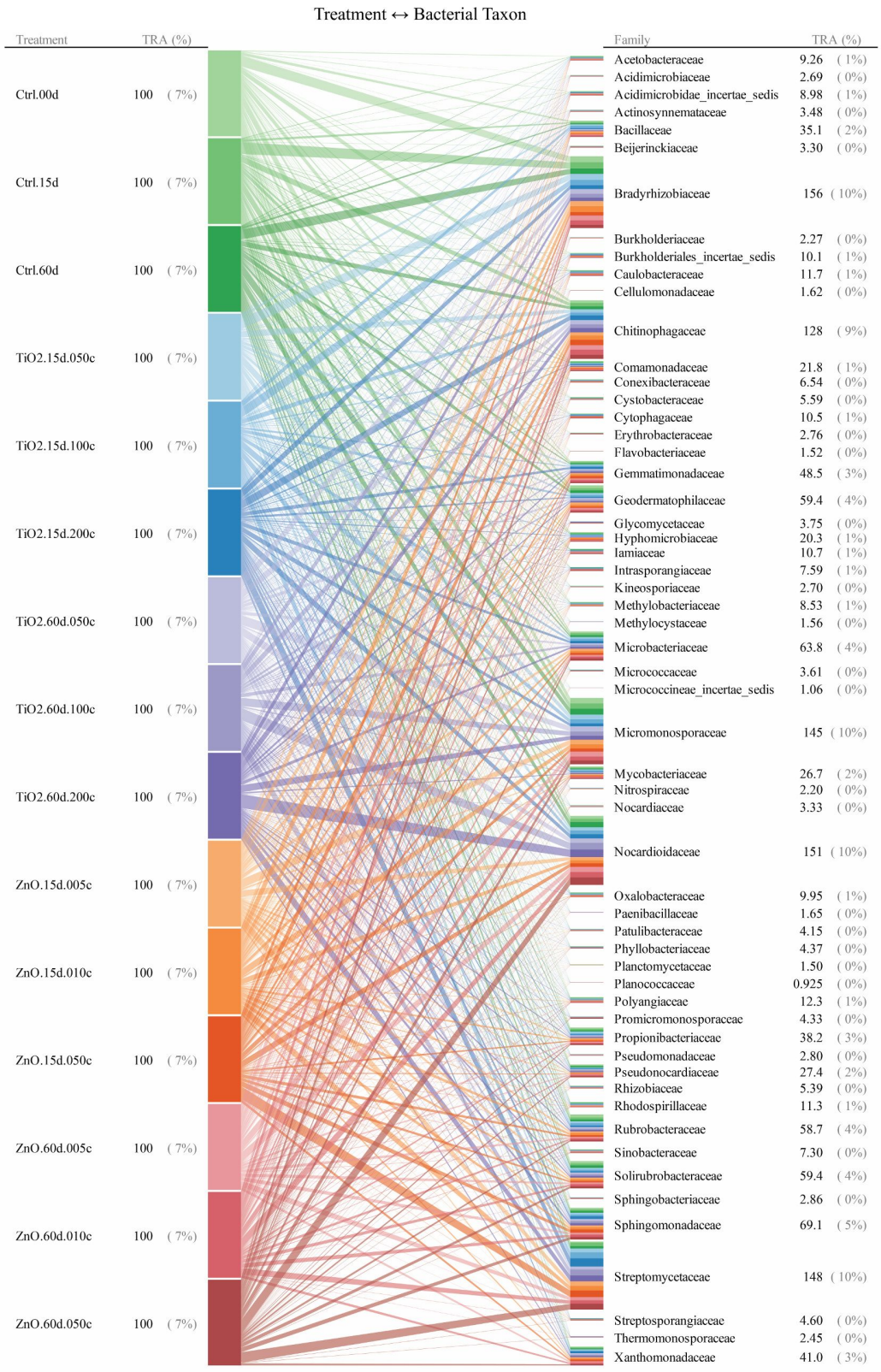


Figure 6: Bipartite graphs for MNP-bacteria interrelationships at family levels. Soil bacteria taxa identified for the above graph are for the 95th–5th percentile range ≥ 10/*n* (*n* denotes the total number of bacterial taxa at a given taxonomic level) in relative abundance. TRA denotes total relative abundance. The treatments are labelled as in Figure 2.

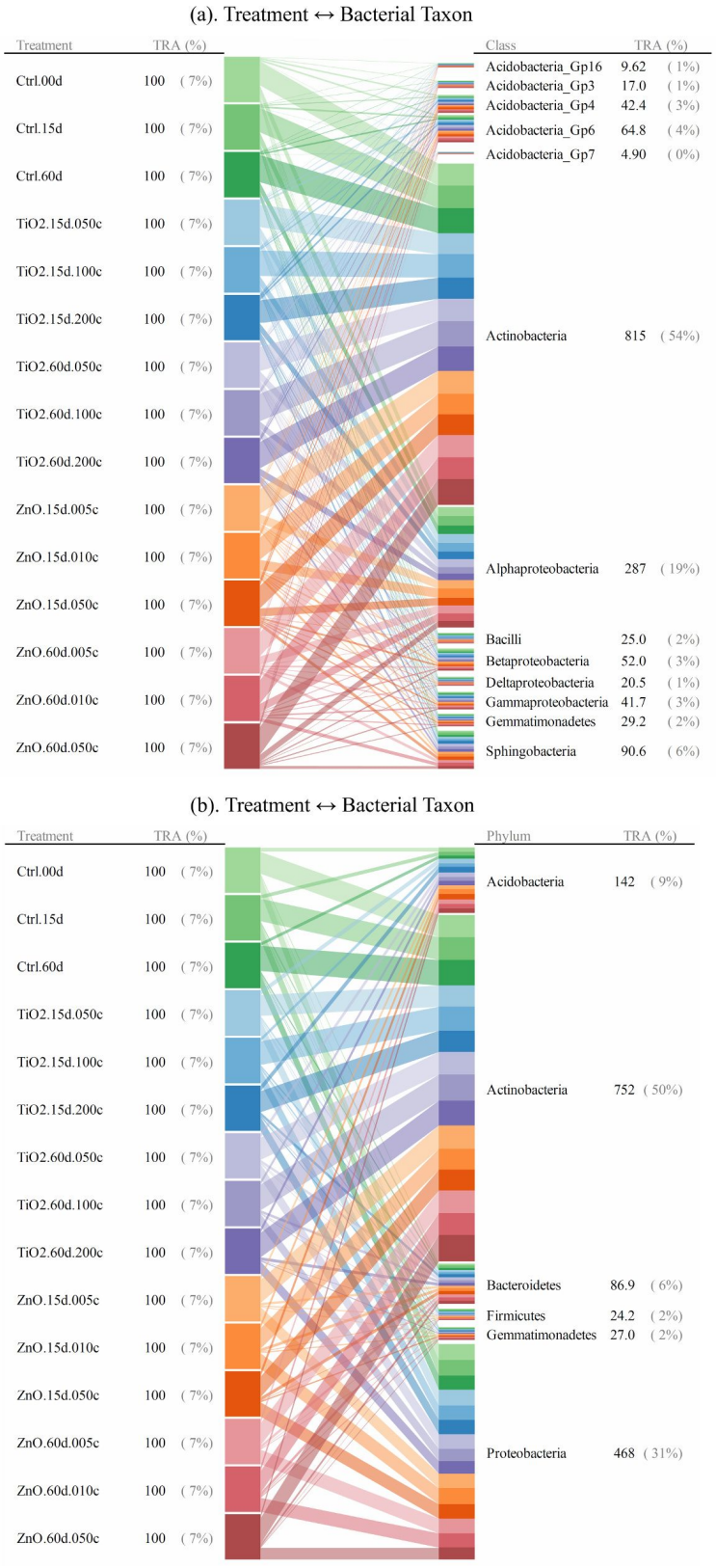


Figure 7: Bipartite graphs for MNP-bacteria interrelationships at (a). class, and (b). phylum levels. Soil bacteria taxa identified for the above graph are for the 95th–5th percentile range ≥ 10/*n* (*n* denotes the total number of bacterial taxa at a given taxonomic level) in relative abundance. TRA denotes total relative abundance. The treatments are labelled as in Figure 2.

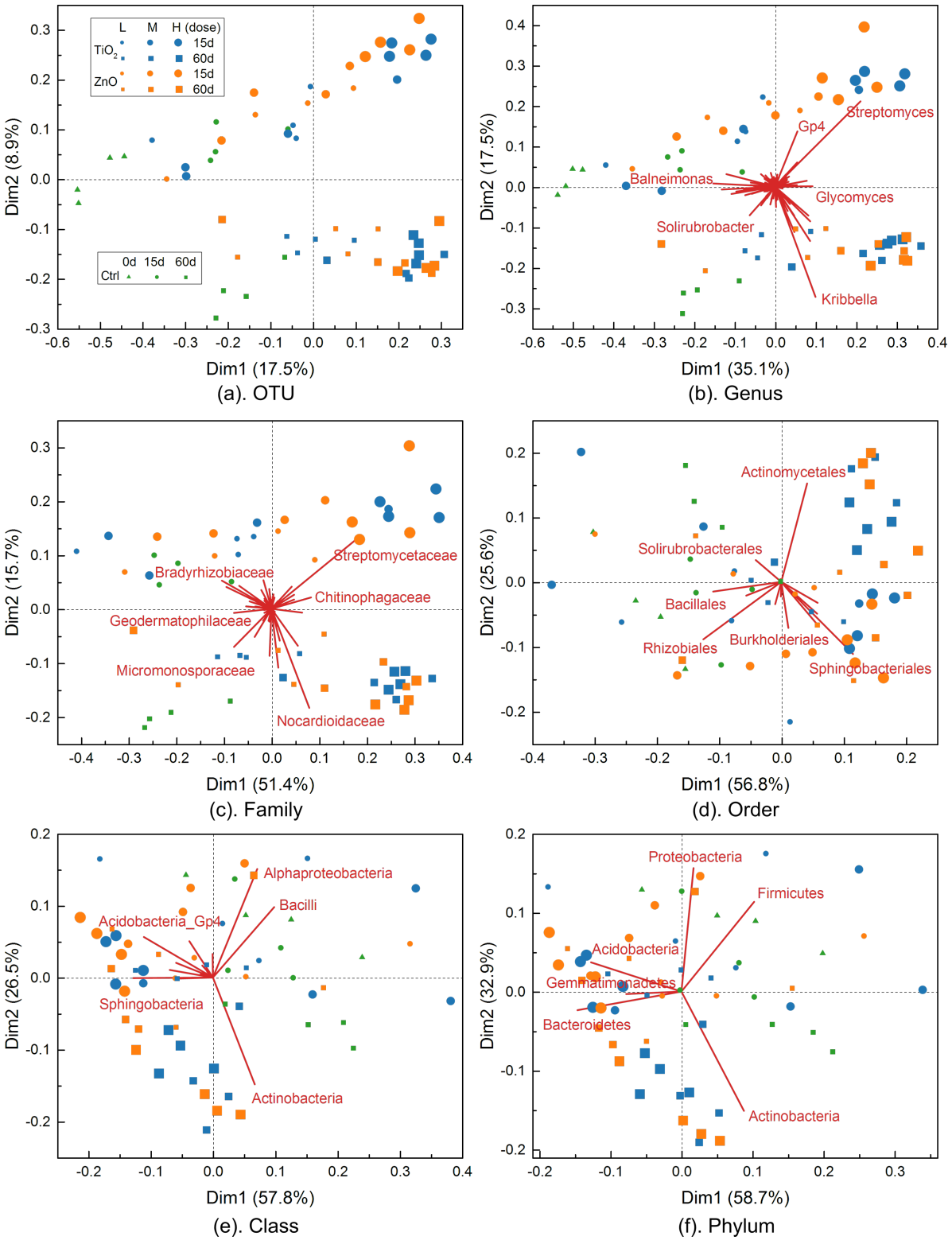


Figure 8: Contribution biplots generated by log-ratio analyses for taxonomic levels from OTU to phylum. The total variance in the complete datasets as accounted by the two principal row coordinates (dim1 and dim2) is provided in the appended parentheses. The contribution vectors (bacterial taxa) were scaled to fit into the scatter plots of the treatments. For the treatments (TiO_2 and ZnO MNPs and controls (Ctrl)), the exposure time is denoted by "##d" with "L", "M", "H" corresponding to doses of 0.5, 1.0, 2.0 mg/g (soil) and 0.05, 0.1, and 0.5 mg/g (soil) for TiO_2 and ZnO MNPs, respectively. The contribution vectors are omitted for the plot of OTU level to avoid cluttering the plot.

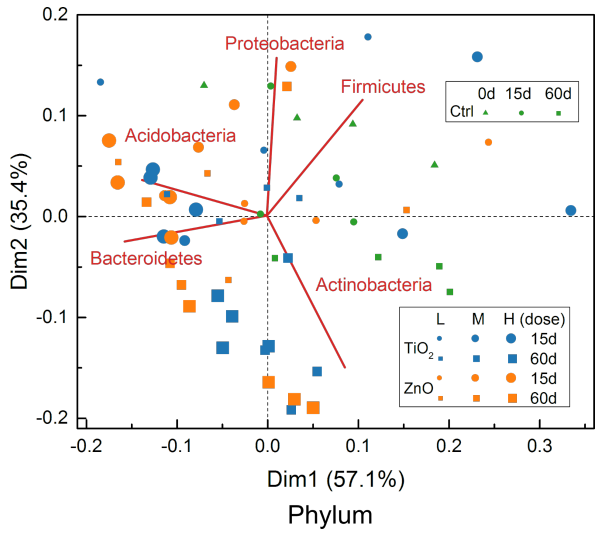


Figure 9: Contribution biplot for phylum level with *Gemmatimonadetes* removed. The treatments are labelled as in Figure 8.

(a). LR Distance

	Phylum	*Class*	*Order*	*Family*	*Genus*	*OTU*
Phylum	1.00	0.96	0.96	0.79	0.76	0.69
Class	0.96	1.00	0.95	0.81	0.80	0.75
Order	0.96	0.95	1.00	0.87	0.83	0.76
Family	0.79	0.81	0.87	1.00	0.94	0.86
Genus	0.76	0.80	0.83	0.94	1.00	0.96
OTU	0.69	0.75	0.76	0.86	0.96	1.00

(b). L_1 Distance

	Phylum	*Class*	*Order*	*Family*	*Genus*	*OTU*
Phylum	1.00	0.99	0.97	0.83	0.81	0.72
Class	0.99	1.00	0.98	0.86	0.84	0.74
Order	0.97	0.98	1.00	0.90	0.87	0.76
Family	0.83	0.86	0.90	1.00	0.98	0.83
Genus	0.81	0.84	0.87	0.98	1.00	0.86
OTU	0.72	0.74	0.76	0.83	0.86	1.00

Figure 10: Distance correlation between taxonomic levels from OTU to phylum using (a) log-ratio (LR) distance and (b) L_1 distance.

be divided into two groups of high consistency. The first group contains phylum, class, and order levels with average distance correlation of 0.96, while family, genus, and OTU formed a second group of average distance correlation of 0.92. Compared to the high intra-group consistencies, the average distance correlation between the two groups dropped to 0.78. The above distance correlation analysis again suggests that family could be a suitable taxonomic level for MNP impact assessment as it is the highest taxonomic level of good consistency to the OTU level. The distance correlation analysis (Figure 10a) also indicates that, in general, levels closer in the taxonomic hierarchy are more consistent with each other. Finally, it is also noted that, the two principal row coordinates (i.e., dim1 and dim2) of the biplots for OTU, genus, and family levels (Figure 8a–c) account for <80% of the total variance in the complete datasets (which can be considered as the information preserved by the biplots). The above explained variance increased to >80% in the biplots for order, class, and phylum levels, indicating that the inter-treatment distances were closely maintained in these biplots [32].

Multidimensional scaling maps

The L_1 distance matrix calculated for the 15 treatments (in quadruplicate) at the OTU level is illustrated in Figure 11 as a hierarchically clustered heatmap [32,38,39] established using average-link [32,38,39]. According to the recommended threshold of $L_1 < 0.5$ [32], three meta-clusters were identified from the heatmap with Cluster II and III mainly comprised of MNPs exposed for 15 and 60 days and Cluster I formed by the remainder (Figure 11). Characterization of Cluster II and III by exposure time is consistent with the contribution biplot for OTU level (Figure 8a) and previous studies [18,19] that also demonstrated significant impact of exposure period on soil bacterial communities. In addition, all high doses of TiO_2 (2.0 mg/g (soil)) and ZnO (0.5 mg/g (soil)) MNPs are found in Cluster II and III, while all controls are found in Cluster I (Figure 8), indicating that both MNPs altered soil bacterial communities at relatively high dose.

Based on the distance matrix calculated for the OTU level, a 2D map (Figure 12) was established using nonmetric multidimensional scaling (NMDS) for direct presentation of inter-treatment (in quadruplicate) distances. The NMDS established for the OTU level (Figure 12) agrees well with the hierarchical clustering result (Figure 11) with the treatments in Cluster II and III located mainly in the first and fourth quadrants, while the treatments contained in Cluster I are scattered in the second and third quadrants. In addition, the NMDS (Figure 12) further demonstrates that there is large variance within the replicates of each treatment, which obscures the inter-treatment distance relationships. The NMDS for OTU level (Figure 12) is also similar to the contribution biplot (Figure 8a) generated for the same level. Although the above NMDS (Figure 12) had a good *stress* of 14.85% [49]; however, *stress* is usually an over-optimistic measure of preserved/lost information [32] compared to the percent of explained variance (which is not defined for NMDS).

The obscureness caused by the quadruplicate of each treatment is avoided in the NMDS using the average-link of L_1 distance between different treatments (Figure 13). Without the interference of replicates, the simplified NMDS clearly shows that the high dose of ZnO (0.5 mg/g (soil)) and TiO_2 MNPs (2.0 mg/g

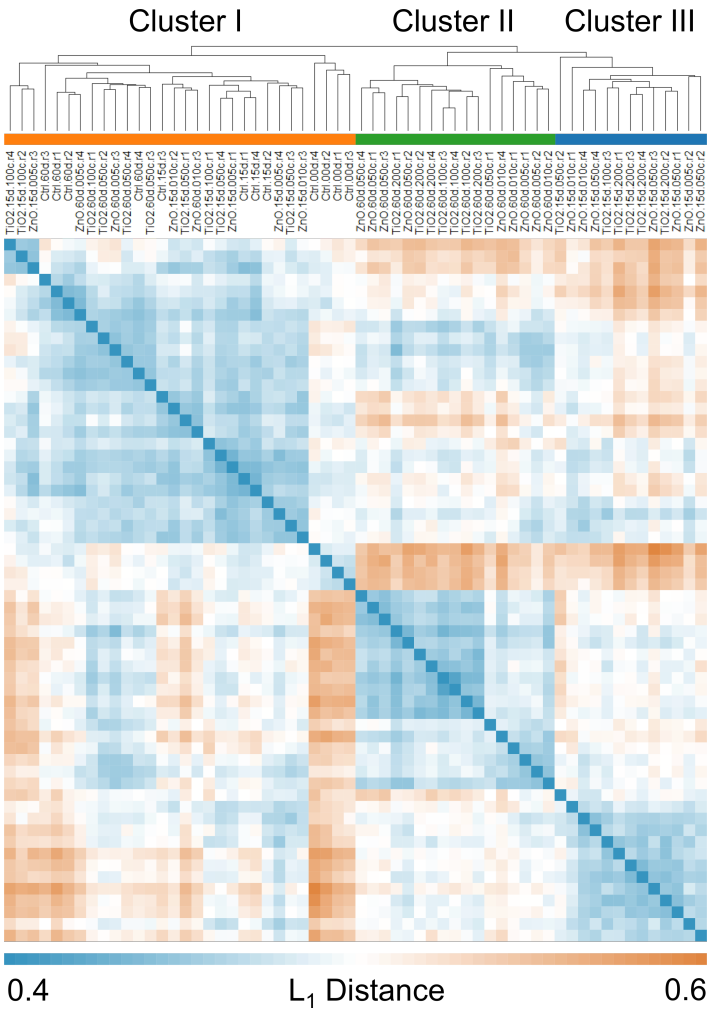


Figure 11: Clusters of treatments obtained via hierarchical clustering based on their L_1 distances calculated at OTU level. Three meta-clusters were identified according to the recommended threshold of $L_1 < 0.5$ [34]. The treatments are labelled as in Figure 2 with an additional ".r#" identifying different replicates.

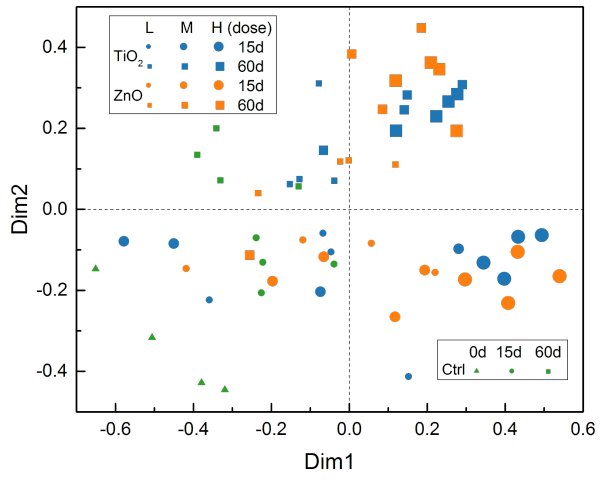


Figure 12: Nonmetric multidimensional scaling (NMDS) for OTU level (*stress* = 14.85%). The treatments are labelled as in Figure 8.

(soil)) have significant impacts on soil bacterial communities at the OTU level (Figure 13a) as they are distant from the controls. Similar behavior of the ZnO and TiO_2 MNPs is also observed in the simplified NMDSs (Figure 13b,c) established for the genus and family levels. However, as the taxonomic hierarchy increased to order, class, and phylum levels, the treatments (controls and MNPs) disperse and mix with each other on the NMDSs (Figure 13d–f), signifying that the taxonomic levels are too high to differentiate the impact of MNPs on soil bacterial communities. The above observations with the NMDSs are consistent with those from the contribution biplots (Figure 8) generated by LRA. In addition, the distance correlations calculated between the six different taxonomic levels based on L_1 distance (Figure 10b) are also similar to those obtained based on the log-transformed relative abundance of bacterial taxa. In the NMDSs, a number of bacterial taxa of significant gradients (vectors of large length) are outlined (Figure 13), indicating that

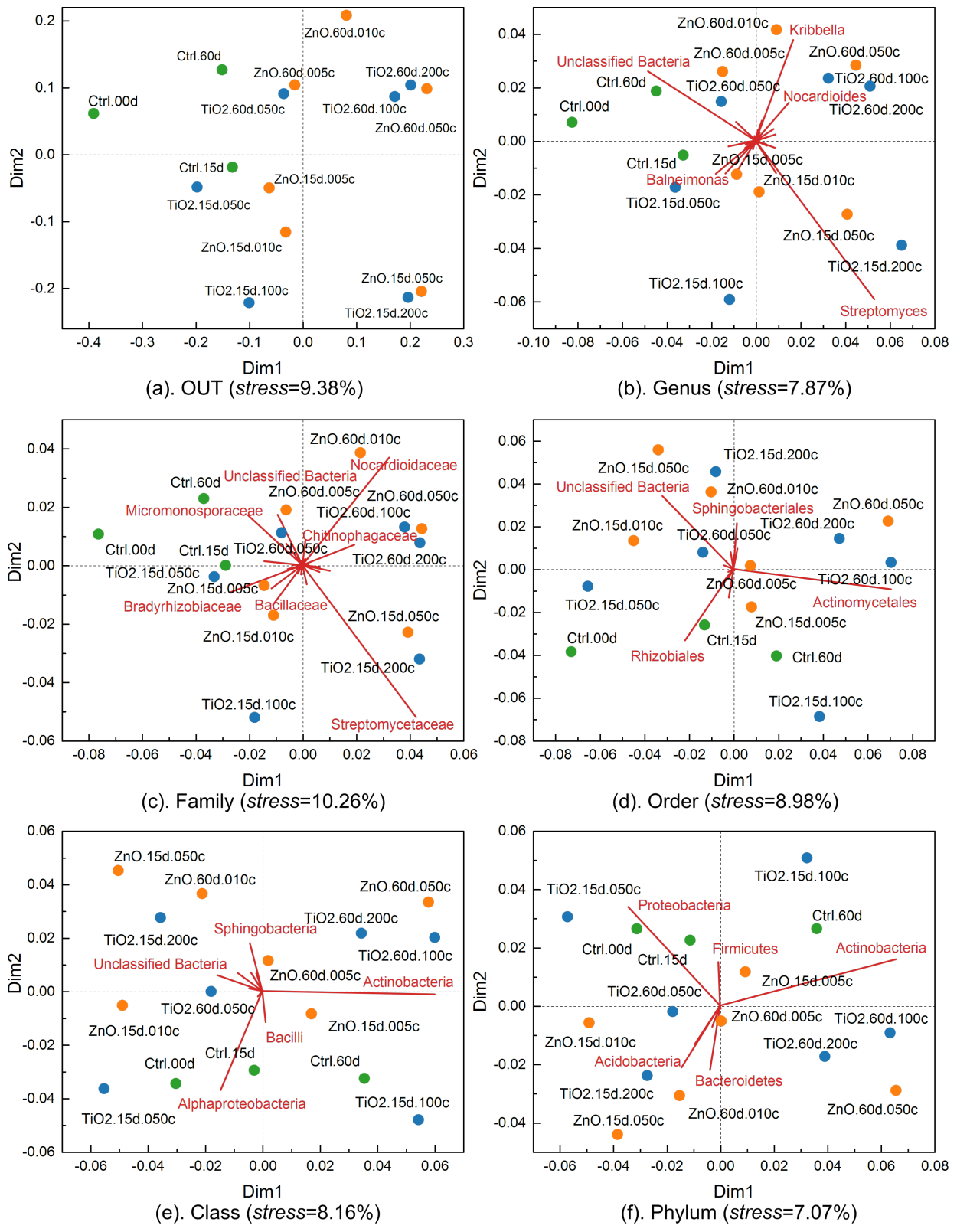


Figure 13: Simplified nonmetric multidimensional scaling (NMDS) for taxonomic levels from OTU to phylum. The gradient vectors of bacterial taxa were scaled to fit into the scatter plots of the treatments. The gradient vectors are omitted for the plot of OTU level to avoid cluttering the plot.

their relative abundance varies significantly across the treatments [32]. However, these gradient vectors are not directly related to the contributions of the corresponding bacterial taxa to treatment separation and the NMDSs are not subcompositionally coherent [32,36,37].

Conclusion

The impact of manufactured nanoparticles (MNPs) on soil bacterial communities was analyzed using a series of visual exploration approaches. The analyzed soil bacterial community dataset contained the counts/relative abundance of a set of hierarchical taxa (at operational taxonomic unit (OTU), genus, family, order, class, and phylum levels) measured for 15 soil treatments with exposure to TiO_2 (at dose of 0.5, 1.0, and 2.0 mg/g (soil)) and ZnO (at dose of 0.05, 0.1, and 0.5 mg/g (soil)) MNPs for periods of 15 and 60 days or 0, 15, and 60 days without exposure to MNPs (i.e., controls). Bipartite graphs were established to illustrate the inter-relationships between MNPs and responses of bacterial taxa. The bipartite graphs were shown to be useful for identifying, from numerous MNP-bacteria interrelationships, those that reflect significant change in relative abundance of bacterial taxa. Contribution biplots of subcompositional coherence property were generated by log-ratio analysis (LRA) [32,36,37], providing joint displays for the separation (distribution) of treatments and the contribution (variance) of bacterial taxa. The LRA contribution biplots and two-dimensional maps, constructed from the dataset using hierarchical clustering and nonmetric multi-dimensional scaling (NMDS), also demonstrated that high doses of ZnO and TiO_2 MNPs caused significant compositional changes in soil bacterial communities. The LRA contribution biplots and the simplified NMDSs, together with the distance correlation analysis for the consistency between MNP impacts summarized at taxonomic levels, suggest that family could be a suitable taxonomic level for MNP impact assessment. Utilization of the above visual data exploration approaches can be particularly useful if deployed as a web-based platform for rapid assessment of the impact of MNPs on bacterial soil communities, as well as other ecological systems to guide the development of safe-by-design nanomaterials.

Acknowledgements

This material is based upon work supported by the National Science Foundation and the Environmental Protection Agency under Cooperative Agreement Number DBI-0830117. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Environmental Protection Agency. This work has not been subjected to EPA review and no official endorsement should be inferred.

References

1. Murty, B. S.; Shankar, P.; Raj, B.; Rath, B. B.; Murday, J. *Textbook of Nanoscience and Nanotechnology;* Springer: Berlin, Germany, 2013.
2. Guo, Z.; Tan, L. *Fundamentals and Applications of Nanomaterials,* 1st ed.; Artech House Publishers: Boston, MA, U.S.A., 2009.
3. The Wilson Center. Inventory Finds Increase in Consumer Products Containing Nanoscale Materials. http://www.nanotechproject.org/cpi/.
4. Nel, A.; Xia, T.; Mädler, L.; Li, N. *Science* **2006,** *311,* 622–627. doi:10.1126/science.1114397
5. Xia, T.; Malasarn, D.; Lin, S.; Ji, Z.; Zhang, H.; Miller, R. J.; Keller, A. A.; Nisbet, R. M.; Harthorn, B. H.; Godwin, H. A.; Lenihan, H. S.; Liu, R.; Gardea-Torresdey, J.; Cohen, Y.; Mädler, L.; Holden, P. A.; Zink, J. I.; Nel, A. E. *Small* **2013,** *9,* 1428–1443. doi:10.1002/smll.201201700
6. Handy, R. D.; Shaw, B. J. *Health Risk Soc.* **2007,** *9,* 125–144. doi:10.1080/13698570701306807
7. Helland, A.; Scheringer, M.; Siegrist, M.; Kastenholz, H. G.; Wiek, A.; Scholz, R. W. *Environ. Sci. Technol.* **2007,** *42,* 640–646. doi:10.1021/es062807i
8. Hristozov, D. R.; Gottardo, S.; Critto, A.; Marcomini, A. *Nanotoxicology* **2012,** *6,* 880–898. doi:10.3109/17435390.2011.626534
9. Grieger, K. D.; Linkov, I.; Hansen, S. F.; Baun, A. *Nanotoxicology* **2012,** *6,* 196–212. doi:10.3109/17435390.2011.569095
10. Colvin, V. L. *Nat. Biotechnol.* **2003,** *21,* 1166–1170. doi:10.1038/nbt875
11. Gerber, C.; Lang, H. P. *Nat. Nanotechnol.* **2006,** *1,* 3–5. doi:10.1038/nnano.2006.70
12. Scown, T. M.; van Aerle, R.; Tyler, C. R. *Crit. Rev. Toxicol.* **2010,** *40,* 653–670. doi:10.3109/10408444.2010.494174
13. Cohen, Y.; Rallo, R.; Liu, R.; Liu, H. H. *Acc. Chem. Res.* **2013,** *46,* 802–812. doi:10.1021/ar300049e
14. Liu, H. H.; Cohen, Y. *Environ. Sci. Technol.* **2014,** *48,* 3281–3292. doi:10.1021/es405132z
15. Gottschalk, F.; Sonderer, T.; Scholz, R. W.; Nowack, B. *Environ. Sci. Technol.* **2009,** *43,* 9216–9222. doi:10.1021/es9015553
16. Klaine, S. J.; Alvarez, P. J. J.; Batley, G. E.; Fernandes, T. F.; Handy, R. D.; Lyon, D. Y.; Mahendra, S.; McLaughlin, M. J.; Lead, J. R. *Environ. Toxicol. Chem.* **2008,** *27,* 1825–1851. doi:10.1897/08-090.1
17. Tiede, K.; Hassellöv, M.; Breitbarth, E.; Chaudhry, Q.; Boxall, A. B. A. *J. Chromatogr. A* **2009,** *1216,* 503–509. doi:10.1016/j.chroma.2008.09.008
18. Ge, Y.; Schimel, J. P.; Holdena, P. A. *Appl. Environ. Microbiol.* **2012,** *78,* 6749–6758. doi:10.1128/AEM.00941-12
19. Ge, Y.; Schimel, J. P.; Holden, P. A. *Environ. Sci. Technol.* **2011,** *45,* 1659–1664. doi:10.1021/es103040t
20. Ju-Nam, Y.; Lead, J. R. *Sci. Total Environ.* **2008,** *400,* 396–414. doi:10.1016/j.scitotenv.2008.06.042
21. Navarro, E.; Baun, A.; Behra, R.; Hartmann, N. B.; Filser, J.; Miao, A.-J.; Quigg, A.; Santschi, P. H.; Sigg, L. *Ecotoxicology* **2008,** *17,* 372–386. doi:10.1007/s10646-008-0214-0
22. Baun, A.; Hartmann, N. B.; Grieger, K.; Kusk, K. O. *Ecotoxicology* **2008,** *17,* 387–395. doi:10.1007/s10646-008-0208-y
23. Asharani, P. V.; Lianwu, Y.; Gong, Z.; Valiyaveettil, S. *Nanotoxicology* **2011,** *5,* 43–54. doi:10.3109/17435390.2010.489207
24. Gagné, F.; Auclair, J.; Turcotte, P.; Fournier, M.; Gagnon, C.; Sauvé, S.; Blaise, C. *Aquat. Toxicol.* **2008,** *86,* 333–340. doi:10.1016/j.aquatox.2007.11.013
25. Du, J.; Wang, S.; You, H.; Zhao, X. *Environ. Toxicol. Pharmacol.* **2013,** *36,* 451–462. doi:10.1016/j.etap.2013.05.007
26. Tong, Z.; Bischoff, M.; Nies, L.; Applegate, B.; Turco, R. F. *Environ. Sci. Technol.* **2007,** *41,* 2985–2991. doi:10.1021/es061953l

27. Madsen, E. L. *Nat. Rev. Microbiol.* **2005,** *3,* 439–446. doi:10.1038/nrmicro1151
28. Falkowski, P. G.; Fenchel, T.; Delong, E. F. *Science* **2008,** *320,* 1034–1039. doi:10.1126/science.1153213
29. Xue, K.; Wu, L.; Deng, Y.; He, Z.; Van Nostrand, J.; Robertson, P. G.; Schmidt, T. M.; Zhou, J. *Appl. Environ. Microbiol.* **2013,** *79,* 1284–1292. doi:10.1128/AEM.03393-12
30. He, S.; Feng, Y.; Ren, H.; Zhang, Y.; Gu, N.; Lin, X. *J. Soils Sediments* **2011,** *11,* 1408–1417. doi:10.1007/s11368-011-0415-7
31. Nogueira, V.; Lopes, I.; Rocha-Santos, T.; Santos, A. L.; Rasteiro, G. M.; Antunes, F.; Gonçalves, F.; Soares, A. M. V. M.; Cunha, A.; Almeida, A.; Gomesa, N. N. C. M.; Pereirae, R. *Sci. Total Environ.* **2012,** *424,* 344–350. doi:10.1016/j.scitotenv.2012.02.041
32. Greenacre, M.; Primicerio, R. *Multivariate Analysis of Ecological Data;* Fundación BBVA: Bilbao, Spain, 2013.
33. Dormann, C. F.; Strauss, R. *Methods Ecol. Evol.* **2014,** *5,* 90–98. doi:10.1111/2041-210X.12139
34. Dormann, C. F.; Fründ, J.; Blüthgen, N.; Gruber, B. *Open Ecol. J.* **2009,** *2,* 7–24. doi:10.2174/1874213000902010007
35. Dormann, C. F.; Gruber, B.; Fründ, J. *R News* **2008,** *8,* 8–11. doi:10.1007/s00357-009-9027-y
36. Greenacre, M. Compositional Data and Correspondence Analysis. *Compositional Data Analysis;* John Wiley & Sons: Hoboken, NJ, U.S.A., 2011; pp 103–113.
37. Greenacre, M.; Lewi, P. *J. Classif.* **2009,** *26,* 29–54. doi:10.1007/s00357-009-9027-y
38. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques,* 3rd ed.; Morgan Kaufmann: Waltham, MA, United States, 2011.
39. Dunham, M. H. *Data Mining: Introductory and Advanced Topics;* Prentice Hall: Upper Saddle River, NJ, U.S.A., 2002.
40. Cox, T. F.; Cox, M. A. A. *Multidimensional Scaling;* CRC Press: Boca Raton, FL, U.S.A., 2010.
41. Székely, G. J.; Rizzo, M. L.; Bakirov, N. K. *Ann. Stat.* **2007,** *35,* 2769–2794. doi:10.1214/009053607000000505
42. Miller, R. J.; Lenihan, H. S.; Muller, E. B.; Tseng, N.; Hanna, S. K.; Keller, A. A. *Environ. Sci. Technol.* **2010,** *44,* 7329–7334. doi:10.1021/es100247x
43. Golub, G. H.; Van Loan, C. F. *Matrix Computations,* 4th ed.; John Hopkins University Press: Baltimore, MD, U.S.A., 2012; Vol. 3.
44. Goslee, S. C.; Urban, D. L. *J. Stat. Software* **2007,** *22,* 1–19.
45. Venables, W. N.; Ripley, B. D. *Modern Applied Statistics with S,* 4th ed.; Statistics and Computing; Springer: Berlin, Germany, 2002. doi:10.1007/978-0-387-21706-2
46. Wasserman, L. *All of Statistics: A Concise Course in Statistical Inference;* Springer: Berlin, Germany, 2004.
47. Sun, L.; Ji, S.; Ye, J. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011,** *33,* 194–200. doi:10.1109/TPAMI.2010.160
48. Peter, J.; Young, W.; Haukka, K. E. *New Phytol.* **1996,** *133,* 87–94. doi:10.1111/j.1469-8137.1996.tb04344.x
49. Wilson, M. K.; Lowe, W. H.; Nislow, K. H. *J. For.* **2014,** *112,* 337–345.

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Analyzing collaboration networks and developmental patterns of nano-enabled drug delivery (NEDD) for brain cancer

Ying Huang[1,2,3], Jing Ma[1,2], Alan L. Porter[*3,4], Seokbeom Kwon[3] and Donghua Zhu[1,2]

Full Research Paper

Open Access

Address:
[1]School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China, [2]Lab of Knowledge Management and Data Analysis (KMDA), Beijing Institute of Technology, Beijing 100081, China, [3]School of Public Policy, Georgia Institute of Technology, Atlanta, GA 30332, USA and [4]Search Technology, Inc., Atlanta, GA 30092, USA

Email:
Alan L. Porter[*] - alan.porter@isye.gatech.edu

* Corresponding author

Keywords:
bibliometrics; brain cancer; collaboration network; nano-enabled drug delivery (NEDD); nanoinformatics

Beilstein J. Nanotechnol. **2015,** *6,* 1666–1676.
doi:10.3762/bjnano.6.169

Received: 01 April 2015
Accepted: 16 July 2015
Published: 31 July 2015

This article is part of the Thematic Series "Nanoinformatics for environmental health and biomedicine".

Guest Editor: R. Liu

Abstract

The rapid development of new and emerging science & technologies (NESTs) brings unprecedented challenges, but also opportunities. In this paper, we use bibliometric and social network analyses, at country, institution, and individual levels, to explore the patterns of scientific networking for a key nano area – nano-enabled drug delivery (NEDD). NEDD has successfully been used clinically to modulate drug release and to target particular diseased tissues. The data for this research come from a global compilation of research publication information on NEDD directed at brain cancer. We derive a family of indicators that address multiple facets of research collaboration and knowledge transfer patterns. Results show that: (1) international cooperation is increasing, but networking characteristics change over time; (2) highly productive institutions also lead in influence, as measured by citation to their work, with American institutes leading; (3) research collaboration is dominated by local relationships, with interesting information available from authorship patterns that go well beyond journal impact factors. Results offer useful technical intelligence to help researchers identify potential collaborators and to help inform R&D management and science & innovation policy for such nanotechnologies.

Introduction

Drug delivery research has grown rapidly over the past two decades and has enabled drug development by designing suitable delivery systems that improve efficacy, lower dosing frequency, and encourage patient convenience and compliance [1]. Within the last ten years, nano-enabled drug delivery (NEDD) has drawn the attention of research and industry areas,

as a key nanotechnology. Nanoparticulate drug-delivery vehicles have been developed using various nanomaterials and components (mainly polymers). Such systems have the ability to encapsulate and carry the payload (therapeutics) and penetrate through biological membranes to deliver that payload to specific target disease sites [2-4]. The outstanding advantage of NEDD is that the applicable nanoparticles can keep the pharmaceutical well protected from degradation and prolong the exposure of the pharmaceutical through controlled release. Thus, NEDD provides a novel approach to medical therapy, including treatment of chronic diseases and genetic disorders [5]. At the present, various kinds of nanoparticles have been developed as drug carriers, such as liposomes, micelles, polymeric conjugates and so on [6-8]. Among these, the brain tumor-targeting drug delivery systems, which increase drug accumulation in the tumor region and reduce toxicity in the normal brain and peripheral tissue, are a promising new approach [9].

Collaboration fosters interactions between different actors within and across fields, which reflects sharing of knowledge and other resources [10]. Linkages generated among actors accelerate communication and information exchange for mutual benefit [11]. In these circumstances, research collaboration facilitates keeping up with advances in methods and findings in related fields. It is vital in interdisciplinary arenas and nano-bioinformatics can bolster intelligence concerning advances and potential collaborators. “R&D landscaping” to understand collaboration and developmental patterns can offer global-level insights [12]. This paper aims to support policy-makers or managers who are making strategic technical decisions regarding NEDD for brain cancer gain useful intelligence on technical and international capabilities. The research employs bibliometric, text analytic, and social network analysis methods to explore the collaboration patterns at the country, institution, and author levels to understand better the international development of NEDD for brain cancer.

This paper highlights three points:

1. The international collaboration index (ICI), calculated using a paper collaboration ratio (PCR) and an international collaboration range (ICR), is applied to measure networking for the top 10 countries at the following stages: 1990–1999, 2000–2009 and 2010–2014;
2. An organization diversity index (ODI) and a country diversity index (CDI) are used to judge the collaboration diversity of leading institutions;
3. The matrix of co-authorship performance, which crosses two dimensions – a paper impact index (PII) and an author contribution index (ACI) – locates the contribution of outstanding domain authors.

Together, these provide a new perspective on scientific collaboration and academic evaluation.

The paper is organized as follows: The first section provides general background on NEDD for brain cancer research. In the second part, search strategy and data are introduced. We focus on the scientific activity and collaboration network at the country, institution, and individual levels in the third section. In the conclusion, we make a brief summary of the research findings and identify promising opportunities for further research.

Search strategy and data

To develop the search strategy of NEDD for brain cancer, we first characterized and classified the essential components, building on a previously developed framework [13].

With the help of knowledgeable colleagues and previous NEDD-related work [14,15], we devised a Boolean, term-based search algorithm for NEDD directed at brain cancer, informed by various reviews and "foresight" pieces. This led us to the following categorization with which to frame our current search, as per Table 1.

We thus obtained 1859 records (language is English and document type is Article), from 1990 to 2014, from the Science Citation Index Expanded (SCI-Expanded) and the Social Sciences Citation Index (SSCI) of the Web of Science (WoS).

Nanomedicine research is a multidisciplinary activity, so exploring the disciplinary distribution is illuminating. Figure 1 offers a science overlay map [16] of NEDD for brain cancer, based on the Web of Science categories of the journals in which the 1859 papers appeared. The map shows that biomedicine and materials science are the most active disciplines. Cognitive science, chemistry and clinical medicine are other prominent disciplines.

Results and Discussion

International collaboration analysis

International scientific collaboration has been a driving force for promoting scientific and technological advancement. In this paper we examine the countries of the authors’ affiliations. Figure 2 shows the number of publications by country, based on the location of all author affiliations (not just first authors), from 1990 to 2014.

Among the publication trends, the USA and China stand out. The USA has led over the past 20 years (Japan had a small advantage in 1998), yet China has dramatically caught up over the last 5 years. According to this trend, China will boast the

Table 1: Search Strategy of NEDD for Brain Cancer in Web of Science.

Set	Category	Records	Search Terms
# 1	T (Target)	63,707	TS = (((brain or "central nervous system" or CNS) near/1 (cancer* or anticancer* or tumor* or tumour* or oncology or neoplasm* or carcinoma*)) or glioma* or glioblastoma*)
# 2	N (nanoparticles and materials)	1,135,180	TS = (nano* or micelle* or liposome* or dendrimer* or metal complex* or hydrogel* or “quantum dots*” or chitosan* or alginate*)
# 3	M (Medicine)	128,626	TS = (temozolomide or procarbazine or carmustine or BCNU or lomustine or CCNU or vincristine or everolimus or irinotecan or cisplatin or carboplatin or methotrexate or etoposide or bleomycin or vinblastine or actinomycin or dactinomycin or cyclophosphamide or ifosfamide)
# 4	P (Pharmaceutical)	40,937	TS = (siRNA or "short interfering RNA" or "small interfering RNA")
# 5	D (delivery systems)	4,936,370	TS = (deliver* or vehicle* or carrier* or vector* or treat* or therap* or "control* releas*" or "sustain* releas*" or transduct* or transfect* or transport* or translocat*)
# 6	Final	1859	#1 AND #2 AND (#3 OR #4 OR #5)

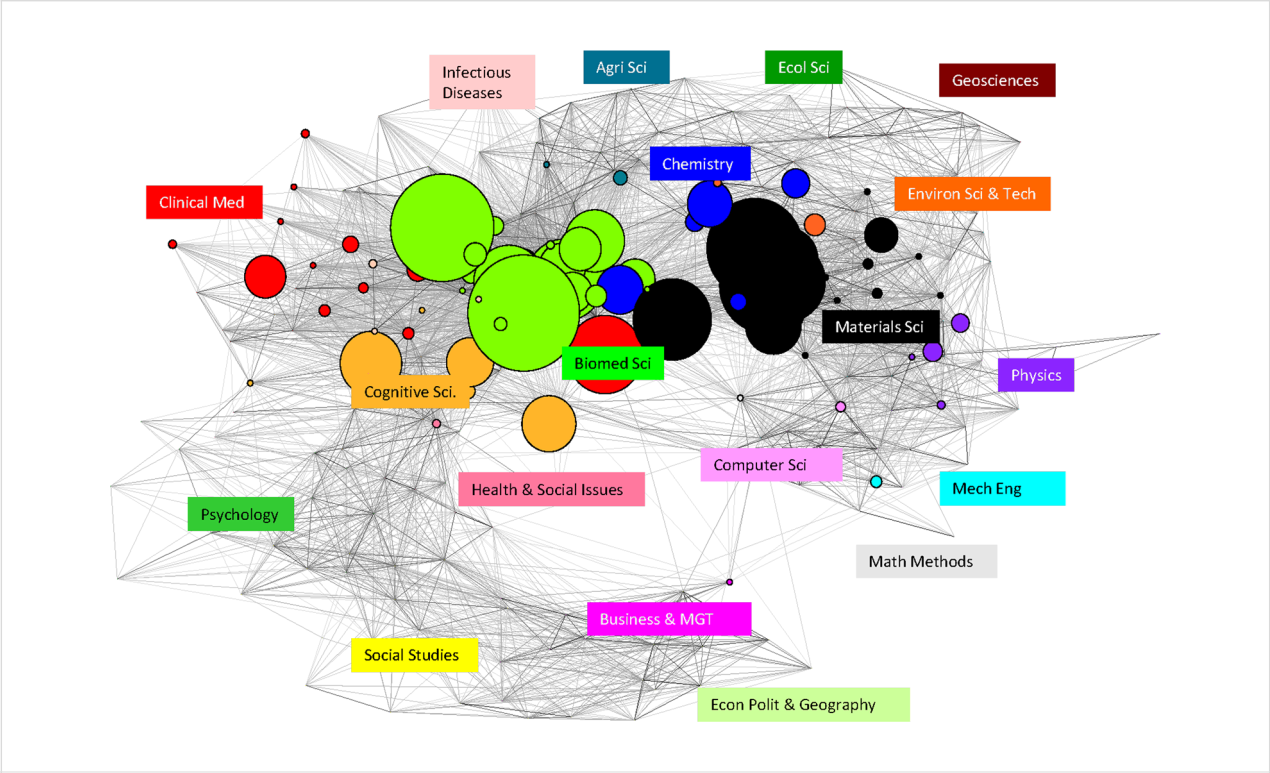


Figure 1: NEDD for brain cancer research across the disciplines.

largest proportion of literature in the near future, and the USA and China will remain the key players in the field of NEDD for brain cancer.

To better understand the various development patterns of the top 10 countries, we introduce centrality analysis models that help answer the question, "What characterizes an important vertex?” [17]. These models are degree centrality (DC), closeness centrality (CC), and betweenness centrality (BC).

For DC, which is defined as the number of links incident upon a node, the USA maintains the highest value, meaning that US researchers have more linkages with researchers in other countries. Germany also has wide academic collaboration networks, especially since 2000.

Based on CC, which is a measure of the total distance to sequentially spread information to all other nodes [18], the USA is located in the core position, making it more likely to collabo-

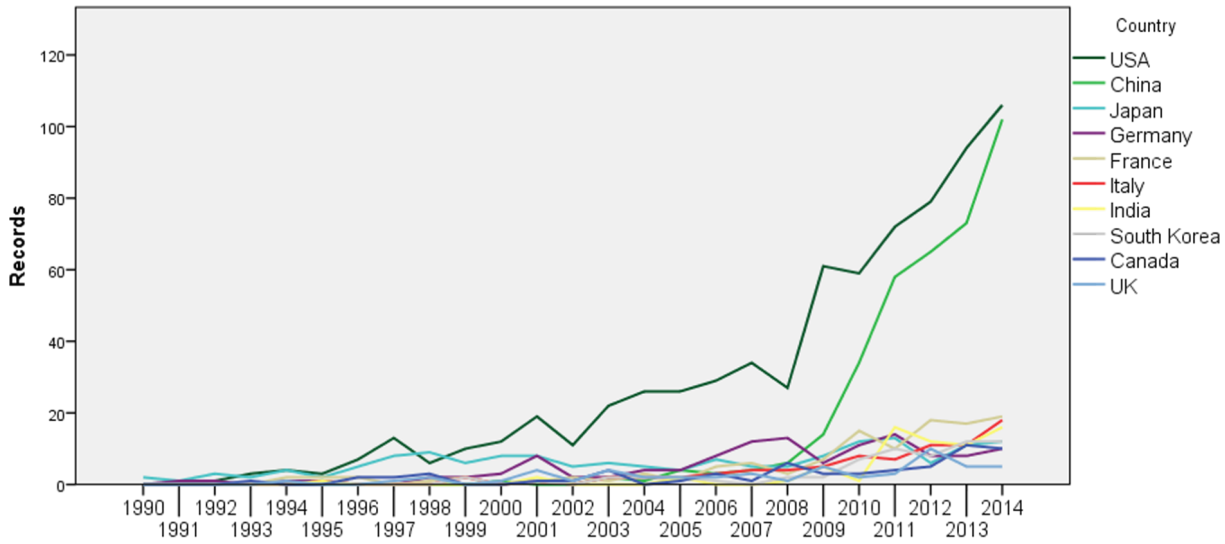


Figure 2: Publication trend of top 10 countries in 1990–2014.

rate with other countries. All other countries share a similar distance among other nodes, from 2000 to 2014.

From the BC perspective, the USA and Germany perform well, acting as a bridge along the shortest path between two other countries. The most striking finding is that, although China is a leader in publication, it plays a quite limited role in connecting other countries (shown as Table 2).

Additionally, the international collaboration index (ICI), calculated by paper collaboration ratio (PCR) and international collaboration range (ICR), is applied to measure the top academic internationalization degree for the top 10 countries within 1990–1999, 2000–2009 and 2010–2014 respectively, as shown in Figure 3.

(1) Paper collaboration ratio (PCR) is defined as how much a country's multinational papers accounted for the country or region's total number of papers. This is derived from "the share of international publications" [19].

$$PCR_N = \frac{\sum_{k \neq N} P_{N,k}}{P_N} \quad (1)$$

In Equation 1, N indicates the country want to calculate, $P_{N,k}$ is the number of papers produced from the cooperation between country 'N' and country 'k'. Thus,

$$\sum_{k \neq N} P_{N,k}$$

Table 2: Centrality analysis for top 10 countries in different stages.

	1990–1999			2000–2009			2010–2014		
	DC	CC	BC	DC	CC	BC	DC	CC	BC
USA	8	0.833	0.208	23	0.706	0.340	35	0.709	0.392
China	0	0.000	0.000	6	0.486	0.011	11	0.519	0.030
Japan	1	0.476	0.000	7	0.522	0.044	11	0.514	0.057
Germany	3	0.588	0.084	16	0.643	0.225	21	0.596	0.099
France	2	0.500	0.000	7	0.522	0.054	19	0.554	0.115
Italy	0	0.000	0.000	5	0.468	0.006	18	0.583	0.091
India	0	0.000	0.000	3	0.439	0.001	11	0.500	0.017
South Korea	0	0.000	0.000	1	0.419	0.000	10	0.519	0.015
Canada	3	0.526	0.003	5	0.500	0.003	12	0.533	0.058
UK	1	0.476	0.000	13	0.581	0.139	11	0.519	0.021

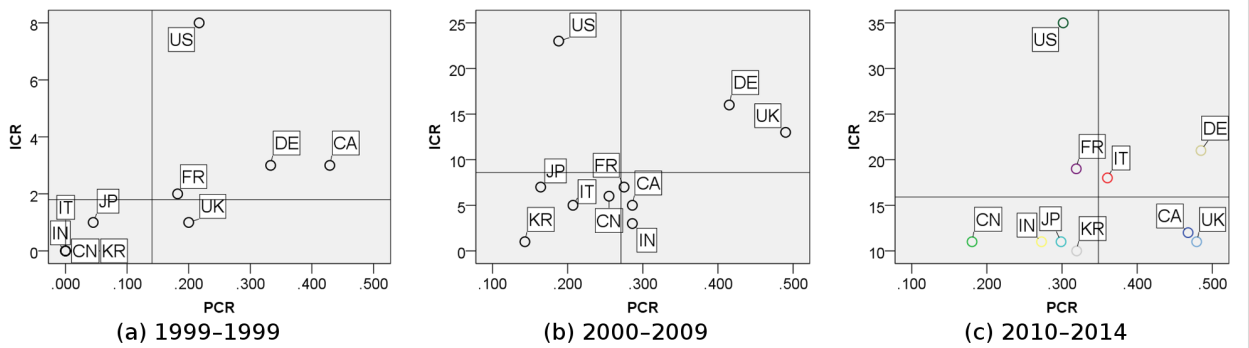


Figure 3: Scatter plot of international collaboration index (ICI) for top 10 countries in (a) 1990–1999; (b) 2000–2009; (c) 2010–2014. The guides indicate the average value of these top 10 countries.

represents the total amount of multinational papers produced from a certain country or region that has taken part in related research by collaboration with the country 'N,' and P_N represents the total amount of papers produced from the country 'N'.

(2) International collaboration range (ICR) is defined as how many partner countries have been involved in collaborations and reflects the breadth of one country or region's international collaboration from a macro view [20].

$$\mathrm{ICR} = TN \tag{2}$$

In Equation 2, TN is the total number of countries or regions with which a country or region has cooperated.

From the scatter plots (Figure 3), we identify some interesting findings:

1. All the top 10 countries show an improvement, both in PCR and ICR, which indicates that international cooperation is becoming more and more frequent in the field of NEDD for brain cancer;
2. The USA always leads the global research, and it has the widest academic collaboration networks and relatively fruitful cooperation outcomes;
3. Compared to other regions, Asian countries, including China, Japan, India and South Korea, are located at the low-ICR and low-PCR area, which means they have relatively less connection with researchers of other nationalities, despite their recent growth in articles published.

Institutional co-authorship analysis

In general, the research levels of a certain country depend on its leading institutions. Figure 4 shows the 12 leading institutions

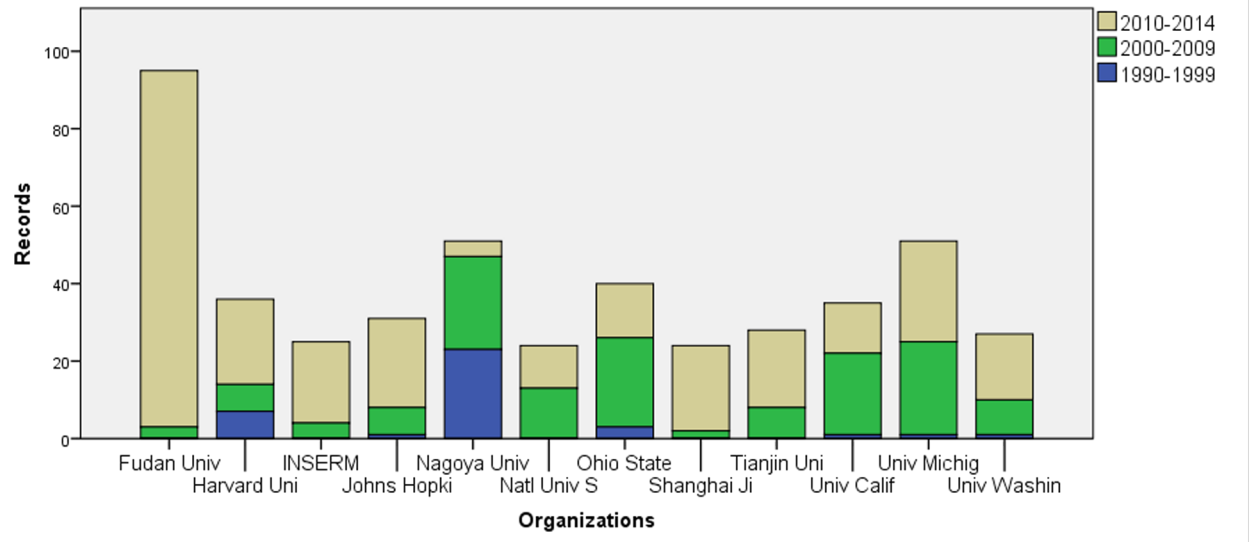


Figure 4: Publications of the top 12 institutions in 1990–2014.

in NEDD for brain cancer research. Most institutions show good performance for the last 5 years, and Fudan University achieves an amazing number of research results, showing that their number of publications between 2010 and 2014 is far greater than any other institution in the same time period. Nagoya University led the domain development previously, but it encountered a serious decline recently and is losing ground.

Among these top 12 institutions, half come from the USA, three are from China, and the remaining three organizations are in Japan, France, and Singapore. Citations that establish links to other works or other researchers are treated as an indicator of impact [21]. From Table 3, we can see that papers published by the National University of Singapore are the most cited by other researchers (63 per paper), and they also reference more previous work (57 records) per publication. Additionally, some other institutions from the USA perform outstandingly in citations as well, including the University of Michigan, Harvard University, and the University of Washington. However, citation is usually skewed, so we introduce the median times cited that is the median value of all times cited to further evaluate the citation behavior. University of California, San Francisco (UCSF) shows most expressive performance in median times cited, and followed by University of Michigan, Ohio State University and National University of Singapore. Even through Harvard University stands out in average times cited, most of the citations are contributed by the few highly cited papers.

In the area of collaboration activity, we introduce the organization diversity index (ODI) and the country diversity index (CDI) to locate the top 12 institutions.

(1) ODI is defined as the index of the collaboration distributions and collaboration times of certain organizations with other organizations through multi-institutional publications. It can be expressed as follows:

$$\mathrm{ODI} = 1 - \frac{\sum_{i \in C} \left(Q^i\right)^2}{TQ^2} \tag{3}$$

In Equation 3, Q^i represents the number of multi-institutional publications involving collaborators from certain institutions 'i'. C represents the set of historical collaborators of the targeted organization, TQ represents the total multi-institutional publications of the organization.

(2) CDI has a definition similar to ODI, but it is set to explore the country level, rather than the institutional level.

In Figure 5 we see that Harvard University, INSERM, Tianjin University, and Ohio State University have wide international academic collaboration and influential research results. In comparison with some other institutions, such as University of Michigan and National University of Singapore, they tend to have more connections with international institutions than domestic organizations. The University of California, San Francisco (UCSF), has a strong partnership with other institutions in its country. Other leading institutions – including Fudan University, Nagoya University, Johns Hopkins University, and the University of Washington – have strong cooperative relations inside their organizations.

Table 3: Publications and citation information for the top 12 institutions.[a]

Organization	Records	Average times cited	Median times cited	Average times citing	Country
Fudan Univ	95	19.9	12.0	13.0	China
Nagoya Univ	53	25.5	14.0	17.0	Japan
Univ Michigan	51	52.9	27.0	38.3	USA
Ohio State Univ	40	35.5	27.0	24.0	USA
Harvard Univ	37	52.0	19.5	46.4	USA
Univ Calif San Francisco	35	38.7	32.0	23.9	USA
Johns Hopkins Univ	31	24.6	10.0	23.1	USA
Tianjin Univ	28	32.0	14.0	26.6	China
Univ Washington	27	47.3	21.0	35.5	USA
INSERM	24	13.8	7.5	11.8	France
Natl Univ Singapore	24	63.2	26.5	57.4	Singapore
Shanghai Jiao Tong Univ	24	13.2	3.0	10.7	China

[a]Citation counts include self-citation.

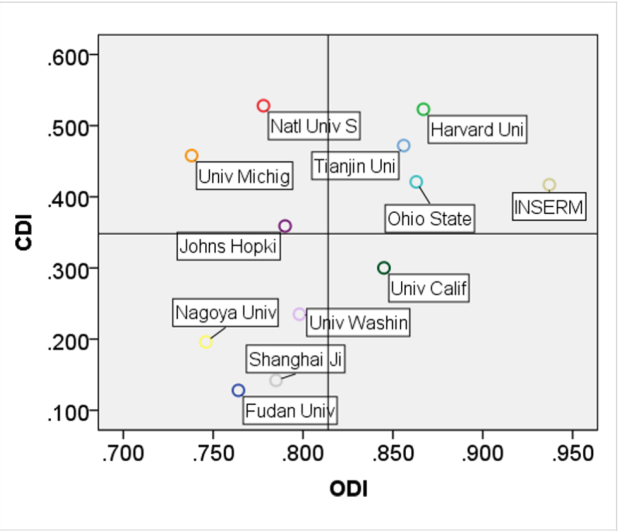


Figure 5: Collaboration activity of the top 12 institutions.

Author activity analysis

No matter the advantages of a country or the influence of an institution, it is the researchers that make them truly great. Exploring the core authors in the NEDD for brain cancer field can help researchers take advantage of leading potential cooperative partners. Figure 6 shows the co-author network of the top 20 authors, in terms of numbers of research papers. From Figure 6, we see that the majority of the authors in the NEDD for brain cancer field have strong connections in the micro-community. In other words, they often come from the same institution (see Table 4). There are five main partnering relationships: Nagoya University–University of California, San Francisco (UCSF) group, Fudan University group, University of Angers group, Ohio State University group, and a University of Michigan group.

Even though the USA ranks first in this new field, none of its authors rank in the top three of the author list, and only two rank in the top 10. Yoshida and Mizuno, both of whom come from the Department of Neurosurgery, Nagoya University School of Medicine in Japan, rank first and second in the list, respectively. Among the top 20 authors, however, US scholars represent 45% of the total, showing that the USA does hold a strong position in NEDD brain cancer research. Besides the representation from the USA, four authors come from China, three from France, three from Japan and one from Germany.

It should be noted that authorship analysis focuses on the productivity of authors and their contributions in their respective fields. In multi-authored papers, the first author position is

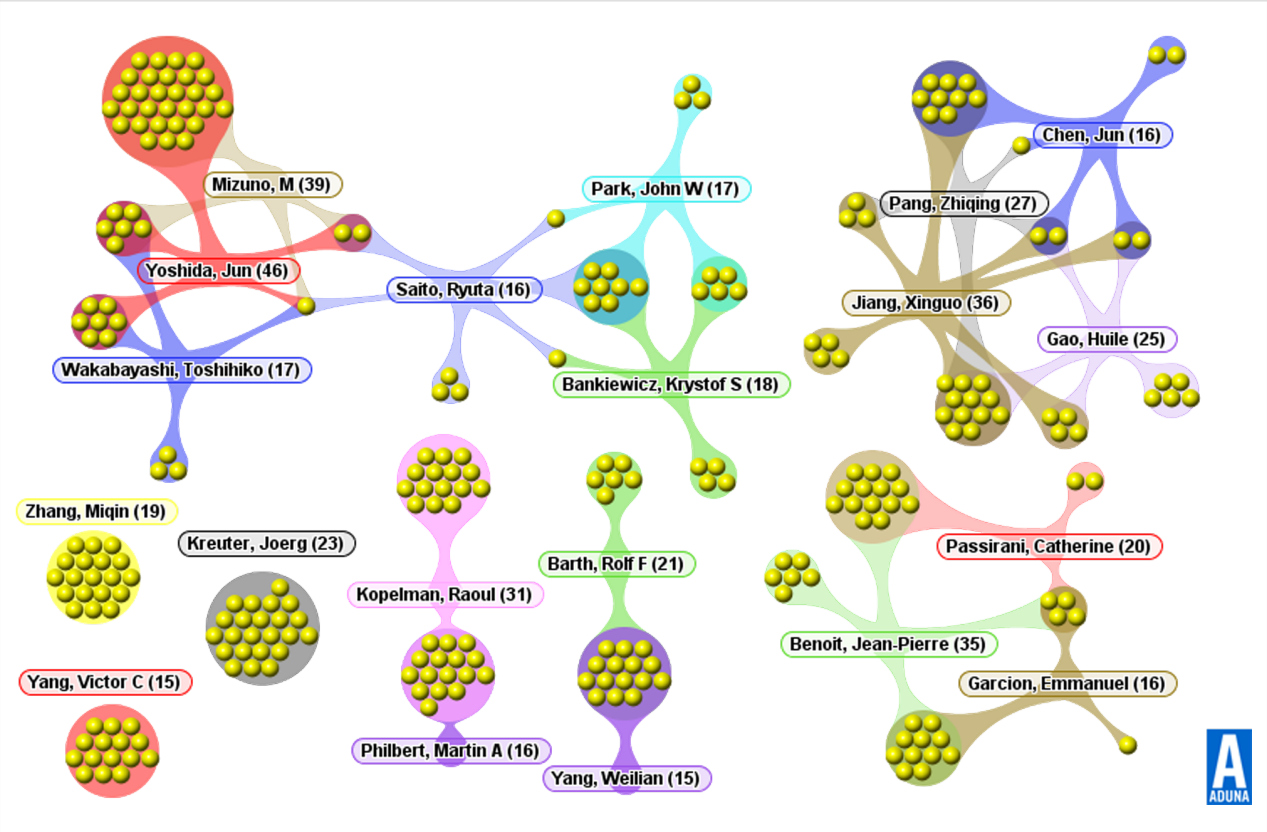


Figure 6: Co-author network of the top 20 authors.

occupied by the individual nominally making the greatest contribution [22]. Authors in the top 20 list, while productive in domain publications, are not often listed as the first author. Only 4 researchers occupy the first position in more than 20% of their respective papers. In addition, publication amounts do not always match the results of the citation evaluation, which can be observed as the average times cited and the h-index (shown as Table 4). What's more, we can figure out that the research areas of these leading researchers tend to emphasize oncology, materials science, pharmacology & pharmacy, chemistry, neurosciences & neurology, research & experimental medicine and engineering, which indicates that nanomedicine research is a multidisciplinary activity. At the same time, researchers coming from the same institution tend to focus on similar research areas and collaborate on within-domain research.

In order to better evaluate author activity, we introduce the paper impact index (PII) and author contribution index (ACI).

(1) Paper impact index (PII) is defined as the sum of the impact factors of all published papers. It can be expressed as follows:

$$\mathrm{PII} = \sum_{i \in A} \mathrm{IF}i \tag{4}$$

Table 4: Top 20 authors in NEDD for brain cancer.

Authors	Records	1st-Author records	Average citations	h-index	Organization	Country	Research area
Yoshida, Jun	46	7	24.59	18	Nagoya Univ	Japan	Oncology; Neurosciences & Neurology
Mizuno, M	39	8	18.74	15	Nagoya Univ	Japan	Oncology; Neurosciences & Neurology
Jiang, Xinguo	36	0	21.36	14	Fudan Univ	China	Materials Science; Pharmacology & Pharmacy
Benoit, Jean-Pierre	35	0	16.4	13	Univ Angers	France	Pharmacology & Pharmacy; Chemistry
Kopelman, Raoul	31	1	61.97	19	Univ Michigan	USA	Chemistry; Materials Science
Pang, Zhiqing	27	2	22.71	14	Fudan Univ	China	Materials Science; Pharmacology & Pharmacy
Gao, Huile	25	14	12.81	9	Fudan Univ	China	Pharmacology & Pharmacy; Materials Science
Kreuter, Joerg	23	3	68.26	18	Univ Frankfurt	Germany	Pharmacology & Pharmacy; Chemistry
Barth, Rolf F	21	3	51.9	17	Ohio State Univ	USA	Oncology; Chemistry
Passirani, Catherine	20	0	15.5	10	Univ Angers	France	Pharmacology & Pharmacy; Chemistry
Zhang, Miqin	19	0	53.68	13	Univ Washington	USA	Materials Science; Chemistry
Bankiewicz, Krystof S	18	0	38.35	13	Univ Calif San Francisco	USA	Neurosciences & Neurology; Oncology
Park, John W	17	1	53.83	14	Univ Calif San Francisco	USA	Neurosciences & Neurology; Oncology
Wakabayashi, Toshihiko	17	2	32.24	11	Nagoya Univ	Japan	Oncology; Research & Experimental Medicine
Chen, Jun	16	0	18.48	11	Fudan Univ	China	Materials Science; Enginerring
Garcion, Emmanuel	16	1	19.56	9	Univ Angers	France	Pharmacology & Pharmacy; Chemistry
Philbert, Martin A	16	0	86.25	15	Univ Michigan	USA	Chemistry; Pharmacology & Pharmacy
Saito, Ryuta	16	7	47.81	11	Univ Calif San Francisco	USA	Neurosciences & Neurology; Oncology
Yang, Victor C	15	0	48.73	12	Univ Michigan	USA	Materials Science; Pharmacology & Pharmacy
Yang, Weilian	15	6	13.6	8	Ohio State Univ	USA	Oncology; Chemistry

In Equation 4, IF_i represents the impact factor (IF) of the journal that published the article 'i' of certain author, as indicated by the journal citation reports (JCR), provided by Thomson Reuters; 'A' represents the set of articles that the author published.

(2) Author contribution index (ACI) is defined as the total contribution of the author in all authored papers. Authorship order only reflects relative contribution (with considerable variability in norms), whereas evaluation committees often prefer other quantitative measures. A reasonable method for quantifying contributions is to give the first author credit for the whole contribution, the second author half, the third a third, and so forth [23]. In this paper, we take the value as follows:

$$\mathrm{ACI} = H1 + 0.5H2 + 0.25H3 + 0.125Hn \qquad (5)$$

In Equation 5, H_1, H_2, H_3 represents the number of a certain author's first-, second- and third-author papers within a period, and H_n represents the number of papers in which his or her name appears after the first three in the authorship order.

The author activities of top 20 authors in NEDD for brain cancer are shown as Figure 7.

From the standpoint of research performance, many authors publish papers in high IF journals, which allows their work to be more widely accessible and more influential on other researchers. Jiang (Fudan University), Pang (Fudan University), Kopelman (University of Michigan), Benoit (University of Angers), and Zhang (University of Washington) have similar activity patterns of marked research influence. From the standpoint of contributions in multi-authored publications, Mizuno (Nagoya University), Yoshida (Nagoya University) and Gao (Fudan University) all published more papers as the first author during our survey period.

Typically, advanced scholars will publish their research results in high IF journals, while promising scholars publish more papers as the first author. According to this logic, the author activity pattern can be divided into four types, based on academic influence and research contribution:

1. High-PII and High-ACI: Prestigious and active researchers
2. High-PII and Low-ACI: Experienced and senior researchers
3. Low-PII and High ACI: Growing and promising researchers
4. Low-PII and Low-ACI: New and emerging researchers.

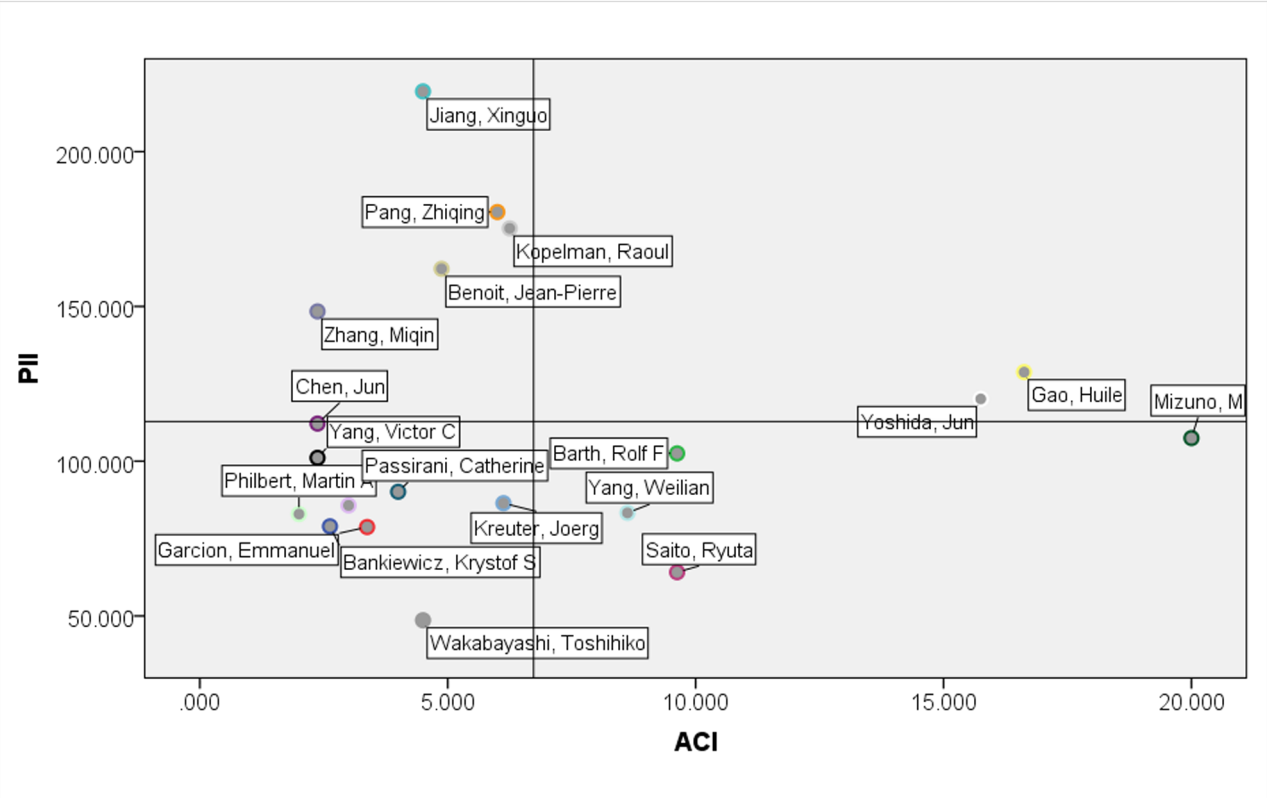


Figure 7: The author activities of top 20 authors in NEDD for brain cancer.

Thus, we see that, in the NEDD for brain cancer field, there is a leading minority of key authors while most of the other authors are still in the stage of exploring this NESTs.

Conclusion

The above analyses reveal some interesting and meaningful findings for the NEDD for brain cancer field:

1. International cooperation is becoming more and more frequent overall, but most countries have different cooperation characteristics, and their academic status varies in different periods.
2. Leading institutes with higher publication numbers perform strongly in terms of citations. American institutes are especially prominent, both in citation behavior and in the collaboration index, as measured by country diversity and organization diversity.
3. Academic researchers tend to seek internal partnerships. Their contributions in published literature should be further evaluated with respect to authorship patterns, even though these publications are accepted by high-impact journals.

NEDD systems are rapidly growing as a key area for nanotechnology application and emerging on a variety of R&D fronts to address a large range of challenges, and curing brain cancer is a high potential application of NEDD that is worth of more exploration. Exploring nano biomedicine research from the respective of social science causes us great interest. Literature informatics, such as our multi-tier R&D landscaping, can help inform science policy makers about collaboration patterns and help technology managers prioritize developmental prospects. Analyzing large compilations of research publication (and/or patent) records can help track developmental trajectories and forecast innovation pathways. Topical analyses within field, not emphasized here, can further aid researchers in identifying potentially useful techniques and research findings in adjacent fields, as well as spotting potential collaborators. The method proposed in the paper can be applied to other research fields to support policy-makers or managers who are making strategic technical decisions with the goal to enhance their technological innovation capabilities and international competitiveness.

Acknowledgements

This research is undertaken at Georgia Institute of Technology, drawing on support from the United States National Science Foundation (NSF) (Award No.1064146), MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Award No.13YJC630042) and the China Scholarship Council (CSC Student ID 201406030005). The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the supporters. We thank Yi Zhang and other colleagues in the "Innovation Co-lab" of Georgia Institute of Technology, Beijing Institute of Technology, and Manchester University, for their advice and feedback.

References

1. Orive, G.; Hernández, R. M.; Gascón, A. R.; Domínguez-Gil, A.; Pedraz, J. L. *Curr. Opin. Biotechnol.* **2003,** *14,* 659–664. doi:10.1016/j.copbio.2003.10.007
2. Ganguly, K.; Chaturvedi, K.; More, U. A.; Nadagouda, M. N.; Aminabhavi, T. M. *J. Controlled Release* **2014,** *193,* 162–173. doi:10.1016/j.jconrel.2014.05.014
3. Wang, A. Z.; Langer, R.; Farokhzad, O. C. *Annu. Rev. Med.* **2012,** *63,* 185–198. doi:10.1146/annurev-med-040210-162544
4. Chaturvedi, K.; Ganguly, K.; Kulkarni, A. R.; Kulkarni, V. H.; Nadagouda, M. N.; Rudzinski, W. E.; Aminabhavi, T. M. *Expert Opin. Drug Delivery* **2011,** *8,* 1455–1468. doi:10.1517/17425247.2011.610790
5. Marcato, P. D.; Duran, N. *J. Nanosci. Nanotechnol.* **2008,** *8,* 2216–2229. doi:10.1166/jnn.2008.274
6. Felice, B.; Prabhakaran, M. P.; Rodriguez, A. P.; Ramakrishna, S. *Mater. Sci. Eng., C* **2014,** *41,* 178–195. doi:10.1016/j.msec.2014.04.049
7. Egusquiaguirre, S. P.; Igartua, M.; Hernández, R. M.; Pedraz, J. L. *Clin. Transl. Oncol.* **2012,** *14,* 83–93. doi:10.1007/s12094-012-0766-6
8. Mundargi, R. C.; Babu, V. R.; Rangaswamy, V.; Patel, P.; Aminabhavi, T. M. *J. Controlled Release* **2008,** *125,* 193–209. doi:10.1016/j.jconrel.2007.09.013
9. Wei, X.; Chen, X.; Ying, M.; Lu, W. *Acta Pharm. Sin. B* **2014,** *4,* 193–201. doi:10.1016/j.apsb.2014.03.001
10. Ma, J.; Wang, X.; Zhu, D.; Zhou, X. *Int. J. Technol. Manag.* **2015,** in press.
11. Ozcan, S.; Islam, N. *Technol. Forecast. Soc. Change* **2014,** *82,* 115–131. doi:10.1016/j.techfore.2013.08.008
12. Wang, X.; Li, R.; Ren, S.; Zhu, D.; Huang, M.; Qiu, P. *Scientometrics* **2014,** *98,* 1745–1762. doi:10.1007/s11192-013-1180-8
13. Robinson, D. K. R.; Huang, L.; Guo, Y.; Porter, A. L. *Technol. Forecast. Soc. Change* **2013,** *80,* 267–285. doi:10.1016/j.techfore.2011.06.004
14. Zhou, X.; Porter, A. L.; Robinson, D. K. R.; Shim, M. S.; Guo, Y. *Nanomedicine* **2014,** *10,* 889–896. doi:10.1016/j.nano.2014.03.001
15. Ma, J.; Porter, A. L. *Scientometrics* **2015,** *102,* 811–827. doi:10.1007/s11192-014-1392-6
16. Rafols, I.; Porter, A. L.; Leydesdorff, L. *J. Am. Soc. Inf. Sci. Technol.* **2010,** *61,* 1871–1887. doi:10.1002/asi.21368
17. Borgatti, S. P. *Soc. Networks* **2005,** *27,* 55–71. doi:10.1016/j.socnet.2004.11.008
18. Newman, M. E. J. *Soc. Networks* **2005,** *27,* 39–54. doi:10.1016/j.socnet.2004.11.009
19. Glänzel, W.; De Lange, C. *Scientometrics* **1997,** *40,* 605–626. doi:10.1007/BF02459304
20. Wang, X.; Huang, M.; Wang, H.; Lei, M.; Zhu, D.; Ren, J.; Jabeen, M. *J. Informetr.* **2014,** *8,* 854–862. doi:10.1016/j.joi.2014.08.004
21. Leydesdorff, L.; Amsterdamska, O. *Sci. Technol. Hum. Val.* **1990,** *15,* 305–335. doi:10.1177/016224399001500303
22. Verhagen, J. V.; Wallace, K. J.; Collins, S. C.; Scott, T. R. *Nature* **2003,** *426,* 602. doi:10.1038/426602a

23. Tscharntke, T.; Hochberg, M. E.; Rand, T. A.; Resh, V. H.; Krauss, J. *PLoS Biol.* **2007,** *5,* e18. doi:10.1371/journal.pbio.0050018

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

The Nanomaterial Data Curation Initiative: A collaborative approach to assessing, evaluating, and advancing the state of the field

Christine Ogilvie Hendren[*1], Christina M. Powers[2,3], Mark D. Hoover[4] and Stacey L. Harper[5,6]

Full Research Paper

Open Access

Address:
[1]Center for the Environmental Implications of NanoTechnology, Duke University, Durham, NC, USA, [2]National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, RTP, NC, USA, [3]current affiliation: Office of Transportation and Air Quality, Office of Air and Radiation, U.S. EPA, Ann Arbor, MI, USA, [4]National Institute for Occupational Safety and Health, Morgantown, WV, USA, [5]Department of Environmental and Molecular Toxicology, Oregon State University, Corvallis, OR, USA and [6]School of Chemical, Biological and Environmental Engineering, Oregon State University, Corvallis, OR, USA

Email:
Christine Ogilvie Hendren[*] - christine.hendren@duke.edu

* Corresponding author

Keywords:
curation; data integration; interoperability; nanoinformatics; nanomaterials

Beilstein J. Nanotechnol. **2015,** *6,* 1752–1762.
doi:10.3762/bjnano.6.179

Received: 26 March 2015
Accepted: 17 July 2015
Published: 18 August 2015

This article is part of the Thematic Series "Nanoinformatics for environmental health and biomedicine".

Guest Editor: R. Liu

Abstract

The Nanomaterial Data Curation Initiative (NDCI), a project of the National Cancer Informatics Program Nanotechnology Working Group (NCIP NanoWG), explores the critical aspect of data curation within the development of informatics approaches to understanding nanomaterial behavior. Data repositories and tools for integrating and interrogating complex nanomaterial datasets are gaining widespread interest, with multiple projects now appearing in the US and the EU. Even in these early stages of development, a single common aspect shared across all nanoinformatics resources is that data must be curated into them. Through exploration of sub-topics related to all activities necessary to enable, execute, and improve the curation process, the NDCI will provide a substantive analysis of nanomaterial data curation itself, as well as a platform for multiple other important discussions to advance the field of nanoinformatics. This article outlines the NDCI project and lays the foundation for a series of papers on nanomaterial data curation. The NDCI purpose is to: 1) present and evaluate the current state of nanomaterial data curation across the field on multiple specific data curation topics, 2) propose ways to leverage and advance progress for both individual efforts and the nanomaterial data community as a whole, and 3) provide opportunities for similar publication series on the details of the interactive needs and workflows of data customers, data creators, and data analysts. Initial responses from stakeholder liaisons throughout the nanoinformatics

community reveal a shared view that it will be critical to focus on integration of datasets with specific orientation toward the purposes for which the individual resources were created, as well as the purpose for integrating multiple resources. Early acknowledgement and undertaking of complex topics such as uncertainty, reproducibility, and interoperability is proposed as an important path to addressing key challenges within the nanomaterial community, such as reducing collateral negative impacts and decreasing the time from development to market for this new class of technologies.

Introduction

The topic of Big Data, and its promise to combine and analyze vast amounts of information to produce new knowledge, has gained widespread interest across many fields and in popular science literature today. The bioinformatics community provides a concrete illustration of the value that mechanisms for synthesizing large and disparate datasets could bring to the broader scientific community. Collaborative approaches to synthesize data add value to the scientific community in terms of a variety of parameters, including: leveraging research investments across multiple initiatives, facilitating trans-disciplinary translation of information, accelerating scientific discovery, and enabling faster risk assessment and commercialization of new technologies. These parameters are especially critical for emerging technologies, such as nanotechnology. The issues addressed in this initiative are certainly not unique to nanomaterials; in fact, they are important to chemistry, materials science and toxicology fields as a whole. However, drawing on existing experience with standards development, data handling and data integration to address viable solutions for complex data integration within the scope of nanomaterial data may serve as a specific case that could ultimately provide insights useful to broader data spheres.

Challenges for the global development of engineered nanomaterials

Researchers and product developers around the globe are currently working toward understanding and controlling the behavior of matter at the nanoscale. Engineered nanomaterials (ENMs), typically classified as materials with at least one dimension between 1 and 100 nanometers that exhibit unique physical, biological, or chemical behavior due to their size, present both the opportunity to harness their novel properties for a wide range of applications, as well as to anticipate and mitigate potential collateral consequences (e.g., accumulation of biopersistent materials in environmental media and latent adverse health effects of a material) [1,2]. Because understanding the behavior of nanomaterials of natural or incidental origin is a critical aspect of investigating the impacts of nanomaterials that are engineered, data are being gathered on all classes of these materials; therefore, throughout the paper we refer to "nanomaterials" to encompass all types (i.e., natural, incidental, engineered), except in cases in which we explicitly state ENM(s). The large variety of potential nanomaterial physicochemical characteristics and applications has led to diverse and rapidly emerging data in terms of materials (both pristine and modified), their interactions in environments (both laboratory-based and natural), and across a broad spectrum of potentially relevant biological interactions. The prospect of integrating nanomaterial datasets is thus difficult in itself. Add to this the fact that protocols for fabricating, measuring and testing nanomaterials are still in the process of being developed. Moreover, nanomaterials are dynamic, often transforming dramatically upon release to the environment, or into the body. Such challenges make the process of integrating diverse nanotechnology-related datasets a seemingly intractable problem. Progress toward defining and achieving a level of "functional interoperability" of datasets, which we define as the level of sameness within a dataset that facilitates sharing and comparison for a given analytical purpose, will require a collaborative effort by the nanomaterial community (i.e., researchers, product developers, funding agencies, regulators). Specifically, community members will need to define the purposes for sharing and to develop and apply complementary approaches to collect, manage and share data in ways that can support those purposes.

Community focus on building effective nanoinformatics approaches

The need for collaborative and dedicated attention to informatics in the nanomaterial community was a focal point of two recent National Research Council (NRC) reports on nanomaterial research progress for environment, health and safety (EHS) [3,4]. A number of efforts to begin enabling interoperability in nanomaterial datasets are already underway that draw on established data management approaches. Examples of specifically funded data repository projects include: the RTI International Nanomaterial Registry (http://www.nanomaterialregistry.org) and the National Cancer Institute (NCI) Nanotechnology Characterization Lab (http://ncl.cancer.gov). The Nanotechnology Knowledge Infrastructure (NKI), one of six signature initiatives of the National Nanotechnology Coordination Office, also provides a resource for federal agencies in the United States to work toward shared data streams (http://www.nano.gov/NSINKI). The Materials Genome Initiative (http://materialsinnovation.tms.org/genome.aspx) is a

broader, but related data management effort to catalogue materials and their key characteristics [5].

Prior to the development of these efforts, the NCI established the National Cancer Informatics Program (NCIP) Nanotechnology Working Group (Nano WG) for nanomaterial researchers with a specific interest in informatics and computational approaches. This working group includes active membership and input from many communities (e.g., nanoEHS, commercial industry, standards community), but began with a particular emphasis on nanomedicine. From this area of emphasis, the NCIP NanoWG is well-positioned to serve as a conduit for sharing experience and best practices of the bioinformatics community with the emerging nanoinformatics community. In doing so, the NCIP NanoWG facilitates the translation of lessons learned in prior efforts to link disparate datasets and probe important community research questions; the group also leads discussions of data issues unique to the uncertainties inherent to nanomaterials and other emerging technologies that have inherent uncertainties. The NCIP NanoWG now encompasses additional stakeholder groups including industry representatives and environmental risk forecasters, all similarly interested in how the novel properties of engineered nanomaterials affect their interactions and behavior.

Since its inception, the NCIP NanoWG has supported the development of the NanoParticle Ontology (NPO) (http://www.nanoontology.org) vocabulary standards, first published in 2011 and periodically updated. In addition, the group recently developed and published data-exchange standards along with tools to enable the use of these standards (ISA-TAB-Nano; ASTM International E2909-13) [6]. To build on these efforts, the NCIP NanoWG is now developing a shared vision for curation of data related to nanoscale materials via the broad, community-inclusive NDCI project presented here.

A vision of nanoinformatics roles and responsibilities

The NCIP NanoWG-lead Nanomaterial Data Curation Initiative (NDCI) explores the critical aspect of data curation within the development of informatics approaches to understanding nanomaterial behavior. The following working definition (expanded from the Nanoinformatics 2020 Roadmap [7]) has been proposed [8]: "*Nanoinformatics is the science and practice of determining which information is relevant to meeting the objectives of the nanoscale science and engineering community, and then developing and implementing effective mechanisms for collecting, validating, storing, sharing, analyzing, modeling, and applying the information, and then confirming that appropriate decisions were made and that desired mission outcomes were achieved,*[…]" with additional steps in the informatics lifecycle including "[…]*conveying experience to the broader community, contributing to generalized knowledge, and updating standards and training*." Successful nanoinformatics endeavors will apply all of the steps in the process.

In the context of the overall working definition of nanoinformatics, the roles and responsibilities of the myriad individuals who are engaged in the development and application of nanotechnology can be viewed as fitting into four categories: data *customers* (who specify the data needs for their intended purposes), data *creators* (who will develop relevant and reliable data to meet the customer needs), data *curators* (who will perform the central roles described in this NDCI work), and data *analysts* (who will develop and apply models for data analysis and interpretation that are consistent with the quality and quantity of the data and that meet customer needs). In some instances, the same individuals may perform all roles, and in the larger global reality the individuals and their roles may extend over significant distances, organizations, and time periods.

The central role of curation

Data curation has been defined as "the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education; curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time" [9]. Data curation has been chosen as the focus of the collaborative initiative because of its central role in facilitating all aspects of the informatics lifecycle. Resources like those noted above that are developing to organize and analyze nanomaterial data represent efforts that can differ widely in terms of data scopes, driving goals, and development phases. Despite these potentially divergent aspects, one commonality shared across all nanoinformatics resources is that data must be *curated* into those resources.

The purpose of this article

This article outlines the NDCI project and lays the foundation for a series of papers on nanomaterial data curation. Ultimately, through this series of papers, the NDCI will: 1) present and evaluate the current state of nanomaterial data curation across the field on multiple specific data curation topics, 2) propose ways to leverage and advance progress for both individual efforts and the nanomaterial data community as a whole, and 3) provide opportunities for similar publication series on the details of the interactive needs and workflows of data customers, data creators, and data analysts.

The specific objectives of the NDCI paper series include:

- to capture a snapshot of current nanomaterial data curation practices and issues,

- to develop recommendations for moving the nanoinformatics community toward increasingly standardized curation practices; and
- to facilitate collaborations between researchers, product developers, and others working with nanomaterials that establish and utilize common datasets for cross-boundary work (e.g., application of data from an academic institution to nanomaterial product development in industry).

In the subsequent sections below, we expand on the rationale and approach for our focus on data curation as an integral piece within the nanomaterial community's efforts to progress towards the functional interoperability of datasets, and we conclude with an invitation for active community collaboration in these efforts.

The NDCI focus on data curation

The motivation

The term nanoinformatics can encompass a vast scope and differ in meaning to different audiences. These scopes and meanings may refer to such diverse data types and uses as: catalogues of self-identified nano-enabled products on the market; efforts to derive nano-specific quantitative structure activity relationships (QSARs); or estimating environmental concentrations based on a mixture of measurements and models. The range of definitions, scopes and purposes of nanomaterial data-driven efforts is broad, but what is shared between these efforts are the needs to leverage limited resources and to understand clearly what the emerging data mean. There are many aspects to consider and optimize in moving toward a true knowledge or data commons as called for in various ways by the NRC, the NNI and the EU Nanosafety Cluster. Multiple focal areas and driving goals must be considered across the data life cycle; multiple roles exist as well, with different orientations toward the data including creators, customers, curators, and analysts. At this nascent stage in the formation of a nanoinformatics community, even in the face of so much disparity, one common aspect shared across all nanoinformatics resources is that in some form, data must be curated into them. Through exploration of sub-topics related to all activities necessary to enable, execute, and improve the curation process, it is our goal that the NDCI will provide a substantive analysis of nanomaterial data curation itself, as well as a platform for multiple other important discussions to advance the field of nanoinformatics.

Scientific data curation, a mature field within library science and a maturing sub-field of most data-driven academic domains, is increasingly a topic of interest within the nanomaterial research and associated nanoinformatics communities [10]. The methods, protocols and parameters guiding data generation within this young area of science are developing in parallel with data characterizing these novel materials, their performance, and their potential impacts. With the innumerable materials, functionalities, and complex application and implication scenarios, testing ENMs on a case-by-case basis is an intractable proposition; leveraging research investments across the community will be critical to enable the type of iterative feedback between disciplines and sectors necessary to meet the important challenges of responsibly commercializing nanotechnologies. By working together from the beginning to tackle difficult data issues including uncertainty, reproducibility, and interoperability of complex datasets, the nanoinformatics community could collaboratively address these challenges. In doing so, the community can help decrease the time from development to market and reduce collateral negative impacts of nano-enabled technologies.

The goals of this initiative are to describe the current baseline of curation practices and to develop recommendations for moving the nanoinformatics community forward. Data curation is a broad term encompassing all aspects involved with assimilating data into centralized repositories or sharable formats. Borrowed from the concept of art curation, the term "curation" is selected to signify that this process entails more than a series of data management tasks, but also includes elements of discernment and judgment inherent to this decision process. The curation practices captured through the NDCI will incorporate aspects of both reasoning and methods for curation steps including: sourcing and parsing of information into datasets; organizing data into cyberinfrastructures; formatting data for current or future interoperability; and identifying implications that commonly adopted data and meta-data formatting conventions may have on defining data quality and therefore impacting future experimental design. A goal of the NCIP NanoWG Nanomaterial Data Curation Initiative is to help establish an understanding of what a wide range of stakeholders in data curation mean when they talk about and undertake this process. In doing so, we can identify synergies and disconnects between different efforts, both of which are necessary to advance toward interoperability of large, disparate nano datasets. There are many ways to orient a discussion on the integration of tools and datasets; nano curation was selected as a focus because the process of understanding how different organizations consume and manage nanotechnology related data will require us to explicitly discuss underlying assumptions and practical approaches to individual efforts. In turn, we can better understand and communicate with the scientific community what would be required to integrate the efforts. Though we will present synthesized recommendations for moving forward, we are also committed to reporting dissenting opinion. Indeed, where disagreement can be identified, we may diagnose the root cause of disconnects between approaches to curation. This in

itself will represent a useful exercise as we map out the landscape of nano data curation and determine what level of interoperability between datasets and systems will be necessary to support a range of goals across the community (e.g., developing new ENM consumer products, designing nanotherapeutics, evaluating potential toxicity of multiple nanomaterial types).

The fundamental driver underlying all the layers of the nano curation discussion is to understand: What is it that must match between materials, systems, and data fields in order to enable comparisons? This project will move through that question by probing what is meant by each part of this fundamental question: What materials? What systems? What data fields? And what comparisons? The answer to these questions, as expected, will be "it depends". Our approach in writing this series of papers will be to systematically illuminate on what it depends, and why.

Critical sub-topics in nanomaterial data curation

A paper will be developed for each of a number of sub-topic groups relevant to nano curation (Table 1). We acknowledge the vast scope of the topics as outlined in Table 1, each of which is complex and relevant to informatics approaches within many other fields. This is a dynamic initiative and the list is provided as a starting point; it may grow and/or change over time through community dialogue and the identification of topical areas that are in need of exploration and clarification. We may also choose to condense and rearrange subtopics, but the list below represents the primary ideas generated collectively by the NCIP NanoWG, and reiterated by the participation of nanomaterial data curation stakeholders (to be discussed below). The currently planned series of NDCI papers is scheduled for production over the next two years, with the first manuscript accepted for publication in the Beilstein Journal of Nanotechnology, the following three in preparation, and the final topic being scoped by a designated author team.

Table 1: NDCI curation sub-topics.

#	sub-topic area	planned focus
1	curation workflows	Addresses workflow aspects such as curation protocols for consuming data from primary literature as well as data transfers between repositories or between data customers and data consumers. Discusses mechanisms for both primary curation of data into repositories and interoperable sharing between resources. A direct comparison of officially documented and/or informally institutionalized curation protocols will provide a clear baseline and allow concrete discussion of next steps for protocol standardization. Also addresses a starting point for the workflows in terms of sourcing, including various approaches for identifying sources: active sourcing, where the data repository does the work (either automated or manual) of identifying data sources, or passive sourcing where the dataset owners are the agents that seek access to the repository.
2	data completeness and quality	Includes discussion of both data quality and data completeness. Completeness is a measure of the raw data, assays, processed data, or derived data. What are different ways data completeness could be defined, and are these completeness criteria shaped of the goals for the data being curated? High quality data could still be sparse or "incomplete", so separately, what approaches are employed to define and evaluate data quality? This sub-topic encompasses issues such as precision, error, and sufficiency of meta-data for reproducibility. Are there differences when evaluating data quality captured from a database versus from the primary literature?
3	curation responsibilities	Covers curation responsibilities, including established and developing roles and division of curation labor and exploring the real challenges associated with quantity vs. quality of data entries. Curation training and performance expectations will also be addressed, as will the roles of other non-curators in defining the curation process (e.g. how might data "customers", such as peer-reviewed journals, influence the process).
4	integration between databases and datasets	How do we define and operationalize integration between databases and datasets? What level of interoperability is required to support data integration in a way that supports various goals for comparison and analysis? Specific topics that can be challenges to interoperability will be discussed, for example, questions such as what is the primary key – the root or kernel that makes an individual record unique? Some infrastructures base the primary key on the nanomaterial, whether on the batch, the lot level, or just the product name. Others utilize a particular study or experiment as the basis around which the structure is oriented. This definition of a unique entry into a database is fundamental to the structure of the database, often differs between different resources, and greatly impacts how data are curated from a source. Finding ways to map across these differences in record definition will be an important consideration.
5	metadata	The way metadata are handled within a database and within data records is critical to every other nanotechnology data curation topic listed. For example, environmental and biological media characterizations are critical for interpretation as well as comparison of data. Temporal metadata are also key; how experimental and characterization timing is incorporated to data collection and infrastructure is integral to enabling reproducibility of data and to achieving functional interoperability between datasets.

In each paper, we will examine each of the sub-topics, identified in Table 1 following this consistent discussion structure:

1. Why this sub-topic is important and relevant to the understanding of nanomaterial data curation, and the subsequent functional interoperability of datasets.
2. How does the purpose of an individual nanomaterial data resource or curation effort (e.g., to inform product development, to identify data gaps for research prioritization) impact (i) the approaches to this aspect of curation and (ii) particular challenges involved with this aspect of curation?
3. What are established handling methods for this sub-topic in mature fields (e.g., biological data curation)?
4. What are key challenges specific to emerging materials/nanomaterials with regard to this sub-topic? Are there any specific use cases to illustrate these issues and make them tangible?
5. What are some recommendations for advancing nanomaterial data curation in support of functional interoperability between datasets and resources: (i) Opportunities to leverage existing nanoinformatics resources (e.g. ISA-TAB-nano) in addressing integration for this sub-topic, or reasons not to do so? (ii) Practical next steps for individual stakeholders or the community as a whole?

Results and Discussion

For each sub-topic paper, information relevant to the discussion topics listed above will be gathered from a group of Stakeholder Liaisons who represent various organizations with activities related to curation of nanomaterial data. The role of the Stakeholder Liaisons will remain consistently defined throughout the NDCI series, but the make-up of the group is envisioned as dynamic. First, with increasing visibility of the project, it is the hope of the authors to gain more interest and widen participation in the Stakeholder Liaison group. While maximum retention will be sought for consistency and comparison across all topics, realistically the NDCI team realizes some individuals may choose to be involved in all papers within the series while others may elect to abstain from a given paper given interest or time constraints. In the interest of maximizing the scope of the baseline view of the nanocuration field, the NDCI will be inclusive of all Stakeholder Liaison responses. Our first step in this project was to identify these stakeholders through a series of inquires sent out by appropriate members of the NCIP NanoWG leadership team. Five organizations responded to our initial invitation recruiting Stakeholder Liaisons and provided answers to a set of foundational questions for this initial framing paper; their responses are presented in Tables 2–4 (see below). It is important to note that all Stakeholder Liaisons have been made explicitly aware that their names and institutions are associated with their responses to these questions, in an effort to foster a transparent discussion; all respondents were also provided the opportunity to review the final draft of this manuscript for as inclusive a process as possible. Several more have agreed to serve as Stakeholder Liaisons going forward on the other sub-topic papers, and we intend to continue expanding upon the initial group as this project moves forward. We will begin each sub-topic paper process by the NCIP NanoWG leadership team posing a set of questions to the Stakeholder Liaison group. A period of one month is allotted for response preparation, and the NDCI team has committed to circulating no more than one set of questions at a time to address the topics in series and to be mindful of the time and effort requirements placed on the Stakeholder Liaisons. As in this article, all stakeholder responses will be presented in the published articles to transparently represent the community perspectives; although as the liaison list grows, due to various limiting considerations of some participating organizations, decisions may be made to forego full liaison transparency in favor of being able to include the input of as broad as possible a swath of nanomaterial data stakeholders. Together, the responses provide a baseline snapshot of current practical experiences, and a range of views that will feed into a synthesized summary of recommendations addressing curation on behalf of the nanoinformatics community. The collection of this diverse and expanding group of stakeholder perspectives will foster development of preliminary recommendations for how to advance nanomaterial curation in principal and in practice, while identifying a community of practice in the process.

Establishing a baseline of nanomaterial curation considerations

For the current article, the NCIP NanoWG leadership team established communication with individuals in the current nanocuration Stakeholder Liaison group and posed three fundamental questions:

1. Briefly describe the scope (goals and research questions) of your data curation efforts.
2. What do you believe are the major challenges in nanoscience/nanotechnology data curation?
3. Within your effort, what data (information) is necessary to directly compare nanomaterials and determine if they are the same material?

As expected, responses showed variety in both purpose (of the resource and the organizations represented) and scope. In response to the first question, the responses show that the purpose of curation encompasses efforts across the life cycle of nanomaterials and the life cycle of datasets generated about nanomaterials (Table 2). Some efforts focus on capturing data at

Table 2: Liaison question #1.

liaison	affiliation	scope of data curation effort
Bill Zamboni	UNC	My research program at UNC is involved in the profiling and translational development of nanoparticle agents. My research program focuses on evaluating the pharmacokinetics (PK) and pharmacodynamics (PD) of nanoparticle agents in preclinical models and in patients. Specifically, we are involved in evaluating the factors that alter the function of the mononuclear phagocyte system (MPS) which then alters the PK and PD of nanoparticle agents in preclinical models and in patients. We have developed phenotypic probes of MPS function that predicts the PK and PD of nanoparticles in animals and patients. We are also developing a high throughput screen (HTS) of the interaction between nanoparticles and the MPS which predicts in vivo PK of the nanoparticles. The MPS HTS can be used to screen and select nanoparticles with high and low MPS uptake prior to going into in vivo studies. We are also evaluating how the MPS may be involved in the clearance and distribution of nanoparticles via capture (i.e. nanoparticle goes to the spleen and then is taken up by the MPS) and/or hijacking (i.e. the nanoparticle is taken up by the MPS cells in the blood and then delivered to tissues while inside the MPS cells).
Christoph Steinbach, Clarissa Marquardt	DaNa database NanoRA	The goal of our project is to provide impartial information and the real knowledge on safety aspects of (manmade) nanomaterials. DaNa in the acronym for DAtabase NAnomaterials but today we prefer talking about our Knowledgebase Nanomaterials and that describes our goals very well: We try to separate publications which are suitable for assessment of safety aspects of nanomaterials from those who are not suitable. So we try to collect not only arbitrary data but scientifically proven knowledge. The need to perform such kind of assessment is documented e.g., in a publication by Hristozov et al. [11].
Marina (Nina) Vance	nanotechnology Consumer Products Inventory	Our curation effort is centered on the nanotechnology Consumer Products Inventory (CPI). The CPI was developed by the Woodrow Wilson International Center of Scholars in 2005 and it is currently the most comprehensive listing of consumer products that contain or claim to contain nanomaterials. The main goal of the CPI is to document the way in which nanotechnology is entering the consumer market. Specifically, we want to provide the science and regulatory communities, as well as consumers, with current and accurate information about nano-enabled consumer products and the nanomaterials they contain.
Christine Ogilvie Hendren	CEINT NIKC (Center for Environmental Implications of NanoTechnology NanoInformatics Knowledge Commons)	Our curation effort is centered around interrogating the data gathered from across the Center for Environmental Implications of Nanotechnology along with comparative literature from throughout the field external to the center. Though our controlled material sourcing has created a rich integrated dataset as a starting point, we have a wide range of data types and fields, representing our focus on complex environmental interactions and transformations as well as impacts across a biological continuum and including ecosystem-wide measures. Our central research goals driving the data integration process are to 1) Probe mechanistic relationships between material and system properties and their combined effects on nanomaterial fate and effect in the environment, 2) Organize our disparate data to provide directional guidance to risk assessors even prior to achieving goal 1, and 3) Test our hypotheses that a amassing data on a small number of semi-empirical functional assays measurements will allow us to further goals 1 and 2. Beyond supporting CEINT mission-focused research questions, two key goals of our data integration project are to build a cyberinfrastructure that captures the data in a way that enables reproducibility and quality control down the road, and to ultimately develop associated tools to involve researchers in self-curation of their data so they can shorten the curation timeline and realize the benefits of analyzing their data together with other comparable datasets.
Julio Cesar Facelli, David Eugene Jones	NanoSifter (University of Utah)	The purpose of the NanoSifter project here at the University of Utah is to create a natural language processing (NLP) tool which is capable of extracting nanoparticle data associated to nanoparticle properties directly from the primary literature. Currently, the tool can extract data associated to hydrodynamic diameter, particle diameter, molecular weight, zeta potential, cytotoxicity, IC_{50}, cell viability, encapsulation efficiency, loading efficiency, and transfection efficiency. We plan to expand the information that NanoSifter can extract, while also improving the precision, recall, and f-measure of this tool.

the point of generation (academic or industrial research), and some focus on capturing data after its packaging and release in publications. Stakeholder representation from across the ENM-product life cycle presents an opportunity to identify and enable information hand-offs that facilitate targeted integration of nanomaterial data. The differences in curation scope will allow

us to explore the extent to which curation practices need to be the same in order to enable data comparison. In addition, we may be able to identify whether or not there are drivers to integrate datasets between organizations with very specific and more general scopes.

The stakeholder responses to the second question we posed on challenges to curation (Table 3) include aspects of every subtopic area to be addressed within this project, including social aspects, such as reluctance to share, data quality issues, ontology development and adoption decisions, and a simple lack of data. Other issues listed pertained to larger epistemological issues pervasive throughout the field of nano science. These included uncertainty about which material and system parameters are appropriate for predicting material behavior and interaction; and the struggle to make near-term decisions based on emerging science.

The stakeholder responses to the open-ended question on comparison of nanomaterials all honed in on the critical question

Table 3: Liaison question #2.

liaison	affiliation	major challenges to curation of nanomaterial data
Bill Zamboni	UNC	The complexity and high variability nature of MPS function in animal models and patients which results in high PK and PD variability of nanoparticles. The current inability to predict nanoparticle PK and PD in vivo based on standard critical micelle concentration (CMC)-like measurements (e.g., size and charge). The need to evaluate the interaction between the MPS and nanoparticles early in development and even before going into in vivo studies.
Christoph Steinbach, Clarissa Marquardt	DaNa database NanoRA	We think we are taking care of one of the most important challenges in nanomaterials data curation: separating valid from invalid data. In this regard, the major challenge is to gain information on the identity of a nanomaterial in a given study, which involves a careful physical-chemical characterization of a nanomaterial. Most of the data we consider invalid has a lack of information on material properties, which also hampers comparability of studies. Moreover, the collection of standard operating protocols (SOPs) or harmonized protocols for nanotox-testing is the second important challenge we want to address within the next four years. From a more information technological point of view, the development of suitable data models and adequate ontological structures to support next generation electronic infrastructures is another challenge.
Marina (Nina) Vance	Nanotechnology Consumer Products Inventory	One major challenge we face is a general lack of support from the nanotechnology industry. Secrecy is inherent to the product development strategy of most companies, which makes it very difficult to provide a detailed characterization of industrial nanomaterials. A potential contributing factor to this problem, which applies specifically to the CPI, is a fear that association to the CPI may negatively affect the image of the consumer products. Another challenge we face in curating the CPI is keeping it up to date with the fluidity of the consumer market. Consumer products come and go daily, their names and models change over time, as do their companies' websites. To attempt to tackle this issue, we have added crowdsourcing capabilities to the CPI, so that interested consumers, manufacturers, or researchers can enter new data or suggest edits to any entry. Now, our main challenge is to catalyze the participation of the CPI contributors.
Christine Ogilvie Hendren	CEINT NIKC (Center for Environmental Implications of NanoTechnology NanoInformatics Knowledge Commons)	Absence of established data-sharing protocols for existing measurement techniques (not to mention those that are currently being developed). Complexity of the interactions of nanomaterials in the environment, and large numbers of influential parameters governing transformations. Wide range of variety in systems studied and particular parameters reported in those systems. How time points are handled with respect to explaining when materials were characterized, measured along the trajectory of a long-term experiment is a challenge; this gets back to our driving goal of creating a database that supports reproducibility and multi-study comparison.
Julio Cesar Facelli, David Eugene Jones	NanoSifter (University of Utah)	In my opinion, there are a number of major challenges in nanoscience/nanotechnology data curation. The first is developing standards and protocols to report data in the literature which the nanoscience/nanotechnology community adheres to and follows. There are so many different ways that properties of nanoparticles can be reported in the literature, which makes the retrieval of such information quite cumbersome. Another major challenge is further development of the nanoparticle ontology (NPO) to add more functionality, metadata, and relationships to the ontology.

begged by asking what materials are the same: what do we mean by “sameness”? Similar definitional questions arose around curation resource purpose (Table 4).

From these initial framing questions alone, it is clear that in order to make progress in integrating data through consistent nano curation processes, and to achieve functional interoperability that will render efforts to establish nanoinformatics fruitful, the nanomaterial community will have to maintain a focus on the need for purpose-based integration. Therefore through interaction with stakeholder liaison that will follow this inaugural publication, and the synthesis of their input, we will distill the recommended tenets of nanomaterial data curation both in terms of baseline requirements for all nanoinformatics activities as well as for a range of purposes.

The experience to date in the NCIP NanoWG and in assembling the NDCI has already begun addressing the third NDCI

Table 4: Liaison Question #3.

liaison	affiliation	data deemed necessary for nanomaterial comparison
Bill Zamboni	UNC	The need to be able to evaluate encapsulated/conjugated and released drug as part of formulation development and as part of in vivo PK studies. The need to evaluate biodistribution differences to tumor, tissues and the MPS. The need to evaluate the bi-directional interaction between nanoparticles and the MPS.
Christoph Steinbach, Clarissa Marquardt	DaNa database NanoRA	A very good question which is extremely hard to answer: What does “same material” mean, not only from the informational point of view but also from the other side, the definition of “same material”? Which set of parameters do you need? Even if you change the size or shape of a particle totally different behavior can be achieved. We have developed a set of criteria (see http://www.nanopartikel.info/files/methodik/DaNa-Literature-Criteria-Checklist_Methodology.pdf) which need to be fulfilled that we accept a certain publication as “knowledge” in the meaning described in the answer to the first question. Here we also describe the material characterization criteria. In fact we are absolutely aware that this does not make finally sure, that we are always talking of the “same” material, but for our purposes it’s enough. We think that a lot of further research is necessary to determine the right “same material” parameters. Furthermore the comparability in nano-sciences does not end with the “same” material as it is shown in certain round robin experiments [12,13]. Does it help when you assume to have the same material and the following experiments show different results because of other factors? I do not know if that leads to a better solution: Perhaps some kind mathematical probability that tells us *x* parameters (out of *y* parameters which can be determined with today’s characterization methods) of one substance are the same for another. The higher the number of same parameters the higher the probability the two substances are the “same”?
Marina (Nina) Vance	Nanotechnology Consumer Products Inventory	Within the CPI, it is very difficult to determine if a nanomaterial present in two or more products is, in fact, the same. We can group nanomaterials of the same composition together, but without a detailed description from the manufacturer, that would be impossible. In order to directly compare nanomaterials within consumer products, we would need, in the very least, the following: Composition, Shape, Size, Composition of coatings, Crystallinity
Christine Ogilvie Hendren	CEINT NIKC (Center for Environmental Implications of NanoTechnology NanoInformatics Knowledge Commons)	This depends on the level of granularity in the comparison. We believe that in order to support comparison and analysis in support of our research goals (elucidate mechanisms governing nanomaterial behavior and translate this into forecasts of risk), what is absolutely required are intrinsic characteristics of the nanomaterial, the surrounding system characteristics (e.g., be the system lab controlled, environmental media, biological systems), and system-dependent or "extrinsic" material characteristics. Only when all of these aspects, and their appropriate corresponding metadata describing preparation and testing protocols, are consistently reported can we know that direct comparison of two datasets is possible.
Julio Cesar Facelli, David Eugene Jones	NanoSifter (University of Utah)	The data (information) that is most necessary to directly compare nanomaterials and determine if they are the same material are the molecular descriptors and biochemical activity of the nanomaterials. The molecular descriptors (e.g., molecular weight, hydrodynamic diameter) and biochemical activity (e.g., cytotoxicity, cell viability, transfection efficiency) of the nanomaterials can be used by data mining and machine learning methods to compare materials and determine their similarity if the materials are discrete compounds. If the materials are not discrete compounds (i.e., polymers), properties such as molecular weight distribution and polydispersity will be the properties to assess for comparison of materials.

goal of facilitating the interdisciplinary and trans-sector collaborations that we believe will be critical ingredients in successful advancement of nanoinformatics efforts. The team-writing experience within the author teams of the NDCI topic papers includes promising aspects that can foster collaborations. For each topic paper, a group of self-selected NCIP NanoWG members are volunteering to lead the topic, assembling author groups that, in the case of the four papers already being undertaken, often consist of people who have never collaborated or published together prior, and soliciting the broad input provided by Stakeholder Liaisons across the nanomaterial data community. New connections are being made between individuals and organizations, and for each topic these new teams are working through the available literature across a variety of academic disciplines, synthesizing the baseline input from Stakeholder Liaisons, and shaping recommendations and future questions for the consideration of the growing nanoinformatics field. Though there are not direct Stakeholder Liaison interactions planned as part of the NDCI, the transparency and sharing of their responses through the NDCI series will offer fertile ground for potential communication and collaboration between like or complimentary groups in future efforts. Lastly, the recommendations emerging from the NDCI series will no doubt include suggestions on opportunities regarding the potential for linkages and collaborations.

We welcome input from the nanomaterial community on the approach for the project laid out in this article and encourage continued feedback as the effort moves forward, including via participation from growing list of nanomaterial data stakeholders. Interested community members can share feedback or join the NCIP by visiting to https://nciphub.org/, and can learn more about the NDIC in particular by visiting https://nciphub.org/groups/nanotechnologydatacurationinterestgroup/wiki/MainPage.

Acknowledgements

The authors gratefully acknowledge the contributions of the NDCI Stakeholder Liaisons whose responses to our initial questions on nanomaterial data curation, and willingness to participate in this transparent process, enable the NDCI to capture a baseline of current nanomaterial data curation views and practices. We also benefited greatly from the thoughtful commentary, input, and organizational support of Dr. Juli Klemm and Mervi Heiskanen of the National Cancer Institute Center for Biomedical Informatics and Information Technology as well as the contributions of Dr. Martin Fritts of SAIC Frederick. We are grateful for the funding from the NIH National Cancer Informatics Program, which supports the facilitation of the NCIP Nanotechnology Working Group, resulting in the collaborations that enable the NDCI and associated products. We also acknowledge the discussions and input from many of our NCIP Nanotechnology Working Group colleagues throughout the growing nanoinformatics community, who have helped shape the concept of this manuscript with their expertise. C.O.H. would like to acknowledge the Center for the Environmental Implications of NanoTechnology (CEINT) funding from National Science Foundation (NSF) and the Environmental Protection Agency (EPA) under NSF Cooperative Agreement DBI-1266252 and EF-0830093. S.L.H would like to acknowledge support provided by the National Institute of Health (grant # ES017552-01A2). The views, opinions, and content in this article are those of the authors and do not necessarily represent the views, opinions, or policies of their respective employers or organizations. Mention of company names or products does not constitute endorsement. The authors declare no competing interests.

References

1. Hansen, S. F.; Maynard, A.; Baun, A.; Tickner, J. A. *Nat. Nanotechnol.* **2008,** *3,* 444–447. doi:10.1038/nnano.2008.198
2. Maynard, A. D. *Nat. Nanotechnol.* **2014,** *9,* 159–160. doi:10.1038/nnano.2014.43
3. National Resource Council. *A Research Strategy for the Environmental, Health and Safety Aspects of Engineered Nanomaterials;* The National Academies Press: Washington, DC, U.S.A., 2012.
4. National Research Council. *Council Research Progress on Environmental, Health and Safety Aspects of Engineered Nanomaterials;* The National Academies Press: Washington, DC, U.S.A., 2013.
5. National Science and Technology Council. *Materials Genome Initiative for Global Competitiveness;* National Science and Technology Council, Office of Science and Technology Policy: Washington, DC, U.S.A., 2011.
6. Thomas, D. G.; Gaheen, S.; Harper, S. L.; Fritts, M.; Klaessig, F.; Hahn-Dantona, E.; Paik, D.; Pan, S.; Stafford, G. A.; Freund, E. T.; Klemm, J. D.; Baker, N. A. *BMC Biotechnol.* **2013,** *13,* 2. doi:10.1186/1472-6750-13-2
7. de la Iglesia, D.; Harper, S.; Hoover, M. D.; Klaessig, F.; Lippell, P.; Maddux, B.; Morse, J.; Nel, A.; Rajan, K.; Reznik-Zellen, R.; Touminen, M. T. *Nanoinformatics 2020 Roadmap;* InterNano Nanomanufacturing Library, 2011. doi:10.4053/rp001-110413
8. Hoover, M. D.; Myers, D. S.; Cash, L. J.; Guilmette, R. A.; Kreyling, W. G.; Oberdörster, G.; Smith, R.; Cassata, J. R.; Boecker, B. B.; Grissom, M. P. *Health Phys.* **2015,** *108,* 179–194. doi:10.1097/HP.0000000000000250
9. Cragin, M. H.; Heidron, P. B.; Palmer, C. L.; Smith, L. C. An Educational Program on Data Curation. In *American Library Association Science & Technology Section Conference,* 2007.
10. Borgman, C. L. *Big data, little data, no data: scholarship in the networked world;* MIT Press: Boston, MA, U.S.A., 2015.
11. Hristozov, D. R.; Gottardo, S.; Critto, A.; Marcomini, A. *Nanotoxicology* **2012,** *6,* 880–898. doi:10.3109/17435390.2011.626534

12. Roebben, G.; Ramirez-Garcia, S.; Hackley, V.; Roesslein, M.; Klaessig, F.; Kestens, V.; Lynch, I.; Garner, C.; Rawle, A.; Elder, A.; Colvin, V.; Kreyling, W. G.; Krug, H. F.; Lewicka, Z.; McNeil, S.; Nel, A.; Patri, A.; Wick, P.; Wiesner, M.; Xia, T.; Oberdörster, G.; Dawson, K. *J. Nanopart. Res.* **2011,** *13,* 2675–2687.
13. Xia, T.; Hamilton, R. F., Jr.; Bonner, J. C.; Crandall, E. D.; Elder, A.; Fazlollahi, F.; Girtsman, T. A.; Kim, K.; Mitra, S.; Ntim, S. A.; Orr, G.; Tagmount, M.; Taylor, A. J.; Telesca, D.; Tolic, A.; Vulpe, C. D.; Walker, A. J.; Wang, X.; Witzmann, F. A.; Wu, N.; Xie, Y.; Zink, J. I.; Nel, A.; Holian, A. *Environ. Health Perspect.* **2013,** *121,* 683–690. doi:10.1289/ehp.1306561

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Nanotechnology in the real world: Redeveloping the nanomaterial consumer products inventory

Marina E. Vance*[1], Todd Kuiken[2], Eric P. Vejerano[3], Sean P. McGinnis[4], Michael F. Hochella Jr.[5], David Rejeski[2] and Matthew S. Hull[1]

Full Research Paper

Open Access

Address:
[1]Institute for Critical Technology and Applied Science, Virginia Tech, 410 Kelly Hall (0194), 235 Stanger St., Blacksburg, VA 24061, United States, [2]Woodrow Wilson International Center for Scholars, One Woodrow Wilson Plaza - 1300 Pennsylvania Ave., NW, Washington, DC 20004, United States, [3]Department of Civil & Environmental Engineering, Virginia Tech, 418 Durham Hall (0246), Blacksburg, VA 24061, United States, [4]Department of Materials Science and Engineering, Virginia Tech, Holden Hall (0237), Blacksburg, VA 24061, United States and [5]Department of Geosciences, Virginia Tech, 4044 Derring Hall (0420), Blacksburg, VA 24061, United States

Email:
Marina E. Vance* - marinaeq@vt.edu

* Corresponding author

Keywords:
consumer products; database; inventory; nanoinformatics; nanomaterials

Beilstein J. Nanotechnol. **2015,** *6,* 1769–1780.
doi:10.3762/bjnano.6.181

Received: 28 March 2015
Accepted: 07 August 2015
Published: 21 August 2015

This article is part of the Thematic Series "Nanoinformatics for environmental health and biomedicine".

Guest Editor: R. Liu

Abstract

To document the marketing and distribution of nano-enabled products into the commercial marketplace, the Woodrow Wilson International Center for Scholars and the Project on Emerging Nanotechnologies created the Nanotechnology Consumer Products Inventory (CPI) in 2005. The objective of this present work is to redevelop the CPI by leading a research effort to increase the usefulness and reliability of this inventory. We created eight new descriptors for consumer products, including information pertaining to the nanomaterials contained in each product. The project was motivated by the recognition that a diverse group of stakeholders from academia, industry, and state/federal government had become highly dependent on the inventory as an important resource and bellweather of the pervasiveness of nanotechnology in society. We interviewed 68 nanotechnology experts to assess key information needs. Their answers guided inventory modifications by providing a clear conceptual framework best suited for user expectations. The revised inventory was released in October 2013. It currently lists 1814 consumer products from 622 companies in 32 countries. The Health and Fitness category contains the most products (762, or 42% of the total). Silver is the most frequently used nanomaterial (435 products, or 24%); however, 49% of the products (889) included in the CPI do not provide the composition of the nanomaterial used in them. About 29% of the CPI (528 products) contain nanomaterials suspended in a variety of liquid media and dermal contact is the most likely exposure scenario from their use. The majority (1288 products, or 71%) of the products do not present enough supporting information to corroborate the claim that nanomaterials are used. The modified CPI has

enabled crowdsourcing capabilities, which allow users to suggest edits to any entry and permits researchers to upload new findings ranging from human and environmental exposure data to complete life cycle assessments. There are inherent limitations to this type of database, but these modifications to the inventory addressed the majority of criticisms raised in published literature and in surveys of nanotechnology stakeholders and experts. The development of standardized methods and metrics for nanomaterial characterization and labelling in consumer products can lead to greater understanding between the key stakeholders in nanotechnology, especially consumers, researchers, regulators, and industry.

Introduction

Advancements in the fields of nanoscience and nanotechnology have resulted in myriad possibilities for consumer product applications, many of which have already migrated from laboratory benches into store shelves and e-commerce websites. Nanomaterials have been increasingly incorporated into consumer products, although research is still ongoing on their potential effects to the environment and human health. This research will continue long into the future.

To document the penetration of nanotechnology in the consumer marketplace, the Woodrow Wilson International Center for Scholars and the Project on Emerging Nanotechnology created the Nanotechnology Consumer Product Inventory (CPI) in 2005, listing 54 products [1]. This first-of-its-kind inventory has become one of the most frequently cited resources showcasing the widespread applications of nanotechnology in consumer products. In 2010, the CPI listed 1012 products from 409 companies in 24 countries. Even though it did not go through substantial updates in the period between 2010 and 2013, it continued being heavily cited in government reports [2] and the scientific literature – the website http://www.nanotechproject.org has been cited over 2,580 times in articles according to Google Scholar – and became a popular indicator of the prevalence of nanotechnology in everyday life and the need to further study its potential social, economical, and environmental impacts [3-6]. The CPI has also been criticized due to its lack of science-based data to support manufacturer claims. Other longstanding suggestions for improvement included: more frequent updates, indications when products were no longer available for purchase by consumers, and the inclusion of more product categories to improve the searchability of the CPI database [7].

Since the creation of the CPI, other nanotechnology-related inventories have been developed around the world. In 2006, a German company launched a freely accessible internet database of nanotechnology products [8]. The website associated with this database was not accessible at the time of this writing and its last available record is from May 2014, when 586 products were listed. In 2007, Japan's National Institute of Advanced Industrial Science and Technology created an inventory of "nanotechnology-claimed consumer products" available in Japan [2]. This inventory is freely accessible online and it acknowledges the CPI in its website. At the time of this writing, the inventory listed 541 product lines and 1241 products; its last update occurred in 2010 [9]. In 2009, two European consumer organizations, the European Consumers Organization (BEUC) and the European Consumer Voice in Standardization (ANEC), joined efforts to develop an inventory of "consumer products with nano-claims" available to consumers in Europe [10]. A new inventory was generated annually from 2009 to 2012, but the 2011 and 2012 versions focused exclusively on products containing silver nanoparticles (nanosilver); the latest version in 2012 listed 141 nanosilver products. This inventory does not provide a searchable online database, but it can be downloaded for free as an Excel spreadsheet. In 2012, the Danish Consumer Council and Ecological Council and the Technical University of Denmark's Department of Environmental Engineering launched "The Nanodatabase", an inventory of products available for purchase that are claimed to contain nanomaterials and are available in the European consumer market [11]. This inventory has been continually updated and it currently lists 1423 products.

These worldwide efforts to understand the transition of nanotechnology from the laboratory bench to the commercial marketplace substantiate the need for applying the concept of nanoinformatics to a nanotechnology-enabled consumer products database, which is to determine the most relevant and useful information needed by a variety of stakeholders and to develop tools for its most effective use [12]. Databases such as the CPI offer information useful and relevant to a variety of stakeholders who are interested in a) understanding which consumer products incorporate nanotechnology and b) developing strategies, tools, and policies that may be needed to ensure safe and responsible use of those products.

Nanomaterials are regulated without specific provisions in the U.S. as hazardous chemical substances and pesticides, under the EPA's Toxic Substances Control Act (TSCA) [13] and the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) [14]. When used as food additives, drugs, or cosmetics, nanomaterials are regulated under the Federal Food, Drug, and Cosmetic Act (FFDCA).

In the European Union, nanomaterials are regulated under the Concerning the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) and the Classification, Labeling, and Packaging (CLP) regulations when those are classified by the Commission as hazardous chemical substances [15]. The Biocidal Products Regulation (BPR) has special provisions for biocidal materials that consist of nanoparticles, aggregates, or agglomerates in which at least 50% of primary particles have at least one dimension between 1 and 100 nm, with no provisions for "novel properties" stemming from their small size [16]. Cosmetics that contain nanomaterials are also regulated by the European Commission, and although the use of nanoscale titanium dioxide is permitted, zinc oxide is not [17]. The German Federal Environment Agency performed an Impact Assessment of a European Register of Products Containing Nanomaterials and determined that when compared to the implementation of a variety of national registries, an unified European registry would bring many advantages, including a lower cost for industries and, ultimately, a registry would benefit consumers, companies, and governments [18].

The objective of this work was to modify the CPI to improve its functionality, reliability, and utility to the diverse group of stakeholders who have come to depend on it as a critical resource for current information on nano-enabled consumer products. Specific objectives were (1) to update the CPI data to gain an insight into the penetration of nanotechnology in the consumer products market over the past decade; (2) to determine and implement improvements to the CPI based on the scientific literature and a survey of nanotechnology experts and CPI users; and to (3) develop a sustainable model to facilitate future CPI maintenance using crowdsourcing tools.

Below, we present a brief history of this inventory over a decade of existence. We also describe the specific changes made in the inventory during this project (referred here as CPI 2.0). Finally, we present an overview of the current data present in the CPI after the completion of this project.

Results and Discussion

CPI growth over time

Table 1 lists the growth of the CPI since 2005. In 2011, before this current project, the CPI described 1314 products. Since then, 489 products that are no longer available or marketed as containing nanotechnology have been archived and 500 products have been added. The new total of 1814 products as of March 2015 represents a thirty-fold increase over the 54 products originally listed in 2005 – which is not a complete representation of the growth of this market, as our methodology has also evolved over time. Based on our review, the CPI is the largest online inventory of nanotechnology consumer products available. Products come from 622 companies located in 32 countries (Supporting Information File 1, Table S1).

The products listed on the CPI 2.0 satisfy three criteria: (1) they can be readily purchased by consumers; (2) they are claimed to contain nanomaterials by the manufacturer or another source; and (3) their claim to contain nanomaterials appears reasonable to CPI curatorial staff.

Although the steady growth of the inventory indicates that the popularity of products claimed to incorporate nanotechnology is continually increasing, not all products have persisted in the consumer market. In the past seven years, 34% of the entries in the inventory have been archived because the product is not currently available in the market or their claim to contain nanotechnology can no longer be verified. One example of a claim that can no longer be verified is a product that is still available for purchase on a manufacturer's website but no longer references, explicitly, the incorporation of nanotechnology into that

Table 1: Number of products in the CPI over time.

Year	Total products	Products added	Products archived	Data collection notes
2005	54	54	0	Beginning of CPI as a static pdf document.
2006	356	302	0	Launch of the online CPI.
2007	580	278	0	Nanoscale silver emerged as most cited nanomaterial.
2008	803	223	0	Health and fitness products represented 60% of the inventory.
2009	1015	212	107	Added archiving function to the CPI.
2010	1015	0	0	No data collected.
2011	1015	0	0	No data collected.
2012	1438	426	0	Beginning of CPI 2.0 project, focus on adding new products.
2013	1628	190	288	Launch of crowdsourcing component. Extensive effort put into adding and archiving products.
2014	1814[a]	238[a]	223[a]	Extensive effort put into adding and archiving products.

[a]The CPI now has crowdsourcing capabilities, so these numbers are a snapshot in time and will not represent the CPI at the time of reading.

product. Even after archiving, a product can return to the main inventory listing if a third party makes the claim that the product indeed contains nanomaterials or if the manufacturer restates their nanomaterial claim.

In the CPI, entries are grouped under eight generally accepted consumer goods categories that are loosely based on publicly available consumer product classification systems (Figure 1) [19]. The Health and Fitness category includes the largest listing of products in the CPI, comprising 42% of listed products (excluding archived products). Within the Health and Fitness category, Personal Care products (e.g., toothbrushes, lotions, and hairstyling tools and products) comprise the largest subcategory (39% of products). Starting in 2012, a large continual effort has been put into periodically checking products for their current availability and current claim to contain nanotechnology. This effort resulted in archiving 316 products in the Health and Fitness category – mainly in the Personal Care and Clothing subcategories – with 86 and 78 products archived between 2012 and 2014, respectively.

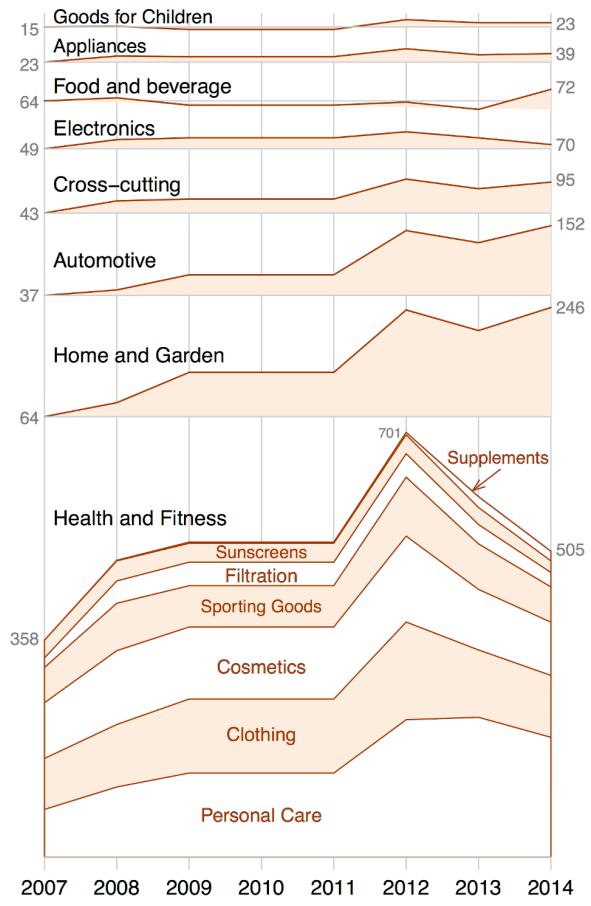


Figure 1: Number of available products over time (since 2007) in each major category and in the Health and Fitness subcategories.

New nanomaterial descriptors

Eight new product descriptors were introduced to facilitate the use of this database by a variety of stakeholders (namely industry and the scientific and regulatory communities):

1. main nanomaterial composition or type,
2. nanomaterial shape and size,
3. nanomaterial coating or stabilizing agent,
4. nanomaterial location within the product,
5. nanomaterial function in the product,
6. potential exposure pathways,
7. "how much we know",
8. "researchers say".

The experimental section of this paper describes all new product descriptors. The results of the five new quantitative descriptors are presented and discussed below. Since the "nanomaterial shape and size", "coating and stabilizing agent", and the "researchers say" categories are text-entry data fields, thus qualitative information at this point, we have not included their analysis in this paper.

Nanomaterial composition

Of the 1814 products listed in the CPI, 47% (846 products) advertise the composition of at least one nanomaterial component and 62 of those products list more than one nanomaterial component (e.g., a product comprised of both silver and titanium dioxide nanomaterials). There are 39 different types of nanomaterial components listed in the inventory (listed in Supporting Information File 1, Table S2), which have been grouped into five major categories in Figure 2 and Figure 4, to improve their legibility: metal, carbonaceous, silicon, not advertised, and other. Nominally, metals and metal oxides comprise the largest nanomaterial composition group advertised in the inventory, listed in 37% of products.

Titanium dioxide (TiO_2), silicon dioxide, and zinc oxide are the most produced nanomaterials worldwide (on a mass basis) and the global annual production of silver nanoparticles represents only 2% of that of TiO_2 [20,21]. However, silver nanoparticles are the most popular advertised nanomaterial in the CPI, present in 438 products (24%). The CPI reports the numbers of different consumer products and product lines available in the market, so there is no implication on mass, volume, or concentration of nanomaterials incorporated into products or the production volume of each product.

Of carbonaceous nanomaterials (89 products), the majority of products listed contains carbon nanoparticles (sometimes described as carbon black, 39 products) and single- or multi-walled carbon nanotubes (CNT, 38 products). Unfortunately,

891 (49%) of the products included in the CPI do not present the composition or a detailed description of the nanomaterial used (Figure 2).

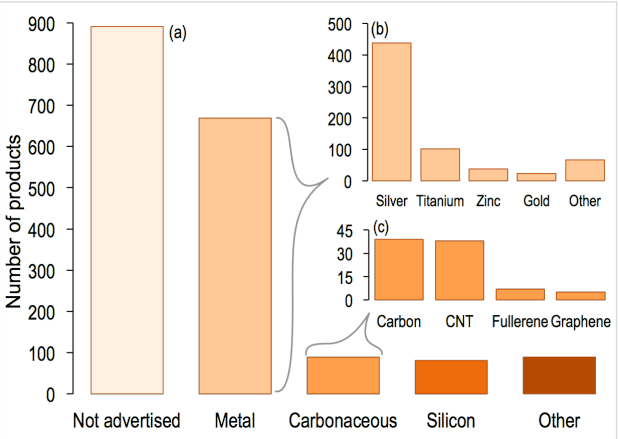


Figure 2: (a) Claimed composition of nanomaterials listed in the CPI, grouped into five major categories: not advertised, metal (including metals and metal oxides), carbonaceous nanomaterials (carbon black, carbon nanotubes, fullerenes, graphene), silicon-based nanomaterials (silicon and silica), and other (organics, polymers, ceramics, etc.). (b) Claimed elemental composition of nanomaterials listed in the metals category: silver, titanium, zinc, gold, and other metals (magnesium, aluminum oxide, copper, platinum, iron and iron oxides, etc.). (c) Claimed carbonaceous nanomaterials (CNT = carbon nanotubes).

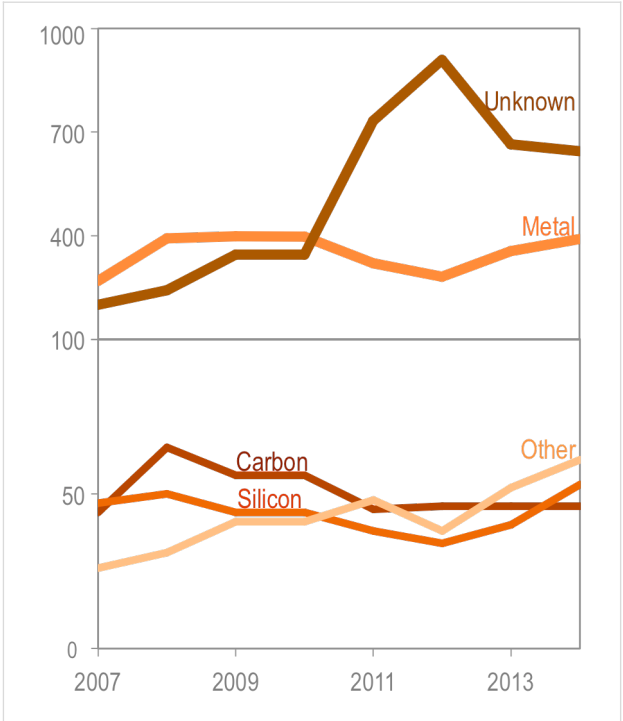


Figure 3: Major nanomaterial composition groups over time. Carbon = carbonaceous nanomaterials (carbon black, carbon nanotubes, fullerenes, graphene). Other = organics, ceramics, polymers, clays, nanocellulose, liposomes, nano micelles, carnauba wax, etc. Note the difference in scale between the top and bottom panels in this plot.

The percentages of nanomaterial compositions in the CPI 2.0 are somewhat in agreement with those of the Danish Nanodatabase. The Nanodatabase also lists a high fraction of products with unknown nanomaterial composition (944 products or 66%) and, among known compositions, silver is also the most frequently advertised nanomaterial component, with 207 products or 14.5% [11]. Silver nanoparticles are popular consumer product additives due to their well-documented antimicrobial properties [22].

Figure 3 shows how the availability of these major nanomaterial composition groups changed over time. Since the start of the CPI 2.0 project (2012), products with unknown (not advertised) nanomaterial compositions have decreased by 12%, which is partially due to these products being archived and of their composition being identified and added to the inventory. Products advertising to contain metal and metal oxide nanomaterials, silicon-based nanomaterials (mostly SiO_2 nanoparticles), and a variety of other nanomaterial components (organics, ceramics, polymers, clays, nanocellulose, liposomes, nano micelles, carnauba wax, etc.) have been growing in popularity. During the same period, carbonaceous nanomaterials have remained stable at around 50 products available in the market.

Of the 846 products listed in the CPI for which we were able to determine a nanomaterial composition, 61 products (7%) advertise to contain more than one main nanomaterial component. Figure 4 presents 11 nanomaterial components that were most frequently listed with others in the same product.

Silver and titanium dioxide are the nanomaterial components most likely to be combined with other nanomaterials in consumer products, with 35 and 30 product combinations, respectively. Silver and titanium dioxide were paired with each other in 10 products (cosmetics and electronics); titanium dioxide and zinc oxide were paired in 10 products (sunscreens, cosmetics, and paints). The European Commission's Cosmetics Regulation has permitted the use of nanoscale titanium dioxide in sunscreens, but not zinc oxide [17].

Calcium and magnesium were listed together in dietary supplements. Nano-ceramics and silver are used in combination in water filtration products, cosmetics, and a humidifier. These results demonstrate the use of nanohybrids [23] in consumer products and indicate that the use of nanotechnology-based consumer products in the home may, in some cases, lead to multiple exposures from a combination of nanomaterial compositions. These results suggest the need to examine nanomaterial toxicity effects that could be synergistic, additive, or even antagonistic.

	Gold	Carbon	Copper	Ceramics	Iron	Calcium	Magnesium	Silicon	Zinc oxide	Titanium	Silver
Gold	7	1		1				1		1	3
Carbon	1	10	1		2					2	3
Copper		1	10		1	1	1	2	1		1
Ceramics	1			11				1		2	7
Iron		2	1		11	1	1			2	1
Calcium			1		1	13	8	1			
Magnesium			1		1	8	13	1			
Silicon	1		2	1		1	1	14	1	2	4
Zinc oxide			1					1	14	10	2
Titanium	1	2		2	2			2	10	30	10
Silver	3	3	1	7	1			4	2	10	35

Figure 4: Major nanomaterial composition pairs in consumer products. Carbonaceous nanomaterials (carbon black, carbon nanotubes, fullerene, and graphene) were combined into the same category (carbon). Grey boxes in the diagonal represent the total times each nanomaterial composition has been listed with other compositions in the same product.

Nanomaterial location

About 29% of consumer products in the CPI (528 products) contain nanomaterials suspended in a variety of fluids (e.g., water, skin lotion, oil, car lubricant). The second largest group in this category – with 307 products – comprises solid products with surface-bound nanoparticles (e.g., hair curling and flat irons, textiles). Figure 5 shows the location of nanomaterials for which a composition has been identified [24].

The majority (64%) of carbonaceous nanomaterials are embedded in solid products, whereas products of all other compositions are more commonly suspended in liquid. Of the few bulk nanomaterials that are available for purchase by consumers, the largest group (42%) consists of metal and metal oxide nanomaterials. Metals and metal oxides were also the largest composition for surface-bound particles and those suspended in liquid products. The majority (67%) of products with nanostructured surfaces consist of nanomaterials of undetermined composition. An example of such product is a liquid or spray products that forms a nanofilm upon application over a surface. Of nanostructured bulk materials, the majority (57%) are silicon-based nanomaterials (e.g., computer processor parts). It is interesting to note that we expect nano-electronics to exist now in massive numbers of consumer products, such as mobile devices, where field effect transistors, the heart of chip technology, have components (sources, gates, collectors, channels) that are now in the nanoscale [25] and would fit into the nanostructured bulk category. However, because most of these products do not advertise their use of nanomaterials, we believe that they are grossly underrepresented in the CPI.

Nanomaterial function

Of the 1814 inventory entries, 1244 were grouped according to the expected benefits of adding such nanomaterials to the product (Figure 6). A significant portion of products in the CPI (31% of products analyzed) utilize nanomaterials – mostly silver nanoparticles, but also titanium dioxide and others – to confer antimicrobial protection. Nanomaterials such as titanium

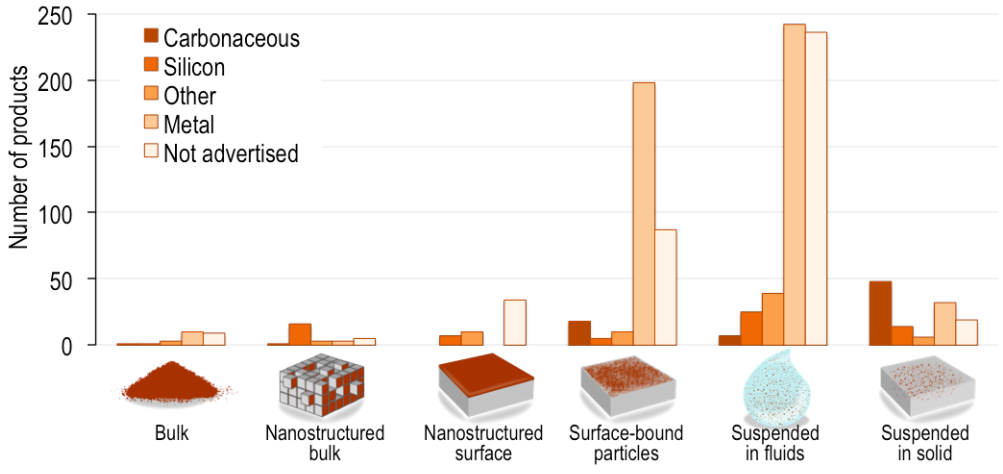


Figure 5: Locations of nanomaterials in consumer products for which a nanomaterial composition has been identified.

dioxide and silicon dioxide are used to provide protective coatings (15%) and for environmental treatment (to protect products against environmental damage or to treat air and water in the home, 15%). Cosmetic products (12%) are advertised to contain a variety of nanomaterials such as silver nanoparticles, titanium dioxide, nano-organics, gold, and others. A wide variety of nanomaterial compositions (silver, nano-organics, calcium, gold, silicon dioxide, magnesium, ceramics, etc.) were also advertised to be used for health applications, such as dietary supplements (11%).

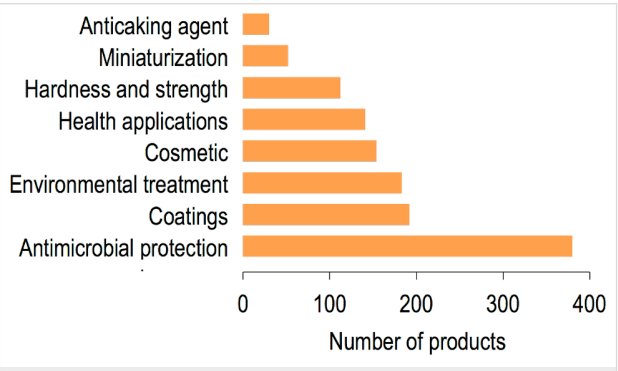


Figure 6: Expected benefits of incorporating nanomaterial additives into consumer products.

Potential exposure pathways

Since critical information such as nanomaterial size and concentration are not known for most products listed on the CPI, the actual health risks of these products remain largely unknown. Nevertheless, the CPI may be useful for inferring potential exposure pathways from the expected normal use of listed products. To investigate this utility, we analyzed a subset of 770 products from the CPI to determine their most likely route(s) of exposure (Figure 7).

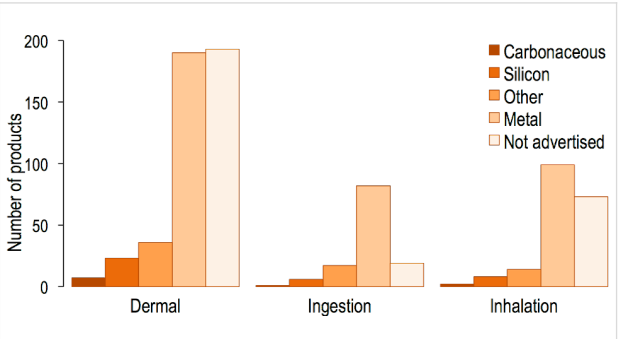


Figure 7: Potential exposure pathways from the expected normal use of consumer products, grouped by major nanomaterial composition categories.

We identified the skin as the primary route of exposure for nanomaterials from the use of consumer products (58% of products evaluated). This is because many entries in the CPI consist of (1) solid products that contain nanomaterials on their surfaces and are meant to be touched or (2) liquid products containing nanomaterial suspensions which are meant to be applied on the skin or hair. Of the products evaluated, 25% present nanomaterials that can possibly be inhaled during normal use (e.g., sprays and hair driers) and 16% contain nanomaterials that may be ingested (e.g., supplements and throat sprays). Hansen et al. developed a framework for exposure assessment in consumer products. In this framework, products that contain nanomaterials suspended in liquid and products that may emit airborne nanoparticles during use are expected to cause exposure [26].

Since metals and metal oxides are the most common nanomaterial composition in the CPI, they are also the most likely materials to which consumers will be exposed during the normal use of product via dermal, ingestion, and inhalation routes. Products containing nanomaterials of unknown composition are most likely to lead to exposure via the dermal route.

Berube et al. [7] offered a critique of the original CPI in 2010, which focused primarily on the lack of data pertinent to the dosages of nanomaterials to which consumers might be exposed through CPI-listed products. This is a valid criticism given that information used to populate the CPI is based primarily on marketing claims made by manufacturers. However, the most recent modifications of the CPI offer a potential remedy for data gaps through the contributions of third-party research teams. These modifications are especially timely as there is a growing number of published studies assessing consumer exposure to nanomaterials released during the use of nanotechnology-enhanced consumer products [27], such as cosmetic powders [28], sprays [29,30], general household products [31], and products for children [32,33]. One challenge is that there are no standardized methods for assessing consumer risks from using nanotechnology-enabled consumer products or a set of agreed-upon metrics for characterizing nanomaterials to determine environmentally relevant concentrations [34]. The development of such standards is seen as a top strategy for safe and sustainable nanotechnology development in the next decade [35]. The Consumer Product Safety Commission recently requested $7 million to establish the Center for Consumer Product Applications and Safety Implications of Nanotechnology to help develop methods to identify nanomaterials in consumer products and to understand human exposure to those materials [36].

How much we know

Through the “How much we know” descriptor, inventory entries are rated according to the reliability of the

manufacturer's claim that products contain nanomaterials. We evaluated 1259 products present in the inventory for the "How much we know" descriptor and the majority (71%) of products are not accompanied by information sufficient to support claims that nanomaterials are indeed used in the products, such as a manufacturer datasheet containing technical information about nanomaterial components (e.g., median size, size distribution, morphology, concentration). Only nine products have been classified in Category 1, "Extensively verified claim" due to the availability of scientific papers or patents describing the nanomaterials used in these products (Figure 8). The experimental section, below, presents a full description of these categories.

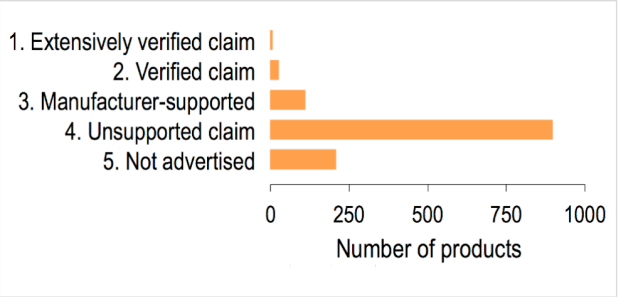


Figure 8: Distribution of products into the "How much we know" categories.

Hansen [37] performed interviews with 26 nanotechnology stakeholders who agreed on an incremental approach to nanomaterial regulation in consumer products, including classification and labeling. The European Commission's Classification, Labeling, and Packaging (CLP) regulation covers nanomaterials that are classified by the Commission as hazardous chemical substances [15]. Becker [38] reported that there are diverging opinions in the nanotechnology industry with regards to labeling, ranging from ''If it's a nano-scale material, people should know, hands down" to not supporting labeling because "it wouldn't accurately inform consumers of anything and would be bad for business because it would scare consumers."

Appropriate nanomaterial labeling containing sufficient technical information (i.e., at a minimum, nanomaterial composition, concentration, and median size) would better inform consumers and highly benefit researchers interested in understanding consumers' exposure and nanomaterial fate and transport in the environment.

Crowdsourcing

Since October 29, 2013, when the modified inventory (CPI 2.0) was released, 557 new user accounts have been requested. Of these, only approximately 10 users who were not directly or indirectly involved in the research team performing the CPI upgrade and maintenance suggested updates or edits to CPI entries. These edits have all been suggested by users from industry and academia.

Future work is needed to better educate users on their role as curators of CPI 2.0 and the importance of the data they contribute. Providing the supporting technical data required to verify the nature and quantity of nanomaterial components in CPI-listed products is a massive undertaking, and no single laboratory can accomplish it on its own or within a short amount of time. A long-term solution is to promote the importance of crowd-sourcing data collection and implementing standard data collection and reporting best practices that can help reliably populate the CPI with much needed supporting data. The new crowd-sourcing capability can also be used to provide high school-, undergraduate- and graduate-level educators with meaningful assignments that can help teach students about the prevalence of nanotechnology in everyday products and will contribute to the continued growth of this resource.

Nanotechnology expert survey

The survey was submitted to 147 people who have published research papers or reports in the applications of nanotechnology in consumer products and its potential impacts, participated in recent conferences in the field, or were notably involved in the field of nanotechnology and the consumer products industry. The survey had a 46% response rate (68 respondents), which is in the expected range for this type of survey [39]. The majority of respondents (59%) had six to ten years of experience working with nanotechnology and 38% of respondents had more than ten years of experience. Half (51%) of respondents work in academic institutions and 25% work in governmental agencies. Most respondents (88%) have previously used the CPI in their work, and all respondents believe they will or may use it again in the future.

Results convey a general belief or hope that the CPI will become more useful after the modifications reported in this publication. When asked the following open-ended questions: "How did you use the CPI in your work?" and "To what end do you think you might use the CPI in the future?", answers could be easily grouped into three main categories: (1) for raising awareness, teaching, or for urging the need for regulation, (2) to justify the need for research in research proposals or papers, and (3) to use the inventory data for research (Figure 9).

Half the respondents (51%) have used the CPI in the past to gather data for research (e.g., searching for consumer products of a certain nanomaterial composition to understand their potential applications or consumer exposure) while 74% believe they

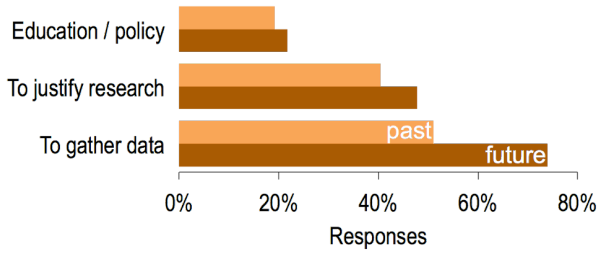


Figure 9: Nanotechnology survey answers on how respondents have used the CPI in the past and how they might use it in the future.

will use the CPI for that purpose in the future. The majority (79%) of survey respondents believed the modified CPI would present more products than its previous version, which indicates their belief in the growing prevalence of nanotechnology in consumer products.

Survey respondents suggested a number of new categories of information for the CPI 2.0, including nanomaterial type or composition, location of nanomaterial within the product, nanomaterial size, relevant scientific publications that describe the products in the inventory, a summary of known toxicity of the advertised nanomaterial, supply chain information, volume produced, and life cycle assessment information.

Most of these suggestions were included in the CPI 2.0 as the new categories described in this work. Others, such as known nanomaterials toxicity were not pursued since toxicity can vary greatly depending on particle size, coating, and exposure route (e.g., inhalation versus ingestion).

Piccinno et al. and Keller et al. provide global estimates for production and major applications of nanomaterials [20,21]. We recommend that future work associated with this inventory or others include information on the production volumes for each product, since this information is presently unavailable.

Additional results from this survey are available in Supporting Information File 1.

Conclusion

The modified version of the Wilson Center's nanotechnology consumer products inventory (CPI 2.0) was released in October 2013. We improved the searchability and utility of the inventory by including new descriptors for both the consumer products and the nanomaterial components of those products (e.g., size, concentration, and potential exposure routes). The updated CPI 2.0 now links listed products to published scientific information, where available, and includes a metric to assess the reliability of the data associated with each entry. Finally, the CPI 2.0 has enabled crowdsourcing capabilities, which allow registered users to upload new findings such as basic product composition information, human and environmental exposure data, and complete life cycle assessments. There are inherent limitations to this type of database, but recent improvements address the majority of issues raised in published literature and in a survey of nanotechnology experts.

Improvements to the CPI were motivated, in part, by the recognition that it represents and will continue to represent an important information resource for a broad range of stakeholders, especially consumers and the academic and regulatory communities. The CPI is a useful interactive database for educating consumers and legislators on the real-world applications of nanotechnology. Michaelson stated that the CPI transformed "the face of nanotechnology away from innovations in the realm of science fiction to the iconic images of everyday consumer products" [2]. The academic community can continue to make use of this inventory to help prioritize, for example, which types of products or nanomaterial components to evaluate in human exposure or toxicity studies, life cycle assessments, and nanomaterial release studies.

The CPI is useful for policy makers interested in regulating nanotechnology in consumer products by understanding their increasing numbers in the market, the main nanomaterial components that are chosen by manufacturers, and the likelihood for exposure. Beaudrie et al. [40] urge that there should be regulatory reforms to improve oversight of nanomaterials throughout their life cycle.

Finally, the current lack of global standardized methods and metrics for nanomaterial characterization and labeling in consumer products is an issue that, if addressed, can lead to greater understanding between the key stakeholders in nanotechnology, especially researchers, regulators, and industry. Further, as we recognize the growing importance of tools like the CPI for the needs of diverse stakeholder groups, steps should be taken to help ensure that those tools are fully developed and refined to meet those needs.

Experimental

Nanotechnology expert survey

To determine potentially useful improvements for the CPI, we developed a web-based survey to gather the informed opinions of nanotechnology experts – mostly in US-based academic institutions, governmental agencies, and research centers. Their answers guided the CPI modifications and provided an idea of the expectations related to the inventory. The survey questions are presented in the Supporting Information File 1.

New descriptors

To improve the utility and searchability of this database, seven product descriptors were created. Entries in the inventory were revised to go beyond a categorization of the consumer products and instead, to include more information on the nanomaterials themselves. We searched for this information mainly on the internet – on manufacturer's websites, retailer's websites, news sites and blogs, patents – and, when available, product labels.

Nanomaterial composition

The main composition of the nanomaterials used. This information, when available, was added to the database in the form of a check-box list, in which more than one nanomaterial composition can be selected for each consumer product.

Nanomaterial shape and size

Because there are many different ways in which manufacturers can measure and describe the shape and size of nanomaterials in consumer products (i.e., units of nanometers or micrometers, thickness of nanofilms, diameter or length of fibers or tubes, diameter or radius of nanoparticles, maximum, median, average, or minimum size), this descriptor was added as a text entry field in the database, which allows for any form of data entry but makes data analysis cumbersome.

Coatings

We created another text entry field in the CPI to include any available information on the coatings or stabilizing agent used along the nanomaterials in each product.

Nanomaterial location

To assist CPI users in understanding the potential for nanomaterial release and exposure scenarios from the use of these consumer products, we created a qualitative descriptor for the location of nanomaterials within each product. We adapted the categorization framework for nanomaterials from Hansen et al. [24] to determine the following nanomaterial locations within products:

- Bulk: Nanomaterials sold in powder form or in liquid suspensions
- Nanostructured bulk: Products or parts that contain nanostructured features in bulk (e.g., nanoscale computer processors)
- Nanostructured surface: Products or parts that contain nanostructured features on their surface (e.g., nanofilm-coated products)
- Surface-bound particles: Nanoparticles added to the surface of a solid product or part (e.g., a computer keyboard coated with silver nanoparticles for antimicrobial protection)
- Suspended in liquid: Nanomaterials suspended in a liquid product (e.g., disinfecting sprays, liquid supplements)
- Suspended in solid: Nanomaterials suspended in a solid matrix, usually plastic or metal (e.g., composites of carbon nanotubes in a plastic matrix to confer strength).

Nanomaterial function

We created a metric to describe the reason why nanotechnology was added to each consumer product or the function it performs within each product. We investigated a subset of 1244 products in the CPI for each product's intended use, the manufacturer claims, and, most importantly, the type or composition of nanomaterials used to infer potential nanomaterial functions (e.g., antimicrobial protection, hardness and strength, pigment).

Potential exposure pathways

Using methodology similar to that applied for the "nanomaterial functions" category, we investigated the CPI entries for possible exposure scenarios resulting from the expected normal use of each consumer product. Entries were only populated when a potential exposure risk was identified.

How much we know

In an effort to verify the data associated with each product listed on the CPI, we created a metric called "How much we know". Products were divided into five categories based on the information available to substantiate manufacturer claims that a particular product contains nanomaterial components (Table 2). Category 4, "Unsupported claim", is the default category for products added to the CPI based soley on a manufacturer's marketing claims. A product can rise in ranking according to the amount of information that is available to corroborate the manufacturer's claim that the product contains nanomaterials. If the manufacturer provides supporting information (e.g., a datasheet containing electron micrographs showing the nanomaterials or a particle size distribution), the product is placed in Category 3, "Manufacturer-supported claim". If a third-party further supports the information provided by the manufacturer, such as through a publication or technical report, then the product can be placed into Category 2, "Verified claim". If a product is backed by multiple science-based sources (e.g., a peer-reviewed scientific paper or patent documentation), it is then placed in Category 1, "Extensively verified claim". Category 5, "Not advertised by the manufacturer", is a special class for products that have been shown to contain nanomaterials but the manufacturer does not advertise this fact anywhere in product labeling or other informational materials. Category 5 has been added in recognition of the fact that not all nano-enabled products are marketed by manufacturers as such.

Table 2: "How much we know" categorization, based on the information available to substantiate manufacturer claims that a particular product contains nanomaterial components.

Category	Manufacturer claims to use nanotechnology	Manufacturer provides supporting information	Third-party information is available	Compelling information from multiple sources is available
1. Extensively verified claim	yes	yes	yes	yes
2. Verified claim	yes	yes	yes	
3. Manufacturer-supported claim	yes	yes		
4. Unsupported claim	yes			
5. Not advertised by manufacturer			yes	

Researchers say

In order to add available scientific information to the inventory, we created a text-entry database field named "Researchers say", which makes it possible to include an extract from a research paper (such as the abstract), author citation, and a link to the paper.

Crowdsourcing

We added a new crowdsourcing capability to the CPI website so that consumers, manufacturers, and the greater scientific community can contribute new information on nanomaterial composition of CPI products to the inventory. New contributors must request an account by completing a form with their contact information, and they must provide a reason why they would like to gain access to this crowdsourcing tool. Accounts are manually reviewed. Access is granted to all requesters who complete the form and have a legitimate purpose for contributing information. Once an account is created, users may sign in and suggest edits to any product (including the archiving of products no longer available or no longer advertising to contain nanomaterials) or suggest new products to the inventory. As a quality control measure, suggestions and new product forms contributed by registered users must be approved by a CPI curator before updates or revisions are posted to the inventory.

Supporting Information

Supporting Information File 1

A compilation of company and product numbers listed by country of origin. A list of all nanomaterial components included in the inventory. Nanotechnology expert survey questions. Additional nanotechnology expert survey results.
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-181-S1.pdf]

Acknowledgements

Funding for this work was provided by the Institute for Critical Technology and Applied Science (ICTAS) at Virginia Tech and the Virginia Tech Center for Sustainable Nanotechnology (VTSuN). We acknowledge the important help of J. Rousso, E. Bruning, S. Guldin, J. Wang, D. Yang, X. Zhou, L. Marr, the VTSuN graduate students in updating inventory entries, and the Laboratory for Interdisciplinary Statistical Analysis at Virginia Tech. We also acknowledge the Center for the Environmental Implications of Nanotechnology, funded under NSF Cooperative Agreement EF-0830093, for helping to inform our understanding of the broad world of manufactured nanomaterials. Co-author M. Hull acknowledges helpful discussions with A. Maynard of the Arizona State University Risk Innovation Lab that provided important motivation for this work.

References

1. The Project on Emerging Nanotechnologies. Consumer Products Inventory. http://www.nanotechproject.org/cpi (accessed March 25, 2015).
2. Michelson, E. S. *Rev. Policy Res.* **2013,** *30,* 464–487. doi:10.1111/ropr.12034
3. Currall, S. C.; King, E. B.; Lane, N.; Madera, J.; Turner, S. *Nat. Nanotechnol.* **2006,** *1,* 153–155. doi:10.1038/nnano.2006.155
4. Kahan, D. M.; Braman, D.; Slovic, P.; Gastil, J.; Cohen, G. *Nat. Nanotechnol.* **2009,** *4,* 87–90. doi:10.1038/nnano.2008.341
5. Maynard, A. D. *Nanotechnology: A Strategy for Addressing Risk;* Woodrow Wilson International Center for Scholars, 2006; p 45.
6. Maynard, A. D.; Aitken, R. J.; Butz, T.; Colvin, V.; Donaldson, K.; Oberdörster, G.; Philbert, M. A.; Ryan, J.; Seaton, A.; Stone, V.; Tinkle, S. S.; Tran, L.; Walker, N. J.; Warheit, D. B. *Nature* **2006,** *444,* 267–269. doi:10.1038/444267a
7. Berube, D. M.; Searson, E. M.; Morton, T. S.; Cummings, C. L. *Nanotechnol. Law Bus.* **2010,** *7,* 152–163.
8. Nano Products and Technologies. http://www.nanoproducts.de (accessed Feb 28, 2015).
9. National Institute of Advanced Industrial Science and Technology. A Nanotechnology-Claimed Consumer Products Inventory in Japan. http://www.aist-riss.jp/ (accessed Feb 28, 2015).
10. The European Consumer Organization. http://www.beuc.org/ (accessed Feb 28, 2015).
11. Danish Consumer Council. The Nanodatabase. http://nanodb.dk/ (accessed Feb 28, 2015).

12. de la Iglesia, D.; Harper, S.; Hoover, M. D.; Klaessig, F.; Lippell, P.; Maddux, B.; Morse, J.; Nel, A.; Rajan, K.; Reznik-Zellen, R.; Tuominen, M. T. *Nanoinformatics 2020 Roadmap;* National Nanomanufacturing Network, 2011. doi:10.4053/rp001-110413
13. Toxic Substances Control Act (TSCA), 15 U.S.C. §2601–2692, Washington, DC, 1976.
14. Environmental Protection Agency (EPA). Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA). http://www.epa.gov/agriculture/lfra.html (accessed March 2, 2015).
15. European Commission. Nanomaterials - Chemicals - Enterprise and Industry. http://ec.europa.eu/enterprise/sectors/chemicals/reach/nanomaterials/index_en.htm (accessed March 2, 2015).
16. European Commission. Nanomaterials under Biocidal Products Regulation - Echa. http://echa.europa.eu/regulations/nanomaterials-under-bpr (accessed March 2, 2015).
17. European Commission. Nanomaterials in Cosmetics. http://ec.europa.eu/growth/sectors/cosmetics/products/nanomaterials/index_en.htm (accessed March 2, 2015).
18. Hermann, A.; Diesner, M.-O.; Abel, J.; Hawthorne, C.; Greßmann, A. *Assessment of Impacts of a European Register of Products Containing Nanomaterials;* Federal Environment Agency (Umweltbundesamt): Dessau-Roßlau, Germany, 2014; p 142.
19. Maynard, A.; Michelson, E. S. *The Nanotechnology Consumer Products Inventory;* Woodrow Wilson International Center for Scholars, 2006.
20. Piccinno, F.; Gottschalk, F.; Seeger, S.; Nowack, B. *J. Nanopart. Res.* **2012,** *14,* 1109. doi:10.1007/s11051-012-1109-9
21. Keller, A. A.; McFerran, S.; Lazareva, A.; Suh, S. *J. Nanopart. Res.* **2013,** *15,* 1692. doi:10.1007/s11051-013-1692-4
22. Nowack, B.; Krug, H. F.; Height, M. *Environ. Sci. Technol.* **2011,** *45,* 1177–1183. doi:10.1021/es103316q
23. Saleh, N. B.; Aich, N.; Plazas-Tuttle, J.; Lead, J. R.; Lowry, G. V. *Environ. Sci.: Nano* **2015,** *2,* 11–18. doi:10.1039/C4EN00104D
24. Hansen, S. F.; Larsen, B. H.; Olsen, S. I.; Baun, A. *Nanotoxicology* **2007,** *1,* 243–250. doi:10.1080/17435390701727509
25. Lu, W.; Lieber, C. M. *Nat. Mater.* **2007,** *6,* 841–850. doi:10.1038/nmat2028
26. Hansen, S. F.; Michelson, E. S.; Kamper, A.; Borling, P.; Stuer-Lauridsen, F.; Baun, A. *Ecotoxicology* **2008,** *17,* 438–447. doi:10.1007/s10646-008-0210-4
27. Royce, S. G.; Mukherjee, D.; Cai, T.; Xu, S. S.; Alexander, J. A.; Mi, Z.; Calderon, L.; Mainelis, G.; Lee, K.; Lioy, P. J.; Tetley, T. D.; Chung, K. F.; Zhang, J.; Georgopoulos, P. G. *J. Nanopart. Res.* **2014,** *16,* 2724. doi:10.1007/s11051-014-2724-4
28. Nazarenko, Y.; Zhen, H. J.; Han, T.; Lioy, P. J.; Mainelis, G. *Environ. Health Perspect.* **2012,** *120,* 885–892. doi:10.1289/ehp.1104350
29. Nazarenko, Y.; Han, T. W.; Lioy, P. J.; Mainelis, G. *J. Exposure Sci. Environ. Epidemiol.* **2011,** *21,* 515–528. doi:10.1038/jes.2011.10
30. Quadros, M. E.; Marr, L. C. *Environ. Sci. Technol.* **2011,** *45,* 10713–10719. doi:10.1021/es202770m
31. Benn, T.; Cavanagh, B.; Hristovski, K.; Posner, J. D.; Westerhoff, P. *J. Environ. Qual.* **2010,** *39,* 1875–1882. doi:10.2134/jeq2009.0363
32. Quadros, M. E.; Pierson, R.; Tulve, N. S.; Willis, R.; Rogers, K.; Thomas, T. A.; Marr, L. C. *Environ. Sci. Technol.* **2013,** *47,* 8894–8901. doi:10.1021/es4015844
33. Tulve, N. S.; Stefaniak, A. B.; Vance, M. E.; Rogers, K.; Mwilu, S.; LeBouf, R. F.; Schwegler-Berry, D.; Willis, R.; Thomas, T. A.; Marr, L. C. *Int. J. Hyg. Environ. Health* **2015,** *218,* 345–357. doi:10.1016/j.ijheh.2015.02.002
34. Holden, P. A.; Klaessig, F.; Turco, R. F.; Priester, J. H.; Rico, C. M.; Avila-Arias, H.; Mortimer, M.; Pacpaco, K.; Gardea-Torresdey, J. L. *Environ. Sci. Technol.* **2014,** *48,* 10541–10551. doi:10.1021/es502440s
35. Savolainen, K.; Backman, U.; Brouwer, D.; Fadeel, B.; Fernandes, T.; Kuhlbusch, T.; Landsiedel, R.; Lynch, I.; Pylkkänen, L. Nanosafety in Europe 2015-2025: Towards Safe and Sustainable Nanomaterials and Nanotechnology Innovations. Finnish Institute of Occupational Health: Helsinki, Finnland, 2013; http://www.ttl.fi/en/publications/Electronic_publications/Nanosafety_in_europe_2015-2025/Documents/nanosafety_2015-2025.pdf.
36. Consumer Product Safety Commission Fiscal Year 2016 Performance Budget Request. http://www.cpsc.gov/Global/About-CPSC/Budget-and-Performance/FY2016BudgettoCongress.pdf (accessed March 2, 2015).
37. Hansen, S. F. *J. Nanopart. Res.* **2010,** *12,* 1959–1970. doi:10.1007/s11051-010-0006-3
38. Becker, S. *J. Nanopart. Res.* **2013,** *15,* 1426. doi:10.1007/s11051-013-1426-7
39. Baruch, Y.; Holtom, B. C. *Hum. Relat.* **2008,** *61,* 1139–1160. doi:10.1177/0018726708094863
40. Beaudrie, C. E. H.; Kandlikar, M.; Satterfield, T. *Environ. Sci. Technol.* **2013,** *47,* 5524–5534. doi:10.1021/es303591x

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

NanoE-Tox: New and in-depth database concerning ecotoxicity of nanomaterials

Katre Juganson*[1,2], Angela Ivask[1,3], Irina Blinova[1], Monika Mortimer[1,4] and Anne Kahru[1]

Full Research Paper

Open Access

Address:
[1]Laboratory of Environmental Toxicology, National Institute of Chemical Physics and Biophysics, Akadeemia tee 23, 12618 Tallinn, Estonia, [2]Department of Chemistry, Tallinn University of Technology, Akadeemia tee 15, 12618 Tallinn, Estonia, [3]Mawson Institute, University of South Australia, Mawson Lakes, 5095 South Australia, Australia and [4]Bren School of Environmental Science & Management, University of California Santa Barbara, Santa Barbara, California 93106-5131, United States

Email:
Katre Juganson* - katre.juganson@kbfi.ee

* Corresponding author

Keywords:
nanoparticles; physico-chemical properties; REACH; Thomson Reuters Web of Science; toxicity mechanisms

Beilstein J. Nanotechnol. **2015,** *6,* 1788–1804.
doi:10.3762/bjnano.6.183

Received: 31 March 2015
Accepted: 30 July 2015
Published: 25 August 2015

This article is part of the Thematic Series "Nanoinformatics for environmental health and biomedicine".

Guest Editor: R. Liu

Abstract

The increasing production and use of engineered nanomaterials (ENMs) inevitably results in their higher concentrations in the environment. This may lead to undesirable environmental effects and thus warrants risk assessment. The ecotoxicity testing of a wide variety of ENMs rapidly evolving in the market is costly but also ethically questionable when bioassays with vertebrates are conducted. Therefore, alternative methods, e.g., models for predicting toxicity mechanisms of ENMs based on their physico-chemical properties (e.g., quantitative (nano)structure-activity relationships, QSARs/QNARs), should be developed. While the development of such models relies on good-quality experimental toxicity data, most of the available data in the literature even for the same test species are highly variable. In order to map and analyse the state of the art of the existing nanoecotoxicological information suitable for QNARs, we created a database NanoE-Tox that is available as Supporting Information File 2. The database is based on existing literature on ecotoxicology of eight ENMs with different chemical composition: carbon nanotubes (CNTs), fullerenes, silver (Ag), titanium dioxide (TiO_2), zinc oxide (ZnO), cerium dioxide (CeO_2), copper oxide (CuO), and iron oxide (FeO_x; Fe_2O_3, Fe_3O_4). Altogether, NanoE-Tox database consolidates data from 224 articles and lists altogether 1,518 toxicity values (EC_{50}/LC_{50}/NOEC) with corresponding test conditions and physico-chemical parameters of the ENMs as well as reported toxicity mechanisms and uptake of ENMs in the organisms. 35% of the data in NanoE-Tox concerns ecotoxicity of Ag NPs, followed by TiO_2 (22%), CeO_2 (13%), and ZnO (10%). Most of the data originates from studies with crustaceans (26%), bacteria (17%), fish (13%), and algae (11%). Based on the median toxicity values of the most sensitive organism (data derived from three or more articles) the toxicity order was as follows: Ag > ZnO > CuO > CeO_2 > CNTs > TiO_2 > FeO_x. We believe NanoE-Tox database contains valuable information for ENM environmental hazard estimation and development of models for predicting toxic potential of ENMs.

Introduction

The production and use of engineered nanomaterials (ENMs) in consumer products is increasing rapidly [1]. As of March 20, 2015 there were more than 1,800 products listed in Consumer Products Inventory [2]. According to this inventory, the most abundant ENMs used in consumer products are silver (438 products), titanium (107), carbon (90), silica (81), zinc (38) and gold (24) with the main applications in antimicrobial protection (381 products), coatings (188) and health products (142). The number of published articles could serve as a good indicator of the potential future use of ENMs. A search performed on March 19, 2015 in Thomson Reuters Web of Science (WoS) with the keywords chosen based on Aitken et al. [3] and Bondarenko et al. [4] and listed in Table S1 (Supporting Information File 1) revealed that the majority of the papers concerned the applications of carbon nanotubes (36,609 papers, 40%), followed by Ag nanoparticles (NPs; 16,970, 19%), TiO_2 NPs (11,802, 13%), and iron oxide NPs (10,479, 11%) while the most common fields of application were sensors (28,027 papers, 31%), catalysis (10,435, 11%) and drug delivery (8,838, 10%) (Figure 1, Table S1, Supporting Information File 1). However, the exact production volumes of ENMs are not publicly available [4]. Piccinno et al. estimated based on a survey sent to companies producing and using ENMs that the most produced ENMs were TiO_2 (550–5,500 t/year), SiO_2 (55–55,000 t/year), AlO_x (55–5,500 t/year), ZnO (55–550 t/year), carbon nanotubes (CNT; 55–550 t/year), FeO_x (5.5–5,500 t/year), CeO_x and Ag (both 5.5–550 t/year), fullerenes and quantum dots (both 0.6-5.5 t/year) [5]. Warningly, the increasing production and use of ENMs leads inevitably to their higher concentrations in the environment. Thus, the risks caused by ENMs both to humans and the environment need to be assessed [6].

Risk assessment of all the ENMs in the market would require the sacrifice of enormous amounts of test organisms of diverse range [7]. Therefore, there is a need to refine, reduce or replace (3R's) animal testing and develop alternative risk evaluation methods [7,8]. Recently, the categorisation of ENMs based on their physico-chemical properties, exposure and use scenarios and biological effects was suggested as a strategy to facilitate regulatory decision making while minimising time-consuming and costly in vivo studies [9]. In addition to high-throughput screening tests, modelling can provide information for rapid assessment of the toxicity mechanisms of ENMs [10]. For instance, models based on dynamic energy budget (DEB) theory have been developed for predicting toxicity mechanisms of ENMs [11]. Also, quantitative (nano)structure-activity relationship (QSARs/QNARs) models have great potential for predicting the harmful effects of ENMs from their physical, chemical, and morphological properties that can be measured

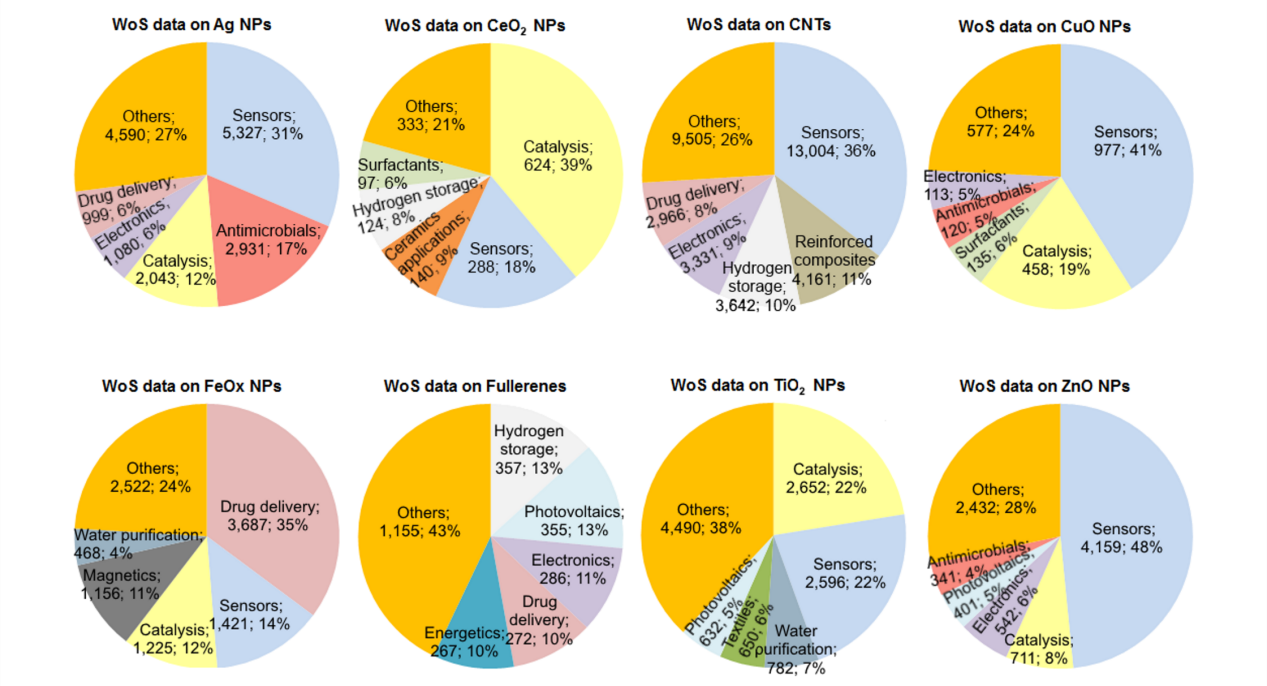


Figure 1: Proposed fields of application of engineered nanomaterials (ENMs) according to the publications in Thomson Reuters WoS. Keywords were selected from the review by Bondarenko et al. [4]. Numbers below each application category indicate the number and share of papers retrieved. The numerical data are presented in Table S1 (Supporting Information File 1). The bibliometric data search was performed in Thomson Reuters WoS on March 19, 2015.

experimentally or computed based on the ENMs structure [12]. Development of in silico methods relies on good-quality experimental data on ENM toxicity as the set of parameters which determine the toxic potential of each type of ENMs in specific test species/taxa is largely unknown [13].

In order to relate the toxic effects of ENMs to their physico-chemical properties and reveal the data gaps, the existing data have to be carefully collected and analysed. One increasingly popular approach in systematically collecting and organising available data on nanomaterials is creating databases. In 2012, Hristozov et al. emphasised that the available data on nanomaterials in environmental, health and safety databases and online chemical databases were very scarce [14]. Recently, a databases working group was established in the framework of European Union NanoSafety Cluster [15] which highlights the importance of development of in-depth databases on ENMs. In addition, nanotoxicity-related databases are developed and supported at national level in EU. For instance, in Germany an application-based nanomaterial database, which includes information on potential toxicological effects of ENMs, has been created in the DaNa project [16,17]. In Denmark, a database that focuses on potential risks of ENM containing products, "The Nanodatabase", has been developed [18]. The latter lists currently 1,425 products and introduces NanoRiskCat that evaluates ENMs risk according to potential exposure and hazard potential of these ENMs to humans and environment [19]. However, the risk estimations are derived from the available literature on the effects of nanomaterials but not on the actual risk assessment of the specific ENM-containing products. Therefore, the risk levels reported in the database do not account for concentrations or the physico-chemical properties of the specific ENMs used in the products. Independent online databases containing nanotoxicological information have also been created in other countries outside Europe. For instance, NanoToxdb: A database on Nanomaterial Toxicity [20] that is by description a comprehensive database containing information on nanomaterials toxicity to *Daphnia magna.* However, it contains altogether only 32 EC_{50} values for 10 different ENMs and contains no references for the toxicity data. Moreover, no information on physico-chemical properties of ENMs except primary particle size has been included in the database and regarding testing conditions, only the test duration is reported in a few cases. As a different approach, some databases, e.g., NHECD (Knowledge on the Health, Safety and Environmental Impact of Nanoparticles) [21] and Hazardous Substances Data Bank [22] comprise nanotoxicological papers.

In this communication we present a nanoecotoxicological database based on existing literature data on ecotoxicity of selected ENMs. In addition to quantitative toxicity data (e.g., EC_{50} values) information on physico-chemical properties of ENMs and testing conditions as well as on reported mechanisms and uptake of ENMs in the organisms was compiled. All the collected data were analysed to give an overview of ENM toxicity across different studied species. The following ENMs based on production volumes, application in consumer products and technological potential were included in the database: carbon nanotubes (CNTs), fullerenes, silver (Ag), titanium dioxide (TiO_2), zinc oxide (ZnO), cerium dioxide (CeO_2), copper oxide (CuO), and iron oxide (FeO_x; Fe_2O_3, Fe_3O_4). Furthermore, all these ENMs, except CuO, are listed by the Organisation for Economic Co-operation and Development (OECD) Working Party on Manufactured Nanomaterials as 'commercially relevant' representative manufactured nanomaterials to be investigated under the OECD sponsorship programme [23]. We believe the database presented in this paper contains valuable information for ENM environmental hazard estimation and development of models, including valid QSAR models, for predicting toxic potential of ENMs.

Methodology

The process of creating the nanoecotoxicological database can be roughly divided into three steps: selecting keywords for literature search, performing the literature search in Thomson Reuters WoS, collecting and classification of information from retrieved papers into a database. As the selection of keywords is critical in this type of data collection, all the keywords used in this study are listed in Table 1. To find different possible types

Table 1: Keywords used for bibliometric data search in Thomson Reuters WoS database.

ENM	Keywords
Ag	(nano* AND ecotoxic* AND silver) OR (nano* AND ecotoxic* AND Ag)
CeO_2	(nano* AND ecotoxic* AND cerium *oxide) OR (nano* AND ecotoxic* AND ceria) OR (nano* AND ecotoxic* AND CeO2)
CNT	(nano* AND ecotoxic* AND carbon nanotu*) OR (nano* AND ecotoxic* AND CNT) OR (nano* AND ecotoxic* AND *CNT)
CuO	(nano* AND ecotoxic* AND copper oxide) OR (nano* AND ecotoxic* AND CuO)
FeO_x	(nano* AND ecotoxic* AND iron *oxide) OR (nano* AND ecotoxic* AND Fe3O4) OR (nano* AND ecotoxic* AND Fe2O3)
fullerene	(nano* AND ecotoxic* AND fulleren*)
TiO_2	(nano* AND ecotoxic* AND titanium *oxide) OR (nano* AND ecotoxic* AND titania) OR (nano* AND ecotoxic* AND TiO2)
ZnO	(nano* AND ecotoxic* AND zinc oxide) OR (nano* AND ecotoxic* AND ZnO)

of 'nano' materials, i.e., nanoparticles, nanomaterials, nanotubes, a truncated search term "nano*" was selected. In order to give equal weight to all ecotoxicological test species, the restricting keyword "ecotoxic*" was used instead of organism-specific keywords. Thus, inevitably some of the ecotoxicological data on ENMs has been unintentionally excluded from the database because not all articles reporting studies on nanotoxicity to environmentally relevant organisms necessarily use terms "ecotoxic", "ecotoxicity" or "ecotoxicology". When performing the search, truncated names, molecular formulas and/or common abbreviations of the 8 NPs were used (Table 1).

Thomson Reuters WoS database – one of the largest international and multidisciplinary databases available, covering the most comprehensive list of journals published in English – was used for the bibliometric data search. Using WoS (all databases, all years) for the keyword searches enabled us to compare the data collected into NanoE-Tox with analyses performed in our previous reviews [4,8,24,25]. The search was performed on a regular basis from October 2012 to January 6, 2015. From each paper that was retrieved using the keywords specified in Table 1, maximum available information on physico-chemical properties of ENMs and the toxicity data were extracted and tabulated. It is important to note that in the earlier papers dating back 10 years from now, the NPs characterisation was often limited to their primary size. In more recent nanotoxicological articles, set of parameters required for characterisation of ENMs generally include chemical composition, purity, primary particle size, shape, surface area, coating, agglomeration and/or aggregation, hydrodynamic size in the aqueous test medium, surface charge, stability and solubility of ENMs. For the current NanoE-Tox database (Supporting Information File 2) we collected the following properties of the pristine NPs: chemical composition, origin (producer/in-house synthesised), shape, coating, primary size (diameter and length if applicable), impurities, surface area, and other reported observations. For the characterisation of ENMs in the test environment the following information was registered: test medium, hydrodynamic size of NPs in the test environment (including the method used for analysis), dissolution (if applicable), and surface charge (ζ-potential). Concerning the toxicity testing, we tabulated the following information: test organism, test medium, test duration, temperature, illumination and other reported conditions, toxicity endpoint/measure (e.g., EC_{50}, LC_{50}, NOEC), obtained toxicity value, and other reported observations. In addition, each paper was analysed to find information concerning (i) specific mechanism of toxicity of the studied ENM (Table S2, Supporting Information File 1) (ii) uptake in the organisms, and (iii) accumulation in cells, tissues and organs (Table S3, Supporting Information File 1). All the collected data were compiled into a Microsoft Excel spreadsheet which was used for creating a database on ecotoxicology of engineered nanomaterials, NanoE-Tox (Supporting Information File 2).

Results and Discussion

During the recent years, the number of peer-reviewed papers related to nanoecotoxicology has increased exponentially. According to Thomson Reuters WoS, 770 nanoecotoxicological peer-reviewed papers that corresponded to keywords "nano* AND ecotoxic*" were published between 2006 and March 2015. The rapidly increasing number of scientific publications on ecotoxicity of ENMs over the past decade, has inspired several review articles summarising the existing data in the field [4,8,13,24-31]. However, each review has focused on specific aspects and parameters of ENMs testing; therefore, it is difficult to get an overview of all the factors (and their values) that might influence the toxicity of ENMs. We have previously collected and analysed ecotoxicological data for seven different NPs (TiO_2, ZnO, CuO, Ag, SWCNTs, MWCNTs and C_{60} fullerenes) and seven organism groups representing different trophic levels (bacteria, algae, crustaceans, ciliates, fish, yeasts and nematodes). Altogether 77 toxicity values were analysed [24]. In our recent review [4], we summarised the recent research on toxicological and ecotoxicological findings for Ag, CuO and ZnO NPs including more than 300 toxicity values. In addition to ecotoxicological test species the toxic effects of studied NPs toward mammalian cells in vitro were reviewed [4]. The bibliographic search performed in the current study by using keywords listed in Table 1 resulted in nearly 500 individual papers. All the papers were thoroughly studied for ecotoxicity data. Unfortunately, many of the retrieved papers either did not concern the NP of interest or were review articles. In addition, the importance of including synonyms in keywords to increase the number of relevant articles in search results was apparent (Table 1). For example, the search using keywords "nano* AND ecotoxic* AND cerium *oxide" resulted in 30 papers, whereas "nano* AND ecotoxic* AND CeO2" resulted in 34 papers; remarkably, only 20 papers overlapped. The latter example was also true for other ENMs.

Analysis of the database: general overview of the sources and contents of the papers

The search in Thomson Reuters WoS using the time span of "all years" indicated that all the papers about ecotoxicity of ENMs have been published within the last ten years. Almost half of the papers retrieved from the initial bibliographic search, 224 of 500 articles from 66 journals, contained relevant nanotoxicological information and were included in NanoE-Tox database (Supporting Information File 2). From these studies 1,518 toxicity values were recorded with test conditions on toxicity testing and physico-chemical parameters of NPs linked to the toxicity data (further designated as 'database entry'). Out of 224

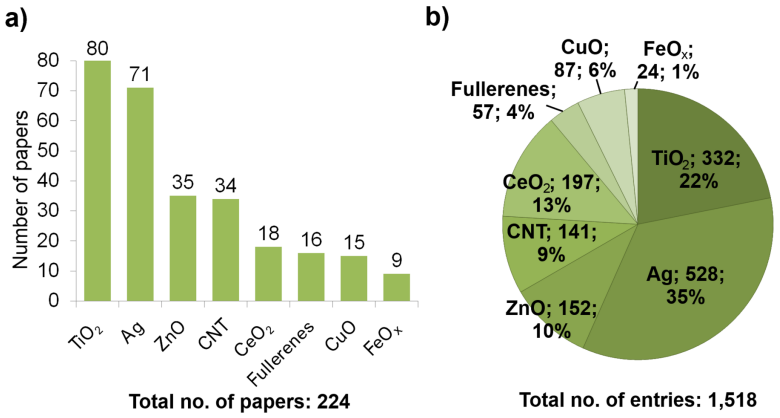


Figure 2: Information in the NanoE-Tox database for different types of engineered nanomaterials (ENMs): (a) number of scientific papers in the database and (b) number and share of entries for each of the tested nanoparticles (ENM; number of entries; share of entries). One entry equals one line that includes all the ENM parameters and toxicity test details. The database entries were selected based on bibliometric data search in Thomson Reuters WoS using the keywords as indicated in Table 1 as of January 6, 2015.

scientific papers that were selected for the database the largest number of papers concerned TiO_2 and Ag (80 and 71, respectively) followed by ZnO and CNTs (35 and 34 papers). For CeO_2, fullerenes and CuO, 15–18 papers were found and the lowest number of papers was retrieved for FeO_x (Figure 2a). From the 1,518 toxicity values (entries) in the database, the highest percentage (35%) concerned Ag followed by TiO_2 (22%), CeO_2 (13%), ZnO (10%), CNTs (9%), CuO (6%), fullerenes (4%) and FeO_x (1%) (Figure 2b).

Chronologically, the first nanoecotoxicological studies included in the database were published in 2006 and concerned TiO_2 NPs and CNTs (Figure 3). The first papers on ecotoxicity of fullerenes and ZnO NPs were published in 2007 followed by CeO_2, CuO and Ag NPs at 2008. While ecotoxicological effects of TiO_2 are still extensively studied, the interest in ecotoxicology of CNTs has slightly decreased. Notably, the most rapid increase rate appears to be in the number of published papers about nanosilver (Figure 3). The information on ecotoxicity of FeO_x particles started to emerge in 2009, i.e., later than for the other selected NPs (Figure 3). These findings are coherent with the literature survey by Kahru and Ivask [8] who showed that according to the citation pattern, the focus of the environment-related research shifted towards nanotoxicology by 2005 and the 'pioneering' NPs in environmental safety studies were CNTs, fullerenes, TiO_2, SiO_2 and ZnO. The analysis of the journals that contributed to the database revealed that more than half of the relevant papers originated from seven journals: Environmental Toxicology and Chemistry (29 papers), Environmental Science & Technology (25), Chemosphere (18), Environmental Pollution (12), Aquatic Toxicology (12), Science of the Total Environment (11), and Journal of Hazardous Materials (10 papers) (Table S4, Supporting Information File 1).

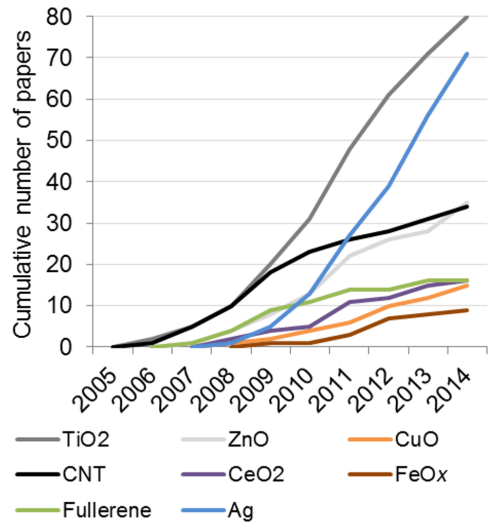


Figure 3: Evolution of nanoecotoxicological information about eight different nanomaterials according to the number of papers in NanoE-Tox database. The database entries were selected based on bibliometric data search in Thomson Reuters WoS using the keywords as indicated in Table 1 as of January 6, 2015.

Analysis of the database: physico-chemical characterisation of nanomaterials

The physico-chemical characteristics of ENMs included in the NanoE-tox database can be divided to intrinsic properties and properties that are specific to the test environment. The intrinsic characteristics are: name, CAS number, origin, shape, initial coating or functionalization, primary size, possible impurities, surface area and other observations, and the test environment-specific characteristics are: media, size, dissolution and zeta potential (Supporting Information File 2). Figure 4 illustrates the distribution of the data on ENM characteristics in NanoE-

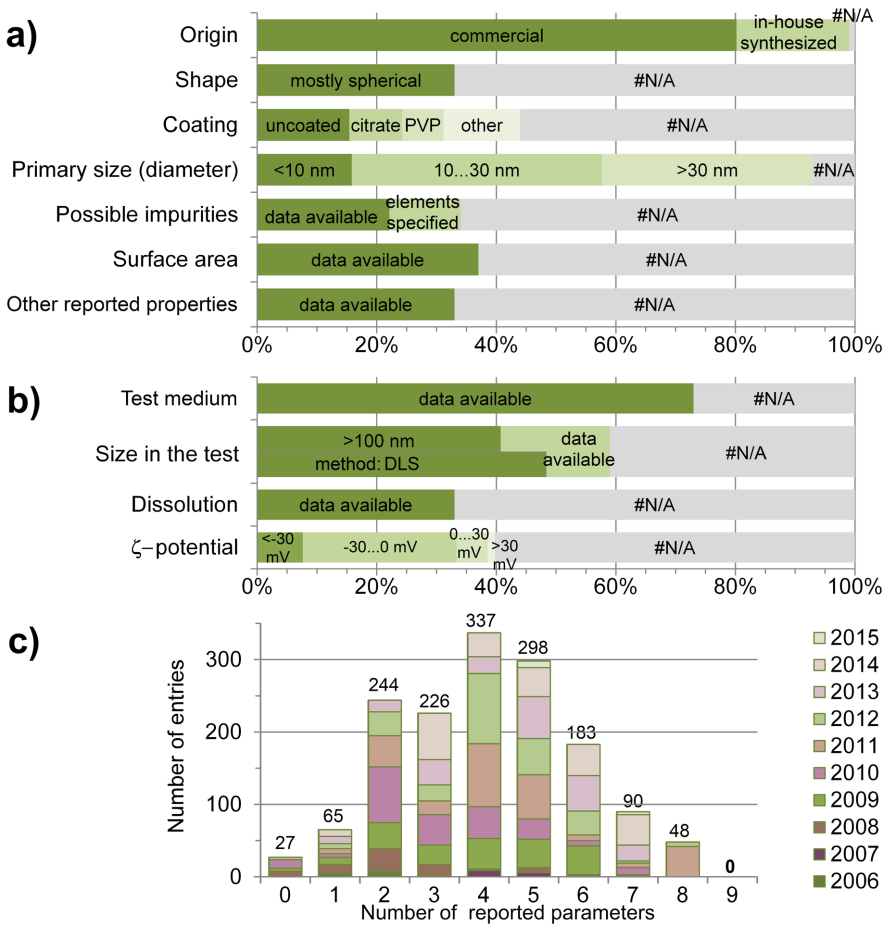


Figure 4: NanoE-Tox database: available data on characterisation of ENMs. Pristine (a) and environment-specific (b) properties as a percentage of all entries (1,518) in the database. Number of ENM parameters (shape, coating, primary size, impurities, surface area, other reported observations, size in the test, dissolution, ζ-potential) by number of entry and by publication year (c) #N/A - data not available. The database entries were selected based on bibliometric data search in Thomson Reuters WoS using the keywords as indicated in Table 1 as of January 6, 2015.

Tox database. Analysis of the papers revealed that in 99% of the entries the origin of the ENMs was known and 80% of the nanomaterials were obtained from commercial sources (Figure 4a). The most common source for all ENMs was Sigma Aldrich, 40% of all commercial particles were obtained from there. TiO_2 particles were mostly purchased from Evonik Industries (former Evonik-Degussa).

Many authors have emphasised that understanding the real risks of ENMs is a challenging task as there are several parameters that might have an influence on the biological effects of ENM [8,24,32-35]. Besides the chemical composition, the most important parameter determining the toxicity of NPs is their small size and size-dependent toxicity has been hypothesised in various papers [36,37]. Indeed, particle size has been considered as one of the most important physico-chemical parameter also in the papers collected in this study as this parameter was reported for 93% of the entries in the database. For all rod-shaped particles, also their length was reported. However, the results showed that most of the particles that were used in the 224 selected papers, were rather heterogeneous as in many cases the primary size was reported as a size range. According to Burello and Worth [38] ENMs with a diameter larger than 20–30 nm act often as bulk materials; thus, the "true nano-effects" are attributable to ENMs with smaller size. Indeed, in a recent paper on toxicity of different sizes of Ag NPs to bacteria, yeast, algae, crustaceans and mammalian cells in vitro Ivask et al. [39] showed that the toxicity of 20, 40, 60 and 80 nm monodisperse citrate-coated Ag NPs could fully be explained by released Ag ions whereas 10 nm Ag NPs proved more toxic than predicted. Analysis of the data in NanoE-Tox database revealed that the particles were smaller than 10 nm in 17% of the entries and in the size range of 10–30 nm in 45% of the entries (Figure 4a). Therefore, more than half of the studies have been performed using ENMs that should have size-dependent nanoeffects but as in most cases the NPs were polydis-

perse (i.e., had a broad size range) these effects were not often observed. Specific surface area that is closely related to the size of ENMs was reported in 37% of the entries (Figure 4a).

Another parameter that has been hypothesised to affect NP toxicity is morphology. For instance, some studies have shown that rod-shaped ENMs or triangular nanoplates could be more toxic than spherical ones [40-42]. However, the shape of ENMs was mentioned only in 33% of the entries and most of the experiments in the collected articles were performed with spherical particles (Figure 4a).

In addition to particle size and morphology, surface coating and/or functionalisation has been considered as an important parameter determining the biological effects of ENMs. For example, it has been discussed that coating on nanosilver plays an important role in Ag NPs toxicity [4,43,44]. However, information on initial coating or functionalisation of NPs was provided only in less than half of the entries. This is alarming because the surface chemistry of ENMs dictates their interactions with biological molecules and cells [45]. Altogether, 44% of the entries in the database contained information on NP coating: 29% of these were coated and 15% uncoated. ENMs were most often modified with citrate (31% of all coatings) and polyvinylpyrrolidone (PVP; 24% of all coatings) (Figure 4a). The high percentage of coated NPs in the database can be explained by the fact that nanosilver which constituted 35% of the database entries is frequently functionalised with different coatings, polyvinylpyrrolidone (PVP) and citrate being the most widely used.

A parameter closely related to NP surface properties is surface charge. It has been shown that positively charged ENMs tend to attach to the cellular surface that is negatively charged and these interactions may cause cell membrane damage [13,46]. In most studies ζ-potential is used as an indication of the surface charge of ENMs and NPs are considered to be stable in aqueous suspension if the ζ-potential is greater than ±30 mV [47]. In NanoE-Tox database, ζ-potential was reported in 40% of the entries. Most of the studies were performed with negatively charged ENMs (8% less than −30 mV, 25% −30…0 mV), 5% of the experiments were done with ENMs that had ζ-potential in the range of 0…+30 mV, and only 1% of the studies used stable positively charged ENMs (greater than +30 mV) (Figure 4b).

Another important parameter affecting toxicity of ENMs is the presence of impurities, for example presence of ‘seeding metals’ (catalysts) in CNTs that may count for observed toxic effects [48]. Purity of ENMs was reported in 34% of the entries; 65% of these cases mentioned purity as a percentage and 35% of the entries identified residual elements. Other reported observations, the most common parameters being crystal structure, density, and absorbance, were specified in 33% of the entries (Figure 4a).

Both in toxicological tests as well as in natural environments, the bioavailability and toxicity of ENMs depends on their fate in respective conditions [24,49]. In aquatic environment, ENMs tend to form agglomerates that might lead to their precipitation from the water phase; on the other hand, metal-based ENMs can release potentially toxic metal ions due to dissolution [50]. Cu^{2+}, Zn^{2+} and Ag^{+}, which can easily be released from respective ENMs are very toxic to a variety of aquatic organisms already at concentrations of milligrams and even micrograms per litre [4]. Analysis of the database entries (Figure 4b) showed that the most often reported ENM characteristic in the toxicity tests was hydrodynamic size (59% of all the entries) that usually (in 82% of the entries) was measured using dynamic light scattering (DLS) method. The data on hydrodynamic sizes indicated that ENMs tend to agglomerate in test conditions as 69% of the reported sizes were larger than 100 nm (in comparison, nearly all respective primary sizes were less than 100 nm). Dissolution of ENMs in toxicity tests was reported in 33% of all the entries. From all the studies using potentially soluble NPs (Ag, ZnO, CuO, CeO_2 and FeO_x) only half (51%) had measured the solubility of the particles.

As emphasised above, one of the goals of generating experimental nanotoxicological data is to apply them in model development that would allow for the comparison of physico-chemical properties of ENMs with their biological effects (QNAR models). It has been proposed that the QNAR models may even partially replace the expensive animal tests for evaluation of ENM related hazards [13]. Currently, there are a few QNAR modelling studies available for NPs [51]. However, these studies are based on relatively limited set of experimental data and therefore, applicable only for a small range of ENMs and organisms. Thus, in order to create a model with reasonable predictive power, several physico-chemical properties as well as data on a variety of NPs have to be included into the modelling to correlate the properties with toxic effects [25]. To evaluate whether the data in NanoE-Tox database might be suitable for (QNAR-)modelling, we analysed how many physico-chemical parameters of ENMs that could later be compared with the toxicological data were reported in each study. Nine physico-chemical parameters—shape, coating, primary size, impurities, surface area, other reported observations, size in the test, dissolution, surface charge (ζ-potential)—were analysed for the rate of being measured, i.e., how many of these were reported in one entry. In most of the studies, 2–6 of these parameters were reported (Figure 4c). Analysis of the data by year of publication revealed that despite of increasing number of nanotoxico-

logical articles being published each year, some of these still report only up to three parameters of ENM. On the other hand, there were no studies where all nine selected physico-chemical properties were explored, and in only 9% of the studies 7–8 parameters were reported. Hence, although the ecotoxicological data on NPs are rapidly increasing, there is still a shortage of accompanying information concerning physico-chemical properties of ENMs that may limit the use of nano(eco)toxicological data for QNARs.

Analysis of the database: ecotoxicological data

According to the European Union (EU) regulation on Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), the potential ecotoxicological effect of all chemical substances (including ENMs) that are produced in a volume of more than one tonne per year and sold in the EU must be evaluated. The amount of tests required depends on the production volume. If it exceeds 1 t/year, short-term tests with aquatic invertebrates (preferred species is *Daphnia*) and plants (algae is preferred) must be conducted. In case of the production volume over 10 t/year additional short-term tests with fish and studies of activated sludge respiration must be performed. Aforementioned aquatic studies must be performed also as long-term experiments for substances produced over 100 t/year; in addition, early life stage toxicity tests on fish, short-term toxicity tests on fish embryo and sac-fry stages and juvenile growth tests on fish must be carried out. With production over 100 t/year also terrestrial tests, short-term toxicity to invertebrates and plants and effects on soil microorganisms, must be performed. Finally, if the production volume for a certain substance exceeds 1,000 t/year, long-term terrestrial toxicity tests must be performed with invertebrates, plants, sediment organisms and birds in addition to all the previously mentioned aquatic and terrestrial studies [52].

To evaluate the compatibility of the toxicological data collected to NanoE-Tox database with the regulatory requirements, we collected the following data: type of test organism, test media, test duration and temperature, illumination conditions, test endpoint, toxicity measure and value. Also specific mechanisms of toxicity and accumulation of NPs in the cells, tissues or organs, and other observations were noted.

Organisms used for evaluation of biological effects of ENMs

Though the exact production volumes of ENMs are unknown, the estimated production of several ENMs exceeds the set 1 t/year limit [5]. Thus, according to legislation, several tests have to be conducted to bring these ENMs to the market. Organism-wise analysis of NanoE-Tox database revealed that information about effects of selected ENMs is available for 116 different test species (Table S5). Most of the experiments have been performed with water flea *Daphnia magna* (337 entries), followed by bacterium *Escherichia coli* (120 entries), unicellular alga *Pseudokirchneriella subcapitata* (107 entries), fish *Danio rerio* (66 entries), naturally luminescent bacterium *Vibrio fischeri* (44 entries), and nematode *Caenorhabditis elegans* (41 entries). In summary, by far the most often used test organisms were crustaceans constituting approximately one third (500/1,518) of all the tested species (Figure 5, Table S5, Supporting Information File 1). The abundance of toxicity data in crustaceans is likely derived from the mandatory reporting of these data according to REACH legislation as stated above. On the other hand, the amount of information about the effects of ENMs on algae – another mandatory test for REACH – is much more limited. With the keywords used in this study (Table 1), no information was found on algal toxicity of fullerenes and iron oxide and only one study evaluated the effect of CuO NPs on algae (Figure 5). The latter indicates that even if there are more publications on algal toxicity of ENMs, which were not retrieved in this study, the effects of ENMs on algae have been poorly studied. The same applies also to articles on effects of ENMs on fish. In NanoE-Tox database, there are no studies on the effect of CuO NPs on fish and only one study reported the effect of CeO_2 NPs and two studies showed the effect of fullerenes and FeO_x NPs to fish. Interestingly, toxicity tests with plants have been conducted with all 8 NPs. While relatively many studies have been performed with bacteria, the majority of them consider the effects towards potentially pathogenic bacterial strains, e.g., *Escherichia coli*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa*, *Staphylococcus aureus* (Table S5, Supporting Information File 1), which is likely driven by the important application area of some types of ENMs (TiO_2, ZnO, CuO, Ag) as antimicrobials [4,53]. About 16% of the entries in the database regard test organisms other than crustaceans, algae, fish, plants and bacteria. Those organisms included yeasts, protists, amphibians, bivalves, cnidarians, echinoderms, insects, nematodes, rotifers, snails and worms (Table S5, Supporting Information File 1). Hence, quite a wide range of test organisms has already been included in the evaluation of biological effects of ENMs. This certainly increases environmental relevance of these studies and the NanoE-Tox database.

Environmentally relevant test conditions

Recently, it has been highlighted that though most of the ENMs end up in the environment, relatively small amount of studies have been conducted in conditions relevant to the nature [54-56]. This was also reflected by the data collected into NanoE-Tox: 79% of the studies were performed in various artificial test media and only 15% in natural waters and 5% in soils, sludge or

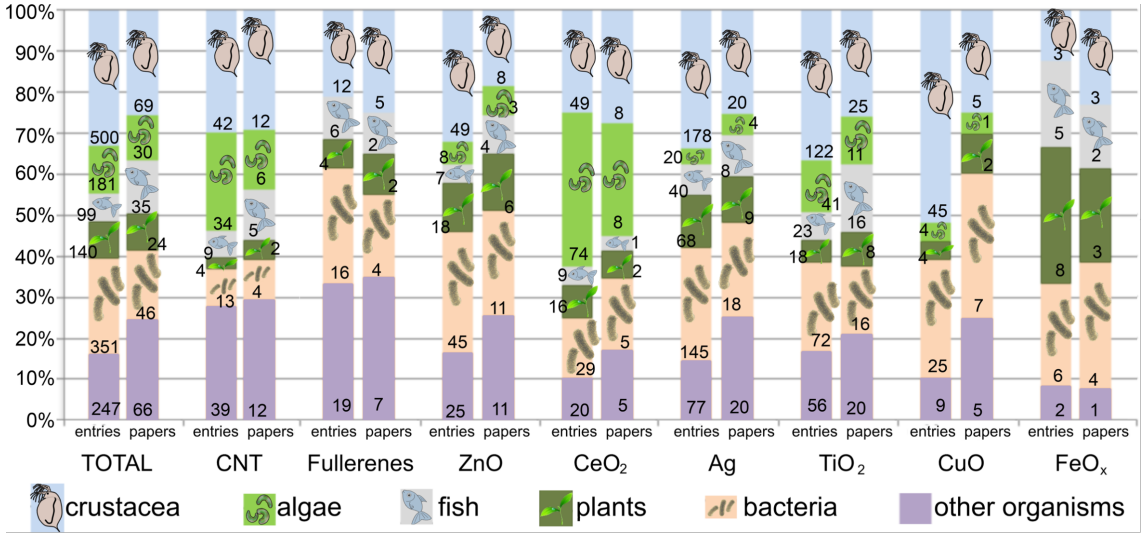


Figure 5: Types of test organisms used for evaluation of biological effects of selected ENMs in NanoE-Tox database. For each ENM, the left column represents the number of entries and right column represents the number of respective publications in the database. The database entries were selected based on bibliometric data search in Thomson Reuters WoS using the keywords as indicated in Table 1 as of January 6, 2015. The number of papers and entries for different ENMs is also presented in Figure 2.

sediments. Generally, the test conditions were relatively well reported in the majority of the analysed papers: the time of exposure (test duration) was reported in nearly all cases, while the test temperature was documented in more than 90% of the entries and information about illumination (illumination conditions/dark) was mentioned in 75% of the entries.

Toxicity endpoints used

The toxicity values for ENMs, irrespective of the endpoint, were based on nominal concentrations of ENMs. As expected, in most of the studies (77% of the entries) the toxicological endpoint was viability (e.g., mortality, immobilisation, growth inhibition, luminescence/fluorescence inhibition) while the effects on viability were classically expressed as half-effective (EC_{50}), half-inhibitory (IC_{50}), or half-lethal (LC_{50}) concentrations. 28% of the entries reported EC_{50} values, 10% LC_{50} values, 20% of the studies reported the concentration that did not exhibit any effect to the test organisms, i.e., NOEC (no observed effect concentration) values. However, some studies did not report any classical toxicity values because only one or two concentrations of NPs were tested by the authors; that did not allow for the establishment of a dose–response curve and, thus, calculations of E(L)C values. In addition, some papers considered the effect of ENMs on reproduction or studied possible malformations caused by ENMs that would be difficult to use for modelling purposes. As a result, the data that could be used as comparative inputs for models to evaluate the ecotoxicologial effects of ENMs is fairly limited in the database.

Analysis of the data consolidated into NanoE-Tox

Nano(eco)toxicological studies have usually two main aims: (i) the assessment of the toxic potential of ENMs, and (ii) the elucidation of the mechanism of toxic action [4,25]. In the following sections we will describe how NanoE-Tox database addresses these aims.

Toxicity of engineered nanomaterials

According to EU's regulation on classification, labelling and packaging of substances and mixtures (CLP) [57], chemical substances can be categorised as acutely or chronically toxic based on the results of standardised toxicity tests (reviewed by Crane et al. [58]) with fish (96 h), crustaceans (48 h) or algae (72 or 96 h). While by legislation acute toxicity has only one category ($E(L)C_{50}$ of the most sensitive organism ≤ 1 mg/L), chronic toxicity can be divided into four sub-categories ($E(L)C_{50}$ ≤1 mg/L; $E(L)C_{50}$ >1 to ≤10 mg/L; $E(L)C_{50}$ > 10 to ≤100 mg/L; $E(L)C_{50}$ > water solubility) that incorporate the degradation rate and bioconcentration factor of the chemical substance. Unfortunately, the latter two are not commonly determined in ecotoxicological studies; thus, in NanoE-Tox database bioconcentration factor has been reported only for FeO_x in fish larvae [59] and TiO_2 in coral tissue [60] and in crustaceans [61]. In order to give an overview of the ecotoxicity data collected for NanoE-Tox database (Figure 6), the hazard classification of ENMs was adjusted accordingly: acutely very toxic and potentially chronically very toxic ($E(L)C_{50}$ ≤ 1 mg/L), potentially chronically toxic ($E(L)C_{50}$ >1 to ≤10 mg/L), potentially chronically harmful ($E(L)C_{50}$ > 10 to

≤100 mg/L) and not classified (E(L)C_{50} > 100). Figure 6 depicts median values of all EC_{50}, LC_{50} and IC_{50} values with minimum and maximum values from NanoE-Tox database. Median EC_{50} values were calculated because these are the most precise estimates derived from the concentration–effect curve [62] and also, median EC_{50} values are often used in the QSAR analysis [63]. Analysis of the sources of the median values showed that most of the data in one data point originated from one (red frame, 19 points) or two (orange frame, 10 points) papers, only 18 median values were derived from 3 or more papers (green frame).

Based on the median toxicity values of the most sensitive organisms (i.e., theoretically representing the weakest link in the ecosystem), the toxicity of selected ENMs decreased in the order Ag > ZnO > FeO_x > CuO > fullerenes > CNTs > TiO_2 > CeO_2. However, when toxicity values that were derived from three or more papers were considered, the order slightly changed: Ag > ZnO > CuO > CeO_2 > CNTs > TiO_2 > FeO_x. The median values reported here are in general agreement with those published previously [4,24,26] (Table 2). However, such evaluation where the median values are derived across all different test conditions and test species is not in accordance with the current legislation. In order to be coherent with legislation, we next analysed the toxicity data obtained in standard tests with fish (96 h), daphnids (48 h) and algae (72 or 96 h) (Figure 7), i.e., the mandatory tests required under CLP [57] for classification of substances, and applied the same hazard ranking criteria as was used in Figure 6. This analysis showed that the most toxic ENM was Ag that could be classified as "acutely very toxic" and "potentially chronically very toxic". ZnO and FeO_x were also ranked as "acutely very toxic" and "potentially chronically very toxic" although less toxic than Ag. It is worth mentioning that the classification of FeO_x NPs was based on only one study (entry in the database), warranting further research of FeO_x NPs for more accurate ecotoxicity evaluation. According to median E(L)C_{50} values from the standard toxicity tests, CuO and CeO_2 NPs, CNTs and fullerenes

Table 2: Comparison of the median E(L,I)C_{50} values for different species in NanoE-Tox database and previous reviews [4,24,26].

ENM	E(L,I)C_{50} range in NanoE-Tox	E(L,I)C_{50} range in other reviews
Ag	0.01–245 mg/L	0.01–38 mg/L [4] 0.04–39 mg/L [24]
CeO_2	8.5–46.6 mg/L	0.1–100 mg/L [26]
CNTs	4.5–338 mg/L	1.0–500 mg/L [24]
CuO	0.32–569 mg/L	2.1–100 mg/L [4] 0.71–127 mg/L [24]
FeO_x	0.23–240 mg/L	#N/A[a]
fullerenes	1.5–11 mg/L	0.25–100 mg/L [24]
TiO_2	6.8–589 mg/L	39–11987 mg/L [24]
ZnO	0.05–3376 mg/L	0.08–121 mg/L [4] 0.055–97.4 mg/L [24]

[a] #N/A: not applicable.

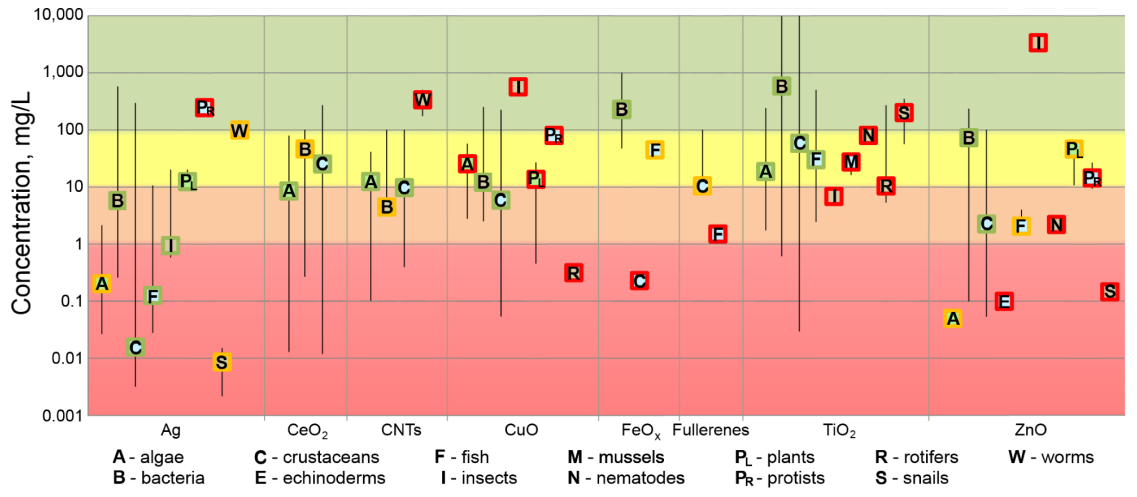


Figure 6: NanoE-Tox database: toxicity of selected nanoparticles to different organisms (data filtered by keyword ecotoxic*). Median E(L,I)C_{50} values ± minimum and maximum values. Colours of the frames surrounding the letters indicate the number of papers from which the respective data originates: red = 1 paper, orange = 2 papers, green ≥ 3 papers. The whiskers indicate the variability of the data. Note the logarithmic scale of y-axis. The E(L,I)C_{50} values used to derive the median values are from 113 papers and usually based on nominal concentration of the compound [44,55,56,61,64-172]. The toxicity ranking is indicated with the coloured background: E(L)C_{50} ≤ 1 mg/L – acutely very toxic, potentially chronically very toxic (red); E(L)C_{50} >1 to ≤10 mg/L – potentially chronically toxic (orange); E(L)C_{50} > 10 to ≤100 mg/L – potentially chronically harmful (yellow); E(L)C_{50} > 100 – not classified (green). The database entries were selected based on bibliometric data search in Thomson Reuters WoS™ using the keywords as indicated in Table 1 as of January 6, 2015.

fell into the category of "potentially chronically toxic" and TiO_2 NPs were ranked as "potentially chronically harmful".

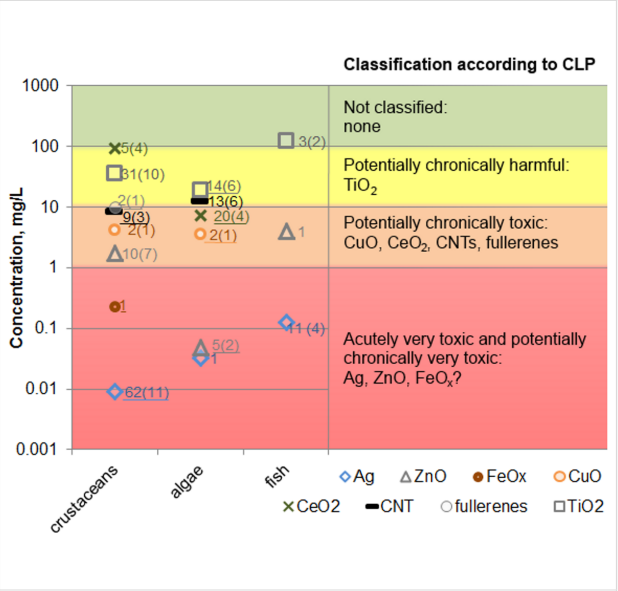


Figure 7: Classification of selected nanoparticles according to European Union CLP legislation based on their toxicity to fish (96 h), daphnids (48 h) and algae (72 or 96 h). Toxicity values were extracted from Figure 6. Classification of NPs is based on the most sensitive organism as described in CLP [57]. The number next to the symbol indicates the number of E(L,I)C_{50} values used to derive the median value and the number in the parenthesis indicates the number of papers from which the respective data originates. Underlined numbers indicate the datapoints (lowest E(L,I)C_{50} value for this ENM) used for classification. Note the logarithmic scale of the y-axis.

Mechanism of toxic action

While after a decade-long research the exact mechanisms of toxic action of ENMs are still debated, the main proposed mechanisms can be outlined as follows: (i) physical interactions of ENMs with cells or cellular components, (ii) production of reactive oxygen species and resulting induction of oxidative stress, and (iii) toxic effect of released ions from metal/metal oxide ENMs [13,25,28]. Analyses of the information in NanoE-Tox database (Table S2, Supporting Information File 1) revealed that the most often reported potential mechanism of toxic action for ZnO [128-132,173], Ag [44,64-73,174-177], and CuO [55,64,73,126-129,173] NPs was the release of metal ions. On the other hand, some studies have also proposed that the toxicity of these ENMs might be at least partially caused by the NPs themselves [73-84,178-181]. However, most of the studies reporting NP-specific effects of Ag, CuO and ZnO used insoluble particles and tested them in higher concentrations compared to the ones commonly reported as toxic. Thus, it can be concluded, in accordance with some previous studies [4,25], that in most cases the observed toxicity of these three ENMs was triggered by toxic metal ions. Other modes of toxic action reported for Ag NPs included destabilisation of cell membranes/mechanical membrane damage [89,175,182,183], oxidative stress [71,73,89,175,176,184,185], DNA damage/genotoxicity [102,186,187], and binding to sulfhydryl groups [100]. Similar effects were also demonstrated in case of ZnO NPs [84-86,188-190]. The mechanism of toxic action of insoluble ENMs like CeO_2 [109,110], CNTs [116,133,191] and TiO_2 [153-156,192] was usually reported as particle-driven mechanical membrane damage. NanoE-Tox database contains only one study suggesting the mechanism of toxicity of fullerenes (oxidative stress) [193] and there are no data about possible mechanism of action of FeO_x NPs.

Additionally, the information collected to the NanoE-Tox database indicated that ENMs were readily ingested by different organisms [55,72,77,119-123,192,194-202] and tended to accumulate in them [55,59,60,69-71,84,122-126,159,176-179,187,189,192,201-214] or on their surface [79,117-119,126,136-140,196,215-218] (Table S3, Supporting Information File 1). Similar findings have been reported in previous studies [24-26,29].

Conclusion

NanoE-Tox database that is available as Supporting Information File 2 of this paper is the first online-available database that contains in-depth nanoecotoxicological information on eight ENMs accompanied by considerable amount of information on ENM physico-chemical properties, testing conditions and, to some extent, also on mechanisms of toxic action. Hence, NanoE-Tox enables the comparison of toxicity of ENMs across different test species and, in addition, could provide valuable input for computational toxicity modeling (e.g., QSARs) and risk assessment.

The analysis of the database entries resulted in coherent data with previously published studies: the most toxic of the selected ENMs were Ag NPs followed by ZnO and CuO NPs and the toxicity of these ENMs was largely triggered by their solubility. Additionally, systematic collection of the data revealed several gaps in the current knowledge about ENM ecotoxicity: (i) in most cases the physico-chemical properties of the investigated NPs were described insufficiently, (ii) relatively few experiments have been performed with algae and fish, and (iii) ecotoxicity tests with standard test organisms were often performed with modified protocols (i.e., duration of the test was either shorter or longer than required by the OECD or ISO standards). Although the NanoE-Tox database is limited to a selected range of articles entered in the Thomson Reuters WoS database by January 6, 2015 and retrieved by using specific keywords, it provides a good overview of the existing ecotoxicological information about Ag, CeO_2, CuO, FeO_x, TiO_2 and ZnO NPs, carbon nanotubes and fullerenes.

Supporting Information

Supporting Information File 1

Supplementary tables.
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-183-S1.pdf]

Supporting Information File 2

NanoE-Tox database.
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-183-S2.xls]

Acknowledgements

We thank Liina Kanarbik for collecting ecotoxicity data about FeO_x ENMs and Meelika Koitjärv for assistance in collecting ecotoxicity data about TiO_2 ENMs. This work was supported by Estonian Research Council's "Environmental Conservation and Environmental Technology R&D Program" project "TERIK-VANT" and by Estonian Ministry of Education and Research (target-financed theme IUT23-5 and PUT748).

References

1. Kessler, R. *Environ. Health Perspect.* **2011,** *119,* a120–a125. doi:10.1289/ehp.119-a120
2. The Project on Emerging Nanotechnologies. http://www.nanotechproject.org/cpi/products/ (accessed Feb 26, 2015).
3. Aitken, R. J.; Chaudhry, M. Q.; Boxall, A. B. A.; Hull, M. *Occup. Med.* **2006,** *56,* 300–306. doi:10.1093/occmed/kql051
4. Bondarenko, O.; Juganson, K.; Ivask, A.; Kasemets, K.; Mortimer, M.; Kahru, A. *Arch. Toxicol.* **2013,** *87,* 1181–1200. doi:10.1007/s00204-013-1079-4
5. Piccinno, F.; Gottschalk, F.; Seeger, S.; Nowack, B. *J. Nanopart. Res.* **2012,** *14,* 1109. doi:10.1007/s11051-012-1109-9
6. Sun, T. Y.; Gottschalk, F.; Hungerbuehler, K.; Nowack, B. *Environ. Pollut.* **2014,** *185,* 69–76. doi:10.1016/j.envpol.2013.10.004
7. Hartung, T.; Sabbioni, E. *Wiley Interdiscip. Rev.: Nanomed. Nanobiotechnol.* **2011,** *3,* 545–573. doi:10.1002/wnan.153
8. Kahru, A.; Ivask, A. *Acc. Chem. Res.* **2013,** *46,* 823–833. doi:10.1021/ar3000212
9. Godwin, H.; Nameth, C.; Avery, D.; Bergeson, L. L.; Bernard, D.; Beryt, E.; Boyes, W.; Brown, S.; Clippinger, A. J.; Cohen, Y.; Doa, M.; Hendren, C. O.; Holden, P.; Houck, K.; Kane, A. B.; Klaessig, F.; Kodas, T.; Landsiedel, R.; Lynch, I.; Malloy, T.; Miller, M. B.; Muller, J.; Oberdorster, G.; Petersen, E. J.; Pleus, R. C.; Sayre, P.; Stone, V.; Sullivan, K. M.; Tentschert, J.; Wallis, P.; Nel, A. E. *ACS Nano* **2015,** *9,* 3409–3417. doi:10.1021/acsnano.5b00941
10. Holden, P. A.; Nisbet, R. M.; Lenihan, H. S.; Miller, R. J.; Cherr, G. N.; Schimel, J. P.; Gardea-Torresdey, J. L.; Univ, C. *Acc. Chem. Res.* **2013,** *46,* 813–822. doi:10.1021/ar300069t
11. Klanjscek, T.; Nisbet, R. M.; Priester, J. H.; Holden, P. A. *Ecotoxicology* **2013,** *22,* 319–330. doi:10.1007/s10646-012-1028-7
12. Fourches, D.; Pu, D.; Tassa, C.; Weissleder, R.; Shaw, S. Y.; Mumper, R. J.; Tropsha, A. *ACS Nano* **2010,** *4,* 5703–5712. doi:10.1021/nn1013484
13. Djurisic, A. B.; Leung, Y. H.; Ng, A. M. C.; Xu, X. Y.; Lee, P. K. H.; Degger, N.; Wu, R. S. S. *Small* **2015,** *11,* 26–44. doi:10.1002/smll.201303947
14. Hristozov, D. R.; Gottardo, S.; Critto, A.; Marcomini, A. *Nanotoxicology* **2012,** *6,* 880–898. doi:10.3109/17435390.2011.626534
15. EU NanoSafety Cluster - Database WG. http://www.nanosafetycluster.eu/working-groups/4-database-wg.html (accessed March 9, 2015).
16. Kühnel, D.; Marquardt, C.; Nau, K.; Krug, H. F.; Mathes, B.; Steinbach, C. *Environ. Sci. Eur.* **2014,** *26,* 21. doi:10.1186/s12302-014-0021-6
17. DaNa 2.0 Information about nanomaterials and their safety assessment. http://nanopartikel.info/en/nanoinfo/knowledge-base (accessed March 23, 2015).
18. The Nanodatabase. http://nanodb.dk/en/search-database (accessed March 3, 2015).
19. Hansen, S. F.; Jensen, K. A.; Baun, A. *J. Nanopart. Res.* **2013,** *16,* 2195. doi:10.1007/s11051-013-2195-z
20. NanoToxdb:A Database on Nanomaterial Toxicity. http://iitrindia.org/envis/Default.aspx (accessed March 2, 2015).
21. NHECD. Knowledge on the Health, Safety and Environmental Impact of Nanoparticles. http://nhecd-fp7.eu (accessed Feb 10, 2015).
22. HSDB a TOXNET database. http://www.toxnet.nlm.nih.gov/newtoxnet/hsdb.htm (accessed March 4, 2015).
23. Organization for Economic Co-operation and Development. *List of Manufactured Nanomaterials and List of Endpoints for Phase One of the Sponsorship Programme for the Testing of Manufactured Nanomaterials: Revision. ENV/JM/MONO;* Paris, 2010.
24. Kahru, A.; Dubourguier, H.-C. *Toxicology* **2010,** *269,* 105–119. doi:10.1016/j.tox.2009.08.016
25. Ivask, A.; Juganson, K.; Bondarenko, O.; Mortimer, M.; Aruoja, V.; Kasemets, K.; Blinova, I.; Heinlaan, M.; Slaveykova, V.; Kahru, A. *Nanotoxicology* **2014,** *8* (Suppl. 1), 57–71. doi:10.3109/17435390.2013.855831
26. Collin, B.; Auffan, M.; Johnson, A. C.; Kaur, I.; Keller, A. A.; Lazareva, A.; Lead, J. R.; Ma, X.; Merrifield, R. C.; Svendsen, C.; White, J. C.; Unrine, J. M. *Environ. Sci.: Nano* **2014,** *1,* 533–548. doi:10.1039/C4EN00149D
27. Holden, P. A.; Klaessig, F.; Turco, R. F.; Priester, J. H.; Rico, C. M.; Avila-Arias, H.; Mortimer, M.; Pacpaco, K.; Gardea-Torresdey, J. L. *Environ. Sci. Technol.* **2014,** *48,* 10541–10551. doi:10.1021/es502440s
28. Ma, H.; Williams, P. L.; Diamond, S. A. *Environ. Pollut.* **2013,** *172,* 76–85. doi:10.1016/j.envpol.2012.08.011
29. Menard, A.; Drobne, D.; Jemec, A. *Environ. Pollut.* **2011,** *159,* 677–684. doi:10.1016/j.envpol.2010.11.027
30. Jackson, P.; Jacobsen, N. R.; Baun, A.; Birkedal, R.; Kühnel, D.; Jensen, K. A.; Vogel, U.; Wallin, H. *Chem. Cent. J.* **2013,** *7,* 154. doi:10.1186/1752-153X-7-154
31. Sánchez, A.; Recillas, S.; Font, X.; Casals, E.; González, E.; Puntes, V. *TrAC, Trends Anal. Chem.* **2011,** *30,* 507–516. doi:10.1016/j.trac.2010.11.011
32. Nel, A. E.; Mädler, L.; Velegol, D.; Xia, T.; Hoek, E. M. V.; Somasundaran, P.; Klaessig, F.; Castranova, V.; Thompson, M. *Nat. Mater.* **2009,** *8,* 543–557. doi:10.1038/nmat2442

33. Albanese, A.; Tang, P. S.; Chan, W. C. W. *Annu. Rev. Biomed. Eng.* **2012,** *14,* 1–16. doi:10.1146/annurev-bioeng-071811-150124
34. Suresh, A. K.; Pelletier, D. A.; Doktycz, M. J. *Nanoscale* **2013,** *5,* 463–474. doi:10.1039/C2NR32447D
35. Handy, R. D.; von der Kammer, F.; Lead, J. R.; Hassellöv, M.; Owen, R.; Crane, M. *Ecotoxicology* **2008,** *17,* 287–314. doi:10.1007/s10646-008-0199-8
36. Sharifi, S.; Behzadi, S.; Laurent, S.; Forrest, M. L.; Stroeve, P.; Mahmoudi, M. *Chem. Soc. Rev.* **2012,** *41,* 2323–2343. doi:10.1039/C1CS15188F
37. Luyts, K.; Napierska, D.; Nemery, B.; Hoet, P. H. M. *Environ. Sci.: Processes Impacts* **2013,** *15,* 23–38. doi:10.1039/C2EM30237C
38. Burello, E.; Worth, A. P. *Wiley Interdiscip. Rev.: Nanomed. Nanobiotechnol.* **2011,** *3,* 298–306. doi:10.1002/wnan.137
39. Ivask, A.; Kurvet, I.; Kasemets, K.; Blinova, I.; Aruoja, V.; Suppi, S.; Vija, H.; Käkinen, A.; Titma, T.; Heinlaan, M.; Visnapuu, M.; Koller, D.; Kisand, V.; Kahru, A. *PLoS One* **2014,** *9,* e102108. doi:10.1371/journal.pone.0102108
40. Pal, S.; Tak, Y. K.; Song, J. M. *Appl. Environ. Microbiol.* **2007,** *73,* 1712–1720. doi:10.1128/AEM.02218-06
41. Sadeghi, B.; Garmaroudi, F. S.; Hashemi, M.; Nezhad, H. R.; Nasrollahi, A.; Ardalan, S.; Ardalan, S. *Adv. Powder Technol.* **2012,** *23,* 22–26. doi:10.1016/j.apt.2010.11.011
42. George, S.; Lin, S.; Jo, Z.; Thomas, C. R.; Li, L.; Mecklenburg, M.; Meng, H.; Wang, X.; Zhang, H.; Xia, T.; Hohman, J. N.; Lin, S.; Zink, J. I.; Weiss, P. S.; Nel, A. E. *ACS Nano* **2012,** *6,* 3745–3759. doi:10.1021/nn204671v
43. El Badawy, A. M.; Silva, R. G.; Morris, B.; Scheckel, K. G.; Suidan, M. T.; Tolaymat, T. M. *Environ. Sci. Technol.* **2011,** *45,* 283–287. doi:10.1021/es1034188
44. Ivask, A.; ElBadawy, A.; Kaweeteerawat, C.; Boren, D.; Fischer, H.; Ji, Z.; Chang, C. H.; Liu, R.; Tolaymat, T.; Telesca, D.; Zink, J. I.; Cohen, Y.; Holden, P. A.; Godwin, H. A. *ACS Nano* **2014,** *8,* 374–386. doi:10.1021/nn4044047
45. Walczyk, D.; Bombelli, F. B.; Monopoli, M. P.; Lynch, I.; Dawson, K. A. *J. Am. Chem. Soc.* **2010,** *132,* 5761–5768. doi:10.1021/ja910675v
46. Fröhlich, E. *Int. J. Nanomed.* **2012,** *7,* 5577–5591. doi:10.2147/IJN.S36111
47. Hartig, S. M.; Greene, R. R.; Dikov, M. M.; Prokop, A.; Davidson, J. M. *Pharm. Res.* **2007,** *24,* 2353–2369. doi:10.1007/s11095-007-9459-1
48. Ge, C.; Li, Y.; Yin, J.-J.; Liu, Y.; Wang, L.; Zhao, Y.; Chen, C. *NPG Asia Mater.* **2012,** *4,* e32. doi:10.1038/am.2012.60
49. Nowack, B.; Bucheli, T. D. *Environ. Pollut.* **2007,** *150,* 5–22. doi:10.1016/j.envpol.2007.06.006
50. Batley, G. E.; Kirby, J. K.; McLaughlin, M. J. *Acc. Chem. Res.* **2013,** *46,* 854–862. doi:10.1021/ar2003368
51. Winkler, D. A.; Mombelli, E.; Pietroiusti, A.; Tran, L.; Worth, A.; Fadeel, B.; McCall, M. J. *Toxicology* **2013,** *313,* 15–23. doi:10.1016/j.tox.2012.11.005
52. Regulation (EC) No 1907/2006 of the European Parliament and of the Council concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH). *Official Journal of the European Union, L396;* 2006; pp 1–849.
53. Suppi, S.; Kasemets, K.; Ivask, A.; Künnis-Beres, K.; Sihtmäe, M.; Kurvet, I.; Aruoja, V.; Kahru, A. *J. Hazard. Mater.* **2014,** *286,* 75–84. doi:10.1016/j.jhazmat.2014.12.027
54. Bour, A.; Mouchet, F.; Silvestre, J.; Gauthier, L.; Pinelli, E. *J. Hazard. Mater.* **2015,** *283,* 764–777. doi:10.1016/j.jhazmat.2014.10.021
55. Blinova, I.; Ivask, A.; Heinlaan, M.; Mortimer, M.; Kahru, A. *Environ. Pollut.* **2010,** *158,* 41–47. doi:10.1016/j.envpol.2009.08.017
56. Blinova, I.; Niskanen, J.; Kajankari, P.; Kanarbik, L.; Käkinen, A.; Tenhu, H.; Penttinen, O.-P.; Kahru, A. *Environ. Sci. Pollut. Res.* **2013,** *20,* 3456–3463. doi:10.1007/s11356-012-1290-5
57. Regulation (EC) No 1272/2008 of the European Parliament and of the Council on classification, labelling and packaging of substances and mixtures. *Official Journal of the European Union, L353;* 2008; pp 1–1355.
58. Crane, M.; Handy, R. D.; Garrod, J.; Owen, R. *Ecotoxicology* **2008,** *17,* 421–437. doi:10.1007/s10646-008-0215-z
59. Chen, P.-J.; Tan, S.-W.; Wu, W.-L. *Environ. Sci. Technol.* **2012,** *46,* 8431–8439. doi:10.1021/es3006783
60. Jovanović, B.; Guzmán, H. M. *Environ. Toxicol. Chem.* **2014,** *33,* 1346–1353. doi:10.1002/etc.2560
61. Dalai, S.; Pakrashi, S.; Chandrasekaran, N.; Mukherjee, A. *PLoS One* **2013,** *8,* e62970. doi:10.1371/journal.pone.0062970
62. Isnard, P.; Flammarion, P.; Roman, G.; Babut, M.; Bastien, P.; Bintein, S.; Esserméant, L.; Ferard, J. F.; Gallotti-Schmitt, S.; Saouter, E.; Saroli, M.; Thiébaud, H.; Tomassone, R.; Vindimian, E. *Chemosphere* **2001,** *45,* 659–669. doi:10.1016/S0045-6535(00)00600-7
63. Cronin, M. T. D.; Jaworska, J. S.; Walker, J. D.; Comber, M. H. I.; Watts, C. D.; Worth, A. P. *Environ. Health Perspect.* **2003,** *111,* 1391–1401.
64. Jo, H. J.; Choi, J. W.; Lee, S. H.; Hong, S. W. *J. Hazard. Mater.* **2012,** *227–228,* 301–308. doi:10.1016/j.jhazmat.2012.05.066
65. Radniecki, T. S.; Stankus, D. P.; Neigh, A.; Nason, J. A.; Semprini, L. *Chemosphere* **2011,** *85,* 43–49. doi:10.1016/j.chemosphere.2011.06.039
66. Engelke, M.; Köser, J.; Hackmann, S.; Zhang, H.; Mädler, L.; Filser, J. *Environ. Toxicol. Chem.* **2014,** *33,* 1142–1147. doi:10.1002/etc.2542
67. Lee, Y.-J.; Kim, J.; Oh, J.; Bae, S.; Lee, S.; Hong, I. S.; Kim, S.-H. *Environ. Toxicol. Chem.* **2012,** *31,* 155–159. doi:10.1002/etc.717
68. Allen, H. J.; Impellitteri, C. A.; Macke, D. A.; Heckman, J. L.; Poynton, H. C.; Lazorchak, J. M.; Govindaswamy, S.; Roose, D. L.; Nadagouda, M. N. *Environ. Toxicol. Chem.* **2010,** *29,* 2742–2750. doi:10.1002/etc.329
69. Schlich, K.; Klawonn, T.; Terytze, K.; Hund-Rinke, K. *Environ. Toxicol. Chem.* **2013,** *32,* 181–188. doi:10.1002/etc.2030
70. Kaveh, R.; Li, Y.-S.; Ranjbar, S.; Tehrani, R.; Brueck, C. L.; Van Aken, B. *Environ. Sci. Technol.* **2013,** *47,* 10637–10644. doi:10.1021/es402209w
71. Oukarroum, A.; Barhoumi, L.; Pirastru, L.; Dewez, D. *Environ. Toxicol. Chem.* **2013,** *32,* 902–907. doi:10.1002/etc.2131
72. Juganson, K.; Mortimer, M.; Ivask, A.; Kasemets, K.; Kahru, A. *Environ. Sci.: Processes Impacts* **2013,** *15,* 244–250. doi:10.1039/C2EM30731F
73. Ivask, A.; Bondarenko, O.; Jepihhina, N.; Kahru, A. *Anal. Bioanal. Chem.* **2010,** *398,* 701–716. doi:10.1007/s00216-010-3962-7
74. Ribeiro, F.; Gallego-Urrea, J. A.; Jurkschat, K.; Crossley, A.; Hassellov, M.; Taylor, C.; Soares, A. M. V. M.; Loureiro, S. *Sci. Total Environ.* **2014,** *466,* 232–241. doi:10.1016/j.scitotenv.2013.06.101
75. Macken, A.; Byrne, H. J.; Thomas, K. V. *Ecotoxicol. Environ. Saf.* **2012,** *86,* 101–110. doi:10.1016/j.ecoenv.2012.08.025

76. Kumar, D.; Kumari, J.; Pakrashi, S.; Dalai, S.; Raichur, A. M.; Sastry, T. P.; Mandal, A. B.; Chandrasekaran, N.; Mukherjee, A. *Ecotoxicol. Environ. Saf.* **2014,** *108,* 152–160. doi:10.1016/j.ecoenv.2014.05.033
77. Scanlan, L. D.; Reed, R. B.; Loguinov, A. V.; Antczak, P.; Tagmount, A.; Aloni, S.; Nowinski, D. T.; Luong, P.; Tran, C.; Karunaratne, N.; Don, P.; Lin, X. X.; Falciani, F.; Higgins, C. P.; Ranville, J. F.; Vulpe, C. D.; Gilbert, B. *ACS Nano* **2013,** *7,* 10681–10694. doi:10.1021/nn4034103
78. Harmon, A. R.; Kennedy, A. J.; Poda, A. R.; Bednar, A. J.; Chappell, M. A.; Steevens, J. A. *Environ. Toxicol. Chem.* **2014,** *33,* 1783–1791. doi:10.1002/etc.2616
79. Laban, G.; Nies, L. F.; Turco, R. F.; Bickham, J. W.; Sepúlveda, M. S. *Ecotoxicology* **2010,** *19,* 185–195. doi:10.1007/s10646-009-0404-4
80. Osborne, O. J.; Johnston, B. D.; Moger, J.; Balousha, M.; Lead, J. R.; Kudoh, T.; Tyler, C. R. *Nanotoxicology* **2013,** *7,* 1315–1324. doi:10.3109/17435390.2012.737484
81. Manusadžianas, L.; Caillet, C.; Fachetti, L.; Gylyte, B.; Grigutytė, R.; Jurkoniene, S.; Karitonas, R.; Sadauskas, K.; Thomas, F.; Vitkus, R.; Férard, J.-F. *Environ. Toxicol. Chem.* **2012,** *31,* 108–114. doi:10.1002/etc.715
82. Mortimer, M.; Kasemets, K.; Vodovnik, M.; Marinšek-Logar, R.; Kahru, A. *Environ. Sci. Technol.* **2011,** *45,* 6617–6624. doi:10.1021/es201524q
83. Santo, N.; Fascio, U.; Torres, F.; Guazzoni, N.; Tremolada, P.; Bettinetti, R.; Mantecca, P.; Bacchetta, R. *Water Res.* **2014,** *53,* 339–350. doi:10.1016/j.watres.2014.01.036
84. Yu, L.-p.; Fang, T.; Xiong, D.-w.; Zhu, W.-t.; Sima, X.-f. *J. Environ. Monit.* **2011,** *13,* 1975–1982. doi:10.1039/c1em10197h
85. Lee, W.-M.; An, Y.-J. *Chemosphere* **2013,** *91,* 536–544. doi:10.1016/j.chemosphere.2012.12.033
86. Fahmy, S. R.; Abdel-Ghaffar, F.; Bakry, F. A.; Sayed, D. A. *Arch. Environ. Contam. Toxicol.* **2014,** *67,* 192–202. doi:10.1007/s00244-014-0020-z
87. Zhu, X.; Zhu, L.; Duan, Z.; Qi, R.; Li, Y.; Lang, Y. *J. Environ. Sci. Health, Part A: Toxic/Hazard. Subst. Environ. Eng.* **2008,** *43,* 278–284. doi:10.1080/10934520701792779
88. Waalewijn-Kool, P. L.; Ortiz, M. D.; van Gestel, C. A. M. *Ecotoxicology* **2012,** *21,* 1797–1804. doi:10.1007/s10646-012-0914-3
89. Gou, N.; Gu, A. Z. *Environ. Sci. Technol.* **2011,** *45,* 5410–5417. doi:10.1021/es200455p
90. Barrena, R.; Casals, E.; Colón, J.; Font, X.; Sánchez, A.; Puntes, V. *Chemosphere* **2009,** *75,* 850–857. doi:10.1016/j.chemosphere.2009.01.078
91. Silva, T.; Pokhrel, L. R.; Dubey, B.; Tolaymat, T. M.; Maier, K. J.; Liu, X. *Sci. Total Environ.* **2014,** *468,* 968–976. doi:10.1016/j.scitotenv.2013.09.006
92. Mallevre, F.; Fernandes, T. F.; Aspray, T. J. *Environ. Pollut.* **2014,** *195,* 218–225. doi:10.1016/j.envpol.2014.09.002
93. Gaiser, B. K.; Biswas, A.; Rosenkranz, P.; Jepson, M. A.; Lead, J. R.; Stone, V.; Tyler, C. R.; Fernandes, T. F. *J. Environ. Monit.* **2011,** *13,* 1227–1235. doi:10.1039/c1em10060b
94. Kim, J.; Kim, S.; Lee, S. *Nanotoxicology* **2011,** *5,* 208–214. doi:10.3109/17435390.2010.508137
95. Rani, P. U.; Rajasekharreddy, P. *Colloids Surf., A* **2011,** *389,* 188–194. doi:10.1016/j.colsurfa.2011.08.028
96. Völker, C.; Boedicker, C.; Daubenthaler, J.; Oetken, M.; Oehlmann, J. *PLoS One* **2013,** *8,* e75026. doi:10.1371/journal.pone.0075026
97. Römer, I.; Gavin, A. J.; White, T. A.; Merrifield, R. C.; Chipman, J. K.; Viant, M. R.; Lead, J. R. *Toxicol. Lett.* **2013,** *223,* 103–108. doi:10.1016/j.toxlet.2013.08.026
98. Li, T.; Albee, B.; Alemayehu, M.; Diaz, R.; Ingham, L.; Kamal, S.; Rodriguez, M.; Bishnoi, S. W. *Anal. Bioanal. Chem.* **2010,** *398,* 689–700. doi:10.1007/s00216-010-3915-1
99. Lapied, E.; Moudilou, E.; Exbrayat, J.-M.; Oughton, D. H.; Joner, E. J. *Nanomedicine* **2010,** *5,* 975–984. doi:10.2217/nnm.10.58
100. Farkas, J.; Christian, P.; Urrea, J. A. G.; Roos, N.; Hassellöv, M.; Tollefsen, K. E.; Thomas, K. V. *Aquat. Toxicol.* **2010,** *96,* 44–52. doi:10.1016/j.aquatox.2009.09.016
101. Kim, J.; Park, Y.; Lee, S.; Seo, J.; Kwon, D.; Park, J.; Yoon, T.-H.; Choi, K. J. *Korean J. Environ. Health Sci.* **2013,** *39,* 369–375. doi:10.5668/JEHS.2013.39.4.369
102. Nair, P. M. G.; Park, S. Y.; Lee, S.-W.; Choi, J. *Aquat. Toxicol.* **2011,** *101,* 31–37. doi:10.1016/j.aquatox.2010.08.013
103. Panacek, A.; Prucek, R.; Safarova, D.; Dittrich, M.; Richtrova, J.; Benickova, K.; Zboril, R.; Kvitek, L. *Environ. Sci. Technol.* **2011,** *45,* 4974–4979. doi:10.1021/es104216b
104. Gnanadesigan, M.; Anand, M.; Ravikumar, S.; Maruthupandy, M.; Vijayakumar, V.; Selvam, S.; Dhineshkumar, M.; Kumaraguru, A. K. *Asian Pac. J. Trop. Med.* **2011,** *4,* 799–803. doi:10.1016/S1995-7645(11)60197-1
105. El-Temsah, Y. S.; Joner, E. J. *Environ. Toxicol.* **2012,** *27,* 42–49. doi:10.1002/tox.20610
106. Ravindran, A.; Prathna, T. C.; Verma, V. K.; Chandrasekaran, N.; Mukherjee, A. *Toxicol. Environ. Chem.* **2012,** *94,* 91–98. doi:10.1080/02772248.2011.617034
107. Völker, C.; Gräf, T.; Schneider, I.; Oetken, M.; Oehlmann, J. *Environ. Sci. Pollut. Res.* **2014,** *21,* 10661–10670. doi:10.1007/s11356-014-3067-5
108. Bernot, R. J.; Brandenburg, M. *Hydrobiologia* **2013,** *714,* 25–34. doi:10.1007/s10750-013-1509-6
109. Rodea-Palomares, I.; Boltes, K.; Fernández-Piñas, F.; Leganés, F.; García-Calvo, E.; Santiago, J.; Rosal, R. *Toxicol. Sci.* **2011,** *119,* 135–145. doi:10.1093/toxsci/kfq311
110. Manier, N.; Bado-Nilles, A.; Delalain, P.; Aguerre-Chariol, O.; Pandard, P. *Environ. Pollut.* **2013,** *180,* 63–70. doi:10.1016/j.envpol.2013.04.040
111. Velzeboer, I.; Hendriks, A. J.; Ragas, A. M. J.; Van de Meent, D. *Environ. Toxicol. Chem.* **2008,** *27,* 1942–1947. doi:10.1897/07-509.1
112. Booth, A.; Størsetha, T.; Altin, D.; Fornara, A.; Ahniyaz, A.; Jungnickel, H.; Laux, P.; Luch, A.; Sørensen, L. *Sci. Total Environ.* **2015,** *505,* 596–605. doi:10.1016/j.scitotenv.2014.10.010
113. Van Hoecke, K.; Quik, J. T. K.; Mankiewicz-Boczek, J.; de Schamphelaere, K. A. C.; Elsaesser, A.; van der Meeren, P.; Barnes, C.; McKerr, G.; Howard, C. V.; van de Meent, D.; Rydzyński, K.; Dawson, K. A.; Salvati, A.; Lesniak, A.; Lynch, I.; Silversmit, G.; de Samber, B.; Vincze, L.; Janssen, C. R. *Environ. Sci. Technol.* **2009,** *43,* 4537–4546. doi:10.1021/es9002444
114. Tomilina, I. I.; Gremyachikh, V. A.; Myl'nikov, A. P.; Komov, V. T. *Inland Water Biol.* **2011,** *4,* 475–483. doi:10.1134/S1995082911040201
115. García, A.; Espinosa, R.; Delgado, L.; Casals, E.; González, E.; Puntes, V.; Barata, C.; Font, X.; Sánchez, A. *Desalination* **2011,** *269,* 136–141. doi:10.1016/j.desal.2010.10.052
116. Long, Z.; Ji, J.; Yang, K.; Lin, D.; Wu, F. *Environ. Sci. Technol.* **2012,** *46,* 8458–8466. doi:10.1021/es301802g

117. Asghari, S.; Johari, S. A.; Lee, J. H.; Kim, Y. S.; Jeon, Y. B.; Choi, H. J.; Moon, M. C.; Yu, I. J. *J. Nanobiotechnol.* **2012,** *10,* 14. doi:10.1186/1477-3155-10-14
118. Artells, E.; Issartel, J.; Auffan, M.; Borschneck, D.; Thill, A.; Tella, M.; Brousset, L.; Rose, J.; Bottero, J.-Y.; Thiéry, A. *PLoS One* **2013,** *8,* e71260. doi:10.1371/journal.pone.0071260
119. Martinez, D. S. T.; Faria, A. F.; Berni, E.; Souza Filho, A. G.; Almeida, G.; Caloto-Oliveira, A.; Grossman, M. J.; Durrant, L. R.; Umbuzeiro, G. A.; Alves, O. L. *Process Biochem.* **2014,** *49,* 1162–1168. doi:10.1016/j.procbio.2014.04.006
120. Kowk, K. W. H.; Leung, K. M. Y.; Flahaut, E.; Cheng, J.; Cheng, S. H. *Nanomedicine* **2010,** *5,* 951–961. doi:10.2217/nnm.10.59
121. Kennedy, A. J.; Gunter, J. C.; Chappell, M. A.; Goss, J. D.; Hull, M. S.; Kirgan, R. A.; Steevens, J. A. *Environ. Toxicol. Chem.* **2009,** *28,* 1930–1938. doi:10.1897/09-024.1
122. Zhu, X.; Zhu, L.; Chen, Y.; Tian, S. *J. Nanopart. Res.* **2009,** *11,* 67–75. doi:10.1007/s11051-008-9426-8
123. Petersen, E. J.; Akkanen, J.; Kukkonen, J. V. K.; Weber, W. J., Jr. *Environ. Sci. Technol.* **2009,** *43,* 2969–2975. doi:10.1021/es8029363
124. Resano, M.; Lapeña, A. C.; Belarra, M. A. *Anal. Methods* **2013,** *5,* 1130–1139. doi:10.1039/c2ay26456k
125. Petersen, E. J.; Pinto, R. A.; Mai, D. J.; Landrum, P. F.; Weber, W. J., Jr. *Environ. Sci. Technol.* **2011,** *45,* 1133–1138. doi:10.1021/es1030239
126. Pradhan, A.; Seena, S.; Pascoal, C.; Cássio, F. *Chemosphere* **2012,** *89,* 1142–1150. doi:10.1016/j.chemosphere.2012.06.001
127. Heinlaan, M.; Ivask, A.; Blinova, I.; Dubourguier, H.-C.; Kahru, A. *Chemosphere* **2008,** *71,* 1308–1316. doi:10.1016/j.chemosphere.2007.11.047
128. Luna-delRisco, M.; Orupõld, K.; Dubourguier, H.-C. *J. Hazard. Mater.* **2011,** *189,* 603–608. doi:10.1016/j.jhazmat.2011.02.085
129. Ko, K.-S.; Kong, I. C. *Appl. Microbiol. Biotechnol.* **2014,** *98,* 3295–3303. doi:10.1007/s00253-013-5404-x
130. Wiench, K.; Wohlleben, W.; Hisgen, V.; Radke, K.; Salinas, E.; Zok, S.; Landsiedel, R. *Chemosphere* **2009,** *76,* 1356–1365. doi:10.1016/j.chemosphere.2009.06.025
131. Li, M.; Zhu, L.; Lin, D. *Environ. Sci. Technol.* **2011,** *45,* 1977–1983. doi:10.1021/es102624t
132. Fairbairn, E. A.; Keller, A. A.; Mädler, L.; Zhou, D.; Pokhrel, S.; Cherr, G. N. *J. Hazard. Mater.* **2011,** *192,* 1565–1571. doi:10.1016/j.jhazmat.2011.06.080
133. Liu, S.; Wei, L.; Hao, L.; Fang, N.; Chang, M. W.; Xu, R.; Yang, Y.; Chen, Y. *ACS Nano* **2009,** *3,* 3891–3902. doi:10.1021/nn901252r
134. Wei, L.; Thakkar, M.; Chen, Y.; Ntim, S. A.; Mitra, S.; Zhang, X. *Aquat. Toxicol.* **2010,** *100,* 194–201. doi:10.1016/j.aquatox.2010.07.001
135. Pereira, M. M.; Mouton, L.; Yépremian, C.; Couté, A.; Lo, J.; Marconcini, J. M.; Ladeira, L. O.; Raposo, N. R. B.; Brandão, H. M.; Brayner, R. *J. Nanobiotechnol.* **2014,** *12,* 15. doi:10.1186/1477-3155-12-15
136. Roberts, A. P.; Mount, A. S.; Seda, B.; Souther, J.; Qiao, R.; Lin, S.; Ke, P. C.; Rao, A. M.; Klaine, S. J. *Environ. Sci. Technol.* **2007,** *41,* 3025–3029. doi:10.1021/es062572a
137. Schwab, F.; Bucheli, T. D.; Lukhele, L. P.; Magrez, A.; Nowack, B.; Sigg, L.; Knauer, K. *Environ. Sci. Technol.* **2011,** *45,* 6136–6144. doi:10.1021/es200506b
138. Planchon, M.; Ferrari, R.; Guyot, F.; Gélabert, A.; Menguy, N.; Chanéac, C.; Thill, A.; Benedetti, M. F.; Spalla, O. *Colloids Surf., B* **2013,** *102,* 158–164. doi:10.1016/j.colsurfb.2012.08.034
139. Lin, D.; Xing, B. *Environ. Sci. Technol.* **2008,** *42,* 5580–5585. doi:10.1021/es800422x
140. Dabrunz, A.; Duester, L.; Prasse, C.; Seitz, F.; Rosenfeldt, R.; Schilde, C.; Schaumann, G. E.; Schulz, R. *PLoS One* **2011,** *6,* e20112. doi:10.1371/journal.pone.0020112
141. Edgington, A. J.; Roberts, A. P.; Taylor, L. M.; Alloy, M. M.; Reppert, J.; Rao, A. M.; Mao, J.; Klaine, S. J. *Environ. Toxicol. Chem.* **2010,** *29,* 2511–2518. doi:10.1002/etc.309
142. Blaise, C.; Gagné, F.; Férard, J. F.; Eullaffroy, P. *Environ. Toxicol.* **2008,** *23,* 591–598. doi:10.1002/tox.20402
143. Scott-Fordsmand, J. J.; Krogh, P. H.; Schaefer, M.; Johansen, A. *Ecotoxicol. Environ. Saf.* **2008,** *71,* 616–619. doi:10.1016/j.ecoenv.2008.04.011
144. Sovova, T.; Koci, V.; Kochankova, L. Ecotoxicity of nano and bulk forms of metal oxides. In *Nanocon 2009, Conference Proceedings,* 2009; pp 341–347.
145. Goix, S.; Lévêque, T.; Xiong, T.-T.; Schreck, E.; Baeza-Squiban, A.; Geret, F.; Uzu, G.; Austruy, A.; Dumat, C. *Environ. Res.* **2014,** *133,* 185–194. doi:10.1016/j.envres.2014.05.015
146. Zhu, X.; Tian, S.; Cai, Z. *PLoS One* **2012,** *7,* e46286. doi:10.1371/journal.pone.0046286
147. Zhu, X.; Zhu, L.; Li, Y.; Duan, Z.; Chen, W.; Alvarez, P. J. J. *Environ. Toxicol. Chem.* **2007,** *26,* 976–979. doi:10.1897/06-583.1
148. Hund-Rinke, K.; Simon, M. *Environ. Sci. Pollut. Res.* **2006,** *13,* 225–232. doi:10.1065/espr2006.06.311
149. Hartmann, N. B.; von der Kammer, F.; Hofmann, T.; Baalousha, M.; Ottofuelling, S.; Baun, A. *Toxicology* **2010,** *269,* 190–197. doi:10.1016/j.tox.2009.08.008
150. Sadiq, I. M.; Dalai, S.; Chandrasekaran, N.; Mukherjee, A. *Ecotoxicol. Environ. Saf.* **2011,** *74,* 1180–1187. doi:10.1016/j.ecoenv.2011.03.006
151. Dalai, S.; Pakrashi, S.; Nirmala, M. J.; Chaudhri, A.; Chandrasekaran, N.; Mandal, A. B.; Mukherjee, A. *Aquat. Toxicol.* **2013,** *138,* 1–11. doi:10.1016/j.aquatox.2013.04.005
152. Hartmann, N. B.; Engelbrekt, C.; Zhang, J.; Ulstrup, J.; Kusk, K. O.; Baun, A. *Nanotoxicology* **2012,** *7,* 1082–1094. doi:10.3109/17435390.2012.710657
153. Clément, L.; Hurel, C.; Marmier, N. *Chemosphere* **2013,** *90,* 1083–1090. doi:10.1016/j.chemosphere.2012.09.013
154. Simon-Deckers, A.; Loo, S.; Mayne-L'Hermite, M.; Herlin-Boime, N.; Menguy, N.; Reynaud, C.; Gouget, B.; Carrière, M. *Environ. Sci. Technol.* **2009,** *43,* 8423–8429. doi:10.1021/es9016975
155. Seitz, F.; Rosenfeldt, R. R.; Schneider, S.; Schulz, R.; Bundschuh, M. *Sci. Total Environ.* **2014,** *493,* 891–897. doi:10.1016/j.scitotenv.2014.06.092
156. Xiong, D.; Fang, T.; Yu, L.; Sima, X.; Zhu, W. *Sci. Total Environ.* **2011,** *409,* 1444–1452. doi:10.1016/j.scitotenv.2011.01.015
157. Pereira, R.; Rocha-Santos, T. A. P.; Antunes, F. E.; Rasteiro, M. G.; Ribeiro, R.; Gonçalves, F.; Soares, A. M. V. M.; Lopes, I. *J. Hazard. Mater.* **2011,** *194,* 345–354. doi:10.1016/j.jhazmat.2011.07.112
158. Cherchi, C.; Chernenko, T.; Diem, M.; Gu, A. Z. *Environ. Toxicol. Chem.* **2011,** *30,* 861–869. doi:10.1002/etc.445
159. Zhu, X.; Chang, Y.; Chen, Y. *Chemosphere* **2010,** *78,* 209–215. doi:10.1016/j.chemosphere.2009.11.013
160. Amiano, I.; Olabarrieta, J.; Vitorica, J.; Zorita, S. *Environ. Toxicol. Chem.* **2012,** *31,* 2564–2566. doi:10.1002/etc.1981
161. Marcone, G. P. S.; Oliveira, A. C.; Almeida, G.; Umbuzeiro, G. A.; Jardim, W. F. *J. Hazard. Mater.* **2012,** *211,* 436–442. doi:10.1016/j.jhazmat.2011.12.075

162. Clemente, Z.; Castro, V. L.; Jonsson, C. M.; Fraceto, L. F. *J. Nanopart. Res.* **2014,** *16,* 2559. doi:10.1007/s11051-014-2559-z
163. Lovern, S. B.; Klaper, R. *Environ. Toxicol. Chem.* **2006,** *25,* 1132–1137. doi:10.1897/05-278R.1
164. Ma, H.; Brennan, A.; Diamond, S. A. *Environ. Toxicol. Chem.* **2012,** *31,* 1621–1629. doi:10.1002/etc.1858
165. Li, S.; Pan, X.; Wallis, L. K.; Fan, Z.; Chen, Z.; Diamond, S. A. *Chemosphere* **2014,** *112,* 62–69. doi:10.1016/j.chemosphere.2014.03.058
166. Griffitt, R. J.; Luo, J.; Gao, J.; Bonzongo, J.-C.; Barber, D. S. *Environ. Toxicol. Chem.* **2008,** *27,* 1972–1978. doi:10.1897/08-002.1
167. Palaniappan, P. R.; Pramod, K. S. *Food Chem. Toxicol.* **2010,** *48,* 2337–2343. doi:10.1016/j.fct.2010.05.068
168. Velayutham, K.; Rahuman, A. A.; Rajakumar, G.; Santhoshkumar, T.; Marimuthu, S.; Jayaseelan, C.; Bagavan, A.; Kirthi, A. V.; Kamaraj, C.; Zahir, A. A.; Elango, G. *Parasitol. Res.* **2012,** *111,* 2329–2337. doi:10.1007/s00436-011-2676-x
169. Libralato, G.; Minetto, D.; Totaro, S.; Mičetić, I.; Pigozzo, A.; Sabbioni, E.; Marcomini, A.; Ghirardini, A. V. *Mar. Environ. Res.* **2013,** *92,* 71–78. doi:10.1016/j.marenvres.2013.08.015
170. Wang, H.; Wick, R. L.; Xing, B. *Environ. Pollut.* **2009,** *157,* 1171–1177. doi:10.1016/j.envpol.2008.11.004
171. Zhu, X.; Zhou, J.; Cai, Z. *Environ. Sci. Technol.* **2011,** *45,* 3753–3758. doi:10.1021/es103779h
172. Franklin, N. M.; Rogers, N. J.; Apte, S. C.; Batley, G. E.; Gadd, G. E.; Casey, P. S. *Environ. Sci. Technol.* **2007,** *41,* 8484–8490. doi:10.1021/es071445r
173. Rousk, J.; Ackermann, K.; Curling, S. F.; Jones, D. L. *PLoS One* **2012,** *7,* e34197. doi:10.1371/journal.pone.0034197
174. Calder, A. J.; Dimkpa, C. O.; McLean, J. E.; Britt, D. W.; Johnson, W.; Anderson, A. J. *Sci. Total Environ.* **2012,** *429,* 215–222. doi:10.1016/j.scitotenv.2012.04.049
175. Niazi, J. H.; Sang, B.-I.; Kim, Y. S.; Gu, M. B. *Appl. Biochem. Biotechnol.* **2011,** *164,* 1278–1291. doi:10.1007/s12010-011-9212-4
176. Gagné, F.; Auclair, J.; Fortier, M.; Bruneau, A.; Fournier, M.; Turcotte, P.; Pilote, M.; Gagnon, C. *J. Toxicol. Environ. Health, Part A* **2013,** *76,* 767–777. doi:10.1080/15287394.2013.818602
177. Stampoulis, D.; Sinha, S. K.; White, J. C. *Environ. Sci. Technol.* **2009,** *43,* 9473–9479. doi:10.1021/es901695c
178. Zhang, D.; Hua, T.; Xiao, F.; Chen, C.; Gersberg, R. M.; Liu, Y.; Ng, W. J.; Tan, S. K. *Ecol. Eng.* **2014,** *70,* 114–123. doi:10.1016/j.ecoleng.2014.04.018
179. Pang, C.; Selck, H.; Banta, G. T.; Misra, S. K.; Berhanu, D.; Dybowska, A.; Valsami-Jones, E.; Forbes, V. E. *Environ. Toxicol. Chem.* **2013,** *32,* 1561–1573. doi:10.1002/etc.2216
180. Manzo, S.; Rocco, A.; Carotenuto, R.; De Luca Picione, F.; Miglietta, M. L.; Rametta, G.; Di Francia, G. *Environ. Sci. Pollut. Res.* **2011,** *18,* 756–763. doi:10.1007/s11356-010-0421-0
181. Roh, J.-y.; Sim, S. J.; Yi, J.; Park, K.; Chung, K. H.; Ryu, D.-y.; Choi, J. *Environ. Sci. Technol.* **2009,** *43,* 3933–3940. doi:10.1021/es803477u
182. Fabrega, J.; Renshaw, J. C.; Lead, J. R. *Environ. Sci. Technol.* **2009,** *43,* 9004–9009. doi:10.1021/es901706j
183. Hu, C.; Li, M.; Wang, W.; Cui, Y.; Chen, J.; Yang, L. *Toxicol. Environ. Chem.* **2012,** *94,* 732–741. doi:10.1080/02772248.2012.668020
184. Dasari, T. P.; Hwang, H.-M. *Sci. Total Environ.* **2010,** *408,* 5817–5823. doi:10.1016/j.scitotenv.2010.08.030
185. Nair, P. M. G.; Park, S. Y.; Choi, J. *Chemosphere* **2013,** *92,* 592–599. doi:10.1016/j.chemosphere.2013.03.060
186. Park, S.-Y.; Choi, J.-H. *Environ. Eng. Res.* **2010,** *15,* 23–27. doi:10.4491/eer.2010.15.1.428
187. Ghosh, M.; J, M.; Sinha, S.; Chakraborty, A.; Mallick, S. K.; Bandyopadhyay, M.; Mukherjee, A. *Mutat. Res., Genet. Toxicol. Environ. Mutagen.* **2012,** *749,* 60–69. doi:10.1016/j.mrgentox.2012.08.007
188. Kumar, A.; Pandey, A. K.; Singh, S. S.; Shanker, R.; Dhawan, A. *Free Radical Biol. Med.* **2011,** *51,* 1872–1881. doi:10.1016/j.freeradbiomed.2011.08.025
189. Hu, C. W.; Li, M.; Cui, Y. B.; Li, D. S.; Chen, J.; Yang, L. Y. *Soil Biol. Biochem.* **2010,** *42,* 586–591. doi:10.1016/j.soilbio.2009.12.007
190. Hao, L.; Chen, L. *Ecotoxicol. Environ. Saf.* **2012,** *80,* 103–110. doi:10.1016/j.ecoenv.2012.02.017
191. Templeton, R. C.; Ferguson, P. L.; Washburn, K. M.; Scrivens, W. A.; Chandler, G. T. *Environ. Sci. Technol.* **2006,** *40,* 7387–7393. doi:10.1021/es060407p
192. Angelstorf, J. S.; Ahlf, W.; von der Kammer, F.; Heise, S. *Environ. Toxicol. Chem.* **2014,** *33,* 2288–2296. doi:10.1002/etc.2674
193. Klaper, R.; Crago, J.; Barr, J.; Arndt, D.; Setyowati, K.; Chen, J. *Environ. Pollut.* **2009,** *157,* 1152–1156. doi:10.1016/j.envpol.2008.11.010
194. Zhang, H.; He, X.; Zhang, Z.; Zhang, P.; Li, Y.; Ma, Y.; Kuang, Y.; Zhao, Y.; Chai, Z. *Environ. Sci. Technol.* **2011,** *45,* 3725–3730. doi:10.1021/es103309n
195. Mouchet, F.; Landois, P.; Puech, P.; Pinelli, E.; Flahaut, E.; Gauthier, L. *Nanomedicine* **2010,** *5,* 963–974. doi:10.2217/nnm.10.60
196. Mwangi, J. N.; Wang, N.; Ingersoll, C. G.; Hardesty, D. K.; Brunson, E. L.; Li, H.; Deng, B. *Environ. Toxicol. Chem.* **2012,** *31,* 1823–1830. doi:10.1002/etc.1888
197. Chan, T. S. Y.; Nasser, F.; St-Denis, C. H.; Mandal, H. S.; Ghafari, P.; Hadjout-Rabi, N.; Bols, N. C.; Tang, X. *Nanotoxicology* **2012,** *7,* 251–258. doi:10.3109/17435390.2011.652205
198. Strigul, N.; Vaccari, L.; Galdun, C.; Wazne, M.; Liu, X.; Christodoulatos, C.; Jasinkiewicz, K. *Desalination* **2009,** *248,* 771–782. doi:10.1016/j.desal.2009.01.013
199. Kim, K. T.; Klaine, S. J.; Cho, J.; Kim, S.-H.; Kim, S. D. *Sci. Total Environ.* **2010,** *408,* 2268–2272. doi:10.1016/j.scitotenv.2010.01.041
200. Galloway, T.; Lewis, C.; Dolciotti, I.; Johnston, B. D.; Moger, J.; Regoli, F. *Environ. Pollut.* **2010,** *158,* 1748–1755. doi:10.1016/j.envpol.2009.11.013
201. Hanna, S. K.; Miller, R. J.; Lenihan, H. S. *J. Hazard. Mater.* **2014,** *279,* 32–37. doi:10.1016/j.jhazmat.2014.06.052
202. Hu, J.; Wang, D.; Wang, J.; Wang, J. *Environ. Pollut.* **2012,** *162,* 216–222. doi:10.1016/j.envpol.2011.11.016
203. Fabrega, J.; Zhang, R.; Renshaw, J. C.; Liu, W.-T.; Lead, J. R. *Chemosphere* **2011,** *85,* 961–966. doi:10.1016/j.chemosphere.2011.06.066
204. Musante, C.; White, J. C. *Environ. Toxicol.* **2012,** *27,* 510–517. doi:10.1002/tox.20667
205. Bourdiol, F.; Mouchet, F.; Perrault, A.; Fourquaux, I.; Datas, L.; Gancet, C.; Boutonnet, J.-C.; Pinelli, E.; Gauthier, L.; Flahaut, E. *Carbon* **2013,** *54,* 175–191. doi:10.1016/j.carbon.2012.11.024
206. Dai, L.; Syberg, K.; Banta, G. T.; Selck, H.; Forbes, V. E. *ACS Sustainable Chem. Eng.* **2013,** *1,* 760–767. doi:10.1021/sc4000434

207. Pakarinen, K.; Petersen, E. J.; Leppänen, M. T.; Akkanen, J.; Kukkonen, J. V. K. *Environ. Pollut.* **2011,** *159,* 3750–3756. doi:10.1016/j.envpol.2011.07.014
208. Pakarinen, K.; Petersen, E. J.; Alvila, L.; Waissi-Leinonen, G. C.; Akkanen, J.; Leppänen, M. T.; Kukkonen, J. V. K. *Environ. Toxicol. Chem.* **2013,** *32,* 1224–1232. doi:10.1002/etc.2175
209. Sun, H.; Zhang, X.; Niu, Q.; Chen, Y.; Crittenden, J. C. *Water, Air, Soil Pollut.* **2007,** *178,* 245–254. doi:10.1007/s11270-006-9194-y
210. Zhang, X.; Sun, H.; Zhang, Z.; Niu, Q.; Chen, Y.; Crittenden, J. C. *Chemosphere* **2007,** *67,* 160–166. doi:10.1016/j.chemosphere.2006.09.003
211. Sun, H.; Zhang, X.; Zhang, Z.; Chen, Y.; Crittenden, J. C. *Environ. Pollut.* **2009,** *157,* 1165–1170. doi:10.1016/j.envpol.2008.08.022
212. Zhu, X.; Wang, J.; Zhang, X.; Chang, Y.; Chen, Y. *Chemosphere* **2010,** *79,* 928–933. doi:10.1016/j.chemosphere.2010.03.022
213. Larue, C.; Castillo-Michel, H.; Sobanska, S.; Trcera, N.; Sorieul, S.; Cécillon, L.; Ouerdane, L.; Legros, S.; Sarret, G. *J. Hazard. Mater.* **2014,** *273,* 17–26. doi:10.1016/j.jhazmat.2014.03.014
214. Larue, C.; Laurette, J.; Herlin-Boime, N.; Khodja, H.; Fayard, B.; Flank, A.-M.; Brisset, F.; Carriere, M. *Sci. Total Environ.* **2012,** *431,* 197–208. doi:10.1016/j.scitotenv.2012.04.073
215. Schwabe, F.; Schulin, R.; Limbach, L. K.; Stark, W.; Buerge, D.; Nowack, B. *Chemosphere* **2013,** *91,* 512–520. doi:10.1016/j.chemosphere.2012.12.025
216. Lin, S.; Reppert, J.; Hu, Q.; Hudson, J. S.; Reid, M. L.; Ratnikova, T. A.; Rao, A. M.; Luo, H.; Ke, P. C. *Small* **2009,** *5,* 1128–1132. doi:10.1002/smll.200801556
217. Li, L.; Sillanpaa, M.; Tuominen, M.; Lounatmää, K.; Schultz, E. *Ecotoxicol. Environ. Saf.* **2013,** *88,* 89–94. doi:10.1016/j.ecoenv.2012.10.024
218. Jarvis, T. A.; Miller, R. J.; Lenihan, H. S.; Bielmyer, G. K. *Environ. Toxicol. Chem.* **2013,** *32,* 1264–1269. doi:10.1002/etc.2180

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Nanocuration workflows: Establishing best practices for identifying, inputting, and sharing data to inform decisions on nanomaterials

Christina M. Powers[1,2], Karmann A. Mills[3], Stephanie A. Morris[4], Fred Klaessig[5], Sharon Gaheen[6], Nastassja Lewinski[7] and Christine Ogilvie Hendren[*8]

Commentary

Open Access

Address:
[1]National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, 109 TW Alexander Drive, Research Triangle Park, NC 27711, USA, [2]Currently: Office of Transportation and Air Quality, Office of Air Quality, 2000 Traverwood Rd, Ann Arbor, MI 48105, USA, [3]RTI International, 3040 Cornwallis Rd., Research Triangle Park, NC 27709, USA, [4]Office of Cancer Nanotechnology Research, National Cancer Institute/NIH, 31 Center Drive, Bethesda, MD 20892, USA, [5]Pennsylvania Bio Nano Systems, LLC, 69 Homestead Drive, Doylestown, PA 18901, USA, [6]Leidos Biomedical Research Inc., Frederick National Laboratory for Cancer Research, 8560 Progress Drive, Frederick, MD 21702, USA, [7]Department of Chemical and Life Science Engineering, Virginia Commonwealth University, 601 W. Main St., P.O. Box 843028, Richmond, VA 23284, USA and [8]Center for the Environmental Implications of NanoTechnology (CEINT), Duke University, P.O. Box 90287, 121 Hudson Hall, Durham, NC 27708, USA

Email:
Christine Ogilvie Hendren[*] - christine.hendren@duke.edu

* Corresponding author

Keywords:
curation; informatics; nanoinformatics; nanomaterials; workflows

Beilstein J. Nanotechnol. **2015,** *6,* 1860–1871.
doi:10.3762/bjnano.6.189

Received: 17 March 2015
Accepted: 07 August 2015
Published: 04 September 2015

This article is part of the Thematic Series "Nanoinformatics for environmental health and biomedicine".

Guest Editor: R. Liu

Abstract

There is a critical opportunity in the field of nanoscience to compare and integrate information across diverse fields of study through informatics (i.e., nanoinformatics). This paper is one in a series of articles on the data curation process in nanoinformatics (nanocuration). Other articles in this series discuss key aspects of nanocuration (temporal metadata, data completeness, database integration), while the focus of this article is on the nanocuration workflow, or the process of identifying, inputting, and reviewing nanomaterial data in a data repository. In particular, the article discusses: 1) the rationale and importance of a defined workflow in nanocuration, 2) the influence of organizational goals or purpose on the workflow, 3) established workflow practices in other fields, 4) current workflow practices in nanocuration, 5) key challenges for workflows in emerging fields like nanomaterials, 6) examples to make these challenges more tangible, and 7) recommendations to address the identified challenges. Throughout the article, there

is an emphasis on illustrating key concepts and current practices in the field. Data on current practices in the field are from a group of stakeholders active in nanocuration. In general, the development of workflows for nanocuration is nascent, with few individuals formally trained in data curation or utilizing available nanocuration resources (e.g., ISA-TAB-Nano). Additional emphasis on the potential benefits of cultivating nanomaterial data via nanocuration processes (e.g., capability to analyze data from across research groups) and providing nanocuration resources (e.g., training) will likely prove crucial for the wider application of nanocuration workflows in the scientific community.

Introduction

A tremendous growth in resources and tools to hold and organize large quantities of data has increased data availability to scientists, engineers, and others in the scientific community. Greater access to data repositories, data sharing platforms, and data visualization tools creates opportunities to compare and integrate information across a variety of diverse fields of study. For fields like nanoscience, or the study of materials at the nanoscale, this opportunity is particularly important given the wide array of disciplines that are inherently involved in synthesizing, testing, regulating, using, and developing new nanomaterial applications (e.g., chemistry, toxicology, ecology, risk assessment, material science). The complexity of developing tools for accessing, sharing, and viewing data relevant to nanomaterials has generated an entire field known as nanoinformatics. This paper is one in a series and focuses on a particular aspect of the nanoinformatics field, namely, the curation of data related to nanoscale materials (nanocuration) [1]. For this purpose, the experiences of three organizations (NCI, RTI and CEINT found in the listing of authors) were compiled into a questionnaire that was submitted to a further four organizations in order to describe current practices. Articles in this series are developed by the Nanomaterials Data Curation Initiative (NDCI), which is part of the National Cancer Informatics Program Nanotechnology Working Group [1]. Other articles in this series discuss several key aspects of nanocuration (temporal metadata, data completeness, database integration), while the specific focus of this article is on the nanocuration workflow, or the process of identifying, inputting, and reviewing nanomaterial data in a data repository (Figure 1).

Discussion

i. Importance and relevance of the workflow to nanocuration

A workflow is a critical component of nanocuration for several reasons. A workflow: 1) defines the process for data curation, 2) allows for comparison across data repositories to determine areas of standardization and bottlenecks, and 3) provides a consistent process for understanding the quality and completeness of a dataset [2]. Defining the process for data curation through the creation of a workflow presents an opportunity for individuals in an organization to establish and standardize the specific steps involved in identifying, inputting, and reviewing nanomaterial data for storage in the associated repository. A focused effort on each step in the workflow facilitates the identification of critical elements within and between each step, such as information transfers from one individual to another, quality control checks, and access rights necessary to input or review data. When individuals in an organization or institution document and define the data curation process, they not only create a valuable resource for future review, revision, and quality assurance/control (QA/QC) measures, but institutionalized workflows also facilitate the creation of training materials. Training materials in turn enable multiple curators to work in parallel, with a streamlined QA/QC process, and thereby mitigate redundant checking of curation decisions. This is critical to nanoinformatics progress, since curation (manual data entry or transfer from a data source) is the primary bottleneck to data collection once a repository structure and language are solidified. Related to the second aspect of the importance of a workflow, comparison between data resources, workflows serve as a written indicator of differences or similarities in underlying assumptions, order of operations, and standardization levels of, for example, data completeness. In comparing workflows from different data repositories, curators may identify common challenges (e.g., acquiring additional experimental design details from authors) or opportunities to leverage resources between repositories. In some instances, such workflow comparisons may lead to the use of common file formats, vocabulary, and structure. Common file, vocabulary, and structure conventions across data repositories in turn facilitates researchers and others utilizing data from across repositories in analyses. Finally, workflows facilitate researchers and other data users understanding the quality and completeness of the curated data. Indeed, in addition to the data quality support provided by the consistent curation practices of a defined workflow, the assessment of data quality and completeness is expressly included in two of the common curation steps articulated in Figure 1. Data quality and completeness is the topic of another article in this series and, thus, will not be discussed at length in this article. Nevertheless, understanding these concepts in various repositories is necessary for researchers or others using the data since different levels of quality or completeness are required for

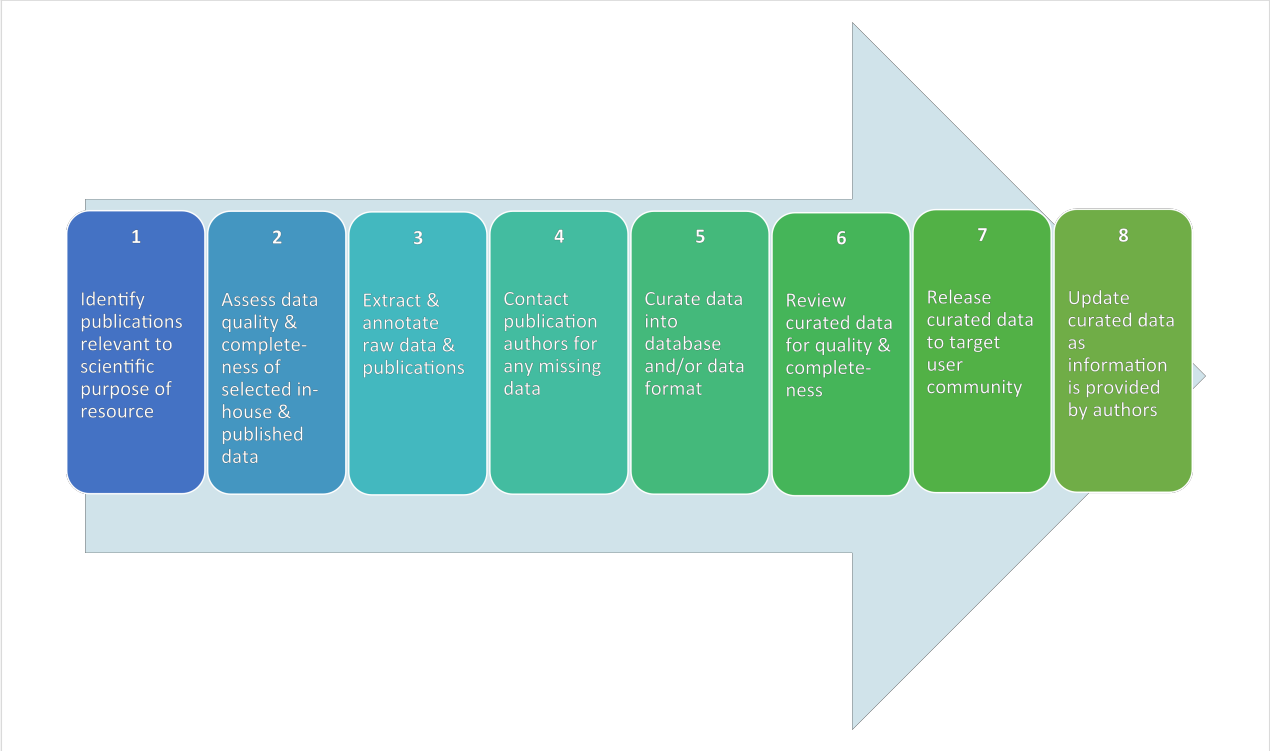


Figure 1: Common steps in nanocuration. The steps commonly included in nanocuration workflows are illustrated, including: 1) Identification of publications relative to the intended scientific purpose; 2) Preliminary assessment of data quality and completeness of selected in-house or publication data for data quality and completeness (with assumption that any in-house data would be pre-identified within a project prior to the wider publication search referred to in Step 1); 3) Data extraction of raw data and/or data from the publication; 4) Communication with publication authors; 5) Curation of data into the intended repository and/or data format (e.g., ISA-TAB-Nano) leveraging common data elements (CDEs) from relevant ontological resources (e.g., NanoParticle Ontology [NPO]); 6) Review of curated data for data quality and completeness; 7) Release of curated data; 8) Update of curated data as additional information is received from the authors. Though shown here in linear fashion, the order of these common steps for an individual process may be flexible and iteration is expected. The specific steps in a workflow may also differ across repositories depending on the intended purpose of the nanomaterial resource.

different uses of data (e.g., research prioritization, screening level decisions about hazard, quantitative risk assessment) [3,4].

ii. Influence of organizational purpose or goals on design and application of a workflow

A discussion of a curation workflow requires an understanding of the curation purpose, (i.e., the objectives of the community sponsoring the data repository and the intended function of the repository). The diversity of communities and organizations involved with nanocuration reflects the multidisciplinary nature of nanotechnology. This diversity also has implications regarding workflow details for each separate curation effort, which inevitably involves validating data sources or characterizing the "quality" of data entries. The three examples that follow demonstrate the interplay.

For instance, the objective of the National Cancer Institute's (NCI) cancer Nanotechnology Laboratory (caNanoLab; https://cananolab.nci.nih.gov/caNanoLab/) data portal is to provide a comprehensive resource for individuals in the biomedical nanotechnology research community to share data that supports the use of nanotechnology in biomedicine (e.g., novel cancer diagnostic or therapeutic tools and technologies). As part of NCI, caNanoLab uses a nanotechnology information object model (nano-OM) to capture standardized nanomaterial composition and characterization concepts [5]. The nano-OM facilitates the use of Common Data Elements (CDEs) for cancer nanotechnology research described in an established data format for nanomaterial data, NanoParticle Ontology [6] (The term Common Data Elements is used in particular by the National Institutes of Health (NIH) in describing their controlled vocabulary approaches, and refers to standardized data types that are consistent across datasets and resources). The use of the nano-OM in caNanoLab supports queries on publications, protocols, nanomaterials and associated compositions and characterizations. These data can be used by modeling and simulation tools to discover data patterns that guide decisions on new biomedical research directions and novel nanomaterials. Users can focus on particular nanomaterial(s) and biological phenomena through selection criteria for literature and research protocol sources that are curated into the repository. Based on the objectives of the repository, the workflow process must

incorporate data and metadata (i.e., information about the data) related to: 1) nanomaterial physicochemical characteristics, 2) in vitro and in vivo assays that analyze nanomaterial properties, biological interactions, toxicity, or efficacy, and 3) information on the protocols used to analyze these nanomaterials and any associated publications.

In contrast, the purpose of RTI International's Nanomaterial Registry (NR; https://www.nanomaterialregistry.org/) is to collect validated data from a broad field of accessible nanomaterial sources relevant to not only medical applications, but also the environmental implications of nanomaterials and their impact on human health and safety. While selection criteria regarding data sources remain a necessary element to the curation workflow, the NR uses an internally defined compliance score (minimal information about nanomaterials [MIAN]) to communicate the relative extent of physicochemical test data completeness to users [7]. This workflow process allows the NR to convey data quality information without restricting the incorporation of data into the repository due to a lack of information on experimental design, conduct, or outcome reported in the literature.

Finally, the Center for Environmental Implications of NanoTechnology (CEINT; http://www.ceint.duke.edu/) generates a wide array of nanomaterial data including characterization of pristine and naturally transformed particles, fate and transport data, toxicity data, and information on ecological impacts not limited to toxicity (e.g., nutrient cycling impacts) from laboratories within the Center and from collaborators. These laboratories represent a variety of scientific disciplines and use or develop well-founded, yet innovative procedures that may eventually be standardized. The CEINT-NIKC (CEINT NanoInformatics Knowledge Commons) focuses on developing the infrastructure and data gathering practices necessary to capture the full value of the Center's multidisciplinary activities for integration and analysis not only of internally generated data, but also with any relevant literature that can also be curated into the system. The expectation is that some of the critical data may reside beyond publicly available peer-reviewed articles, and thus may need to be solicited directly from researchers (e.g., via theses, lab notebooks, spreadsheets). In this case, the primary selection criterion for including data in the repository is that the data are directly relevant to the driving research questions of the Center. The driving research questions focus on: 1) elucidating the characteristics of materials and systems, and 2) mechanisms driving nanomaterial behavior in complex systems; thus, data in the repository span a range of traditionally separate disciplines. Furthermore, the dynamic nature of nanomaterials in terms of changes in chemical identity as they migrate environmentally must be matched by an equally dynamic interaction of these disciplines in regularly evaluating both current and past data. This is not a matter of only data quality, but also of identifying new, useful concepts that bind the disciplines together for a common community purpose. The workflow process thus must be well-defined, yet flexible enough to incorporate new types of data or linkages across data types (e.g., dissolution rate at a particular pH and toxicity in a specific organism).

These three organizations (caNanoLab, NR, and CEINT-NIKC) differ in sourcing data to be curated (established protocols, literature sources, primarily internal or fully external), the intended users (medical researchers conversant with bioinformatics, the general nanotechnology public, and Center investigators), and function (modeling for repeatable experimentation, accessing nanomaterial sources, exploratory research requiring coordination among disciplines). For each, "high quality" means fit-for-purpose and thus the curation workflow is integral to meeting the community's goal. The existence of established workflows in each organization allows for the identification of common challenges associated with the development or use of the workflow process. These challenges include: 1) establishing a minimal information set to include in the workflow, 2) determining a vocabulary (based on standards as much as possible) for the curators to use, and 3) defining how the data quality and validation are ensured in the workflow. In all three cases, the purposes of the repository necessitated that the workflow design include an opportunity to contact the investigators who developed the data (i.e., authors of peer-reviewed articles, Center members) in order to obtain complete and high quality data sets. In addition, the workflow can help facilitate sharing data across these or other resources. For instance, different organizations can incorporate a common data format in their respective workflows. An example data format is ISA-TAB-Nano, which is a file transfer protocol for querying among federated data repositories that are independently maintained by organizations with related, but not necessarily overlapping objectives [8]. Communication among federated repositories allows each separate community to tailor the workflow to their available resources, especially in this fluid period of debates regarding dose metrics, physicochemical characterization data sets, and protocol standardization.

Notably, in some organizations the term "curation" may be used in a less formal sense to simply describe the process used to identify data and integrate it into a data repository system. The process to formalize a curation workflow may take place after an initial phase of simply working through the informal process. The process of formalizing the curation workflow may be particularly important when a group expands or opens their repository to contributions from

stakeholders outside of the research group. NanoDMS (http://biocenitc-deq.urv.cat/nanodms/), an FP7 project in the European Union, represents an example of using an informal curation workflow that may become more formalized during the group's maturation. Ultimately, the purpose of the organization or group that develops the data repository not only drives the development of the workflow process, but may also determine how and when the workflow process is incorporated into the curation effort.

iii. Established methods for workflows in mature fields

Organizations or groups that are working to incorporate or further develop a workflow for nanomaterial data curation may benefit from adapting methods established in other, perhaps more mature, fields (e.g., bioinformatics). In general, other fields utilize one of two approaches: 1) establish specific file formats with standardized vocabularies and fields, or 2) create collection formats at a generalized level to allow for the variation and uncertainty across a field. As a specific example of the first approach, the genomics community has developed a curation workflow that uses standardized file formats for both metadata and raw DNA sequence data for submissions into standard repositories [9]. A validation tool (Picard, http://broadinstitute.github.io/picard/) is then used to verify that the data fits the standard. An example of the second approach can be found within the *C. elegans* field with the WormBase repository (http://www.wormbase.org/#01-23-6). Notably, the genomics and WormBase workflows also take different approaches to the responsibility of entering data into a public repository. The genomics field requires authors to submit their own data using the provided file formats, whereas WormBase has a group of data curators responsible for identifying, entering, and managing data in the repository. Giving authors the responsibility of submitting data in standard formats to established repositories is an avenue for discussion in the nanomaterial community. Indeed, the NCI Alliance for Nanotechnology in Cancer now expects grantees to submit and share data using an established repository, caNanoLab (http://grants.nih.gov/grants/guide/rfa-files/RFA-CA-14-013.html). The extent to which other funding organizations add requirements for authors to share data in specified repositories will likely depend on a variety of factors, including the usability and accessibility of simple workflows for adding data to a repository.

iv. Current practice in nanocuration workflows – Stakeholder responses to questions

To understand how practices in more established fields compare with the current state of nanocuration workflow practices across the field, the NDCI Leadership requested input from several individuals currently involved in developing nanomaterial data repositories. Seven representatives from organizations of different sizes and sectors (e.g., academia, government) responded to requests for input. Three of the respondents are also authors of this article since they represent organizations active in the nanocuration field. While the responding organizations represent a diverse swath of the nanomaterial field, the views presented here are not intended to provide a comprehensive representation of nanocuration workflows; rather, the intent of presenting these stakeholder responses is to help identify challenges and opportunities for improvement in nanocuration workflows by providing a snapshot in time of current practices. Additional details on the process used to contact and gain information from respondents is available in [1]. Briefly, the NDCI requested input from stakeholders in the fall of 2014 and winter of 2015 (November to January) on questions related to: 1) Sourcing data for nanocuration workflows, 2) Entering and reviewing data in a workflow, 3) Creating and revising a workflow, and 4) Interacting with other organizations to develop a workflow or populate their repository. Stakeholder responses are summarized below and in Figures 2–5 with additional details available in Supporting Information File 1.

a. Sourcing data for nanocuration workflows

As shown in Figure 2, two stakeholders consistently use established criteria for selecting data from the peer-reviewed literature to include in their repository, while four others report using loosely established, situation-dependent criteria. Most stakeholders (4 of 7) do supplement information in journal articles with information from other sources (e.g., searching for the paper in other databases) (Figure 2), since this approach provides a valuable source of supplemental data (see Supporting Information File 1 for details). When using sources other than peer-reviewed articles, stakeholders did consistently use established criteria (Figure 2). However, the majority of stakeholders (5 of 7) responded that their workflow does not currently include a quality assurance (QA) process. The two examples of a QA process included: 1) a manual review of data identified through a semi-automatic natural language processing (NLP) data extraction procedure, and 2) a second individual checking the initial curation (see Supporting Information File 1 for details).

b. Entering and reviewing data in a workflow

After determining how to source nanomaterial data for a repository, repository developers may establish guidelines for entering and reviewing data in the workflow. Of the stakeholders who responded to the NDCI request, just over half had individuals who are explicitly identified as a curator enter nanomaterial data (4 of 7 explicitly identified curators, with 3 of the 4 being specifically trained as a curator; Figure 3). In most cases, there

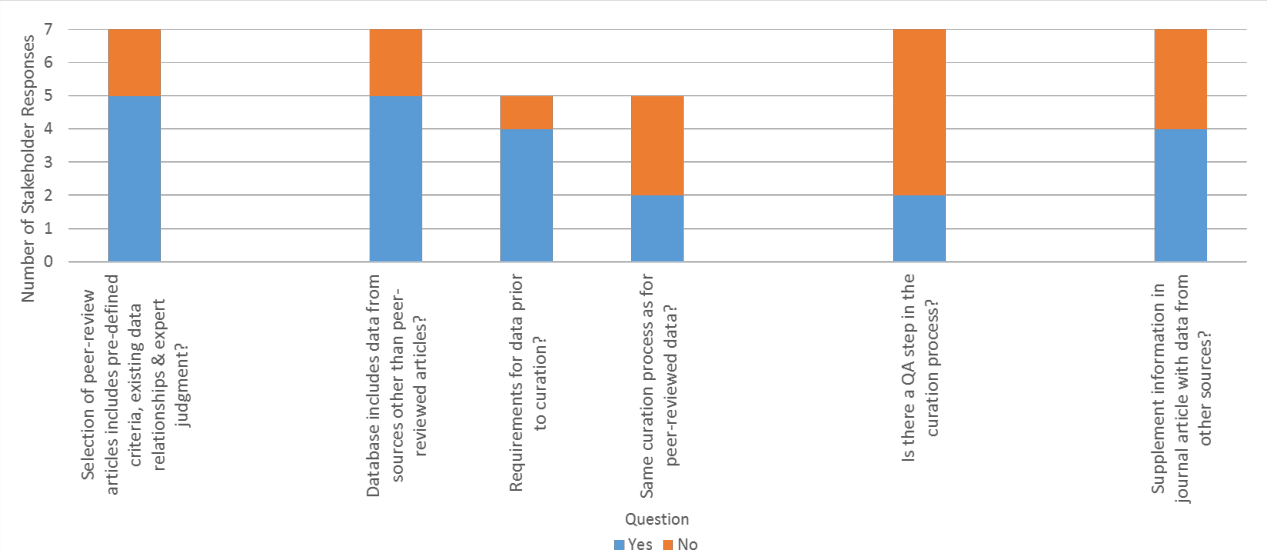


Figure 2: Stakeholder responses regarding sourcing Data. Stakeholder responses to questions related to sourcing nanomaterial data in a workflow for a data repository. Full text of stakeholder responses is available in Supporting Information File 1.

was no process for non-curators to submit data to the repository (Figure 3). One example of a process for others to submit data consisted of researchers sending data in a standardized format (ISA-TAB-Nano) to a single person designated as responsible for data entry. Another stakeholder has a clearly defined and publicly available user's guide for external submissions (see Supporting Information File 1 for details). Most respondents did note plans to develop a formal process for data submission in the future (see Supporting Information File 1 for details). All stakeholders distinguish peer-reviewed data from other types of information; however, not all further distinguish the data type (e.g., protocols, raw or unprocessed data) and some note that their repository only includes in-house data or only includes peer-reviewed data (Figure 3 and Supporting Information File 1). The majority of stakeholders (4 of 7) have a process in place to weed out or deprecate data, although they generally do not have a formal change log in place to document changes (only 1 of 7 stakeholders has a change log) and only two of

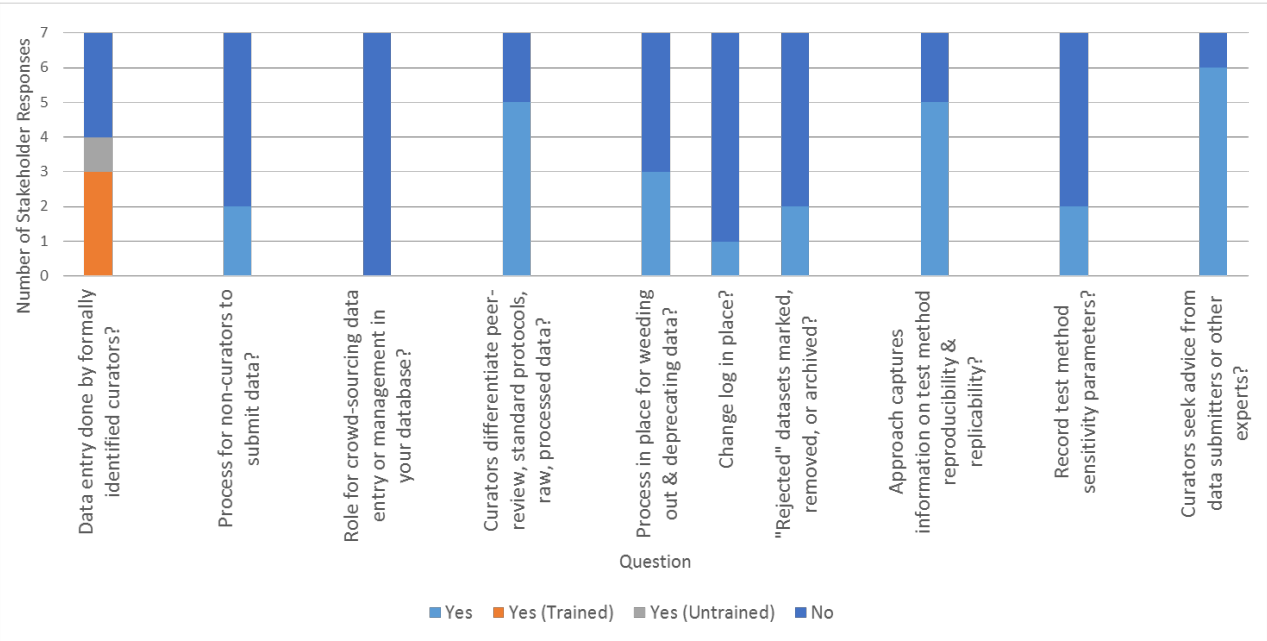


Figure 3: Stakeholder responses regarding data entry and review. Stakeholder responses to questions related to entering and reviewing data in a workflow. Full text of stakeholder responses is available in Supporting Information File 1.

seven explicitly mark and/or remove "rejected data" (Figure 3). Five of the stakeholders currently capture information related to test method reproducibility or replicability (Figure 3), though this typically occurs only through indirect measures (e.g., number of replicates, number of times protocol has been run in-house), or only in instances that data appear "interesting" (see Supporting Information File 1 for details). Only two of the stakeholders who responded currently capture information on test method sensitivity in completing their workflow (Figure 3); in one case this refers to the structural ability to incorporate sensitivity analyses if included in the publication, while in the other the functionality to carry out sensitivity analyses through query was part of the system design. In contrast, almost all stakeholders (6 of 7) consult advisors with relevant expertise if questions arise about data being entered through the workflow (Figure 3).

c. Creating and revising a workflow

As discussed in Section i (Importance of the workflow in data curation), there are a number of advantages to capturing the process for sourcing, entering, and reviewing data into a formal workflow. The majority of the stakeholders stated that they have written a workflow document to capture their process (5 of 7; Figure 4). These documented processes range in their formality and level of development; two stakeholders noted that they only recently developed a written workflow, while another stated that they were in the process of developing the documentation (see Supporting Information File 1 for details). The majority of stakeholders (4 of 7) reported drawing on other resources when creating their workflow. Most stakeholders (5 of 7) do not have a protocol in place to manage changes to their workflow (Figure 4), which might be expected since workflow documentation is in the early stages for this group of respondents. In addition, many (4 of 7) replied that they have not established specific future milestones for workflow improvements. In contrast, most stakeholders (6 of 7) did have a process in place to apply changes in the workflow to previously curated data (Figure 4). Such change processes seem particularly important in a field where the resource infrastructures and the curation processes are still in development.

d. Interacting with other organizations to develop a workflow or populate their repository

Efforts to work with publishers, journal article authors, and others involved in nanocuration can be beneficial in developing a workflow and populating a repository. However, based on stakeholder responses, it may be too early in the development of nanoinformatics infrastructures to see the establishment of such relationships. Respondents stated that there has been little activity to date in the nanocuration field to work with publishers on these issues, although there is recognition of the eventual importance of this aspect (Figure 5). One stakeholder did express interest in discussing the topic with publishers and noted that their organization includes individuals who serve as journal editors, which could facilitate such conversations (see Supporting Information File 1 for details). Compared to efforts to work with publishers, stakeholders indicated that there have been more efforts to contact journal article authors (5 of 7 stakeholders indicated they contacted authors; Figure 5). Yet, stakeholders who did make an effort to contact authors had dichotomized views of how willing authors were to share data or characterization protocols (Figure 5). Several stakeholders stated that authors were generally cooperative (but included caveats), while another stated that authors generally were not helpful. The respondent suggested that a lack of cooperation from authors could be due to a lack of interest in curating their

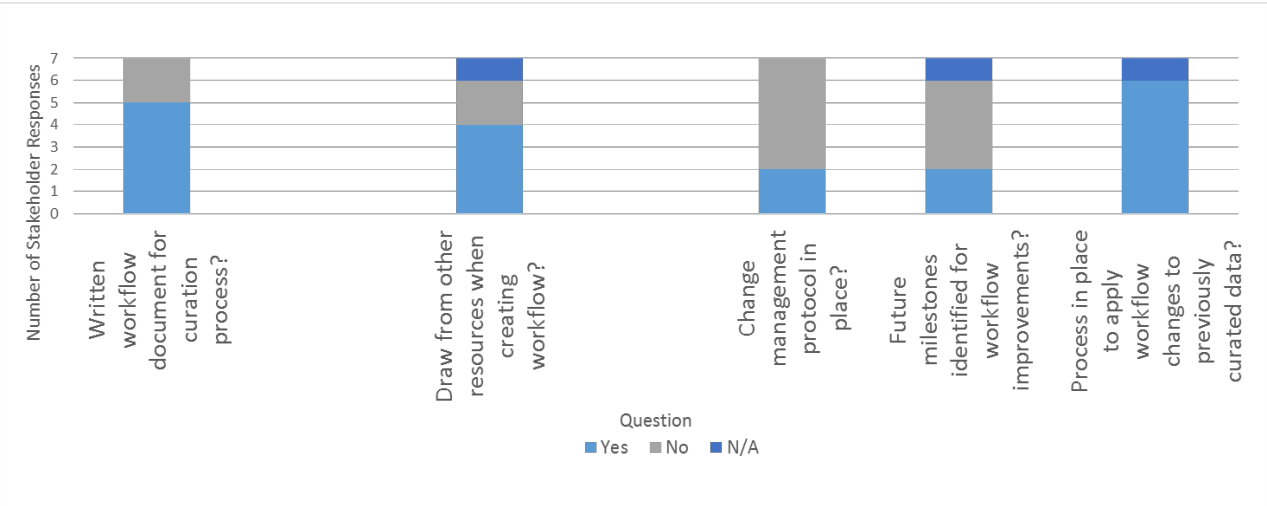


Figure 4: Stakeholder responses regarding creation and revision. Stakeholder responses to questions related to creating and revising a written workflow. Full text of stakeholder responses is available in the Supporting Information File 1.

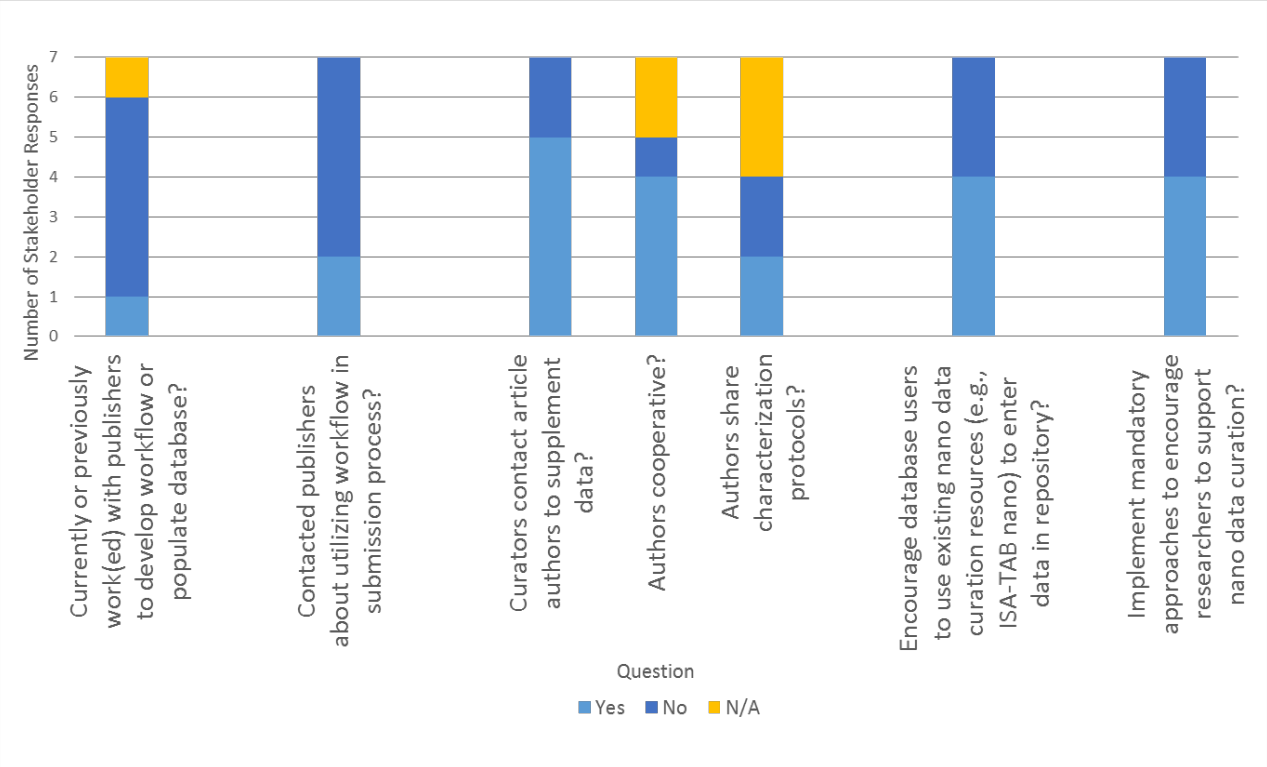


Figure 5: Stakeholder responses regarding working with other organizations. Stakeholder responses to questions related to interacting with other organizations to develop a workflow or populate a data repository. In each panel, the response categories (e.g., “yes”, “no”, “N/A”) for each question are provided in the legend. Questions are listed on the x-axis and the number of stakeholders responding in each category is on the y-axis. Full text of stakeholder responses is available in Supporting Information File 1.

data and/or the fact that authors were no longer in the same position (e.g., a PhD student generated data but had since graduated). One stakeholder noted that concerns about intellectual property rights might limit some authors’ willingness to share characterization protocols, while another suggested using established mechanisms to connect with researchers (e.g., the website ResearchGate) when requesting information from authors (see Supporting Information File 1 for details). In the longer term, curators could avoid the need to contact authors for additional information if researchers also reported their data using existing nanocuration resources (e.g., ISA-TAB-Nano) or other metadata tracking frameworks; however, only four of seven stakeholders stated that they encourage individuals to submit data in a standard format (e.g., ISA-TAB-Nano) (Figure 5). One reason that stakeholders provided for not using a standard format is that the data repository is only used in-house (see Supporting Information File 1 for details). To encourage more support for researchers to use nanocuration resources, stakeholders offered a variety of suggestions. Just over half of the stakeholders supported journals or funding agencies mandating that researchers use standard formats, while the other stakeholders emphasized the need for voluntary training or educational resources to encourage researchers to invest the time necessary for capturing their data in standard formats. Many stakeholders emphasized the need for significant funding to support the establishment and adoption of standardized data sharing mechanisms (Figure 5; see Supporting Information File 1 for details).

v. Key challenges related to curation workflows for emerging and nanomaterials

While current practice in other, more mature fields provides some insight for the development of nanocuration workflows, the stakeholder responses described above indicate there are several challenges that the community will need to address in order to more efficiently and effectively develop nanocuration workflows. Some challenges are perhaps universally applicable to a variety of fields, both emerging and established, while others are more unique to emerging fields such as nanomaterials (Figure 6). Both types of challenges are discussed below in the context of what they imply for the development and application of data curation workflows in the nanomaterial community. The next section provides examples to illustrate the challenges outlined here.

Challenges that may impact a workflow and are generally applicable across the scientific community, include: 1) incomplete data in publications (i.e., an insufficient amount of infor-

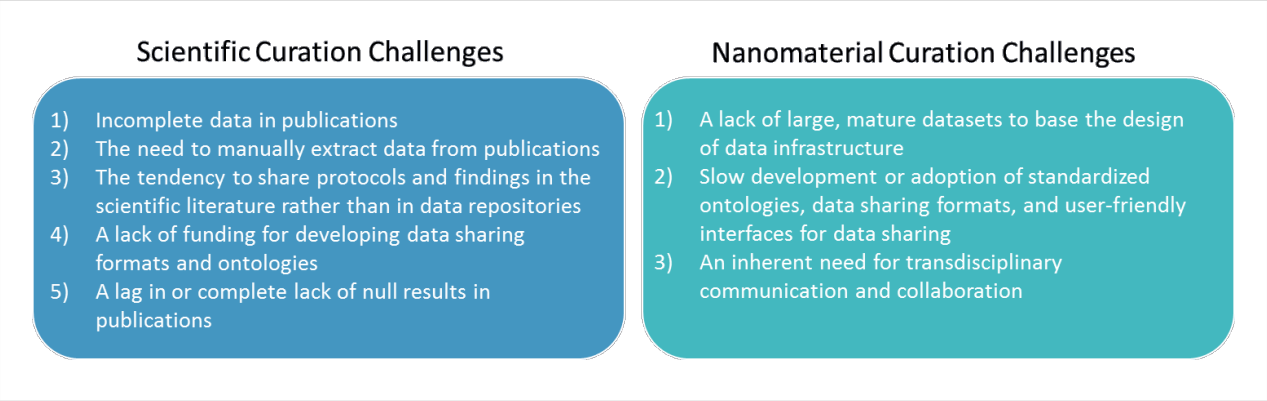


Figure 6: Scientific and nanomaterial curation challenges. Nanomaterial curation challenges expand on curation challenges inherent in general scientific curation.

mation to reproduce an experiment or enable nanomaterial comparisons), 2) the need to extract data manually from publications, 3) a tendency to share protocols and findings in the scientific literature rather than in data repositories, 4) a lack of funding for developing data sharing formats and ontologies, and 5) a lag in or complete lack of null results in publications (i.e., journals rejecting manuscripts with null findings, or researchers not submitting data for publication until it includes at least one positive finding). These challenges generally impact how a workflow is or can be used (e.g., incomplete data in publications may require that the workflow include direct interaction with study authors to the extent possible). However, a workflow alone is unlikely to influence the scientific community to change its practices (e.g., investigators are unlikely to include additional data in publications because those data are required for one or more data repositories). To overcome these challenges in the nanomaterial community, and the scientific community more broadly, community members will need to understand the impact of current practices on data utility and applicability. Greater discussion between community members about the value of large data repositories and data sharing practices may have the greatest potential of driving toward resolution of these challenges. While the incentive of access to larger, interoperable datasets may encourage researchers and funding agencies to extend time, effort, and funds toward curating data into shared repositories, additional incentives will likely be necessary. As expanded on in Section vii, several incentives could drive researcher-contribution of data, including: 1) funds for data sharing by funding organizations, 2) requirements to submit data to central repositories from funding organizations or publishers, and 3) publication credit for dataset submission (e.g., receipt of a digital object identifier for data submissions). Ideally, these actions would be supported by data gathering software (e.g., electronic notebooks) that can export datasets in standard formats (e.g., ISA-Tab-Nano) and require minimal data restructuring by researchers. This would thus facilitate data curation that does not require a concerted effort separate from the research itself.

In contrast to broadly applicable challenges, challenges that are more unique to emerging fields, like nanomaterials, include: 1) a lack of large, mature datasets on which to base the design of data infrastructure, 2) slow development or adoption of standardized ontologies, data sharing formats, and user-friendly interfaces for data sharing, and 3) an inherent need for transdisciplinary communication and collaboration. Nanomaterial data workflows can likely facilitate progress in overcoming these challenges. For instance, by establishing and using a data curation workflow, caNanoLab, the Nanomaterial Registry, and CEINT-NIKC are all developing large data repositories that can guide the development of infrastructure for future nanomaterial data repositories as well as iterate improvements to themselves. The development and use of a workflow also inherently facilitates transdisciplinary communication and collaboration through the incorporation of data from a variety of domains (e.g., physicochemical, environmental transport, toxicity). Indeed, a workflow process is one aspect of a nanoinformatics approach that can actually be defined and followed in advance of a mature field, as a part of intentionally documenting research in pursuit of eventual data standardization. Seeing workflows as a critical part of overcoming some of the current challenges to nanocuration is perhaps one way to emphasize the importance of the nanomaterial community utilizing and further developing this integral piece of data curation. While nanocuration is being discussed in this section in terms of challenges, this effort is a response to the even greater challenge posed by the responsible development of an emerging technology that is fully expected to generate a large number of products and applications. Continuing the current tendency for each organization to maintain its own database with local interpretations of acceptable test protocols and data interpretation will impede the pace of innovation when organizations repeat work already

done, but not accessible to others, or when firms and regulators are not aware of data pertinent to their discussions.

vi. Examples of the identified challenges

Examples of the challenges outlined above help illustrate the importance of these issues and their impact on the goal of understanding nanomaterial interactions and behavior in different media. For instance, data curators at caNanoLab encounter several of the challenges outlined above, and these in turn impede the efficiency and effectiveness of the workflow. Related to the challenge of incomplete information in publications, caNanoLab curators have identified incomplete datasets, missing steps in protocol descriptions, and figures without underlying data or descriptions. Without these details, curators are unable to assess data quality and complete the curation workflow. In some cases, curators can obtain the missing information from study authors, but this slows the workflow process and is not always possible. Related to challenges more specific to the nanomaterial community, caNanoLab curators note that inconsistent terminology and a lack of automated data sharing tools impede the efficient implementation of their workflow.

Data curators at the Nanomaterial Registry have collaborated with CEINT-NIKC researchers to curate some of the Center's findings into the Registry. While this collaboration will ultimately benefit the nanomaterial community by adding to the publicly-accessible repository, it actually highlighted some of the challenges outlined above. Specifically, CEINT-NIKC staff trained to curate the Center's data into the Registry found that: 1) more data could be gathered when speaking directly to the researcher rather than relying on their publications (e.g., publications did not always share all of the physicochemical characterizations available on the nanomaterial tested, which were later captured by speaking with the researcher), and 2) in at least one case the original researcher had moved on from CEINT and targeted communication, with an associated time lag, was needed to retrieve additional information. Collaborators from both the Registry and CEINT concluded that curating from literature is not an optimal solution. This finding, and similar experiences across the nanocuration field, suggests that approaches like the NCI Alliance for Nanotechnology in Cancer that require authors to add data into a public repository may become more common practice moving forward.

vii. Recommendations: Opportunities to leverage existing nanoinformatics resources for workflows and practical next steps for the nanomaterial community

Several opportunities exist to address the challenges discussed above in ways that leverage existing nanoinformatics resources. These opportunities can be broadly categorized in two areas: 1) to empower authors to submit data to repositories using standardized formats (e.g., ISA-TAB-Nano [8]) and nomenclature, and 2) to expand and further develop existing tools and repositories for nanomaterial data. Specific actions that the nanomaterial community can take to make progress in each opportunity area are outlined below to facilitate collaborative efforts in nanocuration.

Related to the first opportunity area, current practices in the nanomaterial community generally demand that curators of data repositories manually enter data from publications in the scientific literature. This practice not only slows down the workflow process, but also can frequently result in incomplete data entries or errors. To address this issue, the community could work to shift the responsibility of data sharing to investigators. Such a shift in responsibility could be spurred on by journal publishers and funding organizations requiring investigators to add their data to specified public repositories. In some instances, data could be added to repositories prior to publication during the data collection process in a non-public format, which could easily be made public later in an article. Entering data into repositories prior to publication could help reduce errors (i.e., minimize forgotten protocol details) and expedite the time to publication by avoiding the need to enter all the data at once, after completion of the study. If the repositories available for nanomaterial data develop methods to facilitate interoperability, then investigators could share their data with multiple stakeholder groups by entering information in a standardized format and ontology in one repository. This idealized scenario will of course take time to realize, but will only become possible through collaborative work in the nanomaterial community to support nanoinformatics. Some of that collaborative work might include the steps discussed below related to the second opportunity area: expanding tools and repositories.

Individuals and organizations in the nanomaterial community could consider mechanisms to enhance resources for development work on the ISA-TAB-Nano data-sharing tool and associated tools (e.g., time, opportunities for user community discussions, budgetary support). Development projects could focus on improving usability of the tool, automating some of the functions, and building data-entry interfaces. Resources for this work will be critical to support continued use of the tools, but to expand use of ISA-TAB-Nano and related tools, the community would benefit from opportunities for training. For example, a series of facilitated web-conferences (e.g., WebEx) or in-person workshops could provide valuable insight to new users. Resources for similar events that focus on more established users could support dialogue between data curators and ISA-TAB-Nano designers so that the tool continues to evolve in

ways most useful to the user community. These discussions could also identify opportunities for workflow standardization across data repositories, as well as identify additional topic areas that would benefit from open dialogues in the nanocuration community. For instance, community users might discuss how natural language processing or other automated approaches might facilitate bringing data into repositories through ISA-TAB-Nano [10].

Recommendations proposed here have been based on the current landscape of the nanoinformatics field, and are focused on potential best practices to catalyze progress given the existence of multiple repositories and resources emerging from a variety of independently funded efforts representing diverse missions. It is not expected that a single unified resource for nanomaterial data analysis would ever be practical or particularly useful, given the established need for different levels of detail, data domains, and functionalities based on the driving purpose of the resource [1]. However, it may well be that some streamlining and optimization would be beneficial as the field matures, such that resources that have developed independently but that share similar analytical purposes, target communities, or sufficient CDEs might be merged into common resources to maximize effectiveness and sustainability.

Conclusion

The curation workflow provides a means not only to share data through nanoinformatics, but also to communicate underlying assumptions about the data within and between organizations. The development and implementation of an explicit workflow process for nanocuration not only plays a role in building a single data repository, but also in providing information about standardization, common bottlenecks, and leverage points that can benefit the community as a whole. Current repositories and tools for sharing data provide a strong foundation for implementation of existing workflows such as those discussed above; however, progress in expanding the development and use of nanocuration workflows would benefit from efforts across the scientific community to address the myriad of challenges that face the implementation of nanocuration workflows (e.g., incomplete data in publications, funding for data sharing tools, use of standardized ontology). We welcome input from the nanomaterial community on the potential next steps to overcome the challenges laid out in this article, and encourage continued input as the effort moves forward. Interested community members can share feedback or join the National Cancer Informatics Program (NCIP) Nanotechnology Working Group by visiting https://nciphub.org/groups/nanowg/overview, and can learn more about the Nanomaterial Data Curation Initiative, in particular, by visiting https://nciphub.org/groups/nanotechnologydatacurationinterestgroup/wiki/MainPage.

Supporting Information

Supporting Information contains all stakeholder responses that are summarized in Section iv (Current practice for nanocuration workflows: Stakeholder responses to questions) and Figures 2–5.

Supporting Information File 1

Stakeholder responses to Nanomaterials Data Curation Initiative (NDCI) questions regarding current nanocuration workflow practices (Note that respondents 5–7 are also authors on this article).
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-189-S1.pdf]

Acknowledgements

Authors are grateful to Mervi Heiskanen (NIH/NCI) for her time and technical expertise to provide tools that supported collaboration on this article. C.O.H. would like to gratefully acknowledge the Center for the Environmental Implications of NanoTechnology (CEINT) funding from National Science Foundation (NSF) and the Environmental Protection Agency (EPA) under NSF Cooperative Agreement DBI-1266252 and EF-0830093. RTI International, developers of the Nanomaterial Registry, would like to thank the National Institutes of Health (NIH) for funding their work, under contract HHSN268201000022C. The views, opinions, and content in this article are those of the authors and do not necessarily represent the views, opinions, or policies of their respective employers or organizations Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. government.

References

1. Hendren, C. O.; Powers, C. M.; Hoover, M. D.; Harper, S. L. *Beilstein J. Nanotechnol.* **2015,** *6,* 1752–1762. doi:10.3762/bjnano.6.179
2. Downs, R. R.; Chen, R. S. *Earth Sci. Inf.* **2010,** *3,* 101–110. doi:10.1007/s12145-010-0051-6
3. de la Iglesia, D.; Cachau, R. E.; García-Remesal, M.; Maojo, V. *Comput. Sci. Discovery* **2013,** *6,* 014011. doi:10.1088/1749-4699/6/1/014011
4. Harper, S. L.; Hutchison, J. E.; Baker, N.; Ostraat, M.; Tinkle, S.; Steevens, J.; Hoover, M. D.; Adamick, J.; Rajan, K.; Gaheen, S.; Cohen, Y.; Nel, A.; Cachau, R. E.; Tuominen, M. *Comput. Sci. Discovery* **2013,** *6,* 014008. doi:10.1088/1749-4699/6/1/014008
5. Gaheen, S.; Hinkal, G. W.; Morris, S. A.; Lijowski, M.; Heiskanen, M.; Klemm, J. D. *Comput. Sci. Discovery* **2013,** *6,* 014010. doi:10.1088/1749-4699/6/1/014010
6. Thomas, D. G.; Pappu, R. V.; Baker, N. A. *J. Biomed. Inf.* **2011,** *44,* 59–74. doi:10.1016/j.jbi.2010.03.001

7. Mills, K. C.; Murry, D.; Guzan, K. A.; Ostraat, M. L. *J. Nanopart. Res.* **2014,** *16,* 1–9. doi:10.1007/s11051-013-2219-8
8. Thomas, D. G.; Gaheen, S.; Harper, S. L.; Fritts, M.; Klaessig, F.; Hahn-Dantona, E.; Paik, D.; Pan, S.; Stafford, G. A.; Freund, E. T.; Klemm, J. D.; Baker, N. A. *BMC Biotechnol.* **2013,** *13,* 2. doi:10.1186/1472-6750-13-2
9. Shumway, M.; Cochrane, G.; Sugawara, H. *Nucleic Acids Res.* **2009,** *38,* D870–D871. doi:10.1093/nar/gkp1078
10. García-Remesal, M.; García-Ruiz, A.; Pérez-Rey, D.; de la Iglesia, D.; Maojo, V. *BioMed Res. Int.* **2013,** *2013,* 410294. doi:10.1155/2013/410294

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Framework for automatic information extraction from research papers on nanocrystal devices

Thaer M. Dieb*[1], Masaharu Yoshioka[1], Shinjiro Hara[2] and Marcus C. Newton[3]

Full Research Paper

Open Access

Address:
[1]Graduate School of Information Science and Technology, Hokkaido University, Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido, 060-0814, Japan, [2]Research Center for Integrated Quantum Electronics, Hokkaido University, Kita 13, Nishi 8, Sapporo 060-8628, Japan and [3]Physics & Astronomy, University of Southampton, Southampton, SO17 1BJ, UK

Email:
Thaer M. Dieb* - diebt@kb.ist.hokudai.ac.jp

* Corresponding author

Keywords:
annotated corpus; automatic information extraction; nanocrystal device development; nanoinformatics; text mining

Beilstein J. Nanotechnol. **2015,** *6,* 1872–1882.
doi:10.3762/bjnano.6.190

Received: 31 March 2015
Accepted: 20 August 2015
Published: 07 September 2015

This article is part of the Thematic Series "Nanoinformatics for environmental health and biomedicine".

Guest Editor: R. Liu

Abstract

To support nanocrystal device development, we have been working on a computational framework to utilize information in research papers on nanocrystal devices. We developed an annotated corpus called " NaDev" (*Na*nocrystal *Dev*ice Development) for this purpose. We also proposed an automatic information extraction system called "NaDevEx" (*Na*nocrystal *Dev*ice Automatic Information *Ex*traction Framework). NaDevEx aims at extracting information from research papers on nanocrystal devices using the NaDev corpus and machine-learning techniques. However, the characteristics of NaDevEx were not examined in detail. In this paper, we conduct system evaluation experiments for NaDevEx using the NaDev corpus. We discuss three main issues: system performance, compared with human annotators; the effect of paper type (synthesis or characterization) on system performance; and the effects of domain knowledge features (e.g., a chemical named entity recognition system and list of names of physical quantities) on system performance. We found that overall system performance was 89% in precision and 69% in recall. If we consider identification of terms that intersect with correct terms for the same information category as the correct identification, i.e., loose agreement (in many cases, we can find that appropriate head nouns such as temperature or pressure loosely match between two terms), the overall performance is 95% in precision and 74% in recall. The system performance is almost comparable with results of human annotators for information categories with rich domain knowledge information (source material). However, for other information categories, given the relatively large number of terms that exist only in one paper, recall of individual information categories is not high (39–73%); however, precision is better (75–97%). The average performance for synthesis papers is better than that for characterization papers because of the lack of training examples for characterization papers. Based on these results, we discuss future research plans for improving the performance of the system.

Introduction

Nanoscale research is a rapidly progressing domain and many research papers containing experimental results have been published. Because it is a very time-consuming task to read through all related papers, several research efforts have been conducted in the nanoinformatics research domain. This includes the construction of databases for sharing the experimental results [1-5], and the set-up of portals for sharing useful information [6-12]. Those approaches try to support data collection processes based on human efforts. It is desirable to have a framework to support information extraction from research papers. This approach is widely used in other research domains. For example, the GENIA corpus [13] was constructed to extract biology-related information (e.g., genome, protein) and the BioCreative IV CHEMDNER corpus [14] was created to extract chemical and drug names. Based on such corpora, several researchers have proposed a variety of methods for the extraction of information from research papers [15-17]. In the nanoinformatics domain, only a few researchers have attempted to automatically extract information from research papers [18-20] and their frameworks are explicitly focused on nanomedicine applications.

Nanocrystal device development [21-26] is an important area of nanoscale research. To support analysis of experimental results in this domain, extracting experimental information from related publications is desirable. We previously constructed an annotated corpus called "NaDev" (*Na*nocrystal *Dev*ice Development corpus) [27,28] for research papers on nanocrystal device development. We also proposed a framework to extract information from research papers by using machine learning tools [29,30]. However, this system was only evaluated using the corpus constructed in our preliminary experiment, which was not sufficient to compare automatic information extraction results with those from human annotators. In addition, in the discussion of constructing NaDev corpus, we found that the paper type (i.e., synthesis or characterization) affected the style of writing, so the information extraction quality varied according to paper type.

In this paper, we propose a framework for automatic information extraction, NaDevEx (*Na*nocrystal *Dev*ice Automatic Information *Ex*traction Framework) from research papers on nanocrystal devices and evaluate the system using the NaDev corpus. Furthermore, we discuss the quality of automatic information extraction compared with that from human annotators and conduct a failure analysis to identify future research issues. In this analysis, we compare the results for synthesis papers with the results for characterization papers to better understand the effect of the type of paper on the system performance.

Before discussing our automatic information extraction experiments using NaDev, we briefly review previous studies on extracting useful information from research papers in other domains and introduce our proposed system for automatic information extraction.

Utilizing information in research papers using text-mining techniques is an increasingly important trend in several domains. In bioinformatics for example, several frameworks for automatic extraction of biomedical entities from research papers have been proposed [15,16]. In the chemical information domain, different approaches compete to extract chemical entities and drug names automatically from the literature [17] using the BioCreative IV CHEMDNER corpus [14]. We can classify approaches to information extraction and named entity recognition into two groups. One is a machine-learning approach that uses a domain corpus, such as GENIA, to find typical patterns for explaining useful terms. The other is a rule-based system that uses rules to extract useful terms (e.g., use a list of chemical symbols to identify chemical compounds). Many recent systems have used a combination of both approaches.

For extracting information from nanocrystal device papers, we have proposed an automatic information extraction framework [29] using machine learning techniques. This approach tries to extract information step-by-step. We call this step-by-step extraction "cascading style extraction" [31].

A preliminary performance check of the automatic information extraction system using the corpus developed for the preliminary experiment confirmed the appropriateness of the general framework. However, the characteristics of NaDevEx were not fully examined. In this paper, we conduct system evaluation experiments for NaDevEx using the NaDev corpus and analyze system performance compared with human annotators' results. We also discuss plans for future research based on this analysis.

Materials and Methods

NaDev corpus

The NaDev corpus [27,28] was constructed to identify experimental information for extraction from nanocrystal device development papers. In order to extract wide varieties of experimental information, NaDev corpus uses full text of research papers instead of abstracts that are commonly used for constructing such corpora. Abstracts usually do not contain detailed explanation about experimental parameters in relation with output evaluation. It is necessary to extract such information to analyze experimental results adequately. In this corpus, eight information categories are annotated as useful informa-

tion in papers related to nanocrystal device development. These information categories are defined as below:

- Source material (SMaterial): Material used as input in the experiment, such as InGaAs.
- Material characteristic feature (MChar): Characteristic feature of the materials, such as hexagonal. Such feature might be a result of manufacturing process or is a characteristic feature of source material.
- Experimental parameter (ExP): Parameter for controlling experiment's conditions, such as diameter or total pressure.
- Experimental parameter value (ExPVal): Value of an experimental parameter, such as 50 nm or 10 atoms.
- Evaluation parameter (EvP): Parameter that is used to evaluate the output of the experiment, such as peak energy.
- Evaluation parameter value (EvPVal): Value of an evaluation parameter, such as 1.22 eV.
- Manufacturing method (MMethod): Method used in the experiment to achieve the desired product, such as selective-area metalorganic vapor-phase epitaxy.
- Target artifact or final product (TArtifact): Final output of the experiment, such as nanowires.

The NaDev corpus has 392 sentences. 2870 terms are annotated using these information categories. Figure 1 shows a sample of the corpus. Table 1 shows the number of categorized terms in NaDev corpus.

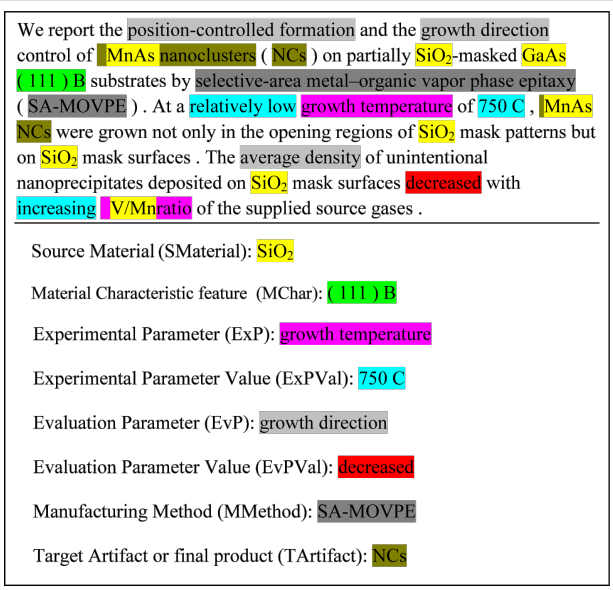


Figure 1: Sample of NaDev corpus.

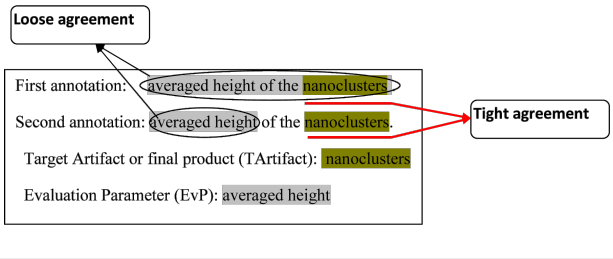


Figure 2: Example of tight and loose agreement.

Corpus construction

The corpus construction guideline [27] was prepared in collaboration with a domain expert in nanocrystal device development by using the results of the annotation experiments by domain graduate students. In each experiment, two graduate students were asked to annotate the same paper independently. Annotated results were compared to check the reliability of the guideline. We used kappa coefficient to test inter-annotator agreement (IAA) [32]. Two metrics were used for the analysis: tight agreement, which considers the term boundary and term category to decide the agreement; and loose agreement, which ignores the term boundary, i.e., when a term overlaps with a correct term of the same information category, we treat it as correct (see Figure 2 for an example).

For the inter-annotator mismatch cases, we had meetings for discussing these cases with the annotators, and collected adequate annotation examples for further reference. Inter annotator mismatches, in most cases occurred due to the difficulty to set correct boundaries of the term, specially, in the EvPVal and ExP information categories.

Corpus evaluation

Even though the corpus construction guideline reached a reliable level with loose agreement [29], it was necessary to evaluate this corpus and finalize it with a domain expert researcher to ensure reliability. We classified the annotations of graduate students into agreed and disagreed annotations. Careless

Table 1: Number of categorized terms in NaDev corpus.

Information category	SMaterial	MMethod	MChar	TArtifact	ExP	EvP	ExPVal	EvPVal	Total
terms	780	136	381	416	262	365	234	296	2870
of total	27%	5%	13%	15%	9%	13%	8%	10%	

mistakes, such as one annotator missed to add an annotation, or typical types of disagreement when annotators misunderstood the guideline, were easily checked in the discussion after each annotation experiment, so they were considered to be agreed annotations.

To improve the consistency of the annotation and to overcome problems found by examining the corpus, the domain expert proposed few modifications to the corpus-construction guideline.

With the revision of the domain expert, we found the corpus contains two types of papers depending on the content and the writing style. Four of the papers focus on the synthesis of new nanomaterials [33-36], and the other focuses on the characterization of nanomaterials [37]. We have made a finalized version of the five papers of the corpus based on the revision of the domain expert. To evaluate the annotation reliability of the graduate students, we compared this finalized version with the original corpus constructed before the evaluation experiment. Evaluation showed that, if we exclude the effect of the guideline modifications made by the domain expert, for synthesis papers, the agreed annotation results obtained through discussion after the annotation experiments have high precision for all information categories (ranging between 96% and 100%). Discussion between annotators after the annotation process is important, because it can resolve mismatches caused by careless mistakes or misunderstanding of the guideline. Recall is also high (ranging between 91% and 100%). For the characterization paper, the precision is high (ranging between 94% and 100%), but the recall is low because of the larger number of disagreed annotations in this case. The lack of deep domain knowledge of the students for the characterization paper seems to have had a considerable effect on the quality of the annotation.

We concluded generally that information categories such as SMaterial, MMethod, and ExPVal tend to be easier to annotate. Conversely, information categories such as the parameters ExP, and EvP, and EvPVal tend to be more difficult to annotate, requiring deeper domain knowledge, particularly for the characterization paper. Most of the disagreed annotations in these categories resulted from difficulties in setting correct boundaries for these information categories.

Automatic information extraction

Our information extraction system uses a cascading style extraction based on machine learning. For example, chemical named entities are useful for identifying source materials (e.g., As), and identification of source material is useful for identifying term boundaries of experimental parameters (e.g., pressure of AsH_3 gas). The order of information categories for extraction was designed by using the overlapping structure between information categories. For example, for experimental parameters and source materials (e.g., pressure of AsH_3 gas), the extraction of source material should be prior to extraction of experimental parameters. Figure 3 shows a procedure to extract these information categories step-by-step.

First, linguistic features such as part-of-speech (POS) tags, orthogonal features, and lemmatization features are generated using the results from a morphological analysis tool [38]. Second, we use domain knowledge tools (i.e., the output of a chemical named entity recognition tool [29], matching results from a physical quantities vocabulary list, and a list of common measurement units [30]) to generate domain knowledge-related features (CNER, PAR, and UNT, respectively). For the latter step, we used CRF++ [39], an implementation of conditional random field (CRF) [40] as a machine learning system that uses part of the corpus as training data for information extraction. In each step, we use all the features generated by the tools, including linguistic features and domain knowledge-related features.

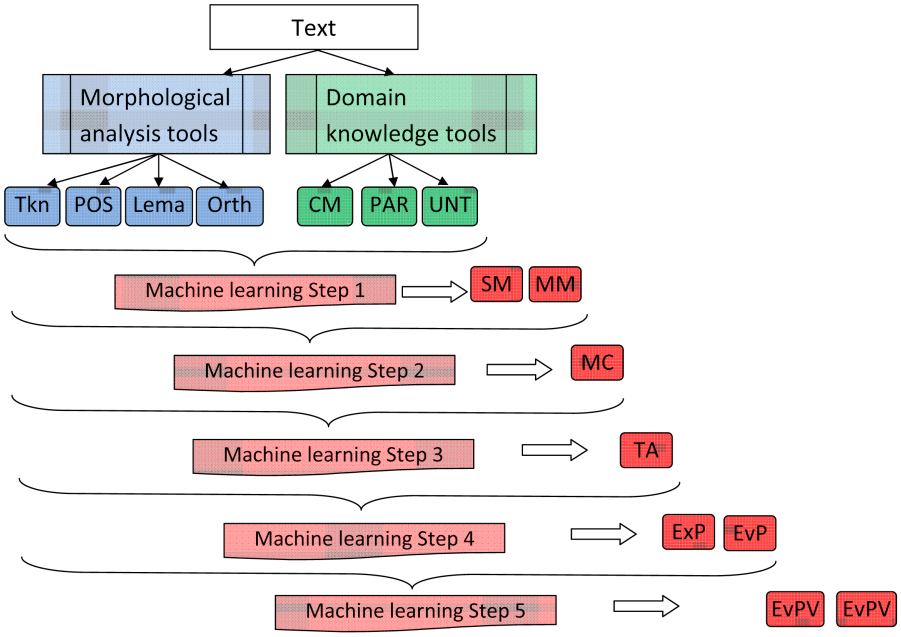
Results and Discussion

System implementation

The NaDevEx system accepts plain text as input and adds annotations to the terms in the text that belong to the information categories defined in the NaDev corpus construction guideline.

Information about the most recent version of the system, which was used for these experiments, is as follows.

- Linguistic features: GPostLL tagger (ver. 0.9.3) [38].
- An orthogonal feature was added using regular expressions based on the definition in [15].
- Domain knowledge-based features: (i) A chemical named entity feature was added using SERB-CNER (Syntactically Enhanced Rule-Based Chemical Named Entity Recognition System) that we developed to annotate chemical entities in nanocrystal device papers. (ii) A parameter identification feature was added based on a list of physical quantities: we compiled a list that contains physical properties of matter (e.g., density, concentration), common parameters found in nanocrystal device papers (e.g., height, conductivity), and several keywords that usually correlate with parameters (e.g., ratio, rate). The list was checked by nanocrystal device researchers as a basic list for physical quantities. (iii) A parameter value identification feature was added based on a list of common measurement units.
- CRF tool: CRF++ (ver.0.58)



Legend: Tkn: token, POS: part of speech, Lema: lemmatization, Orth: orthogonal, CM: chemical named entity, PAR: physical quantity matching, UNT: measurement unit list matching, SM: SMaterial, MM: MMethod, MC: MChar, TA: TArtifact, ExP: ExP, EvP: EvP, ExPV: ExPVal, and EvPV: EvPVal

Figure 3: Outline of our automatic information extraction system.

The input for the CRF++ tool is in IOB format, which identifies the position (beginning, inside, out of) of a token of text related to a term. Figure 4 shows an example of input data for the CRF++ tool.

For the training, NaDevEx first added linguistic features and results of the domain knowledge-based systems to the original texts. Then information about correct annotations was used to train the machine learning system CRF++ in cascading style. For the information extraction, the system used the same tools to add linguistic features and results of domain knowledge and used the learning results of CRF++ in cascading style to generate the final answer.

Experiment plan

In this paper, we evaluate our automatic information extraction system (NaDevEx) and discuss the characteristics of this system by using the NaDev corpus. We design an experiment plan to address the following three main issues:

- system performance analysis compared with human annotators

Tkn	POS	Lema	Orth	CM/PAR/UNT	SM/MM	MC	TA	ExP/EvP	ExPV/EvPV
MnAs	NP	mnas	TwoCaps	B-CM	B-SMaterial	O	B-TArtifact	O	O
thin	JJ	thin	Lowercase	O	O	O	I-TArtifact	O	O
films	NNS	film	Lowercase	O	O	O	I-TArtifact	O	O
have	VHP	have	Lowercase	O	O	O	O	O	O
shown	VVN	show	Lowercase	O	O	O	O	O	O
ferromagnetic	JJ	ferromagnetic	Lowercase	O	O	O	O	B-EvP	O

Legend: Tkn: token, POS: part of speech, Lema: lemmatization, Orth: orthogonal, CM: chemical named entity, PAR: physical quantity matching, UNT: measurement unit list matching, SM: SMaterial, MM: MMethod, MC: MChar, TA: TArtifact, ExP: ExP, EvP: EvP, ExPV: ExPVal, and EvPV: EvPVal

Figure 4: Example of CRF++ input data.

- system performance analysis for each type of corpus paper (synthesis or characterization)
- effect of domain knowledge features on system performance

System performance analysis compared with human annotators

We evaluated our system performance using the NaDev corpus. We used five-fold cross validation and calculated precision, recall, and F-score. In each fold, we trained the system using four of the five papers as training data and evaluated its performance using the fifth paper. Because NaDev gold standards are based on the annotation of the domain expert, those results represent the comparison between NaDevEx performance and the annotation of the domain expert. Because NaDevEx is built using machine-learning techniques, deep domain knowledge is difficult to acquire using NaDevEx. Therefore, we contrast NaDevEx performance with that based on agreement between two novice annotators, as discussed previously. These comparison results represent the ideal level of annotation without deep domain knowledge.

Table 2 contrasts the average performance for each information category between NaDevEx and the human annotation results compared with the annotation of the domain expert. Underlining indicates that the difference between NaDevEx performance and the human annotation results is statistically insignificant at the 5% level ($P \geq 0.05$). The human annotations were made prior to the released version of the guideline [27]. Recall of categories that were subject to new definitions (SMaterial and MChar) is underestimated. If we assume that all the new added annotations based on the released guideline were identified by human annotators, recall of SMaterial and MChar is increased to 0.99 and 0.93, respectively.

From Table 2, the performance of NaDevEx on the SMaterial category is almost comparable with human annotation. For MMethod, MChar, and ExP, performance is comparatively good for precision but not so good for recall. For the other categories, the system performance is not so good for precision and worse for recall. Based on the nature of the machine-learning system, it is easier to extract the terms that appear in the training data than ones that are unique in the test data. However, if there are similar terms (e.g., a term that overlap with one in the training data or terms used in similar context) in the training data, the system can extract such terms.

There are several cases that show the term boundary identification problem, especially for unique compound terms. To check the effect of such problems, we used the loose agreement metric as illustrated in Figure 2.

For human annotators, even though there were many cases of loose agreement between the two annotators, discussion after annotation experiments generally resolved these boundary mismatch issues. Table 3 contrasts the average performance for each information category for NaDevEx and the human annotation results for loose agreement compared with the annotation of the domain expert. Underlining indicates that the difference between NaDevEx performance and the human annotation results is statistically insignificant at the 5% level ($P \geq 0.05$).

The differences between the evaluation results of Table 2 and Table 3 reflect the difficulty of identifying term boundaries. For NaDevEx, performance for loose agreement improves for all information categories in precision and recall, especially for TArtifact, EvP, ExPVal, and EvPVal. This shows that these categories have many problems related to identifying term boundaries. If we accept loose agreement as correct (in most cases we can find appropriate head nouns such as temperature, or pressure in loose matching terms), TArtifact and EvPVal also become almost comparable with human annotation for precision.

Table 2: Average performance of NaDevEx and the human annotation results compared with the annotation of the domain expert.

	human			NaDevEx		
	precision	recall	F-score	precision	recall	F-score
SMaterial	0.97	0.79	0.87	0.95	0.94	0.94
MMethod	1.00	0.91	0.95	0.97	0.73	0.82
MChar	0.93	0.84	0.88	0.94	0.67	0.75
TArtifact	0.99	0.90	0.94	0.88	0.73	0.80
ExP	1.00	0.91	0.94	0.93	0.68	0.76
EvP	0.98	0.91	0.94	0.78	0.55	0.64
ExPVal	0.99	0.97	0.98	0.80	0.53	0.64
EvPVal	1.00	0.86	0.92	0.75	0.39	0.51
Total	0.98	0.86	0.91	0.89	0.69	0.77

Table 3: Average performance of NaDevEx and the human annotation results for loose agreement compared with the annotation of the domain expert.

	human			NaDevEx		
	precision	recall	F-score	precision	recall	F-score
SMaterial	0.99	0.81	0.89	0.98	0.97	0.97
MMethod	1.00	0.91	0.95	0.98	0.73	0.83
MChar	0.94	0.85	0.89	0.96	0.68	0.77
TArtifact	1.00	0.90	0.95	0.96	0.79	0.86
ExP	1.00	0.91	0.95	0.97	0.71	0.79
EvP	0.99	0.92	0.95	0.86	0.60	0.71
ExPVal	1.00	0.97	0.99	0.92	0.62	0.74
EvPVal	1.00	0.86	0.92	0.88	0.46	0.60
Total	0.99	0.87	0.92	0.95	0.74	0.83

In general, Table 2 and Table 3 show that NaDevEx has problems in identifying term boundaries in categories where human annotators have the same difficulty. However, discussion between the annotators after each annotation experiment helped to reduce these difficulties.

In addition, recall of the categories MChar, ExP, EvP, ExPVal, and EvPVal is comparatively worse than that made by the human agreement. For these categories, there are varieties of compound terms that usually contain characteristic technical terms within their boundaries. However, because of the variability in using these technical terms for constructing compound terms, NaDevEx cannot extract such terms appropriately. We discuss this issue in detail in the section “Effect of domain knowledge features on system performance”.

System performance analysis based on type of paper

System performance differs between synthesis papers and characterization papers. Table 4 shows the average performance of NaDevEx for four synthesis papers and one characterization paper including loose agreement cases using five-fold cross validation.

One reason for the lower performance with the characterization paper is a lack of examples of sentences and terms that are frequently used in characterization papers and not in synthesis papers. To discuss this effect, we conducted a 10-fold cross validation that uses four papers and half of the fifth paper as training data, evaluated on the other half of the fifth paper. Table 5 shows the average performance of NaDevEx on four synthesis papers and one characterization paper using 10-fold cross validation including loose agreement.

In this case, because we can use one-half of a paper as training data, the number of terms that are unique to the test data decreased. The performance for 10-fold cross validation is slightly better than that for five-fold cross validation. However, in total, the increased ratio for characterization with loose recall was slightly better than that for synthesis papers.

Table 4: NaDevEx average performance on synthesis and characterization papers using five-fold cross validation.[a]

	average synthesis papers						characterization paper					
	prec	rec	F	L-prec	L-rec	F	prec	rec	F	L-prec	L-rec	F
SMaterial	0.95	0.94	0.94	0.98	0.97	0.97	0.93	0.96	0.95	0.96	0.99	0.97
MMethod	0.97	0.75	0.84	0.98	0.76	0.85	1.00	0.63	0.77	1.00	0.63	0.77
MChar	0.94	0.78	0.85	0.96	0.79	0.86	0.92	0.22	0.36	1.00	0.24	0.39
TArtifact	0.93	0.79	0.85	0.95	0.81	0.87	0.69	0.49	0.57	1.00	0.71	0.83
ExP	0.91	0.77	0.83	0.96	0.81	0.87	1.00	0.31	0.48	1.00	0.31	0.48
EvP	0.80	0.57	0.66	0.88	0.62	0.73	0.73	0.48	0.58	0.77	0.51	0.61
ExPVal	0.81	0.57	0.66	0.95	0.67	0.78	0.76	0.41	0.53	0.82	0.44	0.57
EvPVal	0.74	0.41	0.53	0.87	0.48	0.62	0.79	0.33	0.46	0.90	0.37	0.53
Total	0.90	0.75	0.82	0.95	0.79	0.86	0.82	0.47	0.60	0.93	0.53	0.68

[a]prec: precision, rec: recall, L-prec: loose precision, L-rec: loose recall, F: F-score

Table 5: NaDevEx average performance on synthesis and characterization papers using 10-fold cross validation.[a]

	average synthesis papers						average characterization paper					
	prec	rec	F	L-prec	L-rec	F	prec	rec	F	L-prec	L-rec	F
SMaterial	0.95	0.94	0.94	0.98	0.97	0.97	0.96	0.97	0.96	0.97	0.99	0.98
MMethod	0.96	0.81	0.87	0.96	0.81	0.87	1.00	0.63	0.77	1.00	0.63	0.77
MChar	0.95	0.83	0.89	0.97	0.84	0.90	0.84	0.35	0.46	0.87	0.37	0.49
TArtifact	0.95	0.85	0.90	0.96	0.87	0.91	0.71	0.53	0.61	0.98	0.75	0.85
ExP	0.93	0.81	0.86	0.98	0.86	0.91	0.59	0.33	0.42	0.88	0.46	0.61
EvP	0.80	0.63	0.70	0.88	0.69	0.77	0.77	0.47	0.58	0.87	0.53	0.66
ExPVal	0.81	0.67	0.73	0.93	0.77	0.83	0.69	0.46	0.55	0.78	0.51	0.61
EvPVal	0.75	0.48	0.58	0.88	0.56	0.68	0.78	0.35	0.48	0.93	0.41	0.57
Total	0.91	0.79	0.84	0.96	0.83	0.89	0.80	0.51	0.62	0.93	0.59	0.72

[a]prec: precision, rec: recall, L-prec: loose precision, L-rec: loose recall, F: F-score

Effect of domain knowledge features on system performance

As we have already discussed, it is difficult for the machine learning system to find terms that are unique to the test data. Table 6 shows the number of unique terms in each paper and the system performance for extracting such terms.

For SMaterial, even though there are many terms that are unique to the test data, the system can identify such terms with a considerably higher coverage ratio than is obtained for other information categories. In most cases, those terms are identified as Chemical Named Entities and the system can generalize the training data by using the information that has been

Table 6: Unique term analysis for each paper.[a]

	synthesis papers								
	paper 1			paper 2			paper 3		
	uniq	extracted	coverage	uniq	extracted	coverage	uniq	extracted	coverage
SMaterial	15	8	0.53	6	5	0.83	16	10	0.63
MMethod	0	0	NA	0	0	NA	14	4	0.29
MChar	6	2	0.33	23	7	0.30	25	14	0.56
TArtifact	11	3	0.27	12	4	0.33	17	9	0.53
ExP	8	5	0.63	10	0	0.00	7	3	0.43
EvP	11	3	0.27	27	2	0.07	21	4	0.19
ExPVal	26	10	0.38	13	5	0.38	20	6	0.30
EvPVal	29	13	0.45	33	10	0.30	39	15	0.38
Total	106	44	0.42	124	33	0.27	159	65	0.41

	synthesis paper			characterization paper			
	paper 4			paper 5			corpus average coverage
	uniq	extracted	coverage	uniq	extracted	coverage	
SMaterial	12	0	0.00	7	6	0.86	0.57
MMethod	10	2	0.20	7	2	0.29	NA
MChar	10	1	0.10	68	3	0.04	0.27
TArtifact	13	2	0.15	46	4	0.09	0.28
ExP	11	1	0.09	22	0	0.00	0.23
EvP	52	11	0.21	49	17	0.35	0.22
ExPVal	38	11	0.29	23	8	0.35	0.34
EvPVal	44	10	0.23	52	9	0.17	0.31
Total	190	38	0.20	274	49	0.18	0.29

[a]uniq: number of unique terms in each paper; extracted: number of terms identified by NaDevEx; coverage: coverage percentage of unique terms identified.

provided by the CNER tool, discussed earlier. For the parameters ExP and EvP, precision is good when the system can use parameter list to identify parameter-related terms. However, because of the insufficient coverage of parameter-related terms used in nanocrystal device development, recall of these parameters is worse than the results of human annotators.

These results show that preprocessing annotation based on domain knowledge is generally promising, but coverage of the parameter information based on a list of physical quantities is not enough for nanocrystal device papers. As we have already discussed in the section "System performance analysis compared with human annotators", there are many compound terms that contain particular domain-specific terms within their boundaries for characterizing categories. Figure 5 shows an example of such domain-specific terms.

Human annotators might be able to recognize such domain-specific terms with their domain knowledge. However, NaDevEx lacks such ability, specially with small training examples. It is necessary to evaluate the effectiveness of such a list by using a larger corpus.

Discussion

The performance of NaDevEx is good for precision (95% for loose agreement overall), but is not good for recall (74% for loose agreement in total) at present. For the information category with rich domain-knowledge information (SMaterial), our system performance is almost comparable with that of human annotators. The precision of the system output is generally high: it is good (more than 95%) for MMethod, MChar, TArtifact and ExP but modest (more than 85%) for other categories (EvP, ExPVal, and EvPVal) with loose agreement. In contrast, the recall of the system is low (46–73%), even with loose agreement.

It is necessary to take into account the effect of the corpus size. As we discussed in Table 6, it is difficult to extract unique terms that do not exist in the training data (percentage of the unique terms among total terms is almost 30% (853/2870)). It is better to check the percentage of the unique terms among total terms when the size of the corpus increases. On the contrary, identification of non-unique terms is comparatively easier for such a small size corpus.

There are two possible research approaches to increase recall of the system output. One approach is to increase the corpus size. It is good to use one whole paper for clear understanding of the role of the terms in the paper, but the varieties of terms are not greatly increased because of the repetitive mention of terms. For the next step, it may be better to construct an abstract-based corpus to increase the variety of terms. It is also preferable to have a balanced mixture of synthesis and characterization papers. Another approach is to construct resources for representing domain knowledge. A list of terms that are frequently used in nanocrystal device papers is helpful to extract related terms that are in the list and variations of the terms based on the head terms in the list. There are physical parameters that cannot be extracted using the general physical quantities list (e.g., lattice, (111)B surface), so it is better to use vocabulary lists that include the parameters in this domain.

NaDevEx can be used as a preprocessor to find research papers that contain recent analysis results on nanocrystal devices to support the data collection process. Because NaDevEx is good at identifying source material, we can construct appropriate queries to restrict the output to papers that discuss a particular type of source material. Usage of other information categories may work well for finding related papers in a precision oriented manner, but it may miss papers because of the bad recall performance. A possible solution to this problem is implementing a

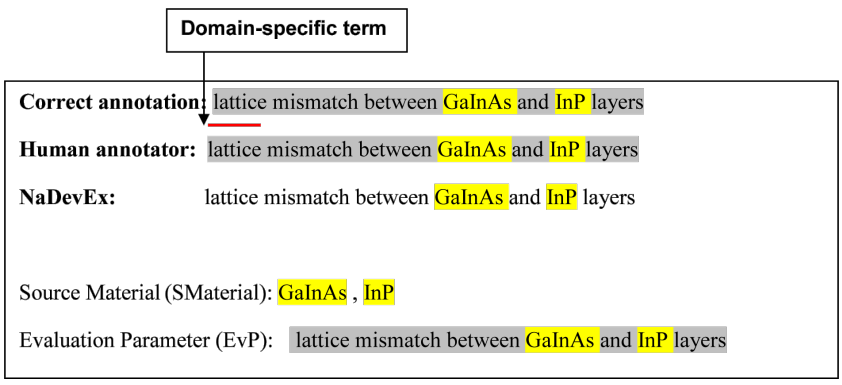


Figure 5: Domain-specific terms in NaDev corpus.

framework that utilizes user-defined keyword lists as a knowledge resource for extracting such information. Another is using simple keyword search to find more papers that may contain such information.

Conclusion

In this paper, we introduce NaDevEx, which automatically extracts useful information from nanocrystal device research papers based on the information categories defined in the NaDev corpus. This system has almost comparable performance with the human annotators for source material information, because of the good performance of the chemical named entity recognition system. For other categories, the precision is good (better than 85% in case of loose agreement), but there is a problem with recall because of the lack of examples, especially for characterization papers. To improve the performance, we discuss future research plans: increasing the corpus size by using abstract texts and constructing resources for representing domain knowledge (e.g., lists of parameters and manufacturing methods).

Acknowledgements

This research was supported partially by a grant to the Hokkaido University Global COE program, “Next-Generation Information Technology Based on Knowledge Discovery and Knowledge Federation”, from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, and JSPS KAKENHI Grant Number 2654011 by Japan Society for the Promotion of Science. We would also like to thank Prof. Takeuchi (Okayama University) and Prof. Kano (Japan Science and Technology Agency) for their discussion and comments about cascading named entity recognition.

References

1. Kozaki, K.; Kitamura, Y.; Mizoguchi, R. Systematization of nanotechnology knowledge through ontology engineering - A trial development of idea creation support system for materials design based on functional ontology. In *Poster notes of ISWC2003,* Sanibel Island, FL, U.S.A.; 2003; pp 63–64.
2. Thomas, D. G.; Pappu, R. V.; Baker, N. A. *J. Biomed. Inf.* **2011,** *44,* 59–74. doi:10.1016/j.jbi.2010.03.001
3. DaNa project. http://www.nanoobjects.info/en/ (accessed July 6, 2015).
4. Guzan, K. A.; Mills, K. C.; Gupta, V.; Murry, D.; Scheier, C. N.; Willis, D. A.; Ostraat, M. L. *Comput. Sci. Discovery* **2013,** *6,* 014007. doi:10.1088/1749-4699/6/1/014007
5. Xiao, L.; Tang, K.; Liu, H.; Yang, X.; Chen, Z.; Xu, R. Information extraction from nanotoxicity related publications. In *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* Shanghai, China, Dec 18–21, 2013; pp 25–30.
6. Kimmig, D.; Marquardt, C.; Nau, K.; Schmidt, A.; Dickerhof, M. *Comput. Sci. Discovery* **2014,** *7,* 014001. doi:10.1088/1749-4699/7/1/014001
7. Gaheen, S.; Hinkal, G. W.; Morris, S. A.; Lijowski, M.; Heiskanen, M.; Klemm, J. D. *Comput. Sci. Discovery* **2013,** *6,* 014010. doi:10.1088/1749-4699/6/1/014010
8. Madhavan, K.; Zentner, L.; Farnsworth, V.; Shivarajapura, S.; Zentner, M.; Denny, N.; Klimeck, G. *Nanotechnol. Rev.* **2013,** *2,* 107–117. doi:10.1515/ntrev-2012-0043
9. Integrated Nanoinformatics Platform for Environmental Impact Assessment of Engineered Nanomaterials. http://nanoinfo.org/ (accessed July 6, 2015).
10. Liu, R.; Hassan, T.; Rallo, R.; Yoram, C. *Comput. Sci. Discovery* **2013,** *6,* 014006. doi:10.1088/1749-4699/6/1/014006
11. Harper, S. L.; Hutchison, J. E.; Baker, M.; Ostraat, N.; Tinkle, S.; Steevens, J.; Hoover, M. D.; Adamick, J.; Rajan, K.; Gaheen, S.; Cohen, Y.; Nel, A.; Cachau, R. E.; Tuominen, M. *Comput. Sci. Discovery* **2013,** *6,* 014008.
12. de la Iglesia, D.; Cachau, R. E.; García-Remesal, M.; Maojo, V. *Comput. Sci. Discovery* **2013,** *6,* 01401. doi:10.1088/1749-4699/6/1/014011
13. Kim, J.-D.; Ohta, T.; Tateisi, Y.; Tsujii, J. *Bioinformatics* **2003,** *19* (Suppl. 1), i180–i182. doi:10.1093/bioinformatics/btg1023
14. BioCreative IV CHEMDNER corpus. http://www.biocreative.org/resources/corpora/bc-iv-chemdner-corpus/ (accessed July 6, 2015).
15. Takeuchi, K.; Collier, N. *Artif. Intell. Med.* **2005,** *33,* 125–137. doi:10.1016/j.artmed.2004.07.019
16. Gaizauskas, R.; Demetriou, G.; Artymiuk, P. J.; Willett, P. *Bioinformatics* **2003,** *19,* 135–143. doi:10.1093/bioinformatics/19.1.135
17. Krallinger, M.; Rabal, O.; Leitner, F.; Vazquez, M.; Salgado, D.; Lu, Z.; Leaman, R.; Lu, Y.; Ji, D.; Lowe, D. M.; Sayle, R. A.; Batista-Navarro, R. G.; Rak, R.; Huber, T.; Rocktäschel, T.; Matos, S.; Campos, D.; Tang, B.; Xu, H.; Munkhdalai, T.; Ryu, K. H.; Ramanan, S. V.; Nathan, S.; Žitnik, S.; Bajec, M.; Weber, L.; Irmer, M.; Akhondi, S. A.; Kors, J. A.; Xu, S.; An, X.; Sikdar, U. K.; Ekbal, A.; Yoshioka, M.; Dieb, T. M.; Choi, M.; Verspoor, K.; Khabsa, M.; Giles, C. L.; Liu, H.; Ravikumar, K. E.; Lamurias, A.; Couto, F. M.; Dai, H.-J.; Tsai, R. T.-H.; Ata, C.; Can, T.; Usié, A.; Alves, R.; Segura-Bedmar, I.; Martinez, P.; Oyarzaba, J.; Valencia, A. *J. Cheminf.* **2015,** *7* (Suppl. 1), S2. doi:10.1186/1758-2946-7-S1-S2
18. Jones, D. E.; Igo, S.; Hurdle, J.; Facelli, J. C. *PLoS One* **2014,** *9,* e83932. doi:10.1371/journal.pone.0083932
19. Garcia-Remesal, M.; Garcia-Ruiz, A.; Perez-Rey, D.; De la Iglesia, D.; Maojo, V. *BioMed Res. Int.* **2013,** 410294.
20. de la Iglesia, D.; García-Remesal, M.; Anguita, A.; Muñoz-Mármol, M.; Kulikowski, C.; Maojo, V. *PLoS One* **2014,** *9,* e110331. doi:10.1371/journal.pone.0110331
21. Kriegel, I.; Scotognella, F. *Beilstein J. Nanotechnol.* **2015,** *6,* 193–200. doi:10.3762/bjnano.6.18
22. Davydova, M.; Kulha, P.; Laposa, A.; Hruska, K.; Demo, P.; Kromka, A. *Beilstein J. Nanotechnol.* **2014,** *5,* 2339–2345. doi:10.3762/bjnano.5.243
23. Capan, I.; Carvalho, A.; Coutinho, J. *Beilstein J. Nanotechnol.* **2014,** *5,* 1787–1794. doi:10.3762/bjnano.5.189
24. Yatsui, T.; Morigaki, F.; Kawazoe, T. *Beilstein J. Nanotechnol.* **2014,** *5,* 1767–1773. doi:10.3762/bjnano.5.187
25. Ikejiri, K.; Sato, T.; Yoshida, H.; Hiruma, K.; Motohisa, J.; Hara, S.; Fukui, T. *Nanotechnology* **2008,** *19,* 265604. doi:10.1088/0957-4484/19/26/265604
26. Fukui, T.; Ando, S.; Tokura, Y.; Toriyama, T. *Appl. Phys. Lett.* **1991,** *58,* 2018–2020. doi:10.1063/1.105026

27. Dieb, T.; Yoshioka, M.; Hara, S. NaDev (Nanocrystal Device development) Corpus Annotation Guideline. *TCS Technical Reports, TCS-TR-B-15-12, July 2015;* Hokkaido University, Division of Computer Science: Hokkaido, Japan, 2015.
28. Dieb, T.; Yoshioka, M.; Hara, S. Construction of tagged corpus for Nanodevices development papers. In *Proceedings of International Conference on Granular Computing (GrC),* Kaohsiung, Taiwan; 2011; pp 167–170.
29. Dieb, T.; Yoshioka, M.; Hara, S. Automatic Information Extraction of Experiments from Nanodevices Development Papers. In *Proceedings of International Conference on Advanced Applied Informatics (IIAIAAI),* Fukuoka, Japan, Sept 20–22, 2012; pp 42–47.
30. Dieb, T.; Yoshioka, M.; Hara, S. Automatic Annotation of Parameters from Nanodevice Development Research Papers. In *Proceedings of the 4th International Workshop on Computational Terminology Computerm 2014,* Dublin, Ireland; 2014; pp 77–85.
31. Kano, Y.; Miwa, M.; Cohen, K. B.; Hunter, L. E.; Ananiadou, S.; Tsujii, J. *IBM J. Res. Dev.* **2011,** *55,* 11:1–11:10. doi:10.1147/JRD.2011.2105691
32. Green, A. Kappa statistics for multiple raters using categorical classifications. In *Proceedings of the 22nd annual SAS User Group International conference,* San Diego, CA, U.S.A.; 1997; pp 1110–1115.
33. Hara, S.; Motohisa, J.; Fukui, T. *J. Cryst. Growth* **2007,** *298,* 612–615. doi:10.1016/j.jcrysgro.2006.10.178
34. Hara, S.; Fukui, T. *Appl. Phys. Lett.* **2006,** *89,* 113111. doi:10.1063/1.2349309
35. Hara, S.; Kawamura, D.; Iguchi, H.; Motohisa, J.; Fukui, T. *J. Cryst. Growth* **2008,** *310,* 2390–2394. doi:10.1016/j.jcrysgro.2007.12.026
36. Wakatsuki, T.; Hara, S.; Ito, S.; Kawamura, D.; Fukui, T. *Jpn. J. Appl. Phys.* **2009,** *48,* 04C137.
37. Ito, S.; Hara, S.; Wakatsuki, T.; Fukui, T. *Appl. Phys. Lett.* **2009,** *94,* 243117. doi:10.1063/1.3157275
38. GPoSTTL. http://gposttl.sourceforge.net (accessed July 6, 2015).
39. CRFpp. http://taku910.github.io/crfpp/ (accessed July 6, 2015).
40. Lafferty, J.; McCallum, A.; Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01,* San Francisco, CA, U.S.A.; 2011; pp 282–289.

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Predicting cytotoxicity of PAMAM dendrimers using molecular descriptors

David E. Jones[1], Hamidreza Ghandehari[2,3,4] and Julio C. Facelli[*1,4]

Full Research Paper

Open Access

Address:
[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT 84112, USA, [2]Department of Bioengineering, University of Utah, Salt Lake City, UT 84112, USA, [3]Department of Pharmaceutics and Pharmaceutical Chemistry, University of Utah, Salt Lake City, UT 84112, USA, and [4]Utah Center for Nanomedicine, Nano Institute of Utah, University of Utah, Salt Lake City, UT 84112, USA

Email:
Julio C. Facelli[*] - julio.facelli@utah.edu * Corresponding author

* Corresponding author

Keywords:
data mining; machine learning; molecular descriptors; poly(amido amine) dendrimers (PAMAM)

Beilstein J. Nanotechnol. **2015,** *6,* 1886–1896.
doi:10.3762/bjnano.6.192

Received: 18 March 2015
Accepted: 20 August 2015
Published: 11 September 2015

This article is part of the Thematic Series "Nanoinformatics for environmental health and biomedicine".

Guest Editor: R. Liu

Abstract

The use of data mining techniques in the field of nanomedicine has been very limited. In this paper we demonstrate that data mining techniques can be used for the development of predictive models of the cytotoxicity of poly(amido amine) (PAMAM) dendrimers using their chemical and structural properties. We present predictive models developed using 103 PAMAM dendrimer cytotoxicity values that were extracted from twelve cancer nanomedicine journal articles. The results indicate that data mining and machine learning can be effectively used to predict the cytotoxicity of PAMAM dendrimers on Caco-2 cells.

Introduction

In silico approaches, such as data mining and machine learning, have been very successful in medicinal chemistry and are commonly used to guide the design of small pharmaceutical compounds [1]. In contrast, although nanomedicine is a rapidly growing field [2], there have been only a few attempts to use data mining techniques in this field. For instance, Liu et al. analyzed a number of attributes of a variety of nanoparticles in order to predict the 24 hour postfertilization mortality in zebrafish [3]. Horev-Azaria and colleagues used predictive modeling to explore the effect of cobalt–ferrite nanoparticles on the viability of seven different cell lines [4]. Sayes and Ivanov used machine learning to predict the induced cellular membrane damage of immortalized human lung epithelial cells caused by metal oxide nanomaterials [5].

As discussed in a previous paper [6], there are a very limited number of databases compiling the properties of nanomedical relevant compounds. We speculate that this has seriously limited the use of data mining techniques in the field of nanomedicine. However, in the above referenced publication, we demonstrated that natural language processing (NLP) techniques can be effectively used to automatically extract nanopar-

ticle property information from the original literature. Here we argued that this development opens the possibility to explore the use of data mining and chemometric techniques to guide the design of new, more effective treatments using nanoparticles. In this paper we apply the methods of data mining and machine learning to predict the cytotoxicity of poly(amido amine) (PAMAM) dendrimers.

Cytotoxicity was the selected criterion because it is of key concern for the nanoscience and nanomedicine community [7,8], considering that high cytotoxicity is a definitive cause for eliminating a material for potential human applications. Reliable prediction of cytotoxicity using in silico approaches possesses the potential for high payoff in nanomaterial development, allowing the concentration of scarce development resources to be directed towards the synthesis and testing of promising materials with expected low levels of toxicity. Cytotoxicity can be determined by a gamut of in vitro toxicity assays focusing on a number of cellular parameters including cell viability, oxidative stress, genotoxicity, and inflammatory response [9]. In this paper, we focus on the cell viability to characterize cytotoxicity [10].

PAMAM dendrimers are good candidates for a data mining methodological study because they are well documented and have the potential to be highly useful as delivery vectors [11]. These nanoparticles are composed of a central core that is surrounded by concentric shells, thus resulting in their well-defined, highly branched structure [12,13]. The generation of the dendrimer is determined by the number of concentric shells that surround the core of the structure. These polymeric nanoparticles can easily be tailored for specific applications. Benefiting from their characteristic scaffold structures, they have been demonstrated to be suitable carriers for a number of diverse bioactive agents, improving the solubility and bioavailability of poorly soluble ones [14,15]. These particular nanoparticles are also promising for use in the treatment of cancer, including oral formulations. In spite of all the desirable properties of dendrimers, there is a significant setback for their use in biomedicine due to their potential toxicological effects, which depend on the structure that is used. It has been shown that cationic PAMAM dendrimers can have surface charge-, generation-, and concentration-dependent toxicity [16-19].

The goal of this research is to demonstrate that data mining methods like the ones used here can be a presynthesis step to identify nondesirable PAMAM dendrimers that have a substantial probability of high toxicity. It would thus be possible to eliminate them from the early stages of the synthetic development pipeline with reasonable confidence. This technique is not meant to replace cytotoxicity assays in the laboratory, but rather to augment these methods. This method will bolster existing cytotoxicity assays by providing the ability to determine relevant compounds with low cytotoxicity and to eliminate weak-candidate PAMAM dendrimers from synthesis and confirmatory testing. This work also illustrates a proof of concept that data mining and machine learning can be applied to PAMAM dendrimers to predict their biochemical properties. This result could potentially be expanded to other nanomaterials in the future.

Results and Discussion

Five different analyses were performed to classify a dendrimer as toxic or nontoxic using different combinations of molecular descriptors and experimental conditions. The first analysis utilized all the molecular descriptors available in MarvinSketch (see Experimental section and Table S1 in Supporting Information File 1). The second analysis involved an automatic feature selection method in which the molecular descriptors that were used had a nonzero rank according to the ChiSquaredAttributeEval method in Weka (see details in the Experimental section). The ChiSquaredAttributeEval method determines the rank of an attribute by calculating the chi-squared statistic with respect to the class [20]. The third analysis used only the molecular descriptors selected by expert advice (see details in the Experimental section): molecular weight, atom count, pI, and molecular polarizability. The fourth analysis included the same molecular descriptors used in the second analysis in addition to the experimental concentration (i.e., the amount in mM of PAMAM dendrimer added to the human colon carcinoma Caco-2 cells culture during the cytotoxicity analysis). The final analysis independently assessed the performance of our best method by randomly splitting the dataset into a training set, including 83 of the values, and a test set, including 20 of the values in the dataset.

The results for the first, second, and third analyses performed to classify dendrimers as toxic/nontoxic are presented in Table 1, Table 2, Table 3 and in Supporting Information File 1, Tables S2–S4. The tables list the average precision, recall, F-measure, and mean absolute error for the toxicity class prediction for all classifiers considered here. The tables also contain the accuracy value for the percentage of correctly classified instances. For all analyses, all classifiers consistently had an accuracy at or above 60.2%.

For the first analysis, Table 1 and Table S2, the J48 and the filtered classifiers show the best results in the 10-fold cross-validation with an accuracy of 74.8%, while bagging, locally weighted learning (LWL), and naive Bayes Tree (NBTree) performed the best with an accuracy of 77.7% in the leave-one-out cross-validation (Table S2). The results from the automatic

Table 1: Results from the 10-fold cross-validation listed by classifier for the first analysis including all molecular descriptors. See Equation 1–4 for the definition of precision, recall, F-measure, and mean absolute error and accuracy.

Classifier	Precision	Recall	F-measure	Mean absolute error	Accuracy
Naive Bayes	0.654	0.660	0.655	0.3370	66.0%
SMO	0.738	0.738	0.725	0.2621	73.8%
J48	0.789	0.748	0.750	0.3077	74.8%
Bagging	0.746	0.738	0.740	0.3211	73.8%
Classification via regression	0.734	0.738	0.730	0.2978	73.8%
Filtered classifier	0.789	0.748	0.750	0.3077	74.8%
LWL	0.775	0.738	0.741	0.2966	73.8%
Decision table	0.678	0.660	0.664	0.3878	66.0%
DTNB	0.691	0.670	0.674	0.3490	67.0%
NBTree	0.696	0.670	0.674	0.3511	67.0%
Random forest	0.736	0.718	0.722	0.3077	71.8%

Table 2: Results from the 10-fold cross-validation listed by classifier for the second analysis including the automatically feature-selected molecular descriptors. See Equation 1–4 for the definition of precision, recall, F-measure, and mean absolute error and accuracy.

Classifier	Precision	Recall	F-measure	Mean absolute error	Accuracy
Naive Bayes	0.654	0.660	0.655	0.3370	66.0%
SMO	0.738	0.738	0.725	0.2621	73.8%
J48	0.789	0.748	0.750	0.3077	74.8%
Bagging	0.746	0.738	0.740	0.3211	73.8%
Classification via regression	0.734	0.738	0.730	0.2978	73.8%
Filtered classifier	0.789	0.748	0.750	0.3077	74.8%
LWL	0.775	0.738	0.741	0.2966	73.8%
Decision table	0.678	0.660	0.664	0.3878	66.0%
DTNB	0.691	0.670	0.674	0.3490	67.0%
NBTree	0.696	0.670	0.674	0.3572	67.0%
Random forest	0.736	0.718	0.722	0.2988	71.8%

Table 3: Results from the 10-fold cross-validation listed by classifier for the third analysis including the molecular descriptors selected by experts. See Equation 1–4 for the definition of precision, recall, F-measure, and mean absolute error and accuracy.

Classifier	Precision	Recall	F-measure	Mean absolute error	Accuracy
Naive Bayes	0.762	0.748	0.750	0.2822	74.8%
SMO	0.738	0.738	0.725	0.2621	73.8%
J48	0.789	0.748	0.750	0.3077	74.8%
Bagging	0.731	0.718	0.721	0.3217	71.8%
Classification via regression	0.762	0.748	0.750	0.3230	74.8%
Filtered classifier	0.804	0.757	0.760	0.3061	75.7%
LWL	0.834	0.777	0.778	0.3008	77.7%
Decision table	0.658	0.650	0.653	0.3980	65.0%
DTNB	0.658	0.650	0.653	0.3969	65.0%
NBTree	0.722	0.689	0.693	0.3454	68.9%
Random forest	0.758	0.748	0.750	0.2973	74.8%

feature selection analysis, using the ChiSquaredAttributeEval and ranker procedures as the attribute evaluator and search method, respectively, are presented Table 2 and Table S3. These results do not differ drastically from those observed in the first analysis, indicating that the use of automatic feature selection does not improve the classification of toxicity in this

study. Alternative automatic feature selection methods using all the WEKA recommended pairings of attribute evaluator and search methods were also tested but did not show any significant improvement in classification prediction performance when using the J48 classifier. These results are presented in Table S7 in Supporting Information File 1. The classification using the features selected by expert advice (Table 3 and Table S4) show that the LWL classifier performed the best with an accuracy of 77.7% in the 10-fold cross-validation. The leave-one-out cross-validation (Table S4) had three classifiers that performed with an accuracy of 78.6% (naive Bayes, bagging, and classification via regression). There is an increase in accuracy across most of the classifiers between the 10-fold and leave-one-out cross-validations. This is an interesting finding because Kohavi noted that k-fold cross-validations typically perform better than leave-one-out cross-validations [21]. This might be an artifact of the dataset not being exactly 50–50 split between toxic and nontoxic samples, thus leading to skewness toward nontoxic predictions.

The decision tree used by the 10-fold and leave-one-out cross-validation J48 classifiers for the first, second, and third analyses is depicted in Figure 1. As shown in the decision tree, the isoelectric point, pI, is the property that is used to classify the dataset. This property represents the pH at which the net charge of an ionizable molecule is zero. The decision tree indicates that if the pI is greater than 12.63, then the dendrimers are toxic. There are 59 PAMAM dendrimers that are classified as toxic of which 21 are misclassified. If the pI is less than or equal to 12.63, then the dendrimers are classified as nontoxic. There are 44 PAMAM dendrimers classified as nontoxic of which 2 are misclassified.

These results indicate that data mining and machine learning can be implemented to predict the cytotoxicity of PAMAM dendrimers on Caco-2 cells with reasonably high accuracy using only molecular descriptors. The misclassifications observed in Figure 1 are much more significant when examining the dendrimers classified as toxic because almost half of these dendrimers are actually nontoxic. This constitutes a substantial quantity of potentially useful dendrimers that are being ruled out, indicating the necessity for further analysis to decrease the number of false positives.

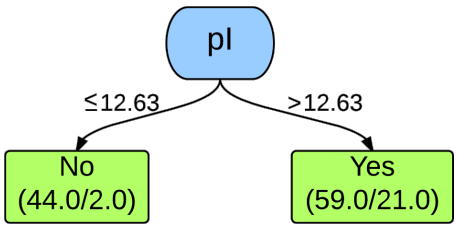


Figure 1: Decision tree for both 10-fold and leave-one-out cross-validation J48 classifier of the first, second, and third analyses. The values indicated on the branches represent the rule or decision used for making the classification. The boxes at the bottom represent the classifications with the number of PAMAM dendrimers classified as such on the left and the number of exceptions (misclassifications) on the right.

Table 4 presents the results using the best performing classifiers from the previous section of the analysis using the expert-selected molecular descriptors with the addition of the concentration of dendrimers used in the experiments. No improvement in prediction was observed when using either the filtered or LWL classifiers, but the J48 prediction accuracy of the classification improved to 83.5%. This substantial improvement in the accuracy of the J48 classifications (from 74% to 83.5 %) shows the importance of including the concentration information from the experimental design in addition to the computed molecular descriptors to properly classify compounds as toxic or nontoxic.

Table 4: Results from the 10-fold cross-validation listed by classifier for the fourth analysis including the expert-selected molecular descriptors with cytotoxicity concentration. See Equation 1–4 for the definition of precision, recall, F-measure, and mean absolute error and accuracy.

Classifier	Precision	Recall	F-measure	Mean absolute error	Accuracy
Naive Bayes	0.755	0.738	0.741	0.2984	73.8%
SMO	0.738	0.738	0.725	0.2621	73.8%
J48	0.838	0.835	0.836	0.2203	83.5%
Bagging	0.836	0.835	0.835	0.2618	83.5%
Classification via regression	0.742	0.738	0.739	0.3157	73.8%
Filtered classifier	0.804	0.757	0.760	0.3061	75.7%
LWL	0.834	0.777	0.778	0.2995	77.7%
Decision table	0.658	0.650	0.653	0.3980	65.0%
DTNB	0.658	0.650	0.653	0.3969	65.0%
NBTree	0.716	0.689	0.693	0.3347	68.9%
Random forest	0.769	0.767	0.768	0.2483	76.7%

The J48 decision tree for the analysis discussed above is depicted in Figure 2. In this case, the pI, molecular weight, and cytotoxicity concentration are the discriminators in the classification. As can be seen, the feature representing the concentration of dendrimers used in the experiments is present in the decision tree for this analysis. The diagram of the decision trees generated from the J48 classifier illustrates important attributes used in the accurate prediction of toxicity for PAMAM dendrimers. The greatest prediction accuracies were achieved after supplementing the expert-selected features with a descriptor representing the experimental conditions by including the concentration under which the cytotoxicity data was acquired. Figure 2 has the same structure at the top level as Figure 1: when the pI is less than or equal to 12.63, 44 PAMAM dendrimers are classified as nontoxic with an exception of 2 that are misclassified. However, when the pI is greater than 12.63, it leads to other options in the classification of the remaining PAMAM dendrimers. The decision made at the next node is determined for a PAMAM dendrimer molecular weight of ≤6908.8 Da or >6908.8 Da. If the molecular weight is >6908.8 Da, 24 PAMAM dendrimers are classified as toxic with four that are misclassified. If the molecular weight is ≤6908.8 Da, there is another option for the molecular weight being ≤3271.9 Da or >3271.9 Da. The final option can be made considering the concentration target for the desired application of the PAMAM dendrimer. In Figure 2, it can be clearly observed that the number of misclassifications (false positives) has been significantly reduced due to this further analysis (from 21 in Figure 1, to 5 in Figure 2). Due to the significant decrease in false positives, the accuracy of the J48 classifier improved. There was a slight increase in the number of false negatives due to this further analysis (from 2 in Figure 1, to 5 in Figure 2).

The classification scheme in Figure 2 identifies three clusters of viable PAMAM dendrimers that have tolerable levels of cytotoxicity: those with a pI less than or equal to 12.63; those with a pI greater than 12.63, but with molecular weights less than or equal to 3271.9 Da that could be used up to concentrations of less than or equal to 0.7 mM; and those with a pI greater than 12.63, with molecular weights between 6908.8–3271.9341 Da that can be used in formulations requiring concentrations less than or equal to 0.01 mM. When designing novel PAMAM dendrimers, these guidelines could be used for developing viable candidates exhibiting low to no cytotoxicity. This demonstrates the importance of combining experimental conditions with molecular descriptors to achieve the greatest prediction accuracy in the classifiers and to find compounds that may be viable under more restrictive conditions. Another important observation is that the properties present in the decision tree diagrams represent the more general properties of charge, size,

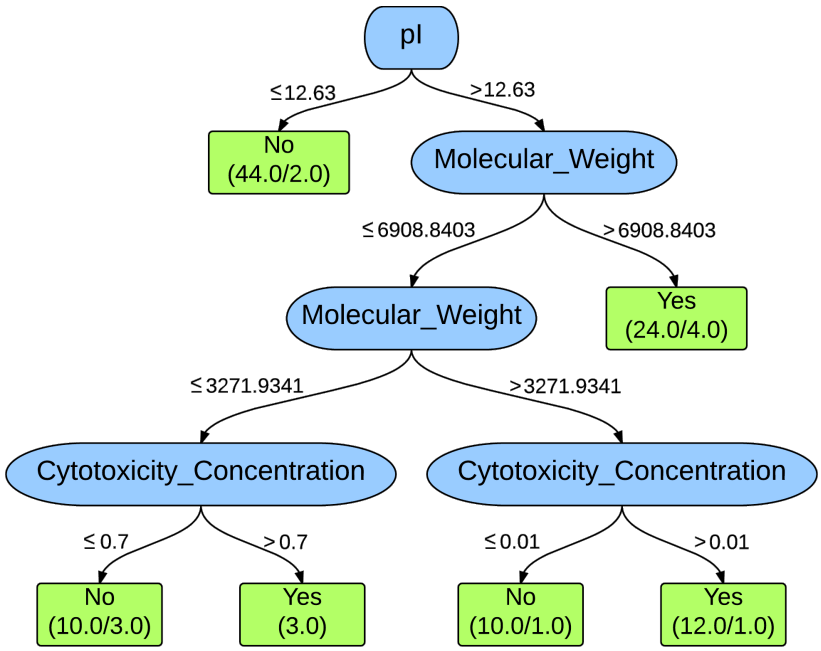


Figure 2: Decision tree for 10-fold cross-validation J48 classifier for the fourth analysis including the molecular descriptors expert-selected with the concentration information of dendrimers used in the experiments. The values present on the branches represent the rule or decision used for making the classification. The boxes at the bottom represent the classifications with the number of PAMAM dendrimers classified as such on the left and the number of exceptions (misclassifications) on the right.

and concentration, which have been hypothesized to be the primary causes of cytotoxicity in Caco-2 cells [22].

Table 5 and Table 6 show the data from the external validation study that was performed to further validate the results presented above. For this study, the dataset was randomly split into a training set consisting of 83 cytotoxicity values, and a test set consisting of 20 cytotoxicity values from the original dataset. Table 5 presents the results from the analysis of this test set using all of the molecular descriptors. For all but one of the classifiers, the predicted accuracy was 65.0%, which is slightly lower than the values obtained for the cross-validation analysis, but the LWL classifier performed very well with an accuracy of 95.0%. This is an interesting finding considering that the highest performance of this classifier in the first four analyses was 77.7%. Table 6 shows the data from the analysis of the test set using only the expert-selected features as well as the cytotoxicity concentration data. Again, the LWL classifier performed with an accuracy of 95.0%, thus no improvement was observed in the classification ability of this algorithm between all molecular descriptors and the expert-feature-selected molecular descriptors with cytotoxicity concentration data. There are two algorithms that exhibited a large improvement between Table 5 and Table 6, namely, the naive Bayes and J48 algorithms. Both of these algorithms improved from a prediction accuracy of 65.0% to 90.0%, which is substantially higher than the values obtained in the cross-validation studies.

These results indicate that data mining and machine learning can be implemented to accurately predict the cytotoxicity of PAMAM dendrimers on Caco-2 cells. According to Figure 2, the results also indicate that the properties such as charge, size, and the desired concentration of the PAMAM dendrimers in the formulation are the important properties in the prediction of cytotoxicity on Caco-2 cells. We believe that the methods used in this work can be expanded to analyze and predict many other biochemically relevant properties of not only unmodified PAMAM dendrimers but also for surface-modified PAMAM

Table 5: Results from the external validation test set analysis listed by classifier using all molecular descriptors. See Equation 1–4 for the definition of precision, recall, F-measure, and mean absolute error and accuracy.

Classifier	Precision	Recall	F-measure	Mean absolute error	Accuracy
Naive Bayes	0.803	0.650	0.617	0.3426	65.0%
SMO	0.803	0.650	0.617	0.3500	65.0%
J48	0.803	0.650	0.617	0.2776	65.0%
Bagging	0.803	0.650	0.617	0.2953	65.0%
Classification via regression	0.803	0.650	0.617	0.3047	65.0%
Filtered classifier	0.803	0.650	0.617	0.2776	65.0%
LWL	0.955	0.950	0.950	0.2510	95.0%
Decision table	0.803	0.650	0.617	0.4206	65.0%
DTNB	0.803	0.650	0.617	0.4182	65.0%
NBTree	0.803	0.650	0.617	0.2945	65.0%
Random forest	0.803	0.650	0.617	0.2784	65.0%

Table 6: Results from the external validation test set analysis listed by classifier including the molecular descriptors expert-selected with cytotoxicity concentration. See Equation 1–4 for the definition of precision, recall, F-measure, and mean absolute error and accuracy.

Classifier	Precision	Recall	F-measure	Mean absolute error	Accuracy
Naive Bayes	0.918	0.900	0.900	0.1868	90.0%
SMO	0.803	0.650	0.617	0.3500	65.0%
J48	0.918	0.900	0.900	0.1768	90.0%
Bagging	0.888	0.850	0.849	0.2408	85.0%
Classification via regression	0.803	0.650	0.617	0.3678	65.0%
Filtered classifier	0.803	0.650	0.617	0.2776	65.0%
LWL	0.955	0.950	0.950	0.2467	95.0%
Decision table	0.803	0.650	0.617	0.4206	65.0%
DTNB	0.803	0.650	0.617	0.4182	65.0%
NBTree	0.803	0.650	0.617	0.3082	65.0%
Random forest	0.888	0.850	0.849	0.2187	85.0%

dendrimers. This method will bolster existing cytotoxicity assays by providing the ability to determine relevant compounds with low cytotoxicity for synthesis and confirmatory testing. This thereby reduces the search space necessary for developing biomedically relevant PAMAM dendrimers. This work not only demonstrates a proof of concept that data mining and machine learning can be applied to PAMAM dendrimers to predict the biochemical property of cytotoxicity, but also indicates that further studies including much larger data sets are necessary to develop reliable and robust classification methods that can be apply to a broader set of compounds, cell cultures and experimental designs.

Conclusion

In this study, classification methods for predicting the Boolean classification of cytotoxicity in Caco-2 cells treated with PAMAM dendrimers were introduced. The results indicate that data mining and machine learning can be used to predict the cytotoxicity of PAMAM dendrimers on Caco-2 cells with good accuracy. In the classification method explored here, it was observed that the properties regarding charge, size, and concentration of the PAMAM dendrimers are the most important properties in the prediction of cytotoxicity and cell viability of Caco-2 cells treated with PAMAM dendrimers. To the authors' knowledge, these results are the first application of data mining and machine learning to predict the cytotoxicity of PAMAM dendrimers on Caco-2 cells using a classification method.

Experimental

The overall workflow of the analysis reported in this paper is presented in Figure 3. The details of the different processes are given in the following subsections.

Nanoparticle selection

The PAMAM dendrimers selected for our study included generations 0, 1, 1.5, 2, 2.5, 3, 3.5, 4, and 4.5 compounds that have been used for transepithelial transport. The full-generation PAMAM dendrimers (generations 0, 1, 2, 3, and 4) are amine- or hydroxy-terminated dendrimers. The half-generation PAMAM dendrimers (generations 1.5, 2.5, 3.5, and 4.5) are carboxyl-terminated dendrimers. For more general property information on the full- and half-generation PAMAM dendrimers, see Table S4 in Supporting Information File 1, which includes the property information for the PAMAM dendrimers analyzed in this study. The toxicity studies used here correspond to assays of these compounds on the human colon carcinoma Caco-2 cell line. The publications containing property data for the nanoparticles selected for this study were gathered from nanomedicine articles available in Scopus and PubMedCentral using the search terms "PAMAM dendrimers AND cytotoxicity AND Caco-2 cells". In order for the PAMAM dendrimer cytotoxicity values to be considered relevant for extraction, both cell viability and treatment concentration information had to be available in the publication. From this literature corpus, 103 PAMAM dendrimer cytotoxicity

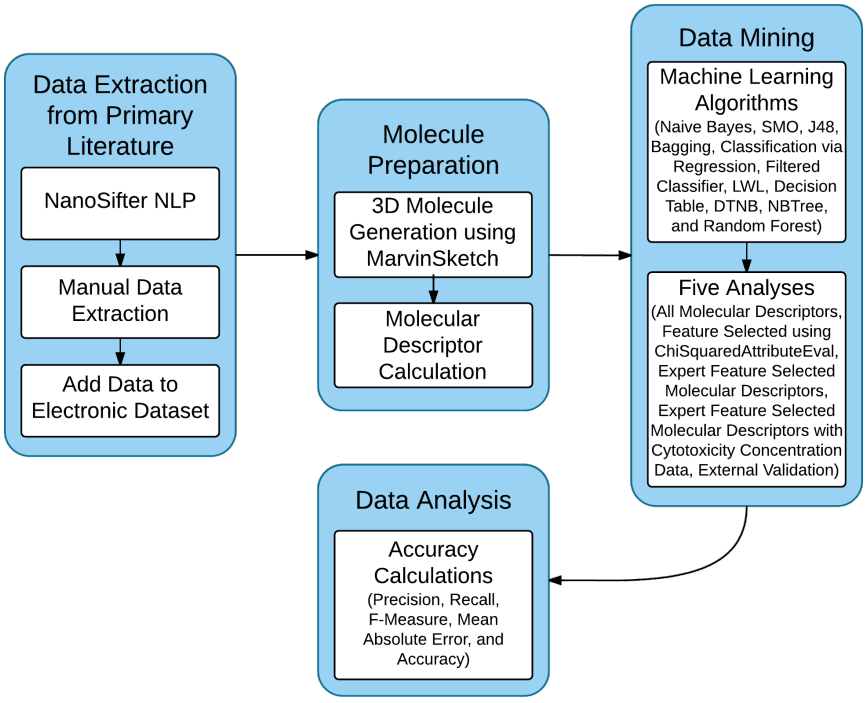


Figure 3: Simplified workflow diagram for the method used in this study.

values were extracted to be included in this study [23-34]. NanoSifter [6], followed by manual revision, was used to extract the cell viability and cytotoxicity treatment concentration information from the journal articles in the corpus described above.

Chemical structure rendering and molecular descriptor calculation

The structures of the PAMAM dendrimers were manually constructed using MarvinSketch by ChemAxon [35,36]. There were a total of 10 PAMAM dendrimer structures created for this study. They included generations 0, 1, 1.5, 2, 2.5, 3, 3.5, 4, and 4.5 PAMAM dendrimers. These models include both amine-terminated (full-generation) and carboxyl-terminated (half-generation) structures, as well as one hydroxy-terminated structure (full-generation but hydroxy-terminated). The molecular descriptors for each molecule were calculated using plugins built into MarvinSketch [36]. The list of the 51 molecular descriptors calculated for each molecule is given along with their corresponding definitions in Supporting Information File 1, Table S1. Among these molecular descriptors, there are 42 structural properties (two mass-related, six atom-count-related, seven bond-count-related, four ring-size-related, 13 ring-count-related, and ten other structural properties) and nine chemical properties (five charge-related and four hydrogen-bonding-related properties).

Data preparation and preprocessing

The data, consisting of the molecular descriptors calculated for all of the molecules considered here and the corresponding cell viability and cytotoxicity data, was uploaded into WEKA [20] to perform the machine learning and data mining analysis using classification methods to discern between toxic and nontoxic compounds. In order to assign a categorical value to each dendrimer cytotoxicity data point, the threshold was established at a cell viability value of 90% (i.e., compounds were considered nontoxic at a certain concentration of PAMAM dendrimer nanoparticles if 90% of the Caco-2 cell population survived after the intervention). Because there is statistical variation in cell viability studies, nontoxic materials can have a few percent above or below 100% cell viability. Hence, the threshold of 90% was set arbitrarily to take into account the usual variability in this type of study.

Prediction of toxicity using classification methods

Five different analyses were performed to classify a dendrimer as toxic or nontoxic using different combinations of molecular descriptors and experimental conditions. The first analysis utilized all the molecular descriptors. The second analysis involved an automatic feature selection using the ChiSquaredAttributeEval and ranker method built into WEKA, where only molecular descriptors with a nonzero rank were included in this analysis. The molecular descriptors with a nonzero rank were H-bond acceptor sites, pI, logP, Harary index, refractivity, bond count, molecular polarizability, rotatable bond count, atom count, logD, aliphatic bond count, chain bond count, chain atom count, aliphatic atom count, exact mass, molecular weight, Wiener index, Randic index, Szeged index, Wiener polarity, Platt index, H-bond donor count, hyper Wiener index, H-bond donor sites, and H-bond acceptor count. The third analysis used only molecular descriptors selected by expert advice: molecular weight, atom count, pI, and molecular polarizability. In this paper we refer to selected by expert advice as the properties that an experienced researcher in nanocarriers, Dr. Ghandehari, expected to be relevant to predict toxicity based on his own knowledge derived from work is his lab and literature precedents. The fourth analysis included the same molecular descriptors as the ones used in the second analysis and the experimental concentration, i.e., the amount in mM of PAMAM dendrimer added to the Caco-2 cells during cytotoxicity analysis. The fifth analysis was an external validation study in which we randomly selected 20 cytotoxicity values from the original dataset of 103 to create a test set. The remaining 83 cytotoxicity values were used as the training set.

In this work we used the following classifiers: naive Bayes, sequential minimal optimization (SMO), J48, bagging, classification via regression, filtered classifier, LWL, decision table, decision table/naive Bayes (DTNB), NBTree, and random forest. We wanted to explore many modeling methods to provide a wide landscape of available techniques. Since the computational cost is low, there is no strong argument to limit this exploration. Naive Bayes is a Bayesian classifier that uses posterior probability to predict the value of the target attribute [37]. That is, by using a given input attribute, the classifier attempts to find the target attribute value that maximizes the conditional probability of the target attribute. SMO is a support vector machine classifier that globally replaces all values and transforms nominal attributes into binary ones [38]. By default it normalizes all attributes. J48 is a decision tree classifier, which is based on the C4.5 algorithm [39]. This method starts with large sets of cases which belong to known classes, then cases are analyzed for patterns that allow for reliable discrimination of classes. The patterns are represented as models, either in the form of decision trees or sets of if/then rules that can be used to classify new cases. Bagging is a hybrid classification method that creates classes and reduces variance by bagging classifiers [40]. Classification via regression performs its classification by binarizing each class and building one regression model for each class [41]. The filtered classifier is an arbitrary classifier that runs on data passed through an arbitrary filter

[20]. LWL uses an instance-based algorithm to assign instance weights [42]. The decision table is a simple decision table majority classifier [43]. DTNB is a decision table/naive Bayes hybrid classifier. During the search, the algorithm determines the need to divide the attributes into two disjoint subsets: one for the decision table, the other for naive Bayes [44]. NBTree is a decision tree/naive Bayes hybrid classifier that builds a decision tree with naive Bayes classifiers at the leaves [45]. All the calculations were performed using WEKA [20].

Two different cross-validation [46] schemes were performed for each classifier. The first one was a 10-fold cross-validation in which the dataset was divided into 10 parts or folds [20]. During each classification run, nine of the folds were used as a training set and one was used as a test set and the results were averaged over the ten runs. The second cross-validation scheme used here was the leave-one-out cross-validation [20]. As this cross-validation method states, one sample is left out as the test set, and the rest of the dataset is the training set. This method runs this through as many iterations as there are samples in the dataset.

The predictions determined by WEKA were evaluated and determined to be true positive, false positive, or false negative by manual inspection. The precision, recall, and F-measure were calculated using the following equations:

$$\text{precision} = TP / \left(TP + FP\right) \tag{1}$$

$$\text{recall} = TP / \left(TP + FN\right) \tag{2}$$

$$\text{F-measure} = \frac{\left(1+\beta^2\right)\cdot \text{precision} \cdot \text{recall}}{\left(\beta^2 \cdot \text{precision}\right) + \text{recall}} \tag{3}$$

$$\text{mean absolute error} = \left(\sum f_i - y_i\right)/n \tag{4}$$

In these equations, TP is true positive, FP is false positive, FN is false negative, and β is the weighting applied to the relationship between precision and recall. The precision and recall were weighted evenly, so β = 1 [6]. The precision, recall, and F-measure of each classifier were calculated for each classification (toxic/nontoxic). Each measure for each classification (toxic/nontoxic) was then averaged. The average value for the precision, recall, and F-measure were recorded. For mean absolute error, f_i is the prediction, y_i is the true value, and n is the number of calculated absolute errors.

Supporting Information

Supporting Information File 1

Supporting Tables.
This document includes all of the tables not present in the text of the document that referenced throughout the document.
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-192-S1.pdf]

Supporting Information File 2

Raw Data.
This is the dataset containing all of the raw data used for all of the analyses in this study.
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-192-S2.xlsx]

Supporting Information File 3

SMILES description of all the dendrimers studied here.
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-192-S3.smiles]

Acknowledgements

The project described was supported by Grant Number T15LM007124 from the National Library of Medicine. Also, this work has been partially funded by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number 1ULTR001067 and NIH grants R01ES024681 and R01EB007470.

References

1. Tropsha, A.; Golbraikh, A. *Curr. Pharm. Des.* **2007,** *13,* 3494–3504. doi:10.2174/138161207782794257
2. Thomas, D. G.; Pappu, R. V.; Baker, N. A. *J. Biomed. Inf.* **2011,** *44,* 59–74. doi:10.1016/j.jbi.2010.03.001
3. Liu, X.; Tang, K.; Harper, S.; Harper, B.; Steevens, J. A.; Xu, R. *Int. J. Nanomed.* **2013,** *8* (Suppl. 1), 31–43. doi:10.2147/IJN.S40742
4. Horev-Azaria, L.; Baldi, G.; Beno, D.; Bonacchi, D.; Golla-Schindler, U.; Kirkpatrick, J. C.; Kolle, S.; Landsiedel, R.; Maimon, O.; Marche, P. N.; Ponti, J.; Romano, R.; Rossi, F.; Sommer, D.; Uboldi, C.; Unger, R. E.; Villiers, C.; Korenstein, R. *Part. Fibre Toxicol.* **2013,** *10,* 32. doi:10.1186/1743-8977-10-32
5. Sayes, C.; Ivanov, I. *Risk Anal.* **2010,** *30,* 1723–1734. doi:10.1111/j.1539-6924.2010.01438.x
6. Jones, D. E.; Igo, S.; Hurdle, J.; Facelli, J. C. *PLoS One* **2014,** *9,* e83932. doi:10.1371/journal.pone.0083932
7. Elsaesser, A.; Howard, C. V. *Adv. Drug Delivery Rev.* **2012,** *64,* 129–137. doi:10.1016/j.addr.2011.09.001

8. Fadeel, B.; Garcia-Bennett, A. E. *Adv. Drug Delivery Rev.* **2010,** *62,* 362–374. doi:10.1016/j.addr.2009.11.008
9. Landsiedel, R.; Ma-Hock, L.; Kroll, A.; Hahn, D.; Schnekenburger, J.; Wiench, K.; Wohlleben, W. *Adv. Mater.* **2010,** *22,* 2601–2627. doi:10.1002/adma.200902658
10. Workshop on Nanoinformatics Strategies. National Science Foundation: Arlington, Virginia, USA, 2007. http://128.119.56.118/~nnn01/Workshop.html
11. du Toit, L. C.; Pillay, V.; Choonara, Y. E.; Pillay, S.; Harilall, S.-l. *Recent Pat. Drug Delivery Formulation* **2007,** *1,* 131–142. doi:10.2174/187221107780831941
12. Bielinska, A.; Kukowska-Latallo, J. F.; Johnson, J.; Tomalia, D. A.; Baker, J. R., Jr. *Nucleic Acids Res.* **1996,** *24,* 2176–2182. doi:10.1093/nar/24.11.2176
13. Meltzer, A. D.; Tirrell, D. A.; Jones, A. A.; Inglefield, P. T.; Hedstrand, D. M.; Tomalia, D. A. *Macromolecules* **1992,** *25,* 4541–4548. doi:10.1021/ma00044a013
14. Kolhe, P.; Misra, E.; Kannan, R. M.; Kannan, S.; Lieh-Lai, M. *Int. J. Pharm.* **2003,** *259,* 143–160. doi:10.1016/S0378-5173(03)00225-4
15. Wood, K. C.; Little, S. R.; Langer, R.; Hammond, P. T. *Angew. Chem.* **2005,** *44,* 6704–6708. doi:10.1002/anie.200502152
16. Greish, K.; Thiagarajan, G.; Herd, H.; Price, R.; Bauer, H.; Hubbard, D.; Burckle, A.; Sadekar, S.; Yu, T.; Anwar, A.; Ray, A.; Ghandehari, H. *Nanotoxicology* **2012,** *6,* 713–723. doi:10.3109/17435390.2011.604442
17. Thiagarajan, G.; Greish, K.; Ghandehari, H. *Eur. J. Pharm. Biopharm.* **2013,** *84,* 330–334. doi:10.1016/j.ejpb.2013.01.019
18. Xu, Q.; Wang, C. H.; Pack, D. W. *Curr. Pharm. Des.* **2010,** *16,* 2350–2368. doi:10.2174/138161210791920469
19. Yellepeddi, V. K.; Kumar, A.; Palakurthi, S. *Expert Opin. Drug Delivery* **2009,** *6,* 835–850. doi:10.1517/17425240903061251
20. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. *ACM SIGKDD Explorations Newsletter* **2009,** *11,* 10–18. doi:10.1145/1656274.1656278
21. Kohavi, R. The Power of Decision Tables. In *Machine Learning: ECML-95;* Lavrac, N.; Wobel, S., Eds.; Lecture Notes in Computer Science, Vol. 912; Springer Verlag: Berlin, Germany, 1995; pp 174–189. doi:10.1007/3-540-59286-5_57
22. El-Sayed, M.; Ginski, M.; Rhodes, C. A.; Ghandehari, H. *J. Bioact. Compat. Polym.* **2003,** *18,* 7–22. doi:10.1177/0883911503018001002
23. Goldberg, D. S.; Vijayalakshmi, N.; Swaan, P. W.; Ghandehari, H. *J. Controlled Release* **2011,** *150,* 318–325. doi:10.1016/j.jconrel.2010.11.022
24. Jevprasesphant, R.; Penny, J.; Jalal, R.; Attwood, D.; McKeown, N. B.; D'Emanuele, A. *Int. J. Pharm.* **2003,** *252,* 263–266. doi:10.1016/S0378-5173(02)00623-3
25. Ke, W.; Zhao, Y.; Huang, R.; Jiang, C.; Pei, Y. *J. Pharm. Sci.* **2008,** *97,* 2208–2216. doi:10.1002/jps.21155
26. Kitchens, K. M.; Foraker, A. B.; Kolhatkar, R. B.; Swaan, P. W.; Ghandehari, H. *Pharm. Res.* **2007,** *24,* 2138–2145. doi:10.1007/s11095-007-9415-0
27. Kitchens, K. M.; Kolhatkar, R. B.; Swaan, P. W.; Ghandehari, H. *Mol. Pharmaceutics* **2008,** *5,* 364–369. doi:10.1021/mp700089s
28. Kolhatkar, R. B.; Kitchens, K. M.; Swaan, P. W.; Ghandehari, H. *Bioconjugate Chem.* **2007,** *18,* 2054–2060. doi:10.1021/bc0603889
29. Najlah, M.; Freeman, S.; Attwood, D.; D'Emanuele, A. *Int. J. Pharm.* **2007,** *336,* 183–190. doi:10.1016/j.ijpharm.2006.11.047
30. Pisal, D. S.; Yellepeddi, V. K.; Kumar, A.; Kaushik, R. S.; Hildreth, M. B.; Guan, X.; Palakurthi, S. *Int. J. Pharm.* **2008,** *350,* 113–121. doi:10.1016/j.ijpharm.2007.08.033
31. Schilrreff, P.; Mundiña-Weilenmann, C.; Romero, E. L.; Morilla, M. J. *Int. J. Pharm.* **2012,** *7,* 4121–4133. doi:10.2147/IJN.S32785
32. Sweet, D. M.; Kolhatkar, R. B.; Ray, A.; Swaan, P.; Ghandehari, H. *J. Controlled Release* **2009,** *138,* 78–85. doi:10.1016/j.jconrel.2009.04.022
33. Teow, H. M.; Zhou, Z.; Najlah, M.; Yusof, S. R.; Abbott, N. J.; D'Emanuele, A. *Int. J. Pharm.* **2013,** *441,* 701–711. doi:10.1016/j.ijpharm.2012.10.024
34. Najlah, M.; Freeman, S.; Attwood, D.; D'Emanuele, A. *Bioconjugate Chem.* **2007,** *18,* 937–946. doi:10.1021/bc060325q
35. Future-proofing Cheminformatics Platforms. ChemAxonhttp://www.chemaxon.com/wp-content/uploads/2012/04/Future_proofing_cheminformatics_platforms.pdf (accessed Aug 24, 2015).
36. *Marvin,* Revision 5.12.4; ChemAxon, 2013.
37. Witten, I. H.; Frank, E.; Hall, M. A. *Data Mining: Practical Machine Learning Tools and Techniques,* 3rd ed.; Morgan Kaufmann: Burlington, MA, U.S.A., 2011. doi:10.1016/B978-0-12-374856-0.00025-0
38. Schölkopf, B.; Burges, C. J. C.; Smola, A. J. *Advances in Kernel Methods;* The MIT Press: Cambridge, MA, U.S.A., 1998.
39. Quinlan, R. *C4.5: Programs for Machine Learning;* Morgan Kaufmann: San Mateo, CA, U.S.A., 1993.
40. Breiman, L. *Mach. Learn.* **1996,** *24,* 123–140. doi:10.1023/A:1018054314350
41. Frank, E.; Wang, Y.; Inglis, S.; Holmes, G.; Witten, I. H. *Mach. Learn.* **1998,** *32,* 63–76. doi:10.1023/A:1007421302149
42. Frank, E.; Hall, M.; Pfahringer, B. Locally weighted naive Bayes. In *Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence,* Acapulco, Mexico; 2003; pp 249–256.
43. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence,* Morgan Kaufmann: San Francisco, CA, U.S.A., 1995.
44. Hall, M. A.; Frank, E. Combining Naive Bayes and Decision Tables. In *Proceedings of Twenty-First International Florida Artificial Intelligence Research Society,* May 15–17, 2008; AAAI Press: Coconut Grove, FL, U.S.A., 2008; pp 318–319.
45. Kohavi, R. Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In *Proceedings of the second international conference on knowledge discovery and data mining,* 1996; pp 202–207.
46. Maimon, O.; Rokach, L. *The Data Mining and Knowledge Discovery Handbook;* Springer: New York, NY, U.S.A., 2005. doi:10.1007/b107408

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

An ISA-TAB-Nano based data collection framework to support data-driven modelling of nanotoxicology

Richard L. Marchese Robinson[1], Mark T. D. Cronin[*1], Andrea-Nicole Richarz[1] and Robert Rallo[2]

Full Research Paper

Open Access

Address:
[1]School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, James Parsons Building, Byrom Street, Liverpool, L3 3AF, United Kingdom and [2]Departament d'Enginyeria Informatica i Matematiques, Universitat Rovira i Virgili, Av. Paisos Catalans 26, 43007 Tarragona, Catalunya, Spain

Email:
Mark T. D. Cronin[*] - M.T.Cronin@ljmu.ac.uk

* Corresponding author

Keywords:
databases; ISA-TAB-Nano; nanoinformatics; nanotoxicology; quantitative structure–activity relationship (QSAR)

Beilstein J. Nanotechnol. **2015,** *6,* 1978–1999.
doi:10.3762/bjnano.6.202

Received: 31 March 2015
Accepted: 27 August 2015
Published: 05 October 2015

This article is part of the Thematic Series "Nanoinformatics for environmental health and biomedicine".

Guest Editor: R. Liu

Abstract

Analysis of trends in nanotoxicology data and the development of data driven models for nanotoxicity is facilitated by the reporting of data using a standardised electronic format. ISA-TAB-Nano has been proposed as such a format. However, in order to build useful datasets according to this format, a variety of issues has to be addressed. These issues include questions regarding exactly which (meta)data to report and how to report them. The current article discusses some of the challenges associated with the use of ISA-TAB-Nano and presents a set of resources designed to facilitate the manual creation of ISA-TAB-Nano datasets from the nanotoxicology literature. These resources were developed within the context of the NanoPUZZLES EU project and include data collection templates, corresponding business rules that extend the generic ISA-TAB-Nano specification as well as Python code to facilitate parsing and integration of these datasets within other nanoinformatics resources. The use of these resources is illustrated by a "Toy Dataset" presented in the Supporting Information. The strengths and weaknesses of the resources are discussed along with possible future developments.

Introduction

Nanotechnology, which may be considered the design and application of engineered nanomaterials with desired properties [1,2], is of increasing importance [3,4]. Nanomaterials may be considered to be any chemicals with (a majority of) constituent particles with one or more dimensions in the nanoscale (typically 1–100 nm) range and engineered nanomaterials may be considered to be any nanomaterials that are intentionally produced. (It should be noted that slightly different definitions of these terms have been proposed by different organisations [1] and the European Commission has recommended a specific definition of a "nanomaterial" for legislative and policy purposes within the European Union [5].)

Nanomaterials have been used and/or have been investigated for use in a diverse range of applications such as sunscreens, cosmetics, electronics and medical applications [2,4,6,7]. In addition to interest in the benefits offered by nanotechnology, concerns have also been raised about the potential risk posed by nanomaterials to human health and the environment [3,4,7]. Various research initiatives have been (and are being) funded to advance scientific understanding of nanotechnology and nanosafety and to enable the appropriate selection, design and regulation of nanomaterials for technological applications [3,8,9]. There is a particular interest in the possibility of using computational approaches as part of the safety assessment of nanomaterials, e.g., to enable "safety by design" [3,7,9,10].

Experimental data are critical to advancing understanding of the properties of nanomaterials and the ability to design nanomaterials with desirable technological properties and acceptable safety profiles [2,9-11]. In order to enable "safety by design", data from toxicity studies need to be related to relevant structural/physicochemical data [10], where the latter may include information about chemical composition as well as a range of other measured properties such as size distribution statistics and zeta potential, to name but two [12]. Being able to relate these data allows for the development of predictive models based on quantitative structure–activity relationships (QSARs) for nanomaterials – so-called quantitative nanostructure–activity relationships ("QNARs") [10] or "nano-QSARs" [13] – as well as "category formation" and "read-across" predictions [9,14,15].

In order to make most effective use of these data, experimental datasets should be made available via a standardised, electronic format that facilitates meaningful exchange of information between different researchers, submission to (web-based) searchable databases, integration with other electronic data resources and analysis via appropriate (modelling) software [9,16-18]. This could entail directly populating files based on a standardised format or direct entry of data into searchable databases using a (web-based) data entry tool [19], followed via data export/exchange in a standardised format. However, in contrast to directly populating standardised, structured files (such as spreadsheets), direct entry of data into (web-based) searchable databases may not be possible for domain experts (e.g., nanotoxicologists in experimental labs) with little or no informatics support. These researchers may not have their own, in-house database systems and data entry to a third party database at the point of data collection may not be practical. Data collected using standardised, structured files may be readily, programatically submitted to (web-based) searchable databases at a later stage in the research cycle.

Standardised, structured files also facilitate programmatic analysis (i.e., entirely new codes and/or configuration files do not need to be developed for each new dataset) for the purposes of computational modelling. They also facilitate integration between datasets, partly due to the ease of programmatic analysis and in part because standardisation makes it clearer when two items of (meta)data in distinct datasets are related. Data integration within searchable databases supports computational modelling via enabling data from multiple sources to be combined, in principle, for more robust, generalisable analysis and via facilitating the identification of data which are relevant to the needs of a given modeller.

Regarding the nature of these standardised, structured files, whilst more complicated file formats based on the eXtended Markup Language (XML) or the Resource Description Framework (RDF) might be considered, a spreadsheet-based file format offers a key advantage: most scientists are likely to be familiar with creating, editing and viewing spreadsheet-based datasets [17,20,21]. Indeed, these kinds of files can be edited and viewed using widely used, non-specialist software (such as Microsoft Excel), whilst (to some extent) a spreadsheet-like interface may be retained within specialist software designed to ensure the files are compliant with the rules of a standardised specification [17,20,22]. However, no claim is being made as to the intrinsic optimality of a spreadsheet-based format: a detailed discussion of the advantages and disadvantages of different file formats is beyond the scope of the current publication and interested readers are referred to the cited literature and the references therein [17,20,21].

The ISA-TAB-Nano specification, comprising a set of interrelated spreadsheet-based tabular file types, was recently proposed as a solution to the requirement for a standardised, electronic format for nanomaterial data [16,17,23]. However, as well as a general specification specifying how different kinds of (meta)data should be recorded in a standardised fashion, additional requirements for nanotoxicology datasets to be most valuable for analysis of trends and development of data driven models exist. These requirements include the need to report the necessary physicochemical parameters, experimental details and other relevant metadata such as provenance [12,24-27]. Whilst the generic ISA-TAB-Nano specification [17,23] specifically calls for relevant provenance information to be provided, and facilitates presentation of other (meta)data, it does not specify all of the (meta)data which should be recorded nor exactly how these (meta)data should be presented.

This article presents a set of resources which were designed for manually harvesting data from the published literature to create ISA-TAB-Nano datasets in order to support analysis and model-

ling of nanotoxicology data, including the integration of these data within online, searchable databases. Specifically, these resources are as follows: a collection of Excel templates for creating ISA-TAB-Nano files containing specific, relevant (meta)data manually harvested from the scientific literature; a corresponding set of business rules for populating these templates which build upon the generic ISA-TAB-Nano specification; a Python program for converting the resulting ISA-TAB-Nano files to tab-delimited text files to facilitate computational analysis and database submission. Since there is a growing interest in the use of ISA-TAB-Nano as a community standard for organising nanomaterial data, from a variety of individual researchers and organizations [3,28-32], it is anticipated that these resources will be of value for the research community.

These resources were developed within the context of the NanoPUZZLES project [33], but their development was informed via discussions with various researchers in the nanoinformatics/nanotoxicology community and consideration of various complementary nanoinformatics resources such as those developed within the MODERN [34] and eNanoMapper [35] projects.

The rest of the article is organised as follows. Section 1 of "Results and Discussion" provides a brief overview of the generic ISA-TAB-Nano specification. Section 2 summarises some challenges associated with the use of this generic specification (especially when used to collect data from the literature), which the current work sought to address. Section 3 summarises the data collection templates and the basis on which they were developed. Section 4 summarises the new business rules which were created for populating these templates. Section 5 provides an overview of the Python program written to facilitate analysis and databases submission of datasets created using these templates. Section 6 presents a "Toy Dataset" created using these templates. Section 7 presents a critical appraisal of the developed resources, discusses links to related research initiatives and resources along with possible future directions for this work. The "take home" messages of this article are summarised under "Conclusion". The challenges, business rules and notable limitations of the presented resources (summarised in sections 2, 4 and 7, respectively) are fully explained in the Supporting Information. The resources described in this article, along with the "Toy Dataset", are publicly available under open licenses (see Supporting Information Files 1–4).

Results and Discussion

1 A brief overview of the generic ISA-TAB-Nano specification

The ISA-TAB-Nano specification [17,23] extends the ISA-TAB specification [18,20,22,36] which was previously proposed as an exchange standard for biological data and metadata based on a standardised metadata representation. Unless noted otherwise, the specification incorporates [17,23] all the business rules (e.g., restrictions on which fields can hold multiple values) associated with the original ISA-TAB specification [36]. The official ISA-TAB-Nano wiki [23] provides the most up to date information regarding the generic ISA-TAB-Nano specification, including detailed descriptions [37-40] and Excel templates for each of the file types described below. Since the original description of the specification in Thomas et al. [17], two revisions (version 1.1 and version 1.2) of the specification had been published on the wiki at the time of writing. The overview provided in the current paper refers to version 1.2 of ISA-TAB-Nano. Since the specification is extensively described elsewhere [17,23], the following overview focuses on the essential background required to understand the following sections of the current paper.

The ISA-TAB-Nano specification describes a set of four linked file types (Investigation, Study, Assay, Material), each of which is a spreadsheet-like table, which are used to record different kinds of (meta)data associated with a given "investigation", which may be considered to correspond to a set of different kinds of experimental studies carried out on a given set of nanomaterials [36]. In addition, the specification describes corresponding business rules governing how these files can be populated. A given "investigation" is associated with a single Investigation file and, potentially, multiple Study, Assay and Material files. The kinds of (meta)data each file type is designed to record and the links between different kinds of files is summarised in Figure 1 and discussed in more detail below.

Investigation file

The Investigation file [37] reports key metadata describing the terms used in the other files as well as reporting overall conclusions derived from the "investigation", if any.

Material file

Each of the nanomaterial samples (implicitly as originally sourced for the "investigation" [17]) is described by a corresponding Material file [40] associated with a unique identifier reported in the "Material Source Name" column and used to label the Material file. A Material file presents chemical composition information along with other descriptive information about the sample such as nominal or manufacturer supplied characteristics reported via end user defined "Characteristics [*characteristic name*]" columns. Since nanomaterials of diverse types (e.g., dendrimers, carbon nanotubes, surface-coated metal oxides) may comprise different components (e.g., core and shell), the initial rows of the Material file are used to describe the overall nanomaterial sample with subsequent rows used to

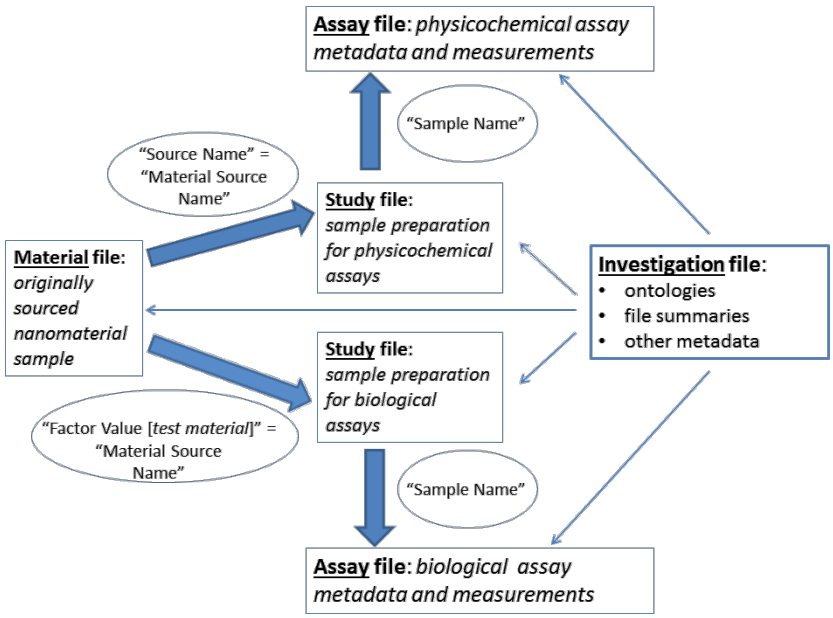


Figure 1: A schematic illustration of the links between ISA-TAB-Nano files. Biological or material samples are prepared for measurements in biological or physicochemical assays respectively. Assay files link measurement values with prepared sample identifiers ("Sample Name" values). Study files describe sample preparation. Material files describe the nanomaterials obtained for testing, denoted via their "Material Source Name" identifiers. N.B. Italic font denotes generic names, e.g., "Factor Value [*test material*]" is replaced with "Factor Value [nanomaterial]" in the NanoPUZZLES in vitro cell-based Study file template.

describe the individual components: the overall sample and different components are each assigned unique values in the "Material Name" column.

Study file

A Study file [38] describes the preparation of samples for analysis via some assay protocol. The identifiers of prepared samples are reported in "Sample Name" columns, with sequentially prepared samples corresponding to identifiers in sequential "Sample Name" columns, and the identifier(s) of the original material(s) from which these samples were prepared is (are) reported in the "Source Name" column. In principle, multiple "Source Name" identifiers might correspond to one or more "Sample Name" identifiers [36]. However, in the simplest case (as adopted in the current work), a single prepared sample corresponds to a single original material, i.e., each row corresponds to a single "Source Name" and a single "Sample Name" identifier. Properties associated with the original material or, more specifically, a prepared sample may be reported via "Characteristics [*characteristic name*]" columns situated after the "Source Name" column or after the relevant "Sample Name" column respectively. Here, it should be noted that the properties recorded via these columns should not include experimental endpoints which would be reported via an Assay file or other information about original nanomaterial samples which would be reported via a Material file.

The transformation of the original material into the prepared sample(s) corresponds to one or more protocols (with corresponding protocol names reported in "Protocol REF" columns), associated with corresponding protocol "parameters" (reported in "Parameter Value [*parameter name*]" columns), and "factors" (reported in "Factor Value [*factor name*]" columns). The concept of "parameters" refers to "variables that are kept constant in an assay experiment", whilst the concept of "factors" refers to "variables that are changed for studying their effects on the measured endpoint" [17]. If the assay is biological (e.g., an in vitro cytotoxicity assay), the originally sourced biological material is considered the original material, with its identifier reported in the "Source Name" column, from which a sample is prepared for testing in an assay and the originally sourced nanomaterial is considered a "factor", since the effect of adding this nanomaterial to the biological sample being prepared for evaluation is studied: the corresponding Material file identifier ("Material Source Name") is reported in an appropriate "Factor Value [*factor name*]" column (e.g., "Factor Value [nanomaterial]). If the assay measures nanomaterial physicochemical parameters (e.g., size by dynamic light scattering, zeta potential), the originally sourced nanomaterial sample is considered the original material, i.e., the "Material Source Name" is reported in the Study file "Source Name" column. It follows that different Study files must be created for samples prepared for biological or physicochemical assays.

Assay file

An Assay file [39] links (a subset of) the prepared samples described in a given Study file to the experimental measurements, of a given type, obtained in a given assay. Each Assay file row corresponds to a given sample, with the "Sample Name" identifier defined in the corresponding Study file being reported in the Assay file "Sample Name" column. Additional columns ("Protocol REF", "Assay Name", "Parameter Value [*parameter name*]", "Factor Value [*factor name*]") in the Assay file identify the assay protocol performed and experimental details associated with the production of a given (set of) data point(s) obtained from that assay for a given sample. (Here, the concepts of "parameters" and "factors" are as defined above for the Study file, although Assay file "parameters" are specific to Assay file protocols and one may choose to report "factors" in the Assay file if they are applicable to the assay procedure used to generate data points for a given prepared sample [17,39].) The corresponding data points are presented in "Measurement Value [*statistic(measurement name)*]" columns, e.g., "Measurement Value [z-average(hydrodynamic diameter)]" for an Assay file describing dynamic light scattering (DLS) size measurements [41,42].

External files

"External" files [17,36], presenting additional information associated with the original nanomaterial samples or assay measurements, can be linked to the appropriate Material and Assay file respectively via additional columns and may also be included within the ISA-TAB-Nano dataset.

Support for (meta)data standardisation

The ISA-TAB-Nano specification promotes standardised reporting of (meta)data in the following ways. (1) It defines a certain number of fixed fields (rows in the Investigation file, or columns in the remaining file types). (2) It describes a syntax for adding additional fields of a given type, e.g., "Parameter Value [*parameter name*]" and "Factor Value [*factor name*]". (3) It supports links between terms added by the end user (e.g., a *parameter name* or the unit for a "Measurement Value [*statistic(measurement name)*]" column entry) and standardised definitions retrieved from ontologies. (An excellent introduction to ontologies can be found in the recent articles of Thomas et al. [2,11] along with an overview of a highly relevant example: the NanoParticle Ontology (NPO) [2].) (4) It supports links to standardised protocol documentation, for sample preparation or assay measurements, for protocol names reported in "Protocol REF" columns in a Study or Assay file. (The ontologies to which various terms are linked are defined using fields in the Investigation file, which also provides links between protocol names and standardised documentation.)

As well as providing some pre-defined fields and stipulating a specific syntax for adding fields of a specific type (e.g., "Factor Value [*factor name*]"), miscellaneous additional fields can be created via adding new "Comment [*name of (meta)data item*]" fields if no appropriate alternative exists.

2 Challenges associated with the generic ISA-TAB-Nano Specification which were addressed in the current Work

Table 1 presents some key challenges associated with the use of the generic ISA-TAB-Nano specification (version 1.2), especially when used to collect data from the published literature, and which were addressed in the work reported in the current article. An in-depth explanation of these challenges, along with a detailed discussion of the manner in which they were addressed via the use of the templates and business rules summarised in sections 3 and 4, respectively, is provided in Supporting Information File 4. It should be noted that not all of these challenges are specific to ISA-TAB-Nano, i.e., some of them might be encountered when collecting data from the literature using other formats, and by no means are all of these challenges specific to collection of data from the published literature, i.e., some of them might be encountered when trying to report primary experimental data according to the generic ISA-TAB-Nano specification. It should also be noted that not all of these challenges are necessarily within the scope of the generic ISA-TAB-Nano specification to resolve, e.g., the definition of appropriate minimum information criteria. The need to address these challenges informed the design of the templates discussed in section 3 and the accompanying business rules, summarised in section 4 and presented in full in Supporting Information File 4, which were applied for the purpose of data collection from the nanotoxicology literature within the NanoPUZZLES EU project. It should be noted that no claim is made that all of these challenges are perfectly addressed via use of the resources presented in the current publication. The strengths and weaknesses of the manner in which these issues are addressed via the templates and business rules developed within NanoPUZZLES are discussed in the context of the detailed explanation of these challenges, which is presented in Supporting Information File 4. In addition, some of these challenges are returned to in the context of considering notable limitations of the resources developed within NanoPUZZLES. These notable limitations are summarised in section 7 and discussed in detail in Supporting Information File 4.

3 NanoPUZZLES data collection templates

General overview of templates

These templates were developed within the NanoPUZZLES project [33] and were specifically designed for collection of nanotoxicology data from the literature to support analysis of

Table 1: Summary of challenges with the generic ISA-TAB-Nano specification which were addressed in the current work.

no.	challenge	Applicable, in principle, to any format rather than being specific to ISA-TAB or ISA-TAB-Nano?	Applicable to ISA-TAB?	Applicable to ISA-TAB-Nano?
1	Standardised reporting of stepwise sample preparation needs to be established.	×	×	×
2	Ambiguity exists regarding where different kinds of information should be recorded.	—	×	×
3	Standardised recording of imprecisely reported experimental variables and measurements is required.	×	×	×
4	Ambiguity exists regarding the creation of "Comment […]" fields.	×	×	×
5	Statistical terms need to be clearly defined.	×	×[a]	×[a]
6	Ambiguity exists regarding how to link to terms from ontologies.	—	—	×
7	Ambiguity exists regarding whether or not "Parameter Value" or "Factor Value" column entries must be constant or not constant respectively.	—	×	×
8	Linking to images reported in publications is challenging.	×	×	×
9	Standardised reporting of multiple component "characteristics", "factors", and "parameters" (e.g. mixtures) needs to be established.	—	×	×
10	A standardised means of linking multiple "external" files to a given Material file is required.	—	—	×
11	Greater clarity regarding the existence of "unused" factors, parameters and measurement names in the Investigation file is required.	—	×[a]	×
12	A standardised approach for dealing with "non-applicable" metadata is required.	×	×	×
13	The concept of an "investigation" should be more tightly defined for the purpose of collecting data from the literature.	—	—	×
14	Clearly defined minimum information criteria are required.	×	×	×

[a]It should be noted that ISA-TAB is not designed to record experimental measurements in Assay files, i.e., the "Measurement Value [*statistic*(*measurement name*)]" Assay file columns and the corresponding Investigation file "Study Assay Measurement Name" field are an ISA-TAB-Nano extension [17,37,39]. However, regarding the issue of clearly defining statistical terms (challenge no. 5), ISA-TAB datasets may include "external" data files (i.e., "external" to the basic Investigation, Study and Assay file types) such as "data matrix" files which may include statistical terms such as "p-value" [36,43]. Standardisation of statistical terms may be achieved via using terms from the STATistics Ontology (STATO) [44]. The challenge noted here (challenge no. 5) regarding clearly defining statistical terms concerns how to appropriately create links to ontologies for these terms in ISA-TAB-Nano datasets.

trends and the development of data driven computational models such as nano-QSARs. These templates are available from the myExperiment online repository [45,46]: file entry "NanoPUZZLES ISA-TAB-Nano Templates" [47]. Version 3 of this file entry corresponds to the version of the templates referred to in the current publication and any corrections and/or extensions of these templates will also be made publicly available via future versions of this file entry.

The motivation for employing non-generic templates, designed to record specific kinds of (meta)data of interest to specific researchers, as opposed to generic templates that merely indicate the kinds of fields which the four ISA-TAB-Nano file types (Investigation, Study, Assay, Material) can contain, is that specific files with specific fields would need to be created at the point of data collection in any case but creating these specific files "on-the-fly" (i.e., at the point of data collection) is problematic. For example, a generic Assay file template would only indicate that certain, unspecified, experimental variables and endpoint values should be recorded using "Parameter Value […]" (or other column type such as "Factor Value […]") and "Measurement Value […]" columns, respectively. However, when collecting certain kinds of data obtained with a given assay, a specific Assay file with specific "Measurement Value

[…]" and "Parameter Value […]"columns (or other column types such as "Factor Value […]") would need to be created to record the (meta)data of interest. Indeed, the Investigation file is designed to associate a given "Study Assay Measurement Type" (e.g., size) and "Study Assay Technology Type" (e.g., dynamic light scattering) with a given "Study Assay File Name". Hence, specific templates (such as those developed in the current work) serve two important purposes: (a) they avoid the end user having to decide which specific fields, of a given type, should be created to record specific items of (meta)data; (b) they communicate to the end user which items of (meta)data should be reported in the dataset, i.e., they effectively define minimum information criteria. However, in case the specific templates do not capture all the experimental (meta)data of interest to a given end user of the dataset, it is important to recognise that the templates may be updated with new fields (in existing templates) or additional specific templates may be created.

The templates developed in the current work were adapted from generic Excel templates made available by the ISA-TAB-Nano developers [23]. The templates presented in this publication are designed to be compatible with version 1.2 of the ISA-TAB-Nano specification [23]. The generic templates were adapted as follows.

1. Predefined "Comment […]" fields were added to the Investigation file template for recording additional important metadata, e.g., "Comment [GLP]" for recording whether or not the corresponding studies were carried out according to Good Laboratory Practice [27,48].

2. Two specific Study file templates were created for sample preparation prior to physiochemical or cell based in vitro assays. (A Study file for sample preparation prior to in vivo assays was under development at the time of writing.)

3. Specific Assay file templates were created for (a) different kinds of physiochemical measurements and, in some cases, (b) for specific assays which might be employed to make those measurements. In some cases, where scenario (b) was not applicable, generic "Measurement Value [*statistic*(*measurement name*)]" columns were created with the *statistic* and/or *measurement name* presented as a generic "[TO DO: ….]" label: these labels should be replaced, as required, with specific *statistic* and *measurement name* values during data collection (as documented in the templates) or columns with these generic headings should be deleted if not applicable. For example, an Assay file template was designed for recording size measurements from a non-predetermined assay type ("a_InvID_PC_size_Method.xls") in addition to some Assay file templates for recording size measurements obtained using specific assay types - such as dynamic light scattering (DLS) ("a_InvID_PC_size_DLS.xls") [41,42]. The former template ("a_InvID_PC_size_Method.xls") includes the column "Measurement Value [[TO DO: appropriate average]([TO DO: appropriate size measurement])]": this would be updated to "Measurement Value [mean of the number distribution(diameter)]", to give but one possible example, during dataset creation. The latter template ("a_InvID_PC_size_DLS.xls") includes the columns "Measurement Value [z-average (hydrodynamic diameter)]" and "Measurement Value [polydispersity index]".

4. Specific Assay file templates were created for recording toxicity data for endpoints that were prioritised within the NanoPUZZLES project.

5. Predefined "Characteristics […]", "Factor Value […]" and "Parameter Value […]" columns were added to these Study and Assay file templates based upon consideration of which experimental variables were expected to affect the associated assay measurements. For example, the Study template for cell based in vitro studies ("s_InvID_InVitro.CB.xls") includes the predefined columns "Characteristics [cell type {EFO:http://www.ebi.ac.uk/efo/EFO_0000324}]" and "Factor Value [exposure medium]".

6. Predefined "Characteristics […]" columns were added to the Material file template for recording important chemical composition information, beyond that specified in the generic templates, along with nominal/vendor supplied values of various other physicochemical parameters, e.g., "Characteristics [Product impurities found {MEDDRA:http://purl.bioontology.org/ontology/MDR/10069178}]", "Characteristics [Major crystalline phase]" and "Characteristics [average size]".

7. Predefined "Comment […]" columns were added to the Material, Study and Assay file templates for recording key metadata that could (a) assist in interpreting the results or (b) allow the quality of the results to be assessed. For example, the template "a_InvID_PC_size_TEM.xls" for recording size by transmission electron microscopy (TEM) contains the columns "Comment [primary particle measurements]" and "Comment [size: from graph]" to address requirements of type (a) and (b) respectively. The "Comment [primary particle measurements]" column was designed to report whether or not the size measurements obtained were explicitly stated, in the publication from which they were extracted, to have been made for the primary particles: in principle, TEM might be used to provide information about agglomerates, aggregates or primary (individual) particles for a given prepared sample [49,50]. The "Comment [size: from graph]" column was predicated on the assumption

that data extracted from graphs (which are not uncommon when collecting data from the literature) are less reliable (i.e. more prone to transcription errors) than data extracted from tables or text.

8. For some fields, drop-down lists with possible field entries were created using the "Data Validation" option in Excel 2010.

9. The fields were colour coded to indicate those fields which were judged to be essential (green), desirable (yellow) or not important for the purposes of the NanoPUZZLES project (red).

10. Some fields (e.g., the Material file "Material Design Rationale" column) which were not considered important for the purposes of the NanoPUZZLES project were simply deleted.

11. Detailed comments were added (via the Excel 2010 "Review" tab) describing how different predefined fields should be populated during data collection.

12. The fields in the Investigation template ("i_InvID.xls") were populated insofar as possible prior to data collection. This included specifying predefined "factors" and "parameters" (c.f. other templates) and defining a set of ontologies from which terms should (preferentially) be obtained during data collection.

13. Some of the fields in the templates were populated with *indicated* values where appropriate. In some cases, these indications might actually be literally entered as values for the corresponding field entries, e.g., "size determination by DLS" entered in the first row of the "Protocol REF" column in the "a_InvID_PC_size_DLS.xls" template. However, in other cases, the suggested entries should not be entered literally, e.g., "size determination by <Assay technology type>" entered in the first row of the "Protocol REF" column in the "a_InvID_PC_size_Method.xls" template, where "<Assay technology type>" would be replaced with the name of the relevant method, such as "environmental scanning electron microscopy" [51,52] for the Assay file ("a_TOY.article_PC_size_ESEM.xls") in the "Toy Dataset" (see section 6) derived from the template "a_InvID_PC_size_Method.xls".

14. NanoPUZZLES specific naming conventions were established (as suggestions, rather than business rules) for creating files based on these templates. For example, "InvID" denotes "Investigation Identifier" and "Method" denotes an assay measurement technique such as dynamic light scattering (DLS).

15. A new "ImageLink" template was created ("ImageLink_NUMBER_for_InvID.xls") for linking to images reported in publications which are not associated with a single file that can be redistributed as part of a dataset or uniform resource identifier (URI). The use of this template is defined by NanoPUZZLES business rule no. 18 (see section 4 and Supporting Information File 4).

Identification of important experimental variables and characterisation data

The experimental variables (for both toxicological and physicochemical assays) and types of physicochemical characterisation data which the templates were designed to capture were based upon considering the well-known MINChar Initiative Parameters List [53], the provisional recommendations developed within the NanoSafety Cluster Databases Working Group [26], other resources developed within the context of the NanoSafety Cluster projects PreNanoTox [54] and MARINA [55] as well as discussions with nanotoxicology researchers and consideration of the published literature regarding toxicologically significant physicochemical characterisation parameters (for nanomaterials) and experimental variables which could significantly affect toxicological or physicochemical measurements [10,12,49,56-63]. However, no claim is made that the templates developed to date within the NanoPUZZLES project would capture all of the experimental variables or relevant characterisation information indicated by the cited proposals or otherwise recognised as important in the nanotoxicology community.

Physicochemical characterisation data captured by the templates

The categories of physicochemical information these templates were designed to capture, along with the corresponding Material and/or Assay file templates, are summarised in Table 2. In keeping with the generic ISA-TAB-Nano specification (version 1.2) [64], information which could be recorded using an Assay file template ("a_.....xls") should only be recorded using the Material file template ("m_MaterialSourceName.xls") if its value was nominal or vendor supplied.

These categories of physicochemical information correspond to all of the kinds of physicochemical information highlighted as being important in the MINChar Initiative Parameters List [53], with the context dependence stressed by this initiative being (partially) captured via recording sample conditions using "Factor Value […]" columns in the physicochemical Study file template ("s_InvID_PC.xls"), e.g., "Factor Value [medium]".

In order to construct these templates, careful consideration was required of exactly how to record different kinds of physicochemical information highlighted as being important. Firstly, this required consideration of which measurements might correspond to different kinds of physicochemical information;

Table 2: Categories of physicochemical information which the NanoPUZZLES ISA-TAB-Nano templates were designed to capture.

category	template(s)	comments
chemical composition (including surface composition, purity and levels of impurities)	"m_MaterialSourceName.xls"	Only chemical composition information associated with the original / vendor supplied nanomaterial should be reported here, i.e., not adsorption data (see below).
crystal structure/ crystallinity	"m_MaterialSourceName.xls"; "a_InvID_PC_crystallinity_Method.xls"	—
shape	"m_MaterialSourceName.xls"; "a_InvID_PC_shape_Method.xls"	Both qualitative descriptions of shape or "aspect ratio" data [60] can be recorded.
particle size/ size distribution	"m_MaterialSourceName.xls"; "a_InvID_PC_size_Method.xls"; "a_InvID_PC_size_DLS.xls"; "a_InvID_PC_size_TEM.xls"	Dynamic light scattering (DLS) [41] or transmission electron microscopy (TEM) [65,66] measurements are captured using the indicated Assay file templates. Otherwise, unless size values are nominal/vendor supplied, size measurements are captured via the generic Assay file template.
surface area	"m_MaterialSourceName.xls"; "a_InvID_PC_surface area_Method.xls"	This was designed to record "specific surface area" values, i.e., surface area per unit mass [58].
surface charge/ zeta potential	"m_MaterialSourceName.xls"; "a_InvID_PC_zetapotential_Method.xls"	Zeta potential is commonly used as a proxy for surface charge [58].
adsorption	"a_InvID_PC_adsorption_Method.xls"	This was designed to record "adsorption constants" [57] and (equilibrium) adsorption percentages [67] for specific small molecule / macromolecular "probe" species.
reactivity	"a_InvID_PC_reactivity.rateofchange_of.X_SeparationTechnique_Method.xls"	The design of this template reflects the fact that, for some reactivity assays, the analysed species needs to be removed prior to making measurements [68].
dissolution	(1) "a_InvID_PC_dissolution.conc_of.X_SeparationTechnique_Method.xls" ; (2) "a_InvID_PC_dissolution.fraction-dissolved_SeparationTechnique_Method.xls"; (3) "a_InvID_PC_dissolution.rate_of.X_SeparationTechnique_Method.xls"	The design of these templates reflects the fact that a number of different kinds of dissolution measurement may be made for inorganic nanoparticles: (1) the (time dependent) concentrations of various species released by dissolution [67,69] (which may be a redox process [69]); (2) the (time dependent) percentage of original nanoparticles dissolved [70]; (3) the (time dependent) dissolution rate [71]. The design of these templates further reflects the fact that dissolution assay protocols typically employ a separation step to isolate the analysed species [61].
molecular solubility	"a_InvID_PC_solubility_Method.xls"	In the current context, the Chemical Methods Ontology definition of "solubility" [72] was used: "the concentration of a solute in a saturated solution". This Assay template was specifically designed for recording molecular "solubility" measurements, e.g., the solubility of fullerene nanoparticles [73].
agglomeration/ aggregation	"a_InvID_PC_AAN_BETapproach.xls"	This template was designed for recording the "average agglomeration number" derived from BET gas adsorption data, size measurements and particle density values [58,74]. However, it should be noted that recording of size information obtained under different experimental conditions (using the Assay file templates noted above) may also convey information about the agglomeration state [58]. In addition, a number of physicochemical Assay files (e.g. "a_InvID_PC_size_Method.xls") contain "Comment […]" columns (e.g., "Comment[primary particle measurements]") designed to record whether or not the reported data are noted to refer to the primary particles (as opposed to agglomerates and/or aggregates) by the authors of the reference from which the data were extracted.
hydrophobicity	"m_MaterialSourceName.xls"; "a_InvID_PC_logP_Method.xls"	—

the "minimum" characterisation parameters reported in various proposals [12,53] are sometimes quite broadly defined, e.g., "Surface Chemistry, including reactivity, hydrophobicity" [53]. Secondly, this required consideration of which corresponding Material file "Characteristics […]" and/or Assay file "Measurement Value […]" columns needed to be defined - as well as, in some cases, which "Parameter Value […]" columns needed to be defined, e.g., "Parameter Value [analyte role]" (i.e., the

dissolved species being measured) for dissolution Assay file templates. No claim is made that the templates developed to date within the NanoPUZZLES project would capture all relevant measurements which might be associated with a given category of physicochemical information listed in Table 2.

Experimental variables captured by the templates

The experimental variables associated with sample preparation prior to applying assay protocols for (1) physicochemical measurements (see above) or (2) cell based in vitro toxicological assays are principally described via "Factor Value […]" columns in two Study file templates: (1) "s_InvID_PC.xls", (2) "s_InvID_InVitro.CB.xls".

For physicochemical studies, these "Factor Value […]" columns record the values of experimental variables associated with the preparation of a nanomaterial sample prior to application of an assay protocol, e.g., "Factor Value [physical state]" (for recording whether or not the sample was prepared as a suspension or a powder), "Factor Value [medium]" (for recording the suspension medium, i.e., not applicable if the "physical state" is a powder), "Factor Value [Sonication]" (for recording whether or not the sample was sonicated [49]).

For cell-based in vitro studies, these "Factor Value […]" columns record the values of experimental variables associated with preparation of the composite sample being tested, i.e., the nanomaterial suspension and the biological component on which the effect of the nanomaterial will be evaluated. Hence, they are designed to capture different kinds of experimental variables: (1) those which are relevant to preparation of the biological sample prior to adding the nanomaterial, e.g., the "Factor Value [culture medium glucose supplement]" in "s_InvID_InVitro.CB.xls" designed to record whether or not the cells were grown in glucose containing "culture medium", which may significantly affect the observed toxicity in some in vitro assays [56]; (2) those which are relevant to the preparation of the nanomaterial sample applied to the biological sample, e.g., "Factor Value [exposure medium]" and "Factor Value [Sonication]" for capturing the "exposure medium" for an in vitro (cell-based) study (otherwise known as the "exposure media" [75,76], i.e., the liquid mixture via which the tested chemical – a nanomaterial in the current context - reaches the cells) and whether or not sonication was applied to the tested nanomaterial suspension respectively; (3) those which are relevant to the combined sample to which the assay protocol is applied, e.g., "Factor Value [cells Exposure Duration]".

Capturing of the experimental conditions under which corresponding physicochemical characterisation and toxicity data were generated is important to assess whether or not characterisation was performed under biologically relevant conditions [77]. For example, whether or not a given size measurement was performed in the same suspension medium used for an in vitro (cell-based) study might be determined via comparing the "Factor Value [medium]" and "Factor Value [exposure medium]" entries in the physicochemical and in vitro (cell-based) Study files, respectively. However, details regarding possible suspension medium additives – such as serum and dispersant aids [78] – would need to be compared with each other by comparing the values in additional "Factor Value […]" fields.

In addition, for the "s_InvID_InVitro.CB.xls" Study file template, "Characteristics […]" columns associated with the "Source Name" column (i.e., positioned after the "Source Name" column but before the "Sample Name" column) are used to describe experimental variables which are inherent to the biological specimen: "cell type", "cell line", "organism" and "strain", as defined in the Experimental Factor Ontology (EFO) [79,80].

Experimental variables specifically associated with assay protocols are recorded in Assay files, principally using "Parameter Value […]" columns, e.g., "Parameter Value [Instrument]", "Parameter Value [negative control]".

It should be noted that the manner in which some of these experimental variables are captured via these templates might be carried out differently by other researchers and may deviate from the expectations of the generic ISA-TAB(-Nano) specification [17,23,36]. Some of the "Factor Value […]" columns (e.g., "Factor Value [physical state]" or "Factor Value [final cell density]" in "s_InvID_PC.xls" and "s_InvID_InVitro.CB.xls" respectively) might be considered to refer to characteristics of the prepared sample. Hence, these kinds of variables might elsewhere be recorded using "Characteristics […]" columns associated with the "Sample Name" column, i.e., positioned after the "Sample Name" column [36]. Other variables recorded via "Factor Value […]" columns (e.g., "Factor Value [Sonication Duration]") might be kept constant in some experiments [81], hence could be considered protocol parameters which would be recorded using "Parameter Value […]" columns [17]. However, the use of "Factor Value […]" columns to record these latter variables was deemed appropriate to account for scenarios in which these variables (e.g., sonication duration) were varied to assess their effect on assay measurements [49]. The fact that certain kinds of variables might be considered, in keeping with the generic ISA-TAB-Nano specification [17] discussed in section 1, "parameters" in one set of experiments and "factors" in another depending upon whether or not they were kept constant or varied to study their effects on the assay measure-

ment values does not lend itself to consistently organising these experimental variables in predefined template columns as developed in the current work.

The potential ambiguity associated with how to record different experimental variables can be illustrated by considering differences between the NanoPUZZLES ISA-TAB-Nano [47] and ToxBank ISA-TAB templates [82,83]: (1) the NanoPUZZLES Study file template "s_InvID_InVitro.CB.xls" contains the column "Factor Value [exposure medium]" for describing the suspension medium via which a tested nanomaterial is applied to the cells in an in vitro study, whereas the ToxBank Study file template "studySample.xml" contains the column "Characteristics[vehicle]" for describing the medium used to dilute a tested compound in an in vitro, in vivo or ex vivo study; (2) the NanoPUZZLES Assay file templates treat the identity of assay controls as "Parameter Value […]" entries (e.g., "Parameter Value [negative control]"), whereas the ToxBank Study file template uses a "Characteristics […]" column ("Characteristics[control]") to assign negative or positive control status to different samples.

Toxicity data captured by the templates

Assay file templates were developed to capture toxicity data associated with two toxicological endpoints which were initially prioritised within the NanoPUZZLES project: cytotoxicity ("a_InvID_cytotoxicity.cell-viability_Method.xls", "a_InvID_cytotoxicity.sub-lethal_Method.xls") and genotoxicity ("a_InvID_genotoxicity_Method.xls"). Cytotoxicity and genotoxicity are amongst the endpoints which are frequently considered when evaluating metal oxide nanoparticles in cell-based in vitro assays [4,84]. A number of nano-QSAR models have been developed for cytotoxicity [13,85-91] and some models have also been developed for nanomaterial genotoxicity [9,92,93].

The genotoxicity Assay file template ("a_InvID_genotoxicity_Method.xls") was designed to capture the most important outputs from different kinds of genotoxicity tests. Specifically, the "Parameter Value [Biomarker]" was designed to record the, test specific, biomarker whose increase relative to control values ("Measurement Value [mean(increase in biomarker level)]") would be determined for nanomaterial exposed samples. For example, "Parameter Value [Biomarker]" might report "micronuclei" or "number of revertants" if the method employed was the micronucleus test [94] or Ames test [95,96] respectively.

Since the results obtained for different sample preparation conditions (e.g., different tested concentrations) are usually used to derive an overall genotoxicity study call (i.e., "positive", "negative" or "equivocal") [94,96], a corresponding "Measurement Value [study call]" was added. Values in this latter column should be associated with "derived sample" identifiers as introduced in NanoPUZZLES business rule no. 10 (see section 4 and Supporting Information File 4 for an in-depth explanation).

The lethal cytotoxicity Assay file template ("a_InvID_cytotoxicity.cell-viability_Method.xls") was designed to record data corresponding to a reduction in cell "viability" (typically interpreted as an increase in "cell death") obtained from cell based in vitro assays such as MTT, MTS, LDH, and colony forming unit (CFU) counting [97-99]. The "percent cytotoxicity" columns ("Measurement Value [mean(percent cytotoxicity)]", "Measurement Value [standard deviation(percent cytotoxicity)]") are designed to record the "percent cytotoxicity" (a measure of cell death relative to controls equal to 100 – "percent viability") [100] associated with specific sample preparations, i.e., a specific value for the administered concentration or dose [101]. Other "Measurement Value […]" columns were designed to record measures of cytotoxicity derived from dose (or concentration) response relationships: the lowest observed effect level (LOEL) [102] (used, in the current work, to denote the lowest concentration/dose at which significant cell death relative to controls is observed), the LC_{50} [103] and LD_{50} [104], i.e., the concentration and dose, respectively, which, in the current context, kills 50% of the treated cells relative to controls. Values in these latter columns should be associated with "derived sample" identifiers as introduced in NanoPUZZLES business rule no. 10 (see section 4 and Supporting Information File 4 for an in-depth explanation).

The sub-lethal cytotoxicity Assay file template ("a_InvID_cytotoxicity.sub-lethal_Method.xls") was designed to record data from cell based in vitro assays designed to detect sub-lethal phenomena which might be quantified in terms of changes in key biomarkers. For example, oxidative stress and inflammation might be detected via measuring the level of glutathione or various cytokine biomarkers respectively [97]. (These sub-lethal phenomena would not be considered "cytotoxicity" by all researchers [84].) The manner in which this template was designed to capture sub-lethal cytotoxicity data is similar to the design of the genotoxicity Assay file template discussed above: the "Parameter Value [Biomarker]" column entries would state, for example, "glutathione" (depending upon the assay), with "Measurement Value […]" columns recording the "increase in biomarker level" (relative to control) as well as the LOEL [102] if this is reported. Values in this latter column should be associated with "derived sample" identifiers as introduced in NanoPUZZLES business rule no. 10 (see section 4 and Supporting Information File 4 for an in-depth explanation).

4 NanoPUZZLES business rules

Within the NanoPUZZLES project [33], a number of project specific business rules were created for the purpose of specifying how the ISA-TAB-Nano templates described in section 3 should be populated with data from literature sources. As noted in section 2, and fully explained in Supporting Information File 4, some of these business rules were specifically designed to address challenges associated with the generic ISA-TAB-Nano specification. A summary of these business rules is provided in Table 3. Supporting Information File 4 presents detailed explanations of how these business rules should be applied and, where appropriate, considers their strengths and weaknesses compared to possible alternatives which might be applied in future work.

These new rules were applied in addition to the rules which are part of the generic ISA-TAB-Nano specification as of version 1.2 [17,23,36-40]. (The new rules took precedence over the generic specification in case of conflicts.) It should also be remembered that additional guidance on creating ISA-TAB-Nano datasets using these templates is provided in section 3 and that guidance on populating individual fields is provided in the Excel-created comments linked to specific column titles. Finally, in keeping with the generic specification, the Investigation file and all corresponding files (Study, Assay and Material files along with all external files when applicable), for a single dataset, were added to a single, flat compressed ZIP archive (see section 5).

5 NanoPUZZLES Python program to facilitate computational analysis and database submission

Excel-based ISA-TAB-Nano templates are presented in this publication and elsewhere [17,23]. However, ISA-TAB-Nano files (Investigation, Study, Assay, Material) are commonly implemented in tab-delimited text format [105], reflecting the fact that ISA-TAB-Nano is an extension of ISA-TAB and ISA-TAB is intended to be implemented using tab-delimited text files (Investigation, Study, Assay) [36]. The authors of the current publication are unaware of any software specifically designed for parsing ISA-TAB datasets [22,82,106], which might be extended to parse ISA-TAB-Nano datasets, or software specifically designed for parsing ISA-TAB-Nano datasets [107,108], which does not require the key file types (Investigation, Study, Assay and, for ISA-TAB-Nano, Material) to be represented in tab-delimited text format. This includes publicly available online resources recently developed within the context of the MODERN project [107]: an ISA-TAB-Nano dataset validator and "Nanomaterial Data Management System" ("nanoDMS") – with the latter program implementing a web-based, searchable database system which is able to, amongst other functionality, import validated ISA-TAB-Nano datasets [30,109,110].

To facilitate database submission and other computational analysis, a Python [111] program was written, within the context of the NanoPUZZLES project, to enable automated conversion of an ISA-TAB-Nano dataset prepared using Excel-based templates to a tab-delimited text version of this dataset. Specifically, this program was designed to take a flat, compressed ZIP archive (e.g., "Investigation Identifier.zip") containing Excel ("xls") versions of an Investigation file, plus corresponding Study, Assay and Material files, and convert this to a flat, compressed ZIP archive (e.g., "Investigation Identifier-txt.zip") containing tab-delimited text versions of these files. Any external Excel-based "xls" files (e.g., "ImageLink" files introduced in the current work) contained in the archive will also be converted to tab-delimited text files and other external files will be transferred to the new archive without modification.

The program has four Open Source dependencies: a Python interpreter [111] along with the xlrd, xlwt [112] and unicodecsv [113] Python modules. For the purposes of code development, Python version 2.7.3, xlrd version 0.93, xlwt version 0.7.5 and unicodecsv version 0.9.4 were employed. All code was tested on a platform running Windows 7. The program does not have a graphical user interface (GUI): input is specified from the command prompt, e.g., "python xls2txtISA.NANO.archive.py –i InvestigationID.zip". The source code and documentation are available via the "xls2txtISA.NANO.archive" project on GitHub [114]. Version 1.2 of the program is referred to in the current publication [115].

Figure 2 provides an overview of the functionality of the program. As part of converting from Excel-based to tab-delimited text versions of ISA-TAB-Nano files, this program carries out basic checks on the datasets (e.g., checking for the presence of at least one file of type Investigation, Study, Assay, Material) and attempts to correct for basic potential errors in the file contents (e.g., removing line endings inside field entries) which might be introduced when manually preparing ISA-TAB-Nano files using Excel templates. However, the program does not carry out any sophisticated "parsing" of the datasets, i.e., no attempt is made to interpret the data in terms of the meaning of individual fields or the contents of individual field entries. No checks are carried out on the consistency of different files. Issues such as case sensitivity, null values and special characters (beyond removing internal line endings) are not addressed. Nonetheless, by facilitating conversion to tab-delimited text format, this enables the datasets to be parsed via more sophisticated tools such as those developed for validating ISA-TAB-Nano datasets within the MODERN project [107,108].

Table 3: Summary of the NanoPUZZLES business rules.

business rule no.	short description
1	A new “investigation” (corresponding to a new dataset comprising a single Investigation file, a set of Study, Assay and Material files and any “external” files if applicable) should be created for each reference (e.g., journal article), unless that reference specifically states that additional information regarding experiments on the same original nanomaterial samples was reported in another reference.
2	The “Factor Value […]” columns in the Study file refer to those values which are applicable to the sample prepared immediately prior to application of an assay protocol.
3	If the entry for a “Characteristics […]”, “Factor Value […]” or “Parameter Value […]” column corresponds to multiple components (e.g., mixtures), record this as a semicolon (“;”) delimited list of the separate components.
4	If the entry for a “Characteristics […]”, “Factor Value […]” or “Parameter Value […]” column corresponds to multiple components, record the entries in corresponding columns as a semicolon (“;”) delimited list with the entries in the corresponding order.
5	Any intrinsic chemical composition information associated with a nanomaterial sample (as originally sourced) should be recorded using a Material file even if it is determined/confirmed using assay measurements reported in the publication from which the data were extracted.
6	Any suspension medium associated with the nanomaterial sample (as originally sourced) should only be described using a Material file “Material Description” column.
7	Any impurities should be described using entries in the relevant Material file “Characteristics [….]” columns.
8	Any original nanomaterial components, which are neither a suspension medium nor described as “impurities” in the reference from which the data are extracted, should be described using separate rows of the Material file as per the generic ISA-TAB-Nano specification.
9	All “Sample Name” values for “true samples” should have the following form: “s_[Study Identifier]_[x]”, e.g., “s_[Study Identifier]_1”[a]
10	Assay file “Measurement Value […]” column entries which correspond to concentration-response curve statistics, or similarly derived measures, should be associated with a “derived sample” identifier rather than a “true sample” identifier.
11	Imprecisely reported experimental variables should be reported using “Factor Value [statistic(original factor name)]” columns created “on-the-fly”.
12	Imprecisely reported measurement values should be reported using “Measurement Value [statistic(measurement name)]” columns created “on-the-fly”.
13	“Comment […]” columns (rows) can be added without restriction to a Study, Assay, Material (Investigation) file as long as they are appropriately positioned and as long as each new “Comment […]” column (row) has a unique name for a given file.
14	All “statistic” names must be entered in the corresponding Investigation file template “Comment [Statistic name]” row.
15	When linking to terms from ontologies, the “preferred name” should be selected and the full ID entered in the corresponding “Term Accession Number” field.
16	“Factor Value […]” column entries are allowed to be constant.
17	Only “Parameter Value […]” column entries associated with a given “Protocol REF” column entry in a Study or Assay file need to be constant.
18	Images should be linked to assay measurements using a new “ImageLink” file type, if the generic ISA-TAB-Nano approach cannot be applied.
19	Any nanomaterial structure representation files, which are not associated with specific Assay file “Measurement Value […]” entries, should be linked to the corresponding Material file using ZIP archives specified in the appropriate “Material Data File” column entry.
20	Empty “Factor Value […]”, “Parameter Value […]” or “Measurement Value […]” columns in Study or Assay files can be deleted without having to update the corresponding Investigation file “Study Protocol Parameters Name”, “Study Factor Name”, or “Study Assay Measurement Name” fields.
21	Non-applicable columns should be populated with “N/A” where this conveys information.
22	“Measurement Value [statistic(measurement name)]” columns in the templates which use a label of the form “[TO DO:…]” for the statistic or measurement name must either be updated, based on the kind of statistic and/or measurement name indicated by the label(s), or deleted.

[a]Here, the “[Study Identifier]” [37] is unique to the corresponding Study file and “[x]” denotes a numeric value which is specific to a given “true sample”, meaning a prepared sample corresponding to a specific set of experimental conditions, in contrast to the “derived sample” concept introduced in NanoPUZZLES business rule no. 10.

As well as the default behaviour of this program described above, two command line options were specifically introduced to enable submission of an ISA-TAB-Nano dataset developed using these Excel templates to a database developed using the nanoDMS software [30,107,109,110]. The first option (“-a”) truncates all ontology identifiers: at the time of writing, “.”

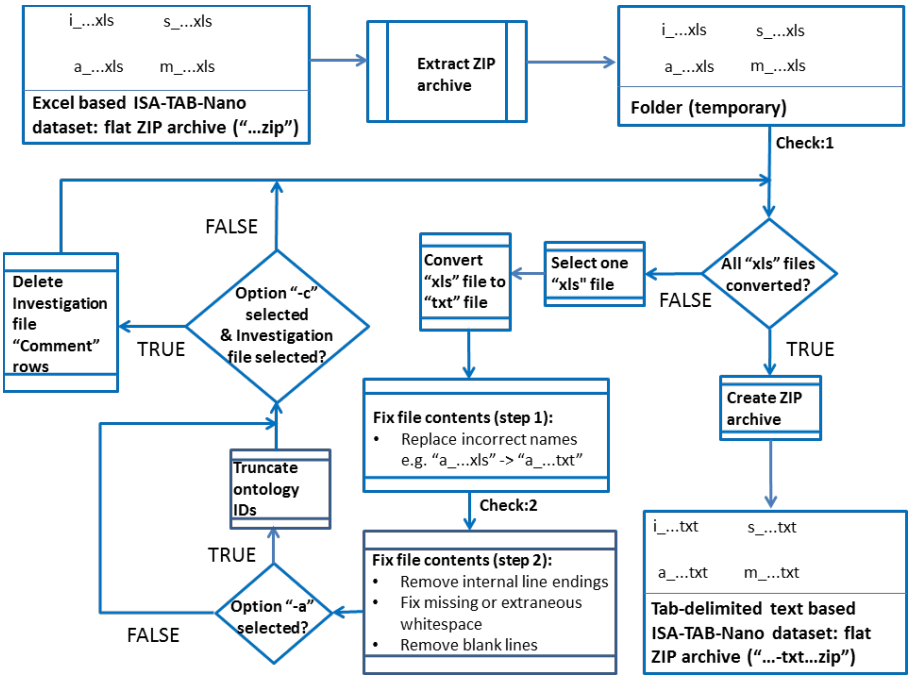


Figure 2: Schematic overview of the steps carried out by the Python program for converting Excel ("xls") based ISA-TAB-Nano datasets to tab-delimited text ("txt") based ISA-TAB-Nano datasets. For simplicity, only one Investigation, Study, Assay and Material file (and no external file such as an image) is included in this hypothetical dataset. In addition to the file processing steps summarised in this schematic, basic checks are carried out on the input: (1) there should be at least one Investigation, Study, Assay and Material file; (2) there should be no duplicate column titles in a Study, Assay or Material file other than those which are explicitly allowed by the ISA-TAB-Nano specification (e.g., "Unit").

characters were not permitted by the nanoDMS system in the headers of the Material, Study or Assay files, i.e., the column heading "Characteristics [shape {NPO:http://purl.bioontology.org/ontology/npo#NPO_274}]" in the Material files generated using the default options would need to be converted to "Characteristics [shape {NPO:NPO_274}]" etc. The second option ("-c") removes all "Comment […]" rows from the Investigation file: at the time of writing, these rows would also (indirectly) trigger errors when trying to load ISA-TAB-Nano datasets into the nanoDMS system. The output files are automatically named according to the options selected.

6 Toy dataset

In order to illustrate the use of all of the NanoPUZZLES template files, a "Toy Dataset" was created based upon these template files in accordance with the business rules summarised in section 4 and discussed in detail in Supporting Information File 4. It must be noted that the (meta)data contained within this "Toy Dataset" are not real, although they are based upon consideration of the nanoscience literature [4,49,51,57,58,60,61,67,68,70,71,73,74,97,116,117]. Indeed, no primary literature reports presenting data corresponding to all of the templates were identified as of the time of writing. An overview of the toy data content of this "Toy Dataset", generated after uploading this dataset into the nanoDMS database [110], is provided below in Figure 5 and Figure 6.

This "Toy Dataset" is available from the Supporting Information in three versions: Supporting Information File 1 corresponds to a flat archive containing files created using the original Excel templates and saved as "xls" files; Supporting Information File 2 is the version of this dataset created using the default options of the Python program described in section 5; Supporting Information File 3 was generated using the "-a" and "-c" flags of this software. This latter version (Supporting Information File 3) could be uploaded into the nanoDMS database [110], which is further discussed in section 7. The following figures provide an overview of the upload procedure for this dataset as well as illustrating the use of the nanoDMS system for retrieving these data: Figures 3–7.

7 Critical appraisal of the current work and possible future directions

Some notable limitations of the NanoPUZZLES templates and business rules introduced in this article

The strengths and weaknesses of the manner in which the challenges associated with the generic ISA-TAB-Nano specifica-

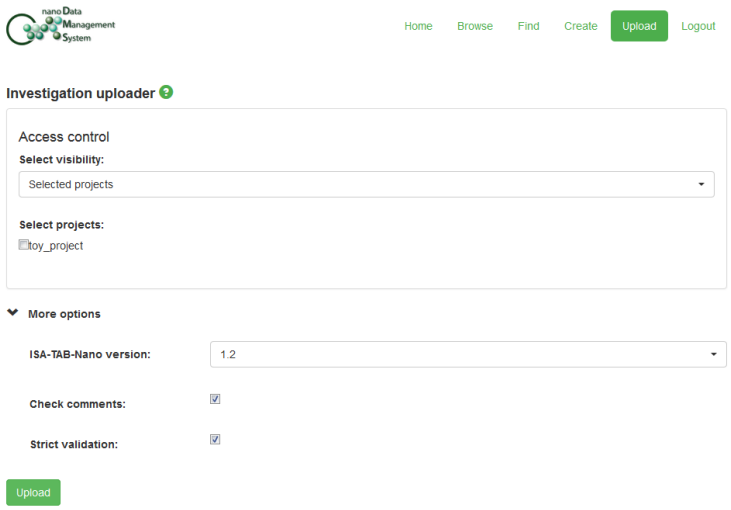


Figure 3: Upload options for loading the suitable version of the “Toy Dataset” (Supporting Information File 3) into the nanoDMS online database, which can be accessed via the cited web-address [110]: ontology identifiers were truncated and Investigation “Comment […]” rows deleted, using the Python program described in section 5, in order to enable this submission. Since these were not real data, the upload settings were selected such that the “Toy Dataset” was not publicly visible after uploading.

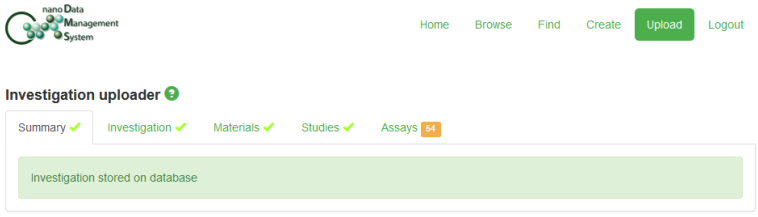


Figure 4: Confirmation that the “Toy Dataset” (Supporting Information File 3) was successfully uploaded: no error messages were generated by the internal ISA-TAB-Nano dataset validator and the warning messages regarding the position of the "Measurement Value [...]" and "Image File" columns reflect the addition of the “Measurement Value […]” column type to ISA-TAB-Nano, as compared to ISA-TAB, Assay files.

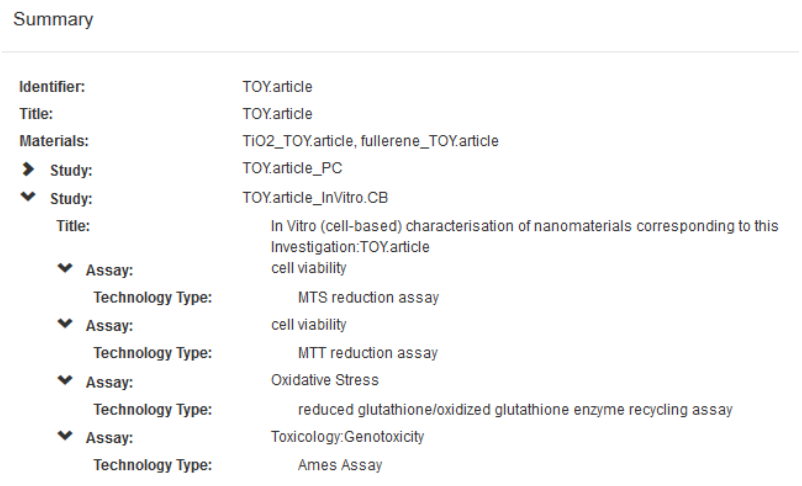


Figure 5: A summary of the in vitro cell-based assay toy data in the “Toy Dataset” (Supporting Information File 3) generated via the nanoDMS system. This summary can be generated via selecting the applicable dataset entry under the "Browse" menu of the nanoDMS system.

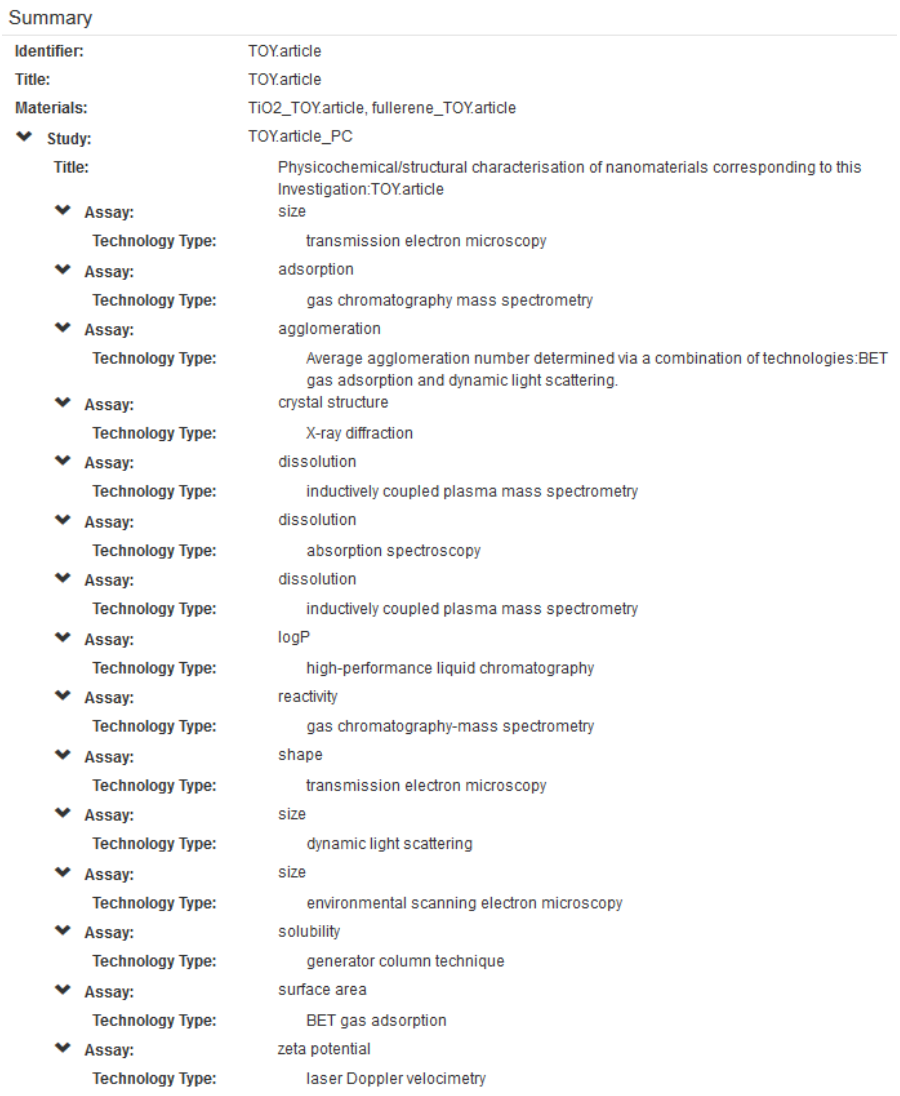


Figure 6: A summary of the physicochemical assay toy data recorded in the “Toy Dataset” (Supporting Information File 3), generated via the nanoDMS system as per Figure 5. This does not include the hypothetical chemical composition and nominal/vendor supplied data recorded in the Material files.

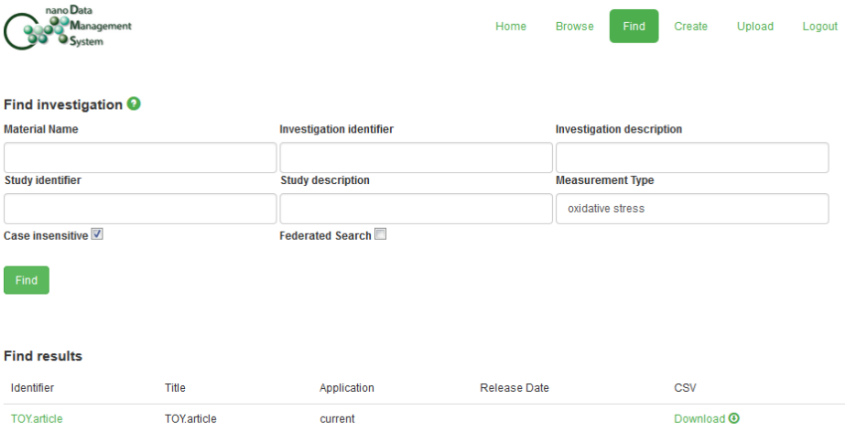


Figure 7: Retrieving the “Toy Dataset” (Supporting Information File 3) via searching for "oxidative stress" data in the nanoDMS system.

Table 4: Summary of some notable limitations of the NanoPUZZLES templates and business rules.

limitation no.	brief description
1	Standardised reporting of stepwise sample preparation is still not handled perfectly.
2	Time dependent physicochemical characterisation data may not be perfectly captured by the templates.
3	Recording of reaction rate constants and quantum yields may need revision.
4	The manner in which chemical composition information is captured via the templates may require revision.
5	There is the possibility of information loss when mapping (raw) data reported in the literature onto predefined "Measurement Value […]" columns.
6	The current templates are not best suited to capturing experimental data for all kinds of samples.
7	The business rules regarding multiple component "characteristics", "factors" or "parameters" (e.g., mixtures) may require revision.
8	The templates are not currently designed to capture data from in vivo toxicology studies.
9	Manually populating the Excel templates is time consuming and error prone.

tion (see section 2) were addressed via the templates and business rules developed within NanoPUZZLES are discussed in Supporting Information File 4. Beyond the need to address these general challenges, the specific strengths and weaknesses related to the design of the NanoPUZZLES templates (section 3) and business rules (section 4) were also discussed in section 3 and Supporting Information File 4, respectively. For example, it was noted in section 3 (under the "Experimental Variables Captured by the Templates" sub-section) that the manner in which certain experimental variables are recorded using the NanoPUZZLES templates may deviate from how other researchers would capture these metadata using ISA-TAB-Nano. Likewise, a possible alternative to the use of "derived sample" identifers (introduced in NanoPUZZLES business rule no. 10) for capturing concentration-response curve statistics, such as an LC_{50} [103], and related data is presented when discussing this business rule in Supporting Information File 4.

Table 4 summarises what are arguably the most notable remaining challenges associated with using these resources (templates and business rules) to collect nanotoxicology data from the literature. An in-depth discussion of these challenges, along with some suggestions for addressing them, is provided in Supporting Information File 4.

Integrating data collected using the NanoPUZZLES templates and business rules into databases

Various options currently exist, or are under development, for submitting the ISA-TAB-Nano files generated using the resources presented in sections 3, 4 and (if relevant) 5 to online, searchable databases. Submission to these databases should assist nano-QSAR researchers in identifying and retrieving data for modelling.

One option, as discussed previously, would be to submit the files to a database developed using the freely available "Nanomaterial Data Management System" ("nanoDMS") software [30,107-110] which was created within the context of the MODERN project. This database system was specifically designed to act as a searchable, online repository for ISA-TAB-Nano files and upload to the system is only allowed if the internal ISA-TAB-Nano dataset validator, also available as a standalone online tool [107], does not generate any error messages. An existing implementation of such a database was publicly available at the time of writing [110] and submission of a suitably prepared version of the "Toy Dataset" described in section 6 was successful (see Figures 3–7). However, as discussed in section 5 and section 6, this submission would currently require some modification of the datasets, i.e., some ontology identifers would need to be truncated and Investigation file "Comment […]" rows would need to be removed.

Another possible option would be to upload datasets generated using these resources into the eNanoMapper database [31,118,119]. This might be achieved via using the eNanoMapper customisable Excel spreadsheet parser to extract data from the Excel files created directly using the NanoPUZZLES templates [120]. Alternatively, it might be possible for an ISA-TAB-Nano parser (under development within eNanoMapper at the time of writing) to parse the tab-delimited text files generated using the program described in section 5. In either case the mapping of the input files onto the internal eNanoMapper data model would be performed in a transparent way, either explicitly via a JSON configuration file or implicitly by the ISA-TAB-Nano parser [31].

A brief illustration of some of the functionality of the nanoDMS database and its use for querying data generated using the NanoPUZZLES templates and business rules is presented in Figures 3–7. However, it should be noted that an in-depth discussion of the complete functionality of the nanoDMS and eNanoMapper databases is beyond the scope of the current

paper. Interested readers are referred to the cited references for further details regarding the nanoDMS [30,109,110] and eNanoMapper [31,118,119] databases.

Conclusion

There is a clear need to capture physicochemical and toxicological nanomaterial data in consistently organised electronic datasets which can be integrated into online, searchable databases to support predictive nanotoxicology. The generic ISA-TAB-Nano specification serves as a useful starting point for constructing such datasets but additional guidance regarding how to capture different kinds of (meta)data, as reported in the nanotoxicology literature, as well as exactly which (meta)data to record in these datasets is required. The publicly available resources presented in the current publication are proposed as means of (partially) addressing these requirements as well as facilitating the creation of ISA-TAB-Nano datasets. These resources are data collection templates, corresponding business rules which extend the generic ISA-TAB-Nano specification, and Python code to facilitate parsing of these datasets and integration of these datasets within other nanoinformatics resources. Nonetheless, various challenges remain with standardised collection of data from the nanotoxicology literature which these resources cannot be claimed to have definitively solved such as the need for standardised recording of stepwise sample preparation and temporal information as well as the wider need to achieve community consensus regarding minimum information standards. Extension of these resources by the nanoinformatics community, ideally working closely with the nanotoxicology community, is anticipated to enhance their value.

Supporting Information

Please note that in addition to the following Supporting Information files, which are versions of the "Toy Dataset" referred to in section 6, the templates and Python program described in this article are publicly available as previously explained [47,114,115].

Supporting Information File 1

"Toy Dataset" (i.e., not real data) created using the data collection templates.
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-202-S1.zip]

Supporting Information File 2

"Toy Dataset" converted using the Python program (default options).
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-202-S2.zip]

Supporting Information File 3

"Toy Dataset" converted using the Python program ("-a", "-c" options).
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-202-S3.zip]

Supporting Information File 4

Additional documentation and discussion.
[http://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-6-202-S4.pdf]

Author contributions

RLMR developed the NanoPUZZLES templates, business rules and Python program for generating text versions of datasets created with these templates. RLMR also prepared the "Toy Dataset" and wrote the first draft of this publication. MTDC and AR identified the need to use a standardised format, specifically ISA-TAB-Nano, within the NanoPUZZLES project and contributed to discussions regarding the kinds of data that should be recorded using the NanoPUZZLES templates. RR contributed to discussions regarding the use of ISA-TAB-Nano, oversaw the development of the ISA-TAB-Nano related tools within the MODERN project and provided feedback on the work carried out within NanoPUZZLES. All authors assisted with drafting the final manuscript.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreements no. 309837 (NanoPUZZLES project) and no. 309314 (MODERN project). RLMR thanks Sharon Gaheen (Leidos Biomedical Research Inc.), Nathan Baker (Pacific Northwest National Laboratory) and Nina Jeliazkova (IdeaConsult Ltd.) for valuable discussions regarding ISA-TAB and ISA-TAB-Nano as well as for commenting on an early draft of this manuscript. RLMR also thanks the ISA Team and Philippe Rocca-Serra (University of Oxford) for helpful correspondence regarding ISA-TAB and for commenting on parts of a draft of this manuscript. RLMR thanks Christoffer Åberg (University of Groningen), Neill Liptrott (University of Liverpool), Claire Mellor (Liverpool John Moores University) and Rafi Korenstein (Tel-Aviv University) for valuable discussions regarding the (nano)toxicology literature. RLMR also thanks Lang Tran and Peter Ritchie (Institute of Occupational Medicine), as well as Rafi Korenstein (Tel-Aviv University), for providing access to the MARINA project data collection templates and the PreNanoTox project database structure respectively, which partially informed the current work. Thanks are owed to Roger Pons and

Josep Cester (Universitat Rovira i Virgili) for their work on the ISA-TAB-Nano tools developed within the MODERN project. Thanks are also owed to the EU NanoSafety Cluster Databases Working Group and the US Nanotechnology Working Group for many valuable discussions.

References

1. Lövestam, G.; Rauscher, H.; Roebben, G.; Sokull Klüttgen, B.; Gibson, N.; Putaud, J.-P.; Stamm, H. *Considerations on a Definition of Nanomaterial for Regulatory Purposes, JRC Reference Reports;* European Commission Joint Research Centre, 2010.
2. Thomas, D. G.; Pappu, R. V.; Baker, N. A. *J. Biomed. Inf.* **2011,** *44,* 59–74. doi:10.1016/j.jbi.2010.03.001
3. Lynch, I. *Compendium of Projects in the European NanoSafety Cluster: 2014 Edition;* 2014.
4. Golbamaki, N.; Rasulev, B.; Cassano, A.; Marchese Robinson, R. L.; Benfenati, E.; Leszczynski, J.; Cronin, M. T. D. *Nanoscale* **2015,** *7,* 2154–2198. doi:10.1039/C4NR06670G
5. Rauscher, H.; Sokull-Klüttgen, B.; Stamm, H. *Nanotoxicology* **2012,** *7,* 1195–1197. doi:10.3109/17435390.2012.724724
6. Hendren, C. O.; Mesnard, X.; Dröge, J.; Wiesner, M. R. *Environ. Sci. Technol.* **2011,** *45,* 2562–2569. doi:10.1021/es103300g
7. Gajewicz, A.; Rasulev, B.; Dinadayalane, T. C.; Urbaszek, P.; Puzyn, T.; Leszczynska, D.; Leszczynski, J. *Adv. Drug Delivery Rev.* **2012,** *64,* 1663–1693. doi:10.1016/j.addr.2012.05.014
8. National Nanotechnology Initiative. http://www.nano.gov/ (accessed March 27, 2015).
9. Richarz, A.-N.; Madden, J. C.; Marchese Robinson, R. L.; Lubiński, Ł.; Mokshina, E.; Urbaszek, P.; Kuzmin, V. E.; Puzyn, T.; Cronin, M. T. D. *Perspect. Sci.* **2015,** *3,* 27–29. doi:10.1016/j.pisc.2014.11.015
10. Lynch, I.; Weiss, C.; Valsami-Jones, E. *Nano Today* **2014,** *9,* 266–270. doi:10.1016/j.nantod.2014.05.001
11. Thomas, D. G.; Klaessig, F.; Harper, S. L.; Fritts, M.; Hoover, M. D.; Gaheen, S.; Stokes, T. H.; Reznik-Zellen, R.; Freund, E. T.; Klemm, J. D.; Paik, D. S.; Baker, N. A. *Wiley Interdiscip. Rev.: Nanomed. Nanobiotechnol.* **2011,** *3,* 511–532. doi:10.1002/wnan.152
12. Stefaniak, A. B.; Hackley, V. A.; Roebben, G.; Ehara, K.; Hankin, S.; Postek, M. T.; Lynch, I.; Fu, W.-E.; Linsinger, T. P. J.; Thünemann, A. F. *Nanotoxicology* **2013,** *7,* 1325–1337. doi:10.3109/17435390.2012.739664
13. Puzyn, T.; Rasulev, B.; Gajewicz, A.; Hu, X.; Dasari, T. P.; Michalkova, A.; Hwang, H.-M.; Toropov, A.; Leszczynska, D.; Leszczynski, J. *Nat. Nanotechnol.* **2011,** *6,* 175–178. doi:10.1038/nnano.2011.10
14. *Guidance on Grouping of Chemicals, Second Edition;* Series on Testing & Assessment, Vol. 194; Organisation for Economic Co-operation and Development, 2014.
15. Gajewicz, A.; Cronin, M. T. D.; Rasulev, B.; Leszczynski, J.; Puzyn, T. *Nanotechnology* **2015,** *26,* 015701. doi:10.1088/0957-4484/26/1/015701
16. Baker, N. A.; Klemm, J. D.; Harper, S. L.; Gaheen, S.; Heiskanen, M.; Rocca-Serra, P.; Sansone, S.-A. *Nat. Nanotechnol.* **2013,** *8,* 73–74. doi:10.1038/nnano.2013.12
17. Thomas, D. G.; Gaheen, S.; Harper, S. L.; Fritts, M.; Klaessig, F.; Hahn-Dantona, E.; Paik, D.; Pan, S.; Stafford, G. A.; Freund, E. T.; Klemm, J. D.; Baker, N. A. *BMC Biotechnol.* **2013,** *13,* 2. doi:10.1186/1472-6750-13-2
18. Sansone, S.-A.; Rocca-Serra, P.; Field, D.; Maguire, E.; Taylor, C.; Hofmann, O.; Fang, H.; Neumann, S.; Tong, W.; Amaral-Zettler, L.; Begley, K.; Booth, T.; Bougueleret, L.; Burns, G.; Chapman, B.; Clark, T.; Coleman, L.-A.; Copeland, J.; Das, S.; de Daruvar, A.; de Matos, P.; Dix, I.; Edmunds, S.; Evelo, C. T.; Forster, M. J.; Gaudet, P.; Gilbert, J.; Goble, C.; Griffin, J. L.; Jacob, D.; Kleinjans, J.; Harland, L.; Haug, K.; Hermjakob, H.; Sui, S. J. H.; Laederach, A.; Liang, S.; Marshall, S.; McGrath, A.; Merrill, E.; Reilly, D.; Roux, M.; Shamu, C. E.; Shang, C. A.; Steinbeck, C.; Trefethen, A.; Williams-Jones, B.; Wolstencroft, K.; Xenarios, I.; Hide, W. *Nat. Genet.* **2012,** *44,* 121–126. doi:10.1038/ng.1054
19. Guzan, K. A.; Mills, K. C.; Gupta, V.; Murry, D.; Scheier, C. N.; Willis, D. A.; Ostraat, M. L. *Comput. Sci. Discovery* **2013,** *6,* 014007. doi:10.1088/1749-4699/6/1/014007
20. Sansone, S.-A.; Rocca-Serra, P.; Brandizi, M.; Brazma, A.; Field, D.; Fostel, J.; Garrow, A. G.; Gilbert, J.; Goodsaid, F.; Hardy, N.; Jones, P.; Lister, A.; Miller, M.; Morrison, N.; Rayner, T.; Sklyar, N.; Taylor, C.; Tong, W.; Warner, G.; Wiemann, S. *OMICS* **2008,** *12,* 143–149. doi:10.1089/omi.2008.0019
21. González-Beltrán, A.; Maguire, E.; Sansone, S.-A.; Rocca-Serra, P. *BMC Bioinf.* **2014,** *15* (Suppl. 14), S4. doi:10.1186/1471-2105-15-S14-S4
22. Rocca-Serra, P.; Brandizi, M.; Maguire, E.; Sklyar, N.; Taylor, C.; Begley, K.; Field, D.; Harris, S.; Hide, W.; Hofmann, O.; Neumann, S.; Sterk, P.; Tong, W.; Sansone, S.-A. *Bioinformatics* **2010,** *26,* 2354–2356. doi:10.1093/bioinformatics/btq415
23. ISA-TAB-Nano Wiki. https://wiki.nci.nih.gov/display/ICR/ISA-TAB-Nano (accessed March 27, 2015).
24. *Guidance Manual for the Testing of Manufactured Nanomaterials: OECD Sponsorship Programme, First Revision;* Series on the Safety of Manufactured Nanomaterials, Vol. 25; Organisation for Economic Co-operation and Development, 2010.
25. *Guidance on Sample Preparation and Dosimetry for the Safety Testing of Manufactured Nanomaterials;* Series on the Safety of Manufactured Nanomaterials, Vol. 36; Organisation for Economic Co-operation and Development, 2012.
26. Aberg, C. NanoSafety Cluster Databases Working Group. Overview and recommendation of data quality: Working draft. http://www.nanosafetycluster.eu/working-groups/4-database-wg/tasks-2/2013-2.html (accessed March 20, 2015).
27. Lubinski, L.; Urbaszek, P.; Gajewicz, A.; Cronin, M. T. D.; Enoch, S. J.; Madden, J. C.; Leszczynska, D.; Leszczynski, J.; Puzyn, T. *SAR QSAR Environ. Res.* **2013,** *24,* 995–1008. doi:10.1080/1062936X.2013.840679
28. Gaheen, S.; Hinkal, G. W.; Morris, S. A.; Lijowski, M.; Heiskanen, M.; Klemm, J. D. *Comput. Sci. Discovery* **2013,** *6,* 014010. doi:10.1088/1749-4699/6/1/014010
29. Mills, K. C.; Murry, D.; Guzan, K. A.; Ostraat, M. L. *J. Nanopart. Res.* **2014,** *16,* 2219. doi:10.1007/s11051-013-2219-8
30. Rallo, R. An ISA-TAB Nano compliant data management system for nanosafety modelling. https://nciphub.org/resources/500 (accessed March 27, 2015).
31. Jeliazkova, N.; Chomenidis, C.; Doganis, P.; Fadeel, B.; Grafström, R.; Hardy, B.; Hastings, J.; Hegi, M.; Jeliazkov, V.; Kochev, N.; Kohonen, P.; Munteanu, C. R.; Sarimveis, H.; Smeets, B.; Sopasakis, P.; Tsiliki, G.; Vorgrimmler, D.; Willighagen, E. *Beilstein J. Nanotechnol.* **2015,** *6,* 1609–1634. doi:10.3762/bjnano.6.165

32. Morris, S. A.; Gaheen, S.; Lijowski, M.; Heiskanen, M.; Klemm, J. *Beilstein J. Nanotechnol.* **2015,** *6,* 1580–1593. doi:10.3762/bjnano.6.161
33. NanoPUZZLES Project Homepage. http://www.nanopuzzles.eu (accessed July 15, 2015).
34. MODERN Project Homepage. http://modern-fp7.biocenit.cat/index.html (accessed March 28, 2015).
35. eNanoMapper Homepage. http://www.enanomapper.net/ (accessed March 28, 2015).
36. Rocca-Serra, P.; Sansone, S.-A.; Brandizi, M.; Hancock, D.; Harris, S.; Lister, A.; Miller, M.; O'Neill, K.; Taylor, C.; Tong, W. Specification documentation: release candidate 1, ISA-TAB 1.0. http://isatab.sourceforge.net/docs/ISA-TAB_release-candidate-1_v1.0_24nov08.pdf (accessed July 21, 2015).
37. ISA-TAB-Nano Wiki: Investigation File Documentation. https://wiki.nci.nih.gov/display/ICR/Investigation (accessed March 28, 2015).
38. ISA-TAB-Nano Wiki: Study File Documentation. https://wiki.nci.nih.gov/display/ICR/Study (accessed March 28, 2015).
39. ISA-TAB-Nano Wiki: Assay File Documentation. https://wiki.nci.nih.gov/display/ICR/Assay (accessed March 28, 2015).
40. ISA-TAB-Nano Wiki: Material File Documentation. https://wiki.nci.nih.gov/display/ICR/Material (accessed March 28, 2015).
41. *Dynamic light scattering - common terms defined (Whitepaper)*; Malvern Instruments Ltd., 2014.
42. Baalousha, M.; Lead, J. R. *Environ. Sci. Technol.* **2012,** *46,* 6134–6142. doi:10.1021/es301167x
43. Rayner, T. F.; Rocca-Serra, P.; Spellman, P. T.; Causton, H. C.; Farne, A.; Holloway, E.; Irizarry, R. A.; Liu, J.; Maier, D. S.; Miller, M.; Petersen, K.; Quackenbush, J.; Sherlock, G.; Stoeckert, C. J.; White, J.; Whetzel, P. L.; Wymore, F.; Parkinson, H.; Sarkans, U.; Ball, C. A.; Brazma, A. *BMC Bioinf.* **2006,** *7,* 489. doi:10.1186/1471-2105-7-489
44. STATistics Ontology (STATO) Homepage. http://stato-ontology.org/ (accessed Aug 3, 2015).
45. Goble, C. A.; Bhagat, J.; Aleksejevs, S.; Cruickshank, D.; Michaelides, D.; Newman, D.; Borkum, M.; Bechhofer, S.; Roos, M.; Li, P.; De Roure, D. *Nucleic Acids Res.* **2010,** *38* (Suppl. 2), W677–W682. doi:10.1093/nar/gkq429
46. myExperiment Homepage. http://www.myexperiment.org/home (accessed March 28, 2015).
47. NanoPUZZLES ISA-TAB-Nano Templates. http://www.myexperiment.org/files/1356.html (accessed Sept 17, 2015).
48. *OECD Principles on Good Laboratory Practice;* Organisation for Economic Co-operation and Development: Paris, France, 1998. doi:10.1787/9789264078536-en
49. Murdock, R. C.; Braydich-Stolle, L.; Schrand, A. M.; Schlager, J. J.; Hussain, S. M. *Toxicol. Sci.* **2008,** *101,* 239–253. doi:10.1093/toxsci/kfm240
50. Kaiser, D. L.; Watters, R. L. *National Institute of Standards and Technology (NIST) Certificate of Analysis for Standard Reference Material® 1898 Titanium Dioxide Nanomaterial;* National Institute of Standards and Technology: Gaithersburg, MD, U.S.A., 2012.
51. Han, B.; Lu, X. *Surf. Coat. Technol.* **2009,** *203,* 3656–3660. doi:10.1016/j.surfcoat.2009.05.046
52. Donald, A. M. *Nat. Mater.* **2003,** *2,* 511–516. doi:10.1038/nmat898
53. The Minimum Information for Nanomaterial Characterization (MINChar) Initiative Parameters List. https://characterizationmatters.wordpress.com/parameters/ (accessed March 28, 2015).
54. PreNanoTox Project Homepage. http://www.prenanotox.tau.ac.il/ (accessed March 28, 2015).
55. MARINA Project Homepage. http://www.marina-fp7.eu/ (accessed March 28, 2015).
56. Marroquin, L. D.; Hynes, J.; Dykens, J. A.; Jamieson, J. D.; Will, Y. *Toxicol. Sci.* **2007,** *97,* 539–547. doi:10.1093/toxsci/kfm052
57. Chen, R.; Zhang, Y.; Darabi Sahneh, F.; Scoglio, C. M.; Wohlleben, W.; Haase, A.; Monteiro-Riviere, N. A.; Riviere, J. E. *ACS Nano* **2014,** *8,* 9446–9456. doi:10.1021/nn503573s
58. Powers, K. W.; Brown, S. C.; Krishna, V. B.; Wasdo, S. C.; Moudgil, B. M.; Roberts, S. M. *Toxicol. Sci.* **2006,** *90,* 296–303. doi:10.1093/toxsci/kfj099
59. Powers, K. W.; Palazuelos, M.; Moudgil, B. M.; Roberts, S. M. *Nanotoxicology* **2007,** *1,* 42–51. doi:10.1080/17435390701314902
60. Donaldson, K.; Poland, C. A. *Curr. Opin. Biotechnol.* **2013,** *24,* 724–734. doi:10.1016/j.copbio.2013.05.003
61. Misra, S. K.; Dybowska, A.; Berhanu, D.; Luoma, S. N.; Valsami-Jones, E. *Sci. Total Environ.* **2012,** *438,* 225–232. doi:10.1016/j.scitotenv.2012.08.066
62. Hoffman, A. J.; Carraway, E. R.; Hoffmann, M. R. *Environ. Sci. Technol.* **1994,** *28,* 776–785. doi:10.1021/es00054a006
63. Pathakoti, K.; Huang, M.-J.; Watts, J. D.; He, X.; Hwang, H.-M. *J. Photochem. Photobiol., B* **2014,** *130,* 234–240. doi:10.1016/j.jphotobiol.2013.11.023
64. ISA-TAB-Nano 1.2 Release Notes. https://wiki.nci.nih.gov/display/ICR/ISA-TAB-Nano%201.2%20Release%20Notes (accessed March 28, 2015).
65. Handy, R. D.; van den Brink, N.; Chappell, M.; Mühling, M.; Behra, R.; Dušinská, M.; Simpson, P.; Ahtiainen, J.; Jha, A. N.; Seiter, J.; Bednar, A.; Kennedy, A.; Fernandes, T. F.; Riediker, M. *Ecotoxicology* **2012,** *21,* 933–972. doi:10.1007/s10646-012-0862-y
66. Ramachandramoorthy, R.; Bernal, R.; Espinosa, H. D. *ACS Nano* **2015,** *9,* 4675–4685. doi:10.1021/acsnano.5b01391
67. Thomas, A. J.; Kuhlbusch, H. F. K. *NanoCare: Health related aspects of nanomaterials (Final Scientific Report), Final Scientific Report;* NanoCare Project, 2009.
68. Hoffman, A. J.; Carraway, E. R.; Hoffmann, M. R. *Environ. Sci. Technol.* **1994,** *28,* 776–785. doi:10.1021/es00054a006
69. Jinnouchi, R.; Toyoda, E.; Hatanaka, T.; Morimoto, Y. *J. Phys. Chem. C* **2010,** *114,* 17557–17568. doi:10.1021/jp106593d
70. Meulenkamp, E. A. *J. Phys. Chem. B* **1998,** *102,* 7764–7769. doi:10.1021/jp982305u
71. Diedrich, T.; Dybowska, A.; Schott, J.; Valsami-Jones, E.; Oelkers, E. H. *Environ. Sci. Technol.* **2012,** *46,* 4909–4915. doi:10.1021/es2045053
72. Chemical Methods Ontology (CHMO): Solubility Definition. http://www.ontobee.org/browser/rdf.php?o=CHMO&iri=http://purl.obolibrary.org/obo/CHMO_0002815 (accessed March 28, 2015).
73. Jafvert, C. T.; Kulkarni, P. P. *Environ. Sci. Technol.* **2008,** *42,* 5945–5950. doi:10.1021/es702809a
74. Hackley, V. A.; Ferraris, C. F. *NIST Recommended Practice Guide: The Use of Nomenclature in Dispersion Science and Technology;* National Institute of Standards and Technology: Gaithersburg, MD, U.S.A., 2001.
75. Doskey, C. M.; van 't Erve, T. J.; Wagner, B. A.; Buettner, G. R. *PLoS One* **2015,** *10,* e0132572. doi:10.1371/journal.pone.0132572

76. Cohen, J. M.; Teeguarden, J. G.; Demokritou, P. *Part. Fibre Toxicol.* **2014,** *11,* 20. doi:10.1186/1743-8977-11-20
77. Crist, R. M.; Grossman, J. H.; Patri, A. K.; Stern, S. T.; Dobrovolskaia, M. A.; Adiseshaiah, P. P.; Clogston, J. D.; McNeil, S. E. *Integr. Biol.* **2013,** *5,* 66–73. doi:10.1039/C2IB20117H
78. Murdock, R. C.; Braydich-Stolle, L.; Schrand, A. M.; Schlager, J. J.; Hussain, S. M. *Toxicol. Sci.* **2008,** *101,* 239–253. doi:10.1093/toxsci/kfm240
79. Malone, J.; Holloway, E.; Adamusiak, T.; Kapushesky, M.; Zheng, J.; Kolesnikov, N.; Zhukova, A.; Brazma, A.; Parkinson, H. *Bioinformatics* **2010,** *26,* 1112–1118. doi:10.1093/bioinformatics/btq099
80. Experimental Factor Ontology Homepage. http://www.ebi.ac.uk/efo/ (accessed Aug 3, 2015).
81. Sayes, C.; Ivanov, I. *Risk Anal.* **2010,** *30,* 1723–1734. doi:10.1111/j.1539-6924.2010.01438.x
82. Kohonen, P.; Benfenati, E.; Bower, D.; Ceder, R.; Crump, M.; Cross, K.; Grafström, R. C.; Healy, L.; Helma, C.; Jeliazkova, N.; Jeliazkov, V.; Maggioni, S.; Miller, S.; Myatt, G.; Rautenberg, M.; Stacey, G.; Willighagen, E.; Wiseman, J.; Hardy, B. *Mol. Inf.* **2013,** *32,* 47–63. doi:10.1002/minf.201200114
83. ToxBank ISA-TAB Templates. https://github.com/ToxBank/isa2rdf/tree/master/isa2rdf/isa2rdf-cli/src/main/resources/toxbank-config (accessed March 28, 2015).
84. Nel, A. E. *J. Intern. Med.* **2013,** *274,* 561–577. doi:10.1111/joim.12109
85. Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Gini, G.; Puzyn, T.; Leszczynska, D.; Leszczynski, J. *Chemosphere* **2012,** *89,* 1098–1102. doi:10.1016/j.chemosphere.2012.05.077
86. Luan, F.; Kleandrova, V. V.; González-Díaz, H.; Ruso, J. M.; Melo, A.; Speck-Planche, A.; Cordeiro, M. N. D. S. *Nanoscale* **2014,** *6,* 10623–10630. doi:10.1039/C4NR01285B
87. Toropova, A. P.; Toropov, A. A.; Benfenati, E.; Korenstein, R. *J. Nanopart. Res.* **2014,** *16,* 2282. doi:10.1007/s11051-014-2282-9
88. Pathakoti, K.; Huang, M.-J.; Watts, J. D.; He, X.; Hwang, H.-M. *J. Photochem. Photobiol., B* **2014,** *130,* 234–240. doi:10.1016/j.jphotobiol.2013.11.023
89. Kleandrova, V. V.; Luan, F.; González-Díaz, H.; Ruso, J. M.; Speck-Planche, A.; Cordeiro, M. N. D. S. *Environ. Sci. Technol.* **2014,** *48,* 14686–14694. doi:10.1021/es503861x
90. Liu, R.; Rallo, R.; George, S.; Ji, Z.; Nair, S.; Nel, A. E.; Cohen, Y. *Small* **2011,** *7,* 1118–1126. doi:10.1002/smll.201002366
91. Toropova, A. P.; Toropov, A. A.; Rallo, R.; Leszczynska, D.; Leszczynski, J. *Ecotoxicol. Environ. Saf.* **2015,** *112,* 39–45. doi:10.1016/j.ecoenv.2014.10.003
92. Toropov, A. A.; Toropova, A. P. *Chemosphere* **2014,** *104,* 262–264. doi:10.1016/j.chemosphere.2013.10.079
93. Toropov, A. A.; Toropova, A. P. *Chemosphere* **2015,** *124,* 40–46. doi:10.1016/j.chemosphere.2014.10.067
94. OECD. *Test No. 487: In Vitro Mammalian Cell Micronucleus Test;* Organisation for Economic Co-operation and Development: Paris, France, 2014.
95. Doak, S. H.; Manshian, B.; Jenkins, G. J. S.; Singh, N. *Mutat. Res., Genet. Toxicol. Environ. Mutagen.* **2012,** *745,* 104–111. doi:10.1016/j.mrgentox.2011.09.013
96. OECD. *Test No. 471: Bacterial Reverse Mutation Test;* Organisation for Economic Co-operation and Development: Paris, France, 1997.
97. Lewinski, N.; Colvin, V.; Drezek, R. *Small* **2008,** *4,* 26–49. doi:10.1002/smll.200700595
98. Domey, J.; Haslauer, L.; Grau, I.; Strobel, C.; Kettering, M.; Hilger, I. Probing the Cytotoxicity of Nanoparticles: Experimental Pitfalls and Artifacts. Bioanalytical Reviews; Springer: Berlin, Germany, 2014; pp 1–14.
99. Thill, A.; Zeyons, O.; Spalla, O.; Chauvat, F.; Rose, J.; Auffan, M.; Flank, A. M. *Environ. Sci. Technol.* **2006,** *40,* 6151–6156. doi:10.1021/es060999b
100. BioAssay Ontology Percent Cytotoxicity Definition. http://bioportal.bioontology.org/ontologies/BAO?p=classes&conceptid=http%3A%2F%2Fwww.bioassayontology.org%2Fbao%23BAO_0000006 (accessed March 20, 2015).
101. Cohen, J. M.; Teeguarden, J. G.; Demokritou, P. *Part. Fibre Toxicol.* **2014,** *11,* 20. doi:10.1186/1743-8977-11-20
102. Lewis, R. W.; Billington, R.; Debryune, E.; Gamer, A.; Lang, B.; Carpanini, F. *Toxicol. Pathol.* **2002,** *30,* 66–74. doi:10.1080/01926230252824725
103. BioAssay Ontology LC50 Definition. http://bioportal.bioontology.org/ontologies/BAO?p=classes&conceptid=http%3A%2F%2Fwww.bioassayontology.org%2Fbao%23BAO_0002145 (accessed March 20, 2015).
104. BioAssay Ontology (BAO): LD50 Definition. http://bioportal.bioontology.org/ontologies/BAO?p=classes&conceptid=http%3A%2F%2Fwww.bioassayontology.org%2Fbao%23BAO_0002117 (accessed March 28, 2015).
105. ISA-TAB-Nano Wiki: Curated Examples. https://wiki.nci.nih.gov/display/ICR/ISA-TAB-Nano%20Curated%20Examples (accessed July 21, 2015).
106. ISA-Tools Software. http://www.isa-tools.org/software-suite/ (accessed July 21, 2015).
107. MODERN Project Web Applications. http://modern-fp7.biocenit.cat/tools.html (accessed March 28, 2015).
108. MODERN Project Software Downloads. http://nanodms.biocenit.cat/downloads.html (accessed July 21, 2015).
109. Pons, R.; Cester, J.; Giralt, F.; Rallo, R. *D2.1. MODERN Data Repository, European Union Seventh Framework Programme Project Deliverable Report D2.1;* 2014.
110. Nanomaterial Data Management System (nanoDMS). http://biocenitc-deq.urv.cat/nanodms (accessed Sept 11, 2015).
111. Python Programming Language. https://www.python.org/ (accessed March 28, 2015).
112. Working with Excel Files in Python. http://www.python-excel.org/ (accessed March 29, 2015).
113. Unicodecsv Python Module. https://pypi.python.org/pypi/unicodecsv (accessed March 29, 2015).
114. xls2txtISA.NANO.archive GitHub Repository. https://github.com/RichardLMR/xls2txtISA.NANO.archive (accessed June 30, 2015).
115. xls2txtISA.NANO.archive GitHub Repository: release version 1.2. https://github.com/RichardLMR/xls2txtISA.NANO.archive/releases/tag/v1.2 (accessed Sept 11, 2015).
116. Kim, J. A.; Åberg, C.; Salvati, A.; Dawson, K. A. *Nat. Nanotechnol.* **2012,** *7,* 62–68. doi:10.1038/nnano.2011.191
117. Shinohara, N.; Matsumoto, K.; Endoh, S.; Maru, J.; Nakanishi, J. *Toxicol. Lett.* **2009,** *191,* 289–296. doi:10.1016/j.toxlet.2009.09.012
118. Jeliazkova, N.; Doganis, P.; Fadeel, B.; Grafstrom, R.; Hastings, J.; Jeliazkov, V.; Kohonen, P.; Munteanu, C. R.; Sarimveis, H.; Smeets, B.; Tsiliki, G.; Vorgrimmler, D.; Willighagen, E. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* 2014; pp 1–9. doi:10.1109/BIBM.2014.6999367

119. eNanoMapper Database. http://data.enanomapper.net (accessed March 29, 2015).
120. eNanoMapper Parsers for Different NM Data Formats. https://github.com/enanomapper/nmdataparser (accessed March 29, 2015).

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Application of biclustering of gene expression data and gene set enrichment analysis methods to identify potentially disease causing nanomaterials

Andrew Williams* and Sabina Halappanavar

Full Research Paper

Open Access

Address:
Environmental Health Science and Research Bureau, Environmental and Radiation Health Sciences Directorate, Health Canada, Ottawa K1A 0K9, Canada

Email:
Andrew Williams* - andrew.williams@canada.ca

* Corresponding author

Keywords:
gene expression; risk assessment; toxicogenomics

Beilstein J. Nanotechnol. **2015,** *6,* 2438–2448.
doi:10.3762/bjnano.6.252

Received: 19 August 2015
Accepted: 30 November 2015
Published: 21 December 2015

This article is part of the Thematic Series "Nanoinformatics for environmental health and biomedicine".

Guest Editor: Y. Cohen

Abstract

Background: The presence of diverse types of nanomaterials (NMs) in commerce is growing at an exponential pace. As a result, human exposure to these materials in the environment is inevitable, necessitating the need for rapid and reliable toxicity testing methods to accurately assess the potential hazards associated with NMs. In this study, we applied biclustering and gene set enrichment analysis methods to derive essential features of altered lung transcriptome following exposure to NMs that are associated with lung-specific diseases. Several datasets from public microarray repositories describing pulmonary diseases in mouse models following exposure to a variety of substances were examined and functionally related biclusters of genes showing similar expression profiles were identified. The identified biclusters were then used to conduct a gene set enrichment analysis on pulmonary gene expression profiles derived from mice exposed to nano-titanium dioxide (nano-TiO_2), carbon black (CB) or carbon nanotubes (CNTs) to determine the disease significance of these data-driven gene sets.

Results: Biclusters representing inflammation (chemokine activity), DNA binding, cell cycle, apoptosis, reactive oxygen species (ROS) and fibrosis processes were identified. All of the NM studies were significant with respect to the bicluster related to chemokine activity (DAVID; FDR p-value = 0.032). The bicluster related to pulmonary fibrosis was enriched in studies where toxicity induced by CNT and CB studies was investigated, suggesting the potential for these materials to induce lung fibrosis. The profibrogenic potential of CNTs is well established. Although CB has not been shown to induce fibrosis, it induces stronger inflammatory, oxidative stress and DNA damage responses than nano-TiO_2 particles.

Conclusion: The results of the analysis correctly identified all NMs to be inflammogenic and only CB and CNTs as potentially fibrogenic. In addition to identifying several previously defined, functionally relevant gene sets, the present study also identified

two novel genes sets: a gene set associated with pulmonary fibrosis and a gene set associated with ROS, underlining the advantage of using a data-driven approach to identify novel, functionally related gene sets. The results can be used in future gene set enrichment analysis studies involving NMs or as features for clustering and classifying NMs of diverse properties.

Introduction

Metadata analysis that leverages genomics data has become increasingly popular as more experiments populate publicly available data repositories such as the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) and European Bioinformatics Institute (EBI; https://www.ebi.ac.uk/arrayexpress/). A systems biology approach through meta-analysis has the potential to reveal relationships and insight on resulting phenotypes that may not be possible to detect through the analysis of any individual experiment [1-12].

Conventional molecular approaches for the study of organismal response to toxicant exposures or diseases involve the study of one gene or a few genes at a time, whereas biological response is driven by a group of genes. Thus, when normal function of a specific biological process is perturbed, alterations and enrichment in the expression of a subset of co-functioning genes associated with that biological process are observed. Toxicogenomic tools such as gene expression profiling have become a widely used strategy for investigating the genome-wide changes relating to molecular mechanisms underlying many complex responses and diseases. The fact that genes interact with each other and are expressed in functionally relevant patterns implies that gene-expression data can be grouped into functionally meaningful gene sets across a subset of conditions [13-32]. The analysis of such predefined gene sets is a powerful alternative to individual gene analysis [13]. However, derivation of meaningful and relevant gene sets from the thousands of genes showing expression changes following exposure to toxicants is challenging.

Gene set data analysis, a computational technique which determines if a predefined set of genes exhibit statistically significant differential expression between two or more experimental conditions (time, dose, tissue, etc.), relies on the knowledge of annotated pathways relevant to the underlying physiology or biology being investigated. A survey conducted by Huang et al. [33] identified 68 different gene set enrichment tools. These methods are applied to manually and computationally curated [29] gene sets to identify enriched functional groupings of genes. These gene set enrichment tools include DAVID [21,22], EASE [34], GoMiner [35], MAPPFinder [36], Onto-express [37] and others, which consist of controlled descriptions of gene functions that are frequently used to define gene sets. Other tools, such as pathway databases including Gene Ontology [38], KEGG [39], BioCyc [40], TfactS [41], CTD [42], and BioCarta (http://www.biocarta.com), have also been applied in gene set analysis. Despite the number of tools available, the effective identification of functional groups of genes relevant to the underlying physiology across several conditions still remains a challenge. As a result, these tools continue to be refined and improved.

Nanomaterials (NMs) are materials manufactured on the nanoscale (1–100 nm) and are the building blocks of nanotechnology. On the nanoscale, materials exhibit unique size-associated properties (optical, magnetic, mechanical, thermodynamic, electrical, etc.), which are harnessed for use in various commercial applications [43]. Current applications of NMs include therapeutic applications (e.g., nanomedicine, drug delivery, diagnostics), agriculture, manufacturing, electronics, cosmetics, textiles, and environmental remediation and protection. Although NMs are synthesized from their corresponding, known, bulk chemical substances, owing to their distinct size-associated properties, their biological or toxicological behavior are often different from their analogous bulk compound. Because of their smaller size and large surface area, NMs are known to have increased ability to interact with cellular membranes, they can easily cross cellular barriers and penetrate deeper regions of tissue (such as the highly vascularized alveolar regions of lungs), and they exhibit increased toxicological activity as compared to the corresponding bulk material or comparatively large particles [43]. A variety of conventional toxicology tools have been assessed using both in vitro and in vivo models for their suitability and applicability for toxicity testing of NMs. However, these tools are single-endpoint-based or targeted in nature, investigate only one type of response at a time, and lack detailed mechanistic information [44]. Given the rate at which nanotechnology is growing, and the limitations of the currently available toxicological testing tools, it is estimated that it will take several decades and millions of dollars to complete the assessment of NMs of various sizes, shapes and surface coatings that require immediate assessment [45]. Therefore, more efficient toxicity testing and prediction tools are needed to provide a comprehensive overview of the biological activities of NMs to rapidly screen the toxicological potential of NMs.

Over the last few years, genome-wide expression analysis tools have been used as an alternative approach to comprehensively investigate the toxicological response induced by various

classes of NMs and to identify the properties of NMs that are responsible for eliciting adverse effects. We have previously used transcriptomics profiling tools to investigate the underlying mechanisms of toxicity induced by nanoparticles of titanium dioxide (nano-TiO_2) [46-48] and carbon nanotubes (CNTs) [49,50] of various sizes and properties. This work identified the properties of nano-TiO_2 that influence their inflammogenic potential [51]. These studies have generated a large repository of gene expression data that reflect the diversity of NM-induced biological response across a variety of experimental conditions. However, the challenge lies in the effective use of these data to discern individual or networks of genes conferring adverse outcomes of regulatory importance or disease phenotypes.

In the present study, we used a meta-analysis approach like that described by Turcan et al. [20] to identify functionally related biclusters of genes showing similar expression profiles, derived from publicly available gene expression data sets describing specific lung diseases (Table 1). One advantage of biclustering is that genes in the same cluster do not have to behave similarly over all experimental conditions. Unlike classical clustering techniques, biclusters can overlap with each other. This is ideal for mining functionally related gene sets as genes can be asso-

Table 1: Publically available datasets.

	GEO accession; reference	Platform	Disease model/nanomaterial
Lung disease models	GSE4231 [57]	UCSF 10Mm Mouse v.2 Oligo Array (GPL1089); UCSF GS Operon Mouse v.2 Oligo Array (GPL3330); UCSF 11Mm Mouse v.2 Oligo Array (GPL3331); UCSF 7Mm Mouse v.2 Oligo Array (GPL3359)	Lung inflammation models
	GSE6116 [58]	Affymetrix Mouse Genome 430 2.0 Array (GPL1261)	Biomarkers to predict female mouse lung tumors
	GSE6858 [59]	Affymetrix Mouse Genome 430 2.0 Array (GPL1261)	Model of experimental asthma
	GSE8790 [60]	Affymetrix Mouse Genome 430 2.0 Array (GPL1261)	Cigarette smoke-induced emphysema
	GSE11037 [11]	Agilent-011978 Mouse Microarray G4121A (GPL891)	Emphysema
	GSE18534 [61]	Affymetrix Mouse Genome 430 2.0 Array (GPL1261)	Mouse small cell lung cancer model
	GSE19605 [62]	Illumina MouseRef-8 v2.0 expression beadchip (GPL6885)	Lung carcinogenesis
	GSE25640 [63]	Affymetrix Mouse Genome 430 2.0 Array (GPL1261)	Pulmonary fibrosis
	GSE31013 [64]	Affymetrix Mouse Genome 430 2.0 Array (GPL1261)	Spontaneous lung tumors
	GSE40151 [65]	Affymetrix Mouse Genome 430 2.0 Array (GPL1261)	Idiopathic pulmonary fibrosis
	GSE42233 [66]	Illumina Mouse WG-6 v2.0 expression beadchip (GPL6887)	Lung cancer
	GSE52509 [67]	Illumina MouseRef-8 v2.0 expression beadchip (GPL6885)	COPD
NM studies	GSE29042 [68]	GPL4134 Agilent-014868 Whole Mouse Genome Microarray 4x44K G4122F	CNT: MWCNT-7
	GSE35193 [48]	GPL7202 Agilent-014868 Whole Mouse Genome Microarray 4x44K G4122F	CB: Printex 90
	GSE41041 [47]	GPL7202 Agilent-014868 Whole Mouse Genome Microarray 4x44K G4122F	TiO_2: UV-Titan L181
	GSE47000 [49]	GPL10787 Agilent-028005 SurePrint G3 Mouse GE 8x60K Microarray	CNT: Mitsui7
	GSE60801 [51]	GPL7202 Agilent-014868 Whole Mouse Genome Microarray 4x44K G4122F	TiO_2: NRCWE-025, NRCWE-030
	GSE60801 [51]	GPL7202 Agilent-014868 Whole Mouse Genome Microarray 4x44K G4122F	TiO_2 Sanding dust: Indoor-R, Indoornano TiO_2
	GSE60801 [51]	GPL7202 Agilent-014868 Whole Mouse Genome Microarray 4x44K G4122F	TiO_2: Sanding dust NRCWE-032, sanding dust NRCWE-033
	GSE60801 [51]	GPL7202 Agilent-014868 Whole Mouse Genome Microarray 4x44K G4122F	TiO_2: NRCWE 001 (no charge), NRCWE 002 (positively charged)
	GSE61366 [50]	GPL10787 Agilent-028005 SurePrint G3 Mouse GE 8x60K Microarray	CNT: NRCWE-26, NM-401

ciated with more than one biological process. Several studies [3,52-55] have shown that biclustering is a useful methodology to uncover processes that are active only over some but not all experimental conditions [56].

In this study, experiments investigating lung diseases (including lung inflammation, emphysema, chronic obstructive pulmonary disease (COPD) or lung cancer) in mice using the whole genome gene expression tools were obtained from GEO. For each study, raw data were downloaded from GEO and normalized as described in the methods below. Biological replicates for each of the experimental conditions were averaged. All studies were merged together and biclustering was employed. Through this analysis, ten biclusters representing ten functional gene sets were identified. Using DAVID [21,22], the biological functions associated with these biclusters were identified. Next, we applied these candidate gene sets/biclusters to nine, publically available, toxicogenomic gene expression studies (Table 1, published studies from our laboratory) to examine the toxicity induced by a variety of NMs (nano-TiO_2, CB and CNTs) to determine the disease significance of the altered gene expression profiles following exposure to NMs. The analysis was restricted to lung disease models since pulmonary response following NM exposure is well characterized.

Results and Discussion

Identification of biclusters of genes from lung disease models

To develop a data-driven view of the mouse lung response following exposure to NMs, publicly available genomic data from GEO that describe characteristic features of select lung diseases were leveraged. Eleven studies encompassing 52 experimental conditions with 8752 common gene symbols were assembled and specific gene sets were extracted using the repeated Bimax [69] biclustering method. A total of ten distinct biclusters were identified. The results of the biclustering are visually summarized in Figure 1.

Bicluster-1 consisted of studies investigating small cell lung carcinoma, spontaneous lung tumor, asthma and pulmonary fibrosis. This bicluster consisted of 19 gene symbols (C1qa, C3ar1, Cd68, Clec4n, Ctsk, Ect2, Fcgr3, Gp2, Igf1, Mmp12, Ms4a6d, Ms4a7, Pbk, Prc1, Saa3, Shcbp1, Spp1, Timp1 and Ube2c). Submitting these gene symbols into the DAVID functional annotation analysis tool (http://david.abcc.ncifcrf.gov) resulted in no significant gene ontology (GO). The top three ranked GO terms based on unadjusted p-values were acute inflammatory response (p-value = 0.0023), extracellular region (p-value = 0.0067) and extracellular region part (p-value = 0.0083). The lung disease models that comprised this bicluster were the model for human small cell lung carcinoma (GSE18534), spontaneous lung tumor (GSE31013), experimental asthma (GSE6858), active pulmonary fibrosis days 7, 14, and 21 (GSE40151) and pulmonary fibrosis (GSE25640).

The second bicluster consisted of twenty gene symbols (4632434I11Rik, Ccna2, Ccnb1, Ccnb2, Cdc20, Cdca8, Cldn4, Hells, Kif22, Mad2l1, Megf10, Melk, Msr1, Mx1, Plk4, Psat1, Rad51, Rrm2, Sprr1a and Uhrf1) with lung disease models such as a model for human small cell lung carcinoma, spontaneous lung tumor, chemical-induced lung carcinogenesis model from GSE6116 (1,5-naphthalenediamine; NAPD) and pulmonary fibrosis. Using DAVID, many GOs and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were found significant (FDR p-value < 0.05). Ten of the twenty gene symbols from this bicluster were elements of the cell cycle GO (FDR p-value = 4.9×10^{-6}) and five were part of the KEGG pathway (FDR p-value = 2.5×10^{-4}).

The bleomycin injury and the bacterial infection models (GSE4231), as well as lung disease models related to pulmonary fibrosis, constituted the third bicluster. This bicluster contained 17 gene symbols (Aif1, Ccl2, Ccl9, Ccr5, Cdkn1a, Chl1, Cxcl9, Cyp7b1, Ereg, Fcgr1, Mt2, Retnla, Sfn, Sfrp1, Slc26a4, Socs3, and Tnc). Nine of the seventeen gene symbols are part of the extracellular region GO (FDR p-value = 0.0056). Other significant GO terms included chemokine receptor binding (FDR p-value = 0.017), extracellular region part (FDR p-value = 0.021) and chemokine activity (FDR p-value = 0.032).

The fourth bicluster contained gene symbols associated with chromatin binding (Arid4b, Atrx, Cnot6, Ezh2, Glmn, Hif1a, Ncl, Npm1, Ofd1, Sdccag1, Ssb, Tfrc, Tpp2, Ttc3, Zfp386) (FDR p-value = 0.019). This bicluster contained lung disease models associated with chemical exposure to known lung carcinogens (NAPD, *N*-1-naphthyl)ethylenediamine dihydrochloride (NEDD), 2,3-benzofuran (BFUR)) (GSE6116), a model for human small cell lung carcinoma, spontaneous lung tumor and cigarette smoke-induced emphysema (GSE8790). Many of the gene symbols found in this bicluster are transcription factors involved in the gene expression regulation and are associated with one form of cancer or another.

The fifth bicluster consisted of 35 gene symbols (1700019G17Rik, Ap1m2, Arg1, Atic, Cdc6, Ckmt1, Cldn7, Ddit4, Fetub, Galnt2, Gatm, Grb7, H1f0, Hdac11, Ildr1, Mapk13, Mcm2, Mcm5, Mcm6, Mrps15, Nup50, Pgls, Plek2, Psmd8, Rbp4, Rfc4, Rgl3, Rrs1, Serpine1, Sh3yl1, Slc25a13, Slc39a11, Spata5, Tk1, and Tmprss4). The lung disease models that formed this bicluster included the model for human small

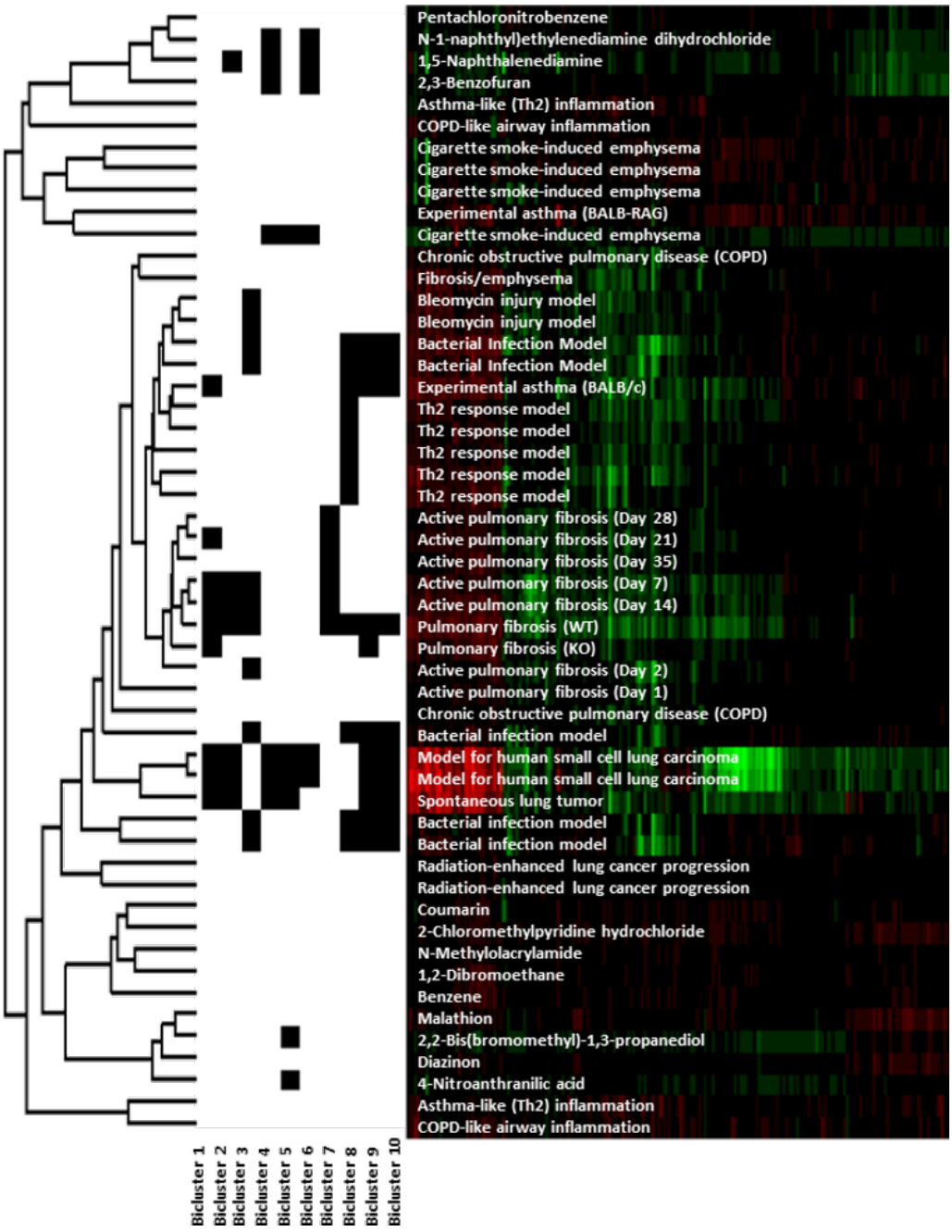


Figure 1: Heatmap of the gene symbols obtained from the bicluster data analysis. The distance metric used for the cluster analysis was 1-correlation estimated using Spearman correlation with average linkage.

cell lung carcinoma, spontaneous lung tumor, cigarette smoke-induced emphysema and two chemical exposures, 2,2-bis(bromomethyl)-1,3-propanediol (BBMP; lung carcinogen) and 4-nitroanthranilic acid (NAAC; which resulted in no observed tumors). DNA replication for the GO term (FDR p-value = 4.1×10^{-3}) and KEGG pathway (FDR p-value = 4.1×10^{-3}) were significant. The only other significant GO term was DNA replication initiation (FDR p-value = 0.028). A few genes showed association with matrix degradation, inflammation and energy metabolism.

The sixth bicluster consisted of models for human small cell lung carcinoma, cigarette smoke-induced emphysema and chemical exposures BFUR, NAPD and NEDD. DAVID annota-

tion analysis of the 23 gene symbols (Atm, Baz1b, Bclaf1, Ccar1, Dek, Dhx9, Epb4.1l3, F5, Hgf, Kif5b, Mier1, Pgm2l1, Plcb4, Ppil4, Rabep1, Smc1a, Stk3, Syncrip, Tcerg1, Ugcg, Usp9x, Zfml, and Zfp292) showed that this group of genes was primarily involved in the acetylation process (FDR p-value = 0.0056). Many GO terms related to the regulation of apoptosis were present in the results obtained by DAVID analysis. However, these results were not statistically significant after the FDR adjustment.

The seventh bicluster contained lung disease models related to pulmonary fibrosis only. DAVID analysis of the gene symbols included in this bicluster (Ccl3, Cd200r1, Chodl, Clec5a, Col24a1, Cxcl10, Emr1, Fxyd4, Gpnmb, Havcr2, Igj, Il1rn, Mmp10, Slc37a2, Syt12, Tgm1, Tlr8, Trem2, Wfdc12, and Zranb3) showed association with pulmonary fibrosis but no significant gene sets were derived. This bicluster can potentially serve as a candidate gene set for pulmonary fibrosis.

The eighth bicluster consisted of models for bacterial infection, Th2 response (GSE4231), asthma (GSE6858) and pulmonary fibrosis (GSE25640) with sixteen gene symbols (C1qb, Ch25h, Clec4a2, Ctss, F7, Fcgr2b, Itgam, Itgb2, Lgmn, Lpxn, Ly86, S100a4, Serpina3g, Serpina3n, Slc7a2, and Tbxas1). These gene symbols resulted in three significant GOs: response to wounding (FDR p-value = 0.0037), defense response (FDR p-value = 0.0063) and inflammatory response (FDR p-value = 0.0045).

The ninth bicluster consisted of the down-regulated gene symbols (Actc1, Cfd, Ckm, Ckmt2, Cox7a1, Cox8b, Csrp3, Eno3, Fmo3, Myh6, Myl1, Myl7, Pln, Pon1, Smpx, Sult1d1, Tnnc1, and Tnni3) and included a bacterial infection model, a model for human small cell lung carcinoma, spontaneous lung tumor, an asthma model and a pulmonary fibrosis model. These gene symbols were significantly associated with KEGG pathway cardiac muscle contraction (FDR p-value < 0.0001) and GO terms such as myosin complex (FDR p-value = 0.02) and regulation of system process (FDR p-value = 0.0015).

The tenth bicluster resulting from the analysis of the genes that were 2-fold down-regulated consisted of lung inflammation and disease models such as the bacterial infection model, a model for human small cell lung carcinoma, the study on spontaneous lung tumor, an asthma model and pulmonary fibrosis. This bicluster consisted of seventeen gene symbols (Aldh3a1, Bmp6, Cyp1a1, Cyp4b1, Eng, Fmo1, Fmo2, Gpr155, Igfbp6, Mapt, Ndrg2, Omd, Pcolce2, Pgam2, Scube2, Slc7a10, and Tnxb). These genes were associated with a variety of functions including fatty acid metabolism; however, DAVID functional annotation analysis of these gene symbols resulted in no statistically significant results to known annotated gene sets. However, several of these genes are associated with reactive oxygen species (ROS), which may not be a well-established gene set.

Application of biclusters to classify NM-induced lung response

Next, gene set enrichment analysis (GSEA) [29] using the bicluster-method-derived genes sets was conducted on the nine publically available studies [47-51,68] that examined NM-induced pulmonary toxicity. These results are presented in Figure 2. Bicluster-3 (genes associated with chemokine activity reflecting pulmonary inflammation) was enriched for most of the NMs. These results are in alignment with other studies in the literature that have shown pulmonary inflammation to be the predominant response following exposure to a variety of NMs. Bicluster-7 was the other significant cluster that was enriched in most of the experiments related to CNTs and CB. This cluster consisted of gene symbols showing strong association with pulmonary fibrosis. CNTs are well known to induce pulmonary fibrosis [50]. Although exposure to CB was not shown to cause lung fibrosis at the tested doses [48], studies have shown that CB exposure enhances bleomycin-induced lung fibrosis [70]. These results suggest that both carbon-based NMs may perturb similar biological processes and functions and factors in addition to the altered expression of a few genes in the gene set may contribute to the initiation of lung fibrosis.

Conclusion

In this study, we examined the applicability of a data-driven approach to identify gene sets from the comprehensive gene expression data using a biclustering method. The results showed that the lung response to NM exposure predominantly reflects responses observed following bacterial infections and bleomycin injury models that involve acute inflammation. The combined biclustering and gene set enrichment analysis also identified CNT and CB as potentially fibrogenic NMs. Although several genes sets associated with acute DNA binding, cell cycle, apoptosis, and ROS response that were specific to different disease models were also observed to be perturbed following exposure to NMs, the implication of such perturbation was not clear from this analysis. In addition the identification of several previously defined, functionally relevant gene sets, the present study also identified two novel genes sets: Bicluster-7 (consisting of genes associated with pulmonary fibrosis) and Bicluster-10 (consisting of genes associated with ROS), underlining the advantage of using a data-driven approach to identify novel, functionally related gene sets. The results can be used in future gene set enrichment analysis studies involving NMs or as features for clustering and classifying NMs of diverse properties.

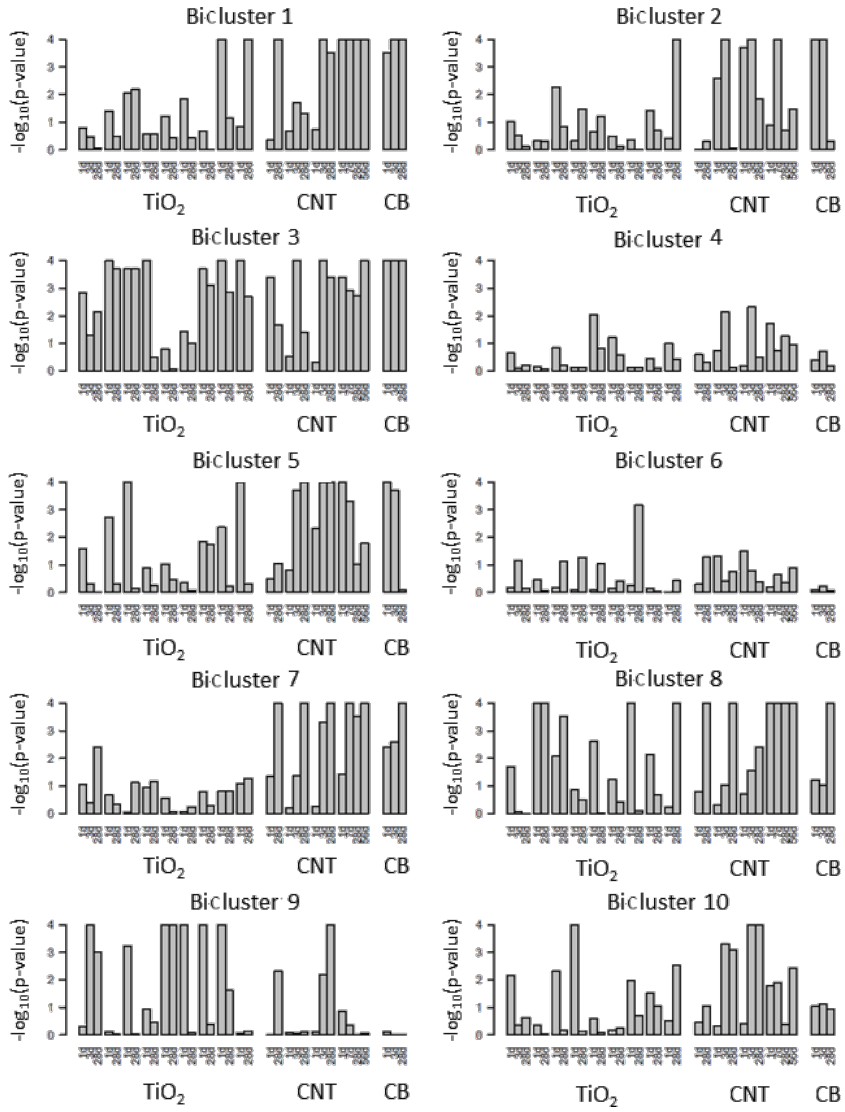


Figure 2: Gene set enrichment results of the NM datasets. Barplots of the $-\log_{10}$(p-value) from the GSEA are presented for each of the NM studies. The studies are ordered in the barplots as follows: TiO_2: UV-Titan L181, NRCWE-025, NRCWE-030, Sanding Dust Indoor-R, Sanding Dust Indoor-nano, Sanding dust NRCWE-032, Sanding dust NRCWE-03, NRCWE 001 (No charge), NRCWE 002 (positively charged); CNT: Mitsui7, NRCWE-26, NM-401, MWCNT-7; CB: Printex 90.

While powerful, data-driven meta-analysis approaches have several limitations. One important limitation is that the analysis is conditional on the subset of studies selected from the public data repositories such as GEO and EBI. Also, the experiments available in these repositories may not be representative of the population. For example, there are other mouse models of lung diseases that were not included in the present study due to lack of publicly available data or failure to meet the criteria set by the present study (time points, mouse strain, microarray platforms used).

The analysis is also limited to the gene symbols that were consistently investigated across the various microarray platforms from the different studies included in the analyses. Furthermore, the bicluster analysis is conditional to the two-fold change cut-off employed to create the binary matrix for the Bimax algorithm and the choice of the Bimax parameters. Modifying the fold cut-off to 1.75- and 1.5-fold, an additional 28 (23 up and 5 down) and 100 (89 up and 11 down) biclusters were identified. However, the interpretations were derived from the 2-fold cut-off as it provides the most conservative approach. The biclusters were stable when varying the minimum number of rows and when varying the minimum number of columns. Here, additional clusters were identified when these parameters were reduced and clusters were eliminated when these parameters were increased. Changes to any of the above

could impact the final results and therefore the interpretation of the data.

Experimental

Lung disease models

The data were obtained from the GEO. The accession numbers for the studies [11,57-67] used in the exploration of novel gene sets are presented in Table 1. These data sets cover a variety of lung diseases and lung injury outcomes, including different lung inflammation models, emphysema, chronic obstructive pulmonary disease and experiments studying lung cancer and lung tumors. Several different microarray platforms including the Illumina expression beadchip were used in these studies. The analysis was restricted to lung disease models since pulmonary responses following exposure to NMs are well characterized.

Data processing and normalization

The $\log_2$ transformation was applied to all signal intensity measurements. For the two color microarray studies, the LOWESS normalization method [71] using the R statistical software environment [72] was applied. For studies using the Affymetrix GeneChips®, the RMA normalization was applied using the justRMA function in the affy [73] R package. Quantile normalization was applied for studies that utilized the Illumina beadchip. This was done using the lumiN function in the lumi [74] R package.

Probes with technical replicates were then averaged using the median. The data for each study was then merged to its appropriate annotation file to obtain the gene symbol. Probes with the same gene symbol were then averaged using the median. The experimental conditions with biological replicates were averaged using the median. The median was used as it is a robust estimate of the central tendency.

For each experimental condition, the data was further normalized by centering to the matched control. The control samples were then removed from the data set. The remaining data is presented relative to the control, equivalently the $\log_2$ of the fold change (estimated using medians) for all the studies. The data were then merged across studies using the gene symbol. The mining the $\log_2$ of the fold changes was done in an attempt to minimize the cross-platform differences. However, platform differences may exist through compression of the fold-change values [75].

Biclustering

The biclustering data analysis was conducted in R using the biclust [69] package. The repeated Bimax [56] method was selected for this analysis. Bimax uses a simple data model that assumes two possible states for each expression level, no change and change with respect to a control experiment. For this analysis, two binary matrices were constructed: one matrix, consisting of zeros and ones, where the ones indicated genes that were 2-fold up-regulated and a second matrix, where the ones identify genes that were 2-fold down-regulated.

The option for the minimum number of rows for the Bimax method was set at 15. The minimum number of columns (which represent the experimental conditions) was set as 5 and the maximum number of columns was set as 15. This resulted in 8 biclusters from the binary matrix representing the up-regulated genes and 2 biclusters were identified for the matrix representing the down-regulated genes.

NM-induced lung response data sets

The data sets examining differential gene expression in mouse lung exposed to CB, nano-TiO_2 or CNTs were compiled from GEO. Since this is a proof-of-concept study, the investigation was limited to those NMs for which lung toxicological response is well characterized. Also, the genomics datasets with multiple doses and post-exposure time points were considered in the analysis. The GEO accession numbers for these studies are presented in Table 1. These studies utilized the two color Agilent microarray reference design [76]. The data were LOWESS normalized and probes with technical replicates were averaged. The annotation file containing the gene symbol was merged with the expression data and probes with multiple gene symbols were averaged using the median expression.

Gene set enrichment

As the NM-induced lung response data sets contained multiple doses, the test statistic from the Attract [19] approach was used. Using this method, the overall F-statistic for the dose effect was estimated for each gene. The F-statistics were then $\log_2$-transformed. A two sample t-test (assuming unequal variances) was then conducted, comparing the mean of the $\log_2$ F-statistics within the bicluster to the mean of the $\log_2$ F-statistics for all genes. The observed t-statistics and p-values are reported in Figure 2.

References

1. Hughes, T. R.; Marton, M. J.; Jones, A. R.; Roberts, C. J.; Stoughton, R.; Armour, C. D.; Bennett, H. A.; Coffey, E.; Dai, H.; He, Y. D.; Kidd, M. J.; King, A. M.; Meyer, M. R.; Slade, D.; Lum, P. Y.; Stepaniants, S. B.; Shoemaker, D. D.; Gachotte, D.; Chakraburtty, K.; Simon, J.; Bard, M.; Friend, S. H. *Cell* **2000,** *102,* 109–126. doi:10.1016/S0092-8674(00)00015-5
2. Rhodes, D. R.; Barrette, T. R.; Rubin, M. A.; Ghosh, D.; Chinnaiyan, A. M. *Cancer Res.* **2002,** *62,* 4427–4433.
3. Ihmels, J.; Friedlander, G.; Bergmann, S.; Sarig, O.; Ziv, Y.; Barkai, N. *Nat. Genet.* **2002,** *31,* 370–377. doi:10.1038/ng941

4. Ramaswamy, S.; Ross, K. N.; Lander, E. S.; Golub, T. R. *Nat. Genet.* **2002,** *33,* 49–54. doi:10.1038/ng1060
5. Xu, L.; Tan, A. C.; Naiman, D. Q.; Geman, D.; Winslow, R. L. *Bioinformatics* **2005,** *21,* 3905–3911. doi:10.1093/bioinformatics/bti647
6. Lamb, J.; Crawford, E. D.; Peck, D.; Modell, J. W.; Blat, I. C.; Wrobel, M. J.; Lerner, J.; Brunet, J. P.; Subramanian, A.; Ross, K. N.; Reich, M.; Hieronymus, H.; Wei, G.; Armstrong, S. A.; Haggarty, S. J.; Clemons, P. A.; Wei, R.; Carr, S. A.; Lander, E. S.; Golub, T. R. *Science* **2006,** *313,* 1929–1935. doi:10.1126/science.1132939
7. Hibbs, M. A.; Hess, D. C.; Myers, C. L.; Huttenhower, C.; Li, K.; Troyanskaya, O. G. *Bioinformatics* **2007,** *23,* 2692–2699. doi:10.1093/bioinformatics/btm403
8. Hassane, D. C.; Guzman, M. L.; Corbett, C.; Li, X.; Abboud, R.; Young, F.; Liesveld, J. L.; Carroll, M.; Jordan, C. T. *Blood* **2008,** *111,* 5654–5662. doi:10.1182/blood-2007-11-126003
9. Dudley, J. T.; Tibshirani, R.; Deshpande, T.; Butte, A. J. *Mol. Syst. Biol.* **2009,** *5,* 307. doi:10.1038/msb.2009.66
10. Daigle, B. J., Jr.; Deng, A.; McLaughlin, T.; Cushman, S. W.; Cam, M. C.; Reaven, G.; Tsao, P. S.; Altman, R. B. *PLoS Comput. Biol.* **2010,** *6,* e1000718. doi:10.1371/journal.pcbi.1000718
11. Thomson, E. M.; Williams, A.; Yauk, C. L.; Vincent, R. *Am. J. Pathol.* **2012,** *180,* 1413–1430. doi:10.1016/j.ajpath.2011.12.020
12. Jackson, A. F.; Williams, A.; Recio, L.; Waters, M. D.; Lambert, I. B.; Yauk, C. L. *Toxicol. Appl. Pharmacol.* **2014,** *274,* 63–77. doi:10.1016/j.taap.2013.10.019
13. Parfett, C.; Williams, A.; Zheng, J. L.; Zhou, G. *Regul. Toxicol. Pharmacol.* **2013,** *67,* 63–74. doi:10.1016/j.yrtph.2013.06.005
14. Barry, W. T.; Nobel, A. B.; Wright, F. A. *Bioinformatics* **2005,** *21,* 1943–1949. doi:10.1093/bioinformatics/bti260
15. Leung, Y. F.; Cavalieri, D. *Trends Genet.* **2003,** *19,* 649–659. doi:10.1016/j.tig.2003.09.015
16. Lu, Y.; Liu, P.-Y.; Xiao, P.; Deng, H.-W. *Bioinformatics* **2005,** *21,* 3105–3113. doi:10.1093/bioinformatics/bti496
17. Qin, H.; Feng, T.; Harding, S. A.; Tsai, C.-J.; Zhang, S. *Bioinformatics* **2008,** *24,* 1583–1589. doi:10.1093/bioinformatics/btn215
18. Tuglus, C.; van der Laan, M. J. *Stat. Appl. Genet. Mol. Biol.* **2009,** *8,* 1–15. doi:10.2202/1544-6115.1397
19. Mar, J. C.; Matigian, N. A.; Quackenbush, J.; Wells, C. A. *PLoS One* **2011,** *6,* e25445. doi:10.1371/journal.pone.0025445
20. Turcan, S.; Vetter, D. E.; Maron, J. L.; Wei, X.; Slonim, D. K. In *Proceedings of the Pacific Symposium,* Kohala Coast, Hawaii, Jan 3–7, 2011; World Scientific Publishing Co Pte Ltd: Singapore, 2011; pp 50–61. doi:10.1142/9789814335058_0006
21. Huang, D. W.; Sherman, B. T.; Tan, Q.; Collins, J. R.; Alvord, W. G.; Roayaei, J.; Stephens, R.; Baseler, M. W.; Lane, H. C.; Lempicki, R. A. *Genome Biol.* **2007,** *8,* R183. doi:10.1186/gb-2007-8-9-r183
22. Huang, D. W.; Sherman, B. T.; Tan, Q.; Kir, J.; Liu, D.; Bryant, D.; Guo, Y.; Stephens, R.; Baseler, M. W.; Lane, H. C.; Lempicki, R. A. *Nucleic Acids Res.* **2007,** *35,* W169–W175. doi:10.1093/nar/gkm415
23. Engreitz, J. M.; Daigle, B. J., Jr.; Marshall, J. J.; Altman, R. B. *J. Biomed. Inf.* **2010,** *43,* 932–944. doi:10.1016/j.jbi.2010.07.001
24. Huang, Y.; Li, H.; Hu, H.; Yan, X.; Waterman, M. S.; Huang, H.; Zhou, X. J. *Bioinformatics* **2007,** *23,* i222–i229. doi:10.1093/bioinformatics/btm222
25. Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D. *Proc. Natl. Acad. Sci. U. S. A.* **1998,** *95,* 14863–14868. doi:10.1073/pnas.95.25.14863
26. Spellman, P. T.; Sherlock, G.; Zhang, M. Q.; Iyer, V. R.; Anders, K.; Eisen, M. B.; Brown, P. O.; Botstein, D.; Futcher, B. *Mol. Biol. Cell* **1998,** *9,* 3273–3297. doi:10.1091/mbc.9.12.3273
27. Segal, E.; Shapira, M.; Regev, A.; Pe'er, D.; Botstein, D.; Koller, D.; Friedman, N. *Nat. Genet.* **2003,** *34,* 166–176. doi:10.1038/ng1165
28. Bar-Joseph, Z.; Gerber, G. K.; Lee, T. I.; Rinaldi, N. J.; Yoo, J. Y.; Robert, F.; Gordon, D. B.; Fraenkel, E.; Jaakkola, T. S.; Young, R. A.; Gifford, D. K. *Nat. Biotechnol.* **2003,** *21,* 1337–1342. doi:10.1038/nbt890
29. Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; Mesirov, J. P. *Proc. Natl. Acad. Sci. U. S. A.* **2005,** *102,* 15545–15550. doi:10.1073/pnas.0506580102
30. Ihmels, J.; Bergmann, S.; Barkai, N. *Bioinformatics* **2004,** *20,* 1993–2003. doi:10.1093/bioinformatics/bth166
31. Wang, X.; Dalkic, E.; Wu, M.; Chan, C. *Curr. Opin. Biotechnol.* **2008,** *19,* 482–491. doi:10.1016/j.copbio.2008.07.011
32. Fehrmann, R. S. N.; de Jonge, H. J. M.; ter Elst, A.; de Vries, A.; Crijns, A. G. P.; Weidenaar, A. C.; Gerbens, F.; de Jong, S.; van der Zee, A. G. J.; de Vries, E. G. E.; Kamps, W. A.; Hofstra, R. M. W.; te Meerman, G. J.; de Bont, E. S. J. M. *PLoS One* **2008,** *3,* e1656. doi:10.1371/journal.pone.0001656
33. Huang, D. W.; Sherman, B. T.; Lempicki, R. A. *Nucleic Acids Res.* **2009,** *37,* 1–13. doi:10.1093/nar/gkn923
34. Hosack, D. A.; Dennis, G.; Sherman, B. T.; Lane, H. C.; Lempicki, R. A. *Genome Biol.* **2003,** *4,* R70. doi:10.1186/gb-2003-4-10-r70
35. Zeeberg, B. R.; Feng, W.; Wang, G.; Wang, M. D.; Fojo, A. T.; Sunshine, M.; Narasimhan, S.; Kane, D. W.; Reinhold, W. C.; Lababidi, S.; Bussey, K. J.; Riss, J.; Barrett, J. C.; Weinstein, J. N. *Genome Biol.* **2003,** *4,* R28. doi:10.1186/gb-2003-4-4-r28
36. Doniger, S. W.; Salomonis, N.; Dahlquist, K. D.; Vranizan, K.; Lawlor, S. C.; Conklin, B. R. *Genome Biol.* **2003,** *4,* R7. doi:10.1186/gb-2003-4-1-r7
37. Draghici, S.; Khatri, P.; Bhavsar, P.; Shah, A.; Krawetz, S. A.; Tainsky, M. A. *Nucleic Acids Res.* **2003,** *31,* 3775–3781. doi:10.1093/nar/gkg624
38. Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. *Nat. Genet.* **2000,** *25,* 25–29. doi:10.1038/75556
39. Kanehisa, M.; Goto, S. *Nucleic Acids Res.* **2000,** *28,* 27–30. doi:10.1093/nar/28.1.27
40. Karp, P. D.; Ouzounis, C. A.; Moore-Kochlacs, C.; Goldovsky, L.; Kaipa, P.; Ahrén, D.; Tsoka, S.; Darzentas, N.; Kunin, V.; López-Bigas, N. *Nucleic Acids Res.* **2005,** *33,* 6083–6089. doi:10.1093/nar/gki892
41. Essaghir, A.; Toffalini, F.; Knoops, L.; Kallin, A.; van Helden, J.; Demoulin, J.-B. *Nucleic Acids Res.* **2010,** *38,* e120. doi:10.1093/nar/gkq149
42. Gohlke, J. M.; Thomas, R.; Zhang, Y.; Rosenstein, M. C.; Davis, A. P.; Murphy, C.; Becker, K. G.; Mattingly, C. J.; Portier, C. J. *BMC Syst. Biol.* **2009,** *3,* 46. doi:10.1186/1752-0509-3-46
43. Maynard, A. D.; Warheit, D. B.; Philbert, M. A. *J. Toxicol. Sci.* **2011,** *120,* S109–S129. doi:10.1093/toxsci/kfq372
44. Geraci, C. L.; Castranova, V. *Wiley Interdiscip. Rev.: Nanomed. Nanobiotechnol.* **2010,** *2,* 569–577. doi:10.1002/wnan.108
45. Lai, D. Y. *Wiley Interdiscip. Rev.: Nanomed. Nanobiotechnol.* **2012,** *4,* 1–15. doi:10.1002/wnan.162

46. Halappanavar, S.; Jackson, P.; Williams, A.; Jensen, K. A.; Hougaard, K. S.; Vogel, U.; Yauk, C. L.; Wallin, H. *Environ. Mol. Mutagen.* **2011,** *52,* 425–439. doi:10.1002/em.20639
47. Husain, M.; Saber, A. T.; Guo, C.; Jacobsen, N. R.; Jensen, K. A.; Yauk, C. L.; Williams, A.; Vogel, U.; Wallin, H.; Halappanavar, S. *Toxicol. Appl. Pharmacol.* **2013,** *269,* 250–262. doi:10.1016/j.taap.2013.03.018
48. Bourdon, J. A.; Halappanavar, S.; Saber, A. T.; Jacobsen, N. R.; Williams, A.; Wallin, H.; Vogel, U.; Yauk, C. L. *Toxicol. Sci.* **2012,** *127,* 474–484. doi:10.1093/toxsci/kfs119
49. Poulsen, S. S.; Jacobsen, N. R.; Labib, S.; Wu, D.; Husain, M.; Williams, A.; Bøgelund, J. P.; Andersen, O.; Købler, C.; Mølhave, K.; Kyjovska, Z. O.; Saber, A. T.; Wallin, H.; Yauk, C. L.; Vogel, U.; Halappanavar, S. *PLoS One* **2013,** *8,* e80452. doi:10.1371/journal.pone.0080452
50. Poulsen, S. S.; Saber, A. T.; Williams, A.; Andersen, O.; Købler, C.; Atluri, R.; Pozzebon, M. E.; Mucelli, S. P.; Simion, M.; Rickerby, D.; Mortensen, A.; Jackson, P.; Kyjovska, Z. O.; Mølhave, K.; Jacobsen, N. R.; Jensen, K. A.; Yauk, C. L.; Wallin, H.; Halappanavar, S.; Vogel, U. *Toxicol. Appl. Pharmacol.* **2015,** *284,* 16–32. doi:10.1016/j.taap.2014.12.011
51. Halappanavar, S.; Saber, A. T.; Decan, N.; Jensen, K. A.; Wu, D.; Jacobsen, N. R.; Guo, C.; Rogowski, J.; Koponen, I. K.; Levin, M.; Madsen, A. M.; Atluri, R.; Snitka, V.; Birkedal, R. K.; Rickerby, D.; Williams, A.; Wallin, H.; Yauk, C. L.; Vogel, U. *Environ. Mol. Mutagen.* **2015,** *56,* 245–264. doi:10.1002/em.21936
52. Cheng, Y.; Church, G. Biclustering of expression data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, ISMB 2000,* San Diego, CA, U.S.A., Aug 19–23, 2000; 2000; pp 93–103.
53. Ben-Dor, A.; Chor, B.; Karp, R.; Yakhini, Z. In *Proceedings of the 6th Annual International Conference on Computational Biology,* Washington, DC, April 18–21, 2002; ACM Press: New York, NY, U.S.A., 2002; pp 49–57. doi:10.1145/565196.565203
54. Tanay, A.; Sharan, R.; Kupiec, M.; Shamir, R. *Proc. Natl. Acad. Sci. U. S. A.* **2004,** *101,* 2981–2986. doi:10.1073/pnas.0308661100
55. Murali, T. M.; Kasif, S. Extracting conserved gene expression motifs from gene expression data. In *Pacific Symposium on Biocomputing,* Lihue, Hawaii, Jan 3–7, 2003; World Scientific Publishing Co Pte Ltd: Singapore, 2003; pp 77–88.
56. Prelić, A.; Bleuler, S.; Zimmermann, P.; Wille, A.; Bühlmann, P.; Gruissem, W.; Hennig, L.; Thiele, L.; Zitzler, E. *Bioinformatics* **2006,** *22,* 1122–1129. doi:10.1093/bioinformatics/btl060
57. Lewis, C. C.; Yang, J. Y. H.; Huang, X.; Banerjee, S. K.; Blackburn, M. R.; Baluk, P.; McDonald, D. M.; Blackwell, T. S.; Nagabhushanam, V.; Peters, W.; Voehringer, D.; Erle, D. J. *Am. J. Respir. Crit. Care Med.* **2008,** *177,* 376–387. doi:10.1164/rccm.200702-333OC
58. Thomas, R. S.; Pluta, L.; Yang, L.; Halsey, T. A. *Toxicol. Sci.* **2007,** *97,* 55–64. doi:10.1093/toxsci/kfm023
59. Lu, X.; Jain, V. V.; Finn, P. W.; Perkins, D. L. *Mol. Syst. Biol.* **2007,** *3,* 98. doi:10.1038/msb4100138
60. Rangasamy, T.; Misra, V.; Zhen, L.; Tankersley, C. G.; Tuder, R. M.; Biswal, S. *Am. J. Physiol.: Lung Cell. Mol. Physiol.* **2009,** *296,* L888–L900. doi:10.1152/ajplung.90369.2008
61. Schaffer, B. E.; Park, K.-S.; Yiu, G.; Conklin, J. F.; Lin, C.; Burkhart, D. L.; Karnezis, A. N.; Sweet-Cordero, E. A.; Sage, J. *Cancer Res.* **2010,** *70,* 3877–3883. doi:10.1158/0008-5472.CAN-09-4228
62. Ochoa, C. E.; Mirabolfathinejad, S. G.; Ruiz, V. A.; Evans, S. E.; Gagea, M.; Evans, C. M.; Dickey, B. F.; Moghaddam, S. J. *Cancer Prev. Res.* **2011,** *4,* 51–64. doi:10.1158/1940-6207.CAPR-10-0180
63. Liu, T.; Baek, H. A.; Yu, H.; Lee, H. J.; Park, B.-H.; Ullenbruch, M.; Liu, J.; Nakashima, T.; Choi, Y. Y.; Wu, G. D.; Chung, M. J.; Phan, S. H. *J. Immunol.* **2011,** *187,* 450–461. doi:10.4049/jimmunol.1000964
64. Pandiri, A. R.; Sills, R. C.; Ziglioli, V.; Ton, T.-V. T.; Hong, H.-H. L.; Lahousse, S. A.; Gerrish, K. E.; Auerbach, S. S.; Shockley, K. R.; Bushel, P. R.; Peddada, S. D.; Hoenerhoff, M. J. *Toxicol. Pathol.* **2012,** *40,* 1141–1159. doi:10.1177/0192623312447543
65. Peng, R.; Sridhar, S.; Tyagi, G.; Phillips, J. E.; Garrido, R.; Harris, P.; Burns, L.; Renteria, L.; Woods, J.; Chen, L.; Allard, J.; Ravindran, P.; Bitter, H.; Liang, Z.; Hogaboam, C. M.; Kitson, C.; Budd, D. C.; Fine, J. S.; Bauer, C. M.; Stevenson, C. S. *PLoS One* **2013,** *8,* No. e59348. doi:10.1371/journal.pone.0059348
66. Delgado, O.; Batten, K. G.; Richardson, J. A.; Xie, X.-J.; Gazdar, A. F.; Kaisani, A. A.; Girard, L.; Behrens, C.; Suraokar, M.; Fasciani, G.; Wright, W. E.; Story, M. D.; Wistuba, I. I.; Minna, J. D.; Shay, J. W. *Clin. Cancer Res.* **2014,** *20,* 1610–1622. doi:10.1158/1078-0432.CCR-13-2589
67. John-Schuster, G.; Hager, K.; Conlon, T. M.; Irmler, M.; Beckers, J.; Eickelberg, O.; Yildirim, A. Ö. *Am. J. Physiol.: Lung Cell. Mol. Physiol.* **2014,** *307,* L692–L706. doi:10.1152/ajplung.00092.2014
68. Guo, N. L.; Wan, Y.-W.; Denvir, J.; Porter, D. W.; Pacurari, M.; Wolfarth, M. G.; Castranova, V.; Qian, Y. *J. Toxicol. Environ. Health, Part A* **2012,** *75,* 1129–1153. doi:10.1080/15287394.2012.699852
69. Kaiser, S.; Santamaria, R.; Khamiakova, T.; Sill, M.; Theron, R.; Quintales, L.; Leisch, F.; De Troyer, E. biclust: BiCluster Algorithms, R package, Version 1.2.0. 2013; http://CRAN.R-project.org/package=biclust (accessed Dec 8, 2015).
70. Kamata, H.; Tasaka, S.; Inoue, K.-i.; Miyamoto, K.; Nakano, Y.; Shinoda, H.; Kimizuka, Y.; Fujiwara, H.; Ishii, M.; Hasegawa, N.; Takamiya, R.; Fujishima, S.; Takano, H.; Ishizaka, A. *Exp. Biol. Med.* **2011,** *236,* 315–324. doi:10.1258/ebm.2011.010180
71. Yang, Y. H.; Dudoit, S.; Luu, P.; Lin, D. M.; Peng, V.; Ngai, J.; Speed, T. P. *Nucleic Acids Res.* **2002,** *30,* e15. doi:10.1093/nar/30.4.e15
72. R Core Team. *R Foundation for Statistical Computing*, Vienna, Austria, 2014.
73. Gautier, L.; Cope, L.; Bolstad, B. M.; Irizarry, R. A. *Bioinformatics* **2004,** *20,* 307–315. doi:10.1093/bioinformatics/btg405
74. Du, P.; Kibbe, W. A.; Lin, S. M. *Bioinformatics* **2008,** *24,* 1547–1548. doi:10.1093/bioinformatics/btn224
75. Wang, Y.; Barbacioru, C.; Hyland, F.; Xiao, W.; Hunkapiller, K. L.; Blake, J.; Chan, F.; Gonzalez, C.; Zhang, L.; Samaha, R. R. *BMC Genomics* **2006,** *7,* 59. doi:10.1186/1471-2164-7-59
76. Kerr, M. K.; Churchill, G. A. *Genet. Res.* **2001,** *77,* 123–128. doi:10.1017/S0016672301005055