Nanoinformatics: spanning scales, systems and solutions

Edited by Iseult Lynch, Georgia Melagraki, Diego Martinez and Kunal Roy
Prediction of cytotoxicity of heavy metals adsorbed on nano-TiO₂ with periodic table descriptors using machine learning approaches

Joyita Roy, Souvik Pore and Kunal Roy*§

Full Research Paper

Address:
Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, 700032, India

Email:
Kunal Roy* - kunalroy_in@yahoo.com

* Corresponding author
§ Phone: +91 98315 94140; Fax: +91-33-2837-1078; kunal.roy@jadavpuruniversity.in

Keywords:
heavy metals; HK-2 cell; ML algorithm; periodic table descriptors; QSAR

Abstract

Nanoparticles with their unique features have attracted researchers over the past decades. Heavy metals, upon release and emission, may interact with different environmental components, which may lead to co-exposure to living organisms. Nanoscale titanium dioxide (nano-TiO₂) can adsorb heavy metals. The current idea is that nanoparticles (NPs) may act as carriers and facilitate the entry of heavy metals into organisms. Thus, the present study reports nanoscale quantitative structure–activity relationship (nano-QSAR) models, which are based on an ensemble learning approach, for predicting the cytotoxicity of heavy metals adsorbed on nano-TiO₂ to human renal cortex proximal tubule epithelial (HK-2) cells. The ensemble learning approach implements gradient boosting and bagging algorithms; that is, random forest, AdaBoost, Gradient Boost, and Extreme Gradient Boost were constructed and utilized to establish statistically significant relationships between the structural properties of NPs and the cause of cytotoxicity. To demonstrate the predictive ability of the developed nano-QSAR models, simple periodic table descriptors requiring low computational resources were utilized. The nano-QSAR models generated good $R^2$ values (0.99–0.89), $Q^2$ values (0.64–0.77), and $Q^2F_1$ values (0.99–0.71). Thus, the present work manifests that ML in conjunction with periodic table descriptors can be used to explore the features and predict unknown compounds with similar properties.

Introduction

Nanoparticles (NPs) have gained much attention due to their widespread applications in different areas, and they are continually designed to yield certain desired properties [1]. With the uninterrupted development of new NPs, engineered nanoparticles in the form of metal oxide nanoparticles are becoming a new area of research. Metal oxides have been used in different
industries, and the market is developing rapidly [2]. According to a recent analysis, approximately 1300 consumer products containing NPs were marketed in 2012. As a common metal oxide nanoparticle material, nanoscale titanium dioxide (nano-TiO$_2$) has been evaluated for diverse applications. TiO$_2$ has been shown to be a promising material for practical applications because it is highly photoreactive, inexpensive, non-toxic, chemically and biologically inert, and photostable. Also, nano-TiO$_2$ exhibits high specific surface area and anti-corrosion and photocatalytic properties [3]. It absorbs UV radiation and shows self-cleaning ability. Nanoparticles have a susceptibility to adsorb other substances to form a mixture leading to a shift of toxicity to living organisms [4]. Hence, many studies have reported cytotoxic characteristics of TiO$_2$ [5,6].

Some NPs are fatal to living cells, and their cytotoxicity may inhibit cell growth cycles, leading to death of organisms. Considering this fact, the cytotoxicity of TiO$_2$ in combination with other pollutants has been evaluated. TiO$_2$ is the most commonly manufactured nanoparticle material. It is assumed that because of the considerably high exposure TiO$_2$ NPs may enter the food chain. Because of current industrialization processes, organisms are also exposed to heavy metal pollutants [7]. Emitted NPs may interact with the pollutants, and this may subsequently lead to bioaccumulation. The contamination of water and soil with heavy metals has increased with anthropogenic and industrial activities [8,9]. TiO$_2$ NPs commonly co-exist with different heavy metals as they are released from wastewater treatment facilities to freshwater bodies, affecting the mode of action and the fate of the contamination. Studies have reported the ability of TiO$_2$ NPs to adsorb heavy metals and to increase their transport rate into hosts, increasing their concentration in the cell. Hu et al. [10] investigated the joint effect of TiO$_2$ NPs and humic acid (HA) on Cd$^{2+}$ bioaccumulation in zebrafish. In another study, Yang et al. [11] showed that TiO$_2$ NPs increased the accumulation of Cd$^{2+}$ in the ciliate Tetrahymena thermophila. Further, Tan et al. [12] showed increased uptake and retention of Cd$^{2+}$ and Zn$^{2+}$ adsorbed on TiO$_2$ NPs in Daphnia magna. Heavy metal contamination affects plant growth and indirectly affects human health via the food chain. Heavy metals have become an important factor limiting crop yields and, thus, threatening food security. Therefore, to improve crop yields, heavy metals need to be removed.

The toxicity of single-substance NPs has been tested extensively; however, the combination of single-substance NPs with other NPs or metals may cause co-exposure effects on living organisms. The extensive use of heavy metals in areas such as medicine and agriculture increased the negative impact of heavy metals on environment and living organisms, raising the need for risk assessment. Unlike other pollutants, heavy metals do not decompose, leading to bioaccumulation and biological hazards [13]. Heavy metals enter the human body through the consumption of fish and plants [14]. To date, heavy metals are removed through various methods. Among all methods available for removing heavy metals and toxic pollutants from waters, adsorption is the most widely used. Therefore, the joint organismal toxicity should be assessed.

Recently, nanoscale quantitative structure–activity relationship (nano-QSAR) models have been successfully applied to investigate the toxicity of NPs. QSAR models for predicting the biological activity of 48 fullerene derivatives [15], 51 manufactured nanoparticles with varying core metals, coatings, and surface attachments [16], and 80 surface-modified multifunctional carbon nanotubes have been reported. Another approach, namely nano-read-across (nano-RA) [17], has been used to determine the cytotoxicity of unknown nanomaterials based on structure similarities with known substances. Materials with similar structures are likely to produce similar toxicity through comparable mechanisms. The development of machine learning (ML) approaches, such as artificial neural networks (ANNs), decision trees, logistic regression (LR), support vector machines (SVM), Naïve Bayes (NB), random forest (RF), and $k$-nearest neighbor ($k$-NN), can be used to construct models that simulate complex relationships [18] and make predictions based on training data.

Using ensemble learning (EL) [19] methods, one can determine the relationship between the response and the predictor as well as solve regression problems. Additionally, such methods overcome problems with weak predictors and can be used to reduce the overfitting of the training data by averaging and incorporating multiple models. Ensemble learning is established with multiple algorithms and is divided into bagging and boosting algorithms. The boosting algorithm is an iterative algorithm that uses a weak model to build a strong model. Both bagging and boosting improve the prediction accuracy of weaker learners. A boosting algorithm combines many models linearly, with each new model depending on the previous one. In the bagging algorithm, replica data sets are generated that minimize prediction variance in machine learning. An iterative algorithm performs a series of repeated steps to gradually improve the model’s performance or to optimize a specific parameter. The algorithms continue to update the model’s parameters based on the training data until a certain stopping criterion is met, such as reaching an optimal solution, or a predefined number of steps are completed. This process is performed during the training of the model, where the model learns from the data by adjusting its parameters to minimize a specific cost or error function. These algorithms play a crucial role in training machine learning models and are fundamental to many optimization and learning techniques. Fine-tuning the model parameters through iterations
helps to improve the model’s performance and makes it more suitable for making accurate predictions for new, unseen data.

The boosting algorithm is an ensemble method that works sequentially by adding predictors to an ensemble, each one correcting its predecessors. In the boosting algorithm, at first, an initial model is developed with the dataset and then the algorithm tries to adjust the model parameters and again develops a model that tries to correct or minimize errors present in the previous model. This process is repeated until a satisfactory model is obtained or the error function is significantly optimized. Through this process, we get a strong learner or model from several weak learners or models by sequentially minimizing the error present in the predecessor models. Here, the weak model represents the models that are developed at an initial stage and contain a significant amount of error. The strong model is indicated by the final model, which contains a significantly low level of error and is able to predict new unknown data more accurately. Bagging (or bootstrap aggregating) is an ensemble method that generates a number of bootstrap datasets by a method called random sampling with replacement, and each dataset is used to train the models separately. The final prediction is obtained by averaging the outcome of each model (for regression models) or by majority voting (for classification models).

The objective of the present study was to construct EL-based regression models (RF, Gradient Boost, Extreme Gradient Boost, and AdaBoost) with periodic table descriptors for predicting the cytotoxicity, in terms of cell viability, of eight heavy metals adsorbed on nano-TiO$_2$. Also, the best algorithm showing the most contributing features responsible for the toxicity to HK-2 (human kidney 2) cell has been determined. To the best knowledge of the authors, this is the first work on ML models using periodic table descriptors to successfully demonstrate the high potential of the proposed modeling approaches.

### Methods and Materials

#### Dataset

The dataset was collected from previously published literature [20]. A mixture of nano-TiO$_2$ powders was added to HK-2 cells in Hyclone DMEM medium supplemented with 10% fetal bovine serum (FBS) and 100 mg penicillin/streptomycin and maintained at 37 °C in the presence of 5% carbon dioxide. Nine concentrations of heavy metal salts were added to a constant amount of nano-TiO$_2$ (25 µmol/L). The details of heavy metal concentrations are given in Table 1.

HK-2 cells were utilized to determine the toxicity in this study using cell viability as the endpoint. HK-2 cells are a sensitive model for examining renal cytotoxicity. They grow in monolayers and are suitable for studying the proximal tubular toxicity of a variety of compounds [21]. The main advantage of HK-2 cells is that they retain the basic morphological and functional properties of proximal tubular epithelial cells [22]. Cell viability was measured by using Equation 1:

$$S = \frac{A_{\text{exp}} - A_{\text{blank}}}{A_{\text{control}} - A_{\text{blank}}}$$  

(1)

Here, $S$ stands for cell survival rate, $A_{\text{exp}}$ is the absorbance value of the experimental group, $A_{\text{control}}$ is the absorbance value of the control group, and $A_{\text{blank}}$ is the absorbance value of the blank control group.

#### Descriptor calculation

Based on the characteristics of metals, we used easily calculable periodic table descriptors. Simple molecular information was generated time-effectively and cost-effectively. The previously used descriptors by Kar et al. [23] are the metal electronegativity ($\chi$), the sum of metal electronegativity for an individual metal oxide ($\sum \chi$), the sum of metal electronegativity for

### Table 1: Different concentrations of heavy metal salt samples in µmol/L.

<table>
<thead>
<tr>
<th>Sample</th>
<th>CdCl$_2$</th>
<th>ZnCl$_2$</th>
<th>CuSO$_4$</th>
<th>NiCl$_2$</th>
<th>Pb(NO$_3$)$_2$</th>
<th>MnCl$_2$</th>
<th>SbCl$_3$</th>
<th>CoCl$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>60</td>
<td>30</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>90</td>
<td>60</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>120</td>
<td>90</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>150</td>
<td>120</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>180</td>
<td>150</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>210</td>
<td>180</td>
<td>600</td>
<td>600</td>
<td>600</td>
<td>30</td>
<td>60</td>
</tr>
<tr>
<td>7</td>
<td>70</td>
<td>240</td>
<td>210</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>35</td>
<td>70</td>
</tr>
<tr>
<td>8</td>
<td>80</td>
<td>270</td>
<td>240</td>
<td>800</td>
<td>800</td>
<td>800</td>
<td>40</td>
<td>80</td>
</tr>
<tr>
<td>9</td>
<td>90</td>
<td>300</td>
<td>270</td>
<td>900</td>
<td>900</td>
<td>900</td>
<td>45</td>
<td>90</td>
</tr>
</tbody>
</table>
an individual metal oxide divided by the number of oxygen atoms present in that metal oxide (Σχ/NO), the number of metal atoms (N\text{Metal}), the number of oxygen atoms (N\text{Oxygen}), the charge of the metal cation in a given oxide (\text{gox}), and the molecular weight (MW). These descriptors are termed “first-generation periodic table descriptors”. The newly introduced sixteen descriptors are denoted as “second-generation periodic table descriptors” [24]. The computed descriptors for all metals are reported in the Excel file in Supporting Information File 1. In addition to being computationally less demanding, periodic table descriptors are size-independent.

**Splitting of data set and hyperparameter tuning**

The dataset was split into training and test sets before building the model. The training set was mainly used to fit the model, and the test set was used to measure the generalization ability of the developed model. Theoretically speaking, the dataset was divided based on a sorted response-based approach using the in-house dataset division tool (https://dtclab.webs.com/software-tools). In this study, the size ratio was set at 3:1 (training set/test set) for dataset division.

In almost any ML algorithm, different models are trained for a dataset and the best-performing model is selected. However, there may be room for improvement, and hyperparameter tuning can significantly improve the model. Here, the optimal values of the hyperparameters of the models were obtained with the GridSearchCV algorithm using the hyperparameter optimizer tool (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home/machine-learning-model-development-guis?pli=1). GridSearchCV tests all combinations of values in the dictionary and evaluates the model using the cross-validation method for each combination. Therefore, we choose the hyperparameter combination with the best average MAE results from the validation sets.

**Feature selection with random forest**

The goal of feature selection techniques is to find the best set of features that allows one to build optimized models. Feature selection using RF is an embedded method. Embedded methods combine the benefits of filter and wrapper techniques. These methods encompass the interaction of features while maintaining reasonable computational cost. In embedded methods, each iteration of the model training process is taken care of, and a few features that contribute the most to the training process are carefully extracted. More precisely, it is measured how much impurity is reduced on averaging (weighted average) through each tree nodes with the selected features. Here, each node is equivalent to the number of training samples associated with it. Through the RF algorithm, we have selected the most contributing eight periodic table descriptors, namely “conc”, “Σχ”, “atomic radius”, “IP_ActivM”, “Mol_Wt”, “χ of metal”, “D3_HeteroNonMetal”, and the total number of atoms in a molecule, from a pool of 43 periodic table descriptors by using the features with the highest Gini importance [25].

The selected first eight descriptors (most contributing features) were further used for modeling using RF, AdaBoost, Gradient Boost, and Extreme Gradient Boost algorithms.

**Model development**

This section introduces four classification models; all of them are ensemble learning models. ML is a subset of artificial intelligence where the machine learns from data and improves performance from past experiences and makes a prediction based on it [26]. In this study, along with RF, Gradient Boost, Extreme Gradient Boost, and AdaBoost were also performed. In the supervised learning approach, a model is trained on labeled datasets. Regression analysis algorithms are trained and learned from both input features and output labels. Regression analysis seeks a mapping function from the input features for a continuous output function. In this study, there is no intention to categorize the dataset, instead it is to be predicted quantitatively. Hence, the supervised regression method is selected to map the function of heavy metals and predict the cytotoxicity of these metals on HK-2 cells with periodic table descriptors. For the model development, after dataset division and feature selection, different ML algorithms are performed. The overall workflow is illustrated in Figure 1.

**Random forest (RF)**

In ensemble learning, RF is often used for its flexibility. Whether it is regression or classification, RF is a versatile learning method that can handle both. It works by building several decision trees in the training phase and generates average forecasts of various decision trees involved. In other words, it combines the results of different decision trees to make the best possible decision. Though the goal variable in classification-based issues is categorical, numerical values are present in regression. One advantage of RF is its capacity to analyze large datasets with great efficiency [27]. It can be regarded as a dimensionality reduction method since it analyzes large input data and finds all important variables. While handling RF datasets, the model emphasizes the importance of parameters, which is a highly helpful aspect [28].

**Adaptive boosting (AdaBoost)**

AdaBoost is one of the best boosting algorithms. It uses an ensemble learning method. This approach of machine learning is based on the idea of creating accurate prediction rules by combining many relatively weaker and inaccurate rules and assisting in alleviating overfitting issues. It is possible to make a...
smarter learner by altering the training data intelligently and constructing many submodels. It includes an unlimited amount of decision trees for input data throughout the training stage. During the creation of the first decision tree, incorrect data are highlighted inside the primary model.

The identical data serve as input for a separate model. This procedure is repeated until a specific number of base learners is generated. It uses a weighted average relying on the subsets to determine whether it should be included in the finalized model. In reality, some data may include linear predictions, and others may not. Therefore, utilizing the ensemble AdaBoost allows us to capture the nonlinear predictions and make a precise prediction for such data [29].

Gradient Boost (GB)
In 2002, Friedman [30] suggested an ensemble learning algorithm for both regression and classification. The GB method is associated with each repetition of the randomly chosen training data set with the fundamental model. Overfitting is inhibited by randomly subsampling the training set data; by doing so, the execution time and model accuracy are also improved. Since every repetition of the model must include small data (as a training set) the regression becomes quicker. The GB approach also requires modification or changes in a few parameters. That is, n-trees should not be too small, and the shrinkage aspect, also recognized as the learning rate, must not be kept too high [31].

Extreme Gradient Boost (XGBoost)
In a similar manner as described in [32], another ensemble ML algorithm, XGBoost of tree boosting, uses a gradient-boosting framework for efficient and scalable implementation performance. Ensemble learning uses multiple predictions that are multiple models for gradient enhancement and yields good adaptability to outliers and continuous variables. It is an efficient tool for dealing both regression and classification problems. The basic idea is to build “N” regression trees to train each subsequent tree using the residual from the previous tree. Models are built recursively until there is no improvement in
the results obtained. The new models predict the residuals of the prior model and then collectively provide the final predictions [33]. The gradient descent algorithm is used to minimize the loss while adding new models. Then, these individual predictors or classifications are combined to give more strong and more precise predictions. The workflow of the ML algorithm is represented in Figure 2. Tuning can be done using the grid search method.

SHAP analysis

The feature importance in the model was determined using the Shapley Additive exPlanation (SHAP) method, using SHAP version 0.41.0. The SHAP framework takes into account the calculation of Shapley values. These values are calculated from the average marginal contribution of each feature from all conceivable coalitions. First, the dataset is incorporated into the model, then the SHAP framework assigns a Shapley value to each feature that contributes to the corresponding output of the model. Therefore, SHAP helps to select the features based on a ranking algorithm [34]. We have selected the features having the highest Shapley values for the training set since the standard method tends to overestimate the continuous variables.

Model validation

A reliable model should pass the threshold values for different internal and external validation metrics. Internal validation generates the generalization ability and robustness of the model. In contrast, external prediction is used to validate the model. The most common metrics to measure internal quality are the coefficient of determination ($R^2$ and $Q^2_{LOO}$). Besides these, we have also calculated the root mean square error (RMSE) of the training set. The mean absolute error ($\text{MAE}_{\text{test}}$), the root mean
Applicability domain (AD) analysis

After building the model, the applicability domain (AD) must be considered. AD represents the domain that can be effectively predicted by the model that is based on the training set data. The samples within the domain of applicability can only explain the reliability of the predicted values. A William’s plot was used to determine the AD of the present work. The leverages were calculated using the in-house Hi_Calculator-v1 Software (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home?pli=1). The distance between the X value of the i-th observation and all X values is represented by the leverage value. It generally considers \(3k’/N\) as the critical value or the standard value \(h^*\). Here, \(k’\) represents the number of descriptors plus 1, and \(N\) represents the number of compounds in the training set. If the leverage value is higher than \(h^*\), the corresponding compound is outside the AD.

Results and Discussion

In this research, we have used four ML models, namely RF, AdaBoost, Gradient Boost, and Extreme Gradient Boost to forecast the toxicity of heavy metals adsorbed on nano-TiO\(_2\) to HK-2 cells using periodic table descriptors (Table 2). The ML models were built using the features selected by the RF algorithm. Model specification and configuration were carried out by optimization of the hyperparameters. The AD was also determined, and all compounds were found to be below the threshold of \(h^* = 0.42\), as shown in the Williams plot in Figure 3. The AD is the chemical space formed based on the descriptors of the training set compounds. The compounds in the chemical space are considered reliable for predictions, while those beyond the AD would not guarantee accurate predictions. The AD plays an important role in determining the uncertainty of the predictions.
of specific molecules based on how close they are to the training set compounds used to develop the model. AD is a valuable tool for the characterization of interpolation spaces based on the modeled descriptors and response functions.

The applicability domain developed here is based on the features of some specific heavy metal salts, that is, CdCl$_2$, ZnCl$_2$, MnCl$_2$, CoCl$_2$, CuSO$_4$, NiCl$_2$, Pb(NO$_3$)$_2$, and SbCl$_3$. The developed model should be applicable to other closely related heavy metal salts.

**Diagnosis based on SHAP value**

The goal of SHAP is to explain the prediction of an instance by computing each feature of the prediction. First, the SHAP value is used to calculate the magnitude of the contribution of each feature and then ranked to obtain the importance ranking of features. Features with large absolute Shapley values are important. Here, we have used the kernel method to calculate the SHAP values [36]. The SHAP analysis and hyperparameter tuning (max_depth: “none” min_samples_leaf, min_samples_split, n_estimators) revealed that concentration, followed by atomic radius and IP_ActivM, ranked highest among the eight features (conc, $\sum\chi$, atomic radius, IP_ActivM, Mol_Wt, $\chi$ of metal, D3_HeteroNonMetal, and atoms in the molecule) in the RF model. The hyperparameter setting n_estimators was kept at a value of 80 for RF, while it was 130, 70, and 170 for Gradient Boost, Extreme Gradient Boost, and AdaBoost respectively. The relative importance of each descriptor for all ML algorithms can be understood using the SHAP analysis (Figure 4). The SHAP methodology identifies the features contributing most to the model prediction. We can find that the conc (concentration of the heavy metal) descriptor contributes the most to the EL algorithms. The Shapley values reflect the average marginal contribution of a feature value across all possible feature coalitions, both in terms of magnitude and direction.

**Results of model validation for all ML methods**

In order to determine if heavy metals and TiO$_2$ nanoparticles had any cytotoxic effects, the selected eight important periodic table-based features were used. The final models developed with RF, AdaBoost, Gradient Boost, and Extreme Gradient Boost were evaluated using MAE$_{train}$, RMSE$_{train}$, $R^2$, and $Q^2$ for the training set and MAE$_{test}$, MSE, RMSE$_{test}$, $Q^2F_1$, $Q^2F_2$ metrics for the test set, and the results are shown in Table 1. According to the results, the MAE$_{test}$ (0.14) was found to be the least for the test set in the RF method, followed by AdaBoost,
Gradient Boost, and XGBoost. The XGBoost gives the highest $R^2$ (0.99) for the training set, while AdaBoost gives the lowest $R^2$ (0.88) with the highest MAE$_{test}$ (0.33). Cross-validation (CV) statistics were obtained based through 20 times fivefold repetitive CV along with 1000 times shuffle split CV (mean ± SEM) method. This is done to protect the model from overfitting when the data is limited. The results of the CV indicate clearly that the models do not memorize the correspondence between the descriptors since the outcome of $R^2$ is highest and the MAE value is lowest for the RF model after the repetitive CV method. This suggests the superiority of the RF model to other models. Figure 5 presents the cross-validation statistics based on 20 times fivefold repetitive CV and 1000 times shuffle split CV on $R^2$ and MAE for the developed ML model.

General mechanism of toxicity

In the process of screening all descriptors from different ML methods, some common descriptors for heavy metals were discovered that are clear indicators of their importance regarding toxicity to HK-2 cells. We found that the concentration of the heavy metal (conc), the atomic radius of the metal, the electronegativity, and the molecular weight of the heavy metal influence the survival rate of the HK-2 cells. It was observed that conc, mol wt, atomic radius, and the total number of atoms in the molecule were of high importance in all the models. The increase of conc, mol_wt, and total atoms in a molecule is believed to increase toxicity. The toxicity of the heavy metals is also time- and dose-dependent. Among many other factors, the valence state plays an important role in toxicokinetics and toxicodynamics. Many studies have shown that an increased concentration of heavy metals is correlated with the severity of hepatotoxicity and nephrotoxicity [37]. Lead causes toxicity through an ionic mechanism followed by the generation of reactive oxygen species (ROS). Another, biomarker for ROS is lipid peroxidation [38] as free radicals cause lipid peroxidation inside the cell membrane. The catalytic properties of the metals are also responsible for an increased toxicity of manufactured nanoparticles [39] (Figure 6). Electronegativity and atomic radius influence the catalytic properties of the metal. Metal cations also catalyze the lipid peroxidation process [40] through enhancement of endocytosis and the intrinsic properties of the heavy metal. The toxicity is associated with internalization and bioaccumulation in the HK-2 cells. The increase in the concentration of heavy metals and their adsorption to nano-
Figure 6: General mechanism of toxicity to HK-2 cell by heavy metals.

TiO$_2$ induces toxicity by increasing the generation of ROS in the HK-2 cells.

Comparison with the previous work
The present work describes the development of a model for heavy metals with different concentrations through simple periodic table descriptors using various ML methods. The results obtained from the ML method suggest that the models have better predictivity than the models developed previously by Sang et al. [20] as shown in Table 3. Sang et al. [20] applied the random forest algorithm and the AdaBoost algorithm for QSAR modeling using quantum mechanical descriptors. In contrast, the present study involved the random forest algorithm, the AdaBoost algorithm along with Gradient Boost and XGBoost algorithms using simple periodic table descriptors that are easy to interpret and can be calculated quickly without the involvement of expert personnel. These descriptors simplify the nanostructure property calculation and determine the nanoscale interactions without much computational intervention. The use of such descriptors saves time; the descriptors are also cost-effective and have a clear and straightforward physical meaning, which facilitates the mechanical interpretation of the QSAR models. A direct comparison was not possible due to different dataset division and descriptors but the results obtained in the present work for the RF method was superior to that of the previous work.

Conclusion
We have performed cytotoxicity modeling of eight heavy metal compounds adsorbed on nanoscale TiO$_2$ regarding HK-2 cells and explored the features responsible for the toxicity mechanism. Many studies have examined the co-exposure of metal and metalloid mixtures with heavy metals. The co-exposure may also be affected by dose variations at the biomarker level. Also, co-exposure in humans was found to lead to more profound renal damage than exposure to each of the elements alone. Hence, to elucidate the features responsible for the toxicity, in the present study, ML algorithms were applied along with periodic table descriptors for QSAR modeling. Experiment-independent periodic table descriptors produced better results than quantum chemical descriptors in previous studies. The periodic table descriptors used in QSAR models have strong

| Table 3: Comparison of the current work with the previous study. |
|---------------------|-----------------|-------------|----------|---------|---------|---------|---------|
| Descriptors         | Method          | $R^2$       | $Q^2_{\text{LOO}}$ | MAE$_{\text{train}}$ | RMSE$_{\text{C}}$ | $Q^2_{F1}$ | $Q^2_{F2}$ | MAE$_{\text{test}}$ | RMSE$_{\text{P}}$ |
| periodic table-based | random forest   | 0.96        | 0.72        | 0.13     | 0.2     | 0.94     | 0.94     | 0.14     | 0.19     |
| (current study)     | random forest   | 0.85        | 0.70        | —        | 0.06    | 0.86     | 0.85     | —        | 0.10     |
| quantum mechanical  | random forest   | —           | —           | —        | —       | —        | —        | —        | —        |
| (Sang et al.)       |                 | —           | —           | —        | —       | —        | —        | —        | —        |
Supporting Information
Supporting Information File 1
Detailed information regarding heavy metals at different concentrations.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-14-77-S1.xlsx]

Funding
This study was funded by Indian Council of Medical Research (ICMR), New Delhi, in the form of a Senior Research Fellowship to JR. SP acknowledges AICTE (All India Council for Technical Education), New Delhi for financial assistance.

ORCID® iDs
Joyita Roy - https://orcid.org/0000-0001-5000-7073
Kunal Roy - https://orcid.org/0000-0003-4486-8074

References
27. El Naqa, I.; Murphy, M. J. What Is Machine Learning?, Machine Learning in Radiation Oncology; Springer International Publishing: Cham, Switzerland, 2015; pp 3–11. doi:10.1007/978-3-319-18305-3_1

Theoretical guidance, which can help scientists design new entities with expected properties. As a part of the model development process, periodic table descriptors can be used in conjunction with other descriptors that are compatible with them. The periodic table descriptors are not only less computationally demanding but also independent of the size of the particles. The ML algorithm with periodic table descriptors has helped to evaluate the cell survival rate of HK-2 cells in less time and at less cost than using expensive quantum chemical descriptors and experimental descriptors. Among all algorithms, the random forest model shows the best prediction ability with $Q^2_{F1} = 0.91$ and $\text{MAE}_{\text{test}} = 0.14$ for the test set. Hence, a good feature selection method reduced the computation time required to train a model. The SHAP analysis also emphasized the most significant features contributing to the model. We have proposed also a generalized mechanism for the most impactful features generated by the model. As a result, periodic table descriptors and machine learning can be used together to decipher features of unknown compounds and predict compounds that are similar.


License and Terms

This is an open access article licensed under the terms of the Beilstein-Institut Open Access License Agreement (https://www.beilstein-journals.org/bjnano/terms), which is identical to the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0). The reuse of material under this license requires that the author(s), source and license are credited. Third-party material in this article could be subject to other licenses (typically indicated in the credit line), and in this case, users are required to obtain permission from the license holder to reuse the material.

The definitive version of this article is the electronic one which can be found at: https://doi.org/10.3762/bjnano.14.77

950
Multiscale modelling of biomolecular corona formation on metallic surfaces

Parinaz Mosaddeghi Amini*, Ian Rouse, Julia Subbotina and Vladimir Lobaskin

Full Research Paper

Address:
School of Physics, University College Dublin, Belfield, Dublin 4, Ireland

Email:
Parinaz Mosaddeghi Amini* - parinaz.mosaddeghiamini@ucdconnect.ie

* Corresponding author

Keywords:
all atomistic; aluminum; bionano interface; coarse grained model; lactose; milk protein; multiscale modelling; protein corona

Abstract

In the realm of food industry, the choice of non-consumable materials used plays a crucial role in ensuring consumer safety and product quality. Aluminum is widely used in food packaging and food processing applications, including dairy products. However, the interaction between aluminum and milk content requires further investigation to understand its implications. In this work, we present the results of multiscale modelling of the interaction between various surfaces, that is (100), (110), and (111), of fcc aluminum with the most abundant milk proteins and lactose. Our approach combines atomistic molecular dynamics, a coarse-grained model of protein adsorption, and kinetic Monte Carlo simulations to predict the protein corona composition in the deposited milk layer on aluminum surfaces. We consider a simplified model of milk, which is composed of the six most abundant milk proteins found in natural cow milk and lactose, which is the most abundant sugar found in dairy. Through our study, we ranked selected proteins and lactose adsorption affinities based on their corresponding interaction strength with aluminum surfaces and predicted the content of the naturally forming biomolecular corona. Our comprehensive investigation sheds light on the implications of aluminum in food processing and packaging, particularly concerning its interaction with the most abundant milk proteins and lactose. By employing a multiscale modelling approach, we simulated the interaction between metallic aluminum surfaces and the proteins and lactose, considering different crystallographic orientations. The results of our study provide valuable insights into the mechanisms of lactose and protein deposition on aluminum surfaces, which can aid in the general understanding of protein corona formation.

Introduction

The interface between biological systems and engineered materials has gained significant attention in recent years because of its wide range of applications, spanning from food to medicine and environmental science [1,2]. This interface plays a crucial role in ensuring the safety and quality of processed and packaged products. The selection of packaging materials and their...
interaction with biological components have emerged as critical determinants impacting the preservation, shelf life, and overall acceptability of dairy products [3]. Consequently, the interface between biologically relevant molecules and nanoscale materials, such as aluminum, has become an increasingly important and intriguing area of research [4]. For long-term storage and preservation of prepared food, the choice of containers and utensils made from specific materials is essential [5]. For example, it was shown that ripened cheese and cheese spreads acquire a higher aluminum content as compared to other milk products [6]. Aside from wrapping and container packaging, aluminum has found a wide popularity in other applications, such as manufacturing of kitchen utensils, cosmetics, and components for medical and scientific equipment [7].

Figure 1 presents a schematic contamination cycle of dairy products, showcasing potential sources and pathways of aluminum pollution. It illustrates the journey of milk from a cow grazing on grass contaminated with heavy metals, highlighting the crucial role of metallic containers, metal-based equipment, and kitchen utensils in maintaining product integrity. The figure further demonstrates the potential to introduce heavy metal contamination, including iron and aluminum, during processing and emphasizes the formation of a milk layer in form of a protein/lactose corona at the outer surface of macroscopic and micro- and nano-sized particulate after packaging. It also highlights the dynamic interactions at the bionano interface associated with potential human health hazards. Through biomolecule adsorption, change of conformation, and surface chemistry, foreign materials engage in a complex interplay of dynamic physicochemical interactions, kinetics, and thermodynamic exchanges that can lead to undesirable outcomes [1,8-10].

In a more general context, the importance in understanding the mechanism of bionano interactions arises from the increasing awareness and concerns regarding the safety of nanoparticles (NPs) in relation to human and animal health. The toxicity of NPs is closely linked to their chemical aggressiveness and varies with their physicochemical properties, including surface area, charge, and reactivity. Understanding the intricate interplay between these properties and the biological systems is vital for assessing and mitigating any potential adverse effects associated with exposure to NPs [11]. To advance in this field, it is crucial to comprehend the underlying forces and molecular...
constituents that govern the interactions between biomolecules and metals. However, traditional safety assessment methods can be costly, time-consuming, and often involve animal studies. In this regard, in silico modelling offers a promising alternative that can predict the interactions of NPs with living organisms. By leveraging computational approaches, in silico modelling provides a humane and cost-effective means of obtaining the necessary information, thus aiding in the evaluation of NP safety and reducing reliance on animal experimentation [12-14]. Data-driven methods that rely on statistical analysis are employed for this purpose, particularly when sufficient data are available. These methods leverage the power of large datasets to identify patterns, trends, and correlations between metal properties and their interactions with biomolecules [15-18]. In recent years, researchers have focused on using physics-based models to understand the mechanisms underlying the formation of NP protein corona, a complex layer of biomolecules that surrounds NPs upon their exposure to biological fluids [19,20]. It is widely recognized that composition and configuration of the protein corona play a crucial role in determining the biochemical reactivity, sensitivity of NPs, as well as their cellular uptake and systemic transfer [21]. However, in order to develop predictive models, a deeper understanding of the interactions at the bionano interface and their relationship to material and protein properties is necessary. Gathering more information on these intricate interactions will facilitate the development of accurate predictive models, thereby advancing our ability to assess the behavior and potential implications of NPs in biological systems. The bionano interface can be broken down into three interconnected components: (i) the surface of the NP, which is influenced by its physicochemical composition, (ii) the interface between the solid NP and the surrounding liquid environment, where notable changes occur upon interaction, and (iii) the contact zone between the solid–liquid interface and biological substrates (Figure 2) [22].

In this work, we study bionano interactions involving metallic aluminum and common dairy biomolecules, namely lactose and the six most abundant milk proteins [23]. The main objective of our analysis is to computationally quantify the relative binding of these proteins on zero-valent aluminum surfaces based on their energy of adsorption and orientation. We employ a three-level multiscale method (as shown in Figure 3) to calculate the energies of adsorption and the content of the corona for these proteins on the selected surfaces. In the section “Results and Discussion”, we provide a detailed explanation of the theoretical model developed to study the interaction between protein and lactose with metals, as well as the rationale behind the parameterization scheme used. Subsequently, we discuss the simulation results and analyze the individual adsorption affinities predicted for molecules representing the biological aspect of the interface, including amino acids (AAs), milk proteins, and...
carbohydrates. Additionally, we examine the preferred orientations of these molecules upon adsorption and investigate the kinetics of competitive adsorption among the proteins and lactose, aiming to understand the process of protein deposition on metallic surfaces. Finally, the key insights gained from this study are summarized, highlighting the implications and potential applications of the findings.

Results and Discussion

Here, we aim to predict the content of a biomolecular corona on a metallic aluminum surface. At the largest scale, our methodology employs a coarse-grained (CG) kinetic Monte Carlo (KMC) method [16] to simulate competitive adsorption of biomolecules onto the aluminum surface. To achieve this, we evaluate individual binding energies at various orientations (represented by heatmaps) for each selected protein immobilized on different fcc planes of the aluminum surface. These heatmaps for individual proteins are acquired through UnitedAtom (UA) simulations [24,25]. While the UA method has been parameterized for a range of rigid surfaces, including metals (Ag, Au, Cu, and Fe), oxides (TiO$_2$, SiO$_2$, and Fe$_2$O$_3$), carbonaceous NPs (graphene, carbon nanotubes, and carbon black), semiconductors (CdSe) [26], and polymers [27], it lacks the set of short-range potentials required for calculating milk protein-aluminum adsorption energies. Here, we compute potentials of mean force (PMF) for Al surfaces derived from explicit all-atom molecular dynamics simulations utilizing a previously established scheme [2,24,28]. These PMFs provide the input required to determine the adsorption energies between milk proteins and aluminum surfaces by using multiscale UA CG model, spanning from the atomistic level of description to the complete mesoscale model of the corona. Figure 3 shows the parameterization and simulation workflow, outlining different stages and components involved in the study.

All-atoms short-range interaction modelling results

All-atom metadynamics simulations were conducted using GROMACS-2018.6 and PLUMED (PLUMED2-2.5.1.conda.5) software packages [29-31]. CHARMM-GUI/Nanomaterial Modeler was employed to construct the topology and force fields of three fcc surfaces of Al: (100), (110), and (111) [32]. The General Amber Force Field (GAFF) was utilized to model side-chains analogues (SCA) within the system [33,34]. The AMBER force field is a widely recognized and extensively validated force field that provides accurate descriptions of molecular systems [35]. We evaluated the short-range PMFs between 22 SCAs and an Al slab in a solvent environment comprising water and salt ions. The system’s pH value was maintained at a neutral level, and the NaCl salt concentration was set to 150 mM, mimicking the overall ionic strength of milk and equivalent to one salt molecule per 10 nm$^3$. The system underwent equilibration for 1.0 ns under constant pressure conditions at 1.0 bar and a temperature of 300 K, following the NPT ensemble, employing Berendsen weak coupling method [36]. Subsequently, a pre-equilibration phase was conducted for
Figure 4: Adsorption free energy profiles of SCAs on three aluminum fcc slabs as a function of the surface separation distance (SSD). These profiles were calculated using all-atom AWT-MetaD. The vertical lines indicate the positions of water and ion layers. (a) Al(100), (b) Al(110), and (c) Al(111).

10 nanoseconds within the NVT ensemble. For the short-range interactions, the cut-off distance was defined as 1.0 nm. In the adaptive well-tempered metadynamics (AWT-MetaD) simulations, the adsorption energy was calculated at a temperature of 300 K, a pressure of 1.0 bar, and a neutral pH within the NVT ensemble. Additionally, we measured the interaction energy as a function of surface separation distance (SSD) as a collective variable, enabling a comprehensive analysis of the AA–NP interactions. For a detailed explanation of the method used in this study, please refer to previous reports [2,24,28] where the method has been described in depth. Figure 4 and dataset [37] show the obtained free energy of adsorption in units of $k_B T$.

The water density profiles obtained from MD simulations for the slab–water system in the context of Al surfaces revealed characteristics that were previously observed for other simulated metallic surfaces [2,28]. The profiles exhibited two distinct regions with elevated water density located approximately 0.15–0.18 nm and 0.42–0.48 nm away from the aluminum surface. These regions corresponded to the first and second water layers adjacent to the metal surface, respectively (as depicted in Supporting Information File 1, Figure S1). Further examination of the ion density profiles indicated the presence of sodium ions within a range of 0.55–0.60 nm and chloride ions within a range of 0.42–0.46 nm from the Al surface. Notably, the positions of the chloride ions align closely with the second
water layer, while sodium ions are located past this layer, as marked by the blue and purple vertical dashed lines in Figure 4. This alignment suggests that the chloride ions integrate into the network of water molecules comprising the second adlayer. Additionally, the analysis of the PMFs revealed a significant minimum at a distance of 0.21–0.25 nm. Figure 5 shows the minimum energy values obtained for each AA on different facets of the aluminum surface (100, 110, and 111) in a bar chart.

A comparison of the adsorption energies on aluminum and iron surfaces reveals distinct preferences for different AAs. On aluminum surfaces, ARG, PRO, TRP, TYR AAs show the strongest attraction (−63.32k_BT to −41.46k_BT), followed by HIE, GLN, PHE, GAN (−43.86k_BT to −20.85k_BT). VAL, THR, SER, CYS, ALA exhibit the weakest attraction (−19.51k_BT to −1.76k_BT). On iron surfaces, charged and aromatic PRO, TYR, ARG, HIS AAs are strongly adsorbed (−91.29k_BT to −43.34k_BT), while hydrophobic VAL, LEU, ALA AAs show a weaker adhesion (−21.70k_BT to 2.86k_BT) [2]. We also show the PMF for glucose with aluminum surfaces, used as the basis for a model of lactose, a sugar highly present in milk, as discussed later, computed using the PMFPredictor software in Figure 6 [38].

Figure 5: Minimum energy of adsorption (k_BT) for each SCA on three Al fcc slabs obtained through all-atom simulations: (a) Al(100), (b) Al(110), and (c) Al(111). Notably, Al(111) exhibits a stronger binding affinity than Al(100) and Al(110).
Figure 6: The interaction potential of glucose with the three Al surfaces predicted using the PMFPredictor Toolkit. The solid lines give the ensemble average of ten versions of the model while the shaded regions indicate the 95% confidence intervals.

Protein–NP interactions

To further understand the adsorption energy and orientation of each individual protein, a primary coarse-graining step was performed. In this part, we use the UA model to predict the protein–NP binding energies. This model takes into account various factors, such as the material’s chemical composition, size, shape, surface roughness, charge, functionalization, and hydrophobicity, when constructing CG models for the bionano interface. The UA model simplifies the protein–NP interactions by representing proteins as rigid structures composed of 20 AA types, each represented by a single bead. This interaction is described through a short-range surface non-bonded potential ($U_{s}^{nb}$) (including van der Waals (vdW) repulsion and solvent effects), a long-range core vdW potential ($U_{i}^{vdW}$), and an electrostatic potential ($U_{el}$). Through interaction potentials for specific AAs with the NP, the overall interaction potential between the NP and the complete protein ($U_{p-NP}$) is expressed in a pairwise additive manner:

$$U_{p-NP} = \sum_{i=1}^{NAA} U_i (d_i (0, \phi))$$

$$= \sum_{i=1}^{NAA} U_i^{el} (d_i (\theta, \phi)) + \sum_{i=1}^{NAA} U_i^{nb} (d_i (\theta, \phi))$$

$$+ \sum_{i=1}^{NAA} U_i^{vdW} (d_i (0, \phi)).$$

The potential $U_{p-NP}$ depends on the distance $d_i$ between the centers of mass of the NP and each AA in the protein. This distance is determined by the protein’s orientation with respect to the NP’s surface, which is defined by two rotational angles ($\phi, \theta$) relative to the protein’s initial orientation. This initial orientation is set by performing a principle axis transformation such that the axis associated with the smallest moment of inertia is aligned to the $z$ axis and the second smallest to the $y$ axis, that is, the $z$ axis is now typically associated with the greatest extent of the protein. Since this does not uniquely specify the orientation, further rotations of 180° are then applied if necessary such that the electric dipole moment is positive along these two axes. This produces a convenient reference state by which other orientations are defined. The specific orientation ($\phi, \theta$) is generated by applying a rotation of $-\phi$ around the $z$ axis followed by a rotation of 180° – 0 around the $y$ axis. The short-range surface non-bonded potentials are extracted from AWT-MetaD simulations, which were described in the section “All-atoms short-range interaction modelling results”. The Hamaker technique is used to approximate the long-range term that results from the vdW forces working through the aqueous medium between the NP core and the $i$-th AA. The electrostatic interaction between the NP and AA is represented by the screened Coulomb potential. More comprehensive information about the theoretical aspects of the UA model can be found in our previous publications [2,25,28,39,40]. The output of the UA simulations contains a collection of rotational configurations and their corresponding $E(\theta_k, \phi_l)$ values. By employing Boltzmann averaging and weighting factors based on the potential energy as a function of distance for each angle, we calculate the average adsorption energy of these configurations. Using this approach, we evaluate the adsorption energies of the entire proteins on aluminum surfaces. To predict the three-dimensional (3D) structures of proteins, we utilize the I-TASSER (Iterative Threading ASSEmbly Refinement) 5.1 software [41], which uses the protein’s AA sequences as an input.

For this study, we have chosen six representative cow milk proteins and lactose, which constitute most of the non-fat milk solids. Table 1 displays properties of the chosen compounds. It includes their UniProt IDs, molecular weights, charges, and the number of AAs in each protein. The charge data was determined through the PROPKA method [42,43] at a pH of 7.0. We model the lactose molecule as a pair of glucose beads; it does not possess a UniProt ID or a count of AA residues. We estimated the concentration of each protein and lactose based on their weight fraction in milk and considering the fact that cow milk has 30–39 g/L of protein and 45–55 g/L of lactose in total. The molar mass of each protein was taken from AlphaFold database [44]. Following this, all proteins underwent a 50 ns equilibration in water using NVT and NPT ensembles.

The UA computations were conducted using nine different Al NPs with varying radii, namely 2, 5, 10, 20, 30, 40, 50, 80, and...
100 nm, to investigate the influence of size and curvature on the adsorption energies. The results and detailed information on the calculation can be found in Supporting Information File 1, Figure S2 and Figure S3, which illustrate the variations in adsorption energies as a function of NP size. Within the range of 2–20 nm the binding energies of ALAC, BLAC, BC, and BSA show an initial increase on all surfaces, followed by a stabilization at larger NP sizes. In contrast, AS1C and AS2C exhibit a continuous rise in binding energy across the entire size spectrum, ranging from $-48.0k_BT$ at 2 nm to $-281.09k_BT$ at 100 nm for AS1C and $-15.26k_BT$ at 2 nm to $-275.60k_BT$ at 100 nm for AS2C, with AS2C exhibiting the most dramatic changes in binding energy as a function of size. This strong size dependence in binding energy for AS2C can be attributed to its rod-like 3D structure and the rigidity assumption in our model. As the size of the NP increases, AS2C can make more extensive contact with the surface. This increased contact area leads to enhanced binding affinity, resulting in the observed stronger binding across the size range. This is not the case for other proteins on the list as they are more compact and, therefore, reach a maximum number of contacts at relatively small NP sizes. Regarding the binding affinity rankings, for the smallest NPs (2 nm), the order from weakest to strongest is observed as AS2C, BSA, ALAC, BLAC, AS1C, and BC on Al(100), with similar rankings observed on Al(110) and Al(111) surfaces. However, for the largest (flattest) NPs (100 nm), the binding affinity ranking changes to ALAC, BSA, BC, AS2C, and AS1C on Al(100), BC, ALAC, BLAC, BSA, AS2C, and AS1C on Al(110), and BLAC, ALAC, BC, BSA, AS2C, and AS1C on Al(111) (see Supporting Information File 1, Figure S2). In reality, protein structures are not rigid, allowing them to adapt to the surfaces upon immobilisation. This can potentially affect their binding behavior. This can be especially significant for caseins, as they belong to the group of flexible milk proteins with no tertiary structure. Globular milk proteins (lactoglobulin and lactalbumin) are expected to be less prone to this shortcoming of the UA model.

Figure 7 shows the output of the UA model for the selected milk proteins on aluminum NPs with a surface size of 80 nm with zeta potential $-5$ mV at pH 7.0. The heatmaps display the adsorption energies for all values of $0$ and $\phi$. Blue areas with lower energies indicate more favorable orientations of the proteins. Each heatmap is accompanied by a 3D representation of the protein on the NP surface, with the AAs closest to the NP’s surface marked. The AAs that are most likely to make contact with the metal surfaces, according to analysis, are LYS, TYR, PHE, GLU, ARG, and ASP.

The rankings of protein adsorption on each aluminum surface are shown in Table 2, highlighting the variations in adsorption energies ($E_{ads}/k_BT$) and the particular protein–surface interactions ($0$ and $\phi$ in degrees). Moreover, the minimum distance ($r_{\text{min}}$ in nm) indicates the closest approach of the protein to the aluminum surface during the adsorption process.

The ranking of adsorption energies highlights the distinct adsorption behaviors of various proteins on different metal fcc surfaces. We can see that AS1C exhibits the highest adsorption energy on Al(100) and Al(111) surfaces, while on Al(110), AS1C, and AS2C show similar adsorption energies. In contrast, on metallic iron, AS1C consistently demonstrates the highest adsorption energy on Fe(100), Fe(110), and Fe(111) surfaces. This result reflects the size and shape of the AS1C protein, which allows it to make the largest number of contacts with the metal as compared to the other proteins. Regarding the most weakly bound proteins, on aluminum surfaces, ALAC consistently exhibits the lowest adsorption energy across all three surfaces, while BLAC shows slightly higher adsorption energies. In contrast, on iron surfaces, ALAC and BLAC demonstrate comparable adsorption energies, with ALAC exhibiting slightly lower energies on Fe(110) and Fe(111) surfaces [2]. We note that generally the binding of proteins to aluminum is weaker than to iron, which may be caused by the smaller lattice constant of fcc iron and higher density of surface atoms.

### Table 1: Characteristics of the selected milk proteins and lactose.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>UniProt ID</th>
<th>Compound name</th>
<th>MWa, Da</th>
<th>Charge, e</th>
<th>Resb</th>
<th>C2 [10^-4], mol/L</th>
<th>Rq [Å]</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS1C</td>
<td>P02662</td>
<td>αs1-casein</td>
<td>24526.00</td>
<td>-8.5</td>
<td>214</td>
<td>4</td>
<td>20.05</td>
</tr>
<tr>
<td>AS2C</td>
<td>P02663</td>
<td>αs2-casein</td>
<td>26016.69</td>
<td>4.5</td>
<td>222</td>
<td>1</td>
<td>40.81</td>
</tr>
<tr>
<td>BC</td>
<td>P02666</td>
<td>β-casein</td>
<td>25107.33</td>
<td>-4.5</td>
<td>224</td>
<td>4</td>
<td>22.53</td>
</tr>
<tr>
<td>ALAC</td>
<td>P00711</td>
<td>α-lactalbumin</td>
<td>16246.61</td>
<td>-5</td>
<td>142</td>
<td>0.9</td>
<td>15.01</td>
</tr>
<tr>
<td>BLAC</td>
<td>P02754</td>
<td>β-lactoglobulin</td>
<td>19883.25</td>
<td>-6</td>
<td>178</td>
<td>2</td>
<td>15.50</td>
</tr>
<tr>
<td>BSA</td>
<td>P02769</td>
<td>bovine serum albumin</td>
<td>69293.41</td>
<td>-4.5</td>
<td>607</td>
<td>0.1</td>
<td>27.69</td>
</tr>
<tr>
<td>LAC</td>
<td>—</td>
<td>lactose</td>
<td>342.3</td>
<td>0</td>
<td>—</td>
<td>1300</td>
<td>4.28</td>
</tr>
</tbody>
</table>

aMolecular weight, bNumber of residues, cConcentrations [mol/L] of the molecules in milk that were used in KMC calculations, dRadius of gyration of the biomolecules in Ångstrom.
Figure 7: Adsorption energy heatmaps obtained from the UnitedAtom model and corresponding 3D representations of the interactions of (a) AS1C, (b) AS2C, (c) BC, (d) BLAC, (e) ALAC, and (f) BSA with Al(110) in the preferred orientations. The figure highlights the closest AAs to the surface of the material.

Supporting Information File 2, Table S2 reports the preferred orientations of all 820 milk proteins based on the lowest energy from the UnitedAtom output. In our investigation of these proteins, we focused on identifying the most strongly adsorbing proteins when exposed to Fe and Al. These proteins, including P19660, A6QP30, G3X745, F1MMI6, E1BBY7, A6QLY7, and Q9N2I2, demonstrated remarkable similarity in their binding behavior towards Fe(100) and Al(100) surfaces, E1BGJ4, A5D7M6, A6QP30, G3X745, and F1N1C7 on Fe(110) and Al(110) surfaces, and F1MMI6 and E1B748 and A6QP30 on Fe(111) and Al(111) surfaces.

In the subsequent step, we predicted the composition of the milk protein layer at the aluminum surfaces. For this analysis, we consider the Al surface as a spherical NP with the protein layer uniformly adsorbed on its entire surface, forming the protein corona.

Competitive adsorption and biomolecular corona

Kinetic Monte Carlo (KMC) simulations as implemented in the CoronaKMC tool [26] were employed to investigate competitive adsorption and to determine the composition of the protein corona. This method models adsorbates as hard spheres, which adsorb and desorb to the surface of the NPs, with different orientations of each protein treated as different potential adsorbates to allow for a more physically realistic model of corona formation for anisotropic proteins. In brief, a standard kinetic Monte Carlo routine is used to advance the simulation from one event, collision of an incoming adsorbate with the NP or desorption of an adsorbed species, to the next, with events occurring with a probability proportional to their rate. In the initial form of the model, adsorption is assumed to occur with unit probability if the incoming species does not overlap with any currently adsorbed species and fails to take place otherwise. We
parameterize this model using adsorption and desorption rate constants extracted from UnitedAtom results as described previously [16,45]. In brief, each potential adsorbate (e.g., a small molecule or a particular orientation of a protein) is projected onto the surface of the NP and a convex hull procedure used to estimate the area of the NP occupied by that adsorbate, A. The adsorbate is then assigned an effective radius \( R_i \) and \( \theta_i \), an orientation \( \theta_j \) is assigned a concentration

\[
C_{i,j} = C_i \frac{\sin \theta_j}{\sum_k \sin \theta_k}
\]

(4)
to ensure that orientations are correctly weighted and the total concentration summed over orientations is correctly reproduced. Scripts to automate this parameterization based on UA output and adsorbate structure files are available as part of the UnitedAtom repository [26].

We further analyze the results for adsorption of milk components obtained from KMC simulations, specifically focusing on the mean absolute and relative abundance of proteins (10−3 nm²) adsorbed on Al surfaces per unit area (nm²). Table 3 shows the abundances of proteins and lactose on Al surfaces.

The simulations were performed using NPs with a radius of 80 nm, and the results are collected in Table 3. It presents the number concentration and mass abundance of proteins adsorbed on three different Al surfaces, namely Al(100), Al(110), and Al(111). Each protein’s adsorption behavior is quantified in terms of its number concentration (expressed in units of 10⁻³ nm⁻²) and mass abundance (represented as a percentage of the total adsorbed mass). These calculations were performed utilizing the most recent KMC method modifications, including an alternative mode in which the acceptance–rejection criteria for incoming adsorbates are altered to allow replacement of pre-existing adsorbates. We should note that Al(111) has the lowest energy of all three surfaces, according to the Materials Project data, so we expect the adsorption profile in real systems to be similar to that predicted for Al(111).

\[
D = \frac{k_B T}{6\eta} \left( \frac{1}{R_{NP}} + \frac{1}{R_A} \right)^2.
\]

(3)
taking the viscosity \( \eta = 8.9 \times 10^{-4} \) Pa·s. We employ SI units in the above calculation, noting that \( k_a \) must then be multiplied by 1000 to convert from units m³ mol⁻¹ to L mol⁻¹. Desorption rates are found by requiring that \( k_d = k_a \times 10^{-3} \times e^{E_{ads}/k_BT} \), where \( E_{ads} \) is the value obtained for that orientation using UnitedAtom [45]. A concentration is then assigned to the adsorbate based on the bulk concentration of that adsorbate, weighted by the relative abundance of that orientation of the adsorbate if necessary. This means that for protein \( i \) with a bulk concentration of \( C_i \) and a set of orientations \( \theta_k \), an orientation \( \theta_j \) is assigned a concentration

\[
C_{i,j} = C_i \frac{\sin \theta_j}{\sum_k \sin \theta_k}
\]

(4)
to ensure that orientations are correctly weighted and the total concentration summed over orientations is correctly reproduced. Scripts to automate this parameterization based on UA output and adsorbate structure files are available as part of the UnitedAtom repository [26].

We further analyze the results for adsorption of milk components obtained from KMC simulations, specifically focusing on the mean absolute and relative abundance of proteins (10⁻³ nm²) adsorbed on Al surfaces per unit area (nm²). Table 3 shows the abundances of proteins and lactose on Al surfaces.

The simulations were performed using NPs with a radius of 80 nm, and the results are collected in Table 3. It presents the number concentration and mass abundance of proteins adsorbed on three different Al surfaces, namely Al(100), Al(110), and Al(111). Each protein’s adsorption behavior is quantified in terms of its number concentration (expressed in units of 10⁻³ nm⁻²) and mass abundance (represented as a percentage of the total adsorbed mass). These calculations were performed utilizing the most recent KMC method modifications, including an alternative mode in which the acceptance–rejection criteria for incoming adsorbates are altered to allow replacement of pre-existing adsorbates. We should note that Al(111) has the lowest energy of all three surfaces, according to the Materials Project data, so we expect the adsorption profile in real systems to be similar to that predicted for Al(111).

\[
D = \frac{k_B T}{6\eta} \left( \frac{1}{R_{NP}} + \frac{1}{R_A} \right)^2.
\]

(3)
taking the viscosity \( \eta = 8.9 \times 10^{-4} \) Pa·s. We employ SI units in the above calculation, noting that \( k_a \) must then be multiplied by 1000 to convert from units m³ mol⁻¹ to L mol⁻¹. Desorption rates are found by requiring that \( k_d = k_a \times 10^{-3} \times e^{E_{ads}/k_BT} \), where \( E_{ads} \) is the value obtained for that orientation using UnitedAtom [45]. A concentration is then assigned to the adsorbate based on the bulk concentration of that adsorbate, weighted by the relative abundance of that orientation of the adsorbate if necessary. This means that for protein \( i \) with a bulk concentration of \( C_i \) and a set of orientations \( \theta_k \), an orientation \( \theta_j \) is assigned a concentration

\[
C_{i,j} = C_i \frac{\sin \theta_j}{\sum_k \sin \theta_k}
\]

(4)
to ensure that orientations are correctly weighted and the total concentration summed over orientations is correctly reproduced. Scripts to automate this parameterization based on UA output and adsorbate structure files are available as part of the UnitedAtom repository [26].

We further analyze the results for adsorption of milk components obtained from KMC simulations, specifically focusing on the mean absolute and relative abundance of proteins (10⁻³ nm²) adsorbed on Al surfaces per unit area (nm²). Table 3 shows the abundances of proteins and lactose on Al surfaces.

The simulations were performed using NPs with a radius of 80 nm, and the results are collected in Table 3. It presents the number concentration and mass abundance of proteins adsorbed on three different Al surfaces, namely Al(100), Al(110), and Al(111). Each protein’s adsorption behavior is quantified in terms of its number concentration (expressed in units of 10⁻³ nm⁻²) and mass abundance (represented as a percentage of the total adsorbed mass). These calculations were performed utilizing the most recent KMC method modifications, including an alternative mode in which the acceptance–rejection criteria for incoming adsorbates are altered to allow replacement of pre-existing adsorbates. We should note that Al(111) has the lowest energy of all three surfaces, according to the Materials Project data, so we expect the adsorption profile in real systems to be similar to that predicted for Al(111).

\[
D = \frac{k_B T}{6\eta} \left( \frac{1}{R_{NP}} + \frac{1}{R_A} \right)^2.
\]

(3)
taking the viscosity \( \eta = 8.9 \times 10^{-4} \) Pa·s. We employ SI units in the above calculation, noting that \( k_a \) must then be multiplied by 1000 to convert from units m³ mol⁻¹ to L mol⁻¹. Desorption rates are found by requiring that \( k_d = k_a \times 10^{-3} \times e^{E_{ads}/k_BT} \), where \( E_{ads} \) is the value obtained for that orientation using UnitedAtom [45]. A concentration is then assigned to the adsorbate based on the bulk concentration of that adsorbate, weighted by the relative abundance of that orientation of the adsorbate if necessary. This means that for protein \( i \) with a bulk concentration of \( C_i \) and a set of orientations \( \theta_k \), an orientation \( \theta_j \) is assigned a concentration

\[
C_{i,j} = C_i \frac{\sin \theta_j}{\sum_k \sin \theta_k}
\]

(4)
to ensure that orientations are correctly weighted and the total concentration summed over orientations is correctly reproduced. Scripts to automate this parameterization based on UA output and adsorbate structure files are available as part of the UnitedAtom repository [26].

We further analyze the results for adsorption of milk components obtained from KMC simulations, specifically focusing on the mean absolute and relative abundance of proteins (10⁻³ nm²) adsorbed on Al surfaces per unit area (nm²). Table 3 shows the abundances of proteins and lactose on Al surfaces.
Table 3: Mean amounts of proteins adsorbed on Al surfaces per unit area: number concentration (per nm$^2$) and mass abundance obtained from KMC simulations with NPs of radius 80 nm. These calculations have been done using the KMC method with displacements.

<table>
<thead>
<tr>
<th>Protein</th>
<th>$N_{ads}$ [$10^{-3}$, nm$^{-2}$]</th>
<th>$M_{ab}$, %</th>
<th>$N_{ads}$ [$10^{-3}$, nm$^{-2}$]</th>
<th>$M_{ab}$, %</th>
<th>$N_{ads}$ [$10^{-3}$, nm$^{-2}$]</th>
<th>$M_{ab}$, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS1C</td>
<td>12.26</td>
<td>57.16</td>
<td>16.70</td>
<td>67.82</td>
<td>27.21</td>
<td>83.19</td>
</tr>
<tr>
<td>BC</td>
<td>4.45</td>
<td>21.24</td>
<td>3.38</td>
<td>14.07</td>
<td>1.91</td>
<td>5.84</td>
</tr>
<tr>
<td>BLAC</td>
<td>2.91</td>
<td>10.99</td>
<td>2.97</td>
<td>9.79</td>
<td>1.00</td>
<td>2.43</td>
</tr>
<tr>
<td>LAC</td>
<td>96.59</td>
<td>6.28</td>
<td>89.13</td>
<td>5.05</td>
<td>84.50</td>
<td>3.62</td>
</tr>
<tr>
<td>ALAC</td>
<td>1.14</td>
<td>3.51</td>
<td>1.13</td>
<td>3.05</td>
<td>1.84</td>
<td>3.60</td>
</tr>
<tr>
<td>AS2C</td>
<td>0.11</td>
<td>0.55</td>
<td>0.04</td>
<td>0.16</td>
<td>3.00</td>
<td>1.09</td>
</tr>
<tr>
<td>BSA</td>
<td>0.02</td>
<td>0.25</td>
<td>0.00</td>
<td>0.05</td>
<td>0.02</td>
<td>0.21</td>
</tr>
</tbody>
</table>

We also compared the protein composition in the corona on aluminum and iron [2], obtained in our previous work using the original KMC approach without molecular displacements. This comparison is shown in Figure 8. AS1C exhibited the highest abundance on both iron and aluminum among the studied proteins, indicating a strong affinity for both metals with both KMC methods as well as its high number concentration in solution. The following AS1C, BC, BLAC, and ALAC also showed fairly equal abundances on the surfaces of iron and aluminum. In contrast, BSA displayed the lowest abundance on both metals because of its larger size and the relatively low molar fraction in milk as compared with other proteins. Figure 8 shows the mass abundance of each protein on both aluminum (Al(100), Al(110), and Al(111)) and iron (Fe(100), Fe(110), and Fe(111)) surfaces. We can also observe that AS1C, BLAC, and ALAC display significantly enhanced presence on Fe surfaces in contrast to Al. Conversely, AS2C shows greater adsorption on Al surfaces as compared to Fe. Overall, we expect a somewhat different corona formed on these metallic surfaces.

Real-life organic media do not consist only of proteins, but they also include many other molecules, for example, sugars and other organic compounds that may bind to NPs along with proteins. It can reasonably be assumed that these molecules may alter both the kinetics and equilibrium state of the corona and, moreover, may play a role in biological outcomes. Thus, it is of interest to include these small molecules in the corona simulation to not only gain further insight into this particular case of...
aluminum in milk, but also to establish a methodology by which more general molecules can be included in these simulations. We choose lactose as a prototypical example of a small molecule capable of binding to NPs, since it is present at a high concentration in milk. We model the lactose molecule as a pair of glucose beads separated by a distance determined by the equilibrium structure of lactose. Although this is not completely rigorous, it demonstrates how the UnitedAtom software can be adapted to model larger molecules other than proteins using the same fragment-based approach. To avoid the need to run a time-consuming parameterization protocol based on metadynamics simulations, we produce PMFs for the glucose bead using a machine-learning technique (PMFPredictor) trained on previous metadynamics results [38]. For the lactose molecule, each constituent glucose bead is assigned a charge of 0, and the Hamaker term is neglected because of the small size of these beads. Following this parameterization, the coarse-grained lactose molecule is processed identically to proteins using the same automated pipeline, that is, UnitedAtom is run to produce a table of orientation-specific binding energies. These are mapped to rate constants for adsorption and desorption. We stress that this procedure is sufficiently generic that essentially arbitrary organic molecules can be included in the simulation by performing a fragment-based decomposition, generating PMFs via traditional or machine-learning approaches, and constructing a coarse-grained representation for input to UA. To simplify this procedure for more complex molecules, we have developed a Python script (MolToFragments.py) employing RDKit [46] to automate splitting larger molecules into suitable fragments and producing coarse-grained input files suitable for UnitedAtom and included it in this repository [26].

The addition of lactose (or other small molecules) to the corona simulation poses a challenge for the form of the CoronaKMC algorithm previously employed because of the high concentration and very small binding area of this small molecule relative to proteins [16,45]. As a consequence of these factors, the original form of the algorithm results in rapid coverage of the NPs with a very large quantity of lactose. This greatly increases the required computational time, which scales as $O(N^2)$ for $N$ adsorbed particles. Moreover, in this original form of the model, a single adsorbed lactose molecule inhibits the adsorption of a large protein, no matter how strongly the protein may adsorb. To counteract these issues, the following features were added to the new version of the CoronaKMC software. First, we implemented a method to accelerate the simulation by adjusting rate constants for quasi-equilibrated processes (e.g., the adsorption of lactose) according to the methodology of Dybeck and co-workers [47]. Second, we added an optional mode in which the acceptance–rejection criteria for an incoming adsorbate are modified such that an incoming adsorbate is no longer immediately rejected if it overlaps with a pre-existing adsorbate. Instead, the incoming adsorbate is accepted with a probability $p$ given by,

$$p(\Delta E) = \frac{\exp[-\Delta E/k_B T]}{1 + \exp[-\Delta E/k_B T]},$$

where $\Delta E$ is the difference in energy between the two states,

$$\Delta E = E_{\text{ads}} - \sum_j E_j,$$

where $j$ is the set of all adsorbed particles that would overlap with this particle, taking $\Delta E = E_{\text{ads}}$ if no overlaps are found. If the adsorbate is accepted, then all the overlapping particles are removed from the NP. We note that this breaks the principle of detailed balance in that it allows for the replacement of a set of adsorbates by a single molecule, but does not allow for the converse in which a set of incoming molecules can displace an adsorbate. We justify this neglect on the basis that the required event of multiple simultaneous collisions on a single target would occur so rarely that it would essentially not be sampled in the course of a simulation. The probabilistic acceptance to regions of the NP without explicit adsorbates present effectively multiplies the adsorption rate by a factor of $p(E_{\text{ads}})$. Thus, to maintain the same equilibrium constant, we must multiply the desorption rate by the same factor, noting that this correction is only significant for very weakly adsorbing particles with $E_{\text{ads}} \sim -3k_B T$. This methodology does not treat adsorption of water to the NP explicitly. Instead, it is assumed that all binding energies are defined relative to the adsorption of water, which is assigned an affinity $E_{\text{ads}} = 0k_B T$, and that the concentration of water is sufficiently high such that any region of the NP without an explicit adsorbate can be assumed to be covered in water.

The results of simulations obtained with the updated CoronaKMC (i.e., including the molecule displacement) are shown in Table 3, and they suggest a notable variation in the abundances of proteins and lactose among different Al crystallographic orientations. Notably, on all surfaces studied, AS1C and BC consistently exhibited the highest protein abundances, while BLAC, LAC, and ALAC demonstrated moderate adsorption levels. In contrast, AS2C and BSA consistently displayed the lowest adsorption among the proteins considered in our simulations. Furthermore, when considering different Al facets, it is evident that the (110) surface consistently exhibited the weakest average adsorption across all proteins. When the displacement is allowed, AS1C gains much more space in the corona by replacing other proteins, mostly BLAC, ALAC, and AS2C.
Figure 9 presents a comparison between the protein abundances in the corona on Al and Fe obtained using the enhanced version of the KMC algorithm with molecular displacements. As discussed earlier, this improved algorithm addresses computational efficiency concerns and more accurately represents long-term scenarios during protein corona formation. As shown in the Figure, these algorithmic improvements have a profound impact on the mass concentration of milk proteins on metallic surfaces, particularly on iron. In the original algorithm (Figure 8), proteins showed comparable mass abundances on both metals. However, the enhanced algorithm reveals a distinct change in the adsorption behavior of the AS1C protein on Fe and Al surfaces, characterized by a substantial increase in mass concentration compared to other proteins. The data in Table 3 show that in terms of mass abundance lactose ranks fourth among the corona components (see Supporting Information File 1, Figure S3). As compared to the algorithm without displacement [2], the protein abundance ranking on iron (NP radius 80 nm) surfaces changes to AS1C ≫ BC ≥ BLAC ≥ ALAC > AS2C ≈ BSA. A comparable affinity ranking is also now observed for aluminum surfaces (80 nm) studied in current work: AS1C ≫ BC ≥ BLAC ≥ ALAC > AS2C ≈ BSA.

Conclusion
In this work, we applied a multiscale computational model to study the adsorption of milk solids on the metallic surfaces of aluminum, widely used in food processing/packaging. The milk model contained the six most common milk proteins and lactose. To account for the size differences of selected milk constituents, we used an improved competitive adsorption algorithm that can potentially achieve a realistic description of biocorona formation processes with diverse adsorbates (e.g., for predicting an eco-corona).

Our computational model predicts strong binding of milk proteins to pure aluminum surfaces, which is in agreement with our previous observations for metallic iron surfaces [2]. For aluminum, we also found that AS1C and AS2C exhibited the strongest binding to the metal, followed by BSA, BC, BLAC, and ALAC, which displayed weaker adsorption. We also found similar protein abundances in the corona for the two metals demonstrated by KMC simulation results. AS1C dominates the adsorption as the most abundant protein on aluminum surfaces, with BSA being the least abundant. We found a small difference in the predicted corona content between the two metals: BC and BLAC prefer Al(100) and Al(110) to iron, while AS1C prefers Fe(100) and Fe(110) over aluminum.

Although the adsorption energy regulates the interaction strength between proteins and surfaces, the mass concentration of proteins in the solution has a major effect on the amount of protein adsorbed onto the surface. Expanding the milk model by adding lactose into the mix did not alter the ranking of protein abundance in the corona. Despite the high concentration in the milk, lactose does not exceed the mass abundance of specific proteins such as AS1C due to its small size. In our model, it essentially forms a thin monolayer on the surface.

Overall, our freely accessible multiscale computational model [26] allows us to make predictions of the binding strength, preferred orientations, and relative abundance of the specified proteins.
molecules on the specified material surfaces or NPs and, thus, gives an insight into the mechanisms of bionano interaction. We can compare different materials in terms of the protein binding affinity and corona content and optimize the processes in food and chemical industry. The presented methodology can be easily extended to other molecules, materials, and contexts involving the bionano interface such as environmental safety, health, medical devices, or toxicology.

**Supporting Information**

Table S1: Adsorption free energies for each SCA on Al surfaces; Figure S1: Water density profiles for aluminium slabs: (a) Al(100), (b) Al(110), (c) Al(111), Figure S2: Influence of the NP size on the adsorption energies; Figure S3: Milk molecules ranking based on mass abundance in the corona, Figure S4: Example of AlINP size-dependent interaction of ALAC: (a) 2 nm, (b) 5 nm, (c) 10 nm, (d) 20 nm, (e) 40 nm, (f) 50 nm, (g) 80 nm, (d) 100 nm; Figures S5–S10: Comparison of interaction of AS1C, AS2Cβ-casein, ALAC, BLAC, BSA, with different Al fcc surfaces: (a) Al(100), (b) Al(110), (c) Al(111); Table S2: Description of 820 milk proteins interaction with Al (100, 110, 111) based on the lowest energy values of the adsorption heatmaps.

**Supporting Information File 1**
Supporting material.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-21-S1.pdf](https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-21-S1.pdf)

**Supporting Information File 2**
820-Milk-protein-table.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-21-S2.pdf](https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-21-S2.pdf)

**Acknowledgements**
The authors extend their gratitude to the Irish Centre for High-End Computing (ICHEC) and the UCD Sonic High-Performance Computing Centre for providing the necessary computational resources for this research. The graphical abstract of this article was created with https://biorender.com/ and it’s content is not subject to CC BY 4.0.

**Funding**
The research presented in this study was supported by funding from Science Foundation Ireland under grant number 16/IA/4506, the European Union Horizon 2020 Programme grant 814572 (NanoSolveIT), and European Union Europe Programme grant 101092741 (nanoPASS).

**ORCID® iDs**
Parinaz Mosaddeghi Amini - https://orcid.org/0009-0003-4603-6830
Ian Rouse - https://orcid.org/0000-0002-3686-7701
Julia Subbotina - https://orcid.org/0000-0002-2227-0787
Vladimir Lobaskin - https://orcid.org/0000-0002-5231-0639

**References**


**Preprint**
A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.3762/bxiv.2023.45.v1

---

**Supporting Information**

Table S1: Adsorption free energies for each SCA on Al surfaces; Figure S1: Water density profiles for aluminium slabs: (a) Al(100), (b) Al(110), (c) Al(111), Figure S2: Influence of the NP size on the adsorption energies; Figure S3: Milk molecules ranking based on mass abundance in the corona, Figure S4: Example of AlINP size-dependent interaction of ALAC: (a) 2 nm, (b) 5 nm, (c) 10 nm, (d) 20 nm, (e) 40 nm, (f) 50 nm, (g) 80 nm, (d) 100 nm; Figures S5–S10: Comparison of interaction of AS1C, AS2Cβ-casein, ALAC, BLAC, BSA, with different Al fcc surfaces: (a) Al(100), (b) Al(110), (c) Al(111); Table S2: Description of 820 milk proteins interaction with Al (100, 110, 111) based on the lowest energy values of the adsorption heatmaps.

**Supporting Information File 1**
Supporting material.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-21-S1.pdf](https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-21-S1.pdf)

**Supporting Information File 2**
820-Milk-protein-table.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-21-S2.pdf](https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-21-S2.pdf)