Chemical glycobiology

Edited by Elisa Fadda, Rachel Hevey, Benjamin Schumann and Ulrika Westerlind
Abstract
Automated glycan assembly (AGA) affords collections of well-defined glycans in a short amount of time. We systematically analyzed how parameters connected to the solid support affect the AGA outcome for three different glycan sequences. We showed that, while loading and reaction scale did not significantly influence the AGA outcome, the chemical nature of the linker dramatically altered the isolated yields. We identified that the major determinants of AGA yields are cleavage from the solid support and post-AGA purification steps.

Introduction
Automated glycan assembly (AGA) is a solid-phase method that enables the rapid synthesis of complex oligo- and polysaccharides from protected monosaccharide building blocks (BBs) [1,2]. Iterative cycles of glycosylation, capping, and selective deprotection afford the support-bound glycan with a programmable sequence (Figure 1A). The protected glycan is then cleaved from the solid support and subjected to post-AGA deprotection steps to reveal the target glycan. AGA is mostly performed on cross-linked polystyrene resins equipped with photocleavable linkers [3], offering orthogonality to all the synthetic steps of the assembly, while selectively releasing the glycan at the end of the synthesis.

In recent years, the implementation of new synthetic strategies [4-7] as well as technological improvements [8,9] permitted access to highly complex carbohydrates [10]. Still, variations in yields are not always ascribable to the AGA process [11-16]. Dissimilar structures are assembled in high purity as indicated by HPLC analysis of the crude products, but isolated in relatively low yields. The optimization procedures are focused on glycan elongation (i.e., glycosylation and deprotection steps), whereas less attention is given to variables associated with the solid support [17]. In contrast, substantial knowledge exists on how loading [18], reaction scale [19], and linkers [20,21] affect the overall yield of solid phase peptide synthesis (SPPS). In the
past decades, several supports and linkers have been developed and commercialized for SPPS, enabling a wide range of applications. Solid supports are available with different linker loadings, with low loading (0.1–0.2 mmol/g) being beneficial to avoid aggregation of long peptide sequences, and high loadings (0.4–0.5 mmol/g) advantageous for more efficient syntheses [21].

Herein, we systematically investigate how variations in linker type, resin loading, and reaction scale influence the productivity of AGA.

Results and Discussion

We selected three glycan sequences as models to analyze the effect of different parameters on the AGA outcome. Each sequence was prepared on four batches of Merrifield resin functionalized with two photolabile linkers (L1 [22] vs L2 [3]), at two linker loadings (low vs high) (Figure 1B). Each AGA experiment was performed at two different reaction scales (15 vs 30 µmol). All AGA runs were performed adjusting the resin amount to the desired reaction scale, while keeping the concentration of all other reagents constant (Figure 1B).

The photolabile linkers L1 [22] and L2 [3] are based on the o-nitrobenzyl scaffold [23,24] and expose a hydroxy group that serves as glycosyl acceptor in the first AGA cycle (Figure 1B). While L1 displays a flexible aliphatic chain terminating with a primary alcohol, L2 carries a secondary benzylic alcohol. Upon irradiation with UV light (λ = 360 nm), L1 releases the glycan equipped with an aminoalkyl spacer at the reducing end, whereas L2 affords the free reducing sugar (α/β mixture). Previous data suggested that UV cleavage of L1 and L2 was equally efficient, permitting the isolation of a tetramannoside in around 60% yield [3]. We wondered whether different glycan sequences were more sensitive to the linker structure. Less reactive donors might highlight differences in the linker nucleophilicity [25]. The aggregation of the growing glycan chains is conceivable to be connected to linker flexibility [18]. The efficiency of UV cleavage is probably influenced by glycan structure, solubility, and aggregation tendency [26]. Lastly, purification of the protected glycan upon cleavage could be affected by the presence or absence of a linker.

L1 or L2 were conjugated to Merrifield resins with initial loadings of 0.5 mmol/g and 1.0 mmol/g to yield supports with low (0.3–0.4 mmol/g) or high (0.7–0.8 mmol/g) loadings (see Supporting Information File 1, section 2.3, module A). The latter allows for a larger synthesis scale, but steric hindrance and chain–chain interactions could negatively influence the AGA outcome, as observed for some peptide sequences [18]. Moreover, high-loading supports might result in inefficient UV cleavage due to quenching. These four supports were studied in AGA experiments performed at 15 and 30 µmol reaction scales. While AGA is commonly performed at a 15 µmol reaction scale, a larger reaction scale is attractive to produce more material in a single AGA run, but might suffer from insufficient mixing [27,28], causing slower kinetics [29], temperature gradients [30], and precipitation [31].
We set off to study the effect of these parameters on the AGA of three different glycan sequences (Figure 2). In an increasing order of complexity, we prepared α-1,6-linked dimannosides (1, 2) [32], branched trisaccharides (3, 4) [12], and linear α-1,4-linked hexaglucosides (5, 6) [15, 33]. Each synthesis was performed with 6.5 equivalents of BB per glycosylation cycle.
using previously reported AGA conditions (see Supporting Information File 1, section 2.3, module C). The outcome of each AGA experiment was analyzed in terms of: i) HPLC purity based on the chromatogram of the crude sample after AGA and UV cleavage, ii) isolated yield of the target compound after photocleavage and HPLC purification (path A).

The syntheses of the α-1,6-linked dimannosides 1 and 2 (Figure 2A) were successful on all resins tested, affording the desired product in complete purity regardless of linker type, loading or reaction scale (Figure 3A, and Figures S2 and S3 in Supporting Information File 1). Isolated yields of 49–59% were obtained in all experiments (Figure 2D), after cleavage of the photolabile unit.

The syntheses of the branched trisaccharides 3 and 4 (Figure 2B, and Figures S4 and S6 in Supporting Information File 1) were less efficient. Even though the target compound was the major product in all experiments, deletion sequences were observed in the chromatograms of the crudes (Figure 3B). MS analysis showed the presence of capped linker (†), capped dimer (‡), and Lev-containing dimer (‡) (see Figures S5 and S7 in Supporting Information File 1). No significant variations were noticed within each series of experiments, with slightly better purities obtained for AGA performed on L2 (to note: for experiments on L2 no capped linker was detectable by HPLC; see Supporting Information File 1). Isolated yields were relatively low for all experiments (14–32% on L1 and 29–44% on L2, Figure 2D). These values are quite low even considering the presence of deletion sequences, suggesting that cleavage and purification are more challenging for these structures. Overall, a slightly better performance of L2 resulting in higher purities and better yields was noticed.

HPLC analysis showed that the β-1,4-hexaglucosides 5a and 6a were produced in excellent purity in all experiments (Figures 2D, 3C, and Figures S8 and S9 in Supporting Information File 1). For these compounds, we explored two different post-AGA procedures: the standard path A based on photocleavage and HPLC purification, and path B involving on resin methanolysis of the ester groups, photocleavage, hydrogenolysis of the remaining PGs, and purification (Figure 2C). The latter is commonly employed for compounds synthesized on L2 because of the poor stability of free-reducing glycans in basic conditions needed for the methanolysis step [33].

The isolated yields of the fully protected compound 5a synthesized on L1 were significantly lower than expected (21–36%, Figure 2D, path A), with little variation within the series. Isolated yields for the linker-free compound 6a prepared on L2 were around 10% higher (46%). The absence of deletion sequences in the HPLC of the crude compounds indicated that cleavage and/or purification are the major bottlenecks of these syntheses.

Higher yields (30–57%) were obtained for compound 6b, isolated after the post-AGA procedure path B (Figure 2D). This is surprising since the path B procedure involved additional deprotection steps. Therefore, we wondered whether methanolysis on resin could improve photocleavage efficiency. However, when we tested the same procedure on L1, target compound 5b was isolated in only 15% yield. These results strongly suggest that the two linkers perform differently depending on the glycan sequences.
Conclusion
Taken together, the results showed minimal variation within each series of experiments, indicating that loading and reaction scale are not significantly affecting AGA of those sequences within the range of conditions explored here. This is a promising observation from the perspective of scaling up AGA. No differences were observed for the AGA of simple disaccharides L2 performed on L1 and L2 with an apparent maximal yield of around 60%, in agreement with previous reports [3]. In contrast, other sequences constructed on L2 were isolated in slightly better yields. This result could be connected to more efficient cleavage of L2 in the presence of complex glycan sequences, easier purification of linker-free compounds, or a combination of both.

Our systematic study identified that the major determinants of AGA yields are cleavage from the solid support and purification steps. These two aspects are strongly connected to the glycan structure, with minimal variations such as presence or absence of a linker playing an important role in the post-AGA process. In some cases, performing post-AGA manipulations on resin dramatically improved the overall yield of the process. Future efforts need to focus on the development of new linkers, more efficient cleaving procedures [34], and the implementation of post-AGA manipulation steps on resin.

Supporting Information
Supporting Information File 1
Experimental procedures and characterization data. [https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-19-77-S1.pdf]

Funding
We thank the Max Planck Society, the German Federal Ministry of Education and Research (BMBF, grant number 13XP5114), and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation – SFB 1449 – 43123613; sub-project C2) for generous financial support.

Conflict of interest statement
The authors declare no conflict of interest.

ORCID® iDs
Marlene C. S. Dal Colle - https://orcid.org/0000-0006-1062-0395
Manuel G. Ricardo - https://orcid.org/0000-0003-2365-2864
Nives Hribnik - https://orcid.org/0000-0003-1347-3192
Peter H. Seeberger - https://orcid.org/0000-0003-3394-8466
Martina Delbianco - https://orcid.org/0000-0002-4580-9597

References
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/bjoc/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/bjoc.19.77
Abstract
Carbohydrate sequencing is a formidable task identified as a strategic goal in modern biochemistry. It relies on identifying a large number of isomers and their connectivity with high accuracy. Recently, gas phase vibrational laser spectroscopy combined with mass spectrometry tools have been proposed as a very promising sequencing approach. However, its use as a generic analytical tool relies on the development of recognition techniques that can analyse complex vibrational fingerprints for a large number of monomers. In this study, we used a Bayesian deep neural network model to automatically identify and classify vibrational fingerprints of several monosaccharides. We report high performances of the obtained trained algorithm (GIAlcomics), that can be used to discriminate contamination and identify a molecule with a high degree of confidence. It opens the possibility to use artificial intelligence in combination with spectroscopy-augmented mass spectrometry for carbohydrates sequencing and glycomics applications.

Introduction
DNA and protein sequencing technologies that aim at determining the structure of a biopolymer have been established decades ago and are commonly used in a routine and automated manner. However, the development of such technology for the sequencing of the third class of biological polymer – glycans, also known as carbohydrates, saccharides, or "sugars" – lags far behind. This lack of dedicated analytical tools (glycomics) is clearly identified as a critical bottleneck, impeding the full development of glycosciences despite their relevance for various strategic fields such as pharmaceutical and food industry; bio-based materials and renewable energy, and their considerable potential impact for the society in regard to the United Nations sustainable development goal [1].

The major roadblock to carbohydrate sequencing is intrinsically due to their unique molecular properties, among biopoly-
mers. In contrast with proteins and DNA, which are linear polymers made of a limited number of building blocks with distinct molecular structures, carbohydrates feature hundreds of building blocks—many of them coming in groups of closely related isomers with ambiguous molecular structures—and they form complex, branched arrangements due to the versatility of the glycosidic bond (position and anomericity). In this context, designing generic carbohydrate sequencing methods is both a major scientific challenge and a strategic priority [2,3].

Few years ago we proposed an original solution by bringing together the best of both sides of the analytical chemistry world: Spectroscopy and mass spectrometry (MS). In short, our technology is based on a mass spectrometric analysis—which is particularly powerful for the analysis of complex biological samples but does not readily elucidate isomers which have the same molecular mass—augmented with an infrared laser-based spectroscopic dimension (MS–IR), thus providing valuable additional isomer resolution [4].

We demonstrated that this multidimensional MS–IR molecular fingerprint is unique to each carbohydrate building block and can be used to resolve their full sequence, including their monosaccharide content and the detail of their linkages (position and anomericity). Based on this basic principle, the identification of an unknown carbohydrate proceeds as follows: the polymer is fragmented in monomers, yet maintaining information on the initial structure and the spectroscopic fingerprint (frequency and intensity of the vibrational modes) of each monosaccharide unit is measured, and subsequently identified by comparison with a library of reference spectra of synthetic monosaccharide standards. In the early days of MS–IR spectroscopy, ca. one hour was necessary to record the IR fingerprint of a single molecule and the identification was made by visual inspection, which was shortly automated by introducing a score derived from the convolution between the spectrum of the analyte of interest and the library of reference spectra. Despite the advantage of being automated, this later approach remains biased: for each molecular species, a single spectrum is arbitrarily chosen by the operator and serves as reference for all future analyses.

The latest MS–IR developments brought the data collection down to few seconds [5]. This is a considerable step towards high throughput carbohydrate analysis, which must be accompanied by fast data analysis, thus excluding manual interpretation. Besides, in the prospective of deploying the technology beyond the molecular spectroscopy community, it is essential to develop an automated, reliable, and robust strategy for the analysis of the spectroscopic data. Machine learning methods appear to be appealing candidates to address this challenge. They have been used for mass spectrometry data analysis since the 2000’s [6] and the idea of using them on vibrational spectra goes back to the early 90’s [7]. Support vector machines (SVM) and decision tree ensemble methods were benchmarked on infrared spectra for cancer classification [8] and many research groups focused their efforts on using machine learning for simulating molecular structures; generating vibrational spectra; and classifying chemical groups based on vibrational features [9,10]. In a recent publication, the random forest approach was proposed to identify the presence of structural features in oligosaccharides based on their gas-phase IR spectra [11]. To the best of our knowledge, machine learning classification studies have not been reported to identify saccharides using MS–IR carbohydrate analysis.

Here, we report a study of a probabilistic deep neural network (Bayesian deep neural networks [12]) to support automated monosaccharide recognition for carbohydrate sequencing. We obtained a highly performing algorithm that we called “GIAlcomics”, specifically trained on carbohydrates.

Methodology

Data production

Our carbohydrate analysis approach is based on the IRMPD spectroscopic scheme (infrared multiple photon dissociation), which is the combination of mass spectrometry and IR spectroscopy. IRMPD is an action spectroscopy method that allows recording IR absorption spectra of isolated gas-phase ions, based on the measurement of the wavelength-dependent laser-induced fragmentation yield. When the frequency of the laser is resonant with a vibrational mode of the molecule, the molecule absorbs the radiation and accumulates internal energy until fragmentation [13]. In previous works we have demonstrated that the monosaccharides or oligosaccharides resulting from the fragmentation of a larger precursor possess a very specific IR fingerprint in the 2–4 microns spectral range, that is highly valuable to resolve all types of isomers [4]. Typical experimental IR fingerprint data are shown in Figure 1: they feature the intensities of the vibrational resonances as a function of their frequency in the mid-IR range. After measuring its mass and its IR fingerprint, an unknown analyte (Figure 1a) is readily identified as “GlcNAc” (for N-acetylglucosamine) by comparison with the reference IR spectra of several candidates of identical mass (Figure 1b, featuring three stereoisomers of C4H4S2NO6).

With the rapid development of our approach, such method now reached a high data output since a single IR fingerprint can be obtained in few seconds. The fast and automatic identification and classification of the data becomes compulsory, which motivates the present study.

For this study, a first set of 33 labelled experimental spectra obtained as described previously [4] were collected for training...
and validation of the model. The standard instrumental conditions for recording MS–IR data consist in a laser-enabled mass spectrometer equipped with a 3D ion trap mass analyzer. The following monosaccharides were analyzed: three stereoisomers of hexosamine of chemical formula C_6H_{13}NO_5, namely glucosamine (GlcN), galactosamine (GalN), mannosamine (ManN); and N-acetyl glucosamine (GlcNAc, chemical formula C_8H_{15}NO_6). One typical spectrum of each of the four monomers is shown in Figure 2. Note that both α and β-anomers coexist in the experimental conditions.

The second set of experimental MS–IR spectra was acquired using different instrumental conditions on a different experimental set-up: it consists of the coupling of an alternative design of mass spectrometer (equipped with a 2D ion-trap mass analyzer) with a higher repetition rate laser and a larger spectral bandwidth [5]. New GlcN spectra were acquired in these conditions. One of them is shown in Figure 3 (orange trace) for comparison with an experimental spectrum of GlcN acquired in standard conditions. Due to the larger spectral bandwidth, the spectrum from set 2 looks significantly different: the peaks are broader and less resolved than in the spectrum from set 1. This set is referred to as exogenous and was not used for training: it is used to illustrate the robustness of the method across significantly variable experimental conditions and instrumental performance.

The third set of experimental IRMPD spectra was acquired in standard conditions and includes 5 new spectra from the monomers GlcN, GalN, and ManN as in sets 1 and 2; as well as 7 spectra from species that do not belong in the training set categories (out of distribution, OOD), including disaccharides, a sulfated monosaccharide, and paracetamol. The outlying molecules represent potential "pollutions" in the analysis. This set of data is referred to as endogenous as it was measured on the same apparatus as the training set.
For efficient training of the algorithms, all three experimental datasets were augmented by producing synthetic variants. These synthetic spectra were generated by modulating the experimental ones with the following relevant sources of experimental fluctuations:

- The signal to noise ratio may vary from one measurement to another as it can emerge from a low amount of molecules. This was simulated by adding a Gaussian white noise with a randomly distributed standard deviation between 0 and 5% of the peak signal.
- The overall intensity of the laser can fluctuate from day to day or thorough the entire spectral range, which results in modulated peaks amplitudes. This was simulated as a linear variation of the signal amplitude across the spectral range. The variation was contained in a uniform distribution bounded by ±10%.
- Spectra can be recorded at increased speed for rapid analytical diagnostics, which traduces into a change in binning. To take this into account, data were binned with downgraded resolution then re-binned with 1 cm$^{-1}$ step. The down sampling factor was randomly picked in a range from 1 to 5.
- Small variations of the calibration of the laser wavenumber may occur from day to day, leading to a shift of few wavenumbers of the vibrational spectrum. This was simulated with a maximum random shift per spectrum of ±10 cm$^{-1}$.

Finally, the synthetic spectra were normalized by z-score and interpolated over 1200 bins in the 2600–3800 cm$^{-1}$ spectral range (1 cm$^{-1}$ step) as input vector for the neural network. An example of a synthetic spectrum generated from an experimental spectrum is shown in Figure 3.

A total of 8000 synthetic spectra were randomly produced (2000 for each monomer category) out of the experimental spectra of set 1. They were shuffled to avoid training batches composed of a unique category of molecules. Finally, 70% of them were used for training of the models, and 30% were used for validation. The composition of the datasets used for training, validation and tests is summarized in Table 1.

**Model architecture**

In this study we opted for a fully connected feed-forward network based on the multi-layer perceptron architecture [14].

---

**Table 1: Composition of the three datasets.**

<table>
<thead>
<tr>
<th>Dataset 1</th>
<th>Dataset 2</th>
<th>Dataset 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>training : 70% validation : 30%</td>
<td>classification tests</td>
<td>discrimination tests</td>
</tr>
<tr>
<td>categories</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>acquisition</td>
<td>standard</td>
<td>low res.</td>
</tr>
<tr>
<td>exp. MS–IR spectra</td>
<td>33</td>
<td>4</td>
</tr>
<tr>
<td>augmented set</td>
<td>8000</td>
<td>8000</td>
</tr>
</tbody>
</table>
with probabilistic approach (Bayesian deep neural network, DNN), which allows quantifying the model uncertainty for the classification results. It is composed of 3 hidden layers of 300, 225, and 100 neurons, respectively, and ReLu (rectified linear unit) activation functions for each layer. Two dropout layers are interleaved after the first and second hidden layers with a dropout setting of 25% to avoid over-fitting issues. The training objective is a classification task between the 4 monomer categories with a cross-entropy loss function.

To account for the probabilistic nature of the deep neural network, we used the variational inference technique. Each deterministic weight parameter was replaced by normal distributions defined by a mean value μ and a standard deviation σ which were optimized using the Bayes-by-Backprop method [15]. We chose this method that constrains the weights posterior distribution to normal distributions instead of the more accurate Markov chain Monte-Carlo (MCMC) method for calculation efficiency. With this approach, a quantitative uncertainty of the model predictions can be achieved by inferring each spectrum category several times with the trained model.

Results and Discussion

Model classification accuracy

Our GlAIcomics model shows a classification accuracy of 100% on the validation set and 99.98% on the test set (S.M : dataset 2 in Table 1). The 8000 synthetic spectra of set 2 were sorted by noise level, amplitude modulation, energy shift, and downsampling. The mean accuracy of the model as a function of these four parameters is shown in Figure 4. Note that all parameters have a uniform distribution over the 8000 samples and can be studied independently. The amplitude modulation and downsampling do not play a major role, with a maximum accuracy variation of 0.5%.

We demonstrated that the neural network is suitable for MS–IR classification in experimental conditions with variable resolution, noise or energy jitter. The question remains on how to discriminate unknown molecules or to identify problematic spectra, such as the few misclassification events in the discussion above. In order to address these points, we further assessed the precision of the model and discussed its epistemic uncertainty in the next section.

Model precision and uncertainty

In the context of analytical chemistry where the fraction of "known molecules" (that is, previously referenced in databases) is expected to be significant compared to unknown ones, it is important to make sure that the model is discriminative and we want to maximize the precision of the model at this task. Indeed, the large amount of positive results would make it difficult to identify false positives. However, a small number of negative results is expected, which makes it doable to assess them systematically. False negative could be identified manually, labelled correctly, and injected back to improve the model.

The third dataset was used to evaluate the model discriminative power. It consists of 1300 spectra produced by augmentation of 12 original experimental spectra that were acquired on the standard instrumental setup and were never used by the models during the training and validation phases. This set contains 3 of the 4 known monosaccharides: ManN, GlcN, and GalN as well as 8 other molecules. For benchmarking purposes, all spectra were annotated with true labels.

By running the model inference for one spectrum multiple times we can measure the variability of its prediction probability for each category. If the model gives consistently a high probability for one category after each inference, then its uncertainty is low, and the spectrum likely belongs to the said category of molecules. On the other hand, if the model predicts a category with highly variable probability, then the uncertainty is high, and the spectrum likely does not belong to any of the classification categories. We ran model inference 200 times on each sample and obtained the mean prediction probability for every cate-
Figure 5: DNN Prediction results for third endogenous dataset (5 hexosamine samples and 7 other molecules). The middle map shows the mean prediction probabilities for each category and the right hand side map shows the 5% to 95% interpercentile range for the prediction probability distributions of each category.

Figure 5: DNN Prediction results for third endogenous dataset (5 hexosamine samples and 7 other molecules). The middle map shows the mean prediction probabilities for each category and the right hand side map shows the 5% to 95% interpercentile range for the prediction probability distributions of each category.

Most known molecules are assigned to the right category with a very sharp probability distribution that can be used as the prediction distribution under the null hypothesis that the model is reliable. For most of the “unknown” molecules the model prediction oscillates between two categories, but the probability distributions are extremely broad which means that the neural network uncertainty is important, and the corresponding results should be considered as unclassified and put aside for manual evaluation.

Finally, the performance of the GlAIcomics deep neural network model was compared with two different off-the-shelf techniques based on decision trees: Random forest (RF), an XGBoost (XGB). The evaluation methods are detailed in Supporting Information File 1. The classification accuracy for the validation subset (30% of set 1) is 100%, 99.95% and 100% for RF, XGBoost and GlAIcomics, respectively. For the test set (dataset 2), the accuracy is 99.91%, 99.61%, and 99.98%, respectively. When the accuracy of the prediction is further investigated as a function of the data augmentation parameters used to model experimental fluctuations, an advantage is found for GlAIcomics and RF over XGBoost. Lastly, the three methods were compared for the discrimination of molecules outside of the known categories. GlAIcomics appears to discriminate samples more efficiently than the two other methods with true and false positive rates above 80% (70% and 50% for RF and XGBoost, respectively).

Conclusion

We have evaluated the performances of a Bayesian deep neural network for automatic analysis and classification tasks on glycans MS–IR fingerprints. It showed robust prediction accuracies on an exogenous dataset. We observed that it is capable to generalize as it could categorize more noisy and distorted spectra. We then benchmarked its discrimination capabilities with a mixture of hexosamines and other molecular spectra: the Bayesian neural network architecture offers an access to the model reliability (through its epistemic error) when it comes to classify the spectra and could be used to discriminate outlying molecules or experimental issues when run on new data samples. Therefore, we conclude that a relatively small Bayesian deep neural network is a suitable solution for analysis and classification of saccharides in the context of MS–IR based carbohydrate sequencing. It can be easily integrated in an experimental data pipeline between the experiment raw spectra recording and the sequencing algorithm. Rejected spectra would be manually reviewed and fed back to the model as new training samples which in turn would reduce the epistemic error. It will therefore speed up the construction of glycans spectroscopic fingerprints database. In MS–IR experiments, the IR data as well as the mass of the molecule are simultaneously acquired,
therefore the mass could readily be used as a prefilter. More generally, all experimental data obtained in a glycomics workflow – such as MS/MS; HPLC; ion mobility; … – could ultimately be included in the algorithm for an optimal coverage of complex carbohydrates.

Supporting Information
Supporting Information File 1
Evaluation of the deep neural network model against two different techniques based on decision trees: Random forest (RF) and XGBoost (XGB).

Funding
This work was supported by Region Auvergne Rhône Alpes (IROLIGO grant), ANR Algaims (ANR-18-CE29-0006-02) and LABEX IMUST (ANR-10-LABX0064).

Author Contributions
B.S. and I.C. ran the spectroscopy experiments. T.R.B. proposed, trained and evaluated the machine learning model. T.R.B., B.S., B.M., E.F., F.L. and I.C. participated to the writing of the article

ORCID® iDs
Baptiste Schindler - https://orcid.org/0000-0002-7376-4154
Baptiste Moge - https://orcid.org/0000-0002-4932-6357
Isabelle Compagnon - https://orcid.org/0000-0003-2994-3961

References
5. Yeni, O.; Schindler, B.; Moge, B.; Compagnon, I. Analyst 2022, 147, 312–317. doi:10.1039/d1an01670a

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/bjoc/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/bjoc.19.134
Studying specificity in protein–glycosaminoglycan recognition with umbrella sampling

Mateusz Marcisz‡1,2, Sebastian Anila‡1, Margrethe Gaardløs1, Martin Zacharias3 and Sergey A. Samsonov*1

Full Research Paper
Address:
1Faculty of Chemistry, University of Gdańsk, Gdańsk, Poland,
2Intercollegiate Faculty of Biotechnology, University of Gdańsk and Medical University of Gdańsk, Gdańsk, Poland and 3Physics Department, Technical University of Munich, Garching, Germany
Email:
Sergey A. Samsonov - sergey.samsonov@ug.edu.pl
* Corresponding author ‡ Equal contributors
Keywords:
glycosaminoglycan; molecular docking; protein–glycosaminoglycan interaction specificity; RS-REMD; umbrella sampling

Abstract
In the past few decades, glycosaminoglycan (GAG) research has been crucial for gaining insights into various physiological, pathological, and therapeutic aspects mediated by the direct interactions between the GAG molecules and diverse proteins. The structural and functional heterogeneities of GAGs as well as their ability to bind specific proteins are determined by the sugar composition of the GAG, the size of the GAG chains, and the degree and pattern of sulfation. A deep understanding of the interactions in protein–GAG complexes is essential to explain their biological functions. In this study, the umbrella sampling (US) approach is used to pull away a GAG ligand from the binding site and then pull it back in. We analyze the binding interactions between GAGs of three types (heparin, desulfated heparan sulfate, and chondroitin sulfate) with three different proteins (basic fibroblast growth factor, acidic fibroblast growth factor, and cathepsin K). The main focus of our study was to evaluate whether the US approach is able to reproduce experimentally obtained structures, and how useful it can be for getting a deeper understanding of GAG properties, especially protein recognition specificity and multipose binding. We found that the binding free energy landscape in the proximity of the GAG native binding pose is complex and implies the co-existence of several binding poses. The sliding of a GAG chain along a protein surface could be a potential mechanism of GAG particular sequence recognition by proteins.

Introduction
Glycosaminoglycans (GAGs) are long linear periodic anionic polydisperse polysaccharides, with repeating disaccharide units comprised of a hexuronic acid (or galactose in keratan sulfate) and a hexosamine (N-acetylglucosamine, GlcNAc or N-acetyl-galactososamide, GalNAc) throughout a regular alternation of 1→4 and 1→3-glycosidic linkages [1-3]. GAGs are mainly located on the cell surface and in the extracellular matrix [4]. Due to their charged nature, they bind a large amount of water [5]. Although GAGs were previously considered just an inert glue surrounding the cell, GAG research in the past few decades has
illustrated the crucial role in cell signaling processes, including regulation of cell growth, proliferation and promotion of cell adhesion, anticoagulation, and wound repair [6-9]. All these processes are mediated through their direct interactions with diverse protein targets such as collagen, chemokines [10,11], and growth factors [12-14], which makes them essential in the cell biology [15,16]. In addition, GAGs also facilitate cell migration, act as shock-absorbers in joints and as a sieve in extracellular matrices and are important in maintaining the compressibility of the cartilage. The participation of GAGs in physiological, pathological, and therapeutic functions results principally from their unique physicochemical and structural features, including high negative charge, high viscosity and lubrication propensities, unbranched polysaccharide structures, low compressibility as well as the ability to attract and imbibe large amounts of water [17].

Unlike proteins or nucleic acids, GAGs are constantly altered by processing enzymes and thus they vary greatly in molecular mass, disaccharide unit composition, and sulfation. Based on their core structure they are categorized into six different classes, viz. heparan sulfate (HS), heparin (HP), hyaluronic acid (HA), chondroitin sulfate (CS), dermatan sulfate (DS), and keratan sulfate (KS). The structural and functional diversities of GAGs are regulated by their sequence, size of the chains, degree of sulfation, and the ability to bind proteins [1,18-21].

This structural diversity of GAGs translates into highly heterogeneous functions and allows them to modulate interactions with various protein molecules in respective biological processes [4]. Most of these interactions are driven by electrostatics and are non-specific in nature, however, some of them are highly specific or selective [22-26].

The structural analysis of GAGs improves the understanding of their biological functions and helps in the development of structure–activity relationships for these important biopolymers [27,28]. Although the composition of the individual saccharide components of GAGs is simple, the structural analysis of GAGs is extremely difficult due to their complex pattern of modification such as epimerization and sulfation [29]. In addition, GAGs’ high flexibility and periodicity render these molecules profoundly challenging to analyze using experimental techniques only [30,31]. Thus, computational approaches could be efficiently used to gain insight into protein–GAG interactions that take place at single-molecule levels [32]. More than a complementary tool, computational approaches provide a better understanding of the role of individual interaction partners (including GAGs, solvent, and ions) by bringing often new and experimentally inaccessible details [33,34]. However, for computational researchers, there are still many challenges to overcome that originate from the physicochemical properties of GAGs, viz. their highly polarized (anionic) nature, their periodicity, and the complexity in decoding their sulfation pattern. Their charged nature necessitates the application of appropriate methods for electrostatics, ions, and solvent, particularly given their abundance in protein–GAG interfaces compared to complexes involving other classes of biomolecules. The periodicity can lead to multipose binding, wherein various configurations of the protein–GAG complex may exhibit similar free binding energies, allowing them to co-exist. Interpreting the “sulfation code”, the amount (net sulfation) and particular positions of the sulfation group (sulfation pattern), could assist in the explanation and prediction of GAG specificity [35]. Computational methodologies like molecular docking and molecular dynamics (MD) have proven to be successful in modelling protein–GAG interactions, particularly examining the fundamental questions related to these interactions such as their specificity, the multipose character of GAG binding and the polarity of the binding poses of these periodic molecules.

In the present work, all-atom MD simulations are conducted to study the dynamics of the protein–GAG complexes, and are complemented by free energy analysis. The free energy analysis of the protein–GAG interactions is important in understanding the nature of the interactions and the stability of the binding pose, including the scenario when several co-existing binding poses are identified. We analyze the binding interactions between the GAGs heparin, heparan sulfate, and chondroitin sulfate, and the proteins basic fibroblast growth factor (PDB ID: 1BFC, https://doi.org/10.2210/pdb1BFC/pdb, [12]), acidic fibroblast growth factor (PDB ID: 2AXM, https://doi.org/10.2210/pdb2AXM/pdb, [13]), and cathepsin K (PDB ID: 3C9E, https://doi.org/10.2210/pdb3C9E/pdb, [36]), and PDB ID: 4N8W, https://doi.org/10.2210/pdb4N8W/pdb, [37]). The third complex is known to exist in two different binding poses which are experimentally well established. In this study, the umbrella sampling (US) approach is used to pull away a GAG ligand from the binding site and then pull it back in. The main focus of our study is to evaluate whether the application of the US approach is able to reproduce experimentally obtained structures, and how useful it is for understanding GAG properties as protein recognition specificity and multipose binding. We also check for any trace of transition from the 3C9E to the 4N8W structure by pulling the ligand from its bound position and allowing the ligand to approach the protein from a very distant position to the binding sites.

Materials and Methods

Structures and parameters

Ligand preparation

GAG structures used in the study consist of two parts: 1. the part from the experimental structure (heparin in the 1BFC [12]
and 2AXM [13] complexes and chondroitin sulfate-4 in the case of 3C9E [36]/4N8W [37]), where the length is dp6 (dp stands for degree of polymerization) and 2, an additional part with different degree of sulfation or sulfation pattern (in case of ligands 1 and 2 for 1BFC and 2AXM dp6 desulfated heparan sulfate was added to the reducing end and non-reducing end of the GAG, respectively; in case of ligand 3 for the 1BFC and 2AXM dp6 desulfated heparan sulfate was added both to the reducing and non-reducing end of the GAG; in case of the ligand 4 for the 3C9E/4N8W complex dp6 chondroitin sulfate-6 was added to the reducing end of the GAG. The starting binding mode for the cathepsin K complex with chondroitin sulfate corresponded to the 3CE9 complex. Literature data for the sulfate groups [38] and GLYCAM06 [39] force field parameters were used for GAGs in the subsequent MD simulations. A \(^1\)C\(_4\) conformation for the IdoA2S ring was chosen as it was shown to be the essentially dominant conformation in the microsecond scale simulations performed by Sattelle et al. as it is energetically more favorable than the \(^2\)S\(_4\) conformation [40].

Complex preparation

The obtained ligands were docked using RS-REMD (replica exchange with repulsive scaling), an MD-based docking method [41], to assure proper binding poses of the whole ligand and ring puckering and to be consistent with further simulations. The docked ligands cover the binding site the same way as ligands in the experimental structures. Additionally, since the ligands used in the study are longer, they expand over the binding site and interact with the other parts of the protein as well. Experimental structures cover only a small part of the actual GAG molecule that interacts with the protein (as GAGs are built of tens to thousands of sugar units), therefore using longer ligands does not represent artificial behavior and may provide details of additional naturally occurring interactions. Comparison of the docked poses and PDB structures are presented in Supporting Information File 1, Figure S1.

MD simulations

All the MD simulations of the complexes obtained by RS-REMD docking were performed in AMBER20 package [42]. A TIP3P truncated octahedron water box with a distance of 20 Å from the solute to the box’s border was used to solvate complexes. Na\(^+\) counterions were used to neutralize the charge of the system. Energy minimization was performed preceding the production US runs (described in the next paragraph).

500 steepest descent cycles and \(10^3\) conjugate gradient cycles with 100 kcal/mol/Å\(^2\) harmonic force restraint on solute atoms were performed. It was followed by \(3 \times 10^3\) steepest descent cycles and \(3 \times 10^3\) conjugate gradient cycles without any restraints and continued with heating up the system to 300 K for 10 ps with harmonic force restraints of 100 kcal/mol/Å\(^2\) on solute atoms. Then, the system was equilibrated for each window at 300 K and \(10^5\) Pa in an isothermal, isobaric ensemble for 100 ps.

US production runs were performed for all of the complexes to pull away ligands from the binding site and then to bring them back to the binding site. US simulations consisted of 40 windows where in each the distance between ligand and the binding site was increased by 1 Å using harmonic restraints with a force constant of 10 kcal/mol/Å\(^2\). Each window consists of 100 ns of US simulation, therefore each US simulation is 4 μs. Distances between the following atoms were chosen as a reaction coordinate in the corresponding complexes: Ca@Leu225-O5@12IdoA(2S) (the GAG sequence numbering is according to the AMBER order, from reducing to non-reducing end and @ means that a particular atom belongs to a particular residue) for basic FGF-ligand 1; Cu@Leu225-O5@1GlcNS(6S) for basic FGF-ligand 2; Cu@Gly275-O5@6IdoA(2S) for basic FGF-ligand 3; Cu@Gly5-O5@12IdoA(2S) for acidic FGF-ligand 1; Cu@Gly5-O5@1GlcNS(6S) for acidic FGF-ligand 2; Cu@Gly5-O5@4IdoA(2S) for acidic FGF-ligand 3; Cu@Arg296-C3@12GlcA for cathepsin K-ligand 4. The reaction coordinate values increased in each subsequent window, with the starting point for each window taken from the previous one.

The overlap between the probability distributions in adjacent windows was analyzed both using bootstrap error analysis and visually for equilibration and production runs. WHAM (weighted histogram analysis method [43]) was performed using Grossfield’s WHAM program [44] to calculate the potential of mean force (PMF). For bootstrap analysis, 0.001 iteration, 300 K as temperature, and 1000 as number of Monte Carlo trials were used.

After completing the last window of US simulation, 500 ns unrestrained MD runs were carried out in the same isothermal isobaric ensemble to relax the system. A time step of 2 fs and a cut-off of 8 Å for electrostatics were used. The particle mesh Ewald method for treating electrostatics [45] and SHAKE algorithm for all the covalent bonds containing hydrogen atoms [46] were implemented in the MD simulations. The cpptraj program of AMBER was used for the analysis of the trajectories [47]. In particular, native contacts command with default parameters was used for the analysis of the contacts between protein and GAG molecules established in the course of the simulation.

Binding free energy calculations

MM/GBSA (molecular mechanics generalized born surface area) model igb = 2 [48] from AMBER20 was used for free
energy calculations on the trajectories obtained from RS-REMD simulations.

GAG binding pose accuracy evaluation
For the evaluation of the binding pose accuracy RMSD and RMSatd values were used. RMSD stands for root mean square deviation and it is defined as the average distance between the atoms of superimposed molecules. RMSatd (root mean square atom-type distance) is very similar to the widely used RMSD but instead of using specific atoms it compares atom types (e.g., any carbon atom to any carbon atom instead of specifically numbered carbon atom to the carbon atom with the same number). RMSatd is more appropriate when used for long and periodic molecules (such as GAGs), when a shift by one periodic unit yields the same pose but would result in high RMSD. The similar issue happens when GAG is rotated by 180°; although it occupies the same binding site and the pose is similar, the RMSD value would be expressed in tens of angstroms, while the RMSatd value would be significantly smaller.

Data analysis and its graphical representation were done with the R-package [49] and VMD [50].

Results
In total, 14 US simulations were performed to investigate the specificity of GAG–protein interactions, capabilities of US simulations to dissociate and reassociate protein–GAG complexes in these systems, and potential use of the US simulations in docking of GAG molecules to proteins. In order to do so, six different heparin systems (3 for basic FGF and 3 for acidic FGF) and one chondroitin sulfate system (with cathepsin K) were prepared. For each of the systems 2 US simulations were set up. First, hybrid GAGs (Figure 1) were prepared and docked using RS-REMD to find the pose in the binding site with the lowest interaction energy. Then, the GAG was pulled away from the binding site until it was shifted 40 Å from the starting position. Afterwards, the GAG was pulled in towards the binding site to observe if it reproduces a pose similar to the starting pose. To describe these unbinding and rebinding processes, analyses of RMSD, binding energy, contacts, and hydrogen bonds were performed. Additionally, after the final pulling step, a short MD run of 500 ns was performed to relax the system and to check if the final pose was energetically stable or if it changed during the relaxation step. The data depicted in the graphs result from the analysis of merged US trajectories. While this representation is not entirely physically sound, as the outcomes for each US window reflect the system’s state under particular conditions with explicitly defined reaction coordinate values, the visualization of these continuous data potentially offers a more comprehensive insight into the complexity of the system related to its dynamic behavior within each window.

Basic FGF
Ligand 1. The RMSD increased gradually up to values of around 40 Å during the unbinding process, and then decreased slowly when it was pulled in. After about the 20th window RMSD stabilized between 15 and 20 Å, suggesting that the GAG did not find the initial pose and was trapped in a different minimum (Figure 2). The same scenario was observed in terms of the binding energy (Supporting Information File 1, Figure S2). When the ligand was pulled away the energy increased and when it was pulled in the energy slowly decreased and converged after about 20 windows. The number of native contacts when the ligand was pulled away rapidly dropped from 1500 to 0 and remained 0 for the rest of the US run (Supporting Information File 1, Figure S3). When pulled in, between 20 and 30% native contacts are restored after the 25th window but not to the original level. Even the additional relaxation MD run did not restore any native contacts. This suggests that the GAG gets close to the binding site but does not return to a similar conformation as the initial (experimental) pose. A similar trend is observed with hydrogen bonds where the number of H-bonds drops when the ligand is pulled away but never gets fully restored after being pulled in to the initial pose. Visual analysis supports the observation that only a small part of the GAG chain from the final pose overlaps with its starting position. The final pose is perpendicular to the initial one (Figure 3).

Ligand 2. RMSD slowly increased when pulled away and then when pulled in it gradually decreased to between 6 and 8 Å. During the additional relaxation step, RMSD was further reduced to 3 to 4 Å suggesting that the GAG finds a pose similar to the starting one (Figure 2). The binding energy gradually increased when the ligand was pulled away (from around −150 kcal/mol to around −30 kcal/mol) (Supporting Information File 1, Figure S1). Then, when pulled in, the energy almost did not decrease at the beginning. Only after the 21st window the energy started to decrease more rapidly but it did not go back to the values of −150 kcal/mol corresponding to the initial pose and oscillated around −120 to −100 kcal/mol. During additional relaxation the energies decreased to the range from −130 to −100 kcal/mol. This shows that after an additional MD run, the binding pose did not only become closer to the original structure but also was stabilized energetically in comparison to the pre-relaxation step. The number of native contacts significantly dropped after the first part of US (from >2000 contacts to below ≈500) and then stabilized at around 200 to 300 contacts in the last windows (Supporting Information File 1, Figure S3). When the ligand was pulled back to the binding site, only some native contacts were restored (≈500), but during the subsequent relaxation the number of restored native contacts increased to more than 1000. In case of H-bonds, at the end of the US 70 to
90% of them were restored. Visually, both the final and the initial poses look very similar (Figure 3), and this is also reflected in very low RMSD values (3 to 4 Å for such long and flexible molecules is considered to reflect high structural similarity).

**Ligand 3.** Similar to the other ligands, when pulled away from the binding site the RMSD of ligand 3 gradually increased and during pulling in it slowly decreased but did not return to the initial pose which is represented in the RMSD value of 12 Å at the end of the US simulation. Additional relaxation MD also did not result in any significant decrease of RMSD (Figure 2). The initial binding energy of ~160 kcal/mol increased very fast at the start of pulling away and finished below ~30 kcal/mol at the end (Supporting Information File 1, Figure S2). During the pulling in of the GAG, the energy decreased slowly and reached ~70 to ~50 kcal/mol at the end of US. However, after the 37th window the binding energy drops below ~120 kcal/mol suggesting a more favorable novel ligand conformation. During relaxation, MD energies only improved slightly which is in
agreement with the high RMSD that suggests that GAG did not return to the initial binding pose. The number of native contacts decreased drastically during the first windows of US from 1800 to 0 in the 13th window (Supporting Information File 1, Figure S2). During pulling in ligand towards the binding site only a small percentage of the native contacts were restored (50 to 150 native contacts in the last windows). After the relaxation MD the number of contacts went up to 200 to 250, but it never reached levels close to the initial ones which also suggests that the GAG did not get close to the binding site. The number of H-bonds at the end of pulling in was similar to the start of pulling away. However, none of the H-bonds at the end of the US simulation were established between the same atoms as at the start. Visually, only a part of the GAG’s final pose overlaps with the initial one. The final pose adapts a perpendicular conformation to the starting GAG chain orientation (Figure 3).

Acidic FGF

Ligand 1. RMSD slowly increased during the first phase (windows from 1 to 8) of pulling away and afterwards with the pace similar to the other systems analyzed in this work. On the way back, RMSD of the ligand steadily decreased reaching values 4 to 5 Å at the end of the pulling in (Supporting Information File 1, Figure S4). During the relaxation step, RMSD remained around the same level and did not decrease further. Binding energy started at $-140$ kcal/mol and increased fast during the first 30 windows (Supporting Information File 1, Figure S5). Afterwards, it oscillated between $-20$ and 0 kcal/mol. During pulling in of the ligand the energy did not change before window 25 when it started to decrease reaching $-90$ kcal/mol at the last window. During the relaxation, the energy remained at a similar level. Interestingly, despite the low RMSD at the end, the final energy is less favorable ($-90$ kcal/mol) than the one observed at the beginning of the US
Figure 3: Graphical representation of the ligands’ starting (in red, licorice) and final (in blue, licorice) positions in regard of the binding site of basic fibroblast factor (in yellow, new cartoon).

Figure 4: Graphical representation of the ligand’s starting (in red, licorice) and final (in blue, licorice) poses in regard of the binding site of acidic fibroblast factor (in yellow, new cartoon).

(−140 kcal/mol). The number of native contacts dropped to zero around the 15th window and remained 0 for the rest of the pulling away (Supporting Information File 1, Figure S6). During pulling in, no restoration of native contacts was observed. Also the number of H-bonds when the ligand was pulled all the way in was slightly lower (70 to 80%) than before it was pulled away. The number of H-bonds and native contacts suggest an overall smaller amount of interactions between the ligand and the receptor, but also the establishment of non-native contacts. Visually, the poses from the start and at the end of the US simulations look very similar (Figure 4), with major differences observed around the part of the GAG that is not bound to the protein.

**Ligand 2.** RMSD increased slowly until the 7th window where it started to increase more rapidly. During pulling in, RMSD did not decrease significantly (although visually the ligand is getting close to the initial binding pose) suggesting a drastically different pose of the ligand (Supporting Information File 1, Figure S4). Additional relaxation also did not improve RMSD significantly. In terms of energy of the system it started around −140 kcal/mol and it dropped to the level between −20 and 0 kcal/mol after the 27th window (Supporting Information File 1, Figure S5). During pulling in the energy did not improve significantly. The number of native contacts dropped from ≈1000 to 0 after the 15th window of pulling away (Supporting Information File 1, Figure S6). Only very few native contacts were restored during pulling in at maximum showing 200 of them. The number of H-bonds during pulling in was slightly lower than during pulling away suggesting less interactions between the ligand and the receptor on the way back than at the start of the US simulation. Visually, the major part of the GAG at the end of the US simulation overlapped with its starting pose. However, the final structure is more bent and shifted by
about 3 rings relative to the initial one (Figure 4). This is also confirmed by relatively high RMSD values that did not improve much during the course of the pulling in US.

**Ligand 3.** RMSD increased slowly during the first few windows but unlike the other ligands in this particular case the scenario for RMSD did not change significantly afterwards. During pulling in the ligand back to the binding site only a low RMSD decrease was observed (Supporting Information File 1, Figure S4). During the relaxation, again, only a minor decrease in RMSD was observed suggesting that a slightly more favorable pose was achieved. In terms of the energy evolution during pulling away, it started around $-140$ kcal/mol and then it increased up to the 30th window where it stabilized below $0$ kcal/mol (Supporting Information File 1, Figure S5). On the way back, we observe only a partial improvement of the binding energy as it reached values from $-80$ to $-70$ kcal/mol at the end of pulling in. However, during relaxation the energy lowered to values from $-130$ to $-120$ kcal/mol suggesting binding almost as strong as at the start of the US. The relatively high RMSD and low energy can be justified by the fact that the obtained pose of the ligand was very different from the initial one but there is a small overlapping part that interacts with the ligand around the binding site which can serve as basis for this strong binding. The number of native contacts at the beginning was 1300 and decreased slowly (Supporting Information File 1, Figure S6). In the second part of pulling ligand away changes in number of native contacts were sudden and drastic but they never went completely to 0. The number of contacts oscillated between 50 and 500. On the way back of the ligand, changes are much more subtle and the number of contacts remained between 200 and 400. During the relaxation no significant changes in the number of native contacts was observed. More H-bonds were present at the beginning of pulling the ligand away than at the end of the pulling in suggesting more interactions between the ligand and the protein at the start than at the end of the US. Visually, the final pose of the GAG is much different from the initial one but there is a small overlapping part that interacts with the GAG maintained a relatively strong binding to the protein. This is particularly true for ligand 3 of acidic fibroblast growth factor, which when pulled back into the binding site led to low binding energies but a drastically different pose (partially perpendicular) of the ligand.

Energy contributions of sulfated and unsulfated parts of the GAG were investigated from per residue decomposition of MM/GBSA analysis (Table 2). In every case, sulfated parts were always contributing more to the receptor binding than unsulfated ones. Usually, the sulfated part contributed 3–5 times stronger than the unsulfated part. However, during pulling in of ligand 3 for basic fibroblast growth factor the sulfated part contributed significantly ($-7.6$ kcal/mol for the sulfated part in comparison to $-10$ kcal/mol for the sulfated part, respectively). More interestingly in this case during the pulling in process the contribution of the sulfated part decreased while the one of the unsulfated part increased. This could be interpreted as that the binding of the unsulfated residues can partially compensate the energy loss due to unbinding of the sulfated residues, suggesting rather non-specific interactions between the protein and the ligand.

**Cathepsin K**

During the pulling away of the GAG RMSD slowly and steadily increased. During pulling in RMSD only lowered slightly reaching $20$ Å which suggests that at the end of US the GAG did not return to a pose similar to the starting one. Relaxation MD neither improved the final conformation. The energy of the system increased from $-120$ kcal/mol to values between $-50$

---

**Table 1:** Pearson correlation coefficients between energies obtained from MM/GBSA analysis and RMSD values of the ligand for all frames of the merged MD trajectories.

<table>
<thead>
<tr>
<th></th>
<th>basic fibroblast growth factor (1BFC)</th>
<th>acidic fibroblast growth factor (2AXM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ligand 1 (away)</td>
<td>0.77</td>
<td>0.81</td>
</tr>
<tr>
<td>ligand 1 (in)</td>
<td>0.60</td>
<td>0.78</td>
</tr>
<tr>
<td>ligand 2 (away)</td>
<td>0.84</td>
<td>0.63</td>
</tr>
<tr>
<td>ligand 2 (in)</td>
<td>0.90</td>
<td>0.35</td>
</tr>
<tr>
<td>ligand 3 (away)</td>
<td>0.80</td>
<td>0.50</td>
</tr>
<tr>
<td>ligand 3 (in)</td>
<td>0.58</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Table 2: Energy contributions in kcal/mol of the sulfated and unsulfated parts of GAGs obtained from MM/GBSA per residue decomposition.

<table>
<thead>
<tr>
<th>Ligand 1 (away)</th>
<th>Ligand 1 (in)</th>
<th>Ligand 2 (away)</th>
<th>Ligand 2 (in)</th>
<th>Ligand 3 (away)</th>
<th>Ligand 3 (in)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sulfated</td>
<td>−13.1</td>
<td>Sulfated</td>
<td>−12.8</td>
<td>Sulfated</td>
<td>−16.1</td>
</tr>
<tr>
<td>Not sulfated</td>
<td>−4.7</td>
<td>Not sulfated</td>
<td>−4.6</td>
<td>Not sulfated</td>
<td>−8.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ligand 2 (away)</td>
<td>−11.9</td>
<td>Ligand 2 (in)</td>
<td>−12.1</td>
<td>Ligand 3 (away)</td>
<td>−4.7</td>
</tr>
<tr>
<td>Not sulfated</td>
<td>−6.1</td>
<td>Not sulfated</td>
<td>−4.7</td>
<td>Not sulfated</td>
<td>−10.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−7.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Acidic fibroblast growth factor (2AXM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ligand 1 (away)</td>
</tr>
<tr>
<td>Sulfated</td>
</tr>
<tr>
<td>Not sulfated</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

and −40 kcal/mol around the 22nd window and then remained at this level to the end of pulling away. On the way back of the GAG to the binding site the energy slowly decreased and reached −80 kcal/mol at the end of pulling in. During the relaxation MD the energy decreased further to −110 kcal/mol which is almost the same value as observed at the starting point suggesting that this significantly different pose is almost as stable as the initial one. The number of native contacts lowers from ≈1500 to 0 after the 25th window of the US simulation. During pulling in some native contacts are being restored but the number greatly varies and never surpassed 500 contacts. The number of H-bonds during pulling in are also lower than in the initial pose. Visually, the final pose is significantly different than the starting one (Figure 5 and Figure 6). In the binding site the part with the 4-sulfation of the final GAG conformation is perpendicular to the starting one, and the part of GAG with the 6-sulfation is close to the second GAG binding site of the cathepsin K. The final pose of the GAG partially overlapped with both experimentally known binding sites. This is most likely the reason why the energy at the start and at the end of US is similar to the one of the initial pose despite the fact that much a smaller part of the GAG is located at the first binding site. Hence a comparison of the binding position of the ligand with both crystal structures (3C9E and 4N8W) were carried out to reveal which binding site is preferred upon the reassociation of the ligand (Figure 6). It can be seen that the binding of the ligand to the protein at the end of the sliding in represents a combination of both the binding positions from the crystal structures and the RMSatd score obtained for the two different crystal structures are 4.1 Å and 8.3 Å for 3C9E and 4N8W complexes, respectively. The dodecamer ligand docked to the protein in such a way that the hexameric part with the 4-sulfation (as observed in the crystal structure) occupied the 4N8W site and the hexameric part with the 6-sulfation bound to the site observed in the 3C9E structure. The comparison of the final structure and that obtained after the docking yielded RMSatd of 10.2 Å, which shows that the pulling back results in the structure more similar to that of the crystal structure than to the initial docked pose. A similar comparison of the final structure obtained at the end of the sliding in process was done for the other complexes with their corresponding crystal structures and the one obtained after docking. The ligands’ RMSatd values (Tables S1 and S2 in Supporting Information File 1) show that in the majority of the complexes the final structure is more similar to the crystal structure than to the initial docked structure. However, the goal of the study was to compare the final structures to the starting positions rather than to the experimental.

Figure 5: Graphical representation of the ligand 4 starting (in red, licorice) and final (in blue, licorice) position in regard of binding site of cathepsin K (in yellow, new cartoon).
Investigation of the protein–GAG recognition in the proximity of the native binding pose

Next, we analyzed if the US approach is able to reproduce the native binding pose when pulling away a ligand by just a disaccharide unit and returning it back to the binding site. These simulations involved approximately a 10 Å shift from the native complex of ligand 1 from the basic FGF and allowed to investigate the near-native free energy landscape and the respective atomic details of the protein–GAG recognition. In the forced dissociation process, the RMSD curve looks similar to the ones from the previously described, longer pulling away trajectories: the RMSD values increase gradually as the ligand pose gets closer to the native pose within the shift of a monomeric unit (first part of the pulling away step, 0.5 μs), yielding a rugged shape of the curve (Figure 7). Interestingly, on the way back, the RMSD values reach minimal values at windows 5, 7, and 8, corresponding to the reaction coordinate values of 5 Å, 3 Å, and 2 Å, respectively (as well as corresponding to the 0.5, 0.7 and 0.8 μs of US, respectively), but then go up at the end of the pulling in process.

This is also reflected by the MM/GBSA binding energy analysis (Supporting Information File 1, Figure S7), where the energy gradually increased in the process of dissociation with...
the exception of the stabilized conformation at window 5 (0.5 μs), where the binding strength is energetically comparable with the one of the native binding pose. The MM/GBSA free energy landscape is very rugged on the way of pulling the ligand in. However, the most favorable energies of very comparable values are observed in windows 6 and 8 (0.6 and 0.8 μs, respectively) suggesting that while for window 8 the proximity to the native pose was energetically favorable, the interaction free energy minimum in window 6 corresponds to a distinct non-native binding pose. This suggests a high heterogeneity of the free energy landscape in the proximity of the native binding pose and a high propensity for multipose binding in the system. The number of native contacts and total H-bonds gradually decrease in the dissociation process (Figure 7), while there is a clear peak of the non-native contacts number and additional H-bonds at window 5 suggesting a partial stabilization of the binding by H-bonds, also in agreement with the MM/GBSA binding energy trend. On the way back, at window 8, the ligand establishes most of the native contacts which also correspond to the increase of the number of H-bonds established. This points out that several stabilized and energetically comparable binding poses co-exist in the system. This is also supported by the absence of significant correlations between the MM/GBSA energy and RMSD to the initial pose (Supporting Information File 1, Figure S8).

Another way to calculate the free energy landscape with less details but in a manner more sound for the applied protocol, is to estimate PMF along the reaction coordinate. Such calculations also support the conclusion that the applied protocol did not succeed in returning the system to the original binding pose (Supporting Information File 1, Figures S9–S11). In comparison to the data from the MM-GBSA approach, the PMF data show even larger differences between the starting and the final poses obtained in the US trajectories.

In turn, results of the analysis of pairwise correlations between the number of established H-bonds, MM/GBSA free energy, native, non-native and total number of contacts differ for pulling out and pulling in processes (Table 3). For the dissociation, as expected the number of native contacts and H-bonds correlate very well with the MM/GBSA energies (Pearson correlation coefficients obtained for all frames of the trajectories are 0.81 and 0.76, respectively), while on the way back, the still high correlation with the H-bond number (0.58) and an essentially decreased one with the native contact number (0.36) mean that hydrogen binding dominates the binding energetics of the system and is an origin of multipose binding.

When correlating the values for MM/GBSA and the number of H-bonds averaged per each US window, the correlation coefficients for the pulling away and pulling back processes are 0.97 and 0.56, respectively. Despite these significant differences in the correlations, implying a more complex free energy landscape topology when the ligand is pulled in, the energies per H-bond calculated from the linear regression model are very similar: $-10.1 \pm 0.6$ kcal/mol and $-10.5 \pm 1.0$ kcal/mol for pulling away and in, respectively. These differences in the correlations, however, can be partially attributed to the arbitrary choice of the US reaction coordinate which can affect the pulling away and pulling in processes and, therefore, the data described here.

Furthermore, we analyzed in detail the most representative H-bonds (with the occupancy higher than 20%) established at different US windows that were the most distinguishable in the pulling away and pulling in processes. In particular, we analyzed the X-ray conformation (PDB ID: 1BFC) based MD trajectory and windows 5 and 10 and windows 6, 8, 10 in the forward and reverse processes, respectively (Figure 7). Interestingly, in all these windows with more favorable binding energies, particular three positively charged residues, K101, K106, and K116, maintained strong H-bonds that have been also described here.

Table 3: Pearson correlation coefficients obtained for all frames of the pulling away and pulling in MD trajectories between the protein–GAG recognition parameters. Native: native contacts; non-native: non-native contacts; total: the sum of native and non-native contacts.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Pearson correlation coefficients obtained for all frames of the pulling away and pulling in MD trajectories between the protein–GAG recognition parameters. Native: native contacts; non-native: non-native contacts; total: the sum of native and non-native contacts.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pulling away</td>
<td>Pulling in</td>
</tr>
<tr>
<td>$N_{native}$</td>
<td>$N_{non-native}$</td>
</tr>
<tr>
<td>$N_{native}$</td>
<td>–</td>
</tr>
<tr>
<td>$N_{non-native}$</td>
<td>–</td>
</tr>
<tr>
<td>$N_{total}$</td>
<td>–</td>
</tr>
<tr>
<td>$N_{H\text{-}bonds}$</td>
<td>0.65</td>
</tr>
<tr>
<td>$\Delta G_{\text{MM/GBSA}}$</td>
<td>0.81</td>
</tr>
</tbody>
</table>
erved in the unrestrained MD simulation of the X-ray structure is that there were several non-charged residues (N8, A17, Y84) among the top residues contributing to H-bonds, while almost exclusively positively charged residues were observed to be substantial H-bond contributors in the US windows. In a microsecond-scale MD simulation of the same X-ray structure, three non-charged residues were identified as the top MM/GBSA free energy contributors [52]. This suggests that despite a very complex free energy landscape in the proximity of the native pose, the native pose can be potentially distinguished by the essential contributions of the non-charged residues to the GAG recognition. Further, this implies a certain degree of specificity and not simply electrostatics-driven interactions in this particular molecular complex. Estimation of the free energy barriers in the completed analysis suggests that the sliding of a long GAG on the protein surface is a feasible process that could underline the natural recognition of the specific GAG patterns by a protein target.

Conclusion

In this study, the US approach was used to pull away a GAG ligand from the binding site and then to pull it back in to the binding site. The goal was to analyze if US is able to reproduce experimentally obtained structures, and if it can contribute to a deeper understanding of GAG properties as protein recognition specificity and multipose binding. Although the US is a powerful method it was shown not to be able to accurately reproduce experimental structures or the most energetically favorable binding poses in the majority of the investigated systems with the particular protocols we applied in this study. The limitations in our study can be attributed to two main factors. Firstly, the relatively short sampling times (100 ns) may have been insufficient to adequately equilibrate the systems, especially given their complex free energy landscape in each US window. Secondly, the selected reaction coordinates for pulling away and pulling in may not inherently suggest unique reversible pathways. To improve the convergence and sampling of the free energy landscape, advanced sampling protocols can be employed or the US simulations can be repeated multiple times. Therefore, the data obtained in this study and the conclusions related to these data are rather qualitative. In the next steps, we plan to apply more advanced sampling protocols. However, even when using the described protocol in some of the systems it was able to bring the ligand back to the binding site (in two cases with comparable accuracy to one of the most powerful GAG docking tools (RS-REMD), which corresponds to RMSD values <4 Å). Additionally, it allowed to observe multipose binding phenomena manifesting other energetically favorable binding poses of the GAG in the binding site. In these cases, although the RMSD values with reference to the experimental structures were high as only a very small part of the final GAG binding pose overlapped with the initial structure, binding energies remained almost at the same level as the ones corresponding to the experimental binding poses. Regarding the specificity, in most cases a partial overlap between the GAG parts in the experimental and the pulled in structures corresponding to the same sulfation pattern/amount was observed. Nevertheless, in one of the simulations of the basic fibroblast growth factor system a less sulfated part contributed comparably to the sulfated one suggesting a potential of non-purely electrostatics dominance in the protein–GAG interactions. The more detailed analysis of the GAG recognition in this system in near-native states points out to the complexity of free energy landscape but at the same identifies the key charged H-bonding contributors to the GAG binding that together with several non-charged residues in the binding interface potentially determine the specificity of the interactions in this complex. The analysis of free energy landscapes in the studied systems suggests that sliding of a GAG along a binding site in a protein target could occur naturally and, therefore, could be a way for a protein to effectively sample different particular GAG recognition patterns.

The findings in this work should contribute to the broadening of the knowledge regarding the specificity of protein–GAG interactions and the limitations of the computational tools employed to analyze them.

Supporting Information

Supporting Information File 1
Additional information and graphical representations.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-19-144-S1.pdf]

Funding

This research was funded by the National Science Centre of Poland (grant numbers UMO-2018/30/E/ST4/00037 and UMO-2018/31/G/ST4/00246).

ORCID® iDs

Sergey A. Samsonov - https://orcid.org/0000-0002-5166-4849

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.3762/bxiv.2023.46.v1

References

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/bjoc/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/bjoc.19.144
Synthesis of the 3’-O-sulfated TF antigen with a TEG-N\textsubscript{3} linker for glycodendrimersomes preparation to study lectin binding

Mark Reihill, Hanyue Ma, Dennis Bengtsson and Stefan Oscarson*  

Abstract
The synthesis of gram quantities of the TF antigen (β-ᴅ-Gal-(1→3)-α-ᴅ-GalNAc) and its 3’-sulfated analogue with a TEG-N\textsubscript{3} spacer attached is described. The synthesis of the TF antigen comprises seven steps, from a known N-Troc-protected galactosamine donor, with an overall yield of 31%. Both the spacer (85%) and the galactose moiety (79%) were introduced using thioglycoside donors in NIS/AgOTf-promoted glycosylation reactions. The 3’-sulfate was finally introduced through tin activation in benzene/DMF followed by treatment with a sulfur trioxide–trimethylamine complex in a 66% yield.

Introduction
In a collaboration project with groups from Universities in Munich and Pennsylvania we are investigating carbohydrate–lectin interactions using programmable glycodendrimersomes based on synthetic glycans. We have earlier synthesized 2-[2-(2-azidoethoxy)ethoxy]ethyl (TEG-N\textsubscript{3}) glycosides of lactose, 3’-Su-lactose and LacdiNAc (β-ᴅ-GalNAc-(1→4)-β-ᴅ-GlcNAc), which have then been used for production of the glycodendrimersomes and interaction studies with various galectins [1,2]. In the continuation of this collaboration, to investigate the binding of siglec-1 and the chimera of 3’-SuTF-binding siglec and TF-binding galectin-3, TEG-N\textsubscript{3} glycosides of the TF antigen (β-ᴅ-Gal-(1→3)-α-ᴅ-GalNAc, 1) and its 3’-O-sulfated analogue (2, Figure 1) were required on a gram scale to allow efficient synthesis of the glycodendrimersomes. The TF antigen is presented on the surface of most human cancer cell types and its interaction with galectins 1 and 3 leads to tumour cell aggregation and promotes cancer metastasis [3-5]. The 3’-O-sulfated analogue is known to bind to siglec-1, 4, and 8 [6] as well as galectin 4 [7,8], but its biological role is not that well investigated.

Compound 2 is a new compound but two syntheses of compound 1 have recently been reported, one using an enzymatic approach and a commercial α-TEG-N\textsubscript{3} GalNAc acceptor [9].
and one using glycosyl bromide donors and silver salt-promoted glycosylations [10].

Results and Discussion

To introduce the 2-[2-(2-chloroethoxy)ethoxy]ethyl (TEG-Cl) spacer both a Fischer synthesis starting from unprotected N-acetylgalactosamine and a Lewis acid-promoted reaction starting from per-acetylated galactosamine were initially tested. As reported [11], the Fischer synthesis gives low yields and α-selectivity. The Lewis acid-promoted reaction, which had worked well to produce β-linked TEG-spacer glycosides with per-acetylated lactose and 2-phthalimidoglucosamine [1,2] worked well with 2-chloroethanol as a spacer (68%, pure α) but failed with the TEG-Cl spacer [12], why we instead decided to use a thioglycoside donor to introduce the spacer. To ensure α-selectivity a di-tert-butylsilyl-4,6-acetal-protected donor, as developed by the Kiso group [13,14], was chosen. After some initial testing the known N-Troc-protected donor 3 [15,16] (Scheme 1) was selected [17].

Since donor 3 possessed a Troc group, which contains 3 chlorine atoms, nucleophilic introduction of an azido group at this stage was predicted to be problematic. Therefore, the azido functionality was installed in the spacer before the glycosylation. Donor 3 underwent an NIS/AgOTf-promoted glycosylation with the TEG-\(\text{N}_3\) acceptor [18], furnishing α-linked 4 in an 85% yield (Scheme 1). H-1 appeared as a doublet at 4.95 ppm with a \(J\) value of 3.6 Hz in the \(^1\text{H}\) NMR spectrum proving the anomeric α-configuration. The presence of Troc-rotamers was also apparent, with a ratio of 19:6 being observed by \(^1\text{H}\) NMR in CDCl\(_3\) at 25 °C. Catalytic amounts of NaOMe (0.005 M) in MeOH were used to remove the acetate from compound 4, taking care not to affect the Troc group, to afford acceptor 5 in a 96% yield.

---

**Figure 1:** Structure of target compounds 1 and 2.

**Scheme 1:** Synthesis of target compounds 1 and 2. Key: a) NIS, AgOTf (20 mol %), 4 Å molecular sieves, CH\(_2\text{Cl}_2\), rt, 40 min, 85%; b) NaOMe (10 mol %), MeOH, rt, 4 h, 96%; c) NIS, AgOTf (19 mol %), AW-300 4 Å molecular sieves, CH\(_2\text{Cl}_2\), rt, 1 h, 79%; d) 1 M Bu\(_4\text{NF/THF}, \text{THF, rt, 2 h, then HF-Py, rt, 3 h; e) Ac}_2\text{O/Py (1:2, v/v), rt, 16 h, 53% over 3 steps; f) 1 M NaOMe/MeOH, MeOH, pH 10, rt, 1 h, 90%; g) Bu\(_3\text{SnO, benzene/DMF (5:1, v/v), 125 °C, 24 h, then SO}_2\text{-NMe}_3, \text{DMF, rt, 72 h, then flash chromatography, then Dowex® 50WX4 (Na\textsuperscript+ form) resin, H}_2\text{O, rt, 16 h, 66%}.}$$
Earlier optimizations of the introduction of the β-linked galactose moiety using 2-azidoethyl 2-acetamido-4,6-O-benzylidene-2-deoxy-α-D-galactopyranoside as acceptor showed an acetylated thioglycoside donor to be the best choice [12], surprisingly better than a benzoylated donor [19], why this donor was the first one tested also with the quite different acceptor 5. An NIS/AgOTf-promoted glycosylation with donor 6 [20] yielded 79% of disaccharide 7. Due to the presence of rotamers, NMR spectra of 7 proved to be difficult to analyse when data were recorded in CDCl3. Changing the NMR solvent to CD3OD greatly reduced the complexity of the spectra [21-23].

Since 7 possessed an azido group as part of the linker, removal of the Troc group under reductive conditions was ruled out due to probable chemoselectivity issues [24,25]. Interestingly, Jacquemard et al. outlined a useful, mild method for removing a range of carbamates using Bu4NF in an article from 2004 [26]. As 7 contained a DTBS group, the possibility of removing both Troc and DTBS groups in a one-pot procedure was tested. Disaccharide 7 was therefore treated with 1 M Bu4NF/THF and after 2 hours, full consumption of the starting material was observed by TLC. However, MALDI-TOF mass spectrometry (super-DHB matrix) revealed that only the Troc group had been removed, with the DTBS substituent proving to be stable under these conditions. Addition of a large excess of HF-Py (40 equiv) proved to be necessary to remove the bulky silyl group. After concentration, the crude product was acetylated (Ac2O/Py, 1:2, v/v), furnishing per-acetylated compound 8 in a 53% yield over one-gram scale. Comparing the C HSQC spectra of compounds 1 and 2, there is a clear downfield shift of the H-3'/C-3' signal from 1 to sulfated 2 (Figure 2). This showed that regioselective 3'-O-sulfation had been achieved, with HRMS also indicating that only one sulfate group was present.

TEG-N3 lactose compound, tin-activation was performed with Bu2SnO in refluxing MeOH, followed by stirring with SO3·NMe3 in 1,4-dioxane to afford the 3'-O-sulfate in 65% yield [1]. Here, however, this choice of solvent in the sulfation step led to the material being insoluble and no reaction was observable by TLC. Changing the solvent of the sulfation reaction to DMF resulted in formation of a homogenous solution, but still no conversion to the sulfated product, even when the temperature was raised to 80 °C [28,29]. Switching the sulfating reagent to SO3·Py or performing the reaction at 150 °C in a microwave did not improve the outcome [30,31].

Since there was no observable sulfation taking place, the tin-activation step was suspected to be the root of the problem. To rectify this, similar to Malleron et al., 1 was refluxed, in a Dean–Stark set-up, with Bu2SnO in benzene/DMF (5:1, v/v) [32]. The solvent in the receiver was drained after 24 hours and the benzene was removed from the reaction mixture in vacuo. Sulfation was then performed through addition of SO3·NMe3 to the DMF solution. Consumption of 1 was observed by TLC after 72 hours and stirring with Dowex® 50WX4 (Na+ form) resin resulted in formation of target 2. Purification by flash chromatography, however, led to isolation of a mixture of 2 and a tin-related impurity (n-butylic chain evident by NMR). Acetylation of this material followed by flash chromatography proved ineffective in removing the unwanted entity. To overcome this problem, flash chromatography was performed before stirring with the ion-exchange resin, with no apparent presence of tin impurities by NMR when the sequence was executed in this order and sulfated target 2 was obtained in a 66% yield on a one-gram scale. Comparing the 1H,13C HSQC spectra of compounds 1 and 2, there is a clear downfield shift of the H-3'/C-3' signal from 1 to sulfated 2 (Figure 2). This showed that regiolective 3'-O-sulfation had been achieved, with HRMS also indicating that only one sulfate group was present.

| Table 1: Summary of conditions attempted to achieve regioselective 3'-O-sulfation. |
|---|---|---|---|---|---|---|---|
| **Solvent(s)** | **Temperature/conditions** | **Set-up** | **Sulfating reagent** | **Solvent** | **Temperature/conditions** | **Result/ yield** |
| MeOH | 95 °C | reflux | SO3·NMe3 | 1,4-dioxane | rt | material insoluble, no reaction |
| MeOH | 95 °C | reflux | SO3·NMe3 | DMF | rt–80 °C | no reaction |
| MeOH | 95 °C | reflux | SO3·Py | DMF | 80 °C | no reaction |
| MeOH | 95 °C | reflux, Dean–Stark | SO3·NMe3 | DMF | 150 °C, microwave | 66% |
| benzene/DMF (5:1, v/v) | 125 °C | reflux, Dean–Stark | SO3·NMe3 | DMF | rt | no reaction |

*aTin-activation was performed with 1.2 equiv of Bu2SnO in all cases for 16–24 h and sulfation reactions proceeded for 24–72 h.*
Conclusion

An efficient synthesis of the important TF and 3'-Su-TF antigens equipped with a TEG-N₃ linker to allow formation of various conjugates has been developed for further interaction studies with lectins (galectins and siglecs). The synthesis of the 3'-Su-TF antigen 2 comprises eight steps from the known N-galactosamine donor 3, where two of the steps, removal of the Troc- and DTBS protecting groups are performed in the same pot and the following acetylation without purification of the intermediate, why the synthesis is high-yielding (20% overall yield) and easily scalable (9 g of protected disaccharide 7 and 1 gram of target 2 were synthesized).

Experimental

General methods

All reactions containing air- and moisture-sensitive reagents were carried out under an inert atmosphere of nitrogen in oven-dried glassware with magnetic stirring. N₂-flushed plastic syringes were used to transfer air- and moisture-sensitive reagents. All reactions were monitored by thin-layer chromatography (TLC) on Merck® DC-Alufolien plates precoated with silica gel 60 F254. Visualisation was performed with UV-light (254 nm) fluorescence quenching, and/or by staining with an 8% H₂SO₄ dip (stock solution: 8 mL conc. H₂SO₄, 92 mL EtOH), ninhydrin dip (stock solution: 0.3 g ninhydrin, 3 mL AcOH, 100 mL EtOH) and/or ceric ammonium molybdate dip (stock solution: 25 g ammonium molybdate tetrahydrate, 0.5 g Ce(SO₄)₂, 50 mL H₂SO₄, 450 mL EtOH).

Chromatography

Silica gel flash chromatography was carried out using Davisil® LC60A (40–63 μm) silica gel or with automated flash chromatography systems, Buchi Reveleris® X2 (UV 200–500 nm and ELSD detection, Reveleris® silica cartiges 40 μm, Büchi Labortecnik AG®) and Biotage® SP4 HPFC (UV 200–500 nm, Biotage® SNAP KP-Sil 50 μm irregular silica, Biotage® AB).

Instrumentation

¹H NMR and ¹³C NMR spectra were recorded on Varian Inova spectrometers at 25 °C in chloroform-d (CDCl₃), methanol-d₄ (CD₃OD), deuterium oxide (D₂O) or DMSO-d₆ ((CD₃)₂SO). ¹H NMR spectra were standardised against the residual solvent peak (CDCl₃, δ = 7.26 ppm; CD₃OD, δ = 3.31 ppm; D₂O, δ = 4.79 ppm; (CD₃)₂SO δ = 2.50 ppm); or internal trimethylsilane, δ = 0.00 ppm. ¹³C NMR spectra were standardised against the residual solvent peak (CDCl₃, δ = 77.16 ppm; CD₃OD, δ = 49.00 ppm; (CD₃)₂SO δ = 39.52 ppm and ¹³C NMR spectra recorded in D₂O are unreferenced. All ¹³C NMR spectra are ¹H decoupled. All NMR data are represented as follows: chemical shift (δ ppm), multiplicity (s = singlet, br s = broad singlet, d = doublet, app d = apparent doublet, t = triplet, q = quartet, dd = doublet of doublets, dt = doublet of triplets, m = multiplet), coupling constant in Hz, integration. Assignments were aided by homonuclear ¹H,¹H (COSY, TOCSY) and ¹H,¹³C heteronuclear (HSQC, HMBC) two-dimensional correlation spectroscopies. ¹³C chem-
2-[(2-Azidoethoxy)ethoxy]ethyl 3-O-acetyl-2-deoxy-4,6-O-di-tert-butylsilylene-2-(2′′-2′′′-trichloroethoxycarbonyl-aminoo)-α-D-galactopyranoside (4): Donor 3 [9,10] (9.3 g, 15 mmol) and the TEG-N3 acceptor (synthesized as described in reference [12], but also commercially available, 3 g, 22 mmol) were placed under N2 together and dissolved in dry CH2Cl2 (300 mL). 4 Å molecular sieves (10.2 g) were added and the resulting suspension was stirred at room temperature for 16 hours. NIS (6.66 g, 29.6 mmol) and AgOTf (760 mg, 2.96 mmol) were then added, and the reaction was stirred at room temperature for 40 minutes. The reaction was then quenched with Et3N, filtered through Celite® and concentrated in vacuo. The resulting residue was taken up in EtOAc (700 mL) and washed with 10% aq Na2SO3 (700 mL) water (700 mL) and brine (700 mL). The organic phase was then dried over MgSO4, filtered and reduced to dryness. Flash chromatography on silica gel (toluene → toluene/EtOAc, 3:2) yielded 4 as an orange syrup (8.74 g, 85%). Rf = 0.4 (toluene/EtOAc, 3:2); [α]D +92 (c 1.0, CHCl3); 1H NMR (500 MHz, CDCl3) δ 5.44 (d, J = 10.2 Hz, 1H, NH), 4.98 (dd, J = 11.1, 2.9 Hz, 1H, H-3), 4.95 (d, J = 3.6 Hz, 1H, H-1), 4.87 (d, J = 12.1 Hz, 1H, CH2(AcTroc), 4.69–4.58 (m, 2H, H-4, CH2(B)Troc), 4.49 (td, J = 10.6, 3.6 Hz, 1H, H-2), 4.26 (dd, J = 12.6, 2.2 Hz, 1H, H-6(A)), 4.15 (dd, J = 12.5, 1.7 Hz, 1H, H-6(B)), 3.90–3.77 (m, 2H, H-5, CH2(A/Linker)), 3.76–3.59 (m, 9H, CH2(B)Linker, 4 × CH2(Linker)), 3.39 (t, J = 5.1 Hz, 2H, CH2(Linker)), 2.07 (s, 3H, CH3(OAc)), 1.08 (s, 9H, C(CH3)3(DTBS)), 1.02 (s, 9H, C(CH3)3(DTBS)); 13C NMR (126 MHz, CDCl3) δ 171.3 (C=O(OAc)), 154.6 (C=O(Troc)), 98.5 (C-1), 95.8 (C3(DTBS)), 74.7 (CH2(Troc)), 71.7 (C-3), 70.9 (CH2(Linker)), 70.8 (CH2(Linker)), 70.5 (C-4), 70.27 (CH2(Linker)), 70.25 (CH2(Linker)), 67.63 (CH2(Linker)), 67.57 (C-5), 67.1 (C-6), 50.8 (CH2(Linker)), 49.3 (C-2), 27.7 (C(CH3)3(DTBS)), 27.4 (C(CH3)3(DTBS)), 23.4 (C(CH3)3(DTBS)), 21.1 (CH3(OAc)). 20.9 (C(CH3)3(DTBS)); HRESIMS m/z: [M + Na]+ calc for C32H33Cl3N4O13Si, 710.2158; found, 710.2158.

2-[(2′′-Azidoethoxy)ethoxy]-5′-deoxy-4,6-O-di-tert-butylsilylene-2-(2′′′-trichloroethoxycarbonylaminoo)-α-D-galactopyranoside (5): Compound 4 (8.65 g, 12.5 mmol) was placed under N2 and dissolved in dry MeOH (250 mL). NaOMe (68 mg, 1.3 mmol) was added, and the reaction was stirred at room temperature for 4 hours. The solution was then neutralised with Amberlite® IR120 (H+ form) resin, filtered and concentrated under reduced pressure. Flash chromatography on silica gel (toluene → toluene/acetone, 7:3) yielded 5 as a gold-coloured syrup (7.83 g, 96%). Rf = 0.4 (toluene/EtOAc, 3:2); [α]D +67 (c 1.0, CHCl3); 1H NMR (500 MHz, CDCl3) δ 5.58 (d, J = 9.8 Hz, 1H, NH), 4.93 (d, J = 12.0 Hz, 1H, CH2(A)Troc), 4.72 (d, J = 12.0 Hz, 1H, CH2(B)Troc), 4.43 (d, J = 3.0 Hz, 1H, H-4), 4.28 (dd, J = 12.5, 2.2 Hz, 1H, H-6(A)), 4.16 (m, 1H, H-6(B)), 4.11 (dd, J = 10.1, 3.6 Hz, 1H, H-2), 3.87–3.79 (m, 2H, H-5, CH2(Linker)), 3.74 (dd, J = 11.4, 3.2 Hz, 1H, H-3), 3.70–3.63 (m, 9H, CH2(B)Linker, 4 × CH2(Linker)), 3.43–3.34 (m, 2H, CH2(A)Linker), 2.53 (d, J = 11.8 Hz, 1H, OH), 1.07 (s, 9H, C(CH3)3(DTBS)), 1.05 (s, 9H, C(CH3)3(DTBS)), 1.03 (s, 9H, C(CH3)3(DTBS)); 13C NMR (126 MHz, CDCl3) δ 155.4 (C=O(Troc)), 98.6 (C-1), 95.7 (C3(DTBS)), 74.9 (CH2(Troc)), 73.0 (C-4), 70.9 (CH2(Linker)), 70.7 (CH2(Linker)), 70.27 (CH2(Linker)), 70.24 (CH2(Linker)), 70.14 (C-3), 67.8 (C-5), 67.1 (CH2(Linker)), 67.2 (C-6), 52.5 (C-2), 50.8 (CH2(Linker)), 27.7 (C(CH3)3(DTBS)), 27.5 (C(CH3)3(DTBS)), 23.5 (C(CH3)3(DTBS)), 20.9 (C(CH3)3(DTBS)); HRESIMS m/z: [M + Na]+ calc for C32H34Cl2N4O13Si, 763.1606; found, 763.1605.

2-[(2′-Azidoethoxy)ethoxy]ethyl 2,3,4,6-tetra-O-acetyl-β-D-galactopyranosyl-(1→3)-2-deoxy-4,6-O-di-tert-butylsilylene-2-(2′′-trichloroethoxycarbonylaminoo)-α-D-galactopyranoside (7): Donor 6 [16] (7.67 g, 17.4 mmol) and acceptor 5 (7.57 g, 11.6 mmol) were placed under N2 together and dissolved in dry CH3Cl2 (230 mL). AW-300 4 A molecular sieves (5.45 g) were added, and the resulting suspension was stirred at room temperature for 23 hours. NIS (5.23 g, 23.2 mmol) and AgOTf (577 mg, 2.25 mmol) were added, and the reaction was stirred at room temperature for 1 hour. Et3N was then added until the pH became neutral, and the mixture was filtered through Celite®. The filtrate was then washed with water (400 mL), brine (400 mL), dried over MgSO4, and concentrated in vacuo. Compound 7 was isolated by flash chromatography on silica gel (toluene → toluene/EtOAc, 1:4) as an orange foam (8.97 g, 79%). Rf = 0.5 (toluene/EtOAc, 3:7); [α]D
+71 (c 1.0, CH$_3$OH); $^1$H NMR (500 MHz, CD$_3$OD) δ 5.40 (d, J = 3.4 Hz, 1H, H-4Gal), 5.21 (m, 1H, H-2Gal), 5.11 (dd, J = 10.4, 3.4 Hz, 1H, H-3Gal), 5.05 (d, J = 12.2 Hz, 1H, CH$_2$(A)Troc), 4.93 (d, J = 3.6 Hz, 1H, H-1GalNTroc), 4.88 (d, J = 7.8 Hz, 1H, H-1Gal), 4.79 (d, J = 2.8 Hz, 1H, H-3GalNTroc), 4.58 (d, J = 12.2 Hz, 1H, CH$_2$(B)Troc), 4.38 (dd, J = 11.1, 3.6 Hz, 1H, H-2GalNTroc), 4.30 (m, 1H, H-6Gal), 4.22–4.05 (m, 4H, H-5Gal-H-6(B)-H-6(A-B)), 3.96 (dd, J = 11.1, 2.8 Hz, 1H, H-3GalNTroc), 3.90–3.79 (m, 2H, H-5GalNTroc-CH$_2$(2A)Linker), 3.77–3.63 (m, 9H, CH$_2$(B)Linker, 4 × CH$_2$(2Linker)), 3.46–3.39 (m, 2H, CH$_2$(Linker)), 2.16 (s, 3H, CH$_3$(OAc)), 2.12 (s, 3H, CH$_3$(OAc)), 2.05 (s, 3H, CH$_3$(OAc)), 1.97 (s, 3H, CH$_3$(OAc)), 1.13–1.09 (m, 18H, 2 × C(CH$_3$)$_3$(DTBS)), $^{13}$C NMR (126 MHz, CD$_3$OD) δ 171.99 (C=O-OAc), 171.93 (C=O-OAc), 171.6 (C=O-OAc), 171.4 (C=O-OAc), 156.2 (C=O-2Gal), 103.8 (C-1Gal), 99.6 (C-1GalNTroc), 97.2 (C(ClO$_3$)), 79.3 (C-3GalNTroc), 75.6 (C$_2$(Troc)), 74.1 (C-4GalNTroc), 72.6 (C-3Gal), 71.8 (C-5Gal), 71.52 (C$_2$(2Linker)), 71.49 (CH$_2$(Linker), 71.1 (C$_2$(2Linker)), 70.5 (C-2gal), 69.0 (C-5GalNTroc), 68.6 (C-4Gal), 68.4 (C-6), 68.1 (CH$_2$(Linker), 62.9 (C-6), 51.8 (CH$_2$(Linker)), 51.4 (C-2GalNTroc), 28.2 (C(C$_2$(CH$_3$)$_3$(DTBS)), 28.0 (C(C$_2$(CH$_3$)$_3$(DTBS)), 24.3 (C(C$_2$(CH$_3$)$_3$(DTBS)), 21.7 (C(CH$_3$)$_3$(DTBS)), 21.0 (C$_3$(OAc)), 20.6 (C$_3$(OAc)), 20.51 (CH$_3$(OAc)), 20.48 (CH$_3$(OAc)). HREIMS m/z: [M + NH$_4$]$^+$ calecd for C$_{33}$H$_{59}$S$_2$N$_4$O$_{18}$Si, 993.3003; found, 993.3003.


2-[2-(2-Azidoethoxy)ethyl] $β$-Galactopyranosyl-1→3)-2-acetamido-2-deoxy-$α$-Galactopyranoside (1); Compound 2 (1.84 g, 1.85 mmol) was dissolved in MeOH (100 mL) and freshly prepared 1 M NaOMe/MeOH was added until the solution reached pH 10. The reaction was stirred at room temperature for 1 hour, then neutralised with Amberlite® 1210 (H$^+$form) resin. The resin was filtered off, washed with MeOH and the filtrate was concentrated in vacuo. After lyophilisation, I was obtained as a light brown/orange solid (900 mg, 90%) and required no further purification. $R_f$ = 0.6 (EtOAc/MeOH, 2:3); $[$α$_D$]$^+$ +76 (c 1.0, H$_2$O); $^1$H NMR (500 MHz, D$_2$O) δ 4.92 (d, J = 3.7 Hz, 1H, H-1GalNTroc), 4.46 (d, J = 7.8 Hz, 1H, H-1Gal), 4.36 (dd, J = 11.0, 3.7 Hz, 1H, H-2GalNTroc), 4.24 (d, J = 3.0 Hz, 1H, H-4GalNTroc), 4.06 (dd, J = 11.1, 3.1 Hz, 1H, H-3GalNTroc), 4.01 (m, 1H, H-5), 3.92 (d, J = 3.4 Hz, 1H, H-4Gal), 3.87 (m, 1H, CH$_2$(A)Linker), 3.81–3.71 (m, 12H, 2 × H-6(A-B)), 4 × CH$_2$(2Linker), 3.70–3.59 (m, 3H, H-3Gal-H-5, CH$_2$(B)Linker), 3.56–3.48 (m, 3H, H-2Gal-CH$_2$(2Linker), 2.04 (s, 3H, CH$_3$(NHAc)); $^{13}$C NMR (126 MHz, D$_2$O) δ 174.5 (C=O-OAc), 174.3 (C=O-OAc), 174.2 (C=O-OAc), 173.9 (C=O-Gal), 173.8 (C=O-Gal), 173.7 (C=O-Gal), 173.3 (C=O-Gal), 173.3 (C-5Gal), 77.7 (C-3Gal), 77.3 (C-3Gal), 74.9 (C-5), 72.5 (C-3Gal), 70.6 (C-4Gal), 69.7 (CH$_2$(B)Linker), 69.51 (CH$_2$(2Linker)), 69.46 (CH$_2$(2Linker)), 69.2 (CH$_2$(2Linker)), 68.7 (C-4GalNTroc), 68.5 (C-4Gal), 66.5 (CH$_2$(B)Linker), 61.1 (C-6), 60.9 (C-6), 50.1 (CH$_2$(Linker)), 48.5 (C-2GalNTroc), 22.0 (CH$_3$(NHAC)). NMR data match those reported in the literature [3,4]. HREIMS m/z: [M + H$^+$]$^+$ calecd for C$_{32}$H$_{53}$N$_4$O$_{13}$, 541.2357; found, 541.2354.

$^*$2-[2-(2-Azidoethoxy)ethyl] $β$-Galactopyranosyl-3-O-sulfo-$β$-Galactopyranosyl-1→3)-2-acetamido-2-deoxy-$α$-Galactopyranoside disodium salt (2); Compound 1 (1.24 g, 2.29 mmol) and Bu$_4$SnO (645 mg, 2.75 mmol) were placed under N$_2$ together. Dry benzene/DMF (380 mL, 5:1, v/v) was added and the reaction was refluxed at 125 °C using a Dean–Stark apparatus. After 24 hours, the solvent in the receiver was drained, and the
Supporting Information

Supporting Information File 1
NMR spectra of compounds 1–5, 7 and 8.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-20-17-S1.pdf]

Acknowledgements

We thank Dr. Yannick Ortin and Dr. Jimmy Muldoon for NMR and MS support.

Funding

The authors kindly acknowledge Science Foundation Ireland awards 13/IA/1959 and 20/FFP-P/884 (to S.O) and a China Scholarship Council Ph.D. scholarship (to H. M).

ORCID® iDs

Mark Reihill - https://orcid.org/0000-0003-3896-9346
Hanyue Ma - https://orcid.org/0009-0009-7307-0388
Stefan Oscarson - https://orcid.org/0000-0002-8273-4918

References


179
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/bjoc/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/bjoc.20.17
Optimizations of lipid II synthesis: an essential glycolipid precursor in bacterial cell wall synthesis and a validated antibiotic target

Milandip Karak, Cian R. Cloonan, Brad R. Baker, Rachel V. K. Cochrane and Stephen A. Cochrane*

Full Research Paper

Address:
School of Chemistry and Chemical Engineering, Queen's University Belfast, David Keir Building, Stranmillis Road, Belfast, BT9 5AG, UK

Email:
Stephen A. Cochrane* - s.cochrane@qub.ac.uk

* Corresponding author

Keywords:
chemical glycosylation; lipid II; peptidoglycan; polyprenyls; total synthesis

https://doi.org/10.3762/bjoc.20.22

Received: 17 November 2023
Accepted: 26 January 2024
Published: 06 February 2024

This article is part of the thematic issue “Chemical glycobiology”.

Guest Editor: B. Schumann

© 2024 Karak et al.; licensee Beilstein-Institut.
License and terms: see end of document.

Abstract
Lipid II is an essential glycolipid found in bacteria. Accessing this valuable cell wall precursor is important both for studying cell wall synthesis and for studying/identifying novel antimicrobial compounds. Herein, we describe optimizations to the modular chemical synthesis of lipid II and unnatural analogues. In particular, the glycosylation step, a critical step in the formation of the central disaccharide unit (GlcNAc-MurNAc), was optimized. This was achieved by employing the use of glycosyl donors with diverse leaving groups. The key advantage of this approach lies in its adaptability, allowing for the generation of a wide array of analogues through the incorporation of alternative building blocks at different stages of synthesis.

Introduction
Lipid II (Figure 1) is an essential bacterial glycolipid involved in peptidoglycan biosynthesis [1]. It is synthesized on the inner leaflet of the cytoplasmic membrane, before translocation to the outer leaflet, where it is then used as the monomeric building block of peptidoglycan biosynthesis. Lipid II is a validated antibiotic target for clinically prescribed antibiotics including vancomycin and ramoplanin [2]. It is also the target for a host of other antimicrobials (mostly non-ribosomal peptides), including the tridecaptins [3], nisin [4], teixobactin [5], clovibactin [6], malacidin [7], and cilagicin [8].

Despite significant progress in the chemical synthesis of lipid II and its analogues, the scarcity of these compounds and their limited structural diversity present significant obstacles to in-depth explorations of their intricate structural and functional characteristics. This scarcity issue is further exacerbated by an
Figure 1: Structure of lipid II, with variable positions shown in red and antimicrobial-binding motifs highlighted with blue arcs. R$^1$ = H or Ac; R$^2$ = H or Ac; R$^3$ = OH, OMe or NH$_2$; R$^4$ = H or COOH; R$^5$ = Gly, Ala, Ala-Ser/Ala or o-Asp; R$^6$ = OH, OMe or NH$_2$. These structural modifications are described in detail by Münch and co-workers [9]. For more details on lipid II-binding antimicrobials, see recent review by Buijs and co-workers [2].

Results and Discussion

In our studies, the initial glycosyl donors and acceptors (Figure 2; compounds 1a–g and 2a, b) were synthesized using established procedures from the literature, commencing with $\beta$-glucosamine and benzyl 2-acetamido-4,6-O-benzylidene-2-deoxy-β-D-glucopyranoside as the starting materials, respectively [40–43]. Imidate donors 1a and 1e were obtained exclusively...

...overwhelming demand that far exceeds existing supply capacities. To date, the chemical, chemoenzymatic, or biochemical synthesis of lipid II and its variants has been achieved by several research groups [10–27]. Nonetheless, considering the current state of knowledge, the chemical synthesis approach emerges as a more viable strategy in contrast to other methodologies, as it offers the potential to generate ample quantities of lipid II analogues suitable for high-throughput screening endeavors. In recent years, a major focus of the Cochrane lab has been the chemical synthesis of bacterial polyprenyls to study the mechanism of action of antimicrobial peptides that kill bacteria through binding to these polyprenyls [21,28–34]. Lipid II has been of particular interest, and during our synthesis of multiple different lipid II analogues, we have developed several optimizations, which we describe herein. The base lipid II syntheses upon which optimizations were made are our previously reported syntheses of Gram-negative lipid II in 2016 [20] and Gram-positive lipid II (11) in 2018 [23]. Building upon these synthetic strategies we have achieved noteworthy enhancements in glycosylation conditions, including improvements in reaction time and yields. This approach enables the systematic assembly of lipid II and analogues containing shorter polyprenyl chains, specifically farnesyl ($C_{15}$), geranylgeranyl ($C_{20}$), and solanesyl ($C_{45}$). Such short chain analogues are valuable in several applications due to their improved solubility in aqueous systems. Assembly is achieved by integrating distinct carbohydrate, peptide, and polyprenyl phosphate building blocks. This modular synthetic method allows for the strategic substitution of constituent building blocks at different synthetic stages and provides a practical avenue for producing substantial amounts of lipid II analogues. Consequently, this approach offers a more feasible means of addressing the demands associated with biophysical screening pursuits.

Prior research in the field of total synthesis of lipid II has elucidated that specific combinations of protecting groups on glycosyl acceptors and donors, as represented by compounds 1a and 2a in Figure 2, are proficient in the efficient generation of lipid II disaccharide [35,36]. Subsequently, significant endeavors have been directed towards the exploration of glycosyl donors such as N-phthaloyl 3,4,6-O-triacyetyl-2-deoxy-2-amino-α-glucopyranosyl-1-bromide, N,2,2,2-trichloroethoxy-carbonyl-3,4,6-O-triacyetyl-2-deoxy-2-amino-α-glucopyranosyl-1-bromide, and N-phthaloyl-2-deoxy-2-amino-3,4,6-O-tri-acetate-α-glucopyranosyl-1-(2,2,2-trichloroacetoimidate), all of which have proven successful in disaccharide synthesis alongside C6-protected acceptors (2a or 2b in Figure 2) [10,11,14,15,37,38]. More recently, an innovative one-pot glycosylation approach using a (2,6-dichloro-4-methoxy-phenyl)(2,4-dichlorophenyl)-protected glycosyl acceptor has been developed, demonstrating satisfactory stability under Schmidt glycosylation conditions [18]. In general, the outcome of glycosylation hinges on the specific pairing of glycosyl donors and glycosyl acceptors employed in the reaction. Notably, when glycosyl donors such as 1e–g, featuring acyl group protection at the C2 position, are combined with acceptors like 2b, which have acyl groups protecting the C6 position, the reaction kinetics become sluggish, resulting in low conversion rates or no conversion [36,39].
as α-anomers, and 1b and 1g as a 1:1 α:β mixture which were then purified to give the desired α-anomers. Thioglycosides 1c, 1d, and 1f were isolated purely as β-anomers due to anchimeric assistance from the C2 N-acetyl or N-Troc groups. In glycosyl acceptors, the first amino acid of the lipid II pentapeptide, Ala, was incorporated as a 2-(phenylsulfonyl)ethyl ester, as previously reported by Saha and co-workers [44]. This modification prevents a deleterious side reaction occurring, wherein during glycosylation, muramic acid esters undergo a 6-exo-trig cyclization with the 4-OH group. Comprehensive experimental protocols detailing the preparation of these glycosyl donors can be found in Supporting Information File 1.

Next, we conducted an extended investigation into glycosylation, employing a diverse range of glycosyl donors (1a–g) and acceptors (2a and 2b), and the comprehensive results are presented in Table 1. Initially, our approach was guided by the established protocols of Kurosu et al., which had previously demonstrated effectiveness in glycosylating glycosyl trichloroacetimidate 1a and C6-benzylated MurNAc derivative 2a [18]. Despite our efforts to optimize the yield of the target product 3a, involving modifications to reaction conditions such as transitioning from 0 °C to room temperature and extending the reaction duration from 3 to 24 hours, we did not observe the anticipated enhancements (51% yield, entry 1, Table 1). This trend persisted when we attempted glycosylation between C6-acetylated MurNAc derivative 2b and 1a, where the desired product 3b remained elusive (Table 1, entry 2). In fact, glycosyl acceptor 2b failed to yield the desired glycosylation product 3d under the conditions tested (Table 1, entries 7 and 8). Moderate yields of 3a were achieved when using glycosyl donors such as 1b–d under standard conditions A or B (Table 1, entries 3–5). Notably, both Troc-protected thio-donors 1c,d exhibited similar behavior in terms of yield. Unfortunately, no target product 3e was obtained under standard glycosylation conditions A or B when C2-acetamido glycosyl donors (e.g., 1e–g) were subjected to the glycosylation reaction (Table 1, entries 6, 8, and 9). A slight improvement in the yield of 3a was observed when switching from TMSOTf to TfOH as the activator (Table 1, entry 5 vs entry 10). However, substituting TMSOTf with BF3·OEt2 did not yield any target product 3a (Table 1, entry 3 vs entry 12). In our observations, we initially noted that at room temperature, the degradation rate of glycosyl donor 1a exceeded the rate of product formation. This led to a complex mixture consisting of the target product 3a, acceptor 2a, and various degraded products of donor 1a. This situation posed challenges, as even prolonged reaction times did not enhance the product yield, and the subsequent purification of the target product became a difficult task. However, when we conducted the reaction at lower temperatures, the degradation of glycosyl donor 1a persists.
slowed down, and the reaction proceeded at a moderate rate. Eventually, we found that the utilization of extra equivalents of 1a and activators, following conditions akin to those employed by Kurosu, resulted in a significant boost in the yield of the target product to 68% (Table 1, entry 11).

Next, a comprehensive synthetic strategy for the preparation of α-phosphoryl GlcNAc-MurNAc-pentapeptide 7, based on established protocols with minor adjustments was completed (Scheme 1) [10,11]. After the successful glycosylation reaction, disaccharide 3a, protected with C2-Troc and C6-benzyl groups, was efficiently deprotected under acidic conditions using ZnCl₂/Zn, followed by in situ re-acetylation of the C2-amino group and C6-alcohol with acetic anhydride, resulting in the formation of disaccharide 4 in a one-pot fashion. The anomeric benzyl protecting group in disaccharide 4 was then removed via a Pd/C-catalyzed hydrogenation reaction, producing a mixture of α/β-anomers of compound 5. It is noteworthy to mention that the benzyl ether in compound 4 exhibited successful cleavage upon treatment with sodium bromate/sodium dithionite in ethyl acetate/water, while other protecting functionalities like acetyl and phenylsulfonylethyl ester groups remained intact [45]. The ratio of α/β-anomers in compound 5 was found to be influenced by the reaction conditions, consistently favoring the β-anomer. Further transformation of compound 5 involved α-selective phosphate formation using dibenzyl N,N-diisopropylphosphoramidite and 5-(ethylthio)-1H-tetrazole. The resulting α-phosphate intermediate was then oxidized with hydrogen peroxide to yield dibenzyl α-phosphate 6, achieving an overall yield of 89% for these two steps. Removal of the 2-(phenylsulfonyl)ethanol protecting group in compound 6 was successfully achieved through treatment with 1,8-diazabicyclo[5.4.0]undec-7-ene, leading to the formation of the α-phosphoryl GlcNAc-MurNAc-monopeptide derivative. Subsequently, coupling this intermediate with tetrapeptide, TFA·H-L-Ala·γ-d-Glu(OMe)-L-Lys(COCF₃)-d-Ala·d-Ala-OMe under mild conditions resulted in the synthesis of dibenzyl α-phosphoryl GlcNAc-MurNAc-pentapeptide 7 (see Supporting Information File 1 for compre-
To avoid loss of valuable material through HPLC purification, crude 7 is used directly in the next step, and purification performed after the final prenyl phosphate coupling and global deprotection.

Finally, the benzyl-protecting groups in compound 7 were cleaved via hydrogenolysis, followed by co-evaporation of the resulting crude product in pyridine. This yielded a monopyridyl salt, setting the stage for the final lipid coupling and deprotection sequence. To establish the vital lipid diphasate linkage, we employed the phosphoroimidazolate method, as previously utilized in other lipid II total syntheses [10,11]. The monopyridyl α-phosphoryl GlcNAc-MurNAc-pentapeptide was activated with CDI, with excess CDI being neutralized using anhydrous methanol. The resulting phosphoroimidazolate mixture underwent a cross-coupling reaction with prenyl monophosphates [46] in DMF/THF over a four-day period, yielding fully protected versions of lipid II and its analogues. Subsequent global deprotection reactions, using aqueous NaOH, led to the formation of lipid II (11), with an overall yield of 16% (from compound 7) following reversed-phase HPLC purification (Scheme 2). Similarly, farnesyl, geranylgeranyl, and solanesyl-lipid II analogues 8–10 were synthesized with overall yields of 13%, 21%, and 11%, respectively, using the corresponding prenyl phosphates (Scheme 2).
Conclusion

In conclusion, we have successfully optimized a modular approach for the synthesis of lipid II and its analogues, including variants with distinct prenyl-chain lengths. The key to this methodology lies in the optimization of glycosylation conditions, utilizing readily available glycosyl donors, which is a pivotal step in constructing the central disaccharide unit. The adaptability of our method is showcased through the generation of new lipid II analogues, such as geranylgeranyl and solanesyl lipid II analogues, which involve the incorporation of distinct prenyl monophosphates during the final phases of the synthesis. Thus, this strategy holds considerable promise for advancing the synthesis of a diverse range of lipid II analogues, opening avenues for further exploration into their biophysical characteristics, as well as their interactions with antibiotics.

Acknowledgements

We extend our gratitude to Professor Alethea Tabor and Professor Stefan Howorka from University College London for their valuable collaboration on this project. We also express our appreciation to the dedicated technical team at CCE-QUB for their unwavering technical support throughout this project.

Funding

We thank the Engineering and Physical Sciences Research Council for financial support of this project (Grant No EP/V032860/1).

Conflict of Interest

The authors declare no conflict of interest.

Supporting Information

Supporting Information File 1
Experimental procedures, characterization data, and selected copies of $^1$H, $^{13}$C, and $^{31}$P NMR spectra. [https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-20-22-S1.pdf]

Data Availability Statement

The data that supports the findings of this study is available from the corresponding author upon reasonable request.
   
   doi:10.1002/anie.200351534

   Herdewijn, P.  
   doi:10.1135/cccc20051615

41. Grann Hansen, S.; Skydstrup, T.  

42. Subramanian, V.; Mouné-Pymbock, M.; Hu, T.; Crich, D.  

43. Xu, C.; Liu, H.; Li, X.  
   doi:10.1016/j.carres.2011.03.033

44. Saha, S. L.; Van Nieuwenhze, M. S.; Hornback, W. J.; Aikins, J. A.;  
   Blaszczak, L.  
   *C. Org. Lett.* **2001**, *3*, 3575–3577. doi:10.1021/o1016692t

45. Adinolfi, M.; Barone, G.; Guarnieri, L.; Iadonisi, A.  


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/bjoc/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/bjoc.20.22
Elucidating the glycan-binding specificity and structure of 
*Cucumis melo* agglutinin, a new R-type lectin

Jon Lundstrøm¹,², Emilie Gillon³, Valérie Chazalet³, Nicole Kerekes¹,², Antonio Di Maio⁴, Ten Feizi⁴, Yan Liu⁴, Annabelle Varrot³ and Daniel Bojar*¹,²

**Full Research Paper**

Abstract

Plant lectins have garnered attention for their roles as laboratory probes and potential therapeutics. Here, we report the discovery and characterization of *Cucumis melo* agglutinin (CMA1), a new R-type lectin from melon. Our findings reveal CMA1’s unique glycan-binding profile, mechanistically explained by its 3D structure, augmenting our understanding of R-type lectins. We expressed CMA1 recombinantly and assessed its binding specificity using multiple glycan arrays, covering 1,046 unique sequences. This resulted in a complex binding profile, strongly preferring C2-substituted, beta-linked galactose (both GalNAc and Fuca1-2Gal), which we contrasted with the established R-type lectin *Ricinus communis* agglutinin 1 (RCA1). We also report binding of specific glycosaminoglycan subtypes and a general enhancement of binding by sulfation. Further validation using agglutination, thermal shift assays, and surface plasmon resonance confirmed and quantified this binding specificity in solution. Finally, we solved the high-resolution structure of the CMA1 N-terminal domain using X-ray crystallography, supporting our functional findings at the molecular level. Our study provides a comprehensive understanding of CMA1, laying the groundwork for further exploration of its biological and therapeutic potential.

**Introduction**

Lectins have long been the subject of intense scientific scrutiny, serving as molecular bridges that span the realms of biochemistry, cellular biology, and biomedicine. These carbohydrate-binding proteins boast a range of functions, acting as recognition modules in cell–molecule and cell–cell interactions, thereby playing vital roles in immune defense, regulation of
growth, and apoptosis [1]. In plants, they serve as essential components in development, immunity, and stress signaling [2,3].

In light of the burgeoning interest in the intersection of glycobiology and biomedicine, the characterization of new lectins has carved out a significant niche in scientific research. Specifically, lectins have emerged as invaluable tools for staining cells and tissues, thereby offering insights into cellular heterogeneity and function. For instance, the use of wheat germ agglutinin (WGA) and concanavalin A (ConA) has been instrumental in selectively staining cells based on their glycan expression [4], including single-cell approaches [5,6]. In the realm of therapeutics, lectins such as mistletoe lectins have shown promise in cancer therapy, by virtue of their ability to induce apoptosis in malignant cells [7]. Further, the creation of lectin arrays [8,9], which employ a diverse set of characterized lectins, has enabled high-throughput glycan profiling, thereby advancing both diagnostic methods and biomarker discovery. Examples include arrays that can rapidly profile alterations in glycosylation patterns, pivotal in many diseases and inflammatory changes [10,11].

Traditionally, lectins are divided into classes based on structural similarity and, by extension, common folds [12]. Still, shared binding specificity does not always follow from structural similarity, exemplified by divergent evolution within lectin families as well as independent emergence of similar binding patterns [13]. Many of the most commonly used lectins for the abovementioned applications are R-type lectins, especially those derived from plants. Examples include SNA (from Sambucus nigra, binding Neu5Acα2-6 [14]) or RCA1 (from Ricinus communis, binding terminal β-linked galactose [15]).

Yet, despite the extensive studies on plant lectins, particularly R-type lectins, there are still significant gaps in our understanding. Further, in general, few melon lectins have been studied in detail. Some reports indicate the presence of chitooligosaccharide-binding (i.e., β1-4 GlcNAc oligomers) lectins from phloem exudates of melons [16,17], as well as R-type lectins in bitter melon [18], yet not much else is known about binding specificities exhibited by lectins derived from melons. In particular, existing research in this area often lacks a comprehensive characterization that includes both functional and structural analysis of these lectins.

Here, we introduce a novel member of characterized melon lectins, namely the Cucumis melo agglutinin (CMA1), an R-type lectin derived from melon. Prior to our study, CMA1 was only a predicted protein from genomic sequencing, with moderate certainty scores on lectin-specific databases. Our comprehensive analysis using glycan array experiments, thermal shift assays, and high-resolution X-ray crystallography not only confirms its classification as a functional R-type lectin but also provides a deep dive into its unique glycan-binding profile and high-resolution 3D structure. Overall, we present a deeply characterized new lectin with a unique binding profile of specifically recognizing C2-substituted galactose in the context of glycans.

Results and Discussion

Identification and production of a new lectin from the melon Cucumis melo

CMA1 is a predicted protein from whole-genome shotgun sequencing of leaves from the melon plant Cucumis melo (variant makuwa, taxon ID: 1194695) [19] and has, to our knowledge, never been studied before. With prediction scores of 0.453 on LectomeXplore [12] and 0.251 on TrefLec [20] (from 0, lowest, to 1, highest), CMA1 is moderately certain in its prior classification as a lectin. CMA1 comprises 291 amino acids and is predicted to fold into two linked β-trefoil domains belonging to carbohydrate-binding module family 13 (CBM13) and placing it into the group of R-type lectins. Both CBM13 domains are likely to exhibit carbohydrate-binding activity due to the conservation of key amino acids in at least one of the three potential binding sites. In contrast to other R-type lectins such as ricin, it lacks a catalytic domain.

As R-type lectins are both a well-investigated family of lectins and widely used in research and beyond, we first wanted to analyze where CMA1 would be situated in the broader context of R-type lectins. A multiple sequence alignment of binding domains of representative R-type lectins (Figure 1a) showed that CMA1 exhibited a binding domain with a sequence relatively similar to those of the plant lectins SNA and ricin. However, we note that, in general, the substantial heterogeneity of binding motifs of even closely related lectins (SNA: Neu5Acα2-6, ricin: Gal/GalNAc) does not allow for a strong a priori hypothesis of what CMA1 would bind, even though R-type lectins in general are thought to prefer the Gal/GalNAc type motif mentioned in the context of ricin [21].

We next aligned the individual units of the tandem repeat CBM13 domains, indicated by the N-terminal (34-158) and C-terminal units (162-286) and compared those to the domains of ricin (Figure 1b). R-type lectins have a characteristic Q-x-W structural motif close to their binding site, which is highly conserved [21]. We report that CMA1 largely follows this trend, with three such binding sites in both N- and C-terminal domain, albeit with imperfect overlap. Based on the location of
Figure 1: Characterizing a new lectin from the melon Cucumis melo. (a) Evolutionary relationships of common R-type lectins. For a range of representative R-type lectins, we aligned their protein sequences via MUSCLE [22] and built a neighbor-joining tree with the resulting alignment distances, which is shown as a cladogram. For each protein, we only used the lectin domain, as annotated by UniProt or InterPro. For each protein, a representative binding specificity, based on literature reports, is provided. (b) Similarity of the two CBM13 domains in CMA1. Using MUSCLE to align the N-terminal (34–158) and C-terminal domains (162–286) of CMA1 and ricin (321–448 and 451–575), we indicated the position of the conserved Q-x-W motif in R-type lectins. (c) Recombinant expression of CMA1 in mammalian cells. SDS-PAGE and anti-His-tag Western blot of fractions from the expression of CMA1 protein in CHO-S cells. Note the smeared band indicating the presence of glycosylation. (d) Recombinant expression of CMA1 in bacteria. SDS-PAGE gels of the His-tag affinity chromatography and cation exchange chromatography from the expression of CMA1 protein in E. coli BL21* cells.
the known binding pocket of the R-type lectin ricin and the respective sequence conservation in CMA1, we postulate binding sites around W\textsuperscript{63} for the N-terminal domain and F\textsubscript{273} for the C-terminal domain of CMA1.

As binding specificities of melon lectins in general (beyond chito-oligosaccharides), and CMA1 in particular, are still unknown, we set out to measure, quantify, and understand the glycan-binding properties of CMA1 in depth, as an archetypal example of melon lectins. For this, we needed to express the lectin recombinantly. As it is a secreted plant protein, we elected to express it in mammalian cell lines, to maximize the chances of a functional protein, because of post-translational modifications that would be lacking in bacteria as well as the oxidative environment of the secretory pathway, as CMA1 exhibits predicted disulfide bridges. A single step of His-tag affinity chromatography was sufficient to yield protein of adequate purity and good yield (>15 mg of eluted protein from 800 mL of cell culture, Figure 1c).

In parallel, we also expressed CMA1 in a bacterial expression system, which allowed us to ascertain whether binding was influenced by lectin glycosylation. The full-length mature protein (6–264) and individual N- or C-terminal domains were expressed using a N-terminal fusion comprising DsbC and a hexa-His tag, cleavable by TEV (Tobacco etch virus) protease. Despite the presence of the DsbC signal peptide, we did not observe periplasmic localization, and all proteins were instead purified from the cytoplasm. Ni-NTA affinity chromatography followed by TEV protease cleavage of the fusion construct and subsequent reverse Ni-NTA affinity chromatography resulted in significant co-purification of E. coli contaminants, necessitating an extra purification step, where cation exchange chromatography allowed us to obtain pure fractions of CMA1\textsubscript{16–291}. Of note, this additional purification step was not necessary for the purification of the CMA1 N-terminal domain (Figure 1d). Expression of the CMA1 C-terminal domain did not yield sufficiently pure and monodisperse protein for further biochemical and structural analyses.

*Cucumis melo* agglutinin binds C2-substituted, beta-linked galactose

We then set out to answer the question whether CMA1 was a functional lectin and, if yes, what its binding specificity was. The standard approach to elucidate lectin binding specificity is via glycan array experiments. Here, tagged soluble lectin is added to, often, immobilized glycans and bound lectin is quantified via fluorescence scanners, which can be paired with glycan information due to the known arrangements of immobilized glycans on the plate. To cover the broadest possible sequence space, we tested our eukaryotically produced CMA1 protein against the two largest glycan arrays at the National Center for Functional Glycomics (NCFG, Figure 2a) and the Glycosciences Laboratory at Imperial College London (ICL, Figure 2b). We note that, together, this encompasses 1,046 unique glycan sequences, spanning all major glycan classes and substantial taxonomic diversity. Next to these unique sequences, even more effects stem from a variety of linkers with which these molecules are immobilized.

In general, we observed two binding preferences that were strongly enriched among bound sequences, namely glycans containing Fuc\textalpha\textbeta\textgamma\texti-1-2Gal or N-acyetyl (GalNAc) moiety (Figure 2d). We thus conclude that CMA1 is highly specific for C2-substituted galactose. We further argue for a preference for a beta-linked epitope as, while we do observe binding to structures containing alpha-linked GalNAc, the binding to their beta-linked counterparts was generally stronger (e.g., GalNAc: 1.57 vs GalNAcb: 2.21, in z-scores (see Experimental section)). In part, this is reminiscent to the LacdiNAc binding specificity of *Clitocybe nebularis* lectin (CNL; Figure 1a) [27].

An important finding from the ICL array was that CMA1 exhibited robust binding to glycosaminoglycans (GAGs; Figure 2e; Supporting Information File 1, table “imperial”), in particular chondroitin sulfate (CS) C and A. Given the preference for terminal binding epitopes described above, the question naturally arose how the binding to these longer-chain glycans works. On the ICL array, CS sequences are typically capped with 4,5-unsaturated hexuronic acid derivatives on their non-reducing end and, thus, do not provide terminal GalNAc epitopes for binding. Further, while CMA1 did also bind to GalNAc-terminated GAGs (e.g., CSC-5, CSA-5), we measured higher binding to similar GAGs without the terminal GalNAc in several cases (Figure 2d,e). While some of the GAG probes varied in their immobilization amounts, we confirmed these results in a GAG-focused array (data not shown). We thus posit a binding to internal GalNAc epitopes for the case of GAG binding, potentially mediated by several binding sites.

This argument is strengthened by the observation that the highest observed binding to CSC and CSA was not with the
Figure 2: Characterizing the binding specificity of CMA1. (a, b) Lectin produced in mammalian cells was analyzed on the NCFG array (a) and the ICL array (b). Representative structures bound by CMA1 are shown via the “Symbol Nomenclature For Glycans” (SNFG), drawn with GlycoDraw [23]. Everything except the assigned binding motif is shown with added transparency. Full array data are available in Supporting Information File 1, tables “cfg” and “imperial”. (c) Enrichment analysis of glycan array data. For both NCFG and ICL array data, we used the get_pvals_motif function from glycowork [24] with the keywords ‘terminal’ and ‘exhaustive’, to obtain significantly enriched motifs. *p < 0.05. (d) Common binding motif on the atomic level. Glycan 3D structures for the binding motifs were obtained from the GLYCAM web server [25,26]. (e) Binding of CMA1 to glycosaminoglycans. We grouped chondroitin sulfate (CS) types (A, B, and C) and plotted CMA1 binding against CS chain length. Shown are mean values with their 95% confidence interval. (f) Comparison of CMA1 and RCA1 binding. Glycans with a z-score of at least 0.5 in at least one lectin were retained and plotted as a hierarchically clustered heatmap via the get_heatmap function of glycowork. Representative glycans are shown.
shortest sequences and required at least three repeats, with longer sequences such as CSC-18 even exhibiting the highest binding on the entire array (although we note that the longest GAG sequences were not generally the best binders, potentially hinting at steric clashes or density effects). Another supporting finding can be seen in the fact that CSB (exhibiting iduronic acid in α-configuration, rather than its epimer, glucuronic acid, in β-configuration) showed virtually no binding to CMA1, further arguing for contacts of the GAG chain with the binding site. Lastly, we note that both CSC and CSA contain sulfated GalNAc, which, together with the observation of GalNAc6Sβ1-4GlcNAc as one of the highest binders on the NCFG array, leads us to speculate that sulfation further enhances CMA1 binding, a pattern that has been observed for several lectins [28].

Overall, this characterized binding specificity seemed distinct from other R-type lectins and we thus further compared it to a typical R-type lectin, Ricinus communis agglutinin (RCA1), on the ICL array. Canonically, RCA1 binds β-linked terminal galactose residues, which is generally what we also found in our array experiments, with Galβ in various substructures and glycan types, particularly in those with multiple branches (Figure 2f). At best, the same sequences showed weak binding to CMA1, as they lacked a C2-substitution (Figure S1, Supporting Information File 2). Conversely, CMA1-favored sequences, containing Fucα1-2Gal or GalNAc epitopes, were on average not bound by RCA1 (the exception being sequences in which there was an additional free Galβ terminus). Similarly, most chondroitin sulfate probes were not bound by RCA1. This gives rise to the conclusion that CMA1 does not merely tolerate but rather actively and strongly prefers C2-substituted Gal, while RCA1 does not even tolerate these substitutions. Interestingly, we also find that fucosylation of the GlcNAc residue (as in Lewis antigen motifs) completely abrogates CMA1 binding (Figure S1, Supporting Information File 2), despite the presence of Fucα1-2Gal, likely due to steric clashes in the binding pocket. We thus conclude that the binding profile of CMA1 is distinct from that of the typical R-type lectin RCA1 and unusual for a R-type lectin in general. We also note that the flexibility of accommodated C2 substituents (from N-acetyl moieties to whole monosaccharides), could make CMA1 an interesting candidate for probing synthetically produced glycans with novel substituents.

It is of course interesting to speculate about the physiological role of CMA1 in melons, yet this is hard to probe. It is noteworthy, however, that the glycan types in which its preferred binding motifs occur (O-glycans, milk glycans, GAGs) are absent from most plants, including melons. We thus hypothesize that the role of this lectin might be to recognize non-self epitopes, such as for protection against pathogens, which is a common function in plant lectins [3].

Validating binding in solution and assessing binding affinity

As CMA1 both exhibited multiple binding sites and robust binding to blood group epitopes (H-antigen), we hypothesized that it would be capable of agglutinating red blood cells, justifying its new name. When testing the protein recombinantly produced in mammalian cells, incubation with rabbit erythrocytes indeed resulted in moderate agglutination (Figure 3a), which also demonstrated the binding to these glycan substructures in a physiological context.

To further strengthen the case for CMA1 binding glycans in solution, and corroborate its binding specificity with orthogonal methods, we used a thermal shift assay. Herein, the binding of ligands is assessed by the stabilization of the protein, measured by a denaturation curve. Both the protein produced in mammalian and in bacterial cells exhibited similar melting temperatures here, of approximately 42 °C (Figure 3b). Then, we tested the binding of CMA1 to GlcNAc, GalNAc, and H type 2 blood group antigen (BGH-T2; Fucα1-2Galβ1-4GlcNAcβ1-3Gal; Figure 3c). This resulted in clear melting points shifts for both GalNAc and BGH-T2 to up to 50 °C, yet importantly not for GlcNAc, demonstrating both binding in solution and a further confirmation of the binding specificity obtained by the array experiments. We note that the functional activity of bacterially produced CMA1 indicates that potential modification by glycosylation is not required for ligand binding.

Next, we set out to quantify the binding affinity of CMA1 to its ligands. Lectins often only exhibit weak to moderate binding affinities, which is somewhat ameliorated by an increased avidity on the side of the lectin but also a dense presentation of the bound glycan epitope on the cell surface. We therefore used surface plasmon resonance (SPR) spectroscopy to derive binding constants for the interaction between CMA1 and GalNAc. A single cycle kinetics approach was applied, resulting in a measured $K_D$ of 1.66 ± 0.08 µM (Figure 3d,e). Inhibiting binding of CMA1 to the GalNAc chip through a dilution series of N-acetylactosamine (LacNAc) via multicycle kinetics allowed us to derive an IC$_{50}$ of 1.4 µM (Figure S2a,b; Supporting Information File 2). No inhibition was observed with chondroitin 6-sulfate tetrasaccharide (CSC), and only very weak inhibition for BGH-T2 but no IC$_{50}$ could be determined as we could not increase the concentration to reach the plateau. For the recombinant CMA1-Nter, no binding could be observed on the GalNAc chip. This suggests either avidity effects in conjunction with the C-terminal domain or a high-affinity site on the C-terminal domain, giving rise to the measured $K_D$ of the
full-length protein. Still, we were able to measure the affinity of CMA1-Nter to GalNAc in solution by isothermal calorimetry (ITC), obtaining a $K_D$ of 940 µM, confirming the low affinity (Figure S2c,d; Supporting Information File 2).

**Structural insights from the N-terminal domain of CMA1**

Given the unusual binding specificity exhibited by CMA1, we were intrigued to elucidate the molecular mechanism that would enable the specific binding of C2-substituted galactose. The natural hypothesis here would be the creation of an additional pocket in the 3D structure of the binding site, accommodating the additional substituent at C2. However, as we observed little to no binding to unsubstituted galactose, we rather hypothesized the existence of specific interactions made with the C2-substituents, that did not exist in other R-type lectins such as RCA1. To determine this, we set out to resolve the detailed three-dimensional structure of CMA1 via X-ray crystallography.

We obtained several hits for the full-length protein after sparse screening using a crystallization robot at the HTX platform, EMBL, Grenoble. Pill-shaped crystals obtained under conditions of a high salt concentration, in particular ammonium sulfate (Figure S3, Supporting Information File 2), did not give rise to any diffraction. Multiple layer plate or needles clusters were obtained in the presence of PEGs, but only showed weak diffraction ($\approx$3.5 Å). Finally, in the presence of 20% PEG 8K, 0.2 M MgCl$_2$, and 0.1 M Tris HCl pH 8.5, single diamond-shaped crystals were obtained after 1–2 days for the N-terminal domain (Figure S3, Supporting Information File 2). High-resolution diffraction of the crystals allowed us to solve the CMA1-Nter structure in complex with LacNAc at 1.3 Å and GalNAc at 1.55 Å (see data and refinement statistics in Table 1).
Table 1: Data collection and refinement statistics.

<table>
<thead>
<tr>
<th>Complex</th>
<th>CMA1-Nter-LacNAc</th>
<th>CMA1-Nter-GalNAc</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data collection</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beamline</td>
<td>Soleil PX1</td>
<td>Soleil PX2</td>
</tr>
<tr>
<td>wavelength (Å)</td>
<td>0.97856</td>
<td>0.98011</td>
</tr>
<tr>
<td>space group</td>
<td>I2</td>
<td>I2</td>
</tr>
<tr>
<td>cell parameters a, b, c (Å)</td>
<td>36.70 36.78 94.79</td>
<td>36.61 36.86 94.81</td>
</tr>
<tr>
<td>α, β, γ (°)</td>
<td>90.00 99.24 90.00</td>
<td>90.00 99.17 90.00</td>
</tr>
<tr>
<td>protein chains in a.u.</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>resolution (Å)</td>
<td>46.78–1.32 (1.34–1.32)</td>
<td>35.68–1.55 (1.58–1.55)</td>
</tr>
<tr>
<td>CC1/2 (%)</td>
<td>99.9 (96.9)</td>
<td>99.8 (85.7)</td>
</tr>
<tr>
<td>Rmerge (within I+/I−) (%)</td>
<td>0.055 (0.369)</td>
<td>0.052 (0.496)</td>
</tr>
<tr>
<td>Rmeas (within I+/I−) (%)</td>
<td>0.059 (0.400)</td>
<td>0.064 (0.618)</td>
</tr>
<tr>
<td>Rpim (within I+/I−) (%)</td>
<td>0.022 (0.153)</td>
<td>0.037 (0.364)</td>
</tr>
<tr>
<td>mean I/σ(I)</td>
<td>25.2 (5.7)</td>
<td>14.4 (2.9)</td>
</tr>
<tr>
<td>completeness (%)</td>
<td>99.8 (96.0)</td>
<td>99.7 (99.9)</td>
</tr>
<tr>
<td>number reflections</td>
<td>399970 (18410)</td>
<td>95115 (4581)</td>
</tr>
<tr>
<td>number of unique reflections</td>
<td>29695 (1434)</td>
<td>18279 (911)</td>
</tr>
<tr>
<td>multiplicity (Å)</td>
<td>13.5 (12.8)</td>
<td>5.2 (5.0)</td>
</tr>
<tr>
<td>Wilson B-factor (Å²)</td>
<td>14.1</td>
<td>19</td>
</tr>
<tr>
<td><strong>Refinement</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>resolution (Å)</td>
<td>46.78–1.32</td>
<td>35.69–1.55</td>
</tr>
<tr>
<td>no. reflections/no. free reflections</td>
<td>28192/1503</td>
<td>17373/905</td>
</tr>
<tr>
<td>Rwork/Rfree (%)</td>
<td>14.35/18.58</td>
<td>16.3/20.4</td>
</tr>
<tr>
<td>R.m.s. bond lengths (Å)</td>
<td>0.0130</td>
<td>0.0127</td>
</tr>
<tr>
<td>Rmsd bond angles (°)</td>
<td>1.721</td>
<td>1.893</td>
</tr>
<tr>
<td>Rmsd chiral (Å³)</td>
<td>0.097</td>
<td>0.092</td>
</tr>
<tr>
<td>no. atoms / Bfac (Å²)</td>
<td>1029/15.1</td>
<td>985/19.95</td>
</tr>
<tr>
<td>protein</td>
<td>26/20.3</td>
<td>30/22.3</td>
</tr>
<tr>
<td>ligand</td>
<td>3/21.9</td>
<td>3/27.0</td>
</tr>
<tr>
<td>cadmium</td>
<td>248.28.7</td>
<td>176.31.8</td>
</tr>
<tr>
<td>water</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Ramachandran allowed (%)</td>
<td>99</td>
<td>100</td>
</tr>
<tr>
<td>favored (%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>outliers</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

aValues in parenthesis refer to the highest-resolution shell.

residues of the N-terminal construct (Val6 to Asp132) could be modelled, and unambiguous electron density permitted us to locate and model four cation binding sites (three in each structure) and one sugar binding site (Figure 4a,b and Figure S4, Supporting Information File 2).

The complexed structures allowed us to shed light on the arrangement of the ligand in the binding site (Figure 4c,d). While lectins such as CMA1 typically can present three binding pockets in their CBM13 domain, we hypothesized that the N-terminal half of CMA1 would in fact only exhibit two functional binding sites. However, only the alpha site was found occupied with a carbohydrate here. It is found in a shallow groove, supporting our data on the lack of a distinct distal binding specificity. We report a tight coordination of the O3 and O4 hydroxy groups of the galactose residue involving Asp21, Asn43, and Gln41 side chains, as well as the Gly24 main chain nitrogen. CH–π stacking and hydrophobic interactions occur between the aromatic ring of Trp36 and the alpha face of the ring as well as the hydroxymethyl moiety of the galactose residue, additionally ensuring specificity for galactoside over glucoside as an equatorial conformation of the O4 hydroxy
group would lead to steric clashes and loss of strong hydrogen bonding.

In the LacNAc-complexed structure (PDB ID 8R8A) [29], the GlcNAc residue did not seem to engage in extensive interactions, with only a hydrogen bond between the N-acetyl moiety and the main chain oxygen of Gly24 and hydrophobic interaction with the aromatic ring of Tyr26 (Figure 4c). Further, beyond the C2 position of galactose, a cavity filled with coordinated water molecules hinted at the binding mode for C2-substituted galactose. Notably, the seemingly inactive beta site was found to be occupied by a cadmium ion (Figure S4, Supporting Information File 2), supporting our ITC and SPR data where no multivalent binding effects were observed for the single-domain N-terminal construct.

In the GalNAc-complexed structure (PDB ID 8R8C) [30], the N-acetyl group of GalNAc extended beyond C2 into the cavity noted in the LacNAc complex. While no direct interactions with the protein backbone were observed, we found one water mole-
cule to mediate hydrogen bonding between the oxygen of the N-acetyl group and the Asn side chain oxygen (Figure 4d). Both GalNAc anomers could be observed, showing interactions through water molecule coordination with the Trp ring nitrogen (alpha anomer) or the Gly main chain oxygen (beta anomer).

**Conclusion**

Our work presents a substantial exploration of the binding specificity and mechanism of the hitherto uncharacterized lectin CMA1 from melons. The binding specificity of CMA1, C2-substituted galactose that is preferentially presented in a β-configuration, enables it to bind to a range of biologically relevant epitopes, such as LacdiNAc, S, blood group H, and chondroitin sulfate motifs. Further, the inhibition of binding by the presence of Lewis antigen motifs additionally narrows it binding specificity. Our binding data and structural information lead us to the conclusion that crucially positioned asparagine residues facilitate this unusual binding specificity that delineates CMA1 from typical R-type lectins such as RCA1. Together, these results advance our knowledge of R-type lectins in general and the range of their binding specificities, but also our knowledge of melon lectins in particular, which has remained limited so far. Further experiments are still required to determine the role of the C-terminal domain, as well as the physiological function of the full-length CMA1 protein.

**Experimental**

**Recombinant protein expression**

For mammalian expression, the gene for CMA1 (A0A1S4E5V9) was synthesized with human-optimized codons and a C-terminal hexa-His tag (GSHHHHHH). We then cloned this gene into a pC1 backbone (U47119; Promega GmbH) for expression in mammalian cells under a constitutive cytomegalovirus (CMV) promoter. Then, the Mammalian Protein Expression core facility at the University of Gothenburg transfected this plasmid into FreeStyle™ CHO-S cells (Cat nr R80007, ThermoFisher Scientific). Cells were cultured in Freestyle™ CHO medium at 37 °C in 5% CO₂ in Optimum Growth™ flasks (Thomson instrument company) at 130 rpm in a Multi-clone to mediate hydrogen bonding between the oxygen of the N-acetyl group and the Asn side chain oxygen (Figure 4d). Both GalNAc anomers could be observed, showing interactions through water molecule coordination with the Trp ring nitrogen (alpha anomer) or the Gly main chain oxygen (beta anomer).

For bacterial expression, the gene of CMA1 (33–291, corresponding to residues 6–291 of the mature protein) with optimized codons for Escherichia coli was synthesized flanked by Ncol and Xhol restriction sites where L² was mutated to valine. The gene was inserted in the homemade plasmid pET40b-TEV where the enterokinase cleaving site was replaced by a TEV cleavage site by site directed mutagenesis. This plasmid was obtained by PCR using pET-40b(+) (Novagen, Merck, #70091) as template and the following primers: forward (ggccgatctgcggccatgcggatcgcgaattttcagggcccaccctctgcgtgtctggtggccgagtcgac) and reverse (GGTACCAGATCTGGGGGTGTCCATGTTGCGTGCGCCG) with complementary sequence underlined. PCR was performed using PrimeSTAR DNA polymerase (Takara #TAK045A); then the product was digested by DpnI and finally transformed in NEB5α strain (New England Biolabs, #C2992H). Both gene and vector were digested by Ncol and Xhol restriction enzymes (New England Biolabs) prior to purification on agarose gel using Monarch Gel extraction kit and supplier instructions (New England Biolabs, #T1020S) and ligation using the DNA ligation kit, Mighty Mix (Ozyme, Takara, #TAK6023Z), at room temperature to form the pET40b-TEV-CMA1 plasmid.

The N-terminal domain of CMA1 (6–132 in mature protein) was amplified by PCR using the following primers: forward (ACGCCATGGTGAGCCGTTCTACGC) and reverse (ATATCTCGAGTTAATCTG CCGTACCCCAGGATTGTGGTTG) with complementary sequence underlined. PCR was performed using PrimeSTAR DNA polymerase (Takara #TAK045A); then the product was digested by DpnI and finally transformed in NEB5α strain (New England Biolabs, #C2992H). Both gene and vector were digested by Ncol and Xhol restriction enzymes (New England Biolabs) prior to purification on agarose gel using Monarch Gel extraction kit and supplier instructions (New England Biolabs, #T1020S) and ligation using the DNA ligation kit, Mighty Mix (Ozyme, Takara, #TAK6023Z), at room temperature to form the pET40b-TEV-CMA1 plasmid.

The N-terminal domain of CMA1 (6–132 in mature protein) was amplified by PCR using the following primers: forward (ACGCCATGGTGAGCCGTTCTACGC) and reverse (ATATCTCGAGTTAATCTG CCGTACCCCAGGATTGTGGTTG) with complementary sequence underlined. PCR was performed using PrimeSTAR DNA polymerase (Takara #TAK045A); then the product was digested by DpnI and finally transformed in NEB5α strain (New England Biolabs, #C2992H). Both gene and vector were digested by Ncol and Xhol restriction enzymes (New England Biolabs) prior to purification on agarose gel using Monarch Gel extraction kit and supplier instructions (New England Biolabs, #T1020S) and ligation using the DNA ligation kit, Mighty Mix (Ozyme, Takara, #TAK6023Z), at room temperature to form the pET40b-TEV-CMA1 plasmid. The purified PCR fragment of 359 bp was digested by Ncol and Xhol restriction enzymes, then ligated into pET40b-TEV vector, and finally transformed in NEB5α strain to form the pET40b-TEV-CMA1-Nter and pET40b-TEV-CMA1-Cter plasmids. All plasmids and new vectors were verified by sequencing (Eurofins Genomics, Ebersberg, Germany). Primers were purchased from Eurofins Genomics (Ebersberg, Germany).

**E. coli BL21*(DE3) [Invitrogen, #C601003] cells were transformed by heat shock at 42 °C with pET40b-TEV-CMA1 and Tuner(DE3) [Novagen, #70623] cells with pET40b-TEV-CMA1Nter prior pre-culturing in lysogeny broth (LB) [Invitrogen, #C601003].**
For purification of bacterial recombinant CMA1, each gram of cell pellet was resuspended with 5 mL of buffer A (20 mM Tris-HCl pH 7.5, 500 mM NaCl). After addition of 1 µL of Denarase® (C-LEcta GmbH, #20804) and moderate agitation on a rotating wheel for a period of 30 min at room temperature, cells were lysed using a cell disruptor (Constant Systems Ltd, UK) under a pressure of 2.5 kbar. The lysate was cleared by centrifugation at 24,000g for 30 min at 4 °C and passed through a 0.45 µm syringe filter prior to affinity chromatography purification using 1 mL HisTrap™ HP column (Cytiva) pre-equilibrated with buffer A and an NGC chromatography system (Bio-Rad). After loading the cleared lysate, the column was washed with buffer A plus 50 mM imidazole (Sigma-Aldrich, Merck, #56749) to remove all contaminants and unbound proteins. CMA1 was eluted by a 20 mL linear gradient from 50 mM to 500 mM imidazole in buffer A. The fractions were analyzed by SDS-PAGE with 15% gel and those containing CMA1 were collected and deprived of imidazole by buffer exchange in buffer A using a Macro and Microsep Advance Spin 3 kDa MWCO centrifugal filter (Pall). The N-terminal His-tag was removed for each glycan, mitigating the effects of outliers. The results can be found in Supporting Information File 1, table “cfg”.

ICL array

For experiments, a standard binding buffer (20 mM Tris-HCl pH 7.4, 150 mM NaCl, 2 mM CaCl₂, 2 mM MgCl₂, 0.05% Tween 20, 1% BSA) was used. CMA1 binding was probed by incubation with a penta-His-488 antibody (5 µg/mL). CMA1 was tested in two concentrations (5 and 50 µg/mL) on Version 5.4 of the printed CFG array, consisting of 585 printed glycans in replicates of six. Results from replicates were combined as average RFU (raw fluorescence unit). For this average, the highest and lowest value was removed for each glycan, mitigating the effects of outliers. The results can be found in Supporting Information File 1, table “imperial” and “rca_imperial”, with the array generated in Supporting Information File 3 according to the MIRAGE guidelines (Minimum Information Required for A Glycomics Experiment) [32].

For both array types, data were transformed into z-scores by subtracting the mean value across the array and dividing the results by the standard deviation.
Agglutination assay
The hemagglutinating activity of CMA1 was determined in V-bottom 96-well plates by a twofold serial dilution procedure in PBS using rabbit red blood cells (Atlantis France). 25 µL of 4% erythrocyte suspension was added to an equal volume of the sample, and the mixture was incubated for 60 min at room temperature. Starting concentrations were: CMA1 0.6 mg/mL, AAL 0.5 mg/mL, ConA 2.5 mg/mL, RCA1 2.5 mg/mL, and SNA 0.5 mg/mL.

Thermal shift assay
Thermal shift assays were performed using a Mini Opticon Real Time PCR machine (BioRad). 0.6 mg/mL protein in PBS was mixed with SYPRO Orange (Sigma-Aldrich, Merck, #S5692) and glycans ligand (10 mM GalNAc; Carboxyn, #MA04390; 10 mM GlcNAc, Carboxyn, #MA00834; 10 mM blood group H type-2 tetrasaccharide; Elicityl, GLY032-2) in a total reaction volume of 25 µL. The temperature was raised by 1 °C/min from 25 to 100 °C, and fluorescence readings were taken at each step.

Surface plasmon resonance spectroscopy
Experiments were performed using a Biacore X100 instrument (Cytiva) at 25 °C in HBS-T running buffer (10 mM HEPES pH 7.4, 150 mM NaCl and 0.05% Tween 20). Biotinylated PAA-GalNAc (Lectinity, GlycoNZ, #0031-BP) was immobilized on CM5 chips (Cytiva #BR100012) that were coated previously with streptavidin (Sigma-Aldrich, Merck, #S4762), following standard protocol. Biotinylated GalNAc was diluted to 2 µg/mL in HBS-T before being injected into one of the flow cells of the chip. An immobilization level of 900 response units (RU) was obtained. A reference surface was always present in flow cell 1, allowing for the subtraction of bulk effects and non-specific interactions with streptavidin. The mammalian-produced CMA1 was injected in single cycle kinetic over the flow cell surface at an injection volume of 25 µL. The temperature was raised by 1 °C/min from 25 to 100 °C, and fluorescence readings were taken at each step.

X-ray crystallography
All consumables for crystallization and crystal handling were purchased at Molecular Dimensions, Calibre Scientific, Rotherham, UK, unless stated otherwise. CMA1 concentrated at 5.7 or 3.5 mg/mL in 20 mM HEPES pH 8, 100 mM NaCl, and 14 mM GalNAc was subjected to crystallization screening using the robotized HTXlab platform (EMBL, Grenoble, France) with 200 nL sitting drops at 20 °C using a 1:1 ratio. Wizard I and II screen (Rigaku) and SaltRX (Hampton Research) screens were used and led to more than 30 hits after one to three days. Pill-like crystals were obtained with high salt concentration that could be reproduced by hand in the laboratory. Plates and needles clusters were obtained with PEG containing solutions. For CMA1-Nter, protein at a concentration of 2.9–3.5 mg/mL was crystallized using hanging drop and vapor diffusion methods with a 2 µL drop in 1:1 ratio at 20 °C. Bipyramidal single crystals were obtained after one or two days in a solution containing 10–12% PEG Smear Medium, 0.1 M MES pH 6.5, 1× divalent (5 mM of CaCl$_2$, MgCl$_2$, CsCl$_2$, CdCl$_2$, NiCl$_2$, and zinc acetate), or 5 mM CdCl$_2$, and in the presence or not of 5 mM GalNAc. Cocrystals of CMA1-Nter in complex with LacNAc (Galβ1-4GlcNAc, Elicityl, #GLY008) were obtained by the addition of 5 mM LacNAc to the protein solution and incubation at room temperature for 30 min prior to crystallization. For both complexes, single crystals were mounted in a cryoloop after transfer in a cryoprotectant solution, composed of 30% PEG Smear Medium and 5 mM CdCl$_2$, and flash-cooled in liquid nitrogen. Crystal diffraction was evaluated, and data were collected on the Proxima 1 and 2 beamlines at the synchrotron SOLEIL, Saint Aubin, France using an Eiger 16M or 9M detector (Table 1) for LacNAc and GalNAc complexed structures, respectively. XDS and XDSME were used to process the data and all further steps were performed using programs of the CCP4 suite version 8.25–27 [33-35]. The model coordinates predicted by Alphafold [36] Monomer v2.0 for the monomer of CMA1 (A0A1S4E5V9) were trimmed to only include the N-terminal domain (residues 33–159), with all B-factors reset to 15 Å$^2$, to be subsequently used as a search model to solve the structure of CMA1-Nter by molecular replacement using PHASER [37]. Multiple iterations of anisotropic restrained maximum likelihood refinement using REFMAC 5.8 [38] and manual building using Coot [39] were performed.
Hydrogen atoms were added in their riding positions during refinement and 5% of the observations were set aside for cross-validation analysis. Upon inspection of the electron density maps, carbohydrate moieties were introduced and checked using Privateer [40]. The final model was validated using the wwPDB validation server (https://validate-rcsb-1.wwpdb.org). Structure figures were made using PyMol 2.5.7 and ChimeraX 1.6 [31]. The parameters for CH−π interactions were defined as previously reported [41,42].

Supporting Information

Supporting Information File 1
Full array data regarding the binding specificity of CMA1. [https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-20-31-S1.xlsx]

Supporting Information File 2
Additional figures. [https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-20-31-S2.pdf]

Supporting Information File 3
Supplementary glycan microarray document (MIRAGE) for the ICL glycan arrays. [https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-20-31-S3.pdf]

Acknowledgements

We acknowledge support from the Mammalian Protein Expression core facility at the University of Gothenburg via the Protein Production Sweden (PPS) framework as well as the synchrotron SOLEIL (Saint Aubin, France) for access to beamline Proxima 1 and 2 (Proposal Number 20210859) and for the technical support of Pierre Legrand and Martin Savko, respectively. The authors would like to thank Iris Lopez and Federico Musso for their technical help in the expression assays of CMA1-Nter and CMA1 purification trials, respectively, and Wengang Chai for preparing the GAG probes on the ICL array.

Funding

This work was funded by a Branco Weiss Fellowship – Society in Science awarded to D.B., by the Knut and Alice Wallenberg Foundation, and the University of Gothenburg, Sweden as well as support from the GLYCONanoPROBES (CA18132) and INNOGLY (CA18103) COST actions awarded to J.L. This work was further supported by the Protein-Glycan Interaction Resource of the CFG and the National Center for Functional Glycomics (NCFG) at Beth Israel Deaconess Medical Center, Harvard Medical School (supporting grant R24 GM137763). The glycan microarray studies were performed in the Carbohydrate Microarray Facility at the ICL Glycosciences Laboratory, which is supported by Wellcome Trust biomedical resource grants (099197/Z/12/Z, 108430/Z/15/Z, and 218304/Z/19/Z) and partially by the March of Dimes Prematurity research centre grant (22-FY18-82). The sequence-defined glycan microarrays contain many saccharides provided by collaborators whom we thank, as well as members of the Glycosciences Laboratory for their contribution in the establishment of the NGL-based microarray system. This work benefited from access to EMBL HTX lab, which has been supported by iNEXT-Discovery, project number 871037, funded by the Horizon 2020 program of the European Commission.

ORCID® iDs
Nicole Kerekes - https://orcid.org/0000-0001-7065-8092
Antonio Di Maio - https://orcid.org/0000-0002-2740-9098
Ten Feizi - https://orcid.org/0000-0001-6495-0329
Annabelle Varrot - https://orcid.org/0000-0001-6667-8162
Daniel Bojar - https://orcid.org/0000-0002-3008-7851

Data Availability Statement
All generated data here can be found in the Supporting Information. The coordinates of CMA1 in complex with LacNAc (PDB ID 8R8A), https://doi.org/10.2210/pdb8R8A/pdb, and GalNAc (PDB ID 8R8C) https://doi.org/10.2210/pdb8R8C/pdb, have been deposited in the Protein Data Bank (PDB).

Preprint
A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.1101/2023.11.30.569503

References


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/bjoc/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/bjoc.20.31