Conformational phase spaces of N-glycans under the computational microscope

Vom Fachbereich Produktionstechnik der

UNIVERSITÄT BREMEN

zur Erlangung des Grades

Doktor der Ingenieurwissenschaften (Dr.-Ing.)

genehmigte

Dissertation von

Isabell Louise Grothaus

Gutachter*innen:

Prof. Dr.-Ing. Lucio Colombi Ciacchi
Prof. Dr. Elisa Fadda, Maynooth University, Ireland

Tag der mündlichen Prüfung: 12.12.2023

Abstract

Conformational phase spaces of N-glycans under the computational microscope

Glycans have an extremely important influence on all living matter on Earth. For instance, their occurrence in eukaryotic cells as post-translational modifications confers diverse functions to the underlying proteins. However, the so-called ‘sugar code’ that draws the connection between glycan structure and function still remains to be deciphered. Its unraveling is hampered by the vast amount of monosaccharide types, chemical substituents and linkages, resulting in a very large variety of complex glycan configurations. In addition, especially N-glycans are typically more flexible than proteins, a consequence of their many freely rotating torsion angles along the glycosidic bonds. The result is a large set of multiple conformations that can be adopted by each single N-glycan, opening a ‘third dimension’ of the sugar code beyond primary sequence and molecular topology. The question remains to which extent this third dimension is biologically relevant, and if there exist relationships between the sequence, the three dimensional structure, and the function of N-glycans in their various biological environments.

The flexibility of N-glycans pushes experimental approaches to their limits when it comes to structure determination. Techniques like NMR and X-ray crystallography can determine time-averaged structures, but can not resolve structural ensembles of individual glycan conformers. Experimental methods are, however, of great importance for the determination of where and which N-glycans are attached to specific proteins. Computational techniques such as MD simulations can utilize this information to construct molecular glycoprotein models and explore the many accessible glycan conformers in order to predict their function at the molecular level. However, so far it could not be confirmed that MD approaches can correctly describe and sample glycan conformer distributions. Moreover, no adequate reduced representation of the high-dimensional phase space of all possible glycan conformations was available.

These shortcomings have hindered the computational study of N-glycans, especially regarding their ‘third dimension’ functionality and the impact of the latter on proteins. Therefore, this work tied into these current issues, both introducing fundamentally new methodological advances in the field of computational structural glycobiology and applying them to systems of high biological relevance.

We laid the foundation for our progress by introducing a novel and generally applicable naming convention for glycan conformers based on their adopted set of torsion angles. This enabled a clear and IUPAC-conform way of addressing quantitatively the third dimension of the sugar code. The automated assignment of conformer strings is made available to the glycobiology community through the python package GlyCONFORMER, allowing for the analysis and visualization of individual glycan distributions. Further, we developed a new enhanced-sampling MD scheme, overcoming the obstacle of convergence when sampling the full conformational phase space of glycan structures. The satisfactory performance of this methodological workflow was first demonstrated for free N-glycans in solution, comparing different glycan configurations and validating different force field parameter sets.

The workflow was then successfully applied to glycoprotein systems, unraveling the impact of glycan conformations on protein behavior. In particular, a pathogenic enzyme of the class trans-sialidase was investigated, studying how surface N-glycosylation and its dynamics can influence enzyme stability and activity. A recombinant trans-sialidase from the African parasite *Trypanosoma congolense* was expressed in CHO Lec1 cells, reproducing the natively expected high-mannose type N-glycans. MALDI-TOF MS experiments determined the N-glycosylation pattern for eight N-glycosylation sites. Removal of these N-glycans via Endoglycosidase H treatment revealed no change in secondary struture in circular dichroism measurements, but led to a decrease in substrate affinity relative to the untreated enzyme, without an impact on the conversion rate. MD simulations could provide mechanistic insights into interactions between the highly flexible N-glycans and some conserved amino acids located at the catalytic site. These interactions led to conformational changes, possibly enhancing substrate accessibility and enzyme-substrate complex stability. Sequence alignments further revealed the conservation of N-glycosylation sequons among different *Trypanosoma* species, pointing to a newly discovered, conserved glycan-mediated enzymatic regulation mechanism, adding a new entry to the sugar code dictionary. Further analysis of the conformational phase spaces of interacting glycans revealed a shift in their conformer distributions, underlining the importance of their structural flexibility.

In addition, interactions of glycans within the catalytic site of the carbohydrate-active enzyme α-mannosidase II were examined, studying how the boat conformation of the to be cleaved saccharide unit is induced in the transition state of the catalytic cleavage reaction. We could identify a large shift in the conformational phase space upon binding of the glycan to the catalytic site. However, this was not sufficient to induce a change in pucker conformation. In this system, a mutual dependence between torsional and pucker degrees of freedom could be excluded. The structural rearrangement is likely induced by a shift of the electron density in the sugar ring induced by the interplay of a binding Zn^{2+} ion and the surrounding amino acids of the binding pocket.

Unraveling hidden correlations between a glycan's structure and its function will enable future studies of how glycan structures drive important biological processes, such as disease mechanisms or enzyme-controlled signaling. For instance, our methodological advances could already validate the performance of the GlycoSHIELD software, able to graft realistic N-glycans on any glycoprotein, e.g. the SARS-CoV-2 Spike protein. Further, structural features of other glycan types, such as the polysaccharide fucoidan, could be examined, revealing previously unknown structural rearrangements upon chemical modifications. Although the limitations of the fixed-charge force fields used in this work remain, we are now able to quantify them and pave the way for improved parameterizations. This is especially necessary because there still exists a lack of polarizable or machine-learning-based force fields for carbohydrates. All the studied systems revealed new facets of the postulated 'third dimension' of the sugar code, providing biochemically relevant examples of how the code's dictionary can be deciphered by means of our newly developed MD methodology.

Zusammenfassung

Untersuchung der konformationellen Phasenräume von N-glykanen unter dem virtuellen Mikroskop

Glykane haben einen äußerst wichtigen Einfluss auf alle Lebewesen unserer Erde. So verleiht ihr Vorkommen in eukaryotischen Zellen als posttranslationale Modifikation einer Vielzahl von Proteinen verschiedene Funktionen. Der so genannte „Zuckercode“, welcher die Verbindung zwischen Glykanstruktur und -funktion herstellt, wurde allerings noch nicht entziffert. Seine Entschlüsselung wird durch die hohe Anzahl an Monosacchariden, chemischen Substituenten und Bindungen erschwert, was zu einer enormen Vielfalt komplexer Glykan-Konfigurationen führt. Darüber hinaus sind insbesondere N-Glykane in der Regel flexibler als Proteine, was auf die vielen frei drehbaren Torsionswinkel entlang der glykosidischen Bindungen zurückzuführen ist. Das Ergebnis ist eine Vielzahl von Konformationen für ein einziges N-Glykan, was eine „dritte Dimension“ des Zuckercodes jenseits von Primärsequenz und molekularer Topologie eröffnet. Es bleibt allerdings die Frage, inwieweit diese dritte Dimension biologisch relevant ist und ob es einen Zusammenhang zwischen der Sequenz, der dreidimensionalen Struktur und der Funktion von N-Glykanen in ihren verschiedenen biologischen Umgebungen gibt.

Die Flexibilität von N-Glykanen bringt experimentelle Ansätze bei der Strukturbestimmung an ihre Grenzen. Techniken wie NMR und Röntgenkristallographie können zwar zeitlich gemittelte Strukturen bestimmen, aber keine strukturellen Ensembles einzelner Glykan-Konformer auflösen. Nichtsdestotrotz sind experimentelle Methoden von großer Bedeutung, um zu bestimmen wo und welche N-Glykane an bestimmte Proteine gebunden sind. Computertechniken wie MD-Simulationen können diese Informationen nutzen, um die Funktion von N-Glykanen auf molekularer Ebene vorherzusagen, indem eine Vielzahl von unterschiedlichen Glykan-Konformern für eine modellierte Glykoproteinstruktur erzeugt wird. Bislang konnte jedoch nicht bestätigt werden, dass MD-Ansätze die Verteilung der Glykan-Konformer korrekt beschreiben und erfassen können. Darüber hinaus war keine reduzierte Darstellung des hochdimensionalen Phasenraums aller möglichen Glykan-Konformationen verfügbar, welche diesen adäquat wiederspiegeln würde.

Diese Unzulänglichkeiten haben die computergestütze Untersuchung von N-Glykanen bisher verhindert, insbesondere im Hinblick auf die Funktionalität ihrer „dritten Dimension“ und deren Auswirkungen auf Proteine. Daher knüpft diese Arbeit an die aktuelle Problematik an, indem sie einerseits grundlegend neue methodologische Fortschritte auf dem Gebiet der computergestützten, strukturellen Glykobiologie einführt und andererseits auf Systeme von hoher biologischer Relevanz anwendet.

Wir legten den Grundstein für unsere Fortschritte mit der Einführung einer neuartigen und allgemein anwendbaren Namenskonvention für Glykan-Konformer, basierend auf den Werten ihrer Torsionswinkel. Dies ermöglichte eine quantitative, eindeutige und IUPAC-konforme Darstellung der dritte Dimension des Zuckercodes. Die automatisierte Zuweisung von Konformer-Strings wird der Glykobiologie-Community durch das Python-Paket GlyCONFORMER zur Verfügung gestellt und ermöglicht die Analyse und Visualisierung individueller Glykanverteilungen. Darüber hinaus haben wir ein neues MD-Verfahren entwickelt, welches das Konvergenzproblem von Glykanstrukturen überwindet, indem der gesamten konformationellen Phasenraums exploriert und abgedeckt werden kann. Die zufriedenstellende Leistung dieses methodischen Arbeitsablaufs wurde zunächst für freie N-Glykane in Lösung demonstriert, wobei verschiedene Glykan-Konfigurationen verglichen und verschiedene Kraftfeldparametersätze validiert wurden.

Die entwickelte Methodik wurde dann erfolgreich auf Glykoproteinsysteme angewandt, um die Auswirkungen der Glykan-Konformationen auf das Proteinverhalten zu entschlüsseln. Insbesondere wurde ein pathogenes Enzym aus der Klasse der Trans-Sialidasen untersucht, um herauszufinden, wie die Oberflächen-N-Glykosylierung und ihre Dynamik die Stabilität und Aktivität des Enzyms beeinflussen können. Eine rekombinante Trans-Sialidase aus dem afrikanischen Parasiten *Trypanosoma congolense* wurde in CHO Lec1-Zellen exprimiert und reproduzierte die nativ erwarteten N-Glykane mit hohem Mannosegehalt. In MALDI-TOF MS-Experimenten wurde das N-Glykosylierungsmuster für acht N-Glykosylierungsstellen bestimmt. Die Entfernung dieser N-Glykane durch die Behandlung mit Endoglykosidase H ergab keine Veränderung der Sekundärstruktur in Circular Dichoism Messungen, führte jedoch zu einer Verringerung der Substrataffinität im Vergleich zum unbehandelten Enzym, ohne Auswirkungen auf die Umwandlungsrate. MD-Simulationen konnten mechanistische Erkenntnisse über die Wechselwirkungen zwischen den hochflexiblen N-Glykanen und einigen konservierten Aminosäuren an der katalytischen Stelle liefern. Diese Wechselwirkungen führten zu Konformationsänderungen, die möglicherweise die Zugänglichkeit des Substrats und die Stabilität des Enzym-Substrat-Komplexes verbesserten. Sequenzalignments zeigten darüber hinaus, dass Glykosylierungssequenzen bei verschiedenen *Trypanosoma*-Spezies konserviert sind, was auf einen neu entdeckten Glykan-vermittelten enzymatischen Regulierungsmechanismus hindeutet. Die Analyse der Konformationsphasenräume interagierender Glykane ergab eine Verschiebung ihrer Konformerverteilungen, was die Bedeutung ihrer strukturellen Flexibilität unterstreicht. Darüber hinaus wurden die Wechselwirkungen von Glykanen innerhalb der katalytischen Bindestelle des kohlenhydrataktiven Enzyms α-Mannosidase II untersucht, wobei herausgefunden werden sollte, wie die Bootkonformation der zu spaltenden Saccharideinheit im Übergangszustand der katalytischen Spaltungsreaktion induziert wird. Wir konnten eine große Verschiebung im Konformationsphasenraum nach der Bindung des Glykans an die katalytische Stelle feststellen. Dies reichte jedoch nicht aus, um eine Änderung der Pucker-Konformation zu bewirken. Zumindest für dieses System konnte eine gegenseitige Abhängigkeit zwischen Torsions- und Pucker-Freiheitsgraden ausgeschlossen werden. Die strukturelle Umordnung wird wahrscheinlich durch eine Verschiebung der Elektronendichte im Zuckerring ausgelöst, die durch das Zusammenspiel eines bindenden Zn^{2+}-Ions und der umgebenden Aminosäuren der Bindungstasche hervorgerufen wird.

Die Entschlüsselung verborgener Korrelationen zwischen der Struktur eines Glykans und seiner Funktion wird künftige Studien darüber ermöglichen, wie Glykanstrukturen wichtige biologische Prozesse steuern, etwa Krankheitsmechanismen oder enzymgesteuerte Signalübertragungen. So konnten unsere methodischen Fortschritte bereits die Leistung der GlycoSHIELD-Software validieren, die in der Lage ist, realistische N-Glykane auf jedes beliebige Glykoprotein, z. B. das SARS-CoV-2-Spike-Protein, zu modellieren. Darüber hinaus konnten die strukturellen Merkmale anderer Glykanarten, wie z. B. des Polysaccharids Fucoidan, untersucht werden, um bisher unbekannte strukturelle Umlagerungen nach chemischen Modifikationen aufzudecken. Obwohl die Beschränkungen der in dieser Arbeit verwendeten Kraftfelder bestehen bleiben, sind wir nun in der Lage, diese zu quantifizieren und den Weg für verbesserte Parametrisierungen zu ebnen. Dies ist vor allem deshalb notwendig, weil es immer noch einen Mangel an polarisierbaren oder auf maschinellem Lernen basierenden Kraftfeldern für Kohlenhydrate gibt. Alle untersuchten Systeme zeigten neue Facetten der postulierten dritten Dimension des Zuckercodes und lieferten biochemisch relevante Beispiele dafür, wie das Wörterbuch des Zuckercodes mit Hilfe unserer neu entwickelten MD-Methodik entschlüsselt werden kann.

For those who can’t be here.

Karl
Helga
Rainer

Acknowledgements

I consider my PhD years as an ultramarathon with great pacemakers who enabled me to climb scientific peaks, and a wonderful supportive crew who pulled me out of valleys. It is an endurance run that you can hardly do alone, but need many experienced companions.

Therefore, I would first like to thank Sørge Kelm, Mario Waespy and Rita Groß-Hardt for their confidence in my abilities and their support at the beginning of my PhD in finding an exciting scientific topic. A special thanks goes to my colleague Jana Rosenau, whose continuous support and collaboration was very important for the experimental results in this dissertation. I would also like to thank the former members of the AG Kelm for taking me in as a student of their own. I am grateful for the wonderful HMI working group that I am privileged to be a part of. We survived the Covid pandemic together and I want to thank you for being great colleagues one can expierence real adventures with: Pia Götz, Maria von Einem, Eric Macke, Aparna Malisetty, Lorenzo Bastonero, Wilke Dononelli, Yendry Corrales Ureña, Saeed Amiri, Krishnanjan Pramanik, Wolf-Achim Kahl, Massimo Delle Piane, Filippo Balzaretti and Susan Köppen. My personal note to Britta Hinz: "Meine Güte siehst du gut aus"! Many thanks to Stefan Schmidt for the technical support.

Special mention goes to Giovanni Bussi from the SISSA in Trieste, Italy, who contributed significantly to the success of this dissertation and provided excellent scientific supervision. I would also like to thank his group members Thorben Fröhlking, Mattia Bernetti, Valerio Piomponi, and Vittorio Del Tatto for the warm welcome in Italy and the never-ending discussions on Zoom.

I am most grateful for the eye-to-eye supervision I received during the last years from my PhD supervisor Lucio Colombi Ciacchi. Thank you for the chance to prove myself in the field of computational biophysics, and for being as fascinated by sugars as I am. I am very glad that with you it is possible not only to study free energy profiles, but also the heights and depths of mountain ranges.

However, a supply of love, food and happiness is also essential. So I have to thank my friends and especially Tine for believing in the importance of my work and continuing to motivate me. I want to thank my parents deeply for giving me the confidence to go my own way. Even though my field of work is so far from your knowledge, you have always motivated me to keep going and achieve my goals. I really appreciate your continuous support and am very happy to call you mom and dad.

The person who has suffered the most during this time is my wife, Jana. Thank you for your patience, the freedom you gave me to live my dream, and your enthusiasm for my research. I am very proud of our little family that we built during this time. Thank you for giving us our son Carl.

Preface

Results obtained during my time as a Ph.D. student from November 2019 to June 2023 are summarized in this dissertation, although results have partly been published *a priori* as follows:

> Chapter 2: S.M. Ayala Mariscal, M.L.Pigazzini, Y. Richter, M. Özel, **I.L. Grothaus**, J. Protze, K. Ziege, M. Kulke, M. ElBediwi, J. Vermaas, L. Colombi Ciacchi, S. Köppen, F. Liu, J. Kirstein, Identification of a HTT-specific binding motif in DNAJB1 essential for suppression and disaggregation of HTT, *Nature Communications* 13, 4692, 2022.
>
> Chapter 3: Y. Tsai, N. Chang, K. Reuter, H. Chang, T. Yang, S., N. Zerrouki, M. Gecht, C. Penet, **I. L. Grothaus**, L. Colombi Ciacchi, K. Khoo, G. Hummer, S. D. Hsu, C. Hanus, M. Sikora, Rapid simulation of glycoprotein structures by grafting and steric exclusion of glycan conformer libraries, *Cell*, under revision.
>
> Chapter 3: **I.L. Grothaus**, G. Bussi, L. Colombi Ciacchi, Exploration, representation and rationalization of the conformational phase-space of N-glycans, *Journal of Chemical Information and Modelling*, 62(20), 4992–5008, 2022.
>
> Chapter 4: J. Rosenau*, **I. L. Grothaus***, Y. Yang, N. D. Kumar, L.C. Ciacchi, S. Kelm, and M. Waespy. N-glycosylation modulates enzymatic activity of Trypanosoma congolense trans-sialidase, *Journal of Biological Chemistry*, 298, 102403, 2022.

Following results published in collaboration with other scientific research groups had a different thematic focus and are therefore not explicitly discussed in this dissertation:

> C. Arend*, **I.L. Grothaus***, M. Waespy, L. Colombi Ciacchi, R. Dringen, Modulation of Multidrug Resistance Protein 1-mediated transport processes by the antiviral drug ritonavir, *Neurochemical Research, 49, 66–84, 2024*

Results obtained or analyzed not by myself are highlighted in the individual chapters of this dissertation.

* Shared co-first authorship

Contents

1 | The uprising of carbohydrates

1.1 The sugar code

The discovery of the double helix and its four unique nucleotides of which our genes are composed heralded the start of the era of coding in biology. It was in 1953 when James Watson and Francis Crick, with the help of Rosalind Franklin and Maurice Wilkins, deciphered the genetic code and were the first to show that biological macromolecules are capable of storing information based on the arrangement of individual monomers in programmed order.[1–4] The elucidated deoxyribonucleic acid (DNA) was considered as the first alphabet of life, with its four nucleotides (letters) giving rise to different genes (words). The 20 amino acids occurring in proteins were considered the second alphabet, being linked to the genetic code and dependent on it.[5] In earlier times the story came to a halt at this point, as no further logical concepts could be identified in other biomolecular classes. It would have been, however, detrimental to consider only those concepts, where the information stored in a sequence is translated directly into another sequence. It would have prevented us from exploring more complex and hidden forms of coding capability, involving additional components like decoders to decipher the code into signals. Fortunately, 40 years after the discovery of the genetic code, carbohydrates entered the scene when proteins were found to specifically recognize sugars like ligand molecules[6], suggesting a decoding capability.[7] In order to test the carbohydrate information-storing potential, the number of possible words (peptides or oligosaccharides) that are possible to spell from a predefined set of letters (amino acids, monosaccharides) can be compared. Considering 20 different amino acids or monosaccharides to form a hexapeptide or hexasaccharide respectively, one gets eight orders of magnitude less peptides ($6.4 * 10^7$) than oligosaccharide structures ($1.44 * 10^{15}$). This reflects the huge size of vocabulary one can achieve with polysaccharides.[7,8]

To be able to appreciate this comparison, one has to understand the atomic buildup of carbohydrates and their fundamental difference compared to nucleic acids and peptides. Monosaccharides are the most simple form of carbohydrates, representing the building blocks for oligosaccharides (composed by less than ten monomers)- and larger polysaccharides, called glycans. Without substituents, they present the chemical formula $\mathrm{C_nH_{2n}O_n}$, where n is larger than three and defines the class of sugar. They consist of several polyhydroxyl groups with an aldehyde or ketone group, whereby we only consider the aldehyde in this dissertation. Monosaccharides can interchange between their open-chain or cyclic form through a reversible nucleophilic addition reaction, whereby the cyclic form is predominant both in solution and in the solid state, and is mandatory to build up oligosaccharides (Figure 1.1).[9] A hemiacetal group forms at carbon C1, called the reducing end.

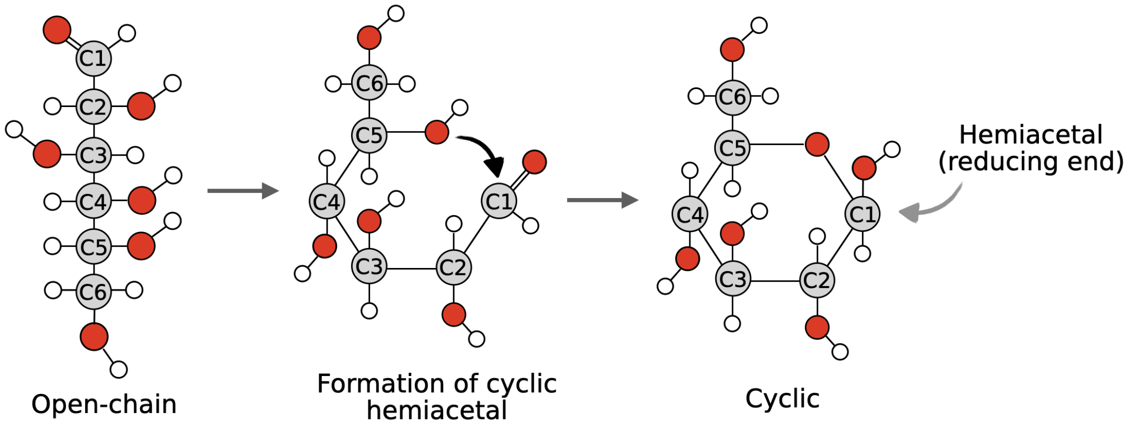


Figure 1.1: **D-glucose in its transition from an open-chain to a cyclic form.** The reversible rearrangement is occurring through the formation of a cyclic hemiacetal. Carbon atoms are represented in gray, oxygen in red and hydrogen in white.

Next to the well known five- and six-carbon sugars, a special sugar class, the sialic acids, consist of a nine-carbon backbone and should be mentioned to stress the diversity of possible carbon atom numbers in monosaccharides (compare Figure 1.2 (i) neuraminic acid).[10] Further, there are different monosaccharide types having an identical number of carbon atoms, but vary in the configuration around their chiral centers (carbon atom attached to four different groups). These different configurations are called isomers. For instance, D-glucose and D-mannose are a special pair of isomers, named epimers. They are called epimeric at carbon C2, as their structures are identical except for the orientation of the hydroxyl group at C2 (Figure 1.2 (ii)).[9] Another form of isomerism is denoted through the prefixes D- and L-, indicating stereoisomers that are mirror-images of themselves, altered in the configuration at C5 for hexoses in the open-chain form (compare the hydroxyl group orientations of L-fucose and D-mannose in Figure 1.2). As almost all sugars considered in this thesis exist in the D-configuration, unless specified differently, the D- prefixes will not be mentioned specifically from here on. Independently of their form, monosaccharides can further be modified by chemical groups like N-acetyl or N-glycolyl, generating derivatives of monosaccharides like N-acetyl glucosamine or neuraminic acid (Figure 1.2 (iii)). Before continuing with the assembly of monosaccharides into chains, it should be noted that in their cyclic form, the C1 carbon atom represents an additional chiral center (anomeric carbon). This gives rise to two possible configurations, termed $\alpha-$ and $\beta-$ anomer, depending on the orientation of the hydroxyl group at C1 and the stereogenic center furthest from the anomeric carbon (C5) (Figure 1.2 (iv)).[9] This characteristic is important in the linkage formation process, as the configuration at C1 defines if the linkage is termed $\alpha-$ or $\beta-$. Glycosidic linkages are enzymatically formed in a condensation reaction, linking two monosaccharide residues via an oxygen atom. In contrast to peptide bonds, any carbon atom of the monosaccharide ring connected to a hydroxyl group can form a glycosidic bond, generating a high degree of linkage possibilities. There is only one restriction, namely that the linkage be formed between the anomeric carbon at the reducing end and any other hydroxyl group, for example in the case of lactose, which is composed of the monosaccharides galactose and glucose, or glycogen, a polysaccharide of several glucose subunits (Figure 1.2 (v)).[9] Due to this, oligosaccharides gain a polarity that is defined by their reducing and nonreducing ends, similar to the amino and carboxyl termini in polypeptides or the 5' and 3' termini in polynucleotides.[9] Further diversification of oligosaccharides is achieved via the introduction of branches, meaning that a sugar is involved in three or more glycosidic linkages, for examples via its C1, C3 and C6 atoms. Important biopolymers like glycogen or N-glycans rely on this branching principle (Figure 1.2 (vi)).

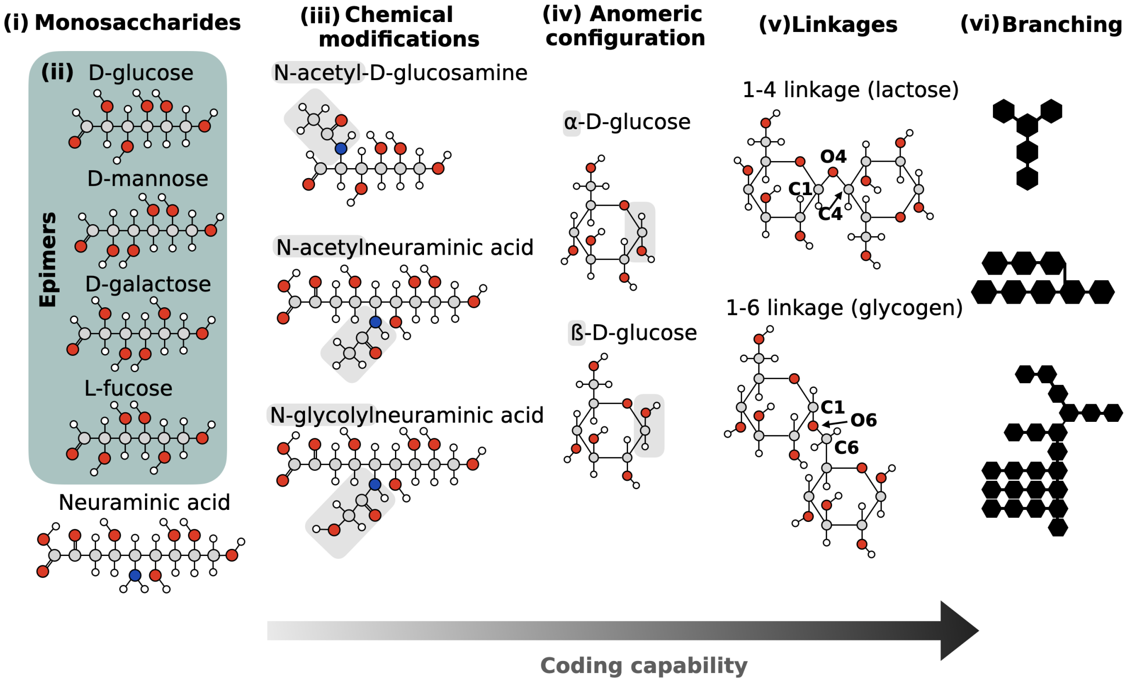


Figure 1.2: **The coding capability of carbohydrates.** Different monosaccharide types are defined by their number of carbon atoms and orientation of hydroxyl groups. They can be further diversified by the addition of chemical groups and consideration of their anomeric configuration. Each monosaccharide can be chained to another one by the formation of glycosidic bonds, formed between the reducing end and any other carbon atom connect to a hydroxyl group. The formation of chains can become even more complex when introducing branches. Carbon atoms are represented in gray, oxygen in red, nitrogen in blue and hydrogen in white.

To summarize the above mentioned structural features, carbohydrates can vary in (i) their carbon number; (ii) their epimers; (iii) chemical modifications; (iv) anomeric configuration; (v) linkage position and (vi) branching position (Figure 1.2).[7] These characteristics allow for more diverse configurations than achievable with nucleotides or amino acids, which can only be linked in one hard coded fashion by either phosphodiester or peptide bonds. The explicit informational properties harbored by glycans were already mentioned by Winterburn and Phelps in the 1970s.[11] However, it took sugars much longer to be considered as the third alphabet of life.[5] The ability to decode the complex glycome (generic term for the entire complement of carbohydrates in an organism) was missing. The story of the decoding machinery had its beginning in the 1930s when Sumner and Howell[12] discovered the sugar specificity of lectins, a class of carbohydrate-binding proteins. Lectins lack any enzymatic activity, are distinct from antibodies and encode sugars without acting on them.[13] The precise recognition however, similar to a key-and-lock mechanism[13], became apparent much later when three-dimensional structures were accessible and structural motifs, chemical groups and type of bonds feasible to analyze.[6] Nowadays, at least

13 different protein folds are known for glycan recognition, assumed to result from the evolutionary adaptation of the protein surface to the ligand structure, where lectins can especially bind larger carbohydrate structures with anomeric extension.[13] Although there are countless different carbohydrate structures, most lectins recognize only a single carbohydrate type like a single epimer structure, realized by a certain special arrangement of amino acids in the binding site.[13] Adhesion of glycans to lectins can trigger diverse signaling pathways in cells, also including those with severe consequences like cell cycle arrest or induction of apoptosis.[13]

In summary, the flow of biological information starting with nucleotides decoding amino acid sequences does not stop after the translation of proteins, but is rather the starting point for enzymes to generate a new *glycan* code, from the third alphabet of life, the monosaccharides.[5] Unraveling the functional meaning and cellular responses of certain glycan words is comparable to the compilation of a dictionary for the glycan vocabulary and has only begun in recent years to be put together.[14] If one wants to understand the fundamentals and far-reaching implications of the sugar code, one must first shed light on the ubiquitous occurrence of glycans inside and outside of cells as well as their routes of synthesis.

1.2 Glycan types and their functionality

In 1861, it was shown by Butlerow that formaldehyde in alkaline solution allowed for the synthesis of molecules like fructose, which could be further converted to glucose and mannose by condensation and rearrangement reactions.[14–16] These prebiotic reaction conditions could be a hint that carbohydrates were utilized already by ancestral primitive microorganisms, existing around three billion years ago.[16] Nowadays, it is a fact that carbohydrates can be found in all three domains of life, indicated by the presence of carbohydrate-active enzymes (CAZymes). The enzyme class comprises glycoside hydrolases, polysaccharide lyases and glycosyltransferases, modifying the glycome at different levels of complexity. The comprehensive existence of CAZymes is underscored by the fact that they account for about 1-3 % of the genome of most organisms.[17] This tremendous enzyme machinery creates many diverse types of glycans, which are particularly different comparing prokaryotic and eukaryotic cells, although the focus will be on the latter.

Carbohydates occur seldom as monosaccharides except for their role as a source of energy. They are found more often as building blocks to form more complex glycan structures. One needs to clearly differentiate between free glycans and those that are conjugated to biomolecules. Freely occuring polysaccharides undertake the task of mechanical support, like cellulose in plants, or represent a long-term energy storage as amylose or amylopectin. Different linkage possibilities and anomeric configurations take effect when comparing the types of monosaccharides present in the above mentioned glycans. They all exclusively consist of repeating units of glucose, however, either linked via $\beta 1 \rightarrow 4$ or $\alpha 1 \rightarrow 4$ bonds in the case of cellulose and amylose, respectively. When amylose strands are additionally cross-linked via $\alpha 1 \rightarrow 6$ bonds, the more complex amylopectin is generated. The chemical modification of glucose by N-acetylation to form N-acetylglucosamine bridges the gap from cellulose in plants to chitin, providing mechanical support in the exoskeleton of insects having the same $\beta 1 \rightarrow 4$ bond type. It should have become clear that these non-conjugated carbohydrates predominantly consist of repeating units of the same monosaccharide type and linkage, not forming a good basis for diverse glycan structures to build a complex glycan vocabulary.

Therefore, more attention should be drawn to conjugated carbohydrates that can be found inside and outside of eukaryotic cells. Conjugated glycans are so called ‘glycosides’, generated through the formation of a linkage between a monosaccharide and an aglycone (organic molecule without sugar residues). The primary sugar unit is then further processed to yield a complex glycan with diverse monosaccharide and linkage types. One can roughly divide conjugated glycans into classes, depending mostly on their monosaccharide-aglycone linkage but also cellular location.

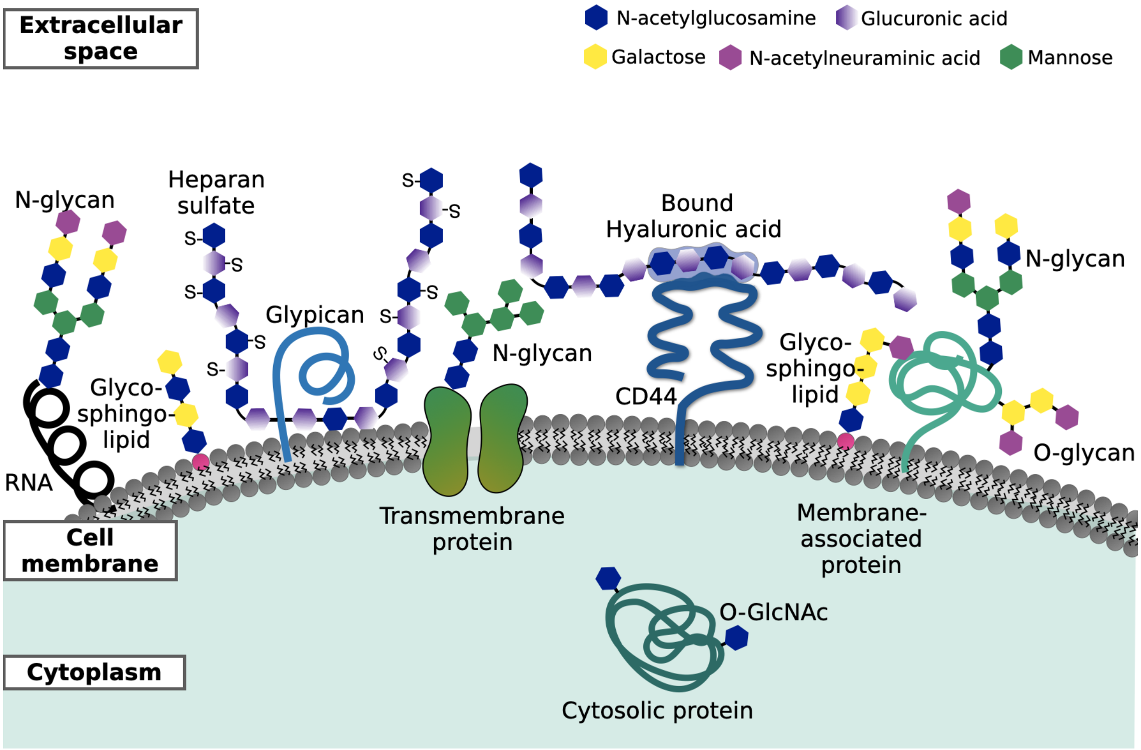


Figure 1.3: **Diversity of glycans inside and outside of cells.** The extracellular glycocalyx consists of glycolipids like glycosphingolipids and proteoglycans like glypican or CD44 with their respective glycosaminoglycans attached, either heparan sulfate or hyaluronic acid. Additionally, the majority of transmembrane or membrane-associated proteins is glycosylated by N- and O-glycans. N-glycosylations were also detected for extracellularly detected RNA. Intracellularly, only small and most often O-GlcNAcs can be found attached to proteins. Monosaccharides are represented as hexagons and colored according to the Symbol Nomenclature For Glycans (SNFG).[18]

Starting at the extracellular site, it can be recognized that every eukaryotic cell is coated by a dense and complex layer of glycans, called the ‘glycocalyx’.[19] It consists mostly of glycosides like glycoproteins, proteoglycans with glycosaminoglycans and glycolipids, although also free glycans do occur (Figure 1.3).[20] In 2021, also ribonucleic acid (RNA) was discovered to harbor N-glycans and being displayed on cell surfaces. However, nothing is yet known about the precise attachment sites or their detailed function.[21] One has to imagine, when looking at a cell from a distance, that there are probably no phospholipids or membrane proteins to be seen, but a hairy undergrowth of various glycan structures. This is due to the size of the glycocalyx spanning 0.5 to 5 µm of the extracellular space around human cells, varying depending on the cell type, organ location and vascular flow.[22–25] Therefore, it appears as if cells would wear a sugar code to depict information on their surface for communication and signaling with its surrounding. Predicting the function of the glycocalyx as an interface between the cell and the extracellular space was troublesome in the past, if only because of its complexity. However, it could be shown that the glycocalyx serves as a physical protective barrier against pathogens, as a mechanosensor for endothelial cells in the blood stream, as a storage compartment. It also influences cell morphology, membrane organization, and cancer progression to only mention a few of its functions.[20]

To shed light on the huddle the glycocalyx represents, the individual conjugated glycan classes should be briefly entangled. First of all, the many different types of phospholipids making up the lipid bilayer of cell membranes are accompanied by glycolipids, where glycosphingolipids (GSLs) are the most abundant subclass in vertebrates.[26] GSLs are formed by at least one monosaccharide which is linked to a ceramide molecule, consisting of a sphingoid base (long-chain aliphatic amine) and a fatty acid moiety. Sphingosine is the most common sphingoid base in mammals and the fatty acid component in ceramindes can vary in length and saturation level. The primary sugar residue (mostly β-linked glucose or galactose) can be further elongated by glycosyltransferases prior to the GSLs being displayed at the outer leaflet of the plasma membrane (Figure 1.3).[27] They are diversified by different glycan structures, whereby more than 400 different ones built from twelve monosaccharide types could be detected, forming constructs involving up to 20 residues.[27,28] GSLs can make up a significant proportion of total lipids in membranes such as the myelin of axons or erythrocytes.[29,30] They are responsible for cell-cell adhesion via *trans* carbohydrate-carbohydrate interactions often mediated by divalent cations, and modulate apoptosis, cell proliferation and intracellular transport.[27,31–34]

Next to lipid components, membranes also consist of transmembrane or membrane-associated proteins. Many of these are glycosylated via co-translational or post-translational modification processes, resulting in the attachment of sugar or glycan molecules to the polypeptide chain. In comparison to other types of protein post-translational modifications like phosphorylation, acetylation or methylation, glycosylation tops all of them in terms of size and complexity, and represents an ubiquitously occurring modification.[35] A special class with enormously large glycans are the proteoglycans, which consist of a protein 'core' and attached glycosaminoglycan chains. Glycosaminoglycans (GAGs) are huge, linear polysaccharides, where a number of 80 monosaccharides per glycan chain is not unusual. They consist of repeated disaccharide units, comprising an amino sugar (e.g. N-acetylglucosamine or N-acetylgalactosamine) and an uronic acid (glucuronic acid or iduronic acid).[36] Common GAGs are hyaluronic acid, keratan sulfates, heparins and heparan sulfates, which can be further modified by the addition of sulfate groups, by fucosylation or sialylation (Figure 1.3).[36] Hyaluronic acid is not directly linked to the protein core but rather servers as a ligand, where all the latter mentioned ones are covalently linked to asparagine, serine or threonine residues. There are only a little over 50 proteoglycans known, whereby many are also secreted into the extracellular matrix (ECM), promoting its assembly and mechanics as well as modulating force transmission.[37,38] One transmembrane proteoglycan example is the CD44 receptor, whose expression is upregulated in cancer cells and promotes migration and invasion processes in metastases.[39] Its hyaluronic acid ligand, upon binding, can activate matrix metalloproteinases and cytoskeleton signaling involved in tumor progression (Figure 1.3).[39,40] Proteoglycans can have several GAGs attached, whereby the amount is not fixed but can vary over time depending on factors like cell type, enzyme availability and environmental conditions.

Glycoproteins are characterized by their more complex type of glycans. However, similarly to proteoglycans, they also occur on the cell surface in the form of transmembrane or membrane-associated proteins. They differ from proteoglycans in the sense that their glycans are much smaller in size and have different structural features (Figure 1.3). However, they share the commonality of glycosidic linkage types, which can not only be formed between sugars but also to hydroxy amino acids such as serine and threonine (O-glycosidic linkage) or asparagine (N-glycosidic linkage).[9] Protein glycosylation is recognized throughout the whole phylogenetic tree, where a total of 13 monosaccharides and 8 amino acids forming over 41 types of glycosidic linkage types.[13,41] In contrast to the very few identified proteoglycans, it is remarkable that over half of all eukaryotic proteins are glycoproteins and that around 90 % of those are N-glycosylated.[42] Due to this excess, N-glycans deserve to be surveyed with more attention. Well-known examples comprise the immunoglobulin IgG, where the glycosylation pattern determines whether an antibody glycoform is pro-inflammatory, containing galactose-deficient N-glycans, or anti-inflammatory, when harboring sialylated N-glycans.[43] This ubiquitous glycosylation form is derived from the covalent tethering of glycans to the polypeptide chain via the terminal NH_2 group of asparagine residues ($\text{Glycan}_{\beta 1\rightarrow N}$). The enzyme-regulated attachment requires the occurrence of asparagine in the amino acid sequence motif N-X-S/T, where X can be any amino acid except proline in order to be recognized for glycosylation (Figure 1.4). There are three structurally different N-glycan types: high mannose, complex and hybrid, whereas the latter one is a mix of the two previous ones. They are all sharing the same oligosaccharide core structure consisting of two N-acetylglucosamine (GlcNAc) residues followed by three branched mannoses (Man): $\text{Man}_{\alpha 1\rightarrow 6}$ $[\text{Man}_{\alpha 1\rightarrow 3}]$ $\text{Man}_{\beta 1\rightarrow 4}$ $\text{GlcNAc}_{\beta 1\rightarrow 4}$ $\text{GlcNAc}_{\beta 1\rightarrow N}$ (Figure 1.4).[44] It is a biantennary glycan with the first mannose residue serving as a branching point for $\alpha 1 \rightarrow 3$ and $\alpha 1 \rightarrow 6$ linkages. The reason for the conservation of this specific core structure is not yet unraveled. Only speculations about its origin can be made, which hint in the direction of intrinsic benefits of the $\text{Man}_{\beta 1\rightarrow 4}\text{GlcNAc}_{\beta 1\rightarrow 4}\text{GlcNAc}_{\beta 1}$ structure for the folding energetics of the underlying protein.[45] Common monosaccharide types are galactose (Gal), L-fucose (Fuc) and sialic acids (Sia) like N-acetylneuraminic acid (Neu5Ac) or N-glycolylneuraminic acid (Neu5Gc), next to Man and GlcNAc. N-glycans can not only be branched in its core but also several times further along the branches, forming tree-like structures. Phosphorylation, acetylation or sulfation are modifications that are typical for proteins, but also confer another level of complexity to glycans after their synthesis, although this fine-tuning of the vocabulary is common to all three alphabets of life.[14] Many different N-glycans can be built from these various structural components, for example over 100 different configurations could be identified in the nematode *Caenorhabditis elegans.*[46] The number of glycans attached to a protein is not only defined by the number of N-X-S/T motifs, but also by the protein conformation. This affects the substrate recognition ability of required CAZymes for processing, enzyme availability and nucleotide sugar metabolism.[44,47] A probable resulting site-specific heterogeneity creates temporally and spatially flexible glycosylation patterns on a protein, further increasing the size of the glycoproteome. There is a differentiation between micro- and macroheterogeneity, where micro heterogeneity corresponds to the variation of glycan structures at a specific glycosylation site and macro heterogeneity defines the site occupancy of the whole protein.[48]

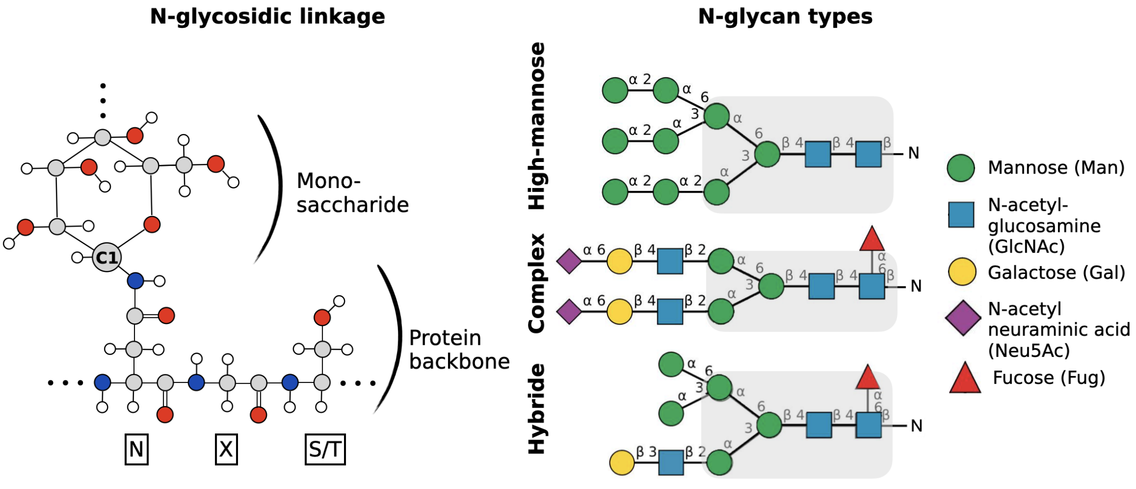


Figure 1.4: **Molecular details of N-glycosidic bonds and different N-glycan types.** N-glycans arise through a covalent bond between the side chain of an asparagine (N) residue and a monosaccharide, as long as the asparagine occurs in the amino acid motif N-X-S/T, where X can be any amino acid except proline, and serine (S) or threonine (T). The glycosidic bond is formed via a nitrogen atom and is therefore termed N-glycosidic linkage. There are different forms of N-glycans, the most dominant ones being the high-mannose type and complex type glycans. The conserved core consisting of five monosaccharides is highlighted in light gray. Carbon atoms are represented in gray, oxygen in red, nitrogen in blue and hydrogen in white. Monosaccharide symbols are used according to the Symbol Nomenclature For Glycans (SNFG) in congruence with the nomenclature of the Consortium for Functional Glycomics.[49] The N-gycan models were drawn using DrawGlycan.[50]

O-linked glycans should also be mentioned. They have a higher diversity compared to N-linked glycans since many sugar core structures are possible and the number of sugar residues in a chain vary from one to many. Glycosidic bonds can be formed between the hydroxyl group of serine or threonine residues to either Fuc, Man, Gal, GlcNAc, glucose (Glc) or xylose (Xyl). Without the necessity of a recognition sequence for the enzymatic attachment of O-glycans, glycosylation predictions solely based on the protein sequence are more difficult.

In general, glycosylations impart a discrete recognitional role to the protein and broaden the range of its functionality.[11] For instance, they are able to regulate the folding, stability and function of the underlying protein, provide target structures for lectins and specific antibodies, and mediate cell-matrix interactions as well as cell-cell recognition.[43,51] Genetic defects in glycosylation, termed congenital disorder of glycosylation (CDG), although rare, often result in embryonic death or a range of severe symptoms and retardation, highlighting the vital role of glycans.[52] The earliest hypotheses, discussing why proteins are glycosylated, were brought up by Edwin Eylar in 1966 and Winterburn and Phelps in 1972.[11,53] They suggested that glycosylation either serves as a marker for extracellular transport, like a postage stamp, or codes for the topological location within the organism.[11] It is indisputable that the majority of glycoproteins occur on the extracellular site of cells; however, glycosylated proteins can also be found in a variety of cellular compartments.

When leaving the glycocalyx and travelling inwards into the cytoplasm, glycosylation of proteins becomes less diverse. There are actually only a few modifications known, all being based on an O-linkage and the attachment of single monosaccharides like GlcNAc, Glc, Fuc or Man to serine or threonine (Figure 1.3).[19] O-GlcNAc is quantitatively probably the most common type of glycoconjugate that can be found on over 6000 proteins, represented in all functional classes and cellular compartments.[54] Its function is hard to summarize as it depends on the specific environment, development stage and of course the protein type. Clinical examples that are influenced by O-GlcNAcylation include Diabetes mellitus Type 2, breast cancer and lung cancer.[55]

Not to be forgotten is the glycosylation machinery, which is located entirely within the cell. It is the basis for the enormous diversity of glycan structures, as these are not directly encoded in the genome but produced and modified by a complex template-independent network of enzymes. It reflects on the fact that the sugar code is not built on the same principles as the genetic code, which is simply replicated based on a template, but rather follows much more intertwined concepts due to its complex generation procedure. There are around 250 – 500 genes that are devoted to the synthesis and remodeling of glycan chains, giving rise to the already mentioned CAZymes, primarily acting in the endoplasmic reticulum (ER) and Golgi apparatus.[13] The biosynthetic pathway of N-glycans in eukaryotes is divided into the production of the dolichol pyrophosphate (Dol-P-P)-linked precursor $Glc_3Man_9GlcNAc_2$ and is followed by the attachment, trimming and subsequent elongation of the glycan at the polypeptide chain.[44] The first phase starts with the subsequent addition of individual monosaccharides (first: two GlcNAc, second: nine Man, third: three Glc) to the lipid-like polyisoprenoid molecule Dol-P-P at the ER membrane facing the cytoplasm. The different monosaccharide types are mostly derived via epimerization from the carbon and energy sources glucose or fructose, transported into the cell and further enzymatically converted.[56] For the elongation of an oligosaccharide by a new sugar residue via a glycosidic linkage, monosaccharides first need to be activated via the addition of nucleoside diphosphates, for example uridine diphosphate (UDP) or guanosine diphosphate (GDP), forming nucleotide sugars.[57] They provide sufficient energy in their phosphor – ester bond to form the 'high – energy' glycosidic linkage by glycosyltransferases, where the individual steps to assemble the 14-sugar glycan $Glc_3Man_9GlcNAc_2$ are performed by a conserved set of asparagine-linked glycosylation enzymes at the cytoplasmic and luminal site of the ER.[58] Transfer of the dolichol-linked precursor en bloc to the asparagine of a recognition sequence in a protein is mediated by the oligosaccharyltransferase (OST), occurring co-translationally and post-translationally in the ER lumen with the release of Dol-P-P (Figure 1.5).[59]

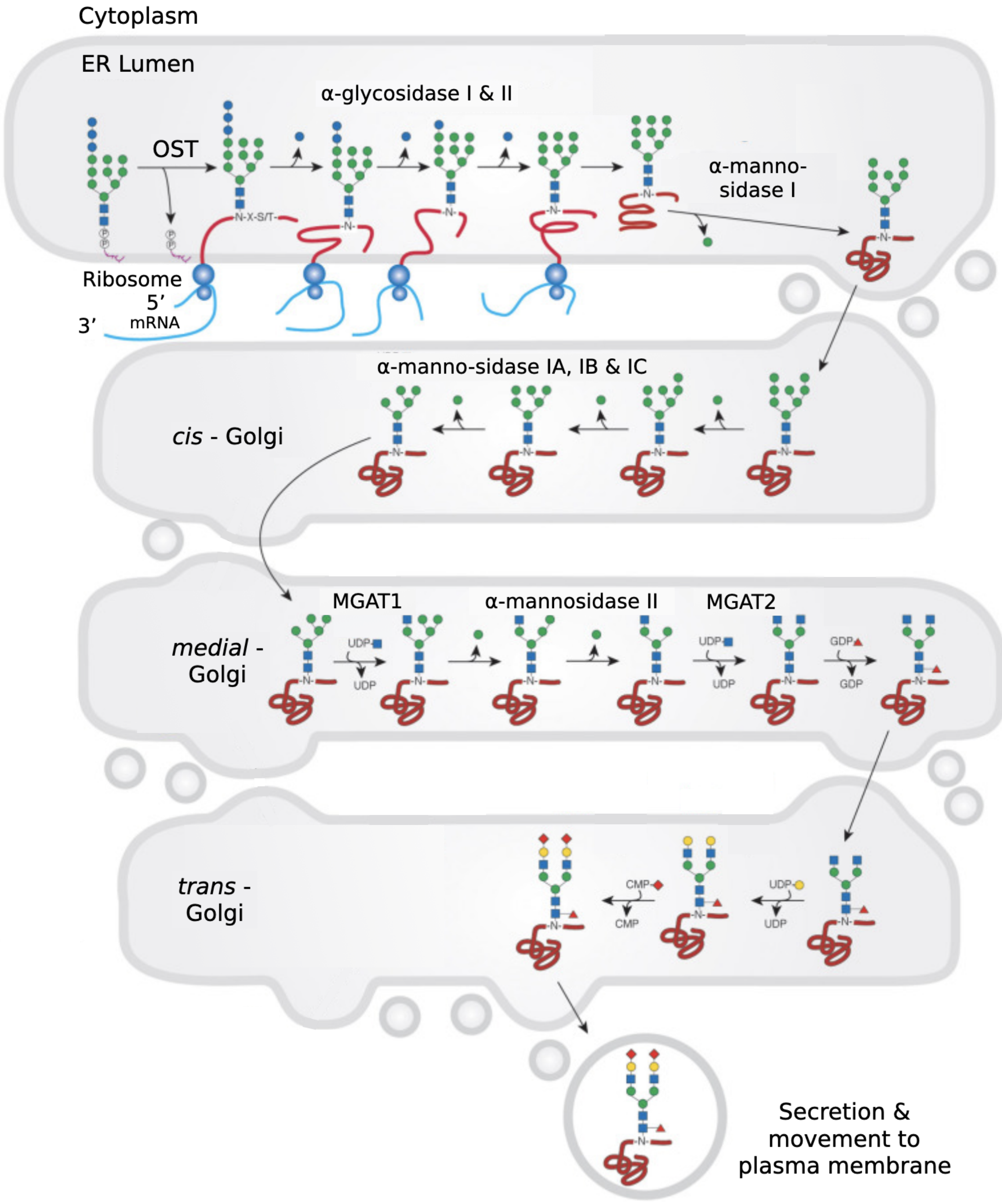


Figure 1.5: **Schematic N-glycosylation synthesis pathway in eukaryotes.** The N-glycan precursor is transferred to the polypeptide chain by oligosaccharyltransferase (OST) in the ER lumen, prior to further trimming and decoration in the ER and Golgi. The action of various enzymes is orchestrated to convert high-mannose type N-glycans into hybride and complex ones. The processed protein is depicted as a coil in red and glycan structures are represented with their monosaccharide symbols according to the SNFG. Adapted from *Essentials of Glycobiology, 4th Edition.*[44]

Following attachment, the N-glycan is trimmed by the interplay of several enzymes, such as α-glucosidase I & II cleaving off the three Glc and α-mannosidase I cleaving off one terminal Man. In most cases the final structure is $\text{Man}_8\text{GlcNAc}_2$, where ER chaperones also regulate folding of the glycoprotein before transfer to the Golgi.[60–62] Shuttling of premature glycoproteins to the *cis*-Golgi leads to additional trimming and formation of $\text{Man}_5\text{GlcNAc}_2$ by the action of α1-2 mannosidases IA, IB, and IC.[62] The resulting small high-mannose type N-glycan is the key building block for subsequent hybrid and complex N-glycan synthesis, performed in the *medial* and *trans*-Golgi. Prior to cleavage of two further mannoses by α-mannosidase II, a GlcNAc residue needs to be attached to the C2 of the α1-3Man via the mannosyl glycoprotein N-acetylglucosaminyltransferase (MGAT1), forming GlcNAcMan$_3$GlcNAc$_2$.[63,64] After the addition of another GlcNAc sugar via MGAT2, all glycans become complex, and GlcNAc, Gal, Fuc or Sia residues are gradually linked. It is particularly important to understand that enzymes like glycosidases and glycosyltransferases are depending on the prior action of other glycosylation enzymes, as their substrate recognition is fine-tuned, although they are also often acting on the same acceptor or donor types and therefore compete for their substrates.[44]

Additionally, glycans along the whole pathway may also escape certain processing steps, which is not necessarily bad, as high-mannose type N-glycans do originate from this mechanism, resulting in structures of the form Man$_{5-9}$GlcNAc$_2$. Therefore, consequences of incomplete glycan processing can result in i) the degradation and recycling of the glycoprotein, ii) secretion of an immature glycoprotein to the plasma membrane or iii) secretion of a mature glycoprotein harboring diverse high mannose or hybrid type N-glycans on its surface. It is worth emphasizing that the outlined N-glycosylation pathway is only an example and the expression of glycosidases and glycosyltransferases is highly flexible, depending on the species, cell type and physiological conditions.

It is the outlined temporal and spatial flexibility of glycosylation patterns that leads to a constantly changing sugar vocabulary on proteins and makes the interpretation of *glycan* words and the creation of a dictionary so difficult. The impact of glycans is as diverse as their structures, whereby recognition and reading by proteins leads to a specific biochemical function or signal. This means that the information transfer must follow certain rules, which we need to understand if we want to decode the sugar code. What has been disregarded until now is that particular residues are recognized not only due to their monosaccharide type, but also their position in a branching structure or their global conformation. An example for such topological recognition specificity is the enzyme α2,6-sialyltransferase that transfers a 2→6-linked sialic acid to one of the termini of complex N-glycans. It preferentially adds to the 1→3-linked branch (with a specificity orders of magnitude larger than on other sites) although the residues on the 1→6-linked branch are chemically identical back to the mannose at the junction (Gal$_{\beta1\rightarrow4}$GlcNAc$_{\beta1-2}$Man$_{\alpha-}$).[65] Another example is the differential conformer selection of lectins[14], where a plant and an animal lectin bind to the same glycan tree, but to different conformations.[66] It becomes visible that there is an enormous impact of the three-dimensional glycan structure on glycan recognition processes and underscores its vital role in deciphering the sugar code.

1.3 Conformational flexibility of glycans

***Note**: Throughout the dissertation the biochemical definition of configuration and conformation is used. A configuration describes the relative position of atoms in a molecule, which can only be altered through cleaving and reforming chemical bonds. A conformation is the shape of a molecule that can be adopted by means of rotation around single bonds.*

It was Emil Fischer in 1894, who discovered the specificity of enzymes like invertase (hydrolyzing saccharose in fructose and glucose) for their substrate configurations (epimers and anomeric configurations), concluding that the correct configuration of the substrate in the protein binding pocket is required to fit like a key to a lock.[67] Almost hundred years later, nuclear magnetic resonance (NMR) and Molecular Dynamics (MD) simulations discovered the conformational flexibility of glycans, enabling them to switch between diverse low-energy conformers, in contrast to existing as one rigid three-dimensional structure.[68–70] It was the beginning of an hypothesis presuming that due to the different conformers adopted by one glycan configuration, there must be a bunch of keys which can be selected by a receptor.[71] The different conformers are part of a conformational phase space that sketches a landscape with maxima and minima of energy, where the movement from one valley to the next is a rapid process. This has long been overlooked in experimental techniques because of the usage of crystallography, as glycans can prevent the crystallization process and are often enzymatically removed prior to experiments.[14]

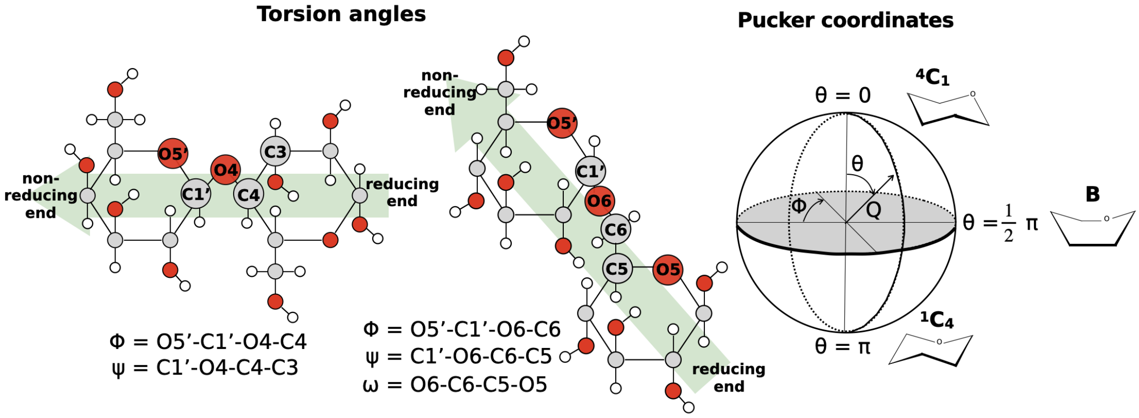


Figure 1.6: **Torsion angles and puckering: Conformational variables describing the key degrees of freedom of polysaccharides.** Depicted are two disaccharides connected via a 1→4 and 1→6 glycosidic linkage, respectively, with torsion angles ϕ, ψ and ω for their geometrical description. Carbon atoms are represented in gray, oxygen in red, nitrogen in blue and hydrogen in white. Green arrows indicate the direction of elongation. Existing puckering conformations can be described by the spherical pucker coordinates ϕ, θ and Q, locating the chair conformers at the poles (C), the boat (B) or skew-boat (S) conformers at the equator and the half-chair (H) and enveloped (E) conformers inbetween (not shown).[72]

Primarily, conformers differ in their torsion angles around the glycosidic linkages between monosaccharides. Depending on the linkage type, there are either two or three torsion angles per linkage, denoted ϕ, ψ and, in the case of 1→6 linkages, ω (Figure 1.6). 1→6 linkages represent a special case, as the C6 carbon is not part of the six-membered saccharide ring, introducing an additional degree of freedom and therefore a third torsion angle. The relative positions of the individual saccharide monomers within each possible conformer are stabilized by hydrogen bonds between the hydroxyl groups of the monomers.[73,74] Torsion angles are defined as ϕ = O5′–C1′–Ox–Cx, ψ = C1′–Ox–Cx–C(x–1) and ω = O6–C6–C5–O5, with x being the carbon number of the linkage at the non-reducing end. An exception are the 2→6 angles between Gal and Neu5Ac, which are defined as ϕ = O6′–C2′–O6–C6, ψ = C2′–O6–C6–C5 and ω = O6–C6–C5–O5.

Furthermore, the second structural feature is characterized by the distortion of the six-membered ring, also called puckering (Figure 1.6). It describes the position of the six atoms within the ring, 4-5 located on the same plane and 1-2 out-of-plane, depending on the conformation. There exist 38 puckering conformations which can be grouped into classes termed chair (C), half-chair (H), enveloped (E), skew-boat (S) or boat (B).[75] Chair and boat conformations exhibit out-of-plane atoms on opposite sites, the former having one atom up and one down ($^1C_4$, $^4C_1$), the latter both atoms up or down (e.g. $^{2,5}B$, $B_{1,4}$). Half-chair (e.g. $^OH_5$, $^OH_1$) and skew-boat (e.g. $^3S_1$, $^5S_1$) conformers display four and three consecutive atoms on a plane, respectively, having two atoms out of plane, one up and one down. Envelope conformers contain only one out-of-plain atom (e.g. ^3E, E_3). All conformers can be unambiguously mapped in a three-dimensional fashion using the spherical pucker coordinates ϕ, θ and Q, introduced by Cremer and Pople.[72] Monosaccharides occur predominantly in the chair conformations $^1C_4$ or $^4C_1$, depending on the sugar type, especially when they are involved in glycosidic linkages. It is believed that due to the equatorial alignment of sugar residues in a glycan accomplished by chair conformers, there are no steric hindrances or repulsions between atoms that are involved in hydrogen bonds, favoring this puckering.[76] Interconversion between pucker conformers depends on the monosaccharide type and the chemical environment, as well as on the exocyclic groups, which where shown to have an immense impact on the determination of energetical barriers between the different puckering conformers.[77]

The flexibility of glycans can be thought as the wacky movement of skydancers, constantly in motion with arms jumping up and down, heads moving from left to right, from front to back. It is unsurprising that nature is touching upon this feature, realizing that different conformations can affect enzymatic reactivity.[77] In the broadest sense, the glycobiology community is aware of the conformational flexibility of glycans, however, we are only beginning to understand its impact on the glycan functionality discussed in the above section. Especially, hard to unveil is the even further diversification of the sugar code due to the bunch of keys a single glycan represents, instead of only one. In most cases, we are in the dark; not able to grasp the still underestimated essential presence of glycan structures on biological macromolecules, due to lack of knowledge or of shortage of appropriate methodologies.

1.4 Structural glycobiology

To shed a little light on the dark, we want to pick up on the conformational flexibility of carbohydrates, also considered as the third dimension of the sugar code.[13] We focus on N-glycans in particular; their omnipresence on many protein surfaces and their sizes of seven to around thirty monosaccharides in eukaryotic cells make them a valuable target. When turning towards the investigation of their three-dimensional structures, one still experiences a lasting lack of suitable experimental methods to capture the many conformers a certain glycan configuration can adopt. Information about the average three-dimensional structures (average of glycan conformations) adopted by N-glycans, characterized by torsion angles along the glycosidic linkages, can in general be obtained by NMR or, in some cases, X-ray diffraction.[78] However, X-ray crystallography is highly unqualified to capture rapid motions in molecules, due to the required crystalline nature of the sample. Even the structural resolution of glycoproteins can be troublesome, as the hydrophilic and flexible N-glycans prevent crystallization and enzymatic cleavage of attached structures or mutation of N-glycosylation sites to prevent their modification is necessary.[79] If glycans remain attached, the maximum amount of structures resolved is limited to those that are in close contact to the protein surface, which might be the case for GlcNAc residues present in the core of N-glycans or glycans that act as ligands in a catalytic site. This issue leads to a great loss of information on glycosylation patterns and glycan conformations, as until now most structures are still resolved via X-ray crystallography.[80] Additional problems, like a missing common naming convention of glycan residues in the past, concur to the fact that only around 0.9 % of PDB entries are glycoproteins, in contrast to their ubiquitous occurrence.[78] A more successful technique in the structure determination of N-glycans is NMR, providing information about the linkage, anomeric configuration and glycan conformation.[81] For example the nuclear Overhauser effect (NOE) can deliver information about linkage conformations, but it is limited to short ranges (only up to 5 Å). Several NOEs are needed to unambigously map one conformer and in the end one only obtains an average three-dimensional structure, as measurements are executed on the millisecond to second time scale, during which the glycan can adopt several different conformations.[81] The same averaging problem also applies to NMR J-coupling constants, providing information about torsion angles of different glycan conformations.[81] One point from which both methods, X-ray crystallography and NMR, are suffering is the large amount of protein sample required, because the expression in bacterial systems, which generally yields the largest amounts, is problematic for glycoproteins.

In principle, only atomic-scale simulations are able to capture the dynamic behavior and deliver full details of the probability distribution of possible conformers in an N-glycan population.[77,82,83] The shortcoming of experimental methods and advantages of MD simulations was already recognized in the 1990s when simulating disccharides, although the early potential energy functions could not match experimental NMR values.[69] Even today the accuracy of employed force fields and the ergodicity of the used methods are the two most crucial points in order to obtain correct conformer distributions and still require refinement.[84,85] In particular, the slow transition between different conformational (rotameric) states prevents efficient phase-space sampling and convergence of conformer distributions in plain MD simulations.[86,87] For this reason, various enhanced-sampling MD techniques have been used to facilitate the crossing of relevant energy barriers and accelerate transition probabilities. These will be described in chapter 2 of this dissertation.

1.5 Objectives - N-glycans from different perspectives

This work aims to address current issues in the field of computational structural glycobiology. Summarizing the points mentioned above, there is a lack of proper glycan conformer sampling in MD simulations because of their high flexibility imposed by the many torsion angles. Along that line, there is no uniform glycan conformer labeling scheme implemented, preventing the possibility to quantify conformations and validate if and which are of greater importance. These global issues prevent the elucidation of the third dimension of the sugar code in MD simulation approaches, although this methodology is one of the most promising techniques in the glycobiology field to study three-dimensional structure-function relationships. In this dissertation the focus will be on both methodological advances as well as their application to solving current scientific questions:

- Chapter 3 covers the development of a novel enhanced sampling scheme applied to free N-glycans in solution in order to tackle their flexibility in MD simulations (Figure 1.7 left).
- Chapter 4 discusses the impact of post-translationally added N-glycans on the structure and function of the parasitic enzyme trans-sialidase from *Trypanosoma congolense*, which is known to be an important virulence factor of the disease trypanosomiasis (Figure 1.7 middle).
- Chapter 5 deals with a putative mutual dependence between torsion angles and puckering of N-glycans, especially when serving as substrate for CAZymes (Figure 1.7 right).

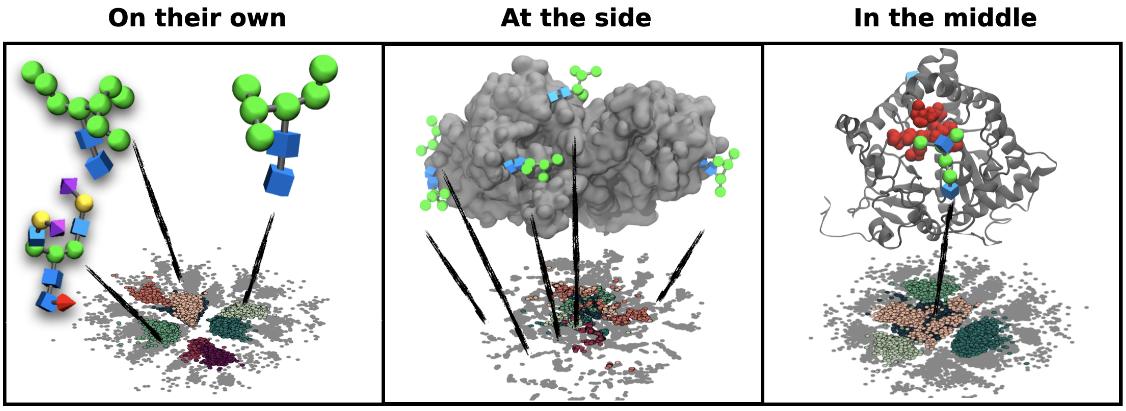


Figure 1.7: **Overview of chapter organization.** Investigation of N-glycans in various physiological environments: first on their own in solution, second when attached as post-translational modifications to proteins, and third when being substrates in catalytic binding sites.

The three chapters present results that were obtained in separately conducted and individually conceptualized studies. However, they are interconnected by their aim to add more insights into the concept of the sugar code, having a common focus on the flexibility of N-glycans and its function in various biological settings (sequence to three-dimensional structure-to-function relationship). The following sections outline the aim of each of the three mentioned chapters in more detail.

1.5.1 On their own - Exploring the flexibility of N-glycans

It may seem arbitrary to analyze a molecular compound on its own, if this only exists in a conjugated fashion. There are, however, several issues associated with the simulation of N-glycans at the molecular level which have to be resolved prior to the investigation of biologically relevant systems.

First, the unraveling of the sugar code is strongly impaired by the lack of standard structural descriptors, such as α-helices and β-sheets for polypeptides. Moreover, the non-linear, branched architecture of glycan chains and the variability of the type of linkages between the sugar monomers prevents the description of their three-dimensional structure in terms of a few conformational variables, as done in proteins via the two-dimensional representation of a Ramachandran plot.[88] Consequently, a glycan conformer labeling scheme is required to facilitate the fundamental study of three-dimensional structure-property relationships in N-glycan systems.

Second, computational studies in the past have shown that even enhanced sampling techniques, including replica-exchange MD (REMD)[89], Hamiltonian replica-exchange MD (H-REMD) with solute scaling (REST2)[90], well-tempered metadynamics [91] and Umbrella Sampling (US)[92], do not necessarily guarantee the convergence of N-glycan distributions when simulated free in solution. The flexibility of individual glycan branches is reminiscent of the conformational variability of disordered peptides. For instance, methods based on bias potentials applied to specific collective variables (CVs), such as well-tempered metadynamics, have so far focused on only few specific torsion angles (e.g. ω)[87], not giving justice to the structural complexity of N-glycans with multiple branches.[87,93–95] Also CV-independent methods such as REMD do not guarantee complete phase-space sampling[87,93,96], and require elaborate pre-calculations when used together with additional bias potentials.[94,97] Hence, there is a need to overcome these difficulties by developing new sampling approaches and achieving converged N-glycan conformer distributions.
Third, even converged trajectories do not guarantee correct conformer distributions per se, as the underlying empirical force field performance is reliant on a correctly parameterized potential energy landscape. There are several biomolecular force fields including carbohydrate parameters that vary in their conceptual setup, the extent of available monosaccharide types and parameterization attempts.[98] A comparison of these is necessary, monitoring their performance in comparison to experimental results as differences were found when simulating free glycans as well as protein bound ones.[84,99] Methodological advances covering the above mentioned issues are introduced in chapter 2 and 3 and their usefulness tested in various application scenarios like the sequence to three-dimensional structure paradigm of N-glycans.

1.5.2 At the side - Importance of surface glycosylation for neglected tropical diseases.

The post-translational modification of proteins by glycan molecules became popular and recognized during the last century. However, it only experienced a great boost during the SARS-CoV-2 pandemic, where studies about the N-glycan shield of the spike protein predicted its involvement in the recognition process via its receptor angiotensin-converting enzyme 2 (ACE2) by modulating the conformational dynamics of the receptor binding domain (RBD). This was made possible by the usage of experimentally determined glycosylation patterns in purely computational MD studies.[100,101] The spike protein is a great example for the exploitation of our sugar code by pathogens like enveloped viruses. As they replicate in human cells and utilize their internal glycosylation machinery, N-glycan patterns on viral receptor-attachment proteins or membrane-fusion proteins located in the outer membrane are identical to our own.[102] Glycans therefore facilitate infections by increasing transmission, viral binding to host cells or pathogenicity as they can mask antigenic sites, evading antiviral therapy due to a lack of antibody generation.[102]

Even pathogens that do not use the intracellular replication machinery of their hosts can still benefit from its N-glycosylated compounds. For instance, trans-sialidase (TS) enzymes, from the Glycoside Hydrolase Family 33, that are expressed on the surface of different species of parasitic *Trypanosoma* (*T.*), unicellular flagellate protozoa, utilize N-glycan compounds for their survival. The membrane-anchored enzymes cleave terminal Sia residues from host-cell glycoconjugates and transfer them to galactose residues on their own surface.[103–105] This surface sialylation has different beneficial functions for the parasite, which is unable to synthesize Sia de novo. During the cyclic trypanosomal life cylce, which alternates between a mammalian host and the tsetse fly as a vector[106,107], TSs particularly promote the survival of the trypanosome in the insect vector and enable it to escape the host's immune system.[108–112] The associated neglected tropical diseases can infect different mammalian hosts depending on the actual trypanosomal species; it includes the human pathogen *T. cruzi*, causing Chagas disease in South America, *T. brucei gambiense and T. brucei rhodesiense*, causing human African trypanosomiasis, also known as sleeping sickness, as well as animal pathogens being responsible for the animal African trypanosomiasis (mainly *T. brucei brucei*, *T. congolense*, *T. vivax*).[103,104,113,114] Typical symptoms of trypanosomiases are weight loss, anemia accompanied by fatigue and immunosuppression. Focusing here on the animal African pathogens; they cause fatal economic losses in agricultural sectors due to reduction of cattle population by 30 – 50 % and meat as well as milk production by 50 %, also leading to increased abortion rates and decreased birth rates in livestock.[115] These effects are causative agents for the overall reduction of benefits from livestock and farming, making it a top unstable form of sustenance. Thus, proper treatment of the disease and especially control of the parasite or vector are important. As TSs represent important virulence factors, they are also promising drug targets or vaccine candidates to combat the fatal diseases caused by trypanosomes. Their detailed study is therefore of high interest and is dealt with in chapter 4.

In addition to their catalytic activity involving the transfer of a sugar residue, of which the mechanism was under study for many years, only little attention has been drawn to the N-glycosylation sites present in the sequences of TSs.[116–119] The existence of high-mannose type N-glycans has been inferred indirectly by concanavalin A (ConA) purification for many TSs in early years.[104,120,121] Also other trypanosomal surface proteins were reported to harbor shorter high-mannose type N-glycans.[122–129] However, the lack of experimental data about detailed glycosylation patterns of TS enzymes has aggravated the study of the impact of surface glycosylation. Only the high number of recent studies providing increased evidence for N-glycans modulating substrate binding and turnover in various enzymes, like human proteases, where the influence was shown to be site-specific depending on the occupancy and diversity of N-glycans [130–133], has motivated us to deep dive into the study of surface glycosylation and its impact on TS activity. First, the dominant N-glycan types at each glycosylation site had to be experimentally determined, in order to be able to build an atomistic model of the glycan shield around the protein. Further, a combination of experimental approaches and MD simulations was employed to investigate the impact of N-glycans on substrate binding, potentially inherent to several members of the TS enzyme family.

1.5.3 In the middle - Induction of glycan conformations by CAZymes

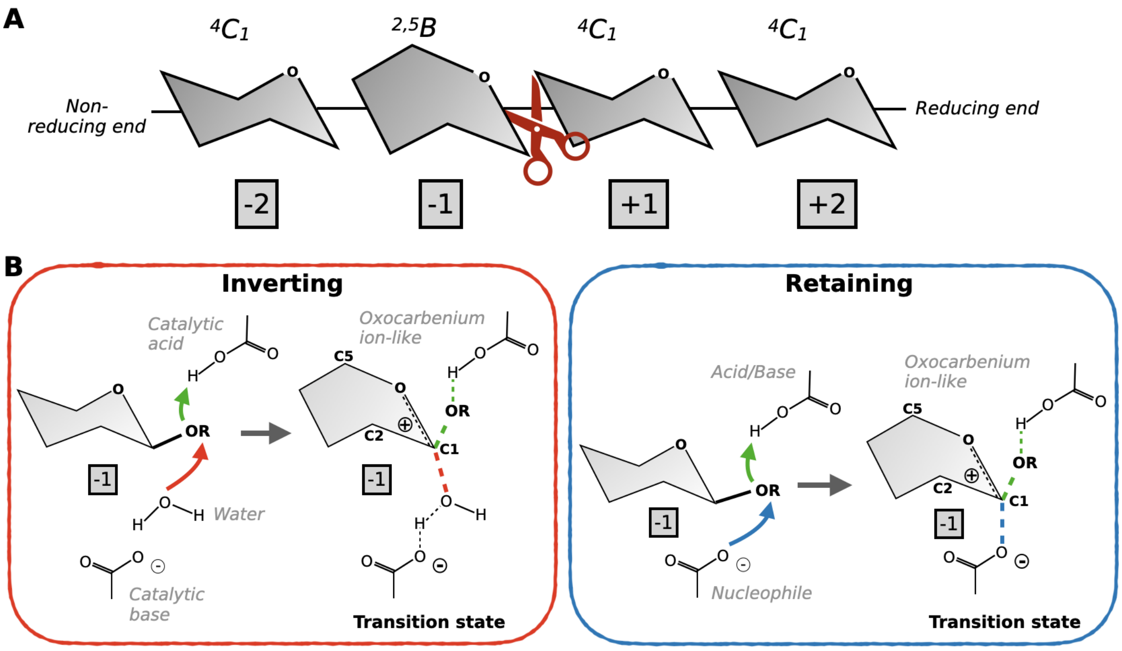


Figure 1.8: **Catalytic mechanisms of glycoside hydrolases.** **A** Bond cleavage occurs between the residues -1 and +1, where the monosaccharide at position -1 (facing the non-reducing end and representing the leaving group) is distorted towards a pucker conformation different from $^4C_1$. **B** Excerpt from the mode of action of inverting and retaining glycoside hydrolases, both of which contain a distorted sugar residue at position -1 in their transition state. An oxocarbenium ion is stabilized through an electron donation from the ring oxygen, leading to a positive charge at the anomeric carbon. This results in a distortion from the relaxed $^4C_1$ conformation into a structure where the C1, C2, C5, and O atoms are as coplanar as possible, which is true for some boat, half-chair and envelope conformations. Redrawn from Aldèvol et al.[134]

Next to their role as post-translational modifications, glycans also serve as ligands for enzymes grouped under the term CAZymes, like glycoside hydrolases (GH) or glycosyltransferases (GT). There are almost 300 families of discreet folds known that have carbohydrate-binding activity within CAZymes; a result of the need to adapt to the many different glycan structures available.[135] The diversity of enzymes is especially required under the consideration that GHs and GTs are solely responsible for the formation or breakage of all glycosidic bonds and the transfer of all carbohydrate residues within a cell. Especially the reaction mechanisms of GHs are of particular interest, as the glycan substrates undergo conformational changes regarding their puckering at monosaccharide position -1 when bound to the catalytic site, being slightly different for different GH families and different substrates (Figure 1.8 **A**).[77,134] The rearrangement is often necessary to achieve an axial linkage or oxocarbenium-ion character, resulting in a higher catalytic efficiency for adjacent bond cleavage, irrespective of the hydrolytic mechanism being inverting or retaining (Figure 1.8 **B**).[77,134] The origin of this conformational pucker change is debated to either arise through the chemical environment of the catalytic pocket or from global conformational changes of the glycan itself.[134] This fundamental research question was left unanswered in the past due to missing structural complexes of CAZymes with their corresponding carbohydrate substrates.[134] Although this issues has been resolved for many examples in recent years, the structural flexibility of glycan substrates as well as the dynamical nature of enzyme scaffolds give rise to major problems when studied by theoretical methods like MD simulations, molecular docking or quantum mechanics/molecular mechanics (QM/MM) approaches. It is either a lack of convergence in MD simulations due to incomplete sampling of the substrate's degrees of freedom, a limited receptor flexibility in molecular docking algorithms, or the too short timescales in QM/MM approaches.[136] As it is assumed that the precise conformation of the carbohydrate substrate has a significant effect on catalysis, the exploration of the complete conformational phase space of the glycan within the binding pocket is indispensable, especially, in order to answer the question whether certain glycan conformers favor monosaccharide pucker conformations at position -1 that are structurally prone to enzymatic cleavage. We aimed at addressing the outlined conformational versus chemical debate by employing α-mannosidase II (GMII) as a model system in chapter 5, knowing that its glycosylation reaction follows a pucker itinerary involving rearrangements from $^4C_1$ over $^OS_2$ to $B_{2,5}$ in its transition state. If the distorted pucker conformation in the transition state is achieved by the enzyme via its chemical environment or by a restricted conformational shape of the glycan in the binding site was examined via enhanced sampling MD simulations, QM calculations as well as employment of dimensionality reduction algorithms.

2 | The computational microscope

Note: *Parts of this chapter are taken from the publication: I.L. Grothaus, G. Bussi, L. Colombi Ciacchi, Exploration, representation and rationalization of the conformational phase-space of N-glycans, Journal of Chemical Information and Modelling, 62(20):4992–5008, 2022.*[137]

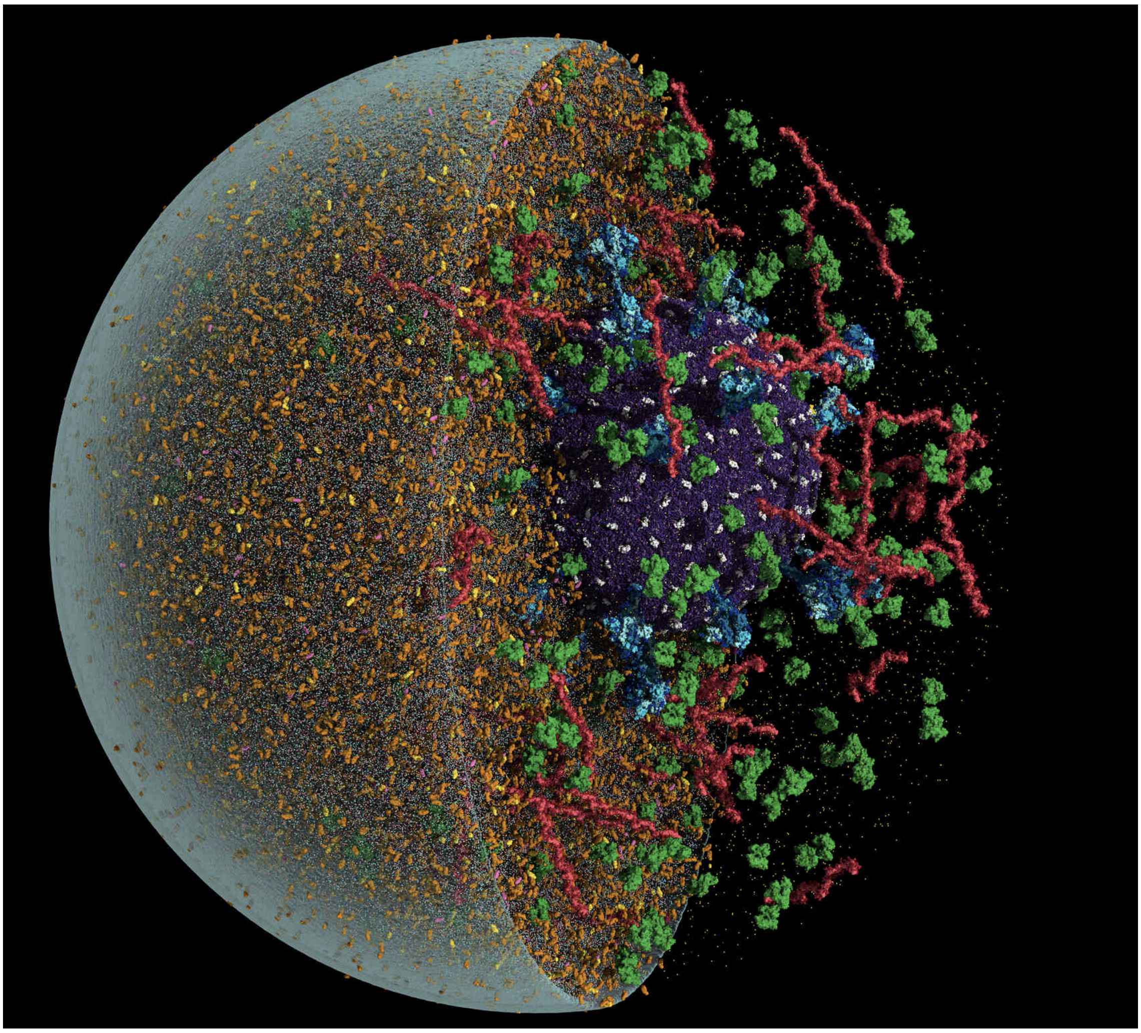


Figure 2.1: **Respiratory aerosol with Delta SARS-CoV-2 virus particle from an all-atom MD simulation.** The coronavirus (purple) with its Spike proteins (light blue) is surrounded by components of the deep lung fluid like mucins (red), albumin (green) as well as lipids (orange). The simulation box comprises over one billion atoms and was run for over 2 ns.[138] *Figure copied with permission from Lorenzo Casalino and Abigail Dommer (Amaro Lab, UC San Diego). Modeling: Abigail Dommer, Lorenzo Casalino, Fiona Kearns, Mia Rosenfeld, Nicholas Wauer, Clare Morris, Rommie Amaro (Amaro Lab, UC San Diego).*

Molecular dynamics simulations were developed in the 1950s[139], as a theoretical and computational technique that explicitly moves atoms in condensed matter over time. It enables the visualization of dynamical processes on an atomistic scale, often giving insights into mechanisms that can not be resolved by wetlab experiments. The progress in the field of MD simulation techniques and the increase in computational power have brought us to a complexity level where we can build up systems containing several millions of atoms and reach time scales up to milliseconds. Compared to the pathetic computational time of few picoseconds (ps) for a protein in the 1970s[140], nowadays, MD simulations can be characterized as a computational microscope depicting whole organelles or even cells. The recent investigation of a SARS-CoV-2 virus in an aerosol particle via all-atom MD simulations, one of the largest biological systems that has been modeled so far, gave important insights into our understanding of airborne disease transmission (Figure 2.1).[138] In order to understand how biology can be modeled with a computer, the mathematical concepts behind MD simulations and the limitations of the method should be discussed. The most prominent issues can be summarized by the following three points[141]:

- accuracy of force field parameters, which influence a correct description of the simulation system,
- limitation of simulation time, which can prevent the exploration of processes of interest,
- dimensionality of output data, which can hamper the interpretation of results.

In contrast to protein-related systems, less attempts have been performed to overcome these limitations in the study of glycans. The following three sections discuss the limitations with reference to the glycan perspective, highlighting the state of art and finally presenting improvements and new routes of investigation.

2.1 The approximation of energies

The basic concept behind MD is to solve Newton's equations of motion in order to move atoms over time, generating a trajectory in phase space. More precisely, Newton's second law can describe the time evolution of the position of N classical particles, depending on the acting force $\mathbf{F}_i(t)$ on each particle i with mass m_i:

$$\frac{\mathbf{F}_i(t)}{m_i} = \mathbf{a}_i(t), \text{ with } \mathbf{a}_i(t) = \ddot{\mathbf{r}}_i(t). \tag{2.1.1}$$

The resulting acceleration $\mathbf{a}_i(t)$ is the second time derivative (indicated by double dots) of the particle's position $\mathbf{r}_i(t)$, which can therefore be calculated via integration. Bold letters always indicate vectors. The forces $(\mathbf{F}_1(t), ..., \mathbf{F}_N(t))$ acting on the particles arise from the interaction among all individual particles in the system and can be derived from the gradient of the potential energy $E_{pot}(\mathbf{r}_1(t), \mathbf{r}_2(t), ..., \mathbf{r}_N(t)) = E_{pot}(\mathbf{r}(t))$:

$$\mathbf{F}_i(t) = -\frac{\partial E_{pot}(\mathbf{r}(t))}{\partial \mathbf{r}_i}, \tag{2.1.2}$$

under the assumption that no energy is lost due to friction or dissipation. Hence, the forces are considered to be purely conservative and the total energy of the system E_{tot} can be described by the sum of the kinetic energy E_{kin} and E_{pot}:

$$E_{tot} = E_{kin} + E_{pot} = \sum_{i=1}^{N} \frac{1}{2} m_i \cdot \mathbf{v}_i^2 + E_{pot}(\mathbf{r}), \text{ with } \mathbf{v}_i = \dot{\mathbf{r}}_i, \tag{2.1.3}$$

where the first time derivative of the particle positions $\mathbf{r}_i$ enters into the formula of E_{kin} and is defined as the velocity $\mathbf{v}_i$. The conservation of energy can be proven by differentiating equation 2.1.3 with respect to time and ($\frac{\partial E_{tot}}{\partial t} = 0$), equalizing to zero. The total energy can also be expressed via the Hamiltonian $\mathcal{H}$ as a function of the position and momentum $\mathbf{p}_i = m_i \cdot \mathbf{v}_i$ of each particle i, similarly to equation 2.1.3:

$$\mathcal{H}(\mathbf{r}, \mathbf{p}) = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + E_{pot}(\mathbf{r}), \tag{2.1.4}$$

with ($\mathbf{p} = \mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_N$). Taking the partial derivative with respect to $\mathbf{p}_i$ and $\mathbf{r}_i$, respectively, yields the Hamilton's equations of motion:

$$\frac{\partial \mathcal{H}}{\partial \mathbf{p}_i} = \frac{\mathbf{p}_i}{m_i} = \dot{\mathbf{r}}_i, \tag{2.1.5}$$

$$\frac{\partial \mathcal{H}}{\partial \mathbf{r}_i} = \frac{\mathrm{d}E_{pot}}{\mathrm{d}\mathbf{r}_i} = -\mathbf{F}_i = -\dot{\mathbf{p}}_i, \tag{2.1.6}$$

which determine the position vectors and momenta of the N particles as a function of time. The integration of these differential equations for a many-body problem is analytically not possible and therefore numerical integration is performed by breaking down the time t into short time steps δt. This implies that the force $\mathbf{F}_i(t)$ acting on each particle in its current positions $\mathbf{r}_i(t)$ is computed each time step and remains constant during δt until the new position at $t + \delta t$ is predicted from equation 2.1.1. Finite-difference approaches are the method of choice for the determination of positions and its time derivatives, using truncated Taylor expansions:

$$\mathbf{r}_i(t + \delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t)\delta t + \frac{1}{2}\mathbf{a}_i(t)\delta t^2 + \mathcal{O}(\delta t^3) \tag{2.1.7}$$

$$\mathbf{v}_i(t + \delta t) = \mathbf{v}_i(t) + \mathbf{a}_i(t)\delta t + \mathcal{O}(\delta t^2), \tag{2.1.8}$$

with $\mathcal{O}(\delta t^N)$ being the order of the truncation error. One example used throughout this dissertation is the Leap Frog algorithm[142] that updates positions at time t and velocities at time $t \pm \frac{1}{2}\delta t$. The determination of velocities at a mid-step, 'leaping' over the positions, have shown to increase the stability and accuracy of the algorithm. First, the current positions $\mathbf{r}_i(t)$ are used to obtain the accelerations $\mathbf{a}_i(t)$ from the forces $\mathbf{F}_i(t)$ acting on the particles according to equation 2.1.1. The velocities at the next mid-step $\mathbf{v}_i(t + \frac{1}{2}\delta t)$ are then derived from the velocities at the previous mid-step $\mathbf{v}_i(t - \frac{1}{2}\delta t)$ and the current accelerations:

$$\mathbf{v}_i\left(t + \frac{1}{2}\delta t\right) = \mathbf{v}_i\left(t - \frac{1}{2}\delta t\right) + \mathbf{a}_i(t)\delta t + \mathcal{O}(\delta t^2). \tag{2.1.9}$$

The updated positions $\mathbf{r}_i(t+\delta t)$ are finally obtained from the previous positions and the updated velocities:

$$\mathbf{r}_i(t+\delta t) = \mathbf{r}_i(t) + \mathbf{v}_i\left(t+\frac{1}{2}\delta t\right) + \mathcal{O}(\delta t^2). \tag{2.1.10}$$

In order to obtain E_{kin} and E_{pot} at the same time, the velocity $\mathbf{v}_i(t)$ can be estimated from the average of the velocities of the previous and next time-step. The velocities at the previous mid-step $\mathbf{v}_i(t-\frac{1}{2}\delta t)$ are most often missing at the beginning of an MD simulation $(t = t_0)$ and can be estimated from the Maxwell-Boltzmann distribution $f_{\mathbf{v}}$, providing the probability for the velocity vector $[v_x, v_y, v_z]$ as a function of temperature:

$$f_{\mathbf{v}}(v_x, v_y, v_z) = \left(\frac{m}{2\pi k_B T}\right)^{\frac{3}{2}} \exp\left[-\frac{m(v_x^2+v_y^2+v_z^2)}{2k_B T}\right], \tag{2.1.11}$$

with k_B being the Boltzmann constant and m the particle mass. The most probable velocity at temperature T can be calculated from the maximum of the distribution, as well as the mean velocity from the weighted integral over all possible velocities.

The various known finite-difference algorithms like Verlet[143], Velocity Verlet[144] or Leap Frog only differ in the truncation error of the Taylor expansion. The derivation of the time step $\delta t = 2$ fs generally used in standard MD simulations is further explained in appendix A.1.

2.1.1 The potential energy function of force fields

When integrating the Newton's equations of motion, in principle only m_i, $\mathbf{a}_i$ and $\mathbf{F}_i$ are required as inputs. However, from equation 2.1.2 it becomes apparent that the potential energy E_{pot} is required to derive $\mathbf{F}_i$. In MD simulations an empirical potential energy function is utilized to determine E_{pot} at every step for every particle in the system. This approach is only an approximation but it often delivers acceptable accuracies due to the high number of parameters that are included in the calculation of E_{pot}, allowing the reproduction of experimental bulk phase properties.[145] The widely used Class I potential energy function consists of a bonded part that corresponds to connected atoms in a molecule and a nonbonded part describing the interatomic electrostatic and van der Waals (vdW) interactions[145]:

$$E_{pot} = E_{bonded} + E_{nonbonded} \tag{2.1.12}$$

The potential energy arising from the connectivity of atoms is described by a sum over all bonds of lengths b, all valence angles θ and all torsion angles, differentiating between proper (ϕ) and improper (φ) torsions:

$$\begin{aligned} E_{bonded} = \sum_{bonds} k_b(b-b_0)^2 + \sum_{angles} k_\theta(\theta-\theta_0)^2 + \sum_{\substack{improper\\torsions}} k_\varphi(\varphi-\varphi_0)^2 + \\ \sum_{torsions}\sum_{n=1}^{6} k_{\phi,n}(1+\cos(n\phi-\delta_n)). \end{aligned} \tag{2.1.13}$$

Herein, $k_b, k_\theta, k_\phi, k_\varphi$ are the force constants, b_0, θ_0, φ_0 the respective equilibrium values, δ_n the phase shift and n the multiplicity of the function.[145] In this representation, bonds and angles are described through simple harmonic potentials according to Hooke's law. Torsion angles are expressed by the first few terms of a Fourier cosine series. The improper term is used to describe the so-called out-of-plane bending of trigonal groups, which is important for the description of aromatic rings to ensure the maintenance of planarity.

In the nonbonded part, the electrostatic term is described by the Coulomb potential (first sum) and the vdW term typically by the Lennard-Jones (LJ) 6-12 potential (second sum) in a pairwise fashion:

$$E_{nonbonded} = \sum_{i<j} \frac{q_i q_j}{4\pi\varepsilon||\mathbf{r}_i - \mathbf{r}_j||} + \sum_{i<j} \epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{||\mathbf{r}_i - \mathbf{r}_j||} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{||\mathbf{r}_i - \mathbf{r}_j||} \right)^{6} \right], \qquad (2.1.14)$$

in which q_i and q_j are the partial atomic charges on each particle i and j, and $||\mathbf{r}_i - \mathbf{r}_j||$ the distance between atoms i and j. ε is the dielectric constant in vacuo, ϵ_{ij} the well depth of the particle pair potential and σ_{ij} the distance where the potential reaches a minimum. The Lennard-Jones potential handles short distance repulsive interactions with the r^{-12} term and attractive dispersive interactions at long ranges with the r^{-6} term. When computing heteroatomic interactions between two different atom types i and j, the Lorentz-Berthelot[146] combination rules are applied:

$$\epsilon_{ij} = \sqrt{\epsilon_{ii} \cdot \epsilon_{jj}}, \qquad (2.1.15)$$

$$\sigma_{ij} = \frac{\sigma_{ii} + \sigma_{jj}}{2}. \qquad (2.1.16)$$

Dispersive vdW and electrostatic interactions between atoms that are separated by less than three bonds are not considered in the non-bonded part and a special 1-4 interaction scaling term is applied to atoms separated by three covalent bonds, reducing their magnitude of interaction. Thereby, bonded atoms separated by two other atoms (three bonds) do not only feel the potential term of the torsion angle but also a nonbonded contribution.

In most biomolecular MD simulations each atom present in the system is simulated explicitly (so-called all-atom approach). Therefore, the aforementioned particles symbolize atoms that are treated as point masses with a fixed atom-centered point charge, which is a very minimalistic description of the quantum mechanical (QM) reality. This radical approximation is necessary in order to access biologically relevant system sizes and time scales, as quantum mechanical calculations, based on the solution of the Schrödinger Equation, exhibit extremely poor computational scaling.[145] Each element (e.g. C) is further divided into specific subgroups and labeled accordingly (OC for a carbon bonded to an oxygen or CH3 for a carbon bonded to three hydrogens), incorporating the chemical environment and linkage of each atom resulting from different hybridization states. Molecules are therefore treated as mechanical systems, where atoms are connected by springs, which can not be broken in the duration of a simulation. The above mentioned contributions to the total potential energy of the system require the careful consideration of long-distance effects, where the concept of periodic boundary conditions and suitable methods to describe long range interactions accurately are explained in the appendix A.2 and A.3.

2.1.2 Biomolecular force fields

The physical model described above for the approximation of potential energies of an all-atom system (equation 2.1.12) is referred to as a force field. It also includes required parameters that enter into the calculation of the energy function such as equilibrium values and force constants. A quite general form of the function is given in equations 2.1.13 and 2.1.14, however it needs to be noted that depending on the force field type, the terms entering into the function can vary although the separation into bonded and nonbonded interactions is conserved. The parameter set is even more force field depended and mostly determines the quality of performances for the applied molecular system. Depending on the parameters, their values are estimated from matching MD simulations to experimental studies or quantum mechanical calculations. Parameters are most often derived for small model compounds and further generalized to build up a molecule from many small functional groups, ensuring transferability of the parameterization to many diverse molecular structures.

The parameterization procedure is usually performed in a sequential fashion: First, force constants of bonds and angles can be derived from infrared radiation (IR), Raman scattering or QM-based normal-mode analysis and corresponding equilibrium values from X-ray crystallography or QM calculations.[147] Subsequently, the partial atomic charges for the Coulomb potential are derived from the electrostatic potential (ESP) of the molecule calculated from *ab initio* QM simulations.[148] Atom-centered point charges that reproduce the QM ESP are determined, under restraints which guarantee the same point charge on equivalent atoms.[149] Next, the ϵ_{ij} and σ_{ij} vdW parameters can be derived from crystal structures but are more commonly taken directly from MD simulations.[147] Especially the accurate description of bulk phase properties is highly dependent to the dispersive interactions. Usually parameters are fine-tuned by comparing to experimental properties like liquid densities, heats of vaporization or free energies of hydration. Lastly the torsional terms are optimized, eminently influencing the performance of a force field due to the large structural changes they can induce.[148] Parameters are obtained from *ab initio* QM calculations, via torsion scanning calculations, where the potential energy surface for the rotation around the torsional axis is determined. In subsequent MD simulations, the torsional force field term is then fitted to reproduce the rotational profile. The performance can be evaluated against experimentally measured frequencies and energy differences.[147,148] As the torsion angle term is fitted last, it is mainly affected by the parameterization of 1-4 interactions (non bonded interactions separated by three bonds), influencing the torsional energy.[148]

The large amount of parameters to fit and the dependency and correlation between the different terms make the parameterization procedure a difficult task. The parameters highly depend on each other and sometimes also lack physical significance as approximations have to be made in order to provide a certain degree of generalization for different types of systems and need to incorporate solvent effects. All these compromises make the force field concept in its current form purely empirical. Its application will therefore necessarily result in foreseeable shortcomings and inaccuracies that can only in part be reduced by continuous refinement of parameter sets.

The outcome of this parameterization ambiguity is a divergent evolution of several different force fields for the same biomolecules, requiring a critical comparison and assessment regarding their performance for different applications. Commonly used biomolecular force fields for the simulation of carbohydrates or protein-carbohydrate systems are for instance the Chemistry at Harvard Molecular Mechanics (CHARMM)[150–152], AMBER[153–155], Groningen Molecular Simulation (GROMOS)[156–159] or OPLS-AA.[160,161] They mostly do not only provide parameters for standard amino acids, but also for lipids, RNA, DNA, carbohydrates and ions, enabling the simulation of complex biological systems. Unfortunatly, the GROMOS and OPLS-AA force fields only provide a small parameter set of monosaccharides and are rather suited for the simulation of unlinked glycans.[162] Since the simulation of glycoconjugates is desired in this work, the focus will mostly lie on the CHARMM and AMBER force fields, providing parameters for most pyranoses and furanoses.[162] The latest CHARMM force field version called CHARMM36 provides parameters for proteins, nucleic acid, lipids and carbohydrates in an additive fashion. Additionally, there is the CHARMM General force field (CGenFF) for any organic molecule that can be combined with the all-atom CHARMM36. The structuring of the AMBER force field family is different, as there exist individual force field names for each biological compound. For instance the latest versions are termed 'ff19SB' for proteins[153], 'GLYCAM06j' for carbohydrates[155], 'lipids21' for lipids[163] and 'gaff2' for small organic molecules[164], whereby they can all be combined within one simulation.

The differences between the functional forms of both force fields, as well as alternative approaches for obtaining force field parameters should shorty be discussed to highlight the general divergences. The current CHARMM36 force field version was derived from the first all-atom CHARMM force field termed CHARMM22 with improvements of side-chain dihedral parameters and reoptimization against high-level QM data.[165,166] The CHARMM force field especially deviates from the standard Class I potential energy function in the bonded part, as it is extended by two extra terms in addition to those in equation 2.1.13. On the one hand, the valence angle between terminal atoms (1,3) is improved by the harmonic Urey-Bradley term:

$$E_{UB} = \sum_{angles 1,2,3} k_{UB}(r_{1,3} - r_{1,3;0})^2, \tag{2.1.17}$$

where k_{UB} represents the force constant, $r_{1,3;0}$ the equilibrium value and $r_{1,3}$ the instantaneous value of the distance.[145] On the other hand, the correction map term CMAP enhances the correct conformational properties and secondary structures in peptide bonds along the ϕ and ψ angles, enabled by a cubic spline potential that corrects the two-dimensional energy surface along ϕ and ψ. It can however not be assumed that these two terms impact the performance of the glycan parameters as the Urey-Bradley term only acts on terminal atoms that likely do not influence the rotation around torsion angles and CMAP is only applied to peptide bonds. Regarding the parameterization philosophy, the CHARMM force field performance is optimized to be relevant in condensed-phase applications, reproducing densities and heats of vaporization of bulk liquids.[145] Partial charges are derived by the supramolecular approach[167], where the minimum interaction energies and distances between functional groups and water molecules are first determined by *ab initio* simulations. Charges are then optimized to reproduce these values in MD simulations of the corresponding compound with the TIP3P water model.[148]

The first all atom force field of the AMBER family for the explicit simulation with water molecules was AMBER ff94. Over the subsequent force field versions, primarily torsion angles parameters have been refined using QM calculations and experimental data sets for small peptides. The functional form of the AMBER family strictly follows that of equations 2.1.13 and 2.1.14 without any additional terms, therefore being almost identical to the CHARMM force field in terms of glycan simulations due to the point discussed above. The two force fields, however, differ regarding the parameterization of partial charges, as the AMBER family tries to obtain atomic charges that reproduce the electrostatic potential computed from QM calculations using Hartree-Fock in the gas phase.[165,168] The problem that these charges do not reproduce polarizability in the bulk phase is counteracted by the usage of the HF/6-31G* level of theory, overestimating the dipole moment by 15-20 % compared to gas phase values.[148]

Nevertheless, the solutions in all parameterization approaches are highly dependent on the geometry-optimized structure used for the QM calculations and of course also on the level of quantum theory used. The resulting uncertainty is reflected in the huge variety of atomic partial charges that can be found in different force fields.[98,165] A large impact can also have the treatment of hydrogen bonds that occur between individual monosaccharides. These are namely not explicitly treated by either force fields through an extra term, but rather just modeled through electrostatic and vdW terms, and are therefore dependent on the correct parameterization of such.

2.2 The sampling problem

2.2.1 Thermodynamic ensembles and their state functions

A molecular dynamics simulation is an ensemble of particles obeying Newton's law, from whose collective motion statistical information can be extract that we call the four laws of thermodynamics. Consequently, MD simulations are based on the concept of statistical mechanics that describes the microscopic interaction of individual particles. Classical thermodynamics is just a collective view of statistical mechanics, where a system is described by bulk macroscopic properties. The systems' properties can be expressed by thermodynamic variables, where there are intensive variables (system size independent) like pressure p, temperature T, chemical potential μ or concentration c, and extensive variables (system-size dependent) like volume V enthalpy H, entropy S or internal energy U. The following paragraphs briefly outline the applied thermodynamic concepts and draw the link between simulated, microscopic particles and macroscopic properties of the system.

Consider a box with walls, having a fixed volume V, filled with N particles, which are moved over time according to Newton's second law as described above, guaranteeing conservation of the total energy E_{tot}. This thermodynamic system depicts a so-called *NVE* ensemble, as the three variables V, N and E_{tot} are kept constant. The ensemble concept originates from statistical mechanics, where a macroscopic system with certain constraints, fixed variables, is described by a large number of microscopic conformations.

An ensemble is the collection of microstates which the system can adopt within the macroscopic constraints. A certain macroscopic state is characterized by a collection of different microscopic conformations. Thermodynamic properties A_{macro} of a macroscopic state can be derived from the behavior of the microstates by calculating statistical averages. These statistical averages can be obtained in two different ways:

> i) Via the calculation of time averages $\langle A \rangle_{time}$, where the property A is derived from the average of individual values A_t, taken on by the system while it visits different microstates over time:
>
> $$A_{macro} = \langle A \rangle_{time} \approx \frac{1}{N_{\delta t}} \sum_{t=1}^{N_{\delta t}} A_t, \tag{2.2.1}$$
>
> with $N_{\delta t}$ being the total number of time steps. The approximate relation becomes an equality only if $N_{\delta t} \to \infty$.
>
> ii) Via the calculation of ensemble averages $\langle A \rangle_{ensemble}$, where the property A is derived from an average over a collection of adopted microstates ($N_{microstate}$) at a fixed time point:
>
> $$A_{macro} = \langle A \rangle_{ensemble} \approx \sum_{k=1}^{N_{microstate}} A(\Gamma_k) P(\Gamma_k), \tag{2.2.2}$$
>
> where Γ_k denotes one point of the phase space of the simulated system. In general, this is a 6N-dimensional space defined by the set of all positions $\mathbf{r}_k$ and momenta $\mathbf{p}_k$ in each microstate k. For our simulations, momenta are not of interest, therefore we only consider the reduced, 3N-dimensional, phase space $\mathbf{\Gamma} = \{\Gamma_k\} = \{\mathbf{r}_k\}$. Within microstate k, $\mathbf{r}_k = \mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_i, ...$ with $\mathbf{r}_i = (x_i, y_i, z_i)$, i being the particle index and x, y, z the particles coordinates. As classical force fields do not allow for a change in configuration, variation of $\mathbf{r}$ only gives rise to different conformations and therefore $\mathbf{\Gamma}$ is also referred to as the conformational space. Later in this work, the conformational phase space is redefined to adapt to the application of glycan systems. $P(\Gamma_k)$ denotes the probability to observe the microstate k.

It is obvious that time averages can be easily obtained from MD techniques by simulating the system over a longer time period, capturing its visited microstates along the progression of the trajectory. On the other hand, ensemble averages are often present in wetlab experiments, for example when the adsorption of a protein sample is measured having a large (of the order of 10^{20}) number of identical molecules dissolved in solution adopting different microstates.

The ergodic hypothesis is a postulate in statistical mechanics combining the two outlined approaches. It states that the ensemble average gives the same solution as the time average for equilibrium properties, providing that the time is long enough and the ensemble large enough:

$$A_{macro} = \langle A \rangle_{time} = \langle A \rangle_{ensemble}. \tag{2.2.3}$$

This fundamental concept allows to predict macroscopic ensemble properties based on single-molecule MD simulations adopting various microstates over time, under appropriate constraints set on the thermodynamic variables (T, p, V, E, N). The choice of the constraints defines the type of ensemble. All of them are thermodynamic ensembles, meaning that they are in their statistical equilibrium. Their choice in an MD simulation depends on the variables that are controlled in experimental studies one wants to compare to. The NVE ensemble, also called microcanonical ensemble, is an isolated thermodynamic system with rigid, adiabatic walls. As it has impermeable walls it can not interact with its surrounding and hence the temperature can not be fixed, although this is crucial in every biological system. More appropriate are the canonical (NVT) or isothermal-isobaric (NpT) ensembles, having a fixed temperature T as well as fixed volume V or pressure p, respectively. The canonical ensemble can be envisioned as a closed thermodynamic system having rigid and diathermic walls, allowing for an energy exchange with an isothermal heat bath to control the temperature. The constraints of the NpT ensemble are even more similar to typical wet lab experiments, where the volume and energy of the system can change.

When solving Newton's equations of motion directly, the NVE ensemble applies, whereas in NVT and NpT ensemble simulations, the additional use of thermostats and barostats are required. For this purpose, the simulation box is coupled to a heat or pressure bath in order to exchange energy and keep the desired variables fixed.[148] For the mentioned ensembles, the temperature and therefore also the energy are not truly constant in the simulation praxis due to several factors including integration errors that sum up over the simulation time, heating due to frictional or external forces as well as drift during equilibration.[169] Commonly applied barostats and thermostats are described in detail in the appendix A.4.

Each ensemble harbors a unique thermodynamic probability distribution of microstates $P(k)$, depending on their energies $E(k)$. Since in our case we only include the positions in our phase-space definition $\mathbf{\Gamma} = \{\mathbf{r}_k\}$ and not the momenta $\{\mathbf{p}_k\}$, the following derivations only include the potential energy $E_{pot}(\mathbf{r}_k)$ and not the kinetic energy $E_{kin}(\mathbf{p}_k)$. For instance, the canonical probability distribution of finding a microstate $\mathbf{r}_k$ with potential energy $E_{pot}(\mathbf{r}_k) = E(\mathbf{r}_k)$ is given by

$$P_{NVT}(\mathbf{r}_k) = \frac{e^{-\beta E(\mathbf{r}_k)}}{Z_{NVT}}, \tag{2.2.4}$$

with $\beta = \frac{1}{k_B T}$. The nominator is the Boltzmann factor and Z_{NVT} the canonical partition function:

$$Z_{NVT} = \sum_k e^{-\beta E(\mathbf{r}_k)}. \tag{2.2.5}$$

In contrast, the probability distribution function of the NpT ensemble

$$P_{NpT}(\mathbf{r}_k) = \frac{e^{-\beta E(\mathbf{r}_k)} e^{\beta p V(\mathbf{r}_k)}}{Z_{NpT}}, \tag{2.2.6}$$

also includes the pressure p and volume of the specific microstate $V(\mathbf{r}_k)$. The isothermal-isobaric partition function Z_{NpT} is

$$Z_{NpT} = \sum_k e^{-\beta E(\mathbf{r}_k)} e^{\beta p V(\mathbf{r}_k)}. \tag{2.2.7}$$

The ensemble-dependent partition function describes the statistical properties of the system under the assumption of thermodynamic equilibrium. It serves as a normalization constant, making sure that the probabilities sum up to 1, partitioning the individual microstates depending on their corresponding energies. For instance, in the canonical ensemble:

$$\sum_k P(\mathbf{r}_k) = \frac{1}{Z} \sum_k e^{-\beta E(\mathbf{r}_k)} = \frac{1}{Z} Z = 1. \tag{2.2.8}$$

The probability distribution of microstates obtained from a molecular dynamics simulation can now be employed to derive the available energy in the system to perform work, and thus to predict the thermodynamic behavior of the system. In order to draw the connection between a microscopic description and a macroscopic state function, one needs to link statistical mechanics to classical thermodynamics.[148] This can be done using the Boltzmann-Planck equation of an ideal gas, relating the number of microstates in a system W to its entropy S:

$$S = k_B \ln W. \tag{2.2.9}$$

This equation shows that the equilibrium state of the system is characterized by the highest number of microstates, because this gives the maximum entropy, according to the second law of thermodynamics. Equation 2.2.9, however, assumes that all microstates are equally probable, which is not the case in a general thermodynamic system, and therefore it should be rewritten as

$$S = -k_B \sum_k P(\mathbf{r}_k) \ln P(\mathbf{r}_k). \tag{2.2.10}$$

Under the assumption of the canonical ensemble (omitting here the subscript NVT), $\ln P(\mathbf{r}_k) = -\beta E(\mathbf{r}_k) - \ln Z$, giving:

$$S = k_B \sum_k P(\mathbf{r}_k) \left(\beta E(\mathbf{r}_k) + \ln Z \right). \tag{2.2.11}$$

This can be rearranged to:

$$\begin{aligned} & S = k_B \beta \frac{\partial \ln Z}{\partial \beta} + k_B \ln Z, \\ & \text{considering that} \\ & \sum_k P(\mathbf{r}_k) E(\mathbf{r}_k) = \langle E \rangle = \frac{1}{Z} \sum_k e^{-\beta E(\mathbf{r}_k)} E(\mathbf{r}_k) = \frac{1}{Z} - \frac{\partial Z}{\partial \beta} = -\frac{\partial \ln Z}{\partial \beta}, \end{aligned} \tag{2.2.12}$$

defining $\langle E \rangle$ as the average potential energy of all possible microstates. By further substituting $\beta = \frac{1}{k_B T}$, one obtains:

$$S = \frac{\partial k_B T \ln Z}{\partial T} = -\frac{\partial F}{\partial T}, \tag{2.2.13}$$

with

$$F = -k_B T \ln Z_{NVT} \tag{2.2.14}$$

expressing the Helmholtz free energy F, the thermodynamic potential of the canonical ensemble. Equation 2.2.14 is a so-called bridge equation, as it relates the thermodynamic potential of the microstate, the partition function, with the thermodynamic potential of the macrostate, i.e. the property whose minimum defines the equilibrium value.[148] The Helmholtz free energy can also be exclusively expressed by thermodynamic variables by noting that:

$$\langle E \rangle = -\frac{\partial \ln Z}{\partial \beta} = \frac{\partial(\beta F)}{\partial \beta} = F + \beta \frac{\partial F}{\partial \beta} = F + TS, \tag{2.2.15}$$

which results in

$$F = U - TS \tag{2.2.16}$$

because the physical constraint of energy conservation imposes that the internal energy U be equal to the ensemble average of the potential energy $\langle E \rangle$. Equation 2.2.16 implies that the free energy of a state does depend on the interatomic interactions of the system through its internal energy term as well as on entropic effects. The Helmholtz free energy is applicable under canonical conditions, however experimental studies representing physiological conditions are mostly conducted under isothermal-isobaric conditions. Therefore, the NpT ensemble is often used in MD simulations, where the thermodynamic potential is the Gibbs free energy, which can also be derived from its partition function via the bridge equation:

$$G = -k_B T \ln Z_{NpT}. \tag{2.2.17}$$

It can further be expressed, similarly to the Helmholtz free energy (equation 2.2.16), considering also the pressure-volume product to incorporate the work associated with a change of volume :

$$G = U + pV - TS = H - TS, \tag{2.2.18}$$

with H being the enthalpy.

Free energy differences are able to predict the spontaneous evolution direction of a thermodynamic process. For example, in a chemical reaction the change in free energy ΔG specifies the maximum amount of energy that can be exchanged during a process and determines if the evolution from an initial state x to a final state y happens spontaneously.[170] States x and y are two subsystems in the NpT ensemble, for instance a ligand in solution (state x) versus a ligand bound to a protein (state y), considering stability of both states as well as a local ergodicity. The process of moving from state x to y is favorable if $\Delta G < 0$, irrespective of the separating barrier, and therefore can happen spontaneously. If $\Delta G > 0$, the reaction requires an external source of energy to go from a low energy state x to a high energy state y, and a change of state is rather unlikely.

The separating barrier between states x and y determines the rate of crossing from one to the other, majorly influencing the reaction kinetics. However, it is important to consider the equilibrium rate, as the process should be at all times in equilibrium with its surroundings to neglect other contributing energy forms like friction.[170] In order to estimate the free energy difference between states x and y, defined as sets of microstates

$k \in x$ and $k \in y$, their bridge equations from equation 2.2.17 can be combined[170]:

$$\Delta G = G_y - G_x = -k_B T[\ln Z_y - \ln Z_x]. \tag{2.2.19}$$

Substituting Z by equation 2.2.7 yields:

$$\Delta G = -k_B T \ln \frac{\sum_{k \in y} e^{-\beta E(\mathbf{r}_k)} e^{\beta p V(\mathbf{r}_k)}}{\sum_{k \in x} e^{-\beta E(\mathbf{r}_k)} e^{\beta p V(\mathbf{r}_k)}}, \tag{2.2.20}$$

which can further be simplified by rearranging equation 2.2.6 and substituting $e^{-\beta E(\mathbf{r}_k)} e^{\beta p V(\mathbf{r}_k)}$:

$$\Delta G = -k_B T \ln \frac{\sum_{k \in y} P(\mathbf{r}_k) \cdot Z}{\sum_{k \in x} P(\mathbf{r}_k) \cdot Z} = -k_B T \ln \frac{P_y}{P_x}. \tag{2.2.21}$$

The ratio $\frac{P_y}{P_x}$, and thus the free energy difference ΔG, can be estimated from MD simulations when enough crossing events between state x and y are ensured to guarantee sufficient statistics. Indeed, it is now even possible to compute individual free energies for different states, e.g.:

$$G_y = -k_B T \ln P_y = -k_B T \ln \frac{N_y}{N_{states}}, \tag{2.2.22}$$

by simply counting how often the system is visiting state y (N_y), divided by the total number of visits of all states N_{states}. Note, however, that only differences' of free energy and not absolute values are physically meaningful. Further advantageous is that the free energy of a thermodynamic system is a state function and only depends on the state of the system, not on its history, regardless of the underlying ensemble.

In MD simulations it is often reasonable to calculate free energy changes along a certain reaction coordinate, also called collective variable (CV). For example, in protein folding a CV along the helicity of a peptide backbone or in ligand binding studies a CV along the distance between a ligand and protein pocket can used. The obstacles that can occur during these approaches are further addressed in section 2.2.2, introducing methods for accurate and converged free energy calculations.

2.2.2 Free energy methods

Molecular dynamics simulations aim at predicting the time evolution of an atomistic system, capturing its multiple conformational changes in order to identify behavioral patterns that provide information about reaction mechanisms like ligand binding or structural changes in biological contexts. As already outlined above, statistical mechanics provides the framework for estimating probabilities of individual microstates, e.g. conformers of a molecule, allowing for the assessment of their phase-space distribution. In order to approximate the microstate probability distribution and subsequently derive macroscopic properties from a molecular simulation, a sufficient sampling of the conformational phase space is required to obtain reliable estimates. This is due to the fact that the ergodic hypothesis (equation 2.2.3) is only valid under the assumption that t is infinite when calculating the time average $\langle A \rangle_{time}$ of a macroscopic property, ensuring that all microstates have been sufficiently visited. Therefore, it needs to be guaranteed that the simulation time is long enough to allow for multiple crossings of relevant free energy barriers separating conformer states. Additionally, only converged simulations reproduce a valid probability distribution P, which is necessary to estimate correct free energies. The prediction of free energy profiles along certain system variables, such as distances between atoms or the helicity of a polypeptide chain, is crucial as they play fundamental roles in anticipating the evolution of the system and can also be compared directly to experiments. The resulting free energy distribution allows for the identification of (meta)stable states, predicting the equilibrium structures of a molecule which is, for instance, extremely important in the field of protein folding.

The outlined requirements most often can not be achieved with standard (unbiased) MD simulations, as the prerequisite of barrier crossings and therefore ergodicity is limited by the available computational power, restricting the accessible timescale in a simulation. Furthermore, with an increasing number of atoms within a molecular system, the degrees of freedom and therefore number of conformations rapidly increase. As biological systems mostly involve several hundreds of molecules like amino acids, lipids and saccharides, they represent complex systems in which the observation of certain effects needs to be enforced. In order to overcome this sampling problem, various approaches have been developed to facilitate barrier crossings during the accessible time scale of MD simulations. There are only certain degrees of freedom exhibiting high free energy barriers, but these are often also the most important for a chemical reaction to occur or a protein to adopt a different conformation. It is therefore necessary to first identify the correct CVs in order to allow the required transition to occur or the required region of the conformational phase space to be explored. Due to the complexity of the simulated systems it is often not advantageous to achieve complete ergodicity of the whole conformational phase space, e.g. sampling all possible folded and unfolded conformations of a protein. It is rather desirable to obtain ergodicity only along the significant CVs to yield quantitative estimates of probability distributions, e.g. differentiating between an active and inactive protein conformation. A CV describing the transition between two states is ideally the committor function, having a value of $\frac{1}{2}$ at the transition point between the two states.[141] It is however very rare that an explicit reaction coordinate can be identified as a committor, and CVs are rather chosen upon chemical intuition.[141] Independently of the methodology applied, the identification of relevant CVs harboring high energy barriers and allowing for restricted phase-space exploration is crucial for the success of the method and the requirement of ergodicity. There are CV-dependent and CV-independent methods, differing in the extent of perturbation of the system. The kinetics of the system is mostly not captured by either of them, only being able to predict equilibrium probability distributions rather than transition rates between states.[141] Furthermore, barrier heights are not invariant to the choice of CV, making kinetic interpretations very dangerous. Calculated free energy barriers should therefore not be used to estimate rates, as their height is highly influenced by the sampling method at hand.

The following paragraphs shortly outline important enhanced sampling techniques under the constraints of a canonical ensemble, especially those that have already been applied to the sampling of glycan structures namely (H-)REMD, REST2 and well-tempered metadynamics (as mentioned in section 1.5.1). We then introduce a newly derived enhanced sampling approach, combining the CV-based method Replica Exchange with Collective-Variable Tempering (RECT) with the CV-independent method REST2 in order to improve in particular the sampling of glycan structures.

2.2.3 Replica exchange

Temperature and Hamiltonian replica exchange

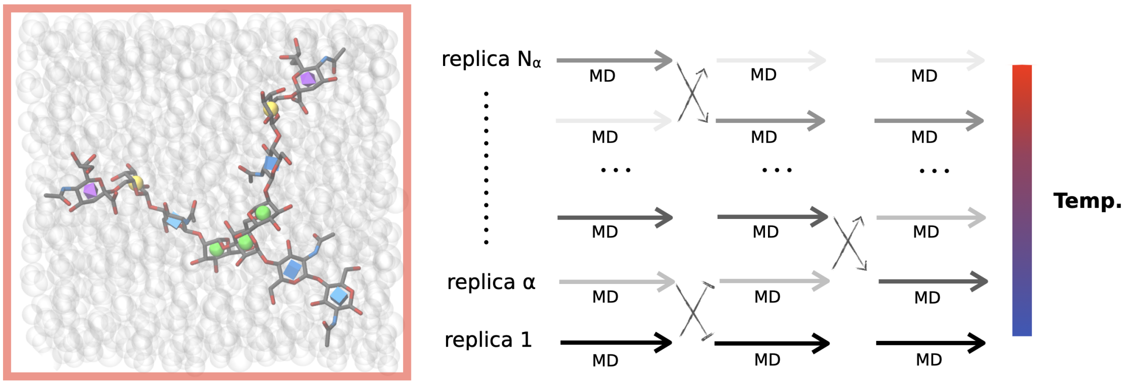


Figure 2.2: **Scheme of a temperature replica exchange MD simulation.** A complex model glycan (blue: GlcNAc, green: Man, yellow: Gal, purple: Neu5Ac) solvated in water (gray) is shown on the left, where the temperature increase is applied to all atoms of the system (red box around simulation box). N_α replicas are simulated in parallel, where coordinates are frequently exchanged between neighboring replica, based on the Metropolis criterion. The temperature is scaled geometrically across the replica ladder to facilitate barrier crossings along degrees of freedom that are stuck in free energy minima.

The replica exchange approach is an overlapping technique, where multiple ensembles are simulated in parallel. They are coupled by exchanging conformations such that the different microstates have defined probability distributions in multiple ensembles.[141] Temperature replica exchange (REMD) was first formulated for MD simulations in 1999, simulating a total of N_α replicas simultaneously at different temperatures (Figure 2.2).[89] It falls under the category of extended-ensemble algorithms, allowing for the exchange of conformation $\mathbf{r}_m$ of replica m with the conformation $\mathbf{r}_n$ of replica n. The different conformations achieve different probability distributions due to a different temperature in each replica, but sharing the same conformational phase space.[141] The probability $P(\mathbf{r}_\alpha)$ of a single conformation $\mathbf{r}_\alpha$ in replica α is given by:

$$P(\mathbf{r}_\alpha) = \frac{e^{-\beta_\alpha E(\mathbf{r}_\alpha)}}{Z_\alpha}, \tag{2.2.23}$$

obeying the Boltzmann distribution. It follows that the joint probability distribution of the extended ensemble P_{REMD} across N_α replicas ($\alpha = 1, ..., N_\alpha$) is given by the product of Boltzmann factors of each replica:[89]

$$P_{REMD} = \exp\left(-\sum_{\alpha}^{N_\alpha} \beta_\alpha E(\mathbf{r}_\alpha)\right) = \prod_{\alpha}^{N_\alpha} P(\mathbf{r}_\alpha), \tag{2.2.24}$$

with $\beta_\alpha = \frac{1}{k_B T_\alpha}$ indicating the different temperatures in each replica. Whether or not an exchange of conformations $\mathbf{r}_m$ and $\mathbf{r}_n$ between replica m and n is permitted, is evaluated by the individual transition probabilities of the forward $P(\mathbf{r}_m, \beta_m; \mathbf{r}_n, \beta_n)$ and reverse process

$P(\mathbf{r}_n, \beta_m; \mathbf{r}_m, \beta_n)$. Imposing detailed balance conditions for the exchange process on the transition probability $P(\mathbf{r}_m, \beta_m; \mathbf{r}_n, \beta_n)$ to converge towards an equilibrium distribution yields:[89]

$$P_{REMD}(X)P(\mathbf{r}_m, \beta_m; \mathbf{r}_n, \beta_n) = P_{REMD}(X')P(\mathbf{r}_n, \beta_m; \mathbf{r}_m, \beta_n), \tag{2.2.25}$$

with $X = (..., \mathbf{r}_m, \beta_m; \mathbf{r}_n, \beta_n, ...)$ and $X' = (..., \mathbf{r}_n, \beta_m; \mathbf{r}_m, \beta_n, ...)$. Since terms from replicas that are not exchanged can be dropped, rearrangement leads to:

$$\frac{P(\mathbf{r}_m, \beta_m; \mathbf{r}_n, \beta_n)}{P(\mathbf{r}_n, \beta_m; \mathbf{r}_m, \beta_n)} = \frac{P_{REMD}(X)}{P_{REMD}(X')} = e^{(-\Delta_{mn}(REMD))}, \tag{2.2.26}$$

with the exchange ratio:

$$\Delta_{mn}(REMD) = (\beta_n - \beta_m)(E(\mathbf{r}_m) - E(\mathbf{r}_n)). \tag{2.2.27}$$

Applying the Metropolis acceptance criterion finally yields the transition probability depending on the individual potential energies in replica m and n:

$$P(\mathbf{r}_m, \beta_m; \mathbf{r}_n, \beta_n) = \begin{cases} 1 & \text{for } \Delta \leq 0 \\ e^{-\Delta_{mn}(REMD)} & \text{for } \Delta > 0 \end{cases}, \tag{2.2.28}$$

giving values in the range from 0 to 1. It can be inferred that to obtain a sufficient exchange probability, the potential energy distributions of replica m and n must overlap. As the width of the energy distribution of each state gets smaller by a factor $N^{1/2}$ with increasing the number of atoms N, the spacing of temperatures across the replica ladder must become smaller, which means that the number of replicas to span the same temperature range needs to increase for larger simulation systems.[141]

Another replica-based approach similar to temperature replica exchange is Hamiltonian replica exchange (H-REMD), which was introduced three years later.[171] Instead of the temperature, the Hamiltonian is scaled over the replica ladder, representing a more general implementation, implying that REMD is only a special case. The probability $P(\mathbf{r}_\alpha)$ of a single conformation $\mathbf{r}_\alpha$ in replica α is similar to equation 2.2.23 given by:

$$P(\mathbf{r}_\alpha) = \frac{e^{-\beta E_\alpha(\mathbf{r}_\alpha)}}{Z_\alpha}, \tag{2.2.29}$$

with β being the same in all replica. The transition probability $P(\mathbf{r}_m, E_m; \mathbf{r}_n, E_n) = P(m \to n) = min\{1, e^{-\Delta_{mn}(H-REMD)}\}$ here depends on the energy in replica m and n:

$$\Delta_{mn}(H - REMD) = \beta[(E_m(\mathbf{r}_n) + E_n(\mathbf{r}_m)) - (E_m(\mathbf{r}_m) + E_n(\mathbf{r}_n))], \tag{2.2.30}$$

imposing the Metropolis acceptance criterion.

Replica exchange with solute scaling - REST 2

The problematic narrow spacing of replicas in (H-)REMD for large systems was tackled by Liu et al.[172] in 2005 introducing replica exchange with solute tempering (REST). This was further refined by Wang et al.[90] in 2011 leading to the updated replica exchange with solute scaling (REST2) scheme. Instead of scaling the whole system, only a smaller subsystem of the simulation box is influenced by the replica ladder, and also the exchange probability depends only on the energy of the subsystem. To be precise, the system is divided into a solute part that is going to be scaled, and a solvent part that is left unscaled. Often the solute part is a protein, peptide or, in this work, a glycan that should be enhanced sampled, whereas the solvent part is represented by the surrounding water molecules and ions (Figure 2.3). The potential energy in replica m of the system is subdivided into three parts that are differently scaled:

$$E_m^{REST2}(\mathbf{r}_m) = \frac{\beta_m}{\beta_0} E_{pp}(\mathbf{r}_m) + \sqrt{\frac{\beta_m}{\beta_0}} E_{pw}(\mathbf{r}_m) + E_{ww}(\mathbf{r}_m), \tag{2.2.31}$$

with E_{pp} being the potential energy of the solute (protein-protein interactions), E_{pw} the potential energy of the protein-water interactions and E_{ww} the solvent energy (water-water interactions). The scaling factor $\frac{\beta_m}{\beta_0}$ especially influences the intramolecular potential energy of the solute, depending on the ratio of the effective temperature T_m in replica m and ground temperature T_0 in the ground replica 0:

$$\frac{\beta_m}{\beta_0} = \frac{1/k_B T_m}{1/k_B T_0} = \frac{k_B T_0}{k_B T_m} = \frac{T_0}{T_m} = \lambda \tag{2.2.32}$$

with λ spanning the range between 1 for the ground temperature and 0 for an infinitely high temperature. In REST2 the Hamiltonian of the system is scaled by λ instead of directly scaling the temperature. However, it is often referred to a replica having an effective temperature, as a doubled temperature is equivalent to a halved energy[173]:

$$P(\mathbf{r}) \propto e^{-\frac{E(\mathbf{r})}{(2\cdot k_B T)}} = e^{-\frac{(E(\mathbf{r})/2)}{k_B T}}. \tag{2.2.33}$$

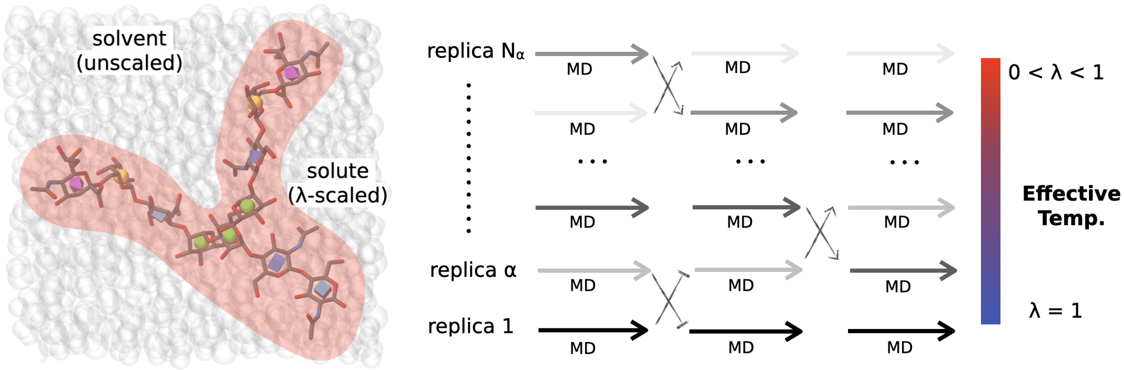


Figure 2.3: **Scheme of a replica exchange with solute scaling (REST2) MD simulation.** A complex model glycan (blue: GlcNAc, green: Man, yellow: Gal, purple: Neu5Ac) solvated in water (gray) is shown on the left, where the system is separated into a solvent (water and ions) and solute part (glycan atoms). Across the replica ladder the Hamiltonian of the solute part is scaled by a factor λ, which is equivalent to increasing the temperature and therefore termed effective temperature (solute part highlighted in red). Replicas are simulated in parallel, and coordinates are exchanged between neighboring replicas, based on the Metropolis criterium. λ is scaled geometrically across the replica ladder to facilitate barrier crossings along degrees of freedom that are stuck in energy minima.

Therefore, a scaling factor smaller than 1 is used to scale the potential energy of the solute part to facilitate the exploration of different conformations by lowering barriers lying between them. The acceptance ratio between replica m and n is given by the Metropolis acceptance criterion

$$P(m \to n) = min\{1, e^{-\Delta_{mn}(REST2)}\} \tag{2.2.34}$$

with

$$\Delta_{mn}(REST2) = (\beta_m - \beta_n)\left[(E_{pp}(\mathbf{r}_n) - E_{pp}(\mathbf{r}_m)) + \frac{\sqrt{\beta_0}}{\sqrt{\beta_m} + \sqrt{\beta_n}}(E_{pw}(\mathbf{r}_n) - E_{pw}(\mathbf{r}_m))\right] \tag{2.2.35}$$

only depending on the scaled energy terms E_{pp} and E_{pw}. Under the approximation that neighboring replica m and n have similar temperatures ($\beta_m \approx \beta_n$), the fluctuations of the reduced potential energy term in replica m,

$$E_{pp} + \frac{1}{2\sqrt{\frac{\beta_0}{\beta_m}}}E_{pw}, \tag{2.2.36}$$

determine the exchanges to replica n, where exchanges are performed under thermodynamic equilibrium. The implementation of choice used in this study[173], in particular, scales non-bonding force field terms, electrostatics (atom charges by $\sqrt{\lambda}$) and vdW (Lennard-Jones parameter ϵ by λ), and only torsion terms (by λ), as scaling of bonds and angles resulted in no beneficial effect.[90]

2.2.4 Metadynamics

Well-tempered metadynamics

Despite the generality of the above mentioned CV-independent methods, sometimes an explicit biasing of a specific CV is more effective, as it is assumed that the system or reaction of interest can be described by few reaction coordinates. Therefore, it is reasonable to use a non-overlapping, single replica approach, in which adaptive biasing potentials are employed, as such well-tempered metadynamics.[91] This is a refinement of the originally introduced metadynamics technique from Laio and Parinello in 2002[174], which is related to the density of state estimation method from Wang and Landau.[175] Assuming that the system of interest is metastable, it is necessary to define a CV $s(\mathbf{r})$ that can differentiate between the different minima, whose distribution can be described by:

$$P(s) \propto e^{-\frac{F(s)}{k_B T}}, \tag{2.2.37}$$

depending on the free energy F along s. The free energy profile $F(s)$ is not known *a priori* and needs to be estimated. Generally, a specific value of s can be estimated from the unbiased free energy function:

$$F(s) = -k_B T \ln \int \mathrm{d}\mathbf{r}\ \delta(s(\mathbf{r}) - s)e^{-\frac{E(\mathbf{r})}{k_B T}} + C, \tag{2.2.38}$$

with C as an arbitrary constant and the Dirac δ describing all the conformations corresponding to s.[176] As discussed, it is to be feared that under standard MD conditions the system is trapped in one of its metastable states, not able to cross existing barriers within the available simulation time. The desirable probability distribution $P(s)$ and the corresponding free energy profile $F(s)$ would not be correct. Therefore, an external bias $B(s)$ is applied in the CV space, allowing the system to emerge from minimum energy basins and explore all regions of the CV. The resulting biased free energy $F'(s)$ is related to the underlying unbiased profile:

$$F'(s) = -k_B T \ln \int d\mathbf{r}\ \delta(s(\mathbf{r}) - s) e^{-\frac{E(\mathbf{r})+B(s(\mathbf{r}))}{k_B T}} + C' = F(s) + B(s) + C' - C, \qquad (2.2.39)$$

showing that the bias needs just to be subtracted from $F'(s)$ to recover the unbiased free energy.[176]

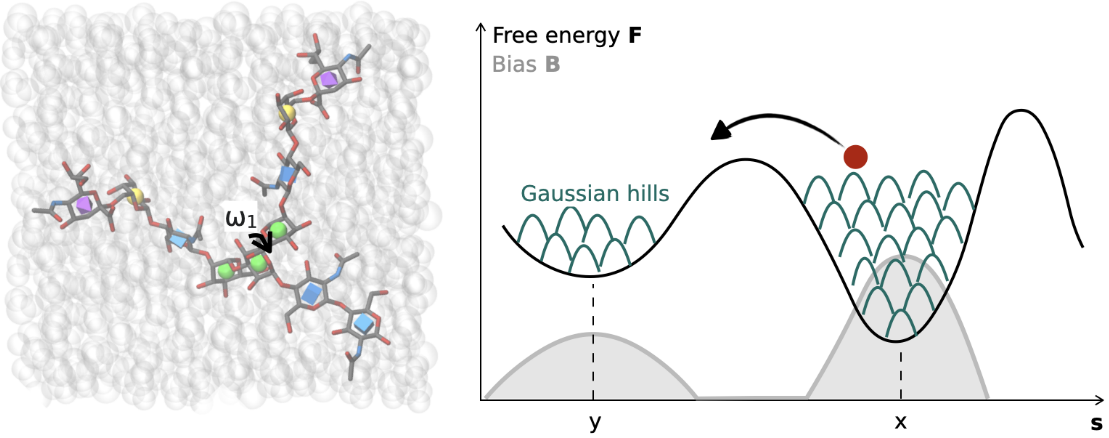


Figure 2.4: **Scheme of a metadynamics simulation.** A complex model glycan (blue: GlcNAc, green: Man, yellow: Gal, purple: Neu5Ac) solvated in water (gray) is shown on the left, where an ω torsion angle is selected as CV (black arrow). Along the progression of the simulation, a bias in the form of small Gaussians (green hills) is iteratively added to the force field torsion angle potential at its current position (red dot). Due to the increase of the energy level, a transition from state x to y via the barrier is facilitated. Inverting the sum of accumulated bias potentials (gray distribution) gives an approximation of the free energy landscape.

Generally speaking, the bias potential $B(s)$ is approaching an optimal shape if it is the negative of its corresponding free energy $B(s) = -F(s)$, ultimately flattening the energy landscape, making existing barriers disappear. Well-tempered metadynamics almost achieves this by an iterative procedure, where a history-dependent bias potential is built in the form of small Gaussians $e^{-\frac{(s-s(t'))^2}{2\sigma^2}}$ that are added in the CV space (Figure 2.4). They are positioned every τ_G time units with a width (σ) and height (w_G), where the bias potential $B(s,t)$ in the simulation is given by a sum of Gaussians:

$$B(s,t) = \sum_{t'=0,\tau_G,\ldots}^{t'<t} w \exp\left[\frac{B(s(t'),t')}{k_B \Delta T}\right] \exp\left[-\frac{(s-s(t'))^2}{2\sigma^2}\right], \qquad (2.2.40)$$

with the deposition rate $w = w_G/\tau_G$. The height of the Gaussian is scaled down by a factor $e^{\frac{B(s(t'),t')}{k_B \Delta T}}$, taking into account the bias potential at the same point where the Gaussian is supposed to be centered. ΔT is an input parameter that has temperature units and is explained more in detail below. In terms of an iterative procedure, this means that at each τ_G a new Gaussian is placed at the current position, with a height depending on the amount of Gaussians already deposited nearby. This prevents the underlying free energy landscape to be overfilled, pushing the system into nonphysical high free energy states, and allows the bias to converge over the simulation time.[176] The scaling of the Gaussian heights disrupts the flat histogram properties $B(s) = -F(s)$, implying that the sum $B(s) + F(s)$ no longer becomes flat but that both properties are rather connected in the long time limit through:

$$B(s, t \to \infty) = -\frac{\Delta T}{T + \Delta T}(F(s) - C(t)), \tag{2.2.41}$$

where the bias is only the fraction $\frac{\Delta T}{T+\Delta T}$ of the negative of the free energy.[91]

Combining the derivation of the bias potential with the distribution that is sampled when the bias is applied (equation 2.2.37 & 2.2.39), we obtain:

$$P'(s, t \to \infty) \propto e^{-\frac{F'(s)}{k_B T}} = e^{-\frac{F(s)+B(s,t)}{k_B T}} = e^{-\frac{F(s)}{k_B (T+\Delta T)}}. \tag{2.2.42}$$

It follows that the bias potential is equivalent to allowing s to be explored at an effective higher temperature $T + \Delta T$.[176] If $\Delta T \to \infty$, standard metadynamics is reproduced, whereas $\Delta T = 0$ implies standard unbiased sampling. During the setup of a metadynamics simulation, the width σ, the deposition pace τ_G, the initial height w_G and the bias factor $\gamma = \frac{T+\Delta T}{T}$, depending on the choice of ΔT, needs to be given.

Gaussians can only be placed in a low-dimensional space of the CV, ensuring the repetitive exploration of the same value s, only differing in the visited microstates $\mathbf{r}_k$. Repetitive exploration of the same values of s results in the repetitive additions of Gaussians to the potential, discouraging the system from visiting these conformations again and exploring new regions of the CV space. This approach would not be possible in the full conformational phase space, as the system would literally never explore the same point twice due to the high dimensionality. It is said that well-tempered metadynamics is limited to approximately biasing three CVs simultaneously, as the bias becomes multi-dimensional and its storage need scales exponential with the number of CVs.

Replica exchange with Collective-Variable Tempering - RECT

The limitation of well-tempered metadynamics to only a few number of CVs that need to be carefully chosen can be overcome by the introduction of a multi-replica approach, combining the overlapping technique H-REMD with non-overlapping well-tempered metadynamics. Replica exchange with Collective-Variable Tempering (RECT)[177] is able to enhance tenths of CVs simultaneously at the cost of simulating several replica of the system in parallel. In the standard well-tempered metadynamics formulation, the bias potential evolves according to the following equation of motion:

$$\dot{B}(s, t) = w \exp\left[\frac{B(s(t), t)}{k_B \Delta T}\right] \exp\left[-\sum_{z=1}^{N_{CV}} \frac{(s^z - s^z(t))^2}{2(\sigma^z)^2}\right], \tag{2.2.43}$$

where z iterates over a total number of CVs, N_{CV}, that are simultaneously biased in a multi-dimensional fashion, if desired. RECT simplifies this by enhanced sampling N_{CV}

degrees of freedom separately, via the application of concurrent one-dimensional history-dependent potentials, where z represents the index for each CV:

$$\dot{B}^z(s^z) = w \exp\left[\frac{B^z(s^z(t),t)}{k_B \Delta T}\right] \exp\left[-\frac{(s^z - s^z(t))^2}{2(\sigma^z)^2}\right]. \tag{2.2.44}$$

The evolution of the individual biases will depend on the marginal probability for each CV z:

$$P(s^z) \propto \int ds^1 ... ds^{z-1} ds^{z+1} ... ds^{N_{CV}} P(s^1, s^2, ..., s^{N_{CV}}), \tag{2.2.45}$$

flattening the distribution for each CV.[177] Selected CVs can however be correlated, as they are usually not orthogonal to each other. This raises the concern that applied concurrent, separate bias potentials are not acting just on a single CV but also affecting the distribution of others.[177] A self-consistent construction of the multiple one-dimensional bias potentials eliminates the effect of an additional effective bias that arises due to the correlation of CVs, ensuring that the marginal probability is flattened.[177] The degree of flatness is controlled by the bias factor γ, which is correlated to the boosting temperature ΔT applied to each CV as in well-tempered metadynamics. An unbiased sampling is achieved by setting $\Delta T = 0$, corresponding to $\gamma = 1$, where a flat histogram is obtained by $\Delta T = \infty, \gamma = \infty$.

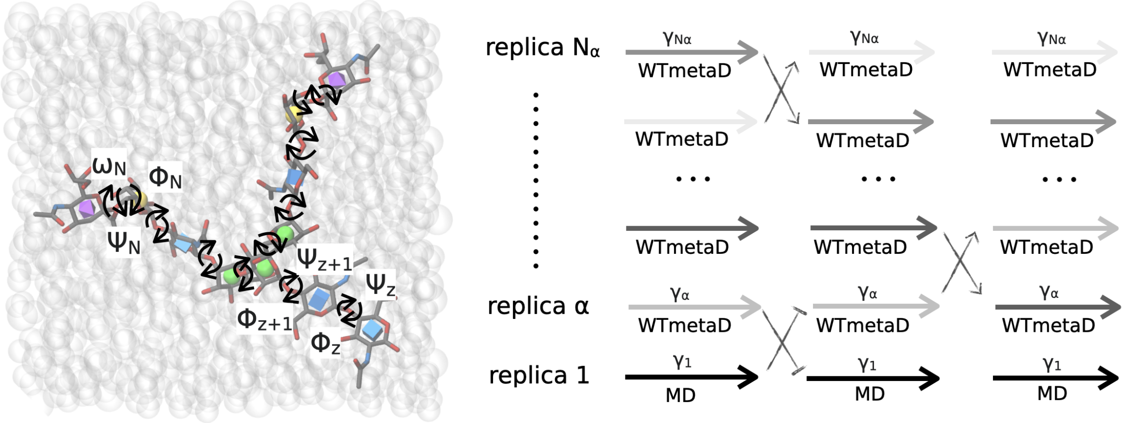


Figure 2.5: **Scheme of a replica exchange with collective-variable tempering (RECT) simulation.** A complex model glycan (blue: GlcNAc, green: Man, yellow: Gal, purple: Neu5Ac) solvated in water (gray) is shown on the left, where all torsion angles N are selected as CVs (black arrows). Similar to a temperature replica exchange simulation, replicas are simulated in parallel, however not the temperature but the bias factor γ of the well-tempered metadynamics (WTmetaD) scheme is scaled over the replica ladder. Along the progression of the simulation, one-dimensional bias potentials are iteratively added to each CV. The highest replica having theoretically ergodic sampling is connected to the unbiased ground replica ($\gamma = 1 =$ standard MD) via the replica ladder. Frequent exchanges of coordinates according to the Metropolis criterion allow for conformations explored in higher replicas to travel down to the ground replica, whose distribution is finally analyzed. The RECT approach combines the replica exchange scheme with well-tempered metadynamics to allow for the simultaneous explicit biasing of dozens of CVs.

Interpolating between these two extreme sampling conditions using a number of increasing γ values can connect ergodic sampling achieved by high γ values with unbiased conditions. This can be exploited in a Hamiltonian replica exchange fashion, scaling γ over the replica ladder and performing a low-dimensional concurrent metadynamics simulation in each replica (Figure 2.5). The ground replica ($\gamma = 1, \Delta T = 0$) represents a standard MD simulation, which is connected through frequent exchanges with higher replicas. These enhance the transition rates above relevant energy barriers trough the acting of multiple bias potentials on selected CVs. To be precise, each CV is biased in each replica, while the bias factor, i.e. the extent of biasing, is different in each replica. The acceptance probability between replica m and n ensuring detailed balance conditions is determined by $P(m \to n) = min\{1, e^{-\Delta_{mn}(RECT)}\}$ with

$$\Delta_{mn}(RECT) = \frac{\sum_z B^z_m(s^z_n) + \sum_z B^z_n(s^z_m)}{k_B T} - \frac{\sum_z B^z_m(s^z_m) + \sum_z B^z_n(s^z_n)}{k_B T}. \qquad (2.2.46)$$

This is equivalent to the exchange probability of H-REMD, with the only difference that exchanges are dependent on the sum of bias potentials instead of on directly modified Hamiltonians. The application of concurrent well-tempered metadynamics to a large number of local CVs ensures importance sampling, while a small set of critical CVs is not necessarily needed to be known *a priori*. There are also other metadynamics-based sampling techniques employing a replica ladder, such as multiple walker, parallel-tempering metadynamics or bias-exchange metadynamics.[178–180] None of these methodologies were ever applied to the study of glycans before.

2.2.5 Tackling the flexibility of N-glycans

The enhanced sampling of N-glycans has already been performed in order to explore the diversity of three-dimensional structures that can be achieved, as their rotation around many torsion angles complicates an accurate prediction of the conformational phase space. The prediction of a correct distribution of glycan conformations in MD simulations becomes a significant task as more and more proteins and membranes are modeled, taking their modifications by the attachment of chemical groups into account. Only focusing on a correctly folded protein is not enough anymore, since especially glycans have shown to be important interaction partners in cellular environments and their conformations decisive for the interaction.[77,134] As outlined already in the introduction, several enhanced sampling techniques have so far failed to capture a converged conformer distribution for different N-glycan types when simulated in an unbound fashion free in solution.[87,93–97]. The difficulties of employing standard MD, Hamiltonian REMD and well-tempered metadynamics for the sampling of glycans will be shown in the following example, using the complex N-glycan A2G2S2 as a model system (Figure 2.6). It consists of four GlcNAc, three Man, two Gal and two Neu5Ac residues, harboring one ω torsion angle that is known to adopt several different conformations, corresponding to multiple energy minima.

When assessing the suitability of a method, it is crucial to monitor the sufficient exploration of all energy minima of all degrees of freedom that contribute to the phase-space exploration. In terms of glycan sampling, this suggests that in particular the rotation around torsion angles needs to be verified, but also the puckering of individual monosaccharides. The transition between energy minima is therefore compared for the representative torsion angles ω_{3-8} and ψ_{1-2} as well as the puckering variable θ of one monosacchride in the glycan structure. Standard MD of A2G2S2 was performed at 310.15 K for 50 ns without any restraints, using the CHARMM36 force field for the glycan atoms and TIP3P as a water model. In the application of H-REMD, 48 replicas were simulated over an effective temperature range of 310.15 K to 500 K for 50 ns. Replica exchanges were attempted every 400 steps with average exchange probabilities of around 10 %. The well-tempered metadynamics simulation was based on the standard MD conditions with a bias on the torsion angle ω_{3-8} with the following parameters: $\gamma = 14, \tau = 500\,\text{steps}, \sigma = 0.35, w_G = 4$.

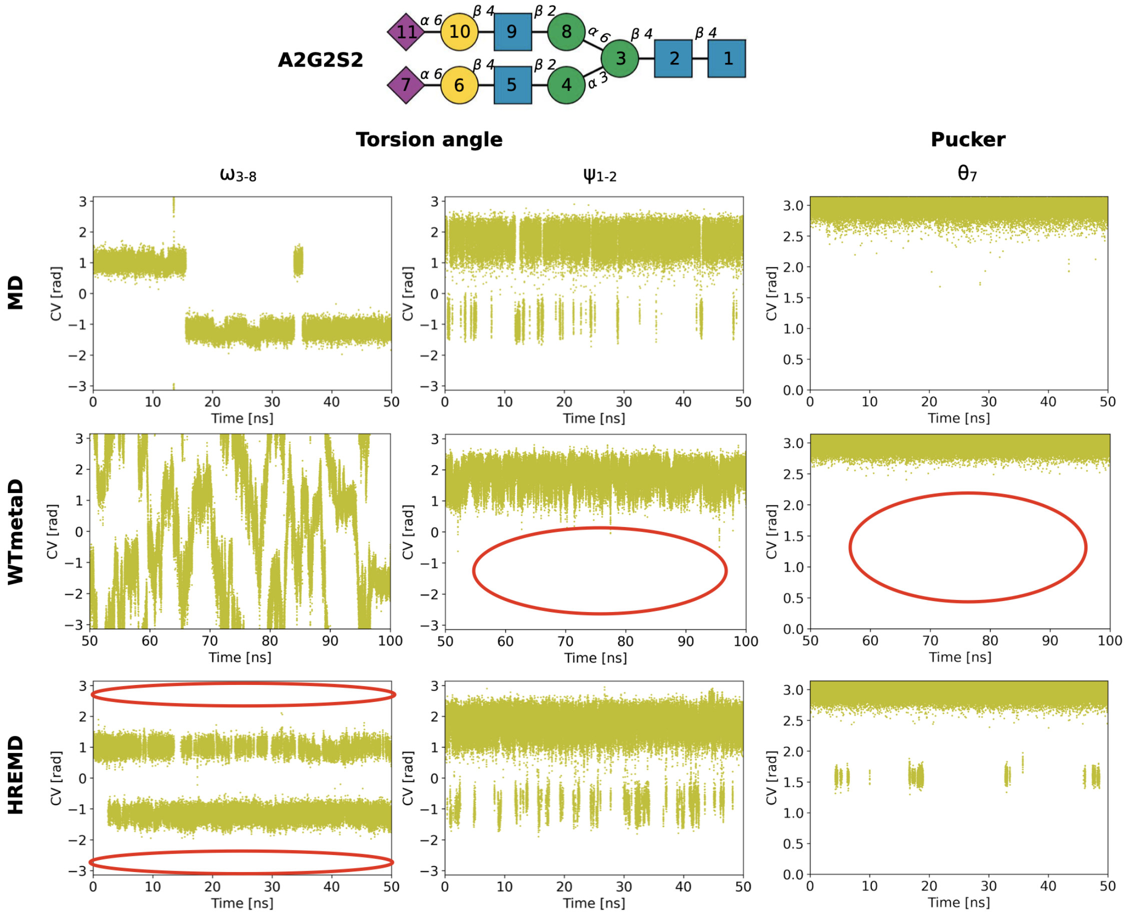


Figure 2.6: **Variable progression under different sampling schemes.** A complex model glycan called A2G2S2 (blue: GlcNAc, green: Man, yellow: Gal, purple: Neu5Ac) was simulated separately under three different sampling conditions: standard MD, well-tempered metadynamics (WTmetaD) biasing the ω_{3-8} torsion angle and H-REMD, scaling the Hamiltonian to achieve an effective temperature up to 500 K in the highest replica. Only few or no transitions could be seen along the torsion angle ω_{3-8} (connecting saccharide units 3 and 8) and pucker coordinate θ_7 (of saccharide units 7) under unbiased conditions. The explicit sampling of ω_{3-8} via well-tempered metadynamics evokes multiple transitions in the CV space, whereby the exploration of ψ_{1-2} and θ_7 is restricted. The simulation time of 50 to 100 ns is shown (instead of 0 to 50 ns), since the adaptive biasing requires some time to evolve to its full potential. The application of H-REMD allows for the exploration of another minimum in θ_7, although ω_{3-8} is lacking full exploration. Red circles indicate regions of missing exploration.

Along the progression of the MD simulation, only few transitions can be found for the flexible ω_{3-8} torsion angle and exploration of only one minimum for θ_7 (Figure 2.6). The ψ_{1-2} torsion angle is more often fluctuating between two states at -1 and 2 rad. Applying well-tempered metadynamics to ω_{3-8} allows for the exploration of the whole CV space along ω_{3-8}, with multiple transitions between states (Figure 2.6), at least for the explicitly biased torsion angle. The transitions along ψ_{1-2} are limited to the global minimum compared to standard MD, a side effect of the biasing of ω_{3-8}, due to which only conformers that reside within one minimum of ψ_{1-2} are sampled. Further, no exploration of the puckering along θ_7 except for the global minimum can be observed in the metadynamics run. In contrast, H-REMD is able to let the system escape the energy minimum along θ_7 by sampling another mininum around 1.5 rad (Figure 2.6). However, the exploration of ω_{3-8} is suffering as only the two major energy basins are visited. Since well-tempered metadynamics does not explore different puckering conformations and unbiased torsion angles, whereas pure replica exchange methods like H-REMD suffer from poor convergence along multiple torsion angles[87], the combination of both methodologies implemented in RECT is suggested. Instead of only biasing the most dominating CVs in the system, all torsion angles should be explicitly sampled via the application of RECT. Furthermore, the number of replicas required to scale within a certain temperature range can be reduced by employing REST2 instead of H-REMD, where only the glycan atoms are defined as the solute region and therefore subjected to a scaled Hamiltonian, leaving the water atoms at ground temperature.

REST-RECT

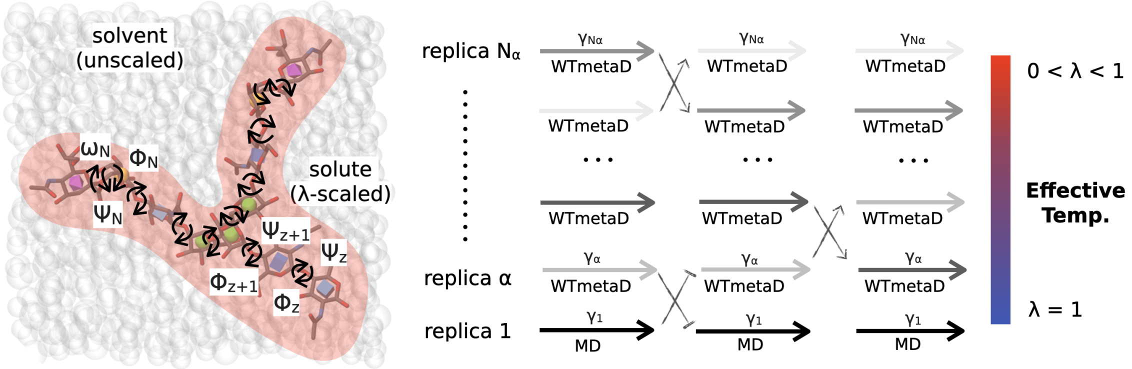


Figure 2.7: **Scheme of a REST-RECT simulation.** A complex model glycan (blue: GlcNAc, green: Man, yellow: Gal, purple: Neu5Ac) solvated in water (gray) is shown on the left, where all torsion angles N are selected as CVs (black arrows). The RECT methodology is applied to all CVs, whereas all glycan atoms are additionally defined as the solute region for sampling via the REST2 algorithm. Over the replica ladder not only the bias factor γ of well-tempered metadynamics is scaled, but also λ of the REST2 scheme acts on the Hamiltonian of the glycan atoms. This combination does not only allow for the explicit biasing of selected CVs (here torsion angles) but also of all other degrees of freedom of the glycan, e.g. puckering of saccharide units.

As the name of the algorithm already suggests, REST-RECT is a combination of the two replica exchange methodologies REST2 and RECT. It combines scaling the solute part of a system by an altered Hamiltonian with concurrent well-tempered metadynamics, where both replica exchange algorithms share the same replica ladder (Figure 2.7). In detail, both λ, scaling the Hamiltonian of the solute, and γ, influencing the bias potentials of the selected CVs, are scaled over the replica ladder simultaneously. By choosing $\gamma = 1$ and $\lambda = 1$ for the ground replica, standard and unbiased MD conditions are ensured, which allows for the direct evaluation of the sampled probability distribution. Using a geometric progression for the increase of γ values and decrease of λ values over the replica ladder, the system becomes highly biased in higher replicas. Especially the increase of γ will gradually flatten the marginal distribution of selected CVs over the replica ladder, where the decrease of λ enhances all degrees of freedom in the solute part and ultimately provides sampling of unidentified CVs that are crucial for the full exploration of the system. Sampling of an ergodic probability distribution from the unbiased ground replica is achieved by exchanges of conformations across the replica ladder, where conformations explored in the most ergodic replica can travel down to the ground replica.

The acceptance probability is composed of contributions from the REST2 and RECT algorithm by

$$P(m \to n) = min\{1, e^{-\Delta_{mn}(REST-RECT)}\} \tag{2.2.47}$$

with:

$$\begin{aligned}\Delta_{mn}REST - RECT = (\beta_m - \beta_n)\left[(E_{pp}(\mathbf{r}_n) - E_{pp}(\mathbf{r}_m)) + \frac{\sqrt{\beta_0}}{\sqrt{\beta_m} + \sqrt{\beta_n}}(E_{pw}(\mathbf{r}_n) - E_{pw}(\mathbf{r}_m))\right] \\ + \frac{\sum_z B^z_m(s^z_n) + \sum_z B^z_n(s^z_m)}{k_BT} - \frac{\sum_z B^z_m(s^z_m) + \sum_z B^z_n(s^z_n)}{k_BT}.\end{aligned} \tag{2.2.48}$$

This formula is very similar to the one from the first application of solute tempering metadynamics, where replica exchange with solute tempering was combined with metadynamics in a similar fashion to enhance the exploration of the free energy surface of the protein G helix by Carlo Camilloni et al.[181] Only the scaling of the E_{pw} term is altered, being the major difference between solute tempering and solute scaling, while the single bias potential of metadynamics is replaced by the sum of potentials from RECT.

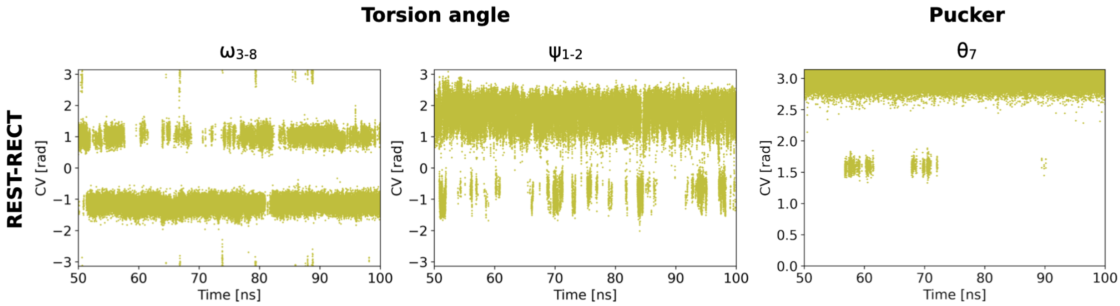


Figure 2.8: **Progression of CVs in a REST-RECT simulation.** When applying the REST-RECT scheme to the complex glycan from Figure 2.6, the algorithm is able to sample along torsion angles and pucker coordinates. The progression along ω_{3-8} and θ_7 shows the repetitive exploration of different CV states. The simulation time of 50 to 100 ns is shown (instead of 0 to 50 ns) as the adaptive biasing requires some time to evolve to its full potential.

The complex N-glycan A2G2S2 introduced above can also be enhanced sampled via REST-RECT, choosing a λ scaling from 1 to 0.4, representing a temperature increase from 310.15 K to 800 K over 12 replicas. Due to the reduced number of replicas necessary to span the same temperature range as compared to H-REMD, we decided to extend the temperature range up to 800K in order to sample the puckering even more effectively. The metadynamics parameter γ was scaled from 1 in the ground replica to 14 in the highest replica, with all torsion angles serving as CVs. A simulation time of 500 ns per replica with exchange attempts every 400 steps between neighboring replicas was chosen. Transitions between states can be observed for both torsion angles as well as puckering coordinates already within 50 ns (Figure 2.8). The sampling of all torsion angles and puckering conformations via REST-RECT should therefore ensure the complete exploration of the conformational phase space for glycan structures. The still rather few transitions observed for the puckering coordinate can be further validated and refined using the sampling of conformations also in higher replica through the Weighted Histogram Analysis Method described below.

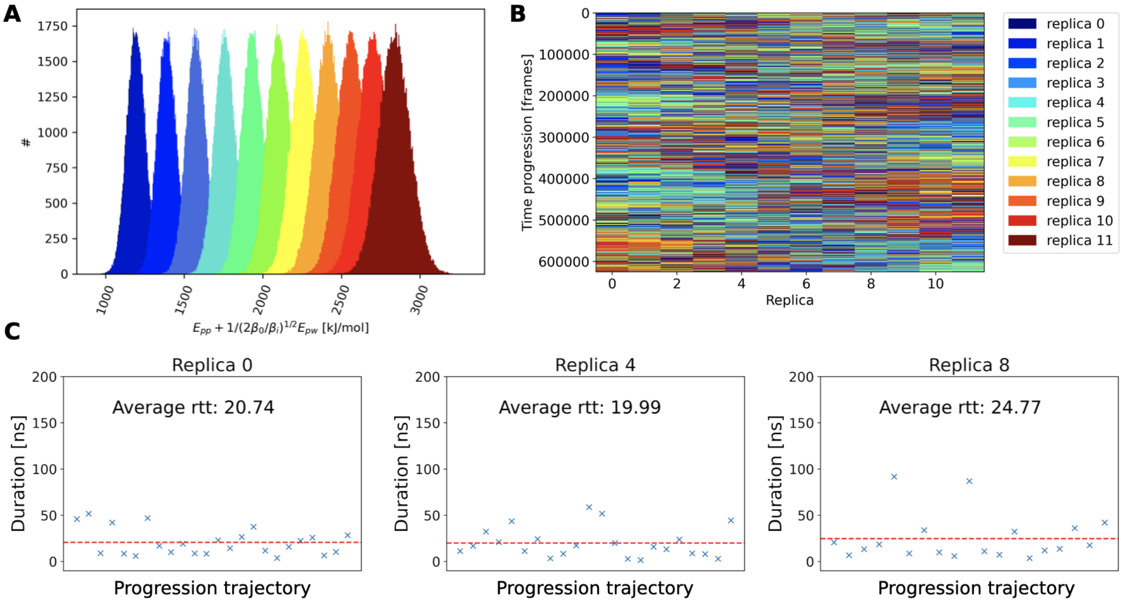


Figure 2.9: **Assessing the performance of a REST-RECT simulation.** The enhanced sampling of glycan A2G2A2 via REST-RECT needs to be validated in terms of sufficient replica exchanges, employing the following quantities: **A** Histogram of the potential energy distribution within each replica, considering only energy terms that are also contributing to the exchange probability. **B** Mixing of replica indices over the time progression of the trajectory. A well intermixed, wide distribution of colors represents many successful replica exchanges. **C** Round trip times for representative replicas with the x-axis showing the progression over time, plotted against the duration of each round trip in ns. Every blue cross indicates the duration of one round trip, whereas the red dotted line marks the average over all recorded round trips in that replica.

A critical point for all replica exchange methods is the assessment of frequent replica exchanges through a sufficient overlap of potential energy distributions in each replica. There are several parameters that should be evaluated, giving a hint for exchanges and coherently indicating convergence of the simulation. First of all the most obvious quantity, the histogram of potential energies in each replica according to equation 2.2.36, can be plotted, as it majorly determines the success of exchanges (Figure 2.9 **A**). It is important to not plot the potential energy of the whole system but only the contributing parts of E_{pp} and E_{pw}. When overlaps of histograms from neighboring distributions can be observed, it indicates that replica exchanges are theoretically possible. The actual exchanges per attempted swap can be visualized plotting the replica indices over the simulation time for the multiple simulations run in parallel, graphically indicating a proper mixing of replicas by a colorful plot (Figure 2.9 **B**). This visual assessment can be supplemented with the calculation of replica round trip times, defined as the time each replica requires to travel up and down the ladder, completing one whole round trip (Figure 2.9 **C**). Each round trip in the different replicas took around 20 ns, which resulted in approximately 25 round trips per simulation, with 10 being sufficient as a rule of thumb (Personal communication from Giovanni Bussi). Last but not least, the exchange probabilities are calculated along the progression of the simulation, where values over 30 % indicate a good exchange rate but still do not necessarily guarantee a sufficient number of round trips for each replica.

2.2.6 WHAM

When running a replica-exchange simulation, often solely the unbiased ground replica is analyzed for its distribution along a collective variable $s(\mathbf{r})$, discarding lots of additional sampling time achieved in higher replicas, due to the bias that is imposed on these distributions. It is possible to reweight these biased distributions $P'(s)$ with the application of the Weigthed Histogram Analysis Method (WHAM)[182], obtaining the unbiased distribution $P(s)$ using a weight $w(s)$:

$$P(s) \propto w(s)P'(s). \tag{2.2.49}$$

It is important to recognize the effect of the bias potential on the individual distributions in each replica α, which are altered according to:

$$P^\alpha(s) \propto e^{-\frac{E(\mathbf{r})+B^\alpha(s(\mathbf{r}))}{k_BT}} = e^{-\frac{B^\alpha(s(\mathbf{r}))}{k_BT}} P(s). \tag{2.2.50}$$

As the different replicas are subjected to different bias potentials $B^\alpha(s(\mathbf{r}))$, the corresponding free energy is given by:

$$F^\alpha(s) = -k_BT \ln P^\alpha(s) = F(s) + B^\alpha(s(\mathbf{r})) + C^\alpha, \tag{2.2.51}$$

with $F(s)$ being the unbiased free energy and C^α a constant that is different in each replica α, ensuring normalization of the biased probability. Rearranging the above equation results in the expression of the unbiased free energy $F(s)$:

$$F(s) = -k_BT \ln P^\alpha(s) - B^\alpha(s) + C^\alpha. \tag{2.2.52}$$

According to the name of the method, the aim is to construct a weighted histogram from which an unweighting is performed to derive $P(s)$. In order to do so, s is divided into a total number of N_{grid} bins, where n_j^α is the number of frames that fall into the jth bin in replica α. The probability $P(n_1^\alpha, n_2^\alpha, ..., n_{N_{grid}}^\alpha)$, is described by the product of multinomial distributions:

$$P^\alpha(n_1^1, ..., n_{N_{grid}}^1, n_1^2, ..., n_{N_{grid}}^{N_\alpha}) = \prod_\alpha \frac{N_{grid}!}{\prod_{j=1}^{N_{grid}} n_j^\alpha!} \prod_{j=1}^{N_{grid}} (P_j^\alpha)^{n_j^\alpha} = \prod_\alpha \frac{N_{grid}!}{\prod_{j=1}^{N_{grid}} n_j^\alpha!} \prod_{j=1}^{N_{grid}} (P_j w_j^\alpha)^{n_j^\alpha} \tag{2.2.53}$$

with P_j being the unbiased distribution in bin j, P_j^α the biased distribution in bin j and replica α and w_j^α the corresponding weight. The weight consists of the bias potential b_j^α and the normalization constant C^α:

$$w_j^\alpha = b_j^\alpha C^\alpha = e^{-\frac{B_j^\alpha}{k_BT}} \cdot C^\alpha, \tag{2.2.54}$$

where frames with a high bias potential are weighted less. The variational parameters P_j and C^α in equation 2.2.53 have to be determined under the constrain of $\sum_j P_j^\alpha b_j^\alpha C^\alpha = 1$, which can be done by finding a set of values for P_j that maximizes the likelihood of observing the trajectory.[183] This can be performed using a set of Lagrange multipliers where, after some manipulation, the WHAM equations emerge:

$$P_j = \frac{N_j}{\sum_\alpha N^\alpha b_j^\alpha C^\alpha} \tag{2.2.55}$$

and

$$C^\alpha = \frac{1}{\sum_j P_j b_j^\alpha}, \tag{2.2.56}$$

where $N^\alpha = \sum_j n_j^\alpha$ and $N_j = \sum_\alpha n_j^\alpha$. It becomes apparent that P_j and C^α can not be explicitly solved as they are dependent on each other. However, a self-consistent scheme can be applied, starting with $C^\alpha = 1$ and recursively iterating until the values of P_j and C^α are not changing anymore. In the application of WHAM, the chronological sequence of replicas is not important, as all conformations from all replicas are counted in the *jth* bin $N^\alpha = \sum_j n_j^\alpha$. Therefore, the individual trajectories of each replica can be concatenated and the weight for each frame i results from:

$$w_i = \frac{1}{\sum_\alpha C^\alpha w_i^\alpha}, \tag{2.2.57}$$

with w_i^α being the weight in each replica α for frame i.

2.3 Reducing dimensions

Note: Notations in this section are mainly chosen with reference to the publication of Helfrecht et al. 2020.[184]

MD provides an extreme amount of data as all coordinates and velocities for each atom of the system can theoretically be saved every 1 or 2 fs, providing information about the conformation and structural arrangement in a time-dependent manner. Considering that simulated systems harbor on average between 10’000 (solvated glycan) to 500’000 atoms (solvated protein), it is hard to decipher which essential features contribute to the observations of interest in such structurally complex systems. Here one can appreciate the value of the already mentioned collective variables, differentiating between conformational states of a molecule and following transitions over time. Hence, the identification of suitable CVs is not only crucial for CV-dependent methods, but also directly related to the interpretation of any kind of MD simulations, as they also allow for a low-dimensional representation of the system.[141] Unlike crystalline solids or clusters of identical atoms, biomolecules mostly have a complex energy landscapes that is far from being symmetrical.[185] It can be taken advantage of the fact that the conformational phase space of most biological systems of interest is characterized by a few high probability areas, along with several metastable states, whereas the rest of the space is factually never sampled and the probabilities tend to be zero.[141] Therefore, CVs are desired which describe the accessible conformations in a projected fashion, spanning a new low-dimensional map.

In the context of simulating N-glycan structures, there is a large number of possible CVs that can be used in order to define the conformational states of the system, where a detailed structural analyses requires the rationalization of this high-dimensional vector space. Dimensionality reduction techniques can help with an efficient graphical representation of the glycan's conformational phase space but can also provide information about the most important structural features to differentiate between conformations. Such projections possibly reveal mutual functional dependencies and hidden correlations among the many CVs.

The outlined problem can be addressed by different machine-learning algorithms, taking the molecular trajectory as a set of high-dimensional vectors and performing a data reduction operation. It should be kept in mind that all of these algorithms result in a projection with a lower information content, keeping only the necessary features to describe the important states of a system. The underlying mathematical models for the different dimensionality reduction techniques have in common that they impose certain assumptions on the high-dimensional data. For instance, it is assumed that sampled data points from a trajectory cluster around few representative three-dimensional structures or that all adoptable conformations lie on a linear or non-linear low-dimensional manifold.[185] The different obstacles going hand in hand with the application of certain dimensionality reduction techniques will be discussed explicitly for the methods applied in this study.

The data set to be analyzed with a dimensionality reduction algorithm should ideally be derived from an unbiased MD simulation, where a series of random high-dimensional vectors $\{\mathbf{r}_t\}$, sampled every t time points, represent the conformations with coordinates $\mathbf{r}$, sampled from the distribution $P(\mathbf{r})$. If, however, enhanced sampling algorithms are employed for the simulation, the system will sample from the biased distribution $P'(\mathbf{r})$, which can be reweighted to $P(\mathbf{r})$ through the calculation of a weight w_t for each frame according to equation 2.2.57. Under the assumption of simulating a free N-glycan in solution, there are different components included in one coordinate framework $\mathbf{r}_t$ that are of different importance to the quantity of interest, which is in this case the conformation of the N-glycan. For instance, the position of water atoms can be generously neglected, but also the pure Cartesian coordinates of the sugar atoms probably hold little information about the three-dimensional glycan structure, as thermal fluctuations make the data set noisy. Chemical intuition can help in this context, considering that especially torsion angles should be able to describe large motions in glycan structures, as it is also true for proteins with their backbone angles.[185] Consequently, it is important to reduce the noise and dimension of the data set *a priori* to enhance the possibility of obtaining an informative projection, as it is nothing else than an illustration of how the random vectors $\{\mathbf{r}_t\}$ are distributed in relation to each other in the low-dimensional space. The notation $\{\mathbf{r}_t\}$, spanning the phase space based on Cartesian coordinates is replaced by $\{\mathbf{X}_i\}$, describing the high-dimensional vector set comprised of all torsion angles (ϕ, ψ and ω) present in a glycan structure, iterating over i frames.

2.3.1 Principle component analysis

There are multiple strategies that have been developed for MD trajectories in recent years to map a high-dimensional feature matrix $\mathbf{X}$ of shape $n_{samples} \times n_{features}$ onto a low-dimensional latent-space matrix $\mathbf{T}$. Their applicability mainly depends on the underlying data structure.[186] In the case of N-glycans, only Principle Component Analysis (PCA)[187] has so far been used to reduce the dimensions of the conformational phase space, however taking Cartesian coordinates as input and therefore being less informative.[188] We focus here on various algorithms differing in their linearity and assumption about the high-dimensional data structure, that should be tried in order to obtain a reasonable low-dimensional projection of glycan conformations. For instance, PCA projects the data onto the linear eigenvector space defined by the k largest eigenvalues obtained by diagonalization of the covariance matrix $\mathbf{C}$ of $\mathbf{X}$. The low-dimensional representation in form of a latent-space matrix $\mathbf{T}$ with shape $n_{samples} \times n_{PCA}$ is defined by:

$$\mathbf{T} = \mathbf{X}\mathbf{P}_{XT}, \tag{2.3.1}$$

where $\mathbf{P}_{XT}$ projects between feature and latent space. $\mathbf{P}_{XT}$ can be estimated from the eigenvalue decomposition of the covariance matrix $\mathbf{C} = \mathbf{X}^{\mathbf{T}}\mathbf{X}$:

$$\mathbf{C} = \mathbf{U_C}\mathbf{\Lambda_C}\mathbf{U_C^T}, \tag{2.3.2}$$

where $\mathbf{\Lambda_C}$ contains the eigenvalues on the diagonal in decreasing order and $\mathbf{U_C}$ the corresponding eigenvectors as columns. The projection matrix $\mathbf{P}_{XT} = \hat{\mathrm{U}}_{\mathbf{C}}$, where $\hat{\mathrm{U}}_{\mathbf{C}}$ only contains the desired top k eigenvectors (the principal components). The highest eigenvalues correspond to the features of maximum data variance and the corresponding eigenvectors are used as axes for e.g. two-dimensional or three-dimensional graphs representing the low-dimensional projection. It is further possible to perform the reverse projection $\mathbf{P}_{TX}$, namely approximating $\mathbf{X}$ in terms of $\mathbf{T}$:

$$\mathbf{X}_{PCA} = \mathbf{T}\mathbf{P}_{TX}, \tag{2.3.3}$$

to assess the performance of the projection. The resulting reconstruction error l is calculated from the difference of the original feature matrix and the approximated:

$$l = ||\mathbf{X} - \mathbf{X}_{PCA}||^2 = ||\mathbf{X} - \mathbf{X}\mathbf{P}_{XT}\mathbf{P}_{TX}||^2, \tag{2.3.4}$$

with $|| \ ||$ being the Frobenius norm.

2.3.2 Diffusion map

In contrast, Diffusion map[189] is a non-linear method, where the connectivity (or diffusion distance) between individual data points (in our case, individual N-glycan conformations) is quantified by the likelihood of transitioning from one to the other, expressed with the help of a diffusion kernel function. Data points of $\mathbf{X}$ are projected onto a two-dimensional matrix $\mathbf{T}$ so that the diffusion distances in the high-dimensional feature space can be approximated by Euclidean distances between points in the reduced space. This ensures the preservation of the local vector-space structure, so-called isometric embedding. In this work, Diffusion maps were computed based on the same algorithm as described by Bottaro et al.[190], using a Gaussian kernel. First, pairwise euclidean distances between frames i and j for all $n_{samples}$ are calculated in the $n_{feature}$-dimensional space, yielding the redundant square matrix $\mathbf{D}$ of shape $n_{samples} \times n_{samples}$. Subsequently, the adjacency matrix $\mathbf{K}$ is computed with Gaussian kernel:

$$K_{ij} = \exp\left(\frac{-\mathbf{D}_{ij}^2}{2\sigma^2}\right), \tag{2.3.5}$$

with σ defining the size of the neighborhood. It considers only three-dimensional structures that are significantly similar, ensuring that the transition probability is small for dissimilar three-dimensional structures. A transition matrix T is constructed iteratively over t iterations:

$$T_{(t+1)} = d_{(t)}^{-1/2} T_t d_{(t)}^{-1/2}, \tag{2.3.6}$$

with d being the diagonal degree matrix of the Gaussian kernel and $T_{(0)} = K$, making it normalized and symmetric[190] in contrast to the original implementation of Diffusion maps.[189] Additionally, this yields results equivalent to the recently introduced bi-stochastic kernel method.[191] It follows that the $n_{samples} \times n_{samples}$ shaped transition matrix T_{ij} gives the probability of a direct transition from sample i to j. An eigenvalue decomposition of T as in the case of PCA is yielding the highest eigenvalues and corresponding eigenvectors, defining the latter as the diffusion components that span the low-dimensional projection.

2.3.3 Sketch-map

The non-linear Sketch-map algorithm[192,193] is based on a different approach named multidimensional scaling (MDS).[194] It does not reproduce an isometric mapping but was developed in order to focus on a projection that includes the most relevant information, discarding additional information about the deposition of points in high-dimensions.[186] The mutual distance of data points in $\mathbf{X}$ is conserved in $\mathbf{T}$ by the application of a sigmoid function focusing on the reproduction of intermediate distances, rather than far-away distances which are dominated by the topology of the high-dimensional space. In practice, the following stress function is to be minimized to produce a mapping:

$$\chi^2 = \left(\sum_{j \neq i} w_i w_j\right)^{-1} \sum_{j \neq i} w_i w_j [F(R_{ij}) - f(r_{ij})]^2, \tag{2.3.7}$$

with w_i being the weight of point i, R_{ij} and r_{ij} the distances between points i and j in high- and low-dimensional space, respectively.[193] The distances are transformed by two sigmoid functions F and f, for the high-dimensional and low-dimensional space, of form:

$$s(r) = 1 - (1 + (2^{(a/b)} - 1)(r/\sigma)^a)^{-b/a}, \tag{2.3.8}$$

where a and b determine the rate at which the functions approach 0 or 1 and σ the switching distance.[193] In F and f the distance σ is the same. Their values are close to 0 when r is smaller than σ and tend towards 1 if r is greater than σ. This ensures that the algorithm pays most attention to points that are close to the parameter σ, which should be determined from analyzing the histogram of pairwise distances R_{ij}, identifying intermediate distances that could correspond to stable conformations. The values a and b are chosen differently for F and f, termed a_D, b_D for F and a_d, b_d for f. For larger sample sizes, sketch-map is not applied to the whole data set but the initial projection is done on selected landmark points, which can be identified through e.g. random sampling or farthest point sampling. The data subset is arranged in the low-dimensional space and the remaining data points projected using an out-of-sample embedding. This is sometimes required, as in the case of sketch-map, since the computational expense of many algorithms scales quadratically or cubically with the number of input points.[186]

2.3.4 Kernel principle covariates regression

Finally, beside these unsupervised techniques, there are also supervised algorithms, which make use of information included in an additional property matrix $\mathbf{Y}$ of shape $n_{samples} \times n_{properties}$. A recent example is the kernel principal covariates regression model (kPcovR)[184], which combines the robust methods of linear regression and PCA. Linear regression aims at determining a linear relation between the input features $\mathbf{X}$ and target properties $\mathbf{Y}$, essentially finding a weight $\mathbf{P}_{XY}$ that minimizes the error of the reconstruction

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{P}_{XY}. \tag{2.3.9}$$

As $\mathbf{P}_{XY}$ minimizes the loss

$$l = ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = ||\mathbf{Y} - \mathbf{X}\mathbf{P}_{XY}||^2, \tag{2.3.10}$$

it can be obtained after matrix rearrangements:

$$\mathbf{P}_{XY} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}, \tag{2.3.11}$$

with $\mathbf{I}$ being a unit vector. λ is a regularization parameter that, when set greater than zero, adds noise to the data in order to reduce overfitting and is hence called ridge regression. The loss in ridge regression extends to

$$l = ||\mathbf{Y} - \mathbf{X}\mathbf{P}_{XY}||^2 + \lambda||\mathbf{P}_{XY}||^2. \tag{2.3.12}$$

For the application of principal covariates regression, linear regression and PCA are combined using a parameter α in order to modulate the weight of each method, corresponding to a combined loss

$$l = \alpha||\mathbf{X} - \mathbf{X}\mathbf{P}_{XT}\mathbf{P}_{TX}||^2 + (1-\alpha)||\mathbf{Y} - \mathbf{X}\mathbf{P}_{XT}\mathbf{P}_{TY}||^2, \tag{2.3.13}$$

with $\mathbf{P}_{XT}\mathbf{P}_{TY} = \mathbf{P}_{XY}$.[184] It tries to find a low-dimensional representation of $\mathbf{X}$ that simultaneously reduces the information loss and error in predicting $\mathbf{Y}$, by diagonalising a modified covariance matrix. The projection matrix becomes:

$$\mathbf{P}_{XT} = \mathbf{C}^{-1/2}\hat{\mathbf{U}}_{\tilde{\mathbf{C}}}\hat{\mathbf{\Lambda}}_{\tilde{\mathbf{C}}}^{1/2}, \tag{2.3.14}$$

with $\tilde{\mathbf{C}}$ being an augmented version of $\mathbf{C}$ and $\hat{\mathbf{\Lambda}}$ a trunctated version of $\mathbf{\Lambda}$.[184] On top, the kernel trick can be applied, finally yielding kPcovR using a radial basis function (Gaussian kernel):

$$k(\mathbf{X} - \mathbf{X}') = \exp(-\gamma||\mathbf{X} - \mathbf{X}'||^2), \tag{2.3.15}$$

for pairs of samples $\mathbf{X}, \mathbf{X}'$. The hyperparameter γ can uncover non-linear relations between samples, improves the latent space projection $\mathbf{T}$, and increases the regression performance.[184] For a more detailed derivation, especially of the Kernel methods, we refer to the original publication of Helfrecht et al. 2020.[184]

2.4 Dependence of experiments and simulations

Note: *Parts of this section are taken from the publication: S.M. Ayala Mariscal, M.L.Pigazzini, Y. Richter, M. Özel, I.L. Grothaus, J. Protze, K. Ziege, M. Kulke, M. ElBediwi, J. Vermaas, L. Colombi Ciacchi, S. Köppen, F. Liu, J. Kirstein, Identification of a HTT-specific binding motif in DNAJB1 essential for suppression and disaggregation of HTT, Nature Communications 13, 4692, 2022.*[195]

The following section aims at highlighting the importance and significance of performing experimental studies in conjunction with computational modeling techniques. Our Huntington study, which revealed atomistic insights into important protein-peptide interactions by interweaving many different scientific approaches, will serve as a motivating example without much methodological detail.

The Huntington's disease (HD) is a neurodegenerative disorder that is induced by the expansion of a glutamine stretch in the first exon of the protein huntingtin (HTT)1. The large ubiquitous protein is trimmed by alternative splicing and caspase-mediated cleavages, yielding the N-terminal first exon of HTT that contains a pathological polyQ expansion (HTTExon1).[196] Above a threshold of Q $\geq$ 39, the disease is fully penetrant with an inverse correlation between the polyQ length and the age of onset. The severity of the disorder depends directly on the polyQ length. The HTTExon1 peptide forms amyloid fibrils whenever an expanded polyQ stretch is present, leading to aggregation of fibrils in both the cytoplasm and nucleus of HD patient's neurons.[197] HTTExon1 consists of a highly conserved N-terminal stretch of 17 amino acids (N17), the polyQ domain that facilitates amyloid formation and a C-terminal proline-rich domain (PRD), composed of two stretches of proline repeats (P1 and P2) (Figure 2.10 **A** left).

Suppression of the pathogenic HTTExon1Q_{48} could be achieved by means of a chaperone complex that consisted of the heat shock protein Hsc70, a class II J-domain protein DNAJB1, and the nucleotide exchange factor Apg218. In order to gain insight into the interaction between HTTExon1Q_{48} and the three chaperones, experimental collaborators performed non-specific cross-linking mass spectrometry that revealed an interaction between the HTT peptide and DNAJB1. DNAJB1 could be shown to bind to the second proline stretch (P2) of the PRD with its C-terminal domain (CTD) and more precisely with the hinge region between CTDI and CTDII involving a Huntingtin binding motif (HBM)(Figure 2.10 **A**). They could pinpoint the interaction with HTTExon1Q_{48} to 9 amino acids in the DNAJB1 sequence that also harbor two positively charged residues, K242 and H244, where in particular the latter one is highly conserved. Using a fluorescence resonance energy transfer (FRET) assay, it was possible to study the fibrilization of HTT, where the addition of the trimeric chaperone complex, Hsc70, DNAJB1 and Apg2 could suppress HTTExon1Q_{48} fibrilization for > 20 h (Figure 2.10 **B** magenta curve). However, substitution of H244 by alanine (H244A) completely abrogated the ability of DNAJB1 to suppress HTT fibrilization together with Hsc70 and Apg2 (Figure 2.10 **B** green and purple curves (DNAJB1H244A + Hsc70 and Apg2)). Our collaborators concluded that mutating H244 by alanine severely limits the ability of the trimeric chaperone complex (DNAJB1, Hsc70, and Apg2) to suppress HTTExon1Q_{48} fibrilization, and this defect could also not be rescued by increasing the concentration of the DNAJB1 variant (Figure 2.10 **B** purple curve).

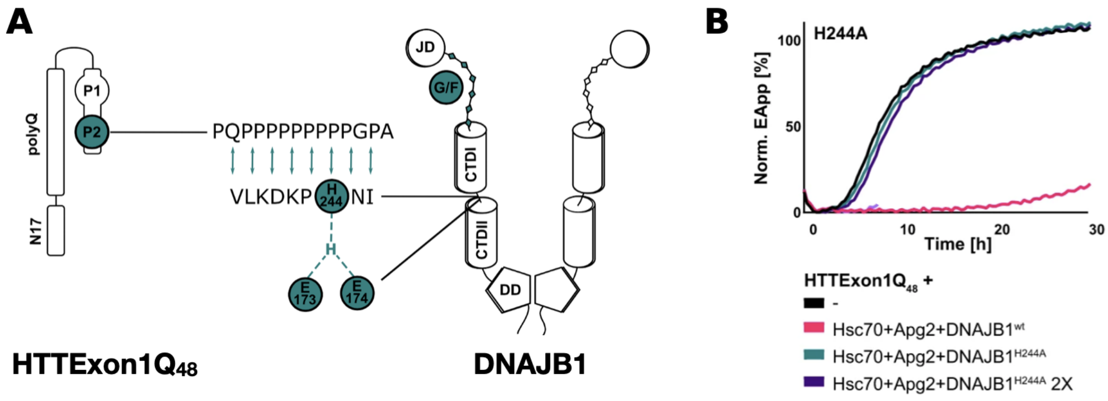


Figure 2.10: **Experimentally determined Huntingtin binding motif (HBM) of DNAJB1 for the Huntingtin peptide Exon 1 (HTTExon1Q_{48}).** **A** Schematic peptide structure of HTTExon1Q_{48}, having the N17 domain at its N terminus, followed by the polyQ region and two polyP regions. The proposed binding motif within DNAJB1 lies between the CTD1 and CTD2 domain, binding to the second polyP region (P2) of HTTExon1Q_{48}. **B** FRET measurements as a readout of HTTExon1Q_{48} aggregation over time in the absence (black curve) and presence of Hsc70, Apg2 and DNAJB1wt (magenta) or variants (green/purple). Figure adopted from Ayala et al. 2022.[195]

Atomistic details and a mechanistic understanding of the conserved binding interface were however so far lacking and prevented a complete understanding of the protein-peptide complex formation. To structurally analyze the complex formed between DNAJB1wt or DNAJB1H244A and HTTExon1Q_{48} computationally, we first needed to predict, refine and then dock a HTTExon1Q_{48} model to DNAJB1. As there is little structural information available for HTTExon1Q_{48}, an initial starting structure was predicted by the homology modeling algorithm I-TASSER[198] serving as input structure, followed by a more intensive refinement via the enhanced-sampling TIGER2h algorithm, performed by Martin Kulke.[199,200]. Representative structures from five identified minimum free energy clusters of HTTExon1Q_{48} were docked to DNAJB1wt, employing the rigid-docking server HDOCK. Only two clusters lead to a reasonable complex formation with DNAJB1wt, involving the HBM and P2 in the binding interface (see for further details Supplementary Figure 3c in Ayala et al. 2022[195]). The two assemblies were further subjected to 500 ns of MD simulation, forming stable complexes with DNAJB1wt (see WT model and contact map in Figure 2.11). Major contacts of amino acids of the P2 of the PRD were formed to

P243, H244, N245, I246, and K248 of DNAJB1wt. Due to hydrogen bonding between E173 and the H244 side chain, only the backbone atoms of H244 are facing P2 of HTTExon1Q_{48} (see enlarged WT molecular model in Figure 2.11). We then docked the minimum structures of HTTExon1Q_{48} to the DNAJB1 mutation variant DNAJB1H244A. Interestingly, the clusters soon detached and moved away from the HBM of DNAJB1H244A, leading to a very faint contact map (Figure 2.11). Due to the missing hydrogen bond between residues 173 and 244, the alanine side chain at position 244 rotated outwards, preventing its backbone to firmly bind to HTTExon1Q_{48}. In summary, our in silico analysis fully supports the experimental finding that the P2 region of HTTExon1Q_{48} forms stable contacts with the HBM of DNAJB1, including residues 238–246 (Figure 2.11). H244 plays a fundamental role in stabilizing the complex via hydrogen bond interactions with its backbone, as observed in the WT model to HTTExon1Q_{48}. Substitutions of H244 by alanine break these bonds and destabilize the complex, explaining the inability of DNAJB1H244A to suppress HTT fibrilization together with Hsc70 and Apg2.

This thematic detour away from glycosylations has clearly demonstrated the far-reaching possibilities that MD simulations provide to interpret experimental results and use them as a starting point for deeper insights. Whenever possible, this mindset was applied in the following result chapters.

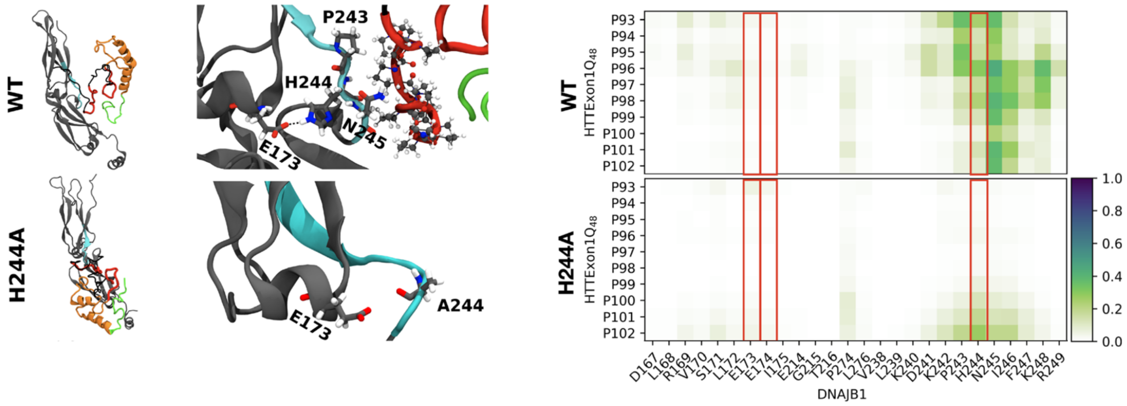
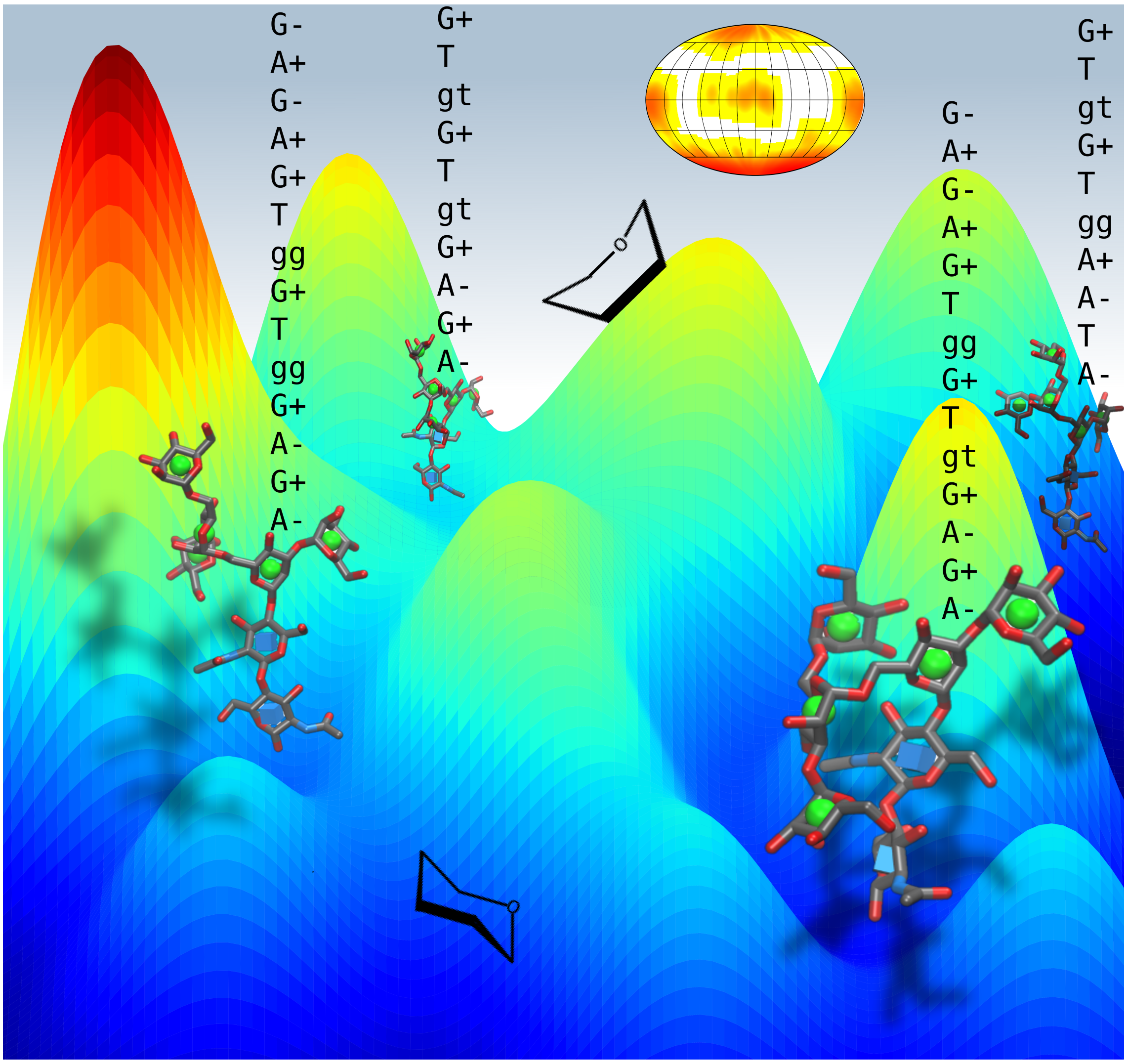


Figure 2.11: **Atomistic simulations explain experimental findings.** MD snapshots of monomeric DNAJB1wt and DNAJB1H244A in complex with HTTExon1Q_{48}(left). Atomistic structure of the HBM of DNAJB1wt or DNAJB1H244A and P2 of HTTExon1Q_{48}, interacting via hydrogen bonds (dotted lines) or non-bonding interactions (middle). DNAJB1 (gray) with highlighted amino acids with a coloring code according to the atom types: hydrogen (white), carbon (cyan), oxygen (red), nitrogen (blue). The HBM is colored in cyan. The domains of HTTExon1Q_{48} are indicated by different colors: polyQ (orange), PRD (P1 and P2: red; residues between P1 and P2: black), N17 (green) and highlighted amino acids in CPK style. Contact maps of DNAJB1wt or DNAJB1H244A and HTTExon1Q_{48}, focus on the HBM and P2 domain (right). Contact maps were constructed by averaging the contacts over the last 400 ns of the MD simulation and additionally averaging over interactions recorded for cluster 1 and 2 of HTTExon1Q_{48} with DNAJB1. Distances were converted by a rational switching function, defining the contact distance (inflection point) at 1 nm, where a value of 1 represents a close contact and 0 no contact. Residues 173, 174, and 244 are highlighted with a red box. Figure adopted from Ayala et al. 2022.[195]

3 | On their own: The phase space



The artistic phase space of N-glycans. Diverse N-glycan conformers of M5 bathing in their free energy landscape, flagged by corresponding conformer labels. They are illuminated by a puckering sun imposing distortions like chair and boot conformations on each saccharide ring.[137]

Note: *Parts of this chapter are taken from the publication: I.L. Grothaus, G. Bussi, L. Colombi Ciacchi, Exploration, representation and rationalization of the conformational phase space of N-glycans, Journal of Chemical Information and Modelling, 62(20):4992–5008, 2022.*[137]

The three main general limitations of MD simulations discussed in chapter 2 also directly apply to the simulations of carbohydrates. In this chapter, possible improvements based on enhanced sampling and dimensionality reduction techniques are proposed and tested to counteract these shortcomings. First the phase space spanned by Cartesian coordinates to describe the conformation of glycans is replaced by a new conformer description that is only dependent on torsion angles, under the assumption that these are the main degrees of freedom. This description is exploited to assess the performance of the enhanced sampling technique REST-RECT, evaluating the convergence of phase-space exploration and simultaneously tackling the problem of limited simulation times. The resulting high-dimensional output data set is further reduced by dimensionality reduction techniques in order to help the interpretation of the obtained results, aiming at a low-dimensional comparison of global glycan conformers that has not been possible so far. The outlined workflow is finally applied to a set of diverse applications comprising the comparison of different N-glycan structures and the performance of current biological force fields.

3.1 Testing

3.1.1 GlyCONFORMER

As detailed in section 1.3, N-glycans are multi-branched structures, characterized by the specific linkages between saccharides monomers. Each glycosidic linkage gives rise to at least two torsion angles (ϕ and ψ), while 1→6 and 2→6 linkages harbor an additional torsion angle ω. Based on these structural characteristics, we constructed an unambiguous labeling scheme to distinguish different conformers of the same N-glycan. The scheme is also applicable to other glycans, independently of their size, number or type of branches and amount of substituents such as fucosylation. Each conformer is identified by a digit string of length N_z, equal to the number of torsion angles in the glycan. For N-glycans, the string begins at the free reducing end, consisting of a β $1 \rightarrow 4$ - linked GlcNAc dimer followed by a mannose residue. For each linkage, the linear string reports digits assigned to ϕ, ψ and ω (if applicable), in this order. In correspondence of a junction (leading e.g. to an α 1→6 and a α 1→3 branch after the first mannose), a string separator is introduced, labeled according to the C atom at the branch origin (e.g. **6**– for 1→6 linkages). The string continues first along the branch of the higher C atom (6 in our case) until reaching the terminal residue, prior to returning to the last junction and following the branch of the next-lower C atom (3 in our case). Additional modifications like the attachment of fucose residues or bisecting GlcNAc residues are included after all other branches are assigned. The separators of primary branches are labelled with bold numbers (**6**– or **3**–), for clarity. The string digits indicate in which intervals of values the torsion angle lies, following the IUPAC nomenclature for dihedrals[201]. Namely, the digits for ϕ and ψ and the corresponding interval of radian values are :

$\mathrm{C} = [-0.52,\ +0.52)$

$\mathrm{G}_{+} = [+0.52,\ +1.57)$

$\mathrm{A}_{+} = [+1.57,\ +2.62)$

$\mathrm{T} = [+2.62, \pi]$ or $[-\pi,\ -2.62)$

$\mathrm{A}_{-} = [-2.62,\ -1.57)$

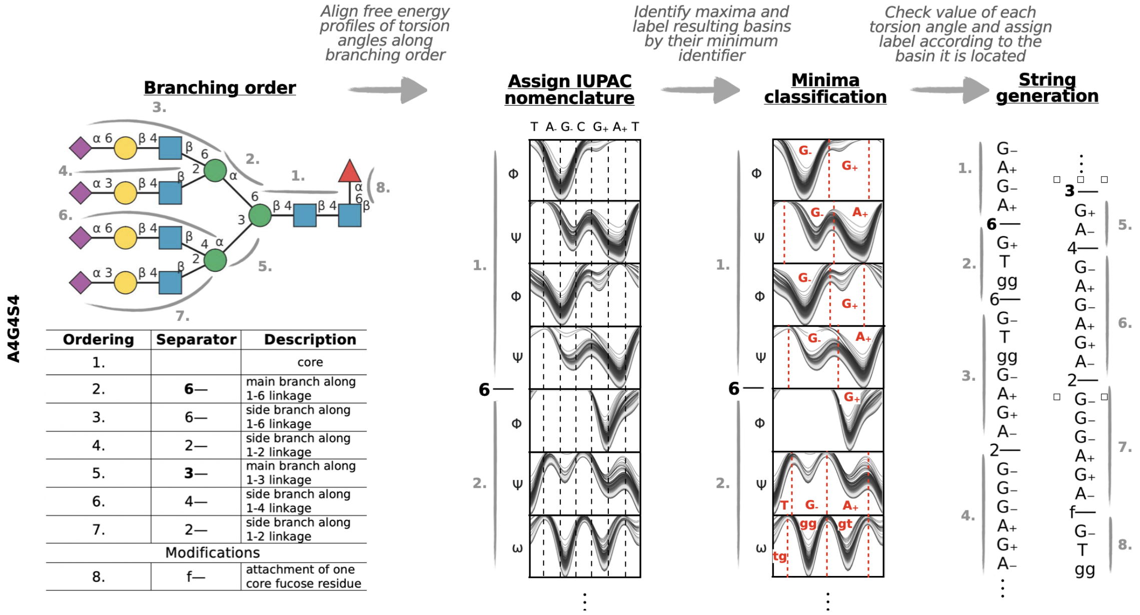
$\mathrm{G}_{-} = [-1.57,\ -0.52)$.

The digits for ω are:

$\mathrm{gg} = [-2.62,\ 0)$

$\mathrm{gt} = [0,\ 2.62)$

$\mathrm{tg} = [2.62,\ \pi]$ or $[-\pi,\ -2.62)$.


Align free energy profiles of torsion angles along branching order
Identify maxima and label resulting basins by their minimum identifier
Check value of each torsion angle and assign label according to the basin it is located
Branching order
Assign IUPAC nomenclature
Minima classification
String generation
A4G4S4
α 6
β 4
β
6
2
α
α 3
β 4
3
4
α
β 4
β 4
β
α
6
1.
2.
3.
4.
5.
6.
7.
8.
T A₋ G₋ C G₊ A₊ T
Φ
Ψ
Φ
Ψ
Φ
Ψ
ω
6—
1.
2.
G₋
G₊
A₊
T
gg
gt
tg
G₋
A₊
G₋
A₊
6—
G₊
T
gg
6—
G₋
T
gg
G₋
A₊
G₊
A₋
2—
G₋
G₋
G₋
A₊
G₊
A₋
⋮
□ □ □
3—
G₊
A₋
4—
G₋
A₊
G₋
A₊
G₊
A₋
2—
□ G₋ □
G₋
G₋
A₊
G₊
A₋
f—
G₋
T
gg

Ordering	Separator	Description
1.		core
2.	**6—**	main branch along 1-6 linkage
3.	6—	side branch along 1-6 linkage
4.	2—	side branch along 1-2 linkage
5.	**3—**	main branch along 1-3 linkage
6.	4—	side branch along 1-4 linkage
7.	2—	side branch along 1-2 linkage
Modifications		
8.	f—	attachment of one core fucose residue

Figure 3.1: **Example of conformer string generation.** Assignment of a conformer string to the complex N-glycan A4G4S4 (blue: GlcNAc, green: Man, yellow: Gal, purple: Neu5Ac, red:Fuc) drawn with DrawnGlycan[50], based on its torsion angle conformations. The various branches of a glycan are ranked according to their type of linkage and calculated free energy profiles of the torsion angles aligned in the corresponding order. Their free energy minimum basins are labeled with respect to the IUPAC nomenclature for torsion angles. For each adopted conformation along a trajectory, the value of each torsion angle in the string is evaluated, assigned to a minimum and correspondingly labeled. The outcome is an ordered string of digits that represents the global conformation of a glycan.

The assignment of each torsion angle to a given interval is performed in the following way. First, the free-energy profile associated with rotation along the torsion angle is calculated from an MD trajectory (most often enhanced sampling MD is necessary in order to achieve converged profiles). The positions of the free-energy minima are then labeled according to the nomenclature above. All angles belonging to the same free-energy basin (around a minimum between the two neighboring maxima) are finally labeled equally to the minimum of their basin (Figure 3.1). As a last step, each recorded set of torsion angles from a frame of a trajectory is translated into a conformer string, built according to the rules above. The scanning of free energy profiles for minima and maxima and the subsequent assignment according to the IUPAC nomenclature are a tedious task if done by hand. Therefore, the workflow was automatized in the python package GlyCONFORMER, categorizing free energy profiles according to the scheme in Figure 3.1. The package can also read in a feature matrix $\mathbf{X}$ of shape $n_{samples} \times n_{features}$, where the features are equal to all torsion angles of the analyzed glycan, and convert the torsion angle values of each frame into the respective conformer strings. This data set can be subsequently used to construct a histogram with the individual conformers selected as bins, yielding a conformer distribution (see below, e.g. Figure 3.3). In order to assess statistical features of this distribution, block averaging can be performed, separating the data set into evenly distributed blocks. The average of all blocks $\bar{X} = \frac{1}{N}\sum_{j=1}^{N} X_j$ is calculated over $N = 10$ blocks, where X_j is the average calculated within each jth block. Error bars are calculated as standard deviations of the sampling distribution (standard error of the mean): $\text{std}(\bar{X}) = \sqrt{\frac{\text{var}(\bar{X})}{N}}$, with the variance of the sampling distribution $\text{var}(\bar{X}) = (\frac{N}{N-1})\left[\frac{1}{N}\sum_{j=1}^{N} X_j^2 - (\frac{1}{N}\sum_{j=1}^{N} X_j)^2\right]$. To further analyze the convergence of conformer populations, moving averages can be calculated for the individual conformers over the simulation time. The outlined analysis represents a fundamental basis for the assessment of glycan simulations, as the conformer string incorporates all degrees of freedom that are necessary to describe the global glycan conformation. The approach is applied to the evaluation of REST-RECT simulations of various N-glycans in the following sections.

In chapter 2 the phase space has been defined as $\mathbf{\Gamma} = \{\mathbf{r}_k\}$ being the $3N$-dimensional phase space of N atoms that is spanned by the Cartesian coordinates of the simulation box. From here on, the conformational phase space expression is redefined and adapted for the application to glycans' three-dimensional structure. Namely, $\mathbf{\Gamma} = \{\mathbf{z}\} = \{\phi_i, \psi_i, \omega_i\}_{i=1,\ldots,N_{linkages}}$, where $\mathbf{z}$ is the vector of all N_z torsion angles present in a certain glycan structure, and ϕ_i, ψ_i, ω_i describing the specific torsion angles of each linkage i. Consequently, the $3N$-dimensional space is reduced to N_z dimensions, which normally range between 10 and 20 for typical N-glycans. Therefore, a microstate corresponds to a certain sequence of torsion angle values, while we define a conformer as a cluster of conformations having all the same conformer strings.

3.1.2 REST-RECT simulations

The proposed REST-RECT methodology was applied to a set of six biologically relevant N-glycans, namely three high-mannose types, M5 FM5 M9, and three complex types, A2G2 A2G2S2 and A4G4S4 (Figure 3.2).

Their three-dimensional structures were constructed using the CHARMM-GUI Glycan Modeller, based on averaged three-dimensional structures from the Glycan Fragment Database[202–205]. Each N-glycan was solvated with a 15 Å thick water layer in a cubic box. For A2G2S2 and A4G4S4, two and four sodium counter-ions were added, respectively, to compensate for the net negative charges resulting from Neu5Ac. MD simulations were performed with the GROMACS code, version 2018.4[206], patched with the PLUMED package, version 2.6[207]. The CHARMM36[150–152] force field was used for the N-glycan molecules in combination with the CHARMM-modified TIP3P water model (mTIP3P)[208]. For the CHARMM36 force field, the recent correction to a previous faulty implementation affecting in particular the ring inversion properties of Neu5Ac has been applied in all simulations[209]. The leap-frog algorithm was used as an integrator with a 2 fs time step and the LINCS algorithm[210] was employed to constrain bonds connected to hydrogen atoms. Temperature control was realized via velocity rescaling[211] using a time constant of 0.1 ps, setting a reference temperature of 310.15 K. The pressure was set to 1 bar with a compressibility of 4.5 x 10^{-5} bar^{-1}, and kept constant via the Parrinello-Rahman barostat with a time constant of 5 ps. The Verlet list scheme[212] was employed with a neighbor list updated every 80 steps. The calculation of electrostatic interactions was done with the Particle Mesh Ewald (PME)[213] method using a cut-off distance of 1.2 nm for the real space contribution.

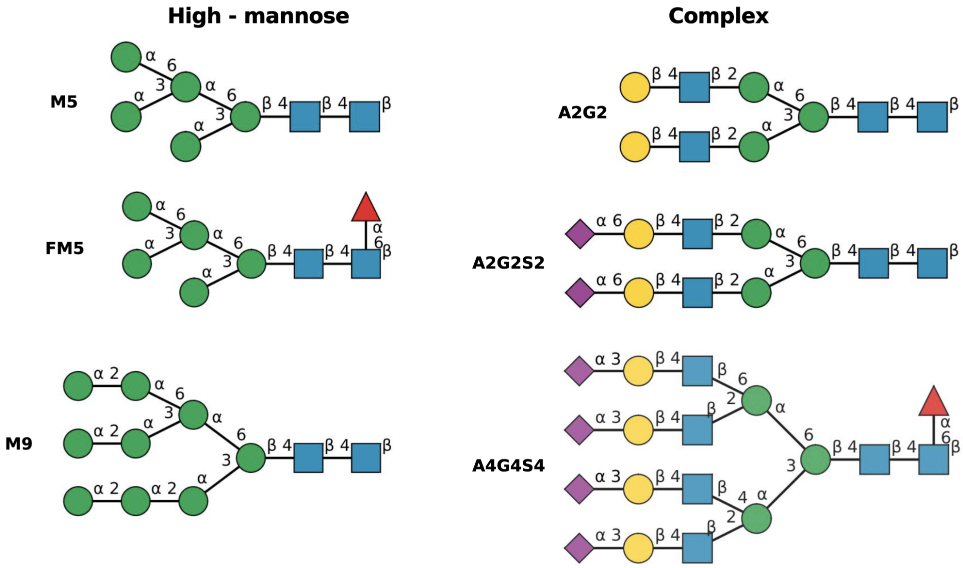


Figure 3.2: **Relevant N-glycan structures.** Model systems employed in this study, namely three high-mannose type *N*-glycans (M5, FM5, M9) and three complex *N*-glycans (A2G2, A2G2S2, A4G4S4). They have been drawn with DrawnGlycan[50] where GlcNAc is blue, Man is green, Gal is yellow, Neu5Ac is purple and Fuc is red.

The following steps were performed in order to equilibrate the systems: First, an energy minimization of water and ions was performed using the steepest-descent algorithm with a tolerance of 1000 kJ mol^{-1} nm^{-1}, restraining the N-glycan atoms. Then, the solvent was equilibrated in one NVT and one NpT run, each lasting 1 ns, with restrained N-glycan atoms. After that, a second energy minimization of all atoms was performed with no constraints, with the same parameters as before. Finally, unrestrained NVT and NpT equilibration runs were performed, lasting 1 ns and 100 ns, respectively. For each simulated system, two different starting conformers, named s1 and s2, were generated by setting the ω angle of the **6**– branch either to a gt (for s1) or a gg (for s2) conformation. Separate simulations starting with the two initial conformations s1 and s2 were performed to validate the convergence of the conformer distributions for the different simulation techniques employed.

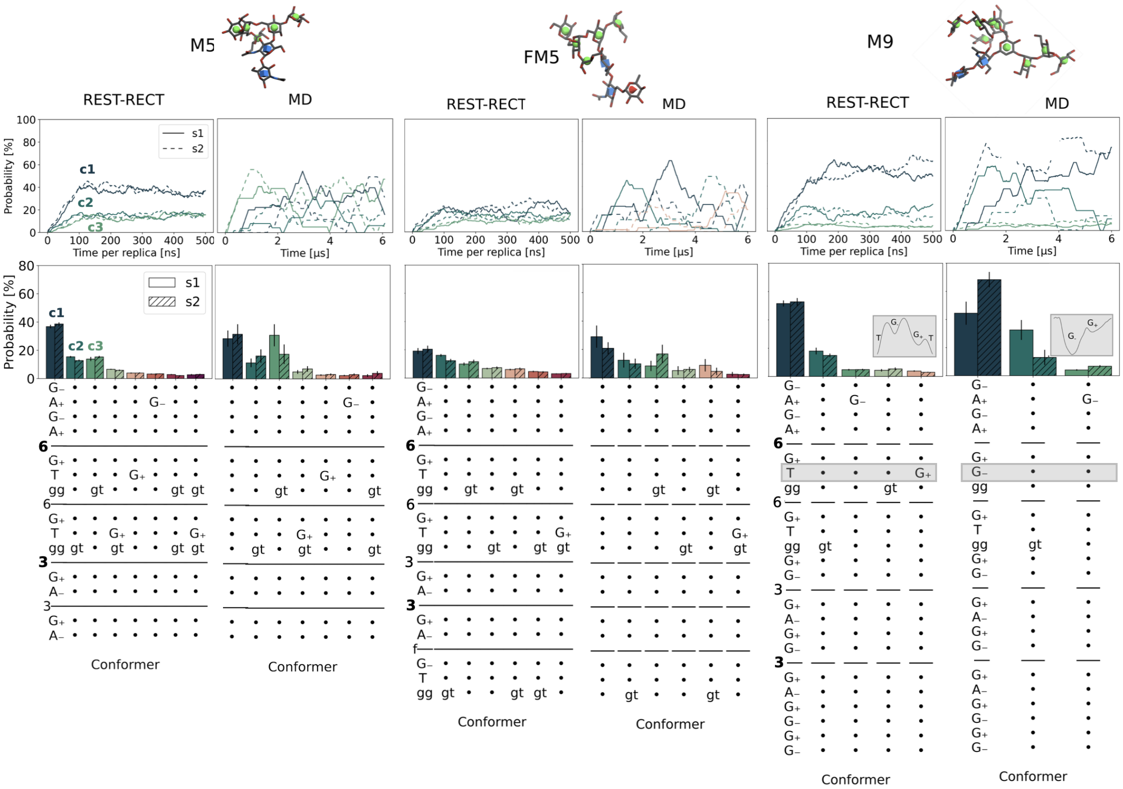


Figure 3.3: **Performance of REST-RECT for high-mannose type N-glycan conformational exploration.** Comparison of REST-RECT with plain MD simulations for systems M5, FM5 and M9 using the CHARMM36 force field. The upper panel includes atomistic structures of each glycan, visualized with the 3D-SNFG tool of VMD[214]. The middle panel shows the moving average for the three most populated conformer clusters using a window size of 100 ns (REST-RECT) and 1.2 μs (MD), corresponding to the same sampling time. Two separate simulations were performed with differing initial starting conformations (s1 and s2). The lower panel reports the resulting conformer distributions. The conformer string is given on the x-axis, where each digit stands for a torsion angle, the letter representing the occupied free-energy minima. Dots are used instead of letters when no change could be observed in comparison with the most populated conformer cluster. The gray boxes highlight a key conformational difference between REST-RECT and MD simulations, with a depicted free energy landscape of the corresponding angle (inserts). Only conformers with a probability higher than 2.5 % are plotted.

In the REST-RECT simulations, the whole N-glycan was defined as the solute, whose Hamiltonian was scaled in replica α by means of a scaling factor λ_α acting on the long range electrostatics, the Lennard-Jones interactions, as well as the dihedral angles. For non-neutral systems a neutralizing background was automatically added via GROMACS. We used 12 replicas and a geometric progression of λ_α values equal to 1, 1, 0.92, 0.84, 0.77, 0.71, 0.65, 0.60, 0.55, 0.50, 0.46, 0.42, spanning an effective temperature ladder from 310.15 K to 800.00 K. Note that both the ground replica ($\alpha = 0$) and the first replica ($\alpha = 1$) were simulated at the same ground temperature $T_0 = 310.15$ K, for convenience of the RECT implementation and analysis. Water and ions were always kept at the ground temperature. Replica exchanges were attempted every 400 steps, following a Metropolis-Hastings acceptance criterion. In the RECT part, all N_z torsion angles, as listed below, of the simulated glycan were defined as CVs and biased simultaneously by N_z one-dimensional potentials in each replica α. N_z amounted to 14, 17, 24, 17, 23 and 37 for M5, FM5, M9, A2G2, A2G2S2 and A4G4S4, respectively. The αth replica was biased with a bias factor γ_α following a geometric progression of values equal to 1, 1.2, 1.46, 1.82, 2.3, 2.94, 3.78, 4.89, 6.34, 8.23, 10.7, 14 over the replica ladder. Gaussian hills were deposited at time intervals of $\tau_G = 1$ ps, with a width of 0.35 rad and a height corresponding to $h_\alpha = (k_B \Delta T_\alpha / \tau) \times \tau_G$, where k_B is the Boltzmann constant, $\Delta T_\alpha = T_0(\gamma_\alpha - 1)$ the boosting temperature and $\tau = 4$ ps the characteristic time for the bias evolution in the RECT part. The geometric progressions of λ_α and γ_α ensured sufficient overlaps of the potential energy distributions at all temperatures, resulting in uniform round trip times for the different replicas (see supporting information of Grothaus et al. 2022[137] for round trip time plots of each N-glycan).

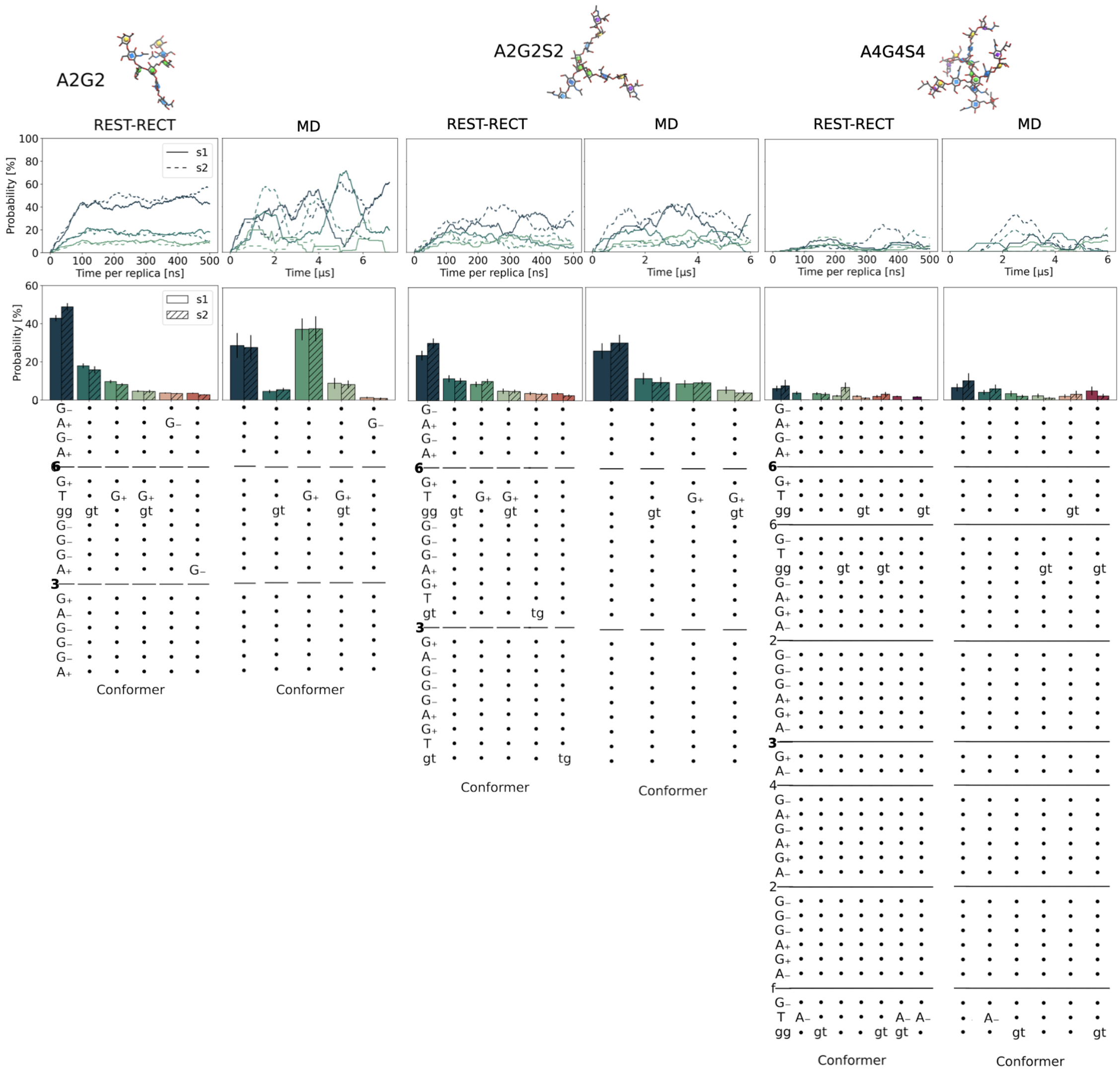


Figure 3.4: **Performance of REST-RECT for complex type N-glycan conformational exploration.** Comparison of REST-RECT with plain MD simulations as in Figure 3.3 for the systems A2G2, A2G2S2 and A4G4S4 using the CHARMM36 force field.

We note that the ground replica was fully unbiased ($\lambda_0 = 1$, $\gamma_0 = 1$), allowing for an unbiased statistical distribution of frames, which could be trivially used in all subsequent analyses. The first replica was biased, but kept at the ground temperature ($\lambda_1 = 1$, $\gamma_1 = 1.2$) to ensure sufficient overlaps between the first two replicas. The inclusion of higher-order replicas in the analyses requires the application of WHAM[182]. However, including replicas up to $\alpha = 4$ in the analyses did not result in a more effective sampling. Higher replicas should not enter the analyses because of the expected low ensemble overlap with the reference replica. Enhanced-sampling REST-RECT simulations were compared to standard MD at 310.15 K to assess the exploration of the conformational phase space spanned by the torsional angles of the N-glycan. A fair comparison of the two methods was ensured by using the same total simulation time, amounting to 6 μs for standard MD and to 500 ns for each replica in the REST-RECT simulations, which included 12 replicas.

The obtained probability distributions for the individual conformers and the moving average for the three most probable ones are shown in Figure 3.3 for the high-mannose type N-glycans and in Figure 3.4 for the complex N-glycans. Labeling of conformers, in line with the official IUPAC nomenclature for dihedral angles[201], was performed as described above with our own GlyCONFORMER python package. The first notable result is that stable and consistent conformer population distributions were obtained already after about 100 ns of REST-RECT simulation, irrespective of the starting conformation s1 or s2, as shown in the upper panels of Figure 3.3 and 3.4. In contrast, especially for high-mannose type N-glycans and A2G2, plain MD simulations displayed large fluctuations, poor convergence and significant dependency upon the starting conformation for individual conformers. This resulted in much larger error bars associated with the conformer distribution histograms (Figure 3.3 lower panel) in comparison with the enhanced-sampling simulations. For A2G2S2 and A4G4S4 (Figure 3.4) the conformer distributions and moving average plots are rather similar, however converged results can be obtained within much short times via REST-RECT, as replicas are parallelized. Furthermore, for M9 standard MD simulations even predicted a different most stable conformer than REST-RECT. The ψ angle in the main 1→6 linkage between two mannose residues (gray box in Figure 3.3, lower panel) remained stuck in a G_{-} free-energy minimum and did not reach the global-minimum T conformation predicted by REST-RECT. The analysis further revealed interesting common patterns of torsion-angle conformations in certain structural elements across the investigated models. For instance, the sequence $G_{-}A_{+}G_{-}A_{+}$ was predicted as the global minimum of the chitobiose core for all glycans, and the $G_{+}Tgg$ sequence characterizes the 1→6 linkages in most cases. Moreover, there was an evident preference for a *gg* conformation of the ω angle, which originates from the *gauche* effect.[83]

Assessing the exploration of the second structural feature, the puckering of saccharide units, was realized by the comparison of two-dimensional free energy profiles along the pucker coordinates θ and ϕ for standard MD and REST-RECT simulations. The two-dimensional pucker plots were calculated using the Mollweide projection, also termed homolographic or elliptical projection (Figure 3.5 **A**). This pseudocylindrical map projection is equal-area, meaning that areas, densities and, thus, free energy values are preserved. Only representative results for glycan A2G2S2 are depicted as it harbors the four most common monosaccharide types. The obtained free energy maps are comparable for all the other N-glycans (Figure 3.5 **B**). Under standard MD simulations, the Gal, Man and GlcNAc residues adopt solely the chair conformation $^{4}C_{1}$, whereby Neu5Ac, next to its minimum around $^{1}C_{4}$, also explores boat and skew-boat conformations. In comparison to REST-RECT, especially the GlcNAc monosaccharide lacks exploration of the chair conformation $^{1}C_{4}$ and several minima along the equator. However, Gal and Man do not explore any other conformations except for $^4C_1$ also under REST-RECT sampling. In terms of convergence, which can only be realistically analyzed for Neu5Ac and GlcNAc (Figure 3.5 **C**), a minimum energy chair conformation is reached after around 100ns and progresses with a stable probability of roughly 97%, where contributions from the other states are only minor. In the limits of the applied CHARMM36 force field, REST-RECT enhances sampling along the variables θ and ϕ for certain monosaccharide types through the scaling of the Hamiltonian, alias increasing the effective temperature.

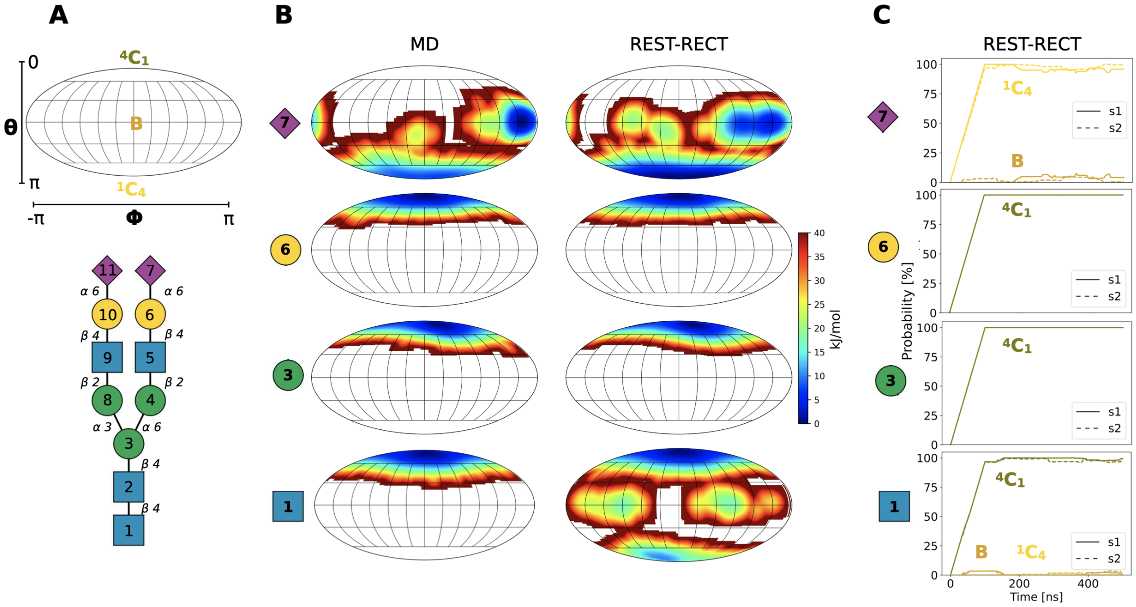


Figure 3.5: **Performance of REST-RECT for complex N-glycan pucker conformations.** **A** 2D schematic representation of the Cremer-Pople puckering coordinates θ and ϕ and structure of the complex glycan A2G2S2. The structure has been drawn with DrawnGlycan[50] where GlcNAc is blue, Man is green, Gal is yellow and Neu5Ac is purple. **B** Free energy profiles of various monosaccharides originating from the sampling of A2G2S2 via REST-RECT and classical MD starting from three-dimensional structure s1. **C** Moving average to assess the convergence of the puckering free energy profiles along θ for the monosaccharides in **B** for two independent start conformations s1 and s2. θ typically harbors three minima, which can be associated with two chair conformations ($^4C_1, ^1C_4$) and one boat or skew-boat conformation (B). θ values were classified as $^4C_1$ when being in the range $[0 - 1.0)$, B between $[1.0 - 2.25)$ and as $^1C_4$ between $[2.25 - \pi]$, with a window size of 10000 datapoints using 50000 in total.

3.1.3 Low dimensions

The set of strings associated with the most-populated conformer clusters of a given N-glycan is an already reduced representation of the highly-dimensional conformational phase space spanned by the Cartesian coordinates. However, such conformer strings are lengthy and become cumbersome when comparing different N-glycan systems. In addition they do not give a measure of the structural proximity among the different conformers within one glycan, as all torsion angles are considered as equally important. We therefore investigated the ability of the dimensionality reduction techniques introduced in chapter 2.3 (PCA, Diffusion map, Sketch map, kPcovR) to deliver two-dimensional representations of the conformer clusters in an efficient and physically meaningful manner, using all torsion angles as input features.

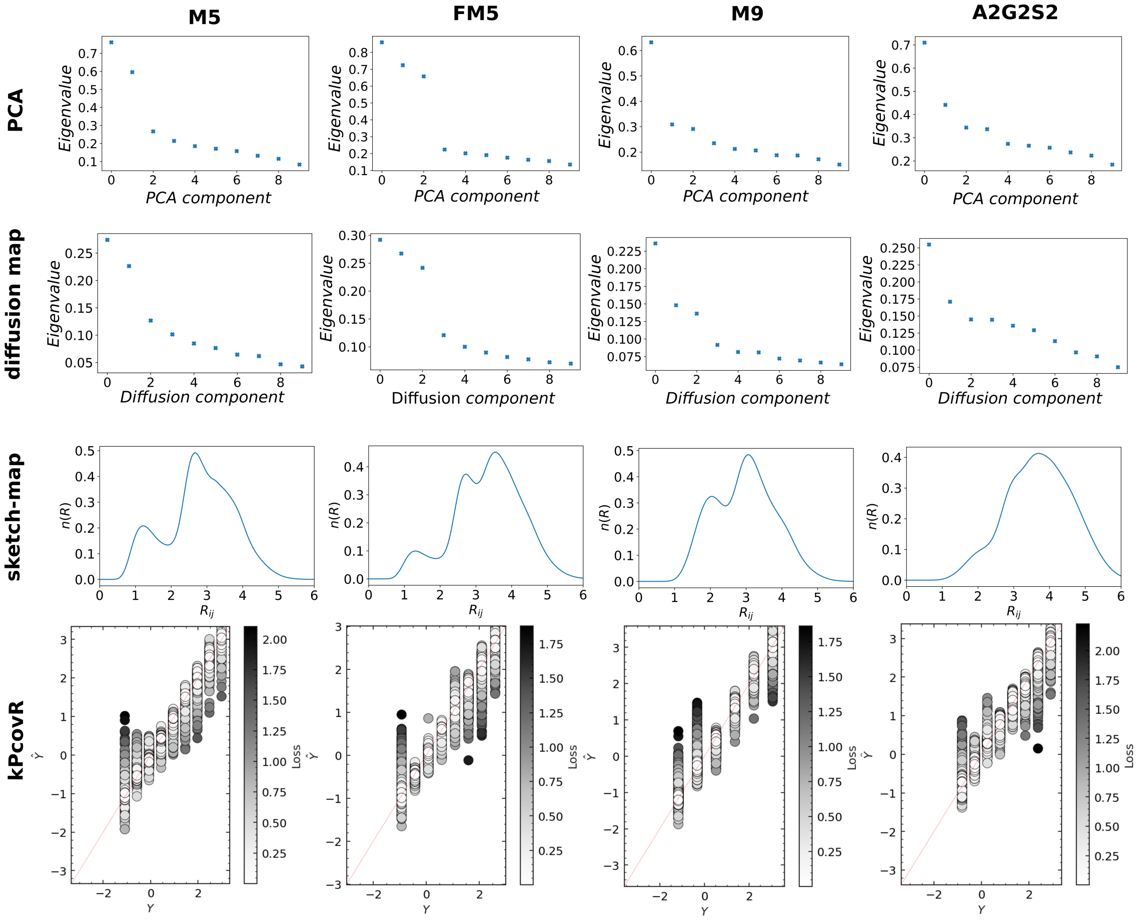


Figure 3.6: **Performance analysis of the four different dimensionality reduction techniques:** Principle component analysis (PCA),diffusion map, sketch-map and kernel principle covariates regression (kPcovR). Eigenvalues of the covariance matrix (y-axis) are plotted against the corresponding components (x-axis) for PCA and diffusion maps. For sketch-map, the histogram of pairwise distances plotted for randomly picked points at a given distance (R_{ij}) is shown. The parameter σ was chosen from this probability distribution, as it represents the switching distance for the sigmoid function, defining which distances are considered close or far in the sketch-map algorithm. At last, regression of the kPcovR algorithm is shown by plotting the property matrix $\mathbf{Y}$ against the reconstructed property variables $\hat{\mathbf{Y}}$. Points are colored according to their associated loss.

We always included 31250 data points from the ground replica of REST-RECT simulations ($n_{samples}$), with the different torsion angles of each N-glycan defined as features ($n_{features}$), resulting in a feature matrix $\mathbf{X}$ with shape $n_{samples} \times n_{features}$. In order to account for the periodicity of the torsion angles, their sin and cos values were used in the feature matrix $\mathbf{X}$ instead of their radian values, in all cases except for Sketch map. PCA calculations were performed with the scikit-learn package[215]. Diffusion maps were calculated with an inhouse python script, where the parameter σ, defining the size of the neighborhood including similar conformational structures, was set equal to 1.7. Sketch maps were calculated with the DimRed module of PLUMED (version 2.6). The matrix of dissimilarities between the frames in the feature space were calculated using the Euclidean distance measure. 500 landmark points were obtained from farthest point sampling and subjected to minimization of the stress function. The switching distance was chosen equal to 2.5 for all N-glycans, which lies roughly in the middle of the range of distances characterized by Gaussian fluctuations (Figure 3.6). We set $a_D = b_D = 4$ in the high-dimensional space, and $a_d = b_d = 2$ in the low-dimensional space, as defined by Ceriotti and co-workers (see chapter 2.3)[193]. The tolerance for the conjugate gradient minimization was set to 10^{-3}, using 20 grid points in each direction and 200 grid points for interpolation. 5 annealing steps were used and the remaining trajectory data were projected on the constructed sketch-map.

The recently developed kPcovR algorithm was employed as described in the tutorials at `https://github.com/lab-cosmo/kernel-tutorials`, using the scikit-cosmo implementation. Besides the feature matrix $\mathbf{X}$, the conformer strings assigned to each frame were employed as properties in the property matrix $\mathbf{Y}$. Prior to fitting, the input was centered and standardized by removing the mean and scaling the data to obtain a unit variance. We note that our data set was used as a whole and not split into separate training and testing data sets. A Gaussian kernel with $\gamma = 1$ was used and mixing parameters α were calculated for each simulated glycan on a subset of 1000 frames for two dimensions, leading to values of 0.1 for M5, 0.5 for FM5, 0.9 for M9 and 0.1 for A2G2S2. The regularization parameter λ was set to 10^{-4} for the linear regression part. The error of the linear regression part was assessed by comparing the true properties $\mathbf{Y}$ and the predicted properties $\hat{\mathbf{Y}}$.

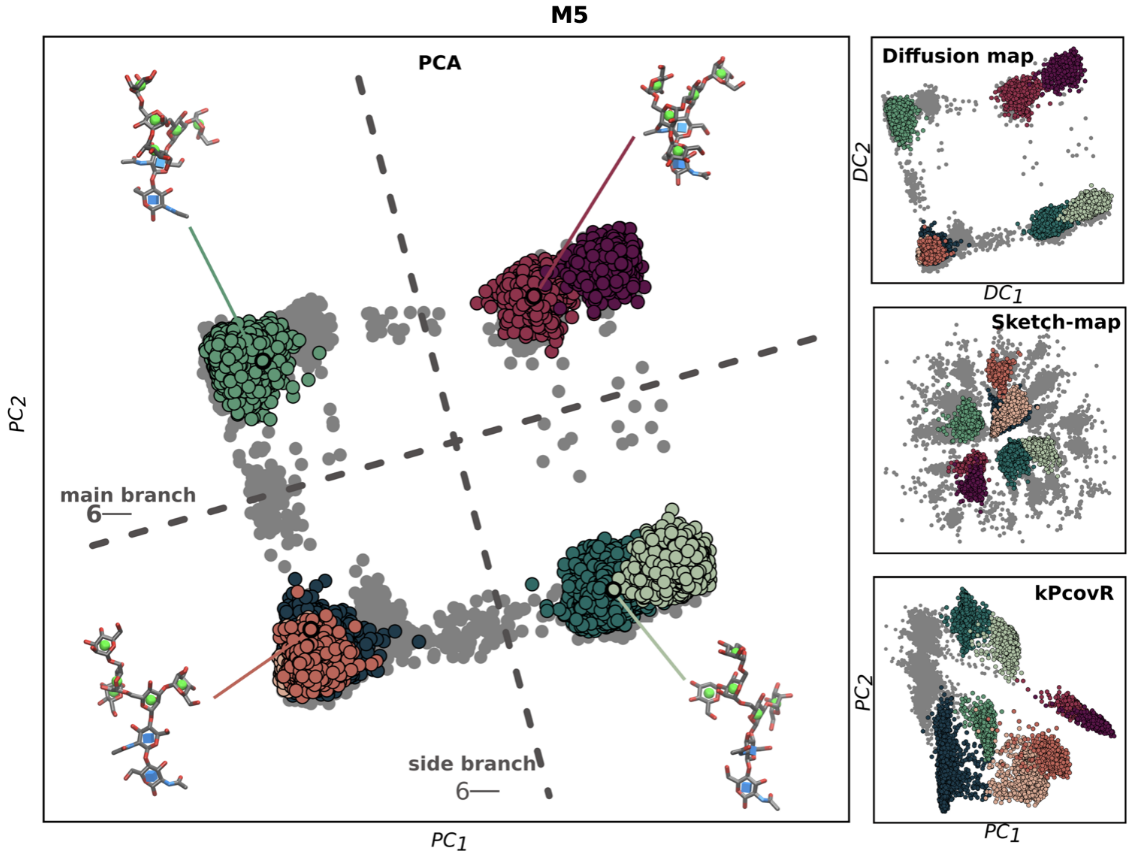


Figure 3.7: **Dimensionality reduction of the conformational phase space of M5.** Comparison of four different dimensionality reduction algorithms to cluster the distinct conformers of N-glycan M5. Principle component analysis (PCA), diffusion map and sketch-map employ all M5 torsion angles as features, whereas kernel principle covariates regression (kPcovR) additionally uses the conformer strings as properties. Each gray point corresponds to one frame and colored points to the respective conformers given in Figure 3.3. Sampling was performed from REST-RECT simulations (only s1).

Example projections of the four dimensionality reduction techniques are depicted for glycan M5 in Figure 3.7, where only selected representations are shown in Figure 3.8 for FM5, M9 and A2G2S2. PCA and diffusion map generated almost identical two-dimensional representations for M5 (Figure 3.7) and gave similar eigenvalue progressions along the PCA/Diffusion components for all analyzed N-glycans (Figure 3.6). There was an obvious gap observed between the first few eigenvalues (two for the case of M5 shown in Figure 3.6) and the remaining ones, indicating that corresponding structural features are more important than others in differentiating the glycan conformers. Having a closer look at the PCA of M5, $PC1$ differentiates conformers along their ω torsion angle in the side branch 6–, and $PC2$ along the main branch **6**–, corresponding to 25.0 % and 19.5 % of variance in the underlying data, respectively. Knowing that the highest variance is included in the rotations around ω torsion angles gave us an *a posteriori* justification for the two selected initialization states s1 and s2, differing in the states of ω in **6**–, situated in very different energetical conformers. The differentiation of conformers from each other by two ω angles gave rise to four main groups of conformer clusters (Figure 3.7). The overlap between clusters in each of those groups originated from the fact that these conformers only differ in ψ angles, which apparently can not be resolved in this two-dimensional representation, revealing the limitations of the PCA and Diffusion map algorithms. In fact, it is interesting to note that the number of highest, well-separated PCA or Diffusion-map eigenvalues is equal to the number of ω angles present in glycan structures M5 (two), FM5 (three) and A2G2S2 (one)(Figure 3.6). In this respect, M9 is an exception, presenting only one well-separated eigenvalue (corresponding to the 6– branch), but harboring two ω angles. Correspondingly, in the two-dimensional PCA map there is a clear separation of cluster conformers along the $PC1$ component, but some overlap along the $PC2$ component (Figure 3.8). The Sketch-map algorithm clustered conformers in a similar way to PCA and diffusion map, differing only in the overall spatial arrangement, but with marginally better separation of the clusters for M5, FM5 and M9 (Figures 3.7 and 3.8). In contrast, for glycan A2G2S2 no separation of conformers could be achieved, basically having all conformers collapsed onto one central point (Figure 3.8). This is probably due to its structural setup and the corresponding conformer distribution, where especially the conformers with a lower probability (colors light orange, orange and red in Figure 3.8 and 3.4) differ only in the ω angle of the terminal Sia residues. This feature can apparently not be mapped properly by the sketch-map algorithm, whereas also the switching distance σ was most complicated to determine, as the histogram of pairwise distances harbors only one maximum (Figure 3.6). In general, sketch-map analysis does not allow for a ranking of most important components and thus for an unbiased identification of the most important structural features of the system. The kPcovR algorithm separated the most-occupied conformers in the most effective way, resulting in clustered clouds with only little overlap to neighboring ones (Figure 3.7 and 3.8). However, the algorithm did not allow for a meaningful interpretation of how conformers are separated or clustered together, since no characteristic feature could be assigned to the kPcovR principal components. It rather seems that the clusters were ranked according to their population probabilities along $PC1$, as suggested by the progression of colors from left to right in the two-dimensional map. The separation along $PC2$ seems to be consistent with the $PC1$ component of the classical PCA projection. The calculated losses for the regressions in kPcovR can only be interpreted as relative values, as $\mathbf{Y}$ consists of discrete conformers, but the linearity between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ along the target (red line) is very clear (Figure 3.6). From this analysis, we conclude that, for the investigated systems, PCA, diffusion map and sketch-map can be used almost interchangeably with respect to their physical meaning, while kPcovR may be useful whenever a two-dimensional representation with well-separated conformer clusters depending on their population distribution is sought for. In the further applications only PCA is employed, as it is the most straightforward and computationally effective algorithm and gives consistent results for all tested N-glycan systems.

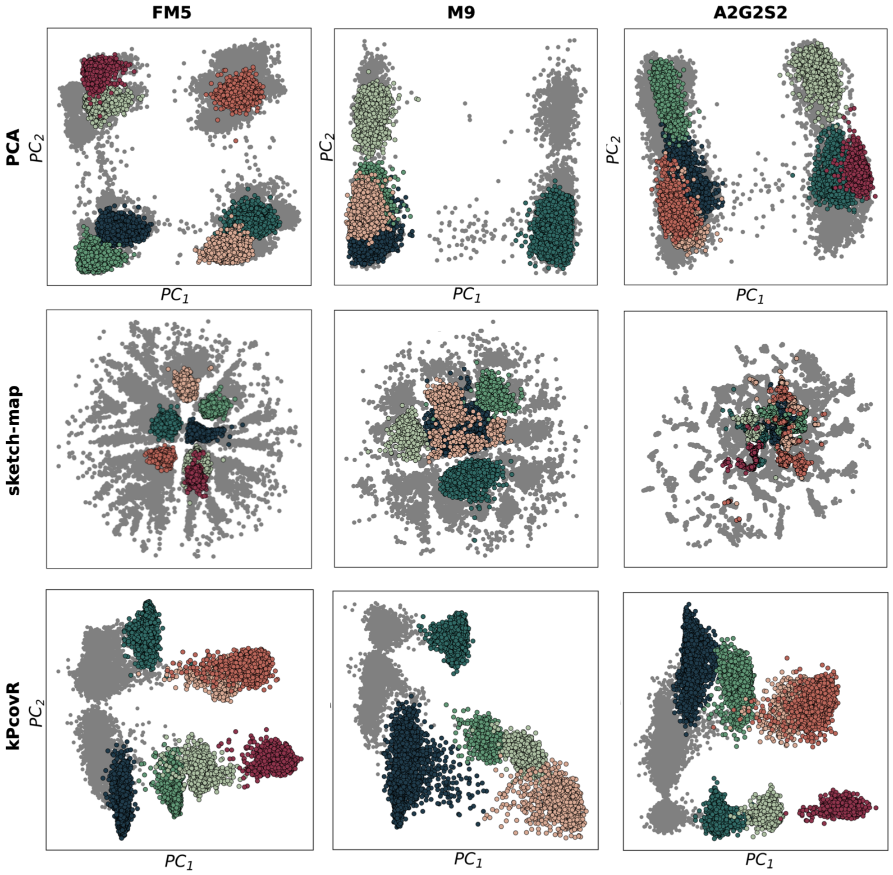


Figure 3.8: **Comparison of dimensionality reduction techniques for various N-glycans.** Each gray point corresponds to one frame and colored points to the respective conformers given in Figure 3.3 and 3.4. Sampling was performed from REST-RECT simulations (only s1). Dimensionality reduction projections for N-glycans FM5, M9 and A2G2S2 using Principle component analysis (PCA), sketch-map or Kernel principle co-variates regression (kPcovR).

We were able to show that REST-RECT simulations provide converged conformer distributions with complete sampling of all torsion and puckering angles within a few hundred nanoseconds of cumulative time, with better accuracy and using less computational time than long standard MD simulations. Sufficiently short round trip times reveal the strength of the RECT method, capable of biasing more than 20 CVs simultaneously while still ensuring adequate diffusion in the replica space. This behavior originates from an adjusted ergodicity by scaled bias factors over the replica ladder, ensuring a proper compensation of free energy barriers by a self-consistent addition of one-dimensional bias potentials.[177] Alternative methods such as temperature REMD[89] are computationally too demanding in solvated systems.[216] Bias-exchange metadynamics[217] would have required more replicas for the same number of biased CVs, and would have only enabled biasing them one at a time.[177] It may be argued that some of the chosen CVs are in fact redundant; in particular, axial ϕ torsion angles in α-linkages occupy only the gauche conformation (G_+) due to the so-called exoanomeric effect. This is due to the favorable overlap of one oxygen lone-electron pair with the antibonding orbital of the adjacent C–O bond[83,218,219], an effect that needs to be appropriately mapped by torsion, Lennard-Jones and Coulomb terms in force-field potentials. However, there is no computational advantage in excluding those CVs from the biased scheme, and is indeed reassuring to see that the results do confirm such background-knowledge details.

Comparison of the newly introduced conformer strings easily reveals differences and similarities, which are immediately traceable to specific linkages and torsion angles. Previous classifications of identified conformers were performed with less clear nomenclature rules. For instance, the groups 'backfold', 'half backfold', 'tight backfold', 'extended-a' and 'extended-b' were defined according to their ψ and ω torsion values of the first 1→6 linkage. These groups can still be differentiated using a low-dimensional representation constructed by PCA (see Figure S11 for A2G2 in supporting information of Grothaus et al. 2022[137]). In line with a previous study, we found that 'half backfold' conformers is in fact part of the 'backfold' cluster.[87] However, this classification system is again limited to the description of only one single branch.

The application of dimensionality reduction techniques to N-glycans has so far been limited to PCA using atom coordinates as input features.[188,220] Nevertheless, clustering of glycan conformers has been already performed in the past, using e.g. end-to-end distances between glycan branches to describe their flexibility and identify the main conformers.[93] Additionally, the usage of spherical coordinates was introduced to describe the dynamical behaviour of the **6**– branch, albeit the procedure was applied to the description of only one single branch.[96] Our successful application of various dimensionality-reduction techniques to represent the high-dimensional phase space of N-glycans highlights their enormous potential in delivering a consistent analysis of the most-populated conformer clusters, while simultaneously providing meaningful information about the most important structural features behind the used descriptive variables. This becomes very important for glycans composed of diverse monosaccharide units arranged in complex branched chains, presenting additional chemical modifications (fucolsylation, sialylation) and including more than two ω torsional angles, making structural relationships not as intuitive as for small glycans like M5.

3.2 Applications

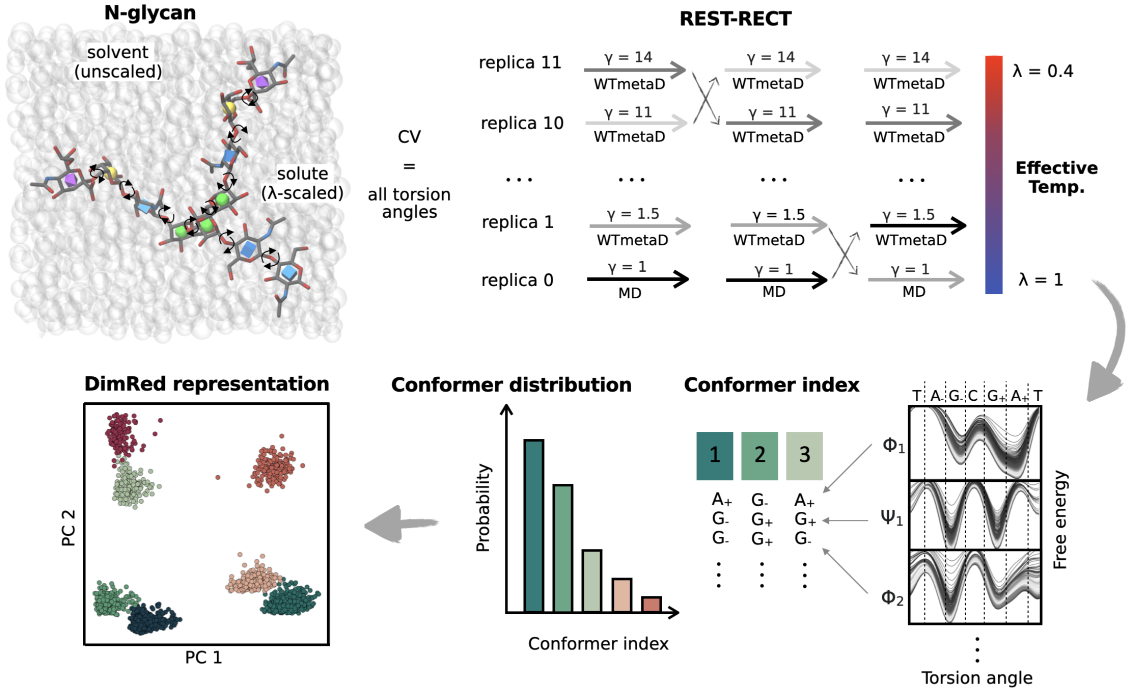


Figure 3.9: **Schematic overview of the joint workflow for the study of N-glycans' conformations.** Free *N*-glycans in solution are simulated employing the enhanced-sampling method REST-RECT to accelerate transitions over barriers for all torsion angles and pucker coordinates. Conformer strings are then constructed based on the free energy landscape of each torsion angle. Individual conformers are grouped together according to these strings, and conformer distributions constructed from the simulated trajectories. Low-dimensional representations of the conformer clusters are finally generated using dimensionality-reduction methods.

After having demonstrated that the REST-RECT methodology accurately samples the phase space of N-glycans and that PCA delivers meaningful low-dimensional representations, a new workflow for computational glycan studies was proposed (Figure 3.9). The practicality of the such is illustrated by its application to two on going research questions, namely:

- How is the global conformation of an N-glycan altered by its size and shape?
- How are biomolecular force fields performing in the reproduction of realistic N-glycan three-dimensional structures?

Finally the usefulness of the method is illustrated on a real-life example, improving the predictive power of the newly developed GlycoSHIELD software.[221]

3.2.1 The influence of glycan size on the conformational phase space

The enzyme machinery in the ER and Golgi apparatus is constantly trimming and elongating glycan structures, leading to a diverse bunch of differently sized and compositioned N-glycans (Figure 1.5). As there is not one final N-glycan structure but also intermediate constructs that can escape further processing steps, the resulting diversity and its impact is puzzling. In order to address this phenomenon from a structural point of view, the global conformation of various N-glycan structures was compared to verify the impact of additional monosaccharide residues. In detail, the conformational landscape of the already mentioned high-mannose type and complex N-glycans was analyzed (Figure 3.2). To this purpose, PCA was employed to represent the free-energy maps associated with the conformational ensembles of five glycan models in two dimensions. Whenever two different feature matrices $\mathbf{X}$ stemming from separate simulations were compared to each other, the corresponding data sets were concatenated prior to PCA calculation. Free energy differences (ΔG) along the principal components 1 and 2 defining the low-dimensional latent-space matrix $\mathbf{T}$ were calculated by constructing two-dimensional histograms with 35 bins and converting the histogram probabilities P according to $G = -k_B T \ln(P)$.

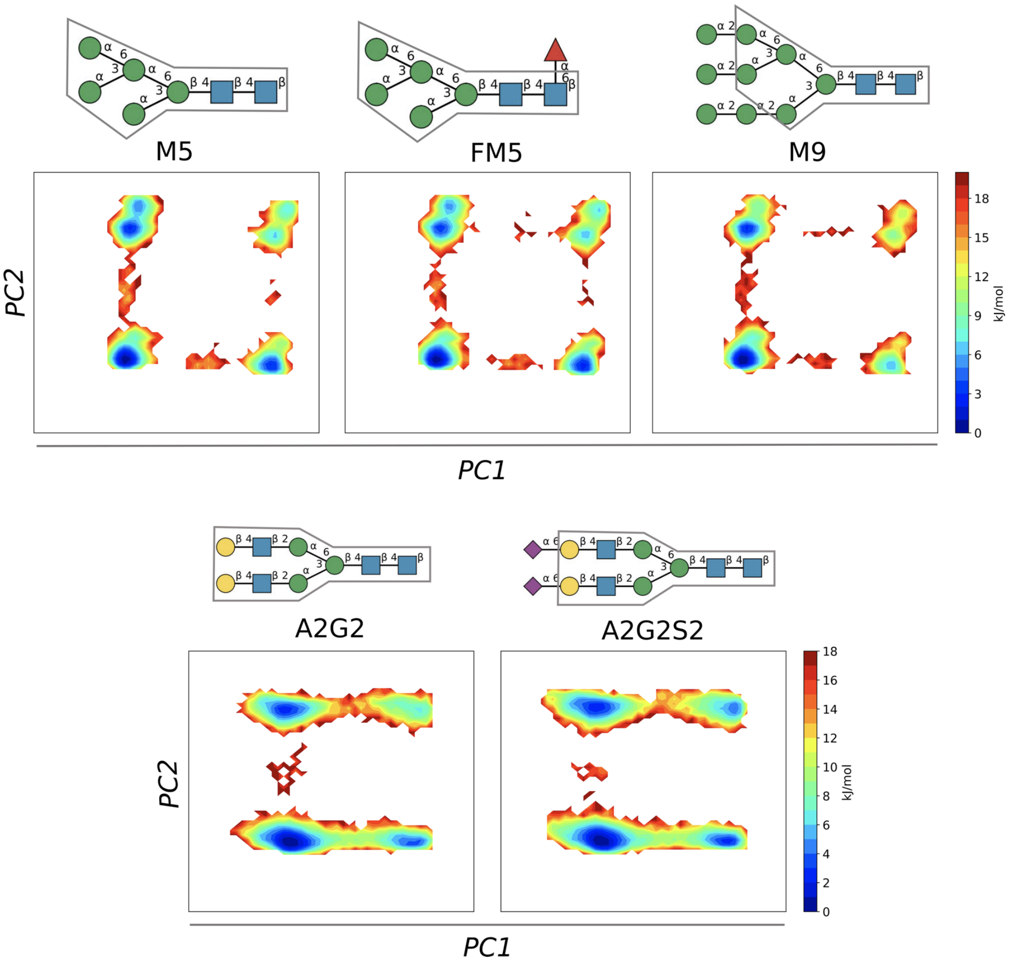


Figure 3.10: **PCA free-energy maps comparing different structural features of N-glycans.** The upper panels compares the three high-mannose type N-glycans M5, FM5 and M9, whereas the lower panels compares A2G2 against its sialylated variant A2G2S2. Only the torsion angles common to all structures (boxes around the schematic glycan models) where used as features in the analysis. A common PCA was performed by concatenating the datasets of M5, FM5, M9 and those of A2G2 and A2G2S2, respectively. Sampling of the phase space was performed with REST-RECT simulations, starting from the s1 conformation. Glycan structures have been drawn with DrawnGlycan[50] where GlcNAc is blue, Man is green, Gal is yellow, Neu5Ac is purple and Fuc is red.

We especially focused on the influence that various chemical modifications (fucosylation, sialylation, variation of branch length) might have on the resulting ensembles. In doing so, we included in the analysis only the structural features common to all compared structures, highlighting the effect of the mentioned modifications on the free-energy maps (Figure 3.10). In general, the analysis showed that the core structure of the high-mannose type N-glycans was only marginally affected by addition of further residues. Fucosylation of M5 (FM5) had no effect at all on the conformational free-energy landscape, whereas elongation of the branches by further mannose units (M9) led to a slight stabilization of the main conformer and destabilization of the secondary minima at larger $PC1$ values (Figure 3.10, upper row). Instead, sialylation of A2G2 (A2G2S2) with additional Neu5Ac units on both branches deepened slightly all three secondary energy minima at the expense of the most-populated region of the conformational phase space in the bottom-right corner of the map (Figure 3.10, lower row). From the here examined N-glycans, it can certainly be stated that apparently no new phase space regions are explored, but rather already visited regions traveled with an altered frequency. It can be concluded that there exists no general rule associated with the elongation of glycan branches. This can lead to the reduction of flexibility, deepening certain energy minima as in the case of high-mannose glycans, or slight increase in conformational diversity, as for the complex N-glycan A2G2S2. The proposed workflow, however, turns out to be a useful tool in dealing with multiple glycan structures simultaneously, being able to evaluate structural differences on a global scale.

The comparison of conformer distributions upon chemical modifications like the addition of core fucosylation or sialylation have also been shown in previous studies to leave the equilibrium distribution almost unaffected[222], although the here used PCA representation visualizes the results much more comprehensively. When analyzing glycan structures on a detailed torsion-angle-based level, the focus lies on the more flexible ω angles, preferably found in their gauche conformation as outlined for the different N-glycans.[74] Both experimental and computational studies of M9 suggested that it is mainly confined in a gauche conformer, meaning that the ω torsion angle in the **6**− branch should be in a *gg* conformation, which is in agreement with our findings.[220,223,224]. The observed stabilization of the global minimum of M9 after elongation of M5 branches by 1→2-linked mannose units (Figure 3.10) is probably due to an increased number of inter-branch hydrogen bonds.[225]

3.2.2 Force field accuracy

The performance of MD simulations always depends to a large extent on the used force fields and their parameters. Force fields are continuously developed and refined for selected systems and cases, such as the stability of protein-carbohydrate complexes, the conformational behaviour of linear polysaccharides, or the ring distortions of monosaccharide units.[84,85,99] Generally speaking, all already mentioned biomolecular force field families have been shown to have good performance in reproducing the behavior and predicting experimental data of polysaccharide systems, with few exceptions.[83,85] However, depending on the saccharide size, the property under investigation and the required level of detail, differences among the force field families do emerge, which can be traced back to how well the steric, electrostatic, and torsional energy terms represent the physical reality and mimic the actual glycan behaviour.[83] The frequent revisions of the force-field parametrization of torsional terms, up to the present day, indeed shows that a correct description of rotational barriers in glycan systems is not at all straightforward.[98] Carbohydrate force fields have been especially compared for the simulation of protein-glycan complexes, where N-glycans were viewed rather on a macroscopic scale without a detailed analysis of their conformer distributions.[84,100,226] We however consider N-glycans on a microscopic scale, estimating the force field performance based on precise distributions of the individual torsion angles and puckering coordinates. This detailed analysis becomes especially important where glycans do not only serve as surface modifications that randomly interact with surrounding amino acids, but undergo specific interactions like a substrate in a protein pocket. The already mentioned example of lectins and glycan processing enzymes indicates the importance of conformer selection by protein binding sites, as also small molecular compounds adopt specific conformations when bound to a protein.[14,66]

Torsion angles

The developed workflow for the simulation and analysis of N-glycans also allows for the assessment of the structural prediction capability of different force fields. Here, we compared the two most widely used force fields for protein and carbohydrate systems, namely CHARMM36 and GLYCAM06j. Enhanced sampling simulations of N-glycan M5, M9, A2G2 and A2G2S2 using REST-RECT were performed with the GLYCAM06j[155] force field in combination with the standard version TIP3P water model (sTIP3P)[227] in addition to the above mentioned simulations using the CHARMM36 force field. We note that the GLYCAM06j force field parameters in a GROMACS format were obtained from the CHARMM-GUI Glycan Modeller, while constructing the glycan structure and simulation box. This Amber force field parameter input generation for GROMACS is available since version 3.6. Comparative plots of REST-RECT and MD simulations as in Figure 3.3, however employing the GLYCAM06j force field, are available as supporting information in Grothaus et al. 2022.[137] Detailed results for A2G2S2 are shown as an example in Figure 3.11, whereas the other glycans are depicted in a reduced representation in Figure 3.12. The comparison revealed very substantial differences for A2G2S2 in terms of both conformer distributions and free-energy landscapes, and even the global-minimum structures were different. CHARMM36 predicted that the majority of conformers (and thus the global free-energy minimum) cluster on the right-bottom region of the PCA map, in contrast to GLYCAM06j, which predicted a global minimum in the left-bottom region. The position of the secondary minima was also different in the two cases. Comparison of the predicted conformer strings indicated that the major differences arise from the ψ angle of the main branch **6**– as well as the two ω angles of the terminal 1→6 linkages between Gal and Neu5Ac (Figure 3.11, lower panels). In the global-minimum three-dimensional structure predicted by CHARMM36, ψ was in a T conformation and ω in a gt conformation. These conformations changed to G_+ and tg in the global-minimum three-dimensional structure predicted by GLYCAM06j, respectively.

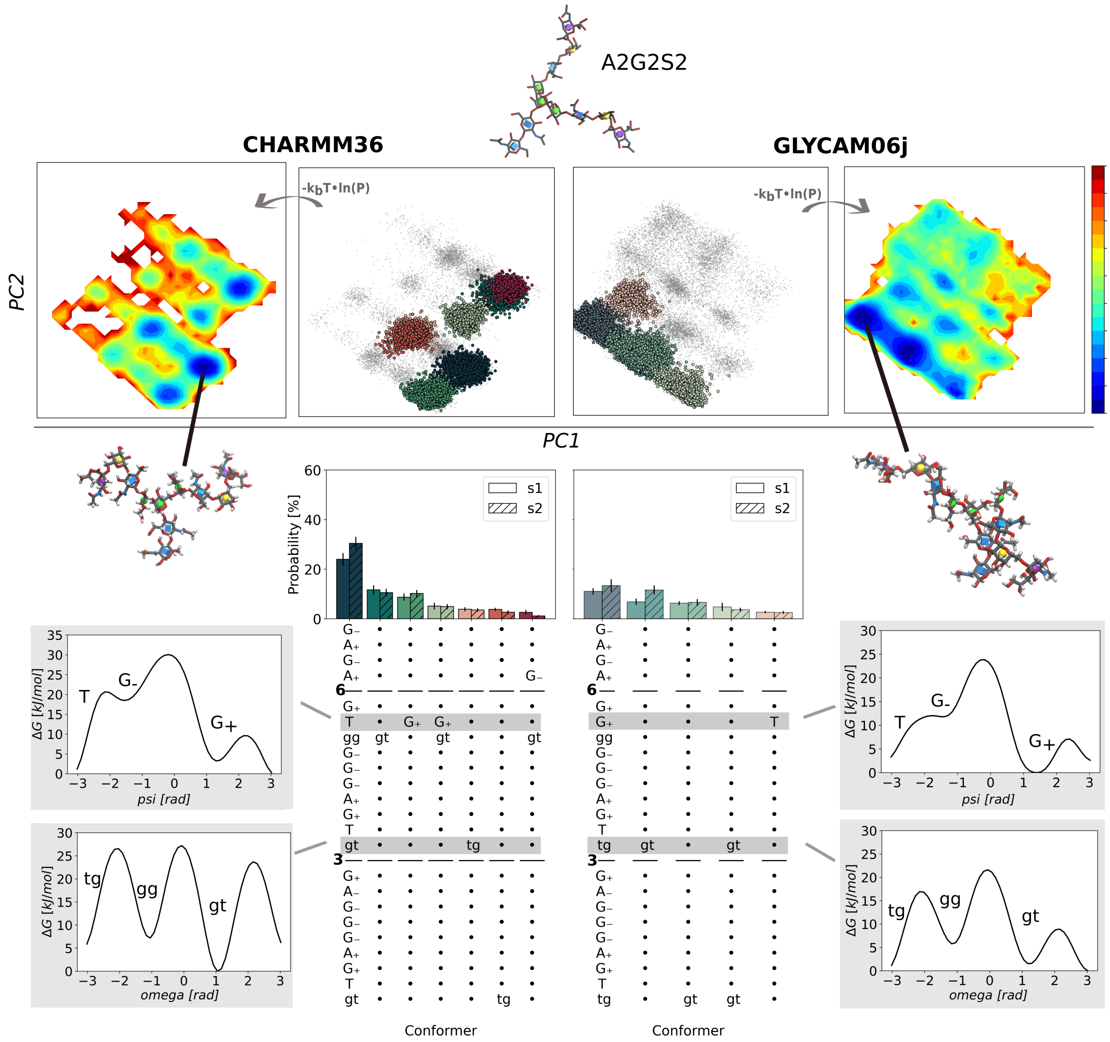


Figure 3.11: **Assessing force field performance for glycan A2G2S2.** Comparison of the conformational phase space (torsion angles) of A2G2S2, predicted by REST-RECT simulations with either the CHARMM36 or the GLYCAM06j force field. The upper panels show the PCA maps of conformer clusters and the corresponding free-energy landscapes. The clusters are colored in accordance to the conformer distributions shown in the lower panels. Free energy profiles along selected torsion angles (indicated by the gray rectangles) are represented besides the conformer strings and labeled with the conformations of the free energy minima. The PCA was constructed by concatenating the datasets of the two force field simulations.

By looking at the one-dimensional profiles along selected torsion angles, it becomes evident that the discrepancies arise from only subtle differences in the force field parametrizations. For instance, the free-energy differences between the T and G_+ conformations of the ψ angle, or between the gt and tg conformations of the ω angle, were less than 5 kJ/mol. However, such small differences have a profound effect on the resulting multi-dimensional free-energy landscape, and lead to rather distant global minima, as observed above. Similar considerations hold for M5, M9 and A2G2 (Figure 3.12) although the conformer distributions were less dramatically different than in the case of A2G2S2, rather having altered relative depths of the same minima. In the next section we will show that these force field differences led to markedly different predictions of NMR spectroscopic fingerprints for the various glycan populations.

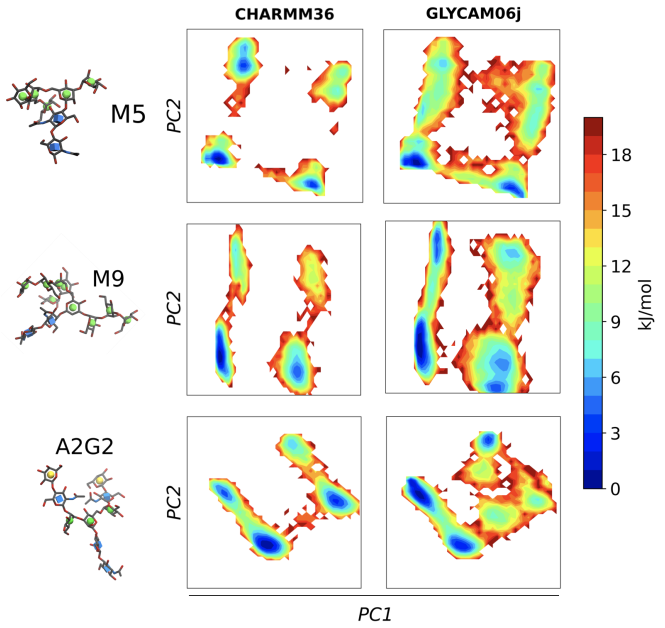


Figure 3.12: **Comparison of glycans' phase spaces explored by the CHARMM36 or GLYCAM06j force field with respect to torsion angles.** Free energy profiles along principle components 1 and 2, for N-glycans M5, M9 and A2G2. The local and global regions of the conformational phase space explored by REST-RECT simulations was assessed when employing either CHARMM36 or GLYCAM06j as force field. PCAs were constructed by concatenating the datasets of the two force field simulations for each glycan type, respectively.

J-coupling calculations

Validation of glycan three-dimensional structures obtained from REST-RECT simulations was carried out by comparing theoretically calculated with experimentally measured scalar ${}^3J_{H,H}$ NMR coupling constants. A comparison of experimental NMR data with the corresponding observables predicted theoretically by the two force fields was performed to ascertain which one can be considered more accurate in terms of torsion angle description. The comparison is meaningful only for ω torsion angles in α 1 $\rightarrow$ 6 linkages, since ϕ and ψ lack the necessary proton pair, whereas the $J_{H5,H6}$ and $J_{H5,H6'}$ constants can be both computed and measured (see Figure 3.13 below for the atom nomenclature).[228] As shown above, the largest variability among the different conformers of N-glycans originated from the ω torsion angles around the 1→6 O-glycosidic linkages, suggesting that the use of ${}^3J_{H,H}$ coupling constants is meaningful and enables a clear validation of the predicted three-dimensional glycan structures.[228] The three protons H5, H6, and H6' harbored by these linkages (see Figure 3.13) give rise to well-defined NMR J-coupling constants, whose values depend on the relative distances between the H nuclear spins, and thus on the conformation of the ω angles. We note that caution must be taken when comparing the results of different force fields because of the inconsistencies in atom labeling conventions. In particular, the H6 (H6S) and H6' (H6R) hydrogens are named 'H61' and 'H62' in CHARMM36, respectively, while the opposite names ('H62' and 'H61') are used in GLYCAM06j.

Theoretical calculations were performed by ensemble averages of the coupling constants computed for all conformers sampled by the REST-RECT simulations, using three different parametrizations of the empirical Karplus equation, namely:

1) The equation of Altona and Haasnoot[229]:

$$ {}^3J_{H,H} = P_1 \cos^2 \omega + P_2 \cos \omega + P_3 + \sum_{i=1}^{4} \triangle_{\chi_i} \left\{ P_4 + P_5 \cos^2(\zeta_i \omega + P_6 \left|\triangle_{\chi_i}\right|) \right\} , $$

where the sum runs over the different substituents (in our case, H, C and two O), the P parameters are taken from the original data set, the electronegativity values Δ_{χ_i} are equal to 0 for H, 0.4 for C and 1.3 for O and the substituent orientations ζ_i are either -1 or 1. The equation is applied to the torsion angles ω = H5–C5–C6–H6 or ω = H5–C5–C6–H6′. For example, the former has electronegativity values of 0.4 for i=1, 1.3 for i=2 and 3, 0 for i=4, with $\zeta_{1,2} = 1$ and $\zeta_{3,4} = -1$.

2) The equations of Stenutz[230]:

$$^3J_{H5,H6'} = 5.08 + 0.47\cos\omega + 0.90\sin\omega - 0.12\cos 2\omega + 4.86\sin 2\omega$$

$$^3J_{H5,H6} = 4.92 - 1.29\cos\omega + 0.05\sin\omega + 4.58\cos 2\omega + 0.07\sin 2\omega$$

with $\omega = O5 - C5 - C6 - O6$.

3) The equations of Tafazzoli[231]:

$$^3J_{H5,H6'} = 5.06 + 0.45\cos\omega - 0.90\cos 2\omega + 0.80\sin\omega + 4.65\sin 2\omega$$

$$^3J_{H5,H6} = 4.86 - 1.22\cos\omega + 4.32\cos 2\omega + 0.04\sin\omega + 0.07\sin 2\omega$$

with $\omega = O5 - C5 - C6 - O6$.

The general equation of Altona and Haasnoot can be applied to different kinds of linkages due to the flexible choice of substituents and was already used before in the evaluation of glycan MD simulations.[96] We decided for the additional use of the two further equations from Stenutz and Tafazzoli, which are derived specifically from J-coupling constants computed with density functional theory for a model aldopyranosyl ring and D-glucose/D-galactose. As all variants of Karplus equations are purely relying on empirical parameters, derivations of the computed J-coupling constants from the experimentally observables are expected. The comparison of the different forms of Karplus equations should reveal the extent of such deviations and facilitate the classification of the obtained values. We stress that the assessment of computed Karplus values is only reliable when the ergodicity of the simulations is fulfilled, which requires complete phase-space sampling by means of converged simulations, as in the case of REST-RECT. Convergence is required because NMR measurements do not give results for a single glycan conformer, but output time-averaged and ensemble-averaged conformational data. The computed coupling constants were therefore averaged over all 62500 frames in each REST-RECT simulation of the different N-glycans (considering only the simulation with starting conformation s1). Block averaging was used to compute error bars, as described above for the probability distributions.

For the main branch of M5, GLYCAM06j led to better agreement between the experimental and theoretical $J_{H5,H6}$ and $J_{H5,H6'}$ frequencies, whereas CHARMM36 performed better for the side branch (Figure 3.13). Regarding M9, CHARMM36 performed better than GLYCAM06j for both ω angles, although both force fields overestimated the $J_{H5,H6'}$ frequency by over 2 Hz. We would like to note that the experimental J-couplings of M9 were only reported as approximations in the original paper[228], but used here due to the lack of other data sources.

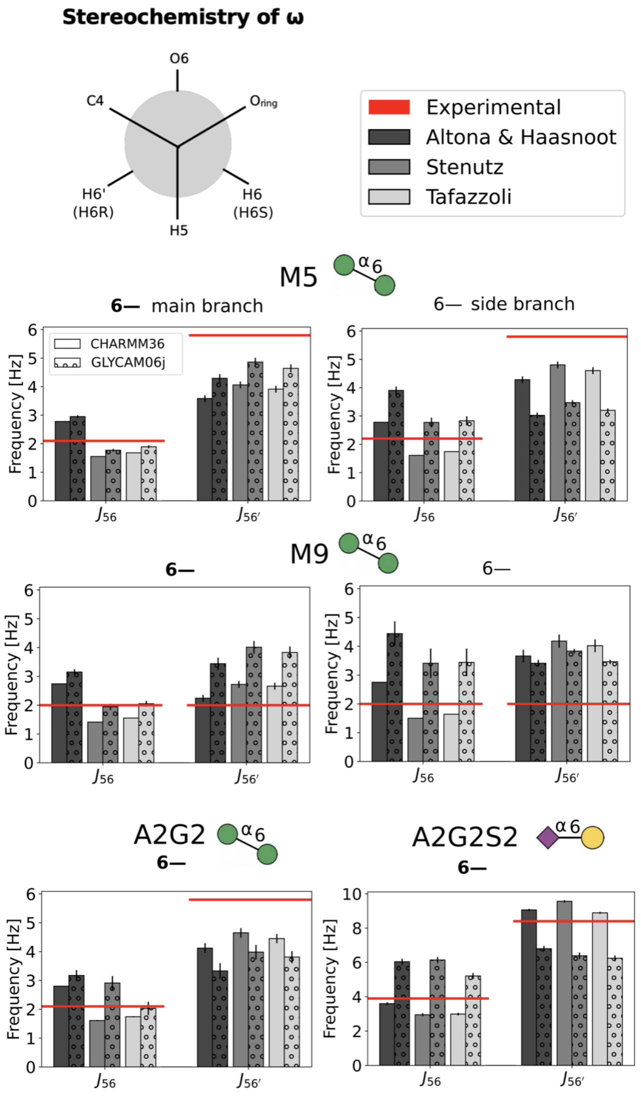


Figure 3.13: **Validation of the ω angle populations by comparison of computed and experimental NMR J-coupling constants for M5, M9, A2G2[228] and A2G2S2[97,232].** The upper-left panel shows the stereochemistry of an ω angle in a *gg* conformation along its C5 and C6 atoms, with labeled protons. The legend on the right reports the color code of the plots below, referring to the three different parametrizations of the Karplus equation used to compute the J-coupling constants. A table including all values explicitly can be found in the supporting information of Grothaus et al. 2022.[137]

For the single ω torsion angle in A2G2, CHARMM36 predicted slightly better frequencies than GLYCAM06j. For the sialylated variant A2G2S2, no experimental parameter of the ω torsion angle between two Man residues was available, therefore the comparison was made for the J-coupling constants of the 1→ 6 linkage between Gal and Neu5Ac. In this case, the experimental values were collected for a different glycan, namely trisaccharide sialyl-α-(2-6)-lactose[97,232], but could be used here as an approximation, because this glycan carries the same terminal branches as A2G2S2. The predicted CHARMM36 values of both $J_{H5,H6}$ and $J_{H5,H6'}$ were in very good agreement with the experimental ones.

Puckering

So far we only focused on the torsion angles of the N-glycan molecules, which were considered as explicit RECT collective variables in our simulations. However, the combination with the REST2 method allowed also good sampling of other structural degrees of freedom, in particular of the puckering conformations of the individual monosaccharide units. Whether the puckering free-energy landscapes represented by the Cremer-Pople parameters were dependent on the force field was investigated by constructing two-dimensional polar free-energy maps in a way that conserves the area defined by intervals of the θ and ϕ pucker coordinates. As usual, the free energy was computed from the histograms of conformer population probability in the ground replica of the REST-RECT simulations (only of starting conformation s1). In Figure 3.14 we show puckering maps for all saccharide units present in the complex N-glycan A2G2S2, comparing simulations using the CHARMM36 and GLYCAM06j force fields.

Analysis of the pucker landscapes along the branches of A2G2S2, which is composed of very diverse monosaccharide units, revealed a strong propensity for the chair conformation $^4C_1$, except for the terminal Neu5Ac units that were in a $^1C_4$ conformation for both force fields. However, the relative boat propensities were quite different for the different units, being very low or absent for Man and Gal units, more evident for GlcNAc units and strongest for Neu5Ac units. We note that GLYCAM06j, in comparison with CHARMM36, generally predicted a broader exploration of the pucker phase space, resulting in an increased appearance of local minima and smaller energy differences between different regions of the maps. Only the terminal Neu5Ac units presented very similar maps for both force fields, with the same distribution of minima and a similar degree of phase-space exploration.

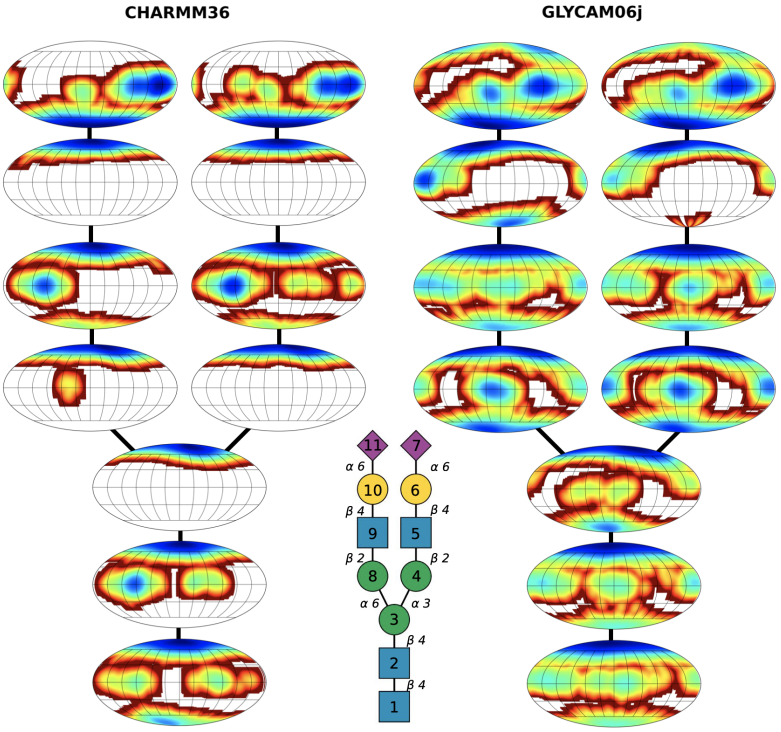


Figure 3.14: **Force field performance regarding saccharide ring distortion.** Free energy surfaces along the Cremer-Pople puckering coordinates θ and ϕ for all saccharide units in A2G2S2, comparing the CHARMM36 and GLYCAM06j force fields. Collective variables were computed from REST-RECT simulations by histogram construction and conversion to free energies. The 2D puckering plots of each monosaccharide are arranged in accordance to their position in the N-glycan structure as depicted in the schematic model. The puckering free energy profile is explained more in detail in Figure 3.5. The schematic glycan has been drawn with DrawnGlycan[50] where GlcNAc is blue, Man is green, Gal is yellow and Neu5Ac is purple.

3.2.3 Concluding remarks

We have performed an in-depth analysis of the N-glycan conformer distributions predicted by the CHARMM36 and GLYCAM06j force fields, focusing on converged free energy profiles of torsion angles that shape the three-dimensional glycan structure and its flexibility.

The observed different phase-space distributions for A2G2S2 and A2G2, as well as the different conformer distributions for M5 and M9, originate from different free energy profiles around the ψ and ω torsion angles in 1→ 6 linkages. Especially the ψ angle of branch **6**– in A2G2 and A2G2S2 is a critical feature, which differentiates between two main conformers, previously named 'backfold' and 'extended'.[96] CHARMM36 consistently produced conformer distributions with only a few high populated states, whereas GLYCAM06j produced broader distributions and flatter associated free-energy landscapes. Overall, neither force field reproduced all sparsely available experimental J-coupling constants with great accuracy (i.e., within the intrinsic error bars of the theoretical method), but CHARMM36 seemed to deliver better structural predictions than GLYCAM06j, especially in the cases of M9 and A2G2S2. To further improve force field performance, the contribution of torsional energy and electrostatic interactions needs to be balanced with great care.[98] The two force fields under investigation, CHARMM36 and GLYCAM06, especially differ in the latter term: CHARMM36 adjusts partial atomic charges to fit solute-water interactions of carbohydrate fragments computed with quantum mechanical methods, whereas the partial charges of GLYCAM06s are derived from the restrained electrostatic potential (RESP) method.[98] This leads to two very different sets of charge values.

The three tested parametrizations of the Karplus equation yielded consistent results, although they are all based on different empirical parameters or functional forms, so that some discrepancies are both expected and unavoidable. In terms of experimental data, $^3J_{H,H}$ coupling constants recorded for a complete N-glycan structure are difficult to obtain or rather difficult to interpret, as the resulting spectra suffer heavily from signal overlaps. It is rather common to determine the torsion angle preference of mono-, di- or trisaccharides, although the influence of the global glycan structure with its interactions is lost and presents only a limiting case.[83,233] It is therefore not surprising that the complete N-glycan structures investigated in this study are not available in the Glycan fragment database.[234] Therefore only very limited experimental data were accessible and rather few parameters for the ω torsion angles of M5, M9, A2G2[228] and partially of A2G2S2[97,232] could be used for comparison. Advanced J-coupling techniques involving isotope labeling of glycan structures, recording of multi-dimensional spectra as well as addressing carbon and nitrogen atoms are summarized in a very detailed recent review.[235] Other NMR observables like the nuclear Overhauser effect (NOE), which has already been used for glycan structure determination [96], are problematic, because three or more NOEs are required for an unambiguous assignment. Moreover, the r^{-6} dependence of the NOE does not allow for a straightforward averaging of conformations.[81]. Rather, a direct calculation of NOEs from converged MD trajectories has been proposed.[236] Three-dimensional structures derived from X-ray crystallography lack dynamical information, and are more helpful in cases where protein-carbohydrate complexes are analyzed and the glycan conformation is restricted by the surrounding amino acids.[81]

In contrast to the torsion angle analysis, no suitable experimental method or data set covering the whole glycan structure could be identified to verify which force field reproduces the natural puckering behavior more correctly. Several computational studies revealed the importance of ring flipping events in determining the polysaccharide conformer distribution[99,237], where the degree of puckering flexibility is influenced by the molecular context and size of the N-glycan.[77,238] It is therefore questionable if experimental datasets for mono-, di- or trisaccharides are a reliable source for judgment. One is left with comparison to e.g. electronic structure calculations, whereby even these are mostly performed for monosaccharides.[77] However, in line with our observations, previous studies of glycan monomers and trimers did notice differences in the puckering landscapes predicted by different force fields, and in particular pointed towards a better performance of the CHARMM36 force field.[99] It is however yet to be determined if especially the restricted puckering of Man and Gal residues in the CHARMM36 force field is due to their embedding in a larger glycan structure or require a force field reparameterization, as QM calculations have predicted the occurrance of boat and various other conformations at least for monosaccharides.[77] Accurate prediction of ring-inversion free-energies is expected to be very important for strongly constrained systems, such as glycan chains bound in protein pockets and subjected to enzymatic reactions, where ring-inversion is often a key step of substrate activation before e.g. hydrolysis of the adjacent linkages.[134] This issue will be investigated in chapter 5.

Regarding the chosen water model, we limited out study to the standard TIP3P model, due to its use for the parameterization of carbohydrate force fields, at least for the CHARMM family. However, more complex water models like TIP4P-Ew or TIP5P have been shown to positively impact the predicted carbohydrate aggregation and protein-carbohydrate interactions.[239–241] A recent review about modeling of complex carbohydrates summarizes the difficulties that might arise from the usage of different water models.[162] As the AMBER family does not depend on a specific water model for the parameterization of non-bonded interactions, more advanced water models could be tested to study glycan structures, although at the expense of larger computational cost.[165,239] The influence of ion parameterization can be largely ignored in this context, as low salt concentrations (less than 100 mM) are adequately modeled by default force field parameters.[165]

The employed enhanced sampling method ensures ergodicity of the performed simulations, their accuracy in predicting experimental observables however remain limited by the functional form and parametrization of the employed force fields. It was already discussed that optimization of parameters is an elaborate task and sometimes restricted to a boost in accuracy for only specific applications, but also new potential energy functions including more parameters like the Class II functional form do not represent a realistic option for further improvements due to the even larger set of variables to parameterize.[145] A different and much more straight forward approach would be the incorporation of available experimental data as a restrain in ensemble-restrained simulations, classically imposing harmonic restraints on each observable, with the experimental reference depicting the center.[165] Another very similar approach is the emerging concept of the maximum entropy principle, either reweighting simulations *a posteriori* or optimizing them on-the-fly by experimental parameters in order to enforce a certain ensemble average.[242] However, care has to be taken to prevent overfitting.[165] The general limitation of these approaches is the requirement of sufficient and accurate experimental data sets like RMSD values, NMR parameters or scattering data. Moreover, these approaches would just improve the performance of single simulations and not lead to a general improvement of transferable force field parameters, loosing the achieved progress with every new simulation system.

The here-examined N-glycans have been the subject of several previous investigations.

A2G2S2 has been studied by Yang and coworkers[94] using REST2 in combination with Hamiltonian bias potentials, employing the CHARMM36 force field. This approach is similar to REST-RECT, the only difference being that Yang and coworkers used biasing profiles on the torsional angles as obtained in preliminary umbrella sampling simulations, whereas in our RECT scheme the compensating profiles are computed on the fly. In fact, the reported free energy profiles for individual torsion angles on ref.[94] are overall in agreement with our CHARMM36 simulations, demonstrating converged phase-space sampling in both studies.

A2G2 was previously investigated by REMD using the GLYCAM06g force field, and the predicted relative populations of the ω angle (O6–C6–C5–C4) in branch **6**– amounted to 71 % and 28 % for the *gg* and *gt* conformers, respectively.[96] Our simulations with the GLYCAM06j force field, however, gave average values of 80 % for *gg*, 11 % for *gt* and 9 % for *tg*. While the force field versions g and j only differ in the atom labeling for consistency with other AMBER force fields or in the addition of parameters for protein-carbohydrate linkage, the simulations by Nishima and coworkers[96] differ from ours with respect to the type of sampling method. We believe that our REST-RECT simulations provide a more complete phase-space sampling, as demonstrated by the very good convergence and the clear independency of the chosen initial configurations. Other earlier investigations of A2G2 using the CHARMM36 force field revealed a distribution of 52 % *gg* vs 48 % *gt* conformations in the ω torsion angle (O6–C6–C5–O5).[87] Our values of 71 % for *gg* and 29 % for *gt* computed with CHARMM36, however, are closer to the experimentally estimated values of 65 % and 35 %, respectively.[87,243] As identical force field parameters were used in both studies, the associated differences can have multiple reasons, namely (i) the use of the sTIP3P water model in contrast to the mTIP3P model used here, again pointing towards the need of better assessing the performance of different solvent models; (ii) incomplete phase-space sampling in the earlier simulations; (iii) the fact that in their simulations Galvelis and coworkers prevented ring inversion for all monosaccharide rings, although there are hints about a possible influence of puckering states on the glycan linkage conformations.[237,244]

3.2.4 GlycoSHIELD

An immediate need for a sufficiently well explored phase space arose during the development of the GlycoSHIELD software, mainly developed by our collaborator Mateusz Sikora from the Max Planck Institute of Biophysics.[221] GlycoSHIELD is a web application (`https://mpibp-hummer.pages.mpcdf.de/glycoshield-md/`) that allows for a quick and easy covalent glycosylation of proteins at identified glycosylation sites, yielding an ensemble of possible glycan conformations. The user can upload a protein structure in pdb format, specify where the residue is supposed to be attached, and select the desired glycan type from a set of diverse glycan structures (Figure 3.15). A conformer library, harboring dozens of conformations for a certain glycan, are step-by-step grafted on the uploaded protein structure, where conformations are discard upon structural clashes with surrounding amino acids. The outcome is a selection of possible glycan conformers virtually attached to the protein surface, helping in the evaluation of the molecular shield that is created by glycosylations and its influence on masking interaction sites of antibodies, drugs or other biomolecules.

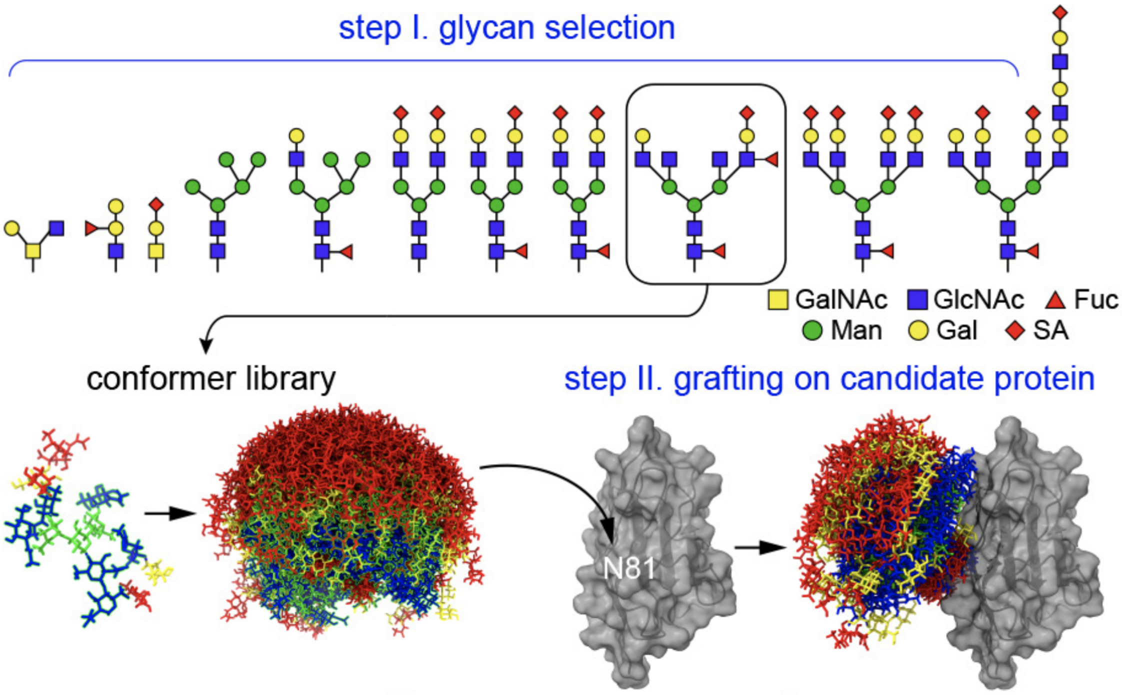


Figure 3.15: **GlycoSHIELD workflow.** The user needs to input a 3D protein structure with identified glycosylation sites, where glycans from the library of conformers not clashing with the protein are grafted onto the surface. Results are exported in pdb format and can be further visualized and analyzed by software packages like VMD. Figure adapted from Mateusz Sikora.[221]

In order to predict reliable results, the conformer library must contain all accessible conformers that a certain glycan is able to adopt. Conformational ensembles can be generated using standard MD simulations, however the exploration of the whole phase space region would be questionable, as argued before in this thesis. Therefore, REST-RECT simulations of the three representative N-glycans, M5 A2G2S2 and A4G4S4, differing in size and composition, were compared to previously performed MD simulations in order to ascertain if conformer distributions from unbiased trajectories are sufficient for a confromer library construction (Figure 3.16). The low-dimensional free energy profiles, depicting explored conformers for each glycan type, revealed that indeed very long standard MD simulations access all conformers that were also sampled during REST-RECT simulations. However, the individual conformer distributions were not sampled correctly, as expected, deviating in the depth of explored minima, especially for M5 and A4G4S4. As GlycoSHIELD only aims at generating a conformational ensemble that includes all possible glycan conformers but does not incorporate their individual distributions, unbiased MD simulations are sufficient in this case.

The novelty of GlycoSHIELD is represented by the fact that the web application can be run from any personal computer and only takes minutes to be performed. In contrast to MD simulations that require expert knowledge and days on high performance computers to obtain such conformational ensembles for bound glycans, the grafting method can be seen as an easy first step in the assessment of glycan influences before performing more elaborate studies.

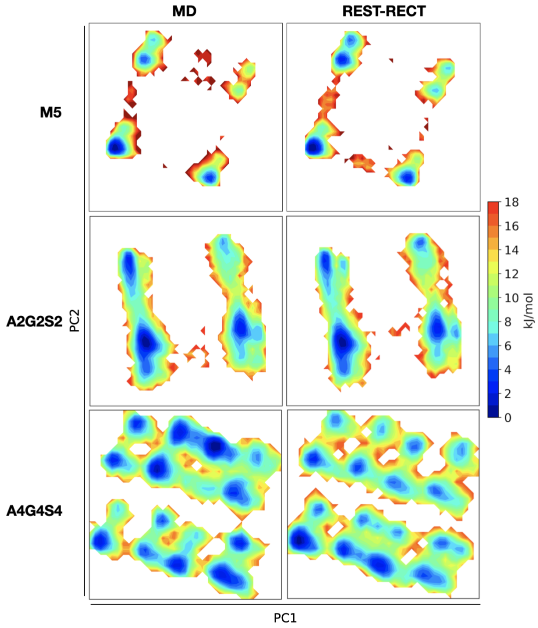


Figure 3.16: **Comparison of glycans' phase spaces explored by unbiased MD or REST-RECT simulations with respect to torsion angles.** Free energy profiles along principle components 1 and 2, for N-glycans M5, A2G2S2 and A4G4S4. PCAs were constructed by concatenating the datasets of the two simulations techniques for each glycan type, respectively, using all torsion angle as input features.

First instance, viruses employ extensive glycosylation to escape the immune system by masking interaction sites of their viral surface proteins, preventing the interaction with e.g. antibodies. Our collaborators compared the performance of GlycoSHIELD to that of classical MD simulations in regards to the algorithm's ability to produce realistic glycan shields. Although the correct conformational distributions of the glycan structures are not reproduced by GlycoSHIELD, which implements just a random selection of conformers that fit the spatial constraints, the methodology is able to capture important features of glycan shielding like the epitope masking of the SARS-CoV-2 Spike protein (Figure 3.17).[221]

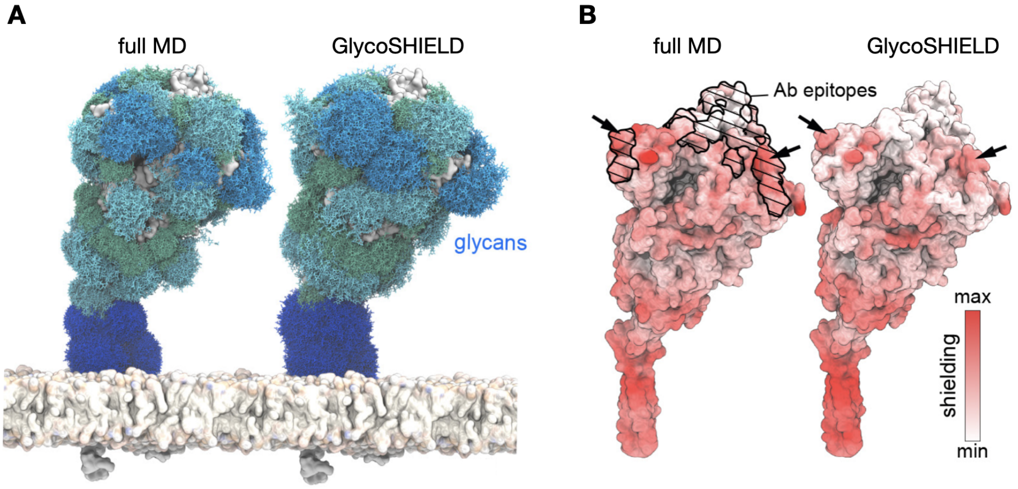
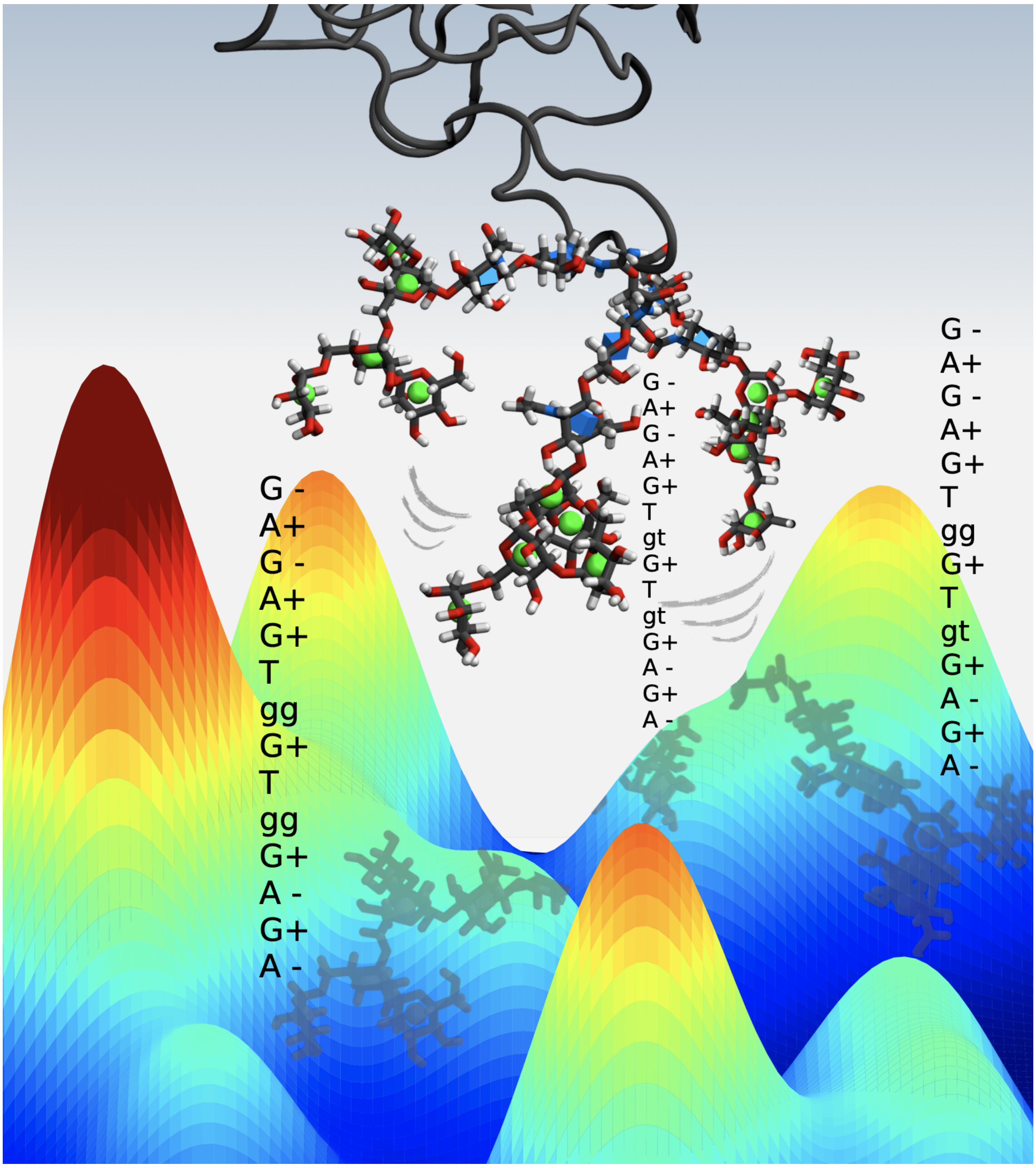


Figure 3.17: **Prediction of SARS-CoV-2 Spike protein epitope accessibility by GlycoSHIELD. A** The Spike protein was either glycosylated and sampled via MD simulations for 10 µs or the glycosylation was reconstructed via GlycoSHIELD. In total, 160 glycan conformers are visualized per glycoslyation site. **B** For both systems, the shielding of the extracellular domain was calculated due to the presence of glycans on the surface. The 3D heatmaps visualize accessibility, where higher color intensities indicate higher shielding. Arrows indicate predicted shielded areas within specific antibody epitopes (black lines and hatched areas). Figure adapted from Mateusz Sikora.[221]

It turned out that the overall morphology of glycan shields obtained was very similar for both methods (Figure 3.17 **A**) and that GlycoSHIELD even correctly predicted the epitope masking areas of N-glycans in comparison to data from MD simulations (Figure 3.17 **B**). GlycoSHIELD represents a very useful tool in the rapid generation of glycosylated proteins in a static fashion that is anyway able to predict realistic glycan shields.

4 | At the side: *#gotglycans*



The artistic phase space of a bound N-glycan. The high-mannose type N-glycan M5 is exploring different conformations although covalently attached to a polypeptide, shifting the free energy landscape compared to a free N-glycan. Each conformer is flagged by corresponding conformer labels.

Note: *Parts of this chapter are taken from the publication: Jana Rosenau*, Isabell Louise Grothaus*, Yikun Yang, Nilima Dinesh Kumar, Lucio Colombi Ciacchi, Sørge Kelm, Mario Waespy, N-glycosylation modulates enzymatic activity of Trypanosoma congolense trans-sialidase, Journal of Biological Chemistry, 298:102403, 2022.*[245] *(* shared co-first authorship)*

Insights obtained in this chapter are due to the collaborative efforts of the working groups Kelm and Colombi Ciacchi at the University of Bremen. Especially Jana Rosenau (formerly Ph.D student in the Kelm lab) and myself were involved with the following contributions: Protein purifications and enzyme activity assays have been alternately performed, with joint effort regarding data analysis and interpretation. The enzyme data for replicate 2 are from Nilima Dinesh Kumar. Circular dichroism experiments and MD simulations have been all performed by myself.

In the early months of the COVID-19 pandemic the Twitter hashtags *#gotglycans* and *#glycotime* became famous, emphasizing the urgent need of including glycan structures in biomolecular simulations and supporting the fact that the century of glycans has begun. The hype goes back to the discovery of the SARS-CoV-2 spike protein glycan shield (Figure 3.17), found to be highly important for the recognition and docking to human cells solely via MD simulations.[100] The computational studies of that time are groundbreaking for future work, highlighting the importance to include post-translational modifications explicitly in all-atom MD simulations, especially glycans. We have already seen in the previous chapter that force fields are being continuously developed to meet the requirements of simulating complex biomolecular systems with different types of molecules, and that they are mostly capable of meeting experimental standards for unbound glycans in solution. The simulation of diverse systems is further facilitated by the automated generation of *in silico* glycoconjugates like glycoproteins via the CHARMM-GUI Glycan Modeler[202], constantly including more features and broadening the number of possible glycan structures to attach.[203–205] Such technical advances and guiding studies are required in order to push forward the understanding of N-glycan's impact on proteins and enzymes, especially computationally. In this chapter, we will have a deeper look on glycosylated trans-sialidase (TS) enzymes, since the effect of N-glycosylation on their activtiy has been neglected as much as the disease in which they play a pivotal role, until now.

Trypanosoma, a parasite genus causing the disease trypanosomiasis, is prevalent in sub – Saharan Africa and South America, depending on the specific trypanosomal species. The parasite is infecting mammals like humans and livestock, where the disease can be either lethal or at least cause fatal economic losses in the agricultural sector (see section 1.5.2). The outer surface of the parasite, and in particular the glycosylphosphatidylinositol (GPI)-anchored enzyme TS, was identified as a major virulence factor for the disease and has been the object of several studies aiming at understanding its function and fundamental biochemical mechanism. The term trypanosomiasis comprises several clinical pictures and diseases, such as the sleeping sickness in Africa and the Chagas disease in South America, all caused by protozoa of the genus *Trypanosoma.* We here focus on the African trypanosome *Trypanosoma congolense* (*T. congolense*), infecting all kinds of animals like cattle, horses, goats, sheeps, pigs and dogs, whereby even atypical human infections could be identified quite recently.[246] The parasite is transmitted between different mammals by tsetse flies serving as vectors, biting the host and ingesting or releasing the parasites through their blood meal. African trypanosomes undergo different life cycle stages, depending on whether they are in the tsetse fly intestines (procyclic trypomastigotes and epimastigotes), vascular host system (metacyclic and bloodstream trypomastigotes) or multiply in other body fluids (bloodstream trypomastigotes).[247] Without going into too much detail, it is important to note that TS enzymes possessed by *T. congolense* are expressed by procyclic insect-infective trypanosomes as well as bloodstream-form trypanosomes in mammalian hosts.[248,249] In both life cycle stages, TS are of utmost importance as they ensure the survival of the parasite under diverse environmental conditions by the preferential transfer of $\alpha 2 \rightarrow 3$-linked Sia residues from host-cell glycoconjugates to terminal β-galactose residues of glycoproteins present on their own surface, thus creating a new $\alpha 2 \rightarrow 3$-glycosidic linkage.[103–105] It was shown that the parasite uses terminal Sia residues to mask its surface, evading the digestive and trypanocidal environment in the tsetse fly gut, as gene technical deletion of endogenous TS expression in trypanosomes and consequent absence of Sia on the surface had lethal effects.[112] When present in its bloodstream-form in the vascular system, *T. congolense* is able to attach to erythrocytes by binding to Sia residues as shown in an *in vitro* study.[250] Furthermore, TS are involved in desialylation of host erythrocytes, which contributes to anemia and therefore can cause direct symptoms.[248,249]

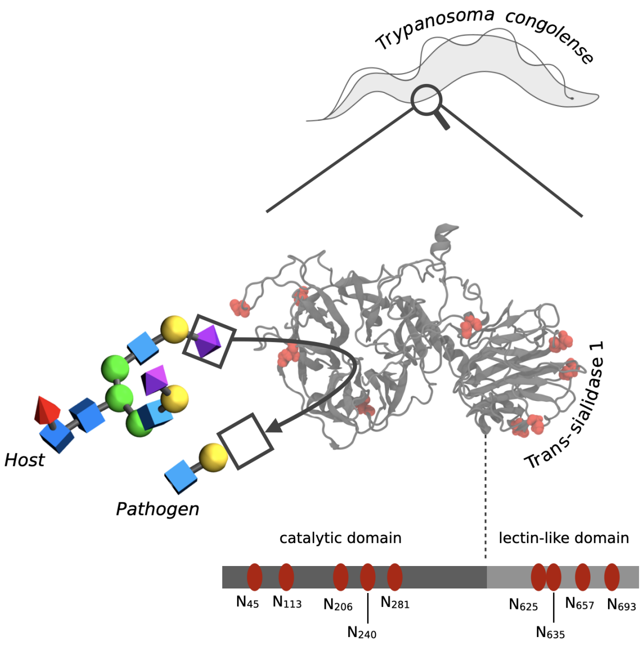


Figure 4.1: **Trans-sialidase as a model for surface modifications via N-glycosylations and concurrent glycan processing.** Molecular model of Trans-sialidase 1 originating from *Trypanosoma congolense* (TconTS1) with highlighted N-glycosylation sites. Asparagine residues in the motif N-X-S/T as putative N-glycosylation sites are highlighted in red. The position of each asparagine is labelled in the amino acid sequence and sorted into the catalytic domain (CD) or the lectin-like domain (LD).

T. congolense harbors 17 different trans-sialidase-like genes, from which 11 can be combined into the *T. congolense* TS (TconTS) family 1, due to their high amino acid sequence similarity (>96 %). The type 1b (TconTS1b) is more closely investigated here, as it has been isolated from procyclic trypomastigotes and possesses one of the highest enzyme activities among the other TconTS families.[251–253] For simplicity, it will be called TconTS1 from here onwards. Like all trans-sialidases, TconTS consist of an N-terminal catalytic domain responsible for the transfer of Sia, and of a C-terminal lectin-like domain whose biological function remains rather unclear (Figure 4.1). The catalytic and lectin-like domain are connected via an

α-helix.[116,117,251] The N-terminus includes a signal sequence for cell secretion, whereas the C-terminus comprises a potential GPI anchor attachment site.[251] An important structural feature of TconTS1 are the predicted nine N-glycosylation recognition sequences (N-X-S/T) (Figure 4.1), distributed across both the catalytic and lectin-like domain.[251] It is noteworthy that TS sequences from African trypanosomal species contain a higher number of putative N-glycosylation sites[251,254] compared to the species from South-American trypanosomes including the structurally closely related *T. rangeli* sialidase (TranSA), although little has been revealed about the impact of glycans on TS enzymes.[245,255,256] Structurally, it has been only postulated that N-glycans could be involved in TconTS oligomerization via binding to a lectin-like domain on the enzyme, creating di- to tetrameric complexes.[257] Their influence on enzyme functionality is still under debate for TS from all species.[116,117,258,259] Previous experiments were performed with recombinant TS expressed by *Pichia pastoris* producing hypomannosylated N-glycans, revealing no differences in Sia transfer activity comparing glycosylated and deglycosylated recombinant TS.[258,259] As these larger fungal N-glycans are different from reported glycan structures on other trypanosomal surface proteins, which usually harbor shorter high-mannose type N-glycans, no conclusion about the influence of native N-glycosylation on TS activity can be drawn from these studies.[122–129] The data available implies that in general not only the impact of putative N-glycans on TS enzymes is unresolved, but also that the type and extent of glycosylation is not known for the different species. For instance, the only crystallized TS from *T. cruzi* (TcruTS) was expressed as a recombinant protein containing several mutations in the amino acid sequence, thus leaving the N-glycosylation pattern unresolved.[117] The crystallization of TranSA revealed that all five potential N-glycosylation sites were occupied with N-glycans, although only the innermost monosaccharide could be detected, giving no hint about the type of N-glycosylation.[255] Only purification via concanavalin A (ConA) columns indirectly confirmed the existence of high-mannose type N-glycans. However, the detailed structures and site-specific patterns remain unresolved with this method.[104,120,121] If one looks at the problem from another angle, namely assessing which N-glycans are theoretically possible due to the N-glycosylation machinery expressed in *Trypanosoma*, one is left with the description provided for *T. brucei*[260], producing all kinds of glycan structures, but there are no database entries on potential oligosaccharyltransferase isoforms for *T. congolense*. Therefore, the N-glycosylation pattern of TconTS1 had to be determined as a first step in order to draw conclusions from subsequent experiments, although only recombinant expression of TconTS was possible. A native expression presents the difficult obstacles of reduced protein yields, expensive culture conditions and requirements of a biosafety level 2 laboratory.

4.1 Are they there?

Since other trypanosomal surface proteins were reported to harbor shorter high-mannose type N-glycans of type M5-9 (5-9ManGlcNAc2)[122–129], we expressed recombinant TconTS1 in leuco-phytohemagglutinin (L-PHA)-resistant Lec1 Chinese hamster ovary (CHO) cells.[261,262] This N-glycosylation mutant cell line is unable to synthesize complex and hybrid N-glycans, and consequently accumulates high-mannose type N-glycans of the composition M5-9 (Figure 3.2), mimicking the situation reported for African trypanosomes.

A recombinant TconTS1, harboring a SNAP-Strep tag (see Figure S1 and S2 in Rosenau et al.[245]), was expressed in monoclonal CHO Lec1 cells, transfected before by Koliwer-Brandl et al.[251], and grown in serum-free CHO medium (Bio&SELL, Feucht, Germany) or Excell medium supplemented with 50 μg/mL gentamicin sulphate (Lonza™ BioWhittaker™, Walkersville, MD, USA) at 37°C and 5% CO2. Due to a transin secretion tag, the protein could be harvested every second day from the cell culture supernatant and subsequently stabilized with 10 mM Tris/HCl pH 8.0, 10 mM EDTA, 10 mM ascorbic acid and 0.02 % sodium azide. Ultracentrifugation was performed at 7,800 rcf for 15 min followed by 40,000 rcf for 45 min at 4 °C to get rid of any cell debris. The clear supernatant was microfiltered (0.22 µm, PES) and concentrated to 50 mL using a Sartorius Vivacell 250 PES Centrifugal Concentrator (Sartorius, Göttingen, Germany) with a Molecular Weight Cut-off (MWCO) of 100 kDa and a pressure of 4 bar. Buffer was exchanged five times with 200 mL of 100 mM Tris/HCl pH 8.0, 150 mM NaCl, 1 mM EDTA and concentrated to a final volume of 10 mL. The concentrate was centrifuged at 21,000 rcf for 30 min and the recombinant protein was purified from the supernatant with Strep-Tactin sepharose (IBA, Göttingen, Germany) according to the manufacturer's protocol, after that the buffer was exchanged to 10 mM potassium phosphate buffer pH 7.4 using a Vivaspin 6 Centrifugal Concentrator (Sartorius) at 2,000 rcf and 4 °C for 20 min.

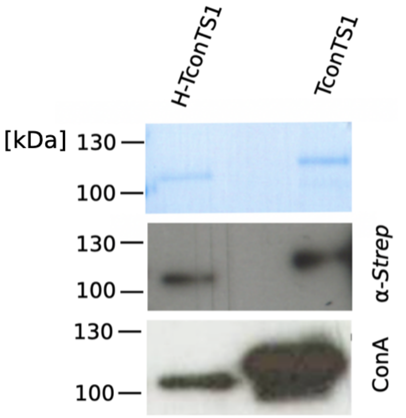


Figure 4.2: **Quality control of protein purification.** TconTS1 and H-TconTS1 (treated with EndoH for 4 h) were analyzed by SDS-PAGE with subsequent Coomassie staining (upper panel with 600 ng of protein), western blot analysis using an anti-*Strep*-tag antibody for the detection of the protein (middle panel with 400 ng of protein) and lectin blotting using ConA for the detection of high-mannose type N-glycans (lower panel with 100 ng of protein). The exposure time was 5 sec for the western blot and 60 sec for the ConA blot.

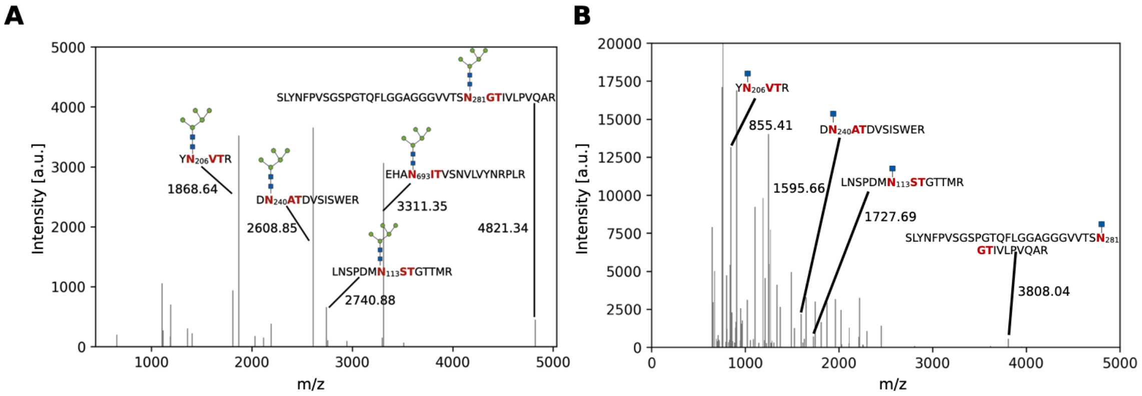
In order to rate the performance of glycosylated TconTS1, a comparison to an unglycosylated variant was required. The most obvious option, introducing point mutations at single N-glycosylation sites, has not been considered in this work, as it was important to ensure a correct folding and function of TconTS1, and co-translational N-glycosylation is known to influence protein folding. Numerous experiments performed by collaborators (Sørge Kelm lab) on site-directed glycosylation knockout using myelin associated glycoprotein (MAG; Siglec-4, unpublished data) have also yielded largely misfolded proteins with loss of function. Furthermore, TconTS exhibits orders of magnitude lower specific activity after expression in bacteria in comparison to expression in CHO-Lec1 cells, providing evidence for misfolding of the enzyme as a consequence of absence of N-glycans.[263] Therefore we aimed for the selective removal of N-glycans, achieved enzymatically by Endoglycosidase H (EndoH) treatment. This cleaves high-mannose type N-glycans within the chitobiose core and leaves only one GlcNAc residue attached to the asparagine in the protein sequence. To this aim, 2 mg of TconTS1 were incubated with

40,000 Units EndoH (New England Biolabs, Frankfurt am Main, Germany) at 37 °C in 2.0 mL 10 mM phosphate buffer at pH 7.4. Subsequent chromatography steps, necessary to remove EndoH from the target protein in the sample, were performed according to the manufacturer's protocols (see caption of Figure 4.4). Finally, the buffer was exchanged again to 10 mM phosphate buffer at pH 7.4. The resulting EndoH-treated enzyme will further be termed H-TconTS1 (hypoglycosylated TconTS1), as described in the following section.

Recombinantly expressed protein samples, TconTS1 and H-TconTS1, were checked for successful purification and deglycosylation via sodium dodecylsulfate polyacrylamide gel electrophoresis (SDS-PAGE), western blot and ConA lectin blot analysis (Figure 4.2). Samples separated via SDS-PAGE were either stained with PageBlue Protein Staining Solution (Thermo Fisher Scientific) or used for western blot or ConA lectin blot analysis. For the specific detection of TconTS1, a polyclonal rabbit anti-Strep (IBA) and a polyclonal, peroxidase-conjugated donkey anti-rabbit antibody (Jackson ImmunoResearch, Cambridgeshire, United Kingdom) were used as primary and secondary antibody in western plot analysis, respectively. In lectin blots, N-glycosylated proteins harboring high-mannose structures were detected employing ConA-biotin (Galab, Hamburg, Germany) and the VECTASTAIN ABC-HRP Kit (Vector Laboratories, Burlingame, CA, United States). A clear band shift of approximately 10 kDa between TconTS1 and H-TconTS1 samples could be inferred from all three blotting techniques, being a result of the removed N-glycans from H-TconTS1, leading to a lower molecular weight and altered migration behavior in the gel (Figure 4.2). Clear bands at 120 and 110 kDa in the western blot confirmed the presence of our desired TconTS samples due to the C-terminal *Strep*-tag in the recombinant TconTS construct. High-mannose type N-glycans of the recombinant protein were detected by ConA lectin blots, where less binding of ConA to EndoH-treated TconTS1 was indicated by a much weaker signal, however still causing a visible band for H-TconTS1(Figure 4.2). Due to the lack of complete N-glycan removal after 4 hours of incubation, different incubation times of up to 48 hours or higher amounts of EndoH enzyme were tested. Complete removal of high-mannose type N-glycans could still not be achieved, probably due to low accessibility of certain N-glycan structures to EndoH. For this reason, subsequent experiments were performed after overnight incubation with EndoH for 16 hours, yielding H-TconTS1 samples which represent a hypoglycosylated version of the target protein.

After having verified that N-glycans were present on recombinantly expressed TconTS1, the site specific glycosylation pattern was analyzed qualitatively by matrix assistant laser desorption ionization – time of flight (MALDI-TOF) mass spectrometry (MS). The distribution of oligosaccharides was evaluated both for TconTS1 and for H-TconTS1 samples (Figure 4.3 **A/B**). As the majority of MALDI-TOF experiments were performed during my Master's thesis in the lab of Sørge Kelm at the University of Bremen under the supervision of Jana Rosenau, the results will only be briefly discussed here to report about the glycosylation pattern of TconTS1, which is important for the remaining investigations in this chapter. A more elaborate analysis and technical details can be found in Rosenau et al.[245]

In order to analyze TconTS1 and H-TconTS1 by MALDI-TOF MS, both enzymes were proteolytically digested to shorter peptide and glycopeptide fragments via trypsin or chymotrypsin treatment. N-glycans could be detected because whenever an asparagine residue in the N-X-S/T motif of a particular glycopeptide was glycosylated, the mass-to-charge (m/z) ratio increased by exactly the mass of the conjugated N-glycan compared with the non-glycosylated peptide. Glycopeptides comprising high-mannose type N-glycans were detected especially in the catalytic domain of TconTS1 with N206 showing the highest diversity (Figure 4.3 **C**).



C

Modification	N_{45}	N_{113}	N_{206}	N_{240}	N_{281}	N_{625}	N_{635}	N_{657}	N_{693}
Non-glycosylated		G	G	G		G/H			H
M5	G	G	G/H	G	G			G	G
M6			G/H						
M7			G						
M8			G/H						
FM5			G						
HexNAc	H	H	H	H				H	H

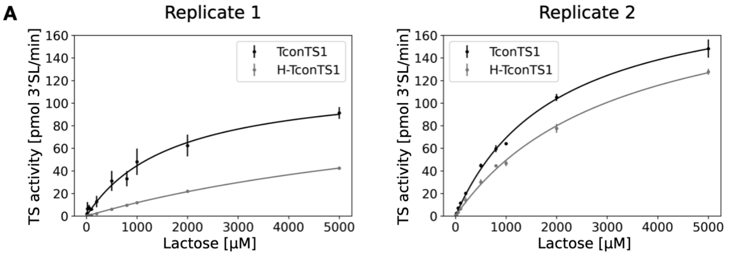
Figure 4.3: **Mapping the N-glycosylation profile of untreated (TconTS1) and hypoglycosylated TconTS1 (H-TconTS1).** MALDI-TOF MS analyses were performed to map N-glycosylation sites of TconTS1. Glycopeptides from protease-digested **A** TconTS1 and **B** H-TconTS1 were ConA-purified to concentrate glycopeptides and reduce the spectrum complexity. Peak lists were extracted from MALDI-TOF mass spectra, plotted with python and annotated with corresponding masses and glycopeptide fragments, respectively. Monosaccharide symbols follow the Symbol Nomenclature for Glycans (SNFG). **C** Summary of N-glycan structures identified for TconTS1 (glycosylated, G) and the EndoH-treated H-TconTS1 (H), digested either with trypsin or chymotrypsin. The spectra were analyzed for masses corresponding to glycopeptides with high-mannose type N-glycans and for non-glycosylated peptides with potential N-glycosylation sites. Spectra of H-TconTS1 were additionally analyzed for glycopeptides with HexNAc residues since at least a residual GlcNAc remains attached to the protein N-glycosylation sites after EndoH treatment. In another approach, glycopeptides from both proteins were purified with ConA after protease digestion and spectra were analyzed for masses of peptides with high-mannose type N-glycans.

The mass difference corresponding to an N-glycan of the composition M5 was predominantly detected at sites N45, N113, N240, N281, N657 and N693 (Figure 4.3 **A**). Interestingly, also non-glycosylated peptides were detected for four glycosylation sites (N113, N206, N240 and N693) that were also found glycosylated, underlining the dynamical and heterogeneous nature of the N-glycosylation process. Peptides containing N625 were only found in their non-glycosylated form. When analyzing H-TconTS1, HexNAc residues were detected for all previously glycosylated sites except for N281, indicating the successful removal of N-glycans from almost all glycosylation sites with one GlcNAc residue remaining attached (Figure 4.3 **B/C**). However, oligomannosidic N-glycosylation at N206 was still detectable after 16 h of EndoH treatment (H entries in N206 column for M5,M6 and M8 in Figure 4.3 **C**), supported by our ConA lectin blot results, demonstrating the binding of ConA to H-TconTS1 at long exposure times (Figure 4.2). It can be concluded that the compiled data reveal N-glycosylation predominantly in the catalytic domain of recombinant TconTS1 with diverse site-specific glycosylation tendencies, including the presence or absence of various high-mannose type structures.

4.2 What are they doing?

After revealing the pattern of N-glycans on the surface of TconTS1, we did investigate whether this had an impact on the functionality of the enzyme. The influence on enzyme activity was monitored in activity assays using fetuin as Sia donor and lactose as acceptor substrate, as previously described.[251,253] The transfer reaction product 3'sialyllactose (3'SL) was quantified by high performance anion exchange chromatography (HPAEC) with pulsed amperometric detection (PAD), allowing for the separation and detection of the educt lactose, the product 3'SL and free Sia residues, which can also be caused by a transfer to water instead of a carbohydrate acceptor. A detailed explanation of HPAEC-PAD and calculated kinetic parameters is given in the appendix C. TconTS1 and H-TconTS1 (50 ng) samples were incubated for 30 minutes at 37 °C with 600 µM fetuin-bound Sia (100 µg dialysed fetuin) and varying concentrations of lactose (0.01-5 mM) in 50 µL of 10 mM potassium phosphate buffer, to calculate the corresponding Michaelis-Menten kinetic parameters K_M and v_{max}. The used Sia concentration corresponds approximately to one third of the Sia concentration reported for glycoproteins in blood serum[264,265]. However, human blood serum has a high proportion of protein-bound Sia with $\alpha 2 \rightarrow 6$-linkages that are not utilized by TconTS.[266,267] The reaction was terminated with 200 µL ice-cold acetone in order to yield protein precipitation, which was further carried out overnight at -20 °C. Subsequently, samples were centrifuged (20 000 rcf, 30 min, 4 °C), the supernatant lyophilized and resuspended in 125 µL water. The HPAEC-PAD system ICS-5000+ (Dionex/Thermo Fisher Scientific) was used to apply 25 µL sample to a CarboPac100 analytical column (250x2 mm, 8.5 µm, Thermo Fisher Scientific), equipped with a guard column (50x2 mm, Thermo Fisher Scientific). Chromatography steps were performed at isocratic conditions with 100 mM NaOH and 100 mM NaOAc for 12 min, followed by a wash step with 100 mM NaOH and 500 mM NaOAc for 5 min and an equilibration step for 8 min to previous conditions. Production of 3'SL was quantified with a purchased 3'SL standard (Carbosynth, Compton, United Kingdom). Data acquisition and evaluation was performed with the Dionex software Chromeleon 7.2 SR5 and parameters of the Michaelis-Menten equation, K_M and v_{max}, were calculated with the curve fit model of SigmaPlot11. Two biological replicates, with three technical replicates each, revealed a lower amount of 3'SL produced by H-TconTS1 relative to TconTS1 (Figure 4.4 **A**). Interestingly, the calculated v_{max} values for the Sia acceptor substrate lactose of about

2.4 µmol 3'SL/(min x mg enzyme) for replicate 1 and 4.1 µmol 3'SL/(min x mg enzyme) for replicate 2 were found to be identical for glycosylated und deglycosylated enzymes within the error range, highlighting the structural integrity of H-TconTS1 (Figure 4.4 **B**).



B

	TconTS1	H-TconTS1
k$_{cat}$ [s^{-1}]	4.44 ± 0.31	4.36 ± 0.16
V$_{max}$ [µmol 3'SL/mg TS x min]	2.42 ± 0.17	2.38 ± 0.09
K$_M$ [µM]	1715 ± 265	9069 ± 466

	TconTS1	H-TconTS1
k$_{cat}$ [s^{-1}]	7.59 ± 0.24	7.64 ± 0.29
V$_{max}$ [µmol 3'SL/mg TS x min]	4.14 ± 0.13	4.17 ± 0.16
K$_M$ [µM]	1989 ± 128	3222 ± 229

Figure 4.4: **EndoH-treated H-TconTS1 shows higher K$_M$ compared to TconTS1. A/B** TS activities for TconTS1 and H-TconTS1 were determined using fetuin as Sia donor and a lactose concentration series as Sia acceptor. Production of 3'sialyllactose was monitored and Michaelis-Menten kinetic parameters, apparent K$_M$ and v$_{max}$ as well as k$_{cat}$ for lactose were evaluated using SigmaPlot11. Data points are means ± standard deviation of three technical replicates for each biological replicate. The two replicates were treated differently in the following way: The H-TconTS1 sample of replicate 1 was purified after EndoH treatment using Strep-Tactin sepharose to remove free glycans and EndoH after that the buffer was exchanged to 10 mM phosphate buffer at pH 7.4 as already described. In contrast, the H-TconTS1 sample of replicate 2 was purified by three chromatography steps, (i) a PD-10 desalting column (Cytiva, Marlborough ,USA) to remove free glycans, (ii) by AffiSep® ConA adsorbent (Galab) to remove remaining proteins with high-mannose type N-glycans and (iii) an amylose affinity purification (New England Biolabs) to remove EndoH by its fused maltose-binding protein.

These results further indicate that the 3'SL production rate of TconTS1 at saturated Sia acceptor concentrations is not influenced by its N-glycosylation state. Differences in v$_{max}$ values between the replicates might be explained by varying amounts of active enzyme after purification and therefore the kinetic parameters were calculated separately for each replicate. The K$_M$ values for lactose differed between TconTS1 (1.7 and 2.0 mM) and H-TconTS1 (9.0 and 3.2 mM), indicating a 1.6- to 5-fold lower Sia acceptor substrate affinity for H-TconTS1 relative to TconTS1 (Figure 4.4). The K$_M$ values determined in this study are similar to the one published by Koliwer-Brandl et al. of 1.7 mM[251], where variations in the K$_M$ of H-TconTS1 between replicates might be a result of differences in the number of N-glycans remaining on these hypoglycosylated TconTS1 preparations.

Comparing our findings to other similar studies shows that most of them did not focus on detailed kinetic parameters and therefore might have missed the effect of deglycosylation. For instance, Haynes et al.[258] studied the influence of N-glycosylation on *T. vivax* TS (TvivTS1) and did not observe an effect on enzyme activity. The same applies to investigations of a mutated variant of TranSA, which expresses TS activity.[259] However, these studies did not determine the K$_M$ values for the substrates used in enzyme reactions. Another study of TranSA, in which the sialidase activity was investigated, did not observe strong effects on K$_M$ when recombinant proteins were expressed in *Escherichia (E.) coli* and compared with the native enzymes isolated from trypanosomes.[268] However, the sialidase activity was determined in the absence of a Sia acceptor substrate such as lactose. For TcruTS, enzymes expressed in *E. coli* still showed transfer activity, although to a lesser extent than observed for the native protein, which might be a result of the absence of N-glycans and/or incorrect protein folding.[256]

4.3 How are they doing it?

The revealed correlations between the N-glycosylation status of TconTS1 and the enzyme's substrate affinity gave rise to the question about how the modulation takes place. In order to guarantee that the post-folding removal of N-glycans did not affect the enzyme stability, we performed circular dichroism experiments to investigate the influence of N-glycans on the TconTS1's secondary structure stability. A detailed introduction into circular dichroism is given in the appendix B. The Applied Photophysics Chirascan spectrometer (Applied Photophysics Limited, Leatherhead, UK) with the Pro-Data Chirascan software (v.4.2.22) was used to evaluate circular dichroism spectra, where at least three repetitive scans over a standard wavelength range of 190 to 250 nm with intervals of 1 nm were performed. Throughout the experiments, Suprasil quartz cells (Hellma UK Ltd.) were used with a pathlength of 0.2 mm. Baseline scans were performed with 10mM phosphate buffer (pH 7.4) only and the baseline subtracted from recorded spectra. Repetitive scans were averaged before a Savitsky-Golay smoothing filter with smoothing windows of three data points was applied. Estimates of secondary structural components were predicted from the circular dichroism spectra using the BeStSel Web server.[269,270]

Temperature-ramping experiments were performed following the suggestions of Norma Greenfield[271] in order to analyze protein stability, unfolding intermediates and the midpoint of the unfolding transition (melting temperature, $\mathrm{T_M}$). In detail, protein samples were heated from 20 °C up to 95 °C with 5 °C temperature steps employing the stepped ramp mode. After 5 min of equilibration time at the respective temperature, at least three spectra were recorded and averaged. The $\mathrm{T_M}$ was calculated from the fraction of protein folded at any temperature (α) defined as:

$$\alpha = \frac{(\theta T - \theta U)}{(\theta F - \theta U)}, \tag{4.3.1}$$

where θT is the ellipticity at any temperature, θU is the ellipticity at the unfolded state and θF at the folded state. $\mathrm{T_M}$ is defined as the temperature at which $\alpha = 0.5$ and also referred to as the melting temperature.[271] In order to calculate α, we chose 195 nm as the wavelength to plot the recorded mean residue ellipticity Θ_{MRE} values against the temperature. Afterwards, the calculated α values were plotted with respect to the temperature and a sigmoid fitting curve was used to obtain a precise $\mathrm{T_M}$ value. As we did not observe a complete unfolding of TconTS1 in any temperature-ramping experiment, θU is defined as the average ellipticity of the two highest temperatures. Accordingly, θF was set as the average ellipticity that was recorded for the two lowest temperatures. Circular dichroism spectra of TconTS1 and H-TconTS1 were analyzed under similar conditions as used for the enzyme activity measurements (35 °C), showing no significant difference over the recorded wavelength range, indicating that both enzymes share the same common secondary structure (Figure 4.5 **A**). In fact, calculated secondary structural elements were identical in both cases, with 37% of β-sheets, 13% of α-helices, 11% of turns and 39% of unstructured (other) components. Spectra recorded during temperature-ramping experiments revealed a high heat stability for both protein preparations, as TconTS1 and H-TconTS1 still kept intact secondary structural elements up to 95 °C (Figure 4.5 **C/D**). The variation of the spectra intensity in the range between 190 and 210 nm during heating indicates a partial unfolding, taking place between 60 °C and 70 °C, with a (T_M) of about 62 °C for both proteins (Figure 4.5 **B**). Thus, TconTS1 and H-TconTS1 exhibit the same secondary element distribution and the same heat stability (as quantified by T_M for partial unfolding).

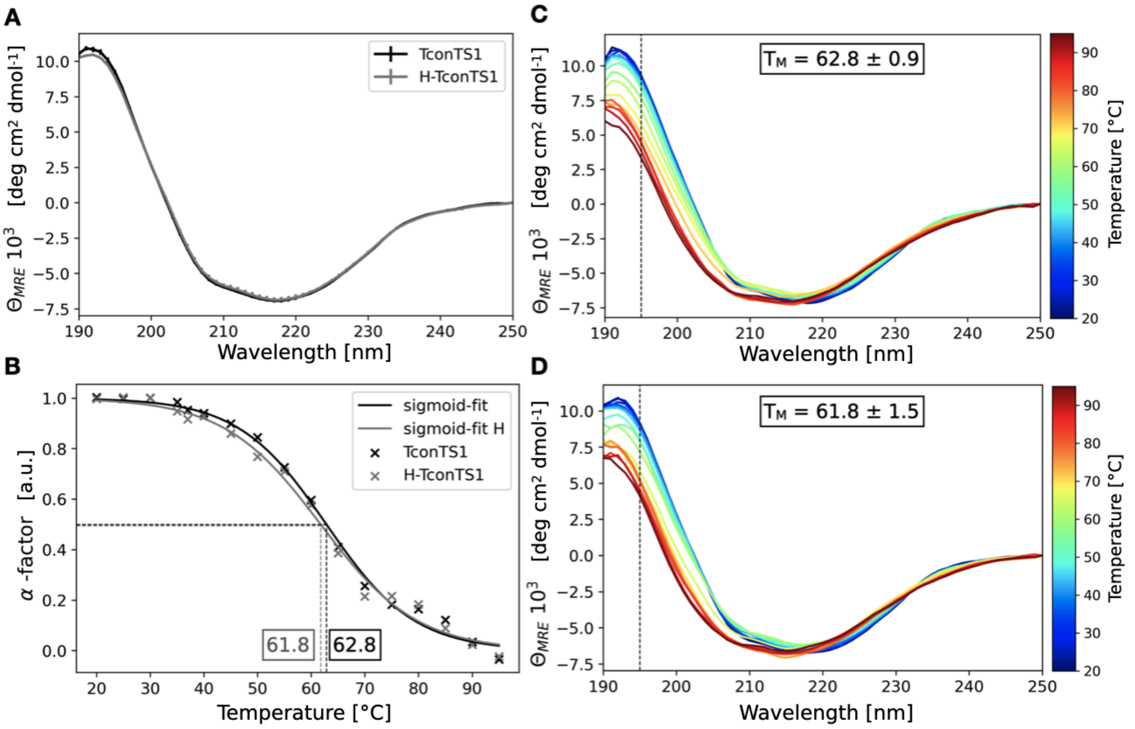


Figure 4.5: **Influence of N-glycans on the TconTS1 secondary structure and stability. A** Circular dichroism spectra of untreated TconTS1 and EndoH-treated H-TconTS1 were measured at 35 °C in 10 mM phosphate buffer at pH 7.4 (means ± standard deviation of two biological replicates). **B** The midpoint of unfolding ($T_M = \alpha - \text{factor} = 0.5$) from a folded state ($\alpha$-factor ≈ 1.0 at 20 °C) to a partially unfolded intermediate (α-factor ≈ 0.0 at 95 °C) was determined for TconTS1 and H-TconTS1 by fitting a sigmoid function to the data. Circular dichroism spectra of temperature-ramping experiments with TconTS1 **C** and H-TconTS1 **D** were recorded in 5 °C steps and a 5 min equilibration time at each step. The midpoint of unfolding (T_M) is determined at 195 nm (dashed line) at the flex point of a sigmoidal function fitting the temperature curve.

In summary, the circular dichroism experiments did not provide evidence for an altered overall secondary structure of H-TconTS1 as an explanation for its observed lower substrate affinity. N-glycosylation-induced changes of the tertiary structure still remain as a possible explanation because they are too subtle to be detected by circular dichroism. The observed thermal stability of TconTS1 might be explained by the high β-sheet content of the protein as well as by an extended interface between catalytic and lectin-like domain stabilized by salt bridges and a well-structured hydrogen bond network, making unfolding rather unlikely.[263] Although N-glycans do not seem to influence the stability of TconTS1 after successful expression, N-glycans are known to be required for proper folding of N-glycosylated proteins, regulated by the calnexin/calreticulin cycle in the ER.[272] Whether this is true for TconTS1 still needs to be investigated, although enzymatic activity of bacterially expressed TconTS provide evidence for glycosylation-dependent misfolding of the enzyme.[263]

As a final step, we employed MD simulations in order to unravel the N-glycan shield at an atomistic resolution, studying N-glycan dynamics on a site-specific level. Due to the lack of experimentally derived structures for TconTS enzymes, homology models were generated by the I-TASSER web server for protein structure and function predictions[198,273] based on the recombinant sequence without the transin signal (see Figure S1 and S2 in Rosenau et al[245]). The numbering of amino acids is in correspondence with the native sequence[251], although TconTS1 was modeled with the engineered SNAP-*Strep* tag for consistency and better comparison with experimental data, requiring restraints to achieve proper folding of this structural part. In detail, a secondary structure restraint as well as a structure template for the SNAP-*Strep* region, generated by I-TASSER beforehand using only the SNAP-*Strep* sequence, were employed. Validation of the TconTS1 homology model, including a discussion of employed templates and their amino acid sequence similarity, can be found in the appendix D.

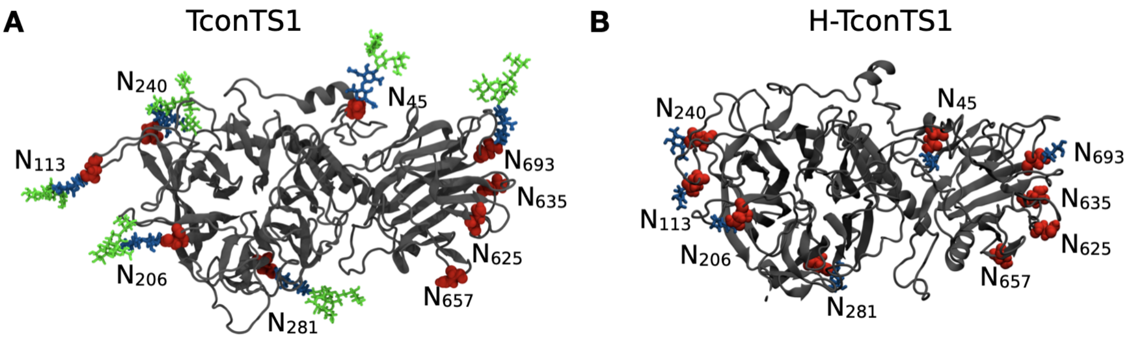


Figure 4.6: **Homology models of glycosylated and deglycosylated TconTS1.** Glycosylations have been chosen in accordance with the structures identified via MALDI-TOF MS. The protein backbone is represented in gray with the New Cartoon style. Carbohydrates in Licorice representation are colored with Man in green and GlcNAc in blue. Asparagine residues of putative N-glycosylation sites are colored in red.

In order to generate an N-glycosylated structural model of TconTS1, M5 glycans were included using the freely accessible CHARMM-GUI Glycan Modeler at positions N45, N113, N206, N240, N281 and N693, as identified by our MALDI-TOF MS experiments. M5 was chosen for all sites, as it represents the simplest and most often found N-glycan in CHO Lec1 cells.[262] The experimentally observed heterogeneity of N-glycosylation at N206 was not considered in our work at this stage. N657 was not glycosylated, although found in our MALDI experiments, because the model building was already completed at the time this glycan was found. A disulfide bond between residues C493 and C503 was formed. In TconTS1, all potential N-glycosylation sites are glycosylated in the catalytic domain, whereas only one out of four potential sites is glycosylated in the lectin-like domain (Figure 4.6 **A**). In the model used to simulate H-TconTS1, single GlcNAc residues were included at positions that were also occupied in the TconTS1 model, mimicking the residual monosaccharide after EndoH treatment (Figure 4.6 **B**). At first, standard MD simulations were performed for TconTS1 and H-TconTS1, in order to observe the dynamical behavior of the covalently linked N-glycans (Figure 4.7). The simulation box was constructed with CHARMM-GUI, filling it with water molecules to obtain a distance of 15 Å between the protein and box edge. 22 K+ ions were added for charge neutralization.

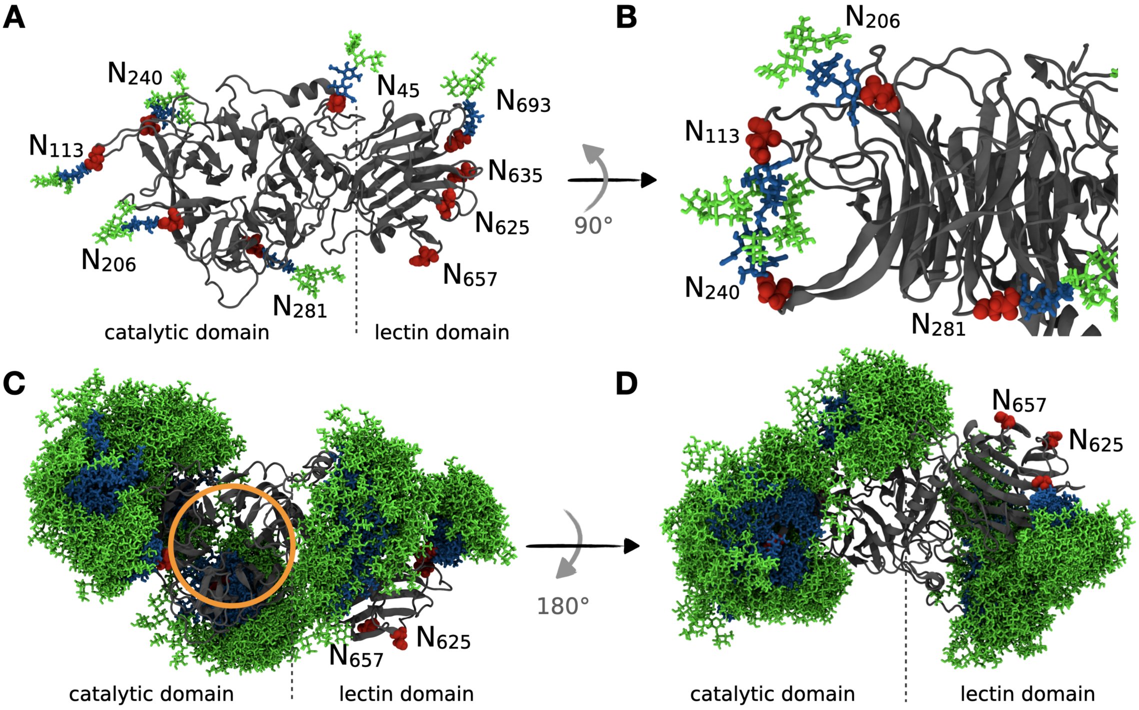


Figure 4.7: **Analysis of the dynamics of TconTS1's N-glycan shield. A** Atomistic model of TconTS1 with M5 N-glycans (Man: green, GlcNAc: blue) at the asparagine residues (red) identified in MALDI-TOF MS experiments. **B** Interactions between N240 and N113 glycans mediated by hydrogen bonds observed during MD simulations. **C** Overlay of all N-glycan positions recorded every 5 ns over a simulation time of 500 ns, with the protein backbone (gray) aligned in all frames and the active site indicated by an orange circle. **D** Same as **C**, with the protein rotated by 180°. The C-terminal SNAP-*Strep* tag is not shown in all structures.

All MD simulations were performed with the GROMACS 2018 version[206], using the CHARMM36m[150] force field for proteins and carbohydrates in combination with the TIP3P water model. The leap-frog algorithm was used as an integrator and the LINCS algorithm was employed to constrain bonds connected to hydrogens atoms.[210] Temperature coupling was performed with velocity rescaling using a τ parameter of 0.1 ps.[211] The Verlet cut-off scheme was employed for van der Waals parameters using PME and the standardized parameters suggested for CHARMM36 in the GROMACS manual version 2019.[212] Energy minimization of water and ions (with restrained protein) was performed using the steepest descent algorithm with a tolerance of 1,000 kJ mol^{-1} nm^{-1}. Equilibration of water (with restrained protein) was done in an NVT and an NPT ensemble for 1 ns, respectively. It followed the energy minimization of the protein (with restrained water and ions) under the same conditions as before. Finally, unrestrained equilibrations were performed under NVT and NPT for 1 ns each. The production runs were performed for 500 ns in the NVT ensemble at 310.15 K, writing coordinates to file every 10 ps. A time step of 2 fs was set for all simulations, if not mentioned otherwise. The systems were analyzed and visualized every 500 ps.

Six out of nine asparagine residues in the motif N-X-T/S are located at the tail of loop regions (Figure 4.7 **A**), which are mostly part of turns or coils framing β-sheet regions. The terminal position and flexibility of these structural elements allow for large motion amplitudes and internal flexibility of the N-glycan trees. These movements enable interactions among glycans in structural proximity, for instance intermolecular hydrogen bonds between the N-glycans at positions N113 and N240 (Figure 4.7 **B**), notwithstanding their distance in the protein sequence. Furthermore, an overlay of the averaged glycan distribution recorded every 5 ns during the simulation (Figure 4.7 **C/D**) revealed a dense glycan coverage (shielding) of the protein, especially for the catalytic domain, except for the direct entrance to the active site (Figure 4.7 **C**).

4.3.1 The substrate free state

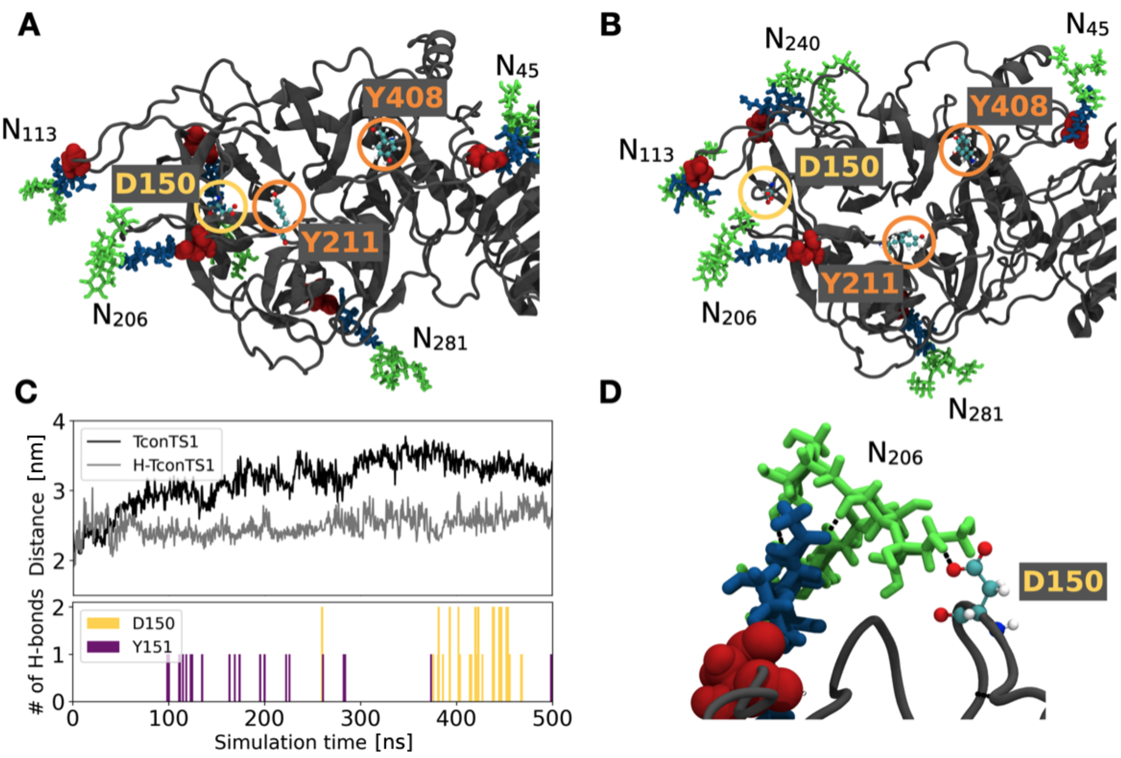


Figure 4.8: **Protein-glycan interactions drive active site rearrangements observed in MD simulations of TconTS1 without substrate.** **A** Amino acids of the catalytic site were in close proximity at the beginning of the simulation (Snapshot at 100 ns). **B** D150 moved out of the catalytic site and formed hydrogen bonds with glycan N206 until the end of the simulation (Snapshot at 350 ns). **C** Time evolution of the distance between the center of D150 and the center of R410 for TconTS1 (black) and H-TconTS1 (gray) as well as numbers of hydrogen bonds for TconTS1 between glycan N206 and D150/Y151. For H-TconTS1, no hydrogen bonds were observed. **D** Detail of the hydrogen bonds (black dashed lines) between D150 and glycan N206 at its terminal mannose branches. D150 is circled in yellow and the ligand-binding residues Y211 and Y408 are circled in orange and represented in ball-and-stick with the following color code: oxygen (red), carbon (cyan), nitrogen (blue), hydrogen (white). The underlying protein structure is represented in cartoon style in gray with asparagine residues of N-glycosylation sites showed as red spheres. Glycan color code: Man (green), GlcNAc (blue).

Interactions of N-glycans with highly conserved amino acids essential for the catalytic activity were analyzed over the 500 ns MD trajectories of TconTS1 and H-TconTS1.[251]

We especially focused on D150, E324 and Y438, known to be directly involved in catalysis, and on R126, R144, Y211, W212, R339, Y408 and R410, which are involved in substrate binding. In TconTS1, D150 was observed to shift from the interior of the active site (Figure 4.8 **A**) towards an exterior position (Figure 4.8 **B**), increasing its distance from R410, which was stationary within the active site (Figure 4.8 **C**). Interestingly, this shift seems to be stabilized by a hydrogen bond formed between D150 and the N-glycan at position N206 (Figure 4.8 **C/D**). Furthermore, detailed analysis revealed that this process was initiated by hydrogen bond formation between Y151 and the N-glycan at N206, already leading to a partial shift of D150 and making it more accessible to interact with the N-glycan (Figure 4.8 **C**). In contrast, for H-TconTS1, D150 was found to be by far less mobile (Figure 4.8 **C**).

4.3.2 The substrate bound state

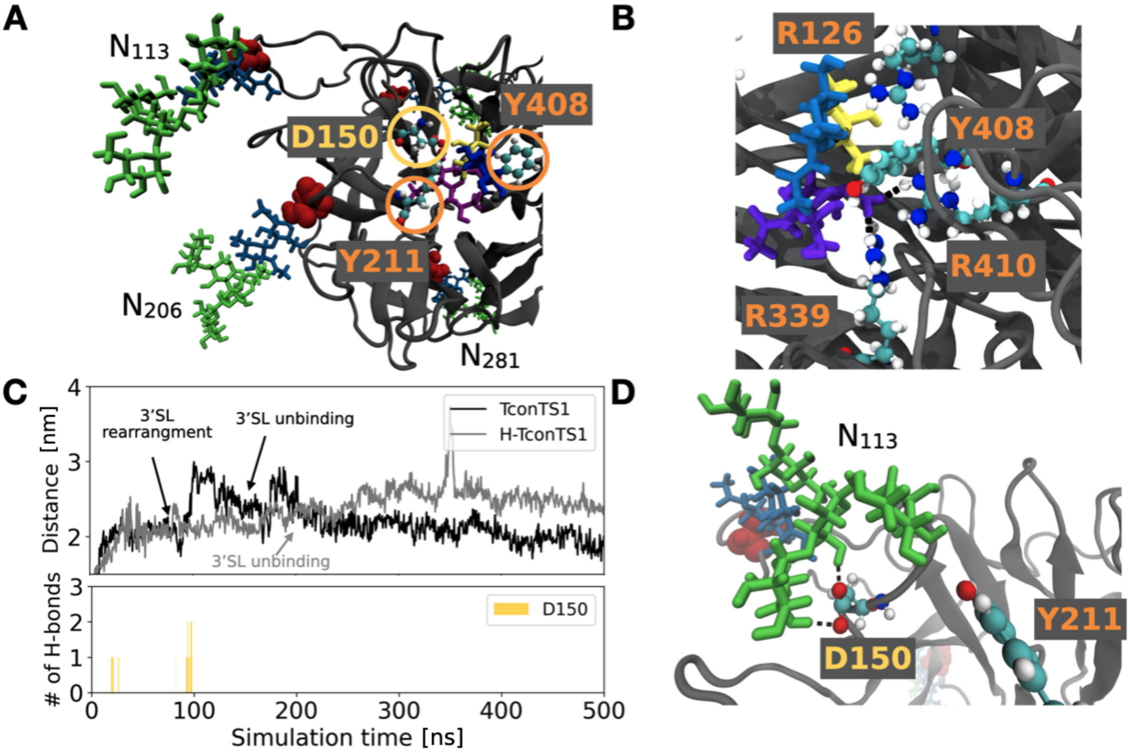


Figure 4.9: **MD simulations of TconTS1 with 3'SL bound to the active center revealed protein-glycan interactions.** **A** Starting structure. **B** Snapshot of 3'SL in the binding pocket, forming hydrogen bonds (black dashed lines) to conserved arginine residues. **C** Distance between the center of D150 and the center of R410 for TconTS1 (black) and H-TconTS1 (gray). Number of hydrogen bonds between D150 and glycan N113 for TconTS1 over the course of the simulation. **D** D150 interacts with glycan N113 via hydrogen bond formation (black dashed lines) at its mannose branches (snapshot at 97 ns). Color code of the amino acids as in Figure 4.8. Glycan color code: Man (green), GlcNAc/Glc (blue), Gal (yellow), Neu5Ac (violet).

To analyze if the observed protein-glycan interactions are also dominant in a substrate bound enzyme state, MD simulations of TconTS1 and H-TconTS1 were performed in complex with the substrate 3'SL. This substrate model was chosen since its composition is similar to the typical terminal branches of complex type N-glycans and was already used in previous enzyme assays.[113] We chose to model the Michaelis complex involving the donor substrate instead of the acceptor substrate in the covalent intermediate state (bound Sia to protein) as the crystal structure of a TcruTS/3'SL complex (PDB entry 1S0I) could perfectly serve as a template. Furthermore, there are no experimental structures of the covalent intermediate state with a bound acceptor substrate and we chose to rather model an accurate approximation rather than a faulty actual structure. 3'SL was positioned in the binding site of the homology-modeled TconTS1 in alignment with the crystal structure of the TcruTS/3'SL complex (PDB entry 1S0I) by VMD. The position of 3'SL was copied to the TconTS structure, and the ligand-protein complex was subjected to CHARMM-GUI for further processing, as described above for the setup up of the simulation cell. In the starting structure, 3'SL was bound at the acceptor substrate binding site between Y211 and Y408, and in close contact to both D150 and the well-conserved arginines R339 and R410 (Figure 4.9 **A/B**). Minimization, equilibration and production runs were performed as described above, using a timestep of 1 or 2 fs to resolve steric clashes. As already seen in the substrate-free simulation, D150 of TconTS1 formed hydrogen bonds with mannose residues of an N-glycan. However, in this case it was N-glycan at N113, which is also structurally in close proximity to the active site, and not N-glycan at N206 (Figure 4.9 **D**). This interaction was observed after 20 ns, when D150 moved slightly off the catalytic site, becoming more accessible for interactions with the N-glycan (Figure 4.9 **C**). Following a structural rearrangement of 3'SL within the binding site after around 70 ns, D150 again interacted with N-glycan N113 and was dragged out of the binding site (Figure 4.9 **C**). In striking contrast, amino acids in H-TconTS1 known to be essential for the direct catalytic activity did not experience any interactions with the residual GlcNAc residues.

4.3.3 The conformer distribution

Interactions of monosaccharides in a glycan tree with surrounding molecules like amino acids are associated to a restricted conformational movement. For instance, observed hydrogen bonds for M5 at position N113 and N206 might restrict the glycan's overall movement and thereby favor certain conformers over other. In order to test the dynamical behavior of TconTS1-bound N-glycans, the conformer distribution of glycan N206 was calculated from the substrate free simulation and compared to an M5 glycan in solution, enhanced sampled via REST-RECT as described in chapter 3 (Figure 4.10 **A**). The conformer distribution of glycan N206 is slightly shifted, favoring conformers with a *gt* configuration at the first ω angle. It can be further recognized that the second ω angle only adopts a *gg* conformation, indicating a certain restriction of the terminal $\alpha 1 \rightarrow 6$ branch. Comparing the individual PCA plots clearly underlines the restricted conformational phase-space sampling, additionally indicating that certain conformers are adopted more frequently than in the free M5 glycan (Figure 4.10 **B**). In summary, bound M5 at position N206 is adopting different glycan conformers compared to its free counterpart, as a consequence of interactions with amino acids Y151 and D150.

In order to test whether even a stiff M5 could interact with catalytic amino acids or if the three-dimensional structure of the N-glycan does not play a role in this context, a classical MD simulation of glycosylated TconTS1 was performed for 200 ns. The same settings as mentioned above were applied, additionally restraining glycan N206 to the most populated conformer of M5 in solution ($G_{-}A_{+}G_{-}A_{+}G_{+}TggG_{+}T$ $ggG_{+}A_{-}G_{+}A_{-}$). Although D150 slightly deviates from its initial position (Figure 4.11 **A**), there could be no hydrogen bonds identified between amino acids D150 or Y151 and glycan N206 (Figure 4.11 **B**). It can be concluded that in agreement with the observed conformer shift for glycan N206 when interacting with catalytic amino acids (Figure 4.10), a flexibility of the here studied N-glycan is apparently relevant for its functional mechanism.

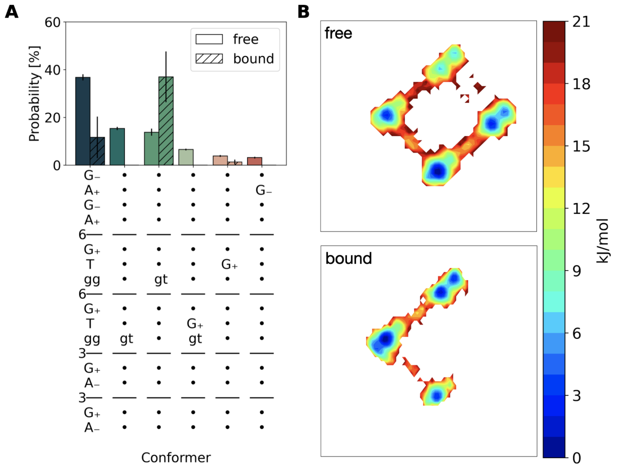


Figure 4.10: **Altered conformer distribution for M5 at N206 in TconTS1. A** Glycan conformer distribution for M5 of a free glycan in solution sampled via REST-RECT and the bound M5 at position N206 of the substrate free simulation. **B** Joint PCA of both data sets and respective free energy landscape.

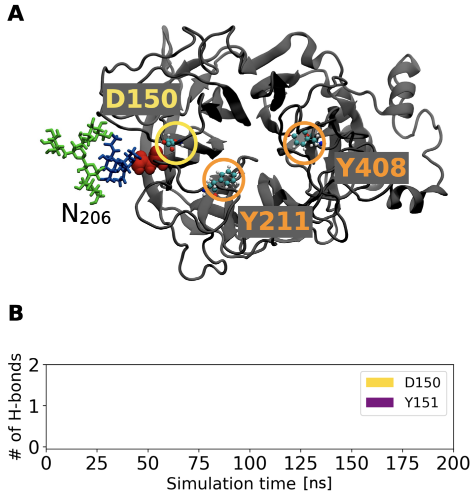


Figure 4.11: **Restricted glycan conformer does not allow hydrogen bond formation. A** Snapshot at around 150 ns of a classical MD simulation of glycosylated TconTS1 (only M5 glycan at position N206 is shown for simplicity). The glycan N206 was restrained to the most populated conformer of M5 in solution: $G_{-}A_{+}G_{-}A_{+}G_{+}TggG_{+}TggG_{+}A_{-}G_{+}A_{-}$ **B** No hydrogen bonds could be detected between D150/Y151 and the conformer-restraint glycan N206. Labeling and coloring as in Figure 4.8.

4.4 Are they conserved?

Diverse high-mannose type N-glycan structures could be identified for almost all nine N-glycosylation sites of TconTS1, positively impacting the substrate affinity of the acceptor substrate lactose in the glycosylated state. Our atomistic simulations of TconTS1 indicate that glycans attached to site N206 (and N113) could play a potential role in positioning the proton donor D150, influencing its ability to act in the enzymatic transfer of Sia, a crucial contribution to substrate conversion.[116,252] Recurring hydrogen-bond interactions of D150 with N-glycans at positions N206 or N113 are the driving force of the conformational change in the catalytic site which could not be detected for H-TconTS1. This fine tuning of critical amino acid side chains' arrangement might facilitate the initial binding of substrates and lead to a higher substrate-binding affinity in line with our kinetic data. The study of the exact mechanism including the whole transfer mechanism with donor and acceptor substrate exceeds the purpose of this study and would require the application of a hybrid quantum mechanics/molecular mechanics (QM/MM) approach. Nevertheless, our observations of hydrogen bond formation between N-glycan N206 and protein residues stand in direct relation to the heterogeneous N-glycosylation structures observed for N206 (M5-8, FM5). In fact, the occurance of not trimmed, large high-mannose type N-glycans such as M8 could be explained by the lack of accessibility of the glycan for glycosidases due to continuous protein-glycan interations. It was previously shown by biochemical *in vitro* assays and computational studies that such protein-glycan interactions decrease glycan accessibility, which might interfere with glycan trimming.[274] This hypothesis is supported by our observation that EndoH was not able to remove all N-glycans from TconTS1 even after 16 h of incubation and therefore also EndoH-mediated removal of N-glycans could be hindered by these interactions. Despite this site-specific analysis, it was interesting to see that especially N-glycans of the catalytic domain were observed to form a highly dynamical 'shield' enclosing the enzyme, while leaving the entrance to the catalytic center open for substrate binding (see Figure 4.7).

Due to the nature of MD simulations, being limited in the number of included atoms, only one protein was simulated at a time, either fully N-glycosylated at all positions that were experimentally identified in this study, or completely deglycosylated with residual GlcNAcs, simulating EndoH treatment. Since experiments such as enzyme assays always represent a cumulative average over all possible protein structures present in the sample, differences in their glycosylation pattern are very likely, but difficult to examine with our qualitative approach. This heterogeneity is not accounted for in our MD simulations and therefore only possible mechanisms for the extreme (fully glycosylated vs. deglycosylated) cases can be derived. The averaging effect of experimental methods especially applies to the hypoglycosylated samples, which were shown to be only partially deglycosylated, mainly at position N206. Therefore, the kinetic effects may even be more pronounced if N-glycans at position N206 would be completely absent. Site-directed mutagenesis of these sites could provide further insights, but it remains an open question whether these mutants would be able to successfully fold into an active enzyme structure.

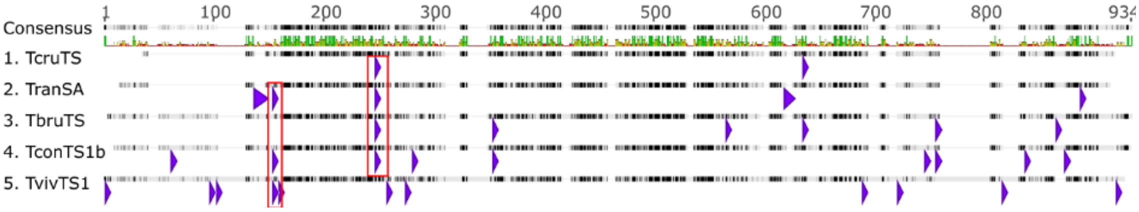
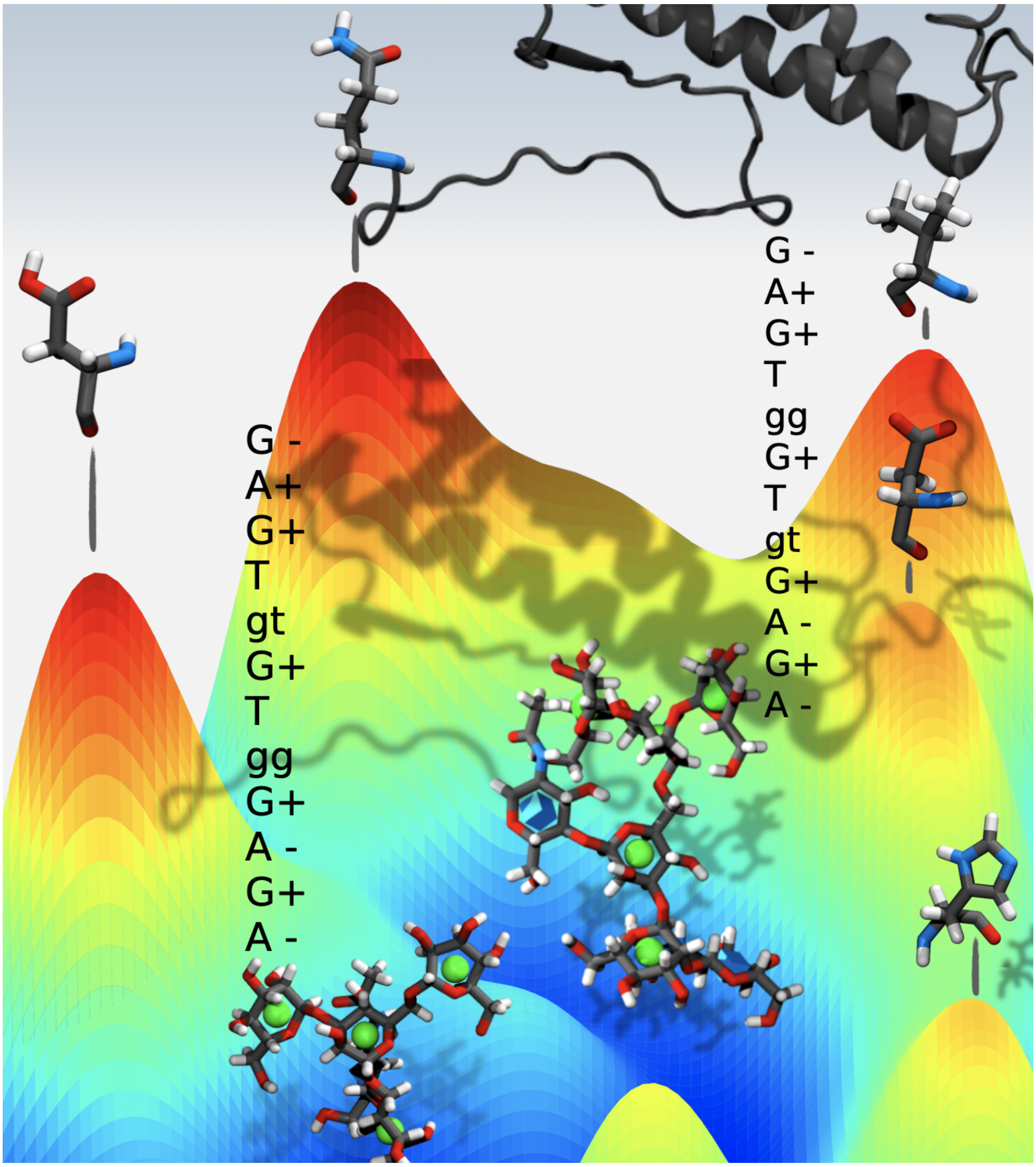


Figure 4.12: **Interspecies conservation of N-glycosyation sites.** The sequence alignment of TS from *T. cruzi* (TcruTS, GenBank ID AAA66352), *T. brucei brucei* (TbruTS, GenBank ID AAG32055), *T. congolense* (TconTS1b, GenBank ID HE583284), *T. vivax* (TvivTS, GenBank ID CCD21087) and the closely related sialidase from *T. rangeli* (TranSA, GenBank ID AAC95493) was generated with the ClustalW Alignment tool of the software Geneious Pro 5.5.9, employing the BLOSUM matrix with a gap opening cost of 10 and gap penalty cost of 0.1. Conserved glycosylation sites are indicated (red box). *Analysis and figure creation were performed by Jana Rosenau.*

Thinking outside the box and comparing N-glycosylation sites of TconTS1 to other trypanosomal species revealed conserved glycosylation sites among TS and other sialidases despite their variations in amino acid sequences. In particular, N206 is conserved in all TS variants of *T. congolense* as already described by Waespy et al.[257] but also in enzymes from *T. cruzi, T. brucei, T. vivax* and sialidase from *T. rangeli*, as revealed by our amino acid sequence alignment (Figure 4.12). Additionally, N113 is also conserved in TconTS, TvivTS and TranSA. Along this line, we propose that there might be a common mechanism of TS activity mediated by N-glycan interactions with amino acids of the active site, in particular with D150 or its equivalents in other TS. The short time scale and lack of N-glycans in previously performed MD simulations probably prevented the observation of the here-observed D150 structural shift.[275–278] Intramolecular glycan-protein interactions are a naturally observed event and have been suggested to regulate the conformation of proteins and their ability to bind to substrates such as collagen.[279,280] This assumption can be confirmed by our study for TS enzymes and highlights the importance of the protein expression system to achieve a biologically-correct post-translational modification pattern, which might also be crucial for accurate protein folding. Furthermore, we underline the importance of including N-glycans in MD simulations, as they can act as key residues conferring function to the protein, as was also shown for the spike protein of SARS-Cov-2.[100]

5 | In the middle: Glycans as substrates



The artistic phase space of an enzymatically bound N-glycan. The high-mannose type N-glycan M5G0 is restricted in its conformational phase space through the surrounding amino acids in the catalytic site. The presence of the protein is shading and altering the free energy landscape. Each conformer is flagged by corresponding conformer labels.

The processing of post-translationally added N-glycans is performed by various enzymes in the ER and Golgi in eukaryotic cells. An important step is the transition from high-mannose type to complex N-glycans via trimming of Man residues by various mannosidases. Several different cancer types like colon, skin and breast cancers are characterized by an altered high amount of complex N-glycans on their cell surfaces, correlated with metastasis growth and disease progression.[63,281,282] Inhibition of certain key enzymes in the glycosylation pathway that bridge the transition from high-mannose type to complex N-glycans, like α-mannosidase II (GMII), reduced the formation of complex N-glycans and could be associated with reduced tumor growth and metastasis.[283] Unfortunately, treatment with the inhibitor swainsonine or its derivatives showed side effects, which could be associated with the simultaneous inhibition of lysosomal α-mannosidase.[284] Despite extensive efforts to develop potent selective inhibitors, no breakthrough has yet been achieved that would have enabled clinical application.[285]

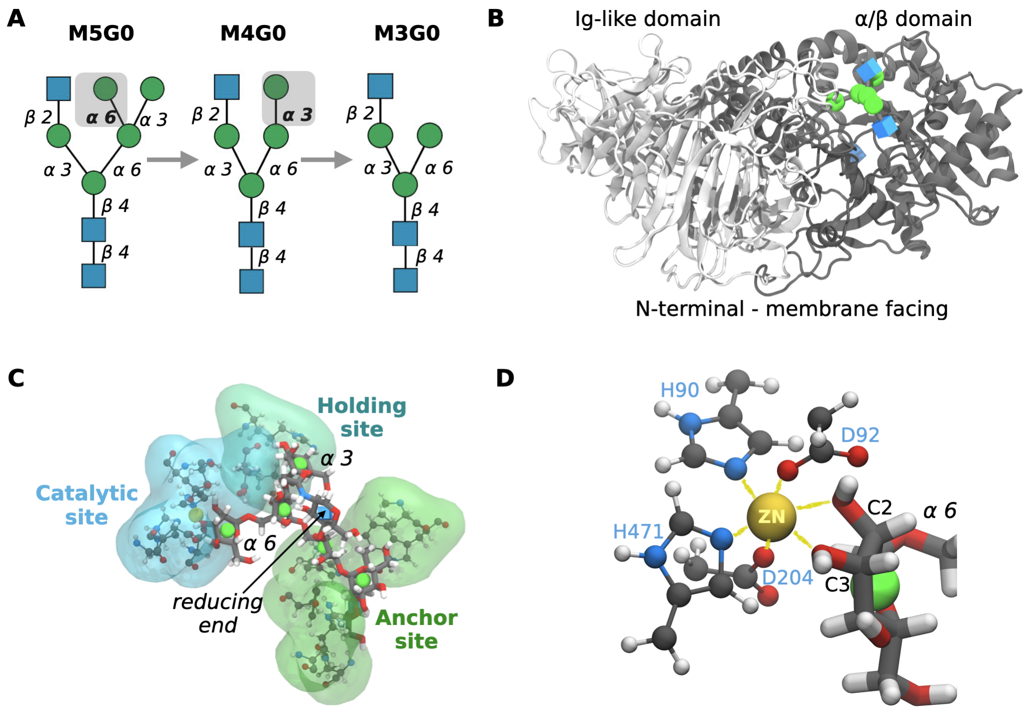


Figure 5.1: **The transmembrane Golgi α-mannosidase II. A** GMII catalyzes the two-step hydrolysis of M5G0 to M3G0, first cleaving the α 1→6 linkage yielding M4G0 and subsequently the terminal α 1→3. **B** Structure of GMII in complex with its substrate M5G0 (PDB entry: 3CZN). The globular structure consists of an Ig-like domain, harboring β-sheets (white), and an α/β domain with the catalytic site (gray). **C** Snapshot of the binding site with M5G0 bound to the anchor (Q64, Y267, H273, P298, W299, R410), holding (R343, D340) and catalytic (H90, D92, D204, D341, H471, D472) sites. **D** Zoom into the catalytic site, where a Zn ion is sixfold-coordinated involving the amino acids H90, D92, D204 and H471 as well as the O2 and O3 atoms of the terminal Man residue. Atomistic glycan structures are represented in Licorice representation and amino acids in CPK. Atoms are colored by their respective element with carbon in gray, oxygen in red, nitrogen in blue, hydrogen in white and zinc in yellow. The glycan color code follows the SNFG regulations.

The mechanisms of distortion of a monosaccharide pucker from its free, low-energy chair conformation to a strained, distorted one in a GH transition state is important to understand in order to reveal the whole catalytic mechanism and gather information for the development of putative inhibitors.[286] The distortion is often necessary in order to achieve are more favorable conformation of the linkage that is to be cleaved, as well as create an oxocarbenium ion-like character. There is probably no universal explanation for this fundamental question, as GHs are characterized by different catalytic mechanisms (inverting/retaining) and harbor diverse ion types in their binding sites. However, we aimed at addressing this issue by employing GMII as a model system due to its well studied catalytic reaction and the available crystal structure of the almost native protein in complex with its unmodified substrate (PDB entry: 3CZN). We especially focused on a putative structural role of the flexible glycan substrate itself, as well as the chemical influence of the Zn ion, which was previously suggested to induce pucker distortion.[286,287]

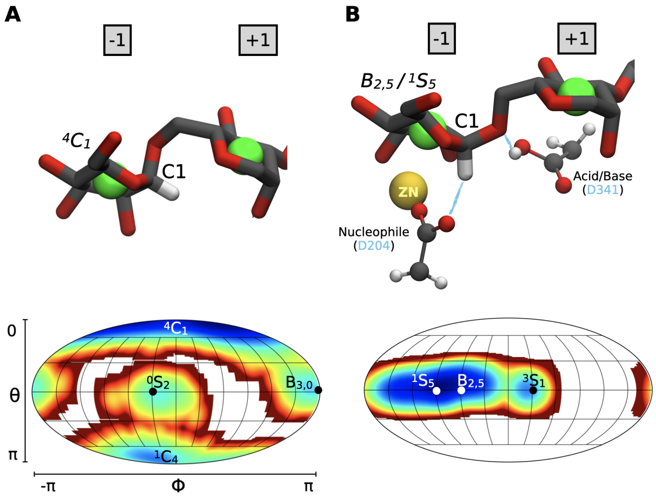


Figure 5.2: **GMII follows a retaining mechanism to achieve the transition state.** **A** Atomistic representation of the α 1-6 branch, when free in solution. The terminal Man residue adopts a $^4C_1$ pucker conformation, corresponding to the global minimum in the ϕ/θ puckering plot. **B** When the α 1-6 branch is binding to the catalytic site of GMII, the terminal Man residue adopts a $B_{2,5}/^1S_5$ pucker conformation in its transition state, forming a deep energy minimum in the puckering plot. Atomistic glycan structures are represented in Licorice representation and amino acids in CPK. Atoms are colored by their respective element with carbon in gray, oxygen in red, hydrogen in white and zinc in yellow. The glycan color code follows the SNFG regulations with Man in green.

GMII belongs to the GH family 38, catalyzing the final removal of two Man residues prior to the incorporation of diverse monosaccharides in downstream steps.[288] This integral membrane protein is situated in the *medial* Golgi, acting on both terminal α1→6 linked and α1→3 linked mannoses, converting glycan M5G0 (GlcNAcMan5GlcNAc2) to M3G0 (GlcNAcMan3GlcNAc2) within the same catalytic binding site (Figure 5.1 **A**).[289] The amino acid sequence of GMII is conserved among many eukaryotes. The first crystal structure was resolved in 2001 for *Drosophila melanogaster*, because the human homologue is difficult to purify in large amounts.[63,289] It consists of two larger domains, an N-terminal α/β domain harboring the glycan binding site and the C-terminal Ig-like domain, whose function is yet to be discovered (Figure 5.1 **B**).[63] The structurally unresolved N-terminus is predicted to serve as a membrane anchor, positioning the protein with its catalytic site facing the Golgi lumen.

The binding site of GMII consists of three distinct sugar binding regions: the anchor, holding and catalytic sites (Figure 5.1 **C**).[289] At the anchor site, the GlcNAc residue of the α1→3 linked branch is stably bound by several conserved amino acids (Y267, W299, P298, H273), correctly orienting the flexible substrate for hydrolysis. As usual, CAZymes are very substrate specific and it is therefore not surprising that the catalytic reaction depends on the presences of the $\beta1 \rightarrow 2$ linked GlcNAc residue, as the activity was found to be 80-fold reduced for the structurally similar glycan M5. The holding site (R343, D340) is able to accommodate the terminal $\alpha1 \rightarrow 3$ linked branch, however not the longer terminal $\alpha1 \rightarrow 6$ branch, suggesting an initial cleavage of the $\alpha1 \rightarrow 6$ branch prior to the $\alpha1 \rightarrow 3$ one.[289] The catalytic site is characterized by a stably bound Zn ion, sixfold coordinated by oxygens 2 and 3 of the terminal Man residue and amino acids H90, D92, D204 and H471 (Figure 5.1 **D**).

It is suggested that the mode of action is similar for the cleavage of the two different linkages, although we focus only on the $1 \rightarrow 6$ linked mannose here.[289] Glycan linkage hydrolysis follows a retaining mechanism in GMII, where the substrate is distorted from the low energy $^4C_1$ pucker over to the transition state conformer $^OS_2/B_{2,5}$, to a $^1S_5$ pucker conformation in the covalent glycosyl–enzyme intermediate state (Figure 5.2). The nucleophile of the reaction is represented by D204 and the acid/base by D341 (Figure 5.2 **B**). The detailed reaction mechanism was already studied by QM/MM calculations, showing that the glycosidic bond dissociates prior to the formation of a covalent bond between the substrate and D204, where the transition state has a clear oxocarbenium ion-like character.[290] We focus on the glycan conformation at the very beginning of the reaction, being interested in how the distortion of the terminal Man residue is induced from a $^4C_1$ conformation free in solution (Figure 5.2 **A**), to a $B_{2,5}/^1S_5$ pucker before reaching the transition state (Figure 5.2 **B**).

5.1 The conformer hypothesis

N-glycans are very flexible molecules, able not only to rotate around their glycosidic bonds but also to adopt different pucker conformations, even when freely simulated in solution (see chapter 3).[77] It stands to reason that a mutual dependence of pucker conformations and certain glycan conformers can be assumed. [134] As a first step, the prevailing glycan conformers in the binding site of GMII need to be determined, as the accessible conformers might differ from those free in solution. If the conformer distribution deviates from the one of a free glycan, it is to be investigated whether there is a direct effect on the puckering propensity of the terminal Man residue at position -1. A sufficient phase-space sampling of the glycan substrate under the additional difficulty posed by the interactions with surrounding amino acids was tackled by the application of the REST-RECT methodology.

An atomistic model of GMII was built up from the crystal structure with PDB entry 3CZN (from *Drosophila melanogaster*), having the native ligand structure of M5G0 resolved, except for one GlcNAc residue at the non-reducing end. We expect this missing residue to have no significant impact on the conformer distribution and therefore did not add it *a priori*, to stay as close as possible to the experimental conditions. The glycan is nevertheless termed M5G0 onwards. Only the α/β domain (amino acids 31-510) harboring the ligand binding site was included in the model (Figure 5.1 **A**), keeping the system as small as possible to save computational resources. An unfolding of the shortened C-terminus could not be detected in subsequent simulations and was therefore not necessary to constraint. Via CHARMM-GUI, the mutation of the nucleophile D204A was reversed to a deprotonated aspartic acid and D341 converted to its protonated form, mimicking the optimal conditions for the initialization of a retaining hydrolysis reaction. Furthermore, missing hydrogens were added and the system solvated in a 12.5 x 12.5 x 12.5 Å^3 water box, containing only the Zn^{2+} as an ion in the binding site. All simulations were performed with GROMACS 2018.4[206], patched with PLUMED 2.6[207]. An energy minimization, restraining the positions of the heavy atoms of the protein, was performed using the steepest-descent algorithm with a tolerance of 1000 kJ mol^{-1} nm^{-1}. An NVT equilibration was subsequently performed for 1 ns with the same restraints as in the minimization step. Following production runs were always performed under the NPT ensemble. The leap-frog algorithm was used as an integrator with a 2 fs time step, and the LINCS algorithm[210] was employed to constrain bonds connected to hydrogen atoms. Temperature control was realized via velocity rescaling[211], using a time constant of 0.1 ps, setting a reference temperature of 310.15 K. The pressure was set to 1 bar with a compressibility of 4.5 x 10^{-5} bar^{-1}, and kept constant via the Parrinello-Rahman barostat with a time constant of 5 ps. The Verlet list scheme[212] was employed with a neighbor list updated every 80 steps. The calculation of electrostatic interactions was done with the PME[213] method using a cut-off distance of 1.2 nm for the real space contribution. A standard MD simulation of the whole system, lasting for 1 ns, was performed with both, the CHARMM36 and the GLYCAM06j force fields, in order to assess their ability to convert the pucker at position -1 in the catalytic site from a chair to a distorted boat or skew boat conformation. Spontaneous happening of this transition is necessary, as the crystal structure displays only a chair conformation for the terminal Man residue due to the D204A mutation. Constraints of the distance between the catalytic proton of D341 and the O6 of the α 1→6 linkage to 1.4 Å, as well as of the distance between the oxygens of D92 and HO2 of the terminal Man residue to 1.6 Å, should induce a distorted pucker at position -1. The first constraint is important for the catalytic reaction and the second mimics an experimentally observed hydrogen bond that does not form under standard MD simulation conditions.[290] It turned out that only the GLYCAM06j force field was able to predict the transition from a chair to a boat conformation as already demonstrated by Petersen et al. 2009[290]. The CHARMM36 force field let the terminal Man to persist in a chair conformation. Therefore, the ff19SB force field[153] was used for the protein and ion atoms, GLYCAM06j for the glycan atoms, and TIP3P as the water model.[208]

For the sufficient phase-space exploration of ligand M5G0 bound in the catalytic site, a REST-RECT simulation of the GMII protein-glycan complex was performed. The simulation setup differed slightly from that of chapter 3, as sampling in this larger molecular system introduced convergence issues (Figure 5.4 **A**). First of all, 16 replicas were used with λ_α values equal to 1, 0.98, 0.95, 0.92, 0.90, 0.87, 0.84, 0.81, 0.78, 0.75, 0.72, 0.69, 0.65, 0.62, 0.58, 0.55, spanning an effective temperature range from 310.15 K to 570 K. A geometric progression of λ_α values turned out to be less efficient regarding replica exchanges. The solute region did not only include the glycan atoms, but also the following amino acids of the binding site: D106, R228, Y267, Y269, D270, H273, R289, D340, R343, R410, D412, D472, T477 (Figure 5.3 **A**). These were identified from REST-RECT test runs, in which they formed hydrogen bonds with monosaccharide atoms, preventing an efficient exploration of glycan conformers, and therefore hindering convergence. It is no coincidence that several of these solute amino acids are also part of the anchor, holding and catalytic sites of GMII, as it is their natural function to constrain the glycan in the binding pocket. Water and ions were always kept at the ground temperature. The RECT part biased all 14 torsion angles simultaneously via one-dimensional bias potentials in each replica α. The αth replica was biased using a bias factor γ_α, with values equal to 1, 1.13, 1.27, 1.43, 1.61, 1.82, 2.05, 2.31, 2.60, 2.93, 3.30, 3.72, 4.19, 4.73, 5.32, 6 over the replica ladder. Several distance constraints were used for different purposes (Figure 5.3 **B**). On the one hand, the glycan as well as the Zn ion had to remain bound to the binding site even at higher temperatures. Therefore, distances between Zn and nitrogen atoms of H90 and H471, as well as Zn and O2 and O3 of the Man residue were kept fixed at 2 Å. On the other hand, the terminal Man residue should be retained in its distorted pucker conformation that is associated to the reactant state prior to the transition, in order to sample the conformational phase space of M5G0 under catalytic conditions. Hence, the D92-glycan and D341-glycan distance constraints mentioned above were used in addition to restraining the pucker coordinate θ to 1.5, only allowing for the sampling of boat and skew boat pucker conformations.

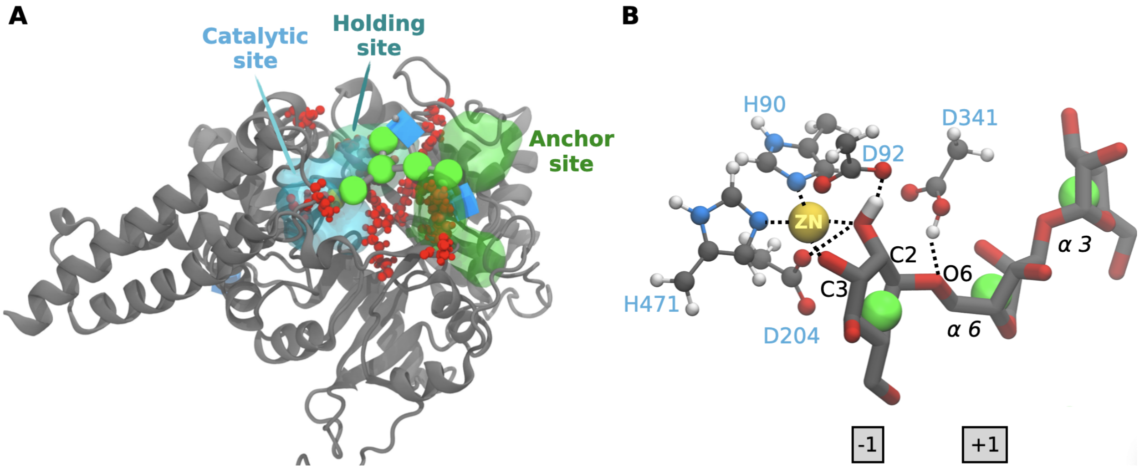


Figure 5.3: **REST-RECT setup of GMII in complex with M5G0.** **A** The α/β domain with residues 31-510, including the binding site and the ligand, was simulated. The solute region included all glycan atoms as well as amino acids of the binding site (highlighted in CPK style in red): D106, R228, Y267, Y269, D270, H273, R289, D340, R343, R410, D412, D472, T477. **B** Harmonic restraints (black dotted lines) were introduced to keep the Zn ion and the terminal Man residue positioned in the binding site. Atomistic glycan structures are shown in Licorice representation and amino acids in CPK. Atoms are colored by their respective element with carbon in gray, oxygen in red, nitrogen in blue, hydrogen in white and zinc in yellow. The glycan color code follows the SNFG regulations with Man in green and GlcNAc in blue.

The REST-RECT simulation was run for 500 ns per replica, yielding a total of 8 µs of cumulative sampling time. The plausibility of the chosen replica exchange parameters was verified by examining the exchanges of replicas over the replica ladder. Including only 12 replicas, the standard REST-RECT settings from chapter 3 resulted in insufficient replica overlaps and no exchanges across the whole replica ladder (Figure 5.4 **A**). However, a colorful mixing of replica indices could be observed after including 16 replicas with an exchange acceptance above 50 % for all replicas. Additionally, the replicas underwent several round trips with round trip times of 11 ns for the lowest and highest replica (Figure 5.4 **B**).

The evidently sufficient replica exchange settings resulted in a converged conformer distribution for M5G0 bound in the catalytic site of GMII (Figure 5.5 **A**). The cumulative average of the three most populated conformers was calculated from the unbiased ground replica, becoming flat after 350 ns of simulation time. The first 200 ns were discarded, as the metadynamics potentials first had to take effect, requiring a longer lead time to push the glycan away from its initial conformation. The subsequently calculated conformer histogram (Figure 5.5 **B**) displays that conformers differ mostly in the second ϕ and both ω angles, as it is also the case for free glycans in solution.

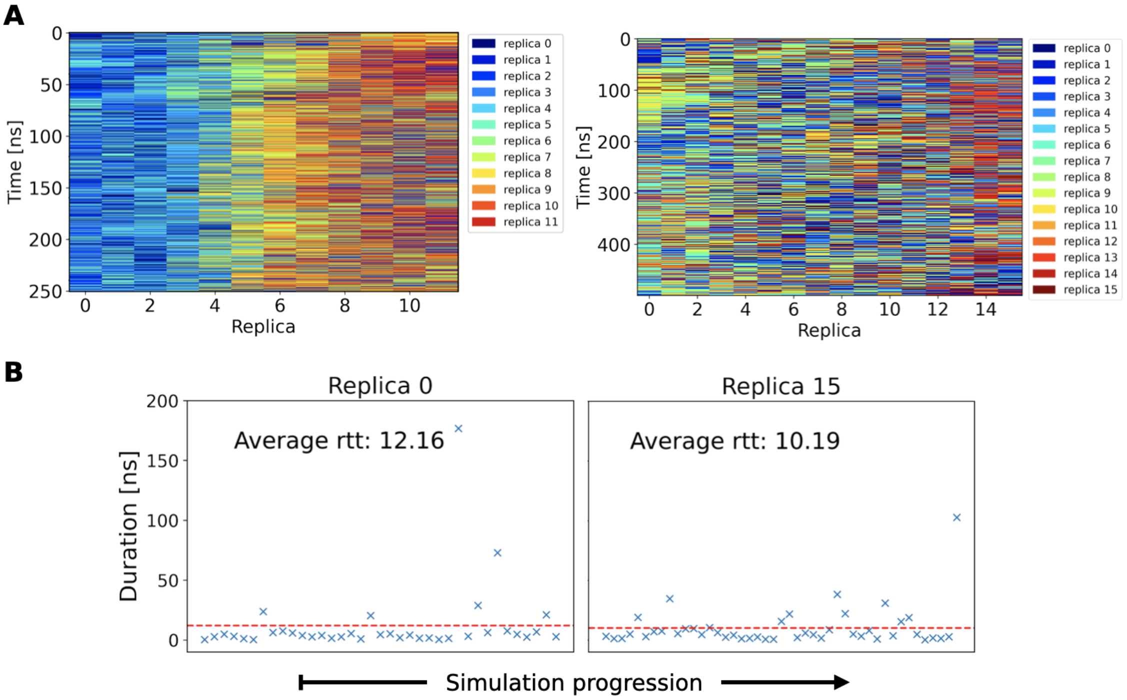


Figure 5.4: **Replica details of REST-RECT simulating GMII in complex with M5G0. A** Replica exchanges visualized by plotting the replica index over time along the replica ladder. 500 exchanges over the course of the simulations are visualized out of 18000 exchange attempts. On the left, a REST-RECT simulation using 12 replicas and standard settings from chapter 3 compared to adapted settings using 16 replicas and altered bias scaling. **B** Duration of round trip times (rtt) for replica 0 and 15 (ground and highest replica), along the progression of the simulation. An average of around 10 ns per round trip (red dotted line) allows for 50 round trips per replica during one REST-RECT simulation.

The temperature and bias factor ranges in this protein-glycan simulation were shorter compared to similar simulations of free glycans. This is because major convergence issues were faced when applying the standard scheme of chapter 3, making it necessary to reduce the biasing to a minimum, while simultaneously ensuring complete sampling of all degrees of freedom. Test calculations revealed that a maximum temperature of 550 K was still sufficient to sample different pucker conformations in conjunction with biasing all torsion angles. Additionally, also the bias factor could be reduced to a maximum amount of six, still allowing for a complete sampling of all torsion angles in the highest replica (Figure 5.5 **C**). The distributions of torsion angle values in the lowest replicas look different from the ones in replica 15, as a much smaller bias potential is applied. However, frequent replica exchanges ensure the mixing of relevant conformers from higher to lower replicas.

The obtained conformer distribution of bound M5G0 was first compared to free M5G0 in solution, validating that a conformer shift is indeed induced by the altered chemical surrounding of the protein pocket (Figure 5.6 **A**). Free M5G0 was sampled via REST-RECT as described in chapter 3 using the GLYCAM06j force field and TIP3P water. The distributions are not majorly different regarding the overall sampled conformers, but populate them to a different extent. For instance, the highest populated conformer of free M5G0 (25 %) only contributes to 6 % for bound M5G0. In contrast, the most populated conformer of bound M5G0 (32 %) is the fifth conformer in the free M5G0 simulation (6 %).

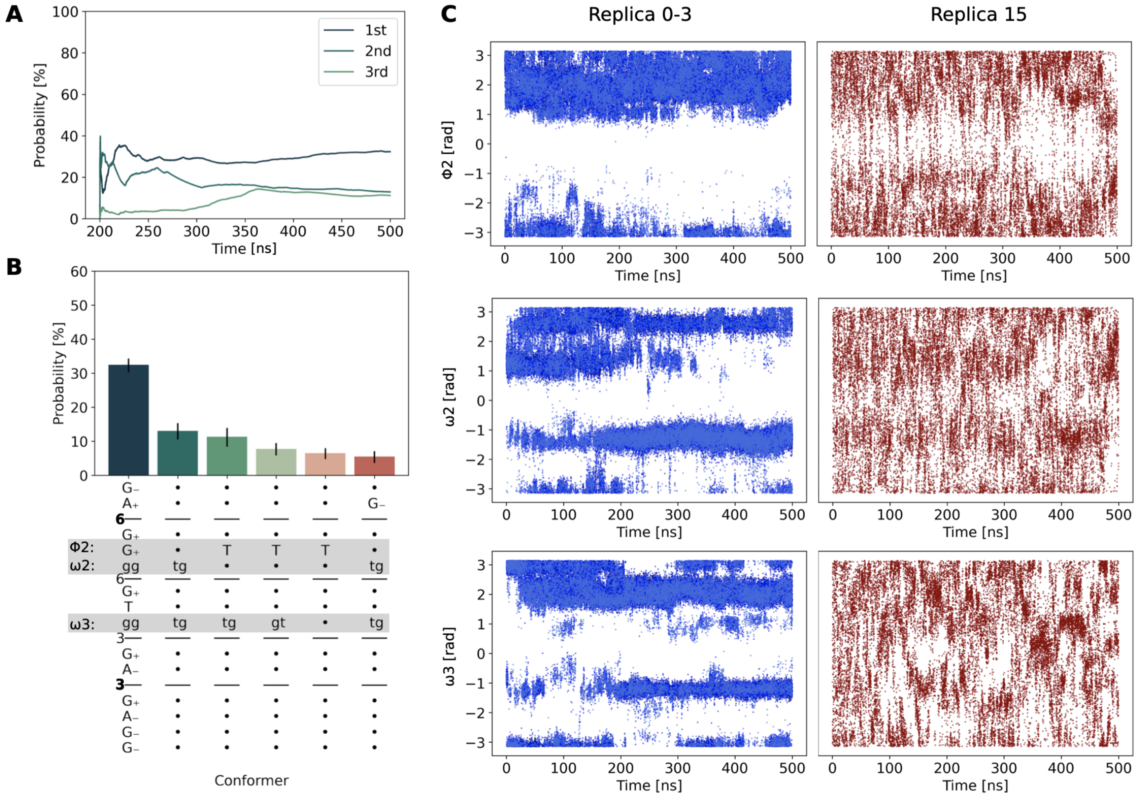


Figure 5.5: **Convergence of the M5G0 conformer distribution. A** Cumulative average of the three most populated conformer clusters according to panel **B**. **B** Conformer distribution for M5G0. The conformer string is given on the x-axis. The gray boxes highlight key conformational differences between conformers. For **A** and **B**, data points recorded up to 200 ns of simulation were discarded, as the REST-RECT sampling required lead time to take action. 75000 datapoints were analyzed. **C** Torsion angle fluctuations of ϕ 2, ω 2 and ω 3 over the simulation time for replica 0-3 and 15, respectively.

The spatial distribution of both data sets is more easily visualized using a joint PCA map, e.g. plotting the free energy surfaces independently (Figure 5.6 **B**). Free M5G0 covers a larger phase space compared to its bound counter part, probably due to a restrained movement resulting from the surrounding amino acids. Moreover, the accessible phase-space region of the bound systems is different in the sense that it includes conformers that are not sampled in the free system (Figure 5.6). Generally, the three most populated conformers of the two systems are not overlapping (Figure 5.6 **C**). When overlaying the initial conformers sampled in the standard MD simulation of 1 ns with induced distorted pucker conformation for terminal Man with data points from free and bound M5G0 molecules, it is surprising to observe that they lie in a region that is only sampled by free M5G0 (Figure 5.6 **D**). It appears that the crystallized conformer is majorly influenced by the D204A mutation, inducing a conformational change in the first ω angle (gt instead of gg) that is no longer sampled in subsequent extended simulations in which the mutation was reversed.

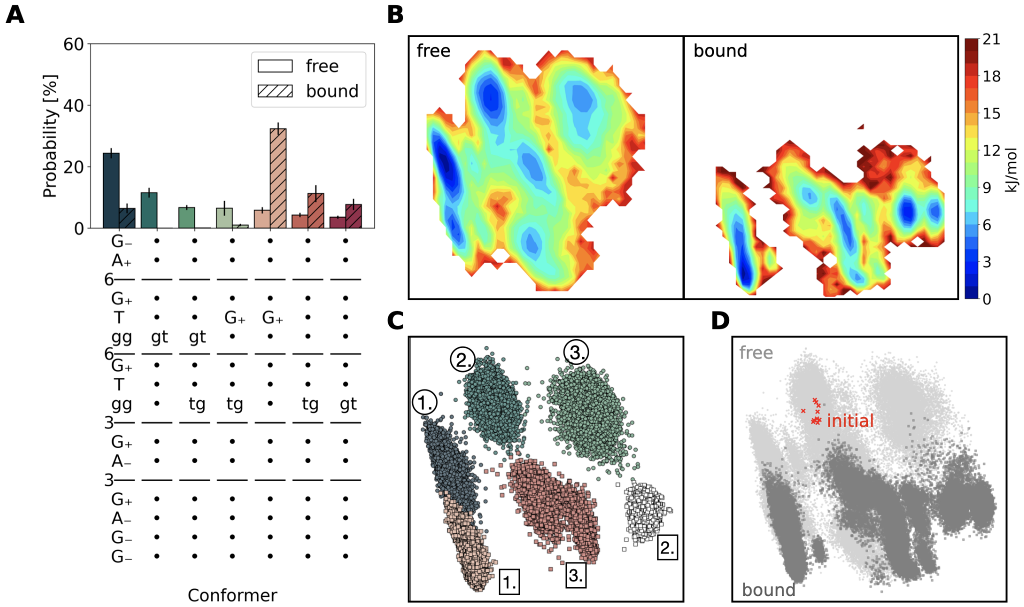


Figure 5.6: **Comparing conformer distributions of free and bound M5G0. A** Conformer distribution from Figure 5.5 plotted together with the distribution obtained from simulating M5G0 free in solution via REST-RECT. **B** Joint PCA, visualizing the conformational free energy landscape for free and bound M5G0. **C** Joint PCA of free and bound M5G0, highlighting the three most populated conformers, respectively. Coloring is in accordance to **A**, with circles for free and squares for bound M5G0. The white coloring for the second conformer of bound M5G0 is due to the absence of this conformer in **A**, as it is not adopted by free M5G0. **D** Joint PCA of the complete distribution of free and bound M5G0, where the initial three-dimensional glycan structure from a classical MD simulation of 1 ns is displayed as red crosses. The terminal Man residue in the catalytic site was restrained to a boat conformation, counteracting the induced chair conformation resulting from the mutation D204A in the experimentally determined structure.

After the confirmation of a shifted conformer distribution of bound M5G0 in its reactant state, it is yet to be answered whether this shift is able to affect the relative pucker propensity in individual monosaccharide units. We addressed this hypothesis from two different perspectives. In a first approach, a free M5G0 glycan in solution was enhanced sampled via REST-RECT, with the pucker coordinate θ harmonically restrained to 1.5 with a 1500 kJ mol^{-1} force constant, in order to enforce a boat/skew boat conformation. The resulting conformer population shows only slight deviations from the one of an unrestrained free M5G0 (Figure 5.7 **A**). Additionally, it does not resemble the histogram obtained from GMII-bound M5G0 (Figure 5.6 **A**), except for the reduction of probability of the fourth conformer.

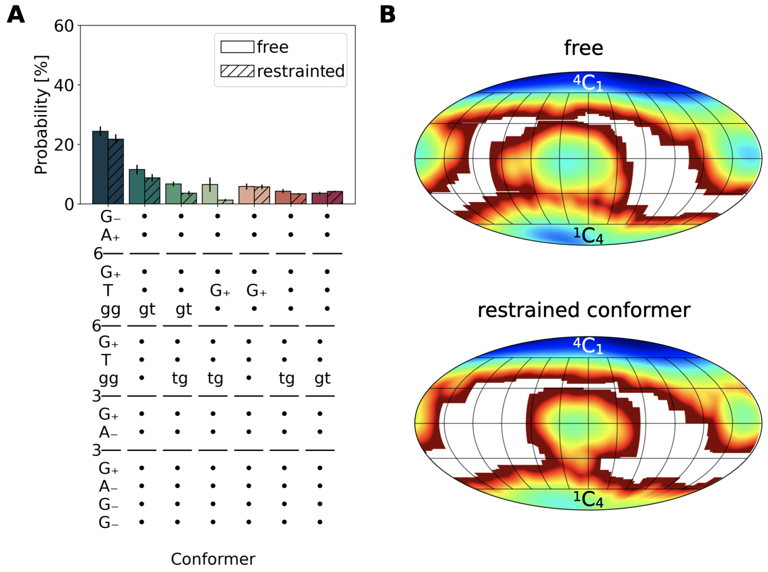


Figure 5.7: **Mutual dependence of torsion and pucker conformations in GMII. A** Conformer distribution of free M5G0 compared to its pucker restraint counterpart, fixing a boat conformation for the terminal α1→6 linked Man. **B** Puckering free energy profiles of the terminal α1→6 linked Man residue from a free M5G0 glycan and from a restraint M5G0 glycan, where the conformer was fixed to the most populated state in the bound GMII simulation ($G_-A_+G_+TggG_+TggG_+A_-G_+A_-G_-G_-$).

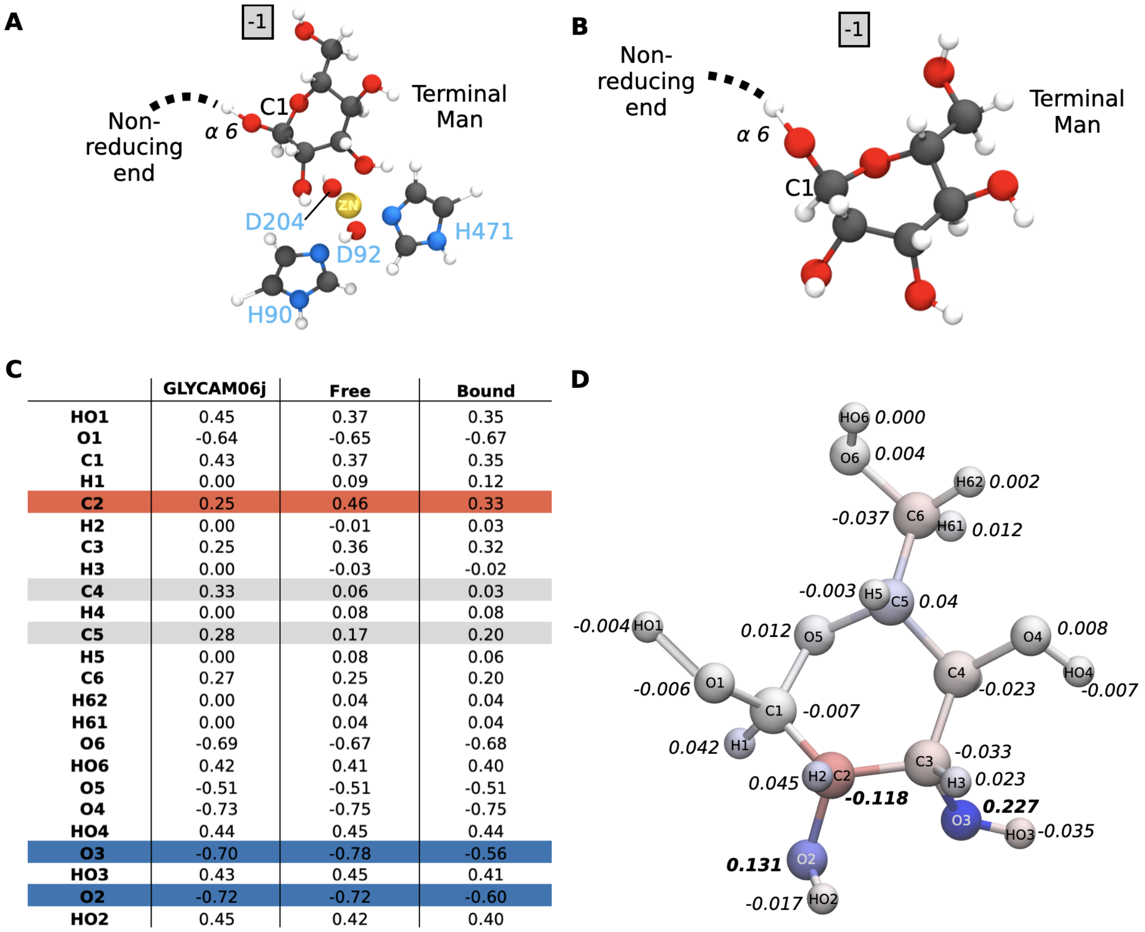
In the second approach, a free M5G0 in solution was restrained to the most populated conformer ($G_-A_+G_+TggG_+TggG_+A_-G_+A_-G_-G_-$) of bound M5G0 (Figure 5.6 **A**). The harmonic restraints on each torsion angle were applied using force constants of 5000 kJ mol^{-1} and the system subjected to a REST2 simulation with a temperature range from 310.15 to 800 K in order to enhance sample the pucker conformers of each monosaccharide unit in the glycan tree. Comparison to the unrestrained REST-RECT simulation of free M5G0 revealed that the puckering free energy of the terminal α1→6 linked Man was not influenced by the restraint (Figure 5.7 **B**). It can therefore be concluded that there exists no mutual dependence between glycan conformation and ring distortion states, at least for terminal monosaccharide residues. Whether the same holds true for central residues in longer polysaccharide chains should be the subject of future studies.

5.2 The charge hypothesis

Given that the conformer hypothesis does not hold, at least for GMII, it is now time to take a closer look at the Zn ion and its influence on the pucker distortion of M5G0. Due to its full d-orbital, Zn^{2+} is an electronically quite stable ion that tends to withdraw electrons from its coordinating atoms.[286] For GH enzymes, it is a rather unusual prosthetic group, with calcium being the the most prominent divalent cation. Previous QM/MM simulations assign a catalytical role to Zn, assisting the lengthening of the glycan's HO2-O2 bond in the oxocarbenium ion-like transition state, counteracting the electron deficiency around the C1 anomeric center.[290] This consideration only takes effect later in the reaction mechanism, once the substrate is already in its distorted pucker shape and does not address the reason for the conformational change. However, there were also experiment-based assumptions regarding the Zn ion playing a role in the pucker distortion, driving the reaction towards catalysis. The presumption is based on a crystallized structure of GMII in complex with noeuromycin adopting a $^{1,4}B$ conformation, although no further mechanism or casual reason was provided.[286] The crystal structure of mutant D204A of GMII (PDB entry: 3CZN) might give another hint in the direction of the Zn ion playing an active role in influencing the pucker conformation.[289] The terminal Man at position -1 adopts a low-energy $^4C_1$ chair conformation, potentially a product of the altered coordination of the Zn ion due to a missing ligation partner. The Zn ion debate is especially interesting given the fact that the hydrolysis reaction of several other α-mannosidase enzymes, comprising GH families 76, 92, 47 and 125, can also take place without the presence of a divalent ion.[134,291–293] In the following, we focused on the hypothesis of whether the presence of a Zn ion could lead to a charge redistribution around the glycan atoms, promoting a conformational change of the puckering.

This hypothesis was addressed by quantum chemical calculations, recalculating the partial charges of a free Man residue versus a GMII-bound one (Figure 5.8 **A,B**). The bound system ($B_{2,5}$ conformation) mimics a reduced catalytic site, containing a Zn ion and fragments of the four ligand amino acids (D92, D204, H90, H471). Aspartic acids were reduced to hydroxyl groups and histidine residues to imidazole rings, resulting in an overall neutral charge of the system. The ab initio QM software ORCA, version 4.2.1, was employed, using the theoretical model HF/6-31G*. First, CHELPG (CHarges from ELectrostatic Potentials using a Grid-based method) charges were calculated for the bound system. Then, the charges were calculated for the free Man unit in exact the same conformation. In a post-processing step, the CHELPG charges were set to the values for chemically equivalent atoms. In the bound system, the charges of the terminal Man atoms were modified so that their sum amounted to zero, discarding the charges computed on the ligand atoms. Partial charges calculated for the free system are comparable to the ones of the original GLYCAM06j force field (Figure 5.8 **C**). Comparing the free to the bound system, especially the Zn coordinated oxygen atoms O2 and O3 become less negatively charged, whereas the directly connected C2 atom becomes less positively charged (Figure 5.8 **C,D**). This charge redistribution effect is induced by Zn pulling electrons from the negatively charged oxygen atoms. Whether this redistribution, not observable under standard force field conditions, is able to induce a change in pucker conformation was further tested by metadynamics simulations of single Man residues with altered partial charges.

The computed CHELPG charges could be directly employed, as they represent nothing else than the so-called RESP charges that are used in the AMBER force field family. Two systems, each having a single Man residue solvated in a TIP3P water box, were built with partial charges corresponding to the free and bound system. The more important θ angle, differentiating between chair and boat conformers, was explicitly biased via metadynamics simulations of 100 ns, in which the bias was positioned every 500 steps, with parameters $\gamma = 6$, $\sigma = 0.1$, and $w_G = 2$. The resulting free energy profiles along θ showed no difference between the two systems, having the same global minimum located at $^4C_1$ (Figure 5.9). Therefore, just changing the partial charges to account for the effect of the Zn^{2+} ion in the GMII binding site is not sufficient to induce a change in pucker conformation. As already suggested by the crystal structure of PDB entry 3CZN, it is most probably the joint action of D204 and the Zn ion that upon interaction with the glycan, alter the potential energy functions around the torsion angles. Therefore, it is possible for M5G0 to position itself in the catalytic site, adopting a different conformer distribution compared to its freely solvated counterpart.[286]



	GLYCAM06j	Free	Bound
HO1	0.45	0.37	0.35
O1	-0.64	-0.65	-0.67
C1	0.43	0.37	0.35
H1	0.00	0.09	0.12
C2	0.25	0.46	0.33
H2	0.00	-0.01	0.03
C3	0.25	0.36	0.32
H3	0.00	-0.03	-0.02
C4	0.33	0.06	0.03
H4	0.00	0.08	0.08
C5	0.28	0.17	0.20
H5	0.00	0.08	0.06
C6	0.27	0.25	0.20
H62	0.00	0.04	0.04
H61	0.00	0.04	0.04
O6	-0.69	-0.67	-0.68
HO6	0.42	0.41	0.40
O5	-0.51	-0.51	-0.51
O4	-0.73	-0.75	-0.75
HO4	0.44	0.45	0.44
O3	-0.70	-0.78	-0.56
HO3	0.43	0.45	0.41
O2	-0.72	-0.72	-0.60
HO2	0.45	0.42	0.40

Figure 5.8: **Recalculating charges of the terminal Man in the catalytic site via Hartree-Fock. A** Atomistic structure of a single bound Man residue in the structurally reduced catalytic site of ManII, only including Zn^{2+} and fragments of the ligand amino acids D92, D204, H90 and H471. **B** Atomistic structure of a single Man residue. **C** Detailed list of obtained CHELPG charges from Hartree-Fock calculations for free and bound M5G0, with a comparison to the available charges in the GLYCAM06j force field. **D** Charge differences calculated for each atom of the single Man residue in the bound vs. free state. Each atom is colored according to its associated charge difference, with red being negative and blue being positive.

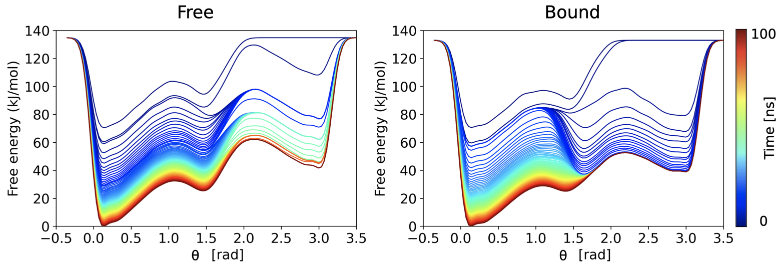


Figure 5.9: **Free energy profiles along pucker coordinate θ for different partial charge sets.** Pucker conformations of glycan M5G0 were enhanced sampled, applying different partial charges, calculate from QM simulations (Figure 5.8).

In summary, both the conformer and the charge hypotheses need to be discarded, not being the reason for a ring distortion in the catalytic site of GMII at position -1. However, the native glycan conformer of M5G0 in its reactant state could be revealed through extensive sampling via REST-RECT simulations, deviating from the suggested crystal structure due to the D204A mutation. This significant difference underlines the importance of careful evaluation of unphysiological crystal structures and the necessity of their refinement prior to the derivation of results. We have been able to successfully apply the REST-RECT methodology to a complex glycan-protein system, ensuring converged sampling of the glycan conformational phase space, independent of the difficulties connected with the protein-glycan interactions. So far, there was no method available for the explicit sampling of conjugated glycans and the prediction of converged conformer distributions. Only Yang et al. 2017[294] studied a glycosylated HIV envelope protein by means of enhanced sampling simulations, although no absolute convergence could be achieved due to insufficient sampling parameters. Consequently, our REST-RECT approach provides solid ground for studying the crucial impact of topological glycan parameters in recognition processes with glycan receptors.[13]

The comparison of the GLYCAM06j and CHARMM36 force field in this glycan-protein context again underlines the importance of parameter improvement, especially for the description of pucker conformers in the CHARMM36 force field. QM calculations of single β-mannose residues suggest a more flat puckering free energy landscape, in line with the GLYCAM06j force field. The known issue of the GLYCAM06j force field of overestimating carbohydrate-environment interactions can rather be neglected in this context, as it is exactly a stable protein-glycan complex we are interested in.[162] To check whether the overestimated electrostatic attractions between hydroxyl groups may have an impact on the obtained conformer distributions for M5G0 bound to GMII, the usage of better water models such as TIP5P, and a corresponding rescaling of Lennard-Jones parameters would be required.[162,295]

6 | Conclusion

Originally, the impact of glycan diversity was discovered via the identification of the ABO blood types by Prof. Watkins and Prof. Morgan in 1952, revealing the antigen's epitopes to be different carbohydrates.[296,297] Today, it is actually estimated that there are around 3000 different motives in complex carbohydrates that can be recognized by glycan-binding proteins, resulting from their enormous coding capability described in the first chapter of this thesis (Figure 1.2).[297] This impressive variety is also found in N-glycan structures and hampers a rational understanding of clear structure-function relationships at the basis of a still mysterious sugar code[298,299]. Detailed studies are required, for instance, in the case of the most abundant human serum antibody immunoglobulin G (IgG), where the interaction with its receptor FcγRIII was shown to be modulated by the type of glycosylation on both proteins, triggering a cellular toxicity response.[300] The compositional heterogeneity of N-glycans that leads to different configurations is further complemented by their structural flexibility, giving rise to different conformations. The question remains open, to what extent these accessible conformers, known as the third dimension of the sugar code, impact the N-glycans' functions.

6.1 N-glycans' 3D structure-function relationships

Expectations

Although N-glycans are ubiquitous in eukaryotic cells and associated to various biological effects, their function is mostly dependent on the monosaccharide composition and putatively associated three-dimensional structure. We claimed that this three-dimensional structure-to-function relationship requires a correct and comprehensive description of the whole ensemble of glycan conformers, motivating a fundamental exploration of the conformational phase space of representative N-glycans in various cellular contexts. Our idea was addressed mainly by atomistic simulations, because they allow for result interpretation at the molecular level and the systems to be investigated are only limited by the availability of force field parameters and structural information. Applying our developed sampling and analysis workflow of chapter 2 and 3 to differently situated N-glycans throughout the dissertation, we expected the elucidation of possible three-dimensional structure-function relationships.

Findings

Comparison of the phase space explored by differently shaped free N-glycans in solution revealed rather subtle differences in conformer distribution for their common monosaccharide residues, except for M9 in comparison to M5. Our findings are in agreement with another computational study that did not identify an effect of core fucosylation and sialylation on the conformational dynamics of complex N-glycans.[222] They, however, revealed an effect of branch galactosylation on the dynamics of the α1-6 branch, shifting the population towards a 'folded-over' conformation.[222] Consequently, conformational effects due to the elongation of glycan structures by monosaccharides are possible and can result in significant biological effect as in the case of IgG Fc glycosylation, where galactosylation of complex N-glycans is associated to aging and immune activation.[222,301,302]

Another three-dimensional structure-function effect could be identified for the protein-bound glycan N206 of TconTS1, which explored a different phase space region as a consequence of protein-glycan interactions. This was estimated to have a significant impact on the enzyme's function. Here, the third dimension of the sugar code took effect, allowing for stable protein-glycan interactions that were not observable when glycan M5 at N206 was restrained to its most populated solution conformer, showing no hydrogen bonding to both D150 and Y151 (Figure 4.11). This clearly underlines the importance of glycan flexibility and corresponding shifted conformer distributions. As glycosylation patterns are known to be dynamical, the occupation of site N206 by other high-mannose type N-glycans like M7 to M9 could also be expected, although the impact of this change in glycoform can not be foreseen and would require further MD simulations. The recent example of the SARS-CoV-2 Spike protein's glycosylation sites affecting the stability of the RBD open conformation, regulating its accessibility to ACE2, underlines the importance to take the dynamics of glycosylation site heterogeneity into account.[303] A dependence has been revealed between high-mannose type N-glycan size at three important glycosylation sites (N165, N234, N343) and the N-glycan's ability to interact with protein domains. A reduction in glycan size resulted in a shift of adopted protein conformers within the RBD. Consequently, there exists an intrinsic dependence between glycan configuration with their associated conformations and site-specific functions. The complete investigation of the sequence to three-dimensional structure-to-function relationships of N-glycans for TS enzymes would additionally require the investigation of TS from other trypanosomal species, harboring the same conserved N-glycosylation site but without any known effects. This could reveal a putatively conserved glycan-mediated enzymatic regulation mechanism, adding a new entry to the sugar code dictionary for a larger portion of *Trypanosoma* species.

The conformation of N-glycans is particularly important when serving as substrates or interaction partners, because recognition by CAZymes or other glycan-binding proteins happens at specific binding interfaces with a certain predefined morphology to which the ligand molecule has to adapt. An effect that goes beyond this expectation was demonstrated with the application of the REST-RECT methodology to GMII in complex with its native substrate M5G0. We were able to show that upon binding, not only a subset of the glycan's conformational phase space could be adopted compared to free M5G0, but protein-carbohydrate interactions even led to the exploration of new regions of the phase space. The fact that only certain conformers can be adopted within the binding site underlines that GMII only allows for the binding of specific conformer keys from a bunch of keys corresponding to all possible conformations. This is in line with the key-and-lock mechanism that was extended for carbohydrates by Barry Hardy in 1997[71], hypothesizing that one glycan configuration harbors multiple keys due to its conformational phase space. However, our results also suggest a kind of induced-fit model that goes beyond the original formulation by Daniel Koshland in 1958[304], in which the protein undergoes conformational changes to adapt to the ligand structure. Namely, we observe a complementary induced-fit, in which the glycan ligand is forced out of its equilibrium conformational phase space due to protein-carbohydrate interactions. Another protein-induced fit was found for the pucker conformation of the terminal Man residue, where our results did not detect a self-induction mechanism related to specific glycan conformers favoring the adaptation of specific pucker coordinates, refuting the assumption of mutual dependence.

Interpretation

The most prominent issues of MD simulations (chapter 2), namely: (i) accuracy of force fields, (ii) limitation of simulation time and (iii) dimensionality of output data, could be mostly tackled for the field of computational structural glycobiology. This provides a suitable framework for the qualitative generation of structural glycan data as well as their quantitative analysis. Our proposed REST-RECT algorithm overcame the shortage of simulation time as it did not only allowed for the enhanced sampling of free N-glycans in solution, but also ensured complete phase-space exploration of protein-bound glycans despite the constraints dictated by amino acid-glycan interactions. In comparison to other successful glycan sampling algorithms such as the replica exchange scheme of Yang and coworkers[94], our algorithm presents an easy-to-use approach without the necessity of pre-calculations or specific CV selection. The great success of converging the protein-bound glycan simulation of GMII in complex with M5G0 is especially pioneering, as the few attempts of enhanced sampling protein-bound glycans were either insufficient in terms of convergence or simply did not validate this critical methodological feature.[294,300,305] Further, it needs to be highlighted that we were the first to include the sampling and convergence of puckering coordinates, being a fundamental feature of especially terminal monosaccharides when serving as substrates for CAZymes. It can be concluded that our suggested workflow was able to reveal important features of the third dimension of the sugar code for all three discussed N-glycan systems, based on the exploration and comparison of conformational phase spaces.

The development of the here-introduced conformer string based on the torsion angles setting of N-glycans especially facilitated the task of unraveling the sugar code for various N-glycan systems. It provides a solid and IUPAC-nomenclature-compliant way of labeling different N-glycan conformers. The developed python script GlyCONFORMER for an automated assignment of conformer strings to any sampled N-glycan structure will hopefully be helpful in the future to other structural glycobiologists and paves the way for a uniform glycan conformer labeling. In conjunction with the complementing clustering and dimensionality reduction analysis, our approach outperforms the previously introduced spherical coordinates as well as simple ϕ/ψ plots, and can provide a new standardized way for reporting glycan conformers in the future.[96] It is especially able to tackle the dimensionality of glycans imposed by the many torsion angles, reducing the barrier to compare structures and draw conclusions from their phase-space distribution. The developed visualization workflow already proved to be useful not only in our own applications, but also in the context of e.g. GlycoSHIELD, validating the reference data sets for grafting N-glycans on protein structures. The effort to validate and improve force field parameters of glycans could be facilitate in the future by our conformer string and visualization technique, as already briefly performed in chapter 3.

6.2 Shortcomings

It remains to say that our work did not intend to improve any force field parametrization, but rather verified them by comparison to experimental studies and particularly highlighting their differences. Despite the usage of the simple water model TIP3P, we could underline the difference between puckering coordinates exploration between the CHARMM36 and GLYCAM06j, next to other subtle differences for glycosidic linkages. We have to conclude that further evaluation would also require the usage of more accurate water models like TIP5P, which was not performed within this dissertation due to the compromise between accuracy and computational efficiency. Further improvement of additive force field terms for e.g. glycosidic linkages, which are mostly dependent on steric, electrostatic and torsional energy terms, might reach a performance plateau at some point, resulting from the lack of polarizability. New approaches like the CHARMM Drude polarizable force field utilize virtual 'Drude' particles that are charged and connected to every polarizable parent atom via an harmonic spring, mimicking the deformation of the electron cloud of the parent atom due to the surrounding electrostatic environment.[145,306] Although parameters for most N-glycan monosaccharides and linkages are still missing, a recent work derived polarizable force field parameters for glucuronic acid and N-acetyl galactosamine in order to be able to model the unsulfated GAG chondroitin.[307] Next to torsion angle values that were comparable to NMR experiments, it was very interesting to see that the Drude model, in comparison to the standard one, is able to sample a larger Cremer–Pople puckering coordinate space, eliminating the restricted sampling of the CHARMM36 force field reported by us in chapter 3 and 5. The application of polarizable force fields to the GMII simulation system in chapter 5 would be especially helpful, in order to at least partially include the effect of the Zn^{2+} ion on the electrostatic potential of the bound glycan ligand. With the usage of fixed-charged force fields, an elaborate reparameterization of at least electrostatic and torsion angle parameters for the glycan atoms, but probably also of other surrounding amino acids, would be necessary.

Furthermore, the field of machine learning also entered the stage in order to improve parameters for existing classical force fields models.[162] Machine learning-based force fields may provide an even higher accuracy and extend the possibility of MD simulations by treating even large molecular systems on a QM level, including the natural description of chemical reactions.[308] This is possible due to the estimation of potential energies, depending on the position and charge of atoms in the system, from e.g. *ab initio* data as training sets, as it is done in the case of the ANI-1 potential or TORCHMD-NET with a deep neural network.[309,310]

6.3 Perspectives

The chicken-and-egg problem in glycobiology

When elaborating about the aspects of the sugar code, it will not have escaped the attentive reader that, especially for N-glycans, a high degree of conservation of certain structural features exists. Especially the first five monosaccharides of the non-reducing end (two GlcNAc and three Man) and their linkage types are identical for all three N-glycan types, high-mannose, complex and hybride, forming the so called core of every N-glycan. The conservation goes back to the internal glycosylation machinery, where $\mathrm{Glc_3Man_9GlcNAc_2}$ is the only precursor for all N-glycans that are able to be synthesized. Independently of the clade, being it Obazoa (fungi, insects, animal, molluscs), Excavates (single-cell flagellate organisms, parasites) or Archaeplastida (plants, algae), all eukaryotes exhibit the same N-glycan core structure.[311]

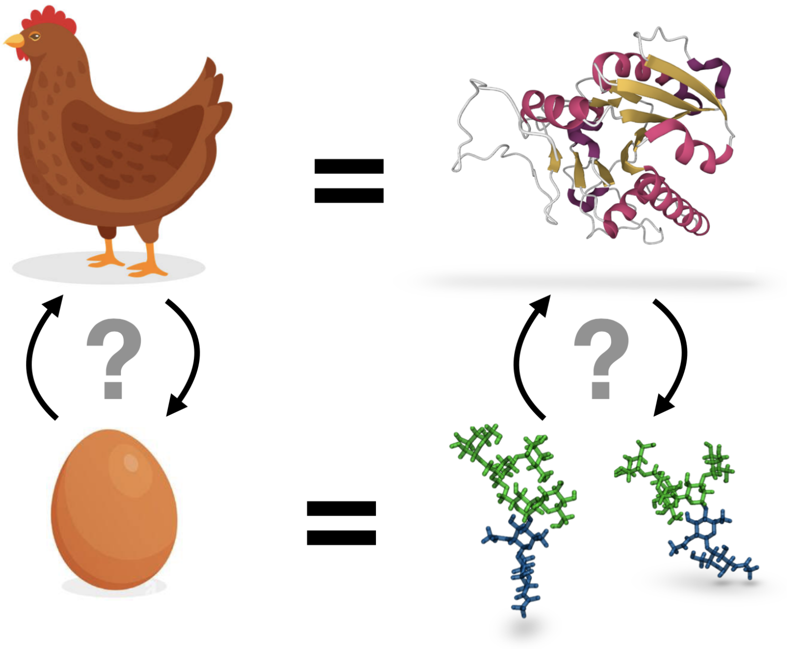


Figure 6.1: **What was first?** Analogy between the classical chicken-and-egg problem and the question whether glycosyltransferases (GTs) shaped the N-glycan structures or if only energetically favorable N-glycan structures shaped the catalytic sites of glycosyltransferases.

Therefore, the question arises what were the major determinants for the evolution of this conserved motive: $\mathrm{Man}_{\alpha1\rightarrow6}$ $[\mathrm{Man}_{\alpha1\rightarrow3}]$ $\mathrm{Man}_{\beta1\rightarrow4}$ $\mathrm{GlcNAc}_{\beta1\rightarrow4}$ $\mathrm{GlcNAc}_{\beta1\rightarrow N}$. One could hypothesize that the core structure evolved with the availability of GTs that are known to only catalyze one specific transfer reaction, only forming one linkage type and therefore being a contingent result of the available enzyme chemistry. In contrast, the conserved structural features could simultaneously also be the most energetically favorable ones and GTs might have evolved to construct exactly these features. This fundamental research question is comparable to the classical chicken-and-egg problem paradox (Figure 6.1).

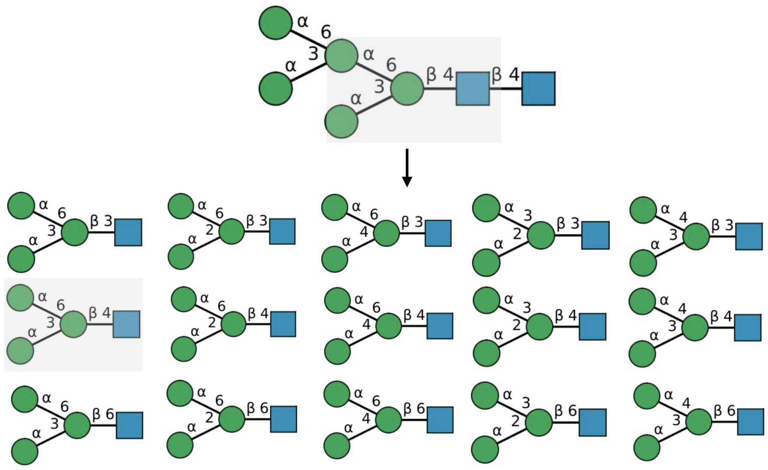


Figure 6.2: **Variations of the conserved N-glycan core.** On top, the smallest biologically relevant N-glycan is shown, where the conserved core junction is highlighted in gray. Below, schematic representation of N-glycan core junctions with varying connectivity between the invariable four monosaccharides. Configurations differ in the carbon that is forming the glycosidic linkage, where all chemically possible structures have been generated.

The outlined issue addresses the mystery of the sugar code from another perspective. For instance, the five-membered monosaccharide core could be studied to verify if its structural conservation is dictated by thermodynamic stability. The next steps in deciphering the conservation of the N-glycan sugar code would involve the computation of the free energy difference between each core structure that is varying in connectivity (Figure 6.2), ranking relative stability of the native configuration against the artificial ones.

The secrets of fucoidan

The topic of thermodynamic stability was also broad up by our collaborator Jan-Hendrik Hehemann from the Max Planck Institute for Marine Microbiology, curious about the three-dimensional structure of the complex extracellular matrix polysaccharide fucoidan, which is secreted by brown algae into the ocean. Depending on the specific species, it exhibits slightly different composition and linkage patterns, having sulfated fucose as its main monosaccharide type in common.[312] Fucoidan can be described as a bioactive macromolecule, associated with diverse pharmacological and pharmaceutical applications such as anti-tumor, anti-coagulant and anti-viral, although the exact mechanisms are undetermined.[313] Its effect was often dependent on the level of sulfation and specific sulfate pattern, posing the question how polysaccharide conformation is changing upon alterations of these properties. For instance, an experimentally predicted molecular model of fucoidan from *Laminaria hyperborea* consists of an α1-3 Fuc backbone with frequent small branches of type α1-2 and α1-4 (Figure 6.3 **A**), where the degree of sulfation was linked to coagulation.[314] Simulation of a short fragment of this polysaccharide (Figure 6.3 **B**) at different levels of sulfation could give insights into potential conformational effects that are linked to specific functions.

Enhanced sampling simulations via REST-RECT of the mentioned fucoidan molecule, sulfated and unsulfated, revealed three main conformers for each variant (Figure 6.3 **C/D**), where the conformer strings are completely different for the two molecules. Intramolecular conformer comparison shows that only one torsion angle is different among the three main conformers, respectively, mainly altering the orientation of the branches. Counterintuitively, the unsulfated fucoidan exhibits a much higher conformer stability for its first two conformers (90 %) compared to sulfated fucoidan (70 %), indicating a reduced flexibility of the unsulfated fucoidan. This finding was surprising as we expected the negatively charged sulfate groups to repel each other, making carbohydrate-carbohydrate interactions unfavorable, and therefore allowing for less flexibility and less accessible conformers.

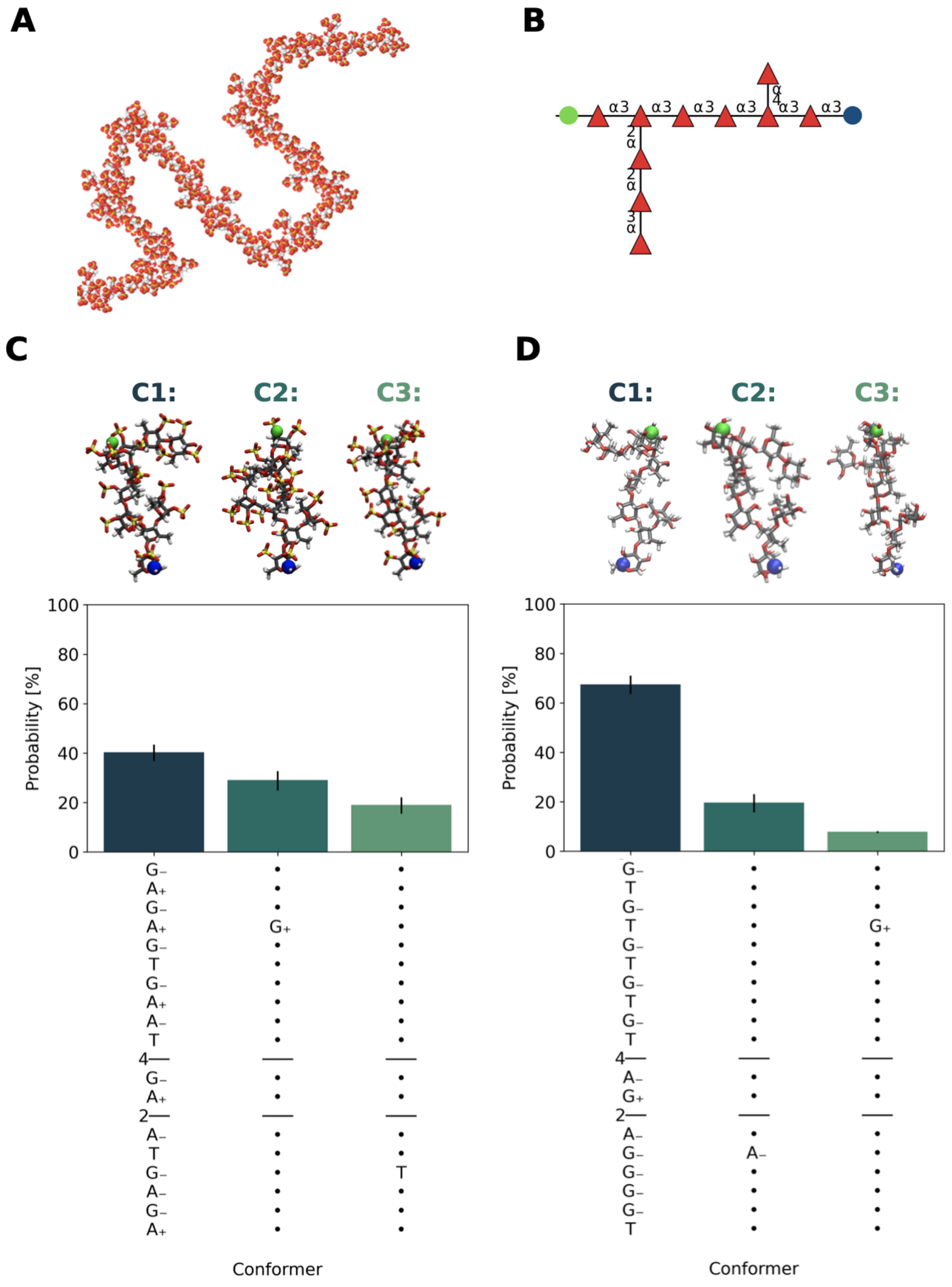


Figure 6.3: **Structural ensemble of fucoidan variants.** **A** Putative full-scale structural model of sulfated fucoidan from *Laminaria hyperborea*. Adapted from Kopplin et al.[314] **B** Two-dimensional sequence model of a ten-residue fragment of fucoidan with two representative branches. **C** Conformer distribution for the sulfated fucoidan molecule in **B**, enhanced sampled via REST-RECT with representative atomistic structures of each conformer. **D** Conformer distribution for the unsulfated fucoidan molecule in **B**, enhanced sampled via REST-RECT with representative atomistic structures of each conformer. The reducing end of the fucoidan fragment is indicated by a blue circle/ball and the opposing end of the α1-3 Fuc backbone by a green circle/ball to enhance visibility.

Comparing the accessible phase-space region of both fucoidan variants revealed almost no overlaps, as already indicated by the dissimilar conformer strings (Figure 6.4). The rigidity of the unsulfated fucoidan is further underlined by only two accessible free energy minima versus four for sulfated fucoidan (Figure 6.4) We can conclude that a varying sulfation level induces a change in three-dimensional structure of fucoidan, at least for the all or nothing cases we have studied here. This result was only possible due to the application of the REST-RECT approach and our invented conformer string. Further implications about possible functions of this conformer rearrangement require the future simulation of larger fucoidan polysaccharides, as depicted in Figure 6.3 **A**. As already demonstrated, we would like to note that our developed framework is not limited to the investigation of N-glycans, but can be extended to glycan classes with other linkage types, such as O-glycans, GPI-anchors, GAGs, glycosphingolipids or, as in this application, polysaccharides.

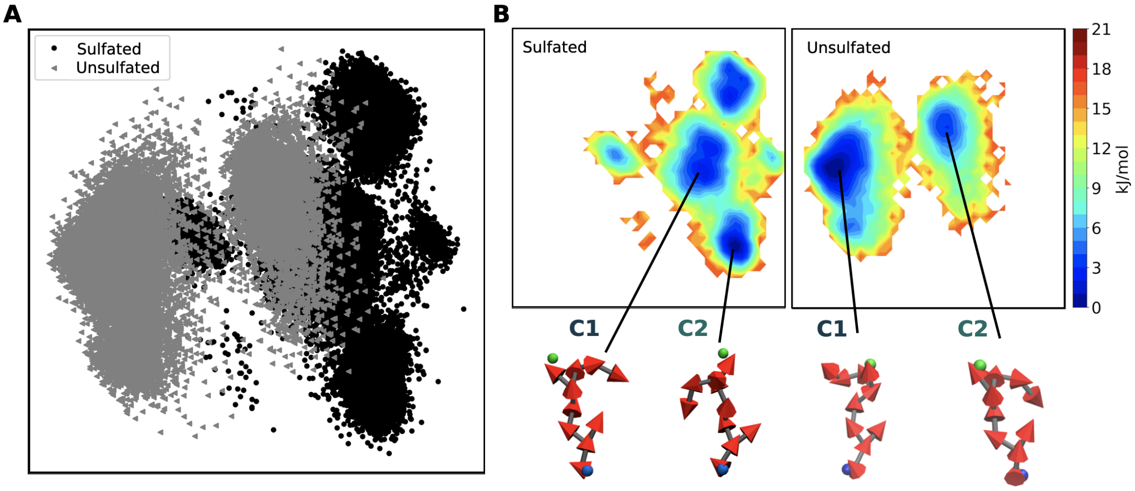


Figure 6.4: **Comparison of sulfated and unsulfated fucoidan conformers.** **A** Joint PCA of the complete distribution of sulfated and unsulfated fucoidan conformers, where each point represents one sampled frame. **B** Joint PCA, visualizing the conformational free energy landscape of sulfated and unsulated fucoidan, separately. The reducing end is indicated by a blue circle/ball and the opposing end of the α1-3 Fuc backbone by a green circle/ball to enhance visibility.

6.4 Closing words

Glycobiologist Robert Woods, developer of the GLYCAM force field parameters, stated in his review about glycan structures from 2018 that:

> "Theoretical improvements in carbohydrate modeling have led to a much greater depth of understanding [...], but they have not profoundly altered many of the conclusions derived from extremely approximate early models, at least regarding their conformational preferences."[83]

Although I partly agree with this statement, it must be added that especially glycan-protein interactions hinder the observation of relevant, bio-active glycan conformers under standard MD conditions, thus justifying the large effort that was put in this dissertation to improve carbohydrate sampling techniques for glycoproteins. Further, I can not help suspecting that despite the here-introduced elucidation of the third dimension of the sugar code, its structural origin leading to a high flexibility has not only protein-specific effects, but the ‘third dimension’ concept also has to be considered at different levels of complexity. On the one hand, there are cases where site-specific and type-specific N-glycans are necessary in order to confer function to the protein as we have seen for TconTS1 in this dissertation, but also for examples from the literature like IgG and the SARS-CoV-2 Spike protein.[300,303] On the other hand, glycan shields that hide the immunogenic protein surface from the humoral immune system are due to the N-glycan abundance rather than their composition, three-dimensional structure, or specific protein-carbohydrate interactions. For instance, coronaviruses that have been circulating among humans for longer time than SARS-CoV-2 have a denser glycan shield with nearly twice the number of glycan sites.[315,316]

Despite our knowledge about certain sequence to three-dimensional structure-to-function relationships of N-glycans that gradually extend the glycan dictionary,I believe it is still not generally possible to predict the far-reaching consequences of site-specific protein glycosylation *a priori*, if it ever will be. It is therefore not superfluous to repeat the immense importance of properly including N-glycan structures in MD simulations, be it indirectly through the usage of tools like GlycoSHIELD or directly by means of enhanced sampling of pre-attached glycan structures.

A | Molecular dynamics simulation

A.1 Choice of the time step

As mentioned in chapter 2, solving the equations of motion via numerical integration requires the introduction of a time step δt. Larger time steps maximize the achievable simulation time. However, forces are kept fixed during each time step, leading to severe artifacts if the time step is chosen too large, hence the dynamics become nonphysical. An important parameter to consider when determining an appropriate time step are the fastest motions in the system in order to ensure the conservation of energy. For biomolecular systems, these are the vibrational frequencies of bonds, being around 3500 cm^{-1} for the fastest bond (O-H). As Nyquist's theorem[317] states that the sampling rate should be at least twice the frequency of the highest frequency in the wave, we would require a time step of around 1 fs to consider the fastest vibrating bond. The reciprocal term of the frequency w, the oscillating period, is:

$$\begin{aligned} &\frac{1}{w} \approx 3 \cdot 10^{-15}\,\mathrm{s} \approx 3\,\mathrm{fs} \\ &\text{with} \\ &w = 3 \cdot (3 \cdot 10^{10})\,\frac{\mathrm{cm}}{\mathrm{s}} \cdot 3500\,\frac{1}{\mathrm{cm}}. \end{aligned} \tag{A.1.1}$$

In order to increase this time step, constrained dynamics can be employed. Here, bond vibrations (mostly O-H bonds) are replaced by holonomic constraints to eliminate fast motions in the evaluation of the time step. Several different algorithms have been proposed for this purpose, with SHAKE[318] and LINCS[210] being the most popular ones in MD simulations.

A.2 Periodic boundary conditions

In order to start a simulation by integrating the equations of motion, not only the desired ensemble needs to be selected, but also special care has to be taken about the simulation box itself. The box size and shape is determined by three basis vectors $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$. Different shapes are available, although we only employed cubic boxes here, which is also the most simple shape to use. Considering a box filled with 100'000 particles, a considerable fraction of them is interacting with the edges. If these would be treated as physical walls, the interaction of particles at the edges would be very different from particles in the bulk and hence altering the thermophysical properties of the system in a nonphysical way.[148] In order to mimic bulk conditions in all parts of the simulation box, periodic boundary conditions (PBC) are applied. They surround the central simulation box with an infinite number of identical copies, which implies that when a diffusing particle leaves the box on one side, it immediately enters from the opposite side, conserving the number of particles.[148]

A.3 Treating long range interactions

Intramolecular interactions like bonds, angles and dihedrals just consider the direct neighbors of the investigated atoms and therefore bonded terms do not require a special treatment. However, non-bonded terms have interaction ranges several orders larger and would require the consideration of intermolecular interactions between atoms beyond the walls of their simulation box. Under PBC conditions, this is especially problematic as the atom would self-interact through its periodic images, leading to wrong energies. Since the largest contribution to the potential energy and forces on any atom results from its nearest neighbors, the first prevention measure is the application of the Minimum Image Convention (MIC), neglecting interactions with distant periodic images.[148] It states that each atom i only feels the nearest periodic image of any other atom, therefore considering only interactions within a spherical radius $r_{cut,MIC} = L/2$ around atom i, with L being the box length. Furthermore, cut-offs are applied as explained in the following sections.

A.3.1 Verlet cut-off - short range electrostatics

The potential energy function assumes a pairwise additivity of interaction energies and therefore requires the consideration of $\frac{1}{2}N(N-1)$ pairs to determine the potential energy for a total number of N atoms for long range intermolecular interactions.[148] As this calculation is the most resource-consuming part of each MD simulation, it is questionable whether the pair potential needs to be computed every step, even for atoms that lie on opposite sides of the simulation cell. Therefore, a spherical cut-off r_{cut} can be introduced, limiting the number of interaction pairs by:

$$r_{ij} \leq r_{cut} \tag{A.3.1}$$

with r_{ij} being the distance between atoms i and j.[148] In order to prevent the repetitive calculation of r_{ij} every step, Verlet[143] proposed a neighbor list which stores nearest neighbors within $r_{cut} + \delta r$ of each atom i, and is only updated every 10 to 100 steps. The reservoir δr should guarantee that an atom j diffusing into r_{cut} of atom i is recognized although the neighbor list was not updated. Instead of the simulation time scaling by $\propto N^2$, the application of the Verlet list method, especially to the LJ term, reduces it to $\propto N$.

A.3.2 Ewald summation - long range electrostatics

The Verlet cut-off scheme can not be applied to speed up the calculation of the electrostatic interactions, as these decay with r^{-1}. Therefore, their interaction radius exceeds half of the box length $L/2$, which is not compatible with the Minimum Image convention. Because of the long range nature of the electrostatic interactions, it would not even be sufficient to consider only the nearest neighbor atoms of the central simulation box, but also the infinite number of periodic images to guarantee energy conservation. The total electrostatic interaction energy would be given by:

$$E_{electrostatic} = \frac{1}{2} \sum_{n}^{\prime} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{q_i q_j}{4\pi\varepsilon(\mathbf{r}_{ij} + \mathbf{n}L)}, \tag{A.3.2}$$

wherein $\mathbf{r}_{ij}$ is the real distance between charges and not the minimum-image, and $\mathbf{n} = \begin{pmatrix} n_x \\ n_y \\ n_z \end{pmatrix}$ the box index vector.[148] Self interactions between q_i and q_j for $i = j$ in the central box ($\mathbf{n} = \mathbf{0}$) are omitted, indicated by the prime symbol. As the sum is only conditionally converging, the Ewald summation[319] or Smooth Particle Mesh Ewald method (SPME)[213] can be employed to obtain a rapid and absolute converging sum. The Ewald summation splits the electrostatic term into series, where they can be differentiated into short-range and long-range interactions as well as two correction terms for self interactions E_{self} and intramolecular interactions within the same molecule E_{corr}:

$$E_{electrostatic} = E_{short-range} + E_{long-range} - E_{self} - E_{corr}. \tag{A.3.3}$$

The short-range interactions are characterized by each charge q_i being surrounded by a symmetric neutralizing diffusive charge cloud (Gaussian charge distribution), having the same magnitude as q_i but of opposite sign.[148] The resulting electrostatic potential of q_i and its charge cloud are given by $\frac{q_i}{r}\mathrm{erfc}(\sqrt{\alpha r})$, wherein α defines the width of the distribution, r the intercharge distance and erfc being the complementary error function. The resulting contribution to $E_{electrostatic}$ from all screened charges in the central box is given by[148]:

$$E_{short-range} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{q_i q_j}{r_{ij}} \mathrm{erfc}(\sqrt{\alpha r_{ij}}), \tag{A.3.4}$$

when α is chosen large enough as interactions are limited to short ranges. The induced screening charge requires a second compensating charge distribution of opposite sign to counteract the first, whereas here the interactions with the periodic images have to be taken into account. The long-range characteristic of this second charge distribution makes it necessary to perform it in reciprocal space applying Fourier transformation, in order to obtain a rapid convergence.[148] The long-range part of $E_{electrostatic}$, characterizing interactions of the second compensating charge distribution, is given by:

$$E_{long-range} = \frac{1}{2} \sum_{k \neq 0} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{4\pi}{Vk^2} q_i q_j \exp[-i\mathbf{k} \cdot (\mathbf{r}_i - \mathbf{r}_j)] \exp(-k^2/4\alpha), \tag{A.3.5}$$

with $\mathbf{k}$ being the reciprocal vector $= 2\pi\mathbf{n}/L^2$. The more efficient SPME method was developed to improve the efficiency of the computation of this reciprocal part of the Ewald term.[148] First, the charges are assigned to grid points by B-spline functions, interpolating the reciprocal space on a regular grid. A Fast Fourier Transform algorithm is then applied to obtain the reciprocal energy term by a single sum over the grid points.

A.4 Thermostats and Barostats

The temperature T of a system can be directly related to the velocity $\mathbf{v}_i$ of its particles by

$$\frac{1}{2}\sum_{i=1}^{N} m_i \cdot \mathbf{v}_i^2 = \frac{3}{2} N k_B T, \tag{A.4.1}$$

where both sides of the equation are equal to the kinetic energy E_{kin}. The instantaneous temperature $T(t)$ at time point t is given by

$$T(t) = \frac{\sum_{i=1}^{N} \frac{m_i}{2} \mathbf{v}_i^2(t)}{N_f k_B}, \tag{A.4.2}$$

wherein $N_f = \frac{3}{2}N$ describes the degrees of freedom of the system.[148] The most simplistic way of achieving $T(t) = T_0$, where T_0 is the target temperature of the simulation box, is by adjusting the velocity with a scaling factor $c(t)$, where the resulting temperature change ΔT can be estimated by:

$$\Delta T = \frac{\sum_{i=1}^{N} \frac{m_i}{2} (c(t) \cdot \mathbf{v}_i(t))^2}{N_f k_B} - \frac{\sum_{i=1}^{N} \frac{m_i}{2} \mathbf{v}_i^2(t)}{N_f k_B} \tag{A.4.3}$$

$$\Delta T = (c(t)^2 - 1) \cdot T(t) \tag{A.4.4}$$

and yields a scaling factor of

$$c(t) = \sqrt{\frac{T_0}{T(t)}}. \tag{A.4.5}$$

A similar approach that is used in MD simulations is the weak-coupling Berendsen thermostat[320], where the coupling is performed every step with a difference in temperatures T_0 and $T(t)$ that is proportional to the rate of change of temperature:

$$\frac{dT(t)}{dt} = \frac{1}{\tau}(T_0 - T(t)), \tag{A.4.6}$$

introducing the coupling parameter τ which adjusts the strength of the coupling. The scaling factor then becomes

$$c(t) = \sqrt{1 + \frac{\delta t}{\tau}\left\{\frac{T_0}{T(t - \frac{1}{2}\delta t)} - 1\right\}} \tag{A.4.7}$$

considering the leap-frog algorithm for integration.[169] The drawback of this method is that it does not generate a canonical distribution, as the fluctuations of the kinetic energy are suppressed by the scaling factor. However, the velocity-rescaling thermostat[211] is a modified version of the Berendsen thermostat which ensures a correct kinetic energy distribution, generating a canonical ensemble, and was therefore used in all MD simulations in this dissertation. It adds an additional stochastic term enabling a correct kinetic energy distribution by modifying the kinetic energy directly:

$$dE_{kin} = \frac{\delta t}{\tau}(E_{kin,0} - E_{kin}) + 2\sqrt{\frac{E_{kin}E_{kin,0}}{N_f}\frac{dW}{\sqrt{\tau}}}, \tag{A.4.8}$$

wherein dW is a Wiener noise and $E_{kin,0}$ the target value of the kinetic energy.[169]

Similar to thermostats, there are also barostats which couple the system to an external pressure in order to rescale e.g. box vectors or coordinates and keep the pressure at a target value $\mathbf{P}_0$. The Berendsen barostat works in the same way as its thermostat, modifying the equations of motion to relax the instantaneous pressure $\mathbf{P}$ according to the rate of pressure change (compare equation A.4.6):

$$\frac{d\mathbf{P}(t)}{dt} = \frac{1}{\tau_p}(\mathbf{P}_0 - \mathbf{P}(t)), \tag{A.4.9}$$

wherein τ_p is the coupling strength to the external pressure. The pressure scaling factor

$$\mu = \left(1 + \frac{\delta t}{\tau_p}\beta_p \left\{\mathbf{P}_0 - \mathbf{P}(t)\right\}\right)^{\frac{1}{3}}, \tag{A.4.10}$$

with β_p being the isothermal compressibility, scales the coordinates and the box edges, leading to a change in volume and hence pressure.[169] Also the Berendsen barostat does not generate a canonical ensemble. However, due to the weak coupling, a smooth change of the system is ensured. A more sophisticated approach is the Parrinello-Rahman barostat[321,322], which is an extended system coupling method, extending the equations of motion by an additional degree of freedom, namely the desired quantity to control:

$$\frac{d^2\mathbf{r}_i}{dt^2} = \frac{\mathbf{F}_i}{m_i} - \mathbf{M}\frac{d\mathbf{r}_i}{dt}, \tag{A.4.11}$$

where the extra term is comparable to a frictional term and $\mathbf{M}$ depends on the box vectors $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ and $\mathbf{W}^{-1}$.[169] The mass parameter matrix $\mathbf{W}^{-1}$ defines the strength of the coupling and relies on the same parameters as the Berendsen barostat, namely β_p, τ_p and additionally on the largest box matrix element L:

$$\mathbf{W}_{ij}^{-1} = \frac{4\pi^2\beta_{p,ij}}{3\tau_p^2 L}. \tag{A.4.12}$$

It allows the box volume as well as its shape to fluctuate, while preserving the canonical ensemble even for small simulation boxes.

B | Secondary structure determination with circular dichroism

Circular dichroism is a spectroscopy technique that is able to obtain conformational information of protein and nucleic acid structures, based on the optical activity of chiral molecules.

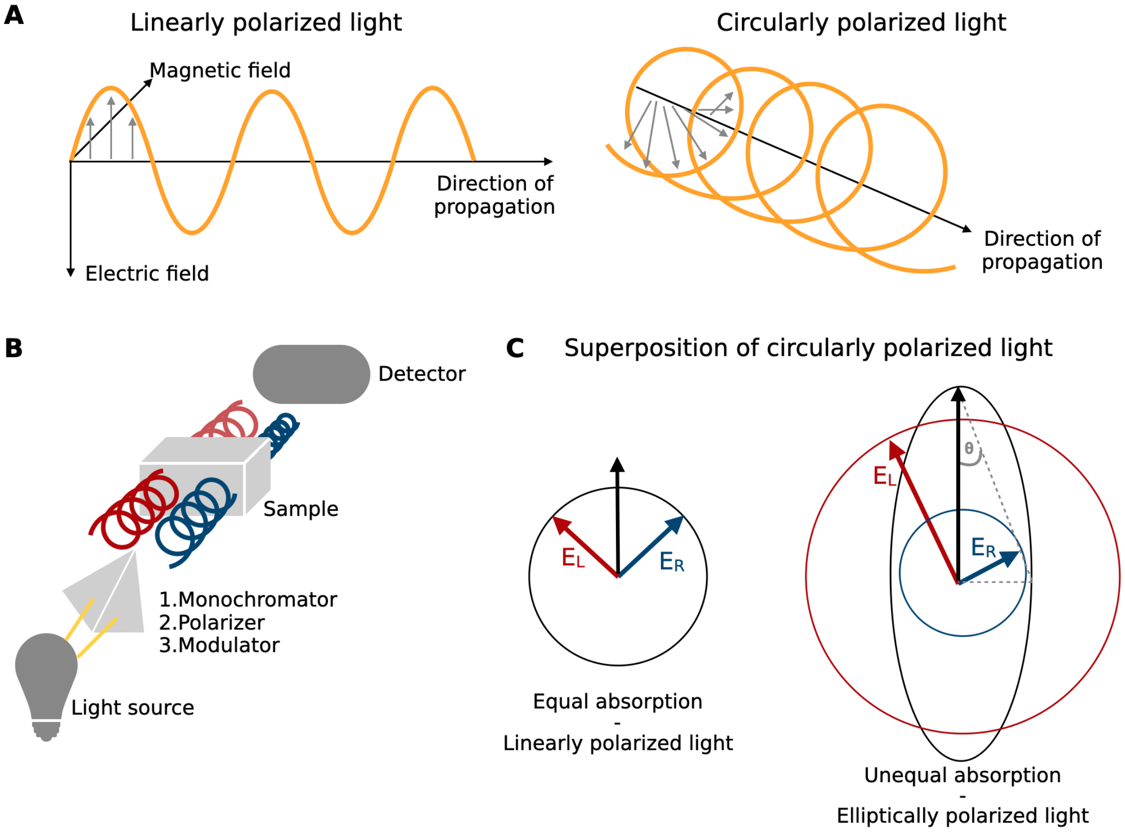


Figure B.1: **Principles of circular dichroism.** **A** Linearly and circularly polarized light along their direction of propagation. **B** Schematic setup of a circular dichroism spectrometer, generating circularly polarized light. Left-handed and right-handed circularly polarized light is passed through a chiral sample that absorbs the two components differently. The differential absorption is recorded by a detector. **C** Superposition of left-handed and right-handed circularly polarized light, that has passed through a sample. Left: Linearly polarized light is produced, as the magnitude of the electric field vectors E_L and E_R (for left and right) is equal. When E_L and E_R are unequal due to differential absorption, elliptically polarized light is generated.

Chirality, lacking microscopic mirror symmetry, arises due to the internal structure of a molecule or its linkage to a chiral center.[323] Its optical activity is characterized by interaction with linearly polarized light, as it rotates the orientation of the plane of polarization about the optical axis. This effect can be measured with a polarimeter like a circular dichroism spectrometer, equipped with a light source, monochromator, linear polarizer and photoelastic modulator (Figure B.1 **B**). Unpolarized light from a light source is passed through a monochromator in order to spatially separate the colors of light, filtering for the wavelengths of interest. A subsequent linear polarizer filters the beam of light, generating linearly, well-defined polarized light. This is characterized by its electric field vector oscillating only in one plane along the propagating direction (Figure B.1 **A**). Circularly polarized light, with the electric field vector rotating with a constant magnitude about its propagation direction (Figure B.1 **A**), is further generated by the modulator, where an alternating electrical field is applied.[323] There exists left-handed and right-handed circularly polarized light, where passage through an optically active sample can result in no absorption, the same amount of absorption, or differential absorption of one of the components ($\mathrm{A_L}$ and $\mathrm{A_R}$). In the two former cases, radiation in the originally polarized plane (linear polarization) would be generated, as the electric field vectors $\mathrm{E_L}$ and $\mathrm{E_R}$ have the same magnitude and opposite rotations (Figure B.1 **C**). The latter case, where one component is more strongly absorbed than the other, $\mathrm{E_L} \neq \mathrm{E_R}$, results in elliptically polarized light, giving rise to the effect of circular dichroism (Figure B.1 **C**). The difference in absorption results from two different refractive indices for left and right circularly polarized light in chiral molecules, leading to a changed velocity and wavelength for both components as they pass through the sample. Subsequently, a wavelength-dependent difference in absorption can be observed. A circular dichroism spectrometer is detecting the two components separately and assesses their differential absorbance:

$$\Delta A = A_L - A_R, \tag{B.0.1}$$

in milliabsorbance [mA] units. Another common output unit is ellipticity θ [mdeg], the angle whose tangent is the ratio of the minor to the major axis of the ellipse (Figure B.1 **C**):

$$\tan\theta = \frac{\mathrm{E_R} - \mathrm{E_L}}{E_R + E_L} \tag{B.0.2}$$

It is related to absorbance as:

$$\theta = 32.98 \cdot \Delta A. \tag{B.0.3}$$

In order to correct for the concentration $c\,$[g/L] of the sample and pathlength $l\,$[cm] of the cuvette used for measuring, the molar absorbance per residue

$$\Delta\epsilon = \epsilon_L - \epsilon_R = \frac{\Delta A}{c \cdot l} = \frac{\theta}{c \cdot l \cdot 32.98}\,[\mathrm{cm^{-1}M^{-1}}] \tag{B.0.4}$$

can be calculated. A more commonly used output unit is mean residue ellipticity Θ_{MRE} (per residue) in the historical unit [deg cm^{2}dmol^{-1}]:

$$\Theta_{MRE} = \Delta\epsilon \cdot 3298 = \frac{\theta \cdot MRW \cdot 0.1}{c \cdot l}, \tag{B.0.5}$$

with MRW being the mean residue weight, derived from dividing the average molecular weight M [g/mol] by the number of amino acids $N - 1$. The formula can also be rearranged to:

$$\Theta_{MRE} = \frac{\theta \cdot 0.1}{c \cdot l \cdot (N-1)}, \tag{B.0.6}$$

with c in [mol/l].[324]

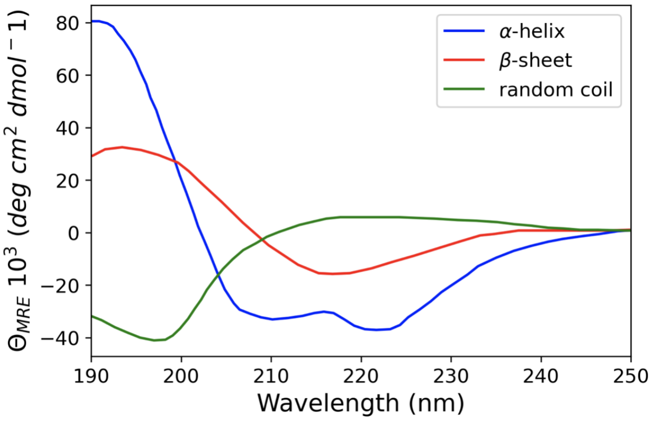


Figure B.2: **Representative CD spectra of proteins only containing one of the three main secondary structure elements.** Redrawn from *Wei et al.*[325]

When the dichroism is measured as a function of wavelength, a circular dichroism spectrum can be generated. The type of information one can obtain depends on the employed wavelength region, where the absorption bands of proteins and peptides lay in the ultraviolet (UV) regime. Spectra recorded in the far UV region between ∼ 180 − 250 nm, corresponding to peptide bond absorption, provide information about the amount of secondary structural elements (Figure B.2). If circularly polarized light is interacting with the protein sample, the electronic structure gives characteristic signals for different secondary structure elements due to different electronic excitation energies. For instance, proteins possessing mostly α-helices produce minima at 208 nm and 222 nm, whereby a maximum is measured at 193 nm.[326] Proteins consisting of antiparallel β-sheets possess a minimum at 218 nm and a maximum at 195 nm.[327] In contrast, random coil structures like in disordered proteins harbor a low ellipticity above 210 nm and produce a minimum between 190 -200 nm.[324,328] In the near UV region between $\sim 260-320$ nm one is able to obtain information about the tertiary structure as it reflects the environments of the aromatic amino acid side chains.

Next to single point measurements for the determination of structural features, circular dichorism can also measure changes along several spectra when recorded under changing conditions. For instance, the thermodynamics of folding and unfolding can be obtained for proteins, when spectra are measured as a function of temperature.[271,329] During the unfolding process, secondary structural components will eventually denature and the associated different absorption between folded states like α-helices and unordered structures like coil yield in a shift of recorded spectra. At characteristic wavelengths, the free energy ΔG of unfolding, the midpoint of the unfolding transition $\mathrm{T_M}$ and the associated constant of folding K can be calculated from spectra measured step wise over increasing temperatures.[271] Under the assumption that the unfolding path only includes the transition from the folded state F to the unfolded state U, K can be derived from their difference in concentration:

$$K = \frac{[\mathrm{F}]}{[\mathrm{U}]}. \tag{B.0.7}$$

ΔG is depended on K:

$$\Delta G = -\mathrm{RT} \ln K, \tag{B.0.8}$$

with R being the gas constant and T the absolute temperature in Kelvin. The midpoint of unfolding ($\mathrm{T_M}$) is defined as the fraction of folded protein to be one half ($\alpha = 0.5$), with α being calculated for different temperatures along the ramping experiment. It can be determined directly from recorded ellipticity values:

$$\alpha = \frac{(\theta T - \theta U)}{(\theta F - \theta U)}, \tag{B.0.9}$$

where θT is the ellipticity at any temperature, θU is the ellipticity at the unfolded state and θF at the folded state.[271] Plotting calculated α values, recorded at a specific wavelength, against increasing temperatures typically results in an inverse sigmoid function, where $\mathrm{T_M}$ can be accurately determined via fitting. Finally, K can also be expressed in terms of α:

$$K = \frac{\alpha}{(1 - \alpha)}. \tag{B.0.10}$$

Compared to NMR spectroscopy or X-ray crystallography, circular dichroism cannot resolve residual-specific structures in a three-dimensional fashion, but rather provide global information that support other complementary techniques. Its advantages are the very rapid measurements without the necessity of advanced sample preparation like the formation of crystals for X-ray crystallography. The only limitation is represented by the choice of buffer that is suitable for dissolution of samples, as chloride ions strongly absorb at wavelength below 195 nm, generating high noisy signals, and are therefore unsuitable for measurements.[323]

C | Quantifying enzyme activity

C.1 Anion exchange chromatography

High performance anion exchange chromatography (HPAEC) with pulsed amperometric detection (PAD) is a common ion chromatography method for the separation and determination of mono- and oligosaccharides. It is a physico-chemical separation process that is based on the distribution of sample molecules between a liquid mobile phase and a solid stationary phase. Anion exchange chromatography is based on a chromatography column (stationary phase) that consists of beads with positive charges on their surfaces (here: quarternary amines) (Figure C.1 **A**).[330] Not only charged but also various neutral mono- or oligosaccharides can be separated, because they are weak acids, meaning that under high pH conditions these molecules become ionized and can be separated in their anionic form, without the necessity for any derivatization. The mobile phase therefore consists mostly of a sodium hydroxide (NaOH) solution, creating a basic environment in order to convert target molecules to oxyanions. As a first step in HPAEC the column is equilibrated with NaOH, prior to the application of a sample containing different carbohydrate species. Positively charged compounds and those that do not become ionized due to the strong basic environment would not interact with the column and directly elute, whereas anionic species are differently retained by the stationary phase, depending on their specific affinity (Figure C.1 **A**). Empirical observations have been made, for instance that an increase of branching in a glycan or increase of number of mannose residues is also increasing the retention time.[331] In order to elute carbohydrates that are charged under neutral pH conditions (like Neu5Ac) from the positively charged stationary column, stronger eluents are required than for sugars that are uncharged at physiological/neutral pH. This can be accomplished by increasing concentrations of sodium acetate (NaOAc) in addition to the standard mobile phase, as acetate ions compete with negatively charged sugars for the binding to the quarternary amines, which results in elution of the sugars from the stationary phase. In the end, different carbohydrate species are eluted at different time points from the column, due to different retention times as a consequence of the interaction with the chromatography column. The subsequent detection of eluted molecules is based on the direct detection technique PAD under high pH conditions. The electroactive carbohydrates are oxidized at the surface of a gold electrode (anode) under a positive potential, generating an electrical current that is integrated over a set time period. The generated current is proportional to the concentration of the sugar molecule. The applied potential needs to be optimized to yield high responses for the analyte of interest but low responses for interfering molecules. Instead of only applying single-potential amperometry, the pulsed version is necessary because the interaction of carbohydrates fouls the electrode over time. Cleaning steps are necessary in-between, meaning the application of high and low potentials after a certain time period of measurement with the working potential.

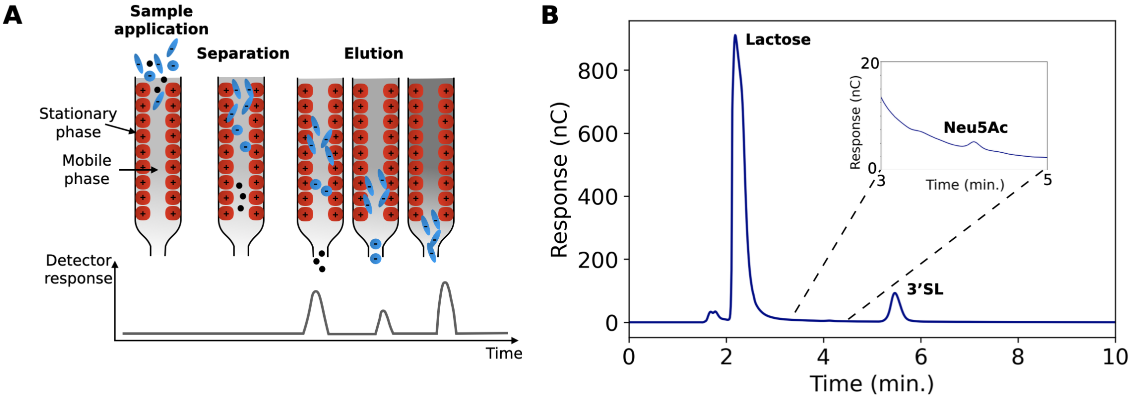


Figure C.1: **Principles of HPAEC-PAD. A** Schematic representation of an anion exchange chromatography, differently retaining the molecules in a sample. The stationary phase (red) is positively charged, whereas the mobile phase (shaded gray) is strongly alkaline, generating oxyanion molecules (blue/black) in the sample. Over the time course of elution, the composition of the mobile phase can vary, adding amounts of competing, negatively charged acetate ions to ensure that no compounds are retained on the column. Whenever carbohydrates elute from the column, the detector is giving a signal that is proportional to the concentration of eluted molecules. **B** Representative spectrum, showing the signal response (nC) upon elution of lactose, Neu5Ac and 3'SL from our measurements.

In our case, we focused on the separation of lactose from the single residues Neu5Ac and 3'SL, where they elute in exactly this order, with lactose being the most neutral carbohydrate and 3'SL the largest negatively charged one (Figure C.1 **B**). However, in order to guarantee unambiguous peak assignment, it is necessary to analyze suitable standards along the samples measured. For the determination of produced 3'SL, measurements of different 3'SL standard concentrations (5, 10, 20, 60, 100, 200, 500, 1000 pmol) were performed. The concentration of lactose was not of interest as it represented the educt in the enzyme reactions and was added in excess. The area under the curve was determined and plotted against the known concentrations. Finally, the 3'SL amount in enzyme samples could be determined.

C.2 Michaelis-Menten kinetics

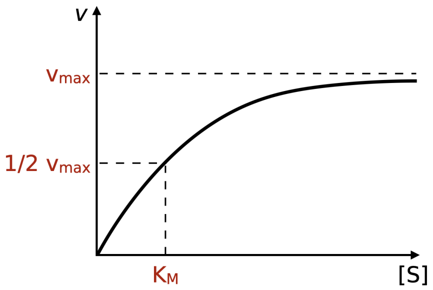


Figure C.2: **Michaelis-Menten saturation curve.** Kinetics of an enzyme reaction, showing the dependence of substrate concentration [S] on the reaction rate v. $\mathrm{K_M}$ is the substrate concentration where $\mathrm{v_{max}}$ is half.

In order to further calculate kinetic parameters from the enzymatic reactions and corresponding measured 3'SL amounts, we employed the famous Michaelis-Menten model.[332] It defines the speed at which a product is formed as the reaction rate and relates it to the concentration of substrate. Under the assumption that the enzymatic reaction is irreversible and the product is not used as a substrate, the kinetic reaction reads: $\mathrm{E} + \mathrm{S} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{ES} \xrightarrow{k_2} \mathrm{P} + \mathrm{E}$, with E being the enzyme, S the substrate, ES the enzyme-substrate complex and P the product. Relating this model to our enzymatic reaction at hand, the enzyme would be TconTS1, the substrate would be lactose and 3'SL would represent the product. Although it is theoretically possible that the produced 3'SL is used as a substrate, making the process reversible, the amount of fetuin-bound Sia in the reaction exceeds the amount of produced 3'SL by orders of magnitude (600 µmol Sia » 0.0001 µmol 3'SL) and therefore back reactions are assumed to be negligible in the short time scale of the experiment. The Michaelis-Menten model further assumes a quasi-steady state of the enzyme-substrate complex that:

$$\frac{\mathrm{d[ES]}}{\mathrm{d}t} = k_1[\mathrm{S}][\mathrm{E}] - [\mathrm{ES}](k_{-1} + k_2) = 0, \tag{C.2.1}$$

as well as a constant enzyme concentration:

$$[\mathrm{E}] + [\mathrm{ES}] = [\mathrm{E}]_0 = \mathrm{const.} \tag{C.2.2}$$

The former requirement is complied by only measuring in the initial phase of product formation, which is characterized by a linear growth of product over time when sufficient amount of substrate is supplied. The latter condition is inherently fulfilled by the experimental setup of the reaction, always using 50 ng of enzyme per tube. Rearrangement of equation C.2.1 yields:

$$0 = k_1[\mathrm{S}]([\mathrm{E}]_0 - [\mathrm{ES}]) - [\mathrm{ES}](k_{-1} + k_2), \tag{C.2.3}$$

$$k_1[\mathrm{S}][\mathrm{E}]_0 = k_1[\mathrm{S}][\mathrm{ES}] + [\mathrm{ES}](k_{-1} + k_2) \tag{C.2.4}$$

$$[\mathrm{S}][\mathrm{E}]_0 = [\mathrm{S}][\mathrm{ES}] + [\mathrm{ES}]\frac{k_{-1} + k_2}{k_1}. \tag{C.2.5}$$

We can define $\frac{k_{-1}+k_2}{k_1}$ as $\mathrm{K_M}$ and further substitute:

$$[\mathrm{S}][\mathrm{E}]_0 = [\mathrm{ES}]([\mathrm{S}] + \mathrm{K_M}), \tag{C.2.6}$$

$$\frac{[\mathrm{S}][\mathrm{E}]_0}{\mathrm{K_M} + [\mathrm{S}]} = [\mathrm{ES}], \tag{C.2.7}$$

The reaction rate describes the formation of product over time, depending on k_2:

$$v = \frac{d[\mathrm{P}]}{dt} = k_2[\mathrm{ES}]. \tag{C.2.8}$$

Substituting equation C.2.7 into the reaction rate yields:

$$v = k_2[\mathrm{E}]_0 \frac{[\mathrm{S}]}{\mathrm{K_M} + [\mathrm{S}]} = \frac{\mathrm{v_{max}}[\mathrm{S}]}{\mathrm{K_M} + [\mathrm{S}]}, \tag{C.2.9}$$

defining $k_2[\mathrm{E}]_0$ as $\mathrm{v_{max}}$. The reaction rate is therefore dependent on the maximum rate achieved by the system, $\mathrm{v_{max}}$, on the residual substrate concentration and on $\mathrm{K_M}$. The constant $\mathrm{K_M}$ is also referred to as the Michaelis constant and is equal to the [S] at which $\mathrm{v_{max}}$ is half-maximum (Figure C.2). It is recognized as a measure for the substrate affinity to the enzmye, where smaller values of $\mathrm{K_M}$ point to higher affinities because less substrate is required to achieve a maximum conversion rate. In contrast to $\mathrm{v_{max}}$, the Michaelis constant is not altered by purity or concentration of enzyme, but of course dependent on the identity of enzyme and substrate. k_2 is also referred to as the turnover number k_{cat}, reflecting the maximum number of substrate molecules converted per enzyme and per second.

D | Homology modeling of TconTS1

Note: *Parts of this chapter are taken from the publication: Jana Rosenau*, Isabell Louise Grothaus*, Yikun Yang, Nilima Dinesh Kumar, Lucio Colombi Ciacchi, Sørge Kelm, Mario Waespy, N-glycosylation modulates enzymatic activity of Trypanosoma congolense trans-sialidase, Journal of Biological Chemistry, 298:102403, 2022.*[245]

Due to missing experimental structures of TS from *T. congolense*, the atomistic structure of TconTS1 was derived using the I-TASSER web server for protein structure and function predictions.[198,273]. The engineered SNAP-Strep was included for consistency and better comparison with experimental data. The threading algorithm mainly employed TranSA (PDB entry: 2agsA, 2A75) as well as TcruTS (PDB entry: 1ms9) as templates. Validation of the homology modeled TconTS1 was performed by an amino acid sequence alignment of recombinant TconTS1 with TranSA (PDB entry 2ags) and TcruTS (PDB entry: 1ms9) revealing that 10 out of 14 amino acids predicted to be important for enzymatic activities are conserved among all compared models (Figure D.1 **A**). Furthermore, 2 of the remaining 4 sites are conserved between TconTS1 and TcruTS (Y211, P379) and only the remaining 2 are not conserved in TconTS1 (A325, Y408). It needs to be noted that especially Y408, part of the lactose holder pair in the binding site, is replaced by tryptophan in TranSA and TcruTS and therefore both amino acids resemble each other due to their hydrophobic character. Coloring of the atomistic structure of TconTS1 by the amino acid sequence alignment from Figure D.1 **A** gives the impression that most conserved residues are located in ß-sheet or α-helix regions (Figure D.1 **B**). Amino acids of loop regions seem to be less conserved, probably also being less important for the overall structural folding and function of the enzyme. Structural alignment of TconTS1 with TranSA and TcruTS reveals a high similarity of all models with respect to the secondary and tertiary structure (Figure D.1 **C**). Only the N-terminal part of TconTS1 is longer compared to TranSA and TcruTS and therefore cannot be aligned. Independent prediction of the secondary structure by I-Tasser as well as the inherent thermal mobility of each residue of TconTS1 are akin to that of TranSA and TcruTS (Figure D.1 **D**). This is because ß-sheets are dominant in the catalytic and lectin domain, whereas an α-helix is connecting both domains.

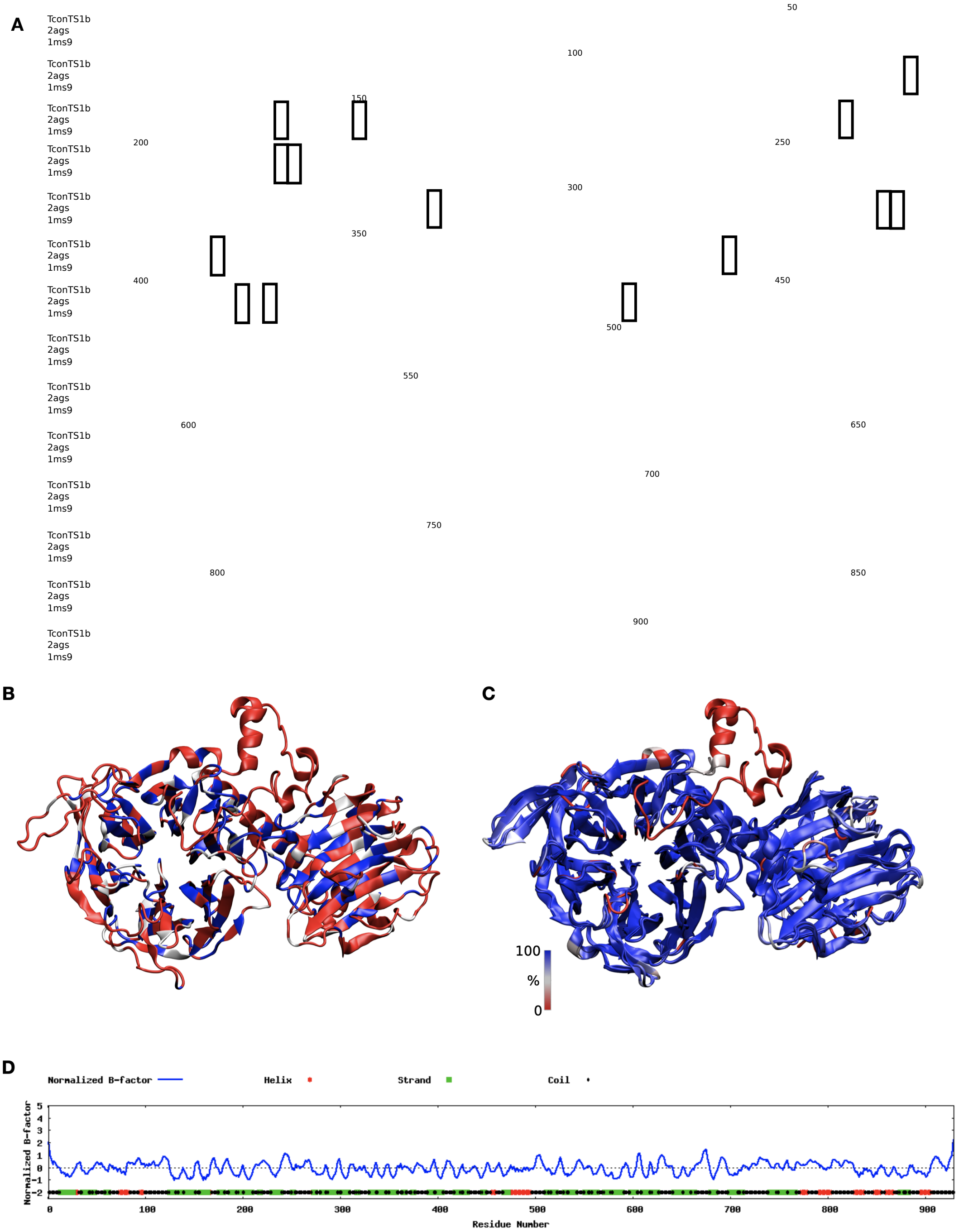


Figure D.1: **Validation of structural model.** **A** Amino acid sequence alignment of recombinant TconTS1, TranSA (PDB entry: 2ags) and TcruTS (PDB entry: 1ms9) by ClustalW using the bioinformatics analysis tool MultiSeq implemented in VMD.[333] Fully conserved amino acids are depicted in blue, partially conserved in white and not conserved in red. Residues of the catalytic domain, which are considered to be important for enzymatic activity, are surrounded by a black box.[252,253] Residue numbering is in correspondence with the native TconTS1 sequence. **B** 3D structure of TconTS1 in a cartoon style, where coloring of each amino acid corresponds to **A**. The C-terminal SNAP-Strep-Tag is not shown for simplicity. **C** Structural alignment of TconTS1, TranSA (PDB entry: 2ags) and TcruTS (PDB entry: 1ms9) by VMD represented in cartoon style. Coloring is in accordance with the Q factor of each residue, where Q is a metric for structural homology implemented in VMD. Blue corresponds to 100 % structural identity and a color shift over white to red, to lower and lower identities. **D** Plot of the normalized B-factor, representing the inherent thermal mobility of each residue with indication of predicted secondary structural elements, generated by I-Tasser.[198,273]

Despite an amino acids sequence identity of only 37/38 % between recombinant TconTS1 (with SNAP-Strep Tag) and TranSA/TcruTS, the I-Tasser homology model validation suggests a similar secondary structure of TconTS1 compared to other TS from different species and therefore predicts a similar tertiary model. Conservation of almost all catalytically involved amino acids in TconTS1 further supports the idea of structural similarity to the other TS. The I-Tasser confidence score of the recombinant TconTS1 model was given with -2.99 (range -5 to 2), where a higher value signifies a higher confidence. It is, however, necessary to note that the artificial SNAP-Strep domain is included in this assessment, and an analysis of only the native TconTS1 results in a confidence score of -0.63, stating a much higher reliability. A confidence score of above -1.5 means that more than 90% of the predictions are correct and therefore our TconTS1 models is considered to be predicted with an overall correct fold.[198].

Bibliography

[1] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, 1953.

[2] Rosalind E. Franklin and R. G. Gosling. Molecular Configuration in Sodium Thymonucleate. *Nature*, 171(4356):740–741, 1953.

[3] M. H. F. Wilkins, A. R. Stokes, and H. R. Wilson. Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids. *Nature*, 171(4356):738–740, 1953.

[4] Matthew Cobb. Sexism in science: did watson and crick really steal rosalind franklin's data? *The Guardian*, June 2015. [Accessed on 2022-10-05].

[5] Hans-Joachim Gabius and Jürgen Roth. An introduction to the sugar code. *Histochemistry and Cell Biology*, 147(2):111–117, 2017.

[6] Nathan Sharon and Halina Lis. History of lectins: from hemagglutinins to biological recognition molecules. *Glycobiology*, 14(11):53R–62R, 06 2004.

[7] Roger A. Laine. Invited Commentary: A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 × 1012 structures for a reducing hexasaccharide: the Isomer Barrier to development of single-method saccharide sequencing or synthesis systems. *Glycobiology*, 4(6):759–767, 1994.

[8] Roger A Laine. The information-storing potential of the sugar code. In *Glycosciences: Status and Perspectives*, pages 1–14. 1996.

[9] Peter H. Seeberger. Monosaccharide diversity. In Ajit Varki, Richard D. Cummings, Pamela Esko, Jeffrey D.and Stanley, Gerald W. Hart, Markus Aebi, Debra Mohnen, Taroh Kinoshita, Nicolle H. Packer, James H. Prestegard, Ronald L. Schnaar, and Peter H. Seeberger, editors, *Essentials of Glycobiology*, chapter 2. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 4th edition, 2022.

[10] Amanda L. Lewis, Xi Chen, Ronald L. Schnaar, and Ajit Varki. Sialic acids and other nonulosonic acids. In Ajit Varki, Richard D. Cummings, Pamela Esko, Jeffrey D.and Stanley, Gerald W. Hart, Markus Aebi, Debra Mohnen, Taroh Kinoshita, Nicolle H. Packer, James H. Prestegard, Ronald L. Schnaar, and Peter H. Seeberger, editors, *Essentials of Glycobiology*, chapter 15. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 4th edition, 2022.

[11] P. J. Winterburn and C. F. Phelps. The Significance of Glycosylated Proteins. *Nature*, 236(5343):147–151, 1972.

[12] J B Sumner and S F Howell. Identification of Hemagglutinin of Jack Bean with Concanavalin A. *Journal of bacteriology*, 32(2):227–37, 1936.

[13] Hans-Joachim Gabius, Sabine André, Jesús Jiménez-Barbero, Antonio Romero, and Dolores Solís. From lectin structure to functional glycomics: principles of the sugar code. *Trends Biochem. Sci.*, 36(6):298–313, 2011.

[14] Hans-Joachim Gabius, Maré Cudic, Tammo Diercks, Herbert Kaltner, Jürgen Kopitz, Kevin H. Mayo, Paul V. Murphy, Stefan Oscarson, René Roy, Andreas Schedlbauer, Stefan Toegel, and Antonio Romero. What is the Sugar Code? *ChemBioChem*, 23(13):e202100327, 2022.

[15] A Butlerow. Bildung einer zuckerartigen Substanz durch Synthese. *Annalen der Chemie und Pharmacie*, 120(3):295–298, 1861.

[16] J Hirabayashi. On the Origin of Elementary Hexoses. *The Quarterly Review of Biology*, 71(3):365–380, 1996.

[17] Gideon J Davies, Tracey M Gloster, and Bernard Henrissat. Recent structural insights into the expanding world of carbohydrate-active enzymes. *Current Opinion in Structural Biology*, 15(6):637–645, 2005. Catalysis and regulation/Proteins.

[18] Robert S. Haltiwanger. Symbol Nomenclature for Glycans (SNFG). *Glycobiology*, 26(3):217–217, 2016.

[19] Karen J. Colley, Ajit Varki, Robert S. Haltiwanger, and Taroh Kinoshita. Cellular organization of glycosylation. In Ajit Varki, Richard D. Cummings, Pamela Esko, Jeffrey D.and Stanley, Gerald W. Hart, Markus Aebi, Debra Mohnen, Taroh Kinoshita, Nicolle H. Packer, James H. Prestegard, Ronald L. Schnaar, and Peter H. Seeberger, editors, *Essentials of Glycobiology*, chapter 4. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 4th edition, 2022.

[20] Leonhard Möckl. The Emerging Role of the Mammalian Glycocalyx in Functional Membrane Organization and Immune System Regulation. *Frontiers in Cell and Developmental Biology*, 8:253, 2020.

[21] Ryan A. Flynn, Kayvon Pedram, Stacy A. Malaker, Pedro J. Batista, Benjamin A.H. Smith, Alex G. Johnson, Benson M. George, Karim Majzoub, Peter W. Villalta, Jan E. Carette, and Carolyn R. Bertozzi. Small RNAs are modified with N-glycans and displayed on the surface of living cells. *Cell*, 184(12):3109–3124.e22, 2021.

[22] Sietze Reitsma, Dick W. Slaaf, Hans Vink, Marc A. M. J. van Zandvoort, and Mirjam G. A. oude Egbrink. The endothelial glycocalyx: composition, functions, and visualization. *Pflugers Archiv*, 454(3):345–359, 2007.

[23] Hans Vink and Brian R. Duling. Identification of Distinct Luminal Domains for Macromolecules, Erythrocytes, and Leukocytes Within Mammalian Capillaries. *Circulation Research*, 79(3):581–589, 1996.

[24] Paul M. A. van Haaren, Ed VanBavel, Hans Vink, and Jos A. E. Spaan. Localization of the permeability barrier to solutes in isolated arteries by confocal microscopy. *American Journal of Physiology-Heart and Circulatory Physiology*, 285(6):H2848–H2856, 2003.

[25] R.T.A. Megens, S. Reitsma, P.H.M. Schiffers, R.H.P. Hilgers, J.G.R. De Mey, D.W. Slaaf, M.G.A. oude Egbrink, and M.A.M.J. van Zandvoort. Two-Photon Microscopy of Vital Murine Elastic and Muscular Arteries. *Journal of Vascular Research*, 44(2):87–98, 2007.

[26] Ronald L. Schnaar, Roger Sandhoff, Michael Tiemeyer, and Taroh Kinoshita. Glycosphingolipids. In Ajit Varki, Richard D. Cummings, Pamela Esko, Jeffrey D.and Stanley, Gerald W. Hart, Markus Aebi, Debra Mohnen, Taroh Kinoshita, Nicolle H. Packer, James H. Prestegard, Ronald L. Schnaar, and Peter H. Seeberger, editors, *Essentials of Glycobiology*, chapter 11. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 4th edition, 2022.

[27] Giovanni D'Angelo, Serena Capasso, Lucia Sticco, and Domenico Russo. Glycosphingolipids: synthesis and functions. *FEBS Journal*, 280(24):6338–6353, 2013.

[28] Manish Sud, Eoin Fahy, Dawn Cotter, Alex Brown, Edward A Dennis, Christopher K Glass, Alfred H Merrill, Robert C Murphy, Christian R H Raetz, David W Russell, and Shankar Subramaniam. LMSD: LIPID MAPS structure database. *Nucleic acids research*, 35(Database issue):D527–32, 2006.

[29] William T Norton. *Myelin.* Springer Nature, 1977.

[30] Donald M. Marcus and Louise E. Cass. Glycosphingolipids with Lewis Blood Group Activity: Uptake by Human Erythrocytes. *Science*, 164(3879):553–555, 1969.

[31] N Kojima and S Hakomori. Specific Interaction between Gangliotriaosylceramide (Gg3) and Sialosyllactosylceramide (GM3) as a Basis for Specific Cellular Recognition between Lymphoma and Melanoma Cells*. *Journal of Biological Chemistry*, 264(34):20159–20162, 1989.

[32] N. Kojima and S. Hakomori. Cell adhesion, spreading, and motility of GM3-expressing cells based on glycolipid-glycolipid interaction. *Journal of Biological Chemistry*, 266(26):17552–17558, 1991.

[33] N Kojima, M Shiota, Y Sadahira, K Handa, and S Hakomori. Cell adhesion in a dynamic flow system as compared to static system. Glycosphingolipid-glycosphingolipid interaction in the dynamic system predominates over lectin- or integrin-based mechanisms in adhesion of B16 melanoma cells to non-activated endothelial cells. *Journal of Biological Chemistry*, 267(24):17264–17270, 1992.

[34] William J. Cook and Charles E. Bugg. Calcium-carbohydrate bridges composed of uncharged sugars. Structure of a hydrated calcium bromide complex of α-fucose. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 389(3):428–435, 1975.

[35] Gerald W. Hart. Glycosylation. *Current Opinion in Cell Biology*, 4(6):1017–1023, 1992.

[36] Catherine L.R. Merry, Ulf Lindahl, John Couchman, and Jeffrey D. Esko. Proteoglycans and sulfated glycosaminoglycans. In Ajit Varki, Richard D. Cummings, Pamela Esko, Jeffrey D.and Stanley, Gerald W. Hart, Markus Aebi, Debra Mohnen, Taroh Kinoshita, Nicolle H. Packer, James H. Prestegard, Ronald L. Schnaar, and Peter H. Seeberger, editors, *Essentials of Glycobiology*, chapter 17. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 4th edition, 2022.

[37] Jeffrey M. Mattson, Raphaël Turcotte, and Yanhang Zhang. Glycosaminoglycans contribute to extracellular matrix fiber recruitment and arterial wall mechanics. *Biomechanics and Modeling in Mechanobiology*, 16(1):213–225, 2017.

[38] Xingyu Chen, Dongning Chen, Ehsan Ban, Kimani C. Toussaint, Paul A. Janmey, Rebecca G. Wells, and Vivek B. Shenoy. Glycosaminoglycans modulate long-range mechanical communication between cells in collagen networks. *Proceedings of the National Academy of Sciences*, 119(15):e2116718119, 2022.

[39] Linda T. Senbanjo and Meenakshi A. Chellaiah. CD44: A Multifunctional Cell Surface Adhesion Receptor Is a Regulator of Progression and Metastasis of Cancer Cells. *Frontiers in Cell and Developmental Biology*, 5:18, 2017.

[40] Lilly Y W Bourguignon, Marisa Shiina, and Jian-Jian Li. Hyaluronan-CD44 interaction promotes oncogenic signaling, microRNA functions, chemoresistance, and radiation resistance in cancer stem cells leading to tumor progression. *Advances in cancer research*, 123:255–75, 2014.

[41] Robert G. Spiro. Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology*, 12(4):43R–56R, 2002.

[42] R Apweiler, H Hermjakob, and N Sharon. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochimica et biophysica acta*, 1473(1):4 8, 12 1999.

[43] Colin Reily, Tyler J. Stewart, Matthew B. Renfrow, and Jan Novak. Glycosylation in health and disease. *Nat. Rev. Nephrol.*, 15(6):346–366, 2019.

[44] Pamela Stanley, Kelley W. Moremen, Nathan E. Lewis, Naoyuki Taniguchi, and Markus Aebi. N-glycans. In Ajit Varki, Richard D. Cummings, Pamela Esko, Jeffrey D.and Stanley, Gerald W. Hart, Markus Aebi, Debra Mohnen, Taroh Kinoshita, Nicolle H. Packer, James H. Prestegard, Ronald L. Schnaar, and Peter H. Seeberger, editors, *Essentials of Glycobiology*, chapter 9. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 4th edition, 2022.

[45] Sarah R. Hanson, Elizabeth K. Culyba, Tsui-Ling Hsu, Chi-Huey Wong, Jeffery W. Kelly, and Evan T. Powers. The core trisaccharide of an N-linked glycoprotein intrinsically accelerates folding and enhances stability. *Proceedings of the National Academy of Sciences*, 106(9):3131–3136, 2009.

[46] Ibh Wilson, K Paschinger, and D Rendic. Glycosylation of model and 'lower'organisms. In *The sugar code. Fundamentals of glycosciences*, pages 139–154. Wiley-VCH: Weinheim, Germany, 2009.

[47] Kelley W. Moremen, Michael Tiemeyer, and Alison V. Nairn. Vertebrate protein glycosylation: diversity, synthesis and function. *Nature Reviews Molecular Cell Biology*, 13(7):448–462, 2012.

[48] Jingzhong Guo, Huiping Tu, and Fouad Atouf. Measurement of macro- and micro-heterogeneity of glycosylation in biopharmaceuticals: a pharmacopeia perspective. *Future Drug Discovery*, 2(4):FDD48, 2020.

[49] Ajit Varki, Richard D. Cummings, Jeffrey D. Esko, Hudson H. Freeze, Pamela Stanley, Jamey D. Marth, Carolyn R. Bertozzi, Gerald W. Hart, and Marilynn E. Etzler. Symbol nomenclature for glycan representation. *Proteomics*, 9(24):5398–5399, 2009.

[50] Kai Cheng, Yusen Zhou, and Sriram Neelamegham. DrawGlycan-SNFG: a robust tool to render glycans and glycopeptides with fragmentation information. *Glycobiology*, 27(3):200–205, 2017.

[51] Danielle Skropeta. The effect of individual N-glycans on enzyme activity. *Bioorg. Med. Chem.*, 17(7):2645 2653, 04 2009.

[52] Stephanie Grünewald, Gert Matthijs, and Jaak Jaeken. Congenital Disorders of Glycosylation: A Review. *Pediatric Research*, 52(5):618–624, 2002.

[53] Edwin H. Eylar. On the biological role of glycoproteins. *Journal of Theoretical Biology*, 10(1):89–113, 1966.

[54] Eugenia Wulff-Fuentes, Rex R. Berendt, Logan Massman, Laura Danner, Florian Malard, Jeet Vora, Robel Kahsay, and Stephanie Olivier-Van Stichelen. The human O-GlcNAcome database and meta-analysis. *Scientific Data*, 8(1):25, 2021.

[55] Meryem Bektas and David S. Rubenstein. The role of intracellular protein O-glycosylation in cell adhesion and disease. *Journal of Biomedical Research*, 25(4):227–236, 2011.

[56] Hudson H. Freeze, Michael Boyce, Natasha E. Zachara, Gerald W. Hart, and Ronald L. Schnaar. Glycosylation precursors. In Ajit Varki, Richard D. Cummings, Pamela Esko, Jeffrey D.and Stanley, Gerald W. Hart, Markus Aebi, Debra Mohnen, Taroh Kinoshita, Nicolle H. Packer, James H. Prestegard, Ronald L. Schnaar, and Peter H. Seeberger, editors, *Essentials of Glycobiology*, chapter 5. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 4th edition, 2022.

[57] Satu Mikkola. Nucleotide Sugars in Chemistry and Biology. *Molecules*, 25(23):5755, 2020.

[58] Eranthie Weerapana and Barbara Imperiali. Asparagine-linked protein glycosylation: from eukaryotic to prokaryotic systems. *Glycobiology*, 16(6):91R–101R, 2006.

[59] Shiteshu Shrimal, Natalia A Cherepanova, and Reid Gilmore. Cotranslational and posttranslocational N-glycosylation of proteins in the endoplasmic reticulum. *Seminars in Cell & Developmental Biology*, 41:71 78, 2015-05.

[60] Megan K. Barker and David R. Rose. Specificity of Processing α-Glucosidase I Is Guided by the Substrate Conformation. *Journal of Biological Chemistry*, 288(19):13563–13574, 2013.

[61] E. Sergio Trombetta, Jan Fredrik Simons, and Ari Helenius. Endoplasmic Reticulum Glucosidase II Is Composed of a Catalytic Subunit, Conserved from Yeast to Mammals, and a Tightly Bound Noncatalytic HDEL-containing Subunit. *Journal of Biological Chemistry*, 271(44):27509–27516, 1996.

[62] Steven W. Mast and Kelley W. Moremen. Family 47 α-Mannosidases in N-Glycan Processing. *Methods in Enzymology*, 415(Annu. Rev. Biochem.732004):31–46, 2006.

[63] Jean M.H. van den Elsen, Douglas A. Kuntz, and David R. Rose. Structure of Golgi α-mannosidase II: a target for inhibition of growth and metastasis of cancer cells. *The EMBO Journal*, 20(12):3008–3017, 2001.

[64] Richard Strasser, Johannes Stadlmann, Barbara Svoboda, Friedrich Altmann, Josef Glössl, and Lukas Mach. Molecular basis of N-acetylglucosaminyltransferase I deficiency in Arabidopsis thaliana plants lacking complex N-glycans. *The Biochemical journal*, 387(Pt 2):385–91, 2004.

[65] Carlito B. Lebrilla, Jian Liu, Göran Widmalm, and James H. Prestegard. Oligosaccharides and polysaccharides. In Ajit Varki, Richard D. Cummings, Pamela Esko, Jeffrey D.and Stanley, Gerald W. Hart, Markus Aebi, Debra Mohnen, Taroh Kinoshita, Nicolle H. Packer, James H. Prestegard, Ronald L. Schnaar, and Peter H. Seeberger, editors, *Essentials of Glycobiology*, chapter 3. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 4th edition, 2022.

[66] M Gilleron, H C Siebert, H Kaltner, C W von der Lieth, T Kozár, K M Halkes, E Y Korchagina, N V Bovin, H J Gabius, and J F Vliegenthart. Conformer selection and differential restriction of ligand mobility by a plant lectin–conformational behaviour of Galbeta1-3GlcNAcbeta1-R, Galbeta1-3GalNAcbeta1-R and Galbeta1-2Galbeta1-R' in the free state and complexed with galactoside-specific mistletoe lectin as revealed by random-walk and conformational-clustering molecular-mechanics. *European journal of biochemistry*, 252(3):416–27, 1998.

[67] Emil Fischer. Einfluss der configuration auf die wirkung der enzyme. *Berichte der deutschen chemischen Gesellschaft*, 27(3):2985–2993, 1894.

[68] M Hricovíni, R N Shah, and J P Carver. Detection of internal motions in oligosaccharides by 1H relaxation measurements at different magnetic fields. *Biochemistry*, 31(41):10018–23, 1992.

[69] J P Carver, D Mandel, S W Michnick, A Imberty, and J W Brady. Conformational analysis o f oligosaccharides, reconciliation of theory with experiment. In *Computer Modeling of Carbohydrate Molecules*, ACS Symposium Series, chapter 16, pages 266–280. American Chemical Society, 1990.

[70] J. P. Carver. Oligosaccharides: How can flexible molecules act as signals? *Pure and Applied Chemistry*, 65(4):763–770, 1993.

[71] Barry J Hardy. The glycosidic linkage flexibility and time-scale similarity hypotheses. *Journal of Molecular Structure: Theochem*, 395-396:187–200, 1997.

[72] D Cremer and J A Pople. General definition of ring puckering coordinates. *J. Am. Chem. Soc.*, 97(6):1354–1358, 1975.

[73] Marcos D Battistel, Hugo F Azurmendi, and Darón I Freedberg. NMR in Glycoscience and Glycotechnology. *New Dev. NMR*, pages 1–19, 2017.

[74] Karl N. Kirschner and Robert J. Woods. Solvent interactions determine carbohydrate conformation. *Proc. Natl. Acad. Sci.*, 98(19):10541–10545, 2001.

[75] J. C. P. Schwarz. Rules for conformation nomenclature for five- and six-membered rings in monosaccharides and their derivatives. *J. Chem. Soc., Chem. Commun.*, 0(14):505–508, 1973.

[76] Norman L Allinger, Mary Ann Miller, Frederic A Van Catledge, and Jerry A Hirsch. Conformational analysis. LVII. The calculation of the conformational structures of hydrocarbons by the Westheimer-Hendrickson-Wiberg method. *Journal of the American Chemical Society*, 89(17):4345–4357, 1967.

[77] Heather B. Mayes, Linda J. Broadbelt, and Gregg T. Beckham. How Sugars Pucker: Electronic Structure Calculations Map the Kinetic Landscape of Five Biologically Paramount Monosaccharides and Their Implications for Enzymatic Catalysis. *J. Am. Chem. Soc.*, 136(3):1008–1022, 2014.

[78] James H. Prestegard. A perspective on the PDB's impact on the field of glycobiology. *J. Biol. Chem.*, 296:100556, 2021.

[79] Peter D. Kwong, Richard Wyatt, Elizabeth Desjardins, James Robinson, Jeffrey S. Culp, Brian D. Hellmig, Raymond W. Sweet, Joseph Sodroski, and Wayne A. Hendrickson. Probability Analysis of Variational Crystallization and Its Application to gp120, The Exterior Envelope Glycoprotein of Type 1 Human Immunodeficiency Virus (HIV-1)*. *Journal of Biological Chemistry*, 274(7):4115–4123, 1999.

[80] Protein Data Bank. Pdb data distribution by experimental method and molecular type, 2022. https://www.rcsb.org/stats/summary [Accessed on 2022-11-07].

[81] Mark R. Wormald, Andrei J. Petrescu, Ya-Lan Pao, Ann Glithero, Tim Elliott, and Raymond A. Dwek. Conformational Studies of Oligosaccharides and Glycopeptides: Complementarity of NMR, X-ray Crystallography, and Molecular Modelling. *Chem. Rev.*, 102(2):371–386, 2002.

[82] Suyong Re, Shigehisa Watabe, Wataru Nishima, Eiro Muneyuki, Yoshiki Yamaguchi, Alexander D. MacKerell, and Yuji Sugita. Characterization of Conformational Ensembles of Protonated N-glycans in the Gas-Phase. *Sci. Rep.*, 8(1):1644, 2018.

[83] Robert J Woods. Predicting the Structures of Glycans, Glycoproteins, and Their Complexes. *Chem. Rev.*, 118(17):8005–8024, 2018.

[84] Anita Plazinska and Wojciech Plazinski. Comparison of Carbohydrate Force Fields in Molecular Dynamics Simulations of Protein–Carbohydrate Complexes. *J. Chem. Theory Comput.*, 17(4):2575–2585, 2021.

[85] Ryan D. Lazar, Farideh B. Akher, Neil Ravenscroft, and Michelle M. Kuttel. Carbohydrate Force Fields: The Role of Small Partial Atomic Charges in Preventing Conformational Collapse. *J. Chem. Theory Comput.*, 18(2):1156–1172, 2022.

[86] Suyong Re, Wataru Nishima, Naoyuki Miyashita, and Yuji Sugita. Conformational flexibility of N-glycans in solution studied by REMD simulations. *Biophys. Rev.*, 4(3):179–187, 2012.

[87] Raimondas Galvelis, Suyong Re, and Yuji Sugita. Enhanced Conformational Sampling of N-Glycans in Solution with Replica State Exchange Metadynamics. *J. Chem. Theory Comput.*, 13(5):1934–1942, 2017.

[88] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, 7(1):95–99, 1963.

[89] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1-2):141–151, 1999.

[90] Lingle Wang, Richard A Friesner, and B J Berne. Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). *J. Phys. Chem. B*, 115(30):9431–8, 2011.

[91] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.*, 100(2):020603, 2007.

[92] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23(2):187–199, 1977.

[93] Suyong Re, Naoyuki Miyashita, Yoshiki Yamaguchi, and Yuji Sugita. Structural diversity and changes in conformational equilibria of biantennary complex-type N-glycans in water revealed by replica-exchange molecular dynamics simulation. *Biophys. J.*, 101(10):L44–6, 2011.

[94] Mingjun Yang, Jing Huang, and Alexander D MacKerell. Enhanced conformational sampling using replica exchange with concurrent solute scaling and hamiltonian biasing realized in one dimension. *J. Chem. Theory Comput.*, 11(6):2855–67, 2015.

[95] Hiraku Oshima, Suyong Re, and Yuji Sugita. Replica-Exchange Umbrella Sampling Combined with Gaussian Accelerated Molecular Dynamics for Free-Energy Calculation of Biomolecules. *J. Chem. Theory Comput.*, 15(10):5199–5208, 2019.

[96] Wataru Nishima, Naoyuki Miyashita, Yoshiki Yamaguchi, Yuji Sugita, and Suyong Re. Effect of bisecting GlcNAc and core fucosylation on conformational properties of biantennary complex-type N-glycans in solution. *J. Phys. Chem. B*, 116(29):8504–12, 2012.

[97] Dhilon S. Patel, Robert Pendrill, Sairam S. Mallajosyula, Goran Widmalm, and Alexander D. MacKerell. Conformational Properties of α- or β-(1→6)-Linked Oligosaccharides: Hamiltonian Replica Exchange MD Simulations and NMR Experiments. *J. Phys. Chem. B*, 118(11):2851–2871, 2014.

[98] Elisa Fadda and Robert J. Woods. Molecular simulations of carbohydrates and protein–carbohydrate interactions: motivation, issues and prospects. *Drug Discovery Today*, 15(15-16):596–609, 2010.

[99] Wojciech Plazinski and Anita Plazinska. Molecular dynamics simulations of hexopyranose ring distortion in different force fields. *Pure Appl. Chem.*, 89(9):1283–1294, 2017.

[100] Lorenzo Casalino, Zied Gaieb, Jory A Goldsmith, Christy K Hjorth, Abigail C Dommer, Aoife M Harbison, Carl A Fogarty, Emilia P Barros, Bryn C Taylor, Jason S McLellan, Elisa Fadda, and Rommie E Amaro. Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Cent. Sci.*, 6(10):1722–1734, 2020.

[101] Yasunori Watanabe, Joel D. Allen, Daniel Wrapp, Jason S. McLellan, and Max Crispin. Site-specific glycan analysis of the SARS-CoV-2 spike. *Science*, 369(6501):330–333, 2020.

[102] Yuqing Li, Dongqi Liu, Yating Wang, Wenquan Su, Gang Liu, and Weijie Dong. The Importance of Glycans of Viral and Host Proteins in Enveloped Virus Infection. *Frontiers in Immunology*, 12:638573, 2021.

[103] Sergio Schenkman, Man-Shiow Jiang, Gerald W. Hart, and Victor Nussenzweig. A novel cell surface trans-sialidase of trypanosoma cruzi generates a stage-specific epitope required for invasion of mammalian cells. *Cell*, 65(7):1117–1125, 1991.

[104] M Engstler, G Reuter, and R Schauer. The developmentally regulated trans-sialidase from Trypanosoma brucei sialylates the procyclic acidic repetitive protein. *Molecular and biochemical parasitology*, 61(1):1 13, 1993-09.

[105] L C Pontes de Carvalho, S Tomlinson, F Vandekerckhove, E J Bienen, A B Clarkson, M S Jiang, G W Hart, and V Nussenzweig. Characterization of a novel trans-sialidase of Trypanosoma brucei procyclic trypomastigotes and identification of procyclin as the main sialic acid acceptor. *The Journal of Experimental Medicine*, 177(2):465–474, 1993.

[106] Lori Peacock, Simon Cook, Vanessa Ferris, Mick Bailey, and Wendy Gibson. The life cycle of Trypanosoma (Nannomonas) congolense in the tsetse fly. *Parasites & Vectors*, 5(1):109, 2012.

[107] Keith Vickerman. Developmental cycles and biology of pathogenic trypanosomes. *British Medical Bulletin*, 41(2):105–114, 1985.

[108] Mao Ming, Marina Chuenkova, Eduardo Ortega-Barria, and Miercio E.A. Pereira. Mediation of Trypanosoma cruzi invasion by sialic acid on the host cell and trans-sialidase on the trypanosome. *Molecular and Biochemical Parasitology*, 59(2):243–252, 1993.

[109] Sergio Schenkman, Michael A.J. Ferguson, Norton Heise, Maria Lucia Cardoso de Almeida, Renato A. Mortara, and Nobuko Yoshida. Mucin-like glycoproteins linked to the membrane by glycosylphosphatidylinositol anchor are the major acceptors of sialic acid in a reaction catalyzed by trans-sialidase in metacyclic forms of Trypanosoma cruzi. *Molecular and Biochemical Parasitology*, 59(2):293–303, 1993.

[110] S Tomlinson, L C Pontes de Carvalho, F Vandekerckhove, and V Nussenzweig. Role of sialic acid in the resistance of Trypanosoma cruzi trypomastigotes to complement. *Journal of immunology (Baltimore, Md. : 1950)*, 153(7):3141–7, 1994.

[111] V L Pereira-Chioccola, A Acosta-Serrano, I Correia de Almeida, M A Ferguson, T Souto-Padron, M M Rodrigues, L R Travassos, and S Schenkman. Mucin-like molecules form a negatively charged coat that protects Trypanosoma cruzi trypomastigotes from killing by human anti-alpha-galactosyl antibodies. *Journal of cell science*, 113 (Pt 7)(7):1299–307, 2000.

[112] Kisaburo Nagamune, Alvaro Acosta-Serrano, Haruki Uemura, Reto Brun, Christina Kunz-Renggli, Yusuke Maeda, Michael A J Ferguson, and Taroh Kinoshita. Surface sialic acids taken from the host allow trypanosome survival in tsetse fly vectors. *The Journal of experimental medicine*, 199(10):1445 1450, 05 2004.

[113] M Engstler, R Schauer, and R Brun. Distribution of developmentally regulated trans-sialidases in the Kinetoplastida and characterization of a shed trans-sialidase activity from procyclic Trypanosoma congolense. *Acta tropica*, 59(2):117 129, 1995-05.

[114] Fabien Guegan, Nicolas Plazolles, Théo Baltz, and Virginie Coustou. Erythrophagocytosis of desialylated red blood cells is responsible for anaemia during Trypanosoma vivax infection. *Cellular Microbiology*, 15(8):1285–1303, 2013.

[115] Brent M Swallow. Impacts of trypanosomiasis in African agriculture, 11 1999.

[116] A Buschiazzo, G A Tavares, O Campetella, S Spinelli, M L Cremona, G París, M F Amaya, A C Frasch, and P M Alzari. Structural basis of sialyltransferase activity in trypanosomal sialidases. *The EMBO journal*, 19(1):16 24, 01 2000.

[117] Alejandro Buschiazzo, María F Amaya, María L Cremona, Alberto C Frasch, and Pedro M Alzari. The crystal structure and mode of action of trans-sialidase, a key enzyme in Trypanosoma cruzi pathogenesis. *Molecular cell*, 10(4):757 768, 2002-10.

[118] Marı'a Fernanda Amaya, Andrew G Watts, Iben Damager, Annemarie Wehenkel, Tong Nguyen, Alejandro Buschiazzo, Gastón Paris, A.Carlos Frasch, Stephen G Withers, and Pedro M Alzari. Structural Insights into the Catalytic Mechanism of Trypanosoma cruzi trans-Sialidase. *Structure*, 12(5):775–784, 2004.

[119] Gustavo Pierdominici-Sottile, Nicole A Horenstein, and Adrian E Roitberg. Free energy study of the catalytic mechanism of Trypanosoma cruzi trans-sialidase. From the Michaelis complex to the covalent intermediate. *Biochemistry*, 50(46):10150–8, 2011.

[120] S Schenkman, L Pontes de Carvalho, and V Nussenzweig. Trypanosoma cruzi trans-sialidase and neuraminidase activities can be mediated by the same enzymes. *The Journal of experimental medicine*, 175(2):567 575, 02 1992.

[121] Markus Engstler, Gerd Reuter, and Roland Schauer. Purification and characterization of a novel sialidase found in procyclic culture forms of Trypanosoma brucei. *Molecular and Biochemical Parasitology*, 54(1):21–30, 1992.

[122] Angela Savage, Rudolf Geyer, Stephan Stirm, Erwin Reinwald, and Hans-Jörg Risse. Structural studies on the major oligosaccharides in a variant surface glycoprotein of Trypanosoma congolense. *Molecular and Biochemical Parasitology*, 11:309–328, 1984.

[123] Susanne E. Zamze, E. Wrenn Wooten, David A. Ashford, Michael A. J. Ferguson, Raymond A. Dwek, and Thomas W. Rademacher. Characterisation of the asparagine-linked oligosaccharides from Trypanosoma brucei type-I variant surface glycoproteins. *European Journal of Biochemistry*, 187(3):657–663, 1990.

[124] S.E. Zamze, D.A. Ashford, E.W. Wooten, T.W. Rademacher, and R.A. Dwek. Structural characterization of the asparagine-linked oligosaccharides from Trypanosoma brucei type II and type III variant surface glycoproteins. *Journal of Biological Chemistry*, 266(30):20244–20261, 1991.

[125] Achim Treumann, Nicole Zitzmann, Andreas Hülsmeier, Alan R Prescott, Andrew Almond, John Sheehan, and Michael A J Ferguson. Structural characterisation of two forms of procyclic acidic repetitive protein expressed by procyclic forms of Trypanosoma brucei. *Journal of Molecular Biology*, 269(4):529–547, 1997.

[126] Angela Mehlert, Nicole Zitzmann, Julia M. Richardson, Achim Treumann, and Michael A.J. Ferguson. The glycosylation of the variant surface glycoproteins and procyclic acidic repetitive proteins of Trypanosoma brucei. *Molecular and Biochemical Parasitology*, 91(1):145–152, 1998.

[127] Silvia Utz, Isabel Roditi, Christina Kunz Renggli, Igor C. Almeida, Alvaro Acosta-Serrano, and Peter Bütikofer. Trypanosoma congolense Procyclins: Unmasking Cryptic Major Surface Glycoproteins in Procyclic Forms. *Eukaryotic Cell*, 5(8):1430–1440, 2006.

[128] James A. Atwood, Todd Minning, Fernanda Ludolf, Arthur Nuccio, Daniel B. Weatherly, Gerardo Alvarez-Manilla, Rick Tarleton, and Ron Orlando. Glycoproteomics of Trypanosoma cruzi Trypomastigotes Using Subcellular Fractionation, Lectin Affinity, and Stable Isotope Labeling. *Journal of Proteome Research*, 5(12):3376–3384, 2006.

[129] Maria Julia Manso Alves, Rebeca Kawahara, Rosa Viner, Walter Colli, Eliciane Cevolani Mattos, Morten Thaysen-Andersen, Martin Røssel Larsen, and Giuseppe Palmisano. Comprehensive glycoprofiling of the epimastigote and trypomastigote stages of Trypanosoma cruzi. *Journal of Proteomics*, 151:182–192, 2017.

[130] Anais Chavaroche, Mare Cudic, Marc Giulianotti, Richard A. Houghten, Gregg B. Fields, and Dmitriy Minond. Glycosylation of a disintegrin and metalloprotease 17 affects its activity and inhibition. *Analytical Biochemistry*, 449:68–75, 2014.

[131] A J Wittwer, S C Howard, L S Carr, N K Harakas, J Feder, R B Parekh, P M Rudd, R A Dwek, and T W Rademacher. Effects of N-glycosylation on in vitro activity of Bowes melanoma and human colon fibroblast derived tissue plasminogen activator. *Biochemistry*, 28(19):7662–9, 1989.

[132] S C Howard, A J Wittwer, and J K Welply. Oligosaccharides at each glycosylation site make structure-dependent contributions to biological properties of human tissue plasminogen activator. *Glycobiology*, 1(4):411–8, 1991.

[133] Shihui Guo, Wolfgang Skala, Viktor Magdolen, Peter Briza, Martin L. Biniossek, Oliver Schilling, Josef Kellermann, Hans Brandstetter, and Peter Goettig. A Single Glycan at the 99-Loop of Human Kallikrein-related Peptidase 2 Regulates Activation and Enzymatic Activity. *Journal of Biological Chemistry*, 291(2):593–604, 2016.

[134] Albert Ardevol and Carme Rovira. Reaction Mechanisms in Carbohydrate-Active Enzymes: Glycoside Hydrolases and Glycosyltransferases. Insights from ab Initio Quantum Mechanics/Molecular Mechanics Dynamic Simulations. *J. Am. Chem. Soc.*, 137(24):7528–47, 2015.

[135] Elodie Drula, Marie-Line Garron, Suzan Dogan, Vincent Lombard, Bernard Henrissat, and Nicolas Terrapon. The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Research*, 50(D1):D571–D577, 2021.

[136] Sheng-You Huang. Comprehensive assessment of flexible-ligand docking algorithms: current effectiveness and challenges. *Briefings in Bioinformatics*, 19(5):982–994, 2017.

[137] Isabell Louise Grothaus, Giovanni Bussi, and Lucio Colombi Ciacchi. Exploration, Representation, and Rationalization of the Conformational Phase Space of N-Glycans. *Journal of Chemical Information and Modeling*, 62(20):4992–5008, 2022.

[138] Abigail Dommer, Lorenzo Casalino, Fiona Kearns, Mia Rosenfeld, Nicholas Wauer, Surl-Hee Ahn, John Russo, Sofia Oliveira, Clare Morris, Anthony Bogetti, Anda Trifan, Alexander Brace, Terra Sztain, Austin Clyde, Heng Ma, Chakra Chennubhotla, Hyungro Lee, Matteo Turilli, Syma Khalid, Teresa Tamayo-Mendoza, Matthew Welborn, Anders Christensen, Daniel GA Smith, Zhuoran Qiao, Sai K Sirumalla, Michael O'Connor, Frederick Manby, Anima Anandkumar, David Hardy, James Phillips, Abraham Stern, Josh Romero, David Clark, Mitchell Dorrell, Tom Maiden, Lei Huang, John McCalpin, Christopher Woods, Alan Gray, Matt Williams, Bryan Barker, Harinda Rajapaksha, Richard Pitts, Tom Gibbs, John Stone, Daniel M. Zuckerman, Adrian J. Mulholland, Thomas Miller, Shantenu Jha, Arvind Ramanathan, Lillian Chong, and Rommie E Amaro. #COVIDisAirborne: AI-enabled multiscale computational microscopy of delta SARS-CoV-2 in a respiratory aerosol. *The International Journal of High Performance Computing Applications*, page 10943420221128233, 2022.

[139] B J Alder and T E Wainwright. Phase Transition for a Hard Sphere System. *The Journal of Chemical Physics*, 27(5):1208–1209, 1957.

[140] J. Andrew McCammon, Bruce R. Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–590, 1977.

[141] Je'rome He'nin, Tony Lelievre, Michael R Shirts, Omar Valsson, and Lucie Delemotte. Enhanced Sampling Methods for Molecular Dynamics Simulations [Article v1.0]. *Living Journal of Computational Molecular Science*, 4(1), 2022.

[142] R W Hockney. Potential calculation and some applications. *Methods Comput. Phys*, 9:135–211, 1970.

[143] Loup Verlet. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review*, 159(1):98–103, 1967.

[144] William C Swope, Hans C Andersen, Peter H Berens, and Kent R Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, 76(1):637–649, 1982.

[145] K. Vanommeslaeghe and A.D. MacKerell. CHARMM additive and polarizable force fields for biophysics and computer-aided drug design. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1850(5):861–871, 2015.

[146] H A Lorentz. Ueber die Anwendung des Satzes vom Virial in der kinetischen Theorie der Gase. *Annalen der Physik*, 248(1):127–136, 1881.

[147] Tamar Schlick. Molecular Modeling and Simulation: An Interdisciplinary Guide. *Interdisciplinary Applied Mathematics*, pages 265–298, 2010.

[148] Gabriele Raabe. *Molecular Simulation Studies on Thermophysical Properties.* Springer, 2017.

[149] Thomas Fox and Peter A Kollman. Application of the RESP Methodology in the Parametrization of Organic Solvents. *The Journal of Physical Chemistry B*, 102(41):8070–8079, 1998.

[150] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L de Groot, Helmut Grubmüller, and Alexander D MacKerell. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods*, 14(1):71–73, 2017.

[151] Olgun Guvench, Elizabeth Hatcher, Richard M. Venable, Richard W. Pastor, and Alexander D. MacKerell. CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses. *J. Chem. Theory Comput.*, 5(9):2353–2370, 2009.

[152] Olgun Guvench, Sairam S Mallajosyula, E Prabhu Raman, Elizabeth Hatcher, Kenno Vanommeslaeghe, Theresa J Foster, Francis W Jamison, and Alexander D MacKerell. CHARMM Additive All-Atom Force Field for Carbohydrate Derivatives and Its Utility in Polysaccharide and Carbohydrate–Protein Modeling. *J. Chem. Theory Comput.*, 7(10):3162–3180, 2011.

[153] Chuan Tian, Koushik Kasavajhala, Kellon A A Belfon, Lauren Raguette, He Huang, Angela N Migues, John Bickel, Yuzhang Wang, Jorge Pincay, Qin Wu, and Carlos Simmerling. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *Journal of Chemical Theory and Computation*, 16(1):528–552, 2019.

[154] James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, 2015.

[155] Karl N Kirschner, Austin B Yongye, Sarah M Tschampel, Jorge González-Outeiriño, Charlisa R Daniels, B Lachele Foley, and Robert J Woods. GLYCAM06: A generalizable biomolecular force field. Carbohydrates: GLYCAM06. *J. Comput. Chem.*, 29(4):622–655, 2007.

[156] Chris Oostenbrink, Alessandra Villa, Alan E Mark, and Wilfred F van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *Journal of computational chemistry*, 25(13):1656–76, 2004.

[157] Maria M. Reif, Philippe H. Hunenberger, and Chris Oostenbrink. New Interaction Parameters for Charged Amino Acid Side Chains in the GROMOS Force Field. *Journal of Chemical Theory and Computation*, 8(10):3705–3723, 2012.

[158] Halvor S Hansen and Philippe H Hünenberger. A reoptimized GROMOS force field for hexopyranose-based carbohydrates accounting for the relative free energies of ring conformers, anomers, epimers, hydroxymethyl rotamers, and glycosidic linkage conformers. *Journal of computational chemistry*, 32(6):998–1032, 2010.

[159] Karina Nester, Karolina Gaweda, and Wojciech Plazinski. A GROMOS Force Field for Furanose-Based Carbohydrates. *Journal of Chemical Theory and Computation*, 15(2):1168–1186, 2019.

[160] William L. Jorgensen, David S. Maxwell, and Julian Tirado-Rives. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996.

[161] Michael J. Robertson, Julian Tirado-Rives, and William L. Jorgensen. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *Journal of Chemical Theory and Computation*, 11(7):3499–3509, 2015.

[162] Elisa Fadda. Molecular simulations of complex carbohydrates and glycoconjugates. *Curr. Opin. Chem. Biol.*, 69:102175, 2022.

[163] Callum J. Dickson, Ross C. Walker, and Ian R. Gould. Lipid21: Complex Lipid Membrane Simulations with AMBER. *Journal of Chemical Theory and Computation*, 18(3):1726–1736, 2022.

[164] Xibing He, Viet H. Man, Wei Yang, Tai-Sung Lee, and Junmei Wang. A fast and high-quality charge model for the next generation general AMBER force field. *The Journal of Chemical Physics*, 153(11):114502, 2020.

[165] Robert B. Best. Biomolecular Simulations, Methods and Protocols. pages 3–19, 2019.

[166] Robert B Best, Xiao Zhu, Jihyun Shim, Pedro E M Lopes, Jeetain Mittal, Michael Feig, and Alexander D MacKerell. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ 1 and χ 2 Dihedral Angles. *Journal of Chemical Theory and Computation*, 8(9):3257–3273, 2012.

[167] Alexander MacKerell. Empirical force fields. In *Computational Methods for Protein Structure Prediction and Modeling*, chapter 2, pages 45–70. Springer New York, NY, 2007.

[168] Christopher I Bayly, Piotr Cieplak, Wendy Cornell, and Peter A Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry*, 97(40):10269–10280, 1993.

[169] Paul Bauer, Berk Hess, and Erik Lindahl. GROMACS 2022.4 Manual. *Zenodo*, 2022.

[170] Vytautas Gapsys, Servaas Michielssens, Jan Henning Peters, Bert L de Groot, and Hadas Leonov. Calculation of binding free energies. *Methods in molecular biology (Clifton, N.J.)*, 1215:173–209, 2014.

[171] Hiroaki Fukunishi, Osamu Watanabe, and Shoji Takada. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *The Journal of Chemical Physics*, 116(20):9058–9067, 2002.

[172] Pu Liu, Byungchan Kim, Richard A Friesner, and B J Berne. Replica exchange with solute tempering: a method for sampling biological systems in explicit water. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13749–54, 2005.

[173] Giovanni Bussi. Hamiltonian replica-exchange in GROMACS: a flexible implementation. *Mol. Phys.*, 112(3-4):379–384, 2013.

[174] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proc. Natl. Acad. Sci.*, 99(20):12562–12566, 2002.

[175] Fugao Wang and D. P. Landau. Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. *Physical Review Letters*, 86(10):2050–2053, 2001.

[176] Giovanni Bussi and Davide Branduardi. Free-energy calculations with metadynamics: Theory and practice. In *Reviews in Computational Chemistry*, volume 28, chapter 1, pages 1–49. WILEY, 2015.

[177] Alejandro Gil-Ley and Giovanni Bussi. Enhanced Conformational Sampling using Replica Exchange with Collective-Variable Tempering. *J. Chem. Theory Comput.*, 11(3):1077–1085, 2015.

[178] Paolo Raiteri, Alessandro Laio, Francesco Luigi Gervasio, Cristian Micheletti, and Michele Parrinello. Efficient Reconstruction of Complex Free Energy Landscapes by Multiple Walkers Metadynamics †. *The Journal of Physical Chemistry B*, 110(8):3533–3539, 2006.

[179] Giovanni Bussi, Francesco Luigi Gervasio, Alessandro Laio, and Michele Parrinello. Free-Energy Landscape for β Hairpin Folding from Combined Parallel Tempering and Metadynamics. *J. Am. Chem. Soc.*, 128(41):13435–13441, 2006.

[180] Fahimeh Baftizadeh, Pilar Cossio, Fabio Pietrucci, and Alessandro Laio. Protein Folding and Ligand-Enzyme Binding from Bias-Exchange Metadynamics Simulations. *Current Physical Chemistrye*, 2(1):79–91, 2012.

[181] Carlo Camilloni, Davide Provasi, Guido Tiana, and Ricardo A. Broglia. Exploring the protein G helix free-energy surface by solute tempering metadynamics. *Proteins: Structure, Function, and Bioinformatics*, 71(4):1647–1654, 2008.

[182] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.*, 13(8):1011–1021, 1992.

[183] Giovanni Bussi and Gareth A Tribello. Analyzing and Biasing Simulations with PLUMED. *Methods in molecular biology (Clifton, N.J.)*, 2022:529–578, 2019.

[184] Benjamin Helfrecht, Rose K Cersonsky, Guillaume Fraux, and Michele Ceriotti. Structure-property maps with Kernel Principal Covariates Regression. *Mach. Learn.: Sci. Technol.*, 1, 2020.

[185] Gareth A. Tribello and Piero Gasparotto. Biomolecular Simulations, Methods and Protocols. *Methods in Molecular Biology*, 2022:453–502, 2019.

[186] Gareth A Tribello and Piero Gasparotto. Using Dimensionality Reduction to Analyze Protein Trajectories. *Front. Mol. Biosci.*, 6:46, 2019.

[187] I.T. Jolliffe. *Principal Component Analysis.* Springer Verlag, 1986.

[188] Rajarshi Roy, Sayan Poddar, Md Fulbabu Sk, and Parimal Kar. Conformational preferences of triantennary and tetraantennary hybrid N-glycans in aqueous solution: Insights from 20 μs long atomistic molecular dynamic simulations. *J. Biomol. Struct. Dyn.*, pages 1–16, 2022.

[189] R R Coifman, S Lafon, A B Lee, M Maggioni, B Nadler, F Warner, and S W Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci.*, 102(21):7426–7431, 2005.

[190] Sandro Bottaro, Alejandro Gil-Ley, and Giovanni Bussi. RNA folding pathways in stop motion. *Nucleic Acids Res.*, 44(12):5883–5891, 2016.

[191] Nicholas F Marshall and Ronald R Coifman. Manifold learning with bi-stochastic kernels. *J. Inst. Math. Its Appl.*, 84(3):455–482, 01 2019.

[192] Gareth A Tribello, Michele Ceriotti, and Michele Parrinello. Using sketch-map coordinates to analyze and bias molecular dynamics simulations. *Proc. Natl. Acad. Sci. U. S. A.*, 109(14):5196–201, 2012.

[193] Michele Ceriotti, Gareth A Tribello, and Michele Parrinello. From the Cover: Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. U. S. A.*, 108(32):13023–8, 2011.

[194] Warren S Torgerson. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4):401–419, 1952.

[195] S. M. Ayala Mariscal, M. L. Pigazzini, Y. Richter, M. Özel, I. L. Grothaus, J. Protze, K. Ziege, M. Kulke, M. ElBediwi, J. V. Vermaas, L. Colombi Ciacchi, S. Köppen, F. Liu, and J. Kirstein. Identification of a HTT-specific binding motif in DNAJB1 essential for suppression and disaggregation of HTT. *Nature Communications*, 13(1):4692, 2022.

[196] Eberhard Scherzinger, Annie Sittler, Katja Schweiger, Volker Heiser, Rudi Lurz, Renate Hasenbank, Gillian P. Bates, Hans Lehrach, and Erich E. Wanker. Self-assembly of polyglutamine-containing huntingtin fragments into amyloid-like fibrils: Implications for Huntington's disease pathology. *Proceedings of the National Academy of Sciences*, 96(8):4604–4609, 1999.

[197] Stephen W Davies, Mark Turmaine, Barbara A Cozens, Marian DiFiglia, Alan H Sharp, Christopher A Ross, Eberhard Scherzinger, Erich E Wanker, Laura Mangiarini, and Gillian P Bates. Formation of Neuronal Intranuclear Inclusions Underlies the Neurological Dysfunction in Mice Transgenic for the HD Mutation. *Cell*, 90(3):537–548, 1997.

[198] Ambrish Roy, Alper Kucukural, and Yang Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*, 5(4):725–738, 2010.

[199] Martin Kulke, Norman Geist, Daniel Moller, and Walter Langel. Replica-Based Protein Structure Sampling Methods: Compromising between Explicit and Implicit Solvents. *J. Phys. Chem. B*, 122(29):7295–7307, 2018.

[200] Norman Geist, Martin Kulke, Lukas Schulig, Andreas Link, and Walter Langel. Replica-Based Protein Structure Sampling Methods II: Advanced Hybrid Solvent TIGER2hs. *J. Phys. Chem. B*, 123(28):5995–6006, 2019.

[201] A. D. McNaught and A. Wilkinson. Iupac compendium of chemical terminology – the gold book, 2009.

[202] Sunhwan Jo, Taehoon Kim, Vidyashankara G. Iyer, and Wonpil Im. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.*, 29(11):1859–1865, 2008.

[203] Sunhwan Jo, Kevin C. Song, Heather Desaire, Alexander D. MacKerell, and Wonpil Im. Glycan reader: Automated sugar identification and simulation preparation for carbohydrates and glycoproteins. *J. Comput. Chem.*, 32(14):3135–3141, 2011.

[204] Sang-Jun Park, Jumin Lee, Dhilon S Patel, Hongjing Ma, Hui Sun Lee, Sunhwan Jo, and Wonpil Im. Glycan Reader is improved to recognize most sugar types and chemical modifications in the Protein Data Bank. *Bioinformatics (Oxford, England)*, 33(19):3051–3057, 2017.

[205] Sang-Jun Park, Jumin Lee, Yifei Qi, Nathan R Kern, Hui Sun Lee, Sunhwan Jo, InSuk Joung, Keehyung Joo, Jooyoung Lee, and Wonpil Im. CHARMM-GUI Glycan Modeler for modeling and simulation of carbohydrates and glycoconjugates. *Glycobiology*, 29(4):320–331, 2019.

[206] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, 2015.

[207] Gareth A. Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.*, 185(2):604–613, 2014.

[208] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins †. *J. Phys. Chem. B*, 102(18):3586–3616, 1998.

[209] Olgun Guvench, Devon Martin, and Megan Greene. Pyranose Ring Puckering Thermodynamics for Glycan Monosaccharides Associated with Vertebrate Proteins. *Int. J. Mol. Sci.*, 23(1):473, 2021.

[210] Berk Hess, Henk Bekker, Herman J C Berendsen, and Johannes G E M Fraaije. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18(12):1463–1472, 1997.

[211] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

[212] Szilárd Páll and Berk Hess. A flexible algorithm for calculating pair interactions on SIMD architectures. *Comput. Phys. Commun.*, 184(12):2641–2650, 2013.

[213] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An N x log(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[214] David F Thieker, Jodi A Hadden, Klaus Schulten, and Robert J Woods. 3D implementation of the symbol nomenclature for graphical representation of glycans. *Glycobiology*, 26(8):786–787, 2016.

[215] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(85):2825–2830, 2011.

[216] Koji Hukushima and Koji Nemoto. Exchange Monte Carlo Method and Application to Spin Glass Simulations. *J. Phys. Soc. Jpn.*, 65(6):1604–1608, 1996.

[217] Stefano Piana and Alessandro Laio. A Bias-Exchange Approach to Protein Folding. *J. Phys. Chem. B*, 111(17):4553–4559, 2007.

[218] Henning Thgersen, Raymond U. Lemieux, Klaus Bock, and Bernd Meyer. Further justification for the exo-anomeric effect. Conformational analysis based on nuclear magnetic resonance spectroscopy of oligosaccharides. *Can. J. Chem.*, 60(1):44–57, 1982.

[219] Saul Wolfe, Whangbo Myung-Hwan, and David J. Mitchell. On the magnitudes and origins of the “anomeric effects”, “exo-anomeric effects”, “reverse anomeric effects”,

and C-X and C-Y bond -lengths in XCH2YH molecules. *Carbohydr. Res.*, 69(1):1–26, 1979.

[220] Syed Shahzad-ul-Hussan, Mallika Sastry, Thomas Lemmin, Cinque Soto, Sandra Loesgen, Danielle A. Scott, Jack R. Davison, Katheryn Lohith, Robert O'Connor, Peter D. Kwong, and Carole A. Bewley. Insights from NMR Spectroscopy into the Conformational Properties of Man-9 and Its Recognition by Two HIV Binding Proteins. *ChemBioChem*, 18(8):764–771, 2017.

[221] Yu-Xi Tsai, Ning-En Chang, Klaus Reuter, Hao-Ting Chang, Tzu-Jing Yang, Sören von Bülow, Noémie Zerrouki, Michael Gecht, Camille Penet, Isabell Louise Grothaus, Lucio Colombi Ciacchi, Kay-Hooi Khoo, Gerhard Hummer, Shang-Te Hsu, Cyril Hanus, and Mateusz Sikora. Rapid simulation of glycoprotein structures by grafting and steric exclusion of glycan conformer libraries. *Cell, under revision*, 2023.

[222] Aoife M Harbison, Lorna P Brosnan, Keith Fenlon, and Elisa Fadda. Sequence-to-structure dependence of isolated IgG Fc complex biantennary N -glycans: a molecular dynamics study. *Glycobiology*, 29(1):94–103, 2018.

[223] E. W. Wooten, R. Bazzo, C. J. Edge, S. Zamze, R. A. Dwek, and T. W. Rademacher. Primary sequence dependence of conformation in oligomannose oligosaccharides. *Eur. Biophys. J.*, 18(3):139, 1989.

[224] Robert J. Woods, Ahammadunny Pathiaseril, Mark R. Wormald, Christopher J. Edge, and Raymond A. Dwek. The high degree of internal flexibility observed for an oligomannose oligosaccharide does not alter the overall topology of the molecule. *Eur. J. Biochem.*, 258(2):372–386, 1998.

[225] Carl A Fogarty and Elisa Fadda. Oligomannose N-Glycans 3D Architecture and Its Response to the FcωRIIIa Structural Landscape. *J. Phys. Chem. B*, 125(10):2607–2616, 2021.

[226] Sairam S. Mallajosyula, Sunhwan Jo, Wonpil Im, and Alexander D. MacKerell. Molecular Dynamics Simulations of Glycoproteins using CHARMM. *Methods in molecular biology (Clifton, N.J.)*, 1273:407–429, 2015.

[227] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[228] S W Homans, R A Dwek, J Boyd, M Mahmoudian, W G Richards, and T W Rademacher. Conformational transitions in N-linked oligosaccharides. *Biochemistry*, 25(20):6342–6350, 1986.

[229] C.A.G. Haasnoot, F.A.A.M. de Leeuw, and C. Altona. The relationship between proton-proton NMR coupling constants and substituent electronegativities—I An empirical generalization of the karplus equation. *Tetrahedron*, 36(19):2783–2792, 1980.

[230] Roland Stenutz, Ian Carmichael, Göran Widmalm, and Anthony S. Serianni. Hydroxymethyl Group Conformation in Saccharides: Structural Dependencies of 2 J HH, 3 J HH, and 1 J CH Spin-Spin Coupling Constants. *J. Org. Chem.*, 67(3):949–958, 2002.

[231] Mohsen Tafazzoli and Mina Ghiasi. New Karplus equations for 2JHH, 3JHH, 2JCH, 3JCH, 3JCOCH, 3JCSCH, and 3JCCCH in some aldohexopyranoside derivatives as determined using NMR spectroscopy and density functional theory calculations. *Carbohydr. Res.*, 342(14):2086–2096, 2007.

[232] Leszek Poppe, Rainer Stuike-Prill, Bernd Meyer, and Herman van Halbeek. The solution conformation of sialyl-$\alpha(2{\rightarrow}6)$-lactose studied by modern NMR techniques and Monte Carlo simulations. *J. Biomol. NMR*, 2(2):109–136, 1992.

[233] Barbara Mulloy, Gerhald W Hart, and Pamela and Stanley. N-glycans. In Ajit Varki, Richard D. Cummings, Hudson H Esko, Jeffrey D.and Freeze, Pamela Stanley, Carolyn R Bertozzi, Gerald W. Hart, and Marilynn E Etzler, editors, *Essentials of Glycobiology*, chapter 47. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 2th edition, 2009.

[234] Sunhwan Jo and Wonpil Im. Glycan fragment database: a database of PDB-based glycan 3D structures. *Nucleic Acids Res.*, 41(Database issue):D470–4, 2012.

[235] Carolina Fontana and Goran Widmalm. Primary Structure of Glycans by NMR Spectroscopy. *Chemical Reviews*, 2023.

[236] G. Chalmers, J.N. Glushka, B.L. Foley, R.J. Woods, and J.H. Prestegard. Direct NOE simulation from long MD trajectories. *Journal of Magnetic Resonance*, 265:1–9, 2016.

[237] Wojciech Plazinski, Mateusz Drach, and Anita Plazinska. Ring inversion properties of 1→2, 1→3 and 1→6-linked hexopyranoses and their correlation with the conformation of glycosidic linkages. *Carbohydr. Res.*, 423:43–48, 2016.

[238] Benedict M. Sattelle and Andrew Almond. Shaping up for structural glycomics: a predictive protocol for oligosaccharide conformational analysis applied to N-linked glycans. *Carbohydr. Res.*, 383(Science 333 2011):34–42, 2014.

[239] Wesley K. Lay, Mark S. Miller, and Adrian H. Elcock. Reparameterization of Solute-Solute Interactions for Amino Acid–Sugar Systems Using Isopiestic Osmotic Pressure Molecular Dynamics Simulations. *J. Chem. Theory Comput.*, 13(5):1874–1882, 2017.

[240] Wesley K. Lay, Mark S. Miller, and Adrian H. Elcock. Optimizing Solute–Solute Interactions in the GLYCAM06 and CHARMM36 Carbohydrate Force Fields Using Osmotic Pressure Measurements. *J. Chem. Theory Comput.*, 12(4):1401–1407, 2016.

[241] Jorg Sauter and Andrea Grafmuller. Solution Properties of Hemicellulose Polysaccharides with Four Common Carbohydrate Force Fields. *J. Chem. Theory Comput.*, 11(4):1765–1774, 2015.

[242] Andrea Cesari, Sabine Reißer, and Giovanni Bussi. Using the Maximum Entropy Principle to Combine Simulations and Solution Experiments. *Computation*, 6(1):15, 2018.

[243] Angeles Canales, Alvaro Mallagaray, Javier Pérez-Castells, Irene Boos, Carlo Unverzagt, Sadine André, Hans-Joachim Gabius, Francisco Javier Cañada, and Jesús Jiménez-Barbero. Breaking Pseudo-Symmetry in Multiantennary Complex N-Glycans Using Lanthanide-Binding Tags and NMR Pseudo-Contact Shifts. *Angew. Chem., Int. Ed.*, 52(51):13789–13793, 2013.

[244] Wojciech Plazinski and Mateusz Drach. The influence of the hexopyranose ring geometry on the conformation of glycosidic linkages investigated using molecular dynamics simulations. *Carbohydr. Res.*, 415:17–27, 2015.

[245] Jana Rosenau, Isabell Louise Grothaus, Yikun Yang, Nilima Dinesh Kumar, Lucio Colombi Ciacchi, Sørge Kelm, and Mario Waespy. N-glycosylation modulates enzymatic activity of Trypanosoma congolense trans-sialidase. *Journal of Biological Chemistry*, 298:102403, 2022.

[246] Mahamat Alhadj Moussa Ibrahim, Judith Sophie Weber, Sen Claudine Henriette Ngomtcho, Djoukzoumka Signaboubo, Petra Berger, Hassane Mahamat Hassane, and Sørge Kelm. Diversity of trypanosomes in humans and cattle in the HAT foci Mandoul and Maro, Southern Chad—A matter of concern for zoonotic potential? *PLoS Neglected Tropical Diseases*, 15(6):e0009323, 2021.

[247] Centers for Disease Control and Prevention. Parasites - african trypanosomiasis (also known as sleeping sickness). March 2020.

[248] Virginie Coustou, Nicolas Plazolles, Fabien Guegan, and Théo Baltz. Sialidases play a key role in infection and anaemia in Trypanosoma congolense animal trypanosomiasis. *Cellular Microbiology*, 14(3):431 445, 02 2012.

[249] Andrew J. Nok and Emmanuel O. Balogun. A Bloodstream Trypanosoma congolense Sialidase Could Be Involved in Anemia during Experimental Trypanosomiasis. *The Journal of Biochemistry*, 133(6):725–730, 2003.

[250] Keith L. Banks. In Vitro Binding of Trypanosoma congolense to Erythrocytes*. *The Journal of Protozoology*, 26(1):103–108, 1979.

[251] Hendrik Koliwer-Brandl, Thaddeus T Gbem, Mario Waespy, Olga Reichert, Philipp Mandel, Eric Drebitz, Frank Dietz, and Sørge Kelm. Biochemical characterization of trans-sialidase TS1 variants from Trypanosoma congolense. *BMC Biochemistry*, 12(1):39, 07 2011.

[252] Evelin Tiralongo, Ilka Martensen, Joachim Grötzinger, Joe Tiralongo, and Roland Schauer. Trans-sialidase-like sequences from Trypanosoma congolense conserve most of the critical active site residues found in other trans-sialidases. *Biological chemistry*, 384(8):1203 1213, 2003-08.

[253] Thaddeus T Gbem, Mario Waespy, Bettina Hesse, Frank Dietz, Joel Smith, Gloria D Chechet, Jonathan A Nok, and Sørge Kelm. Biochemical Diversity in the Trypanosoma congolense Trans-sialidase Family. *PLoS Neglected Tropical Diseases*, 7(12):e2549 12, 12 2013.

[254] Georgina Montagna, M Laura Cremona, Gastón Paris, M Fernanda Amaya, Alejandro Buschiazzo, Pedro M Alzari, and Alberto C C Frasch. The trans-sialidase from the african trypanosome Trypanosoma brucei. *Eur. J. Biochem.*, 269(12):2941 2950, 2002-06.

[255] Maria Fernanda Amaya, Alejandro Buschiazzo, Tong Nguyen, and Pedro M Alzari. The high resolution structures of free and inhibitor-bound Trypanosoma rangeli sialidase and its comparison with T. cruzi trans-sialidase. *J. Mol. Biol.*, 325(4):773–84, 2003.

[256] Oscar E Campetella, Antonio D Uttaro, Armando J Parodi, and Alberto C C Frasch. A recombinant Trypanosoma cruzi trans-sialidase lacking the amino acid repeats retains the enzymatic activity. *Molecular and Biochemical Parasitology*, 64(2):337–340, 1994.

[257] Mario Waespy, Thaddeus T Gbem, Leroy Elenschneider, André-Philippe Jeck, Christopher J Day, Lauren Hartley-Tassell, Nicolai Bovin, Joe Tiralongo, Thomas Haselhorst, and Sørge Kelm. Carbohydrate Recognition Specificity of Trans-sialidase Lectin Domain from Trypanosoma congolense. *PLoS Neglected Tropical Diseases*, 9(10):e0004120, 2015.

[258] Carole L F Haynes, Paul Ameloot, Han Remaut, Nico Callewaert, Yann G J Sterckx, and Stefan Magez. Production, purification and crystallization of a trans -sialidase

from Trypanosoma vivax. *Acta Crystallographica Section F Structural Biology Communications*, 71(5):577–585, 2015.

[259] Haiying Li, Morten I Rasmussen, Martin R Larsen, Yao Guo, Carsten Jers, Giuseppe Palmisano, Jørn D Mikkelsen, and Finn Kirpekar. Automated N-glycan profiling of a mutant Trypanosoma rangeli sialidase expressed in Pichia pastoris, using tandem mass spectrometry and bioinformatics. *Glycobiology*, 25(12):1350–1361, 2015.

[260] Samuel M. Duncan and Michael A.J. Ferguson. Common and unique features of glycosylation and glycosyltransferases in African trypanosomes. *Biochemical Journal*, 479(17):1743–1758, 2022.

[261] P Stanley. Glycosylation Mutants of Animal Cells. *Annual Review of Genetics*, 18(1):525–552, 1984.

[262] Simon J North, Hung-Hsiang Huang, Subha Sundaram, Jihye Jang-Lee, A Tony Etienne, Alana Trollope, Sara Chalabi, Anne Dell, Pamela Stanley, and Stuart M Haslam. Glycomics profiling of Chinese hamster ovary cell glycosylation mutants reveals N-glycans of a novel size and complexity. *J. Biol. Chem.*, 285(8):5759 5775, 02 2010.

[263] Mario Waespy, Thaddeus Termulun Gbem, Nilima Dinesh Kumar, Shanmugam Solaiyappan Mani, Jana Rosenau, Frank Dietz, and Sørge Kelm. Cooperativity of catalytic and lectin-like domain of Trypanosoma congolense trans-sialidase modulates its catalytic activity. *PLoS Neglected Tropical Diseases*, 16(2):e0009585, 2022.

[264] Sillanaukee, Pönniö, and Jääskeläinen. Occurrence of sialic acids in healthy humans and different disorders. *European Journal of Clinical Investigation*, 29(5):413–425, 1999.

[265] Yuko Nagai, Iori Sakakibara, and Hidenao Toyoda. Microdetermination of Sialic Acids in Blood Samples by Hydrophilic Interaction Chromatography Coupled to Post-column Derivatization and Fluorometric Detection. *Analytical Sciences*, 35(5):517–520, 2019.

[266] Yoko Kita, Yoshiaki Miura, Jun-ichi Furukawa, Mika Nakano, Yasuro Shinohara, Masahiro Ohno, Akio Takimoto, and Shin-Ichiro Nishimura. Quantitative Glycomics of Human Whole Serum Glycoproteins Based on the Standardized Protocol for Liberating N-Glycans *. *Molecular & Cellular Proteomics*, 6(8):1437–1445, 2007.

[267] Guinevere S M Lageveen-Kammeijer, Noortje de Haan, Pablo Mohaupt, Sander Wagt, Mike Filius, Jan Nouta, David Falck, and Manfred Wuhrer. Highly sensitive CE-ESI-MS analysis of N-glycans from complex biological samples. *Nat. Commun.*, 10(1):2137, 2019.

[268] A Buschiazzo, O Campetella, and A C Frasch. Trypanosoma rangeli sialidase: cloning, expression and similarity to T. cruzi trans-sialidase. *Glycobiology*, 7(8):1167–73, 1997.

[269] András Micsonai, Frank Wien, Linda Kernya, Young-Ho Lee, Yuji Goto, Matthieu Réfrégiers, and József Kardos. Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proceedings of the National Academy of Sciences*, 112(24):E3095–E3103, 2015.

[270] András Micsonai, Frank Wien, Éva Bulyáki, Judit Kun, Éva Moussong, Young-Ho Lee, Yuji Goto, Matthieu Réfrégiers, and József Kardos. BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra. *Nucleic Acids Research*, 46(Web Server issue):gky497–, 2018.

[271] Norma J Greenfield. Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. *Nature Protocols*, 1(6):2527–2535, 2006.

[272] Maurizio Molinari. N-glycan structure dictates extension of protein folding or onset of disposal. *Nature Chemical Biology*, 3(6):313–320, 2007.

[273] Jianyi Yang and Yang Zhang. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.*, 43(W1):W174–W181, 2015.

[274] Corina Mathew, R. Gregor Weiß, Christoph Giese, Chia-wei Lin, Marie-Estelle Losfeld, Rudi Glockshuber, Sereina Riniker, and Markus Aebi. Glycan-Protein Interactions Determine Kinetics of N-Glycan Remodeling. *bioRxiv*, page 2020.12.01.406371, 2020.

[275] Ozlem Demir and Adrian E Roitberg. Modulation of catalytic function by differential plasticity of the active site: case study of Trypanosoma cruzi trans-sialidase and Trypanosoma rangeli sialidase. *Biochemistry*, 48(15):3398–406, 2009.

[276] Felicity L. Mitchell, Steven M. Miles, João Neres, Elena V. Bichenkova, and Richard A. Bryce. Tryptophan as a Molecular Shovel in the Glycosyl Transfer Activity of Trypanosoma cruzi Trans-sialidase. *Biophys. J.*, 98(9):L38–L40, 2010.

[277] Felicity L. Mitchell, João Neres, Anitha Ramraj, Rajesh K. Raju, Ian H. Hillier, Mark A. Vincent, and Richard A. Bryce. Insights into the Activity and Specificity of Trypanosoma cruzi trans-Sialidase from Molecular Dynamics Simulations. *Biochemistry*, 52(21):3740–3751, 2013.

[278] Isadora A. Oliveira, Arlan S. Gonçalves, Jorge L. Neves, Mark von Itzstein, and Adriane R. Todeschini. Evidence of Ternary Complex Formation in Trypanosoma cruzi trans-Sialidase Catalysis. *J. Biol. Chem.*, 289(1):423–436, 2014.

[279] Adam W. Barb. Intramolecular N-Glycan/Polypeptide Interactions Observed at Multiple N-Glycan Remodeling Steps through [13C,15N]-N-Acetylglucosamine Labeling of Immunoglobulin G1. *Biochemistry*, 54(2):313–322, 2015.

[280] Sheng-Hung Wang, Tsai-Jung Wu, Chien-Wei Lee, and John Yu. Dissecting the conformation of glycans and their interactions with proteins. *Journal of Biomedical Science*, 27(1):93, 2020.

[281] James W. Dennis, Maria Granovsky, and Charles E. Warren. Protein glycosylation in development and disease. *BioEssays*, 21(5):412–421, 1999.

[282] P E Goss, M A Baker, J P Carver, and J W Dennis. Inhibitors of carbohydrate processing: A new class of anticancer agents. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 1(9):935–44, 1995.

[283] P E Goss, C L Reid, D Bailey, and J W Dennis. Phase IB clinical trial of the oligosaccharide processing inhibitor swainsonine in patients with advanced malignancies. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 3(7):1077–86, 1997.

[284] D Bowen, J Adir, S L White, C D Bowen, K Matsumoto, and K Olden. A preliminary pharmacokinetic evaluation of the antimetastatic immunomodulator swainsonine: clinical and toxic implications. *Anticancer research*, 13(4):841–4, 1993.

[285] Zheng Yang Lee, Jason Siau Ee Loo, Agustono Wibowo, Mohd Fazli Mohammat, and Jhi Biau Foo. Targeting Cancer via Golgi α-mannosidase II Inhibition: How Far Have We Come In Developing Effective Inhibitors? *Carbohydrate Research*, 508:108395, 2021.

[286] Douglas A. Kuntz, Huizhen Liu, Mikael Bols, and David R. Rose. The role of the active site Zn in the catalytic mechanism of the GH38 Golgi α-mannosidase II: Implications from noeuromycin inhibition. *Biocatalysis and Biotransformation*, 24(1-2):55–61, 2006.

[287] David R Rose. Structure, mechanism and inhibition of Golgi α-mannosidase II. *Current Opinion in Structural Biology*, 22(5):558–562, 2012.

[288] B Henrissat and A Bairoch. New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *The Biochemical journal*, 293 (Pt 3)(3):781–8, 1993.

[289] Niket Shah, Douglas A. Kuntz, and David R. Rose. Golgi α-mannosidase II cleaves two sugars sequentially in the same catalytic site. *Proceedings of the National Academy of Sciences*, 105(28):9570–9575, 2008.

[290] Luis Petersen, Albert Ardèvol, Carme Rovira, and Peter J Reilly. Molecular mechanism of the glycosylation step catalyzed by Golgi alpha-mannosidase II: a QM/MM metadynamics investigation. *J. Am. Chem. Soc.*, 132(24):8291–300, 2010.

[291] Andrew J. Thompson, Jerome Dabin, Javier Iglesias-Fernández, Albert Ardèvol, Zoran Dinev, Spencer J. Williams, Omprakash Bande, Aloysius Siriwardena, Carl Moreland, Ting-Chou Hu, David K. Smith, Harry J. Gilbert, Carme Rovira, and Gideon J. Davies. The Reaction Coordinate of a Bacterial GH47 α-Mannosidase: A Combined Quantum Mechanical and Structural Approach. *Angewandte Chemie International Edition*, 51(44):10997–11001, 2012.

[292] Katie J Gregg, Wesley F Zandberg, Jan-Hendrik Hehemann, Garrett E Whitworth, Lehua Deng, David J Vocadlo, and Alisdair B Boraston. Analysis of a new family of widely distributed metal-independent alpha-mannosidases provides unique insight into the processing of N-linked glycans. *The Journal of biological chemistry*, 286(17):15586–96, 2011.

[293] Andrew J. Thompson, Gaetano Speciale, Javier Iglesias-Fernández, Zalihe Hakki, Tyson Belz, Alan Cartmell, Richard J. Spears, Emily Chandler, Max J. Temple, Judith Stepper, Harry J. Gilbert, Carme Rovira, Spencer J. Williams, and Gideon J. Davies. Evidence for a Boat Conformation at the Transition State of GH76 α-1,6-Mannanases—Key Enzymes in Bacterial and Fungal Mannoprotein Metabolism. *Angewandte Chemie International Edition*, 54(18):5378–5382, 2015.

[294] Mingjun Yang, Jing Huang, Raphael Simon, Lai-Xi Wang, and Alexander D MacKerell. Conformational Heterogeneity of the HIV Envelope Glycan Shield. *Sci. Rep.*, 7(1):4435, 2017.

[295] Jodi A Hadden, Alfred D French, and Robert J Woods. Unraveling Cellulose Microfibrils: A Twisted Tale: Unraveling Cellulose Microfibrils. *Biopolymers*, 99(10):746–756, 2013.

[296] Winifred M. Watkins and W. T. J. Morgan. Neutralization of the Anti-H Agglutinin in Eel Serum by Simple Sugars. *Nature*, 169(4307):825–826, 1952.

[297] Richard D. Cummings. The repertoire of glycan determinants in the human glycome. *Molecular BioSystems*, 5(10):1087–1104, 2009.

[298] Ruedi Aebersold, Jeffrey N Agar, I Jonathan Amster, Mark S Baker, Carolyn R Bertozzi, Emily S Boja, Catherine E Costello, Benjamin F Cravatt, Catherine Fenselau, Benjamin A Garcia, Ying Ge, Jeremy Gunawardena, Ronald C Hendrickson, Paul J Hergenrother, Christian G Huber, Alexander R Ivanov, Ole N

Jensen, Michael C Jewett, Neil L Kelleher, Laura L Kiessling, Nevan J Krogan, Martin R Larsen, Joseph A Loo, Rachel R Ogorzalek Loo, Emma Lundberg, Michael J MacCoss, Parag Mallick, Vamsi K Mootha, Milan Mrksich, Tom W Muir, Steven M Patrie, James J Pesavento, Sharon J Pitteri, Henry Rodriguez, Alan Saghatelian, Wendy Sandoval, Hartmut Schlüter, Salvatore Sechi, Sarah A Slavoff, Lloyd M Smith, Michael P Snyder, Paul M Thomas, Mathias Uhlén, Jennifer E Van Eyk, Marc Vidal, David R Walt, Forest M White, Evan R Williams, Therese Wohlschlager, Vicki H Wysocki, Nathan A Yates, Nicolas L Young, and Bing Zhang. How many human proteoforms are there? *Nature Chemical Biology*, 14(3):206–214, 2018.

[299] Katrine T. Schjoldager, Yoshiki Narimatsu, Hiren J. Joshi, and Henrik Clausen. Global view of human protein glycosylation pathways and functions. *Nat. Rev. Mol. Cell Biol.*, 21(12):729–749, 2020.

[300] Aoife Harbison and Elisa Fadda. An atomistic perspective on ADCC quenching by core-fucosylation of IgG1 Fc N-glycans from enhanced sampling molecular dynamics. *Glycobiology*, 2019.

[301] Maja Pučić, Ana Knežević, Jana Vidič, Barbara Adamczyk, Mislav Novokmet, Ozren Polašek, Olga Gornik, Sandra Šupraha Goreta, Mark R. Wormald, Irma Redžić, Harry Campbell, Alan Wright, Nicholas D. Hastie, James F. Wilson, Igor Rudan, Manfred Wuhrer, Pauline M. Rudd, Djuro Josić, and Gordan Lauc. High Throughput Isolation and Glycosylation Analysis of IgG–Variability and Heritability of the IgG Glycome in Three Isolated Human Populations*. *Molecular & Cellular Proteomics*, 10(10):M111.010090, 2011.

[302] Sanne E. de Jong, Maurice H. J. Selman, Ayola A. Adegnika, Abena S. Amoah, Elly van Riet, Yvonne C. M. Kruize, John G. Raynes, Alejandro Rodriguez, Daniel Boakye, Erika von Mutius, André C. Knulst, Jon Genuneit, Philip J. Cooper, Cornelis H. Hokke, Manfred Wuhrer, and Maria Yazdanbakhsh. IgG1 Fc N-glycan galactosylation as a biomarker for immune activation. *Scientific Reports*, 6(1):28207, 2016.

[303] Aoife M. Harbison, Carl A. Fogarty, Toan K. Phung, Akash Satheesan, Benjamin L. Schulz, and Elisa Fadda. Fine-tuning the spike: role of the nature and topology of the glycan shield in the structure and dynamics of the SARS-CoV-2 S†. *Chemical Science*, 13(2):386–395, 2021.

[304] D. E. Koshland. Application of a Theory of Enzyme Specificity to Protein Synthesis*. *Proceedings of the National Academy of Sciences*, 44(2):98–104, 1958.

[305] Monique J. Rogals, Alexander Eletsky, Chin Huang, Laura C. Morris, Kelley W. Moremen, and James H. Prestegard. Glycan Conformation in the Heavily Glycosylated Protein, CEACAM1. *ACS Chemical Biology*, 17(12):3527–3534, 2022.

[306] Asaminew H. Aytenfisu, Mingjun Yang, and Alexander D. MacKerell. CHARMM Drude Polarizable Force Field for Glycosidic Linkages Involving Pyranoses and Furanoses. *Journal of Chemical Theory and Computation*, 14(6):3132–3143, 2018.

[307] Poonam Pandey, Asaminew H Aytenfisu, Alexander D MacKerell, and Sairam S Mallajosyula. Drude Polarizable Force Field Parametrization of Carboxylate and N-Acetyl Amine Carbohydrate Derivatives. *Journal of chemical theory and computation*, 15(9):4982–5000, 2019.

[308] Oliver T Unke, Stefan Chmiela, Huziel E Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T Schütt, Alexandre Tkatchenko, and Klaus-Robert Müller. Machine Learning Force Fields. *Chemical reviews*, 121(16):10142–10186, 2021.

[309] J. S. Smith, O. Isayev, and A. E. Roitberg. Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. *Chem. Sci.*, 8:3192–3203, 2017.

[310] Philipp Thölke and Gianni De Fabritiis. TorchMD-NET: Equivariant Transformers for Neural Network based Molecular Potentials. *arXiv*, 2022.

[311] Charlotte Toustou, Marie-Laure Walet-Balieu, Marie-Christine Kiefer-Meyer, Marine Houdou, Patrice Lerouge, François Foulquier, and Muriel Bardor. Towards understanding the extensive diversity of protein N-glycan structures in eukaryotes. *Biological Reviews*, 97(2):732–748, 2022.

[312] Marcel Tutor Ale and Anne S. Meyer. Fucoidans from brown seaweeds: an update on structures, extraction techniques and use of enzymes as tools for structural elucidation. *RSC Advances*, 3(22):8131–8141, 2013.

[313] Ahmed Zayed and Roland Ulber. Fucoidans: Downstream Processes and Recent Applications. *Marine Drugs*, 18(3):170, 2020.

[314] Georg Kopplin, Anne Mari Rokstad, Hugo Me'lida, Vincent Bulone, Gudmund Skjåk-Bræk, and Finn Lillelund Aachmann. Structural Characterization of Fucoidan from Laminaria hyperborea: Assessment of Coagulation and Inflammatory Properties and Their Structure–Function Relationship. *ACS Applied Bio Materials*, 1(6):1880–1892, 2018.

[315] Yongfei Cai, Jun Zhang, Tianshu Xiao, Christy L. Lavine, Shaun Rawson, Hanqin Peng, Haisun Zhu, Krishna Anand, Pei Tong, Avneesh Gautam, Shen Lu, Sarah M. Sterling, Richard M. WalshJr., Sophia Rits-Volloch, Jianming Lu, Duane R. Wesemann, Wei Yang, Michael S. Seaman, and Bing Chen. Structural basis for enhanced infectivity and immune evasion of SARS-CoV-2 variants. *Science*, 373(6555):642–648, 2021.

[316] Maddy L. Newby, Carl A. Fogarty, Joel D. Allen, John Butler, Elisa Fadda, and Max Crispin. Variations within the Glycan Shield of SARS-CoV-2 Impact Viral Spike Dynamics. *Journal of Molecular Biology*, 435(4):167928, 2023.

[317] C.E. Shannon. Communication in the Presence of Noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[318] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, 1977.

[319] P P Ewald. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik*, 369(3):253–287, 1921.

[320] H J C Berendsen, J P M Postma, W F van Gunsteren, A DiNola, and J R Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984.

[321] Shuichi Nosé and M L Klein. Constant pressure molecular dynamics for molecular systems. *Molecular Physics*, 50(5):1055–1076, 1983.

[322] M Parrinello and A Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 1981.

[323] Sharon Kelly and Nicholas Price. The Use of Circular Dichroism in the Investigation of Protein Structure and Function. *Current Protein & Peptide Science*, 1(4):349–384, 2000.

[324] Norma J Greenfield. Using circular dichroism spectra to estimate protein secondary structure. *Nature Protocols*, 1(6):2876–2890, 2006.

[325] Yang Wei, Aby A. Thyparambil, and Robert A. Latour. Protein helical structure determination using CD spectroscopy for solutions with strong background absorbance from 190 to 230nm. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1844(12):2331–2337, 2014.

[326] G Holzwarth and P Doty. The ultraviolet circular dichroism of polypeptides. *Journal of the American Chemical Society*, 87(2):218–28, 1965.

[327] N Greenfield and G D Fasman. Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry*, 8(10):4108–16, 1969.

[328] S.Y. Venyaminov, I.A. Baikalov, Z.M. Shen, C.S.C. Wu, and J.T. Yang. Circular Dichroic Analysis of Denatured Proteins: Inclusion of Denatured Proteins in the Reference Set. *Analytical Biochemistry*, 214(1):17–24, 1993.

[329] Norma J Greenfield. Analysis of the kinetics of folding of proteins and peptides using circular dichroism. *Nature Protocols*, 1(6):2891–2899, 2006.

[330] Jeffrey Rohrer. Carbohydrate analysis by high-performance anion-exchange chromatography with pulsed amperometric detection (HPAE-PAD). *Thermo Fisher Scientific, Technical Note*, 2021.

[331] J S Rohrer. Separation of asparagine-linked oligosaccharides by high-pH anion-exchange chromatography with pulsed amperometric detection: empirical relationships between oligosaccharide structure and chromatographic retention. *Glycobiology*, 5(4):359–60, 1995.

[332] Leonor Michaelis, Maud Leonora Menten, Kenneth A Johnson, and Roger S Goody. The original Michaelis constant: translation of the 1913 Michaelis-Menten paper. *Biochemistry*, 50(39):8264–9, 2011.

[333] Elijah Roberts, John Eargle, Dan Wright, and Zaida Luthey-Schulten. MultiSeq: unifying sequence and structure data for evolutionary analysis. *BMC Bioinformatics*, 7(1):382–382, 2006.

Declaration

I, Isabell Louise Grothaus, confirm that the work presented in this thesis was carried out without any unauthorized third party assistance. No other sources or aids than the ones specified have been used. Text passages which have been included word by word or by content from other sources have been indicated accordingly.

Isabell Louise Grothaus
Bremen, Germany
09.06.2023

Supervision

Results from the supervision of the following students' work have been included in this dissertation:

Paul Spellerberg, Conformational characteristics of N-glycans, 2022.

Annika Niemann, Molekulardynamische Simulationen zur Charakterisierung eines trimeren Chaperonkomplexes für die Disaggregation von Huntingtin, 2022.

Georg Bossenz, Molecular dynamics simulation of selected glycans and verification using Karplus equation variations, 2021.

Yikun Yang, The influence of N-linked glycans on secondary structure and stability of Trans-sialidase 1b from *Trypanosoma congolense*, 2020.