Substituted Oligosaccharides as Protein Mimics: Deep Learning Free Energy Landscapes

Benjamin Bouvier

Substituted oligosaccharides as protein mimics: deep learning free energy landscapes

Benjamin Bouvier*

Enzyme and Cell Engineering, CNRS UMR7025/Université de Picardie Jules Verne, 10, rue Baudelocque, 80039 Amiens Cedex, France.

E-mail: benjamin.bouvier@u-picardie.fr

Abstract

Protein-protein complexes power the majority of cellular processes. Interfering with the formation of such complexes using well-designed mimics is a difficult, yet actively pursued, research endeavor. Due to the limited availability of results on the conformational preferences of oligosaccharides compared to polypeptides, the former have been much less explored than the latter as protein mimics, despite interesting ADMET characteristics.

In this work, the conformational landscapes of a series of 956 substituted glucopyranose oligomers of lengths 3 to 12 designed as protein interface mimics are revealed using microsecond-timescale, enhanced-sampling molecular dynamics simulations. Deep convolutional networks are trained on these large conformational ensembles, to predict the stability of longer oligosaccharide structures from those of their constituent trimer motifs. Deep generative adversarial networks are then designed to suggest plausible conformations for oligosaccharide mimics of arbitrary length and substituent sequences, that can subsequently be used as input to docking simulations. Analyzing the performance of the neural networks also yields insights into the intricate collective effects that dominate oligosaccharide conformational dynamics.

Introduction

Protein-protein (PP) complexes power the majority of cellular processes and have thus long since been recognized for their potential as drug targets. By competing with native protein partners for recognition and binding, a well-designed mimic can potentially inhibit or regulate the formation of a PP complex and the associated biological function.[1] However, this endeavor is much more challenging than the design of traditional drugs, designed to fit inside a binding pocket of known size, shape and chemical character and interact with a handful of well-defined aminoacids only: PP interfaces are usually extensive (involving tens or hundreds of amino acids of varying chemical nature) and predominantly comprise large, flat patches. Alanine-scan experiments have revealed that only a few clusters of amino acids, often termed hotspots, contribute significantly to the binding free energy and thus represent prime drug targets.[2] These hotspots usually correlate with marked evolutionary conservation and isolation from water in the native complex.[3] Recently, more subtle hotspot features have been captured using machine learning, furthering the ability to predict hotspots at PP interfaces without resorting to long and costly scanning experiments.[4–6]

Binding to a PP interface both strongly and specifically requires making simultaneous contacts with several hotspots of different chemical characters and forming a determined pat-

tern in space. In the native protein partner, this is achieved with complementary amino acids conformationally constrained by the protein fold. Indeed, a mimic molecule sufficiently large to simultaneously target multiple hotspots would probably lack the required conformational rigidity: for instance, peptides have been extensively applied as mimics of protein partners targeting PP interfaces, but need to be conformationally constrained using scaffolding groups.[7] However, because PP interfaces are typically much flatter than active sites, the need for conformational rigidity in mimics is not as drastic, which allows some latitude in the mimic design process and is advantageous from the entropic point of view.

Oligosaccharides, the long forgotten third class of biomacromolecules, are coming of age as interesting alternatives to peptides for the design of PP interface mimics: chitosan derivatives,[8] sugar foldamers,[9] sugar amino acids,[10] leptin-based oligopeptides,[11] glycopeptide-antibody chimeras,[12] glycosated dendrimers[13] have all demonstrated their value as modulators of PP interfaces. Oligosaccharides can be obtained from naturally occurring polymers (chitin, cellulose, starch...) and are more rigid than peptides of similar sizes due to the cyclicity of their monomer constituents. Additionally, they avoid some of the unfavorable ADMET profiles of polypeptides (self-aggregation, proteolysis, immune response). The synthesis and purification of carbohydrates have long represented a bottleneck to their widespread adoption, especially for industrial applications. Indeed, synthetic pathways leading to oligosaccharides have had to be painstakingly designed to ensure the specificity of each monomer addition, usually by adding several protecting/leaving group steps (and the associated separation and purification tasks) between each chain extension phase. Fortunately, the development of automated synthesis technologies[14] now provides rapid access to a wide variety of oligosaccharides, allowing the screening of such compounds for drug discovery[15]. These methodological developments have also helped to bolster the bioavailability of oligosaccharide drugs (one of their long-standing limitations) by reducing the compounds to their smallest active components or by combining them with other molecules. Arixtra, auranofin, zanamivir, topiramate, acarbose, elmiron, sulodexide, fucoidan, idrabiotaparinux, fondaparinux are all highly bioavailable intravenously, while other drugs such as pentosan polysulphate can be orally delivered[16]. Interstingly, carbohydrate peptidomimetics often show superior bioavailability compared to the peptides they emulate, whose amide backbone makes them less permeable to membranes[17]. Notwithstanding other challenges (such as the rapid clearance by the organism), oligosaccharide now appear as compelling drug scaffolds.

Selecting a potentially suitable functionalized oligosaccharide mimic to target a given PP interface requires understanding the impact of the oligomer length and the nature of its substituent groups on its conformational preference. While globular proteins, in which inter-residue hydrogen bonds impose secondary and tertiary structures, can be described by a single fold, polysaccharide hydrogen bonds tend to be displaced by water and have a less stringent effect on the overall conformational preference. Oligosaccharides are thus best described as weighted conformational ensembles, making these molecules less amenable than proteins to experimental structural methods.[18] Fortunately, this issue has been partly alleviated by successfully combining molecular dynamics (MD) with NMR experiments;[19] the conformational information garnered by such studies has progressively been compiled into structural oligosaccharide databases, initiating with GlycoMapsDB in 2007[20] and continuing to this day. However, most databases map conformational space with only two dihedral angles per glycosidic linkage[21] (neglecting pyranose degrees of freedom), and even the most recent[22] only contains a total of 2598 distinct conformational maps...

Predicting the conformational preference of oligosaccharides from their structural formulas is thus highly desirable, for applications ranging from the fundamental understanding of oligosaccharide conformational space topologies to the practical design of chemobiological drugs.

For the latter application, the possible differences between the free and protein-bound conformations of an oligosaccharide mimic (which depend on the nature of the partners) is not detrimental: the first steps of the recognition between partners typically occur at sufficiently large distances for the knowledge of the free mimic conformational preferences to remain meaningful[23]. In this paper, I use long-timescale (total simulation time $\sim 4\,\mathrm{ms}$), enhanced-sampling molecular dynamics simulations to extensively characterize the conformational free energy landscape of $\alpha{-}1,4$ glucopyranose oligomers of three different lengths (trimers, hexamers and dodecamers) substituted with 8 possible moieties. I design a recursive convolutional deep learning network, trained on this extensive dataset, to predict the stability of any given conformation of an oligomer of given length and substitution, and examine whether this information can be inferred from the stability of its constituent trisaccharide motifs. Finally, a generative adversarial network is introduced to suggest stable conformations for arbitrary oligomers; it can be used as a source of potential PP interface mimics, for instance to power subsequent high-throughput protein-mimic docking simulations.

Methods

Molecular simulations. The forcefield parameters for the oligosaccharide mimics were derived from the GLYCAM[24] and GAFF[25] forcefields; the atomic charges for the substituents were obtained using the RESP procedure.[26] The oligomers were assembled from the corresponding parameterized fragments using LEaP[27] and ACPYPE[28] (see Supporting Information Available for details).

The trimer, hexamer and dodecamer systems were minimized and equilibrated using the procedure described in Supporting Information Available. They were simulated for respectively 1 µs, 1 µs, 1.5 µs at 300 K and 1 bar; conformations were recorded every 10 ps. Dihedral principal component analyses (dPCA),[29] including all dihedral angles involving non-hydrogen atoms, were performed on these production trajectories. The first two dPCA eigenmodes (as ranked by contribution to variance) were used as collective variables to monitor and enhance conformational sampling. Well-tempered metadynamics simulations[30] of 2 µs, 4 µs, 6 µs were performed along these two variables for trimers, hexamers and dodecamers, respectively. Frames were extracted every 10 ps for all simulations for subsequent analysis. The free energy landscape of each oligomer was obtained as a function of the collective variables from the sum of the Gaussian biasing potentials accumulated during the simulations; its convergence with respect to simulation length was verified (see Supporting Information Available). Molecular dynamics simulations were performed using GROMACS 2021.2[31] and PLUMED 2.5;[32] a dPCA module was implemented inside PLUMED specifically for this study.

Deep learning. An adequate encoding of dataset exemplars is crucial to the performance of deep learning methods. In this work, dihedral angles were encoded as their sine and cosine values, which is a straightforward way of implementing angle periodicity into the networks but requires two input nodes per angle. Each monomer comprises 18 angles (13 for the sugar, 2 for the linker, and 3 spanning sugar and linker), while two consecutive monomers are linked by 6 angles. The 8 possible monomer substituent types (see figure 1) were mapped to integer values between 0 and 7 and encoded into an 8-dimensional latent space vector using embedding nodes. Thus, a conformation of an oligosaccharide of length n was input to the classifier networks as a vector of dimension $n\times(8+2\times18)+(n-1)\times2\times6$. The free energy of each conformation was obtained by interpolating over the relevant free energy surface based on the values of the projections on the dPCA eigenvectors. Conformations with free energy values lower or equal to $2.5\,\mathrm{kcal\,mol^{-1}}$ relative to the global minimum, which form the attraction basins of the main non-metastable minima on the free energy surfaces (see Results and discussion), were labeled as the positive samples of the dataset; the remaining conformations were

labeled as negatives. A conformation was considered to be predicted positive by the classifier networks if the output of the final sigmoid node belonged to the $[0.5, 1]$ range, negative otherwise.

The dataset populations were as follows: 102 400 512 conformations for trimers, 147 600 369 for hexamers, 45 375 075 for dodecamers, and 96 976 131 for the mixed set (consisting of trimers, hexamers and dodecamers in equal proportions). These datasets were randomly partitioned between training and test set according to a 75/25 ratio. All networks used the binary cross-entropy loss function and AdamW optimizer,[33] with learning rates ramping down from 0.1 to 1.0×10^{-4} as training convergence proceeded and a decay rate of 1×10^{-4}. All neural networks were implemented using PyTorch.[34] Additional, non-deep learning tasks (decision trees, random forests and ADABoost classifications) were performed using Scikit-Learn.[35]

Results and discussion

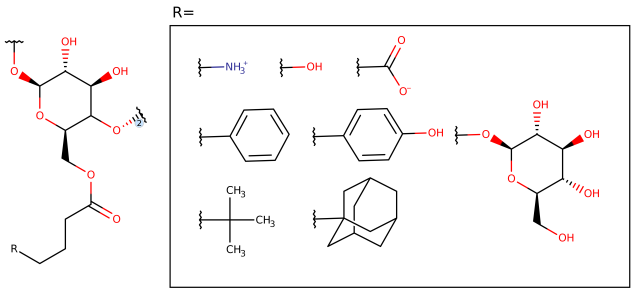

Figure 1: Markush structures of the mimics under study. Left: glucopyranose monomer template; an ester linker connects position 6 on the glucopyranose ring to one of 8 possible functional groups (right panel). Consecutive monomers are linked via $\alpha-1,4$ glycodidic bonds.

Free energy landscapes of oligosaccharides from enhanced sampling simulations. The oligosaccharide mimics were built from substituted glucopyranose monomers assembled along $\alpha-1,4$ bonds. Position 6 on each monomer was subtituted with a 4-carbon ester linker bearing one of 8 possible groups. These were chosen as synthetically amenable isosteres of the different types of amino acid side chains (charged, polar neutral, H-bonding, aromatic, aliphatic, bulky). A glucopyranose substituent was also included to allow the potential reticulation of oligomers (figure 1). All 512 possible trimer combinations of these 8 monomers were constructed. The number of possible combinations for longer oligosaccharides makes it computationally intractable to simulate them all on the microsecond time scale. Thus, a subset of 369 hexamers and 75 dodecamers, with equal representations of all 8 substituents and featuring possibly important patterns (repetitions of 2 and 3 identical substituents), were handpicked for simulation. All selected oligosaccharides were simulated for at least 1 µs, and the resulting trajectories were subjected to dPCA analysis. The first two eigenmodes for each oligomer were used as collective variables in metadynamics simulations of up to 6 µs, revealing the conformational free energy landscapes of the oligomers (see Methods for details).

The free energy minima on all trimer surfaces were identified and classified in terms of stability (free energy difference to the lowest minimum) and topological persistence[36] (relevance of the minimum compared to neighboring ones, evaluated by the height of the barrier separating them: considering a "water level" continuously rising on the free energy surface, how long would it take for both minima to belong to the same "lake"? See Supporting Information Available for details). Figure 2 shows that the minima can be classified into three clusters with respect to these two measures. (i) Low-energy and high-persistence minima are the representatives of the main attraction basins on the surface; they are often surrounded, within each basin, by (ii) numerous other minima of similarly low energies but low persistence (low energy barriers to neighboring minima). Finally, (iii) high-energy regions also feature local minima; they usually correspond to metastable states delimited by low free energy barriers, thus have low persistence.

Computing the average minimum stability and persistence for each trisaccharide reveals substituent-dependent trends (figure 2). Small, charged and/or H-bonding groups (OH, COO^-, $NH3^+$) create strong interactions with very spe-

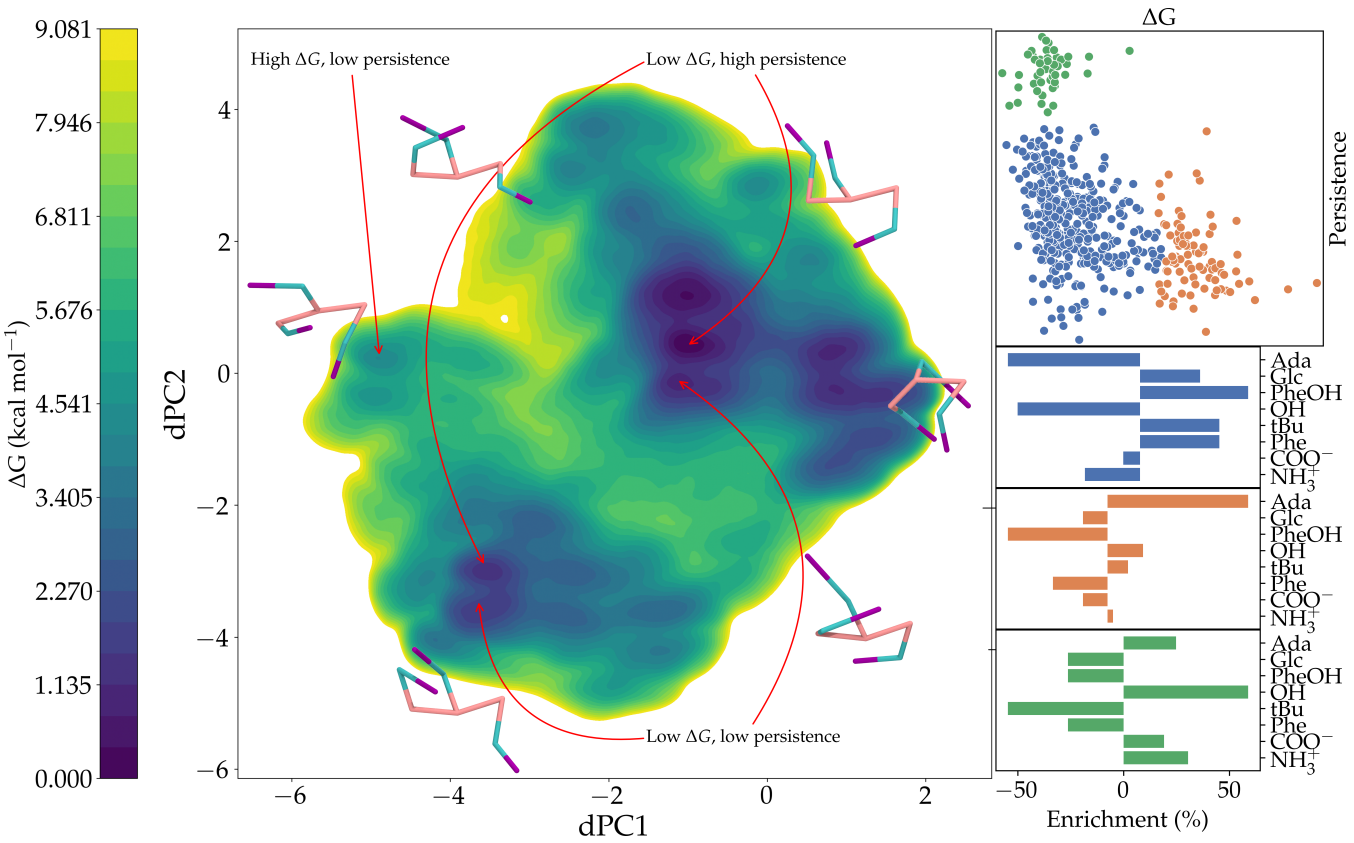

Figure 2: Left: free energy surface of an example trisaccharide (substituent sequence COO^--Phe-Glc) along the two first dPCA eigenmodes, showing the different types of minima in terms of stability (ΔG to the lowest minimum) and persistence (local relevance) as well as cartoons of the corresponding conformations (the centroids of glucopyranose rings, linker and substituent are respectively colored pink, cyan and purple). Upper right: normalized values of stability and persistence averaged over the minima of each trisaccharide, showing the division into three clusters. Lower right: enrichment or depletion of the three clusters in the different substituent types.

cific geometries; trimers bearing such groups tend to have a large number of highly stable, well-separated (highly persistent) minima (green cluster on figure 2). Interestingly, bulky adamantane groups also tend to generate such well-defined minima, but simultaneously favor high-energy, low-persistence metastable minima (orange cluster). This is the signature of Van der Waals interactions: very crowded oligosaccharides with several adamantane groups have specific minima separated by high barriers which sample the strongly repulsive part of the Lennard-Jones potential; on the other hand, in less constrained trimers, adamantane interacts weakly via the dispersive part of the Lennard-Jones potential, resulting in low free energy barriers and shallow minima. Finally, trimers with aromatic groups tend to present extensive attraction basins containing multiple minima separated by low free energy barriers (low ΔG, low persistence – blue cluster): due to limited steric effects and interaction strengths, such systems are less conformationally constrained. Surprisingly, trimers bearing glucose substituents tend to behave similarly despite the bulk and hydrogen-bonding capacities of the latter; this could be due to the coexistence of multiple simultaneous interactions, not all of which need to break when transitioning from a local minimum to its close-lying neighbor on the free energy surface. Examples of trimer free energy surfaces representing these various cases are provided on Supporting Information Available figure S2. Because the dPCA eigenvectors are a complex mixture of individual angles, translating the relative positions of the minima on the free energy surface in terms of conformational differences is difficult, except for very close-lying local minima within a super-

basin which do tend to show a degree of similarity. This can be verified from the local minimum structures on figure 2: the only detectable trend is a loose correlation of the first eigenvector with trimer compacity. Furthermore, the eigenvectors vary considerably from one trimer to the next (the average inner product between eigenvectors of distinct trimers is 0.32 ± 0.12 for the first eigenvector and 0.31 ± 0.13 for the second); the 2D free energy surfaces thus originate from completely different 'slices' of conformational space, making the comparison between surfaces rather futile. These considerations are excellent arguments in favor of using neural networks to derive simplified models of the free energy landscapes: their 'black box' nature isn't really a drawback when applied to such an abstract dataset.

Figure 3 compares the free energy surfaces of trimers to these of hexamers and dodecamers in terms of ΔG and persistence. As previously observed on figure 2, the distribution of ΔG values for trimers is bimodal: predominant minima within the main attraction basins are located less than $2\,\mathrm{kcal\,mol^{-1}}$ above the global minimum, while metastable states are centered around $5\,\mathrm{kcal\,mol^{-1}}$. The range of populated persistence values (discounting the global minima, which by definition have a persistence of 255 – see Supporting Information Available) extends above 100, suggesting multiple, distinct, very stable local minima. When the oligomer length increases, the free energy surfaces become more complex; individual minima tend to fuse into larger attraction basins, shifting the distribution of persistences toward lower values. Similarly, the distributions of ΔG values for longer oligomers progressively lose the bimodality seen in trimers; this is especially apparent for dodecamers, which also feature a much larger proportion of metastable minima. This can be ascribed to the much higher number of possible combinations of simultaneous interactions in such larger systems.

Learning trisaccharide conformational preferences. This initial comparison suggests a complex evolution of the oligosaccharide free energy surfaces when moving from trimers to longer oligomers. Will deep learning methods, known to reliably capture complex trends, reveal collective effects linking the dynamics of larger mimics to that of their smaller constituents? To test this hypothesis, a neural network (termed "trimer classifier") was built and trained to predict whether a trimer of given conformation and substituent sequence is stable (i.e., located in a zone of low relative free energy surrounding one or multiple local minima on the free energy surface). If this network can achieve sufficient performance, multiple instances of it could possibly be combined to predict the stability of longer isomer conformations, based on the sequences of their constitutive trimer patterns.

The trimer classifier network has a fully connected architecture which is represented on figure 4). The input layer comprises 156 input nodes, consisting of (i) embedding nodes which encode the nature of each substituent on the trimer into an 8-dimensional latent space; (ii) fully connected input nodes which hold the sine and cosine values of all relevant dihedral angles of the trimer (see Methods for details). The 156-dimensional vector representing a trimer is passed through 6 fully connected neuron layers with nonlinear ReLU activation functions. The final sigmoid function constrains the output of the network to the $]0, 1[$ range, interpreted as the probability of the input trimer to be representative of a low-lying free energy minimum.

The network was trained for 500 epochs on a training set of $\sim$ 77 million conformations from the metadynamics simulations of all possible trimers, and tested on a separate test set of $\sim$ 26 million conformations (see Methods for details). The learning process converged within 300 epochs (see Supporting Information Available), and achieved near-perfect predictions on the test set (sensitivity 0.96, specificity 0.98, precision 0.97, accuracy 0.97, F1 score 0.96). The network was thus successful in capturing the substituent-induced variety of free energy surface topologies and achieving predictive power on unlearnt samples. Beyond the topology of the network itself, the quality of the training set plays a crucial role: the enhanced sampling simulations performed in this work provide a more thorough description of

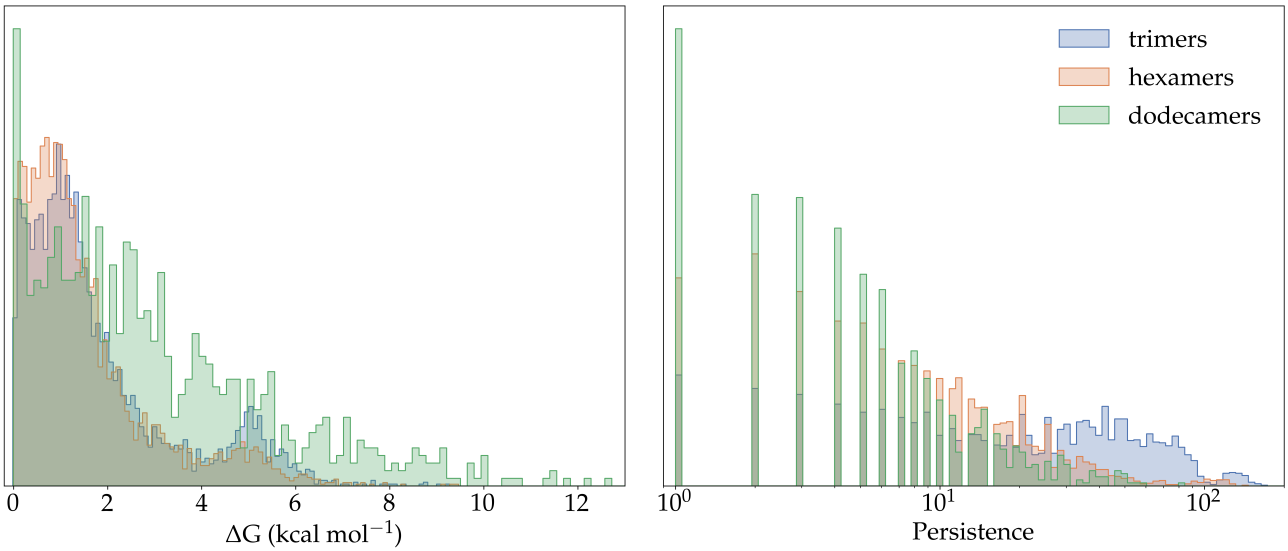

Figure 3: Histograms of the relative stability (left) and persistence (right) of all trimer, hexamer and dodecamer minima.

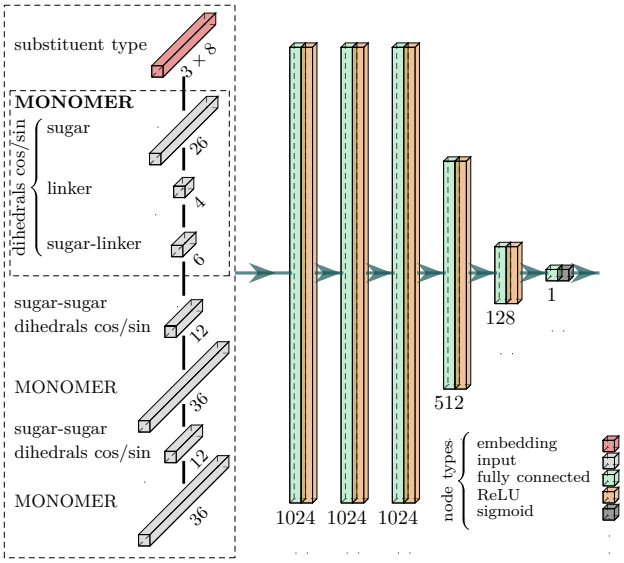

Figure 4: Fully connected classifier network used to learn and predict whether a trisaccharide of given conformation and substutient sequence belongs to a stability basin on the corresponding free energy surface.

conformational space and more accurate free energy estimates (especially for less-populated minima) than Boltzmann-statistics simulations of equivalent computational cost.

Onward to longer oligomers: learning collective effects. To investigate the existence of possible collective effects in longer oligosaccharide mimics, I then examined whether a relationship could be found between the status of a hexamer or dodecamer conformation as a minimum, and the corresponding status of its constituent trimers. These trimers are defined by sliding a window encompassing 3 sugar units over the oligosaccharide with a stride of one unit; an oligomer of length n thus comprises $n-3+1$ individual trimers. For hexamers, a simple decision tree based on the predictions of the 4 constituent trimers as minima or non-minima by the classifier network did not achieve an acceptable performance (F1 = 0.67). In particular, 58 % of conformations predicted as minima were false positives; the prediction of negatives fared somewhat better (23 % of false negatives). More complex methods, using the continuous probability output from the trimer classifier network rather than the binary mininum/non-minimum classification derived from it, did not provide better results: the ADABoost method yielded 66 % of false positives and 24 % of false negatives for both test and training sets (F1 $\approx$ 0.6); a random forest of 10 decision trees yielded a near-perfect classification of the training set (false predictions<1 %, F1 = 0.98), but achieved this result using very large trees with nearly as many leaves as data points and proved incapable of categorizing the test set (45 % false positives, 48 % false negatives, F1 = 0.47). For dodecamers, predictions were strongly biased toward false negatives: a

simple decision tree on the binary predictions of the 10 constitutive trimers yielded 99 % of structures predicted as non-minima; ADABoost on the continuous probabilities output by the trimer classifier yielded similar results for both training and test sets, and a random forest reproduced the training set perfectly but predicted 98 % of test set samples as non-minima regardless of their true nature.

From these performances, which are worse or barely better than those of a random classifier, it appears that trisaccharides bonded together within a larger oligomer retain no memory of their individual conformational preferences. While strong collective effects were expected, the extent of their domination of the conformational space of oligomers is striking. The occurrence of folded conformations in longer oligomers, driven by intramolecular interactions between monomers that are distant along the chain but spatially close, is a typical information that cannot be inferred from the conformations of isolated trimers; however, the respective over- and underprediction of hexamer and dodecamer minima suggest that additional, more complex collective effects on different scales of polymer length might actually coexist. These results are in line with those of previous conformational studies of polysaccharides (relatively scarce considering the importance of these polymers as natural products). In 2002, Rosen at al[37] failed to correlate the preferred conformations of oligosaccharides with those of the pentasaccharide repeating motifs on which they were built; however, their work did not include intrasaccharide degrees of freedom and defined stable conformations as MM3 potential energy minima, neglecting entropic effects which are essential in flexible oligomers. Two decades later, a state-of-the art study by Watanabe et al[38] mapped the conformational ensembles of high-mannose-type oligosaccharides obtained from MD simulations by projecting their ~ 100 major internal degrees of freedom into a kernel Hilbert space of dimensions up to 4 and grouped their free energy minima into 21 clusters; however, within each cluster, major differences in monomer ring puckering states, glycosidic linkages, interresidue hydrogen bonds and end-to-end distances remained – a testimony to how little the understanding of collective effects in oligosaccharides has really progressed in twenty years.

Convolutional deep neural networks are efficient detectors of multiscale collective patterns. These networks, popularized by their ability to identify objects in real-life images regardless of their scale and position, are now applied to datasets of very diverse types and origins.[39] They typically contain a sequence of convolutional layers of decreasing dimensions. Convolutional layers operate within a window of chosen size, which is slid over the input vector with a chosen stride. Input data contained inside the window are convoluted by a number of teachable convolution kernels, which react to collective input patterns spanned by the window. The output dimensionality depends on the number of possible window positions but is typically smaller than the input dimension. Pooling layers can also be used to decrease this dimensionality, by averaging or taking the maximum value of inputs inside the window. Each layer aggregates and convolutes the outputs of the previous one using its own sliding window: as information flows along the network and the dimensionality of the layers progressively decreases, the layers thus achieve a synthetic view of increasingly long spans of the original input vector, and the patterns they detect become more and more global.

The concept of convolutional networks appears perfectly suited to the identification of potentially multiscale collective effects in the conformational spaces of oligosaccharides. For instance, by using a window length of 3 and a stride of 1, the previous assumption of basing the conformational behavior of oligosaccharides on that of their constituent trimers can be replicated. However, unlike the previously performed simple aggregation of each trimer's likelihood as a minimum, the recursive convolution of individual trimer patterns operated by the successive layers allows the detection of complex collective patterns spanning the length of the entire oligosaccharide chain; these include, in longer oligomers, the occurrence of folded conformations stabilized by intramolecular in-

teractions.

However, unlike most datasets where all entries are represented by fixed-size vectors, oligomers of varying length are encoded with varying dimensionalities, which conditions both the size of the network input layer and the number of convolutional/pooling steps required to bring down the dimensionality. To tackle this issue, the convolutional network designed in this work was built around recursive components (figure 5). The input stage consists of two distinct sets of fully connected layers, respectively taking as input the dihedral angles of a monomer (sugar and substituent) and a connector (angles involving atoms connecting two consecutive sugars). These subnetworks are called as many times as needed depending on the polymer length N, and their outputs concatenated into a vector of size l_N. The latter is fed into a recursive convolutional subnetwork, featuring two convolutional layers which reduce the width of the data flow from l_i to l_{i-1} at each iteration. This subnetwork is recursively called $N-2$ times, until its output reaches size l_2. At this point, a fully connected subnetwork is applied, which outputs the probability for the input conformation to be a free energy minimum. Because (i) all monomers share the same fully connected modules regardless of their position in their containing oligomer and the length of the latter and (ii) because patterns in longer oligomers are detected by calling the same convolutional subnetwork multiple times, the network is never aware of the size of the oligomers it is being trained on. This prevents it from taking the 'easy way out' of learning trends that are specific to an oligomer length; instead, it is forced to find more global patterns, which should enable it to better extrapolate to oligomer length it has not been trained upon.

The recursive convolutional network was trained for 600 epochs on a mixed set of ~ 73 million trimer, hexamer and dodecamer conformations in equal proportions (see Methods). The training convergence was reached within 250 epochs (see Supporting Information Available). Interestingly, the trained network performed much better on hexamers and dodecamers than on trimers: the global F1 score of 0.90 can be decomposed into respective scores of 0.96, 0.92 and 0.83 for these three oligomer lengths. The prediction of trimers performs well in terms of specificity and accuracy, but less so in terms of sensitivity and precision. This is caused by true positive and false negative prediction rates which are respectively low and high compared to true negatives and false positives: the network tends to predict as non-minima trimer structures which are actually minima. While not optimal, in practical use this is preferable to flagging as minima conformations which aren't. Hexamer and dodecamer minima are very well predicted; the somewhat lower F1 score for dodecamers is counterbalanced by excellent accuracy and specificity scores which prove the network's ability to avoid minima overprediction. The origin of the performance disparities between trimers, hexamers and dodecamers isn't obvious; because performance does not vary monotonously from short to long polymers, it does not involve the fact that longer polymers benefit from a deeper network (the recursive layers being invoked multiple times). In fact, adding convolutional layers to the recursive subnetwork was tried and did not result in a noticeable increase in performance.

Generating stable olisaccharide conformations on the fly. The recursive convolutional network has proved able to identify stable oligosaccharide conformations regardless of length. However, in the use-case of high-throughput docking simulations, the ability to suggest such conformations on the fly without the need for costly molecular dynamics simulations is desirable. Therefore, a generative adversarial network (GAN) was built and trained to suggest potential free energy minimum structures for oligosaccharides of any length and substitution (figure 6). It consists of two subnetworks: (i) a generator, which takes as input a point in a low-dimensional 'latent space' and generates a corresponding minimum structure; (ii) a discriminator, which takes such generated conformations as input and tries to distinguish them from actual minima. Both subnetworks are trained simultaneously, with opposing goals: the discriminator is fed a train-

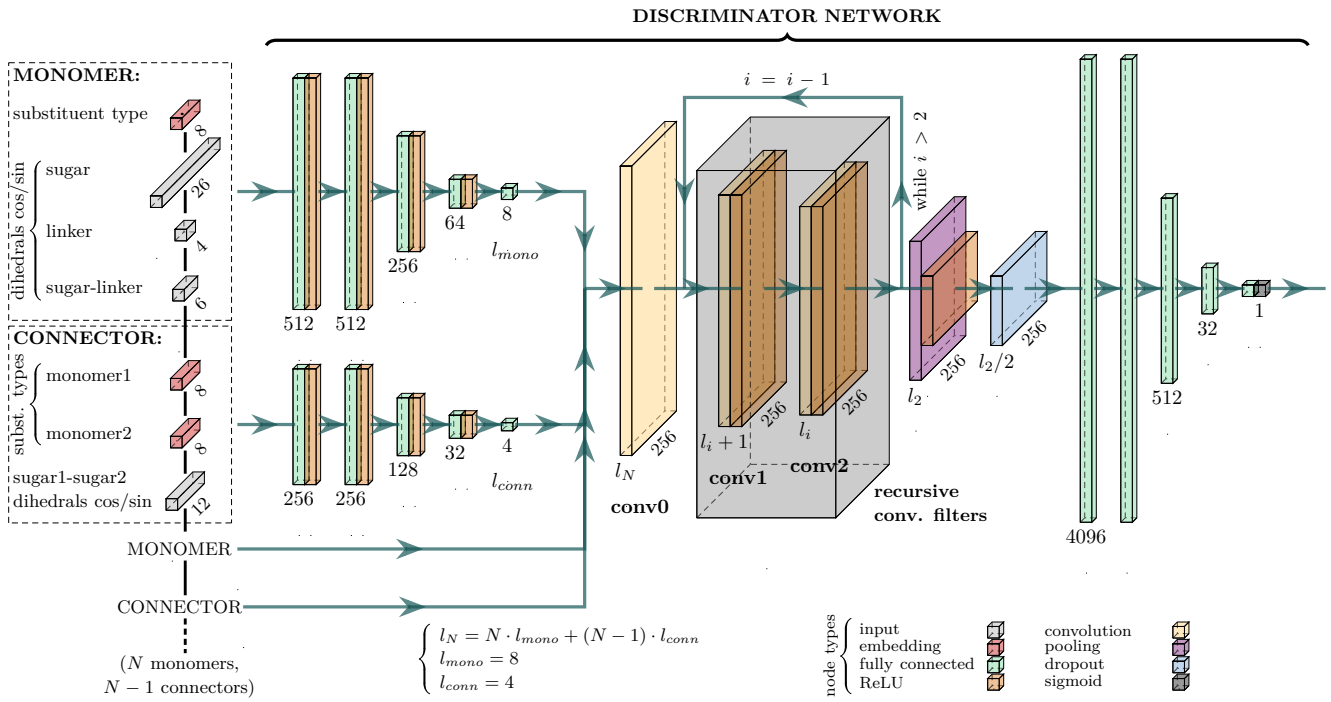

Figure 5: Recursive convolutional classifier network, designed to predict whether an oligosaccharide of given length, substituent sequence and conformation belongs to a free energy minimum basin. The kernel size and stride values for the two layers of the recursive convolutional module are $(l_{mono} + l_{conn}, 1)$ and $(2, 1)$.

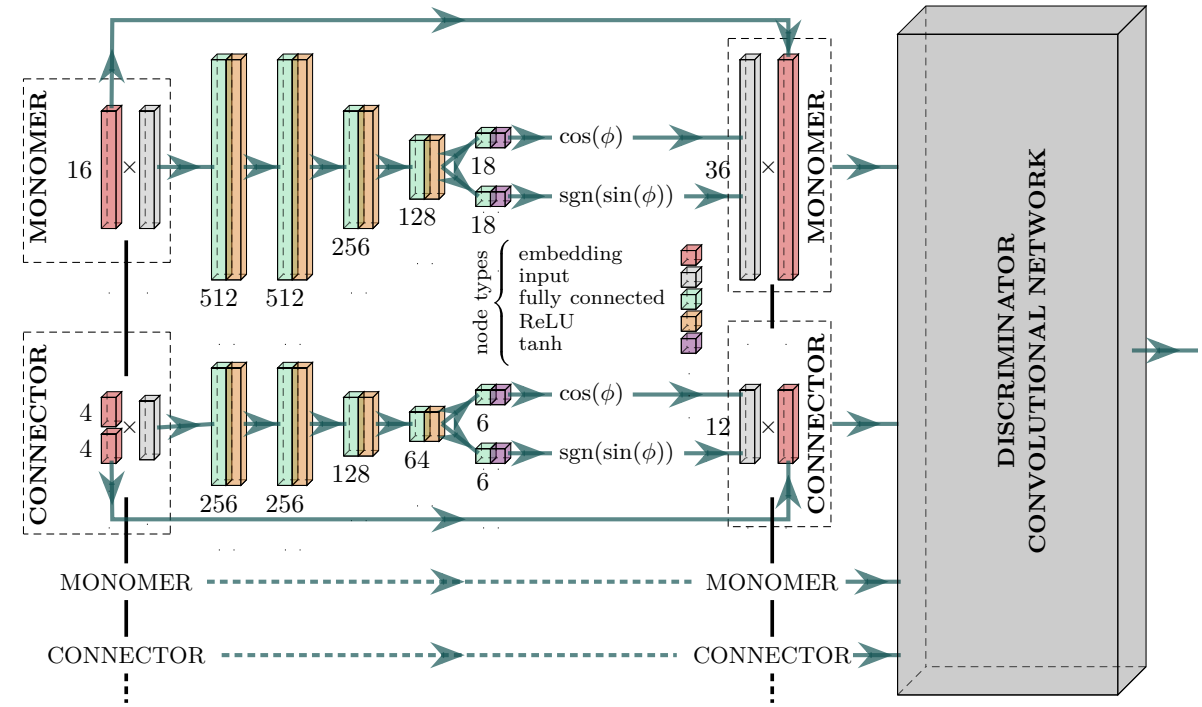

Figure 6: Generative adversarial network. The discriminator subnetwork is identical to the recursive convolutional network on figure 5.

ing set mixing actual minima and generated conformations and gets better at distinguishing one from the other, while the generator gets better at 'fooling' the discriminator by creating ever more realistic conformations. The convergence of this unsupervised learning process is achieved once generator and discriminator reach a stalemate[40]. The capacity of a trained generator to create realistic instances of complex data has been applied to many fields, including bioinformatics[41] and molecular design.[42] Here, the discriminator network employs the previously validated recursive convolutional network. For the generator, a monomer and a connector subnetworks of fully connected neurons are respectively called N and $N-1$ times to generate an N-mer conformation; their output is contrained to the $[-1, 1]$ range using hyperbolic tangent nodes and interpreted as the $(cos\ \phi, sign(sin\ \phi))$ values of all relevant dihedral angles Φ. Latent space dimensions of 16 and 8 for the monomer and connector subnetworks were found to adequately balance generator performance and computational cost. The oligomer substituent sequence was fed into both generator and discriminator networks using embedding nodes, which technically makes the network a conditional GAN;[43] for the generator, the output of the embedding nodes directly multiplies the latent space input.

The GAN was trained for 800 epochs on a set of $\sim$ 27 million trimer, hexamer and dodecamer minima. From 400 epochs on, the binary cross-entropy loss for both generator and discriminator oscillated repeatedly and periodically between two values, indicating that a stalemate between both had been reached in which each subnetwork continually countermeasures the other's action. There is no theoretical guarantee that a given GAN can achieve simultaneous stabilization of both subnetworks, and in practice this is seldom the case even in toy systems[44]; the learning process was thus considered complete, having reached a state of dynamic equilibrium (see figure S5 for additional detail). The generator weights corresponding to the smallest encountered loss were retained; the generator was then decoupled from the discriminator and used to suggest the relevant dihedral angles for 100 stable con-

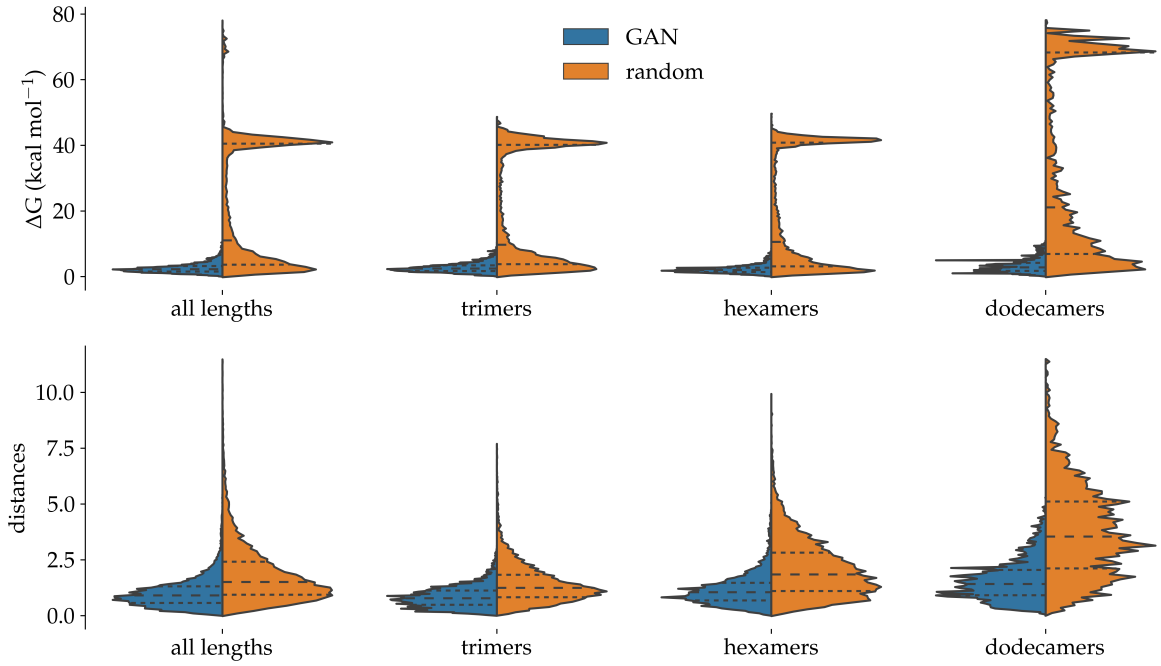

Figure 7: Performance of the GAN for the suggestion of stable conformations of random oligomers, compared to randomly selected conformations. Top row: distribution of ΔG above the global minimum. Bottom row: distribution of Euclidean distances to the closest minimum in the plane spanned by the first two dPCA eigenvectors. Quartiles are indicated by dashed lines.

formations of each of the oligomers under study (see Supporting Information Available for the procedure used to generate conformations from the dihedral angle values).

The performance of the trained generator subnetwork is illustrated on figure 7 by comparing the distributions of free energies and distances to the closest minimum on the surface between randomly chosen conformations and those suggested by the GAN. The distribution of generated free energies for all oligomer lengths has an average of $2.2\,\mathrm{kcal\,mol^{-1}}$ and an upper quartile of $3.3\,\mathrm{kcal\,mol^{-1}}$, in stark contrast to the random distributions in which very high-lying conformations ($>40\,\mathrm{kcal\,mol^{-1}}$) appear prominently. Similarly, the distribution of distances to the nearest minimum for the generated conformations efficiently filters out the high-lying entries in the random distribution (with respective upper quartiles of 1.3 and 2.5). Interestingly, the relative decrease in the spread of distance values is not as large as the one observed for free energies. This is due to the fact that many free energy surfaces feature extensive attraction basins containing multiple minima separated by low barriers: a conformation can thus be relatively structurally distant from a minimum and still belong to the latter's basin.

These statistics prove the generator's ability to generate sensible conformations 'on the fly' that can be used in the subsequent steps of a pipeline: for instance, by populating a library of potential oligosaccharide mimics which will be docked to a given protein target. Indeed, we are currently using the neural networks in this study to build a library containing hundreds of thousands of mimics; it will be queried using a graph representation of the nature and relative position of the hotspots at the target PP interface, rapidly providing the most likely candidates for the inhibition of the corresponding PP complex. With the training set generation and learning protocols validated, new monomer susbtituent types of specific interest can be added straightforwardly, by updating the networks' training based on additional molecular dynamics simulations.

Concluding remarks

A self-consistent model of polysaccharide conformational preference remains in the future. However, two avenues for improvement can help bring this goal closer: (i) enrich the corpus of accumulated conformational data on these molecules, which is still direly underdevelopped considering the natural relevance of oligosaccharides, and (ii) devise models to bridge the gap between local and global structural descriptors, which remains the stumbling block of current studies. The present work contributes to both avenues by providing extensive, long-timescale, enhanced sampling all-atom simulations with accurate conformational free energy estimations, on which the ability of conformational deep learning to detect multiscale spatial patterns is leveraged. It shows that the conformational behavior of longer oligosaccharides can indeed be inferred from their smaller constituent spans, but also demonstrates that the relationship is far from trivial, being dominated by collective effects on spans of different sequence lengths. In addition, the capacity to rapidly suggest potentially stable conformations of oligosaccharides of given lengths and substitutions using GANs is very valuable for the population of libraries of mimics, which should prove beneficial for the rapid and easy preselection of possible inhibitors of PP complexes of interest.

Acknowledgements

The calculations presented herein were performed using HPC resources from GENCI-IDRIS (Grant 2021-A0090711969) and the UPJV MatriCS computing platform.

Data and Software Availability

The following data is provided for download at https://extra.u-picardie.fr/nextcloud/index.php/s/52t2DHYDgswEJ3w: topology files for all mimics; 3D structures of all minima; free energy landscapes along the first two angular principal components for all systems; source files for the dPCA PLUMED module and the

neural networks; weights and biases for the trained networks. The full molecular dynamics trajectories for all systems (2 TB of data) are available upon request.

Supporting Information Available

Simulation details; procedure for identifying free energy minima; procedure for generating conformations from GAN outputs; convergence of learning processes.

References

(1) Wells, J. A.; McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* **2007**, *450*, 1001–1009.

(2) Reichmann, D.; Rahat, O.; Cohen, M.; Neuvirth, H.; Schreiber, G. The molecular architecture of protein–protein binding sites. *Curr. Opin. Struct. Biol.* **2007**, *17*, 67–76.

(3) Bouvier, B.; Grünberg, R.; Nilges, M.; Cazals, F. Shelling the Voronoi Interface of Protein-Protein Complexes Reveals Patterns of Residue Conservation, Dynamics, and Composition. *Proteins* **2009**, *76*, 677–692.

(4) Aumentado-Armstrong, T. T.; Istrate, B.; Murgita, R. A. Algorithmic approaches to protein-protein interaction site prediction. *Algorithms Mol. Biol.* **2015**, *10*, 7.

(5) Macalino, S. J. Y.; Basith, S.; Clavio, N. A. B.; Chang, H.; Kang, S.; Choi, S. Evolution of In Silico Strategies for Protein-Protein Interaction Drug Discovery. *Molecules* **2018**, *23*, 1963.

(6) Nag, S.; Baidya, A. T. K.; Mandal, A.; Mathew, A. T.; Das, B.; Devi, B.; Kumar, R. Deep learning tools for advancing drug discovery and development. *3 Biotech* **2022**, *12*, 110.

(7) González-Muñiz, R.; Bonache, M. Á.; de Vega, M. J. P. Modulating Protein–Protein Interactions by Cyclic and Macrocyclic Peptides. Prominent Strategies and Examples. *Molecules* **2021**, *26*, 445.

(8) Li, R.; Zhu, L.; Liu, D.; Wang, W.; Zhang, C.; Jiao, S.; Wei, J.; Ren, L.; Zhang, Y.; Gou, X.; Yuan, X.; Du, Y.; Wang, Z. A. High molecular weight chitosan oligosaccharide exhibited antifungal activity by misleading cell wall organization via targeting PHR transglucosidases. *Carbohydr. Polym.* **2022**, *285*, 119253.

(9) Tyrikos-Ergas, T.; Fittolani, G.; Seeberger, P. H.; Delbianco, M. Structural Studies Using Unnatural Oligosaccharides: Toward Sugar Foldamers. *Biomacromolecules* **2019**, *21*, 18–29.

(10) Gruner, S. A. W.; Locardi, E.; Lohof, E.; Kessler, H. Carbohydrate-based mimetics in drug design: Sugar amino acids and carbohydrate scaffolds. *Chem. Rev.* **2002**, *102*, 491–514.

(11) Kovalszky, I.; Surmacz, E.; Scolaro, L.; Cassone, M.; Ferla, R.; Sztodola, A.; Olah, J.; Hatfield, M. P. D.; Lovas, S.; Otvos, L. Leptin-based glycopeptide induces weight loss and simultaneously restores fertility in animal models. *Diabetes Obes. Metab.* **2010**, *12*, 393–402.

(12) Banik, S. M.; Pedram, K.; Wisnovsky, S.; Ahn, G.; Riley, N. M.; Bertozzi, C. R. Lysosome-targeting chimaeras for degradation of extracellular proteins. *Nature* **2020**, *584*, 291–297.

(13) Illescas, B. M.; Rojo, J.; Delgado, R.; Martín, N. Multivalent Glycosylated Nanostructures To Inhibit Ebola Virus Infection. *J. Am. Chem. Soc.* **2017**, *139*, 6018–6025.

(14) Panza, M.; Pistorio, S. G.; Stine, K. J.; Demchenko, A. V. Automated Chemical Oligosaccharide Synthesis: Novel Approach to Traditional Challenges. *Chem. Rev.* **2018**, *118*, 8105–8150.

(15) Seeberger, P. H.; Werz, D. B. Automated synthesis of oligosaccharides as a basis for drug discovery. *Nat. Rev. Drug Discovery* **2005**, *4*, 751–763.

(16) Oligosaccharides in drug discovery. 2014; `https://www.glycomar.com/documents/Oligosaccharidesindrugdiscovery_2014_000.pdf`.

(17) Wang, J.; Zhang, Y.; Lu, Q.; Xing, D.; Zhang, R. Exploring Carbohydrates for Therapeutics: A Review on Future Directions. *Front. Pharmacol.* **2021**, *12*, 756724.

(18) Woods, R. J. Predicting the Structures of Glycans, Glycoproteins, and Their Complexes. *Chem. Rev.* **2018**, *118*, 8005–8024.

(19) Yamaguchi, T.; Sakae, Y.; Zhang, Y.; Yamamoto, S.; Okamoto, Y.; Kato, K. Exploration of Conformational Spaces of High-Mannose-Type Oligosaccharides by an NMR-Validated Simulation. *Angew. Chem. Int. Ed.* **2014**, *53*, 10941–10944.

(20) Frank, M.; Lutteke, T.; von der Lieth, C.-W. GlycoMapsDB: a database of the accessible conformational space of glycosidic linkages. *Nucleic Acids Res.* **2007**, *35*, 287–290.

(21) Imberty, A.; Pérez, S. Structure, Conformation, and Dynamics of Bioactive Oligosaccharides: Theoretical Approaches and Experimental Validations. *Chem. Rev.* **2000**, *100*, 4567–4588.

(22) Scherbinina, S. I.; Frank, M.; Toukach, P. V. Carbohydrate Structure Database oligosaccharide conformation tool. *Glycobiology* **2022**, *32*, 460–468.

(23) Leckband, D. E.; Israelachvili, J. N.; Schmitt, F. J.; Knoll, W. Long-Range Attraction and Molecular Rearrangements in Receptor-Ligand Interactions. *Science* **1992**, *255*, 1419–1421.

(24) Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *J. Comput. Chem.* **2008**, *29*, 622–655.

(25) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(26) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247–260.

(27) Case, D. A.; Aktulga, H. M.; Belfon, K.; Ben-Shalom, I. Y.; Berryman, J. T.; Brozell, S. R.; Cerutti, D. S.; Cheatham III, T. E.; Cisneros, G. A.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Giambasu, G.; Gilson, M. K.; Gohlke, H.; Goetz, A. W.; Harris, R.; Izadi, S.; Izmailov, S. A.; Kasavajhala, K.; Kaymak, M. C.; King, E.; Kovalenko, A.; Kurtzman, T.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Machado, M.; Man, V.; Manathunga, M.; Merz, K. M.; Miao, Y.; Mikhailovskii, O.; Monard, G.; Nguyen, H.; O'Hearn, K. A.; Onufriev, A.; Pan, F.; Pantano, S.; Qi, R.; Rahnamoun, A.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shajan, A.; Shen, J.; Simmerling, C. L.; Skrynnikov, N. R.; Smith, J.; Swails, J.; Walker, R. C.; Wang, J.; Wei, H.; Wolf, R. M.; Wu, X.; Xiong, Y.; Xue, Y.; York, D. M.; Zhao, S.; Kollman, P. A. Amber 2022. 2022; `https://ambermd.org`, University of California, San Francisco.

(28) da Silva, A. W. S.; Vranken, W. F. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Res. Notes* **2012**, *5*, 367.

(29) Mu, Y.; Nguyen, P. H.; Stock, G. Energy landscape of a small peptide revealed by

dihedral angle principal component analysis. *Proteins Struct. Funct. Bioinf.* **2004**, *58*, 45–52.

(30) Barducci, A.; Bussi, G.; Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **2008**, *100*, 20603.

(31) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1*, 19–25.

(32) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Comm.* **2014**, *185*, 604–613.

(33) Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2017**, 1711.05101.

(34) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., D'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc., 2019; pp 8024–8035.

(35) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(36) Huber, S. *Data Science – Analytics and Applications*; Springer Fachmedien Wiesbaden, 2021; pp 81–88.

(37) Rosen, J.; Robobi, A.; Nyholm, P.-G. Conformation of the branched O-specific polysaccharide of Shigella dysenteriae type 2: Molecular mechanics calculations show a compact helical structure exposing an epitope which potentially mimics galabiose. *Carbohydr. Res.* **2002**, *337*, 1633–1640.

(38) Watanabe, T.; Yagi, H.; Yanaka, S.; Yamaguchi, T.; Kato, K. Comprehensive characterization of oligosaccharide conformational ensembles with conformer classification by free-energy landscapes via reproductive kernel Hilbert space. *Phys. Chem. Chem. Phys.* **2021**, *23*, 9753–9760.

(39) Chen, L.; Li, S.; Bai, Q.; Yang, J.; Jiang, S.; Miao, Y. Review of Image Classification Algorithms Based on Convolutional Neural Networks. *Remote Sens.* **2021**, *13*, 4712.

(40) Alom, M. Z.; Taha, T. M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M. S.; Hasan, M.; Essen, B. C. V.; Awwal, A. A. S.; Asari, V. K. A State-of-the-Art Survey on Deep Learning Theory and Architectures. *Electronics* **2019**, *8*, 292.

(41) Chen, X.; Li, C.; Bernards, M. T.; Shi, Y.; Shao, Q.; He, Y. Sequence-based peptide identification, generation, and property prediction with deep learning: a review. *Mol. Syst. Des. Eng.* **2021**, *6*, 406–428.

(42) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.* **2019**, *4*, 828–849.

(43) Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, 1411.1784.

(44) Goodfellow, I. NIPS 2016 Tutorial: Generative Adversarial Networks. *arXiv* **2017**, 1701.00160.

Graphical TOC Entry