

Modeling Protein–Glycan Interactions with HADDOCK

Anna Ranaudo, Marco Giulini, Angela Pelissou Ayuso, and Alexandre M. J. J. Bonvin*



Cite This: *J. Chem. Inf. Model.* 2024, 64, 7816–7825



Read Online

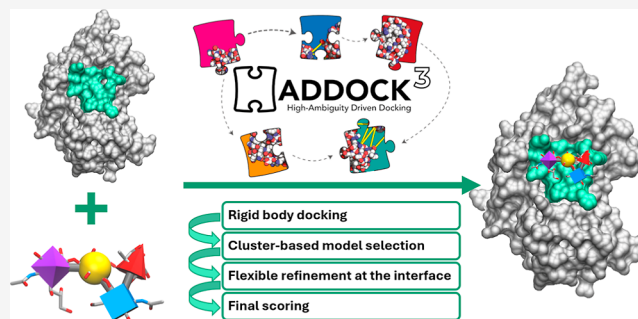
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: The term glycan refers to a broad category of molecules composed of monosaccharide units linked to each other in a variety of ways, whose structural diversity is related to different functions in living organisms. Among others, glycans are recognized by proteins with the aim of carrying information and for signaling purposes. Determining the three-dimensional structures of protein–glycan complexes is essential both for the understanding of the mechanisms glycans are involved in and for applications such as drug design. In this context, molecular docking approaches are of undoubted importance as complementary approaches to experiments. In this study, we show how high ambiguity-driven DOCKing (HADDOCK) can be efficiently used for the prediction of protein–glycan complexes. Using a benchmark of 89 complexes, starting from their bound or unbound forms, and assuming some knowledge of the binding site on the protein, our protocol reaches a 70% and 40% top 5 success rate on bound and unbound data sets, respectively. We show that the main limiting factor is related to the complexity of the glycan to be modeled and the associated conformational flexibility.



INTRODUCTION

Glycans are organic compounds with a polymeric structure consisting of monosaccharides, small building blocks covalently linked to each other in various ways through glycosidic bonds, in linear or branched arrangements. Depending on the number of monosaccharide units, glycans are named disaccharides (2 units), oligosaccharides (3–10 units), and polysaccharides (>10 units).¹ For simplicity, we use here the term “glycan” for referring to compounds with any number of monosaccharide units.

Glycans can form complex structures. This complexity already lies in monosaccharides themselves, which have a high degree of intrinsic chemical variability. The second source of complexity arises from the way monosaccharides are linked to each other as each glycosidic bond can form two possible stereoisomers at the anomeric carbon, i.e., the carbon whose asymmetric center is formed upon cyclization of the monosaccharide. Regioisomers may also exist because of the many hydroxyl groups, which allow two monosaccharides to be linked in several ways.² Monosaccharides, having the ability to create more than two glycosidic bonds, can give rise to branched chains. The frequent occurrence of branched patterns, as opposed to the linear patterns typically found in most peptides and oligonucleotides, contributes to the complexity of the glycans’ structural landscape. Besides the very large amount of glycans that can be built starting from a few monosaccharides, glycans show large conformational variability at room temperature due to the low torsional energy barriers around the glycosidic bonds.³

Glycans are ubiquitously found in living organisms, where they can be freestanding or linked to proteins and lipids, giving rise to glycoproteins and glycolipids, respectively. They are involved in a variety of biological processes that can be classified into three main categories.⁴ First, they can play structural roles by either assisting in creating extracellular scaffolds, such as cell walls and matrices, or by being involved in protein folding and function. Second, they can serve as crucial sources of energy metabolism. Third, they can play the role of information carriers, being recognized by glycan-binding proteins (GBPs). Glycans can bind proteins through non-covalent, reversible interactions for signaling purposes, initiating a range of biological processes in both plants and animals.⁵ The role of glycans has recently been highlighted in the context of the well-known SARS-CoV-2 spike protein. This protein, which enters host cells by connecting to the angiotensin-converting enzyme (ACE2), is surrounded by a layer of glycans, which hides it from the immune system. Casalino et al.⁶ showed that specific glycans play a crucial role in the movement and structure of the part of the spike protein that binds to ACE2. Their removal results in diminished binding to ACE2.

Received: August 2, 2024

Revised: September 23, 2024

Accepted: September 25, 2024

Published: October 3, 2024



It is evident that understanding the way glycans interact with proteins is of crucial importance. Experimental methods such as X-ray crystallography and cryo-electron microscopy face challenges when dealing with glycans because of their intrinsic flexibility and heterogeneity. Computational approaches, such as molecular docking, can provide an alternative, less expensive, and faster way of generating three-dimensional (3D) models of protein–glycan complexes.

Despite the growing attention that is being given to the carbohydrates field,^{7,8} the modeling of protein–glycan complexes by docking has only received limited attention compared to that of other biomolecular complexes. A few protocols have been developed in the past years, such as GlycanDock based on Rosetta,⁹ a heparin-specific protocol based on PIPER/ClusPro,¹⁰ Vina-Carb,¹¹ GlycoTorch Vina,¹² RosettaCarbohydrate,¹³ and ATTRACT.¹⁴ Recently, a new version of AlphaFold,¹⁵ AlphaFold3,¹⁶ that can handle glycosylated proteins, has been published. It, however, does not allow one to model protein–glycan complexes. State-of-the-art protein–ligand docking software encounters challenges in addressing the conformational variability of glycans as they are usually developed for dealing with small, more rigid molecules.⁹

In this study, we use high ambiguity-driven DOCKing (HADDOCK)^{17,18} to address the protein–glycan interaction prediction problem. HADDOCK is an information-driven docking approach that can harvest knowledge about binding sites to drive the docking process (see below). Note that HADDOCK¹⁸ has already been applied to glycan modeling^{19,20} but, to date, without an exhaustive benchmarking. Here, we first benchmark the baseline performance of HADDOCK in predicting the 3D structures of protein–glycan complexes on a bound data set composed of 89 high-resolution experimental complexes from the Protein Data Bank (PDB)²¹ assuming an ideal scenario in which the binding interfaces on both protein and glycan are known in order to drive the docking process. A protocol is then proposed to deal with a realistic scenario in which the bound conformations of the partners are unknown. The GLYCAM-Web web server^{22,23} is used for the generation of glycan unbound structures, while the unbound protein structures are retrieved from the PDB. Finally, to address the conformational flexibility challenge of glycans, we assess whether providing an ensemble of glycan conformations generated through a short conformational sampling carried out with HADDOCK prior to the docking process can improve the performance of the protocol, next to the standard semi-flexible refinement of the interface.

METHODS

Benchmark Data Set Preparation. HADDOCK's^{17,18} performance in reproducing the binding geometries of protein–glycan complexes was evaluated by exploiting an adapted version of the data set provided in GlycanDock.⁹ It is composed of 109 experimentally determined high-resolution (<2.0 Å) protein–glycan complexes collected from the PDB. By discarding the entries containing glycans not yet supported by HADDOCK (refer to <https://rascar.science.uu.nl/haddock2.4/library> for a list of supported glycans), a data set of 89 complexes was obtained, which will be referred to as the *bound* data set from now on. The protein receptors in this data set include 8 antibodies, 21 carbohydrate-binding modules, 18 enzymes, 27 lectins or glycan-binding proteins (GBPs), and 15 viral glycan-binding proteins. The length of the glycans ranges

from 2 to 7 monosaccharide units, 72 of which are linear glycans and the remaining 17 branched ones. This structural diversity is considered in the analysis of the docking performance.

With the aim of evaluating HADDOCK's performance in a realistic unbound docking scenario, a subset of 55 complexes (out of 89) was defined for which unbound protein forms were available in the PDB. The glycans' unbound conformations were generated with the GLYCAM-Web web server.^{22,23} This data set of 55 unbound conformations of both proteins and glycans will be referred to as the *unbound* data set from now on. It contains 47 linear and 8 branched glycans; 25 glycans are composed of three or fewer monosaccharide units, while 30 have more than three units. For the evaluation of HADDOCK's performance on the *unbound* data set, all the complexes including glycans made up of three or fewer monosaccharide units are treated together and referred to with the label **SL-SB** (short linear-short branched). For the larger *bound* data set, linear (**SL**) and branched (**SB**) short glycans are analyzed separately. The complexes including glycans composed of more than three units (**L** for long) are divided into linear (**LL**, 23 cases) and branched (**LB**, 7 cases). Glycans and protein structures were prepared for docking as described in [Supporting Information Text S1](#). Details on the two data sets are reported in [Supporting Information File 1](#), and the Symbol Nomenclature for Glycans^{24,25} (SNFG) is given in [Supporting Information Table S1](#). All HADDOCK-ready bound and unbound conformations, together with restraints, HADDOCK configuration files, and analysis scripts, are provided in the following GitHub repository: <https://github.com/haddocking/protein-glycans>.

HADDOCK General Protocol and Scoring Function.

Docking calculations were performed with HADDOCK3 (<https://github.com/haddocking/haddock3>, DOI:10.5281/zenodo.10527751), the new, modular version of the well-established HADDOCK 2.X software.²⁶ The original HADDOCK protocol consists of three successive steps: (i) full randomization of the orientations and docking by rigid-body energy minimization; (ii) semi-flexible refinement by simulated annealing in torsion angle space during which the interfaces are considered flexible; and (iii) final refinement, either by energy minimization (current default) or by a short molecular dynamics simulation in explicit solvent. HADDOCK3 overcomes this rigid workflow structure as its constituent modules can be freely combined and interchanged by the user. A description of the HADDOCK3 modules used in this study is given in [Supporting Information Text S2](#).

HADDOCK's scoring function (HS) includes the intermolecular electrostatic (E_{elec}) and van der Waals (E_{vdW}) energies (calculated with the OPLS force field²⁷), an empirical desolvation energy term (E_{desolv}),²⁸ the buried surface area (E_{BSA}), and the ambiguous interaction restraint energy (E_{air}) (see below). The combination and weights of those terms depend on the stage of the protocol. In the present study, we compare the default scoring function at the rigid-body stage (eq 1) with one in which the weight of the van der Waals energy term was increased from 0.01 to 1.0 (eq 2) as recommended for small-molecule docking with HADDOCK.²⁹ The scoring function in eq 2 will be referred to with the name *vdW* to be distinguished from the *default* one. For the flexible refinement stage, the default scoring function is used (eq 3).

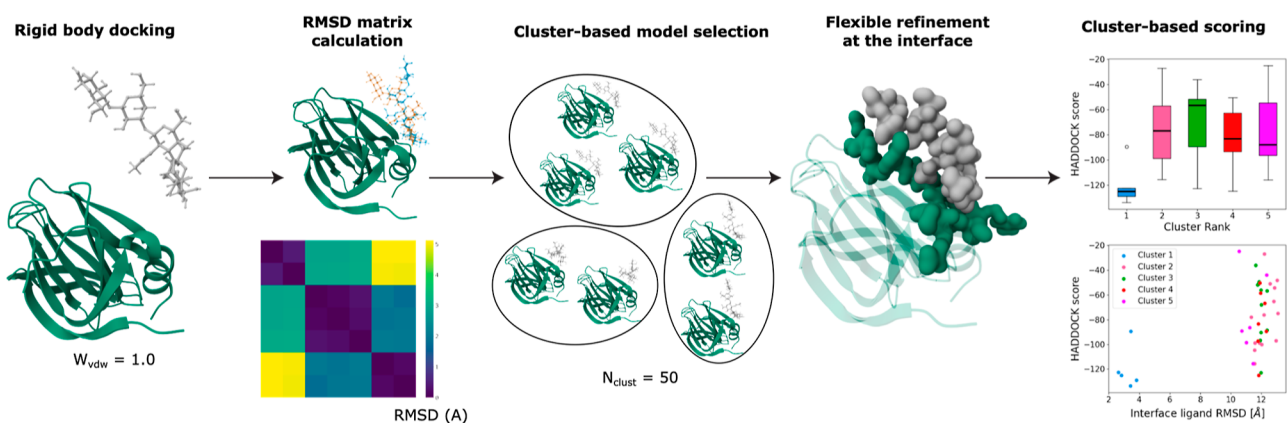


Figure 1. Schematic representation of the docking protocol for the *unbound* data set. First, rigid-body docking is performed. The models are then clustered based on RMSD. The best-scoring models of each cluster are then subjected to a flexible refinement (interface), and the resulting models are again clustered and analyzed.

HS(rigid-body default)

$$= 0.01E_{\text{vdw}} + 1.0E_{\text{el}} + 1.0E_{\text{desolv}} + 0.01E_{\text{air}} - 0.01E_{\text{BSA}} \quad (1)$$

HS(rigid-body vdW)

$$= 1.0E_{\text{vdw}} + 1.0E_{\text{el}} + 1.0E_{\text{desolv}} + 0.01E_{\text{air}} - 0.01E_{\text{BSA}} \quad (2)$$

HS(flexref default)

$$= 1.0E_{\text{vdw}} + 1.0E_{\text{el}} + 1.0E_{\text{desolv}} + 0.1E_{\text{air}} - 0.01E_{\text{BSA}} \quad (3)$$

Restraints to Drive the Docking. Experimental or predicted information about binding sites can be introduced as restraints for guiding the docking process and scoring the generated models (E_{air} in eqs 1–3). Interface information is typically introduced as ambiguous interaction restraints (AIRs),¹⁸ which are defined from lists of residues divided into two groups: active and passive. Active residues are those of central importance for the interaction. They are restrained to be part of the interface throughout the docking and refinement process; if an active residue is not part of the interface in a model, it generates restraint energy (and corresponding forces). Passive residues are those that could contribute to the interaction, being among the possible interaction partners of the active residues defined on the other molecules; if a passive residue is not in the interface of a given model, there is no restraint energy generated.

The functional form of the restraint energy function is similar to the distance restraining function introduced by Nilges,³⁰ namely, a flat bottom function which is harmonic for short violations and then transits to a linear form, thus ensuring stable force computations.

In this work, two scenarios in terms of AIRs were considered: (i) true-interface scenario (**ti-aa**), where active residues, corresponding to the interface residues within 3.9 Å^{31,32} from the partner, are defined for both the protein and the glycan; (ii) true-interface-protein–full glycan passive scenario (**tip-ap**), where active residues are still defined for the protein interface, but all residues of the glycan are listed as passive. By default, HADDOCK randomly discarded 50% of the defined AIRs for each docking model. This is done to

account for possible wrong information (false positives) in the experimental data.

Docking Protocols for Bound and Unbound Data Sets. Two different protocols were used for the *bound* and *unbound* data sets. For the *bound* data set, a workflow consisting of the following six modules was defined for running HADDOCK3 (details in Supporting Information Table S2 and the corresponding config file available in the GitHub repository):

- 1 topaaa: creation of the topologies of the two partners during which any missing atoms are automatically added.
- 2 rigidbody: AIR-driven generation of rigid-body models.
- 3 caprieval: models' quality analysis (see below).
- 4 rmsdmatrix: calculation of the RMSD matrix between all the models based on either all the interface residues (when **ti-aa** AIRs are used) or the protein interface residues and the whole glycan (when **tip-ap** AIRs are used).
- 5 clustrmsd: RMSD-based agglomerative hierarchical clustering of the models using the average linkage criterion and a distance cutoff of 2.5 Å. Only clusters containing four or more models are evaluated.
- 6 caprieval: cluster-based evaluation of the quality of the models.

As the starting structures are already in their bound conformations, no flexible refinement was performed.

For the *unbound* data set (details in Table S3 and the corresponding config file available in the GitHub repository), the workflow consisted of the following 12 modules:

- 1 topaaa: creation of the topologies of the two partners during which any missing atoms are automatically added.
- 2 rigidbody: AIR-driven generation of rigid-body docking models (1000 by default) with increased sampling (200 per conformation) when starting from an ensemble of conformations.
- 3 caprieval: models' quality analysis (see below).
- 4 rmsdmatrix: calculation of the RMSD matrix between all the models based on either all the interface residues (when **ti-aa** AIRs are used) or the protein interface residues and the whole glycan (when **tip-ap** AIRs are used).

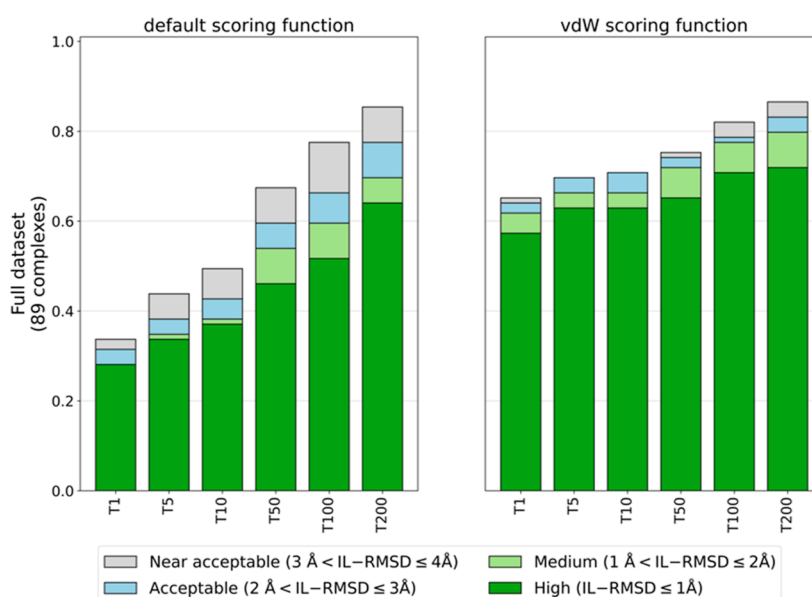


Figure 2. Comparison of bound docking success rates obtained with the *default* ($w_{vdW} = 0.01$) and *vdW* ($w_{vdW} = 1.0$) scoring functions (eqs 1 and 2) as a function of the number of top-ranked models ($T = 1, 5, 10, 50, 100$, and 200) selected using true interface residues of both protein and glycan to define the ambiguous interaction restraints (ti-aa AIRs).

- 5 clustrmsd: RMSD-based agglomerative hierarchical clustering of the models using the average linkage criterion. Here, 50 (150 when the ensemble of glycan conformations is used) clusters are created.
- 6 seletopclusts: selection of the top 5 models of the existing clusters.
- 7 caprieval: cluster-based evaluation of the quality of the models.
- 8 flexref: semi-flexible refinement through a simulated annealing protocol in torsion angle space in which first side-chains and then side-chains and backbone of interface residues are treated as flexible.
- 9 caprieval: models' quality analysis.
- 10 rmsdmatrix: calculation of the RMSD matrix between all the models, as in point 4.
- 11 clustrmsd: RMSD-based clustering of the models as in point 4 but here using a distance cutoff of 2.5 \AA as in the bound scenario.
- 12 caprieval: cluster-based evaluation of the quality of the models.

The workflow for the *unbound* data set with the RMSD clustering of the rigid-body models followed by semi-flexible refinement is schematically represented in Figure 1.

Model Quality Assessment. The quality of the models was evaluated using interface-ligand RMSD (IL-RMSD) with respect to the experimental structures. We did not use the fraction of native contacts (Fnat) as its values are typically quite high irrespective of the pairwise orientation between the molecules and thus unable to highlight small conformational differences, especially for small glycans. The IL-RMSD is calculated by first superimposing the model onto the reference structure by using the backbone atoms of the protein interface residues and then calculating the RMSD on the heavy atoms of the oligosaccharide. This is motivated by the fact that the protein interface is larger compared to the glycan interface and would dominate the RMSD calculation if the standard CAPRI³³ interface RMSD (I-RMSD) metric was used. The IL-RMSD gives a better measure of variations in the position

of the ligand compared to the I-RMSD. The cutoffs used to define the quality of the models based on IL-RMSD are

- high-quality models: IL-RMSD $\leq 1.0 \text{ \AA}$
- medium-quality models: IL-RMSD $\leq 2.0 \text{ \AA}$
- acceptable-quality models: IL-RMSD $\leq 3.0 \text{ \AA}$
- near acceptable-quality models: IL-RMSD $\leq 4.0 \text{ \AA}$

Supporting Information Figure S1 shows examples of docking models falling into these four categories for short- and long-linear glycans and for branched glycans, highlighting how the near acceptable threshold is only appropriate when dealing with long glycans, being too permissive for short monosaccharide chains.

HADDOCK3's performance is evaluated by calculating success rates (SRs), defined as the fraction of complexes having at least one high-, medium-, acceptable-, or near-acceptable-quality model among the top N models ranked according to the HADDOCK score. The cluster-based SR is also calculated by considering the top 4 scoring models within each cluster and assigning a given quality to a cluster if any of the top 4 members of the cluster reaches the corresponding quality cutoff. Note that the clustering step after rigid-body docking considers the 5 top models of each cluster.

Glycans' Conformational Sampling. Conformational sampling of the glycans was carried out with the HADDOCK3 water refinement module (mdref²⁶), starting from the models generated with the GLYCAM-Web web server.^{22,23} Different scenarios were tested in terms of the number of molecular dynamics integration steps and the number of models generated (each simulation starting with a different random seed). The RMSD of the generated models with respect to the bound glycan conformation was calculated with the rmsdmatrix module. The generated models were clustered using RMSD-based hierarchical clustering,^{34,35} requesting 20 clusters. The centers of each cluster were then used to define an ensemble of starting unbound conformations for docking. Further details of the conformational sampling can be found in Supporting Information Text S3.

RESULTS AND DISCUSSION

This section is structured as follows. First, the impact of the rigid-body scoring function on HADDOCK3's performance is discussed based on docking calculations performed on the *bound* data set. The performance of the docking is then analyzed considering structural features of the glycans (length and branching) and the definition of the AIRs. We then focus on the more realistic scenario of *unbound docking*, assessing the best way of selecting the rigid-body models to be refined and the impact of the flexible refinement on the quality of the final models. Finally, the impact and limitations of using an ensemble of glycans as starting points for the docking are discussed.

Bound Docking Performance and the Impact of the Rigid-Body Scoring Function in the Ranking of Models.

We first assessed the accuracy of the rigid-body scoring function in ranking the generated models by comparing the *default* scoring function (eq 1) to the one recommended for small ligands, in which the van der Waals energy weight is increased from 0.01 to 1.0 (*vdW* scoring function, eq 2). This was assessed by running docking calculations on the *bound* data set with true interface restraints (**ti-aa** AIRs). A comparison of success rates (SRs) obtained with the *default* (eq 1) and *vdW* (eq 2) scoring functions is shown in Figure 2.

The *vdW* scoring function with increased van der Waals energy weight ($w_{vdW} = 1.0$) performs much better than the *default* function ($w_{vdW} = 0.01$) with remarkably higher success rates. This highlights the importance of the van der Waals energy term in scoring protein–glycan models. For example, considering T1 and T5 high-quality models, the improvement is around 30%. Even when considering the top 200 models (the default number of models passed to the flexible refinement stage in a standard HADDOCK2.X workflow), a slight improvement is observed in both number and quality of acceptable or better models. Results are consistent if we consider the cluster-based success rate (see Methods), with 66% of top-ranked clusters falling into the acceptable category in the *vdW* settings, compared to 33% of the *default* scenario. Glycans, despite all of their polar groups, have quite hydrophobic properties, especially in their ring structure. A similar behavior was observed in previous work when docking cyclic peptides³⁶ and small ligands.²⁹ Based on these results, all subsequent docking calculations were performed with the *vdW* scoring function.

Impact of Glycan Structural Features and AIR Definition on the Bound Docking Performance. The dependency of the structural features of the glycans on the SR was then assessed using the true-interface AIR restraints (**ti-aa**) and *vdW* scoring function scenario. SRs were calculated grouping the complexes based on the size of the glycans (glycans composed of three or fewer units, labeled with **S**, or more than three, with **L**) and connectivity (linear, **L**, or branched, **B**). The SR, shown in the first column of Supporting Information Figure S2, indicates that HADDOCK3 performs better for long-linear (**LL**) and long-branched (**LB**) glycans. The bound SR for T1 high-quality models is 45, 17, 74, and 73% for **SL**, **SB**, **LL**, and **LB** glycans, respectively (50, 33, 79, and 82%, respectively, for T1 acceptable or better-quality models). When considering a larger number of models, the SRs become rather similar for the various types of glycans (70–80% for T50 and 80–90% for T200), except for the short-branched (**SB**) ones (~70%). The lower performance on

the short-branched glycans and, to a lesser extent, on the **SL** ones could be due to the fact that smaller ligands could be accommodated into the protein-binding site with a greater variability of positions and orientations, while the range of possible docking orientations is more restricted for larger ligands.

The impact of defining the glycan as passive (**tip-ap**) was then investigated. A slight decrease in the SR can be observed (second column of Supporting Information Figure S2), with respect to the **ti-aa** scenario, for **LB** glycans and, to a lesser extent, for **SB** ones. This can be explained by noting that most of the linear-long (**LL**) glycans (27/34) have all of their residues involved in the binding to the protein, whereas the same is true for only half (5/11) of the long-branched (**LB**) ones. Branched glycans are thus more difficult to model when no information about their interface is available. Overall, HADDOCK3's performance on protein–glycan complexes is not much affected by defining the glycan as passive (**tip-ap** AIRs). As experimental interface information for the glycans might not always be available in a realistic scenario, the remainder of the paper will discuss docking calculations only with **tip-ap** AIRs. Note that, for example, nuclear magnetic resonance (NMR) can provide specific information about which groups of a glycan are involved in the binding, as demonstrated for the modeling of the complex between sialic acid and the N-terminal domain of the SARS-CoV2 spike protein using NMR saturation transfer experiments.³⁷

Unbound Docking. In a realistic scenario, the bound conformations of the docking partners are unknown, and only unbound structures or models will be available. As such, conformational rearrangements may be required during binding. We therefore assessed HADDOCK3's performance on the *unbound* data set consisting of 55 complexes. The starting point for the docking was the unbound form of the protein taken from the PDB and models of the glycans obtained with the GLYCAM-Web web server^{22,23} (see Methods section *Benchmark Data Set Preparation*). When dealing with unbound structures, the flexible refinement stage of the workflow might allow some conformational changes to occur. As only a fraction (typically around 20%) of the models is subjected to flexible refinement, ranking and selection of models after the rigid-body docking stage become crucial. We tested two scenarios: (1) the "classical" scenario in which the top 200 ranked models are passed to the flexible refinement stage; (2) a cluster-based selection, made possible by the modular and flexible structure of HADDOCK3. In the second scenario, the rigid-body models are first clustered based on their RMSD similarity; then, the top 5 models of all 50 clusters (150 when the ensemble is used, see below) are selected. In this way, models that might not rank high enough to be selected in an individual model ranking might still get selected and subjected to the flexible refinement, thus increasing the diversity of refined models (potentially at the cost of decreasing the overall number of acceptable or better models in the final set of refined models, which is not an issue provided that the scoring function can identify the near native models). Since we are using an agglomerative hierarchical clustering method allowing the desired number of clusters to be defined, clusters with less than 5 models might be obtained, and for this reason, a different number of models might be subjected to the flexible refinement for each complex. The overall docking performance and the performance by glycan size and branching are shown in Figure 3.

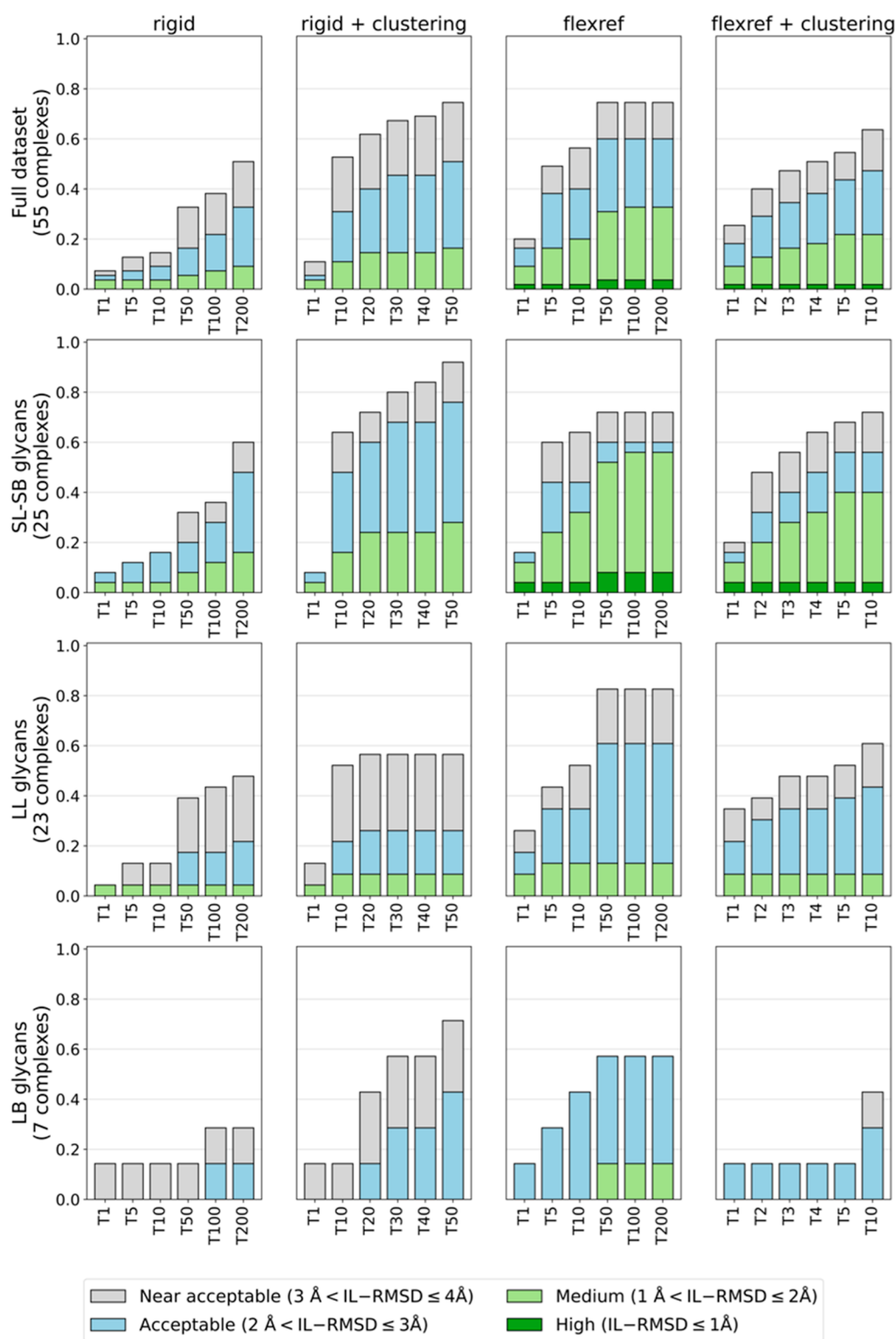


Figure 3. HADDOCK3's performance on the *unbound* data set using the *vdW* scoring function and *tip-ap* AIRs (true interface on the protein defined as active and the glycan residues as passive). The success rates (SRs) (Y axis), defined as the percentage of complexes for which acceptable-, medium-, or high-quality models are generated, are calculated on the top 1 (T1) to top 200 (T200) ranked rigid-body models (column "rigid"), T1 to T50 rigid-body clusters, considering the top 5 models of each cluster (column "rigid + clustering"), the T1 to T200 ranked refined models (column "flexref"), and the T1 to T10 refined clusters, considering the top 4 models of each cluster (column "flexref + clustering"). SRs are shown separately for the whole data set (first row) and for the three categories of complexes grouped by glycan size and connectivity: SL-SB (second row), LL (third row), and LB (fourth row).

The comparison of the rigid-body SR obtained on the single models (column "rigid" in Figure 3) with the SR obtained on the clustered models (column "rigid + clustering" in Figure 3)

reveals that, overall, the selection of clustered models is beneficial to the docking success rate. This is particularly helpful for the long-branched (LB) glycans (fourth row of

Figure 3), with the clustering of the rigid-body models allowing us to retrieve more than 40% acceptable-quality models (70% near-acceptable-quality) compared to 15% acceptable-quality models (30% near-acceptable-quality) for the single model selection. Overall, for all types of glycans, selecting the rigid-body models after clustering is the best way to choose the structures for the refinement stage.

The introduction of flexibility in the interface region strongly improves the quality of the models. Comparing the T200 refined models (column “flexref” in Figure 3) with the selected T50 rigid-body clusters (column “rigid + clustering”), for short (SL-SB) glycans, almost 10% high-quality models are obtained, while there were none before, and almost 60% of the models fall within the medium-quality cutoff, to be compared with around 30% at the rigid-body stage. For the long-linear (LL) glycans, the acceptable-quality success rate increases from around 25% (rigid-body stage) to 61%. The improvement is also substantial for the LB glycans, for which the flexible refinement allows us to obtain medium-quality models, which were not present at the rigid-body stage, and an increase in acceptable-quality SR from 43 to 57%.

Finally, clustering of the refined models (fourth column in Figure 3) slightly improves the success rate compared to T10 single refined models, with around 50% of the glycans having an acceptable or better model in the top 10 clusters.

To demonstrate how flexible refinement affects both glycans' conformations and models' ranking, the best-scoring refined models of three representative complexes are shown in Figure 4, superimposed onto their corresponding rigid-body models and reference structures. These are representative of the SL-SB, LL, and LB groups. The refinement stage results in both better ranking and a better quality (lower IL-RMSD values) of the models. Overall, longer glycans are more challenging to refine than short ones. A high-quality model is produced for the 1C1L complex (SL, IL-RMSD = 0.52 Å after refinement), an acceptable-quality model for 5VX5 (LL, IL-RMSD = 2.22 Å), while for 1OH4 (LB, IL-RMSD = 4.42 Å), the quality is still not acceptable, although it improved after the refinement stage.

We assess the overall impact of the refinement by measuring how much closer to the target structure is the best refined model with respect to the best model prior to the refinement stage; this analysis shows an average improvement of 0.43 Å on our data set with a maximum observed improvement of 3.73 Å. Of course, not all models improve after refinement, but those that do improve also tend to be ranked higher by the HADDOCK scoring function, as shown in Figure 3. Proteins containing long-linear glycans show the most substantial improvement (1.0 Å on average, with a maximum of 3.02 Å). Short- (SL-SB) and long-branched (LB) glycans show a negligible improvement (0.03 and 0.02 Å, respectively), although this is dependent on the specific glycan. For example, considering SL-SB glycans, we observe a maximum improvement of 3.73 Å (PDB 3AOF), while for some other models, the flexible refinement worsens their quality. For LB glycans, a maximum improvement of 2.13 Å is obtained (3AP9). Overall, the flexible refinement affects most LL glycans.

We investigated whether increasing the length of the flexible refinement protocol would further improve the glycan conformations. This did not improve the success rates significantly or in a uniform way as it seemed to depend on the number of models passed to the refinement and the group of glycans considered (data not shown). It further comes at

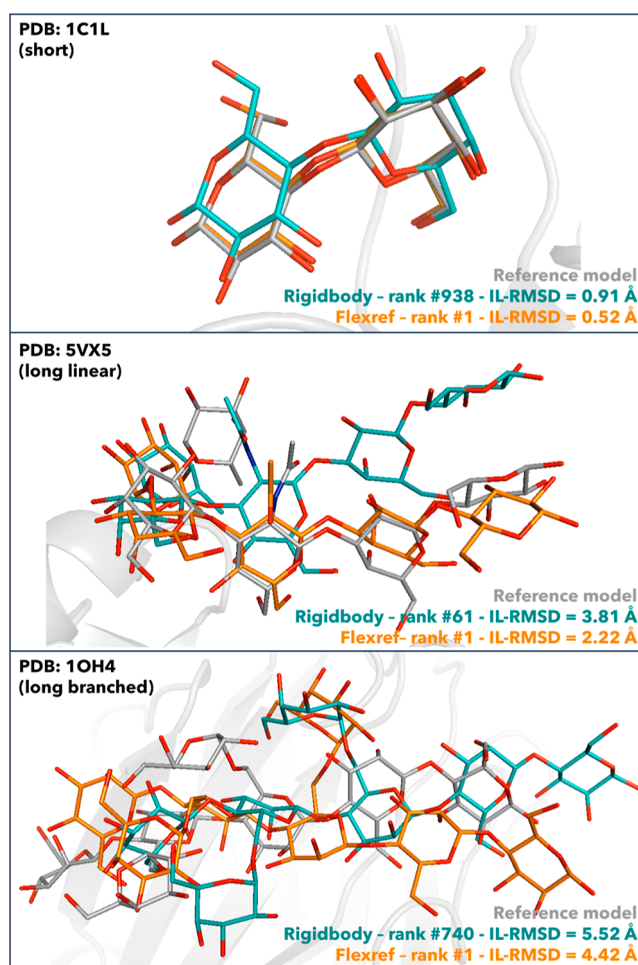


Figure 4. Superimposition of the best-scoring flexible refinement models (orange) and the rigid-body models (teal) to the reference structures (gray) for the complexes 1OH4 (LB), 5VX5 (LL), and 1C1L (SL) and the unbound docking scenario carried out with *vdW* scoring potential and *tip-ap* AIRs. Oxygen atoms of the glycans are shown in red in all the structures, nitrogen atoms in blue, and hydrogens are not shown. Ranking and IL-RMSD values with respect to the reference structures for the flexref and rigid-body models are shown as well.

increased computational costs. As this behavior could not be rationalized in a simple way, this approach was discarded.

Can an Ensemble of Presampled Glycan Conformations Improve the Docking Performance? While the performance in the *unbound docking* scenario is already quite high, unsurprisingly, it does not reach the *bound docking* performance. One limiting factor here could be the unbound glycan conformations used for docking. Comparison of the conformations generated by the GLYCAM-Web web server with respect to the bound ones reveals a strong dependency of the RMSD to the bound form on both the glycan's size and connectivity with mean RMSDs increasing from 0.89 Å for short-linear glycans to 1.74 Å for long-branched ones (see Supporting Information Figure S3).

As HADDOCK can take an ensemble of conformations as the starting point for the docking, we investigated if sampling conformations prior to docking could improve the overall docking performance. We used the water refinement module of HADDOCK, varying the number of models generated and the length of the molecular dynamics sampling. Six different

protocols were investigated (Table S4). The RMSD distributions from the bound glycan conformation of the various protocols are shown as box plots in Supporting Information Figure S4 and compared to the distribution of the original GLYCAM server conformations. From this analysis, the protocol that generates conformations closest to the bound form consists of sampling 400 models, with a 16 times longer refinement protocol (sf400-x16) (which still remains a very short refinement protocol). With this sampling scenario, the average RMSD to the bound conformation decreases from 0.93 to 0.54 Å for SL-SB glycans, from 1.65 to 1.26 Å for LL glycans, and from 1.70 to 1.25 Å for LB glycans. An analysis of the glycosidic linkages indicates that the force field does describe favorable conformations (also during the flexible refinement stage of the docking), but the sampling protocol is too limited to generate conformations very different from those of the starting models generated by GLYCAM (see Supporting Information Figures S8 and S9). Some minor improvements toward the bound form are however observed in some cases.

For docking, we want to limit the number of models in an ensemble to allow for sufficient sampling of each starting model combination without having to increase the sampling too much. To this end, we clustered the sampled glycan conformations using the *rmsdmatrix* and *clustrmsd* modules in HADDOCK3, requesting either 10 or 20 clusters from the ensemble of conformations. RMSDs from the bound conformation were calculated for the centers of each cluster with the *rmsdmatrix* module. Requesting 20 clusters results in the best sampling of the overall glycan RMSD distribution of the ensemble of 400 models, retaining low RMSD conformations. The cluster centers, however, rarely correspond to the best RMSD sampled, which results in some rather limited loss in RMSD to the bound form (see Supporting Information Figure S5). Examples of such conformational ensembles for the same glycans reported in Figure 4 are shown in Supporting Information Figure S6.

The centers of those 20 clusters were provided as an ensemble for unbound docking, following the same rigid-body, cluster-based workflow described above. Only limited improvements are observed in single-structure ranking performance after flexible refinement compared to that of the single unbound conformer protocol (see Supporting Information Figure S7). For example, a higher number of medium-quality models for long-linear glycans is obtained, together with more high-quality models for SL-SB glycans. No improvement is observed for long-branched glycans. This can be attributed to the rather limited conformational sampling during the still short water refinement protocol (Supporting Information Figure S6). Better sampling strategies, possibly based on more extensive MD simulations, will be required. Another issue is related to the selection of the relevant representative conformations while limiting their number for docking purposes.

CONCLUSIONS

In this study, we have assessed HADDOCK's ability in modeling the 3D structures of protein–glycan complexes. First, the baseline performance was evaluated on the rather simple, unrealistic scenario starting from the partners in their *bound* conformations and giving full information about the interface (ti-aa AIRs). This allowed us to improve the rigid-body scoring function for protein–glycan complexes by

increasing the weight of the intermolecular van der Waals energy term to 1.0, as done for the docking of small molecules to proteins.²⁹ An analysis of the *bound docking* performance per type of glycan revealed that longer glycans (in their bound conformation) are easier to model. This is probably a consequence of the lower number of possible orientations that longer glycans can assume when binding the protein. In the more realistic *unbound docking* scenario, in which the glycan conformations were modeled with the GLYCAM-Web web server and free form structures of the proteins are used, conformational changes are required to generate native-like poses. In this case, longer glycans are more difficult to model with only (near-)acceptable predictions generated in most cases.

Making use of the new modular version of HADDOCK3, we introduced a protocol in which rigid-body models are clustered prior to refinement, with representatives of all clusters being passed to the flexible refinement stage of HADDOCK. This strategy enables HADDOCK to retain 51% acceptable or better models after rigid-body selection compared to 33% using the standard single model selection. After flexible refinement, the overall performance reaches an overall success rate of 16% (respectively, 18%) and 38% (respectively, 44%) for T1 and T5 single-structure models (respectively, clusters).

Single-structure or cluster-based selection of models at the end of the workflow shows almost similar performance. This is in line with what was also observed in small-molecule docking with HADDOCK.^{29,38} There is, however, still room to improve the scoring of models. One possible extension could be the incorporation of CH- π stacking interactions into the scoring function, as suggested in previous works.^{39,40}

While the flexible refinement does improve the quality of the models, it is not sufficient in cases in which large conformational changes are required. We therefore investigated whether a limited sampling of glycan conformations prior to docking using the water refinement module of HADDOCK could generate conformations closer to the bound form. While we do observe slight improvements, their impact on the docking performance in an ensemble docking scenario is limited, thus highlighting the need for more extensive conformational sampling strategies, together with methods to identify the most relevant conformations. For example, the use of a database of glycan conformations such as GlycoShape⁴¹ could help in improving HADDOCK's performance in modeling protein–glycan complexes.

ASSOCIATED CONTENT

Data Availability Statement

The software used in this manuscript (HADDOCK3) is publicly available at <https://github.com/haddocking/haddock3>. All input data, analysis scripts, and results presented in the paper can be accessed at <https://github.com/haddocking/protein-glycans>. A tutorial describing the modeling of a protein–glycan complex using HADDOCK3 is hosted at <https://bonvinlab.org/education/HADDOCK3/HADDOCK3-protein-glycan>. The full runs, including docking models from all modules of a workflow, have been deposited in our lab collection (<https://data.sbggrid.org/labs/32/1138>) at the SBGRID data repository.⁴²

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.4c01372>.

Data set used in this study (XLSX)

Details of the preparation of protein and glycan structures; description of HADDOCK3 modules used in this work; details of the glycan conformational sampling protocol; SNFG representation of the glycans; modules and parameters used for bound docking; modules and parameters used for unbound docking; glycan conformational sampling scenarios; example of HADDOCK models satisfying the quality thresholds; HADDOCK3' performance on the bound data set; glycans' RMSD to their bound conformations; impact of mdref on glycan conformations; impact of the clustering on glycans' lowest RMSD; examples of glycan ensembles of conformations; HADDOCK3' performance with the ensembles of glycans; torsion angle analysis of glycosidic linkages: a comparison to HADDOCK flexible refinement models; and torsion angle analysis of glycosidic linkages: a comparison to HADDOCK short molecular dynamics refinement models (PDF)

AUTHOR INFORMATION

Corresponding Author

Alexandre M. J. J. Bonvin – *Bijvoet Centre for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Utrecht 3584CH, The Netherlands*; orcid.org/0000-0001-7369-1322; Email: a.m.j.j.bonvin@uu.nl

Authors

Anna Ranaudo – *Department of Earth and Environmental Sciences, University of Milano-Bicocca, Milan 20126, Italy; Bijvoet Centre for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Utrecht 3584CH, The Netherlands*; orcid.org/0000-0002-8733-7938

Marco Giulini – *Bijvoet Centre for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Utrecht 3584CH, The Netherlands*

Angela Pelissou Ayuso – *Bijvoet Centre for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Utrecht 3584CH, The Netherlands*

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acs.jcim.4c01372>

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

This work was supported by grants from the European High Performance Computing Joint Undertaking project BioExcel (823,830) and The Netherlands e-Science Center (027.020.G13).

Notes

The authors declare no competing financial interest.

ABBREVIATIONS

HADDOCK, high ambiguity-driven DOCKing
IL-RMSD, interface-ligand root mean squared deviation
SL, short-linear
SB, short-branched
LL, long-linear
LB, long-branched
AIRs, ambiguous interaction restraints

PDB, Protein Data Bank.

REFERENCES

- (1) Gold, V. *The IUPAC Compendium of Chemical Terminology*; International Union of Pure and Applied Chemistry (IUPAC): Research Triangle Park, NC, 2019.
- (2) Seeberger, P. H. Monosaccharide Diversity. In *Essentials of Glycobiology*; Varki, A., Cummings, R. D., Esko, J. D., Eds.; Cold Spring Harbor Laboratory Press: New York, 2022; pp 21–32..
- (3) Lebrilla, C. B.; Liu, J.; Widmalm, G.; Prestegard, J. H. Oligosaccharides and Polysaccharides. In *Essentials of Glycobiology*; Varki, A., Cummings, R. D., Esko, J. D., Stanley, P., Hart, G. W., Aebi, M., Mohnen, D., Kinoshita, T., Packer, N. H., Prestegard, J. H., Schnaar, R. L., Seeberger, P. H., Eds.; Cold Spring Harbor Laboratory Press: New York, 2022, pp 33–42..
- (4) Varki, A.; Gagneux, P. Biological Functions of Glycans. In *Essentials of Glycobiology*; Varki, A., Cummings, R. D., Esko, J. D., Eds.; Cold Spring Harbor Laboratory Press: New York, 2022..
- (5) Molina, A.; O'Neill, M. A.; Darvill, A. G.; Etzler, M. E.; Mohnen, D.; Hahn, M. G.; Esko, J. D. Free Glycans as Bioactive Molecules. In *Essentials of Glycobiology*; Varki, A., Cummings, R. D., Esko, J. D., Stanley, P., Hart, G. W., Aebi, M., Mohnen, D., Kinoshita, T., Packer, N. H., Prestegard, J. H., Schnaar, R. L., Seeberger, P. H., Eds.; Cold Spring Harbor Laboratory Press: New York, 2022, pp 539–548..
- (6) Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S.; Fadda, E.; Amaro, R. E. Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Cent. Sci.* **2020**, *6* (10), 1722–1734.
- (7) Woods, R. J. Predicting the Structures of Glycans, Glycoproteins, and Their Complexes. *Chem. Rev.* **2018**, *118* (17), 8005–8024.
- (8) Perez, S.; Makshakova, O. Multifaceted Computational Modeling in Glycoscience. *Chem. Rev.* **2022**, *122* (20), 15914–15970.
- (9) Nance, M. L.; Labonte, J. W.; Adolf-Bryfogle, J.; Gray, J. J. Development and Evaluation of GlycanDock: A Protein–Glycoligand Docking Refinement Algorithm in Rosetta. *J. Phys. Chem. B* **2021**, *125* (25), 6807–6820.
- (10) Mottarella, S. E.; Beglov, D.; Beglova, N.; Nugent, M. A.; Kozakov, D.; Vajda, S. Docking Server for the Identification of Heparin Binding Sites on Proteins. *J. Chem. Inf. Model.* **2014**, *54* (7), 2068–2078.
- (11) Nivedha, A. K.; Thieker, D. F.; Makeneni, S.; Hu, H.; Woods, R. J. Vina-Carb: Improving Glycosidic Angles during Carbohydrate Docking. *J. Chem. Theory Comput.* **2016**, *12* (2), 892–901.
- (12) Boittier, E. D.; Burns, J. M.; Gandhi, N. S.; Ferro, V. GlycoTorch Vina: Docking Designed and Tested for Glycosaminoglycans. *J. Chem. Inf. Model.* **2020**, *60* (12), 6328–6343.
- (13) Labonte, J. W.; Adolf-Bryfogle, J.; Schief, W. R.; Gray, J. J. Residue-centric Modeling and Design of Saccharide and Glycoconjugate Structures. *J. Comput. Chem.* **2017**, *38* (5), 276–287.
- (14) Glashagen, G.; de Vries, S.; Uciechowska-Kaczmarzyk, U.; Samsonov, S. A.; Murail, S.; Tuffery, P.; Zacharias, M. Coarse-grained and Atomic Resolution Biomolecular Docking with the ATTRACT Approach. *Proteins: Struct., Funct., Bioinf.* **2020**, *88* (8), 1018–1028.
- (15) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589.
- (16) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain,

- R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Zidek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* **2024**, *630* (8016), 493–500.
- (17) HADDOCK3, Bonvin's Lab, 2022. <https://github.com/haddock/haddock3>.
- (18) Dominguez, C.; Boelens, R.; Bonvin, A. M. J. HADDOCK: A Protein–Protein Docking Approach Based on Biochemical or Biophysical Information. *J. Am. Chem. Soc.* **2003**, *125* (7), 1731–1737.
- (19) Wu, A. M.; Singh, T.; Liu, J.-H.; Krzeminski, M.; Russwurm, R.; Siebert, H.-C.; Bonvin, A. M. J.; André, S.; Gabius, H.-J. Activity–Structure Correlations in Divergent Lectin Evolution: Fine Specificity of Chicken Galectin CG-14 and Computational Analysis of Flexible Ligand Docking for CG-14 and the Closely Related CG-16. *Glycobiology* **2007**, *17* (2), 165–184.
- (20) Krzeminski, M.; Singh, T.; André, S.; Lensch, M.; Wu, A. M.; Bonvin, A. M. J.; Gabius, H.-J. Human Galectin-3 (Mac-2 Antigen): Defining Molecular Switches of Affinity to Natural Glycoproteins, Structural and Dynamic Aspects of Glycan Binding by Flexible Ligand Docking and Putative Regulatory Sequences in the Proximal Promoter Region. *Biochim. Biophys. Acta, Gen. Subj.* **2011**, *1810* (2), 150–161.
- (21) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.
- (22) Woods Group. *Complex Carbohydrate Research Center; GLYCAM Web*: University of Georgia, Athens, GA, 2023.
- (23) 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* (4), 622–655.
- (24) Varki, A.; Cummings, R. D.; Aebi, M.; Packer, N. H.; Seeberger, P. H.; Esko, J. D.; Stanley, P.; Hart, G.; Darvill, A.; Kinoshita, T.; Prestegard, J. J.; Schnaar, R. L.; Freeze, H. H.; Marth, J. D.; Bertozzi, C. R.; Etzler, M. E.; Frank, M.; Vliegenthart, J. F.; Lütteke, T.; Perez, S.; Bolton, E.; Rudd, P.; Paulson, J.; Kanehisa, M.; Toukach, P.; Aoki-Kinoshita, K. F.; Dell, A.; Narimatsu, H.; York, W.; Taniguchi, N.; Kornfeld, S. Symbol Nomenclature for Graphical Representations of Glycans. *Glycobiology* **2015**, *25* (12), 1323–1324.
- (25) Neelamegham, S.; Aoki-Kinoshita, K.; Bolton, E.; Frank, M.; Lisacek, F.; Lütteke, T.; O'Boyle, N.; Packer, N. H.; Stanley, P.; Toukach, P.; Varki, A.; Woods, R. J.; et al. Updates to the Symbol Nomenclature for Glycans guidelines. *Glycobiology* **2019**, *29* (9), 620–624.
- (26) de Vries, S. J.; van Dijk, A. D. J.; Krzeminski, M.; van Dijk, M.; Thureau, A.; Hsu, V.; Wassenaar, T.; Bonvin, A. M. J. HADDOCK versus HADDOCK: New Features and Performance of HADDOCK2.0 on the CAPRI Targets. *Proteins: Struct., Funct., Bioinf.* **2007**, *69* (4), 726–733.
- (27) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110* (6), 1657–1666.
- (28) Fernández-Recio, J.; Totrov, M.; Abagyan, R. Identification of Protein–Protein Interaction Sites from Docking Energy Landscapes. *J. Mol. Biol.* **2004**, *335* (3), 843–865.
- (29) Basciu, A.; Koukos, P. I.; Mallocci, G.; Bonvin, A. M. J.; Vargiu, A. V. Coupling Enhanced Sampling of the Apo-Receptor with Template-Based Ligand Conformers Selection: Performance in Pose Prediction in the D3R Grand Challenge 4. *J. Comput. Aided Mol. Des.* **2020**, *34* (2), 149–162.
- (30) Nilges, M. A calculation strategy for the structure determination of symmetric dimers by ^1H NMR. *Proteins: Struct., Funct., Bioinf.* **1993**, *17* (3), 297–309.
- (31) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. LIGPLOT: A Program to Generate Schematic Diagrams of Protein–Ligand Interactions. *Protein Eng., Des. Sel.* **1995**, *8* (2), 127–134.
- (32) McDonald, I. K.; Thornton, J. M. Satisfying Hydrogen Bonding Potential in Proteins. *J. Mol. Biol.* **1994**, *238* (5), 777–793.
- (33) Méndez, R.; Leplae, R.; De Maria, L.; Wodak, S. J. Assessment of Blind Predictions of Protein–Protein Interactions: Current Status of Docking Methods. *Proteins: Struct., Funct., Bioinf.* **2003**, *52* (1), 51–67.
- (34) Giulini, M.; Menichetti, R.; Shell, M. S.; Potestio, R. An Information-Theory-Based Approach for Optimal Model Reduction of Biomolecules. *J. Chem. Theory Comput.* **2020**, *16* (11), 6795–6813.
- (35) Sokal, R. R.; Michener, C. D. *A Statistical Method for Evaluating Systematic Relationships*; University of Kansas Science Bulletin, 1958; Vol. 38, pp 1409–1438.
- (36) Charitou, V.; van Keulen, S. C.; Bonvin, A. M. J. Cyclization and Docking Protocol for Cyclic Peptide–Protein Modeling Using HADDOCK2.4. *J. Chem. Theory Comput.* **2022**, *18* (6), 4027–4040.
- (37) Buchanan, C. J.; Gaunt, B.; Harrison, P. J.; Yang, Y.; Liu, J.; Khan, A.; Giltrap, A. M.; Le Bas, A.; Ward, P. N.; Gupta, K.; Dumoux, M.; Tan, T. K.; Schimaski, L.; Daga, S.; Picchiotti, N.; Baldassarri, M.; Benetti, E.; Fallerini, C.; Fava, F.; Gilberti, A.; Koukos, P. I.; Davy, M. J.; Lakshminarayanan, A.; Xue, X.; Papadakis, G.; Deimel, L. P.; Casablanco-Antràs, V.; Claridge, T. D. W.; Bonvin, A. M. J.; Sattentau, Q. J.; Furini, S.; Gori, M.; Huo, J.; Owens, R. J.; Schaffitzel, C.; Berger, I.; Renieri, A.; Naismith, J. H.; Baldwin, A. J.; Davis, B. G.; Davis, B. G. Pathogen–Sugar Interactions Revealed by Universal Saturation Transfer Analysis. *Science* **2022**, *377* (6604), No. eabm3125.
- (38) Koukos, P. I.; Réau, M.; Bonvin, A. M. J. Shape-Restrained Modeling of Protein–Small-Molecule Complexes with High Ambiguity Driven DOCKing. *J. Chem. Inf. Model.* **2021**, *61* (9), 4807–4818.
- (39) Kerzmann, A.; Neumann, D.; Kohlbacher, O. SLICK – Scoring and Energy Functions for Protein–Carbohydrate Interactions. *J. Chem. Inf. Model.* **2006**, *46* (4), 1635–1642.
- (40) Kerzmann, A.; Fuhrmann, J.; Kohlbacher, O.; Neumann, D. BALLDock/SLICK: A New Method for Protein–Carbohydrate Docking. *J. Chem. Inf. Model.* **2008**, *48* (8), 1616–1625.
- (41) Ives, C.; Singh, O.; D'Andrea, S.; Fogarty, C.; Harbison, A.; Satheesan, A.; Tropea, B.; Fadda, E. Restoring Protein Glycosylation with GlycoShape. *bioRxiv* **2023**.
- (42) Meyer, P. A.; Socias, S.; Key, J.; Ransey, E.; Tjon, E. C.; Buschiazzo, A.; Lei, M.; Botka, C.; Withrow, J.; Neau, D.; Rajashankar, K.; Anderson, K. S.; Baxter, R. H.; Blacklow, S. C.; Boggon, T. J.; Bonvin, A. M. J.; Borek, D.; Brett, T. J.; Cafilisch, A.; Chang, C.-I.; Chazin, W. J.; Corbett, K. D.; Cosgrove, M. S.; Crosson, S.; Dhe-Paganon, S.; Di Cera, E.; Drennan, C. L.; Eck, M. J.; Eichman, B. F.; Fan, Q. R.; Ferré-D'Amaré, A. R.; Christopher Fromme, J.; Garcia, K. C.; Gaudet, R.; Gong, P.; Harrison, S. C.; Heldwein, E. E.; Jia, Z.; Keenan, R. J.; Kruse, A. C.; Kvensakul, M.; McLellan, J. S.; Modis, Y.; Nam, Y.; Otwinowski, Z.; Pai, E. F.; Pereira, P. J. B.; Petosa, C.; Raman, C. S.; Rapoport, T. A.; Roll-Mecak, A.; Rosen, M. K.; Rudenko, G.; Schlessinger, J.; Schwartz, T. U.; Shamoo, Y.; Sondermann, H.; Tao, Y. J.; Tolia, N. H.; Tsodikov, O. V.; Westover, K. D.; Wu, H.; Foster, I.; Fraser, J. S.; Maia, F. R. N. C.; Gonen, T.; Kirchhausen, T.; Diederichs, K.; Crosas, M.; Sliz, P. Data Publication with the Structural Biology Data Grid Supports Live Analysis. *Nat. Commun.* **2016**, *7* (1), 10882.