CLIMBS: assessing Carbohydrate-Protein interactions through a graph neural network classifier using synthetic negative data

Authors: Yijie Luo[1*], Fabio Parmeggiani,[1,2,3*]

Affiliations:

[1]School of Biochemistry, University of Bristol; University Walk, Bristol BS8 1TD, UK.

[2]School of Chemistry, University of Bristol; Cantock's Close, Bristol BS8 1TS, UK.

[3]School of Pharmacy and Pharmaceutical Sciences, Cardiff University, Redwood Building, King Edward VII Ave, Cardiff, CF10 3NB, UK

*Corresponding authors. Email: yijie.luo@bristol.ac.uk , parmeggianif@cardiff.ac.uk

I. Abstract

Carbohydrate-protein interactions are essential for biological processes such as cellular signaling and metabolism. However, current prediction methods cannot accurately evaluate the affinity and specificity of proteins for carbohydrates such as glucose and galactose. Here, we develop a machine learning classifier, named CLIMBS, and train it on crystal structures and synthetic data from unsuccessfully designed binders, to effectively assess if carbohydrate-protein complexes represent realistic, native like structures. Compared to other methods, CLIMBS has outstanding accuracy, excellent carbohydrate specificity, sub-second runtime per sample, and no prediction bias towards either negative or positive samples. CLIMBS can work as an additional criterion in carbohydrate-protein interaction modeling for areas such as docking, protein design and enzyme design.

II. Introduction

Carbohydrate recognition by proteins is a fundamental step in a number of biological processes, from glycolysis[1] to assembly of specific glycosylation patterns and formation of polysaccharide materials[2], either for energy storage, as for starch[3], or for structural use, like cellulose[4]. On the other hand, large groups of lectins act as receptors for glycans and poly-saccharides and play key roles in cellular signalling and adhesion [5]. Identification of receptors and binding partners, and specificity predictions have become fundamental in understanding interactions between the glycome and the proteome. Although high throughput [6] and glycan specificity computational models have been developed[7,8] , most of the molecular details for specific interactions remain routinely out of experimental grasp, and addressed through docking and modelling of small molecule-protein interactions[9,10].

Accurately assessing protein-carbohydrate interactions remains challenging for energy-based modelling methods. Even with established scoring functions, such as AMBER [11], Rosetta [12] and AutoDock4 [13] are commonly used in traditional docking methods, weak but key interaction such as CH-π[15] can be difficult to capture when considering the whole complex or interface. Alternatively, machine learning methods have been applied to docking and evaluation of protein-small molecule complexes. GNINA [16] and Sfcnn [17] use convolutional neural networks (CNN) and present complex structures in 3D-grid voxels. CSM-lig [18] and DeepDOCK [19] use graph neural networks (GNN), presenting complex structures as graphs with additional physicochemical characteristics as features. Generative models such as DIFFDOCK [20] and AlphaFold3 [21] have intrinsic confidence models to rank and assess the quality of the generated structures. Some methods have

introduced specific biases to explicitly address protein-carbohydrate interactions, for example the GLYCAM06 energy function [22] and CSM-carbohydrate model [23] .

The development of machine learning methods targeted to protein-carbohydrate interactions has the potential to improve assessment of native-like interface in docking[24] and design [25]. However, for protein carbohydrate complexes, there is no available large dataset which unsupervised learning methods rely on. When looking at available structural information, among 211,432 protein structures in the protein data bank (PDB) [26], only 12.33% (26,705 structures) of them contained carbohydrates as of February 2024. And if we consider redundancy, structures with resolution lower than 3 Å and only structures where the carbohydrate form part of a binding interface, the number drops to 3.84% (8,119 structures). In contrast, supervised learning methods, such as classifiers, require much smaller datasets to achieve reasonable performances[27] , although appropriately labelled. While selected protein carbohydrate-complexes in the PDB can provide positive samples ("binding"), negative samples (or "not-binding") for complexes, where the carbohydrate is positioned unfavorably in a binding pocket, can only be computationally generated. Interestingly, reasonable non binders can be easily generated, since even state of the art protein design methods have achieved success rates of only 0.67% [28] for small molecule binders. This implies that a large synthetic dataset of designed binders for target carbohydrates could provide realistic and challenging negative samples for training supervised learning methods. Building on this hypothesis, we have developed a graph neural network (GNN)-based model: CLIMBS, **cl**ass**i**fier of **m**inimum **b**inding site for **s**ugars (figure 1). We trained it using positive experimental data and synthetic negative data and then compared its performance in discriminating native and

computationally generated complexes with those of established scoring methods. We then explored the potential for retraining on new types of sugars and discriminate complexes with untrained sugars.

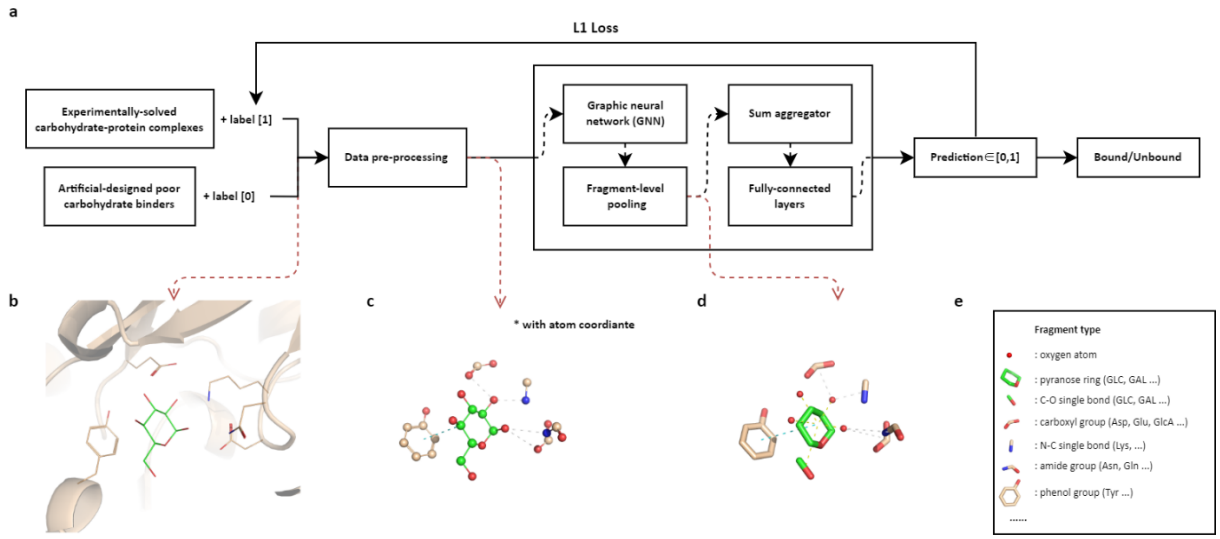


Figure 1 Flowchart of CLIMBS. (a) Overview of CLIMBS. (b) PDB file as input structure data of the model. (c) Structure data becomes an atom-level graph with atom coordinates for each node after data pre-processing. (d) In CLIMBS, structure data becomes a fragment-level graph after a pooling layer. (e) Legend of fragment types shown.

III. Results

Overview of CLIMBS

CLIMBS was developed as a graph neural network classifier to validate atomistic models of carbohydrate-protein complex. CLIMBS takes carbohydrate-protein complex structures as inputs, and it outputs 1 or 0 indicating predictions as “binding” or “non-binding”. The binding interface, rather than the whole complex, is used as the model input to focus on binding site geometry and reduce run time. Atom-level geometric information is used in the form of graphs to capture details of interactions. Due to the structural similarity between carbohydrates, fragment-level information (including groups

of neighbouring atoms, such as hydroxyl groups or carbon rings, see Figure 1) is processed within a pooling layer, facilitating learning across various types of sugars with shared features [29,30]. A structure library was built to include 21 different carbohydrates from experimentally determined carbohydrate-protein complexes that were used as positive samples, while suboptimal computationally-designed carbohydrate-protein complexes were introduced as negative samples (see details in Methods). CLIMBS predicts binding between carbohydrate and proteins with an overall accuracy of 94.78%. The model was trained on 2543 positive and 2389 negative samples with 429 epochs. The binding predictions of different types of sugar-protein complexes were tested on a separate set of complexes containing either carbohydrates used during training or unseen sugars (see table 1); a minimal accuracy of 62.50% was observed for α-D-ribose (not used in training), and a maximal accuracy of 98.77% for N-acetyl-α/β-D-galactosamine. Our carbohydrate-protein complexes library (see supplementary table 1) mirrors the Protein Data Bank, with seven mono-saccharide types (Glc, GlcNAc, galactose, fucose, mannose, sialic acid and fructose) accounting for 86% of all complexes. Carbohydrate less represented in the library such as xylose and GlcA still reach an accuracy similar to the most abundant ones. A robustness check (see methods) revealed that both CH-π and polar interactions contribute to the classification. Moreover, the trained CLIMBS model can be retrained to include new sugar types whilst keeping excellent performance when adding only tens of new samples in the training set.

Table 1 Accuracy of CLIMBS tested by different sugar-protein complexes. 15 types of sugar-protein complex are trained and 6 types are new to the classifier. For trained sugar type, sample number is the size of the test set after splitting (1/5 of the whole dataset). For untrained sugar types, the sample number is the size of the whole dataset. All complexes include Repeat region of unmodified heparan sulphate were derived from Protein Data Bank.

Sugar type in complex	**Accuracy (429 epochs)**	**Tested Sample number**
α/β-D-glucose	96.10%	949
N-acetyl-β-D-glucosamine	98.35%	486
α-D-mannose	95.65%	207
β-D-galactose	94.33%	194
maltose, GLC-(1,4)-βGlc	90.70%	129
α-L-fucose	97.60%	125
N-acetyl-α-neuraminic acid	97.83%	92
N-acetyl-α/β-D-galactosamine	98.77%	81
β-D-fructose	97.14%	70
α-D-ribose (untrained)	62.50%	40
α-L-arabinose (untrained)	78.79%	33
α-D-xylose	90.63%	32
6-O-phosphono-α/β-D-glucose	96.88%	32
GlcNAc-(1,4)-βGlcA & GlcA-(1,3)-βGlcNAc (repeat unit of heparan sulphate, untrained)	79.17%	24
N-acetyl-4-O-sulfo-β-D-galactosamine (untrained)	65.00%	20
β-D-glucuronic acid	89.47%	19
sucrose, GLC-(1,2)- βFru	94.12%	17
N,O6-disulfo-glucosamine	75.00%	8
Repeat region of unmodified heparan sulphate (constructed by 3 GlcA and 3 GlcNAc, untrained)	87.50%	8
	94.78%	**2566**

Performance comparison with other methods

CLIMBS has shown an outstanding performance in predicting sugar binding when compared to other scoring methods, according to runtime and four model performance metrics (accuracy, precision, specificity and sensitivity) (figure 2). A group of N-acetyl-glucosamine, β-glucose, α-glucose, β-D-galactose, β-D-fructose, maltose, sucrose and GlcNAc-GlcA complexes not used for training were selected as test set for performance

comparison (see Supplementary Data 3j). We chose score functions from energy-based methods (Rosetta Energy Function [12], Autodock4 score function [13], HADDOCK3[31]) and a confidence model from a machine learning approach (DIFFDOCK [20]). Each method was calibrated to get the highest accuracy on a carbohydrate binding task (see Supplementary Data 5).

CLIMBS spends only 0.06s to predict one sample in one CPU. It is mainly because the classifier only considers the residues around the protein-sugar interface. CLIMBS has the best accuracy (96.45%) and precision (97.01%), more than 15% higher than other methods. It also has remarkable and highly balanced specificity and sensitivity (96.45% on average and 1.05% in difference). All other four methods have considerable bias, with at least 15% difference between specificity and sensitivity. The DIFFDOCK confidence model, Rosetta Energy Function and the Autodock4 score function biase predictions towards non-bound, while the HADDOCK3 score function prefers to predict complexes as bound.

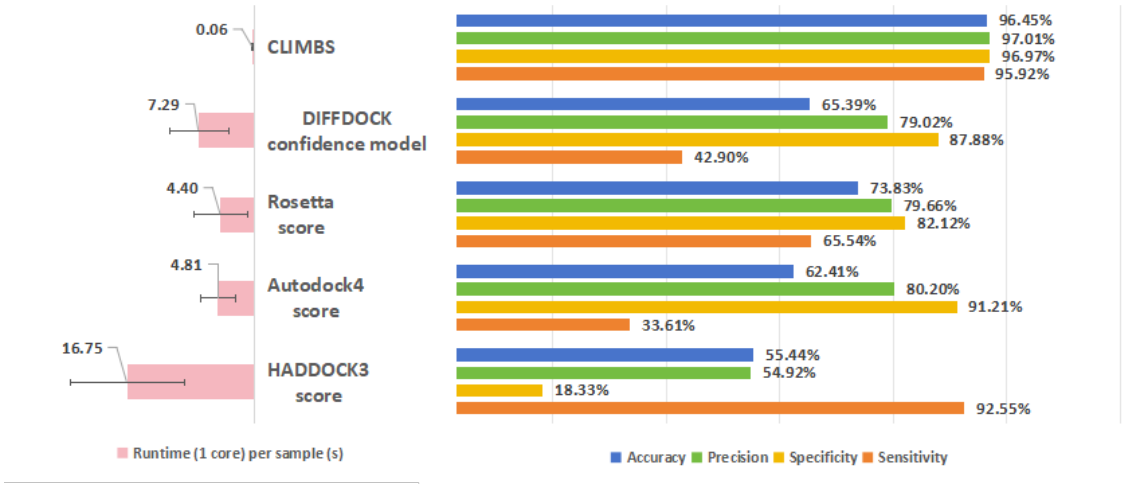


Figure 2 Comparison between CLIMBS and other methods by different indicator: runtime per sample, accuracy: (TP+TN)/(TP+FP+TN+FN), precision: TP/(TP+FP), specificity: TN/(TN+FP) and sensitivity: TP/(TP+FN). TP True Positive, TN True Negative, FP False Positive, FN False negative.

To further investigate the prediction bias of the models, distributions of normalized prediction of positive/negative samples for each method are shown in figure 3. The Wasserstein distance is used to describe the difference between distributions of positive and negative samples. CLIMBS' predictions are separated, with 0.927 Wasserstein distance. For the other four methods, the positive and negative distributions are overlapped in most cases, with 0.123 being the highest Wasserstein distance. Thus, CLIMBS can distinguish well between bound and non-bound samples with various sugar binder datasets.

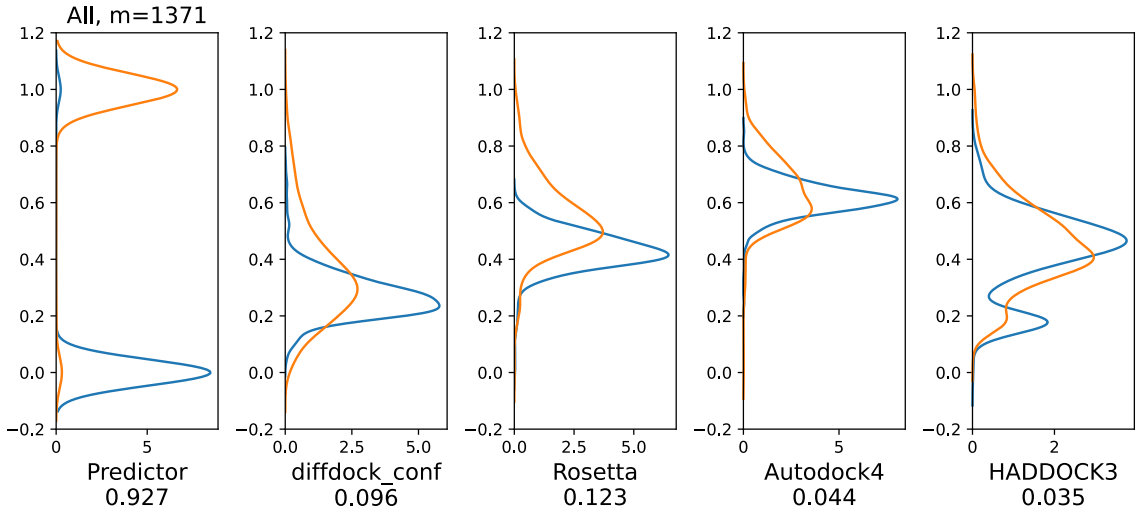


Figure 3 Comparison between CLIMBS and other methods by Wasserstein distance across multiple carbohydrates.. Scores from each method were normalised between 0 and 1, corresponding to the lowest and highest scoring samples. Distributions of normalised scores are plotted as kernel density estimates. Positive samples are shown in orange and negative samples in blue. The number under each subplot is the Wasserstein distance between the normalised score distribution of positive/negative samples for each method. m: tested sample amount. For the dataset see Supplementary Data 3j. For details of each sugar type see Supplementary Data 6.

Sugar-type specificity

Due to the structural similarity of mono-saccharides, interaction specificity is an important performance indicator in modelling sugar-protein interactions. Therefore, we have evaluated CLIMBS specificity performance on the carbohydrate-binding domain of

Marinomonas primoryensis PA14 (MpPA14), which binds to a series of mono-saccharides except N-acetyl-galactosamine (GalNAc). The apparent dissociation constants ($K_d$app) of MpPA14-sugar complexes were measured by isothermal titration calorimetry (ITC), and structures of complexes were solved in previous work. [32]

Six solved complexes [32] and one putative structure with GalNAc (generated by aligning and replacing GlcNAc with GalNAc in a solved MpPA14-GlcNAc complex) were selected as initial structures after Rosetta relaxation. Each initial structure underwent a Rosetta local docking protocol [33] that generated 200 docked structures. CLIMBS was used to evaluate the binding of the initial and the docked structures. The results in table 2 show that the classifier, similarly to Rosetta scoring function, can correctly identify that GalNAc is not binding while the other sugars are bound.

Table 2 Binding between MpPA14 and different sugars. Previous work [32] measured the apparent dissociation constants ($K_d$app) by isothermal titration calorimetry, and the binding free energy (ΔΔG) was calculated based on $K_d$app with a temperature of 298 K. Rosetta and CLIMBS were used to estimate the binding of relaxed solved structures and the structure after running a Rosetta local docking. The local docking test values are the average and standard deviation of the results of 200 docking runs. For CLIMBS the values represent the fraction of times that docking poses have been scored as positive

	Experiment verification		Crystallized structure		Rosetta local docking test	
Sugar	**Kdapp(mM)**	**ΔΔG_298K(kcal/mol)**	**CLIMBS**	**Rosetta ΔΔG (kcal/mol)**	**CLIMBS**	**Rosetta ΔΔG(kcal/mol)**
L-Fucose	0.65	-8.41	1	-7.08	0.90	-9.60±0.76
N-acetyl-glucosamine	1.07	-8.11	1	-4.53	1.00	-3.34±1.27
Glucose	1.4	-7.95	1	-4.95	0.90	-3.07±0.91
Mannose	1.7	-7.84	1	-3.75	1.00	-2.37±0.77
L-Arabinose	2.2	-7.69	1	-4.56	0.98	-0.03±0.57
Galactose	4.1	-7.32	1	-2.88	0.98	-1.89±0.89
N-acetyl-galactosamine	--	--	--	--	0.30	4.81±1.63

Medium-trained model to predict binding of unseen sugars

Several types of sugars have no solved structure in complex with binding protein and CLIMBS is likely to face some samples that including sugar type not used for training. To test the performance on unseen carbohydrates, a model (see Supplementary Data 3d) trained on samples including various sugars except for GlcNAc-GlcA (NAG-BDP),

arabinose (ARA), 5-O-phosphono-ribose (RP5) and 2-N-sulfo-6-O-sulfo-D-glucosamine (ASG) was used to test four unseen sugars with different training epochs (figure 4). As the training epoch increases, the validation loss of trained sugars samples converges to less than 0.1 within 400 epochs. The loss of arabinose samples performs like the trained samples, decreasing along the training. However, the losses of the other untrained sugars samples stop converging between 100 to 200 epochs, and the loss of untrained di-saccharide even begins increasing from 300 epochs. The diverging losses of training set and untrained set depicts that overfitting of the model happens as the training epoch increases, indicating that a medium-trained model (at 280 epochs) is therefore an appropriate choice for predictions on untrained carbohydrates.

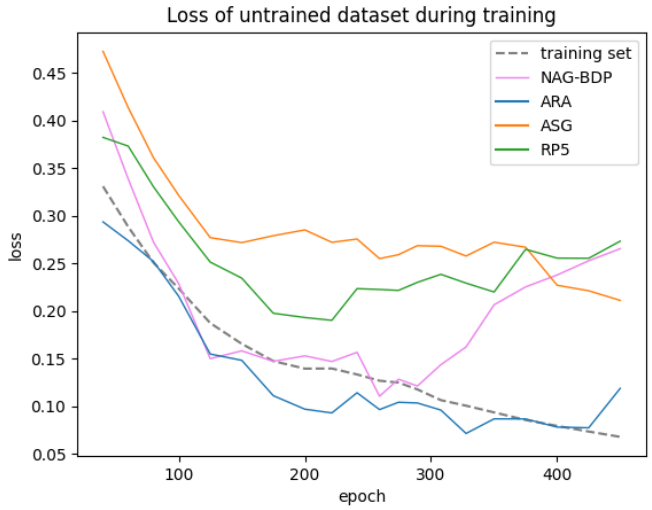


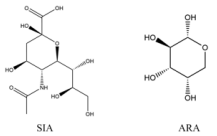
Figure 4 Loss curves of trained and untrained samples during CLIMBS training (see Supplementary Data 3d). The training set included 14 types of sugar binders (see table 1). Four types of untrained samples are GlcNAc-GlcA (NAG-BDP), arabinose (ARA), 5-O-phosphono-ribose (RP5) and 2-N-sulfo-6-O-sulfo-D-glucosamine (ASG) binder.

Retraining the model for new sugars

CLIMBS can be retrained for new sugar types with limited number of samples, which we assessed with a series of retraining tests.

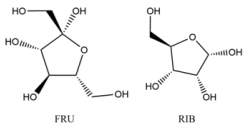
An initial database with samples containing proteins in complex with highly represented carbohydrates in the PDB (Glc, GlcA, GlcNAc, galactose, fucose, mannose and xylose, 5095 samples, see Supplementary Data 3e) was used to train a starting model. Those seven sugars contain fragments common among different sugar (e.g. central ring and hydroxyl groups). Our hypothesis was that, due to shared overall structures, accuracy for new carbohydrate could be rapidly increased with only few samples introducing specific features. Pairs of additional carbohydrates with similar features and absent from the initial dataset, were selected: one was added to the initial database for training and validation, and the second used for monitoring accuracy improvement without training.

a)



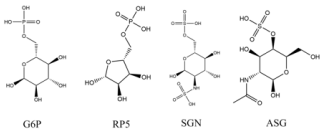
	SIA (37)		ARA (23)	
	accuracy	loss	accuracy	loss
initial database	67.57%	0.318	78.26%	0.21
+12 SIA sample	81.08%	0.226	95.65%	0.088
+24 SIA sample	78.38%	0.248	78.26%	0.232
+48 SIA sample	**91.89%**	0.092	73.91%	0.251

b)



	FRU (17)		RIB (17)	
	accuracy	loss	accuracy	loss
initial database	56.52%	0.426	88.24%	0.108
+12 FRU sample	60.87%	0.414	94.12%	0.195
+24 FRU sample	56.52%	0.429	88.24%	0.136
+48 FRU sample	**95.65%**	0.048	82.35%	0.198

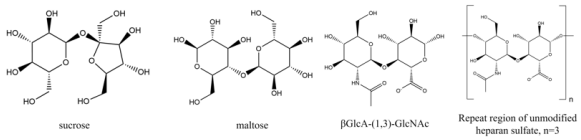
c)



	G6P (14)		RP5 (21)	
	accuracy	loss	accuracy	loss
initial database	71.43%	0.27	76.19%	0.24
+12 G6P sample	**78.57%**	0.23	85.71%	0.15

	SGN (33)		ASG (39)	
	accuracy	loss	accuracy	loss
initial database	60.61%	0.39	74.36%	0.26
+37 SGN sample	**81.82%**	0.19	76.92%	0.23

d)



	Sucrose (58)		Maltose (53)		GlcNAc-GlcA (78)		unmodified HS (8)	
	accuracy	loss	accuracy	loss	accuracy	loss	accuracy	loss
initial database	68.97%	0.31	66.04%	0.34	91.03%	0.09	75.00%	0.22
+12 di sample	72.41%	0.27	73.58%	0.26	89.74%	0.16	87.50%	0.11
+24 di sample	82.76%	0.21	73.58%	0.28	87.18%	0.16	87.50%	0.11
+48 di sample	82.76%	0.2	**77.36%**	0.24	80.77%	0.21	87.50%	0.1
+96 di sample	**87.93%**	0.12	66.04%	0.33	91.03%	0.16	87.50%	0.11

Figure 5 Retraining CLIMBS with new sugars. a. Accuracy and training loss of arabinose (ARA, 37 samples in test set) and sialic acid samples (SIA, 23 samples in test set) from a CLIMBS model trained on initial database plus different number of sialic acid samples (see Supplementary Data 3f). b. Accuracy and training loss of ribose (RIB, 17 samples in test set) and fructose (FRU, 17 samples in test set) samples from a CLIMBS model trained on initial database plus different number of fructose samples (see Supplementary Data 3g). c. (Above) Accuracy and training loss of 6-O-phosphono-glucose (G6P, 14 samples in test set), 5-O-phosphono-ribose (RP5, 21 samples in test set) from a CLIMBS model trained on initial database plus different number of 6-O-phosphono-glucose samples. (Bottom) Accuracy and training loss of N,O6-disulfo-glucosamine (SGN, 33 samples in test set), N-acetyl-4-O-sulfo-galactosamine (ASG, 39 samples in test set) and trained. d. Accuracy and training loss of maltose, sucrose ,GlcNAc-GlcA and repeat regions of unmodified heparan sulfate samples from a CLIMBS model trained on initial database plus different number of maltose and sucrose samples (See Supplementary Data 3i). The ratio of maltose and sucrose is 1:1. HS: heparan sulfate.

Two pyranoses, sialic acid and arabinose, were selected as first candidates with starting accuracy at 67.57% and 78.26%, respectively (table 3a). The lower value for sialic acid might be due to a larger structural complexity (nine-carbon backbone) and the absence of an atom type (CH0 atom) in the database, while all atom types for arabinose are already present in the database. Sialic acid samples were progressively added to the training set, while arabinose was kept untrained (see Supplementary Data 3f). For sialic acid, adding 12 samples is sufficient to reach an approximate 80% accuracy, without decreasing the control accuracy (arabinose). The control accuracy dropped when more than 12 samples were added, indicating overfitting.

As second case, we introduced fructose and ribose (table 3b), which both contain a furanose ring not seen in the initial database (all trained carbohydrate types have a pyranose ring). The initial binding accuracy was high for ribose (88.24%) but not for fructose samples (56.52 %). The lower value for fructose might be due the absence of an atom type (CH0 atom) in the database. Then, fructose samples were added to the training set (see Supplementary Data 3g). The accuracy of fructose samples increases significantly at 48 samples to 95.75%, however showing a small decrease in ribose accuracy.

We also investigated how many samples were needed to retrained CLIMBS for new mono-saccharides that include chemical modifications. Phosphorylation and sulfation of carbohydrate are two frequently observed functional modifications that were not present in the initial database. N,O6-disulfo-glucosamine (SGN) and 6-O-phosphono-glucose (G6P) were selected for model training and testing, while N-acetyl-4-O-sulfo-galactosamine (ASG) and 5-O-phosphono-ribose (RP5) only for testing (see table 3c and

Supplementary Data 3h). Surprisingly, the initial model that wasn't trained on phosphate and sulphate has accuracies between 60% and 76% to predict binding of modified sugars, likely driven by the rest of the carbohydrate structural features. Retraining with less than 40 samples with phosphorylated/sulfated sugar improved binding accuracy to 76%-85%, both for trained and untrained carbohydrates.

At last, we investigated retraining for di-saccharides constructed by seen mono-saccharides. Maltose (glucose-α-1,4-glucose) and sucrose (glucose-α-1,2-fructose) were selected as candidate di-saccharide for training and testing (See table 3d and Supplementary Data 3i). Additionally, a heparan sulphate fragment, GlcNAc-GlcA (β-1,4 and β-1,3 linkage) disaccharide unit was added for testing. The initial model, trained on only mono-saccharides samples, has considerable performance on the di-saccharide structures, despite no glycosidic bond being included in the training set. The accuracy of sucrose and maltose increases from near 70% to 88.93% with 96 extra samples for sucrose and to 77.36% with 48 extra samples for maltose. The accuracy of GlcNAc-GlcA and repeat regions of unmodified heparan sulfate are hardly affected and fluctuates at high value around 90%. It shows that the model trained by mono-saccharides is able to assess di-saccharide and even poly-saccharide accurately. More generally, introducing 48 di-saccharide samples into training can increase the accuracy before overfitting.

IV. Discussion

In this study we have developed CLIMBS, a machine learning classifier to evaluate carbohydrate-protein binding in atomistic models.

CLIMBS shows high accuracy and specificity toward mono-saccharides, di-saccharides

and poly-saccharides. With the atomic level representation, it can work as a reliable benchmark for sugar-binding protein design and sugar-protein docking. Also, the extremely fast runtime makes it feasible to combine with current energy-based methods or machine learning methods, filtering the false positive/negative results and increasing the overall pipeline accuracy.

A sugar-protein complex library was built to include 21 different carbohydrates for model training and testing. Solved sugar-protein complexes and designed carbohydrate binders were collected and labelled as positive and negative samples, while other methods tend to use solved binder structures with scalar labels such as experimental binding affinity. A binary label was chosen because there is not a scalar experimental result for non-binding. On one hand, it means that CLIMBS cannot describe the strength or extent of sugar-protein interactions, although, as in the case of validation of multiple docking poses, a crude but effective scalar value can be established by averaging the entire population. On the other hand, CLIMBS focuses on effectively discriminating good-binding and poor/non-binding, using a method independent from and unrelated to currently used score functions or confidence metrics.

The choice of suboptimal carbohydrate designs as synthetic negative data, was a key element in the setup of the training dataset and in the performance of CLIMBS. It allowed us to build a balanced dataset and to focus training on specific atomic details that might be overlooked in current approaches.

CH-π interactions, which are often fundamental for sugar binding but missed for most energetic-based methods, are considered as one of the interaction types in this model. A robustness check depicts that the classifier considers CH-π interacting residue and other

factors like residue with polar interaction and intra-molecular chemical bond. A fragment-level representation keeps the necessary information with less redundancy, facilitating learning of carbohydrate features. For samples with new sugar types, classifiers with lower trained epoch can be used to avoid potential overfitting, or few tens of new samples can be added to the dataset to retrain. These characteristics make CLIMBS architecture particularly suitable for carbohydrate-protein investigation, since databases are relatively small and different carbohydrates have similar structures, enhancing the recognition of these common features.

Metal ions can be part of carbohydrate binding sites, as for the MpPA14 used for assessing carbohydrate specificity, which contains a calcium ion. 141 carbohydrate-protein complexes in the library contain at least one metal atoms in binding site, but are considered only implicitly in CLIMBS; although the metal ions were removed during the data preprocessing, the residues interacting to the metal ions were kept and provided this information during training. As the result, CLIMBS can work with carbohydrate-protein complexes that include metal ions even in absence of ions in the structure models.

Compared to other score functions and machine learning methods, CLIMBS has shown outstanding accuracy and precision, fast runtime, and no predisposition towards favoring positive or negative predictions. Furthermore, as an entirely geometry-based method, it can distinguish different sugars with highly similar structure similarity that energy-based methods cannot. It is therefore a new scoring tool, orthogonal to other energy and machine learning based methods, that, when combined with docking and design protocols and pipeline, allows to improve selection of protein-carbohydrate models that retain native-like features.

V. Methods

Model architecture

CLIMBS takes a PDB structure file containing a carbohydrate chain and one or more protein chains as input, and it outputs a value of 1 (binding) or 0 (not binding). A protein-sugar interacting graph is generated from data pre-processing. Then, the interacting graph is passed to the embedding module including spherical message passing layers [34], principal neighborhood aggregation layers [35], and a super-node-based hierarchical graph pooling layer [36]. Followed by fully connected layers and a sigmoid function, a binary result is returned.

Data pre-processing elicits the essential details of protein-sugar interactions in the pdb structure file (see Supplementary Data 7). Firstly, it selects the amino acid and carbohydrate residues on the interface. Only the amino acid/saccharides that have polar or CH-π interaction to the saccharide/amino acids are kept (see Supplementary Data 1). Hydrogen atoms are discarded. Secondly, it elicits the minimum binding site from the interface residues: for amino acids, either the terminal functional groups of sidechains or the mainchain backbones are kept; for saccharides, all heavy atoms are kept. Finally, the information of the minimum binding site is generated, including atom coordinates, an atom-level interacting graph $G=\{V,E\}$ and an assignment matrix S. Node matrix V includes the one-hot encoded atom types. Edge matrix E describes three types of interaction between nodes: chemical bond, polar interaction and CH-π interaction. Assignment matrix S assigns atoms into corresponding fragments. Fragments are groups of atoms that commonly exist in different molecules. For example, glucose is constructed by a 6-member ring fragment, 4 oxygen fragments and a carbon-oxygen (single bond)

fragment. The fragment-level pooling layer of CLIMBS can capture the features of fragments shared among different types of sugars.

Given the pre-processed data, the classifier model returns a binary value indicating bound or not (see figure 5). In the beginning, the interaction graph $G=\{V,E\}$ as well as the atom coordinates are passed to spherical message passing (SMP) layers [34] to generate an updated graph $G'=\{h,E\}$. The first SMP layer converts the atom coordinates into three geometrical representations (distances, angles, and torsions) and incorporates them into the interaction graph G. Then, the rest of SMP layers update edges and aggregates them into node feature matrix h. Then, principal neighbourhood aggregation (PNA) [35] layers update $G'=\{h,E\}$ into $G''=\{h',E\}$ by a multiple aggregation function and degree-scalers. After that, a pooling layer coarsens the atom-level graph $G''=\{h',E\}$ into a fragment-level graph $G^*=\{h^*,E^*\}$ with the assignment matrix S. Then, the fragment-level node feature h^* is passed to a sum aggregator and an activation function to output a global feature u. The global feature u is then passed to fully connected layers and a sigmoid function to output a normalized prediction $p \in [0,1]$. Finally, 0 or 1 is outputted by comparing the p to a threshold value (default 0.5, for detail see Supplementary Data 5) to indicate as unbound or bound.

Datasets

Positive samples were derived from the Protein Data Bank. Structures were validated (seen Supplementary Data 4) to ensure high quality structures are used. To increase the reliability and accuracy, redundance cleaning was done for the positive database according to protein-sugar complex RMSD and interface RMSD. All positive samples

were relaxed by Rosetta [37] to alleviate potential clashes.

Negative samples were obtained from computational design. Experimental success rate of designed binders for small molecule is around 0.07% according to the latest RFAA design method [28], though with high scoring for energy- and structure-based metrics, e.g. AF2 pLDDT > 80, designed RMSD <2 Å , low Rosetta ddG (<30 in the case of digoxigenin binder) and at least one hydrogen bond to ligand. Therefore, poorly scoring designs can be reasonably assumed as failures. Here, A Rosetta design protocol (unpublished) was used to generate carbohydrate-protein complexes. Top 10% worst (ranked by binding free energy of Rosetta Energy Function) complexes were selected as negative samples. In general, those designed sugar binders are underperforming because of either lack of interactions (e.g. cavities, shallow clefts), or inaccurate interactions (e.g. clashes and distorted hydrogen bonds).

Model optimization

The training of CLIMBS was supervised by the binary 0/1 label of negative and positive samples. Mean absolute errors (MAE, also known as L1 loss) between the label and the prediction p were used for parameter optimization. The sugar-protein complex structure library comprises a ratio of positive and negative samples which is approximately 1:1. 15 types of mono-saccharides, 3 types of di-saccharides and a linear poly-saccharide with the corresponding binders, in total 13507 samples are in the library (see Supplementary Data 3a). The library was split into training, validation and test sets with the ratio of 2:2:1, while maintaining the 1:1 positive and negative ratio (for training detail see Supplementary Data 2).

Pooling layer selection

Fragment level is the best representation for pooling layer in CLIMBS, evaluated by a small-training assay. We have explored three representations of minimum binding site: atom-level, fragment-level and molecule-level representations, according to the scale of the molecular structure. To figure out which is the best representation before the sum aggregator, classifier models with those three different pooling layers were trained and tested. As a result in Supplementary table 3, the fragment-level pooling layer is the best for both mono-saccharide prediction and di-saccharide prediction.

Robustness check

A robustness check was done to unravel the contribution of factors to the performance of CLIMBS. Model robustness is the ability of model to maintain performance when faces adversarial data compared to the training data. Adversarial samples were generated by slightly modifying the original samples such as with masked edges and extra nodes or edges in the interacting map G. As a result in Supplementary table 4, it reveals a property of CLIMBS model that: the exist of CH-π interacting residue increases the probability to be predicted as positive results (bound), but it is not the only factor, which is consistent to the experimental result from Hudson [15].

VI. Reference

1 Mulichak, A. M., Wilson, J. E., Padmanabhan, K. & Garavito, R. M. The structure of mammalian hexokinase-1. *Nature Structural Biology* **5**, 555-560, doi:10.1038/811 (1998).

2 Williams, G. J. & Thorson, J. S. Natural product glycosyltransferases: properties and applications. *Adv Enzymol Relat Areas Mol Biol* **76**, 55-119, doi:10.1002/9780470392881.ch2 (2009).

3 Pfister, B. & Zeeman, S. C. Formation of starch in plant cells. *Cell Mol Life Sci* **73**, 2781-2807, doi:10.1007/s00018-016-2250-x (2016).

4 McNamara, J. T., Morgan, J. L. & Zimmer, J. A molecular description of cellulose biosynthesis. *Annu Rev Biochem* **84**, 895-921, doi:10.1146/annurev-biochem-060614-033930 (2015).

5 Geijtenbeek, T. B. & Gringhuis, S. I. Signalling through C-type lectin receptors: shaping immune responses. *Nat Rev Immunol* **9**, 465-479, doi:10.1038/nri2569 (2009).

6 Feizi, T. & Chai, W. Oligosaccharide microarrays to decipher the glyco code. *Nature Reviews Molecular Cell Biology* **5**, 582-588, doi:10.1038/nrm1428 (2004).

7 Lundstrom, J., Korhonen, E., Lisacek, F. & Bojar, D. LectinOracle: A Generalizable Deep Learning Model for Lectin-Glycan Binding Prediction. *Adv Sci (Weinh)* **9**, e2103807, doi:10.1002/advs.202103807 (2022).

8 Bojar, D. *et al.* A Useful Guide to Lectin Binding: Machine-Learning Directed Annotation of 57 Unique Lectin Specificities. *ACS Chem Biol* **17**, 2993-3012, doi:10.1021/acschembio.1c00689 (2022).

9 Nieto-Fabregat, F. *et al.* Computational toolbox for the analysis of protein-glycan interactions. *Beilstein J Org Chem* **20**, 2084-2107, doi:10.3762/bjoc.20.180 (2024).

10 Perez, S. Computational modeling of protein-carbohydrate interactions: Current trends and future challenges. *Adv Carbohydr Chem Biochem* **83**, 133-149, doi:10.1016/bs.accb.2023.10.003 (2023).

11 Tian, C. *et al.* ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *Journal of Chemical Theory and Computation* **16**, 528-552, doi:10.1021/acs.jctc.9b00591 (2020).

12 Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J Chem Theory Comput* **13**, 3031-3048, doi:10.1021/acs.jctc.7b00125 (2017).

13 Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* **30**, 2785-2791, doi:10.1002/jcc.21256 (2009).

14 Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J Chem Inf Model* **61**, 3891-3898, doi:10.1021/acs.jcim.1c00203 (2021).

15 Hudson, K. L. *et al.* Carbohydrate-Aromatic Interactions in Proteins. *J Am Chem Soc* **137**, 15152-15160, doi:10.1021/jacs.5b08424 (2015).

16 McNutt, A. T. *et al.* GNINA 1.0: molecular docking with deep learning. *J*

Cheminform **13**, 43, doi:10.1186/s13321-021-00522-2 (2021).
17 Wang, Y., Wei, Z. & Xi, L. Sfcnn: a novel scoring function based on 3D convolutional neural network for accurate and stable protein-ligand affinity prediction. *BMC Bioinformatics* **23**, 222, doi:10.1186/s12859-022-04762-3 (2022).
18 Pires, D. E. & Ascher, D. B. CSM-lig: a web server for assessing and comparing protein-small molecule affinities. *Nucleic Acids Res* **44**, W557-561, doi:10.1093/nar/gkw390 (2016).
19 Liao, Z. *et al.* DeepDock: Enhancing Ligand-protein Interaction Prediction by a Combination of Ligand and Structure Information. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 311-317 (2019).
20 Corso, G. a. S., Hannes and Jing, Bowen and Barzilay, Regina and Jaakkola, Tommi. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. *International Conference on Learning Representations (ICLR)* (2023).
21 Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500, doi:10.1038/s41586-024-07487-w (2024).
22 Kirschner, K. N. *et al.* GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *J Comput Chem* **29**, 622-655, doi:10.1002/jcc.20820 (2008).
23 Nguyen, T. B., Pires, D. E. V. & Ascher, D. B. CSM-carbohydrate: protein-carbohydrate binding affinity prediction and docking scoring function. *Brief Bioinform* **23**, doi:10.1093/bib/bbab512 (2022).
24 Buttenschoen, M., Morris, G. M. & Deane, C. M. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science* **15**, 3130-3139, doi:10.1039/D3SC04185A (2024).
25 Canner, S. W., Shanker, S. & Gray, J. J. Structure-based neural network protein-carbohydrate interaction predictions at the residue level. *Front Bioinform* **3**, 1186531, doi:10.3389/fbinf.2023.1186531 (2023).
26 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242, doi:10.1093/nar/28.1.235 (2000).
27 Jiang, T., Gradus, J. L. & Rosellini, A. J. Supervised Machine Learning: A Brief Primer. *Behav Ther* **51**, 675-687, doi:10.1016/j.beth.2020.05.002 (2020).
28 Krishna, R. *et al.* Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom. *bioRxiv*, 2023.2010.2009.561603, doi:10.1101/2023.10.09.561603 (2023).
29 Zhou, Z. *et al.* in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 4761-4772.
30 Yang, Z. *et al.* FragDPI: a novel drug-protein interaction prediction model based on fragment understanding and unified coding. *Front Comput Sci* **17**, 175903, doi:10.1007/s11704-022-2163-9 (2023).
31 Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. *Journal of the American Chemical Society* **125**, 1731-1737, doi:10.1021/ja026939x (2003).
32 Guo, S. *et al.* Structural Basis of Ligand Selectivity by a Bacterial Adhesin Lectin Involved in Multispecies Biofilm Formation. *mBio* **12**, 10.1128/mbio.00130-00121, doi:doi:10.1128/mbio.00130-21 (2021).
33 Sircar, A., Chaudhury, S., Kilambi, K. P., Berrondo, M. & Gray, J. J. A

generalized approach to sampling backbone conformations with RosettaDock for CAPRI rounds 13-19. *Proteins* **78**, 3115-3123, doi:10.1002/prot.22765 (2010).
34 Liu, Y. *et al.* Spherical Message Passing for 3D Graph Networks. *arXiv pre-print server*, doi:None

arxiv:2102.05013 (2021).
35 Corso, G., Cavalleri, L., Beaini, D., Li\`o, P. & Veli\vkovi\'c, P. Principal Neighbourhood Aggregation for Graph Nets. *arXiv pre-print server*, doi:None

arxiv:2004.05718 (2020).
36 Ying, R. *et al.* Hierarchical Graph Representation Learning with Differentiable Pooling. *arXiv pre-print server*, doi:None

arxiv:1806.08804 (2019).
37 Nivón, L. G., Moretti, R. & Baker, D. A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *PLOS ONE* **8**, e59004, doi:10.1371/journal.pone.0059004 (2013).

VII. Acknowledgements

Yijie Luo is under support of China Scholarship Council. Fabio Parmeggiani is the recipient of an EPSRC early career fellowship (EP/S017542/2).

VIII. Author Contributions:

The research project was conceived by Yijie Luo (YL) and Fabio Parmeggiani (FP). YL developed the code and collected the data. FP supervised the project. YL and FP analyzed the data and wrote the manuscripts.

IX. Competing interests:

The authors declare no competing interests.

X. Data and materials availability:

The code and CLIMBS model will be available as part of the Rosetta ML package.