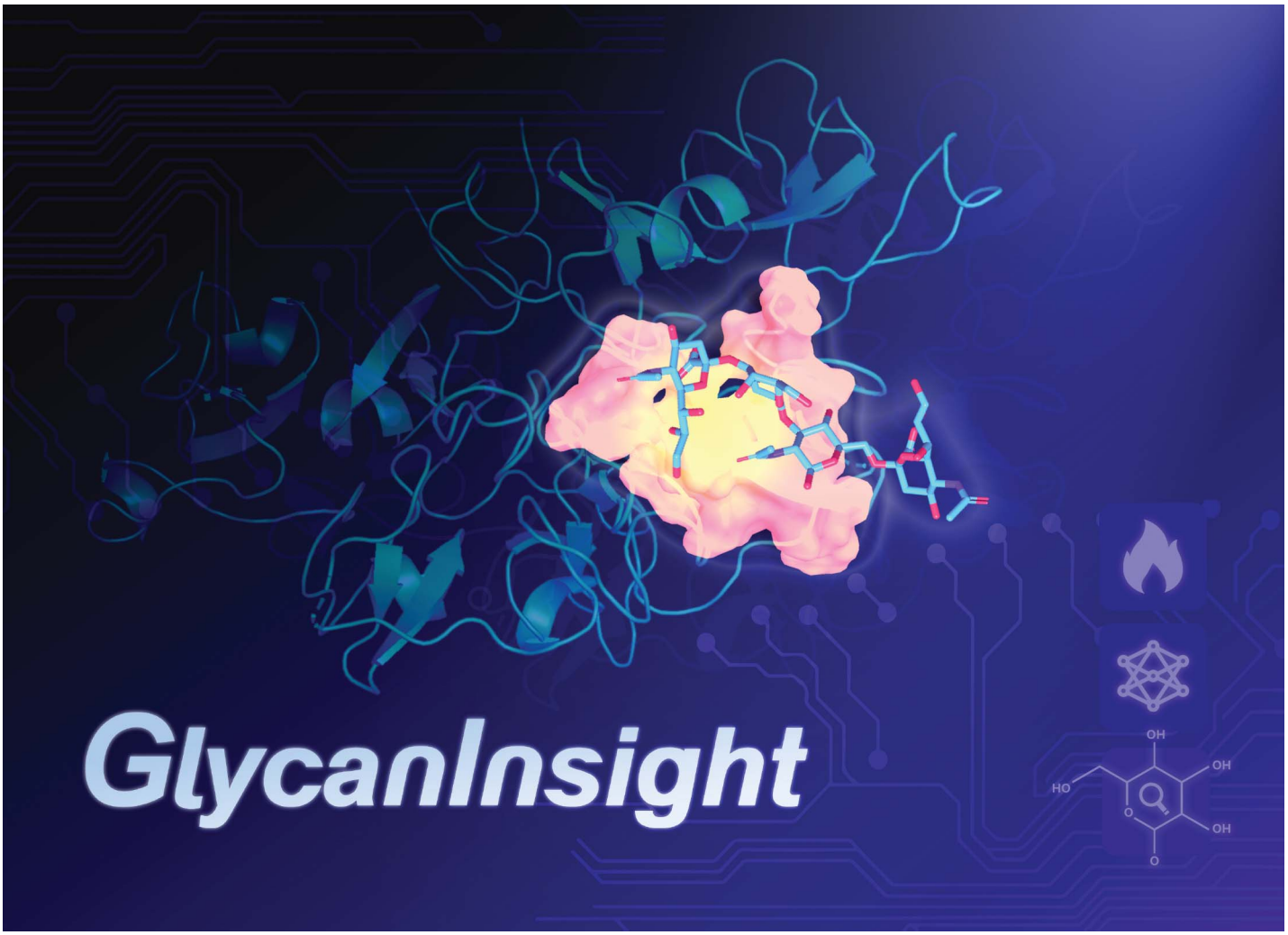

Showcasing research from Professor Cheng's laboratory, Department of Pharmaceutical and Artificial Intelligence Sciences, School of Medicine, Shanghai Jiao Tong University, Shanghai, China.

GlycanInsight: an open platform for carbohydrate-binding pocket prediction and characterization

GlycanInsight is an open platform for carbohydrate-binding pocket prediction and characterization. To provide insight into carbohydrate-protein interaction, it predicts carbohydrate-binding pockets on a protein structure, analyzes pocket characteristics, and suggests putative binding ligands. On the benchmark dataset of experimental structures, GlycanInsight achieves a high Matthews correlation coefficient of 0.63, outperforming existing tools, and maintains robust performance on AlphaFold2-predicted structures. By integrating precise prediction with automated structural annotation and ligand retrieval, GlycanInsight facilitates mechanistic studies and rational design of glycan-targeted therapeutics. The platform is freely accessible at http://www.glycaninsight.cn/.

Image reproduced by permission of Xi Cheng from *Chem. Sci.,* 2025, **16**, 10264.

As featured in:

See Mingyue Zheng, Xi Cheng *et al., Chem. Sci.,* 2025, **16**, 10264.

ROYAL SOCIETY OF CHEMISTRY

15 YEARS ANNIVERSARY

rsc.li/chemical-science

Registered charity number: 207890

Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2025, **16**, 10264

All publication charges for this article have been paid for by the Royal Society of Chemistry

Received 25th March 2025
Accepted 21st May 2025

DOI: 10.1039/d5sc02262b

rsc.li/chemical-science

GlycanInsight: an open platform for carbohydrate-binding pocket prediction and characterization†

Qinyu Chu,‡[abcd] Xinheng He,‡[cd] Xinyi Tan,‡[c] Zhiyong Gu,[acd] Yin Luo,[c] Zifu Huang,[c] Mingyue Zheng *[acd] and Xi Cheng *[bcd]

Carbohydrate–protein interactions underlie key physiological and pathological processes, yet identification of glycan-binding sites remains challenging due to the complexity of glycans and a lack of dedicated computational tools. We present GlycanInsight, a deep learning-based open platform that predicts carbohydrate-binding pockets on protein structures. On the benchmark dataset of experimental structures, GlycanInsight achieves a high Matthews correlation coefficient (MCC) of 0.63, outperforming existing tools, and maintains robust performance on AlphaFold2-predicted structures (MCC = 0.53). GlycanInsight clusters predicted residues into three-dimensional carbohydrate-binding pockets for detailed structural inspection, quantitatively analyzes pocket characteristics, searches for other proteins with similar pockets, and suggests putative binding ligands for the predicted pockets. By integrating precise prediction with automated structural annotation and ligand retrieval, GlycanInsight facilitates mechanistic studies and rational design of glycan-targeted therapeutics. The platform is freely accessible at https://www.glycaninsight.cn/.

Introduction

As a substance generally covering living cells in all organisms, carbohydrates (or glycans) interact with diverse protein families to regulate various biological and pathological processes.[1,2] Understanding how carbohydrates bind to and act on the protein therapeutic targets is of great significance for glycoscience and clinic translation.[3] In these applications, precise knowledge of carbohydrate-binding pockets on the proteins is required. However, experimental identification of carbohydrate-binding pockets is time-consuming and expensive, due to the complexity and flexibility of carbohydrates.[4] Hence, the development of a reliable carbohydrate-binding pocket predictor is important in understanding the carbohydrate–protein interactions.

However, few computational methods have been developed for predicting carbohydrate-binding pockets, and only a few are available as publicly accessible tools, including StackCBPred and PeSTo-Carbs.[5,6] StackCBPred uses a support vector machine model to learn from small datasets, and identifies carbohydrate-binding residues directly from protein sequences. This method utilizes a stacking-based approach with machine learning algorithms to enhance prediction accuracy, particularly by addressing the imbalance commonly found in datasets where non-binding residues vastly outnumber binding residues. PeSTo-Carbs is an extension of PeSTo (Protein Structure Transformer)[7] trained to predict protein–carbohydrate-interacting interfaces. PeSTo is a parameter-free geometric deep learning model, which has exhibited exceptional performance in predicting protein–protein binding interfaces.[7] As a carbohydrate-version of PeSTo, PeSTo-Carbs achieved a Matthews correlation coefficient (MCC) value of 0.475 for carbohydrate-binding interface prediction on an extensive test data set of 343 subunits.[6] In previous work, we have developed DeepGlycanSite, a deep learning-based model capable of accurately predicting carbohydrate-binding residues with the target protein structure.[8] Incorporating geometric and evolutionary features of proteins into a deep equivariant graph neural network with the transformer architecture, DeepGlycanSite has achieved state-of-the-art performance on binding residue prediction for diverse carbohydrates.[8] Although its source code is freely available, the application of DeepGlycanSite still requires knowledge and skills in computer science.

[a]*School of Pharmaceutical Science and Technology, Hangzhou Institute of Advanced Study, Hangzhou 330106, China. E-mail: myzheng@simm.ac.cn*

[b]*Medicinal Chemistry and Bioinformatics Center, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China. E-mail: xichengeva@sjtu.edu.cn*

[c]*State Key Laboratory of Drug Research and Information Management Office, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China*

[d]*University of Chinese Academy of Science, Beijing 100049, China*

† Electronic supplementary information (ESI) available: GlycanInsight-predicted carbohydrate-binding pockets of porcine pancreatic α-amylase. GlycanInsight-suggested ligand information of human galectin-3. GlycanInsight-predicted carbohydrate-binding pocket and experimentally identified carbohydrate-recognition domain of human galectin-3. GlycanInsight-predicted and experimentally identified carbohydrate-binding pockets of porcine pancreatic α-amylase. Carbohydrate-binding pocket prediction and analysis of Siglec-7, Klebsiella phage KP34p57 capsular depolymerase and GspB. See DOI: https://doi.org/10.1039/d5sc02262b

‡ These three authors contributed equally to this work.

The developments in glycoscience suffer from a lack of open, effective and easy-to-use computational tools. Here, we present GlycanInsight, a user-friendly open platform encapsulating DeepGlycanSite for carbohydrate-binding residue prediction. GlycanInsight allows users to employ an experimental structure (or a computational model) of a protein to predict the carbohydrate-binding residues, and clusters predicted residues into three-dimensional pockets for detailed structural inspection. Notably, it quantitatively analyzes pocket characteristics, searches for other proteins with similar pockets and suggests putative binding ligands, including carbohydrates and chemical compounds.

Experimental

Online platform implementation

The platform is based on Bootstrap and Django. On the frontend, Bootstrap is used as the main framework, offering responsive design and a rich component library to ensure the aesthetics and adaptability of the user interface. Django serves as the backend framework, providing secure and highly scalable server-side support. The NGL[9,10] framework is used to render and interactively display protein structures in the web browser. The RCSB dashboard[11] provides comprehensive access to biomolecular databases, enabling users to obtain, download, and analyze the required data within a single platform.

Workflow overview

On the main page of GlycanInsight, users can submit jobs without a login requirement. To perform the carbohydrate-binding pocket prediction of a protein, GlycanInsight needs the structure of the query protein. There are three ways for users to provide the protein information: (1) entering an existing PDB ID from the Protein Data Bank, (2) uploading their own PDB files, or (3) entering an existing UniProt ID from the AlphaFold protein structure database.[12,13] Given the input protein structure information, DeepGlycanSite is utilized by default to calculate the carbohydrate-binding probability of each residue.[8] GlycanInsight can also consider chemical information of a carbohydrate by employing DeepGlycanSite$_{+Ligand}$ instead of DeepGlycanSite. A URL is assigned to each submission so the user can access the results or track the processing status. The web server uses NGL Viewer[9,10] to present the predicted carbohydrate-binding residues in a molecular viewer, and employs RCSB dashboard[11] to show carbohydrate-binding probability values for all protein residues. The evolutionary conservation of each protein residue is estimated based on the UniRef50 sequence database (release 2021_03) using HMMER 3.3.2, which includes the Easel tools.[14–17] All predicted carbohydrate-binding residues are clustered into pockets using DBSCAN.[18] For each predicted carbohydrate-binding pocket, ProBiS[19,20] is used to search for similar ligand-binding pockets and corresponding ligands in a combined dataset consisting of the DeepGlycanSite dataset and the PDBbind database.[8,21] A link is provided to download a compressed file containing protein and pocket PDB files, visualization scripts and a list of prediction results.

Similar binding pocket search

DeepGlycanSite is a residue-centric prediction model, which predicts the carbohydrate-binding probability of each residue. GlycanInsight clusters the DeepGlycanSite-predicted carbohydrate-binding residues into pockets using DBSCAN, which separates areas of high density from areas of low density.[18] Residues with carbohydrate-binding probability values larger than 0.5 are selected for pocket clustering. The center-of-mass distances among selected residues are calculated. Moreover, only heavy atoms are considered in the calculation. Two selected residues are clustered as one pocket when their pairwise distance is less than 8 Å.

We constructed a dataset of 44 703 ligand-binding pockets for similar binding pocket search. We used 8102 carbohydrate-binding proteins from the DeepGlycanSite dataset[8] and 17 696 ligand-binding proteins from the PDBbind database[21] to extract ligand-binding pockets. A residue was classified as part of a ligand-binding pocket if the minimum distance between any of its heavy atoms and any heavy atom of the ligand was less than 4 Å. To define the solvent-accessible surface of these pockets, a probe with a radius of 1.4 Å was rolled over protein atoms represented as van der Waals spheres. Residues situated within 4 Å beneath this surface were then included in the analysis. These pockets were modeled as protein graphs, composed of vertices and edges that encapsulate both geometrical and physicochemical surface properties.[20,22] This representation was designed to capture potential interactions between the protein and its ligands. To make binding ligand suggestions, we employed the ProBiS algorithm, which facilitates local, surface-oriented comparisons of protein graphs. Using a fast maximum clique algorithm, ProBiS identifies all possible similar regions between two proteins.[23] It operates independently of their fold or sequence. Each identified maximum clique, defined by its rotational and translational variations, represents a rigid local similarity and is used to superimpose the protein structures. Subsequently, local backbone alignment of the superimposed structures is performed to uncover additional similarities that may have been overlooked by the maximum clique approach.

Datasets

The T145 test dataset involves 145 carbohydrate–protein complexes. In this dataset, any protein with more than 95% sequence identity to the training (or validation) sets was excluded.[8] We also employed AlphaFold2 [13] and AlphaFold2 Multimer[24] to predict protein structures based on the protein sequences of T145.[8] The top five ranked conformation models for each protein were selected to construct an independent testing set T145$_{AF2}$, consisting of 145 unique proteins and 725 apo structure models.

Evaluation metrics

There are mainly three metrics used to evaluate carbohydrate-binding pocket detection algorithms. The metrics are the

Matthews correlation coefficient (MCC), precision and balanced accuracy. For a given protein, the carbohydrate-binding residues are positives, while the others are negatives. Correctly predicted carbohydrate-binding residues are true positives (TP). Correctly predicted non-carbohydrate-binding residues are true negatives (TN). Incorrectly predicted carbohydrate-binding residues are false positives (FP). Incorrectly predicted non-carbohydrate-binding residues are false negatives (FN).

The MCC is defined in eqn (1):

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP}) \times (\mathrm{TP}+\mathrm{FN}) \times (\mathrm{TN}+\mathrm{FP}) \times (\mathrm{TN}+\mathrm{FN})}} \quad (1)$$

The MCC ranges from −1 to 1. A small value of −1 indicates that no carbohydrate-binding residue is correctly predicted; a large value of 1 indicates that all carbohydrate-binding residues are correctly predicted.

The precision is defined in eqn (2):

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \quad (2)$$

The balanced accuracy is defined in eqn (3):

$$\mathrm{Balanced\ accuracy} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right) \quad (3)$$

Results and discussion

GlycanInsight workflow

As shown in Fig. 1, users can either upload a three-dimensional structure in the PDB format or provide a PDB ID of a protein, in which case GlycanInsight will retrieve the corresponding PDB file from the PDB database. When no experimental structure is available, users can provide a UniProt ID of a protein, in which case GlycanInsight will retrieve the corresponding AlphaFold2-predicted structure from the AlphaFold protein structure database. Users can also specify a carbohydrate ligand for the query protein by uploading a chemical structure file in SDF, PDB or MOL2 format. The two-dimensional chemical structure information of the carbohydrate ligand will be extracted from the file for prediction.

GlycanInsight uses DeepGlycanSite to predict the carbohydrate-binding probabilities for all protein residues. When provided with a file containing the two-dimensional carbohydrate structure, the platform employs the ligand-specific DeepGlycanSite$_{+\mathrm{Ligand}}$ model to incorporate query ligand information. Residues with predicted probabilities larger than 0.5 are identified to bind carbohydrates. To provide a more intuitive insight into these residues, GlycanInsight converts residue-centric predictions to a pocket-centric perspective. To quantitatively characterize each pocket, the platform calculates its geometric center and estimates evolutionary conservation. Assuming similar pockets bind similar ligands, GlycanInsight compares predicted pockets with the reported protein–ligand complexes in the PDB database. All complexes with high alignment scores (indicating similarity to predicted pockets) are

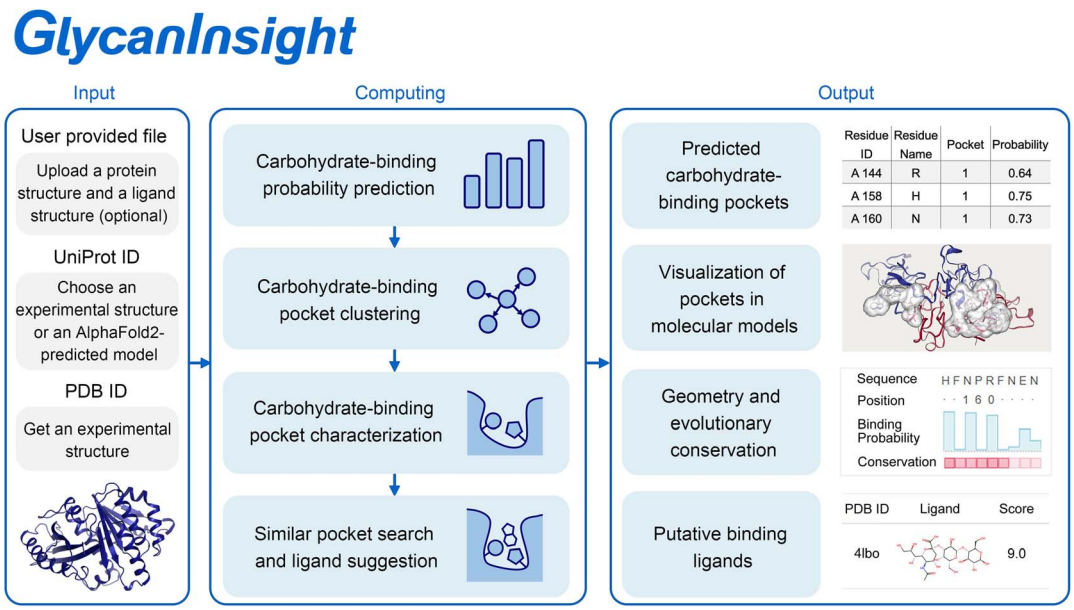

Fig. 1 Overall view of GlycanInsight workflow. For input, users can submit a protein structure in PDB format, a UniProt ID, or a PDB ID to predict carbohydrate-binding pockets. They can also submit a carbohydrate in SDF, PDB or MOL2 formats to offer extra ligand information. For computing, the platform predicts carbohydrate-binding residues and clusters them into three-dimensional pockets for characterization, similar pocket search and putative ligand suggestion. For output, predicted carbohydrate-binding pockets, geometry and evolutionary conservation annotations, similar pockets, and suggested ligands are presented in an interactive graphical interface and are available for download.

Table 1 Carbohydrate-binding pocket prediction on the T145 dataset[a]

Web server	MCC	Precision	Balanced accuracy
StackCBPred	0.02 ± 0.09***	0.05 ± 0.03***	0.53 ± 0.10***
GRaSP-web	0.17 ± 0.46***	0.31 ± 0.32***	0.63 ± 0.15***
Fpocket	0.19 ± 0.32***	0.19 ± 0.28***	0.62 ± 0.20***
PeSTo-Carbs	0.34 ± 0.26***	0.27 ± 0.21***	0.77 ± 0.18*
PrankWeb 3	0.37 ± 0.28***	0.25 ± 0.19***	0.83 ± 0.16
GlycanInsight	0.63 ± 0.29	0.63 ± 0.31	0.83 ± 0.16

[a] Data represent means ± standard deviation. The two-tailed Mann–Whitney *U* test is used to determine the statistical difference between GlycanInsight and an alternative web server. *** indicates *P* is less than 0.001. * indicates *P* is less than 0.1.

listed, and their ligands are suggested as putative ligands for predicted pockets.

GlycanInsight provides a graphical user interface for interactive inspection of all results. It enables users to display predicted results through both structural and sequence perspectives. It also provides detailed information of predicted carbohydrate-binding pockets and putative binding ligands. In addition to its online visualization tools, the system also provides the option to export results as a PyMOL script for offline exploration. Users can download all prediction, analysis, and visualization results from the dedicated output page.

Table 2 Carbohydrate-binding pocket prediction on the $T145_{AF2}$ dataset[a]

Web server	MCC	Precision	Balanced accuracy
StackCBPred	0.02 ± 0.09***	0.05 ± 0.03***	0.53 ± 0.10***
GRaSP-web	0.03 ± 0.35***	0.18 ± 0.23***	0.55 ± 0.07***
Fpocket	−0.07 ± 0.50***	0.13 ± 0.19***	0.57 ± 0.12***
PeSTo-Carbs	0.31 ± 0.25***	0.25 ± 0.19***	0.73 ± 0.17***
PrankWeb 3	0.13 ± 0.24***	0.13 ± 0.12***	0.62 ± 0.12***
GlycanInsight	0.53 ± 0.28	0.48 ± 0.28	0.82 ± 0.15

[a] Data represent means ± standard deviation. The two-tailed Mann–Whitney *U* test is used to determine the statistical difference between GlycanInsight and an alternative web server. *** indicates *P* is less than 0.001.

Prediction performance

We compared the prediction performance of GlycanInsight with five competing tools, using an independent dataset T145 of 145 different carbohydrate-binding proteins with experimental structures.[8] StackCBPred and PeSTo-Carbs are the only accessible carbohydrate-binding pocket prediction web servers. PrankWeb 3[25] and GRaSP-web[26] are two state-of-the-art web servers for ligand-binding pocket prediction. Fpocket[27] is a long-standing web server for ligand-binding pocket prediction. The metrics of MCC, precision and balanced accuracy were used to measure the predictive performance of the assessed web servers. As shown in Table 1, GlycanInsight clearly outperformed the other five web servers. It achieved the average MCC and precision more than 0.6, whereas the competitors had the average MCC and precision less than 0.4.

Structure-based binding pocket prediction requires a protein structure. Although the number of experimentally determined protein structures continues to escalate, it still trails behind the number of recognized protein sequences.[28] Recent breakthroughs in protein structure prediction, particularly the emergence of AlphaFold2 and the AlphaFold protein structure database,[12,13,24] have paved the way for applying structure-based approaches to proteins without experimentally determined structures. This development motivated us to adopt the predicted protein structures in GlycanInsight, allowing users to enter a UniProt ID as the input. We used an AlphaFold2-predicted protein structure dataset

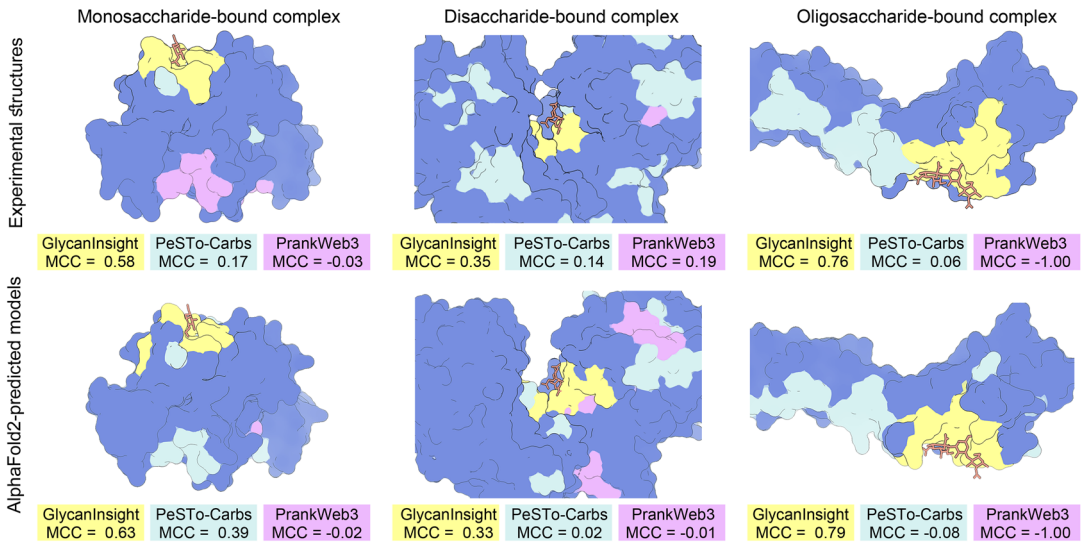

Fig. 2 Saccharide-binding pocket prediction of different approaches for three representatives. Predictions of GlycanInsight (*yellow*), PeSTo-Carbs (*cyan*) and PrankWeb 3 (*pink*) were mapped on the given protein structures. Experimental structures are shown on the top and AlphaFold2-predicted protein models are shown on the bottom. Saccharides are displayed as sticks to indicate the true binding sites of a *N*-acetylglucosamine-binding lectin (PDB ID: 6stn), a chondroitin sulfate unit-binding lyase (PDB ID: 7eiq) and a sialoglycan-binding siglec (PDB ID: 6x3q).

($T145_{AF2}$) to compare GlycanInsight with five competing web servers on carbohydrate-binding pocket prediction. These predicted protein structures were generated based on the protein sequences of T145. As shown in Table 2, GlycanInsight still achieved the average MCC of 0.53 and the average precision of 0.45, remarkably outperforming the other web servers. PeSTo-Carbs had the average MCC of 0.31 and the average precision of 0.24. PrankWeb 3 and the other web servers had the average MCC and precision less than 0.2.

We further analyzed the capabilities of GlycanInsight, PeSTo-Carbs and PrankWeb 3 in predicting binding pockets for different carbohydrates. Fig. 2 displays the prediction results for a *N*-acetyl-glucosamine-binding (*i.e.* monosaccharide-binding) lectin,[29] a chondroitin sulfate unit-binding (*i.e.* disaccharide-binding) lyase,[30] and a sialoglycan-binding (*i.e.* oligosaccharide-binding) siglec.[31] GlycanInsight effectively identified the monosaccharide-, disaccharide- and oligosaccharide-binding pockets on both experimental and AlphaFold2-predicted protein structures, highlighting its generalized applicability. In contrast, PeSTo-Carbs showed an MCC value of 0.39 in a monosaccharide-binding pocket prediction case, but had MCC values less than 0.2 in all the other cases. PrankWeb 3 had MCC values less than 0.2 in all cases.

Case study

We demonstrated the functionality of GlycanInsight with a well-known carbohydrate-binding protein, *i.e.* galectin-3. Galectin-3 is a β-galactoside-binding lectin, regulating cell migration, immune response and tissue remodelling.[32–35] Due to its important biological functions, galectin-3 has been employed as a diagnostic marker and therapeutic target in clinical applications.[34,36] To predict without experimental protein structure,

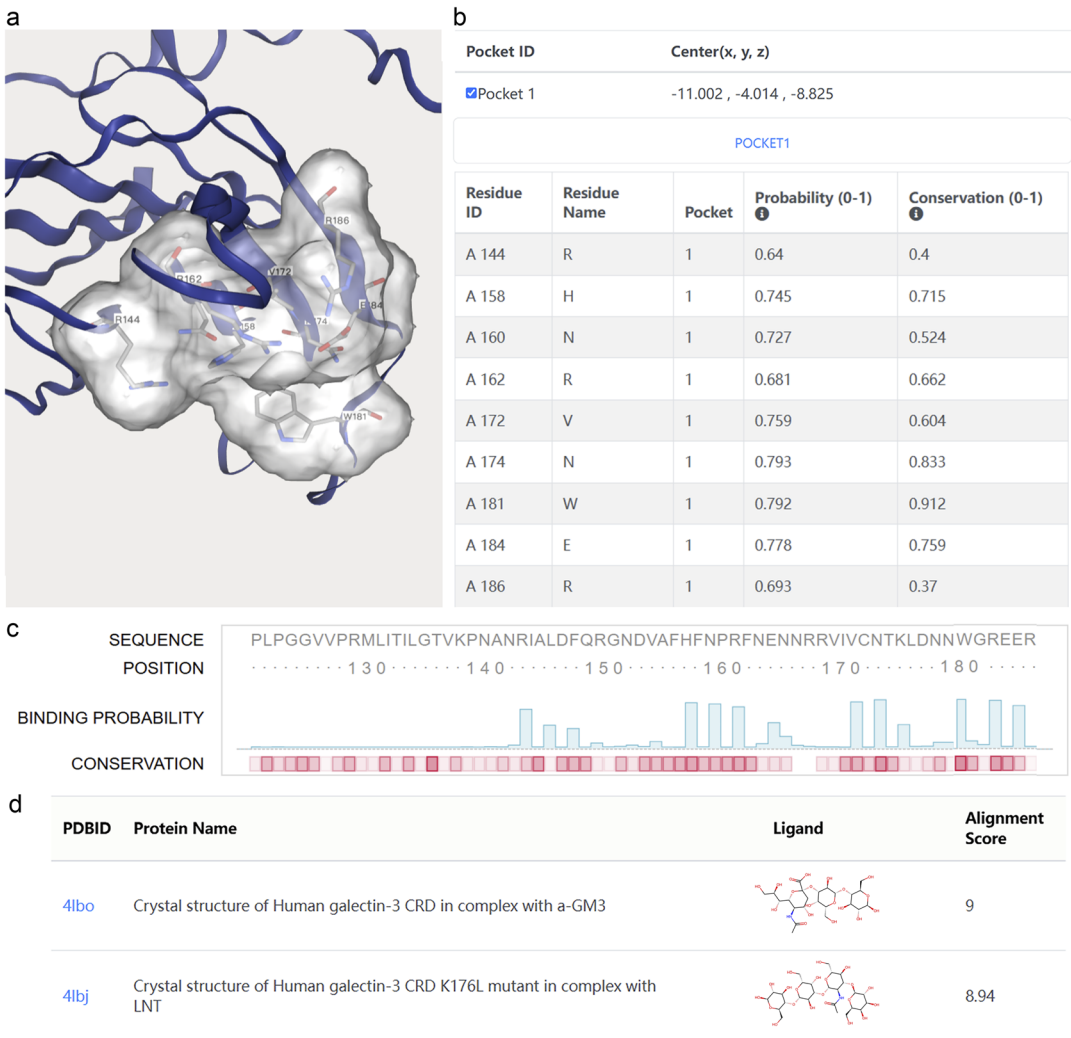

Fig. 3 User-friendly graphical interface of GlycanInsight. (a) An interactive molecular viewer of the predicted carbohydrate-binding pocket on the query protein. The predicted pocket is displayed as a surface and sticks. Users can rotate, zoom, and toggle individual residues. (b) Geometry and composition information of the predicted pocket. The center coordinates of the pocket are presented. Residues in the pocket are listed with carbohydrate-binding probabilities and conservation scores. (c) Sequence annotations of the predicted pocket. A blue bar chart indicates the carbohydrate-binding probability for each residue. Conservation scores of residues are represented by transparency of red color, with more conserved residues appearing as deeper red. (d) Putative binding ligands are suggested for the predicted pocket, which are presented with their chemical structures and scores.

a user can start a new prediction job by choosing "UniProt ID" tab. A user can enter the UniProt ID of P17931 for the human galectin-3, click "Search" button, choose "AlphaFoldDB – P17931" and then click "Download" button to retrieve the predicted structure of the galectin-3. After clicking "Submission" button, the prediction and analysis processes are initiated and take a few minutes. When these processes are completed, results are displayed on the output page.

As shown in Fig. 3a–d, the output page of GlycanInsight is split into a visualization part and three results sections. The visualization part contains a molecular viewer that displays the protein structure in a cartoon representation and exhibits the predicted carbohydrate-binding pockets as surface and sticks (Fig. 3a). The viewer enables users to interactively inspect pockets within the protein structure in detail.

The "Pocket Prediction" section shows quantitative characteristics of the predicted carbohydrate-binding pockets, including geometrical location, residue composition, carbohydrate-binding probability and evolutionary conservation (Fig. 3b). This section is interactively coupled with the molecular viewer. Users can choose to inspect a pocket in the molecular viewer window by choosing the pocket ID in this section. For each predicted pocket, its geometrical center coordinates are calculated, and the residues in composition are listed, along with residue IDs, residue names, carbohydrate-binding probabilities and evolutionary conservation scores. GlycanInsight defines a residue with carbohydrate-binding probability greater than 0.5 as a carbohydrate-binding one, and considers a residue with evolutionary conservation score greater than 0.5 to be a conserved one. In this case, nine residues (R144, H158, N160, R162, V172, N174, W181, E184 and R186) are predicted to form a carbohydrate-binding pocket. This predicted pocket highly overlaps with known carbohydrate-recognition domains of galectin-3,[37,38] consisting of R144, H158, N160, R162, E165, N174, W181, E184 and R186 (Fig. S1†). And most of them exhibited conservation scores greater than 0.5, suggesting a highly conserved carbohydrate-binding motif of galectin-3 (Fig. 3b). GlycanInsight can identify multiple carbohydrate-binding pockets for a single protein, thereby facilitating identification of secondary binding pockets. For example, α-amylases randomly cleave α-glucans and possess several secondary binding pockets to enhance their cleavage efficiency.[39,40] More than one substrate bond is cleaved in a single enzyme–substrate encounter.[41,42] For the Porcine pancreatic α-amylase (UniProt ID: P00690), GlycanInsight identified one catalytic pocket and four secondary carbohydrate-binding pockets, consistent with experimental observations (Fig. S1 and Table S1†).[43–45] These secondary binding pockets may assist the catalytic pocket in binding helical amylose chains.

The "Sequence Analysis" section presents the amino acid sequence view of the predictions, including carbohydrate-binding probability and evolutionary conservation score for each residue (Fig. 3c). Carbohydrate-binding probability is displayed as a column chart. The evolutionary conservation scores are represented by the transparency of red squares. This dual visualization enables intuitive interpretation of the predicted carbohydrate-binding residues.

The "Suggested Ligands" section lists putative binding ligands for the predicted carbohydrate-binding pockets (Fig. 3d). Assuming similar binding pockets tend to bind to similar ligands, GlycanInsight suggests which ligand has the potential to bind to a predicted carbohydrate-binding pocket, *via* comparing each predicted pocket with the known ligand-binding pocket. A table is provided for all ligand candidates. It includes ligand chemical structure diagrams, alignment scores between a predicted pocket and a known binding pocket, PDB ID and name of a protein containing the known binding pocket. In this case study, a total of 157 reported ligand-binding pockets were identified to be similar to the predicted carbohydrate-binding pocket of the galectin-3 (Table S2†). This result reveals that more than 20 proteins share conserved carbohydrate-binding functional motifs with galectin-3 (Table 3). Such carbohydrate-binding motifs can interact with various saccharides, including 7 monosaccharides, 17 disaccharides and 20 oligosaccharides (Table S3†). In particular, lactose (Gal(β1-4)Glc) binds to ten proteins, whereas LacNAc (Gal(β1-4)GlcNAc) interacts with six proteins (Table S3†). In addition to typical saccharides, 25 carbohydrate-based ligand candidates of galectin-3 are also indicated, including two galectin-1 inhibitors, three galectin-7 inhibitors, three galectin-8 inhibitors, and one galectin-9 inhibitor (Table S4†). And 20 reported galectin-3 inhibitors were also listed,

Table 3 Proteins containing similar pocket (with an alignment score more than 5.00) as the predicted-carbohydrate-binding pockets of human galectin-3

Index	Protein name	Best alignment score	PDB ID
1	Galectin-9	8.27	3wv6
2	Galectin-7	8.22	2gal
3	Galectin-1	8.12	6e20
4	Galectin-4	8.10	5duw
5	Porcine adenovirus 4 fiber protein	8.05	2wsv
6	*Cyclocybe cylindracea* galectin	8.00	3wg3
7	Galectin-8	7.98	3vkl
8	Chicken GRIFIN	7.95	5nle
9	Tl-gal	7.93	5glz
10	*Xenopus laevis* galectin-Ib	7.81	3wud
11	Galectin-2	7.77	5ews
12	*Caenorhabditis elegans* galectin LEC-6	7.75	3vv1
13	Congerin II	7.73	1wld
14	Galectin-5	7.65	5jpg
15	Congerin I	7.64	1c1l
16	S-LAC lectin	7.63	1hlc
17	*Xenopus laevis* galectin-Va	7.27	3wuc
18	Galectin-10	7.00	6l6a
19	Galectin-13	6.84	6a63
20	Fungal lectin CGL3	6.73	2r0h
21	Galectin-11	6.22	6n3r
22	Galectin-16	6.17	6ljr

including one with clinical trials (Table S4†). Olitigaltin is an inhaled small molecule inhibitor of galectin-3 with a clinical trial phase II for idiopathic pulmonary fibrosis and COVID-19.[46–49] These findings provide valuable ligand information complementing the carbohydrate-binding pocket analysis for biological study and drug development of galectin-3. See more example cases in the ESI (Fig. S3–S5†). While GlycanInsight enhances the prediction of carbohydrate-binding pockets, its functionalities of similar pocket search and ligand suggestion rely on two key factors: (1) the completeness of the reference pocket-ligand database and (2) the parameter settings (*e.g.*, the alignment score threshold). Users are advised to interpret results in the context of these dependencies, and future updates will prioritize expanding the reference database and introducing user-customizable parameters. All results can be downloaded by clicking the "Download Files" section tab. A PyMol script is also provided for offline inspection of results.

Conclusions

GlycanInsight provides free access to an easy-to-use online service for carbohydrate-binding pocket prediction, inspection and analysis. It is capable of predicting binding pockets for diverse carbohydrates, outperforming alternative web servers. To offer more insights into carbohydrate-binding pockets, GlycanInsight not only provides a graphical user interface for visualizing prediction results, but also performs evolutionary conservation calculation and makes potential binding-ligand suggestion. Knowledge of carbohydrate binding enables research ranging from function annotation to rational drug design. We believe that GlycanInsight can provide valuable insights for glycoscience and carbohydrate-based drug development.

Data availability

The data supporting this article have been included as part of the ESI.† GlycanInsight and all datasets are freely available at **https://www.glycaninsight.cn/**.

Author contributions

Qinyu Chu: methodology, visualization, writing—original draft. Xinheng He: methodology, data curation, writing, review & editing. Xinyi Tan: software, visualization, writing, review & editing. Zhiyong Gu: formal analysis. Yin Luo: software. Zifu Huang: resources. Mingyue Zheng: supervision, funding acquisition, review & editing. Xi Cheng: conceptualization, data curation, methodology, supervision, funding acquisition, review & editing.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was partially supported by Shanghai Municipal Science and Technology Major Project; National Key Research and Development Program of China (2021YFA1301900); the Strategic Priority Research Program of Chinese Academy of Sciences (XDB0830000); Fund of Youth Innovation Promotion Association (2022077); Fund of Shanghai Oriental Elite Talent Program (QNKJ2024005).

References

1 A. Varki, Biological roles of glycans, *Glycobiology*, 2017, **27**, 3–49.
2 B. Ernst and J. L. Magnani, From carbohydrate leads to glycomimetic drugs, *Nat. Rev. Drug Discovery*, 2009, **8**, 661–677.
3 B. A. H. Smith and C. R. Bertozzi, The clinical impact of glycobiology: targeting selectins, Siglecs and mammalian glycans, *Nat. Rev. Drug Discovery*, 2021, **20**, 217–243.
4 M. E. Griffin and L. C. Hsieh-Wilson, Tools for mammalian glycoscience research, *Cell*, 2022, **185**, 2657–2677.
5 S. Gattani, A. Mishra and M. T. Hoque, StackCBPred: A stacking based prediction of protein-carbohydrate binding sites from sequence, *Carbohydr. Res.*, 2019, **486**, 107857.
6 P. Bibekar, L. Krapp and M. Dal Peraro, PeSTo-Carbs: Geometric Deep Learning for Prediction of Protein-Carbohydrate Binding Interfaces, *J. Chem. Theory Comput.*, 2024, **20**, 2985–2991.
7 L. F. Krapp, L. A. Abriata, F. C. Rodriguez and M. Dal Peraro, PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces, *Nat. Commun.*, 2023, **14**, 2175–2185.
8 X. H. He, L. F. Zhao, Y. P. Tian, R. Li, Q. Y. Chu, Z. Y. Gu, M. Y. Zheng, Y. S. Wang, S. N. Li, H. L. Jiang, Y. Jiang, L. Q. Wen, D. Y. Wang and X. Cheng, Highly accurate carbohydrate-binding site prediction with DeepGlycanSite, *Nat. Commun.*, 2024, **15**, 5163–5175.
9 A. S. Rose, A. R. Bradley, Y. Valasatava, J. M. Duarte, A. Prlić and P. W. Rose, NGL viewer: web-based molecular graphics for large complexes, *Bioinformatics*, 2018, **34**, 3755–3758.
10 A. S. Rose and P. W. Hildebrand, NGL Viewer: a web application for molecular visualization, *Nucleic Acids Res.*, 2015, **43**, W576–W579.
11 J. Segura, Y. Rose, J. Westbrook, S. K. Burley and J. M. Duarte, RCSB Protein Data Bank 1D tools and services, *Bioinformatics*, 2021, **36**, 5526–5527.
12 M. A.-O. Varadi, S. Anyango, M. A.-O. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. A.-O. Birney, D. Hassabis and S. A.-O. Velankar, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models, *Nucleic Acids Res.*, 2022, 439–444.

13 J. A.-O. Jumper, R. Evans, A. Pritzel, T. A.-O. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A.-O. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. A.-O. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. A.-O. Steinegger, M. A.-O. Pacholska, T. Berghammer, S. Bodenstein, D. A.-O. Silver, O. Vinyals, A. A.-O. Senior, K. Kavukcuoglu, P. Kohli and D. A.-O. Hassabis, Highly accurate protein structure prediction with AlphaFold, *Nature*, 2021, 583–589.

14 B. E. Suzek, Y. Q. Wang, H. Z. Huang, P. B. McGarvey, C. H. Wu and U. Consortium, UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches, *Bioinformatics*, 2015, **31**, 926–932.

15 S. C. Potter, A. Luciani, S. R. Eddy, Y. Park, R. Lopez and R. D. Finn, HMMER web server: 2018 update, *Nucleic Acids Res.*, 2018, **46**, W200–W204.

16 R. D. Finn, J. Clements, W. Arndt, B. L. Miller, T. J. Wheeler, F. Schreiber, A. Bateman and S. R. Eddy, HMMER web server: 2015 update, *Nucleic Acids Res.*, 2015, **43**, W30–W38.

17 R. D. Finn, J. Clements and S. R. Eddy, HMMER web server: interactive sequence similarity searching, *Nucleic Acids Res.*, 2011, **39**, W29–W37.

18 E. Schubert, J. r. Sander, M. Ester, H. P. Kriegel and X. Xu, DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN, *ACM Trans. Database Syst.*, 2017, **42**, 1–21.

19 J. Konc, M. Depolli, R. Trobec, K. Rozman and D. Janežič, Parallel-ProBiS: Fast parallel algorithm for local structural comparison of protein structures and binding sites, *J. Comput. Chem.*, 2012, **33**, 2199–2203.

20 J. Konc and D. Janežič, ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment, *Bioinformatics*, 2010, **26**, 1160–1168.

21 Z. H. Liu, Y. Li, L. Han, J. Li, J. Liu, Z. X. Zhao, W. Nie, Y. C. Liu and R. X. Wang, PDB-wide collection of binding data: current status of the PDBbind database, *Bioinformatics*, 2015, **31**, 405–412.

22 D. Kuhn, N. Weskamp, S. Schmitt, E. Hüllermeier and G. Klebe, From the Similarity Analysis of Protein Cavities to the Functional Classification of Protein Families Using Cavbase, *J. Mol. Biol.*, 2006, **359**, 1023–1044.

23 J. Konc and D. J. p. Janezic, An improved branch and bound algorithm for the maximum clique problem, *MATCH Communications in Mathematical and in Computer Chemistry*, 2007, **4**, 590–596.

24 R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper and D. Hassabis, Protein complex prediction with AlphaFold-Multimer, *bioRxiv*, 2021, preprint, 2021.2010.2004.463034, DOI: **10.1101/2021.10.04.463034**.

25 D. Jakubec, P. Skoda, R. Krivak, M. Novotny and D. Hoksza, PrankWeb 3: accelerated ligand-binding site predictions for experimental and modelled protein structures, *Nucleic Acids Res.*, 2022, **50**, W593–W597.

26 C. A. Santana, S. C. Izidoro, R. C. de Melo-Minardi, J. D. Tyzack, A. J. M. Ribeiro, D. E. V. Pires, J. M. Thornton and S. d. A. Silveira, GRaSP-web: a machine learning strategy to predict binding sites based on residue neighborhood graphs, *Nucleic Acids Res.*, 2022, **50**, W392–W397.

27 V. Le Guilloux, P. Schmidtke and P. Tuffery, Fpocket: An open source platform for ligand pocket detection, *BMC Bioinf.*, 2009, **10**, 168.

28 A. L. Mitchell, A. Almeida, M. Beracochea, M. Boland, J. Burgin, G. Cochrane, M. R. Crusoe, V. Kale, S. C. Potter, L. J. Richardson, E. Sakharova, M. Scheremetjew, A. Korobeynikov, A. Shlemov, O. Kunyayskaya, A. Lapidus and R. D. Finn, MGnify: the microbiome analysis resource in 2020, *Nucleic Acids Res.*, 2020, **48**, D570–D578.

29 M. Perduca, M. Bovi, L. Destefanis, D. Nadali, L. Fine, F. Parolini, D. Sorio, M. E. Carrizo and H. L. Monaco, Structure and properties of the giant reed (Arundo donax) lectin (ADL), *Glycobiology*, 2021, **31**, 1543–1556.

30 M. Takashima, I. Watanabe, A. Miyanaga and T. Eguchi, Substrate specificity of Chondroitinase ABC I based on analyses of biochemical reactions and crystal structures in complex with disaccharides, *Glycobiology*, 2021, **31**, 1571–1581.

31 B. A. Bensing, H. E. Stubbs, R. Agarwal, I. Yamakawa, K. Luong, K. Solakyildirim, H. Yu, A. Hadadianpour, M. A. Castro, K. P. Fialkowski, K. M. Morrison, Z. Wawrzak, X. Chen, C. B. Lebrilla, J. Baudry, J. C. Smith, P. M. Sullam and T. M. Iverson, Origins of glycan selectivity in streptococcal Siglec-like adhesins suggest mechanisms of receptor adaptation, *Nat. Commun.*, 2022, **13**, 2753–2766.

32 N. C. Henderson and T. Sethi, The regulation of inflammation by galectin-3, *Immunol. Rev.*, 2009, **230**, 160–271.

33 T. Funasaka, A. Raz and P. Nangia-Makker, Galectin-3 in angiogenesis and metastasis, *Glycobiology*, 2014, **24**, 886–891.

34 R. J. Slack, R. Mills and A. C. Mackinnon, The therapeutic potential of galectin-3 inhibition in fibrotic disease, *Int. J. Biochem. Cell Biol.*, 2021, **130**, 2753–2766.

35 L. C. Li, J. Li and J. Gao, Functions of Galectin-3 and Its Role in Fibrotic Diseases, *J. Pharmacol. Exp. Ther.*, 2014, **351**, 336–343.

36 R. Dong, M. Zhang, Q. Y. Hu, S. Zheng, A. Soh, Y. J. Zheng and H. Yuan, Galectin-3 as a novel biomarker for disease diagnosis and a target for therapy, *Int. J. Mol. Med.*, 2018, **41**, 599–614.

37 J. Seetharaman, A. Kanigsberg, R. Slaaby, H. Leffler, S. H. Barondes and J. M. Rini, X-ray Crystal Structure of the Human Galectin-3 Carbohydrate Recognition Domain at 2.1-Å Resolution, *J. Biol. Chem.*, 1998, **273**, 13047–13052.

38 K. Saraboji, M. Håkansson, S. Genheden, C. Diehl, J. Qvist, U. Weininger, U. J. Nilsson, H. Leffler, U. Ryde, M. Akke and D. T. Logan, The Carbohydrate-Binding Site in

Galectin-3 Is Preorganized To Recognize a Sugarlike Framework of Oxygens: Ultra-High-Resolution Structures and Water Dynamics, *Biochemistry*, 2012, **51**, 296–306.

39 M. M. Nielsen, E. S. Seo, S. Bozonnet, N. Aghajari, X. Robert, R. Haser and B. Svensson, Multi-site substrate binding and interplay in barley α-amylase 1, *FEBS Lett.*, 2008, **582**, 2567–2571.

40 B. Kramhoft, K. S. Bak-Jensen, H. Mori, N. Juge, J. Nohr and B. Svensson, Involvement of individual subsites and secondary substrate binding sites in multiple attack on amylose by barley α-amylase, *Biochemistry*, 2005, **44**, 1824–1832.

41 M. M. Nielsen, S. Bozonnet, E. S. Seo, J. A. Mótyán, J. M. Andersen, A. Dilokpimol, M. Abou Hachem, G. Gyémánt, H. Næsted, L. Kandra, B. W. Sigurskjold and B. Svensson, Two Secondary Carbohydrate Binding Sites on the Surface of Barley α-Amylase 1 Have Distinct Functions and Display Synergy in Hydrolysis of Starch Granules, *Biochemistry*, 2009, **48**, 7686–7697.

42 M. M. Nielsen, E. S. Seo, A. Dilokpimol, J. Andersen, M. Abou Hachem, H. Naested, M. Willemoës, S. Bozonnet, L. Kandra, G. Gyémánt, R. Haser, N. Aghajari and B. Svensson, Roles of multiple surface sites, long substrate binding clefts, and carbohydrate binding modules in the action of amylolytic enzymes on polysaccharide substrates, *Biocatal. Biotransform.*, 2008, **26**, 59–67.

43 F. Payan and M. X. Qian, Crystal structure of the pig pancreatic α-amylase complexed with malto-oligosaccharides, *J. Protein Chem.*, 2003, **22**, 275–284.

44 S. B. Larson, J. S. Day and A. McPherson, X-ray Crystallographic Analyses of Pig Pancreatic α-Amylase with Limit Dextrin, Oligosaccharide, and α-Cyclodextrin, *Biochemistry*, 2010, **49**, 3101–3115.

45 M. Machius, L. Vertesy, R. Huber and G. Wiegand, Carbohydrate and protein-based inhibitors of porcine pancreatic alpha-amylase: Structure analysis and comparison of their binding characteristics, *J. Mol. Biol.*, 1996, **260**, 409–421.

46 T. Delaine, P. Collins, A. MacKinnon, G. Sharma, J. Stegmayr, V. K. Rajput, S. Mandal, I. Cumpstey, A. Larumbe, B. A. Salameh, B. Kahl-Knutsson, H. van Hattum, M. van Scherpenzeel, R. J. Pieters, T. Sethi, H. Schambye, S. Oredsson, H. Leffler, H. Blanchard and U. J. Nilsson, Galectin-3-Binding Glycomimetics that Strongly Reduce Bleomycin-Induced Lung Fibrosis and Modulate Intracellular Glycan Recognition, *Chembiochem*, 2016, **17**, 1759–1770.

47 A. Sethi, S. Sanam, S. Munagalasetty, S. Jayanthi and M. Alvala, Understanding the role of galectin inhibitors as potential candidates for SARS-CoV-2 spike protein: in silico studies, *RSC Adv.*, 2020, **10**, 29873–29884.

48 N. Hirani, A. C. MacKinnon, L. Nicol, P. Ford, H. Schambye, A. Pedersen, U. J. Nilsson, H. Leffler, T. Sethi, S. Tantawi, L. Gravelle, R. J. Slack, R. Mills, U. Karmakar, D. Humphries, F. Zetterberg, L. Keeling, L. Paul, P. L. Molyneaux, F. Li, W. Funston, I. A. Forrest, A. J. Simpson, M. A. Gibbons and T. M. Maher, Target inhibition of galectin-3 by inhaled TD139 in patients with idiopathic pulmonary fibrosis, *Eur. Respir. J.*, 2021, **57**, 1–13.

49 E. E. Gaughan, T. M. Quinn, A. Mills, A. M. Bruce, J. Antonelli, A. C. MacKinnon, V. Aslanis, F. Li, R. O'Connor, C. Boz, R. Mills, P. Emanuel, M. Burgess, G. Rinaldi, A. Valanciute, B. Mills, E. Scholefield, G. Hardisty, E. G. Findlay, R. A. Parker, J. Norrie, J. W. Dear, A. R. Akram, O. Koch, K. Templeton, D. H. Dockrell, T. S. Walsh, S. Partridge, D. Humphries, J. Wang-Jairaj, R. J. Slack, H. Schambye, D. Phung, L. Gravelle, B. Lindmark, M. Shankar-Hari, N. Hirani, T. Sethi and K. Dhaliwal, An Inhaled Galectin-3 Inhibitor in COVID-19 Pneumonitis A Phase Ib/IIa Randomized Controlled Clinical Trial (DEFINE), *Am. J. Respir. Crit. Care Med.*, 2023, **207**, 138–149.