Atom-level Machine Learning of Protein-glycan Interactions and Cross-chiral Recognition in Glycobiology

Eric J. Carpenter,[1] Chuanhao Peng,[1] Sheng-Kai Wang,[2] Russell Greiner,[3,4] Ratmir Derda[1]*

1. Department of Chemistry, University of Alberta, Edmonton, Alberta, T6G 2G2, Canada
2. Department of Chemistry, National Tsing Hua University, Hsinchu 30013, Taiwan and Frontier Research Center on Fundamental and Applied Sciences of Matters, National Tsing Hua University, Hsinchu 30013, Taiwan
3. Department of Computing Science, University of Alberta, Edmonton, Alberta, T6G 2E8, Canada
4. Alberta Machine Intelligence Institute (AMII), 1101-10065 Jasper Avenue, Edmonton, Alberta, T5J 3B1, Canada

*Corresponding author

Abstract

Cross-chiral recognition in glycobiology is the interactions between biologically conventional proteins and the enantiomers of biological glycans (e.g., L-proteins binding with L-hexoses) from organisms across all kingdoms of life. By symmetry, it also describes the interactions of chirally mirrored proteins with normal D-glycans. Knowledge of cross-chiral recognition is critical to understanding the potential interactions of existing life forms with artificial mirror-life forms, but currently known rules of protein-glycan interaction are insufficient. To build a methodology for learning such interactions, we constructed machine learning models that predict binding strength between proteins and glycans represented as graphs of atoms, rather than monosaccharides. Atomic *q*-gram and Morgan fingerprint (MF) based representation of glycans made it possible to train ML models that predict lectin binding properties of glycans, glycomimetic compounds, and enantiomers of all natural glycans. Critical to this training was merging disparate data—some with relative fluorescence units (RFU) from glycan microarrays and others with K_d values from ITC—using a universal "fraction bound" parameter *f* at a specific lectin concentration. A fully-connected neural network architecture, MCNet takes a MF and concentration (C) as inputs and returns *f* for 147 lectins. Performance of MCNet is comparable to the GlyNet models, and by proxy to other state-of-the art models that predict strength of protein-glycan interactions. MCNet effectively predicts binding of glycomimetic compounds to Galectins 1, 3, and 7. Breaking from a monosaccharide-based description makes it possible for MCNet to predict cross-chiral recognition. We employed a Liquid Glycan Array to validate some predictions, such as the lack of interactions of L-mannose with D-mannose binding lectins, purified ConA, and DC-SIGN displayed on cells, and weak binding of L-Man to galactose-binding lectins. MCNet's atom-level input makes it possible to agglomerate protein-glycan data from diverse glycans across all kingdoms of life and non-glycan structures (e.g. glycomimetic compounds). The universal fraction bound parameter makes it possible to unify disparate quantitative observations (K_d/IC_{50}, RFU, chromatographic retention times, etc.). We believe that such an approach will facilitate a merger of knowledge from diverse glycobiology datasets and predict protein interactions with uncommon/unnatural glycans not attainable from current ML models.

Introduction

Glycans are a major building block of life and major constituent of the glycocalyx that coats cells in all kingdoms of life.[1] Recognition of glycans by glycan-binding proteins (GBPs or lectins) is critical for cell-to-cell communication, immune responses, self vs. non-self recognition,[2] pathological responses like tumorigenesis,[2] and normal physiological processes like fertilization.[3,4] Unlike DNA, RNA or proteins, glycans form linear and branched oligomers with diverse stereochemistry of the monomers and linkages giving rise to a vast "feasible complexity" of glycans.[5] However, the estimate of "biologically relevant complexity"—the number of distinct glycan structures synthesized by a given organism—is smaller by many orders of magnitude. For example, the human glycome is estimated to be limited to ~7,000 structures created by known glycosynthetic enzymes.[6] Building on this dichotomy, inquiries in glycobiology are skewed to glycans of "biologically relevant complexity". Such glycans dominate glycan microarrays, like those produced by Consortium for Functional Glycomics (CFG).[7] Data from CFG and related arrays is used for training of machine learning (ML) that predict protein-glycan interaction.[8] The resulting ML models, again, are focused on predicting interactions between the biologically relevant glycan motifs and well-understood lectins in their training data. Some researchers may questions the need to study "biologically irrelevant" protein-glycan recognition, while others (e.g. Thomas Kuhn[9]) teach the need to break away from a dogmatic paradigm. For example, the elegant report of Canner and Gray used ML models to ask the fundamental question, what is a glycan binding protein,[10,11] and uncovered the intriguing possibility that organisms contain GBPs that have not been yet acknowledged as such. An interesting example are C-type lectins, it is not known how many of the C-type lectins actually bind to traditional or non-traditional glycans[12,13]. In this manuscript, we train ML models that predict quantitative glycan-GBP interactions; however, our primary goal is to break away from the "working complexity" and instead examine binding in i) the feasible complexity of glycans and ii) glycomimetic structures.

An emerging open-ended question regarding the feasible glycan complexity type pertains to enantiomers of known glycans. Such glycans would be produced by mirror-image life, for which the plausibility of laboratory construction has been discussed for the last few decades[14]. The feasibility of creating mirror-life was boosted by the recent synthesis of critical D-enzymes such as D-DNA-ligase, D-RNA polymerase and enantiomeric ribozymes,[15-17] as well as enzymes for ligation of D-proteins.[18] The pioneering work of Kent and co-workers experimentally proved that enzymes made of D-amino acids do not recognize peptide and peptidomimetic substrates of the natural enzyme, but they effectively process the enantiomers of natural substrates[19]. Mirror-image phage display[20] showed that L-peptides evolved to recognize mirror-image D-proteins do not bind to L-proteins, however, D-peptides binds to L-protein. The resistance of mirror-image proteins[21] and enantiomeric aptamers[15-17] to degradation by natural peptidases and nucleases *in vivo* is well established. These combined observations lead to a general notion of an absence of cross-chiral recognition between peptides, nucleotides and metabolites. This lack of cross-chiral recognition is at the heart of 2024 report in Science cautioning that if a mirror image life forms were constructed, it would be impossible for canonical life forms to seek and destroy them[22]. An elegant rebuttal by David Perrin[23], in the same journal, pointed out a critical flaw in such arguments reminding the reader that antigens, designed from D-peptides elicit high and sustained levels of neutralizing antibodies in animals[24] and D-peptides are also efficient immunostimulatory adjuvants[25]. Perrin cautions to "not scare the reader into overestimating the purported dangers, or thinking that the danger is imminent. The risks of mirror-life organisms are likely to be far less virulent than those posed by drug-resistant pathogens that can evolve or jump zoonotically here on earth."[23] A critical element overlooked by both reports is the overwhelming abundance of glycans and critical role of glycan-GBP interactions in self vs. non-self determination. Mirror-image glycans will be one of the first entities encountered by our immune systems;

however, our understanding of cross-chiral recognition between proteins and glycans is very limited. For example, D-mannose is abundant in the glycocalyx of most organisms; the interaction of its 1,2-bis-equatorial diol with Ca^{2+} in C-type lectins is at the heart of the immune response elicited by these lectins[12,13]. However, the L-mannose enantiomer is rare in biology and its protein binding profile is rarely studied[26]. The core recognition motif of D- and L-mannose is mirror-symmetric, 1,2-bis-equatorial diol. Reflection of this diol produces the same 1,2-bis-equatorial diol (**Figure 1A-B**). Antibodies to L-rhamnose (6-deoxy-L-mannose) are abundant in human serum but recognition of L-mannose by L-rhamnose-binding proteins is poorly understood. The spatial arrangement of three adjacent hydroxyls in L-mannose mimics that in D-galactose (**Figure 1C**). Indeed, there are scarce reports of L-mannose serving as a ligand for galactose-binding proteins[26]. In turn, L-galactose, the enantiomer of common galactose, and a close analog of natural L-fucose was reported as a viable substrate of α1,2-L-fucosyltransferase[27]. Further, the discovery of L-glucosidases points to possible use of L-glucose in glyco-conjugates, and the existence of lectins for their recognition[28]. This critical difference in symmetry inversion between glycans and other biological oligomers highlights that cross-chiral glycan-GBP interaction is more plastic than cross-chiral recognition between poly-peptides, oligonucleotides and metabolites. We propose that ML models that effectively embed atomic composition and chirality of such atoms can be serve as tools for the first evaluation of cross-chiral recognition between glycans and GBPs. Such ML models can also unify our understanding of traditional glycans and the "rare glycans" observed in the microbial world.

Outside of the traditional domain of mammalian N- and O-linked glycans, data for glycan-GBP recognition is not systematic. Even in a human glycome, recognition of heparin sulfate proteoglycans and mucins is only starting to emerge[29,30]. Rules for recognition by GBPs for plant glycans[31] or the eubacterial glycans are built progressively using complex organic synthesis of glycan arrays[32-35] and specialized lectin arrays[36,37]. N-linked glycans in Archaea contain exotic building blocks with different architectures and subunits[38]. Intriguingly, many orthologous functions, including speciation in sexually reproducing Archaea were proposed to be driven by interactions of these exotic glycans with as yet uncharacterized GBPs[38]. One barrier towards integrating all such information is the ever-increasing, but still incomplete, alphabet of all glycan building blocks across all kingdoms of life. A mirror inversion of the building blocks exacerbates the problem by effectively doubling the complexity. An ML model, like SweetTalk[39] and GlycanML[40] which represents glycans as collections of carbohydrate monomers can predict whether bacterial glycans are likely to be immunogenic in humans as long as such glycans contain only "natural" building blocks. However, these ML models cannot interpret L-Mannose and other "unnatural" building blocks. As a step towards bypassing all such problems, we examined the possibility of foregoing the "monosaccharide alphabet" and predicting GBP-binding properties of glycans from their atom-level structures.

Atom-level structural insights such as crystallography and molecular dynamics are workhorses in understanding glycan-protein interactions[41]. These investigations make is clear that a specific conformation of glycan and specific conformation of protein define glycan:GBP recognition[42]. This realization led to development of methods that estimate binding properties by computing conformations of proteins and glycans to find an optimal match (e.g., by docking)[41,43]. Calculation of thermodynamic values by exhaustive computation of the ensemble of all conformers is contrasted by Anfinsen's thermodynamic hypothesis "the folded structure is determined [solely] by the amino acid sequence".[44] It is reinforced by the recent 2024 Nobel prize for AlphaFold[45] that predicts a folded structure and subsequently a binding pose in protein-ligand complexes[46] from primary sequences of the protein and ligand. From these observations, we infer the existence of a lightweight ML model that predicts the interaction between two molecules using only atomic connectivity of these two partners. Here we examine the plausibility of lightweight ML models that predict binding of protein to glycan from the mere atomic connectivity of glycan. A caveat in the Anfinsent

hypothesis is the role of the environment: Anfinsens' RNAse adopted two distinct conformations in two different buffers (folding and denaturing buffer)[44]. Similarly, all molecular interactions depend on the environment. A common shortcut[47-50] is to focus on binding events measured in nearly uniform binding environments (similar temperature, buffer compositions, pH). Models from such measurements are not universal, but they should achieve reasonable performance in predicting binding strengths in a standard environment.

Models predicting quantitative parameters (e.g., K_d or IC_{50}) for interactions of proteins with other proteins, antibodies or small molecules are common due to availability of curated datasets of K_d and IC_{50} values. Interestingly, in glycoinformatics, many ML models focused on classification—binder vs. non-binder[39,50,51]—and similar binary decisions (e.g., has motif / does not have motif).[49,52] We find this binary view of binding unsatisfactory because the binding→0 when c→0. At low enough concentration of a binding partner, the answer is always "no". Our previous report avoided this problem by training a regression ML model that predicts continuous variables, the Relative Fluorescent Units (RFU) measured in glycan microarray experiments. Arthur and coworkers further generalized this problem by extrapolating K_d surrogates from the microarray data[53]. The CarboGrove tool developed by Brian Haab and coworkers elegantly merged data from disparate experimental sources and further associated the binding at a specific concentration with the presence or absence of specific motifs[52,54]. We envisioned that the problem of merging glycobiology datasets for ML[8] can benefit from merger with broader chemical datasets like BindingDB[55] or ChEMBL[56]. However, this merger is not possible if glycans are described as oligomers of monosaccharides[8]. Rather than cataloguing an ever-growing catalogue of monomers, we focus on more productive atomic compositions. We paraphrase the dream of "*I want to know if this oligosaccharide motif binds to the GBP*" to the more fine-detail request, "*Given a GBP and a glycan of a definitive atomic composition, I want to know the fraction of glycan **f** bound to this GBP*." Our manuscript shows the plausibility of training a model that takes a glycan, GBP, and the circumstances as input and returns ***f***. A model that forgoes monosaccharide composition should also be able to handle any glycans across all kingdoms of life. It can account for well-documented effects of spacers, linkers, underlying scaffolds, (**Figure S1** and **S2**) in diverse arrays. It can bridge glycan structure with composition of protein sequences on which such glycans are displayed and it can give the first important glimpse into cross-chiral recognition between glycans and GBPs.

Results

Embedding of all-atomic representation of glycans

As in previous reports[7,48,57,58], the core dataset for the training of ML models was binding of GBP to glycans on microarrays produced by the consortium of Functional Glycomics (CFG, Mammalian Printed Array version 5). In short, this data describes performance of 611 glyco-conjugates printed in discrete spots on the glass plate exposed to 1257 fluorescently-labelled GBP compositions and imaged by fluorescent scanner. The amount of protein bound to each glycan is reported as Relative Fluorescent Units (RFU, **Figure 1D**). We previously trained GlyNet neural network to predicted 1257 RFU values for CFG glycan structures represented as monosaccharides q-grams (MQG)[57]. To test whether all-atomic representation of glycans can be used for prediction of qualitative interactions with GBP, we converted all CFG glycan structures to SMILES strings and tested two types of fixed-length representations of SMILE strings: Morgan fingerprints (MF) and atom q-grams (AQG) using RDKit (**Figure 1E-G**). We included atom-q-gram inspired by genomic k-mer statistical methods[59,60] and to be able to draw an explainable parallel between our prior monosaccharide-tree embedding[57] and the new atom-level embedding. Replacing the MQG inputs of the GlyNet architecture with MF or AQG inputs gave rise to several neural network

architectures with 2-3 hidden layers and 50-100 neurons per layer (**Figure 1H**); each architecture was optimized using Adam[61] and tested in 10-fold cross validation for their ability to predict 1257 RFU values for lectins given a MF or AQG representation of CGF glycans.

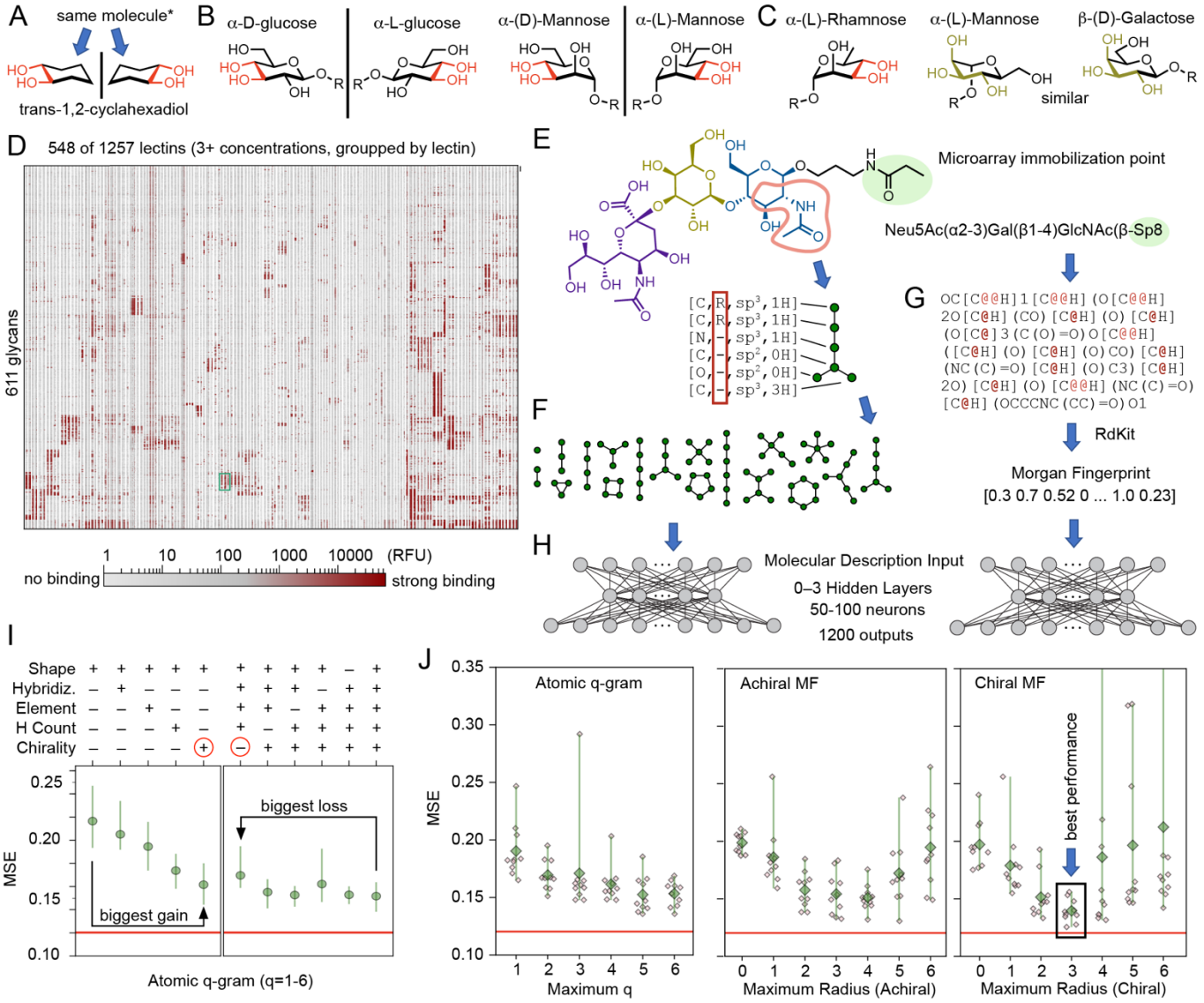

Fig. 1. Machine learning (ML) of glycan:GBP interactions from all-atomic description. (A-C) Examples of preservation of stereochemical motifs between enantiomers, unnatural L-glycans and natural glycans. (D) Glycan:GBP binding data from CFG used for ML. (E-G) Each glycan is converted to counts of (F) atom *q*-grams (AQG) or (G) Morgan fingerprint (MF). Features of nodes in AQG, (F) include element, chiral state, hybridization, and attached hydrogen count, and overall shape. (H) Either MQG or MF counts are used as input into a fully-connected neural network which is trained to produce binding strength estimates at its outputs. (I) Mean square error (MSE) measure the performance of the models that employ different features in AQG nodes. Models trained using AQG with only a minimal set of features rank importance of chirality > H-count > element > hybridization. Similarly, systematic removal of features from AQG show that removal of chirality diminishes performance, whereas removal of hybridization, element identity or H-count is inconsequential. (J) Comparison of AQG of size 1=6 and MF or r=1-6 show that AQG with 6 atoms achieve best performance in AQG, chiral MF3-representation as best across all representations. There is a marked improvement in model trained using chiral vs. achiral MF3. Each value represents MSE measured in separate training fold; green scatter plot shows mean and standard deviation of MSE; the red line denotes the performance of GlyNet with monosaccharide embedding.

Figure 1I-J describes optimal performance for each neural network model after selecting an optimal number of hidden layers, nodes in hidden layers, and weight decay parameters. The distribution of glycans to 10 folds was identical for all models and identical to previously published GlyNet[57]; hence distribution of mean square errors (MSE) across the folds made it possible to compare all models with MSE of the GlyNet model (red line in **Figure 1J**), and by proxy many other state-of-the art models[47-50]. A simple, explainable representation of glycans using AQG of size 6 yielded MSE=0.15, which was only modestly worse than state-of-the-art GlyNet with glycan monomer embedding (MSE=0.12). Systematic subtraction or addition of graph topology, element, hybridization, chirality, number of attached hydrogens (H-count), and chirality of the node (R, S or none) rank chirality as the most important feature and H-count as the most important (**Figure 1I**). Deletion of identity atoms or their hybridization (proxy for bond orders) was inconsequential to predictions. Describing glycans by featureless graphs with 6 nodes yielded MSE=0.22 but adding only chirality information yielded MSE=0.16 (**Figure 1I**). Both observations show an expected critical role of chirality in predicting binding of glycans to lectins. An intriguing observation that was that reasonable prediction of glycan:protein interactions can be achieved by representing glycans as graphs that know only chirality of the node. This observation shows how different glycan:GBP recognition is from a interaction of proteins with classical small molecules that contain few to no chiral centers.

MF of radius 3 with chirality ($MF3_c$) yielded the most optimal performance, which was comparable to performance of GlyNet (**Figure 1J**). This observation clearly demonstrated it is possible to predict glycan-binding performance by replacing a domain-specific monomer composition with universal all-atom-representation. A simple learning architecture like fully connected neural networks can be trained even in relatively small dataset of 600 glycans despite relatively large number of atoms in some glycans (500 non-hydrogen atoms). Elimination of chirality in MF led to notable decrease in performance (**Figure 1J**). Decreasing the maximum radius below 3 was clearly detrimental whereas increasing the radius above three (from 1385-dimensional to 3142-dimensional vectors at $r \leq 4$) decreased median performance and increased fluctuations between the folds. Both observations may be the result of small size of the dataset. A mere 600 glycans is likely not sufficient to learn from high-dimensional MF4-vectors, but it possible that datasets with >>600 glycans and MF4 or higher may yield even better performance than current model.

Expansion of training beyond microarray datasets

All atom representation in principle makes it possible to combine learning for natural glycans and "unnatural" glycomimetic (GM) compounds. To test this hypothesis, we selected 274 GM structures with reported binding strength (association constants, K_a) to Galectin-1, -3, or -7 manually curated from 42 publications (see **Supplementary Table 1**). We embedded GM structures using $MF3_c$ and attempted to train a multi-output neural network that predicted RFU values in some outputs and K_a values for other outputs. The training failed likely due to the disparate structure of the datasets: RFU values were available for CFG glycans, but not GMs while Ka data was available only for the GMs but not the CFG glycans (see supplementary discussion and **Figures S3-S5**). We were able to resolve this problem inspired by the work of Haab and coworkers (CarboGrove[47]), who unified disparate datasets by examining dose response curves. Previous reports employed RFU values to extrapolate K_a values. However, we elected to convert both RFU and K_a to a series of values describing the fraction, *f*, of the glycans bound to GBP at the specific concentrations, C (**Figure 2A-D** and **S6-S8**). K_a is trivially converted to *f*(C) via the Henderson-Hasselbalch equation (**Figure S7**). For the CFG glycans *f* was interpolated between the available concentrations using an array specific minimum and a global maximum RFU (**Figure 2A-D**, detailed in **Figure S6, Supplementary Data 2**). The *f* value is uniquely suited for learning because it is always bounded to the (0;1) region. It allows the qualitative dichotomy where f~0 and f~1 are equivalent to "no binding" and "yes

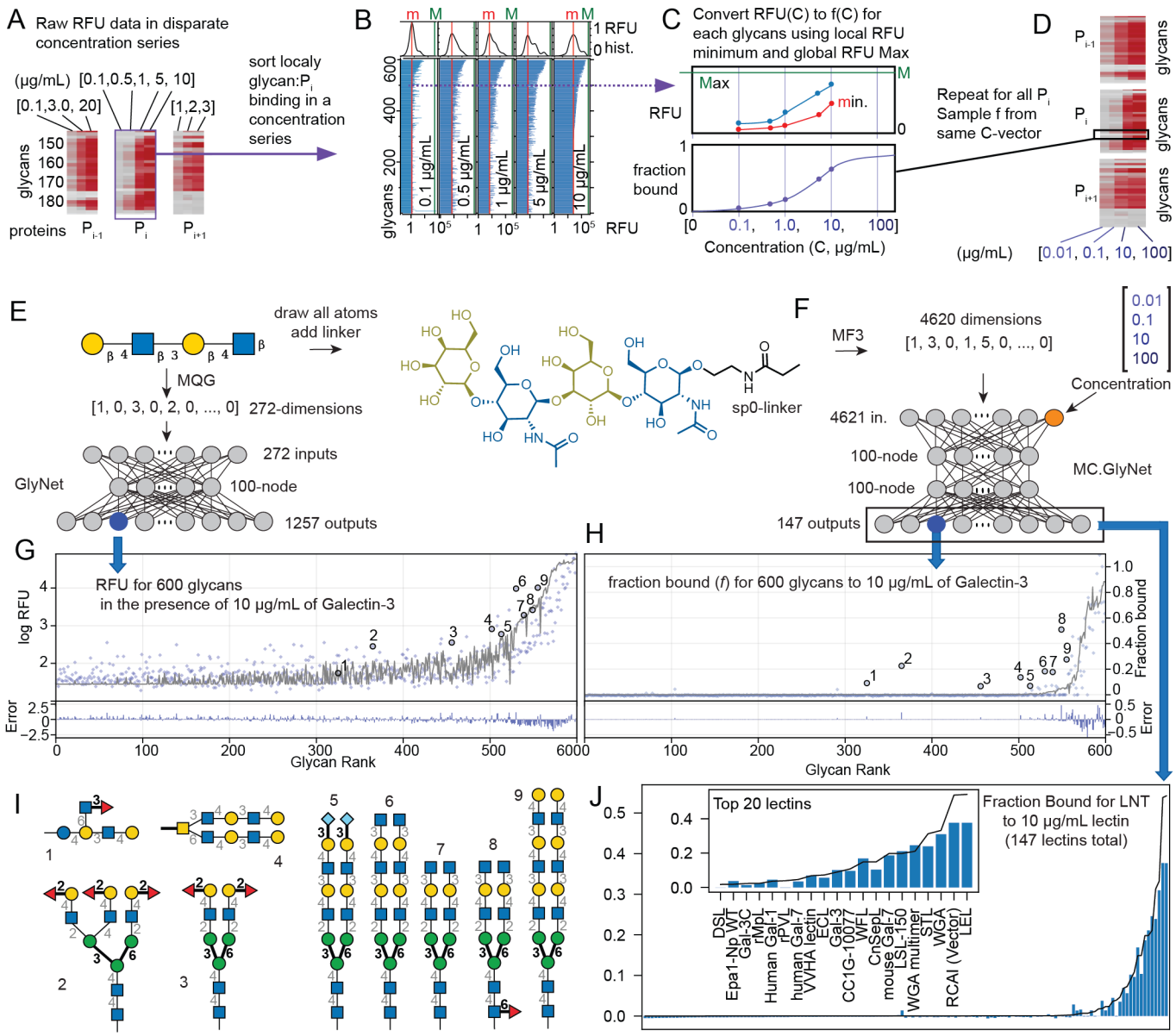

Fig. 2. Workflow of data in training of MCNet. A) 147 lectins tested at multiple concentrations in the CFG v5 data were identified. B) Microarray RFU distributions were processed to extract a background and maximum which were then C) used to rescale into a fraction bound and D) resampled at multiple concentrations. We compared E) our prior monosachharide-tree based RFU model to F) an atom-graph (includes spacer structures) model using Morgan fingerprints and protein concentrations. The results (G & H respectively) show similar trends. In particular, I) a group of glycans predicted by the MF Net model to bind more strongly to Gal-3 than the CFG data, overlaps significantly (6 of 10) with over-predictions in the prior GlyNet model, and are of structures reported[62,63] to have binding. J) Model predictions for binding of LNT to the 147 lectins.

binding": it alos brings an important reminder that such assessment can be made only when concentration is specified. Finally, *f* has an intuitive interpretation (e.g., f=0.5 at a concentration nearing K_a , IC_{50} or EC_{50}) although such generalization needs to be done with caution as highlighted in numerous publications[64-68].

To test *f* as a predictive output, we trained a fully connected neural networks which has the form, ***f*** = *Net*(**M**,C) where inputs **M** and C are MF3 description vector and concentration, *Net* is the neural network

function, and ***f*** the resultant fractions bound. This architecture, referred to as MCNet, has 3931 inputs, one for protein concentration (C) and the 3930-dimensional **M**-vector $MF3_c$ and 147-dimensional output ***f*** vector for 147 lectins from CFG dataset (**Figure 2F**). Models were trained on CFG glycans using the same ten-fold cross-validation structure as above with one clarification: all concentrations were in the same hold-out fold. Due to novelty of the *f*-value, MSE for MCNet cannot be compared directly to MSE of GlyNet or MSE of any state-of-the-art models; instead, we examined individual prediction of glycan binding profiles for lectins with well-known binding properties, with a focus on Galectins (**Figure 2G** vs. **2H**). Supplementary information contains analogous alignment of predicted f-values for other lectins juxtaposed to predictions made by monosaccharide-based models (e.g., GlyNet).

From 599 glycans, MCNet predicted ~10 with modest upward deviation from the predicted trend (i.e., "false positive binders" **Figure 2H**). Examination of the individual structures, we noted that 7 of the 10 glycans contain definitive galectin-binding motifs (LacNAc repeats, **Figure 2I**) and 6 of the 10 of these had the same over-predicted property in the prior GlyNet model, many more than random chance would predict. Atom-level models are learning that similar features of the molecules as GlyNet (**Figure 2E-H**), even though one is looking at smaller groups of atoms and the other at the groups of monosaccharides. Glycans 2 and 3 contained terminal LacNAc with a α-2 fucosylation on terminal Gal (group-H, **Figure 2I**); however α–2 fucosylated lactose have been reported to bind more strongly than lactose by Klassen and coworkers[62] and the α–2 fucosylated LacNAc was confirmed as ligand for Galectin-1 by Surolia and coworkers (**Table S3**)[63]. A 2022 re-run of Galectin-3 binding to CFG arrays by Arthur, Stowel and co-workers (data not used in MCNet training) confirmed group-H glycans on CFG arrays are modest but detectable binders to Galectin-3[53]. The same glycans 2 and 3 in GlyNet model, were predicted to have an order of magnitude higher RFU than the "ground truth" (**Figure 2G**). We conclude that these glycans are experimental false negatives of glycan microarray measurement rather than "false positives" of the ML model. Atom-level MCNet that predicts fraction bound, not only achieved desired performance but also highlighted experimental false negative observations in CFG dataset. This is not a criticism of CFG dataset but merely a reminder of well-established notion that the experimental uncertainty defines a natural upper limit to the predictive performance of machine learning models[68].

We tested the ability of MCNet architecture to make out-of-domain predictions. MCNet1 trained only on CFG data was unable to predict any binding for any glycomimetic molecules (**Figure 3C**). Training MCNet2 only on GM dataset made it possible to predict binding of GM compounds in held-out datasets, but MCNet2 was completely wrong in predicting binding of CFG glycans to galectins (**Figure 3D**). In contrast, the **MCNet3** model trained on all GM and all CFG data effectively predicted GM binding to Gal-1, 3, and 7 for held out GM datasets and all lectin binding for all held out CFG datasets (**Figure 3E,** and supporting file **Per Glycan Prediction Plots.pdf**). Glycan and glycomimetic recognitions are non-overlapping and knowledge of one does not guarantee to knowledge of the other dataset. And conversely, blending of datasets is positively synergistic: accuracy of the prediction of binding of GM to galectins and even binding prediction for CFG glycans to 144 other lectins increased when the MCNet3 model was trained using both CFG and GM molecules. Number of compounds increased from 600 in CFG to 800 in CFG+GM dataset, hence mere 30% increase in the size of the dataset cannot account for an exponential increase in accuracy.

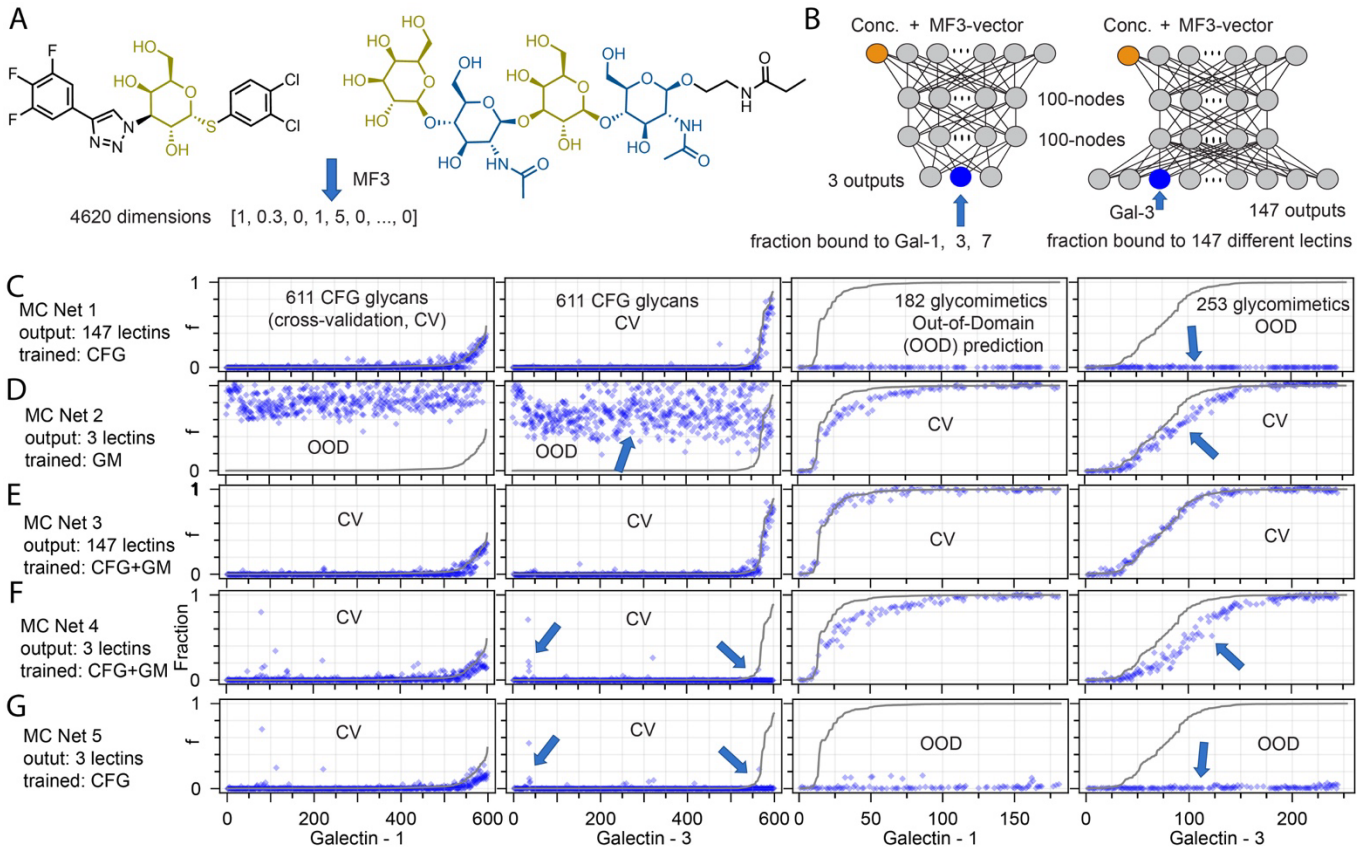

Fig. 3. Predicted binding for the CFG glycans and the glycomimetics (GM) across different model classes. All plots are at ~10 µg/mL. (A) Both glycans and GM compounds can be embedded as MF3 vector. (B) MCNet architectures with either 3 or 147 outputs were used in this test. (C) MCNet1 trained using CFG glycan data (analogous to **Fig. 2F, H**) predict binding of CFG glycans to Galectins in held-out datasets but failed to predict binding of GM to Galectins. (D) MCNet2 trained on GM data exhibits reasonable prediction of GM:Galectin binding but fails in predicting binding properties of CFG glycans. (E) MCNet3 trained using both datasets exhibits desired predictions and exhibits improvement in predictions of GM:Galectin interactions when compared to MCNet2 trained only on GM data or (F) MCNet4 trained only on galectin binding data of GM and CFG compounds. (G) MCNet5 trained only on Galectin:CFG data makes inferior predictions for binding of CFG glycans to Galectin-3 and it can learn only some trends in Galectin-1:CFG dataset.

Control MCNet4, trained on all GM data and only on a subset of CFG data describing binding to Galectin-1, -3, and -7 was also able to predict binding for held out GM and CFG datasets, albeit its accuracy dropped when compared to MCNet3 (**Figure 3G**). This observation illustrates how knowledge from seemingly unrelated glycan:GBP interactions, used in training of MCNet3, enhanced the accuracy of prediction of glycan:Galectin interactions. Another control MCNet5 trained only on a subset of CFG data describing binding to Galectins 1, 3, and 7 to CFG glycans, failed to predict GM binding (**Figure 3G**). We noticed that prediction of Galectin-3 binding deteriorated in MCNet4 and 5 but Galectin-1 predictions remained similar to MCNet1 and 3 (**Figure 3F-G**). It is tempting to blame the ML architecture for these failures, but the reasons are likely more fundamental. These failure of MCNet to extrapolate between CFG and GM datasets indicate that the rules of protein recognition by chiral hydrophilic sp^3-centres in CFG dataset and achiral hydrophobic sp^2-centres in GM dataset cannot be extrapolated from one another.

We anticipated that the learning the dose response should not be different from learning parameters of the curves that describes them. We made a curious observation, however MCNet did not always learn the monotonously increasing sigmoidal nature of the binding curves (**Fig. S9**). Models were trained using

50 f values at 50 input concentrations between 10^{-5} and 100 μM for each molecule (**Fig. S9**). Still, for many GM compounds, models failed to predict an obvious boundary condition: f must become zero at infinitely low concentrations. Instead for many GM compounds f plateaued at arbitrary values. Predictions for some molecules were satisfyingly monotonous, whereas for other we observed unexpected bumps and dips reminiscent of overfitting the monotonous "ground truth" with complex-shaped high-degree polynomial. We anticipate these problems to be resolved in future algorithms that learn f from discrete binding values.

Out-of-domain extrapolation of binding properties to other datasets

MCNet architectures are capable of extrapolating lectin binding not only to glycan structures from GlyTouCan but also to any molecules (e.g., BindingDB, see **Fig. S10**). High-level observations mirror the observations already made in **Figure 3**. Specifically, MCNet1, trained only on CFG glycans, predicts zero binding performance for all molecules that are not *bona fide* glycans. Despite these deficiencies, MCNet1 effectively identifies glycans that are reasonable predictions as ligands for specific lectins. As there are currently no systematic ground truths for these predictions only some of these predictions can be evaluated. An interesting case is the bacterial glycans, which are not part of the training set. MCNet predicts that Galectin-1 and -3 interact with Gal-GalNAc oligomers, structures reminiscent of the E. coli O127 outer polysaccharide (OPS) and the Klebsiella pneumoniae O-antigens and other bacterial polysaccharides reported by Drickamer and co-workers[36], Kiessling, Imperiali and co-workers[37] and Cummings, Arthur, Stowell and co-workers[69,70]. Such data provides a baseline benchmark for all future predictions or experimental validations (see Supplementary file Models.zip).

Mirror image glycan extrapolation

Inversion of all stereocenters in every molecule yields enantiomeric glycans and GM compounds. We tested the ability of MCNet to predict recognition between lectins and enantiomers of the molecules from CFG and GM datasets. Such observations have no "ground truth" only an "expert opinion". For example, with 100% certainty, we can assert that every GM compound should bind to Galectin 1, 3 and 7 significantly better than the enantiomer of the same compound. The absence of cross-chiral recognition arises from the well-understood structural role of the chiral centers around the galactose ring. Interestingly, not all MCNet models predicted a drop in cross-chiral recognition. MCNet2 trained only on GM compounds placed the activity of enantiomers at random, either above or below the performance of the parental compounds (**Figure 4A**). This observation is recurrent in the ML literature and other reports have observed that chirality is one of the most difficult concepts for ML models to understand[71]. In contrast, an MCNet3 model trained on GM and CFG data—the latter contains numerous examples of epimers and diastereomers—was able to predict that nearly of enantiomers of GM compounds are less active than parental compounds (**Figure 4B**). It was important to include data for binding to all 147 lectins: MCNet4 trained using data for 300 GM and 600 CFG glycans binding only to Galectins 1,3,7 has diminished understanding of cross-chiral recognition (**Figure 4C**) when compared to MCNet3 (**Figure 4B**). MCNet appears to agglomerate the knowledge from various proteins effectively. Indirect knowledge of how epimerization of individual atoms influences glycan:protein recognition helps MCNet3 to produce the right answers for out-of-domain cross-chiral recognition.

Epimerization of a glycan may yield another common glycan (e.g. changing a glucose into mannose), but it may also give rise to an atypical glycoform (e.g. β-Glc to α-L-Ido). We examined the ability of MCNet to predict binding-strength for epimers of existing glycans (i.e. molecules that differ from the original by the inversion of one chiral centre, see Supplementary Epimer Heatmaps.pdf). MCNet 1 and 3 generally agree on their predictions with patterns of strongly bound lectins common to both models, but some predictions diverge where binding of the parent glycan in weak. Unsurprisingly, for many glycans

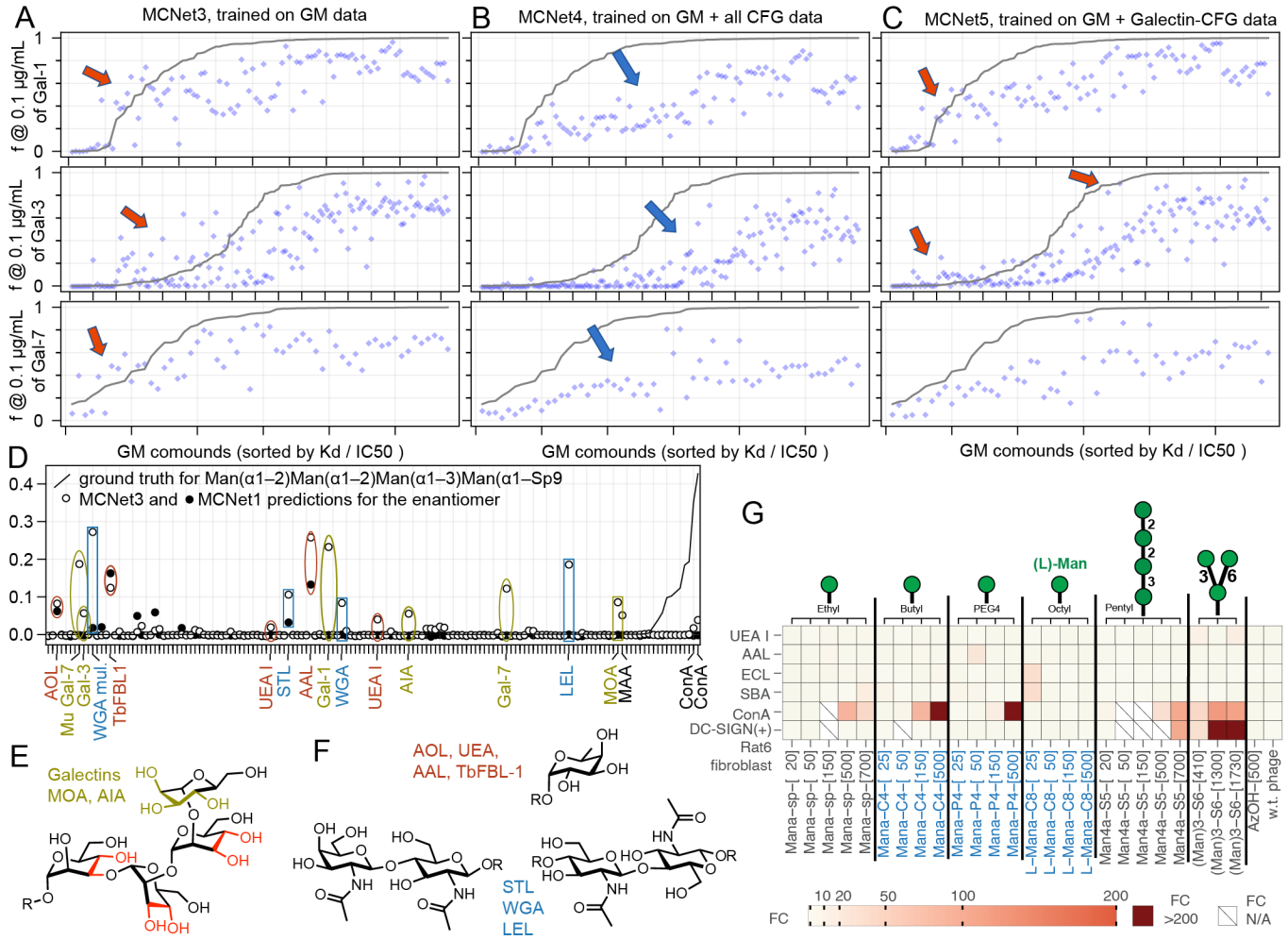

Figure 4. Prediction of Cross-chiral recognition by MCNet. (**A-C**) Solid line describes the ground truth: binding of glycomimetic (GM) compounds to Galectins 1,3 and 7. Blue dots are predicted binding strength of the GM-enantiomer. Expert knowledge of GM:Galectin interactions dictates that all predictions should be below the solid line (i.e., activity of enantiomer must be lower than the activity of the parental compound). Only MCNet3 (**B**) produced answers aligned with "expert knowledge" of cross-chiral recognition. (D) Enantiomers of glycans with terminal Man are predicted by MCNet3 to have decreased binding to ConA and other lectins. (E) Experimental confirmation of the cross-chiral recognition. L-Mannose, displayed at 25 to 500 copies on M13 virion fails to bind to ConA; however, glycans with D-Mannose exhibit an expected performance in the same liquid array.

both MCNet models identified that epimerizations in small glycans that destroy recognition by lectins of the original glycan, while often having no effect in larger glycans, see Supplementary Epimer Heatmaps.pdf. Further summary of observations is available in the supplementary materials.

We then tested the ability of MCNet to predict properties of enantiomers of the CFG glycans. A ground truth for these do not exist because it is unreasonable to propose to manufacture the CFG arrays with 600 enantiomers of existing glycans. However, Peter Kim's mirror-image paradigm[20] postulates that the binding of L-lectins to enantiomers of CFG glycans can be measured by probing D-lectins on existing CFG array. Unfortunately, our attempts to synthesize C-type lectins (DC-SIGN, DC-SIGNR, Langerin) from all D-amino acids are still ongoing and are not trivial. Nevertheless, we established the experimental ground truth for these observations by measuring binding of several lectins to various valences of L-

Mannose and D-Mannose in the same Liquid Glycan Array[72]. Unlike D-Mannose, L-mannose was completely inactive in binding to every tested calcium-dependent lectin (ConA, DC-SIGN on the surface of live cells). Both MCNet1 and MCNet3 were able to corroborate these observations. For Man(α1-6)Man(β-Sp10 and Man(α1-2)Man(α1-2)Man(α1-3)Man(α-Sp9 both MCNet1 and MCNet3 predict that the mirror forms of the glycans have a significant decrease in binding to ConA (**Figure 4D**). MCNet predicted that a tetra-mannose enantiomer is recognized by lectins that traditionally bind galactose (Galectin, MOA, AIA), fucose (AOL, TbFBL-1, UEA-I) and GlcNAc, GlcNAc oligomers and LacDiNAc structures (Stl, WGA, LEL). Recognition by galactose-binding lectins is not entirely surprising given the resemblance of a triol configuration in L-Man and D-Gal, and a preliminary LiGA experiment suggested weak binding between L-Man and prototypical D-galactose binding lectins. Recognition of fucose and GalNac/LacNAc-binding lectins is less clear; we tested one of them and observed no binding of L-Man monomer to UEA. Further experiments are needed to confirm whether a larger oligomer of L-Man possesses UEA binding. The neural network that maximizes learning from diverse glycans and glycomimetics and integrated observations from all lectins outperformed all control ML models that learn on partial datasets (**Figure 4E**). Both predictions highlight the power of complete datasets, such as CFG array datasets for training of ML models that are capable of interpreting outside-of-domain prediction of cross-chiral recognition. The sheer number of chiral centres, diastereomers and near-stereoisomers provides rich training ground for training models focused on predictions of the effect of chirality rather than the effect of addition and substation of atoms and functional groups.

Discussion

In this manuscript, we put contemporary understanding of protein:glycan interaction to a test by aiming to solve "biologically irrelevant" questions. We investigate whether it is possible to predict the binding ability of naturally non-existant or rare glycans, such as chirally mirrored carbohydrates. Such extrapolation is not without its benefits. Glycans across all kingdoms of life appear to be tightly linked to speciation. Species recognize each other and make decision such as "friend" and "danger" by integrating data from glycan:protein interactions; glycan based pathogen-associated molecular patterns (PAMPs) and damage-associated molecular patterns (DAMPs) are central to our immunity. The prediction of strength of glycan:GBP interaction offers a universally comparable property and give rise to taxonomy and speciation orthogonal to the traditional 16S RNA and DNA taxonomy[73]. Synthesis of the mirror-image life forms is not the question of "if" but "when"[14,74] These new species will not use "unknown" glycans but rather a very well-defined enantiomer of known glycans. It is important to understand how this new class of species coated with mirror image glycans will be interpreted using canonical protein receptors; will they be recognized by self/danger recognition receptors? Our manuscript does not resolve this question but builds the ML framework for addressing such question with higher and higher degree of confidence.

The atoms and their connections within a glycan appears to be necessary and sufficient for lightweight ML models to predict the recognition of that glycan by protein receptors. More generally, this atomic connectivity of any molecule is a minimal description of the molecule, and is sufficient to describe its properties. Identity and connectivity of atoms relates to Kolmogorov complexity of a molecule: it is sufficient to compute the properties of a molecule. For example: quantum mechanics (QM) can estimate electronic structure and Boltzmann ensemble of conformations from mere atomic connectivity. ML models can be trained using atomic coordinates; but such models produce properties of specific conformers. This ensemble needs to be computed and integrated over a Boltzmann ensemble to yield a property of such ensemble. We firmly believe that thermodynamic scalar such as strength of binding between glycan and

protein at a given concentration can be predicted without the knowledge of the Boltzmann ensemble of all conformations. We show that binding properties are possible to predict from mere atomic connectivity without any knowledge of any conformations. This observation forms a generalization of well-accepted Anfinsen hypothesis across the entire domain of molecular interactions. It can be phrased simply as "*atomic connectivity of the interacting partners are necessary and sufficient to determine the strength of the binding between these partners*".

The goal of this manuscript is not to achieve a best-in-class performance of an ML model but rather to raise an awareness of the problems for future ML models to focus on. Growing advances in conversion of SMILES to fixed-length representation should be tested alongside MF and AQG in the subsequent manuscripts. This exploration will find the most effective representation of a chemical structure that drives MSE or other loss function to an optimum[40]. We leave such benchmarking and competition for future reports. We find it productive to re-examine what are perhaps the most bare-bone, and interpretable representations of molecules, the atomic *q*-gram (AQG) counts. Predictions of glycan properties from molecules represented with *q*-grams or MF of one atom were surprising. These models are only have counts of atom identities as inputs, with no information about any connections between the atoms. In retrospect, the marginal drop in MSE from 0.12 to 0.20 is not surprising because the 611 CFG glycans are rather diverse. There are only a few isomeric glycans and zero enantiomeric glycans in CFG collection. The binding of large glycans can be correlated with "expanded elemental composition" of such glycan without any consideration for topology. Mere memorization of composition may be a fallacy of models learning from datasets that contain molecules with diverse sets of atoms. The true challenge for molecular interaction is prediction of binding for *bona fide* isomers and stereoisomers. MCNet3 can predict binding to a diverse set of diastereomeric glycans enabling first exploration of mirror-recognition glycobiology. We noticed, however that MCNet struggles to predict obvious expert knowledge such as "binding of Galectin-3 to LacNAc must be ablated when carbons 2, 3, 4 or 5 are epimerized in any ring," (see pages 50, 51 in Epimer Heatmaps.pdf). This is "obvious knowledge" for every glycobiologist. Yet it is not clear how to integrate such obvious expert knowledge into the ML training without exhaustive synthesis and measurement of all epimers of LacNAc.

We proposed that a "simple" total chemical synthesis of a lectin made of D-amino acids and testing of such a lectin on a CFG microarray can confirm or deny many predictions made in this manuscript. Docking[75-80] is another fast way to validate or reject some of the predicted interactions between enantiomeric glycans and common lectins. Such techniques could rapidly advance glycobiological exploration of cross-chiral recognition. The lack of L-mannose binding to all lectins that bind D-mannose (**Figure 4**) was anticipated by Goldstein et al.[81] The dichotomy between L- and D-mannose may rationalized based on the relation of L-mannose and L-rhamnose. L-rhamnose binding lectins do not recognize D-mannose[82-84] and D-mannose-binding proteins rarely bind to L-rhamnose. But counterexamples are known, some lectins bind mannose and rhamnose equally.[85] We also found a report of Tatibouët and co-workers who measured the binding of a multivalent BSA-L-Rha_{17} conjugate to ConA and DC-SIGN with IC_{50} of 3 and 0.6 μM while BSA-D-Man_{16} achieved similar activity (IC_{50} of 0.2 and 0.6 μM). Hence, even extrapolating between L-mannose, L-rhamnose and D-mannose may not be obvious to human experts and needs to be supported by experimental validation. Overall, we believe that focusing on cross-chiral recognition in glycobiology may uncover many unexpected observations, which eventually will reinforce the overall protein-glycan "rules of recognition". An agglomeration of such measurements would form an effective ML training dataset for learning about the role of stereochemistry in glycan recognition and in overall molecular recognition.

Integration of microarray data from multiple protein concentrations into a surrogate of binding strength (e.g. EC_{50})[53] is one approach, but not free of caveats. For example, line shape in dose responses are not always trivial to convert to definitive EC_{50} values and steep, cooperative responses should be delineated from non-cooperative sigmoidal curves (i.e., curves need Hill coefficients). We do not use the experimental errors in our training process, but in models that do, propagation of errors from responses to errors in fits adds another layer of complexity. Both problems are avoided by learning fraction bound instead of Kd. The atom-level glycan representation overcomes the deficiencies observed in monosaccharide-based representations of glycans. For example, the CFG dataset does not contain xylose or L-mannose, hence, a GlyNet-type model or LectinOracle[48] that embeds monosaccharides simply cannot predict binding of oligosaccharides that contain xylose or L-mannose. Embedding an atom-level description of glycans solves this problem and gives the first glimpse into the unseen mirror-universe of glycans: such bizarre mirror-image monomers have never been considered by any ML model due to lack of any training data. These observations can provide an important understanding of the emerging risks (or lack of thereof) from proposals for the synthesis of living mirror-image cells.[22]

The ability to represent a structure does not imply the ability to predict its binding accurately. The atomic q-gram approach has an obvious out-of-domain deficiency: If the CFG and GM dataset does not contain an element (e.g., Br), how can a model even know what a bromine atom is? Any molecule with Br atom appears to be a "cold start problem". An analogous rationale can be made for groups of atoms (functional groups) that are absent from the training dataset. GlyTouCan and BindingDB contains cyclopropane rings, which is completely absent from CFG and GM training dataset, **Table S2**. How is it feasible for model to extrapolate the properties of cyclopropane when the training set does not contain any strained rings? Interestingly, the chirality of individual atoms presents another "cold start problem" and in the absence of any examples of chiral differences, the ML model (e.g., MCNet2) is unable to make any predictions of chiral dependancies on binding.

Chemical intuition taught in an introductory chemistry course can explain the properties of Br from trends in periodic table (e.g., H, F, Cl, Br, I, etc). Even primitive mnemonic features like count of electrons teach student to predict trends in Periodic Table of Elements. Chemical intuition suggests that ML models could take into the account electronic structure of atoms and functional groups (orbital structure, polarizability of the valence shells, Milliken charge) instead of their linguistic description (Bromine, Chlorine, cyclopropane, etc.). Such models might be able to overcome some "cold start" problems. Indeed Jeff Kelly and co-workers profiled 357 quantum mechanical descriptors to predict those that contribute the most to protein-glycan interactions[86]. However, simple electronic structure theory cannot explain long-range effects caused by chirality. The effect of chirality on molecular recognition is easier to teach "by example". This observation highlights critical importance of datasets emanating from glycans, glycobiology and protein:glycan interactions for global efforts in design of "chirality aware" ML models of molecular recognition. Integration of datasets that describe protein:glycan interactions into any ML models introduces the most complete description of the effect of local changes in chirality on the overall molecular recognition; no other fields offer such exhaustive coverage of the properties of diastereomers. Our work helps such integration by proposing strategies for integration of disparate quantitative data via universal fraction bound and encouraging all-atomic molecular representation in the field forged on a monomer-centric alphabet.

Acknowledgements:

We thank Natalia Baranova and Jared Zhang for their efforts in searching for small-molecule galectin inhibitors and reported K_d values in the published literature. We thank Todd Lowary, Matt Macauley and Chris Cairo for their comments and feedback. This work was supported by Canadian Institutes of Health Research (CIHR) (no. 180445), Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (RGPIN-2016-402511) and GlycoNet (CR-29 and TP−22).

References

1. *Essentials of Glycobiology*, (Cold Spring Harbor (NY), 2022).
2. Amon, R., Reuven, E.M., Leviatan Ben-Arye, S. & Padler-Karavani, V. Glycans in immune recognition and response. *Carbohydr Res* **389**, 115-122 (2014).
3. Tecle, E. & Gagneux, P. Sugar-coated sperm: Unraveling the functions of the mammalian sperm glycocalyx. *Mol Reprod Dev* **82**, 635-650 (2015).
4. Abril-Parreno, L.*, et al.* Identification and characterization of O-linked glycans in cervical mucus as biomarkers of sperm transport: A novel sheep model. *Glycobiology* **32**, 23-35 (2022).
5. Guberman, M. & Seeberger, P.H. Automated Glycan Assembly: A Perspective. *J Am Chem Soc* **141**, 5581-5592 (2019).
6. Cummings, R.D. The repertoire of glycan determinants in the human glycome. *Mol Biosyst* **5**, 1087-1104 (2009).
7. 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 (2022).
8. Bojar, D. & Lisacek, F. Glycoinformatics in the Artificial Intelligence Era. *Chem Rev* **122**, 15971-15988 (2022).
9. Kuhn, T. *The Structure of Scientific Revolutions*, (The University of Chicago Press, 2012).
10. Zhang, G.L.*, et al.* The Human Ganglioside Interactome in Live Cells Revealed Using Clickable Photoaffinity Ganglioside Probes. *J Am Chem Soc* **146**, 17801-17816 (2024).
11. Canner, S.W. & Gray, J.J. Protein Interaction of CArbohydrate Predictor (PICAP). in *Society for Glycobiology Annual Meeting* (Amelia Island, Florida, 2024).
12. Zelensky, A.N. & Gready, J.E. The C-type lectin-like domain superfamily. *FEBS J* **272**, 6179-6217 (2005).
13. Brown, G.D., Willment, J.A. & Whitehead, L. C-type lectins in immunity and homeostasis. *Nat Rev Immunol* **18**, 374-389 (2018).
14. Harrison, K., Mackay, A.S., Kambanis, L., Maxwell, J.W.C. & Payne, R.J. Synthesis and applications of mirror-image proteins. *Nat Rev Chem* **7**, 383-404 (2023).
15. Xu, Y. & Zhu, T.F. Mirror-image T7 transcription of chirally inverted ribosomal and functional RNAs. *Science* **378**, 405-412 (2022).
16. Wang, Z., Xu, W., Liu, L. & Zhu, T.F. A synthetic molecular system capable of mirror-image genetic replication and transcription. *Nat Chem* **8**, 698-704 (2016).
17. Wang, M.*, et al.* Mirror-Image Gene Transcription and Reverse Transcription. *Chem-Us* **5**, 848-857 (2019).

18. Weidmann, J., Schnolzer, M., Dawson, P.E. & Hoheisel, J.D. Copying Life: Synthesis of an Enzymatically Active Mirror-Image DNA-Ligase Made of D-Amino Acids. *Cell Chem Biol* **26**, 645-651 e643 (2019).
19. Milton, R.C., Milton, S.C. & Kent, S.B. Total chemical synthesis of a D-enzyme: the enantiomers of HIV-1 protease show reciprocal chiral substrate specificity [corrected]. *Science* **256**, 1445-1448 (1992).
20. Liu, C.*, et al.* Identification of Monosaccharide Derivatives as Potent, Selective, and Orally Bioavailable Inhibitors of Human and Mouse Galectin-3. *J Med Chem* **65**, 11084-11099 (2022).
21. Marinec, P.S.*, et al.* A Non-immunogenic Bivalent d-Protein Potently Inhibits Retinal Vascularization and Tumor Growth. *ACS Chem Biol* **16**, 548-556 (2021).
22. Adamala, K.P.*, et al.* Confronting Risks of Mirror Life. *Science* (2024).
23. Perrin, D. In response to "Confronting risks of mirror life". (Science eLetter, 2024).
24. Briand, J.P.*, et al.* A retro-inverso peptide corresponding to the GH loop of foot-and-mouth disease virus elicits high levels of long-lasting protective neutralizing antibodies. *Proc Natl Acad Sci U S A* **94**, 12545-12550 (1997).
25. Yang, C.*, et al.* Single Dose of Protein Vaccine with Peptide Nanofibers As Adjuvants Elicits Long-Lasting Antibody Titer. *ACS Biomater Sci Eng* **4**, 2000-2006 (2018).
26. Horejsi, V., Matousek, V., Ticha, M. & Trnka, T. Quantitative inhibition of lectin binding to immobilized carbohydrates. in *Sixth International Lectin Meeting*, Vol. 4 (eds. Bog-Hansen, T.C. & Breborowicz, J.) 297-306 (De Gruyter, Poznan, 1984).
27. Ohashi, H., Ohashi, T., Misaki, R. & Fujiyama, K. Arabidopsis thaliana alpha1,2-l-fucosyltransferase catalyzes the transfer of l-galactose to xyloglucan oligosaccharides. *FEBS Lett* **593**, 187-194 (2019).
28. Shishiuchi, R.*, et al.* Discovery of alpha-l-Glucosidase Raises the Possibility of alpha-l-Glucosides in Nature. *ACS Omega* **7**, 47411-47423 (2022).
29. Zong, C.*, et al.* Heparan Sulfate Microarray Reveals That Heparan Sulfate-Protein Binding Exhibits Different Ligand Requirements. *J Am Chem Soc* **139**, 9534-9543 (2017).
30. Wang, L.*, et al.* Efficient platform for synthesizing comprehensive heparan sulfate oligosaccharide libraries for decoding glycosaminoglycan-protein interactions. *Nat Chem* **15**, 1108-1117 (2023).
31. Ruprecht, C.*, et al.* A Glycan Array-Based Assay for the Identification and Characterization of Plant Glycosyltransferases. *Angew Chem Int Ed Engl* **59**, 12493-12498 (2020).
32. Wesener, D.A.*, et al.* Recognition of microbial glycans by human intelectin-1. *Nat Struct Mol Biol* **22**, 603-610 (2015).
33. Stowell, S.R.*, et al.* Microbial glycan microarrays define key features of host-microbial interactions. *Nat Chem Biol* **10**, 470-476 (2014).
34. Zheng, R.B.*, et al.* Insights into Interactions of Mycobacteria with the Host Innate Immune System from a Novel Array of Synthetic Mycobacterial Glycans. *ACS Chem Biol* **12**, 2990-3002 (2017).
35. Geissner, A.*, et al.* Microbe-focused glycan array screening platform. *Proc Natl Acad Sci U S A* **116**, 1958-1967 (2019).
36. Benjamin, S.V.*, et al.* A human lectin array for characterizing host-pathogen interactions. *J Biol Chem* **300**, 107869 (2024).
37. Ghosh, S.*, et al.* Glycan analysis probes inspired by human lectins for investigating host-microbe crosstalk. *bioRxiv*, 1-39 (2024).
38. Eichler, J. N-glycosylation in Archaea-New roles for an ancient posttranslational modification. *Mol Microbiol* **114**, 735-741 (2020).
39. Bojar, D., Powers, R.K., Camacho, D.M. & Collins, J.J. Deep-Learning Resources for Studying Glycan-Mediated Host-Microbe Interactions. *Cell Host Microbe* **29**, 132-144 e133 (2021).

40. Xu, M., Geng, Y., Zhang, Y. & Yang, L. GlycanML: A Multi-Task and Multi-Structure Benchmark for Glycan Machine Learning. *arXiv.org*, 1-14 (2024).
41. Woods, R.J. Predicting the Structures of Glycans, Glycoproteins, and Their Complexes. *Chem Rev* **118**, 8005-8024 (2018).
42. Agostino, M., Yuriev, E. & Ramsland, P.A. A computational approach for exploring carbohydrate recognition by lectins in innate immunity. *Front Immunol* **2**, 23 (2011).
43. Ranaudo, A., Giulini, M., Pelissou Ayuso, A. & Bonvin, A. Modeling Protein-Glycan Interactions with HADDOCK. *J Chem Inf Model* **64**, 7816-7825 (2024).
44. Anfinsen, C.B. Principles that govern the folding of protein chains. *Science* **181**, 223-230 (1973).
45. Jumper, J.*, et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
46. Bryant, P. & Elofsson, A. Peptide binder design with inverse folding and protein structure prediction. *Commun Chem* **6**, 229 (2023).
47. Klamer, Z.L.*, et al.* CarboGrove: a resource of glycan-binding specificities through analyzed glycan-array datasets from all platforms. *Glycobiology* **32**, 679-690 (2022).
48. Lundstrom, J., Korhonen, E., Lisacek, F. & Bojar, D. LectinOracle: A Generalizable Deep Learning Model for Lectin-Glycan Binding Prediction. *Adv Sci (Weinh)* **9**, e2103807 (2022).
49. Hashimoto, K., Takigawa, I., Shiga, M., Kanehisa, M. & Mamitsuka, H. Mining significant tree patterns in carbohydrate sugar chains. *Bioinformatics* **24**, i167-173 (2008).
50. Aoki-Kinoshita, K.F., Ueda, N., Mamitsuka, H. & Kanehisa, M. ProfilePSTMM: capturing tree-structure motifs in carbohydrate sugar chains. *Bioinformatics* **22**, e25-34 (2006).
51. Coff, L., Chan, J., Ramsland, P.A. & Guy, A.J. Identifying glycan motifs using a novel subtree mining approach. *BMC Bioinformatics* **21**, 42 (2020).
52. Klamer, Z.*, et al.* Mining High-Complexity Motifs in Glycans: A New Language To Uncover the Fine Specificities of Lectins and Glycosidases. *Anal Chem* **89**, 12342-12350 (2017).
53. Ho, A.D.*, et al.* An Automated Approach to Assess Relative Galectin-Glycan Affinity Following Glycan Microarray Analysis. *Front Mol Biosci* **9**, 893185 (2022).
54. Klamer, Z. & Haab, B. Combined Analysis of Multiple Glycan-Array Datasets: New Explorations of Protein-Glycan Interactions. *Anal Chem* **93**, 10925-10933 (2021).
55. Liu, T., Lin, Y., Wen, X., Jorissen, R.N. & Gilson, M.K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* **35**, D198-201 (2007).
56. Zdrazil, B.*, et al.* The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res* **52**, D1180-D1192 (2024).
57. Carpenter, E.J., Seth, S., Yue, N., Greiner, R. & Derda, R. GlyNet: a multi-task neural network for predicting protein-glycan interactions. *Chem Sci* **13**, 6669-6686 (2022).
58. Burkholz, R., Quackenbush, J. & Bojar, D. Using graph convolutional neural networks to learn a representation for glycans. *Cell Rep* **35**, 109251 (2021).
59. Vinga, S. & Almeida, J. Alignment-free sequence comparison-a review. *Bioinformatics* **19**, 513-523 (2003).
60. Qian, Y., Zhang, Y. & Zhang, J. Alignment-Free Sequence Comparison With Multiple k Values. *IEEE/ACM Trans Comput Biol Bioinform* **18**, 1841-1849 (2021).
61. Kingma, D.P. & Ba, J. Adam: A Method for Stochastic Optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)* (2014).
62. Shams-Ud-Doha, K., Kitova, E.N., Kitov, P.I., St-Pierre, Y. & Klassen, J.S. Human Milk Oligosaccharide Specificities of Human Galectins. Comparison of Electrospray Ionization Mass Spectrometry and Glycan Microarray Screening Results. *Anal Chem* **89**, 4914-4921 (2017).

63. Bachhawat-Sikder, K., Thomas, C.J. & Surolia, A. Thermodynamic analysis of the binding of galactose and poly-N-acetyllactosamine derivatives to human galectin-3. *FEBS Lett* **500**, 75-79 (2001).
64. Kalliokoski, T., Kramer, C., Vulpetti, A. & Gedeck, P. Comparability of mixed IC(5)(0) data - a statistical analysis. *PLoS One* **8**, e61007 (2013).
65. Landrum, G.A. & Riniker, S. Combining IC(50) or K(i) Values from Different Sources Is a Source of Significant Noise. *J Chem Inf Model* **64**, 1560-1567 (2024).
66. Horovitz, A. & Levitzki, A. An accurate method for determination of receptor-ligand and enzyme-inhibitor dissociation constants from displacement curves. *Proc Natl Acad Sci U S A* **84**, 6654-6658 (1987).
67. Wang, T.Y.*, et al.* Fundamental Determinants of the Accuracy of Equilibrium Constants for Affinity Complexes. *Anal Chem* **95**, 15826-15832 (2023).
68. Kramer, C., Kalliokoski, T., Gedeck, P. & Vulpetti, A. The experimental uncertainty of heterogeneous public K(i) data. *J Med Chem* **55**, 5165-5173 (2012).
69. Wu, S.C.*, et al.* Full-Length Galectin-3 Is Required for High Affinity Microbial Interactions and Antimicrobial Activity. *Front Microbiol* **12**, 731026 (2021).
70. Wu, S.C.*, et al.* Whole microbe arrays accurately predict interactions and overall antimicrobial activity of galectin-8 toward distinct strains of Streptococcus pneumoniae. *Sci Rep* **13**, 5324 (2023).
71. Yoshikai, Y., Mizuno, T., Nemoto, S. & Kusuhara, H. Difficulty in chirality recognition for Transformer architectures learning chemical structures from string representations. *Nat Commun* **15**, 1197 (2024).
72. Sojitra, M.*, et al.* Genetically encoded multivalent liquid glycan array displayed on M13 bacteriophage. *Nat Chem Biol* **17**, 806-816 (2021).
73. Woese, C.R. & Fox, G.E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A* **74**, 5088-5090 (1977).
74. Rohden, F., Hoheisel, J.D. & Wieden, H.J. Through the looking glass: milestones on the road towards mirroring life. *Trends Biochem Sci* **46**, 931-943 (2021).
75. Glashagen, G.*, et al.* Coarse-grained and atomic resolution biomolecular docking with the ATTRACT approach. *Proteins* **88**, 1018-1028 (2020).
76. 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).
77. Mottarella, S.E.*, et al.* Docking server for the identification of heparin binding sites on proteins. *J Chem Inf Model* **54**, 2068-2078 (2014).
78. Boittier, E.D., Burns, J.M., Gandhi, N.S. & Ferro, V. GlycoTorch Vina: Docking Designed and Tested for Glycosaminoglycans. *J Chem Inf Model* **60**, 6328-6343 (2020).
79. 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* **38**, 276-287 (2017).
80. 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* **12**, 892-901 (2016).
81. Goldstein, I.J., Hollerman, C.E. & Smith, E.E. Protein-Carbohydrate Interaction. Ii. Inhibition Studies on the Interaction of Concanavalin a with Polysaccharides. *Biochemistry* **4**, 876-883 (1965).
82. Terada, T.*, et al.* Structural characterization of a rhamnose-binding glycoprotein (lectin) from Spanish mackerel (Scomberomorous niphonius) eggs. *Biochim Biophys Acta* **1770**, 617-629 (2007).
83. Mizgina, T.O.*, et al.* New l-Rhamnose-Binding Lectin from the Bivalve Glycymeris yessoensis: Purification, Partial Structural Characterization and Antibacterial Activity. *Mar Drugs* **22**(2023).

84. Watanabe, Y., *et al.* The function of rhamnose-binding lectin in innate immunity by restricted binding to Gb3. *Dev Comp Immunol* **33**, 187-197 (2009).
85. Wong, J.H. & Ng, T.B. Isolation and characterization of a glucose/mannose/rhamnose-specific lectin from the knife bean Canavalia gladiata. *Arch Biochem Biophys* **439**, 91-98 (2005).
86. Ardejani, M.S., Noodleman, L., Powers, E.T. & Kelly, J.W. Stereoelectronic effects in stabilizing protein-N-glycan interactions revealed by experiment and machine learning. *Nat Chem* **13**, 480-487 (2021).