Interpretable Supervised Learning and Graph-Based Optimization for Glycan-Lectin Binding

by

Joyce An

Submitted to the
Department of Materials Science and Engineering
in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Science

at the

Massachusetts Institute of Technology

May 2022

Signature of Author ……
Department of Materials Science and Engineering
May 6, 2022

Certified by ……...……
Rafael Gomez-Bombarelli
Assistant Professor in Materials Processing
Thesis Supervisor

Accepted by ……..……
James LeBeau
John Chipman Associate Professor of Materials Science and Engineering
Chair, DMSE Undergraduate Committee

ABSTRACT

Non-linear biological macromolecules, such as glycans, participate in a wide range of key structural, metabolic, and regulatory functions in all living organisms. Many of these essential roles involve interactions with glycan-binding proteins called lectins. As a result, there is particular interest in the design of highly specific glycan binders to critical lectins such as dendritic cell-specific ICAM-grabbing non-integrin (DC-SIGN). However, insufficient knowledge of the binding specificity of lectins, combined with the enormous structural complexity of glycans that range from linear to highly branched, serve as a barrier to the rational computational and experimental design of effective glycan binders. Here, using mammalian microarray data from the Consortium of Functional Glycomics, we predict glycan-lectin binding affinity and lectin specificity using an interpretable graph-based supervised learning framework. For the first time, we uncovered both monomers and motifs more precise than the monomer unit critical for lectin specificity. Furthermore, we developed a general graph-based optimization framework for macromolecules that employs the trained regression model ensembles to design glycans with high binding strength, low uncertainty in binding strength prediction, and low probability of human immunogenicity. Our work provides a general framework for iterative, chemistry-informed and topology-agnostic design in the macromolecular chemical space.

TABLE OF CONTENTS

1 INTRODUCTION

Glycans are a diverse class of macromolecules made by all living organisms with critical roles in a wide range of life-sustaining processes. Because glycans participate in many physiological processes through interactions with glycan-binding proteins called lectins, the design of highly specific glycan binders presents an enormous opportunity for the development of novel diagnostics, therapeutics, and vaccines (Amon et al., 2014; Geissner & Seeberger, 2016). However, rational design of novel glycan binders is impeded by incomplete knowledge of the binding specificity of lectins (Blixt et al., 2004).

Existing computational attempts to study glycan interactions and microarrays have encoded glycans as "glycowords" composed of polysaccharide units with characters representing monosaccharides and bonds (Bojar et al., 2021), as well as constituent monomer, dimers, and trimers (Carpenter et al., 2021) to predict glycan immunogenicity and binding strength to lectins. Others have employed frequent subtree mining to identify glycan motifs with a high frequency among strongly binding glycans and a low frequency among weakly binding glycans to predict lectin-glycan binding (Coff et al., 2020; Choletti et al., 2012). Although these prior attempts to study glycan interactions have predictive power, they lack the ability to reveal important motifs for binding, particularly substructures more precise than the individual monomer unit.

The vast diversity in glycan composition and topology presents another barrier in the design and optimization of novel glycan binders. In the design of novel chemical sequences, much existing work focuses on small molecules. Jensen et al. (2017) employed both a graph-based genetic algorithm approach and a graph-based generative model combined with Monte Carlo tree search to optimize the partition coefficient and synthetic accessibility of small molecules. Bengio et al.

(2021) applied GFlowNet, a novel generative method based on flow networks and local flow-matching conditions, to generate diverse drug molecules in large-scale synthesis tasks.

Following the success of optimization approaches to small molecules, several other works extend to the synthetic and biological macromolecular space. For linear macromolecules, such as proteins and linear copolymers, several works have utilized methods such as Monte Carlo Tree Search (MCTS) (Patra et al., 2020) and genetic algorithm (Meenakshisundaram et al., 2017; Schissel et al., 2021), often in conjunction with molecular dynamics (MD) simulations. Patra et al. (2020) interfaced MCTS and MD to identify linear copolymer sequences of up to 30 monomer units with zero interfacial energy between two immiscible homopolymers, while Meenakshisundaram et al. (2017) employed an MD-based genetic algorithm to minimize the interfacial energy at the polymer-polymer interface. Schissel et al. (2021) utilized a predictor-optimizer loop based on directed evolution to design novel abiotic nuclear-targeting miniproteins. However, the aforementioned methods for linear macromolecules do not extend to macromolecules with non-linear topologies, such as branched glycans.

In this work, we predicted glycan-lectin binding strength and specificity using an interpretable supervised learning framework developed in Mohapatra et al. (2022), and developed a graph-based genetic algorithm (GBGA) optimization framework to design glycans binders with improved properties. We employed mammalian glycan array data from the Consortium of Functional Glycomics (CFG) for dendritic cell-specific ICAM-grabbing non-integrin (DC-SIGN) and Concanavalin A (ConA) (Blixt et al., 2004). By training graph neural network (GNN) regression models to predict binding affinity and performing graph attribution on these trained models, we uncovered motifs critical to lectin-glycan specificity, for the first time revealing critical substructures more precise than the individual monomer unit. Finally, using the GBGA approach

we proposed a novel approach to rationally design glycans and macromolecules of diverse topologies.

2 GRAPH REPRESENTATION LEARNING AND GRAPH ATTRIBUTION

2.1 DATASET

After preprocessing to remove any data points with binding strength values more than three standard deviations outside of the mean, the DC-SIGN mammalian glycan microarray dataset consists of 3411 total data points with a mean binding strength in RFU of 156.264 and a standard deviation of 789.929 (Figure 1).

Although previous studies on glycan microarrays have employed log transformation of the RFU values to normalize the dataset, we found that standard scaling from scikit-learn on the raw, untransformed RFU values resulted in stronger model performance compared to log transformation of the RFU values followed by standard scaling, particularly with regards to prediction of the highest binding strengths. Therefore, we employed standard scaling on the original RFU values to transform the data for model training.

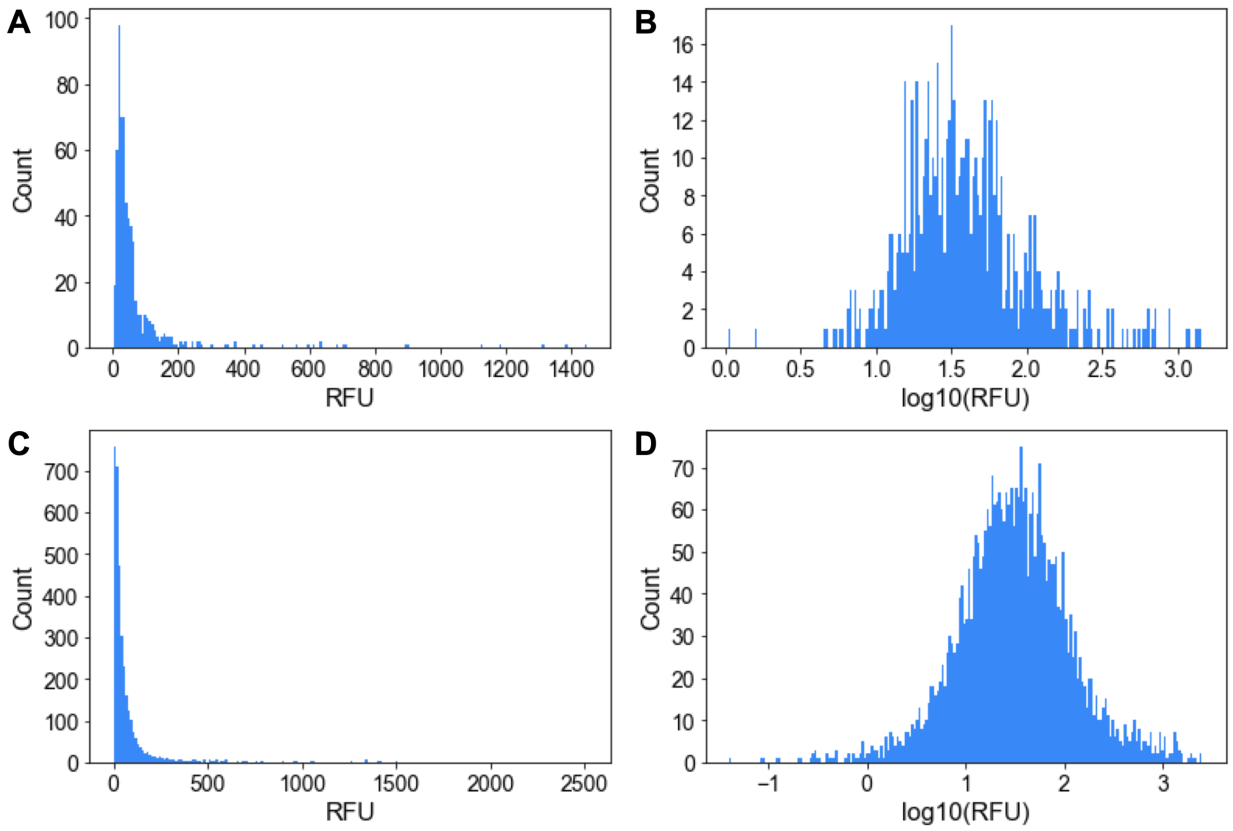


Figure 1. Distribution of binding strength values in relative fluorescence units (RFU) for the mammalian glycan microarray dataset for DC-SIGN. A. The original, non-log transformed average binding strengths for each glycan in RFU. **B.** The log-transformed average binding strengths for each glycan in RFU. **C.** The original, non-log transformed binding strengths for all the replicate data, resulting in 6 RFU values for each glycan. **D.** The log-transformed binding strengths for all the replicate data, resulting in 6 RFU values for each glycan.

2.2 REGRESSION

Despite the critical role lectins like DC-SIGN play in physiological processes such as human immunity, the lack of understanding of the specificity of these lectins serves as a barrier to greater understanding of biological mechanisms and design of promising glycan binders. Comparing glycans in the CFG dataset using a dissimilarity score computation developed in Mohapatra et al. (2021) illustrates the extremely high glycan-lectin specificity of DC-SIGN. Computing the dissimilarity score of all glycans in the dataset with reference to a random glycan with high binding affinity to DC-SIGN, there exists no clear correlation between glycan structure and binding affinity (Figure 2).

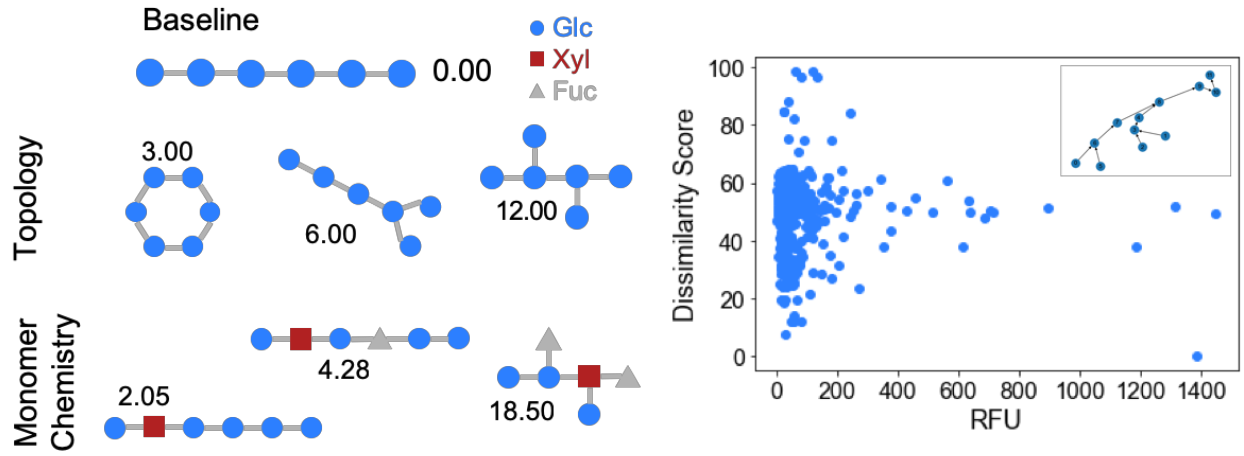


Figure 2. Dissimilarity score computation reveals no clear correlation between structural similarity to a strong binder and binding affinity. Using the reference point of a random glycan with high binding affinity in RFU to DC-SIGN, other glycans with varying RFU values follow a Gaussian distribution for dissimilarity, indicating high glycan-lectin specificity.

We aimed to predict glycan-lectin binding affinity and elucidate lectin specificity using a novel graph-based interpretable supervised learning approach. Representing the glycans as macromolecule graphs, we trained a suite of graph attention models to predict the lectin-glycan binding strength in relative fluorescence units (RFU) (Velickovic et al., 2017). We employed the Attentive FP GNN model architecture due to its superior performance in previous regression tasks and incorporated all replicate data, which includes 6 binding strength values in relative fluorescence units (RFU) for each glycan. For each glycan, monomers and bonds were featurized using circular fingerprints to capture the glycan's local chemistry and global topology (Figure 3A). Regression was performed over minimization of root mean square error (RMSE) on the validation dataset. We trained on 60 percent, validated on 20 percent, and tested on 20 percent of the data set. Hyperparameter optimization was performed on SigOpt over 300 iterations. To determine the final validation and test metrics, we averaged over 25 total model implementations, the top 5 optimized hyperparameter sets with 5 random seeds. On the held-out test data sets for DC-SIGN, the top-performing model ensemble had a root mean square error of 0.782 of the standard deviation of the training data set (Figure 3B, Table 1). It is important to note that in training the models there was unintentional leakage of some of the training points into the validation and test datasets; therefore, the models must be retrained for the final iteration of this work. However, the existing model ensembles can nonetheless serve as a basis for demonstrating the multi-step framework for rational design of novel glycans.

To identify the top-performing model ensemble, we benchmarked against a range of input graph features, including spacer length and physicochemical properties (Figure 4). Due to the well-documented dependence of lectin-glycan binding strength on the length of the spacers that link the glycans to the substrate in the microarray, we incorporated information about the estimated spacer

length in select models. With all edges featurized using circular fingerprints alone, we used 4 sets of graph node features: fingerprints alone, fingerprints and the estimated lengths of the spacer molecules in angstroms, fingerprints and 29 physicochemical descriptors from RDKit (Landrum et al., 2006), and fingerprints, spacer lengths, and physicochemical properties combined. In addition to the 4 sets of node features, we performed a control with estimated spacer lengths alone as both the node and edge features to determine the dependence of binding strength on spacer length alone. Although the models trained with only spacer lengths as node features demonstrate that a positive correlation between spacer length and binding strength exists (Table 1, Figure 4E), neither the presence of estimated spacer lengths nor physicochemical descriptors in the node features significantly impacted model performance, indicating that fingerprint featurization alone may effectively capture glycan features and properties essential for model prediction (Table 1, Figure 4). The top-performing model ensemble employs molecular fingerprints and physicochemical descriptors as input graph features, with the presence of spacers having little impact on model performance.

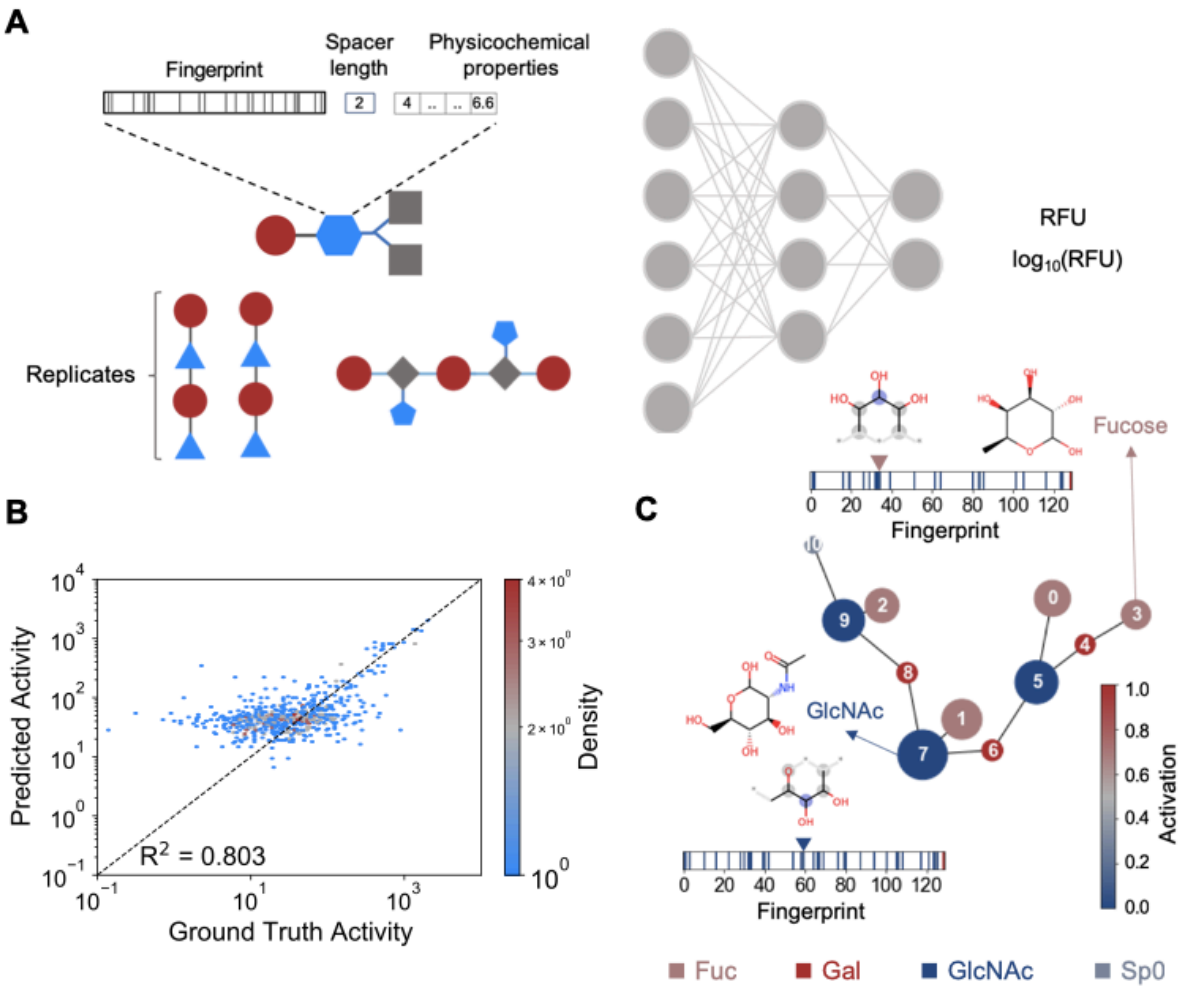


Figure 3. Supervised learning and graph attribution enable state-of-the-art prediction of lectin-glycan binding strength and identification of critical glycan substructures for lectin specificity. A. Combinations of monomers and bond features in the glycan graphs and RFU value transformations tested to find the optimal combinations. **B.** Parity plot obtained from training the optimal GNN model, representing an improvement over existing results in the literature. **C.** For a random glycan with high binding strength, N-acetylglucosamine (GlcNAc) and fucose (Fuc) contribute most to the model's prediction of high binding strength. The highest weighted substructures within both the highlighted GlcNAc and Fuc contain a cis, vicinal hydroxyl group.

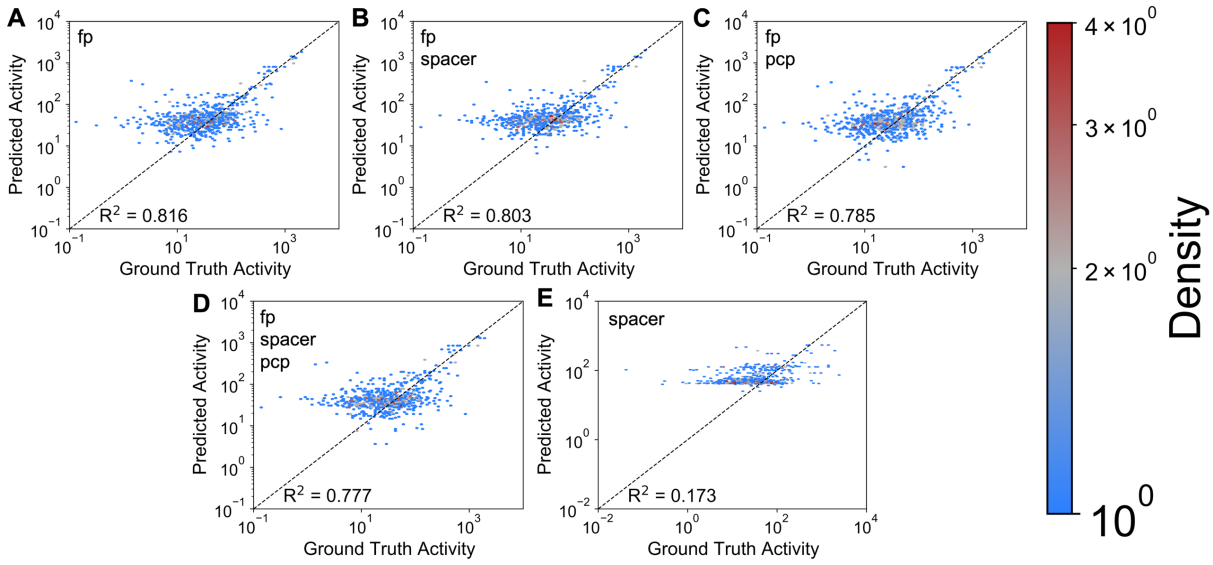


Figure 4. Parity plots for the DC-SIGN mammalian dataset consisting of all 6 RFU values for each glycan, resulting from training using the Attentive FP model architecture, fingerprint featurization, standard scaling normalization, and spacers in the glycan graphs. Each parity plot represents a distinct combination of node features: the presence of absence of fingerprints (fp), the presence or absence of estimated spacer lengths (spacer), and the presence or absence of physicochemical descriptors (pcp) and shows the predictions on the held-out test dataset for the top seed of the top hyperparameter set for each combination.

Include Finger-prints	Include Spacer Length	Include Physico-chemical Properties	Pearson r2		MAE		RMSE		Spearman r		Kendall's Tau		Loss	
			μ	***σ***	**μ**	***σ***	**μ**	***σ***	**μ**	***σ***	**μ**	***σ***	**μ**	***σ***
Yes	Yes	Yes	0.54	0.19	0.28	0.00	0.70	0.20	0.36	0.05	0.26	0.03	**0.12**	0.03
Yes	Yes	No	**0.55**	0.19	**0.28**	0.03	0.74	0.22	**0.34**	0.03	0.25	0.02	**0.12**	0.03
Yes	No	Yes	**0.55**	0.19	0.28	0.03	**0.74**	0.23	**0.36**	0.03	**0.27**	0.02	**0.12**	0.03
Yes	No	No	0.50	0.26	0.29	0.06	0.76	0.27	0.34	0.06	0.25	0.04	0.12	0.04
No	Yes	No	0.14	0.05	0.37	0.02	1.03	0.11	0.24	0.05	0.17	0.02	0.19	0.09

Table 1. The test metrics for the top hyperparameter set for each combination of node features. All models were run with the Attentive FP model architecture, fingerprint featurization, standard scaling normalization, all 6 RFU values for each glycan, and spacers in the glycan graphs. For each metric, the top value among all the model combinations is bolded for clarity.

2.3 GRAPH ATTRIBUTION

Graph attribution helped to decipher opaque structure-property relationships using the pre-trained GNNs, providing significant information in the absence of ground-truth understanding. Using the Integrated Gradients (IG) graph attribution method due to its consistency in previous attribution tasks, we obtained the node weights indicating the importance of each node in the glycans for model decision-making. For a randomly selected glycan with high binding strength to DC-SIGN, we found that N-acetylglucosamine and fucose contributed most to binding specificity.

In addition to the importance of individual monomers, for the first time critical substructures for lectin specificity more precise than the monomer level were also uncovered using attribution by multiplying the node weights with the corresponding monomer fingerprint. The highest weighted substructures within both highlighted monomers of the random glycan contain a cis, vicinal hydroxyl group, which is a characteristic feature present in the binding motifs of sugar ligands that interact with lectins like DC-SIGN, as reported in the literature (Feinberg et al., 2001) (Figure 3C).

We also performed an alternative analysis to determine the most prevalent glycan n-mers (monomers, dimers, and trimers) in glycans with the highest binding affinity, or RFU values, to two lectins: DC-SIGN and ConA. For each n-mer, the number of occurrences of the n-mer in each glycan was multiplied by the RFU value of the corresponding glycan. The sum of these values pooled across all glycans was divided by the total number of occurrences of the n-mer in the dataset to produce an "RFU score" for the n-mer denoting the weighted average of the RFU of values of glycans containing that n-mer. The top monomers contributing to high binding affinity to DC-SIGN include N-acetylglucosamine and fucose, agreeing with findings from graph attribution for the random strong glycan binder to DC-SIGN (Tables 2-4).

Monomer	RFU Score
Fuc	126.955
Man	102.663
GlcNAc	87.112
Gal	81.838
6S-GlcNAc	75.603
6S-Gal	54.124
GalNAc	53.681
3S-Gal	53.379
3S-GalNAc	48.732
6S-Glc	41.474

Table 2. The top 10 glycan monomers out of 32 total monomer types (not including spacers) contributing to high binding strength with DC-SIGN. Only monomers with at least 3 occurrences in the dataset are reported.

Dimer	RFU Score
GalNAc, 6S-GlcNAc	184.543
Fuc, GlcNAc	173.563
Fuc, 6S-GlcNAc	171.625
6S-Gal, 6S-GlcNAc	137.811
Man, Man	110.759
Gal, GlcNAc	97.074
Fuc, Gal	88.076
GlcNAc, GlcNAc	85.438
Man, GlcNAc	83.278
3S-Gal, GlcNAc	69.372

Table 3. The top 10 glycan dimers out of 56 total dimer types (not including those containing spacers) contributing to high binding strength with DC-SIGN. Only dimers with at least 3 occurrences in the dataset are reported.

Trimer	RFU Score
Fuc, GlcNAc, Man	212.008
Fuc, GlcNAc, Gal	211.756
Man, Man, Man	142.495
Gal, GlcNAc, Gal	113.672
GlcNAc, Gal, GlcNAc	110.111
Fuc, GlcNAc, 3S-Gal	103.094
Fuc, GlcNAc, GalNAc	98.058
Fuc, Gal, GlcNAc	95.719
Man, GlcNAc, GlcNAc	92.956
Man, Man, GlcNAc	91.877

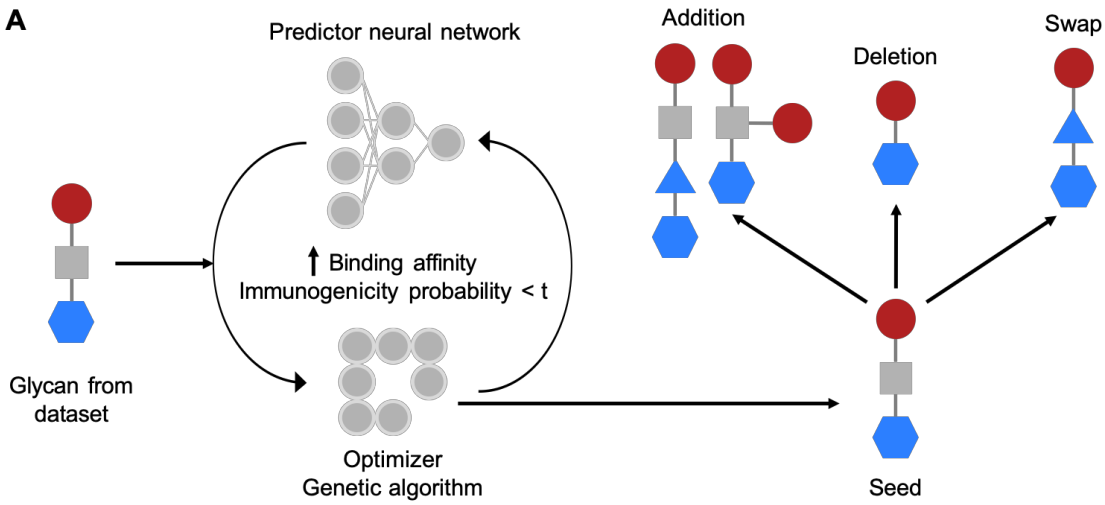
Table 4. The top 10 glycan trimers out of 75 total trimer types (not including those containing spacers) contributing to high binding strength with DC-SIGN. Only trimers with at least 3 occurrences in the dataset are reported.

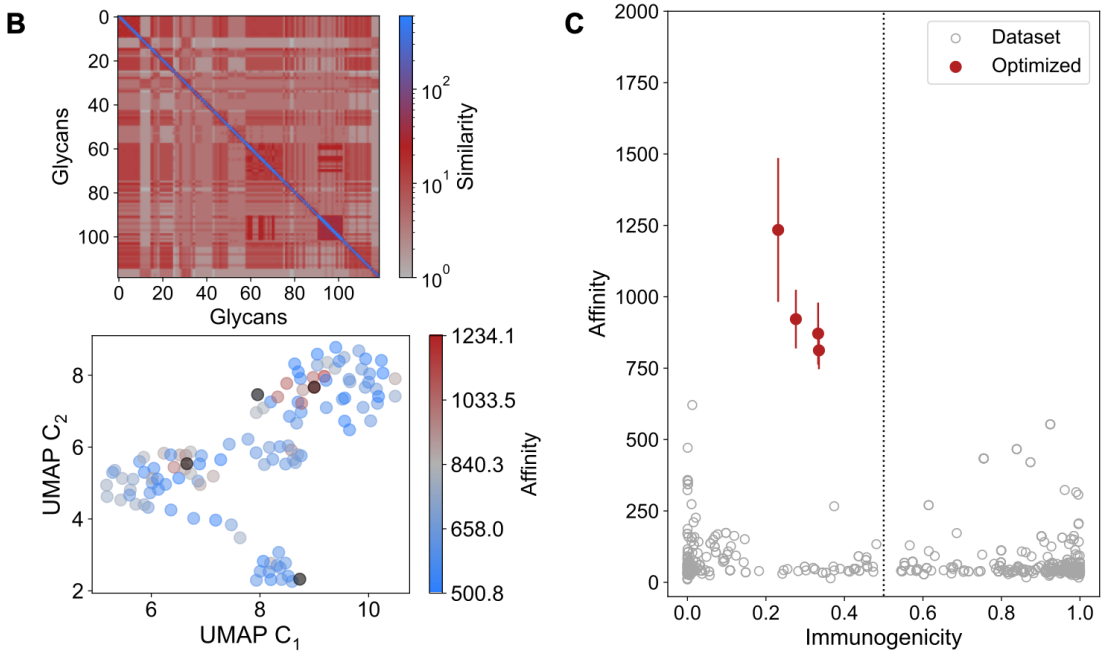
3 GRAPH-BASED OPTIMIZATION OF GLYCANS USING GENETIC ALGORITHM

By leveraging the effective representation of glycans as graphs, we developed GBGA optimization framework for glycans. The approach is applicable to macromolecules of any topology. Taking a featurized glycan graph as input and using an ensemble of trained GNN models representing 5 different hyperparameter sets trained on 2 distinct dataset splits for a specified number of iterations, the optimizer mutates the graph against the objective function of increasing the mean model ensemble prediction of binding affinity to DC-SIGN subtracted by the standard deviation of the ensemble prediction; in other words, the optimizer maximizes the lower bound of the model ensemble prediction, In the meantime, the optimizer maintains the probability of human immunogenicity under a given threshold. Depending on the predicted probability of human immunogenicity of the inputted glycan, two optimization pathways are possible:

1. Initial probability of human immunogenicity under threshold: the algorithm proceeds with a series of mutations, only replacing the glycan seed for structures with both a higher predicted binding affinity than the previous seed and a probability of human immunogenicity below the threshold. For each prediction of human immunogenicity and binding affinity, GBGA employs pre-trained models developed in this work and in (Mohapatra et al., 2022).
2. Initial probability of human immunogenicity above threshold: the algorithm repeatedly mutates the graph, only replacing the glycan seed for structures with both a lower immunogenicity probability than the previous seed and binding affinity no lower than 80% of that of the previous seed. Once the immunogenicity threshold is reached, the algorithm proceeds as described in case 1.

For each mutation, the optimizer randomly chooses between addition, deletion, or swapping of monomers (Figure 5A). In the addition mutation, the optimizer selects an edge at random followed by a node connected to that edge. The optimizer then either inserts the new node directly in the chain adjacent to the selected node, or if the selected node is not saturated with the maximum possible number of neighboring nodes, forms a branch with a specified branching probability. The identity of the added monomer is randomly chosen from 900 commonly occurring monomers in the CFG and GlycoBase datasets (Blixt et al., 2004; Bojar et al., 2021) (Figure 6). In the deletion mutation, the optimizer deletes a node at random, as well as all n edges originally connected to the deleted node. The optimizer then forms $n - 1$ new edges between the previously connected nodes, selecting the specific combination of connected nodes at random (Figure 7). In the swapping mutation, an existing node in the graph and the identity of the replacement node are selected at random (Figure 8). The three mutation types allow for. traversal of the vast compositional and topological diversity of the glycan space.


A
Predictor neural network
Addition
Deletion
Swap
Binding affinity
Immunogenicity probability < t
Glycan from dataset
Optimizer
Genetic algorithm
Seed


B
Glycans
Similarity
UMAP C2
UMAP C1
Affinity
C
Dataset
Optimized
Affinity
Immunogenicity

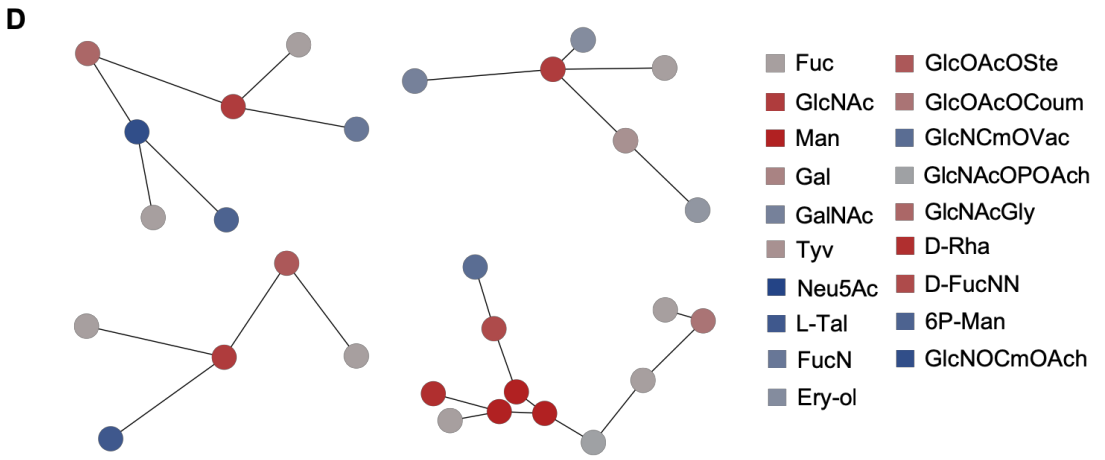

D
Fuc
GlcNAc
Man
Gal
GalNAc
Tyv
Neu5Ac
L-Tal
FucN
Ery-ol
GlcOAcOSte
GlcOAcOCoum
GlcNCmOVac
GlcNAcOPOAch
GlcNAcGly
D-Rha
D-FucNN
6P-Man
GlcNOCmOAch

Figure 5. GBGA generates optimal structures with both high binding affinity and low immunogenicity. A. Glycan graphs are seeded in a GA-predictor optimization loop, where structures undergo a random mutation in each optimization loop against an objective function. **B.** Unsupervised learning of optimized glycans using similarity matrices, UMAP, and weighted clustering aids selection of diverse top sequences. **C.** Relationship between immunogenicity probability and binding affinity for both top 5 optimized glycans and original data set with, t=0.5 as indicated by the dotted line. **D.** Top 5 optimized glycans, with legend in the panel.

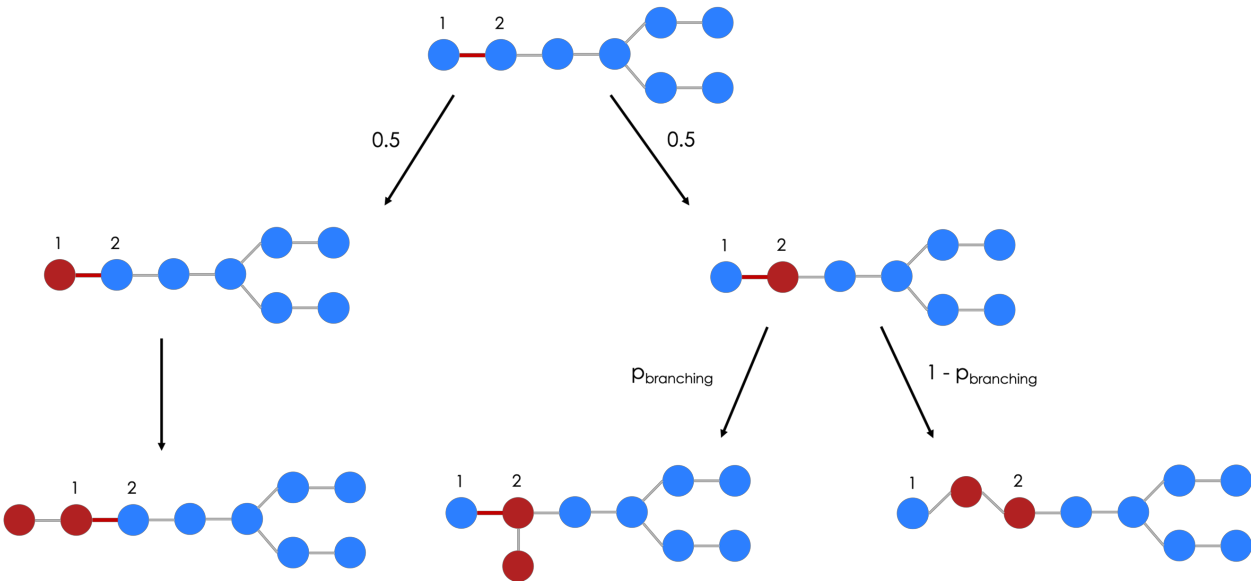


Figure 6. The addition mutation selects an edge at random, followed by a node connected to that edge. A new branch is formed with probability $p_{branching}$.

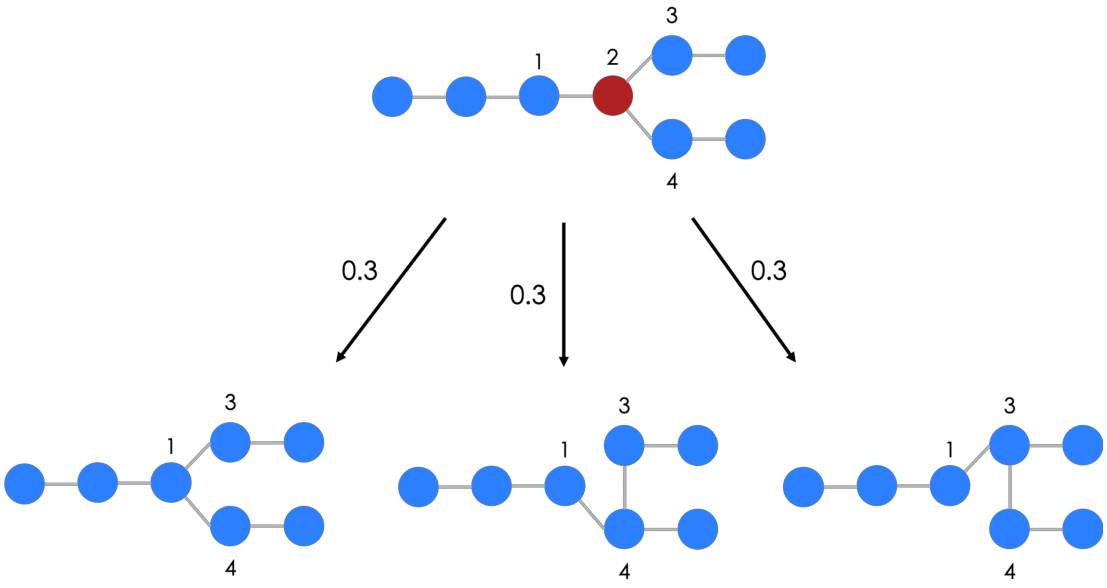


Figure 7. The deletion mutation deletes a node at random, as well as all n edges connected to the deleted node. It forms $n-1$ new edges between the formerly connected nodes.

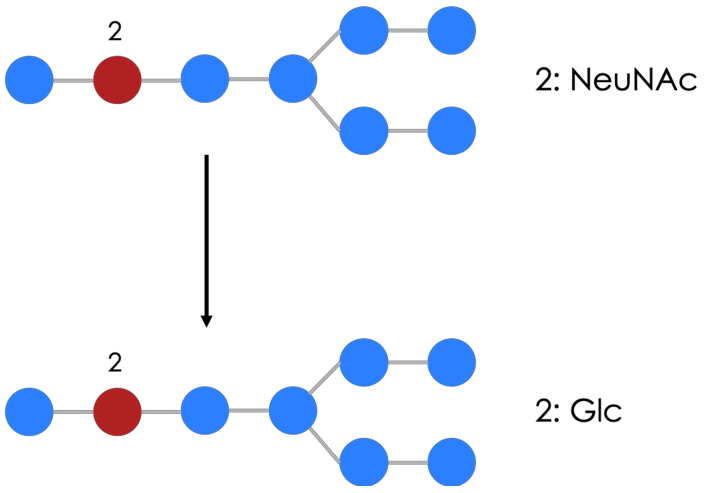


Figure 8. The swapping mutation selects both a node in the graph and a new node identity at random.

Starting with the top 50 strongest binders to DC-SIGN as the original seeds and optimizing over 6000 iterations for each seed, we generated over 1000 glycans with high binding affinity to DC-SIGN and a probability of human immunogenicity under the threshold of 0.5 (Figure 9).

We employed unsupervised learning over the pairwise similarity matrix to select diverse top sequences among the optimized glycan structures. Using two-component UMAP and weighted clustering using k-means, we identified four distinct clusters and selected the glycan with the highest binding affinity from each cluster (McInnes et al., 2018; Likas et al., 2003) (Figures 4B, C). The resulting top four glycans encompass a wide range of topologies and compositions (Figures 4C, D, Figure 9).

The uncertainty of the model ensemble predictions provides insight into the limitations of the GBGA approach. The standard deviation and standard error of the model ensemble predictions of binding affinity for glycans in the dataset are on par with the standard deviation and standard error of glycan microarray data when replicate measurements at multiple spots on the microarray. However, the standard deviation and standard error of the model predictions for the top 50 optimized glycans generated through GBGA is much higher, indicating that while the model ensemble can effectively interpolate among relatively similar glycans within the dataset, it is less successful at extrapolating to predict binding affinity for novel glycans structurally distinct from those it has seen during training (Table 5).

In addition to DC-SIGN, we also optimized glycans that bind with ConA, using a similar model training, optimization and unsupervised learning protocol. Consistent with experimental observations, we noted that most glycans that were predicted to bind with ConA were rich in mannose (Maupin et al., 2012).

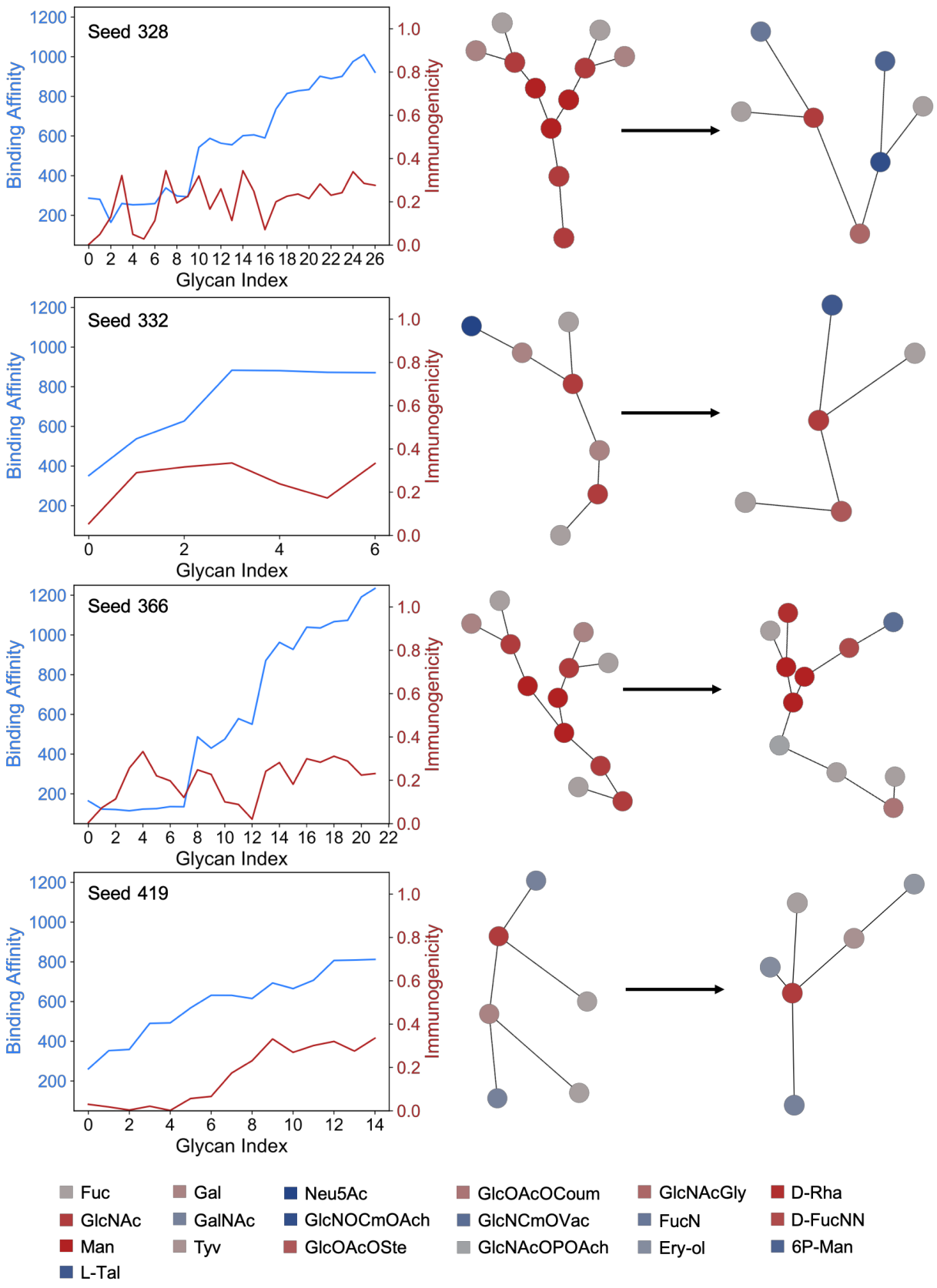

Seed 328
Seed 332
Seed 366
Seed 419
Binding Affinity
Immunogenicity
Glycan Index
Fuc
Gal
Neu5Ac
GlcOAcOCoum
GlcNAcGly
D-Rha
GlcNAc
GalNAc
GlcNOCmOAch
GlcNCmOVac
FucN
D-FucNN
Man
Tyv
GlcOAcOSte
GlcNAcOPOAch
Ery-ol
6P-Man
L-Tal

Figure 8. Evolution of glycan binding affinity to DC-SIGN and probability of human immunogenicity over the glycan index, as well as visualization of each glycan before and after all iterations of GBGA optimization with monomer legend in panel. The glycan index refers to the number of new glycans generated in the optimization loop with improved properties compared to the previous top glycan, and the seed numbers indicate the IDs of the original glycans that enter the optimization loop.

	Average Standard Deviation	Standard Error
Experimental (Dataset)	45.9	18.7
Experimental (Dataset, RFU > 500)	155.4	63.4
Model Ensemble (Dataset)	13.7	4.3
Model Ensemble (Optimized, RFU > 500)	236.2	74.7

Table 5. Average standard deviation and standard error for experimental data at six distinct spots on the glycan microarray for each glycan for both the entire dataset and the top 16 glycan binders to DC-SIGN with RFU > 500, as well as standard deviation and standard error for the model ensemble predictions for both the entire dataset and the top 50 optimized glycans generated through GBGA.

4 FUTURE WORK AND CONCLUSION

We predicted lectin-glycan binding strength with state-of-the-art accuracy and uncovered motifs critical to the model's prediction of binding strength using an interpretable supervised learning framework. In addition, we designed glycans with high binding affinity to DC-SIGN and low probability of human immunogenicity using a novel graph-based optimization framework. Supervised learning with GNNs provides a generalized system to learn over diverse glycan structures and lectin identities. Graph attribution provides explanations for model predictions, for the first time revealing both key monomers and substructures more precise than the individual monomer level. Graph-based optimization leverages the effective representation of macromolecules as graphs to design glycans with a wide range of topologies and monomer compositions. This work illustrates a robust method for the rational design in the general macromolecular space. In the future, we aim to perform graph attribution both across the entire dataset and for the top optimized glycans to elucidate key glycan features responsible for high binding affinity to DC-SIGN in both existing and novel glycan structures. As an ultimate step, we hope to extend this approach to include multiple lectins, as well as experimentally validate the highly specific glycans that we design.

5 ACKNOWLEDGEMENTS

This thesis would not have been possible without the incredible mentorship and support of a few individuals. First, I would like to thank my mentors Professor Rafa Gomez-Bombarelli and Dr. Somesh Mohapatra for their willingness to teach me concepts and commitment to supporting my academic and personal growth. When I first reached out to the lab as a junior with little coding experience, Professor Gomez-Bombarelli generously allowed me the opportunity to work with the group, and from the start, Somesh has patiently helped me learn the fundamentals and gain greater independence as a researcher. Both of my mentors have provided indispensable guidance on not only academic and career-related matters but also life in general, all of which has made the body of research in this thesis possible.

I would also like to thank the entirety of the Gomez-Bombarelli Group for exposing me to a wide range of exciting work through group meetings, supporting my research endeavors, and generally providing a supportive environment for learning. In addition, I am grateful to the DMSE academic office for their support with thesis proposal, presentation, and document logistics, as well as my friends, family, and Course 3 peers and classmates for much-needed morale boosts throughout the research process.

From a contributions standpoint, I would like to thank Dr. Somesh Mohapatra for his critical role in designing and planning all aspects of this research as well as performing the graph attribution work. I would also like to acknowledge Omar Santiago-Reyes for proposing this research idea, setting up a collaboration with the Kiessling Lab at MIT, and preparing the dataset and raw glycan sequences for the work, as well as Dr. Daria Kim from the Kiessling Lab and Professor Laura Kiessling for their critical support in planning this work.

6 REFERENCES

Ron Amon, Eliran Moshe Reuven, Shani Leviatan Ben-Arye, and Vered Padler-Karavani. Glycans in immune recognition and response. *Carbohydrate research*, 389:115–122, 2014.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation. *arXiv*, 2106.04399v2, 2021.

Ola Blixt, Steve Head, Tony Mondala, Christopher Scanlan, Margaret E Huflejt, Richard Alvarez, Marian C Bryan, Fabio Fazio, Daniel Calarese, James Stevens, et al. Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. *Proceedings of the National Academy of Sciences*, 101(49):17033–17038, 2004.

Daniel Bojar, Rani K Powers, Diogo M Camacho, and James J Collins. Deep-learning resources for studying glycan-mediated host-microbe interactions. *Cell Host & Microbe*, 29(1):132–144, 2021.

Eric J Carpenter, Shaurya Seth, Noel Yue, Russell Greiner, and Ratmir Derda. Glynet: A multi-task neural network for predicting protein-glycan interactions. *bioRxiv*, 2021.

Sharath R Cholleti, Sanjay Agravat, Tim Morris, Joel H Saltz, Xuezheng Song, Richard D Cummings, and David F Smith. Automated motif discovery from glycan array data. *Omics: a journal of integrative biology*, 16(10):497–512, 2012.

Lachlan Coff, Jeffrey Chan, Paul A Ramsland, and Andrew J Guy. Identifying glycan motifs using a novel subtree mining approach. *BMC bioinformatics*, 21(1):1–18, 2020.

Hadar Feinberg, Daniel A Mitchell, Kurt Drickamer, and William I Weis. Structural basis for selective recognition of oligosaccharides by dc-sign and dc-signr. *Science*, 294(5549):2163–2166, 2001.

Andreas Geissner and Peter H Seeberger. Glycan arrays: from basic biochemical research to bioanalytical and biomedical applications. *Annual Review of Analytical Chemistry*, 9:223–247, 2016.

Jan H. Jensen. A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. Chemical Science, 10:3567-3572, 2019.

Greg Landrum et al. Rdkit: Open-source cheminformatics. 2006.

Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003.

Jon Lundstrøm, Emma Korhonen, Frederique Lisacek, and Daniel Bojar. Lectinoracle: A generalizable deep learning model for lectin–glycan binding prediction. *Advanced Science*, 9(1):2103807, 2022.

Shane L Mangold and Mary J Cloninger. Binding of monomeric and dimeric concanavalin a to mannose-functionalized dendrimers. *Organic & Biomolecular Chemistry*, 4(12):2458–2465, 2006.

Kevin A Maupin, Daniel Liden, and Brian B Haab. The fine specificity of mannose-binding and galactose-binding lectins revealed using outlier motif analysis of glycan array data. Glycobiology, 22(1):160–169, 2012.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 3(29):861, 2018.

Venkatesh Meenakshisundaram, Jui-Hsiang Hung, Tarak K Patra, and David S Simmons. Designing sequence-specific copolymer compatibilizers using a molecular-dynamics-simulation-based genetic algorithm. Macromolecules, 50(3):1155–1166, 2017.

Somesh Mohapatra, Joyce An, and Rafael Gomez-Bombarelli. Chemistry-informed macromolecule graph representation for similarity computation, unsupervised and supervised learning. Machine Learning: Science and Technology, 2022.

Tarak K Patra, Troy D Loeffler, and Subramanian KRS Sankaranarayanan. Accelerating copolymer inverse design using monte carlo tree search. Nanoscale, 12(46):23653–23662, 2020.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. the Journal of machine Learning research, 12:2825–2830, 2011.

Carly K Schissel, Somesh Mohapatra, Justin M Wolfe, Colin M Fadzen, Kamela Bellovoda, Chia-Ling Wu, Jenna A Wood, Annika B Malmberg, Andrei Loas, Rafael Gomez-Bombarelli, et al. Deep learning to design nuclear-targeting abiotic miniproteins. Nature Chemistry, 13(10):992–1000, 2021.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. arXiv preprint arXiv:1710.10903, 2017.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. Journal of medicinal chemistry, 63(16):8749–8760, 2019.

7 APPENDIX

The following sections demonstrate results of the interpretable supervised learning and graph-based optimization framework for another lectin, Concanavalin A (ConA).

7.1 DATASET

The ConA mammalian glycan microarray dataset consists of a majority of weak or non-binders, along with a small portion of strong binders to form a bimodal distribution of binding strength values (Figure A.1). After preprocessing to remove any data points with binding strength values more than three standard deviations outside of the mean, the total number of average RFU values in the ConA mammalian dataset is 590 with a mean binding strength in RFU of 3242.828 and a standard deviation of 7844.175. The CFG dataset provides 6 binding strength values for each glycan corresponding with 6 different spots on the glycan microarray (Blixt et al., 2004). The total number of replicate data points, 6 RFU values for each glycan, in the dataset is 3364 with a mean binding strength of 3414.378 and a standard deviation of 8078.214.

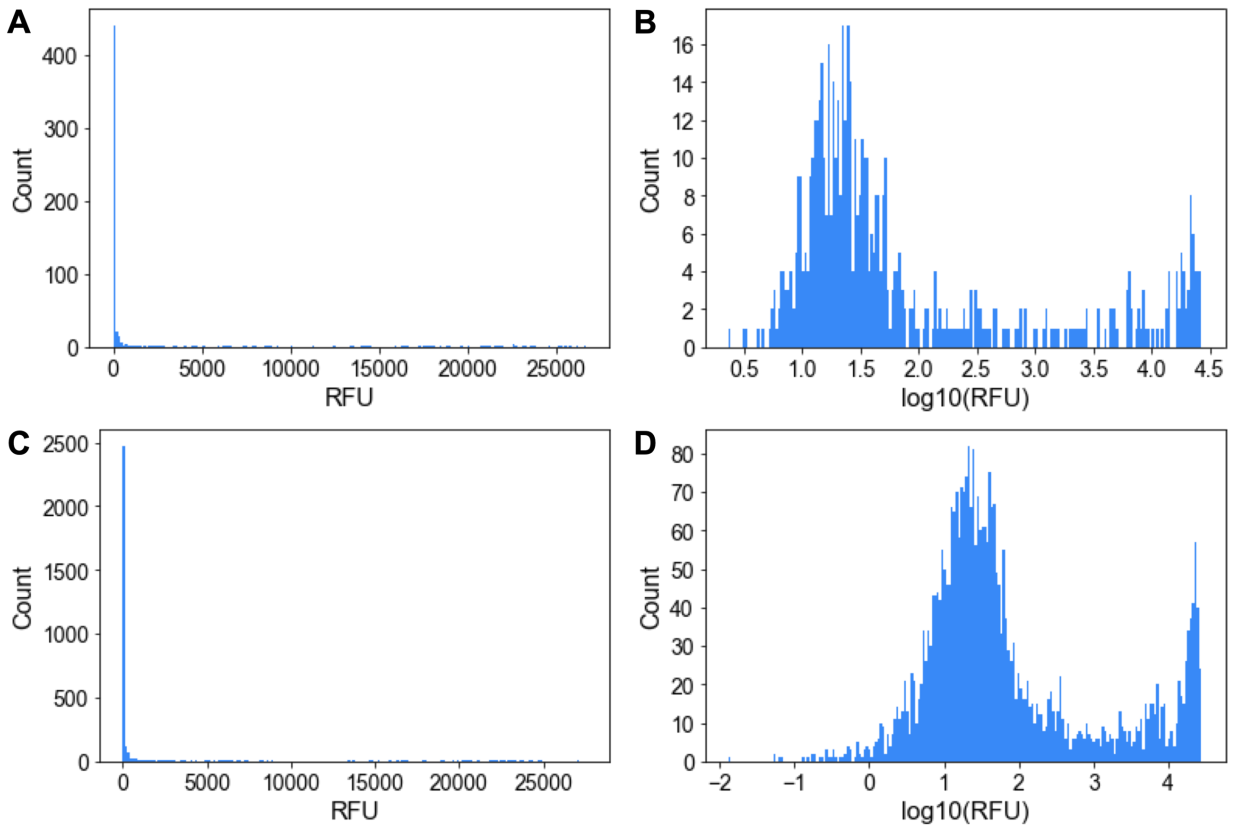


Figure A.1. Distribution of binding strength values in relative fluorescence units (RFU) for the mammalian glycan microarray dataset for ConA. A. The original, non-log transformed average binding strengths for each glycan in RFU. **B.** The log-transformed average binding strengths for each glycan in RFU. **C.** The original, non-log transformed binding strengths for all the replicate data, 6 RFU values for each glycan. **D.** The log-transformed binding strengths for all the replicate data, 6 RFU values for each glycan.

7.2 REGRESSION

Employing the Attentive FP model architecture (Xiong et al., 2019) with fingerprint featurization, we trained on 8 distinct combinations, composed of 2 different sets of binding strength values (average RFU values or all 6 RFU values for each glycan), 2 different normalization types (standard scaling (Pedregosa et al., 2011) alone or log10 transformation followed by standard scaling), and 2 different graph types (with or without spacers).

We employed the same model training process as for DC-SIGN: we utilized a 0.6, 0.2, 0.2 train, validation, test split, performed hyperparameter optimization on SigOpt over 300 iterations, and

obtained final validation and test metrics by averaging over 25 total model implementations for each of the 8 combinations. Among the models trained on the average RFU values for each glycan, the combination that results in the strongest model performance consists of log10 transformation before standard scaling and no spacers in the glycan graphs (Table A.1, Figure A.2). Among the models trained on all 6 RFU values for each glycan, the strongest combination consists of standard scaling alone without log10 transformation and spacers in the glycan graphs (Table A.1, Figure A.2). Comparison between the models trained on the average RFU values versus all RFU values indicates that employing all the replicate data results in stronger model performance. Therefore, we focused on using all 6 RFU values for each glycan in future analysis of additional lectins.

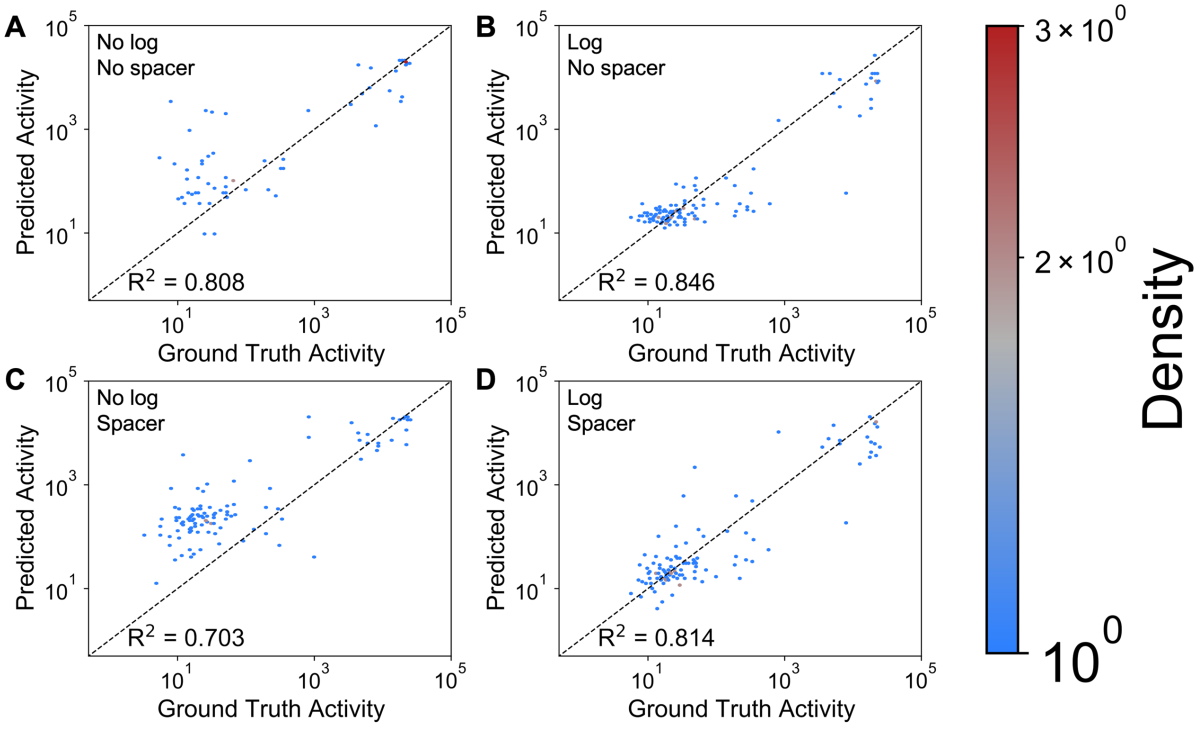


Figure A.2. Parity plots for the ConA mammalian dataset consisting of average RFU values for each glycan, resulting from training using the Attentive FP model architecture and fingerprint featurization. Each parity plot represents a distinct combination of normalization type (either standard scaling alone or log transformation followed by standard scaling) and graph types (including or not including spacers in graphs) and shows the predictions on the held-out test dataset for the top seed of the top hyperparameter set for each combination.

7.3 GRAPH-BASED GENETIC ALGORITHM (GBGA)

We employed the same process for optimizing glycan binders to DC-SIGN for ConA: we took the top 50 glycans with the highest binding affinity to ConA from the dataset, optimized over 6000 iterations, and used unsupervised learning to select diverse top sequences (Figures A.3, A.4). The resulting optimized glycan structures primarily consist of mannose and mannose-adjacent monomers, consistent with literature on the specificity of ConA (Mangold and Cloninger, 2006). Over the course of optimization, the predicted binding affinity generally rises sharply with the first few iterations and gradually stabilizes with a higher number of optimization loops (Figure A.4).

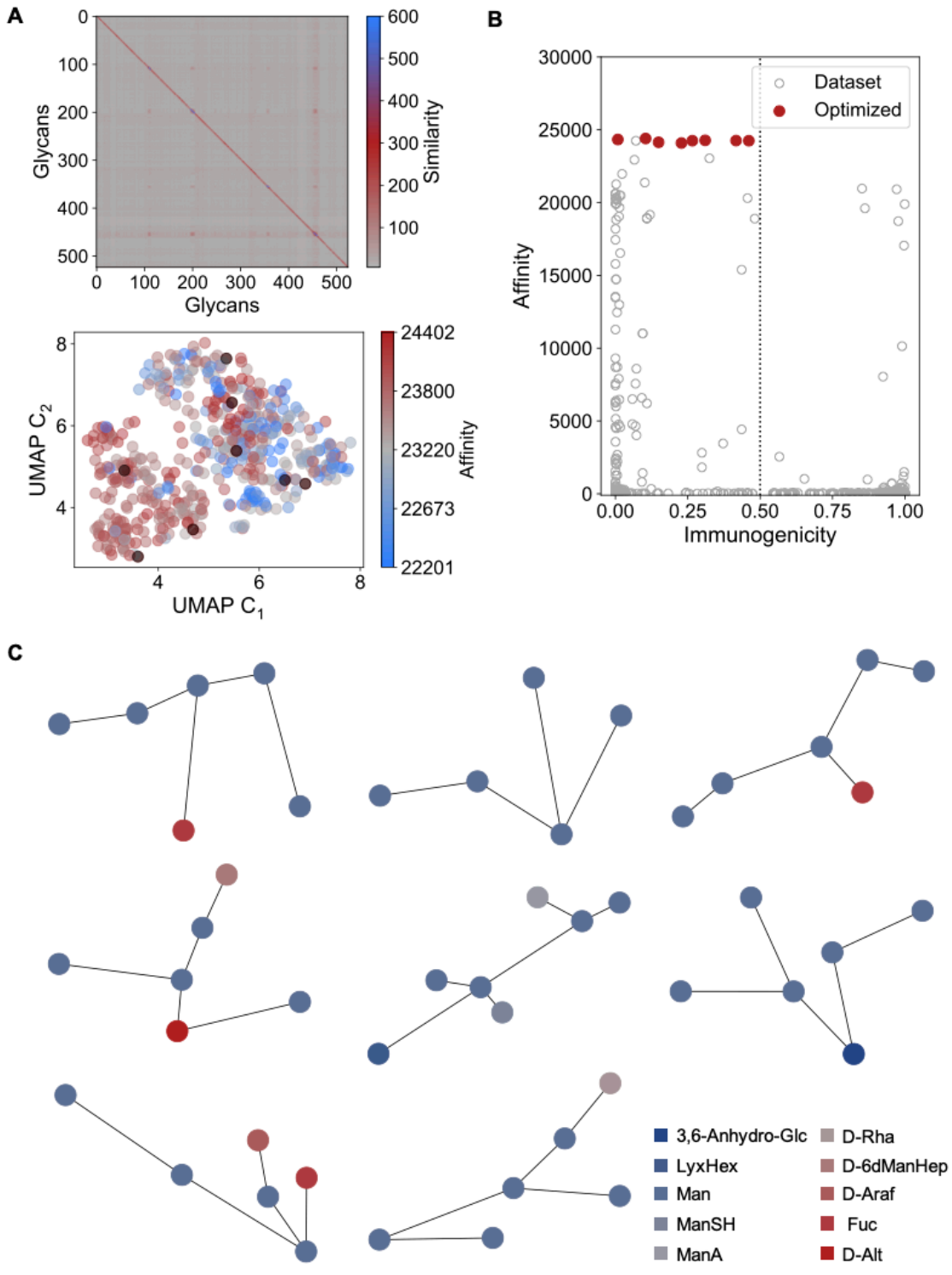


Figure A.3. GBGA generates optimal structures with both high binding affinity to lectin ConA and low immunogenicity. A. Unsupervised learning of optimized glycans using similarity

matrices, UMAP, and weighted clustering aids selection of diverse top sequences. **B.** Relationship between immunogenicity probability and binding affinity for both top 8 optimized glycans and original data set with, t=0.5, as indicated by the dotted line. **C.** Top 8 optimized glycans, with legend in the panel.

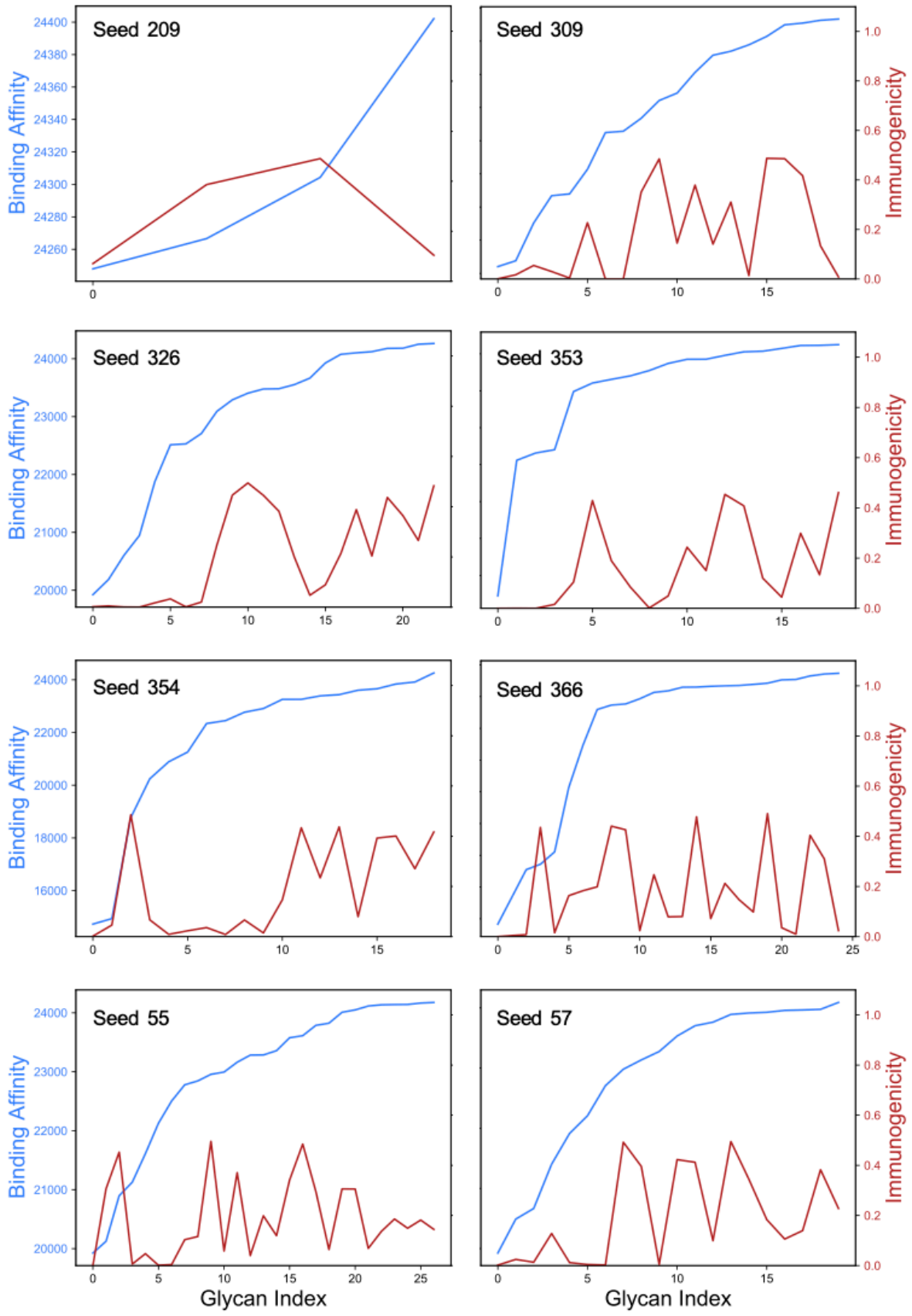

Seed 209
Seed 309
Seed 326
Seed 353
Seed 354
Seed 366
Seed 55
Seed 57
Binding Affinity
Immunogenicity
Glycan Index
Glycan Index

Figure A.4. Evolution of glycan binding affinity to ConA and probability of human immunogenicity over the glycan index, or the number of new glycans generated in the optimization loop with improved properties compared to the previous top glycan. The seed numbers indicate the IDs of the original glycans that enter the optimization loop.