CHEMISTRY-INFORMED MACROMOLECULE GRAPH REPRESENTATION FOR SIMILARITY COMPUTATION AND SUPERVISED LEARNING

Somesh Mohapatra, Joyce An & Rafael Gómez-Bombarelli[*]
Department of Materials Science and Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{someshm,joycean,rafagb}@mit.edu

ABSTRACT

Macromolecules are large, complex molecules composed of covalently bonded monomer units, existing in different stereochemical configurations and topologies. As a result of such chemical diversity, representing, comparing, and learning over macromolecules emerge as critical challenges. To address this, we developed a macromolecule graph representation, with monomers and bonds as nodes and edges, respectively. We captured the inherent chemistry of the macromolecule by using molecular fingerprints for node and edge attributes. For the first time, we demonstrated computation of chemical similarity between 2 macromolecules of varying chemistry and topology, using exact graph edit distances and graph kernels. We also trained graph neural networks for a variety of glycan classification tasks, achieving state-of-the-art results. Our work has two-fold implications – it provides a general framework for representation, comparison, and learning of macromolecules; and enables quantitative chemistry-informed decision-making and iterative design in the macromolecular chemical space.

1 INTRODUCTION

Macromolecules are ubiquitous and indispensable, from constituting what we are made up of to being present in almost everything we use. As biological macromolecules, they form the basis of life, serving as drivers of survival and growth functions. As synthetic macromolecules, humans have engineered the composition and topology to design structural components, sensors, shape-memory materials, drugs, encode messages, and much more (Lutz et al., 2016; Romio et al., 2020; Boydston et al., 2020; Thompson & Korley, 2020).

An individual macromolecule is a result of its monomer composition, connecting bonds, and their spatial arrangement. Monomer and bond are functions of atomic composition, stereochemistry, and arrangement, while spatial arrangement dictates the topology. Experimentalists and theoreticians have explored a vast chemical space by varying monomers, bonds, and topologies – linear and non-linear such as branched, star, and bottle-brush (Figure A.1) (Hiemenz & Lodge, 2007; Johnson et al., 2011; Alvaradejo et al., 2019).

In this work, we propose a graph representation for macromolecules. We use graph edit distances (GEDs) with Tanimoto chemical similarity matrices and propagation graph kernels to compute graph similarity. Further, we train a suite of graph neural network models on different tasks, achieving state-of-the-art results on a data set of glycans.

2 RELATED WORK

Representation. Macromolecules can be represented in line notation similar to simplified molecular-input line-entry system (SMILES) used for small molecules (Lin et al., 2019). Linear

[*] Corresponding author: rafagb@mit.edu.

biological macromolecules, such as proteins and DNA/RNA, are an exception, and represented as sequences of one/three-letter monomer codes. In a recent attempt, glycans (non-linear macromolecules) were represented as sequences, where groups of monosaccharides were clubbed into 'glycowords' and placed in hierarchical brackets (Bojar et al., 2021). Hierarchical fingerprinting is another approach, which follows a hierarchy of atomic, physicochemical, and morphological descriptors (Kim et al., 2018). However, these representations are limited by their coverage of chemical space, ability to support all topologies, and require significant customization for different monomers.

Similarity computation. In recent times, there have been significant advances in similarity computation using GED and graph kernels (Borgwardt et al., 2020; Blumenthal & Gamper, 2020). Development of software packages, such as graphkernels and GraKeL, has provided fast implementations of graph kernels (Sugiyama et al., 2018; Siglidis et al., 2020).

For linear biological macromolecules, such as proteins, DNA/RNA and linear glycans, there are several works for computation of sequence similarities.(Altschul et al., 1990; Bojar et al., 2021) Usually, sequence alignment is done using Smith-Waterman or Needleman-Wunsch algorithm, and scored with substitution matrices, such as BLOSUM62 (Smith & Waterman, 1981; Needleman & Wunsch, 1970; Eddy, 2004). The substitution matrices are based on evolutionary statistics thereby biasing the scoring towards the statistical frequency of a particular monomer's occurrence in the course of evolution, rather than chemical similarity. Apart from sequence alignment in linear macromolecules, edit distances, linear kernels and deep learning methods have been proposed to compute similarity (Jaakkola et al., 2000; Riesen & Bunke, 2009; Bileschi et al., 2019). In the case of non-linear macromolecules, alignment of glycans has been explored using q-grams, tree matching methods and tree kernels (Li et al., 2010; Hosoda et al., 2017; Coff et al., 2020). Unfortunately, the aforementioned methods are limited to biological macromolecules, and do not extend to the general macromolecular chemical space. Moreover, existing tools for biological macromolecules do not allow incorporation of unnatural monomers, and cannot handle non-linear topologies (except for glycan-specific tools).

Machine learning. The field of graph neural networks (GNN) has seen substantial developments in both model architecture and attribution. Different model architectures, such as graph convolutional network (GCN), graph attention network (GAT), message passing neural network (MPNN), SchNet, and Attentive FP, have demonstrated state-of-the-art results across various domains (Zhou et al., 2018; Schütt et al., 2017; Kearnes et al., 2016; Xiong et al., 2020). In a recent work, graph attribution has been studied quantitatively across four metrics – accuracy, stability, faithfulness and consistency - for a wide variety of tasks and model architectures (Sanchez-lengeling et al., 2020).

For macromolecular property prediction, Polymer Genome and similar works using hierarchical fingerprints predict glass transition temperature, dielectric point and other properties (Kim et al., 2018). There have been attempts to extrapolate macromolecular property by training over monomer features (St John et al., 2019; Qiao et al., 2020). GCN over macromolecule graphs with one-hot attributes has been shown to outperform fingerprint-based models (Zeng et al., 2018). In a similar vein, algorithms for graph-representation learning have also been ported to periodic crystals (Xie & Grossman, 2018). While fingerprint-based models are limited by representation capacity, the GCN model, not having been trained on chemical information, cannot extrapolate for unknown monomers.

3 MACROMOLECULE GRAPH REPRESENTATION

We used a generalized text file format to convert a macromolecule structure into machine-readable format (Figure 1A, Appendix C). The text file has 3 sections – SMILES, MONOMERS and BONDS. Under SMILES, monomer and bond codes followed by the stereochemical SMILES are noted (Figure A.5). MONOMERS enumerate an index of all nodes numbered from 1 to n, where n is the total number of monomers, followed by the monomer code. BONDS enumerate an index of connections between monomer indices, followed by bond code.

The macromolecule is represented as a graph, with monomers as nodes and bonds as edges (Figure A.6). Starting from text files, we parsed the macromolecules into NetworkX graphs with node and edge attributes (Hagberg et al., 2008). The monomer and bond molecules were featurized using stereochemical extended connectivity fingerprints (Appendix D) (Rogers & Hahn, 2010). The fingerprint is a unique barcode that captures inherent chemistry of the monomer/bond molecule, by topological exploration of the molecular graph (different from the macromolecule graph). This rep-

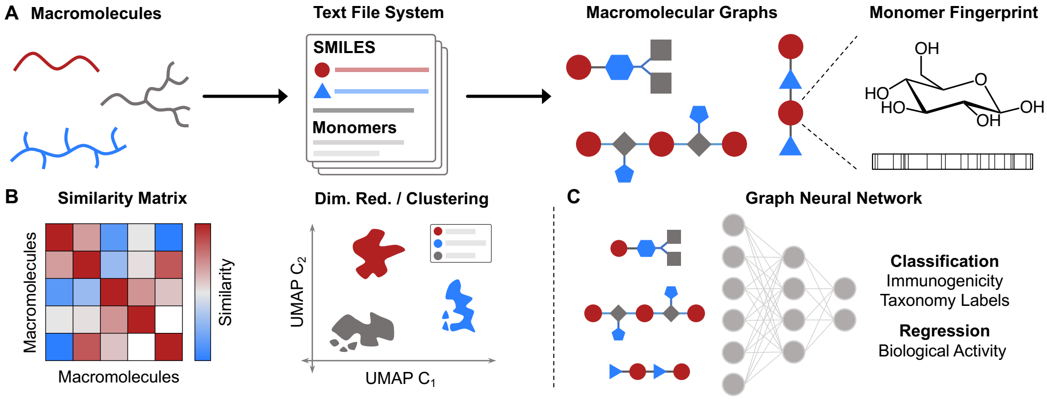


Figure 1: **Macromolecules represented as graphs, enable similarity computation and better machine learning.** **A.** Macromolecular structures are converted into a text file. The text files are parsed into NetworkX graphs. The text file enumerates SMILES for monomers and bonds, node indices corresponding to monomers, and pairs of node indices for bonds. Node and edge attributes are fingerprints of respective molecules. **B.** Pair-wise similarity matrix is obtained for the library of macromolecules. Dimensionality reduction followed by clustering of similarity vectors shows chemically similar regions. **C.** GNNs learn over macromolecule graphs for a variety of tasks.

resentation enables the depiction of macromolecules in their native state with explicit featurization of the stereochemistry and topology, and provides a single framework to represent both natural and synthetic, linear and non-linear macromolecules.

4 Similarity computation using graph edit distance and graph kernel

Leveraging this unique representation, we used exact GED scored with Tanimoto similarity substitution matrices, and graph kernel, to compute similarity between 2 or more macromolecule graphs (Figure 1B). GED computes the similarity between two graphs by assigning scores for node and edge substitution, similar to local sequence alignment for protein and DNA/RNA sequences (Abu-Aisheh et al., 2015; Altschul et al., 1990). Instead of evolutionary statistics-based substitution matrices, we use Tanimoto similarity matrices that compute the similarity between molecular fingerprints (Figure 2A, B). Tanimoto similarity is also applicable for unnatural monomers and provides an accurate measure of chemical similarity, without any evolutionary bias. Since, computing GEDs is costly, we use propagation attribute kernel to obtain similarity matrices for large data sets (Neumann et al., 2016; Siglidis et al., 2020).

We computed similarity matrices and analyzed similarity vectors for a data set of glycans. Propagation attribute kernel, implemented in GraKeL, was used to compute similarity (Figure 2C) (Siglidis et al., 2020) (Appendix E). This kernel makes for an excellent choice for macromolecule graphs as they capture local node information and iteratively propagate this information along the edges. In this manner, the kernel captures the local monomer chemistry and the global topology of the macromolecule. 2-components uniform manifold approximation and projection (UMAP) was used for dimensionality reduction of the similarity vectors, where a similarity vector is the vector listing the similarity of a single macromolecule with the entire library (McInnes et al., 2018). We optimized hyperparameters for UMAP and benchmarked against dimensionality reduction using t-stochastic neighbor embeddings (Appendix F) (Van Der Maaten & Hinton, 2008).

Dimensionality reduction is influenced more by taxonomic classification, such as domain, than immunogenicity. In the plot, colored by domain (Figure 2D), we observed that the arrangement of domains is similar to the evolutionary process, starting from bacteria at the center, then eukarya, followed by viruses at the fringes (Figure A.11). As can be seen, immunogenicity is a result of the glycan belonging to a specific domain, such as bacteria being immunogenic (Figures 2E, F).

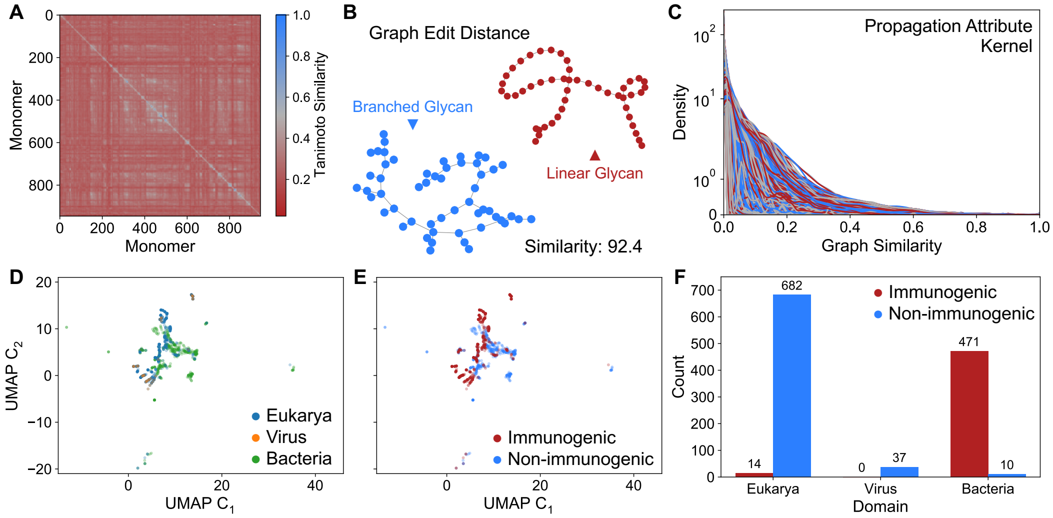


Figure 2: **Glycans have a broad range of chemical similarity. A.** The monomers in glycans are chemically dissimilar, as depicted by low Tanimoto similarity. **B.** Similarity between two glycans was computed using exact graph edit distance, scored with Tanimoto substitution matrices. **C.** Overlay of histograms of similarity vectors for 8899 glycans in the curated data set. The vectors were normalized to the maximum in the respective vector. 2-component UMAP, colored by **D.** domains, and **E.** immunogenicity, for 1313 glycans with immunogenicity labels. **F.** Immunogenic glycans usually belong to bacteria, while non-immunogenic glycans are from eukarya and virus domains.

5 Graph Neural Networks for Classification of Glycans

We trained 5 GNN model architectures over fingerprint and one-hot node and edge attributes to classify glycans for immunogenicity and 8 taxonomic levels (Appendix G). For each task, we evaluated classification metrics obtained by averaging over models retrained for at least top 5 hyperparameter sets with 5 random initialization seeds (Table 1; Tables A.1, A.2). Our models obtained state-of-the-art results and outperformed metrics reported in the literature (Table A.3) (Bojar et al., 2021).

6 Discussion and Future Work

Macromolecule graph representation combined with molecular fingerprints, with graph similarity and GNNs provides for a framework to represent, compute similarity and machine learn macromolecules. This work enables a chemistry-informed approach for the computational study of macromolecules. In the near future, we aim to demonstrate the applicability of our framework on a wide range of macromolecule datasets, including proteins and DNA/RNA.

Table 1: Metrics obtained for most optimal model-attribute combination on test data set.

Task	Model, Attribute	ROC-AUC	F1	Recall	Precision	Accuracy	CE Loss
Immunogenicity	GCN, FP	0.99	0.95	0.95	0.95	0.95	0.11
Domain	Attentive FP, FP	0.99	0.94	0.94	0.94	0.93	0.09
Kingdom	Attentive FP, FP	0.99	0.91	0.89	0.93	0.89	0.06
Phylum	GCN, FP	0.99	0.84	0.80	0.89	0.80	0.03
Class	GCN, FP	0.99	0.74	0.67	0.84	0.67	0.02
Order	GCN, FP	0.98	0.64	0.54	0.78	0.55	0.02
Family	GAT, FP	0.98	0.56	0.46	0.72	0.47	0.01
Genus	GCN, One-hot	0.96	0.51	0.40	0.72	0.41	0.01
Species	GCN, FP	0.97	0.49	0.38	0.68	0.40	0.01

References

Zeina Abu-Aisheh, Romain Raveaux, Jean-Yves Ramel, and Patrick Martineau. An Exact Graph Edit Distance Algorithm for Solving Pattern Recognition Problems. *4th International Conference on Pattern Recognition Applications and Methods*, 2015.

Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

Gabriela Gil Alvaradejo, Hung V.T. Nguyen, Peter Harvey, Nolan M. Gallagher, Dao Le, M. Francesca Ottaviani, Alan Jasanoff, Guillaume Delaittre, and Jeremiah A. Johnson. Polyoxazoline-Based Bottlebrush and Brush-Arm Star Polymers via ROMP: Syntheses and Applications as Organic Radical Contrast Agents. *ACS Macro Letters*, 8(4):473–478, 2019.

Maxwell L. Bileschi, David Belanger, Drew Bryant, Theo Sanderson, Brandon Carter, D. Sculley, Mark A. DePristo, and Lucy J. Colwell. Using Deep Learning to Annotate the Protein Universe. *bioRxiv:10.1101/626507*, 2019.

David B Blumenthal and Johann Gamper. On the exact computation of the graph edit distance. *Pattern Recognition Letters*, 134:46–57, 2020.

Daniel Bojar, Rani K. Powers, Diogo M. Camacho, and James J. Collins. Deep-Learning Resources for Studying Glycan-Mediated Host-Microbe Interactions. *Cell Host and Microbe*, 29(1):132–144.e3, 2021.

Karsten Borgwardt, Elisabetta Ghisu, Felipe Llinares-lópez, Leslie O Bray, and Bastian Rieck. Graph Kernels. *arXiv:2011.03854v2*, 2020.

Andrew J Boydston, Jianxun Cui, Chang-Uk Lee, Brock E Lynde, and Cody A Schilling. 100th Anniversary of Macromolecular Science Viewpoint: Integrating Chemistry and Engineering to Enable Additive Manufacturing with High-Performance Polymers. *ACS Macro Letters*, 9(8): 1119–1129, 2020.

Lachlan Coff, Jeffrey Chan, Paul A Ramsland, and Andrew J Guy. Identifying glycan motifs using a novel subtree mining approach. *BMC bioinformatics*, 21(1):42, 2020.

Sean R Eddy. Where did the BLOSUM62 alignment score matrix come from? *Nature biotechnology*, 22(8):1035, 2004.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural Message Passing for Quantum Chemistry. *arXiv:1704.01212v2*, 2017.

Aric A Hagberg, Daniel A Schult, and Pieter J Swart. Exploring Network Structure, Dynamics, and Function using NetworkX. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman (eds.), *Proceedings of the 7th Python in Science Conference*, pp. 11–15, Pasadena, CA USA, 2008.

Paul C Hiemenz and Timothy P Lodge. *Polymer chemistry*. CRC press, 2007. ISBN 1420018272.

Masae Hosoda, Yukie Akune, and Kiyoko F. Aoki-Kinoshita. Development and application of an algorithm to compute weighted multiple glycan alignments. *Bioinformatics*, 33(9):1317–1323, 2017.

Tommi Jaakkola, Mark Diekhans, and David Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7(1-2):95–114, 2000.

Jeremiah A Johnson, Ying Y Lu, Alan O Burts, Yeon-Hee Lim, M G Finn, Jeffrey T Koberstein, Nicholas J Turro, David A Tirrell, and Robert H Grubbs. Core-clickable PEG-branch-azide bivalent-bottle-brush polymers by ROMP: grafting-through and clicking-to. *Journal of the American Chemical Society*, 133(3):559–566, 2011.

Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick F. Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8): 595–608, 2016.

Chiho Kim, Anand Chandrasekaran, Tran Doan Huan, Deya Das, and Rampi Ramprasad. Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. *Journal of Physical Chemistry C*, 122(31):17575–17585, 2018.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2019.

Greg Landrum. RDKit: Open-source cheminformatics, 2006.

Limin Li, Wai Ki Ching, Takako Yamaguchi, and Kiyoko F. Aoki-Kinoshita. A weighted q-gram method for glycan structure classification. *BMC Bioinformatics*, 11:1–6, 2010.

Tzyy-Shyang Lin, Connor W Coley, Hidenobu Mochigase, Haley K Beech, Wencong Wang, Zi Wang, Eliot Woods, Stephen L Craig, Jeremiah A Johnson, and Julia A Kalow. BigSMILES: a structurally-based line notation for describing macromolecules. *ACS central science*, 5(9):1523–1531, 2019.

Jean-Francois Lutz, Jean-Marie Lehn, E W Meijer, and Krzysztof Matyjaszewski. From precision polymers to complex materials and systems. *Nature Reviews Materials*, 1(5):1–14, 2016.

Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.

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

Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

Marion Neumann, Roman Garnett, Christian Bauckhage, and Kristian Kersting. Propagation kernels: efficient graph kernels from propagated information. *Machine Learning*, 102(2):209–245, 2016.

Bo Qiao, Somesh Mohapatra, Jeffrey Lopez, Graham M. Leverick, Ryoichi Tatara, Yoshiki Shibuya, Yivan Jiang, Arthur France-Lanord, Jeffrey C. Grossman, Rafael Gómez-Bombarelli, Jeremiah A. Johnson, and Yang Shao-Horn. Quantitative Mapping of Molecular Substituents to Macroscopic Properties Enables Predictive Design of Oligoethylene Glycol-Based Lithium Electrolytes. *ACS Central Science*, 6(7):1115–1128, 2020.

Kaspar Riesen and Horst Bunke. Approximate graph edit distance computation by means of bipartite graph matching. *Image and Vision Computing*, 27(7):950–959, 2009.

David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.*, 50(5): 742–754, 2010.

Matteo Romio, Lucca Trachsel, Giulia Morgese, Shivaprakash N. Ramakrishna, Nicholas D. Spencer, and Edmondo M. Benetti. Topological Polymer Chemistry Enters Materials Science: Expanding the Applicability of Cyclic Polymers. *ACS Macro Letters*, 9(7):1024–1033, 2020.

Benjamin Sanchez-lengeling, Jennifer Wei, Brian Lee, Emily Reif, Peter Y Wang, Wei Qian, Kevin Mccloskey, Lucy Colwell, and Alexander Wiltschko. Evaluating Attribution for Graph Neural Networks. *Proceeedings of the Neural Information Processing Systems Conference*, (NeurIPS), 2020.

Kristof T. Schütt, P.-J J. Pieter-Jan P.-J Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert R. K.-R Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in Neural Information Processing Systems*, 2017-Decem(1):992–1002, jun 2017. URL http://arxiv.org/abs/1706.08566www.quantum-machine.org.

Giannis Siglidis, Giannis Nikolentzos, Stratis Limnios, Christos Giatsidis, and Michalis Vazirgiannis. GraKeL : A Graph Kernel Library in Python. *Journal of Machine Learning Research*, 21: 1–5, 2020.

Temple F Smith and Michael S Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

Peter C. St John, Caleb Phillips, Travis W. Kemper, A. Nolan Wilson, Yanfei Guan, Michael F. Crowley, Mark R. Nimlos, and Ross E. Larsen. Message-passing neural networks for high-throughput polymer screening. *Journal of Chemical Physics*, 150(23), 2019.

Mahito Sugiyama, M. Elisabetta Ghisu, Felipe Llinares-López, and Karsten Borgwardt. Graphkernels: R and Python packages for graph comparison. *Bioinformatics*, 34(3):530–532, 2018.

Chase B Thompson and LaShanda T J Korley. 100th Anniversary of Macromolecular Science Viewpoint: Engineering Supramolecular Materials for Responsive Applications—Design and Functionality. *ACS Macro Letters*, 9(9):1198–1216, 2020.

Laurens Van Der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph Attention Networks. *arXiv:1710.10903*, 2017.

Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks. *arXiv:1909.01315*, 2019.

Tian Xie and Jeffrey C. Grossman. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters*, 120(14):145301, apr 2018. doi: 10.1103/PhysRevLett.120.145301.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*, 63(16):8749–8760, 2020.

Minggang Zeng, Jatin Nitin Kumar, Zeng Zeng, Ramasamy Savitha, Vijay Ramaseshan Chandrasekhar, Kedar Hippalgaonkar, and Electronic Materials. Graph Convolutional Neural Networks for Polymers. *arXiv:1811.06231v1*, 2018.

Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph Neural Networks: A Review of Methods and Applications. *arXiv:1812.08434v4*, 2018.

APPENDIX

A MACROMOLECULAR TOPOLOGIES

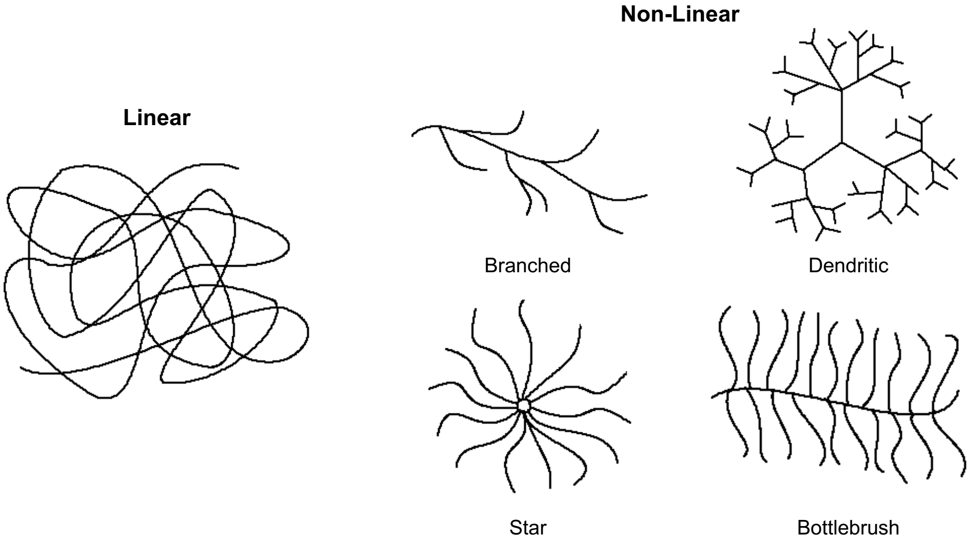


Figure A.1: Macromolecules exist in a diverse array of topologies, including both linear and non-linear, such as branched, dendritic, star, and bottlebrush, architectures.

B GLYCANS DATASET PROCESSING

B.1 DATASET DOWNLOAD

A dataset of 19299 glycans was accessed and downloaded from GlycoBase (accessed on November 2, 2020) (Bojar et al., 2021). The file contained GlycoBase ID, sequence, link (N, O, free, or none), species, and immunogenicity information for each glycan.

For each glycan sequence string the brackets denote branches, with the point of attachment/bonding of the branch as the monomer immediately after the brackets. The first element within the bracket is the monomer most distant from the point of attachment, and the last element within the bracket is the abbreviation of the bond that connects the branch to the original main chain. Nested brackets indicate additional sub-branches off of branches, and multiple sets of brackets next to each other indicate several branches off of the same monomer.

B.2 DATASET PRE-PROCESSING

7 modifications and 152 deletions of glycan sequences were made before the process for situations such as an unequal number of opening and closing brackets and dangling branches without specified connectivity. Additional glycan sequences were removed due to missing SMILES sequences for a number of monomers. The original GlycoBase.csv file was curated to reflect the modification and deletion changes, resulting in a total of 19147 glycans.

Using the curated database, we visualized the distribution of species of origin, link, and immunogenicity of the glycans (Figures A.2, A.3, A.4).

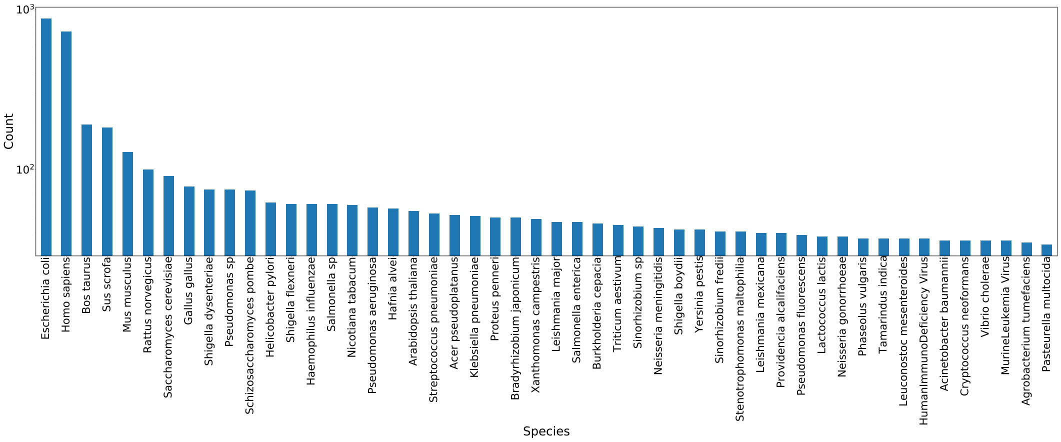


Figure A.2: Top 50 most common glycan species of origin, sorted in descending order of count with y-axis on a logarithmic scale.

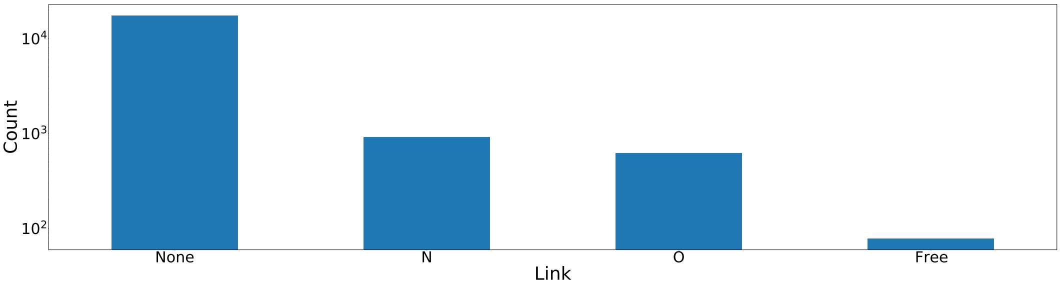


Figure A.3: Four types of glycan links sorted in descending order of count with y-axis on a logarithmic scale.

B.3 SMILES compilation

The chemical composition and formula of the 959 monomers (also known as glycoletters in GlycoBase) and 53 bonds were expressed as isomeric simplified molecular-input line-entry system (SMILES) sequences.

In the dataset retrieved from GlycoBase, all position-specific information about monomer modifications was removed. The supplementary information of SweetTalk contains all the raw glycan sequences before position-specific modification information was removed (Bojar et al., 2021). While insufficient information is provided to directly match the raw sequences with the corresponding edited sequences, the raw sequences can be used to look for trends and most common modification positions for each monomer. A dictionary was created to describe the number of times each monomer appears in all the raw sequences. A couple of terms should be defined for consistency:

- “Monosaccharides” refer to the individual glycoletters without any modifications, such as D-glucose (D-Glu) and galactose (D-Gal). The term “monosaccharide” is simplistic because it also covers alcohols, acids, and other classes of molecules, but the term will be employed for the sake of consistency and clarity.
- “Modifications” refer to substitutions or additions to the original monosaccharide sequence such as a nitrogen-linked acetyl group (NAc) and oxygen-linked methyl group (OMe). In this dataset, monosaccharides can have between 0-4 modifications.
- “Monomers” refer to the combination of the monosaccharide and modification(s), also known as glycoletters in GlycoBase.

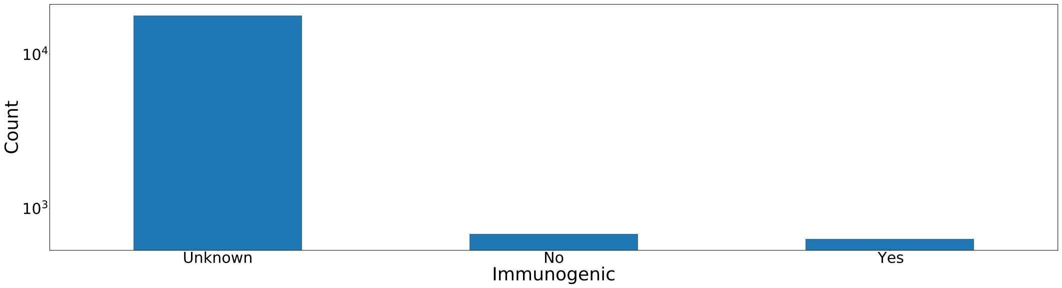


Figure A.4: Three different types of glycan immunogenicity labels sorted in descending order of count, with y-axis on a logarithmic scale.

Using the raw monomers dictionary, the positions of the modifications for each monomer in the corresponding SMILES were set using a list of consistent rules outlined below. The rules are listed in order of priority.

1. For each monomer/glycoletter, if at least one monomer with the same monosaccharide and set of modifications exists in the raw monomers dictionary, assume the set of positions with the highest frequency. If no monomer exists in the raw monomers dictionary with the same monosaccharide and set of modifications, proceed to the following steps.
2. If the monomer has a hydroxyl group at the 2 position, is not a furanose or ketose, and contains a nitrogen-linked modification (N, NBut, NMe, etc.), assign the first occurrence of a nitrogen-linked modification to the 2 position.
3. If the monomer is a hexose or deoxy-hexose and contains a pyruvate acetal (OPyr), assume that the acetal connects the 4 and 6 positions.
4. If the monomer contains an O-linked phosphate (OP) or sulfate (OS) modification, search the raw monomers dictionary for instances of the monosaccharide with only the OP/OS modification and assume the most frequent position.
5. If the monomer contains a variant of the O-linked phosphate modification (OPEtn, OPPEtn, etc.), assign the OP-variant modification the same position as the most common position for the OP modification on the monosaccharide.
6. If the monomer contains two modifications, search the raw glycans dictionary for instances of the monosaccharide with only each modification separately but not at the same time. Assign each modification the position of highest frequency. If either monosaccharide and modification combination does not exist in the raw monomers dictionary and the monomer is a hexose, assign positions in the following order: 2, 4, 3, 6. If the position is already occupied from previous steps or does not exist in a deoxy hexose, skip to the position with next highest priority.
7. If the monomer is a hexose or deoxy-hexose and contains more than two modifications, assign positions in the following order: 2, 3, 4, 6. If the position is already occupied from previous steps or does not exist in a deoxy hexose, skip to the position with next highest priority.
8. Assume all amino acids are connected to the monosaccharide via the oxygen on the carboxyl. If the connectivity is not specified for a group following the amino acid (CysAc, AlaFo), assume that the group following the amino acid is connected to the amino acid via the amine group.
9. For neuraminic acid (Neu), ketodeoxynononic acid (Kdn), pseudaminic acid (Pse), legionaminic acid (Leg), and other similar ketose-based acids, assign modification positions in the following order: 1, 4.
10. For fructofuranose and similar furanoses for which the 1-position is not part of the ring, assign modifications in the following order: 1, 3, 4.

11. For alcohols, follow the same rules that apply to the oxidized form of the alcohol. For example, for glucitol follow the same rules that apply to glucose.

12. Assume all rare monosaccharides (denoted as Sug) are hexoses with no specified stereochemistry.

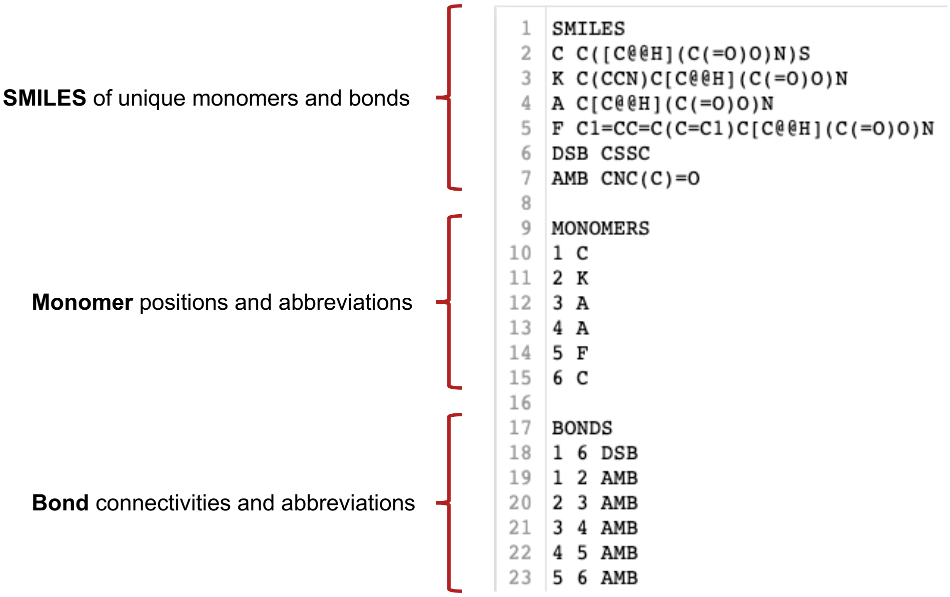
Because the 959 monomers contained some repeat SMILES sequences with different monomer names, 13 redundant names were deleted so that each distinct monomer SMILES only appears once in the SMILES compilation.

The SMILES for the 53 bond types differ in stereochemistry alone (alpha, beta, or unspecified) but all have the same chemical composition of a glycosidic bond. Each bond is expressed as a variation of the SMILES sequence CC(OC)CC, which consists of the glycosidic bond C-O-C with one of the C atoms also connected to both a methyl and ethyl group. The chiral C with the four different attached groups is used to specify the alpha or beta stereochemistry displayed in the bond name. The stereochemistry at the tetrahedral C is consistently S for all alpha bonds, R for all beta bonds, and not specified for all unspecified bonds. The 53 bond names were condensed into 3 distinct bond types differing in stereochemistry alone.

C Text file system

C.1 Format

The text files to convert each macromolecule structure into machine-readable format consist of three sections: the SMILES sequences for each unique monomer or bond, the positions of each monomer, and the two monomer positions connected by each bond. Each section starts with a header to indicate the start of a new part. The first section contains the abbreviation of each unique monomer or bond in the glycan followed by the corresponding SMILES sequence, with each entry on a separate line. The monomers section consists of the monomer position followed by the monomer abbreviation. The bonds section contains the two connectivity positions in the glycan for the bond separated by a space, followed by the bond abbreviation (Figure A.5).


SMILES of unique monomers and bonds
Monomer positions and abbreviations
Bond connectivities and abbreviations

```
SMILES
C C([C@@H](C(=O)O)N)S
K C(CCN)C[C@@H](C(=O)O)N
A C[C@@H](C(=O)O)N
F C1=CC=C(C=C1)C[C@@H](C(=O)O)N
DSB CSSC
AMB CNC(C)=O

MONOMERS
1 C
2 K
3 A
4 A
5 F
6 C

BONDS
1 6 DSB
1 2 AMB
2 3 AMB
3 4 AMB
4 5 AMB
5 6 AMB
```

Figure A.5: Standard text file format with three sections: SMILES of unique monomers and bonds, monomer positions and abbreviations, and bond connectivities and abbreviations.

C.2 Text File Parser

A text file parser converts the macromolecule information stored in the .txt file to a NetworkX graph with monomers expressed as nodes and bonds expressed as edges. The parser goes through the .txt file line by line, stores the monomer information in a dictionary with keys as integer positions and values as monomer abbreviations, and stores the bond information in a dictionary with keys as tuples containing bond connectivities and values as bond abbreviations. Afterwards, the reader uses NetworkX to add each key in the monomer dictionary as a node and each key in the bond dictionary as an edge, storing the abbreviations as attributes for the corresponding node or edge. The resulting NetworkX graphs include both linear and highly branched architectures (Figure A.6).

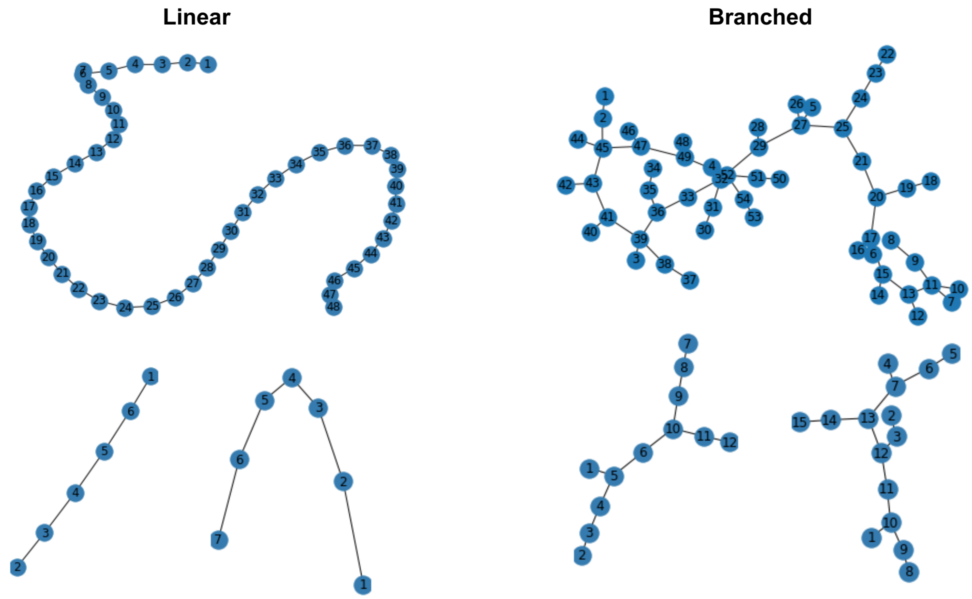


Figure A.6: NetworkX graph representations of glycans that consist of both linear and highly branched architectures of varying molecular weights and complexities.

D Node and edge attributes

D.1 Fingerprints – generation and optimization of hyperparameters

We used RDKit to generate stereochemical extended connectivity fingerprints (Landrum, 2006; Rogers & Hahn, 2010). Radius and number of bits were optimized by calculating mean and standard deviation, and visualizing the distribution of Tanimoto similarity of all monomers in the glycans dataset (Figure A.7A-C). We aimed to obtain fingerprints with lower number of bits and optimal radius that could represent the monomer aptly in both similarity computation and graph neural network models. For 64 bits, we observed that the mean similarity was as high as 0.4 and standard deviation for similarity went up after radius 3 indicating higher hash collision and lesser differentiability. For 128 bits, the mean and standard deviation plateaued around 0.3 and 0.08, respectively. Further, the similarity distribution was qualitatively spread over a larger range, as compared to fingerprints with larger number of bits. For fingerprint bits longer than 128, we observed lower mean and standard deviation, and decreasing spread in the similarity distribution. The decreasing trend indicates that the fingerprints with bits higher than 128 are equally dissimilar, thus, if used, will lead to glycans with equally high dissimilar scores. Hence, we chose fingerprints with 128 bits and radius 3 to generate node attributes for glycans. For edge attributes, we had 3 types of glycosidic bonds differing by the stereochemistry alone, hence, we chose fingerprints with 16 bits and radius 3.

We observed that there were 3 sets with 7 glycans which had exact fingerprints (Figure A.7D). Since, the difference was in the number of carbon atoms in the aliphatic chain, we used the fingerprints as is.

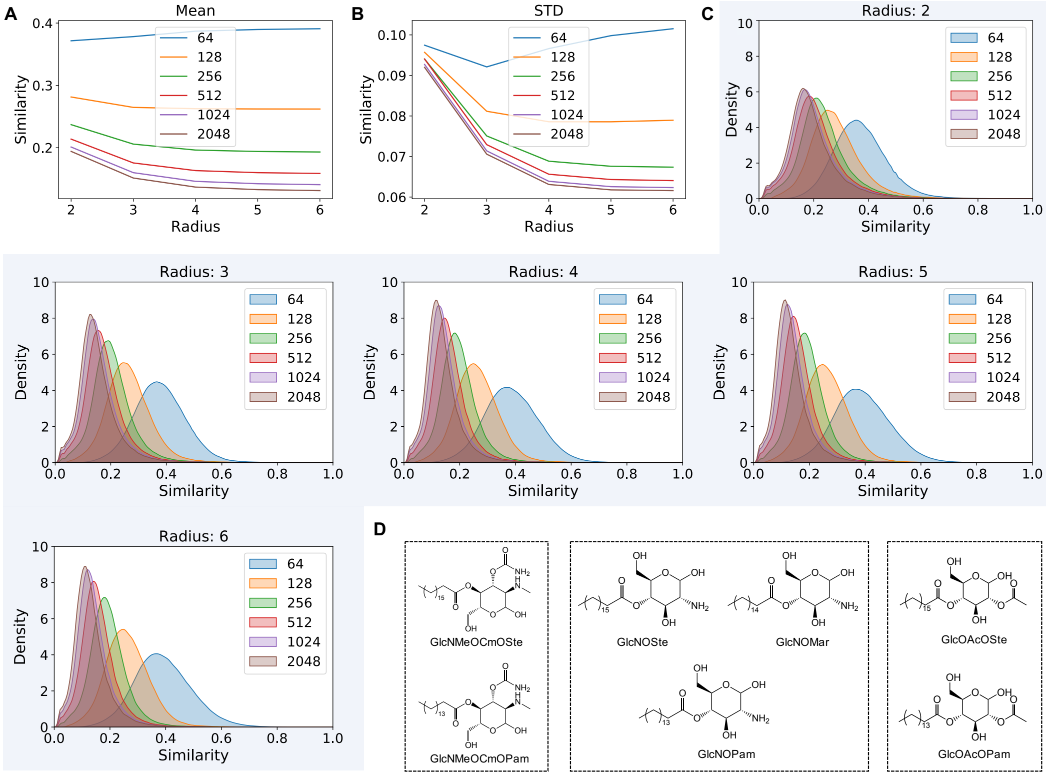


Figure A.7: **A.** Mean, **B.** standard deviation and **C.** distribution of Tanimoto similarity of all monomers in the glycans dataset, calculated using stereochemical extended connectivity fingerprints of different radii and bits. **D.** Glycans in the same box have the exact fingerprint for radius 3 and 128 bits.

D.2 One-hot encoding benchmark

One-hot encodings of the 946 monomer and 3 bond types were also employed as feature types for benchmarking with featurization using molecular fingerprints. The dimensions of the node and edge one-hot encoding features are 946 and 3, respectively.

E Similarity computation

E.1 Analysis of graphs

Most glycan graphs are sparse (Figure A.8). Complete graphs have density of 1, while graphs with density >0.5 have been defined as dense (Borgwardt et al., 2020). The density has been calculated using -

$$d = \frac{2m}{n(n-1)}$$

E.2 Similarity matrix

We computed the $(n \times n)$ similarity matrix for all glycans with labels on at least one taxonomic level using propagation attribute kernel in GraKeL (Figure A.8) (Siglidis et al., 2020). Each pair of

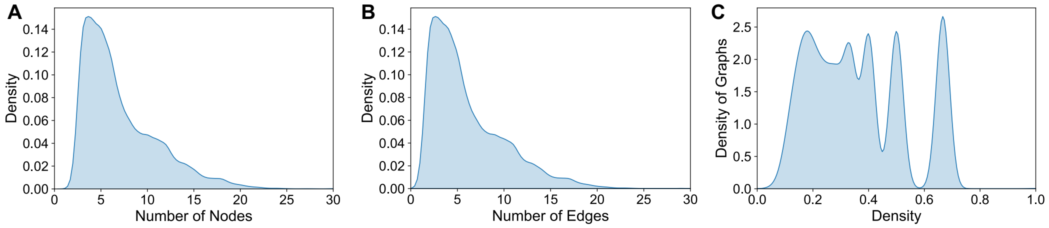


Figure A.8: Distribution of number of **A.** nodes, **B.** edges and **C.** density for glycan graphs.

graph similarity was computed for a maximum of 100 iterations. This resulted in 5% of the pairs being assigned a 0 similarity (10% of all indices in the similarity matrix are 0).

It may be noted that the computation of similarity using graph kernels is way more accessible than graph edit distances. The current computation was done on in minutes (wall time), parallelized across 24 cores. From visual inspection using htop, only about 30% of 2 cores were being used at any particular time, and less than 5% of all other cores were used, although there were 24 jobs running in parallel.

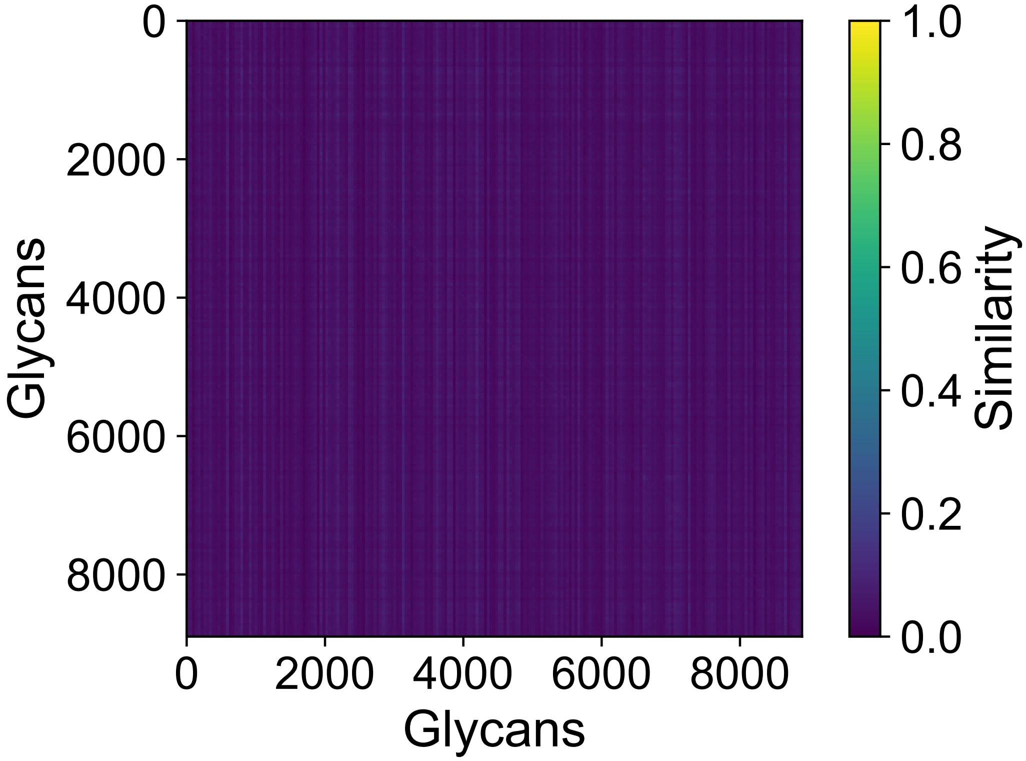


Figure A.9: 2D plot for the $(n \times n)$ similarity matrix of glycans. This is not symmetric because each row is normalized by its maximum.

F Dimensionality reduction

F.1 Hyperparameter optimization for Uniform Manifold Approximation and Projection (UMAP)

Number of neighbors was optimized for 2-component UMAP dimensionality reduction of similarity vectors (Figure A.10) (McInnes et al., 2018). From visual inspection, UMAP with 128 neighbors seems to resolve into optimal size and number of clusters. The subplot shows distinct regions for the immunogenic and non-immunogenic glycans. We note that there is more to similarity than graph kernel distances, and observe that it has been effectively captured using the GNN models (Appendix G).

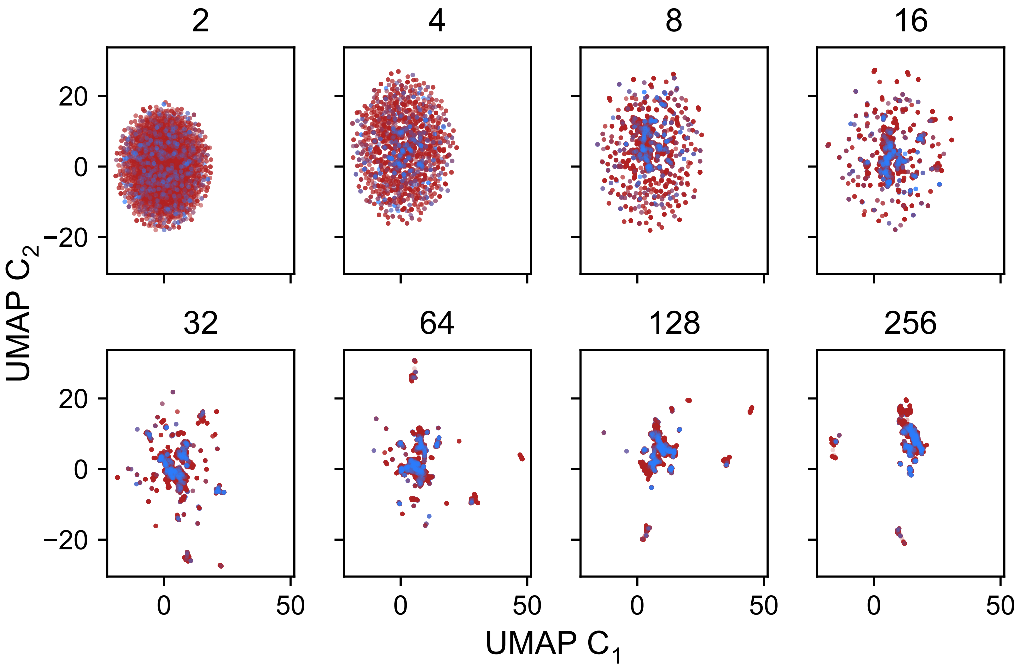


Figure A.10: Visualization of scatter plots for UMAP components, obtained by dimensionality reduction of similarity vectors. The number of neighbors for each UMAP computation has been noted in the title of the respective sub-plot. Coloration is by immunogenicity (red: immunogenic, blue: non-immunogenic glycans).

UMAP dimensionality reduction for higher number of components does not provide more information. The 3 components UMAP looked similar to the 2 components, with slightly more disentangled families (Figure A.11). We limited our visual analysis to 3 components. To check if more components can help in finding distinct clusters, we did dimensionality reduction for $\{2, 3, 5, 10, 30, 50\}$ components and let HDBSCAN - an unsupervised clustering algorithm – to figure out how many clusters are there (McInnes et al., 2017). We noted that the number of clusters are pretty similar, and in low 400s (Figure A.12). The high number of clusters indicates the diversity of the space, and the differences in terms of taxonomy. As a further check to see if the distribution of glycans in different clusters is different, we plotted the histograms of the glycans assigned to each cluster. Across all the components, the histograms seem to be consistent with the number of glycans in each of them (Figure A.13).

F.2 T-SNE benchmark

We benchmarked the dimensionality reduction results obtained from UMAP against a broad range of t-stochastic neighbor embeddings (t-SNE) models (Van Der Maaten & Hinton, 2008). For the differ-

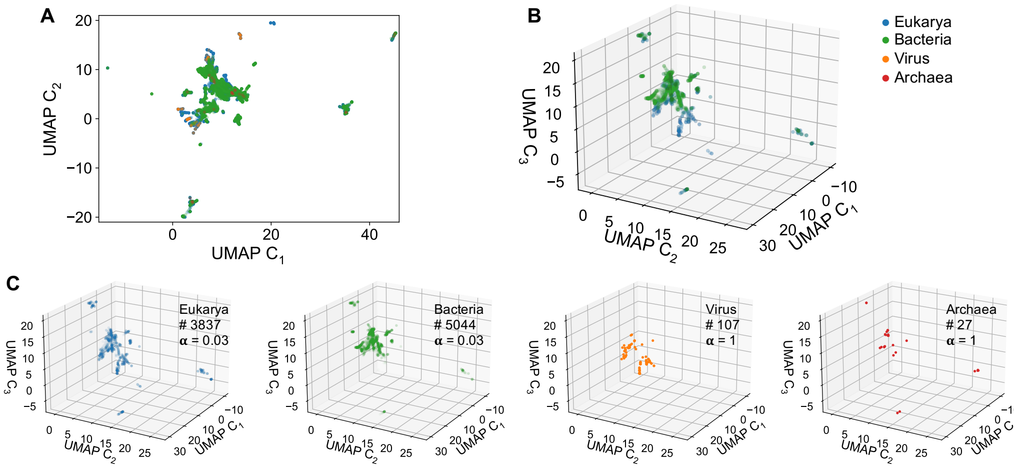


Figure A.11: Visualization of **A.** 2 and **B.** 3 components UMAP, colored by domain. **C.** Glycans corresponding to individual domains are shown, with the text noting the domain, number of glycans, and transparency (α) of each point on a scale of 0 to 1, where 1 is opaque

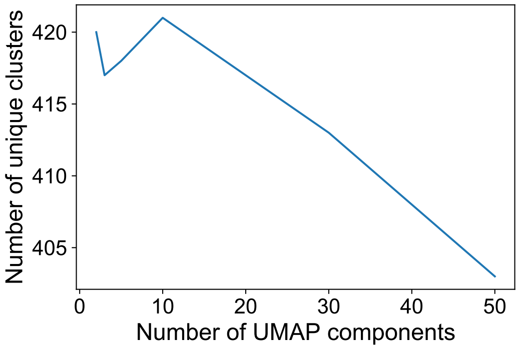


Figure A.12: Number of HDBSCAN unsupervised clusters obtained from UMAP dimensionality reduction for different components.

ent models, we varied perplexity as $\{2, 5, 30, 50, 100\}$, and number of steps as $\{500, 1000, 5000\}$. From the scatter plot, colored by immunogenicity labels, we noted that dimensionality reduction using t-SNE was not able to deduce the differences and getting the glycans into distinct areas (Figure A.14).

G Supervised learning with graph neural networks

G.1 Deep graphs library (DGL) graphs

Following featurization, NetworkX graphs were converted into undirected, unweighted, and homogenous DGL graphs (Wang et al., 2019). For GCN and GAT model architectures, self-loops were added to the DGL graphs to prevent silent performance regression due to zero-in-degree nodes during training.

G.2 Model architectures

We performed graph classification using five distinct graph neural network model architectures detailed below:

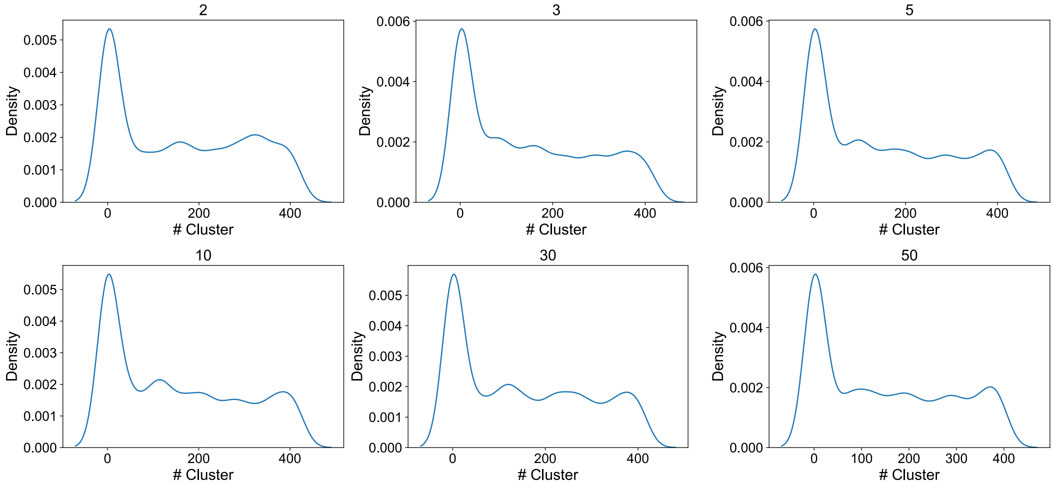


Figure A.13: Distribution of glycans in each cluster for HDBSCAN unsupervised clusters obtained from UMAP dimensionality reduction for different components. The components have been noted as titles for the sub-plots.

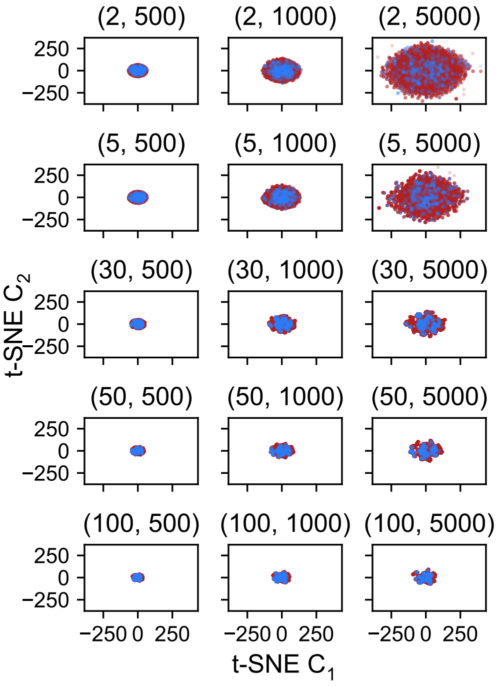


Figure A.14: Visualization of scatter plots for t-SNE components, obtained by dimensionality reduction of similarity vectors. The (perplexity, number of steps) for each t-SNE computation has been noted in the title of the respective sub-plot. Coloration is by immunogenicity (red: immunogenic, blue: non-immunogenic glycans).

- Weave (Kearnes et al., 2016)
- Message Passing Neural Networks (MPNNs) (Gilmer et al., 2017)
- Attentive FP (Xiong et al., 2020)
- Graph convolutional networks (GCN) (Kipf & Welling, 2019)

- Graph Attention Networks (GAT) (Velickovic et al., 2017)

While Weave, MPNN, and Attentive FP utilize both node and edge attributes in prediction, GCN and GAT only consider node attributes. The models were trained using implementations in the DGL LifeSci library (Wang et al., 2019). The optimization was done by minimization of average cross-entropy loss between batches and additional metrics such as F1 score, recall, precision and accuracy were noted.

G.3 Glycan graphs classification

G.3.1 Immunogenicity

Dataset. 1313 glycans in the database have immunogenicity labels, 631 of which are immunogenic and 682 of which are not immunogenic. The training was performed on 60%, validated on 20%, and tested on held-out 20% data.

Models. We classified immunogenicity using 5 model architectures combined with 2 different node and edge featurization types, for a total of 10 model architecture-attribute pairs. For each benchmark, hyperparameter optimization against minimization of binary cross entropy loss was performed on SigOpt, a standardized hyperparameter optimization platform, for 1000 observations and the 10 best sets of hyperparameters were extracted. Each model architecture and featurization combination was trained using the 10 best sets of hyperparameters from SigOpt using 10 distinct random seeds for splitting the dataset into train-validation-test datasets, for a total of 100 trainings per model per attribute type. The tables below report the metrics for the most optimal set of hyperparameters, the values of the most optimal hyperparameters, and mean metrics for each training across all 100 runs. All models achieve stellar performance on all metrics, with little meaningful difference in performance between fingerprint and one-hot encoding featurization (Figures A.15-A.17 ; Table A.1).

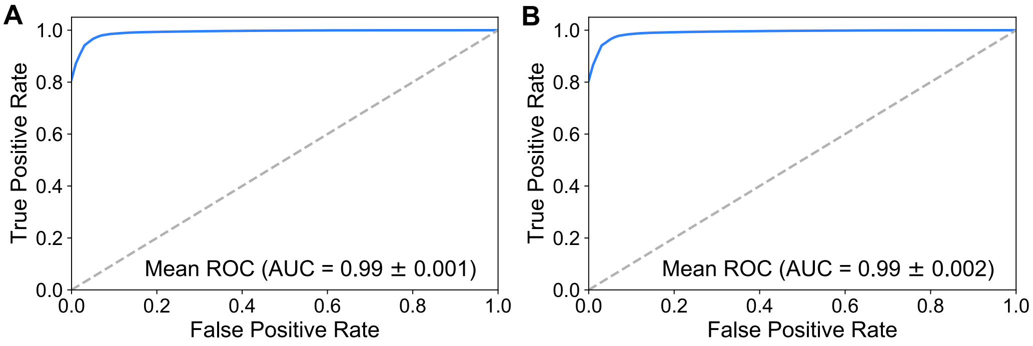


Figure A.15: ROC-AUC curves for fingerprint and one-hot encoding-featurized graphs. **A.** Mean ROC-AUC curve of all 50 fingerprint-featurized experiments (5 model architectures with 10 sets of hyperparameters for each architecture), with the standard deviation shaded in light blue too insignificant to be visible in the graph. **B.** Mean ROC-AUC curve of all 50 one-hot encoding-featurized experiments.

G.3.2 Taxonomy

Dataset. The taxonomy of the glycans was considered on eight levels: domain, kingdom, phylum, class, order, genus, and species. First, the classification of each species into the other seven taxonomic levels was obtained from the supplementary tables of SweetOrigins (Bojar et al., 2021). For each glycan with species information, taxonomic information was added in a new .csv file, with multiple labels on the same taxonomic level separated by commas. Afterwards, any labels with fewer than five glycans were removed, and any species names ending with “sp” that are therefore genus labels in disguise were filtered out to produce the final dataset for taxonomic training. The final

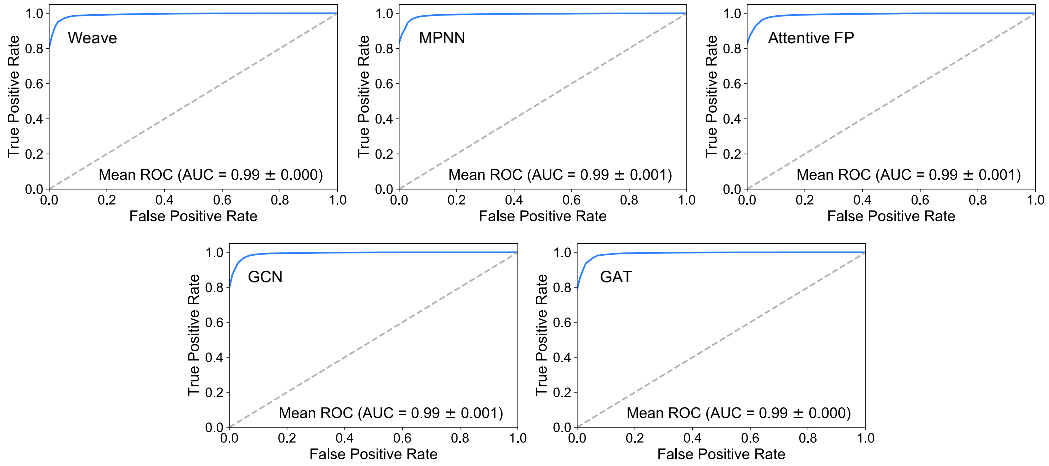


Figure A.16: Mean ROC-AUC curves for fingerprint-featurized experiments for each of the five model architectures (Weave, MPNN, Attentive FP, GCN, and GAT), with the standard deviation shaded in light blue too insignificant to be visible. A standard deviation of 0.000 denotes a value of <0.001.

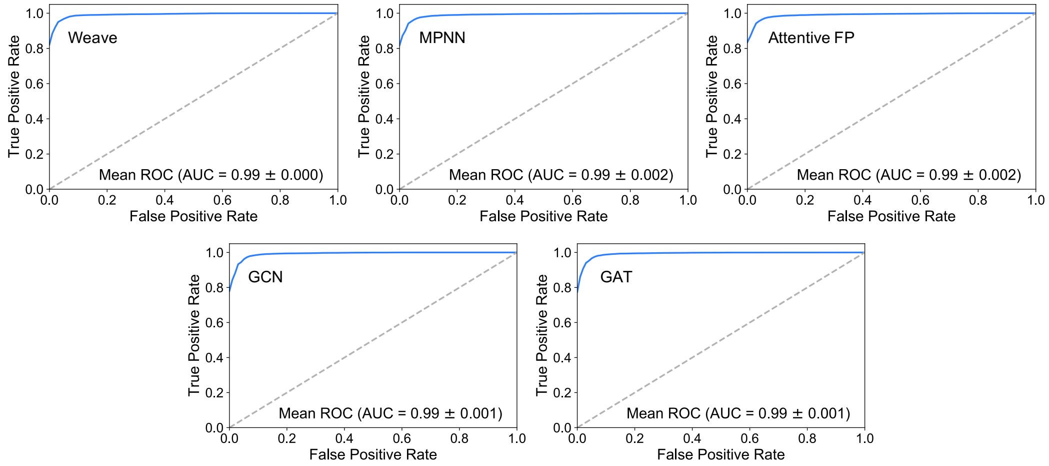


Figure A.17: Mean ROC-AUC curves for one-hot encoding-featurized experiments for each of the five model architectures (Weave, MPNN, Attentive FP, GCN, and GAT), with the standard deviation shaded in light blue too insignificant to be visible. A standard deviation of 0.000 denotes a value of <0.001.

dataset consists of 8899 unique glycans with labels on at least one taxonomic level encompassing 4 domains, 9 kingdoms, 33 phyla, 70 classes, 144 orders, 253 families, 400 genus, and 567 species. The training was performed on 60%, validated on 20%, and tested on held-out 20% data.

Graph labels were generated as one-hot encodings for each taxonomic level, with the length of each one-hot encoding tensor corresponding with the number of unique labels in the taxonomic level. The taxonomy labels accommodate cases in which glycans possess multiple labels within the same taxonomic level.

Models. Benchmarks of multi-label classification were performed using a similar process as with immunogenicity classification for 5 different model architectures, 2 different node and edge attribution types, and 8 different taxonomic levels for a total of 80 model architecture–attribute–taxonomic

Table A.1: Test dataset metrics for the most optimal set of hyperparameters, or the set of hyperparameters that results in the lowest loss. For each metric, the mean μ and standard deviation σ are displayed across all 10 random seeds. "FP" denotes condensed fingerprint featurization, and "One-hot" denotes one-hot encoding featurization.

Model	Attribute Type	ROC-AUC		F1		Recall		Precision		Accuracy		Loss	
		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Weave	FP	0.990	0.005	0.956	0.011	0.955	0.012	0.957	0.011	0.956	0.011	0.133	0.106
	One-hot	0.992	0.005	0.962	0.007	0.962	0.007	0.962	0.007	0.962	0.007	0.105	0.055
MPNN	FP	0.985	0.009	0.955	0.012	0.955	0.012	0.957	0.012	0.955	0.012	0.137	0.100
	One-hot	0.988	0.006	0.953	0.010	0.953	0.010	0.954	0.010	0.953	0.010	0.123	0.052
AttentiveFP	FP	0.990	0.005	0.954	0.011	0.953	0.011	0.955	0.012	0.954	0.011	0.125	0.104
	One-hot	0.990	0.005	0.951	0.011	0.951	0.012	0.952	0.011	0.952	0.011	0.120	0.084
GCN	FP	0.992	0.004	0.953	0.013	0.953	0.013	0.954	0.012	0.953	0.012	0.110	0.086
	One-hot	0.990	0.006	0.955	0.012	0.956	0.012	0.956	0.013	0.956	0.012	0.123	0.110
GAT	FP	0.991	0.006	0.954	0.014	0.954	0.013	0.954	0.014	0.954	0.014	0.109	0.081
	One-hot	0.991	0.004	0.954	0.011	0.954	0.011	0.955	0.010	0.955	0.011	0.119	0.085

Table A.2: The most optimal model architecture and node/edge attribute type that results in the lowest loss for prediction of all taxonomic levels. For each metric, the mean μ and standard deviation σ are displayed across all 5 random seeds. "FP" denotes condensed fingerprint featurization, and "One-hot" denotes one-hot encoding featurization.

Taxonomic Level	Model, Attribute Type	ROC-AUC		F1		Recall		Precision		Accuracy		Loss		Hamming Loss	
		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Domain	Attentive FP, FP	0.993	0	0.938	0.002	0.935	0.004	0.941	0.001	0.925	0.001	0.087	0.002	0.03	0.001
Kingdom	Attentive FP, FP	0.994	0	0.912	0.001	0.891	0.003	0.934	0.002	0.892	0.005	0.056	0.003	0.019	0
Phylum	GCN, FP	0.99	0	0.84	0.009	0.8	0.011	0.885	0.009	0.802	0.01	0.029	0.001	0.009	0
Class	GCN, FP	0.986	0.001	0.741	0.008	0.666	0.01	0.836	0.006	0.67	0.009	0.022	0	0.007	0
Order	GCN, FP	0.979	0.001	0.638	0.012	0.541	0.023	0.778	0.018	0.548	0.016	0.017	0	0.004	0
Family	GAT, FP	0.975	0.003	0.557	0.015	0.456	0.017	0.715	0.015	0.469	0.019	0.013	0	0.003	0
Genus	GCN, One-hot	0.963	0.003	0.513	0.015	0.397	0.015	0.723	0.020	0.412	0.010	0.009	0	0.002	0
Species	GCN, FP	0.968	0.003	0.487	0.021	0.382	0.031	0.675	0.018	0.395	0.020	0.007	0	0.002	0

level combination. For each combination, hyperparameter optimization against minimization of binary cross entropy loss was performed on SigOpt for 1000 observations and the 5 best sets of hyperparameters were extracted. Each model architecture, featurization, and taxonomic level combination was trained using the 5 best sets of hyperparameters from SigOpt using 5 distinct random seeds for splitting the dataset into train-validation-test datasets, for a total of 25 trainings per combination. The pipeline and models achieve state-of-the-art performance on multilabel taxonomic classification on all taxonomic levels (Figures A.18-A.25, Table A.2).

G.4 Benchmarking top GNN models against results reported in literature

For the top-performing models in SI Tables A.1 and A.2, hyperparameter optimization on SigOpt was performed again using the same method as the benchmarks in the work by Bojar et al. (2021) training on 80% of the dataset and reporting metrics on the remaining 20% used as a validation dataset (Table A.3). The experiment on SigOpt was optimized through minimization of the loss, and the remaining metrics (ROC-AUC, F1, recall, precision, and accuracy) were obtained as stored metrics in the SigOpt experiment.

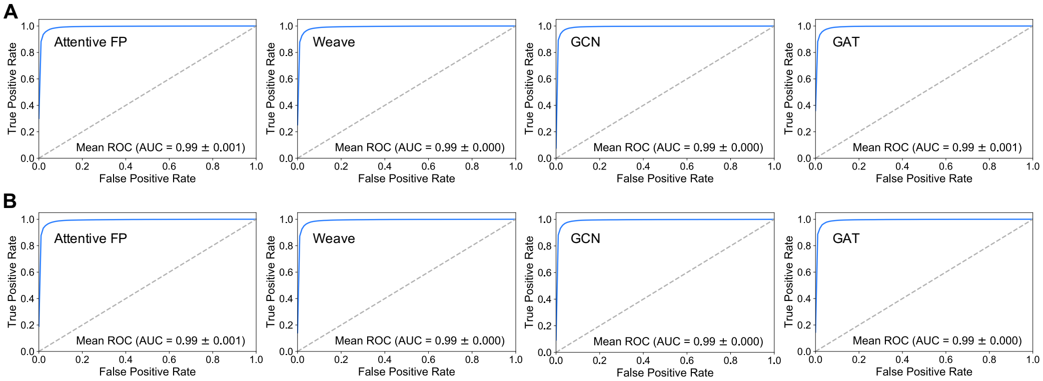


Figure A.18: ROC-AUC curves for classification on the domain level for each of four model architectures (Weave, Attentive FP, GCN, and GAT), with the standard deviation shaded in light blue too insignificant to be visible. A standard deviation of 0.000 denotes a value of <0.001. **A.** Mean ROC-AUC curves for fingerprint-featurized experiments, with each graph displaying the mean and standard deviation across 5 sets of hyperparameters. **B.** Mean ROC-AUC curves for one-hot encoding-featurized experiments.

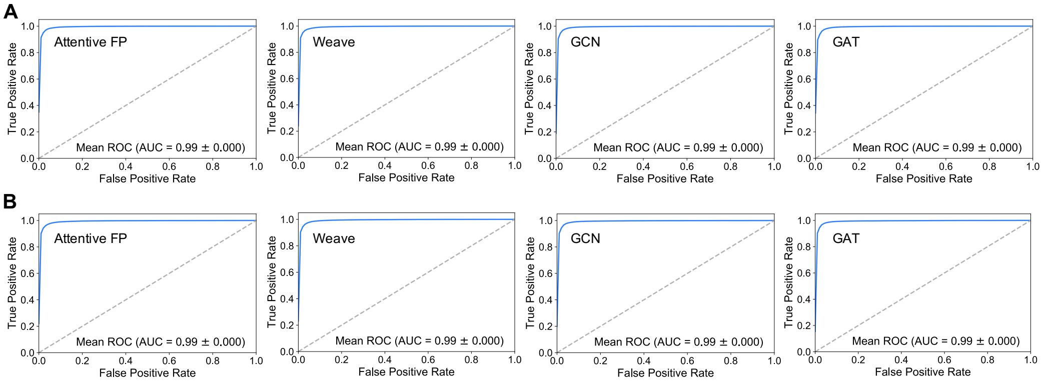


Figure A.19: ROC-AUC curves for classification on the kingdom level for each of four model architectures (Weave, Attentive FP, GCN, and GAT), with the standard deviation shaded in light blue too insignificant to be visible. A standard deviation of 0.000 denotes a value of <0.001. **A.** Mean ROC-AUC curves for fingerprint-featurized experiments, with each graph displaying the mean and standard deviation across 5 sets of hyperparameters. **B.** Mean ROC-AUC curves for one-hot encoding-featurized experiments.

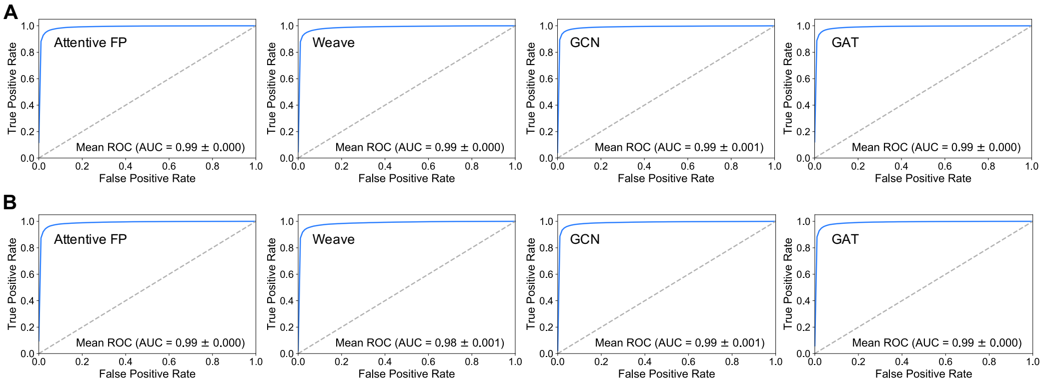


Figure A.20: ROC-AUC curves for classification on the phylum level for each of four model architectures (Weave, Attentive FP, GCN, and GAT), with the standard deviation shaded in light blue too insignificant to be visible. A standard deviation of 0.000 denotes a value of <0.001. **A.** Mean ROC-AUC curves for fingerprint-featurized experiments, with each graph displaying the mean and standard deviation across 5 sets of hyperparameters. **B.** Mean ROC-AUC curves for one-hot encoding-featurized experiments.

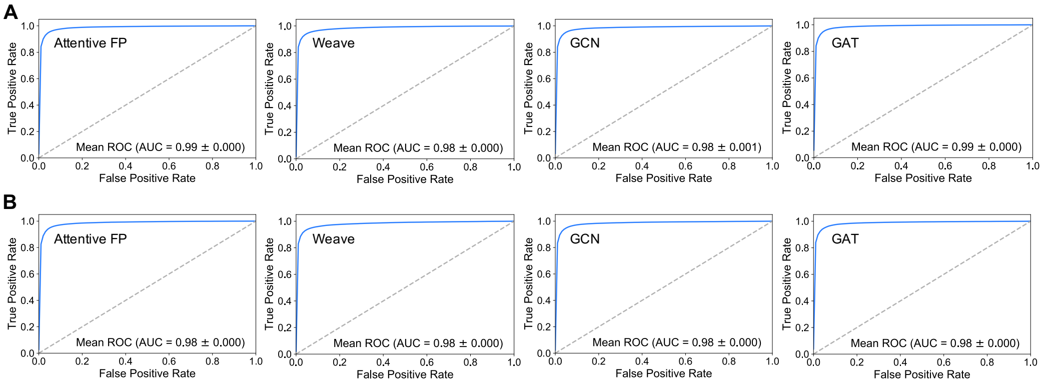


Figure A.21: ROC-AUC curves for classification on the class level for each of four model architectures (Weave, Attentive FP, GCN, and GAT), with the standard deviation shaded in light blue too insignificant to be visible. A standard deviation of 0.000 denotes a value of <0.001. **A.** Mean ROC-AUC curves for fingerprint-featurized experiments, with each graph displaying the mean and standard deviation across 5 sets of hyperparameters. **B.** Mean ROC-AUC curves for one-hot encoding-featurized experiments.

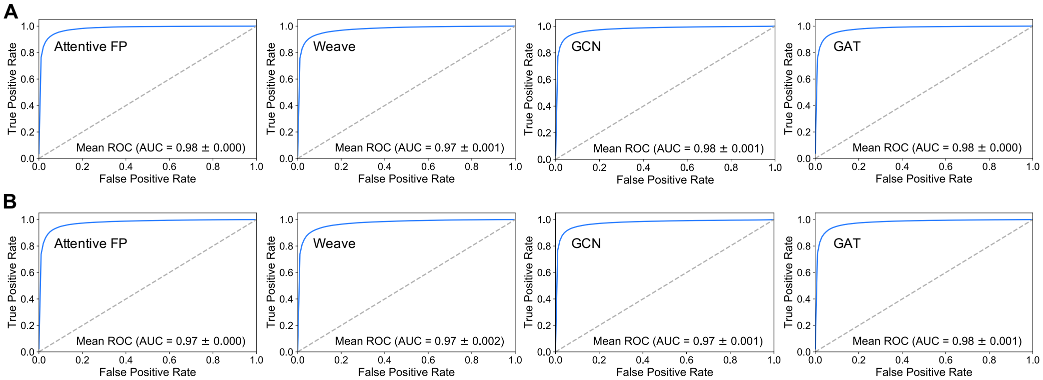


Figure A.22: ROC-AUC curves for classification on the order level for each of four model architectures (Weave, Attentive FP, GCN, and GAT), with the standard deviation shaded in light blue too insignificant to be visible. A standard deviation of 0.000 denotes a value of <0.001. **A.** Mean ROC-AUC curves for fingerprint-featurized experiments, with each graph displaying the mean and standard deviation across 5 sets of hyperparameters. **B.** Mean ROC-AUC curves for one-hot encoding-featurized experiments.

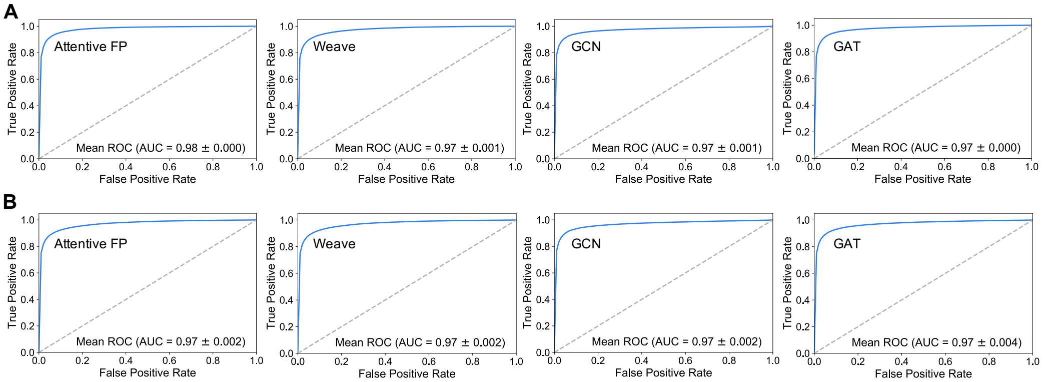


Figure A.23: ROC-AUC curves for classification on the family level for each of four model architectures (Weave, Attentive FP, GCN, and GAT), with the standard deviation shaded in light blue too insignificant to be visible. A standard deviation of 0.000 denotes a value of <0.001. **A.** Mean ROC-AUC curves for fingerprint-featurized experiments, with each graph displaying the mean and standard deviation across 5 sets of hyperparameters. **B.** Mean ROC-AUC curves for one-hot encoding-featurized experiments.

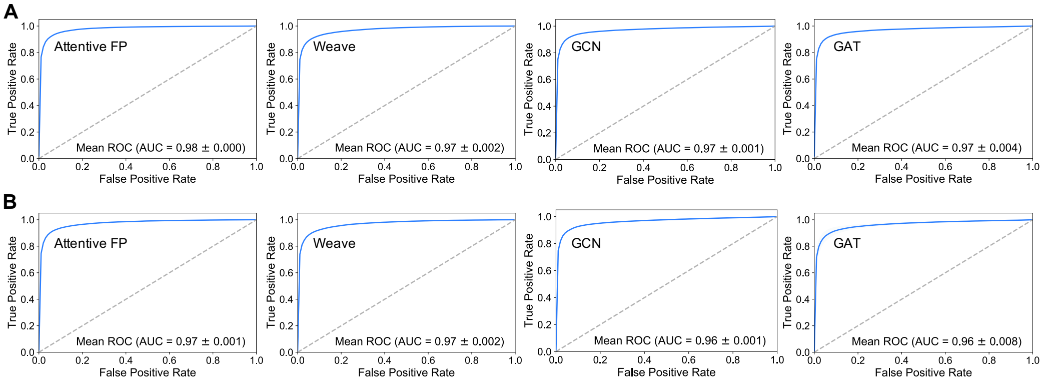


Figure A.24: ROC-AUC curves for classification on the genus level for each of four model architectures (Weave, Attentive FP, GCN, and GAT), with the standard deviation shaded in light blue too insignificant to be visible. A standard deviation of 0.000 denotes a value of <0.001. **A.** Mean ROC-AUC curves for fingerprint-featurized experiments, with each graph displaying the mean and standard deviation across 5 sets of hyperparameters. **B.** Mean ROC-AUC curves for one-hot encoding-featurized experiments.

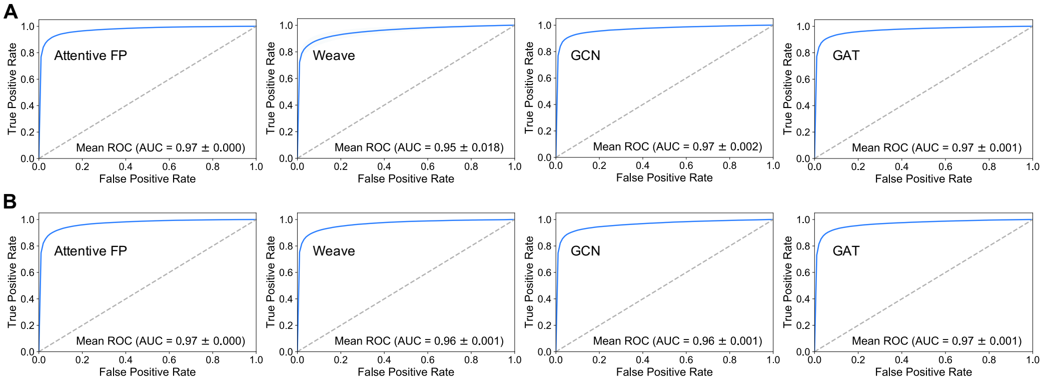


Figure A.25: ROC-AUC curves for classification on the species level for each of four model architectures (Weave, Attentive FP, GCN, and GAT), with the standard deviation shaded in light blue too insignificant to be visible. A standard deviation of 0.000 denotes a value of <0.001. **A.** Mean ROC-AUC curves for fingerprint-featurized experiments, with each graph displaying the mean and standard deviation across 5 sets of hyperparameters. **B.** Mean ROC-AUC curves for one-hot encoding-featurized experiments.

Table A.3: Validation metrics of the top-performing model architecture and attribute combinations for immunogenicity and taxonomic levels obtained from SigOpt using a 0.8, 0.2 train-validation split of the dataset. The subset accuracy values are compared directly with the results for augmented models presented in Table 1 of Bojar et al. (2021). The higher value is bolded for each task.

Class	Paper	Model + Attribute Type	ROC-AUC	F1	Recall	Precision	Accuracy	CE Loss	Hamming Loss
Immunogenicity	This Work	Weave, one-hot	0.999	**0.989**	**0.989**	**0.989**	**0.989**	**0.018**	-
	Bojar, et. al.	-	-	0.929	0.929	0.933	0.931	0.162	-
Domain	This Work	Attentive FP, fp	0.994	0.950	0.945	0.955	**0.940**	**0.081**	0.025
	Bojar, et. al.	-	-	-	-	-	0.931	0.191	-
Kingdom	This Work	Attentive FP, fp	0.997	0.936	0.895	0.922	**0.921**	**0.043**	0.014
	Bojar, et. al.	-	-	-	-	-	0.895	0.325	-
Phylum	This Work	GCN, fp	0.992	0.841	0.804	0.882	**0.808**	**0.028**	0.010
	Bojar, et. al.	-	-	-	-	-	0.801	0.754	-
Class	This Work	GCN, FP	0.986	0.775	0.718	0.841	**0.724**	**0.022**	0.006
	Bojar, et. al.	-	-	-	-	-	0.715	1.173	-
Order	This Work	GCN, fp	0.982	0.663	0.589	0.760	**0.574**	**0.017**	0.005
	Bojar, et. al.	-	-	-	-	-	0.533	2.113	-
Family	This Work	GAT, fp	0.983	0.622	0.527	0.760	**0.544**	**0.011**	0.003
	Bojar, et. al.	-	-	-	-	-	0.466	2.707	-
Genus	This Work	GCN, fp	0.975	0.561	0.470	0.697	**0.470**	**0.009**	0.002
	Bojar, et. al.	-	-	-	-	-	0.385	3.408	-
Species	This Work	GCN, fp	0.976	0.510	0.403	0.694	**0.438**	**0.007**	0.002
	Bojar, et. al.	-	-	-	-	-	0.365	3.955	-