GLYCANML: A Multi-Task and Multi-Structure Benchmark for Glycan Machine Learning

Minghao Xu[1,2,3] **Yunteng Geng**[1 *] **Yihang Zhang**[1 *] **Ling Yang**[1]
Jian Tang[2,3,4,5] **Wentao Zhang**[1 †]
[*]equal contribution [†]corresponding author
[1]Peking University [2]Mila - Québec AI Institute [3]BioGeometry
[4]HEC Montréal [5]CIFAR AI Research Chair
contacts: minghao.xu@mila.quebec, wentao.zhang@pku.edu.cn

Abstract

Glycans are basic biomolecules and perform essential functions within living organisms. The rapid increase of functional glycan data provides a good opportunity for machine learning solutions to glycan understanding. However, there still lacks a standard machine learning benchmark for glycan function prediction. In this work, we fill this blank by building a comprehensive benchmark for **Glycan** **M**achine **L**earning (**GLYCANML**). The GLYCANML benchmark consists of diverse types of tasks including glycan taxonomy prediction, glycan immunogenicity prediction, glycosylation type prediction, and protein-glycan interaction prediction. Glycans can be represented by both sequences and graphs in GLYCANML, which enables us to extensively evaluate sequence-based models and graph neural networks (GNNs) on benchmark tasks. Furthermore, by concurrently performing eight glycan taxonomy prediction tasks, we introduce the **GLYCANML-MTL** testbed for multi-task learning (MTL) algorithms. Experimental results show the superiority of modeling glycans with multi-relational GNNs, and suitable MTL methods can further boost model performance. We provide all datasets and source codes at `https://github.com/GlycanML/GlycanML` and maintain a leaderboard at `https://GlycanML.github.io/project`.

1 Introduction

Glycans are fundamental biomolecules that play crucial roles in maintaining the normal physiological functions and health status of living organisms. They can regulate inflammatory responses [23], enable the recognition and communication between cells [63], preserve stable blood sugar levels [5], *etc.* Thanks to the advance of high-throughput sequencing techniques of glycans [62, 32], a large number of glycan data are accessible, *e.g.*, the more than 240 thousand glycans stored in the GlyTouCan database [51]. This progress enables glycan function analysis by machine learning methods which are essentially data-driven.

There are some existing works that employ machine learning models to predict glycan taxonomy [10, 11], glycan immunogenicity [10, 11], glycosylation [40, 34] and protein-glycan interaction [38]. These works aim to solve one or several related glycan function prediction problems with either sequence models adapted from natural language processing (NLP) [10, 40] or graph neural networks (GNNs) [11, 38]. However, there still lacks a comprehensive benchmark studying the general effectiveness of various machine learning models on understanding diverse glycan functions, which hinders the progress of machine learning for glycan understanding. As a matter of fact, benchmark datasets greatly facilitate the machine learning research of other biomolecules like small molecules [25, 19] and proteins [43, 60].

Preprint. Under review.

Therefore, in this work, we take the initiative of building a **Glycan M**achine **L**earning (**GLYCANML**) benchmark featured with diverse types of tasks and multiple glycan representation structures. The GLYCANML benchmark consists of 11 benchmark tasks for understanding important glycan properties, including glycan taxonomy prediction, glycan immunogenicity prediction, glycosylation type prediction, and protein-glycan interaction prediction. For each task, we carefully split the benchmark dataset to evaluate the generalization ability of machine learning models in real-world scenarios. For example, in glycan taxonomy prediction, we leave out the glycans with unseen structural motifs during training for validation and test, which simulates the classification of newly discovered glycans in nature with novel molecular structures.

The GLYCANML benchmark accommodates two glycan representation structures, *i.e.*, glycan tokenized sequences and glycan planar graphs. For each structure, we adopt suitable machine learning models for representation learning, where sequence encoders such as CNN [22], LSTM [24] and Transformer [53] are employed for glycan sequence encoding, and both homogeneous GNNs [29, 54, 59] and heterogeneous GNNs [20, 45, 52] are used to encode glycan graphs. We evaluate each model on all benchmark tasks to study its general effectiveness. The GLYCANML benchmark also provides a testbed, namely **GLYCANML-MTL**, for multi-task learning (MTL) algorithms, where an MTL method is asked to simultaneously solve eight glycan taxonomy prediction problems which are highly correlated. The performance on this testbed measures how well an MTL method can transfer the knowledge learned from different glycan taxonomies, *e.g.*, transferring between species-level classification and genus-level classification.

Benchmark results show that the RGCN model [45], a typical heterogeneous GNN, performs best on most benchmark tasks, and a simple two-layer CNN can surprisingly achieve competitive performance by using only condensed sequential information of glycan structures. The MTL methods with elaborate task-reweighting strategies can further enhance the performance of glycan taxonomy prediction, showing the potential of MTL for glycan understanding. We hope the GLYCANML benchmark will spark the interest of studying glycoscience with machine learning.

2 Related Work

Glycan machine learning. With the expanding size of experimental glycomics datasets, the integration of machine learning techniques into glycoinformatics shows considerable promise [7, 34]. Early approaches that adapt machine learning into glycomics research include the use of traditional ML algorithms (*e.g.*, SVMs) to learn patterns from mass spectrometry data [31, 35], predict glycosylation sites [12, 33, 48, 42], and classify glycans [61]. Recently, thanks to the advancements in deep learning and new glycomics datasets, there has been a rise in studies applying deep learning to glycan and glycosylation modeling. DeepNGlyPred [40] seeks to identify N-glycosylated sequon from the N-GlyDE dataset [42]. SugarBase [9] is a comprehensive glycan database with metadata and analytical tools, and it supports the development of many glycan representation learning models, such as SweetOrigins [9], SweetTalk [8], SweetNet [11], glyBERT [16] and GNNGLY [1]. Deep learning-based methods like LectinOracle [38] and GlyNet [13] also made notable progress in predicting binding strengths between proteins and glycans.

However, there still lacks a comprehensive benchmark that incorporates diverse types of glycan understanding tasks and different glycan modeling methods like sequence-based and graph-based methods. Also, it is unknown how multi-task learning (MTL) influences the learning of glycan property prediction. In this work, we fill these blanks by introducing the GLYCANML benchmark with multiple task types, multiple representation schemes of glycan structures, and an MTL testbed.

Biological machine learning benchmarks. To evaluate the performance of different machine learning methods in modeling biomolecules, it is necessary to establish large-scale standardized benchmarks. MoleculeNet [58] is a widely-used benchmark of small molecule modeling, evaluating the efficacy of both traditional machine learning and deep learning on predicting molecular properties. A recent work [66] proposes a benchmark for evaluating the performance of Large Language Models (LLMs) on molecular property prediction. In the field of protein modeling, the renowned CASP [30] competition is dedicated to establishing standards for protein structure prediction. Also, benchmark datasets are constructed for machine learning guided protein engineering [43, 17], protein design [18] and protein function annotation [60, 65]. Benchmark datasets are also established for

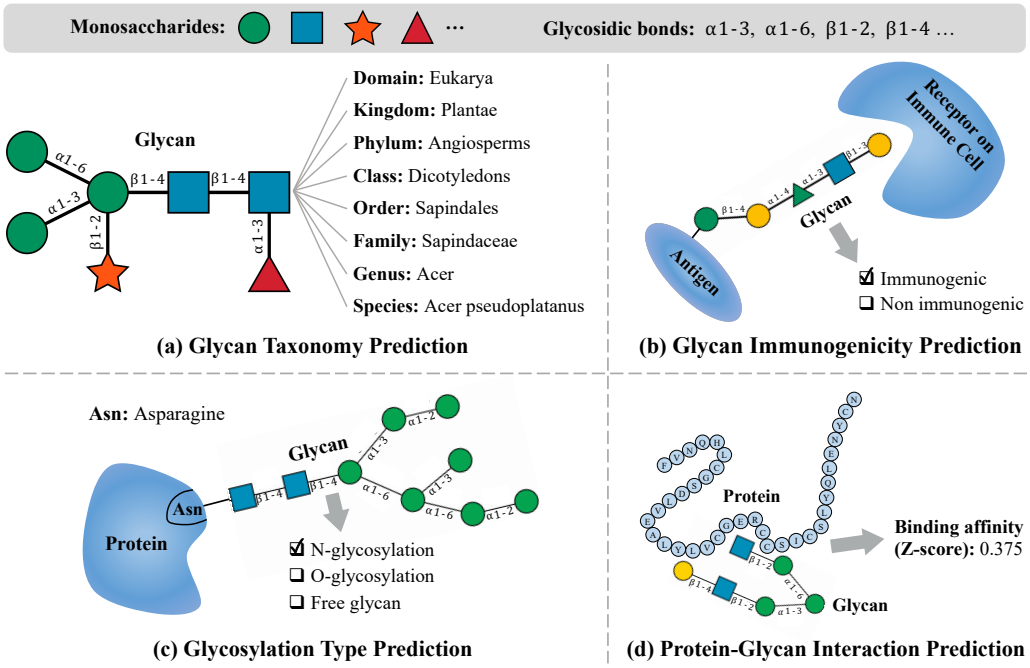

Figure 1: *Illustration of benchmark tasks.* (a) Predicting the biological taxonomy of glycans at eight levels. (b) Judging whether a glycan is immunogenic or not in organisms. (c) Analyzing how a glycan glycosylates its target protein. (d) Given a protein and a glycan, predicting their binding affinity.

other biomolecules like DNAs [26, 39] and RNAs [57]. In this work, we take the initiative of building a glycan machine learning benchmark for comprehensive glycan understanding.

3 Benchmark Tasks

The GLYCANML benchmark consists of 11 benchmark tasks, including glycan taxonomy prediction, glycan immunogenicity prediction, glycosylation type prediction, and protein-glycan interaction prediction, as illustrated in Figure 1. We summarize the information of all tasks in Table 1.

3.1 Glycan Taxonomy Prediction

Scientific significance. The taxonomy of glycans lays the foundation for glycomics research [4]. Biologists commonly classify glycans based on their origin under the hierarchical system of domain, kingdom, phylum, class, order, family, genus and species. Such a systematic classification helps us compare the similarities and differences between glycans, which further facilitates the study of glycan structures and functions. Also, the glycan taxonomy helps us understand the process of biological evolution. By comparing the structures of glycans in different organisms, we can infer their phylogenetic relationships and possible changes that may occur during evolution. Therefore, it could be very helpful to have an accurate glycan taxonomy predictor based on machine learning.

Task definition. In the GLYCANML benchmark, we study glycan taxonomy prediction on domain, kingdom, phylum, class, order, family, genus and species levels, leading to eight individual tasks. These tasks are formulated as classification problems with 4, 11, 39, 101, 210, 415, 922 and 1,737 biological categories, respectively. We report classification accuracy for each task.

Benchmark dataset. We collect the glycans in the SugarBase database [9] that are fully annotated with domain, kingdom, phylum, class, order, family, genus and species labels, with 13,209 glycans in total. We then represent each glycan with the frequencies of popular motifs (*i.e.*, those frequently occurring substructures in glycans), where the motif list proposed by Thomès et al. [50] is employed. Based on such representations, we cluster all glycans in the dataset by K-means ($K = 10$), where 8 clusters are assigned to training, and the remaining two clusters are respectively utilized for validation

Table 1: Benchmark task descriptions. We list each task along with its type, the average number of monosaccharides in each glycan for this task (in $\text{mean}_{(\text{std})}$ format), dataset statistics, and evaluation metric. *Abbr.*, Mono.: Monosaccharides.

Task	Task type	#Mono. per glycan	#Sample	#Train/Validation/Test	Metric
Taxonomy prediction of *Domain*	Classification	$6.39_{(3.51)}$	13,209	11,010/1,280/919	Accuracy (%)
Taxonomy prediction of *Kingdom*	Classification	$6.39_{(3.51)}$	13,209	11,010/1,280/919	Accuracy (%)
Taxonomy prediction of *Phylum*	Classification	$6.39_{(3.51)}$	13,209	11,010/1,280/919	Accuracy (%)
Taxonomy prediction of *Class*	Classification	$6.39_{(3.51)}$	13,209	11,010/1,280/919	Accuracy (%)
Taxonomy prediction of *Order*	Classification	$6.39_{(3.51)}$	13,209	11,010/1,280/919	Accuracy (%)
Taxonomy prediction of *Family*	Classification	$6.39_{(3.51)}$	13,209	11,010/1,280/919	Accuracy (%)
Taxonomy prediction of *Genus*	Classification	$6.39_{(3.51)}$	13,209	11,010/1,280/919	Accuracy (%)
Taxonomy prediction of *Species*	Classification	$6.39_{(3.51)}$	13,209	11,010/1,280/919	Accuracy (%)
Immunogenicity prediction	Binary classification	$7.30_{(3.78)}$	1,320	1,026/149/145	AUPRC
Glycosylation type prediction	Classification	$9.04_{(3.96)}$	1,683	1,356/163/164	Accuracy (%)
Protein-Glycan interaction prediction	Regression	$6.56_{(4.54)}$	564,647	442,396/58,887/63,364	Spearman's ρ

and test. By using such dataset splits, this set of tasks evaluate how well a machine learning model can generalize across structurally distinct glycans.

3.2 Glycan Immunogenicity Prediction

Scientific significance. Predicting the immunogenicity of glycans is of great significance for vaccine design and disease treatment. (1) Glycans are key components in many vaccines, especially in bacterial vaccines. By predicting the immunogenicity of glycans, researchers can design more effective vaccine formulations, thereby improving the protective effect of vaccines [27]. (2) In addition, certain glycans can inhibit tumor growth by activating the immune system [3], and therefore accurately predicting glycan immunogenicity can help optimize tumor treatment strategies.

Task definition. We formulate this task as a binary classification problem, *i.e.*, predicting whether a glycan is immunogenic or not. We evaluate with the AUPRC metric to measure the trade-off between precision and recall of a model on immunogenic glycans.

Benchmark dataset. We select out all glycans in the SugarBase [9] whose immunogenicity is annotated based on evidences in literature, summing up to 1,320 glycans. As in glycan taxonomy prediction, we use the motif-based dataset splitting scheme to derive training, validation and test splits with an 8:1:1 ratio. In this way, we evaluate models' generalization ability across structurally distinct glycans.

3.3 Glycosylation Type Prediction

Scientific significance. Glycans are a class of macromolecules with diverse biological activities, including immune system regulation, antitumor effects, antiviral effects, *etc.* By predicting the type of glycosylation, researchers can better understand the relationship between glycan structure and its functions. Understanding the structure-function relationship is crucial for designing and synthesizing glycan derivatives with specific biological activities [6].

Task definition. Given a glycan, we aim at predicting whether it forms N-glycosylation, O-glycosylation or maintains a free state, formulated as a three-way classification problem. The classification accuracy is used for evaluation.

Benchmark dataset. We traverse the GlyConnect database [2] and select out all glycans with glycosylation annotations, with 1,683 glycans in total. Upon these data, we again employ the motif-based dataset splitting scheme (introduced in Section 3.1) to construct training, validation and test splits with an 8:1:1 ratio. This task again assesses the generalization ability across the glycans with distinct structures.

3.4 Protein-Glycan Interaction Prediction

Scientific significance. The interactions between proteins and glycans play a crucial role in cellular signaling, affecting cell growth, differentiation, and apoptosis [55]. For example, glycans are one of the main components of the extracellular matrix (ECM), which interact with proteins such as collagen,

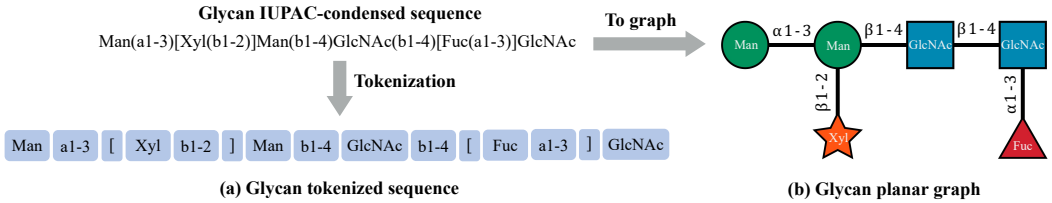

Figure 2: *Illustration of glycan representations.* (a) The glycan tokenized sequence is derived by tokenizing the IUPAC-condensed sequence. (b) The glycan planar graph is constructed by transforming the IUPAC-condensed sequence to graph.

laminin, and fibronectin to form the structural framework of ECM, providing physical support and passing biochemical signals to cells [15]. Understanding these interactions helps reveal how cells respond to external signals.

Task definition. Given a protein and a glycan, this task aims to regress their binding affinity, where the Z-score transformed relative fluorescence unit represents binding affinity. For this task, we adopt the Spearman's correlation coefficient as the evaluation metric to measure how well a model ranks a set of protein-glycan pairs with different binding affinities.

Benchmark dataset. This benchmark dataset is built upon 564,647 protein-glycan interactions deposited in the LectinOracle database [38]. It is desired to have a model that can well generalize to new proteins against training ones, considering the continuous discovery of new proteins by sequencing techniques. Therefore, we split the dataset based on protein sequence similarity. Specifically, we first cluster all protein sequences using MMseqs2 [47] (minimum sequence identity within each cluster: 0.5), and then we derive training, validation and test proteins by splitting all clusters with an 8:1:1 ratio. Finally, samples of protein-glycan pairs are split according to the protein splits.

4 Methods

4.1 Representations

In the GLYCANML benchmark, we adopt two glycan representation structures, *i.e.*, glycan tokenized sequence and glycan planar graph, as illustrated in Figure 2.

Glycan tokenized sequence. A glycan is commonly represented by an IUPAC-condensed sequence. For example, in the sequence "Glc(a1-4)Glc", two glucoses are connected by an alpha-1,4-glycosidic bond, and this structure is the basic component of starch, a typical glycan. To process such sequences with machine learning models, a straightforward way is to tokenize the IUPAC-condensed sequence. Specifically, we regard each monosaccharide (*e.g.*, Glc), each glycosidic bond (*e.g.*, a1-4), and each bracket that indicates glycan branching (*i.e.*, "[" and "]") as a single token, which derives the glycan tokenized sequence, denoted as $x_s = \{s_i\}_{i=1}^{N}$. Various sequence encoders like Transformers [53] can then be applied to such tokenized sequences for glycan representation learning.

Glycan planar graph. Essentially, an IUPAC-condensed sequence describes the branching structure of a glycan, in which the part between brackets "[" and "]" denotes a side branch of the main branch, as illustrated in Figure 2. This structure is well represented by a planar graph $x_g = (\mathcal{V}, \mathcal{E})$, in which nodes $\mathcal{V}$ denotes monosaccharides, and edges $\mathcal{E}$ denotes glycosidic bonds. In this way, graph neural networks (GNNs) are readily used for glycan modeling.

4.2 Baselines

We include three types of models in our benchmark, *i.e.*, sequence encoders for modeling glycan tokenized sequences, and homogeneous and heterogeneous GNNs for modeling glycan planar graphs, with 10 baseline models in total. Their detailed architectures are provided in Table 2.

Sequence encoders. We study the performance of four typical sequence encoders. Inspired by the success of shallow CNNs in modeling biological sequences like protein sequences [46, 60], we investigate (1) a 2-layer shallow CNN model along with (2) a deep residual network (ResNet) [22]

Table 2: Architectures of baseline models. *Abbr.*, Params.: parameters; dim.: dimension; conv.: convolutional; attn.: attention; concat.: concatenate.

Model	Input Layer	Hidden Layers	Output Layer	#Params.
Sequence encoders				
Shallow CNN [46]	128-dim. token embedding	2 × 1D conv. layers (hidden dim.: 128; kernel size: 5; stride: 1; padding: 2)	max pooling over all tokens	191.7K
ResNet [22]	512-dim. token embedding + 512-dim. positional embedding	8 × residual blocks (hidden dim.: 512; kernel size: 3; stride: 1; padding: 1)	attentive weighted sum over all tokens	11.4M
LSTM [24]	640-dim. token embedding	3 × bidirectional LSTM layers (hidden dim.: 640)	weighted sum over all tokens + linear (output dim.: 640) + Tanh	26.7M
Transformer [53]	512-dim. token embedding + 512-dim. positional embedding	4 × Transformer blocks (hidden dim.: 512; #attn. heads: 8; activation: GELU)	linear (output dim.: 512) + Tanh upon `[CLS]` token	21.4M
Homogeneous GNNs				
GCN [29]	128-dim. node embedding	3 × GCN layers	concat. mean & max pooling	67.8K
GAT [54]	128-dim. node embedding	3 × GAT layers (#attn. heads: 2)	concat. mean & max pooling	69.4K
GIN [59]	128-dim. node embedding	3 × GIN layers	concat. mean & max pooling	117.4K
Heterogeneous GNNs				
MPNN [20]	128-dim. node & edge embedding	3 × MPNN layers	Set2Set pooling [56] (#steps: 3)	4.0M
RGCN [45]	128-dim. node embedding	3 × RGCN layers	concat. mean & max pooling	4.2M
CompGCN [52]	128-dim. node embedding	3 × CompGCN layers	concat. mean & max pooling	150.4K

with 8 hidden layers. These two CNN models mainly focus on capturing local information in glycan sequences. To investigate the importance of long context modeling for glycan understanding, we also include (3) a 3-layer bidirectional LSTM [24] and (4) a 4-layer Transformer encoder [53].

Homogeneous GNNs. Upon glycan planar graphs, standard GNNs designed for homogeneous graph modeling can be readily used to learn glycan representations. In our benchmark, three typical homogeneous GNNs, *i.e.*, GCN [29], GAT [54] and GIN [59], serve as baselines, and they are all configured with 3 message passing layers.

Heterogeneous GNNs. As a matter of fact, modeling glycans as homogeneous graphs is suboptimal, in which the rich information within glycosidic bonds is fully ignored. To capture the complete information in glycan graphs, it is more proper to view them as heterogeneous graphs and employ heterogeneous GNNs for representation learning. Therefore, we adapt three popular heterogeneous GNNs, *i.e.*, MPNN [20], RGCN [45] and CompGCN [52], to model glycan graphs, where each model is equipped with 3 message passing layers.

4.3 Model Pipelines

Depending on inputs, the benchmark tasks of GLYCANML can be solved with two model pipelines.

Single-glycan prediction. This pipeline handles the tasks that predict the properties of individual glycans, including glycan taxonomy prediction, glycan immunogenicity prediction, and glycosylation type prediction. For each task, the glycan representation vector is first extracted by a glycan encoder and then passed to an MLP head for task-specific prediction.

Protein-glycan interaction prediction. Because of the additional input of protein, the protein-glycan interaction prediction task requires a different pipeline. Given a protein and a glycan, we first extract the protein representation vector with a protein encoder (*e.g.*, the ESM-1b pre-trained protein language model [44] in this work) and extract the glycan representation vector with a glycan encoder, and these two vectors are then concatenated and sent to an MLP head for interaction prediction.

4.4 Multi-Task Learning

In GLYCANML, the glycan taxonomy prediction tasks classify glycans under the hierarchical system from domain to species. These tasks are highly correlated and well-suited for multi-task learning (MTL) where related tasks are learned together for better generalization performance [64]. Therefore, we integrate eight glycan taxonomy prediction tasks in GLYCANML as a testbed for MTL algorithms, named as the **GLYCANML-MTL** benchmark.

On this benchmark, we study 6 representative MTL methods that focus on loss design and model optimization under the MTL setting. All these methods use the network architecture with hard

parameter sharing [64], where all tasks share a common backbone encoder, and each task owns its individual prediction head. We introduce these methods below, with an abbreviation after each one.

- **Naive MTL (N-MTL):** The most straightforward way to perform MTL is to sum up the losses of all tasks with equal weights and optimize the model with this loss summation. Denoting the losses of GLYCANML-MTL tasks as $\mathcal{L}_i$ ($i = 1, \cdots, 8$), the naive MTL loss is defined as: $\mathcal{L}_{\mathrm{N-MTL}} = \sum_{i=1}^{8} \mathcal{L}_i$.
- **Gradient Normalization (GN) [14]:** However, regarding all tasks equally is suboptimal, considering the varying difficulties of different tasks. Therefore, this method employs a weighted loss summation $\mathcal{L}_{\mathrm{GN}} = \sum_{i=1}^{8} w_i \mathcal{L}_i$, where the weights satisfy: $\sum_{i=1}^{8} w_i = 8$. The main idea of gradient normalization is that different tasks should be trained at similar rates (*i.e.*, similar speed of convergence). To achieve this goal, authors first deem the L2 norm of per-task gradient as the training rate of the task: $r_i = ||\nabla_\theta w_i \mathcal{L}_i||_2$ (θ denotes model parameters), and all tasks are then pushed to have similar training rates by optimizing the loss $\mathcal{L}(w_1, \cdots, w_8) = \sum_{i=1}^{8} ||r_i - \bar{r}||_1$ ($\bar{r} = (\sum_{i=1}^{8} r_i)/8$). For each training step, this loss is first optimized *w.r.t.* loss weights $\{w_i\}_{i=1}^{8}$, and, using the updated loss weights, the MTL loss $\mathcal{L}_{\mathrm{GN}}$ optimizes whole model parameters.
- **Temperature Scaling (TS) [28]:** For classification tasks, the sharpness of categorical distribution represents prediction uncertainty, further implying task difficulty. Inspired by this fact, the TS method seeks to weigh different tasks by scaling their classification logits. In this way, each task loss is defined as $\mathcal{L}_i^{\mathrm{TS}} = -\log(\mathrm{Softmax}(f_\theta(y|x)/\sigma_i^2))$ ($i = 1, \cdots, 8$), where $f_\theta(y|x)$ represents the classification logit of sample x at class y, and σ_i denotes the task-specific temperature parameter for scaling. The temperature parameters are learned along with the whole model.
- **Uncertainty Weighting (UW) [28]:** Kendall et al. [28] shows that the temperature-scaled losses above can be approximated by a weighted summation of unscaled losses: $\mathcal{L}_{\mathrm{UW}} = \sum_{i=1}^{8} \mathcal{L}_i/\sigma_i^2 + \log \sigma_i$, where the weighting parameters $\{\sigma_i\}_{i=1}^{8}$ are learnable. This method also weighs different tasks based on the uncertainty of task predictions.
- **Dynamic Weight Averaging (DWA) [37]:** The loss scales along training can well indicate task convergence. Therefore, this method employs the ratio of consecutive losses to weigh different tasks: $w_i(t) = 8 \cdot \mathrm{Softmax}(\mathcal{L}_i(t)/\mathcal{L}_i(t-1))$, where $\mathcal{L}_i(t)$ denotes the loss of task i at training step t. In this way, more weights are assigned to the tasks with slower convergence.
- **Dynamic Task Prioritization (DTP) [21]:** This method maintains a key performance indicator (KPI) $\kappa_i(t)$ for each task along training (moving average of classification accuracy on our benchmark) and weighs different tasks in a focal loss [36] manner: $w_i(t) = -(1 - \kappa_i(t))^{\gamma_i} \log \kappa_i(t)$, where γ_i is the focusing hyperparameter for task i. Such a task reweighting scheme pays more attention to difficult tasks with low KPI.

5 Experiments

5.1 Experimental Setups

Model setups. For glycan taxonomy, immunogenicity and glycosylation type prediction tasks, upon the glycan embedding extracted by the glycan encoder, we use an MLP with 2 hidden layers and a ReLU nonlinearity in between to perform prediction. For protein-glycan interaction prediction, we use the ESM-1b pre-trained protein language model [44] to extract protein embedding, and, upon the concatenation of protein and glycan embeddings, the binding affinity is predicted by a 2-layer MLP with ReLU activation.

Training setups. We conduct every experiment on three seeds (0, 1 and 2) and report the mean and standard deviation of results. We train with an Adam optimizer (learning rate: 5×10^{-4}, weight decay: 1×10^{-3}) for 50 epochs on taxonomy, immunogenicity and glycosylation type prediction and for 10 epochs on interaction prediction. The batch size is set as 32 for interaction prediction and 256 for other tasks. For model training, we use cross entropy loss to train taxonomy and glycosylation type prediction tasks, use binary cross entropy loss to train immunogenicity prediction, and adopt mean squared error to train interaction prediction. For model selection, 10 times of validation are uniformly performed along the training process, and the checkpoint with the best validation performance is selected for test. For multi-task learning (MTL), the focusing parameter γ of the dynamic task prioritization (DTP) method is set as 2.0, and the model selection of all MTL methods is based on

Table 3: Benchmark results on single-task learning. We report *mean (std)* for each experiment. Three color scales of blue denote the first, second and third best performance. *Abbr.*, Immuno: Immunogenicity; Glycos: Glycosylation.

Model	Domain	Kingdom	Phylum	Class	Order	Family	Genus	Species	Immuno	Glycos	Interaction	Mean Rank
Sequence encoders												
Shallow CNN [46]	93.76$_{(0.66)}$	91.66$_{(2.29)}$	86.87$_{(0.71)}$	70.62$_{(0.76)}$	47.01$_{(1.86)}$	45.27$_{(1.52)}$	38.88$_{(1.24)}$	33.70$_{(1.12)}$	0.776$_{(0.267)}$	97.22$_{(0.44)}$	0.261$_{(0.008)}$	3.5
ResNet [22]	93.29$_{(1.46)}$	89.52$_{(0.72)}$	81.32$_{(3.16)}$	65.29$_{(1.54)}$	41.49$_{(1.62)}$	37.36$_{(0.99)}$	32.93$_{(2.88)}$	26.59$_{(1.89)}$	0.754$_{(0.124)}$	**98.55$_{(0.58)}$**	0.273$_{(0.004)}$	6.0
LSTM [24]	92.78$_{(1.16)}$	88.43$_{(1.55)}$	81.61$_{(1.36)}$	63.47$_{(1.95)}$	41.20$_{(2.84)}$	38.74$_{(1.47)}$	28.87$_{(2.77)}$	26.04$_{(1.73)}$	**0.862$_{(0.016)}$**	96.28$_{(1.42)}$	**0.280$_{(0.001)}$**	6.3
Transformer [53]	91.98$_{(0.27)}$	87.34$_{(1.16)}$	80.49$_{(1.91)}$	62.13$_{(1.47)}$	38.30$_{(1.10)}$	33.30$_{(0.86)}$	27.97$_{(1.61)}$	27.49$_{(1.20)}$	0.729$_{(0.069)}$	95.90$_{(1.45)}$	0.244$_{(0.009)}$	8.5
Homogeneous GNNs												
GCN [29]	94.38$_{(0.51)}$	92.06$_{(0.11)}$	80.30$_{(0.44)}$	63.26$_{(0.64)}$	38.70$_{(1.06)}$	34.93$_{(0.76)}$	33.37$_{(0.49)}$	31.01$_{(0.87)}$	0.688$_{(0.023)}$	95.90$_{(0.11)}$	0.233$_{(0.009)}$	7.2
GAT [54]	94.27$_{(0.41)}$	92.56$_{(0.25)}$	80.81$_{(0.60)}$	62.57$_{(2.29)}$	40.77$_{(2.16)}$	37.50$_{(0.91)}$	36.38$_{(1.10)}$	34.13$_{(0.99)}$	0.685$_{(0.053)}$	94.63$_{(0.39)}$	0.229$_{(0.002)}$	6.6
GIN [59]	94.41$_{(0.17)}$	92.35$_{(0.17)}$	84.04$_{(1.07)}$	67.54$_{(0.31)}$	35.40$_{(1.92)}$	40.04$_{(0.86)}$	34.28$_{(0.44)}$	31.85$_{(2.19)}$	0.716$_{(0.051)}$	97.16$_{(0.19)}$	0.249$_{(0.004)}$	5.1
Heterogeneous GNNs												
MPNN [20]	93.83$_{(0.17)}$	90.39$_{(0.94)}$	82.48$_{(0.94)}$	66.52$_{(1.76)}$	45.85$_{(0.87)}$	41.60$_{(0.64)}$	37.69$_{(2.02)}$	33.80$_{(1.87)}$	0.674$_{(0.119)}$	97.41$_{(0.90)}$	0.217$_{(0.002)}$	5.6
RGCN [45]	**94.78$_{(0.11)}$**	91.80$_{(0.31)}$	**86.94$_{(0.78)}$**	**74.17$_{(0.44)}$**	**47.91$_{(2.07)}$**	**46.35$_{(0.50)}$**	40.30$_{(1.09)}$	38.12$_{(1.15)}$	0.780$_{(0.006)}$	95.14$_{(2.21)}$	0.262$_{(0.005)}$	**2.5**
CompGCN [52]	93.94$_{(0.41)}$	**93.22$_{(0.38)}$**	86.40$_{(0.76)}$	69.28$_{(0.99)}$	45.77$_{(0.85)}$	44.03$_{(1.51)}$	**40.70$_{(1.73)}$**	**40.04$_{(1.32)}$**	0.692$_{(0.006)}$	94.38$_{(3.13)}$	0.257$_{(0.004)}$	3.9

the mean accuracy over all tasks on the validation set. We conduct all experiments on a local server with 100 CPU cores and 4 NVIDIA GeForce RTX 4090 GPUs (24GB). Our implementation is based on the PyTorch [41] deep learning library (BSD-style license) and TorchDrug [67] drug discovery platform (Apache-2.0 license).

5.2 Benchmark Results on Single-Task Learning

In Table 3, we report the single-task performance of 10 representative glycan sequence and graph encoders on benchmark tasks, and the mean rank of each model over the whole benchmark is also presented. Based on these results, we highlight the following findings:

- **Multi-relational GNNs show superiority in glycan modeling.** Two typical multi-relational GNNs, *i.e.*, RGCN and CompGCN, respectively rank first and third place in terms of mean rank. Especially, RGCN achieves the best performance on 5 out of 11 benchmark tasks. Therefore, it is beneficial to model a glycan as a multi-relational graph, where different types of glycosidic bonds are deemed as different relations between monosaccharides.
- **A simple shallow CNN is surprisingly effective.** It is observed that the 2-layer shallow CNN ranks second place in terms of mean rank, and it gains the second-best performance on 4 out of 11 tasks and the third-best performance on 3 out of 11 tasks. We thereby demonstrate that such a shallow CNN model is sufficient to produce informative glycan representations and lead to competitive performance, which aligns with previous findings that shallow CNNs can well model biological sequences like protein sequences [46, 60].
- **It is important to utilize glycosidic bond information.** We can observe clear performance gains of heterogeneous GNNs over homogeneous GNNs on glycan modeling, where in terms of mean rank, three heterogeneous GNNs rank 1st, 3rd and 5th places, while three homogeneous GNNs rank 4th, 8th and 9th places. Compared to homogeneous GNNs that regard all glycosidic bonds as the same, heterogeneous GNNs fully utilize glycosidic bond information by individually treating each type of bonds, leading to obvious benefits.

5.3 Benchmark Results on Multi-Task Learning

In Table 4, we report the benchmark results of different MTL methods against single-task learning. We select shallow CNN and RGCN, *i.e.*, best-performing glycan sequence and graph encoders, as the backbone encoder, and all MTL methods are evaluated on each of them. According to benchmark results, we have the following findings:

- **The temperature scaling (TS) approach performs best.** On both shallow CNN and RGCN, the TS approach achieves the highest mean accuracy, and it outperforms single-task learning with a clear margin (*i.e.*, 1.63% improvement in mean accuracy) when using RGCN as backbone encoder. Therefore, the TS approach can well balance the learning signals from different glycan taxonomy prediction tasks, leading to stable performance gains.

Table 4: Benchmark results on multi-task learning. We report *mean (std)* for each experiment. Two color scales of blue denote the first and second best performance. *Abbr.*, Acc: Accuracy.

Method	Domain	Kingdom	Phylum	Class	Order	Family	Genus	Species	Mean Acc (%)
Backbone encoder: Shallow CNN									
Single-Task	**93.76$_{(0.66)}$**	91.66$_{(2.29)}$	**86.87$_{(0.71)}$**	70.62$_{(0.76)}$	47.01$_{(1.86)}$	**45.27$_{(1.52)}$**	38.88$_{(1.24)}$	33.70$_{(1.12)}$	63.47$_{(0.42)}$
N-MTL	93.25$_{(0.38)}$	91.51$_{(0.54)}$	84.77$_{(0.47)}$	70.58$_{(0.44)}$	47.62$_{(0.41)}$	44.11$_{(1.63)}$	39.32$_{(1.16)}$	**36.74$_{(0.35)}$**	63.49$_{(0.55)}$
GN [14]	92.96$_{(0.49)}$	91.08$_{(0.89)}$	84.95$_{(0.69)}$	70.29$_{(0.66)}$	**48.13$_{(1.21)}$**	42.40$_{(1.49)}$	38.74$_{(0.93)}$	36.49$_{(1.63)}$	63.13$_{(0.25)}$
TS [28]	93.58$_{(0.50)}$	91.55$_{(1.03)}$	85.24$_{(0.62)}$	**71.49$_{(0.58)}$**	47.08$_{(0.60)}$	45.19$_{(2.50)}$	40.19$_{(1.54)}$	35.84$_{(1.20)}$	**63.77$_{(0.19)}$**
UW [28]	92.93$_{(0.94)}$	91.59$_{(1.34)}$	85.56$_{(0.94)}$	69.57$_{(2.24)}$	45.96$_{(2.16)}$	44.40$_{(0.00)}$	38.19$_{(0.22)}$	34.78$_{(1.11)}$	62.87$_{(0.87)}$
DWA [37]	93.25$_{(0.82)}$	**92.31$_{(0.44)}$**	84.73$_{(0.77)}$	69.06$_{(3.36)}$	45.85$_{(3.02)}$	41.57$_{(2.17)}$	35.26$_{(3.79)}$	33.04$_{(0.95)}$	61.88$_{(0.90)}$
DTP [21]	93.54$_{(0.41)}$	90.79$_{(1.26)}$	83.71$_{(0.06)}$	69.02$_{(2.47)}$	46.50$_{(1.24)}$	42.76$_{(2.02)}$	**40.26$_{(1.10)}$**	36.02$_{(1.23)}$	62.83$_{(0.45)}$
Backbone encoder: RGCN									
Single-Task	**94.78$_{(0.11)}$**	91.80$_{(0.31)}$	**86.94$_{(0.78)}$**	**74.17$_{(0.44)}$**	47.91$_{(2.07)}$	46.35$_{(0.50)}$	40.30$_{(1.09)}$	38.12$_{(1.15)}$	65.05$_{(0.21)}$
N-MTL	93.04$_{(1.80)}$	91.59$_{(1.67)}$	85.75$_{(2.57)}$	72.33$_{(1.57)}$	47.99$_{(1.75)}$	45.67$_{(1.64)}$	40.66$_{(1.36)}$	38.59$_{(1.44)}$	64.45$_{(0.40)}$
GN [14]	93.73$_{(0.72)}$	91.88$_{(0.80)}$	85.38$_{(0.66)}$	70.80$_{(0.33)}$	46.32$_{(1.09)}$	44.11$_{(0.13)}$	39.79$_{(1.36)}$	38.67$_{(1.52)}$	63.83$_{(0.65)}$
TS [28]	94.52$_{(0.23)}$	91.62$_{(0.78)}$	86.62$_{(0.86)}$	73.88$_{(1.69)}$	**50.63$_{(0.33)}$**	**47.99$_{(0.22)}$**	**44.87$_{(0.72)}$**	**43.31$_{(0.29)}$**	**66.68$_{(0.23)}$**
UW [28]	93.43$_{(0.56)}$	91.77$_{(0.23)}$	85.60$_{(0.41)}$	73.45$_{(1.61)}$	48.28$_{(1.42)}$	44.90$_{(0.88)}$	42.15$_{(0.71)}$	39.61$_{(0.89)}$	64.90$_{(0.39)}$
DWA [37]	94.09$_{(0.50)}$	91.80$_{(1.11)}$	86.43$_{(1.27)}$	69.42$_{(1.25)}$	45.77$_{(1.03)}$	44.11$_{(1.24)}$	40.08$_{(0.93)}$	37.47$_{(0.44)}$	63.65$_{(0.25)}$
DTP [21]	94.02$_{(0.66)}$	**92.53$_{(0.44)}$**	85.09$_{(0.86)}$	69.60$_{(1.75)}$	46.17$_{(0.70)}$	44.29$_{(0.47)}$	41.49$_{(1.28)}$	39.97$_{(2.11)}$	64.15$_{(0.25)}$

- **MTL methods are not always beneficial.** On shallow CNN, only the naive MTL (N-MTL) and the TS method outperform single-task learning in terms of mean accuracy; on RGCN, only the TS method outperforms single-task learning in terms of mean accuracy. Actually, most MTL methods lead to performance decrease compared to single-task learning. These results suggest the high difficulty of balancing between different glycan taxonomy prediction tasks. More efforts are thus required to boost the MTL performance on the GLYCANML-MTL testbed, which we leave as one of our major future works.
- **MTL is more helpful for difficult tasks.** From domain-level to species-level classification (*i.e.*, from 4-way classification to 1,737-way classification), the task difficulty monotonically increases. We can observe more benefits of MTL on difficult tasks. For example, on RGCN, the TS method outperforms single-task learning on order-, family-, genus- and species-level classification (*i.e.*, tasks with more categories), while the TS method is inferior on domain-, kingdom-, phylum- and class-level classification (*i.e.*, tasks with fewer categories). Therefore, in MTL, more attention is paid to difficult tasks, leading to better performance on these tasks.

6 Conclusions and Future Work

In this work, we build a comprehensive benchmark GLYCANML for glycan machine learning. It consists of diverse types of glycan understanding tasks, including glycan taxonomy prediction, glycan immunogenicity prediction, glycosylation type prediction, and protein-glycan interaction prediction. In GLYCANML, we support two representation methods of glycan structures, *i.e.*, glycan tokenized sequences and glycan planar graphs. Additionally, on eight highly correlated glycan taxonomy prediction tasks, we set up a testbed GLYCANML-MTL to compare different multi-task learning (MTL) algorithms. According to the benchmark results, multi-relational graph neural networks (GNNs) show great promise for glycan modeling, and well-designed MTL methods can further boost model performance.

Limitations and future work. The current GLYCANML benchmark is limited to modeling glycan sequences and 2D glycan graphs, without benchmark datasets of 3D glycan structures. Therefore, in the future, we will go beyond sequence- and 2D-graph-based datasets and methods to 3D-structure-based ones. Also, we will work along with the community to further promote the efforts of MTL for better glycan understanding.

7 Broader Societal Impacts

This work aims at building a comprehensive benchmark for glycan machine learning, incorporating diverse types of glycan understanding tasks, multiple representation methods of glycan structures, and a testbed for multi-task learning algorithms. By evaluating on the proposed benchmark, we can

judge the general effectiveness of a machine learning model on predicting different glycan functions. Therefore, such a benchmark can promote the application of machine learning methods to various real-world glycan-related tasks, such as vaccine design [27] and cancer research [49].

However, we should not ignore the potential harmful aspects brought by glycan machine learning models developed on our benchmark, *e.g.*, designing vaccines with severe adverse reactions. To mitigate such risks, our future works will encourage the responsible usage of the GLYCANML benchmark for real-world problems.

References

[1] Alhasan Alkuhlani, Walaa Gad, Mohamed Roushdy, and Abdel-Badeeh M Salem. Gnngly: Graph neural networks for glycan classification. *IEEE Access*, 2023.

[2] Davide Alocci, Julien Mariethoz, Alessandra Gastaldello, Elisabeth Gasteiger, Niclas G Karlsson, Daniel Kolarich, Nicolle H Packer, and Frédérique Lisacek. Glyconnect: glycoproteomics goes visual, interactive, and analytical. *Journal of proteome research*, 18(2):664–677, 2018.

[3] 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.

[4] Kiyoko F Aoki-Kinoshita. An introduction to bioinformatics for glycomics research. *PLoS computational biology*, 4(5):e1000075, 2008.

[5] Mairead L Bermingham, Marco Colombo, Stuart J McGurnaghan, Luke AK Blackbourn, Frano Vučković, Maja Pučić Baković, Irena Trbojević-Akmačić, Gordan Lauc, Felix Agakov, Anna S Agakova, et al. N-glycan profile and kidney disease in type 1 diabetes. *Diabetes care*, 41(1): 79–87, 2018.

[6] Erhard Bieberich. Synthesis, processing, and function of n-glycans in n-glycoproteins. *Glycobiology of the nervous system*, pages 47–70, 2014.

[7] Daniel Bojar and Frederique Lisacek. Glycoinformatics in the artificial intelligence era. *Chemical Reviews*, 122(20):15971–15988, 2022.

[8] Daniel Bojar, Diogo M Camacho, and James J Collins. Using natural language processing to learn the grammar of glycans. *bioRxiv*, pages 2020–01, 2020.

[9] Daniel Bojar, Rani K Powers, Diogo M Camacho, and James J Collins. Sweetorigins: Extracting evolutionary information from glycans. *bioRxiv*, pages 2020–04, 2020.

[10] 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.

[11] Rebekka Burkholz, John Quackenbush, and Daniel Bojar. Using graph convolutional neural networks to learn a representation for glycans. *Cell Reports*, 35(11), 2021.

[12] Cornelia Caragea, Jivko Sinapov, Adrian Silvescu, Drena Dobbs, and Vasant Honavar. Glycosylation site prediction using ensembles of support vector machine classifiers. *BMC bioinformatics*, 8:1–13, 2007.

[13] Eric J Carpenter, Shaurya Seth, Noel Yue, Russell Greiner, and Ratmir Derda. Glynet: a multi-task neural network for predicting protein–glycan interactions. *Chemical Science*, 13(22): 6669–6686, 2022.

[14] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018.

[15] Miriam Cohen. Notable aspects of glycan-protein interactions. *Biomolecules*, 5(3):2056–2072, 2015.

[16] Bowen Dai, Daniel E Mattox, and Chris Bailey-Kellogg. Attention please: modeling global and local context in glycan structure-function relationships. *bioRxiv*, pages 2021–10, 2021.

[17] Christian Dallago, Jody Mou, Kadina E Johnston, Bruce J Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. Flip: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv*, 2021.

[18] Zhangyang Gao, Cheng Tan, and Stan Z Li. Alphadesign: A graph protein design method and benchmark on alphafolddb. *arXiv preprint arXiv:2202.01079*, 2022.

[19] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

[20] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

[21] Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 270–287, 2018.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[23] Sophia M Hochrein, Hao Wu, Miriam Eckstein, Laura Arrigoni, Josip S Herman, Fabian Schumacher, Christian Gerecke, Mathias Rosenfeldt, Dominic Grün, Burkhard Kleuser, et al. The glucose transporter glut3 controls t helper 17 cell responses through glycolytic-epigenetic reprogramming. *Cell metabolism*, 34(4):516–532, 2022.

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[25] John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.

[26] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

[27] Paulina Kaplonek, Naeem Khan, Katrin Reppe, Benjamin Schumann, Madhu Emmadi, Marilda P Lisboa, Fei-Fei Xu, Adam DJ Calow, Sharavathi G Parameswarappa, Martin Witzenrath, et al. Improving vaccines against streptococcus pneumoniae using synthetic glycans. *Proceedings of the National Academy of Sciences*, 115(52):13353–13358, 2018.

[28] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[29] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*, 2017.

[30] Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)—round xv. *Proteins: Structure, Function, and Bioinformatics*, 2023.

[31] Shotaro Kumozaki, Kengo Sato, and Yasubumi Sakakibara. A machine learning based approach to de novo sequencing of glycans from tandem mass spectrometry spectrum. *IEEE/ACM transactions on computational biology and bioinformatics*, 12(6):1267–1274, 2015.

[32] Kyung Jin Lee, Jin-Hee Jung, Jung Mi Lee, Yangkang So, Ohsuk Kwon, Nico Callewaert, Hyun Ah Kang, Kisung Ko, and Doo-Byoung Oh. High-throughput quantitative analysis of plant n-glycan using a dna sequencer. *Biochemical and Biophysical Research Communications*, 380(2):223–229, 2009.

[33] Fuyi Li, Chen Li, Mingjun Wang, Geoffrey I Webb, Yang Zhang, James C Whisstock, and Jiangning Song. Glycomine: a machine learning-based approach for predicting n-, c-and o-linked glycosylation in the human proteome. *Bioinformatics*, 31(9):1411–1419, 2015.

[34] Haining Li, Austin WT Chiang, and Nathan E Lewis. Artificial intelligence in the analysis of glycosylation data. *Biotechnology Advances*, 60:108008, 2022.

[35] Suh-Yuen Liang, Sz-Wei Wu, Tsung-Hsien Pu, Fang-Yu Chang, and Kay-Hooi Khoo. An adaptive workflow coupled with random forest algorithm to identify intact n-glycopeptides detected from mass spectrometry. *Bioinformatics*, 30(13):1908–1916, 2014.

[36] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[37] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880, 2019.

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

[39] Eric Nguyen, Michael Poli, Matthew G Durrant, Armin W Thomas, Brian Kang, Jeremy Sullivan, Madelena Y Ng, Ashley Lewis, Aman Patel, Aaron Lou, et al. Sequence modeling and design from molecular to genome scale with evo. *bioRxiv*, pages 2024–02, 2024.

[40] Subash C Pakhrin, Kiyoko F Aoki-Kinoshita, Doina Caragea, and Dukka B Kc. Deepnglypred: a deep neural network-based approach for human n-linked glycosylation site prediction. *Molecules*, 26(23):7314, 2021.

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[42] Thejkiran Pitti, Ching-Tai Chen, Hsin-Nan Lin, Wai-Kok Choong, Wen-Lian Hsu, and Ting-Yi Sung. N-glyde: a two-stage n-linked glycosylation site prediction incorporating gapped dipeptides and pattern-based encoding. *Scientific reports*, 9(1):15975, 2019.

[43] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

[44] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[45] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15*, pages 593–607. Springer, 2018.

[46] Amir Shanehsazzadeh, David Belanger, and David Dohan. Is transfer learning necessary for protein landscape prediction? *arXiv preprint arXiv:2011.03443*, 2020.

[47] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

[48] Ghazaleh Taherzadeh, Abdollah Dehzangi, Maryam Golchin, Yaoqi Zhou, and Matthew P Campbell. Sprint-gly: predicting n-and o-linked glycosylation sites of human and mouse proteins by using sequence and predicted structural properties. *Bioinformatics*, 35(20):4140–4146, 2019.

[49] Naoyuki Taniguchi and Yasuhiko Kizuka. Glycans and cancer: role of n-glycans in cancer biomarker, progression and metastasis, and therapeutics. *Advances in cancer research*, 126: 11–51, 2015.

[50] Luc Thomès, Rebekka Burkholz, and Daniel Bojar. Glycowork: A python package for glycan data science and machine learning. *Glycobiology*, 31(10):1240–1244, 2021.

[51] Michael Tiemeyer, Kazuhiro Aoki, James Paulson, Richard D Cummings, William S York, Niclas G Karlsson, Frederique Lisacek, Nicolle H Packer, Matthew P Campbell, Nobuyuki P Aoki, et al. Glytoucan: an accessible glycan structure repository. *Glycobiology*, 27(10):915–919, 2017.

[52] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. Composition-based multi-relational graph convolutional networks. *arXiv preprint arXiv:1911.03082*, 2019.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[54] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[55] Antonio Villalobo, Aitor Nogales-González, and Hans-J Gabius. A guide to signaling pathways connecting protein-glycan interaction with the emerging versatile effector functionality of mammalian lectins. *Trends in Glycoscience and Glycotechnology*, 18(99):1–37, 2006.

[56] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*, 2015.

[57] Xi Wang, Ruichu Gu, Zhiyuan Chen, Yongge Li, Xiaohong Ji, Guolin Ke, and Han Wen. Uni-rna: universal pre-trained models revolutionize rna research. *bioRxiv*, pages 2023–07, 2023.

[58] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[59] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[60] Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Ma Chang, Runcheng Liu, and Jian Tang. Peer: a comprehensive and multi-task benchmark for protein sequence understanding. *Advances in Neural Information Processing Systems*, 35:35156–35173, 2022.

[61] Yoshihiro Yamanishi, Francis Bach, and Jean-Philippe Vert. Glycan classification with tree kernels. *Bioinformatics*, 23(10):1211–1216, 2007.

[62] Maomao Yan, Yuyang Zhu, Xueyun Liu, Yi Lasanajak, Jinglin Xiong, Jingqiao Lu, Xi Lin, David Ashline, Vernon Reinhold, David F Smith, et al. Next-generation glycan microarray enabled by dna-coded glycan library and next-generation sequencing technology. *Analytical chemistry*, 91(14):9221–9228, 2019.

[63] Xiao-Lian Zhang. Roles of glycans and glycopeptides in immune system and immune-related diseases. *Current medicinal chemistry*, 13(10):1141–1147, 2006.

[64] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.

[65] Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. *arXiv preprint arXiv:2203.06125*, 2022.

[66] Zhiqiang Zhong, Kuangyu Zhou, and Davide Mottin. Benchmarking large language models for molecule prediction tasks. *arXiv preprint arXiv:2403.05075*, 2024.

[67] Zhaocheng Zhu, Chence Shi, Zuobai Zhang, Shengchao Liu, Minghao Xu, Xinyu Yuan, Yangtian Zhang, Junkun Chen, Huiyu Cai, Jiarui Lu, et al. Torchdrug: A powerful and flexible machine learning platform for drug discovery. *arXiv preprint arXiv:2202.08320*, 2022.