Recurrent Neural Network-based Prediction of O-GlcNAcylation Sites in Mammalian Proteins

By Pedro Seber and Richard D. Braatz*
Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139
* Corresponding author

Abstract

Glycosylation can modify proteins in positive and essential ways, such as ensuring correct protein folding or increasing the potency of antibodies, or in negative ways, such as promoting cancers. In particular, O-GlcNAcylation has the potential to be an important target for therapeutics, but O-GlcNAcylation sites are not known. In spite of the importance of O-GlcNAcylation, current predictive models are insufficient as they fail to generalize, and many are no longer available. This article constructs MLP and RNN models to predict the presence of O-GlcNAcylation sites based on protein sequences. Multiple different datasets are evaluated separately and assessed in terms of strengths and issues. The models trained in this work achieve considerably better metrics than previously published models, with at least a two-fold increase in F_1 score relative to previously published models; the specific gains vary depending on the dataset. The best model achieves an F_1 score of 36.17%, multiple times greater than any previously published model and 7.6 times higher than when not using any model. We release all of the software used in this work on our GitHub (github.com/PedroSeber/O-GlcNAcylation_Prediction), allowing the reproduction of this article and facilitating future studies in the prediction of O-GlcNAcylation sites.

Author Summary

Glycosylation is a reaction in which a sugar is attached to a functional group of another molecule. This post-translational modification plays a critical role in determining protein structure, function, and stability. On the other hand, improper glycosylation or deglycosylation is associated with many diseases. O-GlcNAcylation is a type of glycosylation in which N-Acetylglucosamine is the first glycan added to an oxygen atom of a protein. A greater understanding of glycosylation could improve treatments, through drugs with novel modes of action or by improving existing therapeutics. In particular, there are no methods for predicting in advance which oxygen sites will be O-GlcNAcylated. Past studies have attempted to predict the presence of O-GlcNAcylation sites using AI models but the associated models have had poor predictive performance. We construct recurrent neural network-based models that improve predictive performance by more than a factor of 7. The software is released as open source, so that researchers can use the software to support their efforts in drug discovery, design, and manufacturing.

1 Introduction

Glycosylation is a co-translational and post-translational modification in which a glycan or glycans are added to proteins. When a glycan is added to the oxygen of an amino acid (typically serine or threonine), this process is called O-linked glycosylation. When the first glycan added is an N-Acetylglucosamine (GlcNAc), this process is called O-GlcNAcylation [23]. O-GlcNAcylation is mediated by the enzymes OGT and OGA and is important functionally and structurally [23] [4]. Conversely, incorrect glycosylation or deglycosylation is associated with multiple diseases such as cancers [25], infections [3], and congenital disorders [10].

Glycosylation has great potential for diagnoses and treatments, which may benefit the biomedical and pharmaceutical industry, physicians, and patients. For example, many forms of carcinoma show increases

in fucosylation, branching, and sialyation [2]. The vast majority of neuroblastomas express disialoganglioside; as such, neuroblastoma may be treated by anti-disialoganglioside monoclonal antibodies according to Phase I-III studies [8] [1]. Antibodies containing poly-α2,8-sialyation have increased half-lives without the introduction of tolerance problems [26]. Conversely, therapeutics may be hindered by glycans not produced by humans. An example is the immunogenic N-glycolylneuraminic acid [20], which is in some CHO-cell-derived glycoproteins [9]. Recent research has shown O-GlcNAcylation can be a powerful target for therapeutics [28], further highlighting the potential of glycosylation.

In spite of how important glycosylation is for biotherapeutics, some challenges remain. Analyses need to take into account the locations of glycosylation sites and the glycan compositions at each of these sites [2]. It is challenging to investigate specific functions of glycosylation due to the wide diversity of glycan sites and distributions [23]. The complex equipment required to determine glycosylation prevents clinical laboratories from widely analyzing patient samples, restricting personalized medicine [2].

Computational tools can aid researchers in better understanding and predicting glycosylation patterns. Models can be classifiers, such as YinOYang [6] or O-GlcNAcPRED-II [11], or regressors – that is, attempt to predict numerical values, such as the models in Refs. [19] and [24]. For example, in the context of glycosylation, classifiers may predict whether an amino acid can be N- or O-glycosylated. Regressors may quantitatively predict the glycan distribution of a glycosylation site.

In this work, we construct recurrent neural network (RNN) classification models to predict the presence of O-GlcNAc sites from mammal data, primarily protein sequences. The model construction procedures employ cross-validation and rigorous unbiased prediction error estimation. Our RNN models achieve significantly higher metrics than previously reported models. Open-source software is provided so that other researchers can reproduce the work or retrain the models as additional or higher quality O-GlcNAcylation data become available.

2 Materials and Methods

This section describes the datasets and methods for constructing the data-driven glycosylation models. Details on how to run the software to make O-GlcNAc site predictions or recreate the results in this article are provided in Supplementary Information.

2.1 Datasets

Three experimental datasets, one at a time, were used to construct the models in this work. Table 1 summarizes the size and features of each dataset. The first dataset, named "Ref. [16] – Original" in this publication, used was taken directly from Ref. [16]. This dataset contained human-selected descriptors based on sequence and structure, but did not contain the protein sequence directly. It had certain issues, such as repeated entries, which likely lead to test-set leakage. Thus, a second dataset, named "Ref. [16] – Modified" in this publication, was built from the raw data in Ref. [16]. This second dataset contained only protein sequences for site prediction. A third, larger dataset, named "Ref. [27] – Modified" in this publication, was built from the processed data from Ref. [27]. Modifications done to this dataset were removing non-mammalian proteins, removing proteins without site information, and splitting entries with multiple isoforms. Due to the presence of isoforms and homologous proteins, care was taken to not include the same sequence multiple times in the processed dataset. This selection was done based on a window size of 5 AA on each side of the central S/T (11 AA total) even for the larger windows, reducing any effects due to site similarity. Thus, test-set leakage was avoided. 20% of each dataset was separated for testing, with the remaining 80% used for cross-validation.

Table 1: Summary of the properties of the datasets used in this work. Datasets came from Refs. [16] and [27] and were modified by us when noted.

Dataset	Unique Entries	Unique Positive Entries	Features Present
Ref. [16] – Original	41,056	565 (1.38%)	Descriptors based on sequence and structure
Ref. [16] – Modified	41,600	535 (1.29%)	Site sequence (21-AA window)
Ref. [27] – Modified	558,168	13,637 (2.44%)	Site sequence (11-AA to 51-AA window)

2.2 Artificial Neural Networks (ANNs)

Multilayer perceptrons (MLPs) are constructed for the Ref. [16] – Original dataset, as it did not contain sequence information, while recurrent neural network (RNN) models are constructed for the other datasets. Model construction was done using PyTorch [21] and other Python packages [7] [22] [17]. For the MLP models, 32 different layer configurations, 4 different learning rates (10^{-2}, 5×10^{-3}, 10^{-3}, 5×10^{-4}), 3 different activation functions (ReLU, tanh, and tanhshrink), and varying loss weights for the positive class were used. For the RNN models, 2 LSTM size configurations (selected based on the number of MLP features in the original dataset), 7 different MLP layer configurations, 2 different learning rates (10^{-2}, 5×10^{-3}), 2 different activation functions (ReLU and tanhshrink), and varying loss function weights for the positive class were used.[1] Moreover, the RNNs trained with the Ref. [27] dataset used an AdamW optimizer with a weight decay parameter of $\lambda = 10^{-2}$ and cosine scheduling [15]. Cosine scheduling has been used primarily in the computer vision field and achieves great results in the context of imbalanced datasets [14] [18]. The best hyperparameters for the MLP or each RNN size are determined by a grid search, testing each combination of layers, learning rate, and activation function. The combination with the highest cross-validation average F_1 score is selected and, for each RNN size, its performance is reported for an independent test dataset.

3 Results

3.1 MLP Models Considerably Surpass Previously Published Models in Terms of Precision and F_1 Score on the Original Dataset of Ref. [16]

All of the data-driven models constructed in this study for the prediction of O-GlcNAcylation sites are trained with hyperparameters selected by cross-validation. Using the "Ref. [16] – Original" dataset, this section compares our MLP model with previously published models: YinOYang [6], O-GlcNAcPRED-II [11], OGTSite [13], and the models in Ref. [16].

The central thesis of Ref. [16] is that O-GlcNAcylation prediction models fail because no model in that article achieved a precision greater than 9%. That conclusion, however, was based on the low precision of the specific models evaluated in that reference. At low, less strict acceptance thresholds, our MLP behaves similarly to YinOYang (Fig. 1). Beginning from a threshold equal to 10^{-6} or higher, our MLP displays greater precision at the same recall level. The precision continues to increase monotonically with threshold, while the F_1 score peaks at a model threshold equal to 0.8. At its maximum F_1 score, our MLP model has a 151% improvement in the F_1 score and 307% in precision relative to the best former models, with a total precision of 35.3% (Table 2). As such, our MLP model shows that the central thesis of Ref. [16] is not valid and that predictive models of the location of O-GlcNAcylation sites can be constructed with reasonable precision.

After the MLP was fully trained, we noticed that the "Ref. [16] – Original" dataset has some issues, such as entries that did not match the raw data and repeated entries, which may lead to test-set

[1] As detailed in the `ANN_train.ipynb` and `ANN_train.py` files in the GitHub repository.

leakage and overoptimistic predictions. Moreover, the models made predictions based on human-selected descriptors, which may be incomplete or biased and are not trivial to obtain, making model usage inconvenient for the end-user. To remedy these issues, we constructed a new dataset from the raw data of Ref. [16], which is called "Ref. [16] – Modified" in this work and analyzed in the next section.

Table 2: Recall, precision, and F_1 score metrics (in %) for the points with highest F_1 score for different O-GlcNAcylation site prediction models. Model "YinOYang" is from Ref. [6], Model "O-GlcNAcPRED-II" is from Ref. [11], Model "OGTSite" is from Ref. [13], and Model "Mauri" is the best model from a collection of models from Ref. [16]. The metrics for these previously published models are from Tables 2 and 4 of Ref. [16]. Model "Our MLP" comes from this work. "No Model" sets all sites to potentially positive – that is, sets the threshold to 0.

Metric	YinOYang (From [6])	O-GlcNAcPRED-II (From [11])	OGTSite (From [13])	Mauri (From [16])	Our MLP (This Work)	No Model
Recall (%)	14.36	65.09	49.09	39.82	22.64	100
Precision (%)	6.89	3.90	6.20	3.10	35.29	1.38
F_1 Score (%)	9.31	7.36	11.01	5.75	27.58	2.72

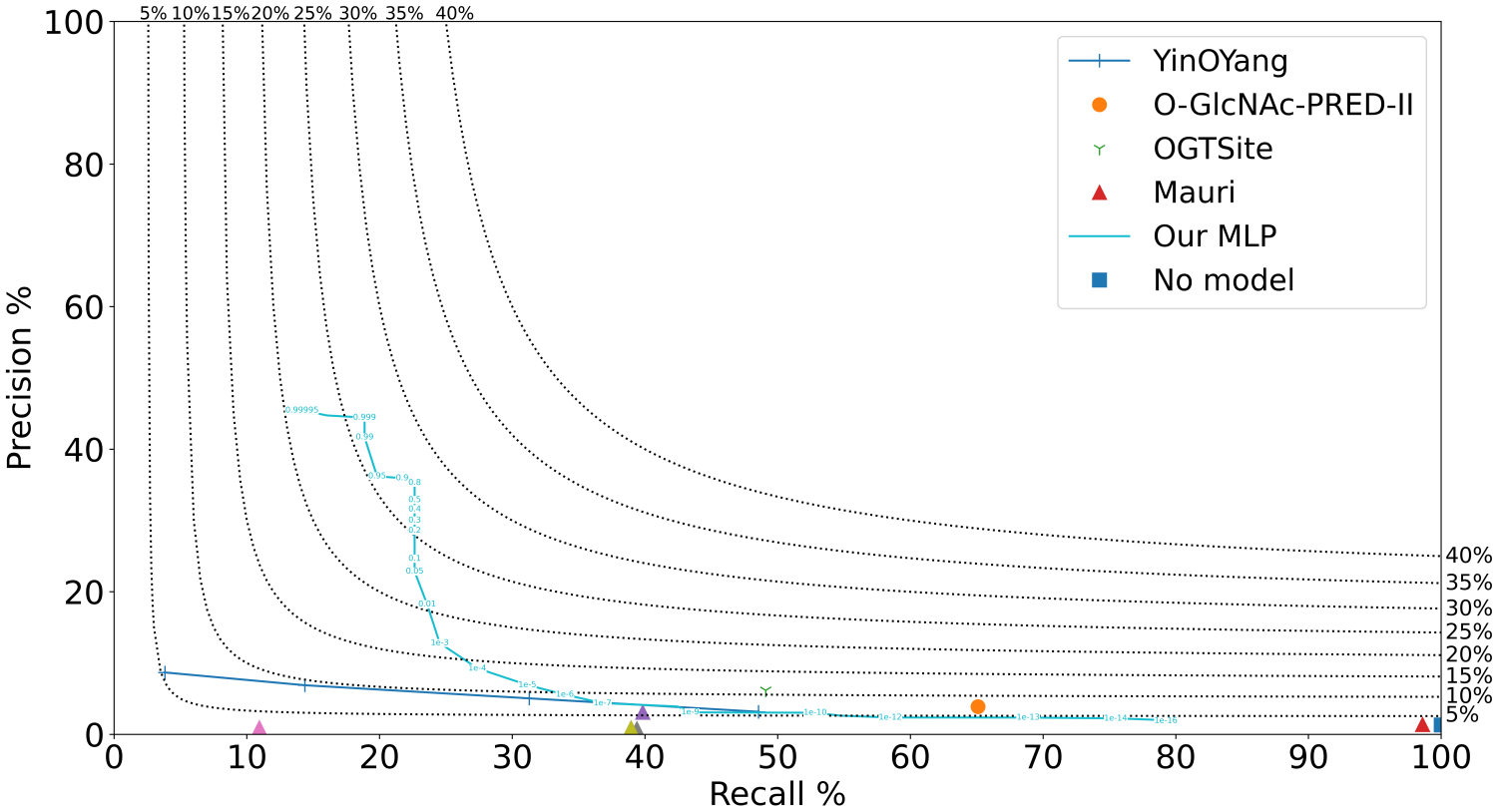

Figure 1: Precision-recall curves for different O-GlcNAcylation site prediction models tested on the original dataset of Ref. [16]. The black dotted lines are isolines of the F_1 score, with the corresponding F_1 scores on the top and right sides of the plot. Model "YinOYang" is from Ref. [6], Model "O-GlcNAcPRED-II" is from Ref. [11], and Model "OGTSite" is from Ref. [13] (complete precision-recall curves are not shown because the models are no longer publicly available). Models "Mauri" are a collection of models from Ref. [16]. Metrics for these previously published models are from Tables 2 and 4 of Ref. [16]. Model "Our MLP" comes from this work; the numbers on its curve represent the minimum threshold for a site to be considered positive. "No Model" sets all sites to potentially positive – that is, sets the threshold to 0.

3.2 RNNs Considerably Surpass Previous Models for the Dataset of Ref. [16] Modified to Prevent Leakage

This section compares our RNN models trained on the "Ref. [16] – Modified" dataset with YinOYang [6]. Our RNNs behave similarly to YinOYang at low thresholds (Fig. 2). Beginning from a threshold of 10^{-10} for the RNN-76 model and 10^{-7} for the RNN-152 model,[2] our RNNs display greater precision at the same recall level. The precision for both RNN models continues to increase nearly monotonically with increasing threshold, while the F_1 score peaks at a threshold of 0.99 for RNN-76 and 0.999 for RNN-152. Moreover, the RNN-76 is strictly superior to the RNN-152 for all thresholds $\geq 10^{-10}$. At its F_1 score maximum, the RNN-76 model has a 134% improvement in the F_1 score and 391% in precision relative to YinOYang (Table 3).

The performance of these models is lower than for the models in Section 3.1, which occurs due to the elimination of test-set data leakage, which biased the metrics upwards.[3] This reduction highlights the importance of constructing test datasets in a manner that avoids the potential for information leakage, when constructing machine learning-based models, to be able to produce accurate assessments of model performance [5] [12].

Table 3: Recall, precision, and F_1 score metrics (in %) for the points with highest F_1 score for different O-GlcNAcylation site prediction models. Model "YinOYang" is from Ref. [6]. Metrics for the YinOYang model were obtained by us. As mentioned above, other previously published models are not available; thus, their performance on this dataset could not be obtained. Models "Our RNN" come from this work. The number that follows each of our RNN models is the LSTM module size. "No Model" sets all sites to potentially positive – that is, sets the threshold to 0.

Metric	YinOYang (From [6])	Our RNN-76 (This Work)	Our RNN-152 (This Work)	No Model
Recall (%)	21.50	15.31	13.27	100
Precision (%)	4.93	24.19	24.53	1.29
F_1 Score (%)	8.02	18.75	17.22	2.55

[2] RNNs with 38 and 228 neurons were also trained and tested, but had cross-validation results inferior to both RNN-76 and RNN-152 and so are not included here.

[3] This drop in performance was not due to the change in features, as tests done on MLPs using this dataset and the same features showed that the new MLPs lost more than 10% of their absolute recall values at the same precision and performed worse than the RNNs (Fig. S1).

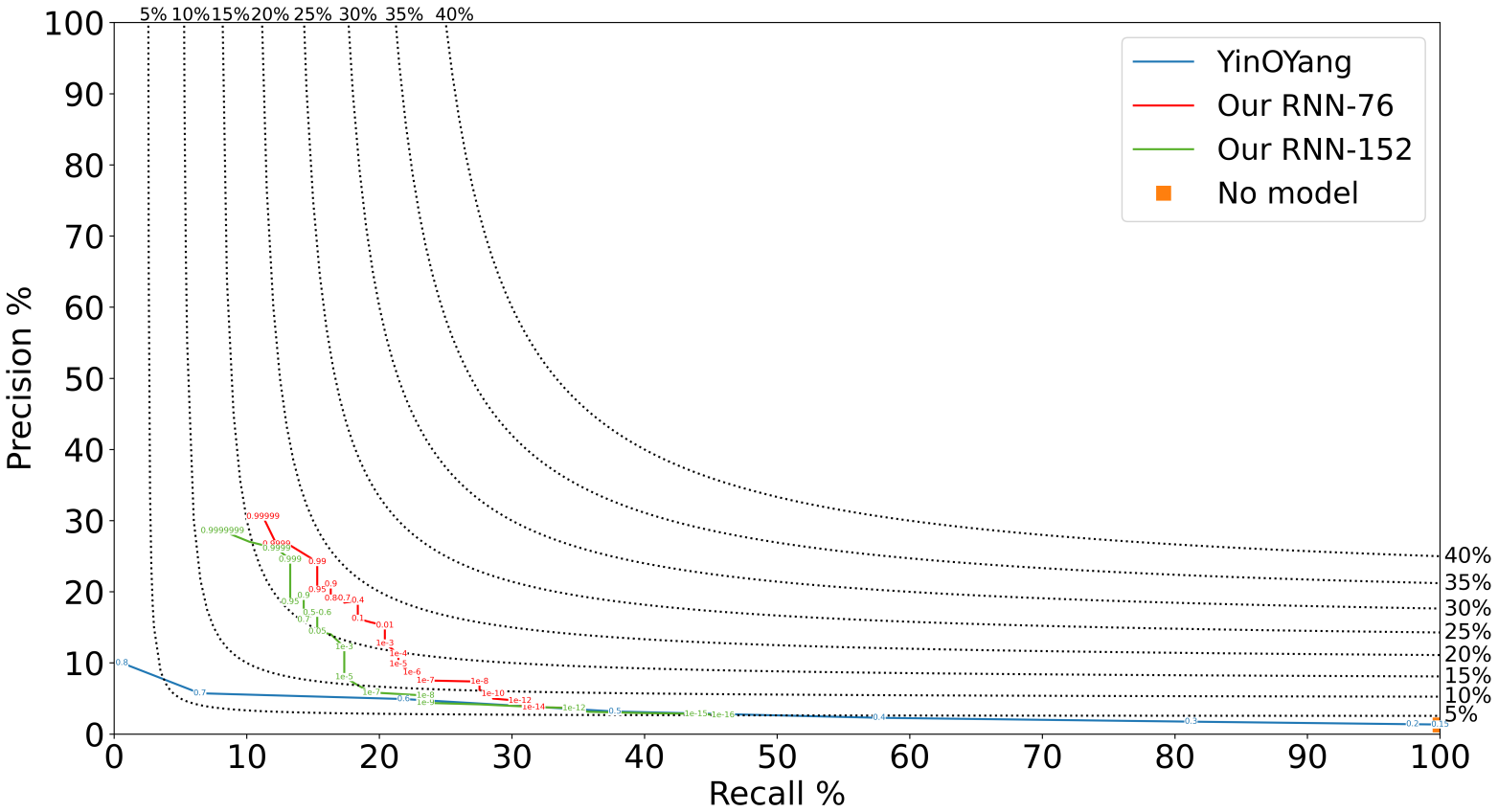

Figure 2: Precision-recall curves for different O-GlcNAcylation site prediction models tested on the modified dataset of Ref. [16]. The black dotted lines are F_1 score isolines, and their corresponding F_1 scores are on the top and right sides of the figure. Model "YinOYang" is from Ref. [6]. Metrics for the YinOYang model were obtained by us. As mentioned above, other previously published models are not available; thus, their performance on this dataset could not be obtained. Models "Our RNN" come from this work. The number that follows each of our RNN models is the LSTM module size used. The numbers on each curve represent the minimum threshold for a site to be considered positive. "No Model" sets all sites to potentially positive – that is, sets the threshold to 0.

3.3 A Larger, Less Imbalanced Dataset Leads to Improved Models

After model training with the Ref. [16] datasets was complete, the dataset from Ref. [27] was located, which is more than an order of magnitude larger than the previously used dataset. Moreover, it contained a larger proportion of positive sites (2.44%), although the dataset was still significantly imbalanced.

Using a slightly modified training procedure (as described in Section 2.2), RNN models are trained on a modified version of the Ref. [27] dataset. Ref. [27] also included a potential motif for O-GlcNAcylation, which was tested on this modified dataset. YinOYang is not compared using this dataset due to its input limitations, as testing this dataset would require over 50 different calls to the online tool. However, this removal allows the testing of multiple window sizes, as YinOYang forces a window size of 10 amino acids on each side of the S/T (21 AA total) which may not produce the highest prediction capability model.

Our RNNs exhibit a much higher recall at the same precision level (Fig. 3) than any models in the previous sections (Fig. 2). Moreover, these RNNs have much higher precision for all recall values lower than 50%. As before, the precision for our RNN models increases monotonically with increasing threshold, while the F_1 score peaks at different thresholds for each model. The performance of the models increases

with increasing window size for sizes up to 20, and it also increases with increasing RNN hidden sizes for models with up to 225 neurons [4] (Fig. 3 and Table 4). A summary of metrics for the Ref. [27] motif and the point for each of our RNNs with the highest F_1 score is in Table 4.

The previously proposed motif is excessively restrictive and fails to adequately capture the sequence needed for O-GlcNAcylation. Out of the 531,628 unique entries in the dataset, only 28 follow the motif, and only 11 out of those 28 are truly positive. Allowing sequences to be treated as positive if at least 4 out of 5 amino acids follow the motif does not significantly improve the results: only 669 sites follow the motif, and only 136 out of those 669 are truly positive. The motif becomes slightly better when sequences are treated as positive if at least 3 out of 5 amino acids follow the motif; 7,787 sites follow the motif and 923 of those are true positives. If sequences are treated as positive if at least 2 out of 5 amino acids follow the motif, there are 54,875 positive sites and 3,232 true positives. Finally, treating sequences with just 1 out of 5 amino acids as positive leads to 233,943 positive sites and 7,898 true positives. Nevertheless, the motif fails to achieve an F_1 score greater than 10% (Table 4).

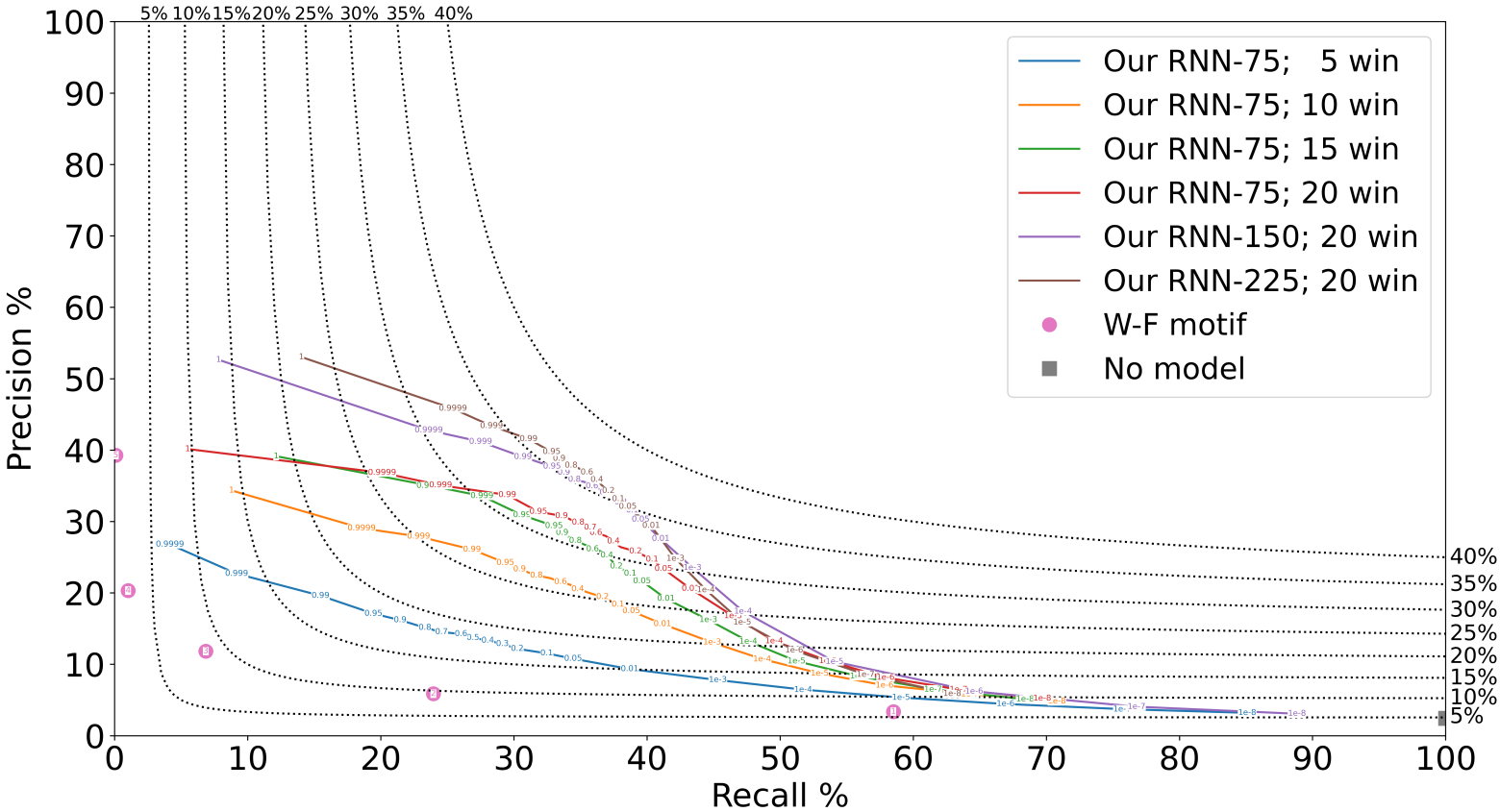

Figure 3: Precision-recall curves for O-GlcNAcylation site prediction models tested using a modified version of Ref. [27]'s dataset. The black dotted lines are F_1 score isolines, and their corresponding F_1 scores are on the top and right sides of the figure. Models "Our RNN-#" come from this work. The number that follows each RNN model is the LSTM module size used; the number after the semicolon represents the window size on each side of the central S/T. The numbers on the "Our RNN-#" curves represent the minimum threshold for a site to be considered positive. The "W-F motif" comes from Ref. [27]; the numbers in the circles represent the minimum number of motif matches for a site to be considered positive. "No Model" sets all sites to potentially positive – that is, sets the threshold to 0.

[4] Models with window sizes equal to 25 performed nearly identically to models with window sizes equal to 20, so the former are not included in Fig. 3 or Table 4. Models with RNN sizes equal to 300 neurons performed nearly identically to models with 150 or 225 neurons, and so are also not shown.

Table 4: Recall, precision, and F_1 score metrics (in %) for the point with highest F_1 score for an RNN trained on data from Ref. [27], and for the motif generated by Ref. [27]. Models "RNN-#" come from this work. The number that follows each RNN model is the LSTM module size used; the number after the semicolon represents the window size on each side of the central S/T. The "W-F Motifs" come from Ref. [27]. The number in parentheses represents the minimum number of amino acids (out of 5) that must follow the motif for a site to be considered positive. "No Model" sets all sites to potentially positive – that is, sets the threshold to 0.

Metric	RNN-75; 5 win (This Work)	RNN-75; 10 win (This Work)	RNN-75; 15 win (This Work)	RNN-75; 20 win (This Work)
Best Threshold	0.90	0.95	0.95	0.80
Recall (%)	21.44	29.35	33.00	34.82
Precision (%)	16.20	24.22	29.38	29.87
F_1 Score (%)	18.45	25.95	31.09	32.16

Metric	RNN-150; 20 win (This Work)	RNN-225; 20 win (This Work)	No Model
Best Threshold	0.50	0.60	N/A
Recall (%)	36.36	35.47	100
Precision (%)	34.58	36.90	2.44
F_1 Score (%)	35.45	36.17	4.76

Metric	W-F Motif (5) (From [27])	W-F Motif (4) (From [27])	W-F Motif (3) (From [27])	W-F Motif (2) (From [27])	W-F Motif (1) (From [27])
Recall (%)	0.08	1.01	6.84	23.94	58.50
Precision (%)	39.29	20.33	11.85	5.89	3.38
F_1 Score (%)	0.16	1.92	8.67	9.45	6.38

4 Discussion

This work constructs MLP and RNN models from multiple sources of literature data on protein O-GlcNAcylation based on human-selected descriptors (Section 3.1) or protein sequences (Sections 3.2 and 3.3). An MLP model trained in this work was compared with previously published models (as reported in Ref. [16]), and this work's MLP model surpasses the previously published models in precision (307% improvement) and F_1 score (151% improvement) metrics (Section 3.1). This study contrasts with the past studies by our use of different model architectures, multiple prediction thresholds, and rigorous cross-validation for hyperparameter selection. An analysis of the dataset of Ref. [16] found multiple issues, however, including test-set leakage, making the results overoptimistic for all models.

To address the dataset issues, a new dataset is constructed from the original dataset of Ref. [16]. This new dataset contains protein sequences instead of descriptors, promoting the training of RNN models and simplifying overall usage for the end-users. Two RNN models trained in this work are compared with YinOYang [6], the only model evaluated in Ref. [16] that is still available (Section 3.2). The correction of the dataset's issues lowered the overall performance of the models tested. Nevertheless, the best RNN model still displayed a 391% improvement in precision and 134% improvement in F_1 score when compared to YinOYang, and the RNN models surpassed an MLP model trained on the same data. This comparison suggests that the RNN model architecture is better suited for prediction of O-GlcNAcylation than the MLP model architecture, which is consistent with the RNN being able to leverage the dynamic data structure of protein site sequences instead of the MLP's limited way of handling structured data.

Finally, a much larger dataset from Ref. [27] is refined and used to train RNN models through a slightly modified methodology. These RNN models displayed superior performance when compared to the other models in this or previous works (Section 3.3). We hypothesize part of this difference is due to the greater number of entries and lower data imbalance found in Ref. [27]'s modified dataset. The use of weight decay and learning rate scheduling contributed to further improving the performance of the RNN models. Ref. [27] also proposes a motif for O-GlcNAcylation, but that motif is excessively restrictive, and so obtaining an O-GlcNAcylation motif remains an open problem.

This software used in this work is publicly available,[5] allowing other researchers to reproduce this work and reuse or improve the code in future studies. The software provides a simple way to install and run the best RNN model (trained on Ref. [27]'s modified data and using a window size of 20 AAs on each side of the central S/T) to predict the presence of O-GlcNAcylation sites based on the local protein sequence.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability statement

The raw data, code used to process these data, and models underlying this article are available in a GitHub repository at https://github.com/PedroSeber/O-GlcNAcylation_Prediction

[5] https://github.com/PedroSeber/O-GlcNAcylation_Prediction.

Appendix

Reproducing the models and plots

The models can be recreated by downloading the datasets and opening the `ANN_train.ipynb` file and running the notebook. The first cell contains default values for the hyperparameters, which can be changed by the user. Alternatively, the `ANN_train.py` file can be run with the appropriate flags (run `python ANN_train.py --help` for details).

The plots can be recreated by running the `make_plot.py` file with the appropriate data version as an input (`python make_plot.py v1` for Fig. 1, `python make_plot.py v3` for Fig. 2, and `python make_plot.py v5` for Fig. 3).

Using the RNN model to predict O-GlcNAcylation sites

The Conda environment defining the specific packages and version numbers used in this work is available as `ANN_environment.yaml` on our GitHub (github.com/PedroSeber/O-GlcNAcylation_Prediction). To use our trained model, run the `Predict.py` file as `python Predict.py <sequence> -t <threshold> -bs <batch_size>`.

Alternatively, create an (N+1)x1 .csv with the first row as a header (such as "Sequences") and all other N rows as the actual amino acid sequences, then run the Predict.py file as `python ANN_predict.py <path/to/file.csv> -t <threshold> -bs <batch_size>`. The results will be saved as a new .csv file.

Supplemental Figures

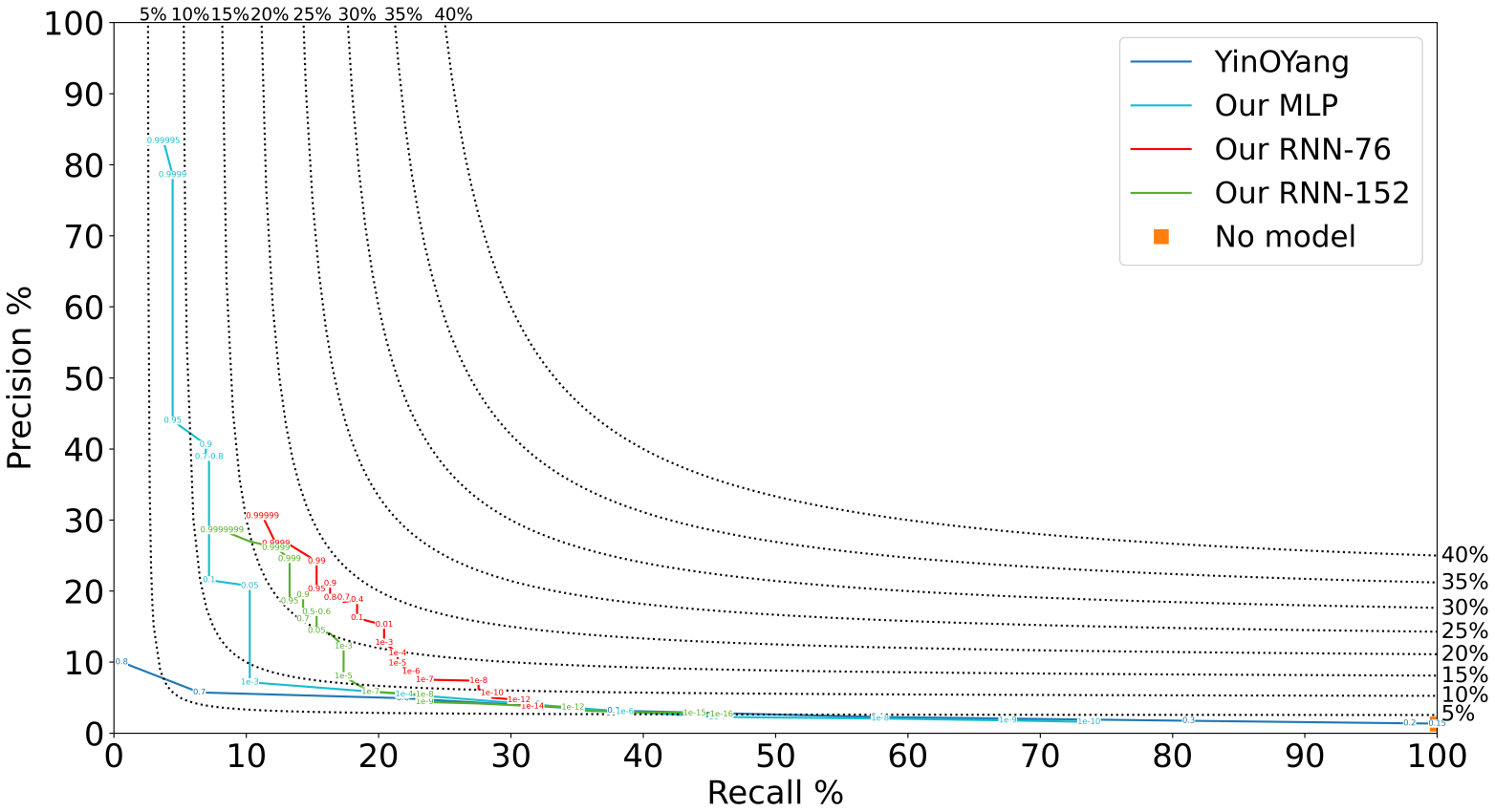

Figure S1: Precision-recall curves for different O-GlcNAcylation site prediction models. The black dotted lines are F_1 score isolines, and their corresponding F_1 scores are on the top and right sides of the figure. Model "YinOYang" is from Ref. [6]. Metrics for the YinOYang model were obtained by us. As mentioned in Section 3.2, other previously published models are not available; thus, their performance on this dataset could not be obtained. Models "Our MLP" and "Our RNN" come from this work. The number that follows each of our RNN models is the LSTM module size used. The numbers on each curve represent the minimum threshold for a site to be considered positive. "No Model" sets all sites to potentially positive – that is, sets the threshold to 0.

REFERENCES

[1] Mahiuddin Ahmed and Nai-Kong V. Cheung. Engineering anti-GD2 monoclonal antibodies for cancer immunotherapy. *FEBS Letters*, 588(2):288–297, 2014.

[2] Andreia Almeida and Daniel Kolarich. The promise of protein glycosylation for personalised medicine. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1860(8):1583–1595, 2016. Glycans in personalised medicine.

[3] Aadil Hussain Bhat, Sudipa Maity, Kuldeep Giri, and Kiran Ambatipudi. Protein glycosylation: Sweet or bitter for bacterial pathogens? *Critical Reviews in Microbiology*, 45(1):82–102, 2019. PMID: 30632429.

[4] Yi-Hsuan Chang, Chia-Lin Weng, and Kuo-I Lin. O-GlcNAcylation and its role in the immune system. *Journal of Biomedical Science*, 27(1):57, 2020.

[5] Alexis Geslin et al. Battery lifetime predictions: information leakage from unblinded training. *ChemRxiv*, 2023.

[6] Ramneek Gupta and Søren Brunak. Prediction of glycosylation across the human proteome and the correlation to protein function. *Pacific Symposium on Biocomputing*, pages 310–22, 2002.

[7] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[8] Wan-Ling Ho, Wen-Ming Hsu, Min-Chuan Huang, Kenji Kadomatsu, and Akira Nakagawara. Protein glycosylation in cancers and its potential therapeutic applications in neuroblastoma. *Journal of Hematology & Oncology*, 9(1):100, 2016.

[9] Cornelis H. Hokke, Aldert A. Bergwerff, Gijs W. K. Dedem, Johannis P. Kamerling, and Johannes F. G. Vliegenthart. Structural analysis of the sialylated N- and O-linked carbohydrate chains of recombinant human erythropoietin expressed in Chinese hamster ovary cells. Sialylation patterns and branch location of dimeric N-acetyllactosamine units. *European Journal of Biochemistry*, 228(3):981–1008, 1995.

[10] J. Jaeken. Chapter 179 – congenital disorders of glycosylation. In Olivier Dulac, Maryse Lassonde, and Harvey B. Sarnat, editors, *Pediatric Neurology Part III*, volume 113 of *Handbook of Clinical Neurology*, pages 1737–1743. Elsevier, Amsterdam, 2013.

[11] Cangzhi Jia, Yun Zuo, and Quan Zou. O-GlcNAcPRED-II: an integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. *Bioinformatics*, 34(12):2029–2036, 2018.

[12] David T. Jones. Setting the standards for machine learning in biology. *Nature Reviews Molecular Cell Biology*, 20:659–660, 11 2019.

[13] Hui-Ju Kao, Chien-Hsun Huang, Neil Arvin Bretaña, Cheng-Tsung Lu, Kai-Yao Huang, Shun-Long Weng, and Tzong-Yi Lee. A two-layered machine learning method to identify protein O-GlcNAcylation sites with O-GlcNAc transferase substrate motifs. *BMC Bioinformatics*, 16(S18):S10, 2015.

[14] Anna Kukleva, Moritz Böhle, Bernt Schiele, Hilde Kuehne, and Christian Rupprecht. Temperature schedules for self-supervised contrastive methods on long-tail data. *arXiv:2303.13664*, 2023.

[15] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv:1608.03983*, 2017.

[16] Theo Mauri, Laurence Menu-Bouaouiche, Muriel Bardor, Tony Lefebvre, Marc F Lensink, and Guillaume Brysbaert. O-glcnacylation prediction: An unattained objective. *Advances and Applications in Bioinformatics and Chemistry*, Volume 14:87–102, 2021.

[17] Wes McKinney. Data structures for statistical computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56–61, 2010.

[18] Sourav Mishra, Toshihiko Yamasaki, and Hideaki Imaizumi. Improving image classifiers for small datasets by learning rate adaptations. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pages 1–6, 2019.

[19] Sooyeon Moon, Sourav Chatterjee, Peter H. Seeberger, and Kerry Gilmore. Predicting glycosylation stereoselectivity using machine learning. *Chemical Science*, 12(8):2931–2939, 2021.

[20] Vered Padler-Karavani, Hai Yu, Hongzhi Cao, Harshal Chokhawala, Felix Karp, Nissi Varki, Xi Chen, and Ajit Varki. Diversity in specificity, abundance, and composition of anti-Neu5Gc antibodies in normal humans: Potential implications for disease. *Glycobiology*, 18(10):818–830, 2008.

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[23] Katrine T. Schjoldager, Yoshiki Narimatsu, Hiren J. Joshi, and Henrik Clausen. Global view of human protein glycosylation pathways and functions. *Nature Reviews Molecular Cell Biology*, 21(12):729–749, 2020.

[24] Pedro Seber and Richard D. Braatz. Linear and neural network models for predicting N-glycosylation in Chinese Hamster Ovary cells based on B4GALT levels. *bioRxiv*, 2023.

[25] Sean R. Stowell, Tongzhong Ju, and Richard D. Cummings. Protein glycosylation in cancer. *Annual Review of Pathology: Mechanisms of Disease*, 10(1):473–510, 2015. PMID: 25621663.

[26] Linde Van Landuyt, Chiara Lonigro, Leander Meuris, and Nico Callewaert. Customized protein glycosylation to improve biopharmaceutical function and targeting. *Current Opinion in Biotechnology*, 60:17–28, 2019.

[27] Eugenia Wulff-Fuentes, Rex R. Berendt, Logan Massman, Laura Danner, Florian Malard, Jeet Vora, Robel Kahsay, and Stephanie Olivier-Van Stichelen. The human o-glcnacome database and meta-analysis. *Scientific Data*, 8(1):25, 2021.

[28] Yi Zhu and Gerald W. Hart. Targeting O-GlcNAcylation to develop novel therapeutics. *Molecular Aspects of Medicine*, 79:100885, 2021.