Computers and Chemical Engineering 189 (2024) 108818

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/cace

Recurrent neural network-based prediction of O-GlcNAcylation sites in mammalian proteins

Pedro Seber, Richard D. Braatz *

Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, United States of America

ARTICLE INFO

Dataset link: https://github.com/PedroSeber/O-GlcNAcylation_Prediction

Keywords:
Glycosylation
O-GlcNAcylation
Computational biology
Machine learning
Deep learning
Recurrent neural networks

ABSTRACT

O-GlcNAcylation has the potential to be an important target for therapeutics, but a motif or an algorithm to reliably predict O-GlcNAcylation sites is not available. Current predictive models are insufficient as they fail to generalize, and many are no longer available. This article constructs recurrent neural network models to predict O-GlcNAcylation sites based on protein sequences. Different datasets are evaluated separately and assessed in terms of strengths and issues. Within a given dataset, results are robust to changes in cross-validation and test data as determined by nested validation. The best model achieves an F_1 score of 36% (more than 3.5-fold greater than the previous best model) and a Matthews Correlation Coefficient of 35% (more than 4.5-fold greater than the previous best model), and, for the F_1 score, 7.6-fold higher than when not using any model. Shapley values are used to interpret the model's predictions and provide biological insight into O-GlcNAcylation.

1. Introduction

Glycosylation is a co- 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 glycan added is an N-Acetylglucosamine (GlcNAc), this process is called O-GlcNAcylation (Schjoldager et al., 2020). Unlike other forms of glycosylation, O-GlcNAcylation does not form complex glycans and may be viewed similarly to phosphorylation (Schjoldager et al., 2020). O-GlcNAcylation is mediated by the enzymes OGT and OGA and is important functionally and structurally (Schjoldager et al., 2020; Chang et al., 2020). Recent research has shown O-GlcNAcylation can be a powerful target for therapeutics (Zhu and Hart, 2021), further highlighting its potential. However, it is challenging to investigate specific functions of O-GlcNAcylation, due to the wide diversity of glycosylation sites (Schjoldager et al., 2020). Conversely, incorrect or loss of O-GlcNAcylation is associated with multiple conditions such as cancers and metastases (Shi et al., 2022), infections (Chang et al., 2020), and heart failure (Umapathi et al., 2021).

In spite of how important O-GlcNAcylation is for human health and biotherapeutics, some challenges remain. As O-GlcNAcylation involves only a single glycan, the main open problems involve predicting where and when an amino acid will be O-GlcNAcylated, and computational tools can aid researchers in better understanding and predicting these phenomena. Models can be classifiers, such as YinOYang (YoY) (Gupta and Brunak, 2002) or O-GlcNAcPRED-II (Jia et al., 2018), or regressors, such as the models in Moon et al. (2021), Seber and Braatz (2023), and Liang et al. (2020). In the context of glycosylation, classifiers may predict whether an amino acid can be O-GlcNAcylated. Regressors may quantitatively predict the glycan distribution of a glycosylation site, and are better suited for predicting distributions of other forms of glycosylation. Predicting the location of O-GlcNAcylation is challenging for multiple reasons, including a low frequency of events (only about 2% of S/T sites are O-GlcNAcylated) and a lack of a motif to guide predictive efforts. The effects of neighboring amino acids likely influence whether an S/T is O-GlcNAcylated, yet not all machine learning architectures are equipped to take into account and leverage this information. Models to predict the presence of O-GlcNAcylation sites have insufficient performance to be helpful tools. A 2021 review (Mauri et al., 2021) found that no published model can achieve a precision $\geq$ 9% on a medium-sized independent dataset, indicating O-GlcNAcylation prediction models fail to generalize successfully despite the high metrics that they can achieve in their respective training datasets. These models also have low F_1 scores and Matthew Correlation Coefficients (MCCs), further indicating their performance is lacking (refer to Section 2.3 for details on these metrics, including why MCC is the best metric). More recent works, which were published after this work was first posted on a preprint server, include LM-OGlcNAc-Site (Pokharel et al., 2023) and O-GlcNAcPRED-DL (Hu et al., 2024).

* Corresponding author.
E-mail address: braatz@mit.edu (R.D. Braatz).

https://doi.org/10.1016/j.compchemeng.2024.108818
Received 30 March 2024; Received in revised form 23 July 2024; Accepted 24 July 2024
Available online 26 July 2024

Although the models in these works perform slightly better than the models evaluated in Mauri et al. (2021) as per their reported MCCs, the performance of these models is still lacking. Both these models were also trained by undersampling from the more-frequent negative class, an improper procedure that inflates training metrics but reduces the generalization performance of a model. Moreover, Pokharel et al. (2023) and Hu et al. (2024) do not include any information on the precision of their models, potentially due to subpar performance in this challenging metric. Backcalculations using these models' own published data show that the precision of Pokharel et al. (2023)'s model on an independent test set (Table 4 of that work) is only 5.42%, and that the precisions of Hu et al. (2024)'s models on independent test sets ("Ind_H_2022" and "Ind_M_2022", Table 4 of that work) are only 1.88% and 3.01% respectively. As such, the central thesis of Mauri et al. (2021) still holds: no published model can achieve a precision ≥ 9% on an independent test set.

In this work, we construct recurrent neural network (RNN) classification models (specifically, we construct bidirectional long short-term memory (LSTM) models) to predict the presence of O-GlcNAcylation sites from mammalian protein sequence data. This method is simpler and less computationally expensive than the ensemble of large-language models used in Pokharel et al. (2023) and the convolutional neural network + LSTM used in Hu et al. (2024). One of the large-language models in Pokharel et al. (2023) also has a non-commercial license, preventing its use in the biopharmaceutical industry. The model construction procedures employ cross-validation and rigorous unbiased prediction error estimation. Our final RNN model achieves significantly higher metrics than previously reported models, and its predictions are interpreted through Shapley values. Open-source software is provided so that other researchers can reproduce the work, retrain models as additional or higher quality O-GlcNAcylation data become available, and use the model to further improve the understanding of O-GlcNAcylation.

2. Materials and methods

2.1. Datasets

Three experimental datasets, one at a time, are used to construct the models. Table 1 summarizes the size and features of each dataset. The first dataset, named "Mauri et al. (2021) – Original" in this publication, is taken directly from Mauri et al. (2021). This dataset contains human-selected descriptors based on sequence and structure, but does not contain the protein sequence directly, and has certain issues, such as repeated entries, which likely lead to test-set leakage. Thus, a second dataset, named "Mauri et al. (2021) – Modified" in this publication, is built from the raw data in Mauri et al. (2021). This second dataset contains only protein sequences for site prediction. Both of the above datasets are included solely to compare the performance of this publication's methodology with that from older publications included in Mauri et al. (2021). These first two datasets are not used in the training or testing of the final model because these datasets are smaller and less complete than the third dataset. A third, larger dataset, named "Wulff-Fuentes et al. (2021) – Modified" in this publication, is built from the processed data from Wulff-Fuentes et al. (2021). S/T residues not marked as O-GlcNAcylated in the original dataset were treated as negative during preprocessing. The dataset was modified to remove non-mammalian proteins, remove proteins without site information, and split 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. All sequences are then one-hot encoded. 20% of each dataset is separated for testing, with the remaining 80% used for cross-validation with five folds. Thus, test-set leakage is avoided. To ensure robustness against variations in the data and avoid biases due to our selection of training and testing sets, a 5-fold nested validation is performed with the best model in this work. In each round of nested validation, 20% of the data are separated as the round's test set. The other 80% are used in five-fold cross-validation for hyperparameter selection (as above). The best model is evaluated against that round's test set. This preprocessing procedure was done using our own code and facilitated by standard Python packages (Harris et al., 2020; Pedregosa et al., 2011; McKinney, 2010).

A fourth set of datasets, named "Ind_H_2022" and "Ind_M_2022", is used solely to compare our model trained with the third dataset and the models from Hu et al. (2024). These two datasets were created by Hu et al. (2024) and used to test their models in that work. To ensure the fairest comparison possible, we also test our best model (trained with the third dataset) on these same datasets.[1]

2.2. Artificial Neural Networks (ANNs)

Multilayer perceptrons (MLPs) are constructed for the Mauri et al. (2021) – Original dataset, as it did not contain sequence information, and recurrent neural network (RNN) models are constructed for the other datasets. A visual diagram of these models is available in Fig. S1. Model construction was done using PyTorch (Paszke et al., 2019). 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.[2] Moreover, the RNNs trained with the Wulff-Fuentes et al. (2021) dataset used an AdamW optimizer with a weight decay parameter of $\lambda = 10^{-2}$ and cosine scheduling (Loshchilov and Hutter, 2017). Cosine scheduling has been used primarily in the computer vision field and achieves great results in the context of imbalanced datasets (Kukleva et al., 2023; Mishra et al., 2019). 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. To interpret the model's predictions, Shapley values are calculated using the shap Python package (Shapley, 1951; Lundberg and Lee, 2017). These interpretable values are also evaluated against the same test set.

2.3. Model evaluation metrics

Binary classification models emit two types of predictions. Because the real data have two potential categories, there are a total of four categories in which a prediction may fall: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). There are multiple metrics that combine some or all of these categories to assess the quality of a model. The simplest of these metrics is accuracy (Eqn. S1), defined as the number of correct guesses divided by the total guesses. However, accuracy is not a suitable metric for imbalanced datasets, as it is possible to achieve high accuracy by simply always predicting the majority class. To correct this issue, two other simple metrics can be used. Recall (aka sensitivity, hit rate, or true positive rate) is the number of true positives divided by all positives in the real data (Eqn. S2). Precision (aka positive predictive value) is the number of true positives divided by all elements classified as positive

[1] This fourth set of datasets is not included in Table 1 because it is not used to train any of our models.

[2] As detailed in the `ANN_train.py` file in the GitHub repository.

Table 1
Summary of the properties of the datasets used in this work for the training of different models. Datasets came from Mauri et al. (2021) and Wulff-Fuentes et al. (2021) and are modified by us when noted.

Dataset	Unique entries	Unique positive entries	Features present
Mauri et al. (2021) – Original	41,056	565 (1.38%)	Descriptors based on sequence and structure
Mauri et al. (2021) – Modified	41,600	535 (1.29%)	Site sequence (21 AA per entry)
Wulff-Fuentes et al. (2021) – Modified	558,168	13,637 (2.44%)	Site sequence (11 AA to 51 AA per entry)

by the model (Eqn. S3). Because recall and precision are opposed to each other, the F_1 score metric was created to balance both and allow assessment of models with a single metric (Eqn. S4). The negative-class equivalents of recall and precision are the specificity (aka selectivity or true negative rate) and the negative predictive value; however, these metrics are not used in this work because it focuses on the discovery of positive O-GlcNAcylation sites. Moreover, the negative imbalance of the datasets makes these last two metrics less suitable for evaluation.

The most used metric to assess binary classification models in multiple fields is the area under the receiver operating characteristic curve (ROC-AUC) (Chicco and Jurman, 2023). However, the ROC-AUC metric suffers from many issues and can lead to overoptimistic and incorrect assessments, especially when working with negatively-imbalanced data (Chicco and Jurman, 2023; Halligan et al., 2015; Lobo et al., 2008). To analyze the quality of binary classification models over multiple thresholds, it is recommended to use the precision–recall (P-R) curve instead (Chicco and Jurman, 2023; Saito and Rehmsmeier, 2015; Chicco, 2017; Ozenne et al., 2015). Finally, to capture in a single metric all four categories in which a prediction may fall, the Matthews correlation coefficient (MCC), also called the (Yule) phi coefficient, may be used (Eqn. S5). It is widely considered the best single metric (Chicco and Jurman, 2023). As a correlation coefficient, the MCC falls between −1 and 1 instead of the typical 0 and 1.

3. Results

3.1. MLP models considerably surpass previously published models in terms of precision, F_1 score, and MCC on the original dataset of Mauri et al. (2021)

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 "Mauri et al. (2021) – Original" dataset, this section compares our MLP model with previously published models: YinOYang (Gupta and Brunak, 2002), O-GlcNAcPRED-II (Jia et al., 2018), OGTSite (Kao et al., 2015), and the models in Mauri et al. (2021).

The central thesis of Mauri et al. (2021) is that O-GlcNAcylation prediction models fail because no model reviewed in that article achieved a precision greater than 9%. While that conclusion was based on the low precision of the specific models evaluated in that reference, models published years later still suffer from the same problem. 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, a 307% in precision relative to the best former models (with a total precision of 35.3%), and a 90.9% improvement in MCC (Table 2). As such, our MLP model shows that the central thesis of Mauri et al. (2021) is not valid and that predictive models of the location of O-GlcNAcylation sites can be constructed with reasonable precision while also surpassing previous models in other metrics.

After the MLP was fully trained, we noticed that the "Mauri et al. (2021) – Original" dataset has some issues, such as entries that did not match the raw data and repeated entries, which may lead to test-set 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 Mauri et al. (2021), which is called "Mauri et al. (2021) – Modified" in this work. This modified dataset uses sequence data instead of human-selected descriptors.

On that corrected dataset, our RNNs behave similarly to YinOYang at low thresholds (Fig. S2). Beginning from a threshold of 10^{-10} for the RNN-76 model and 10^{-7} for the RNN-152 model[3], 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, a 391% improvement in precision, and a 136% improvement in MCC relative to YinOYang[4] (Table S1). Our RNN-76 model also surpasses the MCC reported in Pokharel et al. (2023) and Hu et al. (2024) on an independent test set, although the use of different training/test sets makes this only an indirect surpassing.

The performance of these models is lower than for the models tested with the original dataset of Mauri et al. (2021). This performance loss occurs due to the elimination of test-set data leakage, which biased the metrics upwards.[5] This reduction highlights the importance of constructing test datasets in a manner that avoids the potential for information leakage to be able to produce accurate assessments of model performance (Geslin et al., 2023; Jones, 2019).

3.2. A larger, less imbalanced dataset leads to improved models that surpass previously published models even further

After model training with the Mauri et al. (2021) datasets was complete, the dataset from Wulff-Fuentes et al. (2021) 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 Wulff-Fuentes et al. (2021) dataset. Wulff-Fuentes et al. (2021) also included a potential motif for O-GlcNAcylation, which was tested on this modified dataset. Models with different window sizes were also tested to determine the effect of window sizes on predictive power. In the previous sections and in Mauri et al. (2021), window sizes were restricted to 10 AAs on each side of the S/T (21 AAs total), likely due to YinOYang's fixed window size. However, it is reasonable to believe that AAs further away from the glycosylation site can have an effect on glycosylation;

[3] RNNs with 38 and 228 neurons had inferior cross-validation results, and so are not included.

[4] O-GlcNAcPRED-II and OGTSite are no longer available (their web servers have been down for years and no downloadable version can be obtained); thus, their metrics on the Modified dataset cannot be obtained. Note that these publications have not been retracted. Hu et al. (2024) faced a similar problem in their work.

[5] The performance loss was not due to feature changes. MLPs using this dataset and the same features lost more than 10% of their absolute recall values at the same precision and performed worse than the RNNs (Fig. S2).

Table 2
Recall, precision, F_1 score, and MCC metrics (in %) for the points with highest F_1 score for different O-GlcNAcylation site prediction models. Model "YinOYang" is from Gupta and Brunak (2002), Model "O-GlcNAcPRED-II" is from Jia et al. (2018), Model "OGTSite" is from Kao et al. (2015), and Model "Mauri" is the best model from a collection of models from Mauri et al. (2021). The metrics for these previously published models are from Tables 2 and 4 of Mauri et al. (2021). 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 Gupta and Brunak, 2002)	O-GlcNAcPRED-II (From Jia et al., 2018)	OGTSite (From Kao et al., 2015)	Mauri (From Mauri et al., 2021)	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	27.59	2.72
MCC (%)	8.10	11.73	14.44	7.03	27.56	0.00

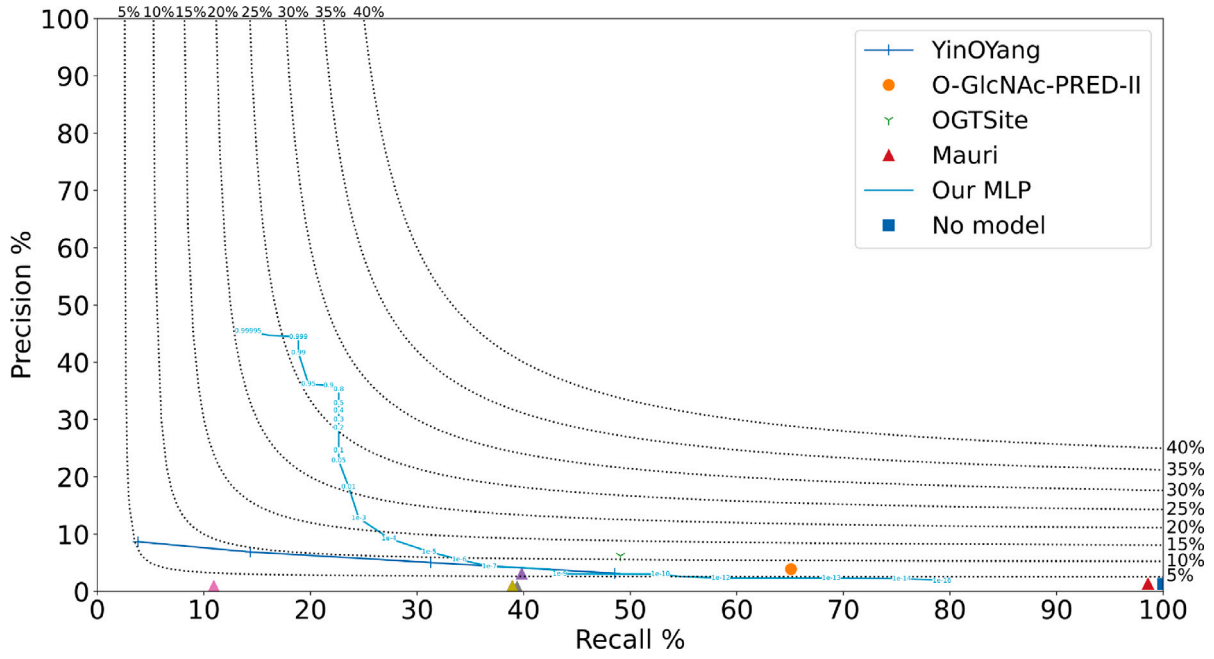

Fig. 1. P-R curves for O-GlcNAcylation site prediction models tested on the original dataset of Mauri et al. (2021). The black dotted lines are isolines of the F_1 score, as labeled on the top and right sides of the plot. Model "YinOYang" is from Gupta and Brunak (2002), Model "O-GlcNAcPRED-II" is from Jia et al. (2018), and Model "OGTSite" is from Kao et al. (2015) (complete P-R curves are not shown because the models are no longer publicly available). Models "Mauri" are a set of models from Mauri et al. (2021). Metrics for these previously published models are from Tables 2 and 4 of Mauri et al. (2021). 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.

thus, models with up to 25 AAs on each side (51 AAs total) were investigated.

Our RNNs exhibit a much higher recall at the same precision level (Fig. 2) than any models in the previous sections (Figs. 1 and S2). 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 models' performance 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[6] (Fig. 2 and Table 3).

Similarly to what occurred with the other datasets (Section 3.1 and Fig. S2), YinOYang performed very similarly to our models at low thresholds, and our models surpassed YoY at thresholds $\geq 10^{-6}$. Furthermore, YoY performed very similarly in this expanded dataset and in the smaller dataset from Section 3.1, indicating it is robust with respect to input data. However, its performance was also considerably low, surpassing an F_1 score of 10% only at one threshold level. Even our RNN with a window size of only 5 AAs – that is, with half the information per sample – is Pareto dominant over YoY. With increasing window sizes, this disparity grows further, highlighting the quality of our RNNs and the positive impact from increasing window sizes. At its F_1 score maximum, the RNN-225 model with a window size = 20 has a 357% improvement in precision, a 257% improvement in the F_1 score, and a 357% improvement in MCC relative to YinOYang (Table 3). Once again, the performance of our model on an independent test set surpasses that attained by the models in Pokharel et al. (2023) and Hu et al. (2024) in an indirect comparison.

To ensure our choice of cross-validation (later, training) and testing data is not biased, and to ensure the chosen hyperparameters are not excessively dependent on the particulars of a cross-validation dataset, five-fold nested validation was performed on the RNN-225; 20 win model (as described in Section 2.1). The difference in performance among the five nested validation folds was negligible (Fig. S3 and Table S2), indicating that the chosen architecture and hyperparameters are robust to variations in the training and testing data. This highlights how this methodology is sound and can be applicable to any O-GlcNAcylation dataset. The best thresholds were also very similar for most folds; for the one fold where that was not the case, it should be noted that thresholds of 0.6 and 0.7 had the third- and second-best

[6] Models with window sizes = 25 or RNN sizes = 300 neurons performed nearly identically to models with window sizes = 20 or RNN sizes = 150 or 225 neurons, so the former are not included in Fig. 2 or Table 3.

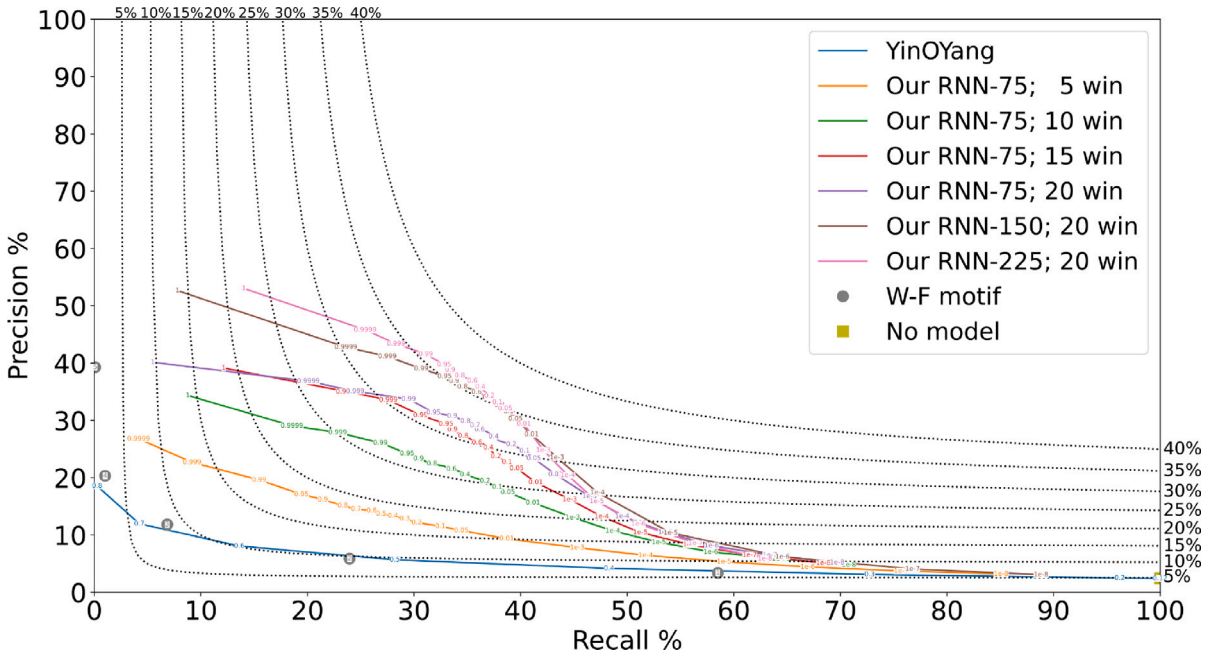

Fig. 2. P-R curves for O-GlcNAcylation site prediction models tested using a modified version of Wulff-Fuentes et al. (2021)'s dataset. The black dotted lines are F_1 score isolines, as labeled on the top and right sides of the figure. Model "YinOYang" is from Gupta and Brunak (2002); its metrics are obtained by us. 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 Wulff-Fuentes et al. (2021); 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.

performance respectively, with absolute F_1 score and MCC differences of less than 0.1%.

The previously proposed motif (from Wulff-Fuentes et al., 2021) 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. Because this motif captured very few sequences overall, we hypothesized that allowing sequences to slightly deviate from the motif could lead to improved results. 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. While these deviations improved the motif's results, the motif fails to achieve an F_1 score or MCC greater than 10%, indicating it is not suitable to predict or describe O-GlcNAcylation (Table 3).

A second motif, very similar to the one tested above, was published in Ma et al. (2022). The authors claim that the sequence they found "tend to be the degenerate consensus motifs for O-GlcNAcylated Ser and Thr residues of human proteins" and that "such motifs seem to be well conserved among species including mice and *Arabidopsis*" (Ma et al., 2022). Once again, this 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 45 follow the motif, and only 14 out of those 45 are truly positive (Table S4). As with the Wulff-Fuentes et al. (2021) motif, minor improvements can be achieved if the motif is relaxed, but this motif tends to perform slightly worse than that of Wulff-Fuentes et al. (2021).

A recent work (Hu et al., 2024) also used deep learning models to predict O-GlcNAcylation. Hu et al. (2024) used two independent test sets in their work, which they labeled "Ind_H_2022" and "Ind_M_2022" and were used to evaluate their human-only and mouse-only prediction models respectively. We test the performance of our model using the same datasets and notice that our model surpasses theirs, respectively, by 151% and 214% in terms of F_1 score and 98.5% and 96.7% in terms of MCC (Table 4). This superiority occurs despite our model's facing two disadvantages: first, Hu et al. (2024) trained one model for each species (human and mouse), whereas our model is trained to work universally on all mammalian sequences. Second, the sequences in these datasets contain only 29 amino acids (corresponding to a window size of 14) per sequence, but our models work best with 41 amino acids (corresponding to a window size of 20) per sequence.

3.3. Interpretation of model predictions using shapley values

The predictions of our models are interpreted via Shapley values by assigning a linear coefficient to each amino acid at each position based on a model's predictions, leading to a $(2 \times \text{window\_size} + 1) \times 20$ matrix of coefficients. Because each position has only one single amino acid, the final threshold for a given sequence is the sum of the $(2 \times \text{window\_size} + 1)$ Shapley values of its amino acids. These allow for elucidation of the effect of each amino acid on the glycosylation chance (Fig. 3). The use of Shapley values with a threshold of 0.10–0.15 leads to minimal losses in performance, indicating the values are descriptive of the models' predictions (Fig. S4 and Table S3).

Some of the most relevant amino acid and position combinations are in agreement with Wulff-Fuentes et al. (2021)'s proposed motif and methodology (e.g.: a *T* at −1 or an A at 2). However, many others are in disagreement. For example, our RNN-225; 20 win considers an H at 9 or an S at −2 as considerably important, but these combinations are marginally less frequent in the positive samples relative to the negative samples. Some other amino acid and position combinations (such as a P at −1 or 4) are slightly more frequent in the positive samples relative to the negative samples, yet our model regards these combinations as critically negative for O-GlcNAcylation. These differences, combined with the superior performance of our models, indicate that the approach of simply counting the most frequent amino acids at each

Table 3
Recall, precision, F_1 score, and MCC metrics (in %) for the point with highest F_1 score for models tested using a modified version of Wulff-Fuentes et al. (2021)'s dataset, and for the motif generated by Wulff-Fuentes et al. (2021). 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. Model "YinOYang" is from Gupta and Brunak (2002); its metrics are obtained by us. The "W-F Motifs" come from Wulff-Fuentes et al. (2021). 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	26.54	31.09	32.16
MCC (%)	16.23	24.59	29.26	30.38

Metric	RNN-150; 20 win (This Work)	RNN-225; 20 win (This Work)	YinOYang (from Gupta and Brunak, 2002)	No Model
Best Threshold	0.50	0.60	0.60	N/A
Recall (%)	36.36	35.47	13.59	100
Precision (%)	34.58	36.90	8.08	2.44
F_1 Score (%)	35.45	36.17	10.13	4.76
MCC (%)	33.76	34.57	7.57	0.00

Metric	W-F Motif (5) (From Wulff-Fuentes et al., 2021)	W-F Motif (4) (From Wulff-Fuentes et al., 2021)	W-F Motif (3) (From Wulff-Fuentes et al., 2021)	W-F Motif (2) (From Wulff-Fuentes et al., 2021)	W-F Motif (1) (From Wulff-Fuentes et al., 2021)
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
MCC (%)	1.70	4.01	7.22	7.23	4.71

Table 4
Recall, precision, F_1 score, and MCC metrics (in %) for the point with highest F_1 score for models tested using the "Ind_H_2022" and "Ind_M_2022" datasets from Hu et al. (2024). Models "RNN-225; 20 win" come from this work and are labeled as per Table 3. Models "Hu et al." are from Hu et al. (2024); their recall and MCC metrics are as reported in Hu et al. (2024), whereas their precision and F_1 are backcalculated by us, as these are not available in that work.

Metric	RNN-225; 20 win; Human (This Work)	Hu et al.; Human (From Hu et al., 2024)	RNN-225; 20 win; Mouse (This Work)	Hu et al.; Mouse (From Hu et al., 2024)
Best Threshold	1	N/A	1	N/A
Recall (%)	12.95	59.20	39.00	80.69
Precision (%)	7.03	1.88	11.88	3.01
F_1 Score (%)	9.12	3.64	18.21	5.80
MCC (%)	7.74	3.90	19.30	9.81

position (a unigram model) is not adequate to describe and predict O-GlcNAcylation, and it is likely that combinatorial effects (such as the secondary structure of, or the net charge near the potential site) play an important role in O-GlcNAcylation.

4. Discussion

This work constructs MLP and RNN models from multiple sources of literature data on protein O-GlcNAcylation based on human-selected descriptors (first part of Section 3.1) or protein sequences (second part of Section 3.1 and all of Section 3.2). According to Mauri et al. (2021), a significant limitation of O-GlcNAcylation models was that they achieved very low precision (< 9%) on independent test sets. Pokharel et al. (2023) and Hu et al. (2024) trained O-GlcNAcylation models using deep learning methods after Mauri et al. (2021) was published, yet neither could reach a precision ≥ 9% on independent test sets. This study contrasts with the past studies by our use of different model architectures, proper handling of dataset imbalance, multiple prediction thresholds, and rigorous cross-validation for hyperparameter selection. An MLP model trained in this work was compared with previously published models (as reported in Mauri et al. (2021)), and this work's MLP model surpasses the previously published models in precision (307% improvement), F_1 score (151% improvement), and MCC (90.9% improvement) metrics (Section 3.1). An analysis of the dataset of Mauri et al. (2021) 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 Mauri et al. (2021). 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 (Gupta and Brunak, 2002), the only model evaluated in Mauri et al. (2021) that is still available (second part of Section 3.1). 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, 134% improvement in F_1 score, and 136% improvement in MCC when compared to YinOYang, and the RNN models surpassed an MLP model trained on the same data (Fig. S2). This comparison suggests that the RNN architecture is better suited for prediction of O-GlcNAcylation than the MLP architecture, which is consistent with the RNN being able to leverage the sequential data structure of protein site sequences over the MLP's limited way of handling structured data.

A much larger dataset from Wulff-Fuentes et al. (2021) 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 previous works, displaying a 357% improvement in precision, 257% improvement in F_1 score, and 357% improvement in MCC over YinOYang (Section 3.2). These RNN models

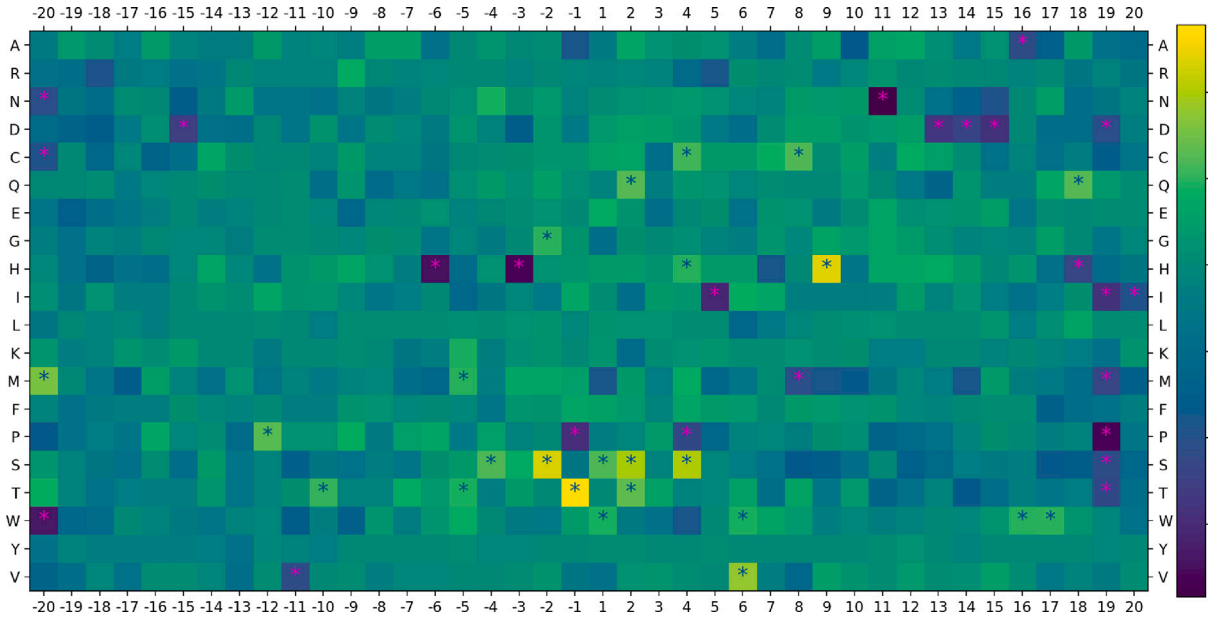

Fig. 3. Heatmap of Shapley values for each amino acid and position from the RNN-225; 20 win model (Section 3.2). Yellow and light green squares indicate AA/position combinations that are more likely to be found in O-GlcNAcylated sequences; blue squares indicate AA/position combinations that are more likely to be found in non-O-GlcNAcylated sequences. Blue and magenta stars represents AA/position combinations at the top 3% or bottom 3% of values respectively. Units for this heatmap are arbitrary and thus not shown.

were also better than the RNN models trained in the previous section, and we hypothesize part of this difference is due to the greater number of entries and lower data imbalance found in Wulff-Fuentes et al. (2021)'s modified dataset. The use of weight decay and learning rate scheduling contributed to further improving the performance of the RNN models in Section 3.2. We then compared the results of our model with these from the models of Hu et al. (2024) on an independent test set created by Hu et al. (2024). Despite the fact that Hu et al. (2024) created one model per species (as opposed to our generalist mammalian model) and that the evaluated datasets had a window size of only 14 (while our model was trained and achieves best results with a window size of 20), our model surpasses the models of Hu et al. (2024) by 151% and 214% in terms of F_1 score and 98.5% and 96.7% in terms of MCC (Table 4). These results highlight the superiority and generalizability of our methodology and model, and also that using a more complex deep learning architecture or features derived from protein structural properties is not a silver bullet for the prediction of O-GlcNAcylation sites.

Wulff-Fuentes et al. (2021) and Ma et al. (2022) also propose similar motifs for O-GlcNAcylation, but their motifs are excessively restrictive. The original formulation of Wulff-Fuentes et al. (2021)'s motif has an F_1 score of only 0.16% and an MCC of 1.70%, and the original formulation of Ma et al. (2022)'s motif has an F_1 score of only 0.21% and an MCC of 1.67%, indicating they are barely better than random guessing. While making them less strict increased their performance, their motifs never achieve an F_1 score or MCC $\geq$ 10% (Tables 3 and S4). The Shapley values extracted from our models, on the other hand, provide interpretability while maintaining most of the superior performance of our models (Section 3.3, Fig. S4, and Table S3). While a few of the most positive or negative Shapley values match the motif proposed in Wulff-Fuentes et al. (2021), many others do not. Given the higher performance of the Shapley value predictions over Wulff-Fuentes et al. (2021)'s motif, this suggests a simple unigram model is not adequate to describe and predict O-GlcNAcylation, and more complex models that take into account interactions are necessary.

The software used in this work is publicly available, 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 Wulff-Fuentes et al. (2021)'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. Instructions are provided in Sections S1 and S2 of the supplemental data or in the README in our GitHub repository.

Funding

This work was supported by a Project Award Agreement from the National Institute for Innovation in Manufacturing Biopharmaceuticals (NIIMBL) from the U.S. Department of Commerce, National Institute of Standards and Technology [70NANB17H002, 70NANB21H086]. P.S. was partially supported by a MathWorks Engineering Fellowship.

CRediT authorship contribution statement

Pedro Seber: Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Richard D. Braatz:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

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

Data and models are publicly available at https://github.com/PedroSeber/O-GlcNAcylation_Prediction.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compchemeng.2024.108818.

References

Chang, Y.-H., Weng, C.-L., Lin, K.-I., 2020. O-GlcNAcylation and its role in the immune system. J. Biomed. Sci. 27 (1), 57.

Chicco, D., 2017. Ten quick tips for machine learning in computational biology. BioData Min. 10 (35).

Chicco, D., Jurman, G., 2023. The matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Min. 16, 4.

Geslin, A., et al., 2023. Battery lifetime predictions: information leakage from unblinded training. ChemRxiv.

Gupta, R., Brunak, S., 2002. Prediction of glycosylation across the human proteome and the correlation to protein function. In: Pacific Symposium on Biocomputing. pp. 310–322.

Halligan, S., Altman, D.G., Mallett, S., 2015. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach. Eur. Radiol. 25, 932–939.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362.

Hu, F., Li, W., Li, Y., Hou, C., Ma, J., Jia, C., 2024. O-GlcNAcPRED-DL: Prediction of protein O-GlcNAcylation sites based on an ensemble model of deep learning. J. Proteome Res. 23, 95–106.

Jia, C., Zuo, Y., Zou, Q., 2018. 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.

Jones, D.T., 2019. Setting the standards for machine learning in biology. Nature Rev. Mol. Cell Biol. 20, 659–660.

Kao, H.-J., Huang, C.-H., Bretaña, N.A., Lu, C.-T., Huang, K.-Y., Weng, S.-L., Lee, T.-Y., 2015. A two-layered machine learning method to identify protein O-GlcNAcylation sites with O-GlcNAc transferase substrate motifs, BMC. 16 (S18), S10.

Kukleva, A., Böhle, M., Schiele, B., Kuehne, H., Rupprecht, C., 2023. Temperature schedules for self-supervised contrastive methods on long-tail data. arXiv:2303.13664.

Liang, C., Chiang, A.W., Hansen, A.H., Arnsdorf, J., Schoffelen, S., Sorrentino, J.T., Kellman, B.P., Bao, B., Voldborg, B.G., Lewis, N.E., 2020. A Markov model of glycosylation elucidates isozyme specificity and glycosyltransferase interactions for glycoengineering. Curr. Res. Biotechnol. 2, 22–36.

Lobo, J.M., Jiménez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Global Ecol. Biogeogr. 17 (2), 145–151.

Loshchilov, I., Hutter, F., 2017. SGDR: Stochastic gradient descent with warm restarts. arXiv:1608.03983.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Vol. 30. Curran Associates, Inc., pp. 4765–4774.

Ma, J., Hou, C., Wu, C., 2022. Demystifying the O-GlcNAc code: A systems view. Chem. Rev. 122, 15822–15864.

Mauri, T., Menu-Bouaouiche, L., Bardor, M., Lefebvre, T., Lensink, M.F., Brysbaert, G., 2021. O-GlcNAcylation prediction: An unattained objective. Adv. Appl. Bioinform. Chem. 14, 87–102.

McKinney, W., 2010. Data structures for statistical computing in Python. In: van der Walt, S., Millman, J. (Eds.), Proceedings of the 9th Python in Science Conference. pp. 56–61.

Mishra, S., Yamasaki, T., Imaizumi, H., 2019. Improving image classifiers for small datasets by learning rate adaptations. In: 2019 16th International Conference on Machine Vision Applications. MVA, pp. 1–6.

Moon, S., Chatterjee, S., Seeberger, P.H., Gilmore, K., 2021. Predicting glycosylation stereoselectivity using machine learning. Chem. Sci. 12 (8), 2931–2939.

Ozenne, B., Subtil, F., Maucort-Boulch, D., 2015. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J. Clin. Epidemiol. 68 (8), 855–859.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., pp. 8024–8035.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pokharel, S., Pratyush, P., Ismail, H.D., Ma, J., KC, D.B., 2023. Integrating embeddings from multiple protein language models to improve protein O-GlcNAc site prediction. Int. J. Mol. Sci. 24, 16000.

Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS One 10, e0118432.

Schjoldager, K.T., Narimatsu, Y., Joshi, H.J., Clausen, H., 2020. Global view of human protein glycosylation pathways and functions. Nature Rev. Mol. Cell Biol. 21 (12), 729–749.

Seber, P., Braatz, R.D., 2023. Linear and neural network models for predicting N-glycosylation in Chinese hamster ovary cells based on B4GALT levels. bioRxiv.

Shapley, L.S., 1951. Notes on the n-person game – II: The value of an n-person game. US Air Force Proj RAND 8.

Shi, Q., Shen, Q., Liu, Y., Shi, Y., Huang, W., Wang, X., Li, Z., Chai, Y., Wang, H., Hu, X., Li, N., Zhang, Q., Cao, X., 2022. Increased glucose metabolism in TAMs fuels O-GlcNAcylation of lysosomal Cathepsin B to promote cancer metastasis and chemoresistance. Cancer Cell 40, 1207–1222.e10.

Umapathi, P., Mesubi, O.O., Banerjee, P.S., Abrol, N., Wang, Q., Luczak, E.D., Wu, Y., Granger, J.M., Wei, A.-C., Gaido, O.E.R., Florea, L., Talbot, C.C., Hart, G.W., Zachara, N.E., Anderson, M.E., 2021. Excessive O-GlcNAcylation causes heart failure and sudden death. Circulation 143, 1687–1703.

Wulff-Fuentes, E., Berendt, R.R., Massman, L., Danner, L., Malard, F., Vora, J., Kahsay, R., Stichelen, S.O.-V., 2021. The human O-GlcNAcome database and meta-analysis. Sci. Data 8 (1), 25.

Zhu, Y., Hart, G.W., 2021. Targeting O-glcNAcylation to develop novel therapeutics. Mol. Aspects Med. 79, 100885.