frontiers | Frontiers in Cell and Developmental Biology

TYPE Original Research
PUBLISHED 10 October 2024
DOI 10.3389/fcell.2024.1456728

OPEN ACCESS

EDITED BY
Xiulan Chen,
Chinese Academy of Sciences (CAS), China

REVIEWED BY
Zilu Ye,
Peking Union Medical College Hospital (CAMS), China
Qingbo Shu,
North Carolina State University, United States

*CORRESPONDENCE
Jia Mi,
jia.mi@bzmc.edu.cn

†These authors have contributed equally to this work

RECEIVED 29 June 2024
ACCEPTED 26 September 2024
PUBLISHED 10 October 2024

CITATION
Zhang L, Deng T, Pan S, Zhang M, Zhang Y, Yang C, Yang X, Tian G and Mi J (2024) DeepO-GlcNAc: a web server for prediction of protein O-GlcNAcylation sites using deep learning combined with attention mechanism.
Front. Cell Dev. Biol. 12:1456728.
doi: 10.3389/fcell.2024.1456728

DeepO-GlcNAc: a web server for prediction of protein O-GlcNAcylation sites using deep learning combined with attention mechanism

Liyuan Zhang[1†], Tingzhi Deng[1,2†], Shuijing Pan[1], Minghui Zhang[1], Yusen Zhang[3], Chunhua Yang[1], Xiaoyong Yang[4], Geng Tian[1] and Jia Mi[1*]

[1]Shandong Technology Innovation Center of Molecular Targeting and Intelligent Diagnosis and Treatment, Binzhou Medical University, Yantai, Shandong, China, [2]National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, Fujian, China, [3]School of Mathematics and Statistics, Shandong University, Weihai, Shandong, China, [4]Department of Comparative Medicine, Department of Cellular and Molecular Physiology, Yale University, New Haven, CT, United States

Introduction: Protein O-GlcNAcylation is a dynamic post-translational modification involved in major cellular processes and associated with many human diseases. Bioinformatic prediction of O-GlcNAc sites before experimental validation is a challenge task in O-GlcNAc research. Recent advancements in deep learning algorithms and the availability of O-GlcNAc proteomics data present an opportunity to improve O-GlcNAc site prediction.

Objectives: This study aims to develop a deep learning-based tool to improve O-GlcNAcylation site prediction.

Methods: We construct an annotated unbalanced O-GlcNAcylation data set and propose a new deep learning framework, DeepO-GlcNAc, using Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN) combined with attention mechanism.

Results: The ablation study confirms that the additional model components in DeepO-GlcNAc, such as attention mechanisms and LSTM, contribute positively to improving prediction performance. Our model demonstrates strong robustness across five cross-species datasets, excluding humans. We also compare our model with three external predictors using an independent dataset. Our results demonstrated that DeepO-GlcNAc outperforms the external predictors, achieving an accuracy of 92%, an average precision of 72%, a MCC of 0.60, and an AUC of 92% in ROC analysis. Moreover, we have implemented DeepO-GlcNAc as a web server to facilitate further investigation and usage by the scientific community.

Conclusion: Our work demonstrates the feasibility of utilizing deep learning for O-GlcNAc site prediction and provides a novel tool for O-GlcNAc investigation.

KEYWORDS

deep learning, O-GlcNAc, CNN, prediciton, attention

Introduction

Protein post-translational modification (PTM) refers to the covalent modification of a protein after synthesized (Conibear, 2020). It plays a crucial role in diversifying protein functions and regulating cellular processes. Among currently known PTMs (Ramazi and Zahiri, 2021), O-linked β-N-acetylglucosaminylation (O-GlcNAcylation) is considered as a critical regulation mechanism (Yang and Qian, 2017). This modification involves the attachment of N-acetylglucosamine (GlcNAc) moieties to serine (S) or threonine (T) residues, a process catalyzed by O-GlcNAc transferase (OGT) and reversed by O-GlcNAcase (OGA) (Yang and Qian, 2017). Among the currently known post-translational modifications (PTMs), O-linked β-N-acetylglucosaminylation (O-GlcNAcylation) is regarded as a critical regulatory mechanism. This modification involves the attachment of N-acetylglucosamine (GlcNAc) moieties to serine (S) or threonine (T) residues, a process catalyzed by O-GlcNAc transferase (OGT) and reversed by O-GlcNAcase (OGA). O-GlcNAcylation plays a vital role as a cellular nutrient and stress sensor, regulating key processes such as signal transduction and cell cycle control (Yang et al., 2020). Its dysregulation has been linked to diseases like cancer, neurodegenerative disorders (Smet-Nocca et al., 2011), and metabolic conditions (Hart et al., 2007; Slawson and Hart, 2011). Identifying O-GlcNAc sites may uncover detailed mechanisms of disease pathology and offer novel therapeutic options. In neurodegenerative diseases, the hyperphosphorylation status of Tau proteins contributes to the neuronal death, and proposed as promising therapeutical targets. The O-GlcNAcylation at residue S400 of the Tau protein may reduce the phosphorylation at S404 which disrupts the GSK3β-mediated sequential phosphorylation process in neuron (Smet-Nocca et al., 2011). Therefore, the elevation of O-GlcNAcylation with O-GlcNAcase inhibitors are proposed as a novel therapy for (Alzheimer's disease) AD (Arnold et al., 1996; Liu et al., 2004; Morris et al., 2015; Bartolome-Nebreda et al., 2021). In metabolic disorders, the dysregulation of gluconeogenesis is one of the processes that is regulated by Peroxisome proliferator-activated receptor gamma coactivator 1-alpha (PGC-1α). The O-GlcNAcylation at Ser333 of PGC-1α is proved to protect PGC-1α from ubiquitination and further proteasomal degradation, which shedding light on new strategies for diabetes treatment (Ruan et al., 2012). Hence, identifying the specific O-glycosylation sites on proteins of interest is crucial for disease and novel drug investigation.

Bioinformatics-based approach has been proved to be advantageous for PTM site identification, with low cost and high throughout capabilities (Meng et al., 2022; Chen et al., 2019). Predicting potential PTM sites prior to experimental validation has become an essential tool for molecular biologists (Wen et al., 2020; Khan et al., 2021). Early predictors like YinOYang (2002) (Gupta and Brunak, 2002) and O-GlcNAcScan (Wang et al., 2011) used machine learning techniques such as artificial neural networks and support vector machines to improve O-GlcNAcylation site identification. Over time, more advanced models like GlycoMine (Li et al., 2015), further improved prediction performance by coupling the Random Forest (RF) algorithm with effective features selected through information gain (IG) and minimum redundancy maximum relevance (mRMR) (Li et al., 2015). Consideration of protein structural features was also proposed for O-GlcNAc prediction, GlycoMinestruct was constructed for O-GlcNAc prediction based on 29 O-linked glycosylated PDB structures, which corresponded to 47 O-linked glycosylation sites (Li et al., 2016). These predictors have demonstrated the effectiveness of bioinformatics approaches in O-GlcNAc prediction, and some of them have been well adopted by researchers. However, several critical issues persist in O-GlcNAc prediction, such as overall unsatisfactory performance and limited availability of online prediction servers. Therefore, more sophisticated models are needed for improving prediction performance. One potential approach to improve prediction accuracy involves leveraging deep learning-based methods, which have demonstrated success in other PTM predictions (Li et al., 2022a; Wang et al., 2022). Deep learning has presented its remarkable performance in comparison to traditional machine learning methods due to its robustness and generalization. Recently, Hu et al. reported an O-GlcNAc predictor based on connection of a convolutional neural network and bidirectional long short-term memory, indicting the potential of deep learning in O-GlcNAc prediction (Hu et al., 2023). However, the performance is still insufficient, and more algorithms are in need to improve current achievement, such as attention mechanism (Mauri et al., 2021).

In this study, we construct an annotated unbalanced O-GlcNAcylation data set and propose a new deep learning framework, DeepO-GlcNAc, using Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN) combined with attention mechanism. We have developed a web server for the prediction of protein O-GlcNAcylation sites, making it freely accessible to the public. To evaluate the generalization performance of DeepO-GlcNAc, we incorporate cross-species data from five organisms—mouse, *Drosophila*, *Caenorhabditis-elegans*, *Arabidopsis*, and rat, our model demonstrates strong robustness. Additionally, we conduct ablation experiments to assess our model's performance. Our model outperforms other architectures such as CNN, CNN-LSTM, CNN-Attention, and CNN-Attention-LSTM, establishing itself as a powerful deep learning-based O-GlcNAc predictor.

A workflow on ensemble DeepO-GlcNAc is showed in Figure 1. A fixed dataset is retained for testing, with the remaining data as training set. PTM data at the protein level are mapped to core sequences using the slide window of 21 size strategy; One-hot encoding was used to digitize discrete features. Each individual model is trained on the processed data under the label setting; each model is subsequently evaluated in the evaluation steps. An online service of DeepO-GlcNAc was constructed. The framework of CNN-Attention-LSTM model is showed in Figure 2.

Materials and methods

Data collection and preparation

We downloaded 4,577 reviewed O-glycosylated protein sequences in dbPTM database (Li et al., 2022b). The obtained 16,691 O-GlcNAcylation sites were experimentally validated. Considering the sequence similarity used in experiments in O-GlcNAcylation site-specific modification assays, we used the

CD-HIT tool to remove protein sequences with greater than 30% homology (Li and Godzik, 2006). As the O-glycosylated sites occur in serine or threonine (S/T), we took S or T as the center and intercepted peptide fragments of length 21. Finally, we obtained protein sequences containing a total of 23,252 S/T sites. They can be represented in the following scheme:

$$P(0) = N_{-10}N_{-9}\ldots N_{-2}N_{-1}O\,N_{+1}N_{+2}\ldots N_{+9}N_{+10}$$

Where the center O denotes serine (S) or threonine (T). If there are fewer than 21 amino acids, we extended these sequences as virtual amino acids with non-existent residual "*" to ensure that the window length of each sequence was fixed at 21. The peptides fragments can be further divided into two classes:

$$P(O) \in \left\{ \begin{array}{c} P^{+}(O), \text{if the center is an O-glycoslated site,} \\ P^{-}(O), \text{otherwise} \end{array} \right\}$$

where $P^{+}(O)$ is an experimentally verified O-glycosylated site, i.e., 2,696 positive samples; $P^{-}(O)$ a non-O-glycosylated site, i.e., 20,556 negative samples.

A total of 23,252 potential sites (Serine and Threonine) are included in the dataset. Among them, 2,696 sites are validated O-GlcNAc sites, and 20,556 sites are considered as negative samples. The dataset was randomly divided into two parts, one part includes 80% of the data as training set and the other with 20% data was used as an independent testing set. A peptide similarity check was performed between two parts with CD-HIT to ensure the testing dataset is independent of training dataset (threshold = 40%) (Li and Godzik, 2006).

To demonstrate the generalization performance of our model, DeepO-GlcNAc was tested on five cross-species benchmark data sets including mouse, *Drosophila*, *Caenorhabditis-elegans*, *Arabidopsis and rats*, the O-GlcNAcylation information in these species were obtained from this website: https://oglcnac.org/atlas/download/. The statistical information on the data is listed in Table 2.

One-hot encoding

The dataset was encoded with One-hot encoding approach, which is a common and popular feature extraction technique that can generate a numerical feature vector from a protein sequence (Meng et al., 2022). According to this method, one amino acid is denoted as a feature vector of 21-dimension such as amino acid alanine (A) is presented as "100000000000000000000" and the dummy amino acid "*" is presented as "000000000000000000001". Therefore, an L*21-dimensional feature vector can be obtained for a protein fragment of length L. In this study, we used window size 21 to generate peptide samples and got a 441 (21 × 21) dimensional feature vector to encode a peptide fragment.

$$B_i = (b_{n1}, b_{n2}, b_{n3}, b_{n4}, \ldots, b_{n21})$$

$$b \in \left\{ \begin{array}{l} A\text{: } 100000000000000000000 \\ C\text{: } 010000000000000000000 \\ \cdots \\ Y\text{: } 000000000000000000010 \\ *\text{: } 000000000000000000001 \end{array} \right\}$$

$$n \in \{A, C, \ldots, Y, *\}$$

We used a weighted cross-entropy loss function (BCE_Loss), which assigns greater importance to positive samples ensuring that the model does not become biased towards predicting the negative class. The "pos_weight" parameter was opted for 4.

$$\text{BCE\_Loss} = -\frac{1}{N}\sum_{i=1}^{N}[w \cdot y_i \cdot \log(\sigma(x_i)) + (1 - y_i) \cdot \log(1 - \sigma(x_i))$$

Convolutional neural networks

A convolutional neural network (CNN) architecture was employed for feature extraction and presentation. Convolutional Neural Networks are originally proposed by Fukushima (1980) as noncognition model, which is one of the earliest algorithms in the field of deep learning. This network mainly consists of four layers of operations (Kim, 2014): convolutional layer, pooling layer, fully connected layer, and output layer. The convolution operation is represented mathematically as shown below:

$$(I \otimes K)_{ij} = \sum_{m=0}^{k_1-1}\sum_{n=0}^{k_s-1} I(i+m, j+m) \cdot K(m, n)$$

Where I is the feature matrix, K is the convolution kernel, i and j represent the $i-th$ and $j-th$ rows and columns of the feature matrix, and m and n represents the $m-th$ and $n-th$ rows and columns of the convolution kernel. Maximum pooling retains the maximum value of each feature, The pooling layer is used to reduce the dimensionality of data, select and filter the features learned, to reduce the complexity of the model and avoid overfitting (Kim, 2014).

Specifically, the CNN model consisted of two convolutional layers (Conv1 and Conv2) followed by rectified linear unit (ReLU) activation functions and max-pooling layers. The first convolutional layer (Conv1) had 10 output channels and a kernel size of 5 × 5, while the second convolutional layer (Conv2) had 20 output channels and a kernel size of 3 × 3. The stride for both convolutional layers was set to 1. The ReLU activation function was applied after each convolutional to introduce non-linearity into the model.

Long short-term memory

Long Short-Term Memory (LSTM) is a type of recursive neural network extension model proposed by Hochreiter and Schmidhuber (1997). The main advantage of LSTM lies in its internal mechanism of gates that control information flow. With the addition of special "gate" structures, LSTM can handle the problem of long-term memory. The LSTM layer exhibits a hidden state dimensionality of 512. Additionally, two fully connected (dense) layers (FC1 and FC2) followed by the LSTM layer. The first fully connected layer manifests an output dimensionality of 288, serving as an intermediary transformation stage between the

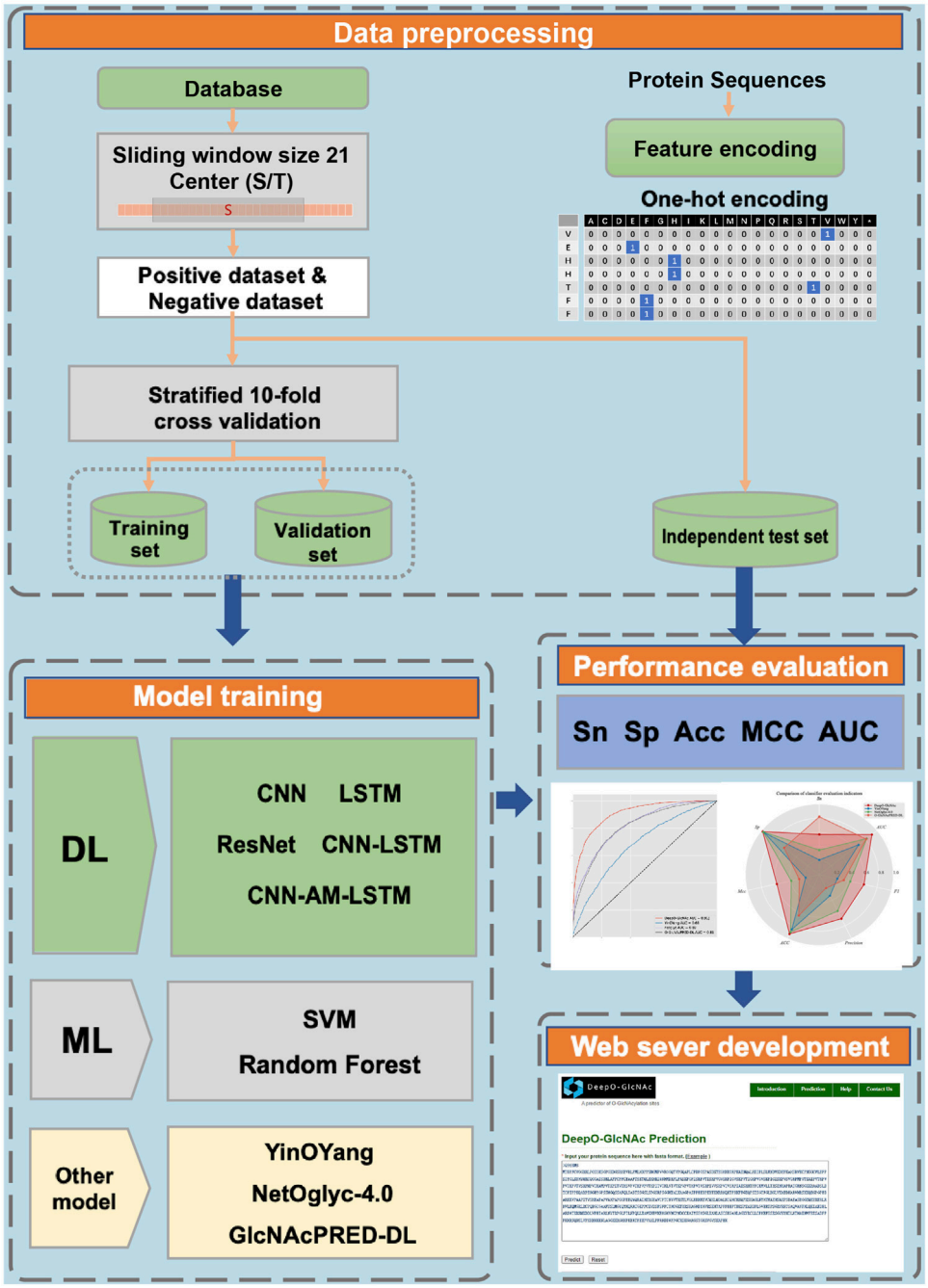

FIGURE 1
Methodology workflow.

LSTM layer and subsequent layers. This dimensionality is selected based on considerations of feature representation and model complexity. The second fully connected layer exhibits an output dimensionality of 2, aligning with the binary classification nature of the task. The ReLU activation function was applied after the first fully connected layer to introduce non-linearity into the model. After the output layer, the log_SoftMax function was employed to compute the logarithm of the softmax probabilities, facilitating model predictions for the binary classification task.

Attention mechanism

Attention Mechanism is widely applied in various fields such as image and natural language processing, due to its ability to achieve fast parallel computations through matrix operations (Ning and Li, 2022). It calculates the attention distribution on input features and outputs the weighted features based on the attention distribution. Therefore combination of Attention Mechanism may benefit for independent CNN or LSTM network models. The SE block (Hu et al., 2017) is adopted as the core structure of attention in this paper,

in order to obtain the importance of each feature channel and the interdependence between feature channels. Weight values are assigned to each feature channel to allow the neural network to focus on these feature channels. For an input of feature channel number C, the weighted feature channels with number C are calculated and then weighted based on the following three operations.

The Squeeze operation uses global average pooling for each channel. It represents the global distribution of responses on feature channels and allows layers near the input to obtain a global receptive field.

$$z_c = F_{sq}(U_c) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j)$$

The Excitation operation, which is a mechanism similar to the gate in recurrent neural networks, generates weights for each feature channel through the parameter w. The Scale operation considers the weights output by Excitation to represent the importance of each feature channel after feature selection, and then scales the original feature channel through multiplication to complete the re-scaling of the original feature along the channel dimension.

$$\widetilde{X_c} = F_{scale}(U_c, s_c) = s_c \cdot U_c$$

Specifically, two SE blocks were employed after first and second convolutional layer, for the two SE modules, input channels are 10 and 20, corresponding to the output channels of the first and the second convolutional layer.

Model evaluation

For deep learning model training, ten-fold cross-validation is performed by dividing the dataset into 10 subsets and using 9 of them as training sets and 1 as test set in turn. Each subset is validated once in the ten-fold cross-validation process (Supplementary Figure S1, Supplementary Table S1). The accuracy of each validation is recorded and the model with the highest accuracy is considered as the optimal model. The independent test set is further used to evaluate the model and compare it with the other tools. Several evaluation metrics are employed in this work, including *Accuracy* (*ACC*), *Matthew's correlation coefficient* (*MCC*), *Sensitivity* (*Sn*), *Specificity* (*Sp*), *Precision*, and *F1*-score which are illustrated as

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 = \frac{2 \times recall \times precision}{recall + precision}$$

where FP FN TP and TN represent the number of false positives, false negatives, true positives and true negatives, respectively. In addition, we use the area under the ROC (AUC) to measure the classifier's ability by plotting the true positive rate (TPR) against the false positive rate (FPR).

Results

Motif conservation analysis of O-GlcNAc sites in human proteins

To illustrate the different distribution and preference of flanking residues surrounding O-GlcNAc sites on human proteins, we used the Probability Logo Generator (pLogo) algorithm. This allowed us to compare the amino acid sequences surrounding observed O-GlcNAc sites with sequences from non-O-GlcNAc sites, utilizing our dataset (Figure 3). Currently, there is no confirmed conservation motif for O-GlcNAc. Our analysis revealed a predominant presence of Threonine (T) and Proline (P) residues in the vicinity of O-GlcNAc sites, whereas Leucine (L) and Cysteine (C) residues were observed around non-O-GlcNAc sites.

Ablation studies on independent test

To evaluate the impact of different model components on performance of DeepO-GlcNAc. We conducted ablation experiments using an independent dataset. The Area Under the Curve (AUC) values of ROC curves indicating that the DeepO-GlcNAc model (AUC = 0.92) outperforms CNN (AUC = 0.79),

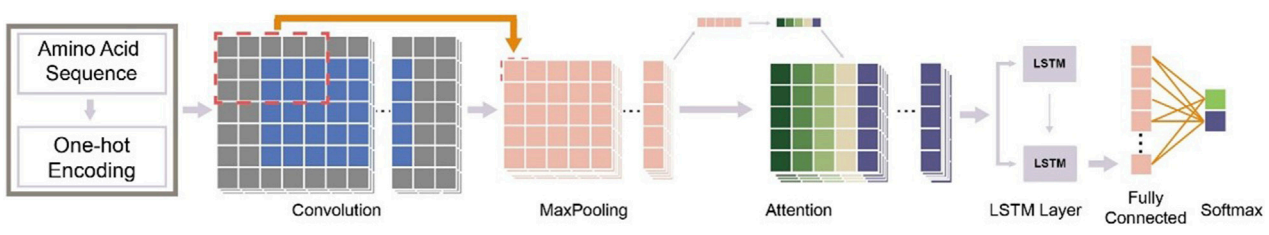

FIGURE 2
Framework of the fusion of CNN-Attention-LSTM (DeepO-GlcNAc). The amino acid sequences are encoded using one-hot encoding, and passed through a series of layers including Convolution, MaxPooling, Attention, LSTM, and Fully connected layers to construct the framework, then activated by Softmax.

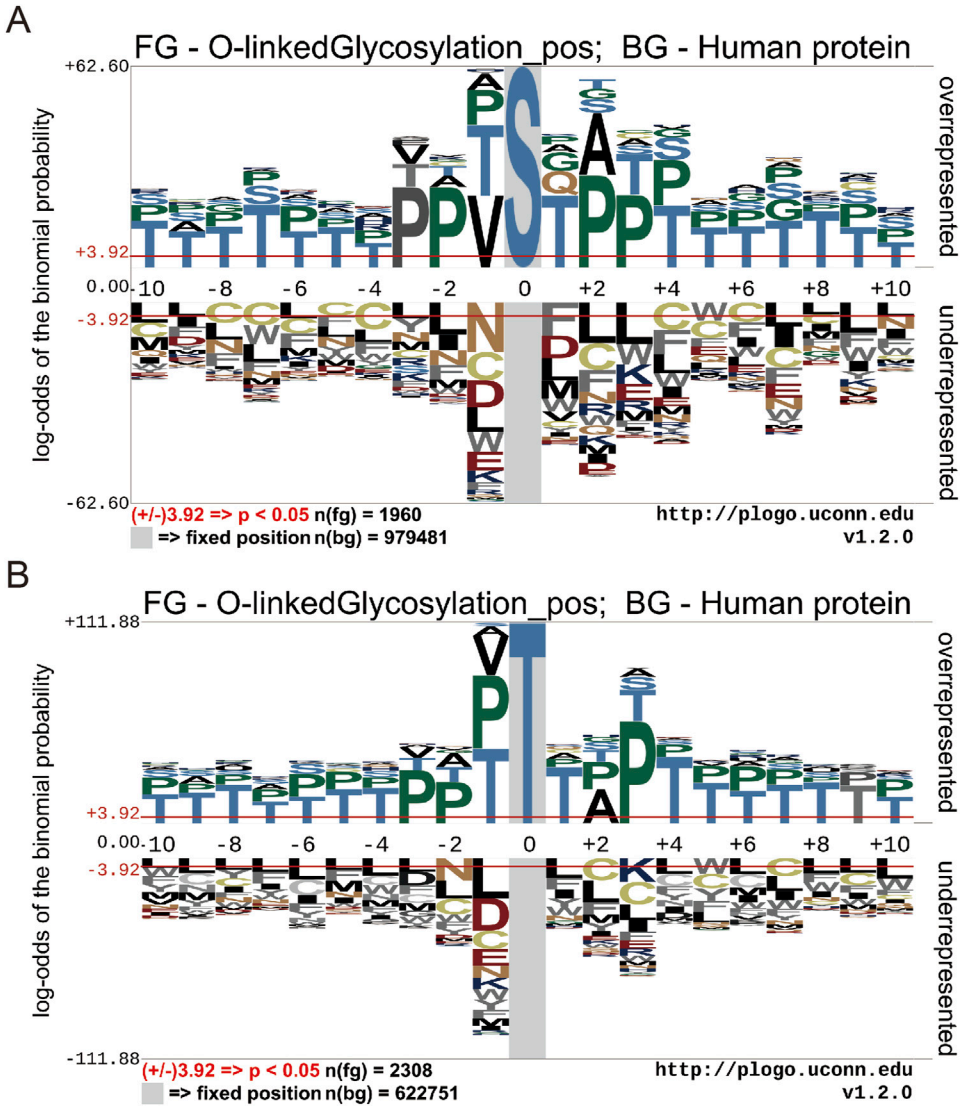

FIGURE 3
Motif conservation analysis of O-GlcNAc sites. **(A)** Motif conservation analysis of O-GlcNAc sites for Threonine (T) **(B)** Motif conservation analysis of O-GlcNAc sites for serine (S). The sequence logos were generated with pLogo with scaled better data visualization. The default values ± 3.92 ($P < 0.05$) were used as the thresholds for significantly overrepresented and underrepresented amino acids, respectively. The red horizontal lines on the sequence logos denote the $P < 0.05$ threshold.

CNN-SE (AUC = 0.87), and CNN-LSTM (AUC = 0.87) models in terms of true positive rate versus false positive rate (Figure 4A). In Precision-Recall curves, the Average Precision (AP) of DeepO-GlcNAc (AP = 0.72) exceeds that of the CNN (AP = 0.44), CNN-SE (AP = 0.62), and CNN-LSTM (AP = 0.62) models (Figure 4B). These results highlight the superior performance of DeepO-GlcNAc, especially in reducing false positives and maintaining higher precision across various recall levels. Compared to CNN, CNN_SE, and CNN_LSTM, DeepO-GlcNAc demonstrated the highest sensitivity (Sn = 0.68), Matthews Correlation Coefficient (MCC = 0.60), accuracy (Acc = 0.92), and F1 score (0.65). Meanwhile, CNN_SE achieved the best specificity (Sp = 0.96) and precision (0.62), as shown in Table 1. These results highlight the superior performance of DeepO-GlcNAc and the importance of the SE module in enhancing model performance.

For each model, the area under the ROC curve and the Precision-Recall curve are reported.

DeepO-GlcNAc demonstrated varied predictive performance across different species

To evaluate the generalization performance of DeepO-GlcNAc, we incorporate cross-species data from five organisms—mouse, *Drosophila*, *Caenorhabditis-elegans*, *Arabidopsis*, and rat. The statistical information on the data is listed in Table 2. As can be seen in Figure 5A, accuracy (ACC) values for all species hovered around 0.6, with the highest for *Arabidopsis* at 0.63, while *rat*, *Caenorhabditis-elegans* and

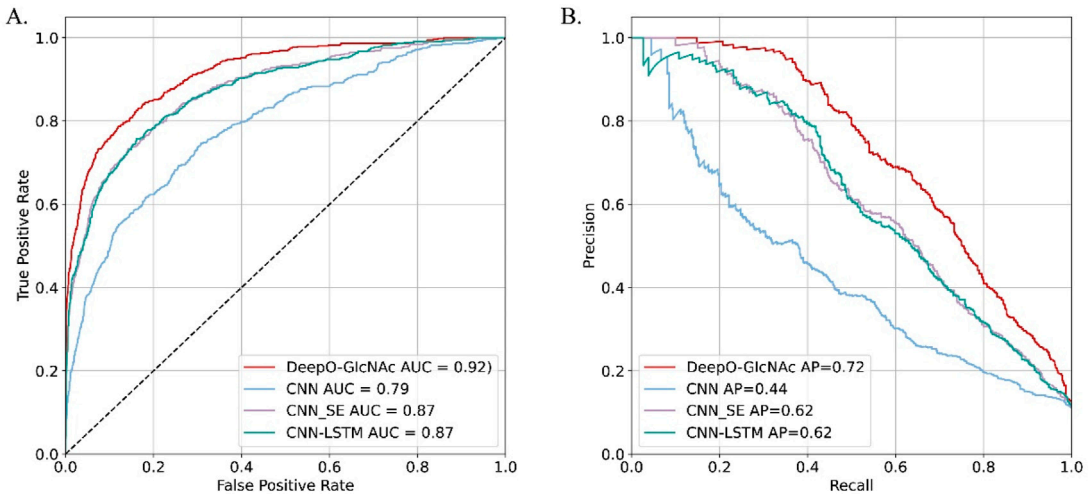

FIGURE 4
Ablation experiments on DeepO-GlcNAc. **(A)** ROC Curves for O-GlcNAc site prediction models. The ROC curves illustrate the performance of various computational models in predicting O-GlcNAcylation sites on the independent dataset including DeepO-GlcNAc, CNN, CNN_SE, CNN_LSTM. **(B)** Precision-recall curves for O-GlcNAc site prediction models. Precision-recall curves assess the precision against recall for the O-GlcNAcylation site prediction models including DeepO-GlcNAc, CNN, CNN_SE, CNN_LSTM.

TABLE 1 Results of the test data in ablation experiments.

	Sn	Sp	MCC	ACC	Precision	F1	AUC
CNN	0.47	0.91	0.35	0.86	0.39	0.43	0.79
CNN-SE	0.49	**0.96**	0.50	0.91	**0.62**	0.54	0.87
CNN-LSTM	0.54	0.95	0.50	0.90	0.57	0.55	0.87
DeepO-GlcNAc	**0.68**	0.95	**0.60**	**0.92**	0.61	**0.65**	**0.92**

Bold indicates the most significant value among the comparisons of different models.

TABLE 2 Statistical information on species apart from human species.

Species	Positive	Negative
mouse	180	11,240
Drosophila	101	6,212
Caenorhabditis-elegans	86	1,538
Arabidopsis	548	24,028
rats	454	10,375

Drosophila showed slightly lower accuracy values at 0.63 and 0.51. Specificity (Sp) was higher in *Arabidopsis* and *rat*, both achieving values above 0.60. Sensitivity (Sn) was highest in *Caenorhabditis-elegans* and mouse, with the former reaching 1.00. The ROC curves (Figure 5B) further reflect the model's performance, where *Caenorhabditis-elegans* displays the highest AUC of 0.83, indicating the most reliable predictions, followed by *Arabidopsis* with an AUC of 0.66. Conversely, *rat* had the lowest AUC of 0.50, suggesting limited predictive power for this species.

Performance comparison between DeepO-GlcNac and current predictors

To demonstrate the predictive capability and robustness of DeepO-GlcNAc, we conducted a performance comparison with other currently available predictors. We compared our model with two available web services and the most-recently released O-GlcNAc prediction tool, including YinOYang, NetOglyc-4.0, and GlcNAcPRED-DL. The independent test dataset was submitted to all the predictors and the results were compared parallelly. DeepO-GlcNAc outperformed all the tools in terms of accuracy (ACC), specificity (Sp), and AUC, Matthew's correlation coefficient (MCC), F1 score and Precision. It achieves an AUC value of 0.92, which is 24% higher than YinOYang, 10% higher than NetOglyc-4.0% and 11% higher than GlcNAcPRED-DL (Figure 6). Our model has the highest precision of 0.61, ACC of 92%, as well as the highest MCC of 0.60. Metrics related to class balance, such as precision-recall curves and the F1-score, DeepO-GlcNAc performs the best. These results highlight the advantages of DeepO-GlcNAc (Table 3). And we also provide the list of the independent dataset as a supporting material including which specific sites and proteins were successfully identified when using a certain tool for benchmarking (Supplementary Table S1).

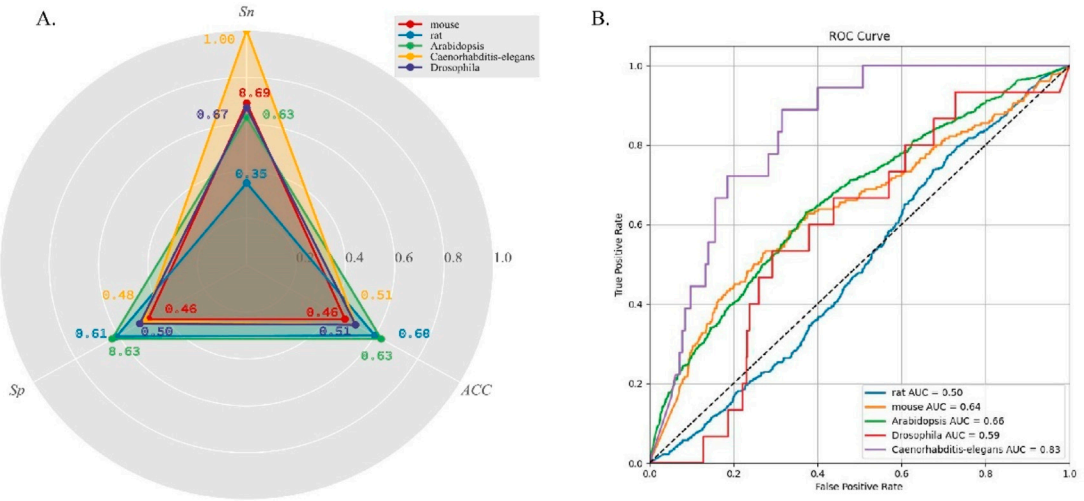

FIGURE 5
(A) Five cross-species datasets testing results on DeepO-GlcNAc. **(B)** Comparison of the ROC curves and AUC values of five cross-species prediction on DeepO-GlcNAc.

Implementation of the DeepO-GlcNAc webserver

To facilitate the usage of DeepO-GlcNAc by other researchers, we have developed a user-friendly web server based on DeepO-GlcNAc. The online service of DeepO-GlcNAc was constructed in an easy-to-use manner using Flask and HTML. The model was deployed in Tencent Cloud, which is equipped with 16 cores, 64 GB memory and a 2 TB hard disk. It was developed using the Windows Sever 2016-Flask-HTML open-source web platform and has been extensively tested on various web browsers including Internet Explorer, Mozilla, Firefox and Google Chrome to provide a robust service. For convenience, the online service of DeepO-GlcNAc was implemented and freely available at http://124.220.189.245:8000/.

Supplementary Figure S2 showcases the user interface of the server, along with an example of prediction output. The server is hosted on the Tencent cloud computing facility. The server utilizes DeepO-GlcNAc to identify O-GlcNAc sites within submitted protein sequences. On the index webpage, users can conveniently submit FASTA formatted protein sequences in the provided textbox. The prediction results include comprehensive information such as the positions of predicted modification sites, corresponding scores, and the overall prediction outcomes. Users also have the option to download the generated prediction results in plain text format for further analysis. In addition, the curated benchmark datasets and the independent test dataset used in our study are available for downloaded from the O-GlcNAc web server (Supplementary Figure S2).

Discussion

In this work, we built a dataset containing 700 experimentally validated O-glycoproteins from humans. Then, we specifically designed three neural network frameworks—CNN, LSTM, and Attention, respectively—to extract protein sequence features. In our analysis, based on the results of ablation experiments, the combination of CNN, LSTM and Attention presented the best performance. Both CNN and LSTM have been proved efficient in PTM prediction (Naseer et al., 2021). CNN excels at capture spatial patterns inherent in the input data. While LSTM networks are adept at capturing sequential patterns and long-range dependencies. We consider that both local sequence patterns and the temporal of these patterns are crucial for O-GlcNAc modification prediction. Therefore, LSTM complements this by processing the spatial feature maps across spatial dimensions to capture temporal dependencies among features for each sequence. Through this fusion, the model gains a comprehensive understanding of both spatial and sequential characteristics associated with O-GlcNAc modification sites, ultimately enhancing its predictive performance. We also deployed an attention mechanism which introduces an adaptive approach where the importance of each channel is individually assessed based on its context. It has been proved that attention mechanism yield substantial performance enhancements in state-of-the-art CNNs (Hu et al., 2020). In our model, the utilization of attention mechanism improved the model's AUC value from 0.87 in the CNN-LSTM architecture to 0.92. This suggests that certain amino acid patterns are more critical for O-GlcNAc prediction. Given that the detailed mechanism of protein O-GlcNAcylation is not fully clear, this information could be valuable for further investigation. Moreover, by suppressing redundant or irrelevant feature maps, the attention mechanism enhances the model's generalization ability and robustness. Consequently, LSTM can focus more on valuable feature information pertinent to the prediction task, thus enhancing the overall performance of the model.

Deep learning-based approaches have been widely applied to various types of PTM prediction, and their advantages have been well demonstrated (Meng et al., 2022; Naseer et al., 2022). However, the application of deep learning in O-GlcNAc prediction has not yet achieved significant success (Mauri et al., 2021). Previous attempts using the CNN for O-GlcNAc prediction did not yield substantial improvements (Hu et al., 2023; Zhu et al., 2022), possibly due to the relatively small dataset sizes and model construction limitations. In our study, we benchmarked five deep learning-based models and

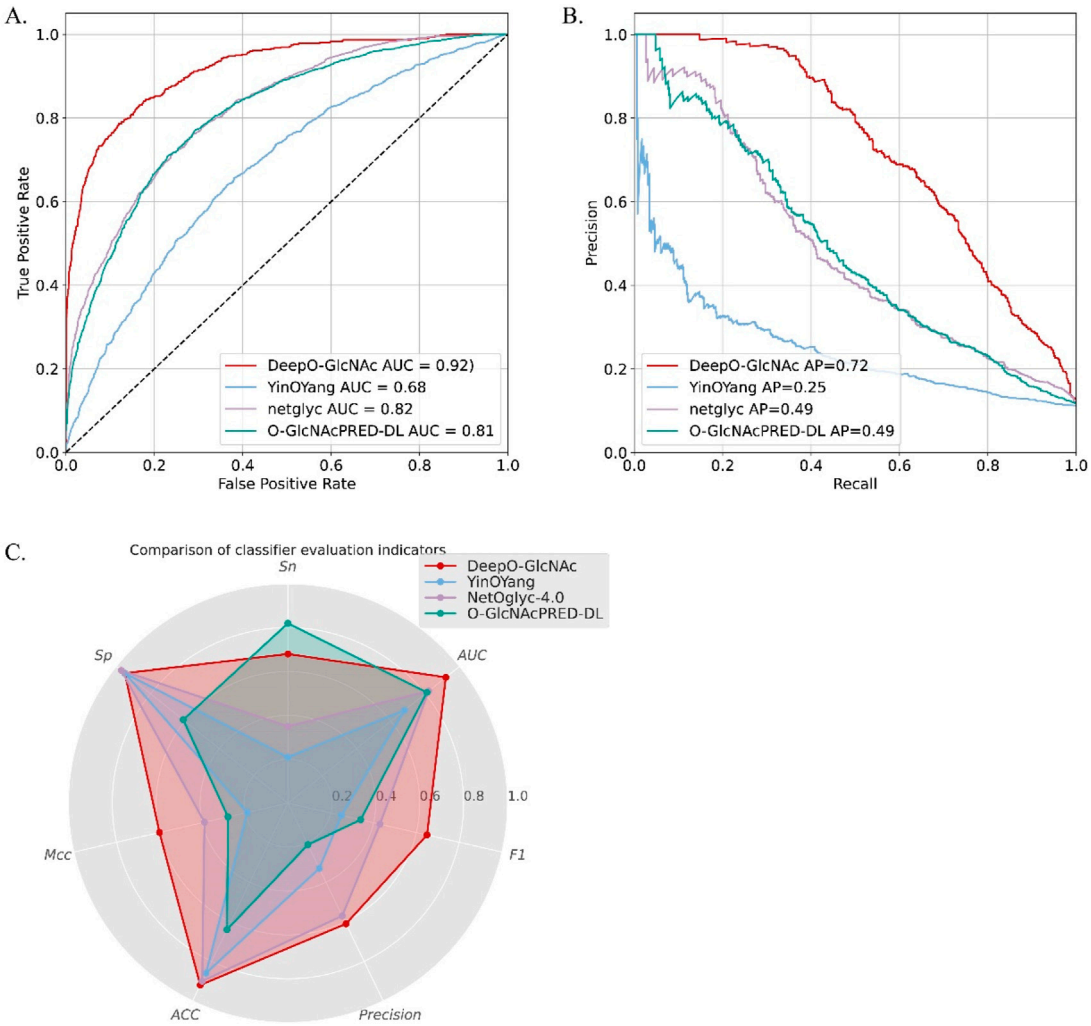

FIGURE 6
The performance comparison between DeepO-GlcNAc to three currently existing predictors, including YinOYang, NetOglyc-4.0, and GlcNAcPRED-DL. **(A)** The ROCs of different predictors are compared based on the independent test dataset. DeepO-GlcNAc presents the best AUC of 0.92 compared to YinOYang (0.68), NetOglyc-4.0 (0.82) and GlcNAcPRED-DL (0.81). **(B)** The PR plots of different predictors are compared. DeepO-GlcNAc presents the best AP of 0.72 compared to YinOYang (0.25), NetOglyc-4.0 (0.49) and GlcNAcPRED-DL (0.49). **(C)** The Radar plot displayed the metrics of Sp, Sn, AUC, F1, Precision, ACC and MCC. DeepO-GlcNAc, YingOyang, NetOglyc-4.0 and GlcNAcPRED-DL are shown in red, blue, purple, and green respectively. DeepO-GlcNAc ranks first in Precision, AUC, F1, ACC, and MCC, and second in Sn among the predictors.

TABLE 3 Performance comparison of DeepO-GlcNAc with other prediction models.

Tool	Sn	Sp	MCC	ACC	Precision	F1	AUC
DeepO-GlcNAc	0.68	0.95	**0.60**	**0.92**	**0.61**	**0.65**	**0.92**
YinOYang	0.21	0.95	0.19	0.86	0.33	0.25	0.68
NetOglyc-4.0	0.35	0.97	0.39	0.90	0.57	0.43	0.82
GlcNAcPRED-DL	**0.82**	0.61	0.28	0.64	0.21	0.34	0.81

Bold indicates the most significant value among the comparisons of different models.

demonstrated the potential of the CNN-Attention-LSTM fusion model through both independent testing and cross-validation. This indicates the feasibility of using deep learning in O-GlcNAc prediction and suggests that further optimization of deep learning models could enhance the prediction performance. Based on the test results of incorporating cross-species data from five organisms—mouse, *Drosophila*, *Caenorhabditis-elegans*, *Arabidopsis*, and rat, our model demonstrates strong generalization performance.

Despite the existence of several algorithms for O-GlcNAc prediction, the availability of public online services is still limited.

Currently, only a few O-GlcNAc prediction services are accessible, which hampers O-GlcNAc research. In this study, we developed a free online service platform based on the CNN-Attention-LSTM model. Compared to the other existing servers, our server demonstrated improved performance in sensitivity, specificity, and precision. Thus, our sever can be a new helpful tool for O-GlcNAc research.

There are still several limitations in our work. In the study, we used a dataset with 2696 O-GlcNAcylation sites experimentally validated. Due to the nature distribution, it is an imbalanced dataset with more negative peptides than positive ones. Even it may better reflect the actual situation, such imbalance may make deep learning algorithms tend to be biased toward the negative class. This could potentially explain why our model exhibited less sensitivity compared to O-GlcNAcPRED-DL, which utilized a balanced dataset. On the other hand, our predictor and another two predictors with imbalanced dataset presents better specificity compared to O-GlcNAcPRED-DL. Whether and how should we deal with such kind of dataset in PTM prediction need to be further investigated. In addition, we employ sliding window method to construct O-GlcNAc sites, which is commonly in PTM predication. However, this method introduces information redundancy and may lead to inefficient resource utilization. Moreover, it captures local sequence information, and neglects the overall structure within the global sequence. This limitation may be improved by optimizing the length of the window.

In summary, the developed DeepO-GlcNAc predictor achieved remarkable performance in O-GlcNAc site prediction. Its success demonstrates the feasibility of using deep learning for O-GlcNAc prediction, and the online predictor service provides a valuable tool for future research in this field.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

LZ: Writing–original draft, Visualization, Investigation, Formal Analysis. TD: Writing–original draft, Validation, Formal Analysis. SP: Writing–original draft, Investigation, Data curation. MZ: Writing–review and editing, Software, Data curation. YZ: Writing–review and editing, Methodology. CY: Writing–review and editing, Supervision, Funding acquisition. XY: Writing–review and editing, Methodology, Conceptualization. GT: Writing–review and editing, Funding acquisition, Conceptualization. JM: Writing–review and editing, Supervision, Funding acquisition, Conceptualization.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was funded by National Natural Science Foundation of China [32170989], Taishan Scholars Construction Engineering, Major Basic Research Project of Shandong Provincial Natural Science Foundation [ZR2019ZD27], Key R&D Program of Shandong Province, China [2023CXPT012], and Natural Science Foundation of Shandong Province [ZR2021MH141].

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2024.1456728/full#supplementary-material

References

Arnold, C. S., Johnson, G. V., Cole, R. N., Dong, D. L., Lee, M., and Hart, G. W. (1996). The microtubule-associated protein tau is extensively modified with O-linked N-acetylglucosamine. *J. Biol. Chem.* 271, 28741–28744. doi:10.1074/jbc.271.46.28741

Bartolome-Nebreda, J. M., Trabanco, A. A., Velter, A. I., and Buijnsters, P. (2021). O-GlcNAcase inhibitors as potential therapeutics for the treatment of Alzheimer's disease and related tauopathies: analysis of the patent literature. *Expert Opin. Ther. Pat.* 31, 1117–1154. doi:10.1080/13543776.2021.1947242

Chen, Z., Liu, X., Li, F., Li, C., Marquez-Lago, T., Leier, A., et al. (2019). Large-scale comparative assessment of computational predictors for lysine post-translational modification sites. *Brief. Bioinform* 20, 2267–2290. doi:10.1093/bib/bby089

Conibear, A. C. (2020). Deciphering protein post-translational modifications using chemical biology tools. *Nat. Rev. Chem.* 4, 674–695. doi:10.1038/s41570-020-00223-8

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* 36, 193–202. doi:10.1007/BF00344251

Gupta, R., and Brunak, S. (2002). Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac. Symposium Biocomput.*, 310–322. doi:10.1142/9789812799623_0029

Hart, G. W., Housley, M. P., and Slawson, C. (2007). Cycling of O-linked beta-N-acetylglucosamine on nucleocytoplasmic proteins. *Nature* 446, 1017–1022. doi:10.1038/nature05815

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735

Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E. (2017). Squeeze-and-Excitation Networks. *arXiv:1709.01507*. 01507. doi:10.48550/arXiv.1709.01507

Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E. (2020). Squeeze-and-Excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 2011–2023. doi:10.1109/TPAMI.2019.2913372

Hu, F., Li, W., Li, Y., Hou, C., Ma, J., and Jia, C. (2023). O-GlcNAcPRED-DL: prediction of protein O-GlcNAcylation sites based on an ensemble model of deep learning. *J. proteome Res.* 23, 95–106. doi:10.1021/acs.jproteome.3c00458

Khan, Y. D., Khan, N. S., Naseer, S., and Butt, A. H. (2021). iSUMOK-PseAAC: prediction of lysine sumoylation sites using statistical moments and Chou's PseAAC. *PeerJ* 9, e11581. doi:10.7717/peerj.11581

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv: 1408.5882*. doi:10.48550/arXiv.1408.5882

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinforma. Oxf. Engl.* 22, 1658–1659. doi:10.1093/bioinformatics/btl158

Li, F., Li, C., Wang, M., Webb, G. I., Zhang, Y., Whisstock, J. C., et al. (2015). GlycoMine: a machine learning-based approach for predicting N-C- and O-linked glycosylation in the human proteome. *Bioinforma. Oxf. Engl.* 31, 1411–1419. doi:10.1093/bioinformatics/btu852

Li, F., Li, C., Revote, J., Zhang, Y., Webb, G. I., Li, J., et al. (2016). GlycoMinestruct: a new bioinformatics tool for highly accurate mapping of the human N-linked and O-linked glycoproteomes by incorporating structural features. *Sci. Rep.* 6, 34595. doi:10.1038/srep34595

Li, Z., Fang, J., Wang, S., Zhang, L., Chen, Y., and Pian, C. (2022a). Adapt-Kcr: a novel deep learning framework for accurate prediction of lysine crotonylation sites based on learning embedding features and attention architecture. *Briefings Bioinforma.* 23. doi:10.1093/bib/bbac037

Li, Z., Li, S., Luo, M., Jhong, J. H., Li, W., Yao, L., et al. (2022b). dbPTM in 2022: an updated database for exploring regulatory networks and functional associations of protein post-translational modifications. *Nucleic acids Res.* 50, D471–d479. doi:10.1093/nar/gkab1017

Liu, F., Iqbal, K., Grundke-Iqbal, I., Hart, G. W., and Gong, C. X. (2004). O-GlcNAcylation regulates phosphorylation of tau: a mechanism involved in Alzheimer's disease. *Proc. Natl. Acad. Sci. U. S. A.* 101, 10804–10809. doi:10.1073/pnas.0400348101

Mauri, T., Menu-Bouaouiche, L., Bardor, M., Lefebvre, T., Lensink, M. F., and Brysbaert, G. (2021). O-GlcNAcylation prediction: an unattained objective. *Adv. Appl. Bioinforma. Chem. AABC* 14, 87–102. doi:10.2147/aabc.s294867

Meng, L., Chan, W. S., Huang, L., Liu, L., Chen, X., Zhang, W., et al. (2022). Mini-review: recent advances in post-translational modification site prediction based on deep learning. *Comput. Struct. Biotechnol. J.* 20, 3522–3532. doi:10.1016/j.csbj.2022.06.045

Morris, M., Knudsen, G. M., Maeda, S., Trinidad, J. C., Ioanoviciu, A., Burlingame, A. L., et al. (2015). Tau post-translational modifications in wild-type and human amyloid precursor protein transgenic mice. *Nat. Neurosci.* 18, 1183–1189. doi:10.1038/nn.4067

Naseer, S., Hussain, W., Khan, Y. D., and Rasool, N. (2021). Optimization of serine phosphorylation prediction in proteins by comparing human engineered features and deep representations. *Anal. Biochem.* 615, 114069. doi:10.1016/j.ab.2020.114069

Naseer, S., Ali, R. F., Fati, S. M., and Muneer, A. (2022). Computational identification of 4-carboxyglutamate sites to supplement physiological studies using deep learning. *Sci. Rep.* 12, 128. doi:10.1038/s41598-021-03895-4

Ning, Q., and Li, J. (2022). DLF-Sul: a multi-module deep learning framework for prediction of S-sulfinylation sites in proteins. *Briefings Bioinforma.* 23. doi:10.1093/bib/bbac323

Ramazi, S., and Zahiri, J. (2021). Post-translational modifications in proteins: resources, tools and prediction methods. *Database* 2021. doi:10.1093/database/baab012

Ruan, H. B., Han, X., Li, M. D., Singh, J. P., Qian, K., Azarhoush, S., et al. (2012). O-GlcNAc transferase/host cell factor C1 complex regulates gluconeogenesis by modulating PGC-1α stability. *Cell Metab.* 16, 226–237. doi:10.1016/j.cmet.2012.07.006

Slawson, C., and Hart, G. W. (2011). O-GlcNAc signalling: implications for cancer cell biology. *Nat. Rev. Cancer* 11, 678–684. doi:10.1038/nrc3114

Smet-Nocca, C., Broncel, M., Wieruszeski, J. M., Tokarski, C., Hanoulle, X., Leroy, A., et al. (2011). Identification of O-GlcNAc sites within peptides of the Tau protein and their impact on phosphorylation. *Mol. Biosyst.* 7, 1420–1429. doi:10.1039/c0mb00337a

Wang, J., Torii, M., Liu, H., Hart, G. W., and Hu, Z.-Z. (2011). dbOGAP - an integrated bioinformatics resource for protein O-GlcNAcylation. *BMC Bioinforma.* 12, 91. doi:10.1186/1471-2105-12-91

Wang, C., Tan, X., Tang, D., Gou, Y., Han, C., Ning, W., et al. (2022). GPS-Uber: a hybrid-learning framework for prediction of general and E3-specific lysine ubiquitination sites. *Briefings Bioinforma.* 23. doi:10.1093/bib/bbab574

Wen, B., Zeng, W. F., Liao, Y., Shi, Z., Savage, S. R., Jiang, W., et al. (2020). Deep learning in proteomics. *Proteomics* 20, e1900335. doi:10.1002/pmic.201900335

Yang, X., and Qian, K. (2017). Protein O-GlcNAcylation: emerging mechanisms and functions. *Nat. Rev. Mol. Cell Biol.* 18, 452–465. doi:10.1038/nrm.2017.22

Yang, Y., Fu, M., Li, M. D., Zhang, K., Zhang, B., Wang, S., et al. (2020). O-GlcNAc transferase inhibits visceral fat lipolysis and promotes diet-induced obesity. *Nat. Commun.* 11, 181. doi:10.1038/s41467-019-13914-8

Zhu, Y., Yin, S., Zheng, J., Shi, Y., and Jia, C. (2022). O-glycosylation site prediction for *Homo sapiens* by combining properties and sequence features with support vector machine. *J. Bioinforma. Comput. Biol.* 20, 2150029. doi:10.1142/s0219720021500293