Using Protein Language Models Embeddings to predict O-GlcNAc glycosylation sites

Adenilson Arcanjo[1,2], Diego Mariano[2], Luana L. Bastos[2], Ana L. A. Bastos[2], Milenna Pirovani[2], Raquel C. de Melo-Minardi[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), Campus Sobral, Sobral, Brazil.

[2] Laboratory of Bioinformatics and Systems, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil.

adenilson.junior@ifce.edu.br, diego@dcc.ufmg.br, luizabastos.luana9@gmail.com, alab@ufmg.br, milennapirovani@ufmg.br, raquelcm@dcc.ufmg.br

Abstract. *O-GlcNAcylation is a post-translational modification (PTM) that involves the covalent bonding of an N-acetylglucosamine (GlcNAc) molecule to serine or threonine amino acid residues in nuclear and cytoplasmic proteins. PTMs dysregulation has been implicated in a wide range of diseases, including cancer, metabolic syndromes, and neurodegenerative disorders. Precise mapping of O-GlcNAc sites is essential for advancing both fundamental understanding and the development of targeted therapeutics. However, their detection remains challenging, which has motivated the development of computational tools to predict these sites with greater accuracy. In this study, we used Protein Language Models (PLMs) to address the challenge of predicting protein residues that are O-GlcNAc modification sites. To evaluate our method, we collected data from the O-GlcNAc Atlas. Our results indicate that our model outperformed competitors in all datasets evaluated. We believe the approach presented here can benefit scientists working on any subject where protein post-translational modifications play a role.*

Keywords*: O-GlcNAcylation, Machine Learning, Protein Language Models, Embeddings.*

1. Introduction

Proteins are essential macromolecules that perform a vast array of biological functions, ranging from enzymatic catalysis and structural support to signal transduction and molecular transport [Morris et al. 2022]. Understanding protein behavior is crucial for deciphering biological mechanisms at the molecular level and for advancing research in fields such as biology, biotechnology, and medicine [Stollar and Smith 2020].

Among the mechanisms that regulate protein function, post-translational modifications (PTMs) are chemical alterations that occur after protein synthesis, playing a fundamental role in expanding the functional diversity of the proteome [Spoel 2018]. By modulating protein structure, localization, stability, and interactions, PTMs regulate multiple aspects of cellular physiology. Dysregulation of PTMs has been implicated in a wide range of diseases, including cancer [Singh et al. 2020], neurodegenerative disorders [Stollar and Smith 2020], and metabolic syndromes [Yang et al. 2023],

highlighting their essential role in maintaining cellular homeostasis. One example of a post-translational modification is O-linked β-N-acetylglucosamine (O-GlcNAc).

O-GlcNAcylation involves the covalent bonding of an N-acetylglucosamine (GlcNAc) molecule to serine or threonine amino acid residues in nuclear and cytoplasmic proteins [Yang and Qian 2017]. Unlike other glycosylation forms that produce extended glycan chains, O-GlcNAcylation resembles phosphorylation in function, regulating critical cellular processes such as transcriptional control and signal transduction [Yang et al. 2020; Yang and Qian 2017]. The modification is catalyzed by O-GlcNAc transferase (OGT) and removed by O-GlcNAcase (OGA) (Figure 1), thereby serving as a sensitive regulator of cellular status responses [Yang et al. 2020].

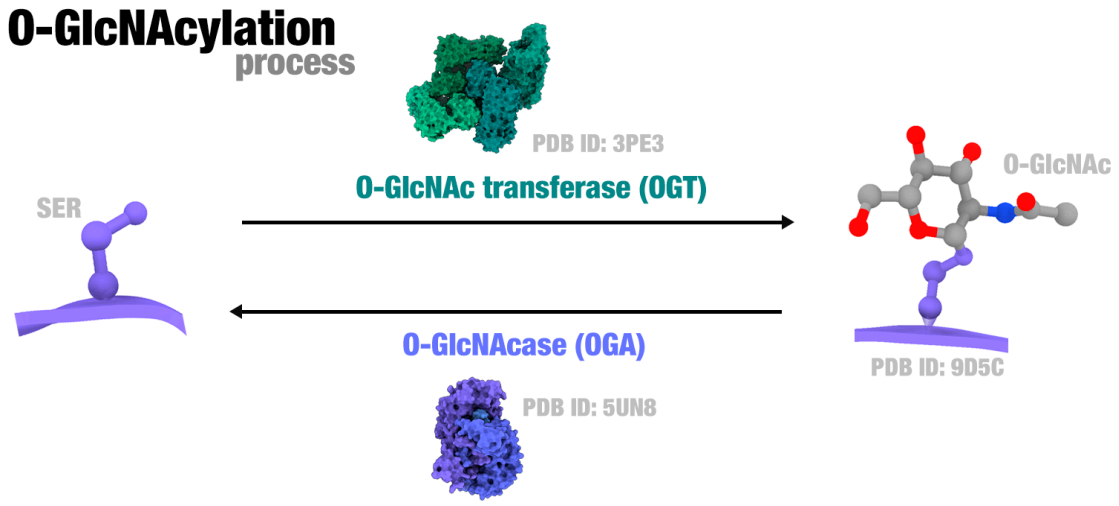

Figure 1. Role of O-GlcNAc transferase (OGT) and O-GlcNAcase (OGA) enzymes in the O-GlcNAcylation process. Figure generated using ChimeraX [Meng et al. 2023].

Aberrant O-GlcNAcylation has been implicated in a spectrum of pathologies, including diabetes mellitus, oncogenesis, and neurodegenerative diseases [Hart et al. 2007; Slawson and Hart 2011; Smet-Nocca et al. 2011]. For example, in Alzheimer's disease, O-GlcNAcylation at serine 400 of the Tau protein diminishes phosphorylation at serine 404, disrupting the sequential phosphorylation cascade mediated by GSK3β and highlighting OGA inhibition as a promising therapeutic approach [Smet-Nocca et al. 2011]. Furthermore, in metabolic disorders such as diabetes, O-GlcNAcylation of the coactivator PGC-1α at serine 333 confers protection against ubiquitin-mediated proteasomal degradation, suggesting novel intervention strategies [Ruan et al. 2012].

Precise mapping of O-GlcNAc sites is therefore essential for advancing both fundamental understanding and the development of targeted therapeutics [Zhang et al. 2024]. Advances in high-throughput techniques, such as mass spectrometry proteomics, have led to the identification of specific O-GlcNAcylation sites in proteins [Zhang et al. 2024; Seber and Braatz 2024]. However, their detection remains challenging, which has motivated the development of computational tools to predict these sites with greater accuracy [Khalid et al. 2024].

In recent years, Large language models (LLMs) have significantly advanced not only the field of Natural Language Processing but also numerous other scientific disciplines [Zhao et al. 2023]. An LLM is a machine learning model trained on immense volumes of text data, allowing it to learn the intricate patterns, context, and relationships inherent in human language. The key innovation enabling this is the Transformer architecture, which captures contextual relationships between elements in a sequence using attention mechanisms [Vaswani et al. 2017]. By training on massive text corpora, models like BERT [Devlin et al. 2019], GPT [Radford et al. 2018; Radford et al. 2019; Brown et al. 2020], and their successors learn to assign each element in a sequence a vector representation, known as an embedding, which captures both semantic and syntactic information. These embeddings serve as compact, information-rich inputs for downstream models in a wide range of tasks.

This same idea has been extended to biological data, particularly protein sequences, through the development of Protein Language Models (PLMs). In this context, amino acid residues are treated analogously to words, and protein sequences are modeled as sentences. These models are trained on massive databases containing hundreds of millions of amino acid sequences, learning to generate embeddings that capture structural, functional, and evolutionary signals from sequence alone.

Among the most prominent models in this domain is the Evolutionary Scale Modeling (ESM) family, developed by Meta AI [Lin et al. 2023]. The ESM-2, the second version of this family of models, was trained on the large protein sequence database UniRef50 [Suzek et al. 2015] and produces residue-level embeddings that capture rich biochemical and evolutionary context. These learned representations have been successfully applied to various prediction tasks, including secondary structure inference, contact map estimation, and the analysis of mutation effects.

A key strength of Protein Language Models is their ability to make accurate predictions without relying on multiple sequence alignments (MSAs), which are often computationally expensive and not always available. As noted by [Weissenow and Rost 2025], models like ESM learn the "language" of proteins, providing a general-purpose, alignment-free framework for modeling protein behavior. Because of the depth of information encoded in their embeddings, even simple classifiers, such as multi-layer perceptrons (MLPs), can achieve competitive results, making this a practical and scalable solution for many bioinformatics problems.

Several computational tools have been developed to predict O-GlcNAc modification sites using a range of machine learning algorithms, including DeepO-GlcNAc [Zhang et al. 2024], O-GlcNAcPRED-DL [Hu et al. 2023], and LM-OGlcNAc-Site [Pokharel et al. 2023], the model by [Khalid et al. 2024], among others. Early models primarily relied on manually curated features derived from the protein sequence, including amino acid composition, sequence motifs, and physicochemical properties. More recent approaches have incorporated position-specific scoring matrices (PSSMs) or leveraged deep learning architectures to improve prediction performance. However, many of these methods depend on multiple sequence alignments (MSAs) or external structural annotations, which may hinder their scalability and general applicability.

We note that the work of [Khalid et al. 2024] used ESM2 embeddings, however their approach had two key limitations: they restricted their model to sequences between 500 and 800 residues, and they used a tokenization that encodes Serine and Threonine as 'G' and all other amino acids as 'N', this approach significantly reduces the valuable contextual information of the protein sequence. In [Pokharel et al. 2023] the authors used embeddings from multiple protein language models: Ankh [Elnaggar et al. 2023], ProtT5 [Heinzinger et al. 2024] and ESM2, having embeddings dimensions of 1536, 2560 and 1024, respectively, totalling an input entry of dimension 5120. These embeddings were processed by three fully connected neural networks that together have ~2.9 million parameters. The use of such a large number of parameters can make the model prone to overfitting, where it may memorize the training data rather than learning generalizable patterns.

In this work, we leverage the power of PLMs embeddings to address the challenge of predicting protein residues that are sites of O-GlcNAc modification. Using as input the embeddings from the smallest model of the ESM2 family, which have dimension 320, we achieve a strong performance in predicting O-GlcNAc sites across various eukaryotic species. Our model uses only ~5.8 thousand parameters and the only restriction is that the input protein sequences must have a maximum length of 1022 residues.

2. Materials and methods

Figure 2 presents an overview of the methodology adopted in this work. Initially, we collected and balanced the dataset, separating it into training (80%), testing (10%), and validation (10%) sets. We then built a neural network model, evaluated it, and finally compared it with similar methods. Details of the procedures adopted will be presented in the following sections.

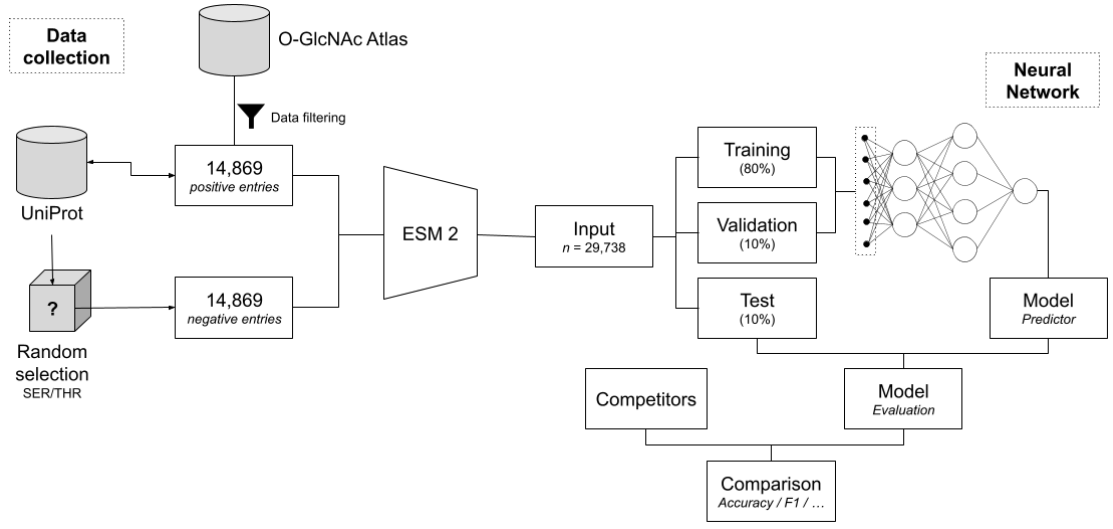

Figure 2. Method overview.

2.1 Data collection

We collected 14,869 O-GlcNAc sites (positive entries) from the O-GlcNAc Atlas [Hou et al. 2025]. The other 14,869 entries were randomly selected from other serine or threonine residues in the sequences of the positive entries to serve as negative controls (for a final dataset of $n = 29{,}738$).

O-GlcNAc Atlas is a curated database of experimentally verified O-GlcNAc sites. The database, last updated on December 17, 2024, contains over 8,000 proteins, including more than 4,000 human proteins. It is divided into unambiguously identified sites and ambiguously identified O-GlcNAc sites, comprising 37,551 and 12,629 sites, respectively. We chose the unambiguously identified sites as our dataset to ensure data reliability and minimize the inclusion of potential false positives.

The database originally consisted of 37,098 entries, each one corresponding to an experimentally observed O-GlcNAc modification site on a serine or threonine residue within a protein. Each entry had a UniProt Accession [The UniProt Consortium 2023] or an MSU ID. The origin of the MSU ID, even after consulting the database's associated publication, could not be ascertained. To ensure data quality and traceability, only entries with a UniProt Accession were kept.

The curation process involved verifying whether the reported O-GlcNAc position in each entry corresponded to a serine or threonine residue in the associated UniProt sequence. Entries were excluded due to missing UniProt accession numbers (5,256), non-numerical O-GlcNAc positions (47), positions exceeding the protein length (104), positions not corresponding to serine or threonine (3,659), and mismatches between the expected and actual residue at the indicated position (995). After this filtering, 30,696 valid entries remained.

Given the length limitation of 1,022 amino acids of the ESM2 model [Lin et al. 2023], proteins exceeding this length were excluded, leaving 14,869 entries in the database. Subsequently, for each protein sequence in the database, the experimentally identified O-GlcNAc sites were designated as positive examples. Other serine or threonine residues present in the sequences were designated as negative examples. This labeling revealed an imbalance, as only 4.7% of serine/threonine residues were identified as O-GlcNAc sites, resulting in a database with 14,869 positive examples and 301,794 negative examples. To address this imbalance, a random subset of 14,869 negative entries was selected. We kept the same number of positive and negative examples for each species, whenever possible. Thus, the final dataset consisted of 29,738 entries, having an equal number of positive and negative examples. This dataset was randomly partitioned into training, validation, and test datasets, with sizes of 80%, 10%, and 10%, respectively. The balance between positive and negative samples for each species was preserved in all datasets, whenever possible.

2.2 Model construction

We built a neural network model to predict O-GlcNAc sites using ESM2 embeddings. The ESM2 embeddings are high-dimensional vector representations of protein sequences generated by the Evolutionary Scale Modeling 2 model [Lin et al. 2023], a family of large protein language models built upon the Transformer architecture. These

models are trained on the large unannotated protein sequence database UniRef50 [Suzek et al. 2015] using a masked language modeling objective, that is, the model learns to predict masked amino acids based on the other amino acids of the sequence. The resulting embeddings, available per amino acid, are information-rich encodings of the protein, including structure and function. For this study, we selected the smallest model of the ESM2 family to avoid overfitting, since the larger models have embeddings of larger dimensions. This model has 8 million parameters distributed along six layers. For each amino acid of a protein sequence, this model creates a 320-dimensional representation. These representations will be used as the input for our model.

For downstream classification tasks utilizing protein language models, three primary strategies are commonly employed: full fine-tuning coupled with a classification head, parameter-efficient fine-tuning (PEFT) coupled with a classification head, or simply using the pre-trained language model as a fixed feature extractor by freezing its parameters and training only a classification head. The last approach is undoubtedly the simpler and quicker way of dealing with protein embeddings while yielding strong results, commonly surpassing the other methods [Sledzieski et al. 2024]. This balance between performance and computational efficiency justified our selection of this method.

Our model consists of a fully connected neural network, with an input layer of 320 neurons, a hidden layer of 18 neurons, and an output layer of one neuron. The first layer dimension was defined by the dimension of the input, the ESM2 embeddings, and the last layer dimension was determined by the nature of the expected output, a single number between 0 and 1, that can be interpreted as the confidence of the model that this particular residue is an O-GlcNAc site. The hidden layer size was chosen to be the geometric mean of the input size and output size, a simple way to select an intermediate size between these two values.

The model was implemented in PyTorch with a training batch size of 32. The Adam optimizer was used with a learning rate of 10^{-3} for 25 epochs, 10^{-4} for the next 25 epochs, and 10^{-5} for the final 25 epochs, for a total of 75 epochs. Binary Cross Entropy Loss was employed. To discourage overfitting, dropout with a rate of 0.5 was applied on the hidden layer. While the training and validation losses appeared to stabilize in later epochs, the model parameters from epoch 47, corresponding to the lowest validation loss, were chosen as the final model. The model performance was assessed by the accuracy, precision, recall, and F1-score.

3. Results and discussion

The model proposed in this study achieved an accuracy of ~76% in O-GlcNAc site prediction tests (Table 1). Additionally, the model demonstrated the capability for generalization, as evidenced by only a minor reduction in evaluation metrics (Table 1) when applied to unseen data, specifically the testing and validation datasets. These metrics indicate the model's ability to identify O-GlcNAc sites, including novel sites that have not been experimentally validated yet.

Table 1. Prediction results

	Accuracy	**Precision**	**Recall**	**F1-score**

Training	78.82%	80.46%	76.11%	78.23%
Validation	76.91%	77.41%	76.01%	76.70%
Testing	76.06%	77.66%	73.17%	75.35%

Among the five most represented species in the database: human (60%), mouse (25.2%), rat (4.1%), yeast (3.7%), and wheat (3.1%), we found that a larger number of database entries corresponds to a higher accuracy on the testing dataset (Figure 3). This finding is consistent with the principles of machine learning, where model performance is expected to improve with the use of more training data.

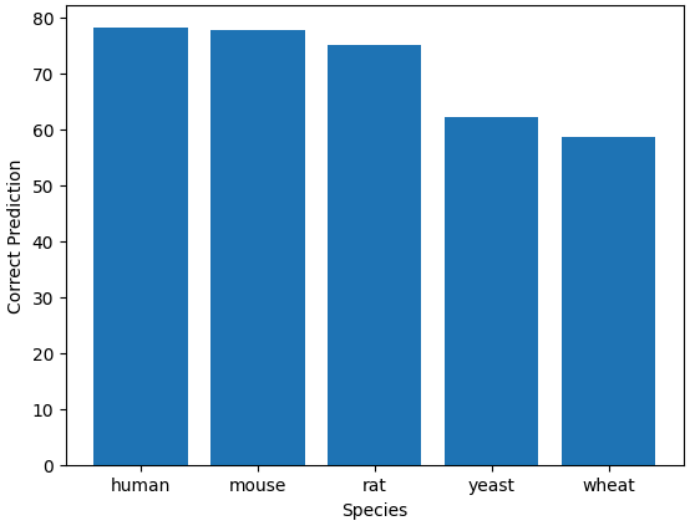

Figure 3. Accuracy of the model in the test dataset, expressed as a percentage, for the species with the most entries.

For a more detailed analysis, we selected the five species with the highest number of entries in our database (Figure 4). We observed a higher number of predicted new O-GlcNAc sites in rat and mouse species. We hypothesize this due to the close phylogenetic proximity between these species, facilitating the transfer of O-GlcNAc sites knowledge within the model. Conversely, the smallest number of predicted discoveries was for wheat and yeast, which aligns with the fact that plant entries comprise no more than 6% and fungal entries no more than 4% of the total entries in the database.

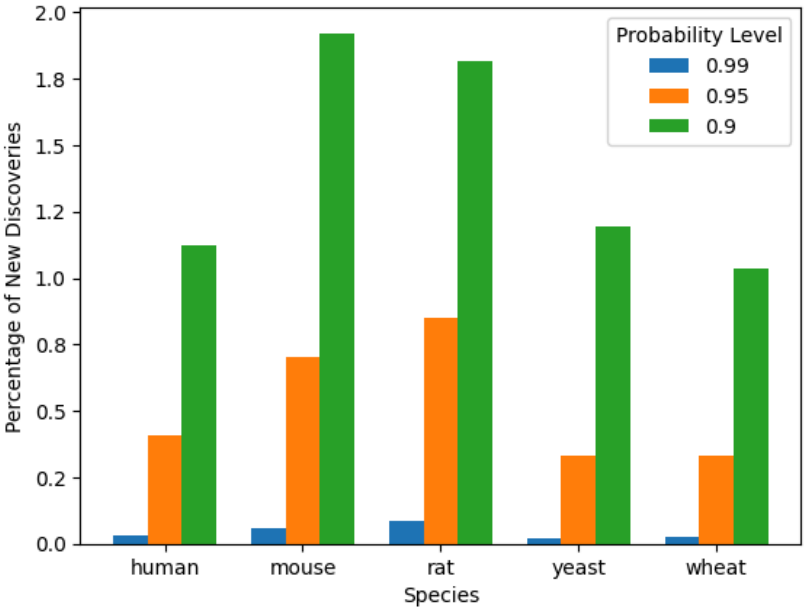

Figure 4. Percentage of Predicted Discoveries by Species and Probability Level.

We also generated lists of putative new O-GlcNAc sites using various probability cutoffs derived from our model. Specifically, we identified 69,798 new entries with a probability of 0.5 or higher, 4,551 entries with a probability of 0.9 or higher, 1,661 entries with a probability of 0.95 or higher, and 122 entries with a probability of 0.99 or higher. These lists can be found in the supplementary material. In Figure 5 experimentally verified and predicted O-GlcNAc sites are shown on the structure of three proteins selected as examples.

The proteins with the highest absolute number of predicted new O-GlcNAc sites, under a probability cutoff of 0.9, were the NUP58 for both human and mouse, with 41 and 40 newly identified post-translational modification sites, respectively. This protein, also known as Nucleoporin 58, is a component of the nuclear pore complex (NPC) and has a molecular mass of 58 kDa. All the molecules that enter or exit the nucleus must either diffuse through or be actively transported by the NPC [Hartono et al. 2019].

Among the proteins with the highest relative number of predicted O-GlcNAc sites, under a probability cutoff of 0.9, there is the Gamma-synuclein for mouse and rat, where at least half of their serine/threonine residues were identified as new O-GlcNAc sites. The normal cellular function of Gamma-synunclein remains unknown. Yet, it is known that it is predominantly located in the peripheral nervous system and retina [George 2001], it is a marker for breast cancer progression [Bruening et al. 2000], and changes in its expression in the retina of Alzheimer's patients have been observed [George 2001].

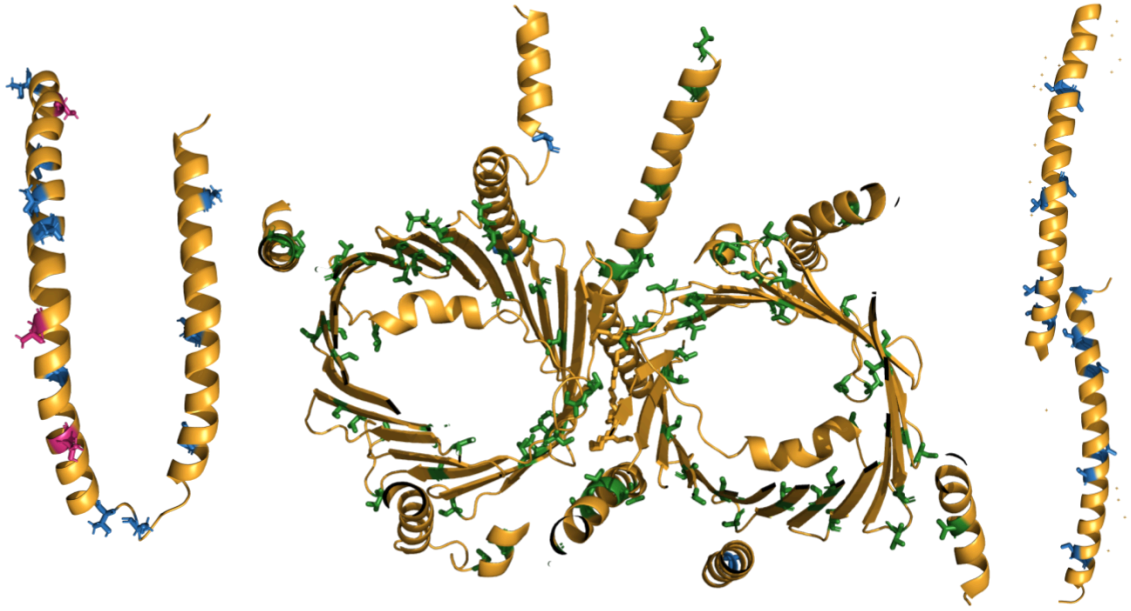

Figure 5. Left: PDB Structure 1XQ8 of Human Micelle-Bound Alpha-Synuclein. Center: PDB Structure 7CK6 of Human Protein Translocase of Mitochondria. Right: PDB Structure 3CI9 of Human HSBP1. In all protein structures, serine and threonine residues are colored accordingly: experimentally verified sites of O-GlcNAc are pink, predicted by our model are colored blue, and others are colored green.

3.1 Comparison to other tools

We compared our model with other available prediction tools using our testing dataset. It is crucial to note that no entries in the testing dataset were used for training or validation of our model, although they could have been used for these purposes in other models.

During our evaluation, YinOYang [Gupta 2001], DeepO-GlcNAc [Zhang et al. 2024] and LM-OGlcNAc-Site [Pokharel et al. 2023] servers were all inaccessible or unresponsive, while [Khalid et al. 2024] does not provide a server or other means of accessing their model. Consequently, O-GlcNAcPRED-DL [Hu et al. 2024] was the only available model for direct comparison. Since O-GlcNAcPRED-DL only accepts human or mouse sequences, we manually submitted the sequences of these organisms present in our testing dataset through their web server (https://oglcnac.org/pred_dl/).

Our model showed superior performance in both cases (Table 2). For human sequences, our model achieved an accuracy of 76.97%, compared to 67.78% for the O-GlcNAcPRED-DL model. Similarly, for mouse sequences, our model achieved an accuracy of 76.49%, while their model achieved an accuracy of 68.28%. We highlight that our model was trained not only for human and mouse sequences, but also for several other species.

Table 2. Comparison to other tools

Model	Human			Mouse		
	Accuracy	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
This study.	76.97%	75.98%	73.12%	76.49%	73.68%	77.78%
O-GlcNAcPRED-DL	67.78%	62.75%	73.86%	68.28%	63.14%	78.17%

4. Conclusion

In this study, we presented a deep learning model for predicting O-GlcNAc sites, a crucial post-translational modification involved in fundamental aspects of cellular biology. We compared it with available models built for the same purpose and obtained superior results, demonstrating improved accuracy in identifying these sites (accuracy of ~77%, compared to ~68% obtained by our competitor). Additionally, our model was built upon the embeddings of a protein language model, which, through training on extensive sequence data, can implicitly encode evolutionary information in a high-dimensional vector representation of protein residues. These information-rich representations allow downstream lightweight models to be strong, impactful tools. We believe this approach can benefit scientists working on any subject where protein post-translational modifications play a role. In future work, we intend to develop user-friendly web interfaces for predicting O-GlcNAc sites in proteins. It is important to emphasize that while the results presented here require experimental validation, they represent a promising alternative for in silico detection of O-GlcNAc sites. Additionally, they may have the potential to advance many areas of molecular biology and biomedicine, including the understanding of important diseases.

Acknowledgments: The authors thank the funding agencies: CAPES, FAPEMIG, and CNPq. The first author thanks IFCE.

Supplementary material: Supplementary tables and figures used in this study are available at https://github.com/LBS-UFMG/OGlcNAc.

Data availability: The dataset is accessible at https://oglcnac.org/.

Funding: This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Conflict of Interest: none declared.

5. References

Brown, T.B., Mann, B., Ryder, N., et al. 2020. Language Models are Few-Shot Learners. http://arxiv.org/abs/2005.14165.

Bruening, W., Giasson, B.I., Klein-Szanto, A.J.P., Lee, V.M.-Y., Trojanowski, J.Q., and Godwin, A.K. 2000. Synucleins are expressed in the majority of breast and ovarian carcinomas and in preneoplastic lesions of the ovary. *Cancer 88*, 9, 2154–2163.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. http://arxiv.org/abs/1810.04805.

Elnaggar, A., Essam, H., Salah-Eldin, W., Moustafa, W., Elkerdawy, M., Rochereau, C., and Rost, B.

2023. Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling. http://arxiv.org/abs/2301.06568.

George, J.M. 2001. The synucleins. *Genome Biology 3*, 1, reviews3002.1.

Gupta, R. 2001. Prediction of glycosylation sites in proteomes: from post-translational modifications to protein function. .

Hart, G.W., Housley, M.P., and Slawson, C. 2007. Cycling of O-linked beta-N-acetylglucosamine on nucleocytoplasmic proteins. *Nature 446*, 7139, 1017–1022.

Hartono, Hazawa, M., Lim, K.S., Dewi, F.R.P., Kobayashi, A., and Wong, R.W. 2019. Nucleoporin Nup58 localizes to centrosomes and mid-bodies during mitosis. *Cell Division 14*, 1, 7.

Heinzinger, M., Weissenow, K., Gomez Sanchez, J., Henkel, A., Mirdita, M., Steinegger, M., and Rost, B. 2024. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics* 6, 4, lqae150.

Hou, C., Li, W., Li, Y., and Ma, J. 2025. O-GlcNAcAtlas 4.0: An Updated Protein O-GlcNAcylation Database with Site-specific Quantification. *Journal of Molecular Biology 437*, 15, 169033.

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. *Journal of Proteome Research 23*, 1, 95–106.

Khalid, A., Kaleem, A., Qazi, W., Abdullah, R., Iqtedar, M., and Naz, S. 2024. Site-specific prediction of O-GlcNAc modification in proteins using evolutionary scale model. *PLOS ONE 19*, 12, e0316215.

Lin, Z., Akin, H., Rao, R., et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science 379*, 6637, 1123–1130.

Meng, E.C., Goddard, T.D., Pettersen, E.F., et al. 2023. UCSF ChimeraX: Tools for structure building and analysis. *Protein Science 32*, 11, e4792.

Morris, R., Black, K.A., and Stollar, E.J. 2022. Uncovering protein function: from classification to complexes. *Essays in Biochemistry 66*, 3, 255–285.

Pokharel, S., Pratyush, P., Ismail, H.D., Ma, J., and Kc, D.B. 2023. Integrating Embeddings from Multiple Protein Language Models to Improve Protein O-GlcNAc Site Prediction. *International Journal of Molecular Sciences 24*, 21, 16000.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving Language Understanding by Generative Pre-Training. .

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language Models are Unsupervised Multitask Learners. .

Ruan, H.-B., Han, X., Li, M.-D., et al. 2012. O-GlcNAc Transferase/Host Cell Factor C1 Complex Regulates Gluconeogenesis by Modulating PGC-1α Stability. *Cell metabolism 16*, 2, 226–237.

Seber, P. and Braatz, R.D. 2024. Recurrent neural network-based prediction of O-GlcNAcylation sites in mammalian proteins. *Computers & Chemical Engineering 189*, 108818.

Singh, M., Bacolla, A., Chaudhary, S., et al. 2020. Histone Acetyltransferase MOF Orchestrates Outcomes at the Crossroad of Oncogenesis, DNA Damage Response, Proliferation, and Stem Cell Development. *Molecular and Cellular Biology 40*, 18, e00232-20.

Slawson, C. and Hart, G.W. 2011. O-GlcNAc signalling: implications for cancer cell biology. *Nature Reviews. Cancer 11*, 9, 678–684.

Sledzieski, S., Kshirsagar, M., Baek, M., Dodhia, R., Lavista Ferres, J., and Berger, B. 2024. Democratizing protein language models with parameter-efficient fine-tuning. *Proceedings of the National Academy of Sciences 121*, 26, e2405840121.

Smet-Nocca, C., Broncel, M., Wieruszeski, J.-M., et al. 2011. Identification of O-GlcNAc sites within peptides of the Tau protein and their impact on phosphorylation. *Molecular bioSystems 7*, 5, 1420–1429.

Spoel, S.H. 2018. Orchestrating the proteome with post-translational modifications. *Journal of Experimental Botany 69*, 19, 4499–4503.

Stollar, E.J. and Smith, D.P. 2020. Uncovering protein structure. *Essays in Biochemistry 64*, 4, 649–680.

Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., and the UniProt Consortium. 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics 31*, 6, 926–932.

The UniProt Consortium. 2023. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research 51*, D1, D523–D531.

Vaswani, A., Shazeer, N., Parmar, N., et al. 2017. Attention is All you Need. *Advances in Neural Information Processing Systems*, Curran Associates, Inc.

Weissenow, K. and Rost, B. 2025. Are protein language models the new universal key? *Current Opinion in Structural Biology 91*, 102997.

Yang, X. and Qian, K. 2017. Protein O-GlcNAcylation: emerging mechanisms and functions. *Nature reviews. Molecular cell biology 18*, 7, 452–465.

Yang, Y., Fu, M., Li, M.-D., et al. 2020. O-GlcNAc transferase inhibits visceral fat lipolysis and promotes diet-induced obesity. *Nature Communications 11*, 181.

Yang, Y.-H., Wen, R., Yang, N., Zhang, T.-N., and Liu, C.-F. 2023. Roles of protein post-translational modifications in glucose and lipid metabolism: mechanisms and perspectives. *Molecular Medicine 29*, 1, 93.

Zhang, L., Deng, T., Pan, S., et al. 2024. DeepO-GlcNAc: a web server for prediction of protein O-GlcNAcylation sites using deep learning combined with attention mechanism. *Frontiers in Cell and Developmental Biology 12*.

Zhao, W., Zhou, K., Junyi, L., et al. 2023. *A Survey of Large Language Models*.