*Glycobiology*, 2023, **33**, 5, 411–422
https://doi.org/10.1093/glycob/cwad033
Advance access publication date 17 April 2023
**Original Article**

# LMNglyPred: prediction of human *N*-linked glycosylation sites using embeddings from a pre-trained protein language model

**Subash C. Pakhrin[1,2], Suresh Pokharel[3], Kiyoko F. Aoki-Kinoshita[4], Moriah R. Beck[5], Tarun K. Dam[6], Doina Caragea[7], Dukka B. KC[3,*]**

[1]School of Computing, Wichita State University, 1845 Fairmount St., Wichita, KS 67260, USA, [2]Department of Computer Science and Engineering Technology, University of Houston-Downtown, Houston, TX 77002, USA, [3]Department of Computer Science, College of Computing, Michigan Technological University, Houghton, MI 49931, USA, [4]Glycan and Life Systems Integration Center (GaLSIC), Soka University, Tokyo 192-8577, Japan, [5]Department of Chemistry and Biochemistry, Wichita State University, 1845 Fairmount St., Wichita, KS 67260, USA, [6]Laboratory of Mechanistic Glycobiology, Department of Chemistry, Michigan Technological University, Houghton, MI 49931, USA, [7]Department of Computer Science, Kansas State University, Manhattan, KS 66506, USA

*Corresponding author: Dukka B. KC, Department of Computer Science, College of Computing, Michigan Technological University, Houghton, MI 49931, USA. Tel: +1906-487-1657; Email: dbkc@mtu.edu

**Protein *N*-linked glycosylation is an important post-translational mechanism in *Homo sapiens*, playing essential roles in many vital biological processes. It occurs at the N-X-[S/T] sequon in amino acid sequences, where X can be any amino acid except proline. However, not all N-X-[S/T] sequons are glycosylated; thus, the N-X-[S/T] sequon is a necessary but not sufficient determinant for protein glycosylation. In this regard, computational prediction of *N*-linked glycosylation sites confined to N-X-[S/T] sequons is an important problem that has not been extensively addressed by the existing methods, especially in regard to the creation of negative sets and leveraging the distilled information from protein language models (pLMs). Here, we developed LMNglyPred, a deep learning-based approach, to predict *N*-linked glycosylated sites in human proteins using embeddings from a pre-trained pLM. LMNglyPred produces sensitivity, specificity, Matthews Correlation Coefficient, precision, and accuracy of 76.50, 75.36, 0.49, 60.99, and 75.74 percent, respectively, on a benchmark-independent test set. These results demonstrate that LMNglyPred is a robust computational tool to predict *N*-linked glycosylation sites confined to the N-X-[S/T] sequon.**

*Key words*: deep learning; *N*-linked glycosylation; post-translation modification; prediction; protein language model.

## Introduction

Post-translational modifications (PTMs) are the predominant factors leading to the diversity of the proteome (Olsen and Mann 2013). Protein *N*-linked glycosylation is one of the most common PTMs in humans that play essential roles in many vital biological processes. Abnormal *N*-linked glycosylation is observed in many common disorders like cancer, inflammation, Alzheimer's disease, and diabetes. The characterization of *N*-linked glycosylation can be used in clinical diagnostics as well as development of therapeutics. In *N*-linked glycosylation, the N-glycans (oligosaccharides) are attached to the nitrogen atom of an asparagine (Asn or N) residue of the protein. It only occurs at the conserved motifs N-X-S or N-X-T sequons, where X can be any amino acid except proline (Gavel and von Heijne 1990; Kowarik et al. 2006; Bagdonaite et al. 2022). However, the presence of such a sequon in the peptide does not sufficiently confirm that it is glycosylated because about one-third to half of the sequons are buried deep inside the proteins and are not accessible to glycosylation enzymes (Nita-Lazar et al. 2004; Petrescu et al. 2004; Zielinska et al. 2010; Schulz 2012). In addition, various artifacts like distance to the next glycosylation site, sequences surrounding a potential glycosylation site, etc., can impact whether the sequon is N-glycosylated or not. In that regard, the presence of this sequon is necessary but not sufficient for *N*-linked glycosylation in both prokaryotes and eukaryotes (Gavel and von Heijne 1990; Nita-Lazar et al. 2004; Petrescu et al. 2004; Wacker et al. 2006).

*N*-linked glycosylation is often identified using mass spectrometry (MS; Agarwal et al. 1969; Medzihradszky 2005) and a lot of progress has been made in experimental techniques used for mapping and quantifying PTMs. In that regard, more than 14,000 unique N-glycosites have been identified in humans (Sun et al. 2019). Although experimental approaches are the most reliable ways to identify N-glycosites, they are often time-consuming, labor-intensive, and still quite limited. Thus, a mechanistic characterization of PTMs including *N*-linked glycosylation is lacking for a large portion of the proteome. Therefore, complimentary computational tools using machine learning and deep learning (DL) are playing an increasingly essential role in the characterization of glycosylation sites.

In this regard, several computational approaches have been developed to predict N-glycosylation sites. For example, NetNGlyc (Gupta and Brunak 2001) uses an artificial neural network (ANN) for *N*-linked glycosylation prediction. It has to be noted that NetNGlyc attempted to solve this problem confined to the sequon (N-X-[S/T]). EnsembleGly (Caragea et al. 2007) uses ensemble support vector machines (SVM) to predict *N*-linked glycosylation sites. It utilizes a Position-

**Received:** September 6, 2022. **Revised:** April 10, 2023. **Accepted:** April 11, 2023

Specific Scoring Matrix (PSSM) that is generated by PSI-Blast (Altschul et al. 1997), physicochemical properties, and a One Hot encoding scheme to encode the dataset. GlycoPP (Chauhan et al. 2012) uses an SVM to predict *N*-linked glycosylation based on amino acid composition (AAC), One Hot encoding, PSSM, Secondary Structures (SS), and accessible surface area (ASA) features. NGlycPred (Chuang et al. 2012) utilizes the sequence, pattern, and structural-based features; and uses a random forest classifier to predict *N*-linked glycosylation sites. The GlycoEP "in silico" (Chauhan et al. 2013) tool uses SVM to predict *N*-linked glycosylation sites. Moreover, GlycoEP datasets were encoded with sequence, evolutionary and structural features. GlycoMine (Li et al. 2015) uses a Random Forest (RF) classifier for *N*-linked glycosylation site prediction, and it is based on heterogeneous functional and sequence-based features. GlycoMine$^{\text{struct}}$ (Li et al. 2016) combines sequence and structural features for predicting *N*-linked glycosylation sites using a random forest. Akmal et al. (Akmal et al. 2017) encoded the dataset with position relative and statistical moments and used artificial neural networks to predict *N*-linked glycosylation sites. GlycoMine_PU (Li et al. 2019) uses a positive unlabeled (PU) learning technique to predict *N*-linked glycosylation sites. N-GlyDE (Pitti et al. 2019) is an SVM-based method that uses sequence and predicted structural features to predict the N-glycosylation sites. Nglyc (Pugalenthi et al. 2020) uses RF with sequence and structural information of amino acids to predict N-glycosylation sites in eukaryotic protein sequences. N-GlycoGo (Chien et al. 2020) uses XGBoost, an ensemble machine learning model, for *N*-linked glycosylation site prediction. PUStackNGly (Alkuhlani et al. 2022) uses a stacking ensemble learning model to detect *N*-linked glycosylation. In the PUStackNGly approach, the Logistic Regression, SVM, ANN, RF were used as base model and prediction from those base model was used to train the SVM meta-classifier model.

Owing to the fact that various DL-based approaches have been proposed in the field of bioinformatics (Quang and Xie 2016; Pakhrin and Pant 2018; Lee et al. 2020; Lv et al. 2021; Pakhrin et al. 2021b; Dhakal et al. 2022; Høie et al. 2022; Pakhrin 2022; Pakhrin et al. 2022; Yang et al. 2022a, 2022b), we have recently seen development of some DL-based approaches for *N*-linked glycosylation sites as well. SPRINT-Gly (Taherzadeh et al. 2019) uses fully connected artificial neural networks to identify *N*-linked glycosylation sites based on sequence, evolutionary, and structural-based features. DeepNGlyPred (Pakhrin et al. 2021a) uses a multi-layer perceptron (MLP) to predict the *N*-linked glycosylation site confined to N-X-[S/T] by encoding a peptide window using sequence-based features (e.g. Gapped-Dipeptide), predicted structural features (e.g. Secondary Structures, Accessible Surface Area, relative solvent accessibility (RSA), torsion angle (Φ, Ψ), and disordered regions.

Although it was known for some time that the presence of the consensus N-X-[S/T] motif does not always lead to glycosylation (Gavel and von Heijne 1990), it has to be noted here that except for NetNGlyc, N-GlyDE, and DeepNGlyPred all these other approaches are evaluated on the asparagine residue without being confined to the N-X-[S/T] sequon. More specifically, the approaches confined to N-X-[S/T] sequon define the glycosylation site prediction problem as a classification problem to classify whether the given sequon is likely to be glycosylated or not. In that regard, the existing approaches can be grouped as approaches confined to the N-X-[S/T] sequon and approaches not confined to the N-X-[S/T] sequon and that currently only a handful of approaches are confined to the N-X-[S/T] sequon. Perhaps this is the reason why the predictive performances of approaches (especially the ones not confined to the N-X-[S/T] sequon) for predicting *N*-linked glycosylation sites tend to be overestimated as the task here is to merely predict each N in the N-X-[S/T] sequon as a glycosylation site. In this regard, NetNGlyc, N-GlyDE, and DeepNGlyPred are important contributions in the field that exploit the fact that this sequon is a necessary but not sufficient condition for *N*-linked glycosylation.

On the other hand, transformer-based language models that are learned from a large corpus of unlabeled data have recently achieved amazing results in the field of natural language processing (NLP) (Vaswani et al. 2017). Due to the availability of a large number of protein sequences in the UniProt knowledgebase and other resources, we now have seen various protein language models (pLMs) being developed (Elnaggar et al. 2021; Rives et al. 2021; Brandes et al. 2022). Considering protein sequences as sentences, Elnaggar et al. developed a pre-trained pLM called ProtT5-XL-UniRef50 (herein called ProtT5) (Elnaggar et al. 2021) based on 2.5 billion protein sequences. The representations of these models have been utilized for various downstream tasks (Littmann et al. 2021; Marquet et al. 2022; Heinzinger et al. 2022; Nallapareddy et al. 2023; Weissenow et al. 2022), and the results demonstrate that the distributed representation learned from the distillation of these language models have useful information that captures the evolutionary context of a sequence, contact map, taxonomy, protein structure, physicochemical properties, and function. Similarly, features from these transformer-based pLMs have been successfully utilized to predict signal peptides (Teufel et al. 2022), lysine glycation sites (Liu et al. 2022), subcellular localization (Thumuluri et al. 2022), protein structural features (Høie et al. 2022), lysine crotonylation sites (Qiao et al. 2021), succinylation site prediction (Pokharel et al. 2022), S-nitrosylation site prediction (Pratyush et al. 2023), and binding residues (Littmann et al. 2021), among others.

Admittedly, some progress has been made in the development of *N*-linked glycosylation site prediction methods confined to N-X-[S/T] sequon. However, for most of these predictors, the input features are still hand-crafted features. As these methods use hand-crafted features, they are heavily biased toward those selected features and they do not exploit the latent representations from the unknown, yet indispensable features. Additionally, to the best of our knowledge, the benefits of the recent advances in large pLMs and the distributed representation learned from the distillation of these language models have not been exploited for *N*-linked glycosylation site prediction. Moreover, besides DeepNGlyPred, the existing approaches for *N*-linked glycosylations site prediction (confined to the sequon) only use traditional machine learning approaches. Hence, in this work, we aim to develop an improved approach to predict *N*-linked glycosylation site prediction confined to N-X-[S-T] sequon by leveraging the vast amount of distilled information learned by these large pLMs combined with advances in DL approaches. Please refer to Table 1 for a list of approaches for glycosylation site prediction along with their ML/DL architecture, the features, and whether the approaches are confined to N-X-[S/T] sequon or not.

**Table 1.** Summary of approaches for *N*-linked glycosylation site prediction approaches.

| Name | ML/DL architecture | Features | Confined to N-X-(S/T)? | Year published |
|---|---|---|---|---|
| NetNGlyc (Gupta and Brunak 2001) | ANN | Cellular role descriptor and subcellular location | Yes | 2001 |
| EnsembleGly (Caragea et al. 2007) | Ensemble support vector machines (SVM) | Position-Specific Scoring Matrix (PSSM), physicochemical properties, and One Hot encoding | No | 2007 |
| GlycoPP (Chauhan et al. 2012) | SVM | Amino acid composition (AAC), One Hot encoding, PSSM, secondary structures (SS), and accessible surface area (ASA) | No | 2012 |
| NGlycPred (Chuang et al. 2012) | Random Forest | sequence, pattern, and structural-based features | No | 2012 |
| GlycoEP (Chauhan et al. 2013) | SVM | Sequence, evolutionary and structural features | No | 2013 |
| GlycoMine (Li et al. 2015) | Random Forest | Structural, functional, and sequence-based features | No | 2015 |
| N-GlyDE (Pitti et al. 2019) | SVM | Gapped dipeptide features<br>Pattern-based surface accessibility (SA) and secondary structure (SS) features | Yes | 2019 |
| SPRINT-Gly (Taherzadeh et al. 2019) | ANN | relative ASA, Secondary Structures, Half-Sphere Exposure, Intrinsically disordered region, Physicochemical properties, Evolutionary Information, Amino acid sequence | No | 2019 |
| Nglyc (Pugalenthi et al. 2020) | Random Forest | Secondary structures, amino acid frequencies, solvent accessibility | Yes | 2020 |
| N-GlycoGo (Chien et al. 2020) | XGBoost | Sequence-, structure-, functional-based features | Yes | 2020 |
| DeepNGlyPred (Pakhrin et al. 2021a) | ANN | PSSM, E (beta-strand), H (helix), and C (coil), predicted accessible surface area (ASA), relative solvent accessibility (RSA), predicted disordered region, gapped dipeptide (GD) | Yes | 2021 |
| PUStackNGly (Alkuhlani et al. 2022) | Stacking Ensemble Learning (ML) | Sequence-based features, profile-based features, structure-based features | No | 2022 |

Although some progress has been made, the current computational approaches are not at the point where they can be used for high-throughput characterization of *N*-linked glycosylation sites. Improving the existing methods for prediction of *N*-linked glycosylation site could help address some of the limitations of MS-based techniques as well as provide a complementary approach to experimental methods. By improving the prediction of *N*-linked glycosylation sites, the community can better characterize *N*-linked glycosylation sites which could also lead to the discovery of novel regulatory mechanisms. Additionally, computational approaches can help guide experimental design by reducing the cost and time required for *N*-linked glycosylation analysis. Moreover, robust computational approaches that have built in feature importance and explainability can provide insights into the mechanisms underlying *N*-linked glycosylation and the factors that influence the site which could then be used to design new therapeutic agents as well as develop tools to detect glycosylated proteins in diagnosis and prognosis of various diseases.

Hence, we propose a novel computational approach called LMNglyPred (**L**anguage **M**odel based ***N***-linked **gly**cosylation site **Pred**ictor) that utilizes embedding from a pLM (i.e. ProtT5) to improve the predictive performance of *N*-linked glycosylation sites. By considering proteins as sentences, we fed the full protein sequence into the pre-trained ProtT5 model to extract fixed-length, high-dimensional per-residue features from the last encoder layer. Subsequently, the high-dimensional contextualized embeddings (i.e. 1,024 feature vector) of the site in interrogation (asparagine, N) are fed into the Deep Neural Network [DNN, essentially a multi-layer perceptron (MLP)]-based classifier for *N*-linked glycosylation site prediction.

Using cross-validation experiments, we found that the final classifier based on the MLP architecture is better than the other architectures compared. To demonstrate its effectiveness, we evaluated the performance of the proposed method LMNglyPred using N-GlyDE's dataset against other existing approaches. Our experiment showed that LMNglyPred achieved better performance in predicting protein *N*-linked glycosylation sites compared to the state-of-the-art predictors, yielding an MCC of 0.717 on N-GlyDE's independent test set. LMNglyPred is a freely available, fast, and reliable approach for prediction of *N*-linked glycosylation sites. All programs and data are available at https://github.com/KCLabMTU/LMNglyPred.

## Results

LMNglyPred utilizes embeddings extracted per residue (1,024 features) for the site of interest (N, asparagine) from ProtT5 using a full-length sequence as input. We use two datasets for training LMNglyPred: N-GlycositeAtlas and N-GlyDE.

**Table 2.** Results of the 10-fold cross-validation on the N-GlycositeAtlas training dataset using different deep and machine learning models. The highest values in each column are highlighted in bold.

| DL and ML model | MCC ± 1 SD | SN ± 1 SD | SP ± 1 SD | ACC ± 1 SD | PRE ± 1 SD |
|---|---|---|---|---|---|
| Random Forest | 0.412 ± 0.018 | 0.707 ± 0.011 | 0.704 ± 0.009 | 0.706 ± 0.009 | 0.705 ± 0.009 |
| LR | 0.495 ± 0.015 | 0.803 ± 0.012 | 0.688 ± 0.005 | 0.746 ± 0.007 | 0.720 ± 0.005 |
| MLP/ANN | **0.524 ± 0.014** | 0.809 ± 0.031 | **0.711 ± 0.031** | **0.760 ± 0.008** | **0.760 ± 0.016** |
| SVM | 0.499 ± 0.014 | 0.736 ± 0.015 | 0.762 ± 0.007 | 0.749 ± 0.007 | 0.743 ± 0.011 |
| XGBoost | 0.426 ± 0.022 | 0.739 ± 0.016 | 0.686 ± 0.017 | 0.713 ± 0.011 | 0.702 ± 0.011 |
| 1D CNN | 0.496 ± 0.023 | **0.810 ± 0.024** | 0.681 ± 0.028 | 0.745 ± 0.012 | 0.718 ± 0.015 |

Protein and peptide redundancies are removed from within and across training and independent test datasets. We performed 10-fold cross-validation on the training dataset(s) to obtain the best hyperparameters for our DL architecture. Finally, we used the hyperparameters obtained from 10-fold cross-validation and trained the model using the overall training set and assessed the trained model on the independent test set and compared the performance against other existing approaches.

## Performance on the N-GlycositeAtlas dataset

### Ten-fold cross-validation on the N-GlycositeAtlas training set

To tune the hyperparameters (parameters whose values are used to control the learning process) (Yang and Shami 2020), and to investigate the performance of various DL/ML models for the training dataset, we performed 10-fold cross-validation on the N-GlycositeAtlas training dataset whose negative sites are sequons from proteins in the endoplasmic reticulum and surface accessible sequons from Golgi apparatus, extracellular, and cell membrane of human glycoproteins. The predictive performance of different DL and ML models using the stratified 10-fold cross-validation on the N-GlycositeAtlas training data set is shown in Table 6. The contextualized embedding of the glycosylated or non-glycosylated token "N" produced by the pre-trained ProtT5 model when fed to MLP achieves the best performance as seen in Table 2 (although the values are relatively close). Intriguingly, the same architecture (MLP) produced the highest result using 10-fold cross-validation on the N-GlyDE training data set as well. This MLP model produced MCC, SN, SP, ACC, and PRE values of 0.524 ± 0.014, 0.809 ± 0.031, 0.711 ± 0.031, 0.760 ± 0.008, and 0.760 ± 0.016, respectively, for the stratified 10-fold cross-validation. As the MLP model produced the best result on 10-fold cross-validation, we selected this architecture as our final model and call it LMNglyPred. The independent test set result and 10-fold cross-validation results produced by LMNglyPred are similar, which demonstrates that this model can be used for *N*-linked glycosylation prediction purposes.

Moreover, to check whether the MLP and SVM are significantly different from a statistical point of view we further performed McNemar's hypothesis test (McNemar 1947; Dietterich 1998). The test comments on whether the two models disagree in the same way (or not). In this test, the default assumption (H0) implies that the two binary classification algorithms disagree to the same amount. Although, when H0 is rejected, it suggests that two binary classifiers disagree in different ways. While performing the test we found *P*-value = 0.179, which is greater than the 0.05 threshold; hence, we accept H0 and there is no difference in the disagreement. Hence, the performance of MLP and SVM are not statistically different.

### Testing on the 10 percent independent test set separated from the N-GlycositeAtlas training set

Finally, to assess the performance of our approach on an independent test set, we trained the model on the overall N-GlycositeAtlas training set and applied it to predict *N*-linked glycosylation sites of proteins in the 10 percent independent test set separated from the overall N-GlycositeAtlas dataset. The total number of samples in each set for N-GlycositeAtlas dataset is shown in Table 6. We achieved MCC, SN, SP, ACC, and PRE values of 0.4959, 76.50, 75.36, 75.74, and 60.99 percent, respectively. Furthermore, MLP was able to classify 1,242 samples as True Negative, 635 samples as True Positive, 406 as False Positive, and 195 as False Negative. Additionally, we balanced the independent test set and the results produced by MLP are shown in Supplementary Table S2 for informative purposes.

## Performance on N-GlyDE dataset

In order to compare our approach against existing approaches, we also trained and tested our approach using N-GlyDE's dataset. It has to be noted here that N-GlyDE's dataset is also confined to N-X-[S/T] sequon.

### 10-fold cross-validation on the N-GlyDE training set

We trained our approach (LMNglyPred) on the N-GlyDE training set. We performed cross-validation on the N-GlyDE training set using 11 different DL/ML models (1D CNN-LSTM, 1D CNN-BiLSTM, 1D CNN, BiLSTM, LSTM, Random Forest, Logistic Regression, Multi-layer Perceptron, Support Vector Machine, XGBoost, Naïve Bayes).

Consequently, all these models were trained on the N-GlyDE training dataset (2,722 training examples) to choose the best-performing DL or machine learning model for *N*-linked glycosylation PTM prediction. The details of these models and their hyperparameters are described in the Methods section. The performance of these models was compared using various metrics including ACC, MCC, precision, sensitivity, and specificity. The detailed results of the 10-fold cross-validation for these models are presented in Table 3, where it can be observed that the Multi-layer Perceptron produces the highest MCC of 0.6576 on 10-fold cross-validation on the N-GlyDE training dataset. The mean MCC, mean accuracy, mean sensitivity, mean specificity, and mean precision obtained from the 10-fold cross-validation for this MLP architecture was 0.657 ± 0.035, 0.958 ± 0.0192, 0.851 ± 0.014, 0.635 ± 0.053, and 0.842 ± 0.018, respectively. It can also

**Table 3.** Results of the 10-fold cross-validation on the N-GlyDE training dataset using different deep and machine learning models. The highest value in each column is highlighted in bold.

| DL and ML model | MCC ± 1 S.D. | ACC ± 1 S.D. | SP ± 1 S.D. | SN ± 1 S.D. | PRE ± 1 S.D. |
|---|---|---|---|---|---|
| 1D CNN-LSTM | 0.622 ± 0.062 | 0.962 ± 0.014 | 0.585 ± 0.063 | 0.837 ± 0.024 | 0.824 ± 0.024 |
| 1D CNN-BiLSTM | 0.632 ± 0.041 | 0.964 ± 0.006 | 0.591 ± 0.055 | 0.841 ± 0.017 | 0.827 ± 0.020 |
| BiLSTM | 0.630 ± 0.045 | 0.959 ± 0.014 | 0.600 ± 0.051 | 0.840 ± 0.018 | 0.829 ± 0.018 |
| LSTM | 0.592 ± 0.055 | 0.953 ± 0.013 | 0.567 ± 0.062 | 0.825 ± 0.022 | 0.817 ± 0.022 |
| LR | 0.650 ± 0.047 | 0.967 ± 0.013 | 0.608 ± 0.031 | 0.848 ± 0.018 | 0.833 ± 0.013 |
| MLP/ANN | **0.657 ± 0.035** | 0.958 ± 0.019 | 0.635 ± 0.053 | **0.851 ± 0.014** | 0.842 ± 0.018 |
| SVM | 0.648 ± 0.053 | **0.974 ± 0.014** | 0.589 ± 0.054 | 0.846 ± 0.021 | 0.827 ± 0.020 |
| XGBoost | 0.603 ± 0.055 | 0.945 ± 0.020 | 0.597 ± 0.047 | 0.829 ± 0.022 | 0.823 ± 0.024 |
| Naïve Bayes | 0.613 ± 0.058 | 0.927 ± 0.013 | 0.647 ± 0.064 | 0.834 ± 0.024 | 0.842 ± 0.026 |
| 1D CNN | 0.647 ± 0.039 | 0.902 ± 0.033 | **0.729 ± 0.028** | 0.845 ± 0.019 | **0.870 ± 0.011** |
| Random Forest | 0.648 ± 0.055 | 0.959 ± 0.012 | 0.622 ± 0.068 | 0.848 ± 0.022 | 0.837 ± 0.025 |

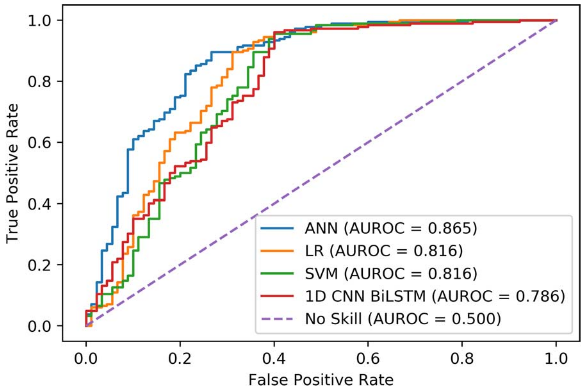

**Fig. 1.** Receiver operating characteristics (ROC) curves of 1D CNN LSTM, ANN, LR, and SVM on the 10 percent test set separated from N-GlyDE training dataset. For each model, the area under the ROC curve is reported.

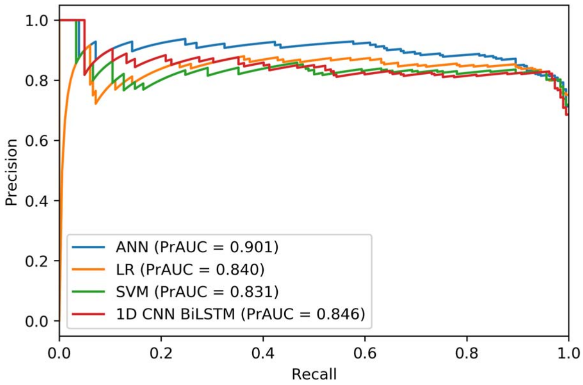

**Fig. 2.** Precision-recall curves of 1D CNN LSTM, ANN, LR, and SVM on the 10 percent test set separated from N-GlyDE training dataset. For each model, the area under the PRAUC curve is reported.

be observed that among the compared approaches the DL methods outperform the machine learning-based model for *N*-linked glycosylation site prediction. Among these approaches, the MLP deep-learning architecture, essentially deep neural network (DNN), has the highest MCC; thus, this model was chosen and trained on the overall training set to predict *N*-linked glycosylation in the N-GlyDE-independent test dataset.

## Training on the N-GlyDE* training set

Owing to the fact that the negative sites in original N-GlyDE dataset have some sequons from mitochondria and nucleus that may not be the “true” negative sites, we separated the original N-GlyDE training set into 90 percent for training and used the remaining 10 percent for independent test set. This is shown in Table 7 as N-GlyDE∗. The performance of various DL and ML architectures while trained on 80 percent (training), 10 percent (validation) and tested on a 10 percent test set separated from the N-GlyDE training dataset is shown in Supplementary Table S1 and the ROC curve is shown in Fig. 1. Furthermore, it also must be noted that none of the 10 percent test set test data is present in the 80 (training) and 10 percent (validation) dataset. It can be observed from Fig. 1 that LMNglyPred, which is based on an MLP approach, has the highest area under the ROC. Figure 2 shows that LMNglyPred has the highest precision-recall area under the curve compared to other DL and machine learning models. So, LMNglyPred is a robust model for the prediction of *N*-linked glycosylation.

## Testing on the 10 percent N-GlyDE independent test set separated from training set

To assess the performance of the MLP model on an independent test set, we trained the model on the 90 percent N-GlyDE training set (also known as N-GlyDE∗) and applied it to predict *N*-linked glycosylation sites on the 10 percent independent test set separated from the N-GlyDE training set. The MLP model produced MCC, accuracy, precision, sensitivity, and specificity values of 0.685, 86.42, 86.75, 94.05 , and 71.00 percent, respectively. Furthermore, while observing the confusion matrix of the MLP classifier it was able to classify 71 as True Negatives and 190 as True Positives. However, it falsely classified 29 as False Positive and 12 as False Negative. The independent test results (10 percent separated from training set) for various machine learning and DL models are presented in Table 4. The results are shown for informative purpose only and the final model (MLP) was selected based on 10-fold cross-validation.

## Comparison of LMNglyPred with other *N*-linked glycosylation site predictors

To assess the performance of LMNglyPred against other approaches, we trained our model on the N-GlyDE training set and applied it to predict *N*-linked glycosylation sites on

**Table 4.** Performance metrices of various machine learning and DL models on the 10 percent independent test dataset separated from the N-GlyDE training dataset. The highest value in each column is highlighted in bold.

| DL and ML model | MCC | SN | SP | ACC | PRE |
|---|---|---|---|---|---|
| 1D CNN-BiLSTM | 0.629 | 0.980 | 0.55 | 0.837 | 0.814 |
| 1D CNN-LSTM | 0.640 | 0.965 | 0.6 | 0.844 | 0.829 |
| BiLSTM | 0.551 | 0.965 | 0.49 | 0.807 | 0.792 |
| LSTM | 0.538 | 0.935 | 0.54 | 0.804 | 0.804 |
| RF | 0.623 | 0.960 | 0.59 | 0.837 | 0.825 |
| LR | 0.648 | **0.965** | 0.61 | 0.847 | 0.857 |
| ANN | **0.685** | 0.940 | **0.71** | **0.864** | **0.867** |
| SVM | 0.629 | 0.980 | 0.55 | 0.837 | 0.814 |
| XGBoost | 0.637 | 0.935 | 0.66 | 0.844 | 0.847 |
| Naïve Bayes | 0.662 | 0.930 | 0.7 | 0.854 | 0.862 |
| 1D CNN | 0.661 | 0.940 | 0.68 | 0.854 | 0.855 |

**Table 5.** Prediction performance of LMNglyPred compared to other available *N*-linked glycosylation site predictors on the independent test set. The highest value in each column is highlighted in bold.

| Predictors | Accuracy | Precision | Sensitivity | Specificity | MCC |
|---|---|---|---|---|---|
| LMNglyPred (Trained on N-GlyDE dataset) | 86.93 | **88.57** | 74.69 | **94.24** | 0.717 |
| N-GlyDE | 74.0 | 61.3 | 82.6 | 68.9 | 0.499 |
| GlycoMine | 72.5 | 61.6 | 70.0 | 73.9 | 0.430 |
| NetNGlyc | 57.2 | 46.0 | 84.4 | 41.1 | 0.265 |
| GlycoEP_Std_PPP | 57.4 | 43.7 | 51.2 | 61.0 | 0.119 |

the N-GlyDE independent test set. The MLP model produced MCC, accuracy, precision, sensitivity, and specificity values of 0.717, 86.93, 88.57, 74.69, and 94.24 percent respectively. These results are better than the performance of the N-GlyDE approach which uses a two-stage prediction approach with a similarity voting algorithm and SVM method based on sequence and structural features. Additionally, while observing the confusion matrix of the MLP classifier, it was able to classify 262 as True Negatives and 124 as True Positives. However, it falsely classified 16 as a False Positive and 42 as a False Negative. This result is likely because the N-GlyDE's 77.14 percent independent test set negative sites are from nucleus, cytosol, and mitochondrion subcellular localization of the proteins. Moreover, ProtT5 contextualized embedding can learn subcellular localization information in its distributed representation.

We further compared the performance of LMNglyPred with other established *N*-linked glycosylation site predictors like N-GlyDE, GlycoMine, NetNGlyc, and GlycoEP_Std_PPP, and the results are shown in Table 5. It should be noted here that we trained LMNglyPred on the N-GlyDE and tested on N-GlyDE test set as the other methods are not trained on N-GlycositeAtlas dataset. Here, we trained LMNglyPred on the N-GlyDE datasets, using the contextual embedding produced by ProtT5 of the glycosylated or non-glycosylated token "N" and then tested on the N-GlyDE independent test set and the results are shown in Table 4. The results for N-GlyDE, GlycoMine, NetNGlyc, and GlycoEP_Std_PPP were adopted from N-GlyDE. It can be observed from Table 5 that LMNglyPred trained on N-GlyDE produced an MCC of 0.717, compared with an MCC of 0.499 for N-GlyDE. From these results, it can be deduced that LMNglyPred performs better than the other compared methods including N-GlyDE. It should be noted that LMNglyPred was trained and tested with exactly same dataset that was used by N-GlyDE and other approaches.

## Visualization using t-SNE plot

Additionally, we investigated the classification efficacy of the features and the learned model visually. Herein, features represent 1024 numeric vectors of glycosylated or non-glycosylated "N" residues extracted from ProtT5 and the learned/trained model refers to the MLP network trained with the N-GlycositeAtlas training dataset. To discern the classification effectiveness of these features as well as the feature vector produced by the penultimate hidden layer of the trained MLP network, we used t-SNE (Maaten and Hinton 2008) to project the features into a two-dimensional space (Fig. 3). For the features extracted from ProtT5 on the glycosylated or non-glycosylated token "N" of N-GlycositeAtlas training set, the positive and negative samples are relatively clustered together (Fig. 3A). Figure 3(B) represents the t-SNE plot of the feature vectors generated from the penultimate hidden layer of the MLP deep-learning architecture when N-GlycositeAtlas training set is used. This shows that negative samples (blue points) are concentrated at the bottom left (third quadrant), whereas positive samples (orange points) are concentrated at the left (second quadrant) which indicates that the pre-trained per residue pLM feature extraction with MLP learns *N*-linked glycosylation patterns and largely clusters positive and negative samples in $\mathbb{R}^2$ space. Hence, this result elaborates that contextualized features produced from pretrained ProtT5 when passed to MLP DL network can cluster positive and negative samples of *N*-linked glycosylation sites in two-dimensional space.

## Discussion

One of the key innovations in LMNglyPred is the incorporation of pLM based features to represent protein sequences. pLM based features have proven to be quite useful in various bioinformatics tasks (Bepler and Berger 2019, 2021;

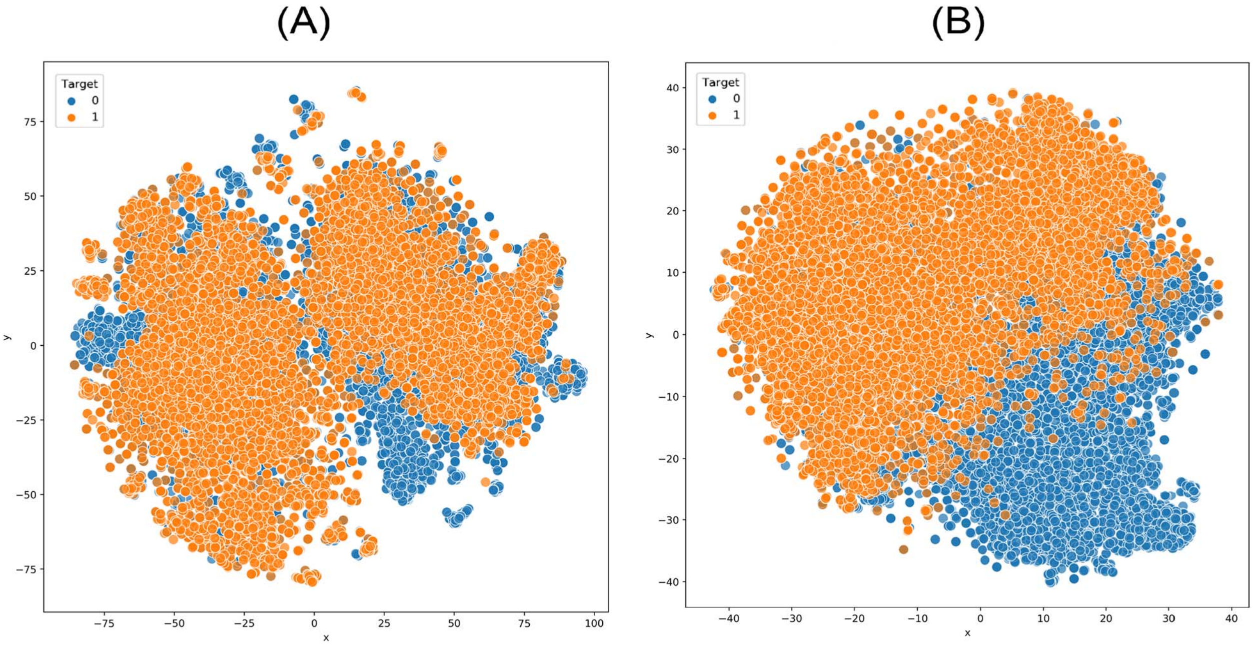

**Fig. 3.** *T*-SNE illustration of the embeddings extracted from ProtT5, and the features transformed by MLP, 1 represents positive *N*-linked glycosylation sites, 0 represents negative (non-) *N*-linked glycosylation sites. (a) Embeddings extracted from ProtT5 (N-GlycositeAtlas training set). (b) Features transformed by MLP (N-GlycositeAtlas training set).

Rao et al. 2019; Zhou et al. 2020; Rao et al. 2021; Zhang et al. 2021; Ferruz and Höcker 2022; Unsal et al. 2022). We had one major goal in the project: to move away from hand-crafted feature extraction for prediction of *N*-linked Glycosylation sites. To achieve this goal, we investigated whether language models learned from a large amount of protein sequences are able to capture the features predictive of glycosylation sites. Additionally, we also wanted to interrogate what type of machine learning approach would work well on these pre-trained feature representations. Additionally, our other contribution is the creation of negative data set where we took the negative sites from secretory pathway (specifically, endoplasmic reticulum and surface accessible sequons from Golgi bodies, cell membrane and extra-cellular compartments). In order to achieve the goal, we used contextualized embeddings learned from a pLM called ProtT5 to extract features for the site of interest. Subsequently, various ML and DL algorithms were evaluated using 10-fold cross validation, and the final model was selected based on the 10-fold cross validation results. The MLP model that we call LMNglyPred achieves the best prediction performance among the compared methods which is likely made possible by using the distilled knowledge from large sets of protein sequences by the pre-trained ProtT5 model that is used to encode the protein sequences.

LMNglyPred neither relies on knowledge of protein structure, nor in the expert-crafted sequence features, nor on time-consuming evolutionary information derived from multiple sequence alignments (MSAs). Instead, the input to the MLP model is a contextual representation of the glycosylated or non-glycosylated token "N" from the pre-trained pLM (ProtT5). This state-of-the-art prediction of *N*-linked glycosylation is likely due to the contextual embeddings of all the amino acids in the protein sequence that are produced by the transformer-based model which makes use of position embedding with a self-attention mechanism. As our results show that our LMNglyPred model outperforms the widely available *N*-linked glycosylation predictors, the LMNglyPred *N*-linked glycosylation predictor is likely to be a very useful tool for the glycobiology community to predict *N*-linked glycosylated sites in proteins. One interesting result portrayed in the t-SNE plot (Fig. 3B) is that our model was largely able to cluster the two classes of glycosylated and non-glycosylated asparagine residues in two-dimensional space.

LMNGlyPred is a new approach that uses information distilled from large pLMs using a DL framework and the results are better than existing approaches. However, there are some limitations associated with LMNGlyPred. One of the limitations of our approach is high FPR. The other limitation of our approach is the likelihood of some "noise" in the N-GlycoSiteAtlas dataset that could be present due to the possibility of spontaneous deamidation (Palmisano et al. 2012). Therefore, LMNGlyPred requires better validated datasets to be useful for high throughput prediction of *N*-linked glycosylation sites. With generation of more quality N-glycosylation datasets and the advances in natural language processing including large pLMs, it is expected that some of these limitations will be addressed in the future.

## Materials and methods

### Predicting protein *N*-linked glycosylation sites

Data preparation is as follows: with the aim to train a DL algorithm to predict *N*-linked glycosylation sites in proteins, we utilized two datasets: N-GlycositeAtlas and N-GlyDE datasets.

### N-GlycositeAtlas-based dataset

The benchmark dataset considered in this work is based on N-GlycositeAtlas (Sun et al. 2019), which is a recently developed large-scale repository for *N*-linked glycosylation that contains 7,204 human glycoproteins. It must be noted here that all the N-glycosylation sites of N-GlyDE are included in the N-GlycositeAtlas database. Initially, all the 7,204 (19 were

**Table 6.** Number of positive and negative glycosylation sites for training and testing proposed in this work.

| Name of dataset | Positive site | Negative site | Total |
|---|---|---|---|
| Training Dataset (+ve from N-GlycoSiteAtlas, −ve from ER + RSA gt 0.5(Golgi + cell membrane + extracellular) | 8,405 | 15,860 | 24,265 |
| Independent Test Set | 830 | 1,648 | 2,478 |

**Table 7.** Number of positive and negative glycosylation sites for training and testing in N-GlyDE.

| Name of dataset | Positive site | Negative site |
|---|---|---|
| N-GlyDE training[a] | 2,023 | 1,001 |
| N-GlyDE test | 167 | 280 |
| N-GlyDE* (90% from original N-GlyDE) training | 1821 | 901 |
| N-GlyDE* test (10% Independent Test Set separated from Training set) | 202 | 100 |

[a] Unable to extract few proteins (originally 2050 and 1030 +ves and −ves)

**Table 8.** Hyperparameters used in the MLP network for the N-GlyDE and N-GlycositeAtlas datasets.

| Name of the parameters | Parameters used |
|---|---|
| Input ProtT5 feature vector length | 1,024 |
| Activation Function | ReLU |
| No. neuron in Dense layers | 64 |
| No. of neuron in the output Dense layer | 2 |
| Activation Function at output layer | softmax |
| Optimizer | Adam |
| Learning rate | 0.001 |
| Objective/loss function | Binary_crossentropy |
| Model Checkpoint | Monitor = “val_accuracy” |
| Reduce learning rate on plateau | Factor = 0.001 |
| Early stopping | patience = 5 |
| Dropout | 0.3 |
| Batch_size | 256 |
| Decision Boundary | 0.5 |
| Epochs | 400 |

obsolete) glycoproteins are extracted from the UniProt and the corresponding 9,235 *N*-linked glycosylation sites from these sequences are extracted and form the positive training data. All these 9,235 glycosylation sites are confined to N-X-[S/T] sequons.

For the creation of negative sites, as it is known that proteins from the endoplasmic reticulum, Golgi apparatus, extracellular, cell membrane undergo *N*-linked glycosylation; hence, 7,875 human glycoproteins are extracted from the DeepLoc-2.0 (Thumuluri et al. 2022) database. Then, N-X-[S/T] sequons are extracted from these proteins and redundant sequences were deleted; moreover, peptides that had more than 30 percent sequence identity are filtered out using CD-HIT (Li and Godzik 2006) which resulted into 17,508 N-X-[S/T] sequons. For model training purpose, under-sampling (Lemaitre et al. 2017) strategy is utilized to balance the dataset by randomly selecting negative sites to match the number of positive sites. The RSA (Tien et al. 2013) of a residue in a protein measures the extent of burial or exposure of that residue in the 3D structure. Hence, we used NetSurfP-2.0 (Klausen et al. 2019) tool to make sure that Golgi apparatus, extracellular, and cell membrane negative sites have RSA greater than 0.5. It also must be noted that none of the positive or negative N-glycosylation sites nor the protein sequences from the N-GlycositeAtlas independent test set are present in the N-GlycositeAtlas training dataset.

Moreover, we made sure that no protein sequence appears in both the training and test sets as ProtT5 (Elnaggar et al. 2021) can learn representation for other sites from the same protein as well which could lead to overestimation of the performance.

## N-GlyDE dataset

This dataset is adapted from N-GlyDE (Pitti et al. 2019), which consists of 2,023 (1,821 + 202) experimentally verified N-glycosylation sites (confined to N-X-[S/T] sequons) from 832 glycoproteins extracted from human proteins in UniProt (ver. 201608). These experimentally verified sites form the positive training data. Additionally, 1,001 (901 + 100) sites that follow the same N-X-[S/T] patterns from the same positive protein sequences but are not experimentally verified as *N*-linked glycosylation sites are considered as negative sites. This makes the N-GlyDE training dataset. Further information about N-GlyDE (Independent Test set) can be found at seminal approach N-GlyDE (Pitti et al. 2019). Table 6 summarizes the number of sites included in each dataset.

## Feature extraction—embeddings from pLM

There are various numerical representations that can be used to encode protein sequences. The recent development in the field is the advent of embeddings (distributed vector representations) that are representations of protein sequences extracted from the last hidden layers of the networks forming the pLM trained on a large set of unlabeled protein sequences. These embeddings have been shown to capture a diversity of higher-level features of proteins and have been used successfully in predicting secondary structure and other tasks (Heinzinger et al. 2022). For the current task, we used embeddings from the pLM ProtT5-XL-Uniref (Elnaggar et al. 2021) (herein called ProtT5). The pLM ProtT5 was trained on unlabeled protein sequences from BFD (Big Fantastic Database; 2.5 billion sequences including meta-genomic sequences) (Steinegger et al. 2019) and UniRef50 (UniProt 2021). ProtT5 has been built in analogy to the NLP (Natural Language Processing) T5 (Raffel et al. 2020) ultimately learning some of the constraints of protein sequences. Features learned by the pLM can be transferred to any (prediction) task requiring numerical protein representations by extracting vector representations for single residues from the hidden states of the pLM (transfer learning). As ProtT5 was only trained on unlabeled protein sequences, there is no risk of information leakage or overfitting to a certain label during pre-training. Essentially, ProtT5 (Elnaggar et al. 2021) outputs fixed-length (1024) vector representations for each residue in a protein sequence. In essence, to predict whether a sequon N-X-[T/S] is glycosylated or not, we extracted a 1024-dimensional vector for each glycosylated or non-glycosylated asparagine (N) residue in the sequon where only the encoder side of ProtT5 was used, and embeddings were extracted from the last hidden layer of the models. We only extracted embeddings from the encoder-side of ProtT5.

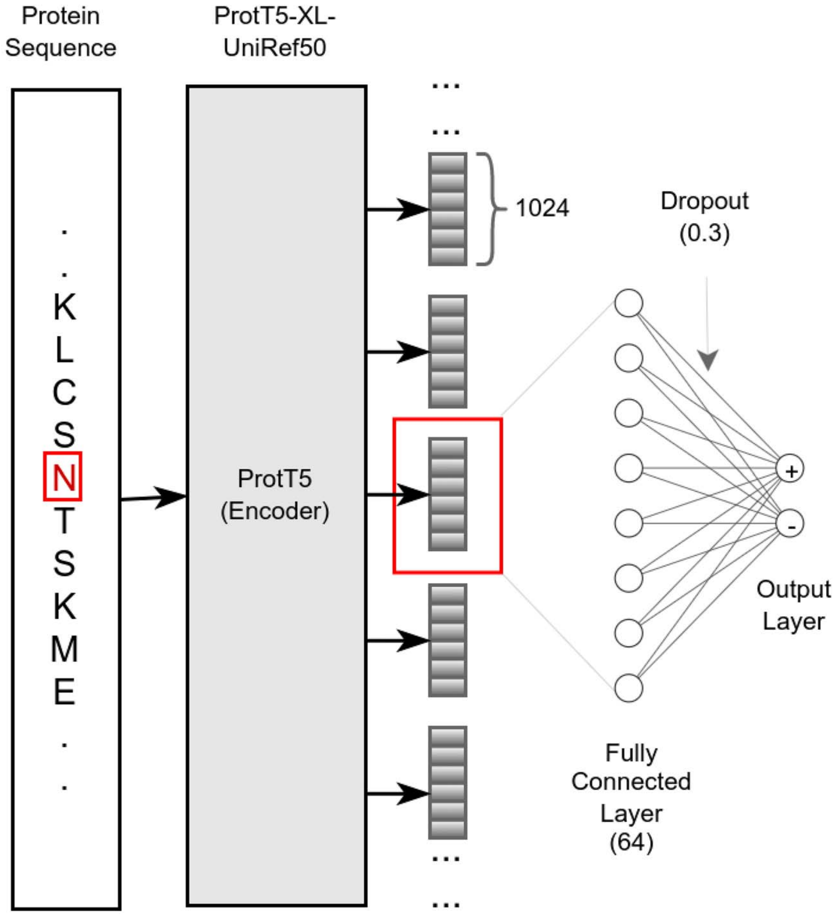

**Fig. 4.** Overall framework of LMNglyPred.

## Model training

As discussed above, *N*-glycosylation occurs on N (Asparagine) residues, and so we extract contextualized embeddings from the ProtT5 model using the full-length protein sequence as input. Finally, the corresponding feature for the site of interrogation (in this case N) is extracted (1,024-dimensional vector) and passed to the subsequent ML/DL model. Using this feature and dataset (both N-GlyDE and N-GlycositeAtlas), we train several ML/DL models to correctly predict *N*-linked glycosylation sites.

The performance of several architectures was evaluated: 1D CNN-LSTM, 1D CNN-BiLSTM, BiLSTM, LSTM, LR, ANN, SVM, XGBoost, Naïve Bayes, 1D CNN, RF, etc. We describe the ANN/MLP/DNN architecture in Fig. 4. As shown in Fig. 4, the features are extracted for the site of interrogation (N, highlighted in red) using full protein sequence as input and the 1024 real-valued feature vectors are fed into a MLP deep-learning architecture consisting of a 64-neuron input layer followed by a 2-neuron output layer. To explore the hyperparameter space, we performed a 10-fold cross-validation grid search on the MLP DL model with the N-GlyDE and N-GlycositeAtlas training dataset. It was done against 1, 2, 3, and 4 dense hidden layers, sigmoid and ReLU activation function, 32, 64, 128, 256, 512, and 1,024 neurons in each layer, RMSprop, and Adam optimizers, and 0.2, 0.3, 0.4, and 0.5 dropout rate, whereas the default learning rate of 0.001 was used. A similar approach was performed for different DL and machine learning algorithms. The optimized hyperparameters using grid search are shown in Table 8.

Based upon grid search, 64-neuron input layers were configured with ReLU activation function. As dropout layer/nodes in the network help alleviate overfitting and improve the generalization, we set the dropout equal to 0.3. Since our task is to train a binary classification model to distinguish *N*-linked glycosylated and non-*N*-linked glycosylated sites. Therefore, in the output dense layer, we set the number of neurons equal to 2. The optimized hyperparameters for the deep-learning architecture are elaborated in Table 8. To avoid overfitting, we have used overfitting reduction techniques like dropout, early stopping, Model Checkpoint, and Reduce learning rate on the plateau. Furthermore, no signs of overfitting and underfitting are present in our trained model as can be seen from Supplementary Figs S1 and S2. The loss curve for the training and validation are following each other as well as the training and validation accuracy curves are also following each other.

The MLP architecture was implemented with TensorFlow 2.3.1 (Abadi et al. 2016) and sklearn 1.0.2. The optimized hyperparameters of MLP model for the *N*-linked glycosylation prediction are elaborated in Table 8.

## Model evaluation and performance metrics

In this study, 10-fold cross-validation was used to evaluate the performance of the model and to determine its robustness and generalizability. During tenfold cross-validation, the data are partitioned into ten equal parts. Then, one part is left out for validation, whereas training is performed on the remaining nine parts. This process is repeated until all parts are used for validation. For the results of 10-fold cross-validation, unless otherwise noted, all performance metrics are reported as the mean value ± standard deviation.

To evaluate the performance of each model, we use accuracy (ACC), sensitivity (SN), specificity (SP), Matthews Correlation Coefficient (MCC), and precision. ACC describes the correctly predicted residues out of the total residues (Equation 1). Meanwhile, SN defines the model's ability to distinguish positive residues (Equation 2), and SP measures the model's ability to correctly identify the negative residues (Equation 3). MCC is the calculated score that considers the model's predictive capability concerning both positive and negative residues (Equation 4). Likewise, precision reveals how many of the correctly predicted cases turned out to be positive (Equation 5):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{1}$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \times 100 \tag{3}$$

$$MCC = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \times 100 \tag{5}$$

## Abbreviations

pLM, protein language model; DL, deep learning; SN, sensitivity; SP, specificity; MCC, Matthews Correlation Coefficient; ACC, accuracy; N, Asparagine; S, Serine; T, Threonine; PTM, post-translational modifications; ANN, artificial neural network; MLP, multi-layer perceptron; SVM, support vector machine; PSSM, Position-Specific Scoring Matrix; AAC, amino acid composition; SS, secondary structures; ASA, accessible surface area; PU, positive unlabeled; RF, Random Forest; DNN, deep neural network; RSA, relative solvent accessibility; MSA, multiple sequence alignment; NLP, natural language processing; 1D CNN, one-dimensional convolutional neural network; LSTM, long short-term memory; AUROC, area under the receiver operating characteristic curve; T5, Text-to-Text Transfer Transformer; Bi-LSTM, bidirectional long short-term memory; ML, machine learning; PrAUC, precision-recall area under the curve; SOTA, state-of-the-art; UniProt, Universal Protein Resource; t-SNE, t-distributed stochastic neighbor embedding

## Acknowledgments

We acknowledge the use of the BeoShock High-Performance Computing resources located at Wichita State University and the N-GlycositeAtlas dataset made freely available for the researchers. We would also like to acknowledge help of Dr. Michael Heinzinger from TUM and Dr. Dennis Livesay for his help and helpful discussion.

## Author contributions

D.B.K., D.C., K.F.A.-K., and T.K.D. conceived of and designed the experiments; S.C.P. performed all the experiments and data analysis, S.P. verified all the programs and models. S.C.P., D.B.K., T.K.D., M.R.B., S.P., D.C., and K.F.A.-K. revised the manuscript. All authors have read and agreed to the published version of the manuscript. D.B.K. oversaw the whole project.

## Supplementary data

Supplementary material is available at *GLYCOB Journal* online.

## Funding

This work was supported by National Science Foundation grants (No. 1608537 to TKD, Nos 2210356 and 1901793 to DKC).

*Conflict of Interest statement.* None declared.

## Data availability statement

All programs and data are available at https://github.com/KCLabMTU/LMNglyPred.

## References

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M. 2016. Tensorflow: a system for large-scale machine learning. *OSDI'16: Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*. 265–283.

Agarwal KL, Kenner GW, Sheppard RC. Feline gastrin. An example of peptide sequence analysis by mass spectrometry. *J Am Chem Soc*. 1969:**91**(11):3096–3097.

Akmal A, Aizaz M, Rasool N, Khan YD. Prediction of N-linked glycosylation sites using position relative features and statistical moments. *PloS one*. 2017:**12**(8):e0181966.

Alkuhlani A, Gad W, Roushdy M, ABM S. PUStackNGly: positive-Unlabeled and StackingLearning for *N*-linked GlycosylationSite prediction. *IEEE Access*. 2022:**10**:12702–12713.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997:**25**(17): 3389–3402.

Bagdonaite I, Malaker SA, Polasky DA, Riley NM, Schjoldager K, Vakhrushev SY, Halim A, Aoki-Kinoshita KF, Nesvizhskii AI, Bertozzi CR, et al. Glycoproteomics. *Nat Rev Methods Primers*. 2022:**2**(1):48.

Bepler T, Berger B. 2019. Learning protein sequence embeddings using information from structure. *International Conference on Learning Representations* 48.

Bepler T, Berger B. Learning the protein language: evolution, structure, and function. *Cell Systems*. 2021:**12**(6):654–669.

Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*. 2022:**38**(8):2102–2110.

Caragea C, Sinapov J, Silvescu A, Dobbs D, Honavar V. Glycosylation site prediction using ensembles of support vector machine classifiers. *BMC Bioinform*. 2007:**8**(1):1–13.

Chauhan JS, Bhat AH, Raghava GP, Rao A. GlycoPP: a webserver for prediction of N- and O-glycosites in prokaryotic protein sequences. *PLoS One*. 2012:**7**(7):e40155.

Chauhan JS, Rao A, Raghava GPS. In silico platform forprediction of N-, O- and C-glycosites in eukaryotic protein sequences. *PLoS One*. 2013:**8**(6):e67008, e67008.

Chien C-HC, Chi-Chang LS-H, Chen C-W, Chang Z-HC, Yen-Wei. N-GlycoGo: predicting protein N-glycosylation sites on imbalanced data sets by using heterogeneous and comprehensive strategy. *IEEE Access*. 2020:**8**:165944–165950.

Chuang G-Y, Boyington JCJ, Gordon Zhu M, Jiang NGJ, Kwong PD, Georgiev I. Computational prediction of N-linkedglycosylation incorporating structural properties and patterns. *Bioinformatics*. 2012:**28**(17):2249–2255.

Dhakal A, McKay C, Tanner JJ, Cheng J. Artificial intelligence in the prediction of protein-ligand interactions: recent advances and future directions. *Brief Bioinform*. 2022:**23**(1):bbab476.

Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput*. 1998:**10**(7): 1895–1923.

Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, et al. ProtTrans: towards cracking the language of Lifes code through self-supervised deep learning and high performance computing. *IEEE Trans Pattern Anal Mach Intell*. 2021:**44**:7112–7127.

Ferruz N, Höcker B. Controllable protein design with language models. *Nat Mach Intell*. 2022:**4**(6):521–532.

Gavel Y, von Heijne G. Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. *Protein Eng Des Sel*. 1990:**3**:433–442.

Gupta R, Brunak S. Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac Symp Biocomput*. 2001:**7**:310–322.

Heinzinger M, Littmann M, Sillitoe I, Bordin N, Orengo C, Rost B. Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genom Bioinform*. 2022:**4**(2):lqac043.

Høie MH, Kiehl EN, Petersen B, Nielsen M, Winther O, Nielsen H, Hallgren J, Marcatili P. NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic Acids Res*. 2022:**50**(W1):W510–W515.

Klausen MS, Jespersen MC, Nielsen H, Jensen KK, Jurtz VI, Sonderby CK, Sommer MOA, Winther O, Nielsen M, Petersen B, et al. NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning. *Proteins*. 2019:**87**(6): 520–527.

Kowarik M, Young NM, Numao S, Schulz BL, Hug I, Callewaert N, Mills DC, Watson DC, Hernandez M, Kelly JF, et al. Definition of the bacterial N-glycosylation site consensus sequence. *EMBO J*. 2006:**25**(9):1957–1966.

Lee J, Yoon W, Kim S, Kim D, Kim S, Ho So C, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020:**36**(4):1234–1240.

Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res*. 2017:**18**:559–563.

Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006:**22**(13):1658–1659.

Li F, Li C, Wang M, Webb GI, Zhang Y, Whisstock JC, Song J. GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome. *Bioinformatics*. 2015:**31**(9):1411–1419.

Li F, Li C, Revote J, Zhang Y, Webb GI, Li J, Song J, Lithgow T. GlycoMine(struct): a new bioinformatics tool for highly accurate mapping of the human *N*-linked and O-linked glycoproteomes by incorporating structural features. *Sci Rep*. 2016:**6**:34595.

Li F, Zhang Y, Purcell AWW, Chou GI, Lithgow K-C, Trevor LC, Song J. Positive-unlabelled learning of glycosylation sites in thehuman proteome. *BMC Bioinform*. 2019:**20**(1):112.

Littmann M, Heinzinger M, Dallago C. Protein embeddings and deep learning predict binding residues for various ligand classes. *Sci Rep*. 2021:**11**(1):23916.

Liu Y, Liu Y, Wang G-A, Cheng Y, Bi S, Zhu X. BERT-Kgly: a bidirectional encoder representations from transformers (BERT)-based model for predicting lysine glycation site for Homo sapiens. *Front Bioinform*. 2022:**2**:834153.

Lv H, Dao FY, Zulfiqar H, Lin H. DeepIPs: comprehensive assessment and computational identification of phosphorylation sites of SARS-CoV-2 infection using a deep learning-based approach. *Brief Bioinform*. 2021:**22**(6).

Maaten, HintonVan der Maaten, Laurens, Geoffrey Hinton. Visualizing data using t-SNE. *Mach Learn Res*. 2008:**9**(11):2579–2605.

Marquet C, Heinzinger M, Olenyi T, Dallago C, Erckert K, Bernhofer M, Nechaev D, Rost B. Embeddings from protein language models predict conservation and variant effects. *Hum Genet*. 2022:**141**(10):1629–1647.

McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*. 1947:**12**(2): 153–157.

Medzihradszky KF. Peptide sequence analysis. *Methods Enzymol*. 2005:**402**:209–244.

Nallapareddy V, Bordin N, Sillitoe I, Heinzinger M, Littmann M, Waman V, Sen N, Rost B, Orengo C. CATHe: detection of remote homologues for CATH superfamilies using embeddings from protein language models. *bioRxiv*. 2023:**39**(1):btad029. https://doi.org/10.1093/bioinformatics/btad029.

Nita-Lazar M, Wacker M, Schegg B, Amber S, Aebi M. The N-X-S/T consensus sequence is required but not sufficient for bacterial *N*-linked protein glycosylation. *Glycobiology*. 2004:**15**:361–367.

Olsen JV, Mann M. Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol Cell Proteomics*. 2013:**12**(12):3444–3452.

Pakhrin, Subas h C. *Deep learning-based approaches for prediction of post-translational modification sites in proteins*. PhD diss. Wichita State University, 2022.

Pakhrin SC, Pant DR. 2018. Multi-armed bandit learning approach with entropy measures for effective heterogeneous networks handover scheme. *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. Greater Noida, India: IEEE. p. 451–455.

Pakhrin SC, Aoki-Kinoshita KF, Caragea D, Kc DB. DeepNGlyPred: a deep neural network-based approach for human *N*-linked glycosylation site prediction. *Molecules*. 2021a:**26**(23):7314.

Pakhrin SC, Shrestha B, Adhikari B, Kc DB. Deep learning-based advances in protein structure prediction. *Int J Mol Sci*. 2021b:**22**(11):5553.

Pakhrin SC, Pokharel S, Saigo H, Kc DB. Deep learning-based advances in protein posttranslational modification site and protein cleavage prediction. *Methods Mol Biol*. 2022:**2499**:285–322.

Palmisano G, Melo-Braga MN, Engholm-Keller K, Parker BL, Larsen MR. Chemical deamidation: a common pitfall in large-scale *N*-linked glycoproteomic mass spectrometry-based analyses. *J Proteome Res*. 2012:**11**(3):1949–1957.

Petrescu AJ, Milac AL, Petrescu SM, Dwek RA, Wormald MR. Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding. *Glycobiology*. 2004:**14**(2):103–114.

Pitti T, Chen CT, Lin HN, Choong WK, Hsu WL, Sung TY. N-GlyDE: a two-stage *N*-linked glycosylation site prediction incorporating gapped dipeptides and pattern-based encoding. *Sci Rep*. 2019:**9**(1):15975.

Pokharel S, Pratyush P, Heinzinger M, Newman RH, Kc DB. Improving protein succinylation sites prediction using embeddings from protein language model. *Sci Rep*. 2022:**12**(1):16933.

Pratyush P, Pokharel S, Saigo H, Kc DB. pLMSNOSite: an ensemble-based approach for predicting protein S-nitrosylation sites by integrating supervised word embedding and embedding from pre-trained protein language model. *BMC Bioinform*. 2023:**24**(1):41.

Pugalenthi G, Nithya V, Chou KC, Archunan G. Nglyc: a random Forest method for prediction of N-glycosylation sites in eukaryotic protein sequence. *Protein Pept Lett*. 2020:**27**(3):178–186.

Qiao Y, Zhu X, Gong H. BERT-Kcr: prediction of lysine crotonylation sites by a transfer learning method with pre-trained BERT models. *Bioinformatics*. 2021:**38**(3):648–654.

Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res*. 2016:**44**(11):e107.

Raffel C, Shazeer N, Roberts A, Lee K, Narang SM, Zhou M, Yanqi LW, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res*. 2020:**21**:1–67.

Rao RB, Nicholas TN, Duan Y, Chen X, Canny J, Abbeel P, Song YS. Evaluating protein transfer learning with TAPE. *Adv Neural Inf Process Syst*. 2019:**32**:9689–9701.

Rao R, Meier J, TOS S, Rives A. Transformer protein language models are unsupervised structure learners. *Biorxiv* 2020:2020–12.

Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA*. 2021:**118**(15):e2016239118.

Schulz BL. Beyond the sequon: sites of N-glycosylation. In: Petrescu S, editors. *Glycosylation*. Rijeka: InTech; 2012. pp. 21–40.

Steinegger M, Mirdita M, Soding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat Methods*. 2019:**16**(7):603–606.

Sun S, Hu Y, Ao M, Shah P, Chen J, Yang W, Jia X, Tian Y, Thomas S, Zhang H. N-GlycositeAtlas: a database resource for mass spectrometry-based human *N*-linked glycoprotein and glycosylation site mapping. *Clin Proteom*. 2019:**16**(35):35.

Taherzadeh G, Dehzangi A, Golchin M, Zhou Y, Campbell MP. SPRINT-Gly: predicting N- and O-linked glycosylation sites of human and mouse proteins by using sequence and predicted structural properties. *Bioinformatics*. 2019:**35**:4140–4146.

Teufel F, Almagro Armenteros JJ, Johansen AR, Gislason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G, Nielsen H. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat Biotechnol*. 2022:**40**(7):1023–1025.

Thumuluri V, Almagro Armenteros JJ, Johansen AR, Nielsen H, Winther O. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Res*. 2022:**50**(W1):W228–W234.

Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. Maximum allowed solvent accessibilites of residues in proteins. *PLoS One*. 2013:**8**(11):e80635.

UniProt C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021:**49**(D1):D480–D489.

Unsal S, Atas H, Albayrak M, Turhan K, Acar AC, Doğan T. Learning functional properties of proteins with language models. *Nat Mach Intell*. 2022:**4**(3):227–245.

Vaswani A, Shazeer N, Parmar N, Uszkoreit N, Jones J, Gomez AN, Kaiser L, Polosukhin I. 2017. Attention is all you need. *Advances in neural information processing systems NIPS* 2017:**30**.

Wacker M, Feldman MF, Callewaert N, Kowarik M, Clarke BR, Pohl NL, Hernandez M, Vines ED, Valvano MA, Whitfield C, et al. Substrate specificity of bacterial oligosaccharyltransferase suggests a common transfer mechanism for the bacterial and eukaryotic systems. *Proc Natl Acad Sci USA*. 2006:**103**(18): 7088–7093.

Weissenow K, Heinzinger M, Rost B. Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. *Structure*. 2022:**30**(8):1169–1177 e4.

Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: theory and practice. *Neurocomputing*. 2020:**415**: 295–316.

Yang L, Wang S, Altman RB. POPDx: an automated framework for patient phenotyping across 392 246 individuals in the UK biobank study. *J Am Med Inform Assoc*. 2022a:**30**(2):245–255.

Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, Compas C, Martin C, Costa AB, Flores MG, et al. A large language model for electronic health records. *NPJ Digit Med*. 2022b:**5**(1): 194.

Zhang Y, Lin J, Zhao L, Zeng X, Liu X. A novel antibacterial peptide recognition algorithm based on BERT. *Brief Bioinform*. 2021:**22**(6):bbab200.

Zhou G, Chen M, Ju CJT, Wang Z, Jiang JY, Wang W. Mutation effect estimation on protein-protein interactions using deep contextualized representation learning. *NAR Genom Bioinform*. 2020:**2**(2):lqaa015.

Zielinska DF, Gnad F, Wisniewski JR, Mann M. Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell*. 2010:**141**(5):897–907.