



WICHITA STATE
UNIVERSITY

UNIVERSITY LIBRARIES

SumoPred-PLM: human SUMOylation and SUMO2/3 sites Prediction using Pre-trained Protein Language Model

Item Type	Article
Authors	Palacios, Andrew Vargas; Acharya, Pujan; Peidl, Anthony Stephen; Beck, Moriah R.; Blanco, Eduardo; Mishra, Avdesh; Bawa-Khalfe, Tasneem; Pakhrin, Subash C.
Citation	Palacios, A.V., Acharya, P., Peidl, A.S., Beck, M.R., Blanco, E., Mishra, A., Bawa-Khalfe, T., & Pakhrin, S.C. (2024). "SumoPred-PLM: human SUMOylation and SUMO2/3 sites Prediction using Pre-trained Protein Language Model." <i>NAR Genomics and Bioinformatics</i> , 6(1). https://doi.org/10.1093/nargab/lqae011
Publisher	Oxford University Press
Download date	2026-05-08 14:16:16
Link to Item	https://doi.org/10.1093/nargab/lqae011

SumoPred-PLM: human SUMOylation and SUMO2/3 sites Prediction using Pre-trained Protein Language Model

Andrew Vargas Palacios^{1,†}, Pujan Acharya^{1,†}, Anthony Stephen Peidl², Moriah Rene Beck³, Eduardo Blanco⁴, Avdesh Mishra⁵, Tasneem Bawa-Khalfe² and Subash Chandra Pakhrin^{1,*}

¹Department of Computer Science and Engineering Technology, University of Houston-Downtown, 1 Main St., Houston, TX 77002, USA

²Department of Biology and Biochemistry, Center for Nuclear Receptors & Cell Signaling, University of Houston, Houston, TX 77204, USA

³Department of Chemistry and Biochemistry, Wichita State University, 1845 Fairmount St., Wichita, KS 67260, USA

⁴Department of Computer Science, University of Arizona, 1040 4th St., Tucson, AZ 85721, USA

⁵Department of Electrical Engineering and Computer Science, Texas A&M University-Kingsville, Kingsville, TX 78363, USA

*To whom correspondence should be addressed. Tel: +1 713 221 5819; Fax: +1 713 223 7407; Email: pakhrins@uhd.edu

†The first two authors and last author contributed equally.

Abstract

SUMOylation is an essential post-translational modification system with the ability to regulate nearly all aspects of cellular physiology. Three major paralogues SUMO1, SUMO2 and SUMO3 form a covalent bond between the small ubiquitin-like modifier with lysine residues at consensus sites in protein substrates. Biochemical studies continue to identify unique biological functions for protein targets conjugated to SUMO1 versus the highly homologous SUMO2 and SUMO3 paralogues. Yet, the field has failed to harness contemporary AI approaches including pre-trained protein language models to fully expand and/or recognize the SUMOylated proteome. Herein, we present a novel, deep learning-based approach called SumoPred-PLM for human SUMOylation prediction with sensitivity, specificity, Matthew's correlation coefficient, and accuracy of 74.64%, 73.36%, 0.48% and 74.00%, respectively, on the CPLM 4.0 independent test dataset. In addition, this novel platform uses contextualized embeddings obtained from a pre-trained protein language model, ProtT5-XL-UniRef50 to identify SUMO2/3-specific conjugation sites. The results demonstrate that SumoPred-PLM is a powerful and unique computational tool to predict SUMOylation sites in proteins and accelerate discovery.

Introduction

Post-translational modifications (PTMs) are the predominant factors leading to the diversity of the proteome (1,2). Protein SUMOylation is one of the most common PTMs in humans that performs essential roles in many vital biological processes like transcription control, chromatin organization, accumulation of macromolecules in cells, regulation of gene expression, and signal transduction (3,4). SUMOylation is also necessary for the conservation of genome integrity (5). Consequently, it is not surprising that a change in SUMOylation dynamics favors the onset of a variety of human diseases including cancer, Alzheimer's disease, Parkinson's disease, viral infections, heart diseases, and diabetes (5–9).

SUMOylation occurs as a modifier in an ϵ -amino group of lysine residues in the target protein through a multi-enzymatic cascade (10). In this reaction, SUMO is connected to a lysine residue in substrate protein by covalent linkage via three enzymes, namely activating (E1), conjugating (E2) and ligase (E3). Also, it can be separated from the target protein by a specific SUMO protease enzyme (11). This covalent SUMO conjugation frequently occurs at a consensus motif ψ -K-X-E where ψ represents lysine, isoleucine, valine, or phenylalanine, K is lysine, X can be any amino acid, and E is glutamic acid (12). However, additional SUMO protein substrates have recently been identified that lack this canonical SUMO consensus motif, making it more challenging to identify SUMOylated protein targets.

SUMOylation requires the Small Ubiquitin-Related Modifier (SUMO) protein, which has structural similarity to ubiquitin and has been discovered in a wide range of eukaryotic organisms (6,13,14). Five SUMO paralogues exist in humans with SUMO1, SUMO2 and SUMO3 expressed ubiquitously in multiple tissue/cell types and consistently the most studied. SUMO2 and SUMO3 are highly homologous with 97% amino acid sequence overlap and frequently referred to as SUMO2/3. Unlike SUMO1, SUMO2/3 includes an internal SUMO-consensus site with lysine 11, which allows for poly-SUMOylation to occur (15). The internal SUMOylation site allows SUMO2/3 to form poly-SUMO chains on target proteins. The SUMO2/3 poly-chain serves as a binding platform for proteins with SUMO-interaction motifs (SIMs) and thereby supports dynamic non-covalent protein complexes. As SUMO2/3 directs both covalent and non-covalent protein interactions, previous biochemical studies suggest a unique protein substrate profile and biological function for SUMO2/3 versus SUMO1. SUMO2/3 specific protein conjugates direct protein degradation, chromatin remodeling, gene expression, and DNA repair (9,16,17). Also, only the SUMO2 knockout is embryonic lethal and is essential for organismal development (18,19). Yet identification of SUMO paralogue specific targets is still in its infancy and is frequently only addressed at an individual protein level.

To date, SUMOylation is most often identified using mass spectrometry and a lot of progress has been made in

Received: June 21, 2023. Revised: November 17, 2023. Editorial Decision: January 12, 2024. Accepted: January 17, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

experimental techniques used for mapping and quantifying PTMs. In that regard, more than 53 000 unique SUMOylation sites have been identified in human proteins (20–22). Although experimental approaches are the most reliable ways to identify SUMOylation sites, they are often time-consuming, labour-intensive, and are still quite limited. Thus, a mechanistic characterization of PTMs including SUMOylation sites is lacking for a large portion of the proteome. Therefore, complementary computational tools using machine learning (ML) and deep learning (DL) are playing an increasingly essential role in the characterization of SUMOylation sites.

Several different SUMOylating site prediction models currently exist (4,23–35). The most utilized program remains GPS-SUMO with the ability to predict SUMO-accepting lysine residues in consensus and non-consensus SUMO sites (36). However, like most prediction models, the input features are still hand-crafted features. Additionally, to the best of our knowledge the benefits of the recent advances in large protein language models (PLMs) and the distributed representation learned from the distillation of these language models have not been explored for SUMOylation site prediction. A recent study evaluated the performance of different models for protein representation, revealing that ProtT5 achieved the best performance in most of the tasks (37). However, ProtT5 has not yet been used for SUMOylation and SUMO2/3 PTM prediction.

Recently, transformer-based language models trained with a large corpus of unlabelled data have achieved stunning results in the field of natural language processing (NLP) (38). Due to the availability of large number of protein sequences in the UniProt knowledge base and other resources, we now have a wide variety of PLMs under development (39–46). Considering protein sequences as sentences, Elnaggar *et al.* developed a pre-trained PLM called ProtT5-XL-UniRef50 (herein called ProtT5) based on 2.5 billion protein sequences (46). The representations of these models have been utilized for various downstream tasks, and the results demonstrate that the distributed representation learned from the distillation of these language models provide useful information that captures the evolutionary context of a sequence, contact map, taxonomy, long-range dependencies, protein structure, physicochemical properties, subcellular localization, and function (47–53). Moreover, long-range dependencies can yield essential insights into the broader context and functional implications of SUMOylation PTM. These dependencies offer indispensable insights into the intricate connections between distant amino acids within a protein, shedding light on how modifications at one site influence the protein's behavior and its interactions with other molecules. Taking these distant relationships into account can enhance the accuracy of algorithms used to predict SUMOylation PTM sites. Additionally, it can also provide valuable information about the protein's three-dimensional structure. For example, it can help pinpoint regions of a protein that, although far apart in the primary sequence, exhibit close interactions within the folded protein structure, which is pertinent for predicting SUMOylation PTM sites. Similarly, features from these transformer-based PLMs have been successfully utilized to predict signal peptides (54), lysine glycation sites (55), N and O-linked glycosylation sites (51,56), phosphorylation sites (57), lysine crotonylation sites (58), subcellular localization (59), protein structural features (60), intrinsic disorder sites (61), and binding residues (52) among others.

Hence, we propose a novel computational approach called SumoPred-PLM (SUMOylation site Prediction using Protein Language Model) that utilizes embeddings from a protein language model (i.e. ProtT5) to improve the predictive performance of SUMOylation sites. By considering proteins as sentences, we feed the full protein sequence into the pre-trained ProtT5 model to extract fixed-length high-dimensional per residue representations from the last encoder layer. Subsequently, the high-dimensional contextualized embeddings (i.e. a vector with 1024 features) of the site in interrogation (Lysine, K) are fed into a Deep Neural Network (DNN), essentially a Multi-layer Perceptron (MLP)-based classifier for SUMOylation and SUMO2/3 site prediction.

Using cross-validation experiments, we found that the classifier based on the MLP architecture performed better compared to other architectures employed. To demonstrate its effectiveness, we evaluated the performance of the proposed method SumoPred-PLM using the GPS-SUMO dataset against quintessential approaches like GPS-SUMO (36). Our experiments showed that SumoPred-PLM achieved better performance in predicting protein SUMOylation sites compared to the state-of-the-art GPS-SUMO predictor, yielding an area under the receiver operating characteristic curve (AUROC) of 0.895. SumoPred-PLM is a freely available, fast and reliable approach for prediction of SUMOylation sites. All programs and data are available at <https://github.com/PakhrinLab/SumoPred-PLM>.

Materials and methods

Predicting protein SUMOylation and SUMO2/3 sites

This section describes the dataset, features extraction method, performance evaluation metrics, feature selection, and the methods for training the model. With the aim to train a DL algorithm to predict SUMOylation and SUMO2/3 sites in proteins, we utilized three different datasets: CPLM 4.0 (22), SUMO2/3 (62) and GPS-SUMO (36), which are described below.

CPLM 4.0 dataset

In this study we utilized the Compendium of Protein Lysine Modifications (CPLM 4.0) dataset that was developed by Zhang *et al.* (22). This dataset consists of 29 different types of lysine PTMs including SUMOylation as a part of the CPLM 4.0 database. To avoid overestimation of the prediction accuracy as well as to maintain diversity, the redundant sites were removed using CD-HIT Suite with a threshold of 30% sequence identity (63). If two or more proteins were found to be modified at the same position and if they have >30% sequence similarity, only one of the proteins was preserved. As a result of this filtering, we obtained 5,695 unique and diverse SUMOylated proteins. Moreover, we separated 5117 SUMOylated proteins for the training. From these training proteins we extracted 26 911 positive samples (SUMOylated sites) and 201 949 negative samples (non-SUMOylate sites). Additionally, we extracted 2988 independent positive and 2988 independent negative SUMOylated sites distributed across 578 proteins for testing. SUMOylation may occur on any lysine (K) amino acids of the SUMOylated protein sequence; however, not all of these sites are SUMOylation sites.

Table 1. Positive and negative SUMOylation sites for training and independent testing derived from CPLM 4.0 dataset

Dataset	Number of Proteins	Positive	Negative
Training	5117	26 911	201 949
Test	578	2988	2988

We considered the experimentally verified SUMOylated sites acquired from the CPLM 4.0 database as positive SUMOylated sites. All other lysine sites within the same substrate are considered as negative SUMOylation sites. The difference between the small number of positive and large number of negative samples makes this benchmark dataset unbalanced. This imbalance can bias the performance of any predictor towards the identification of negative samples (a high true negative rate) over the detection of positive samples (a low true positive rate).

The two commonly used strategies to overcome the imbalance problem are random over-sampling and under-sampling. The idea behind over-sampling is to duplicate the positive samples to increase them to the number of negative samples. While in under-sampling, some of the negative samples are discarded to make the number of negative samples equal to the number of positive samples. The over-sampling procedure could increase the probability of over-fitting the model due to duplication of positive samples while under-sampling often provides a modest solution for a given model. Therefore, we selected an under-sampling procedure to overcome the imbalance problem (64). As a result of under-sampling, we ended up with 26,911 positive and an equal number of negative samples. In this way, we avoid bias in our benchmark towards negative samples and increase our chance to detect more positive samples, or in other words, more SUMOylation sites accurately. Table 1 shows the number of positive and negative sites from CPLM 4.0 Dataset after 30% CD-HIT. Moreover, we explored the dbPTM human SUMOylation dataset (65). [Supplementary Table S1](#) presents the statistics of the training and independent test datasets when 30% psi-cd-hit was applied.

SUMO2/3 dataset

Another dataset we utilized in this study is the human endogenous SUMO2/3 SUMOylation dataset developed by Hendriks *et al.* (62). This dataset consists of 14 869 endogenous SUMO2/3 Sites. We used CD-HIT Suite with a threshold of 30% to remove sequence identity among SUMO2/3 proteins (63). As a result of this filtering, we obtained 3225 SUMOylated proteins. Moreover, we separated 2902 SUMOylated proteins for training. From these training proteins we extracted 10 684 positive sites (SUMO2/3 sites) and 10 684 randomly under-sampled negative sites (non-SUMO2/3 sites) from 131 459 negative sites. Additionally, we extracted 1269 independent positive and 1269 independent negative SUMO2/3 sites distributed across 322 proteins for testing. We considered the experimentally verified SUMO2/3 sites acquired from Hendriks *et al.* database as positive SUMO2/3 sites and all the other lysine sites within the same substrate as negative SUMO2/3 sites. Table 2 shows the number of positive and negative sites from Hendriks *et al.* SUMO2/3 dataset after 30% CD-HIT.

Table 2. Positive and negative SUMO2/3 sites for training and independent testing derived from Hendriks *et al.* dataset

Dataset	Number of proteins	Positive	Negative
Training	2902	10 684	131 459
Test	322	1269	1269

Table 3. Positive and negative SUMOylation sites for training and independent testing derived from the GPS-SUMO dataset

Dataset	Number of Proteins	Positive	Negative
Training	509	891	23 371
Test	39	71	1377

GPS-SUMO dataset

The GPS-SUMO dataset consists of 548 proteins and among them 509 proteins were used for training, and 39 were utilized for independent testing. Eight hundred ninety one experimentally verified human SUMOylation sites were extracted from 509 SUMOylation proteins. All the other lysine's from the same 509 SUMOylated proteins were considered as negative SUMOylation sites. To create a balanced dataset, 891 negative sites were randomly under sampled from 23 371 sites. These experimentally verified sites form the training data. Moreover, 71 experimentally verified independent positive SUMOylated test sites were extracted from 39 different SUMOylation protein which are different than the training proteins. Next, 1377 independent negative lysine sites which do not include the independent SUMOylated positive sites were extracted from the same 39 independent SUMOylation proteins. These experimentally verified sites form the testing dataset. Further information about GPS-SUMO can be found in the seminal approach section of GPS-SUMO (36). Table 3 summarizes the number of sites included in GPS-SUMO dataset.

Feature extraction—embeddings from protein language model

A range of numerical representation schemes can be used to encode protein sequences. A recent development in the field is the advent of embeddings (distributed vector representations), which 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 latent embeddings capture a diversity of higher-level features of proteins and have been used successfully in predicting secondary structure and other tasks (52). In this work, we used embeddings from the PLM, ProtT5-XL-Uniref (herein called, ProtT5) (46). The PLM ProtT5 was trained on unlabeled protein sequences from BFD (Big Fantastic Database; 2.5 billion sequences including meta-genomic sequences) (66), and UniRef50 (67). ProtT5 has been built in analogy to the NLP (Natural Language Processing) T5, ultimately learning some of the constraints of protein sequences (68). 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 using transfer learning. As ProtT5 was only trained on unlabeled protein sequences, there is no risk of information leakage or overfitting to a certain level during pretraining. Essentially, ProtT5 outputs fixed length (1024) vector representations for each residue in a protein

sequence. In essence, to predict whether an amino acid lysine is SUMOylated, SUMO2/3 or not, we extracted a 1024-dimensional vector for each SUMOylated, SUMO2/3 or non-SUMOylated, non-SUMO2/3 lysine residue, where only the encoder side of ProtT5 was used, and embeddings were extracted from the last hidden layer of the models. A similar methodology was applied to extract features from influential Ankh PLM (41). We utilized the Ankh large model because our experimental results show that it can encode more intrinsic information about proteins than the Ankh base model. The only difference from the ProtT5 model was that it produced a per residue contextualized embedding feature vector length of 1536 rather than 1024 produced by ProtT5.

Machine learning and deep learning models

Naïve Bayes (NB) is a simple ML algorithm commonly used for classification tasks (69). It is based on Bayes theorem and assumes that the features are conditionally independent given the class label.

Support Vector Machine (SVM) is a class of supervised machine learning algorithms used for classification and regression tasks (70). The basic idea behind SVM is to find an optimal hyperplane that separates the data into different classes. When the data is not linearly separable, SVM can still classify it by using kernel trick. The kernel trick maps the input data into a higher-dimensional feature space, where it might become linearly separable.

Random Forest (RF) is a popular ensemble learning method used for classification and regression tasks in ML (71). It is an extension of decision trees and combines multiple decision trees to make predictions. For classification tasks, it predicts the class label by taking a majority vote among the individual trees. Each tree's prediction is counted, and the class with the most votes becomes the final prediction.

Logistic Regression (LR) is a ML algorithm used for binary classification tasks (72). It predicts the probability of an instance belonging to a certain class by fitting a logistic (sigmoid) function to the input features. It estimates coefficients to create a linear decision boundary that separates the two classes.

Extreme Gradient Boosting (XGBoost) belongs to the family of gradient boosting method (73). It sequentially adds weak models (decision trees) to iteratively correct the errors made by previous models. It optimizes a specific loss function by finding the best-fitting model in an additive manner.

1D Convolutional Neural Network (1D CNN) is a variant of convolutional neural networks (CNNs) specifically designed for processing one-dimensional sequential data (74). It utilizes one-dimensional convolutional filters to capture local patterns and features in sequential data. The filters slide along the input sequence, performing convolutions and generating feature maps. While traditional CNNs are commonly used for image analysis and computer vision, 1D CNN is particularly suited for tasks involving sequential data, such as time series analysis, speech recognition, and natural language processing. Long Short-Term Memory (LSTM) is a type of recurrent network (RNN) architecture specifically designed to model and process sequential data. It addresses the vanishing gradient problem that occurs in traditional RNNs, allowing for better capturing of long-term dependencies in the data. The hyperparameters and other details are explained in [Supplementary Table S2](#).

Model training

As discussed above, SUMOylation and SUMO2/3 occur on lysine residues, 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 lysine) is extracted (1024-dimensional vector) and passed to the subsequent DL model. Using these representations and datasets (CPLM 4.0, SUMO2/3, and GPS-SUMO), we trained several models to correctly predict SUMOylation, and SUMO2/3 sites in amino acid sequences. The performance of several architectures was evaluated: 1D CNN, 1D CNN-LSTM, 1D CNN-BiLSTM, BiLSTM, LSTM, LR, MLP, SVM, XGBoost, NB and RF. We describe the MLP architecture in Figure 1. As shown in Figure 1, the features are extracted for the site of interrogation (K, highlighted in white) using full protein sequence as input and the 1024 real-valued feature vectors are fed into a MLP deep-learning architecture consisting of 64 neuron input layers followed by 2 neuron output layers. To explore the hyperparameter space, we performed a ten-fold cross-validation grid search on the MLP deep learning model with the CPLM 4.0, and SUMO2/3 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 1024 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 the rest of the deep learning and machine learning algorithms. The optimized hyperparameters using grid search are shown in Table 4. Based upon grid search, 64 neuron input layers were configured with ReLU activation function. As dropout layer/nodes in the network helped alleviate overfitting and improved the generalization capacity, we set the dropout equal to 0.3. Our task was to train a binary classification model to distinguish SUMOylation or SUMO2/3, and non-SUMOylation or SUMO2/3 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 4. 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 Figure S1](#). The loss curve for the training and validation follows each other as well as the training and validation accuracy curves also follow each other.

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 10-fold cross-validation, the data are partitioned into ten equal parts. Then, one part is left out for validation while 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 \pm 1 standard deviation from the mean.

To evaluate the performance of each model, we use accuracy (ACC), sensitivity (SN), specificity (SP) and Matthews correlation coefficient (MCC) (75,76). ACC describes the correctly predicted residues out of the total residues

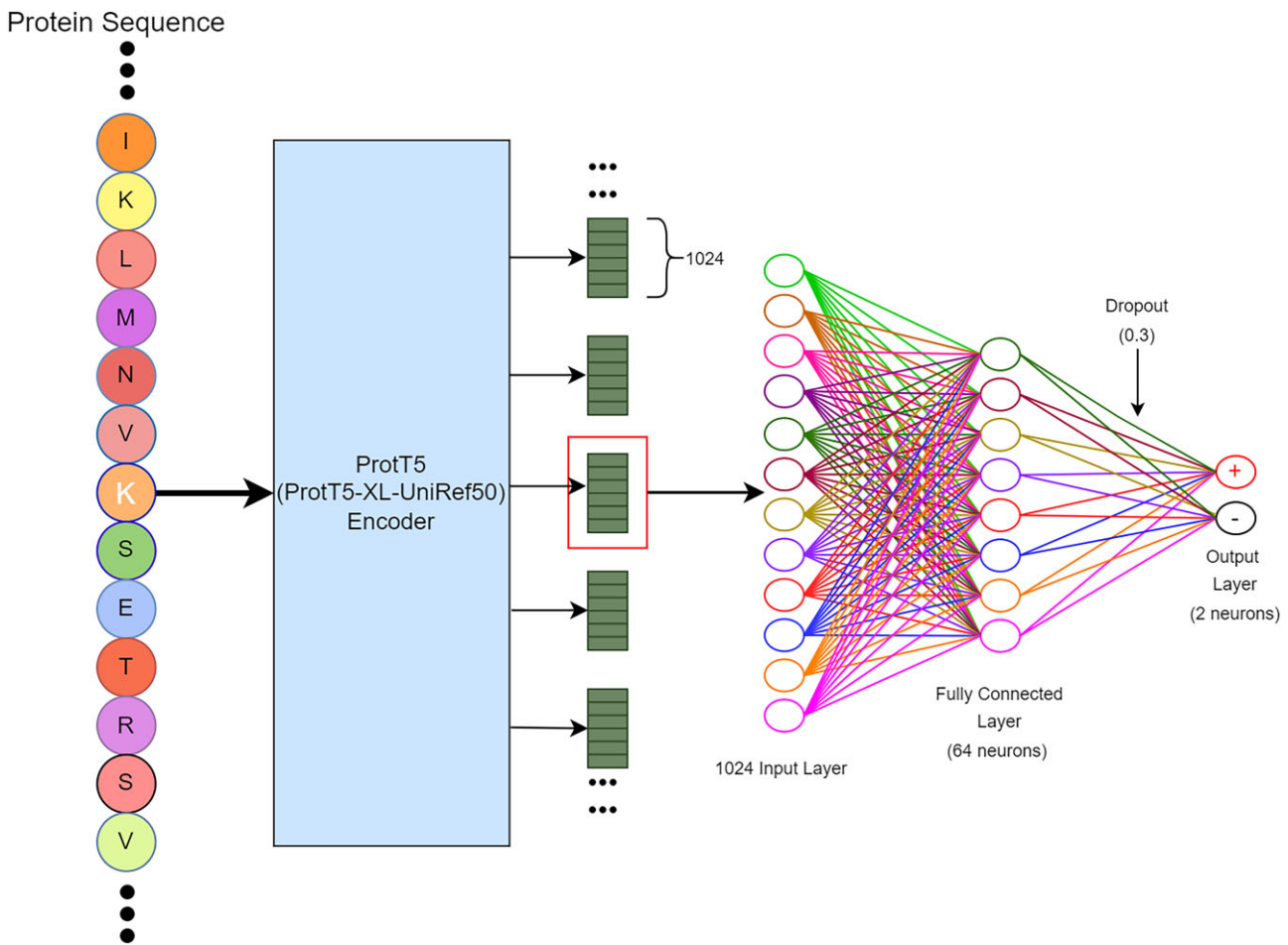


Figure 1. The overall framework of SumoPred-PLM. Beads with letters represent protein sequences. The sky-colored rectangular box represents ProtT5 PLM. Green rectangular boxes are per residue 1024 features representations produced by ProtT5 PLM. The empty circle represents neurons. Each neuron is connected to other nodes via links like a biological axon-synapse-dendrite connection. A dropout of 0.3 means, 30% of neurons are switched off randomly while training the MLP.

Table 4. Hyperparameters used in the MLP network for the SUMOylation, SUMO2/3 and GPS-SUMO datasets

Name of the Parameters	Value Used
No. of layers	1
No. neuron in dense layers	64
No. of neuron in the output layer	2
Activation Function	ReLU
Activation Function at output layer	SoftMax
Optimizer	Adam
Learning rate	0.001
Objective / loss function	Binary cross entropy
Model Checkpoint	Monitor = 'Validation accuracy'
Reduce learning rate on plateau	Factor = 0.001
Early stopping	Patience = 5
Dropout	0.3
Decision Boundary	0.5
Batch size	256
Epochs	400

(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)). On the other hand, MCC considers the model’s

predictive capability concerning both positive and negative residues (Equation (4)).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (1)$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \quad (2)$$

$$Specificity = \frac{TN}{TN + FP} \times 100 \quad (3)$$

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

Results

SumoPred-PLM utilizes per residue embeddings (1024 features) extracted for the site of interest (K) from ProtT5 using a full-length sequence as input. We use three datasets for training SumoPred-PLM: CPLM 4.0, SUMO2/3, and GPS-SUMO. Protein redundancies are removed from within and across training and independent test datasets. We performed

Table 5. Results of the 10-fold cross-validation on the CPLM 4.0 training dataset using different deep and machine learning models encoded with ProtT5 PLM. The highest values in each column are highlighted in bold

Models	MCC \pm 1 S.D.	SN \pm 1 S.D.	SP \pm 1 S.D.	ACC \pm 1 S.D.
MLP	0.478 \pm 0.010	0.757 \pm 0.026	0.720 \pm 0.026	0.738 \pm 0.005
LR	0.444 \pm 0.009	0.730 \pm 0.006	0.713 \pm 0.006	0.722 \pm 0.004
XGBoost	0.390 \pm 0.011	0.701 \pm 0.006	0.689 \pm 0.007	0.695 \pm 0.005
RF	0.331 \pm 0.008	0.578 \pm 0.007	0.748 \pm 0.005	0.663 \pm 0.004
NB	0.229 \pm 0.012	0.448 \pm 0.006	0.768 \pm 0.007	0.608 \pm 0.005
SVM	0.461 \pm 0.022	0.729 \pm 0.017	0.732 \pm 0.015	0.730 \pm 0.011
1D-CNN	0.449 \pm 0.012	0.735 \pm 0.038	0.712 \pm 0.040	0.720 \pm 0.006

10-fold cross-validation on the training dataset(s) to obtain the best hyperparameters for our deep learning 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.

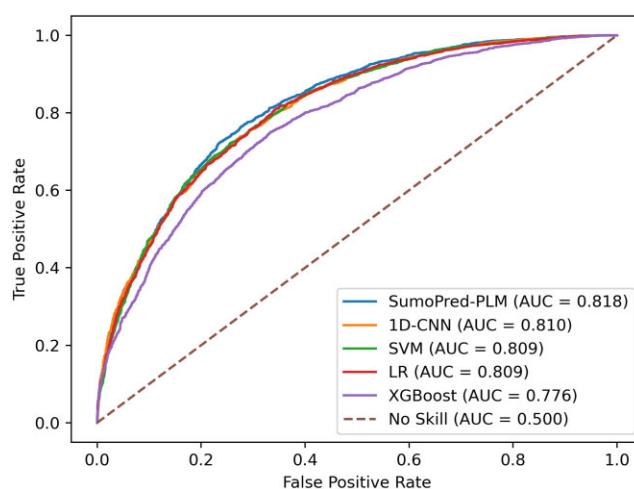
Performance on the CPLM 4.0 dataset

10-fold cross-validation on the CPLM 4.0 training set with ProtT5 features

To tune the hyperparameters (parameters whose values are used to control the learning process) and to investigate the performance of various DL/ML models, we performed 10-fold cross-validation on the CPLM 4.0 training dataset (77). The predictive performance of different DL and ML models using the stratified 10-fold cross-validation on the CPLM 4.0 training data set is shown in Table 5. The contextualized embedding of the SUMOylated or non-SUMOylated token 'K' produced by the pretrained ProtT5 model when fed to MLP achieves the best performance as seen in Table 5. Intriguingly, the same architecture (MLP) produced the highest result using 10-fold cross-validation on the SUMO2/3 training data set as well. This MLP model produced MCC, SN, SP, and ACC values of 0.478 \pm 0.010, 0.757 \pm 0.026, 0.720 \pm 0.026 and 0.738 \pm 0.005 respectively for the stratified 10-fold cross-validation. Since the MLP model produced the best result on 10-fold cross-validation, we selected this architecture as our final model and called it SumoPred-PLM. Furthermore, we conducted a 10-fold cross-validation on the CPLM 4.0 training dataset with the 1D CNN-BiLSTM, 1D CNN-LSTM, BiLSTM and LSTM DL methods. The findings are presented in [Supplementary Table S3](#), revealing subpar performance of these models.

Testing on CPLM 4.0 independent test dataset with ProtT5 feature

To assess the performance of our approach on an independent test set with ProtT5 features, we trained the MLP model on the overall CPLM 4.0 training set and evaluated it with CPLM 4.0 SUMOylation independent test data set. It should be noted that none of the positive or negative SUMOylation sites, nor the protein sequences from the CPLM 4.0 independent test set, are present in the CPLM 4.0 training dataset. We rigidly constrained our experiment with this phenomenon because the PLM can learn representations for other sites from the same protein, which can lead to overestimation of the performance. The total number of samples in each set for CPLM 4.0 dataset is shown in Table 1. Our model achieved MCC, SN, SP, and ACC values of 0.4835, 74.02%, 74.32% and

**Figure 2.** Comparisons of ROC curves of SumoPred-PLM and other models on the SUMOylation CPLM 4.0 independent test dataset. For each model, the area under the ROC curve is reported.

74.17% respectively, on the independent test dataset. Furthermore, MLP was able to classify 2,220 samples as True Negative, 2,212 samples as True Positive, 767 as False Positive and 776 as False Negative. The independent test set result and 10-fold cross-validation results produced by SumoPred-PLM are similar. Moreover, it can be observed from Figure 2 that SumoPred-PLM, which is based on a MLP approach, has the highest area under the receiver operating characteristics curve (ROC). Similarly, Figure 3 shows that SumoPred-PLM has the highest precision-recall area under the curve (PrAUC) compared to other DL and ML approaches. Hence, SumoPred-PLM is a robust computational model for the prediction of SUMOylation PTM in amino acid sequences of proteins. In addition, the SumoPred-PLM MLP model was trained using the dbPTM training dataset (65), utilizing the ProtT5 encoding scheme. Subsequently, the trained model was evaluated with the dbPTM independent test dataset, and the findings are presented in [Supplementary Table S4](#).

Furthermore, McNemar's significant test (78,79) was conducted on the best-performing MLP and SVM classification models. Subsequently, the Chi-square (χ^2) distribution value (0.04) was computed and compared with the alpha (0.05) value. Since the χ^2 value is less than the alpha value, we rejected the null hypothesis, suggesting that there is a significant difference between the SVM and MLP classifiers in predicting the outcomes of an independent test dataset. Moreover, the utility of the recently developed ESM2 (3 billion) PLM (80) on the CPLM 4.0 dataset is illustrated in the [Supplementary Table S5](#).

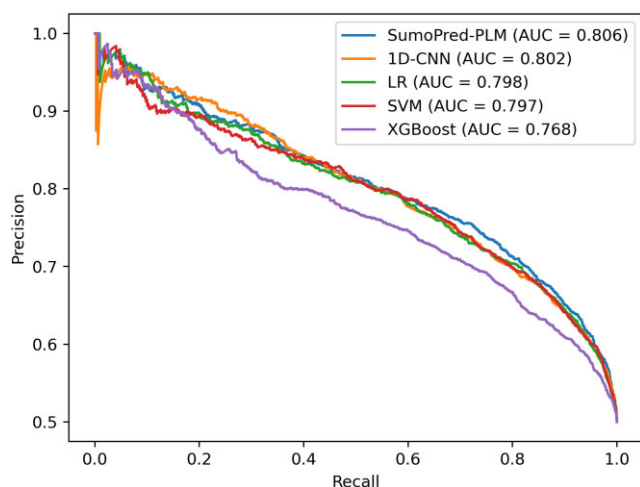


Figure 3. Comparison of precision-recall curves of SumoPred-PLM and other models on the SUMOylation CPLM 4.0 independent test dataset. For each model, the area under the PrAUC is reported.

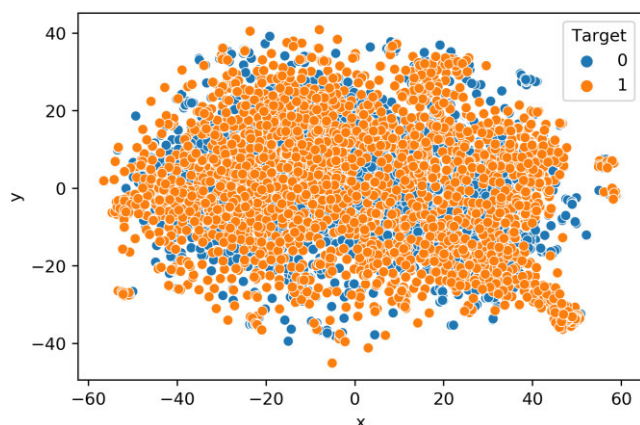


Figure 4. t-SNE illustration of the learned features from ProtT5 language model.

Visualization using t-SNE plot

Additionally, we investigated the classification efficacy of the features and the learned model using t-SNE visualization technique. Herein, features represent the 1024 numeric vectors of SUMOylated, or non-SUMOylated ‘K’ residues extracted from ProtT5, and the learned model refers to the MLP network trained with the CPLM 4.0 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 to project the features into a two-dimensional space (Figure 4) (81). For the features extracted from ProtT5 on the SUMOylated or non-SUMOylated token ‘K’ of CPLM 4.0 training set, the positive and negative samples are relatively clustered together (Figure 4). Figure 5 represents the t-SNE plot of the feature vectors generated from the penultimate hidden layer of the MLP DL architecture when CPLM 4.0 training set is used. This shows that negative samples (blue points) are concentrated at the right while positive samples (orange points) are concentrated at the left, which indicates that the per residue pre-trained PLM feature extraction with MLP learns SUMOylation patterns and largely clusters positive and negative sam-

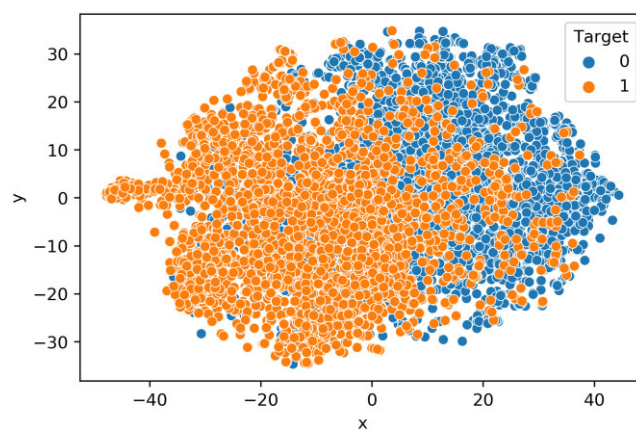


Figure 5. t-SNE illustration of the learned features from the trained MLP model.

ples in two-dimensional space. Hence this result demonstrates that contextualized features produced from pretrained ProtT5 when passed to a MLP deep learning network can cluster positive and negative samples of SUMOylation sites in two-dimensional space.

10-fold cross-validation on the CPLM 4.0 training set with Ankh PLM features

In order to scrutinize the usefulness of recent pre-trained PLM Ankh, we performed 10-fold cross-validation on the CPLM 4.0 training dataset with the embeddings from the Ankh PLM (41). The predictive performance of different DL and ML models using the stratified 10-fold cross-validation on the CPLM 4.0 training data set, where the features are extracted from the Ankh PLM is shown in Table 6. The contextualized embeddings (feature vector length = 1536) of the SUMOylated or non-SUMOylated token ‘K’ produced by the pre-trained Ankh model when fed to MLP achieves the best performance as seen in Table 6. This MLP model produced MCC, SN, SP and ACC values of 0.464 ± 0.010 , 0.752 ± 0.017 , 0.711 ± 0.019 and 0.731 ± 0.005 respectively for the stratified 10-fold cross-validation. These large pretrained PLMs have increased capacity to learn and represent complex patterns of proteins, as well as exhibit better performance in terms of accuracy, generalization, and protein language understanding. Moreover, the token capacity (the maximum number of tokens the model can handle during processing), which affects the model’s ability to handle long sequences of amino acids, is increased in these large pretrained PLMs. Moreover, the 10-fold cross-validation of explored models on CPLM 4.0 training dataset summarizes that Ankh PLM is shorter than the baseline ProtT5 PLM by slight margins, hence we chose pretrained ProtT5 PLM to encode the protein sequence.

Testing on CPLM 4.0 independent test dataset with Ankh feature

To assess the performance of our approach on an independent test set with Ankh features, we trained the MLP model on the overall CPLM 4.0 training set and appraised the trained model with CPLM 4.0 SUMOylation indepen-

Table 6. Results of the 10-fold cross-validation of explored models on the CPLM 4.0 training dataset using Ankh PLM feature encoding. The highest values in each column are highlighted in bold

Models	MCC \pm 1 S.D.	SN \pm 1 S.D.	SP \pm 1 S.D.	ACC \pm 1 S.D.
MLP	0.464 \pm 0.010	0.752 \pm 0.017	0.711 \pm 0.019	0.731 \pm 0.005
LR	0.421 \pm 0.009	0.725 \pm 0.004	0.695 \pm 0.006	0.710 \pm 0.004
XGBoost	0.360 \pm 0.010	0.674 \pm 0.007	0.685 \pm 0.006	0.680 \pm 0.005
RF	0.300 \pm 0.011	0.573 \pm 0.005	0.723 \pm 0.006	0.648 \pm 0.005
NB	0.194 \pm 0.008	0.550 \pm 0.005	0.643 \pm 0.005	0.596 \pm 0.004
SVM	0.461 \pm 0.022	0.729 \pm 0.017	0.732 \pm 0.015	0.730 \pm 0.011
1D-CNN	0.442 \pm 0.010	0.737 \pm 0.006	0.704 \pm 0.007	0.721 \pm 0.005

Table 7. Prediction performance of SumoPred-PLM with ProtT5 and Ankh PLM features on the CPLM 4.0 independent test dataset. The highest values in each column are highlighted in bold

PLM	MCC	SN	SP	ACC
ProtT5	0.483	0.740	0.743	0.741
Ankh	0.472	0.775	0.695	0.735

dent test set. The trained MLP model produced MCC, SN, SP, and ACC of 0.4728, 77.55%, 69.58% and 73.56% respectively, when features from Ankh PLM were used. Furthermore, MLP model trained with Ankh features was able to classify 2077 samples as True Negative, 2315 samples as True Positive, 908 as False Positive, and 670 as False Negative for the CPLM 4.0 independent test dataset. It can be observed from Table 7 that SumoPred-PLM trained with ProtT5 PLM feature representation is better than ANKH PLM feature representation.

Performance on the SUMO2/3 dataset

10-fold cross-validation on the SUMO2/3 training set with ProtT5 features

To further examine the robustness of the proposed model, we performed 10-fold cross-validation on the Hendriks *et al.* SUMO2/3 training dataset. The predictive performance of different DL and ML models using the stratified 10-fold cross-validation on the Hendriks *et al.* SUMO2/3 training data set is shown in Table 8. Intriguingly, the same architecture (MLP) produced the highest performance resulting in MCC, SN, SP and ACC values of 0.481 ± 0.017 , 0.745 ± 0.029 , 0.735 ± 0.027 and 0.740 ± 0.008 , respectively for the stratified 10-fold cross-validation. Since the MLP model produced the best result on 10-fold cross-validation, we selected this architecture as our final model and used it to assess the performance on the Hendriks *et al.* SUMO2/3 independent test set.

Testing on the SUMO2/3 independent test dataset with ProtT5 feature

To further assess the performance, SumoPred-PLM (MLP model) trained on SUMO2/3 dataset was tested with a SUMO2/3 independent test set. The model produced MCC, SN, SP, ACC of 0.4973, 75.88%, 73.83% and 74.86%, respectively on the SUMO2/3 independent test set. Moreover, the SumoPred-PLM was able to classify 937 samples as True Negative, 963 samples as True Positive, 332 as False Positive, and 306 as False Negative on the SUMO2/3 independent test set.

Performance on the GPS-SUMO dataset

Comparison of SumoPred-PLM with State-of-the-Art predictor on GPS-SUMO dataset

To assess the performance of SumoPred-PLM against other approaches, we trained our model on the GPS-SUMO training set and used it to predict SUMOylation sites on the GPS-SUMO independent test set. The MLP model produced MCC, ACC, SN and SP values of 0.3752, 87.56%, 73.23% and 88.30%, respectively on GPS-SUMO independent test dataset. GPS-SUMO produced an area under curve (AUC) of 0.8629 whereas SumoPred-PLM produced an AUC of 0.895 as illustrated in Figure 6. This result is better than the performance of the seminal GPS-SUMO approach, which uses the generation group-based prediction system (GPS) algorithm integrated with Particle Swarm Optimization approach. Furthermore, the MLP classifier was able to classify 1,216 samples as True Negatives, and 52 samples as True Positives. However, it falsely classified 161 samples as False Positive, and 19 samples as False Negative. These results suggest that SumoPred-PLM performs better than the seminal GPS-SUMO method. In addition, it should be noted that SumoPred-PLM was trained and tested with the exact same dataset that was used with the GPS-SUMO approach.

Comparison of SumoPred-PLM with SUMOhydro predictor on SUMOhydro dataset

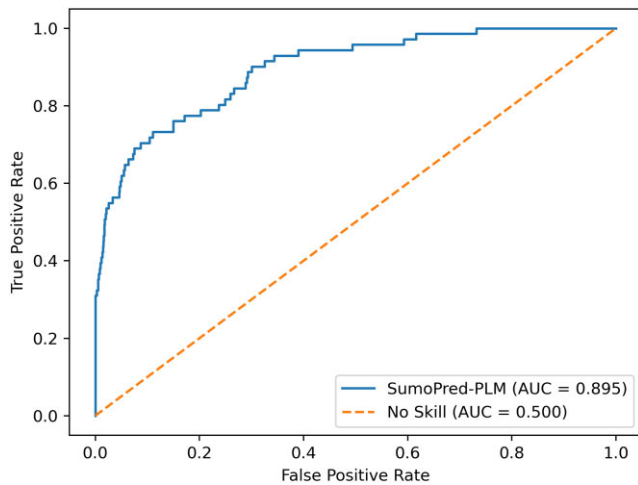
To facilitate a more comprehensive comparison, we have acquired the dataset associated with the SUMOhydro predictor (24). Detailed statistical information pertaining to SUMOhydro dataset is provided in Supplementary Table S6. We extracted the ProtT5 contextualized embedding of the SUMOhydro datasets and then applied these features to the SumoPred-PLM MLP architecture. The results of this analysis are presented in Table 9. It is clear from the results that our predictor outperforms all essential predictors. Additionally, it is important to note that SUMOhydro used a 1:10 ratio of SUMOylation to non-SUMOylation sites in its training dataset. In our experiment, we explored different ratios, and the results indicate that a 1:4 ratio in the training dataset, when combined with ProtT5 embeddings and the MLP architecture, yielded the most optimal outcome. Additionally, SumoPred-PLM MLP model was able to classify 495 samples as True Negative, 17 samples as True Positive, 15 as False Positive, and 7 as False Negative.

Case studies

We performed one case study on the androgen receptor (AR, ANDR_HUMAN UniProt ID: P10275) protein which was not present in training or in the independent test set of CPLM 4.0 dataset. This nuclear steroid receptor is a ligand-activated

Table 8. Comparison of different learning models on Hendriks *et al.* SUMO2/3 training dataset using 10-fold cross-validation, where features were encoded utilizing ProtT5 PLM. The highest values in each column are highlighted in bold

Models	MCC \pm 1 S.D.	SN \pm 1 S.D.	SP \pm 1 S.D.	ACC \pm 1 S.D.
MLP	0.481 \pm 0.017	0.745 \pm 0.029	0.735 \pm 0.027	0.740 \pm 0.008
LR	0.460 \pm 0.020	0.734 \pm 0.017	0.725 \pm 0.015	0.730 \pm 0.010
XGBoost	0.393 \pm 0.030	0.695 \pm 0.016	0.697 \pm 0.018	0.696 \pm 0.014
RF	0.356 \pm 0.017	0.611 \pm 0.010	0.703 \pm 0.012	0.676 \pm 0.08
NB	0.314 \pm 0.015	0.510 \pm 0.017	0.791 \pm 0.009	0.650 \pm 0.008
SVM	0.479 \pm 0.014	0.748 \pm 0.007	0.730 \pm 0.012	0.739 \pm 0.007
1D-CNN	0.449 \pm 0.019	0.726 \pm 0.036	0.721 \pm 0.040	0.724 \pm 0.009

**Figure 6.** ROC curve of SumoPred-PLM on GPS-SUMO independent test dataset.

transcription factor that directs cellular proliferation and differentiation in target tissues (84). Specifically, androgen hormone activated AR binds androgen response elements/ARE on target genes and recruit's coactivator and corepressor proteins to direct gene transcription (85). The AR protein is subject to multiple PTMs including SUMOylation. We and others report AR SUMOylation regulates AR function as our collective whole animal and cell-based studies demonstrate that a disruption of dynamic AR SUMOylation directs aberrant proliferation of prostate and breast cancer cells (86–89). The human androgen receptor protein contains 40 lysine ('K') residues. Biochemical studies first identified canonical SUMO consensus sites that include K387 and K520 on AR (highlighted on the Table S3). The K387 and K520 serve as acceptor sites for both SUMO1 and SUMO2/3 modification of endogenous AR protein in several cell lines. However, mono-SUMO1 and poly-SUMO2/3 chains differentially regulate AR function. SUMO1 modification of AR effects transcriptional activity while SUMO2/3 conjugation to AR directs chromatin enrichment and AR protein stability/degradation (90,91). With 40 lysine residues, we postulated that AR protein may exhibit additional non-consensus SUMO motif and possibly even several SUMO paralogue-specific acceptor sites. Our *in silico* analysis with GPS-SUMO of the primary amino acid sequence of AR identified K387 and K520 and three additional SUMO-acceptor sites (K313, K910, K913, Table S3). However, this platform does not distinguish between SUMO paralogue conjugates. Hence, we next evaluated published mass spectrometry data of the endogenous SUMO2/3 proteome from HeLa cells. The dataset reports that 20% of AR is SUMOylated in

Table 9. Comparison of SumoPred-PLM with other predictors that were trained with SUMOhydro training dataset and tested with SUMOhydro independent test dataset

Method	SN (%)	SP (%)	ACC (%)	MCC
SumoPred-PLM	70.8	97.0	95.8	0.592
JASSA (4)	50.0	94.0	89.3	0.442
SUMOhydro (24)	66.7	93.5	92.3	0.432
SUMOSP2.0 (82)	62.5	92.6	91.2	0.381
seeSUMO-SVM (83)	54.2	95.1	93.3	0.397
seeSUMO-RF (83)	70.8	88.4	87.6	0.351

HeLa cells and eight SUMO2/3-accepting lysine residues of AR are conjugated (five sites analogous to GPS-SUMO and three novel SUMO2/3-acceptors at positions K241, K638, and K823). We then challenged the SumoPred-PLM with the same task of identifying AR SUMOylation sites. As shown in Figure 7, SumoPred-PLM correctly predicts the five validated (K387, K520, K638, K910, K912) SUMO2/3 conjugation sites out of five, thus achieving an accuracy of 100.0% (=5/5). In addition, SumoPred-PLM predicts an additional 14 lysine 'K' sites of ANDR_HUMAN protein as positives. Next, we tested if our platform could identify validated and/or novel SUMO2/3 acceptor sites for AR. The model trained with the SUMO2/3 dataset shows 11 SUMO2/3 acceptors at lysine position 181, 290, 387, 520, 618, 638, 658, 861, 905, 910 and 912. Six of the predicted SUMO2/3 sites are novel and previously reported in Hendriks *et al.*; specifically, K181, 290, 618, 658, 861, 905. Hence, these lysine residues that are predicted to be SUMO2/3 modified await experimental validation. [Supplementary Table S7](#) shows prediction results of SumoPred-PLM for all the lysine in ANDR_HUMAN protein.

Discussion and conclusions

One of the key innovations in SumoPred-PLM is the incorporation of PLM based features to represent protein sequences. PLM based features have proven to be quite useful in various bioinformatics tasks (92–100). Our major goal in the project was to move away from hand-crafted feature extraction for prediction of SUMOylation and SUMO2/3 sites. To achieve this goal, we investigated whether language models learned from a large amount of protein sequences could capture the features predictive of SUMOylation and SUMO2/3 sites. Additionally, we also wanted to investigate what type of machine learning approach would work well on these pre-trained feature representations. Moreover, our other significant contribution is the study of SUMO2/3 data set which was not extensively studied in prior studies. To achieve the goal, we used contextualized embeddings learned from a PLM called ProtT5

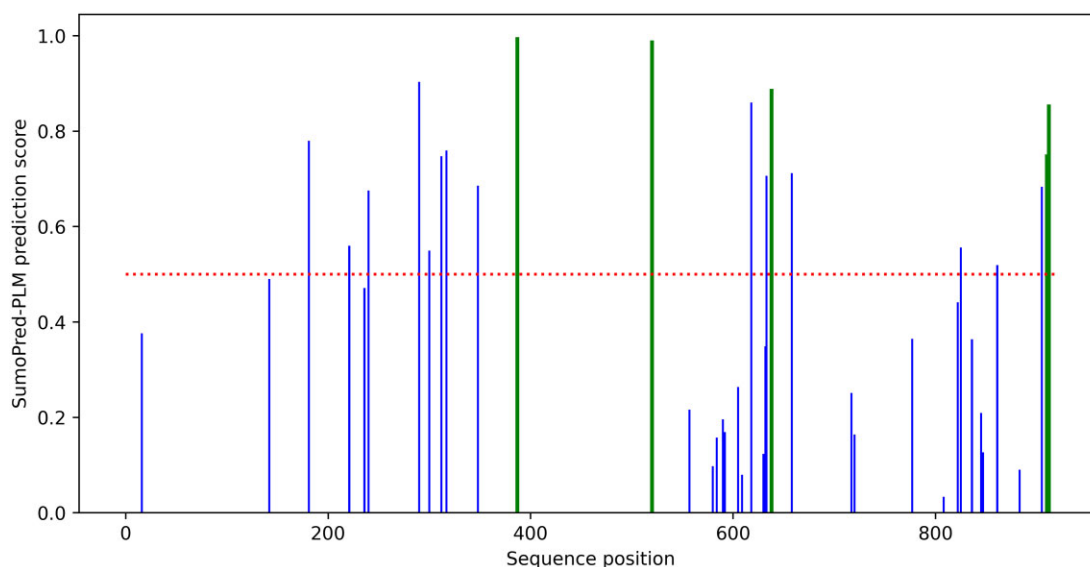


Figure 7. SumoPred-PLM prediction results of human androgen receptor, where sites with a prediction score above 0.5 (shown by the red dotted line) are predicted as SUMOylated sites. Green bars represent the five SUMOylation sites with experimental evidence from protein microarray data.

to extract features for the site of interest. Subsequently, various ML and DL algorithms were evaluated using 10-fold cross validation and the top performing model was selected as the final model. The MLP model, namely SumoPred-PLM, achieves the best prediction performance among the compared methods as it largely benefits from the knowledge obtained from large sets of protein sequences by the pre-trained ProtT5 model that is used to encode the protein sequences.

SumoPred-PLM does not rely on knowledge of protein structure, nor in the expert-crafted sequence features or time-consuming evolutionary information derived from multiple sequence alignments (MSAs). Instead, the input to the MLP model is a contextual representation of the SUMOylated or non-SUMOylated token ‘K’ from the pre-trained PLM (ProtT5). This state-of-the-art prediction of SUMOylation 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. SumoPred-PLM model outperforms the pioneering GPS-SUMO predictor, in the identification of consensus and non-consensus SUMO-acceptor sites. One interesting result portrayed in the t-SNE plot (Figure 5) is that our model was largely able to cluster the two classes of SUMOylated and non-SUMOylated lysine residues in two-dimensional space. SumoPred-PLM is a new approach proposed in this work that uses information distilled from large PLMs to train the DL framework and results an outstanding performance compared to existing approaches. In the future, we will consider using the structural information predicted by AlphaFold2 (101,102) to build models using graph networks (103) for further improving the performance of SUMOylation and SUMO2/3 PTM site prediction.

In addition, we provide a unique service for SumoPred-PLM as a SUMO2/3-specific predictor. To our knowledge, this is the first platform that provides the ability to predict SUMO2/3 paralogue selective acceptor sites. As stated previously, increasing biochemical studies highlight SUMO paralogue differentially effect a protein substrate’s function and stability. Hence, we anticipate that this SUMO2/3 predictor

platform will greatly accelerate the discovery of this SUMO-paralog directed protein effects. For the SUMO2/3 platform, protein machine learning was based on available large-scale SUMO2/3 proteomics data (62). Unfortunately, a similar SUMO1 proteomic analysis is unavailable currently but, when accessible, this dataset can be easily incorporated into the current standing platform.

Data availability

All programs and data are available at <https://github.com/PakhrinLab/SumoPred-PLM> and <https://doi.org/10.6084/m9.figshare.25009160>.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

We acknowledge the use of the BeoShock, Beocat, and Sabine High-Performance Computing resources located at Wichita State University, Kansas State University, and the University of Houston, respectively. We appreciate the email discussions with Dr. Ivo A. Hendriks regarding their generated SUMO2/3 dataset. In addition, we would also like to acknowledge the dbPTM, CPLM 4.0, Hendriks *et al.* SUMO2/3, GPS-SUMO dataset, ProtT5-XL-UniRef50, ESM2 (3B) and Ankh Protein Language Model made freely available for the researchers .

Author contributions: S.C.P, T.B.K., M.R.B., E.B., and A.M. conceived of and designed the experiments; S.C.P., P.A., and A.V.P. performed all the experiments and data analysis, E.B., and A.M. verified all the programs and models. A.S.P. performed case studies. S.C.P, T.B.K., M.R.B., E.B., P.A., A.V.P., A.S.P., and A.M. revised the manuscript. All authors have read and agreed to the published version of the manuscript. S.C.P. oversaw the whole project.

Funding

National Cancer Institute of the National Institutes of Health [R01CA256543 to T.B.K.]; faculty startup fund provided to S.C.P. by U.H.D. The Department of Homeland Security [21STSLA00011-01-0 and 23STSLA00017-01-00 to A.M.].

Conflict of interest statement

None declared.

References

- Olsen, J.V. and Mann, M. (2013) Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol. Cell. Proteomics*, **12**, 3444–3452.
- Jensen, O.N. (2006) Interpreting the protein language using proteomics. *Nat. Rev. Mol. Cell Biol.*, **7**, 391–403.
- Flotho, A. and Melchior, F. (2013) Sumoylation: a regulatory protein modification in health and disease. *Annu. Rev. Biochem.*, **82**, 357–385.
- Beauclair, G., Bridier-Nahmias, A., Zagury, J.F., Saib, A. and Zamborlini, A. (2015) JASSA: a comprehensive tool for prediction of SUMOylation sites and SIMs. *Bioinformatics*, **31**, 3483–3491.
- Kumar, A. and Zhang, K.Y. (2015) Advances in the development of SUMO specific protease (SENp) inhibitors. *Comput. Struct. Biotechnol. J.*, **13**, 204–211.
- Feligioni, M. and Nistico, R. (2013) SUMO: a (oxidative) stressed protein. *Neuromolecular Med.*, **15**, 707–719.
- Droescher, M., Chaugule, V.K. and Pichler, A. (2013) SUMO rules: regulatory concepts and their implication in neurologic functions. *Neuromol. Med.*, **15**, 639–660.
- Lu, L., Shi, X.H., Li, S.J., Xie, Z.Q., Feng, Y.L., Lu, W.C., Li, Y.X., Li, H. and Cai, Y.D. (2010) Protein sumoylation sites prediction based on two-stage feature selection. *Mol. Divers.*, **14**, 81–86.
- Jansen, N.S. and Vertegaal, A.C.O. (2021) A chain of events: regulating target proteins by SUMO polymers. *Trends Biochem. Sci.*, **46**, 113–123.
- Mustfa, S.A., Singh, M., Suhail, A., Mohapatra, G., Verma, S., Chakravorty, D., Rana, S., Rampal, R., Dhar, A., Saha, S., et al. (2017) SUMOylation pathway alteration coupled with downregulation of SUMO E2 enzyme at mucosal epithelium modulates inflammation in inflammatory bowel disease. *Open Biol.*, **7**, 170024.
- Eifler, K. and Vertegaal, A.C. (2015) Mapping the SUMOylated landscape. *FEBS J.*, **282**, 3669–3680.
- Ramazi, S. and Zahiri, J. (2016) Computational prediction of proteins sumoylation: a review on the methods and databases. *J. Nanomed. Res.*, **3**, <https://doi.org/10.15406/jnmr.2016.03.00068>.
- Jentsch, S. and Psakhye, I. (2013) Control of nuclear activities by substrate-selective and protein-group SUMOylation. *Annu. Rev. Genet.*, **47**, 167–186.
- Ramazi, S. and Zahiri, J. (2021) Posttranslational modifications in proteins: resources, tools and prediction methods. *Database (Oxford)*, **2021**, baab012.
- Tatham, M.H., Jaffray, E., Vaughan, O.A., Desterro, J.M., Botting, C.H., Naismith, J.H. and Hay, R.T. (2001) Polymeric chains of SUMO-2 and SUMO-3 are conjugated to protein substrates by SAE1/SAE2 and Ubc9. *J. Biol. Chem.*, **276**, 35368–35374.
- Keiten-Schmitz, J., Schunck, K. and Muller, S. (2019) SUMO chains rule on chromatin occupancy. *Front. Cell Dev. Biol.*, **7**, 343.
- Bouchard, D., Wang, W., Yang, W.C., He, S., Garcia, A. and Matunis, M.J. (2021) SUMO paralogue-specific functions revealed through systematic analysis of human knockout cell lines and gene expression data. *Mol. Biol. Cell*, **32**, 1849–1866.
- Evdokimov, E., Sharma, P., Lockett, S.J., Lualdi, M. and Kuehn, M.R. (2008) Loss of SUMO1 in mice affects RanGAP1 localization and formation of PML nuclear bodies, but is not lethal as it can be compensated by SUMO2 or SUMO3. *J. Cell Sci.*, **121**, 4106–4113.
- Wang, L., Wansleben, C., Zhao, S., Miao, P., Paschen, W. and Yang, W. (2014) SUMO2 is essential while SUMO3 is dispensable for mouse embryonic development. *EMBO Rep.*, **15**, 878–885.
- Medzihradsky, K.F. (2005) Peptide sequence analysis. *Methods Enzymol.*, **402**, 209–244.
- Agarwal, K.L., Kenner, G.W. and Sheppard, R.C. (1969) Feline gastrin. An example of peptide sequence analysis by mass spectrometry. *J. Am. Chem. Soc.*, **91**, 3096–3097.
- Zhang, W., Tan, X., Lin, S., Gou, Y., Han, C., Zhang, C., Ning, W., Wang, C. and Xue, Y. (2022) CPLM 4.0: an updated database with rich annotations for protein lysine modifications. *Nucleic Acids Res.*, **50**, D451–D459.
- Xue, Y., Zhou, F., Fu, C., Xu, Y. and Yao, X. (2006) SUMOsp: a web server for sumoylation site prediction. *Nucleic Acids Res.*, **34**, W254–W257.
- Chen, Y.Z., Chen, Z., Gong, Y.A. and Ying, G. (2012) SUMOhydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties. *PLoS One*, **7**, e39195.
- Yavuz, A.S. and Sezerman, O.U. (2014) Predicting sumoylation sites using support vector machines based on various sequence features, conformational flexibility and disorder. *Bmc Genomics [Electronic Resource]*, **15**, S18.
- Xu, Y., Ding, Y.X., Deng, N.Y. and Liu, L.M. (2016) Prediction of sumoylation sites in proteins using linear discriminant analysis. *Gene*, **576**, 99–104.
- Dehzeni, A., López, Y., Taherzadeh, G., Sharma, A. and Tsunoda, T. (2018) SumSec: accurate prediction of sumoylation sites using predicted secondary structure. *Molecules*, **23**, 3260.
- Sharma, A., Lysenko, A., López, Y., Dehzeni, A., Sharma, R., Reddy, H., Sattar, A. and Tsunoda, T. (2019) HseSUMO: sumoylation site prediction using half-sphere exposures of amino acids residues. *Bmc Genomics [Electronic Resource]*, **19**, 982.
- Khan, Y.D., Khan, N.S., Naseer, S. and Butt, A.H. (2021) iSUMOK-PseAAC: prediction of lysine sumoylation sites using statistical moments and Chou's PseAAC. *PeerJ*, **9**, e11581.
- López, Y., Dehzeni, A., Reddy, H.M. and Sharma, A. (2020) C-iSUMO: a sumoylation site predictor that incorporates intrinsic characteristics of amino acid sequences. *Comput. Biol. Chem.*, **87**, 107235.
- Pakhrin, S.C. and Pant, D.R. (2018) In: *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. IEEE, Greater Noida, India, pp. 451–455.
- Pakhrin, S.C., Aoki-Kinoshita, K.F., Caragea, D. and Kc, D.B. (2021) DeepNGlyPred: a Deep Neural Network-Based Approach for Human N-Linked Glycosylation Site Prediction. *Molecules*, **26**, 7314.
- Pakhrin, S.C., Shrestha, B., Adhikari, B. and Kc, D.B. (2021) Deep learning-based advances in protein structure prediction. *Int. J. Mol. Sci.*, **22**, 5553.
- Pakhrin, S.C., Pokharel, S., Saigo, H. and Kc, D.B. (2022) In: *Deep Learning-Based Advances In Protein Posttranslational Modification Site and Protein Cleavage Prediction*. 2022/06/14 ed.
- Xu, J., He, Y., Qiang, B., Yuan, J., Peng, X. and Pan, X.M. (2008) A novel method for high accuracy sumoylation site prediction from protein sequences. *BMC Bioinf.*, **9**, 8.
- Zhao, Q., Xie, Y., Zheng, Y., Jiang, S., Liu, W., Mu, W., Liu, Z., Zhao, Y., Xue, Y. and Ren, J. (2014) GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. *Nucleic Acids Res.*, **42**, W325–W330.
- Unsal, S., Atas, H., Albayrak, M., Turhan, K., Acar, A.C. and Doğan, T. (2022) Learning functional properties of proteins with language models. *Nature Machine Intelligence*, **4**, 227–245.
- Vaswani, A.S., Parmar, N., Uszkoreit, N., Jones, J., Llion, G., Łukasz, A.N.K. and Polosukhin, I. (2017) Attention Is All You

- Need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*.
39. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., et al. (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. USA*, **118**, e2016239118.
 40. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. and Linial, M. (2022) ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, **38**, 2102–2110.
 41. 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. arXiv doi: <https://arxiv.org/abs/2301.06568>, 16 January 2023, preprint: not peer reviewed.
 42. Cao, Y. and Shen, Y. (2021) TALE: transformer-based protein function Annotation with joint sequence-Label Embedding. *Bioinformatics*, **37**, 2825–2833.
 43. Zhang, S., Fan, R., Liu, Y., Chen, S., Liu, Q. and Zeng, W. (2023) Applications of transformer-based language models in bioinformatics: a survey. *Bioinform. Adv.*, **3**, vbad001.
 44. Li, F.-Z., Amini, A.P., Yang, K.K. and Lu, A.X. (2022) In: *Machine Learning for Structural Biology Workshop, NeurIPS 2022*.
 45. Vig, J., Madani, A., Varshney, L.R., Xiong, C., Socher, R. and Rajani, N.F. (2020) BERTology Meets Biology: interpreting Attention in Protein Language Models. arXiv doi: <https://arxiv.org/abs/2006.15222>, 26 June 2020, preprint: not peer reviewed.
 46. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. (2022) ProfTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, **44**, 7112–7127.
 47. Weissenow, K., Heinzinger, M. and Rost, B. (2022) Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. *Structure*, **30**, 1169–1177.
 48. Nallapareddy, V., Bordin, N., Sillitoe, I., Heinzinger, M., Littmann, M., Waman, V., Sen, N., Rost, B. and Orengo, C. (2023) CATH: detection of remote homologues for CATH superfamilies using embeddings from protein language models. *Bioinformatics*, **39**, btad029.
 49. Heinzinger, M., Littmann, M., Sillitoe, I., Bordin, N., Orengo, C. and Rost, B. (2022) Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genom. Bioinform.*, **4**, lqac043.
 50. Marquet, C., Heinzinger, M., Olenyi, T., Dallago, C., Erckert, K., Bernhofer, M., Nechaev, D. and Rost, B. (2022) Embeddings from protein language models predict conservation and variant effects. *Hum. Genet.*, **141**, 1629–1647.
 51. Pakhrin, S.C., Pokharel, S., Aoki-Kinoshita, K.F., Beck, M.R., Dam, T.K., Caragea, D. and Kc, D.B. (2023) LMNglyPred: prediction of human N-linked glycosylation sites using embeddings from a pre-trained protein language model. *Glycobiology*, **33**, 411–422.
 52. Littmann, M., Heinzinger, M., Dallago, C., Weissenow, K. and Rost, B. (2021) Protein embeddings and deep learning predict binding residues for various ligand classes. *Sci. Rep.*, **11**, 23916.
 53. Pakhrin, S.C. (2022) In: *Deep learning-based approaches for prediction of post-translational modification sites in proteins*. Ph.D. diss. Wichita State University.
 54. Teufel, F., Almagro Armenteros, J.J., Johansen, A.R., Gislason, M.H., Pihl, S.I., Tsirigos, K.D., Winther, O., Brunak, S., von Heijne, G. and Nielsen, H. (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.*, **40**, 1023–1025.
 55. Liu, Y., Liu, Y., Wang, G.A., Cheng, Y., Bi, S. and Zhu, X. (2022) BERT-Kgly: a Bidirectional Encoder Representations From Transformers (BERT)-Based Model for Predicting Lysine Glycation Site for Homo sapiens. *Front Bioinform.*, **2**, 834153.
 56. Pakhrin, S.C., Chauhan, N., Khan, S., Upadhyaya, J., Keller, C., Neuman, L.N., Beck, M.R. and Blanco, E. (2023) Human O-linked Glycosylation Site Prediction Using Pretrained Protein Language Model. bioRxiv doi: <https://doi.org/10.1101/2023.10.23.563673>, 24 October 2023, preprint: not peer reviewed.
 57. Pakhrin, S.C., Pokharel, S., Pratyush, P., Chaudhari, M., Ismail, H.D. and Kc, D.B. (2023) LMPHosSite: a deep learning-based approach for general protein phosphorylation site prediction using embeddings from the local window sequence and pretrained protein language model. *J. Proteome Res.*, **22**, 2548–2557.
 58. Qiao, Y., Zhu, X. and Gong, H. (2021) BERT-Kcr: prediction of lysine crotonylation sites by a transfer learning method with pre-trained BERT models. *Bioinformatics*, **38**, 648–654.
 59. Thummuluri, V., Almagro Armenteros, J.J., Johansen, A.R., Nielsen, H. and Winther, O. (2022) DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Res.*, **50**, W228–W234.
 60. Høie, M.H., Kiehl, E.N., Petersen, B., Nielsen, M., Winther, O., Nielsen, H., Hallgren, J. and Marcatili, P. (2022) NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic Acids Res.*, **50**, W510–W515.
 61. Song, Y., Yuan, Q., Chen, S., Chen, K., Zhou, Y. and Yang, Y. (2023) Fast and accurate protein intrinsic disorder prediction by using a pretrained language model. *Brief Bioinform.*, **24**, bbad173.
 62. Hendriks, I.A., Lyon, D., Su, D., Skotte, N.H., Daniel, J.A., Jensen, L.J. and Nielsen, M.L. (2018) Site-specific characterization of endogenous SUMOylation across species and organs. *Nat. Commun.*, **9**, 2456.
 63. Huang, Y., Niu, B., Gao, Y., Fu, L. and Li, W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.
 64. Lemaitre, G., Nogueira, F. and Aridas, C.K. (2017) Imbalanced-learn: a Python Toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.*, **18**, 559–563.
 65. Li, Z., Li, S., Luo, M., Jhong, J.-H., Li, W., Yao, L., Pang, Y., Wang, Z., Wang, R., Ma, R., et al. (2021) dbPTM in 2022: an updated database for exploring regulatory networks and functional associations of protein post-translational modifications. *Nucleic Acids Res.*, **50**, D471–D479.
 66. Steinegger, M., Mirdita, M. and Soding, J. (2019) Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods*, **16**, 603–606.
 67. UniProt, C. (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
 68. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S.M., Matena, M., Zhou, Y., Li, W. and Liu, P.J. (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, **21**, 5485–5551.
 69. Webb, G.I. (2011) In: Sammut, C. and Webb, G.I. (eds.) *Encyclopedia of Machine Learning*. Springer US, Boston, MA, pp. 713–714.
 70. Smola, A.J. and Schölkopf, B. (2004) A tutorial on support vector regression. *Stat. Comput.*, **14**, 199–222.
 71. Breiman, L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.
 72. Sperandei, S. (2014) Understanding logistic regression analysis. *Biochem. Med. (Zagreb)*, **24**, 12–18.
 73. Friedman, J.H. (2002) Stochastic gradient boosting. *Comput. Stat. Data Anal.*, **38**, 367–378.
 74. LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning. *Nature*, **521**, 436–444.
 75. Yuan, Q., Chen, K., Yu, Y., Le, N.Q.K. and Chua, M.C.H. (2023) Prediction of anticancer peptides based on an ensemble model of deep learning and machine learning using ordinal positional encoding. *Brief Bioinform.*, **24**, bbac630.
 76. Kha, Q.H., Ho, Q.T. and Le, N.Q.K. (2022) Identifying SNARE proteins using an alignment-free method based on multiscale

- convolutional neural network and PSSM profiles. *J. Chem. Inf. Model.*, **62**, 4820–4826.
77. Yang, L. and Shami, A. (2020) On hyperparameter optimization of machine learning algorithms: theory and practice. *Neurocomputing*, **415**, 295–316.
 78. McNemar, Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, **12**, 153–157.
 79. Dietterich, T.G. (1998) Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, **10**, 1895–1923.
 80. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, **379**, 1123–1130.
 81. Maaten, L.v. and Hinton, G. (2008) Visualizing data using t-SNE. *Mach. Learn. Res.*, **9**, 2579–2605.
 82. Ren, J., Gao, X., Jin, C., Zhu, M., Wang, X., Shaw, A., Wen, L., Yao, X. and Xue, Y. (2009) Systematic study of protein sumoylation: development of a site-specific predictor of SUMOsp 2.0. *Proteomics*, **9**, 3409–3412.
 83. Teng, S., Luo, H. and Wang, L. (2012) Predicting protein sumoylation sites from sequence features. *Amino Acids*, **43**, 447–455.
 84. Yang, J., Zhao, Y.L., Wu, Z.Q., Si, Y.L., Meng, Y.G., Fu, X.B., Mu, Y.M. and Han, W.D. (2009) The single-macro domain protein LRP16 is an essential cofactor of androgen receptor. *Endocr. Relat. Cancer*, **16**, 139–153.
 85. Cui, J., Yang, Y., Zhang, C., Hu, P., Kan, W., Bai, X., Liu, X. and Song, H. (2011) FBI-1 functions as a novel AR co-repressor in prostate cancer cells. *Cell. Mol. Life Sci.*, **68**, 1091–1103.
 86. Bahnassy, S., Thangavel, H., Quttina, M., Khan, A.F., Dhanyalayam, D., Ritho, J., Karami, S., Ren, J. and Bawa-Khalife, T. (2020) Constitutively active androgen receptor supports the metastatic phenotype of endocrine-resistant hormone receptor-positive breast cancer. *Cell Commun. Signal.*, **18**, 154.
 87. Bawa-Khalife, T., Cheng, J., Lin, S.H., Ittmann, M.M. and Yeh, E.T. (2010) SENP1 induces prostatic intraepithelial neoplasia through multiple mechanisms. *J. Biol. Chem.*, **285**, 25859–25866.
 88. Bawa-Khalife, T., Cheng, J., Wang, Z. and Yeh, E.T. (2007) Induction of the SUMO-specific protease 1 transcription by the androgen receptor in prostate cancer cells. *J. Biol. Chem.*, **282**, 37341–37349.
 89. Bawa-Khalife, T. and Yeh, E.T. (2010) SUMO losing balance: SUMO proteases disrupt SUMO homeostasis to facilitate cancer development and progression. *Genes Cancer*, **1**, 748–752.
 90. Rytinki, M., Kaikkonen, S., Sutinen, P., Paakinaho, V., Rahkama, V. and Palvimo, J.J. (2012) Dynamic SUMOylation is linked to the activity cycles of androgen receptor in the cell nucleus. *Mol. Cell. Biol.*, **32**, 4195–4205.
 91. Poukka, H., Karvonen, U., Janne, O.A. and Palvimo, J.J. (2000) Covalent modification of the androgen receptor by small ubiquitin-like modifier 1 (SUMO-1). *Proc. Natl. Acad. Sci. USA*, **97**, 14145–14150.
 92. Bepler, T. and Berger, B. (2021) Learning the protein language: evolution, structure, and function. *Cell Syst.*, **12**, 654–669.
 93. Zhou, G., Chen, M., Ju, C.J.T., Wang, Z., Jiang, J.Y. and Wang, W. (2020) Mutation effect estimation on protein-protein interactions using deep contextualized representation learning. *NAR Genom. Bioinform.*, **2**, lqaa015.
 94. Zhang, Y., Lin, J., Zhao, L., Zeng, X. and Liu, X. (2021) A novel antibacterial peptide recognition algorithm based on BERT. *Brief Bioinform.*, **22**, bbab200.
 95. Ferruz, N. and Höcker, B. (2022) Controllable protein design with language models. *Nature Machine Intelligence*, **4**, 521–532.
 96. Singh, J., Paliwal, K., Litfin, T., Singh, J. and Zhou, Y. (2022) Reaching alignment-profile-based accuracy in predicting protein secondary and tertiary structural properties without alignment. *Sci. Rep.*, **12**, 7607.
 97. Singh, J., Litfin, T., Singh, J., Paliwal, K. and Zhou, Y. (2022) SPOT-Contact-LM: improving single-sequence-based prediction of protein contact map using a transformer language model. *Bioinformatics*, **38**, 1888–1894.
 98. Yuan, Q., Chen, S., Wang, Y., Zhao, H. and Yang, Y. (2022) Alignment-free metal ion-binding site prediction from protein sequence through pretrained language model and multi-task learning. *Brief Bioinform.*, **23**, bbab200.
 99. Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F. and Rost, B. (2019) Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinform.*, **20**, 723.
 100. Ferruz, N., Schmidt, S. and Hocker, B. (2022) ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.*, **13**, 4348.
 101. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
 102. Yuan, Q., Chen, S., Rao, J., Zheng, S., Zhao, H. and Yang, Y. (2022) AlphaFold2-aware protein-DNA binding site prediction using graph transformer. *Brief. Bioinform.*, **23**, bbab564.
 103. Yuan, Q., Chen, J., Zhao, H., Zhou, Y. and Yang, Y. (2021) Structure-aware protein-protein interaction site prediction using deep graph convolutional network. *Bioinformatics*, **38**, 125–132.