Journal of proteome research

pubs.acs.org/jpr

Article

LMPhosSite: A Deep Learning-Based Approach for General Protein Phosphorylation Site Prediction Using Embeddings from the Local Window Sequence and Pretrained Protein Language Model

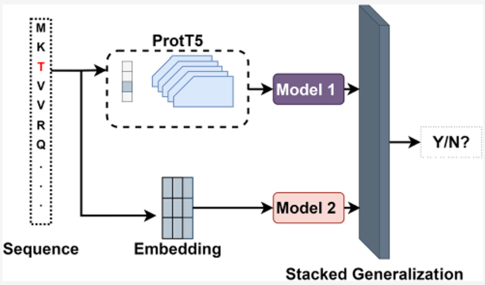
Subash C. Pakhrin, Suresh Pokharel, Pawel Pratyush, Meenal Chaudhari, Hamid D. Ismail, and Dukka B. KC*

Cite This: *J. Proteome Res.* 2023, 22, 2548−2557

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Phosphorylation is one of the most important post-translational modifications and plays a pivotal role in various cellular processes. Although there exist several computational tools to predict phosphorylation sites, existing tools have not yet harnessed the knowledge distilled by pretrained protein language models. Herein, we present a novel deep learning-based approach called LMPhosSite for the general phosphorylation site prediction that integrates embeddings from the local window sequence and the contextualized embedding obtained using global (overall) protein sequence from a pretrained protein language model to improve the prediction performance. Thus, the LMPhosSite consists of two base-models: one for capturing effective local representation and the other for capturing global per-residue contextualized embedding from a pretrained protein language model. The output of these base-models is integrated using a score-level fusion approach. LMPhosSite achieves a precision, recall, Matthew's correlation coefficient, and F1-score of 38.78%, 67.12%, 0.390, and 49.15%, for the combined serine and threonine independent test data set and 34.90%, 62.03%, 0.298, and 44.67%, respectively, for the tyrosine independent test data set, which is better than the compared approaches. These results demonstrate that LMPhosSite is a robust computational tool for the prediction of the general phosphorylation sites in proteins.



KEYWORDS: *post-translational modification, protein language model, phosphorylation, deep learning, stack generalization, score-level fusion, embedding*

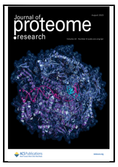
■ INTRODUCTION

Protein phosphorylation is one of the most studied post-translational modifications (PTMs) that plays essential roles in many vital biological processes like cell metabolism, cell motility, apoptosis, replication, transcription, environmental stress responses, DNA repair, immunological responsiveness, and cell cycle control.[1−7] As much as 30% of all eukaryotic proteins are phosphorylated and disruption in the pathway of phosphorylation is associated with the pathological progression of diseases such as Parkinson's, Alzheimer's, cancer, and heart disease,[6,8−10] phosphorylation is commonly found on serine (S), threonine (T), tyrosine (Y), and histidine (H) residues of proteins.[11]

Experimental techniques like low throughput ^{32}P-labeling[12,13] and high throughput mass spectrometry[14,15] are used to detect phosphorylation sites in proteins. However, these techniques are time-consuming and labor-intensive. In this regard, several computational approaches for the prediction of the general phosphorylation sites have been developed. NetPhos[16] is an artificial neural network (ANN)-based approach for prediction of phosphorylation sites. RF-Phos[17] is a random forest-based approach for prediction of general phosphorylation sites. Similarly, PhosPred-RF[18] is another phosphorylation site predictor based on random forest.

Recently, deep learning (DL) architectures have been used to predict various PTMs in proteins. Unlike machine learning (ML)-based models, the DL architecture does not require manual feature extraction. For example, MusiteDeep[19] is a DL-based approach that utilizes one-hot encoding and convolutional neural network (CNN) with an attention layer to predict phosphorylation sites. Likewise, DeepPhos[20] uses densely connected CNN (DCCNN) blocks with different filter sizes and windows to learn multiple representations of sequences to predict phosphorylation sites. Furthermore, Wang and Xu devised a capsule network (CapsNet)[21] -based architecture for prediction of protein phosphorylation sites.

Received: October 16, 2022
Published: July 17, 2023



https://doi.org/10.1021/acs.jproteome.2c00667

Furthermore, DeepPSP[22] merges local window (51) and global window (2000) information using squeeze-and-excitation blocks and LSTM blocks to further improve the phosphorylation PTM site prediction. Among the existing approaches, DeepPSP is quite unique in the sense that this approach integrates both local and global information for predicting the phosphorylation sites. Note that these deep learning methods utilize a one-hot encoding scheme for representation of amino acids. Chlamy-EnPhosSite[23] is an organism-specific phosphorylation site predictor for*Chlamydomonas reinhardtii* based on an ensemble approach that combines long short-term memory and convolutional neural network models in which a supervised word embedding scheme is used to encode the protein sequences. Similarly, PhosIDN[24] uses an integrated deep learning architecture to improve phosphorylation prediction by combining local sequence (using one-hot encoding) and protein−protein interaction (from the STRING database)[25] information. Admittedly, considerable progress has been made in the development of general phosphorylation site prediction methods. However, for all current predictors of phosphorylation sites, the input features are either one-hot encoding (each amino acid is represented with a binary vector), embedding encoding (a learned representation where amino acids that are similar have similar representation), or handcrafted features (physiochemical features extracted from protein sequence).

Recently, the transformer-based[26] language model has shown huge potential to unfold the meaningful latent representation of the sentences/language by implementing a multihead self-attention-based mechanism with masking. By considering the protein sequences as sentences, Elnaggar et al. (ProtTrans)[27] developed a pretrained protein language model (pLM), namely, ProtT5-XL-UniRef50 trained on 2.5 billion protein sequences from the UniRef50 database. The representation learned by this model has been used in various prediction tasks, and the results demonstrate that the information on the evolutionary context of a sequence, contact map, taxonomy, long-range dependencies, protein structure, subcellular localization, physicochemical properties, and function is encoded in their distributed representation.[28−34] Thus, a more effective model of phosphorylation site prediction may be established by using the knowledge distilled by this language model.

In this work, we present a novel general phosphorylation site prediction tool named LMPhosSite that integrates per-residue contextualized embedding (aka embeddings that are obtained based on its context) from the pretrained protein language model (ProtT5) with local word embedding. LMPhosSite utilizes stacked generalization to improve the prediction of phosphorylation sites by combining two different models trained using embeddings obtained from the supervised embedding layer and embeddings from the ProtT5 language model. Initially, these two modules are passed to their corresponding deep learning architectures, and finally, the stacked generalization of these two models is performed using a meta-model. When compared with the other existing phosphorylation site prediction tools, LMPhosSite exhibits an improved prediction performance.

■ MATERIALS AND METHODS

Data Set

To train and evaluate our models, we used the data set from DeepPSP.[22] The data set in DeepPSP were experimentally identified phosphorylation sites collected from the SWISSPROT,[35] dbPTM,[36] phosphoELM,[37] and PhosphoSitePLUS.[38] Subsequently, the CD-HIT[39] tool was used to remove the homologous sequences using a cutoff threshold of 0.5. Finally, the sequences were randomly divided into the training and test set in a ratio of 9:1. The annotated phosphorylation sites from these sequences were defined as positive sites, and any remaining S, T, and Y sites that were not annotated as phosphorylated within the same protein sequence were defined as negative sites. To extract the local information, positive windows were generated with the annotated phosphorylated sites in the middle and an equal number of amino acids on both sides flanking the phosphorylated sites. Negative windows were generated in a similar manner. When the site of interest was located near the N- or C-termini of the protein, pseudoresidues "-" were added to make the window sizes to be of the same length. Duplicates were removed from both the positive and negative data sets. Additionally, we also experimented with a CD-HIT cutoff of 0.3 on the homologous sequences and the corresponding data set and the results are presented in the Supporting Information (Tables S5−S9). However, for comparison purposes with the existing approaches, we use the data set obtained using a CD-HIT cutoff of 0.5.

As S and T residues can be phosphorylated by the same specific kinase (hydroxyl groups in their side chains), the S and T data sets were combined. However, as Y residues undergo different enzymatic processes because of pi-electron and the conjugated electrons[40] (phenol ring on the side chain), the Y data set was kept separate. Like many existing approaches, we developed one model for the combined S and T residues and a separate model for the Y residues. Table 1 shows the total number of positive and negative sites in the training and independent test data sets. To avoid overestimation, we made sure that no protein sequences from the independent test data set are present in the training data set as ProtT5 can learn representation for other sites from the same protein. Also, the number of positive ST sites is slightly different than the original DeepPSP data set as some protein sequences have been updated in the UniProt.

Table 1. Positive and Negative Phosphorylation Sites for Training and Independent Testing

data set	residues	number of proteins	positive	negative
training	ST	12,238	165,483	878,133
	Y	8,742	28,965	134,997
test	ST	1,361	18,551	101,944
	Y	968	3,248	14,503

Feature Encoding

In developing a statistical model for the discrimination of protein phosphorylation sites, one pivotal step is numerical encoding through an encoding scheme that assigns a numerical representation to each amino acid that can accurately reflect the intrinsic correlations with the desired targets.[41] In this study, similar in spirit to DeepPSP, we employed both global and local sequence information for the prediction of protein phosphorylation sites. For extracting global information, we used the entire protein sequence that includes the site of interest as input to a pretrained protein language model to extract per-residue contextualized word embedding. Additionally, for encoding the local information, we utilized a supervised embedding layer to obtain word embedding for the window sequence centered

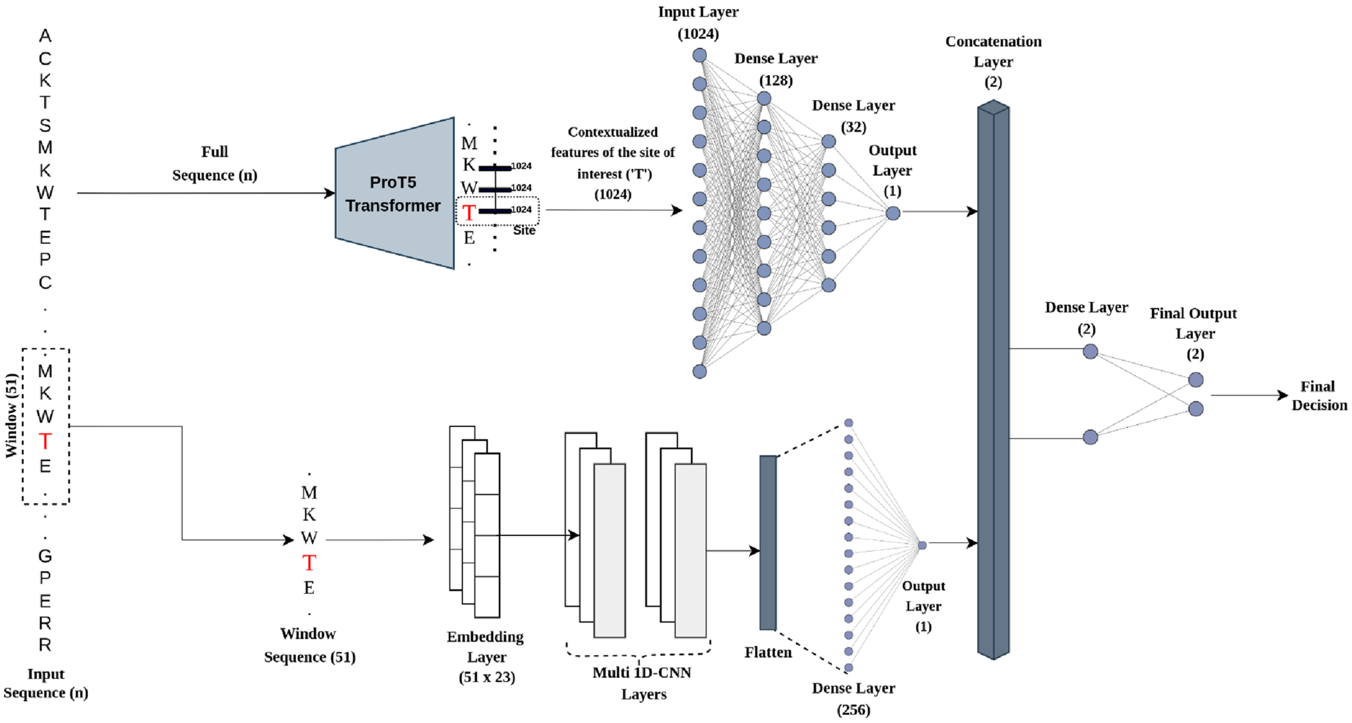


Figure 1. Overall architecture of LMPhosSite: the output scores of two base-models (word embedding (SEL) and ProT5 embedding) are combined using a concatenation layer followed by a basic ANN. The site of interest 'T' is shown in red.

around the site of interest in a supervised manner. We described in detail these two types of embeddings below.

Local Sequence Feature Encoding Using the Supervised Embedding Layer

The post-translational classification approaches such as DeepIPs,[42] LMSuccSite,[43] DeepRMethylSite,[44] and pLMSNOSite[45] showed improved prediction performance using supervised word embedding encoding[46] obtained using a supervised embedding layer (SEL). The SEL learns the representation from integer-encoded sequences through back-propagation in a supervised manner. We used the SEL to encode the local window sequence. First, the 20 canonical amino acids and one pseudoresidue "-" were converted into specific integers ranging from 0 to 20. These make the inputs for the embedding layer that lies at the beginning of our DL architecture. Initially, the weights in the embedding layer were randomly initialized, and these weights were updated during the training process. The key arguments in the embedding layer are vocabulary size, output_dim (size of vector space), and input_length (size of input windows). The output from the embedding layer has the dimension of input_length × output_dim. Based on the 10-fold cross-validation results, for our case, we used a vocabulary size of 23, output_dim of 21, and size of input windows to be 51. This embedding was fed into one of the base-models.

Contextualized Per-Residue Embedding from ProtT5

Recently, embeddings from protein language models (pLMs) were utilized to predict binding residues,[47] signal peptides,[48] subcellular localization,[49] and protein structural features.[50] In this regard, here, we employ a protein language model (pLM) called ProtT5-XL-UniRef50 (ProtT5)[27] to obtain contextualized per-residue embedding for the site of interest. ProtT5 is based on the T5 architecture[51] and trained solely on unlabeled protein sequences from BFD (Big Fantastic Database; 2.5 billion sequences including meta-genomic sequences)[52] and UniRef50.[53] To obtain the contextualized per-residue embedding for the site of interest from ProtT5, the overall protein sequence containing the site of interest was fed into a pretrained ProtT5[27] model and the fixed-length per-residue features were extracted from the last encoder layer. This contextualized (1024 length feature vector) embedding of the site of interest produced by ProtT5 was then fed into the other base-model.

Architecture of LMPhosSite

LMPhosSite integrates two base-models: a model for learning encoding from local sequence information obtained from the window sequence and another model for learning contextualized embedding from the global sequence information obtained using the overall protein sequence, using stacked generalization. Initially, these base-models are trained independently. Finally, score-level fusion of these two models is performed by using a meta-classifier based on a fully connected network. The proposed architecture of LMPhosSite is shown in Figure 1.

Model for Supervised Word Embedding

The input to this model is the (local) window sequence with the site of interest in the middle. For the embedding layer, a window sequence of size 51 (as shown in Figure 2) was taken around the site of interest, and features were extracted subsequently for classification based on this window. The architecture of this model starts with a supervised embedding layer followed by two layers of a one-dimensional convolution neural network (1D-CNN) and a fully connected layer. The first and second layer 1D-CNN were configured to 128 and 64 filters, respectively, and the size of the filter was set to 3. Additionally, the 1D-CNN layer uses the ReLU activation function and a dropout probability of 0.3. All generated feature maps from 1D-CNN were then sent into the MaxPooling layer with a pool size of 2. The MaxPooling layer samples the output feature map of the convolution layers and extracts the most obvious features. Subsequently, the output feature maps from 1D-CNN were flattened and sent to a fully

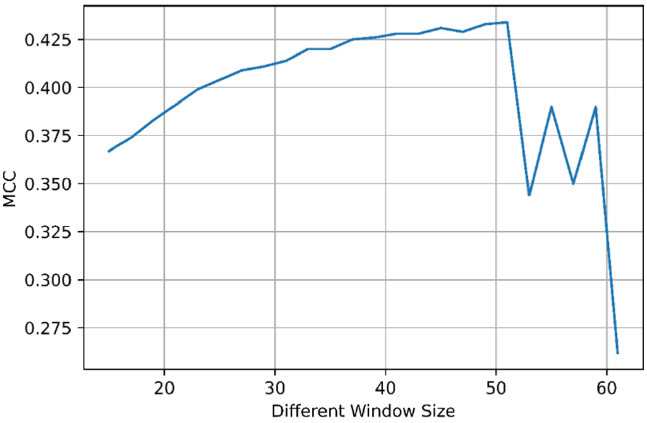


Figure 2. Ten-fold cross-validation mean MCC of the combined ST phosphorylation data set for different window sizes.

connected layer with 256 neurons that has a ReLU activation function and a dropout probability rate of 0.4. The output of this model is a probability (of being phosphorylated) score between 0 and 1.

Note here that the architecture and the associated hyperparameters for this base-model were obtained using 10-fold cross-validation with grid-search. Essentially, several DL architectures such as 1D-CNN, 2D-CNN, LSTM, and BiLSTM were implemented for this model, and 1D-CNN was selected based on the 10-fold cross-validation results.

Model for Contextualized Embedding Obtained from ProtT5

The input to this model is the contextualized embedding of the site of interest (size = 1024) obtained from ProtT5 extracted using the overall sequence. Initially, we extracted a static embedding of size $n \times 1024$, where n is the size of the overall sequence, from the last hidden layer of the ProtT5 architecture. However, we only used embedding of the "site of interest" for our purpose. For instance, given a protein sequence of length of 100 and the site of interest located at the 25th position in the sequence, the ProtT5 will output a contextualized feature vector of size of 1 × 1024 for every 100 amino acid positions (totaling 100 × 1024 size feature matrix). We only utilized the feature vector (dimension: 1 × 1024) corresponding to the site of interest, i.e., 25th amino acid. In our case, the sites of interest are S, T, and Y where they can either belong to the positive set or a negative set. We presented a schematic representation of the overall architecture and how features are extracted for the site of interest using ProtT5 in the Figure 1. The input sequence of length n with the site of interest ("T", denoted in red) was fed to ProtT5. From these embeddings, we only used the 1024 length feature vector of the site of interest ("T") for classification purposes.

The architecture of this model consists of an ANN with two hidden layers and one output layer. The first hidden layer has 128 neurons, and the second hidden layer has 32 neurons. The output of this model is also a probability (of being phosphorylated) score between 0 and 1. Similarly, various architectures such as RF, SVM, XGBoost, logistic regression, and ANN were implemented for this model, and an ANN-based architecture was selected based on 10-fold cross-validation results.

Meta-Classifier for Score-Level Fusion

Next, stacked generalization of the two base-models was performed to obtain a meta-classifier using score-level fusion. The input to this classifier is the two probability outputs from the base-models: the supervised embedding layer model and the ProtT5-based model. The architecture for this model used a basic ANN with one hidden layer having two neurons. All the hyperparameters of individual base-models and the meta classifier used in LMPhosSite are presented Table 2.

Table 2. Hyperparameters of the Proposed Deep Architecture

		hyperparameters				
model	layer	activation function	size[b]	filters	drop-out	MaxPool 1D (pool)
ProtT5	dense[a]	ReLU	128		0.4	
		ReLU	32		0.4	
		sigmoid	1			
embedding encoding	1D CNN	ReLU	3	128	0.3	2
		ReLU	3	64	0.3	2
	flatten					
	dense	ReLU	256		0.4	
		sigmoid	1			
stacked generalization	dense	ReLU	2			
		SoftMax	2			

[a] Dense layers represent the fully connected layers in TensorFlow. [b] The size of convolution layers means the kernel sizes, and the size of dense layers denotes the number of neurons in hidden states.

Note here that the architectures and the associated hyperparameters for the meta-classifier were also obtained using 10-fold cross-validation with grid-search.

Model Evaluation and Performance Metrics

In this study, phosphorylated sites are considered as positives sites, and nonphosphorylated sites are considered as negatives sites. Additionally, phosphorylated sites and nonphosphorylated sites predicted correctly by the model are true positive (TP) and true negative (TN), respectively. The negative sites misclassified as positive sites are false positive (FP) and positive sites misclassified as negative sites are false negative (FN). Ten-fold cross-validation was utilized to evaluate the performance of the models. For the results of 10-fold cross-validation, unless otherwise noted, all performance metrics are reported as the mean value ± one standard deviation. Four metrics including precision (Pre), recall (Rec), F1-score, and Matthew's correlation coefficient (MCC) were used to evaluate the performance of the models (with a probability decision threshold = 0.5). Furthermore, the area under the receiver operating characteristics (ROC) curve and area under the precision-recall curve (PrAUC) were also used as performance metrics. Equation 1 describes precision, recall, F1-score, and MCC where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F1} = 2 \times \frac{\text{PRE} \times \text{RE}}{\text{PRE} + \text{RE}} \tag{3}$$

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

RESULTS

As discussed in the Materials and Methods section, LMPhosSite integrates two types of embeddings using stacked generalization by performing score-level fusion. Using 10-fold cross-validation, we first find the best architectures and hyperparameters for the individual models and subsequently find the best hyperparameters for the meta-classifier using the 10-fold cross-validation strategy.

Essentially, we analyzed the comparative performance of various architectures for base-models and the meta-classifier using 10-fold cross-validation techniques. Finally, we compare the performance of LMPhosSite against existing phosphorylation tools using independent testing. Below, we discuss the results in detail.

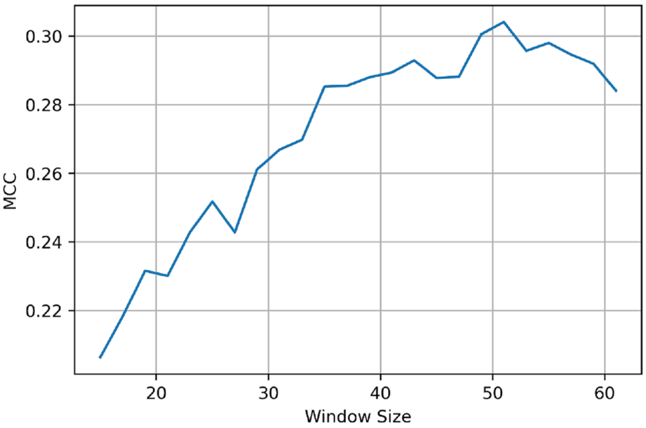


Figure 3. Ten-fold cross-validation mean MCC of the Y phosphorylation data set for different window sizes.

Window Size Selection and Embedding Dimension Selection for Local Sequence Feature

The recent studies[54−56] unravel that neighboring residues can influence the phosphorylation status of the site of interest (in this case serine, threonine, and tyrosine residues). In that regard, to capture local sequence information for the site of interest, a local window sequence centered around the site of interest (S, T, or Y) surrounded by an equal number of flanking residues on both sides is generally taken as input. However, the window size is also a parameter. To determine the window size, we performed a 10-fold cross-validation for the supervised embedding layer model. Essentially, we experimented with various window sizes ranging from 15 to 61 in an increment of 2 for combined ST and Y phosphorylation training data sets (Figure 2) and 3 showed the mean Matthew's correlation coefficient (MCC) produced using different window sizes on the combined ST and Y phosphorylation training data set using 10-fold cross-validation. The detailed results of the analysis are presented in Tables S1 and S2. Further window sizes were not

Table 4. Performance Metrics for Two Base-Models (Em and ProtT5) and Meta-Model (PrT5 + Em) on the Combined ST Phosphorylation Independent Test Set

encoding scheme	MCC	precision	recall	F1-score
ProtT5 (PrT5)	0.3693	0.3733	0.6538	0.4752
embedding (Em)	0.3496	0.3377	0.7029	0.4562
PrT5 + Em (LMPhosSite)	0.3905	0.3878	0.6712	0.4915

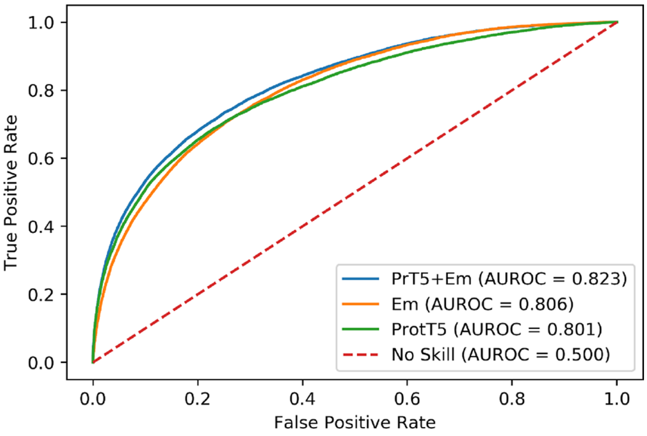


Figure 4. ROC curves of LMPhosSite (PrT5 + Em) and base-models on the combined ST independent test data set. For each model, the area under the ROC curve is also reported.

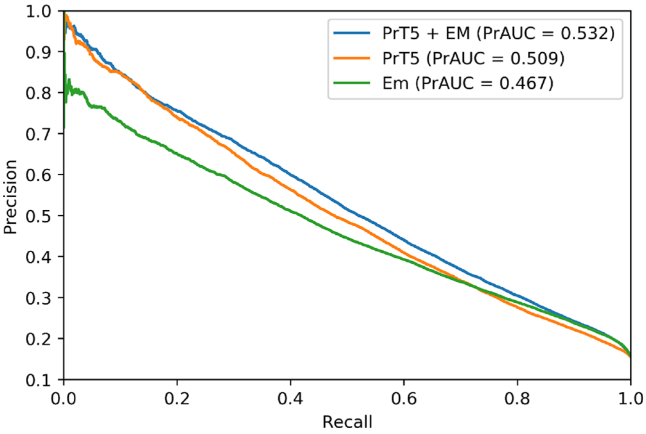


Figure 5. Precision-recall curves of LMPhosSite (PrT5 + Em) and base-models on the combined ST independent test data set. For each model, the area under PrAUC is also reported.

Table 3. Performance Metrics for Two Base-Models (Embedding (Em) and ProtT5 (PrT5)) and Meta-Model (PrT5 + Em) Using 10-Fold Cross-Validation on the Combined ST Phosphorylation Training Data Set[a]

encoding scheme	MCC	precision	recall	F1-score
ProtT5 (PrT5)	0.461 ± 0.005	0.746 ± 0.008	0.697 ± 0.014	0.720 ± 0.004
embedding (Em)	0.434 ± 0.004	0.715 ± 0.021	0.721 ± 0.046	0.717 ± 0.012
PrT5 + Em (LMPhosSite)	0.502 ± 0.004	0.766 ± 0.006	0.721 ± 0.007	0.743 ± 0.002

[a] ± refers to standard deviation.

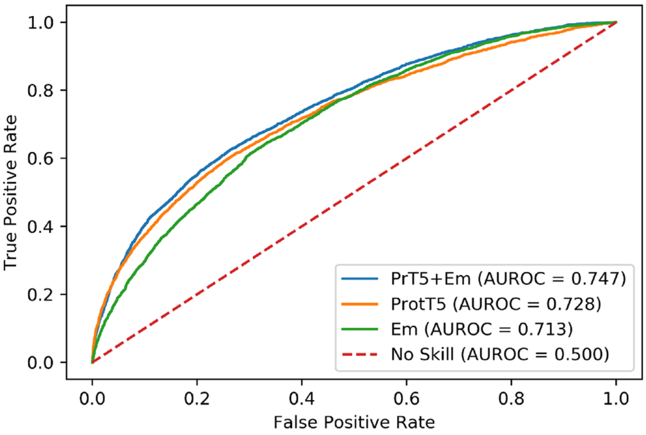


Figure 6. ROC curves of LMPhosSite (PrT5 + Em) and base-models on the Y independent test data set. For each model, the area under the ROC curve is reported.

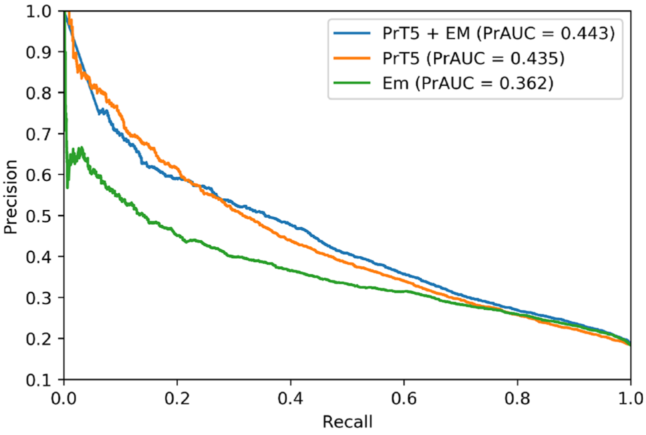


Figure 7. Precision-recall curves of LMPhosSite (PrT5 + Em) and base-models on the Y independent test data set. For each model, the area under the PrAUC is also reported.

Table 5. Prediction Performance of LMPhosSite Compared to Other Existing Predictors on the Combined ST and Y Phosphorylation Independent Data Set[a]

residues	predictors	MCC	precision	recall	F1-score
combined ST	LMPhosSite	**0.3905**	**0.3878**	0.6712	**0.4915**
	DeepPSP	0.3790	0.3741	0.6769	0.4819
	MusiteDeep	0.3342	0.3241	0.7041	0.4439
	Musite	0.2006	0.2237	0.7552	0.3451
	CapsNet	0.2743	0.2405	**0.8855**	0.3783
Y	LMPhosSite	**0.2984**	**0.3490**	0.6203	**0.4467**
	DeepPSP	0.2605	0.3095	0.6561	0.4206
	MusiteDeep	0.2029	0.3488	0.3485	0.3487
	Musite	0.1400	0.2353	0.6786	0.3494
	CapsNet	0.1954	0.2329	**0.8861**	0.3688

[a]Highest values in each column (and category, ST/Y) are highlighted in bold.

analyzed due to the sheer size of the windows and the corresponding increase in the number of pseudoresidues "-" that were required at higher window sizes (for residues near N- and C-termini).

From Figure 2, it can be observed that the best performing window size was 51 (highest MCC, 0.434 for ST and highest MCC, 0.3041 for Y) for the combined ST and Y phosphorylation training data set. Hence, 51 was selected as the value of window size for the combined ST and Y phosphorylation residue for subsequent analysis. Interestingly, the DeepPSP[22] phosphorylation prediction method also uses the same window size (= 51).

Subsequently, the model architectures for the two base-models and the meta-model are selected based on 10-fold cross-validation results.

10-Fold Cross-Validation Results of Base-Models and Meta-Classifier

Here, we discuss the comparative performance of the base-models and the meta-classifier using 10-fold cross-validation on the ST phosphorylation training data set. The results are presented in Table 3. The performance metrics used are precision, recall, MCC, and F1-score.

As mentioned above, the training set consists of 165,483 phosphorylated sites and 165,483 (under sampled from 878,133) nonphosphorylated sites. Consequently, the 10-fold cross-validation was performed on the combined ST phosphorylation training data set (330,966 training examples) to choose the best performing meta-classifier (PrT5 + Em). We also performed experiments using cost-sensitive learning by using all the negative sets, and the result of this analysis is shown in Table S4. The mean MCC of the supervised word embedding base-model (Em) is 0.434 ± 0.004, and the mean MCC of the ProtT5 base-model (ProtT5) is 0.461 ± 0.005. Additionally, it can be observed from Table 3 that the meta-classifier produces a mean MCC, mean precision, mean recall, and mean F1-score of 0.502 ± 0.004, 0.766 ± 0.006, 0.721 ± 0.007, and 0.743 ± 0.002, respectively. Thus, the MCC of the meta-classifier is higher than the MCC of both base-models for the prediction of phosphorylation site prediction, indicating that meta-classifier performs better than the base-models for phosphorylation prediction. Hence, we select this meta-model that combines the two base-models as the final model for the prediction of phosphorylation sites and we call this model as LMPhosSite. As described earlier, the architecture of the meta-classifier is a basic ANN with one hidden layer having two neurons. In conclusion, the meta-model (PrT5 + Em) that combines the local sequence information from the supervised embedding layer (embedding (Em)) and the global information from the ProtT5 model (ProtT5 (PrT5)) produces the best MCC. Note that the architecture for the base-models was independently optimized using 10-fold cross-validation.

Independent Test Results

Subsequently, we assessed the performance of LMPhosSite and the base-models on the independent test set for ST and Y sites. The results of LMPhosSite on independent ST sites are presented in Table 4. As seen from the table, LMPhosSite produces a precision, recall, Matthew's correlation coefficient (MCC), and F1-score of 38.78%, 67.12%, 0.390, and 49.15%, respectively. The LMPhosSite was able to classify 82,275 samples as true negative, 12,452 samples as true positive, 19,650 samples as false positive, and 6,099 as false negative for the ST independent data set. Similarly, when we tested the LMPhosSite on Y phosphorylation sites and tested on the Y phosphorylation test data set, it achieved a precision, recall, Matthew's correlation coefficient (MCC), and F1-score of 34.90%, 62.03%, 0.298, and 44.67%, respectively. The LMPhosSite was able to classify 10,735 samples as true negative, 2,010 samples as true positive, 3,748 samples as false positive, and 1,230 as false negative for the

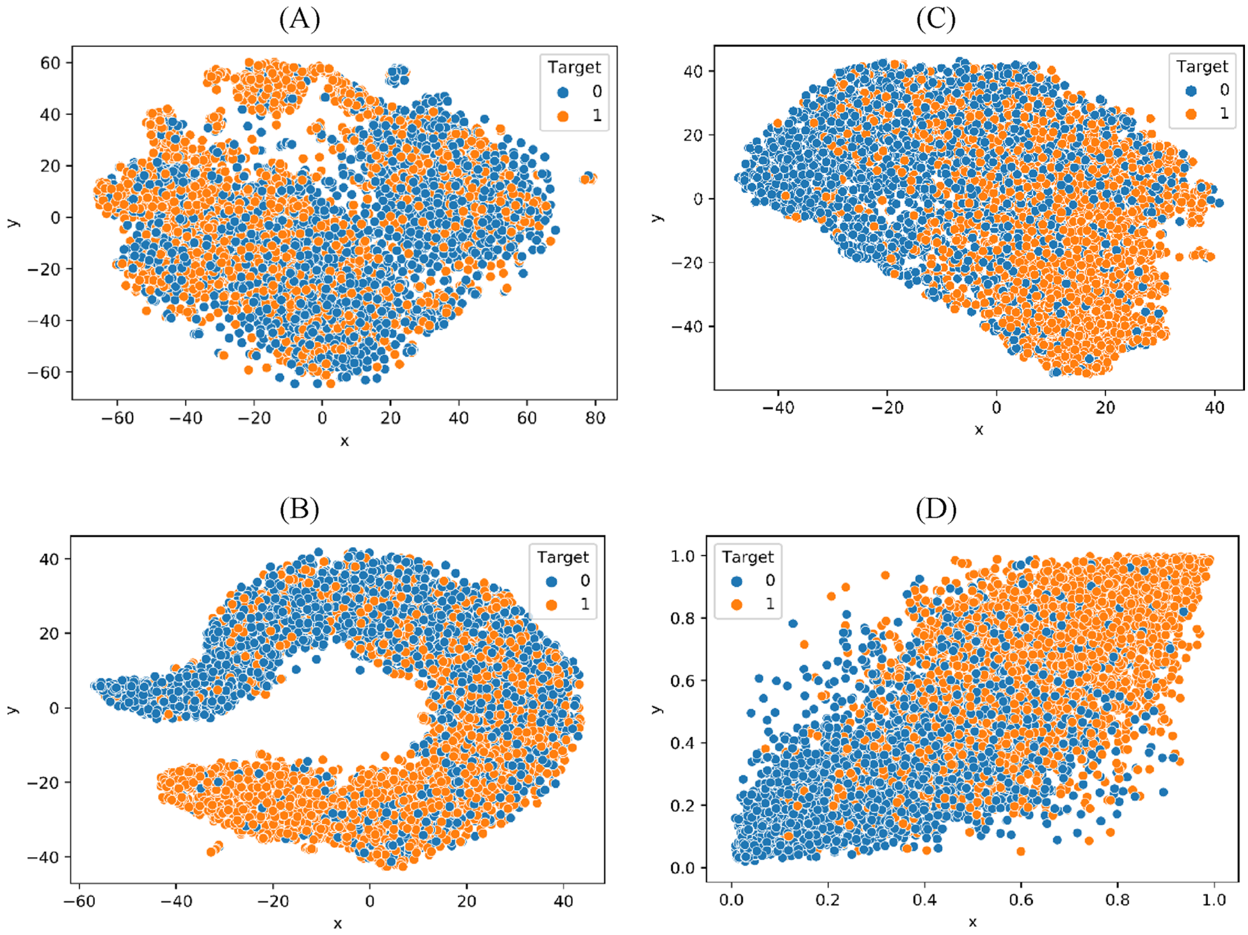


Figure 8. t-SNE illustration of the learned features. (A) Features extracted from the ProtT5 language model, (B) learned features from the ProtT5 base-model, (C) learned features from the embedding base-model, and (D) learned features from stacked generalization (LMPhosSite).

Y independent data set. The independent test results for base-models are shown for informative purposes only, and the final model (meta-classifier) is selected based on 10-fold cross-validation.

Figure 4 shows the receiver operating characteristic (ROC) curve for base-models and LMPhosSite on the combined ST independent test data set. It can be observed that LMPhosSite (PrT5 + Em) has the highest area under the curve (0.823) compared to the base-models. Additionally, we also plotted the precision-recall curve for the base-models and LMPhosSite on the combined ST independent test data set, and the results are shown in Figure 5. It can be observed that LMPhosSite also has the highest precision-recall area under the curve (PrAUC = 0.532).

Moreover, we also plotted ROC and precision-recall curve of the phosphorylation Y independent data set, and the corresponding ROC and precision-recall curves are shown in Figures 6 and 7, respectively. The AUC and PrAUC of LMPhosSite are the highest compared to the base-models for the Y phosphorylation independent data set too.

Comparison with Other Widely Available General Phosphorylation Predictors Based on Independent Testing

We further compared the performance of LMPhosSite with some publicly available predictors including DeepPSP,[22] MusiteDeep,[19] Musite,[57] and CapsNet,[21] and the results are shown in Table 5. DeepPSP, MusiteDeep, and CapsNet utilize state-of-the-art deep-learning algorithms, whereas Musite utilizes a machine-learning algorithm. The results for DeepPSP, MusiteDeep, Musite, and CapsNet were adapted from DeepPSP, and we use the same training and test data set as DeepPSP. It can be observed from Table 5 that LMPhosSite trained on a combined ST phosphorylation data set produced an MCC, precision, recall, and F1-score of 0.3905, 38.78%, 67.12% and 49.15%, respectively.Moreover, LMPhosSite trained on the Y phosphorylation data set produced an MCC, precision, recall, and F1-score of 0.2984, 34.90%, 62.03%, and 44.66%, respectively. LMPhosSite produces the highest MCC (0.3905 for ST and 0.2984 for Y) among the compared approaches for both ST and Y phosphorylation independent testing. Moreover, the confusion matrices of all the predictors are listed in Table S3.

These results demonstrate that LMPhosSite performs better than the compared approaches across all these performance metrics except for recall for combined ST and produces the best results across all these performance metrics except for recall for the Y phosphorylation data set.

t-SNE Plot

Additionally, to discern the classification effectiveness of the features from the base-models as well as the features from the stacked generalization model, we used t-distributed stochastic neighbor embedding (t-SNE)[58] to project these features into the two-dimensional space (Figure 8). The main purpose of t-SNE visualization in this work is to visually observe the separation boundaries between classes in two-dimensional space. The t-SNE plot was generated using 50 as the value for the learning

https://doi.org/10.1021/acs.jproteome.2c00667
J. Proteome Res. 2023, 22, 2548−2557

rate to visualize the learned ProtT5 features and features obtained from the penultimate dense layer of the trained model. Note that these plots were generated using 8,274 (4,137 positive and 4,137 negative) randomly chosen data points.

For the features extracted from the language model ProtT5 of phosphorylated or nonphosphorylated token “S/T”, there are some clusters for positive and negative samples; overall, the positive and negative samples are still not distinct (Figure 8A). Figure 8D represents the t-SNE plot of the feature vectors generated from the concatenation layer of the trained deep-learning architecture, which shows that negative samples (blue points) are concentrated at the third quadrant, while positive samples (orange points) are concentrated at the first quadrant, which indicates that the ensembled DL architecture that processes features from two information sources with MLP as a meta-classifier largely clusters positive and negative samples in $\mathbb{R}^2$ space. For further elaboration, we also provide the t-SNE plot generated from the penultimate layer of trained base-models ProtT5 and embedding encoding in Figure 8B and Figure 8C, respectively.

CONCLUSIONS AND DISCUSSION

Protein phosphorylation is one of the most important post-translational modifications. In this work, we developed a stacked generalization approach called LMPhosSite to predict protein phosphorylation sites. LMPhosSite uses a supervised embedding layer to encode local sequence information and a pretrained protein language model (ProtT5) to encode global sequence information and then integrates these two types of information using a stacked generalization approach. The novelty of the approach is the use of global information extracted from contextualized embedding obtained from a pretrained protein language model (ProtT5) for the site in conjunction with supervised word embedding obtained from the local sequence window. The independent test results show that LMPhosSite achieves better performance than the compared approaches. Hence, it can be concluded that LMPhosSite is a promising general phosphorylation site prediction tool.

The improved performance of the LMPhosSite can likely be attributed to two things: the use of contextual protein language models to extract features from the overall protein sequence for the site of interest and the novel architecture that uses stacked generalization. To our knowledge, LMPhosSite is the first approach that utilizes the distilled information from large pretrained protein language models for the prediction of phoshorylation sites. Similar in spirit to DeepPSP, our model combines the global information (the ProtT5-based embeddings obtained for the site from the whole protein sequence) with the local information (the supervised word embedding obtained from the window sequence using the supervised embedding layer). Since these pLM embeddings can be easily extracted for any protein sequence, LMPhosSite provides robust and fast predictions for phosphorylation sites.

The improvement in the performance of these approaches that utilize language models may be improved by (i) fine-tuning a pretrained protein language model using the proteins of the same characteristics (in this case, phosphorylated proteins), (ii) training newer protein language models using newer language models like GPT-4, and (iii) combining language model-based features with other physio-chemical features. Finally, with the development of newer and more powerful protein language models, the prediction performance of the approaches that make use of distilled information from these language models is likely to improve.

ASSOCIATED CONTENT

Data Availability Statement

LMPhosSite is provided as an open-source tool and is available at https://github.com/KCLabMTU/LMPhosSite and the web server is available at http://kcdukkalab.org/LMPhosSite/.

Supporting Information

The following files are available free of charge at https://github.com/KCLabMTU/LMPhosSite

Table S1. Ten-fold cross-validation results for different windows for combined ST phosphorylation sites. Table S2. Ten-fold cross-validation results for different windows for Y phosphorylation sites. Table S3. Confusion matrices of LMPhosSite when predicting general phosphorylation sites. Table S4. Performance metrics from cost-sensitive imbalance learning on the combined ST phosphorylation independent test set compared with random under-sampling. Table S5. Positive and negative phosphorylation sites for training and independent testing with a CD-HIT cutoff of 0.3. Table S6. Ten-fold cross-validation results for different windows for S/T phosphorylation sites (CD-HIT cutoff of 0.3). Table S7. Ten-fold cross-validation results for different windows for Y phosphorylation sites (CD-HIT cutoff of 0.3) with the Em base-model. Table S8. Performance metrics for two base-models (embedding (Em) and ProtT5 (PrtT5)) and meta-model (PrT5 + Em) using 10-fold cross-validation with a CD-HIT cutoff of 0.3 on the combined ST phosphorylation training data set. Table S9. Performance metrics for two base-models (Em and ProtT5) and meta-model (PrT5 + Em) on the combined ST phosphorylation independent test set with a CD-HIT cutoff of 0.3 (PDF)

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.2c00667.

AUTHOR INFORMATION

Corresponding Author

Dukka B. KC − *Department of Computer Science, Michigan Technological University, Houghton, Michigan 49931, United States;* orcid.org/0000-0001-7443-1928; Email: dbkc@mtu.edu

Authors

Subash C. Pakhrin − *School of Computing, Wichita State University, Wichita, Kansas 67260, United States; Department of Computer Science & Engineering Technology, University of Houston-Downtown, Houston, Texas 77002, United States*

Suresh Pokharel − *Department of Computer Science, Michigan Technological University, Houghton, Michigan 49931, United States*

Pawel Pratyush − *Department of Computer Science, Michigan Technological University, Houghton, Michigan 49931, United States;* orcid.org/0000-0002-4210-1200

Meenal Chaudhari − *Department of Biology, North Carolina A&T State University, Greensboro, North Carolina 27411, United States*

Hamid D. Ismail – *Department of Computer Science, Michigan Technological University, Houghton, Michigan 49931, United States*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jproteome.2c00667

Author Contributions

D.B.K. conceived and designed the experiments; S.C.P. performed all the experiments and data analysis, S.P., M.C., H.D.I., and P.P. checked all the programs and models. H.D.I. developed the Web server. S.C.P., S.P., P.P., D.B.K., M.C., and H.D.I. wrote and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by National Science Foundation (NSF) grant nos. 1901793 and 1564606 (to D.B.K.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge the use of the BeoShock High-Performance Computing resources located at Wichita State University and the DeepPSP data set made freely available for the researchers.

■ ABBREVIATIONS

1D-CNN, one-dimensional convolution neural network; ANN, artificial neural network; BFD, Big Fantastic Database; BiLSTM, bidirectional long short-term memory; CNN, convolutional neural network; DCCNN, densely connected CNN; DL, deep learning; Em, embedding; H, histidine; LSTM, long short-term memory; MCC, Matthew's correlation coefficient; ML, machine learning; MLP, multi-layer perceptron; pLMs, protein language models; PrAUC, area under precision-recall curves; PTMs, post-translational modifications; ReLU, rectified linear activation unit; RF, random forest; ROC, receiver operating characteristics; S, serine; SEL, supervised embedding layer; SVM, support vector machine; T, threonine; t-SNE, t-distributed stochastic neighbor embedding; Y, tyrosine.

■ REFERENCES

(1) Forrest, A. R.; Taylor, D. F.; Fink, J. L.; Gongora, M. M.; Flegg, C.; Teasdale, R. D.; Suzuki, H.; Kanamori, M.; Kai, C.; Hayashizaki, Y.; et al. Phosphoreg DB: the tissue and sub-cellular distribution of mammalian protein kinases and phosphatases. *BMC Bioinformatics* **2006**, 7 (82), 82 DOI: 10.1186/1471-2105-7-82.

(2) Karve, T. M.; Cheema, A. K. Small changes huge impact: the role of protein post translational modifications in cellular homeostasis and disease. *J. Amino Acids* **2011**, *2011*, 207691 DOI: 10.4061/2011/207691.

(3) Gong, W.; Zhou, D.; Ren, Y.; Wang, Y.; Zuo, Z.; Shen, Y.; Xiao, F.; Zhu, Q.; Hong, A.; Zhou, X.; et al. PepCyber:P ~PEP: a database of human protein protein interactions mediated byphosphoprotein-binding domains. *Nucleic AcidsRes.* **2008**, *36* (Database issue), D679–683, DOI: 10.1093/nar/gkm854.

(4) Li, T.; Li, F.; Zhang, X. Prediction of kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach. *Proteins* **2008**, *70*, 404–414, DOI: 10.1002/prot.21563.

(5) Matthews, R. Protein kinases and phosphatasesthat act on histidine, lysine, or arginine residues in eukaryoticproteins: a possible regulator of the mitogen-activated protein kinasecascade. *Pharmacol. Ther.* **1995**, *67*, 323–350, DOI: 10.1016/0163-7258(95)00020-8.

(6) Trost, B.; Kusalik, A. Computational prediction of eukaryotic phosphorylationsites. *Bioinformatics* **2011**, *27* (21), 2927–2935.

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

(8) Temporini, C.; Calleri, E.; Massolini, G.; Caccialanza, G. Integratedanalytical strategies for the study of phosphorylation and glycosylationin proteins. *Mass Spectrom Rev.* **2008**, *27* (3), 207–236.

(9) Nsiah-Sefaa, A.; McKenzie, M. Combined defects inoxidative phosphorylation and fatty acid beta-oxidation in mitochondrial disease. *Bioscience Reports* **2016**, *36* (2), No. e00313.

(10) Cohen, P. The origins of protein phosphorylation. *Nat. Cell Biol.* **2002**, *4*, E127–E130, DOI: 10.1038/ncb0502-e127.

(11) Panni, S. Phospho-peptide binding domains in S. cerevisiae model organism. *Biochimie* **2019**, *163*, 117–127.

(12) Aponte, A. M.; D. P; Harris, R. A.; Blinova, K.; French, S.; Johnson, D. T.; Balaban, R. S. 32P labeling of protein phosphorylation and metabolite association in the mitochondriamatrix. *Methods Enzymol* **2009**, *457*, 63–80, DOI: 10.1016/S0076-6879(09)05004-6.

(13) Beausoleil, S. A.; J. V. e; Gerber, S. A.; Rush, J.; Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **2006**, *24*, 1285 DOI: 10.1038/nbt1240.

(14) Rohira, A. D.; Chen, C. Y.; Allen, J. R.; Johnson, D. L. Covalent small ubiquitin-likemodifier (SUMO) modification of Maf 1 protein controls RNA polymeraseIII-dependent transcription repression. *J. Biol.Chem.* **2013**, *288*, 19288–19295.

(15) Agarwal, K. L.; Kenner, G. W.; Sheppard, R. C. Feline gastrin. An example of peptidesequence analysis by mass spectrometry. *J. Am.Chem. Soc.* **1969**, *91*, 3096–3097, DOI: 10.1021/ja01039a051.

(16) Blom, N.; Gammeltoft, S.; Brunak, S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* **1999**, *294* (5), 1351–1362.

(17) Ismail, H. D.; Jones, A.; Kim, J. H.; Newman, R. H.; Kc, D. B. RF-Phos: A Novel General Phosphorylation Site Prediction Tool Based on Random Forest. *Biomed Res. Int.* **2016**, *2016*, 3281590 DOI: 10.1155/2016/3281590.

(18) Wei, L.; P. X; Tang, J.; Zou, Q. PhosPred-RF: a novel sequence-based predictor for phosphorylation sites using sequential information only. *IEEETrans. NanoBioscience* **2017**, *16*, 240–247, DOI: 10.1109/TNB.2017.2661756.

(19) Wang, D.; Zeng, S.; Xu, C.; Qiu, W.; Liang, Y.; Joshi, T.; Xu, D. MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics* **2017**, *33* (24), 3909–3916.

(20) Luo, F.; Wang, M.; Liu, Y.; Zhao, X. M.; Li, A. DeepPhos: prediction of protein phosphorylation sites with deep learning. *Bioinformatics* **2019**, *35* (16), 2766–2773.

(21) Wang, D. L. Y.; Xu, D. Capsule network for protein post translational modification site prediction. *Bioinformatics* **2019**, *35*, 2386–2394, DOI: 10.1093/bioinformatics/bty977.

(22) Guo, L.; Wang, Y.; Xu, X.; Cheng, K. K.; Long, Y.; Xu, J.; Li, S.; Dong, J. DeepPSP: A Global-Local Information-Based Deep Neural Network for the Prediction of Protein Phosphorylation Sites. *J. Proteome Res.* **2021**, *20* (1), 346–356.

(23) Thapa, N.; Chaudhari, M.; Iannetta, A. A.; White, C.; Roy, K.; Newman, R. H.; Hicks, L. M.; Kc, D. B. A deep learning based approach for prediction of Chlamydomonas reinhardtii phosphorylation sites. *Sci. Rep* **2021**, *11* (1), 12550 DOI: 10.1038/s41598-021-91840-w.

(24) Yang, H.; Wang, M.; Liu, X.; Zhao, X. M.; Li, A. PhosIDN: anintegrated deep neural network for improving protein phosphorylation site prediction by combining sequence and protein-protein interactioninformation. *Bioinformatics* **2021**, *37* (24), 4668–4676.

(25) Szklarczyk, D.; Franceschini, A.; Kuhn, M.; Simonovic, M.; Roth, A.; Minguez, P.; Doerks, T.; Stark, M.; Muller, J.; Bork, P.; Jensen, L. J. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **2010**, *39*, D561–D568, DOI: 10.1093/nar/gkq973.

(26) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. AttentionIs All You Need. *31st Conference on NeuralInformation Processing Systems (NIPS 2017)* 2017.

(27) 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* **2022**, *44* (10), 7112−7127.

(28) 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.

(29) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R.; et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118* (15), No. e2016239118.

(30) Pakhrin, S. C.; Aoki-Kinoshita, K. F.; Caragea, D.; Kc, D. B. DeepNGlyPred: A Deep Neural Network-Based Approach for Human N-Linked Glycosylation Site Prediction. *Molecules* **2021**, *26* (23), 7314.

(31) Pakhrin, S. C.; Pokharel, S.; Saigo, H.; Kc, D. B. Deep Learning-Based Advances In Protein Post translational Modification Site and Protein Cleavage Prediction. *MethodsMol. Biol.* **2022**, *2499*, 285−322.

(32) Pakhrin, S. C.; Shrestha, B.; Adhikari, B.; Kc, D. B. Deep Learning-Based Advances in Protein Structure Prediction. *Int. J. Mol. Sci.* **2021**, *22* (11), 5553 DOI: 10.3390/ijms22115553.

(33) Pakhrin, S. C. *Deep learning-based approaches for prediction of post-translational modification sites in proteins*; Wichita State University, 2022.

(34) Pakhrin, S. C.; Pant, D. R. Multi-Armed Bandit Learning Approach with Entropy Measures for Effective Heterogeneous Networks Handover Scheme. In *2018 International Conference on Advances in Computing, Communication Control and Networking(I-CACCCN)*; Greater Noida, India, 2018.

(35) Apweiler, R. B. A.; Martin, M. J.; et al. Activities at the universal protein resource (UniProt). *Nucleic Acids Res.* **2014**, *42*, D191−D198, DOI: 10.1093/nar/gkt1140.

(36) Lu, C. T.; Huang, K. Y.; Su, M. G.; Lee, T. Y.; Bretana, N. A.; Chang, W. C.; Chen, Y. J.; Chen, Y. J.; Huang, H. D. DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. *Nucleic Acids Res.* **2013**, *41* (Database issue), D295−305, DOI: 10.1093/nar/gks1229.

(37) Diella, F.; Cameron, S.; Gemund, C.; Linding, R.; Via, A.; Kuster, B.; Sicheritz-Ponten, T.; Blom, N.; Gibson, T. J. Phospho.ELM:a database of experimentally verified phosphorylation sites in eukaryoticproteins. *BMC Bioinformatics* **2004**, 79 DOI: 10.1186/1471-2105-5-79.

(38) Hornbeck, P. V.; J. M; Tkachev, S.; Zhang, B.; Skrzypek, E.; Murray, B.; Latham, V.; Sullivan, M. PhosphoSitePlus: A comprehensive resource for investigating the structure and function of experimentally determined post-translational odifications in man and mouse. *Nucleic Acids Res.* **2012**, *40*, D261−D270, DOI: 10.1093/nar/gkr1122.

(39) Huang, Y.; Niu, B.; Gao, Y.; Fu, L.; Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **2010**, *26* (5), 680−682.

(40) Gutzler, R.; Perepichka, D. F. pi-Electron conjugation in two dimensions. *J. Am. Chem. Soc.* **2013**, *135* (44), 16585−16594.

(41) Xu, Y.; Ding, Y.; Ding, J.; Lei, Y.-H.; Wu, L.-Y.; Deng, N.-Y. iSuc-PseAAC: predicting lysine succinylationin proteins by incorporating peptide position-specific propensity. *Sci. Rep.* **2015**, *5*, 10184 DOI: 10.1038/srep10184.

(42) Lv, H.; Dao, F. Y.; 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), bbab244 DOI: 10.1093/bib/bbab244.

(43) Pokharel, S.; Pratyush, P.; Heinzinger, M.; Newman, R. H.; Kc, D. B. Improving protein succinylation sites prediction using embeddings from protein language model. *Sci. Rep.* **2022**, *12* (1), 16933 DOI: 10.1038/s41598-022-21366-2.

(44) Meenal Chaudhari, N. T.; Roy, Kaushik; Robert, H.; Newman, H. S.; C, D. B. K. DeepRMethylSite: a deep learning based approach for prediction of arginine methylation sites in proteins. *Molecular Omics* **2020**, *16* (5), 448−454, DOI: 10.1039/D0MO00025F.

(45) Pratyush, P.; Pokharel, S.; Saigo, H.; D. B.K. 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 Bioinformatics* **2023**, *24* (1), 41 DOI: 10.1186/s12859-023-05164-9.

(46) Bengio, Y.; Ducharme, R.; Vincent, P.; Jauvin, C. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.* **2003**, *3*, 1137−1155.

(47) Littmann, M.; Heinzinger, M.; Dallago, C. Protein embeddings and deep learning predict binding residues for various ligand classes. *Sci. Rep.* **2021**, *11* (1), 23916 DOI: 10.1038/s41598-021-03431-4.

(48) 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.; Nielsen, H. Signal P 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **2022**, *40*, 1023−1025.

(49) Thumuluri, V.; Almagro Armenteros, J. J.; Johansen, A. R.; Nielsen, H.; Winther, O. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Res.* **2022**, *50* (W1), W228−W234.

(50) Høie, M. H.; Kiehl, E. N.; 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.

(51) Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S. M.; Zhou, Michael; Yanqi; Li, W.; Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 5485−5551.

(52) Steinegger, M.; Mirdita, M.; Soding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples many fold. *Nat. Methods* **2019**, *16* (7), 603−606.

(53) UniProt, C. UniProt: the universal protein knowledge base in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D480−D489, DOI: 10.1093/nar/gkaa1100.

(54) Shao, J.; Xu, D.; Tsai, S. N.; Wang, Y.; Ngai, S. M.; Shiu, S. H. Computational identification of protein methylation sites through bi-profile Bayes feature extraction. *PLoS ONE* **2009**, *4* (3), No. e4920.

(55) Chen, Z.; Zhao, P.; Li, F.; Leier, A.; Marquez-Lago, T. T.; Webb, G. I.; Baggag, A.; Bensmail, H.; Song, J. PROSPECT: A web server for predicting protein histidine phosphorylation sites. *J. Bioinform Comput. Biol.* **2020**, *18* (4), 2050018.

(56) Chaudhari, M.; Thapa, N.; Ismail, H.; Chopade, S.; Caragea, D.; Kohn, M.; Newman, R. H.; Kc, D. B. DTL-DephosSite: Deep Transfer Learning Based Approach to Predict Dephosphorylation Sites. *Front Cell Dev Biol.* **2021**, *9*, 662983.

(57) Gao, J.; Thelen, J. J.; Dunker, A. K.; Xu, D. Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol. Cell Proteomics* **2010**, *9* (12), 2586−2600.

(58) Van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *Mach. Learn. Res.* **2008**, *9*, 2579−2605.