DeepPhoPred: Accurate Deep Learning Model to Predict Microbial Phosphorylation

Author

Ahmed, Faisal, Sharma, Alok, Shatabda, Swakkhar, Dehzangi, Iman

Published

2024

Journal Title

Proteins

Version

Accepted Manuscript (AM)

DOI

10.1002/prot.26734

DeepPhoPred: Accurate Deep Learning Model to Predict Microbial Phosphorylation

Faisal Ahmed[1,2], Alok Sharma[3,4], Swakkhar Shatabda[1], and Iman Dehzangi[*5,6]

[1]Department of Computer Science and Engineering, United International University, Dhaka, Bangladesh
[2]Department of Computer Science and Engineering, Premier University, Chattogram, Bangladesh
[3]Laboratory of Medical Science Mathematics, Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo 113-0033, Japan
[4]Institute for Integrated and Intelligent Systems, Griffith University, Nathan, Brisbane, QLD-4111, Australia
[5]Department of Computer Science, Rutgers University, Camden, NJ 08102, USA
[6]Center for Computational and Integrative Biology (CCIB), Rutgers University, Camden, NJ 08102, USA
[*]Corresponding Author: Iman Dehzangi (i.dehzangi@rutgers.edu)

Running Title: "DeepPhoPred: Accurate Phosphorylation Site Predictor"

Abstract

Phosphorylation is a substantial post-translational modification of proteins that refers to adding a phosphate group to the amino acid side chain after translation process in the ribosome. It is vital to coordinate cellular functions, such as regulating metabolism, proliferation, apoptosis, subcellular trafficking, and other crucial physiological processes. Phosphorylation prediction in a microbial organism can assist in understanding pathogenesis and host-pathogen interaction, drug and antibody design, and antimicrobial agent development. Experimental methods for predicting phosphorylation sites are costly, slow, and tedious. Hence low-cost and high-speed computational approaches are highly desirable. This paper presents a new deep learning tool called DeepPhoPred for predicting microbial phospho-serine (pS), phospho-threonine (pT), and phospho-tyrosine (pY) sites. DeepPhoPred incorporates a two-headed convolutional neural network architecture with the squeeze and excitation blocks followed by fully connected layers that jointly learn significant features from the peptide's structural and evolutionary information to predict phosphorylation sites. Our empirical results demonstrate that DeepPhoPred significantly outperforms the existing

*Corresponding author: i.dehzangi@rutgers.edu

microbial phosphorylation site predictors with its highly efficient deep-learning architecture. DeepPhoPred as a standalone predictor, all its source codes, and our employed datasets are publicly available at `https://github.com/faisalahm3d/DeepPhoPred`

Keywords: Post Translational Modification, Microbial Phosphorylation, Evolutionary Information, Structural Information, Classification, Convolutional Neural Network, Imbalance Learning.

1 Introduction

Protein post translational modification (PTM) is a critical regulatory mechanism that refers to proteins' enzymatic changes after their biosynthesis [24]. It is estimated that 5% of the proteome undergo one of more posttranslational modifications [5]. PTMs can occur in different types of amino acids and alter their chemical characteristics. PTMs appear in diverse cellular organelles, including the nucleus, cytoplasm, endoplasmic reticulum, and Golgi body. They typically occur in proteins with critical structures or activities, such as histones, membrane, and secretory proteins[51].

PTM can be of two types - reversible or irreversible. While the former contains covalent changes, the latter includes proteolytic changes and proceeds in one direction. It was shown that most cell proteins are controlled by reversible post translational modification [26].

PTMs affect numerous protein behaviors and features, including protein solubility, folding, localization, protein longevity, protein-protein interactions, cell-cell and cell-matrix interactions, molecular trafficking, and receptor activation [51]. They also significantly contribute to biological processes, including signal transmission, gene expression regulation, gene activation, DNA repair, and cell cycle control [5].

PTMs also play a critical role in normal and pathological cell physiology. They enable cells to respond immediately to external environment changes and signal stimulation. [69]. Therefore, studying and understanding proteins and their PTMs are fundamental to comprehending their functions [65].

So far, more than 200 PTMs, including phosphorylation, prenylation, ubiquitination, methylation, crotonylation, and acetylation have been identified [51]. Among them, phosphorylation is one of the substantial PTM that refers to adding a phosphate group to the amino acid side chain. It is vital for coordinating cellular and organic functions such as regulating metabolism, proliferation, apoptosis, subcellular trafficking, and other crucial physiological processes [9]. The phosphorylation commonly occurs to three amino acids in the protein sequence namely, serine (S), threonine (T) and tyrosine (Y). When phosphorylation occurs in serine (S), threonine (T), and tyrosine (Y) residues, they are

called phospho-serine (pS), phospho-threonine (pT), and phospho-tyrosine (pY) sites, respectively. Phosphorylation is known as the most significant chemical alteration among the intracellular signal transmission pathways [49]. Due to the phosphate group's size and negative charge, phosphorylation is one of the most efficient ways to alter proteins' structure. For the same reason, it works well as a recognition signal for other proteins [30]. Understanding pathogenesis, host-pathogen interactions, drug and antibody design, and the invention of antimicrobial agents can all be aided by phosphorylation studies in microbial organisms [52, 48].

Experimental methods, including Edman degradation [58], 32P-labeling [8], and mass spectrometry [10] for phosphorylation detection, are expensive, time-consuming, and tedious. Hence there is an urgent necessity for computational approaches which are fast, accurate, and frugal.

Machine Learning and Deep Learning have become the most popular methods for rapidly evaluating massive and complicated scientific datasets to retrieve valuable insights. They have also exhibited promising results in bioinformatics and computational biology, including sequence analysis [39, 55, 19, 16], structure prediction [28, 41], interaction and reconstruction [53], biomolecular property and function prediction [54, 44], biomedical image processing [45, 32], diagnosis, and systems biology [39, 15, 31].

During the last two decades, numerous computational methods, such as NetPhos [12], PRED [11], Musite [22], Phospho-SVM [20], PhosphoPredict [57], PhosPred-RF [64], MusiteDeep [63], CapsNet [62], DeepPhos [40], DeepPPSite [3], Chlamy-EnPhosSite [59] and PhosIDN [66] have been proposed to predict phosphorylation sites on eukaryotes or higher order organism. However, to the best of our knowledge, only five machine learning-based tools have been presented to predict phosphorylation sites on prokaryotes or microbial organisms.

The first two bacterial phosphorylation site predictors are NetPhosBac [46] and cPhosBac [37]. The NetPhosBac exploits neural networks as prediction engines. cPhosBac uses k-space amino acid pair (KSAAP) features extracted from the proteins sequence and an SVM as the classifier. Both predictors were trained using the same dataset of 152 and 841 laboratory-verified phosphorylated and non-phosphorylated serine/threonine sites in 119 substrates. Among these two, cPhosBac demonstrates better performance than NetPhosBac. Later, Zhang et al. suggested a prkC-specific phosphorylation site predictor called prkC-PSP [68]. prkC-PSP exploits amino acids' location information as features and SVM as the classification technique. They manually curated 36 phosphorylation and 512 non-phosphorylation sites from the literature, which were experimentally confirmed in the laboratory to

train their predictor.

Later on, Mamun et al. proposed a general microbial phosphorylation site predictor called, MPSite [25]. MPSite utilizes various sequence encoding techniques, including amino acid composition, amino acid frequency composition, binary encoding, and amino acid index properties for feature extraction. Subsequently, it selects the optimal set of features using the Wilcoxon rank-sum test (WR). Finally, the prediction engine is developed employing a Random Forest (RF) classifier. Most recently, Faisal et al. proposed a new machine learning-based tool called RotPhoPred [2, 1], to predict phosphorylation on microbial organisms' serine (S), threonine (T), and tyrosine (Y) residues, integrating structural and evolutionary information. RotPhoPred combines predicted structural features and evolutionary bigram profiles to describe each peptide fragment in the dataset. It then applies the NearMiss-3 undersampling technique to balance the training dataset to avoid bias towards the majority class. Subsequently, they used Rotation Forest classifier to predict microbial phosphorylation sites. So far, RotPhoPred is considered as the most promising microbial phosphorylation site predictor.

Despite all the efforts to improve microbial phosphorylation site prediction in recent years, there are still several shortcomings that require further investigation. First, the overall performance of the aforementioned microbial phosphorylation site predictors is still not satisfactory. The current best predictor fails to detect around 10%, 25%, and 26% pS, pT, and pY sites, respectively. Hence there is eventually room for improvement in prediction performance. Second, all existing predictors rely on manual feature generation mechanisms and use traditional machine learning methods. However, generating domain-specific appropriate features from protein sequences is challenging and requires many trial and error cycles. Most of the time, manual feature engineering leads to insufficient representation and poor prediction results [23, 4, 13]. To the best of our knowledge, automated feature extraction from protein sequences using deep learning approaches has not been explored yet to tackle this problem.

In recent years the Convolutional Neural Network (CNN), with its inherent feature extraction and classification ability, has demonstrated tremendous success in a wide range of bioinformatics applications such as anticancer activity classification of therapeutic peptides [60], sigma promoter identification [7, 34, 56], DNA-protein binding prediction [67], anticancer peptide prediction [23, 4, 13], DNA N6-methyladenine site identification [50, 35] and predicting phenotypes from genotypes [42, 38]. Besides, previous studies show that incorporating proteins' evolutionary and structural information provides more descriptive feature representation, which enhances a model's prediction and classification ability, tremendously [17, 55, 19, 16].

In this study, we propose a new deep learning approach called DeepPhoPred, which consists of two stacked convolutional neural networks with squeeze and excitation blocks that can jointly learn compelling features from peptide's structural and evolutionary characteristics to predict microbial phosphorylation sites accurately. Besides, the data imbalance issue is tackled in this predictor using the Synthetic Minority Oversampling Technique (SMOTE). Our results demonstrate that DeepPhoPred can suppress the state-of-the-art predictors with significantly high margins for different evaluation metrics. It attains sensitivity of 100%, 99.2% and 99.3%, specificity of 98.9%, 99.0% and 98.7%, accuracy of 99.0%, 99.0% and 98.7%, and Matthew's correlation coefficient (MCC) of 0.93, 0.94 and 0.95 for predicting pS, pT and pY sites, respectively. DeepPhoPred enhances the pS, pT, and pY site prediction accuracies by 10.0%, 23.8%, and 21.1% compared to previous studies found in the literature, respectively. DeepPhoPred as a standalone predictor, all its source codes, and our employed datasets are publicly available at `https://github.com/faisalahm3d/DeepPhoPred`.

In summary, this study provides several vital contributions:

1. **Dataset Curation:** Aggregation and curation of substantial benchmark datasets comprising experimentally confirmed microbial phospho-serine (pS), phospho-threonine (pT), and phospho-tyrosine (pY) sites, enhancing the availability of high-quality data for research in this domain.

2. **Novel CNN Architecture:** Introduction of an efficient yet lightweight convolutional neural network (CNN) architecture, employing a relatively low number of learnable parameters for predicting phosphorylation sites. This innovative approach represents the first instance, to the best of current knowledge, of a microbial phosphorylation site predictor based on deep learning that can accurately identify serine, threonine, and tyrosine phosphorylation sites by leveraging structural and evolutionary information embedded in peptides.

3. **Incorporation of Channel Weighting Mechanism:** Implement a channel weighting mechanism within the convolution layers by integrating squeeze and excitation blocks. This inclusion serves to mitigate biasing tendencies toward specific classes during model training, thus enhancing the model's ability to learn more effectively from the data.

4. **Phosphorylation Site Agnostic Model:** Development of a model architecture that demonstrates exceptional performance in predicting pS, pT, and pY sites using the same network configuration. This underscores the versatility and effectiveness of the proposed model across different types of phosphorylation sites.

5. **Comprehensive Experimental Validation:** Extensive experiments were conducted on the benchmark dataset, showcasing performance metrics on par with state-of-the-art predictors. These results serve to validate the efficacy and accuracy of the proposed predictor.

6. **Open Accessibility:** Release the source code and the dataset employed in this study to the public domain. This open availability aims to facilitate and encourage further investigation and exploration in future studies within the scientific community.

2 Methodology

This section presents our proposed methodology to build DeepPhoPred, including the benchmark datasets, peptide encoding, balancing datasets, and architecture and training processes of the deep learning model.

2.1 Benchmark dataset

The benchmark datasets used in this study have been collected from the dbPSP database [48]. dbPSP database holds 19,296 experimentally confirmed phosphorylation sites in 8,586 proteins from 200 prokaryotic organisms of 12 species belonging to bacteria and archaea kingdoms. Initially, we curated three datasets of proteins with phosphorylated serine (S), threonine(T), and tyrosine (Y) residues. Subsequently, we employed CD-HIT tools to remove the protein sequence with more than 40% sequential similarities [21, 36, 29]. The resultant benchmark datasets include 1483, 1220, and 1161 proteins with phosphorylated S, T, and Y sites, respectively.

The pS site prediction problem's dataset consists of 36,513 sites having 2024 phosphorylated (positive) and 34489 no-phosphorylated (negative) sites. The pT dataset comprises 1647 positive and 24592 negative, resulting in 26239 sites. There are 17476 sites, with 1644 and 15832 positive and negative, respectively, in the pY dataset. An overview of three datasets is presented in Table 1. The table shows that the datasets are imbalanced with more negative than positive sites. The imbalance ratios between positive and negative sites are 1:17, 1:15, and 1:10 for pS, pT, and pY datasets.

Table 1: Overview of the entire benchmark dataset for pS, pT, and pY sites prediction problems.

Tasks	Total Sequences	Positive Sites	Negative Sites	Imbalance Ratio (Positive: Negative)
pS	1483	2040	34489	1:17
pT	1220	1647	24952	1:15
pY	1161	1644	15832	1:10

Table 2: Summary of the training dataset **A** for pS, pT, and pY sites prediction problems

Tasks	Total Sequences	Positive Sites	Negative Sites
pS	1260	1700	30283
pT	1037	1394	22373
pY	987	1360	14197

To form the training set, we randomly separate 85% of the proteins for each task, which consists of 1260, 1037, and 987 proteins for pS, pT, and pY sites prediction tasks. This dataset is referred to as A in the rest of the paper and is used to build the prediction models.

The remaining 15% of proteins of each task, which account for 223, 183, and 174 proteins for pS, pT, and pY sites, constitute the independent test set B that is exploited to validate the performance of models. The training set A and independent test set B are summarized in Tables 2 and 3, respectively.

Table 3: Summary of the independent dataset **B** for pS, pT, and pY sites prediction problems

Tasks	Total Sequences	Positive Sites	Negative Sites
pS	223	340	4206
pT	183	253	2579
pY	174	284	1635

For external validation, we formed another dataset (referred to as dataset C) by randomly extracting 715 phosphorylated proteins from Phospho.ELM. The number of positive and negative sites for S, T, and Y residues are summarized in Table 4. We then generate evolutionary and structural profiles for each protein in this dataset in the same ways as in datasets A and B.

Table 4: Number of sites of each of the three phosphorylated residues from dataset **C**.

Residue	Number of instances (Positive)	Number of instances (Negative)
S	1504	9920
T	316	6917
Y	228	4838

2.2 Representation of each site as a peptide

A protein sequence involves multiple phosphorylated and non-phosphorylated sites. This study describes each site by a short chain of amino acids known as a peptide. A peptide P includes n upstream and n downstream neighboring amino acid residues with a central S/T/Y site as follows:

$$P = \{A_{-n}, A_{-(n-1)}, ..., A_{-2}, A_{-1}, S/T/Y, A_1, A_2, ..., A_{+(n-1)}, A_{+n}\} \quad (1)$$

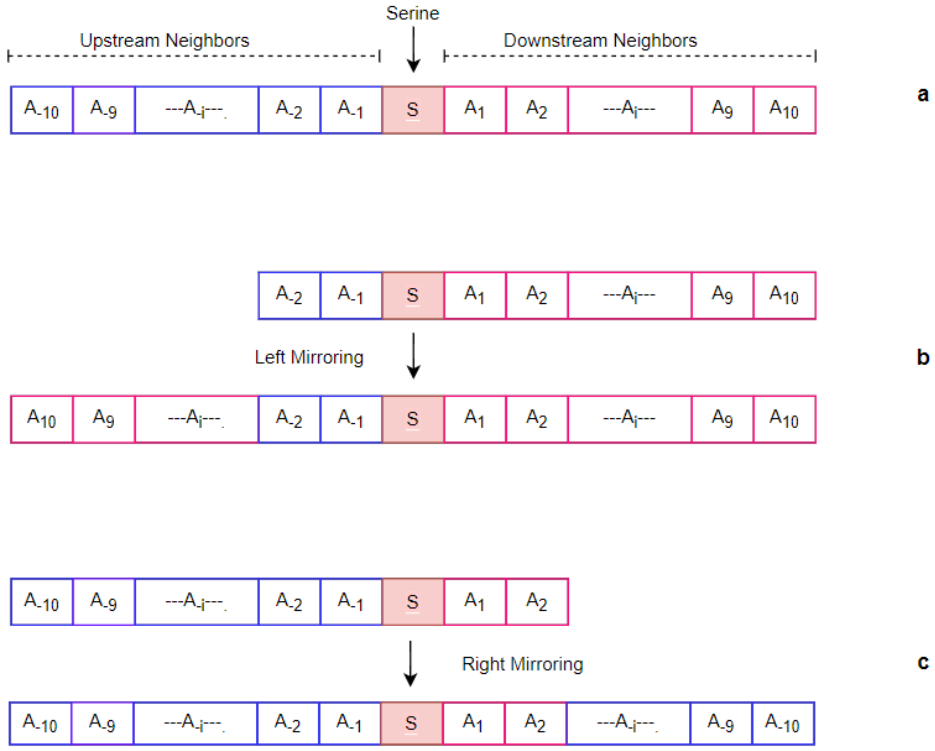

Figure 1: The representation of Serine site as a peptide. (a) when there are enough upstream and downstream neighbors, (b) when there is no sufficient upstream neighbor, and (c) when there is no adequate downstream neighbor.

where A_{-i} and A_{+i} mean the upstream and downstream amino acids respectively. In case of less than n neighboring amino acids on either side of S/T/Y, the mirroring technique is exploited to suffice the absent amino acid positions. The representation of each Serine site as a peptide when $n = 10$ is demonstrated in Figure 1.

2.3 Encoding peptides

Peptides' structural and evolutionary properties are used as inputs to build the proposed deep learning model since they exhibit promising results in other PTM sites prediction. Evolutionary property extracted from position-specific scoring matrix (PSSM) refers to the substitution probability of given amino acid with other amino acids in a protein sequence. On the other hand, protein's structural properties include secondary structures, accessible surface area (ASA), and torsion angles.

The evolutionary information of each protein sequence in our dataset is generated using the PSI-BLAST [6] tool. PSI-BLAST was run for three iterations with a cutoff value of 0.001 on NCBI's

non-redundant (nr) database. It produces an Lx20 dimensional PSSM matrix for each sequence, where L signifies the protein's length and 20 represents different amino acids which form the genetic code. To encode the evolutionary information for each positive or negative site described by the peptide P in our dataset, the affiliated submatrix M is pulled from the PSSM matrix of the protein sequence incorporating P. M is a Wx20 shape matrix, where W is the window size, and 20 represents different amino acids.

Besides, each protein structural properties are predicted using the deep learning-based tool SPIDER2 [28, 41]. It outputs an SPD file that holds a matrix of Lx8 sizes that symbolizes the predicted values of eight structural properties (coil, strand, helix, ASA, ψ, ϕ, θ, τ) for each amino acid residue in a protein of length L. The submatrix N corresponds to the protein fragment p and is segmented from the SPD file to encode the structural properties. Hence, N is a Wx8 dimensional matrix, where W represents the window size, and 8 denotes the eight structural properties mentioned above.

The proposed deep learning model uses M and N matrices that capture evolutionary and structural information, respectively, to learn significant discriminatory features and predict phosphorylation sites.

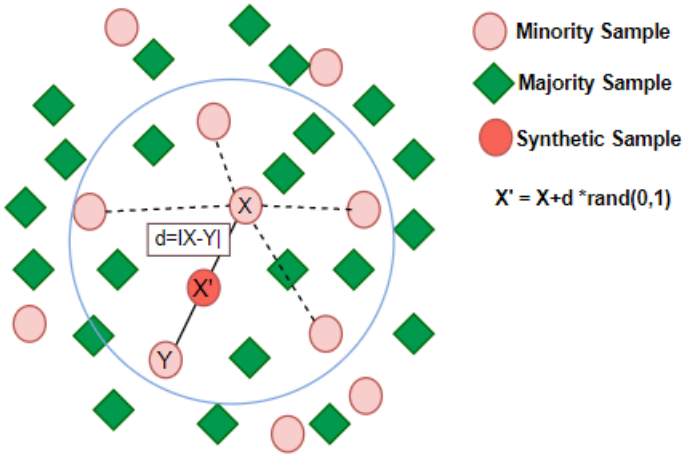

Figure 2: Working process of SMOTE with 5 nearest neighbors

2.4 Balancing the dataset

As discussed in the section 2.1, the benchmark datasets are imbalanced with more negative samples than positive. The model trained on such an imbalanced dataset shows partiality toward the majority class and negatively impacts the model's prediction accuracy. Therefore, it is common practice in machine learning to suspend the imbalance issue from the training data by oversampling [27, 14] or

undersampling [61, 43] techniques. While the former generates synthetic examples for the minority class, the latter cuts down the instances from the majority class to make the sample distribution in each category uniform. Earlier studies on PTM [18] [33] [47] show that the deep learning method performs better with more training data. Therefore, we have balanced the training dataset using the synthetic minority oversampling technique (SMOTE) [14]. SMOTE works as follows:

It randomly selects a sample X from the minority class and discovers its k nearest minority instances based on some distance metric. The distance between X and another sample, Y, which is arbitrarily chosen from the k instance, is then calculated. A synthetic instance X' is then generated by adding the scaled distance value with X. The scaling of the distance value is performed by multiplying it with a random number between 0 and 1. The process is repeated until the dataset contains an equal number of positive and negative instances. The overall strategy for 5 nearest neighbors depicted by the blue circle is shown in Figure 2.

We vectorized the evolutionary and structural encoding matrices and merged them for each peptide in the training set to apply SMOTE techniques. Subsequently, we reversed the vectorization process for each instance in the upsampled dataset to describe it again as two matrices.

2.5 Architecture of the proposed deep learning model

The proposed deep model's architecture is presented in Figure 3.

It's input layer has two heads consisting of 21 (window size) x 20 (number of amino acids) and 21 (window size) x 8 (number of structural properties) dimensional matrices representing the evolutionary and structural information, respectively. Two parallel one-dimensional convolution layers are added next to these input channels, which can jointly learn and extract complex, and discriminative features from the two information sources(structural and evolutionary profiles). Each convolution layer consists of 64 filters with kernel size 5, which is determined through cross-validation. Hence each convolution layer generates a feature map composed of 64 channels.

We do not utilize any pooling layer after the convolution layers for preserving spatial information and residue-related features extracted by the convolution layer. Moreover, our experiments show that it degrades the model's prediction performance. Our dataset is comparatively small, which may cause the overfitting problem. Hence, both l2 (weight decay) and dropout regularization techniques are employed in the convolutional layers to alleviate this issue.

Next, each convolution layer has appended a squeeze and excitation (SE) block to introduce a

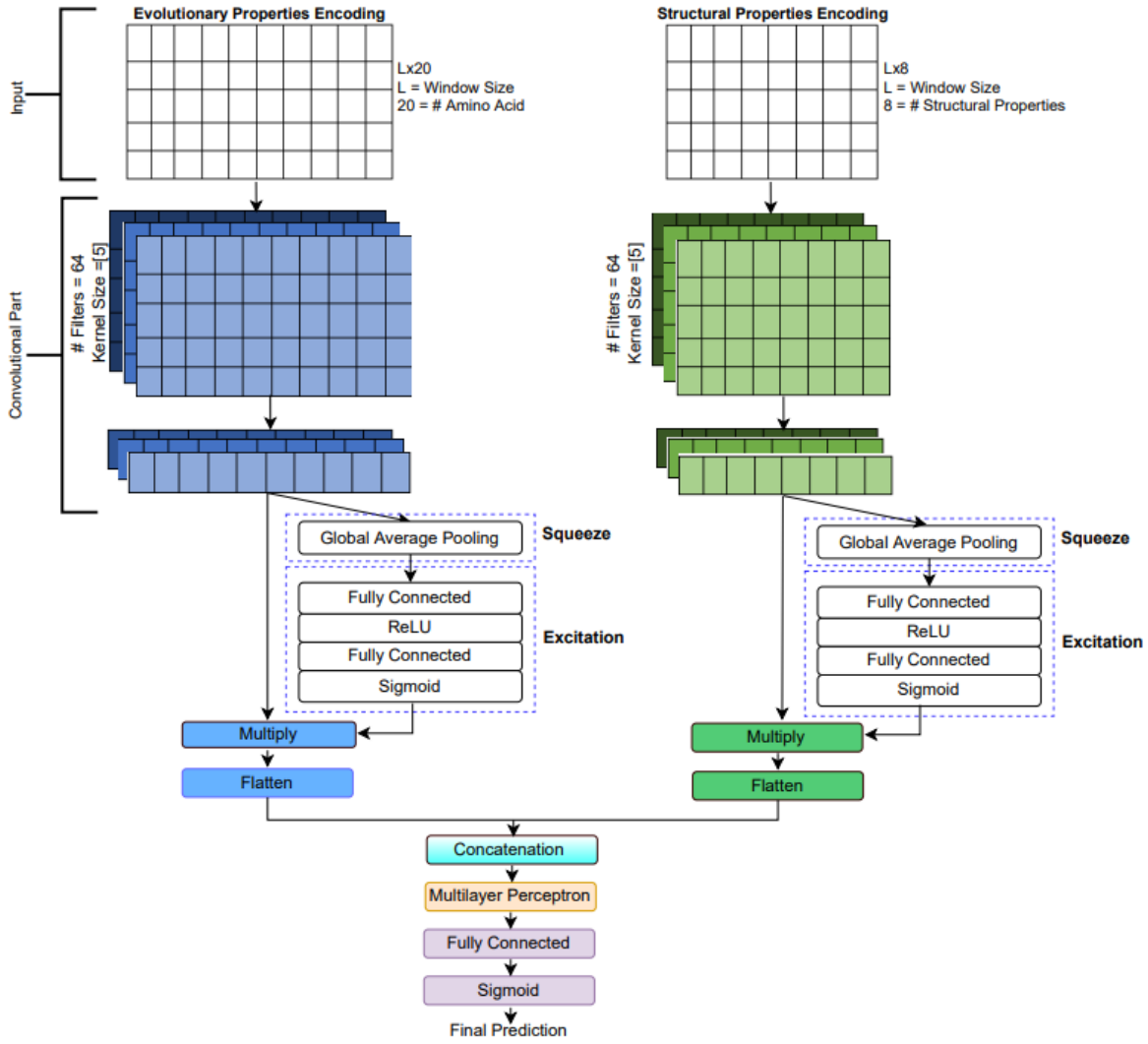

Figure 3: The general architecture of the proposed deep learning model DeepPhoPred

channel attention mechanism by weighting each channel of the feature map adaptively based on their global information. The block consists of three modules: squeeze, excitation, and scaling. The squeeze module gets a global understanding of each channel by employing a global average pooling layer, which reduces the spatial dimension of the whole channel to a singular value by taking the average of all values in that channel. The value indicates the relative importance of the channel. On the other hand, the excitation module learns the weights to scale each channel information of the feature map adaptively. Two fully connected layers operate in this module. The first layer with a Relu activation function reduces the dimensionality of the input tensor by a factor of r (which is set at 16 in this study) and decreases the computational complexity. The second fully connected layer restores the

tensor dimension to the feature map's shape. This layer is followed by the sigmoid activation function, which converts the input tensor's values between 0 to 1. Finally, the scaling module executes element-wise multiplication between the feature map generated by the convolution layer and the output tensor of the sigmoid function to scale the feature map. Thus, the squeeze and excitation block reduces the non-relevant channel information without affecting the effective channels.

The scaled feature maps returned by the squeeze and excitation blocks are flattened into two one-dimensional vectors and later concatenated to form a feature vector. The resulting feature vector is then passed through a multi-layer perceptron consisting of several densely connected neural network layers so that the model can successfully learn the importance of each feature and capture the complex relationship between the features in predicting the phosphorylation sites. The Relu activation function is applied to each dense layer's neuron to draw a non-linear decision boundary in the complex feature space for separating phosphorylation sites from non-phosphorylated sites. The Relu is chosen over other activation functions because of its simplicity, popularity, and effectiveness. The dropout layer is added next to each dense layer to prevent the model from overfitting. The output layer of the network constitutes a dense layer of a single neuron with the sigmoid activation function for binary classification, which returns the probability of the site being phosphorylated.

2.6 Model training

The binary cross-entropy loss function is used as the objective function, which measures the uncertainty associated with the predicted probability distribution as follows:

$$L_c = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)] \tag{2}$$

where, y is the target value of the network, which is 1 for the phosphorylated site and 0 for the non-phosphorylated site, $\hat{y}$ is the predicted probability that the site is positive, and N is the number of training instances. The model's parameters are trained using the Adam optimizer, which computes adaptive learning rates for each parameter based on the estimated first and second moments of the gradients, with an initial learning rate of 0.0001 and an exponential decay rate for the first and second moment estimates of $\beta_1 = 0.5$, $\beta_2 = 0.999$ and batch size of 64 samples. An early stopping strategy was used to reduce the training time. We used the model checkpoint function to choose the best model from all the epochs based on the highest accuracy and lowest loss on the validation dataset. All the

hyperparameters are tuned through the cross-validation approach.

2.7 Validation

The maturity of DeepPhoPred has been evaluated using the stratified 5-fold cross-validation scheme and independent test. In the stratified 5-fold cross-validation approach, we have split the training set into five equal parts, maintaining the same positive and negative class distribution in each part. Subsequently, we built the model using four folds and assessed its performance on the remaining fold using the evaluation metrics mentioned in section 2.8. The process is repeated five times to ensure that each fold has been employed to test the model. Finally, we calculated the mean of each metric. We also use an independent test set to validate our model and assess its generality for unseen data that are not used in the training process.

2.8 Evaluation metrics

The performance of DeepPhPred is measured using six evaluation metrics namely, Sensitivity (Sn), Specificity (Sp), Precision (Pr), Accuracy (Ac) F1-score (F1), and Matthews correlation coefficient (MCC), which are widely accepted for classification problems. They are computed using the following equations:

$$Sn = \frac{TP}{TP + FN} \tag{3}$$

$$Sp = \frac{TN}{TN + FP} \tag{4}$$

$$Pr = \frac{TP}{TP + FP} \tag{5}$$

$$Ac = \frac{TP + TN}{TP + FN + FP + TN} \tag{6}$$

$$F1 = \frac{2 * Pr * Sn}{Pr + Sn} \tag{7}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TN + FN) * (TP + FP) * (TN + FP) * (TP + FN)}} \tag{8}$$

where, TP, TN, FP, and FN denote True Positive, True Negative, False Positive, and False Negative, respectively. True positives indicate the number of instances correctly predicted as Phosphorylation sites. False positive refers to the number of samples that are wrongly identified as phosphorylated while they are no-phosphorylated sites. True negative quantifies the accurately classified non-phosphorylated

sites as non-phosphorylated. Similarly, False Negative is the number of phosphorylated sites misclassified as non-phosphorylated.

3 Results and Discussion

This section presents our results, compares them with previous studies, and discusses their significance.

3.1 Contribution analysis of Squeeze and Excitation (SE) block

While generating the feature map, the convolution layer gives equal priority to each filter which might not be reasonable. Some channels in the feature map might be more critical. Therefore, in this study, we employed a squeeze and excitation (SE) block that introduced a channel attention mechanism and scaled the spatial information of different channels in the feature map based on their relative importance. This section analyses the SE block's contribution to improving the proposed model's prediction performance. To do this, we trained and evaluated the same model after removing the SE block only. The evaluation results reported in Table 5 show that the model performance decreased in cross-validation and independent test when the SE block was removed. As shown in Table 5, the MCC scores of the model with the SE block are 0.18, 0.25, and 0.16 higher than those without the SE block for phosphorylation detecting in serine, threonine, and tyrosine residues, respectively, on the independent test set. The model's precision also decreased significantly from 0.88, 0.90, and 0.93 to 0.60, 0.53, and 0.68 for pS, pT, and pY sites prediction, respectively. These results indicate that the SE block greatly minimizes the false positive rate and enhances the model's prediction performance. The excellent performance of DeepPhoPred in terms of accuracy, precision, and F1 scores stems from two fronts. First, DeepPhoPred combines structural and evolutionary features. As demonstrated in previous works, they are affluent feature descriptors for predicting post-translational modification. Second, DeepPhoPred incorporates squeeze and excitation (SE) blocks in its architecture that bring channel attention mechanisms that are helpful for more discriminative global feature learning. The t-SNE plots in section 3.5, where feature embedding learned by DeepPhoPred is visualized, further explain the high accuracy, precision, and recall performance.

Table 5: Performance of DeepPhoPred with and without SE block: models are trained on dataset A and evaluated on dataset B.

Task	Model	Ac (%)	MCC	Pr	auROC	F1	Sn (%)	Sp (%)	auPR
pS	With SE block	99.0	0.93	0.88	0.99	0.94	100	98.9	1.00
	Without SE block	95.1	0.75	0.60	0.97	0.75	99.4	94.7	0.98
pT	With SE block	99.0	0.94	0.90	0.99	0.95	99.2	99.0	0.99
	Without SE block	91.9	0.69	0.53	0.95	0.69	99.6	91.2	0.99
pY	With SE block	98.7	0.95	0.93	0.99	0.96	99.3	98.7	1.00
	Without SE block	93.2	0.79	0.68	0.96	0.81	100	92.0	0.99

3.2 Contribution analysis for different sequence representations

In this section, we have examined the significance of different sequence representation techniques in phosphorylation prediction. We have trained two individual CNN models, one with the structural and another with the evolutionary information, and compared them with the proposed model (DeepPhoPred) that integrates both encoding schemes for this experiment. Structural and Evolutionary models are designed just by pruning the input head from the DeepPhoPred that captures the evolutionary and structural information, respectively. Other layers of the models were kept unchanged for a fair comparison. The comparative results on 5 folds cross-validation are presented in Table 6. The results show that the combined model outperforms the structural and evolutionary models on cross-validation for each measure in predicting three phosphorylation sites. Significant improvements are observed in accuracy, precision, specificity, MCC, and F1 scores.

For pS prediction, the combined model achieves accuracy, specificity, precision, MCC and F1 values of 99.5%, 99%, 0.99, 0.99, and 1.00, respectively, which are 7.7%, 14.85%, 0.13, 0.15, and 0.8 higher than the structural model and 7%, 12.6%, 0.11, 0.13, and 0.7 higher than the evolutionary model.

For pT, the combined model improves accuracy, specificity, precision, MCC, and F1 score by 8.3%, 15.8%, 0.13, 0.16, and 0.7 compared to the structural model and 5.9%, 11.4%, 0.10, 0.11, and 0.05 compared to the evolutionary model.

For pY, the combined model outperforms the structural model by 9.4%, 17.5%, 0.14, 0.18, and 0.9 in terms of accuracy, specificity, precision, MCC, and F1 score, while they are 9.3%, 17.1%, 0.14, 0.18, and 0.09 for the evolutionary model.

All the results demonstrate that both sequence encoding schemes significantly contribute to the model performance by reducing the false positive rate and increasing its ability to detect non-phosphorylation sites accurately.

The ROC and PR curves of these models on test data are plotted in Figure 4. The curves also

Table 6: 5-fold cross-validation on dataset A for different sequence representations.

Task	Encoding	Ac (%)	MCC	Pr	auROC	F1	Sn (%)	Sp(%)	auPR
pS	Combined	99.5	0.99	0.99	0.99	1.00	100	99.0	1.00
	Structural	91.8	0.84	0.86	0.92	0.92	99.2	84.2	0.98
	Evolutionary	92.5	0.86	0.88	0.92	0.93	98.6	86.4	0.99
pT	Combined	99.4	0.99	0.99	0.99	0.99	99.7	99.1	1.00
	Structural	91.1	0.83	0.86	0.91	0.92	99.0	83.3	0.97
	Evolutionary	93.5	0.88	0.89	0.94	0.94	99.6	87.7	1.00
pY	Combined	98.8	0.98	0.98	0.99	0.99	99.5	98.1	1.00
	Structural	89.4	0.80	0.84	0.89	0.90	98.1	80.6	0.97
	Evolutionary	89.5	0.80	0.84	0.89	0.90	97.9	81.0	0.98

show that the combined model surpasses the individual models since it attains higher AUC values for pS, pT, and pY prediction.

3.3 Comparison with traditional machine learning-based classifiers

In this section, we have compared the results attained by the proposed CNN model with five traditional machine learning models- Rotation Forest (RoF), Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB) and Adaptive Boosting (AdaBoost). The comparative results are presented in Table 7. In the table, CNN shows superior performance over the traditional ML models in all performance metrics in all classification tasks, which suggests the efficacy of the proposed CNN architecture. The sensitivity and specificity of the CNN models are not only higher than the traditional ML models but also more consistent. Consequently, the CNN model yields significantly higher MCC scores (0.93, 0.94, and 0.95) in predicting pS, pT, and pY sites than ML models. In the case of the traditional ML models, we observe a notable imbalance between the sensitivity and specificity scores for all classification tasks, which indicates class bias and results in comparatively lower MCC scores.

The ROC and PR curves for all classification task on the independent test is shown in Figure 5. The figure shows that the CNN model covers a substantially larger region in the PR and ROC curves than conventional machine learning classifiers. It achieves about 25% and 50% higher values in ROC and PR curves than the top traditional ML models for predicting all types of phosphorylation sites. Specifically, it suppresses the best performing ml-classifier (Random Forest) for pS by achieving 0.26 and 0.46 higher areas in ROC and PR curves. It beats the Rotation Forest, which produces the highest AUC values in both ROC and PR curves for detecting pT sites among all ML models by 0.27 and 0.45. Finally, for pY sites, auROC and auPR are 0.32 and 0.46, respectively, higher than the Rotation Forest, which is the best machine predictor. Adaboost shows the worst performance in all cases. Such results

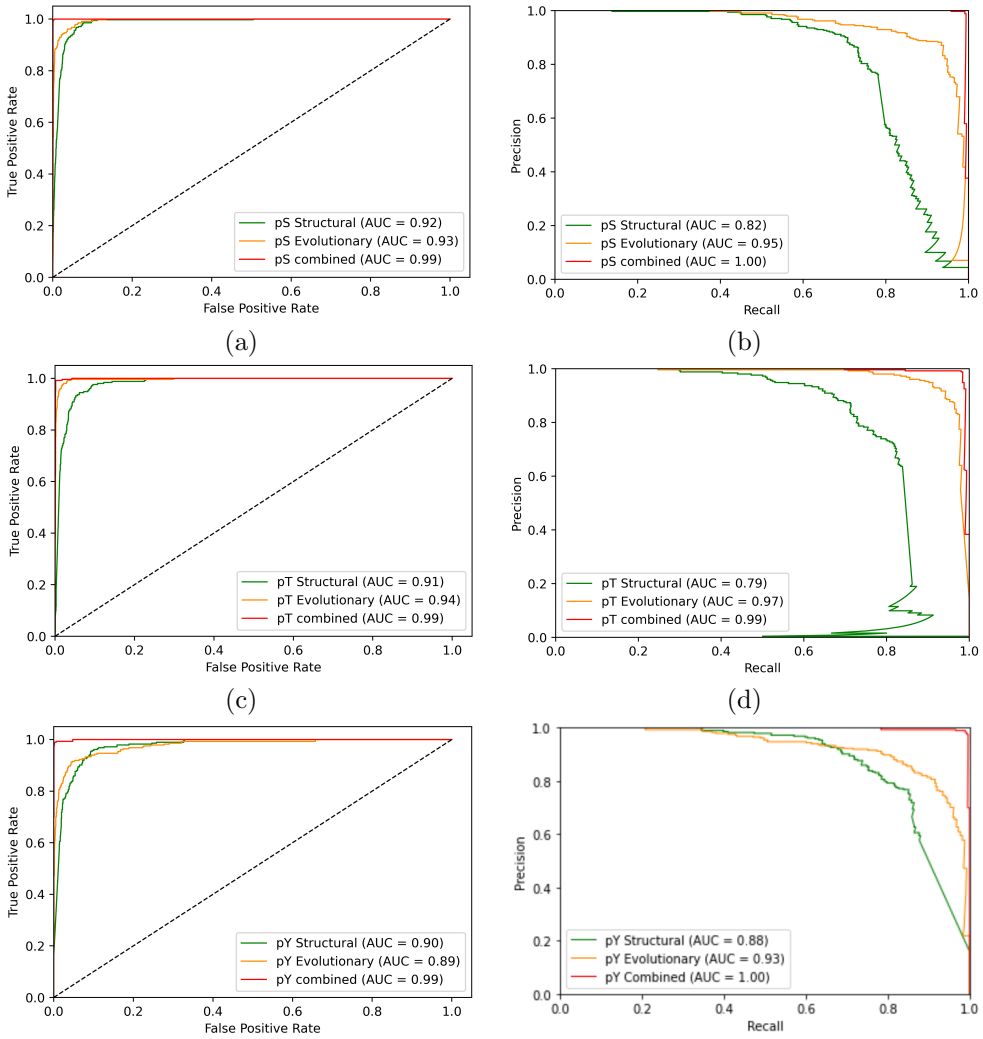

Figure 4: Comparative results for different encoding techniques. (a)ROC curves for pS site, (b)PR curves for pS site, (c)ROC curves for pT site (d) PR curves for pT site, (e) ROC curves for pY site, and (f) PR curves for pY site.

suggest that the traditional ML-based classifiers with the row structural and evolutionary information are insufficient to predict microbial phosphorylation sites accurately. Since the CNN model, with its optimal architecture, possess the ability to learn significant features from two information sources, it achieves outstanding performance in all prediction tasks.

Table 7: Performance of CNN and traditional ML models: models are trained on dataset A and evaluated on dataset B.

Task	Classifier	Ac (%)	MCC	Pr	auROC	F1	Sn (%)	Sp (%)	auPR
pS	CNN	99.0	0.93	0.88	0.99	0.94	100	98.9	1.00
	RoF	95.9	0.66	0.99	0.72	0.63	45.9	100	0.52
	RF	94.4	0.53	0.69	0.73	0.54	44.7	98.4	0.54
	SVM	95.8	0.65	1.00	0.72	0.61	44.1	100	0.53
	GB	84.3	0.29	0.25	0.70	0.34	53.2	86.9	0.33
	AdaBoost	80.8	0.24	0.20	0.68	0.29	52.4	83.1	0.37
pT	CNN	99.0	0.94	0.90	0.99	0.95	99.2	99.0	0.99
	RoF	94.9	0.64	0.98	0.72	0.61	43.9	99.9	0.54
	RF	94.8	0.63	0.99	0.71	0.60	42.7	100	0.53
	SVM	81.3	0.30	0.26	0.72	0.36	59.7	83.4	0.33
	GB	91.8	0.45	0.55	0.70	0.49	43.9	96.5	0.53
	AdaBoost	80.9	0.26	0.24	0.68	0.33	53.4	83.6	0.36
pY	CNN	98.7	0.95	0.93	0.99	0.96	99.3	98.7	1.00
	RoF	91.6	0.62	0.98	0.72	0.61	44.0	99.8	0.53
	RF	86.5	0.43	0.55	0.70	0.50	46.5	93.5	0.52
	SVM	76.8	0.31	0.34	0.69	0.42	57.7	80.1	0.25
	GB	91.3	0.61	1.00	0.71	0.59	41.5	100	0.54
	AdaBoost	73.9	0.23	0.29	0.64	0.37	50.7	77.9	0.18

3.4 Comparing data balancing techniques

To compare the performance of under-sampling and oversampling techniques for data balancing, we have trained the DeepPhoPred using NearMiss and SMOTE methods.

Table 8 presents the comparative results on the independent test set. The SMOTE shows remarkably higher results in this table compared to the NearMiss in all the performance measure metrics. On the other hand, NiearMiss exhibits poor MCC values of 0.64, 0.64, and 0.72 for pS, pT, and pY sites. It fails to detect 8.2%, 10.3%, and 12.7% phosphorylated serine, threonine, and tyrosine sites, respectively. The results are justified since deep learning needs a large dataset to train the model.

The ROC and PR curves for these two methods for predicting S, T, and Y phosphorylation sites are depicted in Figure 6. This figure shows that SMOTE achieves the perfect area under the ROC and PR curves for detecting three types of phosphorylation sites. It produces 0.07, 0.08, and 0.09 higher auROC values and 0.31, 0.34, and 0.22 higher auPR than the NearMiss for pS, pT, and pY,

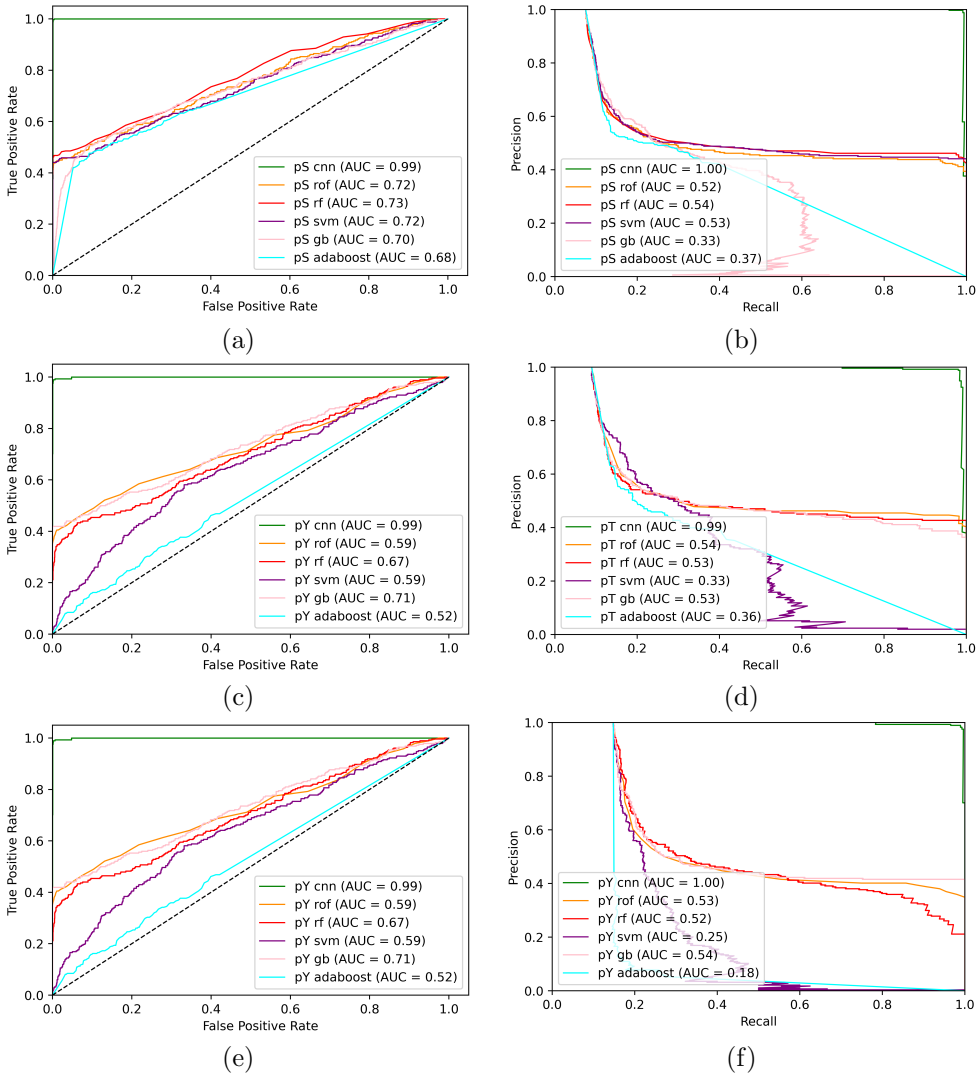

Figure 5: Comparative results for different classifiers. (a) ROC curves for pS site, (b)PR curves for pS site, (c)ROC curves for pT site, (d) PR curves for pT site, (e) ROC curves for pY site, and (f) PR curves for pY site

respectively.

3.5 Selection of optimal window size

385 In this section, we have performed experiments to determine the optimal window size for feature extraction based on the MCC score. The MCC scores for different window lengths (11, 21, 31, 41, and 51) on 5 folds cross-validation are graphically visualized on the bar plots in Figure 7 for pS, pT, and pY sites prediction. This figure reveals that window size 21 is optimal for predicting these three

Table 8: Performance of DeepPhoPred using different data balancing techniques: models trained on dataset A and evaluated on dataset B.

Task	Balancing Method	Sn (%)	Sp(%)	Pr	Ac (%)	F1	MCC	auROC	auPR
pS	SMOTE	100	98.9	0.88	99.0	0.94	0.93	0.99	1.00
	NearMiss	91.8	92.5	0.5	92.4	0.64	0.64	0.92	0.69
pT	SMOTE	99.2	99	0.90	99.0	0.95	0.94	0.99	0.99
	NearMiss	89.7	91.6	0.51	91.4	0.65	0.64	0.91	0.65
pY	SMOTE	99.3	98.7	0.93	98.7	0.96	0.95	0.99	1.00
	NearMiss	87.3	92.9	0.68	91.5	0.77	0.72	0.90	0.78

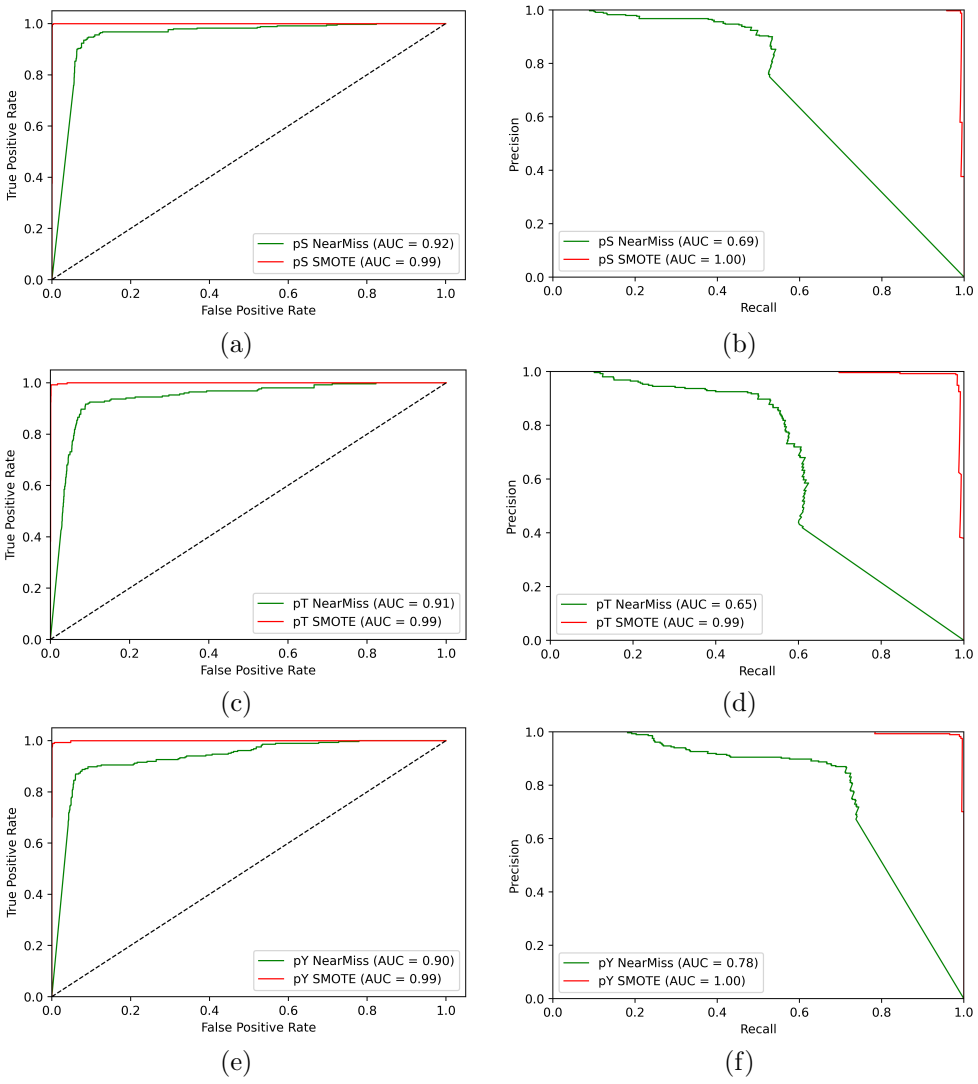

Figure 6: Comparative results for different data balancing technique. (a) ROC curves for pS site, (b)PR curves for pS site, (c)ROC curves for pT site, (d) PR curves for pT site, (e) ROC curves for pY site, and (f) PR curves for pY site

types of phosphorylation sites (pS, pT, and pY) with the highest MCC scores of 0.99, 0.99, and 0.98, respectively. . With a window size of more than 21, the MCC score remains constant, whereas the learnable parameter increases, thus the model complexity. On the other hand, the necessary sequence information may be lost when the window size is less than 21, which causes the model's performance degradation. Our findings are consistent with the results of the previous work[2] and [11], where the authors showed that window sizes of 7 to 10 are optimal for phosphorylation prediction.

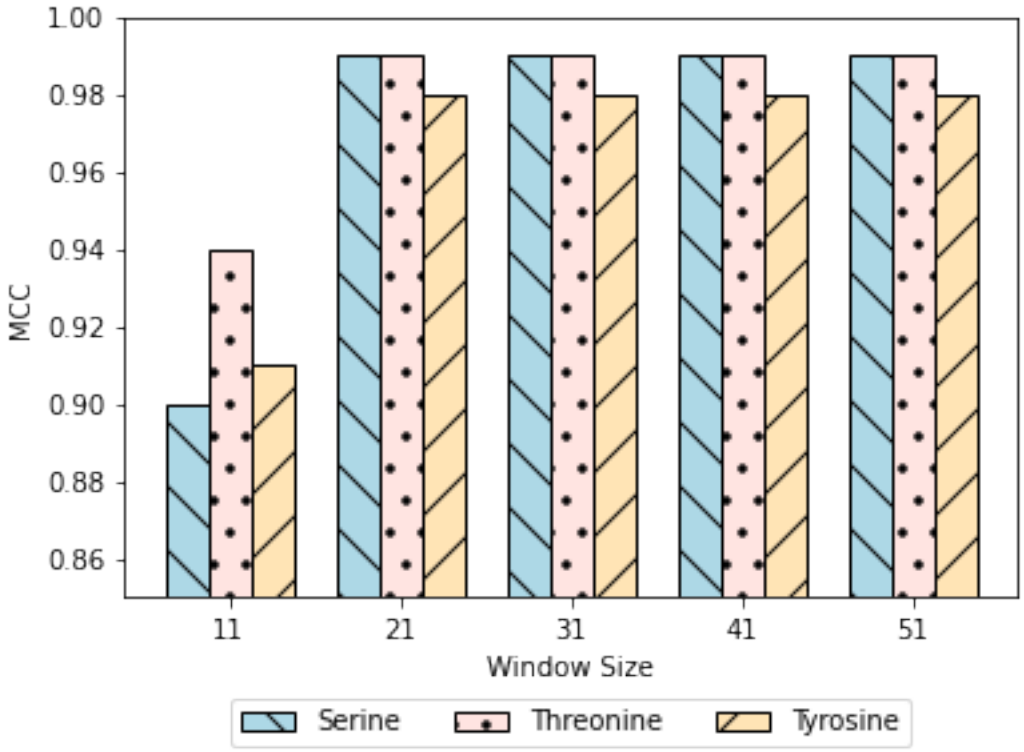

Figure 7: MCC scores for different window sizes for predicting phosphorylation on Serine, Threonine and Tyrosine residues.

3.6 Visualization of the features

We have used the t-distributed stochastic neighbor embedding (t-SNE) to visualize the discriminating ability between the positive and negative samples by the raw input features and learned features by our deep learning model. The scatter plots of two-dimensional components (tsne_1 and tsne_2) achieved by t-SNE for predicting pS, pT, and pY are shown in Figure 8. The plots show that the row structural and evolutionary features distribution is distorted, chaotic, and non-separable between the positive and negative samples. After passing through the convolution and multiple dense layers, the row input

features are transformed into a representation such that there is a transparent partition between the distribution of the positive and negative sites. These results indicate that our model can effectively learn distinguishing features for separating the phosphorylated site from non-phosphorylated ones.

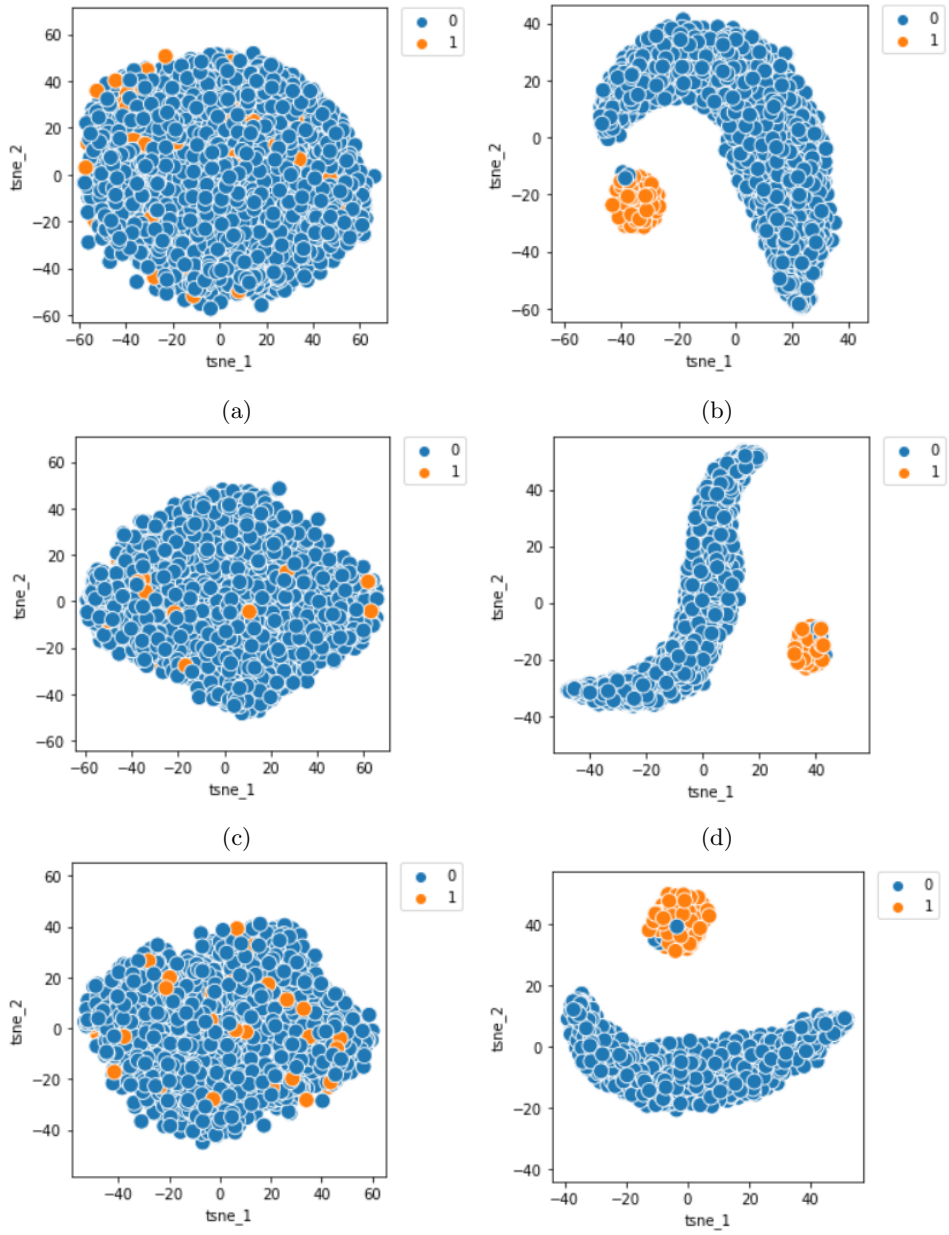

Figure 8: Two-dimensional projection of row input and features learned by DeepPhoPred (left to right) using t-SNE for pS, pT, and pT tasks (top to bottom); blue and orange points are negative and positive samples respectively.

3.7 Comparison with existing predictors

This section compares DeepPhoPred with publicly available existing predictors namely, RotPhoPred, MPSite, and NetPhosBac. For a fair comparison, we uploaded our independent test set to the RotPhoPred, MPsite, and NetPhosBac servers and calculated Sn, Sp, Ac, and MCC for all three tools from the predicted results. The same metrics are also recorded for DeepPhoPred on the independent test set. The comparative results for pS, pT and pY sites prediction are presented in Tables 9-11. The Tables show that DeepPhoPred outperforms MPsite, NetPhosBac, and RotPhoPred in all tasks, with significantly higher scores in all metrics. DeepPhoPred enhances Sn, Sp, Ac, and MCC up to 10%, 6.8%, 8%, and 0.11, 60.9%, 21%, 24%, and 0.82, and 68.8%, 31.1%, 33.9%, and 0.94 over RotPhoPred, MPSite, and NetPhosBack, respectively, for predicting pS. From Table 10, we can observe that DeepPhoPred improves Sn, Sp, Ac, and MCC up to 23.8%, 1.8%, 12.7%, and 0.20, 40.9%, 23.3%, 24.8%, and 0.72 and 89.8%, 6.2%, 13.6%, and 0.91 over RotPhoPred, MPSite, and NetPhosBack, respectively, for predicting pT. Table 11 shows that DeepPhoPred archives up to 21.1%, 4%, 12.3%, and 0.21 improvements in Sn, Sp, Ac, and MCC over RotPhoPred, respectively, in pY site prediction. RotPhoPred is the second best predictor and performs much better than MPSite and NetPhosBac in terms of all metrics in pS, pT, and pY sites prediction. The results demonstrate that DeepPhoPred can identify phosphorylated and non-phosphorylated sites with minimum error. In the case of DeepPhoPred, only 0%, 0.8%, and 0.7% phospho-serine (pS), phospho-threonine (pT), and phospho-tyrosine (pY) residues remain undetected while RotPhosPred fails to detect 10%, 24.6%, and 21.8% pS, pT, and pY sites, respectively. However, MPsite and NetPhosBac show much higher specificity than sensitivity, revealing the models are biased toward the negative class. The results are reasonable because none of these predictors addressed the data skewness issue in their training data. The positive-to-negative sample ratios of the datasets are 1:5 and 1:6 for MPSite and NetPhosBac, respectively. As a result, they can not detect a remarkably higher number of phosphorylated sites. The results suggest that our proposed models can complement the experimental method by helping to initially screen microbial phosphorylation sites. DeepPhoPred as a standalone predictor, all its source codes, and our employed datasets are publicly available at `https://github.com/faisalahm3d/DeepPhoPred`

3.8 Performance on external validation set

Going one step further, we evaluated the robustness of DeepPhoPred on the external validation set B described in section 2.1. The results for pS, pT, and pY prediction tasks are presented in Table 12.

Table 9: Performance of DeepPhoPred and state-of-the-art methods in predicting pS sites: models trained on dataset A and evaluated on dataset B.

Predictors	Sn(%)	Sp(%)	Ac(%)	MCC
NetPhosBac [46]	31.2	67.8	65.1	-0.01
MPSite [25]	39.1	77.9	75	0.11
RotPhoPred [2]	90.0	92.1	91.0	0.82
DeepPhoPred	**100**	**98.9**	**99.0**	**0.93**

Table 10: Performance of DeepPhoPred and state-of-the-art methods in predicting pT sites: models trained on dataset A and evaluated on dataset B.

Predictors	Sn(%)	Sp(%)	Ac(%)	MCC
NetPhosBac [46]	9.4	92.8	85.4	0.03
MPSite [25]	58.3	75.7	74.2	0.22
RotPhoPred [2]	75.4	97.2	86.3	0.74
DeepPhoPred	**99.2**	**99.0**	**99.0**	**0.94**

As shown in this table, the performance of DeepPhoPred is consistent in phosphorylation prediction across all three residues, indicating the high quality and robustness of the models.

To further justify the model's stability, we split the external validation set into 10 parts and measured the performance metrics for each split. Subsequently, each metric's mean and standard deviation are computed and presented in Table 13. As we can observe from the table, the standard deviation is 0 across all metrics, indicating consistent results across the splits.

4 Conclusions

This paper proposes a new computational tool called DeepPhoPred for predicting microbial phosphorylation sites, exploiting protein's evolutionary and structural information. DeepPhoPred is a two-headed CNN architecture embodying squeeze and excitation blocks that can simultaneously learn intrinsic and substantial features from two information sources to detect pS, pT, and pY sites. It produces substantially better results than its peers, as apparent in the comparative results, supporting our starting hypothesis that the fusion of proteins' evolutionary and structural features and application of deep learning techniques can improve microbial phosphorylation site prediction accuracy. This

Table 11: Performance of DeepPhoPred and state-of-the-art methods in predicting pY sites: models trained on dataset A and evaluated on dataset B.

Predictors	Sn(%)	Sp(%)	Ac(%)	MCC
RotPhoPred [2]	78.2	94.7	86.4	0.74
DeepPhoPred	**99.3**	**98.7**	**98.7**	**0.95**

Table 12: Performance of DeepPhoPred on external validation set: models trained on dataset A and evaluated on dataset C.

Task	Sn(%)	Sp(%)	Ac(%)	MCC	auPR
pS	95.0	96.9	96.7	0.87	0.94
pT	96.2	98.4	98.3	0.83	0.96
pY	96.4	97.0	96.9	0.75	0.96

Table 13: Average performance of DeepPhoPred on 10 splits of external validation data: models trained on dataset A and evaluated on dataset B. The figures in the brackets indicate the standard deviation.

Task	Sn(%)	Sp(%)	Ac(%)	MCC
pS	95.07 (0.004)	96.88 (0.000)	96.65 (0.000)	0.87 (0.007)
pT	94.29 (0.079)	98.42 (0.001)	98.24 (0.001)	0.82 (0.052)
pY	95.24 (0.087)	97.08 (0.001)	97.00 (0.000)	0.75 (0.023)

study opens new bases for future research. For our future studies, we aim to develop an interactive and robust web server to access DeepPhoPred with graphical support. We have evaluated the tools' performance on a single dataset. It is vital to examine the performance on more datasets to justify the model's scalability. Moreover, how accurately the models can identify phosphorylation in higher-order organisms is yet to be explored. DeepPhoPred as a standalone predictor, all its source codes, and our employed datasets are publicly available at `https://github.com/faisalahm3d/DeepPhoPred`

Acknowledgements

This article is based on Chapter 4 of Faisal Ahmed's master's thesis [1].

References

[1] Faisal Ahmed. Accurately predicting microbial phosphorylation sites using evolutionary and structural features. Master's thesis, United Internation University, 2022.

[2] Faisal Ahmed, Iman Dehzangi, Md Mehedi Hasan, and Swakkhar Shatabda. Accurately predicting microbial phosphorylation sites using evolutionary and structural features. *Gene*, page 146993, 2022.

[3] Saeed Ahmed, Muhammad Kabir, Muhammad Arif, Zaheer Ullah Khan, and Dong-Jun Yu. Deepppsite: a deep learning-based model for analysis and prediction of phosphorylation sites using efficient sequence information. *Analytical biochemistry*, 612:113955, 2021.

[4] Sajid Ahmed, Rafsanjani Muhammod, Zahid Hossain Khan, Sheikh Adilina, Alok Sharma, Swakkhar Shatabda, and Abdollah Dehzangi. Acp-mhcnn: an accurate multi-headed deep-convolutional neural network to predict anticancer peptides. *Scientific reports*, 11(1):1–15, 2021.

[5] Fasil Ali, Jehangir Shafi Dar, Aqib Rehman Magray, Bashir A Ganai, MZ Chishti, et al. Posttranslational modifications of proteins and their role in biological processes and associated diseases. In *Protein Modificomics*, pages 1–35. Elsevier, 2019.

[6] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[7] Ruhul Amin, Chowdhury Rafeed Rahman, Sajid Ahmed, Md Habibur Rahman Sifat, Md Nazmul Khan Liton, Md Moshiur Rahman, Md Zahid Hossain Khan, and Swakkhar Shatabda. ipromoter-bncnn: a novel branched cnn-based predictor for identifying and classifying sigma promoters. *Bioinformatics*, 36(19):4869–4875, 2020.

[8] Angel M Aponte, Darci Phillips, Robert A Harris, Ksenia Blinova, Stephanie French, D Thor Johnson, and Robert S Balaban. 32p labeling of protein phosphorylation and metabolite association in the mitochondria matrix. *Methods in enzymology*, 457:63–80, 2009.

[9] Fatima Ardito, Michele Giuliani, Donatella Perrone, Giuseppe Troiano, and Lorenzo Lo Muzio. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy. *International journal of molecular medicine*, 40(2):271–280, 2017.

[10] Sean A Beausoleil, Judit Villén, Scott A Gerber, John Rush, and Steven P Gygi. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature biotechnology*, 24(10):1285–1292, 2006.

[11] Ashis Kumer Biswas, Nasimul Noman, and Abdur Rahman Sikder. Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information. *BMC bioinformatics*, 11(1):1–17, 2010.

[12] Nikolaj Blom, Steen Gammeltoft, and Søren Brunak. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of molecular biology*, 294(5):1351–1362, 1999.

[13] Ruifen Cao, Meng Wang, Yannan Bin, and Chunhou Zheng. Dlff-acp: Prediction of acps based on deep learning and multi-view features fusion. *PeerJ*, 9:e11906, 2021.

[14] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[15] Hamed Dashti, Iman Dehzangi, Masroor Bayati, James Breen, Amin Beheshti, Nigel Lovell, Hamid R Rabiee, and Hamid Alinejad-Rokny. Integrative analysis of mutated genes and mutational processes reveals novel mutational biomarkers in colorectal cancer. *BMC bioinformatics*, 23(1):1–24, 2022.

[16] Abdollah Dehzangi, Yosvany López, Sunil Pranit Lal, Ghazaleh Taherzadeh, Abdul Sattar, Tatsuhiko Tsunoda, and Alok Sharma. Improving succinylation prediction accuracy by incorporating the secondary structure via helix, strand and coil, and evolutionary information from profile bigrams. *PloS one*, 13(2):e0191900, 2018.

[17] Abdollah Dehzangi, Sohrab Sohrabi, Rhys Heffernan, Alok Sharma, James Lyons, Kuldip Paliwal, and Abdul Sattar. Gram-positive and gram-negative subcellular localization using rotation forest and physicochemical-based features. *BMC bioinformatics*, 16(4):1–8, 2015.

[18] Yujuan Deng, Yunfang Fu, Huitao Zhang, Xin Liu, and Zhiguo Liu. Protein post-translational modification site prediction using deep learning. *Procedia Computer Science*, 198:480–485, 2022.

[19] Shubhashis Roy Dipta, Ghazaleh Taherzadeh, MD Wakil Ahmad, MD Easin Arafat, Swakkhar Shatabda, and Abdollah Dehzangi. Semal: Accurate protein malonylation site predictor using structural and evolutionary information. *Computers in biology and medicine*, 125:104022, 2020.

[20] Yongchao Dou, Bo Yao, and Chi Zhang. Phosphosvm: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. *Amino acids*, 46(6):1459–1469, 2014.

[21] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012.

[22] Jianjiong Gao, Jay J Thelen, A Keith Dunker, and Dong Xu. Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Molecular & Cellular Proteomics*, 9(12):2586–2600, 2010.

[23] Ali Ghulam, Farman Ali, Rahu Sikander, Ashfaq Ahmad, Aftab Ahmed, and Shruti Patil. Acp-2dcnn: deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network. *Chemometrics and Intelligent Laboratory Systems*, 226:104589, 2022.

[24] Robert Harmel and Dorothea Fiedler. Features and regulation of non-enzymatic post-translational modifications. *Nature chemical biology*, 14(3):244–252, 2018.

[25] Md Hasan, Md Rashid, Mst Khatun, Hiroyuki Kurata, et al. Computational identification of microbial phosphorylation sites by the enhanced characteristics of sequence information. *Scientific reports*, 9(1):1–9, 2019.

[26] Md Mehedi Hasan and Mst Shamima Khatun. Prediction of protein post-translational modification sites: an overview. *Ann Proteom Bioinform*, 2:049–57, 2018.

[27] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE, 2008.

[28] Rhys Heffernan, Kuldip Paliwal, James Lyons, Abdollah Dehzangi, Alok Sharma, Jihua Wang, Abdul Sattar, Yuedong Yang, and Yaoqi Zhou. Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning. *Scientific reports*, 5(1):1–11, 2015.

[29] Ying Huang, Beifang Niu, Ying Gao, Limin Fu, and Weizhong Li. Cd-hit suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5):680–682, 2010.

[30] LN Johnson and D Barford. The effects of phosphorylation on the structure and function of proteins. *Annual review of biophysics and biomolecular structure*, 22(1):199–232, 1993.

[31] Md Saikat Islam Khan, Anichur Rahman, Tanoy Debnath, Md Razaul Karim, Mostofa Kamal Nasir, Shahab S Band, Amir Mosavi, and Iman Dehzangi. Accurate brain tumor detection using deep convolutional neural network. *Computational and Structural Biotechnology Journal*, 20:4733–4745, 2022.

[32] Md Saikat Islam Khan, Anichur Rahman, Md Razaul Karim, Nasima Islam Bithi, Shahab S Band, Abdollah Dehzangi, and Hamid Alinejad-Rokny. Covidmulti-net: a parallel-dilated multi

scale feature fusion architecture for the identification of covid-19 cases from chest x-ray images. *medRxiv*, 2021.

[33] Zaheer U Khan and Dechang Pi. Deepsspred: A deep learning based sulfenylation site predictor via a novel nsegmented optimize federated feature encoder. *Protein and Peptide Letters*, 28(6):708–721, 2021.

[34] Jeehong Kim, Muhammad Shujaat, and Hilal Tayara. iprom-zea: A two-layer model to identify plant promoters and their types using convolutional neural network. *Genomics*, 114(3):110384, 2022.

[35] Nguyen Quoc Khanh Le and Quang-Thai Ho. Deep transformers and convolutional neural network in identifying dna n6-methyladenine sites in cross-species genomes. *Methods*, 204:199–206, 2022.

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

[37] Zhengpeng Li, Ping Wu, Yuanyuan Zhao, Zexian Liu, and Wei Zhao. Prediction of serine/threonine phosphorylation sites in bacteria proteins. In *Advance in Structural Bioinformatics*, pages 275–285. Springer, 2015.

[38] Yang Liu, Duolin Wang, Fei He, Juexin Wang, Trupti Joshi, and Dong Xu. Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Frontiers in genetics*, 10:1091, 2019.

[39] Yosvany López, Alok Sharma, Abdollah Dehzangi, Sunil Pranit Lal, Ghazaleh Taherzadeh, Abdul Sattar, and Tatsuhiko Tsunoda. Success: evolutionary and structural properties of amino acids prove effective for succinylation site prediction. *BMC genomics*, 19(1):105–114, 2018.

[40] Fenglin Luo, Minghui Wang, Yu Liu, Xing-Ming Zhao, and Ao Li. Deepphos: prediction of protein phosphorylation sites with deep learning. *Bioinformatics*, 35(16):2766–2773, 2019.

[41] James Lyons, Abdollah Dehzangi, Rhys Heffernan, Alok Sharma, Kuldip Paliwal, Abdul Sattar, Yaoqi Zhou, and Yuedong Yang. Predicting backbone cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *Journal of computational chemistry*, 35(28):2040–2046, 2014.

[42] Wenlong Ma, Zhixu Qiu, Jie Song, Jiajia Li, Qian Cheng, Jingjing Zhai, and Chuang Ma. A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta*, 248(5):1307–1318, 2018.

[43] Inderjeet Mani and I Zhang. knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, volume 126, pages 1–7. ICML, 2003.

[44] Fereshteh Mataeimoghadam, MA Newton, Abdollah Dehzangi, Abdul Karim, B Jayaram, Shoba Ranganathan, and Abdul Sattar. Enhancing protein backbone angle prediction by using simpler models of deep neural networks. *Scientific reports*, 10(1):1–12, 2020.

[45] Md Ochiuddin Miah, Rafsanjani Muhammod, Khondaker Abdullah Al Mamun, Dewan Md Farid, Shiu Kumar, Alok Sharma, and Abdollah Dehzangi. Clusem: Accurate clustering-based ensemble method to predict motor imagery tasks from multi-channel eeg data. *Journal of Neuroscience Methods*, 364:109373, 2021.

[46] Martin Lee Miller, Boumediene Soufi, Carsten Jers, Nikolaj Blom, Boris Macek, and Ivan Mijakovic. Netphosbac–a predictor for ser/thr phosphorylation sites in bacterial proteins. *Proteomics*, 9(1):116–125, 2009.

[47] Qiao Ning, Xiaowei Zhao, and Zhiqiang Ma. A novel method for identification of glutarylation sites combining borderline-smote with tomek links technique in imbalanced data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021.

[48] Zhicheng Pan, Bangshan Wang, Ying Zhang, Yongbo Wang, Shahid Ullah, Ren Jian, Zexian Liu, and Yu Xue. dbpsp: a curated database for protein phosphorylation sites in prokaryotes. *Database*, 2015, 2015.

[49] Christopher G Proud. Phosphorylation and signal transduction pathways in translational control. *Cold Spring Harbor perspectives in biology*, 11(7):a033050, 2019.

[50] Chowdhury Rafeed Rahman, Ruhul Amin, Swakkhar Shatabda, Md Toaha, and Sadrul Islam. A convolution based computational approach towards dna n6-methyladenine site identification and motif extraction in rice genome. *Scientific Reports*, 11(1):1–13, 2021.

[51] Shahin Ramazi and Javad Zahiri. Post-translational modifications in proteins: resources, tools and prediction methods. *Database*, 2021, 2021.

[52] Md M Rashid, Swakkhar Shatabda, Md Hasan, Hiroyuki Kurata, et al. Recent development of machine learning methods in microbial phosphorylation sites. *Current Genomics*, 21(3):194–203, 2020.

[53] Farshid Rayhan, Sajid Ahmed, Swakkhar Shatabda, Dewan Md Farid, Zaynab Mousavian, Abdollah Dehzangi, and M Sohel Rahman. idti-esboost: identification of drug target interaction using evolutionary and structural features with boosting. *Scientific reports*, 7(1):1–18, 2017.

[54] Alok Sharma, Kuldip K Paliwal, Abdollah Dehzangi, James Lyons, Seiya Imoto, and Satoru Miyano. A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition. *BMC bioinformatics*, 14(1):1–11, 2013.

[55] Swakkhar Shatabda, Sanjay Saha, Alok Sharma, and Abdollah Dehzangi. iphloc-es: Identification of bacteriophage protein locations using evolutionary and structural features. *Journal of theoretical biology*, 435:229–237, 2017.

[56] Muhammad Shujaat, Seung Beop Lee, Hilal Tayara, and Kil To Chong. Cr-prom: A convolutional neural network-based model for the prediction of rice promoters. *IEEE Access*, 9:81485–81491, 2021.

[57] Jiangning Song, Huilin Wang, Jiawei Wang, André Leier, Tatiana Marquez-Lago, Bingjiao Yang, Ziding Zhang, Tatsuya Akutsu, Geoffrey I Webb, and Roger J Daly. Phosphopredict: A bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. *Scientific reports*, 7(1):1–19, 2017.

[58] Laurey Steinke and Richard G Cook. Identification of phosphorylation sites by edman degradation. In *Protein Sequencing Protocols*, pages 301–307. Springer, 2003.

[59] Niraj Thapa, Meenal Chaudhari, Anthony A Iannetta, Clarence White, Kaushik Roy, Robert Newman, Leslie M Hicks, and KC Dukka. Chlamy-enphossite: A deep learning-based approach for chlamydomonas reinhardtii-specific phosphorylation site prediction. 2021.

[60] Patrick Brendan Timmons and Chandralal M Hewage. Ennaact is a novel tool which employs neural networks for anticancer activity classification for therapeutic peptides. *Biomedicine & Pharmacotherapy*, 133:111051, 2021.

[61] Ivan Tomek. Two modifications of cnn. 1976.

[62] Duolin Wang, Yanchun Liang, and Dong Xu. Capsule network for protein post-translational modification site prediction. *Bioinformatics*, 35(14):2386–2394, 2019.

[63] Duolin Wang, Dongpeng Liu, Jiakang Yuchi, Fei He, Yuexu Jiang, Siteng Cai, Jingyi Li, and Dong Xu. Musitedeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Research*, 48(W1):W140–W146, 2020.

[64] Leyi Wei, Pengwei Xing, Jijun Tang, and Quan Zou. Phospred-rf: a novel sequence-based predictor for phosphorylation sites using sequential information only. *IEEE transactions on nanobioscience*, 16(4):240–247, 2017.

[65] Haodong Xu, Yongbo Wang, Shaofeng Lin, Wankun Deng, Di Peng, Qinghua Cui, and Yu Xue. Ptmd: a database of human disease-associated post-translational modifications. *Genomics, proteomics & bioinformatics*, 16(4):244–251, 2018.

[66] Hangyuan Yang, Minghui Wang, Xia Liu, Xing-Ming Zhao, and Ao Li. Phosidn: an integrated deep neural network for improving protein phosphorylation site prediction by combining sequence and protein–protein interaction information. *Bioinformatics*, 37(24):4668–4676, 2021.

[67] Haoyang Zeng, Matthew D Edwards, Ge Liu, and David K Gifford. Convolutional neural network architectures for predicting dna–protein binding. *Bioinformatics*, 32(12):i121–i127, 2016.

[68] Qing-bin Zhang, Kai Yu, Zekun Liu, Dawei Wang, Yuanyuan Zhao, Sanjun Yin, and Zexian Liu. Prediction of prkc-mediated protein serine/threonine phosphorylation sites for bacteria. *PloS one*, 13(10):e0203840, 2018.

[69] Shimin Zhao, Wei Xu, Wenqing Jiang, Wei Yu, Yan Lin, Tengfei Zhang, Jun Yao, Li Zhou, Yaxue Zeng, Hong Li, et al. Regulation of cellular metabolism by protein lysine acetylation. *Science*, 327(5968):1000–1004, 2010.