Gene 851 (2023) 146993

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Accurately predicting microbial phosphorylation sites using evolutionary and structural features

Faisal Ahmed [a,b], Iman Dehzangi [c,d,*], Md. Mehedi Hasan [e], Swakkhar Shatabda [a,*]

[a] Department of Computer Science and Engineering, United International University, Dhaka, Bangladesh
[b] Department of Computer Science and Engineering, Premier University, Chattogram, Bangladesh
[c] Department of Computer Science, Rutgers University, Camden, NJ 08102, USA
[d] Center for Computational and Integrative Biology (CCIB), Rutgers University, Camden, NJ 08102, USA
[e] Tulane Center for Biomedical Informatics and Genomics, Division of Biomedical Informatics and Genomics, John W. Deming Department of Medicine, School of Medicine, Tulane University, New Orleans, LA 70112, USA

ARTICLE INFO

Edited by Lakshminarayan M. Iyer

Keywords:
Post translational modification
Phosphorylation
Evolutionary features
Structural features
Classification
Machine learning

ABSTRACT

Post-translational modification (PTM) is a biological process involving a protein's enzymatic changes after its translation by the ribosome. Phosphorylation is one of the most critical PTMs that occurs when a phosphate group interacts with an amino acid residue along protein sequence. It contributes to cell communication, DNA repair, and gene regulation. Predicting microbial phosphorylation sites can provide better understanding of host-pathogen interaction and the development of anti-microbial agents. Experimental methods such as mass spectrometry are time-consuming, laborious, and expensive. This paper proposes a new approach, called RotPhoPred, for predicting phospho-serine (pS), phospho-threonine (pT), and phospho-tyrosine (pY) sites in the microbial organism by integrating evolutionary bigram profile with structural information and using Rotation Forest as the classification technique. To the best of our knowledge, our extracted features and employed classifier have never been utilized for this task. Comparative results demonstrate that the RotPhoPred surpasses its peers in terms of different metrics such as sensitivity (90.0%, 75.4% and 78.2%), specificity (92.1%, 97.2% and 94.7%), accuracy (91.0%, 86.3%, 86.4%), and MCC (0.82, 0.74 and 0.74) for pS, pT, and pY sites predictions, respectively. RotPhoPred as a standalone predictor and all its source codes are publicly available at: https://github.com/faisalahm3d/RotPredPho.

1. Introduction

Post-translational modification is a biological mechanism in which one or more amino acids of a protein interact with a specific molecular group after its translation process by the ribosome (Rashid et al., 2020). Phosphorylation is one of the most critical and common PTMs. It occurs when a phosphate group is added to an amino acid residue. It most commonly appears in serine (S), threonine(T), and tyrosine(Y). It also happens in arginine, lysine, and histidine residues to a lesser extent (Jamal et al., 2021). Phosphorylation plays an essential role in a wide range of cellular functions, including cell communication, DNA repair, and gene regulation in eukaryote and microbial organisms (Trost and Kusalik, 2011; Chen et al., 2020). Phosphorylation causes dysregulation of cell signalling mechanisms, which results in the development and progress of complex diseases like cancer (Chen and Eschrich, 2014). For example, p53 is a protein where multiple phosphorylation sites are observed to be responsible for tumor development (Loughery and Meek, 2013). Identification of Phosphorylation in prokaryotes cells can provide crucial information for a better understanding of host-pathogen interactions and the development of antimicrobial agents (Shi et al., 2020). Liquid chromatography-tandem mass spectrometry (LC-MS/MS), radioactive chemical labelling, and western blotting are the most common experimental methods for identifying PTMs, including phosphorylation. However, experimental approaches for detecting PTMs are time-consuming, tedious, expensive, and require a skilled workforce. Moreover, the number of protein sequences is increasing exponentially due to advanced sequencing technologies. Therefore, it is unfeasible to identify phosphorylation sites using experimental methods in the wet lab from such a massive protein database. Hence, there is a crucial demand for developing fast and accurate computational tools to identify phosphorylation sites.

During the past few years, several machine learning-based predictors have been proposed to predict phosphorylation sites. The most

* Corresponding authors at: Department of Computer Science, Rutgers University, Camden, NJ 08102, USA (I. Dehzangi).

https://doi.org/10.1016/j.gene.2022.146993
Received 28 July 2022; Received in revised form 5 October 2022; Accepted 14 October 2022
Available online 20 October 2022

Abbreviations list

ASA	acessible surface area
CNN	convolutional neural network
MCC	Mathews correlation coefficient
GBT	gradient boosting trees
LC-MS/MS	Liquid chromatography-tandem mass spectrometry
PTM	Post Translational Modifications
pS	phospho-serine
pT	phospho-threonine
pY	phospho-tyrosine
SVM	Support vector machines
RF	Random Forests
WR	Wilcoxon rank-sum test

promising predictors are PhosPred-RF (Wei et al., 2017), PhosphoSVM (Dou et al., 2014), NetPhos (Blom et al., 1999), PRED (Biswas et al., 2010), Musite (Gao et al., 2010). These methods have used features extracted from proteins' primary sequence or secondary structure. For example, PhosPred-RF (Wei et al., 2017) and PhosphoSVM (Dou et al., 2014) use sequence-based features, whereas PPRED (Biswas et al., 2010) uses evolutionary information to identify phosphorylation sites. Sequence and structural features are combined in NetPhos (Blom et al., 1999) for independent and kinase-specific phosphorylation site prediction. PhosphoPredict (Song et al., 2017) integrates sequence-based and functional features to identify kinase-specific substrates and their corresponding phosphorylation sites. They also used different classification techniques including support vector machines (SVM), Random Forests (RF), gradient boosting trees (GBT), and AdaBoost to build their models.

More recently, several deep learning-based predictors have been proposed to predict Phosphorylation sites. Manual feature extraction is unnecessary for the deep learning-based approaches since they can automatically retrieve complicated patterns from protein sequences. MusiteDeep (Wang et al., 2020), DeepPPSite (Ahmed et al., 2021), DeepPhos (Luo et al., 2019), and Chlamy-EnPhosSite (Thapa and Chaudhari, 2021) are notable deep learning-based Phosphorylation site predictors. MusiteDeep uses one-hot encoding of protein sequence and convolutional neural network (CNN) with attention layer (Wang et al., 2020). DeepPhos utilizes multi-layer CNN architecture consisting of densely connected convolutional blocks with different window and filter sizes (Luo et al., 2019). DeepPPsite is constructed using a stacked long-short-term memory recurrent network (Ahmed et al., 2021), whereas Chlamy-EnPhosSite is an ensemble-based organism-specific predictor developed by combining CNN and LSTM (Thapa and Chaudhari, 2021). DeepPPSite combines five distinct sequence-encoding approaches namely, sequence location information, amino acid composition descriptors, grouped-based features, and physicochemical property-based features. Unlike MisiteDeep and DeepPhos, where binary encoding is used, the embedding layer is employed in Chlamy-EnPhosSite to encode protein sequences.

Among all these approaches, only four computational methods for predicting phosphorylation sites in microbial organisms are available to date. The initial two methods NetPhosBac (Lee Miller et al., 2009) and cPhosBac (Li et al., 2015), are bacteria-specific protein phosphorylation site predictors. The former is created by implementing an artificial neural network algorithm. The latter utilizes k-spaced amino acid pairs (KSAAP) composition for sequence encoding and SVM for classification. The predictors are trained on the same dataset, consisting of 152 experimentally confirmed phosphorylated serine/threonine sites in 119 substrates. The cPhosBac outperforms the NetPhosBac. On the other hand, prkC-PSP was proposed by Zhang et al. as a prkC-specific phosphorylation site predictor (Zhang et al., 2018). It extracts amino acid location information-based features from the protein sequence and use SVM as the classification technique to distinguish probable prkC-specific phosphorylation sites. The dataset contains experimentally identified 36 phosphorylation and 512 non-phosphorylation sites curated manually from the literature. In 2019, Mamun et al., developed a general microbial phosphorylation site predictor named MPsite by using enhanced characteristics of sequence as features and Random Forest as the classification technique (Md Hasan et al., 2019). To build this model, they used Wilcoxon rank-sum test (WR) to select the optimal set of features. The dataset used in this study was collected from the dbPSP, consisting of 2045 pS sites in 1940 proteins and 2174 pT sites in 1534 proteins. MPsite shows more promising performance than the existing microbial phosphorylation site predictors.

Despite all the efforts that have been made so far, there is still room for improving microbial phosphorylation site prediction accuracy. We have observed that only composition-based features were used in the previous studies to predict microbial phosphorylation sites. However, previous research on protein subcellular localization (Dehzangi et al., 2015), bacteriophage protein identification (Shatabda et al., 2017), and protein succinylation and malonylation prediction (Roy Dipta et al., 2020; Dehzangi et al., 2018) have shown that extracting structural and evolutionary information greatly improves prediction performance. Hence, we hypothesize that integrating these features can improve microbial phosphorylation site prediction as well.

In this study, we propose a new machine learning-based predictor called RotPhoPred to accurately predict phospho-serine (pS), phospho-threonine (pT), and phospho-tyrosine (pY) in the protein sequence of microbial organisms, which integrates both structural and evolutionary information. Our predictors coalesce predicted structural features and evolutionary bigram profiles to describe each peptide fragment in the dataset. We also use the NearMiss-3 undersampling technique to balance the dataset to avoid bias towards larger class set. Subsequently, we use Rotation Forest classifier which is an ensemble-based machine learning classifier to predict microbial phosphorylation on serine (S), threonine (T), and tyrosine (Y) residues. We then use 5-folds cross-validation and independent test set to assess the prediction performance of the predictors. The overall flowchart of the proposed method is shown in Fig. 1.

Our results show that RotPhoPred outperforms the existing predictors (NetPhosBac and MPsite). It achieves 90.0%, 91.4%, 91.5%, and 0.82, in terms of Sensitivity, Specificity, Accuracy, and Mathews correlation coefficient (MCC) for predicting pS sites, respectively. It also achieves 75.4%, 99.2%, 86.3%, and 0.74 in terms of sensitivity, specificity accuracy, and MCC for predicting pT sites, respectively. The recorded sensitivity, specificity, accuracy, and MCC for the pY site prediction are 78.2%, 94.7%, 86.4%, and 0.74, respectively. RotPhoPred as a standalone predictor and all its source codes are publicly available at: https://github.com/faisalahm3d/RotPredPho.

The significant contributions of this paper are as follows:

1. The paper proposes the fusion of the evolutionary bigram profile with structural information as features and the utilization of Rotation Forest as the classifier for the first time for microbial phosphorylation prediction.
2. It is the first study for predicting microbial phospho-tyrosine (pY) sites. Previous studies focused only on phosphorylation prediction on serine (S) and threonine (T) residues.
3. Our proposed predictor is generic, which can predict phospho-serine (pS), phospho-threonine (pT), and phospho-tyrosine (pY) sites applying the same feature and classifier.
4. We have conducted extensive experiments on the benchmark datasets of laboratory-verified phosphorylated sites to validate the effectiveness and applicability of the proposed predictor.
5. The predictor can maintain an excellent balance between sensitivity and specificity in a highly imbalanced dataset, as apparent in the experimental results.
6. We have publicly shared our dataset and model so that researchers can quickly reproduce the results for further experiments and

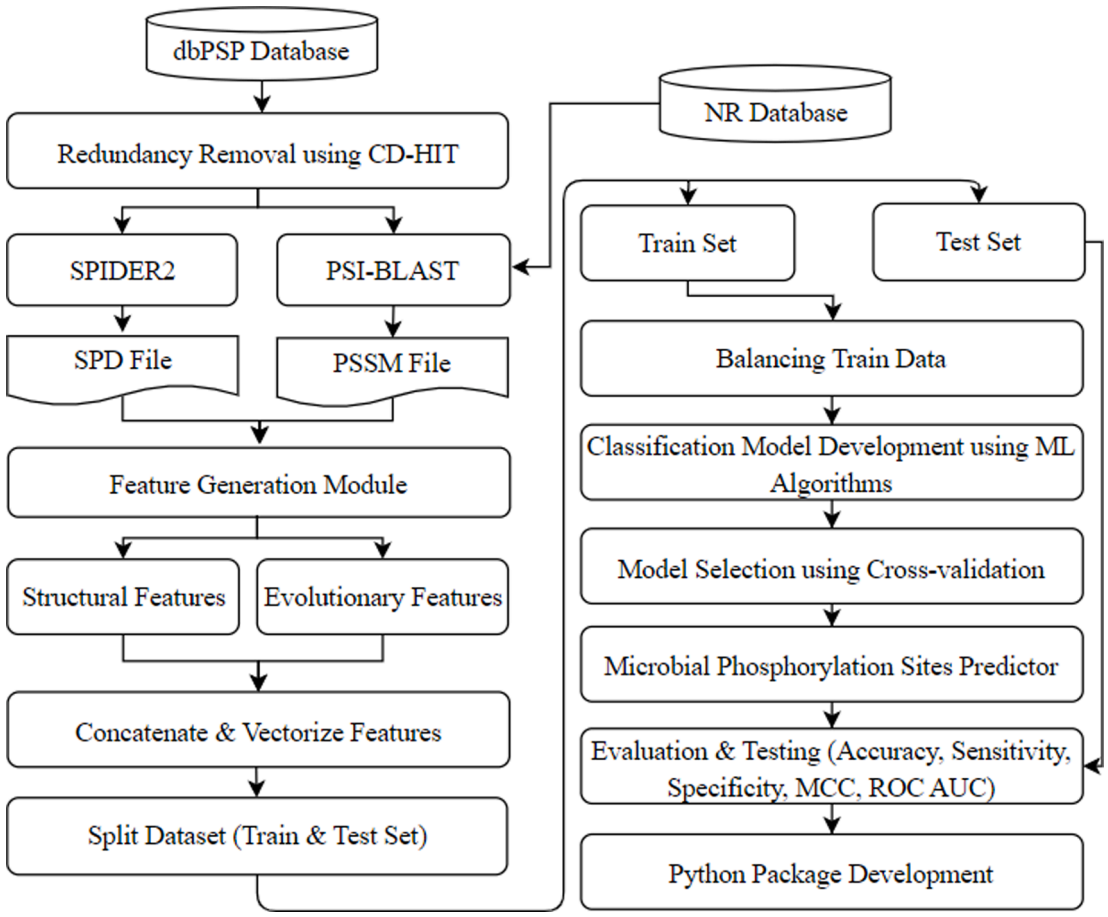

Fig. 1. The overall flowchart of the proposed method.

biologists can easily access the predictor for the initial screening of phosphorylation sites.

2. Material and methods

In this section, we describe the benchmark dataset that is used in this study and present our proposed methodology to build RotPhoPred.

2.1. Benchmark datasets

We have collected the protein sequences with experimentally verified pS, pT, and pY PTMs from the dbPSP database (Shi et al., 2020). The assembled dataset contains redundant proteins. To remove redundancy, we use CD-HIT to remove those proteins with over 40% sequential similarity (Limin et al., 2012; Li and Godzik, 2006; Huang et al., 2010). The final datasets contain 1483, 1220, and 1161 protein sequences for pS, pT, and pY PTMs, respectively. The pT dataset consists of a total 36,513 instances with 2024 phosphorylated (positive) and 34489 non-phosphorylated (negative) sites. There are 26239 samples in the pT dataset, with 1647 positive and 24592 negative sites. The pY dataset comprises of total 17476 instances where 1644 are positive and 15832 are negative. A summary of the datasets used in this study is given in Table 1.

Table 1
The summary of the datasets for pS, pT, and pY identification problems.

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

To avoid overfitted and assess the generality of our model, 10% of the datasets are used to form the independent test sets. The remaining 90% of the datasets are used to train the classifiers (training dataset).

2.2. Features

To extract the evolutionary and structural information, the protein sequences from the benchmark dataset are fed to PSI-BLAST (Altschul et al., 1997) and SPIDER2 (Yang et al., 2017; Heffernan et al., 2015). Using PSI-BLAST, we generate position-specific scoring matrix (PSSM) file, and using SPIDER2, we generate an SPD file. The PSSM calculates the likelihood of replacing each protein's amino acid with the other 20 amino acids based on their location. On the other hand, the predicted secondary structure probabilities, accessible surface area (ASA), and torsion angles for each amino acid residue are described in the SPD file in matrix format. The evolutionary bigram profile and structural features are then constructed from the PSSM and SPD files. Later, the different feature groups extracted from PSSM and SPD files are combined to form a feature vector. After feature vectorization, we split the data into training and test sets. A NearMiss-3 under-sampling technique is then applied to the training set to address the imbalance issue while keeping the test data untouched (Mani and Zhang, 2003).

2.3. Formulation of peptide fragments for each site

In this study, we used a window-based approach to represent each positive or negative site. Each phosphorylation or non-phosphorylation site is described by a peptide fragment P of $2n + 1$ residues with a S/T/Y in the center, n upstream residues at the right, and n downstream

residues at the left as follows:

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

where A_{-i} and A_{+i} represent the upstream and downstream amino acids respectively, and S, T, Y represents serine, threonine, and tyrosine, respectively. If n upstream or downstream residues are not available in the protein to describe the S/T/Y sites, the mirroring technique is used to fill the gap of missing amino acids as shown in Fig. 2.

After analyzing the performance of different window sizes, we choose $n = 10$ since it exhibits the best performance. Fig. 2 demonstrates the overall windowing process for serine residue.

2.4. Evolutionary feature extraction

As it was mentioned earlier, we use PSSM to extract evolutionary information. We generated the PSSM for each protein sequence in our benchmark dataset by running the PSI-BLAST algorithm for three iterations on the non-redundant (nr) database provided by NCBI with a cutoff (E) value of 0.001. The PSSM is an $Lx20$ matrix, where L is the protein sequence's length, and the 20 columns denote different amino acids of the genetic code.

It was shown in previous studies that using bigram, we can extract important disciriminatory information for the classification task from PSSM for similar problems (Sharma et al., 2013; Roy Dipta et al., 2020; Dehzangi et al., 2018; Chandra et al., 2019; Ahmad et al., 2020). Moreover, bigram feature size is independent from the window. It extracts a 400-dimensional feature vector to capture evolutionary information regardless of the number of upstream and downstream residues. As a result, we may expand the number of residues surrounding the S/T/Y site without increasing the number of features. This study generates bigram probabilities for each protein segment to apprehend its evolutionary profile. To generate the bigram profile from the evolutionary information, the submatrix M for the peptide fragment P that describes a phosphorylated or non-phosphorylated site is segmented from the PSSM matrix. M is a $W * 20$ dimensional matrix where W is the window size (W = 21 consisting of 10 upstream, 10 downstream, and one central S/T/Y residue), as mentioned in Section 2.3. Each element $m_{i,j}$ of the matrix M represents the transitional probability of *jth* amino acid at the *ith* position in the peptide fragment P. Then the bigram profile of the submatrix is calculated using the following equation:

$$B_{p,q} = \sum_{k=1}^{20} m_{k,p} m_{k+1,q}, \ \ where \ 1 \leqslant p \leq 20 \ and \ 1 \leqslant p \leq 20 \quad (2)$$

The resulting 20x20 dimensional matrix B represents the PSSM profile bigram of peptide fragment P. Subsequently, the matrix B is converted to a 400-dimensional row vector denoted by F1 as shown in Eq. 3.

$$F1 = [B_{1,1}, B_{1,2}, \ldots, B_{1,20}, B_{2,1}, B_{2,2}, \ldots, B_{2,20}, \ldots, B_{20,1}, B_{20,2}, \ldots, B_{20,20}] \quad (3)$$

2.5. Structural feature extraction

Along with the evolutionary information, the structural properties of the protein have been shown to be effective to predict other PTMs (Dehzangi et al., 2018; Islam et al., 2018; Reddy et al., 2019; Dehzangi et al., 2013; Chowdhury et al., 2017; Roy Dipta et al., 2020). The protein's structural properties include secondary structures, accessible surface area (ASA), and torsion angles.

The secondary structure depicts each amino acid residue in a number of distinct configurations, the most frequent of which are helix, sheet, and coil. Local backbone angles also define the local protein structures through torsion angles between neighboring amino acids. Unlike secondary structure, which provides a coarse-grain description of local configuration in terms of 3 discrete shapes - coil, strand, or helix, local backbone angles give continuous information about the local structure concerning four angles. The four angles include two backbone torsion angles, ψ, and ϕ, which indicate the angles between atoms along the protein backbone, and dihedral angles θ and τ, which represent the rotation angles. The secondary structure and backbone angles describe which amino acids are more dissembled and prone to interact with other macromolecules.

ASA measures how much an amino acid residue area is exposed to solvent (water) in a protein. The amino acid residue on the protein's surface area has a high chance of undergoing PTMs. Hence, ASA is an essential structural property for phosphorylation prediction. As it was mentioned earlier, we used SPIDER2 to predict the values of the parameters mentioned above for each amino acid residue in a protein

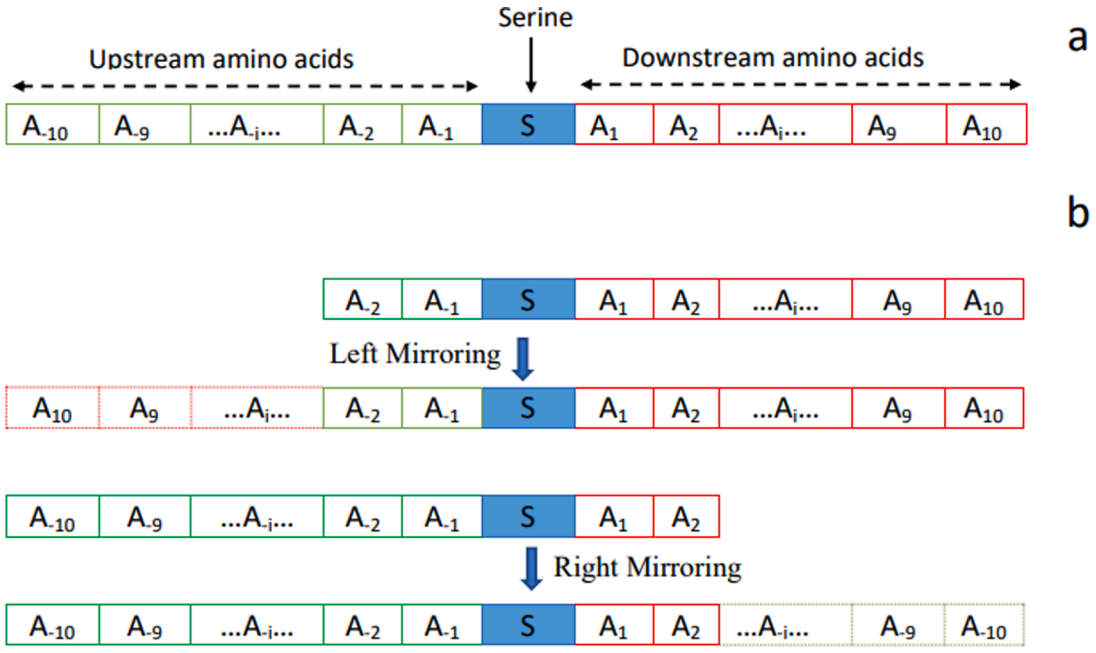

Fig. 2. The representation of serine residue with its upstream and downstream amino acids. (a) Serine residue with enough neighbors on the upstream and downstream sides. (b) Serine residue with inadequate neighbors, either upstream or downstream.

sequence. SPIDER is a deep learning-based tool that achieves promising results for predicting secondary structure, backbone angles, and ASA from protein sequences (Yang et al., 2017; Heffernan et al., 2015). It produces a matrix of $L * 8$ dimensions that contain 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.

To capture the structural information of each peptide segment P in our dataset, we extract the submatrix A from the SPD file of the protein sequence that contains P and flattens it to form a row vector F2. The matrix A is $W * 8$ dimensional, where W is the window size. We investigated different values for window size which among them, using 21 for serine (S), threonine (T), and tyrosine (Y) demonstrates the best performance. Hence, A is a 21x8 dimensional matrix. The resultant row vector is 168 dimensional denoted by F2 as shown in Eq. 4.

$$F2 = [B'_{1,1}, B'_{1,2}, ..., B'_{1,8}, B'_{2,1}, B'_{2,2}, ..., B'_{2,8}, ...B'_{21,1}, B'_{21,2}, ..., B'_{21,8}] \quad (4)$$

2.6. Formation of feature vector

After extracting the bigram profile from PSSM and structural features from SPD, we integrate both feature groups to form a feature vector to predict microbial phosphorylation. The resultant 560 (400 + 168) dimensional feature vector captures the critical structural and evolutionary information essential to discriminating phosphorylated and non-phosphorylated sites.

$$F = [B_{1,1}, ..., B_{1,20}, ..., B_{20,1}, ..., B_{20,20}, B'_{1,1}, ..., B'_{1,8}, ..., B'_{21,1}, ..., B'_{21,8}] \quad (5)$$

2.7. Balancing dataset

The number of non-phosphorylated sites (negative samples) is greater than the number of phosphorylated sites (positive samples) in our benchmark datasets, as shown in Table 1. Such imbalance could influence any machine learning-based predictor to be biased towards the negative sample. Therefore, balancing the training dataset is critical for developing a bias-free predictor. Two main ways to balance the dataset are under-sampling and over-sampling. While the former keeps all samples in the rare class and reduces the abundant type, the latter increase the size of the infrequent category by generating artificial instances. This paper analyzes various balancing techniques from both under-sampling and over-sampling categories, including ADASYS (He et al., 2008), SMOTE (Chawla et al., 2002), Tomek Links (Tomek, 1976), and NearMiss (Mani and Zhang, 2003). Our results demonstrate that NearMiss-3 which is a down-sampling technique exhibited the best performance. Hence, we used the NearMiss-3 technique for balancing our training dataset.

NearMiss-3 selects the given number of closet samples from the majority class (negative) for each instance in the minority class based on the Euclidean distance. Consequently, it picks the more information-rich non-phosphorylated sites, which are vital for designing a powerful decision boundary to differentiate phosphorylated and non-phosphorylated sites (Mani and Zhang, 2003). Moreover, it does not produce artificial samples thus minimizing the computational cost when fitting the model. We implemented NearMiss-3 to select one closet non-phosphorylated site for each phosphorylated site. As a result, the transformed training dataset includes an equal number of phosphorylated and non-phosphorylated sites. Note that the balancing is not performed on the independent test set used to evaluate the performance of the model to avoid overfitting.

2.8. Classification model

In this study, to build RotPhoPred, we use Rotation Forest (RoF) algorithm since it exhibits encouraging performance in similar studies found in the literature (Dehzangi et al., 2015; Roy Dipta et al., 2020; Dehzangi et al., 2010; Bustamam et al., 2019; Wang et al., 2018; Wang et al., 2018; You et al., 2017). Rotation forest is an ensemble learning technique that trains N base classifiers separately in parallel and predicts class labels based on the majority of soft voting (Rodriguez et al., 2006). Unlike Random Forest that uses a random subset of features, the rotation forest uses a transformed feature space to build the individual base learner. To build each base classifier, it randomly splits the feature set into K subsets, and for each subset, a bootstrap sample of size 75% of the original dataset is drawn. Then the Principal Component Analysis (PCA) (Abdi and Williams, 2010) is then performed on the selected samples to transform the feature vector linearly, to enhance diversity among the base learners. Later, the K transformed feature subsets are combined to form a feature vector to train the base classifier. RoF utilizes decision trees as the base learners since they are accurate and sensitive to the rotation of the feature axes. In this study, we used the rotation_forest package available in python3 with 100 decision trees as base estimators as it was shown the effective number in previous studies (Dehzangi et al., 2015; Roy Dipta et al., 2020). The max_features parameter was set to 'auto' and n_jobs, indicating the number of jobs to run parallel, was set to -1 to force the algorithm to use all the functional processors.

2.9. Validation scheme

A wide range of validation schemes, including the k-folds cross-validation and jack-knife test, are reported in the literature to evaluate the efficacy of machine learning-based predictors. In this paper, we have utilized the stratified 5-folds cross-validation to avoid overfitting and assess our performance with different parameter settings. The cross-validation is performed as follows:

1. Divide the dataset into five disjoint folds of equal size by maintaining the percentage of instances from both classes in each fold.
2. Fit the predictor on 4-folds, and evaluate its performance on the remaining fold via different metrics such as sensitivity, specificity accuracy, and MCC.
3. Repeat step 2 five times and calculate each metric's average.

Fig. 3 graphically demonstrates the overall process of the 5-fold cross-validation scheme.

2.10. Evaluation metrics

To evaluate the performance of RotPhoPred, we use six metrics namely, Sensitivity (Sn), Specificity (Sp), Precision (Pr), Accuracy (Ac), F1 Measure (F1), Mathews' correlation coefficient (MCC), and area under the ROC curves (AUC) to evaluate the performance of the proposed method.

Sensitivity measures the predictor's ability to identify phosphorylated sites accurately. It quantifies how many phosphorylated sites the predictor can accurately detect out of the total number of phosphorylated sites. Sensitivity ranges from 0 to 100 percent. The predictor detecting all phosphorylated sites will receive sensitivity of 100.

Specificity assesses how well the predictor performs in detecting non-phosphorylated sites. Hence, it is the ratio of the total number of successfully identified non-phosphorylated sites by the predictor to the actual number of non-phosphorylated sites. The value of specificity can be between 0 to 100 percent.

The predictor's ability to discriminate between phosphorylated and non-phosphorylated sites is measured by accuracy. It summarizes the predictor's overall performance with a single score. Accuracy ranges from 0% to 100%, with 100% indicating the most accurate prediction.

Precision is the ratio of correctly predicted phosphorylated sites to all predicted phosphorylated sites by the model.

F-Measure summarizes the sensitivity and precision through a single measure by computing their harmonic mean. The score reflects how well the predictor can balance the sensitivity and precision. It also ranges from 0 to 1, with 1 denoting perfect balance. It is the most common

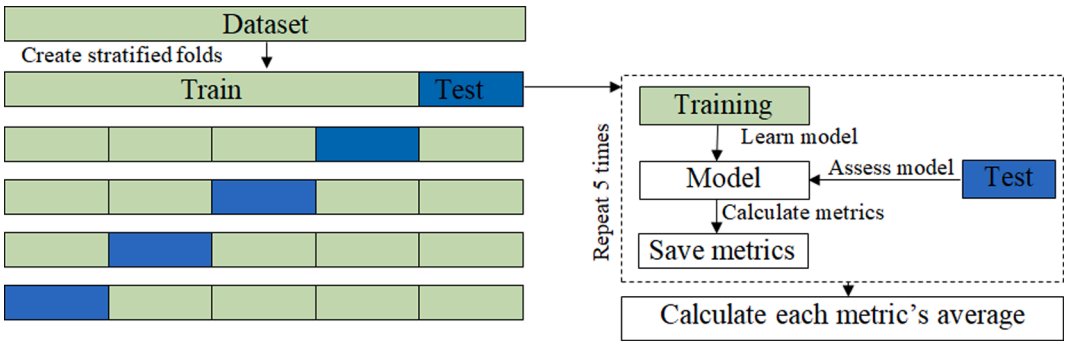

Fig. 3. Schematic overview of five folds cross-validation scheme.

metric used to measure a model's performance developed on an imbalanced dataset.

Matthews correlation coefficient (MCC) is the most reliable statistical metric for binary classifiers when both classes are of interest and the size differs. It considers the actual and predicted classes as two variables and calculates the correlation coefficient between them. It fluctuates from −1 to 1. The higher the correlation between actual and predicted classes, the better the prediction. MCC is 1 for the perfect predictor, indicating a perfect positive correlation. Conversely, when the predictor consistently makes incorrect predictions, the MCC value drops to −1, representing the perfect negative correlation. Respectively, MCC of 0 represents no correlation. The Sn, Sp, Pr, Ac, F1, and MCC are calculated as follows:

$$Sn = \frac{TP}{TP + FN} \tag{6}$$

$$Sp = \frac{TN}{TN + FP} \tag{7}$$

$$Pr = \frac{TP}{TP + FP} \tag{8}$$

$$Ac = \frac{TP + TN}{TP + FN + FP + TN} \tag{9}$$

$$F1 = \frac{2 * Pr * Sn}{Pr + Sn} \tag{10}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TN + FN) * (TP + FP) * (TN + FP) * (TP + FN)}} \tag{11}$$

where, *TP* (True Positive) indicates the number of correctly identified phosphorylation sites, *TN* (True Negative) means the number of correctly identified non-phosphorylation sites, *FP* (False Positive) represents the number of incorrectly identified non-phosphorylation sites as phosphorylated sites, and *FN* (False Negative) denotes the number of incorrectly identified phosphorylation sites as non-phosphorylated sites. We also used the Receiver Operating Characteristics (ROC) curve to evaluate the predictor's performance graphically. ROC curve plots the true positive rate against the false-positive rate for different classification thresholds. The area under the ROC curve can also quantify the predictor's performance. The higher the area under the ROC (AUC) value, the better the predictor. A perfect predictor will achieve the highest score in all of these metrics.

3. Results and discussion

In this section we present our results, compare them with previous studies, and discuss their significance.

3.1. Feature significance analysis

This section investigates the significance of extracted features in phosphorylation site prediction. Three sets of features namely structural, evolutionary, and combination of both have been compared using the Rotation Forest classifier. The results achieved using the 5-folds cross-validation are presented in Table 2 for pS, pT, and pY, prediction tasks. As shown in this table, the evolutionary feature extracted from the PSSM achieves better results for the pS site prediction. In contrast, the structural feature provides more discriminatory information for pT and pY site identification. However, the best prediction performance in terms of sensitivity, specificity, precision, accuracy, and MCC is reported with integrated structural and evolutionary features. This pattern is consistent for all three predictors to identify phosphorylation on serine, threonine, and tyrosine residue. Such consistency justifies the significance of evolutionary and structural features in microbial phosphorylation prediction.

We also plot the ROC curves for structural, evolutionary, and combined features in predicting pS, pT, and pY sites using the independent test set. The plots are shown in Fig. 4. The curves illustrate that the best AUC values of 0.96, 0.90, and 0.91 are achieved respectively for pS, pT, and pY site prediction tasks when structural and evolutionary features are combined.

We created another predictor by extracting the bigram profile from the structural and evolutionary peptide matrixes and training a rotation forest algorithm to examine how the structural information's bigram profile influences the prediction performance. We named it StrucBigram. Table 3 compares StrucBigram with RotPhoPred on the independent test set for predicting pS, pT, and pY sites. The table shows that the performance degrades for all prediction tasks when the bigram feature of structural information is utilized, as indicated by the MCC scores of StrucBigram. RotPhosPred achieves the highest performance for the pS site prediction in all evaluation measures. While predicting pT and pY sites, although StrucBigram attains 13.5% and 3.8% better sensitivity, its performance in other metrics is significantly lower than RotPhosPred. StrucBigram fails to detect 17.5%, 17.6%, and 25.8% non-phosphorylated serine, threonine, and tyrosine sites, respectively. Besides, it misclassified many nonphosphorylated sites as phosphorylated sites, as apparent from the lower precisions of 0.29, 0.33, and 0.38 for the detection of pS, pT, and pY sites. The results also indicate that the StrucBigram predictor is biased towards the positive class. The possible cause behind such biases can be the minimization of features due to the bigram operation on structural information. While the bigram operation is performed, the structural features are reduced from 168 to 64. Consequently, RotPhosPred and StrucBigram are trained on the 568 and 464 features, respectively.

3.2. Comparison with different classifiers

In this section, we analyze the performance of different machine learning algorithms for phosphorylation prediction. To do this, we use

Table 2
Impact of different features on the prediction performance using 5-folds cross validation. The standard deviation among five folds for each metric is presented in the brackets. Bold items indicate the highest values.

Task	Features	Sn(%)	Sp(%)	Pr	Ac(%)	F1	MCC	AUC
pS	Combined	**87.7(0.03)**	**97.6(0.01)**	**0.93(0.01)**	**92.7(0.01)**	**0.92(0.01)**	**0.86(0.02)**	**0.93(0.01)**
	Evolutionary	81.1(0.03)	93.0(0.02)	0.92(0.02)	87.0(0.01)	0.86(0.01)	0.75(0.02)	0.87(0.01)
	Structural	81.1(0.02)	91.6(0.02)	0.91(0.02)	86.4(0.01)	0.86(0.01)	0.73(0.02)	0.86(0.01)
pT	Combined	**79.4(0.02)**	**99.3(0.01)**	**0.99(0.01)**	**89.3(0.01)**	**0.88(0.01)**	**0.80(0.02)**	**0.89(0.01)**
	Evolutionary	50.9(0.01)	98.9(0.01)	0.98(0.02)	74.9(0.01)	0.67(0.01)	0.57(0.02)	0.75(0.01)
	Structural	75.4(0.02)	97.8(0.01)	0.97(0.01)	86.6(0.01)	0.85(0.02)	0.75(0.02)	0.87(0.01)
pY	Combined	**75.2(0.03)**	**99.3(0.01)**	**0.99(0.00)**	**87.3(0.02)**	**0.85(0.02)**	**0.77(0.03)**	**0.87(0.02)**
	Evolutionary	54.2(0.03)	97.1(0.01)	0.95(0.02)	75.6(0.01)	0.69(0.02)	0.57(0.02)	0.76(0.01)
	Structural	67.7(0.02)	95.7(0.01)	0.94(0.01)	81.7(0.01)	0.79(0.01)	0.66(0.02)	0.82(0.01)

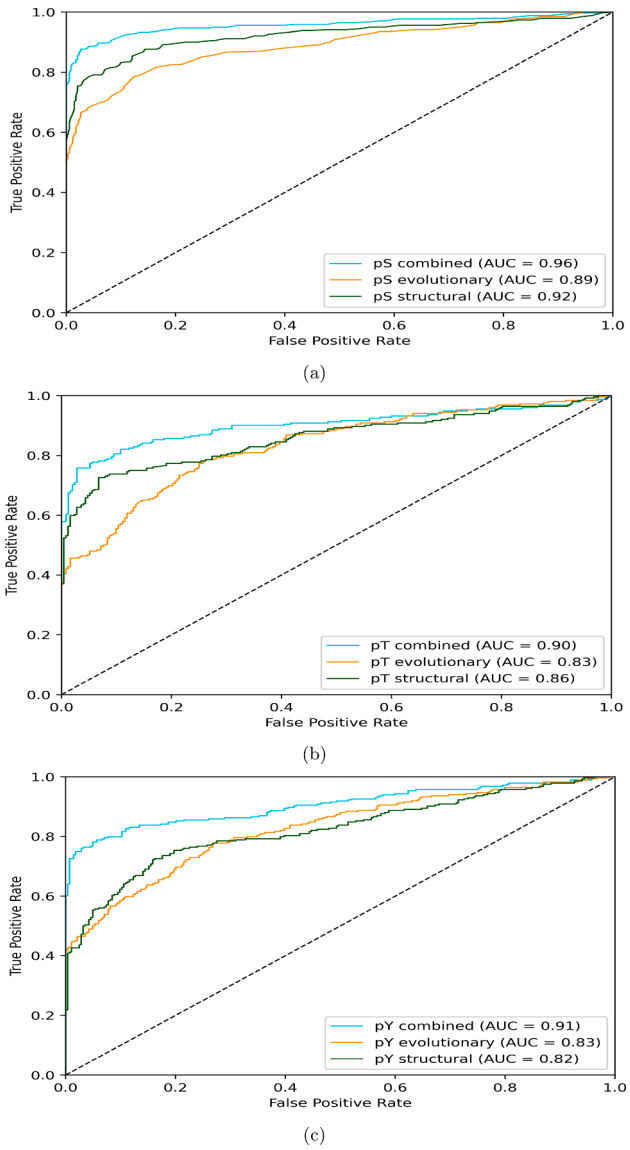

Fig. 4. ROC curves for different feature groups for (a) pS, (b) pT, and (c) pT sites identifications on independent test.

five different classifiers namely, Rotation Forest (RoF), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting Tree (GBT), and Adaptive Boosting (AdaBoost). The hyperparameters of all the classifiers are tuned using cross-validation. We created the Rotation Forest exploiting the *rotation-forest* package available in python3 with 100 decision trees as base estimators. The *max_features* parameter was set to *auto*, and *n_jobs*, indicating the number of jobs to run parallel, was set to −1 to force the algorithm to use all the functional processors. All other machine-learning algorithms are implemented using the *scikit-learn* package in Python. We used a polynomial kernel for SVM to make the samples linearly separable. Besides, set the regularization parameter $c = 1$, which optimizes the hyperplane's margin and minimizes the misclassification of training data. The Gradient Boosting is executed for 100 boosting stages to minimize the log loss with a learning rate of 0.1. The *friedman_mse* splitting criterion was used to measure a split's quality. A decision tree was used as the base estimator for adaptive boosting, which was constructed to execute up to 100 boosting iterations. Each classifier's weight at each boosting iteration was set to 0.1 via the hyperparameter *learning_rate*. We built the random forest model with 100 decision trees and Gini-impurity splitting criteria by setting *n_estimator* = 100 and *criterion* = *gini*, respectively. Other parameters of the classifiers are kept default as in the *rotation-forest* and *scikit-learn* packages. A summary of the hyperparameters settings of different machine learning classifiers used in this study for comparison is shown in Table 4.

For this experiment, we extracted the structural and evolutionary features from each site represented by 21 amino acid residue windows. The experiments using 5-folds cross-validation are shown in Table 5, where the mean values of different performance metrics are reported.

Table 6 shows the independent test results for the various classifiers where the rotation forest (RoF) reporting the best MCC scores for all three predictors for recognizing pS, pT, and pY sites.

The results shows that the RoF beats all the classifiers in sensitivity, accuracy, f1-measure, MCC, and AUC in pS sites identification. The specificity and precision of RoF are also reasonable since they are not significantly lower than the highest values achieved by SVM.

For pT sites predictin task, RoF outperforms all other classifiers in all metrics except the sensitivity. While, Gradient boosting achieves the best sensitivity. However, the sensitivity (75.4%) of the RoF is still comparable to the best results.

Among all the classifiers, RoF shows the highest specificity, precision, accuracy, f1-measure, and AUC values of 94.7%, 0.93, 86.4%, 0.85, and 0.86 in predicting pY sites. It is also competitive in terms of sensitivity. Hence, we use this classifier to build RotPhoPred.

Fig. 5 illustrates the ROC curves of different classifiers for predicting pS, pT, and pY sites on the independent test set. It can be seen from this figure that the AUC values of RoF in predicting pS, pT and pY sites are 0.96, 0.90, and 0.89 respectively which are the highest among the classifiers.

All these results demonstrate the effectiveness of RoF in predicting microbial phosphorylation sites. The secrete behind the superiority of RoF is its ability to do implicit feature selection and introduce diversity in each base classifier by feature transformation using PCA (Rodriguez et al., 2006; Abdi and Williams, 2010).

3.3. Comparison with current state-of-the-art predictors

In this section, we compare RotPhoPred with MPsite and NetPhosBac on the independent test set as the two best microbial phosphorylation

Table 3
Impact of structural bigram profile in the prediction performance on the independent test set. Bold items indicate the highest values.

Task	Model	Sn(%)	Sp(%)	Pr	Ac(%)	F1	MCC	AUC
pS	RotPhoPred	**90.0**	**92.1**	**0.92**	**91.0**	**0.91**	**0.82**	**0.96**
	StrucBigram	88.5	82.5	0.29	83.0	0.44	0.45	0.86
pT	RotPhoPred	75.4	**97.2**	**0.96**	**86.3**	**0.85**	**0.74**	**0.90**
	StrucBigram	**88.9**	82.4	0.33	83.0	0.48	0.48	0.86
pY	RotPhoPred	78.2	**94.7**	**0.94**	**86.4**	**0.85**	**0.74**	**0.89**
	StrucBigram	**82.0**	74.2	0.38	75.5	0.52	0.43	0.78

Table 4
Hyper-parameters summary of different classifiers

Classifiers	Hyper-parameters
Rotation Forset	Base estimator = Decision Tree
	Number of tree = 100
	Maximum features = 'auto'
	Number of jobs = −1
Random Forest	Number of tree: 100
	Splitting criteria: 'gini'
Support Vector Machine	Kernel: 'polynomial'
	Regularization, C: 1.0
Gradient Boosting	Loss: 'log_loss'
	Learning rate = 0.1
	Splitting criteria = '*friedman_mse*'
	Number of boosting stage = 100
Adaptive Boosting	Base estimator = Decision Tree
	Learning rate = 0.1
	Maximum number of estimator = 100

cite predictors. To do this, we fed the independent test as fasta files to MPSite and NetPhosBac servers and collected the predicted result for each site on the dataset. Then, we characterized the performances of these predictors in terms of sensitivity, specificity, accuracy, and MCC. The same metrics are calculated for our method on the independent test set for a fair comparison. The comparative results for the phospho-serine (pS) site are given in Table 7.

The results demonstrate that RotPhoPred outperforms both MPSite and NetPhosBac by achieving the highest sensitivity of 90.0%, specificity of 92.1%, accuracy of 91.0%, and MCC of 0.82. In the phospho-threonine (pT) site, our method also performs better than MPSite and NetPhosBac in terms of sensitivity, specificity, accuracy, and MCC to a large margin, as shown in Table 8. As shown in this table, we enhance the pT performance by 17.1%, 21.5%, 12.1%, and 0.52 in terms of sensitivity, specificity, accuracy, and MCC compared to MPSite as the current best predictor. Since no work has been done to predict the phospho-tyrosine(pY) site, there is no scope to compare the performance of the proposed predictor. Moreover, the independent test set results are consistent with the 5-fold cross-validation approach for both pS and pT classifiers, demonstrating our proposed method's robustness and generality. It is important to note that the superiority of our proposed method comes from the integration of structural and evolutionary features and the use of the RoF classifier as it was discussed in prior sections.

We carefully analyzed the reason behind the poor performances of NetPhosBac and MPSite on the independent test set. We observed that the NetPhosBac was trained on only 152 positive sites, which might overfit or underfit the model. On the other hand, the benchmark training dataset of MPsite was imbalanced with five times higher negative samples than the positive sample for both pS and pT sites. MPsite did not balance the training dataset like us using any imbalance treatment techniques, which may cause it to be biased toward the negative class. Besides, the test set of MPsite (the ratio of positive to negative sites is 1:5) was not highly imbalanced like ours (the ratio of positive to negative sites is 1:⩾15). Maybe because of these reasons, NetPhosBac and MPsite show low MCC scores on our independent test but competitive

Table 6
Comparative results using different machine learning algorithm on independent test.

Task	Classifier	Sn (%)	Sp (%)	Pr	Ac (%)	F1	MCC	AUC
pS	RoF	**90.0**	92.1	0.92	**91.0**	**0.91**	**0.82**	**0.96**
	SVM	79.4	**99.4**	**0.99**	89.4	0.88	0.80	**0.96**
	RF	88.8	92.4	0.92	90.6	0.90	0.81	**0.96**
	AdaBoost	87.6	76.8	0.79	82.2	0.83	0.65	0.90
	GB	**90.6**	79.7	0.82	85.1	0.86	0.71	0.93
pT	RoF	75.4	**97.2**	**0.96**	**86.3**	**0.85**	**0.74**	**0.90**
	SVM	82.9	86.5	0.86	84.7	0.84	0.69	**0.91**
	RF	82.5	84.1	0.84	83.3	0.83	0.67	0.90
	AdaBoost	79.8	72.2	0.74	76.0	0.77	0.52	0.82
	GB	**83.7**	78.6	0.80	81.2	0.82	0.62	0.89
pY	RoF	78.2	**94.7**	**0.94**	**86.4**	**0.85**	**0.74**	**0.89**
	SVM	85.6	73.2	0.76	79.4	0.81	0.59	0.89
	RF	**87.3**	70.8	0.75	79.0	0.81	0.59	0.89
	AdaBoost	83.8	65.5	0.71	74.6	0.77	0.50	0.78
	GB	86.6	66.9	0.72	76.8	0.79	0.55	0.87

Table 5
Comparative results using different machine learning algorithm on 5 folds cross-validation. The standard deviation among five folds for each metric is presented in the brackets. Bold items indicate the highest values.

Task	Classifier	Sn(%)	Sp(%)	Pr	Ac(%)	F1	MCC	AUC
pS	RoF	**87.7(0.03)**	97.6(0.01)	0.97(0.01)	**92.7(0.01)**	**0.92(0.01)**	**0.86(0.02)**	**0.93(0.01)**
	SVM	80.4(0.03)	**99.6(0.01)**	**0.99(0.00)**	90.0(0.02)	0.89(0.02)	0.82(0.03)	0.90(0.02)
	RF	87.2(0.02)	97.4(0.01)	0.97(0.01)	92.3(0.01)	0.92(0.01)	0.85(0.02)	0.92(0.01)
	AdaBoost	84.3(0.02)	85.7(0.01)	0.86(0.01)	85.0(0.00)	0.85(0.00)	0.70(0.00)	0.85(0.00)
	GB	87.5(0.02)	92.2(0.01)	0.92(0.01)	89.9(0.01)	0.88(0.02)	0.80(0.03)	0.90(0.01)
pT	RoF	79.4(0.02)	**99.3(0.01)**	**0.99(0.01)**	89.3(0.01)	0.88(0.01)	0.80(0.02)	0.89(0.01)
	SVM	84.3(0.02)	87.5(0.01)	0.87(0.01)	85.9(0.01)	0.86(0.01)	0.72(0.01)	**0.93(0.01)**
	RF	86.9(0.02)	97.0(0.01)	0.97(0.01)	**91.9(0.01)**	**0.92(0.01)**	**0.84(0.02)**	0.92(0.01)
	AdaBoost	84.3(0.02)	87.5(0.01)	0.87(0.01)	85.9(0.01)	0.86(0.01)	0.72(0.01)	0.86(0.01)
	GB	**87.2(0.01)**	91.3(0.02)	0.91(0.02)	89.2(0.02)	0.89(0.02)	0.76(0.04)	0.89(0.02)
pY	RoF	75.2(0.03)	**99.3(0.00)**	**0.99(0.00)**	87.3(0.02)	0.86(0.02)	0.77(0.03)	0.87(0.02)
	SVM	**83.8(0.01)**	95.7(0.01)	0.95(0.01)	**89.7(0.01)**	**0.89(0.01)**	**0.80(0.02)**	**0.90(0.01)**
	RF	83.4(0.01)	95.2(0.01)	0.95(0.01)	89.3(0.01)	0.87(0.01)	0.79(0.02)	0.89(0.01)
	AdaBoost	79.2(0.02)	80.1(0.01)	0.80(0.01)	79.6(0.01)	0.80(0.01)	0.59(0.02)	0.80(0.01)
	GB	83.1(0.03)	86.4(0.03)	0.86(0.03)	84.6(0.02)	0.84(0.02)	0.70(0.03)	0.85(0.02)

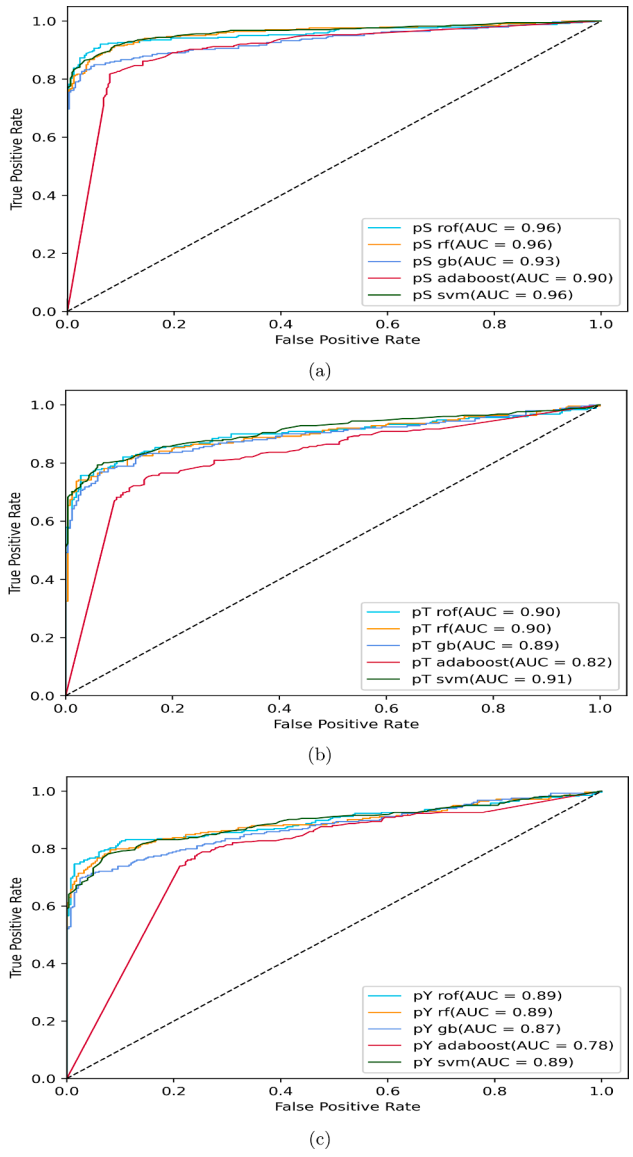

Fig. 5. ROC curves for different classifiers for (a) pS, (b) pT, and (c) pY sites identifications on independent test.

Table 7
Comparison of our method and current predictors for pS site identification on the independent test set.

Predictors	Sn(%)	Sp(%)	Ac(%)	MCC
NetPhosBac (Lee Miller et al., 2009)	31.2	67.8	65.1	−0.01
MPSite (Md Hasan et al., 2019)	39.1	77.9	75	0.11
Proposed	**90.0**	**92.1**	**91.0**	**0.82**

Table 8
Comparison of our method and current predictors for pT site identification on the independent test set

Predictors	**Sn(%)**	**Sp(%)**	**Ac(%)**	**MCC**
NetPhosBac (Lee Miller et al., 2009)	9.4	92.8	85.4	0.03
MPSite (Md Hasan et al., 2019)	58.3	75.7	74.2	0.22
Proposed	**75.4**	**97.2**	**86.3**	**0.74**

scores on their test sets. However, though the models are trained on different training sets, we have evaluated all the predictors on the same independent test set; Hence the comparison is pretty fair. RotPhoPred as a standalone predictor and all its source codes are publicly available at: https://github.com/faisalahm3d/RotPredPho.

4. Conclusion

This paper presents a new microbial phosphorylation site predictor, called RotPhoPred by integrating the structural information and evolutionary bigram profile. We also use Rotation Forest as our employed classifier, which to the best of our knowledge has never been used for this task, to build RotPhoPred. Experimental results on the independent test set demonstrate that RotPhoPred performs better than existing predictors found in the literature for both phospho-serin (pS) and phospho-threonine (pT). Such results indicate that the structural and evolutionary features provide significant discriminatory information to enhance the microbial phosphorylation site prediction task.

We also compared the performance of RoF with other state-of-the-art classifiers using the same set of features. The results demonstrate the performance of using RoF over other classifiers for this task. In the future, we aim at using deep learning models to predict microbial phosphorylation sites more accurately. We also aim to develop a user-friendly and robust web server to provide data interpretation ability using graphical support. RotPhoPred as a standalone predictor and all its source codes are publicly available at: https://github.com/faisalahm3d/RotPredPho

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdi, Hervé, Williams, Lynne J, 2010. Principal component analysis. Wiley Interdiscip. Rev.: Comput. Stat. 2 (4), 433–459.
Ahmad, Md Wakil, Arafat, Md Easin, Taherzadeh, Ghazaleh, Sharma, Alok, Dipta, Shubhashis Roy, Dehzangi, Abdollah, Shatabda, Swakkhar, 2020. Mal-light: Enhancing lysine malonylation sites prediction problem using evolutionary-based features. IEEE Access, 8:77888–77902.
Ahmed, Saeed, Kabir, Muhammad, Arif, Muhammad, Khan, Zaheer Ullah, Dong-Jun, Yu., 2021. Deepppsite: a deep learning-based model for analysis and prediction of phosphorylation sites using efficient sequence information. Anal. Biochem. 612, 113955.
Altschul, Stephen F., Madden, Thomas L., Schäffer, Alejandro A., Zhang, Jinghui, Zhang, Zheng, Miller, Webb, Lipman, David J., 1997. Gapped blast and psi-blast: a new generation of protein database search programs. Nucl. Acids Res. 25 (17), 3389–3402.
Biswas, Ashis Kumer, Noman, Nasimul, Sikder, Abdur Rahman, 2010. Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information. BMC Bioinform. 11 (1), 1–17.
Blom, Nikolaj, Gammeltoft, Steen, Brunak, Søren, 1999. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J. Mol. Biol. 294 (5), 1351–1362.
Bustamam, Alhadi, Musti, Mohamad I.S., Hartomo, Susilo, Aprilia, Shirley, Tampubolon, Patuan P., Lestari, Dian, 2019. Performance of rotation forest ensemble classifier and feature extractor in predicting protein interactions using amino acid sequences. BMC genomics 20 (9), 1–13.
Chandra, Abel, Sharma, Alok, Dehzangi, Abdollah, Shigemizu, Daichi, Tsunoda, Tatsuhiko, 2019. Bigram-pgk: phosphoglycerylation prediction using the technique of bigram probabilities of position specific scoring matrix. BMC Mol. Cell Biol. 20 (2), 1–9.
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.
Ann Chen, Yian, Eschrich, Steven A., 2014. Computational methods and opportunities for phosphorylation network medicine. Transl. Cancer Res., 3(3):266.
Chen, Chi-Wei, Huang, Lan-Ying, Liao, Chia-Feng, Chang, Kai-Po, Chu, Yen-Wei, 2020. Gasphos: protein phosphorylation site prediction using a new feature selection approach with a ga-aided ant colony system. Int. J. Mol. Sci. 21 (21), 7891.
Chowdhury, Shahana Yasmin, Shatabda, Swakkhar, Dehzangi, Abdollah, 2017. idnaprot-es: Identification of dna-binding proteins using evolutionary and structural features. Scient. Reports 7 (1), 1–14.

Abdollah Dehzangi, Somnuk Phon-Amnuaisuk, Mahmoud Manafi, and Soodabeh Safa. Using rotation forest for protein fold prediction problem: An empirical study. In European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, pages 217–227. Springer, 2010.

Abdollah Dehzangi, Kuldip Paliwal, James Lyons, Alok Sharma, and Abdul Sattar. Enhancing protein fold prediction accuracy using evolutionary and structural features. In IAPR International Conference on Pattern Recognition in Bioinformatics, pages 196–207. Springer, 2013.

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

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.

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

Gao, Jianjiong, Thelen, Jay J, Keith Dunker, A., Dong, Xu., 2010. Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. Mol. Cell. Proteom. 9 (12), 2586–2600.

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.

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

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

Md Mofijul Islam, Sanjay Saha, Md Mahmudur Rahman, Swakkhar Shatabda, Dewan Md Farid, and Abdollah Dehzangi. iprotgly-ss: Identifying protein glycation sites using sequence and structure based features. Proteins: Structure, Function, and Bioinformatics, 86(7), 777–789, 2018.

Jamal, Salma, Ali, Waseem, Nagpal, Priya, Grover, Abhinav, Grover, Sonam, 2021. Predicting phosphorylation sites using machine learning by integrating the sequence, structure, and functional information of proteins. J. Transl. Med. 19 (1), 1–11.

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.

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

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.

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

Loughery, Jayne, Meek, David, 2013. Switching on p53: an essential role for protein phosphorylation? BioDiscovery 8, e8946.

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

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.

Md Hasan, Md., Rashid, Mst Khatun, Kurata, Hiroyuki, et al., 2019. Computational identification of microbial phosphorylation sites by the enhanced characteristics of sequence information. Scient. Rep. 9 (1), 1–9.

Rashid, Md.M., Swakkhar Shatabda, Md., Hasan, Hiroyuki Kurata, et al., 2020. Recent development of machine learning methods in microbial phosphorylation sites. Curr. Genom. 21 (3), 194–203.

Hamendra Manhar Reddy, Alok Sharma, Abdollah Dehzangi, Daichi Shigemizu, Abel Avitesh Chandra, and Tatushiko Tsunoda. Glystruct: glycation prediction using structural properties of amino acid residues. BMC bioinformatics, 19(13):55–64, 2019.

Juan José Rodriguez, Ludmila I Kuncheva, and Carlos J Alonso. Rotation forest: A new classifier ensemble method. IEEE transactions on pattern analysis and machine intelligence, 28(10):1619–1630, 2006.

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.

Sharma, Alok, Lyons, James, Dehzangi, Abdollah, Paliwal, Kuldip K., 2013. A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. J. Theoret. Biol. 320, 41–46.

Shatabda, Swakkhar, Saha, Sanjay, Sharma, Alok, Dehzangi, Abdollah, 2017. iphloc-es: Identification of bacteriophage protein locations using evolutionary and structural features. J. Theoret. Biol. 435, 229–237.

Ying Shi, Ying Zhang, Shaofeng Lin, Chenwei Wang, Jiaqi Zhou, Di Peng, and Yu Xue. dbpsp 2.0, an updated database of protein phosphorylation sites in prokaryotes. Scientific Data, 7(1), 1–9, 2020.

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.

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.

Ivan Tomek. Two modifications of cnn. 1976.

Trost, Brett, Kusalik, Anthony, 2011. Computational prediction of eukaryotic phosphorylation sites. Bioinformatics 27 (21), 2927–2935.

Wang, Lei, You, Zhu-Hong, Yan, Xin, Xia, Shi-Xiong, Liu, Feng, Li, Li-Ping, Zhang, Wei, Zhou, Yong, 2018. Using two-dimensional principal component analysis and rotation forest for prediction of protein-protein interactions. Scient. Rep. 8 (1), 1–10.

Wang, Lei, You, Zhu-Hong, Chen, Xing, Yan, Xin, Liu, Gang, Zhang, Wei, 2018. Rfdt: A rotation forest-based predictor for predicting drug-target interactions using drug structure and protein sequence information. Curr. Protein Pept. Sci. 19 (5), 445–454.

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

Wei, Leyi, Xing, Pengwei, Tang, Jijun, Zou, Quan, 2017. Phospred-rf: a novel sequence-based predictor for phosphorylation sites using sequential information only. IEEE Trans. Nanobiosci. 16 (4), 240–247.

Yuedong Yang, Rhys Heffernan, Kuldip Paliwal, James Lyons, Abdollah Dehzangi, Alok Sharma, Jihua Wang, Abdul Sattar, and Yaoqi Zhou. Spider2: A package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. In Prediction of protein secondary structure, pages 55–63. Springer, 2017.

You, Zhu-Hong, Li, Xiao, Chan, Keith C.C., 2017. An improved sequence-based prediction protocol for protein-protein interactions using amino acids substitution matrix and rotation forest ensemble classifiers. Neurocomputing 228, 277–282.

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