scientific reports

Check for updates

OPEN

Computational identification of multiple lysine PTM sites by analyzing the instance hardness and feature importance

Sabit Ahmed[1]✉, Afrida Rahman[1], Md. Al Mehedi Hasan[1], Shamim Ahmad[2] & S. M. Shovan[1]

Identification of post-translational modifications (PTM) is significant in the study of computational proteomics, cell biology, pathogenesis, and drug development due to its role in many bio-molecular mechanisms. Though there are several computational tools to identify individual PTMs, only three predictors have been established to predict multiple PTMs at the same lysine residue. Furthermore, detailed analysis and assessment on dataset balancing and the significance of different feature encoding techniques for a suitable multi-PTM prediction model are still lacking. This study introduces a computational method named 'iMul-kSite' for predicting acetylation, crotonylation, methylation, succinylation, and glutarylation, from an unrecognized peptide sample with one, multiple, or no modifications. After successfully eliminating the redundant data samples from the majority class by analyzing the hardness of the sequence-coupling information, feature representation has been optimized by adopting the combination of ANOVA F-Test and incremental feature selection approach. The proposed predictor predicts multi-label PTM sites with 92.83% accuracy using the top 100 features. It has also achieved a 93.36% aiming rate and 96.23% coverage rate, which are much better than the existing state-of-the-art predictors on the validation test. This performance indicates that 'iMul-kSite' can be used as a supportive tool for further K-PTM study. For the convenience of the experimental scientists, 'iMul-kSite' has been deployed as a user-friendly web-server at http://103.99.176.239/iMul-kSite.

Post-translational modifications (PTM) refers to the covalent addition of certain functional groups to a protein after the translation process[1]. These modifications have significant effects on cellular processes and proteomic research, including cellular signalling, subcellular localization, protein folding, protein degradation, and are also linked to a wide variety of diseases[2,3]. Therefore, identifying and comprehending PTM sites is crucial for scientific investigations in disease identification, prevention, and drug developments[4,5].

There are 20 amino acid residues, such as alanine (A), cysteine (C), lysine (K), arginine (R), etc. Modifications that occur at lysine (K) are named lysine modification or K-PTM. Single or multiple lysine residues may be modified individually or simultaneously where one residue can influence others. In other words, these covalent modifications can aid different K-PTM types, including acetylation, crotonylation, ubiquitination, methylation, butyrylation, succinylation, biotinylation, and ubiquitin-like modifications[6–8]. Though there are several computational tools for predicting various K-PTMs separately, to the best of the authors' knowledge, only three multi-label prediction systems have been developed so far that can take care of the multiplex Lys residues[8–14]. Qiu et al. proposed iPTM-mLys in 2016[5], which could predict four different types of modifications (i.e. acetylation, crotonylation, methylation, and succinylation) simultaneously. The vectorized sequence-coupling model with the random forest algorithm was applied to construct iPTM-mLys[5,15–17]. Hasan and Ahmad proposed mLysPTMpred in 2018[18], where the dataset of iPTM-mLys was utilized to extract the sequence-coupled features, and the cost-sensitive SVM was used as a learning algorithm. The most recent multi-PTM prediction system proposed by Sua et al.[19] has utilized the combination of sequence graph transform (SGT) and convolutional neural networks. All the multi-label predictors, as mentioned earlier, need significant improvement in terms of prediction quality. Furthermore, the number of simultaneous K-PTM prediction capabilities needs to be enhanced. Though there are a few dedicated predictors with multi-PTM prediction capability, all these proposed systems have been trained

[1]Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi 6204, Bangladesh. [2]Computer Science and Engineering, University of Rajshahi, Rajshahi 6205, Bangladesh. ✉email: sabit.a.sirat@gmail.com

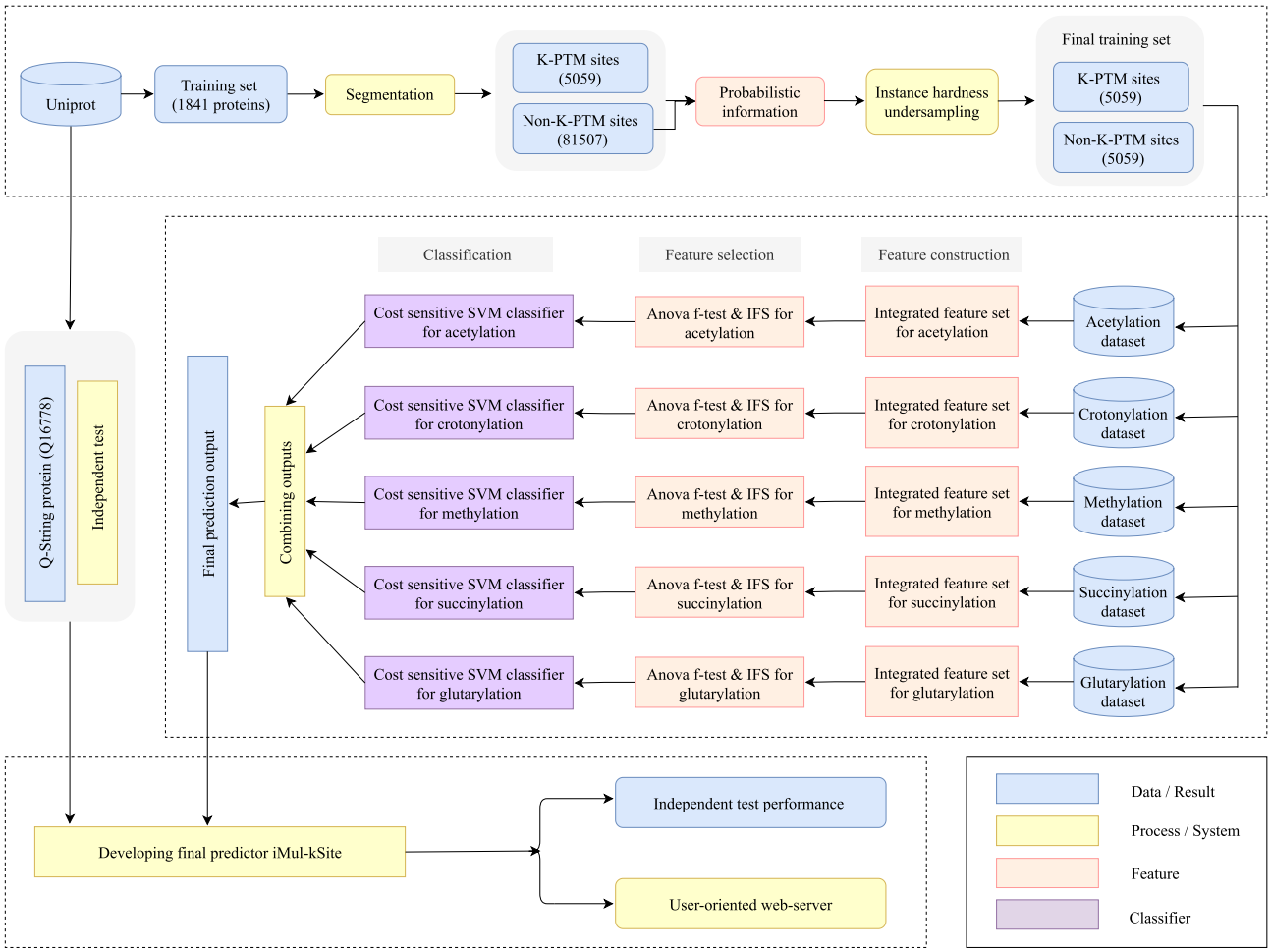

Figure 1. The system flowchart of iMul-kSite.

on the same dataset. Some challenges in this research area include constructing and preprocessing multi-label datasets from raw proteins, lacking multi-label proteins, handling data imbalance, reducing feature dimensions, developing multi-label classifications systems, using proper multi-label evaluation metrics etc. Therefore, adding more types of K-PTMs increases the complexity of this type of research. That might be the reason behind the existence of such a small number of multi-label predictors as well as only one benchmark dataset. Therefore, we have aimed to address these aforestated challenges and construct a highly efficient tool to meet the current demand in the study of post-translational modifications.

In this study, we have proposed a novel multi-label prediction system 'iMul-kSite' to predict five different types of modifications (i.e. acetylation, crotonylation, methylation, succinylation, and glutarylation) concurrently. To develop a successful predictor for PTM sites, one of the main challenges is handling the imbalance in a dataset. Hence, the instance hardness (IH) based undersampling technique has been adopted to remove the redundant samples from the majority class. Another challenge is to elicit features from the input protein sequences as the appropriate features can play a crucial role in better prediction performance[18]. This study has considered several feature encoding methods to develop iMul-kSite, where the amino acid factors, encoded binary features[12,20], pairs of k-spaced amino acids[21], and the vectorized sequence-coupled model[5,10,15] have been aggregated to encode a peptide segment. Afterwards, the analysis of variance (ANOVA) F test statistic along with the incremental feature selection approach has been used to eliminate the redundant and trivial features[22,23]. The support vector machine classifier with the variable cost adjustment process[18] has been implemented to handle the imbalance in each benchmark dataset[24]. A 5-fold cross-validation[18] scheme has been repeated five times for validating the statistical significance of the prediction results, and the average performance of each metric has been reported. A detailed overview is illustrated in Fig. 1.

Methods

Dataset construction. Accurate identification of protein's post-translational modifications often requires a rigorously processed benchmark dataset. As this study is related to the multi-class multi-label classification problem, a few steps have been followed to construct five valid benchmark datasets.

Primary data collection and preprocessing. In the current study, human protein sequences have been utilized for prediction model development and benchmarking. About 9380 protein sequences have been collected from the Universal Protein Resource (UniProt)[25] by applying various constraints (accessed 22 September 2020). Firstly, navigate to the 'Advanced Search' option, select the 'PTM/Processing' and 'Modified residue [FT]' option, keep 'Any assertion method' as 'Evidence'. Then include another query space as 'Organism [OS]', choose 'Homo sapiens (Human) [9606]' from the suggestions as 'Term'. Finally, select the 'Reviewed' option as the third field by adding one more query space. As this study is concerned with a multi-label classification problem, 5 different types of K-PTMs (i.e. acetylation, crotonylation, methylation, succinylation, and glutarylation) have been considered for the dataset construction. After applying a preliminary selection process with the specific keywords of each K-PTM, 1841 proteins have been obtained. For formulating peptide samples meticulously and comprehensively, Chou's scheme[26] has been adopted. According to this scheme, a peptide segment can generally be expressed by,

$$P_\zeta(K) = Q_{-\zeta}Q_{-(\zeta-1)}...Q_{-2}Q_{-1}KQ_{+1}Q_{+2}...Q_{+(\zeta-1)}Q_{+\zeta} \tag{1}$$

where the symbol K denotes the responsible residue 'lysine' at the centre, the subscript ζ being an integer, $Q_{-\zeta}$ and $Q_{+\zeta}$ denotes the ζth leftward and ζth rightward amino acid residues from the centre, and so forth. In this study, primarily a peptide sequence $P_\zeta(K)$ can be categorized into two types,

$$P_\zeta(K) \in \begin{cases} P_\zeta^+(K), \text{if its center is K-PTM site} \\ P_\zeta^-(K), \text{if its center is Non-K-PTM site} \end{cases} \tag{2}$$

where $P_\zeta^+(K)$ contains the positive subset of the peptides and $P_\zeta^-(K)$ contains the negative subset of the peptides with a lysine (K) residue at its centre, and the symbol $\in$ indicated the set theory relationship. For equal-sized K-PTM site formation, $(2\zeta + 1)$-tuple peptide window with K at its centre has been employed. During segmentation, the lacking amino acid at both the right and left end has been filled with the nearest residue[5]. After the peptide fragments have gone through some screening, such as the elimination of sequences in case of redundancy, the primary dataset has been constructed with the following form,

$$S_\zeta(\mathrm{K}) = S_\zeta^+(\mathrm{K}) \cup S_\zeta^-(\mathrm{K}) \tag{3}$$

where the positive subset $S_\zeta^+(\mathrm{K})$ can contain any peptide samples which have one or more modifications (i.e. acetylation, crotonylation, methylation, succinylation, glutarylation) with K at the centre, while the negative subset $S_\zeta^-(\mathrm{K})$ can contain only the false K-PTM samples which have no modifications at all. The sliding window method[10] was adopted to segment the protein sequences with different window sizes where $\zeta = 1, 2, 3, \ldots 24$. Based on the Accuracy value, window size was selected as $(2\zeta + 1) = 49$ where $\zeta = 24$ (i.e. 24 right stream and 24 left stream amino acid residues). It should be mentioned that only the window sizes less than 51 were taken under consideration due to the compelling protein sequence length[10]. Therefore, Eq. (1) has been reduced to,

$$P(K) = Q_{-24}Q_{-23}...Q_{-2}Q_{-1}KQ_{+1}Q_{+2}...Q_{+23}Q_{+24} \tag{4}$$

Following the aforestated process, 5059 K-PTM samples and 81507 Non-K-PTM samples have been obtained.

Data imbalance management and benchmark dataset formation. It can be observed that the primary dataset is highly imbalanced where the ratio between K-PTM and Non-K-PTM sites is 1:16. The instance hardness (IH) based undersampling technique has been employed for reducing this skewness[27]. Later at the classification level, a cost-sensitive SVM classifier has been utilized to address the imbalance in each K-PTM dataset.

Instance hardness undersampling. Smith, Martinez, and Giraud-Carrier have proposed the instance hardness (IH) undersampling technique for binary classification problems[27,28]. In this study, we adopted this technique by measuring the hardness of the sequence-coupling information which have been extracted from the primary dataset by using Eqs. (10), (11) and (12). The detailed methodology of the vectorized sequence-coupling feature extraction technique has been discussed in the "Feature construction" section. From Fig. 1, it can be observed that one or more modifications can occur at 5059 'K-PTM' samples, where 81507 'Non-K-PTM' samples lack any of the modifications. The objective here is to find out the most suitable peptide samples which represent no modification at all. In this work, the hardness of an instance in the coupling feature set measures how likely it is to be misclassified. Higher hardness values indicate that the data samples are noisy or on the border between 'K-PTM' and 'Non-K-PTM' classes, as the learning algorithms would cause them to overfit correctly[28]. For a peptide sample (x_i, y_i), $p(y_i|x_i, h)$ denotes the conditional probability of label y_i for the input feature vector x_i given by the learning algorithms h. The higher the value of $p(y_i|x_i, h)$ is, the more likely h assigns the correct label to x_i, and it is quite opposite for the smaller value of $p(y_i|x_i, h)$[27,28]. The hardness of an instance (x_i, y_i), concerning h, is defined as,

$$I_h[(x_i, y_i)] = 1 - p(y_i|x_i, h) \tag{5}$$

Let $\mathscr{H}$ be the set of weak learners and $p(h|t)$ be the corresponding weight of $h \in H$, where $t = (x_i, y_i) : x_i \in X \wedge y_i \in Y$. Hence, the hardness of an instance in the data sample takes the following form,

Attribute	S_ζ (*acetylation*)	S_ζ (*crotonylation*)	S_ζ (*methylation*)	S_ζ (*succinylation*)	S_ζ (*glutarylation*)
True	4154	208	325	1253	236
False	5964	9910	9793	8865	9882

Table 1. Number of samples in the benchmark dataset for different K-PTMs.

Attribute	1 K-Type	2 K-Types	3 K-Types	4 K-Types	5 K-Types	Non-K-Types
Benchmark dataset	4089	861	77	26	6	5059

Table 2. K-PTM distributions in the training set.

$$\begin{aligned} I[(x_i, y_i)] &= \sum_{\mathscr{H}} (1 - p(y_i|x_i, h))p(h|t) \\ &= \sum_{\mathscr{H}} p(h|t) - \sum_{\mathscr{H}} p(y_i|x_i, h)p(h|t) \\ &= 1 - \sum_{\mathscr{H}} p(y_i|x_i, h)p(h|t) \end{aligned} \tag{6}$$

Following this concept, the imbalanced dataset has been resampled by eliminating the data points from the majority class with high instance hardness values, until the desired balancing ratio of 1:1 has been reached. To estimate the hardness of an instance, we utilized the cost-sensitive support vector machine[29–31] which will be discussed later in this study. It should be mentioned that scikit-learn's library[32] has been used to implement the instance hardness (IH) based undersampling technique. Finally, 5059 positive and 5059 negative samples have been obtained, and the original peptide sequences with the expression of Eqs. (3) and (4) have been retrieved from the returned indices of the resampled dataset. The final benchmark datasets have been constructed by mapping the samples labeled as 'K-PTM' and 'Non-K-PTM' into each individual classes which takes the following form,

$$\begin{cases} S_\zeta(\text{acetylation}) = S_\zeta^+(\text{acetylation}) \cup S_\zeta^-(\text{acetylation}) \\ S_\zeta(\text{crotonylation}) = S_\zeta^+(\text{crotonylation}) \cup S_\zeta^-(\text{crotonylation}) \\ S_\zeta(\text{methylation}) = S_\zeta^+(\text{methylation}) \cup S_\zeta^-(\text{methylation}) \\ S_\zeta(\text{succinylation}) = S_\zeta^+(\text{succinylation}) \cup S_\zeta^-(\text{succinylation}) \\ S_\zeta(\text{glutarylation}) = S_\zeta^+(\text{glutarylation}) \cup S_\zeta^-(\text{glutarylation}) \end{cases} \tag{7}$$

A comprehensive summary of dataset preparation has been presented in Fig. 1. The numbers of samples in the benchmark datasets are outlined in Table 1, and their detailed sequences and positions in the proteins are given in the Supplementary File. The distributions of different types of modifications in the benchmark datasets are tabulated in Table 2. It could be observed that our benchmark datasets contain 4089 samples belonging to one type of K-PTM, 861 to two types, 77 to three types, 26 to four types, and 6 to five types modifications.

Cost-sensitive classifiers. We have handled the imbalance between the K-PTMs and Non-K-PTM sites by utilizing the instance hardness undersampling technique. However, it can be observed from Table 1 that still there exists some skewness between the positive and negative sites of each of the five modifications. Therefore, further adjustments are needed to deal with this issue. We have utilized five cost-sensitive SVM classifiers for mitigating the imbalance problem of five datasets in Table 1. A detailed discussion on the support vector machine prediction algorithm and the proposed model development are presented in the "Support vector machine" and "Model development and validation" sections respectively.

Feature construction. With the evolution of the biological sequences, several encoding methods have been developed for extracting pertinent features hidden in the sequences. After preliminary analysis, it has been observed that the amino acid factors, encoded binary features, pairs of k-spaced amino acids, and the vectorized sequence coupling[12,15,20] technique are more appropriate for representing the protein sequences of the multiple lysine modification sites than any other encoding methods.

Amino acid factors. Five multidimensional attributes[20,33], which include polarity, secondary structure, molecular volume, electrostatic charge, and codon diversity[34], have been constructed from AAIndex by using multivariate statistical analysis[12]. These five transformed properties can be introduced as amino acid factors (AAF)[34]. Since the AAF can reduce the dimensionality of the feature space of physicochemical properties efficiently, it has been utilized in many biological studies[12,34]. The dimensionality of feature vectors has been calculated as follows,

$$D = peptide\ sequence\ length \times number\ of\ factors \tag{8}$$

With a peptide sequence of length 27 and previously described five amino acid factors, $49\times5 = 245$ dimension features have been derived by using this formula.

Binary encoding. Binary encoding[12] can represent the amino acid position and composition by using 20 binary bits for one amino acid[12]. But one additional bit has been conjoined to handle the complexity of sliding windows. For 21 amino acids structured as 'ACDEFGHIKLMNPQRSTVWYZ', each residue inside a sequence fragment can be formed by a 21-dimension binary vector[12]. For instance, residue '*A*', '*G*' and '*Z*' have been encoded as '100000000000000000000', '000000100000000000000' and '000000000000000000001' respectively. According to this concept, each resultant peptide segment is expressed as $49\times21 = 1029$-dimensional feature vectors.

Pairs of k-spaced amino acids. The formation of k-spaced amino acid pairs encoding technique[12,21,35] calculates the occurrence frequencies of the pairs of k-spaced amino acids from a segmented protein sample, that can express the short linear motif information out of it[12,30]. For instance, the encoding of a peptide segment will be a 441-dimensional feature vector if $k = 0$. This can be defined as,

$$(N_{AnA}/N_{Total}, N_{AnC}/N_{Total}, ..., N_{YnY}/N_{Total})_{441} \tag{9}$$

where n stands for any of amino acid, N_{Total} means the occurrence frequency of all k-spaced amino acid pairs[35] and N_{AnA} means the occurrence frequency of the AnA pairs in the segment[20] when $k = 0$. In this study, after merging each of the 441-dimension feature vectors for $k = 0, 1, 2, 3, 4$, a total of 2205-dimensional features have been formed.

Sequence coupling. The composition of pseudo amino acid or PseAAC[10,36,37] has been designed to preserve the sequence pattern information, which is a much harder task for any existing machine learning algorithm[38]. In this study, incorporating sequence coupling information into Chou's general PseAAC has been adopted for extracting features from peptide sequences[5,15,18]. It can be defined as,

$$P(K) = P^{+}(K) - P^{-}(K) \tag{10}$$

where,

$$P^{+}(K) = \left[P^{C^{+}}_{-24}P^{C^{+}}_{-23}\ldots P^{+}_{-1}P^{+}_{+1}\ldots P^{C^{+}}_{+23}P^{C^{+}}_{+24}\right]^{T} \tag{11}$$

$$P^{-}(K) = \left[P^{C^{-}}_{-24}P^{C^{-}}_{-23}\ldots P^{-}_{-1}P^{-}_{+1}\ldots P^{C^{-}}_{+23}P^{C^{-}}_{+24}\right]^{T} \tag{12}$$

where $P^{C^{+}}_{-24}$ in Eq. (11) denotes the conditional probability of amino acid Q_{-24} at the leftmost position given that its adjacent right member is Q_{-23} and so forth[5,18]. In contrast, only P^{+}_{-1} and P^{+}_{+1} are of non-contingent probability as K is the adjoining member of both amino acids at position Q_{-1} and Q_{+1}. All the conditional probability values have been extracted from the positive training dataset. Additionally, all the probability values in Eq. (12) are identical to those of Eq. (11) other than that they can be derived from the negative training dataset. Thus, after omitting K from the center, $(49 - 1) = 48$ dimension features have been obtained.

Feature ensembling. Initially, the four aforestated feature encoding techniques (i.e. AAF, BE, CKSAAP, and sequence coupling) have been implemented separately to encode the training peptides. However, for extracting more PTM-contextual information from the protein sequences, encoded features have been ensembled serially, and scaled through standardization. Finally, $(49\times5) + (49\times21) + (441\times5) + 48 = 3527$ dimension features have been obtained.

Feature selection. Since the dimension of the encoded features is higher, irrelevant, and redundant features should be removed to avoid learning complexity. For this reason, the analysis of variance (ANOVA) F test statistic technique[22,39] has been adopted. It tests the null hypothesis (i.e. all the means of different groups were equal) against the alternative hypothesis (i.e. all the means differed from each other). The one-way ANOVA can be defined as,

$$F = \frac{(n-k)\sum n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2}{(k-1)\sum(n_i - 1)s_i^2} \tag{13}$$

where $n = \sum_{i=1}^{k} n_i$, $\overline{Y}_{i.} = Y_{i.}/n_i$, $\overline{Y}_{..} = Y_{..}/n$

and

$s_i^2 = \sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2/(n_i - 1)$.

It should be mentioned that the dot in $Y_{i.}$ indicates an aggregation over the j index[39]. Where $Y_{i.} = \sum_{j=1}^{n_i} Y_{ij}$ and $Y_{..} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$. The calculated F values are used to rank the features. The discriminative capability of a predictor is better for higher F values.

Support vector machine. The support vector machine (SVM)[29–31], one of the dominant statistical learning algorithms was adopted as a core prediction algorithm. It seeks the optimum hyperplane with the highest margin between two groups[18,40]. Furthermore, it solves the problem of constraint optimization as described below

$$maximize_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \tag{14}$$

Subject to: $\sum_{i=1}^{n} y_i \alpha_i = 0, \quad 0 \le \alpha_i \le C$, for all i= 1, 2, 3, ..., n. After involving the kernel function, the discriminant function of SVM took the following form

$$f(x) = \sum_{i}^{n} \alpha_i y_i k(x, x_i) + b \tag{15}$$

In this paper, the radial basis function kernel[18,41] was applied to construct SVM classifier and given by, $k(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2)$, where $\gamma > 0$[42]. As the benchmark dataset was highly imbalanced, different error cost (DEC)[18] method had been used to tackle the class imbalance problem[24,43]. According to this approach, the SVM soft margin objective function was adjusted to allocate two costs for misclassification[12], such as C^+ for the positive class instances and C^- for the negative class instances

$$C^+ = C * W^+, \qquad C^- = C * W^- \tag{16}$$

In Eq. (16), W^+ is the weight for the positive instances and W^- is the weight for the negative instances and defined by

$W^+ = \frac{M}{2*M_1}, \quad W^- = \frac{M}{2*M_2}$ where M is the total number of elements, M_1 is the number of elements for the positive class, and M_2 is the number of elements for the negative class.

Evaluation metrics. As shown in Table 1 and Supplementary Material, the total number of peptide samples are 10118 in total, of which 4154 are labelled with 'acetylation', 208 with 'crotonylation', 325 with 'methylation', 1253 with 'succinylation', 236 with 'glutarylation', and 5059 with 'Non-K-PTM'. Since a sample can contain more than one labels, metrics for multi-label systems[5,18] have been utilized instead of ordinary metrics for single-label systems[9,10,12,44]. According to Chou's formulation[45], the metrics for multi-label systems can be defined as,

$$\begin{cases} Aiming = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\|Y_i \cap Y_i'\|}{\|Y_i'\|} \right) \\ Coverage = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\|Y_i \cap Y_i'\|}{\|Y_i\|} \right) \\ Accuracy = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\|Y_i \cap Y_i'\|}{\|Y_i \cup Y_i'\|} \right) \\ Absolute - True = \frac{1}{N} \sum_{i=1}^{N} \left(\Delta \|Y_i, Y_i'\| \right) \\ Absolute - False = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\|Y_i \cup Y_i'\| - \|Y_i \cap Y_i'\|}{L} \right) \end{cases} \tag{17}$$

where N and L are the total numbers of the samples and labels in the system respectively[5,18], $\cup$ and $\cap$ denotes the 'union' and 'intersection' in the set theory, $||\ ||$ means the operator acting on the set to calculate the number of its elements, Y_i and Y_i' denotes the subset that contained all the labels experiment-observed and all the labels predicted for the i^{th} sample respectively, and

$$\Delta(Y_i, Y_i') = \begin{cases} 1, \text{if all labels in } Y_i' \text{ and } Y_i \text{ are identical} \\ 0, \text{otherwise} \end{cases}$$

The metrics defined above have been applied effectively in several multi-label based systems[5,18].

Model development and validation. In this study, five separate SVM classifiers[18] have been used to predict acetylation, crotonylation, methylation, succinylation, and glutarylation sites. Each of the classifiers has performed binary classification on the benchmark dataset described in Table 1. For all five K-PTM types, necessary features have been extracted by integrating multiple encoding methods and 100 optimal features with ANOVA F-test have been selected to train the models, as shown in Fig. 1. The radial basis function (RBF) kernel[40,46] has been used for each SVM classifier. As there is a lack of details about the exact 5-way splits of the dataset[40], five complete runs of 5-fold cross-validation have been executed[5,18,47]. The misclassification cost C has been calculated according to Eq. (16) for handling the data imbalance issue. In this study, libSVM's default parameters (i.e. $C = 1$ and $\gamma = 1/$number of features) have been selected to train the model. Eventually, after training the five binary SVM classifiers with the appropriate hyperparameters, multi-label predictor iMul-kSite has been constructed by combining the outputs from these classifiers[40], as depicted in Fig. 1. Five times repetition of the 5-fold cross-validation[40] have produced five sets of values of all metrics, which are defined in the previous section. The average results of each multi-label metric have been taken to evaluate the final model. It should be mentioned that Matlab 2019a and python 3.7.3 have been utilized to implement the system.

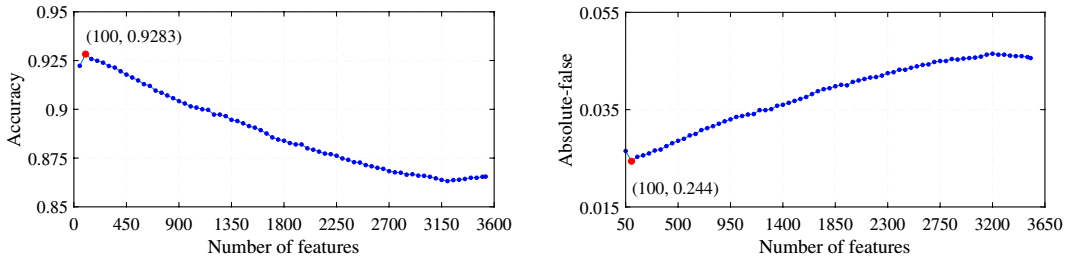

Figure 2. The IFS curves: (**a**) Feature range 50–3527 (Features vs. Accuracy). (**b**) Feature range 50–3527 (Features vs. Absolute-false).

Predictors	Functionality	Aiming(%)	Coverage(%)	Accuracy(%)	Absolute-True(%)	Absolute-False(%)
iPTM-mLys	4 K-PTMs	69.78	74.54	68.37	60.92	13.40
mLysPTMpred	4 K-PTMs	84.82	86.56	83.73	79.73	6.66
CNN + SGT[a]	4 K-PTMs	83.91	83.91	82.75	85.21	4.27
iMul-kSite [b]	4 K-PTMs*	93.18 ± (0.06)	96.13 ± (0.09)	92.70 ± (0.07)	88.77 ± (0.08)	2.97 ± (0.03)
iMul-kSite [c]	**5 K-PTMs**	**93.36** ± (0.05)	**96.23** ± (0.07)	**92.83** ± (0.07)	**88.84** ± (0.11)	**2.44** ± (0.02)

Table 3. Cross-validation performance of the existing predictors. The highest performance is indicated with bold texts. [a]Method proposed by Nie et al.[19] [b,c]Correspond to the iMul-kSite performances on the benchmark datasets containing 4-PTMs and 5-PTMs respectively. *Corresponds to the 4 K-PTMs used in the previous studies i.e. acetylation, crotonylation, methylation and succinylation.

Results

Incremental feature selection. The feature selection procedure has been implemented in two steps. Primarily, all the features have been tested with the analysis of variance (ANOVA) and the features with statistical significance have been obtained[48]. Hence, all of the 3527 features have been ranked according to the calculated F values.

Later, the incremental feature selection (IFS)[12] algorithm has been applied for selecting the optimal number of features[12,48]. For each feature subset of top m ($m = 50, 100, 150, \ldots, 3527$), one SVM classifier with libSVM's default parameter[30,49] has been trained for each K-PTM type and its accuracy and absolute-false rate have been measured by adopting 5-fold cross-validation. As depicted in Fig. 2, the highest accuracy of 92.83% with the lowest absolute-false rate of 2.44% has been achieved with 100 leading features. Finally, the proposed predictor kMul-iSite has been constructed by utilizing the top 100 features.

Prediction performance of iMul-kSite. The performance of the iMul-kSite predictor derived from the aforementioned multi-label metrics is given in Table 3. The values of the five metrics are the average result of five times complete run of 5-fold cross-validation on the benchmark dataset. In Eq. (17), for the first four metrics, the higher the rate is, the better the performance will be, and for the last one, it is entirely the opposite[18]. The rate of the most crucial metric 'Accuracy' for our proposed predictor iMul-kSite is 92.83%. Besides, it has achieved a 93.36% 'Aiming' or 'Precision' rate which represents the average ratio of the predicted labels that hit the target of the original labels. The average ratio of the original labels that are covered by the hits of prediction referred to as 'Coverage' is 96.23%. To the best of the authors' knowledge, no multi-label predictor has achieved a coverage rate of over 90% so far. In addition to that, the experimentally obtained rate of the most stringent and harsh metric 'Absolute-True' is 88.84% which is significant for any multi-label prediction system. Furthermore, the rate of 'Absolute-False' or 'Hamming-Loss' denoting the average ratio of completely wrong hits over the total prediction events is 2.44%.

Comparison with existing multi-label predictors. According to the best of the authors' knowledge, there are only three multi-label prediction systems that can predict multiple K-PTM sites simultaneously. All of these predictors have been constructed for identifying four types of K-PTMs i.e. acetylation, crotonylation, methylation, and succinylation. Qiu et al.[5] have constructed iPTM-mLys, which is the first-ever multi-PTM prediction system for lysine modifications. Hasan and Ahmad[18] have proposed another multi-label prediction system termed as mLysPTMpred. Recently, Sua et al.[19] have constructed a method with the combination of convolutional neural network and sequence graph transform (CNN + SGT). The last two systems have achieved comparatively higher prediction performance than iPTM-mLys. They also have surpassed the milestone of reaching over 80% absolute-true rate.

Predictors	Functionality	Aiming (%)	Coverage (%)	Accuracy (%)	Absolute-true (%)	Absolute-false (%)
iPTM-mLys	4 K-PTMs	67.50	65.00	62.50	55.00	15.00
mLysPTMpred	4 K-PTMs	88.33	87.50	85.83	80.00	6.00
CNN + SGT [a]	4 K-PTMs	65.00	65.00	65.00	85.00	5.00
iMul-kSite	**5 K-PTMs**	**95.00**	**95.00**	**95.00**	**95.00**	**1.67**

Table 4. Performance of different predictors on the Q-string independent test set. The best achievable performance has been indicated with bold texts. [a]Method proposed by Sua et al.[19].

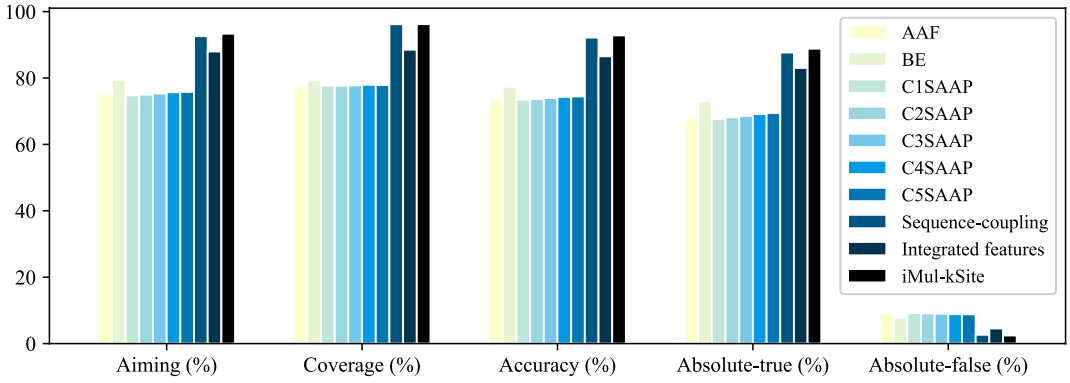

Figure 3. Performance comparison between different feature encoding techniques.

However, we have constructed a novel multi-PTM site predictor iMul-kSite which can predict 5 K-PTM sites concurrently. In addition to that, we have excluded the glutarylation sites from the benchmark dataset and reported the performance of iMul-kSite on the rest of the 4 K-PTMs in Table 3. In comparison with the recently developed multi-label predictor mLysPTMpred[18], it can be observed that the rate of the most crucial metric 'Accuracy' for the proposed predictor iMul-kSite has been increased from 84.82% to 92.83%. Our proposed system has also achieved 8.54% and 9.67% increased aiming and coverage rates respectively. Furthermore, the absolute-true has reached 88.84% and the absolute-false has reached 2.44%. Therefore, the experimental results reported in Table 3 indicate that the constructed multi-label predictor iMul-kSite has achieved better performance than the existing state-of-art multi-PTM predictors even after the inclusion of one more type of PTM site prediction functionality[5,18,19].

It should be mentioned that a Q-string protein sequence (Q16778) has been utilized in iPTM-mLys, mLysPTMpred, and Nie's method for independent test[5,18,19]. Though these multi-PTM predictors do not account for glutarylation sites to be predicted, the independent test results of these predictors have been included in Table 4 for demonstrating the prediction accuracy of the proposed system. According to Eq. (17), the aiming, coverage, accuracy, and absolute-true rates are 95.00%, and the absolute-false rate is 1.67%. The superior performance obtained from both the cross-validation and independent test demonstrates the validity of our proposed model and it could be a high throughput tool for multi-label PTM site identification.

Predictive performance of different feature encoding schemes. The performance obtained by iMul-kSite has been further compared with multiple baseline K-PTM prediction systems, developed using different feature extraction methods, such as the amino acid factors (AAF), binary encoding (BE), pairs of k-spaced amino acids (CKSAAP), and incorporation of sequence coupling information into general PseAAC[12,15,20,34,50,51] to estimate iMul-kSite's K-PTM related information extraction capability. The performances of the specified feature encoding schemes evaluated by 5-fold cross-validation are depicted in Fig. 3.

It may be observed that the amino acid factor (AAF) has acquired a higher absolute-false rate of 9.21% with considerably lower accuracy, absolute-true, aiming, and coverage rate. However, much better results have been picked up by binary encoding (BE) schemes. It has reached 77.35% accuracy with a 79.49% aiming rate and a 79.50% coverage rate. The absolute-false rate is reduced to 7.82% with an absolute-true rate of 73.05%. The composition of the k-spaced amino acid pairs (CKSAAP)[12,52] encoding technique has been adopted for the different combinations of k, in which the '0-spaced ($k = 0$) amino acid pairs' has produced the lowest accuracy, aiming, coverage and absolute-true rate and the highest absolute-false rate. The performances secured by the composition of 1-spaced ($k = 0, 1$), 2-spaced ($k = 0, 1, 2$), and 3-spaced ($k = 0, 1, 2, 3$) amino acid pairs have been improved a little and maximized for the composition of 4-spaced ($k = 0, 1, 2, 3, 4$) amino acid pairs as illustrated in Fig. 3. It has achieved 74.44% accuracy, which is the topmost accuracy among the various combinations of CKSAAP encoding schemes but compared to other feature extraction techniques, it is not a desirable performance. Sequence-coupling, which is one of the most crucial encoding strategies, has attained a higher accuracy rate of 92.20%, an aiming rate of 92.62% with a much lower absolute-false rate of 2.66%. It

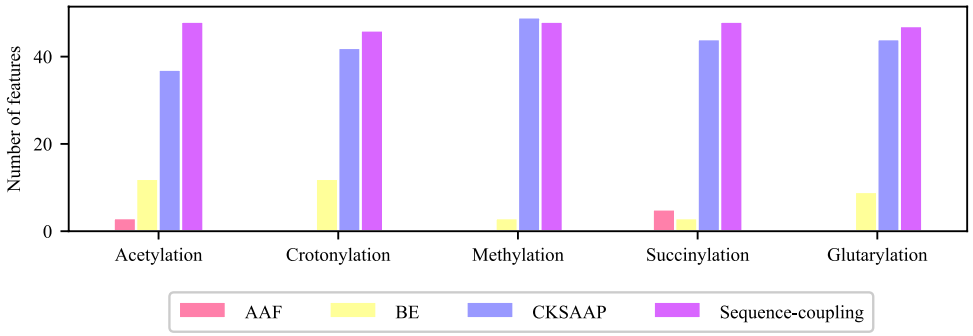

Figure 4. Feature distribution in the optimal feature sets.

Feature name	Acetylation	Crotonylation	Methylation	Succinylation	Glutarylation
AAF (%)	1.23	0.00	0.00	2.04	0.00
BE (%)	1.17	1.17	0.29	0.29	0.88
CKSAAP (%)	1.68	1.91	2.23	2.00	2.00
Sequence-coupling (%)	100.00	95.83	100.00	100.00	97.92

Table 5. Percentage of features selected with ANOVA F-Test and IFS.

has obtained a coverage rate above 90%, which is a rare example in bioinformatics. Therefore, integrating all the feature extraction methods has been considered a successful approach for developing a multi-label predictor. Consequently, the sequence-coupling has been combined with amino acid factor, binary encoding, and the composition of k-spaced amino acid pairs where $k = 0, 1, 2, 3, 4$. But the performances of the integrated features have been degraded and for 3527 dimension features, accuracy has been reduced to 86.55% with the increased absolute-false rate of 4.57%. Later, 100 optimal features have been selected from the high dimension features by conducting ANOVA F-test. By using the libSVM's default parameter value of C and gamma, accuracy and aiming rate have been reached 92.83% and 93.36% respectively[49]. The most uncompromising metric absolute-true rate is 88.84% with a lower absolute-false rate of 2.44%. Figure 3 points out that the model constructed with the informative features termed as 'iMul-kSite' has achieved a discernible performance among all the feature encoding techniques described earlier.

Optimal features analysis. The feature distribution for different K-PTM types is shown in Fig. 4. Moreover, the percentages of each type of feature selected with ANOVA and IFS are illustrated in Table 5 for a better understanding of the importance and dominance of the corresponding features. For the acetylation feature set, out of 100 optimal features, 3 belong to the AAF, 12 belong to the BE, 37 belong to the CKSAAP, and 48 belong to the sequence-coupling. Therefore, the ratios of selected dimensions of these four types of features are 1.23% (3/245), 1.17% (12/1029), 1.68% (37/2205), and 100% (48/48) respectively.

The crotonylation feature set comprises 46 sequence-coupling features, 12 BE features, and 42 CKSAAP features. Figure 4 and Table 5 show that the optimal feature set of crotonylation does not contain any of the AAF features. Hence, the selected dimension ratios of BE, CKSAAP, and sequence-coupling features are 1.17% (12/1029), 1.91% 42/2205), and 95.83% (46/48) respectively. Besides, the methylation feature set consists of 3 BE features, 49 CKSAAP features, and 48 sequence-coupling features, and the ratios of the selected dimensions for each type of feature are 0.29% (3/1029), 2.23% (49/2205), and 100% (48/48) respectively. For the succinylation dataset, 5, 3, 44, and 48 features belong to the AAF, BE, CKSAAP, and sequence-coupling respectively. The dimension ratios for AAF, BE, CKSAAP, and sequence-coupling are 2.04% (5/245), 0.29% (3/1029), 2.00% (44/2205), and 100% (48/48) respectively. For the glutarylation dataset, 9, 44, and 47 features belong to the BE, CKSAAP, and sequence-coupling respectively. The dimension ratios for BE, CKSAAP, and sequence-coupling are 0.88% (9/1029), 2.00% (44/2205), and 97.92% (47/48) respectively.

As reflected in Table 5, the selected feature dimensions for BE, AAF, and CKSAAP have varied over different types of K-PTM site prediction. The sequence-coupling features have a stronger influence on the identification of all of the five K-PTM sites. In contrast, BE, and CKSAAP features have much smaller and almost similar effects on each K-PTM site prediction. AAF features have a slightly better impact on the acetylation and succinylation site prediction but those have barely any effect on the crotonylation, methylation, and glutarylation site prediction. Therefore, it may be concluded that the proposed model augmented the sequence-coupling effect with the essential features of AAF, BE, and CKSAAP has intensified the prediction performance of iMul-kSite.

Analysis on different modifications. The multi-label predictor iMul-kSite has been developed by combining outputs from the five optimized binary classifiers as discussed in the previous section. Though the final

outputs have been evaluated by the multi-label metric system, each of the individual classifiers has been evaluated and tuned depending on the area under curve (AUC) value. From Table 1, it can be seen that the acetylation dataset is quite a balanced dataset. But The imbalance ratio of the number of succinylated sites to that of non-succinylated sites is approximately 1:7. On the other hand, the crotonylation, methylation and glutarylation datasets have higher imbalance ratios (around 1:40) between the number of positive and negative peptides. In this study, the imbalance between the positive and negative sites for different datasets has been handled in two stages. Firstly, the 'K-PTM' and 'Non-K-PTM' sites containing samples have been resampled at the dataset level. Later, the imbalance in each modification dataset has been minimized at the classifier level. It has been observed that, the average AUC of acetylation and succinylation classifiers were 97.64% and 98.44%, respectively. On the other hand, the average AUC values of crotonylation, methylation and glutarylation are 99.98%, 99.89% and 99.96% respectively. It can be concluded that after applying successful data balancing techniques at different levels, the constructed predictor iMul-kSite has demonstrated its superior performance for identifying all five types of different modifications.

Web-server. To aid the experimental researches, a user-oriented web-server for iMul-kSite has been developed. It can be found at http://103.99.176.239/iMul-kSite where proper guidelines for submitting query protein sequences are provided. Users are allowed to submit query sequences either in the input box or in a batch file. For better understanding, a few protein sequences taken from the independent test dataset are included as examples. In addition to that, the benchmark dataset and the training features used for constructing iMul-kSite will be provided upon user request.

Limitations

To improve the efficiency as well as to reduce the computational complexity of identifying 5 K-PTMs simultaneously, we considered instance hardness threshold (IHT) as an undersampling technique and incremental feature selection (IFS) with ANOVA F-Test as feature selection algorithm. Other structural features and evolutionary features might be utilized to improve the performance. Currently, our predictor iMul-kSite can deal with only five modifications i.e. acetylation, crotonylation, methylation, succinylation and glutarylation. We would include more types of modifications in our future study.

Conclusion

Understanding the significance of identifying multiple lysine PTM sites, an efficient and successful predictor iMul-kSite has been developed with five lysine PTM sites prediction capability. After adopting successful data balancing methods, optimized features with the cost-sensitive learning algorithms have improved the prediction performance of the proposed predictor iMul-kSite significantly. Experimental outcomes demonstrate that iMul-kSite is highly promising compared to the existing state-of-the-art multiple lysine PTM site predictors. It is expected to become a high throughput tool for the experimental researchers for further PTM study on the lysine residues. Even experimental scientists may use this web-based tool without knowing its implementation details. Besides, a similar methodology of the proposed predictor can be used in the study of other PTMs such as C-PTM, R-PTM, and S-PTM that correspond to multi-label PTM sites at Cys, Arg, and Ser residues respectively. However, iMul-kSite was designed for five K-PTM types. To extend its prediction capability, other PTM types with new protein sequences can be added in the future.

Received: 29 June 2021; Accepted: 8 September 2021
Published online: 23 September 2021

References

1. Saraswathy, N. & Ramalingam, P. *Concepts and Techniques in Genomics and Proteomics* (Elsevier, Amsterdam, 2011).
2. McDowell, G. & Philpott, A. New insights into the role of ubiquitylation of proteins. In *International Review of Cell and Molecular Biology*, Vol. 325, 35–88 (Elsevier, 2016).
3. Weissman, J. D., Raval, A. & Singer, D. S. Assay of an intrinsic acetyltransferase activity of the transcriptional coactivator CIITA. In *Methods in Enzymology*, Vol. 370, 378–386 (Elsevier, 2003).
4. Chou, K.-C. Impacts of bioinformatics to medicinal chemistry. *Med. Chem.* **11**, 218–234 (2015).
5. Qiu, W.-R., Sun, B.-Q., Xiao, X., Xu, Z.-C. & Chou, K.-C. iPTM-mLys: identifying multiple lysine PTM sites and their different types. *Bioinformatics* **32**, 3116–3123 (2016).
6. Freiman, R. N. & Tjian, R. Regulating the regulators: Lysine modifications make their mark. *Cell* **112**, 11–17 (2003).
7. Xu, Y. & Chou, K.-C. Recent progress in predicting posttranslational modification sites in proteins. *Curr. Top. Med. Chem.* **16**, 591–603 (2016).
8. Jia, J., Liu, Z., Xiao, X., Liu, B. & Chou, K.-C. iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. *Anal. Biochem.* **497**, 48–56 (2016).
9. Rahman, A., Ahmed, S., Rahman, J. & Hasan, M. A. M. Prediction of formylation sites by incorporating sequence coupling into general PseAAC. In *2020 IEEE Region 10 Symposium (TENSYMP)*, 921–924 (IEEE, 2020).
10. Ahmed, S. *et al.* predPhogly-Site: Predicting phosphoglycerylation sites by incorporating probabilistic sequence-coupling information into PseAAC and addressing data imbalance. *PLoS ONE* **16**, e0249396 (2021).
11. Wu, M., Yang, Y., Wang, H. & Xu, Y. A deep learning method to more accurately recall known lysine acetylation sites. *BMC Bioinform.* **20**, 49 (2019).
12. Ju, Z. & He, J.-J. Prediction of lysine glutarylation sites by maximum relevance minimum redundancy feature selection. *Anal. Biochem.* **550**, 1–7 (2018).
13. Bao, W., Yang, B. & Chen, B. 2-hydr_ensemble: Lysine 2-hydroxyisobutyrylation identification with ensemble method. *Chemom. Intell. Lab. Syst.* 104351 (2021).

14. Bao, W. *et al.* Cmsenn: Computational modification sites with ensemble neural network. *Chemom. Intell. Lab. Syst.* **185**, 65–72 (2019).
15. Chou, K.-C. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J. Biol. Chem.* **268**, 16938–16948 (1993).
16. Chou, K.-C. Prediction of human immunodeficiency virus protease cleavage sites in proteins. *Anal. Biochem.* **233**, 1–14 (1996).
17. Lin, W.-Z., Fang, J.-A., Xiao, X. & Chou, K.-C. iDNA-Prot: Identification of DNA binding proteins using random forest with grey model. *PLoS ONE* **6**, e24756 (2011).
18. Hasan, M. A. M. & Ahmad, S. mLysPTMpred: Multiple lysine PTM site prediction using combination of SVM with resolving data imbalance issue. *Nat. Sci.* **10**, 370–384 (2018).
19. Sua, J. N. *et al.* Incorporating convolutional neural networks and sequence graph transform for identifying multilabel protein lysine PTM sites. *Chemom. Intell. Lab. Syst.* **206**, 104171 (2020).
20. Zhe, J. & Wang, S.-Y. Prediction of 2-hydroxyisobutyrylation sites by integrating multiple sequence features with ensemble support vector machine. *Comput. Biol. Chem.* **87**, 107280 (2020).
21. Tung, C.-W. Prediction of pupylation sites using the composition of k-spaced amino acid pairs. *J. Theor. Biol.* **336**, 11–17 (2013).
22. Chen, D., Liu, Z., Ma, X. & Hua, D. Selecting genes by test statistics. *BioMed Res. Int.* **2005**, 132–138 (2005).
23. Ju, Z. & Wang, S.-Y. iLys-Khib: Identify lysine 2-Hydroxyisobutyrylation sites using mRMR feature selection and fuzzy SVM algorithm. *Chemom. Intell. Lab. Syst.* **191**, 96–102 (2019).
24. Veropoulos, K. *et al.* Controlling the sensitivity of support vector machines. *Proc. Int. Joint. Conf. AI* **55**, 60 (1999).
25. Consortium, U. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
26. Chou, K.-C. Prediction of signal peptides using scaled window. *Peptides* **22**, 1973–1979 (2001).
27. Smith, M. R., Martinez, T. & Giraud-Carrier, C. An instance level analysis of data complexity. *Mach. Learn.* **95**, 225–256 (2014).
28. Le, T. *et al.* A cluster-based boosting algorithm for bankruptcy prediction in a highly imbalanced dataset. *Symmetry* **10**, 250 (2018).
29. Vapnik, V. *The Nature of Statistical Learning Theory* (Springer, Berlin, 2013).
30. Ju, Z. & Wang, S.-Y. Prediction of lysine formylation sites using the composition of k-spaced amino acid pairs via Chou's 5-steps rule and general pseudo components. *Genomics* **112**, 859–866 (2020).
31. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
32. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
33. Atchley, W. R., Zhao, J., Fernandes, A. D. & Drüke, T. Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci.* **102**, 6395–6400 (2005).
34. Ju, Z. & He, J.-J. Prediction of lysine propionylation sites using biased SVM and incorporating four different sequence features into Chou's PseAAC. *J. Mol. Gr. Model.* **76**, 356–363 (2017).
35. Ju, Z. & Cao, J.-Z. Prediction of protein N-formylation using the composition of k-spaced amino acid pairs. *Anal. Biochem.* **534**, 40–45 (2017).
36. Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **273**, 236–247 (2011).
37. Du, P., Wang, X., Xu, C. & Gao, Y. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem.* **425**, 117–119 (2012).
38. Zhang, Z. *et al.* Identification of lysine succinylation as a new post-translational modification. *Nat. Chem. Biol.* **7**, 58 (2011).
39. Kutner, M. H. *et al. Applied Linear Statistical Models* Vol. 5 (McGraw-Hill Irwin, New York, 2005).
40. Hasan, M. A. M., Ahmad, S. & Molla, M. K. I. iMulti-HumPhos: A multi-label classifier for identifying human phosphorylated proteins using multiple kernel learning based support vector machines. *Mol. BioSyst.* **13**, 1608–1618 (2017).
41. Ruan, X., Zhou, D., Nie, R. & Guo, Y. Predictions of apoptosis proteins by integrating different features based on improving pseudo-position-specific scoring matrix. *BioMed Res. Int.* 2020 (2020).
42. Ma, Y., Yu, Z., Han, G., Li, J. & Anh, V. Identification of pre-microRNAs by characterizing their sequence order evolution information and secondary structure graphs. *BMC Bioinform.* **19**, 521 (2018).
43. Batuwita, R. & Palade, V. Efficient resampling methods for training support vector machines with imbalanced datasets. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, 1–8 (IEEE, 2010).
44. Chandra, A., Sharma, A., Dehzangi, A., Shigemizu, D. & Tsunoda, T. Bigram-PGK: Phosphoglycerylation prediction using the technique of bigram probabilities of position specific scoring matrix. *BMC Mol. Cell Biol.* **20**, 1–9 (2019).
45. Chou, K.-C. Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst.* **9**, 1092–1100 (2013).
46. Jiang, M. & Cao, J.-Z. Positive-Unlabeled learning for pupylation sites prediction. *BioMed Res. Int.* 2016 (2016).
47. Hasan, M. A. M., Ahmad, S. & Molla, M. K. I. Protein subcellular localization prediction using multiple kernel learning based support vector machine. *Mol. BioSyst.* **13**, 785–795 (2017).
48. Semwal, V. B., Singha, J., Sharma, P. K., Chauhan, A. & Behera, B. An optimized feature selection technique based on incremental feature analysis for bio-metric gait data classification. *Multim. Tools Appl.* **76**, 24457–24475 (2017).
49. Chang, C.-C. & Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**, 27:1-27:27 (2011).
50. Torkamani, A. & Schork, N. J. Accurate prediction of deleterious protein kinase polymorphisms. *Bioinformatics* **23**, 2918–2925 (2007).
51. Ju, Z. & Wang, S.-Y. Computational identification of lysine glutarylation sites using positive-unlabeled learning. *Curr. Genomics* **21**, 204–211 (2020).
52. Chen, Y.-Z., Tang, Y.-R., Sheng, Z.-Y. & Zhang, Z. Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs. *BMC Bioinform.* **9**, 101 (2008).

Acknowledgements

The authors would like to thank Dr. Alok Sharma, Dr. Iman Dehzangi and Dr. Swakkhar Shatabda for their valuable suggestions. All the figures have been generated by using free and open-source software and libraries, such as Diagrams.net (https://www.diagrams.net/), Matlab 2019a and Matplotlib 3.2.2.

Author contributions

Sabit Ahmed conceived the experiments. Both Sabit Ahmed and Afrida Rahman conducted the experiments and wrote the manuscript. Md. Al Mehedi Hasan Hasan supervised the experiments, analyzed the results, and reviewed the manuscript. Shamim Ahmad provided resources and reviewed the manuscript. S. M. Shovan helped to construct the dataset and preparing the manuscript.

Funding

The authors have received no specific funding for this work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-98458-y.

Correspondence and requests for materials should be addressed to S.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com

predPhogly-Site: Predicting phosphoglycerylation sites by incorporating probabilistic sequence-coupling information into PseAAC and addressing data imbalance

Article *in* PLoS ONE · April 2021

DOI: 10.1371/journal.pone.0249396

CITATIONS
6

READS
135

6 authors, including:

Sabit Ahmed
Rajshahi University of Engineering & Technology
7 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Afrida Rahman
Rajshahi University of Engineering & Technology
7 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Md. Al Mehedi Hasan
Rajshahi University of Engineering & Technology
117 PUBLICATIONS **713** CITATIONS
SEE PROFILE

Md. Khaled Ben Islam
Pabna University of Science and Technology
24 PUBLICATIONS **87** CITATIONS
SEE PROFILE

PLOS ONE

RESEARCH ARTICLE

predPhogly-Site: Predicting phosphoglycerylation sites by incorporating probabilistic sequence-coupling information into PseAAC and addressing data imbalance

Sabit Ahmed[1☯]*, **Afrida Rahman**[1☯], **Md. Al Mehedi Hasan**[1], **Md Khaled Ben Islam**[2], **Julia Rahman**[1¤], **Shamim Ahmad**[3]

1 Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh, **2** Computer Science and Engineering, Pabna University of Science and Technology, Pabna, Bangladesh, **3** Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh

☯ These authors contributed equally to this work.
¤ Current address: Institute for Integrated and Intelligent Systems, Griffith University, Brisbane, Australia
* sabit.a.sirat@gmail.com

Citation: Ahmed S, Rahman A, Hasan MAM, Islam MKB, Rahman J, Ahmad S (2021) predPhogly-Site: Predicting phosphoglycerylation sites by incorporating probabilistic sequence-coupling information into PseAAC and addressing data imbalance. PLoS ONE 16(3): e0249396. https://doi.org/10.1371/journal.pone.0249396

Editor: Ozlem Keskin, Koç University, TURKEY

Received: October 1, 2020

Accepted: March 18, 2021

Published: April 1, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0249396

Data Availability Statement: All relevant data are within the paper and its Supporting information files.

Abstract

Post-translational modification (PTM) involves covalent modification after the biosynthesis process and plays an essential role in the study of cell biology. Lysine phosphoglycerylation, a newly discovered reversible type of PTM that affects glycolytic enzyme activities, and is responsible for a wide variety of diseases, such as heart failure, arthritis, and degeneration of the nervous system. Our goal is to computationally characterize potential phosphoglycerylation sites to understand the functionality and causality more accurately. In this study, a novel computational tool, referred to as predPhogly-Site, has been developed to predict phosphoglycerylation sites in the protein. It has effectively utilized the probabilistic sequence-coupling information among the nearby amino acid residues of phosphoglycerylation sites along with a variable cost adjustment for the skewed training dataset to enhance the prediction characteristics. It has achieved around 99% accuracy with more than 0.96 MCC and 0.97 AUC in both 10-fold cross-validation and independent test. Even, the standard deviation in 10-fold cross-validation is almost negligible. This performance indicates that predPhogly-Site remarkably outperformed the existing prediction tools and can be used as a promising predictor, preferably with its web interface at http://103.99.176.239/predPhogly-Site.

Introduction

Post-translational modifications (PTM) refer to specific events after the translation stage, where the covalent inclusion of specific functional groups occurs in a protein [1]. These modifications have enormous impacts on biological processes and proteomic analysis, such as cellular signal transduction, subcellular localization, protein folding, protein degradation, and are

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

also responsible for various kinds of diseases [2]. Therefore, accurate identification and effective comprehension of PTM sites are significant for basic research in disease detection, prevention, and various drug developments [3]. Among the 20 standard constituent amino acid residues of cellular proteins, modifications at lysine residue (K) are commonly known as lysine PTM or K-PTM. According to the literature, several K-PTMs such as acetylation, crotonylation, ubiquitination, phosphoglycerylation, glycation, methylation, butyrylation, succinylation, biotinylation can be aided by these covalent modifications [4–8].

Lysine phosphoglycerylation is one of the reversible post-translational modifications, newly discovered in mouse liver and human cells [8, 9]. The formation of 3-phosphoglyceryl-lysine (pgK) takes place when primary glycolytic intermediate (1,3-BPG) interacts with particular lysine residues [8, 10]. A wide variety of diseases, including heart failure, arthritis, and various types of neurodegenerative disorders can be caused by this phosphoglycerylation. Metabolic labeling with substantial glucose indicates that it can be derived from glucose metabolism [9]. It has significant effects on glycolytic enzyme activities and can build up on cells with high glucose exposure [11]. Potential feedback mechanism that contributes to the creation and redirection of glycolytic intermediates to specific biosynthetic pathways is also established [8–11]. Concerning the crucial role of phosphoglycerylation in such biological processes, the effective way to characterize its functional aspects is to identify phosphoglycerylation sites with higher efficacy. Although high throughput experimental procedures to characterize phosphoglycerylation sites are known to achieve higher accuracy, computational methods are getting popularity as an effective alternative because of their laborsaving, time and cost-efficient characteristics.

Recent studies on identifying phosphoglycerylation sites have introduced several computational tools such as, Phogly-PseAAC [9], CKSAAP_PhoglySite [8], iPGK-PseAAC [12] and Bigram-PGK [11]. The first one has applied a KNN-based predictor with the pseudo amino acid feature source [9], where the second one has implemented a fuzzy SVM based predictor with the formation of k-spaced amino acid pairs feature set [8]. iPGK-PseAAC has utilized the pairwise coupling technique with an SVM classifier [11, 12]. The most recently developed predictor, Bigram-PGK has employed SVM with evolutionary information of the sequences for performance improvement [11]. Among these four predictors, only Bigram-PGK can predict phosphoglycerylation sites with an AUC higher than 0.90. However, the overall performance of this predictor needs further improvement in terms of other measurement metrics to be used as a complementary phosphoglycerylation site identification technique.

For constructing an efficient predictor, appropriate informative patterns connected with phosphoglycerylation need to be extracted. In this study, we are introducing a novel computational tool predPhogly-Site for predicting phosphoglycerylation sites by blending vectorized sequence coupling information with PseAAC [3, 13–16]. After generating necessary features from the protein sequences adopted from Bigram-PGK [11], a cost-sensitive SVM [14, 17–19] classifier has been used to predict phosphoglycerylation sites by minimizing class-level imbalance in benchmark dataset. The workflow of our proposed predictor is shown in Fig 1. For validating the statistical significance of the results, 10-fold cross-validation has been repeated ten times, and the average performances of each evaluation metric have been reported in the Results section. It can be observed that our proposed predictor, predPhogly-Site has achieved superior prediction performance than all the existing predictors. The attained performance of predPhogly-Site in terms of specificity, sensitivity, precision, accuracy, MCC, and AUC are 99.97%, 100%, 99.20%, 99.97%, 99.58%, and 99.99%, respectively. The promising results obtained by predPhogly-Site indicates that it can be used as a high-throughput supporting tool for phosphoglycerylation site prediction.

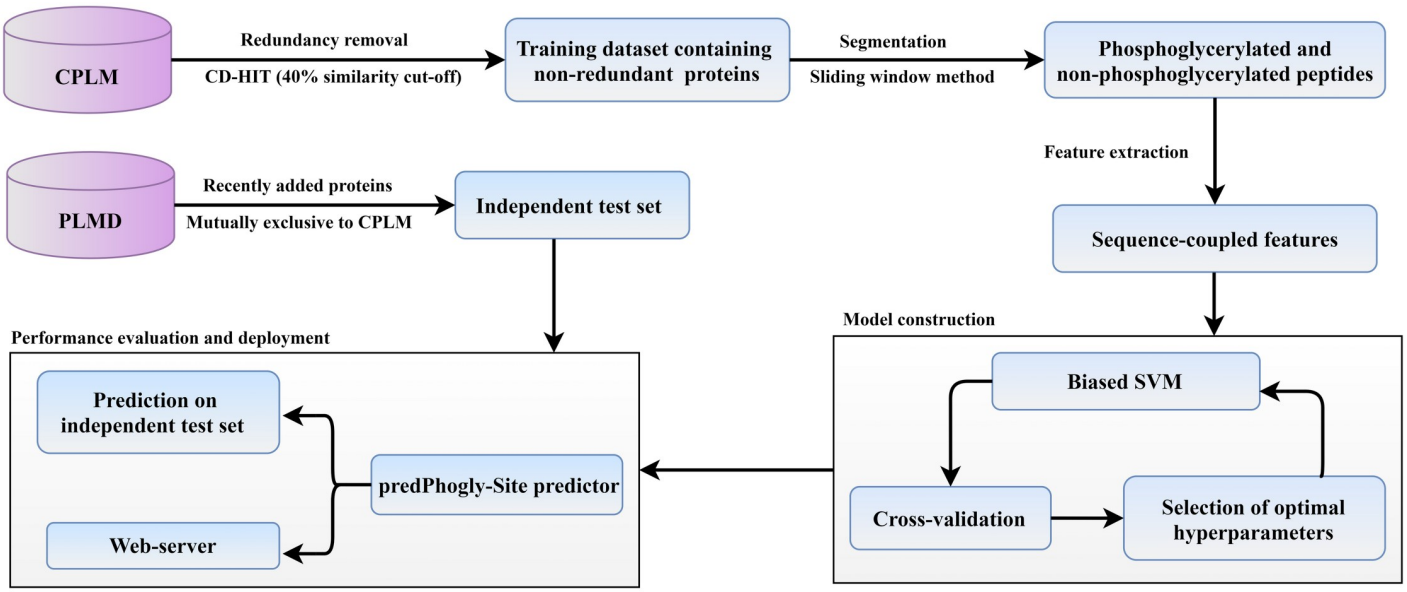

Fig 1. An overview of predPhogly-Site for phosphoglycerylation site prediction.

https://doi.org/10.1371/journal.pone.0249396.g001

Highlighted in a series of recently published predictors [3, 6, 14, 19–23], to develop an efficient predictor with regards to computational biology, one should go through Chou's five-step [14, 24, 25] guidelines: i) generating an acceptable benchmark dataset for training and testing the system, ii) formulating the sequences using proper mathematical representations, iii) developing a prediction approach or introducing a robust prediction algorithm, iv) conducting rigorous cross-validation tests to evaluate predictive accuracy, and v) providing an accessible and easy-to-use web-server. Following these steps, details of materials, methods, results, and analysis will be discussed in the following sections.

Materials and methods

Dataset

In this study, verified annotations of phosphoglycerylation sites were obtained from the CPLM version 2.0 [26], one of the reliable repositories of post-translational modification in lysine residue, and corresponding protein sequences were retrieved from UniProt knowledge-base [27] for developing the prediction model. Subsequently, redundant sequences were discarded with 40% similarity cutoff using CD-HIT [28] for avoiding bias in performance evaluation as this level of redundancy removal was widely accepted [11, 24, 29, 30]. As a result, a total of 91 non-redundant proteins were held out for constructing a benchmark dataset. There were 111 experimentally annotated phosphoglycerylated sites and 3249 non-phosphoglycerylated sites, which was identical to the most recent predictor, Bigram-PGK's [11] dataset (see Table 1). The benchmark dataset containing protein sequences and site positions are given in S1 File. An overview of the dataset preparation as part of the prediction model development is presented

Table 1. Summary of the non-redundant phosphoglycerylation dataset.

Similarity threshold	No. of non-redundant proteins	Phosphoglycerylated sites	Non-phosphoglycerylated sites
40%	91	111	3249

https://doi.org/10.1371/journal.pone.0249396.t001

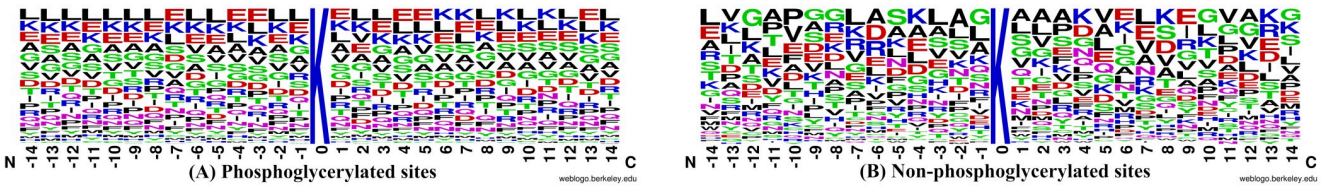

Fig 2. Amino acid frequencies around the K-PTM and non-K-PTM sites.

https://doi.org/10.1371/journal.pone.0249396.g002

in Fig 1. For verifying the statistically significant difference among the positive and negative sites in the obtained dataset, the distribution of amino acid residues in the phosphoglycerylated sites and non-phosphoglycerylated sites are visually analyzed with the help of WebLogo [31] (see Fig 2A and 2B).

To demonstrate the viability of the proposed predictor predPhogly-Site for new proteins, an independent test set was constructed with recent phosphoglycerylation sites, utterly unknown to the benchmark dataset used for prediction model development. Protein sequences with recent phosphoglycerylation sites were collected from the PLMD database [32] (version 3.0), which is an upgraded version of the CPLM database [26], released nearly 03 years later with many newly discovered PTM sites. For ensuring the non-existence of training proteins in the independent test set, we considered only those proteins which were newly added to the PLMD repository much after the creation of the benchmark dataset with verified phosphoglycerylation sites. Therefore, we obtained 33 proteins with 41 phosphoglycerylated sites and 1334 non-phosphoglycerylated sites for the independent test (available as S2 File). Furthermore, the non-existence of recent test sites was verified manually for avoiding accidental bias in performance benchmarking.

Feature construction

To formulate the phosphoglycerylation site sequences more meticulously and comprehensively, Chou's scheme [9, 13, 33] was adopted. According to this scheme, a potential phosphoglycerylation site containing sequence fragment could be expressed as:

$$\Theta_{\zeta}(K) = Q_1 Q_2 \dots Q_{\zeta-1} Q_{\zeta} K Q_{\zeta+1} Q_{\zeta+2} \dots Q_{2\zeta-1} Q_{2\zeta} \quad (1)$$

Where Q_1 to Q_{ζ} denote the leftward and $Q_{\zeta+1}$ to $Q_{2\zeta+1}$ denote the rightward amino acid residues, respectively, while ζ being an integer and centered 'K' indicating "lysine" [14]. Furthermore, the peptide sequences $\Theta_{\zeta}(K)$ can be categorized into two types: $\Theta_{\zeta}^{+}(K)$ and $\Theta_{\zeta}^{-}(K)$, where the first one denotes phosphoglycerylated peptide and the later one denotes non-phosphoglycerylated peptide with a lysine residue at its center [9, 14]. The sliding window method [9] was adopted to segment the phosphoglycerylation protein sequences with different window size where ζ = 1, 2, 3, . . .32. Based on the MCC value, window size was selected as $(2\zeta + 1) = 29$ where $\zeta = 14$ (i.e. 14 rightstream and 14 leftstream amino acid residues). It should be mentioned that, only the window sizes less than 65 were taken under consideration due to the compelling protein sequence length [11]. With a sequence fragment of window size 29, Eq (1) could be expressed as:

$$\Theta(K) = Q_1 Q_2 \dots Q_{13} Q_{14} K Q_{15} Q_{16} \dots Q_{27} Q_{28} \quad (2)$$

At the time of segmentation, for making site sequences' of equal length, the lacking amino acids were filled with 'X' residue [9, 34]. As a result, the phosphoglycerylation dataset had

taken the following form:

$$S_\zeta(\mathrm{K}) = S_\zeta^+(\mathrm{K}) \cup S_\zeta^-(\mathrm{K}) \tag{3}$$

where the positive subset $S_\zeta^+(\mathrm{K})$ could contain only $\Theta_\zeta^+(K)$ samples, while the negative subset $S_\zeta^-(\mathrm{K})$ could contain only $\Theta_\zeta^-(K)$ samples with their center residue K. All the segmented sequences with the expression of Eqs (2) and (3) are provided in S1 File.

For extracting pertinent features hidden in amino acid sequences, different sequence encoding methods such as amino acid composition, pseudo amino acid composition were used initially. However, in the proposed predictor predPhogly-Site, the vectorized sequence-coupled model [3, 14–16, 35] has been incorporated into general PseAAC [3, 14, 33, 35–39] to extract features from the phosphoglycerylation sites conserving the sequence pattern information. According to this conception, the peptide sample in Eq (2) can be expressed as:

$$\Theta(K) = \Theta^+(K) - \Theta^-(K) \tag{4}$$

where,

$$\Theta^+(K) = \begin{bmatrix} \Theta^+(Q_1|Q_2) \\ \Theta^+(Q_2|Q_3) \\ \vdots \\ \Theta^+(Q_{13}|Q_{14}) \\ \Theta^+(Q_{14}) \\ \Theta^+(Q_{15}) \\ \Theta^+(Q_{16}|Q_{15}) \\ \vdots \\ \Theta^+(Q_{27}|Q_{26}) \\ \Theta^+(Q_{28}|Q_{27}) \end{bmatrix} \qquad \Theta^-(K) = \begin{bmatrix} \Theta^-(Q_1|Q_2) \\ \Theta^-(Q_2|Q_3) \\ \vdots \\ \Theta^-(Q_{13}|Q_{14}) \\ \Theta^-(Q_{14}) \\ \Theta^-(Q_{15}) \\ \Theta^-(Q_{16}|Q_{15}) \\ \vdots \\ \Theta^-(Q_{27}|Q_{26}) \\ \Theta^-(Q_{28}|Q_{27}) \end{bmatrix} \tag{5}$$

where, $\Theta^+(Q_1|Q_2)$ denotes the conditional probability of amino acid Q_1 at the leftmost position given that its adjacent right member is Q_2 and the same applies for remaining indices of leftward residues [24]. Similarly, $\Theta^+(Q_{28}|Q_{27})$ denotes the conditional probability of amino acid Q_{28} at the rightmost position given that its adjacent left member is Q_{27} and so forth. In contrast, only $\Theta^+(Q_{14})$ and $\Theta^+(Q_{15})$ are of non-conditional probability as K is the adjoining member of both amino acids Q_{14} and Q_{15} [3, 6, 14, 15, 24]. In order to calculate the probability values of $\Theta^+(Q_{14})$ and $\Theta^+(Q_{15})$, firstly, we have to find the frequency of a given amino acid Q_{14} and Q_{15} from the set of phosphoglycerylated peptides [15]. Then the obtained values should be divided by the frequency of all amino acids occurring at position 14 and 15 respectively. Accordingly, $\Theta^-(K)$ in Eq (5), with its probabilistic components could also be deduced from the set of non-phosphoglycerylated peptides. A few literature on vectorized sequence-coupling model [3, 13, 15, 16] could provide a better understanding of the procedure of probability calculation out of any dataset. Finally, a 28-dimensional feature vector was obtained by using Eqs 4 and 5 for each potential phosphoglycerylated and non-phosphoglycerylated sample.

For better visualization and insights on the sequence-coupling effects at different positions of any sample, we have stored all possible combinations of conditional probability values extracted from the positive subset i.e. $\Theta^+(Q_1|Q_2)$ to $\Theta^+(Q_{13}|Q_{14})$ and $\Theta^+(Q_{16}|Q_{15})$ to $\Theta^+(Q_{28}|Q_{27})$ in one data frame (available in S3 File) and non-conditional probability values for each

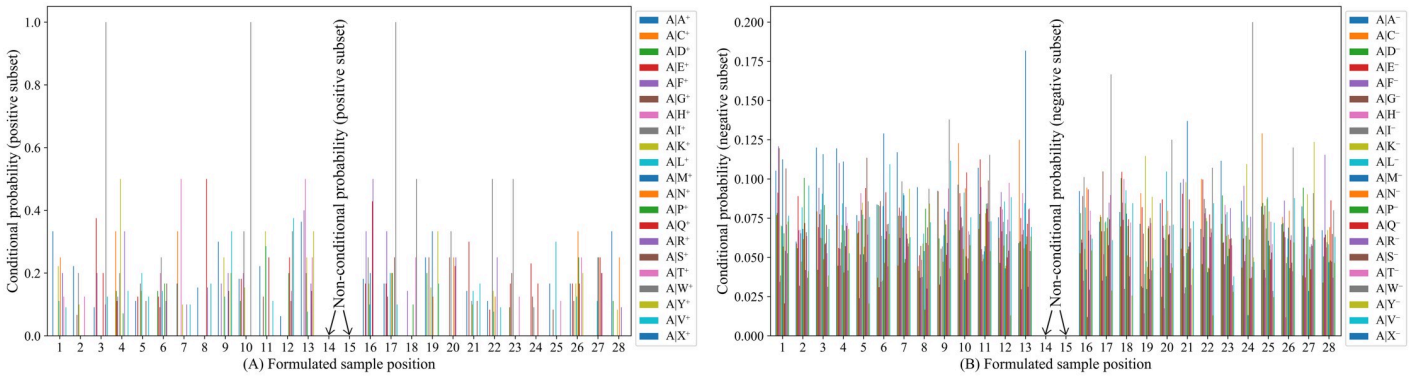

Fig 3. The conditional probability of amino acids at sample positions 1 to 13 and 15 to 28.

https://doi.org/10.1371/journal.pone.0249396.g003

amino acid residue extracted from the positive subset i.e. $\Theta^{+}(Q_{14})$ and $\Theta^{+}(Q_{15})$ in another data frame (available in S4 File) using Pandas library [40], where the columns represent the formulated sample positions and the rows represent the amino acid residues. It should be mentioned that there could be $21 \times 21 = 441$ (including the dummy amino acid residue $'X'$) possible combinations of conditional probability values and 21 non-conditional probability values [15] for each position at any formulated sample. Similarly, the conditional and non-conditional probability values extracted from the negative subset are stored in two separate data frames and provided in S3 and S4 Files, respectively. Fig 3A depicts the conditional probability values of amino acid residue $'A'$ which have been calculated from the positive subset, given that its right member is any of the 21 amino acid residues at sample positions 1 to 13 and the conditional probability values of any of the 21 amino acid residue given that the left member is $'A'$ at sample positions 16 to 28. Similarly, Fig 3B depicts the conditional probability values of amino acid residue $'A'$ which have been calculated from the negative subset, given that its right member is any of the 21 amino acid residues at sample positions 1 to 13 and the conditional probability values of any of the 21 amino acid residue given that the left member is $'A'$ at sample positions 16 to 28. The non-conditional probability values of 21 amino acid residues derived from the positive subset at sample positions 14 and 15 are illustrated in Fig 4A and The non-conditional probability values of 21 amino acid residues derived from the negative subset at sample position 14 and 15 are shown in Fig 4B.

Prediction method and addressing data imbalance

Phosphoglycerylation site prediction problem defined in the previous section is a classification problem. Statistical learning algorithms such as k-nearest neighbor [41], random forest [42] which are widely used in different bioinformatic prediction model development, support vector machine (SVM) [43, 44] is one of the dominant and successful among these algorithms [24, 45]. Apart from that, the structural risk minimization involves a biasing problem where the majority class [24, 46] influences the classification weight. As the set of phosphoglycerylation peptides was highly skewed (i.e. the ratio between positive and negative peptides was approximately 1:29), it could affect the classification model training directly. Inspired by the success of biasing internal decision function during training, as highlighted in recent research [8, 14, 17, 19], different penalty costs C^{+}, and C^{-} were assigned for phosphoglycerylated sites and non-phosphoglycerylated sites, respectively for addressing imbalance issue. Therefore, SVM with cost-sensitivity was applied as a core learning algorithm for prediction model

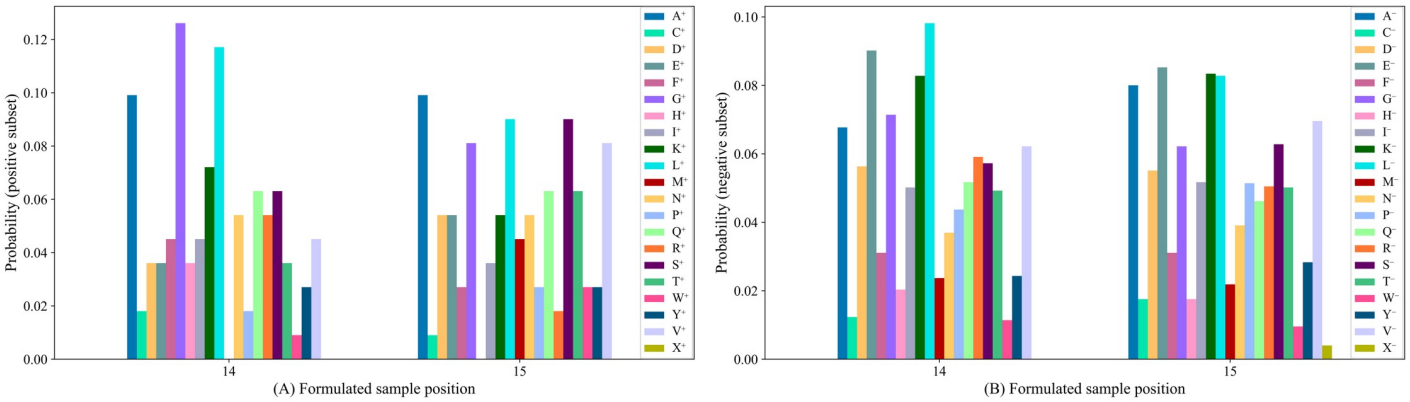

Fig 4. Probabilistic information of 21 amino acids at sample positions 14 and 15.

https://doi.org/10.1371/journal.pone.0249396.g004

development which can be formulated as:

$$min_{w,\xi}^{\frac{1}{2}}\|w\|^2 + C^+\sum_{k=1}^{q}\xi_k + C^-\sum_{k=q+1}^{n}\xi_k \tag{6}$$

(Subject to: $Y_k(w.\varphi(X_k) + a) \geq 1 - \xi_k$ for all, $k = 1, 2, .., n$)
where the training set is denoted by $\{(X_k, Y_k), k = 1, 2, \ldots, n\}$ and first q samples (i.e. $Y_k = 1$, $k = 1, 2, \ldots, q$) are assumed as the positive samples while the rest are assumed as the negative samples (i.e. $Y_k = -1$, $k = q + 1, q + 2, \ldots, n$). The non-linear feature mapping and slack variables are denoted by $\varphi(X)$ and $\xi_k(k = 1, 2, \ldots, n)$, respectively [45, 47]. In our experiments with SVM, as the kernel function, Gaussian RBF was adopted which can be described as: $\Upsilon(X_k, X_j) = \varphi(X_k)^T \varphi(X_j) = exp(-\gamma\|x_i - x_j\|^2)$, where $\gamma > 0$. However, for effective separation of positive and negative samples, addressing the class imbalance problem, misclassification costs $C^+ = \frac{C*n}{2*q}$ and $C^- = \frac{C*n}{2*(n-q)}$ were assigned for phosphoglycerylated sites and non-phosphoglycerylated sites, respectively.

Formulation of evaluation metrics

To objectively assess the prediction performance of predPhogly-Site, we have utilized five widely used statistical metrics, such as accuracy (ACC),sensitivity (Sn), specificity (Sp), precision (pre) and Matthew's Correlation Coefficient (MCC) [20, 24, 30, 45, 47–52]. These matrices can be defined in terms of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) prediction made by the predictor as following:

$$\begin{cases} Sn = \dfrac{TP}{TP + FN} \\ Sp = \dfrac{TN}{TN + FP} \\ Precision = \dfrac{TP}{TP + FP} \\ ACC = \dfrac{TP + TN}{TP + TN + FP + FN} \\ MCC = \dfrac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \end{cases} \tag{7}$$

To the best of our knowledge, state-of-the-art phosphoglycerylation site predictors [8, 9, 11, 12] have also estimated their performance based on these metrics. Thus, performance assessment using these metrics was essential to establish a fair comparative benchmarking. Eventually, we have considered the area under the ROC curve (AUC) [24, 53] in addition to MCC for illustrating the stability and robustness of the prediction model.

Validation of the proposed model

To evaluate the statistical significance of a novel predictor's anticipated performance, three validation schemes, such as k-fold cross-validation, jackknife test, and independent test are widely used [14, 24]. Although the jackknife test can always draw out a unique result for a given dataset and highly desirable, to reduce the computational complexity of model development, researchers prefer k-fold cross-validation over the jackknife test for validating their PTM prediction models [8, 45]. Moreover, existing phosphoglycerylation site predictors validated their anticipated accuracy using k-fold cross validation except Phogly-PseAAC [9]. Even, the most recent predictor, Bigram-PGK [11] validated their model using 10-fold cross-validation and compared with existing predictors. Therefore, to develop and validate our proposed predictor predPhogly-Site, 10-fold cross-validation was adopted. However, as the 10-fold cross-validation involved some arbitrariness, highlighted in [9, 24], to validate the stability, it was repeatedly executed for 10 times. For finding the best performing predictor, a set of prediction models were generated for the hyperparameters C and γ within the grid of $C = \{2^0, 2^1, 2^2, \ldots, 2^8\}$ and $\gamma = \{2^{-1}, 2^{-2}, 2^{-3}, \ldots, 2^{-8}\}$. Using 10-fold cross-validation with 10 repeats, the best model with optimal hyperparameters C and γ were selected (see Table 2) depending on the demonstrated AUC.

The 10-iterations of 10-fold cross-validation were performed according to the following steps:

Step 1: Extract the sequence-coupled features from the segmented sequences provided in S1 File using Eqs (4) and (5).

Step 2: Divide the extracted dataset randomly into 10 disjoint sets.

Step 3: Select 1 set as test set and utilize the remaining 9 sets as training set.

Step 4: Train the RBF kernel based SVM predictor with the training set using the optimal hyperparameters (C, γ) of the respective iteration (see Table 2).

Step 5: Perform prediction on the test set.

Step 6: Repeat steps 2 to 5 until all 10 sets had been used for testing.

Step 7: Merge the prediction outputs and measure the performance with Eq 7.

Step 8: Repeat steps 1 to 7 for 10 times.

Table 2. Selected parameters of 10-fold cross validation (10 iterations).

Iteration	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}
C	2^0	2^0	2^0	2^0	2^0
γ	2^{-1}	2^{-2}	2^{-2}	2^{-2}	2^{-2}
Iteration	6^{th}	7^{th}	8^{th}	9^{th}	10^{th}
C	2^1	2^2	2^2	2^0	2^0
γ	2^{-1}	2^{-2}	2^{-2}	2^{-2}	2^{-2}

https://doi.org/10.1371/journal.pone.0249396.t002

Step 9: Measure the average performance of 10 repetitions with corresponding standard deviations.

The predictive decision-making workflow of predPhogly-Site is available at https://github.com/Sabit-Ahmed/predPhogly-Site as a git repository. For additional validation, an independent test was performed on a set of recent phosphoglycerylation sites. It will be discussed thoroughly in the next section.

Results and discussions

Performance of predPhogly-Site

In this work, we employed SVM with variable cost adjustments [14, 19, 24] for suppressing the imbalance between phosphoglycerylated and non-phosphoglycerylated sites. For separating samples by transforming to higher dimensional feature space, radial basis kernel function [14, 22, 24] was utilized. The average results of the considered statistical performance measures with their standard deviations in 10 repeats are presented in Table 3. As shown in Table 3, the proposed prediction model could predict phosphoglycerylation sites with 99.97% accuracy. In addition to that, its sensitivity, specificity, MCC and AUC measure crossed a benchmark of 99%. Moreover, standard deviations were almost negligible in the case of all the measures. However, for constructing the proposed predictor predPhogly-Site to be deployed as a web service, the benchmark dataset and the prediction model's hyper-parameters with the highest AUC in 10 repetitions (i.e. $C = 2^0$ and $\gamma = 2^{-2}$) were used. An overview of establishing predPhogly-Site is depicted in Fig 1.

Comparative analysis of cross-validation performance

To evaluate the effectiveness of the proposed predictor, predPhogly-Site, we compared it with four state-of-the-art phosphoglycerylation site predictors, such as Phogly-PseAAC [9], CKSAAP_PhoglySite [8], iPGK-PseAAC [12] and Bigram-PGK [11]. Among these predictors, the first three i.e. Phogly-PseAAC, CKSAAP_PhoglySite, and iPGK-PseAAC were benchmarked on the same phosphoglycerylation site dataset which was prepared by Xu et al. [9]. Prediction from Phogly-PseAAC and iPGK-PseAAC could be accessed by their web interface. Though CKSAAP_PhoglySite was also accessible by its Matlab interface, there was no such accessibility option in the most recent predictor, Bigram-PGK. However, Bigram-PGK had collected prediction results from these accessible predictors for its benchmark dataset and reported comparative outcomes for all the considered performance metrics. Thus, for conducting a fair comparison with all these predictors, our primary benchmark dataset, which was not resampled as Bigram-PGK's one, was submitted to the webserver of Phogly-PseAAC and iPGK-PseAAC for getting prediction outcomes. However, CKSAAP_PhoglySite's predictions were obtained through its Matlab interface. After achieving the prediction outcomes from the Phogly-PseAAC, CKSAAP_PhoglySite, and iPGK-PseAAC on the benchmark dataset constructed for this study, the corresponding performance was measured on the same validation set utilized for evaluating our predictor predPhogly-Site (see Section "Validation of the proposed model"). As we adopted different technique for handling the data imbalance issue

Table 3. Cross-validation performance of predPhogly-Site on the benchmark dataset.

Predictor	Sp	Sn	Pre	ACC	MCC	AUC
predPhogly-Site	0.9997 ± 0.0001	1.00±0.00	0.9920±0.0027	0.9997±0.0001	0.9958±0.0014	0.9999±0.00

https://doi.org/10.1371/journal.pone.0249396.t003

Table 4. Cross-validation performance of the existing prediction systems.

Predictor	Sp	Sn	Pre	ACC	MCC	AUC
iPGK-PseAAC	0.9846	0.4595	0.5050	0.9673	0.4648	0.7220
iPGK-PseAAC*	0.9864	0.4555	0.9548	0.8119	0.5692	0.7230
CKSAAP_PhoglySite	0.8941	0.8288	0.2110	0.8920	0.3845	0.8615
CKSAAP_PhoglySite*	0.9420	0.8285	0.8765	0.9043	0.7818	0.8854
Phogly-PseAAC	0.7064	0.6937	0.0747	0.7060	0.1550	0.7000
Phogly-PseAAC*	0.7193	0.6927	0.5518	0.7102	0.3951	0.7062
Bigram-PGK*	0.8973	0.9642	0.8253	0.9193	0.8330	0.9306
predPhogly-Site	**0.9997**	**1.00**	**0.9920**	**0.9997**	**0.9958**	**0.9999**

* Corresponds to the experimental findings reported by the Bigram-PGK study [11].

https://doi.org/10.1371/journal.pone.0249396.t004

and could not obtain the prediction outcomes from the Bigram-PGK predictor on our benchmark dataset, a comparative summary of all the measures was presented in Table 4 in line with Bigram-PGK's experimental findings [11]. As shown in Table 4 and Fig 5, predPhogly-Site achieved a significant improvement over Phogly-PseAAC, CKSAAP_PhoglySite, and iPGK-PseAAC on the same benchmark dataset used in this study. It remarkably outperformed these predictors in sensitivity, specificity, overall accuracy, and AUC. For instance, predPhogly-Site crossed the milestone of 99% in case of sensitivity, specificity, precision, overall accuracy, MCC and AUC.

However, the most recent predictor, Bigram-PGK's [11] performance was relatively higher in most of the metrics. It obtained a sensitivity of 96.42%, an accuracy of 91.93%, an MCC of 83.30%, and an AUC of 93.06% on the dataset utilized in Bigram-PGK [11]. As demonstrated in Table 4, our proposed predictor predPhogly-Site also outperformed Bigram-PGK [11] by 3.58% in sensitivity, 8.04% in accuracy measure, 16.28% in MCC and 6.93% in AUC.

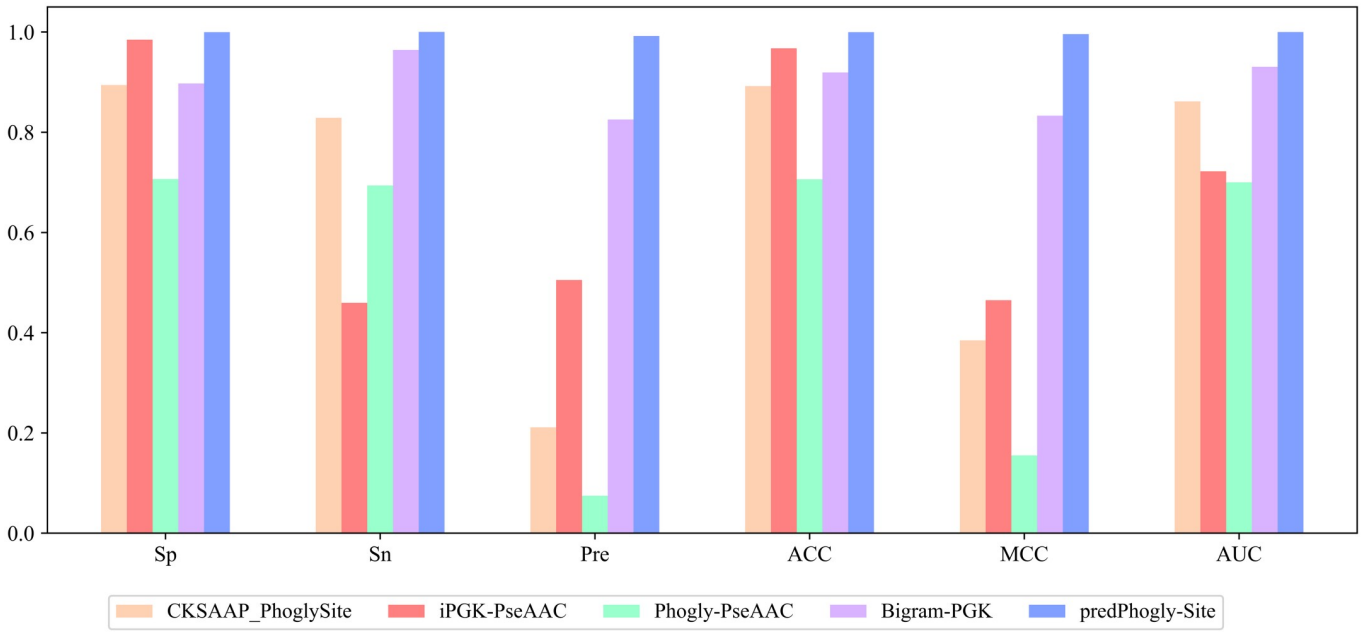

Fig 5. Cross-validation performance of the available predictors.

https://doi.org/10.1371/journal.pone.0249396.g005

Furthermore, the effectiveness of predPhogly-Site over the recent predictors including Bigram-PGK [11] has been demonstrated in Fig 5.

It can be observed that a comparatively higher specificity and precision of 98.64% and 95.48%, respectively, were obtained by iPGK-PseAAC [12] on the Bigram-PGK's [11] resampled dataset. Our proposed predictor, predPhogly-Site, has obtained 1.33% and 3.72% increased performance in both specificity and precision, respectively. Both the results represented in Table 4 and Fig 5 indicate that our proposed predictor predPhogly-Site can identify phosphoglycerylation sites more effectively than any other existing predictors.

It is worth mentioning that among these predictors, Phogly-PseAAC [9] has employed the position-specific amino acid propensity which reflects the position-wise occurrence frequency of each amino acid and the K-Nearest Neighbor (KNN) algorithm for prediction, CKSAAP_PhoglySite [8] has utilized the composition of k-spaced amino acid pairs with the fuzzy SVM, iPGK-PseAAC [12] has applied the pairwise coupling technique with the posterior probability-based SVM and Bigram-PGK [11] have considered the SVM engine with the combination of position-specific scoring matrix and profile bigrams for performance improvement.

It might be intuitive to find some insight into why our proposed predictor predPhogly-Site achieved such superior performance. It was possible because of the effective representation of phosphoglycerylation modification in terms of sequence coupling model among the amino acid residues via the conditional probability (see Figs 3 and 4). Suppressing the imbalance ratio of phosphoglycerylated and non-phosphoglycerylated sites using different error costs based SVM also boosted up the performance improvement.

However, the precision calculation measures the believability of a system when it says a peptide sample is phosphoglycerylated. According to Eq 7, the precision measure depends highly on the false positive rate, and a lower false positive rate results in a higher precision rate. In the Bigram-PGK [11] study, the dataset contained only 111 positive samples and 224 negative samples after applying the k-nearest neighbor cleaning treatment [11] and the experimental findings on the resampled dataset might not reflect the false positive rate properly. Moreover, the existing predictors i.e. iPGK-PseAAC, CKSAAPPhoglySite, and Phogly-PseAAC might not handle the real world imbalanced situation of the dataset appropriately. Hence, when we have uploaded the benchmark dataset containing 111 positive instances and 3249 negative instances (see Table 1) to the web or Matlab interfaces of the existing predictors, the false positive rates have come out higher and results in lower precision rates as compared to the experimental findings reported by the Bigram-PGK study (see Table 4). On the other hand, our proposed predictor has obtained a much lower false positive rate and got a higher precision rate as well as higher sensitivity and specificity for having cost-sensitive SVM as an imbalance management technique. By observing all the performance measurements in this study, it can be concluded that our predictor predPhogly-Site could be a high throughput tool for predicting phosphoglycerylation sites more precisely.

Independent test

Existing phosphoglycerylation site, particularly, the most recent predictor assessed their model using 10-fold cross-validation. However, some researchers [54–57] highlighted the necessity of independent test for assessing prediction model in addition to k-fold (e.g. k = 5,10) cross-validation. Thus, in our work, an independent test was conducted for further evaluation of our proposed model predPhogly-Site on an independent set of phosphoglycerylation sites. The same independent test set was uploaded to the web servers of the existing predictors i.e. iPGK-PseAAC, Phogly-PseAAC and predPhogly-Site for obtaining the prediction results.

Table 5. Prediction performance in Independent test.

Predictor	Sp	Sn	Pre	ACC	MCC	AUC
iPGK-PseAAC	0.9738	0.2927	0.2553	0.9535	0.2494	0.6332
Phogly-PseAAC	0.6837	0.6829	0.0622	0.6836	0.1329	0.6833
CKSAAP_PhoglySite	0.8823	0.7561	0.1649	0.8785	0.3161	0.8192
predPhogly-Site	**0.9993**	**0.9512**	**0.9750**	**0.9978**	**0.9619**	**0.9752**

https://doi.org/10.1371/journal.pone.0249396.t005

However, the prediction results of CKSAAP_PhoglySite on the independent test set were obtained from the Matlab interface. The predictive performance of predPhogly-Site as well as other predictors were summarized in Table 5. However, as Bigram-PGK [11] had no established web-server, so we could not report the performance of these predictors on the independent test set.

As shown in Table 5, predPhogly-Site predicted independent phosphoglycerylation sites with specificity, sensitivity, precision, accuracy, MCC and AUC of 99.93%, 95.12%, 97.50%, 99.78%, 96.19% and 97.52%, respectively, which were almost identical to the cross-validation performance delineated in Table 4. According to the experimental results in Table 5 and the ROC curve illustrated in Fig 6, it was apparent that the proposed predictor predPhogly-Site achieved a significant improvement over their counterparts in terms of all the evaluation metrics.

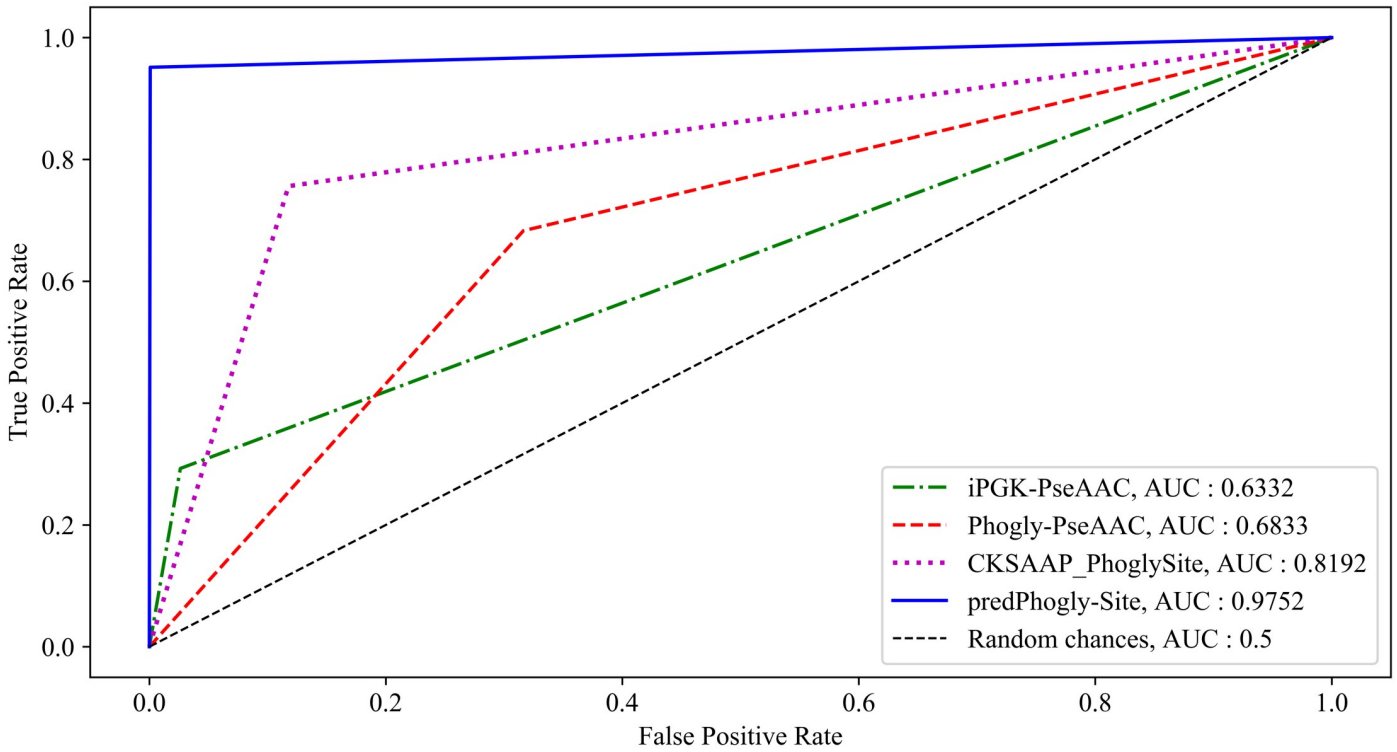

Fig 6. Comparative ROC curves between different prediction methods based on the independent test.

https://doi.org/10.1371/journal.pone.0249396.g006

Web-server

For intensifying user accessibility without the concern of experimental implementations, an easy-to-use web-server for predPhogly-Site has been developed. It can be accessed at http://103.99.176.239/predPhogly-Site. Users can submit one or more query protein sequence(s) directly on the web-server as text input in Fasta format or may prefer to upload as a batch to get their predictions. More detailed guidelines on how to use the web-server as well as the working mechanism of this server can also be found there. After submitting a query protein or as a batch, it may take a few moments to get the prediction result, depending on the availability of server resources. Finally, predPhogly-Site will generate a result page based on the user's submission, i.e., if protein sequences are submitted into the input box, the predictive data will be shown on the result page. Otherwise, it will be sent to the corresponding user through email.

Conclusion

In this study, for identifying phosphoglycerylation sites in protein with higher accuracy, a novel computational tool, predPhogly-Site, has been developed utilizing the coupling effects in a sequence. It exploits probabilistic sequence pattern information with variable cost adjustment in the classifier's decision function for achieving higher predictive performance compared to the existing phosphoglycerylation site predictors. It has achieved significant performance improvement not only in the 10-fold cross-validation, which has been used as the benchmarking technique in the existing predictors but also in an independent test. Moreover, it has also achieved almost identical performance in both 10-fold cross-validation and independent test, which clearly demonstrates its stability. In the 10-fold cross-validation test, it has achieved more than 0.99 in both AUC and MCC, and in case of the independent test, it has achieved nearly 0.97 in the corresponding measures. These experimental outcomes demonstrate that predPhogly-Site is highly promising compared to the existing state-of-the-art phosphoglycerylation site predictors. It is expected to become a high throughput computational tool for PTM researcher for fast exploration of lysine modifications. Even the experimental scientists would be benefited from this web-based tool without going through its mathematical and implementation details. For further performance improvement and usability of this prediction tool, multiple types of post-translational modification with heterogeneous data would be incorporated simultaneously along with prediction interpretation support.

Supporting information

S1 File. Benchmark dataset. The phosphoglycerylated proteins as well as the segmented sequences with respective protein ID and positions have been provided.
(PDF)

S2 File. Independent test dataset. Proteins which have been recently added to the PLMD database and completely unknown to the proposed system.
(PDF)

S3 File. All possible combinations of the conditional probability values derived from the positive and negative subset.
(XLSX)

S4 File. The non-conditional probability values of 21 amino acids derived from the positive and negative subset.
(XLSX)

Author Contributions

Conceptualization: Sabit Ahmed, Afrida Rahman.

Data curation: Sabit Ahmed.

Formal analysis: Sabit Ahmed, Md Khaled Ben Islam, Julia Rahman.

Investigation: Afrida Rahman, Md. Al Mehedi Hasan, Md Khaled Ben Islam, Julia Rahman.

Methodology: Sabit Ahmed, Afrida Rahman.

Resources: Md. Al Mehedi Hasan, Shamim Ahmad.

Software: Afrida Rahman, Shamim Ahmad.

Supervision: Md. Al Mehedi Hasan, Shamim Ahmad.

Validation: Md. Al Mehedi Hasan, Md Khaled Ben Islam, Julia Rahman, Shamim Ahmad.

Visualization: Sabit Ahmed.

Writing – original draft: Sabit Ahmed, Afrida Rahman.

Writing – review & editing: Md Khaled Ben Islam, Julia Rahman.

References

1. Saraswathy N, Ramalingam P. Concepts and techniques in genomics and proteomics. Elsevier; 2011.
2. McDowell G, Philpott A. New insights into the role of ubiquitylation of proteins. In: International review of cell and molecular biology. vol. 325. Elsevier; 2016. p. 35–88.
3. Qiu WR, Sun BQ, Xiao X, Xu ZC, Chou KC. iPTM-mLys: identifying multiple lysine PTM sites and their different types. Bioinformatics. 2016; 32(20):3116–3123. https://doi.org/10.1093/bioinformatics/btw380 PMID: 27334473
4. Freiman RN, Tjian R. Regulating the regulators: lysine modifications make their mark. Cell. 2003; 112(1):11–17. https://doi.org/10.1016/S0092-8674(02)01278-3 PMID: 12526789
5. Reddy HM, Sharma A, Dehzangi A, Shigemizu D, Chandra AA, Tsunoda T. GlyStruct: glycation prediction using structural properties of amino acid residues. BMC bioinformatics. 2019; 19(13):55–64. https://doi.org/10.1186/s12859-018-2547-x PMID: 30717650
6. Jia J, Liu Z, Xiao X, Liu B, Chou KC. iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Analytical biochemistry. 2016; 497:48–56. https://doi.org/10.1016/j.ab.2015.12.009 PMID: 26723495
7. Xu Y, Chou KC. Recent progress in predicting posttranslational modification sites in proteins. Current topics in medicinal chemistry. 2016; 16(6):591–603. https://doi.org/10.2174/1568026615666150819110421 PMID: 26286211
8. Ju Z, Cao JZ, Gu H. Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou's general PseAAC. Journal of Theoretical Biology. 2016; 397:145–150. https://doi.org/10.1016/j.jtbi.2016.02.020 PMID: 26908349
9. Xu Y, Ding YX, Ding J, Wu LY, Deng NY. Phogly-PseAAC: prediction of lysine phosphoglycerylation in proteins incorporating with position-specific propensity. Journal of Theoretical Biology. 2015; 379:10–15. https://doi.org/10.1016/j.jtbi.2015.04.016 PMID: 25913879
10. Moellering RE, Cravatt BF. Functional lysine modification by an intrinsically reactive primary glycolytic metabolite. Science. 2013; 341(6145):549–553. https://doi.org/10.1126/science.1238327 PMID: 23908237
11. Chandra A, Sharma A, Dehzangi A, Shigemizu D, Tsunoda T. Bigram-PGK: phosphoglycerylation prediction using the technique of bigram probabilities of position specific scoring matrix. BMC molecular and cell biology. 2019; 20(2):1–9. https://doi.org/10.1186/s12860-019-0240-1 PMID: 31856704
12. Liu LM, Xu Y, Chou KC. iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. Medicinal Chemistry. 2017; 13(6):552–559. https://doi.org/10.2174/1573406413666170515120507 PMID: 28521678

13. Chou KC. Prediction of signal peptides using scaled window. peptides. 2001; 22(12):1973–1979. https://doi.org/10.1016/S0196-9781(01)00540-X PMID: 11786179

14. Hasan MAM, Ahmad S. mLysPTMpred: Multiple Lysine PTM Site Prediction Using Combination of SVM with Resolving Data Imbalance Issue. Natural Science. 2018; 10(9):370–384. https://doi.org/10.4236/ns.2018.109035

15. Chou KC. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. Journal of Biological Chemistry. 1993; 268(23):16938–16948. https://doi.org/10.1016/S0021-9258(19)85285-7 PMID: 8349584

16. Chou KC. Prediction of human immunodeficiency virus protease cleavage sites in proteins. Analytical biochemistry. 1996; 233(1):1–14. https://doi.org/10.1006/abio.2000.4757 PMID: 8789141

17. Veropoulos K, Campbell C, Cristianini N, et al. Controlling the sensitivity of support vector machines. In: Proceedings of the international joint conference on AI. vol. 55; 1999. p. 60.

18. Lin WZ, Fang JA, Xiao X, Chou KC. iDNA-Prot: identification of DNA binding proteins using random forest with grey model. PloS one. 2011; 6(9). https://doi.org/10.1371/journal.pone.0024756 PMID: 21935457

19. Hasan MAM, Ahmad S, Molla MKI. iMulti-HumPhos: a multi-label classifier for identifying human phosphorylated proteins using multiple kernel learning based support vector machines. Molecular BioSystems. 2017; 13(8):1608–1618. https://doi.org/10.1039/C7MB00180K PMID: 28682387

20. Ju Z, Wang SY. Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou's general pseudo amino acid composition. Gene. 2018; 664:78–83. https://doi.org/10.1016/j.gene.2018.04.055 PMID: 29694908

21. Ju Z, He JJ. Prediction of lysine propionylation sites using biased SVM and incorporating four different sequence features into Chou's PseAAC. Journal of Molecular Graphics and Modelling. 2017; 76:356–363. https://doi.org/10.1016/j.jmgm.2017.07.022 PMID: 28763688

22. Hasan MAM, Li J, Ahmad S, Molla MKI. predCar-site: Carbonylation sites prediction in proteins using support vector machine with resolving data imbalanced issue. Analytical biochemistry. 2017; 525:107–113. https://doi.org/10.1016/j.ab.2017.03.008 PMID: 28286168

23. Bao W, Yang B, Huang DS, Wang D, Liu Q, Chen YH, et al. IMKPse: Identification of protein malonylation sites by the key features into general PseAAC. IEEE Access. 2019; 7:54073–54083. https://doi.org/10.1109/ACCESS.2019.2900275

24. Hasan MA, Ben Islam MK, Rahman J, Ahmad S. Citrullination Site Prediction by Incorporating Sequence Coupled Effects into PseAAC and Resolving Data Imbalance Issue. Current Bioinformatics. 2020; 15(3):235–245. https://doi.org/10.2174/1574893614666191202152328

25. Qiu WR, Xiao X, Lin WZ, Chou KC. iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. BioMed research international. 2014; 2014. https://doi.org/10.1155/2014/947416

26. Liu Z, Wang Y, Gao T, Pan Z, Cheng H, Yang Q, et al. CPLM: a database of protein lysine modifications. Nucleic acids research. 2014; 42(D1):D531–D536. https://doi.org/10.1093/nar/gkt1093 PMID: 24214993

27. Consortium U. UniProt: a worldwide hub of protein knowledge. Nucleic acids research. 2019; 47(D1): D506–D515. https://doi.org/10.1093/nar/gky1049

28. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22(13):1658–1659. https://doi.org/10.1093/bioinformatics/btl158 PMID: 16731699

29. Ju Z, Wang SY. Prediction of lysine formylation sites using the composition of k-spaced amino acid pairs via Chou's 5-steps rule and general pseudo components. Genomics. 2020; 112(1):859–866. https://doi.org/10.1016/j.ygeno.2019.05.027 PMID: 31175975

30. Ning Q, Ma Z, Zhao X. dForml (KNN)-PseAAC: Detecting formylation sites from protein sequences using K-nearest neighbor algorithm via Chou's 5-step rule and pseudo components. Journal of theoretical biology. 2019; 470:43–49. https://doi.org/10.1016/j.jtbi.2019.03.011 PMID: 30880183

31. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome research. 2004; 14(6):1188–1190. https://doi.org/10.1101/gr.849004 PMID: 15173120

32. Xu H, Zhou J, Lin S, Deng W, Zhang Y, Xue Y. PLMD: An updated data resource of protein lysine modifications. Journal of Genetics and Genomics. 2017; 44(5):243–250. https://doi.org/10.1016/j.jgg.2017.03.007 PMID: 28529077

33. Du P, Wang X, Xu C, Gao Y. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. Analytical biochemistry. 2012; 425(2):117–119. https://doi.org/10.1016/j.ab.2012.03.015 PMID: 22459120

34. Qiu WR, Sun BQ, Xiao X, Xu ZC, Chou KC. iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. Oncotarget. 2016; 7 (28):44310. https://doi.org/10.18632/oncotarget.10027 PMID: 27322424

35. Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of theoretical biology. 2011; 273(1):236–247. https://doi.org/10.1016/j.jtbi.2010.12.024 PMID: 21168420

36. Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics. 2005; 21(1):10–19. https://doi.org/10.1093/bioinformatics/bth466 PMID: 15308540

37. Ju Z, He JJ. Prediction of lysine crotonylation sites by incorporating the composition of k-spaced amino acid pairs into Chou's general PseAAC. Journal of Molecular Graphics and Modelling. 2017; 77:200–204. https://doi.org/10.1016/j.jmgm.2017.08.020 PMID: 28886434

38. Min JL, Xiao X, Chou KC. iEzy-Drug: A web server for identifying the interaction between enzymes and drugs in cellular networking. BioMed research international. 2013; 2013. https://doi.org/10.1155/2013/701317 PMID: 24371828

39. Xu Y, Wen X, Wen LS, Wu LY, Deng NY, Chou KC. iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PloS one. 2014; 9(8):e105018. https://doi.org/10.1371/journal.pone.0105018 PMID: 25121969

40. Reback J, McKinney W, jbrockmendel, den Bossche JV, Augspurger T, Cloud P, et al. pandas-dev/pandas: Pandas 1.2.0rc0; 2020. Available from: https://doi.org/10.5281/zenodo.4311557.

41. Wang D, Liu D, Yuchi J, He F, Jiang Y, Cai S, et al. MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. Nucleic Acids Research. 2020;. https://doi.org/10.1093/nar/gkaa275 PMID: 32324217

42. Lv Z, Zhang J, Ding H, Zou Q. RF-PseU: A Random Forest Predictor for RNA Pseudouridine Sites. Frontiers in Bioengineering and Biotechnology. 2020; 8. https://doi.org/10.3389/fbioe.2020.00134 PMID: 32175316

43. Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995; 20(3):273–297. https://doi.org/10.1023/A:1022627411411

44. Vapnik V. The nature of statistical learning theory. Springer science & business media; 2013.

45. Ju Z, Wang SY. Prediction of lysine formylation sites using the composition of k-spaced amino acid pairs via Chou's 5-steps rule and general pseudo components. Genomics. 2020; 112(1):859–866. https://doi.org/10.1016/j.ygeno.2019.05.027 PMID: 31175975

46. Zhang L, Tan B, Liu T, Sun X. Classification study for the imbalanced data based on Biased-SVM and the modified over-sampling algorithm. In: Journal of Physics: Conference Series. vol. 1237. IOP Publishing; 2019. p. 022052.

47. Ju Z, He JJ. Prediction of lysine glutarylation sites by maximum relevance minimum redundancy feature selection. Analytical biochemistry. 2018; 550:1–7. https://doi.org/10.1016/j.ab.2018.04.005 PMID: 29641975

48. Al-Barakati HJ, Saigo H, Newman RH, et al. RF-GlutarySite: a random forest based predictor for glutarylation sites. Molecular omics. 2019; 15(3):189–204. https://doi.org/10.1039/C9MO00028C PMID: 31025681

49. Wu M, Yang Y, Wang H, Xu Y. A deep learning method to more accurately recall known lysine acetylation sites. BMC bioinformatics. 2019; 20(1):49. https://doi.org/10.1186/s12859-019-2632-9 PMID: 30674277

50. Jia C, Zhang M, Fan C, Li F, Song J. Formator: predicting lysine formylation sites based on the most distant undersampling and safe-level synthetic minority oversampling. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2019;. https://doi.org/10.1109/TCBB.2019.2957758 PMID: 31804942

51. Yu J, Shi S, Zhang F, Chen G, Cao M. PredGly: predicting lysine glycation sites for Homo sapiens based on XGboost feature optimization. Bioinformatics. 2019; 35(16):2749–2756. https://doi.org/10.1093/bioinformatics/bty1043 PMID: 30590442

52. Qu K, Han K, Wu S, Wang G, Wei L. Identification of DNA-binding proteins using mixed feature representation methods. Molecules. 2017; 22(10):1602. https://doi.org/10.3390/molecules22101602 PMID: 28937647

53. Malebary SJ, Rehman MSu, Khan YD. iCrotoK-PseAAC: Identify lysine crotonylation sites by blending position relative statistical features according to the Chou's 5-step rule. PloS one. 2019; 14(11): e0223993. https://doi.org/10.1371/journal.pone.0223993 PMID: 31751380

54. Li F, Li C, Marquez-Lago TT, Leier A, Akutsu T, Purcell AW, et al. Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. Bioinformatics. 2018; 34(24):4223–4231. https://doi.org/10.1093/bioinformatics/bty522 PMID: 29947803

55. Adilina S, Farid DM, Shatabda S. Effective DNA binding protein prediction by using key features via Chou's general PseAAC. Journal of theoretical biology. 2019; 460:64–78. https://doi.org/10.1016/j.jtbi.2018.10.027 PMID: 30316822

56. Thapa N, Chaudhari M, McManus S, Roy K, Newman RH, Saigo H, et al. DeepSuccinylSite: a deep learning based approach for protein succinylation site prediction. BMC bioinformatics. 2020; 21:1–10. https://doi.org/10.1186/s12859-020-3342-z PMID: 32321437

57. Liu K, Cao L, Du P, Chen W. im6A-TS-CNN: identifying N6-methyladenine site in multiple tissues by using convolutional neural network. Molecular Therapy-Nucleic Acids. 2020;. https://doi.org/10.1016/j.omtn.2020.07.034 PMID: 32858457

Accurately Predicting Nitrosylated Tyrosine Sites Using Probabilistic Sequence Information

Article *in* Gene · March 2022

DOI: 10.1016/j.gene.2022.146445

5 authors, including:

Afrida Rahman
Rajshahi University of Engineering & Technology
7 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Sabit Ahmed
Rajshahi University of Engineering & Technology
7 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Md. Al Mehedi Hasan
Rajshahi University of Engineering & Technology
117 PUBLICATIONS **713** CITATIONS
SEE PROFILE

Shamim Ahmad
University of Rajshahi
49 PUBLICATIONS **437** CITATIONS
SEE PROFILE

Gene 826 (2022) 146445

Contents lists available at ScienceDirect

Gene

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

Accurately predicting nitrosylated tyrosine sites using probabilistic sequence information

Afrida Rahman [a,1], Sabit Ahmed [a,1], Md. Al Mehedi Hasan [a], Shamim Ahmad [b], Iman Dehzangi [c,d,*]

[a] Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh
[b] Department of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh
[c] Department of Computer Science, Rutgers University, Camden, NJ 08102, USA
[d] Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA

ARTICLE INFO

Edited by John Doe

Keywords:
Post-translational modification
Nitrotyrosine Sites Prediction
Sequence-coupling Model
General PseAAC
Data Imbalance Issue
Support Vector Machine

ABSTRACT

Post-translational modification (PTM) is defined as the enzymatic changes of proteins after the translation process in protein biosynthesis. Nitrotyrosine, which is one of the most important modifications of proteins, is interceded by the active nitrogen molecule. It is known to be associated with different diseases including autoimmune diseases characterized by chronic inflammation and cell damage. Currently, nitrotyrosine sites are identified using experimental approaches which are laborious and costly. In this study, we propose a new machine learning method called PredNitro to accurately predict nitrotyrosine sites. To build PredNitro, we use sequence coupling information from the neighboring amino acids of tyrosine residues along with a support vector machine as our classification technique.Our results demonstrates that PredNitro achieves 98.0% accuracy with more than 0.96 MCC and 0.99 AUC in both 5-fold cross-validation and jackknife cross-validation tests which are significantly better than those reported in previous studies. PredNitro is publicly available as an online predictor at: http://103.99.176.239/PredNitro.

1. Introduction

Post-translational modifications (PTMs) introduce a series of crucial protein modifications after the translation phase (Saraswathy and Ramalingam, 2011). PTMs alter and derivate intra-molecular bonds of amino acids with drastic impacts on proteomic analysis and biological processes, such as cellular signal transduction, metabolism, subcellular localization, protein folding, and protein degradation (McDowell and Philpott, 2016; Weissman et al., 2003; Ghauri et al., 2018; Blantz and Munger, 2002). Hence, efficient identification and appropriate understanding of PTM sites are essential for basic research in the fields of disease detection and prevention, and drug development (Chou, 2015; Qiu et al., 2016). Among 20 fundamental amino acid residues that build proteins, modifications at tyrosine residue (Y) are usually referred to as tyrosine PTM or Y-PTM. There are several tyrosine PTMs such as amidation, phosphorylation, nitration, hydroxylation, sulfation, and ubiquitination (Lee et al., 2006).

Among several forms of PTMs, protein nitrotyrosine is of critical importance. It is generated by the interaction of tyrosine along with nitrate molecules in peroxynitrite (ONOO-), which is reactive and often derived from an aggregation of superoxide radical anion (O2-) and nitric oxide (NO) (Abello et al., 2010) shown in Fig. 1. Nitrotyrosine is regarded as an indicator of inflammation and cell injury. It is also shown to be involved in diseases such as septic shock, Alzheimer, lung cancer, rheumatoid arthritis, celiac disease, cardiovascular disease and asthma (Giasson et al., 2000; Donnini et al., 2008; Brindicci et al., 2010).The experimental method for precisely identifying nitrotyrosine sites is expensive and time-consuming. It is even more sensitive toward proteins that are plentiful. (Hasan et al., 2018). Therefore, there is a demand to develop fast computational approaches to accurately predict nitrotyrosine sites. (Qiu et al., 2017; Rahman et al., 2020).

During the past few years, a wide range of computational methods have been proposed to predict nitrotyrosine sites. The first predictor of nitrotyrosine sites named 'GPS-YNO2' was proposed by Liu et al. (2011). It was developed using four statistical analyses including matrix mutation, weight training, k-means clustering, and motif length selection.

* Corresponding author at: Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh.
E-mail address: i.dehzangi@rutgers.edu (I. Dehzangi).
[1] Contributed equally to this work.

https://doi.org/10.1016/j.gene.2022.146445
Received 8 September 2021; Received in revised form 16 February 2022; Accepted 18 March 2022
Available online 28 March 2022

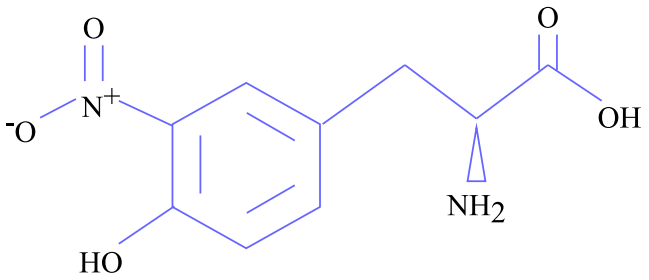

Fig. 1. Chemical structure of 3-Nitrotyrosine.

Table 1
Summary Of dataset.

Dataset	Positive sites	Negative sites
Training dataset	1191	1191
Independent dataset	203	1022

Later on, Xu et al. (2014) constructed 'iNitro-Tyr' using the composition of pseudo amino acid encoding. After that, Ghauri et al. developed 'pNitro-Tyr-PseAAC' (Ghauri et al., 2018) using a backpropagation neural network. In a different study, Xie et al. (2018) constructed 'DeepNitro' using deep learning with four encoding schemes (i.e., positional amino acid distributions, sequence contextual dependencies, physicochemical properties, and position-specific scoring features) to predict nitrotyrosine sites. At the same time, Hasan et al. (2018) developed 'NTyroSite' in which the Wilcoxon-rank sum test was applied as a feature selection technique and Random Forest was applied as a classifier. Most recently, a new machine learning method named 'PredNTS' (Nilamyani et al., 2021) was developed by integrating multiple feature encoding schemes such as K-mer, composition of k-spaced amino acid pairs (CKSAAP), AAindex, and binary as well as using Random Forest as a classifier. It achieved better performance compared to the other previous studies. PredNTS achieved 91% AUC in 5-fold cross-validation test. However, PredNTS does not achieve similar results in terms of sensitivity and MCC. It means that it is better in predicting negative nitrotyrosine samples rather than positive nitrotyrosine samples.

Recognizing the aforementioned scenario, there is a demand for developing novel system for identifying nitrotyrosine sites with higher efficacy. In order to develop an efficient predictor, appropriate sequence patterns associated with tyrosine nitration need to be retrieved. In this study, we propose a new machine learning method called PredNitro to accurately predict nitrotyrosine sites. To build this model, we utilize the vectorized sequence-coupled features to capture the useful information out of the protein sequences and a support vector machine (SVM) as our classification technique (Dehzangi et al., 2015; Vapnik, 2013; Ahmed et al., 2021; Ahmed et al., 2021; Rahman et al., 2020; Ahmed et al., 2021). PredNitro achieves more than 99% AUC in jackkinfe test, k-fold cross-validation test, and independent test. The exploratory results of other crucial metrics demonstrate the superior performance of PredNitro over the other existing approaches. PredNitro is publicly available as an online predictor at: http://103.99.176.239/PredNitro.

2. Materials and methods

2.1. Dataset

The nitrotyrosine dataset for this study was collected from multiple databases (dbPTM, SysPTM2.0, GPS-YNO2) including DeepNitro and iNitro-Tyr, as stated by Nilamyani et al. (2021). It contains 796 proteins with 1406 experimentally validated nitrotyrosine sites. Subsequently, redundant sequences were discarded using CD-HIT with a similarity cutoff of 40% to prevent overfitting issues in performance measurement, since this degree of redundancy reduction is widely acknowledged. For the independent test dataset, 20% of the samples were picked randomly to assure the feasibility of the proposed predictor for new and unseen proteins. Thereafter, a 1:1 ratio of positive to negative samples was chosen from the entire remaining dataset to construct the training set which is identical to the 'PredNTS' (Nilamyani et al., 2021). As a result, the non-redundant training dataset was attained composing 1191 experimentally positive nitrotyrosine samples and 1191 negative nitrotyrosine samples. On the other hand, the independent test dataset was attained composing 203 experimentally positive nitrotyrosine samples and 1022 negative nitrotyrosine samples (See Table 1). The training and independent test dataset with corresponding peptide sequences and site positions are hosted in a GitHub repository at https://github.com/Sabit-Ahmed/PredNitro. An overview of the dataset preparation as well as the general architecture of PredNitro is shown in Fig. 2. For analyzing the statistically verified disparity among the positive and negative nitrotyrosine samples in our dataset, the distribution of amino acid residues in the positive samples and negative samples are visually explored by the guidance of WebLogo in Fig. 3 and Fig. 4 (Crooks et al., 2004).

2.2. Feature construction

In this study, Chou's scheme (Chou, 1993; Ahmed et al., 2021) was implemented to encode more scrupulously and efficiently the sequences of the nitrotyrosine sites and extract features from the peptide segment. Based on the Chou's conception, a tyrosine residue centered peptide can be represented by:

$$\Theta_\zeta(Y) = R_{-\zeta}R_{-(\zeta-1)}\ldots R_{-2}R_{-1}YR_1R_2\ldots R_{+(\zeta-1)}R_{+\zeta} \quad (1)$$

In this equation, $R_{-\zeta}$ and $R_{+\zeta}$ denote the ζ-th leftward and rightward amino acid residues, respectively, while ζ being an integer and 'Y'(center) indicating "Tyrosine" (Ahmed et al., 2021).Again, the peptide sequence $\Theta_\zeta(Y)$ is categorized into two types: $\Theta_\zeta^+(Y), \Theta_\zeta^-(Y)$ are true nitrated peptide and false nitrated peptide with a tyrosine residue at its center (Rahman et al., 2020; Ahmed et al., 2021). To segment the nitrotyrosine protein sequences, the sliding window method was adopted. According (Ghauri et al., 2018) which introduced pNitro-Tyr-PseAAC, using ζ=20 as the window size obtained the best results. Therefore, we use the same window size in this study meaning that the corresponding peptide segment contained $(2\zeta+1) = 41$ amino acid residues. With a sequence fragment of window size 41, Eq. 1 can be presented as:

$$\Theta_{20}(Y) = Q_1Q_2\ldots.Q_{19}Q_{20}YQ_{21}Q_{22}\ldots.Q_{39}Q_{40} \quad (2)$$

In the segmentation process, the absent amino acids are filled with dummy residues denoted by 'X' as it was discussed in Xu et al. (2015), Ahmed et al. (2021) for the processing of site sequences of identical length. Therefore, the nitrotyrosine dataset is taken the following pattern:

$$S_\zeta(\mathrm{Y}) = S_\zeta^+(\mathrm{Y}) \cup S_\zeta^-(\mathrm{Y}) \quad (3)$$

where the positive subset $S_\zeta^+(\mathrm{Y})$ could contain only $\Theta_\zeta^+(Y)$ samples, while the negative subset $S_\zeta^-(\mathrm{Y})$ could contain only $\Theta_\zeta^-(Y)$ samples with their center residue Y. In this study, the vectorized sequence-coupled model has been adopted to extract features from the tyrosine nitrated sites eliciting the sequence pattern details (Chou, 1993; Ahmed et al., 2021). According to Chou's general PseAAC (Chou, 2011), the peptide sample in (2), can be presented as:

$$\Theta(Y) = \Theta^+(Y) - \Theta^-(Y) \quad (4)$$

where

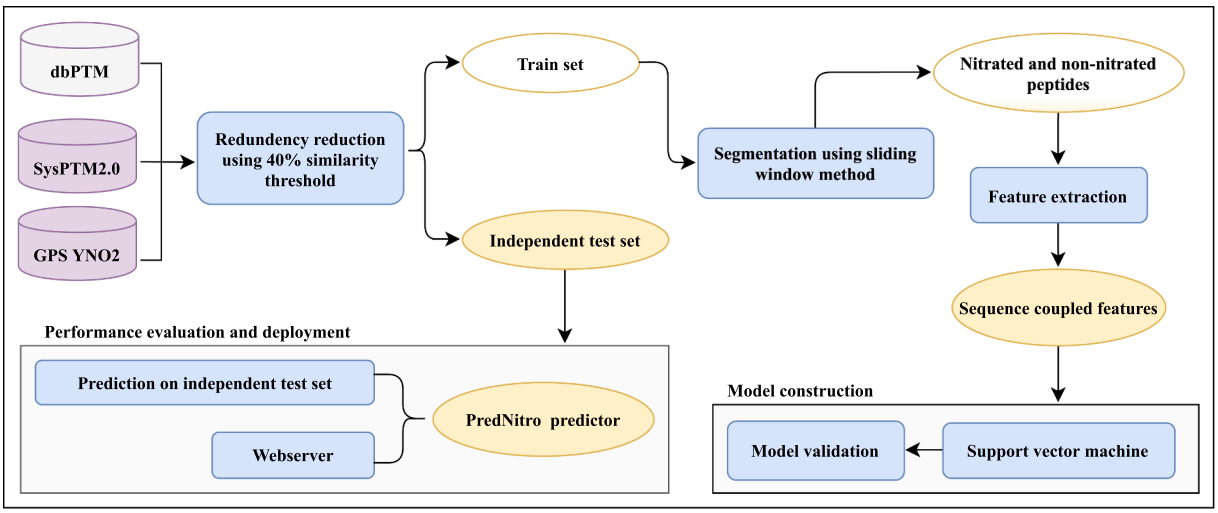

Fig. 2. The general architecture of PredNitro. The nitrotyrosine dataset has been collected from three different databases and splitted it into train set and independent test set. Finally, a new machine learning tool has been developed utilizing the sequence-coupled information and support vector machine called PredNitro to predict nitrotyrosine sites in proteins.

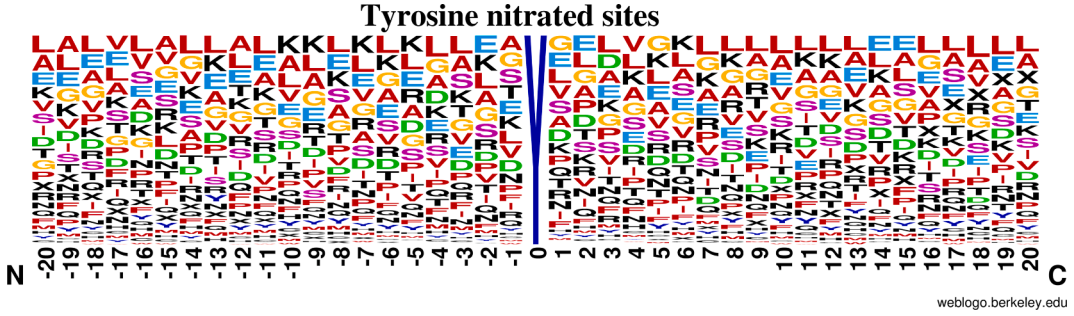

Fig. 3. Frequency plot of positive nitrotyrosine samples.

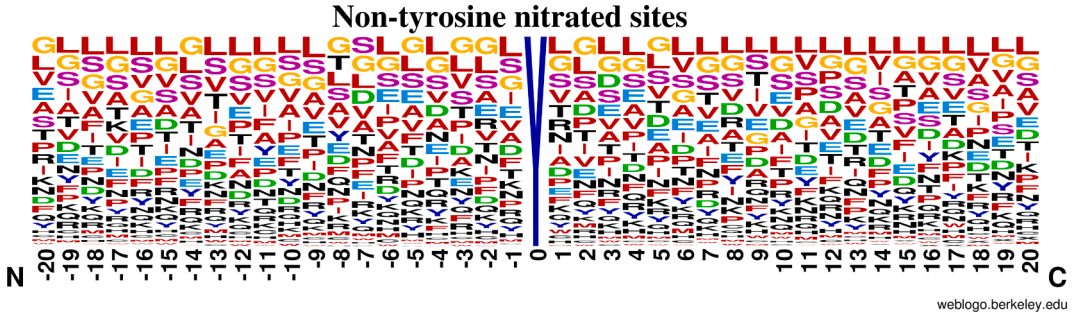

Fig. 4. Frequency plot of negative nitrotyrosine samples.

Table 2
Performances for different classification techniques on the benchmark dataset.

Predictor	Cross-validation Technique	Sn (%)	Sp (%)	ACC (%)	MCC (%)	AUC (%)
KNN	Jackknife	98.07	95.55	96.81	93.65	98.92
	10-Fold	**98.10**	95.56	96.83	93.69	98.96
	5-Fold	97.95	95.53	96.74	93.51	99.00
RF	Jackknife	95.97	97.15	96.56	93.12	98.81
	10-Fold	96.13	97.04	96.58	93.17	98.80
	5-Fold	95.8	96.84	96.32	92.65	98.83
SVM	Jackknife	97.73	98.32	98.03	96.06	**99.48**
	10-Fold	97.74	**98.40**	**98.07**	**96.15**	**99.48**
	5-Fold	97.67	98.34	98.00	96.01	**99.48**

Table 3
Jackknife cross-validation performance comparison on different datasets.

Predictor	Threshold	Sn (%)	Sp (%)	ACC (%)	MCC (%)	AUC (%)
GPS-YNO2*	High	28.89	90.02	82.57	18.84	-
	Medium	40.53	85.02	79.60	21.71	-
	Low	50.09	80.18	76.51	23.35	-
iNitro-Tyr*	-	81.76	85.89	84.52	49.05	-
pNitro-Tyr-PseAAC*	-	85.64	89.40	88.09	62.70	-
Proposed Method[a]	-	93.05	93.86	93.75	77.07	98.20
Proposed Method[b]	-	97.73	98.32	98.03	96.06	99.48

[a] corresponds to the performance derived from the benchmark dataset used in Ghauri et al. (2018).
[b] corresponds to the performance on the benchmark dataset used in Nilamyani et al. (2021).
* corresponds to the performance reported by Ghauri et al. (2018).

Table 4
K-fold cross-validation performance comparison on different datasets.

Predictor	Threshold	Sn(%)	Sp(%)	ACC (%)	MCC (%)	AUC (%)
pNitro-Tyr-PseAAC	-	84.00	93.02	89.10	-	-
Proposed Method[a]	-	93.26 ± 0.12	93.77 ± 0.06	93.70 ± 0.06	77.00 ± 0.20	98.20 ± 0.00
DeepNitro	High	17.70	95.00	84.90	17.20	-
	Medium	29.10	90.00	82.00	19.50	-
	Low	38.50	85.00	78.90	20.60	-
Proposed Method[b]	-	93.46 ± 0.10	93.86 ± 0.02	93.81 ± 0.02	77.52 ± 0.07	98.29 ± 0.01
PredNTS	-	-	-	-	-	91.00
Proposed Method[c]	-	97.67 ± 0.11	98.34 ± 0.11	98.00 ± 0.06	96.01 ± 0.13	99.48 ± 0.00

[a,b] correspond to the 10-fold cross-validation performance derived from the respective dataset used in Ghauri et al. (2018),Xie et al. (2018).
[c] corresponds to the 5-fold cross-validation performance on the dataset used in Nilamyani et al. (2021).

Table 5
Independent test performance comparison with the existing predictors.

Predictor	Sn(%)	Sp(%)	ACC(%)	MCC(%)	AUC(%)
GPS-YNO2	33.40	80.10	72.40	12.20	-
DeepNitro	33.90	80.30	72.60	12.80	-
NTyroSite	44.00	79.30	74.40	19.60	-
PredNTS	52.20	80.90	76.10	28.60	86.00
PredNitro	**100.00**	**88.16**	**90.12**	**74.32**	**99.59**

$$\Theta^{+}(Y)=\begin{bmatrix}\Theta_1^{+}(Q_1|Q_2)\\ \Theta_2^{+}(Q_2|Q_3)\\ \vdots\\ \Theta_{19}^{+}(Q_{19}|Q_{20})\\ \Theta_{20}^{+}(Q_{20})\\ \Theta_{21}^{+}(Q_{21})\\ \Theta_{22}^{+}(Q_{22}|Q_{21})\\ \vdots\\ \Theta_{39}^{+}(Q_{39}|Q_{38})\\ \Theta_{40}^{+}(Q_{40}|Q_{39})\end{bmatrix} \quad (5)$$

$$\Theta^{-}(Y)=\begin{bmatrix}\Theta_1^{-}(Q_1|Q_2)\\ \Theta_2^{-}(Q_2|Q_3)\\ \vdots\\ \Theta_{19}^{-}(Q_{19}|Q_{20})\\ \Theta_{20}^{-}(Q_{20})\\ \Theta_{21}^{-}(Q_{21})\\ \Theta_{22}^{-}(Q_{22}|Q_{21})\\ \vdots\\ \Theta_{39}^{-}(Q_{39}|Q_{38})\\ \Theta_{40}^{-}(Q_{40}|Q_{39})\end{bmatrix} \quad (6)$$

where $\Theta_1^{+}(Q_1|Q_2)$ is the conditional probability of amino acid Q_1 at the leftmost position given that its adjacent right member is Q_2 and so forth (Ahmed et al., 2021). Similarly, $\Theta_{40}^{+}(Q_{40}|Q_{39})$ denotes the conditional probability of amino acid Q_{40} at the rightmost position given that its adjacent left member is Q_{39} and so on. In contrast, only $\Theta_{20}^{+}(Q_{20})$ and $\Theta_{21}^{+}(Q_{21})$ are of non-conditional probability as Y is the adjoining member of both amino acids at position Q_{20} and Q_{21} (Chou, 1993; Ahmed et al., 2021; Rahman et al., 2020; Ahmed et al., 2021; Ahmed et al., 2021). The probability values can be extracted from the set of nitrated peptides using the frequency of a given acid corresponding to their positions. Accordingly, $\Theta^{-}(Y)$ in (4), and its probability components can be deduced from the non-nitrated peptide set in the same way as shown in (6). Finally, a 40-dimensional feature vector was obtained by using Eqs. ()()()(4)–(6) for each potential nitrated and non-nitrated sample.

In order to facilitate visualization and insight into the sequence-coupling effects at various places in each sample, we have stored all conceivable combinations of conditional probability values extracted from the positive training subset i.e. $\Theta^{+}(Q_1|Q_2)$ to $\Theta^{+}(Q_{19}|Q_{20})$ and $\Theta^{+}(Q_{22}|Q_{21})$ to $\Theta^{+}(Q_{40}|Q_{39})$ in one data frame and non-conditional probability values for each amino acid residue retrieved from the positive training subset i.e. $\Theta^{+}(Q_{20})$ and $\Theta^{+}(Q_{21})$ in another data frame, where the columns represent the formulated sample positions and the rows represent the amino acid residues. It should be noted that there are $21 \times 21 = 441$ different combinations of conditional probability values and 21 non-conditional probability values for each position in any formulated sample (including the dummy amino acid residue '*X*') (Chou,

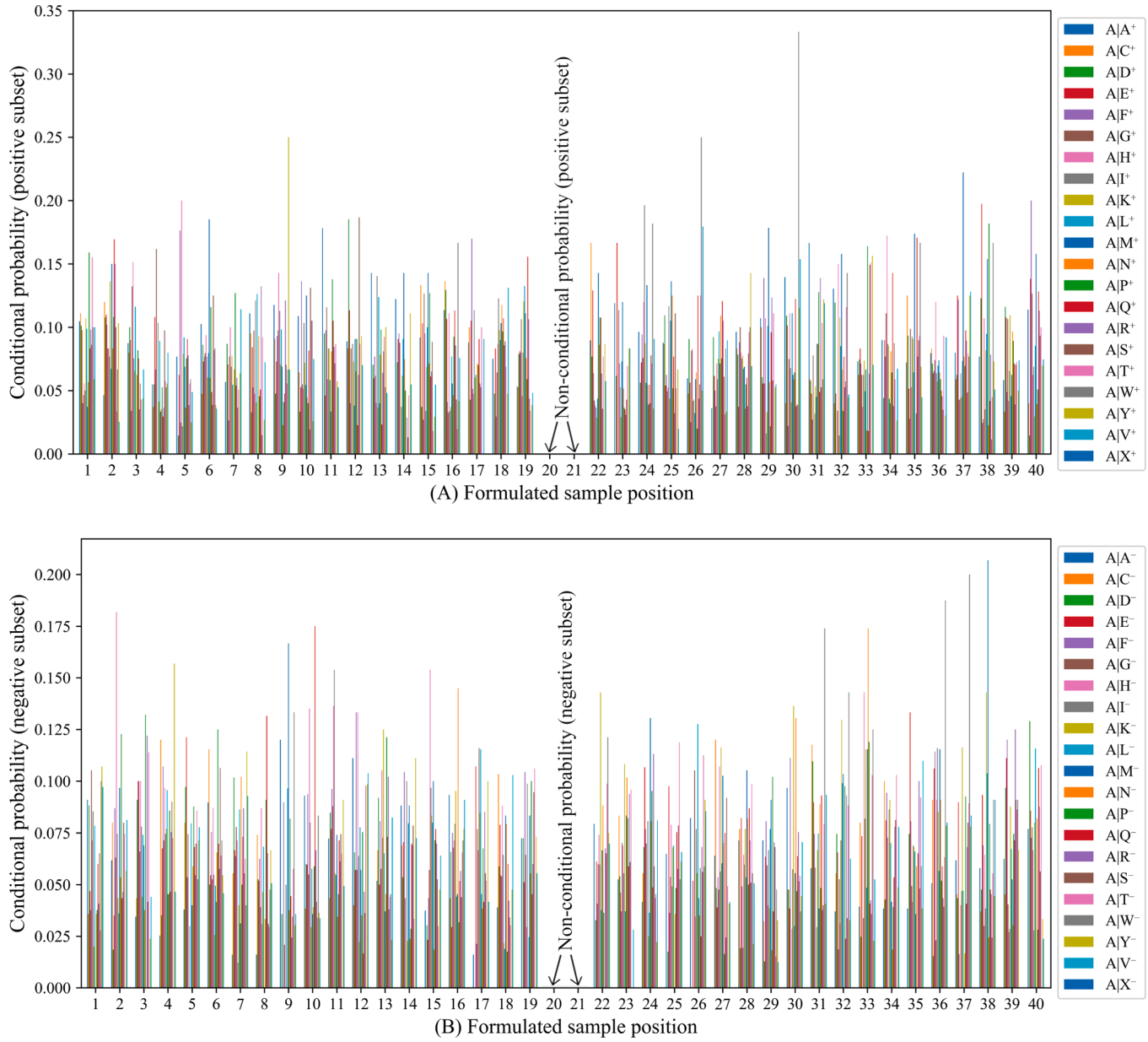

Fig. 5. (A) The conditional probability values of amino acid residue '*A*' which have been calculated from the positive subset, given that its right member is any of the 21 amino acid residues at sample positions 1 to 19 and the conditional probability values of any of the 21 amino acid residue given that the left member is '*A*' at sample positions 22 to 40. (B) The conditional probability values of amino acid residue '*A*' which have been calculated from the negative subset, given that its right member is any of the 21 amino acid residues at sample positions 1 to 19 and the conditional probability values of any of the 21 amino acid residue given that the left member is '*A*' at sample positions 22 to 40.

1993). Likewise, the conditional and non-conditional probability values derived from the negative subset are stored in two distinct data frames. All the corresponding conditional and non-conditional probability values retrieved from the positive and negative subsets are provided in the supplementary material S1.

2.3. Classification algorithm

During the last decade, different types of machine learning models such as random forest (Lv et al., 2020; Shi et al., 2019), logistic regression (Dai et al., 2021), stacking method (Bin et al., 2020), k-nearest neighbor (Wang et al., 2020), and neural network (Xie et al., 2018) had been used to predict the nitration sites (Ghauri et al., 2018, 2011, 2018, 2014, 2018). However, support vector machine (SVM) which is considered as one of the state-of-the-at classification techniques have never been used for this task. SVM aim at enhancing the classification performance by finding the maximal marginal hyperplane (MMH) to separate different classes (Cortes and Vapnik, 1995; Zhang et al., 2019). SVM is widely used in the literature to predict other PTMs and obtained promising results (Ahmed et al., 2021; Ahmed et al., 2021; Ahmed et al., 2021; Rahman et al., 2020; Chandra et al., 2020; Singh et al., 2020; Chandra et al., 2019; Reddy et al., 2019; Chandra et al., 2019). An SVM is designed to solve the following constrain minimization problem:

$$min\tfrac{1}{w,\xi}\|w\|^2 + C^+\sum_{k=1}^{q}\xi_k + C^-\sum_{k=q+1}^{n}\xi_k \tag{7}$$

(Subject to: $Y_k(w.\varphi(X_k) + a) \geqslant 1 - \xi_k$ for all, $k = 1,2,..,n$)where the training set is denoted by $\{(X_k, Y_k), k = 1,2,\ldots,n\}$ and first q examples (i.

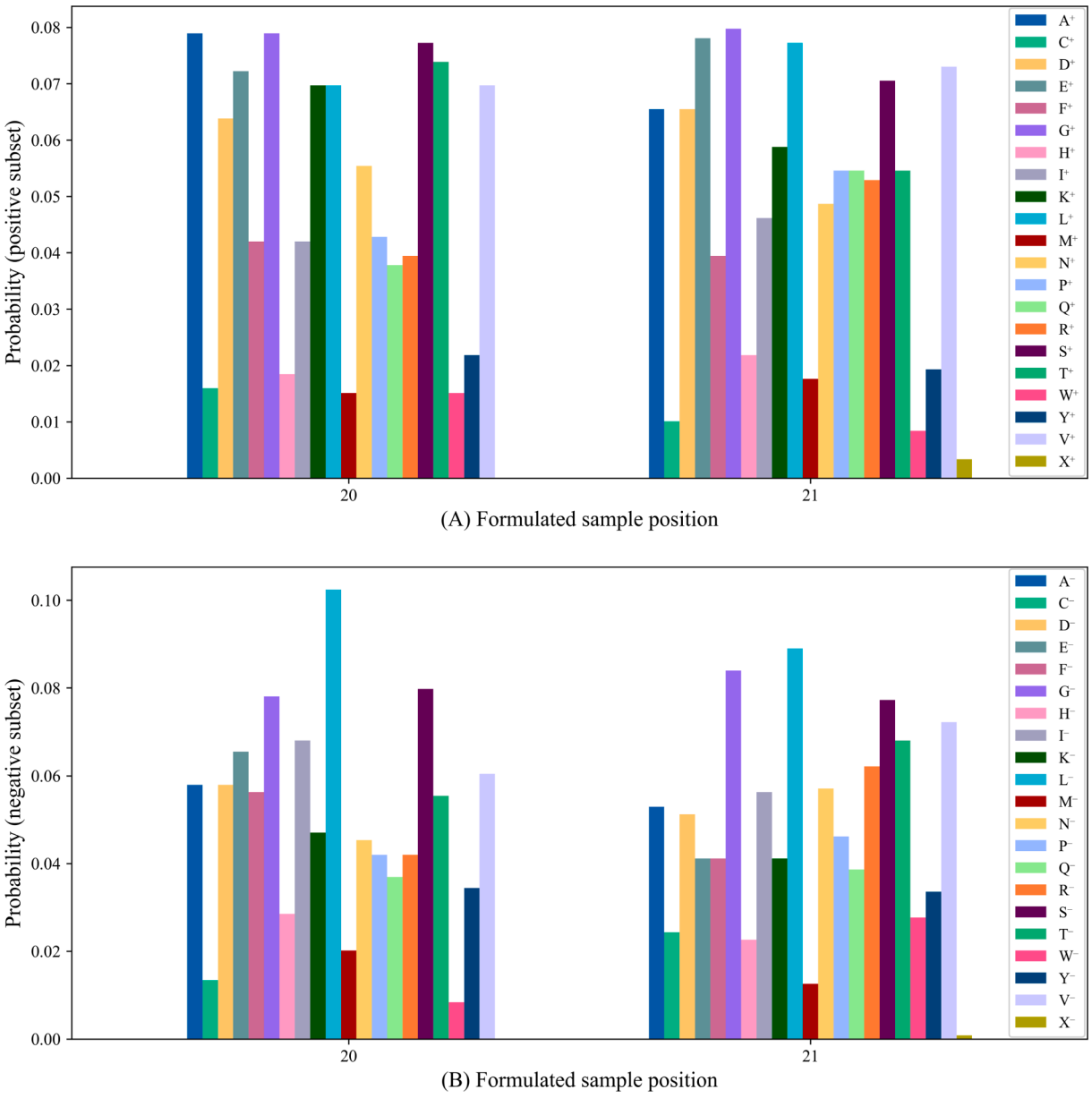

Fig. 6. (A) The non-conditional probability values of 21 amino acid residues derived from the positive subset at sample positions 20 and 21. (B) The non-conditional probability values of 21 amino acid residues derived from the negative subset at sample position 20 and 21.

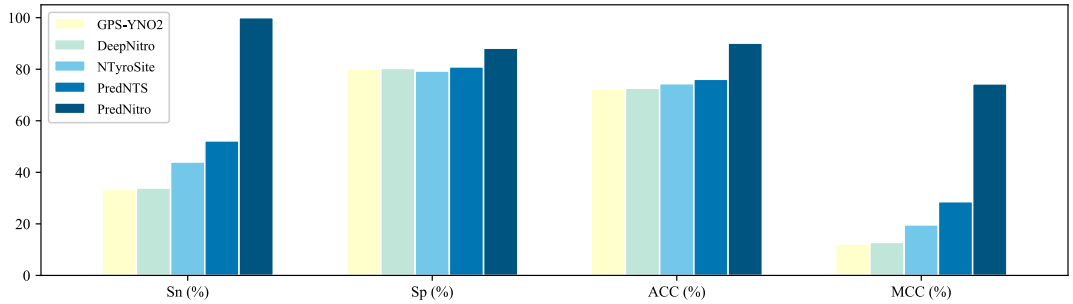

Fig. 7. Performance comparison between PredNitro and the existing methods based on the independent test.

e. $Y_k = 1, k = 1,2,\ldots,q$) are assumed as the positive examples while the rest are assumed as the negative examples (i.e. $Y_k = -1, k = q+1, q+2,\ldots,n$). The non-linear feature mapping and slack variables are denoted by $\varphi(X)$ and $\xi_k (k = 1, 2, \ldots, n)$ respectively (Ju and Wang, 2020). In our experiments, the Gaussian kernel function is described as: $\Upsilon(X_k, X_j) = \varphi(X_k)^T \varphi(X_j) = exp(-\frac{\|X_k - X_j\|^2}{\partial})$, where ∂ is the width of the function.

2.4. Prediction metrics

To investigate the prediction quality of PredNitro, we have utilized four intuitive evaluation metrics, such as accuracy (ACC), sensitivity (Sn), specificity (Sp), and Matthew's Correlation Coefficient (MCC) which have been widely used in the literature for this task (Dehzangi et al., 2018; Dehzangi et al., 2015; Ahmed et al., 2021; Rahman et al., 2020; Chandra et al., 2019). These performance metrics can be calculated as follows:

$$Sn = \frac{TP}{TP + FN} \tag{8}$$

$$Sp = \frac{TN}{TN + FP} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{12}$$

Where TP is true positive, FP is false positive, TN is true negative, and FN is false negative. Here, we also calculate the area under the ROC curve (AUC) and MCC to check the stability and robustness of our system (Ahmed et al., 2021; Rahman et al., 2020).

2.5. Jackknife cross-validation test

Jackknife cross-validation is widely used to evaluate the performance of different models to tackle this problem test (Liu et al., 2011; Ghauri et al., 2018). The jackknife test can provide a unique outcome for a given data set, and it is highly beneficial to reduce the computational sophistication of model creation (Li et al., 2016). Also, the arbitrariness problem and the 'memory' influence can be addressed when using this test (Chou, 2011; Ghauri et al., 2018). In this mode, if the total sample number of the cross-validation is *n*, each sample will be used as a test set once, while the remaining *n-1* samples will be used as training set (Li et al., 2016). In this way, the overall prediction metrics are calculated for each population mixture of size *n-1* using iterative process (Qiu et al., 2014; Liu et al., 2015; Khan et al., 2014). After the completion of N estimation turns, the averaged performance of test sets is calculated as the ultimate outcome for the jackknife cross-validation.

2.6. K-fold cross-validation test

It is often desirable to obtain an appropriate model that can be capable of delivering high efficiency. Although jackknife test is widely used in the literature for this purpose, now-a-days researchers prefer k-fold cross-validation approach over the jackknife test for verifying their PTM prediction model to reduce the computing time of the model creation (Ju et al., 2016; Ju and Wang, 2020). Therefore, here k-fold cross-validation test is also implemented to evaluate our proposed predictor. We performed this model *M* times and reported the average results to guarantee the stability of our reported results. The M-iterations of k-fold cross-validation were performed according to the following steps:

Step 1. Divide the benchmark dataset randomly into k disjoint sets.
Step 2. Select 1 set as test set and utilize the remaining k-1 sets as training set.
Step 3. Train the predictor with the training set using the LibSVM's default parameters ($C = 1, \gamma = 1/\text{numberoffeatures}$).
Step 4. Perform prediction on the test set.
Step 5. Repeat steps 2 to 4 until all k sets had been used for testing.
Step 6. Merge the prediction outputs and measure the performance with Eq. 7.
Step 7. Repeat steps 1 to 6 for M times.
Step 8. Measure the average performance of M repetitions with corresponding standard deviations.

Several tyrosine nitration prediction systems i.e. pNitro-Tyr-PseAAC (Ghauri et al., 2018), and DeepNitro (Xie et al., 2018) have validated their model using k-fold cross-validation approach where k = 10. The most recent predictor PredNTS (Nilamyani et al., 2021) has utilized 5-fold cross-validation technique for model evaluation. Therefore, we adopted both 5-fold and 10-fold cross-validation to be able to directly compare our results with those reported in the literature. The predictive decision-making workflow of PredNitro is publicly available at https://github.com/Sabit-Ahmed/PredNitro as a GitHub repository.

2.7. Independent test

Existing nitrotyrosine site predictors, particularly, the most recent system PredNTS (Nilamyani et al., 2021) assessed their model using 5-fold cross-validation. However, most researchers emphasize on the necessity of independent test for assessing prediction model in addition to k-fold (e.g. k = 5,10) cross-validation (Xie et al., 2018; Ghauri et al., 2018; Nilamyani et al., 2021). Thus, here an independent test consisting of 203 experimentally annotated positive nitrotyrosine sites and 1022 negative nitrotyrosine sites is generated for further evaluation of our proposed model. It should be mentioned that the proteins in the independent test set have not been used for any parameter tuning.

3. Results and Discussions

3.1. Performance of PredNitro

In this study, the sequence-coupling features have been extracted from the benchmark dataset and the support vector machine (Hasan et al., 2017; Hasan and Ahmad, 2018) has been used as prediction algorithm. We also use Radial Basis Function (RBF) as SVM kernel which demonstrates better results than using other kernels (Hasan and Ahmad, 2018; Hasan et al., 2017). To ascertain the statistical importance of the predicted results of PredNitro, three validation techniques, such as k-fold cross-validation (k = 5 and 10), jackknife test, and an independent test, are extensively used (Hasan and Ahmad, 2018; Ahmed et al., 2021). Therefore, our proposed model is evaluated with different types of validation schemes based on the corresponding datasets used in GPS-YNO2, pNitro-Tyr-PseAAC, DeepNitro, and PredNTS studies. Our results demonstrates that PredNitro can predict nitrotyrosine sites with more than 98.0% accuracy and 96.0% MCC on both jackknife test and k-fold cross-validation schemes. In addition, its sensitivity, specificity and AUC measure crossed a benchmark of 97.0% on both cross-validation methods. Furthermore, it achieves 90.12% accuracy with 99.59% AUC on the independent test. In other words, PredNitro consistently outperforms previous studies using all three evaluation methods which demonstrate the generality of this model. The ROC curves of PredNitro in the training and independent test sets are available in supplementary material S2.

3.2. Performance analysis of different classification techniques

Currently, a wide range of machine learning techniques are available, which can effectively classify any nitrotyrosine site containing sample (Ghauri et al., 2018; Hasan et al., 2018; Xu et al., 2014; Xie et al., 2018; Nilamyani et al., 2021). As our benchmark dataset is comparatively small, we have intended to experiment with a several less data demanding machine learning algorithms such as, random forest (Lv et al., 2020; Shi et al., 2019), k-nearest neighbor (Wang et al., 2020), and support vector machine (Ahmed et al., 2021; Ahmed et al., 2021; Ahmed

et al., 2021; Rahman et al., 2020). By utilizing the sequence-coupled features, we have performed the jackknife test, 5-fold and 10-fold cross-validation tests on the benchmark dataset to find the best-suited model for constructing our proposed predictor. From Table 2, it can be observed that the random forest method has obtained the lowest performances in all the cross-validation tests. On the contrary, the k-nearest neighbour algorithm has achieved the highest sensitivity rate of 98.10% in the 10-fold cross-validation. Moreover, its attained sensitivity rates in other validation tests are higher than any other classification method. However, the support vector machine has obtained the highest specificity, accuracy, MCC and AUC measures in all three types of cross-validation tests. Analyzing the outcomes of the three most commonly used classifiers, we have constructed our final prediction system with the SVM classifier because of its high-performance statistics.

3.3. Comparative analysis with existing predictors

At present, there are six main predictors available to predict the nitrotyrosine sites, such as, GPS-YNO2 (Liu et al., 2011), iNitro-Tyr (Xu et al., 2014), pNitro-Tyr-PseAAAC (Ghauri et al., 2018), DeepNitro (Xie et al., 2018), NTyroSite (Hasan et al., 2018), and PredNTS (Nilamyani et al., 2021) for nitrotyrosine site prediction. Particularly, each of these prediction systems has constructed its curated dataset for performance benchmarking. GPS-YNO2, and iNitro-Tyr have used jackknife test while pNitro-Tyr-PseAAC has adopted both the jackknife and 10-fold cross-validation test for model evaluation. DeepNitro and NTyroSite have applied 10-fold cross-validation schemes. On the other hand, PredNTS has validated their model with 5-fold cross-validation technique. For a fair comparison, we have utilized different datasets and validation criteria. The prediction outcome from the jackknife test, k-fold cross-validation, and independent test has been measured with the evaluation metrics described in Eqs. (8)–(12) and reported in Tables 3–5, respectively while corresponding standard deviations where applicable.

As shown in Table 3, the PredNitro achieves significantly better results compared to GPS-YNO2, iNitro-Tyr, pNitro-Tyr-PseAAC in terms of all metrics (i.e accuracy, MCC, sensitivity, specificity) using jackknife cross validation test. For instance, PredNitro outperformed the most recent predictor, pNitro-Tyr-PseAAC, by 7.41% in term of sensitivity, 4.46% in term of specificity, 5.66% in term of accuracy, and 14.37% in term of MCC on the same dataset that they used. AUC which is one of the most important measures has reached above 98.0%. It is important to note that higher sensitivity achieved by our predictor demonstrates that PredNitro is able to predict positive samples significantly better than previous studies.

To scrutinize the results, we further implemented 10-fold cross-validation test for 10 times on the dataset provided by pNitro-Tyr-PseAAC and DeepNitro studies, and again achieved significantly better results compared to these two methods. As shown in Table 4, results obtained by PredNitro are almost similar to the jackknife test performances. This demonstrates the generality of our model for this task. In addition, we have applied 5-fold cross-validation 5 times on the becnhmark dataset used in PredNTS study. When comparing our proposed method with the PredNTS predictor, it can be observed that, the AUC measure has increased from 91.0% to 99.48%. Furthermore, the reported jackknife test results (see Table 3 on the same benchmark dataset are also promising and identical to that of 5-fold cross-validation results. Our proposed method attains over 96.0% for all the prediction metrics.

We also achieve similar results for our independent test set. To conduct a fair comparison, the independent dataset was uploaded to the web-servers of the existing state-of-art predictors (i.e. GPS-YNO2, DeepNitro, NTyroSite and PredNTS) to obtain the prediction outcomes. The predictive performance of PredNitro as well as other predictors are summarized in Table 5 and Fig. 7. As shown in Table 5 and Fig. 7, PredNitro again achieves better results than any other existing predictors. More precisely, it enhance the sensitivity, specificity, accuracy, MCC, and AUC for 47.80%, 7.26%, 14.02% 45.75%, and 13.59% compared to PredNTS as the most recent successful predictor, respectively. Again, significant improvement in sensitivity demonstrates the ability of PredNitro in identifying tyrosine nitration sites. Results represented in Table 5 and Fig. 7 indicate that our proposed predictor PredNitro can be a high throughput tool for the effective identification of the unknown tyrosine nitration sites.

Our results demonstrates that by using effective representation of nitrotyrosine modification in terms of sequence coupling model among the amino acid residues via the conditional probability as well as SVM as our classification technique we are able to significantly enhance tyrosine nitration sites prediction task compared to previous studies (see Fig. 5 and 6).

3.4. Web-server

To increase user accessibility without requiring experimental solutions, we designed an easy-to-use web server for PredNitro which is publicly available at: http://103.99.176.239/PredNitro. Users can enter one or more query protein sequences as text input in Fasta format directly on the web server, or they can upload as a batch to acquire their predictions. There are also more thorough instructions on how to operate the web server as well as the server's operating mechanism. Depending on the availability of server resources, getting the prediction result after submitting a query protein or as a batch may take a few moments. Finally, PredNitro will generate a result page based on the user's input. For example, if protein sequences are entered into the input box, the predictive data will appear on the result page. Otherwise, an email will be sent to the appropriate user.

4. Conclusion

The prediction of nitrotyrosine sites is critically important for attaining a better perception of biological systems. In this study, we develop a novel machine learning tool named PredNitro to accurately predict tyrosine nitrated sites by using a vectorized sequence-coupling model with SVM classifier. By adopting sequence-coupling effect with misclassification cost adjustment, PredNitro acquired extraordinarily higher prediction accuracy compared to the existing nitrotyrosine site predictors. Both in the k-fold cross-validation and jackknife cross-validation test, it obtained a significant improvement in MCC as well as in other crucial metrics (approximately 98.0%, 97.0%, and 0.99 in terms of accuracy, sensitivity, and AUC) which ensures the generality and robustness of our predictor. PredNitro is publicly available as an online predictor at: http://103.99.176.239/PredNitro.

Contributors

A. Rahman, S. Ahmed designed and performed the experiments. A. Rahman, S. Ahmed, and A. Dehzangi wrote the manuscript and validated the results. A. Rahman prepared figures. M. A. M. Hasan and I. Dehzangi mentored and analytically reviewed the paper. S. Ahmad provided the resources for the web-server. All the authors reviewed the article.

Data availability statement

PredNitro as an online predictor and our employed benchmarks are publicly available online at: http://103.99.176.239/PredNitro.

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.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.gene.2022.146445.

References

Abello, N., Barroso, B., Kerstjens, H.A., Postma, D.S., Bischoff, R., 2010. Chemical labeling and enrichment of nitrotyrosine-containing peptides. Talanta 80 (4), 1503–1512.

Ahmed, S., Rahman, A., Hasan, M.A.M., Islam, M.K.B., Rahman, J., Ahmad, S., 2021. predPhogly-Site: Predicting phosphoglycerylation sites by incorporating probabilistic sequence-coupling information into PseAAC and addressing data imbalance. Plos One 16 (4), e0249396.

Ahmed, S., Rahman, A., Hasan, M.A.M., Rahman, J., Islam, M.K.B., Ahmad, S., 2021. predML-Site: Predicting Multiple Lysine PTM Sites with Optimal Feature Representation and Data Imbalance Minimization. IEEE/ACM Trans. Comput. Biol. Bioinform. (01), 1–1.

Ahmed, S., Rahman, A., Hasan, M., Mehedi, A., Ahmad, S., Shovan, S., 2021. Computational identification of multiple lysine PTM sites by analyzing the instance hardness and feature importance. Scient. Rep. 11 (1), 1–12.

Bin, Y., Zhang, W., Tang, W., Dai, R., Li, M., Zhu, Q., Xia, J., 2020. Prediction of neuropeptides from sequence information using ensemble classifier and hybrid features. J. Proteome Res. 19 (9), 3732–3740.

Blantz, R.C., Munger, K., 2002. Role of nitric oxide in inflammatory conditions. Nephron 90 (4), 373–378.

Brindicci, C., Kharitonov, S.A., Ito, M., Elliott, M.W., Hogg, J.C., Barnes, P.J., Ito, K., 2010. Nitric oxide synthase isoenzyme expression and activity in peripheral lung tissue of patients with chronic obstructive pulmonary disease. Am. J. Respirat. Crit. Care Med. 181 (1), 21–30.

Chandra, A., Sharma, A., Dehzangi, A., Ranganathan, S., Jokhan, A., Chou, K.-C., Tsunoda, T., 2018. Phoglystruct: prediction of phosphoglycerylated lysine residues using structural properties of amino acids. Scient. Rep. 8 (1), 1–11.

Chandra, A.A., Sharma, A., Dehzangi, A., Tsunoda, T., 2019. Evolstruct-phogly: incorporating structural properties and evolutionary information from profile bigrams for the phosphoglycerylation prediction. BMC Genom. 19 (9), 1–9.

Chandra, A., Sharma, A., Dehzangi, A., Shigemizu, D., Tsunoda, T., 2019. Bigram-pgk: phosphoglycerylation prediction using the technique of bigram probabilities of position specific scoring matrix. BMC Mol. Cell Biol. 20 (2), 1–9.

Chandra, A.A., Sharma, A., Dehzangi, A., Tsunoda, T., 2020. Ram-pgk: Prediction of lysine phosphoglycerylation based on residue adjacency matrix. Genes 11 (12), 1524.

Chou, K.-C., 1993. A vectorized sequence-coupling model for predicting hiv protease cleavage sites in proteins. J. Biol. Chem. 268 (23), 16938–16948.

Chou, K.-C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theoret. Biol. 273 (1), 236–247.

Chou, K.-C., 2015. Impacts of bioinformatics to medicinal chemistry. Med. Chem. 11 (3), 218–234.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

Crooks, G.E., Hon, G., Chandonia, J.-M., Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome Res. 14 (6), 1188–1190.

Dai, R., Zhang, W., Tang, W., Wynendaele, E., Zhu, Q., Bin, Y., De Spiegeleer, B., Xia, J., 2021. Bbppred: sequence-based prediction of blood-brain barrier peptides with feature representation learning and logistic regression. J. Chem. Inf. Model. 61 (1), 525–534.

Dehzangi, A., Heffernan, R., Sharma, A., Lyons, J., Paliwal, K., Sattar, A., 2015. Gram-positive and gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into chou's general pseaac. J. Theoret. Biol. 364, 284–294.

Dehzangi, A., Lopez, Y., Lal, S.P., Taherzadeh, G., Sattar, A., Tsunoda, T., Sharma, A., 2018. Improving succinylation prediction accuracy by incorporating the secondary structure via helix, strand and coil, and evolutionary information from profile bigrams. PloS one 13 (2).

Donnini, S., Monti, M., Roncone, R., Morbidelli, L., Rocchigiani, M., Oliviero, S., Casella, L., Giachetti, A., Schulz, R., Ziche, M., 2008. Peroxynitrite inactivates human-tissue inhibitor of metalloproteinase-4. FEBS Lett. 582 (7), 1135–1140.

Ghauri, A.W., Khan, Y.D., Rasool, N., Khan, S.A., Chou, K.-C., 2018. pnitro-tyr-pseaac: predict nitrotyrosine sites in proteins by incorporating five features into chou's general pseaac. Current pharmaceutical design 24 (34), 4034–4043.

Giasson, B.I., Duda, J.E., Murray, I.V., Chen, Q., Souza, J.M., Hurtig, H.I., Ischiropoulos, H., Trojanowski, J.Q., Lee, V.M.-Y., 2000. Oxidative damage linked to neurodegeneration by selective α-synuclein nitration in synucleinopathy lesions. Science 290 (5493), 985–989.

Hasan, M.A.M., Ahmad, S., 2018. mLysPTMpred: Multiple Lysine PTM Site Prediction Using Combination of SVM with Resolving Data Imbalance Issue. Natural Science 10 (9), 370–384.

Hasan, M.A.M., Ahmad, S., Molla, M.K.I., 2017. iMulti-HumPhos: a multi-label classifier for identifying human phosphorylated proteins using multiple kernel learning based support vector machines. Mol. BioSyst. 13 (8), 1608–1618.

Hasan, M.A.M., Li, J., Ahmad, S., Molla, M.K.I., 2017. predCar-site: Carbonylation sites prediction in proteins using support vector machine with resolving data imbalanced issue. Analytical biochemistry 525, 107–113.

Hasan, M., Khatun, M., Mollah, M., Haque, N., Yong, C., Dianjing, G., et al., 2018. Ntyrosite: Computational identification of protein nitrotyrosine sites using sequence evolutionary features. Molecules 23 (7), 1667.

Ju, Z., Wang, S.-Y., 2020. Prediction of lysine formylation sites using the composition of k-spaced amino acid pairs via Chou's 5-steps rule and general pseudo components. Genomics 112 (1), 859–866.

Ju, Z., Cao, J.-Z., Gu, H., 2016. Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou's general PseAAC. J. Theor. Biol. 397, 145–150.

Khan, Y.D., Ahmed, F., Khan, S.A., 2014. Situation recognition using image moments and recurrent neural networks. Neural Comput. Appl. 24 (7), 1519–1529.

Lee, T.-Y., Huang, H.-D., Hung, J.-H., Huang, H.-Y., Yang, Y.-S., Wang, T.-H., 2006. dbptm: an information repository of protein post-translational modification. Nucleic acids research 34 (suppl_1), D622–D627.

Li, S., Lu, J., Li, J., Chen, X., Yao, X., Xi, L., 2016. Hydpred: a novel method for the identification of protein hydroxylation sites that reveals new insights into human inherited disease. Mol. BioSyst. 12 (2), 490–498.

Liu, Z., Cao, J., Ma, Q., Gao, X., Ren, J., Xue, Y., 2011. Gps-yno2: computational prediction of tyrosine nitration sites in proteins. Mol. BioSyst. 7 (4), 1197–1204.

Liu, Z., Xiao, X., Qiu, W.-R., Chou, K.-C., 2015. idna-methyl: Identifying dna methylation sites via pseudo trinucleotide composition. Analytical biochemistry 474, 69–77.

Lv, Z., Zhang, J., Ding, H., Zou, Q., 2020. Rf-pseu: A random forest predictor for rna pseudouridine sites. Front. Bioeng. Biotechnol. 8.

McDowell, G., Philpott, A., 2016. New insights into the role of ubiquitylation of proteins. In: International review of cell and molecular biology, Vol. 325, Elsevier, 2016, pp. 35–88.

Nilamyani, A.N., Auliah, F.N., Moni, M.A., Shoombuatong, W., Hasan, M.M., Kurata, H., 2021. Prednts: Improved and robust prediction of nitrotyrosine sites by integrating multiple sequence features. International journal of molecular sciences 22 (5), 2704.

Qiu, W.-R., Xiao, X., Lin, W.-Z., Chou, K.-C., 2014. iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. BioMed research international.

Qiu, W.-R., Sun, B.-Q., Xiao, X., Xu, Z.-C., Chou, K.-C., 2016. iPTM-mLys: identifying multiple lysine PTM sites and their different types. Bioinformatics 32 (20), 3116–3123.

Qiu, W.-R., Jiang, S.-Y., Sun, B.-Q., Xiao, X., Cheng, X., Chou, K.-C., 2017. iRNA-2methyl: identify RNA 2'-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier. Med. Chem. 13 (8), 734–743.

Rahman, A., Ahmed, S., Rahman, J., Hasan, M.A.M., 2020. Prediction of formylation sites by incorporating sequence coupling into general pseaac. in: 2020 IEEE Region 10 Symposium (TENSYMP), IEEE, 2020, pp. 921–924.

Reddy, H.M., Sharma, A., Dehzangi, A., Shigemizu, D., Chandra, A.A., Tsunoda, T., 2019. Glystruct: glycation prediction using structural properties of amino acid residues. BMC bioinformatics 19 (13), 55–64.

Saraswathy, N., Ramalingam, P., 2011. Concepts and techniques in genomics and proteomics. Elsevier.

Shi, F., Yao, Y., Bin, Y., Zheng, C.-H., Xia, J., 2019. Computational identification of deleterious synonymous variants in human genomes using a feature-based approach. BMC medical genomics 12 (1), 81–88.

Singh, V., Sharma, A., Dehzangi, A., Tsunoda, T., 2020. Pupstruct: Prediction of pupylated lysine residues using structural properties of amino acids. Genes 11 (12), 1431.

Vapnik, V., 2013. The nature of statistical learning theory. Springer science & business media.

Wang, D., Liu, D., Yuchi, J., He, F., Jiang, Y., Cai, S., Li, J., Xu, D., 2020. Musitedeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization, Nucl. Acids Res.

Weissman, J.D., Raval, A., Singer, D.S., 2003. Assay of an intrinsic acetyltransferase activity of the transcriptional coactivator CIITA. In: Methods in enzymology, Vol. 370. Elsevier, pp. 378–386.

Xie, Y., Luo, X., Li, Y., Chen, L., Ma, W., Huang, J., Cui, J., Zhao, Y., Xue, Y., Zuo, Z., et al., 2018. Deepnitro: prediction of protein nitration and nitrosylation sites by deep learning. Genomics, proteomics & bioinformatics 16 (4), 294–306.

Xu, Y., Wen, X., Wen, L.-S., Wu, L.-Y., Deng, N.-Y., Chou, K.-C., 2014. iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PloS one 9 (8), e105018.

Xu, Y., Ding, Y.-X., Ding, J., Wu, L.-Y., Deng, N.-Y., 2015. Phogly-PseAAC: prediction of lysine phosphoglycerylation in proteins incorporating with position-specific propensity. J. Theor. Biol. 379, 10–15.

Zhang, L., Tan, B., Liu, T., Sun, X., 2019. Classification study for the imbalanced data based on Biased-SVM and the modified over-sampling algorithm. In: Journal of Physics: Conference Series, Vol. 1237, IOP Publishing, 2019, p. 022052.

predML-Site: Predicting Multiple Lysine PTM Sites with Optimal Feature Representation and Data Imbalance Minimization

Article *in* IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM · September 2021

DOI: 10.1109/TCBB.2021.3114349

CITATIONS
3

READS
35

6 authors, including:

Sabit Ahmed
Rajshahi University of Engineering & Technology
7 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Afrida Rahman
Rajshahi University of Engineering & Technology
7 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Md. Al Mehedi Hasan
Rajshahi University of Engineering & Technology
117 PUBLICATIONS **713** CITATIONS
SEE PROFILE

Julia Rahman
Rajshahi University of Engineering & Technology
27 PUBLICATIONS **87** CITATIONS
SEE PROFILE

predML-Site: Predicting Multiple Lysine PTM Sites with Optimal Feature Representation and Data Imbalance Minimization

Sabit Ahmed, *Student Member, IEEE,* Afrida Rahman, *Student Member, IEEE,* Md. Al Mehedi Hasan, Julia Rahman, Md Khaled Ben Islam, and Shamim Ahmad, *Member, IEEE*

Abstract—Identification of post-translational modifications (PTM) is significant in the study of computational proteomics, cell biology, pathogenesis, and drug development due to its role in many bio-molecular mechanisms. Computational methods for predicting multiple PTM at the same lysine residues, often referred to as K-PTM, is still evolving. This paper presents a novel computational tool, abbreviated as predML-Site, for predicting K-PTM, such as acetylation, crotonylation, methylation, succinylation from an uncategorized peptide sample involving single, multiple, or no modification. For informative feature representation, multiple sequence encoding schemes, such as the sequence-coupling, binary encoding, k-spaced amino acid pairs, amino acid factor have been used with ANOVA and incremental feature selection. As a core predictor, a cost-sensitive SVM classifier has been adopted which effectively mitigates the effect of class-label imbalance in the dataset. predML-Site predicts multi-label PTM sites with 84.18% accuracy using the top 91 features. It has also achieved 85.34% aiming and 86.58% coverage rate which are much better than the existing state-of-the-art predictors on the same rigorous validation test. This performance indicates that predML-Site can be used as a supportive tool for further K-PTM study. For the convenience of the experimental scientists, predML-Site has been deployed as a user-friendly web-server at http://103.99.176.239/predML-Site.

Index Terms—Multi-Label PTM Site Predictor, Post-Translational Modifications, Sequence-Coupling, General PseAAC, K-Spaced Amino Acid Pairs, Binary Encoding, Amino Acid Factor, Support Vector Machine, ANOVA F Test, Incremental Feature Selection, Data Imbalance Issue, Different Error Costs, Sequence Analysis

1 Introduction

POST-translational modifications (PTM) referred to the covalent addition of certain functional groups to a protein after the translation process. [1]. These modifications have significant effects on cellular processes and proteomic analysis, such as cellular signal transduction, subcellular localization, protein folding, protein degradation, and are also responsible for various kinds of diseases [2], [3]. Therefore, identifying and understanding PTM sites is critical for the basic research in disease detection, prevention, and various drug developments [4], [5].

There are 20 amino acid residues, such as alanine (A), cysteine (C), lysine (K), arginine (R), etc. Modifications occur at lysine (K) are named as lysine modification or K-PTM. Single or multiple lysine residues may be modified individually or simultaneously where one residue can influence others. In other words, these covalent modifications can aid different K-PTM types, including acetylation, crotonylation, ubiquitination, methylation, butyrylation, succinylation, biotinylation, and ubiquitin-like modifications [6], [7], [8]. Kim et al. [9], [10] first introduced a method for identifying lysine acetylation sites at the proteomic level. Later, high-resolution mass spectrometry was used by Choudhary et al. [10], [11], to detect 3600 lysine acetylation sites. Various large-scale proteomic technologies, such as stable isotope labeling, high-performance liquid chromatography, high-resolution liquid chromatography–tandem mass spectrometry, etc. were developed to identify lysine crotonylation and succinylation sites [12], [13], [14], [15], [16]. Conventional experimental methods such as, Chip–Chip [17], methylation-specific antibodies [18], and mass spectrometry [19], [20], had been developed to identify methylation sites. However, these approaches are often time-consuming, expensive, and require a high level of technical expertise. As a result, computational methods are getting popularity as an effective alternative because of their laborsaving, time, and cost-efficient characteristics.

Though there are several computational tools for predicting various K-PTMs separately, to the best of the authors' knowledge, only two multi-label prediction systems have been developed so far that can take care of the multiplex Lys residues [8], [10], [21], [22], [23], [24]. One is iPTM-mLys, and the other one is mLysPTMpred, which can predict multiple lysine modification sites as well as their different types [5],

- *Sabit Ahmed, Afrida Rahman and Md. Al Mehedi Hasan were with the Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Kazla, Rajshahi-6204, Bangladesh. E-mail: sabit.a.sirat@gmail.com, afrida.r.samma@gmail.com and mehedi_ru@yahoo.com*
- *Julia Rahman was with the Institute for Integrated and Intelligent Systems, Griffith University, Brisbane, Australia. E-mail: julia.rahman@griffithuni.edu.au*
- *Md Khaled Ben Islam was with the Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna, Bangladesh. E-mail: mdkhaledben@gmail.com*
- *Shamim Ahmad was with the Department of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh. E-mail: shamim_cst@yahoo.com*

Manuscript received August 08, 2020; revised July 06, 2021. (Corresponding author: Sabit Ahmed)

[25]. Both predictors have utilized the vectorized sequence-coupled model as a feature extraction method [5], [25], [26], [27]. The former has employed the random forest algorithm while the last one has used the support vector machine for classification [25], [28]. Both the predictors, as mentioned earlier, need pronounced elevation in terms of the prediction quality. Therefore, a tool with higher efficacy is required to meet the current demand in the study of post-translational modifications.

For developing a successful predictor for PTM sites, one of the main challenges is to elicit features from the input protein sequences as the appropriate features can play a crucial role in better prediction performance [25]. This study considered several feature encoding methods to propose a novel multi-label predictor predML-Site, where the vectorized sequence-coupled model [5], [21], [22], [26], [27], pairs of k-spaced amino acids [29], [30], encoded binary features, and amino acid factors [23], [30] were aggregated to encode a peptide segment. Afterward, the analysis of variance (ANOVA) F test statistic along with the incremental feature selection approach was used to eliminate the redundant and trivial features [29], [31], [32]. The support vector machine classifier with the variable cost adjustment process [25] was implemented to handle the imbalance in dataset [33]. A 5-fold cross-validation [25] scheme was repeated five times for validating the statistical significance of the prediction results, and the average performance of each metric has been reported.

Finally, predML-Site achieved a more desirable performance than both of the iPTM-mLys and mLysPTMpred [5], [25]. It has attained an accuracy of 84.18% with 85.34% aiming rate and 86.58% coverage rate. Besides, the most stringent metric absolute-true rate is increased from 79.73% to 80.56%. The performance obtained through the cross-validation test indicates that predML-Site can be considered as a highly supportive and constructive tool for multi-label site prediction.

Based on the recent collections of publications [5], [8], [22], [25], [29], [34], [35], [36], [37], to develop an efficient predictor with regards to computational biology, one must go through Chou's five-step [24], [25], [38] guidelines: i) generating an acceptable benchmark dataset for training and testing the system, ii) formulating the sequences with the proper mathematical representations, iii) developing a prediction engine or introducing to a powerful prediction algorithm, iv) conducting cross-validation tests properly to evaluate the predictive accuracy, and v) providing an accessible and user-oriented web-server. In accordance with these steps, details of materials, methods, results and analysis will be discussed in Section 2 and 3 respectively.

2 MATERIALS AND METHODS

2.1 Benchmark dataset

In the current study, the K-PTM dataset utilized in iPTM-mLys and mLysPTMpred [5], [25] had been used for prediction model development and benchmarking. This benchmark dataset containing only human protein sequences was collected from the Universal Protein Resource (UniProt) [39]. Various constraints had been applied to derive the dataset by using the advanced search option, such as i) Selected the 'PTM/Processing' and 'Modified residue [FT]' option for 'Fields' [5]. ii) Selected 'Any experimental assertion' for 'Evidence' [5]. iii) Typed 'human' for searching. [25]. iv) Made use of the keywords: 'acetyllysine', 'crotonyllysine', 'methyllysine', or 'succinyllysine' [5]. v) Picked up the proteins consisted of fifty and more amino acids [5], [25]. In this process, 1769 proteins were obtained. For formulating PTM sites in a more meticulous and comprehensive way, Chou's scheme [40] was adopted, similarly iPTM-mLys [5] and mLysPTMpred [25]. According to this scheme, a peptide segment could generally be expressed by

$$P_\zeta(K) = Q_{-\zeta}Q_{-(\zeta-1)}...Q_{-2}Q_{-1}KQ_{+1}Q_{+2}...Q_{+(\zeta-1)}Q_{+\zeta} \quad (1)$$

where the symbol K denoted the responsible residue 'lysine' at the center, the subscript ζ being an integer, $Q_{-\zeta}$ and $Q_{+\zeta}$ denoted the ζ-th leftward and ζ-th rightward amino acid residues from the center, and so forth. Furthermore, a peptide sequence $P_\zeta(K)$ could be categorized into two types [25]

$$P_\zeta(K) \in \begin{cases} P_\zeta^+(K), \text{if its center is K-PTM site} \\ P_\zeta^-(K), \text{otherwise} \end{cases} \quad (2)$$

where $P_\zeta^+(K)$ contained the positive subset of the peptides and $P_\zeta^-(K)$ contained the negative subset of the peptides with a lysine (K) residue at its center, and the symbol $\in$ indicated the set theory relationship [25].
For equal-sized K-PTM site formation, $(2\zeta+1)$-tuple peptide window with K at its center was employed. During segmentation, the lacking amino acid at both the right and left end was filled with the nearest residue [5], [25]. After the peptide fragments went through some screening, such as the elimination of sequences in case of redundancy, a total of four benchmark datasets were constructed similarly to iPTM-mLys [5] with the given form

$$S_\zeta(\text{K-type}) = S_\zeta^+(\text{K-type}) \cup S_\zeta^-(\text{K-type}) \quad (3)$$

where the positive subset $S_\zeta^+(\text{K-type})$ could contain only the true K-type peptide samples, while the negative subset $S_\zeta^-(\text{K-type})$ could contain only the false K-type samples with K at the center [5], [25]. It is to be noted that only the specific type of K-PTM, i.e. 'acetylation', or 'crotonylation', or 'methylation' or 'succinylation' must be used separately and consistently as the 'K-type' described in Eq. (3). A detailed overview of different datasets was presented in Fig. 1. After going through many preliminary tests, window size was chosen as $(2\zeta+1) = 27$, where $\zeta = 13$. Therefore, Eq. (1) had been reduced to

$$P(K) = Q_{-13}Q_{-12}...Q_{-2}Q_{-1}KQ_{+1}Q_{+2}...Q_{+12}Q_{+13} \quad (4)$$

A summary of the benchmark dataset obtained by this process is given in Table 1.

2.2 Feature construction

With the evolution of the biological sequences, several encoding methods have been developed for extracting pertinent features hidden in the sequences. After preliminary analysis, it was observed that the vectorized sequence coupling, pairs of k-spaced amino acids [29], [30], encoded binary features and amino acid factor [23], [30] were more

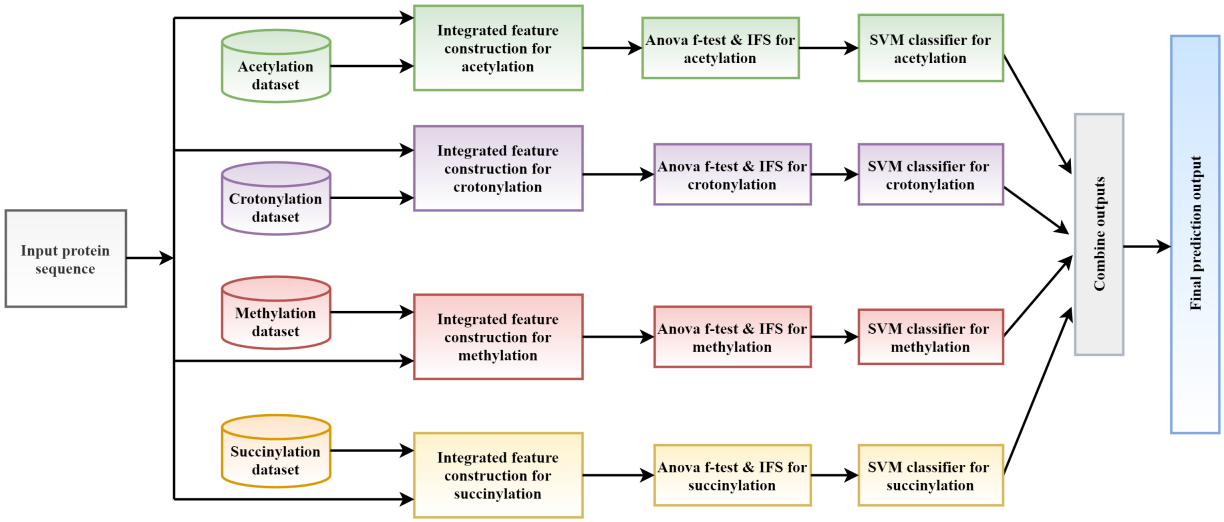

Fig. 1. The system flowchart of predML-Site.

TABLE 1
Four Benchmark Datasets Overview.

Number of samples regarding the PTM types				
Attribute	Ace	Cro	Met	Suc
	S(1)	S(2)	S(3)	S(4)
True	3991	115	127	1169
False	2403	6279	6267	5225

Ace, acetylation; Cro, crotonylation; Met, methylation; Suc, succinylation.

appropriate for representing the protein sequences of the multiple lysine modification sites than any other encoding methods.

2.2.1 Sequence-coupling

The composition of pseudo amino acid or PseAAC [41], [42], [43], [44], [45] was designed to preserve the sequence pattern information, which is a much harder task for any existing machine learning algorithm [15]. In this study, incorporating sequence coupling information into Chou's general PseAAC was adopted for extracting features from peptide sequences [5], [25], [26], [27], [42]. Based on this conception, the segmented proteins of Eq. (4) could be defined as

$$P(K) = P^{+}(K) - P^{-}(K) \tag{5}$$

where

$$P^{+}(K) = \left[P^{C^{+}}_{-13} \;\; P^{C^{+}}_{-12} \ldots P^{+}_{-1} \;\; P^{+}_{+1} \ldots P^{C^{+}}_{+12} \;\; P^{C^{+}}_{+13} \right]^{T} \tag{6}$$

$$P^{-}(K) = \left[P^{C^{-}}_{-13} \;\; P^{C^{-}}_{-12} \ldots P^{-}_{-1} \;\; P^{-}_{+1} \ldots P^{C^{-}}_{+12} \;\; P^{C^{-}}_{+13} \right]^{T} \tag{7}$$

where $P^{C^{+}}_{-13}$ in Eq. (6) denoted the conditional probability [5], [25] of amino acid Q_{-13} at the leftmost position given that its adjacent right member was Q_{-12} and so forth [46]. In contrast, only P^{+}_{-1} and P^{+}_{+1} were of non-contingent probability as K was the adjoining member of both amino acids at position Q_{-1} and Q_{+1}. All the conditional probability values were extracted from the positive training dataset of iPTM-mLys [5], [8], [24], [25]. Additionally, all the probability values in Eq. (7) were identical to those of Eq. (6) other than that they could be derived from the negative training dataset [5]. Thus, after omitting K from the center, $(27-1) = 26$ dimension feature vectors were obtained.

2.2.2 Pairs of k-spaced amino acids

The formation of k-spaced amino acid pairs encoding technique [23], [29], [47], [48] calculated the occurrence frequencies of the pairs of k-spaced amino acids from a segmented protein sample, that could express the short linear motif information out of it [23], [29], [49]. For instance, the encoding of a peptide segment would be a 441-dimensional feature vector if $k = 1$ [29], [48]. This could be defined as

$$(N_{AnA}/N_{Total}, N_{AnC}/N_{Total}, ..., N_{XnX}/N_{Total})_{441} \tag{8}$$

where n stood for any of amino acid including the dummy residue $'X'$, N_{Total} meant the occurrence frequency of all k-spaced amino acid pairs [48] and N_{AnA} meant the occurrence frequency of the AnA pairs in the segment [29], [30], [50] when $k = 1$. In this study, after merging each of the 441-dimension feature vectors for $k = 0, 1, 2, 3$, a total of 1764-dimensional features were formed.

2.2.3 Binary encoding

Binary encoding [23] could represent the amino acid position and composition by using 20 binary bits for one amino acid [23]. But one additional bit was conjoined to handle the complexity of sliding windows. For 21 amino acids structured as $'ACDEFGHIKLMNPQRSTVWYX'$, each residue inside a sequence fragment was formed by a 21-dimension binary vector [23]. For instance, residue $'A'$, $'G'$ and $'X'$ were encoded as '100000000000000000000',

'00000010000000000000' and '00000000000000000001' respectively [23]. According to this concept, each resultant peptide segment was expressed as $27\times21 = 567$-dimensional feature vectors.

2.2.4 Amino acid factor

Five multidimensional attributes [30], [51], which included polarity, secondary structure, molecular volume, electrostatic charge, and codon diversity [35], were constructed from AAIndex by using multivariate statistical analysis [23]. These five transformed properties could be introduced as amino acid factors (AAF) [35]. Since the AAF could reduce the dimensionality of the feature space of physicochemical properties efficiently, it had been utilized in many biological studies [23], [35], [52]. The dimensionality of feature vectors was calculated as follows

$$D = peptide\ sequence\ length \times number\ of\ factors \quad (9)$$

With a peptide sequence of length 27 and previously described five amino acid factors, $27\times5 = 135$ dimension features were derived by using this formula.

2.2.5 Feature ensembling

Initially, the four aforestated feature encoding techniques (i.e. sequence coupling, CKSAAP, BE, and AAF) were implemented separately to encode the training peptides. However, for extracting more PTM-contextual information from the protein sequences, encoded features were ensembled serially, scaled through standardization and obtained $26 + (441\times4) + (27\times21) + (27\times5) = 2492$ features.

2.3 Feature selection

Since the dimension of the encoded features was higher, irrelevant, and redundant features needed to be removed to avoid learning complexity. For this reason, the analysis of variance (ANOVA) F test statistic technique [31], [53] was adopted. It tested the null hypothesis (i.e. all the means of different groups were equal) against the alternative hypothesis (i.e. all the means differed from each other). The one-way ANOVA could be defined as

$$F = \frac{(n-k)\sum n_i(\overline{\mathrm{Y}}_{i.} - \overline{\mathrm{Y}}_{..})^2}{(k-1)\sum(n_i-1)s_i^2} \quad (10)$$

where
$n = \sum_{i=1}^{k} n_i$, $\overline{\mathrm{Y}}_{i.} = Y_{i.}/n_i$, $\overline{Y}_{..} = Y_{..}/n$
and
$s_i^2 = \sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2/(n_i-1)$.

It should be mentioned that the dot in $Y_{i.}$ indicates an aggregation over the j index [53].
where $Y_{i.} = \sum_{j=1}^{n_i} Y_{ij}$ and $Y_{..} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$.
The calculated F values were used to rank the features. for higher F values, the discriminative capability of a predictor is better.

2.4 Support vector machine

The support vector machine (SVM) [49], [54], [55], one of the dominant statistical learning algorithms was adopted as a core prediction algorithm. It seeks the optimum hyperplane with the highest margin between two groups [25], [34], which requires solving constrain problem as below

$$minimize_{w,b}\frac{1}{2}||w||^2$$

Subject to:

$$y_i(w^T x_i + b) \geq 1, i = 1, 2, 3, .. \quad .., n \quad (11)$$

where x_i denotes the input training vectors, y_i denotes the corresponding class labels, w is weight vector, and b denotes the bias or threshold. To permit errors, the optimal problem now becomes:

$$min_{w,b}\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i$$

Subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, i = 1, 2, 3, .. \quad .., n$$
$$\xi_i \geq 0, i = 1, 2, 3, .. \quad .., n \quad (12)$$

where C is penalty parameter and ξ denotes slack variables. Using the Lagrange multipliers [25], [54], the dual formulation expressed in terms of variables α_i can be obtained as

$$maximize_\alpha \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(x_i, x_j)$$

Subject to:

$$\sum_{i=1}^{n} y_i\alpha_i = 0, \quad 0 \leq \alpha_i \leq C$$
$$\text{for all } i = 1, 2, 3, .., n \quad (13)$$

After involving the kernel function [25], the discriminant function of SVM took the following form

$$f(x) = \sum_{i}^{n}\alpha_i y_i k(x, x_i) + b \quad (14)$$

In this paper, the radial basis function kernel [25], [56] was applied to construct SVM classifier and given by,

$$k(x_i, x_j) = exp(-\gamma\|x_i - x_j\|^2) \quad (15)$$

where $\gamma > 0$ [29], [57]. As the benchmark dataset was highly imbalanced, different error cost (DEC) [25] method had been used to tackle the class imbalance problem [33], [58]. According to this approach, the SVM soft margin objective function [25] was adjusted to allocate two costs for misclassification [23], [24], [29], such as C^+ for the positive class instances and C^- for the negative class instances

$$C^+ = C * W^+, \qquad C^- = C * W^- \quad (16)$$

In Eq. (16), W^+ is the weight for the positive instances and W^- is the weight for the negative instances and defined by $W^+ = \frac{M}{2*M_1}$, $W^- = \frac{M}{2*M_2}$
where M is the total number of elements, M_1 is the number of elements for the positive class, and M_2 is the number of elements for the negative class.

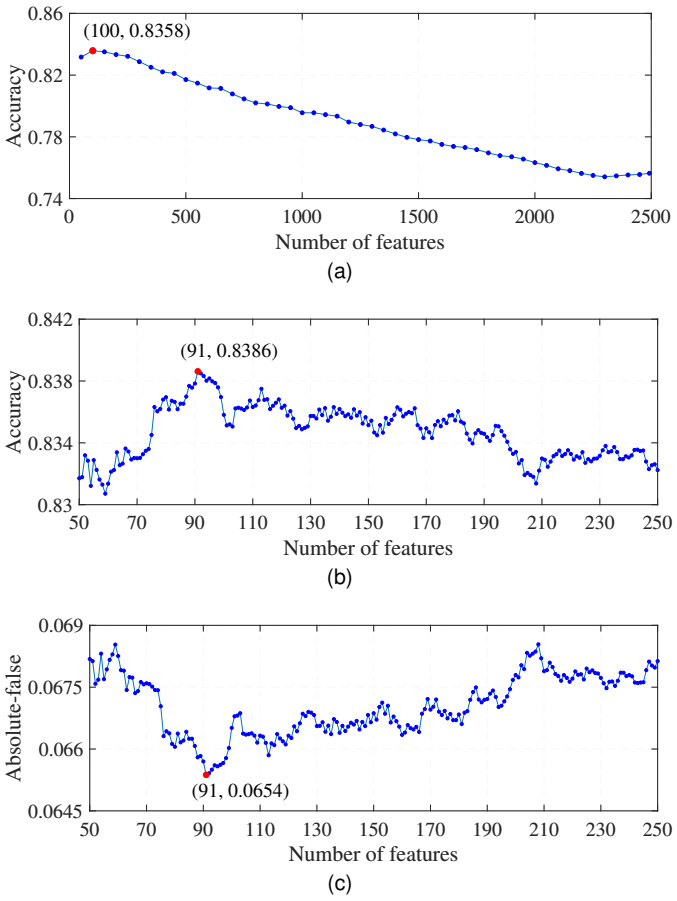

Fig. 2. The IFS curves: (a) Feature range 50 to 2492 (Features Vs Accuracy); (b) Feature range 50 to 250 (Features Vs Accuracy); (c) Feature range 50 to 250 (Features Vs Absolute-false).

2.5 Validation of the proposed model

Among various statistical techniques for validating the performance of prediction models, K-fold cross-validation [25] is the widely used method, particularly for time complexity. Additionally, to establish a fair comparison with the state-of-the-art K-PTM prediction methods, a 5-fold cross-validation test [25], [59], [60] has been executed five times (i.e. 25 iterations in total), because of having insufficient details about the exact 5-way splits [25]. In each 5-fold cross-validation, the training dataset was randomly divided into five approximately equal-sized disjoint sets, [25], [29]. The average results of all metrics were reported with their respective standard deviations for the evaluation of the novel predictor.

2.6 Evaluation metrics

In our benchmark dataset, peptide samples were 6394 in total, of which 3991 were labeled with 'acetylation', 115 with 'crotonylation', 127 with 'methylation', 1169 with 'succinylation', and 1750 with 'non-K-PTM' [5], [25]. Since a sample can contain more than one labels, metrics for multi-label systems were utilized [5], [25] instead of ordinary metrics for single-label systems [10], [21], [22], [23], [24], [29], [61], [62], [63], [64], [65], [66]. For evaluation, we have estimated Aiming, Coverage, Accuracy, Absolute-True and Absolute-False [34], (defined in (17)).

$$\begin{cases} Aiming = \frac{1}{N}\sum_{i=1}^{N}(\frac{\|Y_i \cap Y_i'\|}{\|Y_i'\|}) \\ Coverage = \frac{1}{N}\sum_{i=1}^{N}(\frac{\|Y_i \cap Y_i'\|}{\|Y_i\|}) \\ Accuracy = \frac{1}{N}\sum_{i=1}^{N}(\frac{\|Y_i \cap Y_i'\|}{\|Y_i \cup Y_i'\|}) \\ Absolute-true = \frac{1}{N}\sum_{i=1}^{N}(\Delta\|Y_i, Y_i'\|) \\ Absolute-false = \frac{1}{N}\sum_{i=1}^{N}(\frac{\|Y_i \cup Y_i'\| - \|Y_i \cap Y_i'\|}{L}) \end{cases} \quad (17)$$

where N and L were the total number of the samples and the total number of labels in the system respectively [5], [25], $\cup$ and $\cap$ denoted the 'union' and 'intersection' in the set theory, $||\ ||$ meant the operator acting on the set to calculate the number of its elements [25], Y_i and Y_i' denoted the subset that contained all the labels experiment-observed and all the labels predicted [5], [25] for the i^{th} sample respectively, and

$$\Delta(Y_i, Y_i') = \begin{cases} 1, \text{ if all labels in } Y_i' \text{ and } Y_i \text{ are identical} \\ 0, \text{ otherwise} \end{cases}$$

The metrics defined above have been applied effectively in several multi-label based systems [5], [25].

3 Results and Discussions

3.1 Incremental feature selection

The feature selection procedure was implemented in several steps. First of all, each feature was tested with the analysis of variance (ANOVA) and obtained the features with statistical significance [67]. Hence, all of the 2492 features were ranked according to the calculated F values.

Later, the incremental feature selection (IFS) [29] algorithm was applied in two steps for selecting the optimal number of features as well as saving the computational time [23], [67]. Primarily, for each feature subset of top m (m = 50, 100, 150, ..., 2492), one SVM classifier with LibSVM's default parameter [49], [68] was trained for each K-PTM type and measured its accuracy by adopting 5-fold cross-validation. As depicted in Fig. 2a, the highest accuracy of 83.58% was achieved with 100 leading features. In addition to that, the rate of accuracy was lower with the subset of 50 features and significantly dropped again after taking the subset of 250 features.

Therefore, incremental feature selection with a subset of n (n = 50, 51, 52, ..., 250) features was employed with the identical experimental set-ups as described previously. At that stage, the performance metric accuracy and absolute-false rate were taken into account for evaluating each feature subset. By this procedure, the highest accuracy of 83.86% with the lowest absolute-false rate of 6.54% was obtained with 91 optimal features, as illustrated in Figure 2b and 2c. At the same time, the aiming, coverage, and absolute-true rate reached 85.35%, 85.83%, and 80% respectively. Later, these top 91 features were utilized to construct the proposed system predML-Site.

TABLE 2
Selected C and γ of Five Times Run of 5-Fold Cross-Validation.

Iterations	Type of PTM							
	Acetylation		Crotonylation		Methylation		Succinylation	
	C	γ	C	γ	C	γ	C	γ
1^{st}	$2^{3.6}$	$2^{-9.7}$	$2^{1.5}$	$2^{-9.1}$	$2^{1.3}$	$2^{-9.8}$	$2^{3.5}$	$2^{-9.3}$
2^{nd}	$2^{3.5}$	$2^{-9.7}$	$2^{0.1}$	$2^{-9.5}$	$2^{1.5}$	$2^{-10.1}$	2^{3}	2^{-9}
3^{rd}	$2^{3.5}$	$2^{-9.7}$	2^{0}	$2^{-8.8}$	$2^{0.9}$	$2^{-10.4}$	$2^{3.2}$	$2^{-8.9}$
4^{th}	$2^{3.7}$	$2^{-9.8}$	2^{0}	$2^{-9.5}$	$2^{0.9}$	$2^{-9.4}$	$2^{1.5}$	$2^{-8.4}$
5^{th}	$2^{2.5}$	$2^{-9.1}$	$2^{1.1}$	$2^{-8.9}$	$2^{0.9}$	$2^{-9.8}$	$2^{0.7}$	$2^{-8.1}$

TABLE 3
Selected C and γ for the Web-server.

PTM types	C	γ
Acetylation	$2^{3.6}$	$2^{-9.7}$
Crotonylation	$2^{1.5}$	$2^{-9.1}$
Methylation	$2^{1.5}$	$2^{-10.1}$
Succinylation	$2^{3.5}$	$2^{-9.3}$

3.2 Model development

In this study, four separate SVM classifiers [25] were used to predict the acetylation, crotonylation, methylation, and succinylation sites separately. Each of the classifiers performed binary classification on the benchmark dataset described in Table 1. For all four K-PTM types, necessary features were extracted by integrating multiple encoding methods and selected 91 optimal features with ANOVA F-test and IFS to train the models, as shown in Fig. 1.

The radial basis function (RBF) kernel [34], [69] was used for each SVM classifier, and the grid-search technique was implemented in two steps to find the best parameters. A coarse grid was conducted at first, and after identifying the best regions on that grid (i.e. $C = 2^0, 2^1, 2^2, 2^3, 2^4$ and $\gamma = 2^{-8}, 2^{-9}, 2^{-10}$), a finer grid-search [70] was applied to the neighbourhood. As there was a lack of details about the exact 5-way splits of the dataset [34], five complete runs of 5-fold cross-validation were executed [5], [25], [71]. Each time, the best parameters of each classifier were selected depending on the area under curve (AUC) value [25], which was an essential metric for single-label systems [8], [10], [23], [24], [29]. The misclassification cost C was calculated according to the Eq. (16) for handling the data imbalance issue. The selected parameter values of C (penalty point for soft margin) and γ for each K-PTM type are given in Table 2.

Eventually, after training the four binary SVM classifiers with the appropriate hyperparameters, multi-label predictor predML-Site was constructed by combining the outputs from these classifiers [34], as depicted in Fig. 1. Five times repetition of the 5-fold cross-validation [34] produced five sets of values of all metrics, which were defined in Section 2.6. The average results of each multi-label metric were taken to evaluate the final model.

However, in order to develop the web-server, the C and γ for each classifier were selected from the five sets of parameters based on the highest AUC value. The selected parameters were given in Table 3. It should be mentioned that Matlab 2019a and python 3.7.3 were utilized to implement the system.

3.3 Comparison with the existing predictors

The performance of the predML-Site predictor derived from the aforementioned multi-label metrics was given in Table 4 with corresponding standard deviations. The values of the five metrics were the average result of five times complete run of 5-fold cross-validation on the benchmark dataset [5], [25]. It is to be noted that Table 4 also includes the corresponding performances on the same dataset demonstrated by the existing multi-label predictors named as iPTM-mLys and mLysPTMpred [5], [25].

In Eq. (17), for the first four metrics, the higher the rate was, the better the performance would be, and for the last one, it was entirely the opposite [25], [72]. In comparison with the recently developed multi-label predictor mLysPTMpred, it can be observed that the rate of the most crucial metric 'Accuracy' for the proposed predictor predML-Site had been increased from 83.73% to 84.18%. Besides, the experimentally obtained rate of the most stringent and harsh metric 'Absolute-True' [72] was 80.56%. Very few multi-label predictors in computational biology could reach over 50% for the absolute-true rate [5], [25]. Apart from that, the acquired absolute-true rate of mLysPTMpred and predML-Site could reach over 80%, and the proposed predictor outperformed mLysPTMpred by 0.83%. It should also be mentioned that the rate of 'Absolute-False' or 'Hamming-Loss' [25], [72], [73] denoting the average ratio of completely

TABLE 4
Performance Comparison Between predML-Site and Existing Predictors.

Predictor	Aiming(%)	Coverage(%)	Accuracy(%)	Absolute-true(%)	Absolute-false(%)
iPTM-mLys	69.78	74.54	68.37	60.92	13.40
mLysPTMpred	84.82	86.56	83.73	79.73	6.66
predML-Site	**85.34** ($\pm$0.18)	**86.58** ($\pm$0.17)	**84.18** ($\pm$0.19)	**80.56** ($\pm$0.24)	**6.41** ($\pm$0.08)

TABLE 5
Comparison Between predML-Site and Existing Predictors Based on the Independent Test.

Predictor	Aiming(%)	Coverage(%)	Accuracy(%)	Absolute-true(%)	Absolute-false(%)
iPTM-mLys	67.50	65.00	62.50	55.00	15.00
mLysPTMpred	88.33	87.50	85.83	80.00	6.00
predML-Site	**90.00**	**87.50**	**87.50**	**85.00**	**5.00**

TABLE 6
The Predictive Performance of Various Feature Encoding Techniques on the Benchmark Dataset.

Feature type	Feature count	Aiming(%)	Coverage(%)	Accuracy(%)	Absolute-true(%)	Absolute-false(%)
AAF	135	55.59 ($\pm$0.30)	55.86 ($\pm$0.25)	52.69 ($\pm$0.31)	46.60 ($\pm$0.42)	19.93 ($\pm$0.11)
BE	567	58.17 ($\pm$0.24)	57.92 ($\pm$0.31)	55.39 ($\pm$0.25)	50.05 ($\pm$0.22)	18.78 ($\pm$0.10)
$CKSAAP_{k=0}$	441	54.18 ($\pm$0.21)	54.82 ($\pm$0.23)	51.54 ($\pm$0.21)	45.60 ($\pm$0.22)	20.36 ($\pm$0.07)
$CKSAAP_{k=0,1}$	882	55.58 ($\pm$0.37)	55.94 ($\pm$0.37)	53.02 ($\pm$0.41)	47.53 ($\pm$0.51)	19.68 ($\pm$0.16)
$CKSAAP_{k=0,1,2}$	1323	56.34 ($\pm$0.28)	56.51 ($\pm$0.30)	53.72 ($\pm$0.29)	48.31 ($\pm$0.28)	19.36 ($\pm$0.11)
$CKSAAP_{k=0,1,2,3}$	1764	56.70 ($\pm$0.15)	56.63 ($\pm$0.23)	54.01 ($\pm$0.17)	48.70 ($\pm$0.15)	19.18 ($\pm$0.06)
Sequence-coupling	26	84.80 ($\pm$0.15)	85.80 ($\pm$0.14)	83.45 ($\pm$0.18)	79.72 ($\pm$0.28)	6.69 ($\pm$0.07)
Combined features	2492	78.02 ($\pm$0.20)	77.12 ($\pm$0.16)	75.64 ($\pm$0.18)	71.74 ($\pm$0.20)	9.98 ($\pm$0.07)
Optimal features$_{(predML-Site)}$	**91**	**85.34** ($\pm$0.18)	**86.58** ($\pm$0.17)	**84.18** ($\pm$0.19)	**80.56** ($\pm$0.24)	**6.41** ($\pm$0.08)

wrong hits over the total prediction events is 6.41%, which is 0.25% lower than mLysPTMpred [25].

Furthermore, as for predML-Site, the rate of 'Aiming' or 'Precision' [72], [73] representing the average ratio of the predicted labels that hit the target of the original labels [5], had become 85.34% from 84.82%. The average ratio of the original labels that were covered by the hits of prediction referred to 'Coverage' [72], [73] was also increased from 86.56% to 86.58% compared to mLysPTMpred [25].

Therefore, all the experimental results from Table 4 indicated that the novel multi-label predictor predML-Site achieved better performance than both the iPTM-mLys and mLysPTMpred in terms of 'Aiming', 'Coverage', 'Accuracy', 'Absolute-True' and 'Absolute-False' [5], [25], [34]

It should be mentioned that an independent test was conducted with the protein sequence (Q16778) given in both iPTM-mLys and mLysPTMpred [5], [25]. The labels predicted by predML-Site as well as the experimentally annotated labels were reported in Table 7. According to Eq. (17), the obtained results were: aiming rate = 90.00%, coverage rate = 87.50%, accuracy = 87.50%, absolute-true rate = 85.00% and absolute-false rate = 5.00%, almost identical to the cross-validation performance delineated in Table 4. According to the experimental results in Table 5, it is evident that the proposed predictor predML-Site achieved superior performance regarding the independent test.

3.4 Comparison with different feature extraction methods

The performance obtained by predML-Site was further compared with multiple baseline K-PTM prediction methods, developed using different feature extraction methods, such as the incorporation of sequence coupling information into

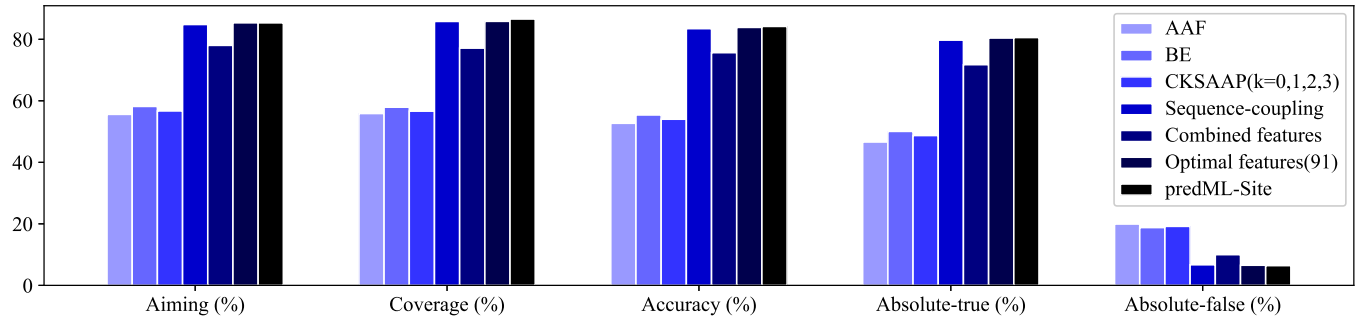

Fig. 3. Comparison of the different feature extraction methods.

general PseAAC [35], the formation of k-spaced amino acid pairs [30], [74], binary encoding and amino acid factors [23], [26], [30], [52], [74] to estimate predML-Site's K-PTM related information extraction capability. The performances of the specified feature encoding schemes evaluated by 5-fold cross-validation are reported in Table 6 with the corresponding standard deviations.

It may be observed that the amino acid factor (AAF) acquired a higher absolute-false rate of 19.93% with considerably lower accuracy, absolute-true, aiming, and coverage rate. However, slightly better results were picked up by binary encoding (BE) schemes. It reached 55.39% accuracy with a 58.17% aiming rate and a 57.92% coverage rate. The absolute-false rate was reduced to 18.78% with an absolute-true rate of 50.05%.

The composition of k-spaced amino acid pairs (CKSAAP) [23], [29], [75] encoding technique was adopted for the different combinations of k, in which the 0-spaced ($k = 0$) amino acid pairs produced the lowest accuracy, aiming, coverage and absolute-true rate and the highest absolute-false rate. The performances secured by the composition of 1-spaced ($k = 0, 1$) and 2-spaced ($k = 0, 1, 2$) amino acid pairs were improved a little and maximized for the composition of 3-spaced ($k = 0, 1, 2, 3$) amino acid pairs as illustrated in Table 6. It achieved 54.01% accuracy, which was the topmost accuracy among the various combinations of CKSAAP encoding schemes but compared to other feature extraction techniques, it was not a desirable performance.

Sequence-coupling, which was one of the most crucial encoding strategies, attained a higher accuracy rate of 83.45%, a coverage rate of 85.80% with a much lower absolute-false rate of 6.69%.

Therefore, integrating all the feature extraction methods was considered as a successful approach for developing a multi-label predictor. Consequently, the sequence-coupling was combined with amino acid factor, binary encoding and the composition of k-spaced amino acid pairs where k = 0, 1, 2, 3. But the performance of the integrated features degraded and for 2492 dimension features, accuracy was reduced to 75.64% with the increased absolute-false rate of 9.98%. That being case, 91 optimal features were selected from the high dimension features by conducting ANOVA F-test and IFS. Finally, the proposed method predML-Site was constructed using the 91 ideal features. With the tuned RBF kernel parameters, it gained the highest accuracy of 84.18%

TABLE 7
Comparison Between Predicted and Experimental Results on Example Protein Q16778.

Sites	Predicted result				Experimental result			
	Ace	Cro	Met	Suc	Ace	Cro	Met	Suc
6	+	+	-	-	+	+	-	-
12	+	+	-	-	+	+	-	-
13	+	+	-	-	+	+	-	-
16	+	+	-	-	+	+	-	-
17	+	+	-	-	+	+	-	-
21	+	+	-	-	+	+	-	-
24	+	+	-	-	+	+	-	-
25	-	-	-	-	-	-	-	-
28	-	-	-	-	-	-	-	-
29	-	-	-	-	-	-	-	-
31	-	-	-	-	-	-	-	-
35	-	+	-	-	-	+	-	-
44	-	-	-	-	-	-	-	-
47	-	-	+	-	-	-	+	-
58	-	-	+	-	-	-	+	-
86	+	-	+	-	+	-	+	-
109	-	-	+	-	-	-	+	-
117	+	-	-	-	-	-	-	-
121	+	-	-	-	+	+	-	-
126	+	-	-	-	-	-	-	-

with an 80.56% absolute-true rate. It also minimized the absolute-false rate, which is 6.41%. Table 6 and Fig. 3 both point out that predML-Site achieved a discernible performance among all the feature encoding techniques described earlier.

3.5 Optimal features analysis

The feature distribution for different K-PTM types was shown in Fig. 4. Additionally, the percentage of each type of feature selected with ANOVA and IFS were tabulated in Table 8 for a better understanding of the importance and dominance of the corresponding features. For the acetylation feature set, out of 91 optimal features, 26 belonged to the sequence-coupling, 24 belonged to the BE, 12 belonged to the AAF, and 29 belonged to the CKSAAP. Therefore, the ratios of selected dimensions of these four types of features are 100% (26/26), 4.23% (24/567), 8.89% (12/135), and 1.64% (29/1764) respectively.

The crotonylation feature set was made of 26 sequence-coupling features, 10 BE features, and 55 CKSAAP features. From Fig. 4 and Table 8, it could be seen that the optimal feature set of crotonylation did not contain any of the AAF features. Hence, the selected dimension ratios of sequence-coupling, BE, and CKSAAP features are 100% (26/26), 1.76% (10/567), and 3.12% (55/1764), respectively. Besides, the methylation feature set consisted of 26 sequence-coupling features, 10 BE features, 1 AAF feature, and 54 CKSAAP features, and the ratios of the selected dimensions for each type of features are 100% (26/26), 1.76% (10/567), 0.74% (1/135), and 3.06% (54/1764) respectively. For the succinylation dataset, 26, 8, 6, and 51 features belonged to the sequence-coupling, BE, AAF, and CKSAAP respectively. The dimension ratios for sequence-coupling, BE, AAF, and CKSAAP are 100% (26/26), 1.41% (8/567), and 4.44% (6/135), respectively.

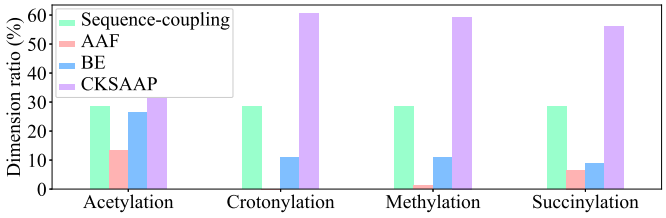

Fig. 4. Feature distribution in the optimal feature sets.

TABLE 8
Percentage of Features Selected with ANOVA F-Test and IFS.

Feature name	Ace	Cro	Met	Suc
AAF (%)	8.89	0.00	0.74	4.44
Sequence-coupling (%)	100	100	100	100
BE (%)	4.23	1.76	1.76	1.41
CKSAAP (%)	1.64	3.12	3.06	2.89

As reflected in both Fig. 4 and Table 8, the sequence-coupling features had a stronger influence on the identification of acetylation, crotonylation, methylation and succinylation sites. In contrast, the BE and CKSAAP features had a smaller effect on identifying all four types of K-PTMs. AAF features had a slightly better impact on the acetylation and succinylation site prediction but it had barely any effect on the crotonylation and methylation site prediction. The selected feature dimensions for BE, AAF, and CKSAAP varied over different types of K-PTM site prediction. Therefore, it may be concluded that the proposed model augmented the sequence-coupling effect with the essential features of CKSAAP, BE, and AAF and intensified the prediction performance.

3.6 Analysis on different modifications

The multi-label predictor predML-Site was developed by combining outputs from the four optimized binary classifiers as discussed in Section 2. Though the final output was evaluated by the multi-label metric system, each of the individual classifiers was evaluated and tuned depending on the area under curve (AUC) value. The average AUC values of 5-times 5-fold cross-validation of each classifier were reported in Table 9.

TABLE 9
Performance of Different Modifications.

	Ace	Cro	Met	Suc
AUC (%)	94.85 (±0.04)	99.97 (±0.01)	99.97 (±0.01)	96.98 (±0.03)

It could be seen that the average AUC of acetylation and succinylation classifiers were 94.85% and 96.98% respectively. On the other hand, the average AUC of crotonylation and methylation is 99.97%. We came to know from Table 1 that crotonylation and methylation datasets had almost similar imbalance ratios (around 50) while acetylation and succinylation datasets had different imbalance ratios. Moreover, the percentage of different types of selected features were almost identical for crotonylation and methylation sites (see Table 8), where the sequence-coupling, BE, and CKSAAP features were considered more important. However, for acetylation sites, sequence-coupling, BE, and AAF features were crucial while only the AAF and sequence-coupling features were vital for the succinylation sites.

3.7 Web-server

To aid the experimental researches, a user-oriented web-server for predML-Site has been developed. It can be found at http://103.99.176.239/predML-Site where proper guidelines for submitting query protein sequences are provided. Users are allowed to submit query sequences either in the input box or in a batch file. For better understanding, a few protein sequences taken from iPTM-mLys [5] are included as example. In addition to that, the benchmark dataset and the training features used for constructing predML-Site will be provided upon user request.

4 Conclusion

An efficient and successful predictor predML-Site was developed with a feature integration process followed by a feature selection technique, understanding the significance of identifying multiple lysine PTM sites. Experimental outcomes demonstrate that predML-Site is highly promising compared to the existing state-of-the-art multiple lysine PTM site predictors. It is expected to become a high throughput tool for researchers for further PTM study on lysine residue. Even the experimental scientists may use this web-based tool without knowing its implementation details. Besides, a similar methodology of the proposed predictor can be used in the study of other PTMs such as C-PTM, R-PTM, and S-PTM that correspond to multi-label PTM sites at Cys, Arg and Ser residues respectively. However, predML-Site currently designed for four K-PTM types. Other PTM types with new protein sequences can be added to extend its capability in the future.

Acknowledgments

Not applicable.

References

[1] N. Saraswathy and P. Ramalingam, *Concepts and techniques in genomics and proteomics*. Elsevier, 2011.

[2] G. McDowell and A. Philpott, "New insights into the role of ubiquitylation of proteins," in *International review of cell and molecular biology*. Elsevier, 2016, vol. 325, pp. 35–88.

[3] J. D. Weissman, A. Raval, and D. S. Singer, "Assay of an intrinsic acetyltransferase activity of the transcriptional coactivator CIITA," in *Methods in enzymology*. Elsevier, 2003, vol. 370, pp. 378–386.

[4] K.-C. Chou, "Impacts of bioinformatics to medicinal chemistry," *Medicinal chemistry*, vol. 11, no. 3, pp. 218–234, 2015.

[5] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, and K.-C. Chou, "iPTM-mLys: identifying multiple lysine PTM sites and their different types," *Bioinformatics*, vol. 32, no. 20, pp. 3116–3123, 2016.

[6] R. N. Freiman and R. Tjian, "Regulating the regulators: lysine modifications make their mark," *Cell*, vol. 112, no. 1, pp. 11–17, 2003.

[7] Y. Xu and K.-C. Chou, "Recent progress in predicting posttranslational modification sites in proteins," *Current topics in medicinal chemistry*, vol. 16, no. 6, pp. 591–603, 2016.

[8] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, "iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset," *Analytical biochemistry*, vol. 497, pp. 48–56, 2016.

[9] S. C. Kim, R. Sprung, Y. Chen, Y. Xu, H. Ball, J. Pei, T. Cheng, Y. Kho, H. Xiao, L. Xiao *et al.*, "Substrate and functional diversity of lysine acetylation revealed by a proteomics survey," *Molecular cell*, vol. 23, no. 4, pp. 607–618, 2006.

[10] M. Wu, Y. Yang, H. Wang, and Y. Xu, "A deep learning method to more accurately recall known lysine acetylation sites," *BMC bioinformatics*, vol. 20, no. 1, p. 49, 2019.

[11] C. Choudhary, C. Kumar, F. Gnad, M. L. Nielsen, M. Rehman, T. C. Walther, J. V. Olsen, and M. Mann, "Lysine acetylation targets protein complexes and co-regulates major cellular functions," *Science*, vol. 325, no. 5942, pp. 834–840, 2009.

[12] H. Yu, C. Bu, Y. Liu, T. Gong, X. Liu, S. Liu, X. Peng, W. Zhang, Y. Peng, J. Yang *et al.*, "Global crotonylome reveals CDYL-regulated RPA1 crotonylation in homologous recombination–mediated DNA repair," *Science advances*, vol. 6, no. 11, p. eaay4697, 2020.

[13] H. Lv, F.-Y. Dao, Z.-X. Guan, H. Yang, Y.-W. Li, and H. Lin, "Deep-Kcr: accurate detection of lysine crotonylation sites using deep learning method," *Briefings in Bioinformatics*, 2020.

[14] B. T. Weinert, C. Schölz, S. A. Wagner, V. Iesmantavicius, D. Su, J. A. Daniel, and C. Choudhary, "Lysine succinylation is a frequently occurring modification in prokaryotes and eukaryotes and extensively overlaps with acetylation," *Cell reports*, vol. 4, no. 4, pp. 842–851, 2013.

[15] Z. Zhang, M. Tan, Z. Xie, L. Dai, Y. Chen, and Y. Zhao, "Identification of lysine succinylation as a new post-translational modification," *Nature chemical biology*, vol. 7, no. 1, p. 58, 2011.

[16] X. Zhao, Q. Ning, H. Chai, and Z. Ma, "Accurate in silico identification of protein succinylation sites using an iterative semi-supervised learning technique," *Journal of theoretical biology*, vol. 374, pp. 60–65, 2015.

[17] D. S. Johnson, W. Li, D. B. Gordon, A. Bhattacharjee, B. Curry, J. Ghosh, L. Brizuela, J. S. Carroll, M. Brown, P. Flicek *et al.*, "Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets," *Genome research*, vol. 18, no. 3, pp. 393–403, 2008.

[18] B. M. Turner, "Cellular memory and the histone code," *Cell*, vol. 111, no. 3, pp. 285–291, 2002.

[19] A. P. Snijders, M.-L. Hung, S. A. Wilson, and M. J. Dickman, "Analysis of arginine and lysine methylation utilizing peptide separations at neutral pH and electron transfer dissociation mass spectrometry," *Journal of the American Society for Mass Spectrometry*, vol. 21, no. 1, pp. 88–96, 2010.

[20] Z. Ju, J.-Z. Cao, and H. Gu, "iLM-2L: A two-level predictor for identifying protein lysine methylation sites and their methylation degrees by incorporating K-gap amino acid pairs into Chou's general PseAAC," *Journal of Theoretical Biology*, vol. 385, pp. 50–57, 2015.

[21] A. Rahman, S. Ahmed, J. Rahman, and M. A. M. Hasan, "Prediction of Formylation Sites by Incorporating Sequence Coupling into General PseAAC," in *2020 IEEE Region 10 Symposium (TENSYMP)*. IEEE, 2020, pp. 921–924.

[22] S. Ahmed, A. Rahman, M. A. M. Hasan, M. K. B. Islam, J. Rahman, and S. Ahmad, "predPhogly-Site: Predicting phosphoglycerylation sites by incorporating probabilistic sequence-coupling information into PseAAC and addressing data imbalance," *Plos one*, vol. 16, no. 4, p. e0249396, 2021.

[23] Z. Ju and J.-J. He, "Prediction of lysine glutarylation sites by maximum relevance minimum redundancy feature selection," *Analytical biochemistry*, vol. 550, pp. 1–7, 2018.

[24] M. A. Hasan, M. K. Ben Islam, J. Rahman, and S. Ahmad, "Citrullination Site Prediction by Incorporating Sequence Coupled Effects into PseAAC and Resolving Data Imbalance Issue," *Current Bioinformatics*, vol. 15, no. 3, pp. 235–245, 2020.

[25] M. A. M. Hasan and S. Ahmad, "mLysPTMpred: Multiple Lysine PTM Site Prediction Using Combination of SVM with Resolving Data Imbalance Issue," *Natural Science*, vol. 10, no. 9, pp. 370–384, 2018.

[26] K.-C. Chou, "A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins." *Journal of Biological Chemistry*, vol. 268, no. 23, pp. 16 938–16 948, 1993.

[27] K.-C. Chou, "Prediction of human immunodeficiency virus protease cleavage sites in proteins," *Analytical biochemistry*, vol. 233, no. 1, pp. 1–14, 1996.

[28] W.-Z. Lin, J.-A. Fang, X. Xiao, and K.-C. Chou, "iDNA-Prot: identification of DNA binding proteins using random forest with grey model," *PloS one*, vol. 6, no. 9, 2011.

[29] Z. Ju and S.-Y. Wang, "Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou's general pseudo amino acid composition," *Gene*, vol. 664, pp. 78–83, 2018.

[30] J. Zhe and S.-Y. Wang, "Prediction of 2-hydroxyisobutyrylation sites by integrating multiple sequence features with ensemble support vector machine," *Computational Biology and Chemistry*, p. 107280, 2020.

[31] D. Chen, Z. Liu, X. Ma, and D. Hua, "Selecting genes by test statistics," *BioMed Research International*, vol. 2005, no. 2, pp. 132–138, 2005.

[32] Z. Ju and S.-Y. Wang, "iLys-Khib: Identify lysine 2-Hydroxyisobutyrylation sites using mRMR feature selection and fuzzy SVM algorithm," *Chemometrics and Intelligent Laboratory Systems*, vol. 191, pp. 96–102, 2019.

[33] K. Veropoulos, C. Campbell, N. Cristianini *et al.*, "Controlling the sensitivity of support vector machines," in *Proceedings of the international joint conference on AI*, vol. 55, 1999, p. 60.

[34] M. A. M. Hasan, S. Ahmad, and M. K. I. Molla, "iMulti-HumPhos: a multi-label classifier for identifying human phosphorylated

proteins using multiple kernel learning based support vector machines," *Molecular BioSystems*, vol. 13, no. 8, pp. 1608–1618, 2017.

[35] Z. Ju and J.-J. He, "Prediction of lysine propionylation sites using biased SVM and incorporating four different sequence features into Chou's PseAAC," *Journal of Molecular Graphics and Modelling*, vol. 76, pp. 356–363, 2017.

[36] M. A. M. Hasan, J. Li, S. Ahmad, and M. K. I. Molla, "predCarsite: Carbonylation sites prediction in proteins using support vector machine with resolving data imbalanced issue," *Analytical biochemistry*, vol. 525, pp. 107–113, 2017.

[37] W. Bao, B. Yang, D.-S. Huang, D. Wang, Q. Liu, Y.-H. Chen, and R. Bao, "IMKPse: Identification of protein malonylation sites by the key features into general PseAAC," *IEEE Access*, vol. 7, pp. 54 073–54 083, 2019.

[38] W.-R. Qiu, X. Xiao, W.-Z. Lin, and K.-C. Chou, "iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach," *BioMed research international*, vol. 2014, 2014.

[39] U. Consortium, "UniProt: a worldwide hub of protein knowledge," *Nucleic acids research*, vol. 47, no. D1, pp. D506–D515, 2019.

[40] K.-C. Chou, "Prediction of signal peptides using scaled window," *peptides*, vol. 22, no. 12, pp. 1973–1979, 2001.

[41] K.-C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, 2005.

[42] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of theoretical biology*, vol. 273, no. 1, pp. 236–247, 2011.

[43] J.-L. Min, X. Xiao, and K.-C. Chou, "iEzy-Drug: A web server for identifying the interaction between enzymes and drugs in cellular networking," *BioMed research international*, vol. 2013, 2013.

[44] P. Du, X. Wang, C. Xu, and Y. Gao, "PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions," *Analytical biochemistry*, vol. 425, no. 2, pp. 117–119, 2012.

[45] Y. Xu, X. Wen, L.-S. Wen, L.-Y. Wu, N.-Y. Deng, and K.-C. Chou, "iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition," *PloS one*, vol. 9, no. 8, p. e105018, 2014.

[46] J. Jia, L. Zhang, Z. Liu, X. Xiao, and K.-C. Chou, "pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC," *Bioinformatics*, vol. 32, no. 20, pp. 3133–3141, 2016.

[47] C.-W. Tung, "Prediction of pupylation sites using the composition of k-spaced amino acid pairs," *Journal of theoretical biology*, vol. 336, pp. 11–17, 2013.

[48] Z. Ju and J.-Z. Cao, "Prediction of protein N-formylation using the composition of k-spaced amino acid pairs," *Analytical Biochemistry*, vol. 534, pp. 40–45, 2017.

[49] Z. Ju and S.-Y. Wang, "Prediction of lysine formylation sites using the composition of k-spaced amino acid pairs via Chou's 5-steps rule and general pseudo components," *Genomics*, vol. 112, no. 1, pp. 859–866, 2020.

[50] Z. Ju, J.-Z. Cao, and H. Gu, "Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou's general PseAAC," *Journal of Theoretical Biology*, vol. 397, pp. 145–150, 2016.

[51] W. R. Atchley, J. Zhao, A. D. Fernandes, and T. Drüke, "Solving the protein sequence metric problem," *Proceedings of the National Academy of Sciences*, vol. 102, no. 18, pp. 6395–6400, 2005.

[52] A. Torkamani and N. J. Schork, "Accurate prediction of deleterious protein kinase polymorphisms," *Bioinformatics*, vol. 23, no. 21, pp. 2918–2925, 2007.

[53] M. H. Kutner, C. J. Nachtsheim, J. Neter, W. Li *et al.*, *Applied linear statistical models*. McGraw-Hill Irwin New York, 2005, vol. 5.

[54] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[55] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[56] X. Ruan, D. Zhou, R. Nie, and Y. Guo, "Predictions of Apoptosis Proteins by Integrating Different Features Based on Improving Pseudo-Position-Specific Scoring Matrix," *BioMed Research International*, vol. 2020, 2020.

[57] Y. Ma, Z. Yu, G. Han, J. Li, and V. Anh, "Identification of pre-microRNAs by characterizing their sequence order evolution information and secondary structure graphs," *BMC bioinformatics*, vol. 19, no. 19, p. 521, 2018.

[58] R. Batuwita and V. Palade, "Efficient resampling methods for training support vector machines with imbalanced datasets," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2010, pp. 1–8.

[59] A. Dehzangi, Y. López, S. P. Lal, G. Taherzadeh, A. Sattar, T. Tsunoda, and A. Sharma, "Improving succinylation prediction accuracy by incorporating the secondary structure via helix, strand and coil, and evolutionary information from profile bigrams," *PloS one*, vol. 13, no. 2, p. e0191900, 2018.

[60] W. Qiu, C. Xu, X. Xiao, and D. Xu, "Computational Prediction of Ubiquitination Proteins Using Evolutionary Profiles and Functional Domain Annotation," *Current Genomics*, vol. 20, no. 5, pp. 389–399, 2019.

[61] C. Jia, M. Zhang, C. Fan, F. Li, and J. Song, "Formator: predicting lysine formylation sites based on the most distant undersampling and safe-level synthetic minority oversampling," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.

[62] Q. Ning, Z. Ma, and X. Zhao, "dForml (KNN)-PseAAC: Detecting formylation sites from protein sequences using K-nearest neighbor algorithm via Chou's 5-step rule and pseudo components," *Journal of theoretical biology*, vol. 470, pp. 43–49, 2019.

[63] Y. Xu, Y.-X. Ding, J. Ding, L.-Y. Wu, and N.-Y. Deng, "Phogly–PseAAC: prediction of lysine phosphoglycerylation in proteins incorporating with position-specific propensity," *Journal of Theoretical Biology*, vol. 379, pp. 10–15, 2015.

[64] L.-M. Liu, Y. Xu, and K.-C. Chou, "iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC," *Medicinal Chemistry*, vol. 13, no. 6, pp. 552–559, 2017.

[65] A. Chandra, A. Sharma, A. Dehzangi, D. Shigemizu, and T. Tsunoda, "Bigram-PGK: phosphoglycerylation prediction using the technique of bigram probabilities of position specific scoring matrix," *BMC molecular and cell biology*, vol. 20, no. 2, pp. 1–9, 2019.

[66] J. Yu, S. Shi, F. Zhang, G. Chen, and M. Cao, "PredGly: predicting lysine glycation sites for Homo sapiens based on XGboost feature optimization," *Bioinformatics*, vol. 35, no. 16, pp. 2749–2756, 2019.

[67] V. B. Semwal, J. Singha, P. K. Sharma, A. Chauhan, and B. Behera, "An optimized feature selection technique based on incremental feature analysis for bio-metric gait data classification," *Multimedia tools and applications*, vol. 76, no. 22, pp. 24 457–24 475, 2017.

[68] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[69] M. Jiang and J.-Z. Cao, "Positive-Unlabeled learning for pupylation sites prediction," *BioMed research international*, vol. 2016, 2016.

[70] C.-W. Hsu, C.-C. Chang, C.-J. Lin *et al.*, "A practical guide to support vector classification," 2003.

[71] M. A. M. Hasan, S. Ahmad, and M. K. I. Molla, "Protein subcellular localization prediction using multiple kernel learning based support vector machine," *Molecular BioSystems*, vol. 13, no. 4, pp. 785–795, 2017.

[72] K.-C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystems," *Molecular Biosystems*, vol. 9, no. 6, pp. 1092–1100, 2013.

[73] X. Cheng, S.-G. Zhao, X. Xiao, and K.-C. Chou, "iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals," *Bioinformatics*, vol. 33, no. 3, pp. 341–346, 2017.

[74] Z. Ju and S.-Y. Wang, "Computational Identification of Lysine Glutarylation Sites Using Positive-Unlabeled Learning," *Current Genomics*, vol. 21, no. 3, pp. 204–211, 2020.

[75] Y.-Z. Chen, Y.-R. Tang, Z.-Y. Sheng, and Z. Zhang, "Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs," *BMC bioinformatics*, vol. 9, no. 1, p. 101, 2008.

Sabit Ahmed Sabit Ahmed has completed his B.Sc. in Computer Science and Engineering from Rajshahi University of Engineering and Technology, Bangladesh. He has done his undergraduate research work with the Machine Learning Research Group, Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Bangladesh. His research interests are computational proteomics and bioinformatics.

Shamim Ahmad Shamim Ahmad completed his B.Sc. and M.Sc. in Computer Science and Engineering from the University of Rajshahi, Bangladesh. He also received his Ph.D. in 2005 from Chubu University, Kasugai, Aichi, Japan. He currently works as a professor in the Department of Computer Science and Engineering, University of Rajshahi, Bangladesh. His research interests include bioinformatics, embedded systems, digital signal processing.

Afrida Rahman Afrida Rahman has completed her B.Sc. in Computer Science and Engineering from Rajshahi University of Engineering and Technology, Bangladesh. She has done her undergraduate research work with the Machine Learning Research Group, Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Bangladesh. Her research interests are computational proteomics and bioinformatics.

M.A.M. Hasan Md. Al Mehedi Hasan completed his B.Sc. and M.Sc. in Computer Science and Engineering from the University of Rajshahi, Bangladesh. He also received his Ph.D. in the field of Bioinformatics and Computational Biology in 2017 from the University of Rajshahi, Bangladesh. He is a founding member of the Machine Learning Research Group in the Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Bangladesh. His research interests include bioinformatics, artificial intelligence, pattern recognition, image processing, machine learning, computer vision, data mining, probabilistic and statistical inference, operating system, computer network, and security.

Julia Rahman Julia Rahman is an Assistant Professor and founding member of the Machine Learning Research Group in the Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Bangladesh. She is currently a Ph.D. candidate in the School of Information and Communication Technology, Griffith University, Australia. Her research interests include bioinformatics, machine learning, and artificial intelligence.

Md Khaled Ben Islam Md Khaled Ben Islam is an Assistant Professor in the Department of Computer Science and Engineering, Pabna University of Science and Technology, Bangladesh. He is also doing collaborative research with the Machine Learning Research Group, Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Bangladesh. His research interests include bioinformatics, machine learning, natural language processing, and IoT.

Prediction of Formylation Sites by Incorporating Sequence Coupling into General PseAAC

Conference Paper · June 2020

DOI: 10.1109/TENSYMP50017.2020.9230745

4 authors:

Afrida Rahman
Rajshahi University of Engineering & Technology
7 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Sabit Ahmed
Rajshahi University of Engineering & Technology
7 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Julia Rahman
Rajshahi University of Engineering & Technology
27 PUBLICATIONS **87** CITATIONS
SEE PROFILE

Md. Al Mehedi Hasan
Rajshahi University of Engineering & Technology
117 PUBLICATIONS **713** CITATIONS
SEE PROFILE

Prediction of Formylation Sites by Incorporating Sequence Coupling into General PseAAC

Afrida Rahman
Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
Email: afrida.r.samma@gmail.com

Sabit Ahmed
Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
Email: sabit.a.sirat@gmail.com

Julia Rahman
Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
Email: juliacse06@gmail.com

Md. Al Mehedi Hasan
Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
Email: mehedi_ru@yahoo.com

***Abstract*—Post-translational modification (PTM) introduces to the biochemical mechanisms in active organisms, which has been observed in increasing functional plurality of proteins as well as operating cellular processes. Lysine formylation is such a PTM site newly invented that has gained much appreciation but a poorly destined role in chromatin composition measurement and gene expression. Our motive is to identify the annotated formyllysine sites more accurately reducing the time complexity. In this study, a novel computational tool, termed as Formyl_Pred, has been developed to predict formyllysine sites. This technique has practically implemented the sequence-coupling information surrounding amino acids of lysine residues along with optimizing slantwise training formyllysine dataset for prediction characteristics enhancement. The performance of Formyl_Pred has been measured from the average of exact 10 runs of the 10-fold cross-validation test obeying the existing methods. After implementation, Formyl_Pred has acquired an accuracy of 99.57% with an area under curve (AUC) of 0.9949. These performances indicate that Formyl_Pred remarkably outperforms the current existing prediction tool.**

***Index Terms*—Formylation Sites Prediction, Post-translational modification, Sequence-coupling Model, General PseAAC, Data Imbalance Issue, Support Vector Machine.**

I. Introduction

Lysine formylation is a irresolute form of protein post-translational modification observed mainly on histones and other nucleoproteins that refers to the process of formyl functional group (-CHO) affluently attaching substrate lysine residues to accomplish the modification through peptide bonds [1, 2]. It is revealed that lysine formylation suppresses the nucleosome cluster as it appears along with the DNA combination. It can arise from oxidative impairment in the cell [3]. Therefore, developing the analysing process of formylation substance layer proteins associated with distinct formylation sites and proper understanding of the molecular mechanism of biological function might be effective to elicit the radical molecular reasons of the corresponding disease. Consequently, the experimental method for accurately detecting formylation sites is expensive and time-consuming. This becomes even more difficult for those proteins which are unknown. Because of the exponential growth of sequenced proteins in recent years, it is necessary to expand rapid and constructive computational methods [4].

According to the previous study, the first predictor of lysine formylation sites was LFPred [5], but there was no online server. And this issue was solved by another predictor CK-SAAP_FormSite developed by Zhe Ju, Shi-Yun Wang [6]. This predictor was constructed by combining the composition of k-spaced amino acid pairs (CKSAAP) as the feature encoding scheme with a biased support vector machine as a classifier and it achieved quite better results. Further processing is needed to find a more efficient encoding method for upgrading the prediction performance of identifying lysine formylation sites. Moreover, the training dataset is quite immitigable. For handling these complications, we have proposed a new approach termed as Formyl_Pred to predict formylation sites using a vectorized sequence-coupling model into chou's general pseudo amino acid composition (PseAAC) and a biased support vector machine (SVM) algorithm. The biased SVM [7] was capable of handling the imbalance between the number of formylation and non-formylation sites. The explorative results signified that the area under curve (AUC) of Formyl_Pred was 0.9949 evaluated by 10-fold cross-validation which achieved superior performance compared to the existing methods.

II. Materials and methods

A. Dataset

The formylation dataset was collected from the PLMD database [8], which comprised 97 formylated proteins as described by Ju and wang [6]. To alleviate redundancy, the pairwise sequence identities of proteins larger than 40% were eliminated by the CD-HIT program [9]. Then the non-redundant dataset was attained composing 182 experimentally

978-1-7281-7366-5/20/$31.00 ©2020 IEEE

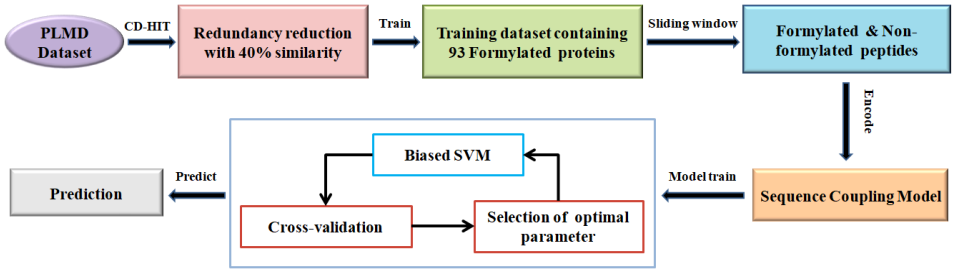

Fig. 1. The system flowchart of Formyl_Pred

substantiated annotated formyllysine sites and the other 1637 non-formyllysine sites along with 93 formylated proteins. To encipher the lysine residue K, the sliding window method having $(2\xi+1)$ sized tuple $(\xi = 11)$ was used in the training dataset to illustrate each site modification as a peptide segment. So, the window size for this method was selected as 23(11 upstream and 11 downstream amino acid residues) and dummy residue was 'X' according to the previous studies [6]. The system flowchart of Formyl_Pred is shown in Fig. 1.

TABLE I
SUMMARY OF DATASET

Sequence identity thresold	Formylated proteins	Formyllysine sites	Non-formyllysine sites
40%	93	182	1637

B. *Feature Construction*

In this study, the sequence-coupling model [10] has been incorporated into Chou's general PseAAC to conserve the sequence-coupled information and extract features from the peptide segment. Based on the Chou's conception, a lysine residue centered peptide can be represented by:

$$\Theta_\zeta(K) = R_{-\zeta}R_{-(\zeta-1)}...R_{-2}R_{-1}KR_1R_2...R_{+(\zeta-1)}R_{+\zeta} \quad (1)$$

In this equation, $R_{-\zeta}$ and $R_{+\zeta}$ denote the ζ-th leftward and rightward amino acid residues, respectively, while ζ being an integer and 'K'(center) indicating "lysine" [11]. Again, the peptide sequence $\Theta_\zeta(K)$ is categorized into two types: $\Theta^+_\zeta(K)$, $\Theta^-_\zeta(K)$ are true formylated peptide and false formylated peptide with a lysine residue at its center [11]. When ζ=11, the corresponding peptide segment contains $(2\zeta+1) = 23$ amino acid residues. According to Chou's general PseAAC [10], the peptide sample in (1), can be expressed as:

$$\Theta(K) = \Theta^+(K) - \Theta^-(K) \quad (2)$$

where

$$\Theta^+(K) = \begin{bmatrix} \Theta^+_{-11}(R_{-11}|R_{-10}) \\ \Theta^+_{-10}(R_{-10}|R_{-9}) \\ \vdots \\ \Theta^+_{-1}(R_{-1}) \\ \Theta^+_{+1}(R_{+1}) \\ \vdots \\ \Theta^+_{+10}(R_{+10}|R_{+9}) \\ \Theta^+_{+11}(R_{+11}|R_{+10}) \end{bmatrix} \quad (3)$$

where $\Theta^+_{-11}(R_{-11}|R_{-10})$ is the contingent probability of amino acid R_{-11} at the leftmost position given that its adjacent right member is R_{-10} and so forth. In contrast, only $\Theta^+_{-1}(R_{-1})$ and $\Theta^+_{+1}(R_{+1})$ are of non-contingent probability as K is the adjoining member of both amino acids at position R_{-1} and R_{+1}. The probability values can be extracted from the set of formylated peptides without any difficulty. Accordingly, $\Theta^-(K)$ in (2), and its probability components can be deduced from the non-formylated peptide set in the same way.

C. *Biased support vector machine*

Nowadays, the support vector machine (SVM) is considered as a widely used supervised learning technique for classification problems[12]. Apart from that, the structural risk minimization involves a biasing problem where the majority class[13] influences the classification weight. As the skewed training formylated peptide set (i.e. the ratio between positive and negative peptides is approximately 1:9) affects the classification directly, according to [7], different penalty costs C^+, and C^- were assigned. Therefore, a cost-sensitive SVM was applied, which can be formulated as follows:

$$min^{\frac{1}{2}}_{w,\xi}\|w\|^2 + C^+\sum_{k=1}^{q}\xi_k + C^-\sum_{k=q+1}^{n}\xi_k \quad (4)$$

(Subject to: $Y_k(w.\varphi(X_k)+a) \geq 1-\xi_k$ for all, $k = 1, 2, .., n$)

where the training set is denoted by $\{(X_k, Y_k), k = 1, 2, ..., n\}$ and first q examples (i.e. $Y_k = 1, k = 1, 2, ..., q$) are assumed as the positive examples while the rest are assumed as the negative examples (i.e. $Y_k = -1, k = q+1, q+2, ..., n$). The non-linear feature mapping and slack variables are denoted by $\varphi(X)$ and $\xi_k(k = 1, 2, ..., n)$ respectively [6]. In our

TABLE II
PERFORMANCE COMPARISON OF BIASED AND STANDARD SUPPORT VECTOR MACHINE.

Approach	Sn(%)	Sp(%)	Precision(%)	ACC(%)	MCC(%)	AUC(%)
Standard SVM	97.64 (±0.37)	99.81 (±0.05)	98.29 (±0.40)	99.59 (±0.06)	97.74 (±0.33)	98.72 (±0.19)
Biased SVM	99.40 (±0.16)	98.58 (±0.02)	96.38 (±0.21)	99.57 (±0.03)	97.64 (±0.16)	**99.49** (±0.09)

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED METHOD WITH THE EXISTING ONE ON THE PLMD DATASET.

Method	Sn(%)	Sp(%)	Precision(%)	ACC(%)	MCC(%)	AUC(%)
CKSAAP_FormSite	71.04 (±1.51)	77.10 (±0.31)	—	76.50 (±0.31)	32.28 (±0.99)	82.34 (±0.34)
Formyl_Pred	99.40 (±0.16)	98.58 (±0.02)	96.38 (±0.21)	99.57 (±0.03)	97.64 (±0.16)	**99.49** (±0.09)

Here, '—' represents the corresponding value is not available in CKSAAP_FormSite.

experiments, the Gaussian kernel function is described as: $\Upsilon(X_k, X_j) = \varphi(X_k)^T \varphi(X_j) = exp(-\frac{\|X_k - X_j\|^2}{\partial})$, where ∂ is the width of the function. To handle the class imbalance problem of formylation and non-formylation sites, two misclassification costs $C^+ = \frac{C*n}{2*q}$ and $C^- = \frac{C*n}{2*(n-q)}$ are assigned for formylated sites and non-formylated sites respectively.

D. *Cross-validation of Proposed Model*

It is always desirable to obtain a suitable model that can be capable of producing high performance, and the grid-search technique was used to select the values of parameters and find the best model for SVM. Here, we implemented 10-fold cross-validation test to train and evaluate our model. To check validity and stability, we iteratively executed the 10-fold cross-validation ten times (i.e. 100 iterations in total). For each time, the formylation dataset was divided into ten disjoint sets by randomization, and afterwards, we obtained the average results of all performance metrics. Based on the highest AUC value achieved, Formyl_Pred was constructed. The parameters C and γ were selected from $\{2^0, 2^1, 2^2, ..., 2^7\}$ and $\{2^{-1}, 2^{-2}, 2^{-3}, ..., 2^{-7}\}$ respectively.

E. *Prediction Metrics*

To justify the prediction quality of Formyl_Pred, we have utilized four intuitive metrics, such as accuracy (ACC),sensitivity (Sn), specificity (Sp) and Matthew's Correlation Coefficient (MCC). However, we have included an additional metric precision to evaluate the integrity of our model. To calculate the specified metrics, four parameters are needed: true positive(TP),false positive(FP),true negative(TN),false negative(FN).The equations of these performance metrics are given below :

$$Sn = \frac{TP}{TP+FN} \tag{5}$$

$$Sp = \frac{TN}{TN+FP} \tag{6}$$

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

$$MCC = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{9}$$

Eventually, we have considered the area under the ROC curve (AUC) to check the robustness of our system.

III. RESULTS AND DISCUSSIONS

A. *Performance of Formyl_Pred*

The maximum AUC value of 10-fold cross-validation(10 times) using the grid-search approach assured the optimal parameters (i.e., $C = 2^2$ and $\gamma = 2^{-2}$) for our proposed model. As demonstrated in Table II, Formyl_pred acquired an MCC value of 0.9764 after applying the sequence-coupled information into the general PseAAC. And other corresponding predictions of Sp, Sn, precision and ACC values are 99.58%, 99.40%, 96.38% and 99.57% respectively. To scrutinize the overall process, we measured the results of biased and standard support vector machines and found that the first one gained the highest AUC value of 0.9949 while the last one obtained an AUC of 0.9872. Therefore, it can be concluded that the biased support vector machine acquired a noticeable performance.

B. *Comparison with CKSAAP_FormSite predictor*

Currently, there are three predictors, namely, LFPred [5], CKSAAP_FormSite [6] and Formator[14] for formylation site prediction, but only CKSAAP_FormSite employed the same dataset derived from PLMD. So, it is equitable to compare the proposed predictor Formyl_Pred with CKSAAP_FormSite predictor. As shown in Table III, the proposed predictor Formyl_Pred achieved much better performance than CKSAAP_FormSite in terms of accuracy, MCC, sensitivity, specificity as well as AUC. For instance, the corresponding value of MCC and AUC significantly improved by 0.6536 and 0.1715 compared to CKSAAP_FormSite. Interestingly, according to Table III, Formyl_Pred has attained a further improvement in

other important statistics, i.e. accuracy, sensitivity and specificity by 23.07%, 28.36% and 21.48% respectively. Furthermore, our proposed method obtained a higher precision rate of 96.38% which was not available in CKSAAP_FormSite. From the result presented in Table III, it is evident that the prediction quality of Formyl_Pred is much better and it outperforms existing predictor CKSAAP_FormSite. For better understanding, the comparison between CKSAAP_FormSite and Formyl_Pred is shown in Fig. 2. Our proposed system has achieved such an outstanding performance as we have successfully trained a biased SVM with the coupling effects, which considers the conditional probability among the amino acid residues.

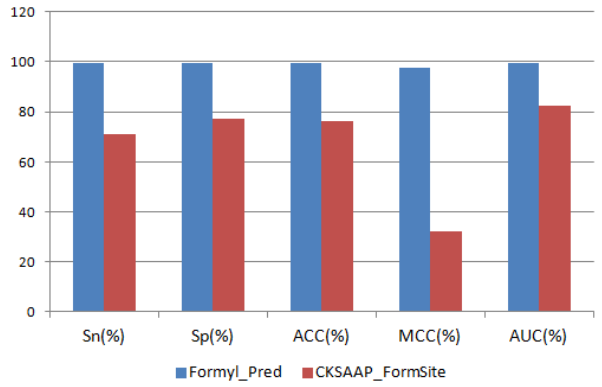

Fig. 2. Comparison between CKSAAP_FormSite and Formyl_Pred

IV. Conclusion

The prediction of formyllysine sites is obligatory for attaining a better perception of biological systems. Considering the time and cost measurement of computational methods, a novel bioinformatics approach named Formyl_Pred was developed to predict lysine formylation sites by using a vectorized sequence-coupling model with SVM classifier. By adopting sequence-coupling effect with misclassification cost adjustment, Formyl_Pred acquired extraordinarily higher success rates compared to the existing formyllysine site predictors. In the cross-validation test, it obtained a significant improvement in AUC as well as in other crucial metrics (accuracy 99.57%, precision 96.38% with an MCC of 0.9764). We have ensured that Formyl_Pred will become a proficient throughput tool or accessorial tool for researchers in the future study of protein formylation. For further improvement, the independent test is needed to make it more appropriate and reasonable. Moreover, multiple types of post-translational modification, along with updated data, will be considered simultaneously in the future.

References

[1] S. Lee, "Post-translational modification of proteins in toxicological research: focus on lysine acylation," *Toxicological research*, vol. 29, no. 2, pp. 81–86, 2013.

[2] T. Wang, Q. Zhou, F. Li, Y. Yu, X. Yin, and J. Wang, "Genetic Incorporation of Nε-Formyllysine, a New Histone Post-translational Modification," *ChemBioChem*, vol. 16, no. 10, pp. 1440–1442, 2015.

[3] T. Jiang, X. Zhou, K. Taghizadeh, M. Dong, and P. C. Dedon, "N-formylation of lysine in histone proteins as a secondary modification arising from oxidative dna damage," *Proceedings of the National Academy of Sciences*, vol. 104, no. 1, pp. 60–65, 2007.

[4] W.-R. Qiu, S.-Y. Jiang, B.-Q. Sun, X. Xiao, X. Cheng, and K.-C. Chou, "iRNA-2methyl: identify RNA 2'-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier," *Medicinal Chemistry*, vol. 13, no. 8, pp. 734–743, 2017.

[5] Q. Ning, Z. Ma, and X. Zhao, "dForml (KNN)-PseAAC: Detecting formylation sites from protein sequences using K-nearest neighbor algorithm via Chou's 5-step rule and pseudo components," *Journal of theoretical biology*, vol. 470, pp. 43–49, 2019.

[6] Z. Ju and S.-Y. Wang, "Prediction of lysine formylation sites using the composition of k-spaced amino acid pairs via Chou's 5-steps rule and general pseudo components," *Genomics*, vol. 112, no. 1, pp. 859–866, 2020.

[7] K. Veropoulos, C. Campbell, N. Cristianini *et al.*, "Controlling the sensitivity of support vector machines," in *Proceedings of the international joint conference on AI*, vol. 55, 1999, p. 60.

[8] H. Xu, J. Zhou, S. Lin, W. Deng, Y. Zhang, and Y. Xue, "PLMD: An updated data resource of protein lysine modifications," *Journal of Genetics and Genomics*, vol. 44, no. 5, pp. 243–250, 2017.

[9] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[10] K.-C. Chou, "Prediction of human immunodeficiency virus protease cleavage sites in proteins," *Analytical biochemistry*, vol. 233, no. 1, pp. 1–14, 1996.

[11] M. A. M. Hasan and S. Ahmad, "mLysPTMpred: Multiple Lysine PTM Site Prediction Using Combination of SVM with Resolving Data Imbalance Issue," *Natural Science*, vol. 10, no. 9, pp. 370–384, 2018.

[12] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[13] L. Zhang, B. Tan, T. Liu, and X. Sun, "Classification study for the imbalanced data based on Biased-SVM and the modified over-sampling algorithm," in *Journal of Physics: Conference Series*, vol. 1237, no. 2. IOP Publishing, 2019, p. 022052.

[14] C. Jia, M. Zhang, C. Fan, F. Li, and J. Song, "Formator: predicting lysine formylation sites based on the most distant undersampling and safe-level synthetic minority oversampling," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.