KPhogly-RAM:Prediction of Phosphoglycerylation Sites Using Residue Adjacency Matrix

Audity Ghosh
Computer Science Engineering
Rajshahi University of Engineering
& Technology, Rajshahi, Bangladesh
audityghosh1999@gmail.com

Farhana Amin
Computer Science & Engineering
Rajshahi University of Engineering
& Technology, Rajshahi, Bangladesh
400fjemin@gmail.com

Mansura Naznine
Computer Science & Engineering
Rajshahi University of Engineering
& Technology, Rajshahi, Bangladesh
naznine31@gmail.com

***Abstract*—Post-translational modification (PTM) is a type of covalent alteration that occurs after the biosynthesis process and is crucial for the study of cell biology. Even though a significant number of proteins have been sequenced, phosphoglycerylation recognition remains a major challenge due to factors such as cost, time, and inefficiency in experimental efforts. Lysine phosphoglycerylation is a newly discovered reversible type of PTM that changes glycolyticenzyme activity and is linked to different diseases. In this study, we developed a new computational approach KPhogly-RAM based on sequence-based information about amino acid residues to identify lysin phosphoglycerylation sites in the protein to understand the functionality and causality of phosphoglycerylation sites. We extracted the residue adjacency matrix for each lysine residue in the protein sequences from the data set and encoded it into fused feature vectors before applying recursive feature elimination to enhance prediction quality. It obtained an accuracy score of 0.8877 in repeated stratified tenfold cross-validation, surpassing most other existing predictors.**

***Index Terms*—phosphoglycerylation,residue adjacency matrix,amino acid composition,tripeptide composition,classifier.**

I. Introduction

Post-translational modification(PTM) is the process that occurs after the translation stage and includes the covalent insertion of particular functional groups in a protein. These alterations have an immense influence on biological processes and proteomics analyses and, they are also accountable for a variety of illnesses. As a result, precise identification and knowledge of PTM locations are essential for fundamental research in disease diagnosis and prevention.

Among the 20 standard component amino acid residues of cellular proteins, modifications at the lysine residue (K) are commonly referred to as lysine PTM or K-PTM, and lysine is one of the most highly reformed residues. According to research, the most prevalent covalent modifications of lysine residues include acetyl, glycosyl, methyl, succinyl, pupyl, and others. Acetylation, phosphoglycerylation, glycation, methylation, butyrylation, succinylation, biotinylation, and other K-PTMs aid in these covalent modifications. Lysine phosphoglycerylation is a reversible post-translational modification discovered for the first time in mouse liver and human cells.3-phosphoglyceryl-lysine (pgK) is generated when the main glycolytic intermediate (1,3-BPG) interacts with particular lysine residues. 3-phosphoglyceryl-lysine inhibits glycolytic enzymes and accumulates in cells exposed to high levels of glucose, resulting in the accumulation and diversion of glycolytic intermediates to other metabolic pathways. Because of the significant importance of phosphoglycerylation biologically, identifying phosphoglycerylation sites with better efficiency is a must.

Computational methods have outperformed the conventional method for distinguishing between modified and unmodified sites. This alteration has been made in order to prevent unproductive laboratory experimentation, such as mass spectrometry, as well as to save money, time, and resources.

Phogly-PseAAC, one of the earliest predictors, used a feature set based on a pseudo amino acid composition and trained a k-nearest neighbors (knn) based predictor. [1].One of the predictors created was CKSAAP-Phoglysite, which trained a fuzzy support vector machine (SVM) and employed the composition of k-spaced amino acid pairs (CKSAAP) for feature development. [2].PhoglyPred, like the CKSAAP Phoglysite predictor, uses protein sequence data to train an SVM classifier using properties such as increased k-mer diversity, position-specific propensity of k-space dipeptide, and modified composition of k-space amino acid pairs with chosen physiochemical attributes [3].

PhoglyStruct [4], EvolStruct-Phogly [5], and Bigram-PGK [6] are the three new techniques for predicting protein phosphoglycerylation sites.Bigram-PGK, the most recently produced predictor, used SVM with evolutionary information from the sequences to increase performance.In addition, five-fold cross-validation was used to evaluate the efficacy of the presented models.

For segment sizes of seven upstream and seven downstream lysine residues, the iPGK-PseAAC classifier considers four tiers of amino acid pairwise couplings. The product is a 50-dimensional vector that contains the frequency of occurrence of each pair and was used to train an SVM classifier. Furthermore, the research was carried out using a data set that had 106 positive (phosphoglycerylation) and 1408 negative locations

979-8-3503-1156-3/22/$31.00 ©2022 IEEE

(non-phosphoglycerylation) [7].

In our work, residue adjacency matrices constructed from the lysin residues were transformed into a fused feature vector using amino acid encoding and a tripeptide composition mentioned in II-C. A synthetic minority over-sampling technique was implemented to handle the imbalanced data set. Recursive feature elimination was performed on the fused vector using gradient boosting classifier stated in II-F. Later our model was trained with different classifiers using repeated ten-fold cross-validation and stratified ten-fold cross-validation with performance measured for each approach. LightGBM using repeated stratified ten-fold outperformed the existing models with 0.8877 accuracy.Also a comparison has been drawn with other models using six-fold cross validation. The whole process has been illustrated in figure 1.

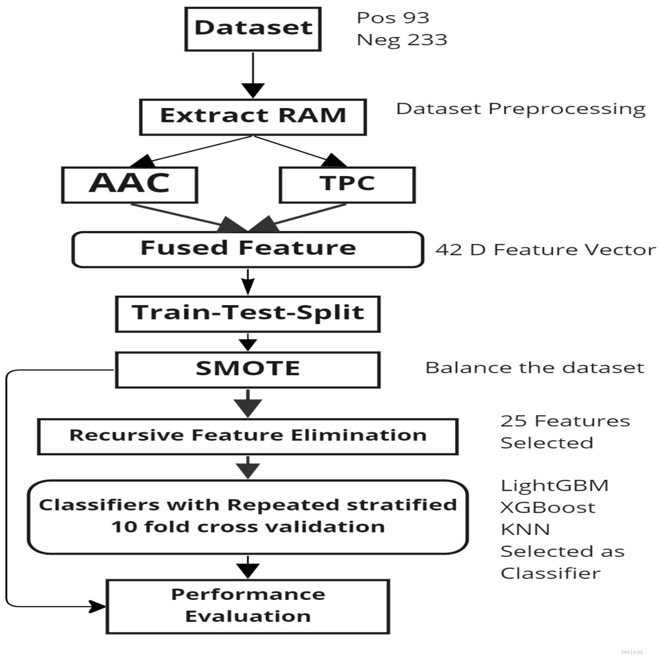


Fig. 1. An overview of KPhogly-RAM for lysin phosphoglycerylation site prediction

II. Materials and Methods

A. Dataset

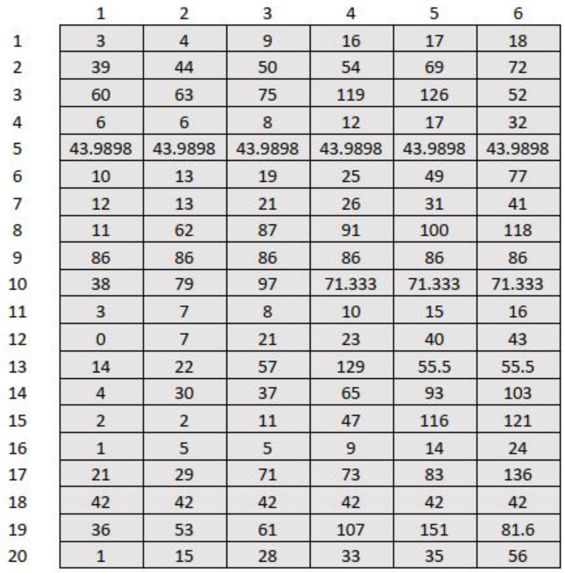
The benchmark dataset utilized in this study was taken from Protein Lysine Modification Database (PLMD) [8]. Chandra et al. [9] used the Cd-hit tool to eliminate sequences with 40% or more sequential similarity from the resulting PLMD dataset, in their work. The resultant dataset contained 91 sequences, from which 3360 lysine residues were extracted with 111 phosphoglycerylation sites and 3249 non-phosphoglycerylation sites,yielding a 1:29 imbalance ratio.To reduce bias during the classification step, redundant instances were deleted from the negative dataset using the k-nearest neighbor (KNN) approach. The Euclidean distance between each sample in the dataset was determined to eliminate the redundant negative samples using the k-nearest neighbors cleaning method.By dividing the number of negative samples by the number of positive samples, the initial number of neighbors, k, was calculated.So to eliminate the imbalance in the classes, the original value of k was 29. Based on Euclidean distances, the goal was to eliminate a negative instance when one of its 29 closest neighbors is a positive instance. The class imbalance remained high even after this filtering step with k=29. With a k value of 69, the number of negative samples was reduced from 3249 to 337. Finally, with k=101, the number of negative samples was decreased to 233. After reducing redundancy, the number of positive samples reduced to to 93, with an imbalance ratio of nearly 1:2 [4], [5].

B. Dataset Preprocessing

RAMseq (Residue Adjacency Matrix from Sequence Data) is an abbreviation for RAM calculated directly from raw protein sequences. A 20 by n matrix is formed by the distance between the target lysine residue and the first, second, third, fourth, and n-th closest amino acid residue of each kind in the sequence. The number n=6 is likely to work well in many data sets [10]. To make the distances positive, absolute values are employed.

$$RAM_{ij} = |AA_j^i - K| \tag{1}$$

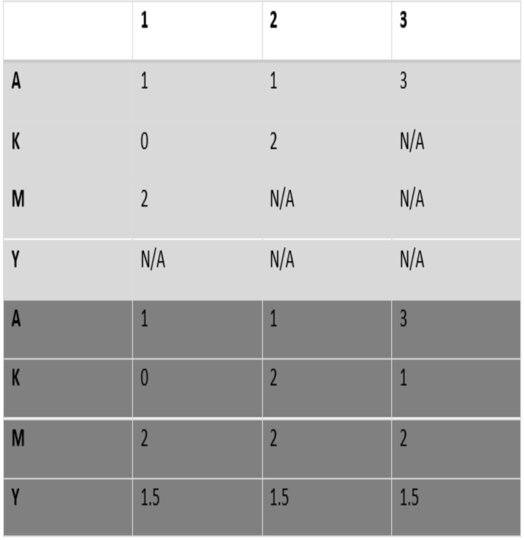
In Equation (1), RAM represents the residue adjacency matrix, K represents the lysine index in sequence, and AA represents the amino acid index. Furthermore, i represents the amino acid type, and j represents the jth closest amino acid of the same type.(Figure 2). depicts this formulation.



	1	2	3	4	5	6
1	3	4	9	16	17	18
2	39	44	50	54	69	72
3	60	63	75	119	126	52
4	6	6	8	12	17	32
5	43.9898	43.9898	43.9898	43.9898	43.9898	43.9898
6	10	13	19	25	49	77
7	12	13	21	26	31	41
8	11	62	87	91	100	118
9	86	86	86	86	86	86
10	38	79	97	71.333	71.333	71.333
11	3	7	8	10	15	16
12	0	7	21	23	40	43
13	14	22	57	129	55.5	55.5
14	4	30	37	65	93	103
15	2	2	11	47	116	121
16	1	5	5	9	14	24
17	21	29	71	73	83	136
18	42	42	42	42	42	42
19	36	53	61	107	151	81.6
20	1	15	28	33	35	56

Fig. 2. 20x6 Residue Adjacency Matrix

The value of n (in the figure 3) is three. It demonstrates how the hole gets filled if there aren't enough of one kind of amino acid to finish the matrix. While row Y's mean is (1 + 1 + 3 + 0 + 2 + 2)/6 = 1.5, row K's mean is (0 + 2)/2 = 1. Tyrosine (Y) is absent from the sequence in this instance, so the average of the whole matrix is computed. Since the third closest lysine residue, Lysine (K), is missing, the mean is derived from the lysine residues that are present. [9].



	1	2	3
A	1	1	3
K	0	2	N/A
M	2	N/A	N/A
Y	N/A	N/A	N/A
A	1	1	3
K	0	2	1
M	2	2	2
Y	1.5	1.5	1.5

Fig. 3. A residue adjacency matrix is created by employing a fictitious protein sequence MAKAKAA with n = 3. [9].

C. Feature Construction

After constructing RAM, two feature construction techniques Amino Acid Composition(AAC) and Tri-Peptide Composition(TPC) were executed to form a fused feature vector from the given RAM.

1) Amino Acid Composition: The frequency of each amino acid type in a protein or peptide sequence is calculated using the Amino Acid Composition (AAC) encoding [11]. The following formula was used to extract AAC features from RAM:

$$X_j = \frac{1}{L}\sum_{i=0}^{1} h_{i,j} \quad (j = 1, 2, ...6) \tag{2}$$

where $h_{i,j}$ denotes the value in the RAM's i^{th} row and j^{th} column in RAM [12].

2) Tripeptide Composition: There are 8000 characteristics in the Tripeptide Composition (TPC) [11].Since the importance of sequence-order information is overlooked by AAC, TPC characteristics are calculated from the RAM in the following way [12] to partially represent the local sequence-order effect:

$$Y_{ij} = \frac{\sum\limits_{k=1}^{L-1} h_{k,i} \times h_{k+1,j}}{\sum\limits_{j=1}^{6}\sum\limits_{k=1}^{L-1} h_{k,i} \times h_{k+1,j}} \quad (1 \le i, j \le 6) \tag{3}$$

where $h_{i,j}$ denotes the value in the RAM's i^{th} row and j^{th} column in RAM [12].

D. Fused Vector

For 233 negative samples and 93 positive samples;for a total of 326 samples a 42-dimensional fused feature vector is gained by combining feature vectors gained from AAC and TPC. We refer to this feature encoding approach as AATP-RAM.

E. Dataset Balancing

After employing cleaning method the imbalanced ratio was nearly 1:2 ratio.Firstly, 20% of the sample dataset was chosen at random to assess the model strength and designated as independent test set. The dataset was split into training and testing sets containing 80% and 20% of the data.To prevent potentially biased predictions, a 1:1 ratio of positive to negative samples was consolidated as the training model from the entire remaining dataset.Synthetic Minority Oversampling Technique(SMOTE) [13] was implemented to over sample the training and testing dataset, 233 positive and 233 negative samples were found in the balanced feature set.

F. Feature Selection Using Recursive Feature Elimination

After getting the combined feature vector, the recursive feature elimination technique was performed using the gradient boosting classifier. It's a backward prediction selection that begins by constructing a model from the complete collection of feature vectors and assigning a significance score to each feature.The model is then rebuilt with the least important feature(s) removed, and the significance scores are computed again.In our study,n(number of feature subsets to evaluate) was set to 25 to get the optimal result.And also the model is trained with all 42 features using different classifiers to better fitting.

G. Classifiers

To train our model, we employed random forest (RF), light gradient boosting machine(LightGBM), eXtreme gradient boosting(XGBoost), and k-nearest neighbors(KNN).LightGBM was utilized as it lowered the processing cost by making the continuous values discrete and our dataset consisted of floating point.Furthermore, Xgboost was used for shrinkage because the dataset was small and prone to overfitting [14]. This method scales newly added weights by a factor of eta after each step of tree boosting and lessens the influence of each individual tree, leaving room for future trees to improve the model while preventing over-fitting. And one of the top classifiers for quickly learning discriminative features from a short dataset is KNN.

III. Results and Discussions

A. Statistical Measures

Four statistical metrics were used to determine predictor's performance : sensitivity, specificity, accuracy. These metrics assess a predictor's ability to detect phosphoglycerylated lysine positions non-phosphoglycerylated lysine sites, having a value between 0 and 1.The accuracy score assesses a predictor's ability to discriminate between phosphoglycerylated and non-phosphoglycerylated locations and ranges from 0 to 1. In terms of equations, the three statistical measures are as follows:

$$Sensitivity(Sn) = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity(Sp) = \frac{TN}{TN + FP} \tag{5}$$

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

True positives, false negatives, true negatives, and false positives are represented by TP, FN, TN, and FP in Equations 4 to 6, respectively.

B. Test Scheme

The stratified ten-fold cross-validation,stratified six-fold cross validation,repeated stratified ten-fold cross validation technique was used to evaluate our predictor's performance [15]. The preceding section's statistical measurements were generated for each fold and then averaged to represent the predictor's overall performance. LightGBM in repeated stratified ten-fold with 42 features outperformed most other existing models with 0.8877 accuracy score whereas LightGBM in stratified ten-fold with 42 features gained 0.8777 accuracy score and LightGBM in stratified six-fold with 25 features gained accuracy score of 0.8751.

When the number of data in each class is equal, the trade-off between the true positive rate and the false positive rate for a predictive model using various probability thresholds is depicted by ROC (receiver operating characteristic) curves.

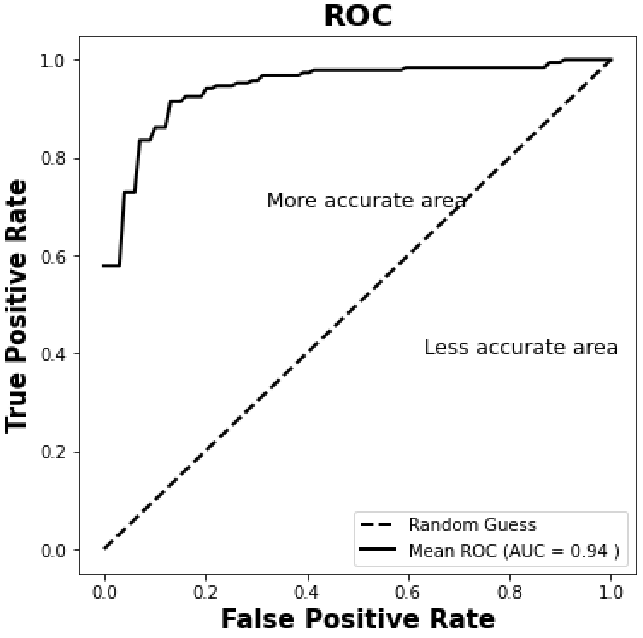


Fig. 4. Receiver operator characteristics (ROC) curve of LightGBM using stratified 6-fold cross-validation with 25 features

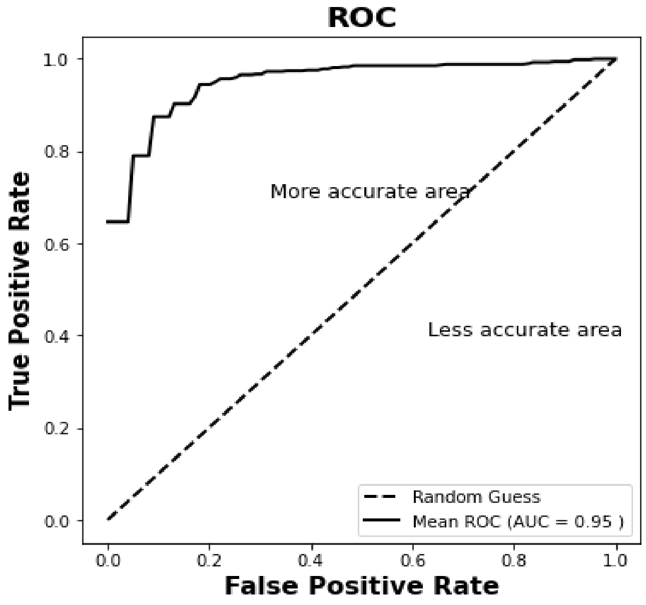


Fig. 5. Receiver operator characteristics (ROC) curve of LightGBM using repeated stratified 10-fold cross-validation with 42 features

C. Different Classifiers Performance Comparison

By using all the features best performance was found for LightGBM with repeated stratified ten-fold cross validation. Table I shows LightGBM, XGBoost, KNN classifier results with stratified ten-fold cross-validation for 42 features.

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT CLASSIFIERS USING STRATIFIED TEN-FOLD CROSS VALIDATION FOR 42 FEATURES

Classifier	Specificity	Sensitivity	Accuracy
$LightGBM*$	0.8923	0.8826	**0.8877**
$LightGBM **$	0.8715	0.8838	0.8777
$XGBoost$	0.8414	0.8887	0.8649
KNN	0.7858	0.8579	0.8218

LightGBM using repeated stratified ten-fold*
LightGBM using stratified ten-fold**

Using recursive feature elimination with gradient boosting and LightGBM classifier produced optimal result for 25 features.Table II shows among classifiers results with stratified six-fold cross-validation for 25 features.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT CLASSIFIERS USING STRATIFIED SIX-FOLD CROSS VALIDATION WITH RFE(25 FEATURES)

Classifier	Specificity	Sensitivity	Accuracy
$LightGBM$	0.8825	0.8671	**0.8751**
$XGBoost$	0.8667	0.8827	0.8749
KNN	0.6962	0.8350	0.7658
RF	0.8667	0.8190	0.8430

Table III furthermore shows classifier results with stratified six-fold cross validation for all 42 features without RFE and XGBoost outperformed all other predictors. To avoid under-fitting, 10 fold cross validation was utilized with six fold cross validation when comparing classifiers.For our dataset, K=10 has been empirically demonstrated to produce test error rates that claim that neither suffer from extremely high bias nor from very low bias.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT CLASSIFIERS USING STRATIFIED SIX-FOLD CROSS VALIDATION WITHOUT RFE(42 FEATURES)

Classifier	Specificity	Sensitivity	Accuracy
$LightGBM$	0.8613	0.8669	**0.8643**
$XGBoost$	0.8721	0.8723	**0.8724**
KNN	0.7809	0.8561	0.8189
RF	0.8669	0.8346	0.8510

D. Performance Comparison With Other Predictors

Comparisons were made with other existing predictors such as CKSAAP-PhoglySite [2], Bigram-PGK [6], and RAM-PGK [9] predictors using a stratified 6-fold cross-validation scheme. Our predictor using LightGBM classifier with 42 features in stratified ten-fold and LightGBM with 42 features in repeated stratified ten-fold outperformed most other existing predictors and achieved the highest accuracy. Even LightGBM with 25 features using recursive feature elimination with six-fold cross validation achieved better results.

TABLE IV
COMPARISON WITH OTHER PREDICTORS

Predictor	Sp	Sn	Accuracy
$CKSAAP\_Phoglysite$	0.6722	0.3494	0.6616
$Bigram\_PGK$	0.6639	0.4055	0.6554
RAM_PGK	0.6436	0.5741	0.6414
$LightGBM*$	0.8923	0.8826	**0.8877**
$LightGBM**$	0.8715	0.8838	**0.8777**
$LightGBM***$	0.8613	0.8669	**0.8643**
$LightGBM****$	0.8825	0.8671	**0.8751**

LightGBM*=repeated stratified ten-fold,42 features
LightGBM**=stratified ten-fold,42 features
LightGBM***=stratified six-fold,42 features
LightGBM****=stratified six-fold,25 features

IV. CONCLUSION

In this work, a new computational method, KPhogly-RAM, has been developed utilizing the residue adjacent matrix to find protein phosphoglycerylation sites with higher accuracy. This novel approach of constructing a residue adjacency matrix from lysin residues and transforming it into a fused vector employing feature construction techniques AAC and TPC played a dominant role in identifying the lysine alteration. The choice of classifiers LightGBM,XGboost and KNN proved successful in distinguishing between modified and unmodified lysine sites. Selection of ten-fold cross validation was effective in avoiding under-fitting.In the repeated stratified ten-fold cross-validation test, the accuracy score was higher i.e 0.8877 with a specificity score of 0.8923 and sensitivity score of 0.8826 which indicates the predictor's stability.Our work has showed that generating residue adjacency matrix from protein residue was successful in building phosphoglycerylation predictor.These experimental results show that KPhogly-RAM outperforms the most other cutting-edge phosphoglycerylation site predictors.

REFERENCES

[1] 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.

[2] Z. Ju, J.-Z. Cao, and H. Gu, "Predicting lysine phosphoglycerylation with fuzzy svm by incorporating k-spaced amino acid pairs into chous general pseaac," *Journal of Theoretical Biology*, vol. 397, pp. 145–150, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022519316001168

[3] Q.-Y. Chen, J. Tang, and P.-F. Du, "Predicting protein lysine phosphoglycerylation sites by hybridizing many sequence based features," *Mol. BioSyst.*, vol. 13, pp. 874–882, 2017. [Online]. Available: http://dx.doi.org/10.1039/C6MB00875E

[4] A. Chandra, A. Sharma, A. Dehzangi, S. Ranganathan, A. Jokhan, K.-C. Chou, and T. Tsunoda, "Phoglystruct: prediction of phosphoglycerylated lysine residues using structural properties of amino acids," *Scientific reports*, vol. 8, no. 1, pp. 1–11, 2018.

[5] A. A. Chandra, A. Sharma, A. Dehzangi, and T. Tsunoda, "Evolstruct-phogly: incorporating structural properties and evolutionary information from profile bigrams for the phosphoglycerylation prediction," *BMC genomics*, vol. 19, no. 9, pp. 1–9, 2019.

[6] 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.

[7] 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.

[8] "PLMD - Protein Lysine Modifications Database." [Online]. Available: http://plmd.biocuckoo.org/

[9] A. A. Chandra, A. Sharma, A. Dehzangi, and T. Tsunoda, "Ram-pgk: Prediction of lysine phosphoglycerylation based on residue adjacency matrix," *Genes*, vol. 11, no. 12, p. 1524, 2020.

[10] N. J. Mapes Jr, C. Rodriguez, P. Chowriappa, and S. Dua, "Residue adjacency matrix based feature engineering for predicting cysteine reactivity in proteins," *Computational and structural biotechnology journal*, vol. 17, pp. 90–100, 2019.

[11] M. Bhasin and G. P. Raghava, "Classification of nuclear receptors based on amino acid composition and dipeptide composition," *Journal of Biological Chemistry*, vol. 279, no. 22, pp. 23 262–23 266, 2004.

[12] J. Wang, H. Zheng, Y. Yang, W. Xiao, and T. Liu, "Preddbp-stack: prediction of dna-binding proteins from hmm profiles using a stacked ensemble method," *BioMed research international*, vol. 2020, 2020.

[13] "SMOTE | Overcoming Class Imbalance Problem Using SMOTE," Oct. 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

[14] "XGBoost: What it is, and when to use it." [Online]. Available: https://www.kdnuggets.com/xgboost-what-it-is-and-when-to-use-it.html

[15] S. Yadav and S. Shukla, "Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification," in *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, 2016, pp. 78–83.