




Article

Accurately Predicting Glutarylation Sites Using Sequential Bi-Peptide-Based Evolutionary Features

Md. Easin Arifat ^{1,†} , Md. Wakil Ahmad ^{1,†}, S.M. Shovan ^{2,†}, Abdollah Dehzangi ^{3,4}, Shubhashis Roy Dipta ¹ , Md. Al Mehedi Hasan ², Ghazaleh Taherzadeh ^{5,*}, Swakkhar Shatabda ^{1,*}  and Alok Sharma ^{6,7,8,9,*}

¹ Department of Computer Science and Engineering, United International University, Dhaka 1212, Bangladesh; marafat152047@bscse.uui.ac.bd (M.E.A.); mahmad152213@bscse.uui.ac.bd (M.W.A.); iamdipta@gmail.com (S.R.D.)

² Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi 6204, Bangladesh; sm.shovan@gmail.com (S.M.S.); mehedi_ru@yahoo.com (M.A.M.H.)

³ Department of Computer Science, Rutgers University, Camden, NJ 08102, USA; i.dehzangi@rutgers.edu

⁴ Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA

⁵ Institute for Bioscience and Biotechnology Research, University of Maryland, College Park, MD 20742, USA

⁶ Institute for Integrated and Intelligent Systems, Griffith University, Brisbane, QLD 4111, Australia

⁷ Department of Medical Science Mathematics, Tokyo Medical and Dental University (TMDU), Tokyo 113-8510, Japan

⁸ Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045, Japan

⁹ School of Engineering and Physics, Faculty of Science Technology and Environment, University of the South Pacific, Suva, Fiji

* Correspondence: gtaherza@umd.edu (G.T.); swakkhar@cse.uui.ac.bd (S.S.); alokanand.sharma@riken.jp (A.S.)

† These authors contributed equally to this work.

Received: 1 July 2020; Accepted: 27 August 2020; Published: 31 August 2020



Abstract: Post Translational Modification (PTM) is defined as the alteration of protein sequence upon interaction with different macromolecules after the translation process. Glutarylation is considered one of the most important PTMs, which is associated with a wide range of cellular functioning, including metabolism, translation, and specified separate subcellular localizations. During the past few years, a wide range of computational approaches has been proposed to predict Glutarylation sites. However, despite all the efforts that have been made so far, the prediction performance of the Glutarylation sites has remained limited. One of the main challenges to tackle this problem is to extract features with significant discriminatory information. To address this issue, we propose a new machine learning method called BiPepGlut using the concept of a bi-peptide-based evolutionary method for feature extraction. To build this model, we also use the Extra-Trees (ET) classifier for the classification purpose, which, to the best of our knowledge, has never been used for this task. Our results demonstrate BiPepGlut is able to significantly outperform previously proposed models to tackle this problem. BiPepGlut achieves 92.0%, 84.8%, 95.6%, 0.82, and 0.88 in accuracy, sensitivity, specificity, Matthew's Correlation Coefficient, and F1-score, respectively. BiPepGlut is implemented as a publicly available online predictor.

Keywords: post-translational modification; lysine Glutarylation; machine learning; extra-trees classifier; bi-peptide evolutionary features

1. Introduction

Post-translational modifications (PTMs) of proteins are associated with various biological processes. They also play a vital role in the diversification of protein functioning in different biological and physiological interactions [1,2]. PTMs are associated with different functions such as systematizing biological activities and regulating localization, and proteins interacting with other cellular molecules. PTMs are also key components of biological processes for the transmission of the genetic code and the control of cellular physiology. In 2016, Trost et al. [3] described the DAPPLE2 tool to predict 20 different types of PTMs from 15 online databases. DAPPLE2 is able to make the prediction task faster than its previous version, DAPPLE [4]. Later on, Li et al. [5] developed a new R package, named PTMscape, that predicts PTM sites based on diverse sets of physicochemical-modified properties. More recently, Chen et al. [6] introduced MUsCADEL tools for the PTMs prediction using deep learning. So far, more than 600 types of PTMs have been identified. Some of the most widely observed PTMs are Acetylation [7], Propionylation [8], Sumoylation [9], Succinylation [10,11], Malonylation [12], and Methylation [13,14] among the main 20 contributing amino acids to build proteins [15].

Lysine Glutarylation is among the recently identified PTMs. Glutarylation occurs when an amino acid along the protein sequence interacts with a glutaryl group. Glutarylated proteins have been identified for many metabolic procedures and mitochondrial functions in both eukaryotic and prokaryotic cells [16]. Among the most important ones, Glutarylation dysregulation has been related in the etiology of metabolic disorders such as cancer [17], mycobacterium tuberculosis [18], diabetes, and brain and liver disorders [19]. Therefore, due to the tangled characteristic and limited knowledge of Glutarylation sites, further analysis for a better understanding of the nature of Glutarylation is required.

During the past few years, a wide range of methods has been proposed to predict Glutarylation sites using many machine learning approaches [20–25]. Recently, many deep learning models have been used to predict different types of PTMs [6,26–29]. In one of the earliest studies, Tan et al. detected 23 Glutarylation sites in 13 unique proteins from HeLa cells [16]. They also examined 683 lysine Glutarylation sites in 191 individual proteins. After that, Xie et al. also identifies 41 Glutarylation sites in 24 Glutarylated proteins. They extracted features based on the composition of amino acids and amino acid interactions and used Support Vector Machine (SVM) as their classifier [18].

In a different study, López et al. proposed structural features and evolutionary information of amino acids to predict the succinylation sites, which is closely related to Glutarylation sites prediction [20,21]. Recently, Zhe et al. [22] developed a predictor tool named GlutPred. To predict the Glutarylation sites, they extracted different kinds of features and applied a maximum relevance minimum redundancy feature selection method. They also used a biased SVM classifier to build GlutPred. At the same time, Yan et al. [23] proposed another predictor, called iGlu-Lys, to tackle this problem. They used a wide range of features and selected the optimal features using special-position information and amino acid pair order. They also used SVM as their preferred classifier. More recently, Huang et al. [24] proposed a new model called MDDGluar. To build this model, they used sequence-based features such as Amino Acid Composition (AAC), Amino Acid Pair Composition (AAPC), and Composition of k-spaced Amino Acid Pairs (CKSAAP). They also employed the SVM classifier to identify the Glutarylation sites. Most recently, Hussam et al. [25] developed another tool, named RF-GlutarySite, that uses sequence-based and physicochemical-based features and employs Random Forest (RF) as a classifier.

Despite all the efforts that have been made so far, the overall performance of the lysine Glutarylation site prediction task remained limited. The main challenge to enhance lysine Glutarylation site prediction performance is the use of features that provide significant discriminatory information. In this paper, we propose a new model called BiPepGlu that uses a bi-peptide-based evolutionary feature extraction concept to enhance lysine Glutarylation prediction performance. We investigate the impact of several classifiers and choose the one with the best performance to build our model. Among them, Extra-Trees (ET) classifier outperforms other classifiers, which is used to build BiPepGlu.

The entire methodology is described in detail in the following sections. An overview of the general architecture of BiPepGlut is given in Figure 1. Our results demonstrate that BiPepGlut is able to significantly enhance lysine Glutarylation prediction accuracy compared to those methods found in the literature. BiPepGlut achieves 92.0%, 84.8%, 95.6%, 0.82, and 0.88 in accuracy, sensitivity, specificity, Matthew's Correlation Coefficient (MCC), and F1-score on the employed independent test, respectively. Such results demonstrate more than 3% improvement for sensitivity, and over 0.3 improvements for MCC compared to those reported in the previous studies. BiPepGlut is implemented as an online predictor and is publicly available at: www.brl.uiu.ac.bd/bioglutarylation/. The data, code, and all the Supplementary Materials used to build the BiPepGlut method are publicly available at: <https://github.com/Wakiloo7/BipepGlut>.

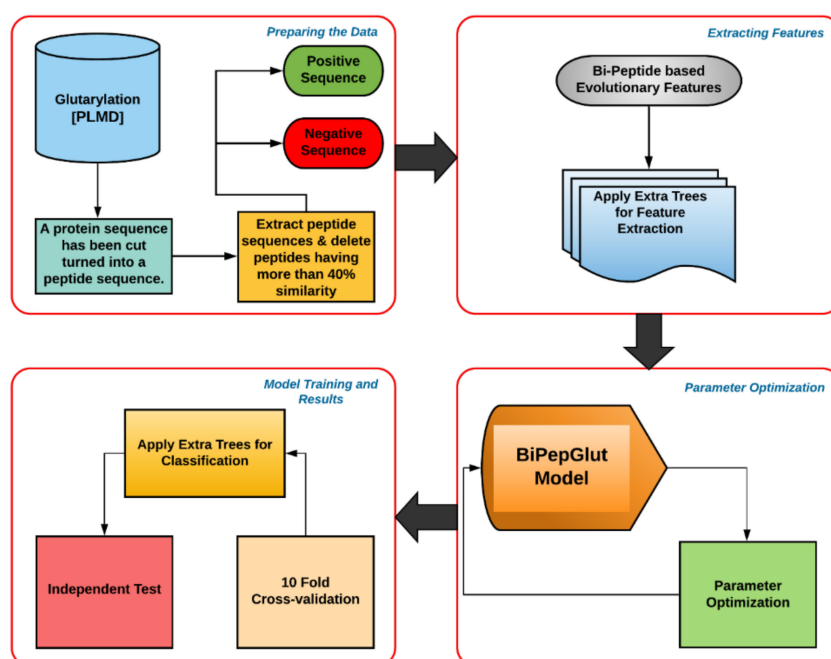


Figure 1. This flow chart demonstrates the general architecture of BiPepGlut. The positive and negative sites were yielded from a public database. Features were then extracted using the bi-peptide-based evolutionary feature extraction technique and then the useful features are selected. After that, the Extra Tree (ET) classifier was trained using our extracted features and then evaluated using 10-fold cross-validation and an independent test set.

2. Materials and Methods

In this section, we present our employed benchmark, how it is prepared for further experimentation, our employed classifiers, proposed feature extraction, and measurement methods.

2.1. Dataset

In this study, we collected a Glutarylation dataset from Protein Lysine Modifications Database (PLMD) [30]. The PLMD repository contains datasets for different PTM sites. All the PTMs recorded in this repository are those that are interacted with the lysine amino acid along the protein sequence. It is mainly because lysine has a high tendency to engage in PTM interaction compared to other amino acids. This dataset contains 211 proteins, which have 715 lysine Glutarylation sites belonging to *Mus musculus* (mouse) and *Mycobacterium tuberculosis* species. Among them, 674 sites in 187 proteins and 41 sites in 24 proteins belong to *Mus musculus* and *Mycobacterium tuberculosis*, respectively. We then cut the protein sequences into peptides by considering window size as 21. This window size has been widely used in the literature and shown to be the best among other window sizes [9,23,25].

For better representation, we use an alphabet notation, where the upstream and downstream lengths are denoted as $\xi = 10$, and the entire window size is $2\xi + 1$ ($2 \times 10 + 1 = 21$). The responsible residue for the Glutarylation site is one letter notation of K (amino acid lysine). Alongside this, a dummy residue (X) has been added on both sides of the proteins when the lysine is in the N-terminus or C-terminus of the proteins and does not have 10 neighboring amino acids in both sides to ensure the uniform length upstream and downstream. This process is shown in Figure 2.

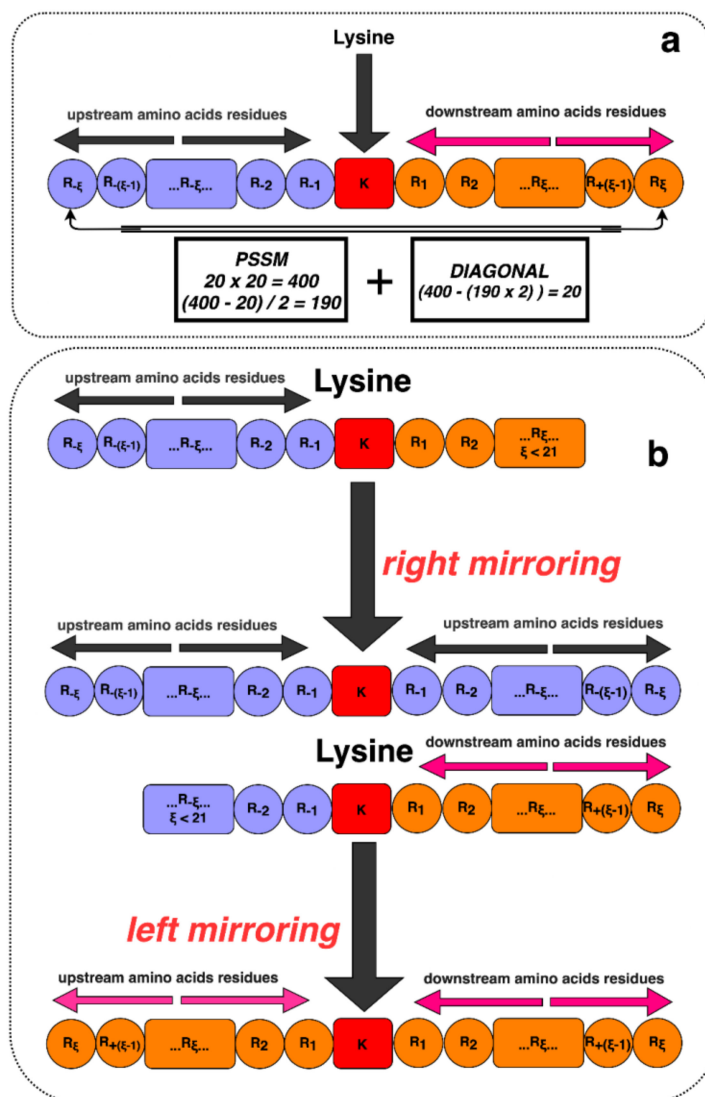


Figure 2. Illustration of lysine residues with its surrounding upstream and downstream amino acids. (a) Lysine residues with sufficient neighboring amino acids. (b) A scenario of adding dummy residues in N-terminus and C-terminus to have insufficient amino acids neighbors on either the upstream or downstream segment.

As a result, we have a total of 723 Glutarylation sites (positive) and 4626 Non-Glutarylation (negative) sites. Later on, we applied CD-HIT [31] over the negative sequences to remove sequences with high sequential similarity. In this case, we use 40% similarity cut-off as it is widely used in the literature [23–25]. Due to the limited availability of positive samples compared to the negative samples, the peptides with positive sites remain untouched. To provide more insight into our employed benchmark, we produce ranking of homology in the positive and negative hits separately using CD-HIT, which is now available at: <https://github.com/Wakiloo7/BiPepGlut/tree/master/CD-HIT>. In this way, we can avoid underfitting our model in predicting positive sites. However, we use both 10-fold

cross-validation and an independent test set to investigate the generality of our model and to avoid bias in our model. As a result of using CD-HIT, the 1923 Non-Glutarylation sequence remains from the original 4626 negative sites. We cross-checked positive sequences in the negative sites to make sure about the validity of our employed benchmark. From the remaining samples, we randomly separate 90% of the samples to build the training set while the remaining 10% to build the independent test set.

2.2. Feature Extraction

Feature extraction is an important step in building an effective and accurate machine learning model. In general, feature extraction is the method of selecting, handling, and managing a set of F features from a given dataset. For our case, the employed data set contains protein sequences. A wide range of feature extraction techniques has been proposed in the literature to extract discriminatory information to represent protein sequences [20,21,32,33]. Most of the extracted features for Glutarylation site predictions are based on the physicochemical or alphabetic sequential properties of the proteins. However, the other sources for feature extraction such as evolutionary-based features have not been adequately explored for the Glutarylation site prediction task [34–36]. In this scenario, we focus on extracting evolutionary-based features using the bi-peptide method to tackle this problem.

2.3. Bi-Peptide-Based Evolutionary Feature Extraction Technique

Peptide is a molecule consisting of two or more amino acids. Peptides are usually shorter than proteins. Our proposed concept includes bi-peptide-based evolutionary feature extraction techniques to predict the Glutarylation sites. This technique has been effectively used for similar studies [33–40]. We extract the features straight from the Position Specific Scoring Matrix (PSSM) as one of the most important resources to extract evolutionary information. PSSM matrix is produced as the output of the Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) [41]. PSI-BLAST aligns a given peptide sequence with a protein database to identify similar sequences and produces PSSMs. These PSSMs specify the substitution score of a given amino acid of a protein sequence compared with other protein sequences. Such a substitution score determines the possibility of a given amino acid is substituted to other amino acids due to evolutionary changes. In this case, we execute PSI-BLAST using three iterations and a cut-off e-value (E) 0.001 to generate the PSSM matrix.

In this study, Glutarylated (positive) and Non-Glutarylated (negative) sites and their neighboring amino acids (10 upstream and 10 downstream of amino acids) were allied to extract the features. In this scenario, these neighboring amino acids are presented with the $P_{\xi}(K)$ segment of sequences. For example, a peptide sample can be presented as:

$$P_{\xi}(K) = R_{-\xi} R_{-(\xi-1)} \dots R_{-2} R_{-1} \odot R_1 R_2 \dots R_{+(\xi-1)} R_{+\xi} \quad (1)$$

The central amino acid expresses as lysine (K) is indexed as ξ . The downstream is indicated as $R_{+\xi}$ and the upstream is denoted as $R_{-\xi}$. A substring of the protein sample is $(2\xi + 1)$, which is the entire length of the peptide sequence. Two categories are shown in this case where each peptide samples fall under them.

$$P_{\xi}(K) \in \begin{cases} P_{\xi}^{+}(K), & \text{if the responsible residue is a Glutarylation site} \\ P_{\xi}^{-}(K), & \text{otherwise} \end{cases} \quad (2)$$

In this scenario, the negative Glutarylated set is denoted as $P_{\xi}^{-}(K)$, and the positive Glutarylated set is denoted as $P_{\xi}^{+}(K)$. As a result, we can introduce our benchmark dataset as:

$$S_{\xi}(K) = S_{\xi}^{+}(K) \cup S_{\xi}^{-}(K) \quad (3)$$

where the Glutarylated set $P_{\xi}^{+}(K)$ is presented in terms of $S_{\xi}^{+}(K)$ and carries the Non-Glutarylated set $P_{\xi}^{-}(K)$, which is presented in terms of $S_{\xi}^{-}(K)$ while \cup describes the union operator. The following techniques are carried to produce the feature vector from our dataset.

(i) The peptide sequence can be presented by P that is constituted as:

$$P = R_1 R_2 R_3 R_4 \dots R_L \tag{4}$$

From the study of Schaffer et al. [41], P can be demonstrated by an $L \times 20$ dimensional matrix, which is shown as:

$$\begin{bmatrix} E_{1 \rightarrow 1} & E_{2 \rightarrow 1} & \dots & E_{L \rightarrow 1} \\ E_{1 \rightarrow 2} & E_{2 \rightarrow 2} & \dots & E_{L \rightarrow 2} \\ \vdots & \vdots & & \vdots \\ E_{1 \rightarrow 20} & E_{2 \rightarrow 20} & \dots & E_{L \rightarrow 20} \end{bmatrix} \tag{5}$$

Here, L refers to the length of P , and $\bar{E}_{i \rightarrow j}$ refers to 20 different amino acids that get propensity of the amino acid residue spread.

(ii) From Equation (5), we generate the transpose matrix as:

$$\begin{bmatrix} E_{1 \rightarrow 1} & E_{2 \rightarrow 1} & \dots & E_{L \rightarrow 1} \\ E_{1 \rightarrow 2} & E_{2 \rightarrow 2} & \dots & E_{L \rightarrow 2} \\ \vdots & \vdots & & \vdots \\ E_{1 \rightarrow 20} & E_{2 \rightarrow 20} & \dots & E_{L \rightarrow 20} \end{bmatrix} \tag{6}$$

with,

$$E_{i \rightarrow j} = \frac{E_{i \rightarrow j} - \bar{E}_j}{SD(\bar{E}_j)} \quad i = 1, 2, 3, \dots, L; j = 1, 2, \dots, 20 \tag{7}$$

where,

$$\bar{E}_j = \frac{1}{L} \sum_{i=1}^L E_{i \rightarrow j} \quad j = 1, 2, \dots, 20 \tag{8}$$

The standard deviation is calculated and denoted where \bar{E} denotes the mean of $E_{i \rightarrow j}$ for $i = 1, 2, \dots, 20$ by the following equation.

$$SD(\bar{E}_j) = \sqrt{\sum_{i=1}^L [E_{i \rightarrow j} - \bar{E}_j]^2 / L} \tag{9}$$

(iii) The newly created matrix $M^T M$ is evaluated by the matrix M as well as the transpose of a matrix, which, in turn, is a 20×20 matrix ($20 \times L \times L \times 20 = 20 \times 20$ matrix) of 400 elements. The transpose matrix M^T multiplies with the main M matrix. The resulting matrix is symmetric. Therefore, the upper or lower triangular plus the main diagonal will have all the information that is extracted in this matrix. The triangular matrix inhibits 210 features where the first 20 comes from the diagonal and the rest 190 $((400 - 20)/2)$ features are from either lower or upper triangle matrices $(190 + 20 = 210$ total), as shown below.

$$\begin{bmatrix} (1) \\ (2) & (3) \\ (4) & (5) & (6) \\ \vdots & \vdots & \vdots \\ (191) & (192) & (193) & \dots & (210) \end{bmatrix} \tag{10}$$

The new matrix was then converted to a vector consisting of 210 elements, which can be represented as follows.

$$P_{evo} = [\Theta_1^E, \Theta_2^E, \dots, \Theta_u^E, \dots, \Theta_{210}^E] \quad (11)$$

2.4. Handling Imbalanced Dataset

As explained in the Dataset Subsection, the number of Glutarylation sites (positive) is lesser than the number of Non-Glutarylation sites (negative). There are significantly more negative samples in our benchmark when compared to positive samples. Such an imbalance may lead the predictor to be biased toward the negative samples. To avoid such a bias, it is necessary to balance the employed dataset. To deal with this issue, various balancing schemes have been introduced in the literature [42–44].

To address this issue, we up-sample positive sites (Glutarylation) instead of down-sampling the negative sites (Non-Glutarylation). Down-sampling may reduce the important usable samples. In this study, we use an oversampling approach by creating well-characterized synthetic data [45–47]. To ensure the little variation based on the property of the dataset, we pick the maximum value of the entire feature vectors. We then multiply the positive sites to 1.0001 and 1.0005, where the new value is much closer to the original value as done in References [13,38]. Initially, we have 723 positive sites. Multiplying 723 positive sites with 1.0001 and 1.0005 ($723 + (723 \times 1.0001) + (723 \times 1.0005) = 2169$), the new values are much closer to the original values. We generate our newly created value in this approach. Therefore, the number of positive sites increases to 2169, while the number of negative sites is 1923, where the ratio between positive and negative is almost ≈ 1 . The overall balancing process is only applied to training data while the test data remain untouched. This is how we make sure to avoid over-fitting. Hence, the balancing strategy also contributes to diminishing bias.

2.5. Classification Techniques

Choosing the most useful classification technique is an essential step in building a machine learning method. In this study, we have applied different kinds of classification methods. These classifiers are also widely used in the literature and demonstrated promising results for similar studies [13,39,48–50]. In this case, we study several ensemble learning methods such as Extreme Gradient Boosting (XGBoost) [48], Extra Tree (ET) Classifier [49], and Random Forest (RF) [25]. We also investigate several meta-classifiers such as Adaptive Boosting (AdaBoost) [39] and a tree-based learning algorithm Light Gradient Boosting Machine (LightGBM) [13]. We also study several of the most popular classifiers such as the Multi-layer Perceptron (MLP) classifier, which is a popular Artificial Neural Network (ANN) model [50].

In this study, the implementation of these classifiers is from the Scikit-learn version 0.19.2. To implement these algorithms through the classification model, we have used the following hyperparameters. Among them, some are default parameters and the rest of the parameters are tuned as required. In XGBoost, we tuned `n_estimators = 300`. For ET classifiers, we used `n_estimator = 10`, `min_sample_split = 2`. For RF classifiers, `max_depth = 2`, `random_state = 42`, and `n_estimators = 300`. For AdaBoost, we use `n_estimators = 300` while, for LightGBM, `num_leaves = 31`, `learning_rate = 40`, and `n_estimators = 40`. Lastly, for MLP, we use one hidden layer and 100 nodes, an activation function as Relu, $\alpha = 1$, `max_iter = 1000`, and `learning_rate_int = 0.001`. During the hyperparameters tuning, among all the classifiers, we identify that the ET classifier attains the best results compared to other classifiers. Extra Trees (ET) classifier uses an ensemble learning method, which is a type of meta estimator that fits many decision trees similar to the RF classifier. In ET, selected features have been chosen randomly by splitting. In many cases, ET improves predictive accuracy and diminishes the chances of over-fitting [49,51].

2.6. Performance Measurements

In this case, we use both 10-fold cross-validation, and an independent test set to study the performance and generality of our proposed model. We also use accuracy, sensitivity, specificity,

MCC, and F1-score as our performance measurements, which are used in previous studies [52,53]. Using these measurements, we will be able to directly compare our results with those reported in the earlier studies. These measurements are formulated as follows.

$$Accuracy (ACC) = \left(1 - \frac{GS_{-}^{+} + GS_{+}^{-}}{GS^{+} + GS^{-}}\right) \times 100 \quad (12)$$

$$Sensitivity (SN) = \left(1 - \frac{GS_{-}^{+}}{GS^{+}}\right) \times 100 \quad (13)$$

$$Specificity (SP) = \left(1 - \frac{GS_{+}^{-}}{GS^{-}}\right) \times 100 \quad (14)$$

$$MCC = 1 - \frac{\left(\frac{G_{-}^{+}}{GS^{+}} + \frac{G_{+}^{-}}{GS^{-}}\right)}{\sqrt{\left(1 + \frac{G_{+}^{-} - G_{-}^{+}}{GS^{+}}\right)\left(1 + \frac{G_{-}^{+} - G_{+}^{-}}{GS^{-}}\right)}} \quad (15)$$

$$F1 - score = 2 \times \frac{(PR \times RE)}{(PR + RE)} \quad (16)$$

where GS^{+} denotes positive (Glutarylation) sites that are correctly classified, GS^{-} denotes negative (Non-Glutarylation) sites that are classified correctly, GS_{-}^{+} indicates Non-Glutarylated peptides that are wrongly classified as Glutarylated, and GS_{+}^{-} , shows the Glutarylated peptides that are incorrectly predicted as Non-Glutarylated. Precision (PR) and Recall (RE) are also examined for the performance analysis along with the F1-score.

3. Results and Discussion

In this section, we will first present how we choose our employed classifier among a wide range of classifiers that we studied in this case. We then compare our results with the state-of-the-art models found in the literature and demonstrate the effectiveness of BiPepGlut. We then analyze our results.

3.1. Building Our Model by Choosing the Most Effective Classifier

In this case, we investigate and compare the performance of six machine learning algorithms: LightGBM [13], RF [25], AdaBoost [39], XGBoost [48], ET classifier [49], and MLP classifiers [50]. The results achieved for this comparison for the 10-fold cross-validation and independent test set are shown in Tables 1 and 2, respectively, where ACC is accuracy, SN is sensitivity, and SP is specificity. As shown in these tables, among these classifiers, LightGBM and ET obtain the best results. Among these two, ET achieves relatively better results.

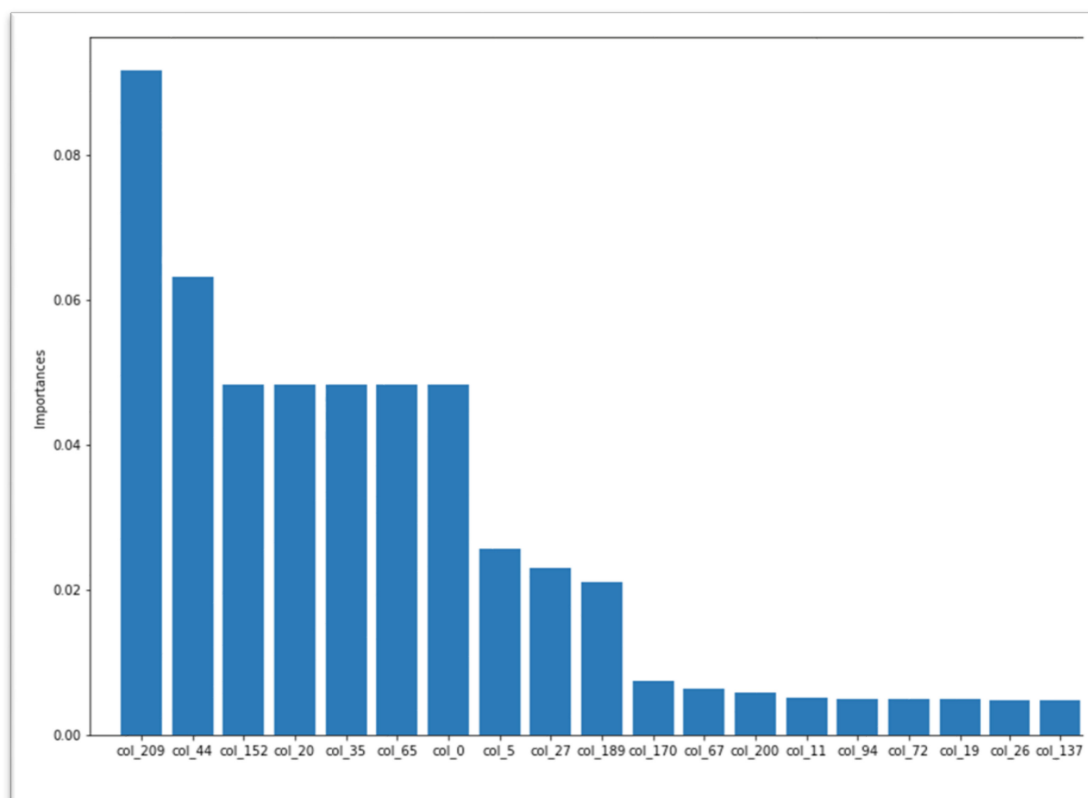
Table 1. Name of measuring matrices used for comparing performances based on a 10-fold cross-validation.

<i>Model</i>	<i>ACC (%)</i>	<i>SN (%)</i>	<i>SP (%)</i>	<i>MCC</i>	<i>F1-Score</i>
<i>RF</i>	80.2%	63.4%	96.9%	0.64	0.76
<i>XGBoost</i>	79.7%	67.8%	91.5%	0.61	0.76
<i>LightGBM</i>	82.9%	74.2%	91.5%	0.67	0.81
<i>AdaBoost</i>	79.2%	74.7%	83.8%	0.59	0.78
<i>ET classifier</i>	81.5%	70.0%	92.9%	0.64	0.79
<i>MLP</i>	78.7%	75.4%	82.0%	0.58	0.78

Table 2. Name of measuring matrices used for comparing performances based on the independent-test set.

<i>Classifier Model</i>	<i>ACC (%)</i>	<i>SN (%)</i>	<i>SP (%)</i>	<i>MCC</i>	<i>F1-Score</i>
<i>RF</i>	79.6%	41.3%	98.9%	0.54	0.58
<i>XGBoost</i>	80.3%	45.7%	97.8%	0.55	0.61
<i>LightGBM</i>	91.2%	78.3%	97.8%	0.80	0.86
<i>AdaBoost</i>	85.4%	76.1%	90.1%	0.67	0.78
<i>ET Classifier</i>	92.0%	84.8%	95.6%	0.82	0.88
<i>MLP</i>	84.7%	76.1%	88.0%	0.64	0.76

BiPepGlut also applies an Extra Trees (ET) classifier by using their Gini importance for computing the importance of features. To do this, we took each feature Gini importance and selected the top-most significant feature, according to their preference. The feature importance chart for our 210 lengths of features is shown in Figure 3. Note that, during model development, we exclusively work with these 210 features instead of skipping any features. Corresponding to all the results, from Tables 1 and 2, the ET [49] classifier obtains better performance compared to other classifiers. It achieves 81.5% in accuracy, 70.0% in sensitivity, 92.9% in specificity, 0.64 in MCC, and 0.79 in F1-score. In addition, the true positive (TP) rates are 1322, 460, and the false positive (FP) rates are 101-fold, 40-fold, and 10-fold cross validation and an independent test set, respectively.

**Figure 3.** Feature importance of 210 features selected for our model development.

We also plot the Receiver Operating Characteristic (ROC) to evaluate the output quality of the BiPepGlut both for 10-fold cross-validation and an independent test set. These plots are shown in Figures 4 and 5, respectively.

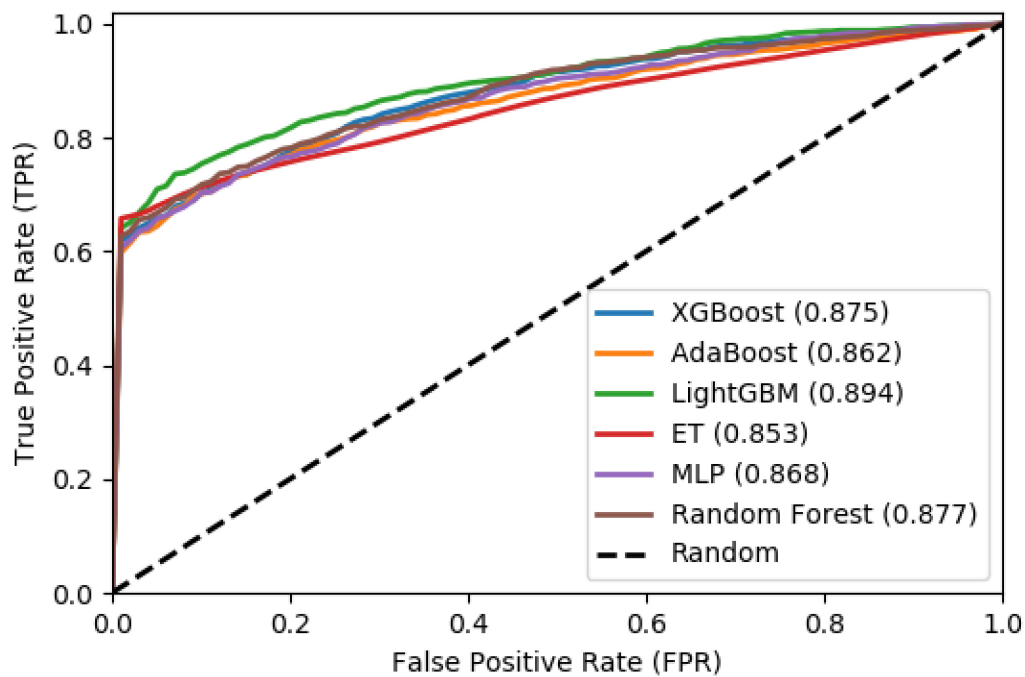


Figure 4. Receiver operator characteristic (ROC) curves using 10-fold cross-validation.

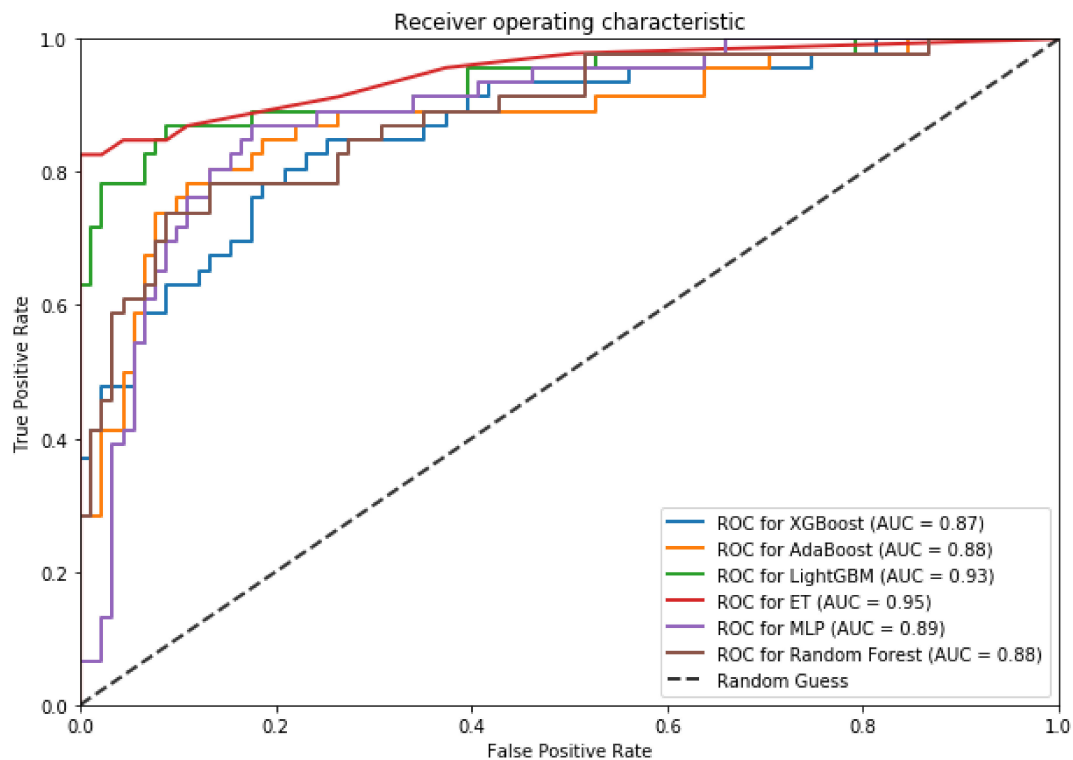


Figure 5. Receiver operator characteristic (ROC) curves using the independent test. The area under the curve (AUC) for each algorithm is indicated in parentheses.

These curves denote the X-axis as a false positive rate and Y-axis as a true positive rate. As shown in Figures 4 and 5, ET performs consistently better than other classifiers. As shown in this figure, ET achieves the area under the curve (AUC) of 0.85 on the 10-fold cross-validation and AUC of 0.95 for

the independent test set. Such results demonstrate the effectiveness of BiPepGlut using ET compared to other classifiers.

3.2. Comparison with State-of-the-Art Models

We compared our method with existing predictors that obtain the best results for the Glutarylation site prediction problem. To the best of our knowledge, we have identified three Glutarylation site predictors with the most promising results. In 2018, GlutPred [22] was developed using multiple feature extraction techniques along with maximum relevance and minimum redundancy feature selections to predict the Glutarylation sites. In the same year, iGlu-Lys [23] was developed using the finest features to predict the Glutarylation sites from the four-encodings method. Later on, RF-GlutarySite [25] was developed using sequence-based and physicochemical features and RF classifiers to predict the Glutarylation sites.

These three predictors are considered as the recent and most accurate predictors for the Glutarylation site prediction problem. To reproduce their results for our benchmark, we uploaded our sequences into their predictors and retrieved the performance of the predictors. Among these predictors, some are using 10-fold cross-validations, and others are using 6-fold, 8-fold, and 10-fold cross-validations during training. Consequently, their results may be exaggerated in independent test sets filtered from the entire data. Reproducing their results, we observed that the outcomes of those studies on independent test sets are better than assumed. Notwithstanding this, BiPepGlut was able to exceed even those obtained results.

We compare our method with these predictors (GlutPred, iGlu-Lys, and RF-GlutarySite). The results are shown in Table 3. As shown in this table, BiPepGlut achieves better results in terms of MCC, and the F1-score on the training set. The MCC and F1-score exceed 0.13 over previous predictor iGlu-Lys and 0.06 compared to RF-GlutarySite. The results of this comparison for the independent test set is shown in Table 4.

Table 3. Comparison of the performance of BiPepGlut to existing Glutarylation predictor using 10-fold cross-validation.

Predictor Tool	ACC (%)	SN (%)	SP (%)	MCC	F1-Score
GlutPred [22]	74.9%	64.8%	76.6%	0.32	0.43
iGlu-Lys [23]	88.4%	50.4%	95.2%	0.51	-
RF-GlutarySite [25]	75.0%	81.0%	68.0%	0.50	0.73
BiPepGlut	81.5%	70.0%	92.9%	0.64	0.79

Table 4. Comparison of the performance of BiPepGlut to an existing Glutarylation predictor using the independent-test set.

Predictor Tool	ACC (%)	SN (%)	SP (%)	MCC	F1-Score
GlutPred [22]	75.4%	51.8%	78.5%	0.22	0.33
iGlu-Lys [23]	88.5%	51.4%	95.3%	0.52	-
RF-GlutarySite [25]	72.0%	73.0%	70.0%	0.43	0.72
BiPepGlut	92.0%	84.8%	95.6%	0.82	0.88

As shown in Table 4, BiPepGlut consistently performed better than other models investigated in this study. Our results demonstrate that the BiPepGlut achieves over 3% better ACC compared to these studies. Sensitivity and the F1-score improved by 11.8% and 0.16, compared to RF-GlutarySite [25]. As shown in this table, BiPepGlut comes with prominent outcomes in all matrices and performs better

than those methods found in the literature. The significant improvement in sensitivity for our model demonstrates that BiPepGlut is able to identify Glutarylation sites by much more than those reported in previous studies. Given the performance of Glutarylation sites prediction, BiPepGlut can be considered as the most successful model compared to other studies found in the literature.

We also illustrate the barplot in Figure 6, which highlights the difference between the performance of BiPepGlut compared to GlutPred [22], iGlu-Lys [23], and RF-GlutarySite [25] in terms of accuracy.

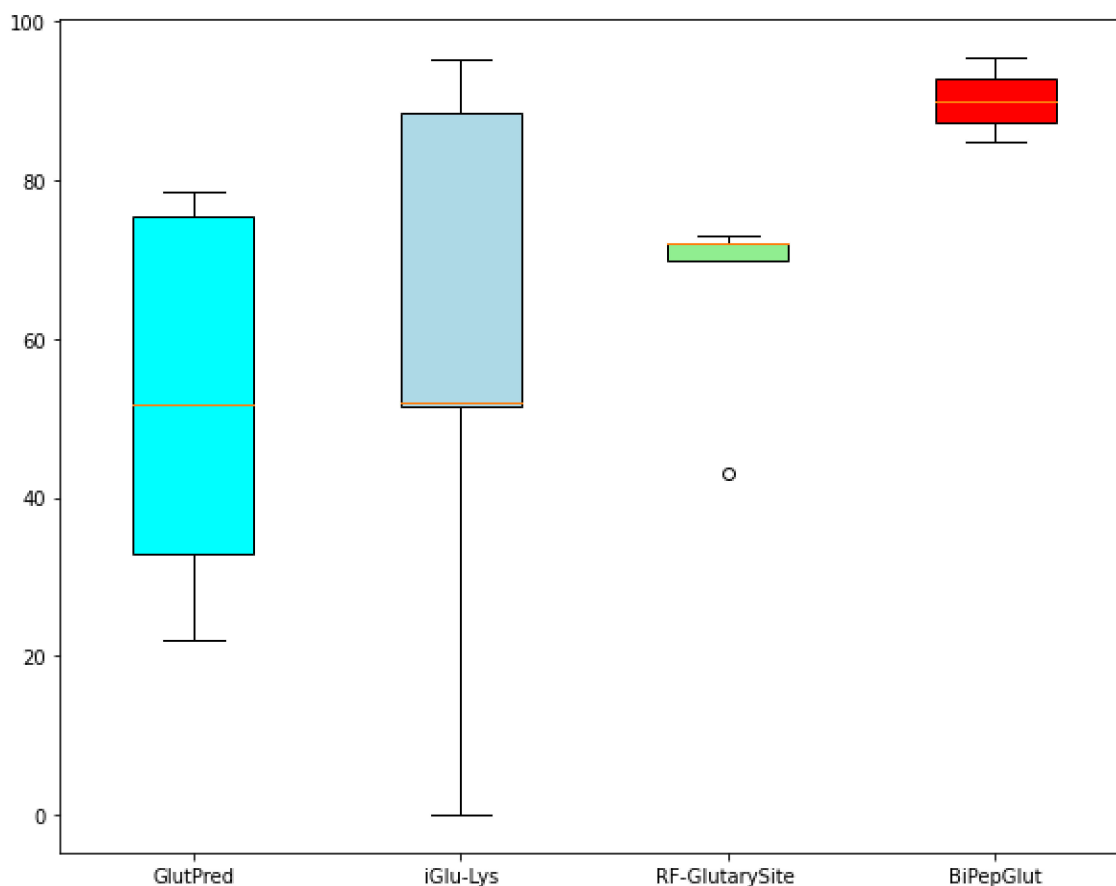


Figure 6. Comparing the results achieved using barplot among our model, BiPepGlut, GlutPred [22], iGlu-Lys [23], and RF-GlutarySite [25].

Results illustrated in this figure demonstrate the effectiveness and accuracy of BiPepGlut in predicting Glutarylation sites compared to those methods found in the literature. BiPepGlut is implemented as an online predictor and is publicly available at: www.brl.uiu.ac.bd/bioglutarylation/. In addition, the data, code, and all the Supplementary Materials used to build BiPepGlut metis are publicly available at: <https://github.com/Wakiloo7/BipepGlut>.

3.3. Web Server Implementation

We implemented BiPepGlut as a user-friendly and easy-to-use webserver. BiPepGlut is publicly available to use at: www.brl.uiu.ac.bd/bioglutarylation/. To use this predictor, the user has to provide a peptide sequence in fasta (.fsa) format. After uploading the sequence in BiPepGlut, PSSM files are generated from the server by using simultaneous iterations of PSI-BLAST where features are extracted and trained using the benchmark dataset. The goal of this predictor is to facilitate Glutarylation prediction. Figure 7 present the screen-shot of our online predictor.

Bioinformatics Research Lab

BiPepGlut
Accurately Predicting Glutarylation Sites using Sequential Bi-Peptide based Evolutionary Features

Enter or copy/paste query protein in Fasta format (Example)

```
MAGIAIKLAKDREAEEGLGSHERAIKYLNQDYETLRNECLEAGALFQDPSFPALPSSLGYKELGPYSSKTR
GIEWKRPTAICADPQFIIGGATRTDICQGALGDCWLLAAIASLTLNEILARVVPPDQSFQENYAGIFHFQF
WQYGEWVEVVDDRLPTKDGELLFVHSAEGSEFWSALLEKAYAKINGCYEALSGGATTEGFEDFTGGIA
EWYELRKPNNLFIKIKALEKGSLLGCSIDITSAADSEAVTYQKLVKGHAYSVTGAEVESSGSLQKLIRIR
NPWQGQVEWTGKWNDCPSWNTVDPEVRANLTERQEDGEFWSFSDFLRHYSRLEICNLTPDLTCDSD
YKWKLTMDGNWRRGSTAGGCRNYPNTFWMNPQYLKLEEEDEEEDGERGCTFLVGLIQKRRRRQ
RKMGEDMHTIGFGIYEVPEELTGGTNIHLGKNFFLTRARERSDTFINLREVLNRFKLPPEYVLPSTFEP
HKDGDFCIRVFSEKKADYQAVDDEIEANIEEIDANEEDIDDGFRRLVQLAGEDAEISAFELQTLRRVLAQR
QDIKSDGFSIETCKIMVMDLDEDEGSGKLGLEFYILWTKIQYQKIYREIDVDRSGTMNSYEMRKALEEAG
FKLPCQLHQVIVARFADDEIIDFDNFVRCLVRLLETFLKIFKQLDPENTGTIQLNLASWLSFSVL
```

Submit

References:
Md. Easin Arafat, Md. Wakil Ahmad, S.M. Shovan, Abdollah Dehzangi, Shubhashis Roy Dipta, Md. Al Mehedi Hasan, Ghazaleh Taherzadeh, Swakkhar Shatabda, & Alok Sharma. "Accurately Predicting Glutarylation Sites using Sequential Bi-Peptide based Evolutionary Features", (Submitted in Genes).

Copyright © Team BiPepGlut, Department of Computer Science and Engineering, United International University 2020

Figure 7. Screen-shot of BiPepGlut homepage.

4. Conclusions and Future Direction

In this study, we proposed a new method called BiPepGlut to predict the Glutarylation sites. To build BiPepGlut, we used bi-peptide-based evolutionary feature representation. We also used the Extra Tree classifier to build this model. Our results demonstrate that BiPepGlut can accurately predict the Glutarylation sites from Non-Glutarylation sites and improve the prediction results.

In the future, we aim to explore a wider range of features and include structural-based features to tackle this problem [39,54]. Such features are shown to be effective in solving similar problems in different studies. We also aim at comparing our extracted features with a wider range of feature extraction methods such as those extracted using iFeature [55] or BioSeq-Analysis [56,57]. In addition, we aim to find larger benchmarks that can allow us to use more advanced and complicated classifiers, such as Deep Learning, Convolutional Neural Network (such as DeepInsight [58]), and Recurrent Neural Network to enhance prediction accuracy even further. It is important to highlight that employing a larger benchmark will also enable us to provide more general and consistent results. Our future direction is to employ a larger benchmark as soon as it becomes available to further investigate the generality of our model. BiPepGlut is implemented as an online predictor and is publicly available at: www.brl.uui.ac.bd/bioglutarylation/. In addition, the data, code, and all the Supplementary Materials used to build BiPepGlut metis are publicly available at: <https://github.com/Wakiloo7/BipepGlut>.

Author Contributions: M.E.A., M.W.A., A.D. designed and performed the experiments. M.E.A., M.W.A., S.R.D. developed the web-server. M.E.A., M.W.A., S.M.S., S.R.D., G.T., A.D., A.S., M.A.M.H., S.S. wrote the manuscript. M.W.A., M.E.A. helped with figures and literature review. G.T., A.S., S.S. mentored and analytically reviewed the paper. All the authors reviewed the article. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Walsh, C.T.; Garneau-Tsodikova, S.; Gatto, G.J., Jr. Protein posttranslational modifications: The chemistry of proteome diversifications. *Angew. Chem. Int. Ed. Engl.* **2005**, *44*, 7342–7372. [CrossRef]

2. Xu, Y.; Ding, J.; Wu, L.Y. iSulf-Cys: Prediction of S-sulfenylation sites in proteins with physicochemical properties of amino acids. *PLoS ONE* **2016**, *11*, e0154237. [[CrossRef](#)]
3. Trost, B.; Maleki, F.; Kusalik, A.; Napper, S. DAPPLE 2: A Tool for the homology-based prediction of post-translational modification sites. *J. Proteome. Res.* **2016**, *15*, 2760–2767. [[CrossRef](#)]
4. Trost, B.; Arsenault, R.; Griebel, P.; Napper, S.; Kusalik, A. DAPPLE: A pipeline for the homology-based prediction of phosphorylation sites. *Bioinformatics* **2013**, *29*, 1693–1695. [[CrossRef](#)]
5. Li, G.X.; Vogel, C.; Choi, H. PTMscape: An open source tool to predict generic post-translational modifications and map modification crosstalk in protein domains and biological processes. *Mol. Omics.* **2018**, *14*, 197–209. [[CrossRef](#)]
6. Chen, Z.; Liu, X.; Li, F.; Li, C.; Marquez-Lago, T.; Leier, A.; Akutsu, T.; Webb, G.; Dakang, X.; Smith, A.; et al. Large-scale comparative assessment of computational predictors for lysine post-translational modification sites. *Brief. Bioinf.* **2019**, *20*, 2267–2290. [[CrossRef](#)]
7. Xie, Z.; Dai, J.; Dai, L.; Tan, M.; Cheng, Z.; Wu, Y.; Boeke, J.D.; Zhao, Y. Lysine succinylation and lysine malonylation in histones. *Mol. Cell. Proteom.* **2012**, *11*, 100–107. [[CrossRef](#)]
8. Kamynina, E.; Stover, P.J. The roles of SUMO in metabolic regulation. *Adv. Exp. Med. Biol.* **2017**, *963*, 143–168.
9. 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. Modell.* **2017**, *76*, 356–363. [[CrossRef](#)]
10. Li, S.; Li, H.; Li, M.; Shyr, Y.; Xie, L.; Li, Y. Improved prediction of lysine acetylation by support vector machines. *Protein Pept. Lett.* **2009**, *16*, 977–983. [[CrossRef](#)]
11. Zhang, Z.; Tan, M.; Xie, Z.; Dai, L.; Chen, Y.; Zhao, Y. Identification of lysine succinylation as a new post-translational modification. *Nat. Chem. Biol.* **2011**, *7*, 58–63. [[CrossRef](#)]
12. Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; Chou, K.C. pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. *J. Theor. Biol.* **2016**, *394*, 223–230. [[CrossRef](#)]
13. Ahmad, M.W.; Arafat, M.E.; Taherzadeh, G.; Sharma, A.; Dipta, S.R.; Dehzangi, A.; Shatabda, S. Mal-light: Enhancing lysine malonylation sites prediction problem using evolutionary-based features. *IEEE Access* **2020**, *8*, 77888–77902. [[CrossRef](#)]
14. Comb, D.G.; Sarkar, N.; Pinzino, C.J. The Methylation of lysine residues in protein. *J. Biol. Chem.* **1966**, *241*, 1857–1862.
15. Martin, C.; Zhang, Y. The diverse functions of histone lysine methylation. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 838–849. [[CrossRef](#)]
16. Hirshey, M.D.; Zhao, Y. Metabolic regulation by lysine malonylation, succinylation, and glutarylation. *Mol Cell Proteomics.* **2015**, *14*, 2308–2315. [[CrossRef](#)]
17. Park, K.J.; Kanehisa, M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics* **2003**, *19*, 1656–1663. [[CrossRef](#)]
18. Xie, L.; Wang, G.; Yu, Z.; You, M.; Li, Q.; Huang, H.; Xie, J. Proteome-wide lysine glutarylation profiling of the *Mycobacterium tuberculosis* H37Rv. *J. Proteome. Res.* **2016**, *15*, 1379–1385. [[CrossRef](#)]
19. Schmiesing, J.; Storch, S.; Dörfler, A.C.; Schweizer, M.; Makrypidi-Fraune, G.; Thelen, M.; Sylvester, M.; Gieselmann, V.; Meyer-Schwwsinger, C.; Koch-Nolte, F.; et al. Disease-linked glutarylation impairs function and interactions of mitochondrial proteins and contributes to mitochondrial heterogeneity. *Cell Rep.* **2018**, *24*, 2946–2956. [[CrossRef](#)]
20. López, Y.; Dehzangi, A.; Lal, S.P.; Taherzadeh, G.; Michaelson, J.; Sattar, A.; Tsunoda, T.; Sharma, A. SucStruct: Prediction of succinylated lysine residues by using structural properties of amino acids. *Anal. Biochem.* **2017**, *527*, 24–32. [[CrossRef](#)]
21. López, Y.; Sharma, A.; Dehzangi, A.; Lal, S.P.; Taherzadeh, G.; Sattar, A.; Tsunoda, T. Success: Evolutionary and structural properties of amino acids prove effective for succinylation site prediction. *BMC Genom.* **2018**, *19*, 923. [[CrossRef](#)]
22. Ju, Z.; He, J.J. Prediction of lysine glutarylation sites by maximum relevance minimum redundancy feature selection. *Anal. Biochem.* **2018**, *550*, 1–7. [[CrossRef](#)]
23. Xu, Y.; Yang, Y.; Ding, J.; Li, C. iGlu-Lys: A predictor for lysine glutarylation through amino acid pair order features. *IEEE Trans. NanoBiosci.* **2018**, *17*, 394–401. [[CrossRef](#)]
24. Huang, K.Y.; Kao, H.J.; Hsu, J.B.K.; Weng, S.L.; Lee, T.Y. Characterization and identification of lysine glutarylation based on intrinsic interdependence between positions in the substrate sites. *BMC Bioinform.* **2019**, *19*, 384. [[CrossRef](#)]

25. AL-Barakati, H.J.; Saigo, H.; Newman, R.H. RF-GlutarySite: A random forest based predictor for glutarylation sites. *Mol. Omics*. **2019**, *15*, 189–204. [[CrossRef](#)]
26. Chen, Z.; He, N.; Huang, Y.; Qin, W.T.; Liu, X.; Li, L. Integration of a deep learning classifier with a random forest approach for predicting malonylation sites. *Genom. Proteom. Bioinf.* **2018**, *16*, 451–459. [[CrossRef](#)]
27. Wu, M.; Yang, Y.; Wang, H.; Xu, Y. A deep learning method to more accurately recall known lysine acetylation sites. *BMC Bioinform.* **2019**, *20*, 49. [[CrossRef](#)]
28. Chaudhari, M.; Thapa, N.; Roy, K.; Newman, R.; Saigo, H.; Dukka, B. DeepRMethylSite: A deep learning based approach for prediction of argininemethylation sites in proteins. *Mol. Omics* **2020**. [[CrossRef](#)]
29. Thapa, N.; Hiroto, S.; Roy, K.; Newman, R.H.; Dukka, K. RF-MaloSite and DL-MaloSite: Two independent computational methods based on random forest (RF) and deep learning (DL) to predict malonylation sites. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 852–860.
30. Xu, H.; Zhou, J.; Lin, S.; Deng, W.; Zhang, Y.; Xue, Y. PLMD: An updated data resource of protein lysine modifications. *J. Gen. Genom.* **2017**, *44*, 243–250. [[CrossRef](#)]
31. Huang, Y.; Niu, B.; Gao, Y.; Fu, L.; Li, W. CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics* **2010**, *26*, 680–682. [[CrossRef](#)] [[PubMed](#)]
32. Dehzangi, A.; Paliwal, K.; Sharma, A.; Dehzangi, O.; Sattar, A. A combination of feature extraction methods with an ensemble of different classifiers for protein structural class prediction problem. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2013**, *10*, 564–575. [[CrossRef](#)] [[PubMed](#)]
33. Dehzangi, A.; López, Y.; Lal, S.P.; Taherzadeh, G.; Michaelson, J.; Sattar, A.; Sharma, A. PSSM-Suc: Accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction. *J. Theor. Biol.* **2017**, *425*, 97–102. [[CrossRef](#)] [[PubMed](#)]
34. Dehzangi, A.; Heffernan, R.; Sharma, A.; Lyons, J.; Paliwal, K.; Sattar, A. Gram-positive and gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into chou's general PseAAC. *J. Theor. Biol.* **2015**, *364*, 284–294. [[CrossRef](#)]
35. Chowdhury, S.Y.; Shatabda, S.; Dehzangi, A. iDNAprot-es: Identification of DNA-binding proteins using evolutionary and structural features. *Sci. Rep.* **2017**, *7*, 14938. [[CrossRef](#)]
36. Ahmad, M.W.; Shovan, S.; Arafat, M.E.; Sifat, M.H.R.; Hasan, M.A.M.; Shatabda, S. Improved performance of lysine glutarylation PTM using peptide evolutionary features. In Proceedings of the 3rd International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE) IEEE, RUET, Rajshahi, Bangladesh, 26–28 December 2019.
37. Sharma, A.; Lyons, J.; Dehzangi, A.; Paliwal, K.K. A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. *J. Theor. Biol.* **2013**, *320*, 41–46. [[CrossRef](#)]
38. Ahmed, M.W.; Arafat, M.E.; Shovan, S.M.; Uddin, M.; Osama, O.F.; Shatabda, S. Enhanced prediction of lysine propionylation sites using Bi-peptide evolutionary features resolving data imbalance. In Proceedings of the IEEE Region 10 Symposium (TENSYP 2020), Dhaka, Bangladesh, 5–7 April 2020.
39. Dehzangi, A.; López, Y.; Lal, S.P.; Taherzadeh, G.; Sattar, A.; Tsunoda, T.; Sharma, A. Improving succinylation prediction accuracy by incorporating the secondary structure via helix, strand and coil, and evolutionary information from profile bigrams. *PLoS ONE* **2018**, *13*, E0191900. [[CrossRef](#)]
40. Shatabda, S.; Saha, S.; Sharma, A.; Dehzangi, A. iPHLoc-ES: Identification of bacteriophage protein locations using evolutionary and structural features. *J. Theor. Biol.* **2017**, *435*, 229–237. [[CrossRef](#)]
41. Schaffer, A. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucl. Acids Res.* **2001**, *29*, 2994–3005. [[CrossRef](#)]
42. 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.* **2016**, *497*, 48–56. [[CrossRef](#)]
43. Chandra, A.; Sharma, A.; Dehzangi, A.; Ranganathan, S.; Jokhan, A.; Chou, K.C.; Tsunoda, T. PhoglyStruct: Prediction of phosphoglycerylated lysine residues using structural properties of amino acids. *Sci. Rep.* **2018**, *8*, 17923. [[CrossRef](#)] [[PubMed](#)]
44. Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; Chou, K.C. iPPBS-Opt: A sequence-Based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets. *Molecules* **2016**, *21*, 95. [[CrossRef](#)] [[PubMed](#)]

45. Jia, C.; Zuo, Y. S-SulfPred: A sensitive predictor to capture S-sulfenylation sites based on a resampling one-sided selection undersampling-synthetic minority oversampling technique. *J. Theor. Biol.* **2017**, *422*, 84–89. [[CrossRef](#)] [[PubMed](#)]
46. Hasan, M.M.; Khatun, M.S.; Kurata, H. A comprehensive review of in silico analysis for protein S-sulfenylation sites. *Protein Pept. Lett.* **2018**, *25*, 815–821. [[CrossRef](#)] [[PubMed](#)]
47. Sun, X.; Li, J.; Gu, L.; Wang, S.; Zhang, Y.; Huang, T.; Cai, Y.D. Identifying the characteristics of the hypusination sites using SMOTE and SVM algorithm with feature selection. *Curr. Proteom.* **2018**, *15*, 111–118. [[CrossRef](#)]
48. Zhang, Y.; Xie, R.; Wang, J.; Leier, A.; Marquez-Lago, T.T.; Akutsu, T.; Song, J. Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework. *Brief. Bioinform.* **2018**, *5*, 2185–2199. [[CrossRef](#)]
49. Iqbal, S.; Hoque, M.T. PBRpredict-Suite: A suite of models to predict peptide-recognition domain residues from protein sequence. *Bioinformatics* **2018**, *34*, 3289–3299. [[CrossRef](#)]
50. Zahiri, J.; Mohammad-Noori, M.; Ebrahimpour, R.; Saadat, S.; Bozorgmehr, J.H.; Goldberg, T.; Masoudi-Nejad, A. LocFuse: Human protein–protein interaction prediction via classifier fusion using protein localization information. *Genomics* **2014**, *104*, 496–503. [[CrossRef](#)]
51. Ismail, H.D.; Newman, R.H. RF-Hydroxysite: A random forest based predictor for hydroxylation sites. *Mol. Biosyst.* **2016**, *12*, 2427–2435. [[CrossRef](#)]
52. Jiao, Y.; Du, P. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quant. Biol.* **2016**, *4*, 320–330. [[CrossRef](#)]
53. Chen, C.Y.; Tang, S.L.; Chou, S.C.T. Taxonomy based performance metrics for evaluating taxonomic assignment methods. *BMC Bioinform.* **2019**, *20*, 310. [[CrossRef](#)] [[PubMed](#)]
54. Dehzangi, A.; Paliwal, K.; Lyons, J.; Sharma, A.; Sattar, A. Enhancing protein fold prediction accuracy using evolutionary and structural features. In Proceedings of the IAPR International Conference on Pattern Recognition Bioinformatics; 17–20 June 2013; pp. 196–207. [[CrossRef](#)]
55. Chen, Z.; Zhao, P.; Li, F.; Leier, A.; Marquez-Lago, T.T.; Wang, Y.; Webb, G.I.; Smith, A.I.; Daly, R.J.; Chou, K.C.; et al. iFeature: A python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* **2018**, *34*, 2499–2502. [[CrossRef](#)] [[PubMed](#)]
56. Liu, B. BioSeq-Analysis: A platform for DNA, RNA, and protein sequence analysis based on machine learning approaches. *Brief. Bioinform.* **2019**, *20*, 1280–1294. [[CrossRef](#)] [[PubMed](#)]
57. Liu, B.; Gao, X.; Zhang, H. BioSeq-Analysis2.0: An updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches. *Nucl. Acids Res.* **2019**, *47*, e127. [[CrossRef](#)] [[PubMed](#)]
58. Sharma, A.; Vans, E.; Shigemizu, D.; Boroevich, K.A.; Tsunoda, T. DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Sci. Rep.* **2019**, *9*. [[CrossRef](#)]

