

## **iProtGly-SS: A Tool to Accurately Predict Protein Glycation Site Using Structural-Based Features**

### Author

Dehzangi, I, Sharma, A, Shatabda, S

### Published

2022

### Book Title

Computational Methods for Predicting Post-Translational Modification Sites

### Version

Accepted Manuscript (AM)

### DOI

[10.1007/978-1-0716-2317-6\\_5](https://doi.org/10.1007/978-1-0716-2317-6_5)

### Rights statement

© 2022 Springer. This is the author-manuscript version of this paper. It is reproduced here in accordance with the copyright policy of the publisher. Please refer to the publisher's website for further information.

### Downloaded from

<http://hdl.handle.net/10072/418549>

### Griffith Research Online

<https://research-repository.griffith.edu.au>

# iProtGly-SS: A Tool to Accurately Predict Protein Glycation Site Using structural-based Features

Abdollah Dehzangi<sup>1,2,\*</sup>, Alok Sharma<sup>3,4,5,6</sup>, Swakkhar Shatabda<sup>7</sup>

<sup>1</sup> Department of Computer Science, Rutgers University, Camden, NJ 08102, USA

<sup>2</sup> Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA

<sup>3</sup> Institute for Integrated and Intelligent Systems, Griffith University, Brisbane, QLD 4111, Australia

<sup>4</sup> Department of Medical Science Mathematics, Tokyo Medical and Dental University (TMDU), Tokyo 113-8510, Japan

<sup>5</sup> Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045, Japan

<sup>6</sup> STEMP, University of the South Pacific, Suva, Fiji

<sup>7</sup> Department of Computer Science and Engineering, United International University, Dhaka 1212, Bangladesh

\* Author to whom correspondence should be addressed.

## Abstract:

Post Translational Modification (PTM) is an important biological mechanism to promote functional diversity among the proteins. So far, a wide range of PTMs has been identified. Among them, Glycation is considered as one of the most important PTMs. Glycation is associated with different neurological disorders including Parkinson and Alzheimer. It is also shown to be responsible for different diseases, including vascular complications of diabetes mellitus. Despite all the efforts have been made so far, the prediction performance of Glycation sites using computational methods remains limited. Here we present a newly developed machine learning tool called iProtGly-SS that utilizes sequential and structural information as well as Support Vector Machine (SVM) classifier to enhance lysine Glycation site prediction accuracy. The performance of iProtGly-SS was investigated using the three most popular benchmarks used for this task. Our results demonstrate that iProtGly-SS is able to achieve 81.61%, 93.62%, and 92.95% prediction accuracies on these benchmarks, which are significantly better than those results reported in the previous studies. iProtGly-SS is implemented as a web-based tool which is publicly available at <http://brl.uiu.ac.bd/iprotgly-ss/>.

## Keywords

Post Translational Modification, Protein Glycation, Feature Selection, Support Vector Machine, Evolutionary Features, Structural Features

## Introduction

Post translation modification (PTM) of proteins is an important mechanism to increase proteomic diversity by covalently attaching small macromolecules to selected amino acid residues after the translation process. PTM plays critical roles in biological reactions ranging from amino acid degradation, cell signaling, regulating the folding process of proteins or their interaction with other molecules to disease causing interactions such as in Parkinson, and many cancers. Therefore, the identification of PTM sites provides detailed insights into the function of proteins and their biological interactions [1]. Proteins undergo a wide range of PTMs. Among them, Glycation is among the most important ones. Glycation is defined as the covalent bonding of sugar molecules with proteins [1,2]. In this process, sugar molecules including glucose, fructose or their other derivatives interact with lysine (K) and arginine (R) residues to form Amadori ketoamine. In the next step, Amadori ketoamine undergoes degradation to form advanced glycation end products (AGEs). The Glycation and its related by-products have been shown to be responsible for many diseases, including neuro-degenerative disorders, such as Alzheimer, and Parkinson [3, 4]. It is also shown to be responsible for complications in vascular complications of diabetes mellitus. Glycation is also shown related in the aging process [5].

Glycation can be detected using experimental methods [6]. Among the experimental methods, Liquid chromatography with mass spectrometry is considered the most successful approach to detect this and other PTMs. However, this method is complicated, expensive, and time-consuming [7, 8]. Hence, it is unable to meet the exponential growth of protein databanks. Therefore, the implementation of fast and accurate computational methods is considered as a feasible alternative to experimental methods to detect Glycation and other PTMs [9]. Among different computational methods, machine learning based approaches have emerged as the most effective and accurate methods to tackle this problem [10].

During the past decade, several machine learning-based methods have been proposed to tackle this problem [11-13]. Among them, GlyNN [14], PreGly [15], and Gly-PseAAC [16] are the most notable ones which demonstrated the best results for this task. GlyNN was proposed in [14]. To build this model, they used the composition of the amino acids around a given lysine as the input features. Consequently, they used Artificial Neural Network (ANN) as their employed classification to predict lysine Glycation sites. Later on, [15] proposed a new method called PreGly to solve this problem. To build this model, they used several sequence-based features including amino acid occurrence frequency, amino acid factors, and the composition of k-spaced amino acid pairs (CKSAAP). They also used Support Vector Machine (SVM) as their classification technique. More recently, [16] used propensity-based features and SVM as the classifier to proposed Gly-PseAAC. Similar to GlyNN and PreGly, Gly-PseAAC also relied solely on sequence-based features to solve this problem. Despite achieving promising results, the lysine Glycation prediction accuracy remains limited.

In a recent study, we have developed iProtGly-SS, a machine learning based model that uses the combination of structural and sequential based information extracted from the protein sequence to accurately predict lysine Glycation sites. We also used SVM as the classification technique to build this model. We investigated the effectiveness of iProtGly-SS on the three most popular benchmarks used for this task. Our results showed that this model is able to significantly outperform previous studies found in the literature to solve this problem. iProtGly-SS achieves 81.61%, 93.62%, and 92.95% accuracies on the employed benchmarks, which in all cases are over 10% better than the results reported in previous studies. It is important to note that iProtGly-SS is the first machine learning-based method to achieve over 90% prediction accuracy for this task on two of the benchmarks used in this study [16].

It is also worth mentioning that iProtGly-SS performance and its achieved results are still better or comparable to more recent machine learning models proposed to predict lysine Glycation sites on the employed benchmarks [17-22].

## **Materials and Algorithm**

In this section, the datasets used for our experiments and the method used to develop iProtGly-SS are explained in detail.

### **Datasets**

For the experimental analysis, we used the three most popular datasets that are widely used for the lysine Glycation prediction task. The first dataset (Dataset A) was collected from CPLM (<http://cplm.biocuckoo.org/>) [10]. It consists of 72 proteins containing 323 Glycation sites. We then used CD-Hit to remove proteins with over 40% sequence similarity. The resulting dataset contains significantly more negative samples compared to positive samples. To avoid bias towards negative samples, we balanced the dataset using random sampling with the ratio of 2:1 (two negative samples for every positive sample). As a result, the final dataset contained 223 positive and 446 negative samples.

The second dataset (Dataset B) was introduced in [14]. This dataset was generated manually by inspecting a large number of proteins that undergone Glycation. The resulting dataset consists of 20 proteins with 89 Glycation sites and 126 non-Glycation sites.

The third dataset (Dataset C) was introduced in [15], which was manually collected from Uniprot [23, 24]. This dataset was originally used to compare the performance of PreGly with other methods. This dataset contains 82 Glycation sites and 117 non-Glycation sites.

### **Algorithm**

To build iProtGly-SS we first studied the impact of several features to investigate their effectiveness and select those with the most important discriminatory information for the lysine Glycation site prediction task. Here we studied 9 feature groups based on four properties:

propensity, composition, secondary structure, and physicochemical properties of the amino acids. We also studied different window sizes to extract our features from the neighboring amino acids. Among them, using a window size of 35 (17 upstream + 17 downstream + 1 central lysine) demonstrated the best results and used to build our model.

Propensity was first used in [16] to build Gly-PseAAC to predict lysine Glycation sites. It was shown in this study that propensity can provide important local information about the interaction of neighboring amino acids [25]. Therefore, it is considered as a sequence-based feature and investigated in this study.

Another sequence-based feature investigated in this study was the composition of amino acids. The composition of the amino acids provides information about the occurrence of the neighboring amino acids to a lysine. Considering 20 standard amino acids plus an unknown amino acid 'X', we extracted 21 features based on the composition property.

On the other hand, we used predicted secondary structure using SPIDER 3.0 to extract structure-based features [26, 27]. Structural information provides information about the local structure of the proteins and how amino acids interact in 3D structure [28, 29]. Structural information has been widely used to predict different types of PTMs, including SUMOylation, Malonylation, Pupylation, phosphoglycerylation, succinylation, and Glycosylation and obtained promising results [30-39].

Finally, we also extracted features based on different physicochemical properties of the amino acids. These features can provide important information about different attributes of the amino acids and how they interact. We studied 6 different physicochemical attributes of the amino acids that were used in [40], namely, polarity, polarizability, normalized van der Waals volume, secondary structure tendency, charge, and solvent accessibility. As a result, we extract 6 feature groups based on these properties.

We also studied a wide range of classifiers which among them, Support Vector Machine (SVM), attained the best results and therefore used as the primary classifier to build iProtGly-SS. SVM is a non-parametric classification technique that aims at finding the maximum marginal hyperplane to minimize the prediction error [41].

We then investigated the impact of each feature group to select those with the highest discriminatory information for lysine Glycation prediction task [15]. To do this, we design an incremental subset feature group selection approach, named INFUSE: INcremental Feature groUp Selection. INFUSE is designed similar to incremental feature selection [42]. In this model, we first investigate the impact of each feature group, separately. We then choose the one with the best performance. In the next step, we add the remaining feature groups to the one selected in the first step and repeat the experiment to identify the best performance. Consequently, we then choose the combination of two with the best performance. We repeat this process for all 9 feature groups to determine the best combination of feature groups for the lysine Glycation site prediction task.

Using INFUSE, we identify that we can obtain the best results by using amino acid composition, secondary structure, and polarity. Hence, we build iProtGly-SS using the combination of these three feature groups and SVM as our classification method.

Table 1 demonstrates the results achieved using iProtGly-SS on our employed datasets compared to PreGly, Gly-PseAAC, and GlyNN using 3-fold, 10-fold, and leave-one-out cross-

validation as our validation methods [14-16]. We compared our results using different evaluation measurements such as Accuracy, Area Under the Curve (AUC), Sensitivity, Specificity, and Matthews Correlation Coefficient (MCC).

Dataset	Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	MCC
Dataset A	iProtGly-SS <sup>a</sup>	<b>81.61±4.7</b>	<b>92.38 ± 1.1</b>	<b>60.09±2.3</b>	<b>0.8676±0.04</b>	<b>0.5619±0.13</b>
	iProtGly-SS <sup>c</sup>	80.72	93.05	56.05	0.8816	0.5474
	Gly-PseAAC <sup>a</sup>	68.69	57.48	74.30	0.7199	0.3166
	Gly-PseAAC <sup>c</sup>	68.91	58.74	73.99	0.7258	0.3198
Dataset B	iProtGly-SS <sup>a</sup>	<b>93.62±4.1</b>	<b>93.68±2.8</b>	<b>93.42±1.6</b>	<b>0.9773±0.02</b>	<b>0.8783±0.03</b>
	iProtGly-SS <sup>c</sup>	94.15	95.79	92.11	0.9841	0.8815
	GlyNN <sup>b</sup>	79.50	78.65	80.15	0.77	0.58
	PreGly <sup>a</sup>	85.51	71.06	95.85	---	0.70
	Gly-PseAAC <sup>a</sup>	68.12	56.06	80.17	0.7705	0.38
Dataset C	iProtGly-SS <sup>a</sup>	<b>92.95±6.0</b>	<b>95.73±1.2</b>	<b>89.02±3.8</b>	<b>0.9672±0.03</b>	<b>0.8543±0.12</b>
	iProtGly-SS <sup>c</sup>	92.96	95.73	89.02	0.9635	0.8545
	PreGly <sup>a</sup>	79.92	64.20	91.57	---	0.5849

<sup>a</sup> 10-fold cross-validation

<sup>b</sup> 3-fold cross-validation

<sup>c</sup> Leave-one-out cross-validation

Table 1: Results achieved using iProtGly-SS on our employed datasets compared to PreGly, Gly-PseAAC, and GlyNN.

As shown in Table 1, iProtGly-SS is able to significantly outperform PreGly, GlyNN, and Gly-PseAAC on all three datasets with respect to all the evaluation measurement. In particular, iProtGly-SS achieves, 81.61%, 93.62%, and 92.95% prediction accuracies which are 12.92%, 8.11%, and 13.03% better than those reported by previous studies in Datasets A, B, and C, respectively. It is also important to note that iProtGly-SS is able to achieve significantly better sensitivity than those reported in other studies [14-16]. This highlights the effectiveness of this method to predict lysine Glycation sites compared to other methods.

## Webserver

1. iProtGly-SS is implemented as an online tool which is publicly available at <http://brl.uiu.ac.bd/iprotgly-ss/>. The input to the webserver is in the SPD3 format (the output of SPIDER 3.0). The partial input SPD3 sample is shown in Figure 1. It is important to note that iProtGly-SS can work with any other predicted secondary structure output file with is in the same format of SPIDER 3.0 [27]. For example, the output of SPIDER 2.0, an earlier version of SPIDER 3.0, which is also used for predicting protein local structure, is in the similar format [28, 29]. Hence, it can also be used as an input to iProtGly-SS. However, for such a case, the performance will not be guaranteed to be exactly the same as what was reported since the local structure prediction accuracy of different tools might be different, which directly impacts the iProtGly-SS performance.

2. iProtGly-SS uses the secondary structure predicted values (P(C), P(E), P(H)) from SPD3 to extract related features as it was mentioned earlier in the Materials and Algorithm Section.

#	AA	SS	ASA	Phi	Psi	Theta(i-1=>i+1)	Tau(i-2=>i+1)	P(C)	P(E)	P(H)
1	M	C	137.7	-101.4	137.8	118.7	-148.8	0.988	0.011	0.001
2	A	C	83.3	-82.7	136.1	113.9	-129.6	0.978	0.021	0.001
3	T	E	45.7	-110.3	139.5	122.6	-158.2	0.364	0.615	0.002
4	K	E	81.8	-112.6	134.5	120.7	-155.5	0.422	0.545	0.005
5	A	E	7.9	-118.7	141.0	126.6	-151.9	0.190	0.799	0.012
6	V	E	14.7	-125.6	137.0	127.1	-158.1	0.034	0.965	0.005
7	C	E	4.7	-119.6	139.2	127.2	-161.1	0.060	0.909	0.040
8	V	E	15.5	-116.0	127.4	118.7	-158.8	0.051	0.946	0.007
9	L	E	14.2	-105.9	133.7	119.2	-156.1	0.098	0.892	0.004
10	K	C	96.4	-120.6	136.1	124.6	-167.9	0.493	0.491	0.004
11	G	C	16.9	-147.6	174.8	129.3	171.8	0.655	0.365	0.005
12	D	C	110.7	-105.5	71.1	109.4	-105.3	0.806	0.184	0.013
13	G	C	43.4	-167.8	-168.7	136.7	32.2	0.943	0.053	0.008
14	P	C	105.4	-76.2	134.1	112.4	-50.9	0.913	0.075	0.012
15	V	E	22.8	-112.3	133.3	120.8	162.9	0.306	0.676	0.005
16	Q	C	110.5	-121.7	138.8	125.6	-165.3	0.530	0.462	0.002

Figure 1: partial SPD3 matrix sample, which is produced as the output of SPIDER 3.0. iProtGly-SS uses predicted secondary structure (P(C), P(E), P(E)) to extract structural based features.

3. Figure 2 demonstrates iProtGly-SS web-based tool and its output for a sample sequence (Q72FV4). As shown in this figure, user can upload the SPD3 for a given protein and submit it. The results will be then shown at the bottom of the page. For each lysine (shown as 'K') in the input protein sequence, iProtGly-SS predict if it undergoes Glycation or not. It also prints the Lysine and its neighboring amino acids with respect to the window size that is adopted for this model (which is equal to 35), the position of the lysine in the protein sequence, and if it is a Glycation site or not. iProtGly-SS is fast, very easy to use, and its generated output is straightforward and easy to interpret.

The screenshot shows the iProtGly-SS web interface. At the top is the Bio-informatics Research Lab header with navigation links for ABOUT, PUBLICATIONS, RESOURCES, and CONTACT. The main content area is divided into two sections: 'iProtGly-SS' and 'Predict Glycation:'. The 'iProtGly-SS' section describes the tool as 'Identifying Protein Lysine Glycation Sites Using Secondary Structural Information' and includes a 'Learn more' button. The 'Predict Glycation:' section has an upload input field for SPD3 files in .s3z format, a 'Choose file' button, and a 'Submit' button. Below this is the 'RESULT' section, which displays a table of prediction results for the protein Q72FV4. The table has three columns: 'Protein', 'Location of 'K'', and 'Prediction on Glycation'. The results show that lysine residues at positions 56, 106, 115, 132, 139, 166, and 184 are predicted to undergo glycation, while those at positions 106, 115, 132, 139, 166, and 184 are predicted not to.

Protein	Location of 'K'	Prediction on Glycation
FLAGGEPFNLSLRGK <sup>56</sup> LLIENVASLXGTTV	56	Glycation <b>Won't</b> Happen
VLGFDPCNQFCHQENAK <sup>106</sup> HEILNCLKYVRDCC	106	Glycation <b>Won't</b> Happen
GHQEMAKNEETLNCLK <sup>115</sup> YVRDGGGFEINFMFLF	115	Glycation <b>Won't</b> Happen
VRPGGGFEPNFMFLFK <sup>132</sup> CEVNGEKAHPLFAFL	132	Glycation <b>Won't</b> Happen
EPNFMIFKCEVNGRKA <sup>139</sup> HLPLFAFLREVLPPTP	139	Glycation <b>Will</b> Happen
LPTFSDDATALMTDFK <sup>166</sup> LTWSEVCHNDSWN	166	Glycation <b>Will</b> Happen
TWSEVCHNDSWNFK <sup>184</sup> ELVGDGVEVRRYSR	184	Glycation <b>Won't</b> Happen

Figure 2: iProtGly-SS and its prediction result for the sample sequence “Q72FV4”

4. To generate SPD3, SPIDER 3.0 requires Position Specific Scoring Matrix (PSSM), which is generated as the output of PSI-BLAST. Generating PSSM is a time taking process (depend on the available computational resources). Hence, to maintain the simplicity of the usage, iProtGly-SS uses SPD3 directly as an input. It also makes it faster. iProtGly-SS produces the output in less than 5 minutes, depending on the length of the input protein. It also makes it independent from those tools (SPIDER 3.0 and PSI-BLAST). iProtGly-SS is written in python, which also makes it easier to adjust and maintain the version control. User can write scripts to automatically upload the SPD3 files to generate the prediction output for more than one input protein sequence.

## Notes

1. The query sequence should be in SPD3 format. It makes the preprocessing before submitting the job harder. We aim to adjust iProtGly-SS in its future versions to get the raw protein sequence as the input and generate SPD3 by itself. It can definitely make it easier for users to work. However, it also makes it much slower. As it was mentioned earlier, generating PSSM is a time taking process. In addition, it makes the version control harder since SPIDE 3.0 and PSI-BLAST consistency also needs to be checked.

2. iProtGly-SS is just available as a web-based tool. Hence, to generate the Glycation prediction sites, proteins are required to be submitted online and one by one. We aim to implement future versions of iProtGly-SS as a standalone tool as well. It will make it easier to use it iteratively for more than one protein in an offline mode.

3. iProtGly-SS also requires regular version control to make sure it is consistent with new versions of Python and the libraries that are used to implement it.

4. As mentioned earlier, iProtGly-SS can work with the output of different tools that predicts the local structure of the proteins and secondary structure in specific as long as they are formatted similar to SPD3. However, its output will no longer be similar and will depend on the prediction accuracy of that tool for the protein secondary structure prediction task.

## References

[1] Khoury, G.A., Baliban, R.C. and Floudas, C.A., 2011. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Scientific reports*, 1(1), pp.1-5.

[2] Wang, D., Liu, D., Yuchi, J., He, F., Jiang, Y., Cai, S., Li, J. and Xu, D., 2020. MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic acids research*, 48(W1), pp.W140-W146.

[3] Castellani, R.J., Harris, P.L., Sayre, L.M., Fujii, J., Taniguchi, N., Vitek, M.P., Founds, H., Atwood, C.S., Perry, G. and Smith, M.A., 2001. Active glycation in neurofibrillary pathology of Alzheimer disease: N $\epsilon$ -(carboxymethyl) lysine and hexitol-lysine. *Free Radical Biology and Medicine*, 31(2), pp.175-180.

- [4] Choudhary, C., Kumar, C., Gnad, F., Nielsen, M.L., Rehman, M., Walther, T.C., Olsen, J.V. and Mann, M., 2009. Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*, 325(5942), pp.834-840.
- [5] Ulrich, P. and Cerami, A., 2001. Protein glycation, diabetes, and aging. *Recent progress in hormone research*, 56(1), pp.1-22.
- [6] Tatjewski, M., Kierczak, M. and Plewczynski, D., 2017. Predicting post-translational modifications from local sequence fragments using machine learning algorithms: Overview and best practices. *Prediction of Protein Secondary Structure*, pp.275-300.
- [7] Thornalley, P.J., Battah, S., Ahmed, N., Karachalias, N., Agalou, S., Babaei-Jadidi, R. and Dawnay, A., 2003. Quantitative screening of advanced glycation endproducts in cellular and extracellular proteins by tandem mass spectrometry. *Biochemical Journal*, 375(3), pp.581-592.
- [8] Zhang, Q., Ames, J.M., Smith, R.D., Baynes, J.W. and Metz, T.O., 2009. A perspective on the Maillard reaction and the analysis of protein glycation by mass spectrometry: probing the pathogenesis of chronic disease. *Journal of proteome research*, 8(2), pp.754-769.
- [9] Eisenhaber, B. and Eisenhaber, F., 2010. Prediction of posttranslational modification of proteins from their amino acid sequence. *Data Mining Techniques for the Life Sciences*, pp.365-384.
- [10] Liu, Z., Wang, Y., Gao, T., Pan, Z., Cheng, H., Yang, Q., Cheng, Z., Guo, A., Ren, J. and Xue, Y., 2014. CPLM: a database of protein lysine modifications. *Nucleic acids research*, 42(D1), pp.D531-D536.
- [11] Xue, Y., Liu, Z., Cao, J. and Ren, J., 2011. Computational prediction of post-translational modification sites in proteins. *Systems and computational biology-molecular and cellular experimental systems*, 5772(6), p.18559.
- [12] Liu, Y., Wang, M., Xi, J., Luo, F. and Li, A., 2018. PTM-ssMP: a web server for predicting different types of post-translational modification sites using novel site-specific modification profile. *International journal of biological sciences*, 14(8), p.946.
- [13] Wang, D., Liu, D., Yuchi, J., He, F., Jiang, Y., Cai, S., Li, J. and Xu, D., 2020. MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic acids research*, 48(W1), pp.W140-W146.
- [14] Johansen, M.B., Kiemer, L. and Brunak, S., 2006. Analysis and prediction of mammalian protein glycation. *Glycobiology*, 16(9), pp.844-853.
- [15] Liu, Y., Gu, W., Zhang, W. and Wang, J., 2015. Predict and analyze protein glycation sites with the mRMR and IFS methods. *BioMed research international*, 2015.
- [16] Xu, Y., Li, L., Ding, J., Wu, L.Y., Mai, G. and Zhou, F., 2017. Gly-PseAAC: identifying protein lysine glycation through sequences. *Gene*, 602, pp.1-7.
- [17] Yu, J., Shi, S., Zhang, F., Chen, G. and Cao, M., 2019. PredGly: predicting lysine glycation sites for Homo sapiens based on XGboost feature optimization. *Bioinformatics*, 35(16), pp.2749-2756.

- [18] Chen, J., Zhang, C., Yang, R. and Zhang, L., 2019, June. A New Machine Learning Based Framework to Identify Protein Glycation Sites Using Comprehensive Features and the mRMR Method. In *2019 Chinese Control And Decision Conference (CCDC)* (pp. 3605-3609). IEEE.
- [19] Chen, J., Yang, R., Zhang, C., Zhang, L. and Zhang, Q., 2019. DeepGly: A Deep Learning Framework With Recurrent and Convolutional Neural Networks to Identify Protein Glycation Sites From Imbalanced Data. *IEEE Access*, 7, pp.142368-142378.
- [20] Reddy, H.M., Sharma, A., Dehzangi, A., Shigemizu, D., Chandra, A.A. and Tsunoda, T., 2019. GlyStruct: glycation prediction using structural properties of amino acid residues. *BMC bioinformatics*, 19(13), pp.55-64.
- [21] Khanum, S., Ashraf, M.A., Karim, A., Shoaib, B., Khan, M.A., Naqvi, R.A., Siddique, K. and Alswaitti, M., Gly-LysPred: Identification of Lysine Glycation Sites in Protein Using Position Relative Features and Statistical Moments via Chou's 5 Step Rule.
- [22] Abrahams, J.L., Taherzadeh, G., Jarvas, G., Guttman, A., Zhou, Y. and Campbell, M.P., 2020. Recent advances in glycoinformatic platforms for glycomics and glycoproteomics. *Current opinion in structural biology*, 62, pp.56-69.
- [23] UniProt Consortium, 2015. UniProt: a hub for protein information. *Nucleic acids research*, 43(D1), pp.D204-D212.
- [24] Lee, T.Y., Huang, H.D., Hung, J.H., Huang, H.Y., Yang, Y.S. and Wang, T.H., 2006. dbPTM: an information repository of protein post-translational modification. *Nucleic acids research*, 34(suppl\_1), pp.D622-D627.
- [25] Tang, Y.R., Chen, Y.Z., Canchaya, C.A. and Zhang, Z., 2007. GANNPhos: a new phosphorylation site predictor based on a genetic algorithm integrated neural network. *Protein Engineering, Design & Selection*, 20(8), pp.405-412.
- [26] Heffernan, R., Dehzangi, A., Lyons, J., Paliwal, K., Sharma, A., Wang, J., Sattar, A., Zhou, Y. and Yang, Y., 2016. Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. *Bioinformatics*, 32(6), pp.843-849.
- [27] Heffernan, R., Yang, Y., Paliwal, K. and Zhou, Y., 2017. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics*, 33(18), pp.2842-2849.
- [28] Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Yang, Y. and Zhou, Y., 2015. Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning. *Scientific reports*, 5(1), pp.1-11.
- [29] Yang, Y., Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A. and Zhou, Y., 2017. Spider2: A package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. In *Prediction of protein secondary structure* (pp. 55-63). Humana Press, New York, NY.

- [30] Islam, M.M., Saha, S., Rahman, M.M., Shatabda, S., Farid, D.M. and Dehzangi, A., 2018. iProtGly-SS: Identifying protein glycation sites using sequence and structure based features. *Proteins: Structure, Function, and Bioinformatics*, 86(7), pp.777-789.
- [31] López, Y., Dehzangi, A., Reddy, H.M. and Sharma, A., 2020. C-iSUMO: A sumoylation site predictor that incorporates intrinsic characteristics of amino acid sequences. *Computational Biology and Chemistry*, 87, p.107235.
- [32] Chandra, A., Sharma, A., Dehzangi, A., Shigemizu, D. and Tsunoda, T., 2019. Bigram-PGK: phosphoglycerylation prediction using the technique of bigram probabilities of position specific scoring matrix. *BMC molecular and cell biology*, 20(2), pp.1-9.
- [33] Dipta, S.R., Taherzadeh, G., Ahmad, M.W., Arafat, M.E., Shatabda, S. and Dehzangi, A., 2020. SEMal: Accurate protein malonylation site predictor using structural and evolutionary information. *Computers in Biology and Medicine*, 125, p.104022.
- [34] Singh, V., Sharma, A., Dehzangi, A. and Tsunoda, T., 2020. PupStruct: Prediction of Pupylyated Lysine Residues Using Structural Properties of Amino Acids. *Genes*, 11(12), p.1431.
- [35] López, Y., Sharma, A., Dehzangi, A., Lal, S.P., Taherzadeh, G., Sattar, A. and Tsunoda, T., 2018. Success: evolutionary and structural properties of amino acids prove effective for succinylation site prediction. *BMC genomics*, 19(1), pp.105-114.
- [36] Arafat, M., Ahmad, M., Shovan, S.M., Dehzangi, A., Dipta, S.R., Hasan, M., Mehedi, A., Taherzadeh, G., Shatabda, S. and Sharma, A., 2020. Accurately Predicting Glutarylation Sites Using Sequential Bi-Peptide-Based Evolutionary Features. *Genes*, 11(9), p.1023.
- [37] Sharma, A., Lysenko, A., López, Y., Dehzangi, A., Sharma, R., Reddy, H., Sattar, A. and Tsunoda, T., 2019. HseSUMO: SUMOylating site prediction using half-sphere exposures of amino acids residues. *BMC genomics*, 19(9), pp.1-7.
- [38] Taherzadeh, G., Dehzangi, A., Golchin, M., Zhou, Y. and Campbell, M.P., 2019. SPRINT-Gly: Predicting N-and O-linked glycosylation sites of human and mouse proteins by using sequence and predicted structural properties. *Bioinformatics*, 35(20), pp.4140-4146.
- [39] Chandra, A., Sharma, A., Dehzangi, A., Ranganathan, S., Jokhan, A., Chou, K.C. and Tsunoda, T., 2018. PhoglyStruct: prediction of phosphoglycerylated lysine residues using structural properties of amino acids. *Scientific reports*, 8(1), pp.1-11.
- [40] Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I. and Kim, S.H., 1999. Recognition of a protein fold in the context of the SCOP classification. *Proteins: Structure, Function, and Bioinformatics*, 35(4), pp.401-407.
- [41] Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine learning*, 20(3), pp.273-297.
- [42] Liu, H. and Setiono, R., 1998. Incremental feature selection. *Applied Intelligence*, 9(3), pp.217-230.