

Predicting Phosphoglycerylation with Transformer features and Deep Learning

Abel Chandra^{1,*}, Alok Sharma^{1,2,*}, Iman Dehzangi^{3,4}, Tatsuhiko Tsunoda^{2,5,6,&}, and Abdul Sattar^{1,&}

¹Institute for Integrated and Intelligent Systems, Griffith University, Brisbane, Australia

²Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

³Department of Computer Science, Rutgers University, Camden, NJ, USA

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

⁵Laboratory for Medical Science Mathematics, Department of Biological Sciences, School of Science, The University of Tokyo, Tokyo, Japan

⁶Laboratory for Medical Science Mathematics, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan

*corresponding authors: abel.chandra@griffithuni.edu.au, alok.fj@gmail.com

& Last authors

Abstract— Understanding protein sequences can advance treatments for various diseases. However experimentally obtaining this information is laborious, time-consuming, and expensive. Traditional machine learning techniques, like support vector machine, random forest and logistic regression, offer potential to fast-track this process but are sometimes limited by data complexity. Deep learning algorithms, in contrast, tend to yield higher performance. In this study, we employed a convolutional neural network to predict protein phosphoglycerylation. Features were extracted from pre-trained transformer models and compared with conventional features, such as evolutionary information and physicochemical/biochemical properties. Our results indicate significant performance improvements across all feature types, with the combination of transformer-based features and the convolutional neural network being especially effective. This methodology holds potential for other protein property prediction tasks. Our software and datasets used in this study are publicly available at <https://github.com/abelavit/DL-Phosphoglycerylation-Prediction.git>.

Keywords— machine learning, deep learning, transformer network, convolutional neural network, proteomics, phosphoglycerylation.

I. INTRODUCTION

Proteomics, the study of all proteins in biological systems, is evolving into a data-rich science due to modern biotechnological advancements [1, 2]. Historically, protein property identification predominantly relied on experimental methods, such as liquid chromatography and mass spectrometry systems. These methods necessitate intricate bioinformatic analysis pipelines, making them challenging, expensive, and time-consuming [3-5]. However, the surge in available data has paved the way for deep learning (DL) technologies, which are increasingly adopted to predict properties of uncharted protein sequence, offering immense value to the scientific community [6-9].

DL empowers computer systems to discern patterns from input data to draw inferences. Unlike traditional machine learning (ML) techniques like random forests (RF) and support vector machines (SVM), DL algorithms inherently learn from data, negating the need for manual feature engineering. Rooted in representation learning, DL employs artificial neural networks that emulate human brain learning processes. Across domains, including computer

vision, natural language processing, and bioinformatics, DL models consistently outperformed conventional ML techniques [10].

Recently, DL architectures like transformer networks [11] (from natural language processing) and convolutional neural networks (CNNs) [12] have made their mark in bioinformatics [13-19]. Transformer networks, equipped with attention mechanisms, grasp inter-positional information in input sequences and excel in tasks like language translation, due to its innovative architecture [11, 20]. Moreover, the CNN architecture captures and preserves spatial hierarchies in sequential data (e.g., protein sequences) thereby extracting features from data that lead to superior performance when compared to the traditional ML methods [21].

Building on this foundation, our study seeks to harness the potency of CNNs, using them as classifiers on features extracted from several tools, including pre-trained transformer models. The goal is to enhance lysine phosphoglycerylation prediction in protein sequences. This approach builds upon and extends our previous work [22], where we utilized these features to train conventional ML classifiers like RF, SVM, logistic regression (LR), and light gradient-boosting machine (LightGBM) to determine the most informative features for phosphoglycerylation prediction.

A. Literature Review

There has been a rise in the study of phosphoglycerylation using computational techniques in the recent years. As a result, a number of predictors have been developed for the prediction of these sites in the protein sequences. One of the earliest predictors is called Phogly-PseAAC [23] which employed pseudo amino acid composition as features to train a k-nearest neighbours algorithm. CKSAAP_Phoglysite predictor [24] was introduced next and it utilized the composition of k-spaced amino acid pairs for feature extraction and trained a fuzzy SVM. Similar feature extraction as CKSAAP_Phoglysite was employed by the PhoglyPred method [25] and the SVM algorithm was trained to build the classifier. Later, iPGK-PseAAC predictor [26] was proposed based on SVM and it used amino acid pairwise couplings as the features. Following these works, Chandra et al developed four separate methods called PhoglyStruct [27], EvolStruct-Phogly [28], Bigram-PGK [29], and RAM-PGK [4].

PhoglyStruct is a multilayer perceptron-based method that used the protein structural features, while EvolStruct-Phogly, Bigram-PGK, and RAM-PGK are all SVM-based methods that used a combination of structural and evolutionary (sequence alignment) features, evolutionary features, and sequence-based information, respectively. iDPGK [30] is a work that was proposed during the same period, in which a SVM-based predictor was employed and the composition of amino acids was used as the feature set.

Most recently, there have been three major phosphoglycerylation predictors proposed. These are predPhogly-Site [31], PLP_FS [32], and BERT_PLPS [33]. The predPhogly-Site method used probabilistic sequence-coupling information to train the SVM algorithm. The PLP_FS predictor used features generated by sequence-based feature extraction methods to fit the SVM algorithm. Finally, the BERT_PLPS predictor used the amino acid sequence features and a transformer-based network (BERT) as the prediction model. From all these proposed methods for the phosphoglycerylation prediction, it can be seen that, except for the most recent method (BERT_PLPS), they rely on traditional ML algorithms. Moreover, the BERT_PLPS method does not incorporate the pipeline to harness the true potential of DL. As a result, there is a huge scope for further improvement of the phosphoglycerylation site prediction in protein sequences.

B. Contributions of the Paper

The major contribution of this paper is to showcase the potential of DL for predicting protein phosphoglycerylation. The major objective is to boost the prediction capability of current methods by the use of pre-trained transformer-based features with architectures such as CNN for classification. A summary of the contributions is as follows:

- Identify a suitable CNN architecture to work with feature vectors.
- Analyse the result obtained for the combination of the feature and the CNN architecture.
- Highlight how the protein sequence representations from the pre-trained transformer models are superior to the traditional features (physicochemical/biochemical properties and evolutionary information).
- Point out the CNN's feature extraction capability for effective phosphoglycerylation prediction and its potential for improving the performance of various protein-related prediction tasks.

C. Paper Organization

The rest of the paper is organized as follows. Section II gives an overview of what protein phosphoglycerylation is and its importance. Section III gives the details of the dataset used in this work, how the features were extracted, how each sample was constructed, data balancing, the evaluation metrics used, and the CNN architecture employed in this work. Section IV provides the results and discussion, and finally section V concludes this work with a hint of future work.

II. PHOSPHOGLYCERYLATION

Phosphoglycerylation is a type of post-translational modification (PTM). Some examples of other types of PTM include methylation [34], glycation [35], acetylation [36], crotonylation [37], phosphorylation [38], succinylation [39], and sumoylation [40]. PTM is the enzymatic change to protein sequences which takes place after the protein translation in the ribosome [29]. It plays a crucial role in biological processes, such as cell functions, to regulate cellular plasticity and dynamics [23]. Phosphoglycerylation is a reversible process that deals with the modification of lysine amino acid residue in the protein sequences. This modification is linked to glycolytic pathways and glucose metabolism [41] and is implicated in a variety of human diseases, such as neurodegenerative disorders [42], coronary heart disease [43], rheumatoid arthritis [44], and multiple sclerosis [45]. Hence, the identification of phosphoglycerylated sites in protein sequences can present valuable information for biomedical research.

III. EXPERIMENTAL SETUP

A. Dataset

The phosphoglycerylation dataset used in this work is adopted from [22]. The protein sequences were originally taken from the Protein Lysine Modification Database (PLMD) available at <http://plmd.biocuckoo.org>. The Cd-hit tool [46] was used to disregard the protein sequences with a sequence similarity of 40% or higher. As a result, the dataset comprised 91 sequences which had a total of 3360 lysine residues. Out of these lysine residues, 111 were found to be experimentally annotated as phosphoglycerylated sites (positive samples), while the remaining 3249 were non-phosphoglycerylated sites (potential negative samples). The dataset composition is shown in Table I.

TABLE I. EXECUTIVE SUMMARY OF THE PHOSPHOGLYCERYLATION DATASET.

Phosphoglycerylation Dataset	
No. of proteins	91
No. of lysine residues	3360
No. of phosphoglycerylated sites	111
No. of non-phosphoglycerylated sites	3249

B. Featurization of Protein Sequences

The dataset [22] included four protein sequence representations: physicochemical/biochemical properties, evolutionary information from multiple sequence alignment, and the representations from two different pre-trained transformer models. These features are referred to as Phy+Bio, BigramPGK, T5, and ESM-1b, respectively, in this work.

The breakdown of each type of protein sequence representations are as follows. The Phy+Bio feature is composed of 10 commonly used physicochemical/biochemical properties pertaining to each amino acid residues in the sequence. These were molecular weight, melting point, length of side chain, isoelectric point, free energy of solution in water, hydrostatic pressure asymmetry index, pK (-COOH), ionization equilibrium constant (pK-a), net charge, and hydrophobicity index. The BigramPGK feature is based on the position-specific

scoring matrices, which were obtained from the PSI-BLAST tool [47], and then the profile bigrams [48] were calculated to produce the final features. The T5 feature was extracted from the pre-trained ProfT5-XL-UniRef50 transformer, which is the best performing model in comparison to the other transformer models proposed in [49]. The ESM-1b feature was extracted from the ESM-1b pre-trained transformer model proposed in [50]. The ESM-1b transformer had protein sequence length limitation for the extraction of the representation, hence the longer sequences were split into multiple parts, taking the fixed lengths shifted to the right, one amino acid at a time, until the end of the sequence had been included in the final part. The representation for each amino acid residue in the protein sequence was then obtained by averaging the overlapping parts. The dimension of the protein sequence representation corresponding to Phy+Bio, BigramPGK, T5, and ESM-1b were $L \times 10$, $L \times 20$, $L \times 1024$, and $L \times 1280$, respectively, where L stands for the length of the protein sequence.

C. Sample Extraction

To extract the lysine residues from the protein sequences, a window of 15 upstream and 15 downstream around the site was employed, which is a commonly used approach for representing an amino acid sample [51, 52]. Each sample can therefore be denoted as:

$$S = \{A_{-15}, A_{-14}, \dots, A_{-1}, K, A_1, \dots, A_{14}, A_{15}\} \quad (1)$$

In (1), A_n represents the upstream amino acid residues, where $1 \leq n \leq 15$, A_n represents the downstream amino acid residues, where $1 \leq n \leq 15$, and K represents the lysine residue at the centre. The window therefore comprises a total of 31 amino acid residues. As commonly practiced, the lysine residues close to the start or finish of the protein sequence which did not have sufficient amino acids to make up the window size were excluded [53]. Moreover, a non-phosphoglycerylated lysine residue was only taken as a negative sample if its corresponding protein sequence contained two or more confirmed phosphoglycerylated sites [54]. As a result, 101 positive samples (label = 1) and 425 negative samples (label = 0) were obtained. The feature vector of each lysine residue corresponding to Phy+Bio, BigramPGK, T5, and ESM-1b representations was therefore 310-dimensional, 400-dimensional (after profile bigram calculation on the 31×20 feature matrix), 31744-dimensional and 39680-dimensional, respectively.

D. Data Balancing

The number of positive samples obtained after the sample extraction stage was less than the number of negative samples. This led to the class imbalance of 1:4.2 between the positive and negative samples which can bias the classification process in favour of the majority class. The random under-sampling technique was adopted where a subset of the negative samples were randomly selected to bring the imbalance ratio down to 1:1.5 [55]. The final number of positive and negative samples were 101 and 152, respectively, to train and test the classifier.

E. Evaluation Metric

In this work, we have used the AUC (Area Under the receiver operating characteristic (ROC) Curve) metric to evaluate the classifier. AUC measures a classifier's ability

to separate positive and negative samples and therefore is seen as a useful metric for evaluating the overall performance of a classifier. It takes on values from 0 to 1, where 0 indicates the worst separability and 1 indicates a perfect separability. AUC measure is prevalent in clinical and computational biology research due to the impediments of achieving high performance in both sensitivity and specificity and increase in one occurs at the cost of the other [56].

F. CNN Classifier

To build the predictor in this study, we employed a 1-dimensional CNN (1D CNN) network using the Tensorflow framework [57]. To efficiently train and test the network, the dataset was split into train (80%) and independent test (20%) sets. Furthermore, from the training set, 80% of the samples were used to train the network and the remaining 20% were used for validation. The network has 3 convolutional layers and 2 fully connected layers. The first convolutional layer has 128 filters of size 5, the second layer has 128 filters of size 3, and the third layer has 64 filters of size 3. The stride of 1 was used in the layers and padding was set to have the layer output size the same as the layer's input size. Moreover, dropouts were used after each of the layers to avoid overfitting. For the fully connected layers, the first layer has 128 neurons, and the second layer has 32 neurons. The network's output has a single neuron for predicting whether the site is phosphoglycerylated or not. The ReLU activation function is used in all the layers of the network, except for the output neuron, which has a sigmoid activation function. The learning rate of the network for the different features (i.e., Phy+Bio, BigramPGK, T5, and ESM-1b) was optimised using the Keras Tuner Bayesian Optimization algorithm. Finally, the network was trained using the Adam optimizer with binary cross-entropy loss and AUC metric and an early stopping with a patience of 5. The architecture of the 1D CNN used in this work is depicted in Fig. 1.

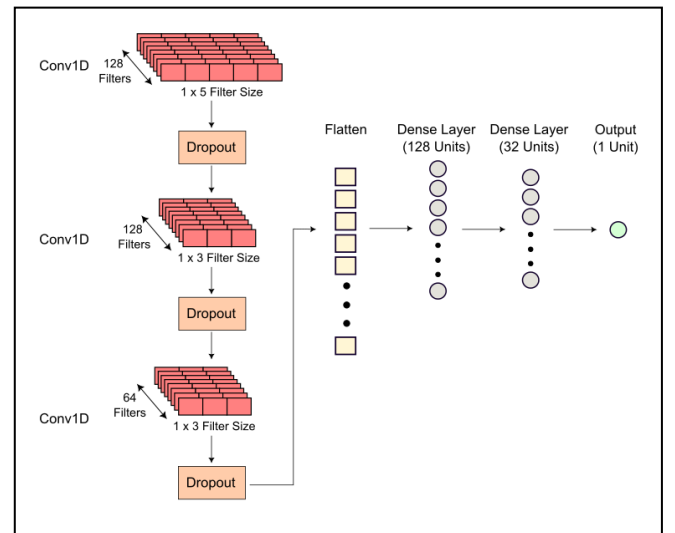


Fig. 1. The 1D CNN architecture employed in this work.

IV. RESULTS AND DISCUSSION

The performances obtained with the use of the DL method (1D CNN) were compared to the performances obtained previously [22] on the same set of features using the traditional machine learning algorithms. For fair comparisons, the same training and independent test sets

were used as the previous work. The AUC values are shown in Table II.

TABLE II. THE PERFORMANCE ON AUC METRIC FOR THE DIFFERENT FEATURES (PHY+BIO, BIGRAMPGK, T5, AND ESM-1B) USING THE TRADITIONAL ML CLASSIFIERS AND THE 1D CNN CLASSIFIER ON THE INDEPENDENT TEST SET. THE HIGHEST VALUES OF EACH ROW ARE HIGHLIGHTED IN BOLD.

Features	Classifiers					
	LR*	SVM (poly)*	SVM (RBF)*	RF*	LightGBM*	1DCNN
Phy+Bio	0.395	0.552	0.371	0.489	0.503	0.637
BigramPGK	0.686	0.666	0.676	0.742	0.742	0.747
T5	0.726	0.726	0.737	0.735	0.592	0.761
ESM-1b	0.803	0.813	0.811	0.719	0.748	0.839
Classifier Average	0.653	0.689	0.649	0.671	0.646	0.746

*Results obtained from previous work.

As seen in Table II, the performance of the 1D CNN network surpasses the performance of the traditional ML classifiers (LR, SVM (poly and RBF kernels), RF, and LightGBM) on all the feature types based on the AUC measure. With respect to the best performing traditional ML method, the AUCs are improved by 15.4% for the Phy+Bio feature over the SVM (poly) classifier, 0.7% for the BigramPGK feature over the RF classifier, 3.3% for the T5 feature over the SVM (RBF) classifier, and 3.2% for the ESM-1b feature over the SVM (poly) classifier. In terms of the average AUC obtained by each classifier on all the features, 1D CNN attained an average AUC of 0.746, which is an increase of 8.3% over the SVM (poly) classifier. Moreover, it can be observed from the measure that the features extracted from the transformer models (T5 and ESM-1b) performed the best in comparison to the evolutionary and physicochemical/biochemical features. This echoes the findings of the previous study [22] that the transformer based features are much more effective for distinguishing between the phosphoglycylated and non-phosphoglycylated lysine residues. Out of all the features with 1D CNN, the ESM-1b transformer-based feature performed the best with an AUC value of 0.839, while the Phy+Bio feature performed the worst with an AUC value of 0.637.

Additionally, we investigated the distribution of the phosphoglycylated and non-phosphoglycylated lysine residues based on their input features and the representation learned through their respective 1D CNNs. Fig. 2 shows the t-SNE visualizations [58] of the training samples of the different features into a two-dimensional space. The sample distributions on the left plots represents the different input features and the right plots shows the representation of the residues from the second-last fully connected layer (32-dimensional) of the trained 1D CNN. It can be seen that the distribution of the phosphoglycylated (in green) and non-phosphoglycylated (in red) lysine residues in the left plots pertaining to the raw input features are relatively clustered together. However, the distribution of the phosphoglycylated and non-phosphoglycylated lysine residues in the right plots, which are the representations from the 1D CNN, are in more distinguishable clusters. This shows that the extraction of features from the input features by the CNN learns patterns relating to the phosphoglycylated and non-phosphoglycylated

characteristics of the residues which demonstrates the significance of the DL network to the prediction task.

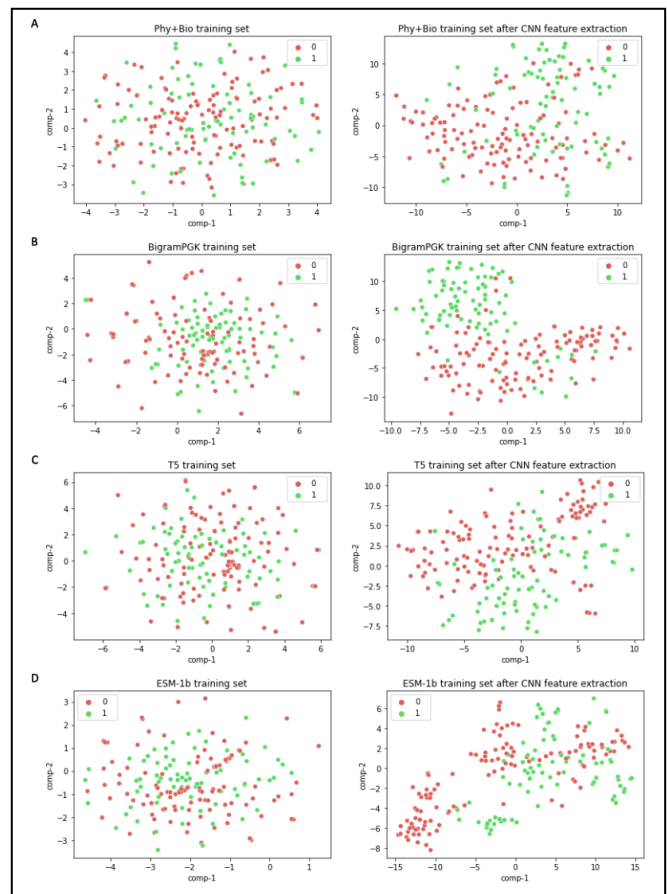


Fig. 2. The t-SNE distribution of the phosphoglycylated and non-phosphoglycylated lysine residues of the training set. Colour green indicates the phosphoglycylated residues and the colour red indicates the non-phosphoglycylated residues. The plot of the raw input features Phy+Bio (A), BigramPGK (B), T5 (C), and ESM-1b (D) are located on the left of the figure and the representation learned through their respective 1D CNNs are located on the right of the figure.

V. CONCLUSION

In this study, we have extended our previous work by using a DL network (CNN) for comparing the effectiveness of the pre-trained transformer-based features against the evolutionary and physicochemical/biochemical based features. As a result, it was seen that the CNN network improved the performances of all the features and demonstrated that the transformer-based features are much more effective for detecting the phosphoglycylated and non-phosphoglycylated sites in the protein sequences. The DL classifier together with the DL features from the pre-trained transformer models holds huge potential for improving the performance of various protein-related prediction tasks. In future, we will investigate the further enhancing of the phosphoglycylation prediction performance by employing 2D CNN architectures after extracting image representations from feature vectors using technologies such as DeepInsight [14].

ACKNOWLEDGMENT

This research is partially supported by Australian Research Council Grant DP180102727. We are grateful to the Griffith University eResearch Service & Specialised

Platforms team for their High Performance Computing Cluster to complete this research.

REFERENCES

- [1] Wen, B., Zeng, W.F., Liao, Y., Shi, Z., Savage, S.R., Jiang, W., and Zhang, B.: 'Deep learning in proteomics', *Proteomics*, 2020, 20, (21-22), pp. 1900335
- [2] Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., López García, A., Heredia, I., Malik, P., and Hluchý, L.: 'Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey', *Artificial Intelligence Review*, 2019, 52, pp. 77-124
- [3] Mann, M., Kumar, C., Zeng, W.-F., and Strauss, M.T.: 'Artificial intelligence for proteomics and biomarker discovery', *Cell systems*, 2021, 12, (8), pp. 759-770
- [4] Chandra, A.A., Sharma, A., Dehzangi, A., and Tsunoda, T.: 'Ram-PGK: prediction of lysine phosphoglycerylation based on residue adjacency matrix', *Genes*, 2020, 11, (12), pp. 1524
- [5] Del Conte, A., Bouhraoua, A., Mehdiabadi, M., Clementel, D., Monzon, A.M., Tosatto, S.C., and Piovesan, D.: 'CAID prediction portal: a comprehensive service for predicting intrinsic disorder and binding regions in proteins', *Nucleic Acids Research*, 2023, pp. gkad430
- [6] Kandathil, S.M., Greener, J.G., and Jones, D.T.: 'Recent developments in deep learning applied to protein structure prediction', *Proteins: Structure, Function, and Bioinformatics*, 2019, 87, (12), pp. 1179-1189
- [7] Meyer, J.G.: 'Deep learning neural network tools for proteomics', *Cell Reports Methods*, 2021, 1, (2)
- [8] Neely, B.A., Dorfer, V., Martens, L., Bludau, I., Bouwmeester, R., Degroove, S., Deutsch, E.W., Gessulat, S., Käll, L., and Palczynski, P.: 'Toward an integrated machine learning model of a proteomics experiment', *Journal of proteome research*, 2023, 22, (3), pp. 681-696
- [9] Zeng, W.-F., Zhou, X.-X., Willems, S., Ammar, C., Wahle, M., Bludau, I., Voytik, E., Strauss, M.T., and Mann, M.: 'AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics', *Nature Communications*, 2022, 13, (1), pp. 7238
- [10] Min, S., Lee, B., and Yoon, S.: 'Deep learning in bioinformatics', *Briefings in bioinformatics*, 2017, 18, (5), pp. 851-869
- [11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I.: 'Attention is all you need', *Advances in neural information processing systems*, 2017, 30
- [12] Fukushima, K.: 'Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position', *Biological cybernetics*, 1980, 36, (4), pp. 193-202
- [13] Gokhale, M., Mohanty, S.K., and Ojha, A.: 'GeneViT: Gene vision transformer with improved DeepInsight for cancer classification', *Computers in Biology and Medicine*, 2023, 155, pp. 106643
- [14] Sharma, A., Vans, E., Shigemizu, D., Boroevich, K.A., and Tsunoda, T.: 'DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture', *Scientific reports*, 2019, 9, (1), pp. 11399
- [15] Sharma, A., Lysenko, A., Boroevich, K.A., Vans, E., and Tsunoda, T.: 'DeepFeature: feature selection in nonimage data using convolutional neural network', *Briefings in bioinformatics*, 2021, 22, (6), pp. bbab297
- [16] Sharma, A., Lysenko, A., Boroevich, K.A., and Tsunoda, T.: 'DeepInsight-3D architecture for anti-cancer drug response prediction with deep-learning on multi-omics', *Scientific Reports*, 2023, 13, (1), pp. 2483
- [17] Nambiar, A., Heflin, M., Liu, S., Maslov, S., Hopkins, M., and Ritz, A.: 'Transforming the language of life: transformer neural networks for protein prediction tasks', in Editor (Ed.) (Eds.): 'Book Transforming the language of life: transformer neural networks for protein prediction tasks' (2020, edn.), pp. 1-8
- [18] Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., and Rives, A.: 'Transformer protein language models are unsupervised structure learners', *Biorxiv*, 2020, pp. 2020.2012.2015.422761
- [19] Wang, W., Peng, Z., and Yang, J.: 'Single-sequence protein structure prediction using supervised transformer protein language models', *Nature Computational Science*, 2022, 2, (12), pp. 804-814
- [20] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805*, 2018
- [21] Qian, Y., Wu, J., and Zhang, Q.: 'CAT-CPI: Combining CNN and transformer to learn compound image features for predicting compound-protein interactions', *Frontiers in Molecular Biosciences*, 2022, 9, pp. 963912
- [22] Chandra, A., Tünnermann, L., Löfstedt, T., and Gratz, R.: 'Transformer-based deep learning for predicting protein properties in the life sciences', *Elife*, 2023, 12, pp. e82819
- [23] Xu, Y., Ding, Y.-X., Ding, J., Wu, L.-Y., and Deng, N.-Y.: 'Phogly-PseAAC: prediction of lysine phosphoglycerylation in proteins incorporating with position-specific propensity', *Journal of Theoretical Biology*, 2015, 379, pp. 10-15
- [24] Ju, Z., Cao, J.-Z., and Gu, H.: 'Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou's general PseAAC', *Journal of Theoretical Biology*, 2016, 397, pp. 145-150
- [25] Chen, Q.-Y., Tang, J., and Du, P.-F.: 'Predicting protein lysine phosphoglycerylation sites by hybridizing many sequence based features', *Molecular BioSystems*, 2017, 13, (5), pp. 874-882
- [26] Liu, L.-M., Xu, Y., and Chou, K.-C.: 'iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC', *Medicinal Chemistry*, 2017, 13, (6), pp. 552-559
- [27] Chandra, A., Sharma, A., Dehzangi, A., Ranganathan, S., Jokhan, A., Chou, K.-C., and Tsunoda, T.: 'PhoglyStruct: Prediction of phosphoglycerylated lysine residues using structural properties of amino acids', *Scientific reports*, 2018, 8, (1), pp. 17923
- [28] Chandra, A.A., Sharma, A., Dehzangi, A., and Tsunoda, T.: 'EvolStruct-Phogly: incorporating structural properties and evolutionary information from profile bigrams for the phosphoglycerylation prediction', *BMC genomics*, 2019, 19, pp. 1-9
- [29] Chandra, A., Sharma, A., Dehzangi, A., Shigemizu, D., and Tsunoda, T.: 'Bigram-PGK: phosphoglycerylation prediction using the technique of bigram probabilities of position specific scoring matrix', *BMC molecular and cell biology*, 2019, 20, pp. 1-9
- [30] Huang, K.-Y., Hung, F.-Y., Kao, H.-J., Lau, H.-H., and Weng, S.-L.: 'iDPGK: characterization and identification of lysine phosphoglycerylation sites based on sequence-based features', *BMC bioinformatics*, 2020, 21, (1), pp. 1-16
- [31] Ahmed, S., Rahman, A., Hasan, M.A.M., Islam, M.K.B., Rahman, J., and Ahmad, S.: 'predPhogly-Site: Predicting phosphoglycerylation sites by incorporating probabilistic sequence-coupling information into PseAAC and addressing data imbalance', *Plos one*, 2021, 16, (4), pp. e0249396
- [32] Sohrawordi, M., Hossain, M.A., and Hasan, M.A.M.: 'PLP_FS: prediction of lysine phosphoglycerylation sites in protein using support vector machine and fusion of multiple F_Score feature selection', *Briefings in Bioinformatics*, 2022, 23, (5), pp. bbac306
- [33] Lai, S., Cao, Y., Wang, P., Ye, L., and Liu, Z.: 'BERT_PLPS: A BERT-based Model for Predicting Lysine Phosphoglycerylation Sites', 2023
- [34] Lan, F., and Shi, Y.: 'Epigenetic regulation: methylation of histone and non-histone proteins', *Science in China Series C: Life Sciences*, 2009, 52, pp. 311-322
- [35] Reddy, H.M., Sharma, A., Dehzangi, A., Shigemizu, D., Chandra, A.A., and Tsunoda, T.: 'GlyStruct: glycation prediction using structural properties of amino acid residues', *BMC bioinformatics*, 2019, 19, (13), pp. 55-64
- [36] Choudhary, C., Kumar, C., Gnäd, F., Nielsen, M.L., Rehman, M., Walther, T.C., Olsen, J.V., and Mann, M.: 'Lysine acetylation targets protein complexes and co-regulates major cellular functions', *Science*, 2009, 325, (5942), pp. 834-840
- [37] Tan, M., Luo, H., Lee, S., Jin, F., Yang, J.S., Montellier, E., Buchou, T., Cheng, Z., Rousseaux, S., and Rajagopal, N.: 'Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification', *Cell*, 2011, 146, (6), pp. 1016-1028
- [38] Cohen, P.: 'The origins of protein phosphorylation', *Nature cell biology*, 2002, 4, (5), pp. E127-E130
- [39] López, Y., Sharma, A., Dehzangi, A., Lal, S.P., Taherzadeh, G., Sattar, A., and Tsunoda, T.: 'Success: evolutionary and structural

- properties of amino acids prove effective for succinylation site prediction', *BMC genomics*, 2018, 19, pp. 105-114
- [40] Lopez, Y., Dehzangi, A., Reddy, H.M., and Sharma, A.: 'C-iSUMO: a sumoylation site predictor that incorporates intrinsic characteristics of amino acid sequences', *Computational Biology and Chemistry*, 2020, 87, pp. 107235
- [41] Bulcun, E., Ekici, M., and Ekici, A.: 'Disorders of glucose metabolism and insulin resistance in patients with obstructive sleep apnoea syndrome', *International journal of clinical practice*, 2012, 66, (1), pp. 91-97
- [42] Yang, Z.R.: 'Predicting sulfotyrosine sites using the random forest algorithm with significantly improved prediction accuracy', *BMC bioinformatics*, 2009, 10, pp. 1-10
- [43] Chen, X., Niroomand, F., Liu, Z., Zankl, A., Katus, H., Jahn, L., and Tiefenbacher, C.: 'Expression of nitric oxide related enzymes in coronary heart disease', *Basic research in cardiology*, 2006, 101, pp. 346-353
- [44] Suzuki, A., Yamada, R., and Yamamoto, K.: 'Citrullination by peptidylarginine deiminase in rheumatoid arthritis', *Annals of the New York Academy of Sciences*, 2007, 1108, (1), pp. 323-339
- [45] Mastronardi, F.G., Wood, D.D., Mei, J., Rajmakers, R., Tseveleki, V., Dosch, H.-M., Probert, L., Casaccia-Bonnel, P., and Moscarello, M.A.: 'Increased citrullination of histone H3 in multiple sclerosis brain and animal models of demyelination: a role for tumor necrosis factor-induced peptidylarginine deiminase 4 translocation', *Journal of Neuroscience*, 2006, 26, (44), pp. 11387-11396
- [46] Li, W., and Godzik, A.: 'Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences', *Bioinformatics*, 2006, 22, (13), pp. 1658-1659
- [47] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J.: 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic acids research*, 1997, 25, (17), pp. 3389-3402
- [48] Sharma, A., Lyons, J., Dehzangi, A., and Paliwal, K.K.: 'A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition', *Journal of theoretical biology*, 2013, 320, pp. 41-46
- [49] Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., and Steinegger, M.: 'ProtTrans: Toward understanding the language of life through self-supervised learning', *IEEE transactions on pattern analysis and machine intelligence*, 2021, 44, (10), pp. 7112-7127
- [50] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., and Ma, J.: 'Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences', *Proceedings of the National Academy of Sciences*, 2021, 118, (15), pp. e2016239118
- [51] Xu, Y., Song, J., Wilson, C., and Whisstock, J.C.: 'PhosContext2vec: a distributed representation of residue-level sequence contexts and its application to general and kinase-specific phosphorylation site prediction', *Scientific reports*, 2018, 8, (1), pp. 8240
- [52] Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.-C.: 'iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset', *Analytical biochemistry*, 2016, 497, pp. 48-56
- [53] Wang, X., Xu, M.L., Li, B.Q., Zhai, H.L., Liu, J.J., and Li, S.Y.: 'Prediction of phosphorylation sites based on Krawtchouk image moments', *Proteins: Structure, Function, and Bioinformatics*, 2017, 85, (12), pp. 2231-2238
- [54] Khalili, E., Ramazi, S., Ghanati, F., and Kouchaki, S.: 'Predicting protein phosphorylation sites in soybean using interpretable deep tabular learning network', *Briefings in bioinformatics*, 2022, 23, (2), pp. bbac015
- [55] Ramazi, S., and Zahiri, J.: 'Post-translational modifications in proteins: resources, tools and prediction methods', *Database*, 2021, 2021, pp. baab012
- [56] Nahm, F.S.: 'Receiver operating characteristic curve: overview and practical use for clinicians', *Korean journal of anesthesiology*, 2022, 75, (1), pp. 25-36
- [57] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., and Devin, M.: 'Tensorflow: Large-scale machine learning on heterogeneous distributed systems', *arXiv preprint arXiv:1603.04467*, 2016
- [58] Van der Maaten, L., and Hinton, G.: 'Visualizing data using t-SNE', *Journal of machine learning research*, 2008, 9, (11)