frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 17 October 2022
DOI 10.3389/fgene.2022.1036862

OPEN ACCESS

EDITED BY
Jia Meng,
Xi'an Jiaotong-Liverpool University, China

REVIEWED BY
Wei Chen,
Chengdu University of Traditional Chinese Medicine, China
Yuan Zhou,
Peking University, China

*CORRESPONDENCE
Leina Ma,
leinama@gmail.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 06 September 2022
ACCEPTED 04 October 2022
PUBLISHED 17 October 2022

CITATION
Zhao J, Jiang H, Zou G, Lin Q, Wang Q, Liu J and Ma L (2022), CNNArginineMe: A CNN structure for training models for predicting arginine methylation sites based on the One-Hot encoding of peptide sequence.
Front. Genet. 13:1036862.
doi: 10.3389/fgene.2022.1036862

CNNArginineMe: A CNN structure for training models for predicting arginine methylation sites based on the One-Hot encoding of peptide sequence

Jiaojiao Zhao[1,2†], Haoqiang Jiang[2†], Guoyang Zou[2], Qian Lin[1], Qiang Wang[3], Jia Liu[4] and Leina Ma[1*]

[1]Cancer Institute of the Affiliated Hospital of Qingdao University and Qingdao Cancer Institute, Qingdao University, Qingdao, China, [2]School of Basic Medicine, Qingdao University, Qingdao, China, [3]Oncology Department, Shandong Second Provincial General Hospital, Jinan, China, [4]Department of Pharmacology, School of Pharmacy, Qingdao University, Qingdao, China

Protein arginine methylation (PRme), as one post-translational modification, plays a critical role in numerous cellular processes and regulates critical cellular functions. Though several *in silico* models for predicting PRme sites have been reported, new models may be required to develop due to the significant increase of identified PRme sites. In this study, we constructed multiple machine-learning and deep-learning models. The deep-learning model CNN combined with the One-Hot coding showed the best performance, dubbed CNNArginineMe. CNNArginineMe performed best in AUC scoring metrics in comparisons with several reported predictors. Additionally, we employed CNNArginineMe to predict arginine methylation proteome and performed functional analysis. The arginine methylated proteome is significantly enriched in the amyotrophic lateral sclerosis (ALS) pathway. CNNArginineMe is freely available at https://github.com/guoyangzou/CNNArginineMe.

KEYWORDS

arginine methylation, deep learning model, amyotrophic lateral sclerosis (ALS) pathway, CNNArginineMe, machine learning

1 Introduction

Protein arginine methylation (PRme) is a common post-translational modification (PTM), which plays a crucial role in pre-mRNA splicing, DNA damage, signaling, mRNA translation, cell signaling, and cell fate decision (Blanc and Richard, 2017; Kumar et al., 2017; Wang S. M. et al., 2019; Abe and Tanaka, 2020; Parbin et al., 2021; Scopino et al., 2021). Arginine contains five potential hydrogen bond donors favourable for interactions with biological hydrogen bond acceptors (Yang and Bedford, 2013). Types of arginine methylation include ω-N^G-monomethyl arginine (MMA), ω-N^G, N^G-asymmetric dimethylarginine (ADMA) and ω-N^G, N^G-symmetric dimethylarginine (SDMA). A family of nine protein arginine methyltransferases (PRMTs) catalyzes the formation

of MMA, ADMA, and SDMA in mammalian cells (Bedford and Clarke, 2009; Yang and Bedford, 2013; Poulard et al., 2016). PRMTs are classified into three groups of enzymes (types I, II, and III) according to their catalyzed types of methylations. All of them produce MMA, and type I PRMTs (PRMT1, PRMT2, PRMT3, Carm1/PRMT4, PRMT6, and PRMT8) form ADMA, while Type II PRMTs (PRMT5 and PRMT9) form SDAM, whereas PRMT7 is the only Type III enzyme, exclusively catalyzing the formation of MMA (Poulard et al., 2016). Arginine methylation has regulatory effects on various physiological processes and pathological conditions; dysregulation of the enzymes is associated with several diseases, such as cancer (Boulanger et al., 2005; Covic et al., 2005; Ratovitski et al., 2015; Fedoriw et al., 2019; Guccione and Richard, 2019; Szewczyk et al., 2020). Therefore, it is essential to accurately predict methylation sites to understand PRme molecular mechanisms.

Traditional experiments used to identify methylation sites—such as mass-spectrometry, methylation-specific antibodies, and ChIP-Chip, are labour-intensive, expensive, time-consuming, and require a high level of technical expertise (Wilkins et al., 1999). With the increase of the identified PRme sites, computational methods have emerged as an efficient strategy to complement and extend traditional experimental methods for PRme site identification.

Eleven computational predictors have been built to predict arginine methylation, including nine machine-learning models and two deep-learning models. In the machine-learning models, MeMo was constructed using sequential features (Chen et al., 2006). Shao *et al.* incorporated a support vector machine (SVM) algorithm with a Bi-profile Bayes feature extraction method (Shao et al., 2009). The model MASA combined the SVM algorithm with protein sequences and structural characteristics (Shien et al., 2009). The model PMeS was based on an enhanced feature encoding scheme (Shi et al., 2012). The predictor iMethyl-PseAAC was formed by incorporating the physicochemical features, sequence evolution, biochemical, and structural disorder information into the general form of pseudo amino acid composition (Qiu et al., 2014). The model PSSMe was based on the information gain optimization method for species-specific methylation site prediction (Wen et al., 2016). The predictor GPS-MSP was developed to predict different PRme types, the first model for predicting each PRme type (Deng et al., 2017). The model MePred-RF integrated the random-forest algorithm with a sequence-based feature selection technique (Wei et al., 2019). Hou and coworkers built a model to predict PRme sites based on composition-transition-distribution features (Hou et al., 2020). In the deep-learning-based models, CapsNet contained a multi-layer CNN for predicting PRme sites, which outperformed other well-known tools in most cases (Wang D. et al., 2019). The deep-learning model DeepRMethylSite was constructed with the integration of One-Hot and embedding integer encodings (Chaudhari et al., 2020). The development of these models has contributed significantly to the discovery of PRme sites.

The limitation of experimentally verified PTM data is often the main reason for inaccurate prediction. With the increase of PRme sites, it is necessary to re-investigate the predictors for PRme sites. We developed and compared several prediction models with the reported predictors in this study. We found that our deep-learning model CNNArginineMe had the best performance. Moreover, we used CNNArginneMe to predict human proteins that contained PRme sites and performed biological function enrichment analysis for these proteins using Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG).

2 Material and methods

2.1 Dataset preparation

Figure 1A shows the construction procedure of the dataset. Specifically, we extracted the human PRme-containing proteins from phosphositePlus v6.5.9.3 (Hornbeck et al., 2015) and UniProt (Consortium, 2017). For each arginine of these proteins, we generated the 51-aa long sequence fragment with the central arginine. It is worth noting that if the central arginine is located at the N-terminus or C-terminus of the protein, the truncated sequence fragment will be padded with "_" to a length of 51 amino acid residues. The related sequence is defined as a positive sample if the central arginine is annotated as methylation. Otherwise, it is defined as a negative sample. We deduplicated the collected fragments. Accordingly, we collected 188,930 Arginine sites, including 9138 PRme sites and 179,792 non-PRme sites (Figure 1A). Because the number of PRme sites is only 5% of non-PRme sites, we randomly extracted 40,000 non-PRme sites as negative samples and considered the 9,138 PRme sites as positive and. We separated the dataset into a ten-fold cross-validation dataset (~90%) and an independent test dataset (~10%). The cross-validation dataset consisted of 8245 positive samples and 35,972 negative samples, and the independent test dataset included 893 positive samples and 4028 negative samples.

2.2 Feature encoding schemes

To create a methylated arginine predictor with high performance, we employ 19 feature encoding schemes, introduced below.

In the one-Hot encoding scheme (Wang et al., 2017), each amino acid is defined as a 20n length vector. Since only one of the 20 bits is 1, it uniquely represents the twenty amino acids. The rest feature encoding approaches (Chen et al., 2018b) include Dipeptide deviation from the expected mean (DDE), dipeptide

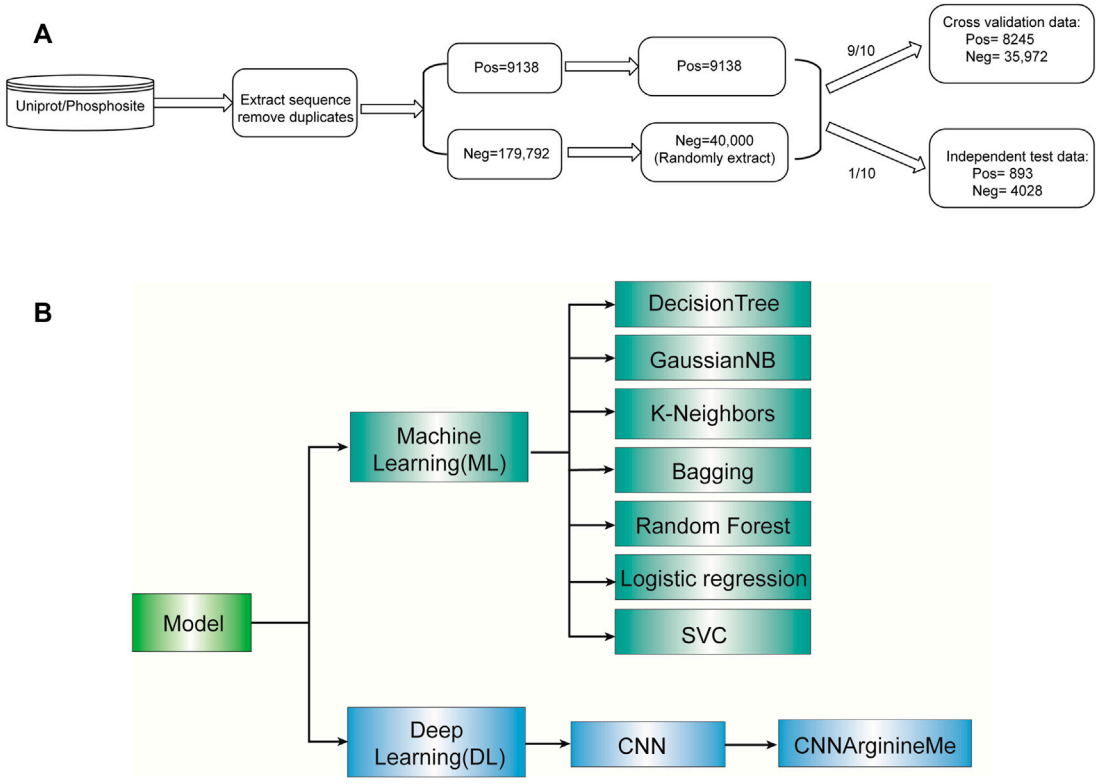

FIGURE 1
Roadmap of this study. **(A)** Flow diagram depicting data pre-processing. **(B)** The algorithms selected for model construction.

composition (DPC), Enhanced Amino Acid Composition (EAAC), Composition of k-spaced Amino Acid Pairs (CKSAAP), Distribution (CTDD); Enhanced GAAC (EGAAC), Transition (CTDT), Composition of k-Spaced Amino Acid Group Pairs (CKSAAGP), Conjoint Triad (CTriad), k-Spaced Conjoint Triad (KSCTriad), binary encoding (BINA), grouped tripeptide composition (GTPC), BLOSUM62, Composition (CTDC), grouped dipeptide composition (GDPC), Z-Scale (ZSCALE), amino acid composition (AAC) and Grouped Amino Acid Composition (GAAC).

2.3 Model construction

We constructed machine-learning models using seven algorithms such as Decision Tree Classifier (Strobl et al., 2009), Gaussian NB (Huang and Hsu, 2002), k-nearest neighbours (Gil-Pita and Yao, 2008), Bagging Classifier (Dong et al., 2006), Random Forest (Pang et al., 2006) (Pang et al., 2006), Logistic Regression (Sperandei, 2014), and SVC (Cai et al., 2003). Their default parameters were used for development, using the corresponding packages in the sklearn of python3 (Supplementary Table S1). We also used Convolutional Neural Network (CNN) algorithm to build deep-learning models (parameters listed in Supplementary Table S2). Each algorithm is described briefly below.

2.3.1 Random forest

The Random Forest classifier is an ensemble of multiple decision tree classifiers, each of which is trained from a different training set and features (Pang et al., 2006).

2.3.2 Support vector classifier (SVC)

SVM is one of the most robust prediction methods based on statistical learning frameworks or VC theory (Cai et al., 2003). Given a set of training examples, each marked as belonging to one of two categories, and an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM maps training examples to points in space to maximize the width of the gap between the two categories. New examples are then mapped into the same space and predicted to belong to a category based on which side of the gap they fall.

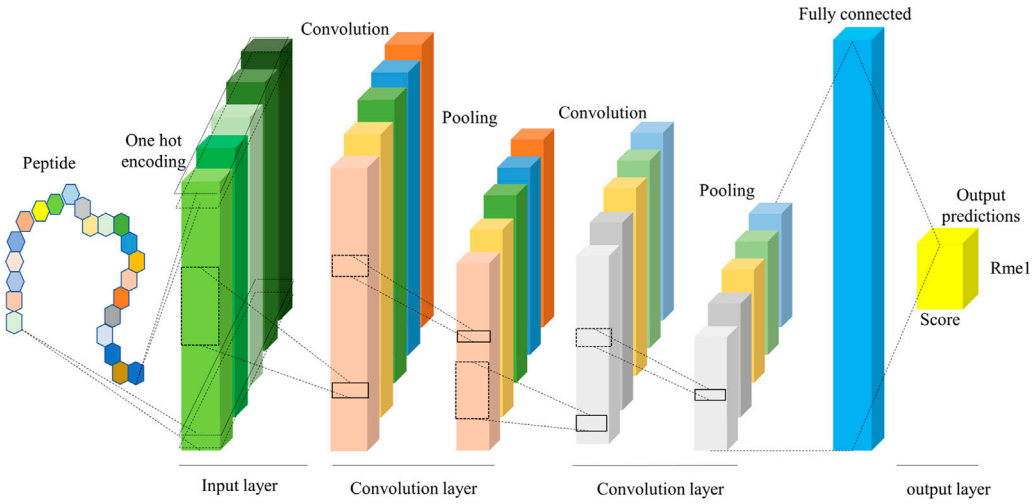

FIGURE 2
The framework of the CNN model. It includes input, convolution, and output layers, where the convolution layer extracts the features from the peptide sequence.

TABLE 1 Prediction Performances of different models integrating the One-Hot encoding approach.

Performances in ten-fold cross-validation

	AUC	Sn (Sp = 0.9)	Sn (Sp = 0.95)	Sn (Sp = 0.99)
Gaussian NB	0.6884	0.4579	0	0
Decision Tree Classifier	0.6383	0.4143	0	0
K-nearest Neighbors	0.7047	0.3149	0.3149	0.1061
Bagging Classifier	0.7678	0.5049	0.3845	0.1554
Random Forest Classifier	0.8167	0.5538	0.4234	0.2144
Logistic Regression	0.8189	0.5332	0.3925	0.1706
SVC	0.8367	0.5812	0.4355	0.2085
CNN	0.8708	0.6642	0.5174	0.2231

Performances in the independent test

	AUC	Sn (Sp = 0.9)	Sn (Sp = 0.95)	Sn (Sp = 0.99)
Gaussian NB	0.6891	0.4528	0	0
Decision Tree Classifier	0.6347	0.409	0	0
K-nearest Neighbors	0.7029	0.314	0.314	0.1136
Bagging Classifier	0.7583	0.4943	0.3567	0.1477
Random Forest Classifier	0.8121	0.5345	0.4131	0.2165
Logistic Regression	0.8199	0.5255	0.3757	0.1409
SVC	0.8365	0.5756	0.4206	0.1929
CNN	0.8671	0.6538	0.5025	0.1902

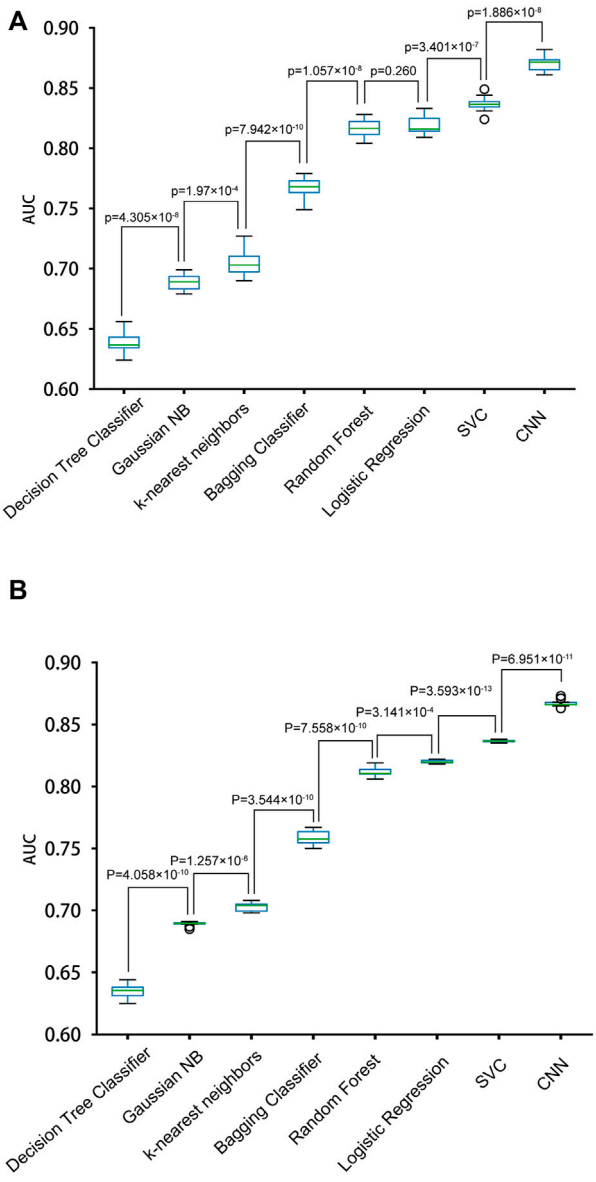

FIGURE 3
Performances of different models with the One-Hot feature for predicting PRme sites. **(A)** The AUC values of different models in ten-fold cross-validation. **(B)** The AUC values of different models in the independent test.

2.3.3 K-Nearest Neighbours algorithm

K-Nearest Neighbours algorithm is a statistical classifier that calculates the distance between the data features to be classified and the training data features and sorts them, takes out the K training data features with the closest distance; then determines the new sample category according to the category of the K similar training data features: if they all belong to the same category, then the new sample also belongs to this category; otherwise, each candidate category is scored, and the category of the new sample is determined according to a specific rule (Gil-Pita and Yao, 2008).

2.3.4 Gaussian NB

Bayes Theorem describes the probability of an event based on prior knowledge of conditions related to the event. Gaussian Naive Bayes is one classifier model that assumes that the prior probability of a feature is usually distributed (Huang and Hsu, 2002).

2.3.5 Decision tree classifier

The decision tree model is a tree structure; each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a category. When running, using training data to establish a decision tree model based on the principle of minimizing the loss function; and when predicting, using the decision tree model to classify new data. It includes three steps: feature selection, decision tree generation, and decision tree pruning (Strobl et al., 2009).

2.3.6 Bagging classifier

The bagging algorithm is representative of parallel integrated learning, which is mainly divided into four steps 1) cleaning the data according to the actual situation; 2) random sampling: repeat T times and randomly select T sub-samples from the sample each time; 3) individual training: Put each sub-sample into individual learner training; 4) classification decision: Use voting method integration to make the classification decision (Dong et al., 2006).

2.3.7 Logistic regression

It is the preferred method for binary classification tasks (Sperandei, 2014). It outputs a discrete binary result between 0 and 1. Moreover, logistic regression measures the relationship between the dependent variable (the label we want to predict) and one or more independent variables (features) by using its inherent logistic function to estimate probability. These probabilities need to be binarized. The task of the logistic function is also known as the sigmoid function, which is an S-shaped curve. It can map any real value to a value between 0 and 1, but it cannot be 0 or 1. Then use a threshold classifier to convert values between 0 and 1 to 0 or 1. Maximum likelihood estimation is a general method for estimating parameters in statistical models.

2.3.8 The deep-learning CNN algorithm

Deep learning is a sub-discipline of machine learning. Deep learning is based on artificial neural networks with representation learning that aim to mimic the human brain. The key difference between deep learning and traditional machine learning algorithms such as support vector machine (SVM) and random forests (RF) is that deep learning can automatically learn features and patterns from data without handcrafted feature engineering (Wen et al., 2020). We took the 1D-CNN Model with One-Hot encoding (CNN_{OH}) as an example to

TABLE 2 Prediction performances of the models integrating different algorithms and various feature encoding approaches in ten-fold cross-validation.

	Random forest	SVC	Logistic regression	CNN
GAAC	0.606	0.609	0.566	
GDPC	0.71	0.658	0.635	
GTPC	0.736	0.667	0.693	
CTDD	0.718	0.693	0.692	
CKSAAGP	0.737	0.676	0.699	
CTDT	0.734	0.709	0.696	
KSCTriad	0.74	0.699	0.726	
CTriad	0.745	0.699	0.726	
EGAAC	0.736	0.713	0.731	
CTDC	0.756	0.72	0.704	
AAC	0.775	0.725	0.701	
ZSCALE	0.805	0.738	0.772	
DPC	0.804	0.759	0.798	
DDE	0.801	0.766	0.798	
CKSAAP	0.8	0.78	0.801	
BLOSUM62	0.807	0.821	0.834	0.848
One-Hot	0.813	0.839	0.822	0.871
EAAC	0.82	0.819	0.841	0.859

illustrate the deep-learning network framework. This model contains four layers, listed below (Figure 2).

1. Input layer. The One-Hot encoding encodes each input sequence of 51 amino acids to a 51 × 21 binary matrix.
2. Convolution layer. It consisted of two convolution sublayers, each followed by a max-pooling sublayer. The first convolution sublayer includes 256 different convolution kernels with a size of 9 × 21. Each kernel is applied to the 51 × 21 matrix and results in a feature vector with the size of 43 (= 51–9+1). Thus, the 256 kernels output a 43 × 256 matrix. Next, a pooling kernel with the size of 2 is applied to the feature matrix and produces a 21 × 256 matrix. In the second convolution sublayer, 32 different convolution kernels with the size 7 × 256 are applied to generate a 15 × 32 matrix, followed by a pooling kernel with size two that produces a 7 × 32 data matrix.
3. Fully connected layer. The 7 × 32 data matrix generated from the convolution layer is nonlinearly transformed to 128 representative features.
4. Output layer. The modification score is calculated based on the 128 features using the "Sigmoid" function.

2.4 Model training

To avoid overfitting, we use early stopping, a widely used method to screen the better models, and use the cross-validation method to get the best prediction model by integrating all the better models.

2.5 Performance evaluation

To evaluate the performance of models, we used Sensitivity (Sn), Specificity (Sp), and the area under the Receiver Operating Characteristic (AUC) as the performance metrics. Sn defines the model's ability to identify positive residues from actual positive residues; the Sp measures the model's ability to identify the negative samples from the actual negative samples; AUC measures the comprehensive performance of the model.

2.6 Statistical methods

The paired student's t-test was used to test the significant difference between the mean values of the two paired populations. The threshold is set to 0.05.

2.7 GO and KEGG analysis

Gene Ontology (GO) analysis for enriched "biological process" terms and enriched genes in KEGG pathways were performed using R (v4.0.4), including clusterProfiler, topGO,

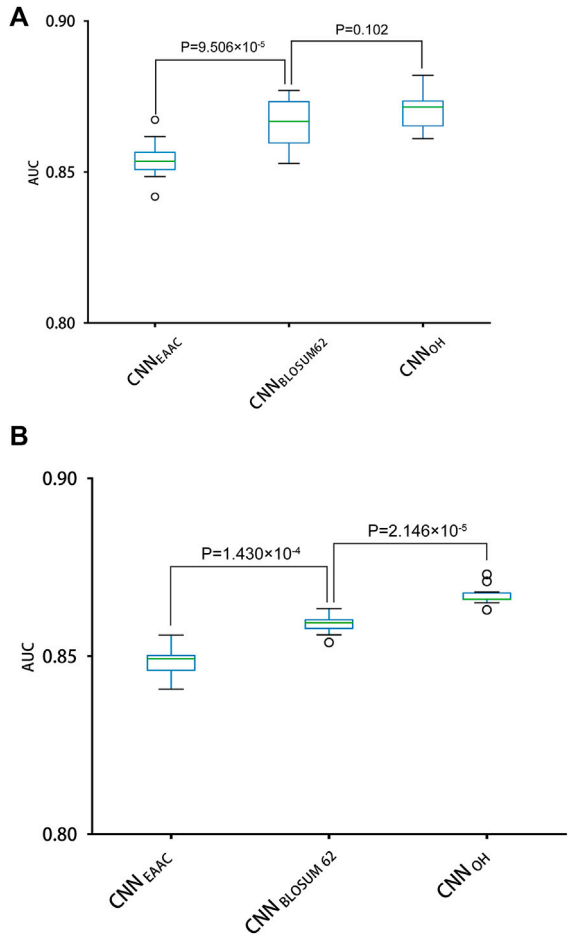

FIGURE 4
Performances of the CNN models with different features for predicting PRme sites. **(A)** The AUC values of different models using ten-fold cross-validation. **(B)** The AUC values of different models in the independent test.

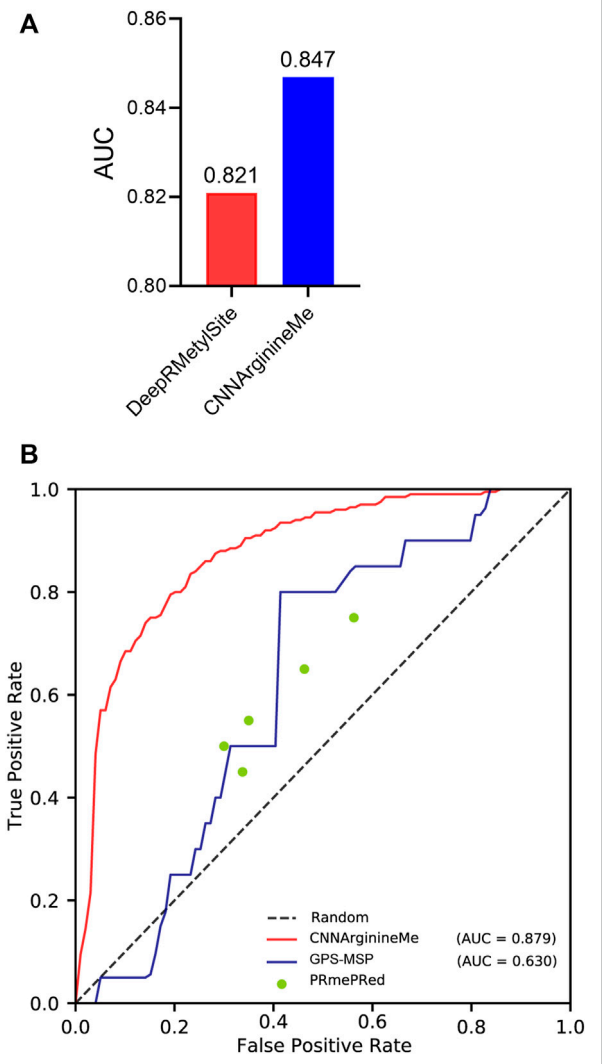

FIGURE 5
The comparison of CNNArginineMe and reported classifiers. **(A)** AUC values of reconstructed CNNArginineMe and DeepRMethylSite. **(B)** The comparison between CNNArginineMe and two reported online prediction models, i.e., GPS-MSP and PRmePRed.

org. Hs.eg.db, AnnotationDbi, stats4, BiocGenerics, Iranges, and enrichplot packages.

3 Results

3.1 The CNN-based model performed better than traditional machine-learning-based models

We constructed eight prediction models by integrating seven machine-learning algorithms and the CNN algorithm with the simple One-Hot encoding approach and compared their performances. The seven machine-learning algorithms included Random Forest, SVC, K-nearest Neighbors Classifier, Gaussian NB, Decision Tree Classifier, Bagging Classifier, and Logistic Regression. The result metrics (AUC, Sn (Sp = 0.9), Sn (Sp = 0.95), Sn (Sp = 0.99)) were used for evaluation in the ten-fold cross-validation and the independent test (Table 1 and Figure 3). The average AUC value and the Sn values of CNN_{OH} model were the largest among the eight models. Therefore, CNN_{OH} is the best model and has excellent predictive ability. Additionally, among the machine-learning models, the average AUC values of SVC, Logistic Regression and Random Forest algorithms were 0.8367, 0.8189, and 0.8167, respectively, which were the three best machine-learning models.

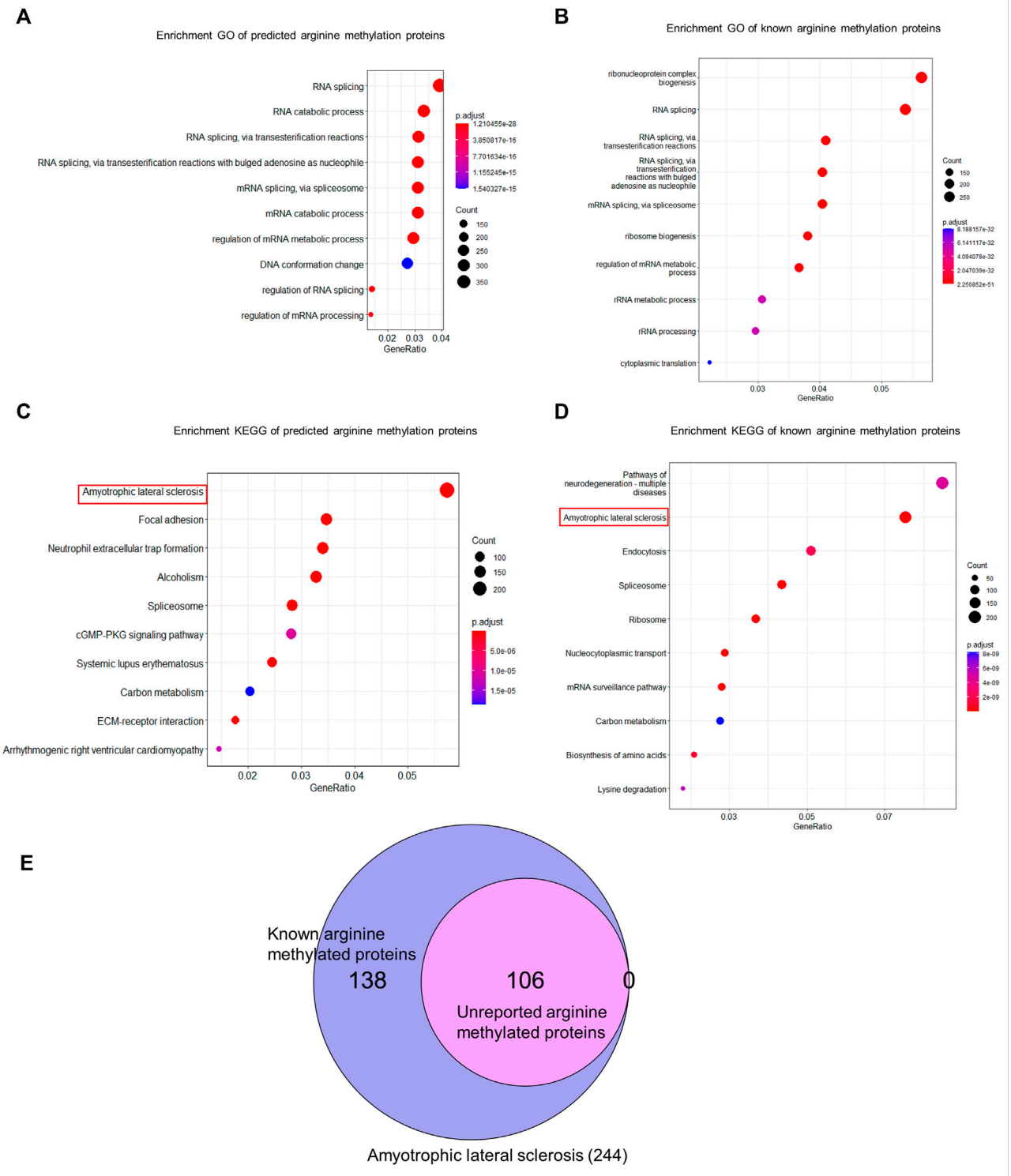

FIGURE 6
GO and KEGG analysis of human arginine methylated proteome. **(A)** GO enrichment analysis of arginine methylated proteome predicted by CNNArginineMe. **(B)** GO Enrichment analysis of known arginine methylated proteins. **(C)** KEGG enrichment analysis of arginine methylated proteome predicted by CNNArginineMe. **(D)** KEGG enrichment analysis of known arginine methylated proteins. **(E)** Venn Diagram of reported and predicted arginine methylated proteins involved in the ALS.

TABLE 3 Number of PRme sites in the ALS-related proteins.

Uniprot ID	Protein name	Number of predicted PRme sites	Number of known PRme sites
Q9Z269	VAPB	2	0
P09651	ROA1	11	8
P12036	NFH	5	0
P22626	ROA2	12	0
P35637	FUS	29	22
P41219	PERI	10	0
P50995	ANX11	3	0
P68366	TBA4A	2	0
Q13148	TADBP	2	1
Q14203	DCTN1	17	0
Q15303	ERBB4	3	0
Q53GS7	GLE1	1	0
Q96CV9	OPTN	3	0
Q96JI7	SPTCS	1	0
Q99700	ATX2	26	1
Q9UMX0	UBQL1	9	0

3.2 Performance comparison of models with different encoding approaches

With the One-Hot encoding approach, we found that the CNN algorithm and three machine-learning algorithms (i.e., SVC, Logistic Regression and Random Forest) had the highest performances compared to others. To evaluate the effect of encoding approaches on the prediction performance, we collected 17 other encoding approaches and compared them with the One-Hot approach, integrated with the three best machine-learning algorithms (see methods for details). Table 2 summarizes the AUC values of these models in terms of ten-fold cross-validation. It can be seen that the machine-learning models with three encoding approaches (i.e., BLOSUM62, One-Hot and EAAC) achieved the largest AUC values. Accordingly, we constructed CNN models using the three encoding approaches. Figure 4 shows that the average AUC value of CNN_{OH} is statistically larger than those of CNN_{EAAC} and $CNN_{BLOSUM62}$ in the independent test, although CNN_{OH} and $CNN_{BLOSUM62}$ had similar AUC values in ten-fold cross-validation. Based on these observations, we chose CNN_{OH} as the predictor of arginine methylation and named it CNNArginineMe.

3.3 Comparison of CNN_{OH} with reported predictors

To examine the predictive quality of the proposed CNNArginineMe, we compare it with reported PRme site predictors. DeepRMethylSite is the latest deep-learning predictor with the best performance compared to other reported ones (Chaudhari et al., 2020). To fairly compare CNNArginineMe and DeepRMethylSite, we used the dataset to construct DeepRMethylSite to rebuild CNNArginineMe and employed its independent test set for evaluation. Figure 5A shows that the AUC value of CNNArginineMe is 0.847, which is higher than that (0.821) of DeepRMethylSite. Furthermore, we selected two more reported predictors developed recently that provide available online prediction websites for comparison, i.e. PRmePRed (Kumar et al., 2017) and GPS-MSP (Deng et al., 2017). Due to the upload limit of the online websites, we randomly 100 sequences from our independent dataset, where the proportion of positive samples was the same as that in the independent test set. We used these 100 sequences to benchmark the three classifiers. Figure 5B shows that CNNArginineMe had the best performance among these models. Therefore, CNNArginineMe has outstanding performance for predicting PRme sites.

3.4 Prediction and functional analysis of arginine methylated proteome

We used CNNArginineMe to predict PRme sites from human proteome with the threshold value corresponding to the specificity of 0.95. We predicted 47888 PRme sites from 19023 proteins, most of which have not been reported. We performed functional analysis of the predicted arginine

methylated proteome using Gene Ontology and the KEGG pathway. The GO enrichment analysis showed that arginine methylated proteome is enriched in RNA splicing, RNA catabolic process, mRNA splicing, mRNA catabolic process, and regulation of mRNA metabolic process (Figure 6A). It is similar to the enrichment of known PRme-containing proteins (Figure 6B). Moreover, based on the KEGG pathway, the predicted arginine methylated proteome is significantly enriched in amyotrophic lateral sclerosis (ALS) (Figure 6C). This same observation could be made for reported PRme-containing proteins (Figure 6D). The ALS pathway contains 244 proteins, of which 106 without methylation annotation were predicted by CNNArginineMe (Figure 6E). According to the Amyotrophic Lateral Sclerosis Online Database, 154 proteins are linked to ALS, and 16 of them are predicted to be arginine methylated (Table 3) (Abel et al., 2012; Yun and Ha, 2020). Out of the 16 proteins, four (i.e., ATX2, FUS, ROA1, and TADBP) contain known PRme sites (Rappsilber et al., 2003; Ong et al., 2004; Guo et al., 2014). Moreover, six of the 16 proteins (i.e., ANXA11, FUS, HNRNPA2B1, HNRNPA1, TARDBP, and VAPB) have common genetic mutations in ALS, suggesting that these mutations may affect arginine methylation (Kabashi et al., 2008; Kim et al., 2013; Picchiarelli et al., 2019; Cadoni et al., 2020; Nahm et al., 2020). In summary, the arginine methylated proteome predicted by CNNArginineMe has similar enrichment features to the known arginine methylated proteins, which may assist the understanding of the functions of arginine methylation.

4 Discussion and conclusion

Many classifiers for predicting various types of PTM sites have been developed by integrating machine-learning or deep-learning algorithms with different encoding features (Chen et al., 2018a; Huang et al., 2018; Chen et al., 2019; Lyu et al., 2020; Zhang et al., 2020; Zhao et al., 2020; Sha et al., 2021; Wei et al., 2021; Zhu et al., 2022). It has been found that the models based on deep-learning algorithms have better prediction performances than those based on traditional machine-learning algorithms. The same observation is also made in this study (Table 2). The CNNArginineMe model integrating the CNN algorithm and the One-Hot encoding approach compares favourably to the machine-learning models integrating distinct algorithms and various encoding features (Table 2). These observations indicate that deep-learning algorithms must be prioritized during model construction to predict PTM sites. In this study, we compared CNNArginineMe with three reported classifiers for predicting PRme sites, i.e., DeepRMethylSite, GPS-MSP and PRmePRed. CNNArginineMe shows superior performance. It may be due to several reasons. Firstly, our dataset for model construction is relatively large, and a deep-learning model with excellent performance requires big data. Secondly, the early stop strategy is used for model construction to avoid overfitting. Nevertheless, CNNArginineMe fails to distinguish between different PRme types. Shortly, we will develop new models for predicting PTM sites with different PRme types.

We used CNNArginineMe to predict arginine methylated proteome and performed GO and KEGG analyses to understand the role of arginine methylation. Our results show that critical proteins of ALS are highly arginine methylated, implying that ALS is related to arginine methylation. Besides, arginine methylation is related to RNA splicing. This observation is consistent with the reports that gene expression is activated or repressed by arginine methylation (Fulton et al., 2019), and splicing fidelity is reduced by inhibiting symmetric or asymmetric demethylation of arginine, mediated by PRMT5 or type I PRMTs (Fong et al., 2019).

In summary, accurate identification of PRme sites could be effective in deciphering the functional and structural characteristics of protein methylation that plays an essential role in cell biology and disease mechanisms, and it will help understand transcriptional regulation, RNA splicing, DNA damage repair, cell differentiation, and apoptosis (Al-Hamashi et al., 2020).

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/guoyangzou/CNNArginineMe.

Author contributions

This study was conceived and designed by LM; HJ and GZ built models, JZ collected data sets and performed bioinformatics analysis; JZ, QL, QW, JL, and LM wrote the paper with comments from all the authors; JZ, HJ, and GZ organized pictures and tables.

Funding

This research was supported by the National Natural Science Foundation of China (grant nos. 81502065, 81672926, and 81972793) and the Natural Science Foundation of Shandong Province, China (ZR2021MC039).

Acknowledgments

We would like to thank Professor Lei Li from China Faculty of Biomedical and Rehabilitation Engineering, University of Health and Rehabilitation Science for his valuable advice and help.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1036862/full#supplementary-material

References

Abe, Y., and Tanaka, N. (2020). Fine-tuning of GLI activity through arginine methylation: Its mechanisms and function. *Cells* 9 (9), E1973. doi:10.3390/cells9091973

Abel, O., Powell, J. F., Andersen, P. M., and Al-Chalabi, A. (2012). ALSoD: A user-friendly online bioinformatics tool for amyotrophic lateral sclerosis genetics. *Hum. Mutat.* 33 (9), 1345–1351. doi:10.1002/humu.22157

Al-Hamashi, A. A., Diaz, K., and Huang, R. (2020). Non-histone arginine methylation by protein arginine methyltransferases. *Curr. Protein Pept. Sci.* 21 (7), 699–712. doi:10.2174/1389203721666200507091952

Bedford, M. T., and Clarke, S. G. (2009). Protein arginine methylation in mammals: Who, what, and why. *Mol. Cell* 33 (1), 1–13. doi:10.1016/j.molcel.2008.12.013

Blanc, R. S., and Richard, S. (2017). Arginine methylation: The coming of age. *Mol. Cell* 65 (1), 8–24. doi:10.1016/j.molcel.2016.11.003

Boulanger, M. C., Liang, C., Russell, R. S., Lin, R., Bedford, M. T., Wainberg, M. A., et al. (2005). Methylation of Tat by PRMT6 regulates human immunodeficiency virus type 1 gene expression. *J. Virol.* 79 (1), 124–131. doi:10.1128/jvi.79.1.124-131.2005

Cadoni, M. P. L., Biggio, M. L., Arru, G., Secchi, G., Orrù, N., Clemente, M. G., et al. (2020). VAPB ER-aggregates, A possible new biomarker in ALS pathology. *Cells* 9 (1), E164. doi:10.3390/cells9010164

Cai, C. Z., Han, L. Y., Ji, Z. L., Chen, X., and Chen, Y. Z. (2003). SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* 31 (13), 3692–3697. doi:10.1093/nar/gkg600

Chaudhari, M., Thapa, N., Roy, K., Newman, R. H., Saigo, H., and Dukka, B, K. C. (2020). *Mol. Omics* 16 (5), 448–454. doi:10.1039/d0mo00025f

Chen, H., Xue, Y., Huang, N., Yao, X., and Sun, Z. (2006). MeMo: A web tool for prediction of protein methylation modifications. *Nucleic Acids Res.* 34, W249–W253. Web Server issue). doi:10.1093/nar/gkl233

Chen, Z., He, N., Huang, Y., Qin, W. T., Liu, X., and Li, L. (2018a). Integration of A Deep learning classifier with A random forest approach for predicting malonylation sites. *Genomics Proteomics Bioinforma.* 16 (6), 451–459. doi:10.1016/j.gpb.2018.08.004

Chen, Z., Liu, X., Li, F., Li, C., Marquez-Lago, T., Leier, A., et al. (2019). Large-scale comparative assessment of computational predictors for lysine post-translational modification sites. *Brief. Bioinform.* 20 (6), 2267–2290. doi:10.1093/bib/bby089

Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., et al. (2018b). iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 34 (14), 2499–2502. doi:10.1093/bioinformatics/bty140

Consortium, T. U. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* 45 (D1), D158–d169. doi:10.1093/nar/gkw1099

Covic, M., Hassa, P. O., Saccani, S., Buerki, C., Meier, N. I., Lombardi, C., et al. (2005). Arginine methyltransferase CARM1 is a promoter-specific regulator of NF-kappaB-dependent gene expression. *Embo J.* 24 (1), 85–96. doi:10.1038/sj.emboj.7600500

Deng, W., Wang, Y., Ma, L., Zhang, Y., Ullah, S., and Xue, Y. (2017). Computational prediction of methylation types of covalently modified lysine and arginine residues in proteins. *Brief. Bioinform.* 18 (4), 647–658. doi:10.1093/bib/bbw041

Dong, L., Yuan, Y., and Cai, Y. (2006). Using Bagging classifier to predict protein domain structural class. *J. Biomol. Struct. Dyn.* 24 (3), 239–242.

Fedoriw, A., Rajapurkar, S. R., O'Brien, S., Gerhart, S. V., Mitchell, L. H., Adams, N. D., et al. (2019). Anti-tumor activity of the type I PRMT inhibitor, GSK3368715, synergizes with PRMT5 inhibition through MTAP loss. *Cancer Cell* 36 (1), 100–114. e125. doi:10.1016/j.ccell.2019.05.014

Fong, J. Y., Pignata, L., Goy, P. A., Kawabata, K. C., Lee, S. C., Koh, C. M., et al. (2019). Therapeutic targeting of RNA splicing catalysis through inhibition of protein arginine methylation. *Cancer Cell* 36 (2), 194–209. e199. doi:10.1016/j.ccell.2019.07.003

Fulton, M. D., Brown, T., and Zheng, Y. G. (2019). The biological Axis of protein arginine methylation and asymmetric dimethylarginine. *Int. J. Mol. Sci.* 20 (13), E3322. doi:10.3390/ijms20133322

Gil-Pita, R., and Yao, X. (2008). Evolving edited k-nearest neighbor classifiers. *Int. J. Neural Syst.* 18 (6), 459–467. doi:10.1142/s0129065708001725

Guccione, E., and Richard, S. (2019). The regulation, functions and clinical relevance of arginine methylation. *Nat. Rev. Mol. Cell Biol.* 20 (10), 642–657. doi:10.1038/s41580-019-0155-x

Guo, A., Gu, H., Zhou, J., Mulhern, D., Wang, Y., Lee, K. A., et al. (2014). Immunoaffinity enrichment and mass spectrometry analysis of protein methylation. *Mol. Cell. Proteomics* 13 (1), 372–387. doi:10.1074/mcp.O113.027870

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. *Nucleic Acids Res.* 43, D512–D520. Database issue). doi:10.1093/nar/gku1267

Hou, R., Wu, J., Xu, L., Zou, Q., and Wu, Y. J. (2020). Computational prediction of protein arginine methylation based on composition-transition-distribution features. *ACS Omega* 5 (42), 27470–27479. doi:10.1021/acsomega.0c03972

Huang, H. J., and Hsu, C. N. (2002). Bayesian classification for data from the same unknown class. *IEEE Trans. Syst. Man. Cybern. B Cybern.* 32 (2), 137–145. doi:10.1109/3477.990870

Huang, Y., He, N., Chen, Y., Chen, Z., and Li, L. (2018). Bermp: A cross-species classifier for predicting m(6)A sites by integrating a deep learning algorithm and a random forest approach. *Int. J. Biol. Sci.* 14 (12), 1669–1677. doi:10.7150/ijbs.27819

Kabashi, E., Valdmanis, P. N., Dion, P., Spiegelman, D., McConkey, B. J., Vande Velde, C., et al. (2008). TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis. *Nat. Genet.* 40 (5), 572–574. doi:10.1038/ng.132

Kim, H. J., Kim, N. C., Wang, Y. D., Scarborough, E. A., Moore, J., Diaz, Z., et al. (2013). Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. *Nature* 495 (7442), 467–473. doi:10.1038/nature11922

Kumar, P., Joy, J., Pandey, A., and Gupta, D. (2017). PRmePRed: A protein arginine methylation prediction tool. *PLoS One* 12 (8), e0183318. doi:10.1371/journal.pone.0183318

Lyu, X., Li, S., Jiang, C., He, N., Chen, Z., Zou, Y., et al. (2020). DeepCSO: A deep-learning network approach to predicting cysteine S-sulphenylation sites. *Front. Cell Dev. Biol.* 8, 594587. doi:10.3389/fcell.2020.594587

Nahm, M., Lim, S. M., Kim, Y. E., Park, J., Noh, M. Y., Lee, S., et al. (2020). ANXA11 mutations in ALS cause dysregulation of calcium homeostasis and stress granule dynamics. *Sci. Transl. Med.* 12 (566), eaax3993. doi:10.1126/scitranslmed.aax3993

Ong, S. E., Mittler, G., and Mann, M. (2004). Identifying and quantifying *in vivo* methylation sites by heavy methyl SILAC. *Nat. Methods* 1 (2), 119–126. doi:10.1038/nmeth715

Pang, H., Lin, A., Holford, M., Enerson, B. E., Lu, B., Lawton, M. P., et al. (2006). Pathway analysis using random forests classification and regression. *Bioinformatics* 22 (16), 2028–2036. doi:10.1093/bioinformatics/btl344

Parbin, S., Damodharan, S., and Rajyaguru, P. I. (2021). Arginine methylation and cytoplasmic mRNA fate: An exciting new partnership. *Yeast* 38 (8), 441–452. doi:10.1002/yea.3653

Picchiarelli, G., Demestre, M., Zuko, A., Been, M., Higelin, J., Dieterlé, S., et al. (2019). FUS-mediated regulation of acetylcholine receptor transcription at neuromuscular junctions is compromised in amyotrophic lateral sclerosis. *Nat. Neurosci.* 22 (11), 1793–1805. doi:10.1038/s41593-019-0498-9

Poulard, C., Corbo, L., and Le Romancer, M. (2016). Protein arginine methylation/demethylation and cancer. *Oncotarget* 7 (41), 67532–67550. doi:10.18632/oncotarget.11376

Qiu, W. R., Xiao, X., Lin, W. Z., and Chou, K. C. (2014). iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. *Biomed. Res. Int.* 2014, 947416. doi:10.1155/2014/947416

Rappsilber, J., Friesen, W. J., Paushkin, S., Dreyfuss, G., and Mann, M. (2003). Detection of arginine dimethylated peptides by parallel precursor ion scanning mass spectrometry in positive ion mode. *Anal. Chem.* 75 (13), 3107–3114. doi:10.1021/ac026283q

Ratovitski, T., Arbez, N., Stewart, J. C., Chighladze, E., and Ross, C. A. (2015). PRMT5- mediated symmetric arginine dimethylation is attenuated by mutant huntingtin and is impaired in Huntington's disease (HD). *Cell Cycle* 14 (11), 1716–1729. doi:10.1080/15384101.2015.1033595

Scopino, K., Dalgarno, C., Nachmanoff, C., Krizanc, D., Thayer, K. M., and Weir, M. P. (2021). Arginine methylation regulates ribosome CAR function. *Int. J. Mol. Sci.* 22 (3), 1335. doi:10.3390/ijms22031335

Sha, Y., Ma, C., Wei, X., Liu, Y., Chen, Y., and Li, L. (2021). DeepSADPr: A hybrid-learning architecture for serine ADP-ribosylation site prediction. *Methods* 203, 575–583. doi:10.1016/j.ymeth.2021.09.008

Shao, J., Xu, D., Tsai, S. N., Wang, Y., and Ngai, S. M. (2009). Computational identification of protein methylation sites through bi-profile Bayes feature extraction. *PLoS One* 4 (3), e4920. doi:10.1371/journal.pone.0004920

Shi, S. P., Qiu, J. D., Sun, X. Y., Suo, S. B., Huang, S. Y., and Liang, R. P. (2012). PMeS: Prediction of methylation sites based on enhanced feature encoding scheme. *PLoS One* 7 (6), e38772. doi:10.1371/journal.pone.0038772

Shien, D. M., Lee, T. Y., Chang, W. C., Hsu, J. B., Horng, J. T., Hsu, P. C., et al. (2009). Incorporating structural characteristics for identification of protein methylation sites. *J. Comput. Chem.* 30 (9), 1532–1543. doi:10.1002/jcc.21232

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochem. Med.* 24 (1), 12–18. doi:10.11613/bm.2014.003

Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol. Methods* 14 (4), 323–348. doi:10.1037/a0016973

Szewczyk, M. M., Ishikawa, Y., Organ, S., Sakai, N., Li, F., Halabelian, L., et al. (2020). Pharmacological inhibition of PRMT7 links arginine monomethylation to the cellular stress response. *Nat. Commun.* 11 (1), 2396. doi:10.1038/s41467-020-16271-z

Wang, D., Liang, Y., and Xu, D. (2019a). Capsule network for protein post-translational modification site prediction. *Bioinformatics* 35 (14), 2386–2394. doi:10.1093/bioinformatics/bty977

Wang, D., Zeng, S., Xu, C., Qiu, W., Liang, Y., Joshi, T., et al. (2017). MusiteDeep: A deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics* 33 (24), 3909–3916. doi:10.1093/bioinformatics/btx496

Wang, S. M., Dowhan, D. H., and Muscat, G. E. O. (2019b). Epigenetic arginine methylation in breast cancer: Emerging therapeutic strategies. *J. Mol. Endocrinol.* 62 (3), R223–r237. doi:10.1530/jme-18-0224

Wei, L., Xing, P., Shi, G., Ji, Z., and Zou, Q. (2019). Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16 (4), 1264–1273. doi:10.1109/tcbb.2017.2670558

Wei, X. W., Sha, Y., Zhao, Y., He, N., and Li, L. (2021). DeepKcrot: A deep-learning architecture for general and species-specific lysine crotonylation site prediction. *IEEE Access* 9, 49504–49513. doi:10.1109/ACCESS.2021.3068413

Wen, B., Zeng, W. F., Liao, Y., Shi, Z., Savage, S. R., Jiang, W., et al. (2020). Deep learning in proteomics. *Proteomics* 20 (21-22), e1900335. doi:10.1002/pmic.201900335

Wen, P. P., Shi, S. P., Xu, H. D., Wang, L. N., and Qiu, J. D. (2016). Accurate *in silico* prediction of species-specific methylation sites based on information gain feature optimization. *Bioinformatics* 32 (20), 3107–3115. doi:10.1093/bioinformatics/btw377

Wilkins, M. R., Gasteiger, E., Gooley, A. A., Herbert, B. R., Molloy, M. P., Binz, P. A., et al. (1999). High-throughput mass spectrometric discovery of protein post-translational modifications. *J. Mol. Biol.* 289 (3), 645–657. doi:10.1006/jmbi.1999.2794

Yang, Y., and Bedford, M. T. (2013). Protein arginine methyltransferases and cancer. *Nat. Rev. Cancer* 13 (1), 37–50. doi:10.1038/nrc3409

Yun, Y., and Ha, Y. (2020). CRISPR/Cas9-Mediated gene correction to understand ALS. *Int. J. Mol. Sci.* 21 (11), E3801. doi:10.3390/ijms21113801

Zhang, L., Zou, Y., He, N., Chen, Y., Chen, Z., and Li, L. (2020). DeepKhib: A deep-learning framework for lysine 2-hydroxyisobutyrylation sites prediction. *Front. Cell Dev. Biol.* 8, 580217. doi:10.3389/fcell.2020.580217

Zhao, Y., He, N., Chen, Z., and Li, L. (2020). Identification of protein lysine crotonylation sites by a deep learning framework with convolutional neural networks. *IEEE Access* 8, 14244–14252. doi:10.1109/ACCESS.2020.2966592

Zhu, Y., Liu, Y., Chen, Y., and Li, L. (2022). ResSUMO: A deep learning architecture based on residual structure for prediction of lysine SUMOylation sites. *Cells* 11 (17), 2646. doi:10.3390/cells11172646