frontiers
in Cell and Developmental Biology

TECHNOLOGY AND CODE
published: 23 February 2021
doi: 10.3389/fcell.2021.617366

pCysMod: Prediction of Multiple Cysteine Modifications Based on Deep Learning Framework

Shihua Li[1,2†], *Kai Yu*[1†], *Guandi Wu*[1†], *Qingfeng Zhang*[1], *Panqin Wang*[2], *Jian Zheng*[1], *Ze-Xian Liu*[1], *Jichao Wang*[3*], *Xinjiao Gao*[4*] *and Han Cheng*[2*]

[1] State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou, China, [2] School of Life Sciences, Zhengzhou University, Zhengzhou, China, [3] CAS Key Lab of Biobased Materials, Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao, China, [4] MOE Key Laboratory for Membraneless Organelles and Cellular Dynamics, Hefei National Laboratory for Physical Sciences at the Microscale, University of Science and Technology of China, Hefei, China

OPEN ACCESS

Edited by:
Jiangning Song,
Monash University, Australia

Reviewed by:
Tzong-Yi Lee,
The Chinese University of Hong Kong, China
Cangzhi Jia,
Dalian Maritime University, China

***Correspondence:**
Jichao Wang
wangjc@qibebt.ac.cn
Xinjiao Gao
gaox@ustc.edu.cn
Han Cheng
chenghan@zzu.edu.cn

† *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Cellular Biochemistry, a section of the journal Frontiers in Cell and Developmental Biology

Received: *14 October 2020*
Accepted: *12 January 2021*
Published: *23 February 2021*

Citation:
Li S, Yu K, Wu G, Zhang Q, Wang P, Zheng J, Liu Z-X, Wang J, Gao X and Cheng H (2021) pCysMod: Prediction of Multiple Cysteine Modifications Based on Deep Learning Framework. Front. Cell Dev. Biol. 9:617366. doi: 10.3389/fcell.2021.617366

Thiol groups on cysteines can undergo multiple post-translational modifications (PTMs), acting as a molecular switch to maintain redox homeostasis and regulating a series of cell signaling transductions. Identification of sophistical protein cysteine modifications is crucial for dissecting its underlying regulatory mechanism. Instead of a time-consuming and labor-intensive experimental method, various computational methods have attracted intense research interest due to their convenience and low cost. Here, we developed the first comprehensive deep learning based tool pCysMod for multiple protein cysteine modification prediction, including *S*-nitrosylation, *S*-palmitoylation, *S*-sulfenylation, *S*-sulfhydration, and *S*-sulfinylation. Experimentally verified cysteine sites curated from literature and sites collected by other databases and predicting tools were integrated as benchmark dataset. Several protein sequence features were extracted and united into a deep learning model, and the hyperparameters were optimized by particle swarm optimization algorithms. Cross-validations indicated our model showed excellent robustness and outperformed existing tools, which was able to achieve an average AUC of 0.793, 0.807, 0.796, 0.793, and 0.876 for *S*-nitrosylation, *S*-palmitoylation, *S*-sulfenylation, *S*-sulfhydration, and *S*-sulfinylation, demonstrating pCysMod was stable and suitable for protein cysteine modification prediction. Besides, we constructed a comprehensive protein cysteine modification prediction web server based on this model to benefit the researches finding the potential modification sites of their interested proteins, which could be accessed at http://pcysmod.omicsbio.info. This work will undoubtedly greatly promote the study of protein cysteine modification and contribute to clarifying the biological regulation mechanisms of cysteine modification within and among the cells.

Keywords: protein cysteine modifications, feature extraction, deep learning, post-translational modifications, prediction

Abbreviations: PTMs, post-translational modifications; Cys, cysteine; H_2S, Hydrogen sulfide; NO, nitric oxide; SVM, support vector machine; ESC, embryonic stem cell; PSO, particle swarm optimization; CKSAAP, composition of k-spaced amino acid pairs; BE, binary encoding profiles; PSSM, position-specific scoring matrix; AAC, amino acid composition; Sp, specificity; Sn, sensitivity; Ac, accuracy; ROC, receiver operating characteristic; AUC, area under ROC curve; MCC, Mathews correlation coefficient; DNN, deep neural network; GCN, graph convolutional neural network.

INTRODUCTION

Post-translational modifications (PTMs) occur at specific amino acids extending the chemical repertoire of the 20 standard amino acids, which reversibly coordinate the signaling networks (Mann and Jensen, 2003; Mertins et al., 2013; Strzyz, 2016). Although cysteine (Cys) appears the least frequently among these common amino acids, it tends to act as a powerful molecular switch to maintain redox homeostasis and regulate a series of cell signaling transductions by PTMs (Marino and Gladyshev, 2011). The susceptibility of Cys to a variety of oxidative post-translational modifications is mainly dependent on the thiol groups, which are considerably more easily oxidized and highly nucleophilic (Brandes et al., 2009; Kumsta et al., 2011). According to different molecular conjugations to the thiol groups, cysteine modification can be classified into different types. Nitric oxide (NO) binding to some cysteine resides causes *S*-nitrosylation (Jia J. et al., 2014) and hydrogen sulfide (H2S) causes *S*-sulfhydration (Mishanina et al., 2015; Yang et al., 2015). Cumulated H_2O_2 reacting with cysteine leads to *S*-sulfenylation (Yang et al., 2015), *S*-sulfinylation (Akter et al., 2018), and *S*-sulfonylation (Lim et al., 2008). Cysteines can also bind metals such as Cu, Zn, and Fe to form iron-sulfur clusters and zinc finger domains (Oteiza, 2012; Rouault, 2015). The thioesterification reaction happened on lipid including *S*-prenylation and *S*-palmitoylation (Roth et al., 2006). These modifications lead to a cascade of biochemical reactions and regulate various physiological and pathological processes, such as autophagy (Carroll et al., 2018), protein stabilization (Kröncke and Klotz, 2009), redox homeostasis (Fra et al., 2017), and cell signaling (Hourihan et al., 2016), demonstrating a close relationship with many human diseases including cancers, diabetes, and so on. In this regard, to dissect the molecular mechanisms and regulatory roles of cysteine modification, it is urgently needed to precisely parse the potential cysteine modification sites and types.

With the rapid development of high-throughput sequencing and excellent specific chemical probes, cysteine modification profiles get unprecedented accumulation. For example, through a low-PH quantitation method, Fu et al. (2019) detected 1,547 sulfhydration sites on 994 proteins. Akter et al. (2018) identified and quantified 387 *S*-sulfinylation sites on 296 proteins in A549 and Hela cells. Recently, with label-free quantification strategy, Shen et al. (2017) identified 2,190 *S*-palmitoylated peptides on 883 proteins in liver. However, because the experimental methods are time consuming and labor intensive, the cysteine modification profiles expanded slowly, which significantly restricted the research on dissecting the molecular functions of cysteine modification. It is necessary to develop *in silico* tools to accurately predict cysteine modification sites, which will definitely promote the experimental identification of sophistical protein cysteine modification sites and types.

There are several computational tools used for predicting distinct cysteine modification types. For *S*-nitrosylation, Xue et al. (2010) collected 504 modification sites and constructed the first tool GPS-SNO for predicting *S*-nitrosylation sites. SNOSite (Lee et al., 2011b) predicted *S*-nitrosylation sites based on 586 experimental sites using support vector machine (SVM). iSNO-ANBPB (Jia C. et al., 2014) mainly adopted an adapted normal distribution bi-profile Bayes (ANBPB) feature extraction model. PreSNO (Hasan et al., 2019) used the LR model to integrate four encoding schemes with support vector machines and RF algorithms to predict SNO sites. In 2018, Xie et al. (2018) developed DeepNitro for the prediction of protein nitration and nitrosylation sites based on deep learning. iSulf-Cys (Xu et al., 2016) is the first program designed for predicting *S*-sulfenylation sites based on 1,105 sites quantified in RKO cells. Ju and Wang (2018) improved the model performance and developed Sulf_FSVM. MDD-Palm (Weng et al., 2017) can identify *S*-palmitoylation sites based on SVM. Recently, Ning et al. (2020) developed GPS-Palm using a deep learning based graphic presentation system for the prediction of *S*-palmitoylation. Although numerous predictors with considerable performance have been developed, the limitations are that all of these tools can predict just one kind of modification type and there is still room for improvement in model performance, while some modification types such as *S*-sulfinylation and *S*-sulfhydration are still lacking excellent predictors.

Previously, we have developed several protein post-translational modification tools for enzyme-specific lysine acetylation (Yu et al., 2020), calpain-specific cleavage site (Liu et al., 2019), and *S*-glutathionylation site (Li et al., 2020) prediction based on deep learning framework and particle swarm optimization (PSO) algorithm, which achieved significantly better performance than exiting tools. Traditional machine learning based method requires careful feature selection and scaling, which limited its performance. However, as a branch of machine learning, deep learning based method can fit high-dimensional features and clarify biological problems better than other algorithms. For example, Xu et al. (2017) constructed a predicting system for histone modification and discovered a potential embryonic stem cell (ESC) fate decision mechanism. DeepBind (Hassanzadeh and Wang, 2016) provided many candidate DNA-binding proteins by predicting DNA and protein-binding events. These results suggested an unprecedented excellent chance to utilize deep learning to solve biological problems. However, a credible deep learning framework is still lacking for comprehensive cysteine modification prediction.

In this work, after integrating the experimentally verified cysteine sites curated from literature and sites collected by other databases and predicting tools, we developed the first comprehensive deep learning based tool pCysMod for multiple protein cysteine modification prediction, including *S*-nitrosylation, *S*-palmitoylation, *S*-sulfenylation, *S*-sulfhydration, and *S*-sulfinylation. Seven sequence features including binary encoding profiles (BE), amino acid composition (AAC), position-specific scoring matrix (PSSM), and composition of k-spaced amino acid pairs (CKSAAP) were used to represent the sequences. These features were extracted and united into a deep learning model, and the hyperparameters were optimized by particle swarm optimization algorithms. Cross-validations indicated our model showed excellent robustness and outperformed existing tools. Besides, we constructed a comprehensive protein cysteine modification prediction web server based on this model to benefit the researches finding the potential modification sites of their interested proteins, which could be accessed at http://pcysmod.omicsbio.info.

METHODS

Benchmark Dataset Preparation

The cysteine modification sites were collected in two major aspects. On the one hand, we curated the experimentally verified sites by searching the literatures from PubMed. For each modification, we used "nitrosylation," "palmitoylation," "sulfenylation," "sulfhydration," and "sulfinylation," together with "cysteine" as our keywords. After traversing all related literatures in PubMed, we manually collected all experimentally verified sites. One the other hand, several databases and predictors with known cysteine modification sites were integrated, including GPS-SNO training dataset (Xue et al., 2010), Deep-Nitro training dataset (Xie et al., 2018), SNOSite training dataset (Lee et al., 2011b), GPS-Palm training dataset (Ning et al., 2020), iSulf-Cys training dataset (Xu et al., 2016), Sulf_FSVM training dataset (Ju and Wang, 2018), and dbPTM database (Huang et al., 2018). Finally, we obtained 23,041 *S*-nitrosylation sites in 10,671 proteins, 2,766 *S*-palmitoylation sites in 1,413 proteins, 4,978 *S*-sulfenylation sites in 3,288 proteins, 2,721 *S*-sulfhydration sites in 1,707 proteins, and 742 *S*-sulfinylation sites in 538 proteins as our final training dataset (**Table 1** and **Supplementary Table S1**).

TABLE 1 | A summary of each type of modification data.

Dataset	Human	Mouse	Rat	Other	Total
Number of *S*-nitrosylation sites (positive data)	10,784	4,103	1,629	2,819	38,670
Number of non-*S*-nitrosylation sites (negative data)	19,335				
Number of *S*-palmitoylation sites (positive data)	748	1,773	74	174	5,532
Number of non-*S*-palmitoylation sites (negative data)	2,766				
Number of *S*-sulfenylation sites (positive data)	2,587	352	1	1,806	9,492
Number of non-*S*-sulfenylation sites (negative data)	4,746				
Number of *S*-sulfhydration sites (positive data)	2,010	0	0	525	5,070
Number of non-*S*-sulfhydration sites (negative data)	2,535				
Number of *S*-sulfinylation sites (positive data)	440	0	208	7	1,310
Number of non-*S*-sulfinylation sites (negative data)	655				

To generate the positive and negative datasets, we retrieved the protein sequence from UniProt database (UniProt Consortium [UC], 2015) for each protein. For each modification, the golden positive dataset was the modification sites from the benchmark dataset, whereas all cysteine sites that were not modified on the same protein were treated as the negative dataset. The sequence box for feature extraction consists of a cysteine in the middle and 15 upstream and downstream amino acids at both ends. For the peptide of less than 31-amino acids, pseudo-amino acids "*" were added to make sure the peptides were of equal length. If the sequence in the negative dataset was the same as the positive set in the same cysteine modification, only the sequence in the positive data set is preserved. In addition, due to the high imbalance between positive and negative samples, we randomly selected the same number of negative samples to ensure that the number of positive peptides was equal to the number of negative peptides (Zhao et al., 2012). At the same time, we used CD-Hit (Fu et al., 2012) with a threshold of 90, 80, and 70% sequence similarity treatment on a short peptide consisting of 31-amino acids, and then performed fivefold cross-validation. In this work, cross-validations were used to evaluate the performance of the model. Since cross-validation is an efficient way of examining the robustness and accuracy of a predicting model, it is unnecessary to divide the benchmark dataset into training set and testing set (Zhang et al., 2020).

Feature Extraction

Binary Encoding Profiles

Binary encoding (BE) (Song et al., 2010) was derived from computational programming, which uses the binary digit, that is, "0" or "1," as the fundamental unit of information. Each printable character can be uniquely represented by combining bits. As mentioned above, each peptide in the benchmark dataset consists of at most 21 types of amino acids, which are ACDEFGHIKLMNPQRSTVWY*. Hence, a 21-dimentional binary vector was used to represent each amino acid. For example, "A" was encoded as (100000000000000000000), "E" was encoded as (000100000000000000000), and the pseudo-amino acid "*" was encoded as (000000000000000000001). In this regard, each peptide was represented by a 651-dimensional vector.

Amino Acid Composition

The amino acid composition (AAC) is an important feature to identify β-barrel membrane proteins (Radivojac et al., 2010; Lee et al., 2011a), which stand for the occurrence frequency of 21-amino acids on any specific peptides. The feature length of this encoding method is 21 for each peptide.

Position-Specific Scoring Matrix

Position-Specific Scoring Matrix (PSSM) was first introduced as an alternative to consensus sequences (Stormo et al., 1982); this feature was derived from a set of functionally related aligned sequences, which is commonly used for computational motif discovery in biological sequences (Stormo, 2000). For a group of given peptides, PSSMs assume that the probabilities for each position are statistically independent and calculate the probability for each specific amino acid at a particular position. The probabilities for a particular position sum up to 1. In this work, we calculated PSSMs for positive dataset and negative dataset, so the dimension of this feature is 62.

Composition of *k*-Spaced Amino Acid Pairs

The encoding scheme based on the Composition of *k*-Spaced Amino Acid Pairs (CKSAAP) (Zhao et al., 2012) is an effective feature extraction method, which can reflect the information of amino acid pair motifs in a set of peptides. The *k*-spaced means two amino acids in a peptide separated by k-amino acids, and CKSAAP encoding calculates the occurrence frequency for each pair. When $k = 0$, it means the occurrence frequency of each pair is composed of adjacent amino acids, and the dimension is 441. In this work, after taking computation and time cost into consideration, we merely adopted $k = 0$, 1, 2, and 3, and the final dimension of this method is 1,764.

Model Construction

Although each modification type has a special benchmark dataset and needs a special model to fit, they have analogous model architectures. Here, we introduce a general deep learning based model architecture used in this work for cysteine prediction. For each modification type, the benchmark peptide dataset was encoded by four feature extraction methods mentioned above. The model received the numerical transferred sequences in the input layer, which consists of four independent DNN submodules to train four input features. Then the four submodules were merged and flattened into a fully connected layer after sufficiently learning the features. Finally, pCysMod output a probability of whether this peptide could undergo particular modification. Early stopping and dropout functions were used to make sure the training set was not over-represented. To optimize the numerous hyperparameters in pCysMod, particle swarm optimization algorithm was applied to generate the maximum performance as previously reported (Yu et al., 2020). The python package "pyswarm"[1] as used.

Performance Evaluation

Four common measurements were adopted to evaluate the performance of pCysMod as previously described (Liu et al., 2012), including specificity (Sp), sensitivity (Sn), accuracy (Ac), and Mathews correlation coefficient (MCC). The detailed descriptions of these four measurements are defined as below:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{1}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{2}$$

$$\mathrm{Ac} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{3}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}^{*}\mathrm{TN} - \mathrm{FP}^{*}\mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{4}$$

We calculated the area under the receiver operating characteristic (ROC) curve (AUC) values to show the model performance. Four-, six-, eight-, and tenfold cross-validations were used to evaluate the robustness and accuracy of pCysMod. Tenfold cross-validation was used to compare the performance of pCysMod with the existing tools.

Implement of the Web Server

pCysMod model was constructed by Keras, with TensorFlow as its backend implementation. The secondary structure and surface accessibility information of the query sequence were calculated by NetSurfP (Petersen et al., 2009), and the disorder information was predicted by IUPred (Dosztanyi et al., 2005). The web server was built in PHP and Python, which could be accessed at http://pcysmod.omicsbio.info.

RESULTS

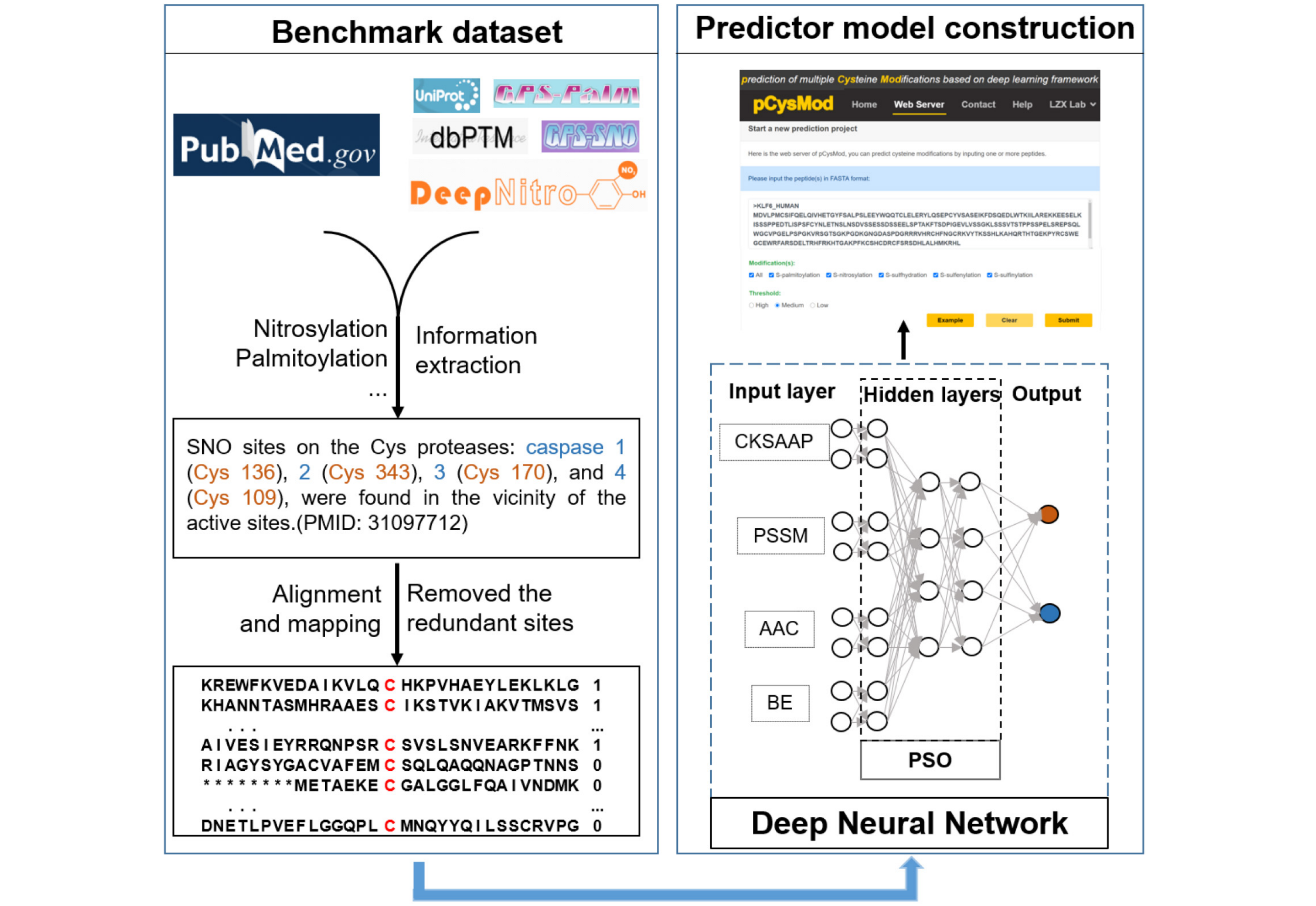
The Construction of Computational Model to Predict Cysteine Modification Sites

Cysteine modification sites were obtained in the literature and other predictive tools (**Figure 1**). After removing redundant sequences and balancing the datasets, we finally obtained 19,335 *S*-nitrosylation-positive sites, 2,766 *S*-palmitoylation-positive sites, 4,746 *S*-sulfenylation-positive sites, 2,535 *S*-sulfhydration-positive sites, and 655 *S*-sulfinylation-positive sites. The number of negative and positive sequences of different modifications was the same and shown in **Table 1**. Then, we developed the first model to predict multiple cysteine modifications named pCysMod. The software was based on deep learning and PSO algorithm. The sequence features were extracted by four methods, including BE, AAC, PSSM, and CKSAAP (**Figure 1**). Furthermore, we used Python, PHP, JavaScript, and HTML to construct pCysMod online server, which can be accessed through http://pcysmod.omicsbio.info.

The Characteristic of Cysteine Modification Sites and Proteins

To better understand the structure of different cysteine modification sites, we used the secondary structure prediction algorithms PsiPred (McGuffin et al., 2000) and IUPred (Dosztanyi et al., 2005) to classify the cysteine sites of all proteins. The *S*-nitrosylation sites and *S*-palmitoylation sites were predominantly distributed in coil, while *S*-sulfenylation, *S*-sulfhydration, and *S*-sulfinylation sites in coil and helix were relatively close (**Figure 2A**), and the cysteine sites were mainly predicted to locate in ordered regions (**Figure 2B**). Furthermore, we used Two Sample Logo (Vacic et al., 2006) to analyze amino acid preference. The difference between *S*-sulfinylation sites and non-*S*-sulfinylation sites are shown in **Figure 2C**. Lysine and asparagine residues were enriched around the *S*-sulfinylation sites, but cysteine residues were deleterious to the modification. In *S*-nitrosylation cysteine modification, the asparagine and glutamic were enriched near the modification

[1] https://github.com/tisimst/pyswarm


Benchmark dataset
PubMed.gov
UniProt
GPS-Palm
dbPTM
GPS-SNO
DeepNitro
NO
OH
Nitrosylation
Palmitoylation
...
Information
extraction
SNO sites on the Cys proteases: caspase 1 (Cys 136), 2 (Cys 343), 3 (Cys 170), and 4 (Cys 109), were found in the vicinity of the active sites.(PMID: 31097712)
Alignment
and mapping
Removed the
redundant sites
KREWFKVEDAIKVLQ C HKPVHAEYLEKLKLG 1
KHANNTASMHRAAES C IKSTVKIAKVTMSVS 1
. . . ...
AIVESIEYRRQNPSR C SVSLSNVEARKFFNK 1
RIAGYSYGACVAFEM C SQLQAQQNAGPTNNS 0
********METAEKE C GALGGLFQAIVNDMK 0
. . . ...
DNETLPVEFLGGQPL C MNQYYQILSSCRVPG 0
Predictor model construction
Input layer
Hidden layers
Output
CKSAAP
PSSM
AAC
BE
PSO
Deep Neural Network

prediction of multiple Cysteine Modifications based on deep learning framework

pCysMod
Home
Web Server
Contact
Help
LZX Lab

Start a new prediction project

Here is the web server of pCysMod, you can predict cysteine modifications by inputing one or more peptides.

Please input the peptide(s) in FASTA format:

```
>KLF6_HUMAN
MDVLPMCSIFQELQIVHETGYFSALPSLEEYWQQTCLELERYLQSEPCYVSASEIKFDSQEDLWTKIILAREKKEESELK
ISSSPPEDTLISPSFCYNLETNSLNSDVSSESSDSSEELSPTAKFTSDPIGEVLVSSGKLSSSVTSTPPSSPELSREPSQL
WGCVPGELPSPGKVRSGTSGKPGDKGNGDASPDGRRRVHRCHFNGCRKVYTKSSHLKAHQRTHTGEKPYRCSWE
GCEWRFARSDELTRHFRKHTGAKPFKCSHCDRCFSRSDHLALHMKRHL
```

Modification(s):

☑ All ☑ S-palmitoylation ☑ S-nitrosylation ☑ S-sulfhydration ☑ S-sulfenylation ☑ S-sulfinylation

Threshold:

○ High ● Medium ○ Low

FIGURE 1 | An overview of the model.

site (**Figure 2C**). Lysine residues also tended to be *S*-sulfenylated and *S*-sulfhydrated, while cysteine residues were enriched in *S*-palmitoylation cysteine modification (**Figure 2C**).

Using the collected human proteins with different cysteine modifications, we conducted GO and KEGG enrichment by clusterProfiler (Yu et al., 2012). We found that the mostly enriched biological processes were catabolic process in *S*-sulfinylation and *S*-nitrosylation, such as carboxylic acid catabolic process and organic acid catabolic process (**Supplementary Figure S1**). *S*-Sulfenylation and *S*-sulfhydration were related to transcription, and *S*-palmitoylation tended to affect transduction (**Supplementary Figure S1**). Based on the enrichment results of GO cellular components, we observed that ribosome was enriched in different cysteine modifications (**Supplementary Figure S1**). GO molecular function and KEGG pathway analyses also indicated that the cysteine modifications other than *S*-palmitoylation were involved in the redox process (**Supplementary Figures S1, S2**). The results were consistent with previous studies, which showed that *S*-nitrosylation, *S*-sulfenylation, *S*-sulfhydration, and *S*-sulfinylation play critical roles in oxidative post-translational modifications (Chung et al., 2013).

Evaluating the Performance of pCysMod

We generated the first model to predict multiple types of cysteine modification based on the method mentioned above. Four-, six-, eight-, and tenfold cross-validations were used to evaluate the accuracy and robustness of pCysMod. The ROC curves and AUC values are displayed in **Figure 3**. The best cross-validation AUC values for *S*-nitrosylation, *S*-palmitoylation, *S*-sulfenylation, *S*-sulfhydration, and *S*-sulfinylation were 0.793, 0.807, 0.796, 0.793, and 0.876. The similar and considerable performance declared the robustness and high accuracy of pCysMod. Since cross-validation is an efficient way of examining the robustness and accuracy of a predicting model, it is unnecessary to divide the benchmark dataset into training set and testing set (Zhang et al., 2020). We tested the predictive performance of different feature extractions. The fivefold cross-validation AUCs were calculated for different features, and the results are visualized in the added **Supplementary Figure S3**, which indicated that combining multiple features can obtain more stable prediction performances. Not only that, in order to avoid the overestimation of prediction performance due to the possible high similarity of the sequences, we used CD-Hit with a threshold of 70, 80, and 90% sequence similarity

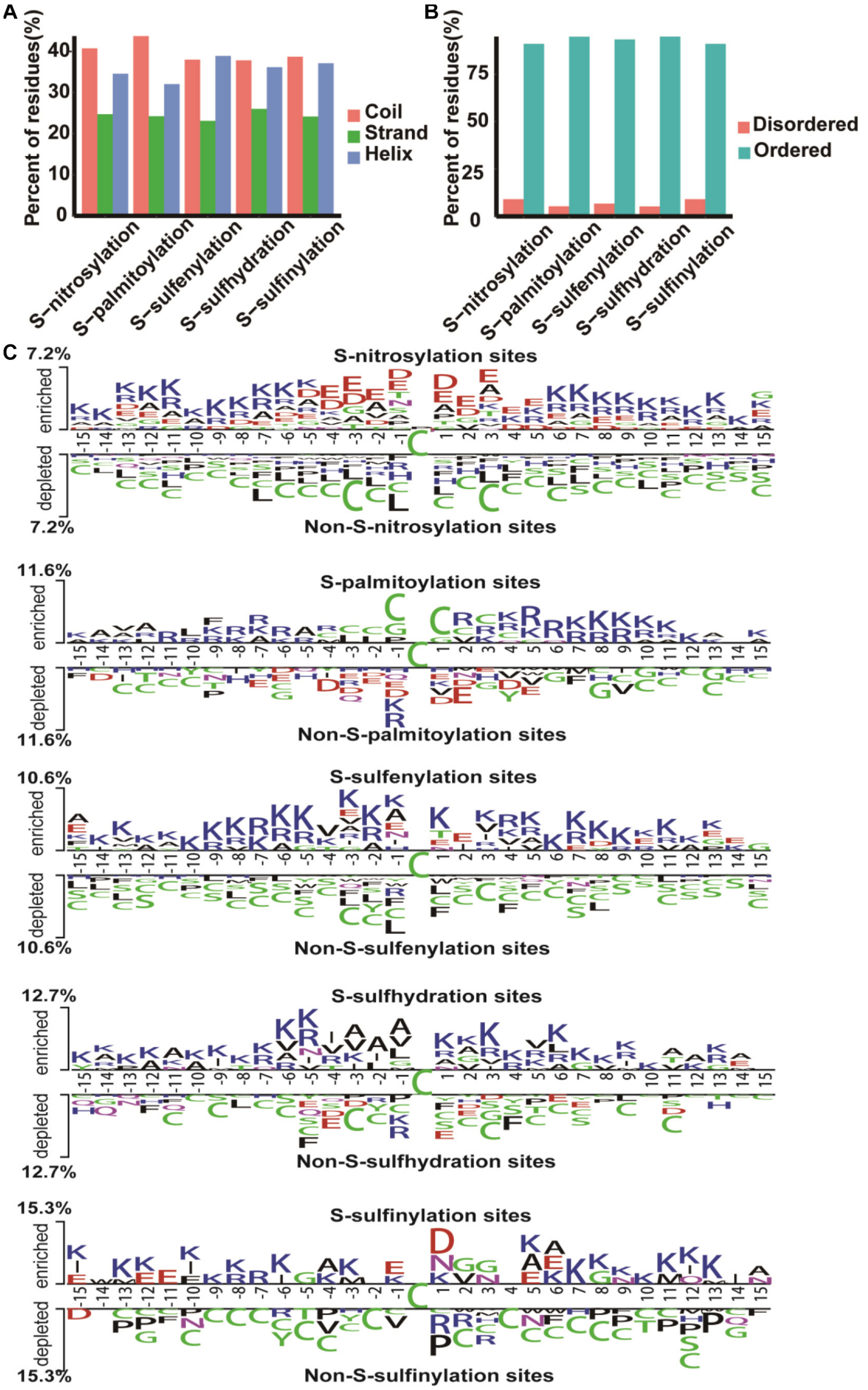


FIGURE 2 | The characteristic of cysteine modification sites and proteins. **(A)** The secondary structure. **(B)** The disorder information of cysteine modification sites. **(C)** Preference for amino acids around the cysteine modification sites and non-cysteine modification sites.

analysis on short peptides composed of 31-amino acids, and then performed fivefold cross-validation based on the clustering results. Compared with only removing redundant peptides, the results showed that not using CD-Hit did not lead to an overestimation of the prediction performance (**Supplementary Table S2**).

We then performed tenfold cross-validation to demonstrate the superiority of pCysMod compared with existing tools, including *S*-nitrosylation site-predicting tools GPS-SNO (Xue et al., 2010), Deep-Nitro (Xie et al., 2018), iSNO-ANBPB (Jia C. et al., 2014), and PreSNO (Hasan et al., 2019), *S*-palmitoylation site-predicting tools GPS-Palm (Ning et al.,

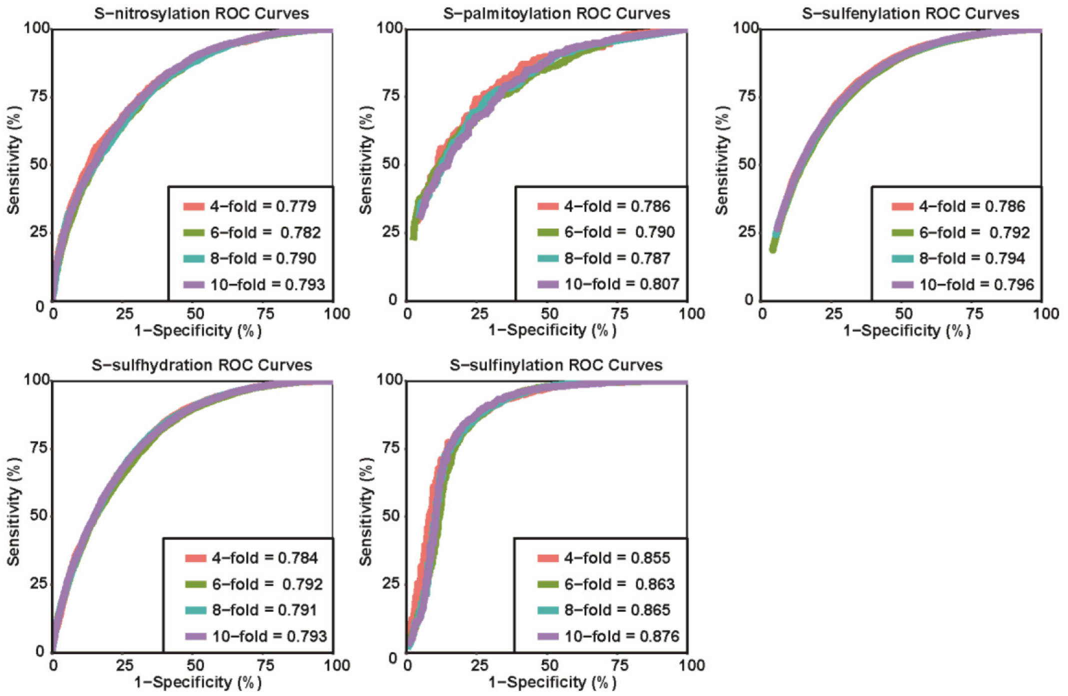


FIGURE 3 | The ROC curves and AUCs of 4-, 6-, 8-, and tenfold cross-validations are shown.

2020) and MDD-Palm (Weng et al., 2017), and *S*-sulfenylation site-predicting tools iSulf-Cys (Xu et al., 2016) and Sulf_FSVM (Ju and Wang, 2018). The performances of these predictors were retrieved from previous reported literatures, which are shown in **Table 2**. Through the comparison, we can conclude that the performance of pCysMod is higher than or equal to existing predictors, showing a considerable predictive power for general cysteine modification prediction.

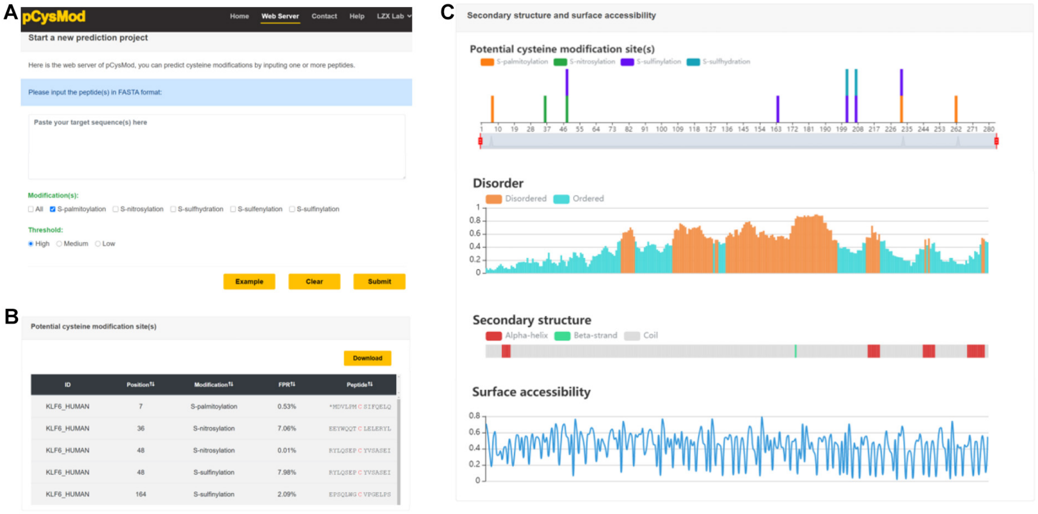
Finally, we have constructed an independent predictor for each modification, with the same basic structure and distinct hyperparameters. At the same time, we tested the cross differentiating capabilities of five cysteine modification predictors, that is, using the constructed model to predict other types of cysteine modification. The prediction results show that, different predictors have specificity for their corresponding modification type (**Supplementary Table S3**). Although the basic structure of each modified model is the same, the internal parameters adjusted by the PSO algorithm are distinct, showing a different modification feature and pattern of each modification type.

TABLE 2 | Performance comparison of pCysMod with other predictors.

CysMod	Predictor	Sn (%)	Sp (%)	Ac (%)	MCC	AUC
S-Nitrosylation	GPS-SNO	53.57	80.14	75.80	0.286	0.524
	DeepNitro	40.0	85.0	77.7	0.236	0.743
	PreSNO	60.4	76.9	75.2	0.252	0.756
	iSNO-ANBPB			67.01	0.351	
	pCysMod	61.09	80.02	70.57	0.420	0.793
S-Palmitoylation	GPS-Palm	68.47	85.04	82.67	0.448	0.855
	MDD-Palm	74.0	74.0	74.0	0.40	0.80
	pCysMod	62.91	80.29	71.66	0.439	0.807
S-Sulfenylation	iSulf-Cys	67.31	63.89	65.59	0.312	0.715
	Sulf_FSVM	68.54	68.03	68.29	0.365	0.747
	pCysMod	75.66	70.08	72.84	0.458	0.796

Implementation of pCysMod Web Server

In order to provide an efficient and convenient way to facilitate basic research, we generated the first comprehensive cysteine modification prediction web server pCysMod. We tested the pCysMod website on various commonly used web browsers, such as Google Chrome, Internet Explorer, and Mozilla Firefox to provide a robust service. The prediction and results pages are shown in **Figure 3**. The input text box required FASTA format protein sequence, and then we should select which type of modification is needed to be predicted and its threshold (**Figure 4A**). The prediction information was organized by two aspects and displayed in the results page, including "Potential cysteine modification sites" (**Figure 4B**) and "Secondary structure and surface accessibility" (**Figure 3C**). The detailed modification sites and types information are displayed in the "Potential cysteine modification sites" section (**Figure 4B**), and the sequence structure properties such as disordered information, secondary structure, and surface accessibility features are shown in the "Secondary structure and surface accessibility" (**Figure 4C**). When multiple protein sequences were submitted, pCysMod will predict and show the first one as a default. By clicking the



Start a new prediction project

Here is the web server of pCysMod, you can predict cysteine modifications by inputing one or more peptides.

Please input the peptide(s) in FASTA format:

Paste your target sequence(s) here

Modification(s):

□ All ☑ S-palmitoylation □ S-nitrosylation □ S-sulfhydration □ S-sulfenylation □ S-sulfinylation

Threshold:

● High ○ Medium ○ Low

Potential cysteine modification site(s)

ID	Position	Modification	FPR	Peptide
KLF6_HUMAN	7	S-palmitoylation	0.53%	*MDVLPM C SIFQELQ
KLF6_HUMAN	36	S-nitrosylation	7.06%	EEYWQQT C LELERYL
KLF6_HUMAN	48	S-nitrosylation	0.01%	RYLQSEP C YVSASEI
KLF6_HUMAN	48	S-sulfinylation	7.98%	RYLQSEP C YVSASEI
KLF6_HUMAN	164	S-sulfinylation	2.09%	EPSQLWG C VPGELPS

Secondary structure and surface accessibility

FIGURE 4 | The web server of pCysMod. **(A)** The prediction page. **(B)** Potential cysteine modification sites. **(C)** Secondary structure and surface accessibility.

selection box, users can choose which protein to display, and this will take 20 s in average. Besides, the proteins and peptides used in this study were uploaded in the web server and users can download the relevant data in the "Help" section. Overall, pCysMod was the first comprehensive cysteine modification prediction web server, which will undoubtedly greatly promote the study of protein cysteine modification and contribute to clarifying the biological regulation mechanisms of cysteine modification within and among the cells.

DISCUSSION

Protein cysteine modifications lead to a series of biochemical reactions, regulate various physiological and pathological processes, such as autophagy (Carroll et al., 2018), protein stabilization (Kröncke and Klotz, 2009), redox homeostasis (Fra et al., 2017), and cell signaling (Hourihan et al., 2016), demonstrating a close relationship with many human diseases including cancers, diabetes, and so on. Although many efforts have been made in this field, the experimental identification of cysteine modification proteins is tedious and laborious and the underlying molecular mechanisms are still unclear. In this regard, to dissect the molecular mechanisms and regulatory roles of cysteine modification, it is urgently needed to precisely parse the potential cysteine modification sites and types.

Through carefully curated previous reported literatures, predictors, and databases, we generated a benchmark dataset that consists of five types of cysteine modification, including *S*-nitrosylation, *S*-palmitoylation, *S*-sulfenylation, *S*-sulfhydration, and *S*-sulfinylation. The cysteine modification sites prefer to enrich in ordered regions. Consistent with previous reports, *S*-nitrosylation, *S*-sulfenylation, *S*-sulfhydration, and *S*-sulfinylation play crucial roles in oxidative post-translational modifications (Chung et al., 2013). Besides, the thioesterification reaction can cause *S*-palmitoylation by reversibly adding one or multiple palmitoyl moieties to cysteine residues (Roth et al., 2006), and *S*-palmitoylation also mediates a series of biochemical reactions, such as metabolism (Shen et al., 2017) and autophagy (Kim et al., 2019).

Then, we generated the pCysMod to predict multiple types of cysteine modification. Four-, six-, eight-, and tenfold cross-validations declared the robustness and high accuracy of pCysMod. Tenfold cross-validation comparison indicated a considerable predictive power for general cysteine modification prediction. We further generated the first comprehensive cysteine modification prediction web server pCysMod to provide an efficient and convenient way to facilitate basic research.

Although pCysMod has performed excellently in predicting cysteine modification, the limitations still exit. Currently, the cysteine modification data are still limited. We will keep collecting more modification types for future plans to generate a more comprehensive cysteine modification predictor. Furthermore, more deep learning methods could be taken into consideration, such as graph convolutional neural network (GCN), capsule network, and attention mechanisms, which may be an important and meaningful approach to help improving the current performance.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

JW, XG, and HC designed and supervised the experiments. SL, KY, and GW performed the experiments and data analysis, and developed the predictor. QZ and PW contributed to data analysis and predictor development. SL, KY, and GW wrote and revised the manuscript with contributions of all authors. All authors reviewed the manuscript.

FUNDING

This work was supported by grants from Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2017ZT07S096 to Z-XL), Pearl River S&T Nova Program of Guangzhou (201906010088), the Natural Science Foundation of China (91953123 to Z-XL, 31601067 to HC, and 21672201 and 92053104 to XG), Key Program for Department of Science and Technology of Qinghai Province (2017-ZJ-Y13), and Tip-Top Scientific and Technical Innovative Youth Talents of Guangdong Special Support Program (2019TQ05Y351).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.617366/full#supplementary-material

REFERENCES

Akter, S., Fu, L., Jung, Y., Conte, M. L., Lawson, J. R., Lowther, W. T., et al. (2018). Chemical proteomics reveals new targets of cysteine sulfinic acid reductase. *Nat. Chem. Biol.* 14, 995–1004. doi: 10.1038/s41589-018-0116-2

Brandes, N., Schmitt, S., and Jakob, U. (2009). Thiol-based redox switches in eukaryotic proteins. *Antioxid. Redox Signal.* 11, 997–1014. doi: 10.1089/ars.2008.2285

Carroll, B., Otten, E. G., Manni, D., Stefanatos, R., Menzies, F. M., Smith, G. R., et al. (2018). Oxidation of SQSTM1/p62 mediates the link between redox state and protein homeostasis. *Nat. Commun.* 9:256. doi: 10.1038/s41467-017-02746-z

Chung, H. S., Wang, S.-B., Venkatraman, V., Murray, C. I., and Van Eyk, J. E. (2013). Cysteine oxidative posttranslational modifications: emerging regulation in the cardiovascular system. *Circulation Res.* 112, 382–392. doi: 10.1161/CIRCRESAHA.112.268680

Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21, 3433–3434. doi: 10.1093/bioinformatics/bti541

Fra, A., Yoboue, E. D., and Sitia, R. (2017). Cysteines as redox molecular switches and targets of disease. *Front. Mol. Neurosci.* 10:167. doi: 10.3389/fnmol.2017.00167

Fu, L., Liu, K., He, J., Tian, C., Yu, X., and Yang, J. (2019). Direct proteomic mapping of cysteine persulfidation. *Antioxid. Redox Signal.* 3, 1061–1076. doi: 10.1089/ars.2019.7777

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Hasan, M. M., Manavalan, B., Khatun, M. S., and Kurata, H. (2019). Prediction of S-nitrosylation sites by integrating support vector machines and random forest. *Molecular Omics* 15, 451–458. doi: 10.1039/C9MO00098D

Hassanzadeh, H. R., and Wang, M. D. (2016). "DeeperBind: enhancing prediction of sequence specificities of DNA binding proteins," in *Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 178–183.

Hourihan, J. M., Moronetti Mazzeo, L. E., Fernández-Cárdenas, L. P., and Blackwell, T. K. (2016). Cysteine sulfenylation directs IRE-1 to activate the SKN-1/Nrf2 antioxidant response. *Mol. Cell* 63, 553–566. doi: 10.1016/j.molcel.2016.07.019

Huang, K.-Y., Lee, T.-Y., Kao, H.-J., Ma, C.-T., Lee, C.-C., Lin, T.-H., et al. (2018). dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications. *Nucleic Acids Res.* 47, D298–D308. doi: 10.1093/nar/gky1074

Jia, C., Lin, X., and Wang, Z. (2014). Prediction of protein S-nitrosylation sites based on adapted normal distribution bi-profile Bayes and Chou's pseudo amino acid composition. *Int. J. Mol. Sci.* 15, 10410–10423. doi: 10.3390/ijms150610410

Jia, J., Arif, A., Terenzi, F., Willard, B., Plow, E. F., Hazen, S. L., et al. (2014). Target-selective protein S-nitrosylation by sequence motif recognition. *Cell* 159, 623–634. doi: 10.1016/j.cell.2014.09.032

Ju, Z., and Wang, S. Y. (2018). Prediction of S-sulfenylation sites using mRMR feature selection and fuzzy support vector machine algorithm. *J. Theor. Biol.* 457, 6–13. doi: 10.1016/j.jtbi.2018.08.022

Kim, S. W., Kim, D. H., Park, K. S., Kim, M. K., Park, Y. M., Muallem, S., et al. (2019). Palmitoylation controls trafficking of the intracellular Ca(2+) channel MCOLN3/TRPML3 to regulate autophagy. *Autophagy* 15, 327–340. doi: 10.1080/15548627.2018.1518671

Kröncke, K. D., and Klotz, L. O. (2009). Zinc fingers as biologic redox switches? *Antioxid. Redox Signal.* 11, 1015–1027. doi: 10.1089/ars.2008.2269

Kumsta, C., Thamsen, M., and Jakob, U. (2011). Effects of oxidative stress on behavior, physiology, and the redox thiol proteome of Caenorhabditis elegans. *Antioxid. Redox Signal.* 14, 1023–1037. doi: 10.1089/ars.2010.3203

Lee, T. Y., Chen, S. A., Hung, H. Y., and Ou, Y. Y. (2011a). Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. *PLoS One* 6:e17331. doi: 10.1371/journal.pone.0017331

Lee, T. Y., Chen, Y. J., Lu, T. C., Huang, H. D., and Chen, Y. J. (2011b). SNOSite: exploiting maximal dependence decomposition to identify cysteine S-nitrosylation with substrate site specificity. *PLoS One* 6:e21849. doi: 10.1371/journal.pone.0021849

Li, S., Yu, K., Wang, D., Zhang, Q., Liu, Z. X., Zhao, L., et al. (2020). Deep learning based prediction of species-specific protein S-glutathionylation sites. *Biochim. Biophys. Acta Proteins Proteom.* 1868:140422. doi: 10.1016/j.bbapap.2020.140422

Lim, J. C., Choi, H. I., Park, Y. S., Nam, H. W., Woo, H. A., Kwon, K. S., et al. (2008). Irreversible oxidation of the active-site cysteine of peroxiredoxin to cysteine sulfonic acid for enhanced molecular chaperone activity. *J. Biol. Chem.* 283, 28873–28880. doi: 10.1074/jbc.M804087200

Liu, Z., Yuan, F., Ren, J., Cao, J., Zhou, Y., Yang, Q., et al. (2012). GPS-ARM: computational analysis of the APC/C recognition motif by predicting D-boxes and KEN-boxes. *PLoS One* 7:e34370. doi: 10.1371/journal.pone.0034370

Liu, Z. X., Yu, K., Dong, J., Zhao, L., Liu, Z., Zhang, Q., et al. (2019). Precise prediction of calpain cleavage sites and their aberrance caused by mutations in cancer. *Front. Genet.* 10:715. doi: 10.3389/fgene.2019.00715

Mann, M., and Jensen, O. N. (2003). Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* 21, 255–261. doi: 10.1038/nbt0303-255

Marino, S. M., and Gladyshev, V. N. (2011). Redox biology: computational approaches to the investigation of functional cysteine residues. *Antioxid. Redox Signal.* 15, 135–146. doi: 10.1089/ars.2010.3561

McGuffin, L. J., Bryson, K., and Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics* 16, 404–405. doi: 10.1093/bioinformatics/16.4.404

Mertins, P., Qiao, J. W., Patel, J., Udeshi, N. D., Clauser, K. R., Mani, D. R., et al. (2013). Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nat. Methods* 10, 634–637. doi: 10.1038/nmeth.2518

Mishanina, T. V., Libiad, M., and Banerjee, R. (2015). Biogenesis of reactive sulfur species for signaling by hydrogen sulfide oxidation pathways. *Nat. Chem. Biol.* 11, 457–464. doi: 10.1038/nchembio.1834

Ning, W., Jiang, P., Guo, Y., Wang, C., Tan, X., Zhang, W., et al. (2020). GPS-Palm: a deep learning-based graphic presentation system for the prediction of S-palmitoylation sites in proteins. *Brief Bioinform.* bbaa038. doi: 10.1093/bib/bbaa038

Oteiza, P. I. (2012). Zinc and the modulation of redox homeostasis. *Free Radic. Biol. Med.* 53, 1748–1759. doi: 10.1016/j.freeradbiomed.2012.08.568

Petersen, B., Petersen, T. N., Andersen, P., Nielsen, M., and Lundegaard, C. (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.* 9:51. doi: 10.1186/1472-6807-9-51

Radivojac, P., Vacic, V., Haynes, C., Cocklin, R. R., Mohan, A., Heyen, J. W., et al. (2010). Identification, analysis, and prediction of protein ubiquitination sites. *Proteins* 78, 365–380. doi: 10.1002/prot.22555

Roth, A. F., Wan, J., Bailey, A. O., Sun, B., Kuchar, J. A., Green, W. N., et al. (2006). Global analysis of protein palmitoylation in yeast. *Cell* 125, 1003–1013. doui: 10.1016/j.cell.2006.03.042

Rouault, T. A. (2015). Mammalian iron-sulphur proteins: novel insights into biogenesis and function. *Nat. Rev. Mol. Cell Biol.* 16, 45–55. doi: 10.1038/nrm3909

Shen, L. F., Chen, Y. J., Liu, K. M., Haddad, A. N. S., Song, I. W., Roan, H. Y., et al. (2017). Role of S-Palmitoylation by ZDHHC13 in Mitochondrial function and Metabolism in Liver. *Sci. Rep.* 7:2182. doi: 10.1038/s41598-017-02159-4

Song, J., Tan, H., Shen, H., Mahmood, K., Boyd, S. E., Webb, G. I., et al. (2010). Cascleave: towards more accurate prediction of caspase substrate cleavage sites. *Bioinformatics* 26, 752–760. doi: 10.1093/bioinformatics/btq043

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics* 16, 16–23. doi: 10.1093/bioinformatics/16.1.16

Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res.* 10, 2997–3011. doi: 10.1093/nar/10.9.2997

Strzyz, P. (2016). Post-translational modifications: extension of the tubulin code. *Nat. Rev. Mol. Cell Biol.* 17:609. doi: 10.1038/nrm.2016.117

UniProt Consortium [UC] (2015). UniProt: a hub for protein information. *Nucleic Acids Res.* 43, D204–D212. doi: 10.1093/nar/gku989

Vacic, V., Iakoucheva, L. M., and Radivojac, P. (2006). Two sample logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* 22, 1536–1537. doi: 10.1093/bioinformatics/btl151

Weng, S.-L., Kao, H.-J., Huang, C.-H., and Lee, T.-Y. (2017). MDD-Palm: identification of protein S-palmitoylation sites with substrate motifs based on maximal dependence decomposition. *PLoS One* 12:e0179529. doi: 10.1371/journal.pone.0179529

Xie, Y., Luo, X., Li, Y., Chen, L., Ma, W., Huang, J., et al. (2018). DeepNitro: prediction of protein nitration and nitrosylation sites by deep learning. *Genomics Proteomics Bioinform.* 16, 294–306. doi: 10.1016/j.gpb.2018.04.007

Xu, Y., Ding, J., and Wu, L. Y. (2016). iSulf-Cys: pzrediction of S-sulfenylation sites in proteins with physicochemical properties of amino acids. *PLoS One* 11:e0154237. doi: 10.1371/journal.pone.0154237

Xu, Y., Wang, Y., Luo, J., Zhao, W., and Zhou, X. (2017). Deep learning of the splicing (epi) genetic code reveals a novel candidate mechanism linking histone modifications to ESC fate decision. *Nucleic Acids Res.* 45, 12100–12112.

Xue, Y., Liu, Z., Gao, X., Jin, C., Wen, L., Yao, X., et al. (2010). GPS-SNO: computational prediction of protein S-nitrosylation sites with a modified GPS algorithm. *PLoS One* 5:e11290. doi: 10.1371/journal.pone.0011290

Yang, J., Gupta, V., Tallman, K. A., Porter, N. A., Carroll, K. S., and Liebler, D. C. (2015). Global, in situ, site-specific analysis of protein S-sulfenylation. *Nat. Protoc.* 10, 1022–1037. doi: 10.1038/nprot.2015.062

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics J Integrative Biol.* 16, 284–287. doi: 10.1089/omi.2011.0118

Yu, K., Zhang, Q., Liu, Z., Du, Y., Gao, X., Zhao, Q., et al. (2020). Deep learning based prediction of reversible HAT/HDAC-specific lysine acetylation. *Brief Bioinform.* 21, 1798–1805. doi: 10.1093/bib/bbz107

Zhang, Z. Y., Yang, Y. H., Ding, H., Wang, D., Chen, W., and Lin, H. (2020). Design powerful predictor for mRNA subcellular location prediction in Homo sapiens. *Brief Bioinform.* 22, 526–535. doi: 10.1093/bib/bbz177

Zhao, X., Zhang, W., Xu, X., Ma, Z., and Yin, M. (2012). Prediction of protein phosphorylation sites by using the composition of k-spaced amino acid pairs. *PLoS One* 7:e46302. doi: 10.1371/journal.pone.0046302

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.