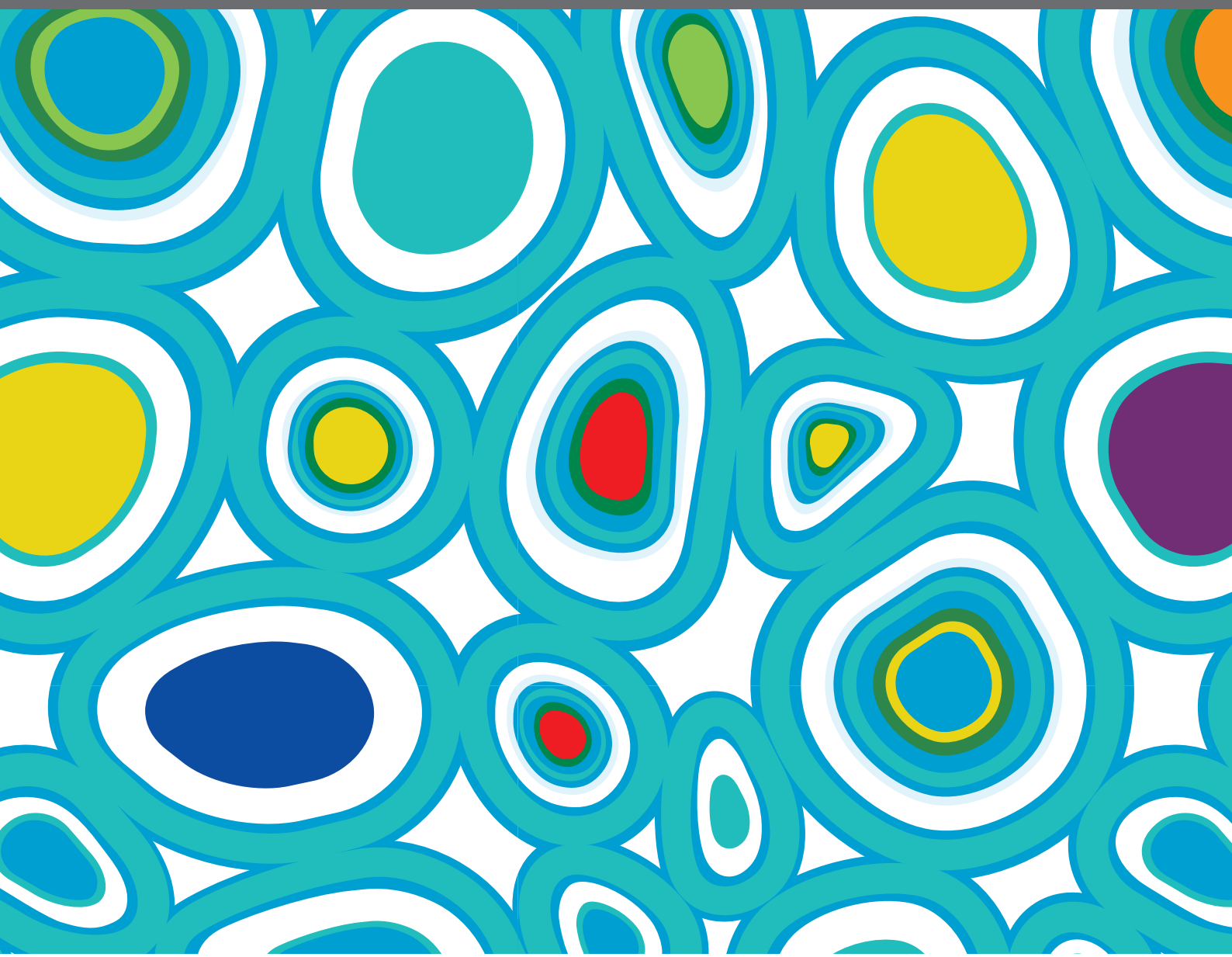
COMPUTATIONAL RESOURCES FOR UNDERSTANDING BIOMACROMOLECULAR COVALENT MODIFICATIONS

EDITED BY: Dong Xu, Hsien-Da Huang, Jiangning Song, Jian Ren and Yu Xue
PUBLISHED IN: Frontiers in Cell and Developmental Biology



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88971-319-6
DOI 10.3389/978-2-88971-319-6

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

COMPUTATIONAL RESOURCES FOR UNDERSTANDING BIOMACROMOLECULAR COVALENT MODIFICATIONS

Topic Editors:
Dong Xu, University of Missouri, United States
Hsien-Da Huang, The Chinese University of Hong Kong, China
Jiangning Song, Monash University, Australia
Jian Ren, Sun Yat-sen University, China
Yu Xue, Huazhong University of Science and Technology, China

Citation: Xu, D., Huang, H.-D., Song, J., Ren, J., Xue, Y., eds. (2021). Computational Resources for Understanding Biomacromolecular Covalent Modifications. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-319-6

Table of Contents

frontiers in Cell and Developmental Biology

EDITORIAL
published: 14 July 2021
doi: 10.3389/fcell.2021.728127

Editorial: Computational Resources for Understanding Biomacromolecular Covalent Modifications

Yu Xue[1]*, *Hsien-Da Huang*[2], *Jiangning Song*[3], *Jian Ren*[4] and *Dong Xu*[5]*

[1] Key Laboratory of Molecular Biophysics of Ministry of Education, Bioinformatics and Molecular Imaging Key Laboratory, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China, [2] School of Life and Health Sciences, Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen, China, [3] Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC, Australia, [4] School of Life Sciences, Sun Yat-sen University, Guangzhou, China, [5] Department of Electrical Engineer and Computer Science and Christopher S. Bond Life Science Center, University of Missouri, Columbia, MO, United States

Keywords: biomacromolecular covalent modification, post-translational modification, DNA modification, RNA modification, machine learning, deep learning, data integration

Editorial on the Research Topic

Computational Resources for Understanding Biomacromolecular Covalent Modifications

Biomacromolecular covalent modifications (BCMs) include protein post-translational modifications (PTMs) and nucleic acid modifications. To date, 670 types of PTMs have been identified, and the most extensively studied PTMs are phosphorylation, ubiquitination, and acetylation. They are involved in regulating almost all biological processes, such as cell cycle, autophagy, and metabolism. More than 150 types of RNA modifications and tens of DNA modifications have been discovered, such as N^6-methyladenosine (m^6A) in messenger RNAs and 5-methylcytosine (5mC) in DNAs, and they play crucial roles in controlling gene expression. There is increasing evidence showing that PTMs are related to many diseases such as cancer and neurological disorders. RNA modification pathways are also found to be dysregulated in human cancer, and as such, epigenomic DNA modifications may shed some light on why certain diseases and tumors develop with aging.

The modification processes of diverse BMCs share some common properties. The deposition of chemical modifications (or marks) onto biomacromolecules is catalyzed by specific enzymes named "writers." The enzymes that remove the modifications are called "erasers." After recognizing the BCM sites, regulator proteins that produce a cellular response are "readers." Identification of these BCM substrates and sites, as well as their "writers," "erasers," and "readers" can provide us a better understanding of how cellular activities are dynamically regulated. Computational algorithms, pipelines, tools, and databases play an increasingly important role in supporting biologists to explore BCM regulation, especially related regulatory mechanisms of protein PTMs and DNA/RNA modifications. We have witnessed substantial progress in computational development for BCM in both breadth and depth in recent years. The breadth covers a dramatically increasing number of BCM types, while the depth of new methods benefits extensively from the recent advancement in machine learning, especially deep learning. This Research Topic highlights these active developments with 8 predictive tools and 1 online database for BCMs in a timely manner. As shown in **Figure 1**, they represent a broad range of BCM types using various computational methods.

OPEN ACCESS

Edited and reviewed by:
Cecilia Giulivi,
University of California, Davis, United States

***Correspondence:**
Yu Xue
xudong@missouri.edu
Dong Xu
xudong@missouri.edu

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

Received: *20 June 2021*
Accepted: *22 June 2021*
Published: *14 July 2021*

Citation:
Xue Y, Huang H-D, Song J, Ren J and Xu D (2021) Editorial: Computational Resources for Understanding Biomacromolecular Covalent Modifications. Front. Cell Dev. Biol. 9:728127. doi: 10.3389/fcell.2021.728127

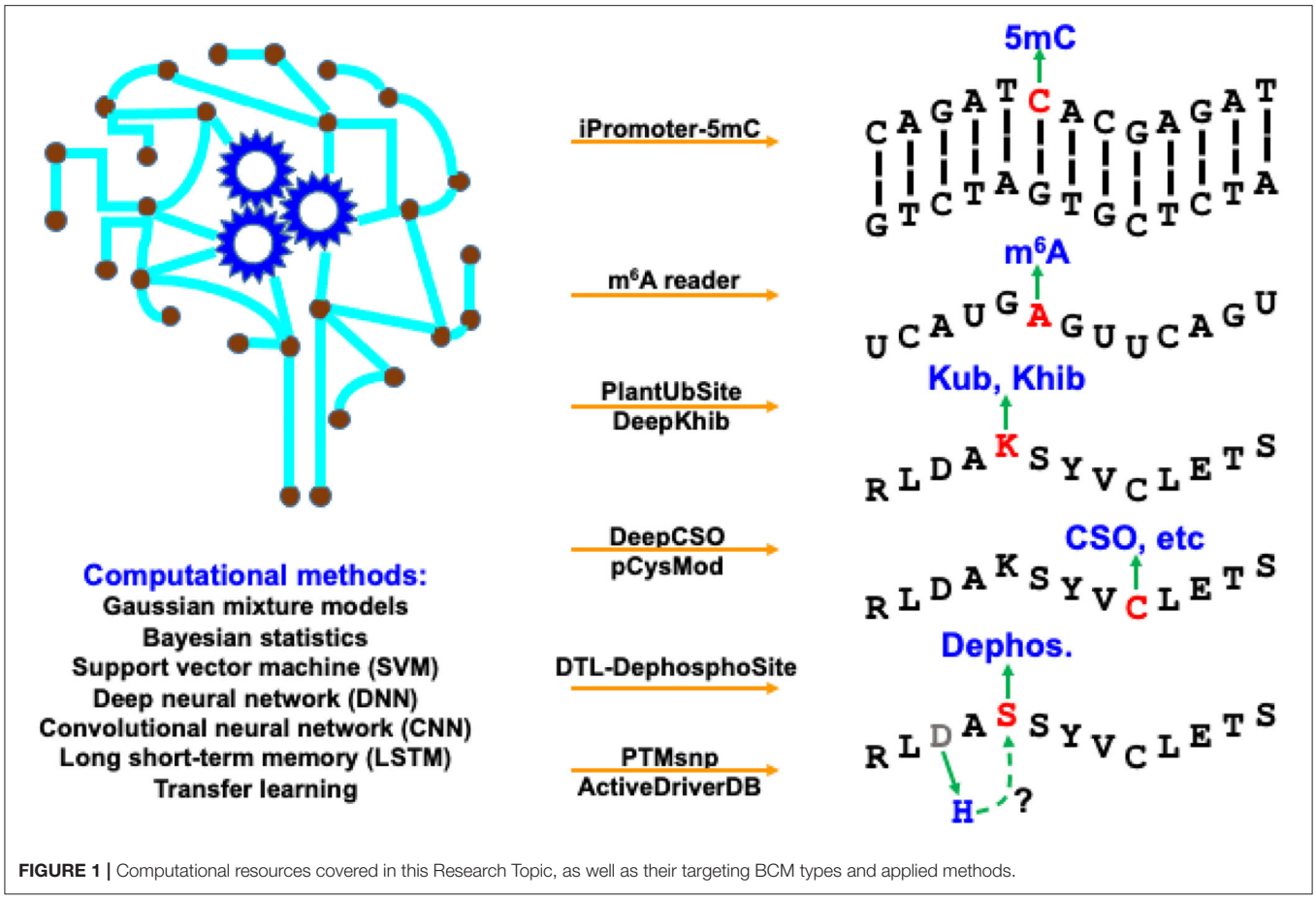


FIGURE 1 | Computational resources covered in this Research Topic, as well as their targeting BCM types and applied methods.

In this Research Topic, two studies developed machine-learning frameworks to predict DNA/RNA modifications. iPromoter-5mC (https://github.com/zlwuxi/iPromoter-5mC) provides a deep neural network (DNN) framework to predict DNA 5-methylcytosine (5 mC) sites (Zhang et al.). The m6A reader (http://m6areader.rnamd.com) adopts a Support Vector Machine (SVM) model to predict reader-specific mRNA N^6-methyladenosine (m^6A) sites (Zhen et al.).

This Research Topic also presents five deep-learning tools for PTM predictions. Two studies target protein modification on the lysine (K) residue using convolutional neural network (CNN), including PlantUbSite (https://github.com/wang-hong-fei/DL-plantubsites-prediction) for predicting plant ubiquitylation sites (Wang et al.) and DeepKhib (http://www.bioinfogo.org/DeepKhib/) for predicting lysine 2-hydroxyisobutyrylation (Khib) sites (Zhang et al.). Two other studies target protein modification on the cysteine (C) residue: DeepCSO (http://www.bioinfogo.org/DeepCSO/) provides a long short-term memory (LSTM) tool to predict cysteine S-sulphenylation (CSO) sites in proteins (Lyu et al.), and pCysMod (http://pcysmod.omicsbio.info/) is a deep neural network (DNN) tool to predict multiple types of protein cysteine modification sites, including S-nitrosylation, S-palmitoylation, S-sulfenylation, S-sulfhydration, and S-sulfinylation (Li et al.). Furthermore, DTL-DephosphoSite (https://github.com/dukkakc/DTLDephos) provides an LSTM framework to predict dephosphorylation sites in proteins (Chaudhari et al.). All these studies demonstrate the excellent predictive power of deep learning for PTM predictions.

PTMsnp and ActiveDriverDB focus on functional annotations of the mutation effects on PTMs. PTMsnp (http://ptmsnp.renlab.org/) is an online service to predict driver mutations that potentially change PTM sites (Peng et al.). The authors use a Bayesian hierarchical model for the prediction, covering 411,574 sites from 33 types of PTMs and 1,776,848 somatic mutations. ActiveDriverDB (https://www.activedriverdb.org/) is an updated database of predicted PTM-specific impact of genetic variations based on Gaussian mixture models and Bayesian posterior probability estimation for proteins and their interaction networks (Krassowski et al.). An interesting estimate of the study indicates the widespread impact of PTM, i.e., 16–21% of pathogenic disease mutations, somatic mutations in cancer genomes and germline variants in the human population potentially affect PTMs and their downstream biological activities.

This Research Topic showcases state-of-the-art computational studies of BCMs. From these excellent papers, it is evident that the field is highly active and more research still is needed. We hope that readers can formulate some good ideas for future development from the papers or utilize the resources for their

biological investigations. Finally, we, as the guest editors of this Research Topic, would like to thank all the authors for their valuable contributions.

AUTHOR CONTRIBUTIONS

YX and DX wrote the first draft. HH, JS, and JR provided critical comments and editorial suggestions for revisions. All the authors agreed on the submitted version.

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.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 28 July 2020
doi: 10.3389/fcell.2020.00614

iPromoter-5mC: A Novel Fusion Decision Predictor for the Identification of 5-Methylcytosine Sites in Genome-Wide DNA Promoters

Lei Zhang, Xuan Xiao and Zhao-Chun Xu**

Computer Department, Jing-De-Zhen Ceramic Institute, Jingdezhen, China

OPEN ACCESS

Edited by:
Yu Xue,
Huazhong University of Science and Technology, China

Reviewed by:
Leyi Wei,
Shandong University, China
Wei Chen,
North China University of Science and Technology, China
Jianbo Pan,
Johns Hopkins Medicine, United States

***Correspondence:**
Xuan Xiao
jdzxiaoxuan@163.com
Zhao-Chun Xu
jdzxuzhaochun@163.com

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

Received: *08 May 2020*
Accepted: *22 June 2020*
Published: *28 July 2020*

Citation:
Zhang L, Xiao X and Xu Z-C (2020) iPromoter-5mC: A Novel Fusion Decision Predictor for the Identification of 5-Methylcytosine Sites in Genome-Wide DNA Promoters. Front. Cell Dev. Biol. 8:614. doi: 10.3389/fcell.2020.00614

The hypomethylation of the whole cancer genome and the hypermethylation of the promoter of specific tumor suppressor genes are the important reasons for the rapid proliferation of cancer cells. Therefore, obtaining the distribution of 5-methylcytosine (5mC) in promoters is a key step to further understand the relationship between promoter methylation and mRNA gene expression regulation. Large-scale detection of DNA 5mC through wet experiments is still time-consuming and laborious. Therefore, it is urgent to design a method for identifying the 5mC site of genome-wide DNA promoters. Based on promoter methylation data of the small cell lung cancer (SCLC) from the database named cancer cell line Encyclopedia (CCLE), we built a fusion decision predictor called iPromoter-5mC for identifying methylation modification sites in promoters using deep neural network (DNN). One-Hot Encoding (One-hot) was used to encode the promoter samples for the classification. The method achieves average AUC of 0.957 on the independent testing dataset, indicating that our predictor is robust and reliable. A user-friendly web-server called iPromoter-5mC could be freely accessible at http://www.jci-bioinfo.cn/iPromoter-5mC, which will provide simple and effective means for users to study promoter 5mC modification. The source code of the proposed methods is freely available for academic research at https://github.com/zlwuxi/iPromoter-5mC.

Keywords: promoter, 5-methylcytosine, fusion decision, predictor, web-server, deep neural network

INTRODUCTION

DNA methylation dominates any cell processes, and plays a particularly important role in regulating expression of gene (Bird, 2007; Deichmann, 2016; Nicoglou and Merlin, 2017). DNA methylation at promoters and enhancers has been associated with cell differentiation, developmental processes, cancer development, and regulation of the immune system (Muller et al., 2019). At present, N6-methyladenine (6mA), N4-methylcytosine (4mC) and 5-methylcytosine (5mC) are the three most well-studied types of DNA methylation (Wei et al., 2019). 5mC is a covalent addition between the methyl group and the 5-carbon of the cytosine ring. In somatic cells, 5mC occurs almost exclusively in the context of paired symmetrical methylation of a CpG site.

Recent study (Michalak et al., 2019) suggests that aberrant levels of 5mC at CpG islands in promoter regions is associated with inactivation of various tumor suppressor genes (TSGs). In

young normal cells, 5mC is low in the promoter regions but high in the genic and intergenic regions. However, in aging and in cancer, a limited number of genomic loci acquire 5mC, especially at the CpG islands in promoter regions of tumor suppressor and Polycomb-repressed gene, resulting in gene silencing and loss of function. In normal tissue, heterochromatin contains repeating elements and is highly methylated. The aberrant promoter methylation can lead to cancer initiation and progression, which has been confirmed in CpG island methylator phenotype (CIMP) cancers (Gessler, 1999; Kang et al., 2002; Mansour, 2014). Thus, promoter methylation can be used as a potential biomarker for cancer diagnosis and for helping determine prognosis, indicating that identification of 5mC modification in promoter regions by analyzing CpG islands in cell systems of a specific cancer could provide a reference for cancer early diagnosis and precise treatment.

Among cancers worldwide, both the incidence and death rate of lung cancer are in the first place. Small cell lung cancer (SCLC) poses approximately 15% of newly increasing clinical cases with lung cancer each year (Siegel et al., 2018). Its pattern is significantly different from other lung cancer, and is closely related to the high expression of E2F target and EZH2 gene of histone methyltransferase. Furthermore, SCLC is famous for its dense cluster of high-level methylation in CpG islands of discrete promoter. Therefore, in this study, we are concentrating on improving the ability to access the methylation status of promoters for a large number of genes or the entire genome in SCLC.

One of the most usual methods for identifying DNA methylation is distinguishing the cytosine-5 methylation within the CpG dinucleotides (Bianchi and Zangi, 2015; Muller et al., 2019). The popular sequencing technology for identifying 5mC sites includes Methylated DNA immunoprecipitation sequencing (MeDIP-seq), Methyl-Binding Domain sequencing (MBD-seq) and DNA methylome profiling at single-base resolution through bisulfite sequencing (MB-seq) (Down et al., 2008). However, these wet-lab methods are expensive and time-consuming. Therefore, it is urgent to develop a number of methods or tools for the accurate detection of DNA 5mC modification sites.

Over the past decade, computational methods have been proposed to identify 5mC modification sites. Bhasin et al. (2005) developed a SVM-based model called "Methylator," for the prediction of 5mC modification sites using the methylated and non-methylated CpG dinucleotide sequences from various sources ranging from plants to humans in MethDB database (Amoreira, 2003). Fang et al. (2006) developed a SVM-based classifier called "MethCGI" using nucleotide sequence contents and transcription factor binding sites as features. Compared with the previous two, the predictor "iDNA-Methyl" (Liu et al., 2015) constructed by using the trinucleotide composition and pseudo amino acid components achieved higher success rates. Recently, a novel computational tool called NanoMod (Liu et al., 2018) was designed to improve the performance of detecting candidate positions with DNA modifications. Based on deep neural networks, a computational approach called DeepCpG (Angermueller et al., 2017) was developed to predict methylation states in single cells.

Though the research about the recognition of DNA 5mC modification sites have had a significant advance in recent years, but still exist shortness. Compared to increasing massive high-throughput data, previous studies are of small sample size. Furthermore, among above-mentioned methods, there are three webservers developed by the researchers: Methylator, MethCGI, and iDNA-Methyl, however, only the latter is available but slow, causing much inconvenience to scholars. Most importantly, there is still no computation tool to identify DNA 5mC modification sites in promoters to detect the biomarkers of a specific cancer. Therefore, in the current study, we are devoted to solve these problems and to develop a tool or software for quickly and precisely identifying DNA 5mC modification sites in promoters.

MATERIALS AND METHODS

Benchmark Datasets

The construction of the high-quality data sets is an essential step in the process of establishing the classification model. In the current study, all the sequence samples were collected from the database named cancer cell line Encyclopedia (CCLE) (Barretina et al., 2012; Li et al., 2019), which provided the location information of gene promoter regions and 5mC modification sites experimented by reduced representation bisulfite sequencing (RRBS) (Ghandi et al., 2019) in cell lines of various cancers. Due to the high incidence rate and mortality rate of lung cancer, here we focused on the small cell lung cancer (SCLC) to reveal the distribution of 5mC modification in promoters.

In accordance with the forward/reverse (±) chain and 5mC modification sites' positions in promoters, we collected the sequence samples from the most recent human assembly GRCh37/hg19 on UCSC Genome Browser. It is noteworthy that the sample sequence containing 5mC modification site described as the base G (guanine) in the reverse chain should convert to the reverse complementary sequence, compatible with the principle that the DNA 5mC methylation tends to occur at cytosine (C). Generally, we considered the base C with the methylation level greater than zero as the true 5mC modification site, otherwise, as the false 5mC modification site.

In order to more succinctly describe the promoter sequence fragment potentially containing 5mC modification site, the sample sequence can be expressed as

$$E_\delta(\mathbb{C}) = E_{-\delta}E_{-(\delta-1)}\cdots E_{-2}E_{-1}\mathbb{C}E_{+1}E_{+2}\cdots E_{+(\delta-1)}E_{+\delta} \quad (1)$$

where the double letter $\mathbb{C}$ represents the cytosine; the subscript δ is an integer, indicating the location of the base in the sequence; $E_{-\delta}$ is the δ -th base upstream from the center and $E_{+\delta}$ is the δ -th base downstream from the center.

The sample sequence thus obtained can be divided into two categories:

$$E_\delta(\mathbb{C}) \in \begin{cases} E_\delta^-(\mathbb{C}) \\ E_\delta^+(\mathbb{C}) \end{cases} \quad (2)$$

where $E_\delta^-(\mathbb{C})$ represents a false 5mC modification segment with $\mathbb{C}$ at its center, $E_\delta^+(\mathbb{C})$ denotes a true 5mC modification segment

TABLE 1 | Distribution of experimental data sets.

Attribute	Total	Training data	Testing data
Positive	69,750	55,800	13,950
Negative	823,576	658,861	164,715

with $\mathbb{C}$ at its center, and the symbol $\in$ denotes "a member of" in the set theory.

Therefore, the benchmark dataset can be formulated by

$$S_\delta = S_\delta^- \cup S_\delta^+ \tag{3}$$

where S_δ^- denotes the negative subset containing the false 5mC modification site samples; S_δ^+, the positive subset containing the true 5mC modification site samples; and symbol $\cup$ represents union in the set theory.

Unbalanced data between the true 5mC modification site samples and the false 5mC modification site samples could more objectively reflect the distribution of 5mC modification in promoters. Therefore, the proportion of positive samples and negative samples was set to about 1:11 in this study. In order to reduce the adverse effects of redundancy and homologous bias, sequences with more than 80% sequence similarity were removed using CD-HIT software.

Finally, we obtained the benchmark dataset S_δ composed of 893,326 methylation sample sequences in promoter regions, of which 69,750 sample sequences belong to the positive sample dataset S_δ^+ and 823,576 sample sequences belong to the negative sample dataset S_δ^-. To investigate the stability and robustness of the prediction model, we randomly selected 80% data in S_δ^+ and S_δ^-, respectively, as training set S_1 for constructing and training the prediction model, and remained 20% as independent testing dataset S_2 to test the constructed model (**Table 1**). These datasets can be downloaded from the website http://www.jci-bioinfo.cn/iPromoter-5mC/download.

Extract Features From DNA Sequences

Feature extraction, fusion and selection are the important steps in machine learning process. Many feature extraction methods for protein, RNA and DNA sequences, including PseAAC, PseKNC, PCPs, PCM, PS(k-mer)NP (Zou et al., 2016), have been proposed to overcome the prediction problem of modification sites. In the current study, we employed two effective feature extraction methods (one-hot and DPF) to extract feature directly form DNA sample sequences.

One-Hot Encoding Method (One-Hot)

One-hot is a simple but effective feature extraction method, especially for deep learning model. The nucleotides A, C, G and T are denoted as one of the four one-hot vectors [1,0,0,0], [0,1,0,0], [0,0,1,0], and [0,0,0,1] (**Figure 1**).

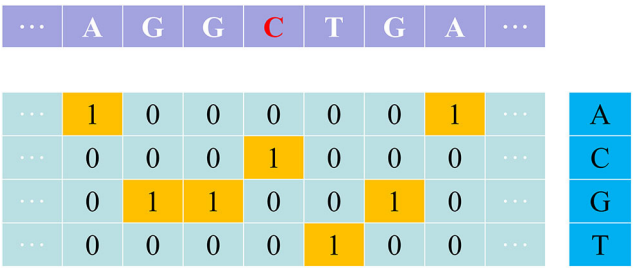


FIGURE 1 | An illustration showing the one-hot encoding method.

The Deoxynucleotide Property and Frequency (DPF)

Deoxynucleotides are the basic structural and functional units of DNA, and the sequence generated by deoxynucleotides determines biological diversity. Therefore, their chemical properties can influence the inherited characteristics of the DNA sequence to a certain extent. Similar to the encoding method of RNA sequences used in identifying 4mC sites, the deoxynucleotide property and frequency (DPF) (Xia et al., 2019; Xu et al., 2019) is an effective sequence encoding scheme for computationally identifying 5mC modification sites.

Each of the four deoxynucleotides has a different chemical property. Given the sample sequence Q represented by Equation (1), the k-th deoxynucleotide in Equation (1) can be converted into a three-dimensional vector, as shown in the Equation (4). Considering that purines have two rings between them and pyrimidines have only one ring, we added the feature of ring structure to feature extraction. Since there is an amino group between A and C, but A keto group between G and T, we added functional group features to feature extraction. In terms of the strength of the hydrogen bond between the base pair, the hydrogen bond between C and G is stronger than the hydrogen bond between A and T, because A is always paired with T by two hydrogen bonds, but C is bound to G by three hydrogen bonds. So we added hydrogen bond features to Q, as shown in the following expression.

$$Q_k = (x_k, y_k, z_k) \tag{4}$$

where x_k represents the "ring structure"; y_k, the "functional group"; z_k, the "hydrogen bond."

x_k, y_k and z_k can be formulated by Equation (5):

$$\begin{aligned} x_k &= \begin{cases} 1 \; if \; Q_k \in \{A \; , \; G\} \\ 0 \; if \; Q_k \in \{C \; , \; T\} \end{cases} \\ y_k &= \begin{cases} 1 \; if \; Q_k \in \{A \; , \; C\} \\ 0 \; if \; Q_k \in \{G \; , \; T\} \end{cases} \\ z_k &= \begin{cases} 1 \; if \; Q_k \in \{A \; , \; T\} \\ 0 \; if \; Q_k \in \{C \; , \; G\} \end{cases} \end{aligned} \tag{5}$$

In order to extract the sequence position information as much as possible (Chen et al., 2017), the cumulative frequency characteristics of deoxynucleotides were adopted:

$$\lambda_k = \frac{\sum_{j=1}^{k} \mathcal{F}(M_j)}{k} \left(1 \le k \le 2\delta + 1\right) \tag{6}$$

where k is the length of the sample sequence, λ_k is the density of the deoxynucleotide Q_k along the subsequence from position 1 to

position k in the sample sequence, and $\mathcal{F}(M_j)$ can be expressed as below.

$$\mathcal{F}(M_j) = \begin{cases} 1 \ if \ M_j = Q_k \\ 0 \ otherwise \end{cases} \quad (7)$$

Then we obtained a feature vector $\vec{v}$ to represent the k-th deoxynucleotide in the sample sequence, as shown in the following formula,

$$\vec{v} = (x_k, y_k, z_k, \lambda_k) \quad (8)$$

The chemical properties of deoxynucleotides reveal the intrinsic relationship between the four different deoxy nucleotides in the sequence and represent the sequence information as discrete vectors by means of 0–1 coding. Therefore, by this method, we represented the sequence with $4 \times L$-D (dimensional) feature vector $\mathcal{W}$ to represent the sample sequence formulated by Equation (1),

$$\mathcal{W} = \left[x_1 y_1 z_1 \lambda_1 \cdots x_{2\delta+1} \ y_{2\delta+1} \ z_{2\delta+1} \ \lambda_{2\delta+1}\right]^T \quad (9)$$

where the symbol T is the transpose operator.

Feature Fusion

Feature fusion usually joins several kinds of different feature vectors into an integrated one, which could express the local and global sequence order information of the given sample sequence. Therefore, in this study, we not only employed one-hot and DPF methods, but also took into account their combination. According with this method, we represented the sequence with $2 \times 4 \times L$-D (dimensional) feature vector.

Framework of the Integrated Predictor

For imbalance problems existing in positive samples and negative samples, the down-sampling method was adopted in the current study. We randomly divided the negative samples from the training dataset S_1 into 11 groups of equal size, one of which can form the balance training subset by combining with the positive samples in the same amount. And then, we could obtain 11 sub-models. After converting into a numeric vector by one-hot, DPF or their combination, a query sequence with the base C in its center, can be input into 11 sub-models for prediction. The 11 prediction results thus obtained can be used to generate the final decision whether the base C is methylated or not by some judging methods, just like a simple majority vote or weighted voting method (**Figure 2**). The integrated predictor obtained by above-mentioned method was named as iPromoter-5mC, which can be used to identify the 5mC modification sites in promoter sequences.

In this study, a simple deep neural network (DNN) framework (Islam et al., 2018) was employed to consturcted the prediction model. The generated feature matrix was fed into the fully connected neural network for training. The fully connected layer of DNN contained 64, 128, 256, 128, 64 neurons in turn, and the activation function was ReLU (Zhuang et al., 2019). For binary problem, the last layer contained two neurons, and sigmoid

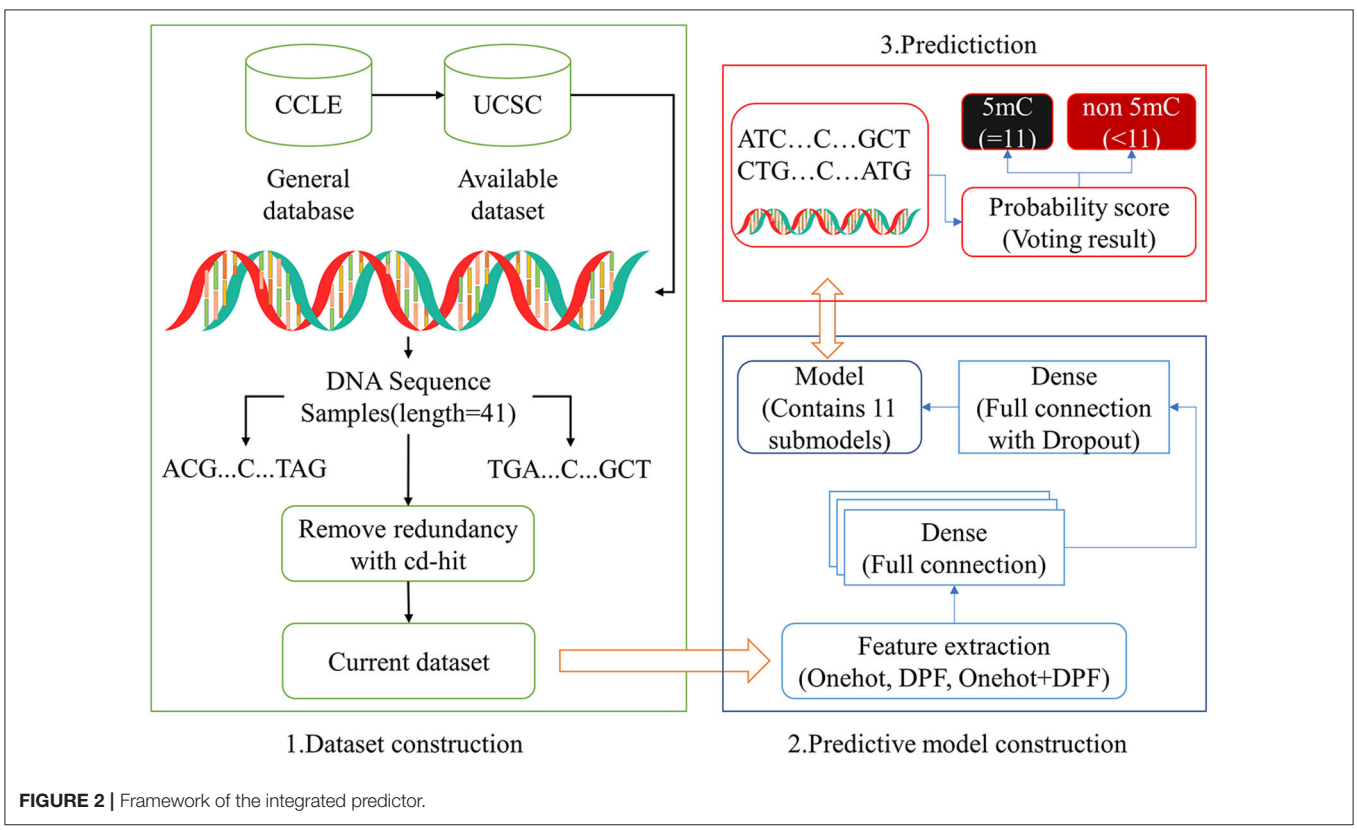


FIGURE 2 | Framework of the integrated predictor.

was selected as the activation function. To prevent overfitting and improve model generalization, a dropout layer was added before the last full connection layer, with a value of 0.3. Five-fold cross validation was conducted to validate the reliability of each sub-model.

Evaluation Metrics

K-fold cross-validation method could effectively utilize limited data, and the evaluation results are as close as possible to the model's performance on the test set. Therefore, we used this method to evaluate the model's performance (Wei et al., 2018; Chen et al., 2019a,b; Dao et al., 2019). For single label system, there are several common evaluation indexes to measure the predictive performance of the predictor, including Sensitivity (Sn), Accuracy (Acc), Specificity (Sp) and Matthew's correlation coefficient (MCC), which can be defined as following,

$$\begin{cases} Sn = 1 - \frac{N_-^+}{N_+}, 0 \le Sn \le 1 \\ Sp = 1 - \frac{N_+^-}{N}, 0 \le Sp \le 1 \\ Acc = 1 - \frac{N_-^+ + N_+^-}{N^+ + N}, 0 \le Acc \le 1 \\ MCC = \frac{1 - \left(\frac{N_-^+ + N_+^-}{N^+ + N}\right)}{\sqrt{\left(1 + \frac{N_+^- N^+}{N^+}\right)\left(1 + \frac{N_-^+ N_+^-}{N}\right)}}, 0 \le MCC \le 1 \end{cases} \quad (10)$$

where N^+ is the total number of 5mC sites actually containing in the sample sequences, i.e., the sum of the quantities of true positive; while N^- denotes the total number of non-5mC site sequences, i.e., the sum of the quantities of true negative; N_-^+ represents the number of true 5mC sites predicted incorrectly as non-5mC sites; N_+^- represents the number of non-5mC sites predicted incorrectly as true 5mC sites.

In addition, we used the Receiver Operating characteristic curve (ROC curve) to exam the performance of the entire integrated predictor model. The true positive rate (Sn) and false positive rate (1-Sp) were set to x-axis and y-axis to plot the ROC curve, respectively. The area under the ROC curve, also known as AUC, was used to quantify the performance of the model.

RESULTS AND CONCLUSIONS

Window Size Analysis

Considering the position specific deoxynucleotide bias, it is necessary to determine the optimal window size δ of sample sequences for identifying 5mC modification sites. Generally, if δ is too small, the residues around the 5mC modification sites cannot carry enough information, leading to poor prediction effect (Xu et al., 2019). Thus, we analyzed the trend of the precision rate of the constructed model with different window size δ. As shown in **Figure 3**, the search step size for δ here was 1nt, with a range of 10–20.

According to the intuitive observation in sub-graphs (A), (B), and (D), when $\delta = 20$, the prediction results generated by the three different methods were the best. In order to distinguish the optimal model obtained by using one-hot, DPF and onehot-DPF, we compared the most important metrics Acc and MCC values, and found that the feature method with the best effect was one-hot, as illustrated in sub-graphs (A) and (D). Therefore,

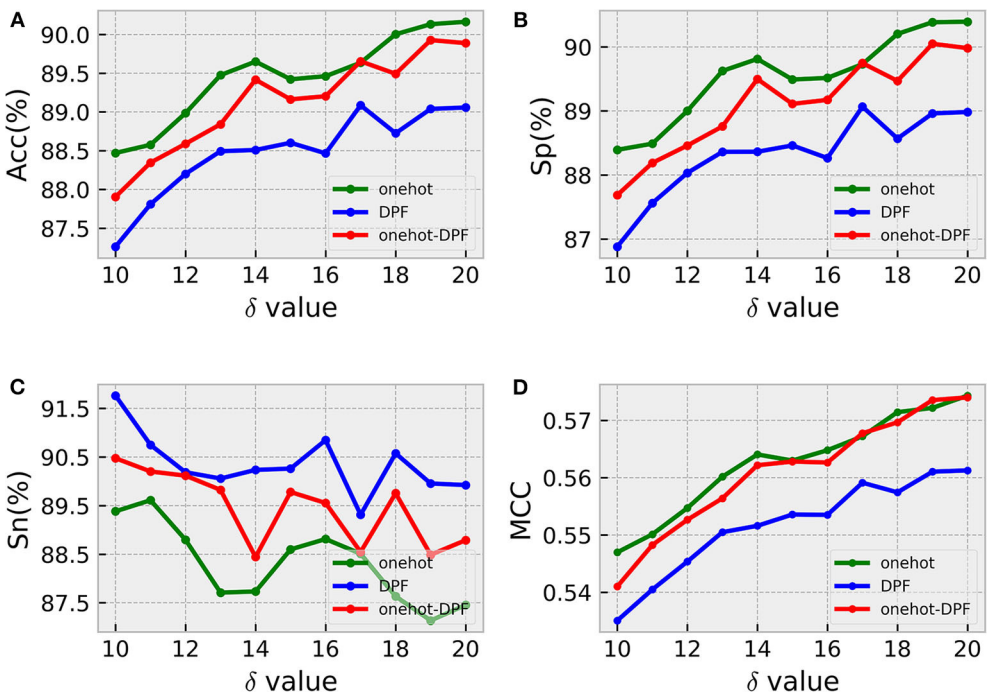


FIGURE 3 | Windows size analysis. Sub-graphs from **(A–D)** represent the ACC, Sp, Sn, MCC values generated by three different feature coding methods under different sliding window sizes, respectively.

the following analysis and calculation were based on δ with 20, indicating the length of the sample sequence formulated by Equation (1) was 41nt.

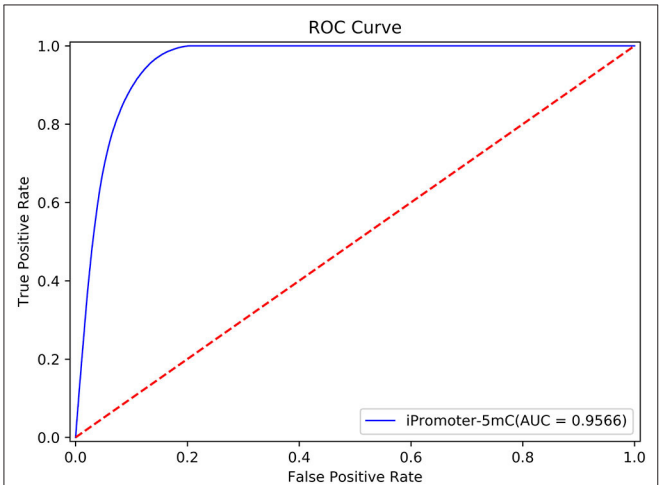


FIGURE 4 | The ROC curve of the S1 dataset on our model.

TABLE 2 | The results obtained by 5-fold cross validation on the training dataset S_1.

Method	Sn (%)	Sp (%)	Acc (%)	MCC
iPromoter-5mC	87.46	90.39	90.16	0.5743

Performance of DNN Models

According to the description in section "Framework of the integrated predictor," we can construct the 11 sub-models based on the training dataset S_1 using one-hot feature extraction method. A simple majority vote strategy was used to integrate all the decisions originated from the 11 sub-models into a final classification result. In the current study, we adopted the strict discriminating standard for identifying 5mC modification sites. If only all the sub-models consider that the potential 5mC sites is a true 5mC modification site, the iPromoter-5mC model could infer the center of this query sequence is a 5mC modification site. After 30 repeated experiments with 5-fold cross validation, we obtained the average values of each metric as the final results of the iPromoter-5mC model, as shown in **Table 2**. The results of the iPromoter-5mC model indicated that the performance of our models was promising, supported by the metric values, such as Sn, 87.46%; Sp, 90.39%; Acc, 90.16%; MCC, 0.5743. To more directly illustrate the performance of the predictor, a ROC curve was plotted using the training dataset S_1, and its corresponding AUC value was calculated. The high AUC value (0.9566) indicates that our predictor iPromoter-5mC has excellent performance and stable performance in predicting the 5mC site (**Figure 4**).

In order to validate the stability of the DNN algorithm model, we compared the performance of the DNN models constructed by one-hot, DPF, and their combination. All the results were displayed as a histogram directly on **Figure 5**. Small discrepancies of every metric value obtained by the three different methods indicated the superior stability of the DNN algorithm model to identify the 5mC modification sites.

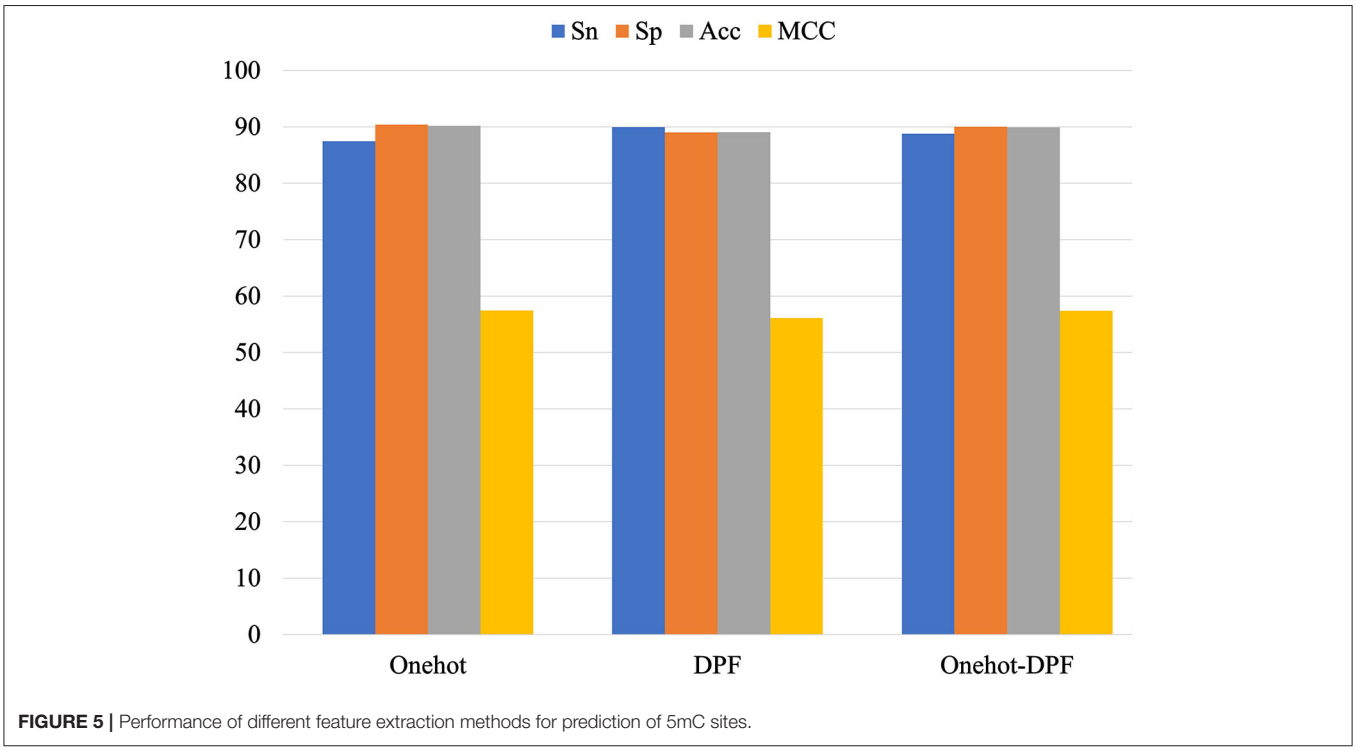


FIGURE 5 | Performance of different feature extraction methods for prediction of 5mC sites.

TABLE 3 | The performance of iPromoter-5mC based on the independent dataset.

Model number	Sn (%)	Sp (%)	Acc (%)	MCC	AUC
1	94.48	86.53	87.15	0.5455	0.9543
2	98.32	83.19	84.37	0.5183	0.9542
3	95.88	85.77	86.56	0.5417	0.9545
4	96.97	84.71	85.66	0.5319	0.9533
5	95.49	85.97	86.72	0.5425	0.9539
6	95.59	85.88	86.64	0.5417	0.9542
7	97.84	83.84	84.93	0.5244	0.9531
8	97.94	83.75	84.86	0.5238	0.9535
9	94.24	86.71	87.29	0.5469	0.9539
10	95.98	85.69	86.49	0.5409	0.9542
11	97.53	84.04	85.09	0.5256	0.9545
iPromoter-5mC	87.77	90.42	90.22	0.5771	0.9570

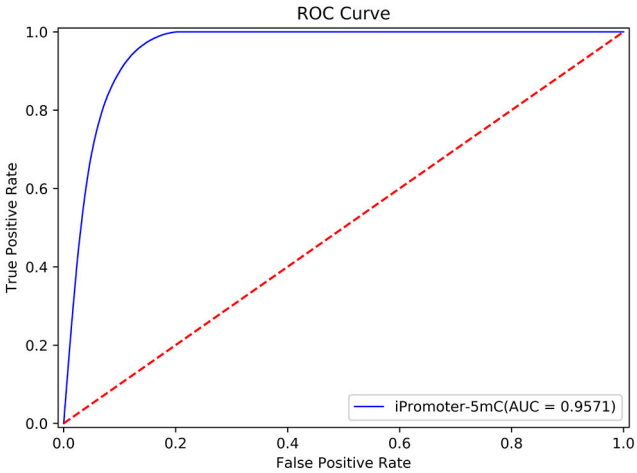


FIGURE 6 | The performance generated by the same prediction framework was still reliable and stable after the expansion of the training data.

The Robustness and Reliability Analysis

Independent test is an effective approach to check the performance of the constructed classification model. Compared with the cross-validation method, it can better verify the robustness and reliability of the prediction models. In the section "Benchmark datasets" in this study, we established the training dataset S_1 and independent testing dataset S_2. Here, we used the independent testing dataset S_2 to further test the performance of the predictor iPromoter-5mC. The results were listed in **Table 3**.

The predictive results of the 11 sub-models using the 5-fold cross-validation method on the independent test dataset S_2 were very stable at about 95, 83, 85%, 0.52 and 0.95 in Sn, Sp, Acc, MCC, and AUC, respectively, indicating that the constructed sub-models are very robust for identifying 5mC modification sites on new data. After integrating all the decisions originated from these sub-models, the independent test performance of this final model were 87.77, 90.42, 90.22%, 0.5771 and 0.9570 in Sn, Sp, Acc, MCC and AUC, respectively. The performance of the predictor iPromoter-5mC was improved, mainly seen in the metrics Acc and MCC. This implied that our designed framework for 5mC modification site prediction is reasonable and efficient, indicating that this method can be extended to realize synthetic problems on accurate prediction of other DNA/RNA modification sites.

To further validate the robustness and reliability of the prediction framework, we implemented 5-fold cross validation on the benchmark dataset S_δ including the training dataset S_1 and the independent test dataset S_2. The results of the ROC curve shown in **Figure 6** showed that the performance generated by the same prediction framework was still reliable and stable after the expansion of the training data, which have laid a solid foundation for establishment of online predictor.

We are also concerned with whether our models are applicable to the data from other cell line or tissues. To do so, we firstly constructed the benchmark dataset according to the 5mC site information in promoter regions of human hepatocarcinoma cell lines (HUH7_LIVER) from database CCLE. This dataset also was divided into the training dataset and the independent test dataset, which were also released on the GitHub and on our online server. And then, we constructed the DNN model using the same method proposed in this study. The results listed in **Table 4** were also promising, indicating that the method using in this study can also be applied to the prediction of 5mC sites in other cancer cell lines.

TABLE 4 | The 5-fold cross validation results on the training set and the independent test set of human hepatocarcinoma cell lines.

Method	Sn (%)	Sp (%)	Acc (%)	MCC	AUC
iPromoter-5mC (training)	80.53	95.79	93.73	0.7408	0.9736
iPromoter-5mC (independent test)	81.22	95.79	93.81	0.7459	0.9735

Comparison With Existing Predictor

Compared with the two early predictors Methylator and MethCGI, the predictor iDNA-Methyl has better prediction performance, which has been demonstrated in the study (Liu et al., 2015). And iDNA-Methyl has own webserver for identifying DNA 5mC sites. Therefore, we compared the performance of iPromoter-5mC with those of iDNA-Methyl. For convenience of comparison, the scores of the four indexes defined in Equation 10 obtained by these two predictors based on the independent test dataset S_2 were listed in **Table 5**. It can be observed from the table that the overall accuracy (Acc) score obtained by the current iPromoter-5mC is significantly higher than that of the existing predictors, as are the other three indicators.

We analyzed its causes and presently summarized as follows: (1) There is the biggest difference between iDNA-Methl and iPromoter-5mC. From the view of the function, iDNA-Methl detected the genome-wide methylation while iPromoter-5mC identified the methylation sites in promoters. (2) Most

TABLE 5 | Comparison of predictors' performance on the independent testing dataset S_2 and sample data from iDNA-Methyl by 5-fold cross validation, respectively.

Success rates	Dataset S2		Sample data from iDNA-Methyl	
	iPromoter-5mC	iDNA-Methyl	iPromoter-5mC	iDNA-Methyl
Sn (%)	87.77	30.62	83.48	61.25
Sp (%)	90.42	90.30	88.04	90.33
Acc (%)	90.22	85.90	86.56	77.49
MCC	0.5771	0.1730	0.7013	0.5471

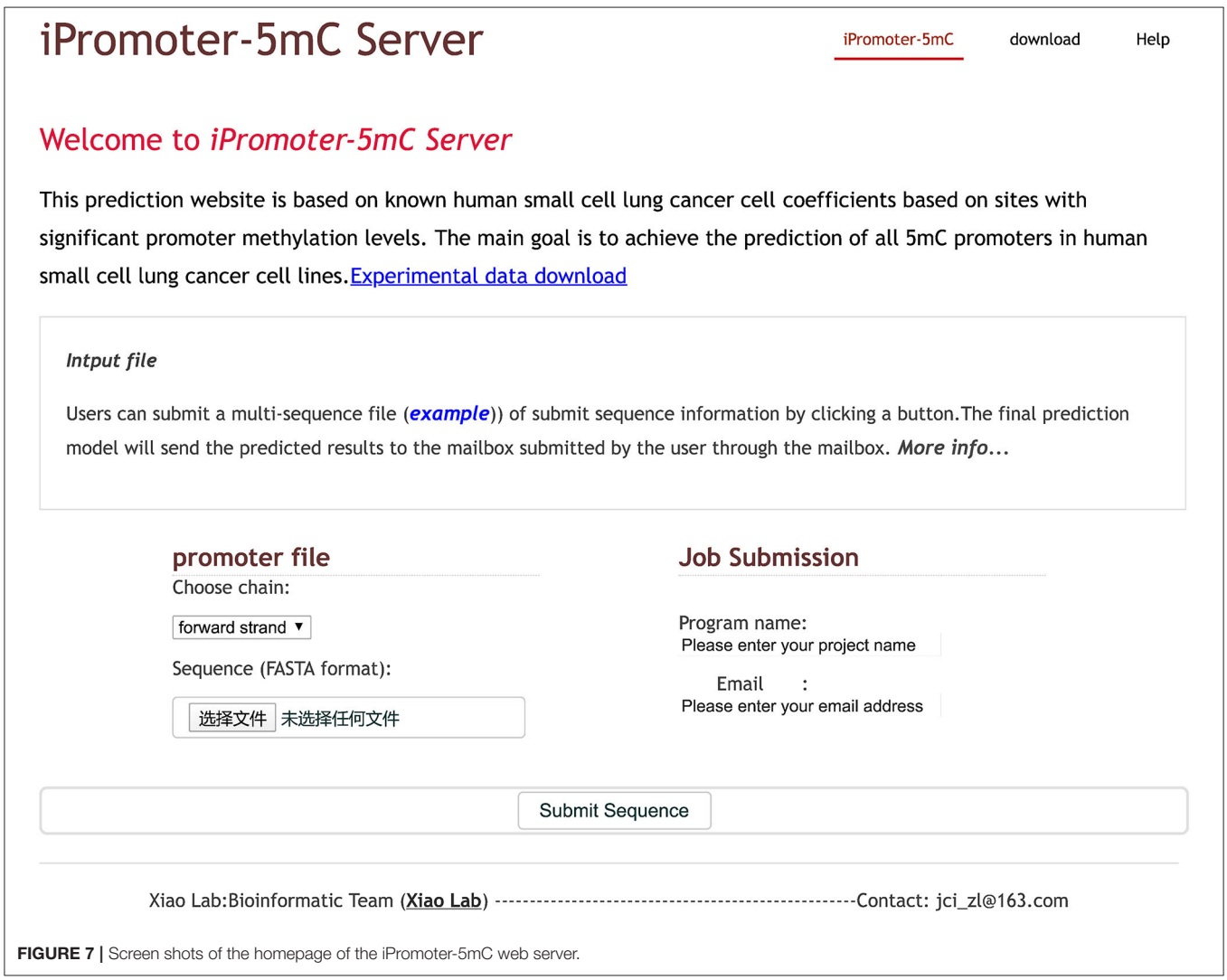


FIGURE 7 | Screen shots of the homepage of the iPromoter-5mC web server.

importantly, the sizes of their benchmark dataset are significantly different. The sample size of iPromoter-5mC is far greater than iDNA-Methl's, which enables our model to obtain better correlation between sequences, causing the phenomenon that the server iPromoter-5mC can identify the 5mC sites of the benchmark dataset from iDNA-Methl effectively while iDNA-Methl cannot. (3) The other reason is that the non-equilibrium degree of the benchmark datasets from these two predictors is significantly different. The unbalance ratio of the positive samples and negative samples from iDNA-Methyl is about 1:2, however, that of the iPromoter-5mC approximately up to 1:11.

In order to further analyze the performance of these two predictors, we implemented experiments to obtain the result by iPromoter-5mC using the sample data from iDNA-Methyl. And we found that the performance of iPromoter-5mC was better than that of iDNA-Methyl (**Table 5**), which also benefits from a large amount of data during our training.

In conclusion, these results indicated that deep learning was better suited for identify 5mC sites on a large dataset, compared to SVM. In fact, parameter optimization of SVM is extremely time-consuming, especially in the case of large amount of data. The predictor iPromoter-5mC can be an outstanding supplemental tool for identifying 5mC sites since the predictor iDNA-Methyl.

Web-Server

A user-friendly web server could provide ease of use for broad scholars to get their desired predictive results without following the complex mathematical calculations. To achieve this, we had developed an online predictor called iPromoter-5mC to identify the 5mC modification sites in promoters, following the principle described below.

For a given promoter sequence, a 41 bp scan window was used to segment the sequence into equal-size sequences. If a DNA query sequence containing potential 5mC modifications sites is in a forward strand, the base C in this DNA sequence will be selected and considered as the fixed length sequence with 41, otherwise, the base G will be found to construct the input sequence, and then be converted to the reverse complementary sequence. After that, users can follow the detailed guide to try out online experience of our web server iPromoter-5mC.

Step 1. Click the link http://www.jci-bioinfo.cn/iPromoter-5mC and then the top page of iPromoter-5mC will be shown in **Figure 7**.

Step 2. Select the strand where the sequence is located from the drop-down list box (the default value is the forward strand).

Step 3. Users can submit the file containing multiple sequences in FASTA format by clicking the submit button.

Step 4. Enter the project name and your e-mail address. The running results will be sent to you by email after finishing the work.

CONCLUSIONS

In this study, we designed a fast and effective DNN model, named iPromoter-5mC, to identify 5mC modification sites in DNA promoter region in cell lines of the small cell lung cancer. The robustness and good performance of the model were verified by feature analysis and various experiments. More importantly, Due to build an easy to use web server can provide users with more convenient, we set up an online web server to identify 5mC modification sites, which can bring great convenience to scholars' research work. The model mentioned in this paper only targets cell lines of lung small cell carcinoma, but the basic method and analysis flow can also be applied to the prediction of 5mC sites of other cancer cell lines.

Although the model in this study achieved higher predictive performance, the future is going to be one that presents many challenges. We are going to continue to study the predictive problem about DNA 5mC methylation. Firstly, with the development of single cell sequencing technology, we will try to accurately predict single-cell DNA 5mC methylation states using deep learning based on single-cell methylation data. Secondly, we plan to design a scheme to achieve accurate classification of DNA 5mC methylation level. Finally, we will construct machine learning models based on other data in cell lines of other cancers to better detect the biomarkers of those cancers.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://www.jci-bioinfo.cn/iPromoter-5mC/download.

AUTHOR CONTRIBUTIONS

XX designed the experiments. LZ constructed the predictor and established the online server. Z-CX wrote the manuscript. All authors read and approved the manuscript. In additional, thank Ang Sun for collecting the data information.

FUNDING

This work was partially supported by the National Nature Science Foundation of China (Nos. 31860312, 31760315, 61300139, 61761023), Natural Science Foundation of Jiangxi Province, China (Nos. 20171ACB20023, 20171BAB202020), the Department of Education of Jiangxi Province (GJJ160866, GJJ180733, GJJ180703), China Postdoctoral Science Foundation Funded Project (Project No. 2017M612949), Jingdezhen technology office program (20192GYZD008-04), Jiangxi province graduate student innovation special fund (YC2019-S388). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

Amoreira, C. (2003). An improved version of the DNA methylation database (MethDB). *Nucl. Acids Res.* 31, 75–77. doi: 10.1093/nar/gkg093

Angermueller, C., Lee, H. J., Reik, W., and Stegle, O. (2017). DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* 18:67. doi: 10.1186/s13059-017-1233-z

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603–607. doi: 10.1038/nature11003

Bhasin, M., Zhang, H., Reinherz, E. L., and Reche, P. A. (2005). Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Lett.* 579, 4302–4308. doi: 10.1016/j.febslet.2005.07.002

Bianchi, C., and Zangi, R. (2015). Molecular dynamics study of the recognition of dimethylated CpG sites by MBD1 protein. *J. Chem. Inf. Model.* 55, 636–644. doi: 10.1021/ci500657d

Bird, A. (2007). Perceptions of epigenetics. *Nature* 447, 396–398. doi: 10.1038/nature05913

Chen, W., Feng, P., Song, X., Lv, H., and Lin, H. (2019a). iRNA-m7G: identifying N(7)-methylguanosine sites by fusing multiple features. *Mol. Ther. Nucl. Acids* 18, 269–274. doi: 10.1016/j.omtn.2019.08.022

Chen, W., Song, X., Lv, H., and Lin, H. (2019b). iRNA-m2G: identifying N(2)-methylguanosine sites based on sequence-derived information. *Mol. Ther. Nucl. Acids* 18, 253–258. doi: 10.1016/j.omtn.2019.08.023

Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 33, 3518–3523. doi: 10.1093/bioinformatics/btx479

Dao, F. Y., Lv, H., Wang, F., Feng, C. Q., Ding, H., Chen, W., et al. (2019). Identify origin of replication in *Saccharomyces cerevisiae* using two-step feature selection technique. *Bioinformatics* 35, 2075–2083. doi: 10.1093/bioinformatics/bty943

Deichmann, U. (2016). Epigenetics: the origins and evolution of a fashionable topic. *Dev. Biol.* 416, 249–254. doi: 10.1016/j.ydbio.2016.06.005

Down, T. A., Rakyan, V. K., Turner, D. J., Flicek, P., Li, H., Kulesha, E., et al. (2008). A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat. Biotechnol.* 26, 779–785. doi: 10.1038/nbt1414

Fang, F., Fan, S., Zhang, X., and Zhang, M. Q. (2006). Predicting methylation status of CpG islands in the human brain. *Bioinformatics* 22, 2204–2209. doi: 10.1093/bioinformatics/btl377

Gessler, M. (1999). WT1 (Wilms' tumor suppressor gene). *Atlas Genet. Cytogenet. Oncol. Haematol.* 3, 177–178. doi: 10.4267/2042/37552

Ghandi, M., Huang, F. W., Jane-Valbuena, J., Kryukov, G. V., Lo, C. C., Mcdonald, E. R. III., et al. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* 569, 503–508. doi: 10.1038/s41586-019-1186-3

Islam, M. M., Tian, Y., Cheng, Y., Wang, Y., and Hu, P. (2018). A deep neural network based regression model for triglyceride concentrations prediction using epigenome-wide DNA methylation profiles. *BMC Proc.* 12:21. doi: 10.1186/s12919-018-0121-1

Kang, Y. H., Lee, H. S., and Kim, W. H. (2002). Promoter methylation and silencing of PTEN in gastric carcinoma. *Lab. Invest.* 82, 285–291. doi: 10.1038/labinvest.3780422

Li, H., Ning, S., Ghandi, M., Kryukov, G. V., Gopal, S., Deik, A., et al. (2019). The landscape of cancer cell line metabolism. *Nat. Med.* 25, 850–860. doi: 10.1038/s41591-019-0404-8

Liu, Q., Georgieva, D. C., Egli, D., and Wang, K. (2018). NanoMod: a computational tool to detect DNA modifications using nanopore long-read sequencing data. *BMC Genomics* 20(Suppl. 1):78. doi: 10.1101/277178

Liu, Z., Xiao, X., Qiu, W. R., and Chou, K. C. (2015). iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition. *Anal. Biochem.* 474, 69–77. doi: 10.1016/j.ab.2014.12.009

Mansour, H. (2014). Cell-free nucleic acids as noninvasive biomarkers for colorectal cancer detection. *Front. Genet.* 5:182. doi: 10.3389/fgene.2014.00182

Michalak, E. M., Burr, M. L., Bannister, A. J., and Dawson, M. A. (2019). The roles of DNA, RNA and histone methylation in ageing and cancer. *Nat. Rev. Mol. Cell Biol.* 20, 573–589. doi: 10.1038/s41580-019-0143-1

Muller, F., Scherer, M., Assenov, Y., Lutsik, P., Walter, J., Lengauer, T., et al. (2019). RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol.* 20:55. doi: 10.1186/s13059-019-1664-9

Nicoglou, A., and Merlin, F. (2017). Epigenetics: a way to bridge the gap between biological fields. *Stud. Hist. Philos. Biol. Biomed. Sci.* 66, 73–82. doi: 10.1016/j.shpsc.2017.10.002

Siegel, R. L., Miller, K. D., and Jemal, A. (2018). Cancer statistics, 2018. *CA Cancer J. Clin.* 68, 7–30. doi: 10.3322/caac.21442

Wei, L., Chen, H., and Su, R. (2018). M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Mol. Ther. Nucl. Acids* 12, 635–644. doi: 10.1016/j.omtn.2018.07.004

Wei, L., Su, R., Luan, S., Liao, Z., Manavalan, B., Zou, Q., et al. (2019). Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics* 35, 4930–4937. doi: 10.1093/bioinformatics/btz408

Xia, C., Xiao, Y., Wu, J., Zhao, X., and Li, H. (2019). "A convolutional neural network based ensemble method for cancer prediction using dna methylation data," in *Proceedings of the 2019 11th International Conference on Machine Learning and Computing - ICMLC '19* (Zhuhai), 191–196. doi: 10.1145/3318299.3318372

Xu, Z. C., Feng, P. M., Yang, H., Qiu, W. R., Chen, W., and Lin, H. (2019). iRNAD: a computational tool for identifying D modification sites in RNA sequence. *Bioinformatics* 35, 4922–4929. doi: 10.1093/bioinformatics/btz358

Zhuang, Z., Shen, X., and Pan, W. (2019). A simple convolutional neural network for prediction of enhancer-promoter interactions with DNA sequence data. *Bioinformatics* 35, 2899–2906. doi: 10.1093/bioinformatics/bty1050

Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 173, 346–354. doi: 10.1016/j.neucom.2014.12.123

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.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 11 August 2020
doi: 10.3389/fcell.2020.00741

m^{6}A Reader: Epitranscriptome Target Prediction and Functional Characterization of *N*^{6}-Methyladenosine (m^{6}A) Readers

Di Zhen[1†], Yuxuan Wu[1†], Yuxin Zhang[1†], Kunqi Chen[1,2], Bowen Song[3,4], Haiqi Xu[1], Yujiao Tang[1,4], Zhen Wei[1,2] and Jia Meng[1,4,5]*

[1] Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, China, [2] Institute of Ageing and Chronic Disease, University of Liverpool, Liverpool, United Kingdom, [3] Department of Mathematical Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, China, [4] Institute of Integrative Biology, University of Liverpool, Liverpool, United Kingdom, [5] AI University Research Centre, Xi'an Jiaotong-Liverpool University, Suzhou, China

OPEN ACCESS

Edited by:
Yu Xue,
Huazhong University of Science and Technology, China

Reviewed by:
Fuyi Li,
Monash University, Australia
Nicolas Reynoird,
INSERM U1209 Institut pour l'Avancée des Biosciences (IAB), France
Lin Zhang,
China University of Mining and Technology, China

***Correspondence:**
Kunqi Chen
kunqi.chen@liverpool.ac.uk;
kunqi.chen@xjtlu.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: 19 May 2020
Accepted: 16 July 2020
Published: 11 August 2020

Citation:
Zhen D, Wu Y, Zhang Y, Chen K, Song B, Xu H, Tang Y, Wei Z and Meng J (2020) m^{6}A Reader: Epitranscriptome Target Prediction and Functional Characterization of *N*^{6}-Methyladenosine (m^{6}A) Readers. Front. Cell Dev. Biol. 8:741. doi: 10.3389/fcell.2020.00741

N^6-methyladenosine (m^{6}A) is the most abundant post-transcriptional modification in mRNA, and regulates critical biological functions via m^{6}A reader proteins that bind to m^{6}A-containing transcripts. There exist multiple m^{6}A reader proteins in the human genome, but their respective binding specificity and functional relevance under different biological contexts are not yet fully understood due to the limitation of experimental approaches. An *in silico* study was devised to unveil the target specificity and regulatory functions of different m^{6}A readers. We established a support vector machine-based computational framework to predict the epitranscriptome-wide targets of six m^{6}A reader proteins (YTHDF1-3, YTHDC1-2, and EIF3A) based on 58 genomic features as well as the conventional sequence-derived features. Our model achieved an average AUC of 0.981 and 0.893 under the full-transcript and mature mRNA model, respectively, marking a substantial improvement in accuracy compared to the sequence encoding schemes tested. Additionally, the distinct biological characteristics of each individual m^{6}A reader were explored via the distribution, conservation, Gene Ontology enrichment, cellular components and molecular functions of their target m^{6}A sites. A web server was constructed for predicting the putative binding readers of m^{6}A sites to serve the research community, and is freely accessible at: http://m6areader.rnamd.com.

Keywords: *N6*-methyladenosine, m^{6}A reader, machine learning (ML), YTH domain, eIF3a

INTRODUCTION

In the exploration of RNA epigenetics, more than 150 types of RNA modification have been identified (Boccaletto et al., 2018). The methylation of adenosine at the N6 position (m^{6}A) is the most prevalent post-transcriptional modification in the mRNA (Meyer and Jaffrey, 2017), which was discovered in a wide range of eukaryotic RNAs (Adams and Cory, 1975) as well as viral RNAs (Gokhale et al., 2016). m^{6}A was considered as a potential mRNA processing regulator in 1970s (Desrosiers et al., 1974), and subsequent studies noticed intensive functions of it (Patil et al., 2018), including cardiac gene expression (Kmietczyk et al., 2019), cell growth,

neuronal development (Chen J. et al., 2019), stress response (Engel et al., 2018), translation initiation, and stabilizing junctional RNA (Liu B. et al., 2018).

Similar to other epigenetic modifications, m^6A is thought to be dynamic and reversible (Song et al., 2019). It can be installed by methyltransferase (writers) or removed by demethylase (erasers). This internal modification also attracts specific binding proteins, namely readers, which bind selectively to m^6A-containing transcripts (Liao et al., 2018). Additionally, m^6A performs many functions through interacting with "reader" proteins (Hazra et al., 2019). The most widely studied readers are YT521-B homology (YTH) family of proteins, which possess the evolutionarily conserved YTH domain that recognizes m^6A mark. The YTH domain consists of 100–150 residues and adopts alpha/beta fold, with 4–5 alpha helices surrounding a curved six-stranded beta sheet (Zhang et al., 2010). In human, five m^6A readers were reported to have the YTH domain, namely YTHDF1,2,3 and YTHDC1,2. However, the YTH domain is not indispensable for m^6A readers, a subunit of translation initiation complex factor EIF3 complex, called EIF3A, was reported as an m^6A reader lacking YTH domain (Meyer et al., 2015).

The m^6A reader YTHDC1 is predominantly found in the nucleus, while YTHDC2 and YTHDF1,2,3 are cytoplasmic (Patil et al., 2016). YTHDC1 and YTHDC2 are unrelated to other members of the YTH family based on amino acid sequence, size or overall YTH domain organization (Patil et al., 2018). By contrast, YTHDF family comprises three paralogs, YTHDF1-3, that share high sequence identity with about 85% of sequence similarity (Hazra et al., 2019). YTHDC1 and three YTHDF proteins contain a single C-terminal YTH domain that binds to m^6A marker by a segment rich of proline, glutamate and aspartate. Compared to other YTH domain-containing proteins, whose YTH domains are embedded in low complexity regions, YTHDC2 has a unique multidomain structure (Hazra et al., 2019). N-terminal R3H domain, central DEAH-box helicase domain and helicase associated 2 domain are also found in YTHDC2 apart from the C-terminal YTH domain. Different from the structures of five YTH domain-containing proteins, EIF3 is a large multiprotein complex comprising 13 subunits (Meyer et al., 2015). The EIF3 binding sites are predominantly mapped at the 5′ untranslated region (5′ UTR) (Lee et al., 2015), whereas the binding sites of YTH domain-containing proteins are usually located near the stop codon.

In addition to different cellular locations and structures, m^6A readers appear to function through various post-transcriptional control mechanisms to regulate RNAs dynamically. Human YTHDC1 has been demonstrated to participate in RNA splicing by interacting with serine/arginine splicing factor SRSF3, which is involved in exon inclusion and exclusion splicing (Ye et al., 2017). As a putative RNA helicase, YTHDC2 enhances the translation of target RNAs and reduces the abundance of target RNAs (Hsu et al., 2017). YTHDF2 is verified to decrease the stability and control the lifetime of its targeted methylated mRNA transcripts (Du et al., 2016), while YTHDF1 ensures efficient protein expression from their shared regions (Wang et al., 2015). YTHDF3, the third member of YTHDF family, has been proposed to share common targets (about 60%) with both YTHDF1 and YTHDF2 (Shi et al., 2017). This suggests potential coordination in regulating gene expression by YTHDF family proteins. YTHDF3 can promote the function of YTHDF1 by interacting with some ribosomal proteins to facilitate mRNA translation. When associating with YTHDF2, YTHDF3 could participate in mRNA decay. In addition to the five members of YTH family, EIF3A plays an important role in biological processes as well. It can act as both repressor and activator of cap-dependent transcript-specific translation through directly binding to m^6A marked mRNA sequence (Lee et al., 2015).

Since the five YTH family proteins (YTHDC1-2 and YTHDF1-3) and EIF3A present distinctive structures and properties, it is worth studying the preferential binding sites in the m^6A marked transcripts for each m^6A reader.

Single base resolution techniques such as miCLIP (Linder et al., 2015) are developed and are fairly effective on screening m^6A sites, and it is usually based on the iCLIP or Par-CLIP approach (Meyer et al., 2015) to identify the binding sites of each m^6A reader. As these wet-lab experiments are costly and laborious, computational methods may provide a viable avenue. To date, a large number of RNA methylation sites have been reported, providing sufficient information for effective computational prediction. A huge amount of data extracted from experiments encouraged the establishment of a number of effective m^6A site predictors, including WHISTLE (Chen K. et al., 2019), SRAMP (Zhou et al., 2016), BERMP (Huang et al., 2018), and Gene2vec (Zou et al., 2019). However, to our knowledge, the prediction dedicated to the target specificity of the readers is absent. In this project, we constructed a predictor, m^6A reader, to distinguish the substrate of each m^6A reader. A comprehensive analysis of these readers was then performed, including the analysis of distribution, conservation, GO enrichment, cellular components and molecular functions of their respective epitranscriptome target sites.

MATERIALS AND METHODS

Collection of m^6A Sites and the Target Sites of m^6A Readers

The transcriptome-wide m^6A sites were collected from 17 different conditions generated from 6 different epitranscriptome profiling approaches of base-resolution or high resolution (**Table 1**).

In this study, we consider the binding sites of six m^6A readers identified by Par-CLIP or iCLIP approaches. Specifically, a total of 16,664 m^6A sites located on 4,722 different genes reported by four experiments were considered as the target sites of YTHDC1, and 1,234 sites on 275 genes identified by two experiments were considered as the target sites for YTHDC2. For the three proteins from YTHDF family, three experiments for each reader proposed 25,597, 28,970, and 7,253 target sites located on 6,714, 6,677, and 3,495 genes for YTHDF1, YTHDF2, and YTHDF3, respectively. Two CLIP experiments conducted on HEK2937T cell line discovered 756 sites located in 470 genes on marked RNA transcripts, which are targeted by EIF3A. The testing datasets

TABLE 1 | Base-resolution or high resolution datasets of m^{6}A sites.

Dataset	Technique	Cell line	GEO	References
S1	miCLIP	MOLM13	GSE98623	Vu et al., 2017
S2		HEK293	GSE63753	Linder et al., 2015
S3		HepG2	GSE73405	Meyer et al., 2015
S4		HEK293T	GSE122948	Boulias et al., 2019
S5		HepG2	GSE121942	Huang et al., 2019
S6		HCT116	GSE128699	van Tran et al., 2019
S7	m^{6}A-CLIP	HeLa	GSE86336	Ke et al., 2017
S8		CD8T	GSE71154	Ke et al., 2015
S9		A549		
S10	MAZTER-seq	HEK293T	GSE122961	Garcia-Campos et al., 2019
S11		ESC		
S12	m^{6}A-REF-seq	HEK293	GSE125240	Zhang et al., 2019c
S13		Brain		
S14		Kidney		
S15		Liver		
S16	PA-m^{6}A-seq	HeLa	GSE54921	Chen K. et al., 2015a
S17	m^{6}A-seq (improved protocol)	A549	GSE54365	Schwartz et al., 2014

and training datasets are strictly segregated under all conditions. Detailed information of the target sites of m^{6}A readers analyzed in this study was summarized in **Table 2**.

Feature Encoding Scheme and Selection

We considered both the conventional sequence-derived features and the genome-derived features.

The sequence-derived features were summarized in the iLearn (Chen Z. et al., 2019; Chen et al., 2020) and BioSeq-Analysis (Liu, 2019; Liu et al., 2019), which can be divided into six different classes. Based on their classification, we chose one method from each class including nucleic acid composition (Lee et al., 2011), binary encoding method (Wu et al., 2015), position-specific tendencies of trinucleotide (He et al., 2018), electron-ion interaction pseudopotentials (He et al., 2019), Autocorrelation and pseudo k-tupler composition (Liu et al., 2015). Also, the chemical property combined with nucleic frequency, which is a popular encoding method in recent years (Bari et al., 2013; Chen et al., 2016a,b, 2017a; Li et al., 2018), was also used in performance testing for m^{6}A reader target prediction.

The genomic features shown in previous projects (Chen K. et al., 2019; Song et al., 2019) are effective in RNA modification prediction. In order to improve the performance of the predictor, 58 mammalian genome features belonging to 9 classes were applied. All the features used were generated by the "GenomicFeatures R/Bioconducter" package using the transcript annotations hg19 TxDb package (Lawrence et al., 2013). The first class involves dummy variables indicating whether the adenosine site overlaps the topological region within the RNA transcript. The second class specifies the relative position of the adenosine site on the region, while the third class tells the length of the target mRNA transcript. Features belonging to the fourth class measure the nucleotide distances to the splicing junction and the nearest neighboring site. The fifth and sixth classes are based on clustering information of modification sites and scores related to conservation (Siepel et al., 2005; Gulko et al., 2015), respectively. The last three feature groups describe RNA secondary structures (Lorenz et al., 2011), genomic properties and attributes of the genes or transcripts, respectively. More details of the genomic features considered in our analysis were presented in **Supplementary Table S1**.

Feature Selection Technique

With multiple features, the dimension of dataset increases, leading to overfitting, information redundancy or increased computational time. To solve this problem, feature selection

TABLE 2 | Target sites of m^{6}A readers identified by Par-CLIP or iCLIP.

Dataset	Reader	Source	Site #	Total #	Gene #	Cell line
D1	YTHDC1	GSE74397 (Roundtree et al., 2017)	482	16,664	4,722	HeLa
D2		GSE58352 (Xu et al., 2014)	2,633			
D3		GSE71096 (Xiao et al., 2016)	2,430			
D4		GSE78030 (Patil et al., 2016)	12,309			HEK293T
D5	YTHDC2	GSE98085 (Hsu et al., 2017)	1,183	1,234	275	HeLa
D6		GSE78030 (Patil et al., 2016)	131			HEK293T
D7	YTHDF1	GSE63591 (Wang et al., 2015)	4,541	25,597	6,714	HeLa
D8		GSE83438 (Gokhale et al., 2016)	2,527			Huh7
D9		GSE78030 (Patil et al., 2016)	20,694			HEK293T
D10	YTHDF2	GSE49339 (Wang et al., 2014)	22,688	28,970	6,677	HeLa
D11		GSE83438 (Gokhale et al., 2016)	5,147			Huh7
D12		GSE78030 (Patil et al., 2016)	6,280			HEK293T
D13	YTHDF3	GSE86214 (Shi et al., 2017)	2,608	7,253	3,495	HeLa
D14		GSE83438 (Gokhale et al., 2016)	177			Huh7
D15		GSE78030 (Patil et al., 2016)	5,082			HEK293T
D16	EIF3A	GSE65004 (Lee et al., 2015)	45	756	470	HEK293T
D17		GSE73405 (Meyer et al., 2015)	731			

is effective in optimizing relevant modeling variables and improving the accuracy of the constructed models. In this study, we performed feature selection using F-score technique (Lin et al., 2014; Dao et al., 2019). Technically, F-score is a wrapper-type feature selection algorithm, used to measure the degree of difference between two real-number data sets. For a given training sample x_d, there are n^+ positive samples and n^- negative samples. The F-score for the i-th feature can be calculated as:

$$F_i = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n^+-1}\sum_{k=1}^{n^+}(\bar{x}_{d,\,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n^--1}\sum_{d=1}^{n^+}(\bar{x}_{d,\,i}^{(-)} - \bar{x}_i^{(-)})^2}$$

where $\bar{x}_i^{(+)}$, $\bar{x}_i^{(-)}$ and $\bar{x}_i$ denote the average frequency of the i-th feature in the positive, negative and the whole samples, respectively; $\bar{x}_{d,\,i}^{(+)}$ and $\bar{x}_{d,\,i}^{(-)}$ represent the value of the i-th feature of the d-th sequence in the positive and negative samples, respectively. A larger F-score value means better predictive ability of a feature. To demonstrate this relative distinguishing ability of every genomic feature, the computed F-score values were rescaled between 0 and 1, and ranked in the descending order. Referring to this ranking, we used incremental feature selection (IFS) and SVM method to complete the selection process (Chen and Lin, 2006; Lin et al., 2014). Specifically, the feature subset begins with the feature with the highest F-score, and the next feature subset contains the last feature subset and one next feature. AUC values of 5-fold cross-validation were obtained for each feature subset.

Machine Learning Approach and Performance Evaluation

To reduce the bias in the experiment, especially when selecting the polyA RNAs during library preparation, we built separate prediction models using full transcript data and mature mRNA data, respectively. In the mature mRNA predictor, only m^{6}A sites located in exon regions are considered.

Since the positive-to-negative ratio of our datasets was highly unbalanced (1:10), we randomly split the negative data into ten parts and combined with the positive dataset with 1:1 positive-to-negative ratio to avoid the unfavorable choice of machine learning classifiers. Subsequently, 10 models were trained and the average outcome score was reported as the performance of the classifier. For each m^{6}A reader, the target sites identified in different experiments were mixed, and then the predictor was trained with 80% of the total sites before being evaluated by the remaining 20% of sites for independent testing. Specifically, the mature mRNA datasets for YTHDF1-3, YTHDC1-2, EIF3a have 39577, 44025, 11065, 24312, 1245, and 1200 training data, and 9895, 11007, 2767, 6078, 311, and 300 testing data. The full transcript datasets for those m^{6}A readers have 40955, 46352, 11605, 26662, 1970, and 1210 training data, and 10239, 11588, 2901, 6666, 492, and 302 testing data.

Machine learning algorithms have been widely applied in many fields of biological research such as predicting structural and functional properties of biological sequences. We applied Support Vector Machine (SVM) (Chang and Lin, 2011) to compare encoding schemes and approaches. To identify a better algorithm for model construction, we compared multiple machine learning algorithms including SVM, Logistic Regression (LR), Random Forest (RF), and XGBoost.

To validate the model performance, besides 5-fold cross-validation, we also applied the cross-sample test, in which the sites reported from one sample (or condition) were reserved for testing purpose and the sites reported in all other samples (or conditions) were used for training. This testing mode directly evaluates the capability of the prediction approach to detect reader-specific target sites under a single biological condition not profiled previously. Besides, four commonly used performance metrics are used for performance evaluation, including Area under the ROC Curve (AUC) (Bradley, 1997), Precision-Recall Curve (PR AUC) (Keilwagen et al., 2014), accuracy (Acc) (Jin and Ling, 2005) and Mathew's correlation coefficient (MCC) (Powers, 2008). The formula of Acc and MCC are as follows:

$$Acc = \frac{TP + TN}{TP + FN + TN + FP}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

where TP is the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives.

Model construction and performance evaluation were conducted in R (Version 3.6.3). Machine learning algorithms were supported by caret package (Kuhn, 2020).

RESULTS AND DISCUSSION

Feature Selection

Due to the high reliability and effectiveness in reflecting intrinsic relation to the targets, sequence-derived features have been widely used and achieved high accuracy in extensive researches focusing on the m^{6}A site prediction. However, genome-derived features have been discovering and currently showing a new perspective in feature extraction (Zhou et al., 2016; Chen et al., 2017a). Here, we extracted genome features from 41 bp sequence data. We employed WHISTLE approach to combine both sequence-derived features and genome-derived features to predict the target specificity of m^{6}A readers. To increase robustness and reduce overfitting of the predicter, feature selection was performed, where those most relevant features to the targets were identified.

Initially, all the genomic features were normalized to ensure the equal contribution of each feature. Then the F-score method was applied to allow all features to be ranked accordingly. Combining IFS and SVM, AUC value of 5-fold cross-validation were obtained for each feature subset. By examining AUC scores, the best performance was achieved by the optimal feature subset. The detailed feature selection results were summarized in **Supplementary Figures S1–S6** for YTHDF1-3, YTHDC1-2 and EIF3A under both the full transcript and mature mRNA transcript, respectively. For example, it can

TABLE 3 | Target prediction performance under cross-condition test.

Mode	Method	YTHDC1	YTHDC2	YTHDF1	YTHDF2	YTHDF3	EIF3A	Average
Full transcript model	m^6A reader	0.974	0.920	0.983	0.983	0.992	1.000	0.975
	Composition	0.769	0.713	0.773	0.778	0.782	0.893	0.785
	MethyRNA	0.763	0.611	0.795	0.794	0.787	0.849	0.767
	EIIP	0.770	0.713	0.768	0.778	0.782	0.894	0.784
	PseKNC	0.733	0.643	0.743	0.755	0.753	0.852	0.747
	AutoCo	0.651	0.586	0.673	0.684	0.737	0.835	0.694
	PSNP	0.777	0.654	0.816	0.816	0.894	0.869	0.804
	onehot	0.750	0.603	0.796	0.795	0.791	0.858	0.766
Mature mRNA model	m^6A reader	0.815	0.730	0.983	0.839	0.883	0.987	0.873
	Composition	0.660	0.503	0.773	0.667	0.707	0.872	0.697
	MethyRNA	0.659	0.631	0.795	0.695	0.733	0.833	0.724
	EIIP	0.670	0.504	0.768	0.667	0.727	0.871	0.701
	PseKNC	0.635	0.593	0.743	0.630	0.706	0.837	0.691
	AutoCo	0.527	0.556	0.673	0.559	0.688	0.820	0.637
	PSNP	0.703	0.675	0.816	0.754	0.858	0.870	0.779
	onehot	0.662	0.622	0.796	0.696	0.757	0.836	0.728

In this test, the sites generated from each sample were used for independent testing, while all other samples were used for training, so the training sites and the test sites were not reported from the same condition. This is often the real scenario of interest where models are constructed to predict target sites under a new biological context. See ***Supplementary Tables S2–S6*** *for more detailed results.*

be observed in **Supplementary Figure S6A** that, the best performance of EIF3A target prediction was achieved with the top 44 features for the mature mRNA model. Therefore, only the top 44 features were used ultimately to build the mature mRNA prediction models for EIF3A target prediction. Likewise, feature selection in target prediction was conducted for every other reader, and the predictors were constructed in the same way.

Performance Based on Different Features

With the nucleotide encoding methods based on chemical properties, extensive studies have achieved high accuracy in the m^6A site prediction. However, for the first time, we explored and compared different sequence encoding schemes for predicting the target specificity of m^6A-binding proteins.

For each m^6A reader, the target sites identified in different experiments were mixed, and then the predictor was trained with 80% of the total sites before being evaluated by the remaining 20% of sites for independent testing. As a comparison, the performance of 5-fold cross-validation on the training data was also reported. As shown in **Supplementary Table S7**, m^6A reader achieved AUC scores of 0.981 and 0.893 in independent testing under the full transcript and mature mRNA models, respectively. This performance is substantially better than other approaches that did not take advantage of genome-derived features.

Subsequently, we evaluated the capability of the proposed method in identifying the reader-specific target m^6A sites under different biological contexts. In this test, the sites generated from each sample were used for independent testing, while all other samples were used for training, so the training sites and the test sites were not reported from the same condition. This is often the real scenario of interest where models are constructed to predict target sites in a new biological context. Besides this cross-condition test, the results of 5-fold cross-validation on the training data were also presented. The detailed evaluation results on every individual sample for every reader are shown in **Supplementary Tables S2–S6**, with a summary of the cross-condition tests presented in **Table 3**. It can be seen that our approach achieved a high accuracy with AUC scores of 0.975 and 0.873 under full transcript and mature mRNA models in the cross-condition test. The performance is again substantially better than the competing methods.

Detect Potential Substrate of m^6A Readers

In order to further confirm the reliability and efficiency of our predictors, we used our predictors to detect m^6A reader binding sites on the unidentified regions. As expected, all m^6A readers

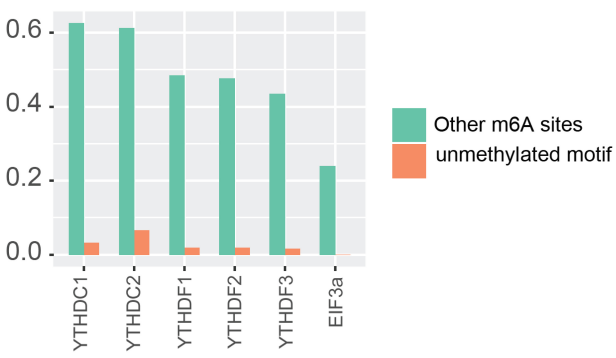


FIGURE 1 | Potential substrate of m^6A readers.

bind to more than 20% m^{6}A sites, while they bind to less than 10% unmethylated motifs as shown in **Figure 1**. The binding preference is significant and reasonable, which demonstrated the high discrimination ability of our predictors. Moreover, we compared the previous binding sites of YTHDF family (**Figure 2A**) and the prediction result of them on unidentified regions (**Figure 2B**). The wet-lab and prediction result shows that readers in YTHDF family have both common and distinct binding sites, suggesting that the binding sites of YTHDF proteins are not exactly identical. This is not consistent with the conclusion in the previous study that YTHDF proteins bind to identical sites on all m^{6}A mRNAs (Zaccara and Jaffrey, 2020). Our result suggests that YTHDF family proteins have similar functions of mediating degradation of m^{6}A mRNAs, and they also have different functions in mRNA regulation simultaneously. This result is consistent with our GO enrichment analysis, and also partially supports that m^{6}A readers' effect on downstream processes are much more heterogeneous and context-dependent across transcripts (Zhang et al., 2020). The predicted probabilities for the targeting of each m^{6}A reader are provided on the download page of the website[1].

Model Comparison

To discover a better machine learning algorithm for our proposed models, we compared the performance of SVM, LR, RF, and XGBoost on mature mRNA and full transcript data for the prediction of target specificity of six m^{6}A readers. In general, the performances of different machine learning algorithms are all very high (>0.8 for mature mRNA models and >0.9 for full transcript models) and have little difference among them as

[1]http://m6Areader.rnamd.com

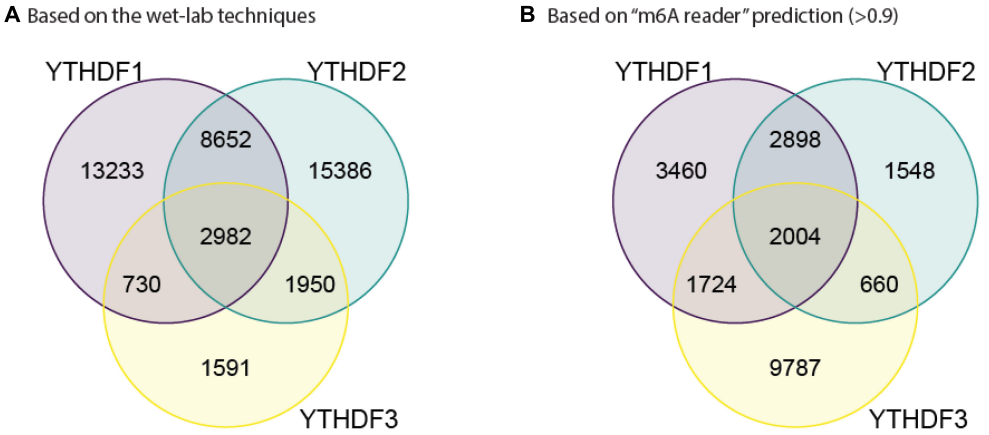


FIGURE 2 | Substrate overlap between YTHDF family.

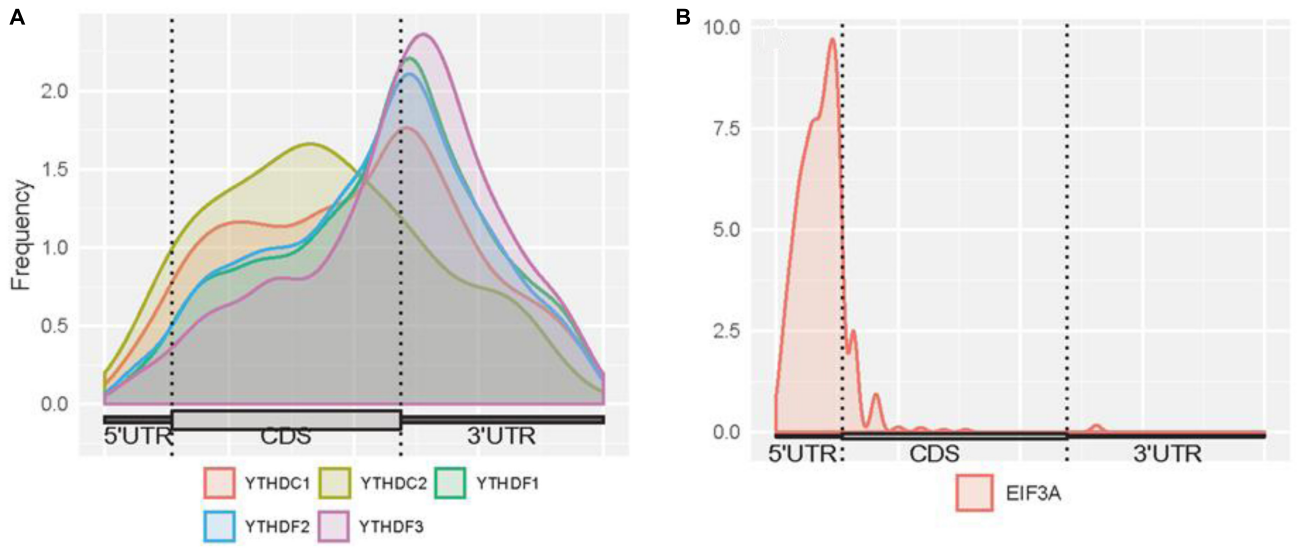


FIGURE 3 | Distribution of m^{6}A readers binding sites on mRNAs. **(A)** Distribution of the binding sites of YTHDC1, YTHDC2, YTHDF1, YTHDF2, and YTHDF3 on mRNAs. **(B)** Distribution of the binding sites of EIF3A on mRNAs. The figures were plotted using the Guitar R/Bioconductor package (Cui et al., 2016).

shown in **Supplementary Table S8**. Therefore, we decided to use SVM classifier for the predictors.

Characterizing the Target Specificity of m^{6}A Readers

Our result suggests that the substrates of m^{6}A readers can be classified, reflecting the distinct biological characteristics of each m^{6}A reader. We thus explored the distribution, conservation, and functional relevance of the substrates of each m^{6}A reader.

Here, we firstly examined the distribution of binding sites for each reader (**Figure 3**). High enrichment of YTHDC1 is observed around stop codons and CDSs. However, it can be noticed that the binding abundance of YTHDC1 is relatively lower than members of YTHDF family in stop codons, while it is highly enriched in CDSs. This is consistent with the fact that YTHDC1 is not only targeting to m^{6}A sites at its C terminus but also directly interacting with pre-mRNA splicing factor SRSF3 or SRSF10, which prefers to reside on the upper stream of m^{6}A sites (Roundtree et al., 2017). The spatial association among those proteins implicates the process of recruiting pre-mRNA splicing factors and inducing mRNA splicing outcomes. Surprisingly, YTHDC2 targets are more enriched in CDSs near stop codons than in 3′ UTR, suggesting that YTHDC2 is distinct from other m^{6}A readers. As YTHDC2 is reported to be the largest protein (∼160 kDa) among all YTH family members and with numerous RNA binding domains (e.g., helicase domain and two Ankyrin repeats, Hsu et al., 2017) apart from YTH domain, besides its acknowledged functions of accelerating translation and degradation of mRNAs as an m^{6}A reader, it is possible that there are potential underlying functions independent from m^{6}A-binding remained to be discovered. For instance, the recent study indicated that YTHDC2 as an RNA induced ATPase moves along the RNA from 3′ to 5′ with helicase activity, and interacts

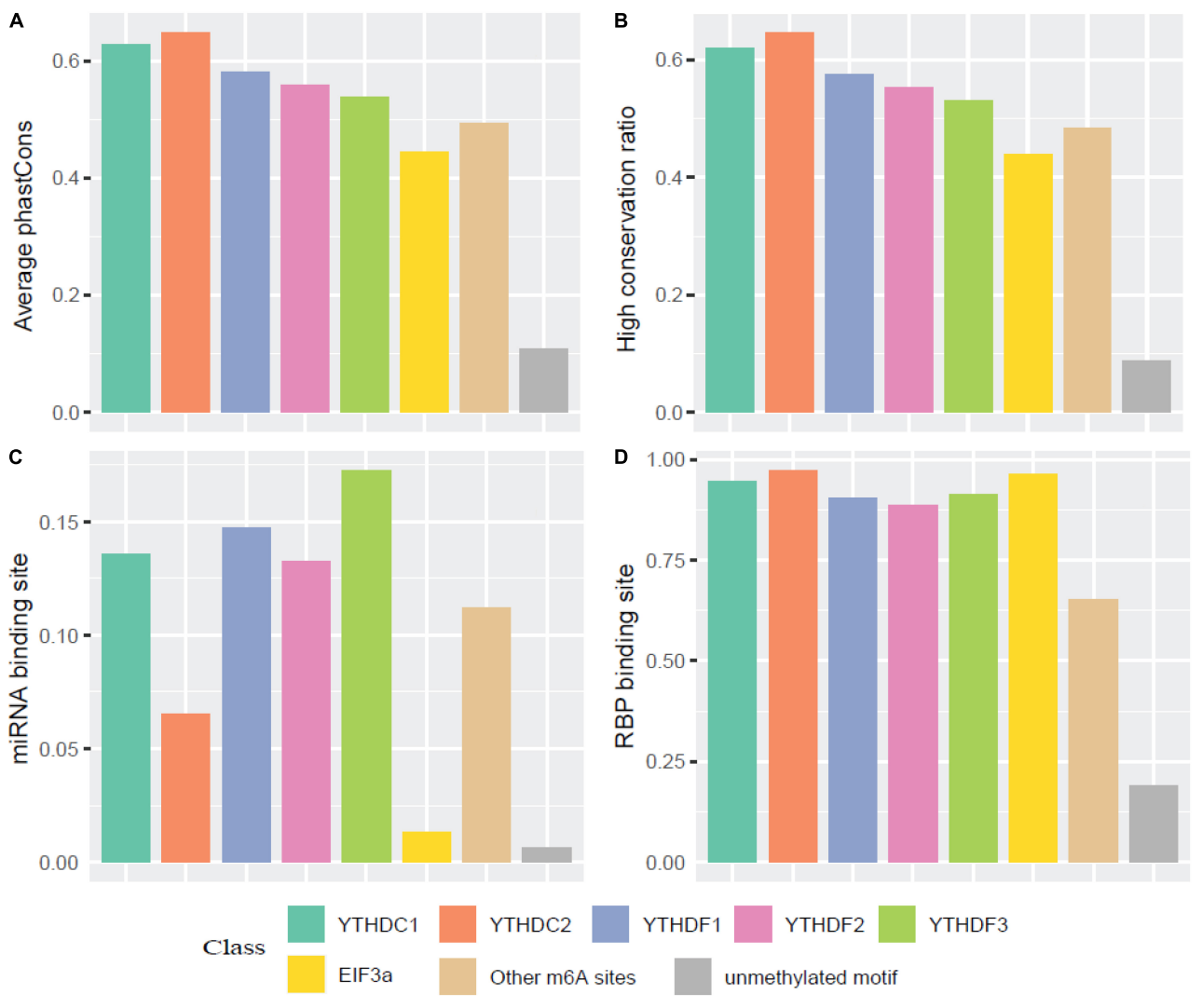


FIGURE 4 | Conservation analysis of the m^{6}A sites targeted by different readers. **(A)** Average phastCons; **(B)** High conservation ratio; **(C)** Frequency of miRNA binding site among the targets of six m^{6}A readers; **(D)** Frequency of RBP binding site among the targets of six m^{6}A readers.

with 5′ to 3′ exoribonuclease XRN1 mediated by two Ankyrin repeats (ANK) on YTHDC2 (Wojtas et al., 2017). Remarkably, YTHDF family shows a similar binding distribution in CDSs and 3′ UTRs with peaks at around stop codons of mRNAs. A similar pattern of results was obtained in previous studies suggesting that YTHDFs directly interplay among one another to collaboratively regulate translation and decay of targeted mRNAs in the cytoplasm (Shi et al., 2017). The binding sites of EIF3A are uniquely enriched at 5′UTRs. This is directly in line with previous findings that the HLH motif of EIF3A interacts predominantly with the m^{6}A residues on the 5′UTR, and EIF3A specifically functions to promote cap-independent translation under diverse cellular stresses.

We then compared the conservation of all m^{6}A readers by phastCons score and high conservation ratio (>0.5). As seen in **Figures 4A,B**, the m^{6}A sites (targeted or not targeted by the studied six readers) are more conservative than unmethylated m^{6}A motifs (DRACH). This suggests that m^{6}A sites and the m^{6}A reader binding sites are more evolutionarily conserved at the gene level, and the occurrence of m^{6}A should be considered of functional importance and maintained under selection pressure. Moreover, the YTH family is more conserved compared with other regulation components, which is similar to the finding that YT521-B homology (YTH) RNA-binding domain in eukaryotes is known to be highly conserved with essential Lys-364, Trp-380, and Arg-478 (Zhang et al., 2010). Additionally, as shown in

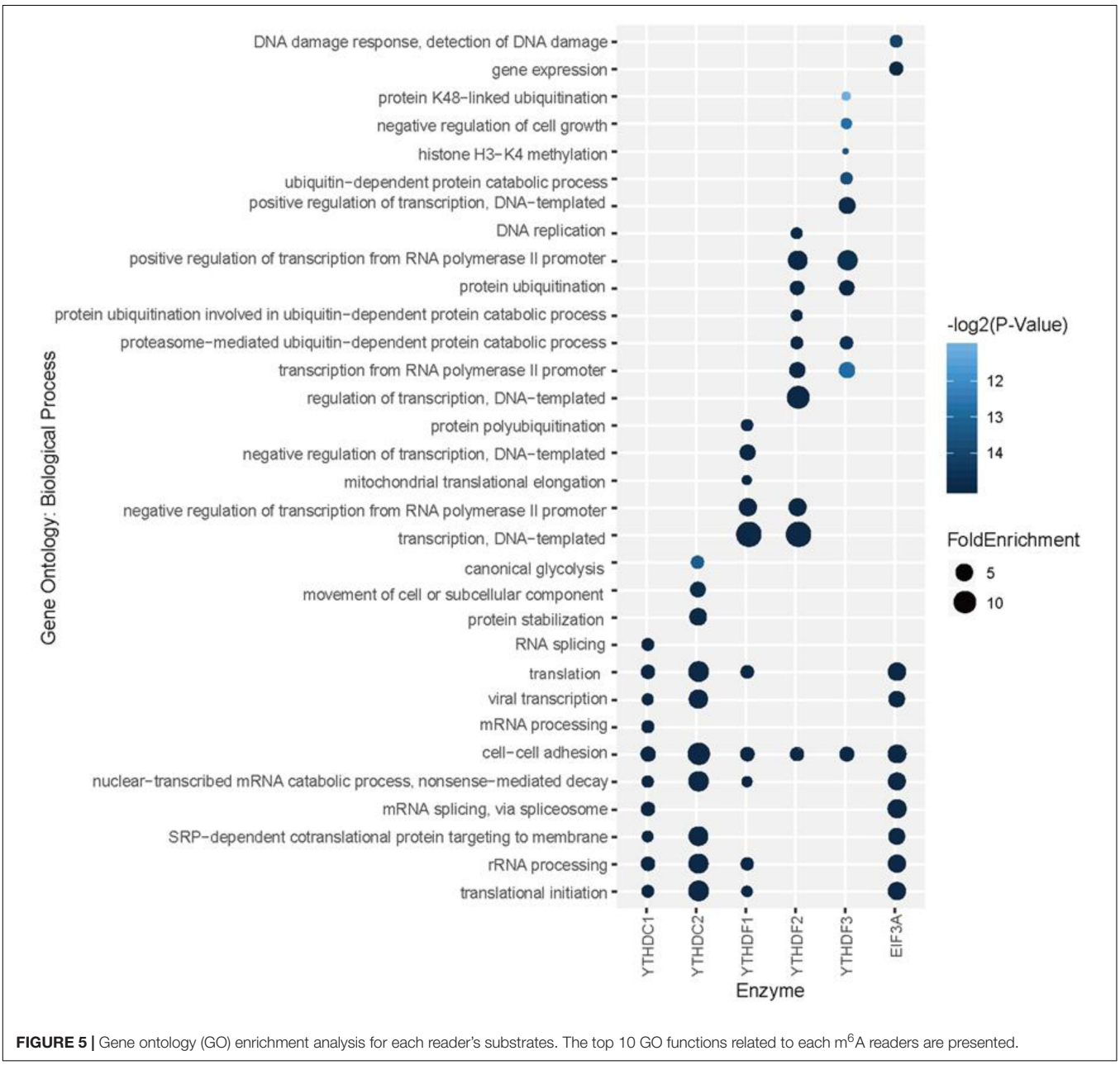


FIGURE 5 | Gene ontology (GO) enrichment analysis for each reader's substrates. The top 10 GO functions related to each m^{6}A readers are presented.

Figure 4C, compared with EIF3a binding sites and unmethylated sites which are mostly not in 3′ UTR, targets of other m^{6}A readers and other untargeted m^{6}A sites are more correlated with the miRNA binding sites. This result agrees well with existing studies investigated that miRNA targets are more enriched in 3′ UTR and m^{6}A peaks prior to the present of miRNA binding for a majority of the time, suggesting that m^{6}A modification functions to enhance initiation of miRNA biogenesis (Meyer et al., 2012; Alarcón et al., 2015). And the relative low overlapping rate between YTHDC2 binding sites and miRNA binding sites could be explained by multiple RNA-binding domains of YTHDC2. Furthermore, the proportions of overlapping of RNA-binding proteins (RBPs) and each m^{6}A reader's binding site are calculated. **Figure 4D** shows that RBPs binding regions overlap with m^{6}A reader binding sites in mRNA more than the other m^{6}A sites, while there are even fewer overlapping regions with unmethylated sites. This is consistent with our knowledge that some RBPs are essential in post-transcriptional control of RNAs including splicing, stabilization, localization and translation of mRNA. In the process of regulating transcription and translation, m^{6}A readers may recruit large numbers of regulators or factors to their targeted RNAs so as to functionally regulate biological processes (Shi et al., 2017).

To explore the association among m^{6}A modification, readers and biological functions, the gene ontology (GO) enrichment analysis was conducted to measure the biological functions of substrates of each reader using DAVID websites (Huang da et al., 2009). The resulting top 10 GO functions related to each m^{6}A readers were illustrated in **Figure 5**. Interestingly, YTHDC1 is involved in mRNA splicing, mRNA processing and nuclear-transcribed mRNA catabolic process, which is consistent with our understanding of its role of mediating nuclear to cytoplasmic export of nascent m^{6}A-containing mRNAs (Roundtree et al., 2017). The targeting of YTHDC2, shown to accelerate the degradation of mRNA and enhance translation efficiency (Hsu et al., 2017), are more related to nonsense-mediated decay, protein stabilization and translational initiation. YTHDF1 targets are enriched under the GO terms of nuclear-transcribed mRNA catabolic process and translation initiation (Wang et al., 2015), suggesting its function in selectively recruiting of ribosomes and facilitating translation. YTHDF2 and YTHDF3 targets are both associated with proteasome-mediated ubiquitin-dependent protein catabolic process, which corresponds to our knowledge of their regulation in the metabolism of cytosolic m^{6}A-modified mRNAs (Wang et al., 2014; Shi et al., 2017). EIF3A, reported to serve as a driver of specialized translation (Lee et al., 2015), is enriched with gene expression, translation and SRP-dependent co-translational protein targeting to the membrane. Moreover, as summarized in **Supplementary Figure S7**, six m^{6}A readers show high enrichment in cytosol, cytoplasm, and membrane. Five of them (YTHDC1, YTHDF1-3, and EIF3a) are enriched in nucleus and nucleoplasm. While YTHDC2 is more enriched in extracellular exosome, extracellular matrix and myelin sheath instead of nucleus or nucleoplasm. All six proteins are enriched in the function of protein binding and poly(A) RNA binding, while they each have other specialized functions. This is consistent with analysis above on the enrichment of biological process and previous relevant literature. All gene ontology enrichment results were shown in **Supplementary Table S9**.

Additionally, we further confirmed the biological meanings of the substrates of all m^{6}A readers. Based on the results of previous GO enrichment analysis (Chen K. et al., 2018), the most significant *p*-values of top 10 terms treated with the negative logarithm were firstly added up, and then those computed results of identified targets were compared with those of randomly selected substrates. With the bootstrap sampling approach, substrates were randomly selected and analyzed for 100 times before the results were summarized as proportions and displayed in pie charts. Conceivably, if our results achieved on real data are more biologically meaningful than random permutation, it is possible that our analysis reliably unveiled the true biological functions. Specifically, there are 88, 100, 73, 68, 80, and 100% chances for each reader to be more enriched in biological functions than random permutation as illustrated in **Figure 6**, suggesting high possibility that our functional prediction for each individual reader is statistically meaningful.

Web Sever for m^{6}A Reader

A web server with a friendly graphical user interface (**Figure 7**) was constructed to properly share the predictive models we constructed for predicting target specificity of the m^{6}A readers. Users may upload the genome ranges in BED format to the website, and a notification email will be sent to the given email address once the job is finished.

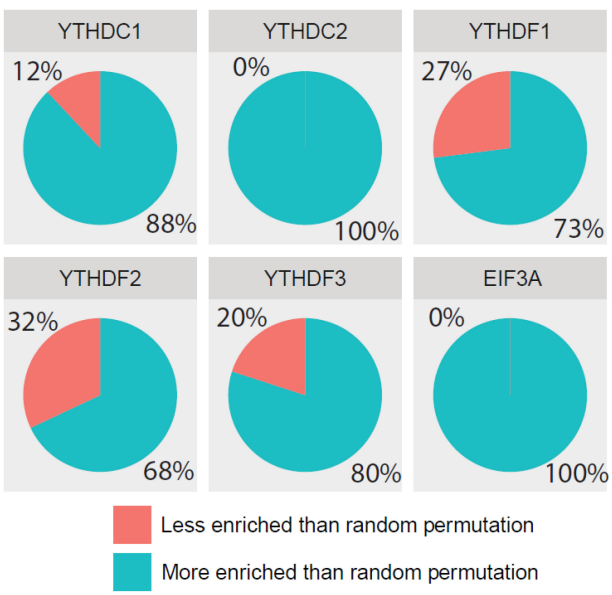


FIGURE 6 | Comparing detection of m^{6}A readers' targets based on biological significance. The most significant *p*-values of top 10 GO terms treated with negative logarithm were added up, and those results of identified targets were compared with those of randomly selected substrates. With the bootstrap sampling approach, substrates were randomly selected and analyzed for 100 times before the results were summarized as proportions and displayed in pie charts.

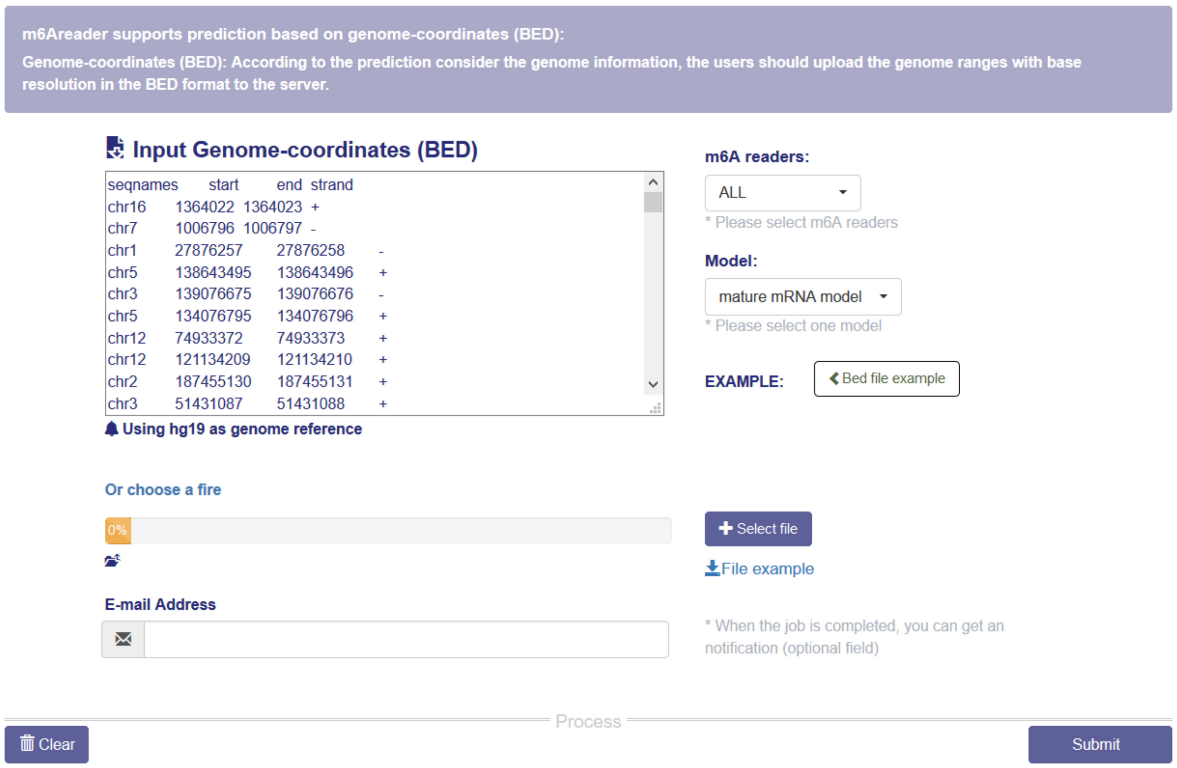


FIGURE 7 | m^6A reader web server. The web server takes genome ranges in BED format as the input, and supports prediction for the target sites of six m^6A readers (YTHDC1, YTHDC2, YTHDF1, YTHDF2 and YTHDF3 and EIF3A). All the materials used in the project, including the training data and codes, are also available on the website.

CONCLUSION

With the great breakthroughs made in RNA modification-mediated regulation of gene expression, studies of emerging transcriptome modifications have driven rapid development of the high-throughput sequencing technologies. With the aid of the invention of m^6A-seq (Dominissini et al., 2012) and MeRIP-seq (Meyer et al., 2012), transcriptome-wide profiling of m^6A is now possible. Based on comprehensive high-throughput sequencing data, MeT-DB (Liu H. et al., 2018) and RMBase (Xuan et al., 2018) were established, providing the site information of RNA modifications. Subsequently, single-based technologies such as m^6A-CLIP (Ke et al., 2015) and miCLIP (Linder et al., 2015) were also developed to precisely identify the positions of m^6A. Complementary to experimental methods, well-established computational models facilitate the analysis of sequencing data and address the challenges presented in the bioinformatics community by predicting potential RNA methylation sites. The exomePeak R/Bioconductor package (Meng et al., 2013, 2014), MACS algorithm (Zhang et al., 2008) and DRME software (Liu et al., 2016) were introduced to analyze epitranscriptome profiling data, which improved our understanding of RNA methylation. Sequence-based site prediction models such as iRNA(m^6A)-PseDNC (Chen W. et al., 2018) and iRNAMethyl (Chen et al., 2015b) applied statistical methods, whereas m^6Apred (Chen et al., 2015c), RAM-ESVM (Chen et al., 2017b), and RNAMethPre (Xiang et al., 2016) integrated machine learning approaches, predicting m^6A sites in different species' transcriptome. Furthermore, potential RNA methylation-disease associations have been revealed by m^6Avar (Zheng et al., 2018) and m^6ASNP (Jiang et al., 2018). With a similar purpose, heterogeneous networks have been used in DRUM (Tang et al., 2019), FunDMDeep-m^6A (Zhang et al., 2019b) and Deepm^6A (Zhang et al., 2019a), showing a new perspective in studying disease-associated RNA methylation.

In this study, we constructed SVM-based models for the prediction of target specificity of m^6A readers (YTHDC1, YTHDC2, YTHDF1, YTHDF2, YTHDF3, and EIF3A). The proposed models rely on 58 genomic features integrated with the sequence features related to chemical properties. After feature selection using the *F*-score method, those models achieved high prediction performance in 5-fold cross-validation and independent testing. Additionally, we compared the performance of different sequence encoding schemes on each reader's substrate prediction. As existing m^6A base-resolution data suffer from the bias of polyA selection, mature mRNA model was also considered besides the full transcript model. Moreover, we compared different machine learning algorithms and showed that four algorithms all demonstrate high performance with little difference in the prediction of target specificity of m^6A readers. We eventually decided to use SVM classifier for our predictors.

It is also worth mentioning that our comprehensive analysis of m^6A readers revealed potential regulatory patterns and biological relationships. We showed that m^6A reader binding sites on mRNAs were concentrated in CDSs and 3′ UTR near stop codons, which is in line with m^6A localization. Although distribution analysis of m^6A readers has been conducted in previous studies and suggested similar binding patterns (Xu et al., 2014; Wang et al., 2015; Hsu et al., 2017), the results we presented were substantially enhanced with the incorporation of multiple datasets. Our result shed lights on the post-transcriptional and translational functions of m^6A readers on m^6A-containing mRNAs with more reliable evidence. Moreover, computed phastCons score and conservation ratio revealed a high conservation of the target sites of m^6A readers, suggesting that they are possibly playing necessary or essential roles in regulating m^6A-containing mRNAs. This is remarkable since we focused on the conservation of binding sites of m^6A readers on mRNAs, rather than the conservation of m^6A motifs itself as widely studied currently (Meyer et al., 2012), thus the biologically meaningful relationship between m^6A readers and m^6A modifications was confirmed. Besides, different from enrichment analysis alone in previous studies (Hsu et al., 2017), we not only unveiled functional relevance through the enrichment of the targets of m^6A readers in biological process, cellular components and molecular functions by GO analysis, but also confirmed that reader-regulated sites are more likely to be biologically significant than randomly selected sites. The combination of statistical analysis and GO analysis ensures the robust detection and critical evaluation of the biological functions with a higher degree of confidence. Furthermore, our GO enrichment analysis result is also consistent with the wet-lab experiment and our prediction on unidentified regions that YTHDF proteins have both similar functions and different functions in the m^6A mRNA regulation. This supports the conclusion made in previous study that m^6A readers' effect on downstream processes are much more heterogeneous and context-dependent across transcripts (Zhang et al., 2020).

However, this study has a number of limitations that could be improved in the future. Firstly, it has been argued that 4SU PAR-CLIP suffers from U-bias in contrast with UV-254 crosslinking or 6SG crosslinking (Ascano et al., 2012), thus other CLIP techniques are recommended to ensure crosslinking efficiency. Secondly, although data from different experiments were combined to build the predictors and 5-fold cross-validation was used to balance the bias-variance tradeoff, data of YTHDC2 and EIF3A substrates are still limited, which may make overfitting of the models possible. Thus, the analysis and prediction will benefit from other data from wet experiments in the future. Thirdly, as genome-derived features improved the performance of predictors dramatically, this suggests that genomic features carry important characteristics of biological data. Considering only 58 of them were involved in the feature selection procedure, it is worth exploring more genomic features so as to allow more effective features to be selected and reduce the bias as much as possible. In the future, it is expected to see the expanded studies of the enzyme target specificity and functional associations of other RNA modifications, such as m^1A and Pseudouridine, on other types of RNAs, such as lncRNA and snRNAs, and in other species, such as mouse and yeast. Additional studies are clearly needed to investigate RNA-sequence-dependent m^6A readers other than YTH domain-containing proteins such as FMR1 (Edupuganti et al., 2017). And it could be quite interesting to explore disease-associated RNA modification based on cellular binding patterns of regulatory proteins on modified RNAs.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

KC conceived the idea, initialized the project, collected and processed the training and benchmark datasets. ZW generated the genomic features. DZ, YW, YZ, and HX built machine learning models. YT and BS designed and built the web server. DZ, YW, and YZ drafted the manuscript. All authors read, critically revised, and approved the final manuscript.

FUNDING

This work has been supported by the National Natural Science Foundation of China (31671373) and XJTLU Key Program Special Fund (KSF-T-01). This work was partially supported by the AI University Research Centre through XJTLU Key Programme Special Fund (KSF-P-02).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2020.00741/full#supplementary-material

REFERENCES

Adams, J. M., and Cory, S. (1975). Modified nucleosides and bizarre 5′-termini in mouse myeloma mRNA. *Nature* 255, 28–33. doi: 10.1038/255028a0

Alarcón, C. R., Lee, H., Goodarzi, H., Halberg, N., and Tavazoie, S. F. (2015). N6-methyladenosine marks primary microRNAs for processing. *Nature* 519, 482–485. doi: 10.1038/nature14281

Ascano, M., Hafner, M., Cekan, P., Gerstberger, S., and Tuschl, T. (2012). Identification of RNA-protein interaction networks using PAR-CLIP. *Wiley Interdiscip. Rev.* 3, 159–177. doi: 10.1002/wrna.1103

Bari, A. T. M. G., Reaz, M. R., Choi, H.-J., and Jeong, B.-S. (2013). *DNA Encoding for Splice Site Prediction in Large DNA Sequence*. Berlin: Springer, 46–58.

Boccaletto, P., Machnicka, M. A., Purta, E., Piatkowski, P., Baginski, B., Wirecki, T. K., et al. (2018). MODOMICS: a database of RNA modification

pathways. 2017 update. *Nucleic Acids Res.* 46, D303–D307. doi: 10.1093/nar/gkx1030

Boulias, K., Toczydlowska-Socha, D., Hawley, B. R., Liberman, N., Takashima, K., Zaccara, S., et al. (2019). Identification of the m(6)Am Methyltransferase PCIF1 reveals the location and functions of m(6)Am in the Transcriptome. *Mol. Cell* 75, 631.e8–643.e8. doi: 10.1016/j.molcel.2019.06.006

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn.* 30, 1145–1159. doi: 10.1016/s0031-3203(96)00142-2

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 1–27. doi: 10.1145/1961189.1961199

Chen, J., Zhang, Y. C., Huang, C., Shen, H., Sun, B., Cheng, X., et al. (2019). m(6)A regulates neurogenesis and neuronal development by modulating histone methyltransferase Ezh2. *Genom. Proteom. Bioin.* 17, 154–168. doi: 10.1016/j.gpb.2018.12.007

Chen, K., Lu, Z., Wang, X., Fu, Y., Luo, G.-Z., Liu, N., et al. (2015a). High-resolution N(6) -methyladenosine (m(6) A) map using photo-crosslinking-assisted m(6) A sequencing. *Angew. Chem. Int. Ed. Engl.* 54, 1587–1590. doi: 10.1002/anie.201410647

Chen, W., Feng, P., Ding, H., Lin, H., and Chou, K.-C. (2015b). iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.* 490, 26–33. doi: 10.1016/j.ab.2015.08.021

Chen, W., Tran, H., Liang, Z., Lin, H., and Zhang, L. (2015c). Identification and analysis of the N6-methyladenosine in the *Saccharomyces cerevisiae* transcriptome. *Anal. Biochem.* 5:13859. doi: 10.1038/srep13859

Chen, K., Wei, Z., Liu, H., de Magalhaes, J. P., Rong, R., Lu, Z., et al. (2018). Enhancing epitranscriptome module detection from m(6)A-Seq data using threshold-based measurement weighting strategy. *BioMed Res. Int.* 2018:2075173. doi: 10.1155/2018/2075173

Chen, W., Ding, H., Zhou, X., Lin, H., and Chou, K.-C. (2018). iRNA(m^6A)-PseDNC: Identifying N6-methyladenosine sites using pseudo dinucleotide composition. *Anal. Biochem.* 561-562, 59–65. doi: 10.1016/j.ab.2018.09.002

Chen, K., Wei, Z., Zhang, Q., Wu, X., Rong, R., Lu, Z., et al. (2019). WHISTLE: a high-accuracy map of the human N6-methyladenosine (m^6A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res.* 47:e41. doi: 10.1093/nar/gkz074

Chen, W., Feng, P., Tang, H., Ding, H., and Lin, H. (2016a). Identifying 2′-O-methylationation sites by integrating nucleotide chemical properties and nucleotide compositions. *Genomics* 107, 255–258. doi: 10.1016/j.ygeno.2016.05.003

Chen, W., Tang, H., Ye, J., Lin, H., and Chou, K. C. (2016b). iRNA-PseU: Identifying RNA pseudouridine sites. *Mol. Ther. Nucleic Acids* 5:e332. doi: 10.1038/mtna.2016.37

Chen, W., Tang, H., and Lin, H. (2017a). MethyRNA: a web server for identification of N(6)-methyladenosine sites. *J. Biomol. Struct. Dyn.* 35, 683–687. doi: 10.1080/07391102.2016.1157761

Chen, W., Xing, P., and Zou, Q. (2017b). Detecting N6-methyladenosine sites from RNA transcriptomes using ensemble support vector machines. *Sci. Rep.* 7:40242. doi: 10.1038/srep40242

Chen, Y.-W., and Lin, C.-J. (2006). "Combining SVMs with Various Feature Selection Strategies," in *Feature Extraction: Foundations and Applications*, eds I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh (Berlin: Springe), 315–324. doi: 10.1007/978-3-540-35488-8_13

Chen, Z., Zhao, P., Li, F., Marquez-Lago, T. T., Leier, A., Revote, J., et al. (2019). iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief Bioinform.* 21, 1047–1057. doi: 10.1093/bib/bbz041

Chen, Z., Zhao, P., Li, F., Marquez-Lago, T. T., Leier, A., Revote, J., et al. (2020). iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief Bioinform.* 21, 1047–1057. doi: 10.1093/bib/bbz041

Cui, X., Wei, Z., Zhang, L., Liu, H., Sun, L., Zhang, S. W., et al. (2016). Guitar: An R/Bioconductor Package for gene annotation guided transcriptomic analysis of RNA-related genomic features. *BioMed Res. Int.* 2016:8367534. doi: 10.1155/2016/8367534

Dao, F. Y., Lv, H., Wang, F., Feng, C. Q., Ding, H., Chen, W., et al. (2019). Identify origin of replication in *Saccharomyces cerevisiae* using two-step feature selection technique. *Bioinformatics* 35, 2075–2083. doi: 10.1093/bioinformatics/bty943

Desrosiers, R., Friderici, K., and Rottman, F. (1974). Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. *Proc. Natl Acade. Sci. U.S.A.* 71, 3971–3975. doi: 10.1073/pnas.71.10.3971

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., et al. (2012). Topology of the human and mouse m^6A RNA methylomes revealed by m^6A-seq. *Nature* 485, 201–206. doi: 10.1038/nature11112

Du, H., Zhao, Y., He, J., Zhang, Y., Xi, H., Liu, M., et al. (2016). YTHDF2 destabilizes m(6)A-containing RNA through direct recruitment of the CCR4-NOT deadenylase complex. *Nat. Commun.* 7:12626. doi: 10.1038/ncomms12626

Edupuganti, R. R., Geiger, S., Lindeboom, R. G. H., Shi, H., Hsu, P. J., Lu, Z., et al. (2017). N6-methyladenosine (m^6A) recruits and repels proteins to regulate mRNA homeostasis. *Nat. Struct. Mol. Biol.* 24, 870–878. doi: 10.1038/nsmb.3462

Engel, M., Eggert, C., Kaplick, P. M., Eder, M., Roh, S., Tietze, L., et al. (2018). The Role of m(6)A/m-RNA methylation in stress response regulation. *Neuron* 99, 389.e9–403.e9. doi: 10.1016/j.neuron.2018.07.009

Garcia-Campos, M. A., Edelheit, S., Toth, U., Safra, M., Shachar, R., Viukov, S., et al. (2019). Deciphering the m(6)A code via antibody-independent quantitative profiling. *Cell* 178, 731e16–747e16.

Gokhale, N. S., McIntyre, A. B. R., McFadden, M. J., Roder, A. E., Kennedy, E. M., Gandara, J. A., et al. (2016). N6-methyladenosine in flaviviridae Viral RNA genomes regulates infection. *Cell Host Microbe* 20, 654–665. doi: 10.1016/j.chom.2016.09.015

Gulko, B., Hubisz, M. J., Gronau, I., and Siepel, A. (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. (in English). *Nat. Genet.* 47, 276–283. doi: 10.1038/ng.3196

Hazra, D., Chapat, C., and Graille, M. (2019). m(6)A mRNA destiny: chained to the rhYTHm by the YTH-Containing Proteins. *Genes* 10:49. doi: 10.3390/genes10010049

He, J., Fang, T., Zhang, Z., Huang, B., Zhu, X., and Xiong, Y. (2018). PseUI: pseudouridine sites identification based on RNA sequence information. *BMC Bioinformatics* 19:306. doi: 10.1186/s12859-018-2321-0

He, W., Jia, C., and Zou, Q. (2019). 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. *Bioinformatics* 35, 593–601. doi: 10.1093/bioinformatics/bty668

Hsu, P. J., Zhu, Y., Ma, H., Guo, Y., Shi, X., Liu, Y., et al. (2017). Ythdc2 is an N(6)-methyladenosine binding protein that regulates mammalian spermatogenesis. *Cell Res.* 27, 1115–1127. doi: 10.1038/cr.2017.99

Huang, H., Weng, H., Zhou, K., Wu, T., Zhao, B. S., Sun, M., et al. (2019). Histone H3 trimethylation at lysine 36 guides m(6)A RNA modification co-transcriptionally. *Nature* 567, 414–419. doi: 10.1038/s41586-019-1016-7

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, 1669–1677. doi: 10.7150/ijbs.27819

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Jiang, S., Xie, Y., He, Z., Zhang, Y., Zhao, Y., Chen, L., et al. (2018). m^6ASNP: a tool for annotating genetic variants by m^6A function. *GigaScience* 7:giy035. doi: 10.1093/gigascience/giy035

Jin, H., and Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans. on Knowl. Data Eng.* 17, 299–310. doi: 10.1109/tkde.2005.50

Ke, S., Alemu, E. A., Mertens, C., Gantman, E. C., Fak, J. J., Mele, A., et al. (2015). A majority of m^6A residues are in the last exons, allowing the potential for 3′ UTR regulation. *Genes Dev.* 29, 2037–2053. doi: 10.1101/gad.269415.115

Ke, S., Pandya-Jones, A., Saito, Y., Fak, J. J., Vågbø, C. B., Geula, S., et al. (2017). m(6)A mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover. *Genes Dev.* 31, 990–1006. doi: 10.1101/gad.301036.117

Keilwagen, J., Grosse, I., and Grau, J. (2014). Area under precision-recall curves for weighted and unweighted data. *PLoS One* 9:e92209. doi: 10.1371/journal.pone.0092209

Kmietczyk, V., Riechert, E., Kalinski, L., Boileau, E., Malovrh, E., Malone, B., et al. (2019). m(6)A-mRNA methylation regulates cardiac gene expression and cellular growth. *Life Sci. Allian.* 2:e201800233. doi: 10.26508/lsa.201800233

Kuhn, M. (2020). *caret: Classification and Regression Training. R package version 6.0-85*. Available: https://CRAN.R-project.org/package=caret (accessed March 20, 2020).

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., et al. (2013). Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* 9:e1003118. doi: 10.1371/journal.pcbi.1003118

Lee, A. S., Kranzusch, P. J., and Cate, J. H. (2015). eIF3 targets cell-proliferation messenger RNAs for translational activation or repression. *Nature* 522, 111–114. doi: 10.1038/nature14267

Lee, D., Karchin, R., and Beer, M. A. (2011). Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.* 21, 2167–2180. doi: 10.1101/gr.121905.111

Li, J., Huang, Y., Yang, X., Zhou, Y., and Zhou, Y. (2018). RNAm5Cfinder: a Web-server for Predicting RNA 5-methylcytosine (m5C) Sites Based on Random Forest. *Sci. Rep.* 8:17299. doi: 10.1038/s41598-018-35502-4

Liao, S., Sun, H., Xu, C., and Domain, Y. T. H. (2018). A family of N(6)-methyladenosine (m(6)A) Readers. *Genom. Proteom. Bioinf.* 16, 99–107. doi: 10.1016/j.gpb.2018.04.002

Lin, H., Deng, E. Z., Ding, H., Chen, W., and Chou, K. C. (2014). iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.* 42, 12961–12972. doi: 10.1093/nar/gku1019

Linder, B., Grozhik, A. V., Olarerin-George, A. O., Meydan, C., Mason, C. E., and Jaffrey, S. R. (2015). Single-nucleotide-resolution mapping of m^6A and m^6Am throughout the transcriptome. *Nat. Methods* 12, 767–772. doi: 10.1038/nmeth.3453

Liu, B. (2019). BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief Bioinform.* 20, 1280–1294. doi: 10.1093/bib/bbx165

Liu, B., Gao, X., and Zhang, H. (2019). BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.* 47:e127. doi: 10.1093/nar/gkz740

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K. C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* 43, W65–W71. doi: 10.1093/nar/gkv458

Liu, B., Merriman, D. K., Choi, S. H., Schumacher, M. A., Plangger, R., Kreutz, C., et al. (2018). A potentially abundant junctional RNA motif stabilized by m(6)A and Mg(2). *Nat. Communi.* 9:2761. doi: 10.1038/s41467-018-05243-z

Liu, H., Wang, H., Wei, Z., Zhang, S., Hua, G., Zhang, S. W., et al. (2018). MeT-DB V2.0: elucidating context-specific functions of N6-methyl-adenosine methyltranscriptome. *Nucleic Acids Res.* 46, D281–D287. doi: 10.1093/nar/gkx1080

Liu, L., Zhang, S.-W., Gao, F., Zhang, Y., Huang, Y., Chen, R., et al. (2016). DRME: count-based differential RNA methylation analysis at small sample size scenario. *Anal. Biochem.* 499, 15–23. doi: 10.1016/j.ab.2016.01.014

Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., et al. (2011). ViennaRNA Package 2.0. *Algorithms Mol. Biol.* 6:26.

Meng, J., Cui, X., Rao, M. K., Chen, Y., and Huang, Y. (2013). Exome-based analysis for RNA epigenome sequencing data. *Bioinformatics* 29, 1565–1567. doi: 10.1093/bioinformatics/btt171

Meng, J., Lu, Z., Liu, H., Zhang, L., Zhang, S., Chen, Y., et al. (2014). A protocol for RNA methylation differential analysis with MeRIP-Seq data and exomePeak R/Bioconductor package. *Methods* 69, 274–281. doi: 10.1016/j.ymeth.2014.06.008

Meyer, K. D., and Jaffrey, S. R. (2017). Rethinking m(6)A Readers, Writers, and Erasers. *Annu. Rev. Cell Dev. Biol.* 33, 319–342. doi: 10.1146/annurev-cellbio-100616-060758

Meyer, K. D., Patil, D. P., Zhou, J., Zinoviev, A., Skabkin, M. A., Elemento, O., et al. (2015). 5′ UTR m(6)A promotes cap-independent translation. *Cell* 163, 999–1010. doi: 10.1016/j.cell.2015.10.012

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell* 149, 1635–1646. doi: 10.1016/j.cell.2012.05.003

Patil, D. P., Chen, C.-K., Pickering, B. F., Chow, A., Jackson, C., Guttman, M., et al. (2016). m(6)A RNA methylation promotes XIST-mediated transcriptional repression. *Nature* 537, 369–373. doi: 10.1038/nature19342

Patil, D. P., Pickering, B. F., and Jaffrey, S. R. (2018). Reading m(6)A in the Transcriptome: m(6)A-Binding Proteins. *Trends Cell Biol.* 28, 113–127. doi: 10.1016/j.tcb.2017.10.001

Powers, D. (2008). Evaluation: from precision, recall and F-Factor to ROC, informedness, markedness & correlation. *Mach. Learn. Technol.* 2, 37–63.

Roundtree, I. A., Luo, G. Z., Zhang, Z., Wang, X., Zhou, T., Cui, Y., et al. (2017). YTHDC1 mediates nuclear export of N(6)-methyladenosine methylated mRNAs. *eLife* 6:e31311. doi: 10.7554/eLife.31311

Schwartz, S., Mumbach, M. R., Jovanovic, M., Wang, T., Maciag, K., Bushkin, G. G., et al. (2014). Perturbation of m^6A writers reveals two distinct classes of mRNA methylation at internal and 5′ sites. *Cell Re.* 8, 284–296. doi: 10.1016/j.celrep.2014.05.048

Shi, H., Wang, X., Lu, Z., Zhao, B. S., Ma, H., Hsu, P. J., et al. (2017). YTHDF3 facilitates translation and decay of N(6)-methyladenosine-modified RNA. *Cell Res.* 27, 315–328. doi: 10.1038/cr.2017.15

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050. doi: 10.1101/gr.3715005

Song, Y., Xu, Q., Wei, Z., Zhen, D., Su, J., Chen, K., et al. (2019). Predict epitranscriptome targets and regulatory functions of N (6)-Methyladenosine (m(6)A) Writers and Erasers. *Evolu. Bioinf. Online* 15:1176934319871290. doi: 10.1177/1176934319871290

Tang, Y., Chen, K., Wu, X., Wei, Z., Zhang, S.-Y., Song, B., et al. (2019). DRUM: inference of disease-associated m(6)A RNA methylation sites from a multi-layer heterogeneous network. *Front. Genet* .10:266. doi: 10.3389/fgene.2019.00266

van Tran, N., Ernst, F. G. M., Hawley, B. R., Zorbas, C., Ulryck, N., Hackert, P., et al. (2019). The human 18S rRNA m^6A methyltransferase METTL5 is stabilized by TRMT112. *Nucleic Acids Res.* 47, 7719–7733. doi: 10.1093/nar/gkz619

Vu, L. P., Pickering, B. F., Cheng, Y., Zaccara, S., Nguyen, D., Minuesa, G., et al. (2017). The N(6)-methyladenosine (m(6)A)-forming enzyme METTL3 controls myeloid differentiation of normal hematopoietic and leukemia cells. *Nature Med.* 23, 1369–1376. doi: 10.1038/nm.4416

Wang, X., Lu, Z., Gomez, A., Hon, G. C., Yue, Y., Han, D., et al. (2014). N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* 505, 117–120. doi: 10.1038/nature12730

Wang, X., Zhao, B. S., Roundtree, I. A., Lu, Z., Han, D., Ma, H., et al. (2015). N(6)-methyladenosine modulates messenger RNA translation efficiency. *Cell* 161, 1388–1399. doi: 10.1016/j.cell.2015.05.014

Wojtas, M. N., Pandey, R. R., Mendel, M., Homolka, D., Sachidanandam, R., and Pillai, R. S. (2017). Regulation of m(6)A Transcripts by the 3′–>5′ RNA Helicase YTHDC2 Is Essential for a Successful Meiotic Program in the Mammalian Germline. *Mol. Cell* 68, 374.e12–387.e12. doi: 10.1016/j.molcel.2017.09.021

Wu, H., Xu, T., Feng, H., Chen, L., Li, B., Yao, B., et al. (2015). Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Res.* 43:e141.

Xiang, S., Liu, K., Yan, Z., Zhang, Y., and Sun, Z. (2016). RNAMethPre: A Web Server for the Prediction and Query of mRNA m^6A Sites. *PLoS One* 11:e0162707. doi: 10.1371/journal.pone.0162707

Xiao, W., Adhikari, S., Dahal, U., Chen, Y. S., Hao, Y. J., Sun, B. F., et al. (2016). Nuclear m(6)A Reader YTHDC1 Regulates mRNA Splicing. *Mol. Cell* 61, 507–519. doi: 10.1016/j.molcel.2016.01.012

Xu, C., Wang, X., Liu, K., Roundtree, I. A., Tempel, W., Li, Y., et al. (2014). Structural basis for selective binding of m^6A RNA by the YTHDC1 YTH domain. *Nat. Chem. Biol.* 10, 927–929. doi: 10.1038/nchembio.1654

Xuan, J. J., Sun, W. J., Lin, P. H., Zhou, K. R., Liu, S., Zheng, L. L., et al. (2018). RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. *Nucleic Acids Res.* 46, D327–D334. doi: 10.1093/nar/gkx934

Ye, F., Chen, E. R., and Nilsen, T. W. (2017). Kaposi's sarcoma-associated herpesvirus utilizes and manipulates RNA N(6)-Adenosine methylation to promote lytic replication. *J. Virol.* 91:e00466-17. doi: 10.1128/jvi.00466-17

Zaccara, S., and Jaffrey, S. R. (2020). A unified model for the function of YTHDF proteins in regulating m(6)A-Modified mRNA. *Cell* 181, 1582.e18–1595.e18. doi: 10.1016/j.cell.2020.05.012

Zhang, S.-Y., Zhang, S.-W., Fan, X.-N., Meng, J., Chen, Y., Gao, S.-J., et al. (2019a). Global analysis of N6-methyladenosine functions and its disease association using deep learning and network-based methods. *PLoS Comput. Biol.* 15:e1006663. doi: 10.1371/journal.pcbi.1006663

Zhang, S.-Y., Zhang, S.-W., Fan, X.-N., Zhang, T., Meng, J., and Huang, Y. (2019b). FunDMDeep-m^{6}A: identification and prioritization of functional differential m^{6}A methylation genes. *Bioinformatics* 35, i90–i98. doi: 10.1093/bioinformatics/btz316

Zhang, Z., Chen, L. Q., Zhao, Y. L., Yang, C. G., Roundtree, I. A., Zhang, Z., et al. (2019c). Single-base mapping of m(6)A by an antibody-independent method. *Sci. Adva.* 5:eaax0250. doi: 10.1126/sciadv.aax0250

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biolo.* 9:R137. doi: 10.1186/gb-2008-9-9-r137

Zhang, Z., Theler, D., Kaminska, K. H., Hiller, M., de la Grange, P., Pudimat, R., et al. (2010). The YTH domain is a novel RNA binding domain. *J. Biol. Chem.* 285, 14701–14710. doi: 10.1074/jbc.M110.104711

Zhang, Z., Luo, K., Zou, Z., Qiu, M., Tian, J., Sieh, L., et al. (2020). Genetic analyses support the contribution of mRNA N6-methyladenosine (m^{6}A) modification to human disease heritability. *Nat. Genet.* doi: 10.1038/s41588-020-0644-z [Epub ahead of print].

Zheng, Y., Nie, P., Peng, D., He, Z., Liu, M., Xie, Y., et al. (2018). m^{6}AVar: a database of functional variants involved in m^{6}A modification. *Nucleic Acids Res.* 46, D139–D145. doi: 10.1093/nar/gkx895

Zhou, Y., Zeng, P., Li, Y. H., Zhang, Z., and Cui, Q. (2016). SRAMP: prediction of mammalian N6-methyladenosine (m^{6}A) sites based on sequence-derived features. *Nucleic Acids Res.* 44:e91. doi: 10.1093/nar/gkw104

Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian N (6)-methyladenosine sites from mRNA. *RNA* 25, 205–218. doi: 10.1261/rna.069112.118

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.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 09 September 2020
doi: 10.3389/fcell.2020.580217

DeepKhib: A Deep-Learning Framework for Lysine 2-Hydroxyisobutyrylation Sites Prediction

Luna Zhang[1†], Yang Zou[2†], Ningning He[2], Yu Chen[1], Zhen Chen[3,4] and Lei Li[1,2*]*

[1] School of Data Science and Software Engineering, Qingdao University, Qingdao, China, [2] School of Basic Medicine, Qingdao University, Qingdao, China, [3] Collaborative Innovation Center of Henan Grain Crops, Henan Agricultural University, Zhengzhou, China, [4] Key Laboratory of Rice Biology in Henan Province, Henan Agricultural University, Zhengzhou, China

As a novel type of post-translational modification, lysine 2-Hydroxyisobutyrylation (K_{hib}) plays an important role in gene transcription and signal transduction. In order to understand its regulatory mechanism, the essential step is the recognition of K_{hib} sites. Thousands of K_{hib} sites have been experimentally verified across five different species. However, there are only a couple traditional machine-learning algorithms developed to predict K_{hib} sites for limited species, lacking a general prediction algorithm. We constructed a deep-learning algorithm based on convolutional neural network with the one-hot encoding approach, dubbed CNN_{OH}. It performs favorably to the traditional machine-learning models and other deep-learning models across different species, in terms of cross-validation and independent test. The area under the ROC curve (AUC) values for CNN_{OH} ranged from 0.82 to 0.87 for different organisms, which is superior to the currently available K_{hib} predictors. Moreover, we developed the general model based on the integrated data from multiple species and it showed great universality and effectiveness with the AUC values in the range of 0.79–0.87. Accordingly, we constructed the on-line prediction tool dubbed DeepKhib for easily identifying K_{hib} sites, which includes both species-specific and general models. DeepKhib is available at http://www.bioinfogo.org/DeepKhib.

Keywords: post-translational modification, lysine 2-hydroxyisobutyrylation, deep learning, modification site prediction, machine learning

OPEN ACCESS

Edited by:
Jian Ren,
Sun Yat-sen University, China

Reviewed by:
Santosh Panjikar,
Australian Synchrotron, Australia
Annamaria Tonazzi,
National Research Council (CNR), Italy

***Correspondence:**
Zhen Chen
chenzhen-win2009@163.com
Lei Li
leili@qdu.edu.cn;
lileime@hotmail.com

[†] 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: *05 July 2020*
Accepted: *17 August 2020*
Published: *09 September 2020*

Citation:
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

INTRODUCTION

Protein post-translational modification (PTM) is a key mechanism to regulate cellular functions through covalent modification and enzyme modification, which dynamically regulates a variety of biological events (Beltrao et al., 2013; Skelly et al., 2016). Recently, an evolutionarily conserved short-chain lysine acylation modification dubbed lysine 2-hydroxyisobutylation (K_{hib}) has been reported, which introduces a steric bulk with a mass shift of +86.03 Da (**Supplementary Figure 1A**) and neutralize the positive charge of lysine (Dai et al., 2014; Xiao et al., 2015). It involves various biological functions including biosynthesis of amino acids, starch biosynthesis, carbon metabolism, glycolysis / gluconeogenesis and transcription (Dai et al., 2014; Huang et al., 2017, 2018a; Li et al., 2017; Meng et al., 2017; Yu et al., 2017; Wu et al., 2018; Yin et al., 2019). For instance, the decrease

of this modification on K281 of glycolytic enzyme ENO1 reduces its catalytic activity (Huang et al., 2018b). The three-dimension structure of the peptide containing K281 in the center was shown as **Supplementary Figure 1B**.

Thousands of K_{hib} sites have been identified in different species including humans, plants and prokaryotes through large-scale experimental approaches (Dai et al., 2014; Huang et al., 2018a), which is summarized in **Supplementary Table 1**. The experimental methods, however, are time-consuming and expensive and thus the development of prediction algorithms *in silico* is necessary for the high-throughput recognition of K_{hib} sites. Two classifiers (i.e., iLys-Khib and Khibpred) have been reported for predicting the K_{hib} sites in a few species (Ju and Wang, 2019; Wang et al., 2020). As many different organisms have been investigated and the number of K_{hib} sites has increased, it is indispensable to compare the characteristics of this modification in different species and investigate whether it is suitable to develop a general model with high confidence. Additionally, the reported models were based on traditional machine-learning (ML) algorithms (e.g., Random Forest (RF)). Recently, the deep learning (DL) algorithms, as the modern ML architecture, have demonstrated superior prediction performance in the field of bioinformatics, such as the prediction of modification sites on DNA, RNA and proteins (Wang et al., 2017; Huang et al., 2018c; Long et al., 2018; Tahir et al., 2019; Tian et al., 2019). We have developed a few DL approaches for the prediction of PTM sites and they all demonstrate their superiority over conventional ML algorithms (Chen et al., 2018a, 2019; Zhao et al., 2020). Therefore, we attempted to compare the DL models with the traditional ML models for the prediction of K_{hib} sites.

In this study, we constructed a convolutional neural network (CNN)-based architecture with one-hot encoding approach, named as CNN_{OH}. This model performed favorably to the traditional ML models and other DL models across different species, in terms of cross-validation and independent test. It is also superior to the documented K_{hib} predictors. Furthermore, we constructed a general model based on the integrated data from multiple species and it demonstrated great generality and effectiveness. Finally, we shared both species-specific models and the general model as the on-line prediction tool DeepKhib for easily identifying K_{hib} sites.

MATERIALS AND METHODS

Dataset Collection

The experimentally identified K_{hib} sites from five different organisms including *Homo sapiens* (human), *Oryza sativa* (rice), *Physcomitrella patens* (moss) and two one-celled eukaryotes *Toxoplasma gondii* and *Saccharomyces cerevisiae*. The data of the species were pre-processed and the related procedure was exemplified using the human data, as listed below (**Supplementary Figure 2**).

We collected 12,166 K_{hib} sites from 3,055 human proteins (Wu et al., 2018). These proteins were classified into 2,466 clusters using CD-HIT with the threshold of 40% according to the previous studies (Li and Godzik, 2006; Huang et al., 2010). In each cluster, the protein with the most K_{hib} sites was selected as the representative of the cluster. On the 2,466 representatives, 9,473 K_{hib} sites were considered positives whereas the remaining K sites were taken as negatives. We further estimated the potential redundancy of the positive sites by extracting the peptide segment of seven residues with the K_{hib} site in the center and count the number of unique segments (Chen et al., 2018a; Xie et al., 2018). The number (9,444) of the unique segments is 99.7% of the total segments, suggesting considerable diversity of the positive segments. The number of the negative sites (103,987) is 11 times larger than that of the positive sites. To avoid the potential impact of biased data on model construction, we referred to previous studies and balanced positives and negatives by randomly selecting the same number of negative sites (Huang et al., 2018c; Tahir et al., 2019). These positives and negatives composed the whole human dataset.

To determine the optimal sequence window for model construction, we tested different sequence window sizes ranging from 21 to 41, referring to the previous PTM studies where the optimal window sizes are between 31 and 39 (Wang et al., 2017; Chen et al., 2018a; Huang et al., 2018b). The window size of 37 corresponded to the largest area under the ROC curve (AUC) through 10-fold cross-validation (**Supplementary Figure 3**) and was therefore selected in this study. It should be noted that if the central lysine residue is located near the N-terminus or C-terminus of the protein sequence, the symbol "X" is added at the related terminus to ensure the same window size of the sequences.

Figure 1 showed the flowcharts for all the species. The dataset of each species was randomly separated into five groups of which four were used for 10-fold cross-validation and the rest for independent test. Each group contained the same number of positives and negatives. Specifically, the cross-validation datasets included 15,156/15,464/10,204/12,354 samples for *H. sapiens*/*T. gondii*/*O. sativa*/*P. patens*, respectively. Accordingly, the independent test sets comprised 3,790/3,866/2,552/3,090 samples for these organisms, separately. These datasets are available at http://www.bioinfogo.org/DeepKhib.

Feature Encodings

The ZSCALE Encoding

Each amino acid is characterized by five physiochemical descriptor variables (Sandberg et al., 1998; Chen et al., 2012).

The Encoding of Extended Amino Acid Composition (EAAC) Encoding

The EAAC encoding is based on the calculation of the amino acid composition (AAC) that indicates the amino acid frequencies for every position in the sequence window. EAAC is calculated by continuously sliding using a fixed-length sequence window (the default is 5) from the N-terminus to the C-terminus of each peptide (Chen et al., 2018b). The related formula is listed below:

$$f(t, win) = \frac{N(t, win)}{N(win)}, t \in \{A, C, D, \cdots, Y\},$$

$$win \in \{window1, window2, \cdots, window37\} \quad (1)$$

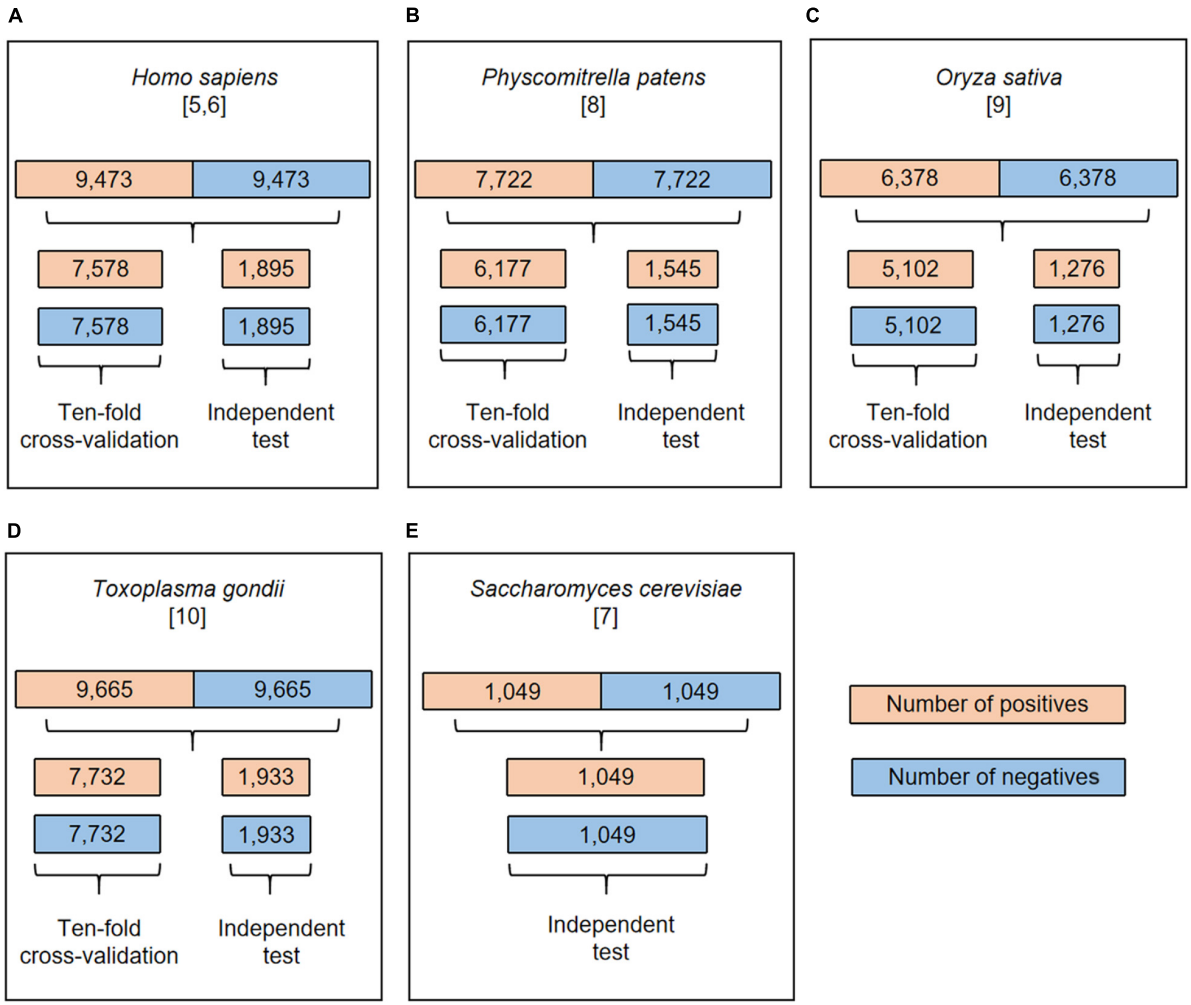


FIGURE 1 | The flowchart of dataset process for *H. sapiens* **(A)**, *P. patens* **(B)**, *O. sativa* **(C)**, *T. gondii* **(D)**, and *S. cerevisiae* **(E)**. All the datasets were separated into cross-validation and independent test datasets except the *S. cerevisiae* dataset.

where N (t, win) is the number of amino acid t in the sliding window win, and N(win) is the size of the sliding window win.

The Enhanced Grouped Amino Acids Content (EGAAC) Encoding

The EGAAC feature (Zhao et al., 2020) is developed based on the grouped amino acids content (GAAC) feature (Chen et al., 2018b, 2020). In the GAAC feature, the 20 amino acid types are categorized into five groups (g1: GAVLMI, g2: FYW, g3: KRH, g4: DE and g5: STCPNQ) according to their physicochemical properties and the frequencies of the groups are calculated for every position in the sequence window. For the EGAAC feature, the GAAC values are calculated in the window of fixed length (the default as 5) continuously sliding from the N- to C-terminal of each peptide sequence.

The One-Hot Encoding

The one-hot encoding is represented by the conversion of the 20 types of amino acids to 20 binary bits. By considering the complemented symbol "X," a vector of size (20+1) bits is used to represent a single position in the peptide sequence. For example, the amino acid "A" is represented by "10000000000000000000," "Y" is represented by "00000000000000000010," and the symbol "X" is represented by "00000000000000000001."

Architecture of the Machine-Learning Models

The CNN Model With One-Hot Encoding

The CNN algorithm (Fukushima, 1980) decomposes an overall pattern into many sub-patterns (features) through a neurocognitive machine, and then enters the hierarchically connected feature plane for processing. The architecture of the CNN model with one-hot encoding (called as CNN_{OH}) contained four layers as follows (**Figure 2A**).

(i) The first layer was the input layer where peptide sequences were represented using the one-hot encoding approach.

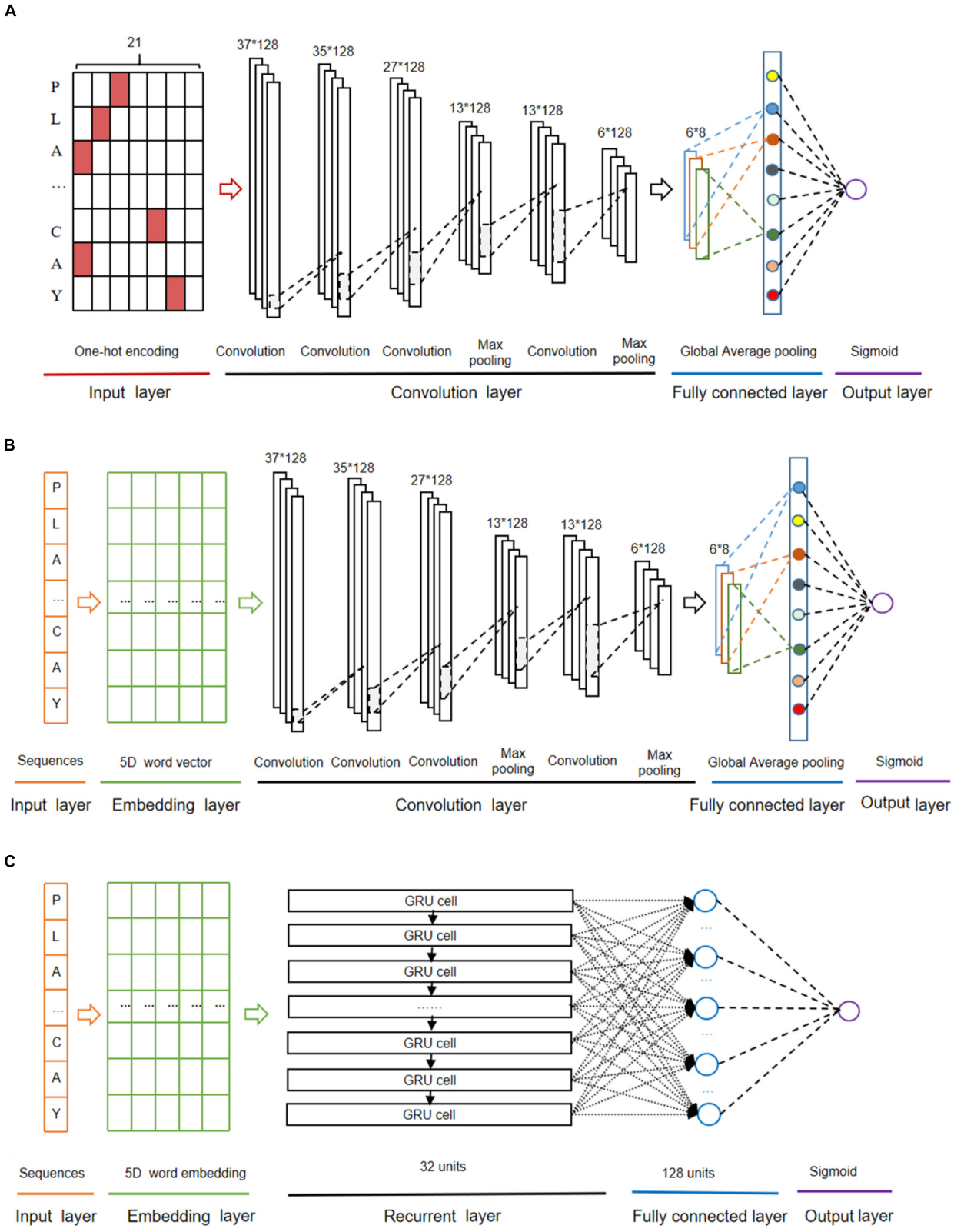


FIGURE 2 | The deep-learning architectures for CNN_{OH} **(A)**, CNN_{WE} **(B)**, and GRU_{WE} **(C)**.

TABLE 1 | Performances comparison of the different classifiers for human K_{hib} prediction.

	Classifier	Sn	Sp	Acc	MCC	AUC
10-fold cross-validation	RF_{EGAAC}	0.727 ± 0.015	0.682 ± 0.017	0.704 ± 0.011	0.409 ± 0.022	0.775 ± 0.011
	RF_{EAAC}	0.744 ± 0.025	0.645 ± 0.023	0.695 ± 0.010	0.391 ± 0.020	0.763 ± 0.008
	RF_{ZSCALE}	0.681 ± 0.016	0.662 ± 0.018	0.672 ± 0.011	0.344 ± 0.023	0.740 ± 0.014
	$RF_{EGAAC+EAAC}$	0.748 ± 0.019	0.691 ± 0.023	0.719 ± 0.012	0.439 ± 0.025	0.789 ± 0.011
	$RF_{EGAAC+ZSCALE}$	0.726 ± 0.019	0.707 ± 0.015	0.716 ± 0.012	0.433 ± 0.025	0.794 ± 0.010
	$RF_{EGAAC+EAAC+ZSCALE}$	0.751 ± 0.016	0.702 ± 0.022	0.727 ± 0.013	0.454 ± 0.026	0.802 ± 0.010
	GRU_{WE}	0.821 ± 0.024	0.683 ± 0.033	0.752 ± 0.009	0.509 ± 0.018	0.830 ± 0.007
	CNN_{WE}	0.849 ± 0.035	0.722 ± 0.042	0.786 ± 0.007	0.578 ± 0.012	0.867 ± 0.005
	CNN_{OH}	0.876 ± 0.025	0.700 ± 0.026	0.788 ± 0.007	0.586 ± 0.014	0.868 ± 0.004
Independent test	RF_{EGAAC}	0.719 ± 0.006	0.676 ± 0.007	0.698 ± 0.002	0.395 ± 0.004	0.767 ± 0.002
	RF_{EAAC}	0.755 ± 0.003	0.638 ± 0.007	0.697 ± 0.003	0.396 ± 0.006	0.764 ± 0.003
	RF_{ZSCALE}	0.680 ± 0.008	0.658 ± 0.009	0.669 ± 0.005	0.337 ± 0.011	0.736 ± 0.003
	$RF_{EGAAC+EAAC}$	0.740 ± 0.006	0.678 ± 0.005	0.709 ± 0.002	0.419 ± 0.005	0.781 ± 0.002
	$RF_{EGAAC+ZSCALE}$	0.728 ± 0.006	0.692 ± 0.006	0.710 ± 0.002	0.420 ± 0.005	0.787 ± 0.002
	$RF_{EGAAC+EAAC+ZSCALE}$	0.752 ± 0.005	0.693 ± 0.004	0.723 ± 0.002	0.446 ± 0.005	0.796 ± 0.002
	GRU_{WE}	0.806 ± 0.015	0.692 ± 0.029	0.749 ± 0.004	0.501 ± 0.007	0.824 ± 0.005
	CNN_{WE}	0.846 ± 0.035	0.719 ± 0.042	0.783 ± 0.006	0.572 ± 0.009	0.865 ± 0.004
	CNN_{OH}	0.874 ± 0.026	0.690 ± 0.035	0.782 ± 0.005	0.575 ± 0.005	0.871 ± 0.001

The data sets for 10-fold cross-validation and an independent test were described in the section "Materials and Methods." The RF classifier with the different encoding approach was named as RF_{EGAAC}, RF_{EAAC}, RF_{ZSCALE}, $RF_{EGAAC+EAAC}$, $RF_{EGAAC+ZSCALE}$, and $RF_{EGAAC+EAAC+ZSCALE}$. The RNN/CNN classifier with the word embedding encoding approach was named as GRU_{WE} /CNN_{WE}, respectively. The CNN classifier with one-hot encoding was named as CNN_{OH}. Ten models were constructed in the 10-fold cross validation and evaluated using the ten different validation datasets and the same independent dataset. Accordingly, the value Sn, Sp, Acc, MCC, and AUC were represented by average ± standard deviation.

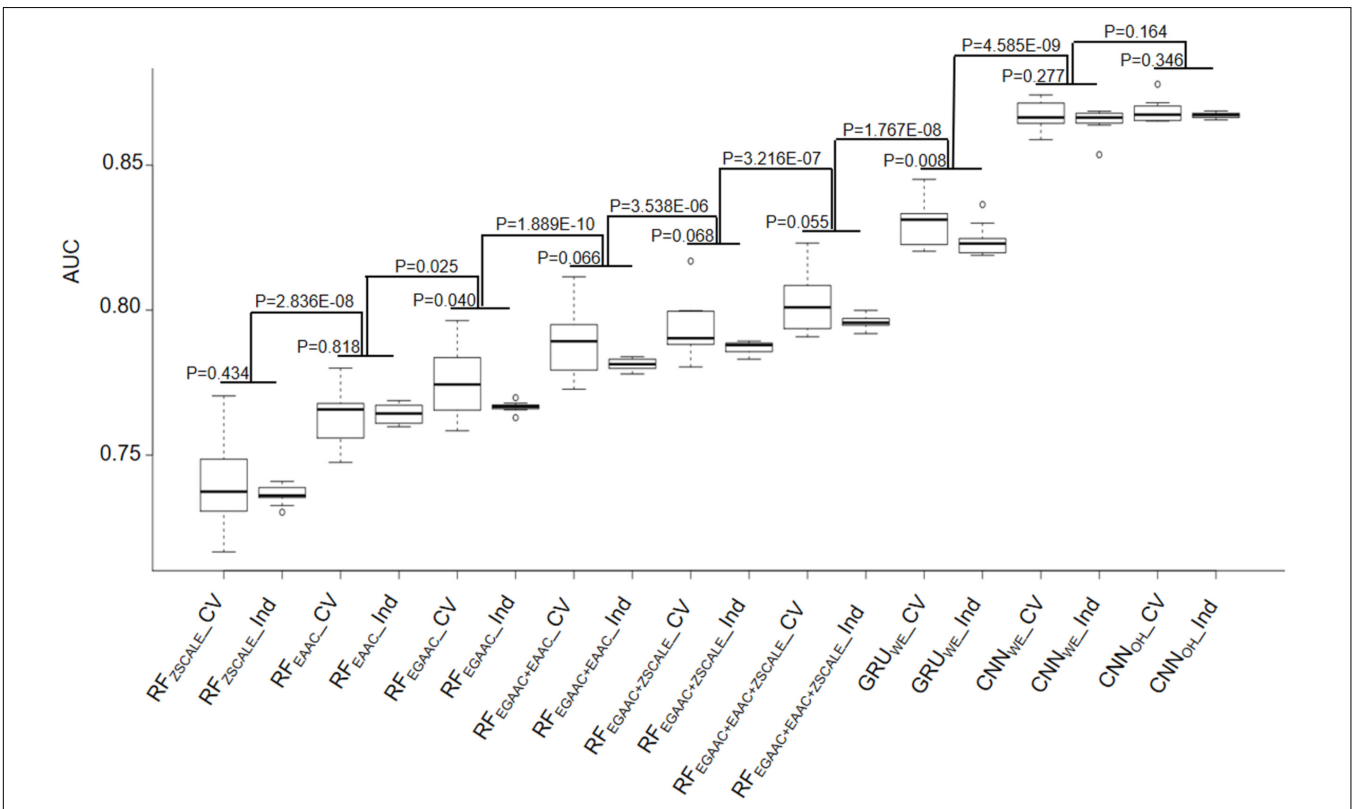


FIGURE 3 | Performance comparison of 10-fold cross-validation and independent test datasets of nine different models.

(ii) The second layer was the convolution layer that consisted of four convolution sublayers and two max pooling sublayers. The convolution sublayers, each sublayer uses 128 convolution filters, the length of which are 1, 3, 9, and 10, respectively. The two max pooling sublayers followed the third and fourth convolution sublayers, individually.
(iii) The third layer contained the fully connected sublayer, which contained a fully connected sublayer with eight neuron units without flattening, and a global average pooling sublayer, which was adopted to correlate the feature mapping with category output in order to reduce training parameters and avoid over-fitting.
(iv) The last layer was the output layer that included a single unit outputting the probability score of the modification, calculated using the "Sigmoid" function. If the probability score is greater than a specified threshold (e.g., 0.5), the peptide is predicted to be positive.

The "ReLU" function (Hahnloser et al., 2000) was used as the activation function of the convolution sublayers and fully connected sublayers of the above layers to avoid gradient dispersion in the training process. The Adam optimizer (Kingma and Jimmy, 2014) was used to optimize the hyper-parameters of this model, which include batch size, maximum epoch, learning rate and dropout rate. The maximum training period was set as 1000 epochs to ensure the convergence of the loss function values. In each epoch, the training data set was separated and iterated in a batch size of 1024. To avoid over-fitting, the dropout of neurons units in each convolution sublayer of the second layer was set 70% and that in the full connection sublayer of the third layer was set 30% (Nitish et al., 2014), the early stop strategy was adopted and the best model was saved.

The CNN Algorithm With Word Embedding

The CNN algorithm with word embedding (CNN_{WE}) contained five layers (**Figure 2B**). The input layer receives the sequence of window size 37 and each residue is transformed into a five-dimensional word vector in the embedding layer. The rest layers are the same as the corresponding layers in CNN_{OH}.

The GRU Algorithm With Word Embedding

The GRU algorithm (Cho et al., 2014) includes an update gate and a reset gate. The former is used to control the extent to which the state information at the previous moment is brought into the current state, whereas the latter is used to control the extent to which the state information at the previous moment is ignored. The GRU algorithm with word embedding (GRU_{WE}) contained five layers (**Figure 2C**). The first, the second and the last layers are the same as the corresponding layers in CNN_{WE}. The third layer is the recurrent layer where each word vector from the previous layer was sequentially inputted into the related GRU unit that contains 32 hidden neuron units. The fourth layer was the fully connected layer that contains 128 neuron units with "ReLU" as the activation function.

The RF Algorithms With Different Features

The Random Forest algorithm (Breiman, 2001) contains multiple decision trees, which remain unchanged under the scaling of feature values and various other transformations, and the output category is determined by the mode of the category output by the individual tree. Each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest. The number of decision trees was set 140. This classifier was developed based on the Python module "sklearn."

Cross-Validation and Performance Evaluation

To evaluate the performance of K_{hib} sites prediction, we adopted four statistical measurement methods. They included sensitivity (Sn), specificity (Sp), accuracy (ACC), and Matthew's correlation coefficient (MCC), listed as follows:

$$Sn = \frac{TP}{TP + FN} \tag{2}$$

$$Sp = \frac{TN}{TN + FP} \tag{3}$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

$$MCC = \frac{TP \times TN - TN \times FP}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{5}$$

In the above equations, TP is true positives, FP is false positives, TN is true negatives, FN is false negatives. In addition, the area under the receiver operating characteristic (ROC) curve (AUC) values was calculated to evaluate the performance of the prediction model.

TABLE 2 | The AUC values of the CNN_{OH} model constructed for *O. sativa, P. patens, T. gondii,* and *H. sapiens*, respectively.

Species	10-fold cross-validation	Independent test
O. sativa	0.823	0.818
P. patens	0.830	0.831
T. gondii	0.862	0.865
H. sapiens	0.868	0.871

TABLE 3 | The AUC values of different CNN_{OH} models in terms of independent test for five distinct organisms.

Prediction models	Independent data sets				
	O. sativa	***P. patens***	***T. gondii***	***H. sapiens***	***S. cerevisiae***
O. sativa	**0.818**	0.788	0.782	0.803	0.721
P. patens	0.761	**0.831**	0.812	0.837	**0.806**
T. gondii	0.781	0.813	**0.865**	0.827	0.776
H. sapiens	0.778	0.818	0.832	**0.871**	0.785
General	**0.802**	**0.840**	**0.860**	**0.868**	**0.789**

The top two models with best performance are bold.

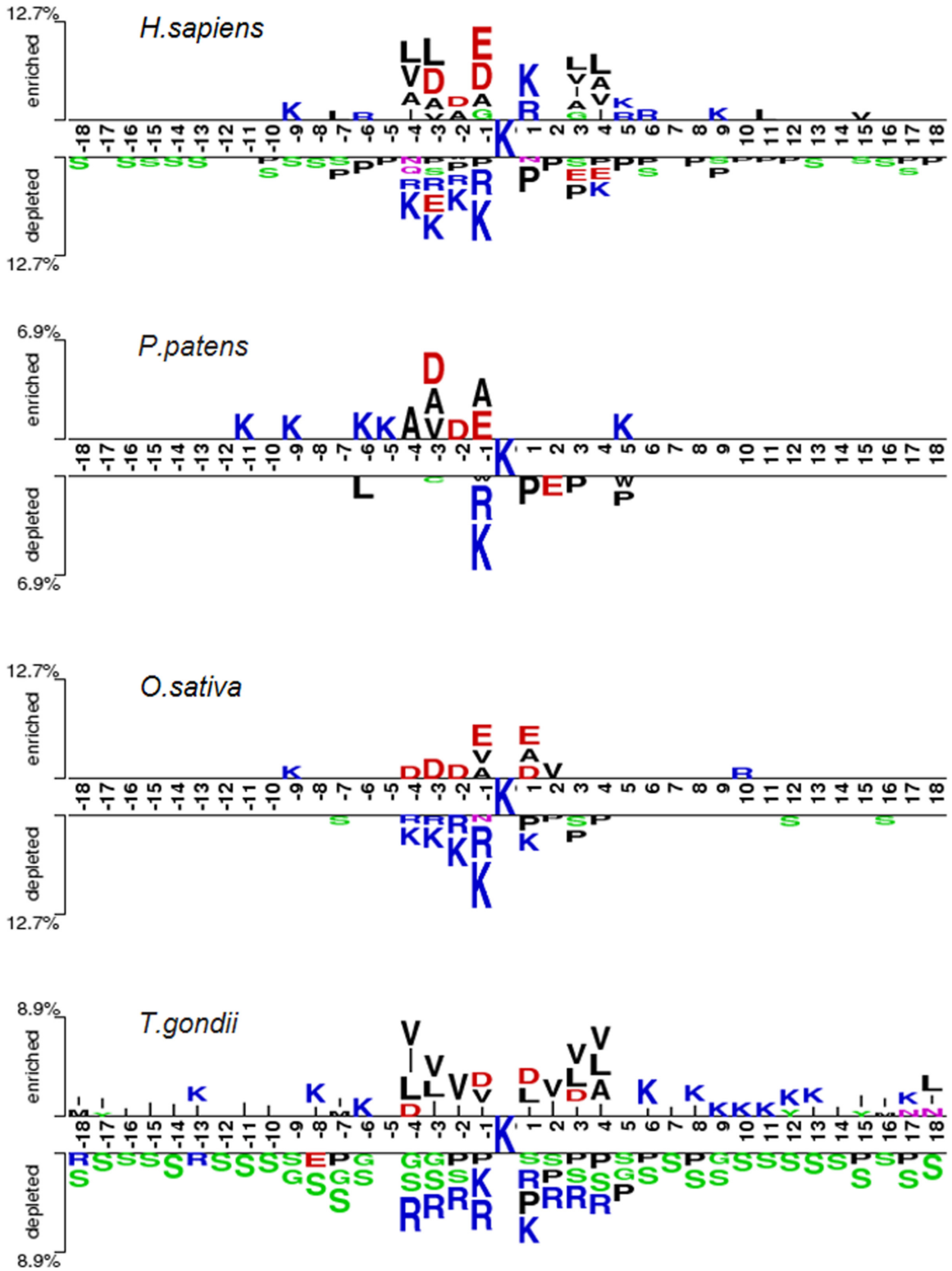


FIGURE 4 | Sequence pattern surrounding the K_{hib} sites, including the significantly enriched and depleted residues based on K_{hib} peptides and non-modification peptides from different species ($P < 0.05$, student's *T*-test with Bonferroni correction). The pattern was generated using the two-sample-logo method (Vacic et al., 2006).

Statistical Methods

The paired student's *t*-test was used to test the significant difference between the mean values of the two paired populations. As for multiple comparisons, the adjusted P value with the Benjamini-Hochberg (BH) method was adopted.

RESULTS AND DISCUSSION

A couple of computational approaches has been developed for the prediction of K_{hib} sites (Ju and Wang, 2019; Wang et al., 2020). Recently, this modification has been investigated across five different species, ranging from single-celled organisms to multiple-celled organisms and from plants to mammals. Additionally, the number of reported sites has been significantly increased. These raised our interest to develop novel prediction algorithms and explore the characteristics of this modification. We pre-processed the data from different species and separated them into the cross-validation dataset and the independent test set (see section "Materials and Methods" for detail; **Figure 1**). We first took the human data as the representative to compare different models and then applied the model with the best performance to other species. The human cross-validation dataset contained 15,156 samples and the independent test set covered 3,790 samples, in each of which half were positives and half were negatives.

CNN_{OH} Showed Superior Performance

We constructed nine models, divided into two categories: six traditional ML models and three DL models (see section "Materials and Methods" for details). The traditional ML models were based on the RF algorithm combined with different encoding schemes. The DL models included a Gated Recurrent Unit (GRU) model with the word-embedding encoding approach dubbed GRU_{WE} and two CNN models with the one-hot and word-embedding encoding approaches named CNN_{OH} and CNN_{WE}, respectively. Both encoding methods are common in the DL algorithms (Chen et al., 2018a; Xie et al., 2018).

The RF-based models were developed with different common encoding schemes, including EAAC, EGAAC and ZSCALE. Among these encoding schemes, EGAAC had the best performance followed by EAAC whereas ZSCALE was the worst in terms of AUC and ACC for both 10-fold cross-validation and the independent test (**Table 1** and **Figure 3**). For instance, EGAAC corresponded to the average AUC value as 0.775, EAAC had the value as 0.763 and ZSCALE had the value as 0.740 for cross validation. Because different encodings represent distinct characteristics of K_{hib}-containing peptides, we evaluated the combinations of the encoding schemes. The combinations showed better performances than individual scheme and the combination of all the three was the best for both cross-validation and the independent test, in terms of AUC, MCC, and ACC (**Table 1** and **Figure 3**). Therefore, the K_{hib} prediction accuracy could be improved by the integration of different encoding schemes.

As the DL algorithms showed superior to the traditional ML algorithms for a few PTM predictions in our previous studies (Chen et al., 2019; Zhao et al., 2020), we examined the DL algorithms for the K_{hib} prediction. Traditionally, CNN is popular for image prediction with spatial invariant features while RNN is ideal for text prediction with sequence features. However, many cases demonstrate that CNN also has good performance when applied to sequence data (Sainath et al., 2013; Tahir et al., 2019). Accordingly, we developed both RNN and CNN models for the K_{hib} prediction with two common encoding approaches: one-hot and word-embedding. Expectedly, all three DL models were significantly better than the traditional ML models constructed above in the cross-validation and independent test (**Table 1** and **Figure 3**). For instance, the average AUC values of the DL models were above 0.824 whereas those of the ML models were less than 0.802.

In these DL models, two CNN models CNN_{OH} and CNN_{WE} had similar performances and both compared favorably to GRU_{WE} (**Table 1** and **Figure 3**). CNN_{OH} had the AUC value as 0.868 for the cross-validation and its values of SN, SP, ACC and MCC were 0.876, 0.700, 0.788, and 0.586, respectively. Here, we chose CNN_{OH} as the 2-Hydroxyisobutyrylation predictor. We evaluated the robustness of our models by comparing their performances between the cross-validation and independent tests. As their performances between these two tests had no statistically different ($P > 0.01$), we concluded that our constructed models were robust and neither over-fitting nor under-fitting.

Construction and Comparison of Predictors for Other Species

We constructed nine models for the human organism and chose CNN_{OH} as the final prediction model. We applied the CNN_{OH} architecture to the other three organisms (i.e., *T. gondii, O. sativa,* and *P. patens*). For each organism, we separated the dataset as the cross-validation set and the independent set. Similar to the human species, the CNN_{OH} models for these species had similar performances between cross-validation and independent test and their AUC values were larger than 0.818 (**Table 2**). It indicates that these constructed models are effective and robust.

As lysine 2-Hydroxyisobutyrylation is conserved across different types of species, we hypothesized that the model built for one species may be used to predict K_{hib} sites for other species. To test this hypothesis, we compared the performances of the CNN_{OH} models in terms of the independent data sets of individual species. Additionally, we built a general CNN_{OH} model based on the training datasets integrated from all the four species. **Table 3** shows that the AUC values of these predictions were larger than 0.761, suggesting that the cross-species prediction had reliable performances. Specifically, given a species, the best prediction performances were derived from

TABLE 4 | The prediction performance of CNN_{OH} compared to iLys-Khib in terms of the same cross-validation and independent test datasets.

Dataset	Model	Sn	Sp	Acc	MCC	AUC
10-fold cross-validation	iLys-Khib	0.745	0.658	0.701	0.404	0.770
	CNN_{OH}	0.830	0.713	0.772	0.547	0.847
Independent test	iLys-Khib	0.725	0.643	0.648	0.186	0.756
	CNN_{OH}	0.861	0.685	0.696	0.281	0.860

A

B

DeepKhib prediction result

Download predictions:

Protein	Position	Sequence	Prediction score	Prediction category
A0AV96	21	AEDSTAAMSSDSAAGSSAKVPEGVAGAPNEAALLALM	0.011984	No
A0AV96	54	LALMERTGYSMVQENGQRKYGGPPPGWEGPHPQRGCE	0.240550	No
A0AV96	77	PPGWEGPHPQRGCEVFVGKIPRDVYEDELVPVFEAVG	0.664412	Very high confidence
A0AV96	109	FEAVGRIYELRLMMDFDGKNRGYAFVMYCHKHEAKRA	0.924083	Very high confidence
A0AV96	121	MMDFDGKNRGYAFVMYCHKHEAKRAVRELNNYEIRPG	0.491399	High confidence
A0AV96	125	DGKNRGYAFVMYCHKHEAKRAVRELNNYEIRPGRLLG	0.084327	No
A0AV96	160	LGVCCSVDNCRLFIGGIPKMKKREEILEEIAKVTEGV	0.599747	High confidence
A0AV96	162	VCCSVDNCRLFIGGIPKMKKREEILEEIAKVTEGVLD	0.044440	No
A0AV96	163	CCSVDNCRLFIGGIPKMKKREEILEEIAKVTEGVLDV	0.043233	No
A0AV96	173	IGGIPKMKKREEILEEIAKVTEGVLDVIVYASAADKM	0.020177	No
A0AV96	190	AKVTEGVLDVIVYASAADKMKNRGFAFVEYESHRAAA	0.597568	High confidence
A0AV96	192	VTEGVLDVIVYASAADKMKNRGFAFVEYESHRAAAMA	0.093939	No
A0AV96	213	GFAFVEYESHRAAAMARRKLMPGRIQLWGHQIAVDWA	0.136066	No
A0AV96	246	VDWAEPEIDVDEDVMETVKILYVRNLMIETTEDTIKK	0.073464	No
A0AV96	263	VKILYVRNLMIETTEDTIKKSFGQFNPGCVERVKKIR	0.908789	Very high confidence
A0AV96	264	KILYVRNLMIETTEDTIKKSFGQFNPGCVERVKKIRD	0.596560	High confidence
A0AV96	278	DTIKKSFGQFNPGCVERVKKIRDYAFVHFTSREDAVH	0.004530	No
A0AV96	279	TIKKSFGQFNPGCVERVKKIRDYAFVHFTSREDAVHA	0.016062	No

Legend:

Label	Score Range	Specificity
Very high confidence	(0.643 - 1)	>99%
High confidence	(0.441- 0.643)	95%-99%
Medium confidence	(0.32 - 0.441)	90%-95%
No	(0,0.32)	<90%

FIGURE 5 | DeepKhib interface for the prediction of K_{hib} sites with the option of organism-specific or general classifiers **(A)** and its application to the prediction **(B)**.

the general model and the model developed specifically for this species. For instance, the human CNN_{OH} model had the best performance followed by the general model in terms of the human independent test whereas the general model had the best accuracy followed by the moss-specific model for the moss independent test. These suggest that on one hand, lysine 2-Hydroxyisobutyrylation of each species has its own characteristics; one the other hand, this modifications across different species share strong commonalities. Therefore, the general model may be effectually applied to any species. Furthermore, we evaluated the generality of the general CNN_{OH} model using the dataset of *S. cerevisiae* that contained 1,049 positive and 1,049 negative samples, which may not be enough for build an effective DL predictor (Chen et al., 2018a). The general model got the AUC value as 0.789, indicating the generality of this model. In other words, the general model is effective to predict K_{hib} sites for any organism.

We identified and compared the significant patterns and conserved motifs between K_{hib} and non-K_{hib} sequences across the different organisms using the two-sample-logo program with *t*-test ($P < 0.05$) with Bonferroni correction (Vacic et al., 2006). **Figure 4** shows the similarities and differences between the species. For instance, the residues R and K at the −1 position (i.e., R&K@P-1) and P at +1 position (i.e., P@P+1) are significantly depleted across the species. On the contrary, K&R@P+1 tend to be enriched for *H. sapiens* but depleted for *T. gondii* whereas both species have the depleted residue Serine across the positions ranging from P-18 to P+18. These similarities between the organisms may result in the generality and effectiveness of the general CNN_{OH} model.

Comparison of CNN_{OH} With the Reported Predictors

We assessed the performance of CNN_{OH} by comparing it with the existing K_{hib} predictors KhibPred (Wang et al., 2020) and iLys-Khib (Ju and Wang, 2019). First, we compared CNN_{OH} with KhibPred for individual species in terms of 10-fold cross-validation (Wang et al., 2020). The average AUC values of CNN_{OH} were 0.868/0.830/0.823 for *H. sapiens/P. patens/O. sativa*, respectively (**Table 2**). On the contrary, the corresponding values of KhibPred were 0.831/0.781/0.825 (Wang et al., 2020). Thus, CNN_{OH} compares favorably to KhibPred. Second, the model iLys-Khib was constructed and tested using 9,318 human samples as the 10-fold cross-validation data set and 4,219 human samples as the independent test set. We used the same datasets to construct CNN_{OH} and compared it with iLys-Khib. CNN_{OH} outperformed iLys-Khib in terms of all the measurements of performance (e.g., Sn, Sp, Acc, MCC, and AUC) for both 10-fold cross-validation and independent test (**Table 4**). For instance, the AUC value of CNN_{OH} was 0.860 for the independent test whereas that of iLys-Khib was 0.756. In summary, CNN_{OH} is a competitive predictor.

Construction of the On-Line K_{hib} Predictor

We developed an easy-to-use Web tool for the prediction of K_{hib} sites, dubbed as DeepKhib. It contains five CNN_{OH} models, including one general model and four models specific to the species (i.e., *H. sapiens, O. sativa, P. patens*, *and T. gondii*). Given a species of interest, users could select the suitable model (e.g., the general model or the model specific to an organism) for prediction (**Figure 5A**). After the protein sequences as the fasta file format are uploaded, the prediction results will be shown with five columns: Protein, Position, Sequence, Prediction score, and Prediction category (**Figure 5B**). The prediction category covered four types according to the prediction scores: no (0–0.320), medium confidence (0.320–0.441), high confidence (0.441–0.643), and very high confidence (0.643–1).

CONCLUSION

The common PTM classifiers are mainly based on the traditional ML algorithms that require the pre-defined informative features. Here, we applied the advanced DL algorithm CNN_{OH} for predicting K_{hib} sites. CNN_{OH} shows its superior performance, because of the capability of the multi-layer CNN algorithm to extract complex features and learn sparse representation in a self-taught manner. Moreover, the general CNN_{OH} model demonstrates great generality and effectiveness, due to the conservation of K_{hib} modification from single-cell to multiple-cell organisms. The outstanding performance of DL in the prediction of K_{hib} sites suggests that DL may be applied broadly to predicting other types of modification sites.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are accessible through http://www.bioinfogo.org/DeepKhib/download.php.

AUTHOR CONTRIBUTIONS

LL conceived this project. LZ and YZ constructed the algorithms under the supervision of LL and ZC. LZ and NH analyzed the data. LL, YZ, YC, and LZ wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported in part by funds from the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 31701142 to ZC) and the National Natural Science Foundation of China (Grant No. 31770821 to LL). LL was supported by the "Distinguished Expert of Overseas Tai Shan Scholar" program. YZ was supported by the Qingdao Applied Research Project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2020.580217/full#supplementary-material

REFERENCES

Beltrao, P., Bork, P., Krogan, N. J., and Van Noort, V. (2013). Evolution and functional cross-talk of protein post-translational modifications. *Mol. Syst. Biol.* 9:714. doi: 10.1002/msb.201304521

Breiman, L. (2001). Random Forests. *Machine Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Chen, Y. Z., Chen, Z., Gong, Y. A., Ying, G., and Parkinson, J. (2012). Sumohydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties. *PLoS One* 7:e39195. doi: 10.1371/journal.pone.0039195

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. *Genom. Proteom. Bioinform.* 16, 451–459.

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

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, 2267–2290.

Chen, Z., Zhao, P., Li, F., Marquez-Lago, T. T., Leier, A., et al. (2020). iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief. Bioinform.* 21, 1047–1057.

Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Comput. Ence.* 2014, 1724–1734. doi: 10.3115/v1/D14-1179

Dai, L., Peng, C., Montellier, E., Lu, Z., Chen, Y., Ishii, H., et al. (2014). Lysine 2-hydroxyisobutyrylation is a widely distributed active histone mark. *Nat. Chem. Biol.* 10, 365–370.

Fukushima, K. (1980). Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* 36, 193–202. doi: 10.1007/BF00344251

Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., and Seung, H. S. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* 405, 947–951. doi: 10.1038/35016072

Huang, H., Luo, Z., Qi, S., Huang, J., Xu, P., Wang, X., et al. (2018a). Landscape of the regulatory elements for lysine 2-hydroxyisobutyrylation pathway. *Cell Res.* 28, 111–125.

Huang, H., Tang, S., Ji, M., Tang, Z., Shimada, M., Liu, X., et al. (2018b). p300-Mediated Lysine 2-Hydroxyisobutyrylation Regulates Glycolysis. *Mol. Cell* 70, 663–678.e.

Huang, Y., He, N., Chen, Y., Chen, Z., and Li, L. (2018c). BERMP: a cross-species classifier for predicting mA sites by integrating a deep learning algorithm and a random forest approach. *Int. J. Biol. Sci.* 14, 1669–1677.

Huang, J., Luo, Z., Ying, W., Cao, Q., and Dai, J. (2017). 2-hydroxyisobutyrylation on histone h4k8 is regulated by glucose homeostasis in saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U.S.A.* 114, 8782–8787. doi: 10.1073/pnas.1700796114

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26, 680–682. doi: 10.1093/bioinformatics/btq003

Ju, Z., and Wang, S.-Y. (2019). iLys-Khib: Identify lysine 2-Hydroxyisobutyrylation sites using mRMR feature selection and fuzzy SVM algorithm. *Chemometr. Intell. Laborat. Syst.* 191, 96–102. doi: 10.1016/j.chemolab.2019.06.009

Kingma, D. P., and Jimmy, B. (2014). *Adam: A Method for Stochastic Optimization*. New York, NY: Cornell University.

Li, Q. Q., Hao, J. J., Zhang, Z., Krane, L. S., Hammerich, K. H., Sanford, T., et al. (2017). Proteomic analysis of proteome and histone post-translational modifications in heat shock protein 90 inhibition-mediated bladder cancer therapeutics. *Sci. Rep.* 7:201.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Long, H., Liao, B., Xu, X., and Yang, J. (2018). A Hybrid Deep Learning Model for Predicting Protein Hydroxylation Sites. *Int. J. Mol. Sci.* 19:2817.

Meng, X., Xing, S., Perez, L. M., Peng, X., Zhao, Q., Redona, E. D., et al. (2017). Proteome-wide Analysis of Lysine 2-hydroxyisobutyrylation in Developing Rice (Oryza sativa) Seeds. *Sci. Rep.* 7:17486.

Nitish, S., Geoffrey, H., Alex, K., Ilya, S., and Ruslan, S. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Machine Lear. Res.* 15, 1929–1958.

Sainath, T. N., Mohamed, A., Kingsbury, B., and Ramabhadran, B. (2013). *Deep convolutional neural networks for LVCSR*. New Jersey: IEEE. doi: 10.1109/ICASSP.2013.6639347

Sandberg, M., Eriksson, L., and Jonsson, J. (1998). New chemical descriptors relevant for the design of biologically active peptides. *a multivariate characterization of* 87 amino acids. *J. Med. Chem.* 41, 2481–2491. doi: 10.1021/jm9700575

Skelly, M. J., Frungillo, L., and Spoel, S. H. (2016). Transcriptional regulation by complex interplay between post-translational modifications. *Curr. Opin. Plant Biol.* 33, 126–132. doi: 10.1016/j.pbi.2016.07.004

Tahir, M., Tayara, H., and Chong, K. T. (2019). iPseU-CNN: Identifying RNA Pseudouridine Sites Using Convolutional Neural Networks. *Mol. Ther. Nucl. Acids* 16, 463–470. doi: 10.1016/j.omtn.2019.03.010

Tian, Q., Zou, J., Tang, J., Fang, Y., Yu, Z., and Fan, S. (2019). MRCNN: a deep learning model for regression of genome-wide DNA methylation. *BMC Genomics* 20:192. doi: 10.1186/s12864-019-5488-5

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

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, 3909–3916.

Wang, Y. G., Huang, S. Y., Wang, L. N., Zhou, Z. Y., and Qiu, J. D. (2020). Accurate prediction of species-specific 2-hydroxyisobutyrylation sites based on machine learning frameworks. *Anal. Biochem.* 602:113793. doi: 10.1016/j.ab.2020.113793

Wu, Q., Ke, L., Wang, C., Fan, P., Wu, Z., and Xu, X. (2018). Global Analysis of Lysine 2-Hydroxyisobutyrylome upon SAHA Treatment and Its Relationship with Acetylation and Crotonylation. *J. Proteome Res.* 17, 3176–3183.

Xiao, H., Xuan, W., Shao, S., Liu, T., and Schultz, P. G. (2015). Genetic Incorporation of epsilon-N-2-Hydroxyisobutyryl-lysine into Recombinant Histones. *ACS Chem. Biol.* 10, 1599–1603. doi: 10.1021/cb501055h

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. *Genom. Proteom. Bioinform.* 16, 294–306.

Yin, D., Zhang, Y., Wang, D., Sang, X., Feng, Y., Chen, R., et al. (2019). Global Lysine Crotonylation and 2- Hydroxyisobutyrylation in Phenotypically Different Toxoplasma gondii Parasites. *Mole. Cell. Proteom.* 18, 2207–2224.

Yu, Z., Ni, J., Sheng, W., Wang, Z., and Wu, Y. (2017). Proteome-wide identification of lysine 2-hydroxyisobutyrylation reveals conserved and novel histone modifications in Physcomitrella patens. *Sci. Rep.* 7:15553. doi: 10.1038/s41598-017-15854-z

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

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.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 30 September 2020
doi: 10.3389/fcell.2020.572195

Incorporating Deep Learning With Word Embedding to Identify Plant Ubiquitylation Sites

Hongfei Wang[1†], Zhuo Wang[1,2†], Zhongyan Li[1,3] and Tzong-Yi Lee[1,3]*

[1] Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen, China, [2] School of Life Sciences, University of Science and Technology of China, Hefei, China, [3] School of Life and Health Sciences, The Chinese University of Hong Kong, Shenzhen, China

Protein ubiquitylation is an important posttranslational modification (PTM), which is involved in diverse biological processes and plays an essential role in the regulation of physiological mechanisms and diseases. The Protein Lysine Modifications Database (PLMD) has accumulated abundant ubiquitylated proteins with their substrate sites for more than 20 kinds of species. Numerous works have consequently developed a variety of ubiquitylation site prediction tools across all species, mainly relying on the predefined sequence features and machine learning algorithms. However, the difference in ubiquitylated patterns between these species stays unclear. In this work, the sequence-based characterization of ubiquitylated substrate sites has revealed remarkable differences among plants, animals, and fungi. Then an improved word-embedding scheme based on the transfer learning strategy was incorporated with the multilayer convolutional neural network (CNN) for identifying protein ubiquitylation sites. For the prediction of plant ubiquitylation sites, the proposed deep learning scheme could outperform the machine learning-based methods, with the accuracy of 75.6%, precision of 73.3%, recall of 76.7%, F-score of 0.7493, and 0.82 AUC on the independent testing set. Although the ubiquitylated specificity of substrate sites is complicated, this work has demonstrated that the application of the word-embedding method can enable the extraction of informative features and help the identification of ubiquitylated sites. To accelerate the investigation of protein ubiquitylation, the data sets and source code used in this study are freely available at https://github.com/wang-hong-fei/DL-plant-ubsites-prediction.

Keywords: ubiquitylation, plant, word embedding, deep learning, transfer learning, convolutional neural network

OPEN ACCESS

Edited by:
Dong Xu,
University of Missouri, United States

Reviewed by:
Peng(Sam) Sun,
Bayer Crop Science (United States), United States
Santosh Panjikar,
Australian Synchrotron, Australia

****Correspondence:***
Tzong-Yi Lee
leetzongyi@cuhk.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: *13 June 2020*
Accepted: *24 August 2020*
Published: *30 September 2020*

Citation:
Wang H, Wang Z, Li Z and Lee T-Y (2020) Incorporating Deep Learning With Word Embedding to Identify Plant Ubiquitylation Sites. Front. Cell Dev. Biol. 8:572195. doi: 10.3389/fcell.2020.572195

INTRODUCTION

As one of the most important posttranslational modification (PTM) processes, ubiquitylation is a modification process in which one or more ubiquitin molecules covalently bind to substrate proteins under the action of a series of enzymes (E1, E2, E3) (Weissman, 2001). The ubiquitin–proteasome pathway (UPP) is the most important protein degradation pathway in eukaryotic cells and participates in various physiological processes, including transcription regulation, cell cycle,

apoptosis, DNA damage repair, metabolism, and immunity (Tu et al., 2012). Moreover, its abnormal regulation is often accompanied with the occurrence of diseases such as cancer, neurodegenerative diseases, and liver diseases (Hoeller et al., 2006; Popovic et al., 2014; Yamada et al., 2018). UPP is closely related to plant physiology, and many studies have proved that ubiquitin–proteasome degradation is involved in plant growth and development, abiotic stress, plant metabolism, and biological stress (Lu et al., 2011;Marino and Rivas, 2012).

Because of the functional significance of ubiquitylation, the identification of new ubiquitylation sites in proteins is highly significant. However, wet laboratory experimental validations are often time consuming and expensive (Nguyen et al., 2016). In contrast, computation-based identification methods, which combine big data and advanced algorithms, can provide an alternative strategy for ubiquitylation site prediction with fast speed and low cost. The population of high-throughput proteomics experiment technology promotes large-scale identification of ubiquitin-conjugated peptides and, then, provides a very large dataset for automatic recognition of ubiquitination sites (Nguyen et al., 2015). Recently, numerous machine learning methods have been proposed for automatic prediction of ubiquitination sites. The Ubipred (Tung and Ho, 2008) is the first online tool that employed the physical and chemical properties of amino acids surrounding ubiquitination sites as features and integrated with support vector machine (SVM) to predict the ubiquitination sites. Then other machine learning methods, such as the k-nearest neighbor and random forest, are also used for ubiquitination site prediction (Radivojac et al., 2010; Cai et al., 2012; Xiang et al., 2013; Jyun-Rong et al., 2016). The hCKSAAP-UbSite (Chen et al., 2013) employed the idea of the composition of k-spaced amino acid pair (CKSAAP), which considers amino acid pair composition features of a specific position. Qiu et al. (2015) believing, through the simple observation of the composition of amino acids, that the sequence order of proteins may be ignored, utilized the pseudo-amino acid composition (PseAAC) to reserve these essential features and developed the iUbiq-Lys. The ubiquitylated protein data are collected from various eukaryotic species, and, considering the features of species evolution, Ubisite (Huang et al., 2016) proposed the position-specific scoring matrices (PSSM), which are calculated through PSI-BLAST. As a promising structural data modeling approach, the deep learning method can extract features from original data automatically without feature engineering, thus some potential and essential features will not be ignored. He et al. (2018) employed the deep learning approach on ubiquitination site prediction and received a well performance on their testing set.

However, the pattern differences between the ubiquitylated proteins of these species are not clear. To the best of our knowledge, no related work focuses on ubiquitylation prediction model development for a particular species. In this work, we first analyzed the pattern differences of ubiquitylated proteins between plants, animals, and fungus. Then an improved word-embedding training scheme based on transfer learning was proposed, connecting with the multilayer convolutional neural network (CNN) for plant ubiquitylation site prediction.

MATERIALS AND METHODS

The workflow of this study is described in **Figure 1**. We collected ubiquitination sites data from the Protein Lysine Modifications Database (PLMD) (Xu et al., 2017), which includes data collected from plants, animals, and fungus. In order to understand the pattern differences of ubiquitylated protein sequences between these species, feature investigations of three species were conducted. Several important sequence features were compared and analyzed to illustrate the pattern differences between plants and other species. Then a novel transfer learning-based word-embedding training scheme was proposed in which two steps of training were conducted. The

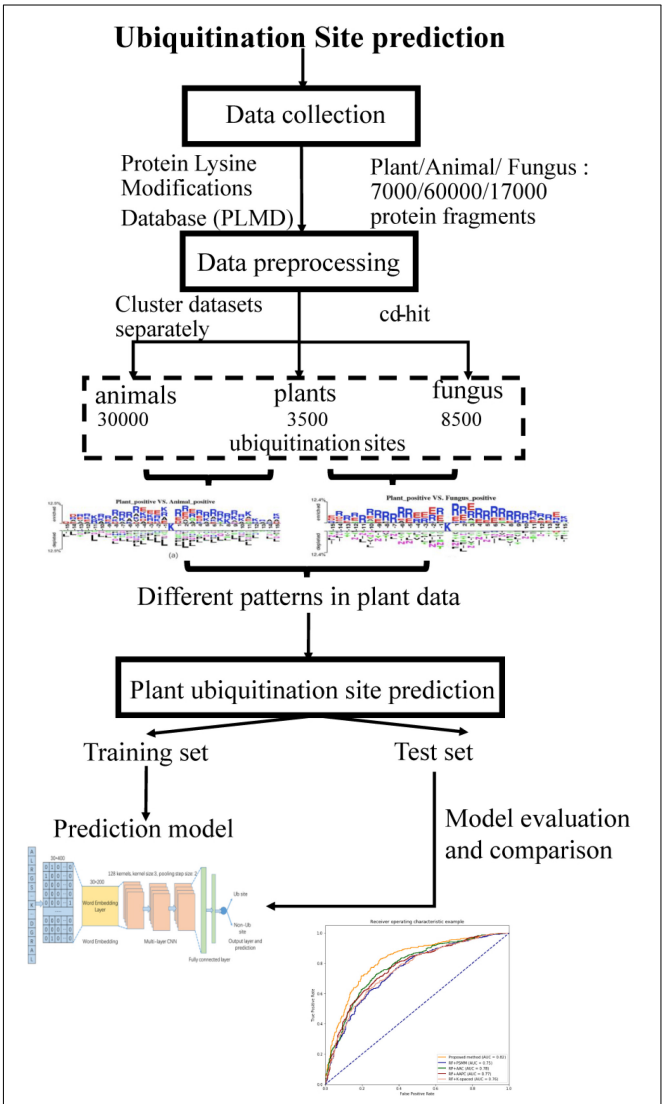


FIGURE 1 | Schematic diagram of the workflow for this study.

original plant protein sequence was used for pretraining of the word2vec network through the skip-gram model, with the optimized parameter transfer as the initial weights of embedding layer and fine-tuning with the subsequent layers together. The trained word-embedding layer captured the sequence features of the plant protein and was appropriated to ubiquitination site prediction at the same time. The multilayer CNN was employed as a classifier and achieved acceptable performance for plant ubiquitination site prediction. Sufficient experiments illustrated that the proposed method outperforms the conventional method on both cross-validation and the independent testing set.

Data Collection and Preprocessing

In this study, the ubiquitination protein sequence is collected from the PLMD database (Xu et al., 2017); the original data contains 121,742 ubiquitination sites from 25,103 proteins. We selected ubiquitination sites from *Arabidopsis thaliana*, *Oryza sativa* subsp *indica*, and *O. sativa* subsp *japonica* for the plant subset, ubiquitination sites from *Homo sapiens* and *Mus musculus* for the animal subset, and *Saccharomyces cerevisiae* for the fungus subset. To construct positive data of modeling, a sliding window with a length of 31 was used to intercept the protein sequences with ubiquitylated lysine residues in the center, where 31 equals 15 amino acids from each side of the lysine residue plus one lysine residue. If the upstream or downstream residues of a protein are less than 15, the lacking residue is filled with a pseudoresidue X. Then, the sequence fragments that contained a window length of 31 amino acids were centered at the lysine residue without annotation of the ubiquitination and were regarded as the negative data of modeling (non-ubiquitylated lysines). We removed the redundant protein fragments to eliminate homology bias using the CD-HIT (Li and Godzik, 2006) with 30% identity to ensure that none of the segments had a larger than 30% pairwise identity in both positive and negative peptides. There are too many negative peptides compared to the positive peptides. In order to keep the data balanced, we selected the same number of negative peptides randomly as positive peptides. Finally, we obtained 7,000 protein fragments for the plant subset, 60,000 protein fragments for the animal subset, and 17,000 protein fragments for the fungus subset.

We obtained 3,500 ubiquitination sites from plants after the preprocessing steps through CD-HIT tools, and then, we selected 3,500 negative samples randomly to keep the data balanced. In this work, we employed both the independent testing and cross-validation method to evaluate the performance of the proposed model. We selected 1,500 protein fragments randomly from the 7,000 samples as the independent testing set, which were used to evaluate the tuned model. In addition, we utilized the 10-fold cross-validation method to test the model performance using the remaining 5,500 samples. The original dataset was randomly partitioned into 10 equal-sized subsamples; a single subsample was retained as the validation data for testing the model, and the remaining nine subsamples were used as training data. The cross-validation process was then repeated 10 times, with each of the 10 subsamples used exactly once as the validation data.

Feature Investigation

Amino Acid Composition

As an important sequence feature, amino acid composition (AAC) can reflect which kind of amino acid is more likely to appear around the ubiquitylated lysine. In this work, we calculated the AAC feature of each peptide using the following equation:

$$A_r = \frac{N_r}{N} \ \mathrm{r} = 1, 2, 3, \ldots, 20$$

where N_{r} denotes the number of amino acid r, and N denotes the length of the protein fragments.

Amino Acid Pairwise Composition

In order to understand the efforts of amino acid complexes for ubiquitination in these species, we calculated the relative frequencies of all possible dipeptides in the sequence. The elements of the feature vector are defined as:

$$D_{r,s} = \frac{N_{r,s}}{N}, \ r, s = 1, 2, \ldots, 20$$

where $N_{\mathrm{r,s}}$ denotes the count of the dipeptide r,s, and N represents the total number of dipeptides in the encoded segment. Consequently, a 400-dimensional vector would be obtained for each segment. Then, heat maps were used to illustrate the dipeptide composition difference between the positive and negative samples, and the value of each pixel was calculated using the following equation:

$$P_{r,s} = \ln \frac{\sum D_{\mathrm{positive}}}{\sum D_{\mathrm{negative}}}$$

Positional Weighted Matrix

Then, we made the positional weighted matrix (PWM) to illustrate the pattern differences of the amino acid distribution around the ubiquitylated lysine between the positive and negative samples, and three heat maps were plotted for these three species, respectively. We define a two-dimensional matrix for each fragment as M^i, whose horizontal axis denotes the positions of protein fragments, and the central position is the targeted lysine, while the vertical axis denotes all these 20 kinds of amino acids. The final PWM for comparison of the positive and negative samples is calculated through the following equation:

$$M_{PA} = \ln \frac{\sum M^i_{\mathrm{positive}}}{\sum M^i_{\mathrm{negative}}}$$

Two Sample Logo

We also employed the Two Sample Logo (Vacic et al., 2006) web server to calculate and visualize the differences between ubiquitylated fragments from different species. Two Sample Logos can be used to determine statistically significant residues around various active sites, protein modification sites, or to find differences between two groups of sequences that share the same sequence motif.

Sequence Encoding

Compared with the traditional machine learning and statistical computation method, the deep learning approach can extract

features automatically from original data without feature engineering (Schmidhuber, 2015). Thus, transferring the amino acid sequences to quantification vectors, which can be processed by a computer program directly, is important (Hua and Quan, 2016). Word embedding is a set of techniques in natural language processing in which words from a vocabulary are represented as vectors using a large corpus of text as the input.

Generally, there are two main word-embedding techniques used in sentence processing. The first method is embedding layer in neural network (Neishi et al., 2017); the essence of embedding layer is a fully connected neural network, which can map the one-hot sequence to a dimensionally specified vector. Some popular deep learning frameworks have predefined functions for this layer. The process of parameter learning of this method is supervised; the parameters are updated with subsequent layers during the learning process under the supervision of a class label. Several PTM site prediction works are based on this scheme. Another word-embedding technique is Word2vec (Mikolov et al., 2013), where similar vector representations are assigned to the words that appear in similar contexts based on word proximity as gathered from a large corpus of documents. After training on a large corpus of text, the vectors representing many words show interesting and useful contextual properties. The training of word2vec is unsupervised because the class label does not participate in the learning process.

In this work, inspired by pretraining and fine-tuning mechanisms of transfer learning, we first employed the original plant protein sequences as training data and pretrained the embedding layer based on the unsupervised skip-gram algorithm. The optimized embedding layer can map each amino acid from a sequence into a vector. The Euclidean distance of vectors can reflect the relative position information of an amino acid. So, the embedding layer can capture the spatial features of the amino acids in the pretraining process. Then, the optimized parameters are transferred as the initial weights of the embedding layer, and fine tuning is done with the subsequent layers together under the supervision of the label of fragments. By contrast, the traditional word-embedding methods often initialize weights randomly and are trained together subsequently, which may ignore the sequence position information. Compared with traditional word-embedding methods, the proposed scheme is more appropriate for plant ubiquitination site prediction.

We employed the skip-gram method (Du et al., 2018) on the construction of word2vec mapping network. The protein sequences were represented as a collection of counts of n-grams, in which n adjacent amino acids were recognized as a word. Inspired by the idea of Hamid and Friedberg (2018), the length of the gram of 1, 2, 3 was tested in our work, and $n = 2$ was optimal, leading to $20^2 = 400$ bigrams. **Figure 2** simply shows the representation learning for bigrams with the skip-gram training. For each protein sequence, we created two sequences by starting the sequence from the first and second amino acids, so that we can consider all of the overlapping bigrams for a protein sequence. We generated the training instances using a context window of size ± 2, where we took the central word as input and used the surrounding words within the context window as outputs. The neural network architecture for training was used on all of instances, then a 200-dimensional vector for each bigram was generated by the neural network. The trained hidden layer weights were transferred as the initial parameters of the embedding layer in the proposed ubiquitination site prediction model.

Word2vec With Convolutional Neural Network

After sequence encoding, one-dimensional CNN was employed to take the bigram encoding vectors as input and predict the label of this fragment whose lysine in the central position can be ubiquitylated or not. The forward calculation of the CNN deep structure is an automatic feature extraction and selection process in each layer. As shown in **Figure 3**, each bigram maps into a 20-dimensional vector so that a sequence of 31 amino acid residues is represented as a 30×20 matrix, which was denoted as X. The next step is the convolutional layer where the filters were used to extract sequence features from the encoding matrix. The process is denoted as

$$C_1 = \delta_1(W_1 \times X + b_1)$$

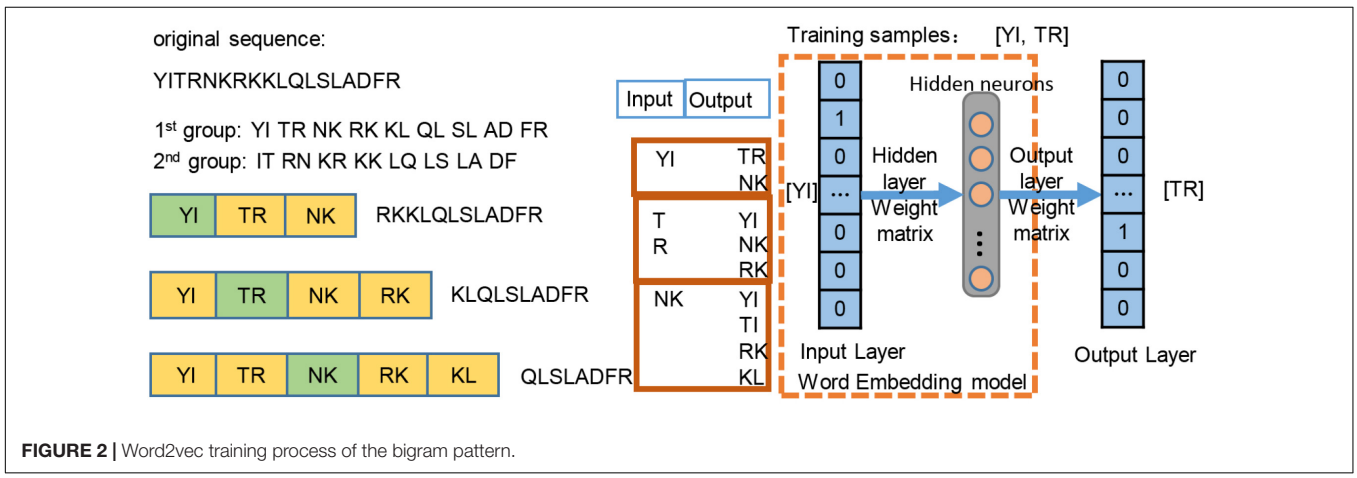


FIGURE 2 | Word2vec training process of the bigram pattern.

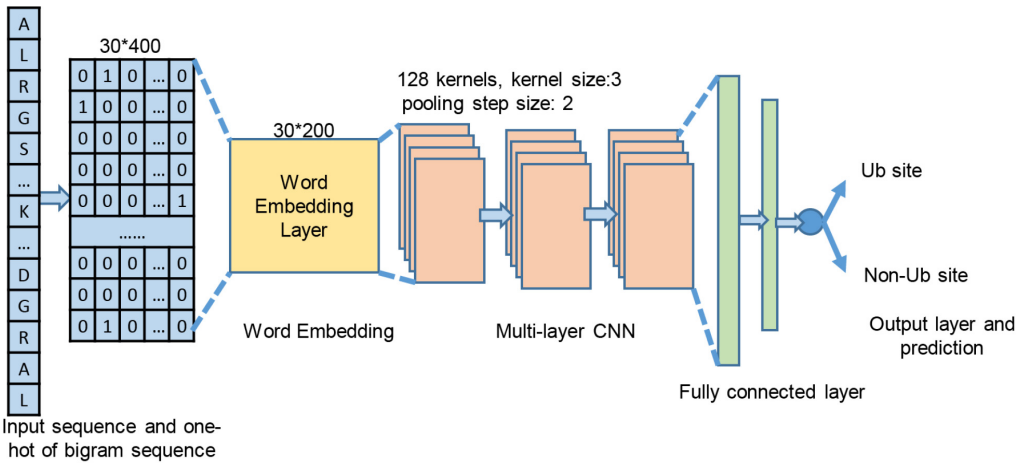


FIGURE 3 | Proposed deep structure for ubiquitination site prediction model.

where δ_1 is the rectified linear (Relu) function, W_1 denotes the weights of the convolution kernel, and b_1 is the bias of this layer. Then, the max pooling function is used for downsampling procedure to reduce the feature dimension.

$$C_{1,\text{out}} = \max \text{pooling}(C_1)$$

The CNN deep structure contained three same sequentially connected blocks, and each block covered a convolution layer with the Relu as its activation function and a max pooling layer. The number of convolution kernels was set as 128, and the convolution kernel size was set at 3. The size of the max pooling windows was 2. Two fully connected layers with 128 and 64 neurons, respectively, are used to integrate features. The output layer contained a single neuron and ends with sigmoid activation to calculate the output x of this layer as

$$\text{Sigmoid}\,(x) = \frac{1}{1+e^{-x}}$$

The backward process of the CNN network is backward propagation, which updates and gets optimal parameters with the following binary cross-entropy loss function.

$$\text{BCE}\left(\hat{y}, y\right) = -\frac{1}{N}\sum_{i=1}^{n}\left[y_i \cdot \log\left(\widehat{y_i}\right) + (1-y_i)\cdot \log\left(1-\widehat{y_i}\right)\right]$$

During the training of the CNN models, the dropout units (the drop rate was set at 0.5) were added after each max pooling layer in the convolutional layer, which are usually required for generalization on unseen data and to avoid overfitting.

Implementation and Training Parameters

The proposed model was achieved through the Keras framework under the Python language. We set the initial learning rate as 0.001, and the RMS prop optimization method was used with $\beta = 0.9$. We initialized the weights of the convolutional network randomly with a Gaussian distribution ($\mu = 0, \sigma = 0.01$). The batch size is 500, and 120 epochs were conducted for each training. All the experiments were performed on a server equipped with Geforce RTX 2080 Ti.

RESULTS

Comparison of Features Between Species

Amino Acid Composition

Figure 4 provides the average of positive and negative segments, respectively, and a histogram for each species was plotted. We can analyze the amino acid composition differences between the positive and negative segments to show different patterns of these spices. For the plant subset, the average percentage of arginine (R) in ubiquitylated protein fragments is doubled in non-ubiquitylated protein segments. By contrast, the arginine differences between ubiquitylated and non-ubiquitylated segments in animals and fungi are not obvious, although the figure of positive samples is 0.7% higher than the negative samples in animals. For animals and fungi, the average percentage of lysine (K) in the positive protein segments is about 1% higher than in the negative samples, and this difference is not obvious in plant samples. What is more, the percentage of leucine (L) in ubiquitylated proteins of animal is 1% higher than in non-ubiquitylated samples; this finding is contrary in plants and fungi. So, the amino acid composition shows really different patterns in different species.

Amino Acid Pairwise Composition

As shown in **Figure 5**, the blue pixels mean this dipeptide is more likely to appear around ubiquitylated lysine than in non-ubiquitylated lysine, while the red means it is less likely to appear in ubiquitylated fragments than in the negative samples. The darker the color, the greater the difference. For the ubiquitination of plants, cysteine (C) is less often composed with other amino acids, such as glycine (G), methionine (M), serine (S), and

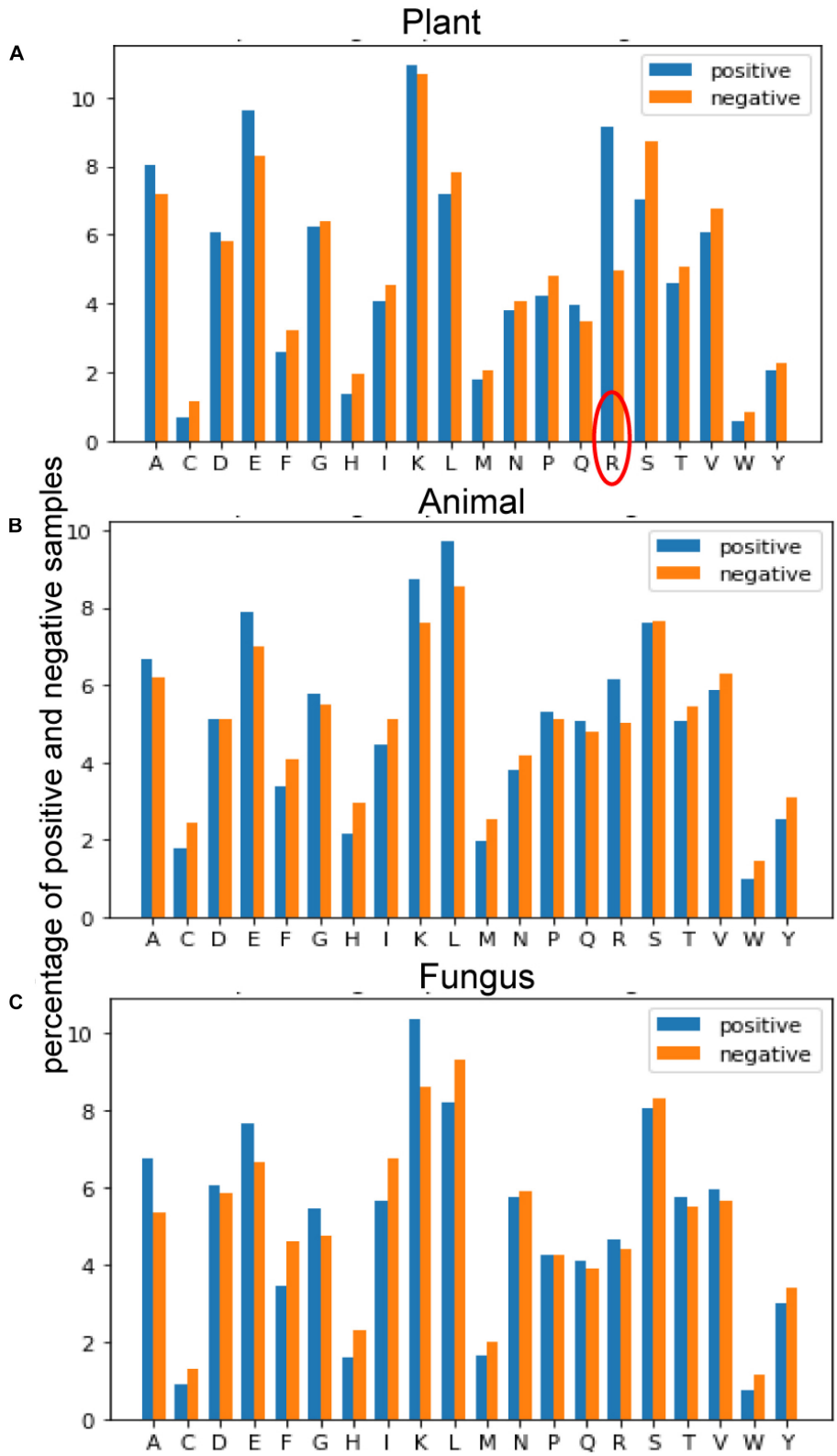


FIGURE 4 | Comparison of the amino acid composition (AAC) features between three species. **(A)** Comparison of AAC features between positive and negative samples of plants. **(B)** Comparison of AAC features between positive and negative samples of plants. **(C)** Comparison of AAC features between positive and negative samples of plants.

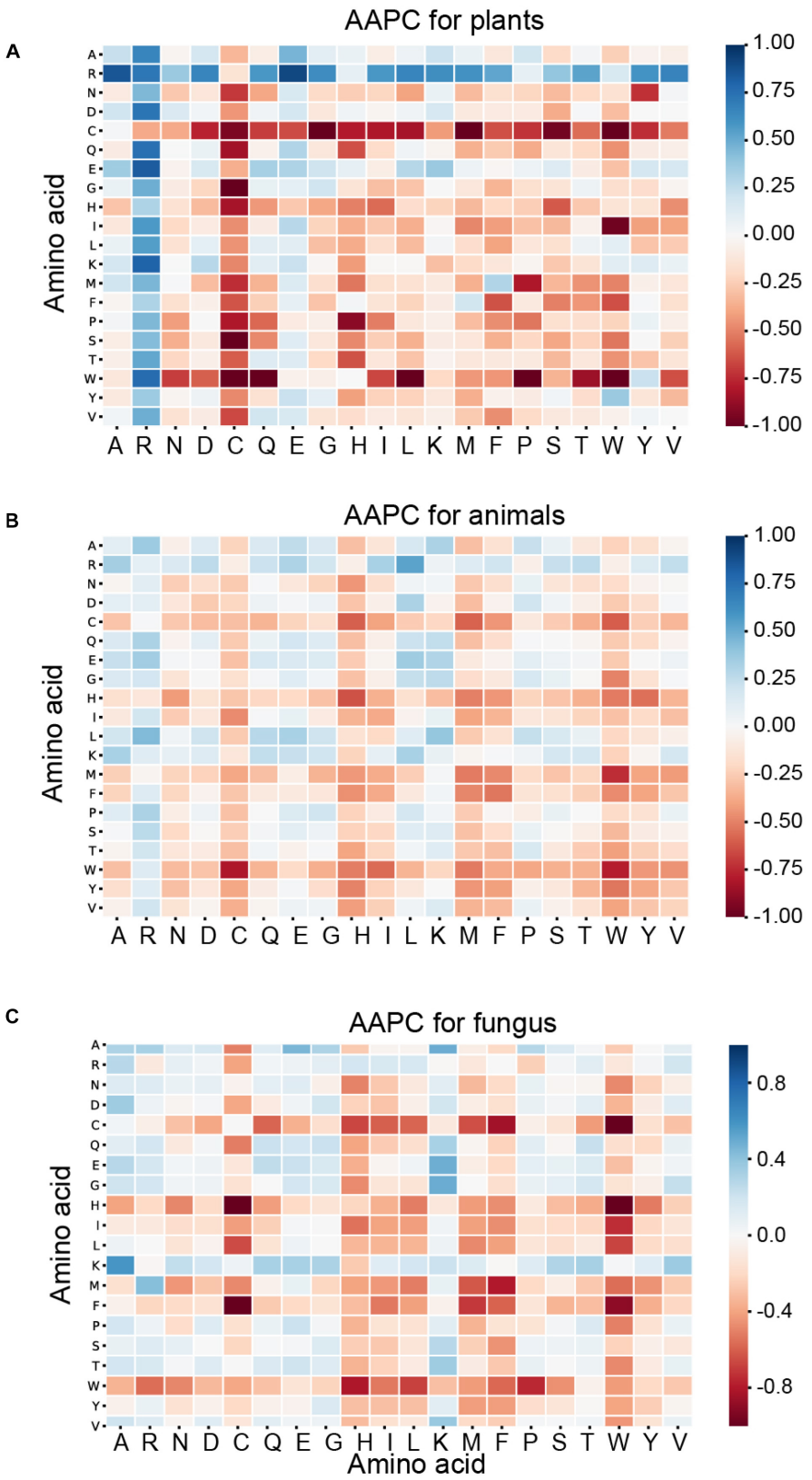


FIGURE 5 | Heatmaps for the amino acid pairwise composition (AAPC) features of three species. **(A)** Heatmap for the AAPC features of plants. **(B)** Heatmap for the AAPC features of animals. **(C)** Heatmap for the AAPC features of fungi.

tryptophan (W). The $P_{r,s}$ of these dipeptides are less than −0.75, which denotes that the distribution of these dipeptides presents obvious differences between the ubiquitylated and non-ubiquitylated fragments in the plant subset. In addition, the pairs that contain arginine (R), especially with alanine (A) and glutamic acid (E), are more likely to appear around the ubiquitylated lysine with $P_{r,s}$ of more than 0.5. However, these phenomena above do not appear in the animal and fungus subsets. For the animal subset, the value of $P_{r,s}$ for the majority of the amino acid combinations range from −0.25 to 0.25, which means that there are no obvious differences between the positive and negative samples, expect that tryptophan (W), combined with cysteine (C), and methionine (M), is less likely to appear in ubiquitylated peptides than in non-ubiquitylated samples. For the fungus subset, cysteine (C), combined with methionine (M), and histidine (H), as well as tryptophan (W), combined with cysteine (C), histidine (H), and phenylalanine (F), are less likely to appear in ubiquitylated peptides than in non-ubiquitylated samples. The $P_{r,s}$ of these amino acid combinations are less than −0.7. The statistical differences of the AAPC feature between the ubiquitylated and non-ubiquitylated fragments show very different patterns in three different species.

Positional Weighted Matrix

As shown in **Figure 6**, blue means that the amino acid is more likely to appear in this position of ubiquitylated fragments, and red means this position is less likely to find this amino acid. For the ubiquitylated segments of the plant, it is more likely that arginine (R) will be found around the ubiquitylated lysine, especially on the 1st to 8th and −9th to −5th positions. In addition, it is clear that histidine (H), cysteine (C), and tryptophan (W) hardly appeared around the ubiquitylated lysine. The feature patterns in fungi and animals are different. For fungi, there is also some lysine (K) often appearing in the preorder of the ubiquitylated lysine, especially on the −9th to −1st position with M more than 0.75. However, lysine (K), which is followed by another lysine (K) in the next position usually is not ubiquitylated with M less than −0.75. For animals, we can find that the glutamic acid (E) more likely appeared on the −1st and −2nd positions near the ubiquitylated lysine with M of more than 0.75, but it less likely to appear on 1st and 2nd positions. The features of specific amino acid distribution in each position also differ in different species.

Two Sample Logo

We employ the Two Sample Logo to show the differences of amino acid distribution in each position between ubiquitylated fragments from different species. The larger fonts denote the amino acid that is more likely to appear in this position with statistical significance. As shown in **Figure 7A**, we set the plant ubiquitylated fragments as positive samples and the animal ubiquitylated fragments as negative samples. We can see that more arginine (R), glutamic acid (E), aspartic acid (D), and alanine (A) appeared around the ubiquitylated lysine in the plants than in the animals, while it is less likely to find leucine (L). Then we set the plant ubiquitylated fragments as positive samples and the animal ubiquitylated fragments as negative samples, which are shown in **Figure 7B**. It is obvious that arginine (R) and glutamic acid (E) are more likely to appear around the ubiquitylated lysine in plants. As for the comparison of ubiquitylated fragments between animals and fungi, there were no obvious patterns except that there is more leucine (L) around the ubiquitylated lysine (**Figure 7C**).

According to the analyses above, the sequence features of the ubiquitylated fragments are really different between these three species. It is significant to build a ubiquitylation site prediction model for a single species, which can avoid the interference of feature differences from other species.

Model Performance Evaluation

The proposed word embedding and CNN-based ubiquitination prediction model is evaluated through a validation test scheme. A 10-fold cross-validation is carried out on the training set for the fine-tuning of the hyper-parameters, as well as for evaluating the reliability of the model. In order to make the experiment results statistically significant, five repeated runs were conducted for each fold cross validation; the mean and standard deviation of the 50 results were regarded as the final result. The independent testing set was used for generalization evaluation and performance comparison with the baseline method. The confusion matrix of the prediction model is shown in **Table 1**, and the performance evolution indexes are defined as follows:

(a) Accuracy that indicates the proportion of correctly classified subjects among the whole subset

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

(b) Precision that quantifies the proportion of samples correctly classified among the classification

$$\text{Precision} = \frac{TP}{TP + FP}$$

(c) Recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances

$$\text{Recall} = \frac{TP}{TP + FN}$$

(d) F-score considers both the precision and recall and evaluate the model performance synthetically

$$F - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

We first compared the proposed model performance with different tuning options through the 10-fold cross-validation scheme. Mean and standard deviation results of the cross validation are calculated, and the comparison results are shown in **Table 2**. The best performance with a mean accuracy of 78.1% and an F-score of 0.782 is given by the proposed model, which combines the transfer word-embedding mechanism and multilayer CNN. By contrast, the traditional one-hot sequence encoding method combined with a 2D CNN classifier obtains the worst performance with only a mean accuracy of 62.3% and an

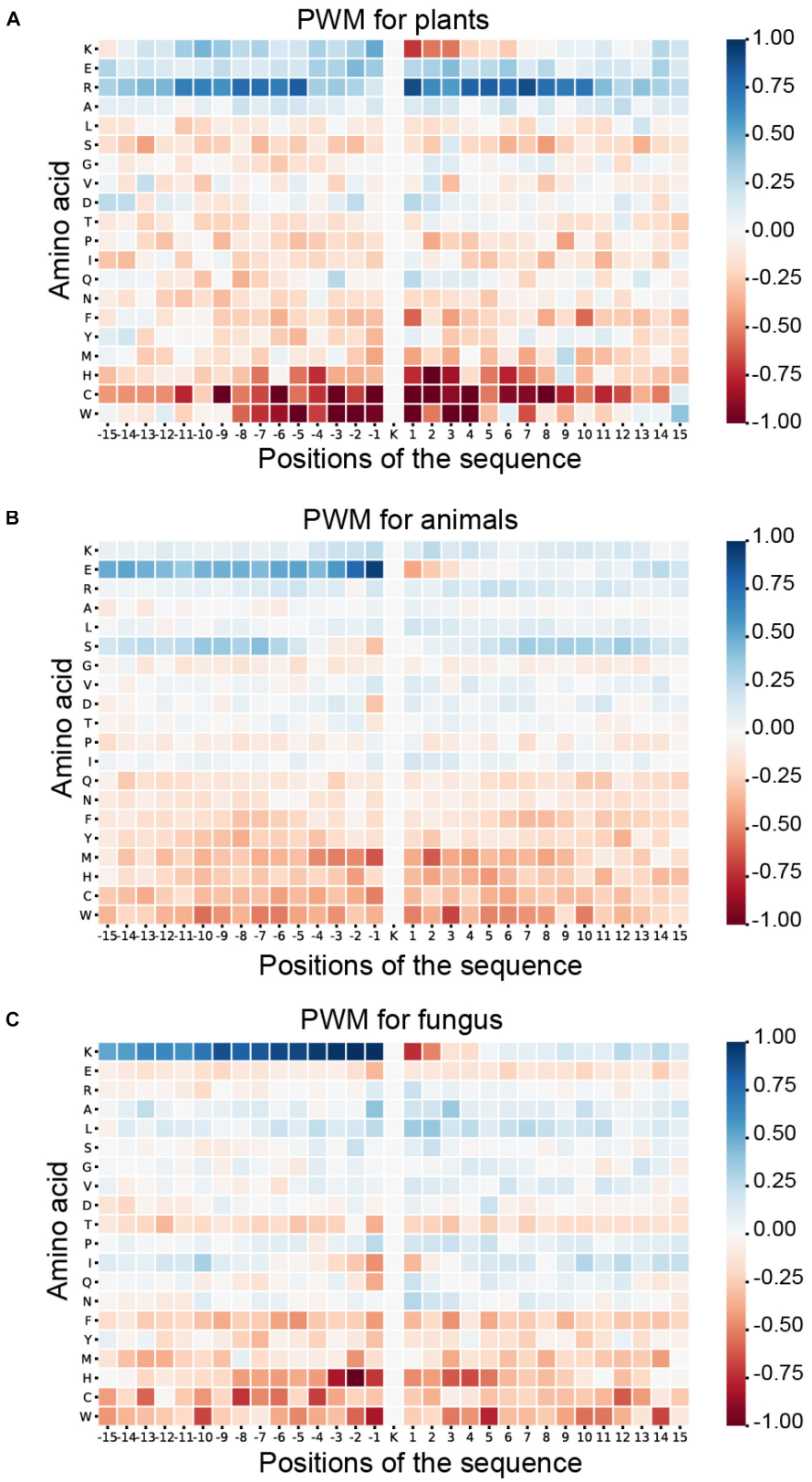


FIGURE 6 | Heatmaps for the positional weighted matrix (PWM) features of three species. **(A)** Heatmaps for the PWM features of plants. **(B)** Heatmaps for the PWM features of animals. **(C)** Heatmaps for the PWM features of fungi.

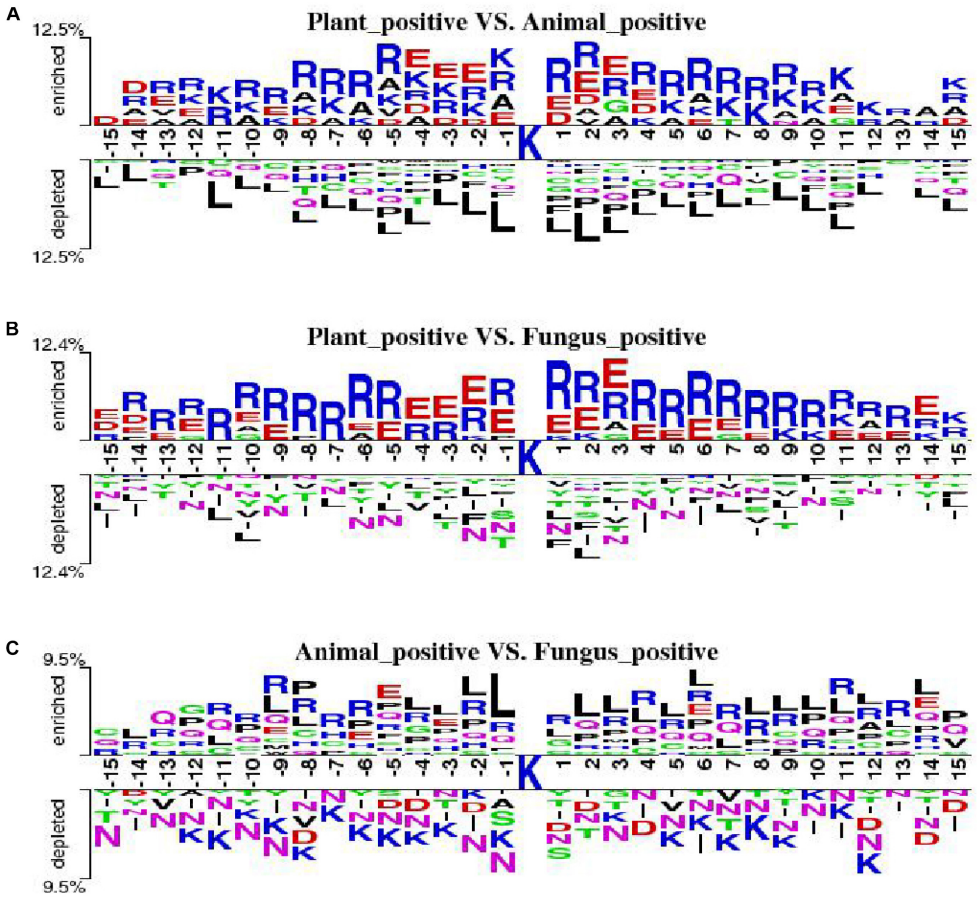


FIGURE 7 | Comparison of the Two Sample Logo of three species. **(A)** Two Sample Logo of positive samples between plants and animals. **(B)** Two Sample Logo of positive samples between plants and fungi. **(C)** Two Sample Logo of positive samples between animals and fungi.

F-score of 0.647. This is mainly because the one-hot encoding matrix is sparse, and the conventional filters cannot capture useful sequence features. The pretrained word2vec encoding model without supervised weights updating also received a poor performance with a mean accuracy of 68.5% and an F-score of 0.6771. The word2vec model was trained on original plant protein sequences, which learned the amino acid bigram patterns of plants. However, without the fine-tuning process, the fixed weights cannot be adjusted to fit the ubiquitination site prediction task well. In addition, the supervised embedding layer with randomly initialized parameters also got a general performance; the effort of pretraining is obvious in ubiquitination site prediction in our proposed method. What is more, our results suggest that the recurrent neural network (RNN) does not contribute much to ubiquitination site prediction; this may because the distant sequence correlation modeling is not useful for this task.

Independent Testing Performance

A series of sequence features were extracted for modeling, including AAC, AAPC, the CKSAAPs, as well as the position-specific scoring matrix (PSSM). Our experiments indicated that the random forest (RF) model outperform other popular algorithms on all these predefined features. **Table 3** shows the comparison between the proposed model and traditional feature-based random forest method on the testing set. The proposed model achieved the best performance with a mean accuracy of 75.6% and an F-score of 0.749. The random forest model also achieved an acceptable performance based on features of k-spaced amino acid pairs, with a mean accuracy of 73.6% and an F-score of 0.717. The PSSM represented the evolutionary profile of the protein sequence; the RF based on the PSSM features can achieved a mean accuracy of 71.1% and an F-score of 0.6942. Then as shown in **Figure 8**, we plotted the ROC curve with AUC of these RF-based model and our model. The proposed model is obvious, overall, in terms of the ROC curve with an 0.81 AUC,

TABLE 1 | Confusion matrix of ubiquitylated site prediction model.

	Predicted positive (Ub)	Predictive negative (non-Ub)
Actual positive (Ub)	True positive (TP)	False negative (FN)
Actual negative (non-Ub)	False positive (FP)	True negative (TN)

TABLE 2 | Cross validation performance comparison between different deep structures and feature encondings.

Model tuning	Accuracy	Precision	Recall	F-score
One-hot encoding + 2D convolutional neural network (CNN)	0.623 ± 0.037	0.662 ± 0.028	0.636 ± 0.019	0.647 ± 0.021
Embedding layer + CNN	0.732 ± 0.006	0.745 ± 0.011	0.692 ± 0.024	0.716 ± 0.029
Fixed word2vec + CNN	0.685 ± 0.024	0.701 ± 0.019	0.653 ± 0.015	0.677 ± 0.022
Transfer embedding + recurrent neural network (RNN)	0.743 ± 0.012	0.749 ± 0.004	0.716 ± 0.017	0.729 ± 0.015
Proposed method	0.782 ± 0.008	0.791 ± 0.013	0.785 ± 0.011	0.782 ± 0.016

TABLE 3 | Performance comparison between different methods on the testing set.

Method	Accuracy	Precision	Recall	F-score
Random forest (RF) with amino acid composition (AAC)	0.703 ± 0.012	0.685 ± 0.026	0.703 ± 0.019	0.694 ± 0.022
RF with amino acid pairwise composition (AAPC)	0.711 ± 0.008	0.706 ± 0.017	0.679 ± 0.021	0.692 ± 0.031
RF with *k*-spaced AAP (*k* = 5)	0.736 ± 0.006	0.721 ± 0.009	0.714 ± 0.015	0.717 ± 0.019
RF with position-specific scoring matrices (PSMM)	0.722 ± 0.014	0.718 ± 0.008	0.706 ± 0.025	0.713 ± 0.018
Proposed method	0.756 ± 0.006	0.733 ± 0.015	0.767 ± 0.017	0.749 ± 0.009

which indicates that the developed classifier has high confidence on plant ubiquitination site prediction.

In order to evaluate the generalization of the proposed model, we also collected data from the dbPTM and iPTMnet databases as an extra testing set. The dbPTM (Huang et al., 2019) and iPTMnet (Huang et al., 2018) contain 107 and 50 proteins of *A. thaliana*, respectively. The CD-HIT, with 30% identity, was employed to remove the redundant protein fragments and eliminate homology bias with the PLMD training data. Finally, 91 positive and 217 negative fragments were used for extra testing. The optimal model in cross validation achieved an accuracy of 74.2%, precision of 73.1%, recall of 73.7%, and F-score of 0.733. The proposed model can also achieve equal performance on other datasets.

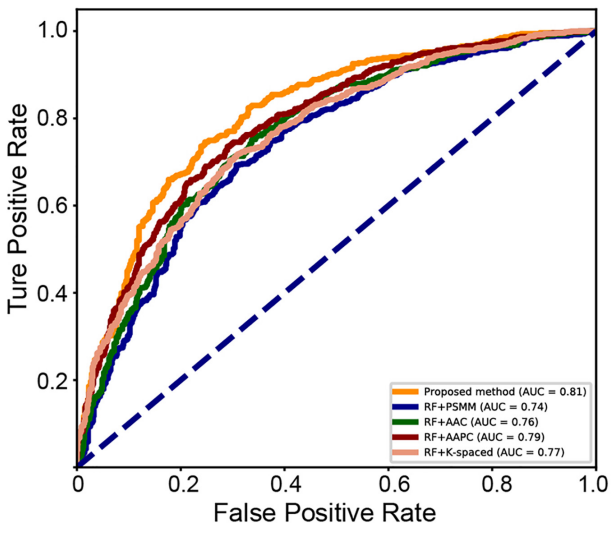


FIGURE 8 | ROC curve of the different methods on the testing set.

Comparison With Other Prediction Tools

We compared the performance of the proposed method with other popular ubiquitylation prediction tools on the independent set. For UbPred (Xiang et al., 2013), iUbiq-Lys (Qiu et al., 2015), and Ubisite (Huang et al., 2016), we uploaded our testing data to their website and counted the confusion matrix of output results to compute the performance indexes. For the Deep ubiquitylation (He et al., 2018) and DeepUbi (Fu et al., 2019), we reproduced their proposed structure with Keras, as well as training steps through our data, then calculated the evaluation indexes. As shown in **Table 4**, our proposed method achieved a balanced and reasonable performance with a mean precision of 73.3%, recall of 76.7%, and F-score with 0.749, although it only achieved a mean accuracy of 75.6%. It can be found that the iUbiq-Lys and Ubisite yielded a high recall and a poor precision, which means that these tools are more likely to classify the suspected samples as positive. Compared with Deep ubiquitylation, the first deep learning-based tool, our method achieved a better overall performance, which is mainly because the word-embedding scheme is more effective to extract the sequence features. The proposed method also outperformed the DeepUbi to some extent because the transfer learning-based method can capture the sequence pattern of plant proteins with word2vec model and the weights of embedding layer just fine-tuned around the pretrained value. In addition, the DeepUbi did not achieve the performance they claimed; this is mainly because the testing experiments are curried out on our plant data with on a small scale. Their proposed structure may need a larger training set to achieve optimal performance due to their training of the embedding layer from a random initial value. Overall, compared with popular tools and methods, our proposed model achieved a better performance on plant ubiquitylation site prediction.

Then predictions were conducted on two types of single plant protein: one contains ubiquitylated substrate sites, and the other has no ubiquitylated sites. The proposed model was compared with three popular ubiquitylation prediction tools, which provide websites for sequence input. The ubiquitylated

TABLE 4 | Performance comparison with other prediction tools.

Tool	Accuracy	Precision	Recall	F-score
UbPred (Xiang et al., 2013)	0.719	0.626	0.738	0.678
iUbiq-Lys (Qiu et al., 2015)	0.799	0.563	0.837	0.671
Ubisite (Huang et al., 2016)	0.752	0.596	0.794	0.681
Deep Ub (He et al., 2018)	0.683 ± 0.021	0.674 ± 0.018	0.703 ± 0.011	0.687 ± 0.024
DeepUbi (Fu et al., 2019)	0.739 ± 0.014	0.733 ± 0.011	0.741 ± 0.021	0.734 ± 0.011
Proposed method	0.756 ± 0.006	0.733 ± 0.015	0.767 ± 0.017	0.749 ± 0.009

TABLE 5 | Performance comparison with other tools on two types of single protein.

UniProt AC	Organism	Sequence length	Number of lysine	Reported ubiquitylated sites	Predicted ubiquitylated sites	
					Tools	Results
O23063	*Arabidopsis thaliana* (Mouse-ear cress)	364	47	142; 222; 225	iUbiq-Lys	3; 4; 103; 217; 225; 363
					UbPred	98; 142; 197
					Ubisite	142; 225; 265; 297
					Proposed model	142; 222; 225; 363
O03042	*Arabidopsis thaliana* (Mouse-ear cress)	479	24	None	iUbiq-Lys	None
					UbPred	8; 32; 201; 356; 474
					Ubisite	474
					Proposed model	None

protein was selected from an independent testing set randomly, and the protein that does not contain a ubiquitylated substrate site was selected from Uniport with no ubiquitylation sites reported. As shown in **Table 5**, the protein with Uniport AC O23063 contains 47 lysine and the positions of 142, 222, 225 are ubiquitylated (Walton et al., 2016). The iUbiq-Lys predicted five ubiquitylated sites, and only one is correct. The UbPred predicted one ubiquitylated site with other two false positive results. The Ubisite identified two sites successfully, while the proposed model can predict all the ubiquitylated sites correctly. It should be noted that the 363 position predicted by the proposed model is a false-positive sample; the performance of the proposed model still has room for improvement for some fragments. The protein with Uniport AC O03042 contains 24 lysine but no ubiquitylated site among them. The UbPred and Ubisite provided wrong predictions, while the iUbiq-Lys and the proposed model can classify them as non-ubiquitylated sites.

The proposed model outperforms traditional machine learning and deep structure mainly because of its two novel characteristics. First, contrastive analyses found pattern differences of ubiquitylated fragments between the three species. Modeling for proteins from a single species can avoid the interference of feature differences from other species. Second, the transfer learning mechanism was employed to pretrain the embedding layer through the original plant protein sequence by the word2vec method, which can capture the sequence features of plant proteins and vectorize them. The Euclidean distance of vectors can reflect the relative position information of the amino acids. The embedding layer can capture the spatial features of amino acids in the pretraining process. So, the model is appropriate for the plant ubiquitination site prediction and achieved a better performance.

CONCLUSION

In this work, we analyzed the sequence features of ubiquitylated protein from plants, animals, and fungi, respectively, then indicated the feature pattern differences between these features. We found that the amino acid distribution around the ubiquitylated lysine of plants differ from other species obviously, such as the clustering of arginine (*R*). The species of the plant was selected as the research target for modeling. A novel transfer learning-based word-embedding model training scheme was proposed. The original plant protein sequence was used for pretraining of the word2vec network through the skip-gram model, then the optimized parameter transfer as the initial weights of the embedding layer, fine-tuning with the subsequent layers together. The multilayer CNN was employed as a classifier and achieved acceptable performance for plant ubiquitination site prediction. Compared with related prediction tools, our method performs excellent suitability for plant ubiquitination site prediction. Considering the pattern differences between different species, in future work, we will integrate more data and establish species-specialized tools for ubiquitination site prediction.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/ supplementary material.

AUTHOR CONTRIBUTIONS

T-YL conceived and headed this project. HW acquired the data, conducted the modeling work, and performed the experiments. ZW analyzed features and helped in the model evaluation works. ZL helped in the data collection and curation. All authors participated in writing or revising the manuscript.

FUNDING

This research was funded by the Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen, China, and the Ganghong Young Scholar Development Fund of Shenzhen Ganghong Group Co., Ltd.

REFERENCES

Cai, Y., Huang, T., Hu, L., Shi, X., Xie, L., and Li, Y. (2012). Prediction of lysine ubiquitination with mRMR feature selection and analysis. *Amino Acids* 42, 1387–1395. doi: 10.1007/s00726-011-0835-0

Chen, Z., Zhou, Y., Song, J., and Zhang, Z. (2013). hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. *Biochim. Biophys. Acta* 1834, 1461–1467. doi: 10.1016/j.bbapap.2013.04.006

Du, L., Wang, Y., Song, G., Lu, Z., and Wang, J. (2018). "Dynamic network embedding: an extended approach for skip-gram based network embedding," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, New York, NY: IJCAI, 2086–2092.

Fu, H., Yang, Y., Wang, X., Wang, H., and Xu, Y. (2019). DeepUbi: a deep learning framework for prediction of ubiquitination sites in proteins. *BMC Bioinformatics* 20:86. doi: 10.1186/s12859-019-2677-9

Hamid, M.-N., and Friedberg, I. (2018). Identifying antimicrobial peptides using word embedding with deep recurrent neural networks. *Bioinformatics* 35, 2009–2016. doi: 10.1093/bioinformatics/bty937

He, F., Wang, R., Li, J., Bao, L., and Zhao, X. (2018). Large-scale prediction of protein ubiquitination sites using a multimodal deep architecture. *BMC Syst. Biol.* 12(Suppl. 6):109. doi: 10.1186/s12918-018-0628-0

Hoeller, D., Hecker, C.-M., and Dikic, I. (2006). Ubiquitin and ubiquitin-like proteins in cancer pathogenesis. *Na. Rev. Cancer* 6, 776–788. doi: 10.1038/nrc1994

Hua, L., and Quan, C. (2016). A shortest dependency path based convolutional neural network for protein-protein relation extraction. *Biomed. Res. Int.* 2016:8479587.

Huang, C. H., Su, M. G., Kao, H. J., Jhong, J. H., Weng, S. L., and Lee, T. Y. (2016). UbiSite: incorporating two-layered machine learning method with substrate motifs to predict ubiquitin-conjugation site on lysines. *BMC Syst. Biol.* 10(Suppl. 1):6. doi: 10.1186/s12918-015-0246-z

Huang, H., Arighi, C. N., Ross, K. E., Ren, J., Li, G., Chen, S.-C., et al. (2018). iPTMnet: an integrated resource for protein post-translational modification network discovery. *Nucleic Acids Res.* 46, D542–D550.

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

Jyun-Rong, W., Wen-Lin, H., Ming-Ju, T., Kai-Ti, H., Hui-Ling, H., and Shinn-Ying, H. (2016). ESA-UbiSite: accurate prediction of human ubiquitination sites by identifying a set of effective negatives. *Bioinformatics* 33, 661–668.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Lu, D., Lin, W., Gao, X., Wu, S., Cheng, C., Avila, J., et al. (2011). Direct ubiquitination of pattern recognition receptor FLS2 attenuates plant innate immunity. *Science* 332, 1439–1442. doi: 10.1126/science.1204903

Marino, D., and Rivas, P. S. (2012). Ubiquitination during plant immune signaling. *Plant Physiol.* 160, 15–27. doi: 10.1104/pp.112.199281

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*. [preprint]. Avaliable at: https://arxiv.org/abs/1301.3781 (accessed June 13, 2020).

Neishi, M., Sakuma, J., Tohda, S., Ishiwatari, S., Yoshinaga, N., and Toyoda, M. (2017). "A bag of useful tricks for practical neural machine translation: embedding layer initialization and large batch size," in *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*), Taipei, 99–109.

Nguyen, V. N., Huang, K. Y., Huang, C. H., Chang, T. H., Bretaña, N., Lai, K., et al. (2015). Characterization and identification of ubiquitin conjugation sites with E3 ligase recognition specificities. *BMC Bioinformatics* 16(Suppl.1):S1. doi: 10.1186/1471-2105-16-S1-S1

Nguyen, V. N., Huang, K. Y., Huang, C. H., Lai, K. R., and Lee, T. Y. (2016). A new scheme to characterize and identify protein ubiquitination sites. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 14, 393–403. doi: 10.1109/tcbb.2016.2520939

Popovic, D., Vucic, D., and Dikic, I. (2014). Ubiquitination in disease pathogenesis and treatment. *Nat. Med.* 20, 1242–1253. doi: 10.1038/nm.3739

Qiu, W.-R., Xiao, X., Lin, W.-Z., and Chou, K.-C. (2015). iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. *J. Biomol. Struc. Dyn.* 33, 1731–1742. doi: 10.1080/07391102.2014.968875

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 Struc. Funct. Bioinform.* 78, 365–380. doi: 10.1002/prot.22555

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Tu, Y., Chen, C., Pan, J., Xu, J., and Wang, C. Y. (2012). The ubiquitin proteasome pathway (UPP) in the regulation of cell cycle control and DNA damage repair and its implication in tumorigenesis. *Int. J. Clin. Exp. Pathol.* 5, 726–738.

Tung, C.-W., and Ho, S.-Y. (2008). Computational identification of ubiquitylation sites from protein sequences. *BMC Bioinformatics* 9:310. doi: 10.1186/1471-2105-9-310

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

Walton, A., Stes, E., Cybulski, N., Van Bel, M., Iñigo, S., Durand, A. N., et al. (2016). It's time for Some "Site"-Seeing: novel tools to monitor the ubiquitin landscape in arabidopsis thaliana. *Plant Cell* 28, 6–16. doi: 10.1105/tpc.15.00878

Weissman, A. M. (2001). Themes and variations on ubiquitylation. *Nat. Rev. Mol. Cell Biol.* 2, 169–178. doi: 10.1038/35056563

Xiang, C., Qiu, J. D., Shi, S. P., Suo, S. B., Huang, S. Y., and Liang, R. P. (2013). Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites. *Bioinformatics* 13, 1614–1622. doi: 10.1093/bioinformatics/btt196

Xu, H., Zhou, J., Lin, S., Deng, W., Zhang, Y., and Xue, Y. (2017). PLMD: An updated data resource of protein lysine modifications. *J. Genet. Genomics* 44, 243–250. doi: 10.1016/j.jgg.2017.03.007

Yamada, T., Murata, D., Adachi, Y., Itoh, K., Kameoka, S., Igarashi, A., et al. (2018). Mitochondrial stasis reveals p62-mediated ubiquitination in Parkin-independent mitophagy and mitigates nonalcoholic fatty liver disease. *Cell Metab.* 28, 588.e5–604.e5.

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.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 10 November 2020
doi: 10.3389/fcell.2020.593661

PTMsnp: A Web Server for the Identification of Driver Mutations That Affect Protein Post-translational Modification

Di Peng[1†], Huiqin Li[1†], Bosu Hu[1†], Hongwan Zhang[2], Li Chen[1], Shaofeng Lin[3], Zhixiang Zuo[2], Yu Xue[3], Jian Ren[1,2] and Yubin Xie[1*]*

[1] Precision Medicine Institute, The First Affiliated Hospital, School of Life Sciences, Sun Yat-sen University, Guangzhou, China, [2] State Key Laboratory of Oncology in South China, Cancer Center, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University, Guangzhou, China, [3] Key Laboratory of Molecular Biophysics of Ministry of Education, Hubei Bioinformatics and Molecular Imaging Key Laboratory, Center for Artificial Intelligence Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China

OPEN ACCESS

Edited by:
Eleonora Napoli,
University of California, Davis,
United States

Reviewed by:
Maria Giuseppina Miano,
Institute of Genetics and Biophysics (CNR), Italy
Przemko Tylzanowski,
KU Leuven, Belgium

***Correspondence:**
Jian Ren
renjian@sysucc.org.cn
Yubin Xie
xieyb6@mail.sysu.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: 11 August 2020
Accepted: 21 October 2020
Published: 10 November 2020

Citation:
Peng D, Li H, Hu B, Zhang H, Chen L, Lin S, Zuo Z, Xue Y, Ren J and Xie Y (2020) PTMsnp: A Web Server for the Identification of Driver Mutations That Affect Protein Post-translational Modification. Front. Cell Dev. Biol. 8:593661. doi: 10.3389/fcell.2020.593661

High-throughput sequencing technologies have identified millions of genetic mutations in multiple human diseases. However, the interpretation of the pathogenesis of these mutations and the discovery of driver genes that dominate disease progression is still a major challenge. Combining functional features such as protein post-translational modification (PTM) with genetic mutations is an effective way to predict such alterations. Here, we present PTMsnp, a web server that implements a Bayesian hierarchical model to identify driver genetic mutations targeting PTM sites. PTMsnp accepts genetic mutations in a standard variant call format or tabular format as input and outputs several interactive charts of PTM-related mutations that potentially affect PTMs. Additional functional annotations are performed to evaluate the impact of PTM-related mutations on protein structure and function, as well as to classify variants relevant to Mendelian disease. A total of 4,11,574 modification sites from 33 different types of PTMs and 1,776,848 somatic mutations from TCGA across 33 different cancer types are integrated into the web server, enabling identification of candidate cancer driver genes based on PTM. Applications of PTMsnp to the cancer cohorts and a GWAS dataset of type 2 diabetes identified a set of potential drivers together with several known disease-related genes, indicating its reliability in distinguishing disease-related mutations and providing potential molecular targets for new therapeutic strategies. PTMsnp is freely available at: http://ptmsnp.renlab.org.

Keywords: protein post-translational modification, genetic mutations, Bayesian hierarchical model, driver genes, disease

INTRODUCTION

Large-scale genome sequencing has uncovered a complex landscape of genetic mutations in multiple patient populations. A major goal of these sequencing projects is to characterize a few disease-related mutations from the majority of neutral passenger mutations. Currently, the most commonly used strategy to prioritize mutations is the frequency-based approach, such as MutSigCV (Lawrence et al., 2013), MuSiC (Dees et al., 2012), and other methods

(Youn and Simon, 2011). These tools can reveal a number of potential driver genes that carry recurrent mutations in a given disease cohort. However, the known driver genes identified from those frequency-based strategies are not sufficient to explain the diverse mechanisms of disease progression. Therefore, several approaches that not only consider recurrent mutations but also combine other functional features, such as evolutionary conservation (Reva et al., 2011), known pathway annotation (Wendl et al., 2011) and protein-protein interaction networks (Vandin et al., 2011; Ciriello et al., 2012), have been proposed.

Among those functional features, one of the most critical factors that can be used in driver gene identification is protein post-translational modifications (PTMs). As key mechanisms to increase proteomic diversity, PTMs can regulate almost all physiological and biochemical processes in mammalian cells. Thus, genetic mutations that occur specifically around the PTM sites (also known as PTM-related mutations) may potentially alter protein functions and disturb regulatory pathways *in vivo*, leading to the development of certain serious diseases, such as cancers. A previous study has reported that mutation of SUMO-conjugated sites in androgen receptor (AR) may result in an increase of AR transcriptional activity, and hence promoting cell proliferation and hypoxia-induced angiogenesis in Prostate cancer (Lin et al., 2004). Meanwhile, experiments have also shown that oncogenic variants altering S768 phosphorylation of EGFR increase its catalytic activity, and S768I mutation can drive tumorigenesis by disrupting EGFR autophosphorylation and rewiring downstream signaling pathways (Huang L. C. et al., 2018). In addition to cancer, Martin et al. have reported that the G553E mutation on huntingtin (HTT) protein can abrogate its post-translational myristoylation and induce cellular toxicity of the protein in cellulo, consequently causing Huntington disease (Martin et al., 2018).

In light of the significant impact of PTM-related mutations on human diseases, several databases have been developed to curate mutations that may potentially affect PTMs. For example, dbPTM collected a subset of PTM-disease associations based on disease-associated non-synonymous SNPs from dbSNP in its 2019 updated version (Huang et al., 2019). Similarly, PhosphpSitePlus provided PTMVar dataset to characterize PTMs that overlap with disease-associated genetic variants and polymorphisms (Hornbeck et al., 2015). Using a similar strategy, other databases such as iPTMnet (Huang H. et al., 2018), PRISMOID (Li et al., 2020), and PTM-SNPs (Kim et al., 2015) were also reported in recent publications. In considering the false positive errors that introduced by the direct mapping of disease-related mutations to PTM sites when deriving disease-related PTM mutations, several studies using predictive tools to extract PTM-related mutations were proposed. For instance, ActiveDriver revealed a set of candidate cancer driver genes harboring mutation hotspots proximal to known phosphorylation, acetylation and ubiquitination sites that may cause the dysfunction of PTM-related mechanisms (Reimand and Bader, 2013; Reimand et al., 2013; Narayan et al., 2016). Besides, MIMP is a machine learning method to predict whether single-nucleotide variants (SNVs) can disrupt existing phosphorylation sites or create new sites (Wagih et al., 2015). Using the MIMP method, ActiveDriverDB is established for collecting human disease mutations and genetic variants that may potentially alter four types of PTMs (Krassowski et al., 2018). In addition, AWESOME utilized 20 PTM prediction tools to predict whether a SNP could change the PTMs level of six common PTM types in a specific protein (Yang et al., 2019). Besides, Simpson et al. developed DeltaScansite to assess the impact of mutations in the flanking regions of phosphosites (Simpson et al., 2019).

Although these reported methods have provided abundant resources of PTM-related mutations, limitations are still existing. First of all, the current methods carried out mutation analysis for one or a few common PTM types, and most other PTM types cannot be covered, thus losing a large amount of PTM-related mutation information. Secondly, most of methods (except ActiveDriver) only consider the impact of mutations on PTM sites alone, and are not associated with specific disease phenotypes, which may preserve a lot of passenger mutations that play a neutral role in disease development. Meanwhile, ActiveDriver only focused on cancer somatic mutations affecting PTMs, but did not extend to other serious diseases. Finally, previous studies mainly developed a database to curate PTM-related mutations obtained by their computational methods for user search, there is still no web-based tool available to annotate rare mutations in new disease research by PTM function. Therefore, existing computational tools are insufficient to assist PTM-mediated disease driver identification, an efficient and easy-to-use mutation analysis tool to discover disease driver mutations that affect a variety of PTM types are in great need to investigate the pathogenesis and development of multiple serious diseases.

In this paper, we introduce PTMsnp, a web server that implements a Bayesian hierarchical model to detect driver proteins with significant PTM-related mutations. PTMsnp has integrated 4,11,574 modification sites from 33 different types of PTMs and 1,776,848 somatic mutations of 33 cancer types. From PTMsnp, one can easily identify significantly PTM-mutated proteins (also known as driver genes) across different cohorts from TCGA. In addition, users can upload their own mutation resources, e.g., cohorts from genome-wide association studies (GWASs), to obtain significantly PTM-mutated proteins as well as potential disease-related mutations that significantly affect PTM status. In order to further evaluate the functional importance of PTM-related mutations, we also integrated multiple computational predictive programs for variant interpretation and clinical classification. To illustrate the functionality of PTMsnp, we applied it to TCGA cancer cohorts and a GWAS dataset of type 2 diabetes cohorts. Several known disease-related genes were successfully identified by PTMsnp, demonstrating that it is practicable to discover putative disease-related genes and hypothesize how they biochemically function in disease development.

MATERIALS AND METHODS

PTMsnp Algorithm

To identify proteins with a significantly high number of PTM-related mutations, we first converted the coordinates of

genetic mutations from the genomic level to the protein level using ANNOVAR (Wang et al., 2010). For analysis, only non-synonymous SNVs that did not create a premature stop codon or remove the existing stop codon were retained. According to previously published literatures (Reimand and Bader, 2013; Reimand et al., 2013; Narayan et al., 2016; Chen et al., 2018), the protein sequence flanking the central PTM site within seven residues was taken as the PTM motif region. The same type of PTM motif regions in the same protein were then merged to create a modification region. Correspondingly, the remaining sequences were merged separately and denoted as background regions. The frequency of each non-synonymous SNV located in the modification region and the background region were separately calculated.

We assumed that, in the patient group, mutations located in the PTM motif regions would probably damage the modification process, thereby influencing protein functions via PTM-related pathways. If such mutations are highly correlated with a given disease lesion, they will probably undergo strong positive selection; therefore, unexpectedly high mutation rates will be observed in these regions. According to this assumption, we developed the following Bayesian hierarchical model to compare the mutation rate between modification regions and background regions.

First, for a given protein, let $Y_1, Y_2, \ldots, Y_k$ represent the count of mutations at each position in the modification region, and let $Y_{k+1}, Y_{k+2}, \ldots, Y_n$ be the same count in the background region. We then modeled the observed counts Y by a Poisson distribution as shown in Equations 1 and 2, where λ_1 and λ_2 are the mutation rates of the modification region and the background region, respectively.

$$Y_{\mathrm{i}} \sim Possion(\lambda_1) \qquad i = 1, 2..., k \tag{1}$$

$$Y_{\mathrm{i}} \sim Possion(\lambda_2) \qquad i = k+1, k+2, ..., n \tag{2}$$

Since the mutation rate may vary markedly in different positions, a prior distribution was applied to λ_1 and λ_2 to capture such fluctuation. As stated in the theory of probability, a gamma distribution is the conjugate prior to the Poisson distribution. Therefore, two gamma distributions with different shape parameters α and scale parameters β were used to describe the distribution of λ_1 and λ_2 in Eqs 3 and 4.

$$\lambda_1 \sim Gamma(\alpha_1, \beta_1) \tag{3}$$

$$\lambda_2 \sim Gamma(\alpha_2, \beta_2) \tag{4}$$

To test the difference between the mutation rates of the background and those of the modification regions, a variable of interest might be the relative mutation rate, which is defined as $R = \lambda_1/\lambda_2$. Given that, a statistical hypothesis was raised as shown below.

$$H_0 : R \leq 1 \tag{5}$$

$$H_1 : R > 1 \tag{6}$$

The *p*-value under the null hypothesis can therefore be calculated from the marginal distribution of *R* given the observed data *Y*. A Markov chain Monte Carlo (MCMC) method was applied to infer such distribution. To control false positives, the Benjamini-Hochberg procedure is applied to each *p*-value. If the corrected *p*-value for a given protein is lower than the significance level, i.e., 0.05, we identify it as a potential disease driver (**Supplementary Methods**).

Database for PTM Sites and Mutations

PTM sites of human proteins were retrieved from the dbPTM (2019 update), iPTMnet (November 2019) database and manually collected from published literatures in PubMed. To unify the heterogeneity of data collected from different sources and ensure site accuracy, we mapped the reported modification sites to UniProtKB protein entries and used sequence comparison to correct the original data information and retain protein isoforms. Each mapped PTM site is attributed with a corresponding literature (PubMed ID) and source.

Somatic mutations were downloaded from the data portals of TCGA (18 July 2019)[1]. To construct an intact set of somatic mutations, mutations generated by four different variant calling workflows were merged and duplicated sites were removed. The ANNOVAR program was applied to annotate the functional consequence of all mutation sites. Only non-synonymous SNVs that did not create a premature stop codon or remove the existing stop codon were retained in our database.

The Processing of WTCCC T2D Dataset

The Wellcome Trust Case Control Consortium (WTCCC) Type 2 Diabetes (T2D) datasets consisted of individual–level genotypes called by BRLMM and Chiamo (The Wellcome Trust Case Control Consortium, 2007) were collected in this study. All SNPs were mapped to GRCh38 (hg38) genomic coordinates according to their RSIDs to facilitate the annotation of SNPs and proteins. Unmapped RSIDs was discarded. For genotypes called by BRLMM, calls with score < 0.5 were retained. For the Chiamo data, the recommended probability threshold for inclusion is > 0.9. After excluding low-quality samples or calls, the valid calls derived from two calling methods are intersected to obtain the reliable genotypes of all samples in T2D. Finally, all genotypes are processed into VCF files and used as input for PTMsnp.

RESULTS

Data Statistics of PTM Sites and Mutations

To assist the functional studies of cancer mutations, PTMsnp provides a database of known PTM sites and somatic mutations. PTM sites of human proteins are mainly derived from dbPTM (2019 update), a database that manually curated PTM peptides from the published literatures and integrated experimentally verified PTM sites from 30 available PTM-related resources such as PhosphoSitePlus (Hornbeck et al., 2015), dbPAF

[1]https://portal.gdc.cancer.gov/

(Ullah et al., 2016), UniProtKB (Boutet et al., 2007), PLMD (Xu et al., 2017), and Phospho.ELM (Dinkel et al., 2011) etc. We also collected additional PTM modification sites in iPTMnet, as well as manually curated from published literatures in PubMed. After strict data correction and filtering, a total of 4,11,574 PTM sites, covering Phosphorylation, Ubiquitination, Acetylation, Methylation, Sumoylation, Malonylation, O(N/C/S)-linked Glycosylation, S-nitrosylation, Glutathionylation, Succinylation, Nitration, Palmitoylation, Myristoylation, Hydroxylation, Crotonylation, Sulfation, Farnesylation, Geranylgeranylation, Gamma-carboxyglutamic acid, Pyrrolidone carboxylic acid, Citrullination, Glutarylation, Amidation, Carbamidation, Oxidation, GPI-anchor, Lipoylation, Neddylation, Carboxylation, and Pyruvate, were curated in our web server. On the other hand, somatic mutations downloaded from the data portals of TCGA were processed to retain non-synonymous SNVs, and finally, 1,776,848 non-synonymous SNVs across 33 cancer types (UCEC, SKCM, COAD, LUAD, STAD, LUSC, BLCA, BRCA, HNSC, GBM, CESC, OV, READ, LIHC, LGG, ESCA, PAAD, PRAD, KIRC, SARC, KIRP, ACC, LAML, UCS, THCA, DLBC, CHOL, THYM, MESO, TGCT, KICH, PCPG, and UVM) were collected in PTMsnp (**Supplementary Table S1**).

Web Server Description

To start PTMsnp, genetic mutations in standard VCF or TAB format need to be inputted in the text area or uploaded via the file selection box (**Figure 1A**). An intact set of somatic mutations from the cancer cohort of TCGA is integrated into the database, and users can also select a cancer type of interest to start analysis. Before calculation, several options, including PTM type, genome assembly version, iteration and burn-in times for the MCMC process, and *q*-value threshold should be set for the PTMsnp program (**Figure 1B**). Besides, users can enter email address to receive email notifications after the calculation is completed. After the submission of an analysis task, a new record will be added to the task monitoring bar at the bottom of the submit page (**Figure 1C**). When a task status is displayed as "complete," the user can click the "view" button to open the result page.

The result page consists of five interactive tables and graphs. The significantly PTM-mutated proteins that may drive the progression of diseases are outputted as a summary table (**Figure 1D**), supporting interactive operations such as filtering and sorting by cancer type, UniProt accession number, protein name and modification type. Each protein is directly linked to the UniProt database according to its accession number for details. The PTM-related mutations located in these proteins can be expanded or collapsed by click each protein record. Original information of PTM-related mutations such as base changes and genotypes are retained, as well as allele frequency obtained from ExAC database. Meanwhile, we scored the pathogenic level of each PTM-related mutation from 0 to 7 by counting the deleterious results of seven functional predictors [SIFT (Kumar et al., 2009), LRT (Chun and Fay, 2009), MutationTaster (Schwarz et al., 2010), MutationAssessor (Reva et al., 2011), FATHMM (Shihab et al., 2013), MetaSVM, and MetaLR (Dong et al., 2015)] curated in the dbNSFP database (Liu et al., 2016). Besides, InterVar (Li and Wang, 2017), and Clinvar (Landrum et al., 2018) are also integrated for clinical interpretation of PTM-related mutations by the ACMG/AMP 2015 guideline (Richards et al., 2015) and known disease association, respectively. For visualization, the distribution of significant PTM-related mutations and mutated PTM types in identified proteins are plotted in a bar graph and a pie chart (**Figure 1E**). In addition, for each identified protein, the mutation sites and known PTM sites together with their functional domains are presented in a schematic biological sequence diagram, where users can freely add or remove PTM tracks (**Figure 1F**). Moreover, to gain further insights into the protein function, we performed Gene Ontology (GO) and pathway enrichment analysis using the clusterProfiler package in R (Yu et al., 2012). The analysis results were illustrated in bar graphs (**Figure 1G**) and bubble plots (**Figure 1H**). All visualization diagrams are available in publication quality for download.

PTMsnp Identifies Known Cancer Genes With Significantly PTM-Related Mutations

To demonstrate how PTMsnp can be used for cancer driver genes detection, we first applied PTMsnp to analyze the somatic mutations from TCGA cohorts across 33 different cancer types. We selected five PTM types, including phosphorylation, acetylation, ubiquitination, methylation and sumoylation, with the largest number of modification sites to analyze the significant PTM-related mutations in cancer patients. PTMsnp identified 9,359 genes with significantly unexpected numbers of PTM-related mutations (P = 0.01, **Figure 2A** and **Supplementary Tables S2, S3**). Known cancer genes collected from the Cancer Gene Census (CGC) (Sondka et al., 2018), Network of Cancer Genes (NCG 6.0) (Repana et al., 2019), ONGene (Liu et al., 2017) as well as TSGene 2.0 (Zhao et al., 2016) database are significantly enriched (n = 2,064, P = 1.455 × 10-8, Fisher's exact test, **Supplementary Table S4**) in our result. Approximately, one–fourth of the identified genes (n = 2,256) contained significant PTM-related mutations in multiple cancer types. Of which, 660 genes were well-known cancer genes, such as CTNNB1, IDH1 (**Figure 3**). These results showed that the significantly PTM-mutated genes identified by PTMsnp may have a broad and important functional impact in the cancer driving mechanism.

Moreover, we found that PTMsnp identified the largest number of significantly PTM-mutated genes in Skin Cutaneous Melanoma (SKCM, **Figure 2A**). The BRAF gene ranked first by the number of PTM-related mutations in SKCM and harbored multiple significant PTM mutations in several cancer types (**Figure 3**). BRAF, also known as serine/threonine-protein kinase B-Raf, can phosphorylate MAP2K1 and thereby activates the MAP kinase signal transduction pathway in living cells. Mutations that activate BRAF functions are present in over 60% of all melanomas (Davies et al., 2002). Studies have shown that BRAF mutations are clustered within the P-loop and activation segment of the kinase domain (Pratilas et al., 2012; **Figure 2B**). These mutations destabilize the interaction between P-loop and the activation segment, which normally locks the kinase in

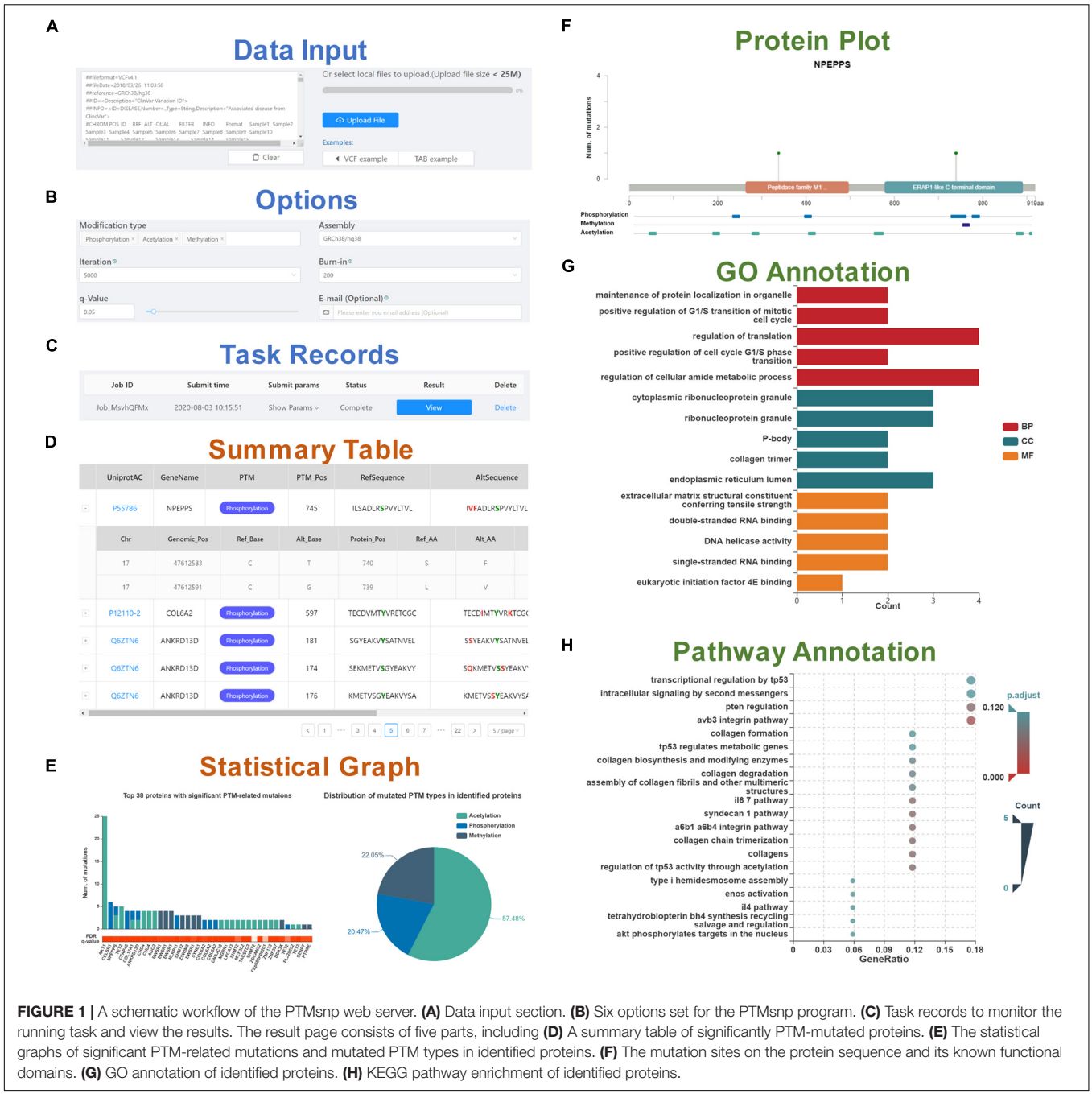


FIGURE 1 | A schematic workflow of the PTMsnp web server. **(A)** Data input section. **(B)** Six options set for the PTMsnp program. **(C)** Task records to monitor the running task and view the results. The result page consists of five parts, including **(D)** A summary table of significantly PTM-mutated proteins. **(E)** The statistical graphs of significant PTM-related mutations and mutated PTM types in identified proteins. **(F)** The mutation sites on the protein sequence and its known functional domains. **(G)** GO annotation of identified proteins. **(H)** KEGG pathway enrichment of identified proteins.

its inactive state until the activation loop is phosphorylated. Consistently, our method has identified a hotspot mutation at V600 of BRAF can significantly altered the modification level of three phosphorylation sites, namely Thr599, Ser602, and Ser605. One of these phosphorylation sites, Thr599, is located in the activation loop and believed to be functional in regulating the activation of BRAF (Lavoie and Therrien, 2015; Kiel et al., 2016). Three other mutations, including D594N, L597Q, and K601E, are also observed to potentially affect the phosphorylation at Thr599 (**Figure 2B**). Existing studies have confirmed that these mutations activate the MAPK pathway in melanoma and are associated with sensitivity to MEK inhibitor drug therapy (Dahlman et al., 2012; Wu et al., 2017). In view of these evidences, we hypothesized that the proto-oncogene BRAF is activated by mutations promoting the phosphorylation of its activation loop, implying the feasibility of applying PTMsnp to analyze cancer mutations from the perspective of affecting PTM modification.

Furthermore, we performed pathway analysis on the identified driver genes using MSigDB C2 Canonical pathways (Liberzon et al., 2015) to explore the biological system driven by PTM-related mutations in SKCM (**Figure 2C**). The top 20 enriched pathways were known to regulate cell proliferation,

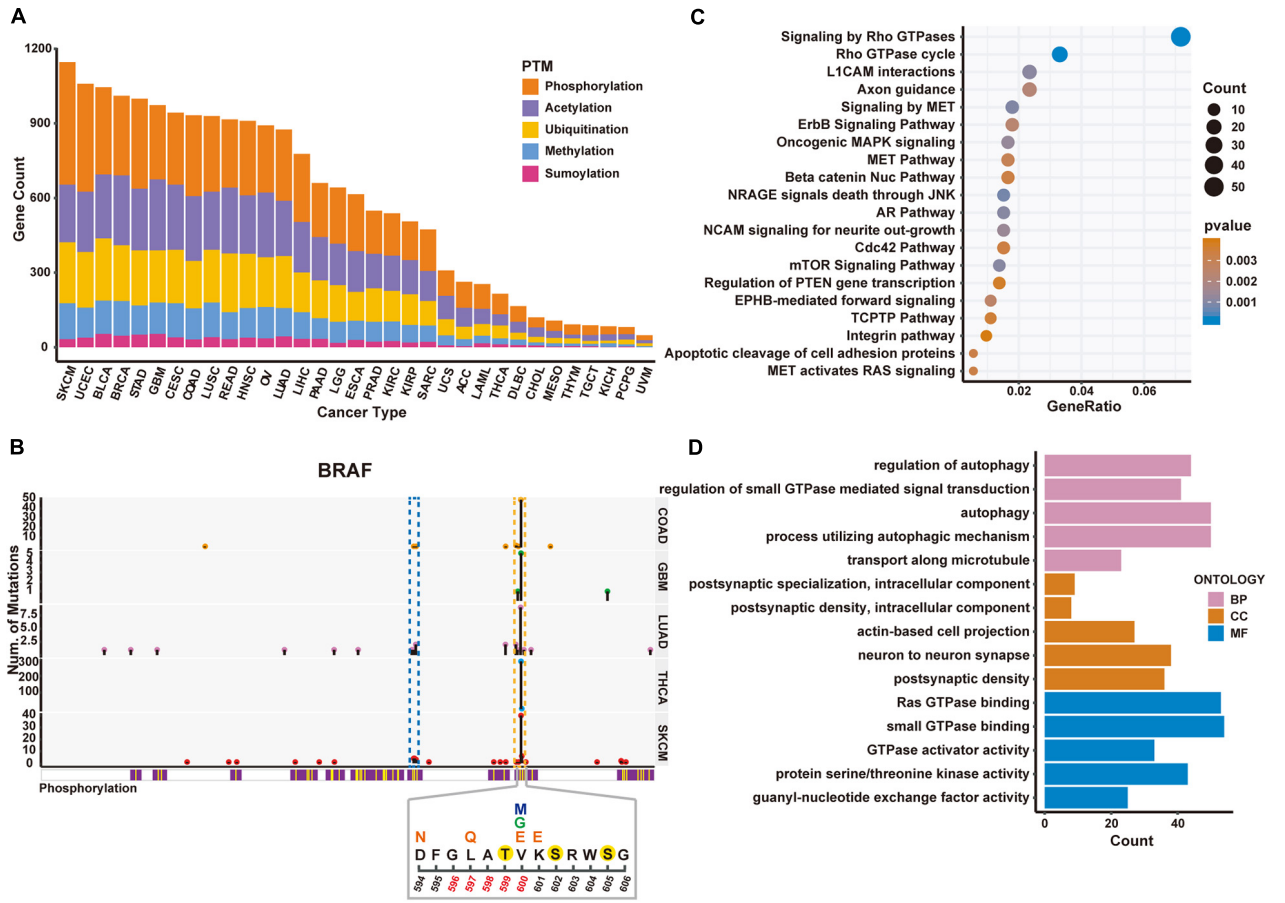


FIGURE 2 | Significantly mutated proteins identified in TCGA cancer cohorts regarding 5 PTM types. **(A)** Number of significantly PTM-mutated genes across five PTM types identified in different cancers. **(B)** Schematic diagram of mutations and protein phosphorylation modification regions within BRAF gene in five cancer types. Upper panel shows the number of mutated samples per position. The blue and yellow dashed boxes represent the P-loop and activation loop on the BRAF protein, respectively. The lower panel shows the mutation and phosphorylation within 594–606 region of the BRAF protein in SKCM. Positions 596–600 are the activation segment. Above the position coordinates is the amino acid sequence. The phosphorylated amino acids are marked with a yellow solid circle. Altered amino acid after mutation is above the original sequence. V600 has three different mutation forms, marked with different colors. **(C)** The enriched pathways of PTM-mutated proteins in SKCM. **(D)** The enriched GO terms obtained from the identified PTM-mutated proteins in SKCM.

migration, differentiation, apoptosis, and cell motility, therefore highlighted altered PTM level may be an important hallmark of cancers (Hanahan and Weinberg, 2011). Similar results were also observed in GO enrichment analysis (**Figure 2D**). These driver genes are enriched in cellular processes such as autophagy whose dysregulation has been linked to many human pathophysiologies including cancer (Chen and Klionsky, 2011; Jiang and Mizushima, 2014). All the above results demonstrated the functional importance of PTM functions in cancer development. Taken together, we suggested that PTMsnp can provide new perspectives on cancer studies, and subsequent experimental validation may help to discover novel mechanisms in cancerogenesis.

PTMsnp Identifies Potential Disease Drivers in GWAS Dataset

In addition, to show the practicability of applying PTMsnp in other disease-related studies, we further performed an analysis on a GWAS dataset of type 2 diabetes (T2D) samples from 1,916 tested individuals. Using PTMsnp, a total of 257 genes (**Supplementary Table S5**) with significant mutations across 12 different PTM types were identified (FDR P =0.05, **Figure 4A**). More than 70% PTM-related mutations are located in phosphorylation regions (**Figure 4B**), which is reasonable when considering the broadness of phosphorylation sites. SLC16A1 has the most frequent PTM-related mutations affecting three types of modifications including phosphorylation, methylation, and ubiquitination (**Figure 4C**). The solute carrier family 16 member 1 (SLC16A1) gene, which encodes the monocarboxylate transporter 1 (MCT1) protein, is a proton-coupled monocarboxylate transporter catalyzing the transportation of many monocarboxylates, such as lactate and pyruvate, across cell membranes. Many studies have revealed that mutations on SLC16A1 are associated with abnormal insulin secretion (Pullen et al., 2012; Al-Khawaga et al., 2019). Moreover, Nikooie et al. (2013) have reported that

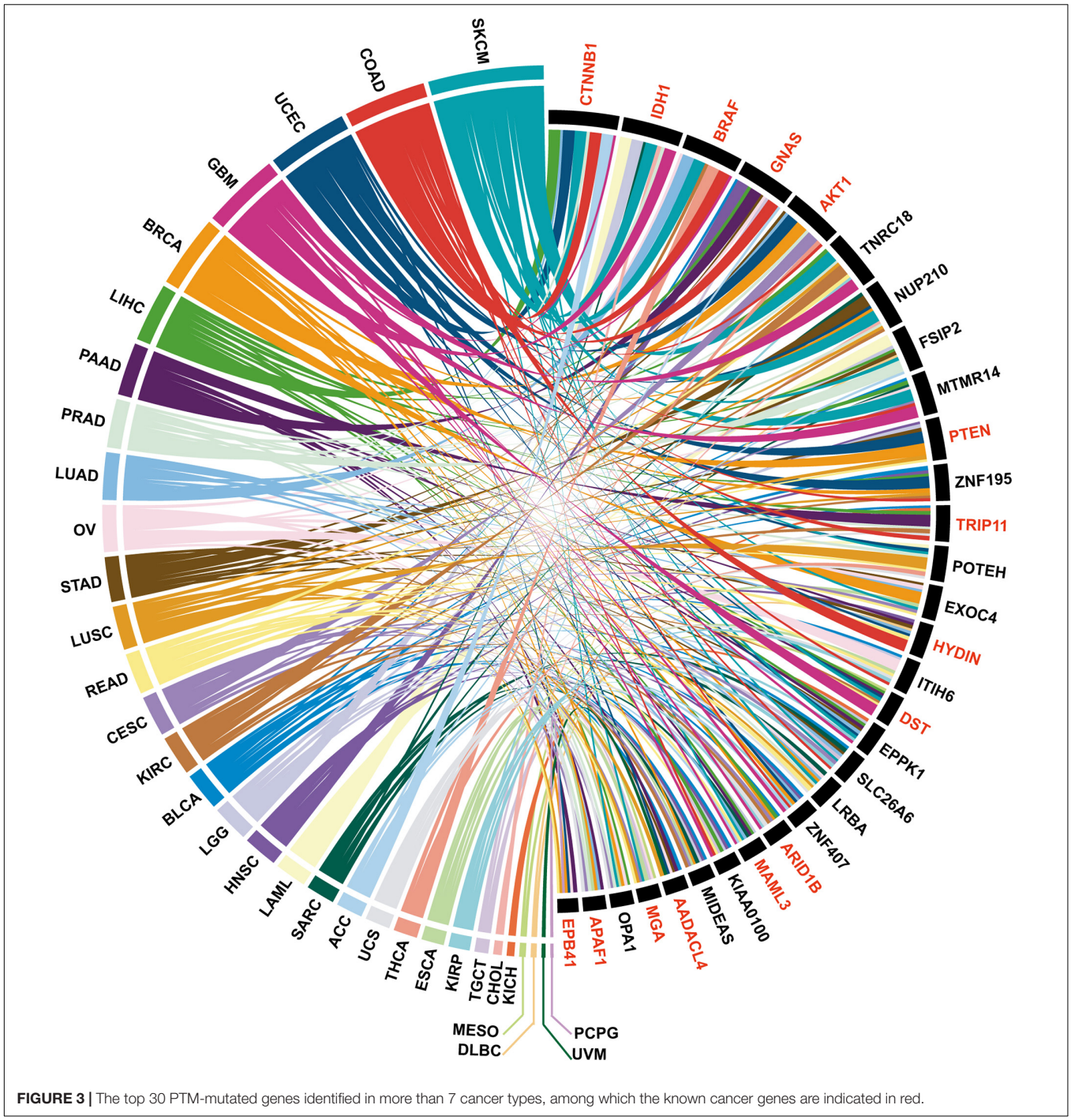


FIGURE 3 | The top 30 PTM-mutated genes identified in more than 7 cancer types, among which the known cancer genes are indicated in red.

the expression of MCT1 is dramatically reduced in diabetes, which may lead to increased insulin resistance. Besides, Zhao et al. (2001) have also found that the overexpression of MCT protein throughout the islet could involve in deranged insulin secretion in some type 2 diabetes. These studies suggested that the abnormal expression of MCT1 may be one of the pathogenic mechanisms of T2D. On the other hand, it has been reported that cAMP can cause the dephosphorylation of MCT1 and thereby reduce its surface expression (Smith et al., 2012). This evidence implies a positive synergy mechanism between MCT1 phosphorylation and its expression. Based on the existing literatures and our results, we speculated that our identified mutations on SLC16A1 can potentially affect its phosphorylation state, and may further lead to abnormal glucose sensing and even insulin resistance in T2D by changing the expression level of MCT1. Therefore, we can reasonably believe that SLC16A1 can serve as a novel PTM-mediated T2D driver genes.

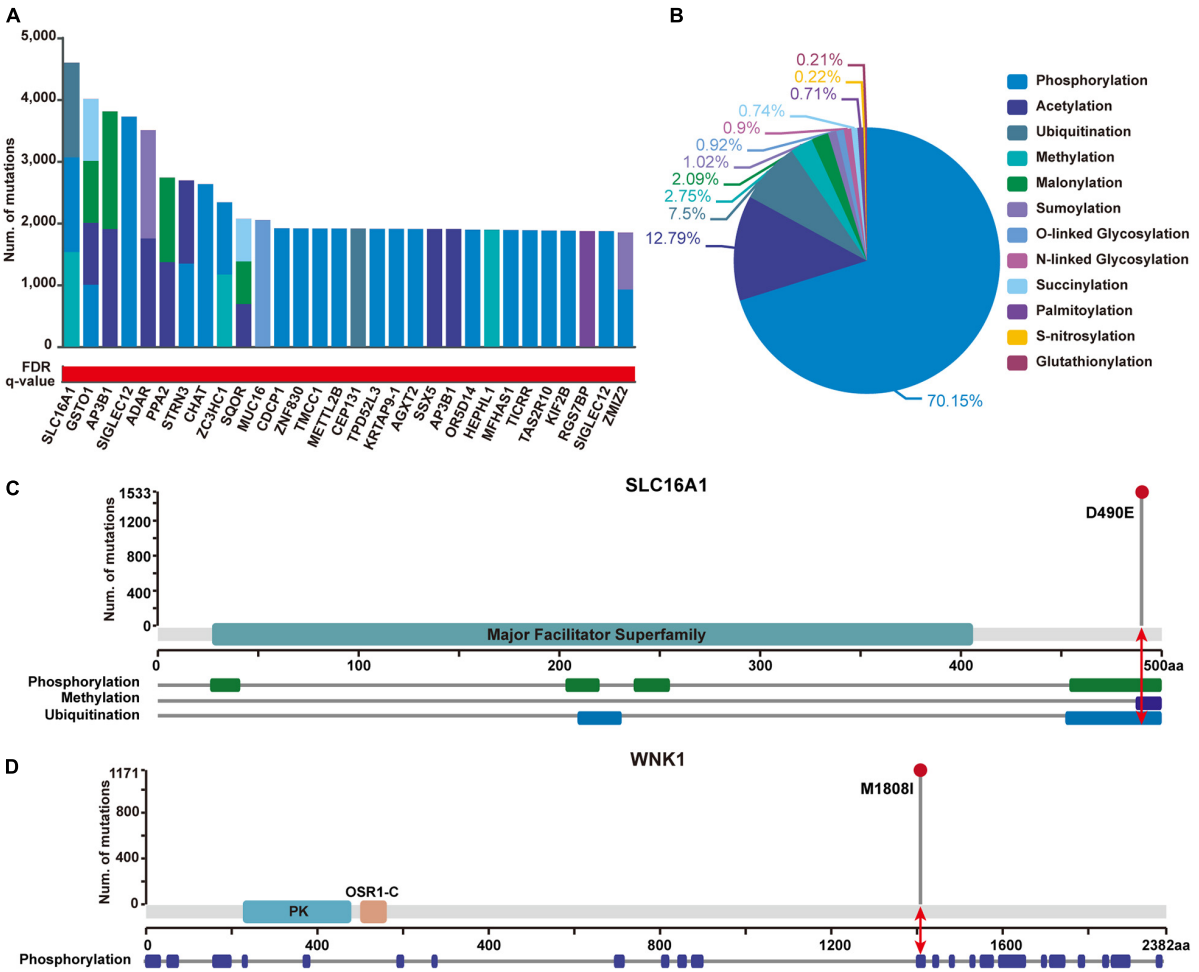


FIGURE 4 | Significantly PTM-mutated proteins identified from a GWAS dataset of Type 2 Diabetes (T2D) samples with 1,916 individuals. **(A)** The top 30 genes ranked by the number of significant PTM-related mutations. Bar height shows the number of samples harboring mutations in each PTM type, respectively. The red and white gradient bar below represents the FDR *q*-value. **(B)** The proportion of PTM-related mutations of each modification type in identified proteins. **(C)** SLC16A1 has the most frequent PTM-related mutations affecting three types of modifications. Upper panel shows the number of mutated samples per position. Protein domain of SLC16A1 are shown in green region along the sequence. The modified regions of three PTM types on SLC16A1 protein are shown below. The modified position where the mutation has occurred is indicated by a red arrow. **(D)** Mutation M1808I were identified to significantly alter phosphorylation status of WNK1. Protein domain of WNK1 are shown in blue and orange (PK, Protein kinase domain; OSR1-C, Oxidative-stress-responsive kinase 1 C-terminal domain). The modified position where the mutation has occurred is indicated by a red arrow.

Furthermore, 23 well-known T2D-related genes were found to carry significant PTM-related mutations in our analysis (**Supplementary Table S6**). Of these genes, With-no-lysine 1 (WNK1) kinase is taken here as an illustrative example (**Figure 4D**). WNK1 is serine-threonine kinase and highly expressed in skeletal muscles. An existing study has shown that insulin can phosphorylate WNK1, thereby activating glucose transporter 4 (GLUT4) translocation and stimulating glucose uptake through the PI3K/Akt signaling cascade. Decreased WNK1 phosphorylation were observed in T2D skeletal muscle, providing a new perspective on WNK1 function in T2D (Kim et al., 2018). Interestingly, we observed that the M1808I mutation on WNK1 was significantly enriched around the phrosphorylation site Thr1810 in T2D patients, implying a pathogenic role of WNK1 in T2D via its aberrant dephosphorylation. Given this observation, it is worthy to perform further experiments to verify the functional role of such mutation regarding to phrosphorylation process.

SUMMARY AND PERSPECTIVES

Genetic mutations in human genomes include both driver mutations that provide selective advantages to disease progression and neutral passenger mutations present due to genome instability. A key challenge facing the biological community is to distinguish only a few driver mutations from the majority of passenger mutations. Previous studies have proven that combining mutations with other important functional

features may provide extra guidance for driver event detection compared to traditional frequency-based methods. PTMs have been successfully used to predict driver mutations in diseases owing to their extensive functions in biological processes. However, the lack of an integrated resource of PTM sites as well as a user-friendly web interface greatly hindered the exploration of PTM-mediated disease progression. The PTMsnp web server was elaborately designed and dedicated for addressing such issues. With the collected PTM dataset, the vast majority of genetic mutations can be further annotated, and potential disease-driven genes can be inferred from the perspective of aberrant PTM status. As applications, we have successfully applied PTMsnp to the detection of cancer driver genes and disease-related genes from type 2 diabetes cohorts. This analysis revealed the prospect of using PTMsnp to explore the underlying pathogenesis of known disease-related mutations and to discover novel cancer drivers for further clinical research.

PTMsnp can be further enhanced in several aspects in the future. First, more genetic mutations such as population mutation datasets can be supported in future updates of PTMsnp. Different PTM processes can be orchestrated by different enzymatic systems, forming a dynamic regulatory cycle in normal cells. The perturbation of such a dynamic regulatory cycle may also lead to certain abnormalities. Therefore, the current algorithm can be further extended to consider mutations in PTM enzymes. In addition, the protein-protein interaction network may also be considered to interpret the impact of genetic mutations on PTM enzyme-substrate interactions, for example, kinase-substrate interactions in phosphorylation. With the ongoing database update and algorithm extensions, we expect PTMsnp to become a useful web server for the biomedical research community and to provide more valuable insights into disease biology and therapy development.

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 authors.

AUTHOR CONTRIBUTIONS

DP performed data analysis, implemented the PTMsnp algorithm, and wrote the manuscript. HL and BH were, respectively, responsible for the front-end page display and back-end logic design of the PTMsnp website. HZ designed original website page. LC and SL manually collected PTM sites from published literatures. ZZ and YX guided the methodology of the research. JR was responsible for supervision, funding acquisition, and writing – review. YBX supervised this work, designed the PTMsnp algorithm, reviewed, and edited the manuscript. All authors have read and approved the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (91753137, 31771462, 81772614, U1611261, 31801105, and 81802438), the National Key R&D Program of China (2017YFA0106700), the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2017ZT07S096), and the Guangdong Basic and Applied Basic Research Foundation (2018A030313323 and 2020A1515010220).

ACKNOWLEDGMENTS

This study is partly based upon data generated by the Wellcome Trust Case Control Consortium (WTCCC). We would like to acknowledge Andrew Hattersley and Mark McCarthy, as well as their support staff and their funding support who contributed to GWASs of type 2 diabetes.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2020.593661/full#supplementary-material

REFERENCES

Al-Khawaga, S., AlRayahi, J., Khan, F., Saraswathi, S., Hasnah, R., Haris, B., et al. (2019). A SLC16A1 mutation in an infant with Ketoacidosis and neuroimaging assessment: expanding the clinical spectrum of MCT1 deficiency. *Front. Pediatr.* 7:299. doi: 10.3389/fped.2019.00299

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, A. (2007). UniProtKB/Swiss-prot. *Methods Mol Biol.* 406, 89–112. doi: 10.1007/978-1-59745-535-0_4

Chen, L., Miao, Y., Liu, M., Zeng, Y., Gao, Z., Peng, D., et al. (2018). Pan-cancer analysis reveals the functional importance of protein lysine modification in cancer development. *Front. Genet.* 9:254. doi: 10.3389/fgene.2018.00254

Chen, Y., and Klionsky, D. J. (2011). The regulation of autophagy - unanswered questions. *J. Cell Sci.* 124(Pt 2), 161–170. doi: 10.1242/jcs.064576

Chun, S., and Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. *Genome Res.* 19, 1553–1561. doi: 10.1101/gr.092619.109

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* 22, 398–406. doi: 10.1101/gr.125567.111

Dahlman, K. B., Xia, J., Hutchinson, K., Ng, C., Hucks, D., Jia, P., et al. (2012). BRAF(L597) mutations in melanoma are associated with sensitivity to MEK inhibitors. *Cancer Discov.* 2, 791–797. doi: 10.1158/2159-8290.Cd-12-0097

Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., et al. (2002). Mutations of the BRAF gene in human cancer. *Nature* 417, 949–954. doi: 10.1038/nature00766

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., et al. (2012). MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* 22, 1589–1598. doi: 10.1101/gr.134635.111

Dinkel, H., Chica, C., Via, A., Gould, C. M., Jensen, L. J., Gibson, T. J., et al. (2011). Phospho.ELM: a database of phosphorylation sites–update 2011. *Nucleic Acids Res.* 39, D261–D267. doi: 10.1093/nar/gkq1104

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., et al. (2015). Comparison and integration of deleteriousness prediction methods for

nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* 24, 2125–2137. doi: 10.1093/hmg/ddu733

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

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. doi: 10.1093/nar/gku1267

Huang, H., Arighi, C. N., Ross, K. E., Ren, J., Li, G., Chen, S. C., et al. (2018). iPTMnet: an integrated resource for protein post-translational modification network discovery. *Nucleic Acids Res.* 46, D542–D550. doi: 10.1093/nar/gkx1104

Huang, L. C., Ross, K. E., Baffi, T. R., Drabkin, H., Kochut, K. J., Ruan, Z., et al. (2018). Integrative annotation and knowledge discovery of kinase post-translational modifications and cancer-associated mutations through federated protein ontologies and resources. *Sci. Rep.* 8:6518. doi: 10.1038/s41598-018-24457-1

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

Jiang, P., and Mizushima, N. (2014). Autophagy and human diseases. *Cell Res.* 24, 69–79. doi: 10.1038/cr.2013.161

Kiel, C., Benisty, H., Llorens-Rico, V., and Serrano, L. (2016). The yin-yang of kinase activation and unfolding explains the peculiarity of Val600 in the activation segment of BRAF. *eLife* 5:e12814. doi: 10.7554/eLife.12814

Kim, J.-H., Kim, H., Hwang, K.-H., Chang, J. S., Park, K.-S., Cha, S.-K., et al. (2018). WNK1 kinase is essential for insulin-stimulated GLUT4 trafficking in skeletal muscle. *FEBS Open Biol.* 8, 1866–1874. doi: 10.1002/2211-5463.12528

Kim, Y., Kang, C., Min, B., and Yi, G. S. (2015). Detection and analysis of disease-associated single nucleotide polymorphism influencing post-translational modification. *BMC Med. Genom.* 8(Suppl. 2):S7. doi: 10.1186/1755-8794-8-s2-s7

Krassowski, M., Paczkowska, M., Cullion, K., Huang, T., Dzneladze, I., Ouellette, B. F. F., et al. (2018). ActiveDriverDB: human disease mutations and genome variation in post-translational modification sites of proteins. *Nucleic Acids Res.* 46, D901–D910. doi: 10.1093/nar/gkx973

Kumar, P., Henikoff, S., Ng, P. C., Chun, S., and Fay, J. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* 4, 1073–1081. doi: 10.1038/nprot.2009.86

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46, D1062–D1067. doi: 10.1093/nar/gkx1153

Lavoie, H., and Therrien, M. (2015). Regulation of RAF protein kinases in ERK signalling. *Nat. Rev. Mol. Cell Biol.* 16, 281–298. doi: 10.1038/nrm3979

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218. doi: 10.1038/nature12213

Li, F., Fan, C., Marquez-Lago, T. T., Leier, A., Revote, J., Jia, C., et al. (2020). PRISMOID: a comprehensive 3D structure database for post-translational modifications and mutations with functional impact. *Brief Bioinform.* 21, 1069–1079. doi: 10.1093/bib/bbz050

Li, Q., and Wang, K. (2017). InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *Am. J. Hum. Genet.* 100, 267–280. doi: 10.1016/j.ajhg.2017.01.004

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.* 1, 417–425. doi: 10.1016/j.cels.2015.12.004

Lin, D. Y., Fang, H. I., Ma, A. H., Huang, Y. S., Pu, Y. S., Jenster, G., et al. (2004). Negative modulation of androgen receptor transcriptional activity by Daxx. *Mol. Cell Biol.* 24, 10529–10541. doi: 10.1128/mcb.24.24.10529-10541.2004

Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-Site SNVs. *Hum. Mutat.* 37, 235–241. doi: 10.1002/humu.22932

Liu, Y., Sun, J., and Zhao, M. (2017). ONGene: a literature-based database for human oncogenes. *J. Genet. Genomics* 44, 119–121. doi: 10.1016/j.jgg.2016.12.004

Martin, D. D. O., Kay, C., Collins, J. A., Nguyen, Y. T., Slama, R. A., and Hayden, M. R. (2018). A human huntingtin SNP alters post-translational modification and pathogenic proteolysis of the protein causing Huntington disease. *Sci. Rep.* 8:8096. doi: 10.1038/s41598-018-25903-w

Narayan, S., Bader, G. D., and Reimand, J. (2016). Frequent mutations in acetylation and ubiquitination sites suggest novel driver mechanisms of cancer. *Genome Med.* 8:55. doi: 10.1186/s13073-016-0311-2

Nikooie, R., Rajabi, H., Gharakhanlu, R., Atabi, F., Omidfar, K., Aveseh, M., et al. (2013). Exercise-induced changes of MCT1 in cardiac and skeletal muscles of diabetic rats induced by high-fat diet and STZ. *J. Physiol. Biochem.* 69, 865–877. doi: 10.1007/s13105-013-0263-6

Pratilas, C. A., Xing, F., and Solit, D. B. (2012). Targeting oncogenic BRAF in human cancer. *Curr. Top. Microbiol. Immunol.* 355, 83–98. doi: 10.1007/82_2011_162

Pullen, T. J., Sylow, L., Sun, G., Halestrap, A. P., Richter, E. A., and Rutter, G. A. (2012). Overexpression of monocarboxylate transporter-1 (SLC16A1) in mouse pancreatic β-cells leads to relative hyperinsulinism during exercise. *Diabetes* 61, 1719–1725. doi: 10.2337/db11-1531

Reimand, J., and Bader, G. D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* 9:637. doi: 10.1038/msb.2012.68

Reimand, J., Wagih, O., and Bader, G. D. (2013). The mutational landscape of phosphorylation signaling in cancer. *Sci. Rep.* 3:2651. doi: 10.1038/srep02651

Repana, D., Nulsen, J., Dressler, L., Bortolomeazzi, M., Venkata, S. K., Tourna, A., et al. (2019). The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol.* 20:1. doi: 10.1186/s13059-018-1612-0

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 39:e118. doi: 10.1093/nar/gkr407

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American college of medical genetics and genomics and the association for molecular pathology. *Genet. Med.* 17, 405–424. doi: 10.1038/gim.2015.30

Schwarz, J. M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* 7, 575–576. doi: 10.1038/nmeth0810-575

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., et al. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* 34, 57–65. doi: 10.1002/humu.22225

Simpson, C. M., Zhang, B., Hornbeck, P. V., and Gnad, F. (2019). Systematic analysis of the intersection of disease mutations with protein modifications. *BMC Med. Genom.* 12(Suppl. 6):109. doi: 10.1186/s12920-019-0543-2

Smith, J. P., Uhernik, A. L., Li, L., Liu, Z., and Drewes, L. R. (2012). Regulation of Mct1 by cAMP-dependent internalization in rat brain endothelial cells. *Brain Res.* 1480, 1–11. doi: 10.1016/j.brainres.2012.08.026

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., and Forbes, S. A. (2018). The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* 18, 696–705. doi: 10.1038/s41568-018-0060-1

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678. doi: 10.1038/nature05911

Ullah, S., Lin, S., Xu, Y., Deng, W., Ma, L., Zhang, Y., et al. (2016). dbPAF: an integrative database of protein phosphorylation in animals and fungi. *Sci. Rep.* 6:23534. doi: 10.1038/srep23534

Vandin, F., Upfal, E., and Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* 18, 507–522. doi: 10.1089/cmb.2010.0265

Wagih, O., Reimand, J., and Bader, G. D. (2015). MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. *Nat. Methods* 12, 531–533. doi: 10.1038/nmeth.3396

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38:e164. doi: 10.1093/nar/gkq603

Wendl, M. C., Wallis, J. W., Lin, L., Kandoth, C., Mardis, E. R., Wilson, R. K., et al. (2011). PathScan: a tool for discerning mutational significance in groups of putative cancer genes. *Bioinformatics* 27, 1595–1602. doi: 10.1093/bioinformatics/btr193

Wu, X., Yan, J., Dai, J., Ma, M., Tang, H., Yu, J., et al. (2017). Mutations in BRAF codons 594 and 596 predict good prognosis in melanoma. *Oncol. Lett.* 14, 3601–3605. doi: 10.3892/ol.2017.6608

Xu, H., Zhou, J., Lin, S., Deng, W., Zhang, Y., and Xue, Y. (2017). PLMD: An updated data resource of protein lysine modifications. *J. Genet. Genom.* 44, 243–250. doi: 10.1016/j.jgg.2017.03.007

Yang, Y., Peng, X., Ying, P., Tian, J., Li, J., Ke, J., et al. (2019). AWESOME: a database of SNPs that affect protein post-translational modifications. *Nucleic Acids Res.* 47, D874–D880. doi: 10.1093/nar/gky821

Youn, A., and Simon, R. (2011). Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* 27, 175–181. doi: 10.1093/bioinformatics/btq630

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

Zhao, C., Wilson, M. C., Schuit, F., Halestrap, A. P., and Rutter, G. A. (2001). Expression and distribution of lactate/monocarboxylate transporter isoforms in pancreatic islets and the exocrine pancreas. *Diabetes* 50, 361–366. doi: 10.2337/diabetes.50.2.361

Zhao, M., Kim, P., Mitra, R., Zhao, J., and Zhao, Z. (2016). TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Res.* 44, D1023–D1031. doi: 10.1093/nar/gkv1268

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.

frontiers in Cell and Developmental Biology

METHODS
published: 01 December 2020
doi: 10.3389/fcell.2020.594587

DeepCSO: A Deep-Learning Network Approach to Predicting Cysteine S-Sulphenylation Sites

Xiaru Lyu[1†], Shuhao Li[2,1†], Chunyang Jiang[1], Ningning He[1], Zhen Chen[3,4], Yang Zou[1] and Lei Li[1,5*]*

[1] School of Basic Medicine, Qingdao University, Qingdao, China, [2] College of Life Sciences, Qingdao University, Qingdao, China, [3] Collaborative Innovation Center of Henan Grain Crops, Henan Agricultural University, Zhengzhou, China, [4] Key Laboratory of Rice Biology in Henan Province, Henan Agricultural University, Zhengzhou, China, [5] School of Data Science and Software Engineering, Qingdao University, Qingdao, China

Cysteine S-sulphenylation (CSO), as a novel post-translational modification (PTM), has emerged as a potential mechanism to regulate protein functions and affect signal networks. Because of its functional significance, several prediction approaches have been developed. Nevertheless, they are based on a limited dataset from *Homo sapiens* and there is a lack of prediction tools for the CSO sites of other species. Recently, this modification has been investigated at the proteomics scale for a few species and the number of identified CSO sites has significantly increased. Thus, it is essential to explore the characteristics of this modification across different species and construct prediction models with better performances based on the enlarged dataset. In this study, we constructed several classifiers and found that the long short-term memory model with the word-embedding encoding approach, dubbed $LSTM_{WE}$, performs favorably to the traditional machine-learning models and other deep-learning models across different species, in terms of cross-validation and independent test. The area under the receiver operating characteristic (ROC) curve for $LSTM_{WE}$ ranged from 0.82 to 0.85 for different organisms, which was superior to the reported CSO predictors. Moreover, we developed the general model based on the integrated data from different species and it showed great universality and effectiveness. We provided the on-line prediction service called DeepCSO that included both species-specific and general models, which is accessible through http://www.bioinfogo.org/DeepCSO.

Keywords: machine learning, modification site prediction, deep learning, Cysteine S-sulphenylation, post-translational modification

OPEN ACCESS

Edited by:
Jiangning Song,
Monash University, Australia

Reviewed by:
Shaoping Shi,
Nanchang University, China
Zexian Liu,
Sun Yat-sen University Cancer Center (SYSUCC), China

***Correspondence:**
Yang Zou
yangzou306@gmail.com
Lei Li
leili@qdu.edu.cn

[†]These authors share first authorship

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

Received: *13 August 2020*
Accepted: *12 November 2020*
Published: *01 December 2020*

Citation:
Lyu X, Li S, Jiang C, He N, Chen Z, Zou Y and Li L (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

INTRODUCTION

Protein Cysteine S-sulphenylation (CSO) is the reversible oxidation of protein cysteinyl thiols to suphenic acids. S-sulphenylation functions as an intermediate on the path toward other redox modifications, such as disulfide formation, S-glutathionylation, and overoxidation to sulfinic and sulfonic acids (Paulsen and Carroll, 2013; Huang J.J et al., 2018). This modification has been reported to influence protein functions, regulate signal transduction and affect cell cycle (Van Breusegem and Dat, 2006; Men and Wang, 2007; Paulsen and Carroll, 2013; Hourihan et al., 2016;

Choudhury et al., 2017; Mhamdi and Van Breusegem, 2018). So far, thousands of CSO sites have been identified from different species including the mammal *Homo sapiens* and the plant organism *Arabidopsis thaliana* using the chemoproteomics approach (Yang et al., 2014; Li et al., 2016; Gupta et al., 2017; Akter et al., 2018; Huang et al., 2019; summarized in **Supplementary Table 1**). Nevertheless, the CSO site detection remains a major methodological issue due to low abundance and dynamic level of CSO-containing proteins *in vivo*. In contrast to the time-consuming and expensive experimental approaches, computational methods for predicting CSO sites have attracted considerable attention because of their convenience and efficiency.

Several computational methods have been developed for the prediction of CSO sites, mainly based on a single human dataset containing 1105 identified CSO sites (Yang et al., 2014). They include MDD-SOH (Bui et al., 2016a), iSulf-Cys (Xu et al., 2016), SOHSite (Bui et al., 2016b), PRESS (Sakka et al., 2016), Sulf_FSVM (Ju and Wang, 2018), S-SulfPred (Jia and Zuo, 2017), Fu-SulfPred (Wang et al., 2019), SulCysSite (Hasan et al., 2017), SOHPRED (Wang et al., 2016), and PredCSO (Deng et al., 2018). Out of them, two are based on protein three-dimensional structures, in which PRESS relies on four different protein structural properties (Sakka et al., 2016) whereas PredCSO is an ensemble model that combines bootstrap resampling, gradient tree boosting and majority voting with the 21 features refined out using a two-step feature selection procedure (Deng et al., 2018). The advantage of both classifiers is the inclusion of accurate structural features but their drawback is the limitation of the available structures. The rest classifiers are based on protein sequences. They can be classified into two clusters in terms of model complexity. The first cluster contains four relatively simple models. ISulf-Cys is an SVM (Support Vector Machine)-based classifier with the integration of three features including binary, PSAAP, and AAindex (Xu et al., 2016). SOHSite is an SVM-based classifier with the combined features of position-specific scoring matrix (PSSM) and AAindex (Bui et al., 2016b). SulCysSite is an RF (Random Forest)-based classifier with the integration of multiple features (Hasan et al., 2017) and Sulf_FSVM is an fuzzy SVM classifier using mRMR feature selection from three kinds of features (Ju and Wang, 2018). The second cluster includes four relatively complex models. MDD-SOH contains two-layered SVMs trained with MDDLogo-identified substrate motifs (Bui et al., 2016a). S-SulfPred is an SVM-based classifier with the balanced training dataset established using one-sided selection undersampling for negative samples and synthetic minority oversampling for positive samples (Jia and Zuo, 2017). Fu-SulfPred contains two layers of forest-based structure with the reconstruction of training datasets for data balance (Wang et al., 2019). SOHPRED was built by integrating four complementary predictors (i.e., a naive Bayesian predictor, an RF predictor, and two SVM predictors), each of which was associated with different training features (Wang et al., 2016). In summary, the characteristics of these sequence-based models are the combination of distinct types of features, or/and the balancing of training data, or/and the integration of different classifiers. Although the developed classifiers have made contribution to the prediction of CSO sites, most of them are currently inaccessible. Moreover, there is a lack of prediction tools for the CSO sites of multiple species. With the growing number of CSO sites verified, it is essential to develop species-specific prediction models with high accuracy or even a general model.

Compared to traditional machine-learning (ML) algorithms (e.g., SVM and RF) used in the prediction approaches described above, the deep-learning (DL) architecture is a promising ML algorithm. In the DL algorithm, a suitable representation of the input data can be transformed into highly abstract features through propagating the whole model. Superposition of hidden layers in neural networks can increase the ability of feature extraction, resulting in a more accurate interpretation of latent data patterns. Indeed, several frequently utilized DL models have been recently applied in the field of Bioinformatics, especially the prediction of post-translational modification (PTM) sites. For instance, deep neural networks were utilized for the prediction of protein nitration and nitrosylation sites (Xie et al., 2018), recurrent neural networks (RNNs) were employed for the prediction of lysine Malonylation sites (Chen et al., 2018b) and convolutional neural networks (CNNs) were used for the prediction of phosphorylation sites and crotonylation sites (Wang et al., 2017; Zhao et al., 2020). Deep learning algorithms have demonstrated their advantages in the application of large data sets, compared to the traditional ML methodology (Chen et al., 2018b). Because of this, the introduction of DL algorithms into the prediction of CSO sites would be a promising move to provide reliable candidates for further experimental consideration.

In this study, we constructed a number *in silico* approaches for the prediction of the CSO sites for *H. Sapiens* and *A. thaliana*. These approaches included the RF and SVM algorithms, one-dimensional CNN (1D-CNN), two-dimensional CNN (2D-CNN) and long short-term memory (LSTM) that is an RNN type. The LSTM model with the word-embedding encoding approach, called $LSTM_{WE}$, compared favorably to the rest approaches with AUC as 0.82 and 0.85 in human and Arabidopsis in terms of cross-validation. Moreover, $LSTM_{WE}$ trained using the data from one organism achieved outstanding performance in predicting CSO sites of other organisms (e.g., AUC = 0.80 for the prediction of Arabidopsis CSO sites using the human model), suggesting that CSO is highly conserved. Therefore, we constructed a general CSO prediction model. These models will facilitate the discovery of new CSO sites and thus will contribute to the understanding of roles and functions of CSO in diverse cellular processes.

MATERIALS AND METHODS

Data Collection and Preprocessing

The experimentally identified CSO sites were derived from two different organisms including *H. Sapiens* and *A. thaliana* (Yang et al., 2014; Li et al., 2016; Gupta et al., 2017; Akter et al., 2018; Huang et al., 2019). The data of the species were pre-processed and the related procedure was exemplified using the *A. thaliana* data, as listed below (**Figure 1A**).

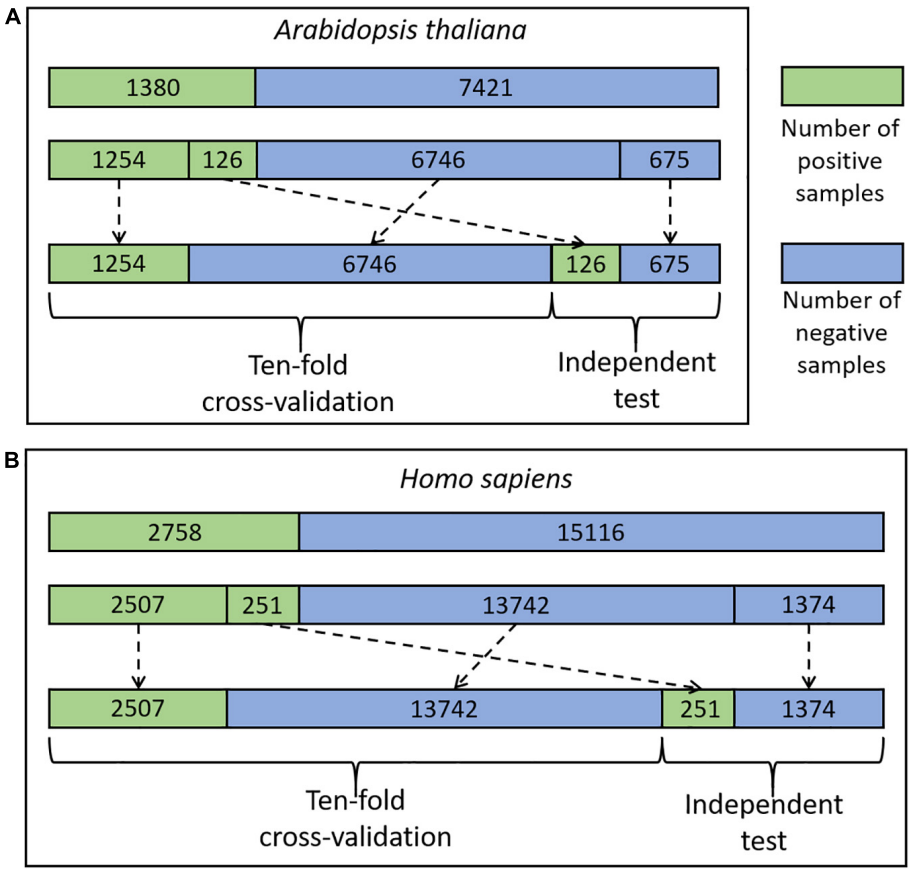


FIGURE 1 | The flowchart of the dataset process for *A. thaliana* **(A)** and *H. sapiens* **(B)**.

We mapped 1537 Arabidopsis CSO sites (Huang et al., 2019) to the UniprotKB database (UniProt Comstortium, 2011) and 1535 sites from 1130 proteins were retained as positive sites. The rest 8819 Cysteine residues in the same proteins were defined as negative sites. Moreover, we truncated these protein sequences into 35-residue segments with the Cysteine located at the center and the positive/negative sites correspond to positive/negative segments, respectively. It should be noted that if the central Cysteine was located around the N-terminus or C-terminus of a protein sequence, the gap symbol "-" was added to the corresponding positions to ensure that the segment had the same length. The segment length was optimized as a hyper-parameter in the Bayesian optimization method (see details in Section of "Optimization Methods for Hyper-Parameters") and finally determined as 33. Furthermore, to reduce the potential influence of the segments with high similarity on the performance of the models to be constructed, we set the identity of any two sequences with less than 40%, referring to previous studies (Bui et al., 2016a; Wang et al., 2016; Xu et al., 2016). When the identity was >40% between two positive segments or two negative segments, one was randomly removed. When the identity was >40% between a positive segment and a negative segment, the positive was retained and the negative was discarded. As a result, 1380 positives and 7421 negatives were retained. Finally, we randomly separated the positive and negative segments into 11 groups of which 10 were used for 10-fold cross-validation (1254 positives and 6746 negatives) and the rest for an independent test (126 positives and 675 negatives) (**Figure 1A**). Similarly, the cross-validation dataset for *H. sapiens* contained 16,249 samples (2507 positives and 13,742 negatives) and the independent test set comprised 1625 samples (251 positives and 1374 negatives) (**Figure 1B**). These datasets are available at http://www.bioinfogo.org/DeepCSO/download.php.

Feature Encoding Schemes

Numerical Representation for Amino Acids (NUM)

The NUM encoding approach maps each type of amino acid residue to an integer (Zhang Y. et al., 2019). Specifically, in the alphabet "AVLIFWMPGSTCYNQHKRDE-", each letter from "A" to "-" is converted to the integers from 0 to 20 in turn. For example, the sequence "VAMR" is encoded as "1,0,6,17." This encoding was used as the input of the first layer for both LSTM and 1D-CNN.

Enhanced Amino Acid Composition

The enhanced amino acid composition (EAAC) encoding (Chen et al., 2018b,c, 2020; Huang Y. et al., 2018) introduces a fixed-length sliding window based on the encoding of amino acid composition (AAC), which calculates the frequency of each type

of amino acid in a protein or peptide sequence (Bhasin and Raghava, 2004). EAAC is calculated by continuously sliding a fixed-length sequence window (using the default value 5) from the N-terminus to the C-terminus of each peptide. The related formula is listed below:

$$f(t, win) = \frac{N(t, win)}{N(win)}, t \in \{A, C, D, \ldots, Y\}, \quad win \in \{window1, window2, \ldots, window35\} \quad (1)$$

where *N(t, win)* is the number of amino acid *t* in the sliding window *win*, and *N(win)* is the size of the sliding window *win*.

Binary Encoding

In the binary encoding (Chen et al., 2018c), each amino acid is represented by a 21-dimensional binary vector that represents 20 amino acids and a complement "-." The corresponding position is set as 1 and the rest position is set as 0. For example, the amino acid "A" is represented by "100000000000000000000," "V" is represented by "010000000000000000000," and the symbol "-" is represented by "000000000000000000001," according to the alphabet "AVLIFWMPGSTCYNQHKRDE-."

AAindex Encoding

AAindex is a database of various indices representing distinct physicochemical and biochemical properties of amino acids and pairs of amino acids.[1] In the 544 physicochemical properties, we retained 531 properties after the removal of properties with "NA." We calculated the performance for each property using the RF classifier based on the 10-fold cross-validation dataset of arabidopsis. We selected the top 36 properties with AUC > 0.7 (**Supplementary Table 3**).

The Composition of k-Spaced Amino Acid Pairs

The composition of k-spaced amino acid pairs (CKSAAP) encoding contains the frequency of the amino acid pair of which both are separated by k-residues (k = 0, 1, 2, 3, 4, 5. We used the default value 5) (Chen et al., 2018c). This scheme represents the short- or long-range interactions amongst the residues along the sequence. The CKSAAP encoding with k = 0 is identical to the di-peptide composition.

The Position-Specific Scoring Matrix

The PSSM encoding was derived from the previous publication (Xie et al., 2018). In brief, we calculated the statistical significance of the differences in the frequencies of symbol occurrence between the positive and negative samples using a two-sample *t*-test (Vacic et al., 2006). Accordingly, the PSSM of significant *P*-values were constructed. By integrating the PSSM of *P*-values with the frequency PSSM for positive and negative samples, we generated the final encoding PSSM that represented the conservation tendency of the positive or negative samples.

Architecture of the Machine-Learning Models

The LSTM Model With the Word Embedding Encoding (LSTM$_{WE}$)

LSTM$_{WE}$ contained five layers, listed as follows (**Figure 2**).

1. Input layer. Each peptide segment is converted into an integer vector with the NUM encoding.

[1] http://www.genome.jp/aaindex/

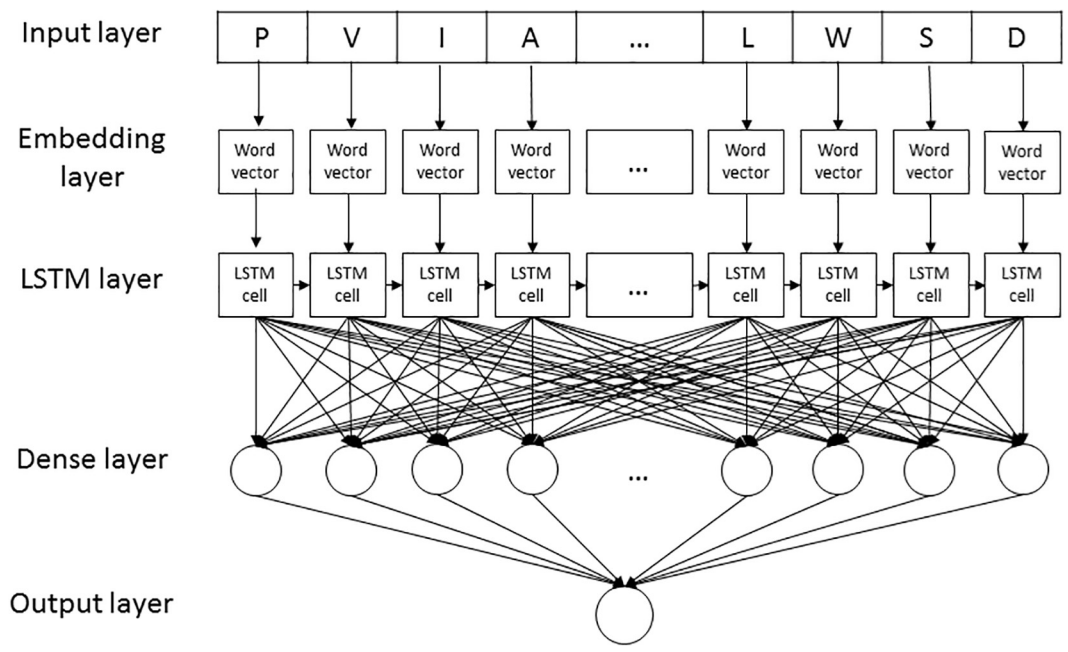


FIGURE 2 | The LSTM$_{WE}$ architecture.

2. Word Embedding (WE) layer. Each integer of the vector from the input layer is encoded into a four-dimension word vector for humans and a five-dimensional word vector for arabidopsis, respectively.
3. LSTM layer. Each of the word vectors is input sequentially into the LSTM cell that contained 32 hidden neuron units.
4. Dense layer. It contains a single dense sublayer that has 16 neurons with the ReLU activation function for humans and 32 neurons for arabidopsis, separately.
5. Output layer. This layer has only one neuron activated by sigmoid function, outputting the probability of the CSO modification.

The 1D-CNN Model With the Word Embedding Encoding

The 1D-CNN model with the word embedding encoding (1D-CNN$_{WE}$) contains five layers (**Supplementary Figure 1**), where the first two layers and last one layer were as same as LSTM$_{WE}$. The third layer was a 1D convolution layer with 22/20 filters for humans/arabidopsis and kernel size as nine. The fourth layer had a single dense sublayer with 16 neurons. The optimal hyper-parameter values were obtained using the Bayesian optimization algorithm.

The 2D-CNN Model With the PSSM Feature

We took advantage of the 2D structure of an input image of CNN architecture and conveniently made similar 2D inputs of PSSM matrixes with the sizes of 20 × 20 s. The purpose of using the 2D-CNN model is to catch the hidden figures inside PSSM profiles. Next, PSSM profiles were connected to the 2D CNN design from the input layer through several hidden layers to the output layer. **Supplementary Figure 2** demonstrated the procedure of inputting a PSSM profile into the CNN model, then passing through a series of convolutional, non-linearity, pooling and fully connected layers and finally outputting the result. This model contained four hidden layers including one 2D convolutional layer, one pooling layer, one flattening layer, and one fully connected layer. Specifically, the first layer contained a PSSM profile on which we applied 2D convolutional operations with some existing parameters including 5 × 5 kernel size, 15 filters and 1 × 1 stride.

The RF Algorithms With Different Features

The RF algorithm integrates multiple decision trees and chooses the classification with the most votes from the trees. Each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest. In this study, we constructed the RF models with six different features, including binary encoding, EAAC encoding, AAindex encoding, CKSAAP encoding, PSSM encoding, and WE. The number of decision trees was selected as 580 *via* the grid search method. These classifiers were developed based on the Python module "sklearn."

The SVM Algorithms With Different Features

We applied the Python-based machine learning package "scikit-learn" to implement the SVM algorithm and adopted the "RBF" kernel function to build the SVM models. The above encoding schemes for RF were applied to the SVM model. In particular, we normalized the feature values that do not range between 0 and 1 (such as PSSM) before inputting the SVM model.

Model Training Strategy

Optimization Methods for Hyper-Parameters

The hyper-parameters of an ML classifier affect prediction performance. Although a lot of combinations of hyper-parameters need to be tested, there are no formal rules to find optimal hyper-parameters. Here we applied two search approaches [grid search and Bayesian optimization (BO)] to automatical adjustment and evaluation of hyper-parameters (**Figure 3**). Grid search is a brute-force method to find the optimal hyper-parameters by training models using each possible combination of hyper-parameters and retaining the hyper-parameters corresponding to the model with the best performance. This method applies to a limited number of hyper-parameters due to the exponential increase in time spent with the number of hyper-parameters. In this study, it was used for the RF-based and SVM-based models. The related grid search spaces (**Supplementary Table 3**) were searched using the GridSearchCV function of the sklearn library in Python. On the contrary, BO provides a principled technique based on Bayes theorem to direct a search of a global optimization problem, which is effective to tune the hyper-parameters of DL models. The BO strategy

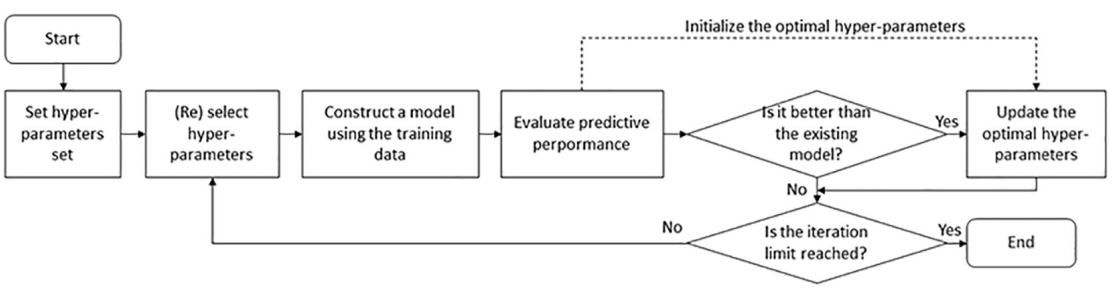


FIGURE 3 | Hyper-parameter optimization procedure for machine-learning classifiers.

was executed using the fmin function of the hyperopt library in Python. The BO related hyper-parameter space contained 10 parameters, including window size, kernel size, and dropout rate (**Supplementary Table 3**). The optimal hyper-parameter combination results for the DL models were listed in **Supplementary Table 4**.

Strategy of Avoiding Overfitting

1The parameters in the DL models were trained and optimized based on binary cross-entropy loss function using the Adam algorithm. The maximum of the training cycles was set through the optimized number of epochs to ensure that the loss function value converged. In each epoch, the training dataset was separated with the batch size as 512 and iterated. To avoid overfitting, the early-stopping strategy was applied, where the training process was stopped early when the training loss did not go down within 50 consecutive iterations. The model with the smallest training loss was saved as the best model. Moreover, the dropout rate of the neuron units was set, which was obtained through the hyper-parameter optimization. **Supplementary Figures 3**, **4** showed the training and validation accuracy and loss curves of the LSTM_{WE} models for different species.

Performance Assessment of the Predictors

Several measures were used to evaluate the prediction performance, including accuracy (ACC), specificity (SP), sensitivity (SN), Matthew's correlation coefficient (MCC). They are defined as follows:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\text{SP} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{SN} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FN} \times \text{FP})}{\sqrt{(\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TP} + \text{FP}) \times (\text{TN} + \text{FN})}}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. Additionally, because the number of positive and negative samples was unbalanced and the above measures were calculated based on the threshold value, a measure that was independent of the threshold value and unaffected by the sample ratio was needed. Therefore, the receiver operating characteristic (ROC) curve and AUC were employed to comprehensively evaluate classification performance. Specifically, due to the low false-positive rate of a predictor is significant in practical application, the area under the ROC curve with <10% false-positive rate (AUC01) was considered.

Statistical Methods

The paired student's *t*-test was used to test the significant difference between the mean values of the

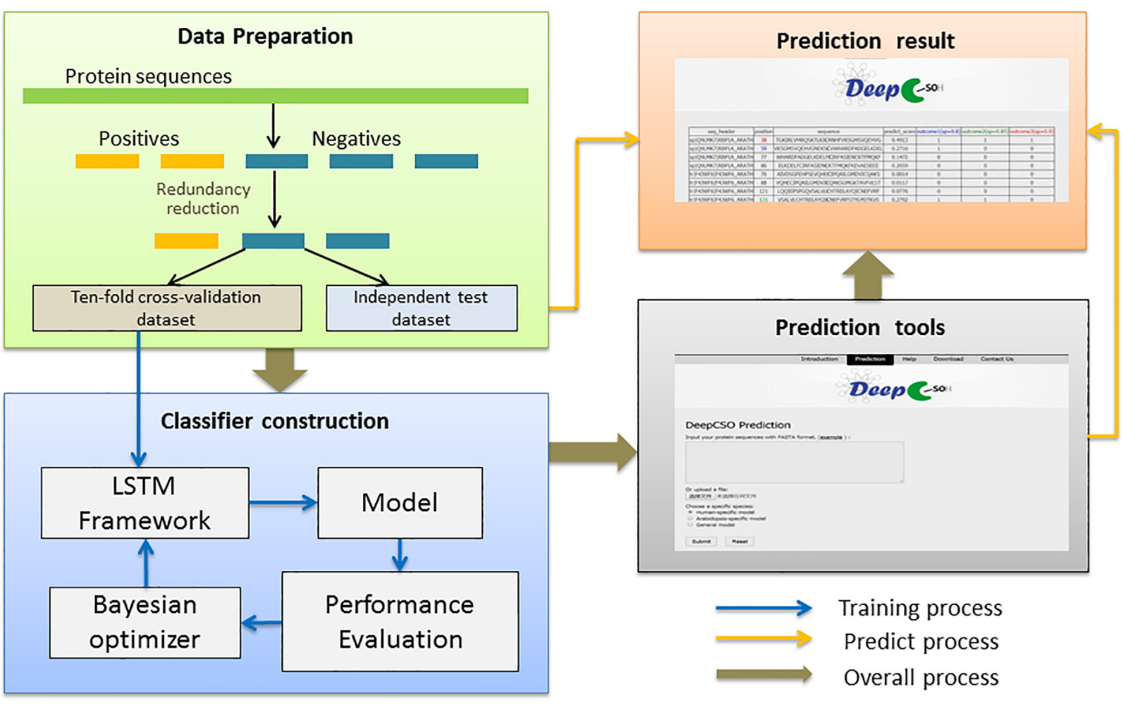


FIGURE 4 | The flowchart of the prediction model construction.

two paired populations. The adjusted *P*-value with the Benjamini-Hochberg (BH) method was adopted for multiple comparisons.

The Flowchart of the Prediction Model Construction

The flowchart of the prediction model construction contained three steps (**Figure 4**). This first step was data collection and preprocessing, in which the sample data were separated into the cross-validation dataset and the independent test dataset for model construction and evaluation. The second step was classifier construction, which involved data decoding, model training, and hyper-parameter adjustment for resulting in a robust predictive model. The third step was the development of the final model as an online prediction tool.

RESULTS AND DISCUSSION

$LSTM_{WE}$ Classifier Performed Favorably to Other Classifiers

Many computational approaches for predicting PTM sites are generally based on traditional ML algorithms (e.g., RF and SVM) combined with various features encoded from peptide sequences. In this study, we constructed both RF-based and SVM-based predictors with different encoding schemes for the CSO site prediction. The encoding schemes include six features [i.e., binary, AAindex, WE, KSAAP, PSSM, and EAAC]. Moreover, deep learning algorithms have recently been applied to the field of PTM site prediction and demonstrated their superior performances (Wang et al., 2017; Chen et al., 2018b). Accordingly, we developed three different DL classifiers, named 1D-CNN_{WE}, 2D-CNN_{PSSM}, and $LSTM_{WE}$.

TABLE 1 | Performances of various classifiers for different species in terms of 10-fold cross-validation.

Classifier[1]	ACC[2]	Sn[2]	Sp[2]	MCC2[2]	AUC[2]	AUC01[2]
Arabidopsis thaliana						
RF_{BINARY}	0.743 ± 0.006	0.449 ± 0.040	0.798 ± 0.001	0.210 ± 0.032	0.696 ± 0.021	0.014 ± 0.002
RF_{EAAC}	0.773 ± 0.007	0.628 ± 0.043	0.799 ± 0.001	0.351 ± 0.033	0.803 ± 0.019	0.024 ± 0.004
RF_{WE}	0.748 ± 0.007	0.474 ± 0.048	0.799 ± 0.001	0.230 ± 0.038	0.728 ± 0.020	0.014 ± 0.002
$RF_{AAINDEX}$	0.744 ± 0.008	0.443 ± 0.053	0.800 ± 0.001	0.206 ± 0.043	0.710 ± 0.025	0.014 ± 0.004
RF_{CKSAAP}	0.749 ± 0.012	0.477 ± 0.078	0.800 ± 0.001	0.234 ± 0.062	0.728 ± 0.032	0.013 ± 0.003
RF_{PSSM}	0.740 ± 0.006	0.419 ± 0.039	0.800 ± 0.000	0.188 ± 0.032	0.670 ± 0.028	0.015 ± 0.004
RF_{E+C+A}	0.760 ± 0.006	0.544 ± 0.040	0.800 ± 0.001	0.287 ± 0.031	0.770 ± 0.016	0.020 ± 0.005
SVM_{BINARY}	0.748 ± 0.009	0.479 ± 0.055	0.798 ± 0.003	0.234 ± 0.043	0.719 ± 0.025	0.017 ± 0.002
SVM_{EAAC}	0.746 ± 0.009	0.458 ± 0.060	0.799 ± 0.001	0.218 ± 0.048	0.704 ± 0.026	0.015 ± 0.004
$SVM_{AAINDEX}$	0.750 ± 0.008	0.486 ± 0.054	0.800 ± 0.000	0.241 ± 0.042	0.724 ± 0.023	0.016 ± 0.004
SVM_{CKSAAP}	0.739 ± 0.007	0.421 ± 0.047	0.798 ± 0.003	0.187 ± 0.037	0.692 ± 0.030	0.013 ± 0.003
SVM_{PSSM}	0.726 ± 0.008	0.330 ± 0.054	0.800 ± 0.001	0.113 ± 0.046	0.590 ± 0.025	0.009 ± 0.003
2D-CNN_{PSSM}	0.766 ± 0.010	0.585 ± 0.064	0.800 ± 0.000	0.319 ± 0.050	0.781 ± 0.030	0.023 ± 0.004
1D-CNN_{WE}	0.783 ± 0.006	0.696 ± 0.041	0.799 ± 0.001	0.401 ± 0.030	0.838 ± 0.019	0.029 ± 0.005
$LSTM_{WE}$	**0.786 ± 0.007**	**0.717 ± 0.044**	**0.799 ± 0.001**	**0.417 ± 0.032**	**0.852 ± 0.018**	**0.030 ± 0.006**
Homo sapiens						
RF_{BINARY}	0.749 ± 0.004	0.466 ± 0.027	0.800 ± 0.000	0.225 ± 0.021	0.720 ± 0.013	0.016 ± 0.002
RF_{EAAC}	0.766 ± 0.006	0.578 ± 0.039	0.800 ± 0.000	0.312 ± 0.030	0.790 ± 0.018	0.020 ± 0.002
RF_{WE}	0.751 ± 0.004	0.480 ± 0.024	0.800 ± 0.000	0.236 ± 0.019	0.732 ± 0.015	0.018 ± 0.001
$RF_{AAINDEX}$	0.750 ± 0.004	0.474 ± 0.025	0.800 ± 0.000	0.231 ± 0.020	0.734 ± 0.017	0.018 ± 0.003
RF_{CKSAAP}	0.753 ± 0.003	0.493 ± 0.018	0.800 ± 0.000	0.246 ± 0.014	0.729 ± 0.016	0.016 ± 0.002
RF_{PSSM}	0.748 ± 0.004	0.462 ± 0.026	0.800 ± 0.000	0.222 ± 0.021	0.707 ± 0.016	0.016 ± 0.001
RF_{E+S+A}	0.761 ± 0.005	0.551 ± 0.033	0.800 ± 0.000	0.291 ± 0.026	0.774 ± 0.012	0.021 ± 0.002
SVM_{BINARY}	0.750 ± 0.005	0.474 ± 0.030	0.800 ± 0.000	0.231 ± 0.024	0.720 ± 0.013	0.017 ± 0.002
SVM_{EAAC}	0.742 ± 0.007	0.421 ± 0.049	0.800 ± 0.000	0.188 ± 0.039	0.680 ± 0.021	0.013 ± 0.002
$SVM_{AAINDEX}$	0.753 ± 0.006	0.498 ± 0.041	0.800 ± 0.000	0.250 ± 0.032	0.737 ± 0.021	0.017 ± 0.001
SVM_{CKSAAP}	0.737 ± 0.005	0.388 ± 0.031	0.800 ± 0.000	0.162 ± 0.025	0.664 ± 0.012	0.012 ± 0.002
SVM_{PSSM}	0.725 ± 0.005	0.316 ± 0.033	0.800 ± 0.000	0.101 ± 0.028	0.578 ± 0.025	0.011 ± 0.002
2D-CNN_{PSSM}	0.766 ± 0.004	0.581 ± 0.029	0.800 ± 0.000	0.314 ± 0.022	0.777 ± 0.011	0.022 ± 0.003
1D-CNN_{WE}	0.778 ± 0.006	0.659 ± 0.036	0.800 ± 0.000	0.373 ± 0.027	0.819 ± 0.012	0.024 ± 0.003
$LSTM_{WE}$	**0.777 ± 0.006**	**0.651 ± 0.038**	**0.800 ± 0.000**	**0.367 ± 0.028**	**0.822 ± 0.011**	**0.024 ± 0.003**

[1]The RF classifiers with the different features were named as RF_{BINARY}, RF_{WE}, etc. The 1D CNN and LSTM classifiers with the word embedding approach were named as 1D-CNN_{WE} and $LSTM_{WE}$, respectively.[2]ACC, Sn, Sp, MCC, AUC, and AUC01 were described in section "Materials and Methods." In the 10-fold cross-validation, 10 models were constructed using the 10 different validation datasets. Finally, the average performance and standard deviation of the 10 models were calculated for the cross-validation dataset. The models with the best performances were highlighted in bold.

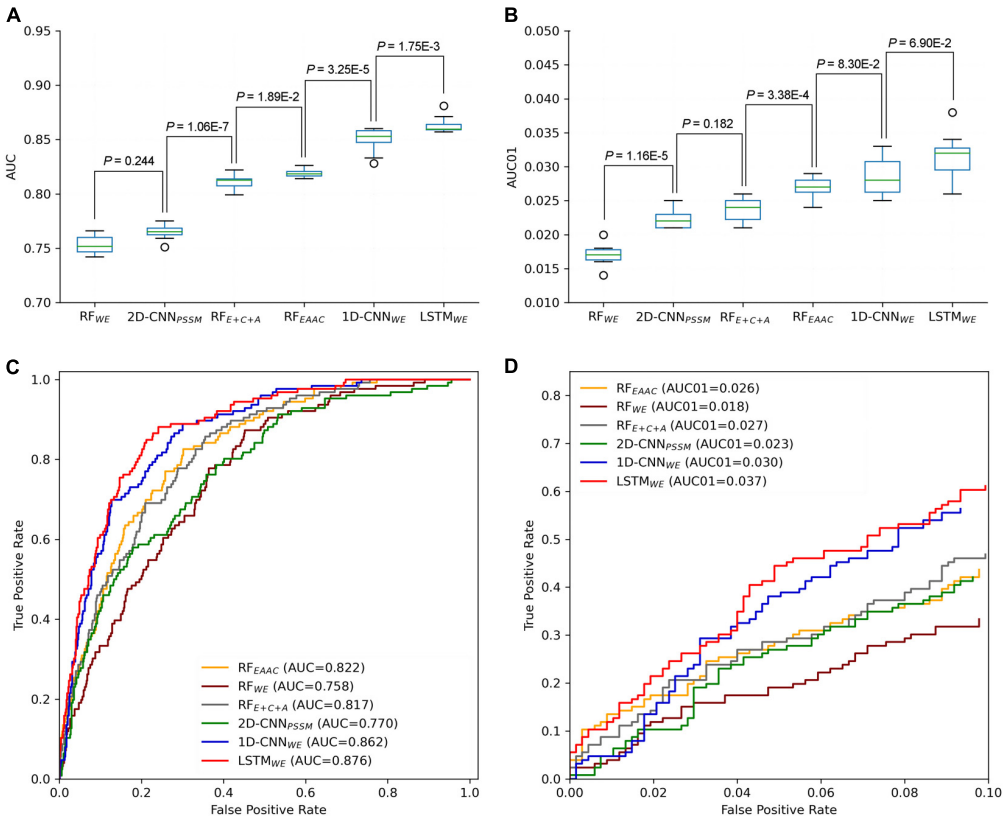


FIGURE 5 | Performance comparison of different CSO predictors on *Arabidopsis thaliana*. The performances of CSO predictors were compared in terms of AUC **(A)** and AUC01 **(B)**, respectively, for 10-fold cross-validation. AUC **(C)** and AUC01 **(D)** curves were generated using the independent test.

We first took the Arabidopsis data to construct and compare different models (Huang et al., 2019). The Arabidopsis cross-validation dataset contained 8000 samples (1254 positives and 6746 negatives) and the independent test set covered 801 samples (126 positives and 675 negatives) (**Figure 1**). We compared the performances of these algorithms in terms of several measures (e.g., ACC, MCC, AUC, and AUC01) for both the 10-fold cross-validation (**Table 1**) and the independent test (**Supplementary Table 5**). In the traditional ML models, RF_{EAAC} showed superior performance than other RF-based and SVM-based models. The previous studies of CSO site prediction showed that the models with the combination of different encoding methods compared favorably to their counterparts with a single encoding approach (Bui et al., 2016b; Xu et al., 2016). Accordingly, we constructed such models and the RF model with the combination of EAAC, CKSAAP, and AAindex, dubbed RF_{E+C+A}, had the best performance. To our surprise, RF_{E+C+A} had inferior performance compared to RF_{EAAC} (**Table 1** and **Supplementary Table 5**).

TABLE 2 | The k-fold cross-validation results of existed tools.

Tools*	Fold	Accuracy	Sensitivity	Specificity	AUC
MDD-SOH	5	0.68	0.7	0.7	
SOHSite	5	0.71	0.72	0.72	
SOHPRED	5		0.727 ± 0.005	0.742 ± 0.001	0.801 ± 0.001
iSulf-Cys	10	0.656 ± 0.007	0.673 ± 0.007	0.639 ± 0.001	0.716 ± 0.009
SulCysSite	10		0.745 ± 0.006	0.744 ± 0.002	0.806 ± 0.002
Sulf_FSVM	10	0.711 ± 0.002	0.733 ± 0.004	0.708 ± 0.002	0.788 ± 0.002
$LSTM_{WE}$	10	0.739 ± 0.006	0.694 ± 0.042	0.744 ± 0.008	0.800 ± 0.011
RF_{EAAC}	10	0.733 ± 0.006	0.607 ± 0.021	0.750 ± 0.007	0.753 ± 0.006
RF_{E+S+A}	10	0.743 ± 0.009	0.728 ± 0.027	0.745 ± 0.009	0.807 ± 0.010

**The cross-validation dataset was derived from Yang's publication (Yang et al., 2014).*

All the models constructed above were based on the imbalanced dataset. To evaluate the effect of the imbalanced dataset on potential overfitting of the classifiers, we reconstructed RF_{EAAC} based on the balanced positive and negative samples. Specifically, because the number of negative samples was around five times larger than that of the positive samples, we randomly separated the negative samples into five parts and created five subsets of training data with a 1:1 positive-to-negative ratio. Subsequently, five RF_{EAAC} models (sub-classifiers) were trained and the average output score from the five sub-classifiers was taken as the final prediction score. **Supplementary Figure 5** showed the performances of the two RF_{EAAC} models based on the balanced and imbalanced dataset, respectively, in terms of the 10-fold cross-validation and the independent test dataset. Because of

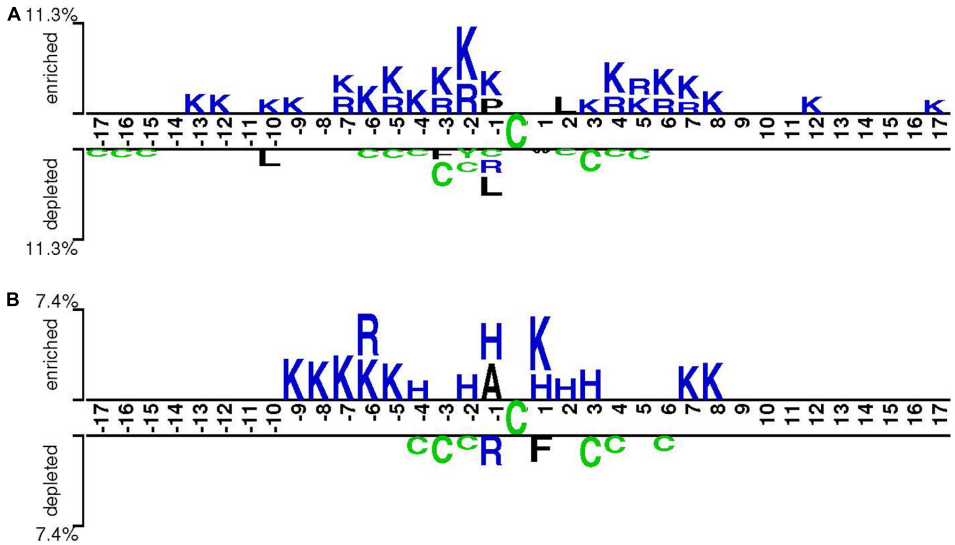


FIGURE 6 | Sequence pattern surrounding the CSO sites, including the significantly enriched and depleted residues based on CSO-containing peptides and non-modification peptides for *H. sapiens* **(A)** and *A. thaliana* **(B)** ($P < 0.05$, *t*-test with Bonferroni correction). The pattern was generated using the two-sample-logo method (Vacic et al., 2006).

the slightly better performance of the RF_{EAAC} model constructed using an imbalanced training dataset, we selected the imbalanced dataset for the construction of the models.

In our previous studies, DL models showed superior performance than traditional ML models (Chen et al., 2018b; Zhao et al., 2020). It is still true for the CSO site prediction. $LSTM_{WE}$ had the best performance among these constructed models in terms of ACC, Sn, MCC, and AUC values for both 10-fold cross-validation and independent test. For instance, its AUC value is 0.852 for the cross-validation and its values of ACC, Sn, Sp, and MCC were 0.786, 0.717, 0.799, and 0.417, respectively (**Table 1** and **Figures 5A,C**). As prediction performance at a low false-positive rate is highly useful in practice, we estimated these predictors using AUC01, where the specificity was determined to be >90%. $LSTM_{WE}$ again showed the largest AUC01 values for both 10-fold cross-validation and the independent test (**Figures 5B,D**). As the encoding approach has a great impact on the traditional ML models (Chen et al., 2018b; Huang Y. et al., 2018; Zhao et al., 2020) and the WE approach integrated with LSTM had the best performance in this study, we attempted to investigate whether the integration of WE and RF had a good performance. Accordingly, we extracted WE layer vector as feature encoding from $LSTM_{WE}$ and trained the RF model, dubbed RF_{WE}. Interestingly, RF_{WE} did not show good performance compared to RF_{EAAC}, $1D\text{-}CNN_{WE}$, or $LSTM_{WE}$. It suggests that the WE encoding approach may be improper for the construction of traditional ML algorithms.

We further constructed the models for the human organism. The Humans cross-validation dataset contained 16,249 samples (2507 positives and 13,742 negatives) and the independent test set covered 1625 samples (251 positives and 1374 negatives) (**Figure 1B**). Similarly, $LSTM_{WE}$ had the best performance (**Table 1**, **Supplementary Table 5**, and **Supplementary Figure 6**). For instance, its values of AUC, ACC, Sn, Sp, MCC, and AUC01 for the 10-fold cross-validation were 0.822, 0.777, 0.651, 0.800, 0.367, and 0.024, respectively. We evaluated the robustness of $LSTM_{WE}$ by comparing their performances between the cross-validation and independent tests for individual organisms. As their performances were not statistically different for each organism (P = 0.18/0.085 for the arabidopsis/humans, respectively), we concluded that the constructed models were robust and neither over-fitting nor under-fitting.

$LSTM_{WE}$ Performed Better Than Reported Classifiers

Six approaches for the prediction of human CSO sites were based on 1105 identified human CSO sites (Yang et al., 2014), including MDD-SOH, SOHSite, SOHPRED, iSulf-Cys, SulCysSite, and Sulf_FSVM. We compare these models and our models (i.e., RF_{EAAC}, RF_{E+C+A}, and $LSTM_{WE}$) to evaluate their prediction performances. Accordingly, we constructed our models using the same dataset derived from the original study (Yang et al., 2014). SulCysSite, $LSTM_{WE}$, and RF_{E+C+A} had the best and similar performances (**Table 2**). The observation that the model with the combined features (i.e., RF_{E+C+A}) had better accuracy than the counterpart with a single feature (i.e.,

TABLE 3 | Evaluation of species-specific and general $LSTM_{WE}$ models using the independent test sets from different species.

Independent test sets	$LSTM_{WE}$ model (AUC value)		
	Arabidopsis-specific	Human-specific	General
A. thaliana	0.876	0.799	0.863
H. sapiens	0.766	0.839	0.834

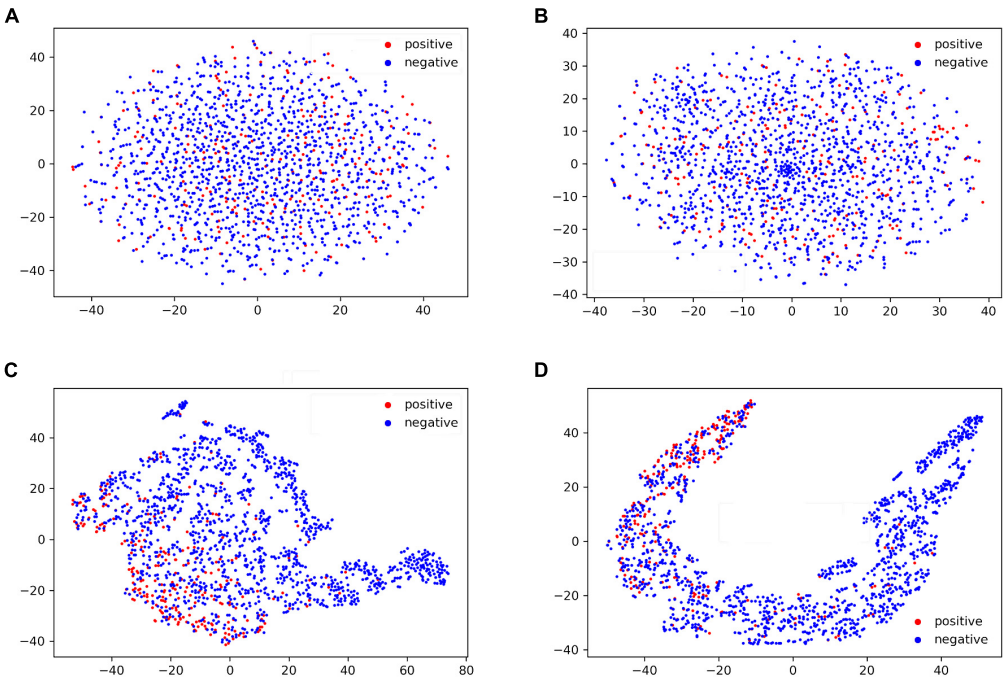


FIGURE 7 | T-SNE visualization of the distributions of peptides in the human independent dataset for the outputs of input layer **(A)**, word embedding layer **(B)**, LSTM layer **(C)**, and dense layer **(D)** of the general $LSTM_{WE}$ model.

RF_{EAAC}) is consistent with the previous studies (Bui et al., 2016b; Xu et al., 2016) but conflicted with our observation above that RF_{EAAC} compared favorably to RF_{E+C+A}. This contradiction derived from the different amounts of the training datasets, where the dataset here was smaller than the datasets described above, indicating that the amount of training data affected the performance of the models. Indeed, based on the small human dataset (1105 positives), RF_{E+C+A} had a better performance than RF_{EAAC}, whereas the performance of RF_{EAAC} was better than that of RF_{E+C+A} with a large amount of the training set (arabidopsis: 1380 positives; human 2758 positives) (**Supplementary Figure 7**). In all comparisons, $LSTM_{WE}$ showed the best performance (**Supplementary Figure**). Additionally, as iSulf-Cys (Xu et al., 2016) is the only accessible model to date, we compared it and $LSTM_{WE}$ using the human independent dataset of this study. The AUC value (0.839) of $LSTM_{WE}$ is significantly larger than that (0.666) of iSulf-Cys (**Supplementary Figure 8**). In summary, $LSTM_{WE}$ performed better than reported classifiers.

Conservation of the CSO Modification and the Development of General $LSTM_{WE}$ Models

Cysteine S-sulphenylation has been identified across various organisms, ranging from yeasts to worms and from plants to humans (Men and Wang, 2007; Hourihan et al., 2016). To understand its conservation, we compared the characteristics of CSO-containing peptides in human and arabidopsis species, respectively, using the two-sample-logo approach (Vacic et al., 2006). **Figure 6** showed that both species shared the enriched basic amino acids R and K and the depleted polar neutral amino acid C. Nevertheless, the amino acid H was enriched for *A. thaliana* whereas the hydrophobic amino acid L was depleted for *H. sapiens*. As the characteristics of CSO-containing peptides were similar between both species, we hypothesized the generalization ability of our developed models. To test this hypothesis, we used the human $LSTM_{WE}$ model to predict the arabidopsis independent test dataset and employed the Arabidopsis $LSTM_{WE}$ model to predict the human independent test dataset. The AUC values were 0.799 and 0.766, respectively, significantly larger than the random prediction (i.e., AUC = 0.5; **Table 3**). Nevertheless, the cross-species prediction had relatively low performance compared to the self-species prediction (AUC = 0.876/0.839 for arabidopsis/human, respectively). As the CSO sites were systematically analyzed in a few species, we developed a general CSO prediction model according to its conservation to boost the investigation for other species. Accordingly, we mixed the training datasets of *H. Sapiens* and *A. thaliana* and constructed the general $LSTM_{WE}$ model and validated it using the independent datasets from both organisms. The performance of the general $LSTM_{WE}$ model was slightly lower than that of the self-species prediction, which may be caused by the interference of the CSO characteristics of other species (**Table 3**). Overall, the conservation of the CSO modification leads to the effective prediction of the general $LSTM_{WE}$ classifier.

To further understand the performance of the general $LSTM_{WE}$ classifier, we visualized the sample distribution, based

on the human independent dataset, from the outputs of the input layer, WE layer, LSTM layer, and dense layer of the general model using the t-SNE algorithm (van der Maaten and Hinton, 2008; **Figure 7**). After the input layer (**Figure 7A**), the positive and negative samples were mixed, as the training goes on (**Figures 7B,C**), positive and negative samples were gradually separated. After the LSTM layer, they were separated (**Figure 7D**). This comparison indicates that the LSTM layer is a powerful method to detect the distinctive features of the positives and negatives. A similar observation is made for the arabidopsis independent test dataset (**Supplementary Figure 9**).

Construction of the Online CSO Predictor

We developed an easy-to-use online tool for the prediction of the CSO sites, dubbed DeepCSO. DeepCSO contains three $LSTM_{WE}$ models: the general model and two species-specific models (i.e., *H. sapiens* and *A. thaliana*). The users could select the general model or species-specific model at the input interface and input the query protein sequences directly or upload the sequence file. After the job submission, the prediction will start and the prediction process may take several minutes. Finally, the prediction results are output in tabular form with five columns: sequence header, position, sequence, prediction score, and prediction results at the specificity levels of 80, 85, and 90%, respectively.

Several Cysteine modification types have been reported in the human organism, such as carbonylation (Wang et al., 2014; Chen et al., 2017, 2018a; Zhang S. et al., 2019), oxidation (Gupta et al., 2017; Akter et al., 2018), succination (Adam et al., 2017), and sulfenylation. Some Cysteine sites can be modified with multiple modification types, which cause PTM cross-regulation. To examine potential PTM cross-regulation at the proteome scale, we downloaded the latest human protein sequences from the Swiss-Prot database (version: 2020_05) and applied the human DeepCSO predictor to predict the potential CSO sites with the annotation of the reported Cysteine modifications (**Supplementary Table 6**). This resource will assist in the investigation of the Cystine co-regulation in the community.

CONCLUSION

The current prediction tools for CSO sites are based on traditional ML methodology that requires experts to pre-define informative features, and no prediction tool has been developed for other than the human organism. In this study, three LSTM-based prediction models were constructed, where two were organism-specific and one was general, and they compared favorably to the reported models. Despite lacking pre-defined features, the deep learning classifier demonstrated superior performance compared to the traditional machine learning methods. This may be due to the self-learning ability of deep learning. The outstanding performance of the general model suggests that the CSO is well conserved and the LSTM-based model has an advantage in long-term memory to capture the key features of the entire sequences.

DATA AVAILABILITY STATEMENT

The 10-fold cross-validation and independent data sets can be found in http://www.bioinfogo.org/DeepCSO/.

AUTHOR CONTRIBUTIONS

LL conceived this project. XL and SL constructed the algorithms under the supervision of LL and YZ.; CJ, XL, and NH analyzed the data. XL, YZ, NH, and ZC, and LL wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported in part by funds from the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 31701142 to ZC), and the National Natural Science Foundation of China (Grant Nos. 31770821 and 32071430 to LL); LL was supported by the "Distinguished Expert of Overseas Tai Shan Scholar" program. YZ was supported by the Qingdao Applied Research Project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2020.594587/full#supplementary-material

REFERENCES

Adam, J., Ramracheya, R., Chibalina, M. V., Ternette, N., Hamilton, A., Tarasov, A. I., et al. (2017). Fumarate hydratase deletion in pancreatic beta cells leads to progressive diabetes. *Cell Rep.* 20, 3135–3148. doi: 10.1016/j.celrep.2017.08.093

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

Bhasin, M., and Raghava, G. P. (2004). Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem.* 279, 23262–23266. doi: 10.1074/jbc.m401932200

Bui, V. M., Lu, C. T., Ho, T. T., and Lee, T. Y. (2016a). MDD-SOH: exploiting maximal dependence decomposition to identify S-sulfenylation sites with substrate motifs. *Bioinformatics* 32, 165–172.

Bui, V. M., Weng, S. L., Lu, C. T., Chang, T. H., Weng, J. T., and Lee, T. Y. (2016b). SOHSite: incorporating evolutionary information and physicochemical properties to identify protein S-sulfenylation sites. *BMC Genomics* 17(Suppl. 1):9. doi: 10.1186/s12864-015-2299-1

Chen, Y., Cong, Y., Quan, B., Lan, T., Chu, X., Ye, Z., et al. (2017). Chemoproteomic profiling of targets of lipid-derived electrophiles by bioorthogonal aminooxy probe. *Redox Biol.* 12, 712–718. doi: 10.1016/j.redox.2017.04.001

Chen, Y., Liu, Y., Lan, T., Qin, W., Zhu, Y., Qin, K., et al. (2018a). Quantitative profiling of protein carbonylations in ferroptosis by an aniline-derived probe. *J. Am. Chem. Soc.* 140, 4712–4720. doi: 10.1021/jacs.8b01462

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

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

Chen, Z., Zhao, P., Li, F., Marquez-Lago, T. T., Leier, A., Revote, J., et al. (2020). iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief. Bioinform.* 21, 1047–1057. doi: 10.1093/bib/bbz041

Choudhury, F. K., Rivero, R. M., Blumwald, E., and Mittler, R. (2017). Reactive oxygen species, abiotic stress and stress combination. *Plant J.* 90, 856–867. doi: 10.1111/tpj.13299

Deng, L., Xu, X., and Liu, H. (2018). PredCSO: an ensemble method for the prediction of S-sulfenylation sites in proteins. *Mol. Omics* 14, 257–265. doi: 10.1039/c8mo00089a

Gupta, V., Yang, J., Liebler, D. C., and Carroll, K. S. (2017). Diverse redoxome reactivity profiles of carbon nucleophiles. *J. Am. Chem. Soc.* 139, 5588–5595.

Hasan, M. M., Guo, D. J., and Kurata, H. (2017). Computational identification of protein S-sulfenylation sites by incorporating the multiple sequence features information. *Mol. Biosyst.* 13, 2545–2550. doi: 10.1039/c7mb00491e

Hourihan, J. M., Moronetti Mazzeo, L. E., Fernandez-Cardenas, 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, J., Willems, P., Wei, B., Tian, C., Ferreira, R. B., Bodra, N., et al. (2019). Mining for protein S-sulfenylation in *Arabidopsis* uncovers redox-sensitive sites. *Proc. Natl. Acad. Sci. U.S.A.* 116, 21256–21261. doi: 10.1073/pnas.1906768116

Huang, J. J., Willems, P., Van Breusegem, F., and Messens, J. (2018). Pathways crossing mammalian and plant sulfenomic landscapes. *Free Radic. Biol. Med.* 122, 193–201. doi: 10.1016/j.freeradbiomed.2018.02.012

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, 1669–1677. doi: 10.7150/ijbs.27819

Jia, C., and Zuo, Y. (2017). S-SulfPred: a sensitive predictor to capture S-sulfenylation sites based on a resampling one-sided selection undersampling-synthetic minority oversampling technique. *J. Theor. Biol.* 422, 84–89. doi: 10.1016/j.jtbi.2017.03.031

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

Li, R., Klockenbusch, C., Lin, L., Jiang, H., Lin, S., and Kast, J. (2016). Quantitative protein sulfenic acid analysis identifies platelet releasate-induced activation of integrin beta2 on monocytes via NADPH oxidase. *J. Proteome Res.* 15, 4221–4233. doi: 10.1021/acs.jproteome.6b00212

Men, L., and Wang, Y. (2007). The oxidation of yeast alcohol dehydrogenase-1 by hydrogen peroxide in vitro. *J. Proteome Res.* 6, 216–225. doi: 10.1021/pr0603809

Mhamdi, A., and Van Breusegem, F. (2018). Reactive oxygen species in plant development. *Development* 145:dev164376. doi: 10.1242/dev.164376

Paulsen, C. E., and Carroll, K. S. (2013). Cysteine-mediated redox signaling: chemistry, biology, and tools for discovery. *Chem. Rev.* 113, 4633–4679. doi: 10.1021/cr300163e

Sakka, M., Tzortzis, G., Mantzaris, M. D., Bekas, N., Kellici, T. F., Likas, A., et al. (2016). PRESS: PRotEin S-Sulfenylation server. *Bioinformatics* 32, 2710–2712. doi: 10.1093/bioinformatics/btw301

UniProt Comstortium (2011). Ongoing and future developments at the universal protein resource. *Nucleic Acids Res.* 39, D214–D219.

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

Van Breusegem, F., and Dat, J. F. (2006). Reactive oxygen species in plant cell death. *Plant Physiol.* 141, 384–390.

van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Wang, C., Weerapana, E., Blewett, M. M., and Cravatt, B. F. (2014). A chemoproteomic platform to quantitatively map targets of lipid-derived electrophiles. *Nat. Methods* 11, 79–85. doi: 10.1038/nmeth.2759

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, 3909–3916. doi: 10.1093/bioinformatics/btx496

Wang, L., Zhang, R., and Mu, Y. (2019). Fu-SulfPred: identification of protein S-sulfenylation sites by fusing forests via chou's general PseAAC. *J. Theor. Biol.* 461, 51–58. doi: 10.1016/j.jtbi.2018.10.046

Wang, X., Yan, R., Li, J., and Song, J. (2016). SOHPRED: a new bioinformatics tool for the characterization and prediction of human S-sulfenylation sites. *Mol. Biosyst.* 12, 2849–2858. doi: 10.1039/c6mb00314a

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: prediction of S-sulfenylation sites in proteins with physicochemical properties of amino acids. *PLoS One* 11:e0154237. doi: 10.1371/journal.pone.0154237

Yang, J., Gupta, V., Carroll, K. S., and Liebler, D. C. (2014). Site-specific mapping and quantification of protein S-sulphenylation in cells. *Nat. Commun.* 5:4776.

Zhang, S., Fang, C., Yuan, W., Zhang, Y., Yan, G., Zhang, L., et al. (2019). Selective identification and site-specific quantification of 4-Hydroxy-2-nonenal-modified proteins. *Anal. Chem.* 91, 5235–5243. doi: 10.1021/acs.analchem.8b05970

Zhang, Y., Xie, R., Wang, J., Leier, A., Marquez-Lago, T. T., Akutsu, T., et al. (2019). Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework. *Brief. Bioinform.* 20, 2185–2199. doi: 10.1093/bib/bby079

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

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.

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:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (1)$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \quad (2)$$

$$\mathrm{Ac} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \quad (3)$$

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

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

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

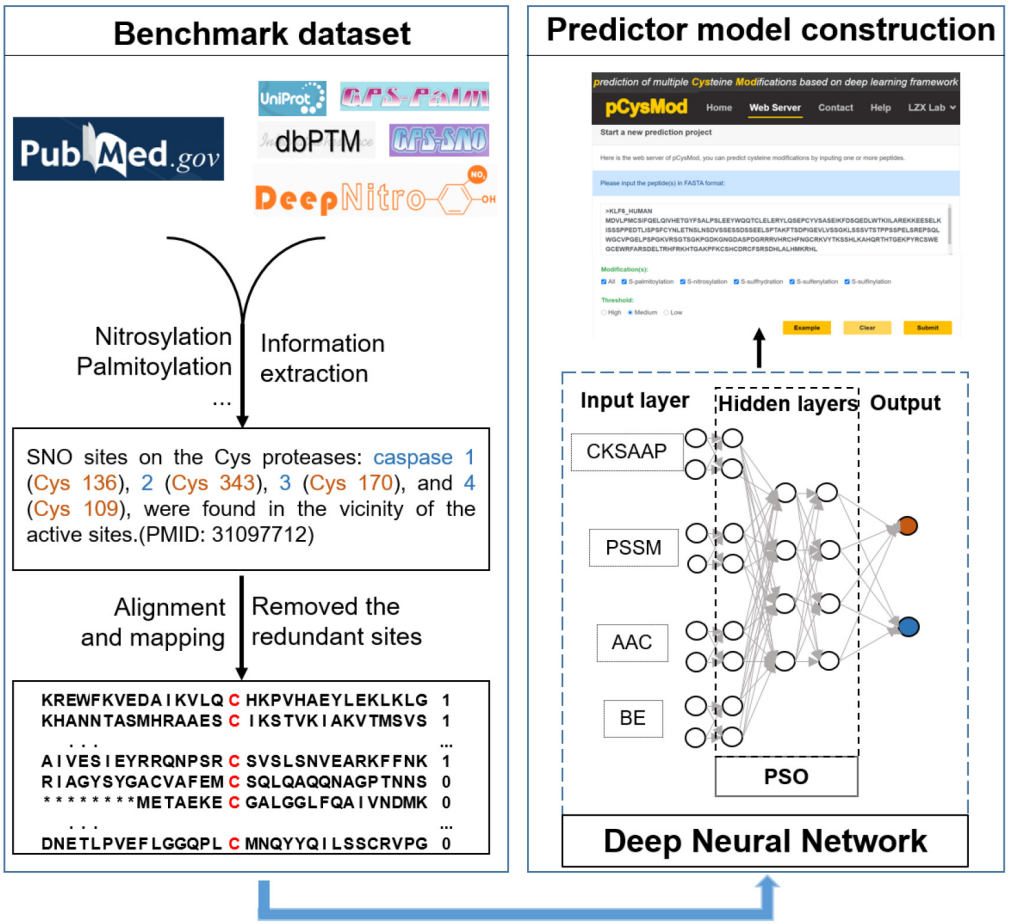


FIGURE 1 | An overview of the model.

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

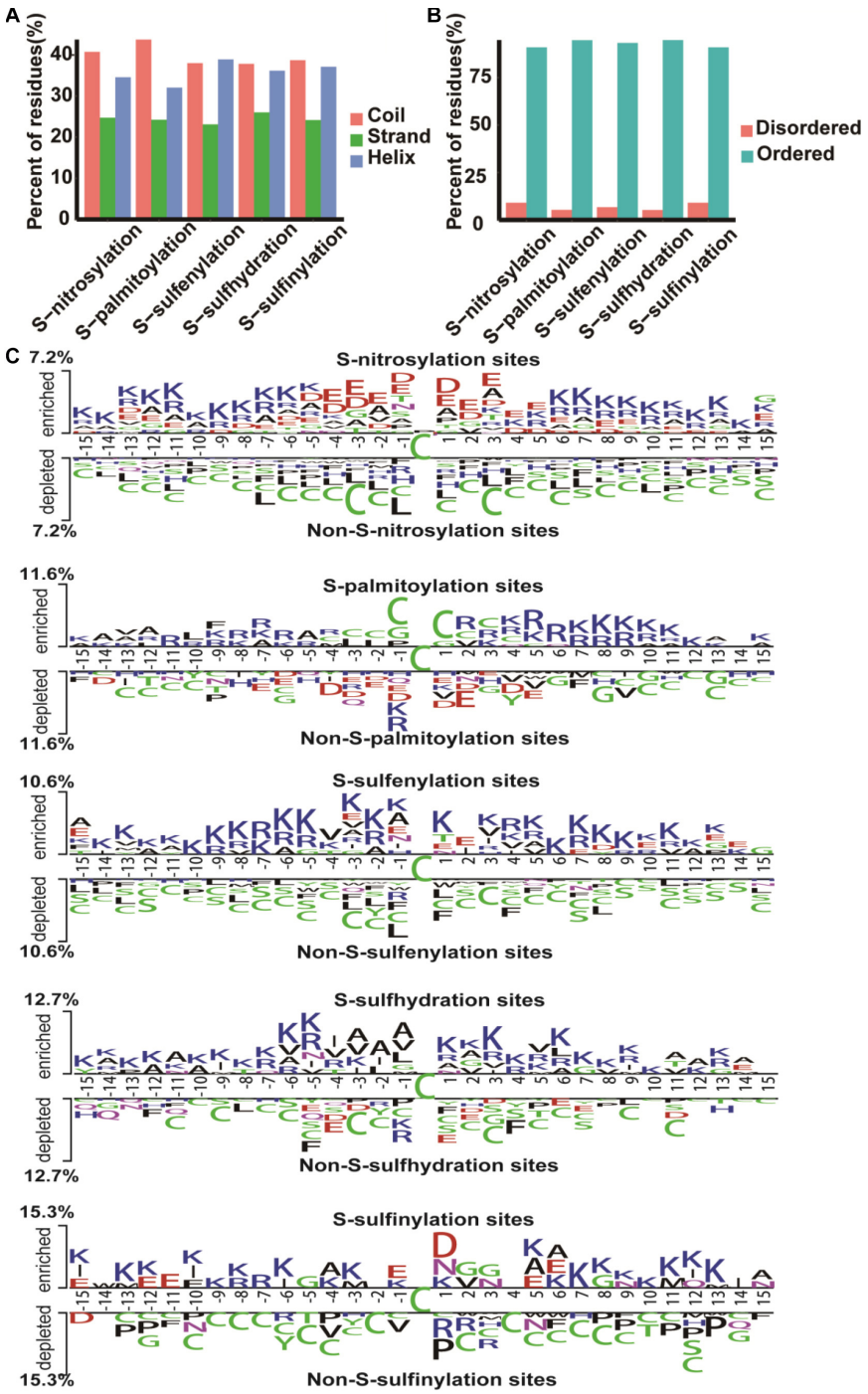


FIGURE 2 | The characteristic of cysteine modification sites and proteins. **(A)** The secondary structure. **(B)** The disorder information of cysteine modification sites. **(C)** Preference for amino acids around the cysteine modification sites and non-cysteine modification sites.

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.,

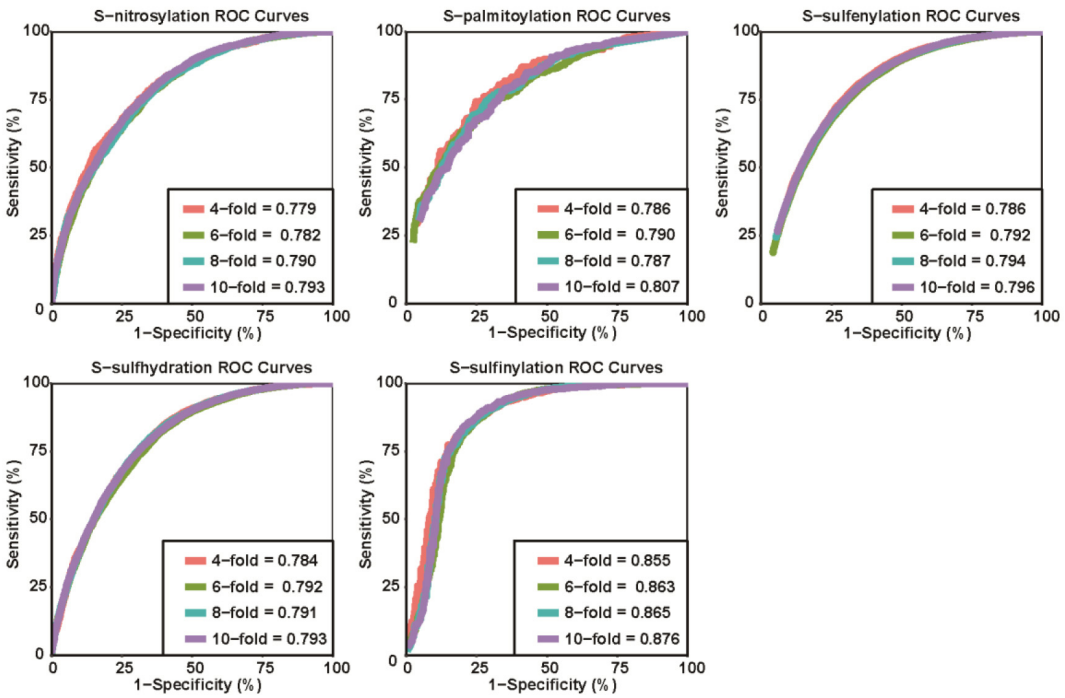


FIGURE 3 | The ROC curves and AUCs of 4-, 6-, 8-, and tenfold cross-validations are shown.

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

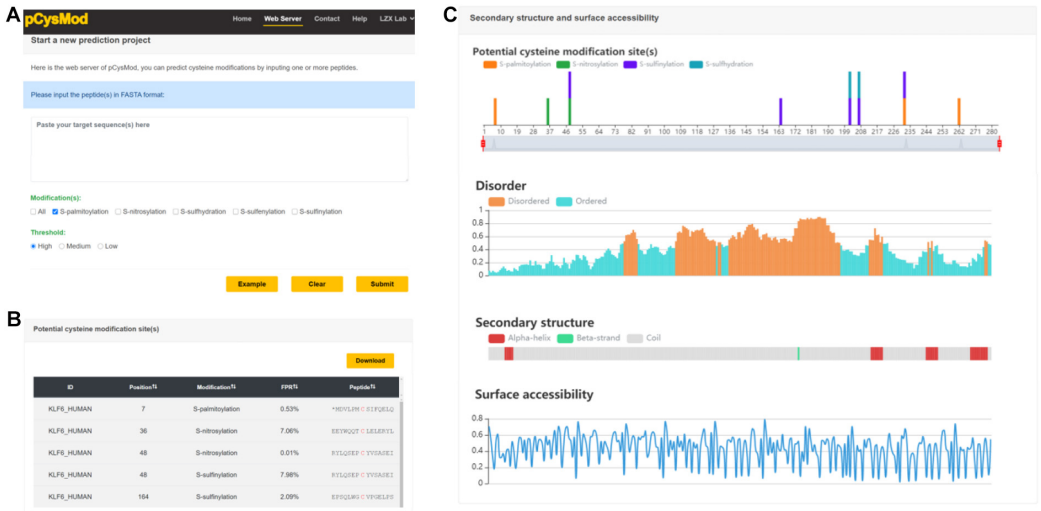


FIGURE 4 | The web server of pCysMod. **(A)** The prediction page. **(B)** Potential cysteine modification sites. **(C)** Secondary structure and surface accessibility.

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. doi: 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.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 23 March 2021
doi: 10.3389/fcell.2021.626821

ActiveDriverDB: Interpreting Genetic Variation in Human and Cancer Genomes Using Post-translational Modification Sites and Signaling Networks (2021 Update)

*Michal Krassowski[1], Diogo Pellegrina[2], Miles W. Mee[2], Amelie Fradet-Turcotte[3,4], Mamatha Bhat[5,6] and Jüri Reimand[2,7,8]**

[1] Nuffield Department of Women's and Reproductive Health, Medical Sciences Division, University of Oxford, Oxford, United Kingdom, [2] Computational Biology Program, Ontario Institute for Cancer Research, Toronto, ON, Canada, [3] Department of Molecular Biology, Medical Biochemistry and Pathology, Universite Laval, Quebec, QC, Canada, [4] Oncology Division, Centre Hospitalier Universitaire (CHU) de Quebec-Universite Laval Research Center, Quebec, QC, Canada, [5] Multiorgan Transplant Program, University Health Network, Toronto, ON, Canada, [6] Division of Gastroenterology & Hepatology, Department of Medicine, University of Toronto, Toronto, ON, Canada, [7] Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada, [8] Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada

OPEN ACCESS

Edited by:
Jian Ren,
Sun Yat-sen University, China

Reviewed by:
Fuyi Li,
The University of Melbourne, Australia
Zexian Liu,
Sun Yat-sen University Cancer Center (SYSUCC), China
Peter Van Hornbeck,
Cell Signaling Technology (United States), United States

***Correspondence:**
Jüri Reimand
Juri.Reimand@utoronto.ca

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

Received: 06 November 2020
Accepted: 08 February 2021
Published: 23 March 2021

Citation:
Krassowski M, Pellegrina D, Mee MW, Fradet-Turcotte A, Bhat M and Reimand J (2021) ActiveDriverDB: Interpreting Genetic Variation in Human and Cancer Genomes Using Post-translational Modification Sites and Signaling Networks (2021 Update). Front. Cell Dev. Biol. 9:626821. doi: 10.3389/fcell.2021.626821

Deciphering the functional impact of genetic variation is required to understand phenotypic diversity and the molecular mechanisms of inherited disease and cancer. While millions of genetic variants are now mapped in genome sequencing projects, distinguishing functional variants remains a major challenge. Protein-coding variation can be interpreted using post-translational modification (PTM) sites that are core components of cellular signaling networks controlling molecular processes and pathways. ActiveDriverDB is an interactive proteo-genomics database that uses more than 260,000 experimentally detected PTM sites to predict the functional impact of genetic variation in disease, cancer and the human population. Using machine learning tools, we prioritize proteins and pathways with enriched PTM-specific amino acid substitutions that potentially rewire signaling networks via induced or disrupted short linear motifs of kinase binding. We then map these effects to site-specific protein interaction networks and drug targets. In the 2021 update, we increased the PTM datasets by nearly 50%, included glycosylation, sumoylation and succinylation as new types of PTMs, and updated the workflows to interpret inherited disease mutations. We added a recent phosphoproteomics dataset reflecting the cellular response to SARS-CoV-2 to predict the impact of human genetic variation on COVID-19 infection and disease course. Overall, we estimate that 16-21% of known amino acid substitutions affect PTM sites among pathogenic disease mutations, somatic mutations in cancer genomes and germline variants in the human population. These data underline the potential of interpreting genetic variation through the lens of PTMs and signaling networks. The open-source database is freely available at www.ActiveDriverDB.org.

Keywords: post-translational modifications (PTM), genome variation, disease genes, cancer drivers, cell signaling, protein interaction networks, databases

INTRODUCTION

Genome-wide sequencing and association studies are rapidly increasing the catalog of human genetic variation such as single-nucleotide variants (SNVs) responsible for phenotypic traits and disease risks (Claussnitzer et al., 2020; Karczewski et al., 2020; The 1000 Genomes Project Consortium, 2015). Sequencing of cancer genomes reveals a complex landscape of somatic variation where a minority of driver mutations enable the oncogenic properties of cells by altering the activity of cancer genes and molecular pathways (Bailey et al., 2018; ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium., 2020; Reyna et al., 2020). Extensive somatic variation found in healthy cells in normal tissues (Blokzijl et al., 2016; Martincorena et al., 2015) adds another dimension of genetic complexity and suggests that populations of cells with distinct genetic makeups are present in every individual. Characterizing the implications of genome variation to cellular and physiological function and disease pathogenesis remains a difficult computational and experimental challenge (Gonzalez-Perez et al., 2013; MacArthur et al., 2014).

Post-translational modifications (PTMs) are core components of signaling networks that expand the functional range of proteins by controlling protein activation, degradation, and protein–protein interactions. PTMs are chemical or polypeptide modifications of amino acids that act as molecular switches. Various enzymes add or remove modifications on substrate proteins or read the modified sites to carry out cellular programs (Pawson, 1995). Signaling networks of PTMs are a major focus of therapy development (Gharwan and Groninger, 2016; Hoeller and Dikic, 2009; Jones et al., 2016). Phosphorylation, acetylation, methylation, and ubiquitination are among the most commonly occurring PTMs in human cells whereas hundreds of classes of PTMs are known (Mann and Jensen, 2003; Montecchi-Palazzi et al., 2008). These PTMs are now routinely mapped using high-throughput techniques and consequently, large public datasets for human proteins are available. Major databases such as PhosphoSitePlus (Hornbeck et al., 2015), UniProt (UniProt Consortium, 2019) and others maintain consistently updated collections of PTM sites derived from high-throughput and focused experimental studies.

PTM sites in human proteins are known to be enriched in somatic driver mutations in cancer genomes (Creixell et al., 2015; Radivojac et al., 2008; Reimand and Bader, 2013; Reimand et al., 2013; Wang et al., 2015) and germline variants implicated in the pathogenesis of human diseases and cancer predisposition (Huang et al., 2018; Li et al., 2010; Reimand et al., 2015). In contrast, PTM sites are depleted of genetic variation in the general human population, indicating the functional importance of conserved PTM signaling and the role of evolutionary constraint (Li et al., 2010; Reimand et al., 2015). Thus, integrative analyses of genetic variation using PTMs is likely to contribute to our understanding of molecular and genetic mechanisms. Besides the amino acid substitutions replacing the central modified residue of a PTM site, a larger class of substitutions affects PTMs by altering the short linear motifs recognized by kinases and other enzymes (Creixell et al., 2015; Reimand et al., 2013; Wagih et al., 2015). For example, the sequence motifs targeted by the ubiquitination system and controlling the degradation of cancer driver proteins are commonly affected by somatic mutations (Martínez-Jiménez et al., 2020; Narayan et al., 2016). As a canonical example of PTM-associated cancer driver mutations, substitutions in the N-terminal phosphosites of the oncogene beta-catenin (CTNNB1) stabilize the protein by disrupting phosphorylation-dependent ubiquitylation (Morin et al., 1997), causing downstream activation of the Wnt pathway and resulting in oncogenesis in diverse cancer types. In a recent study, hotspot somatic substitutions in the splicing factor 3B subunit 1 (SF3B1) at the ubiquitinated residue K700 were shown to abolish ubiquitylation, disrupt its mRNA interactions and cause altered splicing of a subset of transcripts (Zhang et al., 2019), consistent with our earlier analysis (Narayan et al., 2016). As proteomic and genetic datasets grow rapidly, systematic analyses and data resources allow researchers to study potential disease mechanisms involving genetic variation in signaling networks.

We developed the ActiveDriverDB database (www.ActiveDriverDB.org) to facilitate integrative analyses of human genetic variation and PTM sites. We present a major update to our original publication (Krassowski et al., 2017) that includes additional genomic and proteomic datasets, new types of PTMs and improved workflows. We included a phosphoproteomics dataset of SARS-CoV-2 response (Bouhaddou et al., 2020) to enhance integrative analyses of human population variation and infection-specific PTMs. This article describes the major workflows of our database and reviews the recent updates.

RESULTS

The ActiveDriverDB Server

ActiveDriverDB is a web-based database for interpreting protein-coding variation in human genomes using PTM sites (**Figure 1**). Our leading hypothesis is that amino acid substitutions caused by SNVs in PTM sites can alter signaling networks by creating, altering, and disrupting the sites. Genetic variation of PTM sites can affect modification and downstream signaling directly by substituting the modified residue or indirectly by modifying the consensus binding sequences (i.e., short linear motifs) located in the flanking sequences of post-translationally modified residues. Thus, an integrated analysis of PTM sites and genetic variation can evaluate the functional impact of variants and lead to mechanistic insights.

To address this hypothesis, we collected more than quarter of a million unique, experimentally detected PTM sites in the human proteome using the powerful resources available in the public databases PhosphoSitePlus (Hornbeck et al., 2015), UniProt (UniProt Consortium, 2019), Phospho.ELM (Dinkel et al., 2011), and HPRD (Keshava Prasad et al., 2009; **Figures 1A**, **2A,B**). ActiveDriverDB covers seven major types of PTMs with the largest sets of experimental data available for the human proteome. These include 149,299 phosphorylation sites (57%), 87,852 ubiquitination sites (34%), 12,380 methylation sites (4.7%), 11,394 acetylation sites (4.4%), and three types of PTM sites added in the 2021 update of the database: 6,081

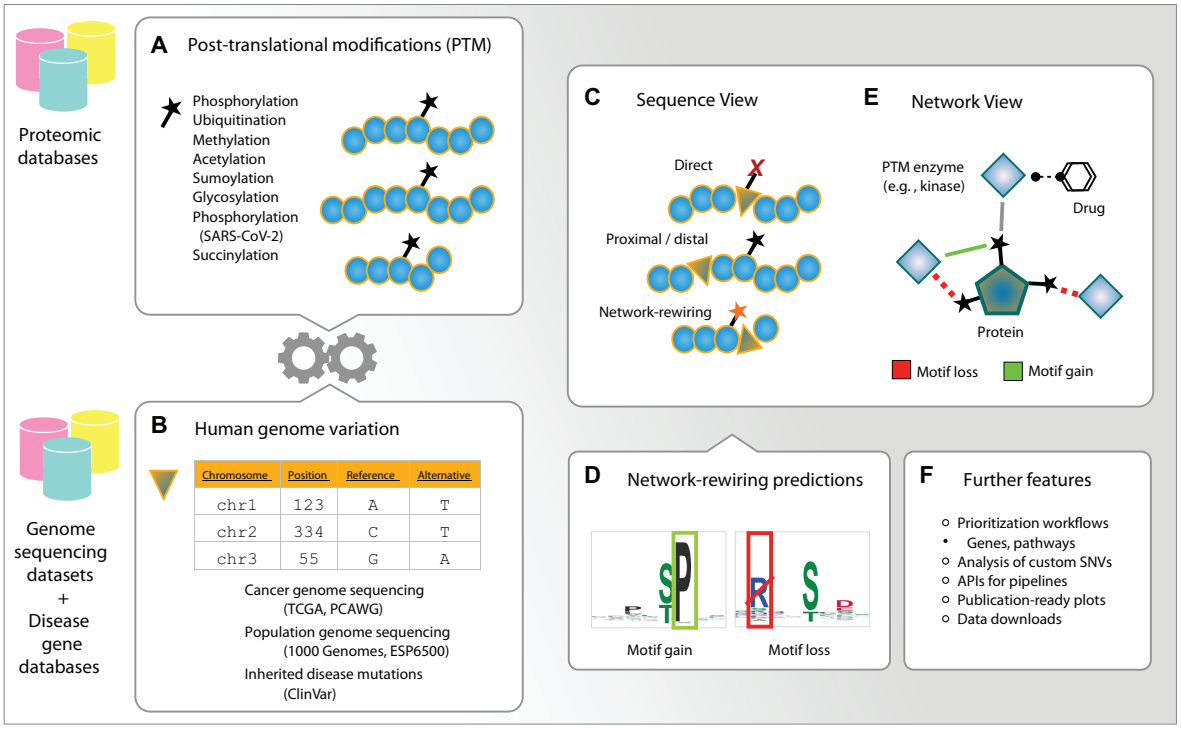


FIGURE 1 | Outline of ActiveDriverDB. ActiveDriverDB is an interactive proteo-genomics database for interpreting human genetic variation using post-translational modification (PTM) sites. **(A,B)** The database integrates PTM sites from experimental studies collected from proteomics databases with amino acid substitutions from genome sequencing projects and curated databeses of disease mutations. **(C)** In the *Sequence View*, substitutions in PTM sites are classified based on their functional impact as direct (at a PTM residue), proximal or distal (within 1–2 or 3–7 positions of a PTM residue), or network-rewiring. **(D)** Network-rewiring substitutions at PTM sites are predicted to disrupt short linear motifs or create new motifs bound by kinases and other enzymes. **(E)** In the *Network View*, proteins and PTM sites are visualized with their interactions with PTM enzymes (e. g., kinases) and the known drugs targeting the enzymes. **(F)** The database also provides prioritized lists of genes and pathways, comprehensive data visualizations and an application user interface (API) for analysing custom variant datasets using computational pipelines.

glycosylation sites (2.3%), 8,049 sumoylation sites (3.1%), and 203 succinylation sites (0.08%). The 261,348 unique PTM sites occur in proteins encoded by 15,570 genes (i.e., 82% of protein-coding genes). Different types of PTMs are known to act in concert in important cellular processes (Dantuma and van Attikum, 2016). Consistently, a fraction of mutated PTM sites (5.5%) is affected by multiple types of PTMs, suggesting that such complex signaling activities may be altered through amino acid substitutions. In this article, we summarize the counts of PTM sites and substitutions in canonical protein isoforms for individual genes, however, our database includes all high-confidence protein isoforms with 552,068 PTM sites. These data show the extent of PTMs in the human proteome and underline their value in interpreting protein-coding genome variation using our database.

We analyzed human genetic variation datasets of three classes using flanking sequences of seven amino acids on both sides of the post-translationally modified residue (**Figures 1B**, **2A,B**). First, we integrated the ClinVar catalog of inherited disease mutations (Landrum et al., 2020) with 237,930 unique amino acid substitutions, of which 65,162 (27%) affected PTM sites. We prioritized 28,976 substitutions classified as *pathogenic* or *likely pathogenic* in ClinVar and found that 6,913 (24%) of these affected PTM sites. When considering the entire ClinVar dataset of disease-associated substitutions, 22% occurred in PTM sites (65,162/237,930). Second, we integrated somatic genome variation of human cancers of nearly 40 types, including the Cancer Genome Atlas (TCGA) PanCanAtlas dataset with ∼10,000 cancer exomes (Ellrott et al., 2018), as well as the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset with ∼2,500 whole cancer genomes (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium., 2020) added in the 2021 update of our database. This resulted in a total of 889,792 unique amino acid substitutions, of which 179,470 (20%) affected PTM sites. Third, we integrated two datasets of genome variation in the human population, the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015) and ESP6500 (Tennessen et al., 2012) with a total of 1,047,196 unique amino acid substitutions, of which 217,932 (21%) affected PTM sites. Together, these genetic maps include 2,049,883 unique amino acid substitutions of which 436,192 (21%) are predicted to affect PTM sites. Our variant impact predictions show the strongest effects on a subset of substitutions in PTM sites: 37,186 (8.5%) substitutions replace the central PTM residue and therefore likely to abolish PTMs, and 35,136 (8.1%) are predicted to create or disrupt kinase-binding

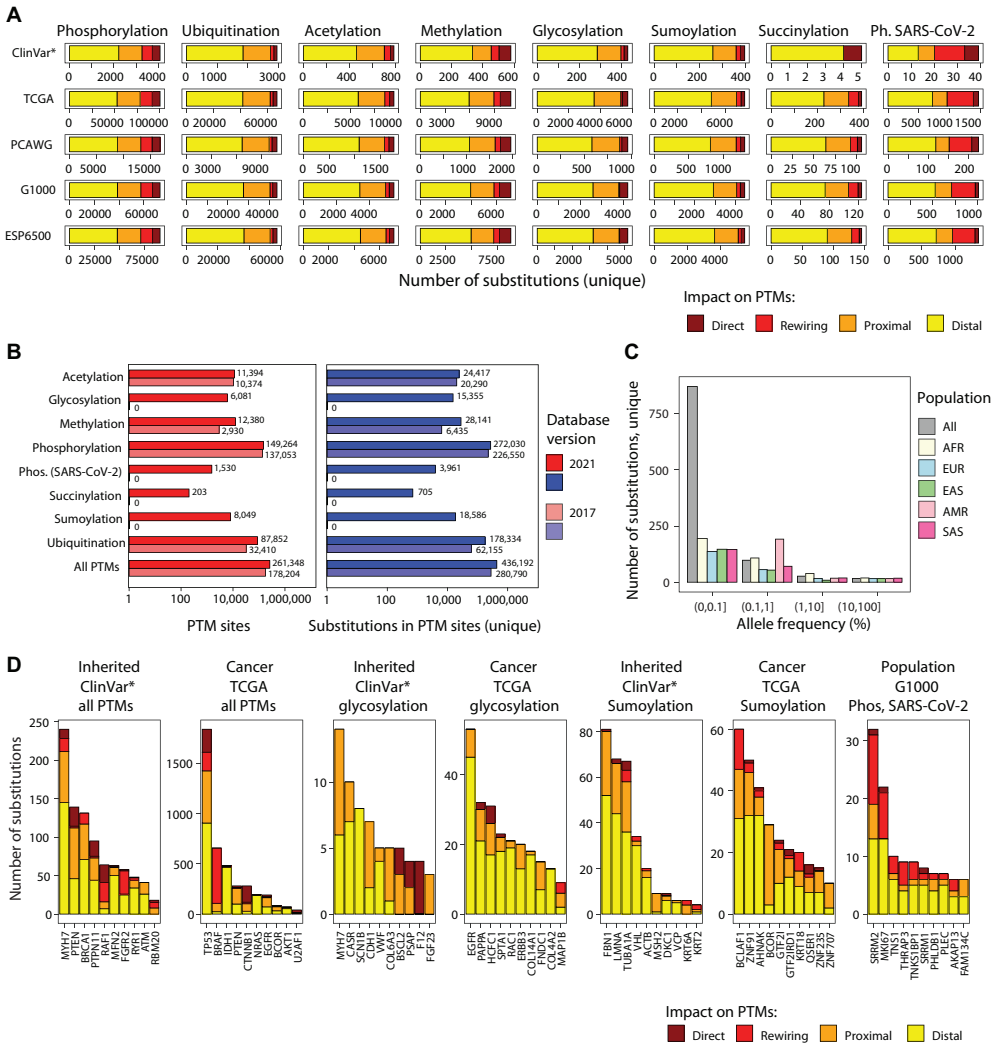


FIGURE 2 | PTM sites and mutations in ActiveDriverDB. **(A)** Summary of genetic variants (i.e., amino acid substitutions) affecting PTM sites in the database. Eight types of PTM sites are shown as horizontal stacked bar plots (left to right) with five genome variation data-stes (top to bottom): interited disease mutations (*ClinVar: only pathogenic and likely pathogenic variants), somatic cancer mutations (TCGA, PCAWG) and human population variation (1000 Genomes, ESP6500). Colors indicate the predicted impact of substitution on PTM sites. Total numbers of unique PTM-associated substitutions in consensus protein isoforms are shown. **(B)** Bar plot shows counts of PTM sites and relatedd substitutions in ActiveDriverDB. The current and previous versions of the database are compared. **(C)** Allele frequency of substitutions in the human population (1000 Genomes) affecting the phosphosites modulated by the SARC-CoV-2 infection in Vero E6 cells. Population cohorts are shown in colors (AFR, African; Admixed American; EAS, East Asian; EUR, European; SAS, South Asian). **(D)** Top genes with PTM-related substitutions in all PTM sites in inheried disease and cancer, genes with glycosylation and sumoylation-associated subtitutions, and top genes in the human population with SARS-CoV-2-specific phosphosites affected by substitutions. Colors indicate the predicted impact of substitutions on PTM sites. Genes were prioritized using ActiveDriver (FDR < 0.05), except for the rightmost group where unique substitution counts were used.

motifs by substituting important amino acid residues within seven positions of PTM sites (Wagih et al., 2015). The majority of substitutions are classified as proximal (30%) or distal (53%) and are located at 1–2 or 3–7 positions from the nearest PTM site, respectively (**Figure 3A**). Most proximal and distal substitutions cannot be interpreted reliably in the context of known kinase-binding motifs; however, these may affect uncharacterized sequence motifs of phosphorylation and other PTM types or cause smaller alterations of sequence motifs (**Figures 3B,C**). The genomic variation of amino acid substitutions in PTM sites provides a wealth of novel hypotheses for further computational and experimental studies to understand genotype–phenotype associations and PTM function.

The Sequence View

The first major workflow of ActiveDriverDB starts with a gene ID of interest provided by the user. The database displays an interactive color-coded overview of the protein sequence where the amino acid substitutions are annotated with respect to their impact on PTM sites and their frequency in the genetic dataset (**Figure 1C**). The user may choose to focus on cancer genomes, inherited diseases, or genome variation in the human population.

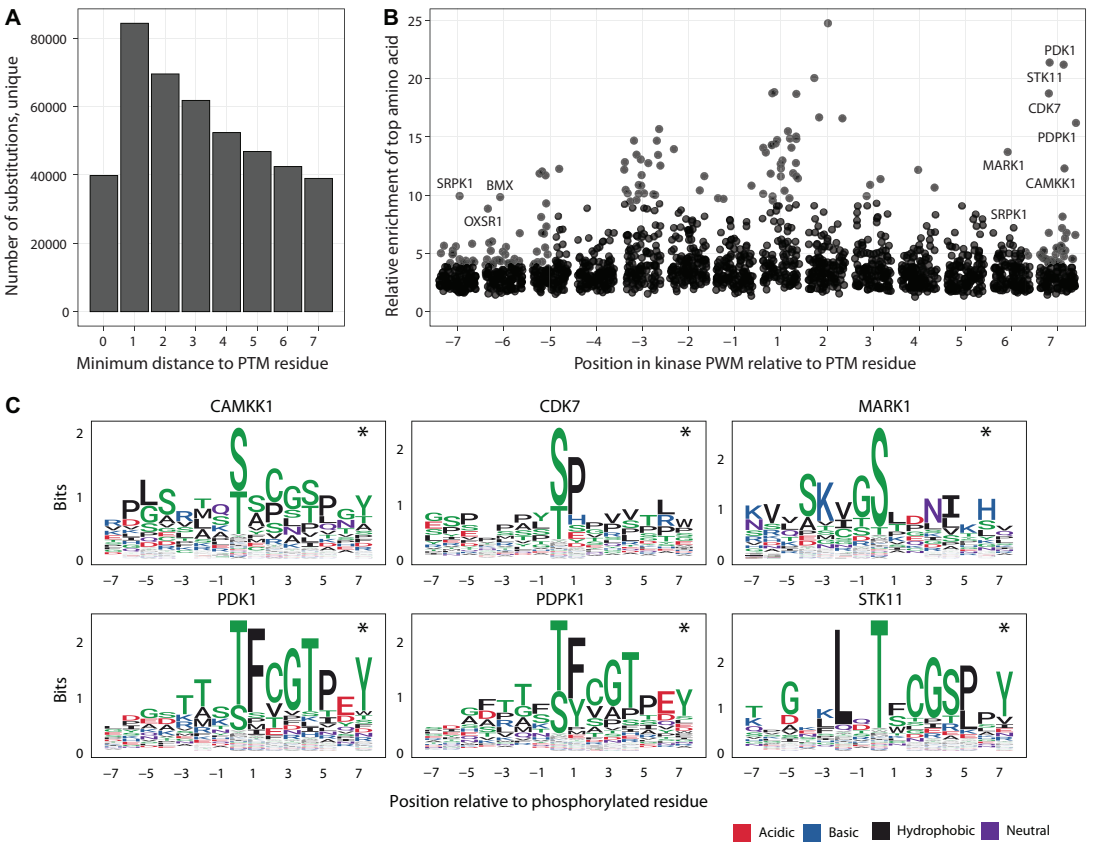


FIGURE 3 | Putative impact of adjacent and distal PTM-flanking residues on kinase binding motifs. **(A)** Histogram of substitutions in PTM sites relative to the distance to the closest modified residues. **(B)** Enrichments of amino acids in the 125 kinase binding site models of position weight matrices (PWMs). Each point represents a position in the consensus binding sequence (short linear motif) of a specific kinase. For each flanking position in the motif (X-axis), the amino acid with the highest enrichment relative to its proteome-wide distribution is shown on the Y-axis, indicating the potential impact of sustitutions at these positions. Kinases with amino acids showing at least eight-fold enrichment at the furthest flanking positions (6th, 7th) are labelled. **(C)** Examples of kinases with enrichments at the 6th and 7th flanking positions of PTM sites. PWM logos show the prevalence of specific amino acids (Y-axis) at the flanking positions (X-axis). Asterisks show the furthest flanking positions from panel A.

The data can be filter based on the disease subtype, type of PTM or the annotations of genetic variants. Four categories are used to classify the PTM-specific impacts of substitutions. *Direct* mutations substitute a central, modified residue of a PTM site with another non-modifiable amino acid residue that will likely disrupt PTMs at the site. *Proximal* and *distal* mutations induce a substitution within 1–2 or 3–7 residues, respectively, from the closest PTM site. For a subset of distal and proximal mutations, we predict that the substitutions have a plausible *network-rewiring* effect since they disrupt an existing short linear motif of a known kinase or other PTM enzyme (i.e., *motif loss*) or create a new sequence motif (i.e., *motif gain*) in the flanking sequence of the PTM site (**Figures 1D**, **2A**). Network-rewiring mutations are predicted using the MIMP method that uses a machine-learning framework of Gaussian mixture models and Bayesian posterior probability estimation to quantify the impact of substitutions on short linear motifs (Wagih et al., 2015). The Sequence View also displays a table of mutations and their impact on PTM sites, information on protein domains (Finn et al., 2017), evolutionary conservation (Pollard et al., 2010) and disorder (Ward et al., 2004), and hyperlinks to external databases. This view allows researchers to construct experimentally testable hypotheses of variant function and associations with phenotypes and disease.

The Network View

The second major workflow starts from a gene of interest in a protein–protein interaction network. The network shows the protein as the central node (i.e., the substrate) and all kinases and other PTM enzymes targeting the protein are shown as peripheral nodes. Approved drugs targeting these PTM enzymes, derived from the DrugBank database (Wishart et al., 2018), are displayed via secondary peripheral interactions of the network. The Network View focuses on enzyme–substrate interactions that occur at individual PTM sites and provides predictions of substitutions causing gains and losses of these interactions through altered sequence motifs, derived from the MIMP method (Wagih et al., 2015). Two types of networks are provided. First, the high-confidence *experimental networks* only include experimentally

validated enzyme–substrate interactions at specific PTM sites collected from databases and previous studies (Hornbeck et al., 2015; Reimand and Bader, 2013; UniProt Consortium, 2019; Wagih et al., 2015). The lenient *MIMP-predicted networks* include computationally predicted interactions at confirmed PTM sites based on the presence of known kinase binding motifs or *de novo* motifs induced by amino acid substitutions (Wagih et al., 2015). This systems-levels overview of PTM-associated mutations helps predict their impact on downstream signaling networks and discover potential avenues for experimental modulation.

Gene and Pathway Prioritization

We statistically analyzed PTM sites and amino acid substitutions to nominate statistically significant cancer driver genes, inherited disease genes, and molecular pathways with enrichments of PTM-associated substitutions (FDR < 0.05), using methods we developed previously (Paczkowska et al., 2020; Reimand and Bader, 2013). The database includes top-ranking genes with frequent PTM-associated mutations in inherited disease and multiple types of cancer (**Figure 2D**). The genes were prioritized using the ActiveDriver method that uses a Poisson statistical model to identify significant over-representations of substitutions at the PTM sites of individual proteins (Reimand and Bader, 2013). For pathway prioritization, genes with enriched substitutions in PTM sites were collapsed into enriched Gene Ontology terms and Reactome molecular pathways using the ActivePathways data fusion method (Paczkowska et al., 2020). Lists of genes and pathways were derived for the combined set of all PTMs, and also separately for each PTM type. To prioritize genes involved in inherited disease, we focused on the mutations with pathogenic or likely pathogenic effects. Gene and pathway prioritization allows researchers to find individual genes and groups of functionally related genes with PTM-associated disease mutations.

Searching, Data Downloads, and Automated Analysis

ActiveDriverDB can be queried interactively and included in automated pipelines. The most common approach is to search the database interactively using a gene symbol or RefSeq ID (e.g., *TP53* or *NM_000345*), or a specific amino acid substitution or a SNV in the GRCh37 version of the human genome (e.g., *IDH1 R132H* or *chr2 209113112 G A*). The database can be queried using names of molecular pathways (e.g., *R-HSA-1640170* or *Cell Cycle*) or diseases (e.g., *Noonan syndrome*) and all genes with such annotations are retrieved. Users can upload a dataset of genetic variants from their experiments to a password-protected area of the database and analyze their data interactively. The upload form supports protein and DNA coordinates of genetic variants. ActiveDriverDB can be used computationally via an Application User Interface (API) of the Representational State Transfer (REST) pattern that provides automated tools to annotate genetic datasets using PTM information. The datasets used in the database are also available for bulk downloads. In this update, we have improved the annotations of PTM sites by adding names of source databases, several classes of protein IDs and flanking sequences of PTM sites. PubMed IDs are available for a subset of sites. The downloadable datasets include PTM sites, PTM-associated substitutions, site-specific enzyme–substrate interaction networks, protein sequences, and disorder predictions. We also provide interactive charts displaying the counts of PTM sites and associated substitutions in the database.

Genetic Variation in Phosphorylation Sites Induced by SARS-CoV-2 Infection

To enable detailed studies of the cellular changes induced by SARS-CoV-2 infection, we incorporated a recent dataset that quantified the proteome-wide phosphorylation changes in response to SARS-CoV-2 infection in Vero E6 cells of green monkeys (*Chlorocebus sabaeus*) (Bouhaddou et al., 2020). We integrated 1,530 unique SARS-CoV-2 modulated phosphosites in proteins encoded by 949 genes that were detected with significant phosphorylation differences in infected *vs.* control cells at the 24-hour post-infection time point (FDR < 0.05 in infected cells; FDR > 0.05 in controls). The majority of these phosphosites occur on serine residues (88%) followed by threonines (11.3%) and tyrosines (0.7%). We filtered a small subset of phosphosites (1%) that mapped to non-phosphorylatable residues in human proteins (i.e., other than S/T/Y) to avoid inclusion of non-human phosphorylation sites and potential sequence alignment artifacts. This dataset enables integrated analyses of human genome variation, PTM sites and signaling networks underlying the SARS-CoV-2 infection and the coronavirus disease (COVID-19) pandemic.

We evaluated the extent of human genome variation and known disease mutations affecting these phosphosites. ActiveDriverDB includes 3,961 amino acid substitutions affecting SARS-CoV-2-modulated phosphosites. These include 2,007 unique substitutions observed in the two human population cohorts (1000 Genomes; ESP6500) and 1,615 unique substitutions detected in somatic cancer genome sequencing projects (TCGA and PCAWG), and 39 unique substitutions with pathogenic or likely pathogenic effects documented in the ClinVar database (**Figure 2A**). We evaluated the impact of these PTM-associated substitutions. A relatively large fraction of substitutions (27%) were predicted to create or disrupt kinase binding motifs according to MIMP (Wagih et al., 2015). A minority of substitutions (5.1%) replaced the phospho-residue with another residue, likely causing direct disruptions of signaling. The remaining substitutions were considered as proximal (17%) or distal (51%) relative to the phosphosites. We also studied the allele frequencies of these PTM-specific substitutions in the human population and found that the majority of variants were of low frequency (i.e., less than 1%) in the 1000 Genomes Project dataset (The 1000 Genomes Project Consortium, 2015), however dozens of variants were more prevalent population-wide (**Figure 2C**). Of the most variable proteins with respect to SARS-CoV-2-specific PTM sites, two are related to alternative splicing (SRRM1, SRRM2) and one to cell cycle regulation (MKI67) (**Figure 2D**). Interestingly, altered SRRM2 phosphorylation has been also observed in HIV-1 infection (Wojcechowskyj et al., 2013). Collectively,

these data suggest that the variable cellular and physiological responses to SARS-CoV-2 infection in humans may have a genetic component that affects the PTM sites and signaling networks that respond to viral infection. Further analysis and experiments may lead to insights to disease mechanisms and therapy options.

Interpreting Genetic Variation Through Protein Glycosylation

Glycosylation is a type of PTM that involves the conjugation of diverse glycan structures to proteins, in particular extracellular components such as receptors and secreted proteins (reviewed in Moremen et al., 2012; Reily et al., 2019). Glycosylation modifications are conducted by approximately 700 enzymes and multiple subtypes are known, whereas N- and O-linked glycosylation are the most common subtypes. Glycosylation is involved in the folding and quality control of proteins and modulates protein function and protein–protein interactions. Glycosylation of extracellular protein domains in cell–cell signaling contributes to developmental processes and the immune system (Moremen et al., 2012). Aberrant glycosylation patterns, often linked to genetic abnormalities of specific glycosylation enzymes, play important roles in autoimmune diseases such as inflammatory bowel disease, diabetes mellitus, systemic lupus, and congenital disorders of glycosylation (Reily et al., 2019). In cancer, glycosylation is involved in the pathways of metastasis, anti-apoptosis and therapy resistance, and the PTM is also used in diagnostic and prognostic biomarkers (Reily et al., 2019). The increasing availability of comprehensive glycoproteomic datasets generated in human samples (Chen et al., 2009; Liu et al., 2005; Wollscheid et al., 2009) enhances the interpretation of disease genes and mutations using this PTM type.

We collected 7,021 experimentally determined glycosylation sites (including 6,081 unique sites) in proteins encoded by 1,683 genes from proteomics databases (Hornbeck et al., 2015; Keshava Prasad et al., 2009; UniProt Consortium, 2019; **Figure 2B**). These include the major subtypes of N-glycosylation (2,680 sites) and O-glycosylation (2,856 sites), a few S- and C-linked glycosylation sites, and 1,437 glycosylation sites with no specified subtype. Interestingly, a fraction of proteins (167 or 10%) has glycosylation sites that co-occur with phosphorylation sites, indicating crosstalk of the underlying signaling networks. In total, we found 15,355 unique amino acid substitutions that affect glycosylation sites, including 429 substitutions with pathogenic or likely pathogenic effects in disease genes in the ClinVar dataset and 6,364 somatic substitutions in cancer genomes (**Figure 2A**). We selected the genes with most significant glycosylation-associated mutations in cancer and inherited disease using ActiveDriver (FDR < 0.05; top 10 genes shown) (**Figure 2D**). In cancer genomes, frequent substitutions at glycosylation sites are apparent in epidermal growth factor receptors and oncogenes EGFR and ERBB3, as well as PAPPA, a secreted protein involved in the activation of insulin-like growth factor pathways (Lawrence et al., 1999). Germline mutations with pathogenic or likely pathogenic effects at glycosylation sites are associated with cardiomyopathies (MYH7), cancer predisposition (CDH1), epilepsy (SCN1B), and others. These examples showcase an integrative analysis of disease mutations with protein glycosylation sites that may offer insights into disease mechanisms.

Interpreting Genetic Variation Through Protein Sumoylation

Sumoylation is a PTM that involves the reversible conjugation of SUMO polypeptides (small ubiquitin-related modifiers SUMO1-4) to consensus sequence sites in target proteins (reviewed in Geiss-Friedlander and Melchior, 2007; Flotho and Melchior, 2013; Celen and Sahin, 2020). Sumoylation plays a key role for the cellular response to stress, such as heat shock and DNA damage (Enserink, 2015). In response to DNA damage, sumoylation acts in concert with ubiquitylation events to orchestrate the recruitment of repair proteins to DNA breaks (Dantuma and van Attikum, 2016). A similar interplay of the two modifiers is observed in hypoxic stress response (Cheng et al., 2007). Sumoylation affects lysine residues primarily in nuclear proteins and is thought to regulate protein activation, inactivation and degradation, and protein–protein interactions. Aberrant sumoylation is implicated in malignancies including ovarian, lung, breast, and prostate cancer (Celen and Sahin, 2020; Geiss-Friedlander and Melchior, 2007). Defects in sumoylation are also associated with neurodegenerative pathologies such as Huntington's, Parkinson's and Alzheimer's diseases (reviewed in Yang et al., 2017). Finally, sumoylation is involved in intrinsic and innate immunity and is a target of viral infection (Hu et al., 2016; Liu et al., 2016).

The updated ActiveDriverDB database includes 8,049 experimentally determined sumoylation sites in 2,478 unique genes primarily collected from PhosphoSitePlus (Hornbeck et al., 2015). Interestingly, more than half of sumoylation sites (4,783 or 59%) co-occur with other types of PTMs, in particular ubiquitination sites. We found 19,226 amino acid substitutions at sumoylation sites (16,914 unique), including 8,450 substitutions in the human population genomics datasets, 8,465 somatic substitutions in cancer genomes, and 397 pathogenic or likely pathogenic substitutions of the ClinVar database, suggesting potential disease mechanisms at mutated sumoylation sites. Driver gene analysis of PTM-enriched amino acid substitutions revealed multiple genes with germline and somatic mutations. In the TCGA cancer genomics dataset, the transcription factors (TFs) BCOR (BCL6 corepressor, FDR = 1.2×10^{-35}) and BCLAF1 (Bcl-2-associated transcription factor 1; ActiveDriver FDR = 9.8×10^{-4}) were significantly enriched in substitutions in glycosylation sites. Both TFs act as transcriptional repressors of apoptosis and are known as cancer driver genes in the COSMIC Cancer Gene Census database (Futreal et al., 2004). Several other TFs of the less-studied zinc finger family were found in the analysis (**Figure 2D**). Sumoylation is known as a mechanism of modulating TF activity, thus somatic substitutions in PTM sites may lead to aberrant TF activity in cancer and cause downstream transcriptional deregulation of cancer hallmark pathways. Further study of these substitutions at PTM sites may refine our understanding of known cancer genes and reveal novel candidates.

Interpreting Genetic Variation Through Protein Succinylation

Succinylation is a PTM that involves the transfer of succinyl groups to lysine residues of substrate proteins via enzyme-dependent and independent means (reviewed in Sreedhar et al., 2020; Trefely et al., 2020). Succinylation has been described only recently (Zhang et al., 2011) and its molecular mechanisms are not fully understood. The highest levels of succinylation are found in mitochondrial proteins, however, high-throughput studies have also detected modifications of cytoplasmic and nuclear proteins. The succinyltransferases CPT1A and KAT2A conduct target-specific modifications while succinyl turnover is controlled by the sirtuin proteins SIRT5 and SIRT7 that regulate bulk succinylation and DNA-damage-dependent succinylation, respectively (Du et al., 2011; Li et al., 2016). The modification is increasingly implicated in transcriptional regulation as histone proteins are often succinylated and site mutations have functional consequences (Smestad et al., 2018; Xie et al., 2012). However, the lysine residues affected by succinylation also undergo other PTMs such as acetylation, methylation and ubiquitylation. Therefore, more research is needed to understand the role of succinylation and its interactions with other PTMs in core cellular processes and human disease (Sreedhar et al., 2020).

Our database includes 203 unique, experimentally determined succinylation sites in proteins encoded by 63 genes, all of which co-occur with other lysine PTMs such as acetylation, methylation, ubiquitylation and sumoylation. Using ActiveDriverDB, we found 772 amino acid substitutions at succinylation sites (705 unique), including 250 substitutions in the human population genomics datasets and 462 somatic substitutions in cancer genomes. In the TCGA cohort of cancer genomes, our analysis highlighted several genes encoding histone proteins (H3J, H2BB, H2BG), reinforcing the role of succinylation in chromatin regulation and suggesting potential PTM-specific driver mutations. In the ClinVar dataset of pathogenic or likely pathogenic mutations, two histone proteins (H3F3A, HIST1H4C) and the copper-zinc superoxide dismutase 1 (SOD1) were highlighted. Mutations in SOD1 are associated with familial amyotrophic lateral sclerosis (Rosen et al., 1993). SOD1 regulates the accumulation of harmful superoxide radicals in cells and coordinated succinylation is required for its function (Lin et al., 2013) whereas mutations impacting its catalytic activity induce the formation of fibrillar aggregates that are toxic for cells (DiDonato et al., 2003). ActiveDriverDB highlights three substitutions flanking the succinylated residue K123 of SOD1 that are annotated as likely pathogenic for amyotrophic lateral sclerosis, suggesting potential hypotheses of these substitutions and altered succinylation in this lethal neurogenerative disease. Further succinylation-associated mutations and putative disease mechanisms are likely to be revealed as larger datasets of these PTM sites are published.

Improved Annotation of Pathogenic Germline Variants of Human Disease

We updated the collection of inherited disease mutations from the ClinVar database (Landrum et al., 2020) and improved the workflow of interpreting these using PTM sites. The new release of ActiveDriverDB includes 237,930 amino acid substitutions associated with human diseases, a four-fold increase compared to the ClinVar dataset included in the previous version of ActiveDriverDB (56,739). The data have been filtered carefully to only include variants with evidence of involvement in human disease. Genetic variants with germline, parental, maternal, and biparental and *de novo* origin are included in the database while variants of somatic and unknown origin are excluded to improve the analysis of inherited disease variants. Variants can be filtered based on clinical significance (such as *pathogenic*, *benign*, *drug response*, *etc.*) and a star rating reflecting the overall strength of evidence. Hyperlinks to the corresponding records in the databases ClinVar and dbSNP allow researchers to quickly access detailed descriptions of the variants and the original publications reporting the evidence of disease associations and pathogenesis. The updated variant filtering and annotations allow higher-confidence interpretation of disease variants with PTM information.

Evaluating the Importance of Distal Flanking Residues of PTM Sites Using Sequence Binding Motifs of Kinases

The majority of substitutions in PTM sites in our database are classified as distal and proximal and are located adjacent to modified residues, especially in the three flanking positions (**Figure 3A**). Only a minority of these substitutions are predicted to have network-rewiring effects since they affect critical sequence residues, however the flanking sequences of PTMs may contain additional functional residues that mediate weaker effects and therefore remain understudied in the database. To quantify the potential effects of proximal and distal substitutions in PTM sites, we systematically analyzed the 130 sequence-binding motifs of kinases used in our database. The motifs are represented as position weight matrices (PWMs) and used for network-rewiring predictions (Wagih et al., 2015). We quantified the PWMs in terms of the strongest amino acid enrichments at each position relative to the proteome-wide distributions of amino acids.

We found that each position of flanking sequence around the PTM sites included at least five-fold enrichment of specific amino acids in several sequence-binding models of kinases (**Figure 3B**). The strongest enrichments of specific amino acids occurred in the flanking windows of three residues around the modified residue. The three flanking positions are also covered by the most substitutions, indicating widespread genetic effects on PTM signaling. However, further positions upstream and downstream of the modified residue also appeared to encode some information with regards to kinase binding. Even when considering only the furthest positions six and seven of the PTM sites, the motifs of 28 kinases included at least five-fold enrichments of certain amino acids whereas more than ten-fold enrichments were observed for six kinases (CAMKK1, CDK7, MARK1, PDK1, PDPK1, and STK11) (**Figure 3C**). The effects measured here likely represent an underestimate since the sequence specificities of many PTM enzymes remain unknown. In summary, this analysis suggests that substitutions at both

proximal and distal flanking positions around the modified PTM sites may affect signaling networks.

Lastly, we asked whether the inclusion of the furthest flanking positions of six and seven from the PTM sites substantially biased our estimates of PTM-associated substitutions seen in known disease genes, in cancer genomes and the human population. Even when excluding the most distal amino acid substitutions at the flanking positions six and seven, a substantial fraction of all human amino acid substitutions is predicted to affect PTM sites. Using this more conservative estimate, PTM sites are affected by 17% of substitutions overall, including 19% of pathogenic or likely pathogenic substitutions in ClinVar and 22% of all ClinVar substitutions, 16% of somatic substitutions in cancer genomes, and 17% of substitutions in the human population genomics datasets. PTM sites, in particular when including the flanking sequences of seven amino acids, are enriched in disease mutations and negatively selected in the human population (Huang et al., 2018; Li et al., 2010; Reimand and Bader, 2013; Reimand et al., 2013; Reimand et al., 2015). Thus, additional functional substitutions likely exist in the flanking sequences of PTMs that cannot be interpreted yet using current proteomics datasets and computational models.

DISCUSSION

The increasing availability of genomic and proteomic technologies expedites the development of diverse applications in research, medicine and society. Human cells and tissues can be profiled at an improved resolution and decreased cost and cause an increasing influx of multi-omics datasets in the public domain. The collection of experimentally validated PTM sites in ActiveDriverDB has grown by 47% compared to the first release of the database in 2017 (261,348 vs. 178,204) while the dataset of disease-associated genome variants has quadrupled in size. Thus, we have the opportunity to interpret an ever-larger number of protein-coding variants in the human genome at an enhanced level of detail. In particular, the network-rewiring impact of variants is likely underestimated currently, since high-confidence short linear motifs are known only for a subset of kinases and other enzymes. Careful computational analysis of short linear motifs in conjunction with known PTM sites is required since such low-complexity motifs are statistically expected to occur frequently across the proteome. As we continue to expand the known repertoire of sequence-binding specificities of diverse PTM enzymes, we are increasingly able to predict the precise network-rewiring effects of substitutions in PTM sites observed in disease genes and the human population. Incorporation of protein structural information may further expand the collection of PTM-associated substitutions since linearly distant amino acids may affect PTMs through spatial interactions in the three-dimensional structures (Kamburov et al., 2015; Iqbal et al., 2020; Hu et al., 2021; Porta-Pardo et al., 2015). However, as the community rapidly generates larger and more sophisticated experimental datasets, the databases that use these for downstream analyses should be updated as well, since the analysis of -omics datasets with outdated annotations has detrimental effects on data interpretation (Wadi et al., 2016). In future updates of the database, we aim to specifically expand the genetic variation datasets mapping the human population, cancer genomes and inherited diseases. ActiveDriverDB and similar resources (Hornbeck et al., 2015; Wang et al., 2015; Li et al., 2020; Yang et al., 2019) allow a diverse community of molecular and cell biologists, geneticists and computational researchers to interpret complex genomic variation data using PTM sites and signaling networks and to explore detailed hypotheses of molecular mechanisms. These can contribute to the development of innovative therapies, biomarkers and precision medicine strategies.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. Processed data can be found here: https://actividriverdb.org/download/.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

MK developed the software, analyzed the data, and performed the data updates. MK, DP, MM, and JR analyzed the data and prepared the figures. AF-T and JR interpreted the data and reviewed the literature. JR wrote the manuscript with significant input from all co-authors. MB and AF-T contributed to project supervision. JR supervised the project. All authors reviewed and edited the manuscript and approved the final version.

FUNDING

MK was supported by the Scatcherd European Scholarship. This work was supported by the Canadian Institutes of Health Research (CIHR) Project Grant to JR, Cancer Research Society (CRS) Operating Grant to AF-T and JR, and the Investigator Award to JR from the Ontario Institute for Cancer Research (OICR). Funding to OICR is provided by the Government of Ontario, Canada.

ACKNOWLEDGMENTS

We are grateful to researchers and developers of databases PhosphoSitePlus, Phospho.ELM, HPRD, ClinVar, DrugBank, UniProt, and others for providing high-quality and frequently maintained datasets. The results published here are in part based upon data generated by the TCGA Research Network as outlined in the TCGA publication guidelines (http://cancergenome.nih.gov/).

REFERENCES

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., et al. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* 174, 1034–1035. doi: 10.1016/j.cell.2018.07.034

Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., et al. (2016). Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* 538, 260–264. doi: 10.1038/nature19768

Bouhaddou, M., Memon, D., Meyer, B., White, K. M., Rezelj, V. V., Correa Marrero, M., et al. (2020). The Global Phosphorylation Landscape of SARS-CoV-2 Infection. *Cell* 182, 685–712e619.

Celen, A. B., and Sahin, U. (2020). Sumoylation on its 25th anniversary: mechanisms, pathology, and emerging concepts. *FEBS J.* 287, 3110–3140. doi: 10.1111/febs.15319

Chen, R., Jiang, X., Sun, D., Han, G., Wang, F., Ye, M., et al. (2009). Glycoproteomics analysis of human liver tissue by combination of multiple enzyme digestion and hydrazide chemistry. *J. Proteome Res.* 8, 651–661. doi: 10.1021/pr8008012

Cheng, J., Kang, X., Zhang, S., and Yeh, E. T. (2007). SUMO-specific protease 1 is essential for stabilization of HIF1alpha during hypoxia. *Cell* 131, 584–595. doi: 10.1016/j.cell.2007.08.045

Claussnitzer, M., Cho, J. H., Collins, R., Cox, N. J., Dermitzakis, E. T., Hurles, M. E., et al. (2020). A brief history of human disease genetics. *Nature* 577, 179–189. doi: 10.1038/s41586-019-1879-7

Creixell, P., Schoof, E. M., Simpson, C. D., Longden, J., Miller, C. J., Lou, H. J., et al. (2015). Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell* 163, 202–217. doi: 10.1016/j.cell.2015.08.056

Dantuma, N. P., and van Attikum, H. (2016). Spatiotemporal regulation of posttranslational modifications in the DNA damage response. *EMBO J.* 35, 6–23. doi: 10.15252/embj.201592595

DiDonato, M., Craig, L., Huff, M. E., Thayer, M. M., Cardoso, R. M., Kassmann, C. J., et al. (2003). ALS mutants of human superoxide dismutase form fibrous aggregates via framework destabilization. *J. Mole. Biol.* 332, 601–615. doi: 10.1016/S0022-2836(03)00889-1

Dinkel, H., Chica, C., Via, A., Gould, C. M., Jensen, L. J., Gibson, T. J., et al. (2011). Phospho.ELM: a database of phosphorylation sites–update 2011. *Nucleic Acids Res.* 39, D261–D267. doi: 10.1093/nar/gkq1104

Du, J., Zhou, Y., Su, X., Yu, J. J., Khan, S., Jiang, H., et al. (2011). Sirt5 is a NAD-dependent protein lysine demalonylase and desuccinylase. *Science* 334, 806–809. doi: 10.1126/science.1207861

Ellrott, K., Bailey, M. H., Saksena, G., Covington, K. R., Kandoth, C., Stewart, C., et al. (2018). Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst.* 6, 271–281e277. doi: 10.1016/j.cels.2018.03.002

Enserink, J. M. (2015). Sumo and the cellular stress response. *Cell Div.* 10:4. doi: 10.1186/s13008-015-0010-1

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., et al. (2017). InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.* 45, D190–D199. doi: 10.1093/nar/gkw1107

Flotho, A., and Melchior, F. (2013). Sumoylation: a regulatory protein modification in health and disease. *Annu. Rev. Biochem.* 82, 357–385. doi: 10.1146/annurev-biochem-061909-093311

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., et al. (2004). A census of human cancer genes. *Nat. Rev. Cancer* 4, 177–183. doi: 10.1038/nrc1299

Geiss-Friedlander, R., and Melchior, F. (2007). Concepts in sumoylation: a decade on. *Nat. Rev. Mole. Cell Biol.* 8, 947–956. doi: 10.1038/nrm2293

Gharwan, H., and Groninger, H. (2016). Kinase inhibitors and monoclonal antibodies in oncology: clinical implications. *Nat. Rev. Clin. Oncol.* 13, 209–227. doi: 10.1038/nrclinonc.2015.213

Gonzalez-Perez, A., Mustonen, V., Reva, B., Ritchie, G. R., Creixell, P., Karchin, R., et al. (2013). Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods* 10, 723–729. doi: 10.1038/nmeth.2562

Hoeller, D., and Dikic, I. (2009). Targeting the ubiquitin system in cancer therapy. *Nature* 458, 438–444. doi: 10.1038/nature07960

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. doi: 10.1093/nar/gku1267

Hu, M. M., Yang, Q., Xie, X. Q., Liao, C. Y., Lin, H., Liu, T. T., et al. (2016). Sumoylation Promotes the Stability of the DNA Sensor cGAS and the Adaptor STING to Regulate the Kinetics of Response to DNA Virus. *Immunity* 45, 555–569. doi: 10.1016/j.immuni.2016.08.014

Hu, R., Xu, H., Jia, P., and Zhao, Z. (2021). KinaseMD: kinase mutations and drug response database. *Nucleic Acids Res.* 49, D552–D561. doi: 10.1093/nar/gkaa945

Huang, K., Mashl, R. J., Wu, Y., Ritter, D. I., Wang, J., Oh, C., et al. (2018). Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* 173, 355-370.e14. doi: 10.1016/j.cell.2018.03.039

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. (2020). Pan-cancer analysis of whole genomes. *Nature* 578, 82–93. doi: 10.1038/s41586-020-1969-6

Iqbal, S., Perez-Palma, E., Jespersen, J. B., May, P., Hoksza, D., Heyne, H. O., et al. (2020). Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. *Proc. Natl. Acad. Sci. U S A* 117, 28201–28211. doi: 10.1073/pnas.2002660117

Jones, P. A., Issa, J. P., and Baylin, S. (2016). Targeting the cancer epigenome for therapy. *Nat. Rev. Genet.* 17, 630–641. doi: 10.1038/nrg.2016.93

Kamburov, A., Lawrence, M. S., Polak, P., Leshchiner, I., Lage, K., Golub, T. R., et al. (2015). Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc. Natl. Acad. Sci. U S A* 112, E5486–E5495. doi: 10.1073/pnas.1516373112

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. doi: 10.1038/s41586-020-2308-7

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., et al. (2009). Human Protein Reference Database–2009 update. *Nucleic Acids Res.* 37, D767–D772. doi: 10.1093/nar/gkn892

Krassowski, M., Paczkowska, M., Cullion, K., Huang, T., Dzneladze, I., Ouellette, B. F. F., et al. (2017). ActiveDriverDB: human disease mutations and genome variation in post-translational modification sites of proteins. *Nucleic Acids Res.* 46, D901-D910. doi: 10.1101/178392

Landrum, M. J., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., et al. (2020). ClinVar: improvements to accessing data. *Nucleic Acids Res.* 48, D835–D844. doi: 10.1093/nar/gkz972

Lawrence, J. B., Oxvig, C., Overgaard, M. T., Sottrup-Jensen, L., Gleich, G. J., Hays, L. G., et al. (1999). The insulin-like growth factor (IGF)-dependent IGF binding protein-4 protease secreted by human fibroblasts is pregnancy-associated plasma protein-A. *Proc. Natl. Acad. Sci. U S A* 96, 3149–3153. doi: 10.1073/pnas.96.6.3149

Li, F., Fan, C., Marquez-Lago, T. T., Leier, A., Revote, J., Jia, C., et al. (2020). PRISMOID: a comprehensive 3D structure database for post-translational modifications and mutations with functional impact. *Brief Bioinform.* 21, 1069–1079. doi: 10.1093/bib/bbz050

Li, L., Shi, L., Yang, S., Yan, R., Zhang, D., Yang, J., et al. (2016). SIRT7 is a histone desuccinylase that functionally links to chromatin compaction and genome stability. *Nat. Commun.* 7:12235. doi: 10.1038/ncomms12235

Li, S., Iakoucheva, L. M., Mooney, S. D., and Radivojac, P. (2010). Loss of post-translational modification sites in disease. *Pac. Symp. Biocomput.* 2010, 337–347. doi: 10.1142/9789814295291_0036

Lin, Z. F., Xu, H. B., Wang, J. Y., Lin, Q., Ruan, Z., Liu, F. B., et al. (2013). SIRT5 desuccinylates and activates SOD1 to eliminate ROS. *Biochem. Biophys. Res. Commun.* 441, 191–195. doi: 10.1016/j.bbrc.2013.10.033

Liu, J., Qian, C., and Cao, X. (2016). Post-Translational Modification Control of Innate Immunity. *Immunity* 45, 15–30. doi: 10.1016/j.immuni.2016.06.020

Liu, T., Qian, W. J., Gritsenko, M. A., Camp, D. G. II, Monroe, M. E., Moore, R. J., et al. (2005). Human plasma N-glycoproteome analysis by immunoaffinity subtraction, hydrazide chemistry, and mass spectrometry. *J. Proteome Res.* 4, 2070–2080. doi: 10.1021/pr0502065

MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature* 508, 469–476. doi: 10.1038/nature13127

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

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., et al. (2015). Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* 348, 880–886. doi: 10.1126/science.aaa6806

Martínez-Jiménez, F. M. F., López-Arribillaga, E., Lopez-Bigas, N., and Gonzalez-Perez, A. (2020). Systematic analysis of alterations in the ubiquitin proteolysis system reveals its contribution to driver mutations in cancer. *Nat. Cancer* 1, 122–135. doi: 10.1038/s43018-019-0001-2

Montecchi-Palazzi, L., Beavis, R., Binz, P. A., Chalkley, R. J., Cottrell, J., Creasy, D., et al. (2008). The PSI-MOD community standard for representation of protein modification data. *Nat. Biotechnol.* 26, 864–866. doi: 10.1038/nbt0808-864

Moremen, K. W., Tiemeyer, M., and Nairn, A. V. (2012). Vertebrate protein glycosylation: diversity, synthesis and function. *Nat. Rev. Mole. Cell Biol.* 13, 448–462. doi: 10.1038/nrm3383

Morin, P. J., Sparks, A. B., Korinek, V., Barker, N., Clevers, H., Vogelstein, B., et al. (1997). Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC. *Science* 275, 1787–1790. doi: 10.1126/science.275.5307.1787

Narayan, S., Bader, G. D., and Reimand, J. (2016). Frequent mutations in acetylation and ubiquitination sites suggest novel driver mechanisms of cancer. *Genome Med.* 8:55. doi: 10.1186/s13073-016-0311-2

Paczkowska, M., Barenboim, J., Sintupisut, N., Fox, N. S., Zhu, H., Abd-Rabbo, D., et al. (2020). Integrative pathway enrichment analysis of multivariate omics data. *Nat. Commun.* 11:735. doi: 10.1038/s41467-019-13983-9

Pawson, T. (1995). Protein modules and signalling networks. *Nature* 373, 573–580. doi: 10.1038/373573a0

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 20, 110–121. doi: 10.1101/gr.097857.109

Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J., and Godzik, A. (2015). A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. *PLoS Comput. Biol.* 11:e1004518. doi: 10.1371/journal.pcbi.1004518

Radivojac, P., Baenziger, P. H., Kann, M. G., Mort, M. E., Hahn, M. W., and Mooney, S. D. (2008). Gain and loss of phosphorylation sites in human cancer. *Bioinformatics* 24, i241–i247. doi: 10.1093/bioinformatics/btn267

Reily, C., Stewart, T. J., Renfrow, M. B., and Novak, J. (2019). Glycosylation in health and disease. *Nat. Rev. Nephrol.* 15, 346–366. doi: 10.1038/s41581-019-0129-4

Reimand, J., and Bader, G. D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mole. Syst. Biol.* 9:637. doi: 10.1038/msb.2012.68

Reimand, J., Wagih, O., and Bader, G. D. (2013). The mutational landscape of phosphorylation signaling in cancer. *Sci. Rep.* 3, 2651. doi: 10.1038/srep02651

Reimand, J., Wagih, O., and Bader, G. D. (2015). Evolutionary constraint and disease associations of post-translational modification sites in human genomes. *PLoS Genet.* 11:e1004919. doi: 10.1371/journal.pgen.1004919

Reyna, M. A., Haan, D., Paczkowska, M., Verbeke, L. P. C., Vazquez, M., Kahraman, A., et al. (2020). Pathway and network analysis of more than 2,500 whole cancer genomes. *Nat. Commun.* 11:729. doi: 10.1038/s41467-020-14367-0

Rosen, D. R., Siddique, T., Patterson, D., Figlewicz, D. A., Sapp, P., Hentati, A., et al. (1993). Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* 362, 59–62. doi: 10.1038/362059a0

Smestad, J., Erber, L., Chen, Y., and Maher, L. J. III (2018). Chromatin Succinylation Correlates with Active Gene Expression and Is Perturbed by Defective TCA Cycle Metabolism. *iScience* 2, 63–75. doi: 10.1016/j.isci.2018.03.012

Sreedhar, A., Wiese, E. K., and Hitosugi, T. (2020). Enzymatic and metabolic regulation of lysine succinylation. *Genes Dis.* 7, 166–171. doi: 10.1016/j.gendis.2019.09.011

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337, 64–69. doi: 10.1126/science.1219240

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Trefely, S., Lovell, C. D., Snyder, N. W., and Wellen, K. E. (2020). Compartmentalised acyl-CoA metabolism and roles in chromatin regulation. *Mol. Metab.* 38:100941. doi: 10.1016/j.molmet.2020.01.005

UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049

Wadi, L., Meyer, M., Weiser, J., Stein, L. D., and Reimand, J. (2016). Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods* 13, 705–706. doi: 10.1038/nmeth.3963

Wagih, O., Reimand, J., and Bader, G. D. (2015). MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. *Nat. Methods* 12, 531–533. doi: 10.1038/nmeth.3396

Wang, Y., Cheng, H., Pan, Z., Ren, J., Liu, Z., and Xue, Y. (2015). Reconfiguring phosphorylation signaling by genetic polymorphisms affects cancer susceptibility. *J. Mol. Cell Biol.* 7, 187–202. doi: 10.1093/jmcb/mjv013

Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., and Jones, D. T. (2004). The DISOPRED server for the prediction of protein disorder. *Bioinformatics* 20, 2138–2139. doi: 10.1093/bioinformatics/bth195

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi: 10.1093/nar/gkx1037

Wojcechowskyj, J. A., Didigu, C. A., Lee, J. Y., Parrish, N. F., Sinha, R., Hahn, B. H., et al. (2013). Quantitative phosphoproteomics reveals extensive cellular reprogramming during HIV-1 entry. *Cell Host Microb.* 13, 613–623. doi: 10.1016/j.chom.2013.04.011

Wollscheid, B., Bausch-Fluck, D., Henderson, C., O'Brien, R., Bibel, M., Schiess, R., et al. (2009). Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nat. Biotechnol.* 27, 378–386. doi: 10.1038/nbt.1532

Xie, Z., Dai, J., Dai, L., Tan, M., Cheng, Z., Wu, Y., et al. (2012). Lysine succinylation and lysine malonylation in histones. *Mol. Cell Proteomics* 11, 100–107. doi: 10.1074/mcp.M111.015875

Yang, Y., He, Y., Wang, X., Liang, Z., He, G., Zhang, P., et al. (2017). Protein SUMOylation modification and its associations with disease. *Open Biol.* 2017:7. doi: 10.1098/rsob.170167

Yang, Y., Peng, X., Ying, P., Tian, J., Li, J., Ke, J., et al. (2019). AWESOME: a database of SNPs that affect protein post-translational modifications. *Nucleic Acids Res.* 47, D874–D880. doi: 10.1093/nar/gky821

Zhang, J., Ali, A. M., Lieu, Y. K., Liu, Z., Gao, J., Rabadan, R., et al. (2019). Disease-Causing Mutations in SF3B1 Alter Splicing by Disrupting Interaction with SUGP1. *Mol. Cell* 76:e87. doi: 10.1016/j.molcel.2019.07.017

Zhang, Z., Tan, M., Xie, Z., Dai, L., Chen, Y., and Zhao, Y. (2011). Identification of lysine succinylation as a new post-translational modification. *Nat. Chem. Biol.* 7, 58–63. doi: 10.1038/nchembio.495

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.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 24 June 2021
doi: 10.3389/fcell.2021.662983

DTL-DephosSite: Deep Transfer Learning Based Approach to Predict Dephosphorylation Sites

Meenal Chaudhari[1], Niraj Thapa[1], Hamid Ismail[1], Sandhya Chopade[1], Doina Caragea[2], Maja Köhn[3], Robert H. Newman[4] and Dukka B. KC[5]**

[1] Department of Computational Data Science and Engineering, North Carolina A&T State University, Greensboro, NC, United States, [2] Department of Computer Science, Kansas State University, Manhattan, KS, United States, [3] Faculty of Biology, Signalling Research Centres BIOSS and CIBSS, University of Freiburg, Freiburg, Germany, [4] Department of Biology, North Carolina A&T State University, Greensboro, NC, United States, [5] Department of Electrical Engineering and Computer Science, Wichita State University, Wichita, KS, United States

Phosphorylation, which is mediated by protein kinases and opposed by protein phosphatases, is an important post-translational modification that regulates many cellular processes, including cellular metabolism, cell migration, and cell division. Due to its essential role in cellular physiology, a great deal of attention has been devoted to identifying sites of phosphorylation on cellular proteins and understanding how modification of these sites affects their cellular functions. This has led to the development of several computational methods designed to predict sites of phosphorylation based on a protein's primary amino acid sequence. In contrast, much less attention has been paid to dephosphorylation and its role in regulating the phosphorylation status of proteins inside cells. Indeed, to date, dephosphorylation site prediction tools have been restricted to a few tyrosine phosphatases. To fill this knowledge gap, we have employed a transfer learning strategy to develop a deep learning-based model to predict sites that are likely to be dephosphorylated. Based on independent test results, our model, which we termed DTL-DephosSite, achieved efficiency scores for phosphoserine/phosphothreonine residues of 84%, 84% and 0.68 with respect to sensitivity (SN), specificity (SP) and Matthew's correlation coefficient (MCC). Similarly, DTL-DephosSite exhibited efficiency scores of 75%, 88% and 0.64 for phosphotyrosine residues with respect to SN, SP, and MCC.

Keywords: post-translational modification, deep learning, transfer learning, dephosphorylation, computational prediction

OPEN ACCESS

Edited by:
Dong Xu,
University of Missouri, United States

Reviewed by:
Peng (Sam) Sun,
Bayer Crop Science, United States
Duolin Wang,
University of Missouri, United States

****Correspondence:***
Robert H. Newman
rhnewman@ncat.edu
Dukka B. KC
dukka.kc@wichita.edu

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

Received: *02 February 2021*
Accepted: *20 May 2021*
Published: *24 June 2021*

Citation:
Chaudhari M, Thapa N, Ismail H, Chopade S, Caragea D, Köhn M, Newman RH and KC DB (2021) DTL-DephosSite: Deep Transfer Learning Based Approach to Predict Dephosphorylation Sites. Front. Cell Dev. Biol. 9:662983. doi: 10.3389/fcell.2021.662983

INTRODUCTION

Protein phosphorylation is an important posttranslational modification (PTM) that regulates many cellular activities and contributes to the etiology and progression of several pervasive diseases, including cancer, diabetes, cardiovascular disease, and neurodegeneration. In eukaryotic cells, phosphorylation, and subsequent dephosphorylation, occurs on serine (S), threonine (T), and tyrosine (Y) residues located on the protein surface. To date, more than two-thirds of the ∼21,000 proteins encoded by the human genome have been shown to be phosphorylated,

making phosphorylation one of the most wide-spread and broadly studied protein PTMs (Ardito et al., 2017). The precise regulation of the phosphorylation status of a protein depends on the opposing activities of protein kinases, which catalyze the transfer of the γ-phosphate of ATP to their downstream substrates, and protein phosphatases, which catalyze the dephosphorylation (i.e., removal of the phosphate group) from the modified site (**Figure 1**). While it is often assumed that any site that can be phosphorylated can also be dephosphorylated, this may not always be the case (Bechtel et al., 1977; Bornancin and Parker, 1997; Keshwani et al., 2012; Senga et al., 2015). Similarly, certain sites may be dephosphorylated more efficiently than others. Though rare, there are instances of phosphorylation sites that are resistant to dephosphorylation. For instance, once phosphorylated, both T197 and S338 in cAMP-dependent protein kinase (PKA) are resistant to dephosphorylation (Bechtel et al., 1977; Keshwani et al., 2012). Similarly, protein kinase G (PKG), protein kinase C (PKC), and calcium/calmodulin-dependent protein kinase 1δ (CAMK1δ) each exhibit phosphatase-resistant states (Bornancin and Parker, 1997; Keshwani et al., 2012; Senga et al., 2015). The relative efficiency of dephosphorylation at a particular site may be, at least partially, dependent on the local protein environment and the ability of phosphatases to recognize the phosphosite.

Phosphorylation site prediction has recently emerged as an important problem in the field of bioinformatics. As a result, many phosphorylation site prediction tools have been developed to predict both general and kinase-specific phosphorylation sites (Lumbanraja et al., 2019; Luo et al., 2019; Haixia et al., 2020; Wang D. et al., 2020; Ahmed et al., 2021; Guo et al., 2021). For instance, to predict general phosphorylation sites based on the primary amino acid sequence of an input protein, Ismail et al. developed the Random Forest (RF)-based phosphosite predictor 2.0 (RF-Phos 2.0) (Ismail et al., 2016). RF-Phos 2.0 assesses the relative importance of hand-selected features to identify putative sites of phosphorylation across many protein families. More recently, Luo et al. developed Deep-Phos, a general and kinase-specific phosphorylation site predictor based on multilayer convolutional neural networks (CNN) (Luo et al., 2019).

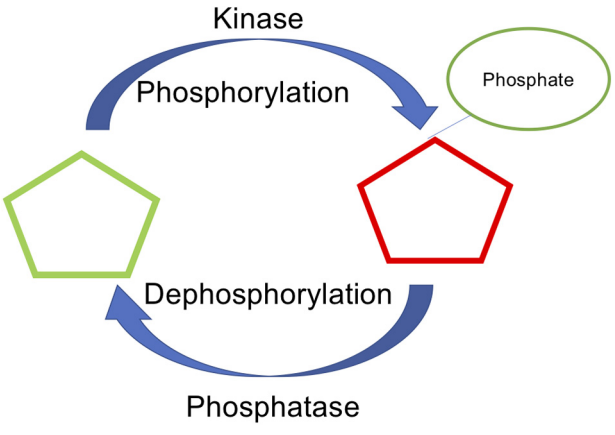


FIGURE 1 | Phosphorylation and dephosphorylation, mediated by kinase and phosphatase as a key reversible post translational modification.

While many phosphorylation site prediction tools have been developed over the past decade to identify putative sites of S, T, and Y phosphorylation (Ismail et al., 2016; Luo et al., 2019; Wang D. et al., 2020), computational prediction of dephosphorylation sites has been much more limited (Wang et al., 2016). Information about dephosphorylation sites is important because it can provide insights into the molecular determinants of phosphatase recognition and may offer clues about the biological half-life of a given phosphorylation event. To date, computational methods for dephosphorylation site prediction have focused on a relatively small group of tyrosine phosphatases consisting of protein tyrosine phosphatase 1B (PTP1B) and the Src homology 2 (SH2) domain-containing phosphatases, SHP-1 and SHP-2 (Wu et al., 2014; Wang et al., 2016; Jia et al., 2017). For instance, Wu et al. developed a method that uses the k-nearest neighbor algorithm to identity the substrate sites of PTP1B, SHP-1, and SHP-2 based on the sequence features of manually collected dephosphorylation sites (Wu et al., 2014). Meanwhile, Wang et al. developed two sophisticated models for predicting the substrate dephosphorylation sites of these phosphatases. The first model, which they termed MGPS-DEPHOS, is modified from the Group-based Prediction System (GPS) while the second model, termed CKSAAP-DEPHOS, utilizes a combination of support vector machine (SVM) and the k-spaced amino acid pairs (CKSAAP) encoding scheme. Finally, Jia et al. (2017) combined the sequence-based bi-profile Bayes feature extraction technique and SVM to predict sites for the same three phosphatases.

One of the primary reasons for the proliferation of phosphorylation site predictors over the past decade is the availability of large databases cataloging experimentally identified phosphorylation sites, such as PhosphoSitePlus and PhosphoELM (Dinkel et al., 2011; Hornbeck et al., 2019). Unfortunately, similar databases have not been available for dephosphorylation sites. However, with the recent curation of the DEPOD database of S, T, and Y dephosphorylation sites, the development of dephosphorylation site predictors is now feasible (Damle and Köhn, 2019). In this study, we compiled a dataset of S, T, and Y dephosphorylation sites from the DEPOD database (Damle and Köhn, 2019) and further extended the available dataset through literature mining, increasing the database more than threefold. We then developed a transfer learning approach utilizing the phosphorylation dataset and a bidirectional long short-term memory (Bi-LSTM) deep learning-based model to predict dephosphorylation sites on proteins. To our knowledge, this is the first study to develop a general dephosphorylation predictor for Y residues and the first to predict general dephosphorylation sites for S/T residues. Our models, which we termed DTL-DephosSite-ST and DTL-DephosSite-Y, performed well when assessed using both five-fold cross-validation and an independent test set.

Here we have developed the first general phosphatase site prediction tool. Unlike phosphatase-specific methods, which are designed to predict both the site of dephosphorylation and

the phosphatase mediating the dephosphorylation event, our general dephosphorylation site prediction method is able to identify putative sites of dephosphorylation irrespective of the phosphatase mediating the dephosphorylation event. This is analogous to the results obtained by MS/MS-based experiments, where information about the responsible phosphatase is not known. Importantly, phosphatase-specific methods are currently restricted to predictions for only three phosphatases (i.e., PTP1B, SHP1, and SHP2), which represent a very small fraction of phosphatases encoded by the human genome. This is likely due, in part, to limited information about the specific phosphatase that mediates a given dephosphorylation event. Therefore, general dephosphorylation site prediction methods offer distinct advantages when the primary goal is to predict whether or not a given site is dephosphorylated.

MATERIALS AND METHODS

Datasets

The human DEPhOsphorylation Database, DEPOD, is a database of dephosphorylation sites that was recently expanded in an updated version in 2019 (Damle and Köhn, 2019). DEPOD accounts for 241 active and 13 inactive human phosphatases in total. Among the active phosphatases, 194 include substrate data. This database provided the starting point to create dephosphorylation datasets for S, T, and Y residues. To this end, we collected all the FASTA sequences from the UniProt database (UniProt Consortium, 2019) and extracted windows with the targeted S/T/Y residue at the center and 16 residues on each side. Negative sequences were extracted using all S/T/Y residues except those that are known positive sites (i.e., all residues except those sites that are known to be dephosphorylated). During the generation of sequences, no fillers (i.e., "-") were used. To minimize the loss of sequences occurring at the ends, a maximum window size of 33 was chosen. Any redundant sequences within and between the positive and negative sites were removed to obtain a non-redundant set. Similar to our previous studies (Chaudhari et al., 2020; Thapa et al., 2020), we used an under-sampling strategy to balance the dataset, which had more negative sites than positive sites prior to balancing (Aridas GLitaFNaCK, 2017). Under-sampling allows random selection of negative sequences to make the number of negative sites equal to the number of positive sequences, thus balancing the dataset.

Once constructed, the dataset was further divided into training and test sets, such that 80% of the data was used to train the models and the remaining 20% of the data was kept aside for independent testing. This training-test dataset, which we termed the DEPOD-19 dataset (**Table 1**), consists of 133 positive sites for S, 58 positive sites for T, and 101 positive sites for Y (**Table 1**).

Though phosphorylation is one of the most wide-spread and well-studied PTMs in eukaryotes, comprehensive lists of dephosphorylation sites are scarce. This is likely due to the lack of computational studies in the field and technical challenges associated with the detection of dephosphorylation sites. Therefore, in order to enlarge the dephosphorylation site dataset (Damle and Köhn, 2019), we did a comprehensive literature review to identify phosphorylated sites that were down-regulated in cells following treatment with various agents. For a given site to be considered dephosphorylated, there must have been no co-stimulation during treatment and the analysis must have been conducted less than an hour after stimulation (to prevent changes in protein expression from substantially contributing to the observed changes in phosphorylation state). Moreover, because many phosphorylation sites have been identified in human cells, we only considered publications using human cells. Finally, to avoid errors stemming from heterogeneity in the phosphorylation patterns in different phases of the cell cycle, our analyses only included cells that had been arrested in the mitotic phase. Using these criteria, we developed the "Downreg" dataset, which consists of 949 dephosphorylation sites in 624 proteins. These included 772 S, 152 T, and 25 Y residues, which represents an ∼3.25-fold increase relative to the DEPOD-19 dataset, as summarized in **Table 1** and **Supplementary Table 2**. A summary of the data sources and the corresponding descriptive statistics for each study (e.g., false discovery rate and data distribution) are included in **Supplementary Table 1** and all the newly added dephosphorylation sites from the "Downreg" dataset have been added in **Supplementary Table 12**.

During sequence extraction, a sub-sequence with window size of 33 centered around the site of interest was created in a manner similar to that described for the DEPOD-19 dataset above. **Supplementary Table 1** summarizes the literature sources and the number of dephosphorylation sites identified. Removal of common sequences within and between the positive and negative sets was performed to obtain a non-redundant dataset. Finally, the "combined dephosphorylation site" dataset was obtained by merging the DEPOD-19 and Downreg datasets and removing any duplicate protein sequences of window size 33. The combined dephosphorylation dataset (ComDephos) is summarized in **Table 1**. For model development, the DEPOD-19 and the ComDephos datasets were used.

Bidirectional LSTM Model

Long Short-Term Memory (LSTM) models are known to provide good performance with sequence data (Hochreiter and Schmidhuber, 1997). LSTM uses different memory cells and an additive gradient function helps to overcome the vanishing and exploding gradient problems in recurrent neural networks (RNN). Importantly, the use of memory cells can keep sequence information in the network for long periods of time.

TABLE 1 | Summary of the training and test datasets used for model development based on sites extracted from the DEPOD-19, Downreg (literature resources) and composite ComDephos datasets.

Dataset	Residue	Train	Test	Total positive	Total negative
DEPOD-19	ST	304	78	191	191
	Y	161	41	101	101
Downreg	ST	1478	370	924	924
	Y	40	10	25	25
ComDephos	ST	1,806	446	1,112	1,112
	Y	201	50	125	125

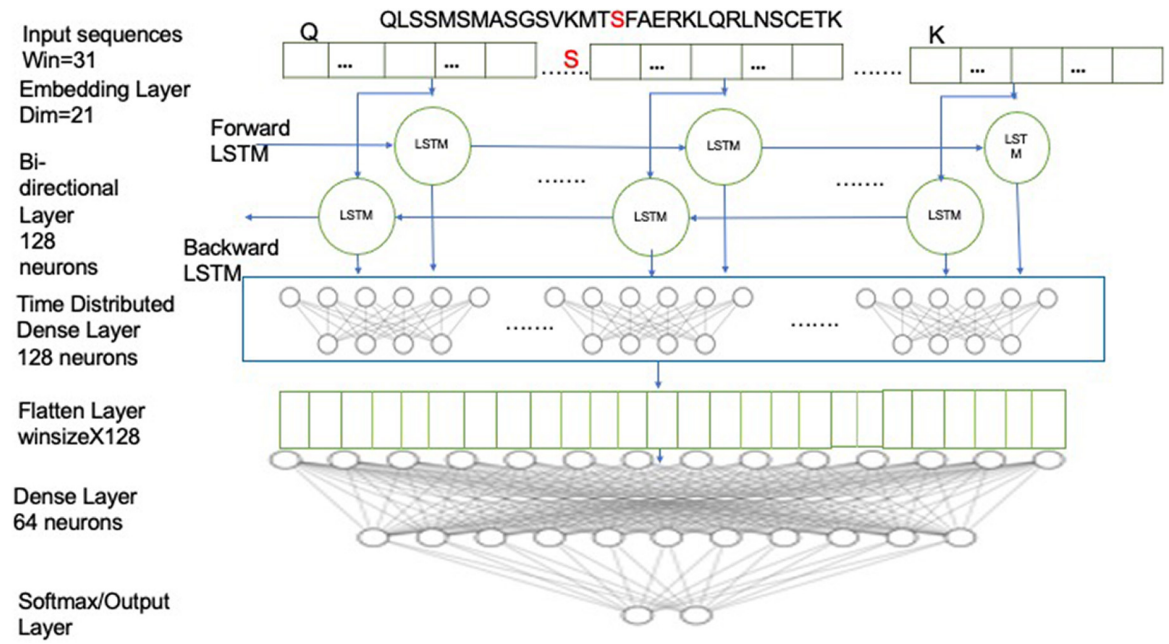


FIGURE 2 | Schematic illustrating the Bi-LSTM deep learning architecture and the parameters used. The input sequence is first fed into embedding layer with dimension of 21, then through two Bi-LSTM layers with 128 neurons and then followed by a time-distributed layer of 128 neurons, which was followed by a flatten layer and then followed by dense layer with 2 neurons with softmax activation.

A single LSTM cell consists of three gates: "input," "forget," and "output" gates (**Figure 2**). The input layer (z_t) consists of a sigmoid layer and a tanh layer. The sigmoid layer filters the previous state to select the relevant cell states for the context while the tanh layer provides a range of values to take to the selected states. The forget layer (r_t) consists of a sigmoid layer, which filters the irrelevant previous cell states by dropping them out. The output layer ($h\flat_t$) employs a tanh layer to provide an update to the selected states, as provided by the input layer (Hochreiter and Schmidhuber, 1997).

The forget gate layer takes previous hidden cells and inputs for each previous cell state. The sigmoid node in the forget gate adds in 0 or 1 to the previous hidden state, deciding whether it would be passed over to the next hidden state. The input gate layer has sigmoid and tanh nodes, where the sigmoid acts as a selection node and selects the values that need to be updated. Meanwhile, the tanh nodes provide a vector of new candidate values for the selected states, acting as the update node. Finally, the output is obtained by adding previous values for old states and updated values for the selective nodes.

In this architecture, we have employed a bidirectional LSTM layer (Bi-LSTM), which uses twice the number of neurons as a conventional LSTM layer. The double neurons create two sets of networks, moving in both the forward and the reverse directions (Schuster and Paliwal, 1997). Thus, a Bi-LSTM layer is able to predict the context of the target residue from the residues from both directions. For example, given a window sequence:

NYTPTSPNYSPTSPSYSPTSPSYSPTSPSYSPS

where the S (red) in the center represents the target residue, the forward LSTM network would predict the probability of having S, given the knowledge of the residues preceding it (i.e., "NYTPTSPNYSPTSPSY") while the backward/reverse LSTM network would predict the probability of having S, given the knowledge of residues following it (i.e., "PTSPSYSPTSPSYSPS"). The window sequences were integer encoded, such that each character in the sequence was replaced by its corresponding integer value. The integer encoded sequences were then fed to the embedding layer, which provides an embedding dimension of 21, which is known to be optimal based on our previous studies (Chaudhari et al., 2020; Thapa et al., 2020). The embedding layer helps in capturing the latent representation of the encodings using a look-up table (Keras, 2015). For model development, a Bi-LSTM layer with 128 neurons was used, with timesteps equivalent to the window size, and return sequences kept as "true." Next, it was followed with a time-distributed layer of 128 neurons. The time-distributed layer applies dense layer operation to every timestep of the 3D tensor (Keras, 2015). This was followed by a flatten layer with a dropout of 0.4 to avoid overfitting and a dense layer of 64 neurons, which was then followed by the output dense layer with 2 neurons with softmax activation. The model was compiled on binary cross-entropy loss using the Adam optimizer (Kingma and Ba, 2014). We used two callbacks while fitting the model: ModelCheckpoint and reduce learning rate on Plateau. ModelCheckpoint obtains the best model with respect to validation accuracy while the reducing learning rate helps in learning the parameters better, especially when the data size is small (Li and Hoiem, 2018). Parameters have been optimized to the settings shown in **Table 2**.

TABLE 2 | Parameters used in LSTM Model for dephosphorylation.

Parameters	Settings
Embedding output dimension	21
Learning rate	0.01
Batch size	512
Epochs	30
LSTM_layer1_neurons	128
Dropout	0.4
Dense_layer_neurons	128, 64, 2

Transfer Learning

As molecular counterparts, phosphorylation and dephosphorylation are closely related to one another but the cellular enzymes catalyzing each event (as well as the molecular determinants underlying recognition of the sites) are different. Moreover, the extensive study of phosphorylation sites has resulted in a comparatively large dataset of phosphosites, while the amount of information about dephosphorylation events has led to a relatively sparse dataset. Taken together, these observations suggest that a transfer learning strategy could be applied to dephosphorylation site prediction.

Recently, deep learning has been used to solve various problems in bioinformatics (Li et al., 2019; Tang et al., 2019; Chaudhari et al., 2020; Thapa et al., 2020; Wang D. et al., 2020; Wang Y. et al., 2020). One of the most serious problems associated with deep learning stems from data dependence. For instance, a significant challenge is posed by the lack of labeled data for the task-of-interest, e.g., dephosphorylation. Indeed, the problem of insufficient training data is an inescapable problem in various areas of bioinformatics. For dephosphorylation, the expense of data acquisition makes it particularly difficult to construct a large-scale, well-annotated dataset.

Previous studies suggest that, when trained on images, deep learning networks tend to learn first-layer features that do not appear to be specific to a particular task (Yosinski et al., 2014). Such first layer features are general in that they are applicable to many datasets and tasks. Exploiting this fact, transfer learning relaxes the hypothesis that the training data and test data are not required to be "independently and identically distributed" and that the model in the target domain does not need to be trained from scratch, which can significantly reduce the burden of training data size (Tan et al., 2018). Transfer of knowledge through shared parameters and weights of the source model and the target domain is one of the strategies in transfer learning (Weiss et al., 2016).

With the exception of a handful of dual specificity kinases and phosphatases, most kinases and phosphatases recognize either S/T or Y residues. Therefore, as is common in phosphorylation site prediction, we considered two models: one for S and T residues and another for Y residues. Thus, distinct phosphorylation and dephosphorylation datasets were formed and designated the Phos-ST and Phos-Y datasets and the Dephos-ST and Dephos-Y datasets.

During transfer learning, three important questions need to be answered: (a) what to transfer, (b) when to transfer, and (c) how to transfer. Therefore, to allow our framework to

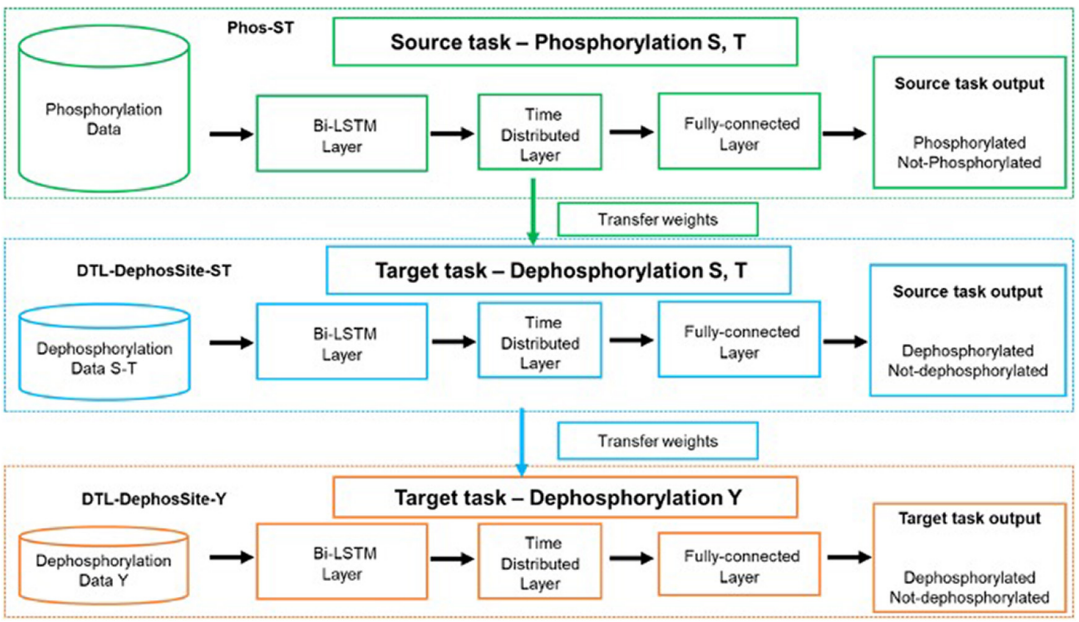


FIGURE 3 | Schematic illustrating the transfer-learning. Green dotted box depicts the training on source task, phosphorylation (S,T), to obtain the Phos-ST model. Once the Phos-Model was obtained the Bi-LSTM model was instantiated with the Phos-Model weights before being trained on the dephosphorylation data. Blue dotted box depicts the transfer learning on the target task, dephosphorylation for ST residues, to obtain the DTL-DephosSite-ST model. During transfer learning, all layers were allowed to re-train and none of the layers were frozen. (We tried various options with various layers frozen but this version produced the best results). Orange dotted box depicts the transfer learning from DTL-DephosSite-ST, to obtain the DTL-DephosSite-Y model.

accommodate smaller datasets, we applied a two-step transfer learning scheme that included a pre-training step and a fine-tuning step (**Figure 3**). The pre-training step results in a source model, which is then available to adapt on the target dataset through fine-tuning.

The pre-training step involves the training of our Bi-LSTM model (as described in section "Bidirectional LSTM Model") on the available phosphorylation data (Wang D. et al., 2020), which are provided in **Supplementary Table 3**. This resulted in a Phos-model that contains learned weights to classify a given motif as phosphorylated or not, specifically the S/T residues. During the fine-tuning step, the weights learned by the source Phos-model were transferred to a new instance of the Bi-LSTM architecture. The model was then trained on the Dephos data containing the S/T residues in the center, thus obtaining a transfer-learned Dephos model for S/T residues. We experimented with different combinations of frozen and re-trained layers and identified a model, where all layers are allowed to re-train, that learned better than others.

Similarly, for the prediction of Y dephosphorylation sites, we experimented with performing transfer learning from Phos-ST-to-Dephos-Y as well as Phos-Y-to-Dephos-Y. These studies suggested that the Dephos-ST-to-Dephos-Y transfer worked the best. Thus, the pre-training step involved training the Dephos-ST model, initialized with transfer-learned weights from Phos-ST on the Dephos-ST dataset. During the fine-tuning step, we retrained all layers on the Dephos-Y dataset. Though varying the layers that were kept frozen or re-trained had less impact in performance, retraining all layers helped in attaining more consistent results.

Finally, we also employed the transfer learned Dephos-Y model on the available phosphatase specific datasets (Wang et al., 2016) for PTP1B, SHP1, and SHP2 (**Supplementary Table 9**).

Performance and Evaluation

To evaluate the performance of each model, we used a confusion matrix to determine Sensitivity (SN), Specificity (SP), Accuracy (ACC) and the Receiver Operating Characteristic (ROC) curve as the performance metrics. The models were evaluated using five-fold cross-validation on the benchmark training dataset and an independent test set.

ACC describes the correctly predicted residues out of the total residues (Eq. 1). Meanwhile, SN defines the model's ability to distinguish positive residues (Eq. 2) and SP measures the model's ability to correctly identify the negative residues (Eq. 3). Matthews Correlation Coefficient (MCC) is the calculated score that takes into account the model's predictive capability with respect to both positive and negative residues (Eq. 4). Likewise, the ROC curve provides a graphical representation of the diagnostic ability of the classifier. The area under the ROC curve (AUC) is used to compare various models, with the models having the highest AUC scores generally performing better in classification than those with lower AUC scores.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (1)$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \quad (2)$$

$$Specificity = \frac{TN}{TN + FP} \times 100 \quad (3)$$

$$MCC = \frac{(TP)\,(TN) - (FP)(FN)}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}} \quad (4)$$

RESULTS AND DISCUSSION

Bidirectional Model on Dephos Datasets (Without Transfer Learning)

To efficiently identify sites that are likely to be dephosphorylated in proteins, we sought to develop a dephosphorylation site prediction tool using the recently expanded DEPOD-19 dataset (**Table 1**). To this end, we first extracted FASTA sequences from the DEPOD-19 dataset. During extraction, we limited the window size to 33 in order to minimize the loss of sequences at the ends of the sequences. We then applied a bidirectional long short-term memory (Bi-LSTM) deep learning strategy to the dataset. During these analyses, we trained on the train dataset and the performance of the resulting model was evaluated using an independent test set (representing 20% of the original dataset) that was kept aside from the training set. These analyses suggest that our preliminary model had reasonable sensitivity (SN) and receiver operating characteristic (ROC) scores of 0.85 and 0.79, respectively. However, this preliminary model suffered with respect to specificity (SP) and Matthew's correlation coefficient (MCC), where it exhibited scores of 0.49 and 0.36, respectively (**Table 3**). A feature-based machine learning strategy employing random forest (RF) yielded similar results (**Supplementary Table 4**).

Though the DEPOD-19 dataset has recently been expanded to include 584 total sites, it still represents a relatively small dataset for model development using machine learning strategies. Therefore, to further expand the dataset, we conducted a comprehensive literature search for dephosphorylation sites. This yielded an additional 1,898 sites whose phosphorylation status decreased within an hour of treatment in mitotically arrested cells (**Table 1**; see section "Materials and Methods" for details). Combining this so called "Downreg" dataset with those sites that had already been curated in the DEPOD-19 dataset resulted in a composite "ComDephos" dataset containing 2,503 total, non-redundant dephosphorylation sites (composed of 1,806 S, 446 T, and 251 Y sites) (**Table 1**). We then repeated our Bi-LSTM-based learning scheme using the newly developed ComDephos dataset and assessed performance based on our independent test set (**Table 3** and **Supplementary Table 5**). This led to marginal

TABLE 3 | Performance of Deep learning model on Depod19 and ComDephos datasets.

Dataset	MCC	Specificity	Sensitivity	ROC_AUC
Depod19	0.36	0.49	0.85	0.79
ComDephos	0.46	0.71	0.76	0.81

Independent test results using the DEPOD-19 and the ComDephos datasets for ST residue.

improvements in model performance using the independent datasets. For instance, while ROC increased marginally (2.5%), SP increased by 44.8% and SN decreased by 10.5%. Together, these changes resulted in a 27.8% increase in overall model performance, as assessed by MCC.

The observed gains are likely due to an increase in the size of the dataset, consistent with several reports that suggest that deep learning models perform well on large datasets and that an increase in the size of the dataset can increase the performance of the resulting model (Zhao, 2017; Feng et al., 2019). However, despite these gains, performance of the model developed using the ComDephos dataset was still relatively poor. Therefore, we asked if model performance could be enhanced using a transfer learning strategy.

Development of S/T Dephosphorylation Site Predictor Using Transfer Learning on the Phosphorylation Site Database

In contrast to dephosphorylation sites, phosphorylation sites have been extensively annotated, totaling 484,110 sites in 20,217 proteins (PhosphoSitePlus; Hornbeck et al., 2019), as 1/31/2021). Given the inherent similarities in the physiochemical properties of the modified sites and the potential differences in the molecular determinants used by kinases and phosphatases to recognize sites of phosphorylation and dephosphorylation, respectively, we reasoned that a transfer learning approach could be applied to develop a model to predict sites of dephosphorylation (**Figure 3**). Therefore, we used the phosphorylation dataset described by Wang et al. (2017). This dataset, which is composed of 31,944 experimentally determined phosphorylation sites and an equal number of negative sites (i.e., S, T, or Y residues that are not known to be phosphorylated), was used to generate a source model (**Supplementary Table 3**). First, we explored the effect that window size had on phosphosite prediction. To this end, progressively smaller window sizes were created, starting with a window size of 33. This was achieved by removing one residue from each end of the sequence in successive steps to yield windows of 33, 31, 29, 27, 25, and 23. We then trained the Bi-LSTM model on the phosphorylation training dataset using each window size and tested on the independent test set (**Supplementary Table 6**). This led to our source phosphorylation model (Phos-Model) for their respective windows, which was used for transfer learning to the target dephosphorylation dataset.

Next, to apply the knowledge gained from phosphorylation site prediction to dephosphorylation, the Bi-LSTM model was instantiated with the Phos-Model weights before being trained on the DEPOD-19 and ComDephos datasets. During transfer learning, all layers were allowed to re-train in the fine-tuning step. This yielded a transfer-learned dephosphorylation model for each window size. To determine the optimal window size, we then conducted five-fold cross-validation of the transfer-learned dephosphorylation dataset based on the ComDephos dataset (**Table 4**). These analyses suggested that window sizes of 29 and 31 led to the best predictors based on MCC. A similar trend was also observed for the phosphorylation dataset (**Supplementary Table 6**) and for a transfer-learned model trained on the DEPOD-19 dataset (**Supplementary Table 7**). Since a window size of 31 performed marginally better with respect to SN and ROC, we selected this window for further analysis. We termed this transfer learned, deep learning-based S/T dephosphorylation site predictor, DTL-DephosSite-ST. Importantly, compared to the S/T model developed using deep learning alone, DTL-DephosSite-ST exhibited an increase in all performance metrics. This resulted in an ∼3.26-fold increase in overall performance for S/T, as assessed by MCC. Likewise, using our independent dataset, DTL-DephosSite-ST outperformed similar transfer-learned dephosphorylation site prediction models that had been trained using either different deep learning architectures, such as conventional LSTM or CNN, or the recently developed DeepPhos (Luo et al., 2019) phosphorylation site predictor, which utilizes densely connected CNNs (**Table 5**). Taken together, these data suggest that DTL-DephosSite-ST effectively predicts putative sites of dephosphorylation on S/T residues.

TABLE 4 | Five-fold cross-validation of various window sizes for prediction of S/T residues following transfer learning using Phos-Model (source) and ComDephos dataset (target).

Window size	MCC ± SD	Specificity ± SD	Sensitivity ± SD	Accuracy ± SD	ROC_AUC
23	0.58 ± 0.05	0.78 ± 0.04	0.80 ± 0.01	0.79 ± 0.02	0.86
25	0.60 ± 0.04	0.78 ± 0.02	0.82 ± 0.03	0.80 ± 0.02	0.86
27	0.60 ± 0.05	0.79 ± 0.04	0.81 ± 0.02	0.80 ± 0.02	0.87
29	**0.61 ± 0.04**	**0.79 ± 0.02**	0.82 ± 0.03	0.80 ± 0.02	0.86
31	**0.61 ± 0.04**	0.77 ± 0.03	**0.83 ± 0.03**	0.80 ± 0.02	**0.87**
33	0.60 ± 0.05	0.78 ± 0.04	0.82 ± 0.03	0.80 ± 0.02	0.87

The highest scores in each metric are highlighted in boldface.

Transfer Learning Dephos-Y

With a transfer-learned S/T dephosphorylation site model in hand, we used a similar strategy to identify putative sites of Y dephosphorylation. Specifically, transfer learning was applied to the Y residues in the ComDephos dataset using DTL-DephosSite-ST as the source model. To obtain the DTL-DephosSite-Y, the model was instantiated with the weights of DTL-DephosSite-ST and all layers were re-trained on the ComDephos-Y dataset. Similar to the results for the S/T models,

TABLE 5 | Comparison between DTL-DephosSite-ST and transfer-learned models developed using other deep learning architectures based on an independent test set.

Architecture	MCC	Specificity	Sensitivity	ROC_AUC
CNN	0.60	0.74	**0.86**	0.89
LSTM	0.64	0.79	0.85	0.86
DeepPhos (DC-CNN): (Luo et al., 2019)	0.64	0.82	0.83	0.89
DTL-DephosSite-ST (Bi-LSTM)	**0.68**	**0.84**	0.84	**0.90**

CNN, Convolutional Neural Network; LSTM, Long short-term memory; DC-CNN, Densely connected CNN; Bi-LSTM, bidirectional LSTM. The highest scores in each metric are highlighted in boldface.

five-fold cross-validation suggested that window sizes of 27 and 31 performed the best, with a window size of 31 exhibiting slightly higher values for the majority of performance metrics (**Table 6**). Interestingly, models that were trained in the same manner using the smaller DEPOD-19 dataset resulted in a more sporadic distribution across windows, with a window size of 27 achieving the best specificity, and a window size of 31 producing the highest values for MCC and Sensitivity (**Supplementary Table 7**). Such a sporadic distribution may suggest that we are approaching a lower limit with respect to the size of the dataset, beyond which transfer learning becomes less effective.

Similarly, models that were trained using different combinations of source models and target datasets (e.g., Phospho-Y as source and ComDephos as target or Phospho-Y as source and DEPOD-19 as target) yielded models that performed well in most metrics, but not as well as the window size 31 Y dephoshorylation model generated using DTL-DephosSite-ST as the source model and the ComDephos dataset as the target dataset (**Supplementary Table 8**). For instance, window sizes of 27 and 31 exhibited similar MCC, with window size of 31 achieving the best specificity, accuracy and ROC scores. Therefore, we chose this model, which we named DTL-DephosSite-Y, for further analysis. Similar to DTL-DephosSite-ST, the newly developed DTL-DephosSite-Y performed well when evaluated using an independent test set (**Table 7**).

TABLE 6 | Five-fold cross-validation of various window sizes for prediction of Y residues following transfer learning using DTL-DephosSite-ST (source) and ComDephos dataset (target).

Window size	MCC ± SD	Specificity ± SD	Sensitivity ± SD	Accuracy ± SD	ROC_AUC
23	0.53 ± 0.09	0.76 ± 0.11	0.76 ± 0.07	0.76 ± 0.04	0.81
25	0.49 ± 0.13	0.76 ± 0.12	0.72 ± 0.09	0.74 ± 0.06	0.79
27	**0.59 ± 0.06**	0.78 ± 0.10	**0.80 ± 0.09**	0.79 ± 0.03	0.82
29	0.50 ± 0.07	0.74 ± 0.06	0.76 ± 0.06	0.75 ± 0.03	0.82
31	**0.59 ± 0.10**	**0.83 ± 0.09**	0.76 ± 0.06	**0.80 ± 0.05**	**0.83**
33	0.58 ± 0.08	0.78 ± 0.07	**0.80 ± 0.04**	0.79 ± 0.05	0.82

The highest scores for each metric are highlighted in boldface.

TABLE 7 | Independent test results of DeepPhos (Luo et al., 2019), DTL-DephosSite-ST and DTL-DephosSite-Y on ComDephos independent set, using the optimized parameters.

Predictor	MCC	Specificity	Sensitivity	Accuracy	ROC_AUC
DeepPhos	0.44	0.48	**0.92**	0.70	0.86
DTL-DephosSite-ST	**0.68**	0.84	0.84	**0.84**	**0.90**
DTL-DephosSite-Y	0.64	**0.88**	0.75	0.82	0.89

Here, results of DeepPhos model is provided to show the performance of a model trained on just Phosphorylation sites. The highest scores in each metric are highlighted in boldface.

CONCLUSION

Here, we describe a strategy that combines deep learning with transfer learning to develop general dephosphorylation site predictors of S/T and Y residues. To our knowledge, the resulting models, termed DTL-DephosSite-ST and DTL-DephosSite-Y, are the first general dephosphorylation site predictors for S/T and Y dephosphorylation, respectively. Deep learning-based models have recently been developed for several important PTMs, including phosphorylation, methylation, acetylation, and succinylation, to name a few (Wang et al., 2017; Luo et al., 2019; Wu et al., 2019; Al-barakati et al., 2020; Chaudhari et al., 2020; Thapa et al., 2020; Ahmed et al., 2021). Similar to previous deep learning-based models, our models did not require any hand selected features during model development. However, unlike many of the other deep learning-based models that were developed using extensive PTM data, the number of experimentally identified dephosphorylation sites was relatively low. As a consequence, our initial attempts to develop dephosphorylation site predictors based solely on deep learning yielded models that did not predict sites efficiently. This is consistent with reports that deep learning does not perform as well on small datasets (Zhao, 2017; Feng et al., 2019). To overcome this limitation, we developed a transfer learning-based approach. Specifically, we generated a source model based on knowledge gained about phosphorylation using a Bi-LSTM deep learning architecture and then applied this information to the ComDephos dataset using transfer learning. The resulting models performed markedly better than those developed using Bi-LSTM alone. This suggests that our approach is able to learn solely through the patterns of motif sequences. Importantly, by utilizing a transfer learning-based strategy, we were able to capitalize on the richness of phosphorylation site datasets in order to improve the efficacy of dephosphorylation prediction. This provides an attractive solution to the scarce data problem and may be applicable in the development of other PTM predictors.

During this project, we also expanded the DEPOD-19 dephosphorylation dataset 3.25-fold to create computational datasets of dephosphorylation. Importantly, this study relies upon the correlation between the cellular processes of phosphorylation and dephosphorylation. We have attempted to measure the level of transferability between phosphorylation and dephosphorylation. Similar correlations are also likely to be found for other PTMs where the forward and reverse reactions are catalyzed by different classes of enzymes, such as methylation/demethylation and acetylation/deacetylation. Prediction of sites of these modifications may thus be amenable to transfer learning. Likewise, PTMs that differ in the molecular characteristics of the PTM itself, but which utilize related enzymes, such as ubiquitin E3 ligases and SUMO E3 ligases, may also be amenable to transfer learning. Finally, all datasets and code developed during this study has been made freely available to the bioinformatics community at https://github.com/dukkakc/DTLDephos to further contribute toward the study of dephosphorylation.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/dukkakc/DTLDephos.

AUTHOR CONTRIBUTIONS

DK, RN, MK, and DC conceived, designed the experiments, and revised the manuscript. MC, HI, NT, and SC performed the experiments and data analysis. MC, DK, RN, and SC wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by National Science Foundation (NSF) grant nos. 1901793, 2003019, and 2021734 to DK. RN is supported by an HBCU-UP Excellence in Research Award from NSF (1901793) and an SC1 Award from the National Institutes of Health National Institute of General Medical Science (5SC1GM130545).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.662983/full#supplementary-material

REFERENCES

Ahmed, S., Kabir, M., Arif, M., Khan, Z. U., and Yu, D.-J. (2021). DeepPPSite: a deep learning-based model for analysis and prediction of phosphorylation sites using efficient sequence information. *Anal. Biochem.* 612:113955. doi: 10.1016/j.ab.2020.113955

Al-barakati, H., Thapa, N., Hiroto, S., Roy, K., Newman, R. H., and Kc, D. (2020). RF-MaloSite and DL-malosite: methods based on random forest and deep learning to identify malonylation sites. *Comput. Struct. Biotechnol. J.* 18, 852–860. doi: 10.1016/j.csbj.2020.02.012

Ardito, F., Giuliani, M., Perrone, D., Troiano, G., and Lo Muzio, L. (2017). The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review). *Int. J. Mol. Med.* 40, 271–280. doi: 10.3892/ijmm.2017.3036

Aridas GLitaFNaCK (2017). Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* 18, 1–5.

Bechtel, P. J., Beavo, J. A., and Krebs, E. G. (1977). Purification and characterization of catalytic subunit of skeletal muscle adenosine 3′:5′– monophosphate-dependent protein kinase. *J. Biol. Chem.* 252, 2691–2697. doi: 10.1016/s0021-9258(17)40514-x

Bornancin, F., and Parker, P. J. (1997). Phosphorylation of protein kinase C-alpha on serine 657 controls the accumulation of active enzyme and contributes to its phosphatase-resistant state. *J. Biol. Chem.* 272, 3544–3549. doi: 10.1074/jbc.272.6.3544

Chaudhari, M., Thapa, N., Roy, K., Newman, R. H., Saigo, H., and B. K. C D. (2020). DeepRMethylSite: a deep learning based approach for prediction of arginine methylation sites in proteins. *Mol. Omics* 16, 448–454. doi: 10.1039/d0mo00025f

Damle, N. P., and Köhn, M. (2019). The human DEPhOsphorylation Database DEPOD: 2019 update. *Database* 2019, 1–7.

Dinkel, H., Chica, C., Via, A., Gould, C. M., Jensen, L. J., Gibson, T. J., et al. (2011). Phospho.ELM: a database of phosphorylation sites–update 2011. *Nucleic Acids Res.* 39, D261–D267.

Feng, S., Zhou, H., and Dong, H. (2019). Using deep neural network with small dataset to predict material defects. *Mater. Des.* 162, 300–310. doi: 10.1016/j.matdes.2018.11.060

Guo, L., Wang, Y., Xu, X., Cheng, K.-K., Long, Y., Xu, J., et al. (2021). DeepPSP: a global–local information-based deep neural network for the prediction of protein phosphorylation sites. *J. Proteome Res.* 20, 346–356. doi: 10.1021/acs.jproteome.0c00431

Haixia, L., Zhao, S., Manzhi, L., Hai Yan, F., and Ming Cai, L. (2020). Predicting protein phosphorylation sites based on deep learning. *Curr. Bioinform.* 15, 300–308. doi: 10.2174/1574893614666190902154332

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780.

Hornbeck, P. V., Kornhauser, J. M., Latham, V., Murray, B., Nandhikonda, V., Nord, A., et al. (2019). 15 years of PhosphoSitePlus® : integrating post-translationally modified sites, disease variants and isoforms. *Nucleic Acids Res.* 47, D433–D441.

Ismail, H. D., Jones, A., Kim, J. H., Newman, R. H., and Kc, D. B. (2016). RF-Phos: a novel general phosphorylation site prediction tool based on random forest. *BioMed. Res. Int.* 2016:3281590.

Jia, C., He, W., and Zou, Q. (2017). DephosSitePred: a high accuracy predictor for protein dephosphorylation sites. *Comb. Chem. High Throughput. Screen* 20, 153–157.

Keras, C. F. (2015). *Keras: Deep Learning for Python*. Available online at: https://keras.io (accessed September 3, 2020).

Keshwani, M. M., Klammt, C., von Daake, S., Ma, Y., Kornev, A. P., Choe, S., et al. (2012). Cotranslational <em&gtcis-phosphorylation of the COOH-terminal tail is a key priming step in the maturation of cAMP-dependent protein kinase. *Proc. Natl. Acad. Sci.U.S.A.* 109:E1221.

Kingma, D. P., and Ba, J. (2014). A method for stochastic optimization. *arXiv*[preprint] arXiv:1412.6980,

Li, Y., Huang, C., Ding, L., Li, Z., Pan, Y., and Gao, X. (2019). Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods* 166, 4–21. doi: 10.1016/j.ymeth.2019.04.008

Li, Z., and Hoiem, D. (2018). Learning without Forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 2935–2947.

Lumbanraja, F. R., Mahesworo, B., Cenggoro, T. W., Budiarto, A., and Pardamean, B. (2019). An evaluation of deep neural network performance on limited protein phosphorylation site prediction data. *Procedia Comput. Sci.* 157, 25–30. doi: 10.1016/j.procs.2019.08.137

Luo, F., Wang, M., Liu, Y., Zhao, X.-M., and Li, A. (2019). DeepPhos: prediction of protein phosphorylation sites with deep learning. *Bioinformatics* 35, 2766–2773. doi: 10.1093/bioinformatics/bty1051

Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* 45, 2673–2681.

Senga, Y., Ishida, A., Shigeri, Y., Kameshita, I., and Sueyoshi, N. (2015). The phosphatase-resistant isoform of CaMKI, Ca2+/calmodulin-dependent protein kinase Iδ (CaMKIδ), remains in its "primed" form without Ca2+ stimulation. *Biochemistry* 54, 3617–3630. doi: 10.1021/bi5012139

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. (2018). "A survey on deep transfer learning," in *Artificial Neural Networks and Machine Learning – ICANN 2018: 2018//*, eds V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, and I. Maglogiannis (Cham: Springer International Publishing), 270–279.

Tang, B., Pan, Z., Yin, K., and Khateeb, A. (2019). Recent advances of deep learning in bioinformatics and computational biology. *Front. Genet.* 10:214. doi: 10.3389/fgene.2019.00214

Thapa, N., Chaudhari, M., McManus, S., Roy, K., Newman, R. H., Saigo, H., et al. (2020). DeepSuccinylSite: a deep learning based approach for protein succinylation site prediction. *BMC Bioinform.* 21:63.

UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515.

Wang, D., Liu, D., Yuchi, J., He, F., Jiang, Y., Cai, S., et al. (2020). MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Res.* 48, W140–W146.

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, 3909–3916. doi: 10.1093/bioinformatics/btx496

Wang, X., Yan, R., and Song, J. (2016). DephosSite: a machine learning approach for discovering phosphotase-specific dephosphorylation sites. *Sci. Rep.* 6:23510.

Wang, Y., Li, F., Bharathwaj, M., Rosas, N. C., Leier, A., Akutsu, T., et al. (2020). DeepBL: a deep learning-based approach for in silico discovery of beta-lactamases. *Brief. Bioinform.* bbaa301. doi: 10.1093/bib/bbaa301

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *J. Big Data* 3:9.

Wu, M., Yang, Y., Wang, H., and Xu, Y. (2019). A deep learning method to more accurately recall known lysine acetylation sites. *BMC Bioinform.* 20:49.

Wu, Z., Lu, M., and Li, T. (2014). Prediction of substrate sites for protein phosphatases 1B, SHP-1, and SHP-2 based on sequence features. *Amino Acids* 46, 1919–1928. doi: 10.1007/s00726-014-1739-6

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). "How transferable are features in deep neural networks?," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Vol. 2, (Montreal, CA: MIT Press), 3320–3328.

Zhao, W. (2017). Research on the deep learning of the small sample data based on transfer learning. *AIP Confer. Proc.* 1864:020018.

Conflict of Interest: MK is a paid consultant for Orion Pharma.

The remaining 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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership