Mini-review

Recent advances in post-translational modification site prediction based on deep learning

Meng, Lingkuan; Chan, Wai-Sum; Huang, Lei; Liu, Linjing; Chen, Xingjian; Zhang, Weitong; Wang, Fuzhou; Cheng, Ke; Sun, Hongyan; Wong, Ka-Chun

Published in:
Computational and Structural Biotechnology Journal

Published: 01/01/2022

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

License:
CC BY-NC-ND

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1016/j.csbj.2022.06.045

Publication details:
Meng, L., Chan, W.-S., Huang, L., Liu, L., Chen, X., Zhang, W., Wang, F., Cheng, K., Sun, H., & Wong, K.-C. (2022). Mini-review: Recent advances in post-translational modification site prediction based on deep learning. Computational and Structural Biotechnology Journal, 20, 3522-3532. https://doi.org/10.1016/j.csbj.2022.06.045

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher's copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.

Download date: 18/01/2025
Mini review

Mini-review: Recent advances in post-translational modification site prediction based on deep learning

Lingkuan Meng\textsuperscript{a,b}, Wai-Sum Chan\textsuperscript{a}, Lei Huang\textsuperscript{a}, Linjing Liu\textsuperscript{a}, Xingjian Chen\textsuperscript{a}, Weitong Zhang\textsuperscript{a}, Fuzhou Wang\textsuperscript{b}, Ke Cheng\textsuperscript{b}, Hongyan Sun\textsuperscript{b,*}, Ka-Chun Wong\textsuperscript{a,*}

\textsuperscript{a} Department of Computer Science, City University of Hong Kong, Hong Kong Special Administrative Region
\textsuperscript{b} Department of Chemistry, City University of Hong Kong, Hong Kong Special Administrative Region

Abstract

Post-translational modifications (PTMs) are closely linked to numerous diseases, playing a significant role in regulating protein structures, activities, and functions. Therefore, the identification of PTMs is crucial for understanding the mechanisms of cell biology and diseases therapy. Compared to traditional machine learning methods, the deep learning approaches for PTM prediction provide accurate and rapid screening, guiding the downstream wet experiments to leverage the screened information for focused studies. In this paper, we reviewed the recent works in deep learning to identify phosphorylation, acetylation, ubiquitination, and other PTM types. In addition, we summarized PTM databases and discussed future directions with critical insights.

© 2022 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Abbreviations:
AAindex, Amino acid index; ATP, Adenosine triphosphate; AUC, Area under curve; Ac, Acetylation; BE, Binary encoding; BLOSUM, Blocks substitution matrix; Bi-LSTM, Bidirectional LSTM; CKSAAP, Composition of k-spaced amino acid Pairs; CNN, Convolutional neural network; CNN\textsubscript{1hot}, CNN with the one-hot encoding; CNN\textsubscript{we}, CNN with the word-embedding encoding; CNNgb, CNN red green blue; CV, Cross-validation; DC-CNN, Densely connected convolutional neural network; DL, Deep learning; DNNS, Deep neural networks; EBGW, Encoding based on grouped weight; EGAAC, Enhanced grouped amino acids content; E. coli, Escherichia coli; IG, Information gain; K, Lysine; KNN, k nearest neighbor; LASSO, Least absolute shrinkage and selection operator; LSTM, Long short-term memory; LSTM\textsubscript{we}, LSTM with the word-embedding encoding; MDCAN, Multilane dense convolutional attention network; MDC, Modular densely connected convolutional networks; ML, Machine learning; MLP, Multilayer perceptron; MMT, Multivariate mutual information; M.musculus, Mus musculus; NMBrbo, Normalized Moreau-Broto autocorrelation; P, Proline; PSP, PhosphoSitePlus; PSSM, Position-specific scoring matrix; PTM, Post-translational modifications; Ph, Phosphorylation; PieAAC, Pseudo-amino acid composition; R, Arginine; RF, Random forest; RNN, Recurrent neural network; ROC, Receiver operating characteristic; SE, Squeeze and excitation; SEV, Split to Equal Validation; S, Serine; ST, Source and target; SUMO, Small ubiquitin-like modifier; SVM, Support vector machines; S.cerevisiae, Saccharomyces cerevisiae; S. typhimurium, Salmonella typhimurium; T, Threonine; Ub, Ubiquitination; V, Tyrosine; ZSL, Zero-shot learning.

* Corresponding authors.
E-mail addresses: hongysun@cityu.edu.hk (H. Sun), kc.w@cityu.edu.hk (K.-C. Wong).

https://doi.org/10.1016/j.csbj.2022.06.045
2001-0370© 2022 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
1. Introduction

Post-translational modifications (PTMs) generally refer to the addition of functional groups (e.g., phosphates, acetates, small proteins, lipids, carbohydrates, etc.) to amino acids during translation [1]. After PTM, amino acids’ chemical properties or structures will be changed, leading to functional changes. To date, over 600 different types of PTMs have been discovered in different proteins [2,3]. It is known that phosphorylation, acetylation, and ubiquitination are the extensively studied PTMs, as quantified with the dbPTM [4] database. PTMs are critical in maintaining protein structures [5], functions [6], metabolic regulation [7], cellular signaling [8], and proteomic diversity [9], whereby our understanding of PTMs is essential to downstream consequences such as diseases. For example, S-nitrosylation is a promising therapeutic target for cancers and neurodegenerative diseases [10–12]; methyl glutamine is associated with the host defense mechanism against microorganisms [13,14]. Different experimental techniques have been developed to reveal the mechanisms underlying PTMs, including chromatin immunoprecipitation (ChIP) [15], western blotting (WB) [16], mass spectrometry (MS) [17,18], and isotope labeling [19]. In the recent decade, MS-based proteomic techniques [20] play a major role in PTM identification, which yield solid data with actual evidence [21]. In addition, computational methods can also explore and predict new modification sites by building a model from those data. In the last few years, machine learning has grown to be a cost-effective and labor-efficient method for the prediction of various PTM sites [22–28]. Specifically, deep learning is an advanced machine learning method that is capable of automatically exploring PTM patterns and capturing high-level abstraction (Fig. 1 [29]). Therefore, it is an appropriate solution to improve the efficiency of PTM sites’ prediction with growing interest in recent years (Fig. 2). A lot of published works focused on adopting deep learning to predict PTM sites for phosphorylation [30], acetylation [31], ubiquitination [32], and many other types of modifications [33,34]. One of the most famous tools is MusiteDeep [30], developed by Wang and Zeng, which leveraged convolutional neural network (CNN) and 2D attention mechanism for phosphorylation sites prediction. DeepPhos [35], which is created by Luo et al., is an efficient phosphorylation sites predictor to identify not only general but also kinase-specific sites. Moreover, Wu et al. [36] and Fu et al. [37] developed deep learning-based methods to predict acetylation and putative ubiquitination sites with promising results.

In this mini-review, we summarized and discussed the most recent (2020–2022) progress made in the prediction of PTMs using deep learning-based methods with a particular emphasis on protein phosphorylation, acetylation, and ubiquitination sites. Moreover, we presented frequently used databases for deep learning-based PTM prediction, along with future directions in the computational identification of PTMs.

2. PTM databases

Available PTM datasets can mainly be retrieved from two sources: databases with various types of data and scientific literature data. The obtained data can be used to train a model for PTM prediction. Table 1 summarizes the leading databases with different data types based on recent literature [38–43].

2.1. UniProt

UniProt [38] is one of the most comprehensive databases with PTM annotations; it contains annotations for a wide variety of PTMs. UniProt data is of high quality and was recognized as an ELIXIR Core Data Resource in 2017 [44]. The database received the CoreTrustSeal certification in 2020. It has four components customized for different uses: UniParc, UniProtKB, UniRef, and UniMES. Notably, the UniProtKB database has become the gateway to protein functional information. Over the last two years, UniProtKB’s sequences have grown to about 190 million [45], despite efforts in sequence redundancy removal at the proteome level. According to the survey, we found that most of the literature collects datasets from UniProtKB as their benchmark datasets. The latest version of the UniProt database can be accessed by visiting https://www.uniprot.org/.

2.2. PLMD

There are 20 types of protein lysine modifications across 176 species in PLMD [43]. The PLMD database was constructed from the CPLA and CPLM databases with manual curations. It contains 284,780 protein lysine modification sites in 53,501 proteins, including 111,253 acetylation sites and 121,742 ubiquitination sites. To the best of our knowledge, it is the largest available database of protein acetylation, along with the largest database of protein ubiquitination sites, which has never been reported in any other ubiquitination sites prediction research. There is a free and open-source version of PLMD 3.0 at https://plmd.biocuckoo.org, which is implemented in PHP and MySQL.

2.3. PhosphoSitePlus

PhosphoSitePlus (PSP) [40] offers comprehensive data information for studying PTMs, such as phosphorylation, SUMOylation, ubiquitination, and others. Manually collected and organized data are curated to constitute this database, which primarily contains human and mouse protein data. At the time of writing, it has harbored 598,976 nonredundant modified sites, including 294,425 phosphorylation sites. The PSP database is versatile, offering a variety of information about the modification sites. PSP is a free database that can be accessed through https://www.phosphosite.org.

3. Phosphorylation site prediction

Phosphorylation is one of the most frequently investigated PTM, referring to the transfer of phosphate groups (PO₄) from adenosine triphosphate (ATP) sites to amino acid chains via the catalysis of various kinases [46]. Typically, phosphorylation of proteins occurs at serine (S), threonine (T), or tyrosine (Y) [47]. Approximately 13,000 human proteins can be phosphorylated, and 230,000 phosphorylation sites in human proteome were reported [48]. In the past decades, phosphorylation studies have gained widespread popularity due to their significance in characterizing signaling pathways [49,50] and cellular processes, such as cell growth [51], cell division [52], and apoptosis [53]. With the development of high-throughput MS-based technology, a single proteomic experiment can detect large-scale phosphorylation. Therefore, various databases have been built to collect annotated phosphorylation sites [38–40]. The application of these databases in recent years
has been enabled through the extensive development of computational methods for phosphorylation sites identification [22,54-58]. In machine learning, we can formulate the phosphorylation site prediction problem as two classification tasks. The first task is the general site prediction, which aims to determine whether a given site can be modified. The second task is the kinase-specific prediction, which determines whether a site can be modified by a particular kinase [29]. In particular, the recent development of deep learning could speed up the progress of phosphorylation site prediction. A well-known deep learning-based predictor, MusiteDeep [30], incorporates one-hot encoding and CNN with attention layers and performs better than previous feature-based models. Another phosphorylation site prediction method, DeepPhos [35], exploits densely connected convolutional neural network (DC-CNN) blocks for predictions. The results of DeepPhos outperform MusiteDeep in not only general sites but also kinase-specific sites.

Fig. 1. Overview of deep learning approaches for PTM prediction. [29].
predictions. Recently, a single unified multi-label classification model, EMBER [58], was released. Unlike the previous deep learning methods, MusiteDeep and DeepPhos, which perform single-label classification, EMBER was designed to predict phosphorylation events for multiple kinases. In this tool, the input sequence is fifteen amino acids in length, of which the eighth site is to be predicted. The sequence is encoded using both one-hot encoding and embedding generated from a siamese neural network. After

Table 1
Summary of PTM databases harbored.

| Database       | Development Year | Number of PTM Sites Deposited | Database Link                      | Annotation                                      | Reference |
|----------------|------------------|-------------------------------|------------------------------------|-------------------------------------------------|-----------|
| UniProt        | 2005             | Varies according to the keyword search | [38]                               | Multiple-type PTM sites for multi-species       | [38]      |
| PLMD           | 2017             | 284,780                       | [43]                               | Protein lysine modification sites for multi-species | [43]      |
| PhosphoSitePlus| 2012             | 598,976                       | [40]                               | Multiple-type PTM sites for multi-species        | [40]      |
| Phospho.ELM    | 2010             | 42,914                        | [39]                               | Phosphorylation sites for Eukaryotic             | [39]      |
| mLbiSiDa       | 2014             | 110,976                       | [41]                               | Uniquitination sites mainly for Human and Mouse  | [41]      |
| DEPOD          | 2015             | 1,215                         | [42]                               | Dephosphorylation interactions                   | [42]      |
encoding, both sequences are fed into their corresponding identical CNNs. In the top layer, the two feature vectors are concatenated, followed by fully connected layers. Finally, the output is a vector of length eight, where each value represents the probability that a family of kinases will phosphorylate an input site. In addition, different tools are also proposed to predict protein-specific phosphorylation sites. In 2020, Chen et al. developed PROSPECT [56] which is a method for phosphorylation site prediction based on one-of-K, EGAAAC, and CKSAAGP encodings [35,59]. The classifier for one-of-K encoding is built with a multi-layer attention-based CNN; and the classifier for EGAAAC encoding employs a multi-layer CNN. In the case of CKSAAGP encoding, the random forest (RF) algorithm is used to train the classifier. After that, an online web server of PROSPECT is developed. In the same year, Wang et al. also presented a web server named MusiteDeep based on their deep-learning models implemented in 2017. The server is capable of providing real-time prediction and batch submission for large-scale protein sequences, as listed in Table 4. Conclusively, we compare the performance of recent deep learning-based phosphorylation predictors in Table 2.

4. Acetylation site prediction

Acetylation is a very common PTM that describes the modification of the acetyl group to amino acid residues. About 63% of mitochondrial proteins can be acetylated at their lysine residues [65]. During the protein acetylation process, the positive charge in lysine residues is neutralized, leading to the regulation of cell lifespan [66], DNA binding [67], the interactions between proteins [68], and the interactions between proteins and membranes [69]. In contrast, dysregulation of lysine acetylation is associated with several diseases, including cancers [70], cardiovascular diseases [71], Parkinson’s diseases [72], and neurodegenerative disorders [73]. Thus, the identification of acetylation sites may benefit the understanding of its molecular mechanism and further experimental design. Proteomic and high-throughput MS-based techniques have identified massive acetylation sites. For example, Choudhary et al. detected 3,600 lysine acetylation sites on 1,750 proteins from a human cell line. [74]; Lundby et al. quantified 15,474 lysine acetylation sites on 4,541 proteins from 16 rat tissues [75]. Several public databases have been developed to facilitate the collection and maintenance of acetylation sites information [38,43]. Therefore, to predict acetylation sites, many computational methods have been proposed [36,76,77]. Among them, deep learning methods are increasingly popular in bioinformatics, which also show encouraging results of acetylation sites identification [78-80]. For example, Wu et al. [36] presented an MLP architecture, DeepAcet, as an acetylation site prediction model. Feature embedding were performed with six methods (One-hot, IG, CKSAAP, PSSM, AAindex, and BLOSUM62); multilayer perceptron (MLP) is then applied to extract features. After adopting 10-fold cross-validation method [81] paired model evaluation on a separate test site, accuracies were reported to be 0.8495 and 0.8487, respectively. Yu et al. also developed a deep neural networks (DNN) based model called DNNAce for acetylation sites prediction [78]. First, they applied eight different encoding methods to extract information from multiple amino acid residues and then fused the encoded feature vectors to create a high-level feature representation. These encodings methods are BE, PseaAC, AAindex, NMBroto, EBGW, MMI, BLOSUM62, and KNN. Next, they employ LASSO to screen the optimal feature subsets to improve the model performance. As a final stage, nine prokaryotic acetylation site datasets are adopted to evaluate the performance and compared to state-of-the-art models such as AdaBoost, Naive Bayes, XGBoost, KNN, RF, SVM, CNN, and LSTM. An evaluation of DNNAce was conducted by comparing its results with ProAcePred [82]. The performance of DNNAce on the remaining eight species was significantly lower than that of ProAcePred except for S. typhimurium species. However, DNNAce outperforms ProAcePred for the other seven species during independent evaluation. Therefore, the advantages of DNNAce are trivial because there is performance discrepancy in training and independent testing. In contrast to deepAcet and DNNAce, which only consider the amino acid sequences and their physicochemical properties, MDC-Kace [80] pays attention to both sequence information and protein structural properties to predict acetylation sites. In MDC-Kace, modular densely connected convolutional networks (MDC), which consist of three independent modules (sequence, physicochemical and structure), is employed to extract features of lysine acetylation sites. In the next step, squeeze and excitation (SE) layer [83] is utilized to weight importance of features to build representation more accurately. Finally, the fused advanced feature is fed into a softmax layer for classification to predict acetylation sites efficiently. The authors compared MDC-Kace with state-of-the-art models (MusiteDeep [30], CapsNet [34], DeepAcet [36], PSKacePred [84], EnsemblePail [85], GPS-PAIL2.0 [86] and ProAcePred [82]) to evaluate its performance. Three species (human, M. musculus, E. coli) datasets have been evaluated by 10-fold cross-validation and independent testing. The results indicate that MDC-Kace has a similar performance as existing acetylation sites predictors.

5. Ubiquitination site prediction

Ubiquitination represents an enzymatic PTM on cellular protein by ubiquitin conjugation [87]. Multiple important cellular processes are related to ubiquitination, including protein degradation [88], cell division [89], and protein stability [90,91]. Ubiquitination serves as a fundamental component of the ubiquitin–proteasome system, mediating more than 80% of protein degradation in eukaryotes [92]. Moreover, aberrant ubiquitination is highly

| Tool name     | Framework     | Encoding strategy | Window size | Average AUC | Reference |
|---------------|---------------|-------------------|-------------|-------------|-----------|
| MusiteDeep    | Keras/TensorFlow | One-hot            | 33          | 0.880       | [30]      |
| PROSPECT     | PyTorch       | One-hot, EGAAAC, CKSAAGP | 27          | 0.770       | [56]      |
| DeepKinZero  | TensorFlow   | Word embedding     | –           | 0.898       | [61]      |
| PhysTransfer | TensorFlow   | BLOSUM62           | 21          | 0.832       | [62]      |
| GPS-PBS      | Keras/TensorFlow | BLOSUM62           | 21          | 0.827       | [57]      |
| DeepPysite   | Keras/TensorFlow | BE, EBGW, CKSAAP, PSPM, IPCP | 21          | 0.909       | [63]      |
| PhosDIN      | Keras/TensorFlow | One-hot, PPI embedding | 21          | 0.939       | [64]      |
| EMBER        | PyTorch       | One-hot            | 15          | 0.928       | [58]      |

Note: -, data not available. AUC: Area under the Curve of ROC.
related to the progression of aging [93] and many diseases; for example, the dysregulation of ubiquitin–proteasome system may contribute to the occurrence of neurodegenerative conditions [94] and inflammatory bowel diseases [95]. Therefore, the identification of ubiquitination sites is an essential step in exploring various ubiquitination-involved mechanisms. In order to identify the ubiquitination sites in proteins, a myriad of experimental [96-98] and computational methods [99-101] have been developed. In recent years, with the continuous growth in high-throughput experimental data [102-104], deep learning [105-107] has been increasingly applied to the prediction of ubiquitination. Fu et al. proposed a deep learning predictor, DeepUbi [37], based on CNN. In this tool, four feature encoding schemes are utilized for feature construction. Under 10-fold cross-validation, DeepUbi is able to achieve an AUC of 0.90, with the accuracy, sensitivity, and specificity being all over 0.85. Compared with DeepUbi, which is trained for general ubiquitination site prediction, DeepTL-Ubi [106] is a species-specific sites predictor which consists of three connected modules: a deep feature extractor, a source label classifier, and a target label classifier. Firstly, a densely connected convolutional neural network (DCCNN) is applied as the deep feature extractor, which is composed of six layers. Features of both source species and target species are extracted simultaneously by the deep feature extractor, mapping samples into a joint feature space. Secondly, the two parallel classifiers are employed to classify source species and target species at the same time. Thirdly, ST(source and target) loss assists the extractor in transferring knowledge from source species to target species by learning relevant features. Finally, as the performance optimization step, the classification loss is minimized to train the two classifiers. DeepTL-Ubi outperforms several existing tools, including Ubisite [108], Ubiprober [24], and MUscADEL [109], as shown in Table 3.

6. Other PTMs

In addition to those discussed, deep learning can also be applied for other PTMs’ predictions, including methylation [110], S-nitrosylation [111], succinylation [112,113], malonylation [114,115], S-sulphenylation [116,117], crotonylation [118-121], 2- hydroxyisobutyrylation [122], glutarylation [123], N-palmitoylation [124] carbonylation [125], and SUMoylation [126]. In particular, crotonylation prediction has demonstrated highly accurate results based on deep-learning methods. Moreover, 2- hydroxyisobutyrylation, as a novel type of PTM, was predicted by deep learning method for the first time in 2020.

Along with predicting conventional PTMs associated with functional group addition, deep learning-based methods have also been applied to predict niche-type PTMs; for instance, Chaudhari et al. developed a transfer learning-based predictor (DTL-DephosSite) for dephosphorylation site prediction [127]. To collect datasets of S, T, and Y dephosphorylation sites, they integrated the experimentally verified datasets from the literature and datasets from the DEPOD database. They then employ bidirectional long short-term memory (Bi-LSTM), which can predict the modification of the target amino acid according to the knowledge of residues from both directions. To the best of our knowledge, it is the first tool that can predict the general dephosphorylation sites for protein S/T residues and Y residues. On the other hand, a novel prediction model focusing on carbonylation, Precar_Deep [125], is recently reported. Carbonylation is an irreversible covalent PTM and is a measure of protein oxidative damage. In this model, CNN and Bi-LSTM are combined under a deep learning framework. The AUC values of the four datasets (K, T, P, and R) reach 0.981, 0.982, 0.987, and 0.976, respectively. The AUC values of the independent test set reach 0.945, 0.978, 0.965, and 0.983, respectively. In addition, there is also a novel small protein-addition type PTM site predictor based on deep learning in 2021. He et al. built an ensemble learning model that adopts CNN and DNN, followed by the output result containing four types of sites. [126] This is the first tool that predicts both ubiquitylation and SUMOylation sites at the same time based on deep learning. PTM prediction tools mentioned in this section, as well as predictors of phosphorylation, acetylation, and ubiquitination, are tabulated in Table 4.

7. Summary and outlook

PTM identification is critical to a better understanding of molecular functions and diseases. Advanced MS-based technology has yielded an extensive list of identified PTMs, providing abundant data to support the development of downstream computational identification methods. Although the traditional machine learning methods can precisely predict the modified sites, deep learning features can be automatically deduced and optimally turned without encoding features ahead of time [29]. Thus, deep learning is highly effective in scientific fields with large and complex datasets. Researchers recently gradually shift their attention from traditional machine learning to deep learning for PTM site prediction (Fig. 2). Furthermore, with the growing number of PTM profiling datasets, deep learning models have been developed for not only phosphorylation, acetylation, and ubiquitination, but also many other PTM types. In this review, we summarized the recently (2020–2022) released deep learning tools and online web servers for protein PTM site prediction (Table 4). Among all these, CNN and cross-validation are the most widely used network model and evaluation strategy, respectively (Fig. 3).

Although several deep learning methods have been built with high performance to predict PTM sites, there is still room for improvement. Most of the existing deep learning algorithms employed CNN, DNN, and LSTM classifiers. However, each classifier has its own advantages and disadvantages. Therefore, further research is required to evaluate more state-of-the-art frameworks such as attention and transformer-based models. On top of that, in many developed tools, although PTM sites are predicted based on certain characteristics, such as sequence information, physical properties, chemical properties, and protein structure properties, there are still other approaches that need to be explored, such as reduced amino acid compositions [128-130]. Additionally, most of web server links are not working, and few methods provide stand-alone versions. After testing all web servers, we found that they were difficult to operate.

Table 3

| Tools          | AUC H.sapiens | AUC M.musculus | AUC R.norvegicus | AUC S.cerevisiae | AUC T.gondii | AUC A nidulans |
|----------------|---------------|----------------|-------------------|------------------|--------------|----------------|
| DeepTL-Ubi     | 0.753         | 0.789          | 0.720             | 0.772            | 0.824        | 0.814          |
| Ubisite        | 0.598         | 0.625          | 0.561             | 0.548            | 0.607        | 0.611          |
| Ubiprober      | 0.624         | 0.661          | 0.644             | 0.699            | 0.630        | 0.638          |
| MUscADEL       | 0.656         | 0.693          | 0.659             | 0.664            | 0.715        | 0.681          |

L. Meng, Wai-Sum Chan, L. Huang et al. Computational and Structural Biotechnology Journal 20 (2022) 3522–3532
| Tool name     | PTM type       | Species                          | Core network model | Evaluation strategy | Benchmark dataset size (modification sites) | Web server/ source code | Published year | Reference |
|--------------|----------------|---------------------------------|--------------------|---------------------|---------------------------------------------|--------------------------|----------------|-----------|
| MusiteDeep   | Multiple       | Human, Escherichia coli         | CNN                | 5-fold CV           | 999,687                                     | https://www.musite.net   | 2017/2020    | [30]      |
| PROSPECT     | Phosphorylation| Escherichia coli                | CNN                | 10-fold CV          | 1,664                                       | *prospect.erc.monash.edu*| 2020          | [56]      |
| DeepKinZero  | Phosphorylation| Human                           | ZSL                | holdout             | 12,901                                      | *https://github.com/    | 2020          | [60]      |
| PhosTransfer | Phosphorylation| –                               | CNN                | 5-fold CV holdout   | 43,785                                      | https://github.com/     | 2020          | [61]      |
| GPS-PBS      | Phosphorylation| Multiple                        | seven-layer DNNs   | 10-fold CV          | 4,458                                       | –                        | 2020          | [62]      |
| DeepPPSite   | Phosphorylation| Mammals and Arabidopsis thaliana|       | 10-fold CV          | 41,436                                      | github.com/saeed344/   | 2021          | [57]      |
| DeepIPs      | Phosphorylation| Human                           | CNN + LSTM         | 5-fold CV           | 10,978                                      | https://lin-group.cn/   | 2021          | [63]      |
| PhosIDN      | Phosphorylation| Human                           | Multi-layer DNNs   | holdout             | more than 160,000                           | https://github.com/     | 2021          | [64]      |
| EMBER        | Phosphorylation| Multiple                        | CNN + RNN          | 5-fold CV           | 8,389                                       | https://github.com/gomezlab/EMBER | 2022          | [58]      |
| DNNAce       | Acetylation    | Multiple                        | DNN                | 10-fold CV and independent test | 96,372                                      | https://github.com/     | 2020          | [78]      |
| Deep-PLA     | Acetylation    | Human and Nonhuman              | CNN                | 5- and 10- fold CV  | 1,331                                       | https://deeppla.cancerbio.info | 2020          | [79]      |
| MDC-Kace     | Acetylation    | Multiple                        | MDC                | 10-fold CV and independent test | 11,583                                      | https://github.com/liangliang/MDC-Kace | 2020          | [80]      |
| DeepTL-Ubi   | Ubiquitination | Multiple                        | CNN                | holdout             | 94,518                                      | https://github.com/USTC-HiLab/DeepTL-Ubi | 2020          | [106]     |
| Wang et al.’s work | Ubiquitination | Multiple                        | CNN                | 10-fold CV          | 121,742                                     | *https://github.com/wang-hong-fei/DeepTL-Ubi | 2020          | [105]     |
| UbiComb      | Ubiquitination | Multiple                        | LSTM               | 10-fold CV          | 121,742                                     | https://nsclbio.jbnu.ac.kr/tools/UbiComb | 2021          | [107]     |
| SSMFN        | Methylation    | Human and Mouse                 | CNN + LSTM         | holdout             | 6,754                                       | *https://github.com/bharuno/SSMFNMethylation-Analysis | 2021          | [110]     |
| Malebary et al.’s work | Methylation | Human                           | CNN                | 10-fold CV and jackknife | 2000                                        | https://github.com/s2018 https://doi.org/1080001/WeBServer.gt | 2022          | [14]      |
| RecSNO       | S-Nitrosylation| –                               | BiLSTM             | 5-fold CV           | 4,762                                       | https://nsclbio.jbnu.ac.kr/tools/RecSNO/ | 2021          | [111]     |
| MDCAN-Lys    | Succinylation  | Human                           | MDCAN              | 10-fold CV          | 77,418                                      | –                        | 2021          | [112]     |
| LSTMCCNNSucc | Succinylation  | Multiple                        | LSTM + CNN         | holdout             | 18,593                                      | https://github.com/     | 2021          | [113]     |
| DeepMal      | Malonylation   | Multiple                        | LSTM + CNN + DNN   | 10-fold CV and independent test | 17,288                                      | https://github.com/QUST-AIBBDRC/DeepMal | 2021          | [114]     |
| K_net        | Malonylation   | Human and Mouse                 | CNN                | 10-fold CV and SEV  | 85,204                                      | –                        | 2020          | [115]     |
| DeepCSO      | S-Sulphenylation | Homo sapiens and Arabidopsis thaliana | LSTM             | 10-fold CV          | 10,354                                      | *https://github.com/deepCSO | 2020          | [116]     |
| DeepSSPred   | S-Sulphenylation| Homo sapiens                     | 2D-CNN             | jackknife           | 7,756                                       | *https://github.com/zaheerkhancs/DeepSSPred | 2021          | [117]     |
| pkcr         | Crotonylation  | Homo sapiens and Papaya         | CNN                | 10-fold CV and independent test | 58,769                                      | *https://github.com/pkcr | 2020          | [119]     |
| Deep-Kcr     | Crotonylation  | Human                           | CNN                | 10-fold CV          | 19,928                                      | https://lin-group.cn/server/Deep-Kcr | 2020          | [120]     |
**Table 4 (continued)**

| Tool name               | PTM type                  | Species                        | Core network model | Evaluation strategy | Benchmark dataset size (modification sites) | Web server/ source code | Published year | Reference |
|-------------------------|---------------------------|--------------------------------|-------------------|---------------------|---------------------------------------------|--------------------------|----------------|-----------|
| DeepKcrot               | Crotonylation             | Multiple                       | CNN novo          | 10-fold CV and independent test | 10,702/1,265/2,044/5,995 | *https://www.bioinfogo.org/deepkcrot.* | 2021 | [121]    |
| nhKcr                   | Crotonylation             | Human                          | CNN rgb           | 10-fold CV and independent test | 180,312                                    | https://nhKcr.erc.monash.edu/ | 2021 | [118]    |
| DeepKhib                | 2-Hydroxyisobutyrylation  | Multiple                       | CNN eff           | 10-fold CV and independent test | 18,946/15,444/12,756/19,330/2,098 | *https://www.bioinfogo.org/DeepKhib.* | 2020 | [122]    |
| DeepGlut                | Glutarylation             | Prokaryotes and Eukaryote      | CNN               | 10-fold CV           | 4,572                                       | *https://github.com/urmisen/DeepGlut.* | 2020 | [123]    |
| NPalmitylDeep-PseAAC    | N-Palmitoylation          | Human                          | DNN               | holdout              | 4,364                                       | https://mega.nz/#F!s9cSiQal1jX00NgmrhxUQoDexmyYuouA | 2021 | [124]    |
| DTL-DephosSite          | Dephosphorylation         | Human                          | Bi-LSTM           | 5-fold CV and independent test | 4,956                                       | https://github.com/dukkakc/DTLDephos | 2021 | [127]    |
| PreCar_Deep             | Carbonylation             | Human and other Mammals        | CNN + BiLSTM      | 10-fold CV and independent test | 5,003                                       | https://github.com/QUST-SHULI/PreCar_Deep | 2021 | [125]    |
| He et al.’s work        | SUMOylation Ubiquitylation| –                              | CNN + DNN         | 10-fold CV           | 280,731                                     | https://github.com/lijingyimm/MultiUbiSUMO | 2021 | [126]    |

**Note:** *, Link is not working at the time of writing. Multiple, more than three species or PTM types. -, data not available.

**Fig. 3.** Sankey diagram depicting the distribution of PTM types, core network models, evaluation strategies, and published years.
By using deep learning based methods, PTM identification can be implemented in a non-invasive, efficient, and low-cost way. However, there is still a caveat before deep learning algorithms can directly diagnose diseases. Typical PTM prediction models lack sufficient interpretations due to the black-box nature of deep learning algorithms. Insufficient interpretability may not be an issue in many areas, but within healthcare, every misdiagnosis can pose a danger to a patient’s health. Therefore, transparent and explainable models [131–133] will be needed, so that the technique can be applied in clinical practice.

CRediT authorship contribution statement

Lingkuan Meng: Writing. Conceptualization. Methodology. Visualization. Wai-Sum Chan: Methodology. Lei Huang: Methodology. Linjing Liu: Methodology. Weitong Zhang: Methodology. Fuzhou Wang: Methodology. Ke Cheng: Methodology. Hongyan Sun: Writing – review & editing. Supervision. Ka-Chun Wong: Writing – review & editing. Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was financially supported by the (1) National Natural Science Foundation of China (Grant No. 32170654, 32122003 and 32000464), (2) the Hong Kong Research Grants Council (No. 11102719, 11302320), (3) Health and Medical Research Fund, the Food and Health Bureau, The Government of the Hong Kong Special Administrative Region [07181426], and (4) City University of Hong Kong (CityU 11202219, CityU 11203520, CityU 11203221).

References

[1] Walsh CT, Garneau-Tsodikova S, Gatto Jr GJ. Protein posttranslational modifications: the chemistry of proteome diversifications. Angew Chem Int Ed Engl 2005;44:7342–72. http://10.1002/anie.200501023.
[2] https://www.uniprot.org/docs/pmlist.
[3] UniProt C., UniProt: a worldwide hub of protein knowledge, Nucleic Acids Res. 2019;47:D506-D515. http://10.1093/nar/gkj0140.
[4] Lee T-Y, Huang H-D, Hung J-H, Huang H-Y, Yang Y-S, Wang T-H. dbPTM: an information repository of protein post-translational modification, Nucleic Acids Res. 2006;34:D622–D627. http://10.1093/nar/gjk083.
[5] Craveur P, Narwani TJ, Rebehmed J, de Brevern AG. Investigation of the impact of post-translational modifications to regulate protein function. Wiley Encyclopedia of Biochemistry 2005;44:7342–72. https://doi.org/10.1002/0470046777.wdbi007.
[6] Lin H, Du J, Jiang H. Post-translational modifications to regulate protein backbone conformation. Amino Acids 2019;51:1065–79. https://doi.org/10.1007/s00726-019-02747-w.
[7] Humphrey SJ, James DE, Mann M. Protein Phosphorylation: a major switch mechanism for metabolic regulation. Trends Endocrinol Metab 2015;26:676–87. https://doi.org/10.1016/j.tem.2015.09.013.
[8] Sharma K, D’Souza RC, Tyanova S, Schaab C, Wisniewski JR, Cox J, et al. DBPhos: predicting N6-methyladenosine sites in RNA sequences via physical–chemical properties. Bioinformatics 2016;32:3116–23. https://doi.org/10.1093/bioinformatics/btw443.
[9] Wang D, Liang Y, Xu D. Capsule network for protein post-translational modification site prediction using logistic regression. Sci Rep 2018;8:31943. https://doi.org/10.1038/s41598-021-87569-1.
[10] Zhou C, Li J, Nie J, Li X, Ni J, Zhang Y, et al. iSNO-AAPair: incorporating amino acid and position features to identify general and species-specific S-nitrosylation sites using a bi-modal deep architecture. IEEE Access 2018;6:31943–5. https://doi.org/10.1109/access.2018.2874882.
[11] He F, Wang R, Li J, Mao L, Xu D, Zhao X. Large-scale prediction of protein ubiquitination sites using a multimodal deep architecture. BMC Bioinformatics 2018;12:109. https://doi.org/10.1186/s12859-018-0628-9.
[12] Long H, Liao B, Xu X, Yang J. A hybrid deep learning model for predicting protein hydroxylation sites. Int J Mol Sci 2018;19. https://doi.org/10.3390/ijms19092817.
[13] Wang D, Liang Y, Xu D. Capsule network for protein post-translational modification site prediction. Bioinformatics 2018;35:2386–94. https://doi.org/10.1093/bioinformatics/bty628.
sequence and protein–protein interaction information. Bioinformatics 2021;37:4668–76. https://doi.org/10.1093/bioinformatics/btab551.

[86] Deng W, Wang C, Zhang Y, Xu Y, Zhang S, Liu Z, et al. GPS-PAIL: prediction of lysine acetylation sites prediction based on elastic net feature optimization. Bioinformatics 2018;34:4102–10. https://doi.org/10.1093/bioinformatics/bty705.

[87] Glickman MH, Ciechanover A. The ubiquitin-proteasome proteolytic sequence and protein–protein interaction information. Bioinformatics 2021;37:4668–76. https://doi.org/10.1093/bioinformatics/btab551.

[88] Deng W, Wang C, Zhang Y, Xu Y, Zhang S, Liu Z, et al. GPS-PAIL: prediction of lysine acetylation sites prediction based on elastic net feature optimization. Bioinformatics 2018;34:4102–10. https://doi.org/10.1093/bioinformatics/bty705.

[89] Glickman MH, Ciechanover A. The ubiquitin-proteasome proteolytic sequence and protein–protein interaction information. Bioinformatics 2021;37:4668–76. https://doi.org/10.1093/bioinformatics/btab551.

[90] Deng W, Wang C, Zhang Y, Xu Y, Zhang S, Liu Z, et al. GPS-PAIL: prediction of lysine acetylation sites prediction based on elastic net feature optimization. Bioinformatics 2018;34:4102–10. https://doi.org/10.1093/bioinformatics/bty705.
Chen Z, Chen Y-Z, Wang X-F, Wang C, Yan R-X, Zhang Z. Prediction of protein ubiquitination using cell extracts and protein microarrays. PNAS 2009;106:2543–8. https://doi.org/10.1021/bi015940q

Du Y, Xu N, Lu M, Li T. hUbiquitome: a database of experimentally verified protein ubiquitination sites in human. Front Cell Dev Biol 2020;8:897. https://doi.org/10.3389/fcell.2020.572195

Lv H, Xiao Z-X. Regulation of p63 protein stability via ubiquitin-proteasome system. J Proteomics 2017;89:1–11. https://doi.org/10.1016/j.jprot.2017.06.007

Wang M, Cui X, Li S, Yang X, Ma A, Zhang Y, et al. DeepMal: Accurate prediction of protein malonylation sites by deep neural networks. Chemom Intel Lab Syst 2020;207: https://doi.org/10.1016/j.chemolab.2020.104722

Sun J, Cao Y, Wang D, Bao W, Chen Y, et al. DeepLSTM: A bidirectional LSTM and CNN-based deep learning method for predicting lysine succinylation sites. Biomed Res Int 2021;2021:9923112. https://doi.org/10.1155/2021/9923112

Wang H, Zhao H, Yu L, Yang L. PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. Bioinformatics 2021;37:8027–33. https://doi.org/10.1093/bib/bbab146

Zuo Y, Li Y, Chen Y, Li G, Yan Z, Yang L. PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. Bioinformatics 2021;37:8027–33. https://doi.org/10.1093/bib/bbab146