Review

Artiffcial Intelligence in Post-Translational Modiffcation Site Prediction: Progress and Future Perspectives

Jiayi Ran , Xiaohan Zhang , Yunze Wang , Silin Chen , Jiaqi Deng , Martin Kosar , Bo Yang , Huiran Wang , Yishu Deng , Tailin Li , Yufei Liu , Lian Wang , Yijiang Guo , Xiaochen Zhang , Huaqiong Huang , Jianya Zhou , Jing Zheng , Zhihao Xu *, Yong Tang *, Jian Liu *

Posted Date: 31 March 2026

doi: 10.20944/preprints202603.2316.v1

Keywords: post-translational modification; artificial intelligence; machine learning; deep learning; proteomics

 doi:10.20944/preprints202603.2316.v1

Review

Artificial Intelligence in Post-Translational Modification Site Prediction: Progress and Future Perspectives

Jiayi Ran [1,2,†], Xiaohan Zhang [1,2,†], Yunze Wang [1,2,†], Silin Chen [1,2,†], Jiaqi Deng [1,2], Martin Kosar [3], Bo Yang [1,2], Huiran Wang [1,2], Yishu Deng [1,2], Tailin Li [1,2], Yufei Liu [1,2], Lian Wang [4], Yijiang Guo [1,2], Xiaochen Zhang [5], Huaqiong Huang [6], Jianya Zhou [7], Jing Zheng [7], Zhihao Xu [8,*], Yong Tang [9,*], Jian Liu [1,2,10,11,12,*]

1 Department of Respiratory and Critical Care Medicine, the Second Affiliated Hospital, and Centre for Infection Immunity and Cancer (IIC) of Zhejiang University-University of Edinburgh Institute (ZJU-UoE Institute), Zhejiang University School of Medicine, Zhejiang University, Hangzhou, 310029, China
2 Edinburgh Medical School: Biomedical Sciences, College of Medicine and Veterinary Medicine, The University of Edinburgh, Edinburgh, UK
3 Zhejiang University-University of Edinburgh Institute, Zhejiang University School of Medicine, Haining, 314400, China
4 Department of Thoracic Surgery, the Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou, China
5 Department of Medical Oncology, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China
6 Key Laboratory of Respiratory Disease of Zhejiang Province, Department of Respiratory and Critical Care Medicine, Second Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou, 310009, China
7 Department of Respiratory Disease, Thoracic Disease Center, The First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, China
8 Department of Respiratory and Critical Care Medicine, the Fourth Affiliated Hospital of School of Medicine, and International School of Medicine, International Institutes of Medicine, Zhejiang University, Yiwu, 322000, China
9 Department of Thoracic Surgery, Shenzhen Nanshan People's Hospital, Shenzhen, 518052, China
10 Key Laboratory of Cancer Prevention and Intervention, China National Ministry of Education, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China
11 Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence, Haining, China
12 Dr. Li Dak Sum & Yip Yio Chin Center for Stem Cell and Regenerative Medicine, Zhejiang University, Hangzhou, China
13 Biomedical and Health Translational Research Center of Zhejiang Provinc, Haining, China
* Correspondence: xuzhihao@zju.edu.cn (Z.X.); tangyong1213@email.szu.edu.cn (Y.T.); jianl@intl.zju.edu.cn (J.L.)
† These authors contributed equally to this work.

Abstract

Post-translational modifications (PTMs) are pivotal in modulating protein function and cellular processes. However, experimental identification of PTM sites remains costly and labor-intensive. Recent advances in artificial intelligence (AI) have empowered accurate and scalable in silico PTM site prediction from large-scale proteomic data. In this review, we provide a comprehensive and up-to-date overview of AI-driven PTM site prediction across more than ten PTM classes, covering single-PTM site prediction, multiple-PTM site prediction, inter-site crosstalk prediction, and functional prediction of modification sites. We systematically analyze and compare key AI frameworks, from conventional machine learning to deep learning, and summarize representative tools. We also identify key challenges and propose future directions for improvement. To foster application and ongoing progress, we provide practical guidelines for method selection and have established a dedicated website, which serves as a community benchmarking resource for the development of PTM site prediction tools. This website will be regularly updated with emerging prediction tools. By integrating comprehensive literature analysis with a dynamic online resource, we aim to provide a robust cornerstone for understanding current capabilities and guiding the future development of PTM site prediction tools, thereby promoting the integration of AI into practical biomedical research applications.

Keywords: post-translational modification; artificial intelligence; machine learning; deep learning; proteomics

 doi:10.20944/preprints202603.2316.v1

1. Introduction

Proteins are the cornerstone of cellular function, driving nearly all cellular activities. Post-translational modification (PTM) refers to the addition of functional groups to specific amino acid residues within a protein, which can alter protein structure, stability, and interactions, expanding the functional proteome beyond the information encoded by the genetic sequence [1]. The earliest discovery of PTM dates back to 1906, when Levene and Alsberg identified phosphate in the protein vitellin [2]. In 1932, phosphoserine was identified in vitellin, followed by the successive discovery of various distinct PTMs, including methylation, acetylation, and ubiquitination [3–5]. Over the decades, great progress has been made in unraveling the regulatory roles and functional implications of PTMs, guiding the exploration of protein regulation and cellular function at the molecular level.

PTMs can act as dynamic regulatory switches that maintain cellular homeostasis, influencing diverse biological processes such as signal transduction, gene regulation, and protein degradation [1,6,7]. They play a critical role in both health and disease and are increasingly recognized as valuable biomarkers and therapeutic targets. Traditional methods for PTM identification, such as enzyme-linked immunosorbent assay (ELISA) and Western blot (WB), rely on modification-specific antibodies and are limited by low sensitivity and poor coverage of unknown modification sites [8]. In contrast, MS enables high sensitivity and high throughput identification of multiple PTM types in a single experiment, supporting comprehensive characterization of large-scale PTM landscapes [9,10]. MS is the gold standard for PTM identification and a primary source for PTM databases. However, MS-based approaches remain time-consuming and costly. To overcome these limitations, statistical methods were developed to predict PTM sites based on sequence similarity and conserved motifs [11]. Although these approaches marked important advances, their limited adaptability and scalability restrict performance across diverse proteomic contexts, highlighting the continued need for rapid, cost-effective, and scalable PTM identification strategies.

Artificial intelligence (AI) has emerged as a powerful approach for PTM site prediction by learning complex patterns from large-scale datasets and enabling flexible prediction of PTM sites beyond conventional computational methods. AI-driven PTM site prediction has advanced rapidly, with continuous breakthroughs enhancing its accuracy and expanding its applications. Early machine learning-based tools, such as Musite, employed support vector machines (SVMs) to predict phosphorylation sites, demonstrating the feasibility of PTM site prediction beyond simple local sequence similarities [12]. The introduction of deep learning marked a major advance, with MusiteDeep leveraging convolutional neural networks (CNNs) to significantly improve prediction accuracy and efficiency [13]. More recently, transformer-based models, exemplified by TransPhos, enabled context-aware modeling of long-range sequence dependencies, further enhancing PTM site prediction performance [14]. In parallel, multi-label predictors such as iPTM-mLys expanded the scope of PTM prediction by enabling simultaneous identification of multiple modification types at the same site [15]. Such AI models have progressively refined PTM site prediction, facilitating robust large-scale proteomic analyses.

In this review, we summarize the recent advances in AI-driven PTM research with a focus on the commonly studied PTM types. We place special emphasis on the use of AI for PTM site prediction, including single- and multiple-PTM site predictions, and we review the use of AI in PTM crosstalk and functional prediction. We also discuss the current challenges in PTM site predictions, propose future improvements, and provide a method selection guideline to facilitate appropriate tool application. In addition, we have developed an interactive website[1] to track and update research articles on AI-based PTM site prediction models.

We aim to provide a comprehensive resource for researchers and facilitate the utilization of AI technologies in PTM research to support the development of more effective and targeted therapeutic strategies.

[1] Companion website: https://cerulean-melomakarona-91047f.netlify.app/. For brevity, its contents are not reiterated below.

 doi:10.20944/preprints202603.2316.v1

2. Overview of Post-Translational Modifications

2.1. Types of Post-Translational Modifications

PTMs are often reversible, enzyme-mediated modifications of proteins, although irreversible and non-enzymatic modifications also occur naturally. More than 650 PTM types have been identified so far [16]. Phosphorylation [17], acetylation [18], methylation [19], glycosylation [20], ubiquitination [21], SUMOylation [22], succinylation [23], and crotonylation [24] are frequently studied PTMs due to their critical roles in cellular regulation (Figure 1). In addition, emerging PTMs such as lactylation [25], malonylation [26], and neddylation [27] are increasingly recognized as important regulators of cellular functions.

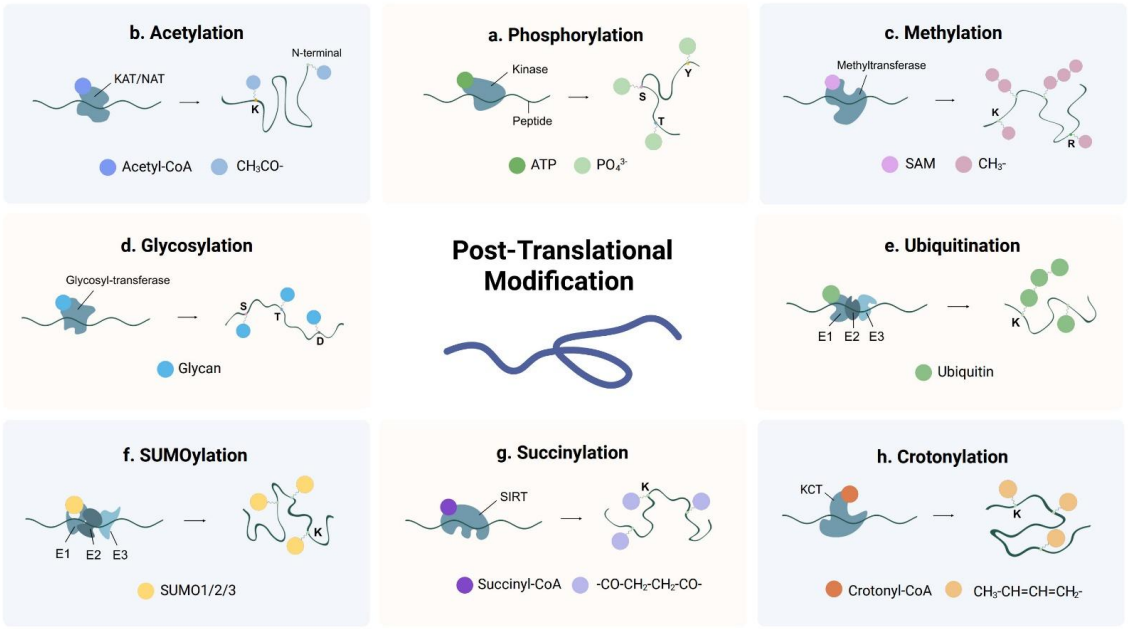

Figure 1. Schematic illustration of eight common types of PTMs. a. Phosphorylation. ATP serves as the phosphate donor, and its phosphate group is transferred to serine, threonine, and tyrosine residues on the target peptide, mediated by kinases. b. Acetylation. The acetyl group (CH_3CO-) from acetyl-CoA is transferred to lysine residues or the N-termini of the peptides, mediated by KATs and NATs, respectively. c. Methylation. Methyl groups (CH_3-) from the donor SAM are transferred to lysine or arginine residues, mediated by protein methyltransferases. Lysine residues can be mono-, di-, or tri-methylated, while arginine residues undergo mono- or di-methylation. d. Glycosylation. Glycans are attached to asparagine (N-linked) or serine/threonine (O-linked) residues, mediated by glycosyltransferases. e. Ubiquitination. A 76-amino acid protein, ubiquitin, is attached to lysine residues of a substrate protein via its C-terminal glycine, sequentially catalyzed by the ubiquitin-activating enzyme (E1), the ubiquitin-conjugating enzyme (E2), and the ubiquitin ligase (E3). f. SUMOylation. A SUMO protein is attached to the lysine residues of the substrate protein under the catalysis of the SUMO-activating enzyme (E1), the SUMO-conjugating enzyme (E2), and the SUMO ligase (E3). g.Succinylation. The succinyl group (-CO-CH_2-CH_2-CO-) from succinyl-CoA is transferred to lysine residues, regulated by SIRT proteins. h.Crotonylation. The crotonyl group (CH_3-CH=CH-CO-) from crotonyl-CoA is transferred to lysine residues, mediated by KCTs. PTM, post-translational modification; ATP, adenosine triphosphate; KAT, lysine acetyltransferase; NAT, N-terminal acetyltransferase; SAM, S-adenosylmethionine; SUMO, small ubiquitin-related modifier; SIRT, silent information regulator; KCT, lysine crotonyltransferase. This figure was created with BioRender.

At the protein level, PTMs regulate conformational dynamics, subcellular localization, and protein–protein interactions, thereby coordinating key cellular processes such as cell cycle progression, metabolism, and apoptosis. Dysregulated PTMs contribute to numerous diseases. For example, tau hyperphosphorylation promotes protein aggregation in Alzheimer's disease [28], while lactylation of ABCF1 at Lys430 has been reported to drive hepatocellular carcinoma progression [29]. These

 doi:10.20944/preprints202603.2316.v1

findings highlight the importance of accurate, site-specific PTM identification for elucidating molecular mechanisms and advancing early diagnosis and precision medicine.

2.2. PTM Databases

PTM databases typically serve as the basis for training PTM site prediction models. Based on the gathered literature, we summarize several commonly used databases, including UniProt, the Protein Data Bank, PhosphoSitePlus, Phospho.ELM, and dbPTM. Table 1 summarizes their key features. These databases provide important information on protein sequence and structure, and on PTM types and sites.

Table 1. Six commonly used PTM databases.

Database	Official Website	Year	Data Sources	Data Scope	Entry Number	Annotations
UniProt	uniprot.org	2005	Experimental data, literature, computational predictions	Protein sequences	574,627	Protein function, domains, variants, PTMs, subcellular localization, biological pathway
PDB	rcsb.org	1971	Experimental data	Protein 3D structures	249,018	Polymer sequences, structural domains, subcellular localization, protein function
PSP	phosphosite.org	2008	Experimental data, literature	Multiple PTMs	606,322	Topology, function, biological pathway
Phospho.ELM	phospho.elm.eu.org	2010	Experimental data, literature	Phosphorylation sites in eukaryotes	42,574	Interaction, domains, kinases
dbPTM	dbPTM	2006	Experimental data, literature	Multiple PTMs	2,845,259	Interaction, function, biological pathway, disease association
CPLM	cplm.biocuckoo.cn	2011	Literature, public databases	Multiple PTMs on lysine residues	592,606	Variants, domains, interaction, function, subcellular localization, biological pathway, physicochemical property

PTM, post-translational modification; PDB, Protein Data Bank; PSP, PhosphoSitePlus; CPLM, Compendium of Protein Lysine Modifications.

2.2.1. UniProt

UniProt provides comprehensive, high-quality, and easily accessible data on protein sequences, structure, and genomic information. As of February 2026, it contains 574,627 sequence entries in UniProtKB/Swiss-Prot, the central knowledge base of UniProt. As a protein database with extensive content and rich annotation, UniProt is recognized by the scientific community and designated as the Global Core Biodata Resource [30]. Among the PTM site prediction models mentioned in this review, UniProt is widely selected as the reference database. Its official website is available online at https://www.uniprot.org/.

2.2.2. Protein Data Bank

Protein Data Bank (PDB) is the oldest database among the five mentioned, having operated consecutively for 54 years with a specialization in providing three-dimensional protein structural data. Since its establishment, PDB has improved in both capacity and accuracy. Its structural information is largely validated by experimental methods, typically macromolecular X-ray crystallography [31]. As of November 2025, it houses > 240,000 biostructures [32]. It allows users to view the overall protein structure and related information and to zoom in to explore specific regions, ligands, and residues in detail [33], offering substantial value for educators and researchers worldwide. The official website of the Protein Data Bank is https://www.rcsb.org/.

2.2.3. PhosphoSitePlus

PhosphoSitePlus (PSP) is a large database covering a wide range of PTM types, including phosphorylation, acetylation, methylation, and other modifications. Version 6.8.2 contains information on 59,791 proteins and 606,322 PTMs of multiple types. Among these sites, over 90% are validated via

 doi:10.20944/preprints202603.2316.v1

high-throughput MS/MS methods [34]. PSP also provides information on genetic variants associated with polymorphisms and disease pathogenicity [35]. The data can be obtained from its official website: https://www.phosphosite.org/.

2.2.4. Phospho.ELM

Since its inception in 2004, Phospho.ELM has served as a comprehensive database providing phosphorylation sites on serine, threonine, tyrosine, and other residues based on the primary literature [36]. This database contains phosphorylation information for eukaryotes, encompassing data obtained from both in vivo and in vitro studies. The current version, also its ninth version (version 9), was updated in 2010. According to the latest available statistics, Phospho.ELM contains 42,574 non-redundant phosphorylation residues of various types and provides more than 11,000 protein sequences containing these PTM sites. A conservation score is available to estimate reliability, and structural information is implemented in this version [37]. It is a free database available through: http://phospho.elm.eu.org/.

2.2.5. dbPTM

As a comprehensive, interactive, and user-friendly database with a special focus on cancer proteomics, dbPTM provides 2,845,259 PTM site records, most of which are experimentally validated and supported by research articles. The 2025 version of dbPTM integrates data from multiple sources, including GPS6.0, Phospho.ELM, PhosphoSitePlus, UniProtKB, and the PDB. Additionally, it incorporates phosphoproteomic data from 13 cancer types to reveal kinase activity in cancer [38]. The website of dbPTM can be accessed at https://awi.cuhk.edu.cn/dbPTM/.

2.2.6. Compendium of Protein Lysine Modifications

Positively charged lysine residues are among the most important amino acid residues that undergo modification. The Compendium of Protein Lysine Modifications (CPLM) offers various types of PTM information with a special focus on lysine sites [39]. Since its initial establishment in 2011, it has been updated across versions CPLA 1.0, CPLM 2.0, and PLMD 3.0. Currently, the fourth version, released in 2017, is available. As of February 2025, it holds 592,606 modifications spanning 29 PLM categories in 219 species. Additionally, it provides researchers with access to annotations and references for further investigation. The official website of CPLM can be accessed at https://cplm.biocuckoo.cn/.

3. Framework for AI-Based PTM Site Prediction Tools

The workflow is illustrated in Figure 2. In practice, end-users supply an input amino acid sequence, which a pre-trained AI model processes to generate predictions of potential PTM sites. Model training begins with the collection of benchmark datasets, consisting of protein sequences with validated modification sites, wherein each residue is labeled as either positive (modified) or negative (unmodified). To enhance predictive accuracy, raw sequence data are often supplemented with structural features. The integrated sequence and structural information are then encoded into numerical feature vectors (embeddings) that capture diverse characteristics of each residue (Table 2) [40–43]. These embeddings serve as inputs for training classification models, encompassing both conventional machine learning algorithms and advanced deep learning architectures.

Conventional machine learning algorithms primarily include linear models (e.g., penalized logistic regression, PLR) [44], tree-based models (e.g., random forest, RF) [45], kernel-based methods (e.g., SVM) [46], and probabilistic frameworks (e.g., Bayesian network) [47]. While these methods show considerable promise in PTM site prediction, their reliance on manually designed features may introduce intrinsic bias. Deep learning architectures leverage multi-layered neural networks, including multi-layer perceptrons (MLPs) [48], CNNs [49], recurrent neural networks (RNNs) [50], graph neural networks (GNNs) [51], and emerging transformer-based models [52]. These models automatically learn feature representations from raw data through successive layers of abstraction, which gives them

 doi:10.20944/preprints202603.2316.v1

an advantage over conventional machine learning methods in processing large-scale, high-dimensional data with minimal manual intervention.

After training, the model can analyze novel protein sequences, assigning a probability score or categorical label to each residue to indicate its predicted propensity for PTM.

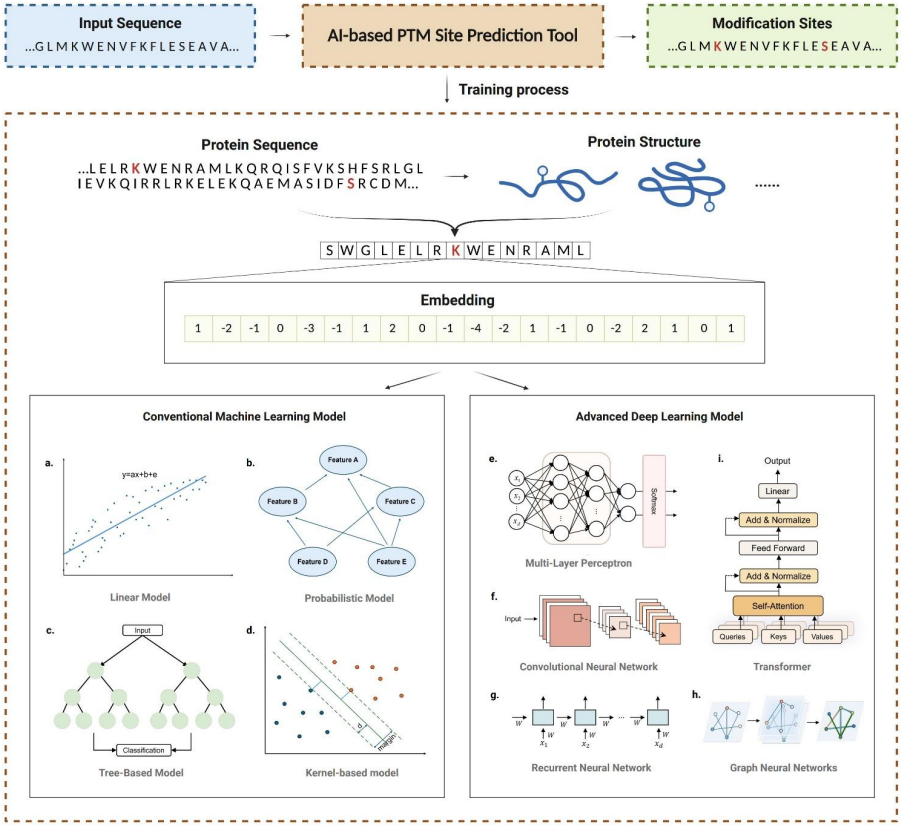

Figure 2. Schematic of an AI-based PTM site prediction workflow. Users provide an amino acid sequence as input to the trained PTM site prediction tool, which outputs predicted modification sites. Model training involves data collection (protein sequence and structural information), embedding of sequence and structural features, and model learning using conventional machine learning (a-d) and deep learning (e-i) architectures. a. Linear model (penalized logistic regression). b. Probabilistic model (Bayesian network). c. Tree-based model (random forest). d. Kernel-based model (support vector machine). e. Multi-layer perceptron. f. Convolutional neural network. g. Recurrent neural network. h. Graph neural network. i. Transformer. AI, artificial intelligence; PTM, post-translational modification. This figure was created with BioRender.

Table 2. Feature extraction methods in PTM site prediction.

Feature Extraction Method	Feature Type	Description
One-hot encoding	Sequence-based	Represents each amino acid as a 20-dimensional binary vector with "1" assigned to the position of the corresponding amino acid and "0" to the rest.
AAC	Sequence-based	Represents a protein sequence by the frequency of each of the 20 amino acids and encodes the information into a 20-dimensional vector.
CKSAAP	Sequence-based	Calculates the frequency of amino acid pairs separated by a distance K and represents each pair frequency as an entry in the feature vector.
AAIndex	Physicochemical Property-based	Involves a set of indices that quantify physicochemical and biochemical properties of amino acids.
PseAAC	Physicochemical Property-based	Transforms protein sequences into numerical vectors by incorporating both the properties of amino acids and the sequence order of the protein.
ASA	Structure-based	Calculates the surface area of a protein accessible to solvent molecules to reflect protein folding, stability, and interactions with other molecules.
PSSM	Evolution-based	Represents a protein sequence of length L as an L×20 matrix, with each row indicating position-specific substitution scores for the 20 amino acids occurring at the corresponding position.

PTM, post-translational modification; AAC, amino acid composition; CKSAAP, composition of k-spaced amino acid pairs; AAIndex, amino acid index; PseAAC, pseudo amino acid composition; ASA, accessible surface area; PSSM, position-specific scoring matrix-based transformation.

 doi:10.20944/preprints202603.2316.v1

4. Single-PTM Site Prediction

Single-PTM site prediction focuses on identifying specific modification types at individual residues within a given sequence. By integrating diverse features with AI-based frameworks, these tools have significantly advanced the discovery of PTM sites. This section highlights methodological developments and representative tools.

To provide more comprehensive insights, we established a website that curates information including databases, sample sizes, model architectures, and performance metrics across a broader range of tools (Figure 3). This website is a PTM site prediction tool library that enables users to search and filter tools based on multiple criteria, and to visually compare their performance across multiple evaluation metrics using radar charts. By integrating prediction tools for a wide range of PTMs, the website serves not only as a comprehensive resource for exploration but also enables systematic performance comparisons that support the development of new tools. In addition, we have summarized the advantages and limitations of representative AI-based PTM site prediction tools in Supplementary Table S1 to help researchers better understand the methodological trade-offs among existing approaches. This table complements the website, offering a more direct reference for method comparison across different PTM types.

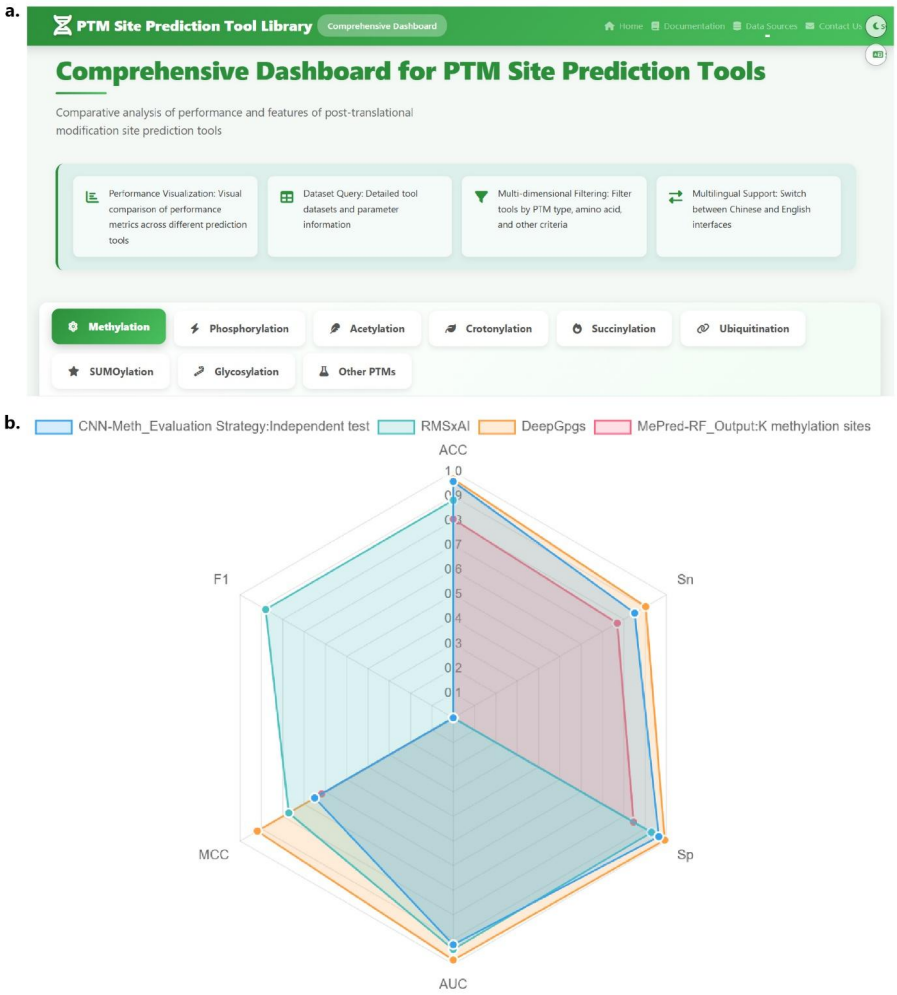

Figure 3. Website interface and performance visualization of the PTM site prediction tool library. The comprehensive dashboard allows users to compare PTM site prediction tools through performance visualization, dataset queries, and multidimensional filtering. It provides access to prediction tools for diverse PTM types, including phosphorylation, acetylation, methylation, glycosylation, ubiquitination, SUMOylation, succinylation, crotonylation, and other novel and uncommon PTMs. a. Interface of the website. b. Example radar chart for visualizing the performance comparison of selected tools. PTM, post-translational modification.

 doi:10.20944/preprints202603.2316.v1

4.1. Phosphorylation Site Prediction

Phosphorylation site prediction tools have evolved over time from simple sequence-based approaches to increasingly complex models. The first application of AI in this field dates back to 1999, when Blom et al. developed NetPhos, an artificial neural network (ANN)-based tool trained on sequence and tertiary structure information to predict serine, threonine, and tyrosine phosphorylation sites [53]. NetPhos laid the foundation for computational prediction of phosphorylation sites. However, its lack of kinase-specific modeling limits the biological interpretability of kinase-substrate interactions. To address this limitation, NetPhosK [54], introduced in 2004, employs a kinase-specific neural network trained on protein kinase A (PKA) phosphorylation datasets, marking an important advancement toward more biologically relevant predictions.

Subsequent studies have continuously improved in phosphorylation site prediction by optimizing algorithms, applying diverse feature representations, and implementing various strategies. MusiteDeep [13] is the first deep learning-based approach for phosphorylation site prediction. It employs a multi-layer CNN with an attention mechanism to automatically learn complex sequence patterns, achieving substantially higher area under the receiver operating characteristic curve (AUC) and mean precision than Musite. DeepPhos [55] also uses CNNs to extract hierarchical sequence features. Unlike MusiteDeep, it leverages densely connected CNN blocks to enhance information flow and to integrate different levels of extracted representations. DeepPhos reportedly outperformed MusiteDeep in both general and kinase-specific phosphorylation site prediction. TransPhos [14] integrates transformer and CNN architectures. The transformer effectively captures long-range dependencies in protein sequences, while the CNN extracts rich, high-dimensional representations. This combination enables more accurate prediction of phosphorylation sites, resulting in reported superior performance compared to many previous models. However, prediction accuracy for tyrosine sites remains relatively lower due to the limited availability of Y-specific training data. Transfer learning from better-represented site types may help improve Y-site prediction.

More recently, PhosHSGN [56], designed based on CNNs and GNNs, enhances prediction by combining sequential and spatial information. This compensates for the limitation of convolutional structures, whose receptive fields are typically restricted to local windows, and enables the incorporation of global protein features and the interplay with other PTMs, achieving an accuracy of 91.71%. Continuing this trend toward multi-modal and advanced feature integration, CaLMPhosKAN [57] introduced a novel fusion approach for general phosphorylation site prediction. It combines codon-aware embeddings (from a codon adaptation language model) with traditional amino acid-aware embeddings, processed via a wavelet-based Kolmogorov-Arnold Network (KAN). This codon-level information captures subtle sequence biases and evolutionary signals often missed by amino-acid-only models, leading to improved predictive accuracy over prior sequence-based tools, particularly in handling noisy or imbalanced datasets.

Recent work has also explored advances in kinase-specific prediction, represented by DCPPS [58]. DCPPS uses the Dynamic Embedding Encoding (DEE) to dynamically capture both semantic and positional information, producing more discriminative features than static embeddings. It further integrates a Cross-Representation Interaction Unit (CRIU) that leverages multi-head cross-attention to effectively combine local and global features for better modeling of kinase-substrate specificity. These designs enable DCPPS to achieve higher predictive performance and scalability, especially on imbalanced kinase-specific datasets.

4.2. Acetylation Site Prediction

One of the earliest AI-based acetylation site prediction tools is NetAcet [59], which was developed using an ANN architecture. In 2010, an ensemble SVM-based algorithm was introduced in EnsemblePail [60], outperforming single SVM classifiers. Eight years later, DeepAce [61] was reported as the first deep learning-based acetylation site predictor. It is built on CNNs and deep neural networks (DNNs) and trained on multi-species datasets to support both general and species-specific predictions.

 doi:10.20944/preprints202603.2316.v1

However, its precision on human and mouse datasets is lower due to data imbalance, and it remains susceptible to overfitting when applied to small datasets.

Several recent models incorporate advanced deep learning innovations. MSTL-Kace [62], developed in 2023, represents a notable advancement by introducing a pre-trained BERT model to generate high-level embeddings. In addition, it employs a two-stage fine-tuning strategy to address limited training data. The model is first fine-tuned on species with larger datasets and subsequently adapted to species with smaller sample sizes. This strategy resulted in substantial performance gains compared with previous models such as DNNAce.

SIPSC-Kac [63], introduced in 2024, covers six prokaryotic species. It integrates AlphaFold-predicted structural information and a swarm intelligence algorithm for feature selection, enabling improved predictive performance, particularly under class imbalance in prokaryotic datasets. Despite these advances, SIPSC-Kac is trained on prokaryotes, limiting applicability to eukaryotic proteins. Another notable tool, TransPTM [64], integrates ProtT5 language model embeddings with GNNs for feature extraction and classification. Notably, it explores acetylation in non-histone proteins, helping to expand available benchmark datasets. DeepCBA [65] employs a hybrid framework combining CNN and Bi-LSTM architectures to capture both local and long-range sequence patterns, together with an attention mechanism to enhance feature weighting. Protein sequences are encoded as tripeptide word vectors generated using the word2vec skip-gram model, improving discriminative capacity. These architectural innovations led to an accuracy of 80.51% and an AUC of 87.36% on independent test sets, indicating strong performance relative to existing predictors.

In 2025, MDDeep-Ace [66] was developed as a multi-species acetylation predictor. It combines CNN-LSTM with the Kullback-Leibler (KL) divergence-based dynamic domain adaptation to address data scarcity in low-sample species, achieving an average AUC of 0.790 and robust cross-species performance.

4.3. Methylation Site Prediction

Protein methylation site prediction has advanced significantly with the progress of AI algorithms, each employing distinct strategies to enhance accuracy and applicability.

Early tools rely on conventional machine learning algorithms. MePred-RF [67] is an RF-based predictor that identifies arginine and lysine methylation sites from sequence information. It integrates six sequence-based feature descriptors and applies the maximal relevance-maximal distance (MRMD) feature selection technique to capture discriminative information effectively. Met-predictor [68], based on SVM, takes a step further by integrating protein tertiary structure information. Notably, Met-predictor also enables prediction of different methylation types (i.e., mono-, di-, and tri-methylation), enabling more precise biological interpretations and downstream analyses.

The emergence of deep learning-based tools marked a major leap. For example, DeepRMethylSite [69] combines CNNs for hierarchical local features and LSTMs for long-range dependencies. This ensemble model achieved improved results with an AUC of 0.82. CNNArginineMe [70] uses pure CNNs. However, by constructing a larger training dataset and applying an early stop strategy to avoid overfitting, it achieved a higher AUC (0.87) than DeepRMethylSite. A key limitation lies in its reliance on one-hot encoding and fixed-window convolution, which restricts the capture of long-range dependencies. This makes independent protein-level testing particularly important. Another predictor, DeepGpgs [71], combines ResNet and LSTM. DeepGpgs introduces two key innovations: a fused Gaussian prior layer for extracting locus information and a gated multi-head attention layer for capturing global sequence context. DeepGpgs not only outperformed other methods in arginine methylation site prediction but also performed strongly in predicting phosphorylation sites, demonstrating strong scalability. A notable recent shift emphasizes interpretability. RMSxAI [72], developed in 2024, applies explainable AI (XAI) techniques to trace model decisions, which enables researchers to better assess and refine predictions. DeepTESite [73] is a recent advanced framework for predicting arginine methylation sites. It integrates a transformer architecture with novel symmetric positional encodings, inspired by the symmetry of arginine methylation. This biologically grounded approach enhances

 doi:10.20944/preprints202603.2316.v1

predictive performance by providing richer, context-aware representations, achieving a competitive accuracy of 0.81.

4.4. Glycosylation Site Prediction

In early computational glycosylation research, NetOglyc [74] pioneered O-linked site prediction using ANNs. It achieved solid performance and laid the groundwork for the field. Building on the early studies, later tools advanced with better algorithms and feature processing. GlycoMine [75] employs RF with minimum redundancy-maximum relevance (mRMR) and information gain (IG)-based feature selection method to identify the most informative features. GlycoMine can predict N-, C-, and O-linked sites and attained an AUC > 0.98 for N- and O-linked glycosylation, outperforming NetNGlyc. Notably, GlycoMine is explicitly developed for the human proteome. Extending it to other organisms typically requires retraining on organism-specific datasets.

SPRINT-Gly [76] improves prediction accuracy by leveraging a larger training dataset. It was trained on the largest N- and O-linked glycosylation site dataset available at the time, which integrated multiple sources with sequence and predicted structural data. This resulted in substantially improved Matthews correlation coefficient (MCC) scores compared to previous methods, reflecting greater robustness.

Deep learning marked further progress. DeepNGlyPred [77] combines sequence, structural, and evolutionary features for human N-linked prediction. Based on this foundation, LMNglyPred [78] advances feature extraction by incorporating embeddings derived from the ProtT5 and achieves outstanding predictive performance. More innovative architectures include DOGpred [79], which employs a 1D-CNN and an attention-based bidirectional GRU (Bi-GRU) network to process spatial and temporal features, respectively. This combination provides the model with superior predictive performance with an AUC of 0.939. A recent work by Hong et al. [80] further advanced O-glycosylation site prediction. They developed a predictor that captures spatial information via local environmental features, combined with sparse recurrent neural networks (SRNN) for sequential modeling and interpretability. By better accounting for spatial context and key influencing factors, this method achieved a reported 1.4-fold F1 score improvement over prior models. For C-linked sites, DeepCSEmbed-C [81] adopts a dual-branch deep learning architecture: an Inception branch processes word embeddings, while a fully connected neural network (FNN) branch handles ProtT5 and evolutionary-scale modeling (ESM) embeddings. DeepCSEmbed-C further applies recursive feature elimination (RFE) and particle swarm optimization (PSO) for effective feature selection, delivering high accuracy (0.952) for glycosylation site prediction.

4.5. Ubiquitination Site Prediction

AI-based ubiquitination site prediction emerged in the late 2000s, with several tools establishing a foundational framework. One of the earliest tools, UbPred [82], uses an RF classifier trained on experimentally verified sites, achieving an AUC of 0.8. Around the same time, CKSAAP_UbSite [83], which employs SVM with composition of k-spaced amino acid pairs (CKSAAP), binary encoding (BE), amino acid index (AAindex), and aggregation propensity features, achieved 1.4% higher accuracy than UbPred. Both tools, however, were not optimized for human-specific prediction, highlighting the need for human-specific predictors.

ESA-UbiSite [84] focuses on human-specific prediction. It combines SVM with an evolutionary screening algorithm (ESA) to incrementally identify effective negative samples from non-validated sites, addressing the common issue of unreliable negatives in PTM datasets. UbiSitePred [85], also based on SVM, improves prediction accuracy by applying the least absolute shrinkage and selection operator (LASSO) for feature selection to eliminate redundancy. UbiSitePred achieved a very high AUC score (reported up to 0.9998) in five-fold cross-validation. However, cross-validation on potentially redundant benchmarks can overestimate true predictive power, leading to such exceptionally high performance. Rigorous evaluation on strictly independent test sets is therefore critical to confirm real-world generalization.

 doi:10.20944/preprints202603.2316.v1

The advent of deep learning has further revolutionized the field. DeepUbi [86], the first deep learning-based predictor, integrates CNNs with 31 physicochemical properties and sequence features. DeepUbi outperformed earlier machine learning-based tools, including UbiPred and CKSAAP_UbSite, demonstrating the advanced capability of deep learning algorithms. Another deep learning-based tool, Caps-Ubi [87], utilizes capsule networks to capture complex amino acid interactions. Capsule networks offer benefits over CNNs by internally modeling hierarchical relationships, resulting in a 2.36% higher accuracy for Caps-Ubi compared to DeepUbi.

Tools focusing on enzyme-specific aspects have also emerged. GPS-Uber [88] combines PLR, DNN, and CNN. Notably, it utilizes transfer learning from general ubiquitination sites to predict E3-specific lysine ubiquitination sites, offering enhanced insights into enzyme-substrate interactions.

4.6. SUMOylation Site Prediction

Various computational tools have been developed over the years to predict SUMOylation sites. Early SUMOylation predictors, such as SUMOhydro (2012) [89], represent an advancement over the mere identification of consensus motifs in earlier studies. SUMOhydro incorporates amino acid hydrophobicity into its binary encoding scheme and utilizes an SVM for classification, outperforming traditional statistical approaches. In 2014, GPS-SUMO [90] was introduced based on the PSO algorithm. Notably, the input data of GPS-SUMO include SUMOylation sites and SUMO-interaction motifs (SIMs). This expands its predictive scope and provides valuable insights into the mechanisms of SUMOylation. Since then, the prediction of protein SUMOylation sites has progressed substantially, moving from conventional machine learning models to advanced deep learning approaches.

Representative machine learning-based tools include SUMOgo [91] and SUMO-Forest [92]. SUMOgo [91] is an RF-based predictor. It integrates sequence features and predicted secondary structure, using a library for SVM (LIBSVM) and mRMR feature selection methods to evaluate feature importance. This design allows the model to optimize prediction accuracy and reduce overfitting based on the ranking. SUMO-Forest [92] employs a cascade forest architecture and adopts genetic algorithm-based weighting to address class imbalance. As a result, it achieves an impressive AUC of 98.38% on a heavily imbalanced dataset (positive: negative = 1:13).

In recent years, deep learning approaches have dominated the field. In 2023 and 2024, Salman Khan et al. consecutively developed two deep learning-based tools: Deep-SUMO [93] and PSSM-Sumo [94]. Deep-SUMO [93] utilizes a multilayer DNN with half-sphere exposure (HSE)-based features and principal component analysis (PCA)-based feature selection, achieving an average accuracy of 96.47% in a 10-fold cross-validation. PSSM-Sumo [94] builds on and improves upon Deep-SUMO by applying PsePSSM and sequential forward selection using an SVM (SFS-SVM) for feature identification. This results in superior performance on the same validation dataset with an accuracy of 98.71%. Despite strong performance, Deep-SUMO and PSSM-Sumo are not fully end-to-end. This increases dependence on feature-generation steps and raises the computational burden for proteome-wide applications. GPS-SUMO 2.0 [95], an updated version of GPS-SUMO, integrates transformer, DNN, and PLR architectures. It adopts a two-stage training process involving pre-training on large-scale protein data and transfer learning on SUMOylation-specific datasets, significantly enhancing prediction performance. SUMO-LMNet [96] is a new deep learning-based predictor specifically designed for the identification of SUMO1 and SUMO2 modification sites. It employs a CNN-based lossless mapping network to capture local and global sequence dependencies, and adopts combined heatmap feature analysis (CHFA)to improve model interpretability. These advances demonstrate the growing role of deep learning in improving the predictive performance for SUMOylation, paving the way for more effective tools.

4.7. Succinylation Site Prediction

As a relatively novel focus of PTM site prediction, the potential of succinylation-related studies is gradually being discovered and explored. The earliest model for succinylation site prediction, iSuc-PseAAC [97], was published in 2015. Constructed upon the conventional machine learning model SVM, it reached an accuracy of 79.94%. PSuccE [98], published in 2018, further adopted an ensemble

 doi:10.20944/preprints202603.2316.v1

SVM strategy, improving accuracy by 9.20% compared with iSuc-PseAAC. One year later, CNN-SuccSite [99] introduced a CNN-based deep learning architecture. It outperformed earlier machine learning models and marked a shift toward deep learning approaches in this field.

Since 2020, AI-based succinylation prediction has expanded rapidly, marked by the release of several high-performing tools. One example is HybridSucc [100], which combines DNN and PLR for modeling ten types of features to improve prediction accuracy. Notably, HybridSucc performs particularly well in human-specific predictions, demonstrating its potential utility in studying human diseases. Deep-KSucc [101] and pSuc-EDBAM [102] were published in 2022, both adapting the CNN-based architecture for prediction tasks. Notably, Deep-KSucc has a parallel structure that combines a dense CNN with ordered-neuron long short-term memory, while pSuc-EDBAM was developed based on ensemble dense blocks. pSuc-FFSEA [103] is another model proposed by the same authors as pSuc-EDBAM. It is the first to use a broad learning system (BLS) within a stacking ensemble framework. This design leads to high robustness and accuracy, with an AUC of 0.85.

In 2023, Ahmed et al. introduced CBL_BLC [104], a pioneering hybrid architecture that integrates a CNN+Bi-LSTM (CBL) and a Bi-LSTM+CNN (BLC). The features from both branches are concatenated and processed through two densely connected layers, enabling the capture of both local and long-range dependencies. CBiLSuccSite [105] refined this by combining CNN and Bi-LSTM with word embedding techniques. Unlike CBL_BLC, which relies on one-hot encoding, CBiLSuccSite leverages word embeddings to automatically extract features from raw input data. This enables the model to more effectively capture intricate dependencies within protein sequences and improves its predictive accuracy to 76%, surpassing that of CBL_BLC. A limitation of CBiLSuccSite is its sequence-only design, which may inadequately capture spatial dependencies that influence succinylation. To address this limitation, iSucc-SnCNs [106] integrats ProtGPT2-based protein language model embeddings with a deep capsule network to process structural and evolutionary representations. This combination enhanced contextual and structural feature extraction, resulting in a 92.92% accuracy and a 0.96 AUC on independent test sets.

4.8. Crotonylation Site Prediction

Crotonylation site prediction has emerged as a developing field in recent years. The earliest study dates back to 2016, when Huang and Zeng proposed CrotPred, a crotonylation predictor based on a discrete hidden Markov model [107]. In leave-one-out cross-validation, CrotPred achieved an accuracy of 0.7823, marking an impressive early attempt in crotonylation site prediction. In 2017, CKSAAP_CrotSite [108] introduced CKSAAP features within an SVM framework, and achieved an accuracy of 98.11%, far surpassing previous tools. Nevertheless, such high performance may reflect overfitting to the training data, as its generalization to independent datasets remains insufficiently validated.

A transition toward deep learning began with Deep-Kcr [109], which is built on a CNN architecture. It integrates sequence, physicochemical, and numerical space features, and uses information gain for feature selection to eliminate redundant features. Deep-Kcr demonstrated excellent prediction performance, surpassing former models such as CKSAAP_CrotSite. DeepCap-Kcr [110] further incorporated CNN, LSTM, and capsule layers to capture hierarchical features more effectively. In independent tests, it outperformed Deep-Kcr, highlighting the advantages of capsule-based modeling.

More recent advances emphasize the integration of more advanced architectures. ILYCROsite [111] combines fuzzy clustering-means (FCM) clustering with generalized regression neural network (GRNN) for undersampling to handle imbalanced datasets. This enhances both predictive performance and generalization ability. LMCrot [112] incorporates ProtT5-derived window-level embeddings. These embeddings are fine-tuned via a residual convolutional Bi-LSTM layer and processed by two parallel architectures: a CNN module for capturing sequence features, and a DNN module for representing physicochemical properties. This multi-branch ensemble design yields improved predictive performance compared with single-representation models. In 2025, DeepMM-Kcr [113] was introduced as a cutting-edge deep learning model for lysine crotonylation site prediction. Its key innovation is the

 doi:10.20944/preprints202603.2316.v1

fusion of transformer-based natural language processing features and traditional hand-crafted features through a multi-head self-attention mechanism. This approach enhances performance by dynamically prioritizing complementary features. On an independent test set, DeepMM-Kcr achieved an AUC of 0.931.

4.9. Novel and Uncommon PTM Site Prediction

Advances in PTM site prediction are reflected not only in improved prediction accuracy but also in the expansion of research scope to encompass newly discovered or less-studied PTMs, such as lactylation, malonylation, and β-hydroxybutyrylation.

AI-based predictors for lactylation have emerged only in the past few years, but advances in AI-based frameworks have rapidly led to high-performance tools. As a pioneering tool, FSL-Kla [114] integrates eight sequence-based and three structure-based features, and employs PCA to mitigate overfitting, demonstrating strong performance on small lactylation datasets. DeepKla [115] introduced a more sophisticated feature extraction pipeline by combining CNN, Bi-GRU, and attention mechanisms. This enables more effective representation of contextual dependencies, contributing to its superior performance (AUC: 0.9722). In 2024, Yang et al. introduced two advanced predictors, ABFF-Kla and EBFF-Kla, representing a significant leap in modeling protein information [116]. A key innovation of their design is the automatic extraction of features from the AlphaFold-predicted protein structure, where ABFF-Kla utilizes attention-based feature fusion and EBFF-Kla employs embedding-based fusion to capture important features.

Malsite-Deep is one of the representative predictors for malonylation sites [117]. It addresses class imbalance using the under-sampling NearMiss-2 method, and a gated recurrent units layer is used to select the optimal feature subset. A DNN then processes these features to make the final prediction. As validated by 10-fold cross-validation and independent testing, this model achieves an AUC value above 0.95.

Another notable tool is SLAM [118], introduced in 2024. SLAM is the first deep-learning predictor specifically designed for lysine β-hydroxybutyrylation. It employs a multi-track encoder that captures both sequence- and structure-derived relationships. The sequence encoder is a hybrid encoder that combines Bi-LSTM, CNN, and a pre-trained protein language model, while the structure encoder leverages GNN. The fused representations are processed through an attention-based decoder and an MLP classifier, enabling accurate prediction of β-hydroxybutyrylation sites. Additionally, a 2026 study introduced BiGKbhb [119], which employs a Bi-GRU architecture based on BLOSUM62 evolutionary encoding to capture contextual information in protein sequences. This model demonstrated strong predictive performance and cross-species transferability on both cross-validation and independent test sets.

In addition to the models described above, other predictors for novel and uncommon PTMs developed since 2021 are summarized on our website, covering modifications such as S-nitrosylation, S-sulfenylation, and formylation that were not previously discussed. The broad range of target types and the increasing diversity of available models underscore the growing potential of computational PTM site prediction moving forward.

5. Multiple-PTM Site Prediction

The above tools focus on predicting a single type of PTM on specific sites of a protein. However, in biological systems, proteins undergo multiple PTMs simultaneously, and these modifications can influence each other to coordinate protein function. Advancements in multiple-PTM site prediction have enabled the simultaneous prediction of different types of PTMs at the same time, providing a more holistic view of the modification landscape.

One of the earliest AI-based tools for multiple-PTM site prediction, ModPred [120], is based on the multiple binary classification (MBC) strategy. MBC involves decomposing a complex classification problem into multiple binary classification tasks, each focusing on predicting one type of PTM [121]. ModPred predicts 23 types of PTMs, with each PTM type being predicted independently using

 doi:10.20944/preprints202603.2316.v1

separate logistic regression models. PTM-ssMP [122] follows a similar framework but employs SVM classifiers and introduces a site-specific modification profile (ssMP) feature, which allows effective extraction, encoding, and integration of local sequence context and proximal PTM information. As a result, PTM-ssMP has demonstrated strong performance across various PTM types. MusiteDeep [123] further advances this paradigm using deep learning. Each predictor within MusiteDeep integrates MultiCNN and CapsNet architectures to generate confidence scores for PTM site predictions separately. These scores are then averaged to provide the final prediction score. Notably, MusiteDeep provides visualization of predicted PTM sites in protein three-dimensional structures, enabling spatial analysis and potential crosstalk exploration.

While the MBC strategy can achieve high accuracy for individual PTM types, it inherently disregards dependencies across PTM types because each classifier operates independently. This often results in overestimated co-occurrence probabilities and limited ability to reflect biological modification patterns. In addition, its separate evaluation across modifications complicates the assessment of overall predictive performance. In contrast, multi-label classification (MLC) predicts multiple PTM types simultaneously within a unified framework, where each modification is treated as a separate label [121]. By learning shared representations across labels, MLC can capture statistical associations and co-occurrence patterns among PTMs, thereby better reflecting the combinatorial nature of protein regulation, although its performance may be affected by label imbalance and increased task complexity.

iPTM-mLys [15], introduced in 2016, is the first multi-label predictor designed to identify multiple lysine PTM types, including acetylation, crotonylation, methylation, and succinylation. It uses an ensemble of RF classifiers to predict both modification occurrence and PTM type, achieving an absolute-true rate above 60%. Since iPTM-mLys, several advanced multi-label PTM site prediction tools have emerged, each contributing unique methodologies to enhance prediction accuracy and applicability. mLysPTMpred [124] was developed two years after iPTM-mLys. It employs a combination of SVM classifiers with cost-sensitive learning via the different error costs (DEC) strategy to address class imbalance. mLysPTMpred achieved an accuracy of 83.73% and an absolute-true rate above 80%, representing a significant improvement. iMul-kSite [125] expands the prediction scope to include glutarylation. It improves feature representation through sequence-coupling analysis and incremental feature selection, demonstrating superior predictive performance with 92.83% accuracy, 93.36% aiming, and 96.23% coverage in cross-validation. More recently, MIND-S [126] advances the field through deep learning, integrating feedforward networks, LSTM, multi-head self-attention, and graph attention layers to capture sequence and spatial features. The model shares parameters across PTM types during training, enabling efficient batch prediction across multiple PTM types, including non-lysine sites. Its use of integrated gradients further provides post-hoc interpretability by quantifying residue contributions to predictions.

From an application perspective, the choice between MBC and MLC depends largely on research objectives and data characteristics. MBC-based approaches can remain advantageous for high-throughput screening, as independent classifiers can achieve strong predictive performance and are often easier to implement when data availability varies across PTM types. In contrast, MLC frameworks are often better suited for biological investigations that require joint modeling of PTM dependencies, such as elucidating combinatorial regulation in disease contexts.

6. Extended Research

The continued advancement of AI-driven PTM site prediction tools has led to significant progress in accurately identifying potential modification sites. However, predicting the presence of PTM sites alone is often insufficient to fully understand the complex biological landscape of PTMs. The biological impact of a PTM determined not only by a single modification event, but also by the complex interactions among different PTM sites, known as PTM crosstalk [127]. Beyond crosstalk, understanding the functional impact of individual PTM sites and how they alter protein functions in cellular processes is also essential for unraveling regulatory networks and disease mechanisms.

 doi:10.20944/preprints202603.2316.v1

6.1. PTM Crosstalk Prediction

Unlike PTM site prediction, which identifies specific residues where modifications occur, PTM crosstalk prediction focuses on relationships between PTM pairs and on structural and functional correlations among multiple modification sites.

Computational studies have laid the foundation for understanding PTM crosstalk, revealing spatial associations and co-evolution of related PTM pairs [128,129]. Based on these insights, AI-driven tools such as PTM-X, PPICT, and DeepPCT have substantially improved PTM crosstalk modeling and prediction. PTM-X [130] employs a naïve Bayes classifier to integrate pairwise PTM features, including sequence and structural distance, co-evolution, and co-localization in disordered regions, thereby generating a prediction score for each PTM pair. Building on this foundation, PPICT [131], an integrated deep-learning architecture, moves beyond sequence-level analysis by combining PPI graphs with sequence, structural, and dynamic features, thereby improving the prediction of complex inter-protein PTM crosstalk. DeepPCT [132], published in 2024, leverages AlphaFold2-predicted protein structures to achieve high-accuracy PTM crosstalk prediction, effectively addressing the limitations of structural information availability. Recently, Ou et al. [133] introduced ProXTalk, a dual-stream framework based on protein large language models. ProXTalk consists of a semantic branch that processes protein sequences and a geometric branch that models spatial relationships. The two branches are integrated via a symmetric cross-attention mechanism. This strategy results in more comprehensive and context-aware representations. On the Inter_3549 benchmark dataset, ProXTalk outperformed PTM-X and PPICT with an AUC of 0.906.

PTM crosstalk prediction complements site prediction by elucidating how modifications interact to regulate protein function, offering insights into their collective impact on cellular processes, and clarifying protein functions in signaling pathways. This is particularly significant in diseases like cancer, where aberrant PTM crosstalk disrupts cellular regulation. However, the extremely limited availability of experimentally validated data has constrained the development of PTM crosstalk tools, with widely used databases such as dbPTM currently containing only 491 PTM crosstalk pairs [38]. Future directions should prioritize expanding experimental datasets through advanced mass spectrometry (MS) techniques to generate validated crosstalk data. Advances in PTM crosstalk prediction may deepen our understanding of PTM regulatory networks, with implications for identifying therapeutic targets in diseases driven by dysregulated PTM crosstalk.

6.2. PTM Functional Prediction

Functional prediction assesses the regulatory effects of PTMs, particularly phosphorylation. This is achieved by analyzing features such as evolutionary conservation, sequence/structural context, dynamics, and protein-protein interaction (PPI), with AI models assigning functional scores to prioritize impactful phosphosites.

Ochoa et al. [134] pioneered this field by constructing a reference human phosphoproteome and integrating 59 features into a phosphosite functional score using a gradient-boosting machine. This framework generates a functional score that ranks phosphosites by regulatory significance. Building on this, FuncPhos-SEQ [135] employs an integrated neural network model using input sequence and PPI data. In FuncPhos-SEQ, CNNs extract kinase-specific motifs, while network embedding and DNNs capture PPI features. These features are then integrated to generate a functional probability score for each phosphosite. FuncPhos-STR [136] extends this by incorporating AlphaFold-derived structural and dynamics information to improve predictive accuracy. It achieved a best AUC of 0.855 and outperformed previously reported models. MMFuncPhos [137] further advances this field through a multi-modal framework that integrates the protein language model ESM-2 with the graph convolutional network. Importantly, this framework also enabled the development of EFuncType through transfer learning. EFuncType represents the first method capable of predicting the directional effect (upregulation or downregulation) of a phosphorylation event on enzyme activity, thereby addressing a key gap in PTM functional annotation.

 doi:10.20944/preprints202603.2316.v1

Functional prediction, as exemplified by these models, complements site prediction by providing a deeper understanding of PTM functions in biological processes. For instance, functional prediction by FuncPhos-SEQ [135] has suggested NADK-S48/50 as functional phosphosites, aiding subsequent experimental validation, which revealed that ERK1/2 can phosphorylate NADK-S48/50 and activate NADK activity. Such advances are valuable for applications such as biomarker discovery and drug development, as they provide insights into key molecular mechanisms and help identify potential therapeutic targets.

7. Challenges and Future Directions

The preceding discussion underscores the rapid expansion of research efforts on AI-based PTM site prediction. Although significant progress has been made, persistent gaps and unresolved issues necessitate deeper investigation. In this section, we outline these challenges and propose research directions that may contribute to future improvements in the field (Figure 4).

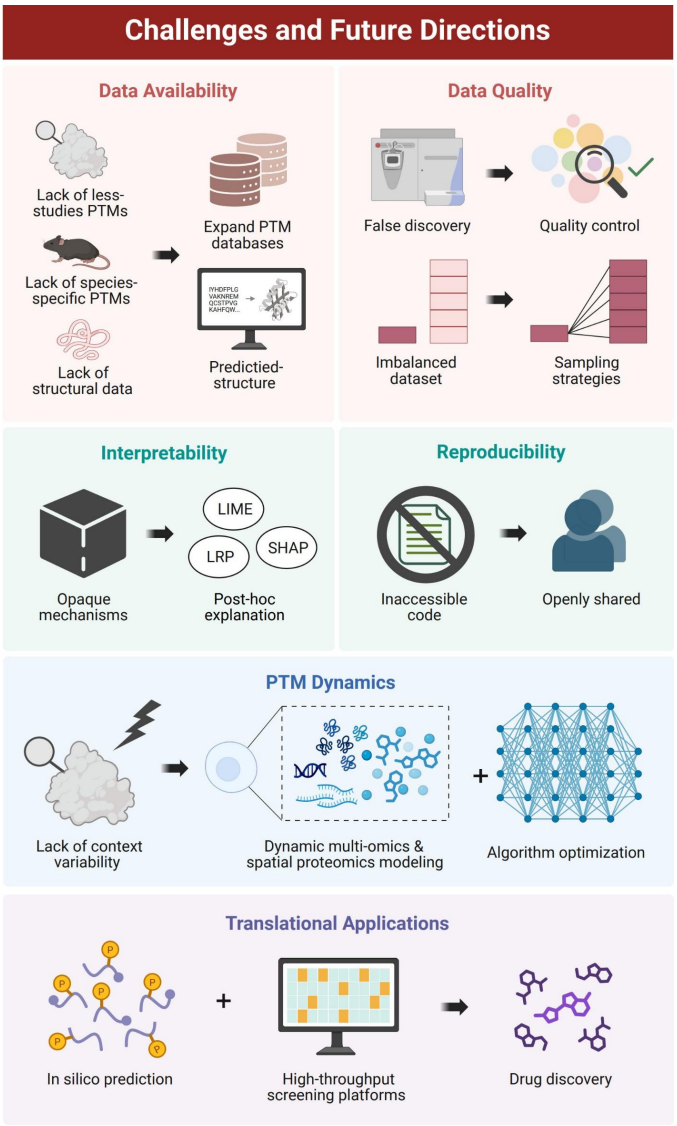

Figure 4. Summary of the challenges and future research directions associated with AI-based PTM site prediction. The diagram synthesizes current challenges in data availability and quality, interpretability, reproducibility, and PTM dynamics, while also highlighting future directions for translational applications. AI, artificial intelligence; PTM, post-translational modification; LIME, local interpretable model agnostic explanations; SHAP, Shapley additive explanations; LRP, layer-wise relevance propagation. This figure was created with BioRender.

 doi:10.20944/preprints202603.2316.v1

7.1. Data Availability

One of the primary challenges in AI-assisted PTM site 852 prediction is the limited availability of comprehensive datasets. 853 Although established databases such as UniProtKB/SwissProt, PhosphoSitePlus, and dbPTM provide valuable repositories of PTM annotations, they are heavily biased toward well-characterized modifications and offer relatively sparse information for less-explored PTMs [1]. Meanwhile, annotation depth in these resources remains predominantly concentrated in model organisms such as humans, mice, and yeast, posing a significant barrier to cross-species generalization [1]. To address these gaps, high-throughput experimental techniques, such as MS and nanopore detection [138,139], hold promise for expanding PTM datasets by enabling more comprehensive profiling of PTMs. These advances could expand the datasets needed to support the development of more robust and generalizable AI prediction models.

Another limitation in PTM site prediction datasets is the limited availability of experimentally determined structural data. PTMs are often influenced by the spatial structure of proteins. However, traditional approaches for determining protein structure, such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, entail significant time and financial costs, resulting in incomplete structural coverage across the proteome [140]. Therefore, many sequence-based models overlook the structural context of PTM sites.

Recent advances in deep learning models such as AlphaFold have improved this situation by enabling the prediction of protein structures with remarkable accuracy [141]. These models have significantly expanded the availability of structural data for PTMs. However, integrating these predicted structures into AI models presents notable challenges. Experimental validation studies have demonstrated that predicted structural models are not uniformly reliable, particularly for multi-domain proteins and intrinsically disordered regions, which are often enriched in PTM sites [142,143]. This may introduce bias during model training. Generating predicted structures at scale can also impose substantial computational costs, especially for large proteomes. Therefore, fully leveraging this promising resource requires specialized strategies for weighting and integrating structural information, as well as scalable computational frameworks to support large-scale processing.

7.2. Data Quality

Experimental methods for detecting PTMs, such as MS, can introduce errors through false discoveries [144]. This compromises the quality of the training data, increasing the risk of AI models overfitting to noise and reducing their generalizability to unseen datasets. In this context, implementing quality control (QC) measures is crucial. These include QC analysis in real time (QC-ART), which can be used to monitor instrument performance or sample quality, and post-collection data validation to eliminate low-quality data [145]. AI-assisted QC also shows considerable promise. Recently, Gao et al. introduced iDIA-QC, an AI-driven tool for classifying data quality in data-independent acquisition MS [146]. This approach offers greater sensitivity and precision compared to traditional data-dependent acquisition in detecting faults in the MS, helping improve data reliability for training AI-based PTM site prediction models.

Class imbalance remains a common problem, where non-PTM sites in training datasets outnumber PTM sites [147]. This imbalance limits the ability of algorithms to capture the underlying data distribution, thereby reducing the sensitivity of AI models in detecting true PTM sites [148]. Techniques such as oversampling, undersampling, or cost-sensitive learning can be used to address this [149]. However, they require careful implementation to avoid introducing new biases, such as overfitting caused by oversampling or loss of information through undersampling [149].

7.3. Interpretability

Interpretability includes two critical aspects, transparency and post-hoc explanation [150]. Transparency describes the extent to which a model's inner workings are understandable to humans. It includes simulatability, decomposability, and algorithmic transparency [150]. Simple conventional

 doi:10.20944/preprints202603.2316.v1

machine learning models, such as decision trees, can be more transparent [151]. However, AI-based PTM predictors today often rely on deep, multi-module models that remain largely opaque to end users due to their complex architectures and non-linear computations, functioning as "black boxes". This complicates the understanding of how specific features in the data contribute to the model's predictions, limiting their practical application in experimental and clinical settings. In such cases, post-hoc explanation, which analyzes information from trained models, can provide insights into how predictions are generated [150,151]. Common post-hoc approaches include local interpretable model-agnostic explanations (LIME), anchors, Shapley additive explanations (SHAP), and layer-wise relevance propagation (LRP) [151,152]. By providing a clearer understanding of how models arrive at their decisions, these techniques can increase researchers' confidence in the reliability of the model's output, thereby facilitating its use in biological discovery and validation. Future studies should therefore consider incorporating advanced post-hoc explanation methods into PTM site prediction models. Such a combination may not only help justify model decisions and build trust, but also assist in identifying and correcting errors, improving models, and generating new insights [153].

7.4. Reproducibility

Reproducibility in AI-based PTM site prediction faces limitations due to a lack of open-source code, which impedes researchers from independently verifying and building upon existing research findings. While tools like MusiteDeep [13] are openly available, some others are accessible only as web servers or appear to have become inaccessible at their original URLs (such as ModPred [120]). Such limited accessibility can reduce standardization and comparability across studies. In the absence of publicly available training code, researchers often rely on online web servers for benchmarking, which can introduce inconsistencies due to differences in server versions, undocumented preprocessing steps, or restricted access to model configurations and datasets. The TRIPOD+AI statement has emphasized the importance of openness, encouraging researchers to report whether analytical code is publicly available and where it can be accessed [154]. To ensure scientific integrity and advance AI-assisted PTM site prediction, the field must emphasize adherence to standardized reporting guidelines and advocate for open access to code and data.

7.5. PTM Dynamics

Another major limitation of current AI-based PTM site-predictors is their inability to capture condition-dependent variability. For example, lysine ubiquitination levels can change in response to DNA damage [155], acetylation has been linked to age-related inflammation [156], nd a ketogenic diet has been found to influence lysine β-hydroxybutyrylation [157]. However, such condition-dependent modifications are rarely reflected in model predictions because many existing predictors still rely primarily on protein sequences. These context-dependent factors can lead to discrepancies between predicted outcomes and actual biological scenarios.

Addressing this limitation will require predictive frameworks that model dynamic cellular contexts. Future research should incorporate condition-aware features by expanding input modalities, including single-cell multi-omics, spatial proteomics, and mechanistic modeling. Such integration could systematically capture cell-to-cell heterogeneity, thereby improving biological realism in PTM site prediction.

In this context, GNNs represent a particularly promising direction, as their architecture enables explicit modeling of complex molecular interactions and spatial dependencies. By representing proteins and cellular components as interconnected systems, GNN-based approaches are well-suited to capture the dynamic regulatory networks underlying PTMs. Future GNN frameworks may integrate multimodal information using heterogeneous graph attention mechanisms [158]. Furthermore, the incorporation of GNN explanation methods, such as SubgraphX, could improve biological interpretability, potentially helping researchers to trace how predicted PTM sites relate to specific signaling pathways, disease states, or cellular perturbations [159].

 doi:10.20944/preprints202603.2316.v1

Ultimately, such developments may substantially improve the biological realism and contextual accuracy of PTM site prediction across diverse physiological and pathological conditions.

7.6. Translational Applications

Beyond methodological advances, the ultimate value of AI-based PTM site prediction lies in its translational impact. Recent studies have begun to bridge in silico prediction and translational research. For example, GPS-Uber has been used to predict ubiquitination sites on cancer proteins, providing candidates for experimental follow-up [88]. Similarly, DeepMVP [160] has been applied to predict PTM-altering variants, identifying cancer-associated PTMs such as reduced inhibitory acetylation on AKT1 and increased inhibitory phosphorylation on TP53, both offering mechanistic hypotheses with potential therapeutic relevance. However, the number of experimentally validated examples remains limited, and substantial efforts are required to translate computational predictions into translational insights.

Looking forward, integrating AI-based PTM site prediction with high-throughput screening platforms could further accelerate the drug discovery pipeline. For example, coupling PTM site prediction with DNA-encoded library screening or fragment-based drug design may help prioritize compounds that target PTM-regulated protein states [161]. Such approaches could yield novel therapeutics for diseases where PTM dysregulation is a hallmark, including cancers, metabolic disorders, and neurodegeneration.

8. Method Selection Guideline

With the rapid expansion of AI-based tools for predicting PTM sites, selecting an appropriate method requires careful consideration of study objectives, data availability, and model characteristics. The following factors provide practical guidance for method selection across diverse PTM types.

8.1. Prediction Scope

Models trained on large datasets can be well suited for broad PTM site prediction across diverse proteins. However, their performance may decline when applied to underrepresented protein families or rare modification types. Researchers should therefore assess whether their target proteins and PTM types fall within the training distribution represented by the selected method.

8.2. Model Architecture

Model architecture influences the biological patterns that a predictor can capture and should therefore be selected to match the prediction task. CNN-based predictors are effective at identifying local sequence features surrounding modification sites, whereas RNNs model sequential dependencies along protein sequences [49,162]. Transformer-based models are well suited for capturing long-range dependencies and global sequence context, enabling flexible representation of complex regulatory relationships [163]. In addition, GNNs are well suited for modeling structural and relational information [51]. Hybrid architectures that integrate multiple modeling strategies may further improve biological representation, but typically impose additional computational demands.

8.3. Interpretability

For applications requiring mechanistic understanding, interpretability is a critical consideration. Conventional machine learning models can offer more transparent decision rules. In contrast, deep learning models usually function as black boxes, which may limit their utility in studies requiring mechanistic explanation despite strong predictive performance.

8.4. Performance Evaluation

Reported performance metrics should be interpreted with caution, as they can vary substantially depending on the specific composition of training and test sets. Cross-validation may yield optimistic

 doi:10.20944/preprints202603.2316.v1

estimates, particularly when redundancy exists between training and test samples, while independent test sets generally offer a more reliable indicator of a model's generalization ability.

9. Conclusions

This review highlights the growing role of AI in PTM site prediction and its potential to advance research on protein function and regulation. In recent years, AI models, from traditional machine learning to deep learning, have made substantial progress in predicting various types of PTM sites. These advancements now extend beyond single-PTM site prediction to multiple-PTM site prediction, PTM crosstalk prediction, and functional assessment of modifications. Despite these developments, key challenges remain, including limitations in data availability and quality, model interpretability and reproducibility, and an incomplete understanding of PTM dynamics. Improvements in dataset scale and quality, algorithmic innovation, and model interpretability will be essential for enhancing predictive accuracy and real-world applicability. To further promote practical implementation, we provide a method selection guideline to help researchers choose appropriate tools based on prediction scope, model architecture, interpretability, and evaluation strategy. This review supports both the development of AI-based prediction tools and their effective application in PTM research, thereby promoting deeper integration of AI into molecular biology and biomedical research.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Table S1: A summary of the advantages and limitations of existing AI-based PTM site prediction tools.

Author Contributions: J.R.: Investigation, Writing—original draft, Writing—review & editing X.Z. (Xiaohan Zhang): Investigation, Writing—original draft, Writing—review & editing Y.W.: Investigation, Writing—review & editing, Software S.C.: Investigation, Writing—review & editing, Software J.D.: Writing—review & editing M.K.: Writing—review & editing B.Y.: Writing—review & editing H.W.: Writing—review & editing Y.D.: Writing—review & editing T.L.: Writing—review & editing Y.L.: Writing—review & editing L.W.: Writing—review & editing Y.G.: Writing—review & editing X.Z. (Xiaochen Zhang): Writing—review & editing H.H.: Writing—review & editing J.Z.: Writing—review & editing J.Z.: Writing—review & editing Z.X.: Writing—review & editing, Funding acquisition Y.T.: Writing—review & editing, Funding acquisition J.L.: Conceptualization, Supervision, Writing—review & editing, Funding acquisition. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work is supported by grants to Dr. Jian Liu from the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2023ZD0507500/2023ZD0507501 and 2023ZD0502900/2023ZD0502902); the National Natural Science Foundation of China (General Grant: 82172899 and 82472637); the National Natural Science Foundation of China (No. 82573076); Key Research and Development Program of Zhejiang Province(No.2025C02094); Zhejiang University; the Open Fund of Zhejiang Provincial Key Laboratory of Pulmonology (KF202302); ZJE seed funding; ZJE 2024 International Campus Talent Special Funding Program; ZJE-UoE Joint Research Project; the Sanming Project of Medicine in Shenzhen (No.SZSM202403006); the Key R&D Program of Zhejiang Respiratory Disease (No.2025C02105 and 2025C02091); and the Zhejiang Provincial "Leading Goose" R&D Programs (No.2025C02105). We acknowledge the help of members in JL's lab. We thank the Biomed-X Laboratory of ZJE Institute, School of Medicine, Zhejiang University, for continuous support.

Conflicts of Interest: No competing interest is declared.

References

1. Shahin Ramazi and Javad Zahiri. Post-translational modifications in proteins: resources, tools and prediction methods. *Database*, 2021:baab012, 2021.

 doi:10.20944/preprints202603.2316.v1

2. PA Levene and CL Alsberg. The cleavage products of vitellin. *Journal of Biological Chemistry*, 2(1):127-133, 1906.
3. Fritz A Lipmann and PA Levene. Serinephosphoric acid obtained on hydrolysis of vitellinic acid. *Journal of Biological Chemistry*, 98(1):109-114, 1932.
4. DMP Phillips. The presence of acetyl groups in histones. *Biochemical Journal*, 87(2):258, 1963.
5. Gideon Goldstein, Margrit Scheid, Ulrich Hammerling, DH Schlesinger, HD Niall, and EA Boyse. Isolation of a polypeptide that has lymphocyte-differentiating properties and is probably represented universally in living cells. *Proceedings of the National Academy of Sciences*, 72(1):11-15, 1975.
6. Ji Min Lee, Henrik M Hammarén, Mikhail M Savitski, and Sung Hee Baek. Control of protein stability by post-translational modifications. *Nature Communications*, 14(1):201, 2023.
7. Luoyi Chen and Min Huang. Oncometabolites in cancer: from cancer cells to the tumor microenvironment. *Holistic Integrative Oncology*, 3(1):26, 2024.
8. Juliane Hermann, Leon Schurgers, and Vera Jankowski. Identification and characterization of post-translational modifications: Clinical implications. *Molecular Aspects of Medicine*, 86:101066, 2022.
9. Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198-207, 2003.
10. Daniel J Geiszler, Daniel A Polasky, Fengchao Yu, and Alexey I Nesvizhskii. Detecting diagnostic features in ms/ms spectra of post-translationally modified peptides. *Nature Communications*, 14(1):4132, 2023.
11. Daniel Schwartz and Steven P Gygi. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nature Biotechnology*, 23(11):1391-1398, 2005.
12. Jianjiong Gao, Jay J Thelen, A Keith Dunker, and Dong Xu. Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Molecular & Cellular Proteomics*, 9(12):2586-2600, 2010.
13. Duolin Wang, Shuai Zeng, Chunhui Xu, Wangren Qiu, Yanchun Liang, Trupti Joshi, and Dong Xu. Musitedeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics*, 33(24):3909-3916, 2017.
14. Xun Wang, Zhiyuan Zhang, Chaogang Zhang, Xiangyu Meng, Xin Shi, and Peng Qu. Transphos: A deep-learning model for general phosphorylation site prediction based on transformer-encoder architecture. *International Journal of Molecular Sciences*, 23(8):4263, 2022.
15. Wang-Ren Qiu, Bi-Qian Sun, Xuan Xiao, Zhao-Chun Xu, and Kuo-Chen Chou. iptm-mlys: identifying multiple lysine ptm sites and their different types. *Bioinformatics*, 32(20):3116-3123, 2016.
16. Anca-Narcisa Neagu, Claudiu-Laurentiu Josan, Taniya M Jayaweera, Hailey Morrissiey, Kaya R Johnson, and Costel C Darie. Bio-pathological functions of posttranslational modifications of histological biomarkers in breast cancer. *Molecules*, 29(17):4156, 2024.
17. Jeffrey A Ubersax and James E Ferrell Jr. Mechanisms of specificity in protein phosphorylation. *Nature Reviews Molecular Cell Biology*, 8(7):530-541, 2007.
18. Maria Shvedunova and Asifa Akhtar. Modulation of cellular processes by histone and non-histone protein acetylation. *Nature Reviews Molecular Cell Biology*, 23(5):329-349, 2022.
19. John M Aletta, Thomas R Cimato, and Murray J Ettinger. Protein methylation: a signal event in post-translational modification. *Trends in Biochemical Sciences*, 23(3):89-91, 1998.
20. Aert F Scheper, Jack Schofield, Raghvendra Bohara, Thomas Ritter, and Abhay Pandit. Understanding glycosylation: Regulation through the metabolic flux of precursor pathways. *Biotechnology Advances*, 67:108184, 2023.
21. Rune Busk Damgaard. The ubiquitin system: from cell signalling to disease biology and new therapeutic opportunities. *Cell Death & Differentiation*, 28(2):423-426, 2021.
22. Alfred CO Vertegaal. Signalling mechanisms and cellular functions of sumo. *Nature Reviews Molecular Cell Biology*, 23(11):715-731, 2022.
23. Rui Shen, Hongyun Ruan, Shuye Lin, Bin Liu, Hang Song, Lu Li, and Teng Ma. Lysine succinylation, the metabolic bridge between cancer and immunity. *Genes & Diseases*, 10(6):2470-2478, 2023.
24. Hongling Zhao, Yang Han, Pingkun Zhou, Hua Guan, and Shanshan Gao. Protein lysine crotonylation in cellular processes and disease associations. *Genes & Diseases*, 11(5):101060, 2024.
25. Hongde Li, Linchong Sun, Ping Gao, and Hai Hu. Lactylation in cancer: Current understanding and challenges. *Cancer Cell*, 42(11):1803-1807, 2024.
26. Lu Zou, Yanyan Yang, Zhibin Wang, Xiuxiu Fu, Xiangqin He, Jiayi Song, Tianxiang Li, Huibo Ma, and Tao Yu. Lysine malonylation and its links to metabolism and diseases. *Aging and Disease*, 14(1):84, 2023.

 doi:10.20944/preprints202603.2316.v1

27. Shizhen Zhang, Qing Yu, Zhijian Li, Yongchao Zhao, and Yi Sun. Protein neddylation and its role in health and diseases. *Signal Transduction and Targeted Therapy*, 9(1):85, 2024.
28. Jonathan X Meng, Yu Zhang, Dominik Saman, Arshad M Haider, Suman De, Jason C Sang, Karen Brown, Kun Jiang, Jane Humphrey, Linda Julian, et al. Hyperphosphorylated tau self-assembles into amorphous aggregates eliciting tlr4-dependent responses. *Nature Communications*, 13(1):2692, 2022.
29. Han Hong, Hexu Han, Lei Wang, Wen Cao, Minjie Hu, Jindong Li, Jiawei Wang, Yijin Yang, XiaoYong Xu, Gaochao Li, et al. Abcf1-k430-lactylation promotes hcc malignant progression via transcriptional activation of hif1 signaling pathway. *Cell Death & Differentiation*, 32(4):613-631, 2025.
30. The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609-D617, 2025.
31. Helen M Berman and Stephen K Burley. Protein data bank (pdb): Fifty-three years young and having a transformative impact on science and society. *Quarterly reviews of biophysics*, 58:e9, 2025.
32. Brinda Vallat, Yana Rose, Dennis W Piehl, Jose M Duarte, Sebastian Bittrich, Chunxiao Bi, Joan Segura, Arthur Zalevsky, Monica R Sekharan, Benjamin M Webb, et al. Rcsb protein data bank: Delivering integrative structures alongside experimental structures and computed structure models. *Nucleic Acids Research*, 54(D1):D489-D498, 2026.
33. Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Henry Chao, Li Chen, Paul A Craig, Gregg V Crichlow, Kenneth Dalenberg, Jose M Duarte, et al. Rcsb protein data bank: Tools for visualizing and understanding biological macromolecules in 3d. *Protein Science*, 31(12):e4482, 2022.
34. Peter V Hornbeck, Jon M Kornhauser, Sasha Tkachev, Bin Zhang, Elżbieta Skrzypek, Beth Murray, Vaughan Latham, and Michael Sullivan. Phosphositeplus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Research*, 40(D1):D261-D270, 2012.
35. Peter V Hornbeck, Bin Zhang, Beth Murray, Jon M Kornhauser, Vaughan Latham, and Elzbieta Skrzypek. Phosphositeplus, 2014: mutations, ptms and recalibrations. *Nucleic Acids Research*, 43(D1):D512-D520, 2015.
36. Francesca Diella, Scott Cameron, Christine Gemünd, Rune Linding, Allegra Via, Bernhard Kuster, Thomas Sicheritz-Pontén, Nikolaj Blom, and Toby J Gibson. Phospho. elm: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, 5(1):79, 2004.
37. Holger Dinkel, Claudia Chica, Allegra Via, Cathryn M Gould, Lars J Jensen, Toby J Gibson, and Francesca Diella. Phospho. elm: a database of phosphorylation sites—update 2011. *Nucleic Acids Research*, 39(suppl_1):D261-D267, 2010.
38. Chia-Ru Chung, Yun Tang, Yen-Peng Chiu, Shangfu Li, Wen-Kai Hsieh, Lantian Yao, Ying-Chih Chiang, Yuxuan Pang, Guan-Ting Chen, Kai-Chen Chou, et al. dbptm 2025 update: comprehensive integration of ptms and proteomic data for advanced insights into cancer research. *Nucleic Acids Research*, 53(D1):D377-D386, 2025.
39. Haodong Xu, Jiaqi Zhou, Shaofeng Lin, Wankun Deng, Ying Zhang, and Yu Xue. Plmd: an updated data resource of protein lysine modifications. *Journal of Genetics and Genomics*, 44(5):243-250, 2017.
40. Farzaneh Esmaili, Mahdi Pourmirzaei, Shahin Ramazi, Seyedehsamaneh Shojaeilangari, and Elham Yavari. A review of machine learning and algorithmic methods for protein phosphorylation site prediction. *Genomics, Proteomics & Bioinformatics*, 21(6):1266-1285, 2023.
41. Austin Spadaro, Alok Sharma, and Iman Dehzangi. Predicting lysine methylation sites using a convolutional neural network. *Methods*, 226:127-132, 2024.
42. Xiaoyang Jing, Qiwen Dong, Daocheng Hong, and Ruqian Lu. Amino acid encoding methods for protein sequences: a comprehensive review and assessment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(6):1918-1931, 2019.
43. Md Mehedi Hasan, Dianjing Guo, and Hiroyuki Kurata. Computational identification of protein s-sulfenylation sites by incorporating the multiple sequence features information. *Molecular BioSystems*, 13(12):2545-2550, 2017.
44. Sunil Kumar and Vaibhav Bhatnagar. A review of regression models in machine learning. *Journal of Intelligent Systems and Computing*, 3(1):40-47, 2022.
45. Barry De Ville. Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(6):448-455, 2013.
46. Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. A review of kernel methods in machine learning. *Mac-Planck-Institute Technical Report*, 156(1), 2006.
47. Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452-459, 2015.

 doi:10.20944/preprints202603.2316.v1

48. Rudolf Kruse, Sanaz Mostaghim, Christian Borgelt, Christian Braune, and Matthias Steinbrecher. Multi-layer perceptrons. In *Computational Intelligence: a Methodological Introduction*, pages 53-124. Springer, 2022.
49. Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):6999-7019, 2021.
50. Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235-1270, 2019.
51. Bharti Khemani, Shruti Patil, Ketan Kotecha, and Sudeep Tanwar. A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. *Journal of Big Data*, 11(1):18, 2024.
52. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.
53. Nikolaj Blom, Steen Gammeltoft, and Søren Brunak. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of Molecular Biology*, 294(5):1351-1362, 1999.
54. Majbrit Hjerrild, Allan Stensballe, Thomas E Rasmussen, Christine B Kofoed, Nikolaj Blom, Thomas Sicheritz-Ponten, Martin R Larsen, Søren Brunak, Ole N Jensen, and Steen Gammeltoft. Identification of phosphorylation sites in protein kinase a substrates using artificial neural networks and mass spectrometry. *Journal of Proteome Research*, 3(3):426-433, 2004.
55. Fenglin Luo, Minghui Wang, Yu Liu, Xing-Ming Zhao, and Ao Li. Deepphos: prediction of protein phosphorylation sites with deep learning. *Bioinformatics*, 35(16):2766-2773, 2019.
56. Jiale Lu, Haibin Chen, and Ji Qiu. Phoshsgn: deep neural networks combining sequence and protein spatial information to improve protein phosphorylation site prediction. *IEEE Access*, 2024.
57. Pawel Pratyush, Callen Carrier, Suresh Pokharel, Hamid D Ismail, Meenal Chaudhari, and Dukka B Kc. Calmphoskan: prediction of general phosphorylation sites in proteins via fusion of codon aware embeddings with amino acid aware embeddings and wavelet-based kolmogorov-arnold network. *Bioinformatics*, 41(4):btaf124, 2025.
58. Mengya Liu, Xin Wang, Zhan-Li Sun, Xiao Yang, and Xia Chen. Dcpps: Prediction of kinase-specific phosphorylation sites using dynamic embedding and cross-representation interaction. *Interdisciplinary Sciences: Computational Life Sciences*, pages 1-18, 2025.
59. Lars Kiemer, Jannick Dyrløv Bendtsen, and Nikolaj Blom. Netacet: prediction of n-terminal acetylation sites. *Bioinformatics*, 21(7):1269-1270, 2005.
60. Yan Xu, Xiao-Bo Wang, Jun Ding, Ling-Yun Wu, and Nai-Yang Deng. Lysine acetylation sites prediction using an ensemble of support vector machine classifiers. *Journal of Theoretical Biology*, 264(1):130-135, 2010.
61. Xiaowei Zhao, Jiagen Li, Rui Wang, Fei He, Lin Yue, and Minghao Yin. General and species-specific lysine acetylation site prediction using a bi-modal deep architecture. *IEEE Access*, 6:63560-63569, 2018.
62. Gang-Ao Wang, Xiaodi Yan, Xiang Li, Yinbo Liu, Junfeng Xia, and Xiaolei Zhu. Mstl-kace: prediction of prokaryotic lysine acetylation sites based on multistage transfer learning strategy. *ACS Omega*, 8(44):41930-41942, 2023.
63. Zhaomin Yao, Haonan Shangguan, Weiming Xie, Jiahao Liu, Sinuo He, Hexin Huang, Fei Li, Jiaming Chen, Ying Zhan, Xiaodan Wu, et al. Sipsc-kac: Integrating swarm intelligence and protein spatial characteristics for enhanced lysine acetylation site identification. *International Journal of Biological Macromolecules*, 282:137237, 2024.
64. Lingkuan Meng, Xingjian Chen, Ke Cheng, Nanjun Chen, Zetian Zheng, Fuzhou Wang, Hongyan Sun, and Ka-Chun Wong. Transptm: a transformer-based model for non-histone acetylation site prediction. *Briefings in Bioinformatics*, 25(3), 2024.
65. Jinsong Ke, Jianmei Zhao, Hongfei Li, Lei Yuan, Guanghui Dong, and Guohua Wang. Prediction of protein n-terminal acetylation modification sites based on cnn-bilstm-attention model. *Computers in Biology and Medicine*, 174:108330, 2024.
66. Yu Liu, Chaofan Ye, Can Lin, Kangkang Mao, and Ming Zhu. Mddeep-ace: species-specific acetylation site prediction based on multi-domain adaptation. *PeerJ*, 13:e19649, 2025.
67. Leyi Wei, Pengwei Xing, Gaotao Shi, Zhiliang Ji, and Quan Zou. Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(4):1264-1273, 2017.
68. Wei Zheng, Qiqige Wuyun, Micah Cheng, Gang Hu, and Yanping Zhang. Two-level protein methylation prediction using structure model-based features. *Scientific Reports*, 10(1):6008, 2020.

 doi:10.20944/preprints202603.2316.v1

69. Meenal Chaudhari, Niraj Thapa, Kaushik Roy, Robert H Newman, Hiroto Saigo, and Dukka BKC. Deeprmethylsite: a deep learning based approach for prediction of arginine methylation sites in proteins. *Molecular Omics*, 16(5):448-454, 2020.
70. Jiaojiao Zhao, Haoqiang Jiang, Guoyang Zou, Qian Lin, Qiang Wang, Jia Liu, and Leina Ma. Cnnarginine: A cnn structure for training models for predicting arginine methylation sites based on the one-hot encoding of peptide sequence. *Frontiers in Genetics*, 13:1036862, 2022.
71. Haiwei Zhou, Wenxi Tan, and Shaoping Shi. Deepgpgs: a novel deep learning framework for predicting arginine methylation sites combined with gaussian prior and gated self-attention mechanism. *Briefings in Bioinformatics*, 24(2):bbad018, 2023.
72. Gaurav Dwivedi, Monika Khandelwal, Ranjeet Kumar Rout, Saiyed Umer, Saurav Mallik, and Hong Qin. Rmsxai: arginine methylation sites prediction from protein sequences using machine learning algorithms and explainable artificial intelligence. *Discover Applied Sciences*, 6(7):329, 2024.
73. Xihong Yuan, Chizhuo Ma, Zhichao Lei, and Chuzheng Wang. Deeptesite: The prediction of protein arginine methylation sites using amino acids sequence symmetric position encodings based on transformer encoder. *IEEE Transactions on Computational Biology and Bioinformatics*,2026.
74. Jan E Hansen, Ole Lund, Niels Tolstrup, Andrew A Gooley, Keith L Williams, and Søren Brunak. Netoglyc: prediction of mucin type o-glycosylation sites based on sequence context and surface accessibility. *Glycoconjugate Journal*, 15(2):115-130, 1998.
75. Fuyi Li, Chen Li, Mingjun Wang, Geoffrey I Webb, Yang Zhang, James C Whisstock, and Jiangning Song. Glycomine: a machine learning-based approach for predicting n-, c-and o-linked glycosylation in the human proteome. *Bioinformatics*, 31(9):1411-1419, 2015.
76. Ghazaleh Taherzadeh, Abdollah Dehzangi, Maryam Golchin, Yaoqi Zhou, and Matthew P Campbell. Sprintgly: predicting n-and o-linked glycosylation sites of human and mouse proteins by using sequence and predicted structural properties. *Bioinformatics*, 35(20):4140-4146, 2019.
77. Subash C Pakhrin, Kiyoko F Aoki-Kinoshita, Doina Caragea, and Dukka B Kc. Deepnglypred: a deep neural network-based approach for human n-linked glycosylation site prediction. *Molecules*, 26(23):7314, 2021.
78. Subash C Pakhrin, Suresh Pokharel, Kiyoko F Aoki-Kinoshita, Moriah R Beck, Tarun K Dam, Doina Caragea, and Dukka B Kc. Lmnglypred: prediction of human n-linked glycosylation sites using embeddings from a pre-trained protein language model. *Glycobiology*, 33(5):411-422, 2023.
79. Ki Wook Lee, Nhat Truong Pham, Hye Jung Min, Hyun Woo Park, Ji Won Lee, Han-En Lo, Na Young Kwon, Jimin Seo, Illia Shaginyan, Heeje Cho, et al. Dogpred: a novel deep learning framework for accurate identification of human o-linked threonine glycosylation sites. *Journal of Molecular Biology*, 437(6):168977, 2025.
80. Seokyoung Hong, Krishna Gopal Chattaraj, Jing Guo, Bernhardt L Trout, and Richard D Braatz. Enhanced o-glycosylation site prediction using explainable machine learning technique with spatial local environment. *Bioinformatics*, 41(2):btaf034, 2025.
81. Md Muhaiminul Islam Nafi. Predicting c-and s-linked glycosylation sites from protein sequences using protein language models. *Computers in Biology and Medicine*, 189:109956, 2025.
82. Predrag Radivojac, Vladimir Vacic, Chad Haynes, Ross R Cocklin, Amrita Mohan, Joshua W Heyen, Mark G Goebl, and Lilia M Iakoucheva. Identification, analysis, and prediction of protein ubiquitination sites. *Proteins: Structure, Function, and Bioinformatics*, 78(2):365-380, 2010.
83. Zhen Chen, Yong-Zi Chen, Xiao-Feng Wang, Chuan Wang, Ren-Xiang Yan, and Ziding Zhang. Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. *PLOS One*, 6(7):e22930, 2011.
84. Jyun-Rong Wang, Wen-Lin Huang, Ming-Ju Tsai, Kai-Ti Hsu, Hui-Ling Huang, and Shinn-Ying Ho. Esaubisite: accurate prediction of human ubiquitination sites by identifying a set of effective negatives. *Bioinformatics*, 33(5):661-668, 2017.
85. Xiaowen Cui, Zhaomin Yu, Bin Yu, Minghui Wang, Baoguang Tian, and Qin Ma. Ubisitepred: A novel method for improving the accuracy of ubiquitination sites prediction by using lasso to select the optimal chou's pseudo components. *Chemometrics and Intelligent Laboratory Systems*, 184:28-43, 2019.
86. Hongli Fu, Yingxi Yang, Xiaobo Wang, Hui Wang, and Yan Xu. Deepubi: a deep learning framework for prediction of ubiquitination sites in proteins. *BMC Bioinformatics*, 20(1):86, 2019.
87. Yin Luo, Jiulei Jiang, Jiajie Zhu, Qiyi Huang, Weimin Li, Ying Wang, and Yamin Gao. A caps-ubi model for protein ubiquitination site prediction. *Frontiers in Plant Science*, 13:884903, 2022.

 doi:10.20944/preprints202603.2316.v1

88. Chenwei Wang, Xiaodan Tan, Dachao Tang, Yujie Gou, Cheng Han, Wanshan Ning, Shaofeng Lin, Weizhi Zhang, Miaomiao Chen, Di Peng, et al. Gps-uber: a hybrid-learning framework for prediction of general and e3-specific lysine ubiquitination sites. *Briefings in Bioinformatics*, 23(2):bbab574, 2022.
89. Yong-Zi Chen, Zhen Chen, Yu-Ai Gong, and Guoguang Ying. Sumohydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties. *PLOS One*, 7(6):e39195, 2012.
90. Qi Zhao, Yubin Xie, Yueyuan Zheng, Shuai Jiang, Wenzhong Liu, Weiping Mu, Zexian Liu, Yong Zhao, Yu Xue, and Jian Ren. Gps-sumo: a tool for the prediction of sumoylation sites and sumo-interaction motifs. *Nucleic Acids Research*, 42(W1):W325-W330, 2014.
91. Chi-Chang Chang, Chi-Hua Tung, Chi-Wei Chen, Chin-Hau Tu, and Yen-Wei Chu. Sumogo: Prediction of sumoylation sites on lysines by motif screening models and the effects of various post-translational modifications. *Scientific Reports*, 8(1):15512, 2018.
92. Ying Qian, Shasha Ye, Yu Zhang, and Jiongmin Zhang. Sumo-forest: a cascade forest based method for the prediction of sumoylation sites on imbalanced data. *Gene*, 741:144536, 2020.
93. Salman Khan, Mukhtaj Khan, Nadeem Iqbal, Naqqash Dilshad, Maram Fahaad Almufareh, and Najah Alsubaie. Enhancing sumoylation site prediction: a deep neural network with discriminative features. *Life*, 13(11):2153, 2023.
94. Salman Khan, Salman A AlQahtani, Sumaiya Noor, and Nijad Ahmad. Pssm-sumo: deep learning based intelligent model for prediction of sumoylation sites using discriminative features. *BMC Bioinformatics*, 25(1):284, 2024.
95. Yujie Gou, Dan Liu, Miaomiao Chen, Yuxiang Wei, Xinhe Huang, Cheng Han, Zihao Feng, Chi Zhang, Teng Lu, Di Peng, et al. Gps-sumo 2.0: an updated online service for the prediction of sumoylation sites and sumo-interacting motifs. *Nucleic Acids Research*, 52(W1):W238-W247, 2024.
96. Cheng-Hsun Ho, Yen-Wei Chu, Lan-Ying Huang, and Chi-Wei Chen. Sumo-lmnet: Lossless mapping network for predicting sumoylation sites in sumo1 and sumo2 using high-dimensional features. *Computational and Structural Biotechnology Journal*, 27:1048-1059, 2025.
97. Yan Xu, Ya-Xin Ding, Jun Ding, Ya-Hui Lei, Ling-Yun Wu, and Nai-Yang Deng. isuc-pseaac: predicting lysine succinylation in proteins by incorporating peptide position-specific propensity. *Scientific Reports*, 5(1):10184, 2015.
98. Qiao Ning, Xiaosa Zhao, Lingling Bao, Zhiqiang Ma, and Xiaowei Zhao. Detecting succinylation sites from protein sequences using ensemble support vector machine. *BMC Bioinformatics*, 19(1):1–9, 2018.
99. Kai-Yao Huang, Justin Bo-Kai Hsu, and Tzong-Yi Lee. Characterization and identification of lysine succinylation sites based on deep learning method. *Scientific Reports*, 9(1):16175, 2019.
100. Wanshan Ning, Haodong Xu, Peiran Jiang, Han Cheng, Wankun Deng, Yaping Guo, and Yu Xue. Hybridsucc: a hybrid-learning architecture for general and species-specific succinylation site prediction. *Genomics, Proteomics & Bioinformatics*, 18(2):194-207, 2020.
101. Huiqing Wang, Hong Zhao, Jing Zhang, Jiale Han, and Zhihao Liu. A parallel model of densecnn and ordered-neuron lstm for generic and species-specific succinylation site prediction. *Biotechnology and Bioengineering*, 119(7):1755-1767, 2022.
102. Jianhua Jia, Genqiang Wu, Meifang Li, and Wangren Qiu. psuc-edbam: Predicting lysine succinylation sites in proteins based on ensemble dense blocks and an attention module. *BMC Bioinformatics*, 23(1):450, 2022.
103. Jianhua Jia, Genqiang Wu, and Wangren Qiu. psuc-ffsea: predicting lysine succinylation sites in proteins based on feature fusion and stacking ensemble algorithm. *Frontiers in Cell and Developmental Biology*, 10:894874, 2022.
104. Shehab Sarar Ahmed, Zaara Tasnim Rifat, M Saifur Rahman, and M Sohel Rahman. Succinylated lysine residue prediction revisited. *Briefings in Bioinformatics*, 24(1), 2023.
105. Thi-Xuan Tran, Nguyen Quoc Khanh Le, and Van-Nui Nguyen. Integrating cnn and bi-lstm for protein succinylation sites prediction based on natural language processing technique. *Computers in Biology and Medicine*, 186:109664, 2025.
106. Shahid Akbar, Ali Raza, Wajdi Alghamdi, Hashim Ali, Quan Zou, and Ximei Luo. Identifying protein succinylation sites using generative transformer and a two-dimensional representation with a deep capsule network. *Iscience*, 28(12), 2025.
107. Guohua Huang and Wenfei Zeng. A discrete hidden markov model for detecting histone crotonyllysine sites. *MATCH Communications in Mathematical and in Computer Chemistry*, 75:717-30, 2016.
108. Zhe Ju and Jian-Jun He. Prediction of lysine crotonylation sites by incorporating the composition of k-spaced amino acid pairs into chou's general pseaac. *Journal of Molecular Graphics and Modelling*, 77:200-204, 2017.

 doi:10.20944/preprints202603.2316.v1

109. Hao Lv, Fu-Ying Dao, Zheng-Xing Guan, Hui Yang, Yan-Wen Li, and Hao Lin. Deep-kcr: accurate detection of lysine crotonylation sites using deep learning method. *Briefings in Bioinformatics*, 22(4):bbaa255, 2021.
110. Jhabindra Khanal, Hilal Tayara, Quan Zou, and Kil To Chong. Deepcap-kcr: accurate identification and investigation of protein lysine crotonylation sites based on capsule network. *Briefings in Bioinformatics*, 23(1):bbab492, 2022.
111. Yun Zuo, Minquan Wan, Yang Shen, Xinheng Wang, Wenying He, Yue Bi, Xiangrong Liu, and Zhaohong Deng. Ilycrosite: Identification of lysine crotonylation sites based on fcm-grnn undersampling technique. *Computational Biology and Chemistry*, 113:108212, 2024.
112. Pawel Pratyush, Soufia Bahmani, Suresh Pokharel, Hamid D Ismail, and Dukka B Kc. Lmcrot: an enhanced protein crotonylation site predictor by leveraging an interpretable window-level embedding from a transformer-based protein language model. *Bioinformatics*, 40(5):btae290, 2024.
113. Yunyun Liang and Minwei Li. A deep learning model for prediction of lysine crotonylation sites by fusing multi-features based on multi-head self-attention mechanism. *Scientific Reports*, 15(1):18940, 2025.
114. Peiran Jiang, Wanshan Ning, Yunshu Shi, Chuan Liu, Saijun Mo, Haoran Zhou, Kangdong Liu, and Yaping Guo. Fsl-kla: A few-shot learning-based multi-feature hybrid system for lactylation site prediction. *Computational and Structural Biotechnology Journal*, 19:4497-4509, 2021.
115. Hao Lv, Fu-Ying Dao, and Hao Lin. Deepkla: An attention mechanism-based deep neural network for protein lysine lactylation site prediction. *iMeta*, 1(1):e11, 2022.
116. Ye-Hong Yang, Jun-Tao Yang, and Jiang-Feng Liu. Lactylation prediction models based on protein sequence and structural feature fusion. *Briefings in Bioinformatics*, 25(2), 2024.
117. Minghui Wang, Lili Song, Yaqun Zhang, Hongli Gao, Lu Yan, and Bin Yu. Malsite-deep: prediction of protein malonylation sites through deep learning and multi-information fusion based on nearmiss-2 strategy. *Knowledge-Based Systems*, 240:108191, 2022.
118. Zhaohui Qin, Huixia Liu, Pei Zhao, Kaiyuan Wang, Haoran Ren, Chunbo Miao, Junzhou Li, Yong-Zi Chen, and Zhen Chen. Slam: Structure-aware lysine β-hydroxybutyrylation prediction with protein language model. *International Journal of Biological Macromolecules*, 280:135741, 2024.
119. Heba M Elreify, Fathi E Abd El-Samie, Moawad I Dessouky, Hanaa Torkey, Said E El-Khamy, and Wafaa A Shalaby. Bigkbhb: a bi-directional gated recurrent unit model for predicting lysine β-hydroxybutyrylation sites. *BMC Genomics*, 27: 102, 2026.
120. Vikas Pejaver, Wei-Lun Hsu, Fuxiao Xin, A Keith Dunker, Vladimir N Uversky, and Predrag Radivojac. The structural and functional signatures of proteins that undergo multiple events of post-translational modification. *Protein Science*, 23(8):1077-1093, 2014.
121. Zhaohui Qin, Haoran Ren, Pei Zhao, Kaiyuan Wang, Huixia Liu, Chunbo Miao, Yanxiu Du, Junzhou Li, Liuji Wu, and Zhen Chen. Current computational tools for protein lysine acylation site prediction. *Briefings in Bioinformatics*, 25(6):bbae469, 2024.
122. Yu Liu, Minghui Wang, Jianing Xi, Fenglin Luo, and Ao Li. Ptm-ssmp: a web server for predicting different types of post-translational modification sites using novel site-specific modification profile. *International Journal of Biological Sciences*, 14(8):946, 2018.
123. Duolin Wang, Dongpeng Liu, Jiakang Yuchi, Fei He, Yuexu Jiang, Siteng Cai, Jingyi Li, and Dong Xu. Musitedeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Research*, 48(W1):W140-W146, 2020.
124. Md Al Mehedi Hasan, Shamim Ahmad, et al. mlysptmpred: multiple lysine ptm site prediction using combination of svm with resolving data imbalance issue. *Natural Science*, 10(09):370, 2018.
125. Sabit Ahmed, Afrida Rahman, Md Al Mehedi Hasan, Shamim Ahmad, and SM Shovan. Computational identification of multiple lysine ptm sites by analyzing the instance hardness and feature importance. *Scientific Reports*, 11(1):18882, 2021.
126. Yu Yan, Jyun-Yu Jiang, Mingzhou Fu, Ding Wang, Alexander R Pelletier, Dibakar Sigdel, Dominic CM Ng, Wei Wang, and Peipei Ping. Mind-s is a deep-learning prediction model for elucidating protein post-translational modifications in human diseases. *Cell Reports Methods*, 3(3), 2023.
127. Pedro Beltrao, Peer Bork, Nevan J Krogan, and Vera van Noort. Evolution and functional cross-talk of protein post-translational modifications. *Molecular Systems Biology*, 9(1):714, 2013.
128. Mario Leutert, Samuel W Entwistle, and Judit Villen. Decoding post-translational modification crosstalk with proteomics. *Molecular & Cellular Proteomics*, 20:100129, 2021.

 doi:10.20944/preprints202603.2316.v1

129. Pablo Minguez, Luca Parca, Francesca Diella, Daniel R Mende, Runjun Kumar, Manuela Helmer-Citterich, Anne-Claude Gavin, Vera Van Noort, and Peer Bork. Deciphering a global network of functionally associated post-translational modifications. *Molecular Systems Biology*, 8(1):599, 2012.
130. Yuanhua Huang, Bosen Xu, Xueya Zhou, Ying Li, Ming Lu, Rui Jiang, and Tingting Li. Systematic characterization and prediction of post-translational modification cross-talk. *Molecular & Cellular Proteomics*, 14(3):761-770, 2015.
131. Fei Zhu, Lei Deng, Yuhao Dai, Guangyu Zhang, Fanwang Meng, Cheng Luo, Guang Hu, and Zhongjie Liang. Ppict: an integrated deep neural network for predicting inter-protein ptm cross-talk. *Briefings in Bioinformatics*, 24(2):bbad052, 2023.
132. Hao Liu, Qingyong Hu, Yanni Ma, Kai Xie, and Yulan Guo. Deeppct: Single object tracking in dynamic point cloud sequences. *IEEE Transactions on Instrumentation and Measurement*, 72:1-12, 2022.
133. Sisi Ou, Wenhao Song, Jinru Li, Songye Gao, Yuxiang Ma, and Xiaohu Shi. Proxtalk: A pllm-driven dual-stream framework for inter-protein ptm crosstalk prediction. In *2025 IEEE International Conference on Bioinformatics and Biomedicine*, pages 339-344. IEEE, 2025.
134. David Ochoa, Andrew F Jarnuczak, Cristina Vieitez, Maja Gehre, Margaret Soucheray, André Mateus, Askar A Kleefeldt, Anthony Hill, Luz Garcia-Alonso, Frank Stein, et al. The functional landscape of the human phosphoproteome. *Nature Biotechnology*, 38(3):365-373, 2020.
135. Zhongjie Liang, Tonghai Liu, Qi Li, Guangyu Zhang, Bei Zhang, Xikun Du, Jingqiu Liu, Zhifeng Chen, Hong Ding, Guang Hu, et al. Deciphering the functional landscape of phosphosites with deep neural network. *Cell Reports*, 42(9), 2023.
136. Guangyu Zhang, Cai Zhang, Mingyue Cai, Cheng Luo, Fei Zhu, and Zhongjie Liang. Funcphos-str: An integrated deep neural network for functional phosphosite prediction based on alphafold protein structure and dynamics. *International Journal of Biological Macromolecules*, 266:131180, 2024.
137. Juan Xie, Ruihan Dong, Jintao Zhu, Haoyu Lin, Shiwei Wang, and Luhua Lai. Mmfuncphos: A multi-modal learning framework for identifying functional phosphorylation sites and their regulatory types. *Advanced Science*, 12(9):2410981, 2025.
138. Eric S Witze, William M Old, Katheryn A Resing, and Natalie G Ahn. Mapping protein post-translational modifications with mass spectrometry. *Nature Methods*, 4(10):798-806, 2007.
139. Xinjia Zhao, Haijuan Qin, Mingliang Tang, Xiaoyu Zhang, and Guangyan Qing. Nanopore: Emerging for detecting protein post-translational modifications. *TrAC Trends in Analytical Chemistry*, page 117658, 2024.
140. Malcolm W MacArthur, Paul C Driscoll, and Janet M Thornton. Nmr and crystallography—complementary approaches to structure determination. *Trends in Biotechnology*, 12(5):149-153, 1994.
141. Isabell Bludau, Sander Willems, Wen-Feng Zeng, Maximilian T Strauss, Fynn M Hansen, Maria C Tanzer, Ozge Karayel, Brenda A Schulman, and Matthias Mann. The structural context of posttranslational modifications at a proteome-wide scale. *PLoS Biology*, 20(5):e3001636, 2022.
142. Jacinto Lopez-Sagaseta and Alejandro Urdiciain. Severe deviation in protein fold prediction by advanced ai: a case study. *Scientific Reports*, 15(1):4778, 2025.
143. Alessio Del Conte, Mahta Mehdiabadi, Adel Bouhraoua, Alexander Miguel Monzon, Silvio CE Tosatto, and Damiano Piovesan. Critical assessment of protein intrinsic disorder prediction (caid)-results of round 2. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1925-1934, 2023.
144. Krishna DB Anapindi, Elena V Romanova, Bruce R Southey, and Jonathan V Sweedler. Peptide identifications and false discovery rates using different mass spectrometry platforms. *Talanta*, 182:456-463, 2018.
145. Ernesto S Nakayasu, Marina Gritsenko, Paul D Piehowski, Yuqian Gao, Daniel J Orton, Athena A Schepmoes, Thomas L Fillmore, Brigitte I Frohnert, Marian Rewers, Jeffrey P Krischer, et al. Tutorial: best practices and considerations for mass-spectrometry-based protein biomarker discovery and validation. *Nature Protocols*, 16(8):3737-3760, 2021.
146. Huanhuan Gao, Yi Zhu, Dongxue Wang, Zongxiang Nie, He Wang, Guibin Wang, Shuang Liang, Yuting Xie, Yingying Sun, Wenhao Jiang, et al. idia-qc: Ai-empowered data-independent acquisition mass spectrometry-based quality control. *Nature Communications*, 16(1):892, 2025.
147. Lijun Dou, Fenglong Yang, Lei Xu, and Quan Zou. A comprehensive review of the imbalance classification of protein post-translational modifications. *Briefings in Bioinformatics*, 22(5):bbab089, 2021.
148. Sedir Mohammed, Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch. The effects of data quality on machine learning performance on tabular data. *Information Systems*, 132:102549, 2025.

 doi:10.20944/preprints202603.2316.v1

149. Shrouk El-Amir and Ibrahim El-Henawy. An improved model using oversampling technique and cost-sensitive learning for imbalanced data problem. *Information Sciences with Applications*, 2:33-50, 2024.
150. Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31-57, 2018.
151. Carl O Retzlaff, Alessa Angerschmid, Anna Saranti, David Schneeberger, Richard Roettger, Heimo Mueller, and Andreas Holzinger. Post-hoc vs ante-hoc explanations: xai design guidelines for data scientists. *Cognitive Systems Research*, 86:101243, 2024.
152. Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. Explainable ai methods-a brief overview. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pages 13-38. Springer, 2020.
153. Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138-52160, 2018.
154. Gary S Collins, Karel GM Moons, Paula Dhiman, Richard D Riley, Andrew L Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B Reitsma, Maarten Van Smeden, et al. Tripod+ ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *British Medical Journal*, 385, 2024.
155. Haoyun Song, Rong Shen, Xiangwen Liu, Xuguang Yang, Kun Xie, Zhao Guo, and Degui Wang. Histone post-translational modification and the dna damage response. *Genes & Diseases*, 10(4):1429-1444, 2023.
156. Ming He, Hou-Hsien Chiang, Hanzhi Luo, Zhifang Zheng, Qi Qiao, Li Wang, Mingdian Tan, Rika Ohkubo, Wei-Chieh Mu, Shimin Zhao, et al. An acetylation switch of the nlrp3 inflammasome regulates aging-associated chronic inflammation and insulin resistance. *Cell Metabolism*, 31(3):580-591, 2020.
157. Junhong Qin, Xinhe Huang, Shengsong Gou, Sitao Zhang, Yujie Gou, Qian Zhang, Hongyu Chen, Lin Sun, Miaomiao Chen, Dan Liu, et al. Ketogenic diet reshapes cancer metabolism through lysine β-hydroxybutyrylation. *Nature Metabolism*, 6(8):1505-1528, 2024.
158. Xiang Wang, Weikang Deng, Zhenyu Meng, and Dewang Chen. Hybrid-attention mechanism based heterogeneous graph representation learning. *Expert Systems with Applications*, 250:123963, 2024.
159. Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations. In *International Conference on Machine Learning*, pages 12241-12252. PMLR, 2021.
160. Bo Wen, Chenwei Wang, Kai Li, Ping Han, Matthew V Holt, Sara R Savage, Jonathan T Lei, Yongchao Dou, Zhiao Shi, Yi Li, et al. Deepmvp: deep learning models trained on high-quality data accurately predict ptm sites and variant-induced alterations. *Nature Methods*, 22(9):1857-1867, 2025.
161. Xudong Wang, Linjie Li, Xuanjing Shen, and Xiaojie Lu. Rational design strategies in dna-encoded libraries for drug discovery. *Angewandte Chemie International Edition*, 64(34):e202511839, 2025.
162. Ibomoiye Domor Mienye, Theo G Swart, and George Obaido. Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information*, 15(9):517, 2024.
163. Shuang Zhang, Rui Fan, Yuti Liu, Shuang Chen, Qiao Liu, and Wanwen Zeng. Applications of transformer-based language models in bioinformatics: a survey. *Bioinformatics Advances*, 3(1):vbad001, 2023.