Genomics, Proteomics & Bioinformatics, 2024, **22(3)**, qzae032
https://doi.org/10.1093/gpbjnl/qzae032
Advance access publication: 10 April 2024
Database

Nphos: Database and Predictor of Protein *N*-phosphorylation

Ming-Xiao Zhao [1,2], Ruo-Fan Ding [3], Qiang Chen [4], Junhua Meng [5], Fulai Li [1], Songsen Fu [1], Biling Huang [1], Yan Liu [2], Zhi-Liang Ji [3,*], Yufen Zhao [1,2,6,*]

[1]Institute of Drug Discovery Technology, Ningbo University, Ningbo 315211, China
[2]Department of Chemical Biology, Key Laboratory for Chemical Biology of Fujian Province, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, China
[3]State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Faculty of Medicine and Life Sciences, Xiamen University, Xiamen 361102, China
[4]Zhejiang Key Laboratory of Pathophysiology, Department of Biochemistry and Molecular Biology, Health Science Center, Ningbo University, Ningbo 315211, China
[5]BGI Genomics, BGI-Shenzhen, Shenzhen 518083, China
[6]Key Laboratory of Bioorganic Phosphorus Chemistry & Chemical Biology, Department of Chemistry, Tsinghua University, Beijing 100084, China

*Corresponding authors: yfzhao@xmu.edu.cn (Zhao Y), appo@xmu.edu.cn (Ji ZL).

Handling Editor: Yu Xue

Abstract

Protein *N*-phosphorylation is widely present in nature and participates in various biological processes. However, current knowledge on *N*-phosphorylation is extremely limited compared to that on *O*-phosphorylation. In this study, we collected 11,710 experimentally verified *N*-phosphosites of 7344 proteins from 39 species and subsequently constructed the database Nphos to share up-to-date information on protein *N*-phosphorylation. Upon these substantial data, we characterized the sequential and structural features of protein *N*-phosphorylation. Moreover, after comparing hundreds of learning models, we chose and optimized gradient boosting decision tree (GBDT) models to predict three types of human *N*-phosphorylation, achieving mean area under the receiver operating characteristic curve (AUC) values of 90.56%, 91.24%, and 92.01% for pHis, pLys, and pArg, respectively. Meanwhile, we discovered 488,825 distinct *N*-phosphosites in the human proteome. The models were also deployed in Nphos for interactive *N*-phosphosite prediction. In summary, this work provides new insights and points for both flexible and focused investigations of *N*-phosphorylation. It will also facilitate a deeper and more systematic understanding of protein *N*-phosphorylation modification by providing a data and technical foundation. Nphos is freely available at http://www.bio-add.org/Nphos/ and http://ppodd.org.cn/Nphos/.

Key words: *N*-phosphorylation; Post-translational modification; Machine learning; Database; Benchmark dataset.

Introduction

Protein *N*-phosphorylation is a natural form of protein phosphorylation [1], in which the phosphate group attacks arginine guanidine (pArg), lysine amino (pLys), and histidine imidazole nitrogen (pHis) to form phosphoramidite bonds. Accumulated evidence has confirmed that protein *N*-phosphorylation plays critical roles in central metabolism [2], chemotaxis regulation [3], aerobic/anaerobic regulation [4], sporogenesis [5], and cell differentiation [6] via the two-component signaling in prokaryotes [7]. Recently, an association of protein *N*-phosphorylation with cancer progression has been discovered [8–10]. However, compared to *O*-phosphorylation (phosphoserine, phosphothreonine, and phosphotyrosine) research, current research on *N*-phosphorylation remains stagnant because of the instability of *N*-phosphates [1] and lack of advancement in detection methods.

Usually, both low-throughput ^{32}P-labeling [11] and high-throughput mass spectrometry (MS) [12] are applied to detect protein *N*-phosphorylation. However, both methods are laborious, time-consuming, and sometimes unstable. The recent introduction of specific antibodies [13,14] and peptide enrichment methods in MS [15–19] largely increased the reliability and efficiency of *N*-phosphosite detection in proteins. Currently, more than 60 protein phosphorylation-related databases have been developed [20], but only several databases, such as UniProt, PhosphoSitePlus [21], dbPTM [22], iPTMnet [23], and dbPSP [24], provide sporadic information on *N*-phosphosites in addition to *O*-phosphosites. Among these databases, PhosphoSitePlus, dbPTM, and iPTMnet are multifunctional comprehensive databases widely used by researchers. In addition to providing phosphosite information, PhosphoSitePlus also provides information on upstream and downstream kinases, phosphatases, and antibodies, as well as the biological processes of phosphosite. In addition to collecting protein phosphosite information, dbPTM includes information on more than 70 other types of post-translational modifications (PTMs). The iPTMnet database contains information on various PTMs. It relies on Protein Information Resources (PIR) (https://proteininformationresource.org/). HisPhosSite was recently introduced to collect 554 experimentally verified and 15,378

Received: 28 December 2022; **Revised:** 3 March 2024; **Accepted:** 1 April 2024.

predicted pHis sites [25], and it is till date the only database providing information on *N*-phosphorylation. However, HisPhosSite does not provide information on pArg and pLys sites.

According to our thorough survey, there were few *in silico* tools specifically developed for the prediction of *N*-phosphosites. Most tools were designed for *O*-phosphorylation, which are mostly kinase-specific [20,26,27], making them inapplicable for *N*-phosphosite prediction because only a few *N*-phosphorylation-specific kinases have been discovered. For instance, McsB was the first catalytic kinase discovered in *Bacillus subtilis* for protein pArg sites [28]. Subsequently, two protein histidine kinases (NME1 and NME2) and several protein histidine phosphatases (such as PHPT1 and LHPP), as well as a handful of substrates, were identified in mammals [29]. In recent years, three models pHisPred [30], PROSPECT [31], and iPhosH-PseAAC [32] have been trained for the large-scale prediction of pHis sites based on 487, 242, and 602 pHis sites, respectively. However, because eukaryotes and prokaryotes might not share common mechanisms of *N*-phosphorylation, implementation of these models for eukaryotic pHis prediction is questionable. Doubtlessly, before technical breakthroughs in experimental methods, computational solutions would serve as applicable, cost effective, and efficient solutions to the large-scale study of protein *N*-phosphorylation.

For this purpose, we created a thorough collection of experimentally verified *N*-phosphosites from public resources in this study. Based on the data, we constructed machine learning models for the human proteome-wide prediction of protein *N*-phosphorylation. Finally, a novel database was constructed to share comprehensive information on protein *N*-phosphorylation. We anticipate that this work will serve as a preferred source for protein *N*-phosphorylation studies of various scales.

Results

Large-scale analysis of protein *N*-phosphorylation

After a thorough literature search, MS processing, and data integration, we eventually collected 11,710 non-redundant *N*-phosphosites in 7344 proteins of 39 species, the majority (approximately 76.67%) of which were determined from human cellular experiments. By amino acid, the protein *N*-phosphosites consisted of 2690 pHis, 4469 pLys, and 4551 pArg sites (**Figure 1A**). By domain of life, 9101 *N*-phosphosites were identified in 5893 eukaryotic proteins, and 2609 sites were identified in 1451 prokaryotic proteins. More statistical information on *N*-phosphosites is presented in Figure 1A.

Subject to the data, *N*-phosphorylation widely occurred in membrane, cytoplastic, and nuclear proteins (Figure 1B). In addition, these proteins covered almost all PANTHER 15 protein categories of scaffold proteins, transcription factors, kinases, membrane-traffic proteins, RNA-binding proteins, and other proteins (Figure 1B). Further Gene Ontology (GO) enrichment analysis revealed that *N*-phosphoproteins ubiquitously participated in various biological processes, especially neutrophil degranulation, apoptotic regulation, protein deubiquitination, cell migration, and RNA splicing (Figure 1B).

In addition, we compared *N*-phosphorylation and *O*-phosphorylation on the proteome scale in humans. The *O*-phosphorylation data were integrated from PhosphoSitePlus [21], iPTMnet [23], and dbPTM [33], and they eventually consisted of 299,033 non-redundant experimentally verified *O*-phosphosites in 23,660 proteins. The number of *O*-phosphosites was approximately 32.31-fold higher than that of known *N*-phosphosites in humans. The average modification abundance per protein was 12.64 for *O*-phosphorylation *vs.* 1.55 for *N*-phosphorylation in humans. Similarly, as noted for *O*-phosphorylation, most proteins only had a single *N*-phosphorylation modification type (Figure 1C). Of 7344 *N*-phosphoproteins, 4373 also underwent *O*-phosphorylation.

Taken together, we conclude that *N*-phosphorylation ubiquitously occurs to almost all protein types and participates in diverse biological processes. Compared to *O*-phosphorylation, protein *N*-phosphorylation is extremely under-detected.

Sequential and structural characterization of *N*-phosphorylation

In general, *N*-phosphoproteins do not show significant tendency regarding protein length, and phosphorylation was not correlated with amino acid usage (Figures S1 and S2). Pattern analysis of 11,710 non-redundant *N*-phosphosite-centered 31 residue peptides helped to characterize the conserved sequential motifs by which kinases recognize the modification site (**Figure 2A**, Figure S3). In eukaryotes, lysine (K) and glutamic acid (E) were enriched near pHis; in particular, K was over-represented at the upstream flank. A particular sequential motif was significantly detected in 2120 pHis sites. For pLys, proline (P) was enriched upstream of phosphosites. An SP-rich motif was observed at the proximal upstream of pLys. Meanwhile, P residues also frequently appeared near pArg. SP or SPS motifs were consistently detected at both flanks of the pArg. By contrast, no significant amino acid preference or conserved motifs were found around the prokaryotic *N*-phosphosites excluding the comparative abundance of arginine (R) upstream of pHis and enrichment of glycine (G) adjacent to pArg (Figure S3). Noteworthy, both *N*-phosphorylation and *O*-phosphorylation are commonly present in the S/P-rich motifs [34], hinting that *N*-phosphorylation shares a similar kinase recognition mechanism with *O*-phosphorylation.

Furthermore, we characterized *N*-phosphorylation structurally. It was not surprising that more than 60% of *N*-phosphorylation events occurred in the exposed region of proteins, with few occurring in the buried region (Figure 2B); however, the preference for the exposed region did not differ from that of non-phosphorylated His/Lys/Arg [enrichment ratio (E-ratio) = 1.0106, $P = 0.2909$]. Meanwhile, *O*-phosphorylation more frequently occurred in the coiled region (E-ratio = 1.2731, $P = 2.20E{-}16$) than non-*O*-phosphosites, and the same was noted for *N*-phosphorylation (E-ratio = 1.0550, $P = 2.24E{-}09$) (Figure 2B). Critically, approximately 70% of *N*-phosphorylation occurred in the ordered region, in line with the distribution of non-phosphorylated His/Lys/Arg. This was significantly different from *O*-phosphorylation, which preferred the disordered region over non-phosphorylated Ser/Thr/Tyr (Figure 2B). Moreover, *N*-phosphorylation, similarly to *O*-phosphorylation, had no strong preference for different protein segments, such as the *N*-terminal, the C-terminal, or the middle regions (Figure 2B).

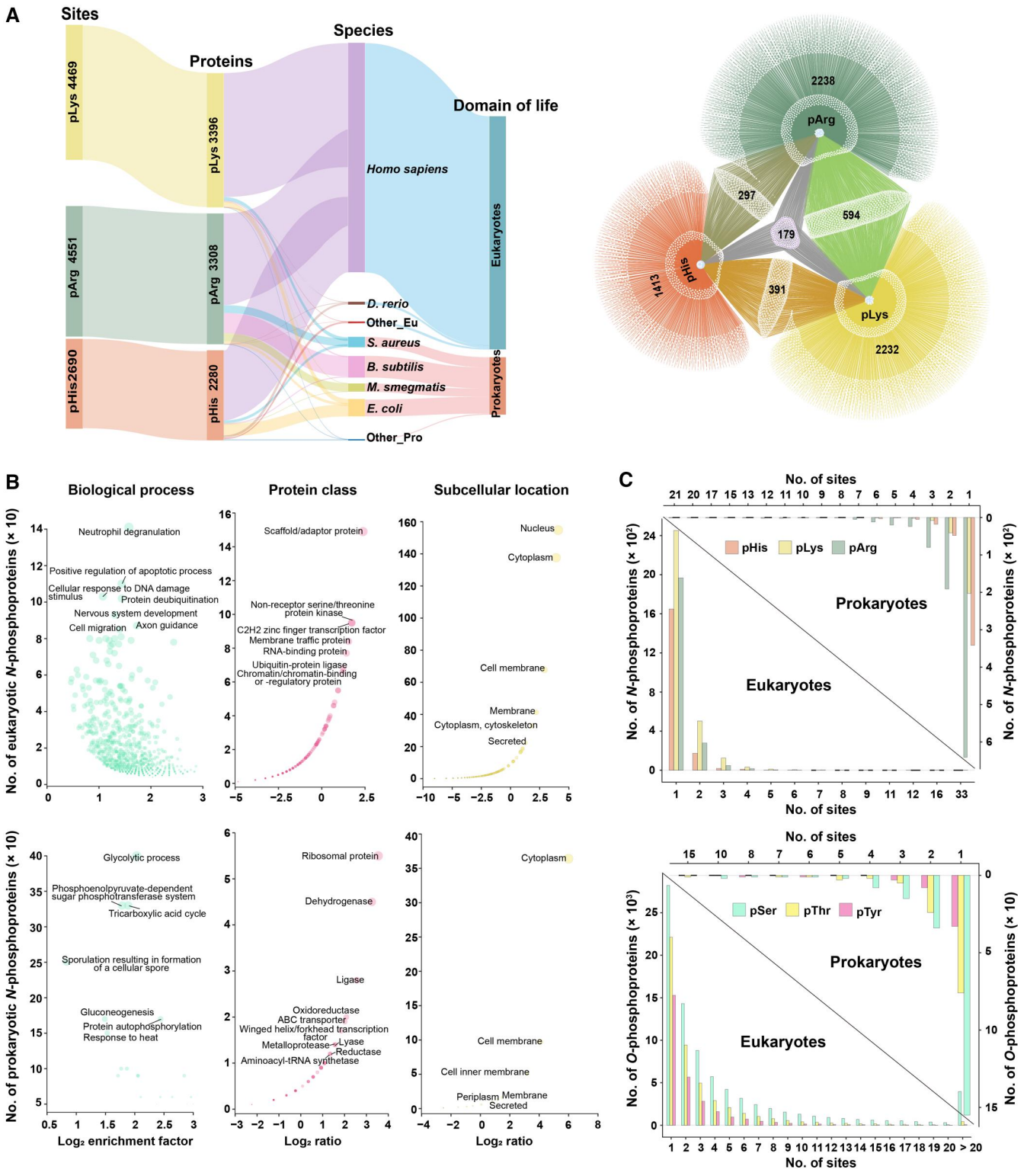

Figure 1 Statistics of experimentally verified protein *N*-phosphorylation

A. Distribution of pHis/pLys/pArg sites by life domain, species, protein, and phosphosite and the Venn network of three *N*-phosphorylation types. **B**. Functional enrichment analyses of *N*-phosphorylated proteins. The protein class analysis was performed using PANTHER 15. The GO and subcellular location analyses were subject to UniProt annotation. Ratio represents the ratio of *N*-phosphoproteins to the total number of proteins in these species, within a specific classification. **C**. Statistics of *O*-phosphorylation and *N*-phosphorylation by proteins or sites. pArg, the phosphate group attacks arginine guanidine; pLys, the phosphate group attacks lysine amino; pHis, the phosphate group attacks histidine imidazole nitrogen; GO, Gene Ontology; pSer, phosphoserine; pThr, phosphothreonine; pTyr, phosphotyrosine; Eu, Eukaryotes; Pro, Prokaryotes; *D. rerio*, *Danio rerio*; *B. subtilis*, *Bacillus subtilis*; *S. aureus*, *Staphylococcus aureus*; *E. coli*, *Escherichia coli*; *M. smegmatis*, *Mycolicibacterium smegmatis*.

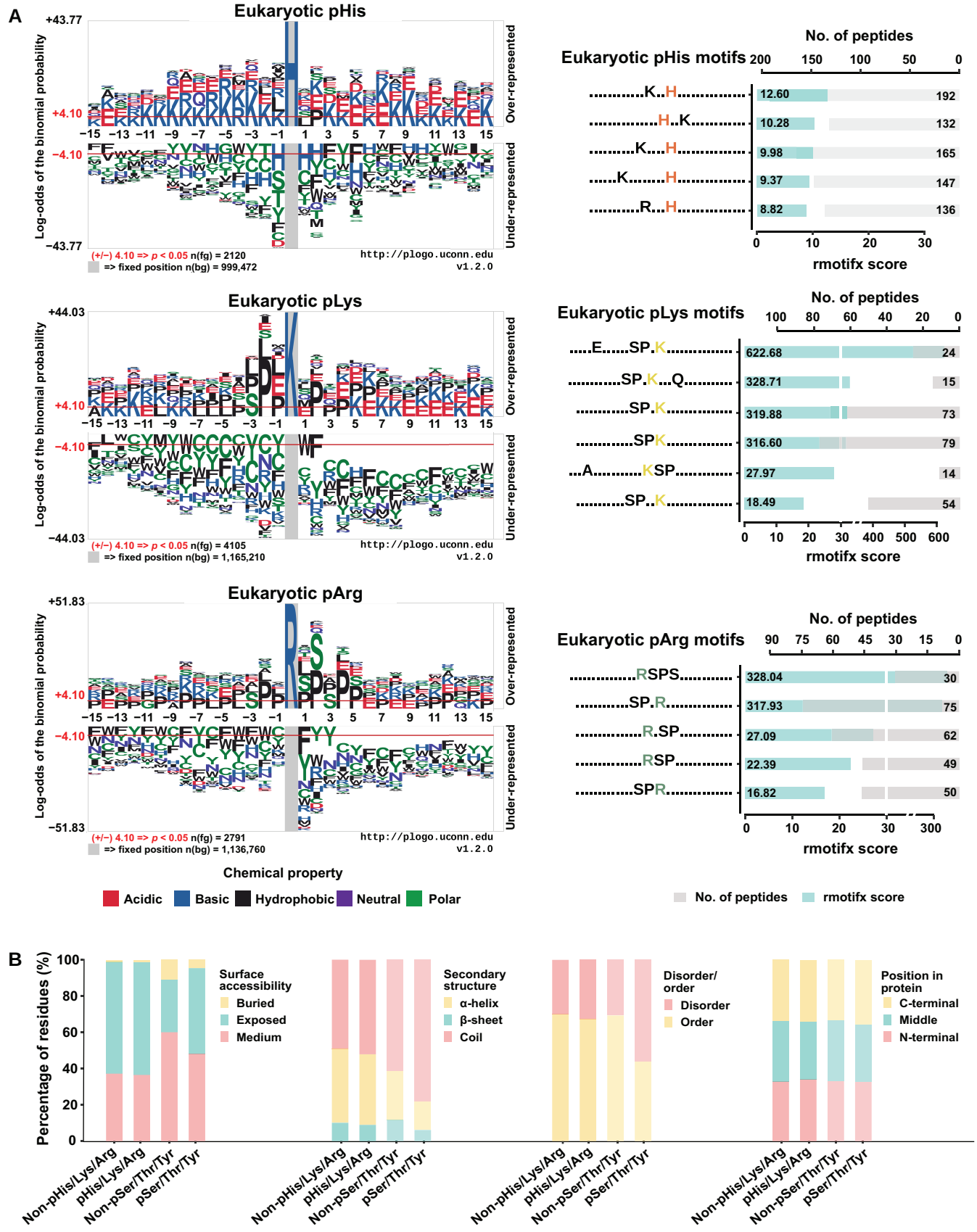

Figure 2 Sequential and structural characteristics of *N*-phosphorylation

A. Sequential features of pHis/pLys/pArg modifications. The conserved motifs were detected with the R package rmotifx [score = $\sum -\log(P)$; score ≥ 5, occurrence ≥ 10]. Logs are calculated in base 10. **B**. The structural propensity analyses of *N*-phosphorylation.

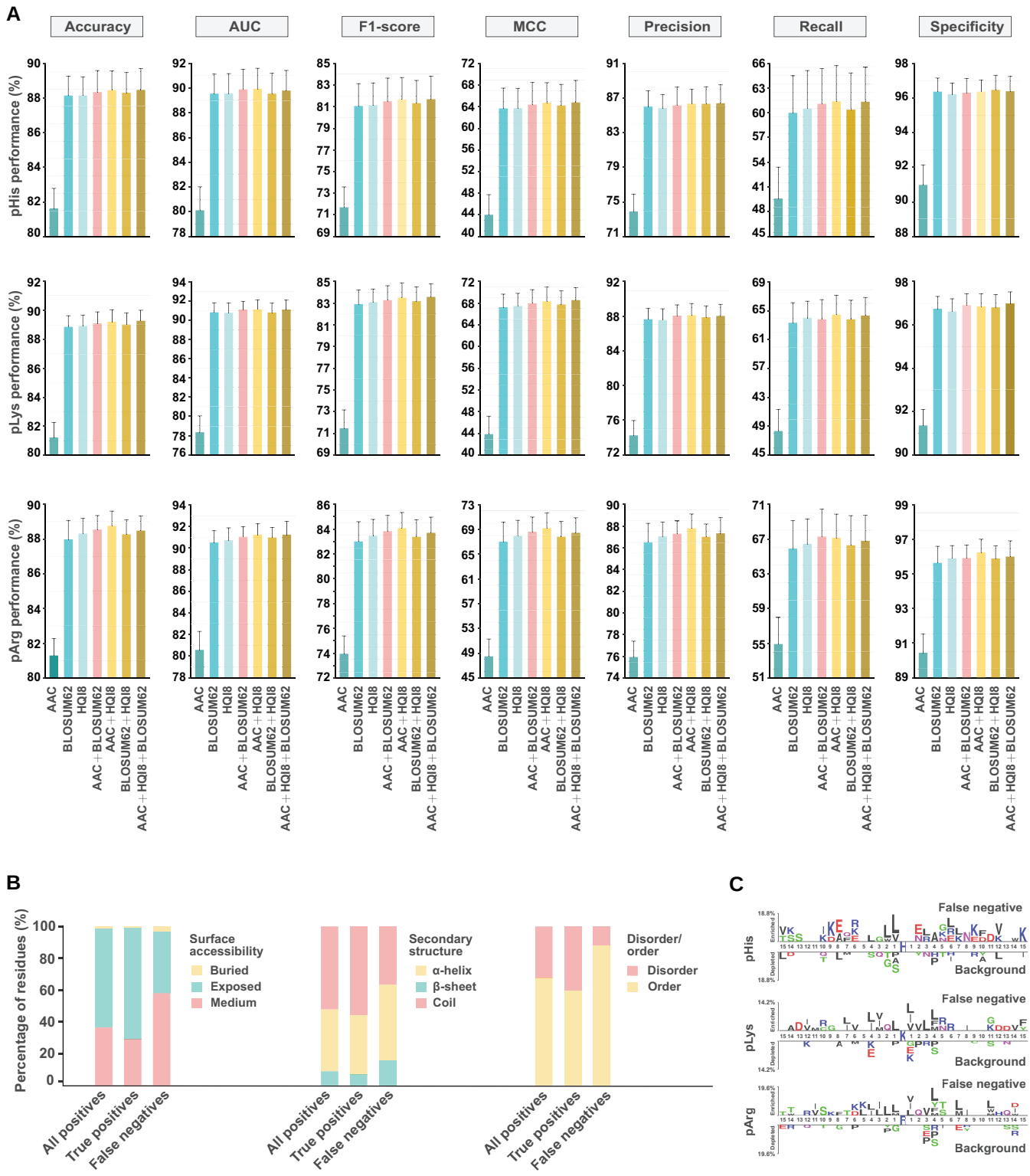

Figure 3 Model optimization and evaluation

A. Selection of optimal model combinations for the prediction of pHis, pLys, and pArg. The GBDT algorithm with the best overall performance was selected. **B**. The structural propensity of all positives, true positives, and false negatives. **C**. The sequence pattern of the false negatives. AAC, amino acid composition; HQI8, high quality indices; MCC, Matthews correlation coefficient; AUC, area under the ROC curve; ROC, receiver operating characteristic; GBDT, gradient boosting decision tree.

Large-scale prediction of human *N*-phosphorylation

Selection of algorithms and feature vectors

To choose the optimal combination of algorithms and feature vectors, we constructed 70 distinct models for each *N*-phosphorylation type separately. The model performance was evaluated and compared thoroughly, and the optimal combinations were ultimately determined. The gradient boosting decision tree (GBDT) algorithm outperformed other algorithms in the prediction of all three *N*-phosphorylation types. For feature vectors, pHis models preferred amino acid composition (AAC) + high quality indices (HQI8), pLys models preferred AAC + HQI8 + BLOSUM62, and pArg models preferred AAC + HQI8 (**Figure 3A**; Table S1). The exact reasons for the different preferences of feature vectors in the prediction of *N*-phosphorylation remained unclear. Accordingly, we re-constructed the GBDT models for *N*-phosphorylation prediction and repeated the hyperparameter optimization to achieve the best performance (Table S2).

Performance evaluation of the GBDT classifiers

Ten-fold cross-validation of the GBDT models based on five randomly shuffled training datasets consolidated the good performance regarding *N*-phosphorylation prediction, achieving mean area under the receiver operating characteristic (ROC) curve (AUC) values of 90.65%, 91.54%, and 91.78% in the prediction of pHis, pLys, and pArg, respectively (Table S3). The good model performance was further confirmed using the five randomly shuffled testing datasets, resulting in mean AUC values of 90.56%, 91.24%, and 92.01% in the prediction of pHis, pLys, and pArg, respectively (**Table 1**).

In addition, we examined the wrongly predicted data. Of all incorrect prediction events, 81.08% were false negatives (19.32%, 37.06%, and 24.70% for pHis, pLys, and pArg, respectively). Compared to all positives and true positives, a high proportion of the false negatives were located in the non-exposed, non-coiled, or ordered regions (Figure 3B), suggesting that they are the minor types of positives. Further sequence pattern analysis of these false negatives also failed to reveal a consistent sequence feature as that of the experimentally verified phosphosites (all positives) (Figures 2A and 3C). Therefore, we speculate that the positives are underrepresented during model learning. The inclusion of more types of positives in model training would significantly improve model performance.

Comparison with other pHis predictors

Previously, three predictors, iPhosH-PseAAC [32], PROSPECT [31], and pHisPred [30], were developed for the prediction of protein pHis sites. PROSPECT is an *Escherichia coli*-specific pHis predictor; iPhosH-PseAAC is a general pHis predictor; pHisPred is a eukaryote-specific and prokaryote-specific pHis predictor; and Nphos is a human-specific pHis/pLys/pArg predictor (**Table 2**). Nphos outperformed these predictors in general with a much better Matthews correlation coefficient (MCC) and F1-score. Although iPhosH-PseAAC deployed the deep learning algorithm (more precisely, multi-layered back propagation neural) to achieve a remarkable recall value (sensitivity for detecting positives) and MCC, the generality and robustness have not been properly evaluated. Importantly, Nphos was constructed on a larger (more than two folds) and more focused (only human pHis) positive dataset than the other predictors, guaranteeing better reliability. Involving all pHis events of different species for model learning could be a major drawback to the earlier predictors, in which the *N*-phosphorylation significantly differed between eukaryotes and prokaryotes regarding both sequential and structural aspects, as we depicted. Other than the exclusive strategy in generating the negatives used in all predictors, Nphos provided an additional constraint on solvent accessibility [setting the relative solvent accessibility (RSA) threshold] of non-pHis and pHis subject to the large-scale phosphorylation characterization, which ensured that the negatives were more interpretable and closer to the actual situation.

Table 1 Performance of GBDT models in the testing datasets

Performance (%)	pHis	pLys	pArg
Accuracy	89.16 ± 0.51	89.06 ± 0.52	88.99 ± 0.52
AUC	90.56 ± 0.40	91.24 ± 0.60	92.01 ± 0.52
F1-score	82.55 ± 0.71	83.13 ± 0.84	83.91 ± 1.03
MCC	66.78 ± 1.28	67.98 ± 1.59	69.43 ± 1.88
Precision	88.00 ± 0.67	88.54 ± 0.88	88.88 ± 0.75
Sensitivity/recall	61.51 ± 1.51	62.70 ± 1.73	64.81 ± 2.28
Specificity	97.17 ± 0.28	97.27 ± 0.41	97.20 ± 0.29

Note: pHis, pLys, and pArg indicate phosphorylation of histidine, lysine, and arginine, respectively. AUC, area under the ROC curve; ROC, receiver operating characteristic; MCC, Matthews correlation coefficient; GBDT, gradient boosting decision tree.

Proteome-wide prediction of human *N*-phosphorylation

Using the well-trained GBDT classifiers, we performed the large-scale prediction of *N*-phosphorylation in the human proteome. The human proteome sequences were derived from the UniProt Knowledgebase (as of February 25, 2022). To obtain highly reliable *N*-phosphosites, a probability threshold of 85% was set. Excluding the experimentally validated sites, 488,825 distinct *N*-phosphosites were predicted in 20,259 human proteins, including 64,409 pHis (15,192 proteins), 214,679 pLys (19,025 proteins), and 209,737 pArg sites (19,364 proteins). This increased the *N*-phosphorylation modification percentages against the background amino acid usage to 21.61%, 32.97%, and 32.75% for pHis, pLys, and pArg, respectively (**Table 3**). These values were comparable to those of experimentally validated *O*-phosphorylation modifications in humans. Importantly, approximately 74.40% of the experimentally verified *N*-phosphoproteins and 87.12% of the predicted *N*-phosphoproteins also underwent *O*-phosphorylation, implying possible crosstalk between *O*-phosphorylation and *N*-phosphorylation.

Nphos, a novel web service for illustrating and predicting protein *N*-phosphorylation

Retrieval of protein *N*-phosphorylation data

This study developed the novel database Nphos (http://www.bio-add.org/Nphos/ or http://ppodd.org.cn/Nphos/) to provide comprehensive information on experimentally verified protein *N*-phosphosites (**Figure 4A**). Nphos offers a foolproof keyword search for rapid data retrieval. The search is initiated by inputting a gene symbol, protein name, or UniProt accession number (AC) on the home page. The hits for the keyword search, if any, are listed by modified positions in descending order, along with the particulars of modified proteins such as UniProt AC, protein name, species, and gene symbol. Clicking the UniProt AC will lead to the detailed information page on *N*-phosphorylation, which is

Table 2 Comparison of models in predicting pHis sites

Model	Species type	Dataset			Window (AA)	Algorithm	Performance			
		Positive	Negative	Test			AUC	Recall	MCC	F1-score
pHisPred	Eu and Pro	Eu: 151 Pro: 336	Eu: 2061 (non-pHis) Pro: 1726 (non-pHis)	20%	31	SVM	Eu: 0.82 Pro: 0.80	–	–	Eu: 0.40 Pro: 0.46
PROSPECT	*E. coil*	244	1420 (non-pHis)	10%	27	CNN	0.821	0.48	0.37	–
iPhosH-PseAAC	General	602	720 (non-pHis)	–	41	MBPNN	–	0.9416	0.86	–
Nphos (this study)	Human	1945	6669 (non-pHis with RSA constraint)	20%	31	GBDT	0.9056	0.6151	0.6678	0.8255

Note: "Negative" and "Positive" refer to the numbers of negative datasets and positive datasets, respectively; "Test" refers to the percentage of the test dataset in the total dataset; "Window" indicates the length of the sequence window for extracting features. AA, amino acid; Eu, Eukaryotes; Pro, Prokaryotes; *E. coil*, *Escherichia coli*; CNN, convolutional neural network; MBPNN, multi-layered back propagation neural; SVM, support vector machine; –, not available.

Table 3 Comparison of phosphorylation ratios in the human proteome

Type	Amino acid	No. of phosphosites	Total No. of residues	Ratio (%)
N-phosphorylation	His	64,409	298,024	21.61
	Lys	214,679	651,188	32.97
	Arg	209,737	640,398	32.75
O-phosphorylation	Ser	179,612	946,894	18.97
	Thr	74,518	608,694	12.24
	Tyr	45,816	302,745	15.13

Note: *N*-phosphosites were predicted by the GBDT models, while the *O*-phosphosites were experimentally verified sites derived from multiple sources described in the main text.

organized into three sections. The "Protein information" section contains information about the protein, including the gene symbol, protein length, motif, function, GO category, subcellular location, sequence, and the crosslink to the Protein Data Bank (PDB) or AlphaFold Protein Structure Database. The motif information can provide clues for the development of *N*-phospho-specific analogs and antibodies. The information on function, GO category, and subcellular location will provide a functional understanding of protein *N*-phosphorylation. The "*N*-phosphosites" section lists the details of *N*-phosphosites along with the source or literature. The "Other PTM sites" section lists all experimentally verified PTM sites, if available, in the protein, including *O*-phosphorylation, acetylation, ubiquitination, and methylation. These PTMs are helpful for illustrating the possible crosstalk with *N*-phosphorylation. The Nphos data are freely available and downloadable by species via the download page.

Interactive prediction of protein *N*-phosphorylation modification

The optimized GBDT models were implemented in Nphos to permit the interactive prediction of *N*-phosphorylation from the primary protein sequence (Figure 4B). To initiate the prediction, the user is asked to input one or more protein sequence(s) in FASTA format (all query sequences must have more than 16 residues) and select a combination option of *N*-phosphorylation types: pHis, pLys, and pArg. Proceeding with the prediction will list the potential *N*-phosphosites, which satisfy the prediction probability of > 0.50, by *N*-phosphorylation types in ascending order according to the residue positions in the protein sequence. According to the prediction probability, the prediction results are roughly divided into three reliability grades: high ($0.85 <$ probability ≤ 1.00), middle ($0.70 <$ probability ≤ 0.85), and low ($0.50 <$ probability ≤ 0.70). The prediction result can be downloaded in a "txt" format. The benchmark datasets used for model construction in this study are also acquirable via the download page to support further intelligent prediction.

Discussion

It has been more than a half-century since protein *N*-phosphorylation was first discovered. Regretfully, compared to protein *O*-phosphorylation or other PTMs such as ubiquitination, acetylation, methylation, and SUMOylation, the functional investigation of *N*-phosphorylation is limited by the technical difficulty in stably detecting *N*-phosphosites. However, continuous efforts have provided valuable data on *N*-phosphorylation. Unfortunately, these data remain out of use after publication, and few attempts have been made to collect these data and reuse them. HisPhosSite [25] is the first protein *N*-phosphorylation repository, and it contains 554 experimentally verified pHis sites in 411 proteins. These sites have been included in Nphos. Therefore, this study represents an important step forward by introducing the novel database Nphos, which thoroughly collects 11,710 experimentally verified *N*-phosphosites in 7344 proteins from 39 species, covering pHis, pLys, and pArg sites. Using this database, we conducted large-scale analyses to characterize the sequential and structural features of *N*-phosphorylation modifications. For instance, we discovered that SP-rich motifs might be required for pLys and pArg, a finding that has not been previously reported.

In this study, we prioritized the use of machine learning algorithms, namely GBDT-based models, to predict pHis, pLys, and pArg sites in the human proteome to extensively expand the knowledge of *N*-phosphorylation. These models are also deployed as an online service for the interactive prediction of *N*-phosphorylation from primary protein sequences. This substantially empowers both large-scale and focused functional studies of *N*-phosphorylation modification.

Doubtlessly, the analyses and predictions made in this study involve the experimentally verified *N*-phosphosites available to date, which remain limited even after thorough collection from public resources. The under-representation of *N*-phosphorylation types explains the comparatively low sensitivity of model prediction. Moreover, the generation of the negatives is crucial for decision-making models. Regretfully, this issue has not been fully solved in this study. The acquirement of additional experimentally validated *N*-phosphosites and the introduction of new learning algorithms such as

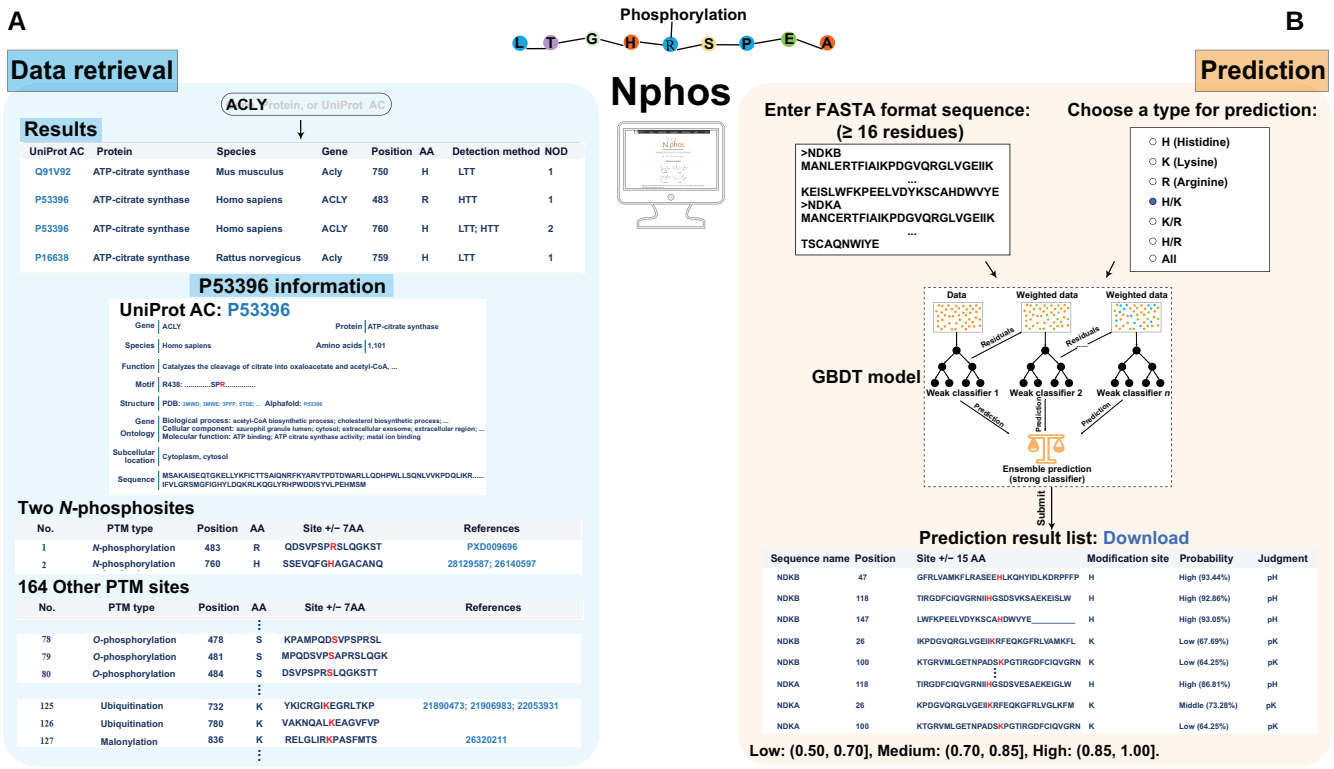

Figure 4 The web services of Nphos

A. Data retrieval of Nphos. **B**. Interactive *N*-phosphosite prediction. AC, accession number; AA, amino acid; PTM, post-translational modification; NOD, number of detections; HTT, high-throughout technology; LTT, low-throughout technology.

label-free learning would significantly improve the accuracy of *N*-phosphorylation prediction. Moreover, the model performance is also affected by the length of the peptide inputted. Some *N*-/*O*-phosphorylation predictors optimize model performance by selecting different peptide lengths [20]. For example, in pHisPred, different machine learning algorithms have different sensitivities to peptide lengths [30]. Therefore, the selection of the optimized sequence length for model construction is expected in the future.

Future development

As a practical matter, all studies on *N*-phosphorylation are limited to the currently identified phosphoproteins. We foresee enormous growth in available *N*-phosphorylation data in the future, highlighting the importance of discovery and better definition of putative *N*-phosphosites. In the future, we plan to expand the available data in Nphos, including information on kinases and phosphatases mediating modifications, protein–protein interactions, crosstalk, domains, disease relationships, and structure visualization. In addition, we will develop new tools for embedded motif analysis, batch searching function, and evolutionary analysis. Moreover, recent applications of deep learning algorithms in various prediction tasks of PTM sites [20,35] have inspired us to take full advantage of cutting-edge artificial intelligent technologies in the accurate prediction of *N*-phosphorylation.

Conclusion

This study provides the most fruitful information on protein *N*-phosphorylation available to date. The model prediction will guide the precise design of antibodies for validating particular *N*-phosphorylation events and further exploring their biological functions. Finally, this work will overcome current experimental constraints on flexible but focused protein *N*-phosphorylation research. In particular, it will assist in the discovery of kinases via studying the relationship between *N*-phosphorylation and *O*-phosphorylation.

Materials and methods

Data collection

Collection of *N*-phosphorylation modifications

We searched PubMed and bioRxiv using multiple keywords such as “pHis”, “pArg”, “pLys”, “protein histidine phosphorylation”, “protein arginine phosphorylation”, “protein lysine phosphorylation”, “phosphoarginine”, “arginine phosphorylation”, “lysine phosphorylation”, “non-canonical phosphorylation”, “phosphorylated lysine”, “*N*-phosphoproteome”, and “pHisphorylation”. The related articles were retrieved, and the relevance to *N*-phosphorylation was manually checked. In total, 115 articles were eventually retrieved as of November, 2021. According to the guidelines of the articles, the raw MS data were downloaded from the ProteomeXchange Consortium (http://www.proteomexchange.org/) [36] for later spectrum unscrambling and data analysis. The details of all raw MS data are presented in Table S4. Furthermore, we extracted the *N*-phosphosites from UniProt (as of June, 2020) by searching the “MOD_RES” field with keywords such as “phosphohistidine”, “phospholysine”, or “phosphoarginine”. Only the entries supported by the literature in PubMed were collected.

Proteome data processing

The raw MS files were processed using Proteome Discoverer software (v2.4; default parameters) to annotate the peptides via searching against combined forward/reversed databases of UniProt Swiss-Prot Proteomes (August, 2020), covering multiple species including *Homo sapiens*, *Danio rerio*, *B. subtilis*, *Staphylococcus aureus*, and *E. coli*. The parameters were set as follows: enzyme, trypsin; precursor mass tolerance, 20 ppm; fragment mass tolerance, 0.6 Da (Ion Trap) or 0.02 Da (Orbitrap); spectrum matching, b, c, y, z ions [electron-transfer/higher-energy collision dissociation (EThcD)], b, y ions [collision-induced dissociation (CID) or high energy collision dissociation (HCD)], c, z ions [electron transfer dissociation (ETD)]; dynamic modification, Phospho/+79.966 Da (H, K, R, S, T, Y); and false discovery rate (FDR) targets, 0.05. The phosphopeptides were detected using the module IMP-ptmRS with the default parameters. All *N*-phosphorylation data were manually checked and the repeated *N*-phosphosites were consolidated to eliminate redundancy and ambiguity.

The background dataset

We also extracted non-phosphorylated His, Arg, and Lys residues from all proteins of the aforementioned six species (Table S4) in UniProt. These non-phosphosites were taken as the background dataset to countermine the possible taxon bias of *N*-phosphorylation.

The data of other PTMs

Other than *N*-phosphorylation, comprehensive information on several PTMs such as *O*-phosphorylation, methylation, acetylation, and ubiquitination was also derived from PhosphoSitePlus [21], iPTMnet [23], and dbPTM [33]. These PTM sites were integrated, and their sequential coordinates were readjusted by referring to the UniProt protein sequences. The normalized PTM data were used as the reference dataset for characterizing *N*-phosphorylation.

Characterization of *N*-phosphosites

Phosphosite enrichment analysis

The E-ratio was calculated as follows:

$$E\text{-}ratio = \frac{m \times N}{M \times n} \quad (1)$$

where m is the number of phosphosites of a particular type, n is the number of non-phosphosites of a particular type, N is the number of all non-phosphosites, and M is the number of all phosphosites. The comparison of enrichment between phosphosites and non-phosphosites was performed using the chi-squared test in R 4.0.3.

Functional enrichment analysis

GO enrichment analysis was conducted using the R package "clusterProfiler" (pvalueCutoff = 0.05, pAdjustMethod = "BH"; BH means Benjamini and Hochberg) by taking the UniProt GO entries of all selected species with *N*-phosphosites as the background (the background dataset). The protein class analysis was performed with the PANTHER 15 (http://pantherdb.org/) [37].

Sequential and structural pattern analysis

The potential sequential motifs around the *N*-phosphosites were extracted using the R package "rmotifx" [38] (score ≥ 5, occurrence ≥ 10). A previous study manifested that the phosphorylation specificity was closely related to the primary sequence surrounding the phosphosite [39], in which the 14 upstream/downstream amino acids around the phosphosite were conserved [40]. Hence, a phosphosite-centered window of 31 continuous amino acids was adopted [41]. Furthermore, the R package "ggseqlogo" was used to generate the sequence logos, colored by residue physicochemical properties. The overall sequential patterns were visualized using the online service pLogo [42] (https://plogo.uconn.edu/). The aforementioned background dataset was used as the control for motif analysis.

The secondary structures and surface accessibility were predicted by SPOT-1D-Single [43] with the default parameters. The order and disorder regions were evaluated by IUPred3 [44] with the default parameters.

Dataset preparation for model construction

The positive dataset

The positives included all experimentally verified *N*-phosphosites in humans. Taking the phosphosite as the central residue, a peptide of 31 continuous amino acids was extracted as the positive. Redundant peptides were excluded. The positives consisted of 8978 distinct 31-residue peptides covering 5798 distinct proteins from the human proteome (**Figure 5**).

The negative dataset

In practice, the determination of non-phosphorylation remains an open question because of the limited number of experimental verified *N*-phosphosites (the positives). As an alternative solution, we generated the negative datasets for non-phosphorylated His, Lys, and Arg separately as follows. (1) The non-phosphorylated His/Lys/Arg residues in the *N*-phosphorylated proteins were selected. (2) The residues in the exposed regions, which were evaluated by the RSA, were excluded. The RSA threshold was determined subject to the dataset balance by amino acid, *i.e.*, RSA ≤ 0.12 for His, RSA ≤ 0.3 for Lys, and RSA ≤ 0.2 for Arg. (3) The negative peptides (the 31-residue sequence centered around non-phosphorylated His, Lys, or Arg) were generated. (4) The wrongly or ambiguously assigned negatives were removed by conducting a homology analysis of the negative sequences against the positive dataset. The homology analysis was performed using CD-HIT [45] by setting the sequence identity threshold to 30%. The information on the resulting negative datasets is given in **Table 4**.

Feature transformation

Every 31-residue peptide (positives and negatives) was converted into a $1 \times X$ feature vector before applying it to model construction (Figure 5). The feature vector was encoded using any combination of three different groups of features: AAC, amino acid index (AAindex)-HQI8, and BLOSUM62.

AAC, which indicates the frequencies of 20 normal amino acids in the peptide, can be calculated as follows:

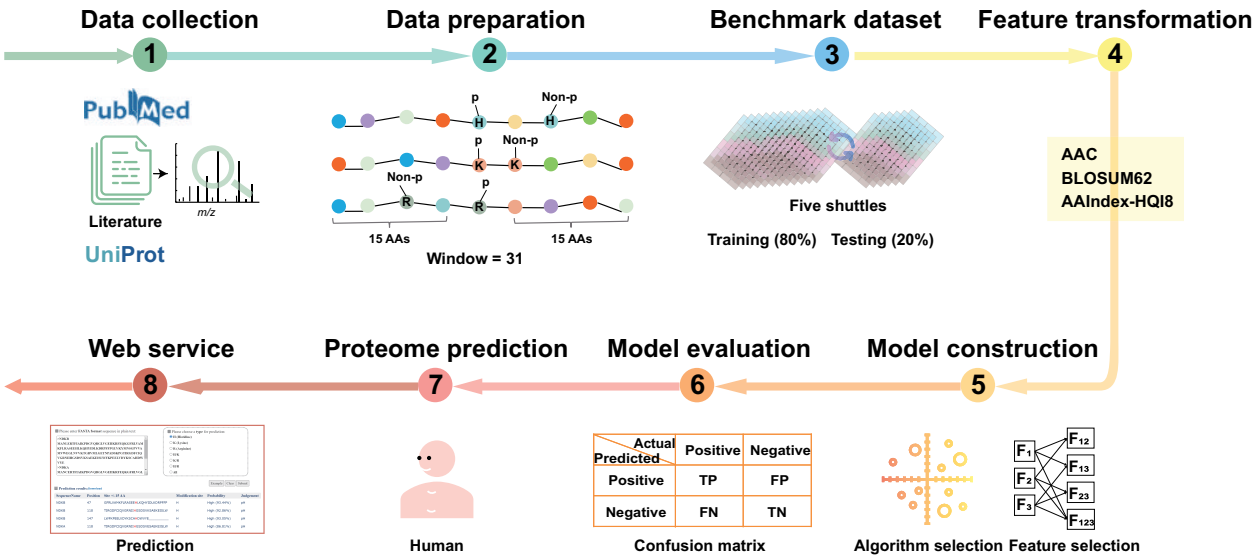

Figure 5 Scheme of the protein *N*-phosphosite predictor
AAindex, amino acid index; TP, true positive; FP, false positive; FN, false negative; TN, true negative; Window, the length of the sequence window for extracting features.

Table 4 Dataset preparation for machine learning

Type of *N*-phosphorylation	No. of positive datasets	No. of negative datasets	Ratio
pHis	1945	6669	1:3.43
pLys	4008	12,970	1:3.24
pArg	2723	7893	1:2.90

Note: Ratio refers to the number of positive datasets *vs.* the number of negative datasets.

$$P_x(a) = \frac{n_x(a)}{\sum_{a=1}^{20} n_x(a)} \quad (a = 1, 2, \cdots, 20) \tag{2}$$

where $n_x(a)$ denotes the number of amino acid x.

AAindex [46] (https://www.genome.jp/aaindex/) is a database containing the physical, chemical, and biological properties of 20 amino acids. At present, there are 566 properties for each amino acid. To overcome the overfitting problems caused by the high number of properties, Sahara et al. used the fuzzy clustering method to cluster and elect central properties, leading to development of the quality index—HQI8 [47]. In this study, we also used HQI8 to extract the physiochemical features of the peptides and further convert them into a 1×248 feature vector.

The BLOSUM62 matrix is a 20×21 dimension matrix of amino acid substitution that is conventionally used to measure the similarity of two peptide sequences. In the matrix, each entry is the logarithm of the odds score, which is found by dividing the frequency of occurrence of the amino acid pair by the likelihood of an alignment of the amino acids by random chance [48].

$$BLOSUM62 = \begin{pmatrix} a_{AA} & \cdots & a_{An} \\ \vdots & \ddots & \vdots \\ a_{mA} & \cdots & a_{mn} \end{pmatrix} \tag{3}$$
$$(n \in A, C, \cdots, W;\ m \in A, C, \cdots, W, -)$$

Each row of the BLOSUM62 matrix $\overrightarrow{R_m}$ represents the conserved substitution of an amino acid by other amino acids:

$$\overrightarrow{R_m} = (a_{mA}, a_{mC}, \cdots, a_{mn}) \quad (n \in A, C, \cdots, W;\ m \in A, C, \cdots, W, -) \tag{4}$$

Noteworthy, the BLOSUM62 matrix is composed of 20 canonical characters (corresponding to 20 amino acids) and one non-canonical character ("–", any one of 20 amino acids). In this manner, the 31-residue peptide was transformed into a 1×620 feature vector $\overrightarrow{F_{BLOSUM62}}$, as follows:

$$\overrightarrow{F_{BLOSUM62}} = (\overrightarrow{s_1}, \overrightarrow{s_2}, \cdots, \overrightarrow{s_{31}}) \tag{5}$$

where s stands for any of 20 amino acids and "–", and $\overrightarrow{s} = \overrightarrow{R_m}$ $(if\ s = m,\ m \in A, C, \cdots, W, -)$.

In this study, the AAC, AAindex-HQI8, and BLOSUM62 feature vectors were computed with self-coded Python scripts. The BLOSUM62 matrix was derived from the work of Henikoff and Henikoff [48]. Combinations of these three feature vectors were generated by connecting them to each other.

The algorithms for model construction

The machine learning models for *N*-phosphorylation prediction were constructed separately for pHis, pLys, and pArg. Overall, 10 different algorithms were adopted for model construction, including AdaBoost, artificial neural network (ANN), CatBoost, decision tree (DT), GBDT [49], k-nearest neighbor (KNN), logistic regression (LR), naive Bayes (NB), quadratic discriminant analysis (QDA), and random forest (RF). All algorithms excluding CatBoost were called via the Python package scikit-learn (v1.0.2). The CatBoost algorithm was called via the Python package catboost (v1.0.4). The calling methods are summarized in Table S5. Parameter refinement was performed using sklearn.model_selection. GridSearchCV (cv = 10).

Model training and performance evaluation

In this study, to determine the globally optimum condition for *N*-phosphorylation prediction, we performed thorough combinational learning based on seven combinations of feature vectors and ten learning algorithms (Figure 5). Consequently, 70 distinct models were constructed and compared for each of the three *N*-phosphorylation types (pHis, pLys, and pArg). The corresponding optimal hyperparameters for every machine learning algorithm are summarized in Table S5, when applicable.

For the construction of each model, the positives and negatives were randomly split into two parts. Specifically, 80% were used for model training and internal evaluation, and the remaining 20% were used for model validation (Figure 5). To enhance the model generalization, the same operation was repeated five times, and the mean performance was taken as the final performance of the model. Meanwhile, the internal evaluation was also performed via 10-fold cross-validation to consolidate the robustness, in which both the positives and negatives were randomly split into ten folds, including nine folds for model construction and one fold for internal evaluation. The 10-fold cross-validation was repeated ten times.

The model performance was evaluated by several parameters, including sensitivity or recall, specificity, precision, F1-score, accuracy, and MCC as follows:

$$Sensitivity\ or\ Recall = \frac{TP}{TP+FN} \tag{6}$$

$$Specificity = \frac{FN}{FP+TN} \tag{7}$$

$$Precision = \frac{TP}{TP+FP} \tag{8}$$

$$F1\text{-}score = \frac{2\times Precision\times Sensitivity}{Precision+Sensitivity} \tag{9}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{10}$$

$$MCC = \frac{TP\times TN - FN\times FP}{\sqrt{(TP+FN)\times(FN+FP)\times(TP+FP)\times(TN+FN)}} \tag{11}$$

where TP, TN, FP, and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively. Furthermore, the ROC curve was prepared, and the AUC was determined.

Database construction

The database was constructed on a system architecture of Linux + Tomcat + Java. MySQL was adopted as the underlying database management system. The web pages were coded with HTML5 technology to support mobile access. We tested the online service on a variety of internet browsers, including Mozilla Firefox, Google Chrome, and Microsoft Edge.

Data availability

Nphos is freely available at http://www.bio-add.org/Nphos/ and http://ppodd.org.cn/Nphos/.

CRediT author statement

Ming-Xiao Zhao: Conceptualization, Data curation, Methodology, Software, Formal analysis, Resources, Investigation, Resources, Validation, Writing – original draft, Visualization. **Ruo-Fan Ding:** Software. **Qiang Chen:** Validation, Writing – original draft. **Junhua Meng:** Software. **Fulai Li:** Resources. **Songsen Fu:** Resources. **Biling Huang:** Resources. **Yan Liu:** Resources. **Zhi-Liang Ji:** Conceptualization, Methodology, Resources, Writing – original draft, Supervision, Writing – review & editing. **Yufen Zhao:** Conceptualization, Supervision, Writing – review & editing. All authors have read and approved the final manuscript.

Supplementary material

Supplementary material is available at *Genomics, Proteomics & Bioinformatics* online (https://doi.org/10.1093/gpbjnl/qzae032).

Competing interests

Junhua Meng is a full-stack engineer of BGI Genomics Co., Ltd. All the other authors have declared no competing interests.

Acknowledgments

The work was supported by the National Key R&D Program of China (Grant No. 2020YFA0608300), the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences (Grant No. YYWT-0901-EXP-16), the Scientific Research Grant of Ningbo University (Grant No. 215-432000282), the Ningbo City Top Talent Project (Grant No. 215-432094250), and the National Natural Science Foundation of China (Grant Nos. 22107055 and 91856126).

ORCID

0000-0002-8821-3654 (Ming-Xiao Zhao)
0000-0003-1667-9866 (Ruo-Fan Ding)
0000-0003-2624-8904 (Qiang Chen)
0009-0001-9021-5114 (Junhua Meng)
0000-0002-5710-7795 (Fulai Li)
0000-0002-3524-8108 (Songsen Fu)
0000-0002-1060-8787 (Biling Huang)
0000-0001-6112-7898 (Yan Liu)
0000-0003-1190-7655 (Zhi-Liang Ji)
0000-0002-8513-1354 (Yufen Zhao)

References

[1] Albert S, Helmut EM. Phosphoamino acid analysis. Proteomics 2001;1:200–6.
[2] Schmidt A, Trentini DB, Spiess S, Fuhrmann J, Ammerer G, Mechtler K, et al. Quantitative phosphoproteomics reveals the role of protein arginine phosphorylation in the bacterial stress response. Mol Cell Proteomics 2014;13:537–50.
[3] Falke JJ, Bass RB, Butler SL, Chervitz SA, Danielson MA. The two-component signaling pathway of bacterial chemotaxis: a

molecular view of signal transduction by receptors, kinases, and adaptation enzymes. Annu Rev Cell Dev Biol 1997;13:457–512.
[4] Unden G, Bongaerts J. Alternative respiratory pathways of *Escherichia coli*: energetics and transcriptional regulation in response to electron acceptors. Biochim Biophys Acta 1997; 1320:217–34.
[5] Perego M. Kinase-phosphatase competition regulates *Bacillus subtilis* development. Trends Microbiol 1998;6:366–70.
[6] Ward MJ, Zusman DR. Regulation of directed motility in *Myxococcus xanthus*. Mol Microbiol 1997;24:885–93.
[7] Stock AM, Robinson VL, Goudreau PN. Two-component signal transduction. Annu Rev Biochem 2000;69:183–215.
[8] Zheng J, Dai X, Chen H, Fang C, Chen J, Sun L. Down-regulation of LHPP in cervical cancer influences cell proliferation, metastasis and apoptosis by modulating AKT. Biochem Biophys Res Commun 2018;503:1108–14.
[9] Hindupur SK, Colombi M, Fuhs SR, Matter MS, Guri Y, Adam K, et al. The protein histidine phosphatase LHPP is a tumour suppressor. Nature 2018;555:678–82.
[10] Lapek JD Jr, Tombline G, Kellersberger KA, Friedman MR, Friedman AE. Evidence of histidine and aspartic acid phosphorylation in human prostate cancer cells. Naunyn Schmiedebergs Arch Pharmacol 2015;388:161–73.
[11] Aponte AM, Phillips D, Harris RA, Blinova K, French S, Johnson DT, et al. ^{32}P labeling of protein phosphorylation and metabolite association in the mitochondria matrix. Methods Enzymol 2009; 457:63–80.
[12] Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol 2006; 24:1285–92.
[13] Kee JM, Oslund RC, Perlman DH, Muir TW. A pan-specific antibody for direct detection of protein histidine phosphorylation. Nat Chem Biol 2013;9:416–21.
[14] Fuhs SR, Meisenhelder J, Aslanian A, Ma L, Zagorska A, Stankova M, et al. Monoclonal 1- and 3-phosphohistidine antibodies: new tools to study histidine phosphorylation. Cell 2015; 162:198–210.
[15] Potel CM, Lin MH, Heck AJR, Lemeer S. Widespread bacterial protein histidine phosphorylation revealed by mass spectrometry-based proteomics. Nat Methods 2018;15:187–90.
[16] Hu Y, Weng Y, Jiang B, Li X, Zhang X, Zhao B, et al. Isolation and identification of phosphorylated lysine peptides by retention time difference combining dimethyl labeling strategy. Sci China Chem 2019;62:708–12.
[17] Fu S, Fu C, Zhou Q, Lin R, Ouyang H, Wang M, et al. Widespread arginine phosphorylation in human cells—a novel protein PTM revealed by mass spectrometry. Sci China Chem 2020;63:341–6.
[18] Hu Y, Jiang B, Weng Y, Sui Z, Zhao B, Chen Y, et al. Bis(zinc(II)-dipicolylamine)-functionalized sub-2 μm core-shell microspheres for the analysis of *N*-phosphoproteome. Nat Commun 2020;11:6226.
[19] Adam K, Fuhs S, Meisenhelder J, Aslanian A, Diedrich J, Moresco J, et al. A non-acidic method using hydroxyapatite and phosphohistidine monoclonal antibodies allows enrichment of phosphopeptides containing non-conventional phosphorylations for mass spectrometry analysis. bioRxiv 2019;691352.
[20] Zhao MX, Chen Q, Li F, Fu S, Huang B, Zhao Y. Protein phosphorylation database and prediction tools. Brief Bioinform 2023; 24:bbad090.
[21] Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res 2015;43:D512–20.
[22] Li Z, Li S, Luo M, Jhong JH, Li W, Yao L, et al. dbPTM in 2022: an updated database for exploring regulatory networks and functional associations of protein post-translational modifications. Nucleic Acids Res 2022;50:D471–9.
[23] Huang H, Arighi CN, Ross KE, Ren J, Li G, Chen SC, et al. iPTMnet: an integrated resource for protein post-translational modification network discovery. Nucleic Acids Res 2018; 46:D542–50.
[24] Shi Y, Zhang Y, Lin S, Wang C, Zhou J, Peng D, et al. dbPSP 2.0, an updated database of protein phosphorylation sites in prokaryotes. Sci Data 2020;7:164.
[25] Zhao J, Zou L, Li Y, Liu X, Zeng C, Xu C, et al. HisPhosSite: a comprehensive database of histidine phosphorylated proteins and sites. J Proteomics 2021;243:104262.
[26] Yang H, Wang M, Liu X, Zhao XM, Li A. PhosIDN: an integrated deep neural network for improving protein phosphorylation site prediction by combining sequence and protein–protein interaction information. Bioinformatics 2021;37:4668–76.
[27] Wang C, Xu H, Lin S, Deng W, Zhou J, Zhang Y, et al. GPS 5.0: an update on the prediction of kinase-specific phosphorylation sites in proteins. Genomics Proteomics Bioinformatics 2020;18:72–80.
[28] Fuhrmann J, Schmidt A, Spiess S, Lehner A, Turgay K, Mechtler K, et al. McsB is a protein arginine kinase that phosphorylates and inhibits the heat-shock regulator CtsR. Science 2009; 324:1323–7.
[29] Fuhs SR, Hunter T. pHisphorylation: the emergence of histidine phosphorylation as a reversible regulatory modification. Curr Opin Cell Biol 2017;45:8–16.
[30] Zhao J, Zhuang M, Liu J, Zhang M, Zeng C, Jiang B, et al. pHisPred: a tool for the identification of histidine phosphorylation sites by integrating amino acid patterns and properties. BMC Bioinformatics 2022;23:399.
[31] Chen Z, Zhao P, Li F, Leier A, Marquez-Lago TT, Webb GI, et al. PROSPECT: a web server for predicting protein histidine phosphorylation sites. J Bioinform Comput Biol 2020;18:2050018.
[32] Awais M, Hussain W, Khan YD, Rasool N, Khan SA, Chou KC. iPhosH-PseAAC: identify phosphohistidine sites in proteins by blending statistical moments and position relative features according to the Chou's 5-step rule and general pseudo amino acid composition. IEEE/ACM Trans Comput Biol Bioinform 2021;18:596–610.
[33] Huang KY, Lee TY, Kao HJ, Ma CT, Lee CC, Lin TH, et al. dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications. Nucleic Acids Res 2019; 47:D298–308.
[34] Amanchy R, Periaswamy B, Mathivanan S, Reddy R, Tattikota SG, Pandey A. A curated compendium of phosphorylation motifs. Nat Biotechnol 2007;25:285–6.
[35] Meng L, Chan WS, Huang L, Liu L, Chen X, Zhang W, et al. Mini-review: recent advances in post-translational modification site prediction based on deep learning. Comput Struct Biotechnol J 2022;20:3522–32.
[36] Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol 2014; 32:223–6.
[37] Mi H, Ebert D, Muruganujan A, Mills C, Albou LP, Mushayamaha T, et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res 2021;49:D394–403.
[38] Wagih O, Sugiyama N, Ishihama Y, Beltrao P. Uncovering phosphorylation-based specificities through functional interaction networks. Mol Cell Proteomics 2016;15:236–45.
[39] Tegge W, Frank R, Hofmann F, Dostmann WR. Determination of cyclic nucleotide-dependent protein kinase substrate specificity by the use of peptide libraries on cellulose paper. Biochemistry 1995;34:10569–77.
[40] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database — 2009 update. Nucleic Acids Res 2009;37:D767–72.
[41] Zhou S, Blechner S, Hoagland N, Hoekstra MF, Piwnica-Worms H, Cantley LC. Use of an oriented peptide library to determine the optimal substrates of protein kinases. Curr Biol 1994; 4:973–82.

[42] O'Shea JP, Chou MF, Quader SA, Ryan JK, Church GM, Schwartz D. pLogo: a probabilistic approach to visualizing sequence motifs. Nat Methods 2013;10:1211–2.
[43] Singh J, Litfin T, Paliwal K, Singh J, Hanumanthappa AK, Zhou Y. SPOT-1D-Single: improving the single-sequence-based prediction of protein secondary structure, backbone angles, solvent accessibility and half-sphere exposures using a large training set and ensembled deep learning. Bioinformatics 2021; 37:3464–72.
[44] Erdos G, Pajkos M, Dosztanyi Z. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. Nucleic Acids Res 2021;49:W297–303.
[45] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012; 28:3150–2.
[46] Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 2008;36:D202–5.
[47] Saha I, Maulik U, Bandyopadhyay S, Plewczynski D. Fuzzy clustering of physicochemical and biochemical properties of amino acids. Amino Acids 2012;43:583–94.
[48] Henikoff S, Henikoff J. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 1992;89:10915–9.
[49] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189–32.

Genomics, Proteomics & Bioinformatics, 2024, 22, 1–13
https://doi.org/10.1093/gpbjnl/qzae032
Database