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

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

[1] *Institute of Drug Discovery Technology, Ningbo University, Ningbo 315211, China*

[2] *State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Faculty of Medicine and Life Sciences, Xiamen University, Xiamen 361102, China*

[3] *Department of Chemical Biology, Key Laboratory for Chemical Biology of Fujian Province, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, China*

[4] *Zhejiang Key Laboratory of Pathophysiology, Department of Biochemistry and Molecular Biology, Health Science Center, Ningbo University, Ningbo 315211, China*

[5] *Key Laboratory of Bioorganic Phosphorus Chemistry & Chemical Biology, Department of Chemistry, Tsinghua University, Beijing 100084, China*

[6] *BGI Genomics, BGI-Shenzhen, Shenzhen 518083, China*

* Corresponding authors

E-mail: yfzhao@xmu.edu.cn (Zhao Y) and appo@xmu.edu.cn (Ji Z).

Running title

Zhao MX et al/Web Server of Nphos

Total word counts (from “Introduction” to “Materials and methods”): 4426

Total references: 49

Total figures: 5

Total tables: 4

Total supplementary figures: 3

Total supplementary tables: 5

The counts of letters in the article title and running title: 58 and 33

The count of keywords and words in Abstract: 5 and 182

Abstract

Protein *N*-phosphorylation widely present in nature and participates in various biological functions. 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 of hundreds of learning models, we chose and optimized gradient boosting decision tree (GBDT) models to predict three types of human *N*-phosphorylation, achieving mean areas under the receiver operating characteristic curve (AUC) 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 and http://ppodd.org.cn/Nphos/.

KEYWORDS: *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]. An association of protein *N*-phosphorylation with cancer progression [8-10] was recently found. However, compared to *O*-phosphorylation (pSer, pThr, and pTyr) 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. 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. 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 modification (PTM). 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 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, few *in silico* tools were specifically developed for the prediction of *N*-phosphosites. Most tools were designed for *O*-phosphorylation, which is mostly kinase-specific [20, 26, 27], making them inapplicable for

N-phosphosite prediction because only 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]) were trained for the large-scale prediction of pHis sites based on 487, 242, and 602 pHis sites. However, because eukaryotes and prokaryotes might not share common mechanisms of *N*-phosphorylation, implementation of these models for eukaryote 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*-phosphorylation sites 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 be the 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 most (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 1**A). By domain of life, 9101 *N*-phosphosites were

identified in 5893 eukaryotic proteins, and 2609 sites were identified in 1451 prokaryotic proteins. More statistics on *N*-phosphosites are 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 participate 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 proteomic scale. 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 33.31 fold higher than that of known *N*-phosphosites. The average modification abundance per protein for *O*-phosphorylation was 12.64, versus 1.55 for *N*-phosphorylation. 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.

Putting all these results 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 had no significant tendency regarding protein length, and phosphorylation was not correlated with amino acid usage (Figure S1 and Figure S2). Pattern analysis of 11,710 non-redundant *N*-phosphosite-centered 31 residue peptides helped characterized the conserved sequential motifs by which kinases

recognize the modification site (**Figure 2**A and Figure S3). In eukaryotes, lysine (K) and glutamic acid (E) were enriched near pHis; in particular, lysine 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, proline residues also frequently appear 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 abundant 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, and few occurred in the buried region (Figure 2B); however, the preference for the exposed region did not differ from that of non-phosphorylated His/Lys/Arg (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 as 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 as *O*-phosphorylation, had no strong preference for different protein segments, the *N*-terminal, the C-terminal, or the middle regions (Figure 2B).

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 + BLOSUM62 + HQI8, and pArg models preferred AAC + HQI8 (**Figure 3**A and Table S1). The exact reasons for the different preferences of feature vectors in the prediction of *N*-phosphorylation were 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 areas under the receiver operating characteristic (ROC) curve (AUCs) 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 test datasets, resulting in mean AUC 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 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 they were 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) (Figure 2A and Figure 3C). Therefore, we speculate that the positives were

under-represented during model learning. The inclusion of more types of positives in the model training would significantly improve the 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- 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 three-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 in 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.

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 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 4**A). 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 position 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 organized into three sections. The section of “Protein Information” section contains information about the protein, including the gene symbol, protein length, motifs, functions, 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*-phosphorylation sites" section lists the details of *N*-phosphosites along with the source or literature. The "Other PTMs 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 the 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, that satisfy the prediction probability of ≥ 0.5, by *N*-phosphorylation type in ascending order according to the residue position in the protein sequence. According to the prediction probability, the prediction results are roughly divided into three reliability grades: high (> 0.85), middle (0.70 – 0.85), and low (0.50 – 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 collected 11,710 experimentally verified *N*-phosphorylation sites in 7344 proteins from 39 species, covering pHis, pLys, and pArg sites. Using these data, 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 are remain limited even after thorough collection from the 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 label-free learning would significantly improve the accuracy of *N*-phosphorylation prediction. Moreover, the model performance is also affected by the length of the inputted peptide. 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 modified site kinases and phosphatases, 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 inspire us to take full advantage of cutting-edge artificial intelligent technologies in the accurate prediction of *N*-phosphorylation.

Conclusion

This study provided the most fruitful information on protein *N*-phosphorylation available to data. 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 (version 2.4, 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*, *Bacillus subtilis*, *Staphylococcus aureus*, and *Escherichia coli*. The parameters were set as follows: enzyme, trypsin; precursor mass tolerance, 20 ppm; fragment mass tolerance, 0.6 (IonTrap) or 0.02 Da (Orbitrap); spectrum matching, b, c, y, z ions (EthcD), b, y ions (CID or HCD), c, z ions (ETD); dynamic modification, Phospho / +79.966 Da (H, K, R, S, T, Y); and 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 mentioned (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 post-translational modifications

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 enrichment ratio (E-ratio) was calculated as follows:

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

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 assessed 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: 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, occurrences ≥ 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 region was 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 8678 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. (i) The non-phosphorylated His/Lys/Arg residues in the *N*-phosphorylated proteins were selected. (ii) 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. (iii) The negative peptides (the 31-residue sequence centered around non-phosphorylated His, Lys, or Arg) were generated. (iv) The wrongly or ambiguously assigned negatives were removed by conducting homologous analysis of the negative sequences against the positive dataset. The homologous 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 × 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:

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

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].

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

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

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

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 $\vec{s} = \overrightarrow{R_m}\ (if\ \ s = m,\ m \in A, C, \cdots, W, -)$.

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 via simple attachment the vectors 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:

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

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

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

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

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

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

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.

Competing interests

The authors have declared no competing interests.

Acknowledgments

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

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

References

1. Albert S, Helmut EM. Phosphoamino acid analysis. Proteomics. 2001;1:200–206.
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(2):537-550.
3. Falke J, Bass R, Butler S, Chervitz S, Danielson M. 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(3):217-234.

5. Perego M. Kinase-phosphatase competition regulates Bacillus subtilis development. Trends Microbiol. 1998;6(9):366-370.

6. Ward M, Zusman D. Regulation of directed motility in Myxococcus xanthus. Mol Microbiol. 1997;24(5):885-893.

7. Stock A, Robinson V, Goudreau P. 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(2):1108-1114.

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(7698):678-682.

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(2):161-173.

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(10):1285-1292.

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(7):416-421.

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(1):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(3):187-190.

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(6):708-712.

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(3):341-346.

18. Hu Y, Jiang B, Weng Y, Sui Z, Zhao B, Chen Y, et al. Bis(zinc(II)-dipicolylamine)-functionalized sub-2 mum core-shell microspheres for the analysis of N-phosphoproteome. Nat Commun. 2020;11(1):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(2):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(Database issue):D512-520.

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(D1):D471-D479.

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(D1):D542-D550.

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(1):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(24):4668-4676.

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(1):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(5932):1323-1327.

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(Suppl 3):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(4):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(2):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(D1):D298-D308.

34. Amanchy R, Periaswamy B, Mathivanan S, Reddy R, Tattikota S, Pandey A. A curated compendium of phosphorylation motifs. Nat Biotechnol. 2007;25(3):285-286.

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;20:3522-3532.

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(3):223-226.

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(D1):D394-D403.

38. Wagih O, Sugiyama N, Ishihama Y, Beltrao P. Uncovering Phosphorylation-Based Specificities through Functional Interaction Networks. Mol Cell Proteomics. 2016;15(1):236-245.

39. Tegge W, Frank R, Hofmann F, Dostmann W. Determination of cyclic nucleotide-dependent protein kinase substrate specificity by the use of peptide libraries on cellulose paper. Biochemistry. 1995;34(33):10569-10577.

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(Database issue):D767-772.

41. Songyang Z, Blechner S, Hoagland N, Hoekstra M, Piwnica-Worms H, Cantley L. Use of an oriented peptide library to determine the optimal substrates of protein kinases. Curr Biol. 1994;4(11):973-982.

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(12):1211-1212.

43. Singh J, Litfin T, Paliwal K, Singh J, Hanumanthappa A, 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(20):3464-3472.

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(W1):W297-W303.

45. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the

next-generation sequencing data. Bioinformatics. 2012;28(23):3150-3152.

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(Database issue):D202-205.

47. Saha I, Maulik U, Bandyopadhyay S, Plewczynski D. Fuzzy clustering of physicochemical and biochemical properties of amino acids. Amino Acids. 2012;43(2):583-594.

48. Henikoff S, Henikoff J. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992;89(22):10915-10919.

49. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001:1189-1232.

Figure legends

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*-phosphoproteins. Eu, Eukaryotes; Pro, Prokaryotes; *B. subtilis*, *Bacillus subtilis*; *S. aureus*, *Staphylococcus aureus*; *E. coli*, *Escherichia coli*; *M. smegmatis*, *Mycolicibacterium smegmatis*. **B.** The 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 protein in these species, within a specific classification. **C.** Statistics of *O*-phosphorylation and *N*-phosphorylation by proteins or sites.

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

A. Sequential features of pHis/pLys/pArg modifications. The conserved motifs were detected with motif-x (score = $\sum$-log(P), score $\geq$ 5, occurrences $\geq$ 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 gradient boosting decision tree (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.

Figure 4 The web services of Nphos

A. Data retrieval of Nphos. NOD, number of detections; HTT, high throughout technology; LTT, low throughout technology. **B.** Interactive *N*-phosphosite prediction.

Figure 5 Scheme of the protein *N*-phosphosites predictor

AA, amino acid; AAC, amino acid composition; AAindex-HQI8, amino acid index-high quality indices; TP, true positive; FP, false positive; FN, false negative; TN, true negative; Windows, the length of the sequence window for extracting features.

Table 1 The performance of GBDT models in the testing datasets

Table 2 Comparison of models in predicting pHis sites

Table 3 Comparison of phosphorylation ratios in the human proteome

Table 4 Dataset preparation for machine learning

Supplementary material

Table S1 Classifiers optimal results

Table S2 The fine-tuning of hyperparameters

Table S3 The performance of GBDT models before and after hyperparameter fine-tuning

Table S4 The MS raw data of protein *N*-phosphorylation

Table S5 Optimal parameters of the classifiers

Figure S1 Protein length distribution of *N*-/*O*-phosphorylation in the proteomes of eukaryotes and prokaryotes

The dotted line represents the mean of the overall distribution of protein length.

Figure S2 Correlation analyses between amino acid usage and the occurrence rate of phosphorylation

A. Results for eukaryotes. **B.** Results for prokaryotes.

Figure S3 Motif analyses around the pHis/pLys/pArg sites

Logs are calculated in base 10. The conserved motifs were detected using motif-x (score = $\sum$-log(P), score $\geq$ 5, occurrences $\geq$ 10)

Figure 1

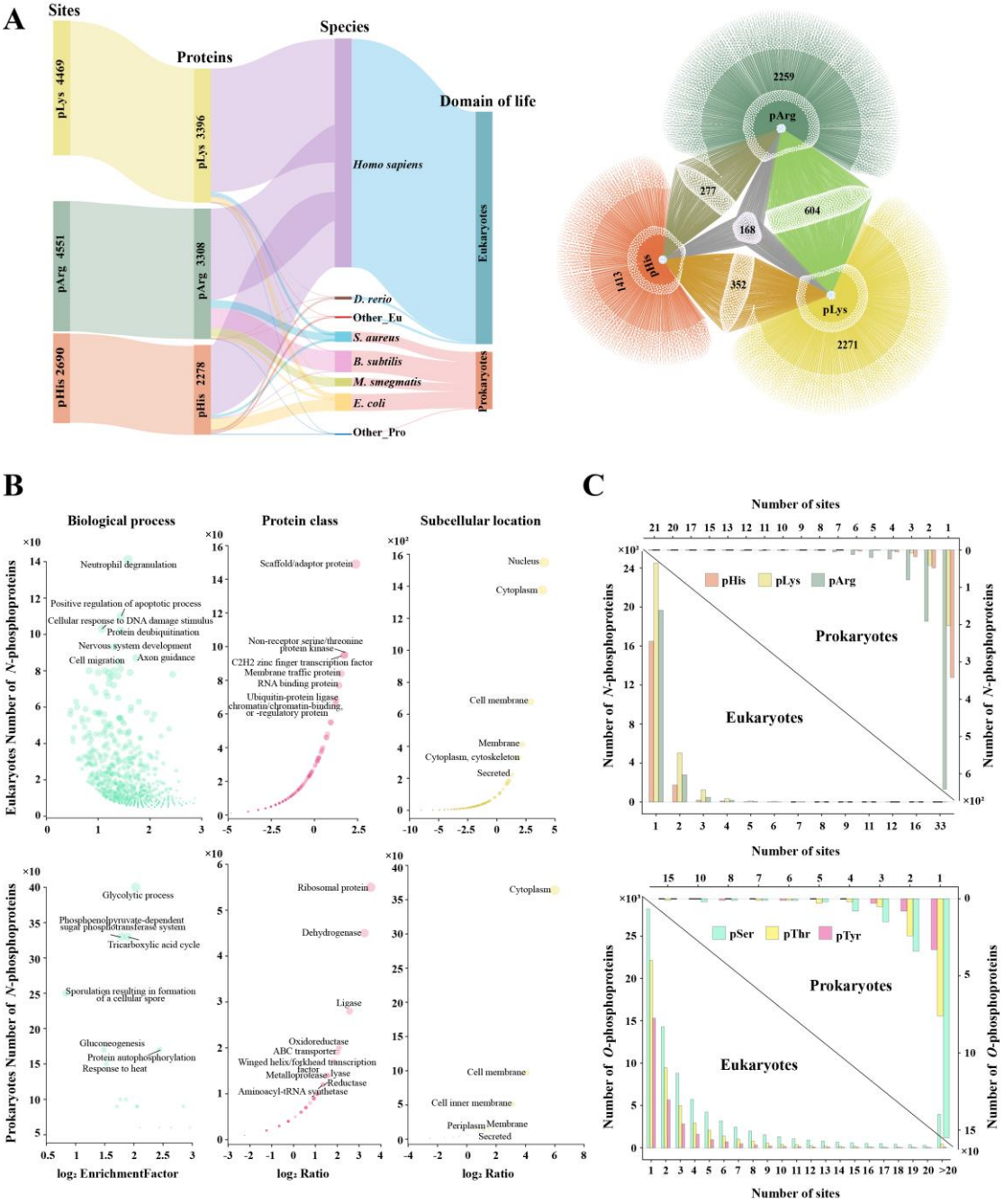
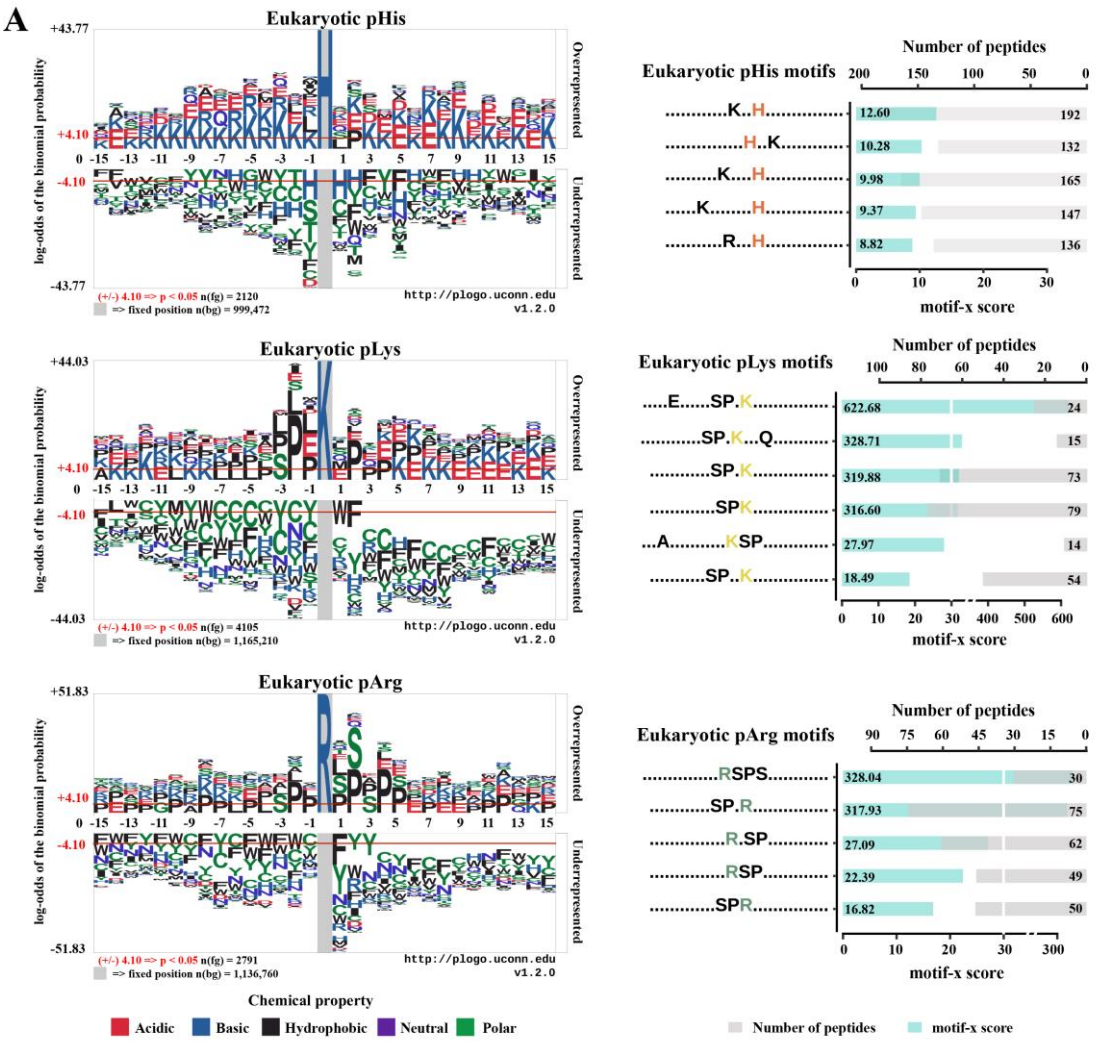

A
Sites
Proteins
Species
Domain of life
pLys 4469
pArg 4551
pHis 2690
pLys 3396
pArg 3308
pHis 2278
Homo sapiens
D. rerio
Other_Eu
S. aureus
B. subtilis
M. smegmatis
E. coli
Other_Pro
Eukaryotes
Prokaryotes
2259
pArg
277
604
168
1413
pHis
352
pLys
2271
B
Biological process
Protein class
Subcellular location
Eukaryotes Number of N-phosphoproteins
Prokaryotes Number of N-phosphoproteins
Neutrophil degranulation
Positive regulation of apoptotic process
Cellular response to DNA damage stimulus
Protein deubiquitination
Nervous system development
Cell migration
Axon guidance
Scaffold/adaptor protein
Non-receptor serine/threonine protein kinase
C2H2 zinc finger transcription factor
Membrane traffic protein
RNA binding protein
Ubiquitin-protein ligase
chromatin/chromatin-binding, or -regulatory protein
Nucleus
Cytoplasm
Cell membrane
Membrane
Cytoplasm, cytoskeleton
Secreted
Glycolytic process
Phosphoenolpyruvate-dependent sugar phosphotransferase system
Tricarboxylic acid cycle
Sporulation resulting in formation of a cellular spore
Gluconeogenesis
Protein autophosphorylation
Response to heat
Ribosomal protein
Dehydrogenase
Ligase
Oxidoreductase
ABC transporter
Winged helix/forkhead transcription factor
Metalloprotease
Lyase
Reductase
Aminoacyl-tRNA synthetase
Cytoplasm
Cell membrane
Cell inner membrane
Periplasm
Membrane
Secreted
log₂ EnrichmentFactor
log₂ Ratio
log₂ Ratio
C
Number of sites
pHis
pLys
pArg
Prokaryotes
Eukaryotes
Number of N-phosphoproteins
pSer
pThr
pTyr
Number of O-phosphoproteins

Figure 2


A
Eukaryotic pHis
Eukaryotic pLys
Eukaryotic pArg
log-odds of the binomial probability
Overrepresented
Underrepresented
(+/-) 4.10 => p < 0.05 n(fg) = 2120
=> fixed position n(bg) = 999,472
(+/-) 4.10 => p < 0.05 n(fg) = 4105
=> fixed position n(bg) = 1,165,210
(+/-) 4.10 => p < 0.05 n(fg) = 2791
=> fixed position n(bg) = 1,136,760
http://plogo.uconn.edu
v1.2.0
Chemical property
Acidic
Basic
Hydrophobic
Neutral
Polar
Eukaryotic pHis motifs
Eukaryotic pLys motifs
Eukaryotic pArg motifs
Number of peptides
motif-x score
.....K..H.....
.....H..K.....
.....K....H.....
.....K.....H.....
.....R...H.....
.....E.....SP.K.....
.....SP.K...Q.....
.....SP.K.....
.....SPK.....
...A.....KSP.....
.....SP..K.....
.....RSPS.....
.....SP.R.....
.....R.SP.....
.....RSP.....
.....SPR.....
Number of peptides
motif-x score

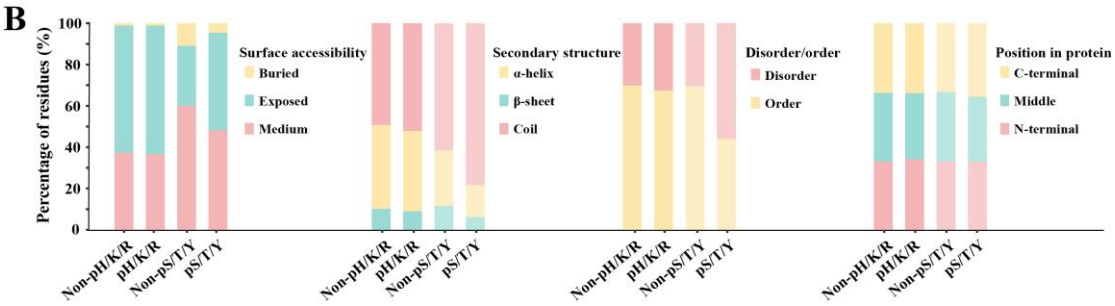

B
Percentage of residues (%)
Surface accessibility
Buried
Exposed
Medium
Secondary structure
α-helix
β-sheet
Coil
Disorder/order
Disorder
Order
Position in protein
C-terminal
Middle
N-terminal
Non-pH/K/R
pH/K/R
Non-pS/T/Y
pS/T/Y

Figure 3

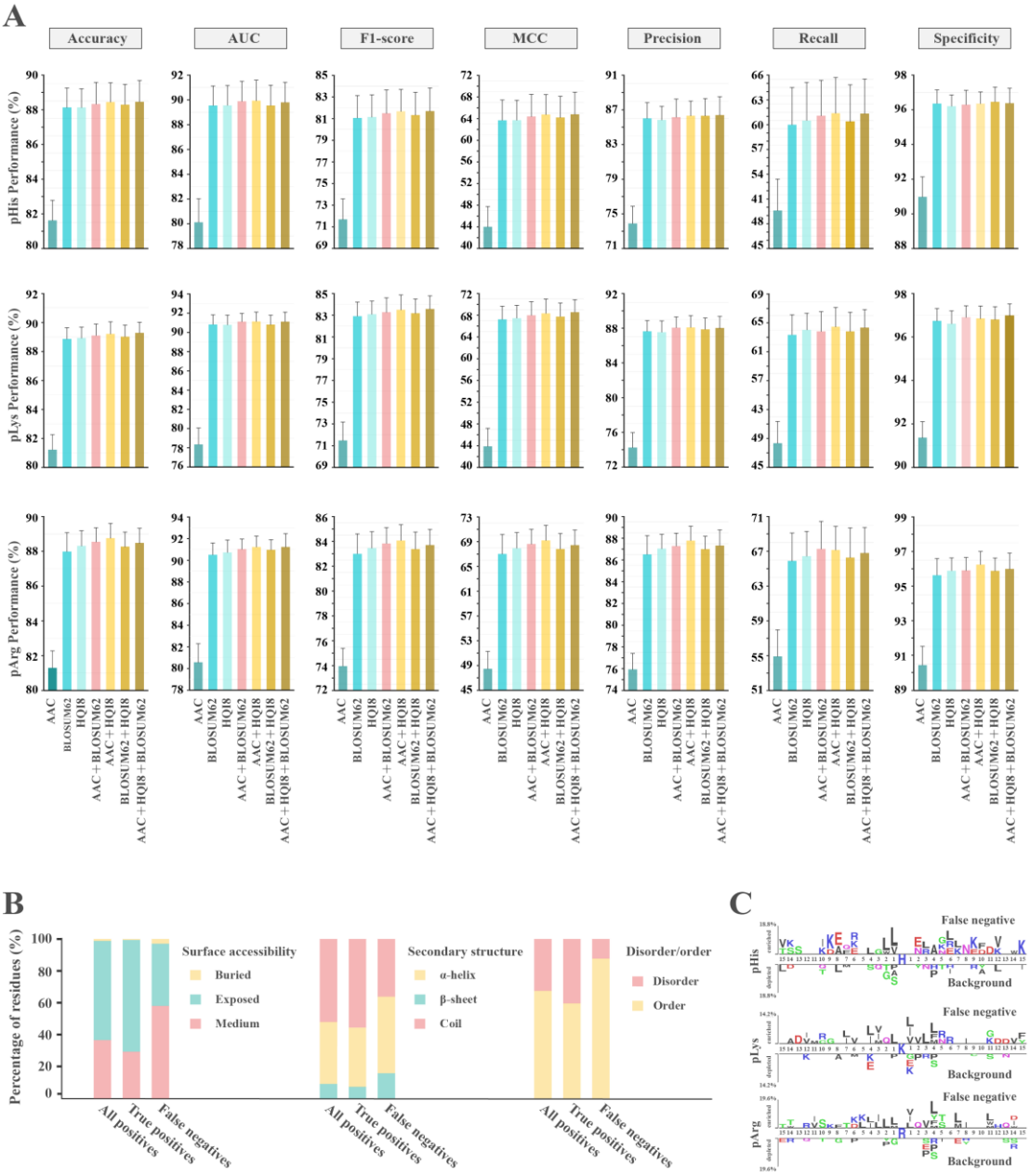

A
Accuracy
AUC
F1-score
MCC
Precision
Recall
Specificity
pHis Performance (%)
pLys Performance (%)
pArg Performance (%)
AAC
BLOSUM62
HQI8
AAC + BLOSUM62
AAC + HQI8
BLOSUM62 + HQI8
AAC + HQI8 + BLOSUM62
B
Percentage of residues (%)
Surface accessibility
Buried
Exposed
Medium
Secondary structure
α-helix
β-sheet
Coil
Disorder/order
Disorder
Order
All positives
True positives
False negatives
C
pHis
pLys
pArg
False negative
Background

Figure 4

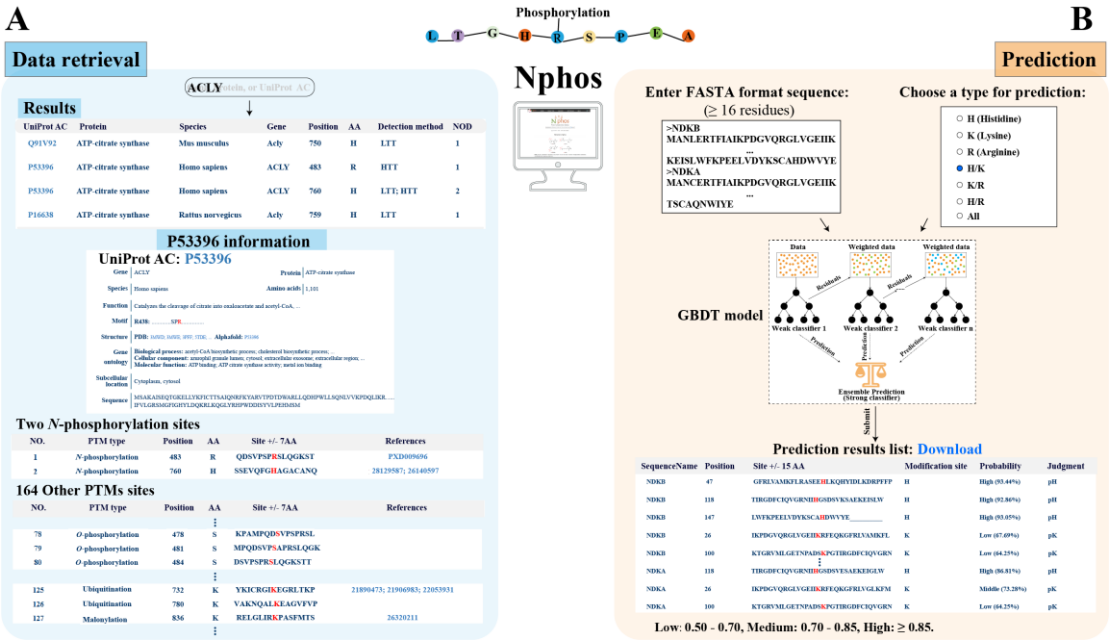


Figure 5

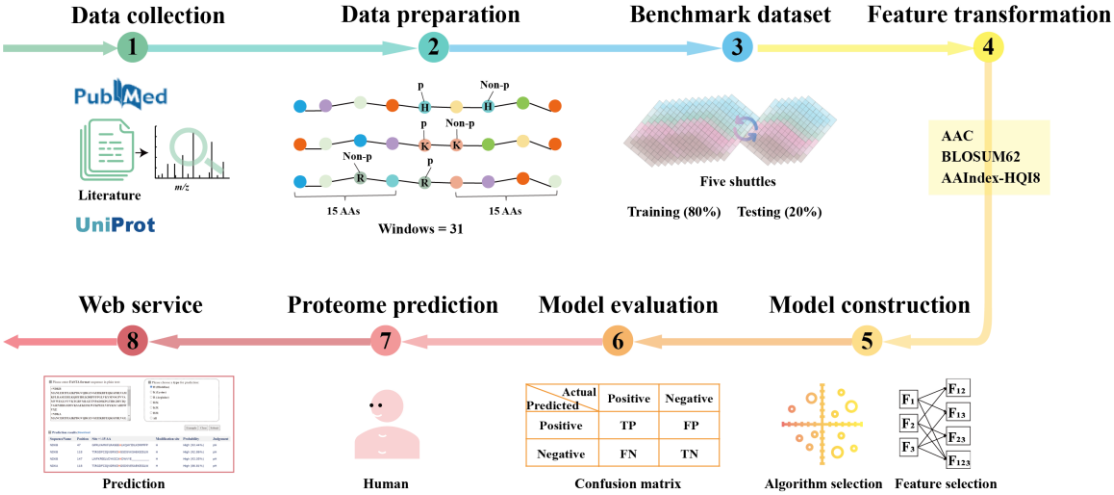


Table 1 The performance of GBDT models in the testing datasets

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 or 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: AUC, Area under curve; MCC, Matthews correlation coefficient

Table 2 Comparison of models in predicting pHis sites

Models	Species type	Dataset			Windows	Algorithm	Performance			
		Positive	Negative	Test [a]			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) [b]	**Human**	**1945**	**6669** (non-pHis with RSA constraint)	**20%**	**31**	**GBDT**	**0.9056**	**0.6151**	**0.6678**	**0.8255**

Note: AUC, Area under curve; CNN, Convolutional neural network; Eu, Eukaryotes; *E. coil*, *Escherichia coli*; GBDT, Gradient boosting decision tree; MBPNN, Multi-layered back propagation neural; MCC, Matthews correlation coefficient; Negative, the number of negative dataset; Positive, the number of positive dataset; Pro, Prokaryotes; RSA, Relative solvent accessibility; SVM, Support vector machines; Windows, the length of the sequence window for extracting features.

a: “Test” refers to the percentage of the test dataset in the total dataset.

b: The results of this study are shown in bold.

Table 3 Comparison of phosphorylation ratios in the human proteome

Types	Amino acid	Phosphosites	Total residues	Ratio (%)
	His	64,409	298,024	21.61
***N*-phosphorylation**	**Lys**	214,679	651,188	32.97
	Arg	209,737	640,398	32.75
	Ser	179,612	946,894	18.97
***O*-phosphorylation**	**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 the multiple sources described in the main text.

Table 4 Dataset preparation for machine learning

***N*-Phos Type**	**Positive**	**Negative**	**Ratio**
pHis	1945	6669	1 : 3.43
pLys	4008	12,970	1 : 3.24
pArg	2723	7893	1 : 2.90

Note: N-Phos, *N*-phosphorylation; Negative, the number of negative dataset; Positive, the number of positive dataset; Ratio, the ratio of positive and negative dataset.