Briefings in Bioinformatics, 2022, **23(2)**, 1–15

https://doi.org/10.1093/bib/bbab574

Review

GPS-Uber: a hybrid-learning framework for prediction of general and E3-specific lysine ubiquitination sites

Chenwei Wang†, Xiaodan Tan†, Dachao Tang, Yujie Gou, Cheng Han, Wanshan Ning, Shaofeng Lin, Weizhi Zhang, Miaomiao Chen, Di Peng and Yu Xue

Corresponding author: Yu Xue, Department of Bioinformatics and Systems Biology, MOE Key Laboratory of Molecular Biophysics, Hubei Bioinformatics and Molecular Imaging Key Laboratory, Center for Artificial Intelligence Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China. Tel.: +86-27-87793903; Fax: +86-27-87793172; E-mail: xueyu@hust.edu.cn.

†These authors contributed equally to this work.

Abstract

As an important post-translational modification, lysine ubiquitination participates in numerous biological processes and is involved in human diseases, whereas the site specificity of ubiquitination is mainly decided by ubiquitin-protein ligases (E3s). Although numerous ubiquitination predictors have been developed, computational prediction of E3-specific ubiquitination sites is still a great challenge. Here, we carefully reviewed the existing tools for the prediction of general ubiquitination sites. Also, we developed a tool named GPS-Uber for the prediction of general and E3-specific ubiquitination sites. From the literature, we manually collected 1311 experimentally identified site-specific E3-substrate relations, which were classified into different clusters based on corresponding E3s at different levels. To predict general ubiquitination sites, we integrated 10 types of sequence and structure features, as well as three types of algorithms including penalized logistic regression, deep neural network and convolutional neural network. Compared with other existing tools, the general model in GPS-Uber exhibited a highly competitive accuracy, with an area under curve values of 0.7649. Then, transfer learning was adopted for each E3 cluster to construct E3-specific models, and in total 112 individual E3-specific predictors were implemented. Using GPS-Uber, we conducted a systematic prediction of human cancer-associated ubiquitination events, which could be helpful for further experimental consideration. GPS-Uber will be regularly updated, and its online service is free for academic research at http://gpsuber.biocuckoo.cn/.

Keywords: Post-translational modification, lysine ubiquitination, ubiquitin-protein ligase, site-specific E3-substrate relation, deep learning

Introduction

As one of the most indispensable post-translational modifications (PTMs), lysine ubiquitination regulates a wide spectrum of biological processes including protein degradation and turnover, membrane trafficking, cell cycle and deoxyribonucleic acid (DNA) damage repair [1–3]. In 1978, Ciehanover *et al.* discovered a 76-amino acid protein, ubiquitin, which can be covalently attached

Chenwei Wang is a postdoc scientist at Huazhong University of Science and Technology. He mainly focuses on the development of new algorithms to predict functional PTM events from multi-omic data. He developed a number of computational methods including iCMod, cMAK, iFPS and iFIP to predict functional kinases, substrates and interacting partners involved in regulating circadian, autophagy and ageing. He was also a major developer of EPSD for collecting known protein phosphorylation sites in eukaryotes.

Xiaodan Tan is a master student at Huazhong University of Science and Technology. She mainly focuses on the collection and annotation of protein lysine modifications.

Dachao Tang is a PhD student at Huazhong University of Science and Technology. His major research is identification of potentially important proteins involved in regulating autophagy.

Yujie Gou is a PhD student at Huazhong University of Science and Technology. She focuses on the development of new deep learning frameworks to process biomedical imaging data.

Cheng Han is a PhD student at Huazhong University of Science and Technology. He is working on the computational prediction of PTM sites.

Wanshan Ning is a postdoc scientist at Huazhong University of Science and Technology. His major research interest is focused on using artificial intelligence methods to analyze sequence, multi-omics and imaging data. He built a new hybrid-learning architecture named HybridSucc for predicting general and species-specific succinylation sites. He also developed GPS-Palm to predict palmitoylation sites, HUST-19 to predict COVID-19 clinical outcomes and POC-19 to prioritize protein biomarkers of COVID-19.

Shaofeng Lin is a PhD student at Huazhong University of Science and Technology. His major research interest is focused on the integration of PTM data. He was a major developer of EPSD. He was also a major developer of iUUCD 2.0, a database of ubiquitin and ubiquitin-like conjugations.

Weizhi Zhang is a PhD student at Huazhong University of Science and Technology. He mainly focuses on developing machine learning algorithms based on multi-omic data. He developed a new method named iCAL to predict cancer mutations that change autophagy selectivity.

Miaomiao Chen is a PhD student at Huazhong University of Science and Technology. She is working on the prediction of phosphorylation sites using deep learning algorithms.

Di Peng is a postdoc scientist at Huazhong University of Science and Technology. His major research interests are focused on experimentally discovering new PTM regulators, substrates and sites in the regulation of diverse biological processes, with the combination of computational predictions.

Yu Xue is a professor at Huazhong University of Science and Technology. He has started to work in the field of PTM Bioinformatics since 2004. He is interested in using both computational and experimental approaches to exploit how functional PTM events can be precisely orchestrated to regulate various biological processes, such as autophagy, circadian and cell fate determination. He is also involved in the establishment of a new interdisciplinary field, artificial intelligence biology (AIBIO) in China.

Received: September 3, 2021. **Revised:** December 11, 2021. **Accepted:** December 14, 2021

to lysine residues in protein substrates through a cascade of biochemical reactions catalyzed by ubiquitin-activating enzymes (E1s), ubiquitin-conjugating enzymes (E2s) and ubiquitin-protein ligases (E3s) [4, 5]. E3s are structurally diverse enzymes and play a critical role in determining the substrate specificity and efficiency of ubiquitination reactions [6, 7]. The aberrances in E3s and ubiquitinated targets have been associated with numerous human diseases, such as cancer, autoimmune diseases, metabolic syndromes and neurodegenerative diseases [7–9]. Thus, identification of E3-specific targets and site-specific E3-substrate relations (ssESRs) is fundamental for understanding the molecular mechanisms and regulatory roles of lysine ubiquitination.

Conventionally, biochemical identification of E3-specific targets and ubiquitination sites is low-throughput (LTP), labor-intensive and time-consuming. During the past years, a number of high-throughput (HTP) experimental methods have been developed, such as yeast two-hybrid screening, phage display, global protein stability profiling, affinity purification-tandem mass spectrometry (AP-MS/MS) and Gly–Gly (diGly) remnant affinity purification [10–12]. For example, in 2008, Yen *et al*. developed a fluorescence-based system called global protein stability profiling, which could monitor the protein turnover under different physiological and disease conditions [13]. Using this method, Yen *et al*. systematically identified 359 highly potential substrates of the Skp1-cullin-F-box (SCF) ubiquitin ligase, and most of the known SCF targets were covered [14]. With the help of AP-MS/MS, Low *et al*. identified 221 potential $\mathrm{SCF}^{\beta\mathrm{TrCP}}$ substrates that contained the DpSGXX(X)pS motif, a primary degron to be specifically recognized by $\mathrm{SCF}^{\beta\mathrm{TrCP}}$ [15]. In addition, Elia *et al*. identified 33 503 ubiquitination sites using the diGly remnant affinity purification strategy and discovered EXO1 as a new SCF target in response to DNA damage [16].

Besides the LTP and HTP experimental assays, computational prediction of E3-substrate interactions (ESIs) or ubiquitination sites has also emerged to be a highly useful approach. For the prediction of ESIs, in 2017, Li *et al*. incorporated multiple types of informative features including orthologous ESI, enriched domain pair, enriched Gene Ontology (GO) term pair, network topology and E3 recognition motif (aka ‘primary degron’) and developed a naïve Bayesian-based method named UbiBrowser [17]. Recently, UbiBrowser 2.0 was released to cover more species, and prediction of deubiquitinase-substrate interactions was also implemented [18]. In parallel, Chen *et al*. integrated transcriptomics-, proteomics-, network- and pathway-based associations and used recursive feature elimination and random forest (RF) algorithms to develop a new method for predicting ESIs [19]. Through further experiments, they validated 3 and 5 potentially new targets of $\mathrm{SCF}^{\mathrm{SKP2}}$ and $\mathrm{SCF}^{\mathrm{FBXL6}}$, respectively [19]. For the prediction of general or species-specific ubiquitination sites, various tools have also been developed, including UbiPred [20], UbPred [21], UbSite [22], CKSAAP_UbSite [23], WPNNA [24], UbiProber [25], hCKSAAP_UbSite [26], RUBI [27], iUbiq-Lys [28], UbiSite [29], ESA-UbiSite [30], PTM-ssMP [31], PTMscape [32], ModPred [33], deepUbiquitylation [34], DeepUbi [35], MUscADEL [36], DL-plant-ubsites-prediction [37], MusiteDeep [38], UbiSite-XGBoost [39], UbiComb [40], CNNAthUbi [41], DeepTL-Ubi [42] and MultiLyGAN [43]. Although numerous efforts have been taken in computational analysis of ubiquitination, prediction of exact ssESRs remains to be a great challenge.

Here, we first provided a brief review of currently available tools for predicting general and species-specific ubiquitination sites. Then, we developed an online service named group-based prediction system for ubiquitin E3 ligase-substrate relations (GPS-Uber), which could predict general and E3-specific lysine ubiquitination sites from protein sequences. For training models in GPS-Uber, seven sequence- and three structure-based features were considered, and three machine learning algorithms including two-dimensional (2D) convolutional neural network (CNN), deep neural network (DNN) and penalized logistic regression (PLR) were integrated into a hybrid-learning architecture. Compared with other existing tools, GPS-Uber showed a highly competitive accuracy, with an area under the curve (AUC) value of 0.7649 for the prediction of general ubiquitination sites. With the help of transfer learning, 111 individual E3-specific predictors were also constructed (Figure 1). To investigate the potential relationships between ubiquitination and cancer, the ubiquitination sites of known cancer proteins were predicted by GPS-Uber at the E3 group level and could serve as a useful resource for further experimental consideration. Taken together, we anticipate that GPS-Uber could be helpful to facilitate the research on E3-mediated ubiquitination.

Methods

Data collection and preparation

First, the combinations of keywords including ‘ubiquitination’, ‘ubiquitinated’ and ‘ubiquitylation’ were added with suffixes such as ‘lysine’, ‘residue’, ‘site’ and ‘proteomic’ to search experimentally identified ssESRs from PubMed. Only known ssESRs in *Homo sapiens* were collected, because much fewer ssESRs were identified in other species. Through the literature biocuration, we obtained 1311 known ssESRs between 1117 human ubiquitination sites of 391 proteins and 177 E3s (Supplementary Table S1). More details on collection of known ssESRs were shown in the Supplementary Methods.

In 2017, we developed the protein lysine modification database (PLMD), which contained 121 742 experimentally identified lysine ubiquitination sites in 25 103 proteins [44]. For the prediction of general ubiquitination sites, these sites were taken as the benchmark data set. A widely used clustering program, CD-HIT [45], was adopted to classify this data set into different clusters

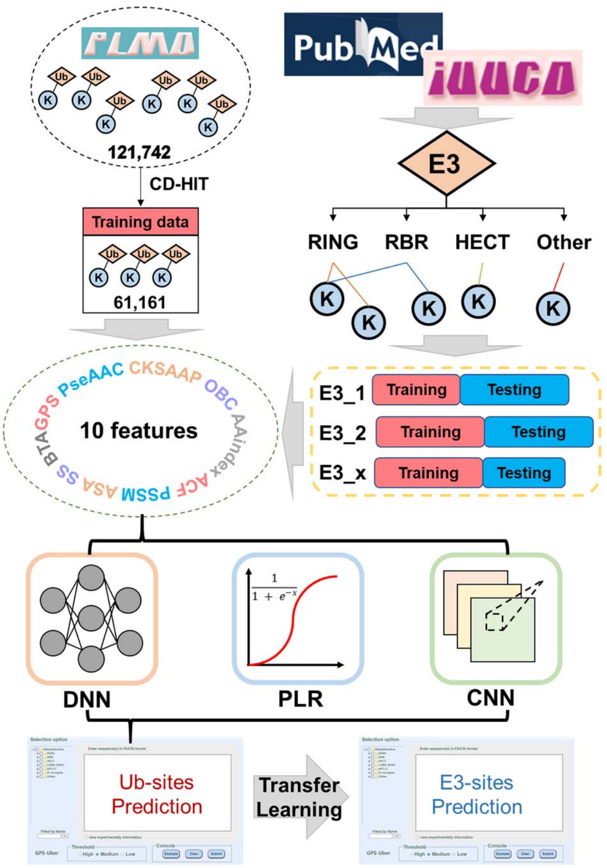


Figure 1. The experimental procedure of this study. First, experimentally identified ubiquitination sites were taken from PLMD [44], and homologous sites were eliminated through the CD-HIT clustering [45] to generate the initial training data set which contained 61 161 ubiquitination sites. Then, 10 types of sequence- and structure-based features, including GPS, PseAAC, CKSAAP, OBC, AAindex, ACF, PSSM, ASA, SS and BTA, were encoded for model training with three algorithms including DNN, PLR and CNN. Meanwhile, known E3-specific ubiquitination sites were manually collected from PubMed and classified to various E3 clusters based on the information of iUUCD [1]. Transfer learning was performed for different E3 clusters to construct E3-specific models based on the general model. Finally, a user-friendly online service was developed for researchers in this field.

with a threshold of 40% sequence similarity. To avoid the homologous redundancy, only one representative sequence in each cluster was extracted into training data. Then, we defined a ubiquitination site peptide USP(m, n) as a lysine residue flanked by upstream m residues and downstream n residues, and USP(10,10) was chosen in this study for rapid training. As previously described [46], the USP(10,10) items around known ubiquitination sites were regarded as positive data, whereas USP(10,10) items from other non-ubiquitinated lysine residues were taken as negative data. For lysine residues located near to N- or C-terminus of the protein sequences, one or multiple characters '*' were added to complement the USP(10,10) items. Prior to model training, the redundant USP(10,10) items were removed.

Before the E3-specific training, the hierarchical classifications of human E3s at different levels, including class, group, subgroup, family and single E3, were downloaded from integrated annotations for Ubiquitin and Ubiquitin-like Conjugation Database (iUUCD) (http://iuucd.biocuckoo.org/) [1], and the 1311 known ssESRs were classified into different E3 clusters at group, subgroup, family and single E3 levels. Only E3 clusters with ≥3 ubiquitination sites were kept for further training. For each E3 cluster, positive and negative data were generated the same as that in general training. Finally, we got 111 E3 clusters with ≥3 known sites.

Performance evaluation measurements

For the evaluation of our methods, four widely used measurements, including sensitivity (Sn), specificity (Sp), accuracy (Ac) and Matthew correlation coefficient (MCC), were calculated as below:

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

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

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

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

For the prediction of general ubiquitination sites, 4-, 6-, 8- and 10-fold cross-validations were performed to evaluate the accuracy and robustness of finally determined models, using the training data set that contained 61 161 known ubiquitination sites. For the comparison of GPS-Uber with other existing tools, a timestamp-based strategy [38] was adopted to split the initial benchmark data set into a secondary training data set containing 55 426 sites reported before 2016, and an independent testing data set with 5735 sites released after 2016. Additional models were generated with this training data set using the algorithm of GPS-Uber, and the testing data set was then used to evaluate the performance of GPS-Uber and other tools. The initial benchmark data set, the secondary training data set and the independent testing data set were freely downloadable at: http://gpsuber.biocuckoo.cn/userguide.php. For predicting E3-specific sites, the robustness of models with ≥30 sites was tested with 10-fold cross-validations for 20 times, and leave-one-out (LOO) validations were performed for other models with <30 sites. For each model, the receiver operating characteristic (ROC) curve was illustrated based on Sn and 1-Sp scores, from which the AUC value was further calculated.

The algorithm of GPS-Uber

In 2020, we developed a hybrid-learning architecture called HybridSucc which combined a PLR algorithm with a DNN algorithm for the prediction of succinylation sites [47]. Through the integration of conventional machine learning and deep learning algorithms, the performance of the predictor was significantly improved. Later, a parallel CNN framework was constructed, and a new tool, GPS-Palm, was released for S-palmitoylation site prediction with a promising accuracy with graphical features [48].

In this study, a novel hybrid-learning framework was designed to incorporate PLR, DNN and CNN algorithms. First, seven types of sequence features including the peptide similarity encoded by the GPS method [46], pseudo amino acid composition (PseAAC), composition of *k*-spaced amino acid pairs (CKSAAPs), orthogonal binary coding (OBC), physicochemical properties in the Amino Acid index (AAindex) database , autocorrelation functions (ACFs) and position-specific scoring matrix (PSSM), and three structural features including accessible surface area (ASA), secondary structure (SS) and backbone torsion angles (BTAs) were encoded to one-dimensional (1D) vectors for PLR and DNN and 2D matrices for CNN (Figure 1) [20–30, 34, 35, 37, 40–43, 47, 49–52]. More details on feature encoding were shown in Supplementary Methods. For each feature, a PLR model with the ridge (L2) penalty was constructed by scikit-learn v0.23.2, and the 'lbfgs' solver was adopted for parameter optimization. A four-layer DNN framework was implemented in Keras 2.4.3 (http://github.com/fchollet/keras) based on tf-nightly-gpu 2.5.0 dev20201028 for the same encoded vectors. Similarly, an 11-layer 2D CNN framework containing four convolutional and four pooling layers was realized for the encoded matrices. The rectified linear unit (*ReLU*) was adopted as the activate function, which was defined as:

$$ReLU(x) = \begin{cases} x, x \geq 0 \\ 0, x < 0 \end{cases}$$

In the output layer, one neuron with the *sigmoid* function was taken to calculate the final score for a given USP(10,10):

$$sigmoid(y) = \frac{1}{1 + e^{-y}}$$

To rapidly determine the optimal parameters of deep learning networks, we randomly extracted 1/10 ubiquitination sites from the training data set for general prediction, and different combinations of parameters were tested with this data set to determine the combination with the highest AUC value. The optimized parameters including number of neurons, dropout ratio, learning rate and pool size were provided in Supplementary Table S2. For each USP(10,10), 10 features ($f_1, f_2, f_3, \ldots, f_{10}$) were separately scored by DNN (D_1, D_2, D_3, …, D_{10}), PLR (P_1, P_2, P_3, …, P_{10}) and CNN (C_1, C_2, C_3, …, C_{10}). Then a 30-dimensional vector containing 30 scores was generated as follows:

$$V = (D_1, D_2, D_3, \ldots, D_{10}, P_1, P_2, P_3, \ldots, P_{10}, C_1, C_2, C_3, \ldots, C_{10})$$

To integrate the information from various features and algorithms, the vector *V* was then used as the secondary feature and a new four-layer DNN model was constructed to get a final score.

For E3-specific prediction, the ssESRs in each cluster were used to fine-tune the general models through the transfer learning strategy, and the optimized models were assigned to corresponding E3 predictors. For each predictor, three thresholds including high, medium and low were determined based on *Sp* values of 95%, 90% and 85%, respectively. In the online service of GPS-Uber, the medium threshold was chosen as the default.

A computer with the NVIDIA GeForce RTX 3090 GPU, a Genuine Intel(R) CPU @ 2.30GHz CPU and 128 GB RAM were used for the training of computational models.

The hypergeometric test

For the enrichment analysis of E3-specific substrates, GO annotation files (released on 1 May 2021) [53] were downloaded from the Gene Ontology Resource (http://geneontology.org/), containing 19 762 human proteins with at least one GO term. For each GO term *t* with E3 group *e*, we defined the following:

N = number of genes annotated by at least one GO term.
n = number of genes annotated by GO term *t*.

M = number of *e*'s substrates annotated by at least one GO term.
m = number of *e*'s substrates annotated by GO term *t*.

The enrichment ratio (E-ratio) of *t* was then computed, and the *P*-value was calculated with the hypergeometric distribution as follows:

$$\mathrm{E-ratio} = \frac{m}{M} / \frac{n}{N}$$

$$p - value = \sum_{m'=m}^{n} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} \ (\text{E-ratio} \geq 1), \text{ or.}$$

$$p - value = \sum_{m'=0}^{m} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} \ (\text{E-ratio} < 1).$$

The hypergeometric test was also adopted for the GO-based enrichment analyses of cancer proteins predicted to be ubiquitinated by E3 groups. A total of 707 cancer proteins were downloaded from the Cancer Gene Census in Catalogue of Somatic Mutations in Cancer (COSMIC) (https://cancer.sanger.ac.uk/census, v94) [54].

The analysis of primary degrons

Previously, it was reported that E3 recognition motifs act as primary degrons to determine the ubiquitination specificity at the substrate level [55]. The Eukaryotic Linear Motif Database (ELM, http://elm.eu.org) provides a comprehensive dataset of experimentally characterized short linear motifs, including known E3-specific primary degrons [56]. Here, we downloaded the file 'elm_classes.tsv' that contained 317 motif classes and associated regular expressions from ELM. The information from the columns 'ELMIdentifier', 'FunctionalSiteName' and 'Description' was extracted, and 27 known degrons were reserved for 11 E3 clusters in GPS-Uber if available (Supplementary Table S3). Using the 're' module of Python, the sequence profile of each E3-specific degron was used to search the protein sequences of the corresponding E3-specific substrates, using the 391 proteins containing 1311 known ssESRs. More details on collection of known ssESRs were shown in Supplementary Methods. For each identified degron motif, the distance to its proximal ubiquitination site modified by the same E3 was counted.

The gene expression and proteomic data

Files (*.mRNAseq_Pre-process.Level_3.*) containing the mRNA expression levels of 37 cancer types of The Cancer Genome Atlas (TCGA) program were downloaded from BROAD Institute (http://gdac.broadinstitute.org/runs/stddata__latest/) [57]. We mapped this data set to 177 E3s and 391 substrates of the E3-specific data set to get their mRNA expression profiles. Time-course proteomic data generated from a previously study [58] were also used, containing 6205 proteins mutually quantified from normal rat kidney cells treated with 16-nm silica nanoparticles at 60 μg/ml for 0, 8, 16, 20 and 24 h. For rat proteins, their human orthologs of E3s and substrates were computationally identified by reciprocal best hits [59].

The data visualization

For each E3 group, the USP(10,10) items in positive data were directly uploaded to the web service of pLogo (https://plogo.uconn.edu/), and corresponding negative data were selected as background. Then the sequence logo was generated automatically. The heat map was diagrammed by a previously developed tool HemI [60], and Cytoscape [61] was used to visualize networks. In addition, the functional domain and predicted ubiquitinated sites of RAC1 were illustrated using DOG 2.0 [62].

Results

A summary of available methods for the prediction of ubiquitination sites

Besides the large-scale identification of ESIs and ubiquitination sites with HTP experimental methods [10–14, 16], computational predictions also provided an alternative approach to facilitate the research of ubiquitination. Because fewer studies have been conducted on the prediction of ESIs [17–19], here we mainly focused on review of the 28 available methods for predicting general or species-specific ubiquitination sites (Supplementary Table S4).

In 2008, Tung *et al*. developed the first ubiquitination site predictor named UbiPred [20]. After the evaluation of different features and classifiers, the combination of 31 informative physicochemical properties from AAindex and support vector machine (SVM) algorithm was adopted for training the final model [20]. In the next 10 years, SVM has been widely used for predicting ubiquitination sites. For example, Chen *et al*. designed CKSAAP_UbSite based on the CKSAAP feature, and SVM was used to predict yeast ubiquitination sites [23]. For the prediction of human ubiquitination sites, the authors released hCKSAAP_UbSite, in which additional features including binary amino acid compositions, AAindex properties and protein aggregation propensity were encoded to construct SVM classifiers [26]. In 2013, Chen *et al*. combined PseAAC, *k*-nearest neighbor (KNN) and AAindex to construct UbiProber for both general and species-specific predictions [25]. Using an iterative approach, Walsh *et al*. reported RUBI as a rapid genome-scale predictor for lysine ubiquitination, whereas bi-directional recurrent neural networks were incorporated with SVM to integrate the sequence- and structure-based features [27]. Later, iUbiq-Lys was released by incorporating PseAAC, PSSM and gray system model [28]. By developing UbiSite with a two-layered SVM model, Huang *et al*. adopted four widely used features including PseAAC, PSSM, positional-weighted matrix (PWM) and ASA and extracted substrate motifs using the MDDLogo [29]. To evaluate the performance of different features, Nguyen *et al*. developed a new framework using SVM, and the motif-based models derived from MDDLogo exhibited the best accuracy [52]. Also, Wang *et al*. constructed an SVM-based method known as ESA-UbiSite, in which 31 AAindex properties were selected by an optimization approach [30]. In 2018, Liu *et al*. reported a comprehensive web server called PTM-site-specific modification profile (ssMP), which provided predictions for multiple types of PTM sites including lysine ubiquitination sites. For each PTM type, ssMP was generated from both local sequence and proximal PTMs, and SVM classifier was then adopted to construct the computational model [31]. Through the integration of various features including AAindex, ASA, SS, BTA and PSSM, Li *et al*. developed an R package named PTMscape for the prediction of various PTM sites including lysine ubiquitination sites, based on linear SVM [32].

Besides SVM, machine learning algorithms including RF and KNN were also adopted to predict lysine ubiquitination. In 2010, Radivojac *et al*. used RF to construct UbPred, which integrated 586 sequence features. Based on the same RF algorithm [21], Zhao *et al*. integrated four features including PseAAC, PSSM, AAindex and disorder score and developed an ensemble model via voting [49]. Also, Lee *et al*. reported UbSite based on a radial basis function network, which combined the features of PseAAC, CKSAAP, PSSM and ASA [22]. Using a feature selection procedure, 456 features including PSSM, AAindex and disorder score were extracted by Cai *et al*., and KNN algorithm was adopted to develop a novel ubiquitination site predictor [50]. Similar features were also incorporated by WPNNA, a new classifier based on an optimized KNN algorithm [24]. Moreover, Pejaver *et al*. designed a LR-based tool named ModPred for the prediction of >20 types of PTM sites, and four feature types including sequence-based, physicochemical, structural properties and evolutionary properties were integrated for model training [33]. Later, Nguyen *et al.* used profile hidden Markov model to build several models based on identified motifs of existing sites [51]. In addition, the eXtreme gradient boosting (XGBoost) algorithm was adopted by Liu *et al.* to develop UbiSite-XGBoost, a new predictor for general ubiquitination sites, and various features including PseAAC, CKSAAP, AAindex, PsePSSM, BLOSUM62, adapted normal distribution bi-profile Bayes and encoding based on grouped weight were integrated [39].

Recently, with the accumulation of ubiquitination sites, advances in deep learning provided a great opportunity for big data training. In 2018, He *et al.* constructed deepUbiquitylation, which combined DNN and CNN to encode three features as OBC, AAindex and PSSM [34]. Later, Fu *et al.* designed a CNN-based framework DeepUbi [35]. The performances of four features including OBC, AAindex, PseAAC and CKSAAP were evaluated, and the combination of OBC and CKSAAP obtained the highest AUC value. Meanwhile, a new computational tool, MUscADEL, was reported by Chen *et al.* for lysine PTMs prediction [36]. An extended RNN framework was constructed with a word embedding layer to extract sequence features. More recently, DeepTL-Ubi was constructed with a densely connected CNN, and transfer learning was performed to extend the prediction for multiple species with the feature of OBC [42]. However, MultiLyGAN released by Yang *et al.* adopted conditional Wasserstein generative adversarial network to eliminate data imbalance, and the RF algorithm was used to generate models for multiple lysine modifications [43]. Beyond general prediction, the development of tools to predict plant ubiquitylation sites is also prevalent. In 2020, Wang *et al.* released a CNN-based architecture called DL-plant-ubsites-prediction, which implemented a word-embedding method based on features of PseAAC, CKSAAP, PWM and sequence logo [37]. In parallel, MusiteDeep was developed to provide efficient predictors for numerous types of PTM sites including ubiquitination sites [38]. In MusiteDeep, two CNN-based networks were integrated to generate the final model for each PTM type with OBC feature [38]. By integrating CNN with long short-term memory, Siraj *et al.* constructed UbiComb for the prediction of plant ubiquitination sites [40]. In addition, an *Arabidopsis thaliana*-specific predictor CNNAtuUbi was designed by Wang *et al.* using a CNN framework [41]. No tools have been developed for the prediction of exact ssESRs from the protein sequences.

The data statistics of known E3-specific ubiquitination sites

Considering that PLMD provides no information on upstream E3s, we collected 1311 experimentally identified ssESRs from the literature for the development of E3-specific models. Using the hierarchical classifications of iUUCD [1], these ssESRs were hierarchically clustered at different levels, including 6 groups, 4 subgroups, 15 families and 93 single E3s, and positive and negative data sets were generated for each cluster.

In our results, four groups of Really Interesting New Gene (RING), Cullin RING, RING-between RING–RING (RBR) and Homologous to the E6AP Carboxyl Terminus (HECT) covered 98.17% (1287) of total ssESRs, and RING contained the largest positive data set with 754 ssESRs in 221 proteins (Figure 2**A**). In contrast, only 9 ssESRs were reported to be ubiquitinated by Recognition components of the N-end rule pathway (N-recognin) E3s, and 17 ssESRs from 6 ubiquitinated proteins were classified into the other group. Obviously, these sites were abundant in eight E3 clusters at the family level, including RING/RING, RBR/RBR, RING/U-box, HECT/HECT, Cullin RING/DDB1-CUL4-X-box/DDB1-binding WD40 protein (Cullin RING/D-CX/DWD), Cullin RING/Skp1-Culline-F-box protein/F-box (Cullin RING/SCF/F-box), Cullin RING/BTB, Cul3 and RBX1 form a Cul3-based ligase/BTB_3-box (Cullin RING/BCR/BTB_3-box) and Cullin RING/ECS/Suppressors Of Cytokine Signalling_Von-Hippel Lindau_BC-box (Cullin RING/ECS/SOCS_VHL_BC-box) (Figure 2**A**). Again, the RING/RING family was matched with most substrates of 676 ssESRs. The comparison of substrates across the four major E3 groups with ≥30 protein substrates demonstrated a low coverage among different groups. Only 7 proteins were known to be modified by 3 types of E3 groups, and 21 substrates were shared by RING and Cullin RING groups (Figure 2**B**). In addition, the sequence logos of these groups were generated for the investigation of potential substrate motifs, which demonstrated diverse patterns for E3 groups (Figure 2**C**). For example, besides a high frequency of serine (S) detected at position +4 for both RING and Cullin RING groups, glutamic acid (E) and proline (P) showed high probabilities at position −3 and + 3 of RING, respectively, whereas a signature of arginine (R) was found at position −6 of Cullin RING. Similarly, different patterns of amino acids were detected with RBR and HECT groups, such as aspartic acid (D) at position −3 for RBR and valine (V) at position −4 for HECT (Figure 2**C**). The results suggested that different E3s prefer to recognize different sequence profiles for substrate ubiquitination.

Next, GO-based enrichment analyses were conducted to detect the biological processes regulated by four major E3 groups with ≥30 substrates (Figure 2**D**). Interestingly, the coverage of biological processes was much higher than substrates, such as the process of ‘positive regulation of transcription by RNA polymerase II’ (GO: 0045944), which was enriched in top five enriched biological processes of RING, Cullin RING and HECT at the same time. Also, ‘negative regulation of apoptotic process’ (GO: 0043066) and ‘protein deubiquitination’ (GO:0016579) were detected simultaneously with two

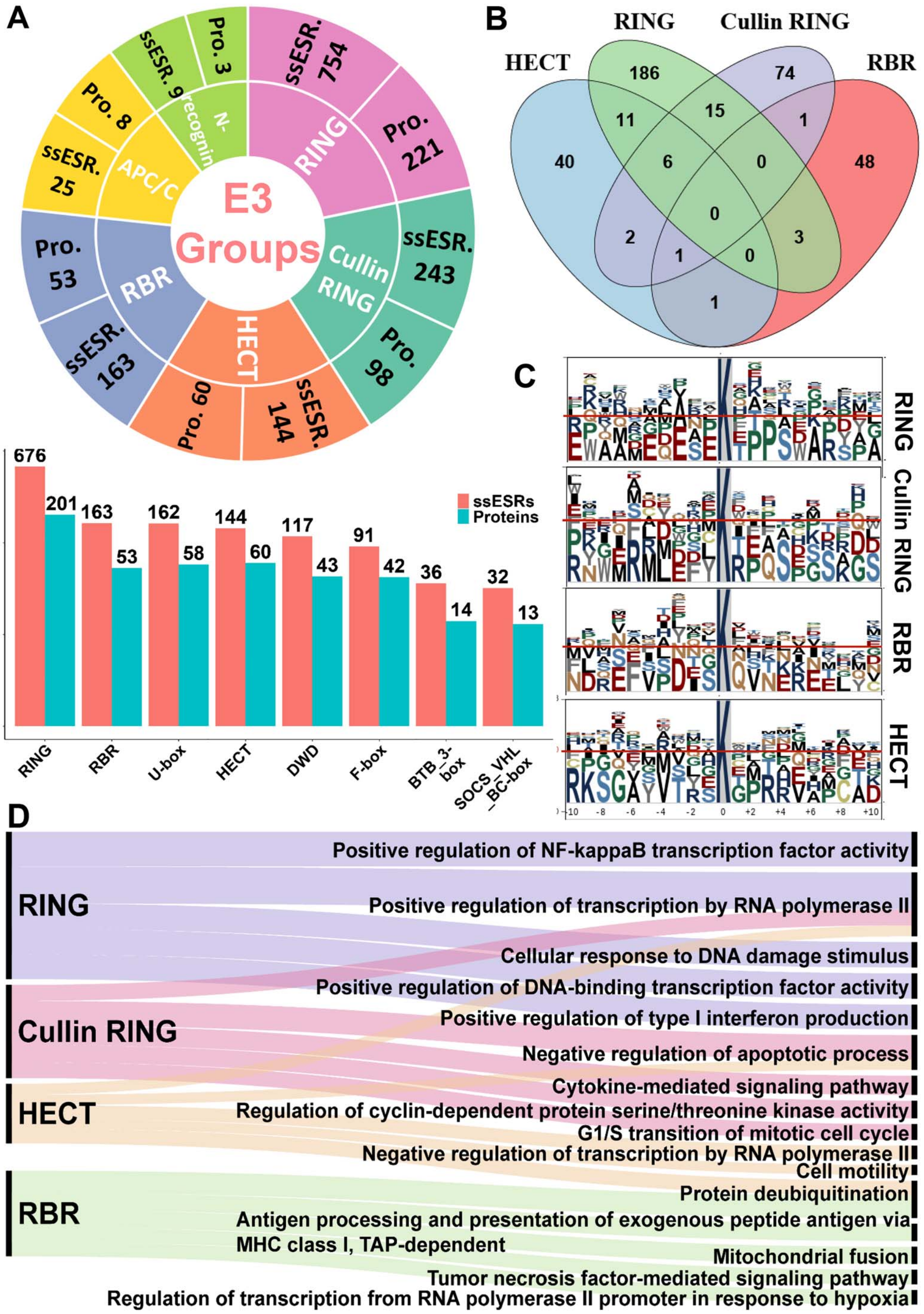


Figure 2. The analysis of the known E3-specific ubiquitination sites. (**A**) The number of known substrates of E3 groups and families with ≥30 known ubiquitination sites. More details are shown in Supplementary Table S1. (**B**) The overlap of protein substrates from four major E3 groups with ≥30 substrates. (**C**) The sequence logos of four major E3 groups. (**D**) The GO-based enrichment analysis of protein substrates from four major E3 groups.

different E3 groups. The results indicated that a considerable number of processes were mutually regulated by different types of E3s. Using the mRNA expression data from TCGA [57], the correlation of the 177 E3s and 391 substrates was analyzed. The average Spearman's rank correlation coefficient (ρ) was calculated as 0.0487 (Supplementary Figure S1**A**), indicating a weak correlation of E3s and their targets at the transcriptional level. Moreover, we re-analyzed the time-course quantitative proteomic data from a recently published study [58], and the average ρ of 0.0614 supported a weak correlation of E3s and their targets at the translational level (Supplementary Figure S1**B**).

Development of a hybrid-learning framework for the prediction of ubiquitination sites

In the past two decades, various features have been adopted to construct the predictors for ubiquitination

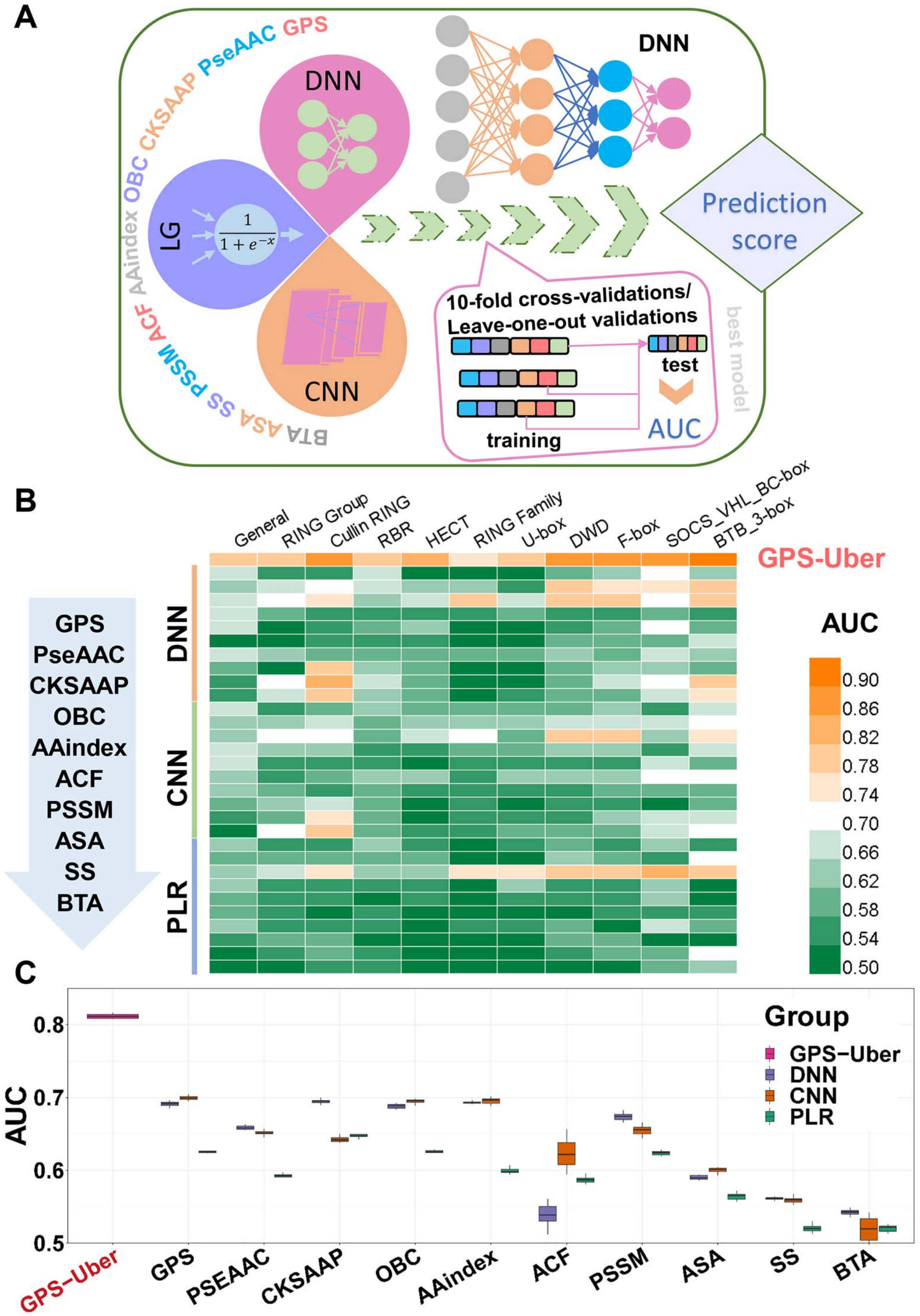


Figure 3. The hybrid-learning framework of GPS-Uber, as well as the performance values of different features. (**A**) For each USP(10,10) item, 10 types of features were encoded for model training with three algorithms including DNN, PLR and CNN. Then a vector containing 30 predicted scores was generated as input for an additional DNN framework to produce the final score. The 4-, 6-, 8- and 10-fold cross-validations and LOO were performed to evaluate the robustness of models. (**B**) For each feature, the AUC values of DNN, CNN and PLR were calculated for the general and E3-specific predictors with ≥30 known ubiquitination sites. (**C**) The distribution of AUC values from 10-fold cross-validations for the general predictor, on different types of features.

sites, and performance improvement was observed with the combination of different features [20–43, 49–52]. Meanwhile, algorithms based on conventional machine learning and deep learning were both widely used, whereas a systematic evaluation was yet to be performed. More importantly, although numerous tools were developed, prediction of exact ssESRs was still a great challenge. Here, we designed a new architecture, GPS-Uber, for the prediction of both general and E3-specific ubiquitination sites (Figure 3**A**). For each USP(10,10) in the training data set, 10 types of sequence- or structure-based features were individually encoded into 1D vectors and 2D matrices, respectively. For each feature, three models were separately constructed based on PLR, DNN and CNN algorithms, using the corresponding encoded vector or matrix. As a result, 30 scores were generated through the combination of features and algorithms, and vectors containing these scores were adopted by a new DNN model as inputs to obtain the final prediction scores for all USP(10,10) items.

The predictor for general ubiquitination sites was first constructed, and transfer learning was then adopted for implementation of E3-specific models. In total, 112 individual predictors were constructed by GPS-Uber. Using the training data set with 61 161 known ubiquitination sites, 10-fold cross-validations were conducted to evaluate the performance of models with ≥30 ubiquitination sites (Supplementary Table S5). The results demonstrated that the AUC values of sequence features were generally higher than those of structural features, especially in GPS-based peptide similarity, PseAAC and CKSAAP (Figure 3**B**). In general, deep learning algorithms showed higher AUC values than PLR (Figure 3**C**). For general prediction, most of the features exhibited small fluctuations of AUC values through 10-fold cross-validations (Figure 3**C**), supporting the robustness and stability of computational models. The final model that combined the three algorithms and 10 features reached an AUC value of 0.8106, which was significantly improved compared with single algorithms or features (Figure 3**C**). At the group level, Cullin RING got the highest AUC value of 0.8967 compared with RING (0.8188), RBR (0.8074) and HECT (0.8204), whereas the AUC values of E3 families ranged from 0.7767 (RING) to 0.9396 (BTB_3-box) (Figure 3**B**). Meanwhile, LOO validations were conducted for predictors with <30 ubiquitination sites (Supplementary Table S5), and the incorporation of 10 types of features and three types of algorithms significantly improved the prediction performance for all data sets, further supporting the superiority of GPS-Uber.

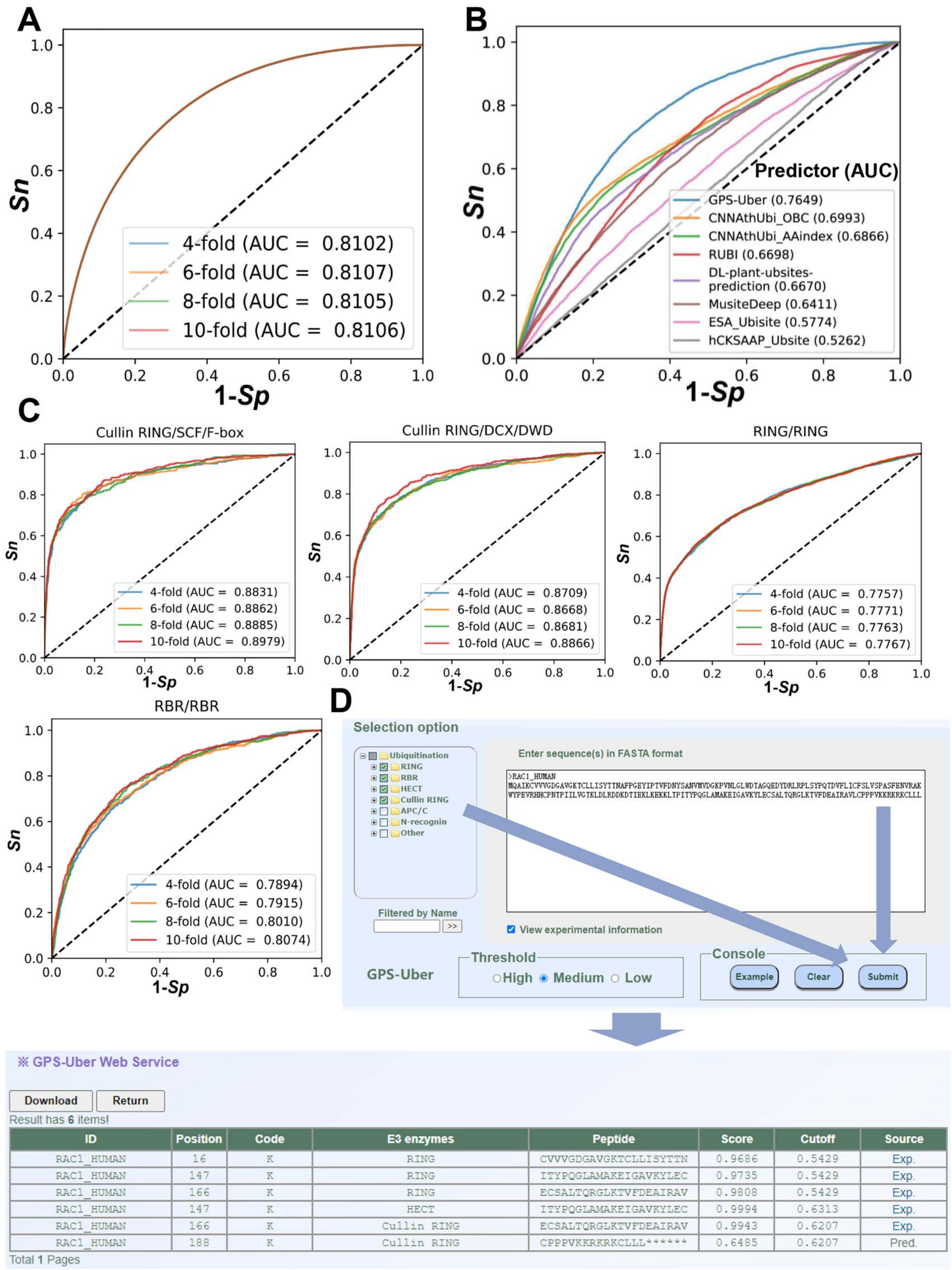
Performance evaluation and comparison

Besides 10-fold cross-validations, 4-, 6- and 8-fold cross-validations were also performed using the initial training data set. For the general model, the AUC values were calculated as 0.8102, 0.8107, 0.8105 and 0.8106 for 4-, 6-, 8- and 10-fold cross-validations, respectively (Figure 4**A**). To demonstrate the superiority of GPS-Uber, a comparison was conducted between GPS-Uber and other existing tools with the independent testing data. Although 28 methods were reported for predicting general or species-specific ubiquitination sites, applicable online services or executable codes were provided by only six tools, including hCKSAAP_UbSite [26], RUBI [27], ESA_Ubisite [30], DL-plant-ubsites-prediction [37], CNNAthUbi [41] and MusiteDeep [38]. These tools were designed for general prediction and no E3-specific model was provided. Using a timestamp-based method [38], 55 426 sites reported before 2016 were used for additional model training by GPS-Uber, whereas 5735 sites released after 2016 were directly submitted to GPS-Uber and other existing tools for an unbiased comparison. For each tool, the ROC curve was illustrated and AUC value was calculated (Figure 4**B**). The results demonstrated that GPS-Uber exhibited a highly competitive accuracy compared with other tools, and the AUC values were 0.7649, 0.6993/0.6866, 0.6698, 0.6670, 0.6411, 0.5774, 0.5262 for GPS-Uber, CNNAthUbi, RUBI, DL-plant-ubsites-prediction, MusiteDeep, ESA_Ubisites and hCKSAAP_UbSite, respectively (Figure 4**B**).

To further evaluate the robustness of GPS-Uber, we conducted an additional validation. For the initial training data set containing 61 161 ubiquitination sites, the USP(10,10) items were randomly separated into five equal parts, with the same distribution of positive data versus negative data. Then, four parts were taken as a new training data set, and the 10-fold cross-validation was performed for parameter optimization, whereas the remaining one part was taken as an independent testing data set for performance evaluation. This procedure was repeated five times until each of the five parts was used as the testing data set for one time. The AUC values ranged from 0.7449 to 0.7536 (Supplementary Figure S2**A**), supporting the stability and superiority of GPS-Uber. In addition, we tested 1D CNN directly using the vectors encoded from the 10 features, and the performance was slightly reduced against 2D CNN when integrated in GPS-Uber (Supplementary Figure S2**B**).

Of note, GPS-Uber provided multiple unique predictors to predict E3-specific ubiquitination sites for the first time, whereas additional 4-, 6- and 8-fold cross-validations were also conducted for models with ≥30 ubiquitination sites. Due to the page limitation, the ROC curves of four E3 families including Cullin RING/SCF/F-box, Cullin RING/DCX/DWD, RING/RING and RBR/RBR were shown (Figure 4**C**). For Cullin RING/SCF/F-box, the AUC values of 4-, 6-, 8- and 10-fold cross-validations were 0.8831, 0.8862, 0.8885 and 0.8979. Similar results were observed for Cullin RING/DCX/DWD with AUC values of 0.8709, 0.8668, 0.8681 and 0.8866, respectively, whereas satisfying performance values were also detected with RING/RING and RBR/RBR. To investigate the site specificity of E3s, four E3s that belonged to HECT group (NEDD4 and ITCH) and RING group (MDM2 and STUB1) were selected. For each E3-specific predictor, the training data sets of the remaining three E3s were individually used to test its performance. From the results, it could be found that each E3-specific predictor only exhibited a much higher accuracy on its own training data set, supporting a strong specificity of E3s on recognition their target sites (Supplementary Figure S2**C**). Taken together, our results indicated the promising robustness and accuracy of GPS-Uber for both general and E3-specific predictions.

For convenience, a user-friendly web server was developed for GPS-Uber (Figure 4**D**). The clickable hierarchical



Enter sequence(s) in FASTA format

```
>RAC1_HUMAN
MQAIKCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVMVDGKPVNLGLWDTAGQEDYDRLRPLSYPQTDVFLICFSLVSPASFENVRAK
WYPEVRHHCPNTPIILVGTKLDLRDDKDTIEKLKEKKLTPITYPQGLAMAKEIGAVKYLECSALTQRGLKTVFDEAIRAVLCPPPVKKRKRKCLLL
```

☑ RING ☑ RBR ☑ HECT ☑ Cullin RING □ APC/C □ N-recognin □ Other

☑ View experimental information

Threshold ○High ● Medium ○ Low

※ GPS-Uber Web Service

Result has 6 items!

ID	Position	Code	E3 enzymes	Peptide	Score	Cutoff	Source
RAC1_HUMAN	16	K	RING	CVVVGDGAVGKTCLLISYTTN	0.9686	0.5429	Exp.
RAC1_HUMAN	147	K	RING	ITYPQGLAMAKEIGAVKYLEC	0.9735	0.5429	Exp.
RAC1_HUMAN	166	K	RING	ECSALTQRGLKTVFDEAIRAV	0.9808	0.5429	Exp.
RAC1_HUMAN	147	K	HECT	ITYPQGLAMAKEIGAVKYLEC	0.9994	0.6313	Exp.
RAC1_HUMAN	166	K	Cullin RING	ECSALTQRGLKTVFDEAIRAV	0.9943	0.6207	Exp.
RAC1_HUMAN	188	K	Cullin RING	CPPPVKKRKRKCLLL******	0.6485	0.6207	Pred.

Total 1 Pages

Figure 4. Performance evaluation and comparison of GPS-Uber with other existing tools. (**A**) The ROC curves and AUC values of *n*-fold cross-validations of the general model. (**B**) Comparison of the general model of GPS-Uber with other existing predictors using an independent testing data set. (**C**) The accuracy values of E3-specific predictors for E3 families, including Cullin RING/SCF/F-box, Cullin RING/DCX/DWD, RING/RING and RBR/RBR. (**D**) Interface of the online service of GPS-Uber.

classification tree of E3s was located in the left panel, which enabled various combinations for the prediction of different E3s. Then, single or multiple protein sequences in FASTA format could be submitted after the selection of E3s, and the prediction results would be presented after a few seconds, which contains potential ubiquitination sites with seven types of information, including 'ID', 'Position', 'Code', 'E3 enzymes', 'Peptides', 'Score' and 'Cutoff' (Figure 4D). Also, we implemented an option 'View experimental information', which could be ticked to additionally present a column of 'Source' in the prediction page (Figure 4D). The experimental evidence of predicted sites could be viewed by clicking on 'Exp.' if available (Figure 4D). Furthermore, an option 'Filtered by Name' was added to enable the rapid search of E3s in the left hierarchical tree (Figure 4D).

Moreover, an additional module was implemented for the prediction using gene names, protein names and/or UniProt accession numbers for eight model species including *H. sapiens*, *Mus musculus*, *Rattus norvegicus*, *Danio rerio*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *A. thaliana* and *Saccharomyces cerevisiae*, and could be accessed at http://gpsuber.biocuckoo.cn/online_name.php (Supplementary Figure S3). In addition, the Ras-related C3 botulinum toxin substrate 1 (RAC1) protein was used as an example for new users. The online service of GPS-Uber was implemented using PHP and JavaScript and has been tested on multiple web browsers including Google Chrome 92.0, Mozilla Firefox 89, Opera 77.0 and Safari 14.1.1. In summary, GPS-Uber was designed to provide a handy resource for the research of ubiquitination.

Prediction of potential cancer-associated ubiquitination events

Considering that numerous signaling pathways were regulated by ubiquitination in human, and aberrant E3s or ubiquitination events have been reported to be associated with cancers, we considered whether GPS-Uber could be used to reveal new relationships between ubiquitination and cancers, which would provide novel insight for the treatment of cancers. A total of 707 cancer proteins maintained in COSMIC [54] were downloaded as input for GPS-Uber, and the ubiquitination sites regulated by the six E3 groups were predicted with the medium threshold. Strikingly, only two cancer proteins were excluded with no site predicted, and 674 (95.33%) proteins were predicted to be ubiquitinated by three or more types of E3s (Figure 5**A**). Moreover, the statistic of identified ubiquitination sites showed that 644 proteins were matched with >10 sites (Figure 5**B**), which might be partly related to the length of protein sequences (Supplementary Table S6). These results demonstrated the prevalence and importance of cancer-associated ubiquitination events.

GO-based enrichment analyses were performed for 154 cancer proteins with predicted ubiquitination sites of all the six E3 groups, and transcription-related pathways were enriched as the dominating process, including 'regulation of transcription, DNA-templated' (GO:0045893 and GO:0006355), 'regulation of transcription by RNA polymerase II' (GO:0045944 and GO:0000122) and 'chromatin remodelling' (GO:0006338) pathways (Figure 5**C**). In addition, biological processes associated with cell cycle and cellular homeostasis were also detected. Similar results were observed when same analyses were conducted with individual E3 category (Figure 5**D**). Interestingly, phosphorylation-related pathways such as 'positive regulation of kinase activity' (GO: 0033674) and 'peptidyl-tyrosine phosphorylation' (GO:0018108) were also identified, which suggested cancer-related crosstalk between ubiquitination and phosphorylation. The human RAC1, an important cancer protein, was selected by GPS-Uber as an example for E3-specific ubiquitination sites prediction. It has been reported that TRAF6, an E3 belonging to RING/RING family, could aggravate ischemic stroke through the ubiquitination of RAC1 at K16 [63]. In addition, K147 in RAC1 was found to be ubiquitinated by IAPs, which were also classified into RING/RING family [64]. Beyond RING/RING family, E3s from Cullin RING/SCF/F-box family could also regulate the degradation of RAC1 through the ubiquitination at K166 [65]. The prediction of GPS-Uber showed highly consistent with these experimental results when categories for RING/RING and Cullin RING/SCF/F-box were chosen, and more information could be provided by using other predictors (Figure 5**E**).

Discussion

Since the discovery of ubiquitin in 1978, the underlying mechanisms of ubiquitination were always the hotspots in the field of PTMs [2, 4, 66]. A broad spectrum of biological processes and diseases, such as protein degradation and cancers, has been reported to be regulated by ubiquitination. As a reversible covalent modification, multiple enzymes were involved in the processes of ubiquitination and deubiquitination, and the substrate specificity of ubiquitination was largely controlled by E3s. Thus, the identification of ubiquitination sites and upstream E3s plays a crucial role in understanding the molecular mechanisms of ubiquitination. Besides conventional LTP experimental strategies, the development of a variety of HTP assays enabled the large-scale identification of ubiquitination sites and led to the emergence of various databases, such as mUbiSIDa [67] and PLMD [44]. Based on these data resources, numerous computational methods have been developed with different features and algorithms and facilitated the rapid identification of potential ubiquitination sites.

In this study, 10 types of well-used features were first integrated to construct a model for general ubiquitination sites prediction. For the integration of conventional machine learning and deep learning algorithms, a hybrid-learning architecture based on PLR, DNN and CNN was constructed, and 10-fold cross-validation was performed with the final model, exhibiting an AUC value of 0.7649 on the independent testing data set (Figure 4**B**). Compared with six existing tools, GPS-Uber showed a highly competitive accuracy on the prediction of general ubiquitination sites (Figure 4**B**). Since the existing tools were mainly focused on general predictions, and predicting E3-specific ubiquitination sites was still not available. To fill this gap, a total of 1311 experimentally identified ssESRs were collected from 637 LTP studies, and E3s were carefully mapped to human proteome. In 2017, we developed a database called iUUCD for ubiquitin and ubiquitin-like conjugations that contained comprehensive annotations and hierarchical classifications for 1153 known E3s from multi-species [1]. In this study, E3s were manually classified at four levels using the information of iUUCD, and transfer learning

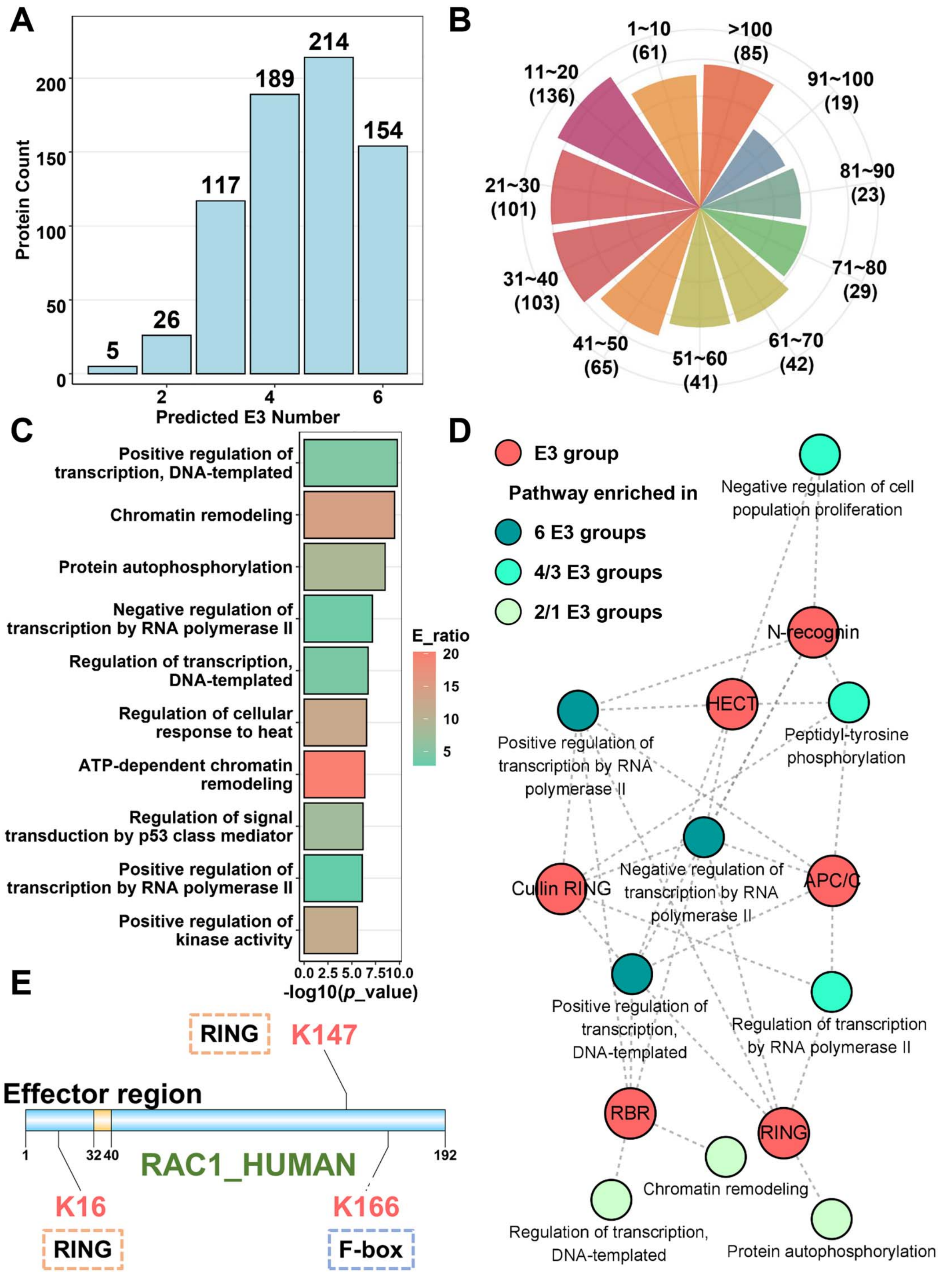


Figure 5. Cancer-associated ubiquitination events predicted by GPS-Uber. (**A**) The distribution of cancer proteins predicted to be ubiquitinated by six E3 groups with GPS-Uber. (**B**) The distribution of numbers of predicted ubiquitination sites from cancer proteins. (**C**) The GO-based enrichment analysis of 154 cancer proteins predicted to be ubiquitinated by all the six E3 groups. (**D**) A network of pathways predicted to be regulated by different E3 groups. (**E**) Predicted ubiquitination sites and upstream E3 families of RAC1_HUMAN with GPS-Uber were visualized by DOG 2.0 [62].

was then conducted with each E3 category for model construction. At last, we implemented the online service of GPS-Uber, which provided 111 E3-specific predictors and one general predictor for ubiquitination sites. In GPS-Uber, E3 clusters with ≥3 ubiquitination sites were kept, and the reliability of these models might be relatively lower. However, including these E3 clusters could provide a more comprehensive prediction and facilitate further experimental validations. For example, we previously developed a tool named GPS 2.0 for the prediction of kinase-specific phosphorylation sites [68]. A predictor, CK1/VRK, was trained by only four sites. Based on the prediction of GPS 2.0, Choi *et al.* successfully validated a novel phosphorylation site, Ser6 in hnRNP

A1, to be specifically modified by VRK1, and such a phosphorylation event plays a critical role in telomere maintenance [69]. Since previous studies suggested that ubiquitination takes part in the regulation of various cancers, we speculated that whether GPS-Uber could be used to reveal novel cancer-associated ubiquitination signatures. Predictors of six E3 groups were adopted for predicting potential ubiquitination sites in known cancer proteins, and the GO-based enrichment results were highly consistent with known studies. Transcription-related pathways were shown to be regulated by all types of E3s, whereas similar functions have already been observed by other researchers [70, 71].

In GPS-Uber, the basic hypothesis is that short peptides around lysine residues provided the major specificity of E3-specific ubiquitination. Following evaluations revealed a promising accuracy of GPS-Uber on predicting E3-specific ubiquitination sites (Figures 3**B**, 4**C**, and Supplementary Figure S2**A**), supporting the existence of such a modification specificity at the site level. However, ubiquitination sites are not the binding sites of E3s, which specifically recognize primary degrons in substrates for the interaction [55]. Previously, it was reported that ubiquitination sites tend to appear very close to primary degrons (often within 20 residues) [55]. Using the sequence motifs of 27 known degrons for 11 E3 clusters (Supplementary Table S3), potential degron sequences were detected from their known E3-specific substrates, and the distances to their proximal ubiquitination sites modified by the same E3s were counted. From the results, it was found that most of the E3-specific ubiquitination sites do not locate close to their corresponding primary degrons, whereas only 36.00%, 32.14% and 27.27% of anaphase-promoting complex/cyclosome (APC/C)-, BRAC1- and VHL-specific ubiquitination sites had a proximal primary degron within 20 residues (Supplementary Figure S4**A**). In addition, we extended USP(m, n) to USP(15, 15), USP(20, 20), USP(25, 25) and USP(30, 30) for four E3 clusters, including RING/RING/MDM2, RING/RING/BRCA1, RING/RING/TRAF6 and Cullin RING/SCF and compared the performance to USP(10, 10). From the results, it could be found that the AUC values were slightly increased with longer flanking regions (Supplementary Figure S4**B**–**E**), indicating that considering potential primary degrons did not significantly improve the accuracy for the prediction of E3-specific ubiquitination sites.

Because only the features of flanking sequences around ubiquitination sites were considered in GPS-Uber, the ESIs should be predetermined by HTP experiments or computational predictions together with following LTP experimental validations. Incorporation of the state-of-the-art computational methods for predicting ESIs into GPS-Uber will be our future plan, and such an integration will be crucial to accurately predict both ESIs and ssESRs and provide more useful clues to identify functionally important ubiquitination events *in vivo*. Also, we will extend the benchmark data sets used in this study. More general and E3-specific ubiquitination sites will be integrated to improve the performance of GPS-Uber, and new features and algorithms should be considered. Moreover, a similar strategy could be adopted with deubiquitinating enzymes for the research of the whole ubiquitination system. In addition, since the crosstalk between ubiquitination and other PTMs like phosphorylation has been confirmed to be important in many cellular processes [72, 73], an improved algorithm incorporated with relationships among different PTM types will be useful to further improve the prediction accuracy. Nevertheless, GPS-Uber will be continuously maintained and improved for academic research.

Key Points

- We reviewed existing tools for predicting general ubiquitination sites, including the information of various features, algorithms and training data sets.
- We developed a novel hybrid-learning framework for predicting lysine ubiquitination sites, which integrated 10 types of features and three types of machine learning algorithms including penalized logistic regression (PLR), deep neural network (DNN) and convolutional neural network (CNN).
- We constructed 111 individual E3-specific predictors through further transfer learning and developed a new tool named GPS-Uber for predicting both general and E3-specific lysine ubiquitination sites, exhibiting a higher accuracy than other existing tools.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgement

The authors thank Drs Han Cheng, Haodong Xu and Hang Xu for their helpful comments.

Funding

Chinese Postdoctoral Science Foundation (2020 M682395, 2018 M642816 and 2019 T120648), Natural Science Foundation of China (31930021 and 31970633), Fundamental Research Funds for the Central Universities (2019kfyRCPY043) and Changjiang Scholars Program of China.

References

1. Zhou J, Xu Y, Lin S, *et al.* iUUCD 2.0: an update with rich annotations for ubiquitin and ubiquitin-like conjugations. *Nucleic Acids Res* 2018;**46**:D447–53.
2. Simoneschi D, Rona G, Zhou N, *et al.* CRL4(AMBRA1) is a master regulator of D-type cyclins. *Nature* 2021;**592**:789–93.
3. Pohl C, Dikic I. Cellular quality control by the ubiquitin-proteasome system and autophagy. *Science* 2019;**366**:818–22.

4. Ciehanover A, Hod Y, Hershko A. A heat-stable polypeptide component of an ATP-dependent proteolytic system from reticulocytes. *Biochem Biophys Res Commun* 1978;**81**:1100–5.
5. Scheffner M, Nuber U, Huibregtse JM. Protein ubiquitination involving an E1-E2-E3 enzyme ubiquitin thioester cascade. *Nature* 1995;**373**:81–3.
6. Zheng N, Shabek N. Ubiquitin ligases: structure, function, and regulation. *Annu Rev Biochem* 2017;**86**:129–57.
7. Bernassola F, Chillemi G, Melino G. HECT-type E3 ubiquitin ligases in cancer. *Trends Biochem Sci* 2019;**44**:1057–75.
8. Popovic D, Vucic D, Dikic I. Ubiquitination in disease pathogenesis and treatment. *Nat Med* 2014;**20**:1242–53.
9. Manasanch EE, Orlowski RZ. Proteasome inhibitors in cancer therapy. *Nat Rev Clin Oncol* 2017;**14**:417–33.
10. Iconomou M, Saunders DN. Systematic approaches to identify E3 ligase substrates. *Biochem J* 2016;**473**:4083–101.
11. O'Connor HF, Huibregtse JM. Enzyme-substrate relationships in the ubiquitin system: approaches for identifying substrates of ubiquitin ligases. *Cell Mol Life Sci* 2017;**74**:3363–75.
12. Rayner SL, Morsch M, Molloy MP, *et al.* Using proteomics to identify ubiquitin ligase-substrate pairs: how novel methods may unveil therapeutic targets for neurodegenerative diseases. *Cell Mol Life Sci* 2019;**76**:2499–510.
13. Yen HC, Xu Q, Chou DM, *et al.* Global protein stability profiling in mammalian cells. *Science* 2008;**322**:918–23.
14. Yen HC, Elledge SJ. Identification of SCF ubiquitin ligase substrates by global protein stability profiling. *Science* 2008;**322**: 923–9.
15. Low TY, Peng M, Magliozzi R, *et al.* A systems-wide screen identifies substrates of the SCFbetaTrCP ubiquitin ligase. *Sci Signal* 2014;**7**:rs8.
16. Elia AE, Boardman AP, Wang DC, *et al.* Quantitative proteomic atlas of ubiquitination and acetylation in the DNA damage response. *Mol Cell* 2015;**59**:867–81.
17. Li Y, Xie P, Lu L, *et al.* An integrated bioinformatics platform for investigating the human E3 ubiquitin ligase-substrate interaction network. *Nat Commun* 2017;**8**:347.
18. Wang X, Li Y, He M, *et al.* UbiBrowser 2.0: a comprehensive resource for proteome-wide known and predicted ubiquitin ligase/deubiquitinase-substrate interactions in eukaryotic species. *Nucleic Acids Res* 2021, gkab962. doi: https://doi.org/10.1093/nar/gkab962.
19. Chen D, Liu X, Xia T, *et al.* A multidimensional characterization of E3 ubiquitin ligase and substrate interaction network. *iScience* 2019;**16**:177–91.
20. Tung CW, Ho SY. Computational identification of ubiquitylation sites from protein sequences. *BMC Bioinformatics* 2008;**9**:310.
21. Radivojac P, Vacic V, Haynes C, *et al.* Identification, analysis, and prediction of protein ubiquitination sites. *Proteins* 2010;**78**: 365–80.
22. Lee TY, Chen SA, Hung HY, *et al.* Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. *PLoS One* 2011;**6**:e17331.
23. Chen Z, Chen YZ, Wang XF, *et al.* Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. *PLoS One* 2011;**6**:e22930.
24. Feng KY, Huang T, Feng KR, *et al.* Using WPNNA classifier in ubiquitination site prediction based on hybrid features. *Protein Pept Lett* 2013;**20**:318–23.
25. Chen X, Qiu JD, Shi SP, *et al.* Incorporating key position and amino acid residue features to identify general and species-specific ubiquitin conjugation sites. *Bioinformatics* 2013;**29**:1614–22.
26. Chen Z, Zhou Y, Song J, *et al.* hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. *Biochim Biophys Acta* 2013;**1834**: 1461–7.
27. Walsh I, Di Domenico T, Tosatto SC. RUBI: rapid proteomic-scale prediction of lysine ubiquitination and factors influencing predictor performance. *Amino Acids* 2014;**46**:853–62.
28. Qiu WR, Xiao X, Lin WZ, *et al.* iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. *J Biomol Struct Dyn* 2015;**33**: 1731–42.
29. Huang CH, Su MG, Kao HJ, *et al.* UbiSite: incorporating two-layered machine learning method with substrate motifs to predict ubiquitin-conjugation site on lysines. *BMC Syst Biol* 2016;**10**(Suppl 1):6.
30. Wang JR, Huang WL, Tsai MJ, *et al.* ESA-UbiSite: accurate prediction of human ubiquitination sites by identifying a set of effective negatives. *Bioinformatics* 2017;**33**:661–8.
31. Liu Y, Wang MH, Xi JN, *et al.* PTM-ssMP: a web server for predicting different types of post-translational modification sites using novel site-specific modification profile. *Int J Biol Sci* 2018;**14**: 946–56.
32. Li GXH, Vogel C, Choi H. PTMscape: an open source tool to predict generic post-translational modifications and map modification crosstalk in protein domains and biological processes. *Mol Omics* 2018;**14**:197–209.
33. Pejaver V, Hsu WL, Xin FX, *et al.* The structural and functional signatures of proteins that undergo multiple events of post-translational modification. *Protein Sci* 2014;**23**: 1077–93.
34. He F, Wang R, Li JG, *et al.* Large-scale prediction of protein ubiquitination sites using a multimodal deep architecture. *BMC Syst Biol* 2018;**12**.
35. Fu H, Yang Y, Wang X, *et al.* DeepUbi: a deep learning framework for prediction of ubiquitination sites in proteins. *BMC Bioinformatics* 2019;**20**:86.
36. Chen Z, Liu X, Li F, *et al.* Large-scale comparative assessment of computational predictors for lysine post-translational modification sites. *Brief Bioinform* 2019;**20**:2267–90.
37. Wang H, Wang Z, Li Z, *et al.* Incorporating deep learning with word embedding to identify plant Ubiquitylation sites. *Front Cell Dev Biol* 2020;**8**:572195.
38. Wang D, Liu D, Yuchi J, *et al.* MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Res* 2020;**48**: W140–6.
39. Liu Y, Jin S, Song L, *et al.* Prediction of protein ubiquitination sites via multi-view features based on eXtreme gradient boosting classifier. *J Mol Graph Model* 2021;**107**:107962.
40. Siraj A, Lim DY, Tayara H, *et al.* UbiComb: a hybrid deep learning model for predicting plant-specific protein ubiquitylation sites. *Genes* 2021;**12**:717.
41. Wang XF, Yan RX, Chen YZ, *et al.* Computational identification of ubiquitination sites in *Arabidopsis thaliana* using convolutional neural networks. *Plant Mol Biol* 2021;**105**:601–10.
42. Liu Y, Li A, Zhao XM, *et al.* DeepTL-Ubi: a novel deep transfer learning method for effectively predicting ubiquitination sites of multiple species. *Methods* 2021;**192**:103–11.
43. Yang Y, Wang H, Li W, *et al.* Prediction and analysis of multiple protein lysine modified sites based on conditional Wasserstein generative adversarial networks. *BMC Bioinformatics* 2021;**22**:171.
44. Xu H, Zhou J, Lin S, *et al.* PLMD: an updated data resource of protein lysine modifications. *J Genet Genomics* 2017;**44**:243–50.

45. Fu L, Niu B, Zhu Z, *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2.
46. Wang C, Xu H, Lin S, *et al.* GPS 5.0: an update on the prediction of kinase-specific phosphorylation sites in proteins. *Genomics Proteomics Bioinformatics* 2020;**18**:72–80.
47. Ning W, Xu H, Jiang P, *et al.* HybridSucc: a hybrid-learning architecture for general and species-specific succinylation site prediction. *Genomics Proteomics Bioinformatics* 2020;**18**:194–207.
48. Ning W, Jiang P, Guo Y, *et al.* GPS-palm: a deep learning-based graphic presentation system for the prediction of S-palmitoylation sites in proteins. *Brief Bioinform* 2021;**22**: 1836–47.
49. Zhao X, Li X, Ma Z, *et al.* Prediction of lysine ubiquitylation with ensemble classifier and feature selection. *Int J Mol Sci* 2011;**12**: 8347–61.
50. Cai Y, Huang T, Hu L, *et al.* Prediction of lysine ubiquitination with mRMR feature selection and analysis. *Amino Acids* 2012;**42**: 1387–95.
51. Nguyen VN, Huang KY, Huang CH, *et al.* Characterization and identification of ubiquitin conjugation sites with E3 ligase recognition specificities. *BMC Bioinformatics* 2015;**16**(Suppl 1):S1.
52. Nguyen VN, Huang KY, Huang CH, *et al.* A new scheme to characterize and identify protein ubiquitination sites. *IEEE/ACM Trans Comput Biol Bioinform* 2017;**14**:393–403.
53. Huntley RP, Sawford T, Mutowo-Meullenet P, *et al.* The GOA database: gene ontology annotation updates for 2015. *Nucleic Acids Res* 2015;**43**:D1057–63.
54. Forbes SA, Beare D, Gunasekaran P, *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 2015;**43**:D805–11.
55. Guharoy M, Bhowmick P, Sallam M, *et al.* Tripartite degrons confer diversity and specificity on regulated protein degradation in the ubiquitin-proteasome system. *Nat Commun* 2016;**7**: 10239.
56. Kumar M, Michael S, Alvarado-Valverde J, *et al.* The eukaryotic linear motif resource: 2022 release. *Nucleic Acids Res* 2021, gkab975. doi: https://doi.org/10.1093/nar/gkab975.
57. Liu J, Lichtenberg T, Hoadley KA, *et al.* An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 2018;**173**:400–416.e11.
58. Ruan C, Wang C, Gong X, *et al.* An integrative multi-omics approach uncovers the regulatory role of CDK7 and CDK4 in autophagy activation induced by silica nanoparticles. *Autophagy* 2021;**17**:1426–47.
59. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science* 1997;**278**:631–7.
60. Deng W, Wang Y, Liu Z, *et al.* HemI: a toolkit for illustrating heatmaps. *PLoS One* 2014;**9**:e111988.
61. Shannon P, Markiel A, Ozier O, *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.
62. Ren J, Wen L, Gao X, *et al.* DOG 1.0: illustrator of protein domain structures. *Cell Res* 2009;**19**:271–3.
63. Li T, Qin JJ, Yang X, *et al.* The ubiquitin E3 ligase TRAF6 exacerbates ischemic stroke by ubiquitinating and activating Rac1. *J Neurosci* 2017;**37**:12123–40.
64. Oberoi-Khanuja TK, Rajalingam K. IAPs as E3 ligases of Rac1: shaping the move. *Small GTPases* 2012;**3**:131–6.
65. Zhao J, Mialki RK, Wei J, *et al.* SCF E3 ligase F-box protein complex SCF(FBXL19) regulates cell migration by mediating Rac1 ubiquitination and degradation. *FASEB J* 2013;**27**:2611–9.
66. Swatek KN, Komander D. Ubiquitin modifications. *Cell Res* 2016;**26**:399–422.
67. Chen T, Zhou T, He B, *et al.* mUbiSiDa: a comprehensive database for protein ubiquitination sites in mammals. *PLoS One* 2014;**9**:e85744.
68. Xue Y, Ren J, Gao X, *et al.* GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol Cell Proteomics* 2008;**7**: 1598–608.
69. Choi YH, Lim JK, Jeong MW, *et al.* HnRNP A1 phosphorylated by VRK1 stimulates telomerase and its binding to telomeric DNA sequence. *Nucleic Acids Res* 2012;**40**:8499–518.
70. Zhang F, Yu X. WAC, a functional partner of RNF20/40, regulates histone H2B ubiquitination and gene transcription. *Mol Cell* 2011;**41**:384–97.
71. Han TY, Guo M, Gan MX, *et al.* TRIM59 regulates autophagy through modulating both the transcription and the ubiquitination of BECN1. *Autophagy* 2018;**14**:2035–48.
72. Worden EJ, Hoffmann NA, Hicks CW, *et al.* Mechanism of cross-talk between H2B ubiquitination and H3 methylation by Dot1L. *Cell* 2019;**176**:1490–501.
73. Vu LD, Gevaert K, De Smet I. Protein language: post-translational modifications talking to each other. *Trends Plant Sci* 2018;**23**: 1068–80.