Computational and Structural Biotechnology Journal 21 (2023) 1014–1021

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

# Deep-learning based approach to identify substrates of human E3 ubiquitin ligases and deubiquitinases

Yixuan Shu [a,1], Yanru Hai [a,1], Lihua Cao [a], Jianmin Wu [a,b,*]

[a] Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Center for Cancer Bioinformatics, Peking University Cancer Hospital & Institute, Beijing 100142, China
[b] Peking University International Cancer Institute, Peking University, Beijing 100191, China

ARTICLE INFO

*Article history:*
Received 19 August 2022
Received in revised form 16 January 2023
Accepted 16 January 2023
Available online 18 January 2023

*Keywords:*
Ubiquitination
Deep learning
Ubiquitin proteasome system
E3-substrate interactions
DUB-substrate interactions
Pan-cancer analysis

ABSTRACT

E3 ubiquitin ligases (E3s) and deubiquitinating enzymes (DUBs) play key roles in protein degradation. However, a large number of E3 substrate interactions (ESIs) and DUB substrate interactions (DSIs) remain elusive. Here, we present DeepUSI, a deep learning-based framework to identify ESIs and DSIs using the rich information present in protein sequences. Utilizing the collected golden standard dataset, key hyperparameters in the process of model training, including the ones relevant to data sampling and number of epochs, have been systematically assessed. The performance of DeepUSI was thoroughly evaluated by multiple metrics, based on internal and external validation. Application of DeepUSI to cancer-associated E3 and DUB genes identified a list of druggable substrates with functional implications, warranting further investigation. Together, DeepUSI presents a new framework for predicting substrates of E3 ubiquitin ligases and deubiquitinates.

## 1. Introduction

Ubiquitination and deubiquitination are common post-translational modifications in eukaryotic cells, participating in a variety of important biological processes [1], including regulation of protein homeostasis, cell cycle, DNA repair, cell proliferation, and apoptosis. The dysregulation of protein ubiquitination and deubiquitination could lead to various diseases including cancer, nervous system and immune system diseases [2,3]. The ubiquitin-proteasome system (UPS) is a multi-component system, consisting of ubiquitin, ubiquitin-activating enzymes (E1), ubiquitin-conjugating enzymes (E2), ubiquitin ligases (E3), 26 S proteasome and deubiquitinating enzymes (DUB). E1 enzyme, E2 enzyme and E3 ligases mediate the ubiquitination process in a signaling cascade catalytic manner, with substrates covalently bound to ubiquitin and then transferred to the 26 S proteasome complex for degradation. DUB enzymes are cysteine protease proteins that reverse the ubiquitination process by removing ubiquitin from target proteins [4].

Among the components of the ubiquitination system, E3s and DUBs have the most prominent roles in tumorigenesis and cancer development [5]. During ubiquitination, E3 ligases and DUBs specifically recognize target substrates, and determines the fate of substrates. Therefore, recognition of E3s/DUBs substrate interactions (ESIs/DSIs) is critical for characterizing the ubiquitination process. A number of databases, including E3Net [6] and UbiNet [7], have been developed to provide a comprehensive collection of experimentally validated E3s-substrates interactions. However, identification of substrates for E3s and DUBs based on biochemical assay methods (e.g. two-hybrid screening, co-immunoprecipitation and mass spectrometry) is often time-consuming and resource-intensive, and some E3s are difficult to characterize using the standard experimental techniques [8]. Therefore, few bioinformatic tools have been developed to predict ESIs/DSIs based on machine learning models. Wang et al. developed UbiBrowser based on Bayesian model [9] to predict ESIs/DSIs, considering enriched domains, GO term pairs, protein-protein interactions, and inferred E3 recognition consensus

*Abbreviations:* UPS, ubiquitin-proteasome system; CNN, convolutional neutral network; E1, ubiquitin-activating enzymes; E2, ubiquitin-conjugating enzymes; E3, ubiquitin ligases; DUB, deubiquitinating enzymes; ESI, E3-substrate interaction; DSI, DUB-substrate interaction; GSP, gold standard positive dataset; ROC, receiver operating characteristic; AUROC, area under the ROC curve; PR, precision recall; AUPRC, area under the PR curve; TCGA, The Cancer Genome Atlas

[*] Correspondence to: Center for Cancer Bioinformatics, Peking University Cancer Hospital & Institute, 52 Fu-Cheng Road, Hai-Dian District, Beijing 100142, China.
*E-mail address:* wujm@bjmu.edu.cn (J. Wu).

[1] These authors contributed equally.

https://doi.org/10.1016/j.csbj.2023.01.021

motif. Chen et al. developed ESINet [10] using random forest model to predict ESIs from proteomic data, transcriptomic data, protein-protein interactions, and pathway-based associations. However, protein sequences have not been fully utilized in these tools. It is known that the structures of a protein are encoded by its sequences [11], and the interaction between enzymes and substrates could be dependent on structural pairing [12]. Therefore, it is feasible and necessary to make full use of protein amino acid sequence information to explore the interaction patterns of ubiquitination related enzymes and substrates.

Recently, with the great success of deep learning methods in biomedical field [13,14], especially in sequence-based analysis [15,16], deepDegron [17] and Degpred [18] have been developed to recognize the potential binding sites (termed degrons [19]) in a substrate, by deep learning-based modeling (forward neural network for deepDegron, and BERT for Degred) using protein sequence information. However, due to the limited data available for training, degron cannot be identified for a majority of E3 ligases by current approaches [17].

Therefore, we developed a deep learning-based framework DeepUSI to predict substrates of E3 ubiquitin ligases and deubiquitinases (DeepESI/DeepDSI for predicting ESIs/DSIs respectively) within human proteome, based on the most comprehensive dataset collected to date. The performance of DeepUSI was thoroughly evaluated by multiple metrics, and internal and external validation. Our pan-cancer analysis identified 24 cancer-associated UPS genes, and a list of druggable substrates were found by application of DeepUSI, which warrants further investigation for drug repurposing opportunities.

## 2. Material and methods

### 2.1. Data collections

To construct gold standard positive (GSP) dataset, 2926 human ESIs from UbiBrowser 2.0 [9] and a set of 1790 ESIs from ESINet [10] (integrated in DegPred [18]) were merged. After deduplication, 2854 ESIs were used as training dataset for ESI modelling, while non-redundant DegPred manual ESIs (ESIs collected from literature by DegPred with duplicate entries in the training dataset excluded) were used as the independent test dataset. Regarding GSP dataset for training DSI model, we used 864 experimentally validated human DSIs from UbiBrowser 2.0, while DSIs from DUBase [20] (with duplicate entries in the training dataset excluded) were used as the independent dataset for testing the generalization performance of DeepDSI.

Due to a lack of the negative datasets verified by experiments, we constructed the gold standard negative dataset using protein physical interaction data from the BioGRID [21] (4.4.212, released on July 25th, 2022). A total of 39,749 unique E3 physical interactions (with known E3 substrates excluded), and 7819 DUB physical interactions (with known DUB substrates excluded) were obtained. In addition, we downloaded the curated protein sequences from Uniprot SwissProt [22] (February 2022 release) as the reference amino acid sequences.

### 2.2. Model performance metrics

To comprehensively evaluate the prediction performance of DeepUSI, we used multiple metrics, including Area Under Receiver Operating Characteristics curve (AUROC), Area Under Precision Recall curve (AUPRC), and F1 score, which were calculated as explained below:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall (also known as Sensitivity)} = \frac{TP}{TP + FN}$$

$$F1 = \frac{2{*}\text{Precision}{*}\text{Recall}}{\text{Precision} + \text{Recall}}$$

$$1 - \text{Specificity} = \frac{FP}{TN + FP}$$

Where TP represents the number of ESIs/DSIs identified correctly, TN represents the number of non-ESIs/non-DSIs identified correctly, FN denotes the number of undetected ESIs/DSIs by prediction, FP denotes the number of non-ESIs/non-DSIs mis-identified as ESIs/DSIs by prediction.

ROC curve presents the effectiveness of models on a data set by showing sensitivity and specificity using different thresholds, with the “Sensitivity” on the y-axis and “1-Specificity” on the x-axis [23]. PR curve takes “Precision” as the y-axis and “Recall/Sensitivity” as the x-axis. F1 score is defined as the harmonic average of “Precision” and “Recall”, with the threshold 0.5 of prediction score used to define TP.

### 2.3. Model implementation and hyperparameters

DeepUSI was developed based on CNN framework to recognize substrates of E3 ubiquitin ligases and deubiquitinases, which was implemented using the DeepPurpose library [24]. Hyperparameters are critical for machine learning model training, in which epoch is the most important one [25]. To determine the appropriate number of epochs, repeated experiments were carried out using datasets with different proportions of positive and negative samples. It was shown that both AUROC value (Fig. S1A) and cross-entropy loss (Fig. S1B) could converge on the validation set within 20 epochs. Thus, 20 epochs for model training were applied in the following analyses. Moreover, an 'early stop' strategy was employed, in which the model with the best performance (based on the validation set) within all iterations will be selected to prevent overfitting due to too many iterations.

In addition, we used 0.001 as the learning rate and 128 as the batch size, with Adam method used to optimize the gradient descent, according to the literature experience. Moreover, we used dropout regularization to reduce overfitting with 0.1 as drop rate in the neutral network of decoder. More hyperparameters are listed in Supplementary Table 1.

### 2.4. Proportion allocation and random sampling

The proportion of positive and negative samples is one of the important issues for classification, which is often overlooked. When training data were imbalanced, the developed model could be likely biased towards to the categories with larger number of samples. Our data sets were found to be imbalanced due to much more negative samples for ESIs (Supplementary Table 2a). To systematically explore what proportion of positive and negative training samples could lead to an optimal prediction performance, we tested six different proportions of negative samples with the total ESIs (1:1, 1:2, 1:3, 1:4, 1:5 and all negative samples included). Followingly, each proportion of positive and negative samples were divided with a ratio of 7:1:2 for training, internal validation, and internal testing, respectively. We found that different proportions of positive and negative samples have little effect on AUROC, whereas it showed a significant impact on the PR curve and F1 score, which were higher (indicating a better prediction precision) when the proportion is closer to 1:1 (Fig. S1C).

To evaluate the impact of random sampling of negative samples, 1: 1 random sampling were also conducted for both ESI (Fig. S2A) and DSI (Fig. S2B) negative samples. As expected, random sampling of negative samples has little impact on model performance (Fig. S2).

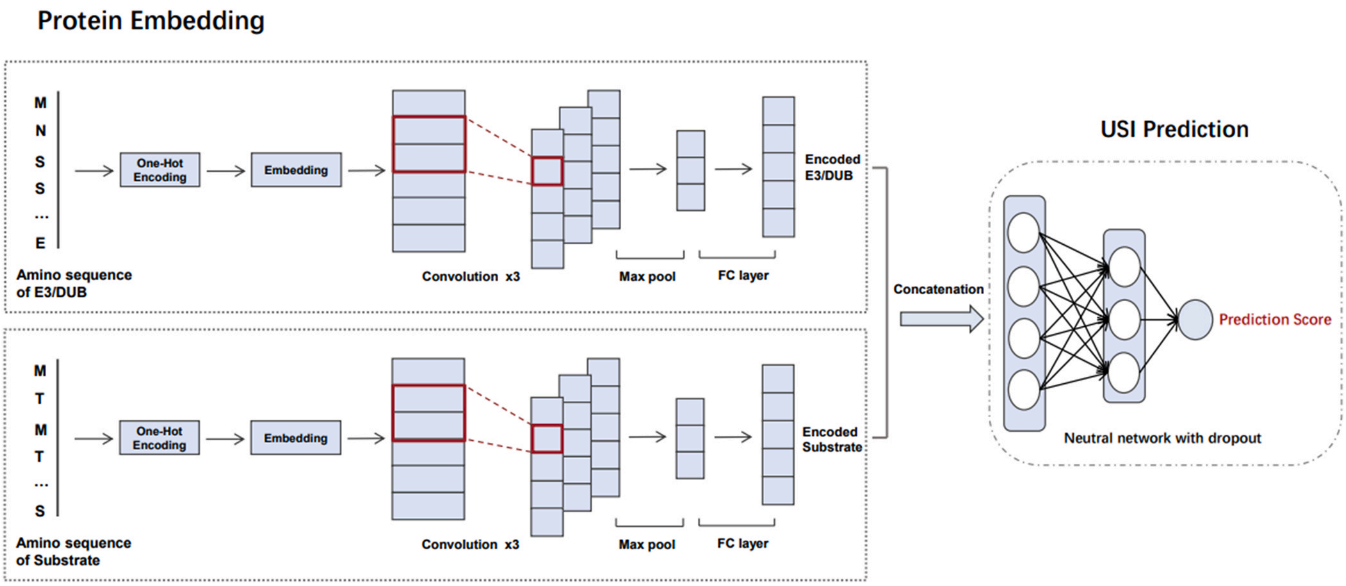


**Fig. 1.** Schematic of the DeepUSI framework. DeepUSI takes amino acid sequence as input feature, which was decomposed and one-hot encoded with an embedding layer for feeding into convolutions, followed by a global max pooling layer and a fully connected layer. Convolution operations were performed three times to extract features. The neutral network with dropout was used to generate the final USI (ESI/DSI) prediction score based on the pair of embeddings of input proteins.

Together, our evaluation results suggested that when a model trained with a balanced proportion of positive and negative samples, it can better fit the characteristics of different classification.

### 2.5. Performance evaluation by cross validation and an independent dataset

To evaluate the predictive performance of our models, cross validation was applied to the models trained with the balanced dataset (positive: negative =1:1) and imbalanced dataset (all negative data included) for both ESIs (Supplementary Table 2b) and DSIs (Supplementary Table 2c) modeling. Ten-fold cross validation was used to obtain the final ESI and DSI prediction model, while five-fold cross validation was also calculated for comparing the reported performance with the published methods [10,26] (Fig. S3A-B). Furthermore, to evaluate the generalization performance, both models, trained using the balanced and the imbalanced dataset by ten-fold cross validation respectively, were tested on an independent dataset for further comparison (Supplementary Table 2d-e). Although both models showed little difference of AUROC in the original dataset, the ESI model trained using the balanced dataset showed a better AUROC in the independent dataset (Fig. S3C-D). Therefore, the model trained using the balanced dataset by ten-fold cross validation was selected as the final prediction model.

### 2.6. Identifying the optimal threshold of prediction score

To identify the optimal threshold to stratify positive and negative interactions, Youden Index was applied. Briefly, it measures the difference between true positive rate (TPR) and false positive rate (FPR) for each threshold, and the threshold was selected as the "best cutoff", when TPR and FPR has the largest difference as shown below:

index = argmax (TPR−FPR)

best_cutoff = thresholds [index]

Where ‘thresholds’ represents the list of different threshold values; TPR and FPR represents the true positive rate and false positive rate using each threshold value; index represents the index value corresponding to the maximum difference between TPR and FPR.

### 2.7. Predicting novel human ESIs and DSIs

Based on the final ESI/DSI models, we made a human proteome-scale predictions, with 11,159,034 possible substrate interactions for 710 E3 ligases and 1,890,405 possible substrate interactions for 118 DUBs tested respectively. Utilizing the aforementioned optimal thresholds, 6,238,978 ESIs and 1,178,694 DSIs predicted to be positive were. The predicted ESIs/DSIs with top 5% scores were defined as high confidence ones.

### 2.8. Pan-cancer analysis to identify cancer-associated UPS genes

By integrating data from the literature [17] and ubiquitination-related databases (iUUCD 2.0 [27], UbiNet 2.0 [7] and UbiBrowser 2.0 [9]), we collected a total of 819 ubiquitin-proteasome system (UPS) genes. Based on pan-cancer RNA-seq data [28] from TCGA, we compared the expression levels of these UPS genes between tumor and paired adjacent normal tissues to identify UPS genes that were significantly up-regulated in tumor tissues. Additionally, based on RNA-seq profiles of 11 normal tissues from the GTEx project [29], we investigated the expression levels of these UPS genes in a large set of normal tissues. Combining the analysis results, we screened the UPS genes that are significantly up-regulated in tumor tissue, and with no or low expression in normal tissues of the same tissue type in the GTEx dataset.

Next, we evaluated the effect of the above UPS genes on tumor cell proliferation by correlating the expression level of each UPS gene with the proliferation marker Ki-67 respectively [30]. In addition, for each TCGA cancer type (n = 15), we evaluated the prognostic value of these UPS genes, by associating gene expression levels with patient survival (including overall survival, disease-free interval, progression-free interval, and disease-specific survival). Finally, a list of 24 candidate UPS genes was summarized by combing the results of the previous parts. These genes met the following conditions simultaneously: i) Significantly up-regulated in tumor tissue and with no or low expression in the corresponding normal tissue; ii) positively correlated with Ki-67 expression; iii) significantly associated with poor patient prognosis.

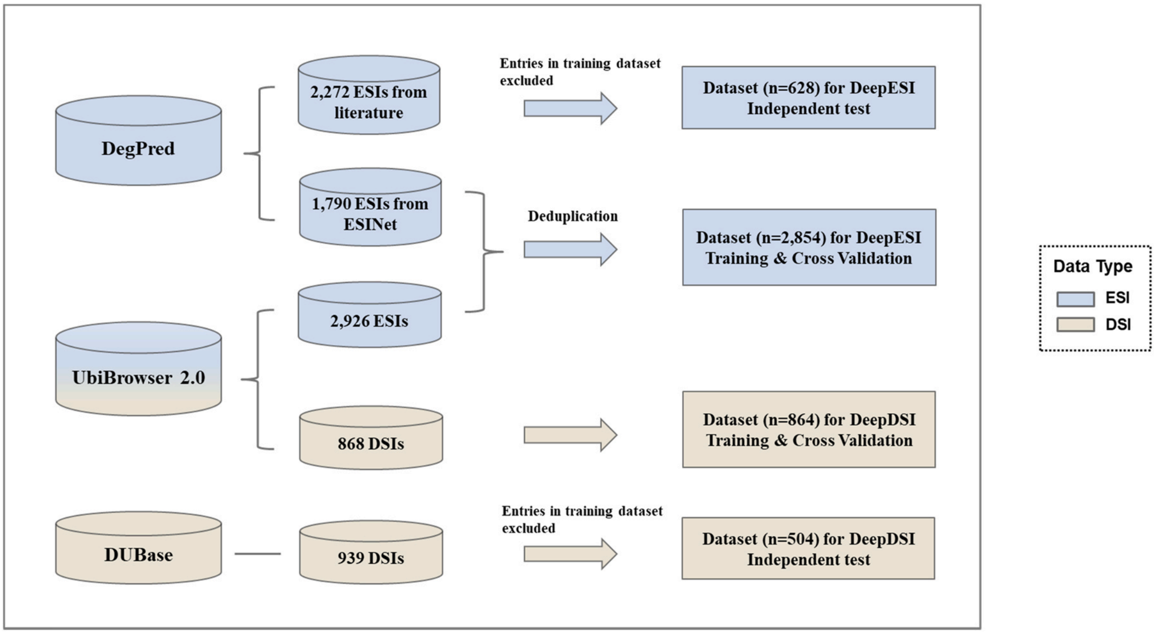
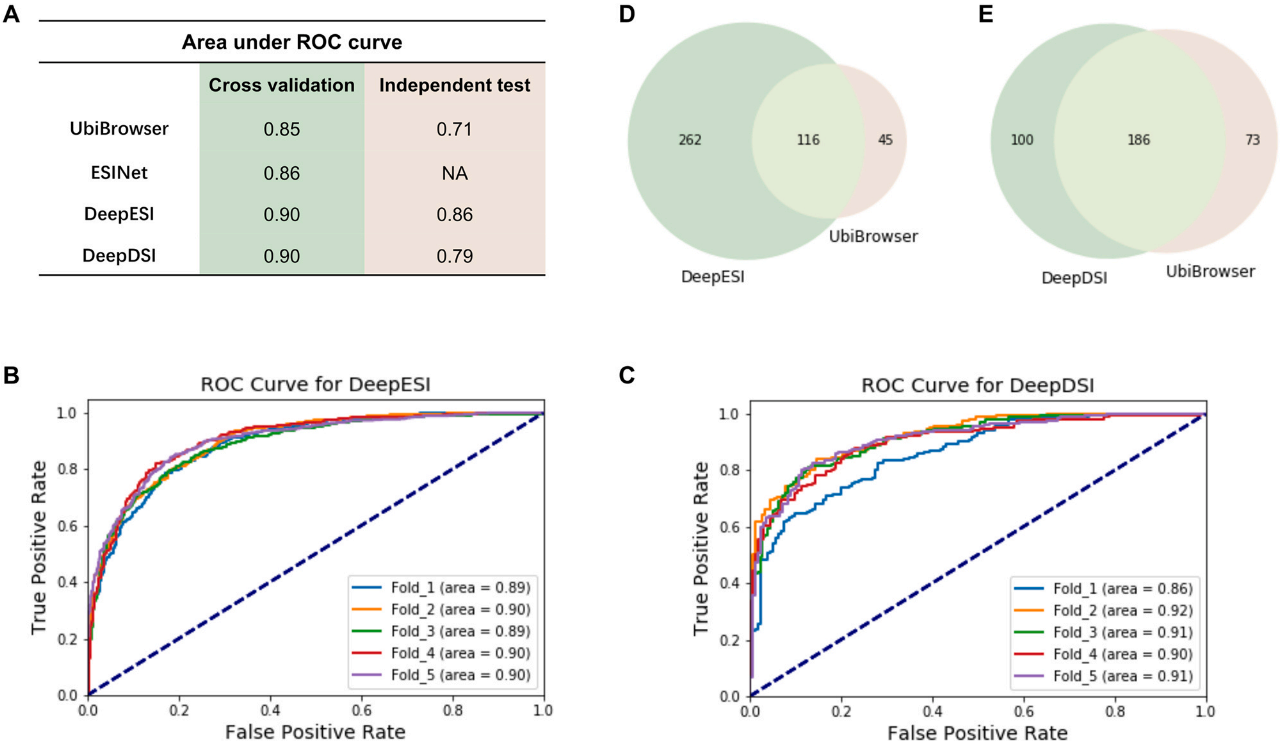

DegPred
2,272 ESIs from literature
1,790 ESIs from ESINet
2,926 ESIs
UbiBrowser 2.0
868 DSIs
DUBase
939 DSIs
Entries in training dataset excluded
Deduplication
Entries in training dataset excluded
Dataset (n=628) for DeepESI Independent test
Dataset (n=2,854) for DeepESI Training & Cross Validation
Dataset (n=864) for DeepDSI Training & Cross Validation
Dataset (n=504) for DeepDSI Independent test
Data Type
ESI
DSI

**Fig. 2.** Flow chart of collection and processing for GSP datasets. Data sources, number of entries, and subsequent use were shown in this figure. Data types were indicated by different colors.



Area under ROC curve

| | Cross validation | Independent test |
|---|---|---|
| UbiBrowser | 0.85 | 0.71 |
| ESINet | 0.86 | NA |
| DeepESI | 0.90 | 0.86 |
| DeepDSI | 0.90 | 0.79 |

**Fig. 3.** Comparison with existing models. A. Performance comparison between UbiBrowser and ESINet based on cross validation using training dataset, and independent test using external dataset (datasets explained in Fig. 2). B-C. ROC curves showing the performance of DeepESI (B) and DeepDSI (C) evaluated by five-fold cross validation (training dataset). D-E. Venn diagrams showing the relationship of ESIs (D) and DSIs (E) identified by DeepUSI and UbiBrowser based on the corresponding independent test dataset.

## 3. Results and discussion

### 3.1. DeepUSI framework

As one of the most known deep learning frameworks, CNN (convolutional neural network) could extract features from biological sequences efficiently and have been successfully applied in DNA and protein sequence-based applications [31,32]. Thus, we developed DeepUSI based on CNN framework to recognize substrates of E3 ubiquitin ligases and deubiquitinases. A pair of protein amino acid sequences were used as input, representing an E3/DUB and a protein of interest respectively. Each amino acid sequence was decomposed to individual character and one-hot encoded with an embedding layer, which were then fed into CNN convolutions. Followingly, the pair of generated latent vectors from input protein sequences were concatenated and fed into a fully-connected neutral

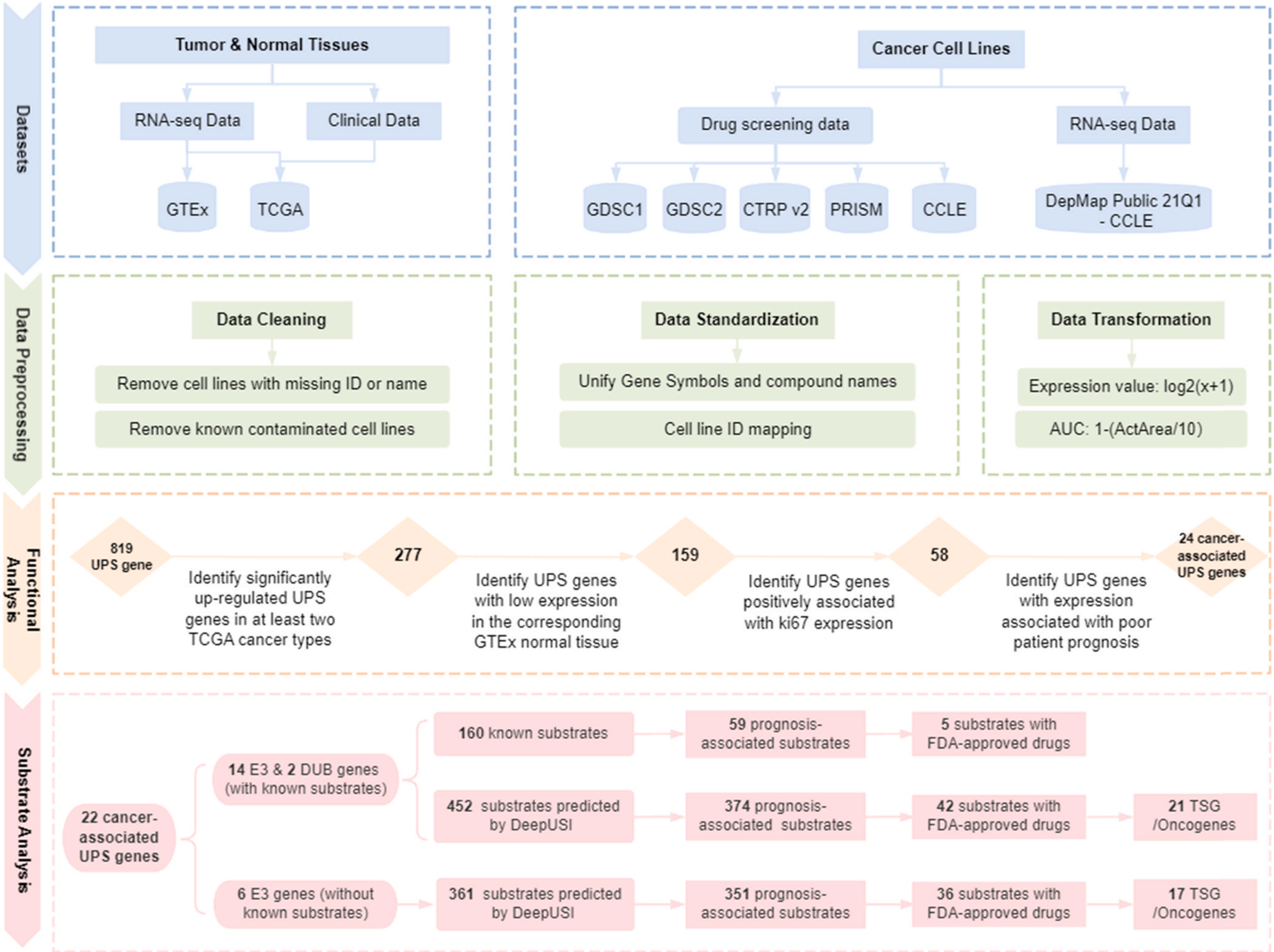

Datasets
Tumor & Normal Tissues
RNA-seq Data
Clinical Data
GTEx
TCGA
Cancer Cell Lines
Drug screening data
RNA-seq Data
GDSC1
GDSC2
CTRP v2
PRISM
CCLE
DepMap Public 21Q1 - CCLE
Data Preprocessing
Data Cleaning
Remove cell lines with missing ID or name
Remove known contaminated cell lines
Data Standardization
Unify Gene Symbols and compound names
Cell line ID mapping
Data Transformation
Expression value: log2(x+1)
AUC: 1-(ActArea/10)
Functional Analysis
819 UPS gene
Identify significantly up-regulated UPS genes in at least two TCGA cancer types
277
Identify UPS genes with low expression in the corresponding GTEx normal tissue
159
Identify UPS genes positively associated with ki67 expression
58
Identify UPS genes with expression associated with poor patient prognosis
24 cancer-associated UPS genes
Substrate Analysis
22 cancer-associated UPS genes
14 E3 & 2 DUB genes (with known substrates)
6 E3 genes (without known substrates)
160 known substrates
59 prognosis-associated substrates
5 substrates with FDA-approved drugs
452 substrates predicted by DeepUSI
374 prognosis-associated substrates
42 substrates with FDA-approved drugs
21 TSG /Oncogenes
361 substrates predicted by DeepUSI
351 prognosis-associated substrates
36 substrates with FDA-approved drugs
17 TSG /Oncogenes

**Fig. 4.** The workflow of pan-cancer analysis of 819 UPS genes. After data collection and preprocessing, pan-cancer functional analysis identified 24 cancer-associated UPS genes by applying multiple filters. Substrate analysis by DeepUSI for 22 E3s and DUBs provide candidates for further investigation, including prognostic association, targets of FDA approved drugs, and cancer-related genes.

network, which output was a prediction score (between 0 and 1) for the likelihood to be an ESI or DSI (Fig. 1).

To help building models with more power, we integrated multiple datasets and constructed the largest ESI and DSI gold standard positive (GSP) datasets by far based on our knowledge (see Material and methods 2.1). There were 2854 ESIs and 864 DSIs collected in total for model training, while 628 additional ESIs and 504 additional DSIs were used as independent datasets to test the generalization performance of DeepESI and DeepDSI respectively (Fig. 2).

### 3.2. Comparison with existing prediction models

In comparison with the published prediction models [9,10], both DeepESI and DeepDSI showed a better AUROC than their reported performance metrics (Fig. 3A), by both cross validation (training dataset, Fig. 3B-C) and external validation using respective independent test dataset (Fig. S3C-D). Furthermore, we utilized the independent test datasets (Fig. 1) for a direct comparison. Based on our proteome-scale predictions (see Material and methods 2.7) and proteome-wide predictions from UbiBrowser (top 20% scores), 116 out of 628 (18.47%) ESIs in the independent test dataset were successfully predicted by both DeepESI and UbiBrowser, and DeepESI strikingly identified 262 (42.52%) additional true ESIs compared with only 45 (7.2%) additional true ESIs identified by UbiBrowser (Fig. 3D). A higher proportion of true DSIs (186 out of 504, 36.9%) were identified by DeepDSI and UbiBrowser in common, while similar portions of specific DSIs (100 vs 73, i.e. 19.84% vs 14.48%) were identified respectively (Fig. 3E). The improved performance of DeepUSI could be due to our design based on CNN models, which mined and utilized global sequence information (determining protein structure and function) in comparison with classical machine learning methods. Additionally, the most comprehensive gold standard datasets we collected by far may help model training.

We continued the comparison using an E3 ligase with biological implications on its potential substrates. TRAIP is an E3 ubiquitin ligase and well-known as a key regulator of interstrand cross-link repair, protecting genome stability in response to replication stress [33]. There were 695 high-confidence TRAIP substrates predicted by DeepUSI and 317 predicted by UbiBrowser, in which 52 predicted substrates were in common. Gene ontology analysis [34] identified 8 out of 52 (15.4%) common substrates were in the cellular response to DNA damage stimulus pathway (GO:0006974) with a statistical enrichment (adjusted $P = 9.25E-08$), including tumor suppressor gene TP53 [35] (Supplementary Table 3a). Moreover, 35 out of 643 DeepUSI-specific substrates (5.4%) were found to be in the same pathway (Supplementary Table 3b) with a statistical enrichment as well (adjusted $P = 4.02E-19$). However, only 5 out of 265 UbiBrowser-specific substrates (1.9%) were identified to be in this pathway (Supplementary Table 3c) with no statistical enrichment (adjusted $P = 0.078$). Therefore, to some extent, it indicates that the substrates

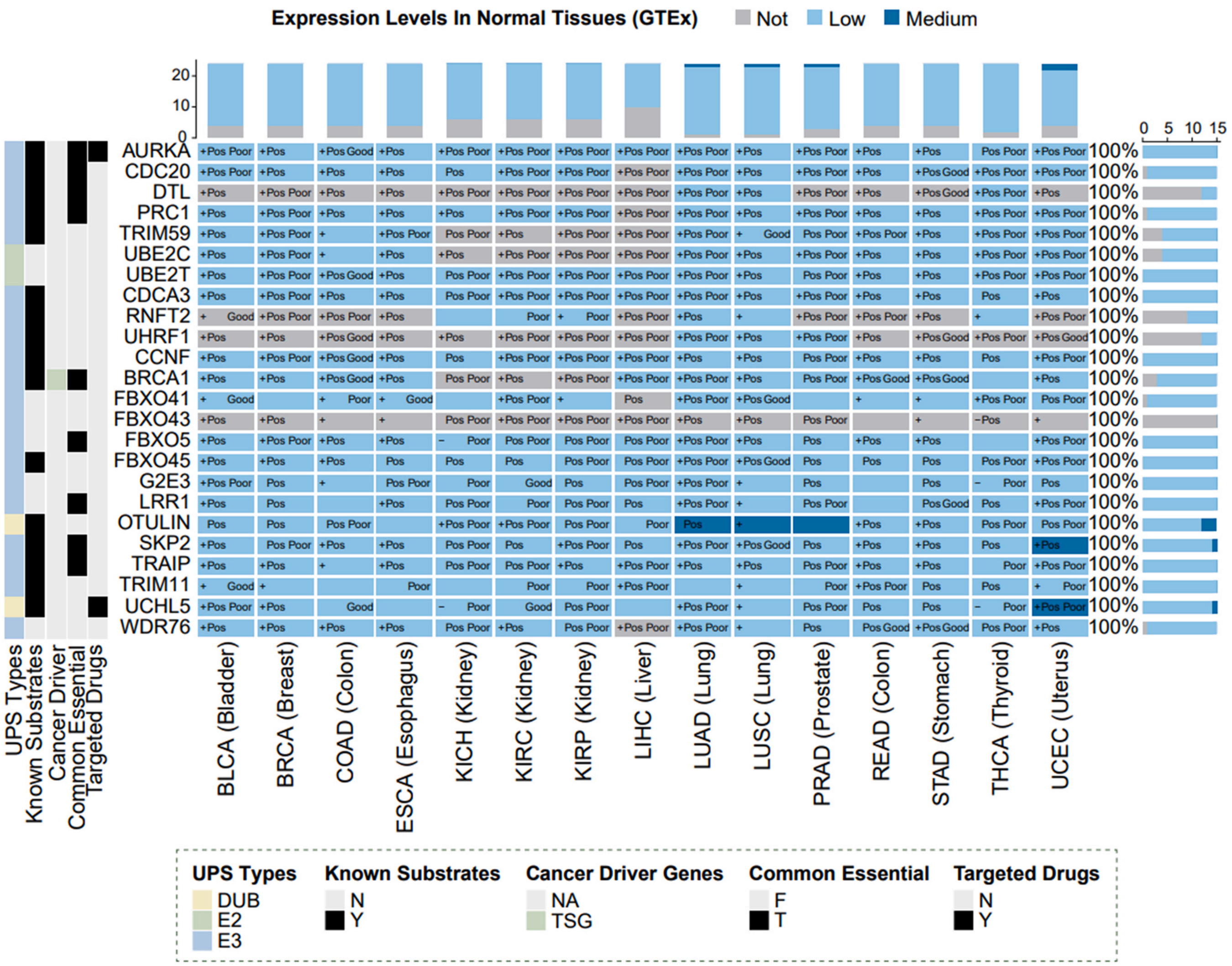


**Fig. 5.** Key UPS genes identified by pan-cancer analysis. The color of each cell indicates the expression level of the UPS gene in GTEx normal tissues. "+" indicates a gene significantly up-regulated in the tumor tissue for a specified cancer type, while "-" indicates significantly down-regulated gene. "Pos" indicates a gene showing significant positive correlation with the expression of Ki-67 for a specified cancer type. "Good" indicates a gene associated with better prognosis for a specified cancer type, while "Poor" indicates association with poor prognosis. The histogram on the right shows the distribution of expression levels of each gene in 11 GTEx normal tissues. The upper histogram shows the distribution of expression levels of 24 UPS genes in each normal tissue. The multi-colored annotation bars in the left panel indicate a gene is: a UPS component, with a known substrate, tumor driver gene (oncogene or tumor suppressor gene), essential gene, with available drugs.

predicted by DeepUSI could be more consistent with its corresponding ubiquitin enzymes regarding biological functions.

### 3.3. Key UPS genes screened by pan-cancer analysis

To identify the UPS genes associated with tumorigenesis and cancer development, we systematically analyzed 819 UPS genes in 15 cancer types using published TCGA cancer datasets and GTEx datasets for matched normal tissue (See Material and methods 2.8 and Fig. 4). As a result, a list of 24 potential cancer-associated UPS genes were identified, with the following criteria simultaneously met in at least two cancer types: i) significantly up-regulated in tumors, with no expression or low expression in matched normal tissues; ii) positively correlated with expression of Ki-67, a cell proliferation marker; 3) high-expression was significantly associated with poor patient prognosis. Among these 24 cancer-associated UPS candidate genes (Fig. 5), 14 E3s and two DUBs have known substrates with experimental evidence, confirming the ability to induce degradation of substrate proteins. However, there are six E3s without known substrates reported, which could therefore benefit from predicted substrates for further experimental validation and investigation.

Next, to reveal substrates with functional implications, we performed pan-cancer prognostic analysis for these known/predicted substrates. Among 160 known substrates of 16 cancer-associated UPS candidate genes (14 E3s and 2 DUBs), 59 substrates were found to be statistically associated with tumor prognosis in the opposite direction to E3 in the same cancer type, while 6 substrates statistically associated with tumor prognosis in the same direction as DUB in the same cancer type, which implied their prognostic significance may be regulated through ubiquitination (Supplementary Table 4a). In total, there are 60 pairs of ESIs and 6 DSIs between these substrates and UPS genes, in which 46 ESIs and 4 DSIs were predicted by DeepUSI. Among these identified substrates with functional implications, five proteins (DNMT1, ESR1, ICAM1, IL3RA, PGR) were found to be targets of FDA approved drugs based on the DrugBank knowledgebase [36]. For example, E3 ligase UHRF1, and its substrate DNMT1, were reported to have a coordinated function in maintaining DNA methylation in cells [37]. DNMT1 was also found to be associated with patient outcome in multiple malignant tumors including triple-negative breast cancer [38], pancreatic cancer [39], and gastric cancer [40]. Therefore, its approved target therapy procainamide, used to treat ventricular arrhythmias, might be a good candidate for drug repurposing in cancer research.

Moreover, there are 452 novel substrates with prognostic association identified for these 16 UPS genes by DeepUSI (Supplementary Table 4b). Within these novel substrates, 42 substrates have FDA-approved drugs with other indications. Moreover, 21 out of these 42 substrates are known tumor suppressors or oncogenes, indicating the potential of drug repurposing in cancer.

Lastly, for the remaining six cancer-associated E3 candidates without known substrates, we applied the same strategy and identified 361 novel substrates (36 have known targeted drugs, 17 known cancer-related genes) with functional implications by DeepUSI and prognostic analysis (Supplementary Table 4c), which provides a valuable catalog of candidate substrates for future investigation.

## 4. Conclusions

Ubiquitination is an important type of post-translational modifications required by many cellular processes [41], and it is essential to recognize the substrates of human E3s and DUBs. Here, we developed DeepESI for ESI prediction and DeepDSI for DSI prediction, by a deep learning-based approach integrating experimentally validated ESIs and DSIs, to make full use of the information contained in global protein sequences. Compared with existing prediction methods, DeepUSI showed a better performance based on amino acid sequence and convolutional neural network, providing new insights for E3 and DUB substrate recognition.

Moreover, the predicted list of ESIs and DSIs provides a rich resource for future investigation of ubiquitination networks. Through a pan-cancer functional analysis of human E3s and DUBs, 24 ubiquitination-related genes were identified with a potential key role in tumorigenesis and cancer development. Further prognostic analysis of their known and DeepUSI-predicted substrates revealed a list of actionable targets with potential drug repurposing opportunities in cancer.

## CRediT authorship contribution statement

**Yixuan Shu**: Conceptualization, Data curation, Formal analysis, Writing – original draft. **Yanru Hai**: Conceptualization, Data curation, Formal analysis. **Lihua Cao**: Methodology. **Jianmin Wu**: Conceptualization, Supervision, Methodology, Writing – review & editing.

## Declaration of Competing Interest

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

## Acknowledgments

This study was supported by Beijing Municipal Science and Technology Commission (Z201100008320006), National Natural Science Foundation of China (82273253), PKU-Baidu Fund (2019BD012), and the third round of Public Welfare Development and Reform Pilot Projects of Beijing Municipal Medical Research Institutes (Beijing Medical Research Institute, 2019-1).

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.01.021.

## References

[1] Narayanan S, Cai C-Y, Assaraf YG, Guo H-Q, Cui Q, Wei L, et al. Targeting the ubiquitin-proteasome pathway to overcome anti-cancer drug resistance. Drug Resist Updates 2020;48:100663https://doi.org/10.1016/j.drup.2019.100663

[2] Ceccarelli DF, Tang X, Pelletier B, Orlicky S, Xie W, Plantevin V, et al. An allosteric inhibitor of the human Cdc34 ubiquitin-conjugating enzyme. Cell 2011;145:1075–87. https://doi.org/10.1016/j.cell.2011.05.039

[3] Sun T, Liu Z, Yang Q. The role of ubiquitination and deubiquitination in cancer metabolism. Mol Cancer 2020;19:146. https://doi.org/10.1186/s12943-020-01262-x

[4] Woo B, Baek K-H. Regulatory interplay between deubiquitinating enzymes and cytokines. Cytokine Growth Factor Rev 2019;48:40–51. https://doi.org/10.1016/j.cytogfr.2019.06.001

[5] Han S, Wang R, Zhang Y, Li X, Gan Y, Gao F, et al. The role of ubiquitination and deubiquitination in tumor invasion and metastasis. Int J Biol Sci 2022;18:2292–303. https://doi.org/10.7150/ijbs.69411

[6] Han Y, Lee H, Park JC, Yi G-S. E3Net: a system for exploring E3-mediated regulatory networks of cellular functions. O111.014076 Mol Cell Proteom 2012;11. https://doi.org/10.1074/mcp.O111.014076

[7] Li Z, Chen S, Jhong J-H, Pang Y, Huang K-Y, Li S, et al. UbiNet 2.0: a verified, classified, annotated and updated database of E3 ubiquitin ligase–substrate interactions. Database 2021;2021. https://doi.org/10.1093/database/baab010

[8] 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. https://doi.org/10.1007/s00018-017-2529-6

[9] Wang X, Li Y, He M, Kong X, Jiang P, Liu X, 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. https://doi.org/10.1093/nar/gkab962

[10] Chen D, Liu X, Xia T, Tekcham DS, Wang W, Chen H, et al. A multidimensional characterization of E3 ubiquitin ligase and substrate interaction network. IScience 2019;16:177–91. https://doi.org/10.1016/j.isci.2019.05.033

[11] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci USA 2021;118:e2016239118https://doi.org/10.1073/pnas.2016239118

[12] Roda S, Santiago G, Guallar V. Mapping Enzyme-substrate Interactions: Its Potential to Study the Mechanism of Enzymes. Advances in Protein Chemistry and Structural Biology vol. 122. Elsevier; 2020. p. 1–31. https://doi.org/10.1016/bs.apcsb.2020.06.001

[13] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems 32. Curran Associates Inc; 2019. p. 8026–37. https://doi.org/10.48550/arXiv.1912.01703

[14] Ahn JC, Connell A, Simonetto DA, Hughes C, Shah VH. Application of artificial intelligence for the diagnosis and treatment of liver diseases. Hepatology 2021;73:2546–63. https://doi.org/10.1002/hep.31603

[15] Kim J, Park S, Min D, Kim W. Comprehensive survey of recent drug discovery using deep learning. IJMS 2021;22:9983. https://doi.org/10.3390/ijms22189983

[16] Chen Z, Zhao P, Li C, Li F, Xiang D, Chen Y-Z, et al. *iLearnPlus:* a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. Nucleic Acids Res 2021;49. https://doi.org/10.1093/nar/gkab122

[17] Tokheim C, Wang X, Timms RT, Zhang B, Mena EL, Wang B, et al. Systematic characterization of mutations altering protein degradation in human cancers. Mol Cell 2021. https://doi.org/10.1016/j.molcel.2021.01.020

[18] Hou C, Li Y, Wang M, Wu H, Li T. Systematic prediction of degrons and E3 ubiquitin ligase binding via deep learning. BMC Biol 2022;20:162. https://doi.org/10.1186/s12915-022-01364-6

[19] Guharoy M, Bhowmick P, Sallam M, Tompa P. Tripartite degrons confer diversity and specificity on regulated protein degradation in the ubiquitin-proteasome system. Nat Commun 2016;7:10239. https://doi.org/10.1038/ncomms10239

[20] Elu N, Osinalde N, Ramirez J, Presa N, Rodriguez JA, Prieto G, Mayor U. Identification of substrates for human deubiquitinating enzymes (DUBs): An up-to-date review and a case study for neurodevelopmental disorders. Semin Cell Dev Biol 2022;132:120–31. https://doi.org/10.1016/j.semcdb.2022.01.001

[21] Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res 2006;34:D535–9. https://doi.org/10.1093/nar/gkj109

[22] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 2004;32:115D–9D. https://doi.org/10.1093/nar/gkh131

[23] Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000;16:412–24. https://doi.org/10.1093/bioinformatics/16.5.412

[24] Huang K, Fu T, Glass LM, Zitnik M, Xiao C, Sun J. DeepPurpose: a deep learning library for drug–target interaction prediction. Bioinformatics 2021;36:5545–7. https://doi.org/10.1093/bioinformatics/btaa1005

[25] Pérez-Enciso M, Zingaretti LM. A guide for using deep learning for complex trait genomic prediction. Genes 2019;10:553. https://doi.org/10.3390/genes10070553

[26] Li Y, Xie P, Lu L, Wang J, Diao L, Liu Z, et al. An integrated bioinformatics platform for investigating the human E3 ubiquitin ligase-substrate interaction network. Nat Commun 2017;8:347. https://doi.org/10.1038/s41467-017-00299-9

[27] Zhou J, Xu Y, Lin S, Guo Y, Deng W, Zhang Y, et al. iUUCD 2.0: an update with rich annotations for ubiquitin and ubiquitin-like conjugations. Nucleic Acids Res 2018;46:D447–53. https://doi.org/10.1093/nar/gkx1041

[28] Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. Nature 2020;578:82–93. https://doi.org/10.1038/s41586-020-1969-6

[29] Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The genotype-tissue expression (GTEx) project. Nat Genet 2013;45:580–5. https://doi.org/10.1038/ng.2653

[30] Scholzen T, Gerdes J. The Ki-67 protein: from the known and the unknown. J Cell Phys 2000;182:311–22. https://doi.org/10.1002/(SICI)1097-4652(200003)182:3<311::AID-JCP1>3.0.CO;2-9

[31] Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018;15:20170387https://doi.org/10.1098/rsif.2017.0387

[32] Song B, Li Z, Lin X, Wang J, Wang T, Fu X. Pretraining model for biological sequence data. Brief Funct Genom 2021;20:181–95. https://doi.org/10.1093/bfgp/elab025

[33] Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Curr Protoc Bioinformatics 2016;54:1.30.1–1.30.33. https://doi.org/10.1002/cpbi.5

[34] Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res 2011;39. https://doi.org/10.1093/nar/gkr483

[35] Liu J, Guan D, Dong M, Yang J, Wei H, Liang Q, et al. UFMylation maintains tumour suppressor p53 stability by antagonizing its ubiquitination. Nat Cell Biol 2020;22(9):1056–63. https://doi.org/10.1038/s41556-020-0559-z

[36] Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P. et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 2006;34:D668–72. https://doi.org/10.1093/nar/gkj067

[37] Unoki M, Brunet J, Mousli M. Drug discovery targeting epigenetic codes: the great potential of UHRF1, which links DNA methylation and histone modifications, as a drug target in cancers and toxoplasmosis. Biochem Pharmacol 2009;78:1279–88. https://doi.org/10.1016/j.bcp.2009.05.035

[38] Wong KK. DNMT1: A key drug target in triple-negative breast cancer. Semin Cancer Biol 2021;72:198–213. https://doi.org/10.1016/j.semcancer.2020.05.010

[39] Wong KK. DNMT1 as a therapeutic target in pancreatic cancer: mechanisms and clinical implications. Cell Oncol 2020;43:779–92. https://doi.org/10.1007/s13402-020-00526-4

[40] Li H, Li W, Liu S, Zong S, Wang W, Ren J, et al. DNMT1, DNMT3A and DNMT3B polymorphisms associated with gastric cancer risk: a systematic review and meta-analysis. EBioMedicine 2016;13:125–31. https://doi.org/10.1016/j.ebiom.2016.10.028

[41] Morreale FE, Walden H. Types of ubiquitin ligases. Cell 2016;165:248-248.e1https://doi.org/10.1016/j.cell.2016.03.003