Supplementary Materials

TransPTM: a Transformer-Based Model for Non-Histone Acetylation Site Prediction

Lingkuan Meng, [1] Xingjian Chen,[3] Ke Cheng,[2] Nanjun Chen,[1] Zetian Zheng,[1] Fuzhou Wang,[1] Hongyan Sun, [2,*] and Ka-Chun Wong [1,4,*]

[1] Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong,
[2] Department of Chemistry, City University of Hong Kong, Kowloon, Hong Kong,
[3] Cutaneous Biology Research Center, Massachusetts General Hospital, Harvard Medical School, MA 02138, United States,
[4] Shenzhen Research Institute, City University of Hong Kong, Shenzhen, China
[*]Corresponding authors. kc.w@cityu.edu.hk, hongysun@cityu.edu.hk

Contents

1.Histone and non-histone protein acetylation sequence data comparison....................................... 1

2. Hyper-parameters of comparative classifiers for TransPTM 4

3. 5-fold cross-validation on entire dataset 8

4. Hyper-parameters selection for ProtT5 language model 8

1.Histone and non-histone protein acetylation sequence data comparison

There are critical biological differences between histone and non-histone protein acetylation. Although these two acetylation processes both occur on lysine residues, it is advisable to study them separately due to the different biological functions they affect and varying sequence contexts of the proteins.

Histone acetylation is an essential part of gene regulation. It involves the covalent addition of acetyl groups to the conserved lysine residue of histone N-terminal tails. These reactions are typically catalysed by serval types of histone acetyltransferases (HATs). To acetylate lysine residues on protein sequences, HATs need to first recognize the features around the target lysine residues. Histone acetylating abilities of HATs are categorized based on substrate sequence similarities, which are typically high among family members [1]. Some of the major families identified include: GCN5, p300, MYST and others. For example, H3K9, refers to the 9th lysine residue which is highly conserved on the tail of the histone H3 protein [2], can be recognized and acetylated by HATs such as GCN5 and PCAF [3].

However, the diversity of non-histone proteins is significantly greater than that of histones. HAT families mainly exhibit specificity towards histone proteins, which may lead to issues with generalization and limits their ability to recognize all non-histone acetylation sites. For example, it is identified that KAT13D (CLOCK), which is not a HAT, can directly acetylates argininosuccinate synthase (ASS1) at K165 and K176 residues [4].

To visualize the differences between histone acetylation and non-histone acetylation protein sequences, we downloaded all acetylated histone sequences from the UniProt database. We then extracted sequences adjacent to the acetylation sites based on the 25 amino acid length window size and compared them with our positive dataset of non-histone acetylation sequences. We analyzed these datasets using three methods: amino acid residue proportion analysis, Two Sample Logo analysis, and multiple sequence alignment (MSA) analysis.

We calculated the occurrence composition for amino acid residues in the histone and non-histone protein acetylation sequences (Figure S1A). Also, a Two Sample Logo was utilized to analyze the occurrence of amino acids around histone and non-histone protein acetylation sites (Figure S1B). From Figure S1A, we can observe that there is notable difference in residue composition between histone and non-histone protein acetylation sequences. The histone acetylation sequences contain significantly higher amounts of alanine (A), glycine (G), and lysine (K) than those found in non-histone protein acetylation sequences. Figure S1B further illustrates that the compositional and positional information of acetylated histone sequences and acetylated non-histone protein sequences have statistically significant differences. We can see that, compared to the non-histone side (lower panel), the histone side (upper panel) displays less variety of amino acid residues, mainly consisting of alanine (A), glycine (G), and lysine (K). This trend aligns with the results shown in Figure S1A. On the non-histone side (lower panel), however, a more diverse array of amino acid residues is presented, reflecting the lack of conservation of the non-histone acetylation sequences. We also applied multiple sequence alignment (MSA) to analysis the conservation features of histone and non-histone protein acetylation sequences using Jalview [5] software. The results are showed in the Figure S1C. We can obverse that the overall conservation score of histone acetylation sequences (score: 28) is more than 3 times higher than non-histone protein sequences (score: 9).

In summary, the biological functions of histones and non-histone proteins are different, and the contexts of the acetylated lysine (K) sites on their sequences are also distinct. Compared to the histone acetylation sequences, the amino acid composition near the non-histone proteins acetylation sites is more diverse, more irregular, and less conserved, necessitating a classifier with enhanced capabilities, particularly in terms of generalization and interpretability. This is why we have designed and implemented a classifier specifically tailored to non-histone proteins.

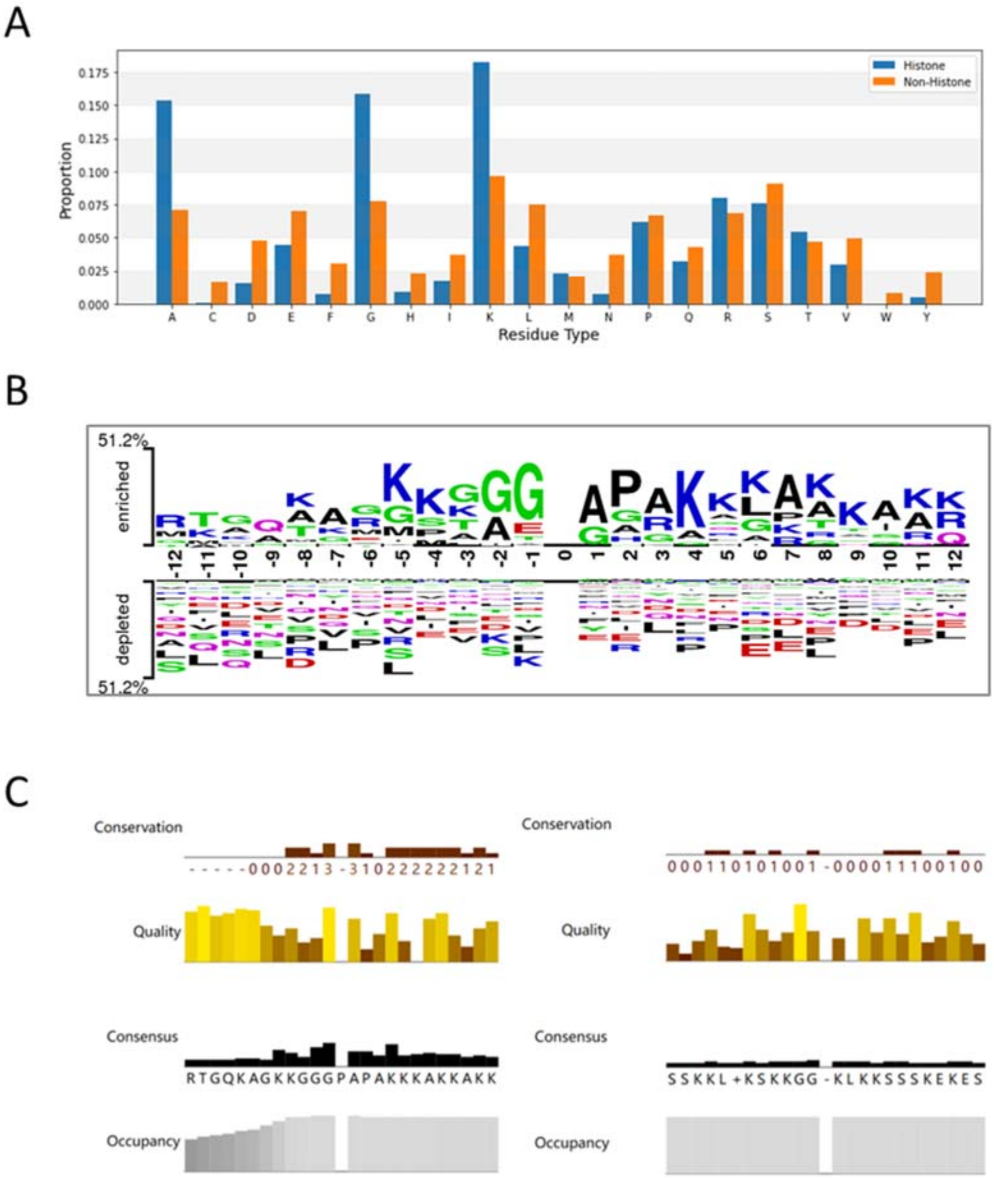


Fig. S1. Comparison of histone and non-histone protein acetylation sequences. (A) The percentage of 20 standard amino acid residues in the histone and non-histone protein acetylation sequences. (B) Two Sample Logo ($p < 0.05$) of the compositional bias in the histone and non-histone protein acetylation sequences. (C) Multiple Sequence Alignment (MSA) analysis of histone and non-histone protein acetylation sequences. Conservation is calculated according to Livingstone and Barton [6] and reflects the number of physicochemical properties shared by all amino acids in a column. Quality score reflects the total likelihood of observing mutations between amino acids aligned at the given column, based on the BLOSUM62 [7] substitution matrix. Consensus annotation row shows the modal residue in each column (or + if more than one residue is observed) and the proportion of sequences that contain that residue. Occupancy measures the number of sequences aligned at each position.

2. Hyper-parameters of comparative classifiers for TransPTM

In our study, we utilized seven distinct classifiers: Random Forest (RF), Support Vector Machine (SVM), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and three existing acetylation site prediction tools. Each of these models was configured with specific hyper-parameters as follows:

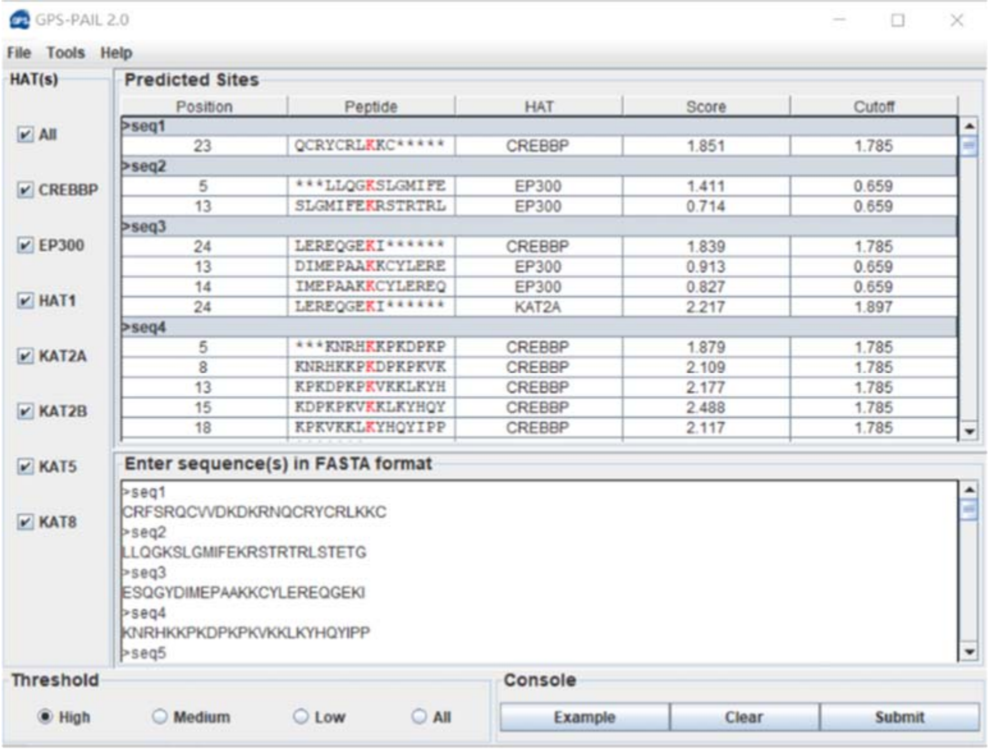
Random Forest (RF): We utilized 23 trees with a max_features of 100.

Support Vector Machine (SVM): The kernel was configured as ‘linear', with C (regularization parameter) set to 30 and tol (tolerance for the stopping criterion) set to 0.001.

Convolutional Neural Network (CNN): The CNN model is composed of two convolutional layers, the first containing 32 filters and the second 64 filters, each followed by a ReLU activation function. We employed the Adam optimizer with a learning rate of 0.001, and the model underwent training for 50 epochs.

Long Short-Term Memory (LSTM): The LSTM model has 3 layers, each with 128 units. We employed the Adam optimizer with a learning rate of 0.001, and the model underwent training for 50 epochs.

GPS-PAIL: In the GPS-PAIL 2.0 software, we input our non-histone acetylation independent testing set and selected the "ALL" and "High" settings under "HAT(s)" and "Threshold" tabs, respectively. This indicates that we apply GPS-PAIL to predict acetylation sites to all the seven distinct HATs, while generate predicted scores for lysine residues with a stringent threshold with Sp value of 95% [8]. (https://pail.biocuckoo.org/online.php)



HAT(s)

☑ All
☑ CREBBP
☑ EP300
☑ HAT1
☑ KAT2A
☑ KAT2B
☑ KAT5
☑ KAT8

Predicted Sites

Position	Peptide	HAT	Score	Cutoff
>seq1				
23	QCRYCRLKKC*****	CREBBP	1.851	1.785
>seq2				
5	***LLQGKSLGMIFE	EP300	1.411	0.659
13	SLGMIFEKRSTRTRL	EP300	0.714	0.659
>seq3				
24	LEREQGEKI******	CREBBP	1.839	1.785
13	DIMEPAAKKCYLERE	EP300	0.913	0.659
14	IMEPAAKKCYLEREQ	EP300	0.827	0.659
24	LEREQGEKI******	KAT2A	2.217	1.897
>seq4				
5	***KNRHKKPKDPKP	CREBBP	1.879	1.785
8	KNRHKKPKDPKPKVK	CREBBP	2.109	1.785
13	KPKDPKPKVKKLKYH	CREBBP	2.177	1.785
15	KDPKPKVKKLKYHQY	CREBBP	2.488	1.785
18	KPKVKKLKYHQYIPP	CREBBP	2.117	1.785

Enter sequence(s) in FASTA format

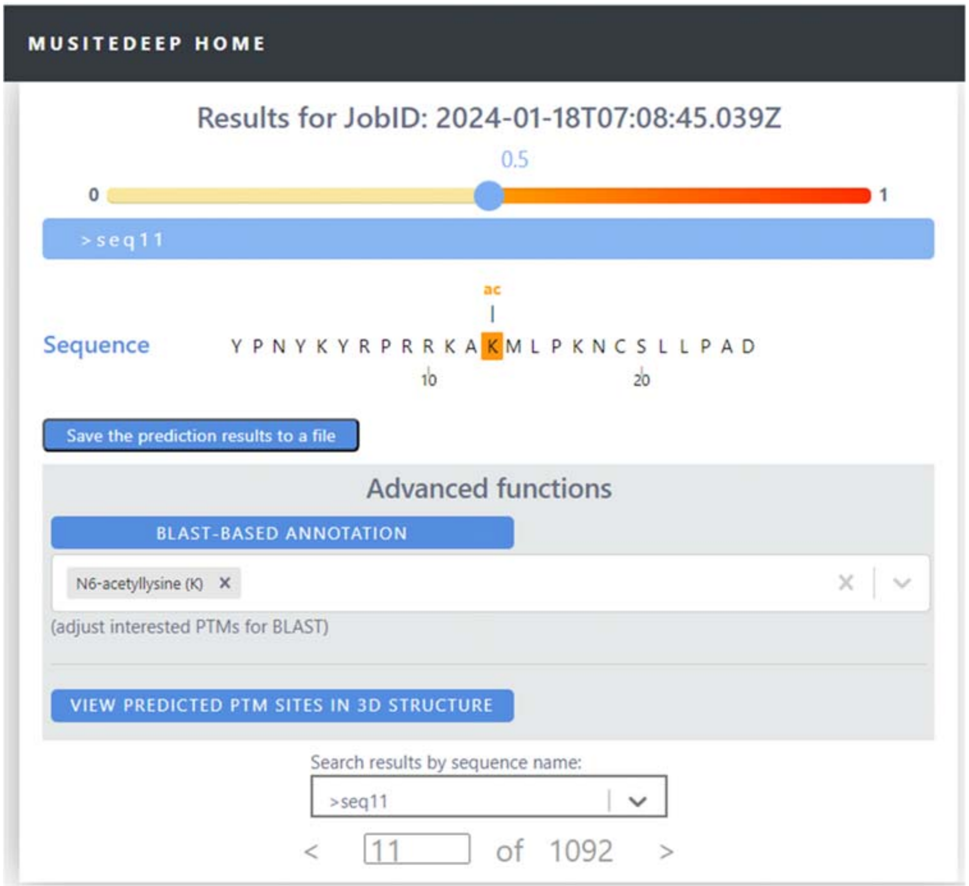
```
>seq1
CRFSRQCVVDKDKRNQCRYCRLKKC
>seq2
LLQGKSLGMIFEKRSTRTRLSTETG
>seq3
ESQGYDIMEPAAKKCYLEREQGEKI
>seq4
KNRHKKPKDPKPKVKKLKYHQYIPP
>seq5
```

Threshold

● High ○ Medium ○ Low ○ All

Console

MusiteDeep: In the MusiteDeep webserver, we input our non-histone acetylation independent testing set and selected the "0.5" and "N6-acetyllysine(K)" settings under "Threshold" bar and "PTM types" tab, respectively. (https://www.musite.net/)



MUSITEDEEP HOME

Results for JobID: 2024-01-18T07:08:45.039Z

0.5

0 1

>seq11

Sequence Y P N Y K Y R P R R K A K M L P K N C S L L P A D

N6-acetyllysine (K)

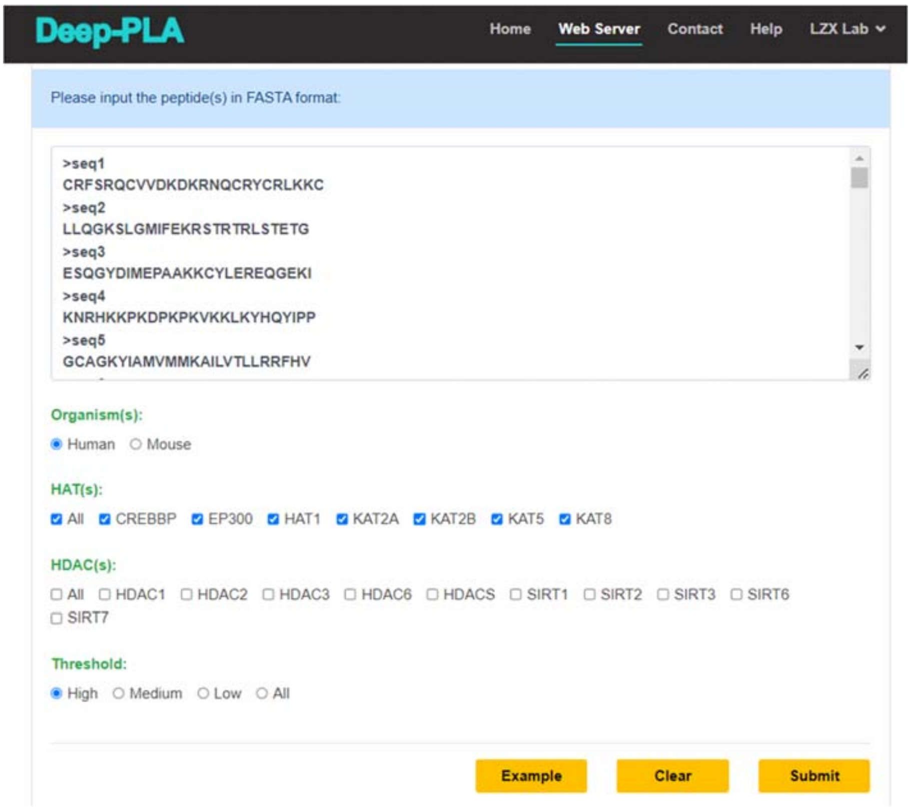
(adjust interested PTMs for BLAST)

Search results by sequence name: >seq11

< 11 of 1092 >

ac
10
20
Save the prediction results to a file
Advanced functions
BLAST-BASED ANNOTATION
VIEW PREDICTED PTM SITES IN 3D STRUCTURE

Deep-PLA: In the Deep-PLA webserver, we input our non-histone acetylation independent testing set and selected the "ALL" and "High" settings under "HAT(s)" and "Threshold" tabs, respectively. (http://deeppla.omicsbio.info/webserver.php)



Please input the peptide(s) in FASTA format:

```
>seq1
CRFSRQCVVDKDKRNQCRYCRLKKC
>seq2
LLQGKSLGMIFEKRSTRTRLSTETG
>seq3
ESQGYDIMEPAAKKCYLEREQGEKI
>seq4
KNRHKKPKDPKPKVKKLKYHQYIPP
>seq5
GCAGKYIAMVMMKAILVTLLRRFHV
```

Organism(s):

● Human ○ Mouse

HAT(s):

☑ All ☑ CREBBP ☑ EP300 ☑ HAT1 ☑ KAT2A ☑ KAT2B ☑ KAT5 ☑ KAT8

HDAC(s):

□ All □ HDAC1 □ HDAC2 □ HDAC3 □ HDAC6 □ HDACS □ SIRT1 □ SIRT2 □ SIRT3 □ SIRT6
□ SIRT7

Threshold:

● High ○ Medium ○ Low ○ All

The hyper-parameters for RF and SVM were selected through a combination of randomized search and 5-fold cross validation to optimize these parameters, ensuring a balance between model complexity and the risk of overfitting. Meanwhile, the hyper-parameters for the CNN and LSTM were chosen through manual selection, allowing us to leverage our domain knowledge and experience to enhance efficiency and computational resource utilization. Comparison with different methods should base on same learning dataset. The results will be unfairness if we use different training data. However, we couldn’t access the source codes of other existing tools.

To evaluate the model, we employed an alternative approach: testing it on the independent set that was not included in the training dataset. Our criterion for selecting these hyperparameter combinations for each model was to choose the one that achieves the highest AUC. This approach aligns with our methodology for training the TransPTM model, where we also prioritized maximizing the AUC.

Classifiers	Hyper-parameters	Run Time (s)
RF	'max_features': 100 'n_estimators': 23	46.51
SVM	'c': 30 'tol': 0.001	499.86
CNN	Activation Function: Relu Loss Function: Cross Entropy Training Epochs: 50 Optimizer: Adam	33.78
LSTM	Activation Function: Relu Loss Function: Binary Cross Entropy Training Epochs: 50 Optimizer: Adam	25.01
GPS-PAIL (software)	‘Threshold’: High ‘HATs’: ALL	-
MusiteDeep (webserver)	‘Threshold’: 0.5 ‘PTM Types’: N6-acetyllysine(K)	-
Deep-PLA (webserver)	‘Threshold’: High ‘HATs’: ALL	-
ProtT5 (fine-tuned)	Activation Function: Relu Loss Function: Binary Cross Entropy Training Epochs: 50 Optimizer: Adam	336.41
TransPTM	Activation Function: Relu Loss Function: Binary Cross Entropy Training Epochs: 100 Optimizer: Adam	261.00

Table. S1. Hyper-parameters of comparative classifiers for TransPTM. Note: -, data not available; run time are excluded as evaluations were conducted via software or webservers.

3. 5-fold cross-validation on entire dataset

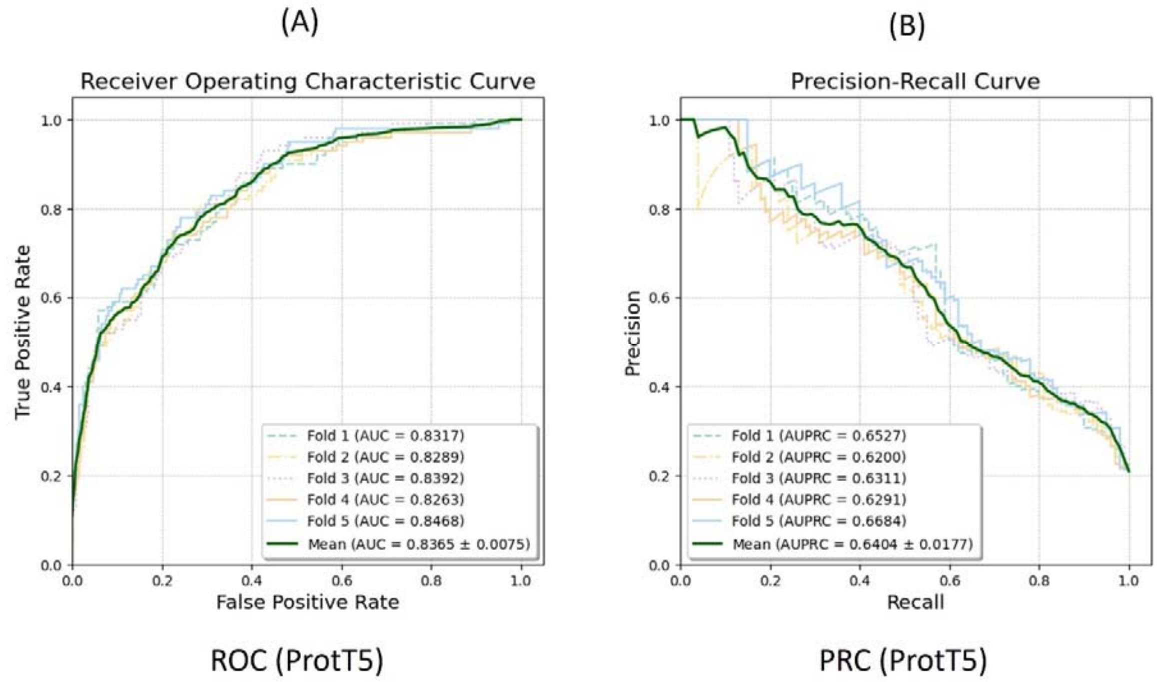


Fig. S2. Performance of TransPTM (ProtT5) on the entire dataset using 5-fold cross-validation. (A) ROC curves, (B) Precision-recall curves.

4. Hyper-parameters selection for ProtT5 language model

To optimize hyperparameters, we designed two types of classifiers for the ProtT5 model: CNN and MLP. For each classifier, we experimented with three different learning rates: 1e-4, 1e-5, and 1e-6. In line with the criteria used for other baseline models, which prioritize maximizing AUC, the best-performing setup was determined to be the CNN classifier with the learning rate of 1e -5. This hyper -parameter combination achieved an accuracy of 0.74, an F1 score of 0.41, an MCC of 0.31, an AUC of 0.71, and an AUPRC of 0.24. These results have been listed into Table S2.

Classifier	Learning rate	Accuracy	F1 score	MCC	AUC	AUPRC
CNN	1e-4	0.7473	0.3867	0.2729	0.6769	0.2259
	1e-5	0.7436	**0.4118**	0.3094	**0.7056**	0.2440
	1e-6	0.8013	0.2439	0.1300	0.5625	0.1649
MLP	1e-4	0.7738	0.3778	0.2612	0.6587	0.2205
	1e-5	0.8132	0.3929	0.2863	0.6563	0.2331
	1e-6	**0.8562**	0.3938	**0.3197**	0.6392	**0.2497**

Table. S2. Hyper-parameters selection for ProtT5 language model. CNN: activation function: Relu, loss function: binary cross entropy, training epochs: 50, optimizer: adam. MLP: activation function: Relu, loss function: binary cross entropy, training epochs: 50, optimizer: adam.

References

[1] R. Marmorstein, Structure of histone acetyltransferases11Edited by P. W. Wright, Journal of Molecular Biology 311(3) (2001) 433-444.

[2] S.K. Kurdistani, S. Tavazoie, M. Grunstein, Mapping Global Histone Acetylation Patterns to Gene Expression, Cell 117(6) (2004) 721-733.

[3] Q. Jin, L.R. Yu, L. Wang, Z. Zhang, L.H. Kasper, J.E. Lee, C. Wang, P.K. Brindle, S.Y.R. Dent, K. Ge, Distinct roles of GCN5/PCAF-mediated H3K9ac and CBP/p300-mediated H3K18/27ac in nuclear receptor transactivation, The EMBO Journal 30(2) (2011) 249-262-262.

[4] R. Lin, Y. Mo, H. Zha, Z. Qu, P. Xie, Z.-J. Zhu, Y. Xu, Y. Xiong, K.-L. Guan, CLOCK Acetylates ASS1 to Drive Circadian Rhythm of Ureagenesis, Molecular Cell 68(1) (2017) 198-209.e6.

[5] A.M. Waterhouse, J.B. Procter, D.M.A. Martin, M. Clamp, G.J. Barton, Jalview Version 2—a multiple sequence alignment editor and analysis workbench, Bioinformatics 25(9) (2009) 1189-1191.

[6] C.D. Livingstone, G.J. Barton, Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation, Bioinformatics 9(6) (1993) 745-756.

[7] S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, Proceedings of the National Academy of Sciences 89(22) (1992) 10915-10919.

[8] W. Deng, C. Wang, Y. Zhang, Y. Xu, S. Zhang, Z. Liu, Y. Xue, GPS-PAIL: prediction of lysine acetyltransferase-specific modification sites from protein sequences, Scientific Reports 6(1) (2016) 39787.