# A Multimodal Deep Architecture for Large-Scale Protein Ubiquitylation Site Prediction

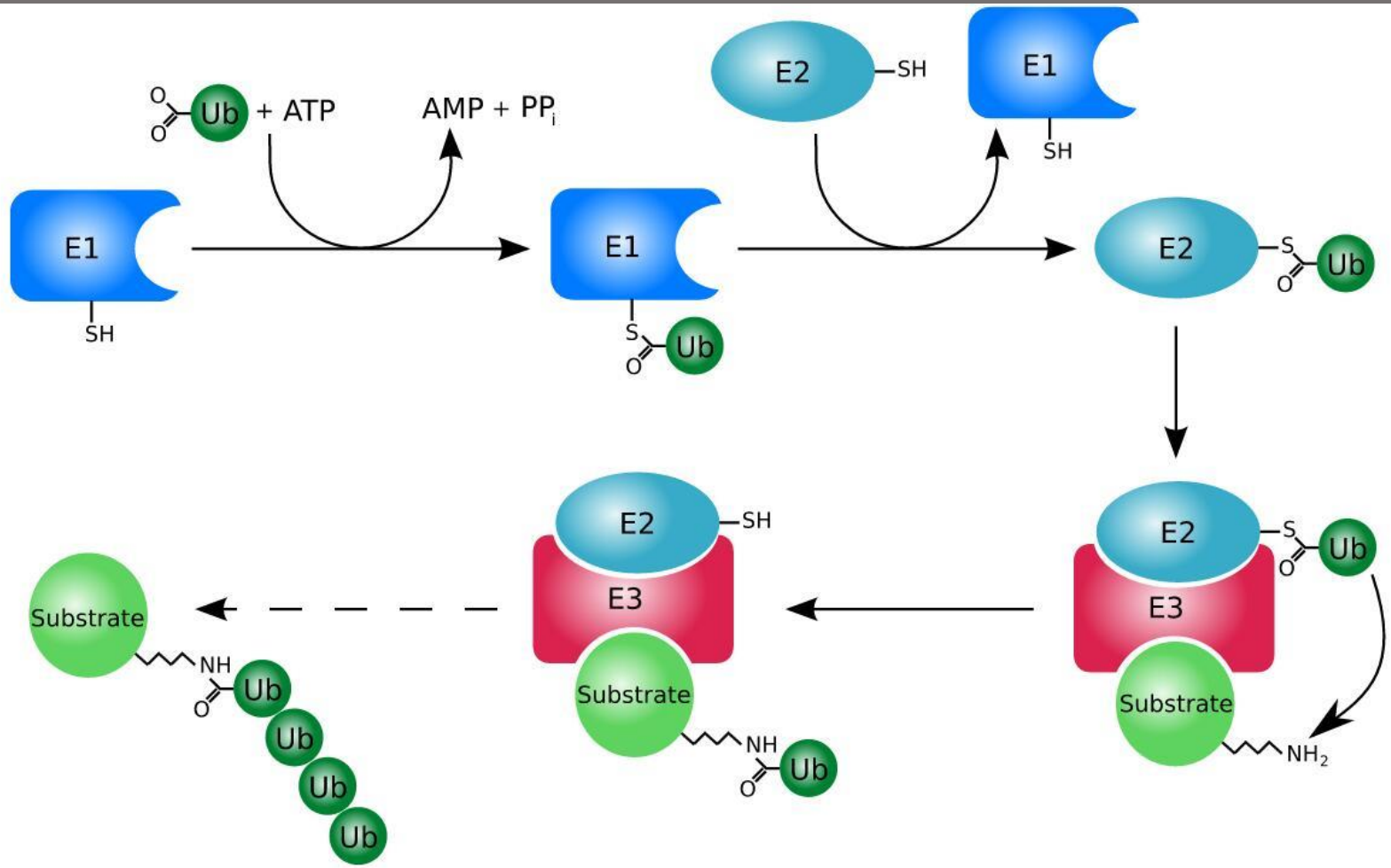
Fei He[1], Lingling Bao[1], Rui Wang[1], Jiagen Li[1], Dong Xu[2], Xiaowei Zhao[1]

[1]School of Information Science and Technology
Institute of Computational Biology
Northeast Normal University Changchun,
China

[2]Department of Electrical Engineering and Computer Science, Informatics Institute, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA

# Content

# 1 Introduction




**The conjugation of ubiquitin to a substrate protein on a particular lysine**

**Involves three types of enzymes (activating, ligases and conjugatin enzymes)**

**Related to various cellular functions**

# 1 Introduction

## Experimental techniques

- Laborious
- Expensive
- Time consuming

- Mass spectrometry
- CHIP-CHIP analysis

## Computational approaches

- ESA-Ubisite
- UbiProber
- iUbiq-Lys
- Ubisite

- ✓ Convenient
- ✓ Efficient
- ✓ Guidance

# 1 Introduction

Testing results from original literatures

Testing results on large scale dataset PLMD

| Tool | Metrics | | | | Testing data scale (sites) |
|---|---|---|---|---|---|
| | Accuracy | Sensitivity | Specificity | MCC | |
| ESA-Ubisite[1] | 92% | 66% | 94% | **0.48** | 2197 |
| | 61.26% | **46.14%** | **63.34%** | **0.064** | 52373 |
| UbiProber[2] | N/A | N/A | N/A | **0.73** | 298 |
| | 55.06% | **62.40%** | **54.05%** | **0.107** | 52373 |
| iUbiq-Lys[3] | 90.06% | 80.99% | 99.06% | **0.82** | 7855 |
| | 84.63% | **3.35%** | **96.88%** | **0.005** | 52373 |
| Ubisite[4] | 73.69% | 85.10% | 69.69% | **0.483** | 14396 |
| | 73.63% | **29.62%** | **79.64%** | **0.073** | 52373 |


1. J. R. Wang, W. L. Huang, M. J. Tsai, K. T. Hsu, H. L. Huang, and S. Y. Ho, "ESA-UbiSite: accurate prediction of human ubiquitination sites by identifying a set of effective negatives," Bioinformatics, vol. 33, no. 5, pp. 661, 2017

2. X. Chen, J. D. Qiu, S. P. Shi, S. B. Suo, S. Y. Huang, and R. P. Liang, "Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites," Bioinformatics, vol. 29, no. 13, pp. 1614, 2013.

3. W. R. Qiu, X. Xiao, W. Z. Lin, and K. C. Chou, "iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model," Journal of Biomolecular Structure & Dynamics, vol. 33, no. 8, pp. 1731, 2015.

4. C. H. Huang, M. G. Su, H. J. Kao, J. H. Jhong, S. L. Weng, and T. Y. Lee, "UbiSite: incorporating two-layered machine learning method with substrate motifs to predict ubiquitin-conjugation site on lysines," Bmc Systems Biology, vol. 10 Suppl 1, no. Suppl 1, pp. 6, 2016.

# 1 Introduction

Challenges on large-scale protein ubiquitylation site prediction

1. Weakness of handcrafted features
   - The conventional feature engineering leads to produce biased and incomplete features
2. Heterogeneity among different features
   - (inspired by iris features) Most existed tools combine multiple modal features
3. Extreme imbalanced data problem
   - Only a small size of lysine can be attached to ubiquitin

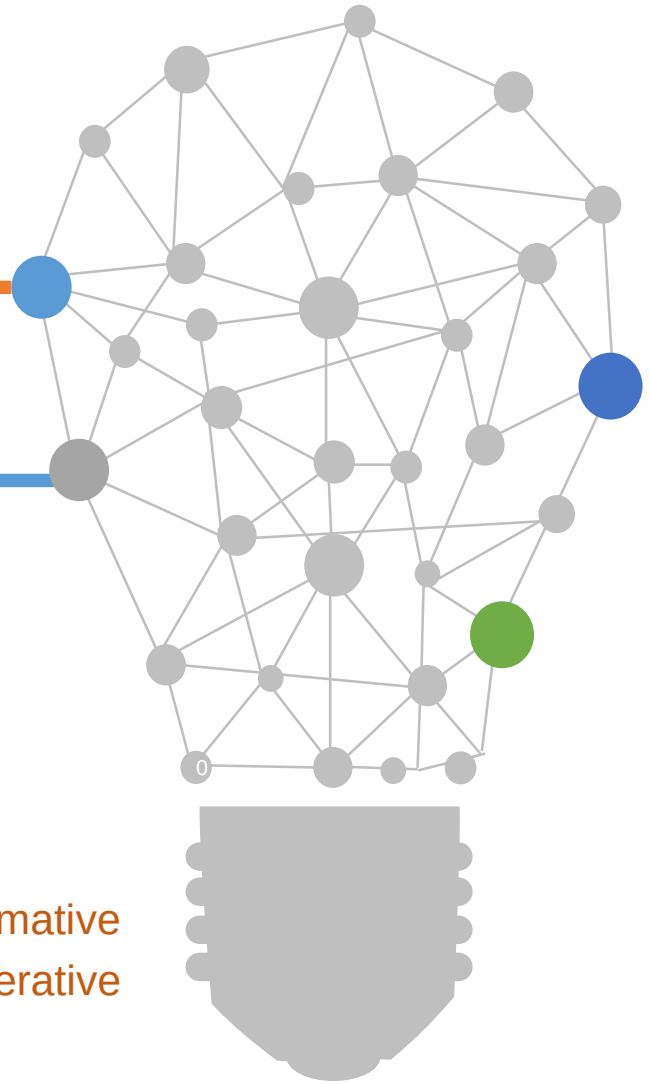
# 1 Introduction

**Deep network**

Detect underlying informative patterns in protein sequence

**Multi-modal fusion**

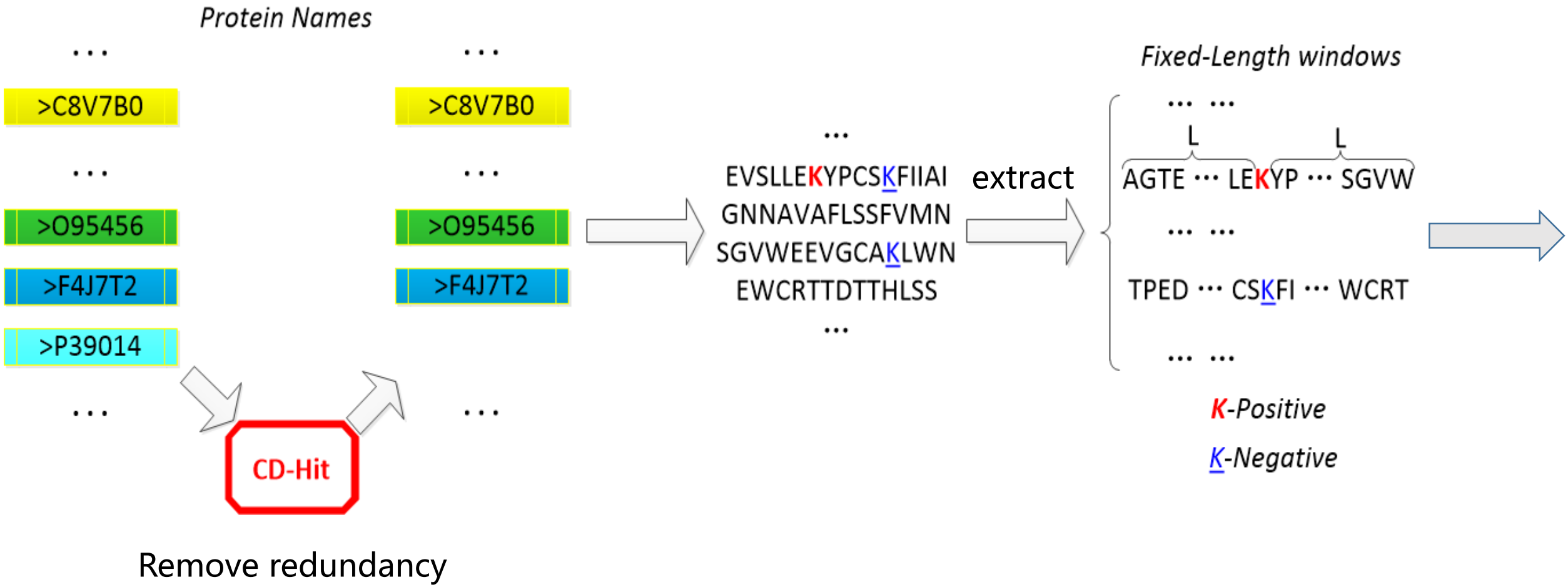
Involve three types of informative modalities and fuse their generative deep representations



**Bootstrap strategy**

Control the model unbiased and stable on imbalanced data

# 2 Method

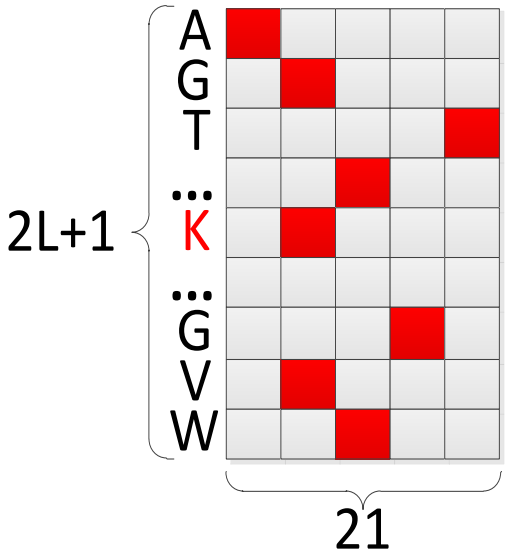



Remove redundancy

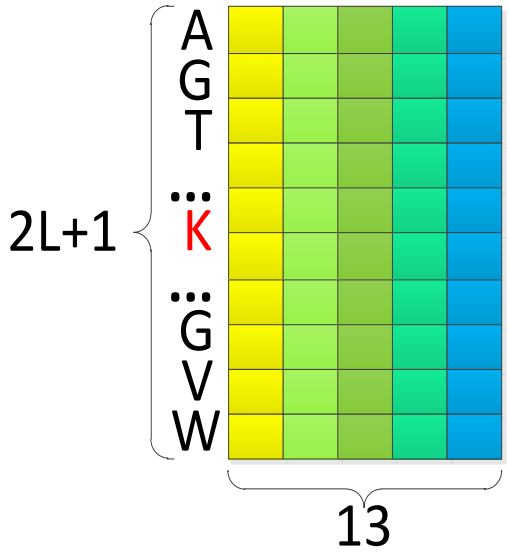
# 2 Method

## Three different categories of protein modalities

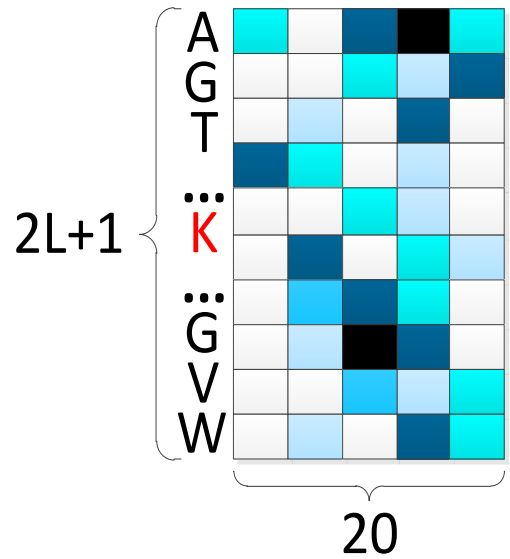
*One-Hot* vector




physico-chemical properties



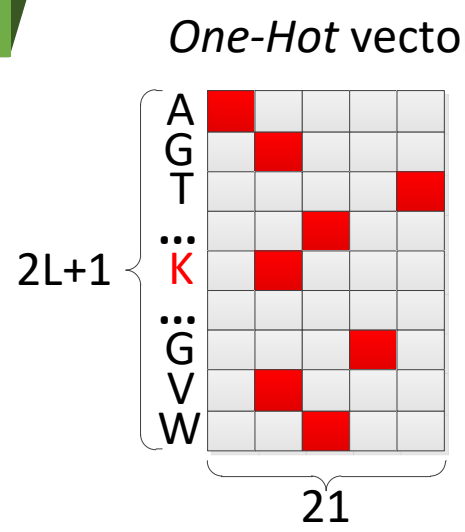

position-specific scoring matrix（PSSM）




Sequence information → PCP for amino acid → Evolutionary profiles

Information Sources

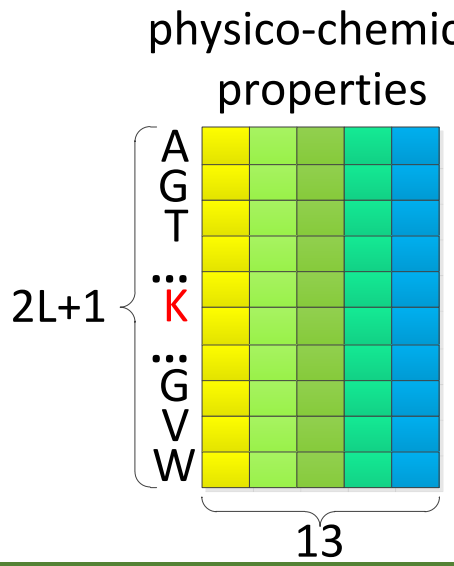
# 2 Method




- Each amino acid is encoded into 21 dimensional binary vector
- The index corresponding to the amino acid is 1, and other positions are 0
- A dash will be filled in absent positions and be encoded to 0.05 (1/21)

A: [1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
C: [0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
D: [0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
E: [0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
F: [0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
G: [0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
H: [0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
I: [0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0]
K: [0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0]
L: [0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0]
M: [0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0]
N: [0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0]
P: [0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0]
Q: [0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0]
R: [0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0]
S: [0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0]
T: [0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0]
V: [0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0]
W: [0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0]
Y: [0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0]
X: [0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1]

# 2 Method

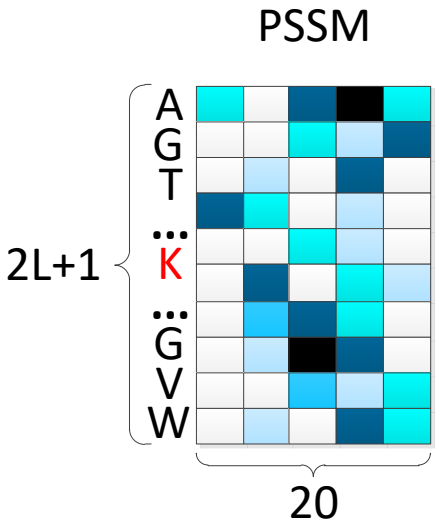



- There is a strong connection between physico-chemical properties of amino acids and ubiquitylation sites[1,2]
- Select top 13 physico-chemical properties in light of the literature[3]

| No | PCP | Description |
|---|---|---|
| 1 | EISD860102 | Atom-based hydrophobic moment |
| 2 | ZIMJ680104 | Isoelectric point |
| 3 | HUTJ700103 | Entropy of formation |
| 4 | KARP850103 | Flexibility parameter for two rigid neighbors |
| 5 | JANJ780101 | Average accessible surface area |
| 6 | FAUJ880111 | Positive charge |
| 7 | GUYH850104 | Apparent partition energies calculated from Janin index |
| 8 | JANJ780103 | Percentage of exposed residues |
| 9 | JANJ790102 | Transfer free energy |
| 10 | PONP800102 | Average gain in surrounding hydrophobicity |
| 11 | CORJ870101 | NNEIG index |
| 12 | VINM940101 | Normalized flexibility parameters, average |
| 13 | OOBM770101 | Average non-bonded energy per atom |


[1] C. W. Tung, and S. Y. Ho, "Computational identification of ubiquitylation sites from protein sequences," Bmc Bioinformatics, vol. 9, no. 1, pp. 310, 2008.
[2] P. Radivojac, V. Vacic, C. Haynes, R. R. Cocklin, A. Mohan, J. W. Heyen, M. G. Goebl, and L. M. Iakoucheva, "Identification, analysis, and prediction of protein ubiquitination sites," Proteins-structure Function & Bioinformatics, vol. 78, no. 2, pp. 365-380, 2010.
[3] X. Chen, J. D. Qiu, S. P. Shi, S. B. Suo, S. Y. Huang, and R. P. Liang, "Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites," Bioinformatics, vol. 29, no. 13, pp. 1614, 2013.

# 2 Method




| | | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | L | -1 | -3 | -3 | -4 | -1 | -3 | -3 | -4 | -3 | 3 | 3 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 |
| 2 | D | -3 | -2 | 1 | 7 | -5 | -1 | 3 | -3 | -2 | -5 | -5 | 0 | -4 | -5 | -3 | -1 | -2 | -5 | -4 | -4 |
| 3 | I | 1 | -3 | -3 | -4 | -2 | -3 | -3 | -3 | -4 | 3 | 2 | -3 | 2 | -1 | -3 | -1 | 0 | -3 | -2 | 3 |
| 4 | Q | 3 | 0 | 1 | 0 | -2 | 1 | 0 | -1 | -2 | -3 | -3 | 2 | -2 | -3 | -2 | 1 | 1 | -4 | -1 | -1 |
| 5 | R | 1 | 3 | 3 | 0 | -3 | 0 | 0 | -1 | 1 | -2 | -2 | 1 | -2 | -3 | -2 | 0 | 1 | -4 | -3 | -2 |
| 6 | G | -1 | -5 | -3 | -4 | -5 | -4 | -5 | 7 | -4 | -6 | -6 | -4 | -5 | -6 | -5 | -3 | -4 | -5 | -6 | -5 |
| 7 | A | 3 | 1 | -2 | -2 | -3 | 2 | 2 | 0 | -2 | -4 | -4 | 4 | -3 | -4 | -3 | 0 | -2 | -4 | -3 | -2 |
| 8 | T | 0 | 3 | 0 | 1 | -4 | 3 | 1 | -3 | -2 | -4 | -4 | 3 | -3 | -4 | -3 | 1 | 0 | -4 | -3 | -3 |
| 9 | L | -2 | -2 | -4 | -5 | -3 | -3 | -3 | -5 | -4 | 3 | 3 | -4 | 0 | -1 | -4 | -3 | -1 | -3 | 0 | 4 |
| 10 | F | -5 | -5 | -6 | -6 | -5 | -6 | -6 | -6 | -3 | -3 | -2 | -6 | -3 | 9 | -6 | -5 | -5 | -1 | 4 | -4 |
| 11 | N | 1 | 1 | 3 | 1 | -3 | 1 | 1 | -1 | -2 | -2 | -2 | 0 | -2 | -3 | -2 | 1 | 1 | -4 | -3 | 0 |
| 12 | R | 2 | -1 | 1 | 1 | -2 | 0 | 1 | 0 | 1 | -2 | -1 | -1 | -2 | -2 | -2 | 0 | 0 | -3 | 0 | -1 |
| 13 | A | 1 | -2 | 6 | -1 | -4 | -2 | -1 | -2 | 1 | -4 | -4 | 0 | -3 | -5 | -3 | 1 | 0 | -5 | -4 | -1 |
| 14 | C | -4 | -7 | -6 | -7 | 11 | -6 | -7 | -6 | -6 | -4 | -4 | -6 | -5 | -6 | -6 | -4 | -4 | -6 | -6 | -4 |
| 15 | A | 5 | -3 | -2 | -3 | -2 | -1 | -3 | 1 | -3 | -2 | -2 | -2 | -2 | -4 | -3 | 1 | 0 | -4 | -3 | 0 |
| 16 | A | 3 | -1 | 1 | -2 | -3 | 3 | -1 | -1 | -2 | -1 | -3 | -1 | -2 | -4 | -3 | 3 | 1 | -4 | -3 | -1 |
| 17 | C | -4 | -7 | -6 | -7 | 11 | -6 | -7 | -6 | -6 | -4 | -4 | -6 | -5 | -6 | -6 | -4 | -4 | -6 | -6 | -4 |
| 18 | H | -5 | -3 | -2 | -4 | -6 | -3 | -3 | -5 | 11 | -7 | -6 | -4 | -5 | -4 | -5 | -4 | -5 | -6 | -1 | -6 |
| 19 | D | 2 | -2 | 0 | 0 | -2 | 0 | -2 | 1 | -2 | 0 | 0 | -1 | 1 | -2 | -2 | -1 | 0 | -3 | -2 | 2 |
| 20 | T | 1 | -3 | 2 | 0 | -3 | 0 | -2 | 5 | -3 | -4 | -3 | -2 | -3 | -4 | -3 | 0 | 1 | -4 | -4 | -3 |
| ... | ... | | | | | | | | | | | | | | | | | | | | |

- Demonstrate the evolutionary profiles of the protein sequences
- Search database : Swiss-Prot
- Parameter : num_iterations 3, e-value 0.001
- Logistic normalization: $1/(1+ e^{-x})$

# 2 Method

**One hot vector**

3- layer 1D CNN to detect local structural feature maps

3-layer DNN to combine all feature maps

**Physico-chemical properties**

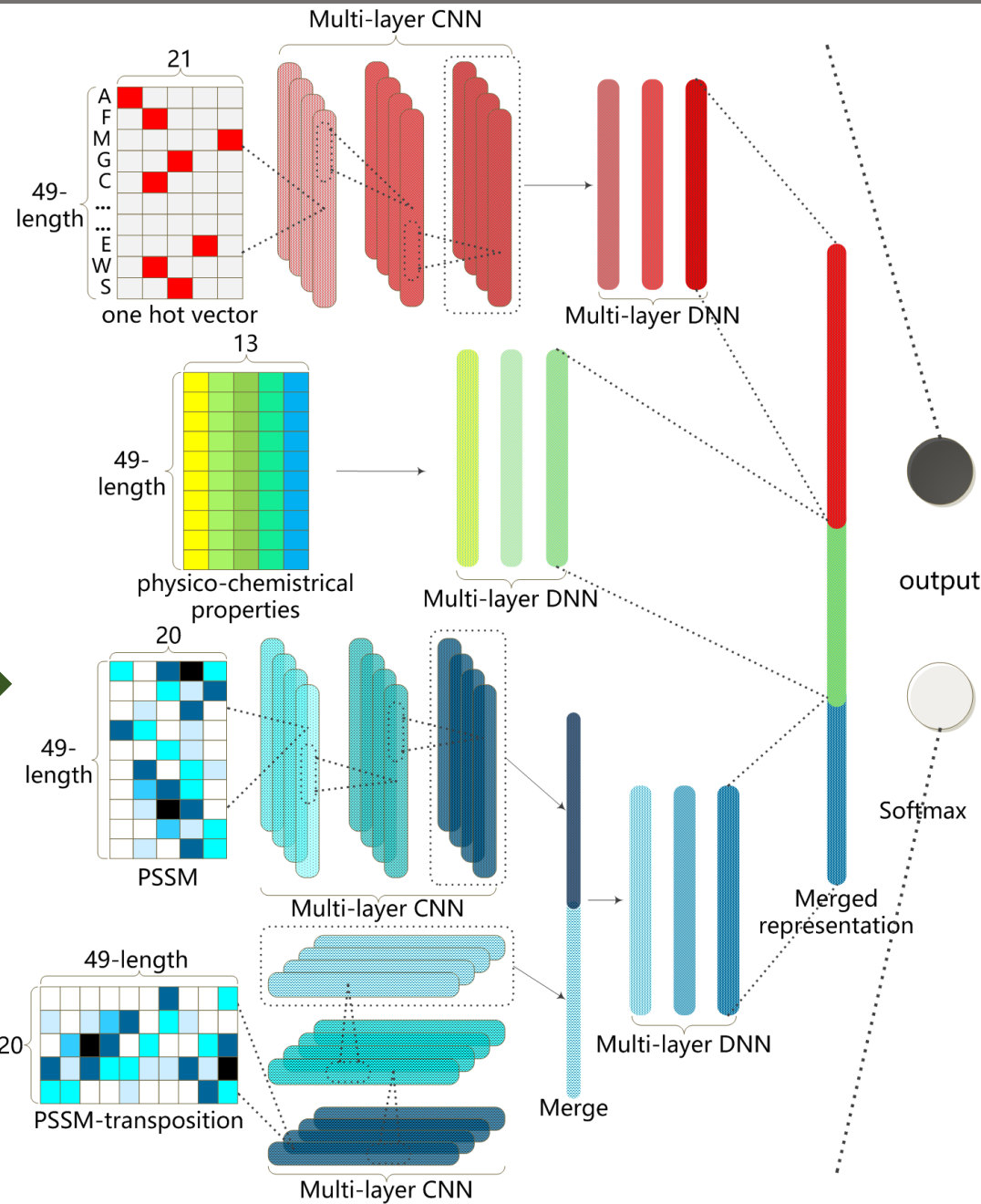
3-layer DNN to interconnect all properties for their joint effect

**PSSM**

3-layer 1D CNN to detect evolutionary profile among amino acids

Another 3-layer 1D CNN to detect evolutionary information cross positions from trans-positional PSSM

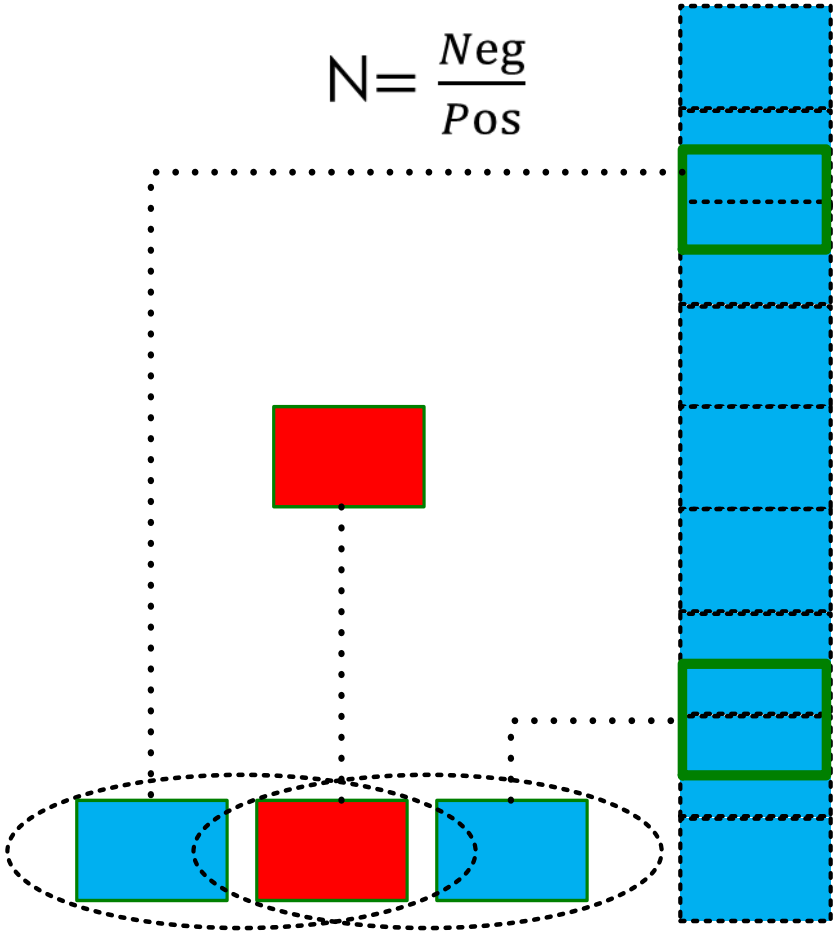
2-layer DNN to merge the feature maps



# 2 Method

## bootstrapping

$$N = \frac{Neg}{Pos}$$



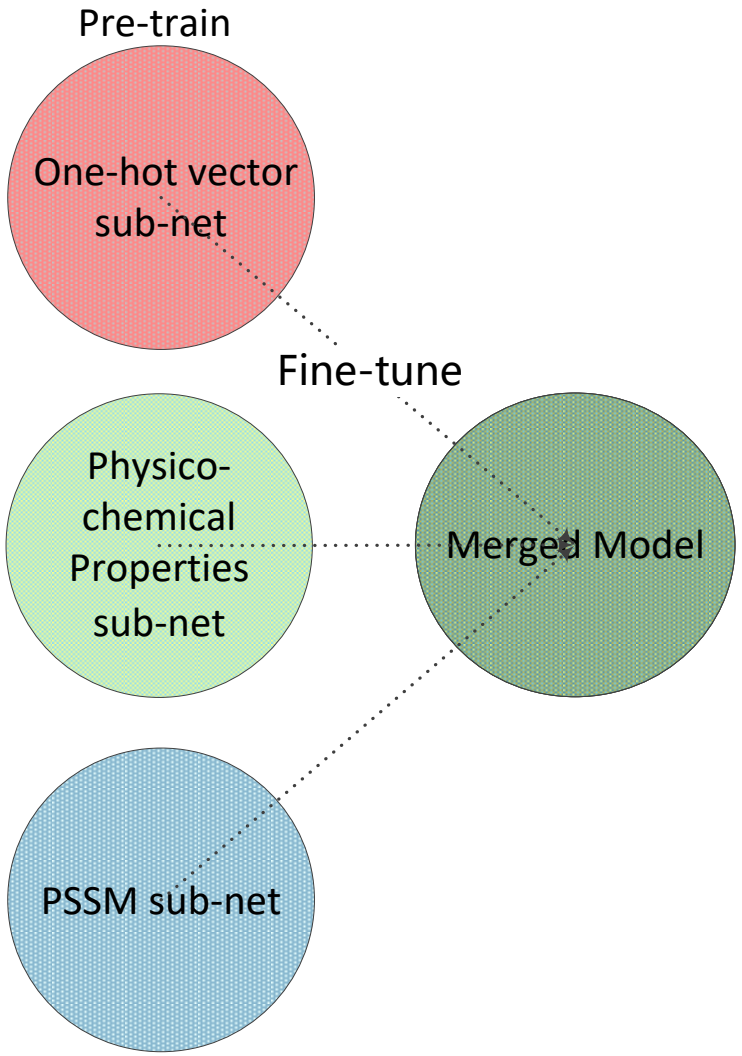
Run on:

Theano 0.9

Keras 1.1.0

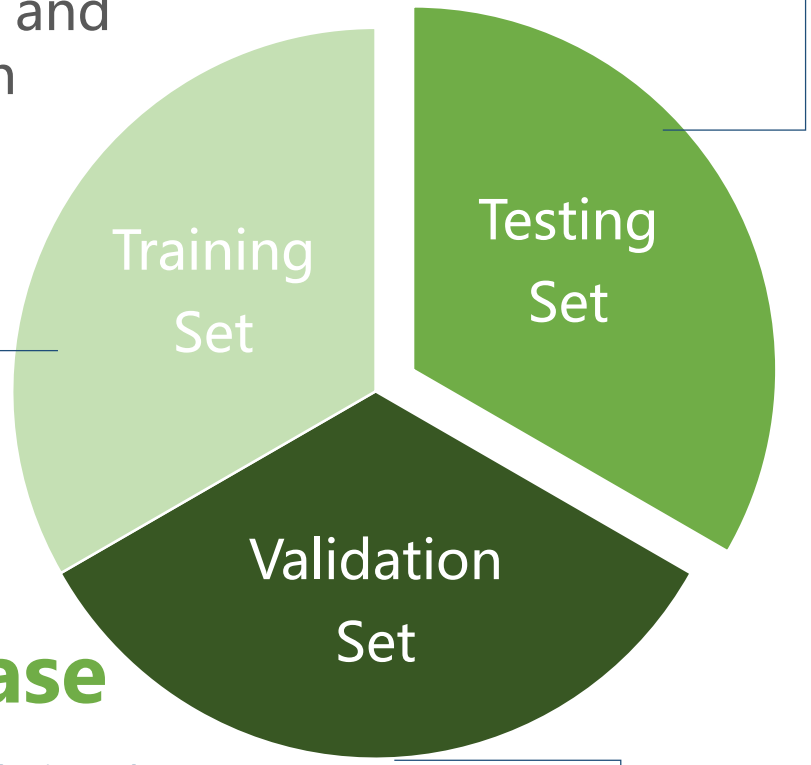
GTX1080Ti

## Pre-train



# 3 Material

12,100 protein sequences with 54,586 Ubiquitylation sites and 320,083 non-ubiquitylation sites




1345 proteins with 6293 Ubiquitylation sites and 46,080 non-ubiquitylation sites

extract 30% of training set as validation samples

## Protein Lysine Modification Database

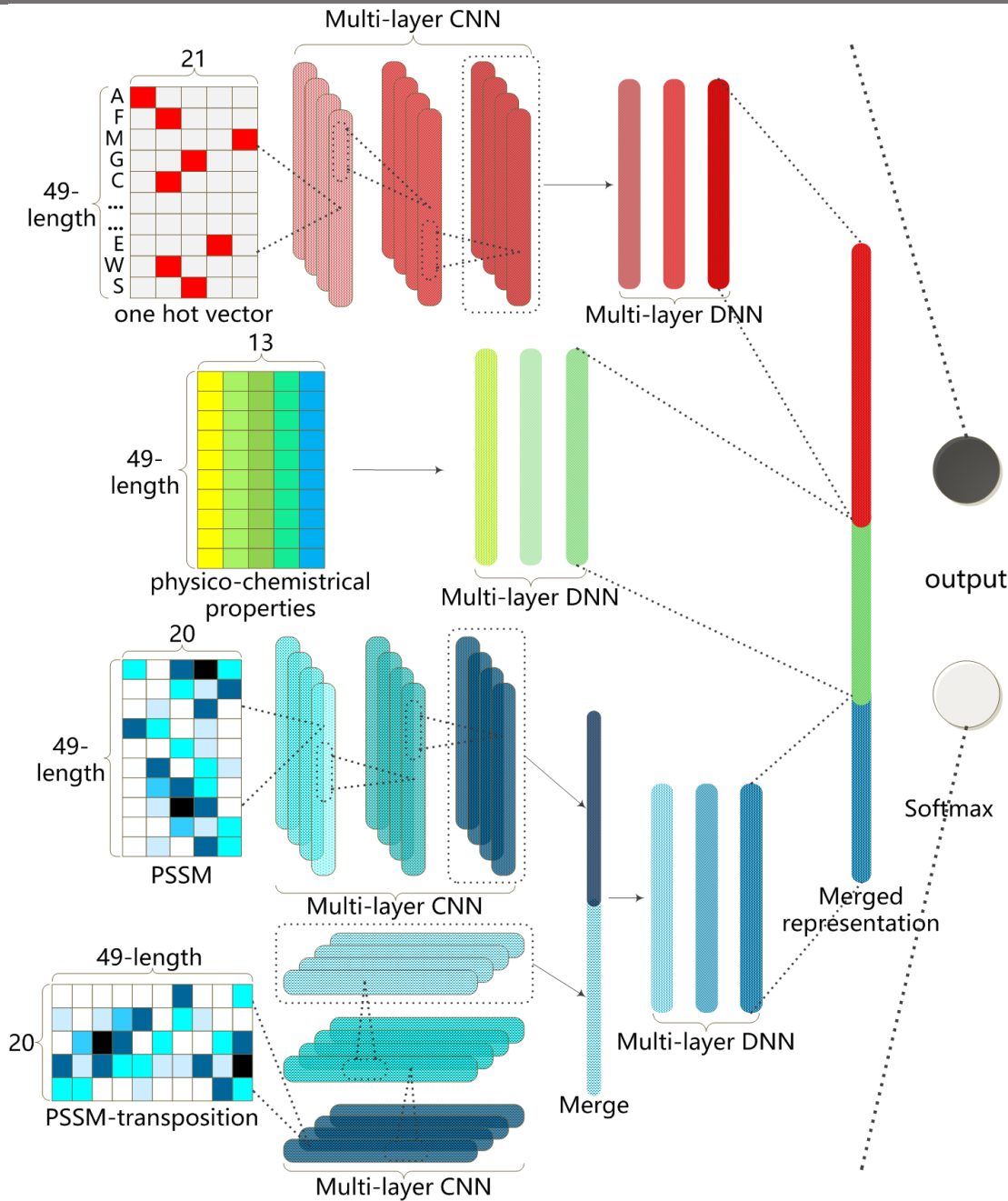
The available largest scale protein ubiquitylation dataset.
Extended from CPLA 1.0 and CPLM 2.0
Never mentioned in any other protein ubiquitylation site prediction research.

# 2 Method

| Subnet | Layer | Hyper-parameters | | | |
|---|---|---|---|---|---|
| | | Activation function | Size | Filters | Dropout |
| One hot vector | 1D Convolution | softsign | 2 | 200 | 0.4 |
| | | softsign | 3 | 150 | 0.4 |
| | | softsign | 5 | 150 | 0.4 |
| | | softsign | 7 | 100 | 0.4 |
| | Dense | relu | 256 | -- | 0.3 |
| | | relu | 128 | -- | 0 |
| | | relu | 128 | -- | -- |

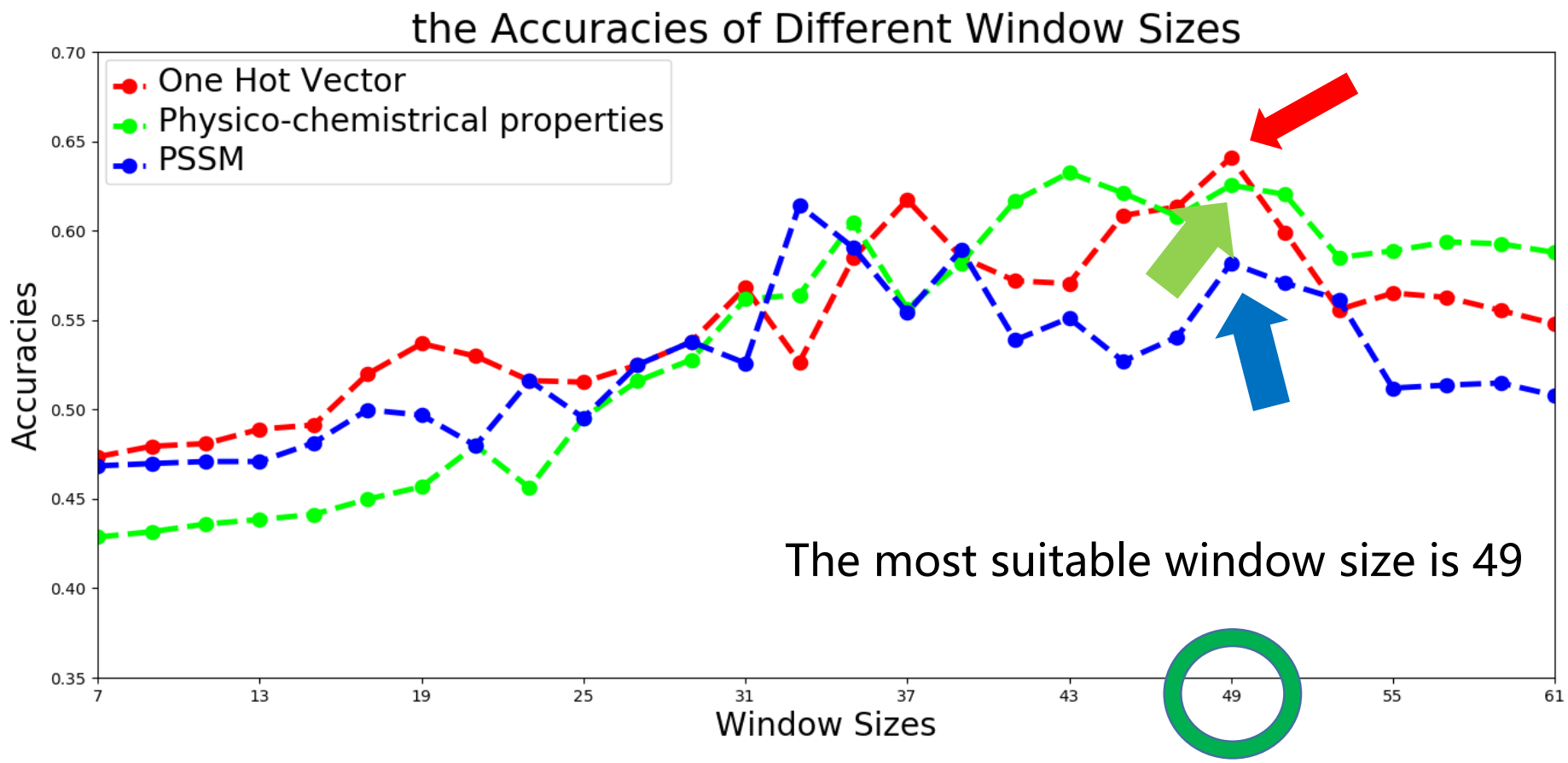
| Subnet | Layer | Hyper-parameters | | | |
|---|---|---|---|---|---|
| | | Activation function | Size | Filters | Dropout |
| Phsico-chemical properties | Dense | softplus | 1024 | -- | 0.2 |
| | | softplus | 512 | -- | 0.4 |
| | | softplus | 256 | -- | 0.5 |
| | | relu | 128 | -- | -- |

| Subnet | Layer | Hyper-parameters | | | |
|---|---|---|---|---|---|
| | | Activation function | Size | Filters | Dropout |
| PSSM profile | 1D Convolution | relu | 1 | 200 | 0.5 |
| | | relu | 8 | 150 | 0.5 |
| | | relu | 9 | 200 | 0.5 |
| | 1D Convolution | relu | 1 | 200 | 0.5 |
| | | relu | 3 | 150 | 0.5 |
| | | relu | 7 | 200 | 0.5 |
| | Dense | relu | 128 | -- | 0.3 |
| | | relu | 128 | -- | 0 |

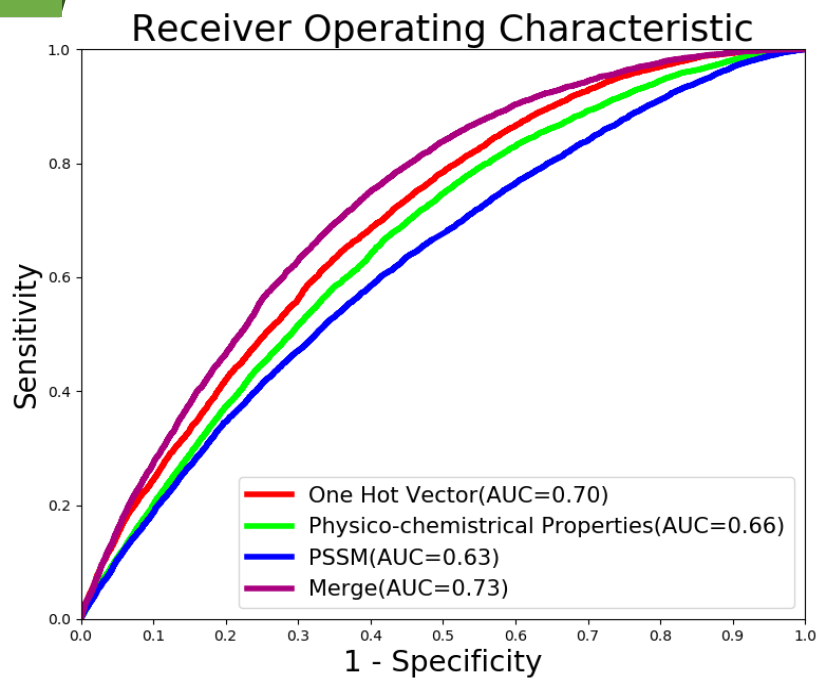


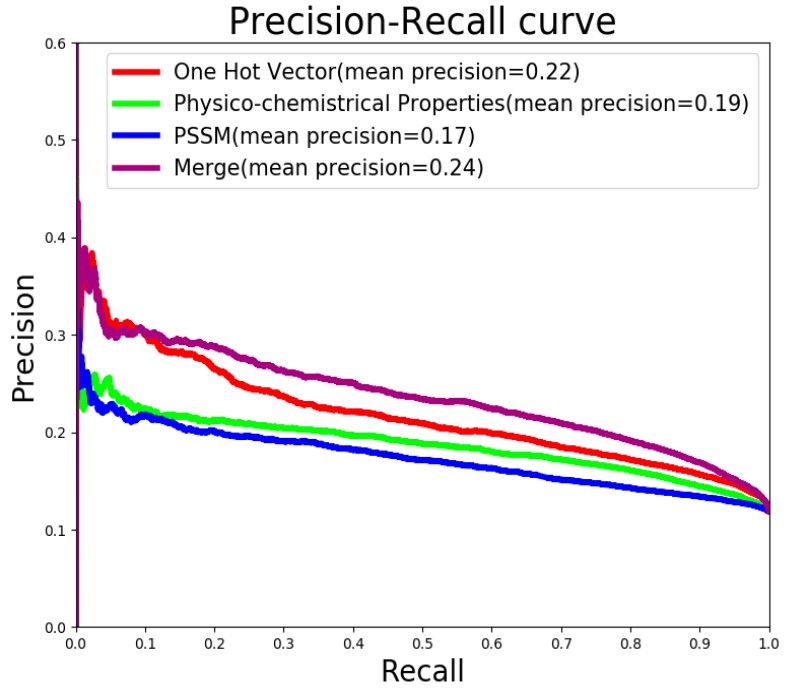
# 4 Results

The accuracy of validation samples using different window sizes on three modalities



# 4 Results




**merged model achieved the better AUC and mean precision than uni-modality**

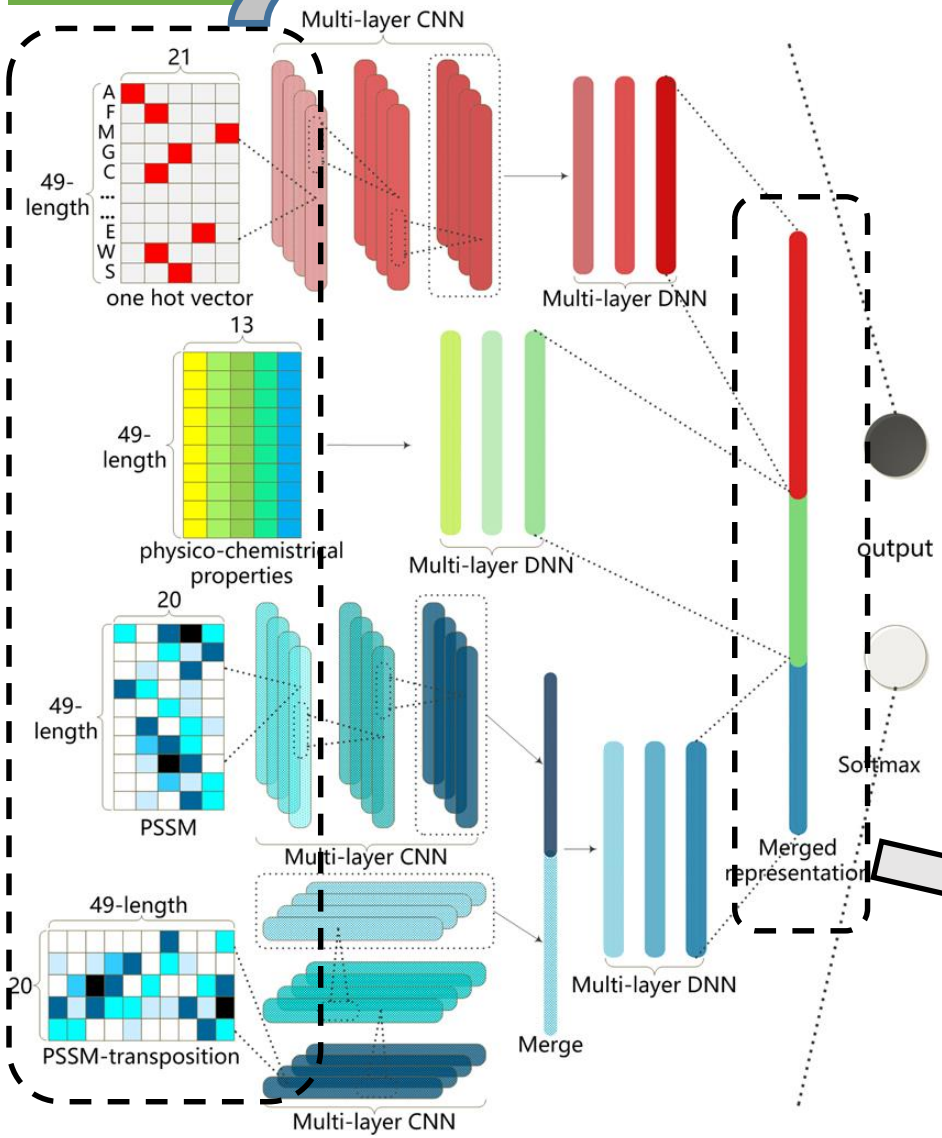
**→ make all input modalities at full capacity**

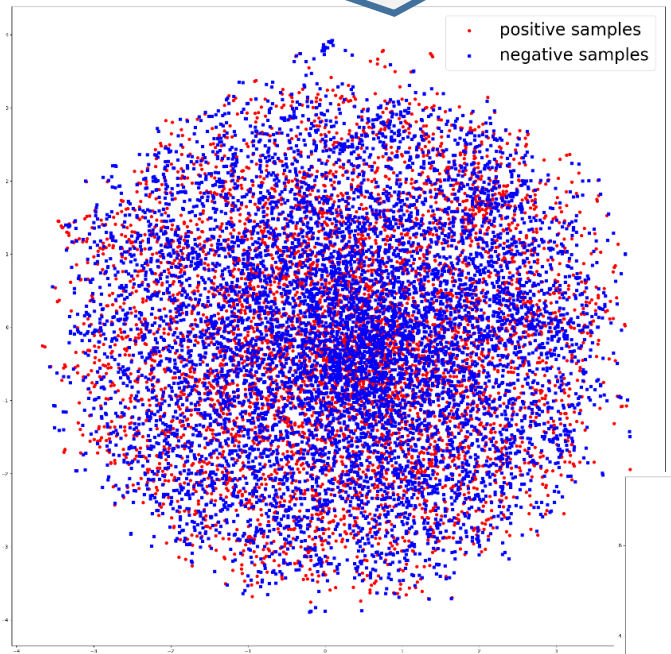
**one hot vector performed the best among the three input modalities**

**→ detect underlying expressions from raw protein sequence fragments**

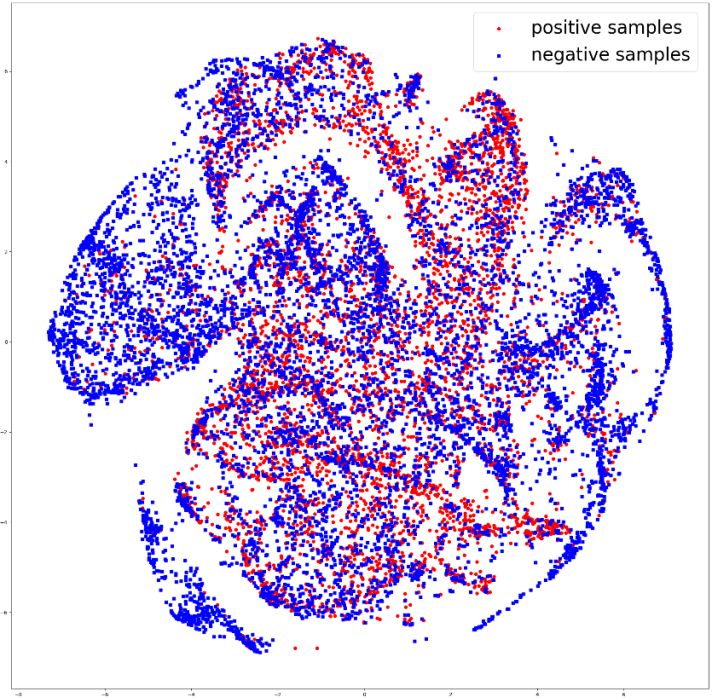
# 4 Results

t-SNE visualization




t-SNE visualization



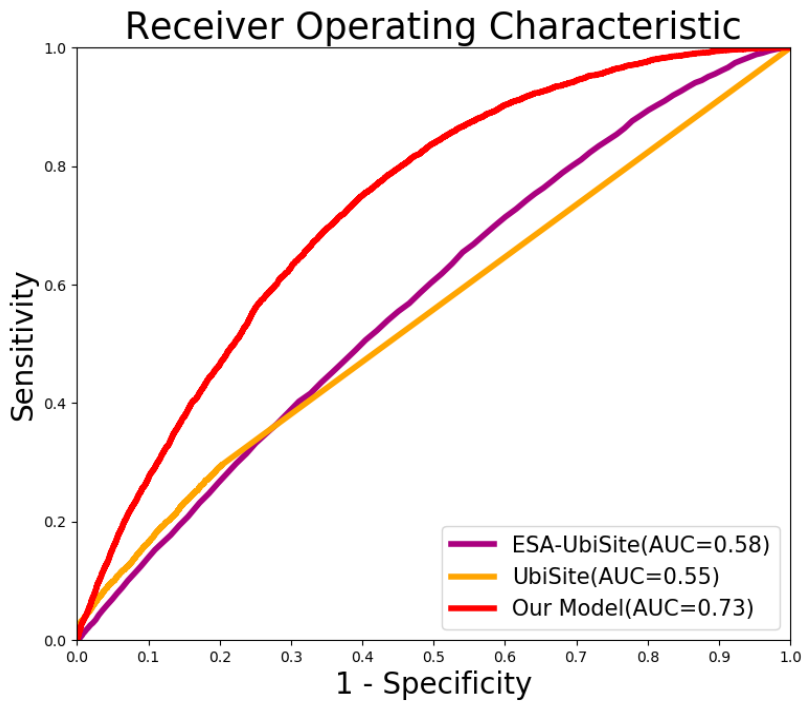
# 4 Results

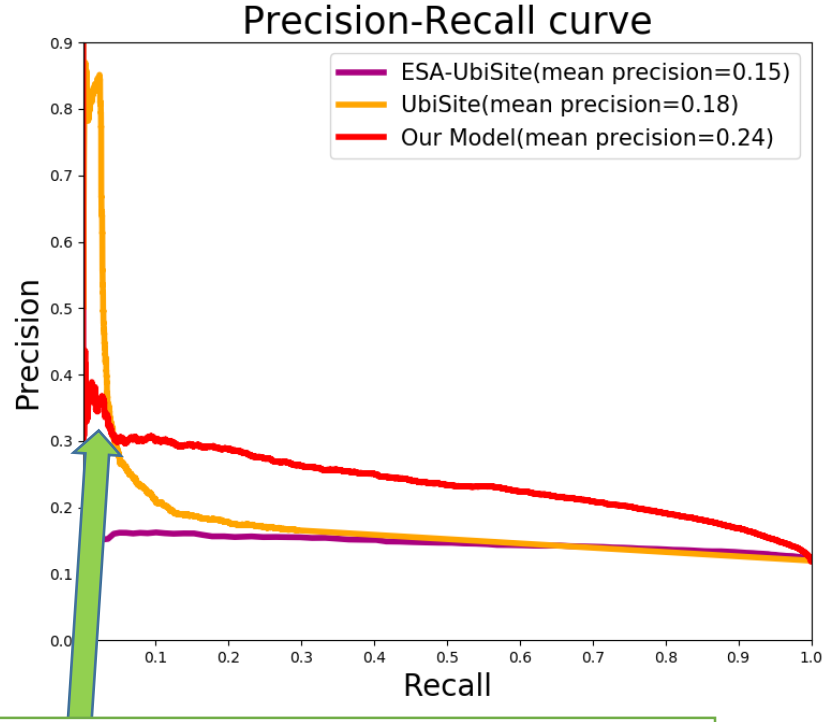
**Comparative results with other protein Ubiquitylation site prediction tools on PLMD**

| Tool | Metrics | | | |
|---|---|---|---|---|
| | Accuracy | Sensitivity | Specificity | MCC |
| **ESA-Ubisite** | 61.26% | 46.14% | 63.34% | 0.064 |
| **UbiProber** | 55.06% | 62.40% | 54.05% | 0.107 |
| **iUbiq-Lys** | 84.63% | 3.35% | 96.88% | 0.005 |
| **Ubisite** | 73.63% | 29.62% | 79.64% | 0.073 |
| **Our deep architecture** | **66.43%** | **66.67%** | **66.40%** | **0.221** |

# 4 Results

The ROC and precision-recall curves comparing proposed deep architecture and two other protein ubiquitylation site prediction tools




Only under a certain minor recall, Ubisite achieved higher precision

**Our model performed at a higher ROC**

**Our model obtained better AUC and mean precision**

**Our deep architecture has evident overall advantages**

# 5 Summary

Encode each sample into three informative modalities including one hot vector, physico-chemical properties and PSSM

Establish a multimodal deep architecture fusing these encoding modalities was for robust classification

Experimental results have proved our effectiveness on the available largest scale data PLMD

The success of our method is mainly due to the data-driven feature detection in deep learning, the multimodal fusion of deep representations, and the bootstrapping algorithm.

# Thanks for listening

**Source codes:**
**https://github.com/jiagenlee/deepUbiquitylation**

hef740@nenu.edu.cn