MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization

Duolin Wang.[1,2] Dongpeng Liu,[2] Jiakang Yuchi,[2] Fei He,[1,3] Yuexu Jiang,[1,2] Siteng Cai,[2] Jingyi Li,[3] Dong Xu[1,2*]

[1] Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA
[2] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211 USA
[3]Department of Computer Science and Information Technology, Northeast Normal University, Changchun, Jilin 130117, China

* To whom correspondence should be addressed. Tel: 573-808-2219; Fax: 573-882-8318; Email: xudong@missouri.edu.

Text S1. Architecture details of the deep-learning framework.

We used the combination of the two previously used networks, MultiCNN [1] and CapsNet [2], to build the deep-learning framework.

The MultiCNN architecture:

1. 1D Conv1: 200 filters; kernel size: 1; stride: 1; activation function: ReLU; dropout rate:0.75

2. 1D Conv2: 150 filters; kernel size: 9; stride: 1; activation function: ReLU; dropout rate: 0.75

3. 1D Conv3: 200 filters; kernel size: 10; stride: 1; activation function: ReLU; dropout rate:0.75

4. Two-dimensional attention layer

The two-dimensional attention layer consists of two independent attention layers, Attention 1 and Attention 2 in Figure 1 of the main manuscript, which have the same architectures but different hyperparameters. Let h_t be a hidden state output from the Conv3, t = 1,2, …, T (T = 33 for Attention 1 and T = 200 for Attention 2). The output H' of each attention layer is a weighted sum of the input hidden states:

$$H' = \sum_{t=1}^{T} h_t \alpha_t \quad (1)$$

where α_t is the SoftMax weight of each hidden state h_t, which is calculated by:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)} \quad (2)$$

$$e_t = f(f(h_t W)U^T) \quad (3)$$

where e_t is generated from the hidden state h_t by a feedforward neural network function (3). W represents the attention hidden matrix, U represents the attention hidden vector, and f represents the linear activation function.

Attention 1: W in Equation (3) has 200 by 200 hidden units; U in Equation (3) has 200 hidden units. L1 regularization on W is 0.151948

Attention 2: W in Equation (3) has 33 by 8 hidden units; U in Equation (3) has 8 by 1 hidden unit; L1 regularization on W is 2

5. Fully connected layer: 149 hidden units.

6. Fully connected layer: 8 hidden units.

7. SoftMax output layer: 2 hidden units.

Loss function: cross-entropy

The CapsNet architecture:

1. 1D Conv1: 200 filters; kernel size: 1; stride: 1; activation function: ReLU; dropout rate:0.75.

2. 1D Conv2: 200 filters; kernel size: 9; stride: 1; activation function: ReLU; dropout rate: 0.75.

3. PrimaryCaps: 480 filters, which are reshaped into 60 channels of 8D capsules; kernel size: 20; stride: 1; activation function: ReLU; dropout rate:0.75.

4: Dynamic Routing: 3 routing iterations.

5. PTMCaps: positive capsule: 10 hidden units; negative capsule: 10 hidden units.

Loss function: margin loss function.

The actual output of each network contains two scalar neurons. One represents the prediction score of the positive class, and the other represents the prediction score of the negative class. Only the prediction score of the positive class will be reported as a single prediction score for each network. The final prediction score is calculated by averaging the two prediction scores obtained by the two independent networks

Table S1. An example of the prediction result file.

ID	Position	Residue	PTMscores	Cutoff = 0.5
>sp|P97756 Calcium/calmodulin-dependent protein kinase				
sp|P97756	3	R	Methylarginine:0.026	None
sp|P97756	4	S	Phosphoserine:0.847; O-linked_glycosylation:0.096	Phosphoserine:0.847
sp|P97756	5	P	Hydroxyproline:0.662	Hydroxyproline:0.662
sp|P97756	12	P	Hydroxyproline:0.06	None
			…	
>sp|A9QT41 NF-kappa-B essential modulator				
sp|A9QT41	2	S	Phosphoserine:0.103; O-linked_glycosylation:0.052	None
sp|A9QT41	3	R	Methylarginine:0.041	None
sp|A9QT41	4	T	Phosphothreonine:0.627; O-linked_glycosylation:0.041	Phosphothreonine:0.627
sp|A9QT41	5	P	Hydroxyproline:0.466	None
			…	

Prediction results for the selected PTM models: methylarginine, phosphoserine, O-linked_glycosylation, and hydroxyproline. The first line of the file is the header. The following lines are prediction results for each sequence. For each sequence, the first line is the sequence title (starts with ">"), and each of the following lines contains 5 columns separated by tabs. They are: the protein identifier (continuous characters before the first space except ">"), the position of the potential PTM sites (Position); the amino acid code of the residue at the position (Residue); the potential PTMs and their predicted confidence scores (PTMscores); the predicted PTMs whose scores are higher than the present cut-off (Cutoff=0.5). The default cut-off is 0.5, which can be changed by users.

Table S2. Timestamp benchmark data.

PTM types	Training set	Test set
	# of positive/negative fragments	# of positive/negative
Phosphoserine/threonine	135556/2803647	8759/230755
Phosphortyrosine	9427/93291	499/5540
N-linked glycosylation	90344/511755	20522/120384
O-lined glycosylation	4216/103771	218/6248
N6-acetyllysine	22355/274668	683/11371
Methylarginine	4675/99946	269/6859
Methyllysine	2781/45524	154/2001
S-palmitoylation-cysteine	3812/26573	151/684
Pyrrolidone-carboxylic-acid	1394/10528	230/891
Ubiquitination	3707/49963	514/6621
SUMOylation	1225/23932	65/1310
Hydroxylysine	356/2650	9/37
Hydroxyproline	2773/11761	422/814

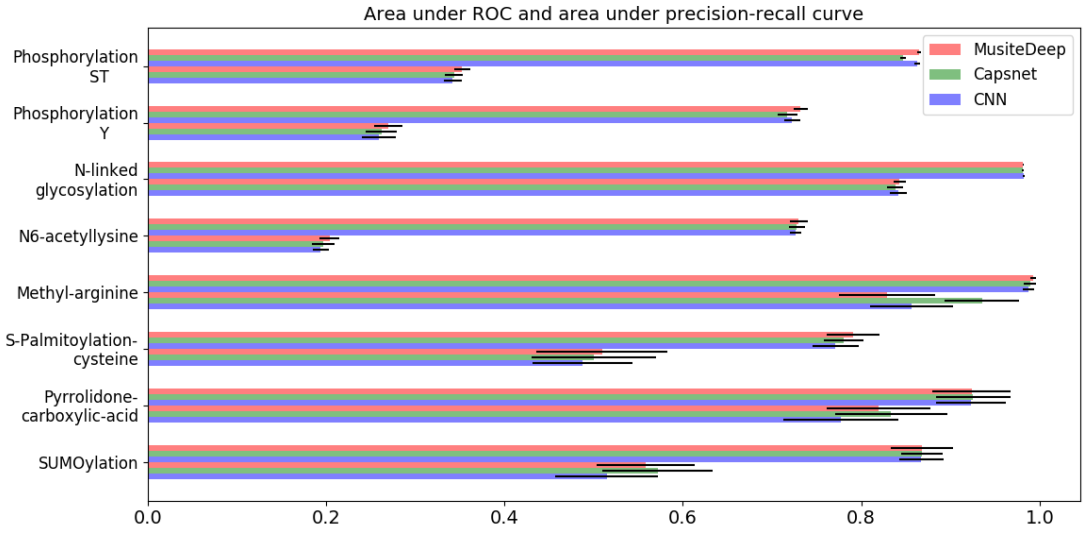
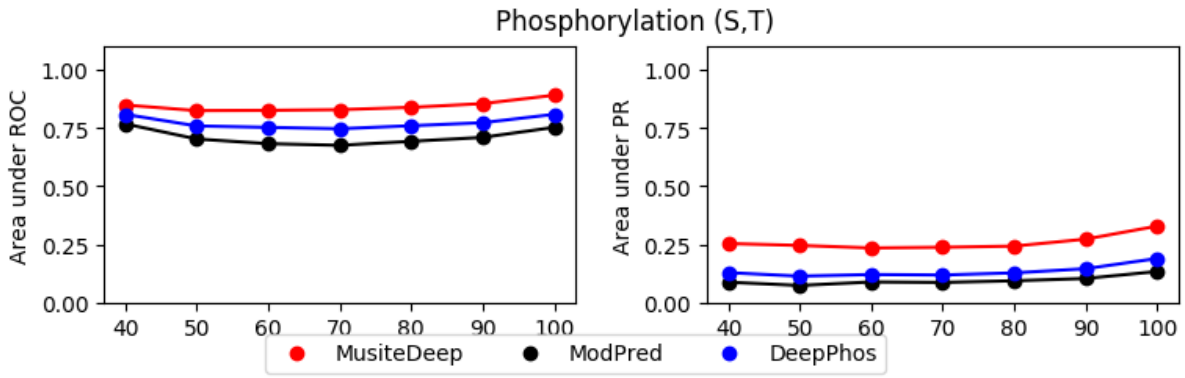
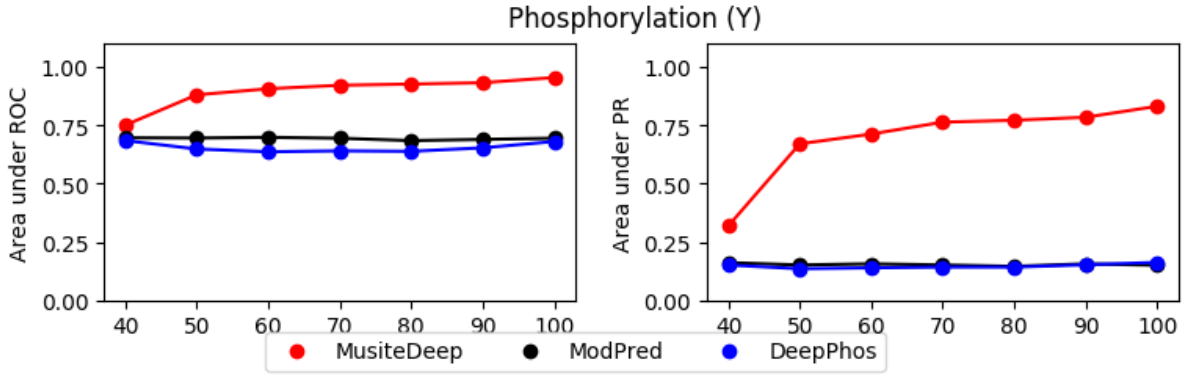
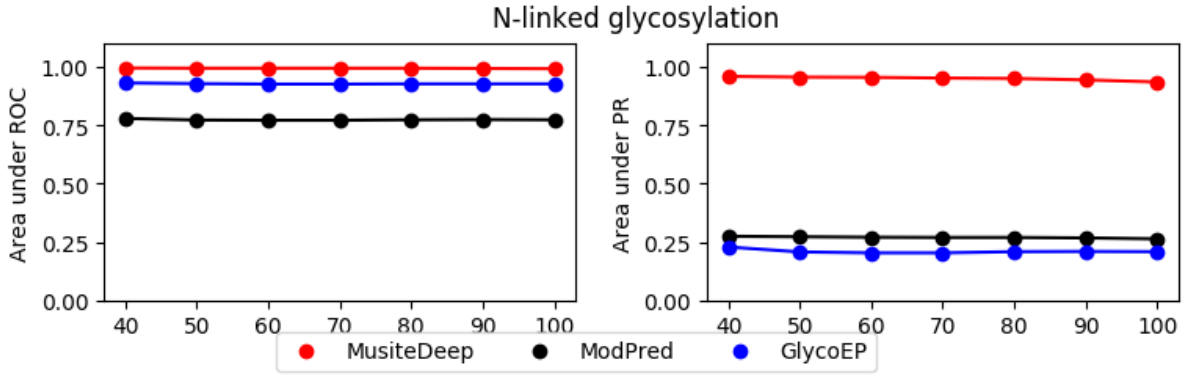
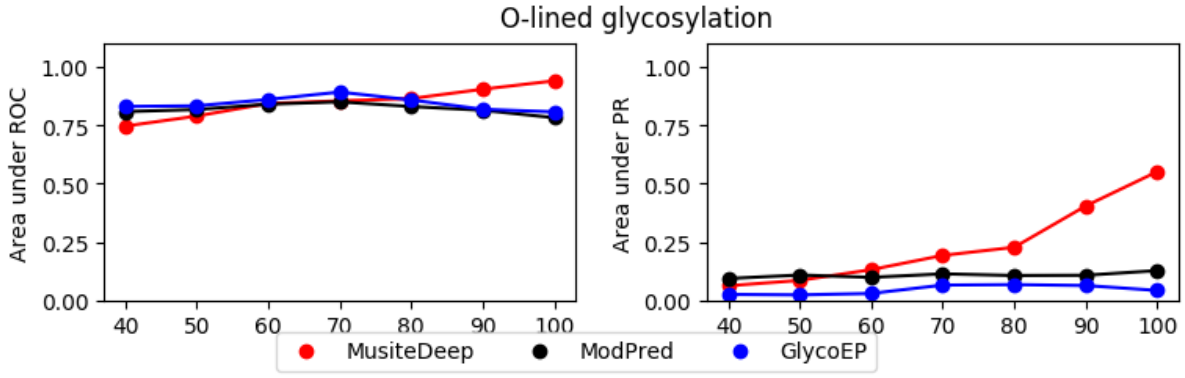


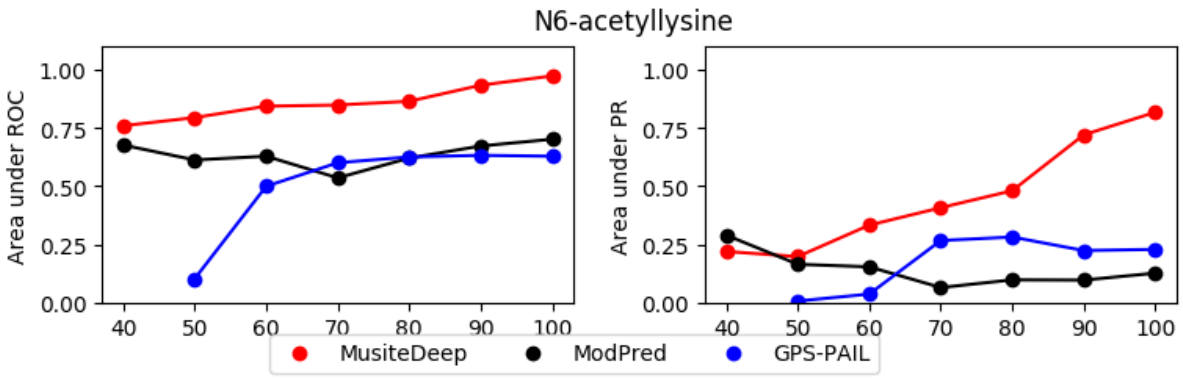
Figure S1. Performance on 10-fold cross-validation dataset. MusiteDeep: the upgraded method used in the MusiteDeep server. CNN: the method in our first work [1]. Capsnet: the method in our second work [2]. For each PTM type, the upper three bars represent the average and +/- standard deviation of the area under the ROC curves, and the lower three bars represent the average and +/- standard deviation of the area under the precision-recall curves for each 10-fold cross-validation dataset provided in [2].

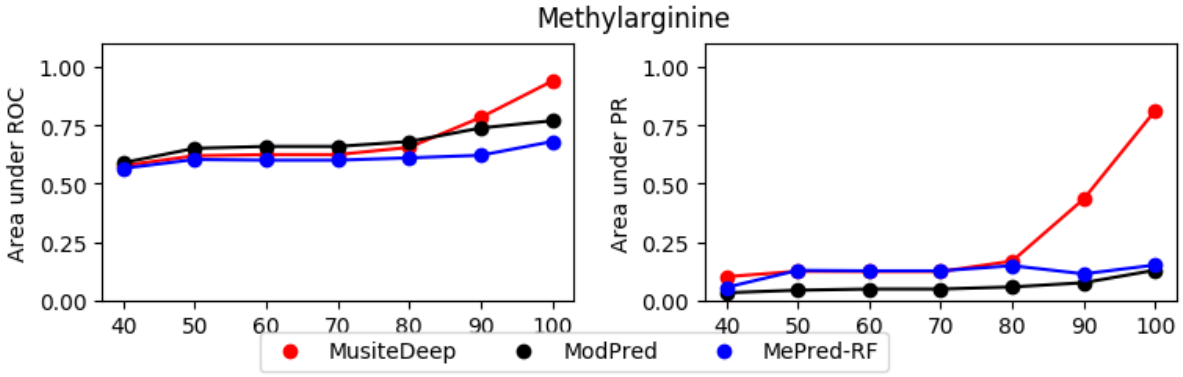

Phosphorylation (S,T)
Area under ROC
Area under PR
1.00
0.75
0.50
0.25
0.00
40
50
60
70
80
90
100
MusiteDeep
ModPred
DeepPhos

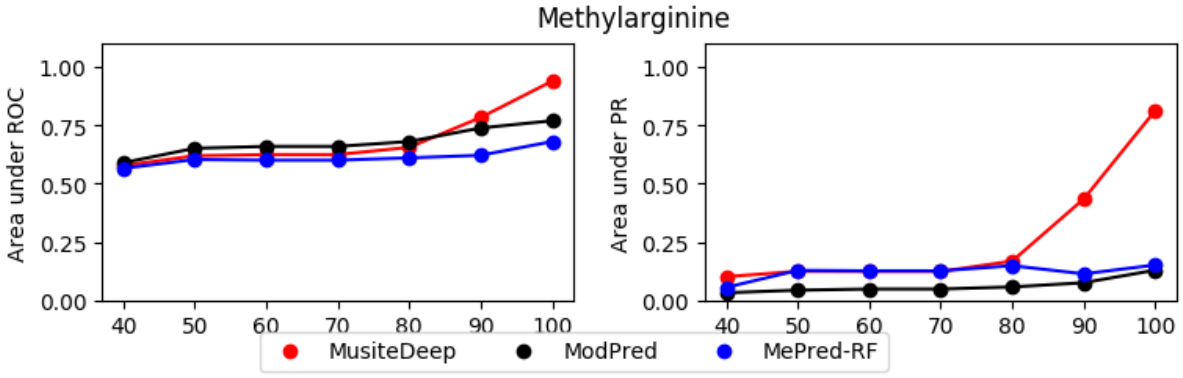

Phosphorylation (Y)
Area under ROC
Area under PR
1.00
0.75
0.50
0.25
0.00
40
50
60
70
80
90
100
MusiteDeep
ModPred
DeepPhos

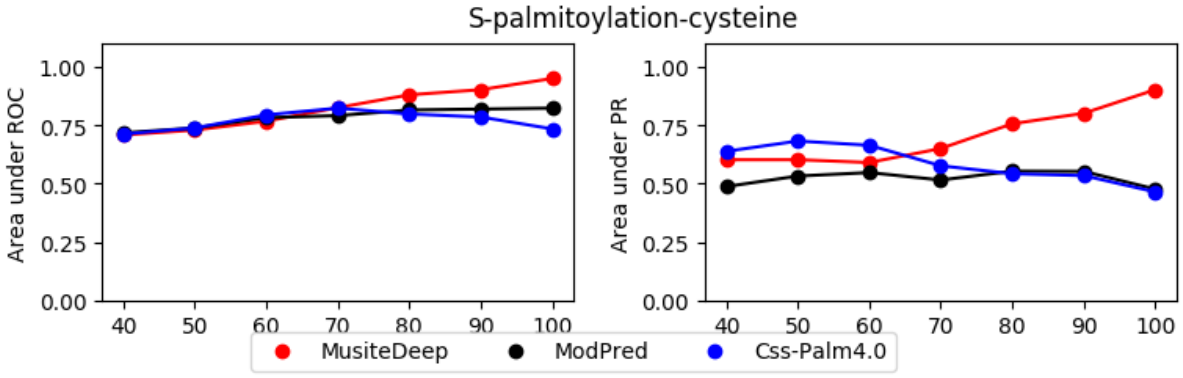

N-linked glycosylation
Area under ROC
Area under PR
1.00
0.75
0.50
0.25
0.00
40
50
60
70
80
90
100
MusiteDeep
ModPred
GlycoEP

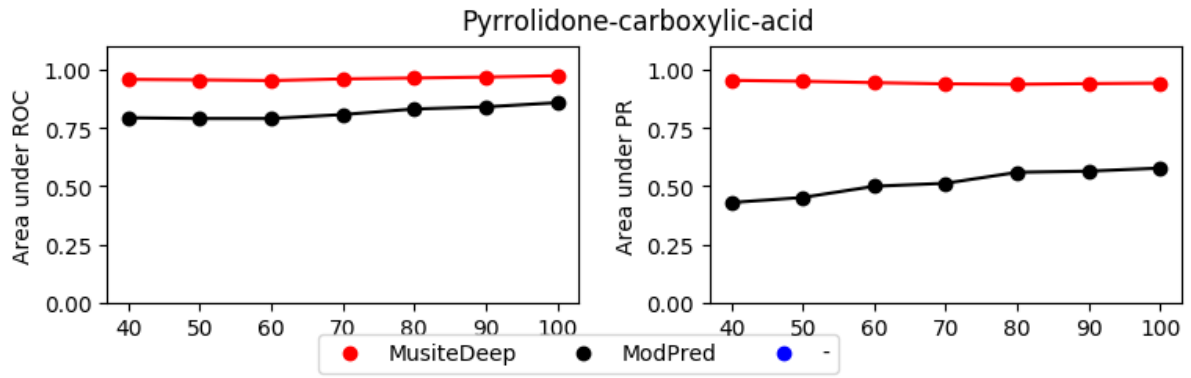

O-lined glycosylation
Area under ROC
Area under PR
1.00
0.75
0.50
0.25
0.00
40
50
60
70
80
90
100
MusiteDeep
ModPred
GlycoEP

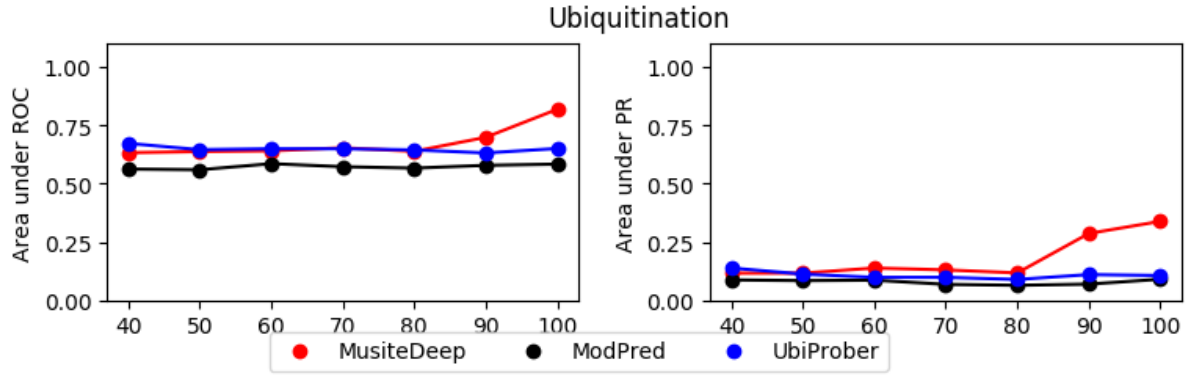

N6-acetyllysine
Area under ROC
Area under PR
1.00
0.75
0.50
0.25
0.00
40
50
60
70
80
90
100
MusiteDeep
ModPred
GPS-PAIL

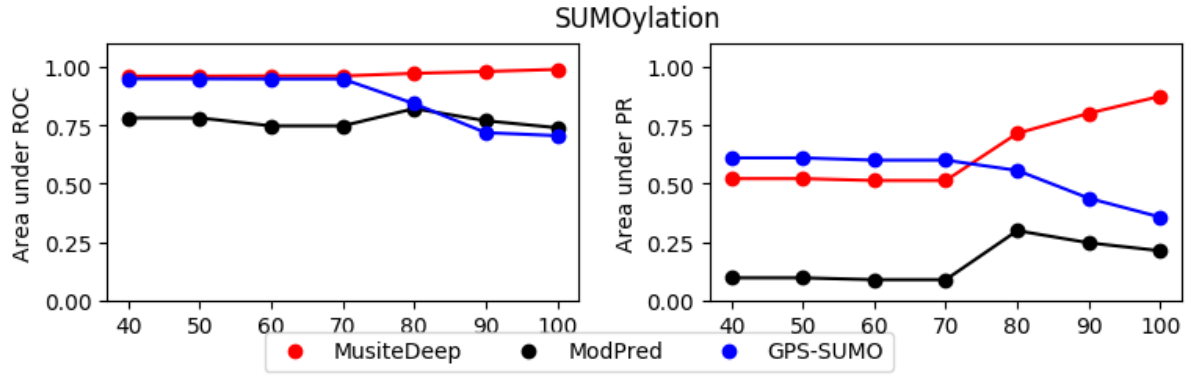

Methylarginine
Area under ROC
Area under PR
1.00
0.75
0.50
0.25
0.00
40
50
60
70
80
90
100
MusiteDeep
ModPred
MePred-RF

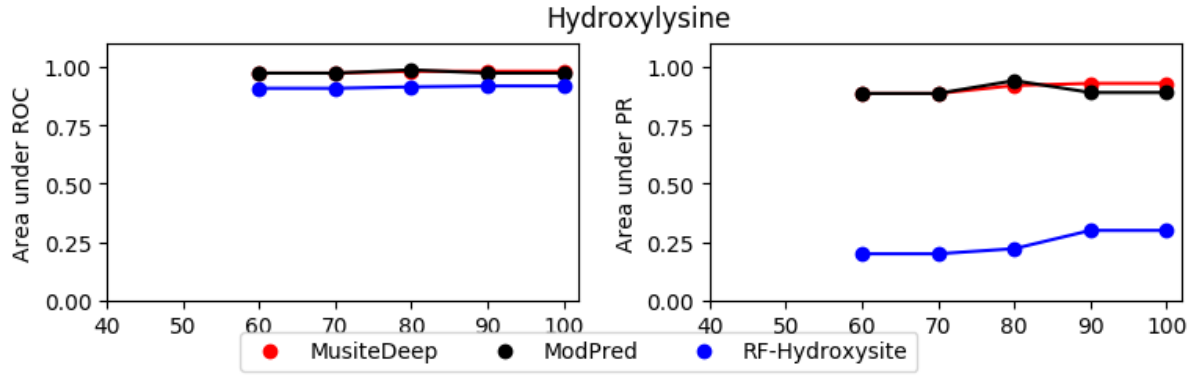

Methylarginine
Area under ROC
Area under PR
1.00
0.75
0.50
0.25
0.00
40
50
60
70
80
90
100
MusiteDeep
ModPred
MePred-RF


S-palmitoylation-cysteine
Area under ROC
Area under PR
1.00
0.75
0.50
0.25
0.00
40
50
60
70
80
90
100
MusiteDeep
ModPred
Css-Palm4.0


Pyrrolidone-carboxylic-acid
Area under ROC
Area under PR
1.00
0.75
0.50
0.25
0.00
40
50
60
70
80
90
100
MusiteDeep
ModPred
-


Ubiquitination
Area under ROC
Area under PR
1.00
0.75
0.50
0.25
0.00
40
50
60
70
80
90
100
MusiteDeep
ModPred
UbiProber


SUMOylation
Area under ROC
Area under PR
1.00
0.75
0.50
0.25
0.00
40
50
60
70
80
90
100
MusiteDeep
ModPred
GPS-SUMO


Hydroxylysine
Area under ROC
Area under PR
1.00
0.75
0.50
0.25
0.00
40
50
60
70
80
90
100
MusiteDeep
ModPred
RF-Hydroxysite

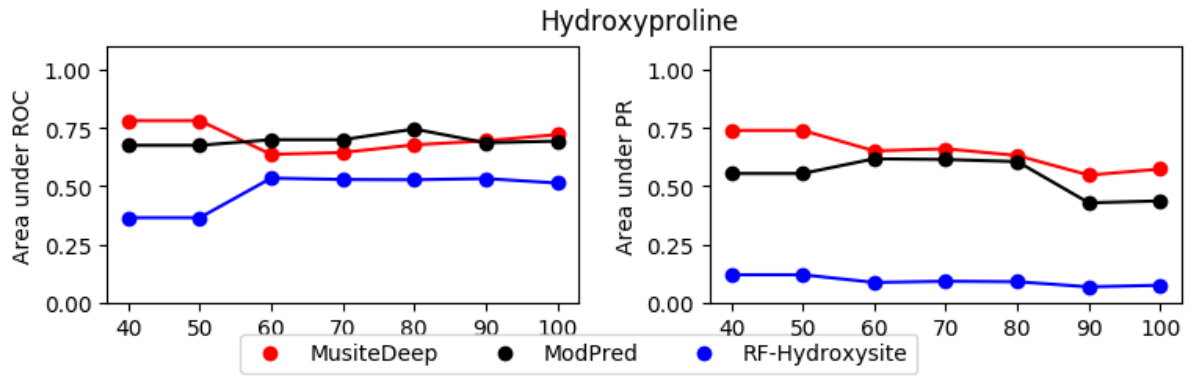


Figure S2. Performance of test subsets with different levels of sequence similarities to the training data evaluated by area under ROC and area under PR. The test subsets that have no more than 40%, 50%, 60%, 70%, 80%, 90% and 100% similarities with the training data were generated by Blastp (2.2.25). The area under ROC and area under PR for each similarity level were shown as dots with colours corresponding to the different methods. For hydroxylysine, no test data remained under the 40% and 50% sequence similarity levels.

REFERENCES

1. Wang, D., et al., *MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction.* Bioinformatics, 2017. **33**(24): p. 3909-3916.
2. Wang, D., Y. Liang, and D. Xu, *Capsule network for protein post-translational modification site prediction.* Bioinformatics, 2019. **35**(14): p. 2386-2394.