W140–W146 *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* *Published online 23 April 2020*
doi: 10.1093/nar/gkaa275

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] and 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 and [3]School of Information Science and Technology, Northeast Normal University, Changchun, Jilin 130117, China

Received February 15, 2020; Revised March 25, 2020; Editorial Decision April 07, 2020; Accepted April 08, 2020

ABSTRACT

MusiteDeep is an online resource providing a deep-learning framework for protein post-translational modification (PTM) site prediction and visualization. The predictor only uses protein sequences as input and no complex features are needed, which results in a real-time prediction for a large number of proteins. It takes less than three minutes to predict for 1000 sequences per PTM type. The output is presented at the amino acid level for the user-selected PTM types. The framework has been benchmarked and has demonstrated competitive performance in PTM site predictions by other researchers. In this webserver, we updated the previous framework by utilizing more advanced ensemble techniques, and providing prediction and visualization for multiple PTMs simultaneously for users to analyze potential PTM cross-talks directly. Besides prediction, users can interactively review the predicted PTM sites in the context of known PTM annotations and protein 3D structures through homology-based search. In addition, the server maintains a local database providing pre-processed PTM annotations from Uniport/Swiss-Prot for users to download. This database will be updated every three months. The MusiteDeep server is available at https://www.musite.net. The stand-alone tools for locally using MusiteDeep are available at https://github.com/duolinwang/MusiteDeep_web.

INTRODUCTION

Protein post-translational modifications (PTMs) generally refer to the additions of various functional groups to amino acid residues (1). PTM is a key mechanism to increase proteomic diversities and plays an important role in the regulation of protein functions (2). Therefore, identifying and understanding PTMs are critical in the studies of biology and diseases. To date, as the accumulation of PTM experimental data, dozens of bioinformatics tools have been developed for PTM site prediction, which provide a fast and low-cost approach in contrast to experimental methods. Many of these tools apply a machine-learning algorithm and provide a prediction for a particular PTM type. Here, we briefly review several representative tools. Musite (3) was proposed by our group and applies a support vector machine (SVM) for phosphorylation site prediction, using the K nearest neighbor (KNN) score, disorder scores and amino acid frequencies as features. GlycoEP (4) used SVM for N-, O- and C-linked glycosylation site prediction. Besides raw sequence features, GlycoEP extracted amino acid composition (AAC) profiles, position-specific scoring matrix (PSSM) profiles, secondary structures, and surface accessibilities as features. GPS-PAIL (5) is a tool to predict HAT-specific lysine acetylation sites, which used their previously proposed GPS2.2 algorithm (6) as well as features from an amino acid substitution matrix (e.g. BLOSUM62) and protein-protein interaction. GPS-SUMO (7) is another GPS algorithm-based tool to predict the sumoylation sites. MePred-RF (8) applied a random forest algorithm with a complex sequence-based feature selection technique to predict methylarginine and methyllysine sites. RF-Hydroxysite (9) applied a random forest algorithm for hydroxylysine and hydroxyproline site prediction that combined information from physicochemical, structural, evolutionary and sequence-order features. Css-Palm4.0 (10) applied a clustering and scoring strategy for palmitoylation site prediction. UbiProber (11) was designed for ubiquitin site prediction that uses KNN, physicochemical property, and AAC features to train an SVM-based predictor. DeepPhos recently applied densely connected convolutional neural network (CNN) blocks for phosphorylation site prediction (12). Because the dysregulation of PTM plays impor-

[*]To whom correspondence should be addressed. Tel: +1 573 808 2219; Fax: +1 573 882 8318; Email: xudong@missouri.edu

tant roles in the development and progression of diseases (13–15), a number of databases were developed to annotate existing or predicted PTM sites with SNPs or diseases, such as dbPTM (16), PTMcode (17) and AWESOME (18). They also provide predictions by assembling third-party tools. For example, the recently developed AWESOME applied 20 available tools for different types of PTM site predictions, and the method described in this paper was included for phosphorylation site prediction.

Despite the availability of these tools and databases, the number of webservers that provide a general prediction for many different PTM types is quite low. PTM-ssMP (19) is one, but it relies on known modification profiles, which cannot make *de-novo* prediction outside the profiles. ModPred (20) provides a general webserver for 23 different modifications. However, it was published in 2014 and applies the logistic regression method, with a performance in need of improvement by more advanced machine learning methods. Most of the other existing predictive webservers were developed for a single type of PTM prediction, such as GlycoEP and MePred-RF. All these servers limit the number of submissions (mostly a single sequence), and none of them supports batch submission for any large-scale prediction, due to calculations of more complex features, such as PSSM, which takes significant computation time.

Here, we introduce a new webserver, MusiteDeep, to provide a general deep-learning framework for protein PTM site prediction and visualization. The method was first introduced in (21). To the best of our knowledge, MusiteDeep was the first deep-learning method in phosphorylation (one of the most studied PTMs) site prediction. It takes raw protein sequences as input and uses CNN (22) with a novel two-dimensional (2D) attention mechanism. It has been widely benchmarked by other predictors (12,23–25) and has always ranked in first or second place. It has also been adopted in several other services (18,26–27). Later, we upgraded the deep-learning framework with the capsule network (28) and extended it to predict for more PTM types, which showed superior performance for almost all the cases with small training samples (29). In this webserver, we updated the previous framework by combining the two previous networks and utilizing more advanced ensemble techniques, while simultaneously providing prediction and visualization for multiple PTMs. The server provides a real-time prediction for a large number of proteins by using only CPU without GPU resources. Users can interactively review predicted PTM sites in the context of known PTM annotations and protein 3D structures through a homology-based search. In addition, the server maintains a local database (updated every three months) to provide pre-processed PTM annotations from Uniport/Swiss-Prot for download. Comparing with the existing web services, MusiteDeep has some obvious advantages in accuracy, speed and scale, and it also provides some unique functions for analyzing prediction results.

MATERIALS AND METHODS

Method overview

Deep-learning framework. The framework of MusiteDeep for protein PTM site prediction is shown in Figure 1. We treated the PTM site prediction problem as a binary classification problem, and for each type of PTM, we trained an independent predictor. In the training process, the framework accepts raw protein sequences as input, and then 33-length residue fragments centered at the target sites were extracted and coded by the one-of-K coding method. We considered the 20 common amino acids and a hyphen character '-', which is used to pad the positions when the valid fragment length is less than 33. Therefore, each position is represented by a 21D vector, with value 1 at the index corresponding to the amino acid or the hyphen character, and a 0 at all other indexes; meanwhile, uncommon amino acids are filled with 0.05. The final output of the framework is a confidence score of the PTM prediction. In regard to the deep-learning architecture, we used the combination of the two previously proposed networks, i.e., MultiCNN (21) and CapsNet (29) as shown in Figure 1 right. The MultiCNN model is constructed by three one-dimensional (1D) CNN layers, a two-dimensional (2D) attention layer, and two fully connected layers; the CapsNet is constructed by two 1D CNN layers, a PrimaryCaps layer and a PTMCaps layer. We trained both networks separately, and in the prediction procedure, a final prediction score is calculated by averaging the prediction scores obtained by the two independent networks. The architecture details can be found in the Supplementary Text S1.

Bootstrapping and weight averaging. Because nearly all PTM types have more negative samples than positive samples, to address the unbalanced issues during training, a bootstrapping technique was applied. As shown in Figure 1 left, the training fragments were partitioned into N subsets, each containing the same number of positive and negative fragments; here, N was determined by the integer part of the negative to positive ratio. The network was trained iteratively, and the number of iterations was set as N, or in practical use, an upper limit (30 by default). In each training iteration, one subset was used to train the network through the Adam stochastic optimization (30) based on mini-batches. After training for N iterations, one classifier was obtained. The early stopping strategy was applied in each iteration, and when the loss of a validation set did not decrease in some number of epochs (one forward and backward pass over the entire subset), the training procedure for that subset would stop. It has been demonstrated that an ensemble model created by averaging the weights from a continuous training procedure leads to wider optima and better generalization (31). Therefore, during the bootstrapping procedure, we applied the weight averaging strategy. We treated each iteration trained on one subset as one training cycle and we saved the weights generated from each iteration denoted as W_i, and their loss on the validation set is denoted as L_i. The final weight W of the classifier can be calculated by the weighted average of the weights from all the iterations as follows:

$$W = \sum_{i=1}^{N} \frac{\exp\left(1/L_i\right)}{\sum_{i=1}^{N} \exp\left(1/L_i\right)} W_i \qquad (1)$$

To further improve the performance, we trained an ensemble of 10 × N_C models using a nested cross-validation. The original training data was divided into 10 equal-sized

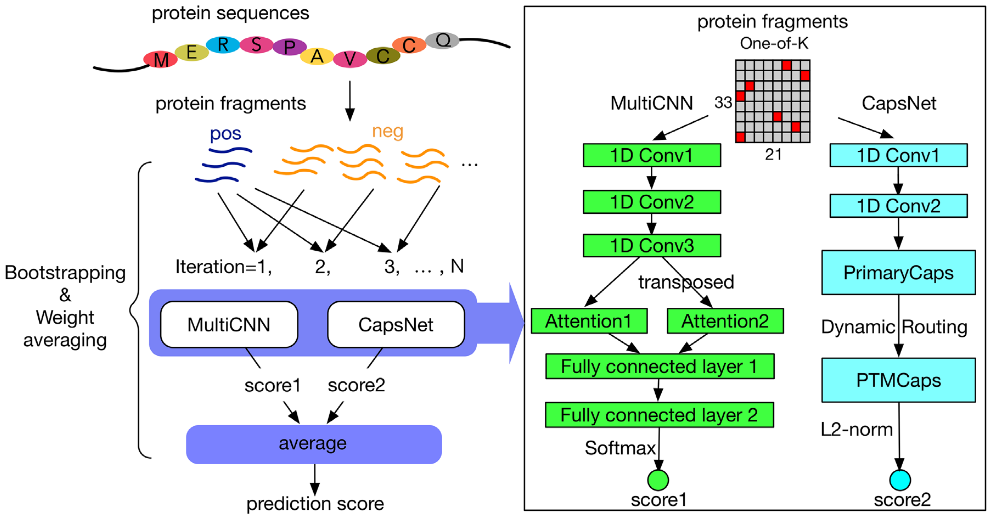


Figure 1. Flowchart of the MusiteDeep framework. pos: positive fragments. neg: negative fragments.

subsets. Each time, nine subsets were used to train a model, and the remaining subset was used as validation to monitor its training process. Ten different training sets were used during the training. In the meantime, on the same training set, we repeated the bootstrapping procedure for N_C times, and N_C independent classifiers were obtained. N_C is a hyper-parameter which can be determined by users, and $N\_C = 2$ was used in this work. The final prediction score was the average of the scores obtained from the $10 \times N\_C$ ensemble models.

Transfer learning. For PTM types that have small training samples, we applied a transfer learning technique to further improve the performance. The lower layers of CNN serve as feature extraction layers, and the extracted features can be generalized to a different dataset. Consider the phosphotyrosine as an example. Because it is catalyzed by different kinase groups with phosphoserine or phosphothreonine, we and other researchers have trained separate models for them. However, they share some common features: For example, the phosphorylation always happened in the disorder region. Therefore, in this case, the transfer learning technique can be applied to use a larger dataset to capture the common features and to use the smaller dataset to expatiate its specified features. In particular, we trained a base network on the phosphoserine and phosphothreonine data; then, we used the pre-trained weights of the base network to initialize the weights for the phosphotyrosine predictor. Finally, we fine-tuned the weights of the predictor using the phosphotyrosine data.

Through the combination of the two architectures and the ensemble techniques, the overall performance has been improved, which is shown in the Supplementary Figure S1 by comparing the previous methods on the 10-fold cross-validation datasets provided in (29).

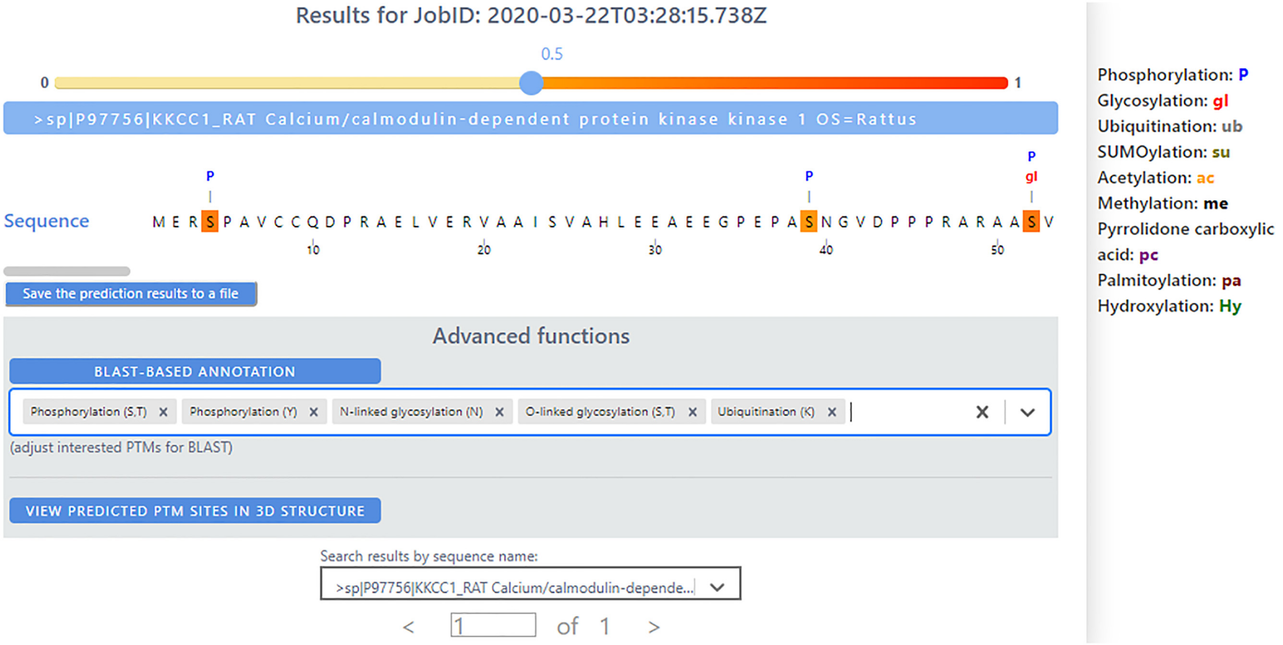
Performance evaluation

Because different tools used different thresholds to present the prediction results, to avoid the effects of distinct threshold values, we used the area under the ROC curves (AUC) and the area under the precision-recall curves to evaluate the performance. Especially, when the negative data is very large, the precision-recall curves are more appropriate to evaluate the performance of a predictor.

RESULTS

MusiteDeep web server

Server inputs and outputs. The input of the server is protein sequences in the FASTA format. The server provides two options for input, paste mode and upload mode. For a small task with up to 10 sequences or 500 amino acids, users can paste their sequences to the input panel, and the job will start immediately and return the result in real time once submitted. For a larger-scale task, users can upload a FASTA file with up to 10 MB. In this mode, once the job is submitted, it will be placed in a queue for processing. One user can process up to five such jobs at the same time. The job can be accessed later by the provided URL or by checking the user's job history. It is important to notice that since we used the browser's local storage to remember the identity of a user, the job history only lists previous jobs that were successfully submitted by the user using the same browser; however, access through the URL has no such restriction. All the jobs will be saved on the server for 72 hours with up to 100MB per user. MusiteDeep does not have a complex feature calculation procedure so it can handle a large number of proteins by only CPU. It takes less than three minutes to predict for 1000 sequences per PTM type. Users can simultaneously select multiple models for prediction from the drop-down list. After a job is finished, the output can be visualized for each input sequence one by one, which can be retrieved by its sequence name or its index number. The output is shown in Figure 2. The predicted PTMs are labeled using their abbreviations on the top of the corresponding positions. Multiple labels are shown on top of one position if that position is predicted to have multiple PTMs. The highlighted colors of the predicted sites corre-



Results for JobID: 2020-03-22T03:28:15.738Z

>sp|P97756|KKCC1_RAT Calcium/calmodulin-dependent protein kinase kinase 1 OS=Rattus

Sequence M E R S P A V C C Q D P R A E L V E R V A A I S V A H L E E A E E G P E P A S N G V D P P P R A R A A S V

0.5
0
1
P
P
P
gl
10
20
30
40
50
Save the prediction results to a file
Advanced functions
BLAST-BASED ANNOTATION
Phosphorylation (S,T)
Phosphorylation (Y)
N-linked glycosylation (N)
O-linked glycosylation (S,T)
Ubiquitination (K)
(adjust interested PTMs for BLAST)
VIEW PREDICTED PTM SITES IN 3D STRUCTURE
Search results by sequence name:
>sp|P97756|KKCC1_RAT Calcium/calmodulin-depende...
< 1 of 1 >
Phosphorylation: P
Glycosylation: gl
Ubiquitination: ub
SUMOylation: su
Acetylation: ac
Methylation: me
Pyrrolidone carboxylic acid: pc
Palmitoylation: pa
Hydroxylation: Hy

Figure 2. An example visualization of the prediction results.

spond to their prediction confidence levels. Upon hovering the mouse on the predicted sites, the detailed information of the prediction will be shown. A user may adjust the prediction confidence threshold by using the slider to obtain more or fewer predicted sites. Besides interactive visualization, the results can be downloaded as plain text, with information of protein identifier, position, residue, PTMscore, and the predicted PTMs whose scores are higher than the user-defined or default cut-off. An example is shown in Supplementary Table S1. In the server, we also provide a REST API for this service along with a template Python program to demonstrate how to use the API.

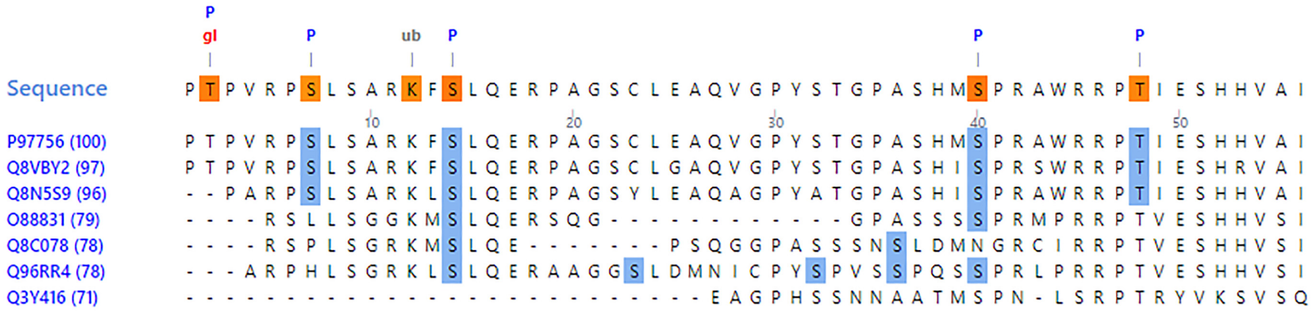
View predicted PTM sites in the context of known PTM annotations. MusiteDeep provides a homology-based search against proteins in UniProtKB/Swiss-Prot, and presents known PTM annotations at the aligned positions, as shown in Figure 3. Protein accession identifier (ACCID), Blast sequence identity, and the known PTM annotations on homologous proteins are presented for each input sequence. Upon hovering the mouse over the colored sites, the specific annotation will be shown. Proteins can be accessed in UniProtKB/Swiss-Prot database by clicking their ACCIDs. In the server, we also provide a REST API for this service along with a template Python program to demonstrate how to use the API.

View predicted PTM sites in the context of protein 3D structure. MusiteDeep provides visualizations of the predicted PTM sites in 3D protein structures by integrating G2S (32), a tool to annotate genomic variants on protein structures, and an NGL viewer (33), which is a web application for molecular visualization. First, a query sequence with the predicted PTM sites is searched by G2S; then, its homologous proteins that have 3D structures in RCSB PDB (34) will be retrieved with the mapping between sequence positions and structural positions. The retrieved information is shown in Figure 4.

The query protein and its predicted PTMs can be viewed in the 3D structure context, as shown in Figure 5. The hover text shows the information of the predicted site, which contains its position on the query sequence, its amino acid types at the query sequence position and at the PDB structural position, and its predicted PTM types (in abbreviation). In this example, multiple PTMs are viewed at the same time, which gives the sense of 3D spatial locations of PTM sites. Users can propose several hypotheses accordingly, such as whether the predicted sites are physically near each other or whether they form a potential PTM cross-talk. i.e. whether these PTMs of multiple residues work together to determine a particular functional outcome. In the example of Figure 5, all the predicted PTM sites locate closely on the structure, which only takes place in the loop region. Since most PTMs tend to occur on the protein surface, we also provide the user an option to add the molecule surface by checking the 'Add surface' box.

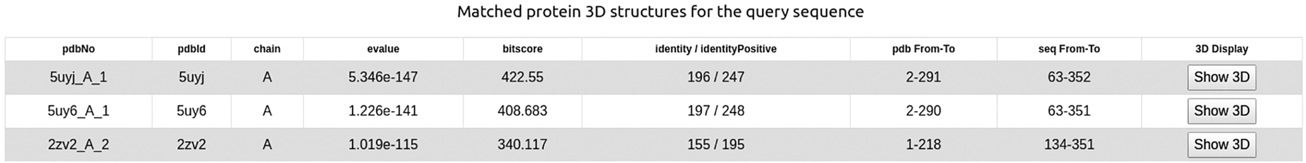
Annotated PTM sequences from UniProtKB/Swiss-Prot. The UniProtKB/Swiss-Prot database contains publicly available and expertly annotated protein sequences (35), including the PTM annotations. MusiteDeep maintains a local database of UniProtKB/Swiss-Prot, which provides preprocessed PTM annotations for users to download and is updated every 3 months.

Web server implementation. The webserver consists of a three-layer architecture of front-end, server-end, and business logic layers. The front-end of the server is implemented with JavaScript libraries, React and jQuery to provide an interactive user interface; the back-end is implemented with



```
                P
                gl        P         ub  P                                                   P               P
                |         |         |   |                                                   |               |
Sequence      P T P V R P S L S A R K F S L Q E R P A G S C L E A Q V G P Y S T G P A S H M S P R A W R R P T I E S H H V A I
                                10                  20                  30                  40                  50
P97756 (100)  P T P V R P S L S A R K F S L Q E R P A G S C L E A Q V G P Y S T G P A S H M S P R A W R R P T I E S H H V A I
Q8VBY2 (97)   P T P V R P S L S A R K F S L Q E R P A G S C L G A Q V G P Y S T G P A S H I S P R S W R R P T I E S H R V A I
Q8N5S9 (96)   - - P A R P S L S A R K L S L Q E R P A G S Y L E A Q A G P Y A T G P A S H I S P R A W R R P T I E S H H V A I
O88831 (79)   - - - - R S L L S G G K M S L Q E R S Q G - - - - - - - - - - - - G P A S S S S P R M P R R P T V E S H H V S I
Q8C078 (78)   - - - - R S P L S G R K M S L Q E - - - - - - - P S Q G G P A S S S N S L D M N G R C I R R P T V E S H H V S I
Q96RR4 (78)   - - - A R P H L S G R K L S L Q E R A A G G S L D M N I C P Y S P V S S P Q S S P R L P R R P T V E S H H V S I
Q3Y416 (71)   - - - - - - - - - - - - - - - - - - - - - - - - - - E A G P H S S N N A A T M S P N - L S R P T R Y V K S V S Q
```

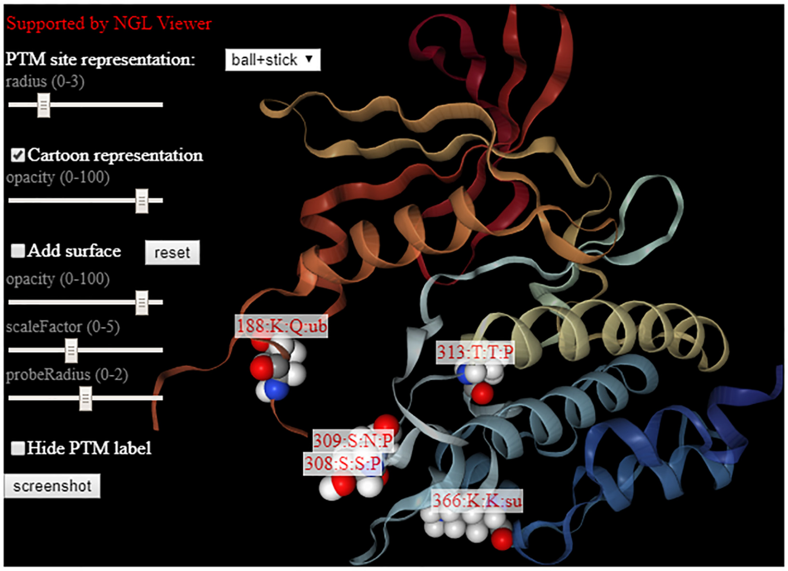
Figure 3. An example output of the Blast-based annotation function.



Matched protein 3D structures for the query sequence

pdbNo	pdbId	chain	evalue	bitscore	identity / identityPositive	pdb From-To	seq From-To	3D Display
5uyj_A_1	5uyj	A	5.346e-147	422.55	196 / 247	2-291	63-352	Show 3D
5uy6_A_1	5uy6	A	1.226e-141	408.683	197 / 248	2-290	63-351	Show 3D
2zv2_A_2	2zv2	A	1.019e-115	340.117	155 / 195	1-218	134-351	Show 3D

Figure 4. An example output retrieved from G2S.



Supported by NGL Viewer

PTM site representation: ball+stick

radius (0-3)

☑ Cartoon representation

opacity (0-100)

□ Add surface reset

opacity (0-100)

scaleFactor (0-5)

probeRadius (0-2)

□ Hide PTM label

screenshot

Figure 5. View predicted PTM sites in the 3D structure context.

Table 1. Performance of MusiteDeep vs. ModPred and a representative method in each PTM

	Area under ROC/area under precision-recall curve		
PTM types	MusiteDeep	ModPred	Other
Phosphoserine/threonine	**0.896/0.329**	0.753/0.134	DeepPhos: 0.809/0.190
Phosphotyrosine	**0.958/0.864**	0.695/0.151	DeepPhos: 0.681/0.163
N-linked glycosylation	**0.993/0.937**	0.774/0.264	GlycoEP: 0.928/0.210
O-lined glycosylation	**0.943/0.539**	0.783/0.128	GlycoEP: 0.808/0.043
N6-acetyllysine	**0.978/0.858**	0.702/0.127	GPS-PAIL: 0.629/0.229
Methylarginine	**0.941/0.844**	0.770/0.130	MePred-RF: 0.681/0.152
Methyllysine	**0.951/0.850**	0.670/0.108	MePred-RF: 0.782/0.514
S-palmitoylation-cysteine	**0.961/0.922**	0.824/0.478	Css-Palm4.0: 0.735/0.465
Pyrrolidone-carboxylic-acid	**0.979/0.947**	0.860/0.578	-
Ubiquitination	**0.804/0.279**	0.584/0.091	UbiProber: 0.651/0.107
SUMOylation	**0.990/0.881**	0.740/0.213	GPS-SUMO: 0.706/0.357
Hydroxylysine	**0.982/0.930**	0.974/0.891	RF-Hydroxysite: 0.919/0.300
Hydroxyproline	**0.732/0.627**	0.694/0.437	RF-Hydroxysite: 0.514/0.075

KOA, a next-generation web framework for node.js; the deep-learning framework was implemented in the business logic layer by Python. We used mongoDB for the server's database. The REST APIs are implemented by KOA and is available at the API section of the server. The stand-alone tools used to run several services on a local machine are available at GitHub (https://github.com/duolinwang/MusiteDeep_web) with detailed documentation.

Independent benchmarking

To demonstrate the performance of the PTM site prediction of MusiteDeep in practical use, we compared MusiteDeep with existing tools by using a timestamp-based dataset. Because each type of PTM may have dozens of predictive tools, to compare with all of them is very challenging. Therefore, besides ModPred, which can cover all of our provided PTM types, we selected the most representative public available tools based on their performance and publication citation for each PTM comparison. The training data was constructed by extracting the protein sequences from UniProtKB/Swiss-Prot before 2010, and the timestamp-based testing data was constructed by the newly released data after 2010. All the annotations generated by computational predictions were removed. For each PTM, all the residues annotated by UniProtKB/Swiss-Prot with the same type of PTM were treated as positive sites, while the residues with the same amino acids excluding the PTM annotations were regarded as the negative sites. The statistics of the data are shown in Supplementary Table S2. The performance of the whole testing data is evaluated by the area under the ROC curves and the area under the precision-recall curves, as shown in Table 1. We also evaluate the performance on test subsets that have different levels of sequence similarities to the training data. The results are shown in Supplementary Figure S2, which shows that most prediction performances hold well with low sequence similarities.

CONCLUSIONS

In this paper, we present MusiteDeep, a web server for protein PTM site prediction and visualization. The method behind the server is a combination of our previously proposed two deep-learning models implemented with two ensemble techniques. Since the predictor does not require the calculation of complex features, the server is capable of providing real-time prediction and batch submission for large-scale protein sequences. The output is presented at the amino acid level for multiple PTMs at the same time. All the submitted jobs will be saved in the server for 72 h with up to 100MB for users to retrieve. Besides prediction, MusiteDeep provides facilities for users to interactively review the predicted PTM sites in the context of known PTM annotations and protein 3D structures. In addition, the server maintains a local database providing pre-processed PTM annotations from Uniport/Swiss-Prot for users to download. Compared with the existing web services, MusiteDeep has some obvious advantages in accuracy, speed and scale. Besides the web server, we provide user-friendly Web APIs to access several services through Python programs and stand-alone tools, which allow users to run MusiteDeep on local machines.

Some limitations of this work include only providing models for PTM types that have enough data in UniProtKB/Swiss-Prot and focusing exclusively on a general modification induced by adding functional groups to the side chain of intermediate residues. In our future work, we plan to extend the framework to more PTM types, including the peptide cleavage and N-terminal PTMs, such as N-terminal acetylation and proteolysis. We will also expand our training data to include more databases on PTMs. The challenge that remains is how to combine the data from different sources and how to control the potential errors. These challenges will be the topics of our future studies.

DATA AVAILABILITY

MusiteDeep is available as a web server at https://www.musite.net. The stand-alone tools to run several MusiteDeep services on a local machine are available in the GitHub repository (https://github.com/duolinwang/MusiteDeep_web).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Institutes of Health [R35-GM126985]. Funding for the open access charge: National Institutes of Health [R35-GM126985].

Conflict of interest statement. None declared.

REFERENCES

1. Knorre,D.G., Kudryashova,N.V. and Godovikova,T.S. (2009) Chemical and functional aspects of posttranslational modification of proteins. *Acta Naturae*, **1**, 29–51.
2. Prabakaran,S., Lippens,G., Steen,H. and Gunawardena,J. (2012) Post-translational modification: nature's escape from genetic imprisonment and the basis for dynamic information encoding. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **4**, 565–583.
3. Gao,J., Thelen,J.J., Dunker,A.K. and Xu,D. (2010) Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol. Cell. Proteomics*, **9**, 2586–2600.
4. Chauhan,J.S., Rao,A. and Raghava,G.P. (2013) In silico platform for prediction of N-, O- and C-glycosites in eukaryotic protein sequences. *PLoS One*, **8**, e67008.
5. Deng,W., Wang,C., Zhang,Y., Xu,Y., Zhang,S., Liu,Z. and Xue,Y. (2016) GPS-PAIL: prediction of lysine acetyltransferase-specific modification sites from protein sequences. *Sci. Rep.*, **6**, 39787.
6. Liu,Z., Yuan,F., Ren,J., Cao,J., Zhou,Y., Yang,Q. and Xue,Y. (2012) GPS-ARM: computational analysis of the APC/C recognition motif by predicting D-boxes and KEN-boxes. *PLoS One*, **7**, e34370.
7. Zhao,Q., Xie,Y., Zheng,Y., Jiang,S., Liu,W., Mu,W., Liu,Z., Zhao,Y., Xue,Y. and Ren,J. (2014) GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. *Nucleic Acids Res.*, **42**, W325–W330.
8. Wei,L., Xing,P., Shi,G., Ji,Z. and Zou,Q. (2019) Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **16**, 1264–1273.
9. Ismail,H.D., Newman,R.H. and Kc,D.B. (2016) RF-Hydroxysite: a random forest based predictor for hydroxylation sites. *Mol. Biosyst.*, **12**, 2427–2435.
10. Ren,J., Wen,L., Gao,X., Jin,C., Xue,Y. and Yao,X. (2008) CSS-Palm 2.0: an updated software for palmitoylation sites prediction. *Protein Eng Des Sel.*, **21**, 639–644.

11. Chen,X., Qiu,J.D., Shi,S.P., Suo,S.B., Huang,S.Y. and Liang,R.P. (2013) Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites. *Bioinformatics*, **29**, 1614–1622.
12. Luo,F., Wang,M., Liu,Y., Zhao,X.M. and Li,A. (2019) DeepPhos: prediction of protein phosphorylation sites with deep learning. *Bioinformatics*, **35**, 2766–2773.
13. Santos,A.L. and Lindner,A.B. (2017) Protein posttranslational modifications: roles in aging and age-related disease. *Oxid Med Cell Longev.*, **2017**, 5716409.
14. Wan,L., Xu,K., Chen,Z., Tang,B. and Jiang,H. (2018) Roles of post-translational modifications in spinocerebellar ataxias. *Front. Cell Neurosci.*, **12**, 290.
15. Wang,Y.C., Peterson,S.E. and Loring,J.F. (2014) Protein post-translational modifications and regulation of pluripotency in human stem cells. *Cell Res.*, **24**, 143–160.
16. Huang,K.Y., Lee,T.Y., Kao,H.J., Ma,C.T., Lee,C.C., Lin,T.H., Chang,W.C. and Huang,H.D. (2019) dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications. *Nucleic Acids Res.*, **47**, D298–D308.
17. Minguez,P., Letunic,I., Parca,L. and Bork,P. (2013) PTMcode: a database of known and predicted functional associations between post-translational modifications in proteins. *Nucleic Acids Res.*, **41**, D306–D311.
18. Yang,Y., Peng,X., Ying,P., Tian,J., Li,J., Ke,J., Zhu,Y., Gong,Y., Zou,D., Yang,N. *et al.* (2019) AWESOME: a database of SNPs that affect protein post-translational modifications. *Nucleic Acids Res.*, **47**, D874–D880.
19. Liu,Y., Wang,M., Xi,J., Luo,F. and Li,A. (2018) PTM-ssMP: A Web Server for Predicting Different Types of Post-translational Modification Sites Using Novel Site-specific Modification Profile. *Int. J. Biol. Sci.*, **14**, 946–956.
20. Pejaver,V., Hsu,W.L., Xin,F., Dunker,A.K., Uversky,V.N. and Radivojac,P. (2014) The structural and functional signatures of proteins that undergo multiple events of post-translational modification. *Protein Sci.*, **23**, 1077–1093.
21. Wang,D., Zeng,S., Xu,C., Qiu,W., Liang,Y., Joshi,T. and Xu,D. (2017) MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics*, **33**, 3909–3916.
22. LeCun,Y., Bengio,Y. and Hinton,G.J.n. (2015) Deep learning. *Nature*, **521**, 436–444.
23. Xu,Y., Song,J., Wilson,C. and Whisstock,J.C. (2018) PhosContext2vec: a distributed representation of residue-level sequence contexts and its application to general and kinase-specific phosphorylation site prediction. *Sci. Rep.*, **8**, 8240.
24. Maiti,S., Hassan,A. and Mitra,P. (2020) Boosting phosphorylation site prediction with sequence feature-based machine learning. *Proteins*. **88**, 284–291.
25. Fenoy,E., Izarzugaza,J.M.G., Jurtz,V., Brunak,S. and Nielsen,M. (2019) A generic deep convolutional neural network framework for prediction of receptor-ligand interactions-NetPhosPan: application to kinase phosphorylation prediction. *Bioinformatics*, **35**, 1098–1107.
26. Yu,K., Zhang,Q., Liu,Z., Zhao,Q., Zhang,X., Wang,Y., Wang,Z.X., Jin,Y., Li,X., Liu,Z.X. *et al.* (2019) qPhos: a database of protein phosphorylation dynamics in humans. *Nucleic Acids Res.*, **47**, D451–D458.
27. López-García,G., Jerez,J.M., Urda,D. and Veredas,F.J. (2019) *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
28. Sabour,S., Frosst,N. and Hinton,G.E. (2017) In: Jordan,MI, LeCun,Y and Solla,SA (eds). *Advances in Neural Information Processing Systems*. pp. 3856–3866.
29. Wang,D., Liang,Y. and Xu,D. (2019) Capsule network for protein post-translational modification site prediction. *Bioinformatics*, **35**, 2386–2394.
30. Kingma,D.P. and Ba,J.L. (2014) Adam: a method for stochastic optimization. arXiv: https://arxiv.org/abs/1412.6980, 22 December 2014, pre-print: not peer reviewed.
31. Izmailov,P., Podoprikhin,D., Garipov,T., Vetrov,D. and Wilson,A.G.J.a.p.a. (2018) Averaging weights leads to wider optima and better generalization. arXiv: https://arxiv.org/abs/1803.05407, 14 March 2018, pre-print: not peer reviewed.
32. Wang,J., Sheridan,R., Sumer,S.O., Schultz,N., Xu,D. and Gao,J. (2018) G2S: a web-service for annotating genomic variants on 3D protein structures. *Bioinformatics*, **34**, 1949–1950.
33. Rose,A.S. and Hildebrand,P.W. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.*, **43**, W576–W579.
34. Burley,S.K., Berman,H.M., Bhikadiya,C., Bi,C., Chen,L., Di Costanzo,L., Christie,C., Dalenberg,K., Duarte,J.M., Dutta,S. *et al.* (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.*, **47**, D464–D474.
35. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M., Bansal,P., Bridge,A.J., Poux,S., Bougueleret,L. and Xenarios,I. (2016) UniProtKB/swiss-prot, the manually annotated section of the uniprot knowledgebase: how to use the entry view. *Methods Mol. Biol.*, **1374**, 23–54.