analytical chemistry

Article

Cite This: Anal. Chem. 2019, 91, 9724–9731

pubs.acs.org/ac

MS/MS Spectrum Prediction for Modified Peptides Using pDeep2 Trained by Transfer Learning

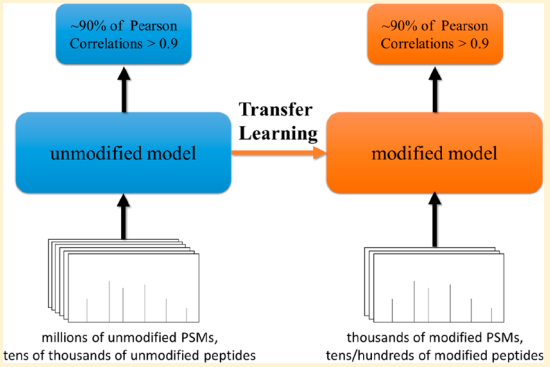
Wen-Feng Zeng,*[,†,‡] Xie-Xuan Zhou,[‡,§] Wen-Jing Zhou,[†,‡] Hao Chi,[†,‡] Jianfeng Zhan,[‡,§] and Si-Min He[†,‡]

[†]Key Laboratory of Intelligent Information Processing and [§]State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences, 100190 Beijing, China

[‡]University of Chinese Academy of Sciences, 100190 Beijing, China

Supporting Information

ABSTRACT: In the past decade, tandem mass spectrometry (MS/MS)-based bottom-up proteomics has become the method of choice for analyzing post-translational modifications (PTMs) in complex mixtures. The key to the identification of the PTM-containing peptides and localization of the PTM-modified residues is to measure the similarities between the theoretical spectra and the experimental ones. An accurate prediction of the theoretical MS/MS spectra of the modified peptides will improve the similarity measurement. Here, we proposed the deep-learning-based pDeep2 model for PTMs. We used the transfer learning technique to train pDeep2, facilitating the training with a limited scale of benchmark PTM data. Using the public synthetic PTM data sets, including the synthetic phosphopeptides and 21 synthetic PTMs from ProteomeTools, we showed that the model trained by transfer learning was accurate (>80% Pearson correlation coefficients were higher than 0.9), and was significantly better than the models trained without transfer learning. We also showed that accurate prediction of the fragment ion intensities of the PTM neutral loss, for example, the phosphoric acid loss (−98 Da) of the phosphopeptide, will improve the discriminating power to distinguish the true phosphorylated residue from its adjacent candidate sites. pDeep2 is available at https://github.com/pFindStudio/pDeep/tree/master/pDeep2.



Most biological processes are regulated by the post-translational modification (PTM) state of proteins. In the past decade, tandem mass spectrometry (MS/MS)-based bottom-up proteomics has become the method of choice for analyzing PTMs in biological samples.[1,2] For the automatic and large-scale analysis of PTMs, identification of the modified peptides and localization of the modified sites rely on comparison of the theoretical spectra and the experimental ones. Hence accurate prediction of the MS/MS spectra of modified peptides is certainly necessary. Many tools, such as PeptideART,[3] OpenMS-Similator,[4] MS2PIP,[5] MassAnalyzer,[6] and MS2PBPI,[7] were developed to predict MS/MS spectra of peptides. Efforts have also been made to make a prediction of modified peptides feasible. For example, MassAnalyzer used the mobile proton properties of the PTM to assist in training the PTM model, and it supported quite a few PTMs;[6] MS2PIP was built based on the boosting method and it required the PTM mass to predict the spectra of modified peptides, and it also supported a few common PTMs.[5] For commons PTMs, it is possible to train an accurate model using traditional machine learning method because we could collect sufficient MS/MS spectra of common PTMs, but it is difficult to train a good model for low-abundant PTMs. Hence, it is worth further developing methods for both common and low-abundant PTMs.

In our previous work of pDeep,[8] it was shown that the deep neural network or deep learning could capture the fragmentation properties of amino acids, and hence significantly improve the accuracies of spectrum prediction for the unmodified peptides. It is reasonable to investigate whether deep learning also works for the spectrum prediction for modified peptides. To train a deep learning model for modified peptides, especially for peptides with low-abundance PTMs, one of the biggest problems is the lack of sufficient benchmark PTM data. The identification of low-abundance PTMs also suffers from high subgroup false discovery rates (FDR).[9] Therefore, for model training, the most ideal data for constructing benchmarks are the synthetic peptides with PTMs. Thanks to the ProteomeTools project,[10] we now have quite a bit of data with various PTMs to train and test the deep learning model. But, nonetheless, the scale of synthetic PTM data is not large enough; we still need new techniques to train accurate PTM models.

Transfer learning is a good choice for training a model with limited benchmark data. As the name implies, transfer learning can transfer the knowledge or learned parameters from one

Received: March 11, 2019
Accepted: June 25, 2019
Published: June 25, 2019

DOI: 10.1021/acs.analchem.9b01262
Anal. Chem. 2019, 91, 9724–9731

domain with sufficient training data to another related domain of interest with limited data. It has been used in many tasks, including classification of skin cancer[11] and diagnosis of eye diseases.[12] Most of the image-based models were first trained using the ImageNet data set (>1 million images)[13] and then fine-tuned by the transfer learning technique using only thousands of disease images. These models could achieve quite high accuracies.

In this paper, we used transfer learning to train the deep learning model for spectrum prediction of modified peptides with limited data based on the pretrained pDeep2 model. Using the public synthetic PTM data sets, including the synthetic phosphopeptides and 21 synthetic PTMs from ProteomeTools, we showed that the accuracies of transfer-learning-based models were better than those without using transfer learning. We showed that, for the phosphorylation analysis, accurate prediction of the fragment ion intensities of the phosphoric acid loss (−98 Da) of the phosphopeptide will be helpful in distinguishing the true phosphorylated site from its adjacent sites.

METHODS

Feature Design for Modified Peptides. In pDeep,[8] each amino acid is represented by a one-hot indicator vector with dimension 20. For a PTM, the most representative feature is its chemical structure, but feature embedding for the complex chemical structure is not well solved yet. Here, we used the chemical composition as a compromise solution to represent a PTM. Most of the common PTMs, like oxidation, phosphorylation, and acetylation, are composed of H, C, N, O, S, and P elements, and a few common PTMs contain metallic elements or other chemical elements; hence, we used a chemical composition vector with dimension 8 to represent common PTMs (see Table 1). For example, the carbamidomethylation often happens on Cys and its chemical composition is $H_3C_2N_1O_1$; hence, its feature vector is [3,2,1,1,0,0,0,0]. The 8-dimension PTM feature is appended to the feature vector of its modified amino acid to generate the feature for the modified amino acid. If there is no PTM on an amino acid, an 8-dimension zero vector is used to represent the "zero PTM", and is also appended to the feature of the unmodified amino acid. It is worth noting that the chemical composition feature cannot be used to represent complex PTMs such as glycosylation because different glycans may share the same chemical composition.

Table 1. 8-Dimension Chemical Composition Feature for the PTM[a]

feature	description
#H	number of hydrogen elements in the PTM
#C	number of carbon elements in the PTM
#N	number of nitrogen elements in the PTM
#O	number of oxygen elements in the PTM
#S	number of sulfur elements in the PTM
#P	number of phosphorus elements in the PTM
reserved	reserved dimension for future use
reserved	reserved dimension for future use

[a]If we need to consider Na, Ca, or other elements, the model is flexible enough to add new dimensions.

Improvement from pDeep to pDeep2. The improved version of pDeep,[8] named as pDeep2, was also built based on two-layer bidirectional long−short-term memory (BiLSTM). But pDeep2 is faster and more flexible than the original version of pDeep because we considered new Tensorflow[14] APIs and new features, as shown below.

(1) The original version of pDeep could predict only the MS/MS spectra of peptides which were not longer than a predefined max length. Here, we used the "dynamic_rnn" API of Tensorflow to process peptides without any length limitation. Besides, for a peptide shorter than the max length, "dynamic_rnn" can also avoid unnecessary calculation on the time step which exceeds the peptide length in the BiLSTM. (2) We considered PTM features in pDeep2, making it possible to predict the MS/MS spectra prediction of modified peptides. (3) We considered the instrument type and collision energy (or normalized collision energy, NCE) in the model, making pDeep2 adaptive for different instruments. We used a one-hot indicator vector to represent different Orbitrap-based instruments. pDeep2 now supports Lumos, Elite, Velos, and Q-Exactive (Q-Exactive related) instruments.

The structure of pDeep2 model was shown in Figure 1a. For a peptide, its amino acid features, PTM features (including the "zero PTM"), precursor charge state, the instrument type, and the collision energy are concatenated and then fed into the Bi-Dy-LSTM layers, and the outputs are the relative intensities of b/y ions with +1 and +2 charge states. The training parameters of the unmodified pDeep2 model were as follows: number of hidden bidirectional dynamic LSTM (Bi-Dy-LSTM) layers = 2; hidden layer size of the LSTM cell = 256; dropout = 0.2; epoch = 100; mini-batch size = 1024; learning rate = 0.001; the loss function was mean absolute error. All models were trained on a Lenovo ThinkStation P310 with an NVIDIA TITAN Xp GPU (12 GB graphical memories).

Transfer Learning for Modified Peptides. Transfer learning is very suitable to train the PTM model because, for a modified peptide, there is only one or a few amino acids modified by PTM, while the majority remain unmodified. The fragmentation properties of the unmodified amino acids have already been well learned from large-scale unmodified PSMs; their properties or knowledge can be transferred to train the PTM model without the requirement of too many benchmark PTM data. The training process is quite simple to implement in deep learning. We first trained an accurate pDeep2 model based on the data sets of unmodified PSMs, which was called the pretrained model or unmodified model. Then we used the transfer learning technique to train PTM models based on the pretrained model, as shown in Figure 1c. When transfer learning is used, the learned parameters of the first Bi-Dy-LSTM layer were fine-tuned to fit the new input PTM features, and the output layer was fine-tuned to fit the new outputs, while intermediate hidden layers were frozen (Figure 1c). The training parameters for transfer learning were as follows: epoch = 20; mini-batch size = 1024; learning rate = 0.001; the loss function was mean absolute error.

In the PTM model, pDeep2 can also consider b/y ions with neutral losses (NLs) of PTMs (Figure 1b), which may occur at phosphorylation on Ser and Thr (−98 Da NL mass), oxidation on Met (−64 Da NL mass), or other PTMs. To predict the intensities of PTM NL fragment ions using transfer learning, we first appended additional virtual PTM NL fragment ions of all b/y backbones into the output vector in the pretrained model (see Figure 1b), and initialized their intensities as zeros. And then, from each spectrum in the PTM data sets, we extracted the intensities of the PTM NL fragment ions of the PTM-containing

DOI: 10.1021/acs.analchem.9b01262
Anal. Chem. 2019, 91, 9724−9731

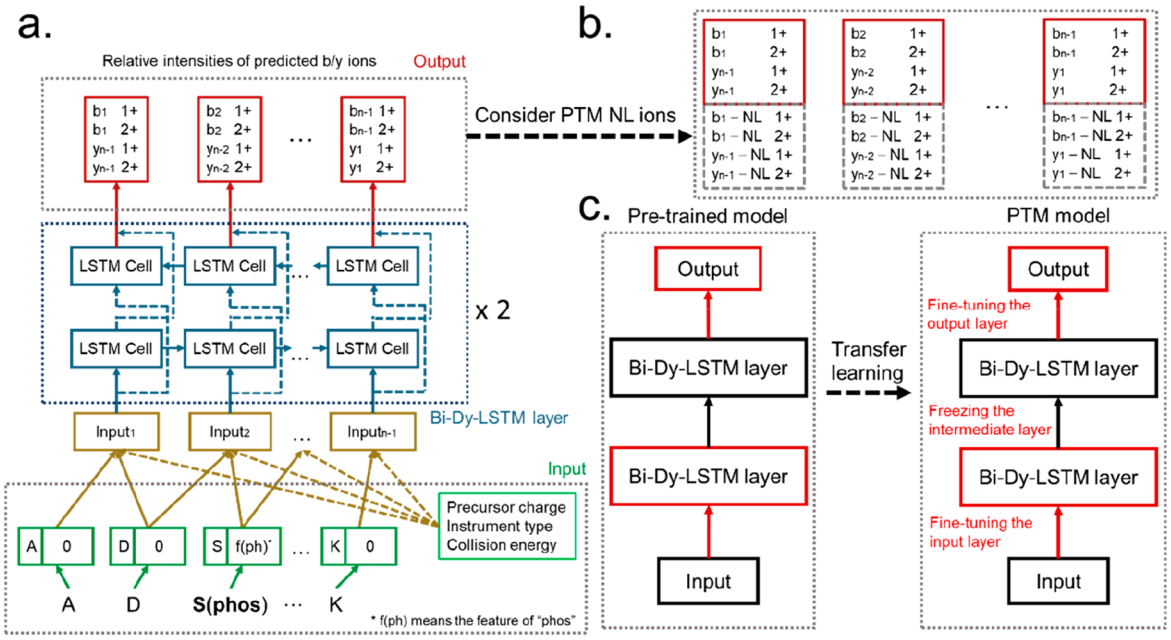


Figure 1. PDeep2 model and transfer learning. (a) pDeep2 consists of the input layer, the output layer, and two bidirectional dynamic LSTM (Bi-Dy-LSTM) layers. The input layer concatenates features of amino acids, PTMs, the precursor charge state, the instrument type, and the collision energy into a numeric tensor. The output layer is a time-distributed fully connected layer which maps the output of the Bi-Dy-LSTM layers onto the relative intensities of fragment ions. (b) To consider PTM neutral loss (NL) ions, we additionally add b/y-NL into the output layer. (c) The transfer learning in pDeep2. When using transfer learning, only the parameters of the first Bi-Dy-LSTM layer and the output layer are fine-tuned.

Table 2. HCD Data Sets Used To Train and Test the Pretrained Unmodified Model[a]

data description	lab	instrument	NCE	no. of PSMs	usage	$P_{PCC>0.75}$	$P_{PCC>0.90}$	median PCC
mouse brain[16]	Mann	QEHF	27	221 106	train	98.8%	94.0%	0.988
fission yeast[17]	Mann	QE	25	118 950	train	96.8%	86.1%	0.977
HEK-293T[18]	Gygi	QE	25	244 177	train	95.0%	88.5%	0.986
HPM[b]: adult CD4T cells[19]	Pandey	Elite	32	77 901	train	97.8%	92.1%	0.989
HPM: adult CD4T cells[19]	Pandey	Velos	41	41 465	train	97.3%	91.3%	0.991
HPM: adult CD4T cells (gel)[19]	Pandey	Velos	41	56 172	train	96.9%	89.6%	0.989
HPM: adult lung[19]	Pandey	Velos	39	29 493	train	98.8%	94.7%	0.992
ProteomeTools[20]	Kuster	Lumos	25	687 865	train	98.8%	95.8%	0.993
ProteomeTools[20]	Kuster	Lumos	30	867 860	train	99.2%	96.0%	0.992
ProteomeTools[20]	Kuster	Lumos	35	652 251	train	99.2%	96.1%	0.993
HeLa (46 fraction)[21]	Olsen	QE	28	280 353	test	97.1%	85.5%	0.972
HEK-293T[21]	Olsen	QE	28	295 693	test	97.1%	84.9%	0.971
HPM: milk[22]	Kuster	Velos	40	28 973	test	98.9%	94.0%	0.986
HPM: rectum[22]	Kuster	Velos	30	86 120	test	98.9%	91.7%	0.974
HPM: stomach[22]	Kuster	Elite	40	120 380	test	98.9%	90.7%	0.973
HPM: fetal gut[19]	Pandey	Velos	35	85 659	test	95.3%	87.5%	0.978
HPM: fetal brain[19]	Pandey	Velos	39	93 631	test	97.0%	90.6%	0.985
ProteomeTools[20]	Kuster	Lumos	25	1 217 654	test	92.4%	85.4%	0.985
ProteomeTools[20]	Kuster	Lumos	30	1 075 764	test	99.3%	96.3%	0.992
ProteomeTools[20]	Kuster	Lumos	35	1 056 958	test	99.1%	95.0%	0.989
HeLa (chymotrypsin)[21]	Olsen	QE	28	183 114	test	95.5%	79.1%	0.958
HeLa (Glu-C)[21]	Olsen	QE	28	158 453	test	93.9%	73.6%	0.949
HeLa (Lys-C)[21]	Olsen	QE	28	236 652	test	95.9%	81.7%	0.967

[a]PCCs between experimental and predicted spectra were also listed. [b]"HPM": human proteome map.

b/y backbones for training. If there were no NLs on a PTM, the intensities of all PTM NL fragment ions will be zero.

In this paper, we used the Pearson correlation coefficient (PCC) as a similarity metric to compare the predicted spectrum with the experimental one. We used the $P(\mathrm{PCC} > x)$ or $P_{\mathrm{PCC}>x}$ as a criterion to evaluate the performance of predictions. $P_{\mathrm{PCC}>x}$ refers to the proportion of PCCs that is greater than a given value x. For example, $P_{\mathrm{PCC}>0.75}$ = 95% means there are 95% PCCs greater than 0.75.

RESULTS

Pretrained Model Preparation for Transfer Learning. The unmodified pDeep2 model was trained and tested on a total of ~8 000 000 high-quality unmodified peptide-spectrum

DOI: 10.1021/acs.analchem.9b01262
Anal. Chem. 2019, 91, 9724−9731

Table 3. Data Sets of Synthetic Peptides with PTMs

data ID	instrument	PTM type	PTM site	NCE	no. of PSMs
Acetyl@K	Lumos	acetylation	K	25,30,35	15 207
Biotinyl@K	Lumos	biotinylation	K	25,30,35	7362
Butyryl@K	Lumos	butyrylation	K	25,30,35	10 781
Crotonyl@K	Lumos	crotonylation	K	25,30,35	16 361
Dimethyl@K	Lumos	dimethylation	K	25,30,35	11 251
Formyl@K	Lumos	formylation	K	25,30,35	15 190
Glutaryl@K	Lumos	glutarylation	K	25,30,35	15 166
GlyGlycyl@K	Lumos	glyglycylation (digested ubiquitin)	K	25,30,35	15 243
Hydroxyisobutyryl@K	Lumos	hydroxyisobutyrylation	K	25,30,35	13 546
Malonyl@K	Lumos	malonylation	K	25,30,35	5 613
Methyl@K	Lumos	methylation	K	25,30,35	14 709
Propionyl@K	Lumos	propionylation	K	25,30,35	16 246
Succinyl@K	Lumos	succinylation	K	25,30,35	14 335
Trimethyl@K	Lumos	trimethylation	K	25,30,35	8894
Citrullin@R	Lumos	citrullination	R	25,30,35	7944
Dimethyl-asym@R	Lumos	asymmetrically dimethylation	R	25,30,35	9733
Dimethyl-sym@R	Lumos	symmetrically dimethylation	R	25,30,35	9091
Methyl@R	Lumos	monomethylation	R	25,30,35	10 165
Nitrotyr@Y	Lumos	nitration	Y	25,30,35	16 152
Phospho@Y	Lumos	phosphorylation	Y	25,30,35	37 692
Hydroxy@P	Lumos	hydroxylation	P	25,30,35	8029
PhosVelos	Velos	phosphorylation	S,T,Y	HCD-40	48 109

matches (PSMs) from various laboratories and instruments; the data set information is shown in Table 2. All these data were searched using pFind3[15] with the open-search mode (tolerance = ±20 ppm for both precursors and fragments), and only unmodified PSMs were kept at 0.01% FDR. The PSM in ProteomeTools was further filtered out once the peptide was not consistent with synthesizing templates. For peptides in ProteomeTools, we randomly split them into train and test parts according to different RAWs (no. of training RAWs/no. of testing RAWs = 4:6).

The test results (usage = "test" in Table 2) of pDeep2 were quite good; the $P_{PCC>0.75}$ was higher than 90%, $P_{PCC>0.90}$ was higher than 80%, and the $P_{PCC>0.75}$ and $P_{PCC>0.90}$ in the training data sets were very close to those in the testing data sets, demonstrating that the model was well trained.

pDeep2 Model for Common PTMs. The pDeep2 model is flexible enough to be extended to predict the MS2 spectra of peptides with common PTMs. From data sets of ordinary MS runs, we can collect quite a bit of PSMs with common PTMs such as oxidation on Met, deamidation on Asn, formylation at the N-terminus of peptides, and pyro-Glu on Gln at the N-terminus of peptides. Here, the sources of training and testing data for common PTMs are the same as the data sets described in Table 2. Test results are shown in the Supporting Information. Highest accuracies were often achieved when transfer learning was used in pDeep2 for common PTMs. When we had sufficient common PTM data for training, we could obtain quite a good model without using transfer learning. But training an accurate model for uncommon PTMs-containing peptides is difficult because of the limited data scale; hence, transfer learning is necessary.

Performance of Transfer Learning for 21 Synthetic PTMs. We used two published benchmark PTM data sets to further train and test the transfer learning model of pDeep2: 21 synthesized PTMs from ProteomeTools (data generated by Lumos)[10] and synthesized phosphopeptides (data generated by Velos),[23] as shown in Table 3. RAW data files were searched using pFind3[15] with the restricted search mode, both of the precursor and fragment tolerances were set as ±20 ppm, and the false discovery rate (FDR) was set as 0.1% at the peptide level. The protein sequence database was generated by concatenating the sequences of synthetic template peptides and sequences from the SwissProt human database. The synthetic peptide templates could be downloaded from PRIDE with ID PXD009449 and PXD000138. The variable modifications were carbamidomethylation on Cys and oxidation on Met, and other variable modifications were set on the basis of the synthesized PTMs. After searching, only the identified PSMs whose peptides were consistent with the synthetic templates were kept for further training and testing. Furthermore, to ensure the matching quality of all PTM spectra, a PSM would be removed if the number of matched peaks was less than its peptide length. Finally, for each PTM data set from ProteomeTools, different peptides were divided into training and testing parts in the proportion of 8:2. In the PhosVelos data set with 96 RAW HCD files, we used PSMs from 77 RAW files for training and PSMs from the remaining 19 RAW files for testing.

We first tested the performance of transfer learning for GlyGlycyl@K and Methyl@K, on which there are no PTM NLs (except for the water loss and ammonia loss). In the test data set of GlyGlycyl@K@HCD30, the test $P_{PCC>0.90}$ of the pretrained model was 52.8% (the curve of "pretrained" in Figure 2a). After transfer learning was used, the test $P_{PCC>0.75}$ and $P_{PCC>0.90}$ on GlyGlycyl@K@HCD30 achieved 99.3% and 90.6%, respectively (the curve of "transfer" in Figure 2a), while the test $P_{PCC>0.90}$ was only 62.3% without transfer learning (the curve of "no transfer" in Figure 2a). Analyses of Methyl@K@HCD30 showed similar results (Figure 2b). It is worth mentioning that, for the pretrained model, the predicted intensities of PTM NL fragment ions were always zeros. Therefore, to ensure the fairness of comparisons among "transfer", "no transfer", and "pretrained" models, PTM NL fragment ions were not considered in the model comparisons.

DOI: 10.1021/acs.analchem.9b01262
Anal. Chem. 2019, 91, 9724−9731

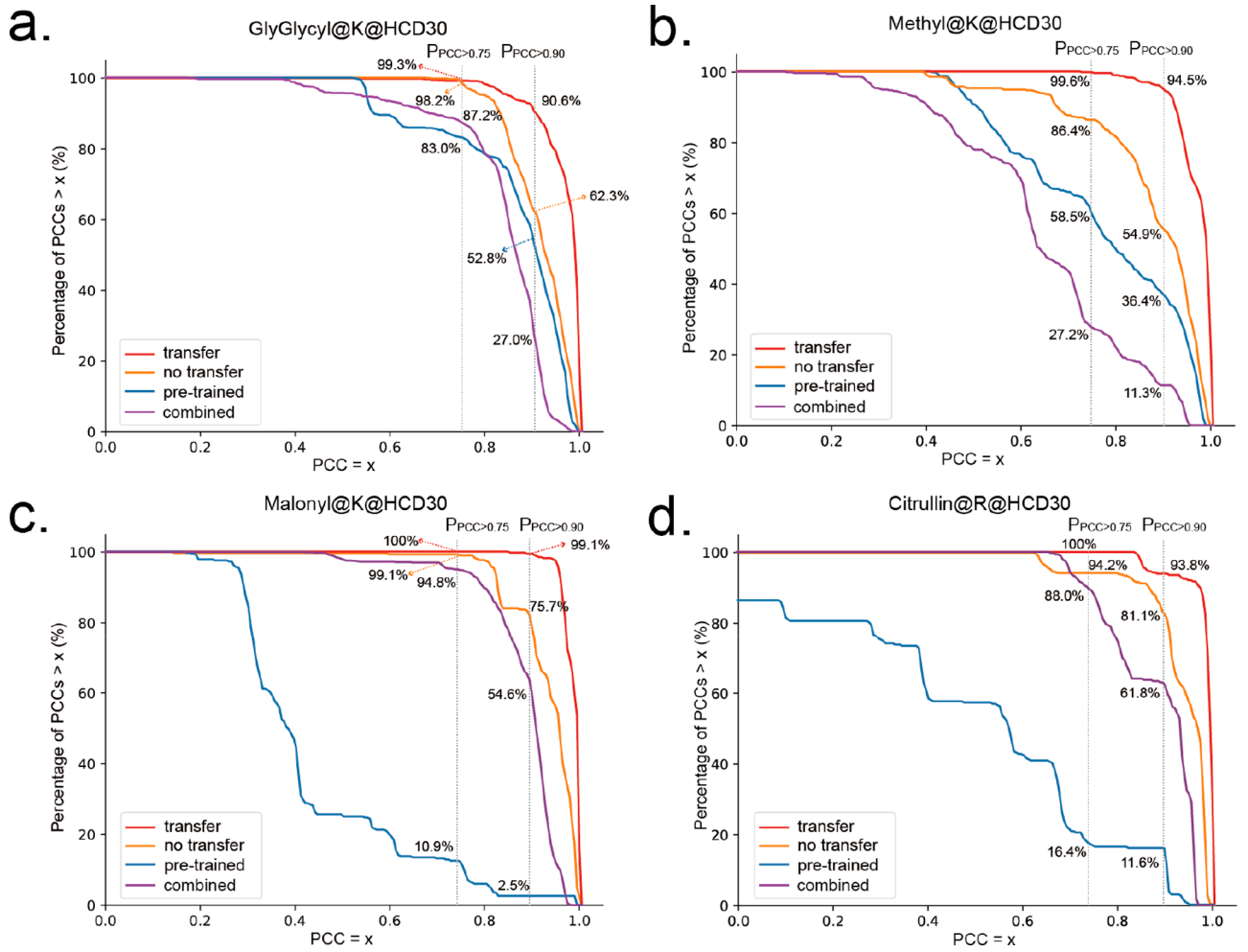


Figure 2. Test results on four selected PTM data sets. (a) Test results on GlyGlycyl@K at HCD30. (b) Test results on Methyl@K at HCD30. (c) Test results on Malonyl@K at HCD30. (d) Test results on Citrullin@R at HCD30. Here, "transfer" means the model was trained with transfer learning, "no transfer" means the model was trained by using only modified data without transfer learning, "pretrained" means predicting the PTM data directly using the pretrained unmodified model, and "combined" means the model was trained by combining the unmodified and modified data.

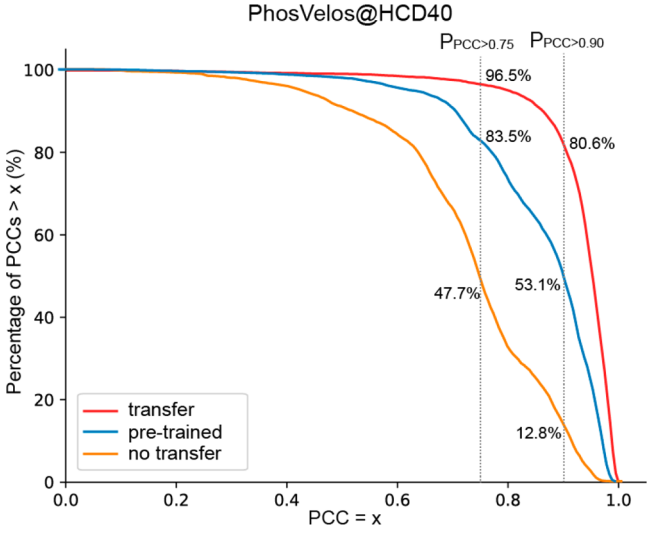


Figure 3. Test results on the PhosVelos data set. The meanings of "transfer", "pretrained", and "no transfer" are similar to those in Figure 2.

We then analyzed the performance of transfer learning for the Malonyl@K and Citrullin@R data sets. Unlike GlyGlycyl@K and Methyl@K, Malonyl@K has a 43.9393 Da NL and Citrullin@R has a 43.0058 Da NL under the HCD-MS/MS.[10] In the test data set of Malonyl@K@HCD30, if we used the pretrained unmodified model to predict the intensities of the b/y backbone ions, the $P_{PCC>0.75}$ and $P_{PCC>0.90}$ would be as low as 10.9% and 3.4%, as shown in Figure 2c. The reasons for the poor performance of the pretrained model may be that the NL of a PTM will make the fragmentation of the modified peptide quite different from that of its unmodified form. After transfer learning is used, the $P_{PCC>0.75}$ and $P_{PCC>0.90}$ increased to 100% and 99.1%, respectively (see Figure 2c). We also showed that if we trained the PTM model without transfer learning, the test $P_{PCC>0.90}$ dropped down to 75.7%. The test results on Citrullin@R@HCD30 were quite similar (Figure 2d). Without transfer learning, the test $P_{PCC>0.90}$ on Citrullin@R@HCD30 was only 81.1%, which was lower than the $P_{PCC>0.90}$ of the model trained with transfer learning. Test results of other synthetic PTMs were shown in the Supporting Information. We also trained the model by combining unmodified and modified PSMs ("combined" in Figure 2); the performance was also not that high compared to that with transfer learning. From all the test results, we can see that for PTMs with or without NLs, transfer learning can always improve the prediction accuracies. We also tried different transfer learning methods, including tuning the first layer ("tune-

DOI: 10.1021/acs.analchem.9b01262
Anal. Chem. 2019, 91, 9724−9731

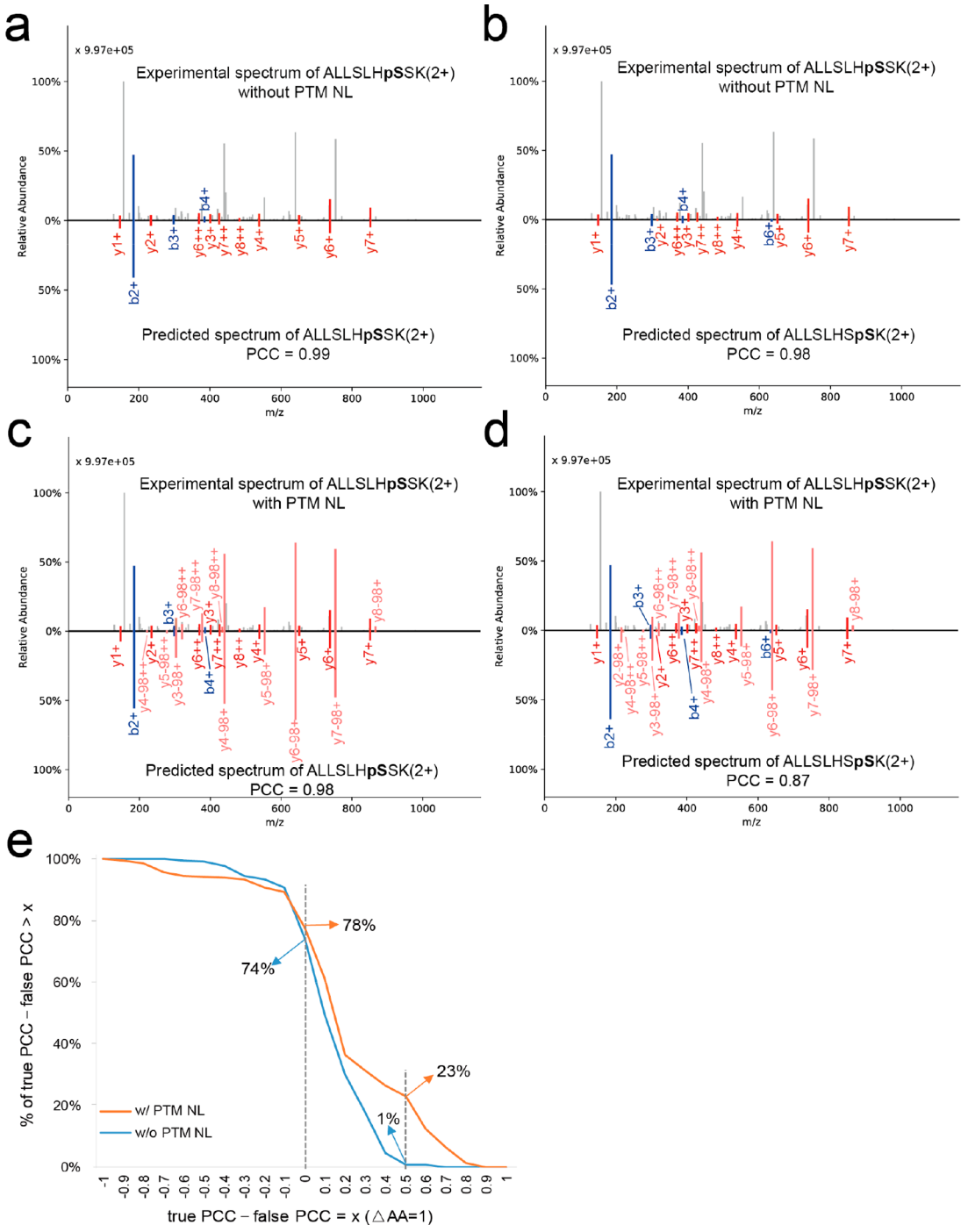


Figure 4. Prediction of PTM NL ions for phosphorylation site localization. The experimental spectrum in (a), (b), (c), and (d) was generated from the synthetic phosphopeptide "ALLSLHpSSK (2+)". (a) PCC = 0.99 for the predicted spectrum of "ALLSLHpSSK (2+)" without predicting the PTM NL ions. (b) PCC = 0.98 for the predicted spectrum of "ALLSLHSpSK (2+)" without predicting the PTM NL ions. (c) PCC = 0.98 for the predicted spectrum of "ALLSLHpSSK (2+)" with predicting the PTM NL ions. (d) PCC = 0.87 for the predicted spectrum of "ALLSLHSpSK (2+)" with predicting the PTM NL ions. (e) Predicting PTM NL ions improve the discriminating power for distinguishing true phosphorylation sites from false sites. "ΔAA = 1" means the false candidate phosphorylation sites are adjacent to the true sites. Here, a total of 283 true and false phosphopeptide pairs with "ΔAA = 1" were statistically analyzed. "true PCC" means the PCC calculated from the true phosphorylation site, and "false PCC" means the PCC calculated from the false site.

first"), tuning the last layer ("tune-last"), tuning the first and the last layers ("tune-first-last", used in this work), tuning all layers ("tune-all"), and in-turns transfer learning (Figure S1); results are shown in the Supporting Information. The results showed that almost all these transfer learning methods had similar performance, except for "tune-last"; it might be because we introduced a PTM feature for each PTM, but the weights of the input layer that connected to the PTM features could not be well trained by tuning only other layers, showing that fine-tuning the input layer using PTM data is very necessary.

To test the impact of different data scales on transfer learning, we also divided each PTM data set into training and testing parts

DOI: 10.1021/acs.analchem.9b01262
Anal. Chem. 2019, 91, 9724−9731

in the proportions of 5:5 and 2:8, and the results are shown in Supporting Information. The results showed that even when we used only 20% data to train the model, transfer learning could also obtain quite acceptable predictions on test data sets. It also showed that the more training data we used, the more accurate the model would be.

Performance of Transfer Learning for SyntPetic Phosphopeptides. Phosphorylation is one of the most important PTMs in life activities, so an accurate prediction of phosphopeptides will be very useful. We used synthetic phosphopeptides to train and test our phosphorylation model, and the synthetic phosphorylated amino acids are Ser, Thr, and Tyr (the PhosVelos data set in Table 3). When we trained the model without transfer learning, the $P_{PCC>0.90}$ was only 12.8%, as shown in Figure 3. Transfer learning could still significantly improve the accuracies of the PTM model. The test $P_{PCC>0.90}$ was improved to 80.6% and test $P_{PCC>0.75}$ was improved to 96.5% after transfer learning was used, as shown in Figure 3.

Importance of Predicting PTM NL Fragment Ions of Phosphopeptides. Unlike the water loss and ammonia loss, b/y ions with the phosphoric acid loss (−98 Da) from the phosphoserine (pS)- or phosphothreonine (pT)-containing peptides will reach very high intensities in the spectrum because the phosphate group is often the most labile part in CID- or HCD-MS/MS.[24,25] It has been shown that considering the b/y−98 Da fragments of phosphopeptides would be helpful to localize the phosphorylation sites.[26] So predicting the intensities of b/y ions with phosphoric acid loss will be helpful as well.

In Figure 4, we used a synthesized phosphopeptide "ALLSLHpSSK" from the test PhosVelos data set to show the importance of the intensities of the phosphoric acid loss for distinguishing the true phosphorylation site from the false sites. On "ALLSLHpSSK", besides the ground truth seventh phosphorylation site (pS7), there are two more candidate phosphorylation sites in this sequence, pS4 and pS8. Distinguishing pS7 from pS8 is a very difficult task because when just the ion mass or m/z information is used, only the y2 and b7 ions can be used as the signatures to distinguish pS7 from pS8 (here, we will use pS7 and pS8 to represent "ALLSLHpSSK" and "ALLSLHSpSK" for simplification). However, the b7 ion may be hardly detected in the HCD-MS/MS, so the y2 ion becomes the unique signature. Without prediction of the intensities of phosphoric acid loss ions, the predicted spectra of pS7 and pS8 were both very similar to the experimental one (PCC = 0.99 vs PCC = 0.98, Figure 4a,b). When the b/y−98 Da fragment ions were considered, differences between pS7 and pS8 could be obviously observed (PCC = 0.98 vs PCC = 0.87, Figure 4c,d). The predicted intensities of y4−98+, y5−98+, y6−98+, and y7−98+ of pS8 were significantly lower than those of pS7, resulting in a lower PCC for pS8. Furthermore, we also performed statistical analysis for the phosphopeptides which contained at least two adjacent candidate phosphorylation sites, as shown in Figure 4e. When PTM NL ions were not considered, the true PCCs (PCCs of true or synthesized phosphorylation sites) were just slightly better than the false PCCs (PCCs of the false phosphorylation sites), there were only ∼1% ΔPCCs (ΔPCC refers to the true PCC minus the false PCC) were ≥0.5. This percentage was increased to ∼23% after the PTM NL ions were considered. These analyses showed that the predicted intensities of PTM NL ions, as well as b/y backbone ions, may provide complementary information to assist in localizing the PTM sites.

CONCLUSIONS

In our previous work, we used deep learning to build the pDeep model to predict the spectra of unmodified peptides. And in this work, we showed that transfer learning can be used to build an accurate model for predicting the spectra of peptides with common PTMs or low-abundance PTMs, even if we only had a limited scale of benchmark modified PSMs. As the spectra with PTM neutral losses of phosphopeptides could be accurately predicted, we demonstrated that pDeep2 will assist in determining the modified sites of phosphorylations. We believe the accurate prediction of spectra for modified peptides will lead to new methods for PTM site localization.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.9b01262.

In-turns transfer learning (PDF)

pDeep2 for common PTMs (XLSX)

Transfer vs no transfer vs pretrained vs combined for PT21-PTMs (XLSX)

Comparison of different transfer learning methods (XLSX)

Transfer learning with different data scales using PT21-PTMs (XLSX)

AUTHOR INFORMATION

Corresponding Author

*E-mail: zengwenfeng@ict.ac.cn (W.-F.Z.).

ORCID

Wen-Feng Zeng: 0000-0003-4325-2147

Wen-Jing Zhou: 0000-0002-5154-6156

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank all pfinders (http://pfind.ict.ac.cn/members.html) for downloading the raw data sets from PRIDE and analyzing these raw data sets using pFind3. This work was supported in part by the National Key Research and Development Program of China (Grant 2016YFA0501301 to S.-M.H.) and the ICT Innovation Program (Grant Y806111000 to S.-M.H.).

REFERENCES

(1) Abu-Farha, M.; Elisma, F.; Zhou, H.; Tian, R.; Zhou, H.; Asmer, M. S.; Figeys, D. *Anal. Chem.* **2009**, *81* (12), 4585−4599.

(2) Gruber, W.; Scheidt, T.; Aberger, F.; Huber, C. G. *Cell Commun. Signal* **2017**, *15* (1), 12.

(3) Li, S.; Arnold, R. J.; Tang, H.; Radivojac, P. *Anal. Chem.* **2011**, *83* (3), 790−796.

(4) Wang, Y.; Yang, F.; Wu, P.; Bu, D.; Sun, S. *BMC Bioinformatics* **2015**, *16*, 110.

(5) Degroeve, S.; Maddelein, D.; Martens, L. *Nucleic Acids Res.* **2015**, *43* (W1), W326−330.

(6) Zhang, Z. *Anal. Chem.* **2011**, *83* (22), 8642−8651.

(7) Dong, N. P.; Liang, Y. Z.; Xu, Q. S.; Mok, D. K.; Yi, L. Z.; Lu, H. M.; He, M.; Fan, W. *Anal. Chem.* **2014**, *86* (15), 7446−7454.

(8) Zhou, X. X.; Zeng, W. F.; Chi, H.; Luo, C.; Liu, C.; Zhan, J.; He, S. M.; Zhang, Z. *Anal. Chem.* **2017**, *89* (23), 12690−12697.

(9) Fu, Y.; Qian, X. *Mol. Cell. Proteomics* **2014**, *13* (5), 1359−1368.

DOI: 10.1021/acs.analchem.9b01262
Anal. Chem. 2019, 91, 9724−9731

(10) Zolg, D. P.; Wilhelm, M.; Schmidt, T.; Medard, G.; Zerweck, J.; Knaute, T.; Wenschuh, H.; Reimer, U.; Schnatbaum, K.; Kuster, B. *Mol. Cell. Proteomics* **2018**, *17* (9), 1850−1863.
(11) Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J.; Swetter, S. M.; Blau, H. M.; Thrun, S. *Nature* **2017**, *542* (7639), 115−118.
(12) Kermany, D. S.; Goldbaum, M.; Cai, W.; Valentim, C. C. S.; Liang, H.; Baxter, S. L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. *Cell* **2018**, *172* (5), 1122−1131 e9 .
(13) Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F.-F. ImageNet: A Large-Scale Hierarchical Image Database. *CVPR* 2009.
(14) Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. *OSDI* **2016**, *12*, 265−283.
(15) Chi, H.; Liu, C.; Yang, H.; Zeng, W. F.; Wu, L.; Zhou, W. J.; Wang, R. M.; Niu, X. N.; Ding, Y. H.; Zhang, Y.; et al. *Nat. Biotechnol.* **2018**, *36*, 1059−1061.
(16) Sharma, K.; Schmitt, S.; Bergner, C. G.; Tyanova, S.; Kannaiyan, N.; Manrique-Hoyos, N.; Kongi, K.; Cantuti, L.; Hanisch, U. K.; Philips, M. A.; et al. *Nat. Neurosci.* **2015**, *18* (12), 1819−1831.
(17) Kulak, N. A.; Pichler, G.; Paron, I.; Nagaraj, N.; Mann, M. *Nat. Methods* **2014**, *11* (3), 319−324.
(18) Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; Gygi, S. P. *Nat. Biotechnol.* **2015**, *33* (7), 743−749.
(19) Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; et al. *Nature* **2014**, *509* (7502), 575−581.
(20) Zolg, D. P.; Wilhelm, M.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Delanghe, B.; Bailey, D. J.; Gessulat, S.; Ehrlich, H. C.; Weininger, M.; et al. *Nat. Methods* **2017**, *14* (3), 259−262.
(21) Bekker-Jensen, D. B.; Kelstrup, C. D.; Batth, T. S.; Larsen, S. C.; Haldrup, C.; Bramsen, J. B.; Sorensen, K. D.; Hoyer, S.; Orntoft, T. F.; Andersen, C. L.; et al. *Cell Syst.* **2017**, *4* (6), 587−599 e4 .
(22) Wilhelm, M.; Schlegl, J.; Hahne, H.; Gholami, A. M.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; et al. *Nature* **2014**, *509* (7502), 582−587.
(23) Marx, H.; Lemeer, S.; Schliep, J. E.; Matheron, L.; Mohammed, S.; Cox, J.; Mann, M.; Heck, A. J.; Kuster, B. *Nat. Biotechnol.* **2013**, *31* (6), 557−564.
(24) Potel, C. M.; Lemeer, S.; Heck, A. J. R. *Anal. Chem.* **2019**, *91* (1), 126−141.
(25) Ulintz, P. J.; Yocum, A. K.; Bodenmiller, B.; Aebersold, R.; Andrews, P. C.; Nesvizhskii, A. I. *J. Proteome Res.* **2009**, *8* (2), 887−899.
(26) Yang, H.; Chi, H.; Zhou, W. J.; Zeng, W. F.; Liu, C.; Wang, R. M.; Wang, Z. W.; Niu, X. N.; Chen, Z. L.; He, S. M. *J. Proteome Res.* **2018**, *17* (1), 119−128.

DOI: 10.1021/acs.analchem.9b01262
Anal. Chem. 2019, 91, 9724−9731