Deep-Learning-Derived Evaluation Metrics Enable Effective Benchmarking of Computational Tools for Phosphopeptide Identification

Authors

Wen Jiang, Bo Wen, Kai Li, Wen-Feng Zeng, Felipe da Veiga Leprevost, Jamie Moon, Vladislav A. Petyuk, Nathan J. Edwards, Tao Liu, Alexey I. Nesvizhskii, and Bing Zhang

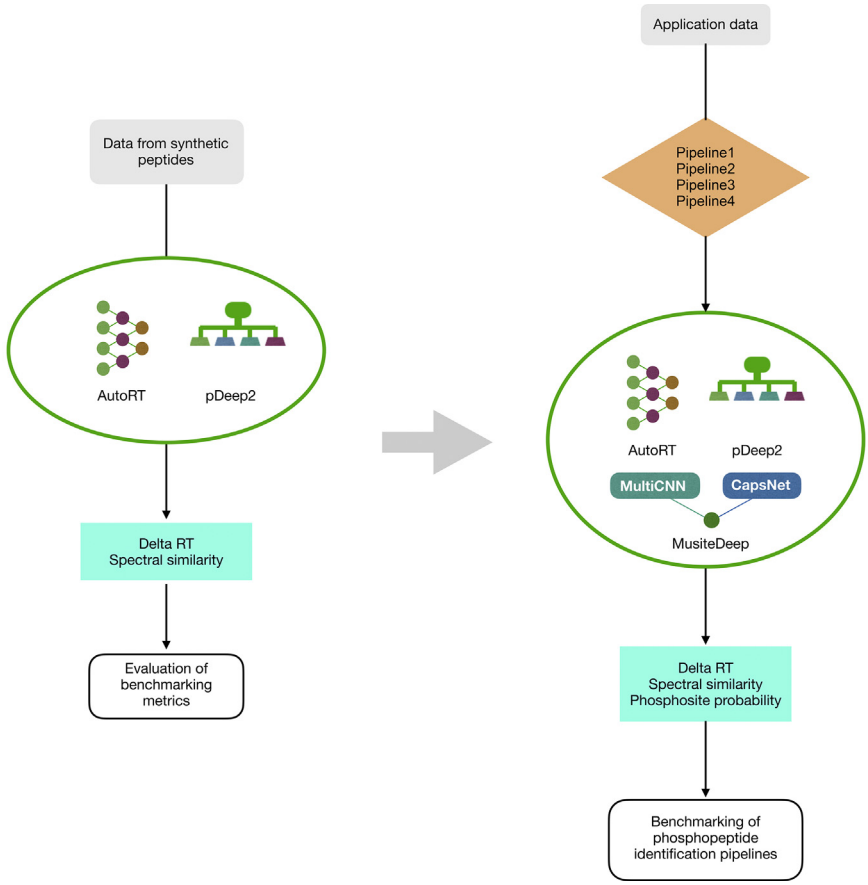
Correspondence

bing.zhang@bcm.edu

In Brief

Tandem mass spectrometry (MS/MS)-based phosphoproteomics is a powerful technology for global phosphorylation analysis. However, applying different computational pipelines to the same dataset may produce substantially different phosphopeptide identification results, underscoring a critical need for benchmarking. We present three deep-learning-derived benchmark metrics. The benchmark metrics demonstrated in this study will enable users to select computational pipelines and parameters for routine analysis of phosphoproteomics data and will offer guidance for developers to improve computational methods.

Graphical Abstract



Highlights

- Computational method selection substantially affects phosphopeptide identification.
- Deep-learning-derived metrics effectively discriminate correct and incorrect PSMs.
- Novel metrics enable computational method comparison on real application data.

2021, Mol Cell Proteomics *20*, 100171

https://doi.org/10.1016/j.mcpro.2021.100171

Deep-Learning-Derived Evaluation Metrics Enable Effective Benchmarking of Computational Tools for Phosphopeptide Identification

Wen Jiang[1], Bo Wen[1], Kai Li[1], Wen-Feng Zeng[2], Felipe da Veiga Leprevost[3], Jamie Moon[4], Vladislav A. Petyuk[4], Nathan J. Edwards[5], Tao Liu[4], Alexey I. Nesvizhskii[3], and Bing Zhang[1,*]

Tandem mass spectrometry (MS/MS)-based phosphoproteomics is a powerful technology for global phosphorylation analysis. However, applying four computational pipelines to a typical mass spectrometry (MS)-based phosphoproteomic dataset from a human cancer study, we observed a large discrepancy among the reported phosphopeptide identification and phosphosite localization results, underscoring a critical need for benchmarking. While efforts have been made to compare performance of computational pipelines using data from synthetic phosphopeptides, evaluations involving real application data have been largely limited to comparing the numbers of phosphopeptide identifications due to the lack of appropriate evaluation metrics. We investigated three deep-learning-derived features as potential evaluation metrics: phosphosite probability, Delta RT, and spectral similarity. Predicted phosphosite probability is computed by MusiteDeep, which provides high accuracy as previously reported; Delta RT is defined as the absolute retention time (RT) difference between RTs observed and predicted by AutoRT; and spectral similarity is defined as the Pearson's correlation coefficient between spectra observed and predicted by pDeep2. Using a synthetic peptide dataset, we found that both Delta RT and spectral similarity provided excellent discrimination between correct and incorrect peptide-spectrum matches (PSMs) both when incorrect PSMs involved wrong peptide sequences and even when incorrect PSMs were caused by only incorrect phosphosite localization. Based on these results, we used all the three deep-learning-derived features as evaluation metrics to compare different computational pipelines on diverse set of phosphoproteomic datasets and showed their utility in benchmarking performance of the pipelines. The benchmark metrics demonstrated in this study will enable users to select computational pipelines and parameters for routine analysis of phosphoproteomics data and will offer guidance for developers to improve computational methods.

Phosphorylation, one of the most common posttranslational modifications (PTMs), is a reversible mechanism that regulates cellular processes such as cell growth, development, and aging through protein kinases and phosphatases (1). Protein phosphorylation dysregulation has been recognized in several diseases, especially cancer (2–4). Tandem mass spectrometry (MS/MS)-based phosphoproteomics provides a high-throughput method to study protein phosphorylation in complex biological samples (5). For example, the Clinical Proteomic Tumor Analysis Consortium (CPTAC) and the International Cancer Proteogenome Consortium (ICPC) have applied phosphoproteomics to the studies of more than ten cancer types (6–14). These and other studies have demonstrated the power of phosphoproteomics in revealing novel biological insights and identifying new effective biomarkers and drug targets for disease prognosis and treatment.

Translating phosphoproteomic data into novel biological and clinical insights relies on effective data analysis (15), and the first step is phosphopeptide identification and phosphosite localization from MS/MS data. Multiple database search tools, such as MaxQuant, MS-GF+, pFind, MSFragger, and X!Tandem, can be used to identify peptide-spectrum matches (PSMs) (16–21). Similarly, multiple tools such as Ascore, PhosphoRS, Andromeda, Mascot Delta Score, pSite, PTMProphet, and LuciPHOr allow scoring of phosphorylation sites to determine phosphosite localizations (22–28).

From the [1]Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, Texas, USA; [2]Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany; [3]Department of Pathology, University of Michigan, Ann Arbor, Michigan, USA; [4]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington, USA; [5]Department of Biochemistry and Molecular & Cellular Biology, Georgetown University, Washington, District of Columbia, USA

*For correspondence: Bing Zhang, bing.zhang@bcm.edu.

https://doi.org/10.1016/j.mcpro.2021.100171

Applying different computational pipelines to the same dataset, however, may produce variable results, with respect to phosphopeptide identifications and phosphosite localizations. This discrepancy has direct impact on all downstream data analyses and interpretation. Therefore, there is a critical need to compare performance of different computational algorithms. New tools are typically assessed in corresponding publications by comparing to previously published tools (22–26). Performed by the tool developers, such evaluation may be biased, a phenomenon called the self-assessment trap (29). Efforts have been made to systematically compare computational pipelines for peptide identification from MS/MS data in an unbiased manner (30–35). An excellent study has comprehensively compared six search engines in combination with several localization scoring algorithms for phosphopeptide identification and site localization (30). In that study, a phosphoproteomic dataset from HeLa cells was used to compare the total number of identifications, and a dataset of synthetic peptides with known sequences and phosphorylation positions was used to compare the quality of identifications. It is desirable to directly compare both number and quality of identifications reported by different computational pipelines in individual application datasets, but quality comparison in real application datasets is difficult due to the lack of ground truth or appropriate evaluation metrics.

Deep learning is a subdiscipline of machine learning based on artificial neural networks. Deep learning can automatically learn patterns from large datasets without handcrafted features. With the exponential growth of MS/MS proteomic data, deep learning has been applied to various areas of proteomics research (36). In particular, deep learning has been highly successful in predicting many peptide properties, including retention time (RT) of a peptide defined as the time point when the peptide elutes from the liquid chromatography (LC) column in an LC-MS/MS system (37–42), fragment intensities of a peptide (38, 41, 43), and phosphorylation sites (44, 45).

In this study, we introduce three deep-learning-derived features as potential metrics for benchmarking computational pipelines for phosphopeptide identification. The first feature is predicted phosphosite probability, which is computed by MusiteDeep with high accuracy as previously reported (44, 46). The second feature is Delta RT, which is defined as the absolute RT difference between RTs observed and predicted by AutoRT (37). The third feature is spectral similarity, which is defined as the Pearson's correlation coefficient (PCC) between spectra observed and predicted by pDeep2 (43). Phosphosite probability predicted by MusiteDeep is independent of experimental conditions and thus can be directly used in our benchmarking study. In contrast, customization of the published AutoRT and pDeep2 models through modified encoding schemes and transfer learning is required for application to individual experiments. We use a synthetic peptide dataset (33) to customize AutoRT and pDeep2 models and to evaluate performance of Delta RT and spectral similarity in discriminating correct and incorrect PSMs. Based on the evaluation results, we use all the three deep-learning-derived features as evaluation metrics to benchmark four computational pipelines on a tandem mass tag (TMT) dataset from the CPTAC human Uterine Corpus Endometrial Carcinomas (UCEC) study (10). The four pipelines include the MS-GF+/Ascore pipeline used in the UCEC publication (19, 22), the CPTAC common data analysis pipeline (CDAP) (MS-GF+/PhosphoRS) (21), the widely used MaxQuant (17), and the more recently published FragPipe pipeline (MSFragger/Philosopher) (20, 47). To demonstrate the general applicability of our benchmarking approach, we further analyzed a TMT dataset from a mouse cell line (mouse_TMT) and a label-free dataset from human cell culture (human_LFQ) using MaxQuant and FragPipe and compared the results using the three evaluation metrics.

EXPERIMENTAL PROCEDURES

Data Sources

Preprepared data for AutoRT base model training and testing were downloaded from the GitHub website (https://github.com/bzhanglab/AutoRT/tree/master/example/data). The MaxQuant search results of the phosphoproteomic data from HeLa cell protein extract used to train and test AutoRT base_phospho model were downloaded from PRIDE (https://www.ebi.ac.uk/pride/) with accession number PXD015087. In our study, we used only a subset of data including 11 raw files (03275_A4_P038189_U09 - 03275_H3_P038189_U08), which were the phosphopeptides analyzed by a nano-flow LC-MS/MS system.

The MaxQuant search results of the synthetic dataset used to evaluate Delta RT and spectral similarity were downloaded from PRIDE (https://www.ebi.ac.uk/pride/) with accession number PXD000138, which included 96 peptide libraries generated from 96 seed peptides. In each library, the two amino acids directly before and after a phosphorylated amino acid in the seed peptide were permutated, leading to libraries of size 2400 peptides or 120 peptides depending on the location of the phosphorylated amino acid. We used the first ten raw files (1.raw–10.raw), which theoretically have 21,720 different peptides (supplemental Table S1).

The raw files of a large-scale TMT10-labeled phosphoproteomic dataset (16 TMT10-plexes, 12 fractions in each plex) from the CPTAC UCEC study were downloaded from the CPTAC data portal (https://cptac-data-portal.georgetown.edu/study-summary/S043). The raw files of an MS3-based TMT10-labeled phosphoproteomic dataset from murine cell lines (mouse_TMT) were downloaded from PRIDE (https://www.ebi.ac.uk/pride/) with accession number PXD015284. One experiment with six fractions was downloaded (02277_A01_P024190_S00_U01_R1 - 02277_F01_P024190_S00_U06_R1). The raw files of a label-free quantification phosphoproteomic dataset from human cell cultures (human_LFQ) were downloaded from PRIDE (https://www.ebi.ac.uk/pride/) with accession number PXD007145. One experiment with three fractions was downloaded (20160408_QE5_nLC5_AH_Bench_2mg_phos_LFQ_oneshot_1C/X/N).

All datasets and their usage in this study are summarized in supplemental Table S1.

Pipelines for the Analysis of Phosphoproteomic Data

For benchmarking, the raw files were searched using four computational pipelines, which are described in Table 1. The first one is MS-

TABLE 1
UCEC search results from four pipelines

Experiment	Pipeline	Search engine	Localization tool	Localization probability/score threshold	Identified localized phosphopeptides	All identified phosphopeptides	Identified localized PSMs	All identified PSMs
16 TMT experiments	MS-GF+/Ascore	MS-GF+	Ascore	13	80,521	103,775	733,325	935,065
	CDAP	MS-GF+	PhosphoRS	0.99	69,400	157,565	423,117	954,927
	MaxQuant	Andromeda	PTM Score	0.75	82,911	120,126	784,940	1,028,857
	FragPipe	MSFragger	PTMProphet	0.75	110,739	142,992	1,175,527	1,393,090
1 TMT Experiment	MS-GF+/Ascore	MS-GF+	Ascore	13	20,200	24,856	29,035	36,440
	CDAP	MS-GF+	PhosphoRS	0.99	14,326	32,382	18,629	42,424
	MaxQuant	Andromeda	PTM Score	0.75	24,412	31,818	35,052	45,908
	FragPipe	MSFragger	PTMProphet	0.75	31,462	37,542	46,949	56,431

GF+ v9881 combined with a localization tool Ascore v1.0.6858, which we referred to as MS-GF+/Ascore pipeline. The second one is MS-GF+ v2017.01.27 combined with a localization tool PhosphoRS, which we referred to as CDAP. The third one was MaxQuant-Andromeda v1.6.5.0, which we referred to as MaxQuant. The fourth one is Philosopher, v3.3.12 (database search with MSFragger, PeptideProphet/ProteinProphet) with a localization tool PTM-Prophet, which we referred to as FragPipe.

Peptide Identification and Phosphosite Localization

For phosphoproteomic data used to train and test AutoRT base_phospho model, search results were downloaded from PRIDE (https://www.ebi.ac.uk/pride/) with accession number PXD015087 (MaxQuant v1.6.2.352, UniProtKB Human Reference Proteome database v22.07.13).

For the synthetic dataset used to evaluate Delta RT and spectral similarity, search results were downloaded from PRIDE (https://www.ebi.ac.uk/pride/) with accession number PXD000138 (MaxQuant version 1.3.0.3, human IPI v3.72 supplemented with synthesized libraries).

All three benchmarking datasets were processed using MaxQuant-Andromeda v1.6.5.0 with default MaxQuant parameters. For the UCEC dataset and human_LFQ dataset, the raw files were searched against RefSeq human protein sequence database downloaded on June 29, 2018 (hg38; 41,734 entries) (48). For the mouse_TMT dataset, the raw files were searched against UniprotKB mouse reference database, downloaded on June 27, 2017 (16,889 entries). For database searching, the following parameters were applied: 20 ppm and 4.5 ppm as first search peptide tolerance and main search peptide tolerance; Trypsin/P as enzyme with two missed cleavage sites; Carbamidomethylation of cysteine as a fixed modification; protein N-terminal acetylation, oxidation of methionine, phosphorylation of serine, threonine, and tyrosine as variable modifications; peptide length with at least seven amino acids. For TMT datasets, fixed TMT modification on the peptide N terminus and Lys residues was considered. The FDR was set to 1% on the site, PSM, and protein levels. A minimum score for modified peptides was set to 40. The cutoff of phosphosite probability estimated by MaxQuant was required to be 0.75 or higher. The MaxQuant output file (msms.txt) was utilized for further analyses.

Besides, all the three benchmarking datasets were processed by FragPipe pipeline. The UCEC dataset was also processed by MS-GF+/Ascore and CDAP. If not specified, the parameters were same as MaxQuant. For FragPipe, the partially tryptic search used a ±20 ppm precursor. For MS-GF+/Ascore, the partially tryptic search used a ±10 ppm parent ion tolerance. A minimum of six unique peptides per 1000 amino acids of protein length were required for achieving 1% at the protein level within the full dataset. Oxidation of methionine was not considered. For CDAP, the semi-tryptic search used a ±20 ppm parent ion tolerance.

RT Data Preparation and AutoRT

AutoRT base model was trained on unmodified peptides from the PXD006109 dataset (37, 49). They were preprepared as training and test data, which contained 123,111 and 13,680 unique unmodified peptide sequences with RTs in the format of AutoRT input files (supplemental Table S1). In order to apply AutoRT to phosphopeptides, we retrained the AutoRT base model using the preprepared training data and encoded oxidation and phosphorylation into the base model. We used "1," "2," "3," "4" to represent oxidation (M), phosphorylation (S), phosphorylation (T), and phosphorylation (Y). The training data were split into 90% and 10% for training and validation automatically by the model. Based on test mean square error (MSE), the top ten best neural architectures were selected and saved. The test MAE of AutoRT base model with 13,680 independent unmodified test peptides from the PXD006109 dataset was 0.5 min. When applying the base model directly to phosphopeptides, the test MAE was 6.1 min.

To enhance the performance of AutoRT on phosphopeptides, we used a two-step transfer learning strategy (manuscript in preparation). Specifically, we prepared transfer learning training and test data using MaxQuant search results of a phosphoproteomic dataset we referred to as the Nano_flow dataset (50). We followed the data preparation rules of AutoRT: when a peptide has multiple PSMs, which means this peptide has multiple RTs, only the peptide with the RT range smaller than 3 min would be kept, and there should be no peptide sequence overlaps between training and test data to avoid overfitting problem. Peptide sequence length limitation is 48 amino acids. The transfer learning training and test data contained 30,360 phosphopeptides and 1000 phosphopeptides separately (supplemental Table S1). We trained the ten models as the basis for transfer learning using the phosphopeptides to improve the base model to base_phospho model (50). The test MAE of the AutoRT base_phospho model with 1000 independent test phosphopeptides from the Nano_flow dataset was 0.6 min. Of note, this AutoRT base_phospho model can be applied to both phosphopeptides and nonmodified peptides.

Delta RT

Before predicting phosphopeptide RTs of a specific dataset, AutoRT base_phospho model was fine-tuned by some highly

confident phosphopeptide identifications from the dataset through transfer learning. In order to get highly confident identifications, we performed database search using multiple computational pipelines and selected the PSMs reported by all pipelines. All PSMs were processed according to the data preparation rules of AutoRT as mentioned above. In total, 90% of highly confident identifications were used to fine-tune AutoRT base_phospho model to create experiment-specific model. Then we predicted RTs of the 10% of the highly confident PSMs (test PSMs, positive controls) and PSMs uniquely identified by individual pipelines (unique PSMs) using the experiment-specific model. Delta RT was the absolute RT difference between RTs observed and predicted by experiment-specific AutoRT model. Smaller Delta RTs indicated better identification and localization performance.

Spectrum Data Preparation and pDeep2

Data used to train and test the pDeep2 base model (pretrain-180921-modloss.ckpt) and pDeep2 base_phospho model (pretrain-180921-modloss-transfer-Phos.ckpt) were described in (43). The original pDeep2 base model and base_phospho model were trained and tested using data from various laboratories. The authors considered instrument types and normalized collision energy (NCE) in the model, making pDeep2 adaptive for different instruments (43). We used original pDeep2 base model and base_phospho model, which can predict the intensities of four types of ions (b, y, b-Modloss, y-Modloss).

Spectral Similarity

Before MS/MS spectrum prediction of a specific phosphoproteomic dataset, pDeep2 base_phospho model was fine-tuned by some highly confident phosphopeptide identifications from the dataset through transfer learning, as described in the Delta RT section. After transfer learning, we used the experiment-specific models to predict MS/MS spectra and calculated the spectral similarities of the predicted spectra and the experimental spectra. We considered PCC, Spearman's Correlation Coefficients (SPC), cosine similarity (cos), Kendal rank correlation coefficient (kdt), and spectral angle (SA) as possible measurements of spectral similarity, and they led to similar conclusions. PCC was selected in this study to represent spectral similarity. The closer the spectral similarity is to 1, the more similar the input peptide sequence and modification information are to the real ones.

General Phosphorylation Probability

We provided RefSeq human and UniProtKB mouse reference peptide libraries described above as input for the MusiteDeep server (https://www.musite.net), which returned the general phosphosite probabilities of serine, threonine, and tyrosine in the peptide sequences (44). The MusiteDeep prediction results of human and mouse libraries are listed in supplemental Table S2.

Benchmark Metrics Evaluation Data

One large-scale synthetic proteomics reference library was used to evaluate two deep-learning-derived metrics, Delta RT and spectral similarity (33). We prepared the training and test data for fine tuning AutoRT base_phospho model, pDeep2 base model, and pDeep2 base_phospho model. Only high scoring PSMs (MaxQuant Delta Score $\geq$ 6 and Score $\geq$ 40 for both unmodified peptides and phosphopeptides and phosphosite localization probability >0.75 for phosphopeptides) were included in this analysis. For AutoRT, we split these PSMs into 6854 unique peptides with RTs for training and 783 unique peptides with RTs for testing. For pDeep2 base_phospho model, there were 7205 PSMs for training and 2506 PSMs for testing. For pDeep2 base model, there were 8484 PSMs for training and 2233 PSMs for testing. There were no overlapping peptide sequences in training and test data to avoid overfitting. The group alias of the positive test phosphopeptides is M_SeqT_LocT (phosphopeptides with true sequences and true phosphorylation localization). The group alias of the positive test unmodified peptides is UM_SeqT (unmodified peptides with true sequences).

In order to evaluate the ability of Delta RT and spectral similarity to distinguish different types of negative controls, we split the negative PSMs into three groups including wrongly identified phosphopeptides (M_SeqF), correctly identified wrongly localized phosphopeptides (M_SeqT_LocF), and wrongly identified unmodified (UM_SeqF). Since the PSM numbers in the M_SeqT_LocF group were small (671 PSMs for Delta RT; 130 PSMs for spectral similarity), we further simulated two additional negative phosphopeptide groups with correctly identified wrongly localized phosphopeptides. We simulated the negative PSMs from the positives. For each positive PSM in which the peptide included two or more Ser/Thr/Tyrs, we kept the retention time and the ion intensity unchanged and moved the correctly localized phosphate group to a randomly chosen wrong localization to create the M_Sim1_LocF group. For M_Sim2_LocF, we limited the move of the phosphate group to the wrong localization closest to the correct one to create the most challenging negative group for discrimination.

Phosphoproteomic Data for Benchmarking

One large-scale phosphoproteomic dataset from CPTAC (TMT-10plex labeled data from >100 uterine cancer samples, UCEC) was used for benchmarking (10). The raw data including 16 experiments which were processed using four computational pipelines including MS-GF+/Ascore, CDAP, MaxQuant, and FragPipe (10, 17, 19, 22, 23, 47).

Search results were filtered based on the recommended cutoffs of each pipeline. One experiment was used for benchmarking (01CPTAC_UCEC_P_PNNL_20170922_B1S1). Since MS-GF+/Ascore pipeline did not consider oxidation on methionine, we considered the overlap PSMs among three pipelines (MaxQuant, FragPipe, CDAP) as highly confident PSMs, which was used to prepare for the training and test data for AutoRT base_phospho model and pDeep2 base_phospho model fine-tuning. For AutoRT and pDeep2, the highly confident data from the experiment were split into 7864 phosphopeptides for training, 851 phosphopeptides for testing, which also served as positive controls. The PSMs identified by only one tool were unique PSMs used for benchmarking the four tools. For AutoRT, the peptide length limitation of training PSMs, test PSMs, and unique PSM was 48 amino acids. Moreover, unique PSMs were filtered out when the same sequences appeared in the training or test PSMs to avoid overfitting.

In order to demonstrate the general applicability of our approach, a TMT dataset from a mouse cell line (mouse_TMT) and a label-free dataset from human cell culture (human_LFQ) were reanalyzed (51, 52). One experiment from each dataset was used for benchmarking. Search results were filtered based on the recommended cutoffs of each pipeline and were shown in supplemental Table S3. The overlap PSMs between the two pipelines were treated as highly confident PSMs and prepared for the training and test data for AutoRT base_phospho model and pDeep2 base_phospho model transfer learning. The mouse_TMT ground truth from the experiment was split into 6851 unique phosphopeptides for training, 738 unique phosphopeptides for testing. The human_LFQ ground truth from the experiment was split into 12,484 unique phosphopeptides for training, 1950 unique phospho-peptides for testing. The PSMs identified by only one tool were unique PSMs, which were used for comparison between MaxQuant and FragPipe. For AutoRT, the peptide length limitation of training PSMs, test PSMs, and unique PSM was 48 amino acids. Moreover, unique PSMs were filtered out when the same sequences were in the training or test PSMs to avoid overfitting.

Experimental Design and Statistical Rationale

For UCEC, global phosphoproteomics data including 16 experiments were processed by MaxQuant as one batch. For mouse_TMT and human_LFQ, phospho-proteomics data from one experiment were processed by MaxQuant as one batch. AutoRT base_phospho model and pDeep2 base_phospho model (or base model for unmodified peptides) required transfer learning using highly confident data from the specific experiment in which RTs need to be predicted. In the evaluation part of Delta RT and spectral similarity, the positive and negative test PSMs from synthetic peptides were used to evaluate the ability of the deep-learning-derived metrics to distinguish wrongly identified or localized peptides. In the application part, Delta RT, spectral similarity and phosphosite probability were used to evaluate the performance of phosphoproteomics pipelines. Data analysis was performed using Python language (Jupyter Notebook v 5.6.0, Python 3). Two-sided Wilcoxon rank sum test was performed, and *p*-values were corrected with Bonferroni correction.

RESULTS

An Overview of the Study

Figure 1 provides a general overview of our study design. Phosphosite probability prediction is independent of experimental conditions, and it has been shown that MusiteDeep can predict general phosphosites with high accuracy, with area under the receiver operating characteristic curve (AUROC) scores of 0.896 for phospho-serine and phospho-threonine and 0.958 for phospho-tyrosine (44). Therefore, phosphosite probability predicted by MusiteDeep (supplemental Table S2) was used directly as a benchmark metric. In contrast, it was unclear whether Delta RT and spectral similarity could accurately distinguish correct and incorrect phosphopeptide identifications and phosphosite localizations. Figure 1*A* illustrates our approach to evaluate these two potential benchmark metrics. One large-scale synthetic proteomics reference library (33) was used to fine-tune AutoRT (37) base_phospho model, pDeep2 (43) base_phospho model, and pDeep2 base model, and to evaluate Delta RT and spectral similarity for discriminating correct and incorrect PSMs (Experimental Procedures). Figure 1*B* shows the application of Delta RT, spectral similarity, and MusiteDeep-predicted phosphosite probability to compare four computational phosphoproteomics pipelines. We randomly selected one TMT experiment from the CPTAC UCEC phosphoproteomic dataset to fine-tune the AutoRT base_phospho model and pDeep2 base_phospho model and to benchmark the four pipelines. The common PSMs reported by multiple pipelines are likely to be high-quality identifications. For Delta RT-based and pDeep2-based evaluation, the common PSMs were split into two sets, one for fine-tuning experiment-specific models and the other served as positive controls in benchmarking. For predicted phosphosite probability-based evaluation, all common PSMs were used as positive controls in benchmarking. To assess the quality of the pipelines, the experiment-specific AutoRT model, the experiment-specific pDeep2 model, and MusiteDeep were applied to PSMs uniquely identified by each pipeline. Delta RTs, spectral similarities, and predicted phosphosite probabilities for the five groups of PSMs were compared against each other.

UCEC Search Results From Four Pipelines

The CPTAC UCEC phosphoproteomic data including 16 TMT10-labeled experiments was searched against the RefSeq human protein database using four computational pipelines including MS-GF+/Ascore, CDAP, MaxQuant, and FragPipe (10, 17, 19, 22, 23, 47). Search engines, localization tools, and localization probability filtering thresholds used in these pipelines, together with search results with and without localization probability filtering, are shown in Table 1. The four pipelines identified a total of 248,868 phosphopeptides and 2,246,161 PSMs without phosphosite localization probability filtering, and 157,275 phosphopeptides and 1,595,981 PSMs with phosphosite localization probability filtering (Fig. 2, *A* and *B*). Among the localized phosphopeptides and PSMs, only 22.3% and 11.4%, respectively, were commonly reported by all four pipelines. MaxQuant and FragPipe had more overlap at the phosphopeptide- and PSM-levels than other overlaps of two other pipelines. In the pairwise Jaccard similarity index analysis, MaxQuant and FragPipe showed the highest similarity, with Jaccard indexes of 0.59 and 0.47 at the phosphopeptide- and PSM-levels respectively (0.54 and 0.44 without phosphosite localization probability filtering) (supplemental Fig. S1, *A* and *B*). While MS-GF+/Ascore and CDAP used the same identification algorithm MS-GF+, Jaccard indexes of the identifications reported by the two pipelines were just 0.36 and 0.25 at the phosphopeptide- and PSM-levels, respectively (0.38 and 0.28 without phosphosite localization probability filtering). This may be explained by either different parameters used in database searching or difference introduced by different localization tools. Overall, FragPipe identified the most localized phosphopeptides and PSMs. CDAP identified the most phosphopeptides but after extremely conservative localization probability filtering, it reported the fewest localized phosphopeptides. The proportion of unlocalized phosphopeptides from CDAP was 56%, which was much higher than the proportions from the other three pipelines, ranging from 22% to 31% (Fig. 2*C*).

Benchmark Metric Evaluation

In order to evaluate performance of Delta RT and spectral similarity in discriminating correct and incorrect PSMs, the modified and unmodified peptides previously reported in a synthetic dataset (33) were analyzed separately. PSMs from the search results were divided into correct ones, *i.e.*, M_SeqT_LocT, UM_SeqT, and incorrect ones, *i.e.*, M_SeqT_LocF, M_SeqF, UM_SeqF. Group names with prefix "UM_" mean unmodified peptides; "M_" means modified peptides. Group names with "_SeqT" or "_SeqF" mean peptides with correct or

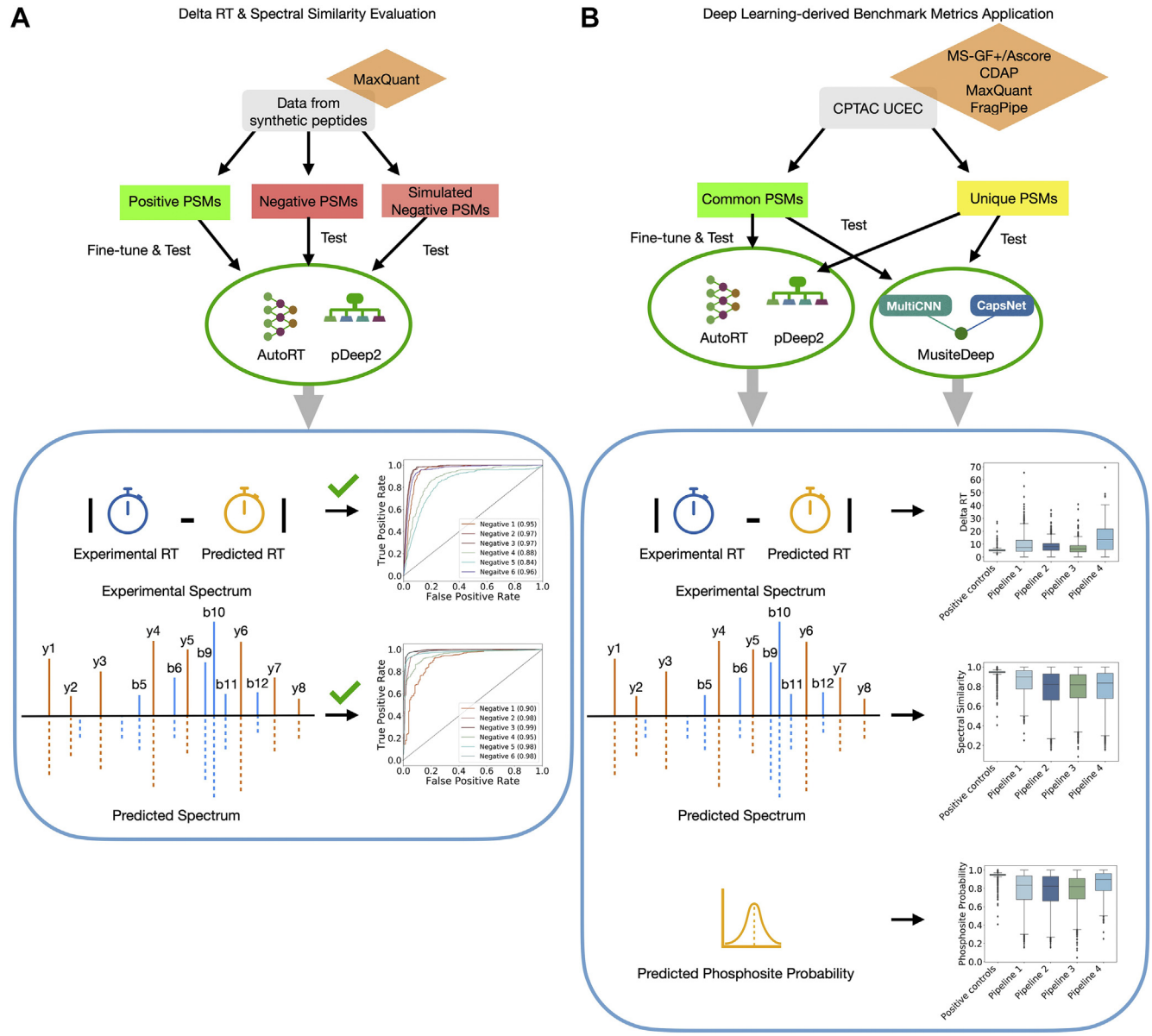


FIG. 1. **Overview of the study design.** *A*, workflow for evaluating two deep-learning-derived metrics. The workflow includes data preparation, model training, computation of Delta RT and spectral similarity, and evaluation using area under the receiver operating characteristic curves (AUROCs). *B*, benchmarking workflow, which includes data preparation, model training, computation of Delta RT, spectral similarity, and phosphosite probability, and comparison of PSMs identified by all computational pipelines (common PSMs) and those uniquely identified by each pipeline (unique PSMs) using the three benchmark metrics.

incorrect sequences; "_LocT" or "_LocF" means peptides with correct or incorrect phosphorylation localization (Table 2).

Correct PSMs were further split into training and positive testing PSMs. For negative testing PSMs, we supplemented incorrect PSMs from the search results with simulated negative PSMs. As described in the Experimental Procedures section, the negative test PSMs were classified into a few different groups in order to evaluate the performance of the metrics in distinguishing different types of negative PSMs from positive PSMs. Detailed information of training and testing PSMs is described in Table 2 and Figure 3*A*.

In order to enhance and evaluate the performance of AutoRT on phosphopeptides, we performed a two-step transfer learning on AutoRT base model, which was trained and tested on 123,111 and 13,680 unique unmodified peptides from the PXD006109 dataset, respectively (37, 49) (supplemental Table S1, Experimental Procedures section). In the first-round transfer learning, we improved AutoRT base model to AutoRT base_phospho model using a Nano_flow dataset (PXD015087) (50), in which 30,360 and 1000 unique phosphopeptides were used for training and testing, respectively. Then, we fine-tuned AutoRT base_phospho model to

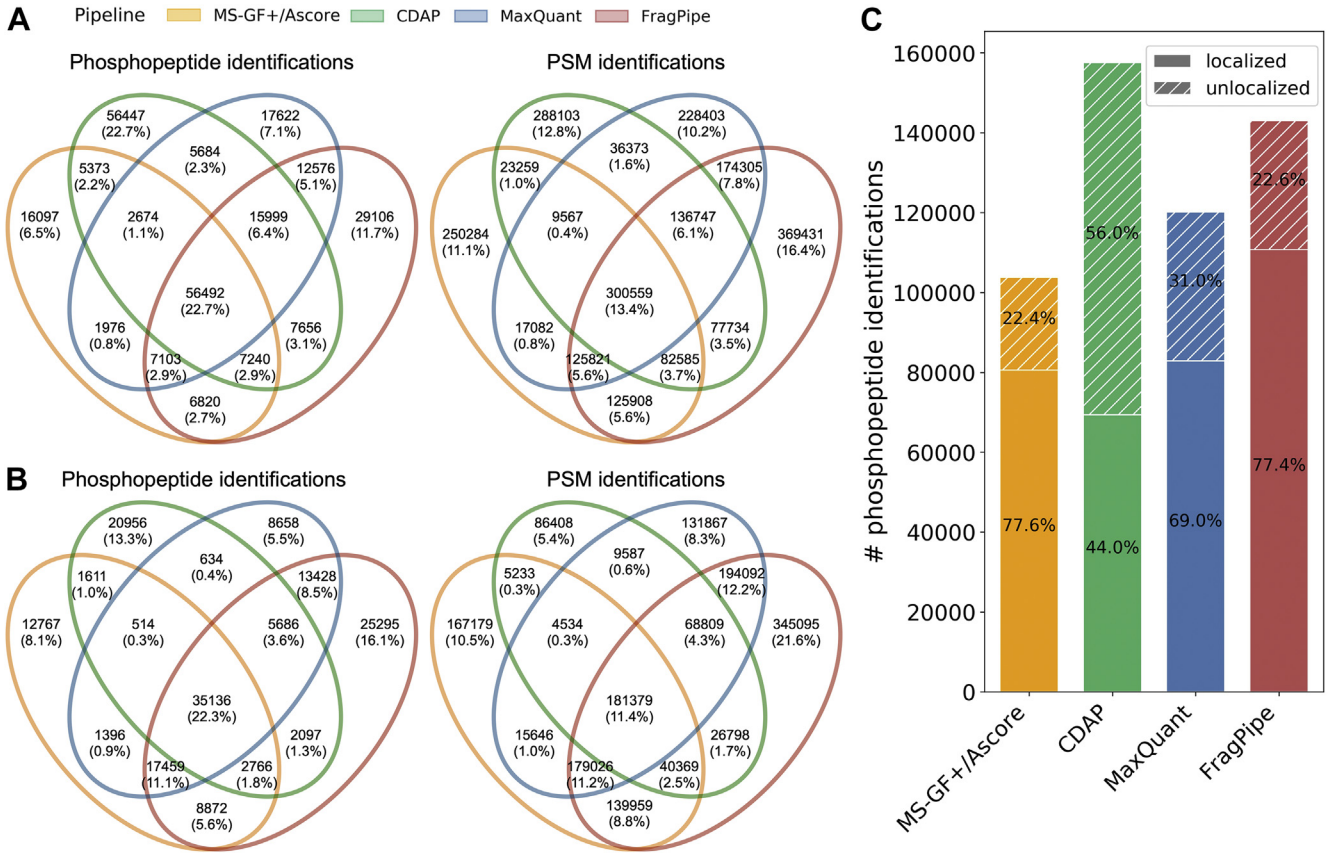


FIG. 2. **UCEC search results from MS-GF+/Ascore, CDAP, MaxQuant, and FragPipe.** *A* and *B*, Venn diagrams of search results of 16 experiments in UCEC from MS-GF+/Ascore, CDAP, MaxQuant, and FragPipe without (*A*) and with (*B*) localization probability filtering. *C*, proportions of localized and unlocalized phosphopeptides from the four pipelines.

AutoRT experiment-specific model using 6887 unique synthetic peptides from PXD000138 (3549 modified peptides and 3338 unmodified peptides) using a second round of transfer learning. Evaluation using M_SeqT_LocT showed that the test Median Absolute Error (MAE) of the experiment-specific model for phosphopeptides was 0.44 min and the correlation of predicted RT and observed RT was 0.97 (Fig. 3*B*), and these numbers were comparable to those from the experiment-specific model for unmodified peptides evaluated using UM_SeqT. These results demonstrated high accuracy of the experiment-specific AutoRT model in RT prediction for both unmodified peptides and phosphopeptides.

Using the experiment-specific AutoRT model, we predicted RTs of the peptides involved in test PSMs and calculated Delta RTs. The positive test PSMs had significantly smaller Delta RTs than all groups of negative test PSMs (Wilcoxon rank sum test, Bonferroni correction, p-value $\leq$ 1e-4) for both modified (Fig. 3*C*) and unmodified peptides (Fig. 3*D*). Then, we further quantified the ability of Delta RT to discriminate between positive test PSMs and different types of negative test PSMs by calculating AUROC. For negative unmodified peptides and phosphopeptides with wrong sequences, the AUROCs were 0.95 and 0.96, respectively. Importantly, for negative phosphopeptides with wrong localizations of phosphorylation sites, the AUROCs were 0.84 to 0.88 (Fig. 3*E*). These data suggested that Delta RT provided excellent discrimination between the positives and different types of the negatives, even for the most difficult scenario in which phosphopeptides had correct sequences but wrong localizations.

For pDeep2, we employed transfer learning to fine-tune the pDeep2 base_phospho model/base model to generate experiment-specific models using ~15,700 PSMs from the synthetic dataset and evaluated the experiment-specific pDeep2 models using ~4700 positive test PSMs. The median PCCs for phosphopeptides and unmodified peptides were 0.97 and 0.98, respectively (Fig. 3, *F* and *G*). These results demonstrated outstanding performance of pDeep2 on both unmodified peptides and phosphopeptides. Next, we predicted the MS/MS spectra of peptides involved in all test PSMs. For both modified (Fig. 3*F*) and unmodified (Fig. 3*G*) peptides, the PCCs from the positive test PSMs were significantly higher than those from the negative test PSMs (Wilcoxon rank sum test, Bonferroni correction, p-value $\leq$ 1e-4). We further quantified the ability of PCC to discriminate between positive test PSMs and different types of negative test

TABLE 2
Training and test PSMs from the synthetic peptide dataset (PXD000138)

Phosphopeptides						
Seq	Phosphorylation number	Site	Group alias	Test PSMs classification	Peptide number	Usage
T	T	T	M_SeqT _LocT		3549	Train
T	T	T	M_SeqT _LocT	Positive	390	Test
T	T	F	M_SeqT _LocF	Negative	488	Test
T	T	F	M_Sim1_LocF	Negative	4499	Test
T	T	F	M_Sim2_LocF (closest site)	Negative	3252	Test
F			M_SeqF	Negative	7572	Test

Unmodified peptides				
Seq	Group Alias	Classification	PSM Number	Usage
T	UM_SeqT		3338	Train
T	UM_SeqT	Positive	358	Test
F	UM_SeqF	Negative	567	Test

PSMs using AUROC. For negative unmodified peptides and phosphopeptides with wrong sequences, the AUROCs were 0.89 and 0.98, respectively (Fig. 3*H*). For correctly identified phosphopeptides with wrong phosphorylation localization, the AUROCs were also very high (0.93–0.96).

Together, our evaluation results suggest that both Delta RT and spectral similarity are qualified metrics for benchmarking phosphopeptide identification.

Benchmarking of Computational Pipelines

We compared four computational phosphoproteomics pipelines using three independent metrics, Delta RT, spectral similarity, and MusiteDeep-predicted phosphosite probability. One TMT experiment from the CPTAC UCEC phosphoproteomic dataset was selected to fine-tune AutoRT and pDeep2 and to benchmark the four pipelines. In this TMT experiment, FragPipe reported the largest number of localized phosphopeptides, followed by MaxQuant, MS-GF+/Ascore, and CDAP (Fig. 4*A*).

RTs reported by MaxQuant were the same as those reported by the other pipelines for the same PSMs (supplemental Fig. S2*A*), indicating that RTs reported by different pipelines for the same PSMs were consistent. Common PSMs reported by multiple pipelines were used for both model training and positive control in benchmarking. Specifically, among the ~10,000 common PSMs, 90% were used for training of an experiment-specific RT prediction model through transfer learning based on the AutoRT base_phospho models (7864 peptides for training), and the remaining 10% (815 peptides) were saved as positive control PSMs in benchmarking. For the positive control PSMs, the correlation of predicted RTs and observed RTs of the positive control PSMs was 0.998 (Fig. 4*B*). These results suggested high accuracy of the predictions. We also classified the union of peptides identified by any of the four search engines, except for those involved in the training PSMs, into different groups based on the number of search engines identifying each peptide. Peptides identified by more search engines had significantly lower Delta RTs (supplemental Fig. S3*A*), further supporting the validity of Delta RT as a benchmark metric.

To assess the quality of the pipelines, we focused on the PSMs uniquely identified by each pipeline (Fig. 4*A*). In this analysis, unique PSMs were further filtered by removing those involving a sequence overlapping with the training sequences and those involving a sequence longer than 48 amino acids due to the peptide length limitation of AutoRT. Unique PSMs from FragPipe showed the highest correlation of predicted RTs and observed RTs (SPC = 0.975, Fig. 4*C*), followed by those from MS-GF+/Ascore (SPC = 0.970), MaxQuant (SPC = 0.956), and CDAP (SPC = 0.905) (supplemental Fig. S2*B*). As expected, the common PSMs showed the lowest Delta RTs (median Delta RT: 0.74) compared with the unique PSMs. Among different groups of pipeline-specific PSMs, those from FragPipe showed significantly lower Delta RTs than the ones from the other three pipelines (Fig. 4*D*, Wilcoxon rank sum test, Bonferroni correction, p-value $\leq$ 1e-4). For unique PSMs from MS-GF+/Ascore, CDAP, MaxQuant, and FragPipe, their median Delta RTs were 1.33 min, 1.16 min, 1.20 min, and 1.06 min. In total, 90% of common PSMs had Delta RTs lower than 2.21 min. For unique PSMs from MS-GF+/Ascore, CDAP, MaxQuant, and FragPipe, there were 69.3%, 70.8%, 69.8%, and 76.2% Delta RTs lower than 2.21 min.

For spectral similarity, an experiment-specific ion intensity prediction model was trained through transfer learning based on the pDeep2 base_phospho model and then tested using the same split of training and test data as AutoRT. High similarities (median PCC: 0.97) between predicted and observed spectra for the test positive control PSMs (*i.e.*, common PSMs in Fig. 4*E*) demonstrated the high accuracy of ion intensity prediction using this model. PSMs supported by more search engines were associated with significantly higher

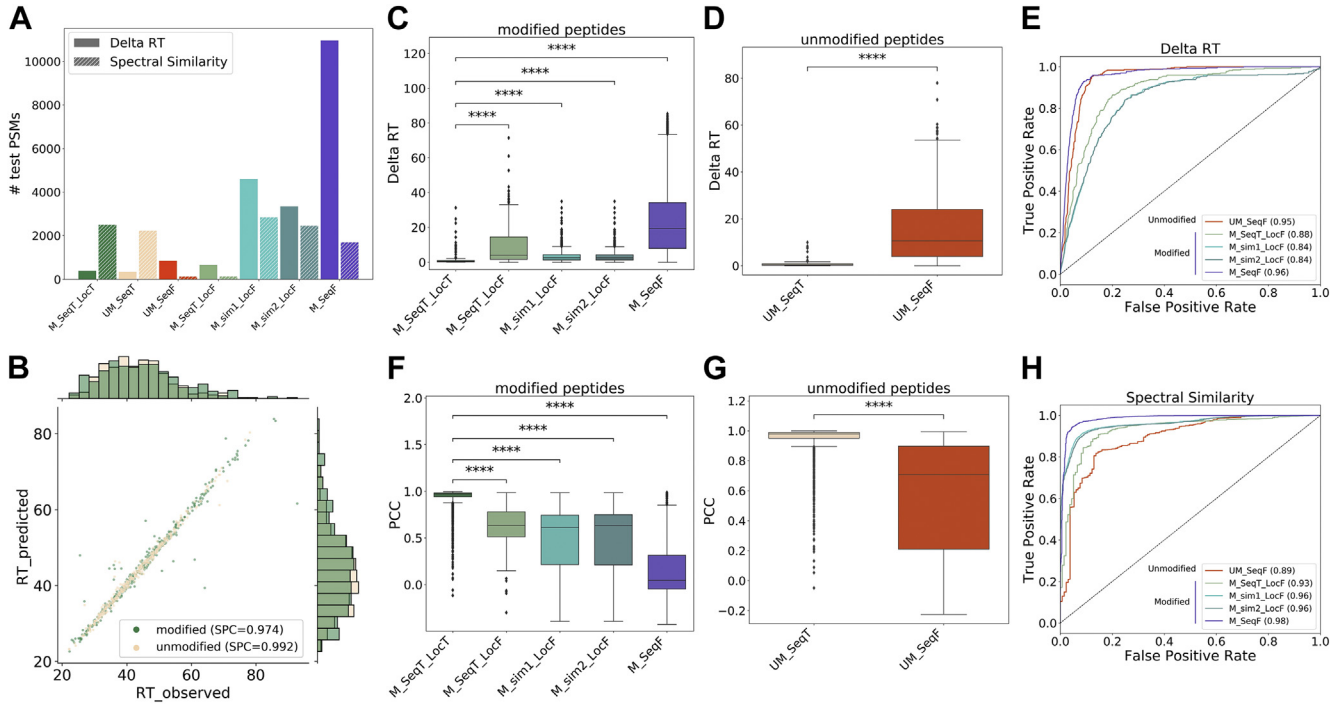


FIG. 3. **Evaluation of Delta RT and spectral similarity as potential benchmark metrics.** *A*, the numbers of positive and negative test PSMs used to evaluate Delta RT and spectral similarity. *B*, scatter plot comparing predicted RTs and observed RTs from positive test PSMs (modified and unmodified peptides). *C* and *D*, Delta RT distributions for positive and negative test modified (*C*) and unmodified (*D*) peptides. *E*, test results of Delta RT on negative test PSMs with unmodified and modified peptides. Group names with prefix "UM_" mean unmodified peptides. Group names with prefix "M_" mean modified peptides. Group names with "_SeqT" or "_SeqF" mean peptides with correct or incorrect sequences. Group names with "_LocT" or "_LocF" mean peptides with correct or incorrect phosphorylation localization. *F* and *G*, spectral similarity distributions for positive and negative test modified peptides and unmodified peptides. *H*, test results of spectral similarity on negative test PSMs with unmodified and modified peptides. For *C*, *D*, *F* and *G*, ns: $p >$ ns: 5.00e-02 $< p \leq$ 1.00e+00; *: 1.00e-02 $< p \leq$ 5.00e-02; **: 1.00e-03 $< p \leq$ 1.00e-02; ***: 1.00e-04 $< p \leq$ 1.00e-03; ****: $p \leq$ 1.00e-04 (Wilcoxon rank sum test, Bonferroni correction). *Centerlines* in the boxplots indicate medians, box limits indicate upper and lower quartiles.

PCCs (supplemental Fig. S3*B*). The common PSMs showed the highest spectral similarities compared with the unique PSMs. Among different groups of pipeline-specific PSMs, those from FragPipe showed significantly higher PCCs than the ones from the other three pipelines (Fig. 4*E*, Wilcoxon rank sum test, Bonferroni correction, *p*-value $\leq$ 1e-4). For unique PSMs from MS-GF+/Ascore, CDAP, MaxQuant, and FragPipe, their median PCCs were 0.91, 0.92, 0.93, and 0.94. In total, 90% of common PSMs had PCCs higher than 0.92. For unique PSMs from MS-GF+/Ascore, CDAP, MaxQuant, and FragPipe, there were 48%, 50%, 50%, and 60% PCCs higher than 0.92.

In predicted phosphosite probability-based benchmarking, the median probability of the peptides corresponding to common PSMs was 0.877, which was clearly higher than those corresponding to the unique PSMs. Among the four groups of pipeline-specific peptides, those uniquely reported by FragPipe and MaxQuant had similar probability distributions, and both sets had significantly higher probability distributions than those uniquely reported by MS-GF+/Ascore or CDAP (Fig. 4*E*, Wilcoxon rank sum test, Bonferroni correction, *p*-value $\leq$ 1e-4). For unique PSMs from MS-GF+/Ascore, CDAP, MaxQuant, and FragPipe, their median phosphosite probabilities were 0.833, 0.805, 0.853, and 0.853, respectively. In total, 90% of common PSMs had phosphosite probabilities higher than 0.721. For unique PSMs from MS-GF+/Ascore, CDAP, MaxQuant, and FragPipe, there were 69.0%, 62.0%, 80.0%, and 78.4% higher than 0.721.

To study the impact of peptide length on PSM quality as indicated by Delta RT, we grouped the PSMs on the basis of peptide length (supplemental Fig. S3*C*). Although all the four pipelines had the same maximum peptide length setting of 50 amino acids, it was noticeable that FragPipe reported more confident longer peptides than the other tools. For the common PSMs, average Delta RTs were low and stable for phosphopeptides shorter than 25 amino acids and tended to be more variable and larger for longer phosphopeptides. The similar pattern was observed for PSMs uniquely reported by FragPipe, although the Delta RT stability was extended to peptides with length longer than 25 amino acids. For PSMs uniquely reported by the other three pipelines, average Delta RT was not only higher but also more variable even for peptides shorter than 25 amino acids. These results suggested that the quality of unique PSMs reported by FragPipe was

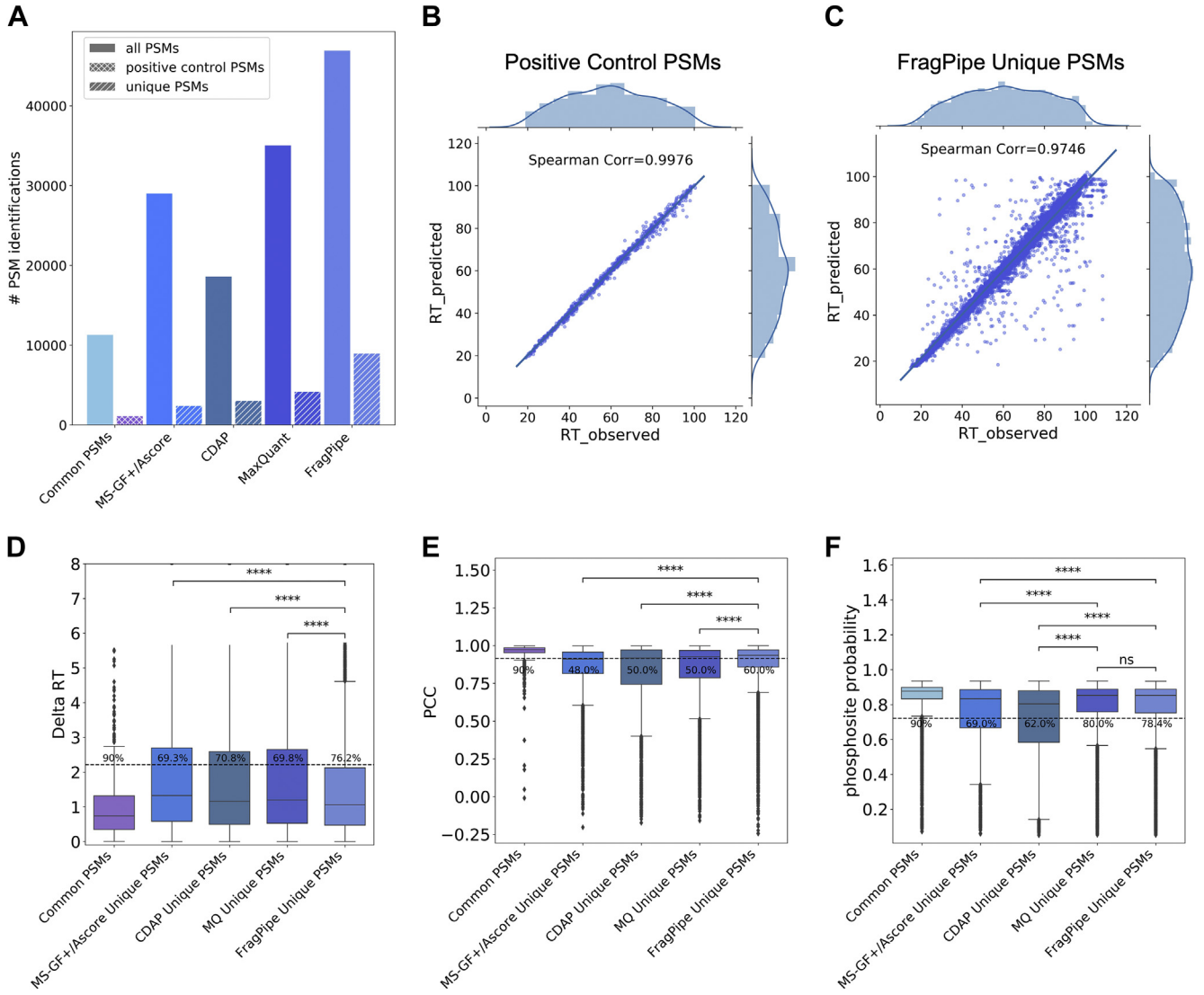


FIG. 4. **Benchmarking of MS-GF+/Ascore, CDAP, MaxQuant, and FragPipe using a TMT human dataset.** *A*, the numbers of all common PSMs, test common PSMs used as positive controls in Delta RT-based benchmarking, and all PSMs and filtered unique PSMs identified by MS-GF+/Ascore, CDAP, MaxQuant, and FragPipe in one UCEC experiment (filtering method: peptide length ≤ 48 aa, no sequence overlaps in ground truth). *B*, scatter plot comparing predicted RTs and observed RTs from positive control PSMs. *C*, scatter plots comparing predicted RTs and observed RTs in the unique PSMs from FragPipe. *D–F*, Delta RT (*D*), spectral similarity (*E*) and phosphosite probability (*F*) distributions for common PSMs and unique PSMs from MS-GF+/Ascore, CDAP, MaxQuant, and FragPipe. A horizontal baseline showing the lower 90% of Delta RTs (*D*), and the higher 90% of spectral similarity (*E*) and phosphosite probability (*F*) of common PSMs. The ratios of PSMs with Delta RTs lower than the baseline (*D*) and those with spectral similarity (*E*) phosphosite probabilities (*F*) higher than the baseline in the pipeline-unique PSMs were labeled. ns: $p >$ ns: 5.00e-02 $< p \leq$ 1.00e+00; *: 1.00e-02 $< p \leq$ 5.00e-02; **: 1.00e-03 $< p \leq$ 1.00e-02; ***: 1.00e-04 $< p \leq$ 1.00e-03; ****: $p \leq$ 1.00e-04 (Wilcoxon rank sum test, Bonferroni correction). For boxplots, *centerline* indicates the median, box limits indicate upper and lower quartiles.

similar to the common PSMs. Notably, although modification localization might be more difficult for longer peptides due to more location options, FragPipe was able to maintain relatively low average Delta RTs in this difficult part.

Together, these results indicated that the FragPipe statistically significantly outperformed the other three in both sensitivity and quality.

General Applicability of the Benchmarking Approach

Since FragPipe outperformed the other three pipelines, and MaxQuant is the most widely used tool, we reanalyzed two additional large-scale phosphoproteomic datasets, a TMT-based mouse (mouse_TMT) and a label-free human (human_LFQ) datasets using these two pipelines and compared their performance using the three benchmark metrics (supplemental Table S3).

FragPipe identified more localized phosphopeptides and PSMs than MaxQuant in both datasets (Fig. 5, *A* and *B*). The proportions of localized and unlocalized phosphopeptides from MaxQuant and FragPipe were similar (Fig. 5*C*). For AutoRT, the common PSMs reported by MaxQuant and FragPipe were treated as highly confident identifications and 90% of them were used for training of the experiment-specific RT prediction models (mouse_TMT: 6851 peptides for training; human_LFQ: 12,484 peptides for training), and the remaining 10% (mouse_TMT: 738 peptides; human_LFQ: 1950 peptides) were saved as positive controls in benchmarking (Fig. 5*D*). PSMs uniquely identified by each pipeline were filtered in the similar way as described above. For each dataset, the training data were used to fine-tune the AutoRT base_phospho model and pDeep2 base_phospho models, and the experiment-specific models were used to predict RTs and ion intensities for the positive control PSMs and the pipeline-specific PSMs. For the positive control PSMs, the correlations of predicted RTs and observed RTs of the positive control PSMs were 0.97 and 0.96 for the mouse_TMT and human_LFQ datasets, respectively (Fig. 5, *E* and *F*).

The positive control PSMs showed the lowest Delta RTs (median RT: 0.63 min for mouse_TMT, 1.45 min for human_LFQ) compared with the two groups of unique PSMs (Fig. 5*G*). For unique PSMs from MaxQuant and FragPipe, their median Delta RTs were 0.79 min and 0.74 min in mouse_TMT dataset, and 2.99 min and 2.78 min in human_LFQ dataset. In total, 90% of common PSMs had Delta RTs lower than 1.84 min and 5.31 min for the mouse_TMT and human_LFQ datasets, respectively. For unique PSMs from MaxQuant and FragPipe, there were 80.7% and 85.9% Delta RTs lower than 1. 84 min in the mouse_TMT, and there were 68.5% and 74.5% Delta RTs lower than 5.31 min in the human_LFQ. The FragPipe-specific PSMs showed significantly lower Delta RTs than the MaxQuant-specific PSMs in both the mouse_TMT dataset (Wilcoxon rank sum test, Bonferroni correction, p-value = 3.3e-3) and the human_LFQ dataset (Wilcoxon rank sum test, Bonferroni correction, $p \leq$ 1e-4).

For pDeep2, the same split of training and test data from mouse_TMT and human_LFQ were used to train and test the experiment-specific ion intensity prediction models (Fig. 5*D*). The positive control PSMs showed the highest spectral similarities (median PCC: 0.94 for mouse_TMT, 0.96 for human LFQ) compared with the two groups of unique PSMs (Fig. 5*H*). For unique PSMs from MaxQuant and FragPipe, their median PCCs were in 0.92 and 0.90 in mouse_TMT dataset, and 0.95 and 0.94 in human_LFQ dataset. In total, 90% of common PSMs had spectral similarities higher than 0.70 and 0.89 for the mouse_TMT and human_LFQ datasets, respectively. For unique PSMs from MaxQuant and FragPipe, there were 82% and 85% spectral similarities higher than 0.7 in the mouse_TMT, and there were 74% and 71% spectral similarities higher than 0.89 in the human_LFQ. For spectral similarities, there was no significant difference for mouse_TMT dataset, and MaxQuant-specific PSMs had higher spectral similarities for human_LFQ dataset.

For predicted phosphosite probability, peptides reported by both pipelines showed the highest probabilities, and peptides uniquely reported by FragPipe had significantly higher probabilities than those uniquely reported by MaxQuant in both datasets (Wilcoxon rank sum test, Bonferroni correction, $p \leq$ 1e-4) (Fig. 5*H*). In total, 90% of common PSMs had phosphosite probabilities higher than 0.782 and 0.788 for the mouse_TMT and human_LFQ datasets, respectively. For unique PSMs from MaxQuant and FragPipe, there were 80.5% and 87.9% higher than 0.782 in the mouse_TMT, and there were 79.5% and 83.0% higher than 0.788 in the human_LFQ.

In summary, FragPipe identified more phosphopeptides and PSMs than MaxQuant in both datasets. Quality evaluation based on both Delta RT and phosphosite probability showed that FragPipe outperformed MaxQuant in both the mouse_TMT and human_LFQ datasets. Although MaxQuant outperformed FragPipe in the human_LFQ dataset according to the evaluation based on spectral similarity, a simple voting strategy would give preference to FragPipe for both datasets.

DISCUSSION

Application of four computational pipelines to the CPTAC UCEC phosphoproteomic data revealed a high-level of discrepancy among these pipelines. Although several studies have compared and evaluated different computational algorithms for phosphoproteomic data analysis using synthetic datasets (30, 53–55), it remains difficult to compare performance of different tools in actual application projects such as the CPTAC UCEC study. The key contribution of our work is the design and demonstration of deep-learning-derived metrics that can be used in actual application phosphoproteomic projects to directly evaluate performance of different computational pipelines.

Deep learning enables highly accurate prediction of many peptide properties (36). Because these properties are typically not used in the traditional computational algorithms for proteomic data analysis, they provide systematic and unbiased metrics for algorithm performance evaluation (37). Among the three metrics considered in this study, phosphosite probability prediction is independent of experimental condition. Predictions of RTs and MS/MS spectra depend on experimental condition, and such dependency was addressed by using experiment-specific data to fine-tune the universal base models for individual experiments through transfer learning. Previous studies used synthetic data to draw conclusions on the performance of computational algorithms. In contrast, we used synthetic data to draw conclusions on the performance of the benchmarking metrics,

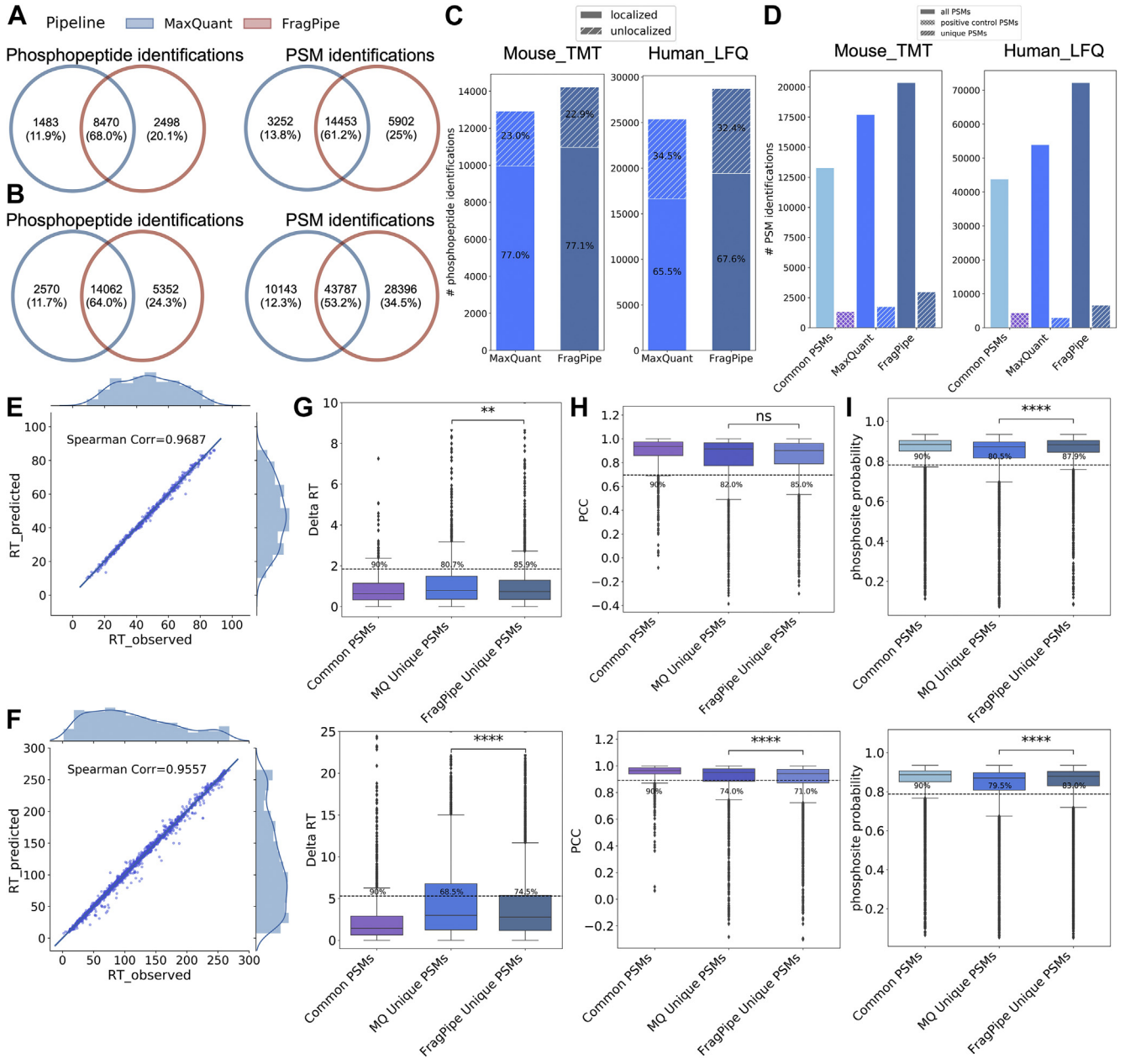


FIG. 5. **Benchmarking of MaxQuant and FragPipe using a TMT mouse dataset and a label free human dataset.** *A* and *B*, Venn diagrams of MaxQuant and FragPipe search results from one experiment in the mouse_TMT dataset (*A*) and one experiment in the human_LFQ dataset (*B*). *C*, proportions of localized and unlocalized phosphopeptides reported by the MaxQuant and FragPipe for the mouse_TMT and human_LFQ datasets, respectively. *D*, the numbers of all common PSMs, test PSMs used as positive controls in Delta RT-based benchmarking, and all PSMs and filtered unique PSMs identified by MaxQuant and FragPipe in the mouse_TMT and human_LFQ datasets, respectively (filtering method: peptide length ≤ 48 aa, no sequence overlaps in ground truth). *E* and *F*, scatter plots comparing predicted RTs and observed RTs from common PSMs in mouse_TMT and human_LFQ datasets, respectively. *G–I*, Delta RT (*G*), spectral similarity (*H*), and phosphosite probability (*I*), distributions for common PSMs and unique PSMs in mouse_TMT and human_LFQ datasets from MaxQuant and FragPipe. A horizontal baseline showing the lower 90% of Delta RTs (*G*) and the higher 90% of spectral similarity (*H*), phosphosite probability (*I*) of common PSMs. The ratios of PSMs with Delta RTs lower than the baseline (*G*) and those with spectral similarity (*H*) phosphosite probabilities (*I*), higher than the baseline in the pipeline-unique PSMs were labeled. ns: $p >$ ns: 5.00e-02 $< p \leq$ 1.00e+00; *: 1.00e-02 $< p \leq$ 5.00e-02; **: 1.00e-03 $< p \leq$ 1.00e-02; ***: 1.00e-04 $< p \leq$ 1.00e-03; ****: $p \leq$ 1.00e-04 (Wilcoxon rank sum test, Bonferroni correction). For boxplots, *centerline* indicates the median, box limits indicate upper and lower quartiles.

which enabled direct comparison of computational tools in actual application datasets. In our evaluation using synthetic data, both delta RT and spectra similarity showed excellent performance for discriminating correct PSMs from incorrect PSMs.

In three independent datasets including both TMT and label-free data from mouse and human studies, we found that FragPipe achieved higher sensitivity and quality compared with the other pipelines. These results encourage broader adoption of the relatively new FragPipe in future phosphoproteomic studies. MaxQuant identified fewer PSMs than FragPipe, but it outperformed MS-GF+/Ascore and CDAP and features a user-friendly interface. Both MS-GF+/Ascore and CDAP use MS-GF+ for phosphopeptide identification, and MS-GF+ has not been updated for 5 years, which may contribute to the relatively poor performance. Interestingly, although CDAP used the most conservative phosphorylation probability cutoff in site localization, which filtered out 56% phosphopeptides, the overall PSM quality of CDAP was the lowest among the four pipelines, likely due to the less stringently filtered phosphopeptide identifications.

An obvious limitation of the study is that our benchmarking only included four pipelines among the many possible pipelines, and only one set of parameters was used for each pipeline. We note that the primary goal of the study is not to identify the best pipeline or the best parameter setting. Instead, we expect that the benchmarking method demonstrated in this study could enable researchers to perform similar comparisons of different pipelines of interest or to compare parameter settings of the same pipeline by themselves on their own studies. Although FragPipe showed the best performance is our analyses, we encourage readers to use our benchmarking method to choose the most appropriate pipelines for their datasets. Moreover, although the study focused on comparing different pipelines, our benchmarking strategy can also be used to evaluate different parameter settings of the same pipeline in order to optimize performance. For example, our results suggest that the performance of CDAP may be significantly improved by increasing the filtering stringency in phosphopeptide identification and reducing the filtering stringency in phosphosite localization. Hence, our benchmarking method may also help developers to improve computational pipelines.

This study demonstrated the utility of Delta RT and spectral similarity as effective metrics for systematic benchmarking of computational pipelines for phosphoproteomic data analysis. We recently showed that incorporating Delta RT and spectral similarity into the PSM rescoring algorithm Percolator can improve peptide identification for immunopeptidomic data (56). Therefore, in addition to serving as benchmarking metrics, we expect that these deep-learning-derived features may also be used directly to improve phosphopeptide identification and site localization algorithms in the future.

DATA AVAILABILITY

The phosphoproteomics data that support the findings of this study are available in PRIDE (https://www.ebi.ac.uk/pride/) with the identifier PXD015087 (50), PXD000138 (33), PXD007145 (52), PXD015284 (51). Preprepared data for AutoRT base model training and testing were downloaded from the GitHub website (https://github.com/bzhanglab/AutoRT/tree/master/example/data) (37). The phosphoproteomic data from uterine corpus endometrial carcinoma study (UCEC) were downloaded from the CPTAC data portal (https://cptac-data-portal.georgetown.edu/study-summary/S043) (10). All code used in this work is available at https://github.com/bzhanglab/Benchmark_Phospho_Identification.

Supplemental data—This article contains supplemental data.

Acknowledgments—This study was supported by the National Cancer Institute (NCI) CPTAC awards U24 CA210954 and U24 CA210955, the Cancer Prevention & Research Institutes of Texas (CPRIT) award RR160027, and funding from the McNair Medical Institute at The Robert and Janice McNair Foundation. B. Z. is a CPRIT Scholar in Cancer Research and a McNair scholar. Portions of the analysis was performed at the Environmental Molecular Sciences Laboratory (grid.436923.9), a U.S. Department of Energy National Scientific User Facility located at the Pacific Northwest National Laboratory operated under contract DE-AC05-76RL01830.

Author contributions—B. Z., B. W., and W. J. conceptualization; W.-F. Z. model improvement; W. J. investigation; W. J., K. L., F. d. V. L., J. M., N. J. E., V. A. P., A. I. N., and T. L. methodology; W. J. and B. Z. writing—original draft; A. I. N., V. A. P., and T. L. writing—review and editing.

Conflict of interest—The authors declare no competing interests.

Abbreviations—The abbreviations used are: AUROC, area under the receiver operating characteristics; CDAP, CPTAC common data analysis pipeline; COS, cosine similarity; CPTAC, Clinical Proteomic Tumor Analysis Consortium; FDR, false discovery rate; ICPC, International Cancer Proteogenome Consortium; KDT, Kendal rank correlation coefficient; LC, liquid chromatography; MAE, median absolute error; MS/MS, tandem mass spectrometry; NCE, normalized collision energy; NL, neutral loss; PCC, Pearson's correlation coefficient; PSM, peptide-spectrum match; PTM, posttranslational modification; RT, retention time; SA, spectral angle; SPC, Spearman's correlation coefficient; TMT, tandem mass tag; UCEC, uterine corpus endometrial carcinoma.

Received April 4, 2021, and in revised from, September 16, 2021
Published, MCPRO Papers in Press, November 1, 2021, https://doi.org/10.1016/j.mcpro.2021.100171

REFERENCES

1. Hunter, T. (1995) Protein kinases and phosphatases: The Yin and Yang of protein phosphorylation and signaling. *Cell* **80**, 225–236
2. Blume-Jensen, P., and Hunter, T. (2001) Oncogenic kinase signalling. *Nature* **411**, 355–365
3. Rikova, K., Guo, A., Zeng, Q., Possemato, A., Yu, J., Haack, H., Nardone, J., Lee, K., Reeves, C., Li, Y., Hu, Y., Tan, Z., Stokes, M., Sullivan, L., Mitchell, J., *et al*. (2007) Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell* **131**, 1190–1203
4. Zanivan, S., Meves, A., Behrendt, K., Schoof, E. M., Neilson, L. J., Cox, J., Tang, H. R., Kalna, G., van Ree, J. H., van Deursen, J. M., Trempus, C. S., Machesky, L. M., Linding, R., Wickstrom, S. A., Fassler, R., *et al*. (2013) In vivo SILAC-based proteomics reveals phosphoproteome changes during mouse skin carcinogenesis. *Cell Rep.* **3**, 552–566
5. Ficarro, S., McCleland, M., Stukenberg, P., Burke, D. J., Ross, M. M., Shabanowitz, J., Hunt, D. F., and White, F. M. (2002) Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nat. Biotechnol.* **20**, 301–305
6. Krug, K., Jaehnig, E. J., Satpathy, S., Blumenberg, L., Karpova, A., Anurag, M., Miles, G., Mertins, P., Geffen, Y., Tang, L. C., Heiman, D. I., Cao, S., Maruvka, Y. E., Lei, J. T., Huang, C., *et al*. (2020) Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. *Cell* **183**, 1436–1456.e31
7. Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., Wang, X., Qiao, J. W., Cao, S., Petralia, F., Kawaler, E., Mundt, F., Krug, K., Tu, Z., Lei, J. T., *et al*. (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55–62
8. Vasaikar, S., Huang, C., Wang, X., Petyuk, V. A., Savage, S. R., Wen, B., Dou, Y., Zhang, Y., Shi, Z., Arshad, O. A., Gritsenko, M. A., Zimmerman, L. J., McDermott, J. E., Clauss, T. R., Moore, R. J., *et al*. (2019) Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. *Cell* **177**, 1035–1049.e19
9. Zhang, H., Liu, T., Zhang, Z., Payne, S. H., Zhang, B., McDermott, J. E., Zhou, J.-Y., Petyuk, V. A., Chen, L., Ray, D., Sun, S., Yang, F., Chen, L., Wang, J., Shah, P., *et al*. (2016) Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell* **166**, 755–765
10. Dou, Y., Kawaler, E. A., Cui Zhou, D., Gritsenko, M. A., Huang, C., Blumenberg, L., Karpova, A., Petyuk, V. A., Savage, S. R., Satpathy, S., Liu, W., Wu, Y., Tsai, C. F., Wen, B., Li, Z., *et al*. (2020) Proteogenomic characterization of endometrial carcinoma. *Cell* **180**, 729–748.e26
11. Clark, D. J., Dhanasekaran, S. M., Petralia, F., Pan, J., Song, X., Hu, Y., da Veiga Leprevost, F., Reva, B., Lih, T.-S. M., Chang, H.-Y., Ma, W., Huang, C., Ricketts, C. J., Chen, L., Krek, A., *et al*. (2019) Integrated proteogenomic characterization of clear cell renal cell carcinoma. *Cell* **179**, 964–983.e31
12. Gillette, M. A., Satpathy, S., Cao, S., Dhanasekaran, S. M., Vasaikar, S. V., Krug, K., Petralia, F., Li, Y., Liang, W.-W., Reva, B., Krek, A., Ji, J., Song, X., Liu, W., Hong, R., *et al*. (2020) Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell* **182**, 200–225.e35
13. Satpathy, S., Jaehnig, E. J., Krug, K., Kim, B.-J., Saltzman, A. B., Chan, D. W., Holloway, K. R., Anurag, M., Huang, C., Singh, P., Gao, A., Namai, N., Dou, Y., Wen, B., Vasaikar, S. V., *et al*. (2020) Microscaled proteogenomic methods for precision oncology. *Nat. Commun.* **11**, 532
14. Huang, C., Chen, L., Savage, S. R., Eguez, R. V., Dou, Y., Li, Y., da Veiga Leprevost, F., Jaehnig, E. J., Lei, J. T., Wen, B., Schnaubelt, M., Krug, K., Song, X., Cieslik, M., Chang, H. Y., *et al*. (2021) Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. *Cancer Cell* **39**, 361–379.e16
15. Savage, S. R., and Zhang, B. (2020) Using phosphoproteomics data to understand cellular signaling: A comprehensive guide to bioinformatics resources. *Clin. Proteomics* **17**, 27
16. 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., Wang, Z. W., Chen, Z. L., Sun, R. X., Liu, T., Tan, G. M., *et al*. (2018) Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.4236
17. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372
18. Craig, R., and Beavis, R. C. (2004) TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466–1467
19. Kim, S., and Pevzner, P. A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277
20. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., and Nesvizhskii, A. I. (2017) MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520
21. Rudnick, P. A., Markey, S. P., Roth, J., Mirokhin, Y., Yan, X., Tchekhovskoi, D. V., Edwards, N. J., Thangudu, R. R., Ketchum, K. A., Kinsinger, C. R., Mesri, M., Rodriguez, H., and Stein, S. E. (2016) A description of the clinical proteomic tumor analysis consortium (CPTAC) common data analysis pipeline. *J. Proteome Res.* **15**, 1023–1032
22. Beausoleil, S. A., Villen, J., Gerber, S. A., Rush, J., and Gygi, S. P. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **24**, 1285–1292
23. Taus, T., Kocher, T., Pichler, P., Paschke, C., Schmidt, A., Henrich, C., and Mechtler, K. (2011) Universal and confident phosphorylation site localization using phosphoRS. *J. Proteome Res.* **10**, 5354–5362
24. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805
25. Savitski, M. M., Lemeer, S., Boesche, M., Lang, M., Mathieson, T., Bantscheff, M., and Kuster, B. (2011) Confident phosphorylation site localization using the Mascot Delta Score. *Mol. Cell. Proteomics* **10**. M110.003830
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., and He, S. M. (2018) pSite: Amino acid confidence evaluation for quality control of de novo peptide sequencing and modification site localization. *J. Proteome Res.* **17**, 119–128
27. Shteynberg, D. D., Deutsch, E. W., Campbell, D. S., Hoopmann, M. R., Kusebauch, U., Lee, D., Mendoza, L., Midha, M. K., Sun, Z., Whetton, A. D., and Moritz, R. L. (2019) PTMProphet: Fast and accurate mass modification localization for the trans-proteomic pipeline. *J. Proteome Res.* **18**, 4262–4272
28. Fermin, D., Walmsley, S. J., Gingras, A. C., Choi, H., and Nesvizhskii, A. I. (2013) LuciPHOr: Algorithm for phosphorylation site localization with false localization rate estimation using modified target-decoy approach. *Mol. Cell. Proteomics* **12**, 3409–3419
29. Norel, R., Rice, J. J., and Stolovitzky, G. (2011) The self-assessment trap: Can we all be better than average? *Mol. Syst. Biol.* **7**, 537
30. Locard-Paulet, M., Bouyssié, D., Froment, C., Burlet-Schiltz, O., and Jensen, L. J. (2020) Comparing 22 popular phosphoproteomics pipelines for peptide identification and site localization. *J. Proteome Res.* **19**, 1338–1345
31. Quandt, A., Espona, L., Balasko, A., Weisser, H., Brusniak, M.-Y., Kunszt, P., Aebersold, R., and Malmström, L. (2014) Using synthetic peptides to benchmark peptide identification software and search parameters for MS/MS data analysis. *EuPA Open Proteomics* **5**, 21–31
32. Kapp, E. A., Schütz, F., Connolly, L. M., Chakel, J. A., Meza, J. E., Miller, C. A., Fenyo, D., Eng, J. K., Adkins, J. N., Omenn, G. S., and Simpson, R. J. (2005) An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis. *Proteomics* **5**, 3475–3490
33. Marx, H., Lemeer, S., Schliep, J. E., Matheron, L., Mohammed, S., Cox, J., Mann, M., Heck, A. J., and Kuster, B. (2013) A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nat. Biotechnol.* **31**, 557–564
34. K. Dagda, R., Sultana, T., and Lyons-Weiler, J. (2010) Evaluation of the consensus of four peptide identification algorithms for tandem mass spectrometry based proteomics. *J. Proteomics Bioinform.* **3**, 39–47
35. Navarro, P., Kuharev, J., Gillet, L. C., Bernhardt, O. M., MacLean, B., Rost, H. L., Tate, S. A., Tsou, C. C., Reiter, L., Distler, U., Rosenberger, G., Perez-Riverol, Y., Nesvizhskii, A. I., Aebersold, R., and Tenzer, S. (2016) A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* **34**, 1130–1136
36. Wen, B., Zeng, W. F., Liao, Y., Shi, Z., Savage, S. R., Jiang, W., and Zhang, B. (2020) Deep learning in proteomics. *Proteomics* **20**, e1900335
37. Wen, B., Li, K., Zhang, Y., and Zhang, B. (2020) Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nat. Commun.* **11**, 1759

38. Gessulat, S., Schmidt, T., Zolg, D. P., Samaras, P., Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A., Reimer, U., Ehrlich, H. C., Aiche, S., Kuster, B., and Wilhelm, M. (2019) Prosit: Proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518
39. Maboudi Afkham, H., Qiu, X., The, M., and Kall, L. (2017) Uncertainty estimation of predictions of peptides' chromatographic retention times in shotgun proteomics. *Bioinformatics* **33**, 508–513
40. Guan, S., Moran, M. F., and Ma, B. (2019) Prediction of LC-MS/MS properties of peptides from sequence by deep learning. *Mol. Cell. Proteomics* **18**, 2099–2107
41. Tiwary, S., Levy, R., Gutenbrunner, P., Salinas Soto, F., Palaniappan, K. K., Deming, L., Berndl, M., Brant, A., Cimermancic, P., and Cox, J. (2019) High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat. Methods* **16**, 519–525
42. [preprint] Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L., and Degroeve, S. (2021) DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nat. Methods* **18**, 1363–1369
43. Zeng, W. F., Zhou, X. X., Zhou, W. J., Chi, H., Zhan, J., and He, S. M. (2019) MS/MS spectrum prediction for modified peptides using pDeep2 trained by transfer learning. *Anal. Chem.* **91**, 9724–9731
44. Wang, D., Liu, D., Yuchi, J., He, F., Jiang, Y., Cai, S., Li, J., and Xu, D. (2020) MusiteDeep: A deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Res.* **48**, W140–W146
45. 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
46. 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
47. da Veiga Leprevost, F., Haynes, S. E., Avtonomov, D. M., Chang, H. Y., Shanmugam, A. K., Mellacheruvu, D., Kong, A. T., and Nesvizhskii, A. I. (2020) Philosopher: A versatile toolkit for shotgun proteomics data analysis. *Nat. Methods* **17**, 869–870
48. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., *et al*. (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745
49. Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J., and Mann, M. (2018) BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat. Methods* **15**, 440–448
50. Bian, Y., Zheng, R., Bayer, F. P., Wong, C., Chang, Y. C., Meng, C., Zolg, D. P., Reinecke, M., Zecha, J., Wiechmann, S., Heinzlmeir, S., Scherr, J., Hemmer, B., Baynham, M., Gingras, A. C., *et al*. (2020) Robust, reproducible and quantitative analysis of thousands of proteomes by micro-flow LC-MS/MS. *Nat. Commun.* **11**, 157
51. Wiechmann, S., Saupp, E., Schilling, D., Heinzlmeir, S., Schneider, G., Schmid, R. M., Combs, S. E., Kuster, B., and Dobiasch, S. (2020) Radiosensitization by kinase inhibition revealed by phosphoproteomic analysis of pancreatic cancer cells. *Mol. Cell. Proteomics* **19**, 1649–1663
52. Hogrebe, A., von Stechow, L., Bekker-Jensen, D. B., Weinert, B. T., Kelstrup, C. D., and Olsen, J. V. (2018) Benchmarking common quantification strategies for large-scale phosphoproteomics. *Nat. Commun.* **9**, 1045
53. Hoopmann, M. R., Kusebauch, U., Palmblad, M., Bandeira, N., Shteynberg, D. D., He, L., Xia, B., Stoychev, S. H., Omenn, G. S., Weintraub, S. T., and Moritz, R. L. (2020) Insights from the first phosphopeptide challenge of the MS resource pillar of the HUPO human proteome project. *J. Proteome Res.* **19**, 4754–4765
54. Lee, D. C., Jones, A. R., and Hubbard, S. J. (2015) Computational phosphoproteomics: From identification to localization. *Proteomics* **15**, 950–963
55. Wiese, H., Kuhlmann, K., Wiese, S., Stoepel, N. S., Pawlas, M., Meyer, H. E., Stephan, C., Eisenacher, M., Drepper, F., and Warscheid, B. (2014) Comparison of alternative MS/MS and bioinformatics approaches for confident phosphorylation site localization. *J. Proteome Res.* **13**, 1128–1137
56. Li, K., Jain, A., Malovannaya, A., Wen, B., and Zhang, B. (2020) DeepRescore: Leveraging deep learning to improve peptide identification in immunopeptidomics. *Proteomics* **20**, e1900334