UniSpec: Deep Learning for Predicting the Full Range of Peptide Fragment Ion Series to Enhance the Proteomics Data Analysis Workflow

Joel Lapin[2,3], Xinjian Yan[1], and Qian Dong[1*]

[1]Mass Spectrometry Data Center, Biomolecular Measurement Division, National Institute of Standards and Technology, 100 Bureau Dr., Gaithersburg, Maryland 20899, United States

[2]Department of Physics, Georgetown University, Washington, DC 20057, United States

[3]Associate, Mass Spectrometry Data Center, Biomolecular Measurement Division, National Institute of Standards and Technology, 100 Bureau Dr., Gaithersburg, Maryland 20899, United States

*** Corresponding author: Email: qian.dong@nist.gov**; Tel: 301-975-2569

ABSTRACT

We present UniSpec, an attention-driven deep neural network designed to predict comprehensive collision-induced fragmentation spectra, thereby improving peptide identification in shotgun proteomics. Utilizing a training dataset of 1.8 million unique high-quality tandem mass spectra (MS2) from 0.8 million unique peptide ions, UniSpec learned with a peptide fragmentation dictionary encompassing 7919 fragment peaks. Among these, 5712 are neutral loss peaks, with 2310 corresponding to modification-specific neutral losses. Remarkably, UniSpec can predict 73% -77% of fragment intensities based on our NIST reference library spectra, a significant leap from the 35%-45% coverage of only b and y ions. Comparative studies with Prosit elucidate that while both models are strong at predicting their respective fragment ion series, UniSpec particularly shines in generating more complex MS2 spectra with diverse ion annotations. Integration of UniSpec's predictions into shotgun proteomics data analysis boosts the identification rate of tryptic peptides by 48% at a 1% false discovery rate (FDR) and 60% at a more confident 0.1% FDR. Using UniSpec's predicted in-silico spectral library, search results closely matched those from search engines and experimental spectral libraries used in peptide identification, highlighting its potential as a stand-alone identification tool. The source code and Python scripts are available on GitHub (https://github.com/usnistgov/UniSpec) and Zenodo (https://zenodo.org/records/10452792), and all datasets and analysis results generated in this work were deposited in Zenodo (https://zenodo.org/records/10052268).

INTRODUCTION

Tandem mass spectrometry (MS/MS) based proteomics enables high-throughput characterization and quantification of tens of thousands of peptides in a single experiment[1-2]. Despite the remarkable advancements in mass spectrometers over the past decade, automated data analysis software tools still struggle to keep pace with the increasingly complex field of proteomics research[3-4]. Usually, this process involves predicting theoretical spectra from known protein sequences present in complete proteome databases and then comparing these predictions with observed spectra. Such in-silico spectra provide broad coverage of peptide-spectrum matches (PSMs) but may lack more distinctive features, such as fragment intensities or non-canonical fragment ions[5-6], that can help mitigate inevitable false positive matches. Spectral library searching involves comparing query spectra to previously identified spectra in the libraries using observed fragment intensities, facilitating the use of more sensitive and fast spectral search algorithms[7-8]. However, the method is limited to observed peptides whose spectra are in the library. Furthermore, the development and application of large-scale spectral libraries suitable for proteomic studies have been a challenging undertaking[9].

Machine learning (ML) based MS/MS prediction models have undergone a series of technological advances in recent years[10-11], and the proliferation of high-throughput peptide MS/MS data in modern shotgun proteomics has had an invaluable impact on their developments. Significant correlations between fragment spectral properties and peptide sequences are the impetus for the adoption of ML techniques for spectral prediction. Prototype ML-based models include MS2PIP[12], utilizing a tree-based xgboost model, and PeptideArt[13], with a shallow two-layer feed-forward neural network. However, as data volume and complexity grow, deep learning (DL), i.e., the application of much larger neural networks with many more parameters, is the next frontier. The strength of DL lies in its powerful ability to decode complex fragmentation patterns from huge spectral data sets, thereby eliminating the need for specialized feature engineering[14]. The advent of DL has redefined fragment intensity prediction, demonstrating the potential of rescoring PSM identifications in database search engines[15-17]. It also

greatly facilitates library searching[18-20], enabling the construction of in-silico spectral libraries of any target peptide, complementing or even replacing experimental libraries.

In the literature, two types of deep learning models for peptide spectrum prediction have been reported. The more common approach, represented by pDeep[21], Prosit[22], AlphaPeptDeep[23], and DeepDIA[18], is to predict the intensities of fragment ion series. These models predict MS2 peak intensities by pre-defining 100 to 600 unique canonical ion categories, such as b_2, y_3, b_2^{2+}, y_3^{2+}, etc. These fragment ions are then assigned theoretical mass-to-charge ratio (*m/z)* values for the specific peptide being predicted. These models require their training spectra to include annotations for the ions they predict. In contrast to predicting intensities for predefined fragment ions with accurate *m/z* data, PredFull[24] predicts fixed-width *m/z* bins with a spectral dispersion of 0.1 Da, which represents MS2 spectra at low resolution[25]. Although the prospect of full spectrum prediction without needing peak annotations for training is alluring, high-resolution fragment analysis is becoming a requirement for confident peptide identifications and detection of polymorphisms and post-translational modifications in proteins[25].

Here, we present UniSpec (Universal Spectra), a novel deep learning model conceptualized to universally predict all detectable fragments series in higher-energy collisional dissociation (HCD) spectra, which is a beam-type collision-induced dissociation (CID) technique specific to the orbitrap mass spectrometers. The model outputs spectra with 7919 intensity predictions, based on extensive MS2 peak annotations in the peptide spectral libraries developed by the National Institute of Standards and Technology (NIST, peptide.nist.gov). We used UniSpec to conduct PSM rescoring and spectral library search and found that UniSpec effectively improved the rate and reliability of peptide identifications, validating that the model is indeed able to effectively simulate the fragmentation spectra of complex peptides in real proteomics studies.

METHODS

Data Preprocessing and Datasets. Detailed preprocessing methods are described in Document S1. Briefly, we utilized publicly available NIST high-resolution mass spectral libraries (peptide.nist.gov). All spectra used in this study are fully annotated and come from "selected" libraries, which contain spectra with the highest scores for given peptide sequences and collision energies (CE) that have multiple identifications. All these spectra report CE in terms of normalized collision energy (NCE) and applied electron volts (eV), the latter being used as model input because it was found to be more reliable than NCE for learning fragmentation patterns and generating MS2 spectra (Document S1). If eV is not provided, our code automatically converts from NCE to eV based on the instrument type. Observed CE differences between different instrument types contained in spectral libraries are estimated by obtaining best-matching CE offsets, which were used as inputs for CE adjustments. Additionally, we removed 13% of low-confidence spectra based on quality metrics described in Document S1.

We compiled a cohesive spectral data collection tailored for model development, consisting of 1.8 million unique tandem mass spectra of approximately 0.8 million unique peptides. The CEs of these spectra cover > 10 observed NCE settings (Figure S1). Seven modifications were included: cysteine carbamidomethylation, methionine oxidation, N-terminal pyroglutamate from glutamine, N-terminal pyroglutamate from glutamate, N-terminal acetylation, pyro-carbamidomethyl N-terminal cysteine, and serine/threonine/tyrosine phosphorylation. In dataset splitting, sequences in the validation and test sets were compared to the training set (and to each other) using the

normalized Levenshtein distance score, any sequences with a score below 70 were considered to be significantly different from the training sequences. Overall statistics for all datasets are provided in Table S1. Most spectra involved charge states from 2+ to 4+ (maximum 8+), CEs from 15 to 70 eV (maximum 100 eV), and peptides from 6 to 30 residues (maximum 40).

High-Resolution HCD Fragmentation Dictionary. We developed an ion dictionary to serve as the model's output space (Table S2). The construction process, which surveys the entirety of the training data, involves the permutation of a, b, y, and precursor ion series with charges ranging from 1+ to 5+, up to 5 isotope peaks per ion, 23 observed neutral losses (NL), and up to 39-mer fragments. All ion series except a-ions include NL. Beyond the permutation list, we added all observed internal ions, immonium ions, and other miscellaneous annotated fragments. Amongst all candidate ions, only those that occurred at least 100 times were selected, which represents 99.8% of all annotated fragments in the training data set. The effect of this cutoff on the dictionary size and coverage of the raw datasets is visualized in Document S2. This process built a dictionary of a total of 7919 fragment peaks, including a-ions (226), b-ions (2026), y-ions (3643), precursor ions (131), internal ions (1836), and immonium ions and fragment ions arising from side chain cleavage (57). Among these, 5712 are neutral loss peaks, and 2310 of them are modification-specific neutral losses.

Model. This study uses an attention-based model, similar to the encoder structure of the transformer architecture[27]. The transformer was originally developed for natural language processing (NLP), which treats words or characters as separate tokens. Deep learning problems that work with amino acid sequences draw many parallels with NLP, including the inputs and tensor shapes that are operated on, and thus techniques from NLP can be transferred over to this problem. The model embeds amino acid sequence, modifications, charge, and eV into a single input tensor, and outputs a one-dimensional vector of prediction intensities for the 7919 fragment peaks in our dictionary. Specific details on the UniSpec model design and implementation can be found in Document S3.

Loss and Evaluation Metric. We calculate the model's training loss using the negative cosine similarity between prediction vector x and target vector y. For any observed spectrum, the target vector is constructed by placing the observed intensity into a 7919 dictionary-based vector. This is done only for peaks whose annotations exist in our ion dictionary, all other peaks are ignored for training purposes. The loss varies from -1 to 0, where -1 means x and y are identical in their corresponding peak intensities. We also take the square root of all spectral intensities before calculating cosine similarity, tempering the effect of highly intense peaks. This spectral transformation is a common practice in MS/MS data analysis[28].

In the post-training evaluation of the model, we evaluate the similarity between the observed and predicted spectra, both represented as variable length peak lists of m/z and intensity. Peaks in both spectra were aligned and matched by their *m/z* and specified mass accuracy tolerances in ppm (further details of peak matching are described in Document S4). We define the evaluation metric in Equation 1 as the cosine similarity score (CSS).

$$CSS = \frac{\Sigma_i x_i^{TP} y_i^{TP}}{\|x\|\|y\|} \qquad \text{Equation 1}$$

The numerator is the dot product of all matched peak intensities between the two vectors x and y, i.e., the sum of the product of matched peaks. The denominator is the product of each vector's L2 norm, which is the square root of the sum of the squares of all intensities in the vector, whether matched or unmatched. True positives (TP) refer to all the matched peaks between the two vectors. Predicted peaks that do not match any observed peaks are categorized as false positives (FP), and unmatched observed peaks are labeled as false negatives (FN). When evaluating annotated spectral data, such as our validation or test set, we removed all precursor (retaining its NL peaks) and unannotated peaks from both target and predicted spectra before scoring CSS.

RESULTS

Evaluation of the Model's In-Sample and Out-of-Sample Performance. A schematic representation of our UniSpec model, from the creation of training datasets and fragmentation dictionary to the implementation of model architecture, is shown in Document S5.

To evaluate the accuracy of the model's predictions, we compared its out-of-sample performance on testing data with the in-sample performance on training data. We predicted spectra from TestCommon (i.e., 49700 test spectra whose peptide sequences were included in the training data) and TestUniq (i.e., 6106 test spectra highly dissimilar with the training sequences and spectra) and calculated their CSS. Figure 1a shows that overall predictions on TestCommon yield a median CSS of 0.951, while predictions on TestUniq yield a median CSS of 0.923, which was less than 3% of the training accuracy, assuaging concerns about overfitting. We also extended this analysis, splitting the CSS results across different charge states and sequence lengths (Figure S2). In all out-of-sample cases, the median CSS remained within a 3% margin of the in-sample accuracy, except for highly charged > 5+ peptides and peptides ranging in size from 36 to 40 residues, where the difference was less than 5% and 4%, respectively. Notably, these highly charged and longer peptides yield the greatest reduction in CSS and appear to be the primary challenge for UniSpec's predictions.

Based on our previous assessments[29-31] of the effect of CE consistency on fragmentation patterns and model performance, we made CE adjustments for the datasets and used eV as the model input (Document S1). Here, we demonstrated in Figure 1b that UniSpec, trained with CE-consistent spectral data, achieved good performance and high agreement (median CSS of 0.92) over a wide range of eVs (15 to 74) derived from more than 10 different observed NCE settings of 25 to 41.

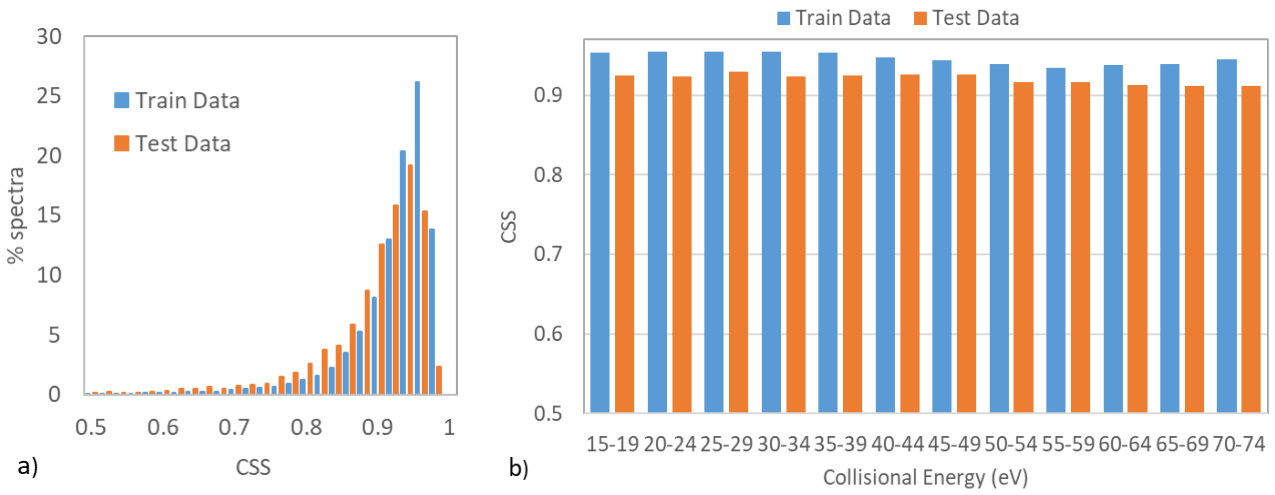

Figure 1. Comparing CSS on testing data to training data. (a) UniSpec's performance on the test set, (b) CSS distribution over collision energies from eV of 15 to 74 that derived from NCE of 25 to 41.

Evaluation by Comparison with Prosit[22]. To directly compare the capabilities of UniSpec with existing models, we selected Prosit, a widely used model in the field. Two datasets were utilized, with one being our test dataset (Set 1) and the other being Prosit's internal dataset, the ProteomeTools HCD Spectral Library at NCE 25 and 35 (Set 2) available at https://www.proteometools.org/index.php?id=53. Set 2 resembles Prosit's training spectra, which contain only b and y fragment ions in largely synthetic peptide spectra. We conducted two comparative analyses using our CSS metrics, namely: a) scoring only b and y ions for both datasets and b) scoring all ions (only on Set 1). To ensure an equitable comparison, we limited the scored peptides to align with Prosit's specified peptide input range: peptide lengths of 7-30, charge states of 1-6, and NCE values between 10-50.

Figure 2 displays the complementary cumulative distribution function (1 - CDF) of the model's CSS. Figure 2a illustrates that UniSpec achieves marginally higher CSS than Prosit across both Set 1 (4919 spectra) and Set 2 (NCE 25: 477416 spectra and NCE 35: 476185 spectra) when scoring only b and y ions. The disparity between the two models for b and y ions, though, is relatively trivial; both models have high CSS, especially on the ProteomeTools library. While performance is equally good on b and y ions, UniSpec has a definitive edge when additional ion annotations are factored into the CSS. In Figure 2b, 94% of the UniSpec predicted CSS exceeded 0.8, with a median of 0.931; only 5% of the Prosit spectra had a higher score at the same CSS cutoff, with a median of 0.640. Further investigation of the two predictions revealed that Prosit's CSS was reduced by 0.286 on average compared to UniSpec. This is primarily due to the missing of common non-canonical ions in its predictions, especially the neutral loss peaks, which highlights the importance of the extended ion predictions. Examples are given in Figure S3.

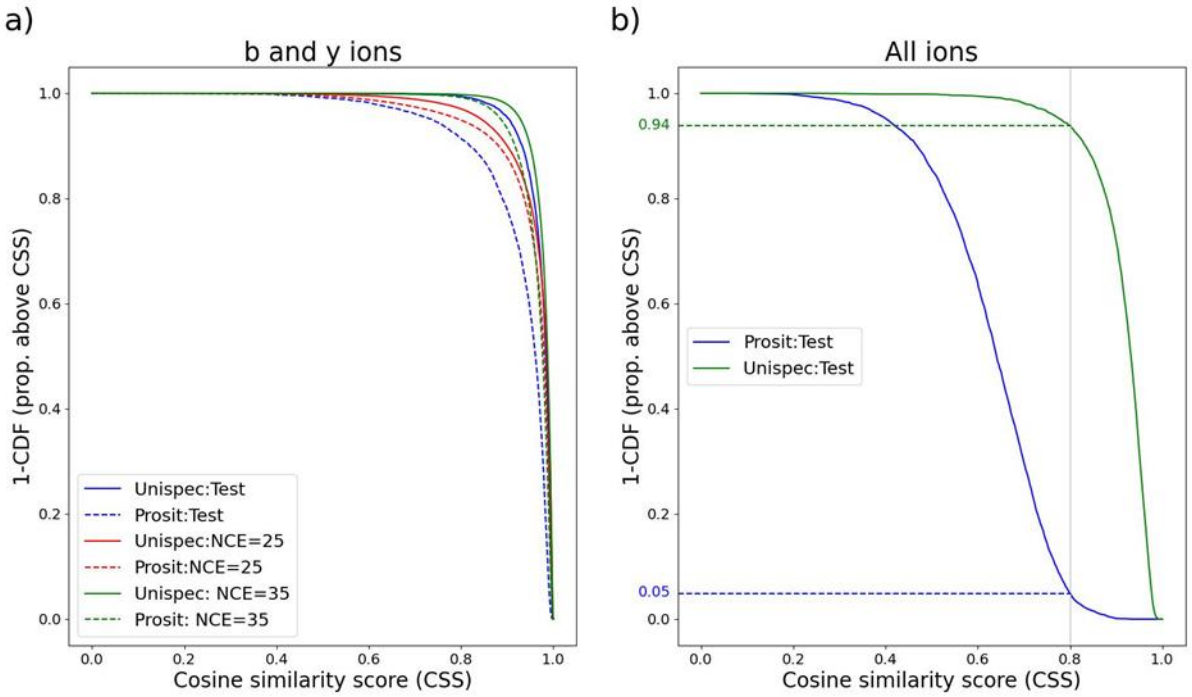

Figure 2. Complementary cumulative distribution functions (1-CDF) of CSS for Prosit and UniSpec. The y-axis denotes the proportion of predictions above the CSS in the x-axis. (a) Predictions of respective models on Set 1 and Set 2 for b and y ions only. Solid lines are the CSS for UniSpec, while dashed lines are for Prosit. (b) CSS performance of the respective models on all ions (excluding unannotated and precursor peaks) for our test set.

Evaluation by Measuring CSS for Different HCD Fragment Ion Series. Existing DL models for mass spectrometry mainly predict the b and y ion series. However, alternate fragmentation mechanisms produce a range of different fragment ions. An analysis of the HCD MS2 spectra from the NIST Human Spectral Library (peptide.nist.gov) reveals the widespread presence of many other fragment types in the experimental spectra, including neutral losses, a-type ions, internal ions, and more. Some precursor ions remain in fragment spectra due to incomplete dissociation, and there are also unknown ion signals. A visual comparison in Figure S4 shows that the ion types 2 to 6 accounted for an average of 29.3% of the total spectral ion intensity (including unannotated signals) in unmodified peptides (a), increasing to 34.6% and 42.8% in oxidatively modified (b) and phosphorylated (c) peptides, respectively. In particular, we observed a significant increase in the neutral loss contribution from 10.6% for unmodified peptides to 20.7% for oxidized peptides to 30.6% for phosphorylated peptides. This suggests that non-canonical ion types, particularly neutral losses, offer vital information about the HCD fragmentation spectra of modified peptides. Thus, predicting all HCD fragments could enhance the reliability of identifying complex biological peptides.

To measure the CSS of various HCD fragment ion predictions on the test set, we divided the dictionary peaks into four subsets, where each subset expanded on the previous one. Subset 1 contains 239 b and y ions for peptide lengths up to 40 and charge states up to 8+. Subset 2 adds a-type ions, precursor ions, and neutral losses, totaling 2176 ions. Subset 3 has 3143 ions, integrating internal and other ion types, while Subset 4, with all 7919 dictionary peaks, includes 4776 isotope peaks.

Isotope peaks of fragment ions are clearly distinguished in high-resolution mass spectrometry spectra[32]. As an ion increases in mass, its higher mass isotopes become more intense, even more so than its monoisotopic peak, leading to difficulties in correctly detecting monoisotopic peaks. UniSpec's prediction of the isotopic peaks for MS2 fragments allows accurate determination of monoisotopic masses using isotopic distributions.

The mean CSS for each subset of the dictionary is shown in Table S3. Subset 1 has the highest CSS of 0.969. As more diverse ions are introduced in subsequent subsets, CSS reduces, with Subset 4 dropping to 0.928, indicating the challenge of predicting isotopic contributions. Nevertheless, the results show the model retains a relatively high level of performance when the ion types are greatly increased, indicating its ability to predict intensities for a wide range of HCD fragment ions.

As a demonstration of UniSpec's broad ion prediction capability, we show in Figure S5 an example of a covalently modified peptide from our test set with predominant oxidation-specific neutral losses (CSS of 0.923). UniSpec correctly predicted the two highly selective cleavages of the N-terminal side to proline residues, resulting in the two dominant fragments, y_5 at *m/z* 588.2924 and a characteristic neutral loss from y_{10}. Indeed, for this doubly charged precursor ion with an oxidized methionine, the model successfully generates (under low proton mobility in this case) all major neutral losses of methane sulfenic acids (CH_3SOH, 64 Da) from the side chain of methionine sulfoxide residues.

Application of UniSpec in PSM Rescoring of Search Engine Results. In this section, we assessed the ability of UniSpec in PSM rescoring against Prosit. Our data analysis workflow (Document S6), built on Percolator from the Crux Mass Spectrometry Toolkit[33-34], rescores Tide search engine results based on fragment ion intensity predictions. Here, Percolator differentiates between true and false positives by representing each PSM with features and assigns a statistical confidence q value. Using UniSpec and Prosit, MS2 spectra were predicted for all target and decoy PSMs obtained in a Tide search. Predictions for the former were done in-house, while predictions for the latter were obtained from the Prosit prediction service website (https://www.proteomicsdb.org/prosit/). New features for rescoring include the cosine similarity score (CSS) and the fractions of matched ion peaks and intensities in the experimental spectra (Table S4). They were combined with the 18 default features for input to Percolator. The methods employed in the calculation of MS2 predictions and PSM features for the rescoring comparisons between UniSpec, Prosit, and Tide are outlined in Table S5. With UniSpec predictions, one feature set, UniSpec-by, was derived only from b- and y-type ions used by Prosit for comparison with the same fragment coverage, and the other, UniSpec-all, was derived from all ions predicted by UniSpec.

We applied our workflow to the public yeast LC/MS dataset PXD028735[35] to study the effects of UniSpec MS2 prediction rescoring. Table 1 indicates a substantial rise in PSM identification rate and unique peptides with prediction-based rescoring versus search engine rescoring at 0.1% FDR (0.001 q-value) and 1% FDR (0.01 q-value). Compared with Tide, UniSpec-all improves the PSM identification rate by 47.6% at 1% FDR and 59.5% at a more confident 0.1% FDR, which is the largest improvement among the three methods. Similarly, UniSpec-all increased unique peptides by 39.3% and 49.8% at FDRs of 1% and 0.1%, respectively, outperforming UniSpec-by and Prosit. Furthermore, Figures 3a and 3b showed that all prediction-based rescoring not only improved Tide's baseline identification rate at other q-value cutoffs but also achieved better PSM results at the tighter 0.1% FDR compared to those at Tide's 1% FDR (Also see Figure S6 for increased peptide identification

rates). The rescoring results demonstrated the significantly improved sensitivity and reliability of the identifications based on the integration of UniSpec fragment ion intensity predictions.

Table 1. Number of PSMs and unique peptides identified by different rescoring methods on yeast dataset PXD028735.

Rescoring methods	Identified PSMs				Identified Peptides			
	0.1% FDR	% incr.	1% FDR	% incr.	0.1% FDR	% incr.	1% FDR	% incr.
Tide	22322		29037		14798		18775	
Prosit	30236	35.5	41823	44.0	19458	31.5	25685	36.8
UniSpec-by	32908	47.4	42304	45.7	20959	41.6	25935	38.1
UniSpec-all	35609	59.5	42866	47.6	22174	49.8	26160	39.3

Note: %incr. percentage increase

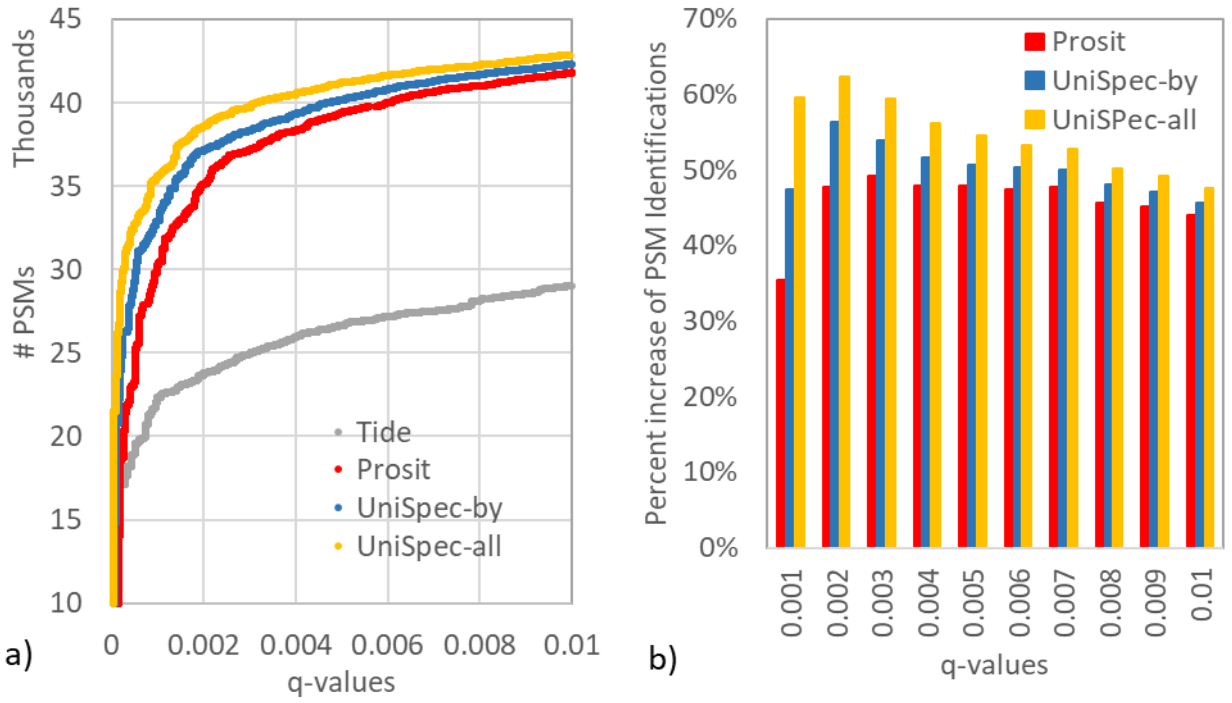

Figure 3. (a) PSM identification rates were obtained *before* and *after* prediction-based rescoring for various q-values. (b) The percentage of increases by Prosit, UniSpec-by, and UniSpec-all as compared to Tide. Note: *before* refers to tide + percolator rescoring; *after* refers to three rescoring methods, Prosit + Percolator, UniSpec-by + Percolator, UniSpec-all + Percolator.

The above substantial increase in identifications validated our extended features of cosine similarity score and fractions of observed ion peaks and intensities that are matched with prediction. These features provided greater separation between target and decoy matches, as is evident when comparing Percolator scores of Tide versus Prosit, UniSpec-by, and UniSpec-all rescoring results (Figure S7). After the prediction-based rescoring, the scores of target and decoy PSMs distinctly diverge, confirming the effectiveness of our features. While experimenting

with individual features, we initially found that using only CSS to represent the predicted features improved identifications by about 20%. This number doubled when adding ion fractions calculated from matched ion peaks and intensities, resulting in an identification increase of more than 40% at the 1% FDR PSM level (Figure S8). We proved that the combination of CSS with various fractions of predicted ion peaks and intensities can better characterize MS2 predictions, allowing Percolator to achieve greater rescoring performance.

To discern differences between rescoring using UniSpec and Prosit, we analyzed the top five peptide ions from both models based on Percolator scoring (Table 2). UniSpec's highest scores were mostly 3+ oxidized peptides, while Prosit's were primarily 2+ unmodified, with one exception being cysteine carbamidomethylation. On average, peptides from UniSpec had a size of 28 amino acids and yielded 213 observed fragment peaks; peptides from Prosit measured 22 amino acids with 58 observed fragment peaks. The top five Percolator scores from UniSpec ranged from 9.4 to 10.6, while for Prosit, these scores were between 3.4 and 4.1. The highest-scored peptides from Prosit received Percolator scores of 4.4 to 4.7, but their counterparts scored between 5.4 and 7.2 (Table 2). Figure S9 compares the top-scoring peptide spectrum predicted by UniSpec with that predicted by Prosit. Although both models accurately predict b and y ions, UniSpec can predict important additional ion intensities, including methionine oxidation-specific losses and other ion types, with CSS 65% better than Prosit. A similar comparison of the top-ranking peptide spectrum predicted by Prosit can be found in Figure S10. These comparisons based on top Percolator scores were not intended to characterize the overall quality of UniSpec or Prosit predictions, as that would require in-depth analysis involving many factors, but rather to highlight how UniSpec's comprehensive fragment ion predictions for real-world complex peptides result in better Percolator scoring outcomes. Overall, The DL fragment ion intensity predictions not only improved PSM identification rates but also bolstered the reliability of identified PSMs. Such advancements pave the way for more accurate and efficient proteomics data analysis.

Table 2. Comparison of the characteristics of the top five peptide ions identified by UniSpec (yellow) and Prosit (red), respectively, based on Percolator scores.

DL model	spectrum precursor *m/z*	top ranking peptides	z	mods	Percolator score		observed spectrum peaks
					UniSpec	Prosit	
UniSpec	1045.8358	SADEVINMANDSEYGLAAGIHTSNINTALK	3	MetOX	**10.64**	4.12	153
UniSpec	1020.8052	VSLDDLQQSIEEDEDHVQSTDIAAMQK	3	MetOX	**9.79**	3.51	415
UniSpec	1102.7757	DYEEELANDQEEEEGQGEGHENQSQEQR	3		**9.72**	3.77	142
UniSpec	1002.5044	APEALFHPSVLGLESAGIDQTTYNSIMK	3	MetOX	**9.59**	3.81	208
UniSpec	1063.2113	IALSRPNVEVVALNDPFITNDYAAYMFK	3	MetOX	**9.42**	3.37	147
Prosit	1333.1207	AEQGEHDENISPAQAAELVGEDLSR	2		6.85	**4.76**	65
Prosit	1014.0074	NVIAETGAGQHGVATATACAK	2	CysCAM	7.21	**4.44**	85
Prosit	1022.5079	SISYNPSQHSVLVNEANGK	2		5.40	**4.44**	40
Prosit	1233.1462	KPADLASLLLNSAGDAQGDEAPALK	2		6.76	**4.41**	53
Prosit	1230.101	VNGQLWDLDRPFEGEANEEIK	2		5.43	**4.40**	49

Note: MetOX, oxidized methionine; CysCAM, carbamidomethylated cysteine; z, charge states; mods, modifications.

Benchmarking Proteome-Scale In-Silico Spectral Libraries on a HeLa Dataset. Given the high similarity between the UniSpec predicted and observed spectra, we now investigate the feasibility of using in-silico libraries to identify peptides from real shotgun proteomics experiments, as this is another mainstream deep learning application.

Utilizing the in-silico human spectral library generated by UniSpec (Document S7), we analyzed high mass accuracy peptide spectra from a human HeLa dataset (PXD022287)[36], encompassing 107972 spectra obtained with an Orbitrap Fusion mass spectrometer. The HeLa dataset was reanalyzed with MSPepSearch against the UniSpec predicted library (UniSpec-Pred, 6247620 spectra) and the NIST human peptide spectral library (NIST-Exp, 911783 spectra, peptide.nist.gov). Our analysis included four separate searches: two for UniSpec-Pred (target and decoy) and two for NIST-Exp (target and decoy), and calculated q-values using the refined separated target/decoy approach[37-38]. These searches were benchmarked against MS-GF+[39] results and executed with settings akin to the original analysis[36], which yielded 75196 target PSMs, as described in Document S8.

Figures S11 (a) and (b) illustrate MSPepSearch scores for target and decoy PSMs from NIST-Exp and UniSpec-Pred, respectively. The distinct bimodal distribution of target matches compared to their decoy counterparts suggests that the generation of both decoy libraries was unbiased. Notably, NIST-Exp target PSM scores are higher than UniSpec-Pred, which is partly attributed to the library search algorithm that counts all MS/MS ion signals, including unannotated peaks that are not present in the predictions. Nevertheless, both approaches yielded consistent FDR estimates, evident from the overlapping curves in (c). This was a strong indication of UniSpec-Pred successfully modeling the patterns learned from the experimental mass spectra.

Our results in Fig. S11c and Fig. 4 indicate that UniSpec-Pred performance closely matches that of MS-GF+ and NIST-Exp. At a q-value cutoff of 0.01, MS-GF+, NIST-Exp, and UniSpec-Pred identified 43589 (58%), 44935 (60%), and 45036 (60%) PSMs, respectively, out of the total 75196. The three methods achieved high agreement in identifying >85% (39049) identical PSMs and >88% (25661) same peptides, as shown in Fig. 4. In addition to the common identifications by all three, we further investigated other peptides discovered by individual methods alone, mainly 1488, 178, and 287 peptides uniquely identified by MS-GF+, NIST-Exp, and UniSpec-Pred, respectively. Among them, 1250 (84%) of MS-GF+, 39 (22%) of NIST-Exp, and 175 (61%) of UniSpec-Pred were only represented in their respective methods. Most of the remaining peptides identified by each method can be found in the low-scoring regions below the q-value threshold of the other two methods. Not surprisingly, we found some representative cases where peptides identified by UniSpec-Pred have very high MSPepSearch scores but low MS-GF+ scores. Two such examples are provided in Figure S12, showing that MS-GF+ failed to identify peptides with complex dominant b-type or internal ions, whereas UniSpec predictions correctly identified these non-canonical ions or unusual patterns.

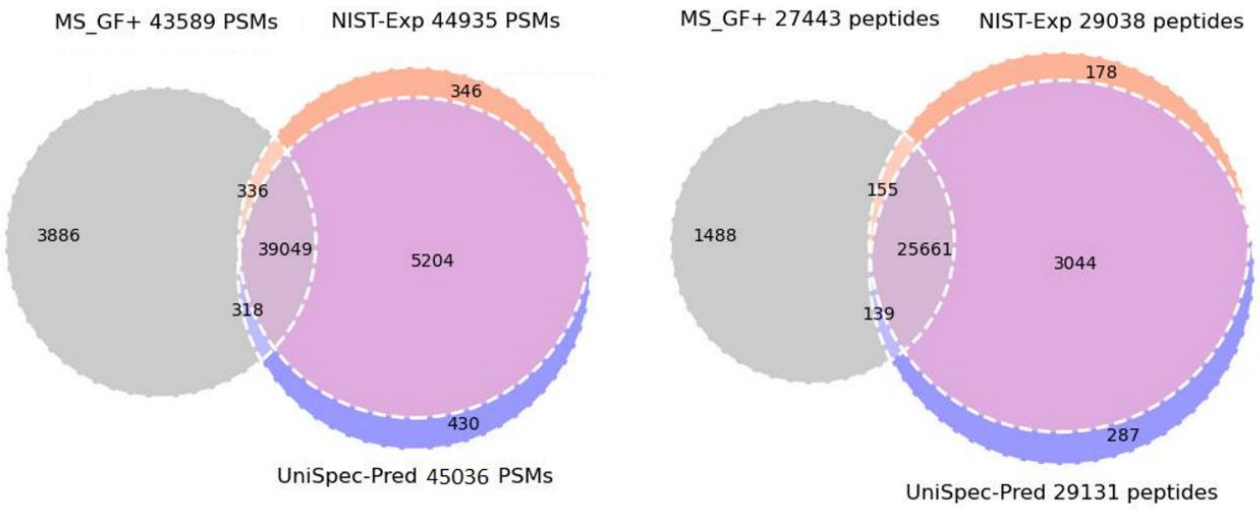

Figure 4 illustrates a performance comparison, at a q-value cutoff of 0.01, of the MS-GF+ database search and the MSPepSearch library search using the spectral libraries of NIST Human peptides and UniSpec predictions on the HeLa dataset. The same PSMs identified by three searches amount to 39049 (left panel), corresponding to 25661 unique peptides (right panel). Note that for ease of display, the regions of overlap for all 3 methods are depicted at 1/100th of their size proportional to their value compared to all other regions.

Here, we aimed to evaluate the practicality of UniSpec in real-world proteomics rather than a broad method comparison. Results indicate that UniSpec's predictions align with a major reference spectral library and renowned sequence database search engine. UniSpec effectively simulates observed ion intensity spectra, enhancing peptide identification, notably in complex peptide spectra.

CONCLUSION

UniSpec is a comprehensive DL predictor that can predict the intensity of the entire HCD MS/MS fragment ion series, going beyond existing tools limited to the b-ion and y-ion series. Key efforts include implementing an attention-based model similar to the Transformer encoder structure, data preprocessing of NIST reference library spectra suitable for model training, and building a peptide HCD fragmentation dictionary based on detailed MS2 peak annotations. Leveraging these innovations, UniSpec can correctly simulate the intensity patterns of a wide variety of fragment ions observed in large-scale shotgun proteomics experiments, ensuring reliable spectrum predictions.

Our evaluation in this paper provided an overall assessment of UniSpec, demonstrating its ability to correctly predict peak intensities of major canonical and minor fragment ions. In comparison with Prosit, both models excel at predicting the major b and y ion series in peptide spectra. However, UniSpec is trained on a diverse range of proteomic spectral data and an extensive MS2 fragmentation dictionary to closely model high-resolution experimental reference spectra. Its ability to predict highly charged, extended sequence fragments and large neutral losses in complex peptides enhances its utility in generating large-scale spectral libraries of entire proteomes.

While the current evaluation highlights the potential of UniSpec, there is still room for its improvement as an MS2 predictor. The training data could be extended to our high-quality datasets of peptides with higher charge (>=5+) and larger size (>=36 amino acids). Future efforts should also focus on extending its scope to non-tryptic peptides, to spectra chemically modified by tandem mass tag (TMT), and to more post-translational modifications. Considering UniSpec's impressive peptide identification performance, DL-predicted MS2 spectra and proteome-scale spectral libraries present immense promise for proteomics research.

Associated Content

Supporting Information:
Supplemental_1_documents.docx: Data strategy and deep learning datasets (Document S1); Ion dictionary size vs. occurrence cutoff (Document S2); Detailed Description of the Model Architecture (Document S3); Peak match criteria in the calculation of cosine similarity score (Document S4); Implementation of the UniSpec attention model architecture (Document S5); MS2 fragment intensity based PSM rescoring (Document S6); Generation of in-silico spectral libraries (Document S7); Data analysis and false discovery rate (FDR) estimate (Document S8). Supplemental_2_Figures.docx: Distribution of Collision Energy in the Training Dataset (Figure S1); Comparing CSS on Testing Data to Training Data (Figure S2); Comparison of Prosit and UniSpec Predictions and Their CSS on the Test Set (Figure S3); The contribution of each ion type to the overall intensity coverage in HCD spectra from the NIST experimental spectral library (Figure S4); An example of a covalently modified peptide from our test set with predominant oxidation-specific neutral losses (Figure S5); Comparison of Four Rescoring Methods by Increased Peptide Identification Rates (Figure S6); Percolator Scores for Target and Decoy PSMs (Figure S7); Comparison of Identified PSMs by Different UniSpec Rescoring Features (Figure S8); Observed and Predicted Spectra of a Top-Scoring Peptide in UniSpec (Figure S9); Observed and Predicted Spectra of a Top-Scoring Peptide in Prosit (Figure S10); MSPepSearch Score Distribution and the PSM Results for Three Different Search Methods (Figure S11) ; Two UniSpec Prediction Case Studies (Figure S12). Supplemental_3_Tables.xlsx: Statistics of the datasets used in developing UniSpec (Table S1); Content of high-resolution HCD fragmentation dictionary (Table S2a); Neutral losses are used in the dictionary (Table S2b); Commonly observed immonium ions and miscellaneous fragments relative to individual amino acid residues (Table S2c); Cosine similarity scores (CSS) are compared by four subsets of the ion dictionary (Table S3); Percolator default and extended features (Table S4); The methods employed in the MS2 predictions and PSM features for the rescoring comparisons between UniSpec+Percolator, Prosit+Percolator, and Tide+Percolator (Table S5).

ACKNOWLEDGMENTS

Special thanks to the reviewers for their valuable comments, which greatly helped improve the paper. The authors would also like to thank William Wallace, Karl Irikura, Trina Mouchahoir, and Arun Moorthy for their discussions on the manuscript. This work was supported solely with NIST funds. Certain commercial instruments or software are identified in this paper to specify the experimental procedure and data analysis workflow adequately. Such identification is not intended to imply endorsement by NIST, nor is it intended to imply that the materials or instruments identified are necessarily the best available for the purpose.

References

1. Cravatt, B. F.; Simon, G. M.; Yates, J. R., 3rd The biological impact of mass-spectrometry-based proteomics. *Nature*, **2007**, 450 (7172), 991–1000.
2. Michalski, A.; Cox, J.; Mann, M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J Proteome Res*. **2011**, 10 (4), 1785–1793.
3. Parker, R.; Tailor, A.; Peng, X.; Nicastri, A.; Zerweck, J.; Reimer, U.; Wenschuh, H.; Schnatbaum, K.; Ternette, N. The Choice of Search Engine Affects Sequencing Depth and HLA Class I Allele-Specific Peptide Repertoires. *Mol. Cell. Proteomics*, **2021**, 20, 100124.
4. Azevedo, R.; Jacquemin, C.; Villain, N.; Fenaille, F.; Lamari, F.; Becher, F. Mass Spectrometry for Neurobiomarker Discovery: The Relevance of Post-Translational Modifications. Cells, 2022, 11(8), 1279.
5. Park, C. Y.; Klammer, A. A.; Käll, L.; MacCoss; M. J.; Noble, W. S. Rapid and accurate peptide identification from tandem mass spectra. *J. Proteome Res.*, **2008**, 7 (7), 3022–3027.
6. Jones, A. R.; Siepen, J. A.; Hubbard, S. J.; Paton, N. W. Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics*, **2009**, 9 (5), 1220–1229.
7. Yen, C. Y.; Houel, S.; Ahn, N. G.; Old, W. M. Spectrum-to-spectrum searching using a proteome-wide spectral library. *Mol. Cell. Proteomics*, **2011**, 10 (7), M111.007666.
8. Griss J. Spectral library searching in proteomics. *Proteomics*, **2016**, 16 (5), 729–740.
9. Deutsch E. W. Tandem mass spectrometry spectral libraries and library searching. *Methods in molecular biology* (Clifton, N.J.), **2011**, 696, 225–232.
10. Cox, J. Prediction of peptide mass spectral libraries with machine learning. *Nat Biotechnol*. **2023**, 41 (1), 33-43.
11. Wen, B.; Zeng, W. F.; Liao, Y.; Shi, Z.; Savage, S. R.; Jiang, W.; Zhang, B. Deep Learning in Proteomics. *Proteomics.* **2020**, 20, 21-22.
12. Degroeve, S.; Martens, L. MS^{2}PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics*. **2013**, 29, 3199–3203
13. Arnold, R. J.; Jayasankar, N.; Aggarwal, D.; Tang, H.; Radivojac, P. A machine learning approach to predicting peptide fragmentation spectra. *Pac Symp Biocomput*. **2006**, 219-230.
14. Shinde, P. P.; Shah, S. A review of machine learning and deep learning applications. *Fourth international conference on computing communication control and automation (ICCUBEA)*, **2018**, (pp. 1-6). IEEE.
15. C, Silva; A. S.; Bouwmeester, R.; Martens, L.; Degroeve, S. Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions. *Bioinformatics* (Oxford, England), **2019**, 35 (24), 5243–5248.
16. Zolg, D. P.; Gessulat, S.; Paschke, C.; Graber, M.; Rathke-Kuhnert, M.; Seefried, F.; et al. INFERYS rescoring: boosting peptide identifications and scoring confidence of database search results. *Rapid Commun. Mass Spectrom.* **2021**, e9128.
17. Wilhelm, M.; Zolg, D. P.; Graber, M.; Gessulat, S.; Schmidt, T.; Schnatbaum, K.; Schwencke-Westphal, C.; Seifert, P.; de Andrade Krätzig, N.; Zerweck, J.; Knaute, T.; Bräunlein, E;, Samaras, P.; Lautenbacher, L.; Klaeger, S.; Wenschuh, H.; Rad, R.; Delanghe, B.; Huhmer, A.; Carr, S. A.; Kuster, B. Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nature communications*. **2021**, 12 (1), 3346.
18. Yang, Y.; Liu, X.; Shen, C.; Lin, Y.; Yang, P.; Qiao, L. In-silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nature communications*, **2020**, 11 (1), 146.
19. Van Puyvelde, B.; Willems, S.; Gabriels, R.; Daled, S.; De Clerck, L.; Vande Casteele, S.; Staes, A.; Impens, F.; Deforce, D.; Martens, L.; Degroeve, S.; Dhaenens, M. Removing the Hidden Data Dependency of DIA with Predicted Spectral Libraries. *Proteomics*, **2020**, 20 (3-4).

20. Searle, B. C.; Swearingen, K. E.; Barnes, C. A.; Schmidt, T.; Gessulat, S.; Küster, B.; Wilhelm, M. Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nature communications*, **2020**, 11 (1), 1548.
21. Zhou, X. X.; Zeng, W. F.; Chi, H.; Luo, C.; Liu, C.; Zhan, J.; He, S. M.; Zhang, Z. pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. *Anal. Chem.*, **2017**, 89(23), 12690–12697.
22. 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.; Wilhelm, M. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat Methods.* **2019**,*16* (6), 509-518.
23. Zeng, W. F.; Zhou, X. X.; Willems, S.; Ammar, C.; Wahle, M.; Bludau, I.; Voytik, E.; Strauss, M. T.; and Mann, M. AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics. *Nat. Commun.* **2022**, 13 (1), 7238.
24. Liu, K.; Li, S.; Wang, L.; Ye, Y.; Tang, H. Full-Spectrum Prediction of Peptides Tandem Mass Spectra using Deep Neural Network. *Anal Chem*. **2020**, 92 (6), 4275–4283.
25. Mann, M.; Kelleher, N. L. Precision proteomics: the case for high resolution and high mass accuracy. *Proceedings of the National Academy of Sciences of the United States of America*, **2008**, 105 (47), 18132–18138.
26. Navarro, G. A. guided tour to approximate string matching. *ACM Computing Surveys*, **2001**, *33* (1), 31–88.
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv Neural Inf Process Syst.*, **2017**, 5998--6008.
28. Stein, S.E.; Scott, D.R. Optimization and testing of mass spectral library search algorithms for compound identification, *J. Am. Soc. Spectrom.* **1994**, *5* (9), 859–866
29. Dong, Q.; Lapin, J.; Mak, T.; Slotta D.; Sheetlin, S.; Stein, S.; Wallace, W.; Geer, L. Increasing the spectrum prediction accuracy of machine learning models through improving collision energy consistencies in proteomic datasets, 69th ASMS Conference on Mass Spectrometry and Allied Topics, 2021, Philadelphia, Pennsylvania, October 31 - November 4, 2021.
30. Dong, Q.; Lapin, J.; Mak, T.; Slotta D.; Geer, L. Increasing the spectrum prediction accuracy of machine learning models through improving collision energy consistencies in proteomic datasets, 70th ASMS Conference on Mass Spectrometry and Allied Topics, 2022, Minneapolis, Minnesota, June 5-9.
31. Dong, Q.; Lapin, J.; Mak, T.; Slotta D.; Geer, L. In-depth Analysis of Collision Energy Inconsistencies in Large-scale Proteomics Datasets for Improving the Spectrum Prediction Accuracy of Deep Learning Models, The ACS Fall 2022 conference: Sustainability in a Changing World, 2022, Chicago, Illinois, August 21–25.
32. Goldfarb, D.; Lafferty, M. J.; Herring, L. E.; Wang, W.; Major, M. B. Approximating Isotope Distributions of Biomolecule Fragments. *ACS omega*, **2018**, 3 (9), 11383–11391.
33. Park, C. Y.; Klammer, A. A.; Käll, L.; MacCoss, M. J.; Noble, W. S. Rapid and accurate peptide identification from tandem mass spectra. *J. Proteome Res.*, **2008**, 7 (7), 3022–3027.
34. McIlwain, S.; Tamura, K.; Kertesz-Farkas, A.; Grant, C. E.; Diament, B.; Frewen, B.; Howbert, J. J.; Hoopmann, M. R.; Käll, L.; Eng, J. K.; MacCoss, M. J.; Noble, W. S. Crux: rapid open source protein tandem mass spectrometry analysis. *J. Proteome Res.*, **2014**, 13 (10), 4488–4491.
35. Van Puyvelde, B.; Daled, S.; Willems, S.; Gabriels, R.; Gonzalez de Peredo, A.; Chaoui, K.; Mouton-Barbosa, E.; Bouyssié, D.; Boonen, K.; Hughes, C. J.; Gethings, L. A.; Perez-Riverol, Y.; Bloomfield, N.; Tate, S.; Schiltz, O.; Martens, L.; Deforce, D.; Dhaenens, M. A comprehensive LFQ benchmark dataset on modern day acquisition strategies in proteomics. *Scientific data*, **2022**, *9* (1), 126.

36. Zeng, X.; Ma, B. MSTracer: A Machine Learning Software Tool for Peptide Feature Detection from Liquid Chromatography-Mass Spectrometry Data. *J Proteome Res*. **2021,** *20* (7), 3455-3462.
37. Jung, K. Statistical methods for proteomics. *Methods Mol Biol.* **2010**, 620, 497-507.
38. Navarro, P.; Vázquez, J.; A refined method to calculate false discovery rates for peptide identification using decoy databases. *J Proteome Res*. **2009**, *8* (4), 1792-6.
39. Kim, S.; Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun.*, **2014**, 5, 5277.

For Table of Contents Only