scientific reports

Check for updates

OPEN

End-to-end deep attention-based multitask pipeline for predicting uncertainty-quantified peptide properties from mass spectrometry data

Usman Tariq[1], Bilal Shabbir[1] & Fahad Saeed[1,2,3]✉

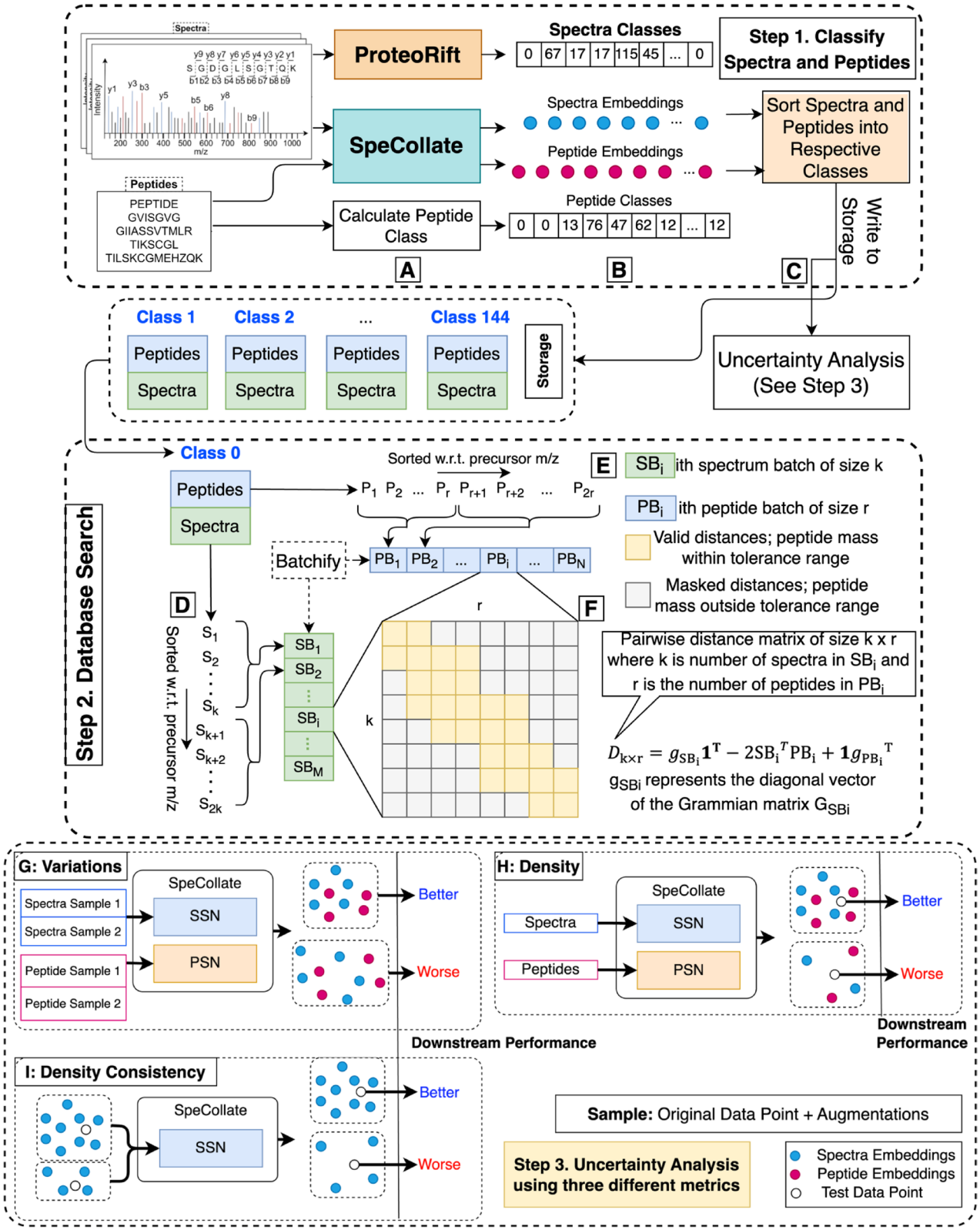
Mass-based filtering significantly reduces the peptide candidate pool for subsequent scoring in database search algorithms. While useful, filtering based on one property may lead to exclusion of non-abundant spectra and uncharacterized peptides – potentially exacerbating the *streetlight* effect. Here we present *ProteoRift*, a novel attention and multitask deep-network, which can *predict* multiple peptide properties (length, missed cleavages, and modification status) directly from spectra 77.8% of the time. Integrating ProteoRift into an end-to-end pipeline significantly reduces the search space compared to mass-only filtering. This delivers 8x to 12x speedups while maintaining peptide deduction accuracy comparable to established algorithmic techniques. We also developed uncertainty estimation metrics, which can distinguish between in-distribution and out-of-distribution data (ROC-AUC 0.99) and predict high-scoring mass spectra against the correct peptide (ROC-AUC 0.94). These models and metrics are integrated in an end-to-end pipeline available at https://github.com/pcdslab/ProteoRift.

Keywords Deep learning, Bioinformatics, Mass spectrometry, Uncertainty

Mass-based filtering is a critical scalability component in all existing database search methods for shotgun proteomics[1,2]. The predominant approach involves searching tandem mass spectrometry (MS/MS) spectra against a protein sequence database using a dedicated peptide identification algorithm[3–7]. Despite advancements in computational techniques, a significant proportion of spectra remain unidentified[7–9], and post-translational modification unaccounted. Recent calls[9–11] to develop better methods necessary to study non-abundant proteins associated with human aliments[12] including rare diseases[13], neurological disorders[14–16], and cancers[17,18]– highlights the importance of continued computational advancement for human health. Technological limitations have resulted in inaccuracies, including misidentification[6], search engine discrepancies[1,3,5,6,19–21], bias towards abundant peptides (the 'streetlight effect')[9,10,22], and dark matter[23] of proteomics.

Precursor mass filtering is a core computational step in mass spectrometry-based proteomics. It works by only comparing an experimental MS/MS spectrum against theoretical peptides whose calculated mass is within a given tolerance of the measured precursor mass. Currently all search-engines[5,6,24] inherently rely on mass-filtering based search-space restriction approach. While useful, this fundamental limitation necessitates either the expulsion of a large number of peptides from further analysis due to misassigned precursors[25–28]– or unreasonable search-times because of large databases[29–31]. Prior research[6,23,25,32–35], spanning more than a decade, has examined the impact of mass-filtering strategies on the number of peptide identifications. Performing search using mass filters with narrow-tolerance windows, while accurate[36], results in less than 50% of the data identified[25]. Open-search methods[25–28] have tried to solve this problem by using larger precursor searches which results in more identifications at the expense of lower sensitivity, and reduce scalability[1]. Despite these advancements, mass-based filtering may miss potential modifications arising from biological processes (unknown PTM's), chemical reactions[25](iron adducts, carbamylation), instrumental artifacts such as incorrect isotope selection[37], and labile modifications[8], all of which remain poorly understood. Existing algorithms,

[1]Knight Foundation School of Computing and Information Sciences, Florida International University (FIU), Miami, FL, USA. [2]Biomolecular Sciences Institute (BSI), Florida International University (FIU), Miami, FL, USA. [3]Department of Human and Molecular Genetics, Herbert Wertheim School of Medicine, Florida International University, Miami, FL, USA. ✉email: fsaeed@fiu.edu



using precursor and fragment ion masses, also result in an excessive number of candidate peptides, most of which do not get confidently matched especially for multiple non-model biological entities for meta-proteomics investigations[29,38–41].

Advances in machine-learning (ML) models have made it possible to develop more accurate and deeper pipelines for MS data analysis[3,19,42–45] which could outperform traditional (algorithmic) databases and denovo searches[19,46,47]. To this end, we recently introduced SpeCollate[3] which is a deep learning framework designed to improve peptide identification from mass spectrometry (MS) data by using a cross-modal similarity network. It encodes MS/MS spectra and peptide sequence information into embeddings to identify peptides. By using deep

◀**Fig. 1**. Complete ML database-search pipeline using ProteoRift and SpeCollate. Step 1 (**A**) Spectra and peptide embeddings are generated using SpeCollate while the ProteoRift predictive model assigns a class (out of 144) for each spectrum. (**B**) Class for a peptide can be calculated from the amino acid sequence. (**C**) Store spectra-peptide classes. Step 2: (**D**, **E**) Batchify spectra and peptides to fit them in server memory. (**F**) Calculate distance matrix of spectra and peptide embeddings in each batch. Step 3: (**G**) Per-feature sample variations and their correlation with downstream performance. A sample, the original data point plus augmentation, has lower feature variations when the middle is generally more confident about the original data point. (H) The density of training data around a given test data point indicates higher model confidence and hence improved downstream performance. (**I**) Density consistency measures how consistent the input data density is with the embedding density. This feature is important in estimating how well the model can capture the data structure and relations.

neural networks, SpeCollate learns complex relationships between the experimental MS spectra and theoretical peptide sequences, which allows it to match and identify peptides, even in the presence of noise or incomplete data. Regular usage of these ML models in systems biology labs continues to be challenging[48], influenced by multifaceted factors[49–51] such as black-box nature of most ML models, absence of confidence metrics for the deduced peptides, and lack of complete end-to-end pipelines. ML techniques that can improve on various components of the pipelines and introduce novel ways of computations, including new ways of reducing search-space, can increase the throughput and utility of the MS data. Some of those efforts include de novo tools like PepNet[52] to predict peptide length, and AHLF[53], capable of predicting phosphorylation status directly from spectra. These studies highlight the potential of learned spectral representations for peptide property estimation.

In this paper, we designed and developed a novel deep-learning model, called *ProteoRift*, to predict multiple peptide features (peptide length, missed cleavages, and modifications) which could potentially be used as filters for search space-reduction (Fig. 1, and Fig. 6). These predictions are established directly from spectra embedding – before any database-search. Our proposed network inputs mass to charge ratio (m/z), and intensity, and indexes values of spectra using the *attention network*. Precursor mass, modification status, and charge value of the spectra is used to predict all the features simultaneously using the *multitasking* technique. In contrast to existing methods[52,53] attention-based multitask networks provide significant advantages over independently trained solo predictors by leveraging shared knowledge across related tasks (like predicting peptide length and PTMs), pushing the models to learn robust and universal data representation. This act of joint learning across different states enables strong regularization and reduction in task-specific overfitting risk leading to potentially better generalization to new, unseen mass spectrometry data. Finally, predicting in a joint fashion using a multitask approach offers superior computational efficiency both for training and inference time for proteomics pipelines. Unlike many existing peptide property prediction tools, which are primarily designed for downstream applications such as spectral library generation, rescoring of peptide-spectrum matches, or post-search annotation. ProteoRift utilizes property prediction upstream of the database search as an explicit candidate-pruning strategy by estimating peptide length, missed cleavages, and modification status directly from spectra before exhaustive spectrum-peptide comparison. Therefore, the primary contribution of ProteoRift is demonstrating that predicted spectral properties can be used to reduce the candidate search space that is comparable to mass-based filtering[43,54–56]. Theoretical analysis shows that successfully applying filters reduces the search space size by more than 90% (Supplementary Fig. 1) which is then confirmed and demonstrated with experiments using real-world data (Figs. 2, 3, 4 and 5 and Supplementary Figs. 5, 6, 7). Our extensive experimentation with open data shows that ProteoRift can predict peptide lengths by up to 92% precision, missed cleavages by 97% precision, and modification status by up to 97% precision. ProteoRift is further integrated with the SpeCollate[3] ML workflow to perform peptide deductions experiments. As a consequence of more precise search-space reduction using other spectra properties, the resulting database-search engine exhibits, on average, 8x – 12x speedups while maintaining peptide accuracies at par with traditional search engines[5,6]. In addition, we developed three novel MS-specific metrics to quantify the uncertainty associated with spectra embeddings, and their inferred peptides using re-trained SpeCollate[3] model. Our results demonstrate that the proposed metrics can distinguish between *in-distribution* and *out-of-distribution* data (ROC-AUC 0.99) and predict high-scoring mass spectra against the correct peptide (ROC-AUC 0.94). These measures provide insight into the model's stability, confidence, and data representation ability, aiming to empower confident usage of ML tools in systems biology settings.

Results

Methods overview

ProteoRift is a deep attention-based multitask network that enables prediction of peptide properties directly from the spectra. To showcase the power of predictive analytics, we integrated two independently trained ML models (*ProteoRift*, and *SpeCollate*) to achieve an end-to-end database search pipeline shown in Fig. 1 which enables comparison of results directly to traditional algorithms such as Tide/Crux and MSFragger. In the first step, an attention-based network generates spectra, and peptide embeddings. Thereafter, a multi-task network is used to predict peptide-length, missed cleavages, and modifications status directly from spectra *before* any peptide deduction. After predicting the multi-task properties, we used the SpeCollate architecture to generate the final embeddings for both spectra and peptides in the latent space, which were later used for the database search. The SpeCollate model was fine-tuned to enhance the quality of its embedding representations and improve overall identification performance. The second step consists of binning spectra, and peptides using the predictions from the first step. These batches are then used for performing database-search for peptide deduction using spectra

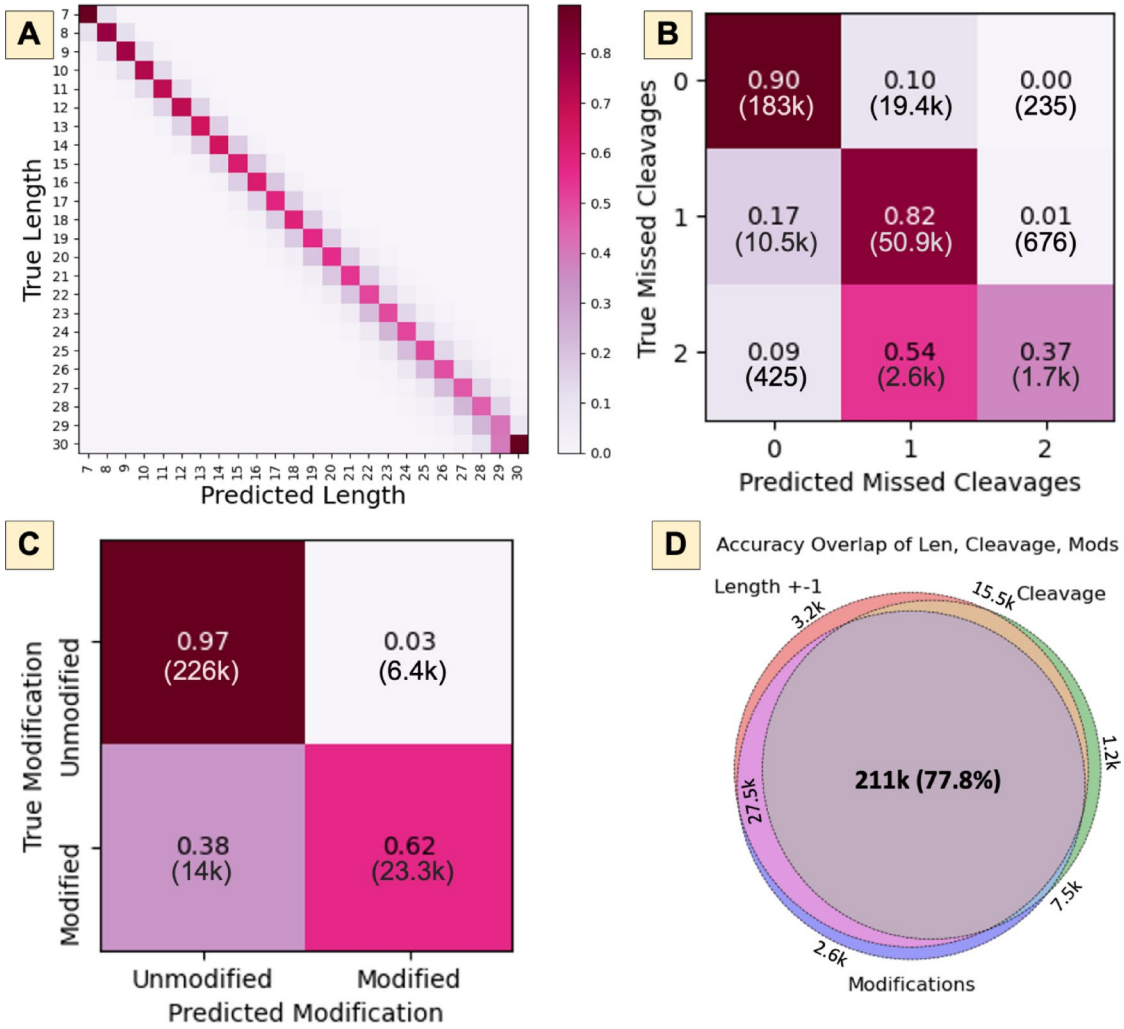


Fig. 2. Confusion matrix of peptide length missed cleavages, and modification status predictions for testing data sets. (**A**) Reasonable prediction recall is exhibited by the model for length between 7 and 30. When exact length is observed, the recall is high for shorter lengths. For larger length, which is a more computational harder problem, the model performance gradually reduces due to accumulated error of the MS peaks. However, when the length constraint is relaxed to ± 1 of the prediction, recall value is >90% is observed for most lengths. More info in Supplementary Fig. 5. (**B**) Prediction with high recall is observed for both commonly found real world data with 0 and 1 missed cleavages. (**C**) Extremely high recall of more than 97% are observed for large number of spectra/peptides for unmodified peptides. While pertaining to training data with small number of modified spectra, the recall of the model is still reasonable enough for predictions. More PTM specific data may result in better trained model. (**D**) The recall when all three heads are used for prediction simultaneously is illustrated in the Venn diagram. ProteoRift accurately predicts all three properties 77.8% of the time.

and peptide in single batch i.e., no intra-batch-deductions are performed. Provided that the first step resulted in a correct prediction, the peptides and the spectra that are *supposed* to match would be in the same batches. The third step is to develop novel metrics to quantify the uncertainty associated with spectra embeddings and peptide deduction. These confidence metrics are reported back with the results of the spectra-to-peptide matches. This pipeline is then used for all experimentation as well as comparison with existing methods. To the best of our knowledge, we demonstrate the first ML model that can predict peptide properties into the database search process, which can be used as an alternative to mass-filtering, directly from the spectra and enables more scalable, accurate, and trustworthy peptide deductions.

To gain insights into our predictive models we performed 3 distinct experiments. The first set of experiments were performed for evaluating the results of the predictive model i.e. how accurate is our developed *ProteoRift* model when trying to predict the length, missed cleavages, and modifications. The second set of experiments were performed for the database search that is batched according to the model predictions. This set of experiments informed us about the performance of the end-to-end pipeline as compared to existing algorithmic techniques, when these properties are used instead of a mass filter. Furthermore, these experiments inform us about the potential search-space, and processing time reduction. The last set of experiments are completed to assess the

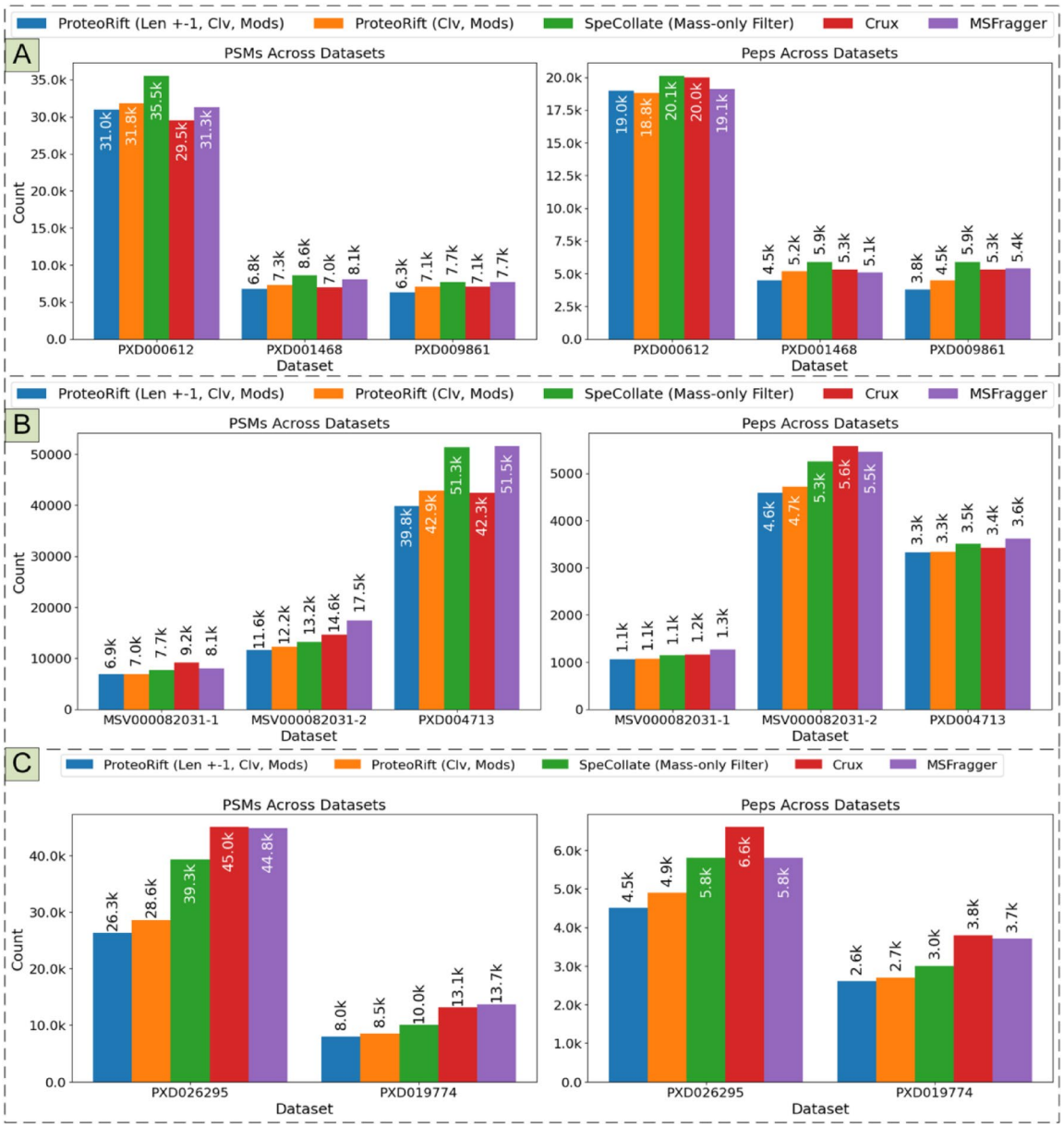


Fig. 3. Comparison of the number of PSMs and peptides identified under 1% FDR for ProteoRift with different paramteres. ProteoRift (Len+/−1, Clv, Mods) uses all three heads for prediction for Length (+/−1), modification status and missed cleavages, ProteoRift (Clv, Mods) only uses modification status and missed cleavages. The filtered spectra/peptide batches are then searched against the database using SpeCollate ML model. SpeCollate (mass-only), Crux and MSFragger use their respective mass filters. The search results are shown for: (**A**)PXD000612[52], PXD001468[23], and PXD009861[53] and searched against human proteome. (**B**) Meta-Proteomics database searches results using msv000082031 and PXD004713 dataset, and the RefUP + + database[57], encompassing 2,259 genera. (**C**) In addition, ProteoRift is evaluated using PXD019774[54] (Immunopeptides), and PXD026295[55]. The results demonstrate that ProteoRift can be used as predictive filtering technique which when combined with our SpeCollate model leads to PSMs and peptides identified under 1% FDR that are comparable to established algorithmic search-engines.

confidence of the ML database search embedding and apprises us about the confidence metrics when compared with the ground truth data sets.

Predictive performance evaluation of the *ProteoRift model*

On the test dataset, we measure the precision values of all classes for each head. To ensure rigorous evaluation, the datasets were partitioned into training, validation, and testing sets at the peptide sequence level, rather than at the spectral level. This peptide-centric splitting strategy guarantees the absence of overlapping peptide sequences across the respective subsets, thereby facilitating an unbiased assessment of model performance.

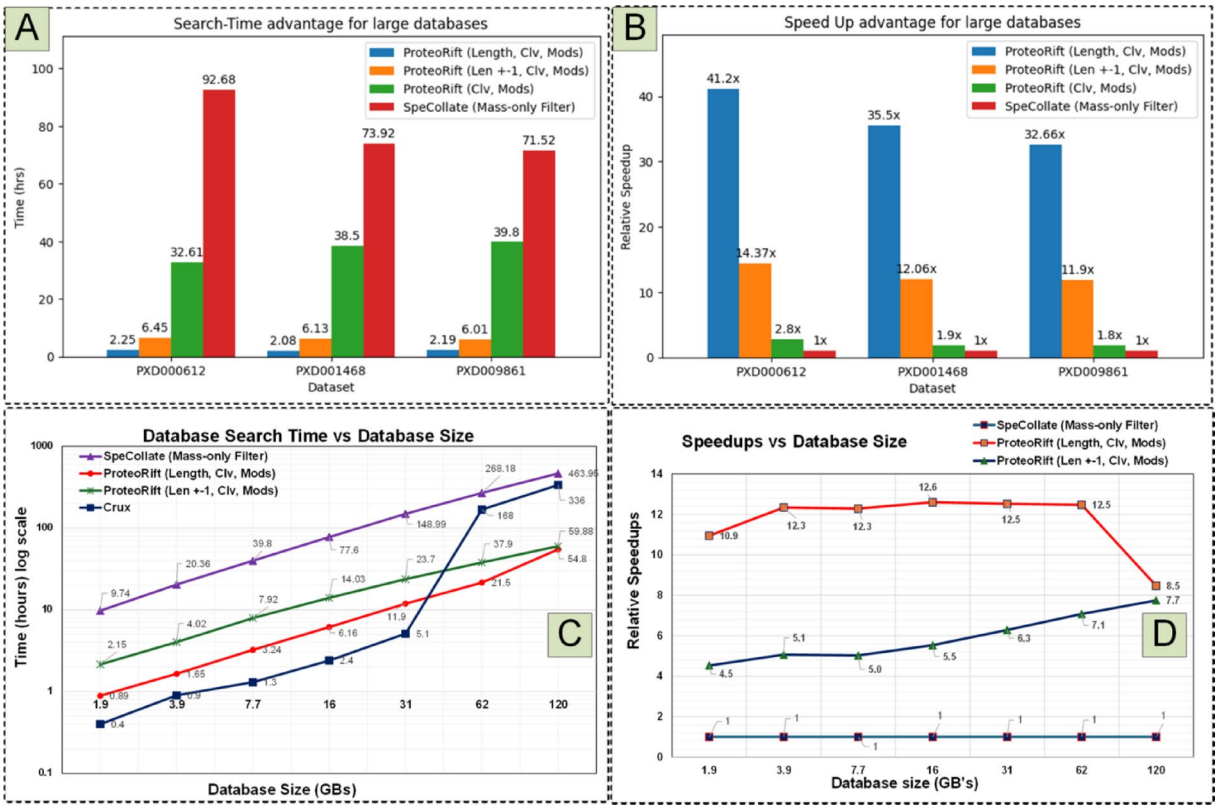


Fig. 4. Comparison of the search time and resulting speedup for ProteoRift with different parameters is shown. ProteoRift (Len+/−1, Clv, Mods) uses all three heads for prediction for Length (+/−1), modification status and missed cleavages, ProteoRift (Clv, Mods) only uses modification status and missed cleavages. The filtered spectra/peptide batches are then searched against the database using SpeCollate ML model. SpeCollate (mass-only), Crux and MSFragger use their respective mass filters. (**A** and **B**): Extreme reduction in database search time is observed when PXD00061252, PXD00146823, and PXD00986153 data sets were searched against proteogenomic database using six-frame translation (6FT) of the human genome (GRCh38.p14, hg38). When all filters are applied, we obtain up to 41x speedup while 14x speedup when enabling the length error margin of ± 1. (**C** and **D**): Time of execution and speedup analysis with increasing database size for a Meta-Proteomics database. The search is performed against a meta-proteomics database published by Proteostorm (RefUP + + database). The mass spectra from the msv000082031 spectral dataset were used. When all filters are applied, we obtain up to 12x speedup while 8x speedup when enabling the length error margin of ± 1.

Performance of length head

Our model predicts the length with high recall which gradually drops for large values when the length requirement is strictly enforced (Fig. 2A). This recall is slightly reduced with increasing length of the peptides due to accumulation of noisy MS peaks caused by a larger number of b-ion and y-ion fragments leading to overlapping peaks, or reduced number of samples of large length available for training. While the first bottleneck is not in our control, the second bottleneck is effectively alleviated by oversampling based on class count and the associated weights.The high accuracy between predicted and true lengths shows that the model has learned meaningful sequence embeddings that capture features relevant to peptide length. Our results show that the total recall value is >90% when the length requirement is relaxed to within ± 1 for the majority of the data sets (Supplementary Fig. 5) showcasing that the model has learned useful embeddings and that the head is sufficiently trained. Considering these results, we assert that $predicted\_length = actual\_length \pm 1$ to be accurate for any subsequent experiments and the model.

Performance of cleavage head

Our experiments also demonstrate that cleavage head predicts with high recall (Fig. 2B). As shown in the Figure, the 0 and 1 missed cleavages are predicted with 90% and 82% respectively. The recall value drop for 2 missed cleavages is expected due to a small number of training examples in our data set. The class counts determine the associated weights (Supplementary Fig. 3D&E). For our experiments we kept the ratio of 2 missed cleavages constant (~2% randomly selected) in the database for training. Even with added importance weight, the recall remains less than ideal but it does not seem to affect the search results substantially (Sect. 2.3) due to rarity of 2 missed cleavages in the real world[57].

(A) Uncertainty metrics performance on retrieving high-quality examples from in-distribution data.				
Score Type	***ROC-AUC***	***F1-Score***	***Precision***	***Recall***
Variation	0.84	0.77	0.75	0.79
Density	0.74	0.68	0.66	0.70
Consistency	0.72	0.67	0.65	0.69
Variation+Density	0.88	0.80	0.79	0.81
Variation+Consistency	0.84	0.74	0.81	0.68
Density+Consistency	0.83	0.76	0.78	0.74
Variation+Density+Consistency	0.94	0.87	0.88	0.86
(B) Uncertainty metrics performance on retrieving in-distribution data from a mix of in-distribution and out-of-distribution data				
Score Type	***ROC-AUC***	***F1-Score***	***Precision***	***Recall***
Variation	0.96	0.96	0.96	0.96
Density	0.85	0.75	0.80	0.71
Consistency	0.99	0.96	0.95	0.97
Variation+Density	0.91	0.84	0.83	0.85
Variation+Consistency	0.99	0.97	0.97	0.97
Density+Consistency	0.99	0.96	0.97	0.96
Variation+Density+Consistency	0.99	0.99	0.99	0.99
C) Confusion matrix for peptide identification accuracy of spectra predicted in-distribution vs out-of-distribution.				

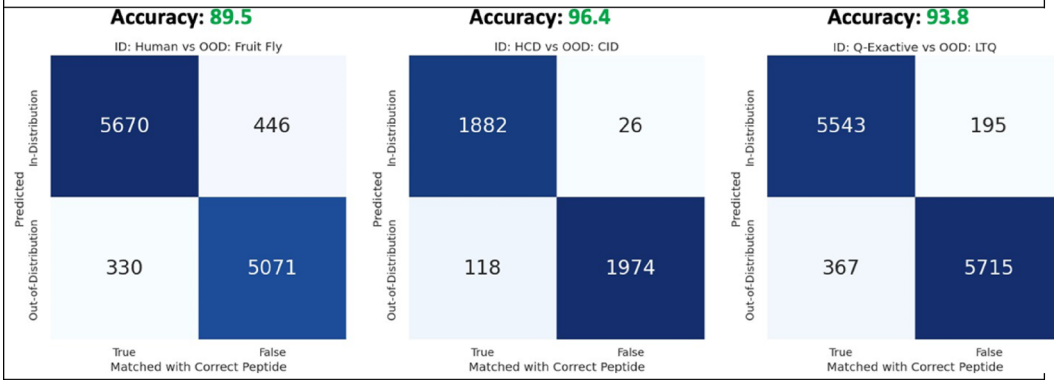


Fig. 5. (**A**) Performance of uncertainty metrics on retrieving high-quality examples from in-distribution data. The values indicate that combining multiple metrics improves the ability to detect high-quality examples significantly with the ROC-AUC value of 0.94. (**B**) Performance of uncertainty metrics on retrieving in-distribution data from a mix of in-distribution and out-of-distribution examples. The values indicate that density consistency is the major contributor towards distinguishing between the two classes with an ROC-AUC of 0.99. (**C**) Peptide identification accuracy of spectra predicted in-distribution vs. out-of-distribution. Three different combinations of in-distribution and out-of-distribution datasets are used to compare the downstream performance with the predicted accuracy. For Human vs. Fruitfly, we use datasets PXD009861 and PXD030281. For HCD vs. CID, we use ProteomeTools HCD and CID libraries. For Q-Exactive vs. LTQ, we use datasets PXD000612 and PXD000125. As can be seen in the figure above, the in-distribution data points can be predicted with all degree of accuracy for data sets with different species, fragmentation strategy, and the MS instrument – making these uncertainty metrics an efficacious tool for ML prediction, and deduction of MS based omics data sets.

Performance of modification-status head

Our experiments also demonstrate that the model predicts with exceptional performance (>97% recall) for unmodified spectra as shown in Fig. 2C. While the performance of the more difficult problem of identifying modification status is close to 62%, it is better than random chance. This behavior can be explained with a smaller number of training samples that are modified (372k) vs. unmodified (2.3e+06) which we deal with count/weighted strategy (Supplementary Fig. 3D&E). This behavior is expected since increasing number of modifications lead to significant changes in the spectral patterns. While the model did learn and performed better than random modification status on unseen data; more data may lead to better performance in the future.

Overall predictive performance

The average performance of three heads will determine the accuracy of the overall predictive model, and its effect on the search-space. We perform the aggregate performance evaluation of the model by measuring the total number of correct measurements for all heads. As shown in Fig. 2D, we can accurately predict all three properties 77.8% of the time. Considering these results, we will use the following combinations of parameters for ProteoRift for subsequent experiments: ProteoRift (Length +/−1, Mod, Cleavage) means usage of all filters while

relaxing the length to +/−1, ProteoRift (Mod, Cleavage) is equivalent to filtering without length as a parameter. While various iterations of the filtering is possible, we believe the current strategy leads to good estimation of the results for various data sets and parameters.

Effect of predicted search-space on peptide deduction completed using end-to-end database-search pipeline

These sets of experiments were completed to estimate the effect of the filters on the search-space as well as the accuracy of the peptides deduced using modified search-space. For our predictive filtering to be effective, the peptide deductions on modified search-spaces must be at par with other algorithmic techniques that use mass-only filtering. Evaluation in the previous section shows the performance of various predictive heads. To perform evaluation of the effect of the predictive filtering on peptide deduction quality, space-reduction, and execution time, we incorporate various permutations of the filter from most relaxed to most stringent. These filter configurations include: (1) Apply mass-only filters, (2) apply missed cleavages and modification filters, (3) all filters (length, missed cleavages, modification status) in conjunction with a relaxed length filter with a margin of ± 1. While various configurations are possible, we believe that this arrangement of experiments successfully demonstrates the range of deduction accuracies expected for a given dataset.

Evaluating generalizability

We collected Pride Archive (PXD) datasets, pre-processed and wrangled them into a suitable format for machine-learning workflows for our experimentation and evaluation of proteomics search: PXD000612[58], PXD001468[25], and PXD009861[59]. In addition, we have evaluated our strategy for PXD019774[60], PXD026295[61] and PXD010595[43] (tryptic peptides). When extending our experiments to Meta-Proteomics database searches, we use msv000082031 dataset. Proteome and MetaProteome database are downloaded from ProteoStrom Github (RefUP++database[62]) encompassing 2,259 genera, and Uniport proteome ID UP000005640 (H. sapiens). In addition, we used PXD013712 (Gammarus pulex) and PXD010712 (Drosophila melanogaster (fruit fly)) to assess out-of-distribution data performance. The results from these experiments ensure that the model generalizes well and is not only memorizing data. Below we compare ProteoRift as an alternative filtering technique to mass-only filtering and compare the peptide deduction results.

Performance on Proteomics data sets

Since ProteoRift can be integrated with any proteomics database search-engine, our goal is to thoroughly assess its effectiveness when deployed across multiple different experiments. For varying filter configurations, we evaluated three distinct datasets, namely PXD000612[58], PXD001468[25], and PXD009861[59]. In these trials, we consider two primary metrics: the number of peptides to spectrum matches (PSM) and identified peptides under a 1% False Discovery Rate (FDR). As shown in Fig. 3A, the PSM results with SpeCollate (mass-only filter) is better than both Crux and MSFragger for PXD000612 and PXD001468 and comparable to MSFragger for PXD009861. For PXD000612, Mod+cleavage filter performs better than both Crux and MSFragger for PSM and performs comparably for the other two data sets. When ProteoRift (Length +/−1, Cleavages, and Mod) are used, which is arguably the most stringent settings, the PSM identified are still comparable to all mass-only filter based search strategies, while exhibiting considerable speedups (Fig. 4A and B). The number of peptides identified when using mod+cleavage filters is comparable to MSFragger and Crux for all data sets.

Performance on Meta-proteomics data sets

When extending our experiments to Meta-Proteomics database searches, we use the RefUP++database[62], encompassing 2,259 genera. We analyze the number of peptides identified (at 1% FDR) with different configurations for the Meta-Proteomics database. Our results are shown in Fig. 3B, demonstrates that the number of peptide-to-spectrum matches, and peptides identified for msv000082031-1, msv000082031-2 and PXD004713. For PXD004713 the number of PSM's using filtering technique is comparable to Crux (42.9k vs. 42.3k) when only modification and cleavage predictive filtering is used. While the number of PSM is higher for MSFragger (51.5k), the number of unique peptides identified (3.6k) are comparable to both Crux and ProteoRift (3.4k vs. 3.3k) when predictive filtering is used. For msv000082031-1 the PSM (7.0k vs. 9.2k vs. 8.1k) and peptides (1.1k vs. 1.2k vs. 1.3k) for Crux and MSFragger but slightly reduced for ProteoRift. Venn diagrams (Supplementary Fig. 6 A, Supplementary Data File 3, Supplementary Data File 4) show significant overlap between the peptides identified using Crux, MSFragger, and ProteoRift (combined with SpeCollate). These results showcase, that using alternate filtering mechanism has still enabled ProteoRift/SpeCollate duo to identify large number of PSM and peptide identifications both overlapping other engines like Crux and MSFragger, and also few that are identified by the SpeCollate engine alone. We have also provided the Length Distribution (Supplementary Fig. 12), Length Density (Supplementary Fig. 13), and Entrapment Analysis (Supplementary Fig. 11) which showcases the reliability of the results. Lastly, these results are still very encouraging, considering that the number of peptides identified with filtered subsets is only slightly lower than with mass-only filtering, and ProteoRift can significantly reduce the search space while having a minimal impact on result quality – all while achieving impressive speedups.

Performance on additional datasets

We also evaluated our strategy for PXD019774[60], PXD010595[43] (tryptic peptides), and PXD026295[61] to ensure that ProteoRift is applicable to a wide range of MS omics scenarios.

PXD019774[60] employed mass spectrometry (MS)-based proteogenomic large-scale profiling to identify potential immunogenic human leukocyte antigen (HLA) Class I- associated peptides both in melanoma, as well as EGFR mutant lung adenocarcinoma. Our results shown in Fig. 3C shows that the number of peptides

identified using MSFragger and Crux are 3.7 K and 3.8 K and Specollate deducts 3 K peptides using mass only filters. ProteoRift identified 2.6 K peptides when considering ProteoRift (Len +/−1, Clv, Mods) filters, which is comparable to the 2.7 K peptides identified when ProteoRift (Clv, Mods) filters were applied. The Venn diagram and UpSet plot (Supplementary Fig. 6B, Supplementary Fig. 16a) indicates that the identified peptides for this data set has large overlap with peptides identified using Crux and MSFragger, and as expected, complete much larger overlap with SpeCollate results.

PXD010595[43] (tryptic peptides) is one of the largest and diverse data sets available, and for this reason results are depicted separately in Supplementary Fig. 7. These tryptic peptides are acquired by solid phase synthesis, and measured on an Orbitrap Fusion mass spectrometer. For each peptide pool, an inclusion list was generated to target peptides using five fragmentation methods (HCD, CID, ETD, EThCD, ETciD) with ion trap or Orbitrap readout. When the experiment was executed with a dataset with all fragmentation methods, the number of peptides identified using ProteoRift (Len +/−1, Clv, Mods) is lower than Crux or SpeCollate (Supplementary Fig. 7 A). However, when ProteoRift (Clv, Mods) is used the number improves (80.4k) and is comparable to the number of peptides identified using Crux (86.4k) while MSFragger (91.1 K) identifies more peptides. As Proteorift is trained on HCD fragmentation data, we further ran the experiment for PXD010595[43] using only HCD data. For HCD data, our results (Supplementary Fig. 7 A) demonstrate that the peptides identified using ProteoRift (Cleavage, Modification) are still comparable to Crux while MSFragger identifies more peptides. Venn diagrams and UpSet plot (Supplementary Fig. 7B, Supplementary Fig. 16f, Supplementary Data File 3, Supplementary Data File 4) for identified peptides show large numbers of peptides that are common between Curx, SpeCollate and ProteoRift. Surprisingly, we observe novel peptides when using ProteoRift, which are not identified by either Crux or SpeCollate. The number of peptides identified using SpeCollate is higher than Crux or ProteoRift in all cases.

On another data set PXD02629555, our results (Fig. 3C) indicate that SpeCollate with mass only filter exhibits the 39.3k PSM's, and Crux and MSFragger exhibit 45k and 44.8k PSMs respectively. While the number of PSM identified using ProteoRift are 28.6k with modification and cleavage filter and 26.3k when all filters are used. For PXD026295[55], our results (Fig. 3C) indicates that SpeCollate with mass only filter exhibits the comparable number of peptides (5.8k) with MSFragger (5.8k) while Crux deducts 6.6k peptides. The number of peptides identified using ProteoRift are comparable to SpeCollate and MSFragger both of which identified similar numbers (4.9k vs. 5.8k) showcasing the effectiveness of our approach. Venn diagram and UpSet plot (Supplementary Fig. 6B, Supplementary Fig. 16b) depicts a large number of peptides that are common between ProteoRift and mass-based filtering-based Crux, MSFragger and SpeCollate approaches. Interestingly there were number of peptides deduced using ProteoRift that were not identified by either Crux or SpeCollate – indicating our smart filtering technique may result in search-space that are missed by mass only filtering and may result in novel peptides in addition to being more efficient.

Evaluation of robustness to substantial dataset divergence

Our model performance evaluation strategy (Sect. 2.3, Fig. 2) was accomplished by partitioning at the peptide sequence level, rather than at the spectral level, resulting in no data overlap between the training and testing data sets, ensuring no data leakage.

(a) Robustness evaluated using divergent, out-of-training-distribution-set species.

To accomplish this, we performed experiments using PXD013712 (Gammarus pulex) and PXD010712 (Drosophila melanogaster (fruit fly)) to evaluate ProteoRift's performance on unseen and divergent data sets. For PXD013712 the same database as the original publication[63] was used with Deamidation, Carbamidomethylation and Oxidation as PTM search parameters. PXD010712[64] was searched using UniProt Taxon ID 7227 and no dynamic PTMs.

Results (Supplementary Fig. 9 A) on PXD013712 (Gammarus pulex) shows that the number of peptides identified using SpeCollate are 6.6k while Crux and MSFragger identified 9.7k and 8.5k. When subjected to additional filtering using ProteoRift, the number of peptides is less when more restrictive parameters are used ProteoRift (Len +/−1, Clv, Mod) but better in performance when this filter is relaxed to just cleavage and modifications (2.3k vs. 5.0k).

Results (Supplementary Fig. 9 A) for PXD010712 (Drosophila melanogaster (fruit fly)) shows that the number of peptides identified using SpeCollate(17.2k) is comparable to both Crux and MSFragger (17.8k and 17k). When subjected to additional filtering using ProteoRift, the number of peptides is 2.3k when more restrictive parameters are used ProteoRift (Len +/−1, Clv, Mod) and 7.8k when the filter is relaxed just to cleavages and modifications. More relaxation of the filter will result in a greater number of peptides and fine tuning to specific data sets will help in identifying more peptides.

Interestingly, Venn diagrams and UpSet plots (Supplementary Fig. 9B, Supplementary Fig. 16c, Supplementary Fig. 16 d Supplementary Data File 3, Supplementary Data File 4) for identified peptides show a large number of peptides that are common between Curx, MSFragger, SpeCollate and ProteoRift. In addition, maximal overlap between SpeCollate and ProteRift is observed (> 90% in most cases) since SpeCollate is used as the underlying search-engine as compared to Crux or MSFragger. SpeCollate[3] has been shown to identify a greater number of peptides, and a larger fraction of unique peptides that may get missed by traditional algorithmic techniques. These results also show that combination of algorithmic and ML models can increase the variety of peptides that may otherwise get missed when using a single technique for deductions.

(b) Robustness evaluated using divergent, out-of-training-set fragmentation methods.

ProteoRift is trained only on HCD data sets. In order to test the robustness and learnability of the model, we tested it on PXD010595[43] (tryptic peptides) measured on an Orbitrap Fusion mass spectrometer using five fragmentation methods (HCD, CID, ETD, EThCD, ETciD) with ion trap or Orbitrap readout. When the experiment was executed with a dataset with all fragmentation methods, the number of peptides identified using ProteoRift (Len +/−1, Clv, Mods) is lower than Crux or SpeCollate (Supplementary Fig. 7 A – left side, Supplementary Fig. 16e). However, when ProteoRift (Clv, Mods) is used the number improves (80.4k) and is comparable to the number of peptides identified using Crux (86.4k), but is lower than SpeCollate and MSFragger. This showcases that ProteoRift has learned useful embeddings and is robust enough to identify peptides when subjected to out-of-distribution data.

Remarks: These results demonstrate that ProteoRift can filter, with varying degrees of success, data from species and fragmentation that are much different than the training sets, showcasing the robustness of the model when subjected to out-of-distribution data. Supplementary Figure (6, 7, and 9) shows large overlap between ProteoRift and SpeCollate search results, and comparable overlaps with Curx and MSFragger. Our results also demonstrate that deduced peptides using ProteoRift results in consistent length distribution (Supplementary Fig. 12) and density (Supplementary Fig. 13) across multiple data sets. In addition, average peptide lengths are found to be nearly identical for both matched and mismatched spectra ID/peptide ID pairs (Supplementary Fig. 14) demonstrating ProteoRift's consistent performance across different lengths and various data sets. Missed cleavage distribution (Supplementary Data File 6 and https://osf.io/tk2r6) was also found to be consistent for ProteoRift across different data sets and tools (Supplementary Fig. 15). Some unique and novel peptides are also identified (Peptides with 1% FDR: https://osf.io/2sndq, Peptides with 5% FDR: https://osf.io/vq4ny, Unique Proteorift Peptides: Supplementary Data File 5 and https://osf.io/ke9dy, Mismatch Analysis: https://osf.io/t3dch) which showcases an interesting property of alternate-filtering model, and is subject of future investigations.

Reduction in search-space and speed up analysis

After training the model, we wanted to establish if predictive filtering effectively reduced the search-space when deducing peptide using database-search engines. To accomplish this, we performed two separate experiments. In the first experiment, we created a proteogenomic database using six-frame translation (6FT) of the human genome (GRCh38.p14, hg38, NCBI RefSeq accessed Feb 2022). Thereafter, PXD000612[58], PXD001468[25], and PXD009861[59] data sets were searched against this proteogenomic database. As compared to mass-only strategy, ProteoRift takes a fraction of the time while exhibiting more than 41x speedups (Fig. 4A, B). Such a large database may be useful for MS omics of non-model organisms and this experiment demonstrates that ProteoRift is able to reduce the search-space even when most of the database entries are not relevant. We performed the second experiment against a meta-proteomics database published by Proteostorm. The mass spectra used were from the msv000082031 spectral dataset. For the Meta-Proteomics search experiments, the database size varies from 2 GB to 120 GB. As shown in Fig. 4C and D, the speedup after using filters increases gradually as the database size increases. This is likely due to the reduced number of out-of-core computations when additional filters are applied leading to a more optimal memory access pattern. In general, our experiments demonstrate that when all filters are applied, we obtain up to 8x to 12x speedups. Speedups of up to 7.7x are obtained when enabling the length error margin of ± 1. This is predominantly attributed to a significant reduction in out-of-core computations, a frequent computational bottleneck in large-scale database searches. When the size of the database is small, Crux which is a highly optimized code base results in smaller execution times (Fig. 4C). However, the time it takes for Crux increases rapidly especially when the database size approaches physical memory of the machine. One could argue that having better disk-based access to Crux will increase the scalability of the tool for meta-proteomics data sets. However, this increased execution time is due to increase in the candidate peptides, as compared to smarter/predictive filters, and demonstrates that mass-based filtering is an inefficient way of reducing search space. Further, increase in candidate peptide is not specific to Crux and the same characteristics are observed for SpeCollate (Fig. 4A). Models such as ProteoRift can predict other properties of the data and significantly reduce the search space even when the database sizes are large, and with comparable accuracy to what one might expect from other serial algorithms. These results show that predictive filtering *does* result in smaller search-space and better execution times especially for larger databases. Crux and SpeCollate are used for a fair comparison.

Remarks: In addition to speeding up the search process, narrower space sliced by ProteoRift reduces the likelihood of high-scoring mismatches and may be interpreted differently between different tools. To ensure that these matches were in fact correct, further analysis was performed by calculating the total number of peptides and their scan IDs identified by each tool, identified unique scan ID's that are found only in one tool's results, and lastly the percentage of shared scan IDs between tools that are assigned to the same peptide. All of these results are completed using 1% false discovery rate (FDR). Based on our analysis, when the same spectrum ID appears in both SpeCollate and ProteoRift, it matches the same peptide on average 98% of cases (Supplementary Fig. 10), with a small number of false positives (Supplementary Table 3). The results were further confirmed by using FDRBench[65] which shows a low entrapment rate of 0.82% and 1.17% for PXD019774 and PXD000612 respectively(Supplementary Fig. 11), and demonstrates the overall reliability and confidence in our tools. Supplementary data is available as mismatch analysis: https://osf.io/t3dch and entrapment analysis: https://osf.io/xhztm, https://osf.io/6s9kc.

Predictive power of confidence and uncertainty metrics and their correlation with real-world datasets

For our experiments, different combinations of in-distribution and out-of-distribution datasets are used to compare the downstream performance with the predicted accuracy. The three data sets include estimation of different species (Human vs. Fruitfly, datasets PXD009861 and PXD030281), fragmentation strategies

(ProteomeTools HCD and CID libraries), and finally different instrument (Q-Exactive vs. LTQ, we use datasets PXD000612 and PXD000125).

Aleatoric uncertainty
Detailed analysis of the metrics performance is given in Fig. 5A. As can be seen in the table, the per-sample feature variation (variation) metric yielded a promising ROC-AUC score of 0.84, surpassing the other individual metrics. This finding illustrates that the model's confidence in an example's embedding position is a crucial indicator for identifying high-quality spectra. Combining the variations and density metrics resulted in a significant enhancement in all performance metrics emphasizing that the two metrics combined capture different aspects of uncertainty providing overall better performance in retrieving high-quality examples. The combination of all three metrics yielded the best performance across all metrics, with a ROC-AUC score of 0.94. This reveals that a holistic view, encompassing variations in embedding, data distribution around the embedding, and consistency between input and embedding densities, provides the most potent approach for identifying high-quality spectra.

Epistemic uncertainty
The standout performance of the consistency metric (ROC-AUC of 0.99) reinforces its innovative design. Assessing the alignment between the density of input data, and embeddings offers insight into the consistency of the model's behavior, providing a crucial aspect of uncertainty evaluation. Detailed results are given in Fig. 5B.

Remarks. To further validate our results, we perform a database search on both in-distribution and out-of-distribution datasets (Supplementary Fig. 8D) which showcases that the predictive power of the proposed metrics is a good estimate of the uncertainty associated with the data. Further, we selected three datasets where each dataset is a mixture of in-distribution and out-of-distribution data. Confusion matrices are calculated to determine what percentage of in-distribution predicted spectra match with the correct peptide. As shown in Fig. 5C, the predicted in-distribution data points perform vastly better than the predicted out-of-distribution data with an accuracy of up to 96.4%. Evaluation of uncertainty on in-distribution vs. out-of-distribution data on varying experimental parameters (species, fragmentation, instrument) showcases the robustness of the proposed metrics, and that predictive power is a good estimate of uncertainty.

Discussion

ProteoRift is a machine-learning method that can predict multiple peptide properties (length, missed cleavages, and modification status) directly from spectra. While useful for search-space reduction, exclusive reliance on mass-only filtering can limit peptide identification, particularly in complex proteomes. Incorporating additional spectral properties beyond mass has the potential to provide a better alternative while enhancing search sensitivity and scalability. We demonstrated that these properties when used for search-space reduction in an end-to-end database search pipeline, results in superior speeds (8x to 12x faster) and comparable results for both proteomics, and meta-proteomics experiments. Evaluations using ProteoRift (with SpeCollate as underlying engine) on a variety of datasets including immunopeptidomic, tryptic, and meta-proteomics exhibit comparable performance to mass-only tools such as SpeCollate, MSFragger and Crux. This highlights ProteoRift's potential as a viable alternative filtering strategy, suitable for integration within end-to-end machine learning pipelines. Further, our evaluation of robustness to substantial dataset divergence exhibits that ProteoRift can be used for a number of systems biology experiments, and the basic model presented in this paper is amenable to fine-tuning for specific datasets. In addition to the end-to-end pipeline, we also introduced certainty metrics which enable confidence estimation of the results especially for identification of novel and uncommon peptides – potentially reducing the skepticism of results obtained using black-box machine-learning models.

Our experiments demonstrate that the model works well for a variety of HCD mouse and human data sets. To optimize performance on out-of-distribution datasets, *transfer learning* might be required. For datasets that closely resemble the original training data (i.e. human, HCD, known PTMs), partial fine-tuning, where only the final layers are retrained, might be sufficient. For datasets that diverge significantly (e.g. different species, fragmentation methods etc.), fine-tuning of the entire model may be necessary. In both cases, leveraging the pre-trained model as a starting point helps preserve valuable learned representations while enabling the model to adapt to domain-specific features[66]. ProteoRift is intended to complement existing algorithmic tools and may result in a greater number of identified peptides. The uncertainty metrics convey the degree of confidence associated with the inferences, which can then guide the determination of whether fine-tuning is warranted.

Limitations of the current approach

ProteoRift is currently a prune-and-search method which can result in reduction in accuracy with successive pruning attempts. Accumulation of more filtering decreases the search-space and time for larger databases, and results in decreased false-positives. However, this also results in decrease in the number of identified peptides due to summation of the errors by three different heads. This model prediction error can be attributed to the imbalance in data (Supplementary Fig. 3) where the distribution of features is highly skewed, and some labels have very few training examples e.g. unmodified peptides are much greater in number than modified peptides. ProteoRift is a supervised learning technique and the amount and variability of the labelled data available is also one of the limitations of the method e.g. Most of the training has been accomplished on the HCD data which result in degradation of the prediction for CID/ETD data. This limitation also translates in the evaluation of the model that is currently limited to the precursor charge value of the spectra to a maximum of 3+. Although the accuracy of the peptide properties inferred are limited by variability of the available labelled data, ProteoRift is a generalizable model that can formulated as a continual learning framework[67] for further learning as data becomes available. While the model architecture is agnostic to fragmentation patterns, it is primarliy trained

and learned spectral representations on HCD datasets characterized by b- and y-type ions. For fragmentation methods that differ substantially from HCD, such as ETD-based approaches (which produce predominantly c/z ions), fine-tuning on appropriately annotated data is recommended. Similarly, ProteoRift has been trained and tested exclusively on tryptic datasets, and when additional datasets were included, only peptides consistent with tryptic digestion rules were considered. Therefore, the current implementation is best suited to tryptic digestion workflows, whereas extending it to non-tryptic proteases would require fine-tuning on enzyme-specific datasets. Moreover, ProteoRift is based on a one-to-one mapping between an MS/MS spectrum and a peptide, which aligns naturally with DDA-style acquisition. DIA data (where spectra result from multiplexed precursor isolation) violate this assumption and is outside the scope of this study, and cannot be applied directly to raw DIA MS/MS windows in its current form. While, in principle, DIA data could be incorporated following deconvolution or pseudo-spectrum generation to approximate DDA-like spectra, such extensions were not explored here and are part of our future work. Any further improvement in the results can be accomplished by using minor fine tuning of the model with relevant data which is made straightforward with our extensive documentation and open-source code. Another limitation of our approach lies in the potential for misclassification during spectrum identification. I.e. a spectrum may be incorrectly matched to a wrong length or missed cleavage simply because it was confined to the associated bin (for length or miscleavage). This then results in comparatively less number of PSM and peptides matches as compared to mass-only filtering strategies (Supplemental Fig. 10, 11, 12, 13, Supplemental Tables 3, and Mismatch Analysis Peptides from https://osf.io/t3dch). Furthermore, Investigating ProteoRift's performance with various backbone models, beyond SpeCollate, is an area for future study.

We believe that machine-learning tools are the enablers of new exciting science. However, the research community needs more synergistic efforts in integrating ML scoring functions with existing algorithmic or all-ML pipelines. Additionally, improved scoring techniques using machine-learning, for both de novo and database search, are steps in the right direction; other parts of the workflow are equally important for discovery of non-abundant peptides and proteins and is part of our future work. In this paper, we demonstrated that predictive filtering using ProteoRift when combined with SpeCollate results in a ML pipeline for peptide deduction that is comparable to Crux and MSFragger for a large number of data sets. Our development of uncertainty measures will enable scientific investigations and adaptation of ML tools in systems biology labs. Such ML computational infrastructure will provide alternate ways to compute and process mass spectrometry data, potentially reducing the *streetlight* effect, and resulting in discovery of uncommon peptides and proteins – related to human disease and health.

Methods

In this section, we will present the design and implementation of our deep attention-based multitask network (ProteoRift). To demonstrate that the predictive powers of our model are at least as useful as the mass-filtering, we devised an end-to-end peptide database search pipeline which results in deduced peptides and the associated confidence in those deductions. Four significant contributions of this work is the design and development of: (1) attention-based network to generate spectra and peptide embeddings, (2) multi-task network to predict peptide-length, missed cleavages, and modifications status directly from spectra, (3) novel metrics to quantify the uncertainty associated with spectra embeddings, and (4) predictive filtering based peptide database search that takes into account the confidence of spectra embeddings.

ProteoRift: a machine-learning model to predict peptide properties directly from the MS spectra

ProteoRift is designed as a multitask learning formulation that is used in conjunction with a pre-trained SpeCollate model. During ProteoRift training, the SpeCollate model was fine-tuned by unfreezing its final layers. This combined fine-tuning and training strategy ensures that the multitask formulations are correctly learned with pre-training of SpeCollate. ProteoRift utilizes SpeCollate's spectral features as input and embeddings for the subsequent database search.

Spectra and peptide embeddings

All raw MS data were converted to MGF format using the ProteoWizard msconvert tool[68]. Following conversion, the top 150–200 most intense peaks were selected for each spectrum. No further denoising, peak merging, or imputation of missing peaks was performed. All data were processed within a DDA framework, where each isolated precursor MS/MS spectrum corresponds to a single peptide identification.

Our first challenge is to translate spectra and peptides' high-dimensional sparse vectors into embeddings that can place semantically similar spectra and peptide inputs close together in the embedding space. We have designed and developed an embedding network shown in Fig. 6. The network uses two self-attention layers with eight head attentions to embed mass spectra with dropout 0.3. Using self-attention layers for embedding mass spectra is useful in capturing long-distance relations between peaks e.g., complimentary b and y peaks as well as short-distance relations such as a neutral loss or isotopic peak (Supplementary Fig. 2). For successfully capturing all the information in the spectrum, m/z, and intensity, sequences are first discretized i.e., m/z values are binned/rounded in 0.1 Da sized bins, and intensity values are discretized between 0 and 1000 before passing them through the embedding layers. Since attention layers process the entire sequence at the same time, the sequence information needs to be encoded separately. For this purpose, we use the sinusoidal harmonics[69]. The m/z and intensity embedding as well as the sequence encoding are of the same size 256 and are added together before being processed by the attention layer.

We did not perform a separate validation study to quantify the impact of alternative preprocessing strategies on prediction accuracy. Instead, we adopted a standard and widely used preprocessing protocol (as mentioned above) to ensure consistency across datasets. The goal of this study was to evaluate the effectiveness of property-

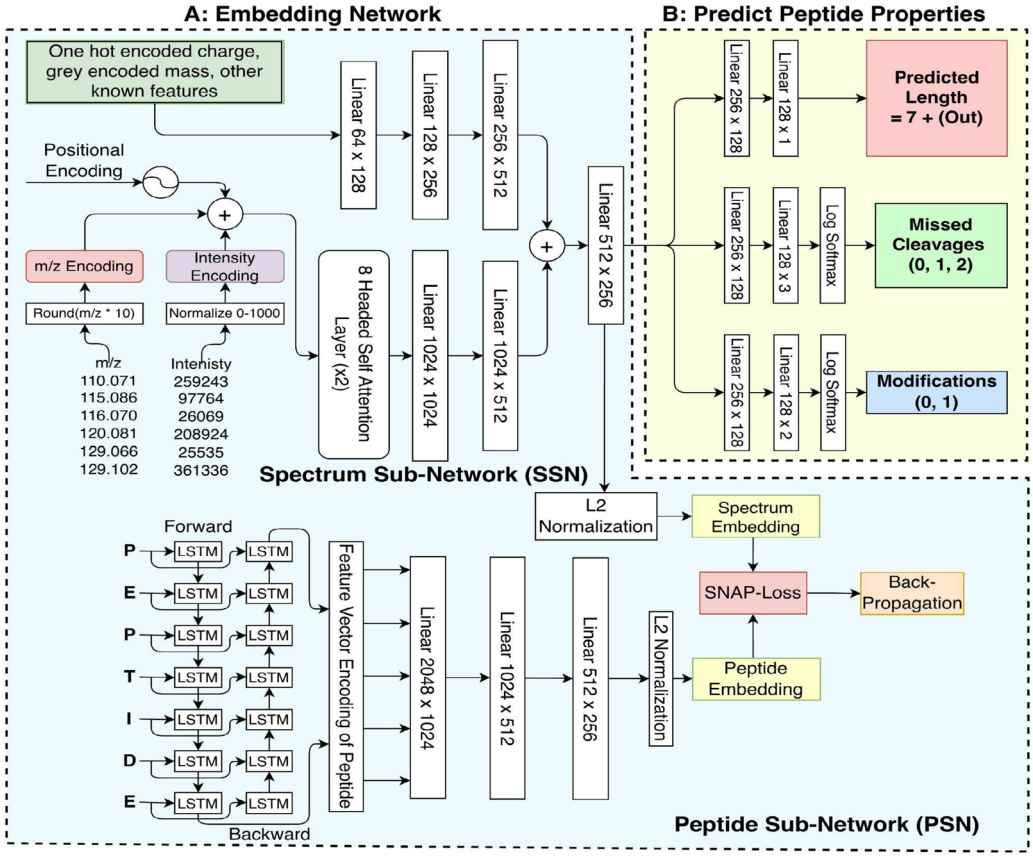


Fig. 6. ProteoRift Network architecture that is designed, developed and trained for spectra and peptide embeddings. (**A**)**:** Embedding network. in spectrum sub-network (SSN), spectra m/z and intensity values are treated as two separate sequences, and embedded using separate embedding layers. Sinusoidal positional encoding is also added to the embeddings to maintain sequence order information. An intermediate spectrum feature vector is added with charge and mass feature vectors and passed through a fully connected layer of 512×256. In Peptide Sub-Network (PSN), peptide embeddings are generated using two Bi-LSMT layers followed by three fully connected layers. (**B**)**:** Predict peptide properties. From the spectrum embeddings, three branches of fully connected layers predict the peptide length, missed cleavages, and modifications. All the output branches calculate loss separately using the LogSoftmax function and a weighted sum of three losses is calculated before calculating gradients. The predicted length outputs a value that corresponds to the length of the predicted peptide, Missed Cleavages block outputs 0, 1 or 2 for zero, one and two missed cleavages, and modification outputs either 0 for no modification and 1 for a positive modification status.

based filtering within a database search framework rather than to optimize MS preprocessing parameters. We have clarified these preprocessing steps in the revised manuscript.

This is followed by two fully connected layers of size 1024×1024 and 1024×512. The feature vector of mass and charge values is added to the intermediate output after the second fully connected layer. The combined feature vector is passed through another fully connected layer to generate spectrum embedding of length 256. Peptide embedding is generated using 2 Bi-Directional Long Short-Term Memory (Bi-LSTM) networks. Amino acid level embedding is generated using an embedding layer, and fed to the Bi-LSTM which generates output vectors of size 2048 per peptide. Note that only the final amino acid output is kept and fed to the following layers. Next, we use three fully connected layers of size 1024, 512, and 256. The final layer is preceded by L2 normalization to generate normalized vectors of length 256 for each peptide. For the section on the performance of additional datasets, we replaced the attention layers with linear feature processing layers. This modification was introduced to improve computational efficiency and accelerate for performing database search on large datasets.

Predicting peptide properties

From the spectrum embeddings, three branches of fully connected layers predict the peptide length, missed cleavages, and modifications as shown in Fig. 6B. For the length prediction, the two fully connected layers of 256×128 and 128×1 are used. The output layers are of length 24 because we predict lengths from 7 to 30 AAs. The missed cleavage branch contains two fully connected layers of size 256×128 and 128×3 as we only consider the max of two missed cleavages. Finally, the modification branch consists of two fully connected layers of size 256×128 and 128×2. For the current version of ProteoRift, we only predict whether the spectrum is modified

or not. In future versions, we will aim to predict the exact number of modifications as well as the specific known modifications that exist in the spectrum. A composite loss function was used. Mean squared error (MSE) was used for peptide length prediction, while cross-entropy loss was applied to missed cleavage and modification status. The respective loss weights were set to 3 (MSE), 1 (missed cleavage), and 1 (modification).

Experimental setup overview

Training and Validation Data. The model is trained using high-quality spectra with selected label peptides. Three datasets i.e., NIST, MassIVE, and DeepNovo datasets are used (Supplemental Data File 1). We further applied filters to only keep peptides with 2 + and 3 + charge, 7–30 peptide lengths, and precursor mass < 5000Da. The spectral libraries are preprocessed to extract the MS/MS spectra along with the target labels from the corresponding peptide i.e., peptide sequence length, missed cleavages, and modification status. Peptide sequence length is calculated as the number of amino acids in the peptide regardless of the modifications on any amino acid. Missed cleavages are calculated as the number of K and R amino acids within the sequence except if followed by P or when at the end of the peptide sequence. Modification status is calculated as true or false depending on whether there is a dynamic modification (one or more) on any number of amino acids. We obtained ~ 2.7 M spectra with known peptide labels out of which ~.37 M were modified, and the rest were unmodified. These modifications encompass common PTMs such as phosphorylation (on serine, threonine, and tyrosine residues), Iodoacetic acid derivatized residue, Carbamidomethylation, Acetylation, and Monohydroxylated residue. The dataset contained ~ 1.7 M spectra with 2 + precursor charge and ~ 0.5 M spectra with 3 + precursor charge. In terms of missed cleavages, the training dataset contains ~ 2M, ~ 0.7 M, and ~ 50k spectra with zero, one, and two missed cleavages respectively. ProteoRift is trained with the train/test/val split at the peptide-level (since we had the labels) and there were no overlapping peptides between the train/testing data sets. The data is randomly split into training (70%), validation (20%) and testing (10%) sets. We tune hyper parameters on the validation set and report test set results for the best run (tuned parameters) achieved on the validation results. The precursor charge value of the spectra was restricted to a maximum of 3+, and this limitation is expected due to the limited variability in the labelled training data. Distribution of peptide length missed cleavages, and modifications of training data used in this study (Supplementary Fig. 3 A, B and C). For our training data more than 50% of the data is for peptide length between 7 and 13. And more than 90% of our data is for peptide length less than 21. Therefore, only 10% of the data covers lengths from > 21 and is similar to other data distribution reported in the literature[70,71]. Similarly, the training samples for 0 missed cleavages, and unmodified peptides are much more in number as compared to 1, or 2 missed cleavages or modified peptides. To ensure that our models work for all lengths, we assign class weights inversely proportional to the class counts. The same oversampling strategy is used for missed cleavages, and modified peptides respective to class counts and associated class weights (Supplementary Fig. 3D, and E).

Runtime environment and experimental parameters

The experiments are performed using the PyTorch framework 1.13.0 and Python 3.10.8. For fast training, we utilized NVIDIA V100s (32 GB SMX2) GPUs with 32GB of memory provided by Expanse SDSC supercomputer installed in nodes with up to 4 GPU nodes, 40 CPU cores, and 384 GBs of memory (10 CPU cores and 96 GBs per GPU node). In addition, NVIDIA A6000 GPUs with 48GBs of memory were utilized as part of our in-house cluster (dragon).

Unless otherwise stated the following tools, and parameters were used for experimentation. Crux (version 4.1-e4b2f4d-2021-10-13), Proteowizard (version 3.0.21287), Comet (version 2019.01 rev. 5) were used. FDR was calculated by using Crux inbuilt Percolator (version 3.05.nightly-137-e806a0c, Build Date Oct 14 2021). Machine learning parameters include embedding_dim = 1024, encoder_layers = 2, num heads = 8, and dropout rate of 0.3. Other parameters include spec_batch_size = 16,384, pep_batch_size = 16,384, search_spec_batch_size = 256, precursor_tolerance = 7 (Precursor tolerance used for database search), maximum charge = 5 and 5 top scoring PSMS are kept for evaluation. More detailed information can be found in Supplemental Data File 2.

Network training and evaluation

The training process begins with a forward pass of the batch containing encoded spectra. Three sets of labels are used i.e., peptide length, missed cleavages, and modifications status. The network generates three separate outputs for each feature respectively which are treated as multi-class classification. The cross-entropy loss function is applied to each feature separately to calculate three loss values. Before the backpropagation is performed, a weighted sum of the three losses is calculated where the weights are determined through a randomized search. Adam optimizer is used for gradient updates with a learning rate of 0.0001 and weight decay of 0.00005. A dropout of 0.3 is used after the attention, and the fully connected layers during the training step. The dropout layers are disabled during the evaluation step. The network is trained for 500 epochs and achieves 92% peptide length precision, 97% missed cleavage precision, and 97% modification precision as shown in Supplementary Fig. 4. Minor increase in the testing precision for length and missed cleavage is due to artificial class weights in the training set to ensure effective learning. However, testing data is real-world data that may not encounter less-probable peptides (e.g. with 2 or more missed cleavages) frequently resulting in increased testing precision performance.

Implementing database search with predictive filters

To implement the database search with the new filters we first pre-assign the database peptides and the experimental spectra into one of the 144 (7–30 peptide length × 0, 1, or 2 missed cleavages × 2 modifications statuses) classes based on their features. For the database peptides, the features are known and hence each peptide gets classified into a non-overlapping class. Spectra are classified based on their predicted features using

the ProteoRift model. A spectrum from a given class only gets searched against peptides in the same class. Since each class is non-overlapping, 144 parallel searches can be performed simultaneously. The general filtered database search flow is shown in Fig. 1 step 2.

In the first step, spectra and peptides are indexed into their respective classes based on their properties. Each spectrum and peptide get assigned to a unique class. Spectra and peptides within each class are batched. As shown in Fig. 1 step 2, spectra in each class e.g., 7-1-0 (Peptide length: 7, 1 missed cleavage, unmodified) are sorted with respect to the precursor mass, and split into batches (*SB*) of size 1024. Similarly, peptides are sorted with respect to their precursor mass, and candidate peptide lists are generated for each SB. To generate a peptide list for a given SB, peptides in the mass range [*MinSB - PreTol*: *MaxSB + PreTol*] are obtained. Here, *MinSB* and *MaxSB* are the minimum and maximum precursor masses in a given SB respectively. *PreTol* is the user-specified precursor tolerance configuration. Next, each peptide list is split into batches (*PB*) of 1024 peptides. The size can be adjusted according to the available memory. For each *PB* in a peptide list, a distance matrix against the *SB* is calculated. A mask is calculated for each spectrum where peptides outside the mass filter range are ignored. Next, the values in the distance matrix are sorted for each spectrum in increasing order of L2 distance, and k smallest distance peptides are kept for each spectrum as these are the highest-scoring peptides. The inverse of the L2 distance is reported as the match score. The process is repeated for each SB and results are written to a percolator[62] input (pin) file. A similar process is repeated for the decoy database for FDR analysis. Percolator/PeptideProphet is then used to estimate false-discovery rate, and a subset of PSMs and peptides estimated to have FDR < 1% is reported.

Uncertainty analysis

The utility of database-search ML models might be subject to various degrees of vagueness influenced by multifaceted factors[46–48]. This imprecision encapsulates two key dimensions: *aleatoric* and *epistemic*[72,73]. Aleatoric imprecision originates from the inherent noise and stochasticity in the data, whereas epistemic imprecision mirrors the model's limitations, signifying the unlearned or unknown aspects within the model's purview. Such uncertainty estimation enables the reliability of the model's predictions. For example, elevated uncertainty in peptide or spectrum embeddings may signal model ambiguity in their representations, potentially indicating a misaligned peptide-spectrum match. Further, such uncertainty quantification provides insights into embedding areas that can be improved either by further training or acquiring more quality or diverse data. For example, higher epistemic uncertainty can indicate regions in the input space where the model is *under-confident* due to a lack of sufficient data. This can point out opportunities for further data collection or model improvement. Similarly, higher aleatoric uncertainty can point out the need for higher-quality data or preprocessing for noise removal. Most importantly, uncertainty estimation allows for risk-aware decision-making. In high-stakes applications such as clinical proteomics — where incorrect peptide identification could lead to misleading conclusions — being aware of the uncertainty associated with each prediction can help avoid potentially costly or harmful decisions.

Our objective is to design uncertainty metrics that can be used to assess the confidence in the peptide deduction i.e., how confident the scientist should be in the inference especially when the identified peptide is novel. We estimate the aleatoric and epistemic uncertainty of SpeCollate's embeddings using our proposed metrics, and analyze how they can inform us about the model's performance and its output confidence. To this end, we developed three different metrics for estimating the uncertainty of the embeddings: (1) We assess the certainty of embedding location by introducing controlled augmentations to the input spectra and then measuring the variation in the output embeddings. (2) The density of the training data around each embedding is measured using a von Mises-Fisher (vMF) Mixture Model to determine how much data the model has seen around a given data point. (3) We evaluate whether the density is consistent between the input spectra and their embeddings, indicating whether the model can maintain data structure and relationships.

Aleatoric uncertainty

To determine how well our metrics estimate aleatoric uncertainty for a given embedding, we use a mass spectrometry dataset that is similar to SpeCollate's training data i.e., ProteomeTools HCD data. We want to see how well these metrics can predict the downstream performance of the in-distribution embeddings e.g., in our case, how well the mass spectra embeddings can be identified in a peptide database search. Ideally, higher confidence along with a higher ranking of the spectrum would indicate the model is certain about its decision. To quantify this, we rank peptide embeddings from the human peptide database against each spectrum. A true label is assigned to a spectrum if the correct peptide is ranked the highest. The label definition is given below:

$$Label = \{1, Correct peptide is ranked highest, 0, Otherwise.$$

We train a gradient-boosting decision tree (GBDT) model with per-sample feature variations (variations for short), density, and density consistency as input and the above-described labels as output. We perform hyperparameter tuning for the max depth, learning rate, and number of estimators. The optimal values selected are shown in supplementary Table 1.

Epistemic uncertainty

To estimate epistemic uncertainty, we aim to classify in-distribution and out-of-distribution data points using our uncertainty metrics. For in-distribution and out-of-distribution data, we use ProteomeTools HCD and CID datasets. As SpeCollate is solely trained on HCD data, CID data is an obvious choice for out-of-distribution examples to validate our uncertainty metrics. We assign the following labels to our curated dataset with nearly 1 million examples for each class:

$$Label = \{1, Id - distributiondatapoint, 0, Out - of - distributiondatapoint.$$

We train a GBDT model on this data with a 70-20-10 split of train validation and test subsets. The optimal parameters selected after hyperparameter tuning are given in supplementary Table 2.

Below we will discuss the proposed uncertainty measures and present their mathematical formulation.

Per-sample feature variation

The incorporation of augmentation techniques to estimate the sample variance is a method employed to quantify the level of certainty a model has regarding the position of its predicted embedding. The principle lies in the deliberate modification of each data point in multiple ways. The following process is followed: each spectra is augmented with Insilco by randomly removing peaks at 0%, 5%, 10%, and 15% for charge states of + 2, +3 and + 4. This results in a total number of augmentations per spectra equal to 12. The sample used in these experiments then consists of original data point plus augmentations listed. All augmentations for the uncertainty estimation are completed using the listed procedure. Variance is calculated as the trace of the sample covariance matrix. The principle lies in the deliberate modification of each data point in multiple ways, e.g. random omission of a certain percentage of peaks or adjusting the charges associated with spectra. The distribution of per-sample feature variance with and without augmentations is illustrated in Supplementary Fig. 8A.

Given an original spectrum *s*, we induce *n* augmentations thereby creating a set of altered spectra denoted by $\{s_1, s_2, \ldots, s_n\}$. Each derived spectrum embedding is a 256-dimensional vector. The variance, a statistical measure of the extent to which the data points deviate from their expected values is computed for each of these dimensions.

Mathematically, sample variance σ can be defined as follows:

$$\sigma = \frac{1}{n-1}\sum_{i=1}^{256}\sum_{j=1}^{n}\left(s_j\left[i\right] - \mu_i\right)^2$$

where $s_j\left[i\right]$ denotes the i-th element in the j-th augmentation,μ_i is the mean of the elements for the i-th dimension, and $n$is the number of augmentations.

To compile these individual variances into a single value, a cumulative sum is calculated across all dimensions. This total sum, or sample variance, is indicative of the model's confidence in the location of the predicted embedding; a smaller variance suggests a higher degree of certainty.

In essence, this technique provides valuable insights into the model's confidence in its predictions, which can be crucial in the interpretation of the downstream performance. Only a trained model is required to estimate the variance. Sample feature variation is shown in Fig. 1 (step 3, part G).

Density

The density of the embedding space around a spectrum embedding is the measure of how many training examples the model has seen close for a given example. We use the *von Mises-Fisher* distribution to estimate the density of the embedding space around a given data point. As the spectra and peptide embeddings generated by *SpeCollate* are L2-normalized, the *von Mises-Fisher* distribution, when combined into a mixture model (vMFMM), is an ideal choice for modeling data due to its directional properties.

Assuming we have a set of N training examples $\{s_1, s_2, \ldots, s_n\}$ and these are normalized to unit length; let's say we fit a vMFMM to this data with k components. Each component k has a center μ_k and a concentration parameter κ_k and a weight w_k. The density p(s) for any point s in this mixture model can be computed as:

$$p\left(s\right) = \sum_{k=1}^{K} w_k \cdot p_k\left(s\right)$$

Where $p_k\left(s\right)$ is the probability density of spectrum embedding s in the k-th von Mises-Fisher component, defined as:

$$p_k\left(s\right) = C_m\left(\kappa_k\right) \cdot e^{\kappa_k \cdot \mu_k^T x}$$

where $C_m\left(\kappa\right)$ is the normalization constant in *m* dimensions, given by:

$$C_m\left(\kappa\right) = \frac{\kappa^{\frac{m}{2}-1}}{\left(2\pi\right)^{\frac{m}{2}} I_{\frac{m}{2}-1}\left(\kappa\right)}$$

and $I_{\frac{m}{2}-1}\left(\kappa\right)$ is the modified Bessel function of the first kind.

Therefore, to measure the density of training examples around a given point, we would compute $p\left(s\right)$ using the above equations. Supplementary Fig. 8B shows the density distribution of the input spectra and its embeddings without any augmentations which show similar distributions. This gives an idea of how familiar the model is with the region around a given embedding. Density estimation is visualized in Fig. 1 (Step 3, part H).

Density consistency

We propose a novel metric, herein referred to as the *Density Consistency (DC)*, which is designed to assess the preservation of density consistency between the input data and the model's output embeddings. The DC

is a measure of a model's ability to maintain the structural and relational characteristics of the data in its transformation to the embedding space.

Given a data point x and its embedding s the DC is mathematically formulated as follows:

$$DC = |D_i(x) - D_e(s)|$$

In this formula, $D_i(x)$ is the density of the input data around a given point x, which is estimated using domain specific knowledge, and $D_e(s)$ is the density of the embeddings around a given point s, which is estimated using the von Mises-Fisher Mixture Model (vMFMM) as outlined in the previous section. The absolute value of the difference between these two densities yields the DC. In the case of our specific application, MS/MS spectra, the density of the input data $D_i(x)$ is calculated using a cosine similarity metric. This serves to gauge the similarity between pairs of input spectra. This metric, therefore, allows us to quantitatively evaluate the model's performance in terms of its ability to maintain the integrity of data relationships and structure in the transformation from the input space to the embedding space. Supplementary Fig. 8 C shows scatter plot between the spectral density and the embedding density of our data and illustrates a similar and highly correlated distribution. Therefore, we argue that the lower the value of DC, the more effectively the model preserves the original data structure in its output embeddings. Density consistency is shown in Fig. 1 (Step 3 part I).

Data availability

No new data was collected for this study. We collected Pride Archive (PXD) datasets, pre-processed and wrangled them into a suitable format for machine-learning workflows for our experimentation and evaluation of proteomics search from ProteomeXchange (PXD000612, PXD001468, PXD009861, PXD019774, PXD026295, PXD010595), and MassIVE (msv000082031) dataset. Proteome database files were downloaded from RefUP++ database (https://github.com/miinslin/ProteoStorm), and Uniport proteome ID UP000005640 (H. sapiens). The parameter, log, and result files in this study are available in https://osf.io/puefz/. Peptides with 1% FDR: https://osf.io/2sndq, Peptides with 5% FDR: https://osf.io/vq4ny, Unique Proteorift Peptides: https://osf.io/ke9dy, Mismatch Analysis: https://osf.io/t3dch, PXD019774_MSFragger_Crux_Analysis(https://osf.io/n527v), Entrapment analysis: https://osf.io/xhztm, https://osf.io/6s9kc.All code, associated models and weights are made open-source at: [https://github.com/pcdslab/ProteoRift](https:/github.com/pcdslab/ProteoRift).

Received: 19 September 2025; Accepted: 2 March 2026
Published online: 13 March 2026

References

1. Haseeb, M. & Saeed, F. High performance computing framework for tera-scale database search of mass spectrometry data. *Nat. Comput. Sci.* **2021 1**, 550–561 (2021).
2. Haseeb, M. & Saeed, F. GPU-acceleration of the distributed-memory database peptide search of mass spectrometry data. *Sci. Rep.* **13**, 18713 (2023).
3. Tariq, M. U. & Saeed, F. SpeCollate: Deep cross-modal similarity network for mass spectrometry data based peptide deductions. *PLoS One* **16**, e0259349 (2021).
4. Tariq, M. U., Ebert, S. & Saeed, F. Making MS Omics Data ML-Ready: SpeCollate Protocols In (ed. Lisacek, F.) (2024).
5. McIlwain, S. et al. Crux: Rapid open source protein tandem mass spectrometry analysis. *J. Proteome Res.* **13**, 4488–4491 (2014).
6. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods* **14**, 513–520 (2017).
7. Sun, J. et al. Open-pFind enhances the identification of missing proteins from human testis tissue. *J. Proteome Res.* **18**, 4189–4196 (2019).
8. Polasky, D. A. et al. MSFragger-Labile: A flexible method to improve labile PTM analysis in proteomics. *Mol. Cell. Proteomics* **22**, 100538 (2023).
9. Kustatscher, G. et al. Understudied proteins: Opportunities and challenges for functional proteomics. *Nat. Methods* **19**, 774–779 (2022).
10. Kustatscher, G. et al. An open invitation to the Understudied Proteins Initiative. *Nat. Biotechnol.* **40**, 815–817 (2022).
11. Dunham, I. Human genes: Time to follow the roads less traveled?. *PLoS Biol.* **16**, e3000034 (2018).
12. Haynes, W. A., Tomczak, A. & Khatri, P. Gene annotation bias impedes biomedical research. *Sci. Rep.* **8**, 1362 (2018).
13. Nguengang Wakap, S. et al. Estimating cumulative point prevalence of rare diseases: Analysis of the Orphanet database. *Eur. J. Hum. Genet.* **28**, 165–173 (2020).
14. Bakos, J., Zatkova, M., Bacova, Z. & Ostatnikova, D. The role of hypothalamic neuropeptides in neurogenesis and neuritogenesis. *Neural Plast.* https://doi.org/10.1155/2016/3276383 (2016).
15. Huang, Z. et al. Brain proteomic analysis implicates actin filament processes and injury response in resilience to Alzheimer's disease. *Nat. Commun.* **14**, 2747 (2023).
16. Johnson, E. C. B. et al. Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. *Nat. Med.* **26**, 769–780 (2020).
17. Chen, F., Chandrashekar, D. S., Varambally, S. & Creighton, C. J. Pan-cancer molecular subtypes revealed by mass-spectrometry-based proteomic characterization of more than 500 human cancers. *Nat. Commun.* **10**, 1–15 (2019).
18. Leiserson, M. D. et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015).
19. Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 8247–8252 (2017).
20. Diament, B. J. & Noble, W. S. Faster SEQUEST searching for peptide identification from tandem mass spectra. *J. Proteome Res.* **10**, 3871–3879 (2011).
21. Craig, R. & Beavis, R. C. TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466–1467 (2004).
22. Seydel, C. Diving deeper into the proteome. *Nat. Methods.* **19**, 1036–1040 (2022).
23. Skinner, O. S. & Kelleher, N. L. Illuminating the dark matter of shotgun proteomics. *Nat. Biotechnol.* **33**, 717–718 (2015).
24. Lazear, M. R. Sage: An open-source tool for fast proteomics searching and quantification at scale. *J. Proteome Res.* **22**, 3652–3659 (2023).

25. Chick, J. M. et al. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat. Biotechnol.* **33**, 743–749 (2015).
26. Nesvizhskii, A. I. Proteogenomics: Concepts, applications and computational strategies. *Nat. Methods* **11**, 1114–1125 (2014).
27. Burke, M. C. et al. The hybrid search: a mass spectral library search method for discovery of modifications in proteomics. *J. Proteome Res.* **16**, 1924–1935 (2017).
28. Solntsev, S. K., Shortreed, M. R., Frey, B. L. & Smith, L. M. Enhanced global post-translational modification discovery with MetaMorpheus. *J. Proteome Res.* **17**, 1844–1851 (2018).
29. Muth, T., Renard, B. Y. & Martens, L. Metaproteomic data analysis at a glance: Advances in computational microbial community proteomics. *Expert Rev. Proteomics* **13**, 757–769 (2016).
30. Muth, T. et al. The MetaProteomeAnalyzer: A powerful open-source software suite for metaproteomics data analysis and interpretation. *J. Proteome Res.* **14**, 1557–1565 (2015).
31. Schiebenhoefer, H. et al. A complete and flexible workflow for metaproteomics data analysis based on MetaProteomeAnalyzer and Prophane. *Nat. Protoc.* **15**, 3212–3239 (2020).
32. Nesvizhskii, A. I. et al. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol. Cell. Proteom.* **5**, 652–670 (2006).
33. Griss, J. et al. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat. Methods* **13**, 651–656 (2016).
34. Ning, K., Fermin, D. & Nesvizhskii, A. I. Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets. *Proteomics* **10**, 2712–2718 (2010).
35. Nielsen, M. L., Savitski, M. M. & Zubarev, R. A. Extent of modifications in Human proteome samples and their effect on dynamic range of analysis in shotgun proteomics* S.. *Mol. Cell. Proteomics* **5**, 2384–2391 (2006).
36. Frank, A., Tanner, S., Bafna, V. & Pevzner, P. Peptide sequence tags for fast database search in mass-spectrometry. *J. Proteome Res.* **4**, 1287–1295 (2005).
37. Zhang, S. et al. High-resolution quadrupole improves spectral purity and reduces interference from non-target ions in isobaric multiplexed quantitative proteomics.. *Anal. Chim. Acta* **1325**, 343135 (2024).
38. Haseeb, M. & Saeed, F. Efficient shared peak counting in database peptide search using compact data structure for fragment-ion index. in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 275–278 (IEEE, 2019). 275–278 (IEEE, 2019). (2019).
39. Wei, M.-M. et al. Integrating Transformer and Graph Attention Network for circRNA-miRNA interaction prediction.. *IEEE J. Biomed. Health Inform.* **29**, 6105–6113 (2025).
40. Huang, Y.-A. et al. Consensus representation of multiple cell–cell graphs from gene signaling pathways for cell type annotation.. *BMC Biol.* **23**, 23 (2025).
41. Wei, M., Wang, L., Su, X., Zhao, B. & You, Z. Multi-hop graph structural modeling for cancer-related circRNA-miRNA interaction prediction.. *Pattern Recogn.* **170**, 112078 (2026).
42. Altenburg, T., Muth, T. & Renard, B. Y. yHydra: Deep learning enables an ultra fast open search by jointly embedding MS/MS spectra and peptides of mass spectrometry-based proteomics. http://biorxiv.org/lookup/doi/ (2021). https://doi.org/10.1101/2021.12.01.470818 doi:10.1101/2021.12.01.470818.
43. Gessulat, S. et al. Prosit: Proteome-wide prediction of peptide tandem mass spectra by deep learning.. *Nat. Methods* **16**, 509–518 (2019).
44. Meyer, J. G. Deep learning neural network tools for proteomics. *Cell Rep. Methods* https://doi.org/10.1016/j.crmeth.2021.100003 (2021).
45. Cox, J. Prediction of peptide mass spectral libraries with machine learning. *Nat. Biotechnol.* **41**, 33–43 (2023).
46. Qiao, R. et al. Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. *Nat. Mach. Intell.* **3**, 420–425 (2021).
47. Yilmaz, M., Fondrie, W., Bittremieux, W., Oh, S. & Noble, W. S. De novo mass spectrometry peptide sequencing with a transformer model. in *International Conference on Machine Learning* 25514–25522PMLR, (2022).
48. Karunratanakul, K., Tang, H. Y., Speicher, D. W., Chuangsuwanich, E. & Sriswasdi, S. Uncovering thousands of new peptides with sequence-mask-search hybrid de novo peptide sequencing framework. *Mol. Cell. Proteomics.* **18**, 2478–2491 (2019).
49. Blundell, C., Cornebise, J., Kavukcuoglu, K. & Wierstra, D. Weight uncertainty in neural network. in *International conference on machine learning* 1613–1622PMLR, (2015).
50. Gal, Y. & Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. in *international conference on machine learning* 1050–1059PMLR, (2016).
51. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances neural Inform. Process. systems* **30**, (2017).
52. Liu, K., Ye, Y., Tang, H. & PepNet A fully convolutional neural network for De novo Peptide Sequencing. (2022).
53. Altenburg, T., Giese, S. H., Wang, S., Muth, T. & Renard, B. Y. Ad hoc learning of peptide fragmentation from mass spectra enables an interpretable detection of phosphorylated and cross-linked peptides. *Nat. Mach. Intell.* **4**, 378–388 (2022).
54. Zeng, W.-F. et al. AlphaPeptDeep: A modular deep learning framework to predict peptide properties for proteomics. *Nat. Commun.* **13**, 7238 (2022).
55. Kalhor, M., Lapin, J., Picciani, M. & Wilhelm, M. Rescoring peptide spectrum matches: Boosting proteomics performance by integrating peptide property predictors into peptide identification. *Mol. Cell. Proteomics.* **23**, 100798 (2024).
56. Dorfer, V., Maltsev, S., Winkler, S. & Mechtler, K. CharmeRT: Boosting peptide identifications by chimeric spectra identification and retention time prediction. *J. Proteome Res.* **17**, 2581–2589 (2018).
57. Deng, J., Julian, M. H. & Lazar, I. M. Partial enzymatic reactions: A missed opportunity in proteomics research. *Rapid Commun. Mass Spectrom.* **32**, 2065–2073 (2018).
58. Sharma, K. et al. Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep.* **8**, 1583–1594 (2014).
59. Bittremieux, W., Meysman, P., Noble, W. S. & Laukens, K. Fast open modification spectral library searching through approximate nearest neighbor indexing. *J. Proteome Res.* **17**, 3463–3474 (2018).
60. Qi, Y. A. et al. Proteogenomic analysis unveils the HLA Class I-presented immunopeptidome in melanoma and EGFR-mutant lung adenocarcinoma. *Mol. Cell. Proteomics* **20**, 100136 (2021).
61. Ino, Y. et al. Evaluation of four phosphopeptide enrichment strategies for mass spectrometry-based proteomic analysis. *Proteomics* **22**, 2100216 (2022).
62. Beyter, D., Lin, M. S., Yu, Y., Pieper, R. & Bafna, V. Proteostorm: An ultrafast metaproteomics database search framework. *Cell Syst.* **7**, 463–467 (2018).
63. Cogne, Y. et al. Comparative proteomics in the wild: Accounting for intrapopulation variability improves describing proteome response in a *Gammarus pulex* field population exposed to cadmium. *Aquat. Toxicol.* **214**, 105244 (2019).
64. Cooper, J. C. et al. Altered localization of hybrid incompatibility proteins in *Drosophila*. *Mol. Biol. Evol.* **36**, 1783–1792 (2019).
65. Wen, B. et al. Assessment of false discovery rate control in tandem mass spectrometry analysis using entrapment. *Nat. Methods* (7), https://doi.org/10.1038/s41592-025-02719-x (2025).

66. Bushuiev, R. et al. Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-025-02663-3 (2025).
67. Wang, L., Zhang, X., Su, H. & Zhu, J. A comprehensive survey of continual learning: Theory, method and application. *IEEE Trans. Pattern Anal. Mach. Intell.* https://doi.org/10.1109/TPAMI.2024.3367329 (2024).
68. Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
69. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods*. **4**, 923–925 (2007).
70. Da Silva Antunes, R. et al. Urinary peptides as a novel source of T Cell allergen epitopes. *Front. Immunol.* **9**, 886 (2018).
71. Huang, C. et al. Combined transcriptomics and proteomics forecast analysis for potential biomarker in the acute phase of temporal lobe epilepsy. *Front. Neurosci.* **17**, 1145805 (2023).
72. Hüllermeier, E. & Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Mach. Learn.* **110**, 457–506 (2021).
73. Der Kiureghian, A. (ed, O.) Aleatory or epistemic? Does it matter? *Struct. Saf.* **31** 105–112 (2009).

Author contributions
FS and UT conceived and designed the project. FS, BS and UT acquired the MS data, wrangled and processed the data to make it suitable for ML models. FS, UT and BS analyzed the data, designed the experiments, conducted the experiments, and presented the results in a comprehensive manner. UT wrote the ML model code, scripts and pipelines which were modified by BS for further investigation and results. FS, UT, and BS wrote the paper, and all authors contributed to the revisions and approved the final version. FS contributed to all aspects of the project.

Funding
This research was supported by the NIGMS of the National Institutes of Health (NIH) under award number: R35GM153434. The authors were further supported by the National Science Foundations (NSF) under the award number: NSF OAC-2312599. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health and/or National Science Foundation. This work used the NSF Extreme Science and Engineering Discovery Environment (XSEDE) Supercomputers through allocations: TG-CCR150017 and TG-ASC200004.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-43215-2.

Correspondence and requests for materials should be addressed to F.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.