Carafe enables high quality *in silico* spectral library generation for data-independent acquisition proteomics

Bo Wen[1], Chris Hsu[1], Wen-Feng Zeng[2], Michael Riffle[1], Alexis Chang[1], Miranda Mudge[1], Brook Nunn[1], Matthew D. Berg[1], Judit Villén[1], Michael J. MacCoss[*1], and William S. Noble[*1,3]

[1]Department of Genome Sciences, University of Washington
[2]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Germany
[3]Paul G. Allen School of Computer Science and Engineering, University of Washington

Abstract

Data-independent acquisition (DIA)-based mass spectrometry is becoming an increasingly popular mass spectrometry acquisition strategy for carrying out quantitative proteomics experiments. Most of the popular DIA search engines make use of *in silico* generated spectral libraries. However, the generation of high-quality spectral libraries for DIA data analysis remains a challenge, particularly because most such libraries are generated directly from data-dependent acquisition (DDA) data or are from *in silico* prediction using models trained on DDA data. In this study, we developed Carafe, a tool that generates high-quality experiment-specific *in silico* spectral libraries by training deep learning models directly on DIA data. We demonstrate the performance of Carafe on a wide range of DIA datasets, where we observe improved fragment ion intensity prediction and peptide detection relative to existing pretrained DDA models.

1 Introduction

Data-independent acquisition (DIA) has emerged as a powerful strategy for collecting liquid chromatography coupled tandem mass spectrometry (LC-MS/MS) data in bottom-up quantitative proteomics [1, 2]. The growth of this acquisition strategy has been enabled by significant technical advances both in the precursor isolation specificity and sensitivity [3, 4]. These advances have made DIA not only practical but preferable for the systematic and reproducible sampling of peptides for quantitative analysis. DIA produces product ion data independent of whether a precursor signal is detectable and is analogous to targeted parallel reaction monitoring (PRM). However, in contrast to PRM, DIA data is collected across a mass-to-charge (m/z) range and not on predefined target peptides. Furthermore, DIA has been enabled by a shift in computational strategy from searching MS/MS or MS2 spectra against peptide sequences to searching peptides against the LC-MS/MS data [5]. This strategy, referred to as peptide-centric searching, eliminates the assumption that each MS2 spectrum is produced by a single peptide, and most tools that implement the peptide-centric method assume that each peptide's fragment information is present in multiple continuous MS2 spectra at a retention time defined by the hydrophobicity of the sequence. Each peptide is scored against the data, often as extracted ion chromatograms, and not against specific spectra. This effectively multiplexes the MS2 spectra acquisition enabling the number of detectable peptides to exceed the number of spectra. Overall, the maturity of the hardware, combined with advances in computational methods, have facilitated systematic sampling of analytes and improved the quantitative analysis of complex peptide mixtures from protein digests.

*Correspondence: maccoss@uw.edu, wnoble@uw.edu

To analyze these data, software tools benefit from the use of existing metadata in the form of "spectral libraries" for analyzing DIA data. In this setting, a "spectral library" refers to a list of peptides, derived from the proteome of interest, each with estimates of the peptide's retention time (RT), fragment ion intensities, and any other information like ion mobility collisional cross-section that can be used to improve the assignment of a peptide sequence to a signal in the data. In practice, accurate estimates of these quantities are critical to achieving accurate peptide detection from DIA data [6, 7]. Accordingly, significant efforts have been made to generate high quality spectral libraries, typically by making use of data generated via data-dependent acquisition (DDA) [8, 9]. Best results are achieved when the spectral library is generated with the same instrument settings as those used for DIA data generation [10]. In practice, generating a comprehensive spectral library by DDA requires multiple biochemical fractions and, thus, is time-consuming and isn't particularly amendable to reuse.

One potential solution to this challenge comes from machine learning. In recent years, deep learning models have been developed that predict both peptide RT and fragment ion intensity highly accurately [11–16]. One of the major applications of these deep learning predictions is to generate *in silico* spectral libraries for DIA analysis [7, 12, 17, 18]. Indeed, it has been shown that these predicted spectral libraries yield comparable or superior performance compared to empirical spectral libraries generated from DDA experiments [7, 12, 19], potentially obviating the need to generate a new empirical spectral library for each project.

Unfortunately, both types of spectral libraries—empirical and predicted—suffer from a mismatch between the DDA data used to generate the library and the DIA data analyze with that library. In particular, fragment ion intensities differ systematically between these two acquisition strategies. This difference arises largely due to differences in collision energy optimization between DDA and DIA [20–22]. Specifically, in DDA, collision energy is typically optimized based on both precursor charge state and mass-to-charge ratio (m/z) for each peptide, whereas in DIA, collision energy is typically optimized based on a fixed charge state and the center m/z of an isolation window during peptide fragmentation. Consequently, the same peptide will yield different fragment ion intensities in DDA and DIA settings. In addition, the settings of liquid chromatography (LC) coupled with mass spectrometry used in DDA and DIA experiments are often different when both types of data are generated on different instruments or different laboratories, leading to differences in RTs between the two types of data that are difficult to calibrate effectively.

To address this mismatch, several methods have been developed to use DIA data in the construction of spectral libraries. For example, Searle *et al.* demonstrated that incorporating gas phase fractionated (GPF) DIA data into an empirical spectral library information leads to a significant boost in the number of detected peptides relative to a DDA-based library [20]. Similarly, the MSLibrarian method improves the peptide detection power of a predicted spectral library by leveraging information from DIA data [22]. In brief, MSLibrarian first applies DIA-Umpire [23] on a given DIA dataset to generate pseudo-DDA spectra and then performs a traditional database search on the pseudo-DDA spectra. Based on the identified peptides and their matched pseudo-DDA spectra, MSLibrarian then optimizes peptide fragment ion intensity prediction parameters using the Prosit machine learning model [12] and calibrates retention time predictions. MSLibrarian also reduces the size of the *in silico* generated spectral libraries by only including proteins identified by a database search using a relaxed false discovery rate (FDR) threshold and limiting the number of fragment ions.

In the context of predicted spectral libraries, the three-stage approach adopted by MSLibrarian can be improved. First, the method used to optimize prediction parameters does not account for different peptide fragmentation strategies. Second, rather than first training a deep learning model using DDA data and then optimizing that model's prediction parameters to account for the DIA settings, it is preferable to train a fragment ion intensity prediction model using DIA data directly. However, this direct approach is complicated because DIA spectra are highly chimeric. Thus, obtaining DIA training data with accurate fragment ion intensities and without interference is difficult.

To address these challenges, we developed a tool, Carafe, that generates high quality *in silico* spectral libraries by training directly on DIA data. Our approach uses a model architecture that is similar to the existing tool AlphaPeptDeep [16]. However, to enable training on chimeric spectra, we developed a two-pronged approach to accurately detect DIA fragment ion peaks that suffer from interference, i.e., peaks that are generated from fragmentation of two or more precursors. We then adapted our model training framework to support masking of these interfered peaks during training. We demonstrate the performance of Carafe on

both global proteome and phosphoproteome datasets.

2 Results

2.1 An overview of Carafe

Carafe consists of three modules for generating an *in silico* spectral library tailored to a specific DIA LC-MS/MS experiment setting of interest (Figure 1a). The first module generates training data for both fragment ion intensity and RT prediction from DIA data. The DIA data could be a single shot DIA-MS run generated from a sample from a model organism (e.g., a human cell line) under a specific LC-MS/MS setting of interest. This module accepts as input peptide detection results from DIA tools in a tab-separated values format and currently supports output produced by DIA-NN [24]. The second module is used to train the RT and fragment ion intensity prediction models using the training data generated by the first module. In practice, this step involves fine-tuning models that were previously trained using DDA data (Methods). The third module generates the *in silico* spectral library tailored to the specific LC-MS/MS setting used to generate the training DIA data by using the trained models from the second module. Once the model is trained, it can be used to generate *in silico* libraries for any species. Currently, Carafe is capable of generating spectral libraries in Apache Parquet column-oriented data storage format, row-oriented tab-separated values (TSV) data format compatible with DIA-NN, EncyclopeDIA [20] and Spectronaut, the blib format supported by Skyline [25], and the HUPO-PSI standardized spectral library format mzSpecLib [26].

When training a fragment ion intensity prediction model using DDA spectra, the training data is typically generated by matching peptides with their DDA spectra by assuming each DDA spectrum is generated from a single peptide. Therefore, for a given peptide spectrum match, all matched fragment ion peaks are considered to be solely contributed from the matched peptide (Figure 1b). However, in DIA, spectra are naturally chimeric. Consequently, some of the matched fragment ion peaks may be produced by multiple peptides (Figure 1b). A key feature of Carafe is its ability to accurately detect DIA fragment ion peaks that suffer from interference. Peaks with interference are often highly abundant because multiple fragment ions contribute to their intensities. Masking out such peaks during training is critical to achieving good performance. Carafe implements two distinct interference detection methods, and a peak is labeled as "shared" if either method detects interference (Figure 1b–c). The first, spectrum-centric approach identifies peaks in a single MS2 spectrum that are associated with at least two different detected peptides, whereas the second, peptide-centric approach looks for peaks that show correlation with the other fragment ions for a given peptide (Figure 1c, details in Methods). The second approach does not rely on detecting multiple peptides from the same MS2 spectrum to determine shared peaks. During model training, all shared peaks are masked out and thus do not contribute to the training loss. In principle, this peak masking training strategy could be easily adapted to work with any existing fragment ion intensity prediction framework by updating the loss function to support peak masking.

To evaluate the performance of the shared peak detection methods, we compared the Pearson correlation between observed fragment ion intensities and the fragment ion intensities predicted using the pretrained DDA model in AlphaPeptDeep for peaks detected as interfered by one method, the other, both, or neither on a TripleTOF 5600 human DIA dataset. Overall, we observed that 48.4% of detected peaks were detected as shared peaks by Carafe. Among these shared peaks, the vast majority (94.9%) were detected only by the peptide-centric method, with 2.2% detected only by the spectrum-centric method and 2.9% detected by both methods. Scatter plots of observed versus predicted fragment ion intensities (Supplementary Figure S1**The performance of the shared peak detection methods.** The detected peaks were divided into four groups: detected as interfered by one method, the other, both or neither (unshared). Each panel plots the predicted versus observed peak intensities, and the blue lines represent a smooth trend calculated using a loess method, highlighting the general pattern of correlation across the intensity range. The number of points and the Pearson correlation between observed and predicted fragment ion intensities is shown at the top of each panelfigure.1) show that the correlation is highest (0.82) for peaks detected as non-interfered by both methods, whereas the correlation is lowest (0.32) for peaks detected as interfered by both methods. This result supports the idea that peak interference can lead to reduced accuracy in fragment ion intensity prediction.

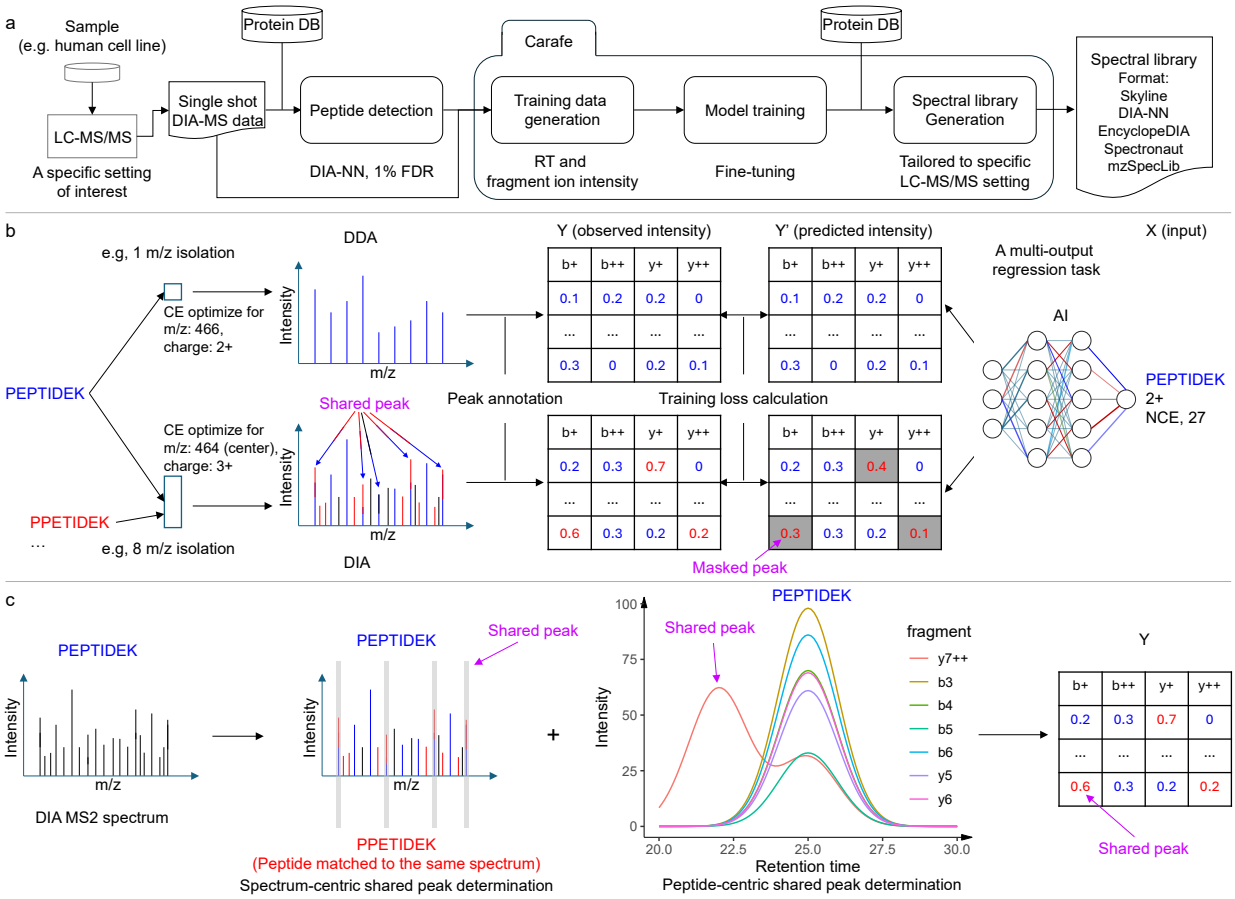

Figure 1: **Overview of Carafe.** (a) Using Carafe to generate experiment-specific *in silico* spectral library for DIA data analysis. (b) The difference of fragment ion intensity model training between using DDA data and using DIA data. (c) Shared peak determination in Carafe: the spectrum-centric approach (left) and the peptide-centric approach (right). The two methods are complementary and are performed independently.

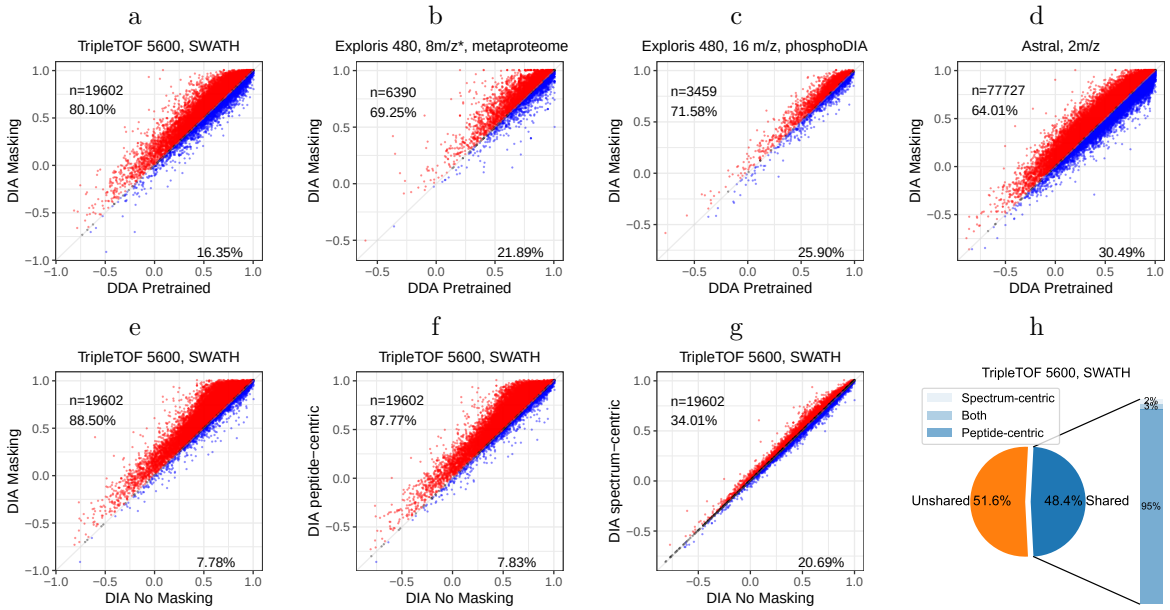

Figure 2: **Fine-tuning with DIA data improves fragment ion intensity prediction. (a–d)** Fragment ion intensity prediction performance comparison between a pretrained DDA model (x-axis) and models trained using DIA data on four different datasets (y-axis). **(e–g)** The performance of the shared peak detection methods in fragment ion intensity prediction. Four models were trained using the same human training data but with different peak masking strategies: no masking, spectrum-centric masking, peptide-centric masking, or both strategies. Each panel compares the Spearman correlation between observed and predicted fragment ion intensities for two different masking strategies, evaluated using the TripleTOF 5600 yeast DIA dataset. Each dot represents a peptidoform (peptide sequence + modification + precursor charge state). The values are Spearman correlations between observed fragment ion intensities and predicted intensities. Only peaks determined to be non-interfered were used in the correlation calculation. The percentages on each plot indicate the proportion of spectra for which one method achieves a higher or lower correlation than the other. **(h)** The proportion of shared peaks detected on the TripleTOF 5600 DIA dataset. "*": staggered isolation window.

2.2 Carafe yields accurate fragment ion intensity and RT predictions for diverse DIA datasets

To evaluate the accuracy of Carafe's fragment ion intensity predictions, we compared the performance of Carafe with the pretrained DDA model from AlphaPeptDeep, thereby testing whether the DIA fine-tuning step in Carafe is beneficial. For this analysis, we used four DIA datasets generated on three MS instrument types (TripleTOF 5600, Orbitrap Exploris 480, and Orbitrap Astral) from two vendors: two global proteome datasets, one metaproteome dataset and one phosphoproteome dataset. Each dataset contains DIA data generated from two different types of samples (a human sample and a yeast or a metaproteome sample). For each dataset, we used a single DIA run from human as training data and a single DIA run from yeast or the metaproteome as testing data. As our overall performance measure, we use the Spearman correlation between the observed and predicted fragment ion intensities, computed only with respect to unmasked peaks.

The results of this experiment show the benefits of fine-tuning with DIA data (Figure 2a–d). For example, on the TripleTOF 5600 dataset, we observe that fine-tuning the model yielded an improved Spearman correlation for 80.10% of the peptides, with an overall median increase in correlation of 0.05. We also observe substantial rates of improvement for the other three datasets (69.25% for the metaproteome data, 71.58% for the phosphoproteomics DIA data, and 64.01% for Orbitrap Astral data). Notably, the pretrained AlphaPeptDeep DDA model only used data from Thermo Scientific and Bruker instruments but provides a good starting point when fine-tuning with data from a TripleTOF instrument, which is from a different

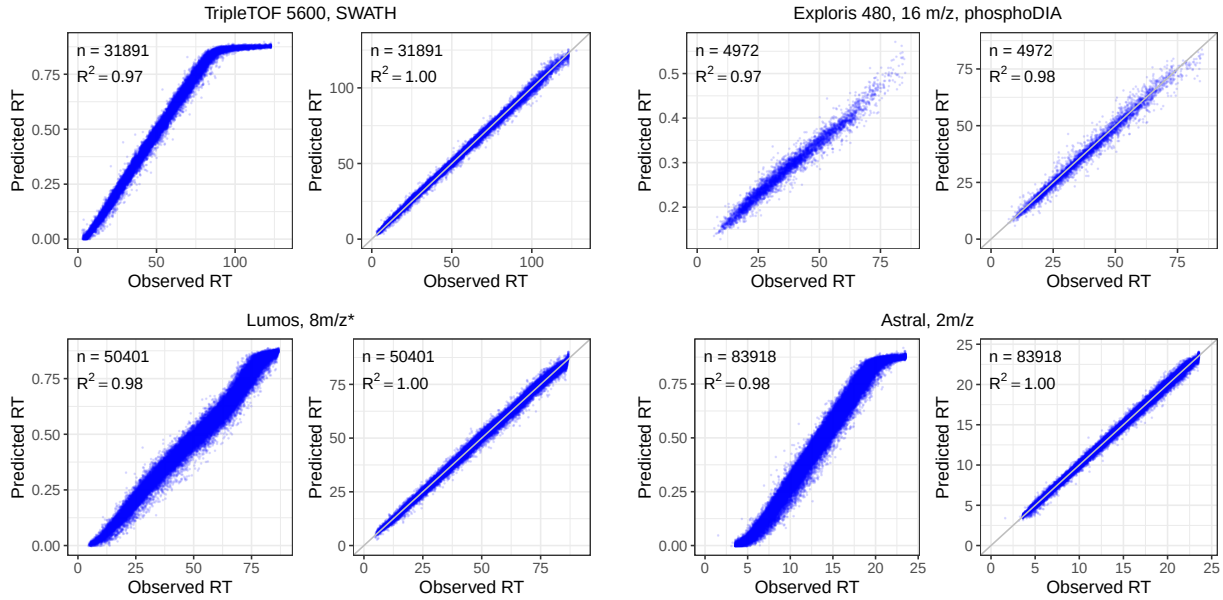

Figure 3: **Fine-tuning with DIA data improves retention time prediction.** Each panel plots observed RT versus RT predicted using the pretrained RT model (left panel) or fine-tuned RT model (right panel) in Carafe. On each dataset, the fine-tuned RT model was trained using the human DIA data. Both types of RT models were evaluated using the corresponding yeast DIA data. "*": staggered isolation window.

instrument vendor SCIEX (Figure 2a).

To further investigate the benefits provided by our peak masking methodology, we also performed an ablation experiment, in which we eliminated one, the other, or both of our two interference detection methods from Carafe. We compared the performance of the primary Carafe model with each of the three ablation models (peptide-centric only, spectrum-centric only, or no masking) by calculating the proportion of peptides with improved peak intensity predictions, as measured by Spearman correlation. As shown in Figure 2e-g, the best performance is achieved when both methods are used. Between the two interference detection methods, the peptide-centric approach appears to be more effective than the spectrum-centric one, as evidenced by the fact that the peptide-centric method alone outperforms the spectrum-centric method alone, and adding the spectrum-centric method to the peptide-centric method only slightly improves the performance. This observation is expected because most of the shared peaks are determined by the peptide-centric method as shown in Figure 2h. These results suggest that the two interference detection methods are complementary.

Finally, we also confirmed the quality of our RT predictions by repeating a similar setup (train on human and test on yeast) on four DIA datasets generated using different LC settings, including three global proteome DIA datasets and one phosphoproteomics DIA dataset. As illustrated in Figure 3, the RT models fine-tuned with DIA data showed superior performance over the pretrained DDA model, as evidenced by the improved squared correlation coefficient (R^2) values. The fine-tuned RT models consistently achieved R^2 values exceeding 0.98 across the datasets. Remarkably, the fine-tuned RT models consistently achieved much better linear correlation and tightening of retention time accuracy across the datasets while the pretrained DDA model showed non-linear correlation on some of the datasets. On two of the DIA datasets (TripleTOF 5600 and Astral), the predictions from the pretrained DDA RT model suggest that the peptides eluted very late have very different elution behavior between the data used to train the DDA RT model and the DIA data. However, after fine-tuning the pretrained DDA RT model using the human DIA data, the elution behavior was learned effectively by the model as evidenced by consistent linear correlation across the whole LC gradient.

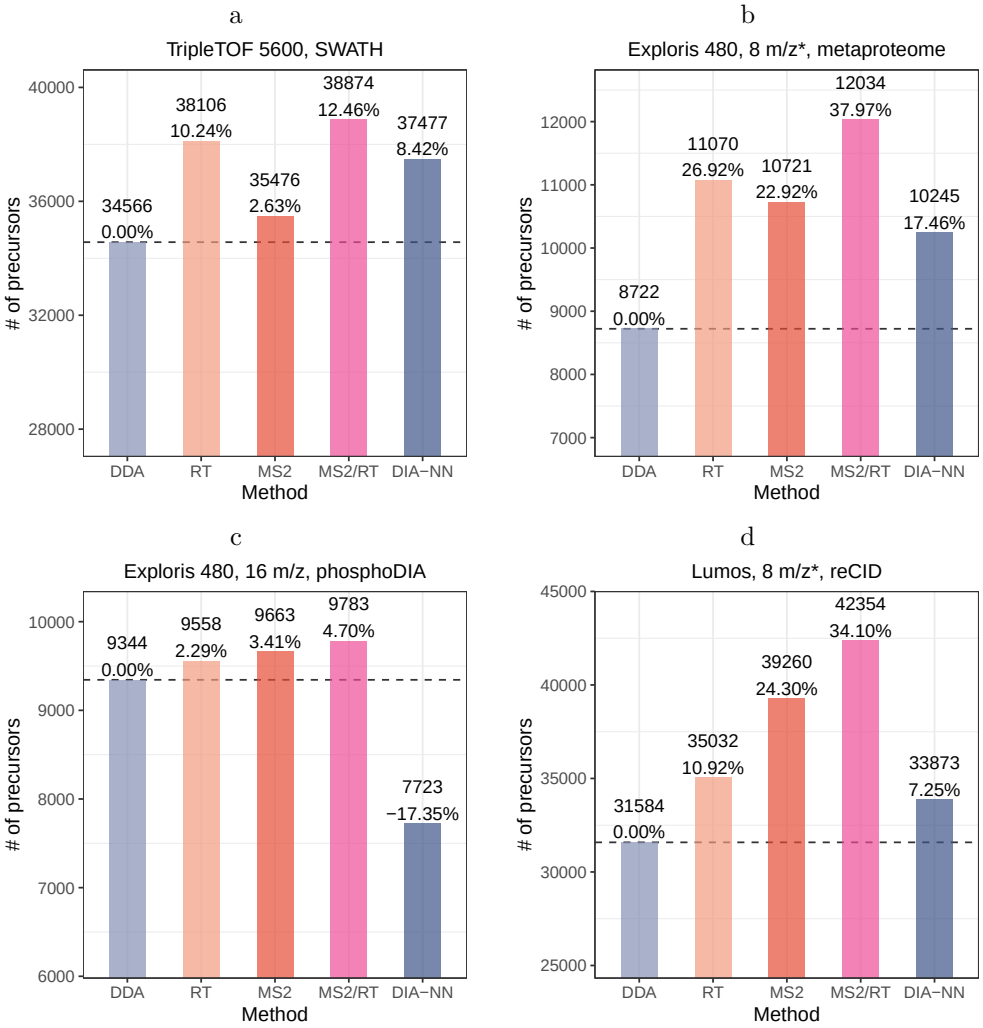

Figure 4: **Fine-tuning fragment ion intensity and RT prediction models improves peptide detection on DIA datasets.** *In silico* library comparison generated using different methods on four different datasets. **(a)** A yeast dataset generated on a TripleTOF 5600 instrument. **(b)** A metaproteomics dataset generating on an Exploris 480 instrument. **(c)** A yeast phosphoproteome DIA dataset generated on an Exploris 480 instrument. **(d)** A yeast dataset generated using resonance-type collision-induced dissociation (reCID) fragmentation method on a Lumos instrument. Each dataset includes human DIA data as training data which was generated using the same LC and MS settings as the yeast or metaproteome data. For each dataset, the number of precursors accepted at the 1% precursor-level FDR are shown. The number on each bar is the percent improvement when comparing to the library generated using the pretrained DDA models. "*": staggered isolation window. For x-axis, DDA: spectral library generated using the pretrained DDA model from AlphaPeptDeep; MS2: spectral library generated using fine-tuned fragment ion intensity prediction model but pretrained RT model from AlphaPeptDeep; RT: spectral library generated using fine-tuned RT prediction model but pretrained fragment ion intensity model from AlphaPeptDeep; MS2/RT: spectral library generated using fine-tuned fragment ion intensity and RT prediction models; DIA-NN: spectral library generated using DIA-NN.

2.3 Carafe increases the number of peptides detected with DIA-NN

Having established the accuracy of Carafe's predictions, we next evaluated the utility of its *in silico* spectral libraries on four different types of DIA datasets. For each dataset, we first compared three different spectral libraries, one generated using the DDA-trained AlphaPeptDeep model (referred to as "DDA" in Figure 4), one generated by Carafe's fully trained model using DIA data (referred to as "MS2/RT" in Figure 4), and one built into DIA-NN (referred to as "DIA-NN" in Figure 4). The latter corresponds to using DIA-NN in its "library-free" mode. As above, we trained each model using human data and evaluated the model using the corresponding yeast or metaproteome data. In this setting, the performance measure is the total number of precursor (i.e., a combination of peptide sequence, modification and precursor charge) detected by DIA-NN at a 1% precursor level FDR threshold using the given spectral library.

The results of this experiment show that Carafe offers a consistent gain in statistical power to detect precursors relative to either of its competitors (Figure 4). Comparing the DDA-trained AlphaPeptDeep model to Carafe's fully DIA-trained model, we observe an increase of 4.7%-38.0% in the number of detected precursors. Furthermore, when comparing with DIA-NN library free search, we identified 3.7%-26.7% more precursors using Carafe fine-tuned models on the four DIA datasets.

Next, to better understand the source of Carafe's improved statistical power, we created two additional *in silico* spectral libraries: one library in which the RT prediction model is fine-tuned using DIA data and the fragment ion intensity prediction model is taken directly from AlphaPeptDeep (referred to as "RT" in Figure 4), and a second library that is the converse (i.e., only the fragment ion prediction model is fine-tuned, referred to as "MS2" in Figure 4). The results of this experiment (Figure 4) clearly show that both types of fine-tuning are necessary: in all four cases, the fully trained Carafe outperforms either the RT-only or MS2-only models. Furthermore, both of the semi-trained models consistently outperform the DDA-trained model. However, the relative performance of RT-only and MS2-only differs by dataset, suggesting that both types of prediction are necessary and neither is consistently more important than the other.

Notably, to demonstrate the generalizability of Carafe, we generated a DIA dataset using resonance-type collision-induced dissociation (reCID) fragmentation method on a Lumos instrument for both a human cell line sample and a yeast sample. The reCID method is expected to generate spectra with very different fragment ion intensity patterns compared with spectra generated using the widely used beam-type CID method (HCD) [27]. In this study, all other datasets were generated using HCD. As shown in Figure 4d, Carafe significantly outperforms both the DDA-trained model and DIA-NN built-in model. Fine-tuning the fragment ion intensity prediction model alone on the human DIA data improves the number of detected precursors on the yeast DIA data by 24.3% compared with the pretrained DDA model.

2.4 Carafe produces improved fragment ion intensities and RTs relative to a library derived from GPF DIA data

To improve the quality of spectral libraries for DIA data analysis, the search tool EncyclopeDIA [20] implements an alternative approach to incorporating DIA data into a spectral library. In brief, a set of narrow window GPF DIA runs are first searched against an *in silico* spectral library generated using a tool such as Prosit [12] to generate a chromatogram library that is tailored to a specific LC-MS/MS setup. In addition to the retention times and fragment ion intensities stored in a typical spectral library, this chromatogram library also includes information about peak shapes and known interferences. Notably, this chromatogram library is substantially smaller than a standard proteome-wide spectral library, because the chromatogram library only contains entries for peptides that can be confidently detected from the GPF DIA data. The chromatogram library is then used to analyze wide window DIA data for peptide quantification. The wide window DIA data is typically generated using a wide isolation window scheme using the same LC setting and the same MS instrument as the GPF DIA data. Empirical evidence suggests that using GPF DIA data in this way yields improved correlation to spectra in DIA data, relative to correlating DDA data to spectra in DIA data [20].

We aimed to compare Carafe's spectral libraries, which are produced by a model fine-tuned using DIA, against the EncyclopeDIA's DIA-based chromatogram libraries. Accordingly, we generated four different types of spectral libraries for a single DIA dataset generated using a Q Exactive HF-X instrument. The first two libraries are *in silico* spectral libraries generated using Prosit and Carafe models, respectively. The

next two libraries are DIA-based chromatogram libraries produced by searching a set of GPF DIA runs using EncyclopeDIA against the Prosit-predicted *in silico* library ("Prosit+GPF") and against the Carafe-predicted *in silico* spectral library ("Carafe+GPF"). We then searched a set of wide window DIA runs from this dataset against each of the four spectral libraries using EncyclopeDIA.

Our results suggest that the Carafe library improves upon the Prosit library. We observed that, subject to a 1% peptide-level FDR threshold, searching using the Carafe library yields 22.6% more peptides than searching using Prosit (Figure 5a). Similarly, between the two GPF run-derived libraries, Carafe+GPF yields 3.2% more peptides than Prosit+GPF.

On the other hand, we also observe that Prosit+GPF outperforms Carafe alone slightly, yielding a 1.6% increase in the number of detected peptides. In practice, Prosit+GPF differs from the Carafe library in two ways. First, Prosit+GPF contains only those peptides that can be confidently detected in the GPF data (44,903 precursors), whereas the Carafe library contains all peptides in the reference proteome (661,012 precursors). Second, Prosit+GPF contains two types of information—peak shape and known interferences—that are absent from the Carafe library. To isolate the effects of each of these differences, we therefore created two additional libraries: "Carafe-reduced" is a reduced version of the Carafe library that only contains the peptides that also appear in the Prosit+GPF library; and "Prosit+GPF-reduced" is the same as Prosit+GPF, except that the peak shape and known interference information is eliminated from the library.

We hypothesized that the difference in library size largely accounts for the observed difference in statistical power between Prosit+GPF and Carafe. Indeed, comparing Prosit+GPF and Carafe-reduced, we find that 4.2% more peptides were identified using the Carafe-reduced library. Furthermore, as shown in Figure 5a, removing peak shape and known interference information from the Prosit+GPF library only slightly reduced the number of identified peptides. These results indicate that the quality of predicted RTs and fragment ion intensities in the Carafe library is higher than the quality of RTs and fragment ion intensities derived from GPF DIA data.

We repeated this analysis on other three DIA datasets generated using three different instruments from two different mass spectrometer vendors. As shown in Figure 5b–d, comparing Prosit+GPF and Carafe-reduced libraries, we find that 18.1%, 6.6% and 0.7% more peptides were identified using the Carafe-reduced library on the TripleTOF 5600, Orbitrap Astral and Orbitrap Fusion Lumos datasets, respectively. Remarkably, the Carafe library outperforms the Prosit+GPF library on both the TripleTOF 5600 and the Astral datasets.

Taken together, these results demonstrate the superior quality of Carafe libraries fine-tuned with DIA data in comparison to DIA-based chromatogram libraries, while also suggesting that using GPF data to limit the number of peptides in the library is helpful in some cases.

3 Discussion

We have demonstrated that Carafe enables high-quality *in silico* spectral library generation tailored to specific LC-MS/MS setups by training deep learning models directly on DIA data. Carafe provides superior peptide detection on a wide range of DIA datasets generated from different mass spectrometer vendors. Compared with spectral libraries generated using DDA data trained models, Carafe's libraries improve the number of detected peptides by 4.7%-38.0%. Furthermore, our results suggest that the quality of Carafe's predictions is higher than the quality of RTs and fragment ion intensities derived from GPF DIA data. The major innovation of Carafe is its ability to train a fragment ion intensity prediction model directly on DIA data. This is achieved through an interference peak detection method used for training data generation and a peak masking strategy used in model training to handle naturally chimeric spectra in DIA data. In principle, this peak masking training strategy could be easily adapted to work with any existing fragment ion intensity prediction framework by updating the loss function to support peak masking. Thus, our method is general and can be implemented into DIA search engines to improve peptide detection by generating experiment-specific, high-quality *in silico* spectral libraries.

With transfer learning, we show that peptides detected on a single human cell line MS run are sufficient to train high quality models for experiment-specific spectral library generation for samples from different organisms and that this strategy generalizes well across different instruments and settings. This approach will remove the burden of generating a new empirical spectral library for a new instrument or an instrument

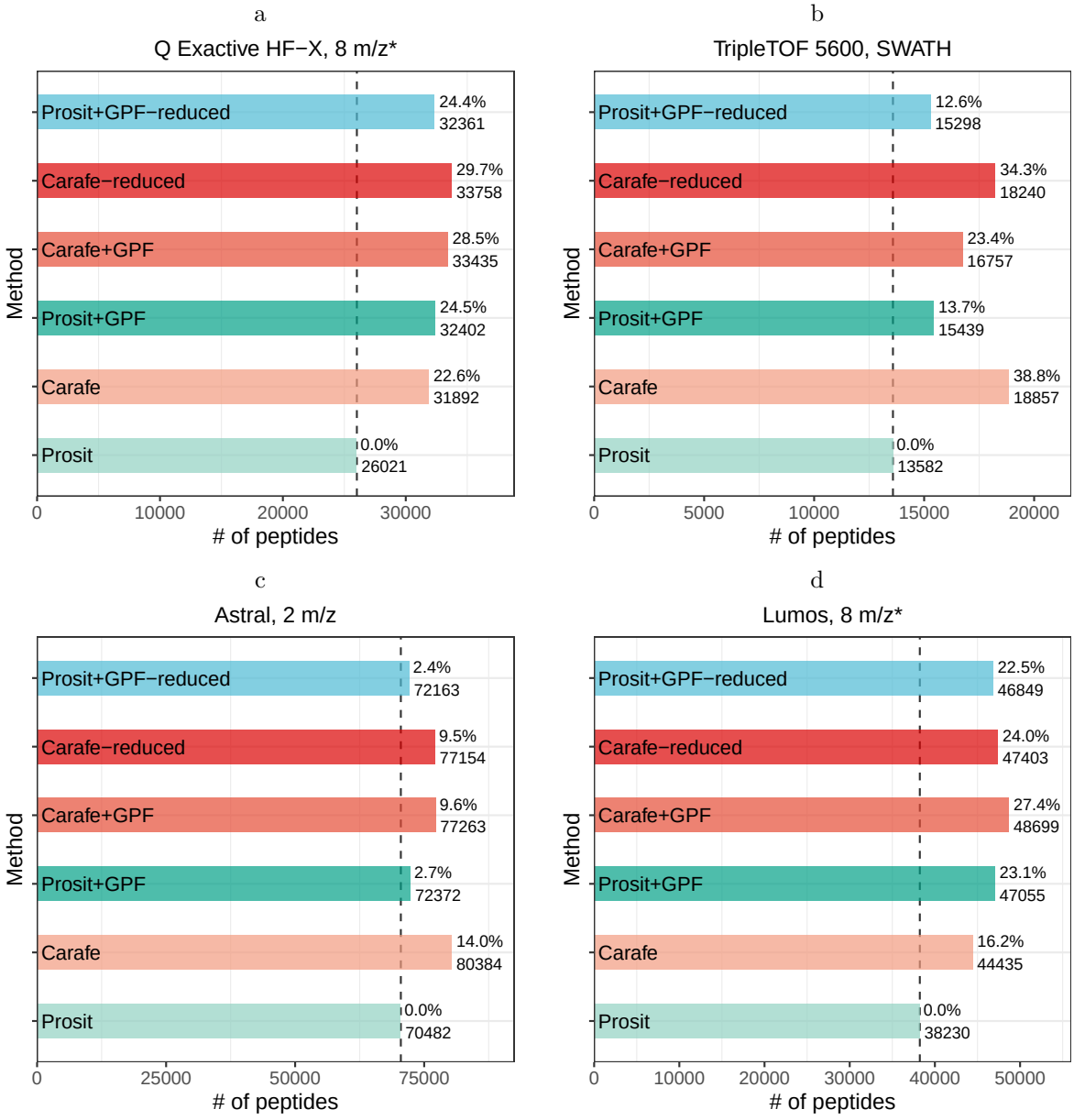

Figure 5: **Comparing Carafe generated libraries with DIA-based chromatogram libraries.** Peptides identified using different spectral libraries on **(a)** a dataset generated using a Q Exactive HF-X instrument, **(b)** a dataset generated using a TripleTOF 5600 instrument, **(c)** a dataset generated using an Orbitrap Astral instrument and **(d)** a dataset generated using an Orbitrap Fusion Lumos instrument. For each dataset, the number of peptides accepted at the 1% peptide-level FDR is shown. The ratio on each bar is the percent improvement when comparing to the library generated using the Prosit model. "*": staggered isolation window.

with different MS settings (such as different fragmentation methods). The proposed method is also fast to run on a CPU since the training data from a single MS run is small. To generate an experiment-specific *in silico* spectral library using Carafe, we recommend generating a single MS DIA run with the same LC and MS instrument settings as the DIA runs to be analyzed using a complex cell lysate from a model organism, such as a yeast or a human cell line sample. Thereafter, the user can fine-tune Carafe models using peptides detected from the single MS run.

Because spectral libraries are essential for DIA data analysis, we expect that Carafe will be a valuable tool for the proteomics community. In the future, we plan to further improve the speed of Carafe when running on a CPU platform. We also plan to support more instruments such as the DIA data generated from Stellar and timsTOF instruments.

4 Methods

4.1 Deep learning models

For fragment ion intensity prediction, we used the transformer framework implemented in AlphaPeptDeep [16], with modifications to support peak masking. The model consists of an embedding layer, a positional encoder layer, and four transformer layers followed by two fully-connected layers. The inputs to the embedding layers include amino acid sequences, modifications, and metadata including peptide charge state, normalized collision energy, and instrument type. These inputs are encoded separately and are concatenated for the following layers. The output of the last layer for each peptide is a two-dimensional matrix in which each row represents a backbone position along the peptide and each column represents a fragment ion type with a specific charge state (e.g., b1+). To accommodate shared fragment ions, we modified the AlphaPeptDeep loss function to mask selected peaks. With peak masking enabled, the masked peaks do not contribute to the overall model loss. We started with a pretrained AlphaPeptDeep model, which had been trained using large-scale DDA data using model weights downloaded from `https://github.com/MannLabs/alphapeptdeep`. We then fine-tuned this model using a training set derived from DIA spectra. The training parameters were set as follows: epoch, 20; warmup epoch, 10; learning rate, 0.0001; batch size, 512. L1 loss was used for training. Both MS instrument type and normalized collision energy (NCE) were extracted from the corresponding training DIA data and used as metadata inputs to the model.

For peptide RT prediction, the hybrid CNN+LSTM model from AlphaPeptDeep was used. We fine-tuned its pretrained models using peptides detected from DIA data. The training parameters were set as follows: epoch, 40; warmup epoch, 10; learning rate, 0.0001; batch size, 1024. L1 loss was used for training.

4.2 Training and testing data generation

We trained two different types of prediction models (RT and fragment ion intensity prediction) for each dataset. For each dataset, we trained using the human DIA data and tested using the paired yeast or metaproteome DIA data. For each training or testing DIA MS run, we used DIA-NN (version 1.8.1) [24] in library-free mode to detect peptides at 1% precursor-level FDR. For the global proteome datasets, all accepted peptides were used in training and testing data generation, whereas for the phosphoDIA dataset, two additional filters were applied: PTM.Q.Value ≤ 0.01, and PTM.Site.Confidence ≥ 0.75.

To prepare training data for fragment ion intensity prediction, for each precursor identified by DIA-NN, we first extracted its matched MS2 spectrum at the apex time point from the corresponding DIA data file. We then calculated all matched fragment ions by comparing the peptide against the MS2 spectrum using a predefined mass tolerance (e.g., 10 ppm). During the matching, b and y ions were considered for global proteome data while b and y ions plus phosphorylation neutral loss (-98 Da) ions were considered for phosphoproteome data. We then used a two-stage strategy to identify fragment ions that are shared with other peptides (Figure 1c). First, in the spectrum-centric shared peak determination step, we examined all other peptides matched to this MS2 spectrum. If a peak from the MS2 spectrum was shared by any other peptide, then the peak was labeled as "shared." Note that some shared peaks may not be detected using this method, if some of the peptides that generated the DIA spectrum were not identified. Therefore, we implemented a second, peptide-centric strategy for detecting shared peaks. Specifically, for each matched fragment ion from the spectrum-centric strategy, we extracted its extracted ion chromatogram (XIC) based

on the peak boundaries determined by DIA-NN. For each matched fragment, we first performed Savizky-Golay smoothing [28] or weighted moving average smoothing on the fragment ion chromatograms and then calculated the Pearson correlation between the smoothed XIC and each of the smoothed chromatograms from the other matched fragment ions. We assigned an aggregate score to each matched fragment ion by summing up the Pearson correlation scores between this fragment and all other matched fragments, and we identified the top-scoring fragment ion based on this aggregate score. Finally, a peak was labeled as shared if the correlation of a fragment to the top-scoring fragment ion was less than a predefined threshold (by default, 0.75). For each peak with correlation score higher than the predefined threshold, we further calculated a peak shape score. Specifically, at each boundary time point of the peak, if the intensity of the peak at the boundary time point was higher than the median intensity of all detected peaks at the boundary time point, and the intensity was also higher than 10% (we increased the cutoff from 10% to 25% when the highest intensity of the peak was less than 50% of the highest intensity of all detected peaks) of the highest intensity of the peak between its left boundary and right boundary, we assigned a score of 1 to the peak when only one of the two boundaries meets the criteria, and a score of 2 when both boundaries meet the criteria. If the peak shape score is 2, then the peak was also labeled as shared.

Overall, any peak identified as a shared peak by either of the two strategies (spectrum-centric or peptide-centric) was labeled as shared. In addition, any peak that matched to more than one peptide fragment ion (e.g., a b-ion and y-ion from the same peptide with similar m/z values) was also taken as a shared peak. Intensities of unmatched fragment ions were set to zero, unless the fragment m/z falls outside the m/z scan range, in which case the ion was masked during model training. Finally, for each annotated spectrum, the matched fragment ion intensities were linearly normalized so that the highest matched peak had a height of 1. If the highest matched peak was a shared peak or the percentage of shared peaks exceeding a specific threshold (e.g., 50%), then the spectrum was not used for training or testing.

To prepare training data for RT prediction, we followed the procedure outlined in a previous study [15]. Specifically, for a peptide form (peptide sequence + modification) with multiple precursors identified with different charge states, if the difference between the maximum and minimum RT values was less than a predefined threshold (by default, 3 minutes), then the average RT value was used as the observed RT of this peptide form. Otherwise, this peptide form was removed and was not used for training or testing.

4.3 Spectral library generation using Carafe

To generate an experiment-specific *in silico* spectral library, a single DIA run was required and was analyzed using DIA-NN. Then a fragment ion intensity prediction model and a retention time prediction model were trained using the method described above. In the RT model fine-tuning step, 10% of peptides were used for testing. For a given protein database, we first performed *in silico* protein digestion based on a specified enzyme, the maximum allowed missed cleavage sites, and peptide length limits. Next, for each digested peptide, different peptidoforms were generated based on specified fixed and variable modifications, as well as a specified precursor change range (in default, 2-4). The fixed modification was set as Carbamidomethyl (C) in default. No variable modification was applied for global proteome data while phosphorylation (STY) was set as variable modification for phosphoproteome data in default. Finally, we applied the trained models to predict fragment ion intensity and RT for all the peptidoforms to generate a spectral library. During spectral library generation, only precursors and fragment ions with m/z values within the scan range of the corresponding DIA data used for model training were included in the library. For each precursor, in default the top 20 fragment ions were selected based on the predicted fragment ion intensity.

4.4 Proteome sample preparation

4.4.1 Human cell culture and treatment

HeLa S3 cells were cultured at 37 °C and 5% CO2 in Dulbecco's modified Eagle's medium (DMEM) supplemented with 4.5 g/L glucose, l-glutamine, 10% fetal bovine serum (FBS), and 0.5% streptomycin/penicillin. Cells were grown to an 80% confluency. At the time of harvest, cells were left attached to plates, rinsed three times quickly with ice-cold PBS then flash-frozen in liquid nitrogen prior to storage at -80 °C.

4.4.2 Yeast growth

The S288C *S. cerevisiae* strain was selected for all downstream applications. Yeast were grown on a plate overnight, and single colonies were inoculated into media containing yeast extract peptone dextrose (YEPD). The cultures were grown to OD 0.6 before harvested, pelletted, and frozen at -80 °C until use.

4.4.3 Protein aggregation capture (PAC) sample preparation

Protein from HeLa and yeast samples were prepared for LC-MS using a protocol modified based on magnetic bead based tryptic digestion methods reported previously [29, 30]. Briefly the cells were lysed in a buffer for 5 minutes containing 2% SDS, 100mM Tris buffer pH 8.5, and ThermoFisher protease inhibitors. Each sample was briefly sonicated on the Branson probe sonicator for 5 seconds, and protein concentration was determined using the bicinchoninic acid (BCA) method calibrated using bovine serum albumin (Pierce, Thermo Fisher scientific) [31]. Samples were then diluted to a final concentration of 1 µg/µL. The diluted sample lysates were reduced in 20 mM dithiothreitol and alkylated in 40 mM iodoacetamide. ReSyn Hydroxyl beads were added to the reduced and alkylated samples at a ratio of 4 µL per 25 µg of protein. Protein aggregation onto the hydroxyl beads was induced by adding acetonitrile to a final concentration of 70%. The bead-bound proteins were further subjected to washes containing three 95% acetonitrile and two 70% ethanol washes. After the final wash, the samples were briefly centrifuged to remove any residual ethanol, and trypsin in 50 mM ammonium bicarbonate was added in at a ratio of 33:1 (protein to trypsin) for digestion at 47°C for 3 hours. The resulting sample peptides were eluted off of the beads, dried down by a centrifuge vacuum speedvac, and frozen in the -80°C until further use. Frozen peptide samples were resuspended to a final concentration of 500 ng/µL in 0.1% formic acid prior to mass spectrometry analysis.

4.5 Phosphoproteome sample preparation

4.5.1 Human cell culture treatment and lysis

HeLa S3 cells were cultured at 37°C and 5% CO2 in Dulbecco's modified Eagle's medium (DMEM) supplemented with 4.5 g/L glucose, l-glutamine, 10% fetal bovine serum (FBS), and 0.5% streptomycin/penicillin. To generate bulk phosphopeptides for method comparisons, cells were grown to 80% confluency, incubated in serum-free medium for 6 h prior to treatment with or without 1 mM pervanadate for 15 min, followed by the addition of 10% FBS for 15 min. At the time of harvest, cells were left attached to plates, rinsed three times quickly with ice-cold PBS then flash-frozen in liquid nitrogen prior to storage at −80°C. Cells were harvested by scraping frozen cells from plates in 8 M urea, 50 mM HEPES, 75 mM NaCl, pH 8.0. Cells were sonicated with six 20 s pulses at 12 W with equal rests in ice. The lysate was clarified by centrifugation at 7197 x g for 25 min at 20°C. The protein concentration was estimated using the bicinchoninic acid method (Pierce, ThermoFisher Scientific). Proteins were reduced with 5 mM dithiothreitol (DTT) for 30 min at 55°C, alkylated with 15 mM iodoacetamide for 15 min at room temperature in the dark, and then quenched with 5 mM DTT for 15 min at room temperature.

4.5.2 Yeast growth and lysis

The haploid S288C derivative BY4742 *S. cerevisiae* strain was used for global phosphopeptide enrichment. Yeast were grown in yeast extract peptone with 2% glucose (YEPD) at 30°C. Overnight cultures were diluted to OD_{600} of 0.1 then grown to OD_{600} 0.8-0.9 before cells were harvested by washing with ice cold water and snap freezing in liquid nitrogen. Cells were lysed by bead-beating with 0.5 mm glass beads at 4°C in urea lysis buffer (8 M urea, 50 mM HEPES pH 8.0, 75 mM NaCl). Lysates were cleared by centrifugation at $21,000 \times g$ for 10 minutes at 4°C, and protein concentration was determined by bicinchoninic acid assay (Pierce, ThermoFisher Scientific). Proteins were reduced with 5 mM dithiothreitol for 30 minutes at 55°C, alkylated with 15 mM iodoacetamide for 30 minutes at room temperature in the dark, and the alkylation was quenched with an additional 5 mM dithiothreitol for 15 minutes at room temperature.

4.5.3 Digest of HeLa and yeast for global phosphopeptide enrichment

Protein lysates were diluted five-fold in 50 mM ammonium bicarbonate and digested by trypsin at a final trypsin-to-protein ratio of 1:100 by mass. Proteins were digested at 37°C for 15 hours with mixing. Digests were quenched with 0.5% TFA (pH < 2). Quenched digests were centrifuged at 7000 x g for 5 min at room temperature to remove precipitates. Peptides were desalted on Waters SEP-PAK C18 cartridges. Briefly, columns were activated by the sequential addition of 1 column volume (CV) of methanol; 3 CVs of 100% acetonitrile (ACN); 1 CV of 70% ACN, 0.25% acetic acid (AA); 1 CV of 40% ACN, 0.5% AA; 3 CV 0.1% TFA. Acidified digests were then loaded followed by reload of the flowthrough. The column was washed with 3 CV of 0.1% TFA and 1 CV of 0.5% AA. Peptides were eluted with 0.75 CV of 40% ACN, 0.5% AA followed by 0.5 CV of 70% ACN, 0.25% AA. Peptides were dried by vacuum centrifugation and stored at −20°C until enrichment. Phosphotyrosine containing peptides were initially depleted by phosphotyrosine-specific enrichment (data not included here) [32]. Remaining peptides from flow throughs were desalted, dried, and stored at −20°C for global phosphopeptide enrichment.

4.5.4 Global phosphopeptide enrichment

Desalted peptides were resuspended in 80% ACN, 0.1% TFA at 900 µL per 250 µg peptides. Precipitates were removed by centrifugation at 21000 x g for 5 min at 4°C and peptides were added to the 96-well plate for R2-P2 as described previously [33] except with the following modifications: peptide binding was performed in a deep well plate in 900 µL and phosphopeptides were eluted in 100 µL instead of 50 µL of 2.5% ammonium hydroxide, 50% ACN followed by acidification with 60 µL of 10% formic acid, 75% ACN. The optional filtering step was performed in which eluates were passed through two layers of C8 filter material in a 200 µL pipette tip as described previously [33]. Peptides were dried by vacuum centrifugation then resuspended in 4% formic acid, 3% acetonitrile for mass spectrometry measurement.

4.6 Metaproteomics sample preparation

The marine microbiome sample for metaproteomic analysis was collected June 4, 2021 at 1:00 pm PDT in East Sound, WA according to methods detailed in [34]. The sample filter (0.22 µm 47mm polyethersulfone) containing the bacterial fraction of the water column (0.22 - 1.0 µm) was processed using mechanical lysis in 100 µL 5% SDS solution followed by three subsequent rinses of the filter with 100 µL nanopure water. The resulting 400 µL whole cell solution was collected in microfuge tubes and sonicated (Branson 250 Sonifier; 20 kHz, 30 × 10 s on ice). Samples were then evaporated using a SpeedVac to a final concentration of 5% SDS in 100 µL. Enolase (0.16 µL 100 ng μL^{-1} enolase per 1 µg protein) was added to the sample at the start of the S-trap protocol to ensure proper sample digestion. The sample (20 µg protein) was treated with benzonase (0.5 µL 250 unit μL^{-1} 10 minutes at 95°C), reduced with 20 mM dithiothreitol for 10 minutes at 60°C and 5 minute cool down to room temperature, alkylated with 40 mM iodoacetamide for 30 minutes in the dark, acidified to pH < 2 (1.2% aqueous phosphoric acid), and then processed on an S-trap column according to manufacturer's recommendations. Proteins were digested with Promega modified trypsin (2 µg for 1:10 ratio, 4 hours 37°C). Purified peptides were evaporated to dryness and resuspended in 2% acetonitrile (ACN), 0.1% formic acid with final concentration of 0.5 µg protein μL^{-1}.

4.7 Liquid chromatography analysis

Liquid chromatography separation of peptides was done prior to the data acquisition on the mass spectrometer. Peptides were separated using a 15 cm PepSep column (150µm ID, 1.9µm particle size) heated to 45°C. Reversed-phase separation on a Neo Vanquish LC system was used in conjunction with a Thermo Astral and Thermo Exploris 480. Total proteome measurements made on the Neo Vanquish with the Thermo Astral were separated by a 24 minute gradient at 1.3 µL/min flow consisting of 4 to 6% Buffer B in 0.7 minutes, 6 to 6.5% Buffer B in 0.3 minutes, 6.5 to 40% Buffer B in 20 minutes, 40 to 55% Buffer B in 0.5 minutes, 55 to 99% Buffer B in 3.5 minutes for the column wash. Total proteome data acquired on the Thermo Exploris were separated on a 90 minute gradient consisting of 4 to 6% Buffer B with 1.3 µL/min flow for 0.7 minutes, 6 to 6.5% Buffer B with 1.3 µL/min flow for 0.3 minutes, 6.5 to 5% Buffer B with 0.8 µL/min flow for 82.7 minutes, 50% Buffer B with 1.3 µL/min flow for 0.5 minutes, 50 to 99% Buffer B with 1.3 µL/min flow for

2.3 minutes, and 99% Buffer B for 3.5 minutes for the column wash. For phospho-enriched samples on the Exploris, the Neo Vanquish gradient was changed to 86.2 minutes and consisted of 1 to 40% Buffer B with 0.8 µL/min flow for 82.7 minutes, 40 to 99% Buffer B with 1.3 µL/min flow for 0.5 minutes, and 99% Buffer B with 1.3 µL/min flow for 3 minutes for the column wash. For total proteome samples collected on the Thermo Lumos Tribrid Orbitrap mass spectrometers, an EvoSep One LC system running the Extended 88 minute method was used or an EASY-nLC 1200 system running a 90 minutes LC gradient was used. Specific details regarding LC conditions are provided in the Supplemental Table 1.

4.8 Data independent acquisition mass spectrometry

Human and yeast data were acquired using data-independent acquisition (DIA) on the Thermo Astral, Thermo Lumos Fusion Tribrid, and Thermo Exploris 480 mass spectrometers. Within a batch of runs, a chromatogram library containing 4 (on the Astral) or 6 (on the Lumos Fusion Tribrid and Exploris 480) independent injections of the samples were analyzed. On the Orbitrap Astral, each library injection consisted of a cycle of one 240,000 resolution full MS spectrum with a mass range of 375-985 *m/z*, custom injection time at 50 milliseconds (ms), and standard target AGC. The full MS1 spectrum was followed by a data-independent MS2 spectrum using the Astral analyzer with 1 *m/z* isolation window with 27% HCD collision energy, 10 ms injection time, 250% target AGC, and varying 125 *m/z* wide precursor mass ranges (400−525, 525−650, 650−775, 775−900 *m/z*) for each injection. On the Lumos Fusion tribrid and Exploris 480, each of the 6 library injections consisted of one cycle of 30,000 resolution full MS spectrum with a mass range of 395−1005 *m/z*, automatic injection time, and standard target AGC. The full scan was followed by a data-independent MS2 spectrum at 30,000 resolution using the Orbitrap analyzer with 4 *m/z* staggered isolation window with 27% HCD collision energy, "auto" maximum injection time on the Exploris 480 and 54 ms maximum injection time on the Fusion Lumos Tribrid, 1000% target AGC, and varying 100 *m/z* wide precursor mass ranges (400-500, 500-600, 600-700, 700-800, 800-900, 900, 1000 *m/z*) for each injection.

The individual human and yeast sample runs of DIA data on the Orbitrap Astral consists of one cycle of a 240,000 resolving power spectrum on the orbitrap with custom injection time of 50 milliseconds and standard target AGC, and an MS2 spectrum collected subsequently using the Astral detector using 2 *m/z* non-staggered isolation window with 5 ms injection time, 400-900 *m/z* precursor mass range, 27% HCD collision energy, and 250% target AGC. For the Exploris 480 and Fusion Lumos Tribrid, the human and yeast sample DIA acquisition contained one cycle of 30,000 resolution with a scan range of 395-1005 *m/z*, auto injection time and standard AGC. The following MS2 spectrum was performed using the Orbitrap at 8 *m/z* staggered isolation windows with 15k resolution. The precursor MS1 spectrum was 400-1000 *m/z* and 27% HCD collision energy remained the same for both instruments. On the Exploris 480, the *m/z* range was 200-2000 with automatic injection time and target AGC of 1000%. On the Fusion Lumos Tribrid, the *m/z* range was 150-2000 with an injection time of 22 ms for a target AGC of 800%. Individual metaproteomic samples were only collected on the Exploris 480 using the 8 *m/z* staggered window and 15k resolution DIA method. For phospho-enriched samples, the above parameters remained the same except the MS1 precursor spectrum range of the MS2 was modified to 450-1050 *m/z* on the Thermo Exploris 480 mass spectrometer with an isolation window of 16 *m/z*. For the reCID experiment, the MS2 spectrum was only performed at 8 m/z with a staggered window and 15k resolution with 27% reCID activation energy. The other parameters remained consistent for these runs on the Fusion Lumos tribrid for human and yeast samples. More details regarding DIA methods are provided in the Supplemental Table 1.

4.9 MS data conversion

Raw MS/MS data were were converted to mzML format files using MSConvert in ProteoWizard (version 3.0.24031) [35]. The vendor's peak picking was enabled for both staggered and unstaggered isolation window DIA data while demultiplexing was enabled for the staggered isolation window DIA data. The MS/MS data in mzML format were then used for downstream analysis.

4.10 Protein databases

Protein sequences for human (UP000005640, 20597 proteins) and yeast (UP000002311, 6060 proteins) were downloaded from UniProt (02/2024). The metaproteome database was downloaded from [34].

4.11 DIA-NN analysis

DIA-NN (version 1.8.1) [36] analysis was performed using the following parameters: fixed modification, carbamidomethyl (C); no variable modification was set except phosphorylation was set for the phosphoproteomics data; enzyme, Trypsin/P with one missed cleavage site allowed; peptide length range, 7–35; precursor charge range, 2–4. The setting of "N-term M excision" was disabled; library generation, "IDs, RT & IM profiling". For the phosphoproteomics data, the maximum number of variable modifications was set to 1 and both parameters "Mass accuracy" and "MS1 accuracy" were set to 20. The precursor FDR was set to 1%. All other parameters were set to their default values. For single run DIA data, the "Q.Value" from the main report was used as precursor q-value for downstream analysis. For datasets with multiple runs, the "Lib.Q.Value" from the main report was used as precursor q-value for downstream analysis.

4.12 EncyclopeDIA analysis

For EncyclopeDIA (version 2.12.30) [20, 37] analysis, we first generated two *in silico* spectral libraries on a yeast protein database (6060 protein sequences) for each DIA dataset . Specifically, we used Oktoberfest (version 0.6.2) with Prosit models (fragment ion intensity prediction model: Prosit_2020_intensity_HCD, retention time model: Prosit_2019_irt) [12, 38] to generate an *in silico* spectral libraries for each DIA dataset. In the spectral library generation step, carbamidomethyl of cysteine was considered as a fixed modification, and no variable modifications were considered. Precursor charges 2 to 4 were considered. The NCE parameter was set to 27. Trypsin (without proline suppression) with one missed cleavage allowed was used and only peptides with lengths between 7 and 35 amino acids were considered in Oktoberfest. We also used Carafe to generate an *in silico* spectral library for each dataset. The Carafe models were fined-tuned using the human sample-derived DIA data from each dataset. In the spectral library generation step, we considered carbamidomethyl of cysteine as a fixed modification, and no variable modifications were considered. Precursor charges 2 to 4 were considered. Trypsin (without proline suppression) with one missed cleavage allowed was used, and only peptides with lengths between 7 and 35 amino acids were considered in Carafe. The *in silico* spectral libraries generated by Oktoberfest and Carafe were then converted to "dlib" format libraries using EncyclopeDIA. The "dlib" format libraries were later used for EncyclopeDIA analysis.

For each DIA dataset, we next generated a DIA-based chromatogram library by searching a set of GPF (narrow window) DIA runs against each *in silico* spectral library using EncyclopeDIA.

Finally, we searched the wide window DIA runs from each DIA dataset against each of the four libraries using EncyclopeDIA. In the analysis, the V2 scoring of EncyclopeDIA was enabled except for the TripleTOF 5600 dataset. Both precursor and fragment mass tolerance parameters were set to 40 ppm for the TripleTOF 5600 dataset. The EncyclopeDIA analysis was run through the nf-skyline-dia-ms workflow (`https://nf-skyline-dia-ms.readthedocs.io`).

4.13 Data availability

The MS/MS datasets generated in this study have been deposited to Panorama Public (ProteomeXchange identifier: PXD056793) and are available at `https://panoramaweb.org/Carafe.url`. The Q Exactive HF-X dataset and the TripleTOF 5600 dataset were downloaded from PRIDE [39] with the accession number PXD028735 [40].

4.14 Code availability

Carafe is available at `https://github.com/Noble-Lab/Carafe`. A Nextflow workflow for Carafe is available at `https://nf-carafe-ai-ms.readthedocs.io/`. The source code of the customized AlphaPeptDeep used in this study is available at `https://github.com/wenbostar/alphapeptdeep_dia`.

Acknowledgments This work was supported by National Science Foundation award 2245300, National Institutes of Health awards R24 GM141156 and U19 AG065156, and by the Intelligence Advanced Research Projects Activity (IARPA) TEI-REX program through the Army Research Office contract W911NF2220059.

Conflict of interest The MacCoss Lab at the University of Washington receives funding from Agilent, Bruker, Sciex, Shimadzu, Thermo Fisher Scientific, and Waters to support the development of Skyline, a quantitative analysis software tool. MJM is a paid consultant for Thermo Fisher Scientific.

References

[1] J. D. Venable, M. Q. Dong, J. Wohlsclegel, A. Dillin, and J. R. Yates, III. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature Methods*, 1:39–45, 2004.

[2] L. C. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Molecular and Cellular Proteomics*, 11(6):O111.016717, 2012.

[3] Lilian R. Heil, Eugen Damoc, Tabiwang N. Arrey, Anna Pashkova, Eduard Denisov, Johannes Petzoldt, Amelia C. Peterson, Chris Hsu, Brian C. Searle, Nicholas Shulman, Michael Riffle, Brian Connolly, Brendan X. MacLean, Philip M. Remes, Michael W. Senko, Hamish I. Stewart, Christian Hock, Alexander A. Makarov, Daniel Hermanson, Vlad Zabrouskov, Christine C. Wu, and Michael J. MacCoss. Evaluating the performance of the astral mass analyzer for quantitative proteomics using data-independent acquisition. *Journal of Proteome Research*, 22(10):3290–3300, 2023.

[4] Florian Meier, Andreas-David Brunner, Max Frank, Annie Ha, Isabell Bludau, Eugenia Voytik, Stephanie Kaspar-Schoenefeld, Markus Lubeck, Oliver Raether, Nicolai Bache, et al. diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nature methods*, 17(12):1229–1236, 2020.

[5] Ying S Ting, Jarrett D Egertson, Samuel H Payne, Sangtae Kim, Brendan MacLean, Lukas Käll, Ruedi Aebersold, Richard D Smith, William Stafford Noble, and Michael J MacCoss. Peptide-centric proteome analysis: an alternative strategy for the analysis of tandem mass spectrometry data. *Molecular & Cellular Proteomics*, 14(9):2301–2307, 2015.

[6] F. Yu, G. C. Teo, A. T. Kong, K. Fröhlich, G. X. Li, V. Demichev, and A. I. Nesvizhskii. Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. *Nature Communications*, 14(1):4154, 2023.

[7] Yi Yang, Xiaohui Liu, Chengpin Shen, Yu Lin, Pengyuan Yang, and Liang Qiao. In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nature communications*, 11(1):146, 2020.

[8] G Rosenberger, CC Koh, T Guo, HL Röst, and Kouvonen P data. A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Scientific data*, 2014.

[9] Tiansheng Zhu, Yi Zhu, Yue Xuan, Huanhuan Gao, Xue Cai, Sander R Piersma, Thang V Pham, Tim Schelfhorst, Richard RGD Haas, Irene V Bijnsdorp, et al. DPHL: a DIA pan-human protein mass spectrometry library for robust biomarker discovery. *Genomics, Proteomics and Bioinformatics*, 18(2):104–119, 2020.

[10] R. Bruderer, O. M. Bernhardt, T. Gandhi, Y. Xuan, J. Sondermann, M. Schmidt, D. Gomez-Varela, and L. Reiter. Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Molecular & Cellular Proteomics*, 16(12):2296–2309, 2017.

[11] X. Zhou, W. Zeng, H. Chi, C. Luo, C. Liu, J. Zhan, S. M. He, and Z. Zhang. pDeep: predicting MS/MS spectra of peptides with deep learning. *Analytical Chemistry*, 89(23):12690–12697, 2017.

[12] S. Gessulat, T. Schmidt, D. P. Zolg, P. Samaras, K. Schnatbaum, J. Zerweck, T. Knaute, J. Rechenberger, B. Delanghed, A. Huhmer, U. Reimer, H. Ehrlich, S. Aiche, B. Kuster, and M. Wilhelm. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods*, 16(6):509, 2019.

[13] S. Tiwary, R. Levy, P. Gutenbrunner, F. S. Soto, K. K. Palaniappan, L. Deming, M. Berndl, A. Brant, P. Cimermancic, and J. Cox. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature Methods*, 16(6):519–525, 2019.

[14] K. Liu, S. Li, L. Wang, Y. Ye, and H. Tang. Full-spectrum prediction of peptides tandem mass spectra using deep neural network. *Analytical Chemistry*, 92(6):4275–4283, 2020.

[15] Bo Wen, Kai Li, Yun Zhang, and Bing Zhang. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nature Communications*, 11(1):1759, 2020.

[16] Wen-Feng Zeng, Xie-Xuan Zhou, Sander Willems, Constantin Ammar, Maria Wahle, Isabell Bludau, Eugenia Voytik, Maximillian T. Strauss, and Matthias Mann. Alphapeptdeep: a modular deep learning framework to predict peptide properties for proteomics. *Nature Communications*, 13(1):7238, 11 2022.

[17] Bo Wen, Wen-Feng Zeng, Yuxing Liao, Zhiao Shi, Sara R Savage, Wen Jiang, and Bing Zhang. Deep learning in proteomics. *Proteomics*, 20(21-22):1900335, 2020.

[18] P. Sinitcyn, H. Hamzeiy, F. Salinas Soto, D. Itzhak, F. McCarthy, C. Wichmann, M. Steger, U. Ohmayer, U. Distler, S. Kaspar-Schoenefeld, et al. MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nature Biotechnology*, 39(12):1563–1573, 2021.

[19] Ronghui Lou, Weizhen Liu, Rongjie Li, Shanshan Li, Xuming He, and Wenqing Shui. Deepphospho accelerates dia phosphoproteome profiling through in silico library generation. *Nature Communications*, 12(1):6685, 2021.

[20] B. C. Searle, L. K. Pino, J. D. Egertson, Y. S. Tin, R. T. Lawrence, B. X. MacLean, J. Vill'en, and M. J. MacCoss. Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nature Communications*, 9:5128, 2018.

[21] Lindsay K Pino, Seth C Just, Michael J MacCoss, and Brian C Searle. Acquiring and analyzing data independent acquisition proteomics experiments without spectrum libraries. *Molecular & Cellular Proteomics*, 19(7):1088–1103, 2020.

[22] Marc Isaksson, Christofer Karlsson, Thomas Laurell, Agnete Kirkeby, and Moritz Heusel. MSLibrarian: optimized predicted spectral libraries for data-independent acquisition proteomics. *Journal of Proteome Research*, 21(2):535–546, 2022.

[23] Chih-Chiang Tsou, Dmitry Avtonomov, Brett Larsen, Monika Tucholska, Hyungwon Choi, Anne-Claude Gingras, and Alexey I Nesvizhskii. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature methods*, 12(3):258–264, 2015.

[24] Vadim Demichev, Christoph B Messner, Spyros I Vernardis, Kathryn S Lilley, and Markus Ralser. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature methods*, 17(1):41–44, 2020.

[25] B. MacLean, D. M. Tomazela, N. Shulman, M. Chambers, G. L. Finney, B. Frewen, R. Kern, D. L. Tabb, D. C. Liebler, and M. J. MacCoss. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*, 26(7):966–968, 2010.

[26] J. Klein, H. Lam, T. Mak, W. Bittremieux, Y. Perez-Riverol, R. Gabriels, J. Shofstahl, H. Hecht, P.-A. Binz, S. Kawano, et al. The Proteomics Standards Initiative standardized formats for spectral libraries and fragment ion peak annotations: mzSpecLib and mzPAF. *chemRxiv*, 2024. doi.org/10.26434/chemrxiv-2024-p71gj.

[27] Damien B Wilburn, Alicia L Richards, Danielle L Swaney, and Brian C Searle. Cider: a statistical framework for interpreting differences in cid and hcd fragmentation. *Journal of proteome research*, 20(4):1951–1965, 2021.

[28] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[29] Tanveer S Batth, MaximA X Tollenaere, Patrick Rüther, Alba Gonzalez-Franquesa, Bhargav S Prabhakar, Simon Bekker-Jensen, Atul S Deshmukh, and Jesper V Olsen. Protein aggregation capture on microparticles enables multipurpose proteomics sample preparation. *Molecular and Cellular Proteomics*, 18(5):1027a, 2019.

[30] Christopher S Hughes, Sophie Moggridge, Torsten Müller, Poul H Sorensen, Gregg B Morin, and Jeroen Krijgsveld. Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nature protocols*, 14(1):68–85, 2019.

[31] PK etal Smith, R Il Krohn, GT Hermanson, AK Mallia, FH Gartner, MDea Provenzano, EK Fujimoto, NM Goeke, BJ Olson, and DC Klenk. Measurement of protein using bicinchoninic acid. *Analytical biochemistry*, 150(1):76–85, 1985.

[32] Alexis Chang, Mario Leutert, Ricard A Rodriguez-Mias, and Judit Villén. Automated enrichment of phosphotyrosine peptides for high-throughput proteomics. *Journal of proteome research*, 22(6):1868–1880, 2023.

[33] Mario Leutert, Ricard A Rodríguez-Mias, Noelle K Fukuda, and Judit Villén. R2-p2 rapid-robotic phosphoproteomics enables multidimensional cell signaling studies. *Molecular systems biology*, 15(12):e9021, 2019.

[34] Brook L Nunn, Emma Timmins-Schiffman, Miranda C Mudge, Deanna Plubell, Gabriella Chebli, Julia Kubanek, Michael Riffle, William S Noble, Elizabeth Harvey, Tasman A Nunn, et al. Microbial metagenomes across a complete phytoplankton bloom cycle: High-resolution sampling every 4 hours over 22 days. *bioRxiv*, pages 2024–09, 2024.

[35] M. C. Chambers, B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T. A. Baker, M. Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E. W. Deutsch, R. L. Moritz, J. E. Katz, D. B. Agus, M. J. MacCoss, D. L. Tabb, and P. Mallick. A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology*, 30(10):918–920, 2012.

[36] V. Demichev, C. B. Messner, S. I. Vernardis, K. S. Lilley, and M. Ralser. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods*, 17(1):41–44, 2020.

[37] B. C. Searl, K. E. Swearingen, C. A. Barnes, T. Schmidt, S. Gessulat, B. Küster, and M. Wilhelm. Generating high quality libraries for dia ms with empirically corrected peptide predictions. *Nature Communications*, 11(1):1–10, 2020.

[38] M. Picciani, W. Gabriel, V.-G. Giurcoiu, O. Shouman, F. Hamood, L. Lautenbacher, C. B. Jensen, J. Müller, M. Kalhor, A. Soleymaniniya, et al. Oktoberfest: Open-source spectral library generation and rescoring pipeline based on Prosit. *Proteomics*, page 2300112, 2023.

[39] Y. Perez-Riverol, A. Csordas, J. Bai, M. Bernal-Llinares, S. Hewapathirana, D. J. Kundu, A. Inuganti, J. Griss, G. Mayer, M. Eisenacher, E. Pérez, J. Uszkoreit, J. Pfeuffer, T. Sachsenberg, S. Yilmaz, S. Tiwary, J. Cox, E. Audain, M. Walzer, A. F. Jarnuczak, T. Ternent, A. Brazma, and J. A. Vizcaíno. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Research*, 47(D1):D442–D450, 1 2019.

[40] Bart Van Puyvelde, Simon Daled, Sander Willems, Ralf Gabriels, Anne Gonzalez de Peredo, Karima Chaoui, Emmanuelle Mouton-Barbosa, David Bouyssié, Kurt Boonen, Christopher J. Hughes, Lee A. Gethings, Yasset Perez-Riverol, Nic Bloomfield, Stephen Tate, Odile Schiltz, Lennart Martens, Dieter Deforce, and Maarten Dhaenens. A comprehensive lfq benchmark dataset on modern day acquisition strategies in proteomics. *Scientific Data*, 9(1), March 2022.