DIAMeter: Matching peptides to data-independent acquisition mass spectrometry data

Yang Young Lu¹, Jeff Bilmes², Ricard A Rodriguez-Mias¹, Judit Villén¹, and William Stafford Noble^{*1,3}

¹Department of Genome Sciences, University of Washington, Seattle, WA, USA

²Department of Electrical Engineering, University of Washington, Seattle, WA, USA

³Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA

Abstract

Tandem mass spectrometry data acquired using data independent acquisition (DIA) is challenging to interpret because the data exhibits complex structure along both the mass-to-charge (m/z) and time axes. The most common approach to analyzing this type of data makes use of a library of previously observed DIA data patterns (a “spectral library”), but this approach is expensive because the libraries do not typically generalize well across laboratories. Here we propose DIAMeter, a search engine that detects peptides in DIA data using only a peptide sequence database. Unlike other library-free DIA analysis methods, DIAMeter supports data generated using both wide and narrow isolation windows, can readily detect peptides containing post-translational modifications, can analyze data from a variety of instrument platforms, and is capable of detecting peptides even in the absence of detectable signal in the survey (MS1) scan.

1 Introduction

Data-independent acquisition (DIA) mass spectrometry (MS) is a powerful technique to study the proteome.^{1,2} Relative to data-dependent acquisition (DDA), DIA offers a broader dynamic range and more reproducible peptide detection. However, to achieve accurate quantification, DIA methods typically employ wide isolation windows to acquire a full mass range of fragmentation spectrum (MS2) data. As a result, multiple peptides are usually co-isolated and co-fragmented in the same MS2 scan. The resulting complex spectra cannot be effectively analyzed using conventional database search engines^{3–7} that follow a “one-peptide-per-spectrum” paradigm.

A common class of methods for analyzing DIA data relies on spectral libraries that store, for each charged peptide, fragmentation and retention time information.^{8–13} However, spectral libraries are expensive to produce, requiring considerable effort and resources devoted to sample preparation and data acquisition. Furthermore, the resulting libraries are typically not reusable across different laboratories or different instrument platforms.¹⁴ Some of these challenges can be addressed by generating spectral libraries *in silico*;^{15–18} however, the resulting predictions depend on the instrument and specific acquisition parameters and are typically less accurate than experiment-specific libraries.¹⁹ Furthermore, current state-of-the-art *in silico* prediction methods provide limited support for post-translational modifications (PTMs).¹⁶

An alternative class of library-free methods circumvent some of the challenges associated with generating and using spectral libraries.^{12, 20–22} Library-free methods detect peptides from DIA data by using a peptide sequence database, rather than a library. The library-free approach is particularly valuable when production

*Corresponding author: Box 355065, Foege Building, S220B, 3720 15th Ave NE, Seattle, WA 98195-5065, william-noble@uw.edu

Figure 1: **Overview of DIAMeter.** The algorithm takes as input DIA data from a single run and a peptide sequence database. The output is a list of peptides, ranked by q-value, which is defined as the minimal FDR threshold at which that peptide is accepted. The Tide search step reports the top k peptide-spectrum matches (PSMs) per spectrum for charge states +1 through +5. In this work, $k = 5$. The edge features computed during composite scoring are described in Section 2.2.

of a spectral library is impractical. These methods can be subdivided into *spectrum-centric* and *peptide-centric* approaches. Spectrum-centric methods, such as DIA-Umpire,²¹ extract coeluting and co-varying precursor and fragment ion groups into pseudospectra. These pseudospectra are then fed into a conventional database search engine designed for DDA-based peptide identification. The extraction of pseudospectra depends heavily on the quality of the precursor signals in the precursor (MS1) scans; hence, DIA-Umpire by design cannot detect peptides with undetectable precursor signals, which commonly arise due to limitations of intrascan dynamic range.²² Unlike spectrum-centric methods, peptide-centric methods such as PECAN²² query the DIA data for the best supporting evidence of detection for each peptide in the database. However, empirical evidence suggests that existing library-free methods are not as sensitive as library-based ones.^{13,23}

Here we introduce DIAMeter, which detects peptides directly from DIA data without dependence on a spectral library (Figure 1). The input to DIAMeter includes m/z-centroided DIA data and a proteome FASTA database. DIAMeter then searches the DIA data using a standard DDA search engine, allowing multiple peptide-spectrum matches (PSMs) per DIA spectrum. The PSMs are then augmented with auxiliary features describing various types of evidence supporting the detection of the associated peptide, including two different peptide-spectrum match scores, analysis of the precursor (MS1) signal, comparison of observed and predicted chromatographic retention time, and coelution of precursor and peptide fragment ions. PSMs are filtered using a weighted combination of these scores, and the PSM feature vectors are then processed by the Percolator machine learning post-processor²⁴ to induce a ranking on peptides along with statistical confidence estimates, where highly ranked peptides are detected in the DIA data with stronger confidence.

Compared to existing DIA analysis methods, DIAMeter offers four primary advantages (Table 1). First, DIAMeter by design can incorporate DIA data acquired using wide precursor isolation windows (*e.g.*, up to 25 Da). In this setting, the MS2 spectra are more complex because multiple peptides are more frequently co-fragmented. Second, DIAMeter can readily detect peptides with PTMs, as long as they are included in the database. Third, DIAMeter makes no assumptions about instrument-specific fragmentation patterns, instrument resolution, or patterns of background noise. Fourth, DIAMeter uses precursor information as a feature but does not require such evidence to detect a peptide. The open source, Apache licensed, Python version of DIAMeter is accessible at <https://bitbucket.org/noblelab/diameter>.

	DIAmeter	Spectrum-centric DIA-Umpire	Peptide-centric PECAN	<i>in silico</i> library Prosit+EncyclopeDIA
Supports wide isolation windows	✓	✓		
Detects PTMs	✓	✓	✓	
Instrument independent	✓	✓		
Detects peptides with undetectable precursor	✓		✓	✓

Table 1: Comparison of library-free peptide detection methods for DIA data.

2 Approach

The DIAMeter workflow (Figure 1) consists of the following steps.

1. Construct a bipartite graph between the observed DIA data and the list of theoretical precursors (i.e., charged peptides) in the database using a DDA search engine.
2. Compute a set of features for each edge in the graph, and eliminate some edges based upon a composite score aggregated from these features.
3. Run the remaining edges, with associated features, through a modified version of the Percolator algorithm.
4. Use target-decoy competition at the peptide level to estimate a q-value (i.e., the minimal FDR threshold at which a given peptide is accepted) for each ranked peptide.

These steps are described in more detail in Sections 2.1–2.4.

2.1 Bipartite graph construction

We conceptualize the DIAMeter approach using an undirected bipartite spectrum-to-precursor graph $G = (U, V, E)$ that matches DIA spectra and a precursor database (Figure 1). Specifically, we define U to be the set of all MS2 spectra, each of which corresponds in a one-to-one fashion to a specific acquisition range in a single MS1 scan. For each spectrum $u \in U$, we let t_u and d_u denote its corresponding MS scan and acquisition range, respectively. The precursor database V is an unordered list of charged peptide sequences. For each precursor $v \in V$, we let m_v , c_v and p_v denote its corresponding precursor m/z value, charge state, and peptide sequence, respectively. Conversely, $u \in U$ and $v \in V$ can also be indexed by u_{t_u, d_u} and v_{p_v, c_v} , respectively.

The edges E represent potential peptide-spectrum matches. Thus, for $v \in V$ and $u \in U$, an undirected edge $e = (u, v) \in E$ connects u and v if and only if the theoretical m/z value of the precursor v falls within a specified acquisition range d_u . Because each spectrum $u \in U$ may potentially be matched to a large number of precursors when the DIA isolation window is wide, we only keep the top k PSMs for each spectrum based upon their primary score. In this work, we use the SEQUEST XCorr score²⁵ as the primary score, computed using the Tide search engine.²⁶ We specify the parameter `--top-match 1000` and then keep the top-scoring k PSMs for each charge state $c \in C$ where $C = \{1, 2, \dots, 5\}$. In this way, for each spectrum $u \in U$, at most $|C| \cdot k$ incident edges are added to the bipartite graph.

2.2 Calculation of edge features

For each edge $e = (u, v) \in E$, DIAMeter computes a set of eight features. These include three features generated by the Tide search engine—charge, delta.cn, delta.lcn—plus five edge features, described in detail below, each of which quantifies one type of evidence for or against the match between the spectrum u and the precursor v .

XCorr with Tailor calibration. The XCorr score is calibrated using the non-parametric Tailor calibration method,²⁷ by specifying the parameter `--use-tailor-calibration T`. In brief, for each edge $e = (u, v) \in E$, the Tailor method calibrates the score $\text{XCorr}_{u,v}$ by dividing by the 99th quantile of the observed, spectrum-specific score distribution. Specifically, say that spectrum u is scored with respect to

N candidate peptides, each with charge state c_v , during the database search. The resulting edge scores are sorted in decreasing order $\text{XCorr}_u^{(1)}, \dots, \text{XCorr}_u^{(N)}$, and the 99th quantile of this score distribution is obtained by identifying the edge score $\text{XCorr}_u^{([N/100])}$ at position $[N/100]$, where $[\cdot]$ denotes the round operator. The Tailor calibrated XCorr score is defined as

$$w_{u,v}^T = \frac{\text{XCorr}_{u,v}}{\text{XCorr}_u^{([N/100])}}.$$

Precursor intensity rank The observed precursor intensity can provide important evidence supporting the existence of a particular peptide in the sample, although as noted previously, lack of detectable signal in the precursor scan does not necessarily imply that the peptide has not been detected in the DIA data. For each edge $e = (u, v) \in E$, we denote I_{u,m_v}^P as the maximum observed precursor intensity at the precursor m/z value m_v in the MS scan t_u , within a user-specified tolerance. Using the raw precursor intensity directly is potentially problematic because the precursor intensity may exhibit biases specific to retention time, and because we don't want a peptide with undetectable precursor signal to receive a score of zero. Therefore, we use the log-rank of precursor intensity instead of the raw intensity, represented as $\log |\{i \in I_u^P | i \geq I_{u,m_v}^P\}|$, where I_u^P is the set of all observed peak intensities in the MS scan t_u . In addition to the monoisotope of the precursor, we also consider $M + 1$ and $M + 2$ isotopes. Therefore, the precursor intensity rank score is defined as

$$w_{u,v}^P = -\log |\{i \in I_u^P | i \geq I_{u,m_v}^P\}| - \log |\{i \in I_u^P | i \geq I_{u,m_v+1/c_v}^P\}| - \log |\{i \in I_u^P | i \geq I_{u,m_v+2/c_v}^P\}| \quad (1)$$

Fragment matching p-value. The fragment-matching p-value, analogous to the matching score in MS Amanda,⁷ measures how likely the theoretical fragments are to match the observed MS2 peaks by chance. Specifically, for each edge $e = (u, v) \in E$, by denoting $N_{u,v}$ and N_v as the number of matched fragment peaks and all theoretical peaks, respectively, the fragment-matching p-value score is defined as the negative logarithm of the probability of matching at least $N_{u,v}$ peaks using the cumulative binomial distribution:

$$w_{u,v}^F = -\log \sum_{k=N_{u,v}}^{N_v} \binom{N_v}{k} p^k (1-p)^{N_v-k}$$

where p_u is the probability to match one peak in u by chance. This value can be calculated as the fraction of the m/z range that is covered by the observed peaks. Analogous to the peak-picking in MS Amanda,⁷ the m/z range is divided into 10 equal length segments, and in each segment, the 10 most intense peaks are preserved. Let $z_u^{(1)}, \dots, z_u^{(N_u)}$ be the sorted m/z values, where $N_u = 100$ is the total number of peaks remaining in the spectrum u . The fragment matching p-value p_u is calculated as

$$p_u = \frac{\left| \bigcup_{i=1}^{N_u} [z_u^{(i)} - \tau, z_u^{(i)} + \tau] \right|}{(z_u^{(N_u)} + \tau) - (z_u^{(1)} - \tau)},$$

where τ specifies the matching tolerance.

Difference between observed and predicted retention time. An important factor to validate a candidate peptide is the difference between the observed and predicted retention times (RT). We use the scan number as a proxy for the observed RT, defining the set of observed RTs for all spectra as $\text{RT}_U^{\text{obv}} = \{t_u | u \in U\}$. In parallel, we use the machine learning method DeepRT(+)² to define the set of predicted RTs for all precursors as $\text{RT}_V^{\text{pred}} = \{t_v | v \in V\}$, where t_v indicates the predicted RT for the precursor $v \in V$. For each edge $e = (u, v) \in E$, the normalized observed and predicted RT can be represented as

$$\text{NormRT}_u = \frac{\left| \{t' \in \text{RT}_U^{\text{obv}} | t' \leq t_u\} \right|}{\left| \text{RT}_U^{\text{obv}} \right|},$$

and similarly for NormRT_v . Thus, the RT difference score is defined as

$$w_{u,v}^R = -|\text{NormRT}_u - \text{NormRT}_v|.$$

Note that DeepRT can only make RT predictions for unmodified peptides and peptides with methionine oxidation or phosphorylation. For peptides with other types of modifications, DIAMeter uses the default value of $w_{u,v}^R = -0.5$.

Precursor and fragment coelution. Coelution of precursor ions and corresponding fragment ions is an important piece of evidence supporting detection of a specific peptide.²¹ In DIAMeter, this score is calculated as the normalized dot product between the corresponding elution profiles. Specifically, for each edge $e = (u, v) \in E$, the precursor elution profile spans nearby spectra within Δt scan cycles ($\Delta t = 2$ in our study), calculated as $\text{Elu}_{m_v}^P = [I_{u_{t_u-\Delta t, d_u}, m_v}^P, \dots, I_{u_{t_u+\Delta t, d_u}, m_v}^P]$, where I_{u, m_v}^P is the maximum observed precursor intensity at the precursor m/z value m_v in the MS1 scan t_u within some specified tolerance (see Section 3.4 for details). Similarly, by denoting $\mathcal{F}_v$ as the set of all fragment m/z values of the precursor v , for each fragment m/z $f \in \mathcal{F}_v$, the fragment elution profile is calculated as $\text{Elu}_f^F = [I_{u_{t_u-\Delta t, d_u}, f}^F, \dots, I_{u_{t_u+\Delta t, d_u}, f}^F]$, where $I_{u, f}^F$ is denoted as the maximum observed fragment intensity at the fragment m/z value f in the MS2 scan u , also within some specified tolerance. We then define the set of normalized dot products between the precursor and its fragments as

$$\text{Elu}_{u,v} = \left\{ \left\langle \frac{\text{Elu}_{m_v}^P}{\|\text{Elu}_{m_v}^P\|_2}, \frac{\text{Elu}_f^F}{\|\text{Elu}_f^F\|_2} \right\rangle \mid f \in \mathcal{F}_v \right\},$$

where $\langle \cdot, \cdot \rangle$ indicates the inner product operator. The final precursor/fragment coelution score $w_{u,v}^E$ is the average of the top 3 values in $\text{Elu}_{u,v}$.

All of DIAMeter’s hyperparameters are summarized in Supplementary Table S2.

2.3 Edge re-scoring and filtering

Using the four scores described above, DIAMeter computes an aggregate score function and uses it to eliminate some edges in the bipartite graph. First, each score is linearly rescaled so that the 1st quantile and the 99th quantile map to the range $[0, 1]$. Thereafter, the aggregated score is $w_{u,v} = w_{u,v}^T + \alpha^P \cdot w_{u,v}^P + \alpha^F \cdot w_{u,v}^F + \alpha^R \cdot w_{u,v}^R + \alpha^E \cdot w_{u,v}^E$, where the four non-negative hyperparameters $\alpha^P, \alpha^F, \alpha^R$, and α^E balance the contributions from the corresponding score components. Subsequently, for each spectrum $u \in U$ and charge state $c \in \{1, \dots, 5\}$, an edge is eliminated if its aggregated score is less than the aggregated score of the top-scoring edge according to XCorr. Formally, we denote $v_{u,c}^{(1)}$ as the precursor corresponding to the maximum XCorr score in the set $\{v \in V \mid (u, v) \in E, c_v = c\}$. The filtering step then discards an edge (u, v) if $w_{u,v} < w_{u, v_{u,c}^{(1)}}$.

2.4 Percolator tailored for DIA

Finally, all of the remaining edges are provided as input to the Percolator machine learning algorithm. Percolator’s semi-supervised algorithm learns to re-rank these edges by iteratively training a series of support vector machine classifiers, where negative examples are decoy edges and positive examples are high-scoring target edges.

Rather than using the FDR estimates provided by Percolator, DIAMeter calculates its own peptide-level FDR estimates. This choice is motivated by Percolator’s use of a spectrum-level target-decoy competition, in which only the highest-scoring PSM per spectrum is retained. In general, spectrum-level results are not useful for DIA analysis; therefore, we prefer to carry out the competition at the level of precursors, rather than PSMs. Hence, DIAMeter calls Percolator using the option `--tdc F`, which disables target-decoy competition. DIAMeter then selects the top-scoring edge per target-decoy pair; i.e., among all edges associated with a given precursor and that precursor’s corresponding decoy, only a single score is retained. The FDR estimate for a

given score threshold τ is then calculated as usual:

$$\text{FDR} = \min \left(1, \frac{|\{p : s(p) > \tau \wedge p \text{ is a decoy}\}|}{\max(1, |\{p : s(p) > \tau \wedge p \text{ is a target}\}|)} \right),$$

where $s(p)$ is the Percolator score associated with precursor p .

One potential challenge associated with this approach to precursor-level FDR estimation is created by Percolator’s cross-validation procedure. This procedure equally divides the PSMs into three splits, from which the PSMs in two splits are used for training and the PSMs in the third split for prediction.²⁸ This approach leads to a potential for “leakage” of information, in the case where, say, a decoy appears in the training set but its corresponding target is in the test set. To avoid this problem, we modified Percolator so that all PSMs related to the same target/decoy pair are assigned to the same cross-validation split.

3 Methods

3.1 Datasets

We used three different datasets to evaluate the performance of DIAMeter, in comparison to the three representative state-of-the-art methods mentioned in Table 1. The datasets were chosen to include different instrument platforms, wide and narrow isolation window sizes, fixed-window and variable-window acquisition schemes, and different dynamic ranges (Table S1). Full details about the three datasets are provided in Supplementary Section 3.2. For all three datasets, the database search was performed with the following parameters: fully tryptic digest with up to two missed cleavages, static modifications of carbamidomethyl for cysteines, and up to three variable modifications of oxidation of methionine.

3.2 Data

LFQBench This is a widely used DIA benchmark for peptide identification and quantification,²⁹ which consists of lysate mixtures from human, yeast, and *E. coli* mixed in predefined proportions. We carried out the evaluation by using LFQBench samples acquired on two mass spectrometers (TripleTOF 5600 and TripleTOF 6600) with two acquisition schemes (32 fixed windows of 25 m/z effective precursor isolation and 64 variable windows optimized for roughly equal precursor partition) over a 120-minute gradient. The data as .wiff files, as well as the FASTA database combining human, yeast, and *E. coli* proteins, were downloaded from the <ftp://ftp.pride.ebi.ac.uk/pride/data/archive/2016/09/PXD002952>.

Plasma This dataset is derived from a large cohort designed to assess the heritability and environmental effects on blood plasma protein abundance.³⁰ The data was acquired on a TripleTOF 5600 mass spectrometer with 32 fixed windows of 25 m/z effective precursor isolation over a 120-minute gradient. The data as .wiff files, as well as the human protein FASTA database, were downloaded from <ftp://ftp.pride.ebi.ac.uk/pride/data/archive/2015/01/PXD001064>.

OxMetYeast This dataset consists of three mass spectrometry experiments where different isolation window sizes were explored in combination with gas phase fractionation, using a yeast lysate with selective oxidation of methionine. The resulting DIA data, as .raw files, include (1) one run covering the 400–1200 m/z mass range acquired using 20 m/z isolation windows; (2) two runs acquired using 10 m/z isolation windows to cover 400–800 and 800–1200 m/z mass ranges; (3) four runs acquired using 5 m/z isolation windows covering 400–600, 600–800, 800–1000, and 1000–1200 m/z ranges, respectively. The dataset has been deposited to the PRIDE repository³¹ under data set identifier XXX.¹

To prepare the yeast sample, 25 μ g of tryptic yeast peptides, prepared as described by Studer *et al.*,³² were resuspended in 30 μ l of oxidizing solution containing 100 mM DMSO, 5% acetonitrile and 1 M HCl, and reacted at room temperature for 20 minutes. Peptides were then diluted two-fold with 0.1% TFA and desalted over a 2-layer C18 Stage tip (Empore, 3M),³³ that had been previously conditioned with 20 μ l methanol, followed by 20 μ l 50% acetonitrile 0.5% acetic acid, and 20 μ l 0.1% TFA. Peptides on the Stage

¹Submitted 23 Jan 2021, awaiting PRIDE ID.

tip were washed 3 times with 20 μ l of 0.1% TFA and eluted with 50% acetonitrile 0.5% acetic acid. Eluted peptides were then lyophilized.

Lyophilized peptides were resuspended in 50 μ l of 5% ACN, 5% formic acid and subjected to liquid chromatography coupled to tandem MS. A total of 1.5 μ l of sample was loaded onto a 100 μ m $\times$ 35 cm fused silica capillary tubing packed with 1.9 μ m ReproSil-Pur C18 (Dr. Maisch) reversed-phase material and separated using a gradient of 5% to 36% acetonitrile in 0.125% formic acid delivered at 300 nl/min over 73 minutes on an EASY-nLC 1200 nano-flow UHPLC system (Thermo Scientific). Peptides were online analyzed on an Orbitrap Eclipse Tribrid MS (Thermo Scientific) with a total 90-minute acquisition time. All methods consisted of an MS1 scan at 60K resolution every 3 seconds with 100% normalized AGC target and a maximum injection time of 50 ms followed by DIA MS2 scans events acquired at 30K resolution with a maximum injection time of 54 ms and 4e5 target AGC.

File conversion The .raw files generated by Orbitrap Eclipse Tribrid were converted to .mzXML (used in DIAMeter and DIA-Umpire) and .mzML (used in EncyclopeDIA) files by using the msconvert provided with ProteoWizard.³⁴ The .wiff files generated by TripleTOF instruments were converted to .mzXML and .mzML files by using the qtofpeakpicker provided with ProteoWizard,³⁴ with the default options: `--resolution 20000 --area 1 --threshold 10 --numberofpeaks 0`.

Noise reduction. DIAMeter applies a noise reduction strategy³⁵ to TripleTOF 5600 data by following the assumption that peaks occurring in a single scan, without any corresponding peaks in preceding or succeeding scans, are noise. Specifically, the MS2 peak is eliminated if neither of the adjacent scans contains the same peak within a specified tolerance. For the Orbitrap and TripleTOF 6600 instruments, noise reduction is not applied due to their high signal-to-noise ratio.

3.3 Consistent FDR estimation across methods

One significant challenge when comparing different peptide detection methods is that each method chooses to estimate FDR in its own fashion. Therefore, an improvement in statistical power for method *A* over method *B* may arise either because *A* outperforms *B* or because *B* makes more conservative FDR estimates than *A*.

To address this challenge, we use a consistent “pseudo-target” scheme to evaluate all methods. The approach, similar to previously described “entrapment” strategies,³⁶ is as follows. First, given a target peptide database, a decoy peptide database is generated of the same size. Both sets of sequences are provided to the peptide detection method, but both are labeled as targets. The method then uses its own internal procedure to generate a decoy database, search the concatenated target-decoy database, and induces a ranking on the peptide as output. We discard the reported FDR and instead use the pseudo-targets to estimate FDR, as in Section 2.4. The inclusion of pseudo-targets leads to a slight loss of statistical power but is only done for evaluation purposes.

3.4 Benchmark settings

DIAMeter The first step of DIAMeter involves a database search that considers, for each spectrum, all possible precursors whose corresponding m/z value falls within the DIA isolation window. To do so, the precursor mass for each DIA spectrum is calculated based on the center m/z value of its corresponding isolation window, and searched using the Tide search engine²⁶ with the parameter `--precursor-window-type mz --precursor-window <isolation window radius>`. Afterwards, the bipartite spectrum-to-precursor graph is constructed based on the top $k = 5$ PSMs for each spectrum and for each charge state. In the subsequent feature calculation step, a precursor and fragment mass tolerance of 10 ppm is used for all Orbitrap datasets, and 30 ppm for all TripleTOF datasets.

The DIAMeter hyperparameters are selected via cross-validation over the 8 experiments involved in the evaluation. For each experiment regarded as the test set, we treated the remaining 7 experiments as the training set and select the best-performing hyperparameters from a grid in which each of the α values ($\alpha^P, \alpha^F, \alpha^R, \alpha^E$) can take any value in the set $\{0, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 12.8\}$. The performance of the chosen hyperparameters is evaluated by the number of accepted peptides at a 1% peptide-level FDR.

Figure 2: **Performance comparison across various methods.** The performance is evaluated by the number of detected peptide, at a 1% peptide-level FDR, in eight experiments from three different datasets.

DIA-Umpire For DIA-Umpire, the preprocessed .mzXML files were analyzed by the signal extraction module of DIA-Umpire (version 2.0) to generate pseudo spectra in MGF format, at all quality tiers, which were subsequently searched using the Tide search engine.²⁶ We kept the search settings as close as possible to the DIAMeter settings to facilitate a fair comparison. The complete list of DIA-Umpire parameter settings is provided in Supplementary Table S3.

PECAN We compared against Walnut, a re-implementation of PECAN provided by EncyclopeDIA (version 0.9.0),¹³ with similar setting as DIAMeter settings. Full parameters are in Section 3.2.

Prosit+EncyclopeDIA For EncyclopeDIA,¹³ we used an *in silico* library generated by Prosit.¹⁷ The fragmentation prediction in Prosit is adjusted based on the optimal normalized collision energy setting to account for DIA-specific fragmentation, as suggested previously.¹⁹ For the Prosit-generated spectral library, by design peptides with length in the range of [7, 30] and charge state $C = \{2, 3\}$ are considered.

4 Results

4.1 DIAMeter confidently detects more peptides

We systematically evaluated the performance of DIAMeter relative to DIA-Umpire, PECAN and Prosit+EncyclopeDIA by counting the number of distinct peptides detected in various datasets at a 1% FDR threshold (Figure 2). For this analysis, FDR is estimated using the pseudo-targets described in Section 3.3.

We first focused on the LFQbench dataset, which allows us to consider how different instrument platforms impact the sensitivity of peptide detection. This dataset is comprised of four mass spectrometry experiments acquired on two different types of mass spectrometer with two different acquisition schemes. At every setting, DIAMeter consistently detects more peptides than the three competing methods, with an average increase of 14.6% relative to DIA-Umpire, 77.7% relative to PECAN, and 48.5% relative to Prosit+EncyclopeDIA. The relatively large performance improvement relative to PECAN and Prosit+EncyclopeDIA may arise because these methods are tailored for Orbitrap instruments. For example, the Prosit prediction model was trained exclusively on Orbitrap data.¹⁷

Not surprisingly, the TripleTOF 6600 provides improved peptide detection over its predecessor, the 5600, consistently across the four analysis methods. For example, for DIAMeterswitching from the TripleTOF 5600 to the 6600 yields an average increase of 46.2% in the number of detected peptides. The only notable

Figure 3: **Performance comparison as a function of FDR threshold.** Performance of each method is evaluated by the number of detected peptides as a function of peptide-level q-values in eight experiments from three different datasets.

exception to this trend is PECAN, where no peptide is confidently detected within 1% FDR threshold on the TripleTOF 6600. We hypothesized that this failure might be due to the large size of the peptide database. This hypothesis is supported by the observation that, when we randomly discarded 70% of the targets and their corresponding decoys from the database, the number of detections increased from 0 to 844.

Next, we tested how the performance of each of the four methods changes as we vary the size of the precursor isolation window. In general, employing a wide isolation window drastically increases the complexity of the resulting spectra and in turn decreases detection sensitivity.²² To test how isolation window size impacts detection performance, we used the OxMetYeast dataset, which consists of three mass spectrometry experiments acquired using isolation windows of 20 m/z, 10 m/z, and 5 m/z. As expected, all four methods detect more peptides as the window size decreases. For example, DIAMeter detects 8903, 14017, and 19622 peptides, respectively. On average across the four methods, the number of detected peptides increases by 23.5% when we switch from 20 m/z to 10 m/z and by 32.6% when we switch from 10 m/z to 5 m/z. Among the four methods, DIAMeter performs best overall for narrower windows, and second best (behind Prosit+EncyclopeDIA) for 20 m/z windows.

Finally, to investigate the robustness of the four analysis methods to samples showing a large dynamic range of peptide abundance, we used a human plasma dataset acquired on a TripleTOF 5600. DIAMeter detected 1811 unique peptides in this dataset, outperforming DIA-Umpire, PECAN, and Prosit+EncyclopeDIA, which reported 1551, 941, and 1038 unique peptides, respectively. Thus, overall, DIAMeter shows robust peptide detection performance across a variety of datasets, in comparison with three methodologically distinct methods for DIA analysis. On the other hand, investigation of detection rates at other FDR thresholds (Figure 3) suggests that, in some cases, the relative performance among methods depends on the selected FDR threshold. In particular, in most tripleTOF experiments, DIAMeter consistently performs well across a range of q-value thresholds, whereas in Orbitrap experiments, DIAMeter performs best for q-values ≤ 0.01 .

4.2 DIAMeter can detect peptides with undetectable MS1 signal

A central challenge in the analysis of any protein tandem mass spectrometry data set is the huge dynamic range of protein abundances in many complex biological samples. In DIA analysis, this variability in protein abundance means that some peptides without detectable precursor signals may still be identifiable on the basis of their fragmentation scans.²² We hypothesized that these low abundance peptides may be particularly problematic for a method like DIA-Umpire, which is designed to construct pseudo-spectra on the basis of precursor signals.

To test this hypothesis, we compared the peptides detected by DIA-Umpire and DIAMeter in three representative datasets acquired on Orbitrap, TripleTOF 5600 and TripleTOF 6600 mass spectrometers. We reasoned that if a peptide exhibits low precursor signal, then DIA-Umpire is unlikely to construct a

Figure 4: **Examples of peptides confidently detected by DIAMeter**. The figure shows peptides detected on different mass spectrometers, in terms of the observed fragment ions matched in the MS2 spectrum as well as the MS1 chromatograms in ± 5 scans.

pseudo-spectrum that corresponds to that peptide. Therefore, for each dataset, we searched for peptides that are detected with high confidence ($<1\%$ FDR) by DIAMeter but for which DIA-Umpire produces no corresponding pseudo-spectrum. In this analysis, we say a peptide corresponds to a pseudo-spectrum if the two m/z values are within the user-specified tolerance (in units of ppm) and if the retention time associated with the pseudo-spectrum is within ± 15 s of the peak peptide detection. We further classified the detected precursors with no corresponding pseudo-spectra into two types: those for which no m/z match is available at any retention time (Type 1), and those for which a match exists, but only outside of the ± 15 s RT window (Type 2).

Our analysis shows that DIAMeter detects many Type 1 and Type 2 precursors, regardless of the mass spectrometer type. In the data acquired on an Orbitrap (OxMetYeast with 20 m/z isolation window), among 11,355 precursors detected by DIAMeter with high confidence, 1064 (9.4%, 526 Type 1 and 538 Type 2) do not coincide with any pseudo-spectra extracted by DIA-Umpire at any quality tier. Among the 538 Type 2 precursors, 477 are identified by DIA-Umpire as different precursors with low confidence (average q -value 47.4%). Analogously, in the experiments acquired by the TripleTOF 5600 and TripleTOF 6600 (LFBench with fixed window scheme), among 9968 and 14,476 confidently detected precursors reported by DIAMeter, 117 (1.2%, 29 Type 1 and 88 Type 2) out of 9968 and 205 (1.4%, 32 Type 1 and 173 Type 2) out of 14,476 do not coincide with any pseudo-spectra extracted by DIA-Umpire at any quality tier, respectively. Furthermore, 87 out of 88 and 164 out of 173 detected Type 2 precursors are identified by DIA-Umpire as different precursors with low confidence (average q -value 62.3% and 58.5%) in two settings, respectively.

Finally, we showcase confidently identified peptides detected in different instrument platforms (Figure 4). These peptide exhibited the lowest q -value among all peptides that lacked any corresponding pseudo-spectra. It is worth mentioning that DIAMeter can confidently detect precursors with both M+1 and M+2 isotope signals but without any monoisotopic signal (Figure 4(b)), which might be caused by glutamine deamidation. Thus, we demonstrated the credibility of peptides without detectable precursor signals, in terms of the observed fragment ions in the matched MS2 spectrum as well as the coeluting ion chromatograms.

4.3 Necessity of each DIAMeter component

A key component of DIAMeter is its use of a variety of sources of evidence to support detection of a single peptide, including the quality of the match between the observed and theoretical spectra, properties of the precursor isotope distribution, the difference between observed and predicted retention time, and properties of the peptide chromatogram. To evaluate the extent to which each of these features contributes to the peptide detection, we ran an ablation study, in which we modified DIAMeter to include various subsets of the full collection of score features. In particular, we considered 10 variants of DIAMeter: five in which we eliminate

Figure 5: **Evaluation of the marginal contribution of each edge feature on different instrument platforms.** Each bar shows the number of peptide detected by a modified variant of DIAMeter. Modifications use either individual edge features or all but one of the edge features.

a single feature (the XCorr score, the fragment matching p-value, the precursor intensity, the RT difference score, and the chromatographic score), and five in which we retain only a single feature. In each case, we run DIAMeter using the reduced features for edge re-scoring, edge filtering, and Percolator postprocessing. We carried out this analysis on the OxMetYeast (20 m/z window) and LFQBench experiments, representing three different instrument platforms.

The results of this analysis (Figure 5) suggest that all five features contribute positively to DIAMeter’s performance. In particular, the Tailor calibrated XCorr score appears to be the most informative feature, from both the inclusion and exclusion perspective, suggesting that peak matching provides the most important evidence in DIA data. Conversely, the fragment matching p-value appears to be the least informative feature on its own. Nonetheless, in combination with other features, inclusion of the fragment matching p-value facilitates better peptide detection, particularly on TripleTOF 6600 instruments.

It is worth mentioning that the utility of some features appears to vary by instrument. For example, exclusion of the precursor/fragment coelution score or the RT difference score makes a negligible contribution to peptide detection for data generated on the Orbitrap, but removing this feature from DIAMeter compromises performance on the TripleTOF instruments. Finally, the inclusion of 5 features in DIAMeter does not fully capture the sample/instrument/acquisition variations. It would be an interesting pursuit in the future to extend DIAMeter to incorporate richer information such as sample complexity, dynamic range, mass accuracy, and m/z windows

4.4 Comparison of FDR control methods

We implemented the pseudo-targets FDR estimation strategy described in Section 3.3 in order to ensure a fair comparison of different DIA analysis methods; however, this approach also allows us to compare the different FDR estimation procedures employed by these methods. To do so, we plot the nominal q-values reported by each method against the q-values estimated using the hidden pseudo-targets. The results on three different instrument platforms (Figure 6) suggest that, in general, three of the DIA analysis methods tend to be conservative, in the sense that the nominal q-values are larger than the pseudo-target q-values. The exception is Prosit+EncyclopeDIA, which yields the reverse direction for two out of three datasets. Note that these results cannot conclusively demonstrate either the validity or invalidity of the nominal q-values because the FDR is defined as the *expectation* (roughly, the average over all possible datasets) of the proportion of false discoveries at a given score threshold. Nonetheless, the results are suggestive that some of these methods (in particular, DIA-Umpire and PECAN) could potentially gain statistical power simply by improving their methodology for FDR estimation.

Figure 6: **Comparison of FDR control methods.** The figure plots the nominal peptide-level q-values reported by each method versus the corresponding q-values estimated by using pseudo-targets, on three datasets acquired by different instrument platforms.

5 Discussion

In this work, we introduced a novel and versatile method, DIAMeter, to detect peptides directly from DIA data without dependence on a spectral library. Compared to existing DIA analysis methods, DIAMeter offers several advantages. First, DIAMeter by design can incorporate DIA data acquired using wide precursor isolation windows. Second, DIAMeter is based upon a conventional DDA database search engine, which is readily able to detect peptides with PTMs. Third, DIAMeter works well on a variety of instrument platforms, because the method makes no assumptions about instrument-specific fragmentation patterns, instrument resolution, or patterns of background noise. Forth, DIAMeter by design can detect peptides with weak or even undetectable precursor signals.

This work points to several promising directions for future research. One possible extension is to accelerate DIAMeter by converting the matching paradigm from our current spectrum-to-peptide matching approach to the spectrum-to-peptides matching approach employed by MSFragger,³⁷ with the aid of a fragment index. Another promising direction could be customizing the spectrum matching score function to incorporate prior knowledge about peptide fragmentation.³⁸

Acknowledgments This work was supported by National Institutes of Health awards R01 GM121818, R35 GM119536, and P41 GM103533, as well as a grant from the Keck Foundation.

References

- [1] J. D. Venable, M. Q. Dong, J. Wohlschlegel, 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] A. Doerr. DIA mass spectrometry. *Nature Methods*, 12(1):35, 2014.
- [3] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20:3551–3567, 1999.
- [4] R. Craig and R. C. Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, 20:1466–1467, 2004.
- [5] S. Kim, N. Mischerikow, N. Bandeira, J. D. Navarro, L. Wich, S. Mohammed, A. J. Heck, and P. A. Pevzner. The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Molecular and Cellular Proteomics*, 9(12):2840–2852, 2010.

- [6] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant. Open mass spectrometry search algorithm. *Journal of Proteome Research*, 3:958–964, 2004.
- [7] Viktoria Dorfer, Peter Pichler, Thomas Stranzl, Johannes Stadlmann, Thomas Taus, Stephan Winkler, and Karl Mechtler. MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *Journal of Proteome Research*, 13(8):3679–3684, 2014.
- [8] H. L. Röst, G. Rosenberger, P. Navarro, L. Gillet, S. M. Miladinovic, O. T. Schubert, W. Wolski, B. C. Collins, J. Malmstrom, L. Malmstrom, and R. Aebersold. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Methods*, 32(3):219–223, 2014.
- [9] R. Bruderer, O. M. Bernhardt, T. Gandhi, S. M. Miladinovic, L. Y. Cheng, S. Messner, T. Ehrenberger, V. Zanotelli, Y. Butscheid, C. Escher, O. Vitek, O. Rinner, and L. Reiter. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Molecular & Cellular Proteomics*, 14:1400–1410, May 2015.
- [10] 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.
- [11] 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.
- [12] Jian Wang, Monika Tucholska, James D R Knight, Jean-Philippe Lambert, Stephen Tate, Brett Larsen, Anne-Claude Gingras, and Nuno Bandeira. MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. *Nature Methods*, 12(12):1106–1108, 2015.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] S. Tiwary, R. Levy, P. Gutenbrunner, F. S. Soto, K. K. Palaniappan, L. Deming, M. Berndt, A. Brant, P. Cimermanic, 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.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] Yuanyue Li, Chuan-Qi Zhong, Xiaozheng Xu, Shaowei Cai, Xiurong Wu, Yingying Zhang, Jinan Chen, Jianghong Shi, Shengcai Lin, and Jiahui Han. Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. *Nature Methods*, 12(12):1105–1106, 2015.

- [21] C.-C. Tsou, D. Avtonomov, B. Larsen, M. Tucholska, H. Choi, A.-C. Gingras, and A. I. Nesvizhskii. DIA-Umpire: a comprehensive computational framework for data-independent acquisition proteomics. *Nature Methods*, 12(3):258–264, 2015.
- [22] Y. S. Ting, J. D. Egertson, J. G. Bollinger, B. Searle, S. H. Payne, W. S. Noble, and M. J. MacCoss. PECAN: a library free peptide detection tool for data-independent acquisition tandem mass spectrometry data. *Nature Methods*, 14(9):903–908, 2017.
- [23] F. Zhang, W. Ge, G. Ruan, X. Cai, and T. Guo. Data-independent acquisition mass spectrometry-based proteomics and software tools: A glimpse in 2020. *Proteomics*, page 1900276, 2020.
- [24] L. Käll, J. Canterbury, J. Weston, W. S. Noble, and M. J. MacCoss. A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4:923–25, 2007.
- [25] J. K. Eng, A. L. McCormack, and J. R. Yates, III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5:976–989, 1994.
- [26] B. Diamant and W. S. Noble. Faster SEQUEST searching for peptide identification from tandem mass spectra. *Journal of Proteome Research*, 10(9):3871–3879, 2011.
- [27] P. Sulimov and A. Kertész-Farkas. Tailor: A nonparametric and rapid score calibration method for database search-based peptide identification in shotgun proteomics. *Journal of Proteome Research*, 19(4):1481–1490, 2020.
- [28] V. Granholm, W. S. Noble, and L. Käll. A cross-validation scheme for machine learning algorithms in shotgun proteomics. *BMC Bioinformatics*, 13(Suppl 16):S3, 2012.
- [29] P. Navarro, J. Kuharev, L. C. Gillet, O. M. Bernhardt, B. MacLean, H. L. Rost, S. A. Tate, C. C. Tsou, L. Reither, U. Distler, G. Rosenberger, Y. Perez-Riverol, A. Nesvizhskii, R. Aebersold, and S. Tenzer. A multicenter study benchmarks software tools for label-free proteome quantification. *Nature Biotechnology*, 34(11):1130–1136, 2016.
- [30] Y. Liu, A. Buil, B. C. Collins, et al. Quantitative variability of 342 plasma proteins in a human twin population. *Molecular Systems Biology*, 11(2):786, 2015.
- [31] J. A. Vizcaíno, A. Csordas, N. del Toro, J. A. Dienes, J. Griss, I. Lavidas, G. Mayer, Y. Perez-Riverol, F. Reisinger, T. Ternent, Q. Q. Xu, R. Wang, and H. Hermjakob. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Research*, 44:D447–D456, 2016.
- [32] R. A. Studer, R. A. Rodriguez-Mias, K. M. Haas, et al. Evolution of protein phosphorylation across 18 fungal species. *Science*, 354(6309):229–232, 2016.
- [33] Juri Rappsilber, Matthias Mann, and Yasushi Ishihama. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using stagetips. *Nature protocols*, 2(8):1896, 2007.
- [34] D. Kessner, M. Chambers, R. Burke, D. Agnus, and P. Mallick. Proteowizard: open source software for rapid proteomics tools development. *Bioinformatics*, 24(21):2534–2536, 2008.
- [35] A. Hu, Y. Y. Lu, J. Bilmes, and W. S. Noble. Joint precursor elution profile inference via regression for peptide detection in data-independent acquisition mass spectra. *Journal of Proteome Research*, 18(1):86–94, 2018.
- [36] V. Granholm, W. S. Noble, and L. Käll. On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics. *Journal of Proteome Research*, 10(5):2671–2678, 2011.

- [37] A. T. Kong, F. V. Leprevost, D. M. Avtonomov, D. Mellacheruvu, and A. I. Nesvizhskii. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods*, 14(5):513–520, 2017.
- [38] W. Bai, J. Bilmes, and W. S. Noble. Bipartite matching generalizations for peptide identification in tandem mass spectrometry. In *ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 327–336, Seattle, WA, 2016.

A Supplement

LFQbench dataset						
Sample ID	Instrument platform	Acquisition scheme	MS1 scan	MS2 scan	Gradient	m/z range
HYE124_32fix_L150206_001	TripleTOF 5600	32 fixed (25 m/z)	2180	69760	120	400-1200
HYE124_32fix_L150211_002	TripleTOF 6600	32 fixed (25 m/z)	2196	70304	120	400-1200
HYE124_64var_L150206_007	TripleTOF 5600	64 variable	2270	145280	120	400-1200
HYE124_64var_L150211_008	TripleTOF 6600	64 variable	2171	139008	120	400-1200

OxMetYeast dataset						
Sample ID	Instrument platform	Acquisition scheme	MS1 scan	MS2 scan	Gradient	m/z range
e01306	Orbitrap Eclipse Tribrid	40 fixed (20 m/z)	1790	71600	90	400-1200
e01307	Orbitrap Eclipse Tribrid	40 fixed (10 m/z)	1786	71440	90	400-800
e01308	Orbitrap Eclipse Tribrid	40 fixed (10 m/z)	1773	70920	90	800-1200
e01309	Orbitrap Eclipse Tribrid	40 fixed (5 m/z)	1781	71240	90	400-600
e01310	Orbitrap Eclipse Tribrid	40 fixed (5 m/z)	1778	71120	90	600-800
e01311	Orbitrap Eclipse Tribrid	40 fixed (5 m/z)	1772	70880	90	800-1000
e01312	Orbitrap Eclipse Tribrid	40 fixed (5 m/z)	1760	70400	90	1000-1200

Plasma dataset						
Sample ID	Instrument platform	Acquisition scheme	MS1 scan	MS2 scan	Gradient	m/z range
yanliu_L130104_007_SW	TripleTOF 5600	32 fixed (25 m/z)	2358	75456	120	400-1200

Table S1: Properties of the three datasets—LFQbench, Plasma and OxMetYeast—used for evaluation.

DIAMeter				
	Parameters	Orbitrap	TripleTOF 5600	TripleTOF 6600
Signal Extraction	k	5		
tide-index	missed-cleavages	2		
	mods-spec	C+57.02146,3M+15.994915		
	enzyme	trypsin		
tide-search	precursor-window	10	30	
	precursor-window-type	ppm		
	mz-bin-width	0.02		
	mz-bin-offset	0		

Table S2: DIAMeter parameters

DIA-Umpire (version 2.0)				
	Parameters	Orbitrap	TripleTOF 5600	TripleTOF 6600
Signal Extraction	RPmax		25	
	RFmax		300	
	CorrThreshold		0.2	
	DeltaApex		0.6	
	RTOverlap		0.3	
	AdjustFragIntensity		TRUE	
	BoostComplementaryIon		TRUE	
	MS1PPM		30	
	MS2PPM		40	
	SN		1.5	
	MS2SN		1.5	
	EstimateBG		TRUE	
	MinMSIntensity		300	
	MinMSMSIntensity		30	
	NoMissedScan		1	
	MaxCurveRTRange		1.5	
	RemoveGroupedPeaks		TRUE	
	RemoveGroupedPeaksRTOverlap		0.3	
	RemoveGroupedPeaksCorr		0.3	
	MinNoPeakCluster		1	
	MaxNoPeakCluster		1	
	IsoPattern		0.5	
	StartCharge		1	
	EndCharge		5	
	MS2StartCharge		2	
	MS2EndCharge		4	
	MinFrag		10	
tide-index	missed-cleavages		2	
	mods-spec		C+57.02146,3M+15.994915	
	enzyme		trypsin	
tide-search	precursor-window	10	30	
	precursor-window-type		ppm	
	mz-bin-width		0.02	
	mz-bin-offset		0	

Table S3: DIA-Umpire parameters.

PECAN / Walnut (version 0.9.0)			
Parameters	Orbitrap	TripleTOF 5600	TripleTOF 6600
acquisition		DIA	
filterPeaklists		TRUE	
ptol	10	30	
ftol	10	30	
maxMissedCleavage		2	
minCharge		1	
maxCharge		5	

Table S4: Parameters for Walnut (a re-implementation of PECAN in EncyclopeDIA¹³).

Prosit+EncyclopeDIA (version 0.9.0)				
	Parameters	Orbitrap	TripleTOF 5600	TripleTOF 6600
Prosit	intensity prediction model	Prosit_2019_intensity		
	iRT prediction model	Prosit_2019_irt		
EncyclopeDIA	acquisition	DIA		
	filterPeaklists	TRUE		
	ptol	10	30	
	ftol	10	30	
	maxMissedCleavage		2	
	minCharge		2	
	maxCharge		3	

Table S5: Prosit+EncyclopeDIA parameters.