Pioneer and Altimeter: Fast Analysis of DIA Proteomics Data Optimized for Narrow Isolation Windows

Nathan T. Wamsley[1], Emily M. Wilkerson[1], Ben Major[1,2], Dennis Goldfarb[1,3*]

[1]Department of Cell Biology and Physiology, Washington University School of Medicine, St. Louis, Missouri, United States of America.
[2]Department of Otolaryngology, Washington University School of Medicine, St. Louis, Missouri, United States of America.
[3]Institute for Informatics, Data Science & Biostastistics, Washington University School of Medicine, St. Louis, Missouri, United States of America.

*Corresponding author(s). E-mail(s): d.goldfarb@wustl.edu;
Contributing authors: n.t.wamsley@wustl.edu; emilyz@wustl.com; bmajor@wustl.edu;

Abstract

Advances in mass spectrometry have enabled increasingly fast data-independent acquisition (DIA) experiments, producing datasets whose scale and complexity challenge existing analysis tools. Those same advances have also led to the use of narrow isolation windows, which alter MS2 spectra via fragment isotope effects and give rise to systematic deviations from spectral libraries. Here we introduce Pioneer and Altimeter, open-source tools for fast DIA analysis with explicit modeling of isolation-window effects. Altimeter predicts deisotoped fragment intensity as a continuous function of collision energy, allowing a single spectral library to be reused across datasets. Pioneer re-isotopes predicted spectra per scan and combines an intensity-aware fragment index, spectral deconvolution, and dual-window quantification for fast, spectrum-centric DIA analysis. Across instruments, experimental designs, and sample inputs, Pioneer enables high-confidence identification and precise quantification at scale, completing analyses 2–6x faster and maintaining conservative false-discovery rate control.

1 Introduction

Mass spectrometry (MS)–based proteomics enables high-throughput quantification of peptides and proteins across complex biological samples. Over the past decade, faster and more sensitive mass analyzers, advances in data analysis, and models that predict peptide properties have increased the depth and throughput of proteomic measurements by more than an order of magnitude [1, 2]. These developments have enabled the widespread adoption of data-independent acquisition (DIA) for quantitative proteomics. In DIA, mass spectrometers isolate and fragment sequential m/z windows for comprehensive and reproducible precursor sampling, but co-fragmentation of multiple precursors produces chimeric spectra that complicate analysis [3–6]. While ongoing gains in instrument speed have enabled progressively narrower isolation windows to alleviate this problem, existing analysis methods do not exploit the unique properties of narrow windows and struggle to keep pace with the scale of current datasets.

Large-scale proteomics datasets now commonly arise from genetic screens, population-level clinical studies, and single-cell analyses, often with over a hundred samples analyzed per day [7–12]. Even as per-sample acquisition times continue to decrease, individual raw files can reach several gigabytes, reflecting the dense sampling enabled by rapid scans. With scan rates approaching 270 Hz, the relevant precursor m/z range can be cycled through with isolation windows as narrow as 2 m/z [13–15]. In DIA, however, isolation windows are not centered on specific

precursor features. As a result, they can exclude a substantial portion of a precursor's isotopic envelope, distorting fragment isotope distributions and the observed spectra [16–18].

Most DIA analysis tools rely on empirical or predicted spectral libraries [19–26]. In many cases, these libraries are derived from data-dependent acquisition (DDA) experiments—either because predictive models are trained on DDA spectra or because empirical libraries are constructed from DDA runs—where isolation windows are centered on precursor features rather than applied independently [27–30]. The consequence is discrepancies between the library spectra and the DIA data under analysis. Unfortunately, many widely used DIA software packages are closed source, making modification of library generation or scoring models impractical. Consistent with this limitation, recent entrapment analyses indicate that some DIA tools underreport false discovery rates [31]. However, the underlying mechanisms are difficult to identify and resolve without algorithmic transparency.

Here we introduce Pioneer and Altimeter, two open-source tools for fast analysis of DIA proteomics data that explicitly model the effects of isolation windows. Pioneer is an end-to-end workflow that performs spectrum-centric identification and quantification, while Altimeter is a deep learning model that predicts total fragment intensity rather than the monoisotopic peak intensity. Pioneer adapts predicted spectra to different isolation windows without fine-tuning, transfer learning, or re-prediction. Using rigorous false discovery rate control and extensive benchmarking, we demonstrate that Pioneer and Altimeter provide a transparent framework for DIA analysis optimized for the capabilities of current-generation instrumentation.

2 Results

2.1 A unified approach to isotope-aware DIA analysis

We developed Pioneer to perform fast DIA analysis while also accounting for precursor isolation throughout the workflow (Fig. 1A, Extended Data Fig. 1). This was motivated by the fact that isolating different subsets of a precursor's isotopic envelope produces MS2 spectra with distinct intensity patterns (Fig. 1B). While a fragment's total relative intensity does not change, the transmission of different precursor isotopes redistributes signal between the monoisotopic peak and its higher isotopes. This redistribution leads to deviations from the expected fragment ratios learned by models trained on DDA data. With the widespread adoption of narrow isolation windows, these effects are no longer rare edge cases: roughly a third of 2^+ precursors are substantially split between adjacent windows at a width of 2 m/z (Fig. 1C).

Pioneer addresses this mismatch by using spectral libraries generated by Altimeter (deployed on the Koina server [36]), which predicts total fragment intensities rather than monoisotopic intensities. Predicted spectra can then be re-isotoped on a per-scan basis according to a fitted quadrupole transmission profile, enabling direct adaptation to the window used during acquisition (Fig. 1D).

To efficiently identify candidate precursors, Pioneer uses an intensity-aware fragment index that limits each MS2 spectrum to a small set of plausible matches (Fig. 1E). The index, inspired by MSFragger [37], contains only the top predicted fragments per precursor to keep it compact while preserving sensitivity. For each MS2 spectrum, matching fragments contribute rank-based scores to candidate precursors according to their predicted rank within their respective precursor. Scores are accumulated across all peaks in the spectrum, and precursors exceeding a tunable threshold advance.

To further refine candidate precursors and perform quantification, Pioneer models each MS2 spectrum as a linear combination of re-isotoped library spectra using robust regression (Fig. 1F). This formulation accounts for chimeric fragmentation by assigning weights to co-isolated precursors. Because each scan may isolate only a portion of a precursor's isotopic envelope, Pioneer normalizes weights by the isolated fraction to recover precursor-level intensity and place adjacent windows on a common scale. After correction, Pioneer performs dual-window quantification, combining weights from adjacent windows into a single chromatographic trace with twice the data points (Fig. 1G). Together, these components prevent changes in isolation windows from altering precursor quantification.

2.2 Decoupled, continuous prediction of total fragment intensity

DIA analysis benefits from predictive models that can be optimized for each dataset and, ideally, each run. However, adapting predictions to run-specific collision energies or isolation windows typically requires repeated neural network inference or fine-tuning; both are computationally expensive. We therefore designed Altimeter to decouple run-specific adaptation from model evaluation through a representation that can be adjusted after inference.

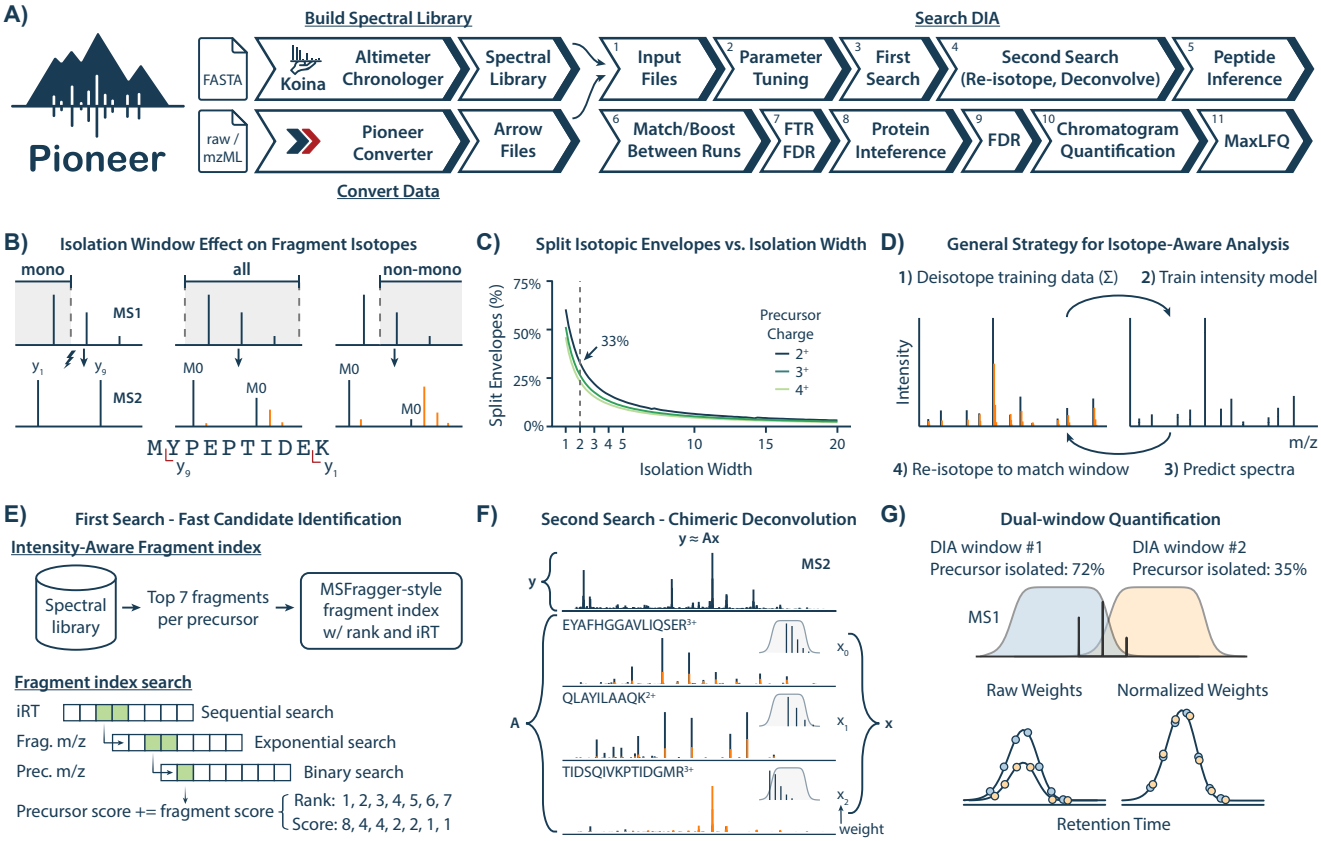

Fig. 1: **Pioneer workflow for DIA analysis. A)** Overall pipeline for building predicted spectral libraries, converting data, and searching DIA runs. Chronologer predicts retention times [32]. MaxLFQ is used to infer protein abundances from peptide abundances [33]. False discovery (FDR) and false transfer rates (FTR) are rigorously controlled. **B)** Example of how isolation windows that capture the monoisotopic, all, or non-monoisotopic precursors alter fragment isotope distributions. Non-monoisotopic fragment peaks are shown in orange. **C)** Theoretical frequency of splitting a precursor's isotopic envelope on a tryptic human digest where both windows isolate ≥25% of the envelope. **D)** General strategy for isotope-aware analysis. Training spectra are deisotoped (summed) and used to train Altimeter, a fragmentation model that predicts total fragment intensity, which are subsequently re-isotoped to match experimental isolation windows. **E)** An intensity- and iRT-aware fragment index is constructed during spectral library prediction using the top fragments per precursor. Candidate precursors are scored by summing rank-based fragment scores, and precursors with sufficient scores undergo an initial classification step to distinguish targets from decoys. **F)** Observed spectra are modeled as a linear combination of predicted spectra from candidate precursors while minimizing Huber loss. Predictions are re-isotoped to the acquisition window (inset shows window position relative to the precursor). The resulting weights are used for quantification. **G)** Dual-window quantification using precursor isolation percentages to normalize fragment intensities. Quadrupole transmission profiles are fit during parameter tuning. Chromatograms are smoothed by a Whittaker–Henderson filter [34, 35].

As an initial step, we reprocessed ProteomeTools [38, 39] to account for precursor isolation and fragment isotopes (Fig. 2A). The addition of higher isotopes doubled the annotated MS2 signal compared to monoisotopic peaks (Fig. 2B). Across the dataset, different subsets of precursor isotopes were isolated depending on their mass and charge (Extended Data Fig. 2A,B). We addressed these inconsistencies by modeling the quadrupole transmission function and using it to compute theoretical fragment isotope distributions. Doing so enabled deisotoping and recovery of total fragment intensities (Fig. 2C; Extended Data Fig. 2C). At the same time, some fragment annotations were unreliable, including cases with poorly fit isotope distributions or those outside the scan range, and were excluded from the loss function via masking (Fig. 2D; Extended Data Fig. 2D).

Crucially, while harmonizing normalized collision energy (NCE) across runs, we observed that fragment intensities varied smoothly with NCE and that low-order splines accurately modeled this relationship (Extended Data Fig. 2E–H). This led us to design a different model architecture: rather than using NCE as an input to the neural network to predict intensities at a single energy, Altimeter predicts spline coefficients for each fragment. NCE is then supplied to the spline to evaluate intensity at arbitrary energies without additional inference (Fig. 2E; Extended Data Fig. 2I,3A). Because these predictions represent total fragment intensities, Pioneer can re-isotope library spectra to match scan-specific precursor isolation efficiencies (Extended Data Fig. 3B). This design substantially improves speed and scalability. A spectral library needs to be predicted only once for a given search space—defined by the sequence database and digestion and modification settings—after which it can be reused across datasets.

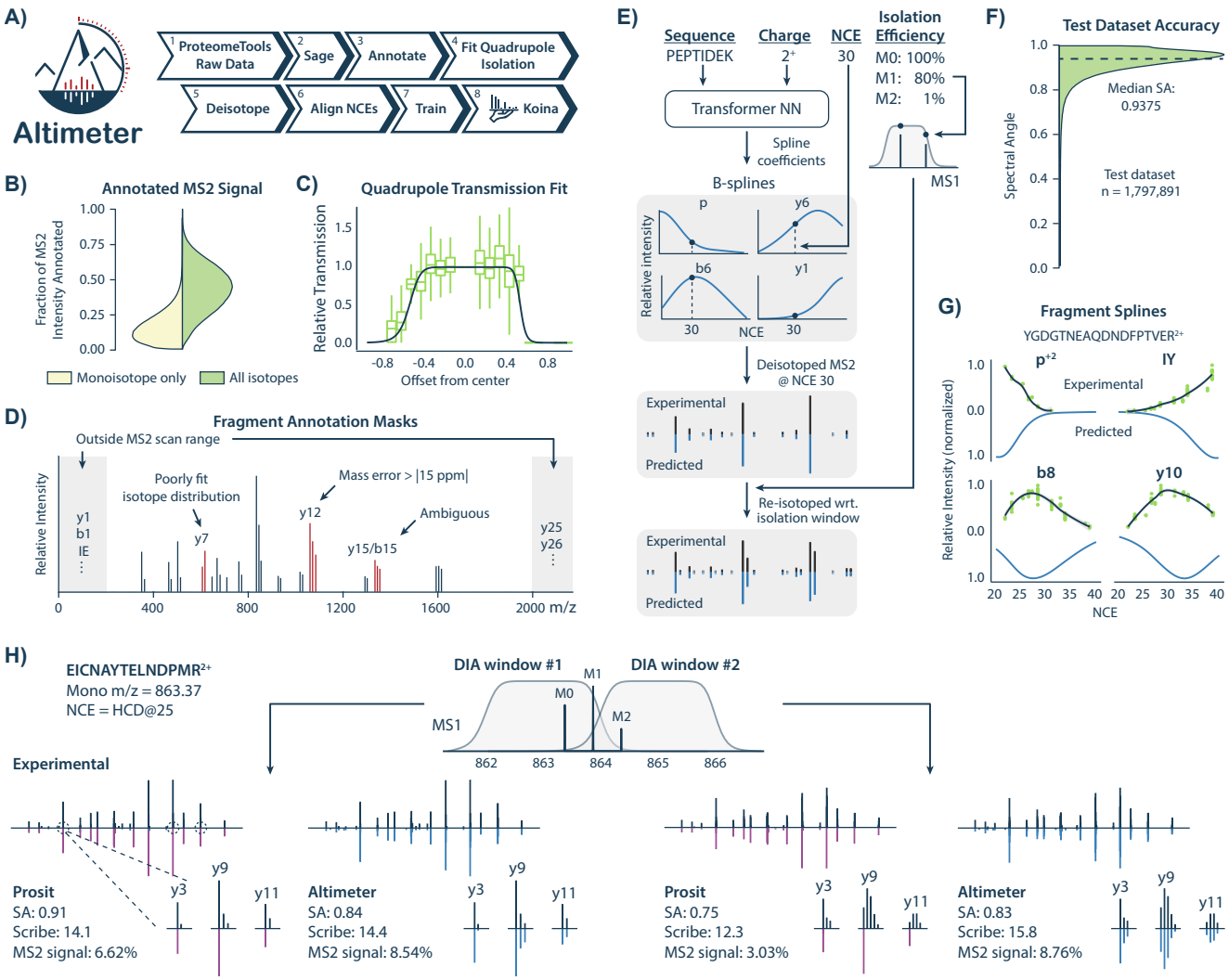

Fig. 2: **Altimeter deep learning model for peptide fragmentation prediction. A)** Model development workflow, from raw spectra to training and deployment. **B)** Fraction of MS2 spectrum intensity annotated in the training data, with and without higher isotopes. **C)** Quadrupole transmission profile fit for a set of runs having the same quadrupole calibration date. Transmission is relative to the central isotope in the window. Boxplots represent estimated transmission rates for neighboring isotopes, calculated from their relative MS2 contribution compared to theoretical precursor isotope abundances. **D)** Cartoon schematic of fragment annotation masks: outside scan range, isotope misfit, high mass error, and ambiguous assignments. **E)** Schematic of the Altimeter model: peptide sequence and charge are input to a transformer network, which outputs B-spline coefficients that model intensity profiles across normalized collision energies (NCE). Monoisotopic spectra are predicted by evaluating the splines at the desired NCE. Spectra can be further re-isotoped given the relative isolation efficiency of each precursor isotope. **F)** Spectral angle accuracy distribution on the held-out test dataset. **G)** Example fragment B-splines for a peptide from the test set, showing experimental (above) and Altimeter-predicted (below). Intensities and spline predictions were rescaled by the maximum across NCEs for visualization. **H)** Example prediction for a peptide precursor spanning two 2 m/z-wide isolation windows on an Orbitrap Astral. Experimental and predicted spectra are compared for Altimeter and Prosit, with spectral angle (SA), Scribe score, and percentage of annotated MS2 signal indicated.

To assess prediction accuracy, we evaluated Altimeter on a held-out test set comprising 10% of the data and achieved a median spectral angle of 0.9375 on the deisotoped spectra (Fig. 2F). Predicted fragment splines closely matched smoothed experimental trends across collision energies, demonstrating that the model accurately captures the continuous dependence of fragmentation on NCE (Fig. 2G). Figure 2H shows a representative example in which a precursor spans two adjacent scans. Despite substantial differences in the observed fragment isotope peaks, Altimeter accurately predicts intensities in both spectra, accounts for a larger fraction of the MS2 signal than Prosit, and achieves higher Scribe [29, 40] scores with comparable spectral angles despite additional matching peaks.

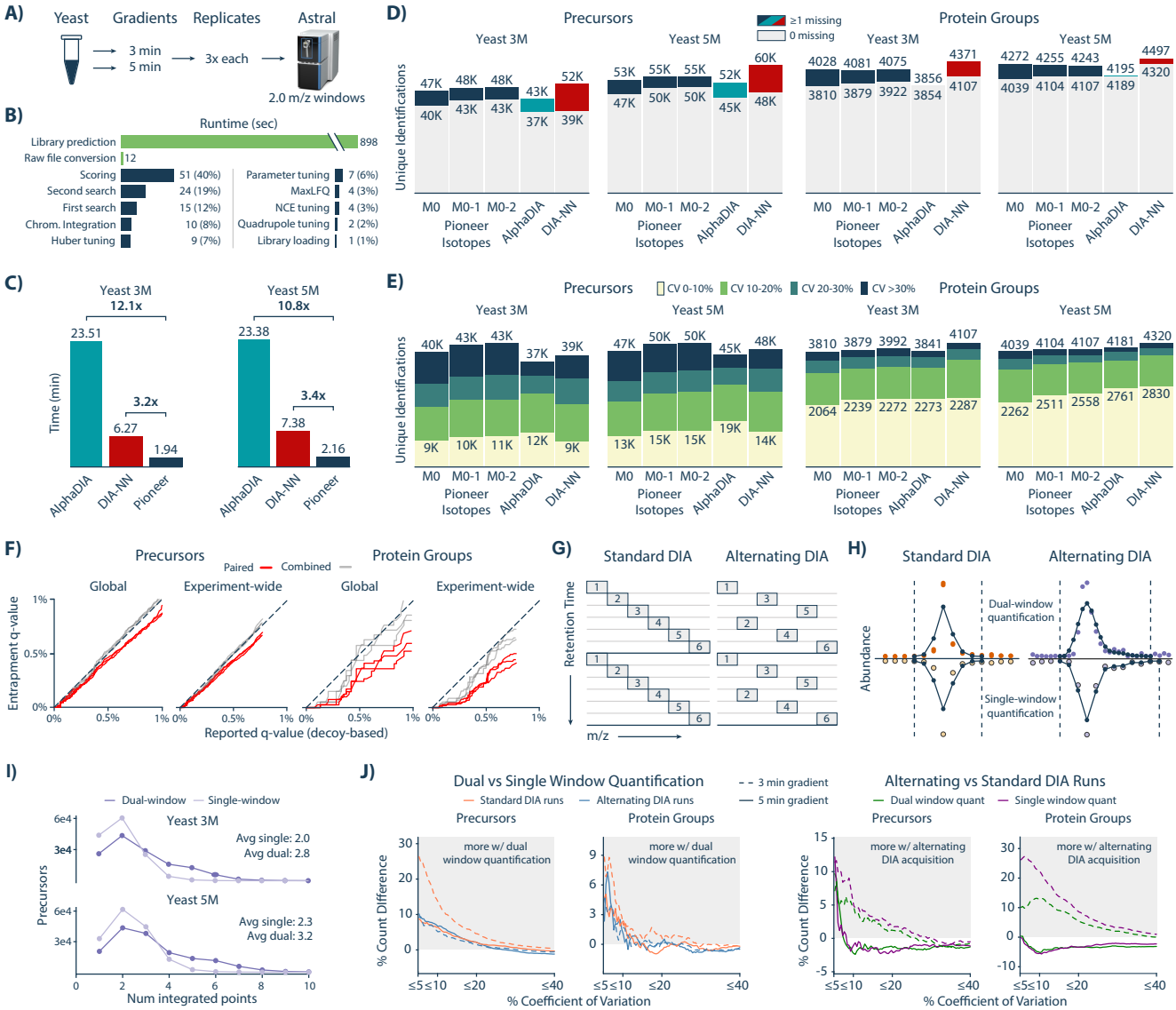

Fig. 3: Pioneer performance on narrow-window DIA benchmarks A) Experimental design schematic. Tryptic yeast digests were analyzed using 3- and 5-min gradients (3M and 5M) in triplicate on an Orbitrap Astral. **B)** Runtime breakdown for the Yeast 5M experiment. Library prediction and raw file conversion were performed independently from the search. **C)** Search times for AlphaDIA, DIA-NN, and Pioneer. **D)** Precursor and protein group identifications stratified by completeness. Pioneer was tested with different fragment isotope settings (M0, M0-1, M0-2) and compared to AlphaDIA v2.0.3 and DIA-NN v2.2.0. **E)** As in C, showing percent coefficient of variation (%CV) distributions for complete identifications. **F)** Entrapment analyses repeated with three independent libraries. **G)** Schematic of standard and alternating DIA acquisition schemes. Yeast experiments from **A** were repeated with alternating acquisition. **H)** Example chromatograms for the same precursor across the combinations of acquisition (standard/alternating) and quantification (dual-/single-window) methods. **I)** Distribution of integrated points per precursor for dual- versus single-window quantification. **J)** Identifications stratified by completeness for standard versus alternating acquisition. **K)** Percent change in identifications at varying ≤%CV thresholds using dual window quantification and alternating DIA.

2.3 Isotope-aware features improve analysis of narrow-window DIA data

We first tested Pioneer on a narrow-window yeast dataset acquired using 3- and 5- minute gradients (Yeast 3M and 5M), intentionally short to limit points across chromatographic peaks and thereby measure the value of Pioneer's dual-window quantification method (Fig. 3A). Analysis outpaced data acquisition, with a search time of 136 seconds for the triplicate 5-minute gradient runs (Fig. 3B). When benchmarked against DIA-NN [21] and AlphaDIA [41], Pioneer was over 3x and 10x faster, respectively (Fig. 3C). Pioneer achieved comparable identification rates to DIA-NN and AlphaDIA (Fig. 3D). Incorporating the M+1 fragment isotope (M0-1) increased precursor identifications by 6–8%, while inclusion of the M+2 isotope (M0-2) provided a slight further improvement. Although DIA-NN identified 10% more unique precursors and protein groups overall, restricting to measurements without missing values resulted in Pioneer identifying 4–10% more precursors and narrowing the gap at the protein

level. Quantitatively, Pioneer yielded comparable numbers across CV thresholds, while AlphaDIA produced more low-CV precursors despite having fewer identifications (Fig. 3E).

To validate FDR control, we performed shuffle-based entrapment analysis using three independent libraries with the paired and combined methods [31]. The paired estimates were consistently lower than Pioneer's reported q-values across all four controlled levels [42]: global and experiment-wide precursors and protein groups (Fig. 3F and Extended Data Fig. 4A). Moreover, the combined method—which provides an upper bound on the true error rate—was typically at or below reported q-values, further indicating that Pioneer's FDR estimates are conservative on these datasets.

Next, we examined the impact of parameter choices on identifications and quantitative precision. Increasing the Huber smoothing parameter, δ, shifts the spectral deconvolution objective toward ordinary least squares. Pioneer's automatic tuning selected a near-optimal δ value, resulting in a 4–5% increase in precursor identifications passing a $\leq$20% CV threshold (Extended Data Fig. 4B). Enabling match-between-runs improved data completeness, increasing precursors without missing values by nearly 10% in the Yeast 3M experiment and protein groups by 2%, with comparable gains in Yeast 5M (9% for precursors and 3% for protein groups; Extended Data Fig. 4C).

Next, we evaluated dual-window quantification in the context of standard and alternating DIA acquisition schemes (Fig. 3G). Dual-window quantification exploits the fact that precursors at isolation window boundaries are captured by adjacent scans. Under standard DIA acquisition, windows are scanned sequentially by m/z, so extra points behave more like technical replicates rather than improving capture of chromatographic peak shape. Alternating DIA addresses this limitation by acquiring odd-numbered windows followed by even-numbered windows, causing adjacent m/z's to be acquired halfway through the cycle and spacing the points equally (Fig. 3H). Dual-window quantification increased the average number of points per peak by 40% and yielded more IDs at lower CVs (Fig. 3I,J).

The alternating window DIA results, however, were mixed. The Astral reorders DIA scans by grouping isolation windows according to quadrupole rod polarity to improve robustness and quantitative performance [43]. As a result, standard acquisition already separates some adjacent windows across sub-cycles, reducing the benefit of an alternating scheme. In addition, the ideal alternating pattern could not be fully implemented due to firmware constraints and acquired 5% fewer scans due to overhead (Extended Data Fig. 4D–G). Under these restrictions, alternating DIA did not improve CVs for the 5-minute gradient, but substantially improved them for the 3-minute gradient, consistent with the expectation that shorter gradients benefit more from more frequent chromatogram sampling.

2.4 Pioneer has consistent performance across three-proteome benchmarks

We next compared Pioneer and DIA-NN on three-proteome benchmarks (Fig. 4A) [7, 44]. Pioneer completed analyses 2.3–4.9x faster than DIA-NN (Fig. 4B) and scaled well with increasing CPU threads, though with diminishing returns (Extended Data Fig. 5A). Although Pioneer identified 11–34% fewer unique precursors and 9–17% fewer unique protein groups than DIA-NN (Fig. 4C,D), differences were smaller for IDs without missing values. Pioneer identified 11% more complete precursors on the MTAC 3P dataset but 1–22% fewer on the remaining datasets, and 2–17% fewer complete protein groups overall. Both tools showed similar quantitative precision and accuracy, with Pioneer exhibiting a broader dynamic range in observed fold changes (Fig. 4E,F).

Protein abundance distributions are expected to be approximately log-normal [45]. In contrast, DIA-NN v2.2.0 produced heavy-tailed or bimodal distributions in some benchmarks, most prominently in the MTAC 3P and Guzman Exploris datasets (Fig. 4G); this behavior was not observed with DIA-NN v1.8.1 (Extended Data Fig. 5B–D). When applying Perseus's standard imputation method, which assumes log-normality, imputed DIA-NN abundances frequently exceeded observed values (Fig. 4H) [46]. We therefore tested for false-positive differentially expressed proteins before and after imputation (Fig. 4I). In the MTAC 3P experiment, DIA-NN reported more human false positives than Pioneer without imputation (11 vs. 2), and this gap increased further after imputation (22 vs. 4).

2.5 Pioneer supports diverse large-scale and low-input datasets

Having established performance on controlled benchmarks, Pioneer was next evaluated on diverse, real-world experiments, where it achieved greater data completeness and faster runtimes despite fewer unique identifications.

In a 312-run yeast knockout dataset (Fig. 5A) [7], Pioneer identified 66K unique precursors and 4,544 protein groups, compared to 91K and 5,048 for DIA-NN (Fig. 5B). Despite reduced identification depth, Pioneer achieved higher data completeness at the precursor level, with 77% of intensity values present across the full matrix compared

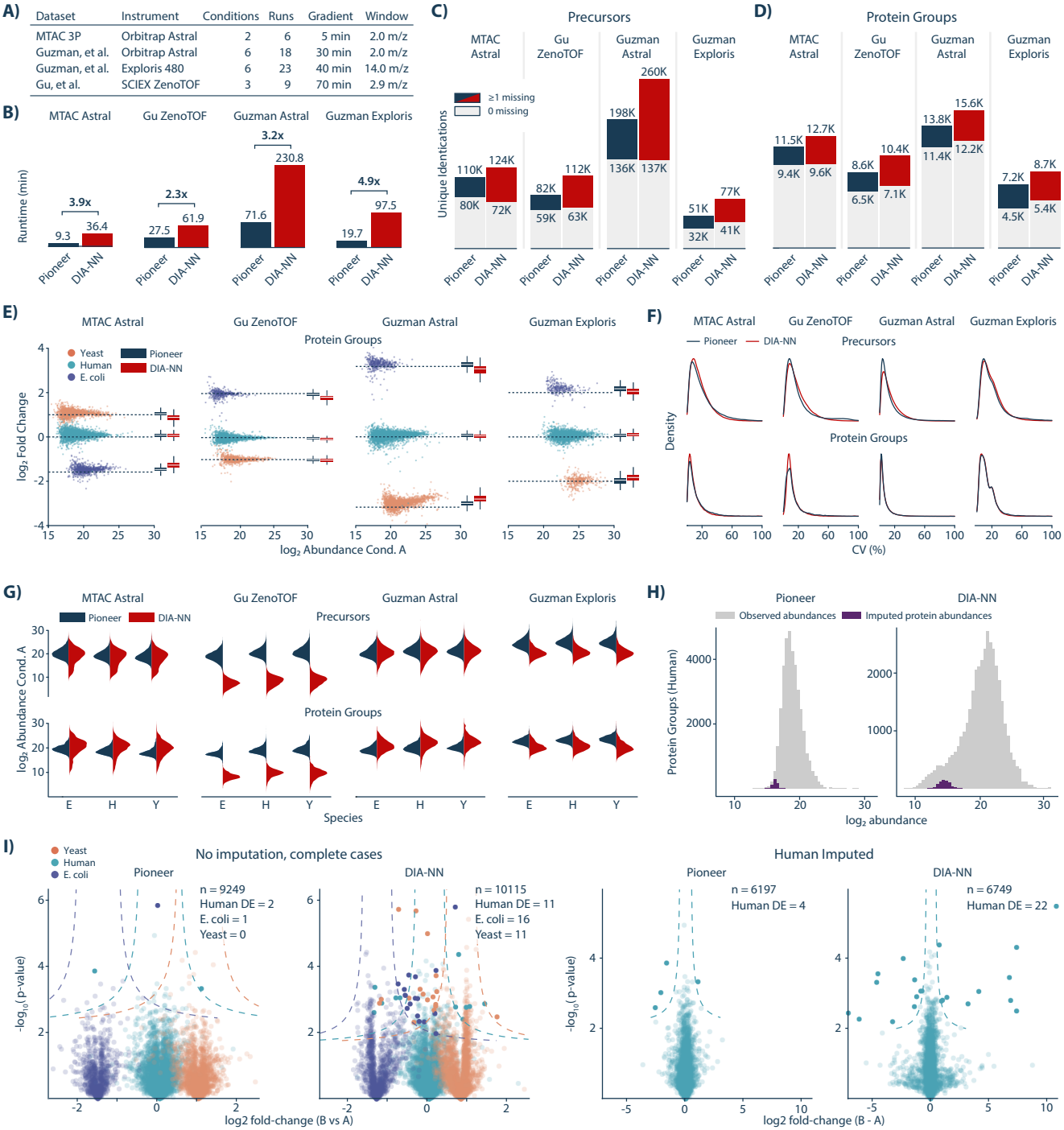

Fig. 4: **Pioneer performance on quantitative three-proteome benchmarks. A)** Overview of three-proteome benchmark experiments. **B)** Analysis runtimes for Pioneer and DIA-NN v2.2.0. Times exclude library generation and file format conversion for both tools. **C)** Unique precursor identifications with zero or $\geq$1 missing values. **D)** As in **C**, but for protein groups. **E)** Log_2-fold change (conditions B vs A) versus $\log_2$ abundance for each protein group for Pioneer. Boxplots summarize Pioneer and DIA-NN results. **F)** Density plot of %CVs for precursors and protein groups. **G)** Log_2 abundance distributions. **H)** MTAC 3P example run showing human protein abundances and Perseus-imputed values with default parameters (width 0.3, downshift 1.8). Imputation was performed only when at least one condition had no missing values. **I)** Differential expression analysis (Perseus) for the MTAC 3P benchmark. For each species, differential expression was tested relative to its median $\log_2$ fold change using a two-sample t-test with a 5% FDR and $s_0 = 0.1$. Left: no imputation, including all species; right: analysis restricted to human protein groups after imputing missing values.

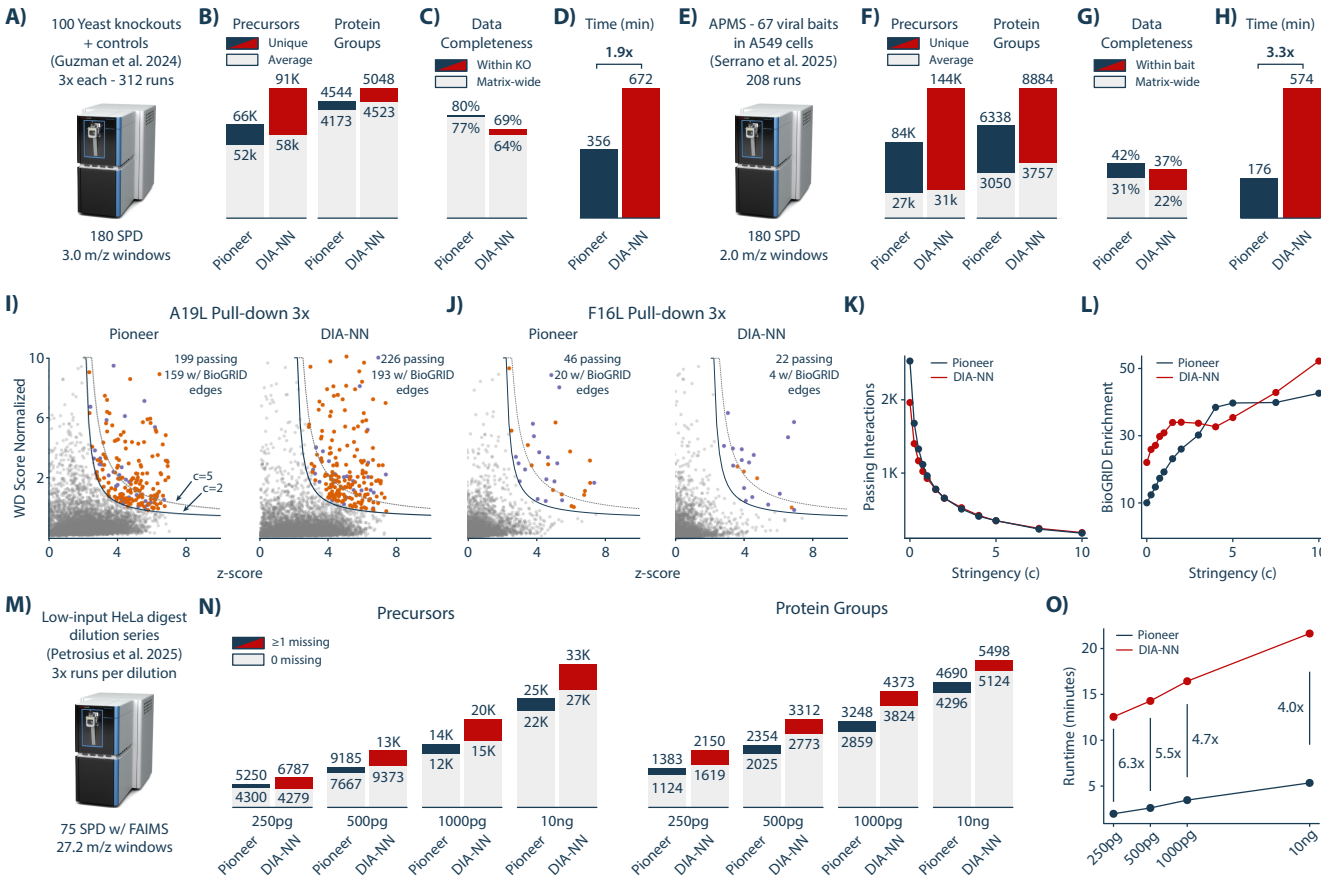

Fig. 5: **Pioneer performance across diverse experimental datasets. A)** Schematic of the yeast knockout (KO) experiment from Guzman et al [7]. **B)** Number of unique and mean-per-run identifications). **C)** Precursor-level data completeness. Grey bars indicate the fraction of non-missing intensity values across the full quantitative matrix, while blue and red bars indicate within-condition completeness, defined as precursor–knockout pairs quantified across all three replicates. **D)** Runtime comparison for the yeast KO experiment. **E)** Schematic of the APMS experiment, in which vaccinia virus proteins were used as baits [12]. **F)** Number of unique and mean-per-run identifications for the APMS experiment. **G)** As in **C**, for the APMS experiment. **H)** Runtime comparison for the APMS experiment. **I-J)** WD-scoring analysis for two example baits, A19L (I) and F16L (J). Because vaccinia bait proteins are absent from interaction databases, we assessed prey-prey interactions among co-purified human proteins. Y-axis: WD-scores normalized to 95th percentile. X-axis: prey abundance Z-score. Blue dots: proteins passing c≥2 threshold. Red dots: passing proteins with a prey partner also passing and present in the BioGRID physical multi-validated network [47]. **K)** Bait-prey pairs passing each stringency threshold. **L)** Enrichment of BioGRID-validated prey-prey interactions over a null model (100 permuted networks) at each stringency threshold. **M)** Schematic of the single-cell-equivalent experiment from Petrosius et al. [10]. **N)** Identifications stratified by zero versus ≥1 missing values across the dilution series. **O)** Runtime comparison across dilution points.

to 65% for DIA-NN (Fig. 5C), and higher within-condition completeness (80% vs 69%). Pioneer completed this analysis in 6 h, corresponding to a 1.9x speedup (Fig. 5D).

In an affinity-purification dataset (Fig. 5E) [12], Pioneer identified 84K unique precursors and 6,338 protein groups, compared to 144K and 8,884 for DIA-NN (Fig. 5F). Again, Pioneer achieved higher data completeness at the precursor level, with 31% of intensity values present across the full matrix compared to 22% for DIA-NN, and higher within-condition completeness (42% vs 37%; Fig. 5G). Pioneer completed the full analysis 3.3x faster (Fig. 5H). Interaction recovery was evaluated using prey–prey interactions, as validated viral bait–human prey interactions are unavailable, and yielded similar numbers of BioGRID pairs despite reduced identification depth (Fig. 5I,J). At matched stringency thresholds, Pioneer and DIA-NN showed comparable enrichment for validated interactions relative to permuted control networks (Fig. 5K,L).

Finally, Pioneer was evaluated on a single-cell-equivalent dilution series (Fig. 5M) [10]. Compared to DIA-NN, Pioneer reported fewer precursor and protein group identifications across dilution points, with the largest differences at low sample input (29–43% and 17–55%, respectively; Fig. 5N). Differences in complete identifications were smaller (0–25% more precursors and 19–44% more protein groups) and decreased with increasing input at the protein group level. However, Pioneer completed its analyses 4-6.3x faster. This advantage exceeded that for the bulk samples, but diminished with increasing sample input (Fig. 5O).

2.6 Robust control of false-discovery and false-transfer rates

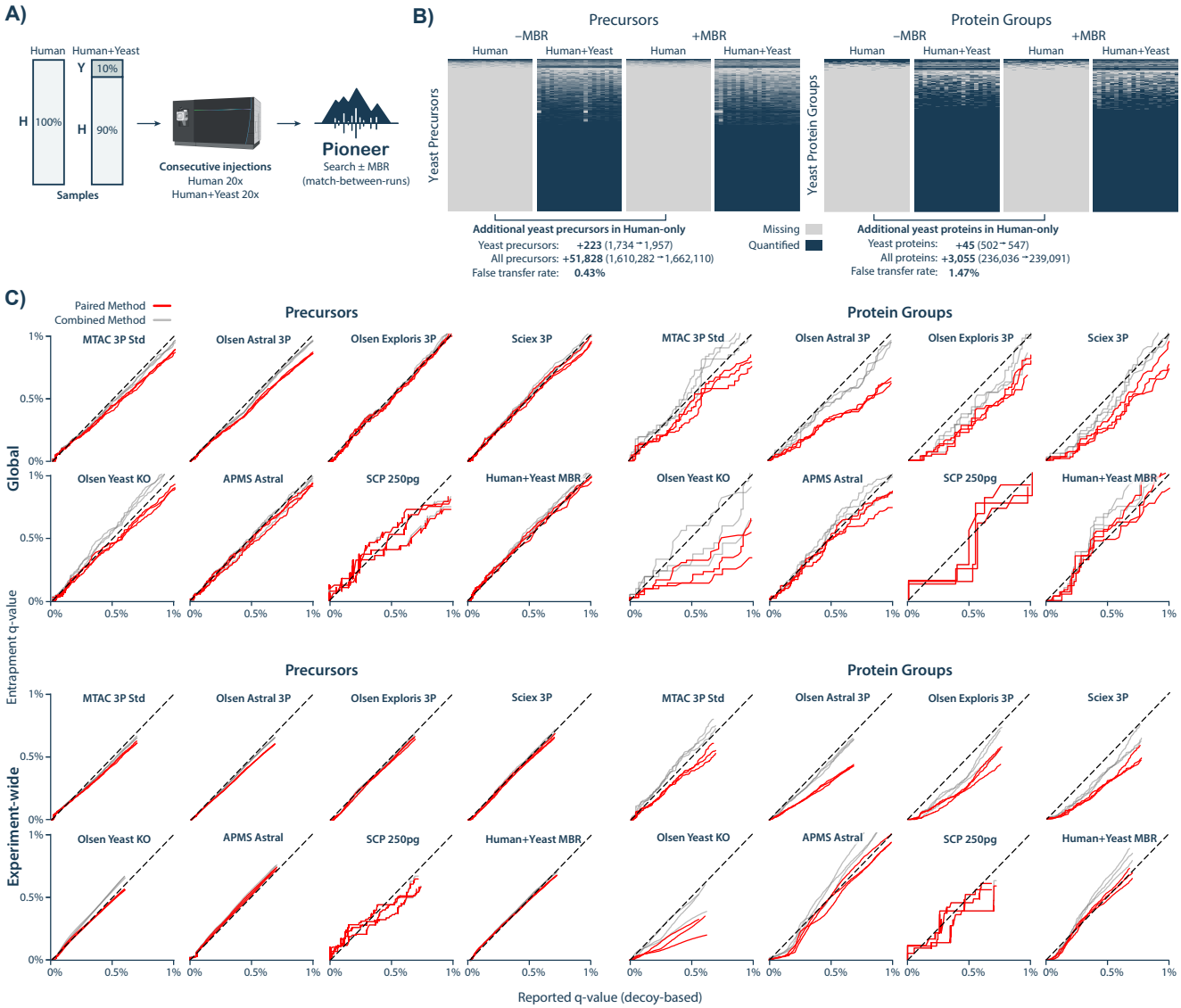

Fig. 6: **False-transfer rate evaluation and entrapment analyses. A)** Experimental design for false-transfer rate (FTR) estimation: twenty injections of a tryptic human digest followed by twenty injections of a human-yeast mixture. Data were searched using Pioneer with and without match-between-runs (MBR). **B)** Yeast identifications in human-only and two-proteome runs with MBR on or off. The FTR is the net gain of spurious yeast IDs in human-only runs divided by the total net increase in identifications across all runs with MBR. **C) C)** Entrapment analyses for the remaining datasets in the main text not previously evaluated by entrapment. All four FDRs (global and experiment-wide, at precursor and protein group levels) were jointly filtered at 1% thresholds.

Finally, we validated Pioneer's false-transfer rate (FTR) and FDR control. To estimate FTR, we generated a two-proteome MBR dataset following a previously established strategy [48] (Fig. 6A). Yeast identifications gained by MBR in human-only runs were interpreted as false transfers and divided by the total increase in identifications. The estimated FTR was 0.45% at the precursor level and 1.47% at the protein group level; both near the 1% target (Fig. 6B).

Entrapment analyses across the remaining datasets confirmed appropriate control of FDRs. Entrapment q-values closely tracked decoy-based q-values at global and experiment-wide levels (Fig. 6C, Extended Data Fig. 6). Both the combined and paired entrapment methods showed no evidence of inflated false discovery rates. Notably, these searches represent the most memory-intensive analyses performed in this study; maximum memory usage observed across all searches was 41 GB, occurring for the Olsen Exploris 3P dataset.

3 Discussion

Pioneer and Altimeter are designed to address central challenges in DIA proteomics: maintaining consistent identification, accurate quantification, and rapid analysis at scale despite the intrinsic complexity of DIA data.

Several studies have noted that DIA MS2 spectra differ from DDA due to increased co-fragmentation, and the absence of charge-dependent collision-energy optimization [49–51]. Our results further show that differences in precursor isotope isolation alter fragments' isotopic envelopes and contribute substantially to this divergence. This has consequences for both library construction and quantification. Empirical libraries are implicitly tied to the isolation windows under which they were generated. For a given precursor, a library may reflect only a single window, leading to preferential identification in that window and misquantification in others. Changes to isolation windows within a dataset would further exacerbate these effects and introduce systematic quantitative biases. Pioneer corrects for these effects, making quantification robust to differences in isolation windows.

In practice, effective isolation can change even when the acquisition method is held constant. Drift in quadrupole transmission over time, instrument recalibration or maintenance, and acquisition across multiple instruments all alter precursor isolation. The same factors also introduce differences in collision energy. Transfer learning and fine-tuning can partially absorb such shifts, but they are computationally expensive and typically applied globally rather than per run. In contrast, Altimeter decouples prediction from acquisition by using peptide sequence and charge as its core inputs, with collision energy and isolation effects applied during analysis, enabling per-run adaptation without retraining. This architectural separation positions Altimeter as a potential foundation model, where a shared sequence- and charge-based representation could support distinct downstream decoders for different fragmentation methods or collisional cross-section prediction.

In addition to its modeling implications, this decoupling enables efficient, scalable analysis workflows. Because our spectral libraries are predicted independently of data acquisition, they can be generated ahead of time, and raw data conversion can proceed on-the-fly as files complete. After the full data collection, only the Pioneer search remains, which motivated our focus on search speed when benchmarking runtime. This separation of prediction from acquisition reduces end-to-end latency and makes Pioneer well suited for large, continuously growing datasets.

Beyond search throughput, downstream analyses depend on the quality of the resulting quantitative matrix. In our benchmarks, Pioneer achieved high data completeness, reducing reliance on imputation. When applied without careful validation, imputation can introduce false positives in downstream analyses, whereas the distributions produced by Pioneer closely match expected behavior, mitigating these risks.

Together, these results establish Pioneer and Altimeter as a rigorously validated, open-source and cross-platform solution for DIA proteomics that combines quantitative accuracy with strict control of false discovery rates. Pioneer is designed to scale to large datasets while remaining efficient in both runtime and memory usage, enabling deployment on standard compute infrastructure without reliance on local GPUs. This software suite provides a practical foundation for modern large-scale DIA studies.

4 Methods

4.1 MTAC Benchmark Experiments

Three-proteome and yeast benchmarks (Figures 3 and 4) were performed at the Mass Spectrometry Technology Access Center at the McDonnell Genome Institute (MTAC@MGI) at Washington University School of Medicine. Commercial protein digest standards were obtained for *Homo sapiens* (Pierce HeLa Protein Digest Standard, Thermo Scientific, Cat. No. 88329, Lot UF286759), *Saccharomyces cerevisiae* strain BY4741 (Pierce Yeast Digest Standard, Thermo Scientific, Cat. No. A47951, Lot AB407858), and *Escherichia coli* (MassPREP E. coli Digest Standard, Waters Corporation, Part No. 186003196, Lot W07052404). Two three-proteome mixtures were prepared by combining the digests at the following mass ratios: Condition A contained 50% human, 20% yeast, and 30% *E. coli*, while Condition B contained 50% human, 40% yeast, and 10% *E. coli*. Stock solutions were diluted to 80 ng/µL in 98% water, 2% acetonitrile, and 0.1% formic acid. A 5 µL injection (400 ng on column) was used for all samples.

Samples were analyzed on an Orbitrap Astral mass spectrometer (Thermo Scientific) coupled to a Vanquish Neo LC system with a µPAC Neo 5.5 cm column (75 µm ID) maintained at 50°C. Mobile phase A was water with 0.1% formic acid; mobile phase B was 80% acetonitrile with 0.1% formic acid. Peptides were separated at 1.5 µL/min using a 5-minute active gradient from 12.6% to 45% B. MS1 scans were acquired in the Orbitrap at 240,000 resolution (380–980 m/z, 5 ms max IT, 500% AGC). DIA scans were acquired in the Astral analyzer using 280 windows of 2 Th isolation width spanning 392–948 m/z (150–2000 m/z scan range, 25% NCE, 5 ms max IT,

0.6 s cycle time). Three technical replicates were acquired for each condition. A yeast-only digest was analyzed in triplicate using both 3-minute and 5-minute active gradients.

4.2 Alternating DIA scheme

To design an alternating isolation window method compatible with the Orbitrap Astral, we first acquired data using sequential m/z windows (standard method) and extracted the realized MS2 scan order from the raw file. Two sub-cycles were detected, corresponding to quadrupole rod polarity groups. We then constructed a revised method in which isolation windows were reordered to maximize temporal separation between m/z-adjacent windows while preserving consecutive acquisition within each polarity block. This produced the most alternating schedule achievable without violating polarity grouping or triggering scan reordering.

In the standard method, approximately half of m/z-adjacent isolation windows (2 m/z apart in center mass) were acquired consecutively, and the remainder were separated by approximately one-half of a scan cycle (∼140 scans). In the alternating method, no m/z-adjacent windows were acquired consecutively; separations were distributed across the cycle, with approximately one quarter separated by one quarter of a cycle (∼70 scans), one half by one half of a cycle (∼140 scans), and one quarter by three quarters of a cycle (∼210 scans).

4.3 Spectral Library Generation

Spectral libraries were generated by tryptic *in silico* digestion of the UniProt reference proteomes for *Homo sapiens* (UP000005640), *Saccharomyces cerevisiae* (UP000002311), and *Escherichia coli* strain K12 (UP000000625), using canonical sequences without isoforms. Parameters common to all tools included cleavage before proline, 1 missed cleavage, carbamidomethylation of cysteine as a fixed modification, 1 variably oxidized methionine, 7-40 AA peptide length, and precursor charges 2–3. For all datasets, the instrument scan m/z ranges were fully covered by their corresponding libraries. Altimeter library precursor and fragment m/z ranges were automatically detected and trimmed to match the experiment files. DIA-NN libraries were generated using its built-in model with the default precursor m/z range, 300–1800, and fragment m/z range, 200–1800. AlphaDIA libraries were generated using AlphaPeptDeep [27] with precursor m/z range 400–1200, fragment m/z range 200–2000, b/y ions with maximum charge 2, and NCE 25 with the Astral instrument model.

4.4 DIA Software Analyses

Pioneer analyses were performed using version 0.6.4 with default parameters unless otherwise indicated. DIA-NN analyses were performed using version 2.2.0 with default parameters, except protein inference was set to "Protein Names". AlphaDIA analyses were performed using version 2.0.3 with default parameters; MBR was enabled, FDR was controlled at the protein level using heuristic inference, and quantification used directLFQ normalization. For all three tools, results were filtered to a 1% FDR threshold. Run-to-run normalization was disabled for Pioneer and DIA-NN because several datasets (e.g., three-proteome mixtures and APMS) violate the assumption of global proteome stability across runs, and identification benchmarks are unaffected by normalization. Analyses were performed on a computer running Windows 10 with a 3.0 GHz 18-core Intel i9-10980XE CPU, 128 GB 2666 MHz DDR4 RAM, using 24 threads.

4.5 Statistics and Imputation

Throughout our analysis, we reported the coefficient of variation (CV) using a bias-corrected formula

$$\left(1 + \frac{1}{4n}\right) \frac{s}{\bar{x}} \tag{1}$$

where n is the sample size, s is the sample standard deviation, and $\bar{x}$ is the sample mean [52]. We calculated the CV on the raw abundances and not on $\log_2$ transformed abundances. For the differential expression analysis of the MTAC three-proteome experiment Fig. 4H-I, imputation and statistical testing were carried out in Perseus [46]. Missing values were imputed by drawing from a normal distribution with a width of 0.3 standard deviations and a downshift of 1.8 standard deviations from the mean of the observed abundance distribution [53]. Imputation was restricted to protein groups for which at least one condition had no missing values. Differential expression was assessed in Perseus using a two-sample t-test ($s_0 = 0.1$, 5% FDR). For the human-only differential expression analysis in (Fig. 4L) we filtered out non-human protein-groups before imputation.

4.6 Protein-Protein Interaction Scoring and Evaluation

To validate protein-protein interactions from APMS data [12] (Fig. 5E-L), we applied WD-scoring as described by Sowa et al. and Huttlin et al. [54, 55]. WD-scores were normalized to the 95th percentile across all bait-prey pairs. We applied a combined WD-score and Z-score threshold to identify high-confidence interactions [12]. For each prey protein, we calculated the Z-score as $Z = (\bar{X} - \mu)/\sigma$, where $\bar{X}$ is the mean prey abundance for the given bait, and μ and σ are the global mean and standard deviation of the prey's abundance across all samples. Prey proteins were considered interactors if they satisfied:

$$\mathrm{WD}_{\mathrm{norm}} \geq 1 + \frac{c}{Z - 2} \tag{2}$$

where c is a stringency parameter. Prey proteins with $Z \leq 2$ were excluded. We evaluated performance across stringency thresholds from $c = 0$ to $c = 10$. Missing values were not imputed; mean abundances were calculated from detected replicates only.

To assess biological validity, we identified putative prey-prey pairs found in the BioGRID physical multi-validated interaction network (version 5.0.252) [47]. Because vaccinia virus proteins served as baits and were absent from BioGRID, we restricted our analysis to prey-prey interactions among human proteins. We constructed 100 permuted networks by filtering BioGRID to human-to-human interactions and randomly shuffling the pairings. At a given stringency parameter, c, we counted prey-prey pairs where both members passed the scoring threshold. Enrichment was calculated as the ratio of BioGRID-validated pairs observed in the true network to the mean number observed across the permuted networks.

4.7 Entrapment Analysis

To evaluate false discovery rate control, we performed entrapment analysis following Wen et al. [31]. We generated target peptide sequences by *in silico* tryptic digestion of the protein FASTA database and created decoy sequences by shuffling target peptides while keeping the C-terminal amino acid fixed, re-shuffling up to 20 times if needed to avoid duplicates. We then applied the same shuffling procedure to these original target and decoy peptide sequences to generate paired entrapment peptides with an entrapment ratio of $r = 1$. During library construction, each target precursor received an entrapment pair ID linking it to its corresponding entrapment precursor.

We independently generated three replicate spectral libraries using this procedure to assess variability in the entrapment estimates, then searched DIA data against each library at both 1% and 5% q-value thresholds. We estimated the false discovery proportion using both the combined and paired methods [31].

For protein group-level analysis, the paired method required a matching strategy to account for differences in protein group composition between target/decoy and entrapment identifications. Because protein inference can assign different sets of protein accessions to target versus entrapment peptides, direct one-to-one pairing is not always possible. For each entrapment protein group, we therefore matched it to the target protein group sharing the maximum number of protein accessions.

4.8 False Transfer Rate Estimation

To estimate the false transfer rate (FTR) of match-between-runs, we prepared two sample types: HeLa digest alone and HeLa digest spiked with yeast digest at a 9:1 mass ratio. Commercial protein digest standards were obtained from Thermo Scientific Pierce: *Homo sapiens* (Cat. No. 88329) and *Saccharomyces cerevisiae* (Cat. No. A47951). Stock solutions were prepared at 200 ng/µL and a 1 µL injection (200 ng on column) was used for all samples.

Samples were analyzed on an Orbitrap Eclipse mass spectrometer (Thermo Scientific) coupled to an UltiMate 3000 RSLCnano system in trap-elute configuration. Peptides were trapped on a Pharmafluidics µPAC trap column and separated on a µPAC Neo 50 cm column at 300 nL/min using a 58-minute active gradient from 12% to 40% B (12% to 20% B over 42 min, then 20% to 40% B over 16 min). Mobile phase A was water with 0.1% formic acid; mobile phase B was 100% acetonitrile with 0.1% formic acid. MS1 scans were acquired in the Orbitrap at 30,000 resolution (390–1010 m/z, 54 ms max IT, 100% AGC, 30% RF lens). DIA scans were acquired in the Orbitrap at 30,000 resolution using 59 windows of 10 m/z isolation width with 1 m/z overlap, spanning 400–1000 m/z (145–1450 m/z scan range, 30% NCE, 54 ms max IT, 1000% AGC, 3 s cycle time).

We acquired 20 human-only replicates followed by 20 HeLa+yeast replicates and searched together with MBR enabled and disabled. Since yeast peptides are absent from human-only samples, any yeast precursor identified in these runs represents a false transfer from human+yeast runs. We calculated the FTR as:

$$\mathrm{FTR} = \frac{N_{\text{yeast}}^{\text{MBR}} - N_{\text{yeast}}^{\text{noMBR}}}{N_{\text{total}}^{\text{MBR}} - N_{\text{total}}^{\text{noMBR}}} \tag{3}$$

where N_{yeast} denotes yeast identifications in human-only runs and N_{total} denotes total identifications across all runs (both human-only and human+yeast).

Data Availability. All mass spectrometry datasets used or generated in this study are publicly available via one of the PRIDE [56], MassIVE, or jPOST repositories [57].

Previously published datasets reanalyzed in this work include the ProteomeTools synthetic peptide dataset (PXD021013, PXD010595, PXD004732, PXD006832); three-proteome benchmark datasets acquired on Orbitrap Astral and Orbitrap Exploris instruments from Guzman et al. (PXD046444); the Guzman et al. yeast knockout benchmark dataset (PXD046386); the affinity-purification mass spectrometry (APMS) dataset (MSV000096304); the SCIEX three-proteome benchmark dataset (JPST002949); and the single-cell-equivalent proteomics dataset (MSV000095333).

Datasets generated in this study and used in the main figures have also been deposited in PRIDE, including the two-proteome match-between-runs benchmark dataset (PXD074320), MTAC yeast benchmark datasets acquired using 3-minute and 5-minute gradients with both standard and alternating DIA acquisition schemes (PXD074310), and the MTAC three-proteome benchmark dataset (PXD074310).

Additional datasets analyzed exclusively in Extended Data figures include a plasma proteomics dataset generated using the Seer Proteograph platform (PXD067119) and a HeLa benchmark dataset comparing uPAC and ionOpticks columns (PXD074355).

Code Availability. All software developed in this study is open source. Pioneer is available at https://github.com/nwamsley1/Pioneer.jl and is released under the GNU Affero General Public License v3 (AGPL-3.0). The Altimeter code is available at https://github.com/GoldfarbLab/Altimeter and is released under the MIT License. Altimeter is also deployed as a hosted inference service via Koina at https://koina.wilhelmlab.org/. Processed training data used for Altimeter model development are publicly available via Zenodo (DOI: 10.5281/zenodo.15875053).

Acknowledgments. We thank the Mass Spectrometry Technology Access Center (MTAC) at Washington University in St. Louis for access to instrumentation and support for the Orbitrap Astral experiments. We also thank Ludwig Lautenbacher for his assistance in deploying Altimeter on the Koina inference server.

References

[1] Stewart, H. I. *et al.* Parallelized acquisition of orbitrap and astral analyzers enables high-throughput quantitative analysis. *Anal. Chem.* **95**, 15656–15664 (2023).

[2] Peters-Clarke, T. M., Coon, J. J. & Riley, N. M. Instrumentation at the leading edge of proteomics. *Anal. Chem.* **96**, 7976–8010 (2024).

[3] Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **32**, 219–223 (2014).

[4] Lu, Y. Y., Bilmes, J., Rodriguez-Mias, R. A., Villén, J. & Noble, W. S. DIAmeter: matching peptides to data-independent acquisition mass spectrometry data. *Bioinformatics* **37**, i434–i442 (2021).

[5] Tsou, C.-C. *et al.* DIA-umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12**, 258–64, 7 p following 264 (2015).

[6] Ting, Y. S. *et al.* PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nat. Methods* **14**, 903–908 (2017).

[7] Guzman, U. H. *et al.* Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition. *Nat. Biotechnol.* **42**, 1855–1866 (2024).

[8] Kverneland, A. H. *et al.* Fully automated workflow for integrated sample digestion and evotip loading enabling high-throughput clinical proteomics. *Mol. Cell. Proteomics* **23**, 100790 (2024).

[9] Xuan, Y. *et al.* B-146 unlocking the potential of large-cohort proteomics studies with the orbitrap astral mass platform. *Clin. Chem.* **70**, hvae106.506 (2024).

[10] Petrosius, V. *et al.* Quantitative label-free single-cell proteomics on the orbitrap astral MS. *Mol. Cell. Proteomics* **24**, 100982 (2025).

[11] Szyrwiel, L., Gille, C., Mülleder, M., Demichev, V. & Ralser, M. Fast proteomics with dia-PASEF and analytical flow-rate chromatography. *Proteomics* **24**, e2300100 (2024).

[12] Serrano, L. R. *et al.* Affinity purification mass spectrometry on the orbitrap-astral mass spectrometer enables high-throughput protein-protein interaction mapping. *J. Proteome Res.* **24**, 2006–2016 (2025).

[13] Hsu, C. *et al.* Evaluation of a prototype orbitrap astral zoom mass spectrometer for quantitative proteomics-beyond identification lists. *J. Proteome Res.* **24**, 5742–5754 (2025).

[14] Heil, L. R. *et al.* Evaluating the performance of the astral mass analyzer for quantitative proteomics using data-independent acquisition. *J. Proteome Res.* **22**, 3290–3300 (2023).

[15] Lancaster, N. M. *et al.* Fast and deep phosphoproteome analysis with the orbitrap astral mass spectrometer. *Nat. Commun.* **15**, 7016 (2024).

[16] Goldfarb, D., Lafferty, M. J., Herring, L. E., Wang, W. & Major, M. B. Approximating isotope distributions of biomolecule fragments. *ACS Omega* **3**, 11383–11391 (2018).

[17] O'Brien, J. J. *et al.* Conditional fragment ion probabilities improve database searching for nonmonoisotopic precursors. *J. Proteome Res.* **22**, 334–342 (2023).

[18] Rockwood, A. L., Kushnir, M. M. & Nelson, G. J. Dissociation of individual isotopic peaks: predicting isotopic distributions of product ions in MS n. *J. Am. Soc. Mass Spectrom.* **14**, 311–322 (2003).

[19] Peckner, R. *et al.* Specter: linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics. *Nat. Methods* **15**, 371–378 (2018).

[20] Frejno, M. *et al.* Unifying the analysis of bottom-up proteomics data with CHIMERYS. *Nat. Methods* **22**, 1017–1027 (2025).

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

[22] Bernhardt, O. M. *et al.* Spectronaut: A fast and efficient algorithm for mrm-like processing of data independent acquisition (swath-ms) data. Presented at the 60th ASMS Conference on Mass Spectrometry and Allied Topics, Vancouver, BC, Canada (2012).

[23] Sinitcyn, P. *et al.* MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nat. Biotechnol.* **39**, 1563–1573 (2021).

[24] Yu, F. *et al.* Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. *Nat. Commun.* **14**, 4154 (2023).

[25] Searle, B. C. *et al.* Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nat. Commun.* **11**, 1548 (2020).

[26] Searle, B. C. *et al.* Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nat. Commun.* **9**, 5128 (2018).

[27] Zeng, W.-F. *et al.* AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics. *Nat. Commun.* **13**, 7238 (2022).

[28] Lapin, J., Yan, X. & Dong, Q. UniSpec: Deep learning for predicting the full range of peptide fragment ion series to enhance the proteomics data analysis workflow. *Anal. Chem.* (2024).

[29] Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).

[30] Wang, M. *et al.* Assembling the community-scale discoverable human proteome. *Cell Syst.* **7**, 412–421.e5 (2018).

[31] Wen, B. *et al.* Assessment of false discovery rate control in tandem mass spectrometry analysis using entrapment. *Nat. Methods* **22**, 1454–1463 (2025).

[32] Wilburn, D. B. *et al.* Deep learning from harmonized peptide libraries enables retention time prediction of diverse post translational modifications. *bioRxivorg* (2023).

[33] Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).

[34] Eilers, P. H. C. A perfect smoother. *Anal. Chem.* **75**, 3631–3636 (2003).

[35] Biessy, G. Whittaker–Henderson smoothing revisited: A modern statistical framework for practical use. *Astin Bull.* **56**, 1–31 (2026).

[36] Lautenbacher, L. *et al.* Koina: Democratizing machine learning for proteomics research. *Nat. Commun.* **16**, 9933 (2025).

[37] Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520 (2017).

[38] Zolg, D. P. *et al.* Building proteometools based on a complete synthetic human proteome. *Nature Methods* **14**, 259–262 (2017).

[39] Zolg, D. P. *et al.* Procal: A set of 40 peptide standards for retention time indexing, column performance monitoring, and collision energy calibration. *PROTEOMICS* **17**, 1700263 (2017).

[40] Searle, B. C., Shannon, A. E. & Wilburn, D. B. Scribe: Next generation library searching for DDA experiments. *J. Proteome Res.* **22**, 482–490 (2023).

[41] Wallmann, G. *et al.* AlphaDIA enables DIA transfer learning for feature-free proteomics. *Nat. Biotechnol.* 1–10 (2025).

[42] Rosenberger, G. *et al.* Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat. Methods* **14**, 921–927 (2017).

[43] Petzoldt, J. *et al.* Enabling up to 270 Hz acquisition speed on the Orbitrap Astral Zoom MS (2025). Poster PO313, presented at the 73rd ASMS Conference on Mass Spectrometry and Allied Topics.

[44] Gu, K. *et al.* Improving proteomic identification using narrow isolation windows with zeno SWATH data-independent acquisition. *J. Proteome Res.* **23**, 3484–3495 (2024).

[45] Zubarev, R. A. The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics* **13**, 723–726 (2013).

[46] Tyanova, S. *et al.* The perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).

[47] Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535–9 (2006).

[48] Lim, M. Y., Paulo, J. A. & Gygi, S. P. Evaluating false transfer rates from the match-between-runs algorithm with a two-proteome model. *Journal of Proteome Research* **18**, 4020–4026 (2019).

[49] Pino, L. K., Just, S. C., MacCoss, M. J. & Searle, B. C. Acquiring and analyzing data independent acquisition proteomics experiments without spectrum libraries. *Molecular & Cellular Proteomics* **19**, 1088–1103 (2020).

[50] Wen, B. *et al.* Carafe enables high quality in silico spectral library generation for data-independent acquisition proteomics. *Nature Communications* **16**, 9815 (2025).

[51] Heil, L. R. *et al.* Building spectral libraries from narrow window data independent acquisition mass spectrometry data. *Journal of Proteome Research* **21**, 1382–1391 (2022).

[52] Sokal, R. R. & Rohlf, F. J. *Biometry: Principles and practice of statistics in biological research* 3 edn (W.H. Freeman, New York, NY, 1994).

[53] Keilhauer, E. C., Hein, M. Y. & Mann, M. Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). *Mol. Cell. Proteomics* **14**, 120–135 (2015).

[54] Sowa, M. E., Bennett, E. J., Gygi, S. P. & Harper, J. W. Defining the human deubiquitinating enzyme interaction landscape. *Cell* **138**, 389–403 (2009).

[55] Huttlin, E. L. *et al.* The BioPlex network: A systematic exploration of the human interactome. *Cell* **162**, 425–440 (2015).

[56] Perez-Riverol, Y. *et al.* The PRIDE database at 20 years: 2025 update. *Nucleic Acids Res.* **53**, D543–D553 (2025).

[57] Okuda, S. *et al.* jPOSTrepo: an international standard data repository for proteomes. *Nucleic Acids Res.* **45**, D1107–D1111 (2017).