Holistic Integration of Deep Learning Models for Mass Spectrometry-Based Peptide Identification

by

Kevin Li Yang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
at the University of Michigan
2025

Doctoral Committee:

Professor Alexey I. Nesvizhskii, Chair
Professor Kristina Håkansson, Florida State University
Assistant Professor Todd Charles Hollon
Associate Professor Alla Karnovsky
Associate Professor Arvind Rao

Kevin L. Yang

yangkl@umich.edu

ORCID iD: 0000-0002-8569-4579

© Kevin L. Yang 2025

Acknowledgements

Alexey, thank you for guiding me through the PhD process. I've always been more of a dreamer than a doer, but you helped me organize my jumbled research ideas into tangible products. Fengchao, thank you for being an awesome officemate and answering all my proteomics and programming questions over the years; I can finally say I know how Git works! Danny, thanks for recruiting me to the lab; who knew that an impromptu chalk talk on your phone in the Pretzel Bell basement could have such a lasting impact. To all my lab members, thank you for the scientific and non-scientific discussions and fun social gatherings, including that disastrous Covid-inducing trip to Bill's Beer Garden.

My scientific journey has introduced me to the best mentors I could ask for. Dr. Wynshaw-Boris and Milena, you helped me discover my love for biology, and even though it took forever for me to grasp what a transcription factor is, now I think about proteins so much! Dr. Bowman, thanks for letting me tinker around with that RNA-seq dataset forever and for explaining PCA plots better than anyone else. Anna, thanks for keeping me company in the mouse facility, throughout all those PCR plates, and during my first flight of beers. Dr. Capra and Ling, just one year in the lab showed me bioinformatics was the right path for me.

I'd like to thank my collaborators from the Mathias Wilhelm and Lennart Martens groups for welcoming me into the machine learning community with open arms. Working with you on publications and hackathons and sharing scientific conversations have truly been a pleasure. I've learned a lot from you all.

My friends got me through this rewarding but undeniably trying period of my life. I've enjoyed being with them on countless trips, catching up on phone calls, and partying at Necto Fridays. I want to highlight a few friends. Ford, thanks for keeping the weekly lunch tradition alive; let's go camping again so you can remind me how little I know about camping. Camille, thank you for being so reliable; it means a lot that you drove me to urgent care when I couldn't fall asleep for 72 hours. Nicole, thank you for being a social butterfly and rallying the troops for fun events! Vania and Ashvin, I'm glad we could discover Ninji in Mario Party together. Maria, even though I was down bad while writing this thesis, at least I don't have 300 Anki cards waiting for me.

My family has supported me throughout my life. Steven, the pandemic would have been so boring without you at home. Mama, thanks for being not just the best mom, but also teaching me Unix (but I still don't use Vim). Baba, you taught me that a scientist's greatest attribute is their curiosity. I'm glad you convinced me to be a scientist.

Finally, I am thankful for my funding from the National Institutes of Health grants R01-GM094231 and U24-CA271037, the Proteogenomics of Cancer Training Program 5T32-CA140044-12, and the University of Michigan Rackham Predoctoral Fellowship for financing my recent coffee shop addiction.

Table of Contents

Acknowledgements ii

List of Tables ix

List of Figures x

List of Appendices xiii

Abstract xiv

Chapter 1 Introduction 1

1.1 The role of peptides and proteins 1

1.2 Bottom-up proteomics 3

1.2.1 Mass spectrometry 3

1.2.2 Fragmentation methods 6

1.2.3 Data acquisition 8

1.2.4 Multiplexing approaches: TMT labeling 10

1.2.5 Separation techniques 11

1.2.6 Why bottom-up: top-down vs bottom-up proteomics 12

1.3 Peptide identification 12

1.3.1 Database and spectral library searching 12

1.3.2 False discovery rate control 14

1.4 Open science in proteomics 17

1.5 Machine and deep learning 20

1.6 Motivations for peptide-spectrum match rescoring 21

Chapter 2 MSBooster: Improving Peptide Identification Rates Using Deep Learning-Based Features 23

2.1 Abstract 23
2.2 Introduction 23
2.3 Results 26
2.3.1 MSBooster and the FragPipe computational workflow 26
2.3.2 HLA peptide identification 28
2.3.3 Neoantigen Discovery 32
2.3.4 Direct identification from DIA data 35
2.3.5 Single cell proteomics 37
2.3.6 timsTOF PASEF data with ion mobility separation 39
2.3.7 Multiple correlated features 42
2.4 Discussion 43
2.5 Updates since publication 44
2.6 Methods 46
2.6.1 MSBooster workflow 46
2.6.2 Calculated MS/MS, RT, and IM features 47
2.6.3 Spectral similarity calculation 48
2.6.4 Retention time and ion mobility calibration 49
2.6.5 Kernel density estimation of predicted retention time and ion mobility distributions 50
2.6.6 Determination of the best features 52
2.6.7 MSFragger search and FDR control 53
2.6.8 MaxQuant search and FDR control 54
2.6.9 MaxQuant analysis of Sarkizova et al. 54
2.6.10 Deep learning predictions 55
2.6.11 HLA motif analysis 56
2.6.12 Mass shift RT/IM calibration 56

2.6.13 Statistical analysis and figure generation 57
2.6.14 Hardware 57
2.7 Data availability 57
2.8 Authors, affiliations, and contributions 58
Chapter 3 Accessing and Benchmarking State-of-the-Art Models from the Koina Model Repository in FragPipe 59
3.1 Abstract 59
3.2 Introduction 60
3.3 Results 62
3.3.1 Integration of Koina into FragPipe 62
3.3.2 Benchmarking PSM Rescoring using Koina and MSBooster 64
3.3.2.1 HLA immunopeptidomics 65
3.3.2.2 Phosphoproteomics 70
3.3.2.3 Astral DIA 72
3.3.2.4 TMT 73
3.3.3 A Heuristic Approach to Optimize Model Selection in MSBooster 78
3.4 Discussion 81
3.5 Methods 83
3.5.1 FragPipe analysis and database searches 83
3.5.2 MSBooster speed benchmarks 84
3.5.3 NCE calibration in MSBooster 84
3.5.4 AlphaPeptDeep timsTOF transfer learning 85
3.5.5 Heuristic best model search scoring 86
3.6 Data availability 87

3.7 Code Availability 88
3.8 Authors, affiliations, and contributions 88
Chapter 4 Peptide-Spectrum Match Rescoring with Full Spectrum Prediction 90
4.1 Abstract 90
4.2 Introduction 90
4.3 Results 93
4.3.1 Fragment ion annotation of tryptic and nontryptic peptides 93
4.3.2 Full spectrum prediction models accurately predict nontryptic fragment ion intensities 97
4.3.3 HLA PSM rescoring 101
4.3.4 PredFull predicts unknown fragment ions 104
4.4 Discussion 106
4.5 Methods 108
4.5.1 Koina servers 108
4.5.2 PredFull fragment annotation 109
4.5.3 MSBooster algorithm with auxiliary ions 110
4.5.4 Spectral similarity features 111
4.5.5 FragPipe workflows 112
4.6 Data availability 113
Chapter 5 The Future of Peptide-Spectrum Match Rescoring 114
5.1 Conclusions 114
5.2 Future Directions 117
5.2.1 Beyond Percolator: dealing with missing values 118
5.2.2 Transfer learning 119
5.2.3 Incorporating retention time predictions before PSM rescoring 123
Appendices 124

Bibliography 154

List of Tables

Table 2.1 Mel15 patient-specific peptides 35

Table 3.1 Model prediction limitations 64

Table 3.2 Retention time model benchmarking 65

Table 3.3 MS/MS spectra model benchmarking 66

Table 3.4 RT and MS/MS combined model benchmarking 66

Table 4.1 Differences in relative intensities of fragment ion types in tryptic and nontryptic peptides 96

Table 4.2 Differences in peak counts of fragment ion types in tryptic and nontryptic peptides .. 97

Table 4.3 Peptide IDs from full spectrum models 103

List of Figures

Figure 1.1 Schematics of a timsTOF and Orbitrap-based mass spectrometer 4

Figure 1.2 Fragmentation of the peptide backbone 7

Figure 1.3 Database searching 13

Figure 2.1 MSBooster workflow 27

Figure 2.2 HLA rescoring 29

Figure 2.3 Neoantigen discovery in melanoma tissue 33

Figure 2.4 Melanoma DIA rescoring with MSFragger-DIA 36

Figure 2.5 Single cell rescoring 38

Figure 2.6 timsTOF HeLa rescoring 41

Figure 3.1 Koina integration in FragPipe 63

Figure 3.2 HLA rescoring with various models 68

Figure 3.3 Phosphoproteomics rescoring with various models 71

Figure 3.4 Comparison of TMT models for the LUAD dataset 76

Figure 3.5 Comparison of TMT models for the human-yeast dataset 77

Figure 3.6 Heuristic best model search 80

Figure 4.1 Fragment ion annotation metrics in A*02:01 data 95

Figure 4.2 Fragment ion similarity distributions 99

Figure 4.3 The relationship between b/y and auxiliary ion similarity scores for class I data from Kraemer et al., 2023 101

Figure 4.4 Peptide IDs from full spectrum models on an HLA DIA dataset 102

Figure 4.5 Investigating unannotated predicted peaks 105

Figure 5.1 Transfer learning a TMT model 122

Appendix Figure A.1 Timing of MSBooster steps and Percolator rescoring 124

Appendix Figure A.2 HLA immunopeptidome PSM rescoring 125

Appendix Figure A.3 Feature correlations 126

Appendix Figure A.4 Score distributions for unmodified and modified HLA peptides 127

Appendix Figure A.5 Melanoma DIA rescoring with DIA-Umpire 128

Appendix Figure A.6 Inverse ion mobility dependence 1/K0 on mass and charge 128

Appendix Figure A.7 PSM score distributions for MSBooster features 129

Appendix Figure A.8 Target-decoy PSM distributions for different charge states and features 130

Appendix Figure A.9 IM rescoring of timsTOF HeLa dataset 130

Appendix Figure A.10 Utility of multiple correlated features 131

Appendix Figure A.11 HLA motif with multiple correlated features 132

Appendix Figure A.12 Updated nanoPOTS rescoring 132

Appendix Figure A.13 A separate RT calibration curve for mass offsets 133

Appendix Figure A.14 Retention time calibration for pyro-glu from Q 134

Appendix Figure A.15 Peptides reported when using different MSBooster features 135

Appendix Figure A.16 A plotted calibration curve between the experimental and predicted RT scales 136

Appendix Figure B.1 MSBooster interface and Koina figures in FragPipe 137

Appendix Figure B.2 HCD vs CID PSM score distributions 138

Appendix Figure B.3 Tryptic peptide IDs on a timsTOF 139

Appendix Figure B.4 AlphaPeptDeep transfer learning with HLA-I peptides on the timsTOF 139

Appendix Figure B.5 Quantification accuracy in the human-yeast dataset 140

Appendix Figure B.6 Spectral similarity of unique peptides in the human-yeast dataset 140

Appendix Figure B.7 Feature values for top PSMs for each model tested during the heuristic model search 144

Appendix Figure B.8 Peptides identified using multiple models 146

Appendix Figure C.1 Fragment ion annotation metrics in lung cancer data 147

Appendix Figure C.2 Fragment ion annotation metrics in Atlas data 148

Appendix Figure C.3 The relationship between b/y and auxiliary ion similarity scores for class II data from Marcu et al., 2021 148

Appendix Figure C.4 The relationship between b/y and auxiliary ion similarity scores for class II data from Kraemer et al., 2023 149

Appendix Figure C.5 The relationship between b/y and auxiliary ion similarity scores for class I DIA data from Pak et al., 2021 149

Appendix Figure C.6 The relationship between b/y and auxiliary ion similarity scores for tryptic data from Wang et al., 2019 150

Appendix Figure C.7 The relationship between b/y and auxiliary ion similarity scores for AspN data from Wang et al., 2019 150

Appendix Figure C.8 Identification numbers from full spectrum models (all other datasets) ... 151

List of Appendices

Appendix A: Supplemental Material for Chapter 2 124

Appendix B: Supplemental Material for Chapter 3 137

Appendix C: Supplemental Material for Chapter 4 147

Appendix D: Relevant parameters in command line MSBooster 152

Abstract

Tandem mass spectrometry (MS/MS) coupled with liquid chromatography and/or ion mobility separation is an essential experimental technique for high-throughput discovery of the proteome, or the collection of proteins in a biological system. Computational tools have revolutionized the field, automating the grueling manual annotation of MS/MS spectra with the peptides that produced them. Assigning peptide sequences to spectra is error-prone even for software solutions, requiring statistical methods to limit the false-positive rate. Powered by prodigious public proteomics data repositories, deep learning (DL) models capable of predicting peptide properties have been leveraged to compare experimental observations against predicted values, producing similarity metrics which can increase the sensitivity of peptide-spectrum matches (PSMs). This helps us achieve deeper insights into the proteome.

The contribution of this work is our software MSBooster, an open-source PSM rescoring enhancer fully integrated into the end-to-end proteomics data analysis pipeline FragPipe. Following an introduction to proteomics, mass spectrometry, and machine learning in Chapter 1, each chapter builds on MSBooster.

In Chapter 2, we present the algorithm behind MSBooster; showcase the unique benefits of the tool for immunopeptidomics, data-independent acquisition (DIA), single cell, and trapped ion mobility spectrometry data; and provide expected rates of improvements in peptide identification across the different tasks. We demonstrate that large human leukocyte antigen (HLA) peptide search spaces and chimeric spectra present challenges for database searching, but DL-based rescoring can improve the precision of peptide identification. Notably, we reanalyze a patient-specific melanoma tissue sample, identifying neoantigens for personalized immunotherapy.

We then explore the use of specialized models in specific proteomics niches in Chapter 3, performing a benchmarking study of models available in a DL-democratizing model repository. We identify high-performing models for HLA, phosphoproteomics, tandem mass tag-labelled, and DIA Astral (Thermo Fisher Scientific) experiments. By swapping out the predictor while keeping all other up- and downstream processing steps the same in FragPipe, we isolate the prediction step to directly assess how well each model performs in the PSM rescoring context. To combat potential choice overload arising from the growing list of peptide prediction models, we present a heuristic model selection algorithm that maximizes peptide identifications. Fascinatingly, we uncover a disjunction between the loss functions used to train retention time models and actual identification improvements.

To explore the "dark fragmentome" in Chapter 4, we extract additional information from MS/MS spectra beyond b- and y-ions with full spectrum prediction and showcase how that further improves HLA peptide identification. We also determine the prevalence of "unknown" peaks in full spectrum prediction. Though we found a non-negligible proportion of predicted peaks we were unable to annotate, considering said unknown peaks may actually decrease peptide identification rates.

Chapter 5 summarizes our work and looks ahead to the future of PSM rescoring and DL models in bottom-up proteomics, with a focus on missing score imputation, transfer learning, and other steps in the data analysis pipeline that can benefit from peptide property prediction. Overall, MSBooster provides a practical solution for incorporating chemical knowledge into the assessment of PSM confidence. Its generalizability facilitates its use with diverse data, incorporating DL predictions into the characterization of everything from drug-binding sites to umami peptides.

Chapter 1 Introduction

1.1 The role of peptides and proteins

When French chemist Antoine Fourcroy precipitated gelatin from blood plasma and albumin from egg whites for the first time in the late 18th century, he initiated our centuries-long pursuit of protein knowledge[1,2]. Many proteins are household names: insulin, collagen, antibodies, and – perhaps infamously – the SARS-CoV-2 spike protein. Described as the "agents of action" or "workhorses" of the cell, proteins travel tirelessly and perform their functions throughout our bodies: protein pumps regulate ion concentrations and pH levels across cell membranes, which is essential for all cells and especially the action potentials of neurons and muscle cells[3]; tumor suppressor proteins such as p53 regulate the cell cycle, DNA repair, and ultimately cell death[4,5]; scaffold proteins and protein enzymes catalyze metabolic reactions and ensure the smooth progression of signaling pathways; DNA- and RNA-binding proteins, lipoproteins, and glycoproteins directly interact with all other macromolecules. It is clear that proteins serve a myriad of functions and are essential in virtually every biological process.

While estimates of the number of protein-coding genes in humans fall between 20 and 30 thousand, the number of proteins surpasses this by orders of magnitude[6,7]. Polymorphisms in protein-coding regions of the genome result in alternate "proteoforms" and contribute to phenotypic differences amongst individuals. But proteoforms extend beyond single amino acid variants. Most genes produce multiple proteoforms from alternative splicing of their messenger RNA (mRNA)[8]. Alternative open reading frames within protein-coding genes, long noncoding RNAs, and untranslated regions produce unique protein sequences, many of which are stable

products with known functions[9-11]. Beyond that, proteins are regulated through the actions of chemical post-translational modifications (PTMs) that modulate their functions. One notable example is the reversible addition of phosphate groups to activate kinases. Phosphorylation is just one of more than 400 known PTMs[12]. When we consider the slew of technical chemical artifacts, artificially designed multiplexing labels, and chemical probes that arise from sample preparation in the lab, we achieve an astounding number of possible proteoforms, with some estimates reaching the billions.

Differences in the collection of proteins (known as the "proteome") between individuals results in phenotypic variation. If a protein is either mutated or expressed too much or too little, this can result in disease. Because of ubiquitous protein involvement in cellular processes, many drugs exist and are in active development to therapeutically target them. Drug databases contain information on hundreds of potential druggable protein targets[13,14].

Peptides are shorter polypeptides than proteins, with 50 amino acids being the general consensus cutoff distinguishing the two. Peptides also have known biological functions. Immunopeptides are produced by enzymatic degradation of proteins. They are bound by major histocompatibility (MHC) complexes to be presented to T cells, eliciting an immune response. MHC-bound peptides are classified as class I or class II depending on which type of MHC molecule they bind, likewise designated as class I and II. MHC I and II molecules typically bind peptides derived from intracellular and extracellular proteins, respectively, though there are plenty of exceptions[15,16]. Other peptides take on more active physiological roles. These bioactive peptides can be antimicrobial or antioxidative and serve as templates for synthetic medicines[17,18]. For example, the 51-amino acid long insulin has been engineered to last longer than its naturally occurring template.

1.2 Bottom-up proteomics

Studying the proteome and its complexity is essential for our understanding of biology. It is also insufficient to infer protein levels solely based on transcriptomics data: while mRNA levels can be used to deduce absolute protein abundance within an order of magnitude, mRNA-protein correlations for all genes within a tissue fall well below 0.5 in some tissue studies. Negative mRNA-protein correlations between the same gene across different tissues have been observed, with one study finding that 10% of genes can be negatively correlated with protein levels[19-21]. Besides, transcriptomics does not provide information on the PTM status of proteins. This necessitates high-throughput and accurate methods for protein identification and quantification.

1.2.1 Mass spectrometry

One popular technique for proteome exploration is mass spectrometry (MS), which uses mass spectrometers to measure the mass of peptides in a sample. Mass spectrometers have three main components: first the ion source, followed by a mass analyzer, and finally a detector[22,23]. The basic steps and components of mass spectrometers are illustrated in Figure 1.1. Mass spectrometers detect charged molecules (ions). Therefore, before entering the instrument, peptides must undergo an ionization step. In electrospray ionization (ESI), a high voltage is applied to peptides in liquid solution, forming charged droplets, in the nm – μm size range, as they flow through the tip of a capillary tube[24]. ESI can produce positively or negatively charged ions, but we will focus on positive ESI (ESI+). As the solvent droplets evaporate, they shrink until just the gas-phase, ionized peptides remain, which are then directed into the mass analyzer. ESI is used in tandem with liquid chromatography (LC) columns (covered in Section 1.2.5), due

to the compatibility between the typical LC flow rate and the ESI desolvation process, enhancing sensitivity and analytical throughput.

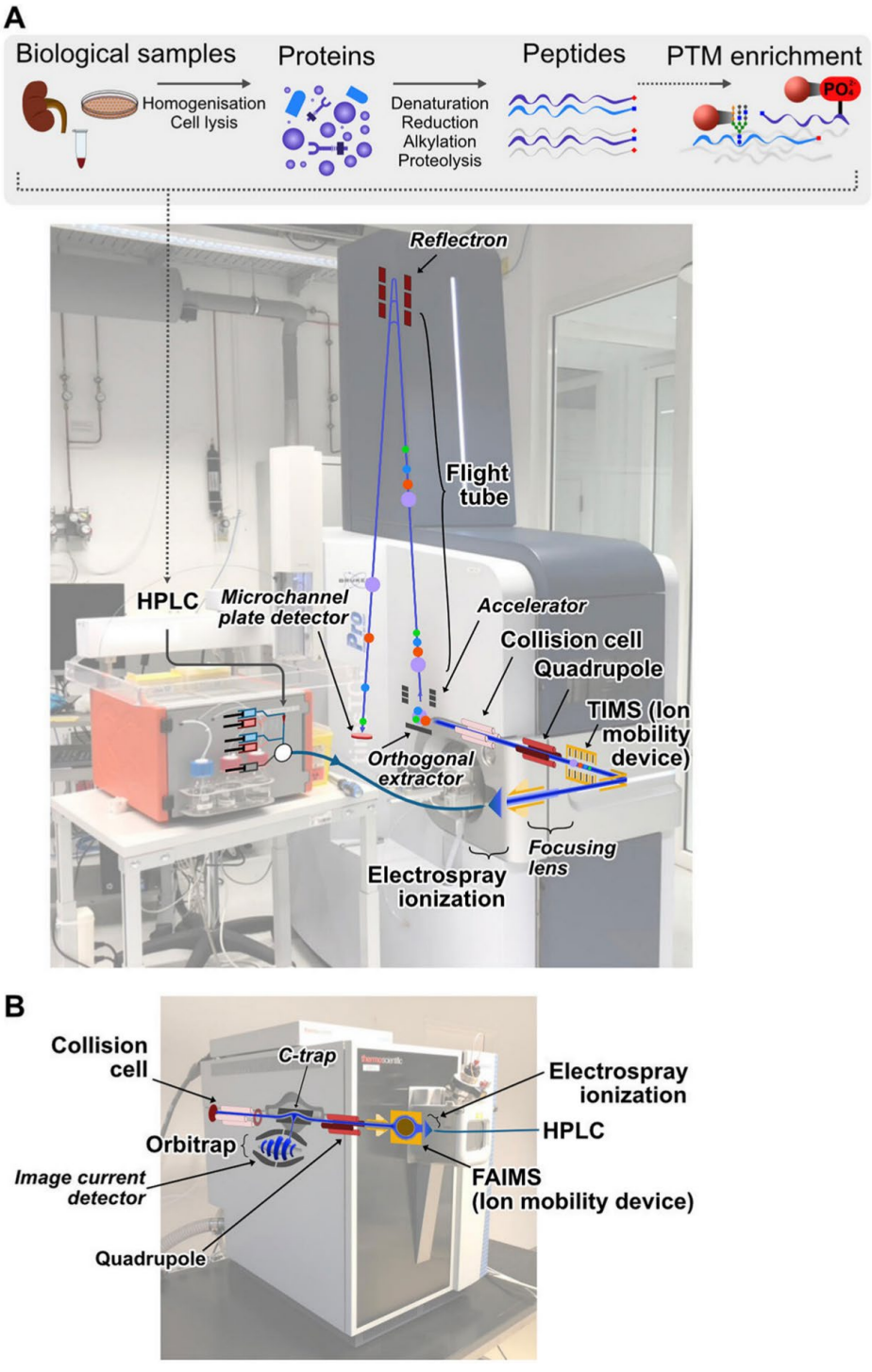


Figure 1.1 Schematics of a timsTOF and Orbitrap-based mass spectrometer

Used with permission of Portland Press, Ltd., from "A beginner's guide to mass spectrometry–based proteomics", *The Biochemist*, Sinha and Mann[22], Volume 42, Copyright © 2020; permission conveyed through Copyright Clearance Center, Inc.

Once inside the mass spectrometer, peptide "precursors" are differentiated by their mass-to-charge (m/z) values. Several mass analyzers exist to separate these distinct precursors, each having distinct resolutions, scanning speeds, and costs[25]. In orbitrap analyzers, precursors in an electrostatic field orbit a central electrode in complex spiraling patterns[26]. An image current detector then records the oscillation frequency between the two halves of the outer electrode. Using Fourier transform, the complex combination of precursor oscillations can be deconvolved into a series of peaks represented in an "MS spectrum". Each peak is described by its precursor m/z and intensity, correlating to the abundance of the precursor in the orbitrap. In a time of flight (TOF) analyzer, equal amounts of kinetic energy are imparted onto selected precursors. Because of each precursor's distinct mass, they have different velocities when travelling down a path until they reach the microchannel plate detector, which releases electrons upon being hit by an ion. This release of electrons is converted into an MS spectrum with peaks defined by their m/z (determined by the ions' time of flight) and abundance/intensity. Thermo Fisher's 2023 release of the new Astral (ASymmetric TRAck Lossless) analyzer combined ideas from the previous generation of analyzers. Precursor ion packets travel up and down a 30 meter-long track, oscillating between asymmetric ion mirrors, one of which is tilted just slightly. Through the course of their 24-26 oscillations between the mirrors, precursor ions of distinct m/z values are separated before reaching a high resolution detector. The Astral improves many aspects of mass spectral data acquisition and quality, including speed of spectrum acquisition, resolution to distinguish close masses, and sensitivity to detect low abundance precursors[27,28].

When combining ionization, analyzers, and detectors, it becomes clearer how peptides can be identified with mass spectrometry: first, peptides must be ionized before entering the

mass spectrometer. Basic amino acids are more likely to hold a positive charge, so for ESI+, proteins are typically digested into peptides containing basic residues. One popular enzyme for this task is trypsin, which cuts peptide bonds to the right ("C-terminal") of lysines and arginines, producing "tryptic" peptides. Once ionized and in the gas phase, they become separated by the mass analyzer before being detected and represented as an MS spectrum. Each MS spectrum represents the collection of precursor m/z values detected by the instrument at a given time. Still, it is impossible at this stage to determine what peptides are present. Shuffling the amino acids around within a peptide sequence would create a peptide with the exact same mass, which even a mass spectrometer with perfect resolution could not distinguish. Furthermore, plenty of tryptic peptides from the human proteome are isobaric or near-isobaric (i.e. having the same or nearly the same mass so as to be indistinguishable at a given mass resolution). Hundreds of precursors can have similar m/z values due to single amino acid variants, PTMs, and isotopes.

1.2.2 Fragmentation methods

To whittle down possible identities for a given precursor, we use a "fragmentation" step, in which the peptide bonds between consecutive pairs of amino acids in a peptide are cleaved. This produces a ladder of peptide fragments which can be detected via a second mass analyzer/detector step and displayed in an "MS2 spectrum," with each peak corresponding to a distinct fragment produced during the fragmentation step. The MS1 and MS2 scans provide a wealth of information to clarify a peptide's identity. Several fragmentation methods have been developed, each utilizing different physical processes to break apart the peptide.

In collision induced dissociation (CID), peptides are fragmented by colliding with a neutral gas (e.g. argon, helium, or nitrogen). Kinetic energy is converted to internal energy during the collisions, breaking covalent bonds throughout the peptide and specifically those

along the peptide backbone. The predominant resulting fragments in CID are b ions if they come from the left hand side of the fragmentation event (N-terminal) or y ions if they are from the right hand side (C-terminal)[29-31] (Figure 1.2).

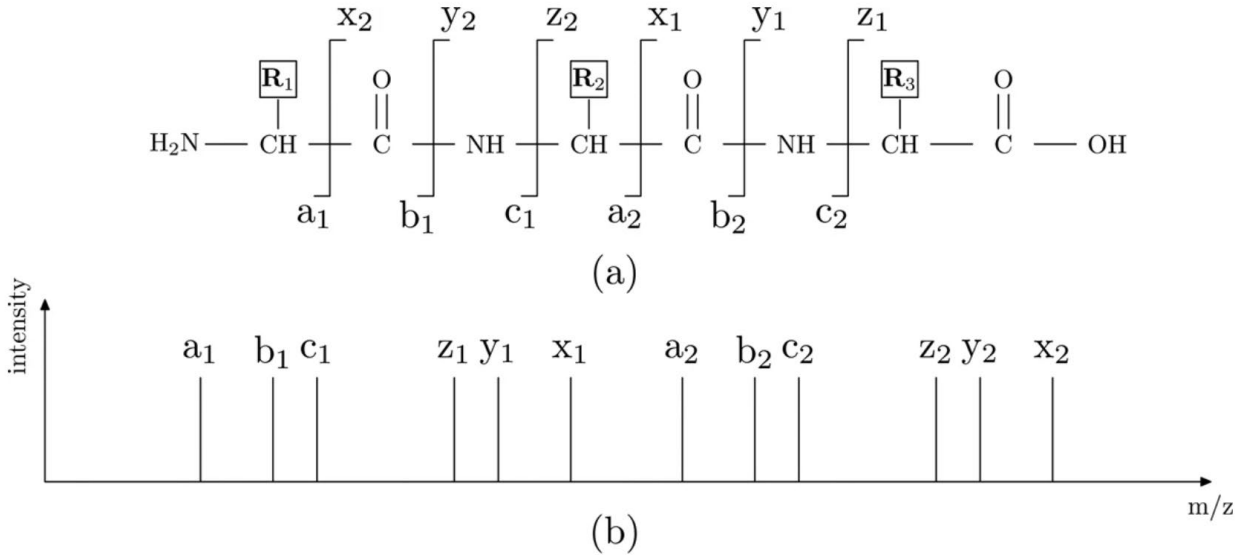


Figure 1.2 Fragmentation of the peptide backbone

Reproduced with permission from Springer Nature, from "Better score function for peptide identification with ETD MS/MS spectra", *BMC Bioinformatics*, Liu et al.[32], Volume 11, Copyright © 2010.

There exist variations of CID. In "resonance excitation" CID, a resonant frequency is applied to the peptide ions in an ion trap; each ion usually fragments only once[33]. "Beam-type" CID methods, such as higher-energy collision dissociation (HCD), quickly impart a large amount of kinetic energy to the precursor ions in a densely packed cell full of neutral gas in a quadrupole TOF or Orbitrap mass analyzer[34]. Because of this quick burst of energy, stronger bonds are more likely broken in HCD as opposed to CID, and multiple fragmentation events may occur. Mechanistic differences of HCD cause pronounced changes in their MS/MS spectra when compared to CID spectra; for example, low-mass b-ions such as the b1 ion may be unstable and fragment into smaller products. One reanalysis of publicly available paired CID-HCD scans showed that their spectra were quite different, with a median Spearman's correlation around 0.5

and many paired spectra having near-zero correlations[33,35]. Though HCD is a variant of CID, we will use the term "CID" to refer exclusively to resonance excitation CID.

Electron transfer dissociation (ETD) functions by the transfer of an electron from an anion (e.g. fluoranthene) to a multiply-charged peptide cation. ETD fragments the peptide backbone at a different location than CID/HCD, producing mainly c and odd-electron zdot ions (Figure 1.3). ETD is a popular choice for highly charged peptides and proteins, as its gentle nature helps to not displace labile modifications such as phosphorylation and glycosylation as in CID/HCD, allowing for accurate localization of these PTMs. A downside of the gentler ETD is its decreased fragmentation efficiency[36,37], complicating peptide identification. To increase fragmentation coverage, ETD can be combined with HCD (EThcD), producing b/y/c/zdot ions and leveraging the benefits of both fragmentation techniques[38-40].

1.2.3 Data acquisition

There are various modes of acquiring data in the mass spectrometer; the choice of acquisition mode affects spectral complexity and quantitative reproducibility. Mass spectrometrists use targeted approaches when they know beforehand which peptides they want to quantify; the number of proteins of interest targeted is typically no more than a few hundred[41]. Untargeted proteomics approaches can be used to dig into the whole proteome, potentially identifying thousands of proteins. We will focus on untargeted methods to identify proteins from a predefined list, such as a reference database of reviewed human protein sequences.

Data dependent acquisition (DDA) has long been the default method for mass spectrum acquisition due to its simplicity. Because precursors continuously enter the mass spectrometer and cannot be held indefinitely, for every MS1 scan, only a subset of all precursor m/z values can be selected for fragmentation before MS2 scans. For example, for every MS1 scan, the top

10 highest intensity precursor peaks could be selected for fragmentation. Because MS1 peak intensity correlates with the precursor abundance, DDA can induce a bias towards quantifying only the most abundant precursors. This can be combated with dynamic exclusion: once a precursor m/z has been selected, it is added to a "blacklist" and is not sequenced again for a set period (e.g. 30 seconds). Other smart precursor ion selection algorithms have been developed to improve upon dynamic exclusion[42]. At the sample processing level, highly abundant proteins can be filtered out before mass spectrometry; such purification steps are necessary in blood serum and cerebrospinal fluid samples to access lower abundance proteins. Without this purification, albumins and immunoglobulins would overshadow other proteins[43,44].

DDA suffers from a stochasticity problem. The top N precursors will vary from sample to sample, leading to missing values, with some datasets experiencing 50% missing values among replicates[45]. Data independent acquisition (DIA) combats this reproducibility problem by selecting all precursors in a larger window for simultaneous fragmentation. Many DIA variants exist, but Sequential Window Acquisition of all THeoretical Mass Spectra (SWATH-MS) illustrates the underlying idea well: for every MS1 scan from 400-1200 m/z, 32 MS2 scans are performed. Each MS2 scan covers a 25 Thomson (Th) window. Windows overlap by 1 Dalton (Da), ensuring that precursor isotopes are captured[46]. By sequencing all precursors in a window, DIA achieves superior reproducibility to DDA[47]. Furthermore, because it does not focus on just the highest abundance precursors, it provides a more complete view of the proteome. Multiple precursors fragment together in DIA, creating complex MS2 spectra. Such spectra require advanced computational algorithms to disentangle. These computational tools are discussed further in Chapter 1.3.

DDA and DIA are not completely distinct from each other and in fact share many ideas and challenges. DDA spectra are simpler than those of DIA but still tend to contain multiple precursors. When the mass spectrometer isolates a narrow m/z window around selected precursors, many precursors of nearly the same mass sneak in with the selected ions. These spectra are known as "chimeric" spectra. Well above 50% of DDA spectra are chimeric, depending on the isolation width[48,49]. Meanwhile, DIA methods with narrow windows have been developed. Narrow window DIA is feasible thanks to the fast scanning speeds of the Astral mass spectrometer[50]. Wide window acquisition is a DDA technique that purposefully expands the isolation window to access more precursors surrounding the top N precursors[51,52]. Improved instrumentation and necessity for deeper proteome coverage (such as in single cell experiments) have driven development of these novel acquisition methods that blur the line between DDA and DIA.

1.2.4 Multiplexing approaches: TMT labeling

Depending on the experimental setup, running a sample in the mass spectrometer can take anywhere from a few minutes to a few hours. Fast approaches are needed if proteomics is ever to be used in clinical diagnostics. Multiplexing is a popular method to sequence multiple samples in the same run. While several commonplace multiplexing reagents exist, we will only address tandem mass tags, or TMT labels[53,54]. In a multiplexed experiment, peptides from each distinct sample are mixed with a different TMT reagent, with the TMT label's reactive group primarily binding the N-termini and lysines of peptides, although TMT overlabeling can occur on other residues. While each TMT label has the same mass, heavy isotopes of carbon and nitrogen are introduced at different locations so that when the TMT-labeled peptides are fragmented, "reporter ion" groups of distinct masses are formed in the low m/z range below 150

m/z. Reporter ions are used for quantification, and MS3 methods can further improve quantitative accuracy[55]. TMT reagents are great for rapid quantitative proteomics, and chemists are pushing the label even further, with one study achieving 35 simultaneous plexes[56].

1.2.5 Separation techniques

Before mass spectrometry analysis, proteomic samples are often separated or "fractionated" to avoid overloading the instrument. Two-dimensional gel electrophoresis (2-DE) was the workhorse of proteomic fractionation for many years. After separating proteins by their isoelectric point (the pH at which the protein has zero net electrical charge) and molecular weight, researchers could isolate a single spot in the gel for enzymatic digestion into peptides before MS sequencing. 2-DE is falling out of fashion because it is low throughput and labor intensive, requiring monitoring of the electric field and pH gradient of the gel. It is also difficult cutting the gels to isolate pure samples of single proteins[57-59].

Liquid chromatography (LC) is now the most used separation technique used before tandem mass spectrometry (LC-MS/MS). The most popular variant is reversed-phase LC (RPLC), in which peptides move through an LC column and exit/elute out of the column at different retention times (RTs) based on their hydrophobicity. A polar mobile phase makes polar peptides elute first, while nonpolar peptides preferentially bind the nonpolar stationary phase (e.g. chains of carbon molecules). The polarity of the mobile phase decreases over time to coax out the more hydrophobic peptides[60]. Two orthogonal forms of LC are sometimes used for greater peptide separation before MS sequencing, leading to cleaner MS spectra[61]. LC gradient timespans can range from 2 hours to 5 minutes, with shorter gradients finding utility in processing clinical samples[62].

Another popular separation technique is ion mobility, first enabled by Bruker Daltonics' line of trapped ion mobility spectrometry TOF (timsTOF) mass spectrometers. In TIMS, precursors are pushed by the opposing forces of a gas and an electric field. The electric field strength gradually lowers until all precursors have exited the TIMS tunnel. The time they exit the tunnel is called their ion mobility (IM) and is inversely proportional to their collisional cross section, a measure of their average surface area. IM provides orthogonal separation to LC and is an increasingly popular tool used with MS and LC for comprehensive sequencing of the proteome[63-65].

1.2.6 Why bottom-up: top-down vs bottom-up proteomics

Throughout this discussion of MS-based proteomics, we have focused on peptide detection, from which proteins can be inferred based on the presence of their theoretical peptides after enzymatic digestion. This is the "bottom-up" approach. "Top-down" proteomics allows for direct detection of intact proteins, which has proven invaluable for protein isoform and PTM localization studies. Bottom-up proteomics is far more popular due to top-down's lower throughput, difficulty of MS spectra interpretation, and issues with intact protein solubility[66,67]. Top-down is well-suited for simple protein mixtures, but bottom-up thrives as an untargeted approach for whole proteome exploration.

1.3 Peptide identification

1.3.1 Database and spectral library searching

Considerable progress has been made in annotation of MS/MS spectra with peptide sequences. "Peptide-spectrum match" (PSM) assignment used to be a manual task, where m/z differences between peaks in the spectra were matched to amino acid masses. Manual validation is still commonplace to gain certainty that a PSM is accurate, but PSM annotation has largely

become a computational task. This is accomplished through a "database search" of an in-silico digestion of a protein database (Figure 1.3).

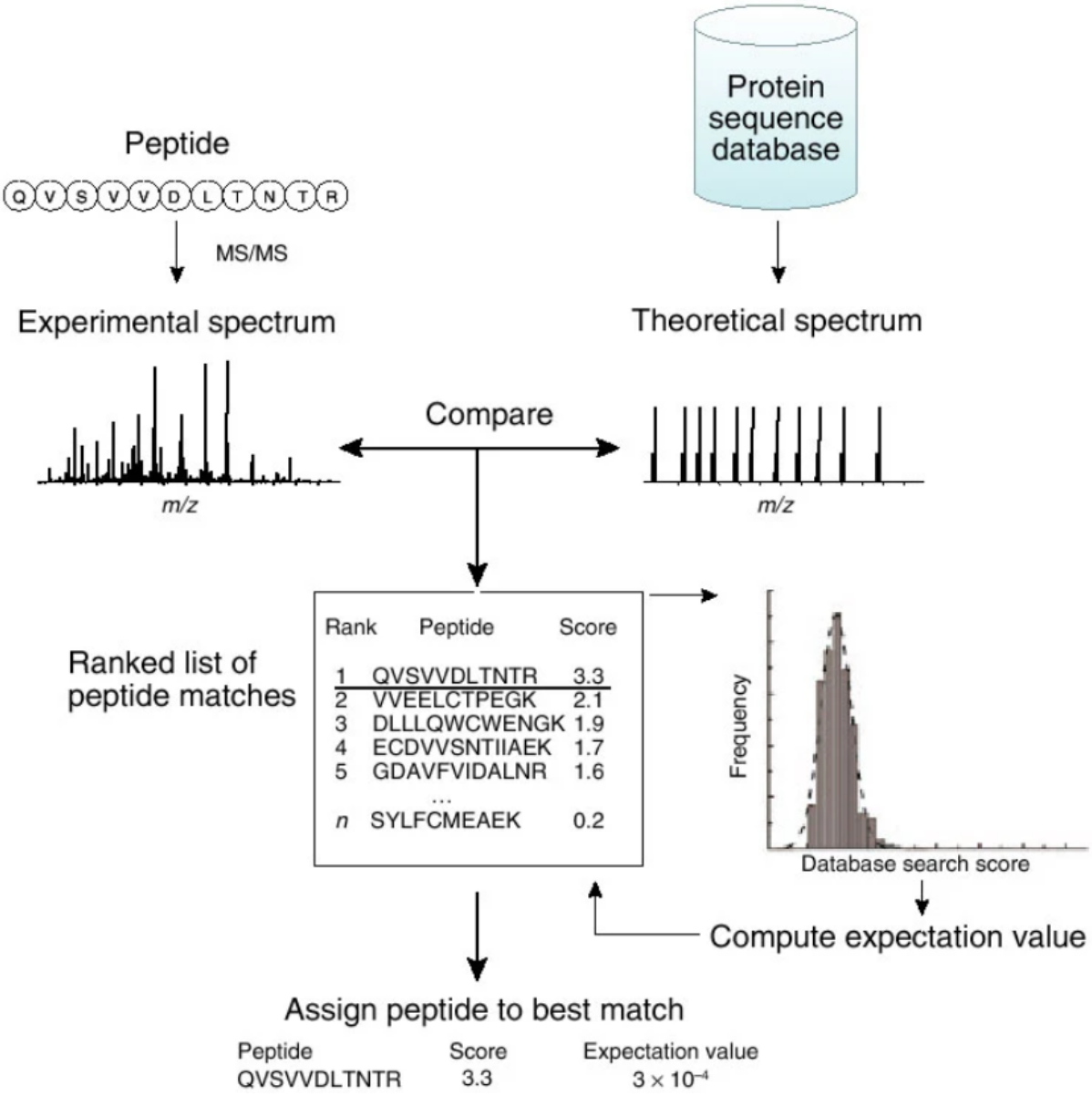


Figure 1.3 Database searching

Reproduced with permission from Springer Nature, from "Analysis and validation of proteomic data generated by tandem mass spectrometry", *Nature Methods*, Nesvizhskii et al.[68], Volume 4, Copyright © 2007.

In a database search, a relevant database of protein sequences (e.g. all canonical proteins from a human proteome) is cut at specific amino acids, following the digestion rules of the selected enzyme. For example, if trypsin was used, each protein sequence will be cut after

lysines and arginines. This produces a collection of peptides that could hypothetically be in the sample. The list of potential candidates for a spectrum is whittled down by filtering it around a narrow m/z range of the detected precursor m/z in the MS1 scan. Each peptide also undergoes in-silico fragmentation, where its theoretical fragment m/z values are calculated, assuming each peptide bond can be broken. Theoretical spectra of all peptide candidates are scored against experimentally acquired MS2 spectra, and the highest scoring peptide is chosen. Many database search tools exist, each with their own scoring functions[49,69-74].

Many database search tools do not consider the relative intensities of matched peaks in their search scores, though a simple visual inspection would reveal that experimental peaks display intensities that differ by orders of magnitude. To better incorporate intensity-based information, "spectral library" search engines compare previously identified peptide spectra against newly acquired spectra and pick the peptide match resulting in the highest score (e.g. dot product)[75-77]. Many organism-specific spectral libraries are readily downloadable, though they may not be the perfect fit if they were generated with instrumentation or parameters different from those of the data to be analyzed. DIA data can similarly be analyzed by spectral libraries created from DDA experiments of the same samples or even from the DIA data itself[78-80]. Spectral library searching comes with the drawbacks of being limited to the smaller search space of whatever peptides are in the spectral library and potentially sacrificing some sample to generate a DDA library for DIA experiments[75,79].

1.3.2 False discovery rate control

Not each MS/MS scan is assigned a PSM; in fact, some experiments have more unannotated spectra than annotated. A database search tool has certain parameters (e.g. minimum theoretical peaks matched, top N intensity peaks to keep, minimum search score to report) that

can influence the annotation rate. It may be that the signal-to-noise ratio of the peaks is too low, or the true peptide has a PTM or other novelty resulting in it not being in the digested protein database. In the best case, these spectra remain unannotated; in the worst case, they are assigned a false identification[81]. Spectral interpretation is ambiguous, as peptides often do not fragment enough to produce the entire expected fragment ion ladder; furthermore, additional fragmentation events such as neutral losses of water or ammonia produce peaks unaccounted for by the search algorithm[82], which may erroneously match theoretical peaks from other peptide candidates.

The peptide search space is so staggeringly large that mistakes in annotation are unavoidable. For example, one can consider a false positive (FP) that shares a substring with the true peptide, yet differs by a few amino acids, resulting in a mass close to the true precursor's. These isobaric and near-isobaric cases are abundant: the phosphorylation and sulfation modifications (79.966Da and 79.957Da, respectively) are difficult to distinguish without high mass resolution. An oxidation modification has the same mass as an alanine to serine or phenylalanine to tyrosine substitution. The true peptide identity may be in the protein database, but because database searches cannot consider all possible PTMs and protein variants, it is likely that some real peptides in the sample are not in the protein database. Rather than trying to assign PSMs to every spectrum, it is more feasible to accept a certain level of FPs. Proteomics searches often set a false discovery rate (FDR) to allow an acceptable rate of FPs. For example, at a PSM-level FDR of 1%, 1% of the final accepted PSMs would be FPs.

To calculate and control the FDR, proteomics algorithms can use "decoys," reversed or shuffled versions of "target" peptides in the database. Any spectra annotated with a decoy is known to be false. We can plot the distribution of database search scores for target PSMs and

compare it to the decoy PSMs' score distribution, and if the decoy generation step was successful, the latter should accurately estimate the distribution of false target PSM scores. This approach assumes that at a given FDR threshold, the number of decoy PSMs exceeding this threshold matches the number of false target PSMs; in other words, decoy PSMs approximate FP target PSMs.

In practice, the target-decoy approach can be implemented through a concatenated search or separate database searches[83,84]. In the concatenated approach, one database containing all target and decoy peptides is searched at once, and the highest-scoring peptide is assigned. In direct target-decoy competition, each spectrum is searched twice, once against a target database, once against a decoy database. The highest scoring of the two peptides is assigned.

Database search scores are no longer used as the exclusive variables in determining FDR cutoffs. Other factors, such as the error between the measured precursor mass and peptide candidate mass or the difference between the first and second best PSM scores, can further improve peptide identification by providing orthogonal information. PeptideProphet is a tool that uses linear discriminant analysis and the expectation maximization algorithm to distinguish between true and false PSMs[85,86]. Percolator uses a semi-supervised approach and linear SVM to learn a hyperplane between true and false PSMs using an arbitrary number of PSM features[87], and a newly added procedure helps to rigorously control its FDR[88].

FDR control is critically important. If underestimated, false biological conclusions may be drawn. Because FDR calculation differs among software, improper FDR control can bias certain algorithms, making them seem better when in reality, they are reporting more false identifications (IDs). An "entrapment" approach can be used for increased confidence in the FDR control[89-91]. In an entrapment experiment, entrapment sequences are generated by shuffling

the sequences in the original peptide database or by concatenating sequences from another organism; decoy entrapment sequences are also generated. The entrapment database can be ten times as large as the original. Entrapment sequences are treated as targets and reported in the final filtered list of IDs. By using a well-controlled entrapment search, Wen et al. found that many software tools only provided a lower-bound FDR estimate, where an upper-bound estimate is desired[92].

FDR control can be extended to the peptide and protein level as well, enabled by tools that incorporate these multiple levels of statistical control[93]. Protein-level FDR is not always used, such as for studies of the HLA peptidome or modified peptide analysis[94]. If protein-level identification is desired, accepted peptides can be assigned to proteins; this is known as "protein inference." Various statistical and deep learning approaches have been proposed[95]. One simple approach to protein inference is based on "Occam's Razor," which dictates that the simplest solution is the correct one; in this case, the minimum set of proteins that can account for all peptides identified is the most likely solution. Greater stringency can be applied by requiring at least two or more peptides to match a protein to report that protein.

1.4 Open science in proteomics

Many advancements in biomedical research have been enabled by data sharing. Not only does it provide transparency in the peer review process, but it also is core to algorithmic development and testing. One popular framework used to guide open science is the FAIR Guiding Principles of findability, accessibility, interoperability, and reusability[96]. These guidelines have been embraced by the proteomics community, with much ongoing work to make data-sharing robust and easy to follow[97]. To fulfill these requirements, data must be accessible and easily searchable via a unique identifier, and they must include descriptive metadata[98-100].

These mandates address the plague of "link rot," in which links to data, references, or websites become dead or point to the incorrect resource as maintenance falls to the wayside[101].

A plethora of well-maintained data repositories exists for the dissemination of proteomic data generated in labs across the globe. Not only do these sources support sharing of the raw data files generated by MS instruments, but they also encourage sharing of intermediate data analysis files and final PSM annotations. Many of these databases are members of the ProteomeXchange consortium, which was created to harmonize independent proteomics data repositories and summarize the data at scale[102-105]. ProteomeXchange provides a central location for researchers to search and download the hundreds of new datasets made available each month from any of its member databases. To name a few of the consortium members, PeptideAtlas and MassIVE (https://massive.ucsd.edu/) streamline the process of reanalyzing spectra, which has been a monumental help to the Human Proteome Project's goal of providing evidence for the existence of all proteins predicted from human genes[106,107]. MassIVE.quant provides an infrastructure specific to storing quantitative proteomics data[108].

Integration of these resources requires a common language to be used in the metadata files. Spearheaded by the Proteomics Standards Initiative, the sample and data relationship format (SDRF) has gained traction in the proteomics community by standardizing the ontological terms used in metadata descriptions[109-111]. Protein and peptide sequence representation is standardized through ProForma notation[112]. The Universal Spectrum Identifier (USI) generates a code for each mass spectra for unique identification and quick searching through the USI service in the Proteomics IDs Database (PRIDE)[113].

This wealth of publicly available proteomics knowledge also enables the creation of databases which summarize the knowledge and integrate it with other information. UniProt,

which maintains collections of reviewed and unreviewed proteins across species, provides a resource from which to download up-to-date protein fasta files for database searching; it integrates many sources of information into one place: each protein can be associated with its function, known variants, PTMs, and 3D structure, amongst other information[114-116]. ProteomicsDB connects proteomics data with drug response data and other types of omics measurements across organisms; it allows users to upload their own expression data temporarily, enabling the use of analytical tools such as drug sensitivity prediction and heatmap clustering[117,118].

Finally, the computational proteomics community has curated specific datasets with which new analysis tools can be tested. These datasets provide a powerful ground truth and relieve the overhead of having to generate new data for every study. These benchmarking datasets have helped software developers compare their new tools with existing ones, with a particular focus on quantitative accuracy[119-121]. This is achieved through measuring out specific dilution levels of spike-in proteins against a constant background. Proteomics software comparison has been semi-automated with ProteoBench (https://proteobench.readthedocs.io/en/v0.2.6/) which provides input files from diverse datasets. Developers can download the files for data processing. After uploading their results, ProteoBench applies the same parameters (e.g. minimum ions for protein quantification) to all previously submitted results, allowing for fair comparison. Users can even upload results from the same software with different parameters, and the website's data visualization of multiple metrics makes it clear how the tools stack up. This is just one of the more recent examples of how community-driven efforts for open science in proteomics have benefited algorithm development.

1.5 Machine and deep learning

The fields of machine and deep learning (ML and DL, respectively) provide us with powerful techniques for prediction and classification tasks. The recent publication of the AlphaFold model for protein structure prediction highlights that the power of artificial intelligence extends to the protein domain[122]. The success of ML/DL models in proteomics is not too surprising: words are to sentences as amino acids are to peptides; as such, many natural language processing techniques and neural network architectures are adaptable to protein sequences. RNNs, LSTMs, and transformers, which excel at handling sequential data, extrapolate well to protein sequences[123]. Another requirement for ML to be successful is that the data must follow some sort of pattern, and this is the case in proteomics. Previous efforts to encode amino acids into latent spaces (multidimensional spaces that represent data in an abstract way) have found that they cluster in biologically intuitive ways corresponding to their physical properties (positively/negatively charged, hydrophobic, etc.)[124]. Furthermore, MS2 spectra, RT, and IM values from the same peptide are highly similar when identified multiple times through the same experiment or lab[125]; this is a foundational principle of spectral libraries, where IDs from DDA experiments help inform where to look for the same peptides in DIA experiments. Even across different instruments, LC gradients, and experimental setups, peptides repeatedly exhibit similar properties. Clustering of peptide spectra across entire repositories storing terabytes of data show that those produced by the same peptide clump together[126,127].

Peptides are predictable, so naturally many models have been trained to predict their properties, including MS2 spectra, RT, IM, ionization efficiency, strength of MHC binding[128], and many others. Earlier studies using ML manually calculated these properties as functions of the peptide sequence, analogous to how the first predictions of protein folding and structure explicitly encoded the process based on the physics of molecular modelling[129]: the “mobile

proton model" dictates where a proton is likely to spend its time on a peptide, influencing what bonds are most often broken[130]. First-generation ML models tabulated features of peptides and their subsequences to generate spectral predictions[131-133]. RT and IM are also determined by similar inherent physicochemical properties. Sophisticated DL architectures have since shown increased accuracy without having to manually calculate these features, instead letting the models learn their own rules, albeit resulting in less intuitive "black box" models[134-140].

Proteomics data repositories provide a goldmine of data for training these models. ML/DL scientists often use stringent quality filters to retain only the best data where they can be confident in the accuracy of the selected peptide/protein IDs, and synthetic datasets with known ground truth labels help increase the amount of high quality data[134,141]. This wealth of data combined with rapidly progressing DL architectures helps the models approach the theoretical upper limit of accuracy, as determined by the consistency of repeated peptide measurements within experiments[125,142].

1.6 Motivations for peptide-spectrum match rescoring

Deep learning predictions can serve as approximations for the values of properties we expect peptides to exhibit. By comparing predicted values with the observed experimental values, we can calculate similarity metrics. Tools such as Percolator can combine these new features with database search scores and other basic PSM descriptors to generate new functions, which can be used to "rescore" the PSMs to better separate true and false positives. PSM rescoring is a promising supplement to database searches, adding the discriminative power of ML/DL predictions without further complicating the database search step. It also provides more flexibility than spectral library approaches that are limited to only previously identified peptides.

Rescoring pipelines have been extensively reviewed[143-145]. They differ in aspects such as types of features calculated, but another important distinction from a user perspective is how they fit in with the rest of their respective analysis pipelines. Rescoring tools distinguish themselves primarily by their design goals, which users should keep in mind when choosing a platform: some value supporting multiple search engines with their distinct data formats[146,147]. Others are optimized for specific types of data, such as immunopeptidomics data where novel noncanonical peptides require strict FDR control[148-151]. Comfortability with computational interfaces also influences the choice of rescoring pipeline, where some users like to interact with intermediate stages of data analysis via Jupyter Notebooks[147], while others primarily value the final identification and quantification results; software setup plays an important role, where some users are comfortable with Anaconda environments, Nextflow, and Docker setup, while others prefer a simple installation script and graphical user interface[146,152]. Finally, an important consideration is the choice of open-source versus vendor-specific solutions[49,153-155].

Chapter 2 MSBooster: Improving Peptide Identification Rates Using Deep Learning-Based Features

This chapter has been adapted from its homonymous publication in Nature Communications[156].

2.1 Abstract

Peptide identification in liquid chromatography-tandem mass spectrometry (LC-MS/MS) experiments relies on computational algorithms for matching acquired MS/MS spectra against candidate peptides using database search tools, such as MSFragger. Here, we present a new tool, MSBooster, for rescoring peptide-to-spectrum matches using additional features incorporating deep learning-based predictions of peptide properties, such as LC retention time, ion mobility, and MS/MS spectra. We demonstrate the utility of MSBooster, in tandem with MSFragger and Percolator, in several different workflows, including nonspecific searches (immunopeptidomics), direct identification of peptides from data independent acquisition data, single-cell proteomics, and data generated on an ion mobility separation-enabled timsTOF MS platform. MSBooster is fast, robust, and fully integrated into the widely used FragPipe computational platform.

2.2 Introduction

Proteomics database search engines output a list of peptides that best match an MS/MS spectrum. The quality of these matches are used as input to computational post-processing tools such as PeptideProphet[85,86,157] and Percolator[87,158]. These tools combine various search engine scores, such as the hyperscore and expectation value, with other properties that are useful for discrimination, such as the difference between the theoretical mass of the peptide and the mass

derived from the measured m/z. The distributional differences of scores for decoy and target peptides are used in the modeling process to determine the optimal combination of individual features, as well as to calculate posterior probabilities of correct identification and estimate FDR. These tools significantly boost the sensitivity of peptide and protein identification at a fixed FDR compared with filtering the data using individual scores reported by the search engine[157].

Although post-processing tools are in many computational pipelines, including FragPipe, they do not incorporate prior knowledge regarding peptide separation coordinates (RT and IM) or fragment ion intensities. High confidence PSMs from previously published studies are stored in public repositories and can be leveraged via spectral library searching[77,141,159-162], in which known fragment ion intensities help differentiate true from false PSMs. However, relying on experimentally derived spectral libraries is often limiting, as these libraries are inherently incomplete. For instance, protein expression varies from biological condition to condition, cell type to cell type, and genetic background to background, so that libraries can be incomplete even for organisms with large amounts of previous MS/MS data available. Thus, approaches for predicting MS/MS spectra[131] and using predicted spectra from available protein sequence data to improve the sensitivity of peptide identification in LC-MS/MS proteomics have been explored[163-165]. The difference between the experimental and predicted RTs is also known to provide additional discriminating power[166-169]; RT differences have previously been incorporated into Percolator and PeptideProphet modeling[169,170]. However, the use of RT and MS/MS spectral predictions was initially limited, in part because of the limitations of first-generation prediction algorithms.

More recently, a wave of DL models has been trained to predict the physicochemical properties of peptides and MS/MS spectra[133,134,171-174]. By training on millions of available

peptides, these models can learn general rules to make accurate predictions for new peptides, assuming they are not vastly different from those on which the models were trained. The use of DL-based RT and spectral predictions has been shown to be particularly useful for DIA data analysis[175-177] and for improving the identification rates in immunopeptidome studies concerned with the analysis of HLA binding peptides[153,178-180]. Unfortunately, current PSM rescoring tools that take advantage of DL-based predictions may be difficult for some users to adopt. For example, MaxQuant with Prosit rescoring requires users to upload their database search results to a web server. Rescoring may be performed locally if the users have GPU access, which is not always the case. DeepRescore[150] requires Docker and Nextflow, which may be difficult for users with less computational experience to install.

Here, we present the DL-based PSM rescoring tool MSBooster, a new addition to the widely used FragPipe computational platform. MSBooster provides a fully automated and integrated solution for the use of DL predictions for improved peptide and protein identification. It uses a DL model to predict the RT, IM and MS/MS spectra of peptides, followed by the generation of additional features for PSM rescoring with Percolator[87]. No external prediction of spectral libraries is required, bypassing concerns about uploading data to shared servers and data privacy. We demonstrate the flexibility of MSBooster and its performance in several different workflows, including HLA immunopeptidome nonspecific searches, DIA quantitative proteomics, single-cell proteomics, and IM-enabled timsTOF mass spectrometry. We also explored the behavior of spectral and RT features in the analysis of single cell proteomics data and investigated the potential benefits of using multiple correlated similarity metrics in Percolator. Finally, we assess and discuss the utility of incorporating IM predictions into PSM rescoring.

2.3 Results

2.3.1 MSBooster and the FragPipe computational workflow

FragPipe (https://fragpipe.nesvilab.org/) is a comprehensive computational platform that automates all steps of proteomic analysis, including peptide-spectrum matching with MSFragger[69,181,182], PSM validation with PeptideProphet[85] or Percolator[87], protein inference with ProteinProphet[183], and FDR filtering (by default 1% FDR at the PSM, ion, peptide, and protein levels) using Philosopher[93]. FragPipe supports the generation of spectral libraries using EasyPQP (https://github.com/grosenberger/easypqp) and the extraction of quantification from DIA data (using DIA-NN[175,176]). DIA-Umpire[184] is included in FragPipe as one of the modules to generate pseudo-MS/MS spectra from the DIA data. Alternatively, peptides can be identified directly from DIA data using MSFragger-DIA[185]. FragPipe has an easy-to-use graphical user interface and includes a data visualization module (FragPipe-PDV), which is an extension of a previously described PDV viewer[186].

Within FragPipe, MSBooster is positioned between MSFragger and Percolator (Figure 2.1) and is enabled by default in most FragPipe analysis workflows (Methods Section 2.6.7), where a FragPipe workflow is the order in which software are to be executed, along with optimized parameters for each tool. MSBooster's role can be divided into the separate steps of peptide extraction from PSM files, input file formatting for a DL model, feature calculation using observed and predicted peptide properties, and addition of the new features to the PSM files. In a typical workflow, MSFragger performs the database search, and reports the list of PSMs and associated search scores in a "pin" file. Without MSBooster, these pin files are passed directly to Percolator (Figure 2.1a). When MSBooster is enabled, it extracts the set of peptides reported in the pin file and creates an input file for a DL model, which in turn generates predictions of the

physicochemical properties of peptides, namely RT, IM, and/or MS/MS spectra (Figure 2.1b). Within FragPipe, we chose DIA-NN[175,176] to predict these properties, as it is already included for DIA quantification. We also show compatibility of Prosit predictions with standalone MSBooster in Results Section 2.3.2. Importantly, predictions are performed only for the relatively small set of PSM candidate peptides reported by MSFragger, rather than the whole in-silico digested proteome. Thus, predictions can be made for each dataset on-the-fly, without the need for time-consuming full spectral library prediction.

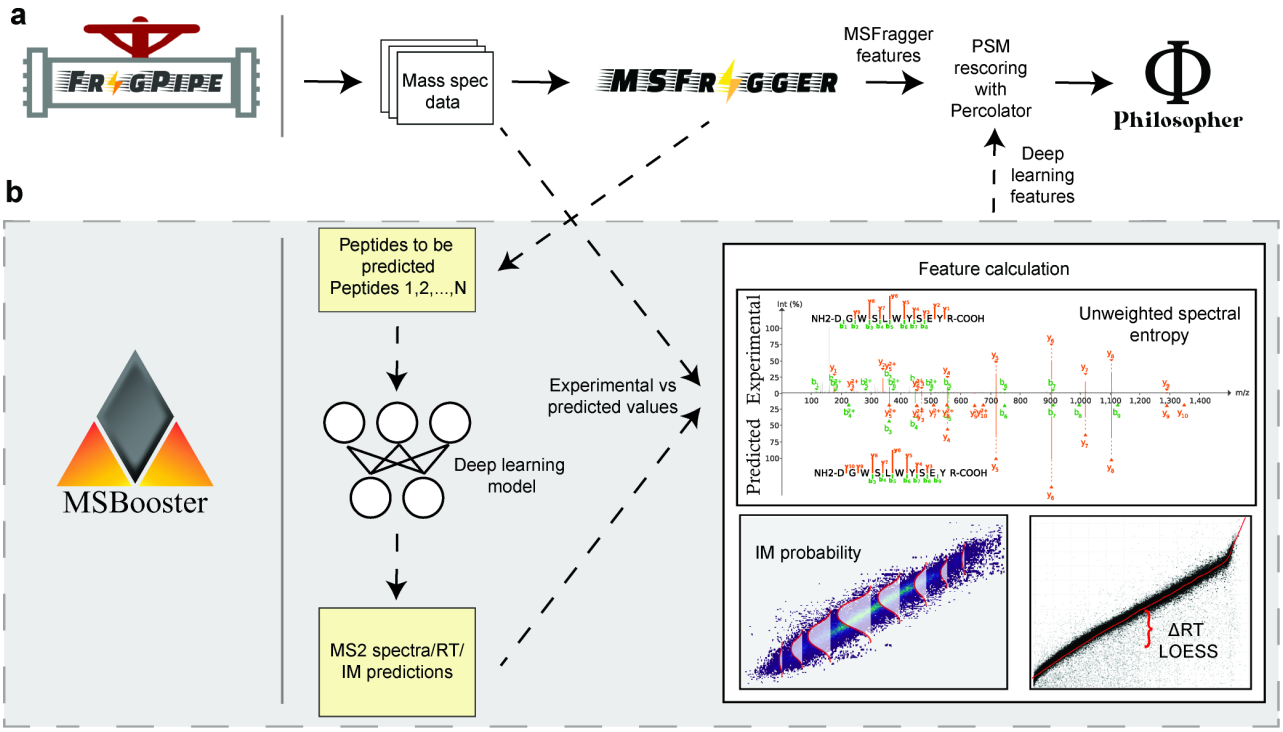


Figure 2.1 MSBooster workflow

The original workflow without MSBooster (a) and new workflow with MSBooster (b) are depicted. Files generated when running MSBooster are depicted in yellow. Dashed arrows are steps run only when using MSBooster. The default features used by MSBooster are shown in "Feature calculation". Features reported by MSFragger, such as hyperscore and charge, are combined with deep learning features calculated in MSBooster for Percolator rescoring before filtering and reporting in Philosopher.

MSBooster then generates features based on the agreement between the experimental and predicted values and adds them to the original pin files. Finally, it passes these extended pin files to Percolator, which learns a linear support vector machine (SVM) to differentiate true targets from decoys[87]. Percolator assigns an SVM score and then a posterior error probability to each PSM. DL-based predictions are done for a limited number of peptide candidates: by default, either for a single (top scoring) peptide per MS/MS spectrum (DDA data), top 3 (narrow window GFP-DIA data) or top 5 (conventional DIA data) when using MSFragger-DIA[185]. MSBooster resulted in only a minor increase in the overall computational run time (Appendix Figure A.1).

2.3.2 HLA peptide identification

Immunopeptidomics studies identify and quantify peptides that are presented by antigen presenting cells. These studies are increasingly required in biomedicine but are associated with computational challenges. Human leukocyte antigen (HLA) peptidome data is a promising candidate for DL-based rescoring due to an expanded nonenzymatic search space, resulting in a higher probability of a high-scoring false match. Because certain major histocompatibility complexes (MHCs) preferentially bind certain peptide motifs, this represents a system in which we know what kinds of peptides should be identified based on their sequences, allowing us to monitor if MSBooster correctly promotes true target PSMs. To demonstrate the performance of MSBooster on HLA peptides, three fractions of an A*02:01 monoallelic cell line[187] were processed using different combinations of features in MSBooster. MS/MS data was acquired on an Orbitrap Exploris 480 (Thermo Fisher Scientific) with higher energy collisional dissociation (HCD). Spectral and RT features increased the number of peptide IDs by 20.4% and 16.6% respectively at 1% FDR, whereas the combination of the two feature types increased the number of IDs by 31.4% (Figure 2.2a).

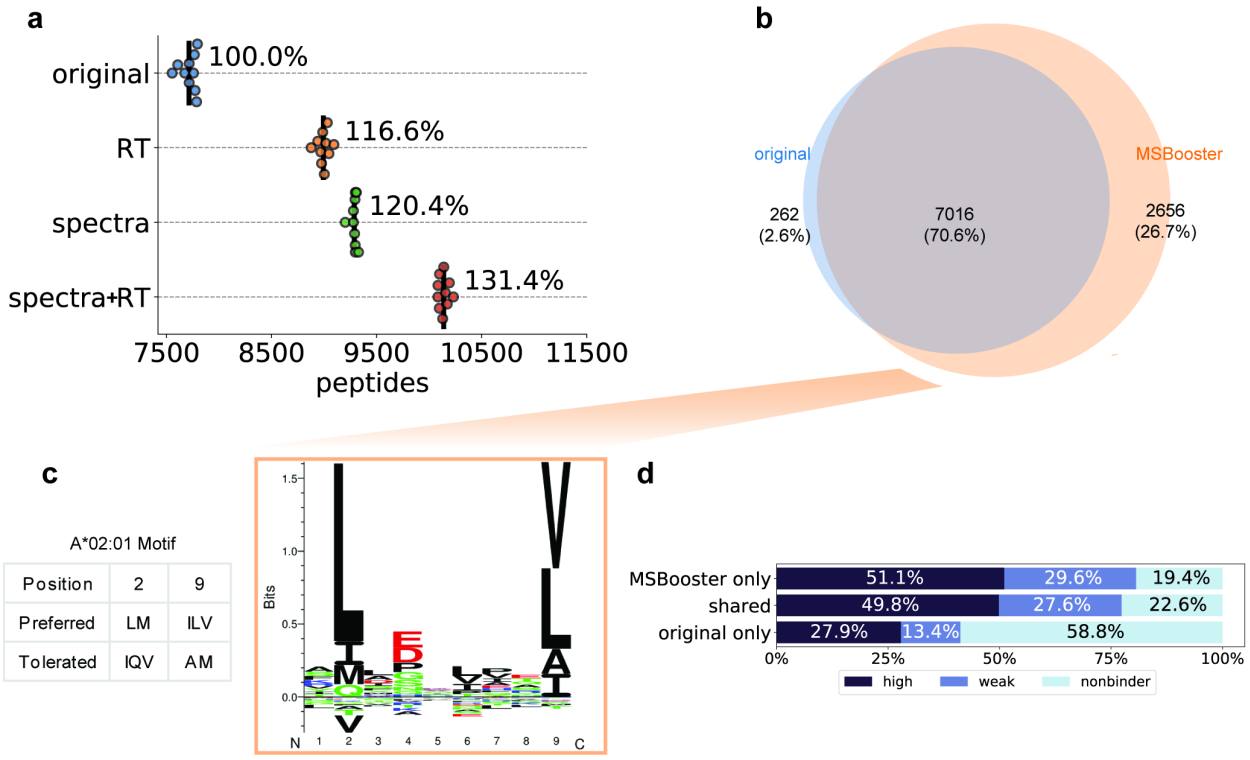


Figure 2.2 HLA rescoring

(a) Swarmplot of the number of HLA peptides reported at 1% FDR using the MSFragger pin files (original), files with the spectral similarity feature added (spectra), retention time similarity feature (RT), or both types of features (spectra+RT). Each dot represents the number reported for each of 10 Percolator runs. Black lines show the average number of peptides reported across 10 Percolator runs. (b) Venn diagram of HLA peptides between lengths 7 and 12 when using either original MSFragger features or additional deep learning features. (c) GibbsCluster-generated motif assigned to the MSBooster-specific peptide subset from (b). The A*02:01 motif was collected from the Immune Epitope Database. (d) Percent of peptides from each subset of (b) that are predicted by NetMHC 4.0 to bind the A*02:01 serotype. Strength of the ligand binding decreases from "high" to "weak" to "nonbinder".

HLA peptide rescoring with DL features has previously been explored[179]. The authors showed an average increase of 159% in peptide identification across 92 monoallelic cell lines[188] when using MaxQuant coupled with Prosit rescoring (Appendix Figure A.2a). While this may make the 31.4% increase with MSBooster on the A*02:01 data seem minimal, this discrepancy may simply reflect the moderate performance of MaxQuant in nonspecific searches[189] and its use of only the Andromeda score for ranking PSMs before FDR filtering[70]. Even without DL-based rescoring, MSFragger provides multiple discriminative scores (Appendix Figure A.3a). To provide a more accurate comparison, we rescored PSMs using only the hyperscore –

MSFragger's database search score – as a starting point (Appendix Figure A.2b). Giving Percolator access to only to the hyperscore, we reported an average of 2949 peptides. Adding spectral and RT features to the hyperscore provided a 183.8% increase in the number of identified peptides, a statistic more in line with the 159% increase previously reported[179]. Importantly, adding other features reported by MSFragger (Appendix Figure A.3a) also gives a 161.5% boost compared to using MSFragger's hyperscore alone, indicating the utility of non-DL features. Using all MSFragger computed and all DL-based from MSBooster features together results in a 243.7% boost compared with using hyperscore alone.

We then specifically compared performance between MSFragger/MSBooster and MaxQuant/Prosit on the A*02:01 data. MaxQuant initially reported 1569 peptides. After PSM rescoring with Prosit, it reported 10680 peptides, a 681% increase. Compare this to the 10138 peptides reported with MSBooster rescoring averaged across ten Percolator runs (Appendix Figure A.2b). To see if MaxQuant's performance could be attributed to Prosit's predictions, we configured MSBooster to accept Prosit-predicted spectral libraries (Methods Section 2.6.10). Indeed, using Prosit in lieu of DIA-NN allowed MSBooster to achieve 10798 peptides on average, 118 more than MaxQuant with Prosit (Appendix Figure A.2b).

While most reported peptides passing 1% peptide-level FDR were shared regardless of whether MSBooster was used (7016 peptides), adding MSBooster resulted in 2656 more identified peptides while only losing 262 (Figure 2.2b). To verify that the added peptides were credible, we identified their HLA sequence motifs using GibbsCluster[190] (Figure 2.2c, Appendix Figure A.2c). MHC binding in the A*02:01 cell line relies on anchors at position 2 and the C-terminus, according to the Immune Epitope Database[191]. The 262 peptides lost after rescoring produced two clusters in GibbsCluster. The first cluster of 132 peptides followed the expected

motif, but the second cluster of 60 peptides was not enriched for the expected amino acids at position 2 (Appendix Figure A.2c). Therefore, many of the peptides removed with the help of MSBooster were likely false positives. In contrast, peptides gained with MSBooster generated one cluster of 2533 peptides that faithfully followed the expected sequence motif for the cell line (Figure 2.2c, Appendix Figure A.2c). To further validate the new peptides, we examined their binding affinities with A*02:01 MHC using predictions from NetMHC[128] (Figure 2.2d). We found that 2143 (80.7%) of the gained peptides and 5430 (77.4%) of the shared peptides were predicted as either strong or weak binders by NetMHC, while this percentage dropped to 41.3% (108 peptides) for the MSBooster-removed peptides. This further supports the idea that peptides gained with DL-based rescoring are more reliable than those that are removed.

An important feature of MSBooster is its ability to handle peptides with post-translational modifications (PTMs) that are not predicted by the DL spectral prediction model. In DIA-NN v1.8, cysteine carbamidomethylation, methionine oxidation, N-terminal acetylation, phosphorylation, and ubiquitination are supported. A multitude of other biologically relevant PTMs exist; for example, cysteinylation is an important PTM to consider in immunopeptidomics, as it plays a role in T cell recognition[192]. Rather than precluding the inclusion of other PTMs in the search or rescoring steps, MSBooster obtains the predicted spectrum for the unmodified peptide and shifts the m/z values of the PTM-containing fragments while retaining their predicted intensities. RT values are the same as those of their peptide counterparts, excluding the new PTMs (e.g. a jointly biotinylated and phosphorylated peptide will use the RT of the phosphorylated peptide). To explore how fragment peak shifting affects the results, we examined the distributions of spectral and RT feature scores for accepted PSMs after Philosopher filtering (Appendix Figure A.4a-b). Each group of PSMs contained only the PTM listed (i.e. the PSMs in

the carbamidomethylated C group were matched to peptides that only contained that PTM, and no oxidized M). As expected, unmodified, carbamidomethylated C, and oxidized M PSMs had high spectral similarities and low RT differences, since DIA-NN included them in the training set. Interestingly, although acetylated N-terminal peptides were in the training set, their spectral similarity score distributions were lower than those of the other peptides. Cysteinylation had a similar distribution to that of PTMs on which DIA-NN was trained. This could mean that cysteinylation does not have a major impact on fragment intensity or that cysteinylated PSMs with lower scores were excluded after FDR filtering. PSMs with pyro-glutamation events from Q had the worst distribution of the PTMs considered. The RT shift in pyro-Glu peptides is expected and has been previously recapitulated[193]. While most PTMs showed an increase in peptides containing those modifications (Appendix Figure A.4c), the exception is pyro-glutamation from Q, where reported peptides dropped from 105 to 25. This decrease is mainly driven by the increased RT difference, as no unique peptides are reported after only rescoring with the RT feature, while rescoring with only spectral similarity results in only 32 lost peptides and 9 gained peptides. This is in stark contrast to the 82 lost and 2 gained peptides after using the MSBooster default of spectral and RT feature rescoring. Overall, our analysis shows that while MSBooster will not exclude any PTM-containing peptides, those PTMs not yet supported by the DL prediction module that drastically affect the peptides' properties will be heavily penalized.

2.3.3 Neoantigen Discovery

The discovery of patient-specific neoantigens, such as those derived from genomic alterations in cancer cells, potentially represents a step up in difficulty in data analysis. Such clinical samples can include up to six MHC alleles instead of the one MHC allele from the monoallelic A*02:01 cell line[187]. We tested MSBooster on a tissue sample from a patient with

metastatic malignant melanoma (Mel15)[194]. This data was acquired on a Q Exactive instrument (Thermo Fisher Scientific, Bremen) with HCD fragmentation. Considering just canonical peptides from the UniProt database, MSBooster increased peptide IDs from 34648 to 41236, a 19.0% boost (Figure 2.3a). Importantly, when considering just noncanonical peptides with variants derived from exome sequencing[194], peptide IDs increased from 14 to 18 (Figure 2.3b, Table 2.1).

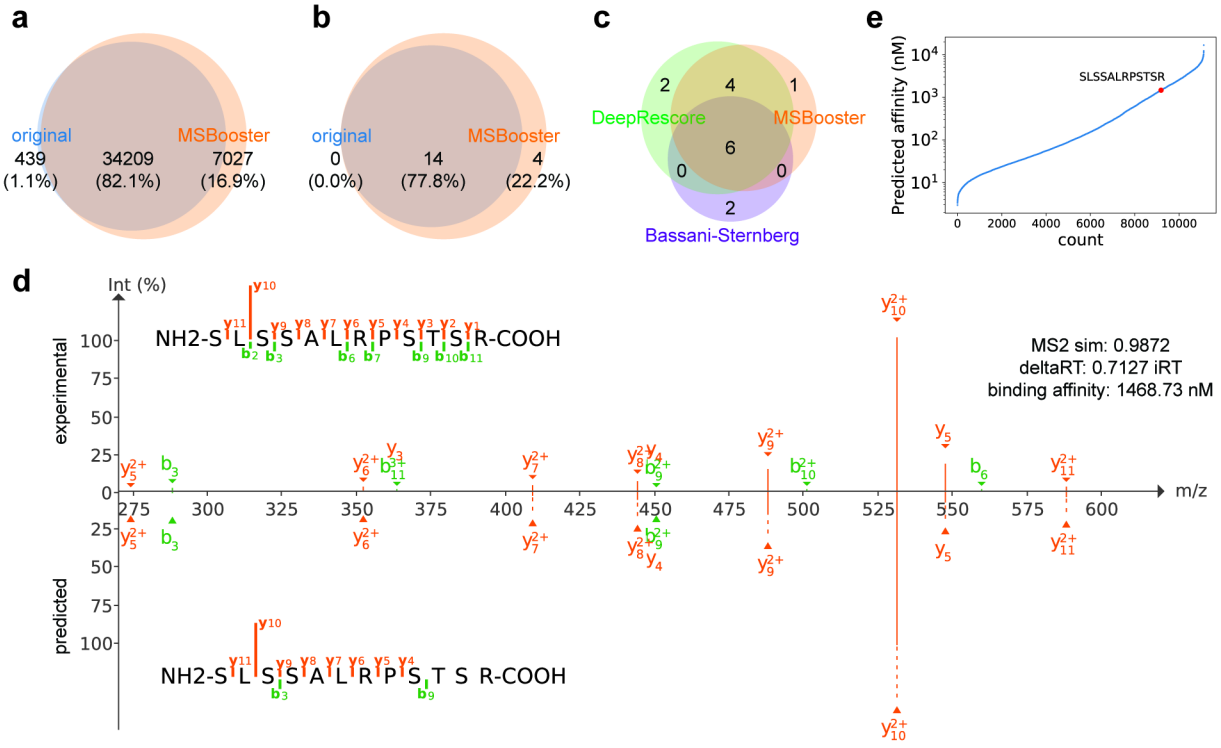


Figure 2.3 Neoantigen discovery in melanoma tissue

(a-b) Venn diagrams of peptides identified without (original) and with MSBooster. These peptides are categorized as canonical peptides from the reference database (a) or noncanonical neoantigens derived from mutations detected by exome sequencing (b). (c) Venn diagram of neoantigens proposed in DeepRescore, Bassani-Sternberg et al., or our study with MSBooster. Peptides in (a-b) were of lengths 7-25, while peptides in (c) were filtered between lengths 8 and 12. (d) PDV visualization of experimental and DIA-NN predicted spectra. y1, y2, b1, and b2 ion intensities are not predicted by DIA-NN and are therefore excluded from visualization. (e) NetMHCpan 4.1 binding affinities of peptides predicted to bind A*03:01. The newly detected peptide SLSSALRPSTSR is shown in red.

The newly reported neoantigens are credible for multiple reasons. First, three of the peptides (RTYSLSSALR, SLSSALRPSTSR, and SYVTTSTRTYSLSSALRPSTSRS) contain

the same single amino acid variant VIM^{G41S} as six other neoantigens already reported without MSBooster. Furthermore, the 14 original neoantigens had an average MS2 similarity of 0.94 and average delta RT of 4.1, compared to the 4 new neoantigens with 0.92 average MS2 similarity and 1.4 average delta RT. We also compared our identified neoantigens to those reported in prior analyses of the same data[150,194] (Figure 2.3c). Consistent with prior work, we only considered length 8-12 peptides here. Both MSBooster and DeepRescore[150] rejected two peptides reported in the original Bassani-Sternberg study[194] – ASWVVPIDIK, which MSFragger did not report, and GRTGAGKSFL (MS2 similarity: 0.81, delta RT: 8.77), which did not pass 1% FDR. Similarly, DeepRescore suggested two peptides that MSBooster did not – DVFPEGTRVGL, which MSFragger did not report, and RLFLGLAIK (MS2 similarity: 0.74, delta RT: 4.12), which did not pass 1% FDR. MSBooster reported one unique peptide in the allowed length range, SLSSALRPSTSR. The best spectral similarity across all PSMs of this peptide was 0.9872, and the lowest delta RT was 0.7127 iRT (Figure 2.3d). Its predicted binding affinity to one of Mel15's alleles A*03:01 was 1468.73 nM, designating it as a weak binder (Figure 2.3e).

Previous studies have verified neoantigens by synthesizing them and comparing spectra from the original and synthetic datasets. Comparison of the experimental and predicted spectra functions as a similar quality control measure. To enable researchers to manually verify PSMs, we have incorporated PDV[186] in FragPipe (creating FragPipe-PDV viewer) for visualization of experimental spectra. In addition, we have added support for loading spectral predictions from MSBooster to FragPipe-PDV, enabling generation of mirror plots to compare experimental and predicted spectra. For the peptide SLSSALRPSTSR, there is high concordance between the spectra, especially with the strong y10^{2+} ion (Figure 2.3d). Manual comparison of the spectra further corroborates that neoantigens proposed by MSBooster are likely true peptides.

Peptide	unweighted spectral entropy	delta RT loess	expectation value
ETLKPGTCVKR	0.9902	1.17262811	6.06553E-06
FVPPTAISHF	0.9528	3.561454358	0.004053639
GPGTASRPSSSRSYVTTSTRTYSLS	0.5778	2.699850881	1.33343E-05
GRIAFFLKY	0.9955	2.511929565	2.096E-05
KLILWRGLK	0.9852	8.616834672	0.006782847
KLKLPIIMK	0.9468	1.883031029	0.000474699
LPIQYEPVL	0.9285	5.580532295	0.006920069
RIKQTARK	0.9593	1.643467155	0.02839955
RLFKGYEGSLIK	0.9608	2.547394134	7.14126E-08
RTYSLSSALR	0.9819	0.294304948	3.28318
SLSSALRPSTSR	0.9872	0.712666839	0.7447078
SLSSALRPSTSRSL	0.9894	6.30303471	2.67227E-05
SLSSALRPSTSRSLY	0.9885	5.637047438	3.59958E-09
SLSSALRPSTSRSLYASSPGGVYAT	0.9139	2.494140328	3.86893E-06
SRFLSQLDK	0.9458	3.96211603	0.0430567
SYVTTSTRTYSLSSALRPSTSRS	0.7695	0.712666839	0.8668054
TYSLSSALRPSTSRSL	0.9915	6.30303471	3.36601E-05
TYSLSSALRPSTSRSLY	0.9943	5.91032802	0.000069728

Table 2.1 Mel15 patient-specific peptides

Peptides identified from a patient-specific protein database file from melanoma issue[194]. Peptides highlighted in yellow are those only identified when MSBooster is enabled. The remaining peptides were identified both with and without MSBooster.

2.3.4 Direct identification from DIA data

DIA offers the benefit of monitoring all precursors within the specified mass range and their fragments across retention time, thereby avoiding the stochasticity of DDA which can only produce MS/MS scans for a limited number of precursors. We extended MSBooster to rescoring of peptide IDs from DIA data. In FragPipe, peptide identification from DIA data can be performed in two ways: (1) with MSFragger-DIA, which identifies peptides from DIA MS/MS scans by direct database searching; and (2) by first processing the DIA MS files using DIA-Umpire[184] to extract pseudo MS/MS spectra, followed by searching with MSFragger as regular DDA data. We tested both approaches on a dataset of six melanoma cell lines[195]. MS/MS data was acquired on an Orbitrap Fusion Lumos Tribrid (Thermo Fisher Scientific) mass spectrometer with HCD fragmentation. Using MSFragger-DIA, MSBooster features increased peptide and protein IDs by 16.6% and 8.9%, respectively (Figure 2.4a-b). Using the DIA-Umpire based

workflow, the number of peptide and protein IDs increased by 16.6% and 9.0%, respectively (Appendix Figure A.5).

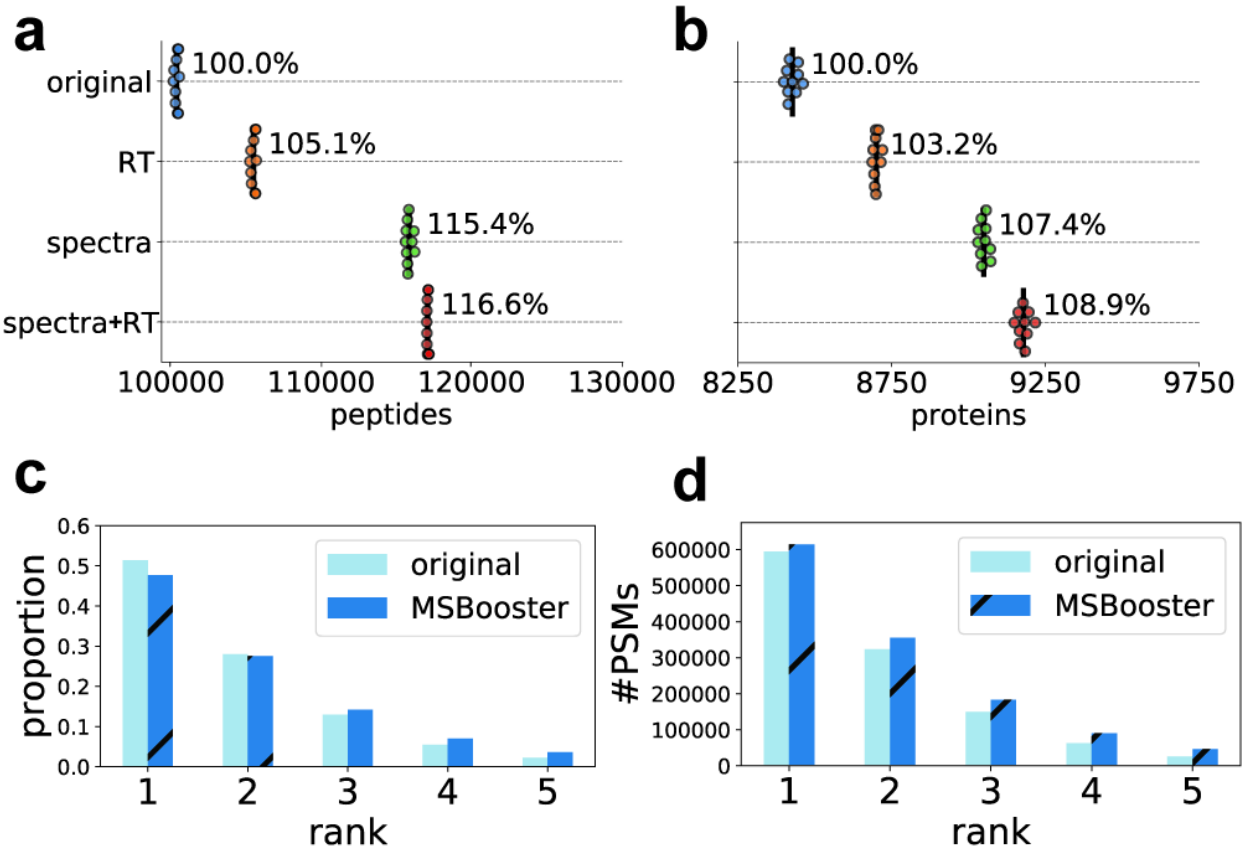


Figure 2.4 Melanoma DIA rescoring with MSFragger-DIA

(a-b) Swarmplots of the number of peptides (a) or proteins (b) reported at 1% FDR. (c-d) The proportion of PSMs (c) or total number of PSMs (d) from each of the five ranks reported. The darker, diagonally dashed bars represent results after spectral and RT rescoring, while the lighter, solid bars represent the results without using deep learning features.

The benefit of rescoring MSFragger-DIA results with MSBooster applies not only to the top scoring PSMs, but also to lower-ranking PSMs. By default, MSFragger-DIA reports up to 5 PSMs per MS/MS scan. While the initial MSFragger rankings were based on hyperscore, other features provided orthogonal information (Appendix Figure A.3b) and helped rescue true PSMs with lower hyperscores. With MSBooster, while the total number of PSMs passing the 1% FDR increases across all ranks, a higher proportion of accepted PSMs are from ranks 3 and below,

while the relative proportion from ranks 1 and 2 decreases (Figure 2.4c-d). MSBooster effectively rescues those lower-ranking PSMs that display characteristics of true positive PSMs.

2.3.5 Single cell proteomics

Single cell proteomics provides a view of the proteomes of individual cells. The lower level of maturity of technological platforms, along with the increased stochasticity of peptide identification due to cell-to-cell variability, makes single cell proteomics another promising area for DL-based PSM rescoring. We tested MSBooster on single cell data from the nanoPOTS platform[196] generated using an Orbitrap Fusion Lumos Tribrid instrument with collision-induced dissociation (CID) fragmentation. Briefly, we analyzed the data obtained from 1, 3, 10, or 50 cells. Single cell MS/MS spectra differ from bulk cell spectra in terms of the number of fragments matched and the degree of fragment ion intensity suppression[197]. We found a trend (Figure 2.5a) that with more cells there was a gradual increase in the median spectral similarity among confidently identified target PSMs (i.e. PSMs with expectation values "e-values" lower than the lowest decoy PSM e-value). As a reference, bulk secretome data obtained from An et al.[198], also generated on an Orbitrap Fusion Lumos instrument, demonstrated a higher median spectral similarity score.

There was a decrease in the median RT difference between 1 and 3 cells, due to an insufficient number of PSMs for optimal RT calibration in MSBooster with one cell only (Figure 2.5b). However, the median RT difference did not decrease past 3 cells, because the RT difference should not change once there are sufficient PSMs for RT calibration. The bulk cell RT score distribution was excluded from the comparison, because the RT score depends on the individual LC set up. Despite the increasing concordance between experimental and predicted values with increasing cells, we did not notice a monotonic relationship between the cell number

and MSBooster performance (Figure 2.5c-d). The single-cell (1 cell) experiment gained 4.7% and 2.8% peptide and proteins, respectively, with spectral and RT rescoring. The few-cell data (3, 10, and 50 cells) gained up to 10.6% peptides (in the 50 cell dataset) and 10.6% proteins (in the 3 cell dataset). The RT feature outperformed the spectral feature in many instances.

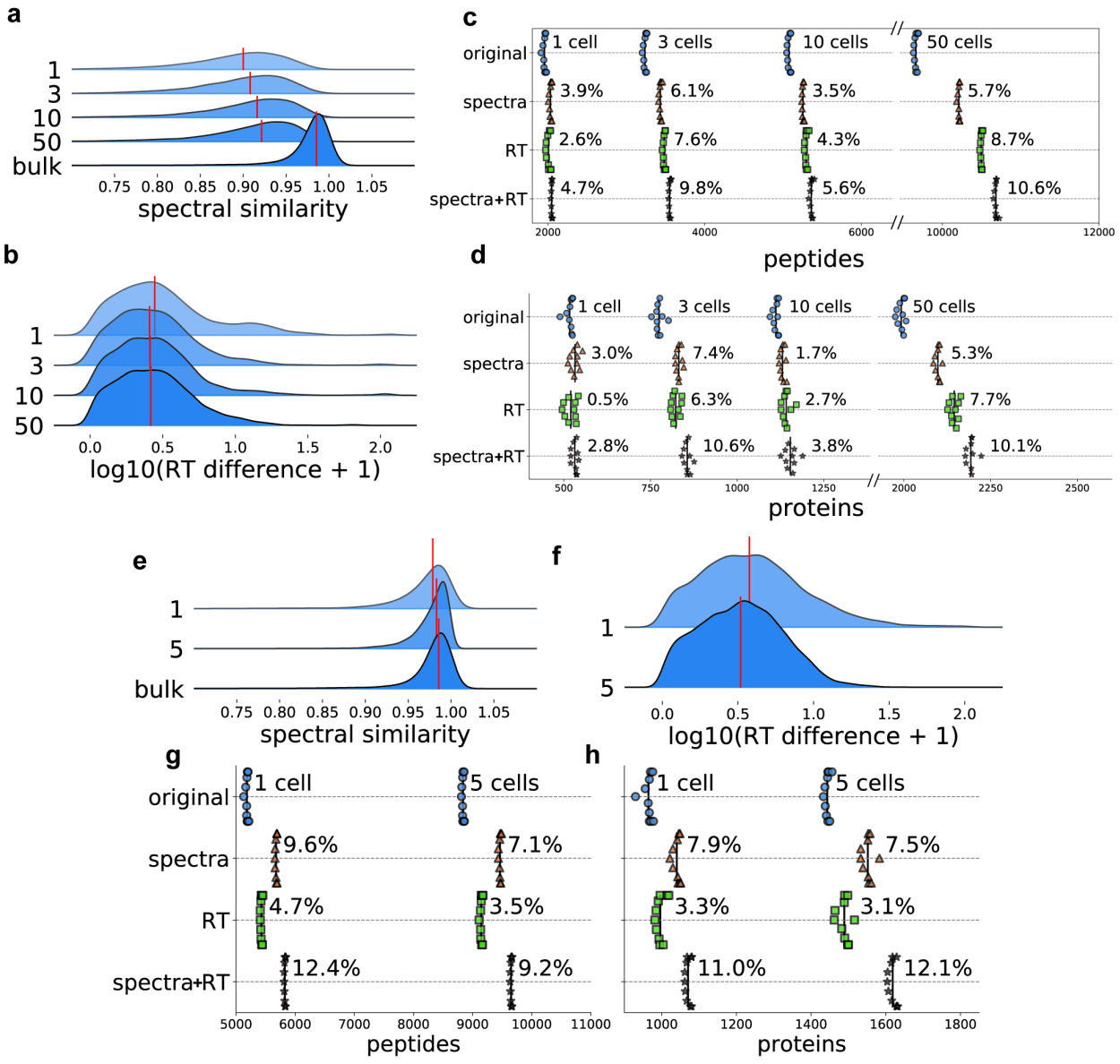


Figure 2.5 Single cell rescoring

(a-d) Results for nanoPOTS data[196]. (a-b) Ridge plots showing the distribution of the spectral (a) and RT (b) feature scores of confident target PSMs for different numbers of cells. The red line indicates the median value. The bulk cell sample is from PXD026436[198], produced on an Orbitrap Fusion Lumos. The RT feature was log normalized for

better visualization. (c-d) Swarmplots of the number of reported peptides (c) and proteins (d) when using different features for Percolator rescoring. (e-h) are the same as (a-d), but for the DISCO data[178].

Because single-cell proteomics methods are rapidly developing, we tested another dataset produced on a Q Exactive MS with Orbitrap mass analyzer with a different sample processing protocol (DISCO)[178] and HCD fragmentation to see whether MSBooster performance was consistent between different single cell protocols. Data from 1 and 5 cells were available. While we see similar trends of increasing median spectral similarity and decreasing median RT difference with an increasing number of cells, median spectral similarity is already above 0.95 for single cells in these data (Figure 2.5e-f). In comparison, even 50 cells in the nanoPOTS data had a median spectral similarity below 0.95 (Figure 2.5a). The lower MS/MS spectral prediction performance in the nanoPOTS data is likely due to differences between CID and HCD spectra. We also noted significant differences in the decoy PSMs' spectral similarity distributions (median of 0.49 vs 0.31, nanoPOTS vs DISCO) and differing numbers of PSMs reported per replicate (mean of 2924 vs 19878, nanoPOTS vs DISCO). We can see the effect of the higher similarity in the DISCO dataset reflected in the number of peptide and protein IDs achieved with rescoring (Figure 2.5g-h). In this dataset, the spectral similarity feature always outperformed the RT feature. The DISCO dataset experienced greater gains in peptide and protein IDs compared to the nanoPOTS dataset at the single cell level, possibly in part due to the greater spectral similarity between the DISCO experimental and predicted spectra.

2.3.6 timsTOF PASEF data with ion mobility separation

Next, we evaluated the performance of MSBooster on data from a HeLa tryptic digest standard analyzed using Parallel Accumulation–Serial Fragmentation (PASEF) on a timsTOF Pro (Bruker) mass spectrometer[65], which couples trapped ion mobility spectrometry (TIMS) to a time-of-flight (TOF) detector. Precursors were fragmented with CID, with amount of collision

energy as a function of ion mobility. Using both spectral and RT features, we achieved 3.9% and 2.7% increases in peptide and protein identification, respectively (Figure 2.6a-b). While this seems to be a minor increase, it highlights that FragPipe's default workflow for conventional tryptic searches performs well even without DL-based rescoring. In this case, giving Percolator only the hyperscore feature can recover most peptides and proteins without the help of the other features reported by MSFragger and MSBooster (Figure 2.6a-b).

Ion mobility (IM) is an additional method for separating precursors prior to MS/MS sequencing. As such, DL models have been extended to predict IM or related collisional cross section values[171,176] for peptide ions. To assess the utility of predicted IM for PSM rescoring, we ran MSBooster with IM features analogous to its RT features. We observed a negligible increase in the number of identified peptides and proteins, below 0.5% (Figure 2.6a-b), with the addition of the IM score. The weakness of the IM features may be explained by the high dependence of the IM on the precursor mass and charge. Because decoy PSMs still have the same charge and similar mass as the unknown true target precursor, predictions of their inverse IM (1/K0) values are still highly correlated with the experimental value (Figure 2.6c-d; Appendix Figure A.6). There is not as much of a spread for IM as for RT prediction (Figure 2.6e-f). Overall, the target PSMs showed a slightly different distribution of IM feature scores from the decoy PSMs (Appendix Figure A.7e).

Because IM features only marginally improved IDs, we further explored this data to see if we could improve our features. We noted a higher density of decoy PSMs at lower 1/K0 values (around 0.7-0.8) compared to confident target PSMs (Figure 2.6c-d, Appendix Figure A.8). Similarly, decoy PSMs tend to have lower masses and shorter lengths than target PSMs for lower charge states. We hypothesized that adding these simpler features may improve performance.

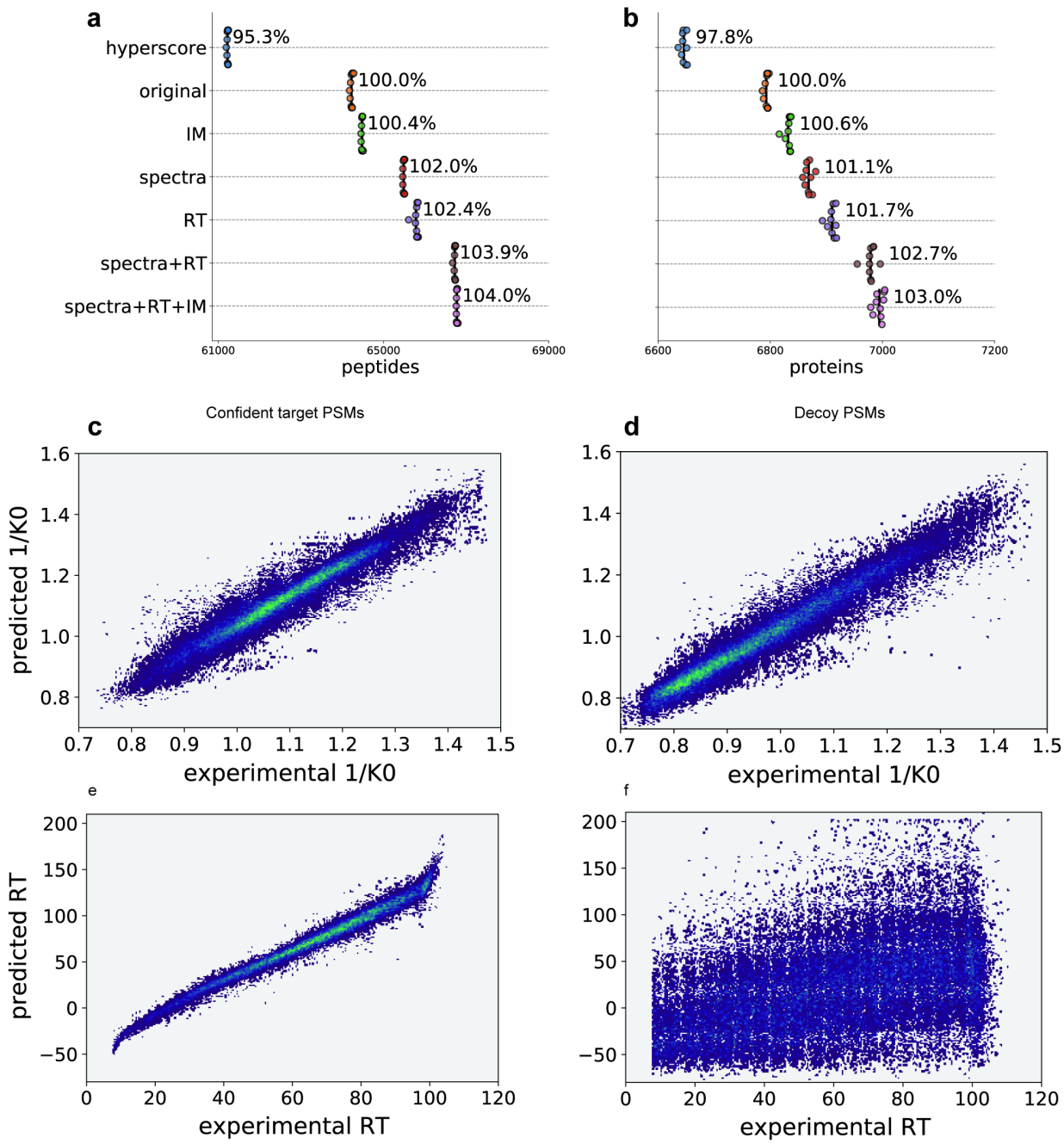


Figure 2.6 timsTOF HeLa rescoring

(a-b) Swarmplot of peptides (a) and proteins (b) reported at 1% FDR. (c-f) Scatter density plots showing the relationships between DIA-NN predicted and experimental IM (c-d) and RT (e-f) values in seconds for peptides with charges 2 and above. Confident target PSMs are shown in c and e, decoy PSMs in d and f. The brighter colors correspond to higher densities of PSMs.

Peptide mass and length have already been incorporated as features in the MSFragger-generated pin file. We tested adding one-hot-encoded charge and 1/K0 values (not to be confused with the

DL-based IM features in MSBooster) as features for Percolator as well (Appendix Figure A.9). While 1/K0 by itself did not improve over only using spectral and RT features, the combination of charge and 1/K0 led to a 4.4% increase over baseline peptide IDs without MSBooster. Simply adding charge and 1/K0 values provides greater boosts compared to our hand-crafted IM feature, but the improvement is still minimal.

2.3.7 Multiple correlated features

Because MSBooster can calculate several variants of spectral, RT, and IM features, we evaluated whether there was value in using multiple correlated features for PSM rescoring (Appendix Figure A.10; all tested features are described in Methods Section 2.6.2). This idea was spurred by the finding that even correlated features may not be truly redundant and may work well to provide better separation between classes[199]. For all datasets, we annotated Percolator input files with either single features ("spectra+RT" and "spectra+RT+IM") or all available features ("multiple spectra+RT" and "multiple spectra+RT+IM"). In most analyses, the use of multiple correlated features resulted in a minor (<1%) increase in identification numbers. Occasionally, the numbers decreased by an equally small amount. To investigate this further, we returned to the HLA immunopeptidome dataset. With the use of multiple correlated features, compared to using only a single feature for each feature type (as in Figure 2.2), 165 peptides were lost, and 184 were gained. Both groups of lost and gained HLA peptides had comparable binding rates (123/165 lost were binders vs 134/184 gained were binders; $\chi^2(1, N=349) = 0.1325$, $p = 0.716876$, chi-square test) and similar sequence motifs (Appendix Figure A.11). Single features are used by default, but because it is difficult to rule out all scenarios where using correlated features may be beneficial, we provide an option for FragPipe users to enable their use (Appendix D).

2.4 Discussion

MSBooster is a new addition to the FragPipe computational platform that provides a boost to the number of identified PSMs, peptides, and proteins by generating DL-based features for PSM rescoring with Percolator. It automatically runs a DL model to acquire predictions for RT, IM, and MS/MS spectra, and generates features based on these predictions to expand the list of scores for each PSM that are useful for discriminating between true and false matches. A notable benefit of MSBooster is that the whole peptide library need not be predicted; only high-scoring candidates identified by MSFragger are evaluated, saving a lot of time, especially for nonspecific searches. Importantly, MSBooster is fully incorporated into FragPipe: a single checkbox enables DL prediction and addition of new features for Percolator rescoring, allowing even inexperienced users to immediately see improved peptide and protein IDs.

We evaluated the improvements provided by MSBooster in various experimental workflows available in FragPipe. We observed robust gains across applications, especially in analyses exhibiting a large search space, such as HLA immunopeptidomics, or multiple peptide candidates per MS/MS scan, such as direct identification from DIA data searches. We found that MSBooster has similar performance to DeepRescore[150] when rescoring a patient-specific melanoma dataset, while also proposing unique neoantigens. MSBooster in FragPipe also outperforms MaxQuant rescoring when both have access to a Prosit predicted library. Several factors may contribute to MSBooster performance, including MS2 spectral quality, deviation between the experimental data acquisition parameters and those of the training data for the prediction models, and the number of PSMs available for rescoring. We present general guidelines for what level of gains are expected from multiple popular applications of MSBooster.

Interestingly, we observed only a marginal impact of adding IM features when analyzing standard HeLa tryptic digest timsTOF PASEF data. However, this does not mean that IM

features will not be useful in other scenarios. IM may improve the resolution of peptidoforms with isobaric modifications like glycopeptides[200] or assist with PTM site localization[201,202]. The minimal strength of IM features observed in this work may also be due to insufficient prediction accuracy, suboptimal feature curation in MSBooster, or the limitations of a linear SVM model in Percolator. Thus, more flexible models[203] for PSM rescoring could be investigated.

2.5 Updates since publication

Since the first release of MSBooster v1.0 in Fragpipe 17, many improvements in flexibility have been made. For example, tools such as ionbot and CHIMERYS can report multiple PSMs per spectrum, potentially allowing consideration of more candidate sequences or co-fragmenting peptides[155,204]. MSFragger has been extended to include a DDA+ mode[205], which is optimized for identification of co-fragmenting precursors present in DDA data. MSBooster is seamlessly incorporated in the MSFragger-DDA+ pipeline.

Second, we have published a standalone command line version with exclusive parameters (Appendix D) to be incorporated into various pipelines outside of FragPipe (https://github.com/Nesvilab/MSBooster). This would be applicable for users interested in de novo sequencing[206-208] or using MSBooster in conjunction with other PSM rescoring tools besides Percolator[203,209]. As these interests arise, MSBooster can be adapted to work with PSM table formats besides Percolator pin files, such as pepXML files.

Third, we have extended MSBooster to use prediction models outside of DIA-NN; fragmentation type, mass spectrometer, and collision energy are inputs not considered by DIA-NN that can have a pronounced influence on peptide fragmentation mechanisms. We have already shown Prosit to be compatible with standalone MSBooster in our HLA example. Increased accessibility to more models warrants a revisit of our single cell analysis. At the 2024

International Single Cell Mass Spectrometry conference, we presented updated rescoring of the nanoPOTS data with the Prosit CID spectral prediction model[179]. By using a more appropriate model (rather than DIA-NN, which predicts HCD spectra), we show that single-cell spectra from this dataset can achieve nearly 0.95 spectral similarity (Appendix Figure A.12a). We also note a jump from 7-8% increases in peptide IDs in single- and few-cell data to 11.5% in the 50 cell replicates (Appendix Figure A.12b), which may represent a sweet spot where enough precursors are captured to mimic bulk cell spectra prediction yet too few PSMs are called for accurate PSM score distribution modeling, as is the case with bulk cell tryptic digest analysis where we see lesser gains in peptide IDs. Chapter 3 provides details of how more models are supported natively in FragPipe.

Finally, we report a method to perform RT/IM calibration for user-defined mass shifts (Methods Section 2.6.12). Learning multiple calibration curves can be beneficial when using models that do not support a given mass shift of interest, and that mass shift causes a somewhat consistent shift in RT/IM. We show it performs well in learning a separate LOESS function for photoaffinity probes with large masses[210] (Appendix Figure A.13). We can immediately see that peptides bound by the probes systematically elute later. Learning a separate curve is not a perfect substitute for training a model on the PTMs of interest, as it is also visually apparent that PSMs for probe-bound peptides have a wider spread around their curve than the non-probe PSMs do. However, using multiple calibration curves may also save PSMs that would otherwise be lost. We reanalyzed the monoallelic cell line data, this time using a separate calibration curve for pyro-glu from Q (Appendix Figure A.14a). Appendix Figure A.4b showed that the -17 Da PTM suffered when using MSBooster. Learning a separate calibration function for it successfully rescues many of the modified peptides lost upon rescoring with older MSBooster versions that

lacked this capability (Appendix Figure A.14b). While using one RT curve for all PSMs produced 50 pyro-glu from Q PSMs (35 unique peptide sequences), using a separate curve for the PTM produced 316 total PSMs (153 peptides) in the final FDR-filtered PSM list. These 316 PSMs had a median delta RT score of 4.7 minutes, compared to the median of 7.1 minutes when using a single LOESS curve. Though not tested here, increased stringency can be applied by adding group FDR, which considers regular PSMs/peptides separately from noncanonical peptides or peptides with rare PTMs[211-213].

2.6 Methods

2.6.1 MSBooster workflow

Workflows with and without MSBooster are depicted in Figure 2.1. In DDA experiments, MS/MS spectra are searched using MSFragger. For peptide identification from DIA data, either full MS/MS spectra are searched using MSFragger-DIA, or DIA-Umpire extracted pseudo-MS/MS spectra are searched using MSFragger as conventional DDA files. MSFragger produces pepXML and pin files, and the latter are used as input into MSBooster. The pepXML files are not used by MSBooster but are necessary for converting the Percolator output files into pepXML files for subsequent protein inference analysis using ProteinProphet. To obtain a list of peptides for DL model prediction, MSBooster iterates through all pin files to obtain all target and decoy peptides matched to at least one PSM. Peptides with the same sequence but different PTMs and/or charges are treated as different peptides. This list is then passed to a DL model to obtain a prediction file. This strategy is significantly faster than predicting the entire in-silico digested proteome, as only peptides reported as potential hits by MSFragger (top ranking peptides per spectrum) are submitted for prediction. Peptides with PTMs not supported by the DL model rely

on predictions for the peptide without the unsupported PTMs. MSBooster adds a shift in fragment m/z to accommodate the new PTMs, but the fragment intensities remain the same.

The core of MSBooster is the feature calculation step. First, the predictions from the DL model are loaded. Then pairs of mzML (or MGF) and pin files are sequentially loaded for processing and DL-extended PSM feature table generation. For DIA data, because multiple peptides may contribute to a single MS/MS scan, the experimental spectra are revised after each PSM has its features calculated. Borrowing from MSFragger-DIA, the highest intensity experimental MS/MS peak within the fragment error tolerance of the reported predicted fragment is removed from the experimental spectrum. The predicted spectra of lower-ranking PSM peptides can no longer have their fragments matched to these removed peaks, because using the same MS/MS peaks for multiple PSMs from the same scan can lead to spurious hits. Once all PSMs from a single pin file are loaded, RT and IM calibrations are performed. In the final step, MSBooster iterates through the pin file row by row, and calculates and adds the desired features. This process is repeated until all pin files have DL features calculated and added. Multiple features were tested and are discussed below.

2.6.2 Calculated MS/MS, RT, and IM features

Available features in MSBooster v1.0 are listed below, along with an equation and source for less commonly used metrics. By default in FragPipe, the unweighted spectral entropy and delta RT LOESS only are used. More features are calculated in v1.3 and are discussed in Methods Section 4.5.4.

- Spectral similarity: Unweighted spectral entropy requires fragment ion intensity vectors to sum to 1. Other metrics require unit normalization (i.e. the sum of squared intensities equals 1). p stands for predicted fragment, P for predicted intensity vector, m for matched

experimental fragment, and M for matched experimental intensity vector. By default, the top 12 highest predicted intensity fragments are used.

- Bray Curtis[214]: $1 - \frac{\Sigma_{i=1}^{n} |p_i - m_i|}{\Sigma_{i=1}^{n} p_i + m_i}$
- Pearson's correlation: If no matched experimental fragment ions, value is -1
- Dot Product
- Unweighted spectral entropy[215]: $1 - \frac{2S_{PM} - S_P - S_M}{ln\,4}$, where entropy S = $-\sum_{i=1}^{n} f_i \, ln \, f_i$, and S_{PM} is the sum of predicted and matched vectors S_P and S_M divided by 2

- Retention time similarity
  - delta RT loess: Described in Methods Section 2.6.4 (Retention time and ion mobility calibration).
  - delta RT loess normalized: Similar to delta RT loess but divided by the interquartile range of predicted RTs for high confidence PSMs in the experimental RT vicinity.
  - RT probability Unif Prior: Described in Methods Section 2.6.5 (Kernel density estimation of predicted retention time and ion mobility distributions).
- Ion mobility similarity
  - Features are like those for RT but done separately for each charge. Only charges +1-7 are supported.

2.6.3 Spectral similarity calculation

To calculate the spectral similarity, the highest intensity fragment ions within the m/z error tolerance of the predicted fragment ions are obtained. Therefore, similarity calculations are

performed using vectors of the same length. If no peak is detected in the experimental spectrum within the m/z error tolerance of the predicted peak, the experimental vector is assigned a 0 at that position. Predicted and matched fragment ions from the experimental MS/MS spectra were normalized before similarity calculation.

2.6.4 Retention time and ion mobility calibration

Using the Apache Commons Math 3 API, local regression (LOESS) is used to calibrate the experimental to the predicted RT and IM values. This is followed by monotonic regression using a Java implementation of the pair adjacent violators algorithm (https://github.com/sanity/pairAdjacentViolators). A different ion mobility model is trained for each charge. The resultant model maps each experimental RT to a value on the predicted RT scale; a calibrated RT value is the experimental RT mapped to the predicted scale. To train the regression models, a subset of PSMs (5000, by default) with expectation values below a preset threshold (10e-3.5) is used. This threshold was chosen from the observation that of the various datasets MSBooster was tested on, no decoys were detected below that expectation value. If between 50 and 5000 PSMs have sufficiently low expectation values, that number of PSMs is used. If fewer than 50 target PSMs with sufficiently low expectation values are available, linear regression is performed instead. For both DDA and DIA, only rank 1 PSMs are considered for the regression. PSMs from peptides with no prediction (e.g. unsupported by the model for being too long) are filtered out. Only one PSM with the lowest expectation value is used from each unique peptide sequence (with PTMs).

A revised LOESS fitting algorithm is used for RT/IM calibration. There are two rounds of fitting, with each round having a step for LOESS and an isotonic regression step for enforcing monotonicity. In the second round, the mean deviation from the calibration curve is calculated for all PSMs used for the LOESS fitting. Any PSM with a delta RT/IM score more than two

standard deviations away from the mean is considered an outlier and removed from the calibration process. The second LOESS calibration step is applied, but weights are applied based on the local median delta RT/IM in a window surrounding each PSM, thereby allowing PSMs with larger deviations to exert more weight; this allows the algorithm to capture sudden changes in slope between experimental and predicted values, which is often a byproduct of different LC setups. The final isotonic regression step is then applied.

To calculate the difference between the predicted and experimental RT, the experimental RT is first calibrated to the predicted scale using the regression model, followed by calculating the difference between the calibrated RT for that MS/MS scan and the predicted RT for that peptide. The same is done for the IM.

To determine the best bandwidth for LOESS fitting, a grid search is performed by splitting the data into five partitions. There are five rounds of fitting. In each round, four splits are used to learn LOESS curves with varying bandwidths (0.01, 0.05, 0.1, and 0.2), and the resulting curves are evaluated on the held out partition. The mean squared error (MSE) for each bandwidth is recorded, and the bandwidth resulting in the lowest MSE is recorded. The average of the bandwidths used for each fitting round is the final bandwidth used when learning a LOESS curve on all data.

2.6.5 Kernel density estimation of predicted retention time and ion mobility distributions

The following discussion uses RT, but the same applies to IM. Empirical distributions of predicted RT/IM values were generated using Statistical Machine Intelligence and Learning Engine (Smile) implementation of kernel density estimation (KDE) with a Gaussian kernel (https://haifengl.github.io/api/java/smile/stat/distribution/KernelDensity.html). The bandwidth of the kernel is estimated by Silverman's rule of thumb, as implemented by Smile. Briefly, the

algorithm works by replacing each point in the distribution with a Gaussian curve of equal amplitude, then summing all the individual curves into one total distribution that is a smoothed version of the empirical distribution. The empirical range is divided into equally sized bins (widths of 1 minute for RT, 0.01 1/K0 units for IM, by default). For each PSM, its predicted RT is placed into a bin with all predicted RTs of PSMs from the same experimentally observed RT minute. The number of times its RT value is added to the bin is weighted by its expectation value; that is, a higher-confidence PSM with a low expectation value will have its predicted RT added to the bin more times than a lower-confidence PSM with a high expectation value. After all predicted RTs are placed in their respective bins, KDE is used to generate empirical distributions. These distributions can be used to estimate the probability of having a PSM with a predicted RT value given its experimental RT and are not subject to the monotonic constraint of the LOESS model. For example, to get the "RT probability" value for a peptide scanned in the 60^{th} minute, the KDE distribution from the 60^{th} minute bin, D, is queried. If the peptide has a predicted RT R_P, then the reported value for this feature is the KDE estimated probability (P_E) of having predicted RT R_P from the 60^{th} minute bin's probability distribution, D. The same procedure is applied to the IM to generate distributions of the predicted IM, separating the PSMs by charge state. The features from MSBooster that use these probabilities also add a uniform prior distribution to the KDE-generated distribution. This uniform prior helps to dampen the effects of bins with fewer entries. For example, if an experimental RT bin contains a single PSM, not using a uniform prior would result in an artificially high probability for that PSM. The uniform prior is decided by sorting the RT bins in ascending order by the number of PSMs contained in each. A bin is chosen based on a preset percentile (10^{th} percentile, by default). The number of PSMs U at the RT bin at this specific percentile is chosen. The uniform prior

probability P_U is distributed equally across the predicted RT range. If a PSM with empirical probability P_E is placed in a bin that contain E PSMs, the value of the "RT probability uniform prior" feature is described by equation (1):

$$\frac{\mathrm{P_U} * \mathrm{U}}{\mathrm{U} + \mathrm{E}} + \frac{\mathrm{P_E} * \mathrm{E}}{\mathrm{U} + \mathrm{E}} \quad (1)$$

2.6.6 Determination of the best features

Several metrics exist for calculating the similarities between experimental and predicted spectra. Although cosine similarity is commonly used, several features were tested to determine which metrics could provide the greatest gains in the identification numbers (Appendix Figure A.15). Percolator is non-deterministic because of the random splitting of PSMs for training and testing, which can be controlled with a random seed. Thus, Percolator was run ten times for each feature, and the number of peptides reported after Philosopher filtering was counted. For spectral similarity features, the greatest boosts were consistently obtained with "unweighted spectral entropy" similarity[215]. For the RT features, "delta RT loess" tended to do the best. Interestingly, "delta RT loess normalized" performed better when there were a small number of cells in the nanoPOTS data[196] (Appendix Figure A.15e-h). We tested a linear regression feature for RT calibration, "delta RT linear," on the HLA and 50 cell datasets (Appendix Figure A.15a). While it performed similarly to "delta RT loess normalized" on the HLA dataset, we found that it may performed sub-optimally for the 50-cell data, where there exists a non-linear relationship between the experimental and predicted RT scales (Appendix Figure A.16). For the IM features, the "IM probability uniform prior" feature performed the best. The distributions of each score for all PSMs, targets and decoys, in the different datasets are shown in Appendix Figure A.7.

2.6.7 MSFragger search and FDR control

Database searches were performed using MSFragger v3.4 in FragPipe v17.2 with Philosopher v4.1.1. For neoantigen detection, MSFragger v3.7, FragPipe v19.2, and Philosopher v5.0.0 were used for visualization of spectra with FragPipe-PDV viewer[186]. All searches used a UniProt fasta from March 18, 2022 except for the neoantigen search, which used a fasta with both canonical and Mel15 specific protein sequences derived from exome sequencing; this fasta was the same used by Li et al.[150] and generated by NeoFlow[216]. The workflows used for each dataset are as follows: HLA immunopeptidome[187] (nonspecific-HLA-C57 workflow); melanoma neoantigen[194] (nonspecific-HLA with carbamidomethylated cysteine added as a variable modification); melanoma DIA data[195] with MSFragger-DIA (DIA_SpecLib_Quant) and with DIA-Umpire (DIA_DIA-Umpire_SpecLib_Quant); HeLa timsTOF[65], single cell proteomics with nanoPOTS[196] or DISCO[178], and secretome[198] (Default). All workflows included oxidation of methionine and N-terminal acetylation as variable modifications. The workflows besides Default also included pyro-glutamation from glutamine and glutamic acid. The HLA workflow had carbamidomethylated cysteine as a fixed modification with the mass difference between cysteinylation and carbamidomethylation (61.98 Da) as a variable modification; the neoantigen workflow included both carbamidomethylation and cysteinylation as variable modifications. A maximum of 3 variable modifications were allowed. Peptide length was set to 7-25 for nonspecific workflows and 7-50 for all others. All workflows used 20 ppm for precursor and fragment error tolerance, with mass calibration and parameter optimization enabled. MSBooster, Percolator, ProteinProphet, and Philosopher were enabled. The HLA workflow was revised to add "—mods M:15.9949" to the Philosopher filter to perform group-specific FDR estimation[217] using the following three categories: unmodified peptides, peptides with oxidized M only, and

peptides with any other modification. The nanoPOTS data were analyzed in separate experiments based on the number of cells (1, 3, 10, or 50). Peptide and protein IDs reported are at 1% FDR, unless otherwise noted. We attempted to calculate FDR specifically for nonreference targets/decoys in our neoantigen dataset[218]; however, too few nonreference decoys were reported to accurately calculate a group-specific FDR. Therefore, the numbers reported are those nonreference target peptides in the original peptide.tsv files.

2.6.8 MaxQuant search and FDR control

MaxQuant v2.1.0.0[70] was used to search the HLA immunopeptidome data[187]. The search tolerance was 20 ppm. For the MaxQuant-only search, oxidation of methionine, n-terminal acetylation, pyroglutamation of glutamine and glutamic acid, and cysteinylation minus carbamidomethylation of cysteine were specified as variable modifications. Carbamidomethylation of cysteine was specified as a fixed modification. FDR at all levels was set to 0.01. For the MaxQuant search to be used for Prosit rescoring, because the only PTMs supported by the base non-TMT Prosit model are carbamidomethylation of cysteine and oxidation of methionine, the former is set as the only fixed modification, the latter as the only variable modification. All FDR levels were set to 1, as is required for Prosit rescoring.

2.6.9 MaxQuant analysis of Sarkizova et al.

Results of the analysis of 92 monoallelic HLA Class I cell lines with MaxQuant[179] was downloaded. For each allele, the number of peptides with non-NA scores were counted and compared before and after Prosit rescoring. The average across all cell lines was calculated. Results are shown in Appendix Figure A.2a.

2.6.10 Deep learning predictions

DIA-NN v1.8[175,176] was used to predict RT, IM, and MS/MS spectra because of its speed and ease of execution within FragPipe. DIA-NN reports the top 12 most intense singly and doubly charged b- and y-ions. Predictions were made for each unique combination of peptide sequence, modifications, and charge. DIA-NN v1.8 supports the predictions for peptides with carbamidomethylated cysteine, oxidized methionine, N-terminal acetylation, phosphorylation, and ubiquitination. For other PTMs such as pyro-glutamation, DIA-NN did not adjust MS/MS fragment peak intensities, but MSBooster shifted the peaks to the appropriate m/z values. The RTs and IM values for peptides with unsupported PTMs remained the same as for counterparts without the PTM.

Prosit[179] was used for rescoring of HLA immunopeptidome data[187], both with MaxQuant and MSBooster. When running with MaxQuant output, the "rescoring" pipeline at https://www.proteomicsdb.org/prosit/ was used. The msms.txt from the MaxQuant with FDR = 1 search was used as input, along with the individual RAW files. The size of the unique set of peptides with q-values less than 0.01 are shown in Appendix Figure A.2b. When used in combination with MSBooster, a command line version of MSBooster was used to extract peptides from the pin files (Appendix D), analogous to how it is done in FragPipe with DIA-NN. The peptides are formatted into an input file for the "spectral library" pipeline in https://www.proteomicsdb.org/prosit/. For both MaxQuant and MSBooster, the "Prosit_2020_intensity_hcd" and "Prosit_2019_irt" models are used for MS/MS spectral and RT predictions, respectively. "Prosit_2020_intensity_cid" was used for the updated single cell analysis in Appendix Figure A.12. The resultant "msp" file from Prosit is read by MSBooster, and spectra for peptides with PTMs besides carbamidomethylated cysteine and oxidized

methionine are generated via m/z shifting, analogous to how it is done when used in conjunction with DIA-NN.

2.6.11 HLA motif analysis

Swarmplots and Venn diagrams were generated with no filtering of peptides based on length. Before using peptides in GibbsCluster 2.0 (https://services.healthtech.dtu.dk/service.php?GibbsCluster-2.0)[190] or the NetMHC software (https://services.healthtech.dtu.dk/service.php?NetMHC-4.0; https://services.healthtech.dtu.dk/services/NetMHCpan-4.1)[128], they were filtered to be between lengths 7 and 12 for Klaeger et al.[187], or between lengths 8 and 12 for Bassani-Sternberg et al.[194]. Position weight matrices were generated using GibbsCluster. The binding affinity of the peptides to the A*02:01 MHC was determined using NetMHC 4.0 using the default settings. Similarly, the binding affinity of peptides from Bassani-Sternberg et al.[194] to A*03:01 was determined using NetMHCpan 4.1, which accommodates peptides longer than length 11, such as our detected neoantigen.

2.6.12 Mass shift RT/IM calibration

A LOESS curve can be learned for each user-specified mass shift. When an MSFragger parameter file (fragger.params) is provided to MSBooster, the tool reads in the mass offsets (not variable modifications) and learns a curve for all the mass offsets together. On the command line, a comma-separated list of arbitrary mass shifts can be passed (Appendix D), each of which will have its own LOESS curve. If between 10 and 100 PSMs that pass all criteria listed in section 2.6.4 are available, a linear instead of LOESS regression will be used. If fewer than 10 PSMs pass the criteria, the mass shift in question will default to using the regular LOESS curve to calculate RT scores.

2.6.13 Statistical analysis and figure generation

Figures were generated in Jupyter Notebooks using Python 3.7.6, Anaconda 2020.02, Conda 4.8.2, Joypy 0.2.6, Jupyterlab 1.2.6, Matplotlib 3.1.3, Matplotlib-venn 0.11.7, Numpy 1.18.1, Pandas 1.3.0, and Seaborn 0.10.0. The scatter density plots for Figure 2.6 and Figure A.6 require a separate Anaconda environment with Python 3.8.3. Statistical tests such as t-tests were all two-sided and performed with SciPy 1.4.1.

2.6.14 Hardware

FragPipe was run and timed using Java 16.0.1. A command line version of MSBooster was run on a Windows desktop with 12 logical CPU cores (Intel(R) Core™ i7-8700 CPU @ 3.20 GHz) and 32 GB of memory. This was essential for automating the testing with different MSBooster features.

2.7 Data availability

MS/MS datasets used in this study can be found at the ProteomeXchange Consortium and the PRIDE partner repository[219] or at the MassIVE repository with the following accession codes:

- HeLa timsTOF DDA PXD010012[65]: https://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD010012
- HLA peptidome MSV000087743[187]: https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=a1638beae5d04267a99f92c550c60b34
- Melanoma neoantigen PXD004894[194]: https://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD004894

- Melanoma DIA PXD022992[195]:

  https://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD022992

- Single cell nanoPOTS MSV000085230[196]:

  https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=3013fc11dc4e4b6dae49a244d92854a7

- Single cell DISCO PXD019958[178]:

  https://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD019958

- Secretome PXD026436[198]:

  https://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD026436

All MSFragger produced pepXML, MSBooster-annotated pin, and fasta files are available at https://doi.org/10.5281/zenodo.8034585 and https://doi.org/10.5281/zenodo.7843558. MHC allele binding motifs were acquired at the Immune Epitope Database (https://www.iedb.org/).

2.8 Authors, affiliations, and contributions

Kevin L Yang[1], Fengchao Yu[2], Guo Ci Teo[3], Kai Li[1], Vadim Demichev[4,5], Markus Ralser[4,6,7], Alexey I Nesvizhskii[1,2]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA.
[2]Department of Pathology, University of Michigan, Ann Arbor, MI, USA.
[3]Department of Pathology, University of Michigan, Ann Arbor, MI, USA.
[4]Department of Biochemistry, Charité Universitätsmedizin, Berlin, Germany.
[5]Department of Biochemistry, University of Cambridge, Cambridge, UK.
[6]Nuffield Department of Medicine, The Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK.
[7]Max Planck Institute for Molecular Genetics, Berlin, Germany.

K.L.Y. developed the algorithm, wrote the software, and analyzed the results. F.Y. assisted with the algorithm and software development. F.Y. and G.C.T. provided support for integration of MSBooster in FragPipe. K.L. provided support for integration of MSBooster in FragPipe-PDV. V.D. and M.R. provided resources pertaining to DIA-NN. K.L.Y. and A.I.N. wrote the manuscript. A.I.N. conceived and supervised the study.

Chapter 3 Accessing and Benchmarking State-of-the-Art Models from the Koina Model Repository in FragPipe

This chapter has been adapted from *Koina: Democratizing Machine Learning for Proteomics Research* to focus on the MSBooster/FragPipe aspect. A preprint is posted on bioRxiv[220], and a publication with revisions has been submitted to Nature Communications.

3.1 Abstract

Tandem mass spectrometry enables exploration of the proteome, producing a complicated and heterogeneous view of biological systems. A single deep learning model that can predict peptide properties covering all post-translational modifications, LC and MS instrumentation, and experimental parameters may never exist. As such, diverse proteomics data requires diverse predictive models for accurate peptide-spectrum match rescoring. Though accessibility is at the forefront of model development, there are few practical solutions to access models trained by different labs in one central location. Here, we describe an update to MSBooster that leverages predictions from models hosted on the Koina web server and model repository. To demonstrate the ease of switching between models, we benchmarked state-of-the-art peptide prediction models on various datasets. We found that generally, PeptDeep performs best for phosphoproteomics data, Prosit for HLA data, and DIA-NN for Astral DIA data, while Prosit and DIA-NN's predictions performed comparably on TMT datasets. Foreseeing the choice overload that can accompany the wave of new models, we developed a heuristic approach for approximating the best-performing models for the data at hand.

3.2 Introduction

Deep learning-based PSM rescoring relies on accurate model predictions. In Chapter 2, we saw how features comparing experimentally observed values of peptide properties with predicted values made a powerful impact on distinguishing true from false IDs. We also described how the default peptide prediction model in FragPipe, DIA-NN, can repurpose its predictions, extrapolating them to types of data missing from DIA-NN's training set. Of course, relying on DIA-NN's predictions is only a quick solution. Indeed, we alluded to the power of diverse Prosit models twice in the last chapter: the first instance was when Prosit, trained on a large HLA dataset, improved HLA PSM rescoring (Appendix Figure A.2). The second example used Prosit CID instead of DIA-NN, which was trained only on HCD data, on a single cell dataset acquired with CID fragmentation (Appendix Figure A.12). The limitations of a single default model were very apparent when rescoring modified peptides that drastically alter peptide properties (Appendix Figure A.4). This showcases a major flaw in that users may unknowingly penalize certain classes of peptides, but users should not be expected to run their analyses both with and without MSBooster to see what peptides are lost and gained. Even worse, misusing models could inadvertently increase false positive IDs. This is the case in one recent study performing PSM rescoring on the citrullination modification[221]. It found that MSBooster increased both true and false positive IDs, resulting in reduced precision. Clearly, if MSBooster is to be broadly applicable, it needs access to diverse models.

Though the genesis of ML/DL tools has reshaped proteomics, it is not trivial to include them in FragPipe. Many models are accessible via web servers[133,136], but these are not easy to connect to our software suite. For example, Prosit was first only available via ProteomicsDB (https://www.proteomicsdb.org/prosit/)[118,134]. To use it in conjunction with MSFragger, users would have to ask MSBooster to produce an input file and upload it manually to the web server

along with their raw file, which can be large, multiple gigabyte files. After waiting in a processing queue, results could be downloaded and read by MSBooster. The process was tedious, certainly not a long term solution. A more feasible approach for model integration could be Python/R packages[135-137,222], but this suffers from each model having its own dependencies with version requirements that necessitate distinct virtual environments. Python installation and environment maintenance can be finicky, with each user having a unique compute environment that makes troubleshooting difficult. Docker containers can provide an enclosed, controlled environment, yet Docker installation presents yet another barrier for non-technical FragPipe users. Furthermore, many models require graphical processing units (GPUs), the workhorse of deep learning, for fast computation, another important consideration in making DL models accessible. Many FragPipe users do not have access to GPUs, and running these models on local CPUs would be prohibitively slow.

Application programming interfaces, or APIs, present a promising solution for MSBooster to access these models. MS2PIP provides a RESTful API (https://iomics.ugent.be/ms2pip/api/v2/), granting any program that can make an HTTP request access to its predictions. As new models are published every month, there is a need for an expedited solution to access the models via APIs. The recent development of the Koina web server and model repository (https://koina.proteomicsdb.org/) provides a central location where any model trained with a major ML/DL framework (e.g. PyTorch/TensorFlow models, XGBoost Python models) can reside[220]. Hosted on a NVIDIA Triton inference server, Koina is a scalable solution for hosting the ever-growing list of proteomics prediction models. The documentation streamlines setting up a model, requiring minimal rewriting from the original model code to the Triton server format; we demonstrate this for the PredFull full spectrum model[223] on a local inference server in

Chapter 4. And not only does Koina make supporting new models in MSBooster easy, but it also allows users to send prediction requests to the public server, alleviating the high resource demands of GPU computing and democratizing DL for all scientists needing peptide predictions. Koina was created with FAIR principles in mind and presents an attractive solution for model hosting and accession, not only for MSBooster, but for many other tools too[147,224,225].

Here, we showcase improvements to MSBooster resulting from access to a greater model collection. After laying out how Koina is integrated into FragPipe, we perform a benchmarking study on HLA, phosphoproteomics, Astral DIA, and TMT data, illustrating how MSBooster with the Koina API can easily test various models to generate guidelines for which ones work best for each datatype. Anticipating an influx of new models, we also present a real-time heuristic approach for finding high performance models for any user-provided data.

3.3 Results

3.3.1 Integration of Koina into FragPipe

The integration of Koina into FragPipe now enables MSBooster to access many more models for prediction besides the previous default, the prediction module of DIA-NN[175], allowing users to compare and mix-and-match models within the same software environment. The FragPipe graphical user interface (GUI) Validation Tab has been extended to include a Koina section (Appendix Figure B.1a). MSBooster takes as input the list of PSMs reported by MSFragger in .pin format (Figure 3.1a) and only requests predictions for those peptides supported by each individual model (Table 3.1). It filters out peptides too long for prediction, and if there are peptides with PTMs unsupported by a model, the unmodified version is queried, and its fragment m/z values are adjusted as previously described[156]. MSBooster can query and parse predictions from Koina for thousands of peptides per second using the default public server

at https://koina.wilhelmlab.org (Figure 3.1b). When timing on a high-performance Linux server, the DIA-NN prediction module runs faster than the Koina models. When timing the models on a consumer laptop, DIA-NN predicts RT slower than the Koina models, and AlphaPeptDeep (hereby shortened to PeptDeep) even predicts MS/MS spectra faster than DIA-NN, suggesting that the performance of DIA-NN is more dependent on the local hardware than Koina is.

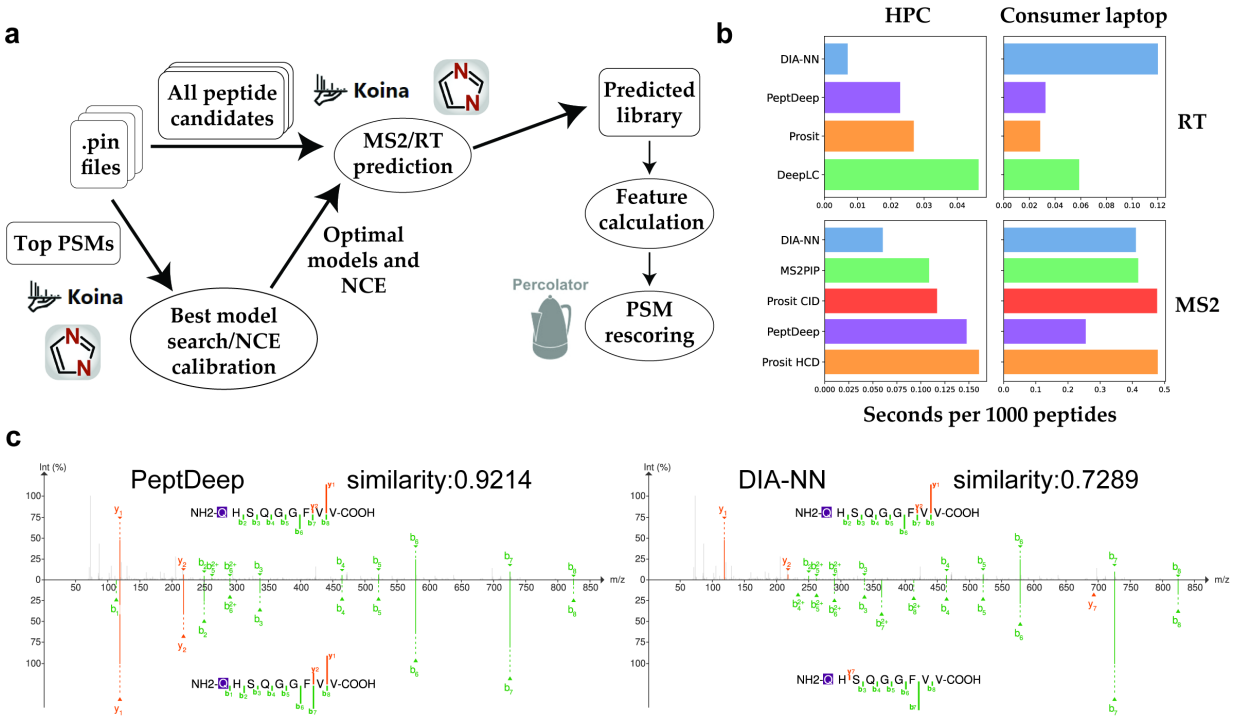


Figure 3.1 Koina integration in FragPipe

(a) MSBooster workflow of FragPipe for PSM rescoring. All peptide candidates are extracted from .pin files and predicted by either DIA-NN or models available on Koina. These models can be specified manually, and NCE calibration is performed automatically if the model accepts NCE as a parameter. MSBooster can also use a heuristic algorithm to automatically choose the best-performing MS/MS+RT model combination. The one thousand top PSMs are selected by default for the best model search and NCE calibration steps. (b) Timing performed using MSBooster. Figures are separated by the type of model (RT or MS/MS) and computing environment used. (c) Spectra 20200317_QE_HFX2_LC3_DDA_CM647_R01.42711.42711.2 from Pak et al., 2021 was visualized using either PeptDeep or DIA-NN predicted spectra, with the experimental spectra on the top and predicted spectra on the bottom. Unweighted spectral entropy similarity was used as the MS/MS similarity metric.

To account for differences between the learned normalized collision energy (NCE) patterns of the predictors and the NCE settings on individual instruments, MSBooster performs an NCE calibration step when using Koina models (Figure 3.1a, Methods Section 3.5.3). To

assess the quality of the rescoring, MSBooster generates multiple QC plots, including one showing the rate of Koina prediction (Appendix Figure B.1b) and one showing the distributions of the similarity scores for PSMs at each NCE value (Appendix Figure B.1c).

	DIA-NN	Prosit	AlphaPeptDeep	MS2PIP	DeepLC
AA length limit	No	30	No	30	60
Phospho	Yes	No	Yes	No	Yes
n-term TMT optional	Yes	No			

Table 3.1 Model prediction limitations

We consider if long peptides (AA length limit), phosphopeptides (Phospho), and peptides without N-terminal TMT labelling (n-term TMT optional) are allowed. This table does not consider the Prosit PTMs model, which was made available on Koina after the time of writing.

A predicted library is saved as a Mascot Generic Format (MGF) file if a Koina model was used or as a binary file if DIA-NN's prediction module was used. The libraries can be loaded into the FragPipe-PDV[186] viewer of FragPipe to compare experimental and predicted spectra (Figure 3.1c). Here, it becomes easier to visualize the differences between models' predictions. For example, DIA-NN does not predict intensities for fragments shorter than three amino acids long, as they are less informative for DIA-NN's peptide-centric approach. Figure 3.1d shows one example in which the pyroglutamated peptide [U:Gln->pyro-Glu]-QHSQGGFVV exhibits a strong y1 fragment intensity, which is matched by PeptDeep's prediction but not by DIA-NN's, resulting in a nearly 0.2 drop in unweighted spectral entropy similarity, which ranges from 0 to 1.

3.3.2 Benchmarking PSM Rescoring using Koina and MSBooster

Koina in MSBooster enables users to compare model performance systematically by keeping all other upstream and downstream data processing steps and parameters the same in FragPipe. Because of differences in model architecture and training data, we hypothesize that the optimal model for rescoring will depend on the specifics of the dataset. To investigate this, we

considered various experiments to pinpoint patterns between specific proteomics data types and optimal models. Specifically, we considered phosphoproteomics data from Arabidopsis thaliana[226] and mouse pancreatic ductal adenocarcinoma cell lines[227], DDA and DIA human leukocyte antigen (HLA) data across different instruments[180,228-230], DIA data generated on the new Orbitrap Astral mass spectrometer (Thermo Scientific)[50,231], and TMT 11 plex-labeled data[232,233]. We considered the spectral similarity and RT difference features calculated by MSBooster individually and together for PSM rescoring and recorded the numbers of peptides identified in Tables 3.2-3.4. We present our detailed findings in the following sections.

	original	DIA-NN_rt	AlphaPept_rt_generic	Deeplc_hela_hf	Prosit_2019_irt	Prosit_2020_irt_TMT
Arabidopsis Phospho	28118±27(100.0%)	28646±27(101.9%)	29082±22(103.4%)	28715±27(102.1%)		
Mouse Phospho	36500±17(100.0%)	36773±16(100.7%)	36839±16(100.9%)	36806±17(100.8%)		
I Marcu	14401±78(100.0%)	15282±72(106.1%)	15200±71(105.6%)	15050±66(104.5%)	14950±85(103.8%)	
II Marcu	10999±26(100.0%)	11262±29(102.4%)	11250±43(102.3%)	11255±23(102.3%)	11288±34(102.6%)	
I Pak DDA	42427±150(100.0%)	45480±102(107.2%)	44882±110(105.8%)	44721±62(105.4%)	44769±69(105.5%)	
I Pak DIA	12775±154(100.0%)	14543±141(113.8%)	14482±95(113.4%)	14208±62(111.2%)	14746±101(115.4%)	
I Ritz DIA	8426±100(100.0%)	9571±170(113.6%)	9234±129(109.6%)	9515±90(112.9%)	9483±149(112.5%)	
I Phulphagar timsTOF	25698±378(100.0%)	28702±89(111.7%)	28977±92(112.8%)	28522±87(111.0%)	28889±85(112.4%)	
II Phulphagar timsTOF	30483±59(100.0%)	31554±49(103.5%)	31387±31(103.0%)	31447±38(103.2%)	31623±32(103.7%)	
timsTOF Meier	64710±17(100.0%)	66366±19(102.6%)	67216±12(103.9%)		65617±21(101.4%)	
Astral DIA Coon	50678±64(100.0%)	57673±74(113.8%)	56218±72(110.9%)	57645±82(113.7%)	56961±46(112.4%)	
Astral DIA Olsen	188249±190(100.0%)	202056±110(107.3%)	200041±152(106.3%)	201965±201(107.3%)	199953±121(106.2%)	
TMT11 Lung APOLLO	91495±41(100.0%)	93876±47(102.6%)				92787±50(101.4%)

Table 3.2 Retention time model benchmarking

Retention time models from Koina are compared with DIA-NN's RT prediction model and running without MSBooster (original). The RT feature used was delta RT loess. Each entry was run 10 times with different Percolator random seeds for training/testing splits. Each cell is formatted as average peptides ± standard deviation (percent improvement over original). Models with the highest percentage gains are highlighted in yellow.

3.3.2.1 HLA immunopeptidomics

It is well established that HLA immunopeptidomics data benefits greatly from PSM rescoring with predicted libraries[135,146,150,153,179,234,235], in part because the larger nonspecific peptide search space leads to many more ambiguous IDs[236]. They have been widely studied because of increased interest in HLA peptides for use as immunotherapeutics[237,238]. Recent developments in DIA and instrumentation (such as the development of the Bruker timsTOF and Thermo Scientific Orbitrap Astral instruments) have provided deeper and more reproducible

views of the HLA peptidome[239,240]. We benchmarked all models on a variety of HLA datasets (Figure 3.2, Tables 3.2-3.4).

	original	DIA-NN_ms2	AlphaPept_ms2_generic (QE or Lumos)	AlphaPept_ms2_generic timsTOF	ms2pip_2021_HCD
Arabidopsis Phospho	28118±27(100.0%)	29283±22(104.1%)	29274±18(104.1%)		
Mouse Phospho	36500±17(100.0%)	37132±16(101.7%)	37152±13(101.8%)		
I Marcu	14401±78(100.0%)	15273±47(106.1%)	16074±38(111.6%)		15122±54(105.0%)
II Marcu	10999±26(100.0%)	11567±24(105.2%)	11701±10(106.4%)		11304±20(102.8%)
I Pak DDA	42427±150(100.0%)	51310±88(120.9%)	53571±45(126.3%)		52242±52(123.1%)
I Pak DIA	12775±154(100.0%)	17894±135(140.1%)	17531±82(137.2%)		16485±79(129.0%)
I Ritz DIA	8426±100(100.0%)	11840±119(140.5%)	11754±91(139.5%)		11214±129(133.1%)
I Phulphagar timsTOF	25698±378(100.0%)	28999±89(112.8%)	30470±71(118.6%)	30226±75(117.6%)	
II Phulphagar timsTOF	30483±59(100.0%)	32599±49(106.9%)	32600±38(106.9%)	32337±31(106.1%)	
timsTOF Meier	64710±17(100.0%)	65964±11(101.9%)	66123±17(102.2%)	66434±20(102.7%)	
Astral DIA Coon	50678±64(100.0%)	60741±52(119.9%)	60586±41(119.6%)	59716±39(117.8%)	56203±49(110.9%)
Astral DIA Olsen	188249±190(100.0%)	214763±140(114.1%)	214399±97(113.9%)	209683±152(111.4%)	203879±82(108.3%)
TMT11 Lung APOLLO	91495±41(100.0%)	93270±53(101.9%)			

	Prosit_2020_intensity HCD	Prosit_2020_intensity CID	Prosit_2023_intensity timsTOF	Prosit_2020_intensity TMT
I Marcu	16020±46(111.2%)	16568±39(115.0%)		
II Marcu	11744±19(106.8%)	11264±23(102.4%)		
I Pak DDA	52777±90(124.4%)	45708±105(107.7%)		
I Pak DIA	18179±100(142.3%)			
I Ritz DIA	12133±132(144.0%)			
I Phulphagar timsTOF	30336±122(118.0%)		30606±80(119.1%)	
II Phulphagar timsTOF	32518±91(106.7%)		33244±30(109.1%)	
timsTOF Meier	65970±20(101.9%)		66521±15(102.8%)	
Astral DIA Coon	60392±37(119.2%)		60169±56(118.7%)	
Astral DIA Olsen	214122±107(113.7%)		213226±104(113.3%)	
TMT11 Lung APOLLO				93306±31(102.0%)

Table 3.3 MS/MS spectra model benchmarking

MS/MS models from Koina are compared with DIA-NN's MS/MS prediction and original.

	DIA-NN + DIA-NN	Tested model combination	Tested model peptides
Arabidopsis Phospho	29507±27(104.9%)	AlphaPept_rt_generic + AlphaPept_ms2_generic	29790±14(105.9%)
Mouse Phospho	37148±17(101.8%)	AlphaPept_rt_generic + AlphaPept_ms2_generic	37280±16(102.1%)
I Marcu	16205±31(112.5%)	DIA-NN + Prosit_2020_intensity_CID	17125±34(118.9%)
II Marcu	11765±15(107.0%)	Prosit_2019_irt + Prosit_2020_intensity_HCD	11840±17(107.6%)
I Pak DDA	52680±77(124.2%)	DIA-NN + AlphaPept_ms2_generic	54421±36(128.3%)
I Pak DIA	19175±107(150.1%)	Prosit_2019_irt + Prosit_2020_intensity_HCD	19575±42(153.2%)
I Ritz DIA	13017±267(154.5%)	DIA-NN + Prosit_2020_intensity_HCD	12748±111(151.3%)
I Phulphagar timsTOF	30923±60(120.3%)	AlphaPept_rt_generic + Prosit_2023_intensity_timsTOF	32362±55(125.9%)
II Phulphagar timsTOF	33471±34(109.8%)	Prosit_2019_irt + Prosit_2023_intensity_timsTOF	34058±42(111.7%)
Astral DIA Coon	64514±41(127.3%)	Deeplc_hela_hf + AlphaPept_ms2_generic	63605±37(125.5%)
Astral DIA Olsen	222999±130(118.5%)	Deeplc_hela_hf + AlphaPept_ms2_generic	222037±77(117.9%)
TMT11 Lung APOLLO	94871±34(103.7%)	DIA-NN + Prosit_2020_intensity_TMT	94991±43(103.8%)
TMT11 Lung APOLLO	94871±34(103.7%)	Prosit_2020_irt_TMT + Prosit_2020_intensity_TMT	94047±20(102.8%)

Table 3.4 RT and MS/MS combined model benchmarking

The FragPipe default (DIA-NN for both MS/MS and RT predictions) was compared against the combinations of other high performing models listed in Tables 3.2 and 3.3. If DIA-NN performed the best in both RT and MS/MS predictions, the second best models were selected for comparison.

Overall, a combination of DIA-NN and Prosit RT models performed the best over baseline (skipping MSBooster), while the Prosit models excelled in the MS/MS category. The largest gain enabled by MSBooster occurred in a DIA dataset[180], where Prosit rescoring added 53.2% more peptide IDs. In addition, we found that combining the best individual RT and MS/MS models outperformed the current FragPipe default of using DIA-NN's predictions to compute both the MS/MS and RT features for all datasets except one (Figure 3.2e). Interestingly, we also note that HLA class I benefited more from predictions than HLA class II, which aligns with previous literature[179] (Figure 3.2a-c, f-g, Tables 3.2-3.4). This is evident across all RT and MS/MS models. A potential reason for this stems from HLA I peptides being more often singly charged than HLA II peptides: singly charged peptides have decreased fragmentation efficiency, causing lower performance of traditional peak counting-based scoring functions.

Our findings highlight how specialized models outperform generic ones. First, we note the differences in performance when rescoring spectra from different fragmentation methods. Distinct collision-induced dissociation (CID, resonance-type CID) and higher-energy collision dissociation (HCD, beam-type CID) models exist for Prosit and MS2PIP. However, only the HCD MS2PIP model was currently available on Koina at the time of writing, while the CID model is on the MS2PIP server at https://iomics.ugent.be/ms2pip/[222]. Therefore, our comparison is limited to using the Prosit models. Marcu et al.[228] used CID to fragment precursors in their class I sample (Figure 3.2a). We found that the Prosit CID model performed best here, achieving 15% more peptide IDs over baseline and outperforming both its Prosit HCD counterpart (11%) and DIA-NN (6%). Likewise, the Prosit HCD model performed better on HCD-fragmented data[180,228] (Figure 3.2b-c). Whichever model is superior is also evident in the distribution of

target PSMs' spectral similarity score distribution, which is shifted to lower similarities when the wrong fragmentation model is used (Appendix Figure B.2).

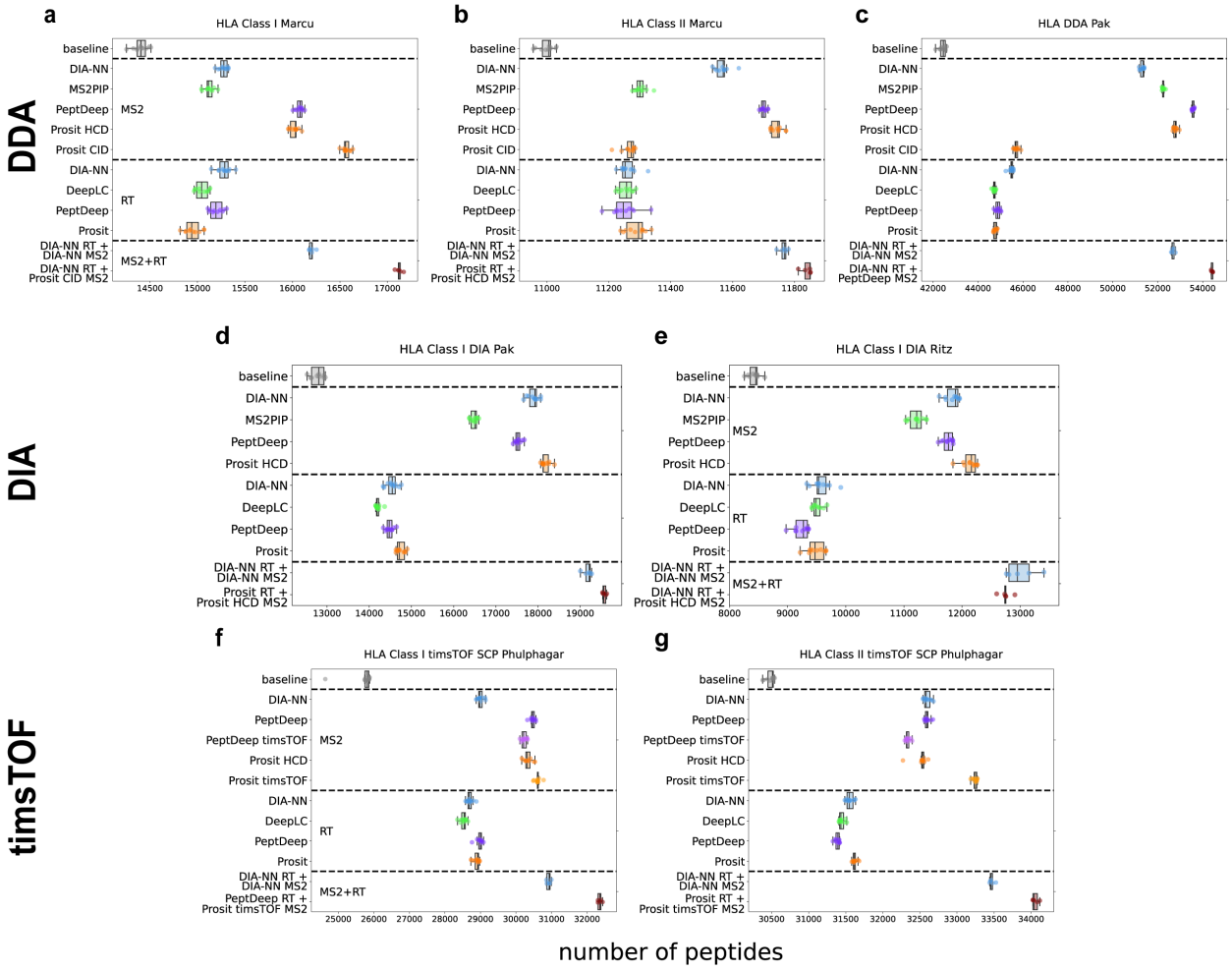


Figure 3.2 HLA rescoring with various models

The datasets used are (a-b) Marcu et al., 2021 class I and II, (c-d) Pak et al., 2021 class I DDA and DIA, (e) Ritz et al., 2017 class I DIA, and (f-g) Phulphagar et al., 2023 class I and II on a timsTOF single-cell proteomics system. "PeptDeep timsTOF" is the PeptDeep model with "timsTOF" set as its instrument metadata, while "PeptDeep" used either "Lumos" or "QE" as its instrument metadata, depending on what was listed in the mzML files or manuscript text of the respective datasets.

Specialized models also improve the rescoring of MS/MS spectra from different mass spectrometers. Strangely, although Prosit timsTOF improves over DIA-NN and the Prosit HCD model trained on Orbitrap data for both class I (Prosit timsTOF: 19.1% improvement over baseline, Prosit HCD: 18%, DIA-NN: 12.8%) and II data from Phulphagar et al.[230], this does not

extend to the PeptDeep models, where "Lumos" PeptDeep predictions identified more peptides than "timsTOF" PeptDeep predictions did (Figure 3.2f-g). We believe this discrepancy may arise because PeptDeep and Prosit were trained on spectra of different NCEs. While Prosit timsTOF was trained on energies ranging from 20-70 eV, PeptDeep timsTOF was only trained on 32-52 eV[135]. More importantly, all timsTOF spectra were annotated as 30 eV during the PeptDeep training phase. Though spectra do differ across Orbitrap and timsTOF instruments, there are certain hotspots across energy levels where they are highly correlated[234,241]. Specifically, precursors fragmented with lower energies (around 20% and 20 eV on Orbitrap and timsTOF instruments, respectively) produce nearly identical spectra across instruments. Phulphagar et al. employed a scheme where precursor IM and fragmentation energy were inversely correlated, with energy ranging from 55-10 eV. Though MSBooster's NCE calibration step seeks the best NCE value for spectral prediction, the lower bound of this experimental NCE range is far below that of what PeptDeep timsTOF was trained on, potentially resulting in suboptimal predictions. However, because low energy fragmented spectra are similar between Orbitrap and timsTOF, the PeptDeep model predicting for a Lumos instrument performs better since it was trained on spectra with HCD energy as low as 20%. Indeed, when we rescored another dataset acquired on a timsTOF with similar ion mobility settings as the data PeptDeep was trained on[65], both PeptDeep and Prosit timsTOF models outperformed their Orbitrap counterparts (Appendix Figure B.3).

To confirm our hypothesis that annotating spectra in the training set with the correct NCE would improve PeptDeep timsTOF performance, we transfer-learned new PeptDeep models using the same datasets Prosit timsTOF was trained on and tested their performance across a range of low collision energy values on HLA class I timsTOF data[230] (Appendix Figure B.4,

Methods Section 3.5.4). Models were trained in one of two ways. The first, named "nce30", was trained by annotating all PSMs as 30eV, as was implemented in PeptDeep's timsTOF training. By excluding informative NCE metadata in the training, this transfer-learned model was still unable to achieve the number of peptides identified by the PeptDeep Lumos model, despite having observed the fragmentation of many new non-tryptic peptides. The second model was dubbed the "variable" model and used the NCE information provided for each peptide. Here, we were able to identify similar peptide numbers to the Lumos model as early as 15eV, yet the variable model was still unable to surpass the Lumos model's performance. We conclude that encoding all training spectra as 30eV resulted in PeptDeep timsTOF learning an average representation of spectra across collision energies that did not extrapolate well to lower energies; when it is properly encoded, NCE influences peptide spectral predictions strongly enough to cause noticeable changes in the final number of peptide IDs, though we were unable to train a PeptDeep timsTOF model that outperformed what was learned by the Lumos model.

Koina models outperformed DIA-NN's predictions the most on the CID and timsTOF datasets. This improved ability to identify peptides provides a compelling reason for incorporating Koina into FragPipe, as models trained on diverse datasets often improve upon DIA-NN's predictions.

3.3.2.2 Phosphoproteomics

Along with DIA-NN, the only Koina models available at the time of writing that support phosphopeptide prediction are PeptDeep and DeepLC. We benchmarked model performance for two datasets. First, we processed phosphoproteome data from 30 different Arabidopsis thaliana tissues[226] and counted how many phosphopeptides were identified. DIA-NN's MS2 feature achieved an average improvement of 3.9% over baseline performance, while PeptDeep's MS2

feature had a 4.1% improvement (Figure 3.3a). PeptDeep's improved MS/MS predictions are also evident in Figure 3.3b, where its predictions for high-scoring "confident target" PSMs are concentrated closer to the maximum similarity of 1. The RT feature tells a similar story (DIA-NN +1.8%, DeepLC +2.3%, PeptDeep +3.7%) (Figure 3.3a).

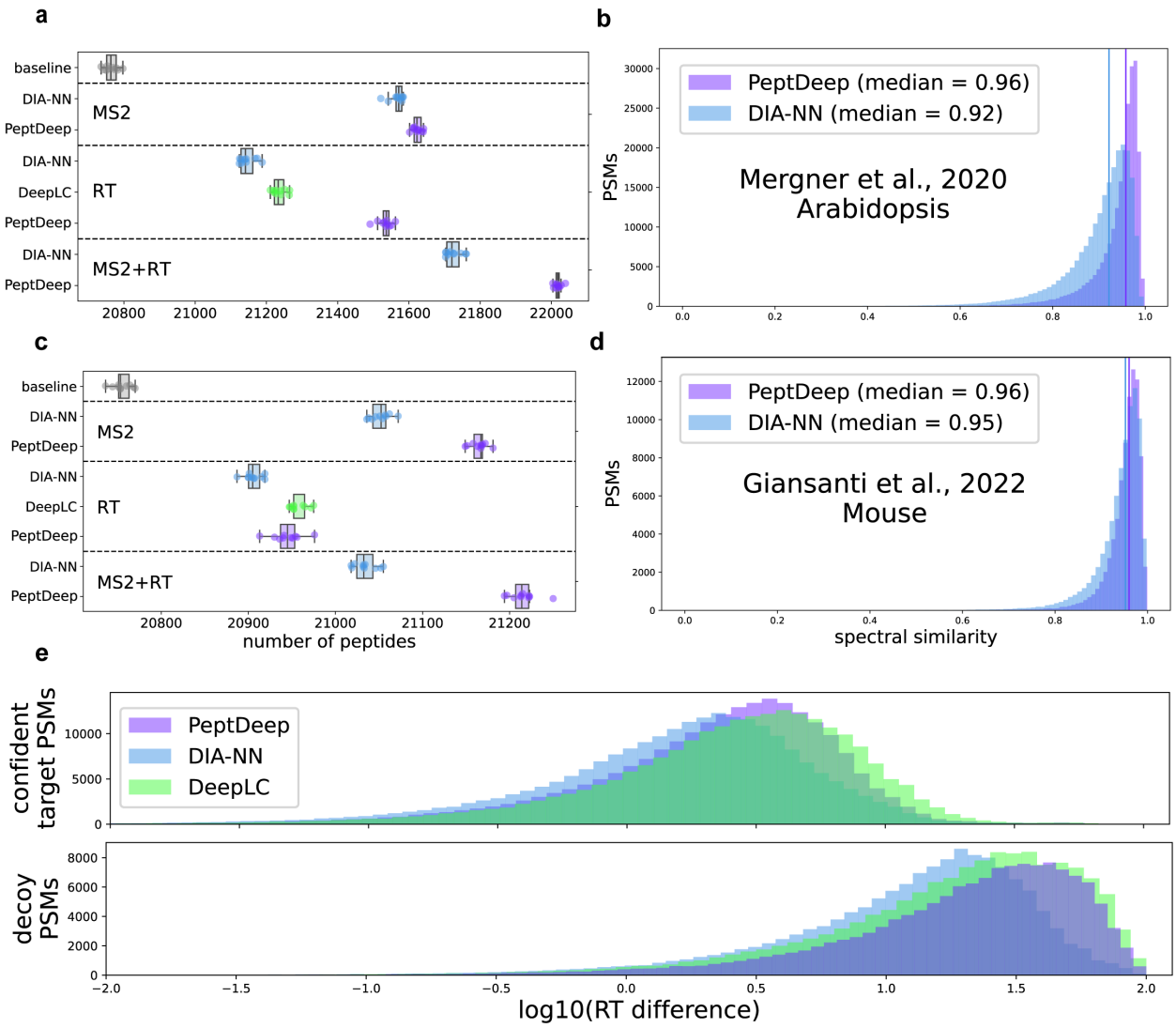


Figure 3.3 Phosphoproteomics rescoring with various models

(a) Numbers of peptides identified using different models on the Arabidopsis thaliana data across 30 tissues[226]. Each model was tested ten times, each time with a different Percolator random seed. "Baseline" is the peptides identified when excluding MSBooster before Percolator rescoring. MSBooster calculated and added the unweighted spectral entropy feature (MS2), delta RT loess feature (RT), or both (MS2+RT). (b) The distribution of spectral similarity scores for confident target PSMs (those PSMs with lower expectation values or "evalues" than the lowest evalue assigned to a decoy PSM in the same pin file). The features were calculated using predictions either from PeptDeep or DIA-NN. (c-d) The same as (a-b) but for the 8 mouse pancreatic ductal adenocarcinoma cell lines[227]. (e) The distribution of delta RT loess scores for confident target and decoy phosphorylated PSMs from the Arabidopsis

thaliana dataset. The log10 of each value plus a small pseudocount was added for clearer visualization of the distribution differences. Features were calculated using predictions from DIA-NN, PeptDeep, or DeepLC.

It is less apparent from the distributions of the RT feature which model performs best (Figure 3.3e). DIA-NN's RT predictions improved IDs the least, yet its confident target PSMs have the smallest RT differences. However, DIA-NN also assigned decoy PSMs smaller RT differences compared to the other models, which decreased target-decoy separation and ultimately led Percolator to find the RT feature less informative. Using the RT and MS2 features together, we achieved a 4.6% increase over baseline using DIA-NN and a 6.0% increase using PeptDeep (Figure 3.3a). This is in line with the magnitude of improvement expected for a closed MSFragger search on tryptic DDA data[156]. Comparison of the three models on 8 mouse pancreatic ductal adenocarcinoma (mPDAC) cell lines[227] showed similar results (Figure 3.3c-d). DeepLC performed similarly to PeptDeep here, identifying 20958 and 20945 phosphopeptides on average, respectively. Based on these findings, we believe that PeptDeep should be used instead of DIA-NN for phosphoproteomics analysis.

3.3.2.3 Astral DIA

Recently, the Orbitrap Astral mass spectrometer (Thermo Scientific) has piqued the interest of the proteomics community for its wide dynamic range, sensitivity of detecting low abundance precursors, and accurate and precise quantification[27,28]. Its higher acquisition rates compared to Orbitrap analyzers allow for narrower isolation windows, producing spectra of reduced complexity in DIA without sacrificing throughput. With the rising interest in narrow window DIA (nDIA), we analyzed two short LC-gradient datasets using 2 Th isolation windows with MSFragger-DIA (Methods Section 3.5.1). The first dataset included three technical replicate injections of Hap1 human cell lysate[231]. As there are no Astral-specific models on Koina, we investigated whether Orbitrap or timsTOF models showed better performance. We

found that for both Prosit and PeptDeep, the Orbitrap mode performed better on this dataset, though only by a small margin (Prosit: 60392 vs 60169; PeptDeep: 60586 vs 59716) (Table 3.3). Regardless, DIA-NN, which is instrument-agnostic, performed the best both for MS2 and RT features. DIA-NN's MS2 and RT features also performed the best when combined. They achieved a 27.3% increase in peptide IDs over baseline. Meanwhile, combining the next best features of DeepLC for RT and PeptDeep for MS/MS only achieved a 25.5% increase over baseline. Similar findings were obtained from the second dataset of 46 fractionations in technical triplicates of HEK293 cells[50]. Here, using DIA-NN for both features increased peptide IDs by 18.5% and outperformed PeptDeep MS2 and DeepLC RT combined (17.9%). This lower percent increase in IDs may be attributed to the fractionated dataset's already deeper proteome coverage at baseline, identifying more than 3 times the number of peptides as the first dataset. From these analyses, we find that Astral nDIA experiments may benefit most simply from DIA-NN predictions, though this leaves room for improvement if an Astral-specific model is trained.

3.3.2.4 TMT

At the time of writing, the only two models available in FragPipe for TMT-based prediction are DIA-NN and Prosit, both of which were trained on hundreds of thousands of sequences[232]. We examined two datasets to see whether these models differed in not only the number of peptide IDs, but also in their quantification quality. We analyzed 36 fractions of one plex of TMT11-labeled peptides from a study of lung adenocarcinoma (LUAD) tumors[233]. On this dataset, Prosit outperformed DIA-NN's MS/MS predictions, but not by a significant amount (93306 vs 93269 TMT-labeled peptides, t-test: $p>0.05$) (Figure 3.4a). Meanwhile, DIA-NN greatly outperformed Prosit's RT predictions (93877 vs 92787, t-test: $p<0.05$). Using both DIA-NN features improves over baseline by 3.7%, in contrast to the 2.8% improvement with Prosit.

Using DIA-NN for RT and Prosit for MS2 prediction, we identified 94991 peptides on average, a 3.8% improvement over baseline and significantly more than using DIA-NN to predict both peptide properties (t-test: $p<0.01$). Combining the DIA-NN and Prosit models is enabled by MSBooster's access to DIA-NN and Koina.

DIA-NN does not assume peptides are TMT-labeled at the N-terminus (Table 3.1). To adapt Prosit predictions to accommodate peptides lacking N-terminal TMT, we simply assign them the RTs of the same peptide sequences but with N-terminal TMT and shift the m/z of their b-ions by the TMT label's mass while retaining the same intensities. To explore how well this strategy works for peptides lacking N-terminal TMT, we compared MSBooster feature scores between peptides with and without the N-terminal label in a subset of confident target PSMs, 2.3% of which were unlabeled at the N-terminus. We found that while both DIA-NN and Prosit exhibit higher accuracy for N-terminally labeled peptides, Prosit exhibited more noticeable differences between the two groups for both the RT difference and MS2 similarity features (Figure 3.4c-h). In addition, Prosit does not support prediction for peptides longer than length 30. MSBooster assigns spectral similarity values of 0 and predicted indexed RTs (iRTs) of 0 to PSMs without peptide predictions, further penalizing Prosit's performance; this is most clearly seen in Figure 3.4d where many confident target PSMs having spectral similarities of 0, and in Figure 3.4g with the horizontal line (around 55 calibrated predicted RT units) of PSMs straying far from the RT calibration curve. In contrast to this "regular" search, we ran a "restricted" database search, setting N-terminal TMT as a fixed modification and limiting the peptide digest length to 30, to determine if the two models performed more similarly without these biases. Without MSBooster's features, the baseline number of TMT-labeled peptides decreased by more than 2000 (Figure 3.4b). Again, we found that DIA-NN performed best for the RT feature and

Prosit for the MS2 feature, though the gap in RT feature performance decreased (Prosit regular vs restricted: 1.4% vs 2.0% increase over baseline; DIA-NN: 2.6% vs 2.7%) and the gap in MS2 feature performance increased (Prosit regular vs restricted: 2.0% vs 2.0%; DIA-NN: 1.9% vs 1.7%). This suggests that when N-terminal TMT labeling is incomplete or when many longer peptides are expected, the flexibility of DIA-NN is beneficial.

When analyzing a separate TMT11-labeled human-yeast protein mixture[232], our initial findings that DIA-NN was best for RT and Prosit for MS/MS did not hold. This dataset had four settings combining either HCD or CID fragmentation with Orbitrap (OT) or ion trap (IT) analyzers. Here we found that DIA-NN and Prosit performed comparably in the number of TMT peptide IDs (Figure 3.5a). Interestingly, though Prosit TMT considers the fragmentation mode during prediction and DIA-NN does not (nor was it trained on any CID-fragmented TMT data), the models performed comparably on CID data. N-terminal labeling rates for confident target PSMs were 0.05%, 0.04%, 0.04%, and 0.05% for CID IT, CID OT, HCD IT, and HCD OT, respectively.

We next assessed whether the unique PSMs identified by the two models differed by their reporter ion intensities. In the LUAD dataset, DIA-NN found 2744 unique TMT PSMs to Prosit's 1392 (Figure 3.4i). We compared the pooled channel's MS2-based reporter ion intensities between the unique sets and found a statistically significant difference between their median intensities (DIA-NN: 1.58e4, Prosit: 1.22e4; $p<0.01$, Mann-Whitney U test). However, given the wide dynamic range of reporter ion intensities spanning five orders of magnitude in this dataset, this difference is not particularly meaningful. Indeed, none of the four MS3-based acquisition methods in the yeast dataset showed a significant difference between DIA-NN's and Prosit's quantification (Figure 3.5b-e) ($p>0.01$, Mann-Whitney U test).

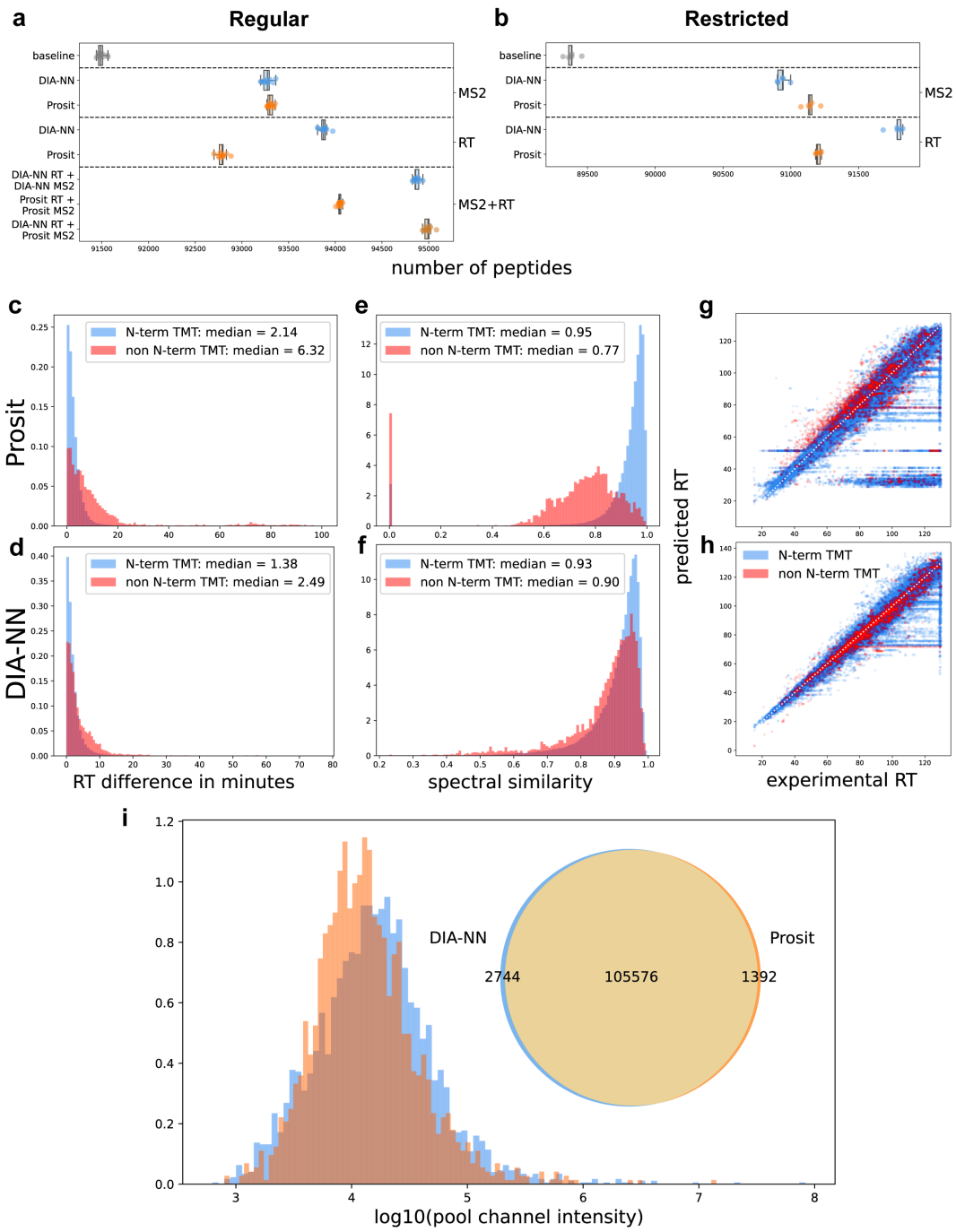


Figure 3.4 Comparison of TMT models for the LUAD dataset

(a) Number of TMT11-labeled peptides identified with 10 iterations of PSM rescoring. (b) The number of TMT peptides identified when the database search is “restricted”. In-silico digest was done only up to 30 amino acids long, and N-terminal TMT was a fixed modification. (c-f) Histograms depicting feature scores for confident target PSMs separated by if they are N-terminally labeled (blue) or not (red). c-d depicts the RT difference score, while e-f depicts the spectral similarity score. (g-h) Scatterplots of confident target PSMs. The x-axis denotes the experimental RT in minutes, while the y-axis shows the predicted RT, after it has been calibrated to the same scale as the experimental RT. Predictions are from Prosit (g) or DIA-NN (h). A white dotted line of the equation x=y shows where a perfect alignment of experimental and predicted RT would lie. (i) 2744 and 1392 unique TMT PSMs

were identified by DIA-NN and Prosit. The 126 channel was a pooled sample and non-zero reporter ion intensities from this channel are visualized here.

Finally, we checked the validity of the unique peptides from each model. The yeast dataset is a good benchmark because each channel has a known amount of yeast protein spiked into a constant amount of human protein, resulting in the expected ratios of 1:1:1:2:2:2:8:8:8:0 for yeast proteins in channels 127N to 131C compared to the 126 channel. Therefore, we focused on peptides originating from the yeast proteome to assess quantification accuracy.

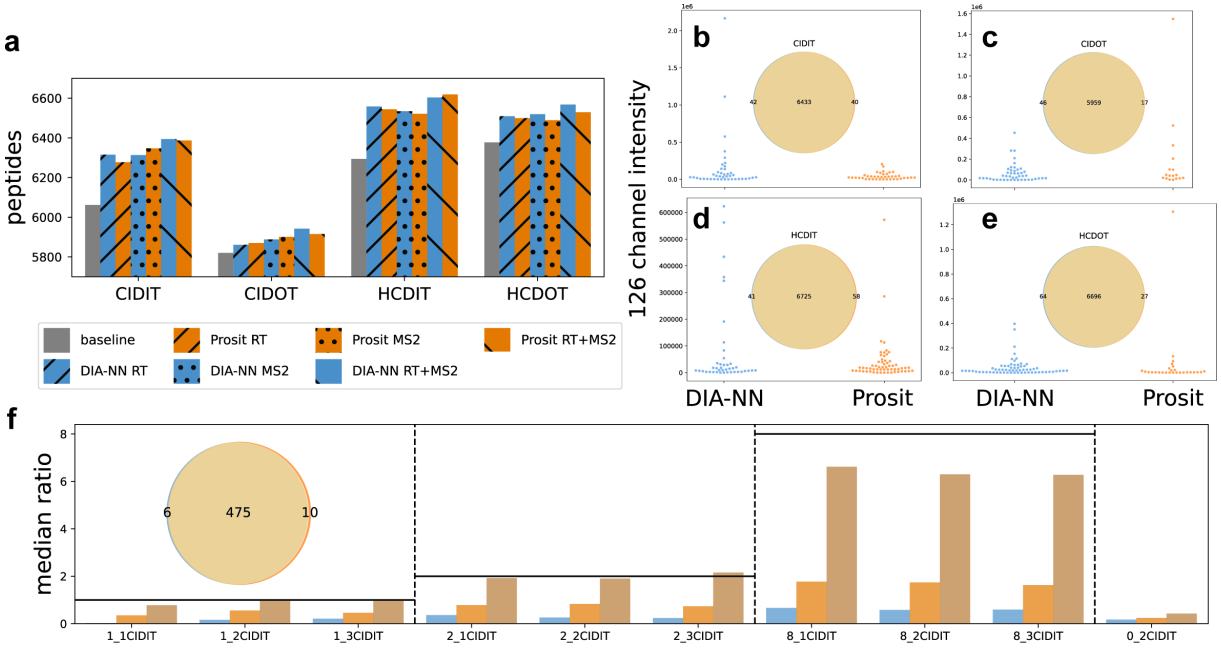


Figure 3.5 Comparison of TMT models for the human-yeast dataset

(a) TMT11 peptides identified in a yeast/human proteome mixture. Four combinations of settings were produced by matching CID/HCD fraction with orbitrap (OT) or ion trap (IT) analyzers. Runs using DIA-NN's predictions are in blue and runs using Prosit's are in orange, with separate hatch marks designating whether RT, MS2, or both feature types were used. (b-e) Unique PSMs from each model were selected and the reporter ion intensities in the 126 channel were visualized in a swarmplot. Both human and yeast PSMs were included here. (f) Quantified CID IT yeast peptides from DIA-NN and Prosit were compared via Venn diagram. For each subset of this Venn diagram, the median ratio of the peptides in each channel to the reference 126 TMT channel is visualized, with a horizontal black line indicating what the expected ratio is.

In the data acquired by CID IT, the two models had 475 quantified yeast peptides in common (Figure 3.5f). DIA-NN and Prosit had 6 and 10 unique quantified peptides, respectively. The shared, DIA-NN-specific, and Prosit-specific sets all showed increasing median ratios across the triplicate channels, with the shared set most closely following that

expected 1:2:8 ratio. The average median ratio for DIA-NN was 0.13, 0.29, and 0.61; for Prosit it was 0.45, 0.78, and 1.71; for the shared set it was 0.93, 1.91, and 6.40. Similar trends exist in the three other acquisition methods (Appendix Figure B.5). Though peptides solely identified by Prosit seemed to exhibit higher median ratios than those from DIA-NN, these differences were not statistically significant ($p>0.01$, Mann-Whitney U test). The unique peptides from both models displayed much lower median ratios than was expected. While some may be false positives, most have high spectral similarities, suggesting that these are real IDs (Appendix Figure B.6). Potential explanations for these reduced ratios include TMT reporter intensity suppression and insufficient unique peptides in each group making it difficult to achieve an accurate median ratio. Overall, most yeast peptides found by DIA-NN and Prosit overlapped, and many that are unique to each model are likely valid. We find insufficient evidence that these unique sets are different in any meaningful way. Therefore, we suggest that these models can be used interchangeably, but DIA-NN should be favored when there are many PSMs matched to peptides longer than 30 amino acids or with unlabeled N-termini.

3.3.3 A Heuristic Approach to Optimize Model Selection in MSBooster

We have observed empirically which models work best for phosphorylation, HLA, Astral DIA, and TMT data. However, it is outside the scope of this manuscript to exhaustively consider all types of proteomics data, and even within the same type of data, there is variability in the best-performing models. As Koina continues to grow and support more models, it may become overwhelming for users to determine which combinations of models work best in their data. To ease this choice overload, we have incorporated an optional module in MSBooster that attempts to determine the best MS/MS, RT, and IM models for a dataset. Importantly, the selected models may be from different frameworks, a benefit conferred by Koina's one-stop model shop.

Rather than have users rescore their entire dataset using each available model, we implemented a heuristic approach that selects models based on the agreement between their predicted and experimental values on a subset of the data (Methods Section 3.5.5). When developing this algorithm, we initially chose the MS/MS model with the highest median similarity and the RT model with the lowest median RT difference but found that the algorithm would sometimes report the wrong models, according to our empirical results. Figure 3.6 depicts one such case for an HLA DDA class I dataset[228]. Though our algorithm correctly chose Prosit CID as the best MS/MS model (Figure 3.6a), it also chose Prosit rather than DIA-NN as the best RT model (Figure 3.6b). Though Prosit had the lowest median delta RT (1.38 minutes) and DIA-NN the largest (2.03 minutes), it was DIA-NN's RT predictions that resulted in the most peptide IDs on average, with a 6.1% increase over baseline. Compare this to Prosit which performed the worst, with a 3.8% improvement over baseline and 312 fewer peptides than when using DIA-NN's predictions. Appendix Figure B.7 shows similar analyses for the other datasets.

It was clear that the median metric often produced incorrect predictions for the RT model search, so we tested several other metrics and summarized their performance across all datasets used here (Figure 3.6c-d, Methods Section 3.5.5). The "top consensus" metric most often found the optimal RT model. By taking the ten largest delta RT values, the algorithm could better determine the best-performing RT model than the median method could. The top 10 consensus method achieved a "heuristic summary score" of 0.899, while the median method scored 0.793 (Methods Section 3.5.5). The best MS/MS method was achieved by taking the bottom 100 lowest spectral similarity scores (0.981 heuristic summary score), but the median similarity method performed very similarly (0.978), and the methods chose different models in only two out of the thirteen datasets. Therefore, there was insufficient justification to move away from the

median method, so we continue to use it to determine the best MS/MS model. The final metrics chosen were “median” for MS2 models and “top consensus” with 10 PSMs for RT models.

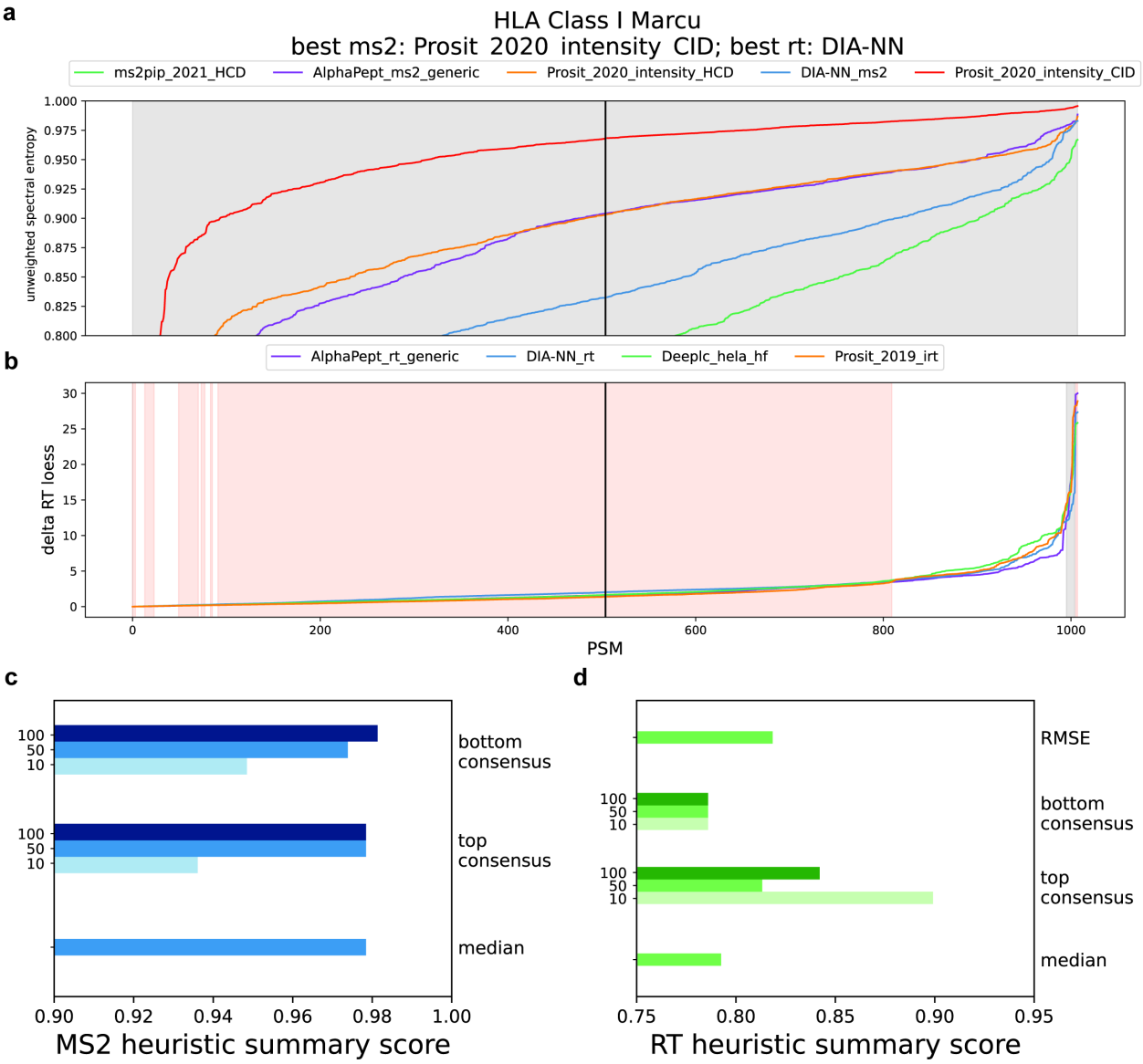


Figure 3.6 Heuristic best model search

Predicted MS/MS spectral similarity (a) and delta RT values (b) for the top 1000 PSMs of an HLA DDA class I dataset (Marcu et al., 2021) were compared to their observed values. The calculated spectral similarity and RT difference scores were ordered in increasing order and plotted for each model. The black line indicates the position of the median value. Grey regions indicate positions at which the model that empirically performed the best would be selected by the heuristic. Red regions indicate positions at which the model selected by the heuristic does especially worse than the empirically best model (i.e. one standard deviation above the mean performance of the heuristic-selected model is less than one standard deviation below the mean performance of the empirically best model). The performance of multiple heuristic methods was summarized across all datasets for MS2 models (c) and RT models (d). For the top and bottom consensus methods, the values of N PSMs tested were 10, 50, and 100 PSMs. Greater detail explaining how the summary score was calculated is provided in Methods Section 3.5.5.

We tested PSM rescoring using all models together by calculating a separate score for each applicable model and combining them in Percolator rescoring. Using multiple MS/MS and RT models resulted in even more peptides than when using a single model each for MS/MS and RT (Appendix Figure B.8). Single models recovered 42-86% of the total peptides added on top of DIA-NN's prediction by multiple model mode. This suggests that for most cases here, a single well-chosen model identified the plurality of peptides without spending the time to predict all peptide candidates multiple times. A mode leveraging the predictive power of multiple Koina models simultaneously can be added to MSBooster in the future.

3.4 Discussion

In this study, we illustrated how Koina's integration in FragPipe enables a benchmarking study of state-of-the-art models without having to run each model in a separate analysis pipeline, thereby isolating predictor performance from confounding factors such as database search and FDR control. We gained insights into which models perform well on specific datatypes and touched on important considerations for training and evaluating these models. Our results align with a recent study that evaluated strategies for training peptide-property models with scarce data[242], suggesting that single-task learning outperforms multi-task learning. While it is convenient to have a single model to handle all predictions, creating multiple specific models allows for greater flexibility in learning weights specific to the data at hand.

Our results in Table 3.4 show not only that Koina models often outperform DIA-NN, but also that models from different frameworks often produce the best results, highlighting the importance of storing diverse models in one central location. They can easily be combined in FragPipe with a dropdown menu, allowing users to experiment with all their favorite models.

Interestingly, the only datasets in our testing where DIA-NN reigned supreme were DIA datasets, suggesting that it was optimized for DIA data during its training.

Anticipating the future availability of many additional models, we developed a heuristic best model search module in MSBooster. We found that the model with the highest MS/MS median similarity or lowest median RT difference did not always lead to the best PSM rescoring results. We caution readers that a model may outperform other models in popular metrics, but this does not mean it can be assumed to be the best model for PSM rescoring or for any task that relies on peptide predictions. These findings also inform the choice of loss functions used when training new prediction models. New RT models could show better performance when trained with a loss function that focuses more on overall or outlier performance. Optimizing predicted spectral libraries for PSM rescoring is a complex problem that benefits from examining ML-based feature scores for high-scoring targets and decoys, as well as those PSMs on the border of the FDR cutoff. An algorithm that consider both targets and decoys, such as one inspired by the mixture model approach of PeptideProphet[86], may provide a better estimate of target-decoy separation and improve our heuristic approach to model-searching.

Finally, we showed that using multiple models to calculate the same similarity features can further increase peptide IDs. Our heuristic search algorithm assigns a score to each model, so it is possible to extend MSBooster to calculate the same features for multiple models for a dataset. Thousands of peptides can be predicted, processed into data structures, and stored in memory per second in MSBooster, meaning that the added time of using multiple models is negligible. A potential side effect of including highly linearly correlated features may be less intuitive Percolator feature weights. Nonlinear methods such as those in Mokapot may remedy this[203].

3.5 Methods

3.5.1 FragPipe analysis and database searches

All analyses were done using FragPipe 21.0 with MSFragger 4.0[69] for database searching, DIA-Umpire for pseudo-MS/MS spectra generation for HLA DIA searches[184], Percolator 3.6.4[87] for PSM rescoring, PTMProphet for phosphorylation localization[243], ProteinProphet[95] for protein assignment, Philosopher 5.1.0[93] for FDR filtering and reporting, and IonQuant 1.10.15[244] and TMT-Integrator 5.0.7[245] for TMT quantification and summarization. MSBooster 1.2.2[156,220] was used to turn deep-learning predictions into features before PSM rescoring, while an updated MSBooster 1.2.30 was used for timing and to demonstrate heuristic model searching. Databases for searches were downloaded within the FragPipe GUI from UniProt with reviewed Swiss-Prot sequences only, and all included common contaminants and reversed sequence decoys. The human database was downloaded on March 18, 2022 (20410 entries); mouse December 19, 2023 (25658 entries); and Arabidopsis June 2016 (the fasta available from Mergner et al., 2020; 48477 entries).

All workflows were adapted from those available in FragPipe, with specific workflows and parameters for each tool provided at https://github.com/Nesvilab/MSBooster/tree/master/Koina%20manuscript%20resource. Data was searched using workflows adapted from those packaged with FragPipe. We relied on FragPipe headless mode (https://fragpipe.nesvilab.org/docs/tutorial_headless.html) for analyses. In-silico digest produced peptides starting from length 7 to length 25 for non-specific searches or 50 for tryptic searches. 2 missed cleavages were allowed except for in Astral DIA searches, where 1 missed cleavage was allowed. 150 top intensity peaks were retained for tryptic searches, 300 for non-specific, and 500 for Astral DIA searches. In HLA searches, experimental fragment peak intensities were square-root

transformed. Carbamidomethylation of cysteine was a fixed modification except for those HLA searches in which there was no alkylation step, and methionine oxidation and N-terminal acetylation were set as variable modifications. Pyro-glutamation of glutamine and glutamic acid were included as variable modifications in HLA searches. TMT searches used a fixed lysine modification and variable N-terminal and serine modifications for the TMT11 tag.

MSBooster was run on the command line to produce output files specific to each prediction model. Version 1 of the Koina models were tested. Critically, the "editedPinSuffix" parameter (Appendix D) was set to an empty string so that the next processing steps would recognize the pin files. Percolator was run with "--subset-max-train 500000" for Astral DIA data only. Results were filtered with Philosopher at 1% FDR at the PSM and peptide levels. 1% protein FDR filtering was also applied for all searches besides HLA searches.

3.5.2 MSBooster speed benchmarks

Timing was recorded on two computers: (1) Linux server: Intel(R) Xeon(R) Gold 6354, 3.00 GHz, 36 cores, 72 logical processors, and 755GB RAM; (2) Windows laptop: Intel(R) Core(TM) i7-9750H, 2.60 GHz, 6 cores, 12 logical processors, and 16GB RAM. Each dataset was predicted ten times using multithreading with 55 threads as specified in the MSBooster parameter file, and the average timing per 1000 peptides is shown in Figure 3.1b. The maximum heap space was set with the Java -Xmx parameter, either as 256GB on the Linux server or 8GB on the Windows laptop.

3.5.3 NCE calibration in MSBooster

The top 1000 PSMs ranked by expectation value across all pin files are extracted and predicted at all integer values in a certain range (20-40%, by default). Predicted and experimental spectra are compared with the unweighted spectral entropy similarity metric[156,215], and the NCE

value that produces the highest median similarity is selected when calling the model to predict all other peptide candidates from the pin files.

3.5.4 AlphaPeptDeep timsTOF transfer learning

Transfer learning a timsTOF model was accomplished with peptdeep 0.2.0. Nontryptic data from Adams et al. was downloaded from PXD043844 and their reanalysis of tryptic data was downloaded from MSV000092462[234]. Each PSM (peptide sequence, fragment masses, and fragment intensities) was extracted from the hdf5 files from PXD043844. PeptDeep requires certain formatting of the input data for training, including MaxQuant msms.txt files. Only charge 1 fragments were listed in these msms.txt files. Fragments and intensities from the hdf5 files, which include charge 2 fragments, were used to replace those listed in the msms.txt files for each PSM, as designated by a distinct raw file/scan number pair.

Two models were trained. The "nce30" model was trained as described in the PeptDeep supplementary data[135], where all PSMs were labeled as 30eV for timsTOF training. The "variable" model was trained using the average collision energies extracted from accumulatedMsmsScans.txt files. Both models were trained for 20 epochs with batch size of 1024, learning rate of 1e-4, 5 warm-up epochs and dropout of 0.1.

We predicted spectral libraries using the NCE30 and variable models from 10-30eV. Input files containing the peptides to predict were generated by MSBooster. The resulting libraries were saved in mgf files and provided to MSBooster to add the spectral similarity score to the pin files for HLA class I timsTOF data38. Downstream analysis by Percolator, ProteinProphet, and Philosopher filtering was repeated thrice for each NCE.

3.5.5 Heuristic best model search scoring

Multiple methods were developed to approximate the best model for a dataset without having to rescore all PSMs and process them through downstream tools. First, the 1000/*P* PSMs with the lowest expectation values (as calculated by MSFragger using the database search hyperscore[69]) are selected from each pin file, where *P* is the number of pin files. Next, all PSMs are sorted in ascending order of their unweighted spectral entropy (MS/MS) or delta RT loess (RT) score. Then, each heuristic method is applied. Loess alignment is applied to mitigate the effects of the experimental setup on the predictions.

- Median: The median score calculated using the predictions of each model is determined. The MS/MS model with highest median similarity and RT model with lowest median RT difference are chosen as the best models.
- Top consensus: N PSMs with the largest values are selected and sorted. Largest values for the PSM features when testing MS/MS models mean those with the highest spectral similarity; when testing RT models, large values mean the greatest deviations from the RT calibration curve. The values of N tested are 10, 50, and 100. At every position from 1 to N, the model with the best PSM feature value (greatest MS/MS similarity or smallest RT deviation) gets a vote. The model with the most votes is selected.
- Bottom consensus: This method is like the top consensus method, but it focuses on the PSMs with lowest MS/MS similarity and smallest RT deviation.
- RMSE: The root mean squared error of RT deviations.

Once each heuristic method had produced its pick for best model, the average number of peptides identified empirically across 10 runs with different Percolator random seeds was divided by the average peptides identified by the model with the highest average. This was

applied separately to RT and MS/MS models. The square of this ratio was averaged across multiple datasets to produce the final heuristic summary score shown in Figures 3.6c-d. The average squared ratio value, though less intuitive than an average ratio value, was better suited for our need to apply a heavier penalty to smaller ratios. For phosphoproteomics datasets, all peptides were considered, not only phosphopeptides. For TMT datasets, all peptides were considered, not just those quantified.

To determine the regions of gray, red, and white in Figures 3.6a-b, the model with best spectral similarity/RT score was recorded at each position of the sorted lists. If the model with the best score at a PSM position was also the best PSM rescoring model, this position is assigned the color gray. If the model with the best score at a position does not match the best PSM rescoring model, the position will either be colored white or red. Red indicates that the mean peptides identified by this model plus one standard deviation is less than the mean peptides for the empirically best model minus one standard deviation, based on results listed in Tables 3.2 and 3.3. White indicates that the best model at the position and best PSM rescoring model have comparable performance (i.e. the mean peptides for this model plus one standard deviation is greater than or equal to the mean peptides for the empirically best model minus one standard deviation).

3.6 Data availability

Data for the FragPipe analysis was downloaded from the ProteomeXchange Consortium with accession codes PXD013868, PXD030983, PXD020186, PXD022950, PXD010012, PXD049028, PXD046372, and PXD036025, and from the MassIVE repository with codes MSV000081439 and MSV000091456.

3.7 Code Availability

A tutorial for using Koina in FragPipe can be found at https://fragpipe.nesvilab.org/docs/tutorial_koina.html.

3.8 Authors, affiliations, and contributions

Ludwig Lautenbacher[1,2]; Kevin L. Yang[3]; Tobias Kockmann[4]; Christian Panse[4,5]; Wassim Gabriel[1]; Dulguun Bold[1]; Elias Kahl[1]; Matthew Chambers[6]; Brendan X. MacLean[6]; Kai Li[3]; Fengchao Yu[7]; Brian C. Searle[8,9,10]; Damien Beau Wilburn[10]; Mohammad Reza Zare Shahneh[11]; Yuhui Hong[12]; Haixu Tang[12]; Mingxun Wang[11,13]; Ralf Gabriels[14,15]; Robbin Bouwmeester[14,15]; Robbe Devreese[14,15]; Jesse Angelis[1]; Eduard Sabidó[16,17]; Tobias K. Schmidt[18]; Alexey I. Nesvizhskii[3,7]; Mathias Wilhelm[1,2]

These authors contributed equally: Ludwig Lautenbacher; Kevin L. Yang
These authors jointly supervised this work: Alexey I. Nesvizhskii, Mathias Wilhelm

[1]Computational Mass Spectrometry, Technical University of Munich (TUM), Freising, Germany
[2]Munich Data Science Institute, Technical University of Munich, 85748, Garching, Germany
[3]Gilbert S. Omenn Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA
[4]Functional Genomics Center Zurich (FGCZ) - University of Zurich | ETH Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
[5]Swiss Institute of Bioinformatics (SIB), Quartier Sorge - Batiment Amphipole, CH-1015 Lausanne, Switzerland
[6]Department of Genome Sciences, University of Washington, Seattle, WA 98195
[7]Department of Pathology, University of Michigan, Ann Arbor, MI, USA
[8]Pelotonia Institute for Immuno-Oncology, The Ohio State University Comprehensive Cancer Center, Columbus, Ohio
[9]Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio
[10]Department of Chemistry and Biochemistry, The Ohio State University, Columbus, Ohio
[11]Department of Computer Science, University of California Riverside, Riverside, CA, USA
[12]Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, IN 47408, United States
[13]Virtual Multi-Omics Laboratory, The Internet, Riverside, CA, USA
[14]VIB-UGent Center for Medical Biotechnology, VIB, 9052 Ghent, Belgium
[15]Department of Biomolecular Medicine, Ghent University, 9052 Ghent, Belgium
[16]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Dr. Aiguader 88, Barcelona 08003, Spain
[17]Universitat Pompeu Fabra (UPF), Dr. Aiguader 88, Barcelona 08003, Spain
[18]MSAID GmbH, Garching, Germany

L.L. developed Koina and the complementary Python and Bioconductor packages. K.L.Y. and F.Y. developed the integration with FragPipe and performed the benchmarking of different models using FragPipe. D.B. developed a prototype of Koina. L.L. & E.K. developed the

documentation webpage. M.C. and B.M. developed the integration with Skyline. K.L. added support for Koina predictions in FragPipe-PDV. B.C.S. developed the integration with EncyclopeDIA. L.L., T.K., C.P., W.G., T.S., D.B.W., M.S., Y.H., H.T., Mi.W., R.G., R.V., R.D., J.A. contributed models to Koina. T.S. conceived the approach. A.N., M.W. supervised the research, providing guidance throughout the project. L.L. and K.L.Y. wrote the manuscript and prepared visualizations. All authors reviewed, edited, and approved the manuscript.

Chapter 4 Peptide-Spectrum Match Rescoring with Full Spectrum Prediction

4.1 Abstract

As the main products of collision-induced dissociation fragmentation, b- and y-ions account for around half the total ion current of the average peptide MS/MS spectra in tandem mass spectrometry experiments. Accordingly, peptide sequence determination primarily relies on these ion types. This ignores a large portion of the remaining peaks, many of which are legitimate fragment ions produced as fragmentation byproducts. Many machine learning models can accurately predict b- and y-ion intensities which are used to calculate similarity metrics to better differentiate correctly and incorrectly assigned peptide-to-spectrum matches (PSMs). Few models cover the myriad of other known fragment ion types, nor have they yet to be regularly used to improve peptide identification. We hypothesize that "full spectrum" models can provide additional benefits over only using b- and y-ions for PSM rescoring in immunopeptidomics experiments, in which peptides are known to generate internal ions. Here, we compare two full spectrum models against each other and traditional b/y-ion predicting models, assessing their prediction accuracy and ability to increase immunopeptide IDs. After, we characterize the "dark fragmentome" of predicted peaks of unknown origin, testing for the prevalence of unannotated predicted peaks and whether they hold any additional value for PSM rescoring. Our novel rescoring method is available for beta testing in the command line version of MSBooster.

4.2 Introduction

So far, our discussions of PSM rescoring have been limited to b- and y-type ions for HCD and CID fragmentation. Indeed, b/y-ions are prominently displayed in CID fragmentation

of peptides[30,31], and they are indispensable for sequence determination. Despite this, b/y-ions only account for an estimated 45-64% of total ion current in the typical unmodified peptide spectra[223,246,247]. The remaining ion current is attributed to a myriad of other fragment types, produced by the interesting chemistry of fragmentation and the mobile proton model[130,248,249]: "neutral losses" of ammonia and water are common, manifesting as losses of 17 and 18 Daltons in singly charged fragments[250]. The extent of neutral loss occurrence is highly dependent on proton mobility, with proline residues in particular promoting neutral losses[251]. "Immonium ions" are low mass ions originating from side chain residues[252-254]. "Internal ions," as the name suggests, are internal subsequences of the peptide.

Non-b/y ions (now referred to as "auxiliary" ions) provide valuable information for spectral interpretation, prompting a discussion of how best to leverage their information. Both manual and automated computational approaches for PSM annotation have benefitted from their consideration. One notable case is MSFragger-Labile[255], a mode of MSFragger that considers the unique fragmentation characteristics of labile PTMs, those whose partial or entire mass can separate from its bound amino acid during fragmentation. It can use "diagnostic ions" (partial fragments of PTMs) to trigger a labile search, which uses fragment and peptide remainder ions for database search scoring and PTM localization. Consideration of neutral losses and diagnostic ions can substantially improve peptide identification rates – adding CN_2H_2 loss to a mouse liver tissue search with R-linked ADP-ribosylation increased PSM IDs by 10%. PTM-specific neutral losses are extensively characterized and can be used to confirm the presence of said PTMs[221,252,256]. Immonium ions help disambiguate the presence of single amino acids, including differentiating isobaric leucine and isoleucine residues[257] and localizing phosphorylation events to tyrosine[252]. Internal ions are particularly interesting in immunopeptidomics searches: produced

by endogenous protease degradation, HLA peptides often lack a C-terminal basic residue and do not produce a strong y-ion ladder[258]. These peptides frequently produce internal ions from two sequential fragmentation events. The Spectrum Mill MS Proteomics software suite (https://proteomics.broadinstitute.org/millhome.htm) uses internal ions to fill out their ion ladder, counting cleavage sites from y-, b-, and internal ions to calculate a backbone cleavage score to use as a quality control metric enforcing the presence of stretches of unambiguous sequence. Furthermore, internal ions are abundant in top-down proteomics spectra; one study highlighted the example of an MS2 spectra of the helicase protein, which had 550 N/C-terminal fragment ions and more than 34,000 internal fragments[259]. Internal ions have found success in protein scoring for top-down experiments[260,261].

Many algorithms benefit from including auxiliary ions, but in other cases, auxiliary ions are computationally removed from the recorded spectra. For example, MSFragger removes water and ammonia neutral loss peaks during writing of calibrated mzML files. This can decrease the likelihood of random peak matches to the wrong peptide candidates, not giving decoy and false target peptides a chance to increase their search scores.

Predicting auxiliary ion intensities could improve peptide identification rates, increasing specificity compared to simply confirming their presence in spectra. Before testing this hypothesis, we need spectral prediction models that consider the full spectrum of ion types. Two popular "full spectrum" prediction models are PredFull[223] and UniSpec[247]. PredFull discretizes the 0-2000 m/z range into 0.1 Dalton bins and predicts the intensity of each bin. In contrast, by searching HCD spectral libraries from the National Institute of Standards and Technology (http://peptide.nist.gov/), Lapin et al. found a set of peaks that appeared at least 100 times in the libraries, comprised of the fragment ion types previously mentioned and many isotopic peaks;

UniSpec predicts this set of 7919 peaks that annotate 99.8% of the total peaks in their libraries. While UniSpec explicitly predicts fragment ions known to be produced during peptide fragmentation, PredFull takes a more hands-off approach, allowing the model to discover fragmentation products by itself. DL models excel at finding patterns in unstructured data; for the spectral prediction task, PredFull could potentially discover ion types not specified by UniSpec. While these two models independently demonstrated their accuracy, they have yet to be compared against each other.

Here, we perform initial tests of how we can use auxiliary ions for PSM rescoring. We focus on HLA peptides, a promising candidate given their prominent role when searching immunopeptidome data with Spectrum Mill. After implementing support for UniSpec and PredFull in MSBooster, we combine the models with traditional b/y-ion models. By assigning the models to predict different ion types, we improve overall identification rates. We then annotate PredFull's predicted bin intensities with their likely fragment identities to explore if it predicted anything unaccounted for by UniSpec and whether unknown ions hold any utility in PSM rescoring. This study presents the first comparison of full spectrum models and a novel algorithm for PSM rescoring with the many ions produced from peptide fragmentation pathways.

4.3 Results

4.3.1 Fragment ion annotation of tryptic and nontryptic peptides

Before using auxiliary ions for PSM rescoring, we wanted to confirm their existence in HLA peptide spectra. We employed UniSpec's annotation scheme with its 7919 possible identities. Annotations are grouped as follows: a-ions ("a"), b-ions with or without neutral loss ("b" and "b-NL"), immonium ions ("imm"), internal ions with or without neutral loss ("int" and "int-NL"), precursor peaks with or without neutral loss ("p" and "p-NL"), and y-ions with or

without neutral loss ("y" and "y-NL"). Any remaining unannotated peaks in the experimental spectra were excluded from this analysis. For each PSM accepted at 1% PSM and peptide FDR in a dataset, we calculated two metrics. The first is the relative contribution of each ion type to the total spectral ion intensity (after excluding unannotated peaks). The second, named "normalized count," is the number of peaks annotated as each ion type, normalized by the peptide length (e.g. if 10 y-ions were detected in a spectrum assigned to a peptide 8 amino acids long, the y-ion normalized count score would be 1.25). Peptides were stratified by those ending with arginine or lysine (tryptic) and those that do not (non-tryptic); as expected in HLA data, nontryptic peptides heavily outnumbered tryptic ones (Tables 4.1 and 4.2). We analyzed three representative HLA datasets to understand general distribution patterns of ion type intensities and counts: a lung cancer dataset ("lung cancer")[262], one fraction of a monoallelic cell line ("A*02:01")[187], and HLA class II data from the HLA Ligand Atlas ("Atlas")[228].

Spectral annotation identified all considered fragment ion types in all datasets (Figure 4.1, Appendix Figures C.1-2). Immonium ions contributed the least to total annotated intensity with a median of less than 2% across datasets, while y-ions dominated with 53% being the median y-ion proportion of intensity in the A*02:01 data. As expected, precursor ions were the least frequently annotated with less than 0.15 normalized counts across datasets, while y-ions with and without neutral losses were frequently detected with as many as 1.9 median y-NL normalized counts in the lung cancer data. Immonium peak annotation rates are heavily dependent on how much of the low m/z range is captured in the MS2 scans: the A*02:01 and Atlas datasets did not venture below 100 m/z and had 0.11 and 0.25 normalized immonium ion counts, respectively. The lung cancer dataset's MS2 m/z range delved into the 50 m/z range and had a median of 0.80 normalized immonium ion counts. Interestingly, internal fragment ions

were abundant and high intensity in both tryptic and nontryptic peptide spectra, never falling below third place for both metrics across datasets; this result is in line with prior work showing their prevalence in bottom-up studies[259].

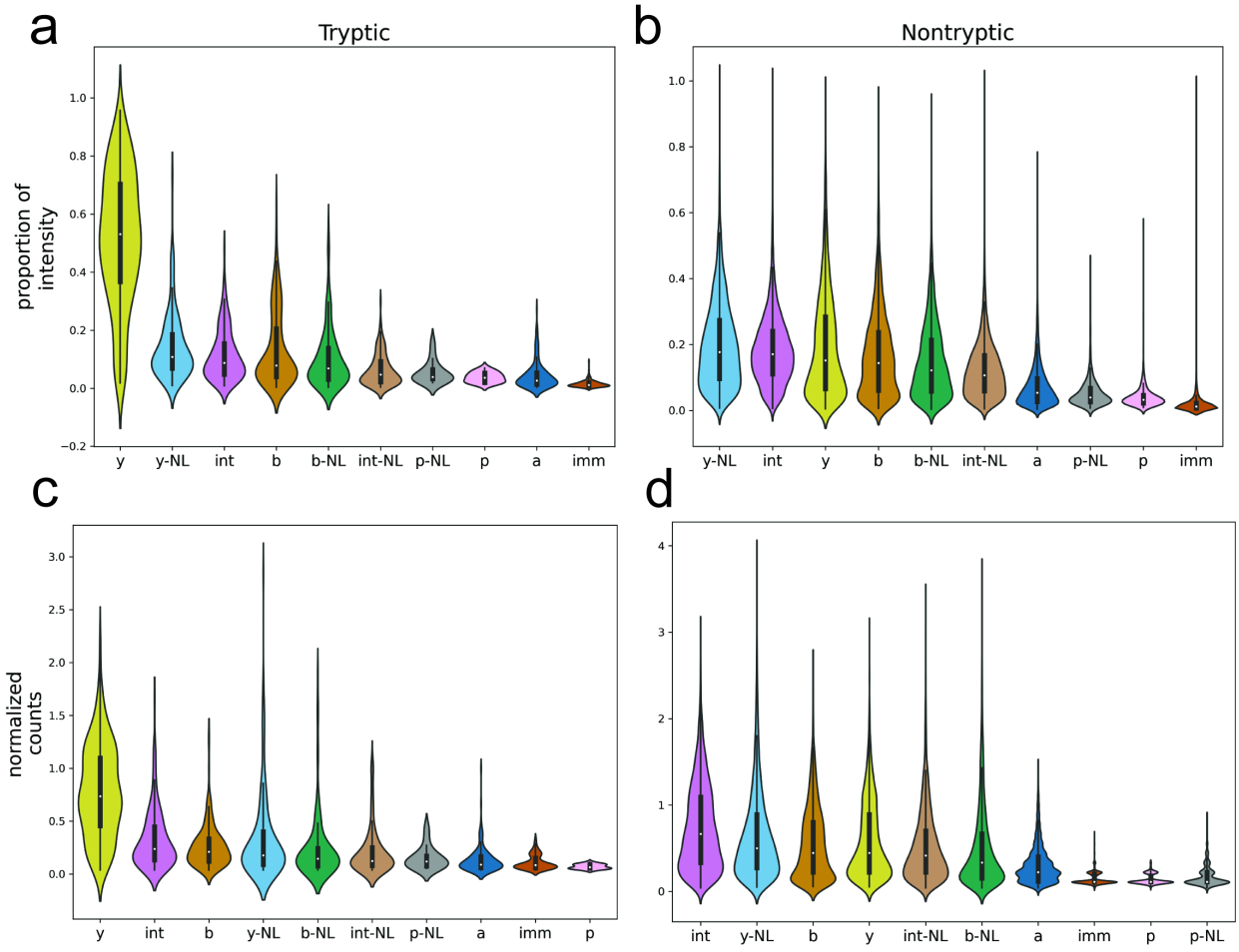


Figure 4.1 Fragment ion annotation metrics in A*02:01 data

Violin plots showing the distribution of proportion of intensities (a, b) and normalized counts (c, d) in tryptic (a, c) and nontryptic (b, d) peptide spectra passing 1% PSM and peptide-level FDR. Ions are sorted from left to right in decreasing order of the metrics. The Silverman method was used to determine the bandwidth of the kernel density estimation. One file corresponding to the first fraction (AG20210507_SR_HLA_A0201_W632_2IPs_Fxn01.mzML[187]) was analyzed.

To compare changes from tryptic to nontryptic spectra, we calculated the difference in median proportion of intensity and normalized counts for each ion type between the two groups (Tables 4.1 and 4.2). Unsurprisingly, y and y-NL ions took the biggest hit when peptides lost their basic C-termini, allowing other ion types to have increased relative intensities.

ion type	Lung cancer I 169 tryptic 14670 nontryptic	A*02:01 227 tryptic 4710 nontryptic	Atlas II 1280 tryptic 5799 nontryptic
a	0.021277	0.0276	0.012023
b	0.036213	0.065644	0.038233
b-NL	0.042953	0.053081	0.028947
imm	0.010453	0.000677	-0.002672
int	0.011968	0.083421	0.028735
int-NL	0.013169	0.0601	0.006461
p	0.005065	-0.003261	-0.002651
p-NL	0.009161	0.000761	-0.003847
y	-0.086645	-0.37945	-0.099414
y-NL	-0.057309	0.068498	-0.031189

Table 4.1 Differences in relative intensities of fragment ion types in tryptic and nontryptic peptides

Differences are calculated by subtracting the median proportion of intensity of each ion type in tryptic peptide spectra from the median proportion of intensity in nontryptic peptides. Each column also designates the number of peptide spectra in each of the tryptic and nontryptic groups.

Interestingly, ion types besides y and y-NL ions consistently saw increased normalized counts in nontryptic peptides. The normalized counts score is unlike proportion of intensity in that it is not a "zero-sum game," meaning that decreased counts of one ion type does not necessarily increase counts of others. Rather, we observed that nontryptic peptides may naturally produce more auxiliary ions. Because of their abundance, we included them in PSM rescoring tests.

ion type	Lung cancer I 169 tryptic 14670 nontryptic	A*02:01 227 tryptic 4710 nontryptic	Atlas II 1280 tryptic 5799 nontryptic
a	0.125	0.131313	0.080882
b	0.242424	0.233918	0.2
b-NL	0.214286	0.186905	0.110068
imm	0.383333	0.025966	-0.047619
int	0.063763	0.428571	0.106667
int-NL	0.042735	0.289216	-0.011765
p	0.038961	0.046528	-0.015686
p-NL	0.0625	-0.006944	-0.033654
y	-0.484848	-0.290643	-0.55
y-NL	-0.607143	0.322134	-0.291679

Table 4.2 Differences in peak counts of fragment ion types in tryptic and nontryptic peptides

Differences are calculated by subtracting the median normalized counts of each ion type in tryptic peptide spectra from the median normalized counts in nontryptic peptides.

4.3.2 Full spectrum prediction models accurately predict nontryptic fragment ion intensities

It is invaluable to check prediction models for their accuracy. UniSpec and PredFull were shown to accurately predict tryptic spectra in their respective publications[142], but they have not been specifically tested on nontryptic data. UniSpec is available on the public Koina server (https://koina.wilhelmlab.org:443/v2/models/), and we added support for PredFull on our own private lab Koina server (Methods Section 4.5.1). We evaluated six HLA datasets[228,262], including one DIA[180] and one phospho-HLA dataset[263]. We also included in our comparison one tryptic dataset as a control and an AspN-digested dataset from the same study[21], hypothesizing that auxiliary ions may be useful for nontryptic enzymatic digestions (AspN cuts N-terminal of aspartic acid). We calculated the accuracy of their spectral predictions, considering b/y and auxiliary ions separately. For the phospho-HLA dataset, we only considered the prediction

performance on phosphopeptides. To filter out potential falsely assigned PSMs, we restricted our analysis to those PSMs that passed a 1% PSM and peptide-level FDR (and 1% protein-level FDR in the tryptic and AspN datasets) and were within a 1 minute deviation from their predicted, calibrated RT. The Prosit[179] and AlphaPeptDeep[135] models were trained on HLA data and known to perform well on HLA PSM rescoring (Table 3.3), making them good benchmarks against which to compare b/y-ion accuracy. Because UniSpec and PredFull were not trained on HLA data, we also included DIA-NN[175] for a fairer comparison.

We first focused on b/y-ion similarity. We observed that for typical HLA data and the trypsin/AspN data, UniSpec performed on par with the Prosit and AlphaPeptDeep predictions, always staying within 0.02 similarity points of the HLA-trained models and often outperforming DIA-NN (Figure 4.2a). PredFull performed the worst of the models, though it stayed within 0.09 similarity points of Prosit/AlphaPeptDeep in most comparisons with a noticeable deficit in the HLA-II peptides. The performance gap was clear for the Atlas Class I data (due to none of the models besides Prosit CID being trained on CID data) and the phosphorylation data (all models natively support the PTM besides PredFull, which employed m/z shifting). We felt confident that full spectrum models could perform well on new data, given the data is similar to their training data.

Next, we assessed the accuracy of UniSpec and PredFull's auxiliary ion predictions. The distributions of spectral similarity scores were quite similar between b/y and auxiliary ions, wherein most cases b/y-ions were predicted slightly more accurately (Figure 4.2b-c); this suggests that these full spectrum prediction models did not develop a strong bias towards b/y-ions and potentially that auxiliary ion intensities, like b/y-ion intensities, are also consistent across repeated spectral acquisitions of the same peptide, though this claim warrants more study.

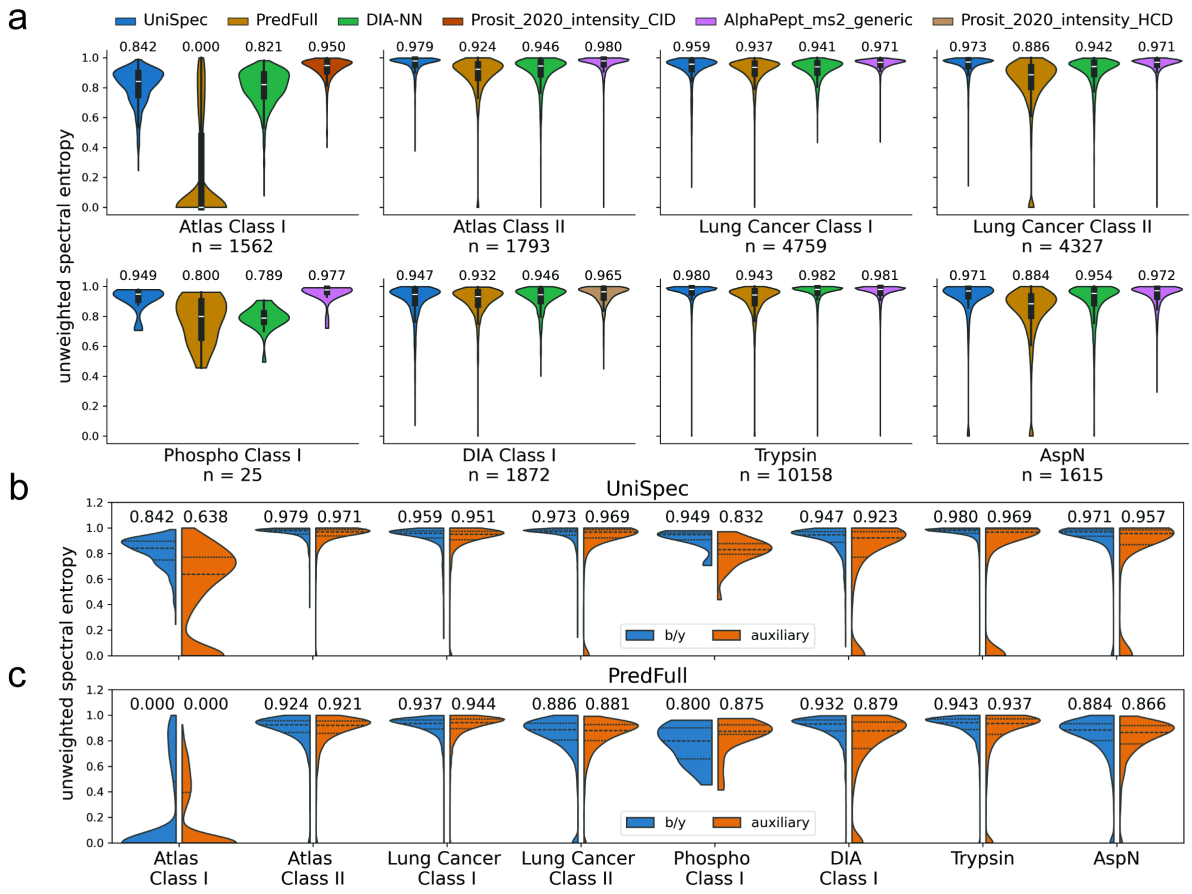


Figure 4.2 Fragment ion similarity distributions

(a) After filtering out PSMs with a delta RT score greater than 1 minute from a 1% FDR-filtered list of PSMs, b/y-ion similarities are visualized in violin plots. Predicted similarities from various models are shown and color-coded consistently across subfigures. The median similarity for each model is listed above their respective plots. Distributions are estimated via kernel density estimation using "Scott" bandwidth estimation. Box-and-whisker plots are shown inside each violin plot, with the box showing the quartiles and whiskers stretching to the furthest points within 1.5 times the interquartile range. The HLA datasets analyzed are Atlas Classes I and II[228], Lung Cancer Classes I and II[262], Phospho Class I[263], and DIA Class I[180]. The Trypsin and AspN datasets are from Wang et al., 2019[21]. A single file from each dataset was analyzed, except for the Phospho-HLA data where all files were used. Sample sizes are provided below each dataset label. (b-c) Split violin plots showing b/y (blue) and auxiliary (orange) ion similarity scores for UniSpec (b) and PredFull (c). Median scores are designated by solid lines, quartiles by dashed lines.

One of the primary purposes of including auxiliary similarity scores is to further discriminate true and false positive PSMs. While it is known that some false positive (FP) PSMs achieve high b/y similarity scores, we hypothesized that it is less likely that they will also achieve high auxiliary similarity scores. To test this, we used decoy PSMs for our known FP group and compared them to likely true positive (TP) target PSMs that passed a 1% FDR and

delta RT cutoff (the same PSMs used in Figure 4.2). We selected those decoy PSMs with either the top 10% highest b/y or auxiliary scores out of all decoy PSMs. We plotted the relationship between the FP and TP PSMs' b/y and auxiliary scores and confirmed that while TP PSMs were likely to score well in both metrics, FP PSMs struggled to simultaneously excel in both (Figure 4.3, Appendix Figures C.3-7). It is important to note that TP PSMs were not selected based on their experimental-predicted spectral similarity, as this would bias our conclusions, though the MSFragger hyperscore is used in Percolator rescoring and is a function of the number and intensities of matched b- and y-ions. Indeed, only 161 of 831 decoy PSMs were members of the top 10% of decoys in both metrics (Figure 4.3a), meaning that fewer than 20% of high b/y scoring decoy PSMs also scored well in auxiliary similarity, and vice versa. Of course, adding an auxiliary similarity score allows FP PSMs that score poorly in b/y similarity to potentially score well instead in auxiliary similarity. The effect of adding auxiliary similarity on PSM rescoring and peptide IDs is directly assessed in the next section.

TP PSMs are sometimes prone to having auxiliary similarity scores of 0 (Figure 4.3c, Appendix Figures C.3-7). This can happen if an experimental spectrum contains none of the predicted auxiliary fragments or if auxiliary fragments are all predicted to be too low intensity. By default, only up to the 30 highest predicted intensity peaks (including b/y and auxiliary ions) are retained; fewer are used if fewer than 30 are within 1% of the predicted base peak intensity (the intensity of the highest intensity peak). If b/y-ions dominate the top 30 spots, then no predicted auxiliary ions make the cut, also resulting in a similarity score of 0. PredFull is less prone to auxiliary scores of 0 (Figure 4.3f), in part due to its unrestricted prediction of bin intensities rather than annotated peaks, allowing PredFull to predict more peaks.

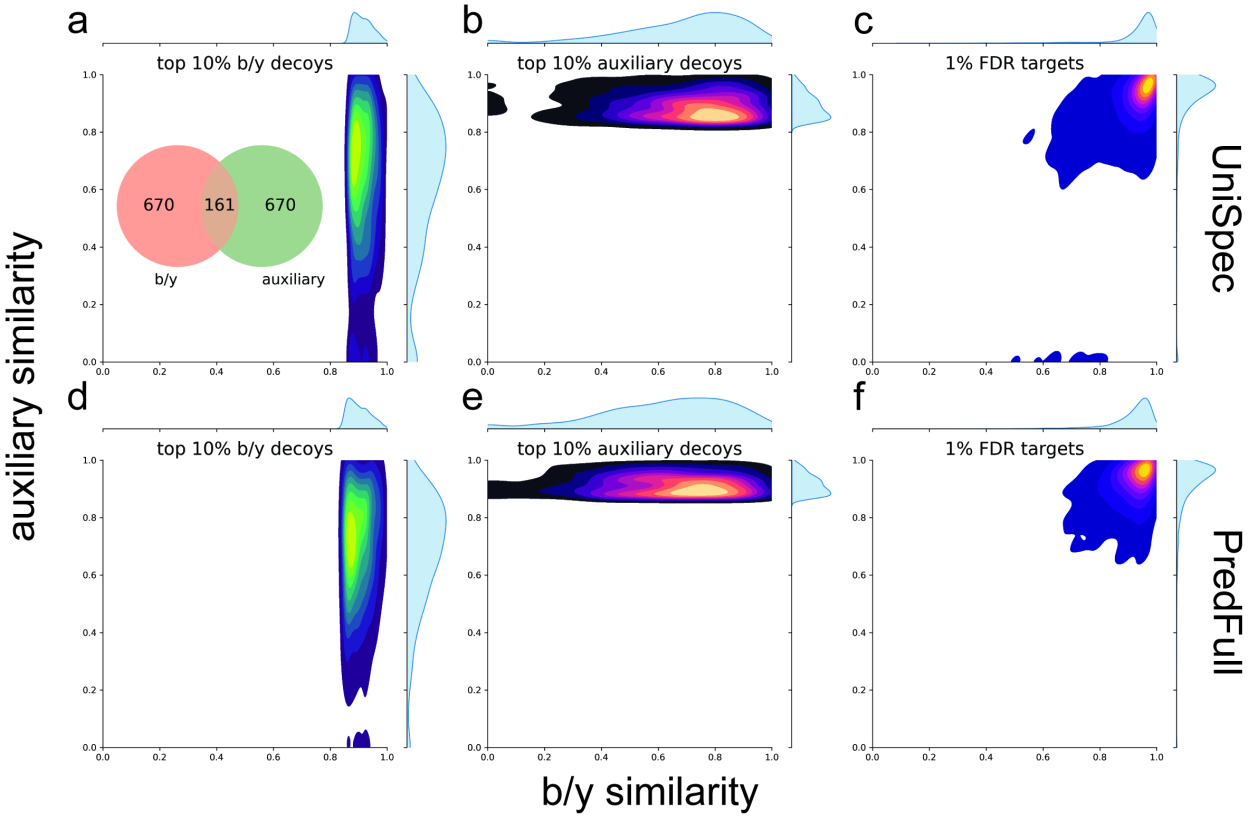


Figure 4.3 The relationship between b/y and auxiliary ion similarity scores for class I data from Kraemer et al., 2023

Kernel density estimations (KDEs) are shown for b/y (x-axis) and auxiliary similarities (y-axis). Bivariate densities are visualized for predicted similarities from UniSpec (a-c) and PredFull (d-f), with brighter areas showing higher areas of density. The top tenth percentile of decoys for b/y (a, d) and auxiliary ion (b, e) similarities are plotted, as well as target PSMs under the 1% FDR and 1 minute delta RT thresholds (c, f). The KDE of each similarity metric is shown along the axes. A Venn diagram showing the overlap of the top tenth percentiles of each score for decoys is shown in (a).

4.3.3 HLA PSM rescoring

Adding new features for auxiliary ion similarity is not trivial. UniSpec and PredFull oftentimes produced lower spectral similarity scores than other "traditional" models like Prosit, AlphaPeptDeep, and DIA-NN did. We hypothesized that delegating the prediction of b/y-ions to traditional models while limiting full spectrum models to predicting auxiliary ions could increase peptide IDs. To test this, we had MSBooster calculate metrics in a few different ways. First, we queried UniSpec or PredFull alone for predictions. We divided b/y and auxiliary ions, calculating the same scores for the two ion sets; this mode was called "divide". "Traditional" mode used a

traditional model for b/y prediction; the chosen model was determined by which model between DIA-NN and the one from MSBooster's heuristic best model search achieved more peptide IDs (Figure 4.4). In traditional mode, these predicted b/y-ions and the predicted auxiliary ions from a full spectrum model were combined into a single spectrum before spectral similarity calculations (Methods Section 4.5.3). Finally, the traditional and divide modes were combined, where one set of features was calculated using only b/y-ions and the other set of features with only auxiliary ions. RT features were included in all rescoring tasks.

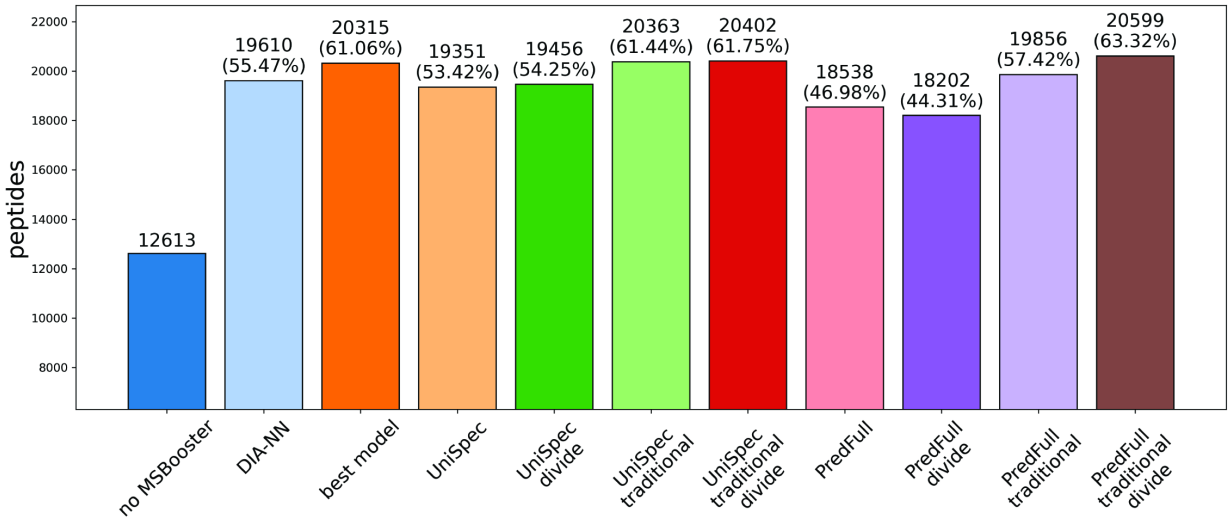


Figure 4.4 Peptide IDs from full spectrum models on an HLA DIA dataset

Peptide IDs for each PSM rescoring mode are shown, with the exact number of peptides and percent gain over "no MSBooster" shown above each bar. "Best model" is the model selected by MSBooster's heuristic best model search. "Divide" indicates b/y and auxiliary ion similarities were calculated separately. "Traditional" indicates a full spectrum model and traditional model were used together to synthesize a single similarity score. "Traditional divide" combines the two ideas, using a traditional model to predict b/y-ions and a full spectrum model to predict auxiliary ions. The model chosen for traditional mode was the best-performing model between DIA-NN and "best model." In this case, Prosit HCD was selected.

We rescored the 6 HLA datasets and counted the peptide IDs. For the tryptic and AspN datasets, both peptide and protein-level IDs were considered. First, we noted that the heuristically selected best model outperformed DIA-NN in all datasets besides the tryptic one (Figure 4.4, Table 4.3, Appendix Figure C.8), highlighting how when all models are robustly trained on tryptic peptides, it is more difficult for MSBooster's heuristic search to predict the

best model. Next, we found that separately calculating b/y similarity scores with traditional models and auxiliary similarity scores with full spectrum models provided the greatest boosts in peptide IDs compared to only using full spectrum models or combining traditional and full spectrum predictions into a single spectral similarity score.

Dataset	No MSBooster	DIA-NN	Best model	UniSpec	PredFull	Gain from auxiliary ions
HLA-I phospho	30430	31.77%	42.09%	43.05%	42.85%	0.96%
HLA-I DIA	12613	55.47%	61.06%	61.75%	63.32%	2.26%
HLA-I Ligand Atlas	14423	13.47%	20.50%	20.48%	20.38%	-0.02%
HLA-II Ligand Atlas	11014	6.91%	7.92%	8.10%	8.01%	0.18%
HLA-I lung cancer	99705	26.18%	35.20%	36.50%	36.05%	1.30%
HLA-II lung cancer	59837	12.12%	14.32%	14.85%	14.43%	0.53%
Tryptic peptides	123616	2.53%	2.33%	2.48%	2.47%	-0.05%
AspN peptides	49740	6.08%	7.00%	6.98%	7.00%	0%
Tryptic proteins	9962	2.84%	2.54%	2.70%	2.67%	-0.14%
AspN proteins	8064	3.63%	4.28%	4.04%	4.29%	0.01%

Table 4.3 Peptide IDs from full spectrum models

Results from Figure 4.4 and Appendix Figure C.8 are summarized here. “UniSpec” is shorthand for “UniSpec traditional divide,” “PredFull” for “PredFull traditional divide.” The “gain from auxiliary ions” column is the percent difference between the maximum of UniSpec/PredFull and the maximum of DIA-NN/Best model.

Finally, comparing UniSpec/PredFull’s “traditional divide” mode with just using the single best traditional model saw the biggest gains in peptide IDs with HLA-I data, with the greatest gain being in the DIA dataset where PredFull improved 2.26% over Prosit HCD alone. Interestingly, PredFull even outperformed UniSpec (20599 vs 20402 peptides), despite having the overall lowest prediction accuracy out of all models tested (Figure 4.2). The HLA-I data from the Atlas dataset did not see improvement with full spectrum models, as they were not trained on CID-fragmented data. HLA-II datasets saw more mild increases from 0.18-0.53%. There was no benefit and sometimes even a negative impact on peptide and protein IDs from using full spectrum prediction with tryptic and AspN data. This is potentially due to the rate of spurious matches to predicted auxiliary ions in FP PSMs outweighing the benefit of adding predicted auxiliary similarity scores for TP PSMs. We concluded that the most promising use case of full

spectrum prediction in our tested datasets was on HCD-fragmented HLA data. This is consistent with Tables 4.1 and 4.2, which showed that nontryptic peptides tend to have more and higher intensity auxiliary peaks.

4.3.4 PredFull predicts unknown fragment ions

PredFull predicts intensities for 0.1 Da bins rather than for defined fragment ions. A result of this is that after annotating each predicted bin with its most likely fragment annotation, there remain many unannotated "unknown" peaks. In a predicted library of 506,896 peptide candidates from the HLA DIA dataset, 1,017,725 of 9,976,065 peaks (10%) with relative intensity at least 1% of the base peak intensity were of unknown origins. When using PredFull for rescoring in the "divide" and "traditional divide" modes, we also included a third category of similarity scores based solely on predicted unknown fragment ions. We required at least two predicted unknown fragment ions in the top 30 highest predicted intensity peaks to calculate a similarity score for a PSM; having zero or one unknown peaks resulted in a score of zero. Setting a minimum of two fragments, we found that 4.9% of PSMs had non-zero unknown similarity scores (Figure 4.5a). The actual distribution of the unknown score showed many PSMs with high predicted similarities (Figure 4.5b). However, many decoy PSMs also attained high similarities, with very little distinction between decoy and target score distributions (target PSMs here can include both true and false IDs). By removing the unknown scores from PredFull rescoring, we identified 20641 peptides, 42 more than when they were included (Methods Section 4.5.3). Despite this, we wanted to see if we could revise MSBooster's scoring functions to improve TP/FP discrimination. To simulate a more stringent score, we modified MSBooster to output zero scores for PSMs with fewer than 3 predicted unknown fragment ions. As expected, the

proportion of zero scores increased (Figure 4.5c). The distributions of target and decoy scores also became more distinct (Figure 4.5d).

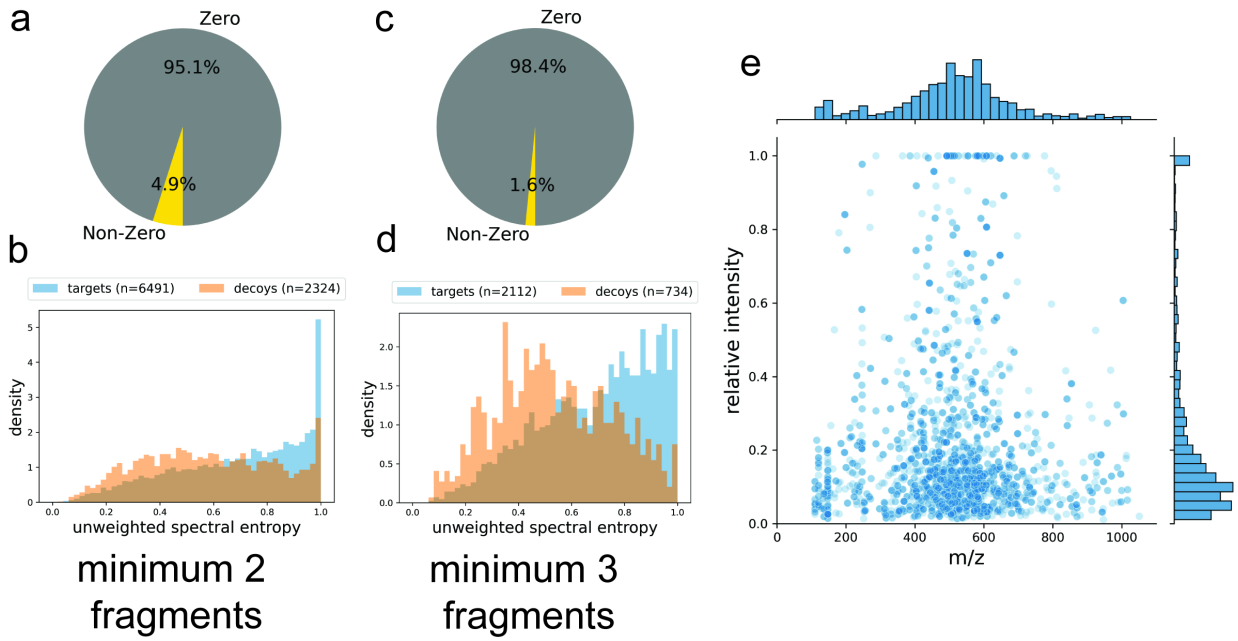


Figure 4.5 Investigating unannotated predicted peaks

(a) The majority of PSMs from the HLA DIA dataset had unweighted spectral entropy similarity scores of zero for unknown ions. Zero scores occur when fewer than 2 unknown fragments were predicted in the top 30 highest intensity peaks, with an intensity threshold of 1% of the base peak intensity. They can also occur when fewer than 2 of the predicted peaks have matches in the experimental spectrum. We only analyzed the Q1 files output by DIA-Umpire[184], corresponding to pseudospectra of precursor ions with more than 2 isotopic peaks detected in MS1 scans. (b) The unweighted spectral entropy similarity scores for unknown ions are plotted for target PSMs (in blue) and decoy PSMs (in orange). Sample sizes for the number of PSMs considered are included in the legend. The y-axis shows the density of scores rather than the absolute number of PSMs. (c-d) Similar figures as (a-b), but for when MSBooster was run requiring at least 3 predicted unknown fragment ions to calculate similarity scores. (e) A scatter plot of unknown peaks' m/z and relative intensity (base peak intensity set to 1) from predicted spectra. This analysis uses those PSMs from (c-d) with non-zero scores, further filtered to PSMs with greater than 0.8 similarity scores in both the b/y and unknown ion categories. Histograms for the m/z values and relative intensities are displayed on the x- and y-axes, respectively.

We wanted to characterize these unknown fragments. To target likely TP PSMs, we focused on target PSMs with a minimum of 0.8 unweighted spectral entropy similarity for both the b/y-ions and unknown ion scores (with a minimum of three predicted unknown ions). Most unknown peaks were predicted to have relative intensities less than 20% of the base peak intensities in their respective spectra (Figure 4.5e). There were also many unknown peaks that

were the highest or nearly highest intensity peaks predicted, residing in the 200-800 m/z range. They were not annotated as precursor peaks (with or without neutral losses) by the UniSpec annotation method. They also inhabited a higher m/z range than where most immonium ions are expected. Further investigation with a more rigorous annotation scheme is needed to elucidate the identities of these high intensity predicted peaks.

4.4 Discussion

UniSpec and PredFull are unique in that they predict the broad variety of fragments produced during HCD fragmentation. After confirming the enrichment of auxiliary ions in nontryptic peptides from HLA data and the accuracy of full spectrum prediction models, we demonstrated that target PSM IDs passing FDR filtering simultaneously achieved high spectral similarity scores for b/y and auxiliary ions, whereas most decoy PSMs that managed to obtain a high b/y-ion similarity underperformed in their auxiliary ion similarity score, and vice versa. Auxiliary ions were most helpful when used to calculate a similarity score separate from the main b/y-ions. Peptide IDs increased the most when full spectrum models were used in tandem with other traditional models to predict auxiliary and b/y-ions, respectively, once again showcasing the power of MSBooster's model flexibility. Interestingly, PredFull outperformed UniSpec on one HLA dataset, showing that despite its hands-off approach of not explicitly defining expected fragment ion types, it still performed well in generating predictions for PSM rescoring. Though not currently available on the public Koina server, PredFull has demonstrated its utility and can be included in future Koina updates, though more investigation is needed to determine under what conditions it can outperform UniSpec.

Our method consistently increased HCD-fragmented HLA peptide IDs, with improvement ranging between 0.18 and 2.26%. Like other models not trained on HLA peptide

data, UniSpec and PredFull were able to extrapolate their predictions by learning from the small portion of nontryptic peptides present in tryptic digestions. These findings serve as a baseline for the first iteration of our algorithm using full spectrum models. There is still plenty of room for improvement. While we use all but p and p-NL peaks in our scoring, we can also filter out isotopic peaks from the predicted spectra. MS2 fragments in bottom-up proteomics are typically in a mass range where theoretical calculations of the fragment isotopic envelope dictate that the monoisotopic fragment peak should be the highest intensity[264]. MSBooster spectral similarity scores do not enforce that the monoisotopic fragment peak is present to also use its higher isotope peaks, allowing FP PSMs to increase their similarity scores by matching a higher isotope peak while ignoring the monoisotopic one. Furthermore, isotopic peak intensities are theoretically calculable[264], potentially making their intensity prediction from UniSpec redundant for our task. To demonstrate the utility of intensity information, we can also compare the PSM rescoring performance of spectral similarity metrics against calculating a MSFragger-like hyperscore for auxiliary ions.

While exploring the "dark fragmentome" of PredFull's unannotated fragment ions, we discovered that some PSMs with abundant unknown ions showed high agreement between these predicted ions and their experimental spectra. High intensity fragments could correlate to fragments with multiple neutral losses[265,266]. While PredFull can learn rare fragmentation patterns without prior knowledge of fragmentation rules, we must not over-interpret MS/MS spectra and try to assign identities to every peak. Artifact peaks and coeluting precursor fragments in spectra from the training set can propagate in unexpected ways, making complete spectral annotation a wild goose chase, especially for lower intensity fragments. The sparsity of non-zero scores for unknown ions also makes them unlikely to play a huge role in PSM

rescoring; their annotation is more of a cerebral exercise than one of practical application. This scarcity supports UniSpec's claim that by mining the NIST HCD library, they can generate an ion collection that covers most peaks in HCD spectra[247]; furthermore, it shows that spectral prediction models guided by prior knowledge of fragmentation mechanisms extract most of the meaningful data from MS/MS spectra. A similar approach of near-exhaustive spectral mining can be used to extend full spectrum prediction to other fragmentation methods.

We only focused on HLA rescoring, but these models may hold promise in other applications. One study synthetically added neutral loss peaks to predicted spectra to aid phosphorylation localization[267]. While their method synthetically shifted m/z peaks from unphosphorylated peptide predictions, UniSpec was specifically trained on phosphopeptides, and its predicted phosphorylation neutral losses may improve upon this study's idea. In addition, PredFull supports electron transfer dissociation (ETD) spectra prediction. ETD is a complimentary fragmentation method to HCD for greater protein sequence coverage with non-trypsin proteases and middle-down proteomics, as well as a complimentary method in peptidomics[38]. We found no studies using PredFull for ETD prediction; an assessment of whether these predictions improve ETD rescoring may provide interesting results.

4.5 Methods

4.5.1 Koina servers

A description of how MSBooster uses the Koina server is provided in Results Section 3.3.1. Briefly, MSBooster formats peptides into JSON files to send to a Koina server (https://koina.proteomicsdb.org/) via HTTP requests. Predictions are sent back to MSBooster in JSON format and saved internally as well as printed to mgf files. Requests are sent in batches of 1000 peptides for all models except for UniSpec, which sends batches of 100 peptides; this

reduced batch size is set by default because the UniSpec implementation on Koina sends back m/z and intensity vectors up to length 200 per peptide corresponding to the highest predicted intensity peaks, much longer than other models' vectors yet still small in comparison to the almost 8000 peaks it can predict. Prediction batches are sent and processed in a multithreaded manner.

PredFull is not yet available on the public Koina server. We set up a local, private server following the guide at https://github.com/wilhelm-lab/koina?tab=readme-ov-file#hosting-your-own-server. Our code edits to support PredFull are available at https://github.com/yangkl96/koina/tree/predfull. Our local server is powered by a NVIDIA GeForce RTX 4090 GPU. For consistency, all model predictions were sent to our local server.

4.5.2 PredFull fragment annotation

PredFull only outputs predicted intensities with a minimum threshold of 0.001, along with the associated m/z values for the bins they belong to. No ion annotations are provided, so MSBooster assigns them upon reading in PredFull predictions. MSBooster reads in a file containing all fragments predicted by UniSpec. Each predicted peak is assigned an ion annotation falling within ±0.05 Da of the bin m/z; if no annotation falls within this range, the fragment is designated as "unknown." Adopting a similar annotation method to Neuhauser et al., 2012[246], we defined a fragment ion hierarchy so that any predicted peak close in m/z to multiple fragment ion types will resolve its ambiguity by being assigned to the highest priority ion type in the hierarchy. The hierarchy in descending priority order is "imm", "y", "b", "a", "y-NL", "b-NL", "a-NL", "int", "int-NL", "unknown".

4.5.3 MSBooster algorithm with auxiliary ions

PSM rescoring with auxiliary ions is available in MSBooster v1.3.4. Spectra, auxiliary spectra, and RT models are separately called for each peptide candidate from MSFragger. The auxiliary spectra model can be substituted in for the main spectral model if the former supports a peptide the latter does not; for example, a 31-amino-acid-long peptide not predicted by Prosit will default to using monoisotopic b/y-ion peaks from UniSpec's spectral prediction. Predicted spectra are normalized so the highest intensity peak is set to 1. Precursor peaks with and without neutral losses are filtered out, if the (auxiliary) spectra model predicts them. For spectral models, up to the top 20 remining peaks are retained if they are all at least 1% of the base peak intensity; for auxiliary spectra models, 30 peaks are retained, following the same intensity threshold filtering parameters. Predicted spectra from the main and auxiliary spectra models are saved in separate mgf files.

The default in MSBooster is to calculate similarity scores using all fragment ion types together; the "divideFragments" parameter is set to 0 by default to specify this behavior (Appendix D). When combining b/y and auxiliary ions predicted by separate models (e.g. DIA-NN for b/y, UniSpec for auxiliary), the top 30 intensity predicted peaks from the union of both models' predictions are used. To calculate b/y and auxiliary similarity scores separately, divideFragments is set to 2. To calculate b/y, auxiliary, and unknown ion similarity scores separately when using PredFull, divideFragments is set to 3. Any ion category with fewer than 2 peaks in the final list of top 30 peaks is assigned a similarity score of 0 for that PSM.

Because PredFull's predicted m/z values only have precision up to 0.1 Da, the highest intensity peak in the experimental spectrum in the range of ±0.05 Da of each predicted peak from PredFull is set as the matched peak intensity for spectral similarity calculation. The remaining

peaks predicted by other models use the ppm or Da tolerance specified in the MSFragger parameter file to search for the highest intensity matched peak from the experiment spectrum. To compare PredFull performance without the unknown fragment similarity scores, divideFragments is set to 2, and "unknown" was added to the "ignoredFragmentIonTypes" parameter.

4.5.4 Spectral similarity features

In addition to the unweighted spectral entropy similarity score, three other similarity scores have been added: weighted spectral entropy similarity, hypergeometric probability, and intersection. These scores are calculated separately for each fragment ion group (i.e. b/y, auxiliary, unknown). Because the hypergeometric probability and intersection scores rely on calculating all hypothetical fragment m/z values for a given fragment ion group, they are not applicable to unknown fragment ions and not calculated or added to the .pin files.

MSBooster's weighted spectral entropy similarity score is derived from findings that it outperforms its unweighted counterpart when spectral entropy is low[215]. This is the case for many short peptides, such as HLA peptides. Our version multiplies the unweighted spectral entropy similarity score by the square root of the spectral entropy of the predicted spectra. Unlike the implementation in Li et al., 2021, we calculate both unweighted and weighted scores for all PSMs.

The hypergeometric probability is derived from the idea that the matched experimental peaks of the highest intensity predicted fragments should be enriched for non-zero intensities. For a given set of ion types, all possible theoretical peak m/z values are calculated, resulting in a set of size *N*. Then, theoretical peaks are matched to the experimental spectrum; the number of non-zero intensity matched peaks is *M*. The number of non-zero intensity experimental peaks

matched to the *P* predicted peaks is designated as *X*. A hypergeometric distribution with population size *N*, number of possible successes *M*, and sample size *P* is created, and the score is the probability associated with *X* successes.

The intersection score is derived from the idea that the experimental and predicted spectra should agree which theoretical fragments have the highest intensities. After calculating all theoretical m/z values corresponding to the fragment ion set, they are matched to the experimental spectrum. Up to the top 20 (MSBooster default) highest intensity matched peaks are kept. The intersection score is the number of peaks shared between the matched and predicted peak m/z values.

4.5.5 FragPipe workflows

FragPipe 22.0 was used for all analyses in conjunction with an updated MSBooster v1.3.4. A reviewed human protein database from March 18, 2022, with contaminant proteins and reversed decoy sequences added by FragPipe, was used for all searches unless otherwise noted. Ligand Atlas data was processed with the "Nonspecific-HLA" workflow with no changes. Lung cancer data was searched with a combined lung cohort protein database provided with the publication. Two oxidation events on methionine and three phosphorylation events on serine/threonine/tyrosine were added as variable modifications. ProteinProphet[95] was enabled. For these two datasets, class I and II files were searched separately, but with the same length restrictions of 7-25 amino acids. HLA DIA data was processed with the "Nonspecific-HLA-DIA" workflow with DIA-Umpire[184], with spectral library generation and quantification disabled. Phospho-HLA was run similarly to the "Nonspecific-HLA" workflow, but with peptide length restricted to 8-15 amino acids, the only variable modifications set to two methionine oxidations and two phosphorylation events on serine/threonine/tyrosine, and ProteinProphet

enabled. Tryptic data was searched with a modified “Default” workflow, not allowing in-silico digestion after proline. AspN data was also searched with a modified “Default” workflow, allowing three missed cleavages and in-silico digesting proteins N-terminal of aspartic acid.

4.6 Data availability

A*02:01 data[187] was downloaded from ProteomeXchange[104] with accession number PXD027165, glioblastoma data from the HLA Ligand Atlas[228] with PXD027165, lung cancer data[262] with PXD034772, phospho-HLA data[263] with PXD013831, HLA DIA data[180] with PXD022950, and trypsin and AspN data[21] with PXD010154.

Chapter 5 The Future of Peptide-Spectrum Match Rescoring

5.1 Conclusions

The results of peptide identification directly influence the final interpretation of bottom-up proteomics data, affecting everything from protein quantification to PTM localization. ML methods have extracted reproducible properties of peptides across their repeated measurements in giant proteomics data repositories; next-generation DL methods have built upon their predecessors and excelled at peptide property prediction. The additional information from comparing experimental measurements and model predictions supplements basic PSM descriptors and database search scores. Knowing this, many popular analysis pipelines have added ML/DL methods to their PSM rescoring steps, digging deeper into the proteome than previously possible.

In Chapter 2, we introduced MSBooster, a PSM feature calculator to be used before Percolator PSM rescoring. By testing PSM rescoring with MSBooster enabled, we found that the greatest gains were observed for the most difficult problems; that is, where DL-based PSM rescoring truly excels is for those problems that pose a challenge for search engines. Not only can these ML-informed scores increase peptide IDs, but they can also improve the reliability of the IDs, demonstrated when MSBooster rejected HLA peptides that did not match expected binding motifs. Furthermore, MSBooster helped identify a novel patient-specific peptide in the melanoma dataset, showcasing MSBooster's utility in therapeutic development. MSBooster's scores are propagated to the final PSM tables, and spectral predictions are displayed in mirror

plots by FragPipe-PDV. In summary, MSBooster scoring is applicable to diverse experimental data types, and its inclusion in FragPipe gives other tools access to its scores.

Using DIA-NN as the default predictor in FragPipe produces accurate predictions for peptides like those in its training dataset. Though MSBooster employs solutions such as m/z peak shifting and separate RT/IM calibration curves for types of peptides previously unencountered by DIA-NN, extrapolation is not possible when PTMs or fragmentation methods significantly change peptide properties. In Chapter 3, we described the integration of Koina into MSBooster, giving users access to the growing list of peptide prediction models developed every year. The public Koina server is fast, not posing a huge bottleneck in analysis time, especially if predicted libraries are reused (Appendix D). Koina makes MSBooster more flexible, allowing users to mix-and-match models. We also created a heuristic model search to automatically pick high-performing models for any given dataset. After benchmarking multiple datasets, we set guidelines for suggested models for each data type. As more models are trained with improved training datasets and DL architectures, these guidelines will eventually become obsolete. In fact, we argue that out of our findings in Chapter 3, the ones with the potential to have the longest lasting impact concern the philosophy of model training. First, the allure of diverse training data weakens if said data is improperly encoded. Second, we find that model accuracy when tested in isolation does not necessarily translate to the final observed peptide gains; we propose modifications to the loss function that may address this discrepancy. Third, we note the lack of standardization in supported peptides by different models (Table 3.1). Though MSBooster hard codes these limitations and filters out peptides impossible to predict, the proteomics community could benefit from standardizing the peptides supported by future models. Consider how carbamidomethylation of cysteine is handled across models: some assume all cysteines are

alkylated, while others require users to specify the PTM in input files, though it is not clear whether unmodified cysteine is present in the training data. Even if standardization is not possible, solutions are needed for users to quickly check each model's restrictions, saving them the time and effort of digging for such details in the literature. Koina has recently expanded model documentation on its website to do just that.

Most spectral prediction models ignore the wealth of informative peaks generated by collision-induced dissociation of peptides. We put these auxiliary fragments to work in Chapter 4, showing marginal improvements in HLA peptide rescoring. Once the algorithm is refined, the newly gained peptides can be assessed to assure they are consistent with expected binding motifs, and the rescoring can be challenged with entrapment experiments to ensure proper FDR control. Auxiliary ion rescoring can be added in the future to FragPipe, which includes many workflows specific to immunopeptidomics processing. The feature is available via the command line for beta testing (Appendix D). Furthermore, we performed the first assessment of full spectrum prediction tools against each other. UniSpec and PredFull approach the problem from very different angles, but both successfully increase HLA peptide IDs when used with traditional b/y-ion models. PredFull's unknown predicted peaks show high similarity to their matched experiment peaks and likely correspond to real fragment ions, but their sparseness suggests UniSpec's annotation scheme of almost 8000 fragment ions captures most HCD fragmentation products.

Overall, we believe that MSBooster provides a holistic solution to PSM rescoring for diverse datatypes by not only supporting the multitude of state-of-the-art peptide property ML models, but also cleverly selecting models *ad hoc* and delegating them to the different tasks of b/y-ion, auxiliary ion, RT, and IM prediction. DL-informed similarity scores not only improve

the sensitivity of peptide ID but also can highlight those peptides that display high concordance between their experimental and predicted values. In particular, spectral (and sometimes RT) similarity often distinguishes true and false IDs better than traditional database search scores. When deciding what biomarkers to select for further evaluation in a clinical setting, we believe that prioritization of high-scoring peptides or proteins that contain those peptides may speed up biological discovery and therapeutic development.

Wide-ranging studies have used MSBooster, showing its effectiveness across various applications. First, it has been used in protein quantitation and PTM characterization studies[268,269]. Others have corroborated MSBooster's indispensability in FragPipe for cancer neoantigen discovery[270], but our tool is broadly applicable to any disease or small molecule drug binding investigations[271]. Apart from cancer[272], MSBooster has been employed in studies of fish malnutrition[273] and hair regrowth[274]. It also helped scientists searching for protein-level evidence of novel single amino acid variants[275] and microproteins originating from small open reading frames[276]. Other applications include interactomics[277], proteolysis[278], developmental biology[279], microbes[280], and food science[281]. We look forward to seeing what other creative works published in the future will make use of FragPipe and MSBooster.

5.2 Future Directions

As an open-source tool, MSBooster will continue to evolve to meet users' needs. Indeed, Results Section 2.5 shows how we have had to change the MSBooster algorithm alongside quickly developing proteomics technologies. We present the following ideas that have repeatedly popped up during its development but have yet to be addressed fully and have the potential for immediate impact.

5.2.1 Beyond Percolator: dealing with missing values

The Percolator[87] algorithm is a major component of many PSM rescoring workflows, given its flexibility in accepting an arbitrary number of features for rescoring. We exclusively use Percolator with MSBooster in FragPipe, but it comes with known limitations. Importantly, it does not accept missing values in its input. Missing values can occur when using a prediction model with an upper bound on peptide length, or if it lacks support for noncanonical amino acids. In the absence of a prediction, MSBooster can assign an arbitrary value to replace the missing value, such as setting spectral similarity to 0. Figure 3.4 highlights how Prosit TMT, which typically performs well, artificially has decreased performance when the sample includes longer peptides. Though these model input limitations did not pose serious issues for our analyses, they limit the adoption of DL-based PSM rescoring for partial enzymatic digestion/limited proteolysis experiments and studies of proteins containing noncanonical residues, such as selenoproteins. Models can be retrained with increased support, but because there exist few guidelines for the kinds of peptides supported by models, some method of handling missing values could be beneficial. Decision-tree methods, many of which are supported by Scikit-learn and Mokapot[203], tolerate missingness and are one possible workaround.

If we wish to continue with Percolator, we could consider missing value imputation, a statistical method for data imputation. The type of data missingness dictates how we can impute it[282,283]. Missing features from longer peptides could be classified as "missing at random" (MAR), which occurs when the probability of missingness of a variable relies on other variables (i.e. MSBooster cannot calculate features for longer peptides, and peptide length is a feature passed to Percolator). Missingness from unsupported amino acids could also be considered as MAR, though because we do not expect incorporation of these residues to change anything about

the distribution of the Percolator features and peptide sequences are not use in the SVM algorithm, there is a case for classifying this as "missing completely at random." These types of missingness are amenable to imputation techniques, unlike "missing not at random," which would occur if the missingness of MSBooster features depended on the feature values themselves (i.e. values are missing because they come from a different distribution than present values).

Another interesting solution would take MSBooster features as "privileged information" to be used during training to inform the decision boundary[284], thereby learning from DL-based similarity features without requiring them for unsupported peptides. Learning using partially available privileged information has been shown to outperform the mean and iterative imputation methods on some datasets.

5.2.2 Transfer learning

MSBooster supports a multitude of models, but FragPipe still lacks a solution allowing users to train their own. ML model development has long been relegated to expert ML practitioners, as many pillars of ML training must be adhered to; failure to do so may produce faulty models, whether that be from overfitting, training on low quality data, or other reasons. Proteomics labs specializing in DL endeavor to train and test models for reliable predictions. Such a tailored approach takes time, necessitating prioritization of the most popular data types. Many labs focus on less popular but critical types of proteomics data, such as rare PTMs, while others develop chemical probes or test new instrumentation and fragmentation methods. To truly democratize DL, solutions are needed to produce high quality models for the users' data at hand.

There is a particular focus in the proteomics DL community on PTM prediction. Some models have encoding schemes that allow them to extrapolate their predictions to arbitrary

PTMs. PTMs can simply be encoded by setting a 0 or 1 at a specific position along the peptide sequence input vector, known as "one-hot embedding." They can also be encoded by their atomic composition, the number of each atom comprising the modification. The latter approach has been adopted by tools such as AlphaPeptDeep[135], DeepLC[136], and MoSTERT-2S[285], allowing them to predict peptides with arbitrary PTMs. This approach is based on "zero-shot learning," the idea that models can generalize to classes previously unseen[286]. Zero-shot learning has displayed varying levels of success, performing well on PTMs chemically similar to those in the training data yet struggling on PTMs with very different structures or atomic compositions[136,285].

Because no single model can extrapolate accurately to all PTMs, users may find that zero-shot models do not work for them. If they have sufficient data or access to their PTM in data repositories, they can train a model themselves using "transfer learning" (TL), which is implemented in AlphaPeptDeep[135] and has been used to generate fully predicted spectral libraries[287]. TL is a technique that uses a base model as a foundation while providing examples from a new class of data for training[288]. It allows a new PTM-specific model to be trained on top of a base model that has already been trained on a large dataset. The new transfer learned model has the benefit of already having learned general peptide fragmentation patterns, freeing up the training regimen to focus on the specifics of the new PTM. TL has a reduced training data size requirement. DL models are notoriously needy, oftentimes requiring millions of training points; TL needs orders of magnitude fewer datapoints (hundreds or thousands, depending on how different the target space is). The term "fine-tuning" is similar to TL but is typically used to mean the slightest tuning of model weights with limited data to fit the nuances of a related task, such as calibrating for differences on the same MS instrument and LC setup for experiments across time; pDeep3 shows this can be accomplished quickly, on the order of tens of seconds[172].

We propose adding a TL module in FragPipe built atop the open-source AlphaPeptDeep TL code (https://github.com/MannLabs/alphapeptdeep, with our own fork for testing at https://github.com/Nesvilab/alphapeptdeep/tree/main). We envision users training their own models using data collected from public repositories or their own data at hand. PSMs that pass FDR filtering can be combined with MS/MS data as input for a TL script, either run locally or on a publicly hosted FragPipe TL server. The newly learned weights can be substituted for the base model's weights for the final prediction step. Koina could be extended to support these new model weights, as model weights are currently pulled from a Zenodo data repository.

While we have not yet added an end-to-end solution for TL, we have performed pilot studies. In one such study, we performed TL on TMT data from a pancancer study hosted by the Clinical Proteomic Tumor Analysis Consortium (CPTAC, https://proteomic.datacommons.cancer.gov/pdc/). The training dataset was comprised of data from non-clear and clear cell renal cell carcinoma and colon adenocarcinoma tissue samples, covering 11,503,870 N-terminally labeled, 7,398,090 lysine-labeled, and 1,352,801 serine-labeled PSMs. We used command line MSBooster to generate AlphaPeptDeep input files containing all peptide candidates from an MSFragger search of TMT-labeled yeast data[289]. We used our newly transfer learned AlphaPeptDeep model (AlphaPeptDeep TMT) to predict these peptides. The predictions were formatted as an mgf file for MSBooster to read. PSM rescoring resulted in a final number of proteins on par with those obtained from other TMT models, DIA-NN and Prosit TMT[232] (Figure 5.1a). Furthermore, AlphaPeptDeep TMT was the only model explicitly trained on TMT overlabeling of serine and extrapolated well to the new PTM localization (Figure 5.1b).

We also trained a TMT model on the yeast data itself, a paradigm which we call "self-learned." By dividing confidently identified TMT peptidoforms into two non-overlapping groups

(with sizes of 28,875 and 29,058 PSMs), we created two unique PSM sets on which to train two TMT models. Each model predicted properties for PSMs assigned to peptidoforms not in their training set, avoiding training-testing data leakage. Training on these smaller datasets produced the most successful PSM rescoring by identifying the most proteins out of all models (Figure 5.2a).

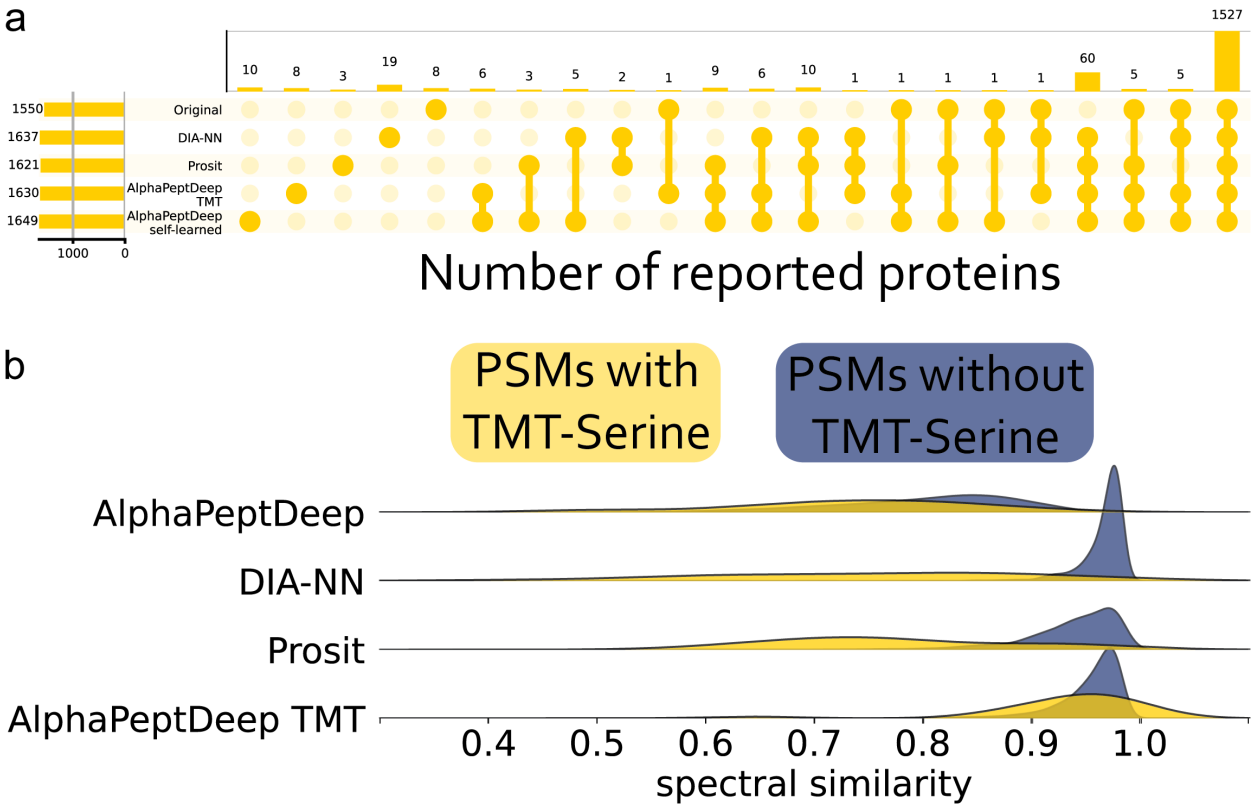


Figure 5.1 Transfer learning a TMT model

(a) An upset plot showing proteins identified when using various models for PSM rescoring. "Original" skips the MSBooster step. (b) Distributions of the unweighted spectral entropy similarity score for TMT-labeled peptide PSMs. Peptide PSMs with TMT on serine are shown in yellow, while all others are shown in blue. AlphaPeptDeep is the generic model not trained on any TMT data.

Fully implementing a TL pipeline in FragPipe requires us to heed cybersecurity and data privacy needs[290]. An interesting research question will be deciding just how many PSMs are needed for proper TL and how best to visualize quality control metrics to reassure users not only that their newly trained models are accurate, but also that they do not overfit on their data if they

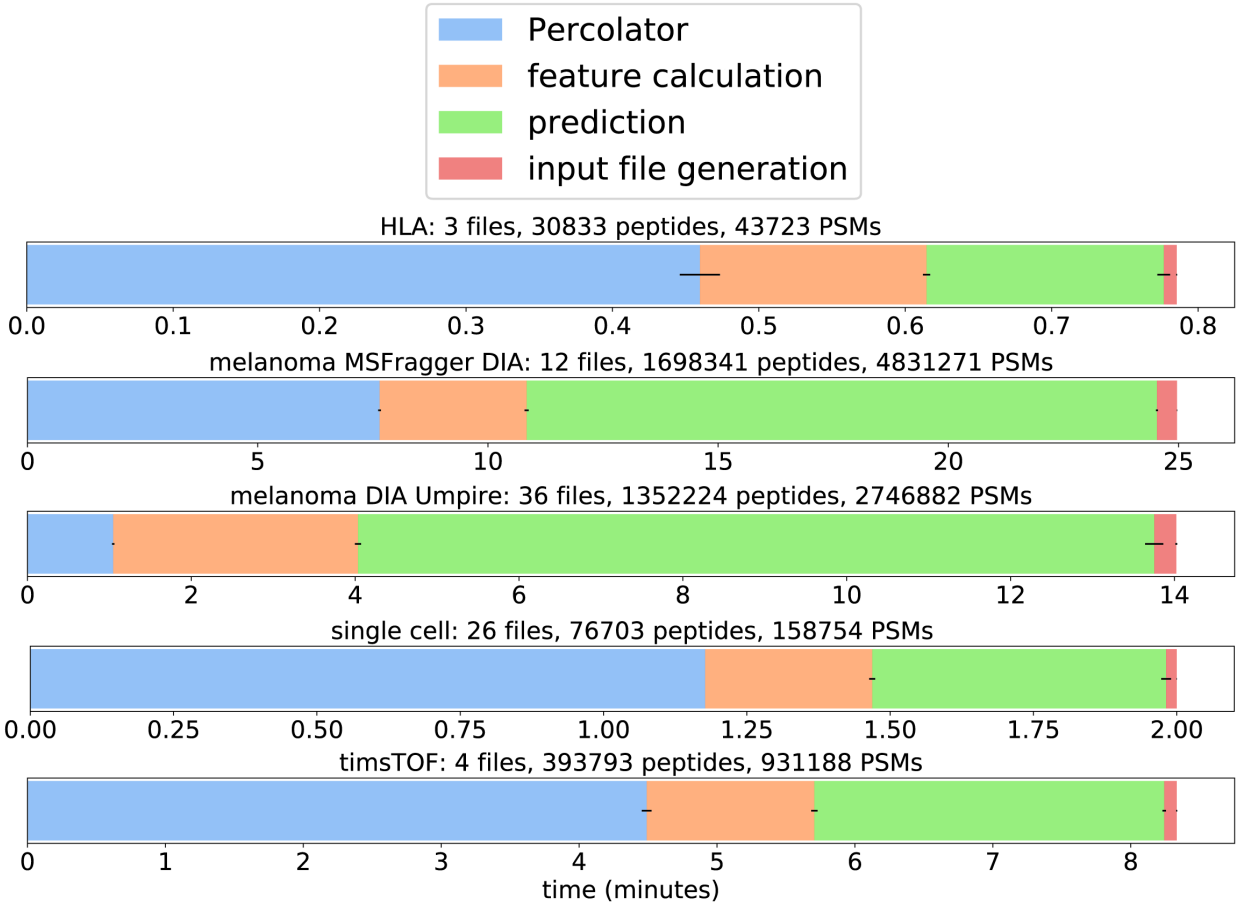
used their own data for training. We envision a future FragPipe that lets users regardless of ML/DL expertise to explore peptide prediction model training, balancing the practical considerations of ease of training, demands of computational resources and time, and confidence in accurate predictions.

5.2.3 Incorporating retention time predictions before PSM rescoring

MSBooster's retention time difference feature has strong discriminative power and is often assigned large weights during Percolator rescoring. This is because MSFragger proposes many peptide candidates with large deviations from the RT calibration curve which can be penalized during PSM rescoring. A natural extension is to instead use predicted RT as a filter, only allowing peptides with a predicted calibrated RT within a certain window of the current experimental RT to be scored against a spectrum. DIA and targeted proteomics software already use similar ideas successfully, referring to a spectral library to determine where along the LC gradient to expect peptides[175,291,292]. Adding RT as a prefilter would decrease the search space for MSFragger at every time step, potentially greatly reducing computational search time. One benefit of our current approach is that we avoid making predictions for entire spectral libraries, instead predicting peptide properties on an *ad hoc* basis for just MSFragger's proposed peptide candidates. To incorporate RT into the database search, we must consider how to predict them in a timely manner. Predictions could be saved for reuse, adding new peptides to an RT database as they are considered in searches for the first time. Differences between reversed-phase and other LC setups also present a challenge for this idea, but new models can easily be trained.

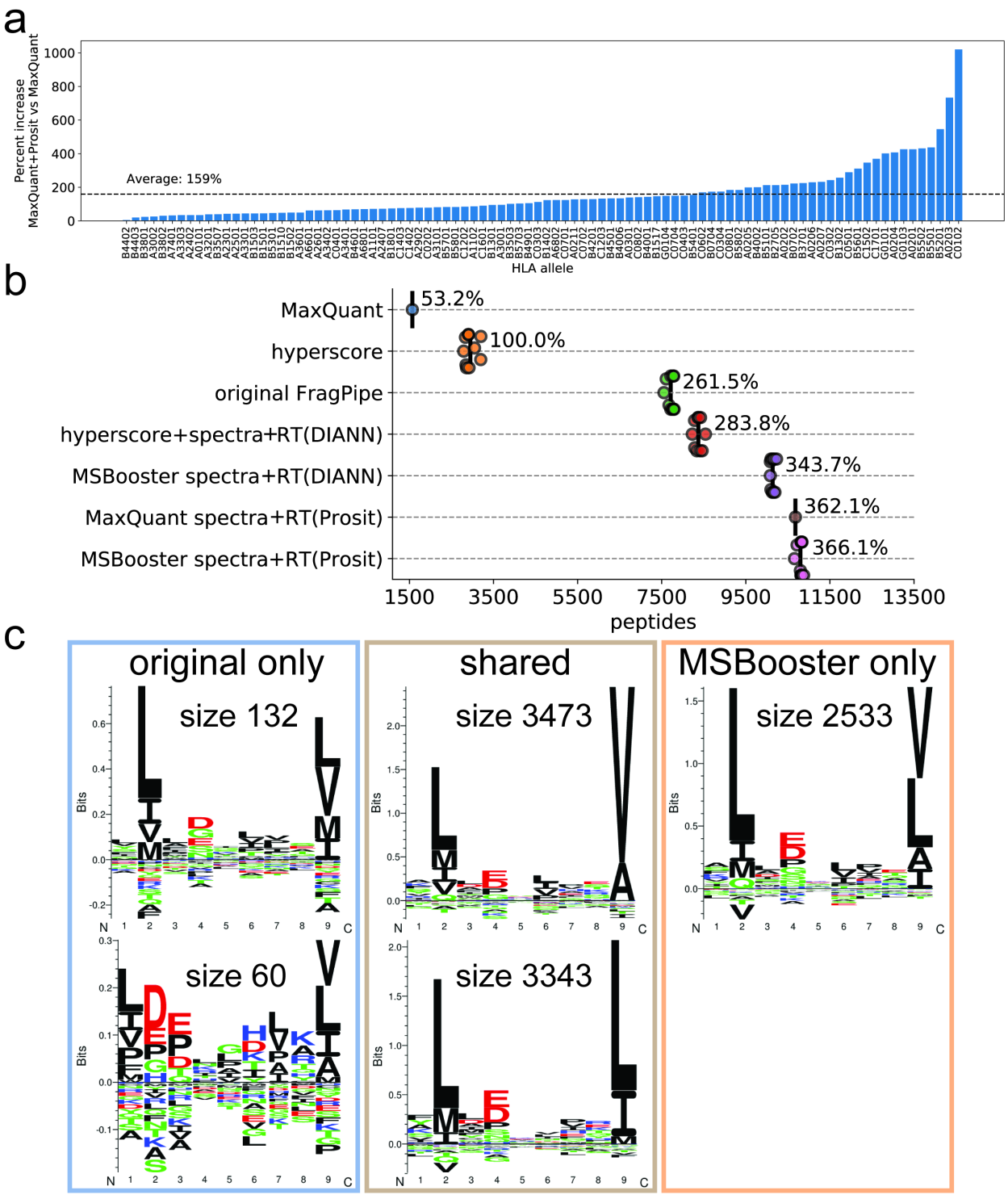
Appendices

Appendix A: Supplemental Material for Chapter 2



Appendix Figure A.1 Timing of MSBooster steps and Percolator rescoring

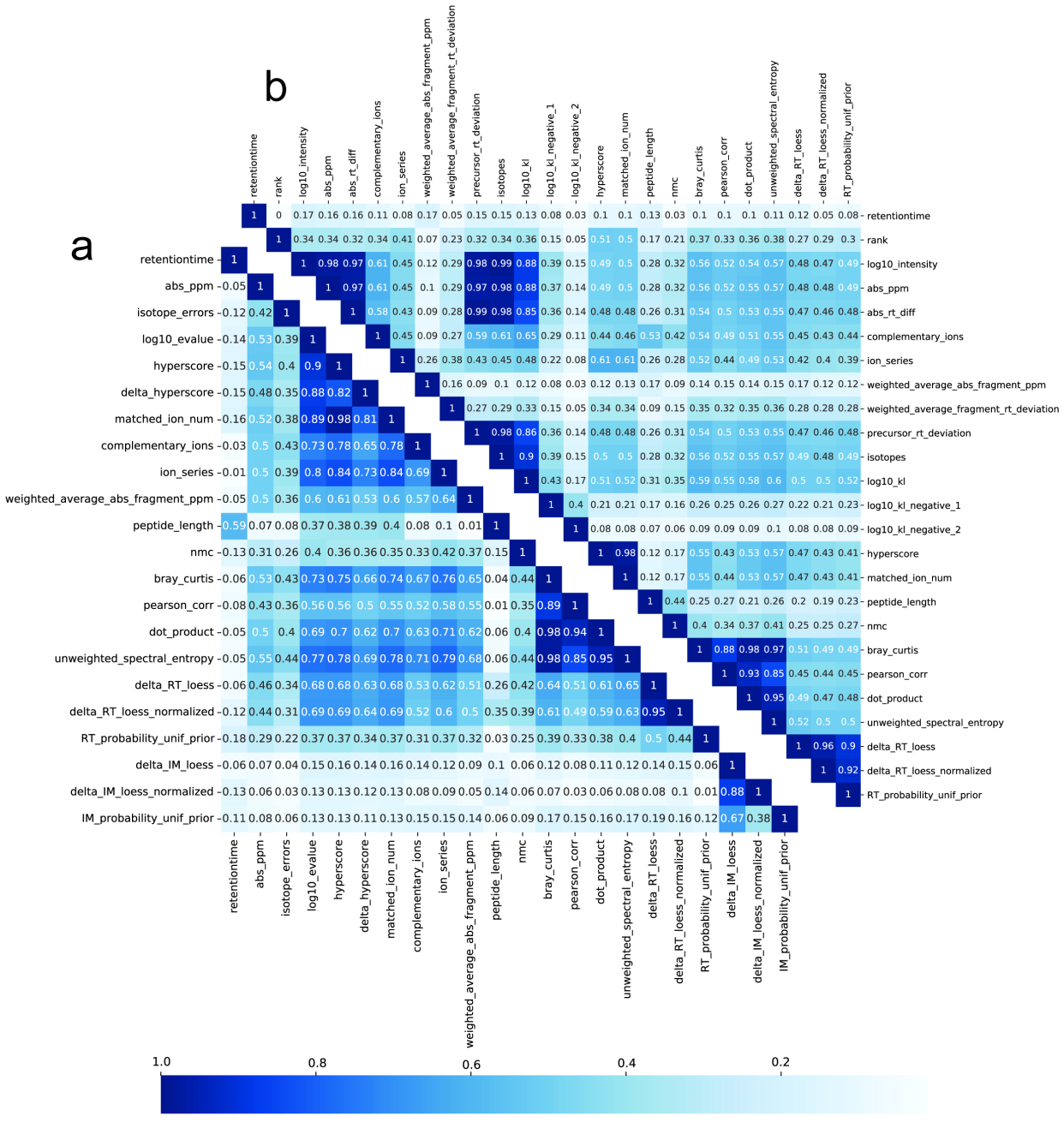
Error bars show the standard deviation from running the software ten times. For each subfigure, the dataset, the number of mzml/pin file pairs, the number of unique peptides in the pin files, and the total number of PSMs are listed. Data are presented as mean values +/- the standard deviation.



Appendix Figure A.2 HLA immunopeptidome PSM rescoring

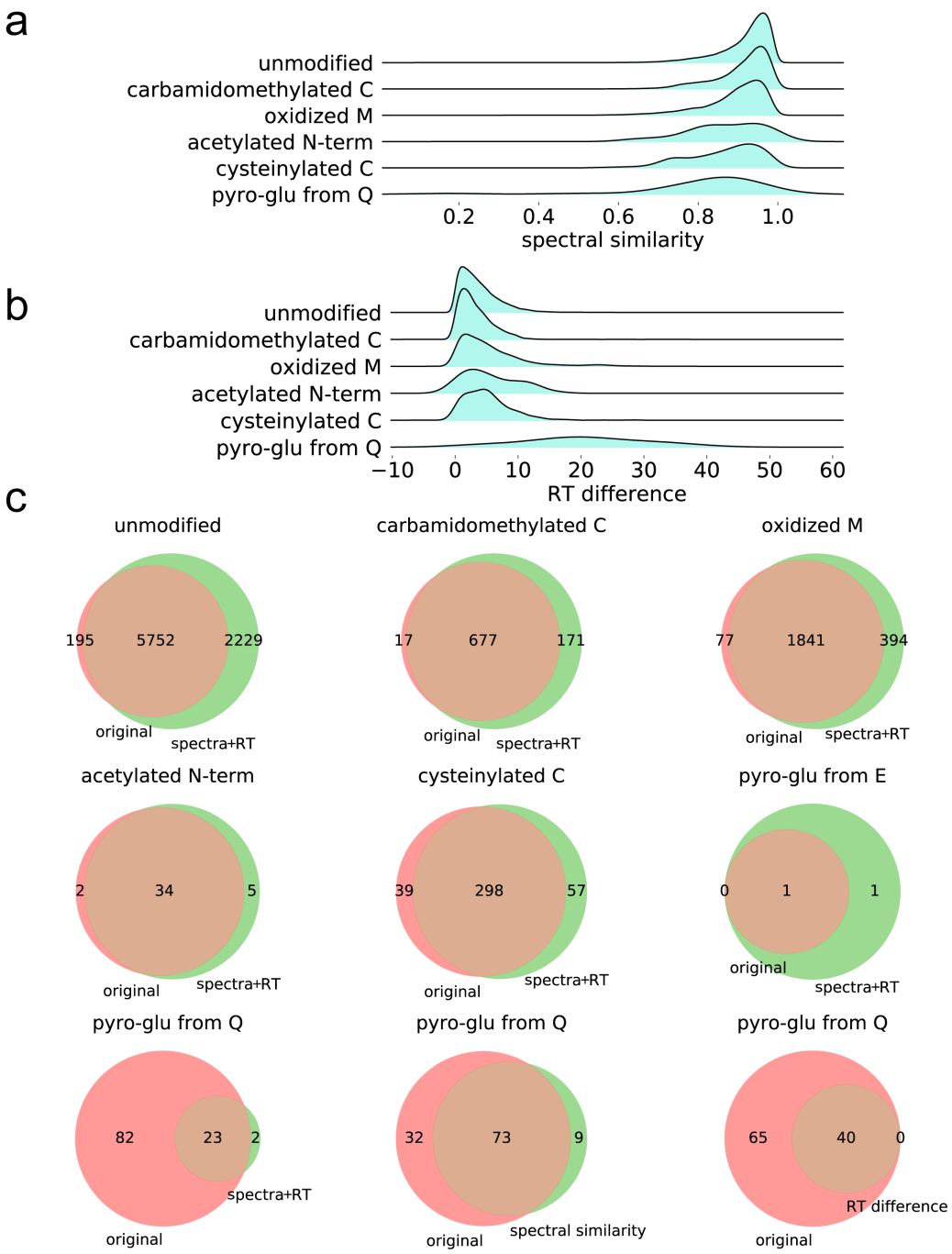
(a) The percent increase in peptides identified when rescoring MaxQuant results with Prosit. Each bar represents the results for a different monoallelic cell line from Sarkizova et al., 2020. (b) Swarmplot of the number of HLA peptides reported in comparison to only using hyperscore or in comparison to MaxQuant rescoring with Prosit. “Original FragPipe” indicates results without using deep learning features, only using

hyperscore and other features calculated by MSFragger. “MSBooster spectra+RT (DIANN)” has the MSBooster generated features added on and is equivalent to “spectra+RT” in Figure 2.2a. (c) GibbsCluster-generated motifs assigned to each peptide subset from the Venn diagram in Figure 2.2b.



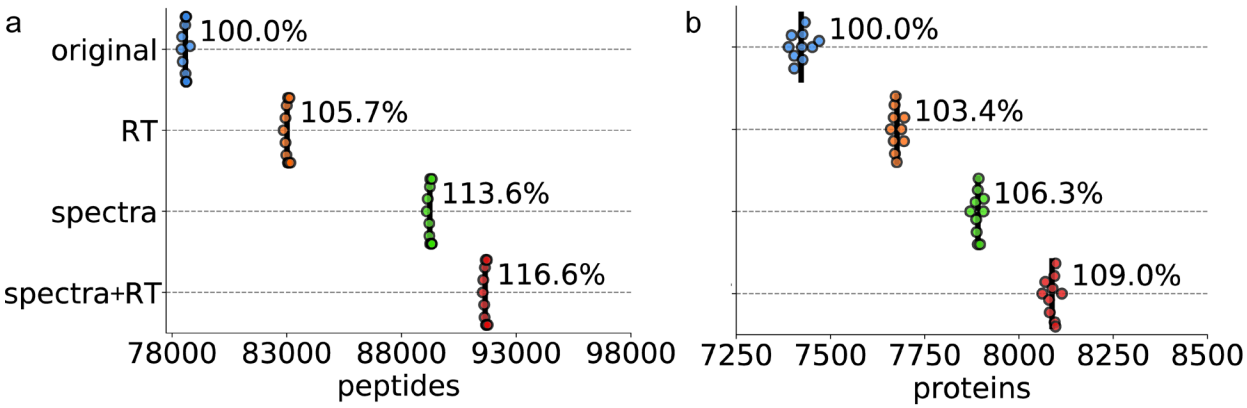
Appendix Figure A.3 Feature correlations

Spearman’s correlation was calculated between MSBooster features and all features reported by MSFragger (a) and MSFragger-DIA (b). (a) was produced from one timsTOF file, (b) from one melanoma file.



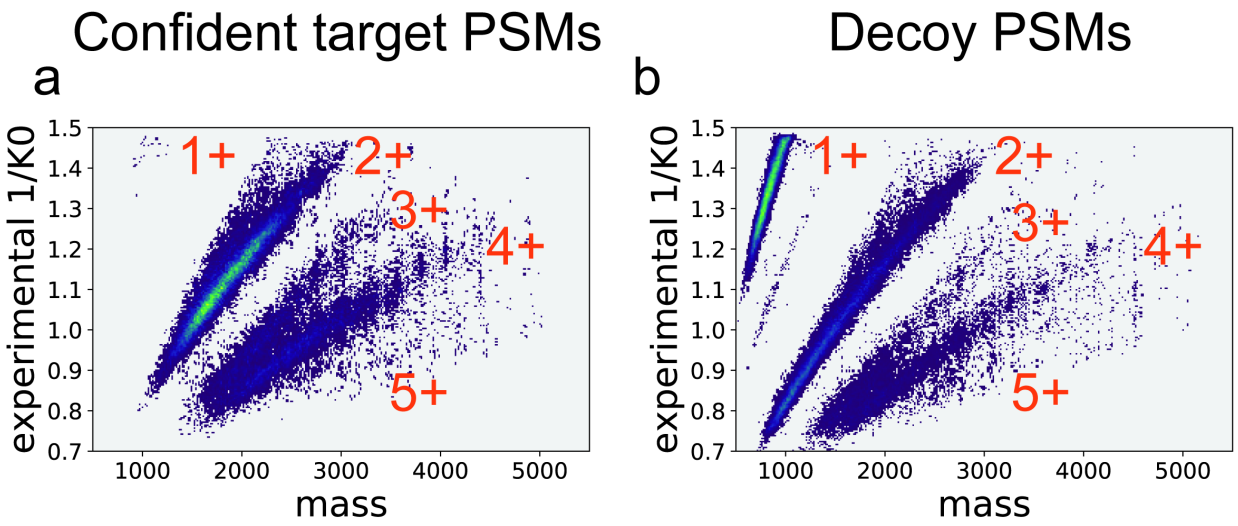
Appendix Figure A.4 Score distributions for unmodified and modified HLA peptides

(a) Spectral similarity values for accepted PSMs at PSM FDR < 1% and peptide FDR < 1%. (b) The same as (a) but for RT difference values. Note that these values are not log-normalized as in other figures. Pyro-glutamation on E was included in the search, but only one was detected and therefore excluded from visualization. (c) Venn diagram of the number of peptides with each modification before and after MSBooster rescoring. The bottom row is for pyroglutamation from Q, either rescored with spectral similarity, RT difference, or both together (spectra+RT).



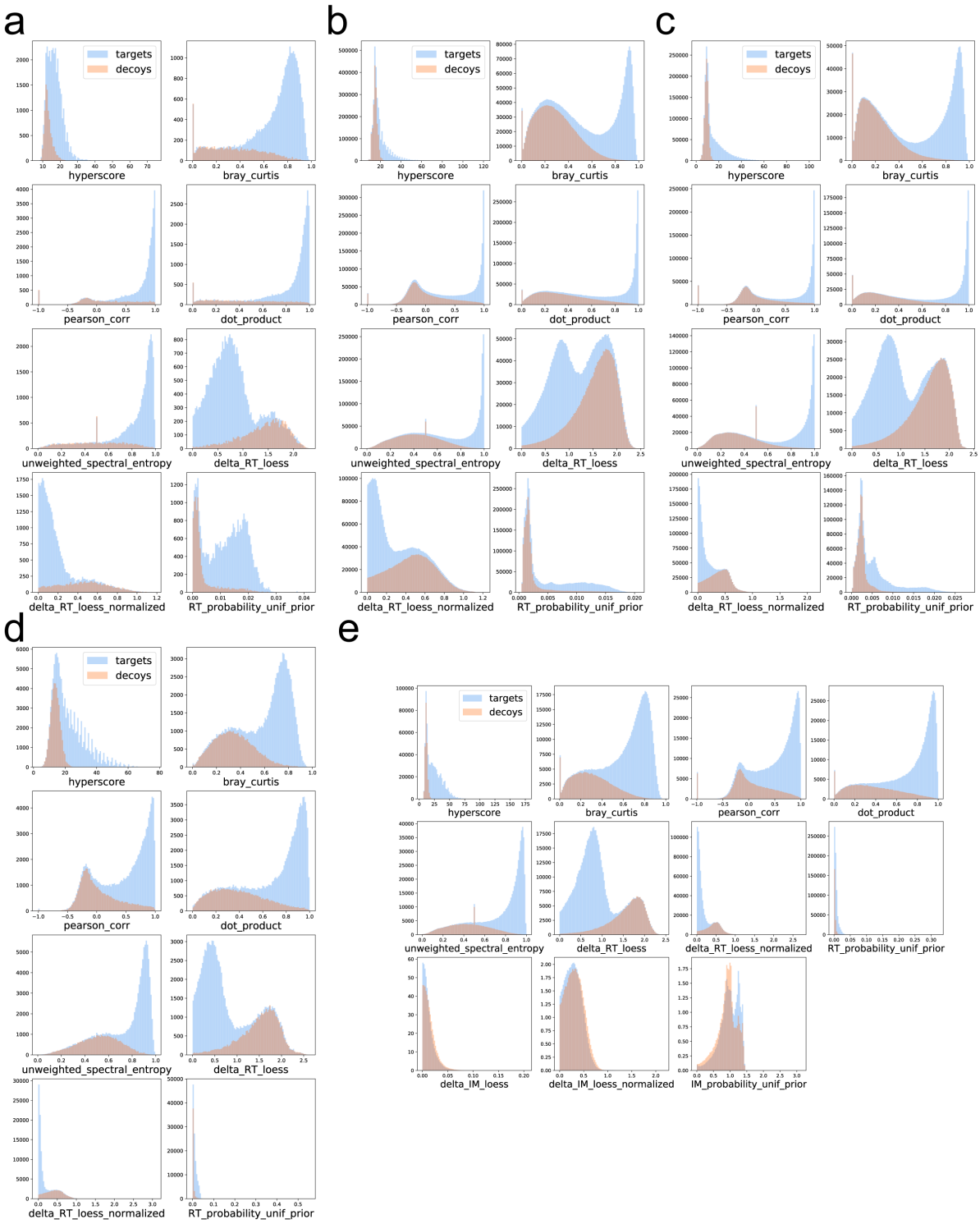
Appendix Figure A.5 Melanoma DIA rescoring with DIA-Umpire

Swarmplots of the number of peptides (a) or proteins (b) reported at 1% FDR.



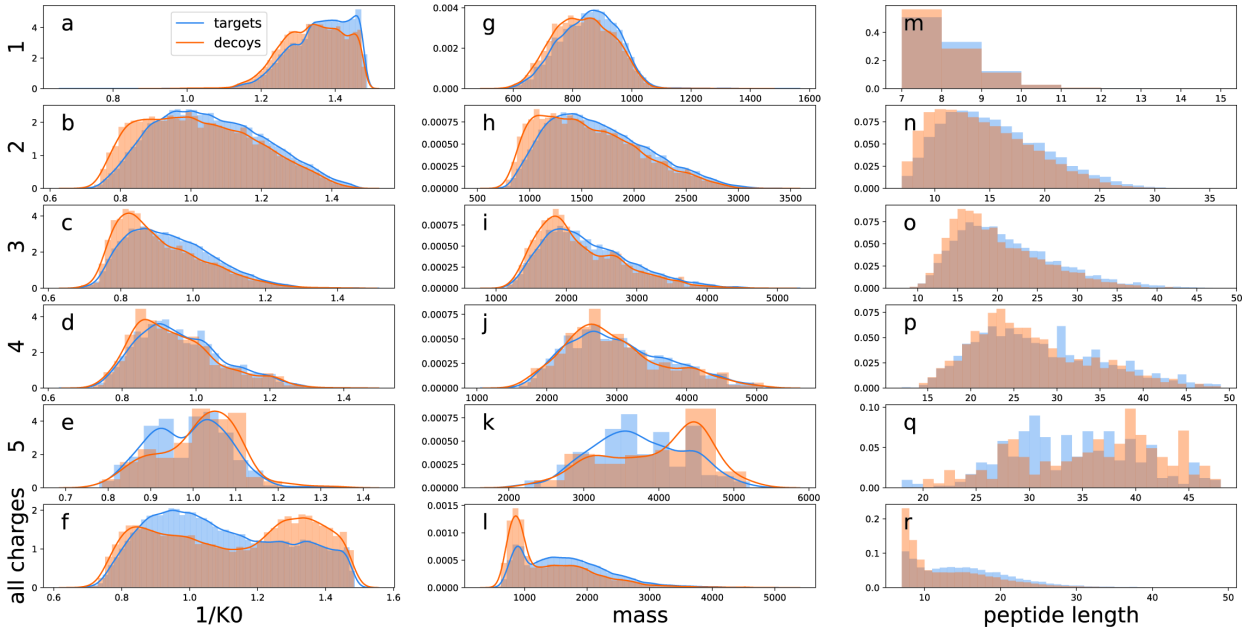
Appendix Figure A.6 Inverse ion mobility dependence 1/K0 on mass and charge

(a-b) Scatter density plots are shown for (a) confident target PSMs and (b) decoy PSMs, where brighter colors represent higher density. Figures a-b were generated using data from a single DDA PASEF run.



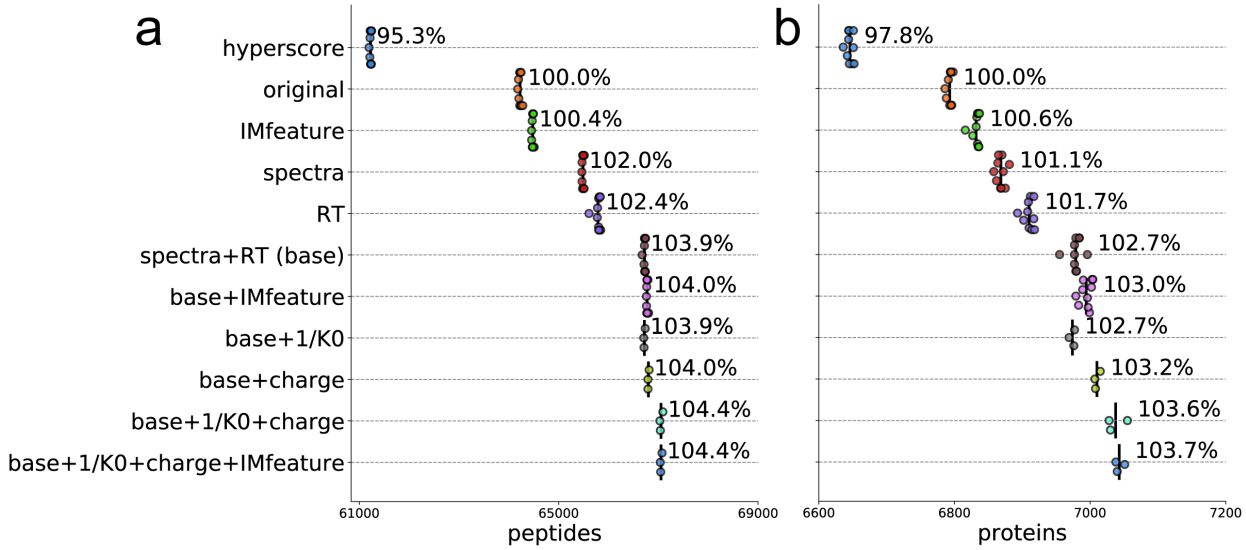
Appendix Figure A.7 PSM score distributions for MSBooster features

Target PSMs are in blue, decoy PSMs in orange. The datasets analyzed were HLA (a), melanoma processed by MSFragger DIA (b), melanoma processed by DIA-Umpire with MSFragger (c), single cell nanoPOTS (d), and timsTOF (e). RT and IM features were log-transformed to better visualize target-decoy separation using the conversion log10(value + 1). Spectral and RT figures have counts on the y-axis, IM has density. Targets are depicted in blue, decoys in orange.



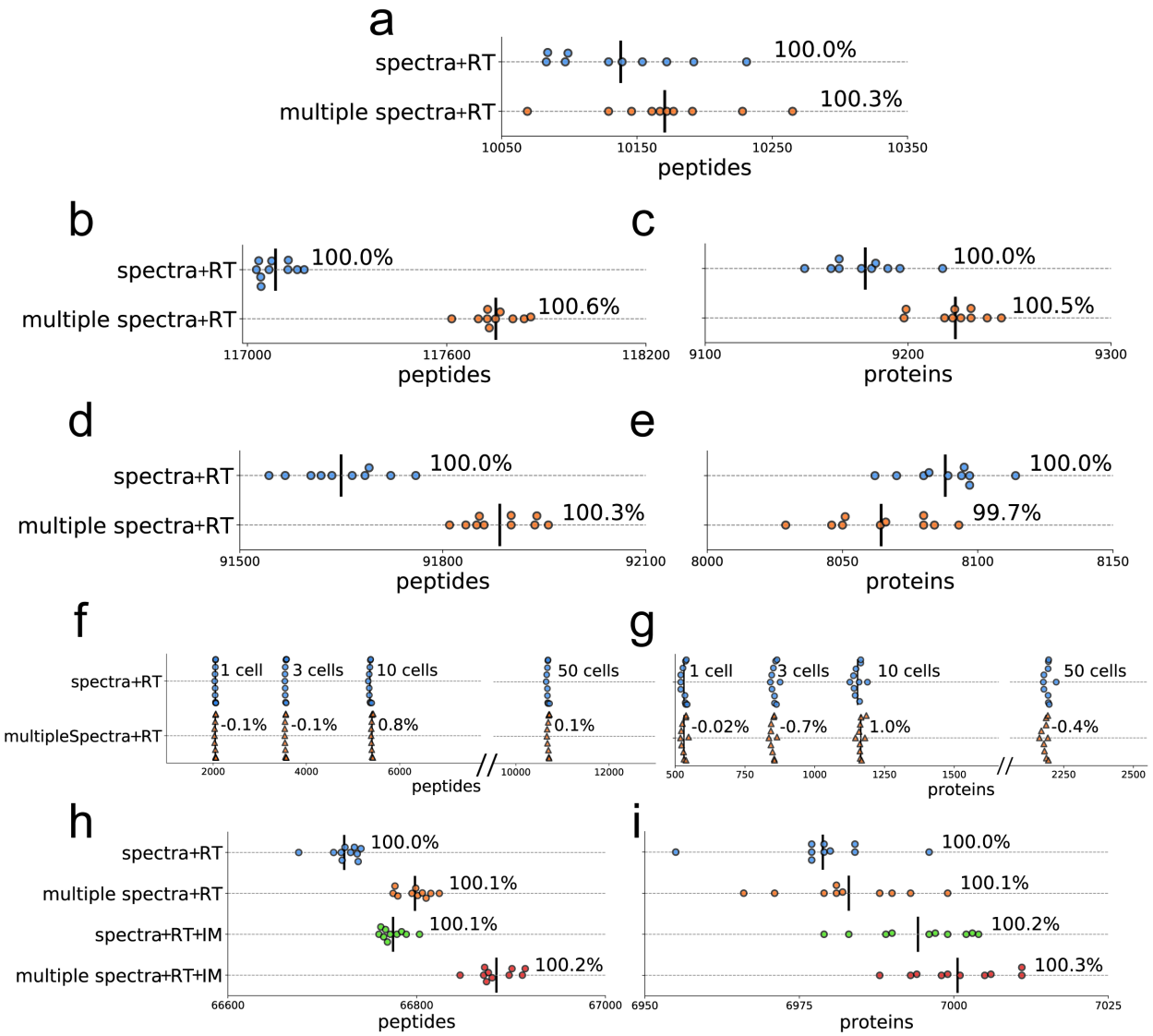
Appendix Figure A.8 Target-decoy PSM distributions for different charge states and features

Target-decoy distributions for inverse ion mobility (a-f), mass (g-l), and peptide length (m-r) are shown for precursors of charge 1-5 or all charges together. All subfigures show relative density rather than PSM counts to better show target-decoy separation. Targets are depicted in blue, decoys in orange.



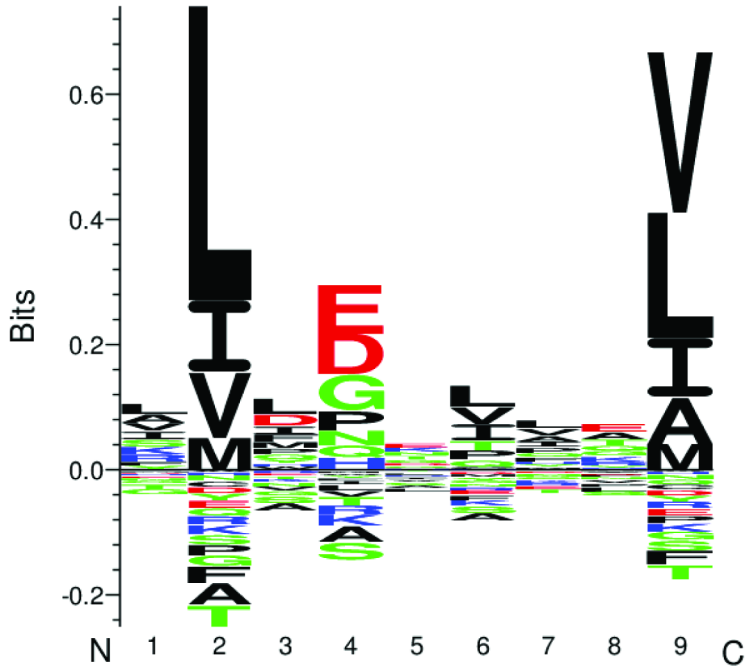
Appendix Figure A.9 IM rescoring of timsTOF HeLa dataset

Swarmplot of peptides (a) and proteins (b) reported at 1% FDR for the timsTOF Hela dataset. The "IMfeature" used was IM probability uniform prior. 1/K0 is the experimental inverse ion mobility. Charge was one-hot encoded between charges 1 and 7. "Base" represents the default in FragPipe with only spectral and RT similarity features.



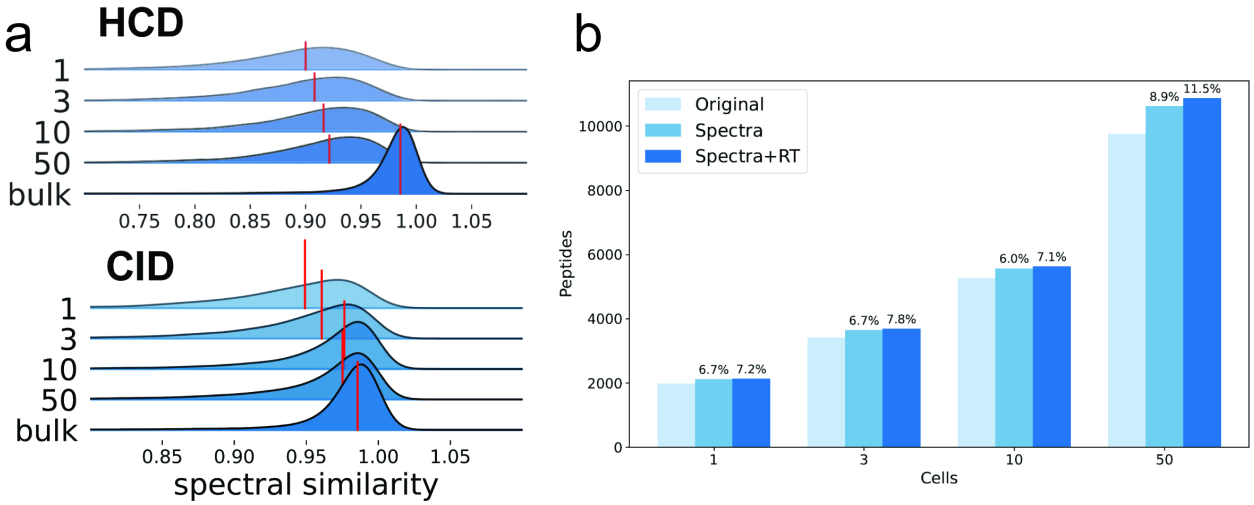
Appendix Figure A.10 Utility of multiple correlated features

Single features (unweighted spectral entropy + delta RT LOESS = spectra+RT; unweighted spectral entropy + delta RT LOESS + IM probability uniform prior = spectra+RT+IM) are compared against multiple features for Percolator rescoring. The numbers reported are for HLA peptides (a), melanoma peptides (b) and proteins (c) with MSFragger-DIA, melanoma peptides (d) and proteins (e) with DIA-Umpire and MSFragger, single cell peptides (f) and proteins (g), and timsTOF peptides (h) and proteins (i).



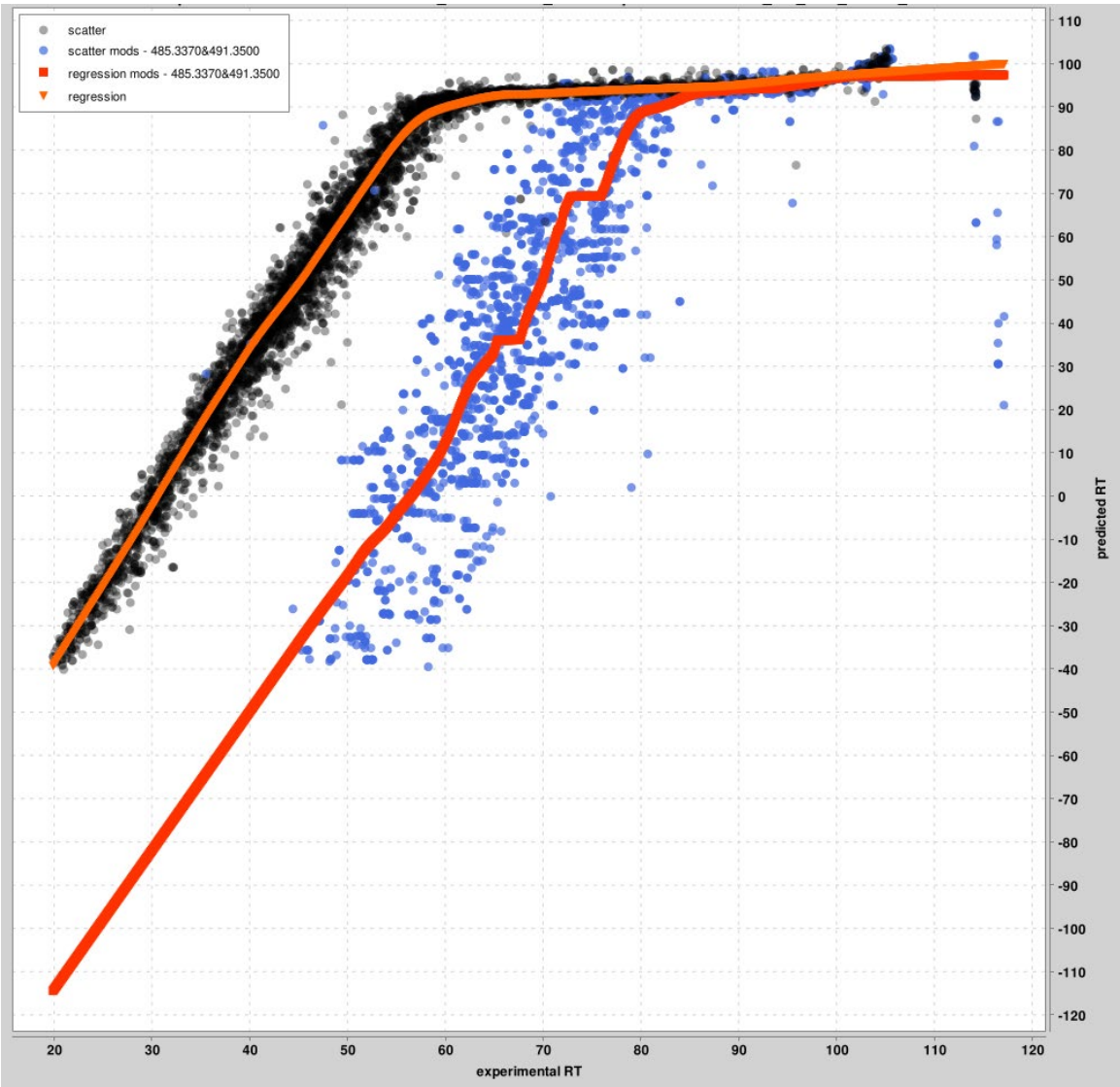
Appendix Figure A.11 HLA motif with multiple correlated features

GibbsCluster-generated motifs for peptides gained from using multiple correlated features in the A*02:01 monoallelic cell line data from Klaeger et al., 2021.



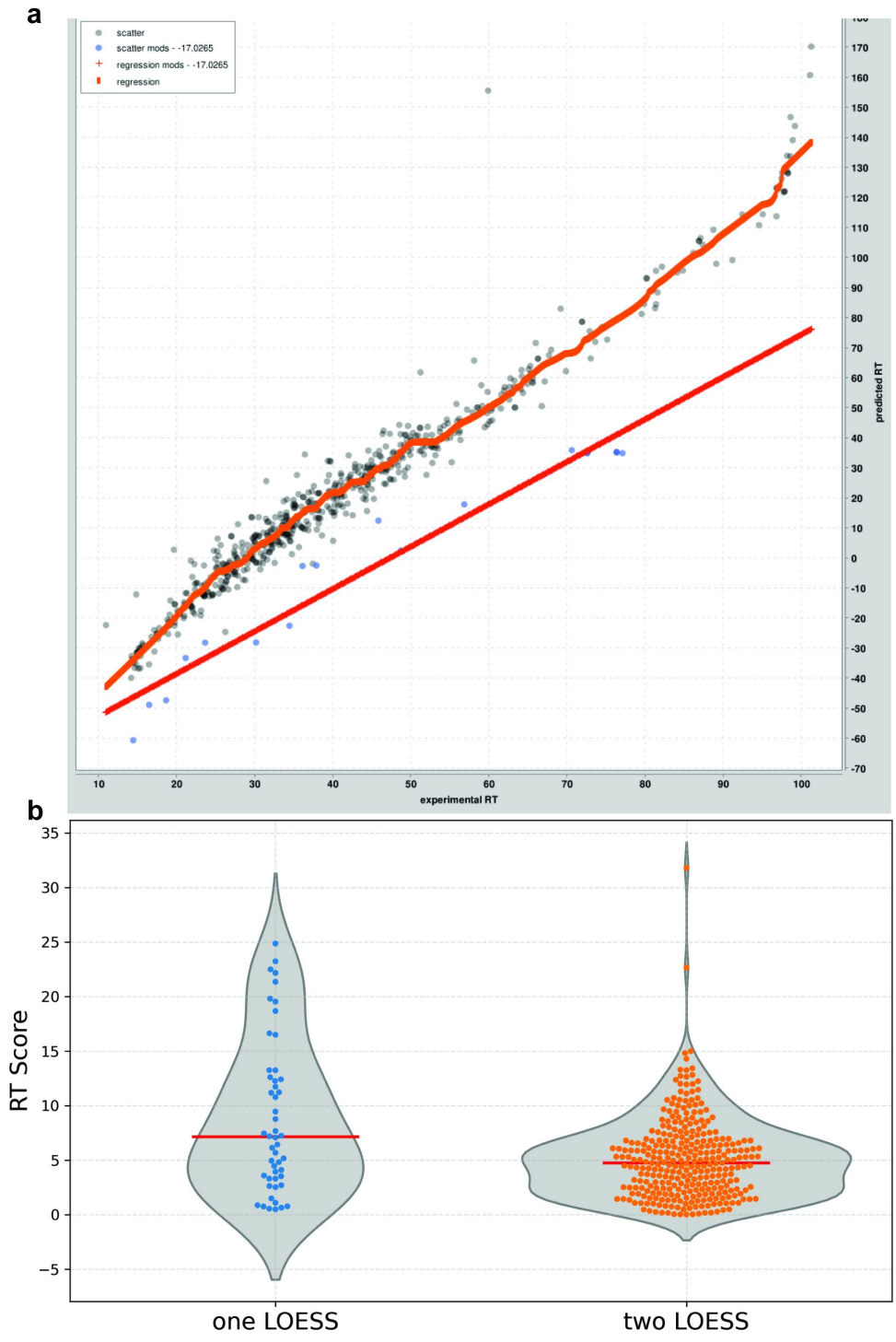
Appendix Figure A.12 Updated nanoPOTS rescoring

(a) A comparison of confident target unweighted spectral entropy scores when using DIA-NN's HCD predictions or Prosit CID's predictions. Various numbers of cells were sequenced via the nanoPOTS platform (Williams et al., 2020), and bulk cell sequencing was from An et al., 2021. The red line indicates the median spectral similarity. (b) Bar plots of peptides identified when using the Prosit CID model. The percentages displayed show the percent gain over original (no MSBooster) for each cell number group.



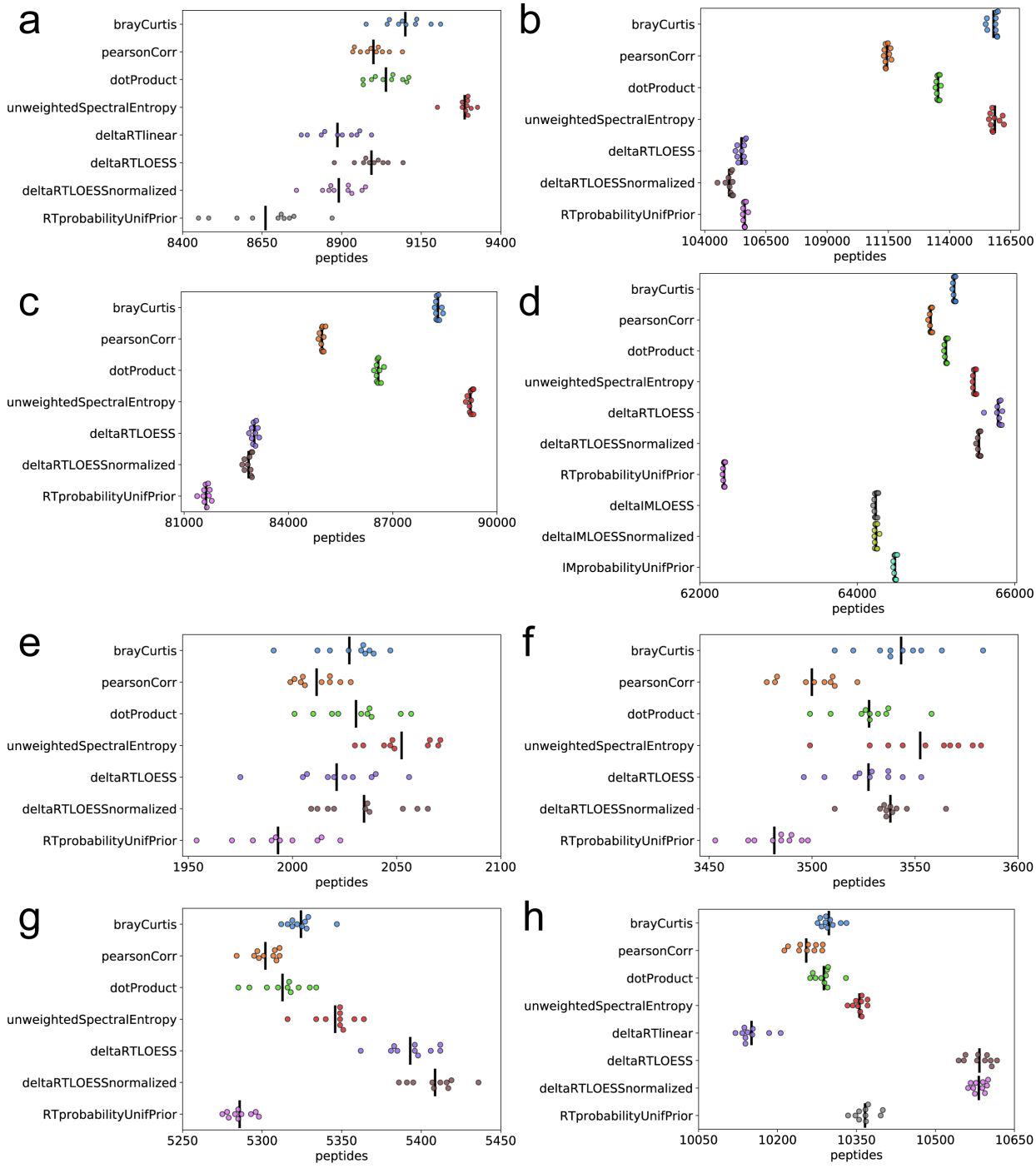
Appendix Figure A.13 A separate RT calibration curve for mass offsets

Masses set as mass offsets in FragPipe experiments have a separate calibration curve (in red) from other peptides not containing the mass offsets (calibration curve in orange). Each dot of the scatter plot shows a point used for RT calibration plotted at its experimental and predicted RTs. Mass offset PSMs are plotted in blue, others in black. Data is from one replicate run of Probe 4 (AMJ 14) from Wozniak et al., 2024. Light (485 Da) and heavy (491 Da) tags are pooled together and share a calibration curve.



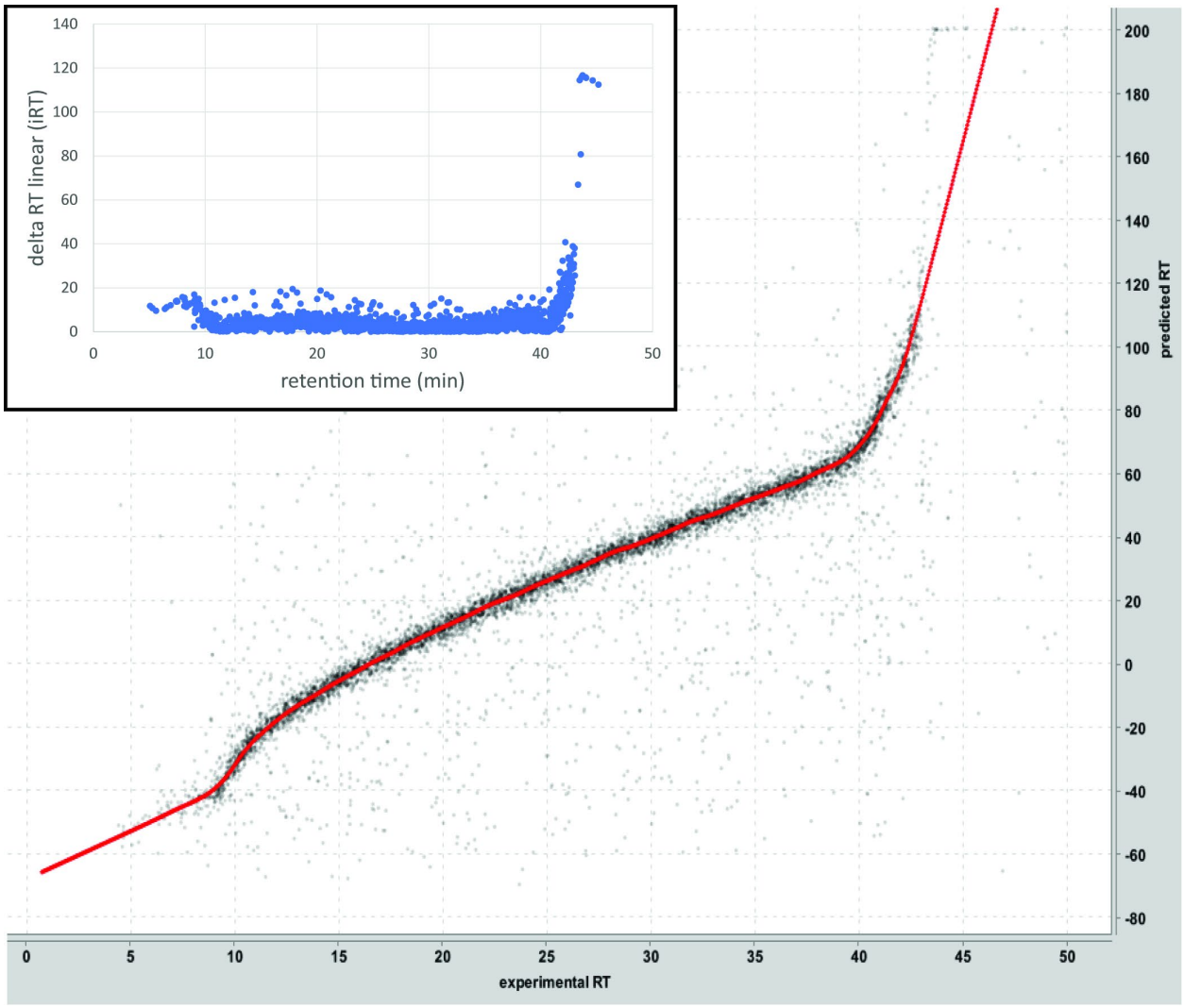
Appendix Figure A.14 Retention time calibration for pyro-glu from Q

(a) Separate LOESS curves are plotted for pyro-glu from Q PSMs (blue points with red line) and all other PSMs (grey points with orange curve). (b) RT scores for pyro-glu from Q PSMs that pass 1% PSM and peptide FDR when using one LOESS curve for all PSMs or two curves. The unit of the RT score is minutes. The red lines indicate median RT scores for each group.



Appendix Figure A.15 Peptides reported when using different MSBooster features

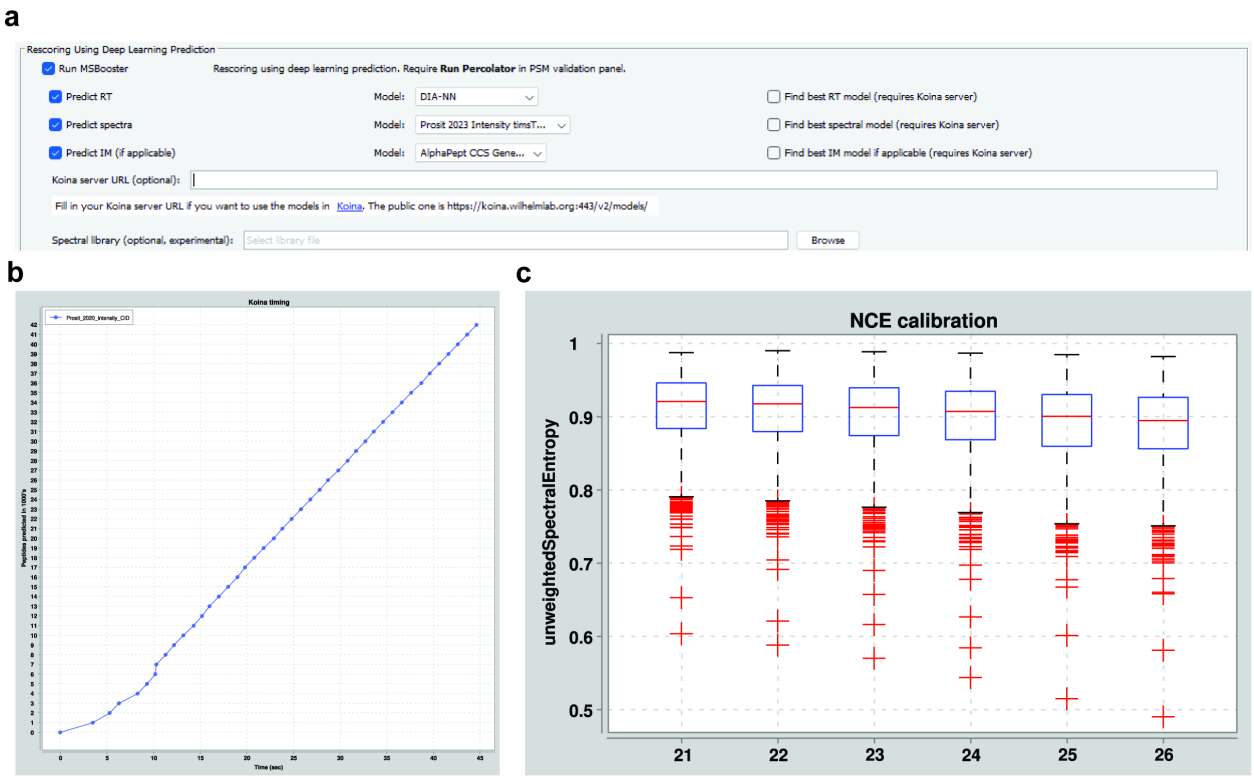
Swarmplots for HLA (a), melanoma MSFragger DIA (b), melanoma DIAU (c), timsTOF (d), 1 cell (e), 3 cells (f), 10 cells (g), 50 cells (h). Each dataset was processed 10 times by Percolator, each time with a different random seed 1-10. The black line indicates mean peptides reported over the 10 runs.



Appendix Figure A.16 A plotted calibration curve between the experimental and predicted RT scales

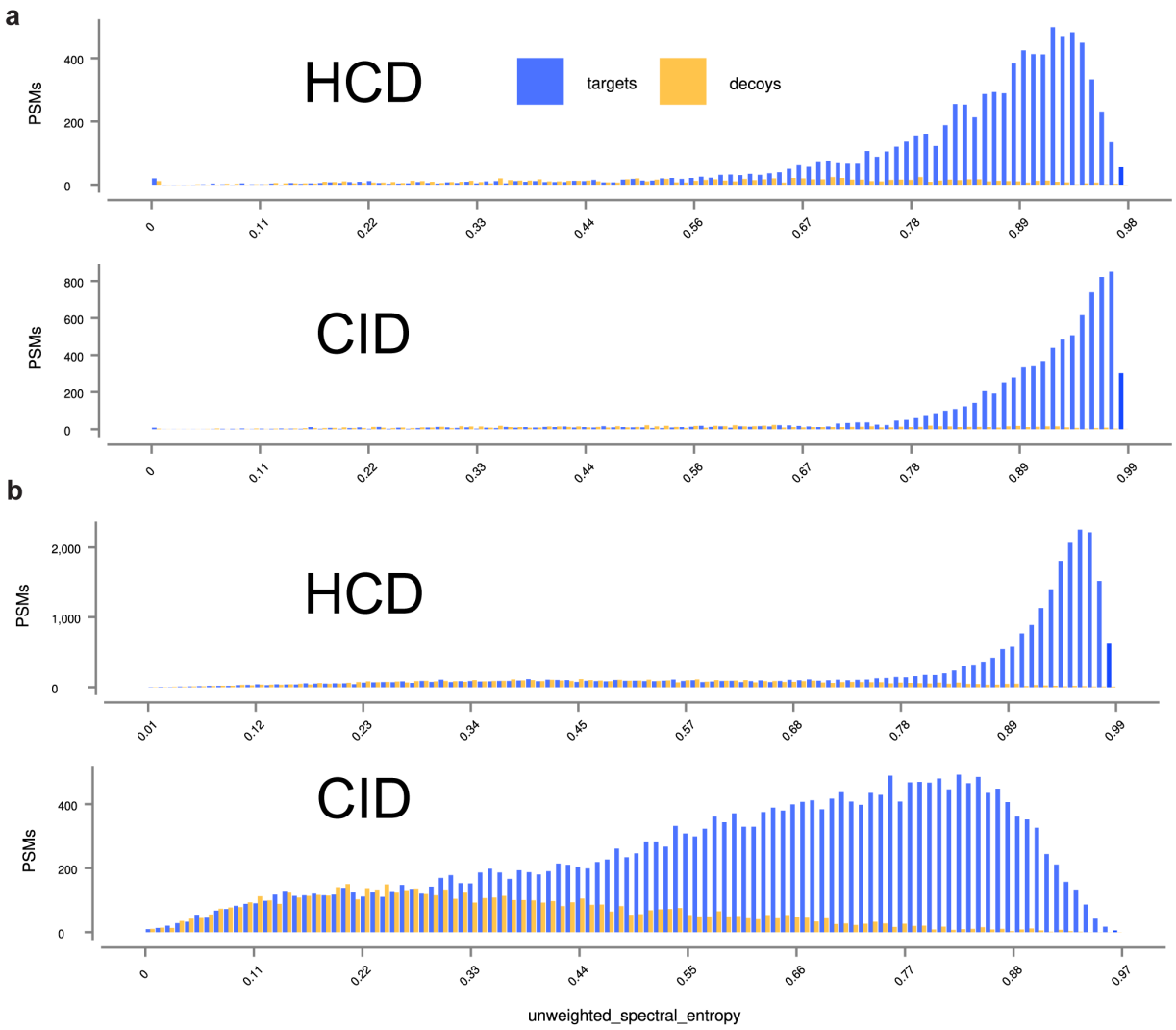
The data shown here is from one replicate of a 50-cell run from the nanoPOTS data (Williams et al.). The red line represents the values estimated by LOESS regression. Individual black dots are those PSMs with spectral similarity greater than 0 and log(expectation value) less than a user-defined threshold, which is by default -3.5. The inset shows the relationship between the experimental RT and difference between predicted and experimental RT when using only a linear regression for calibration (delta RT linear).

Appendix B: Supplemental Material for Chapter 3



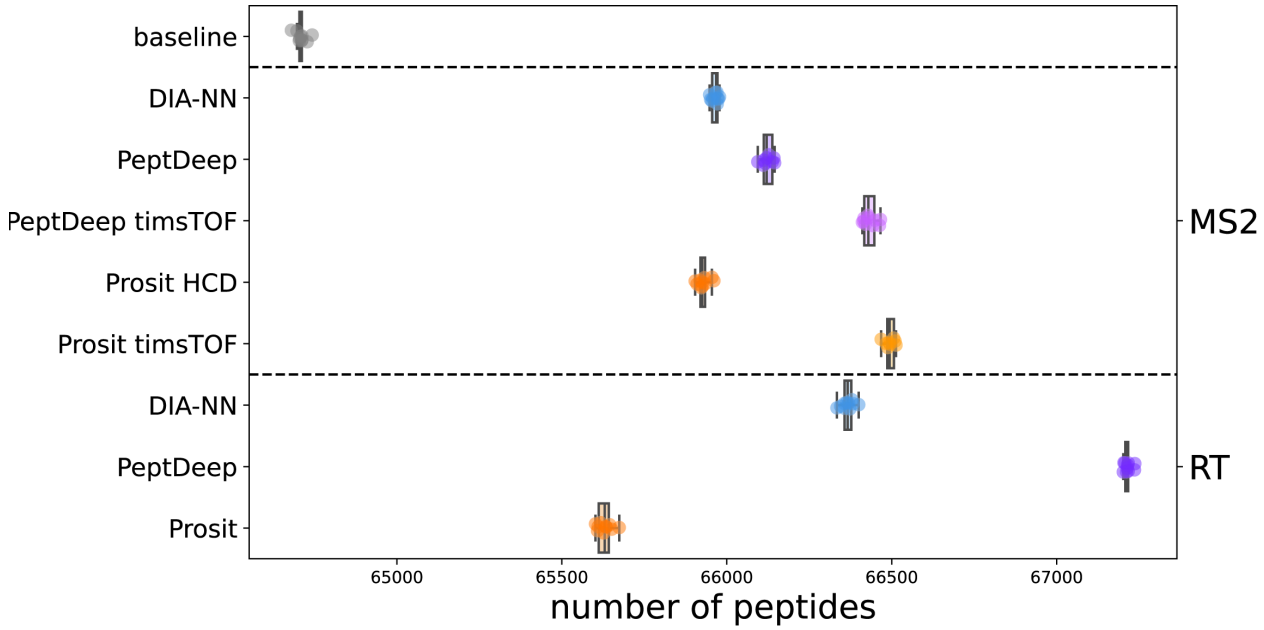
Appendix Figure B.1 MSBooster interface and Koina figures in FragPipe

(a) Screenshot of the MSBooster section under the FragPipe Validation tab. RT, IM and spectral features can be used together or separately, with a dropdown menu to choose a model. If "Find best RT/spectra/IM model" is checked, then a heuristic algorithm is enabled to attempt to determine the combination of models that maximizes peptides identified. One can also manually set an RT model but enable search for the best spectra model, and vice versa. A textbox is included to set the path to the Koina server, which may be a private locally hosted one, or the public one provided. (b) Peptides are submitted in batches of 1000 to the Koina server. Total time taken is recorded on the x-axis. (c) Similarity box-and-whisker plot across sequential normalized collision energy (NCE) values for the PSMs collection during the NCE calibration step. The top whisker is the third quartile plus 1.5 times the interquartile range.



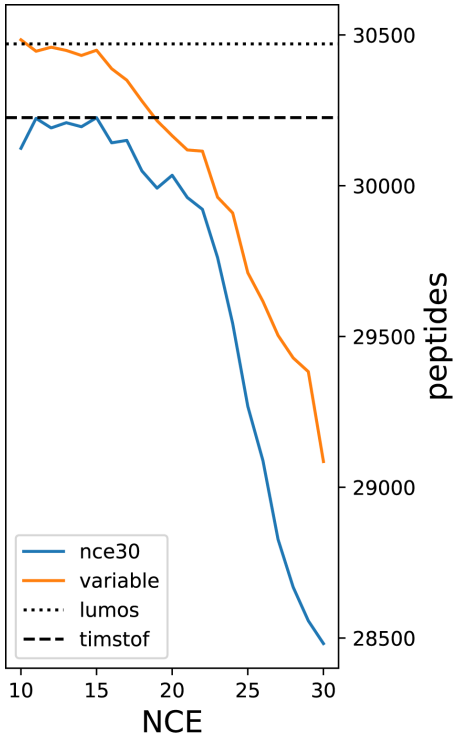
Appendix Figure B.2 HCD vs CID PSM score distributions

Target and decoy PSM distributions for the unweighted spectral entropy MS/MS similarity feature when comparing experimental fragment ion intensities with predicted intensities from the Prosit_2020_intensity HCD and CID models. Distributions are shown for PSMs from one pin file each from a) Marcu et al., 2021 and b) Pak et al., 2021.



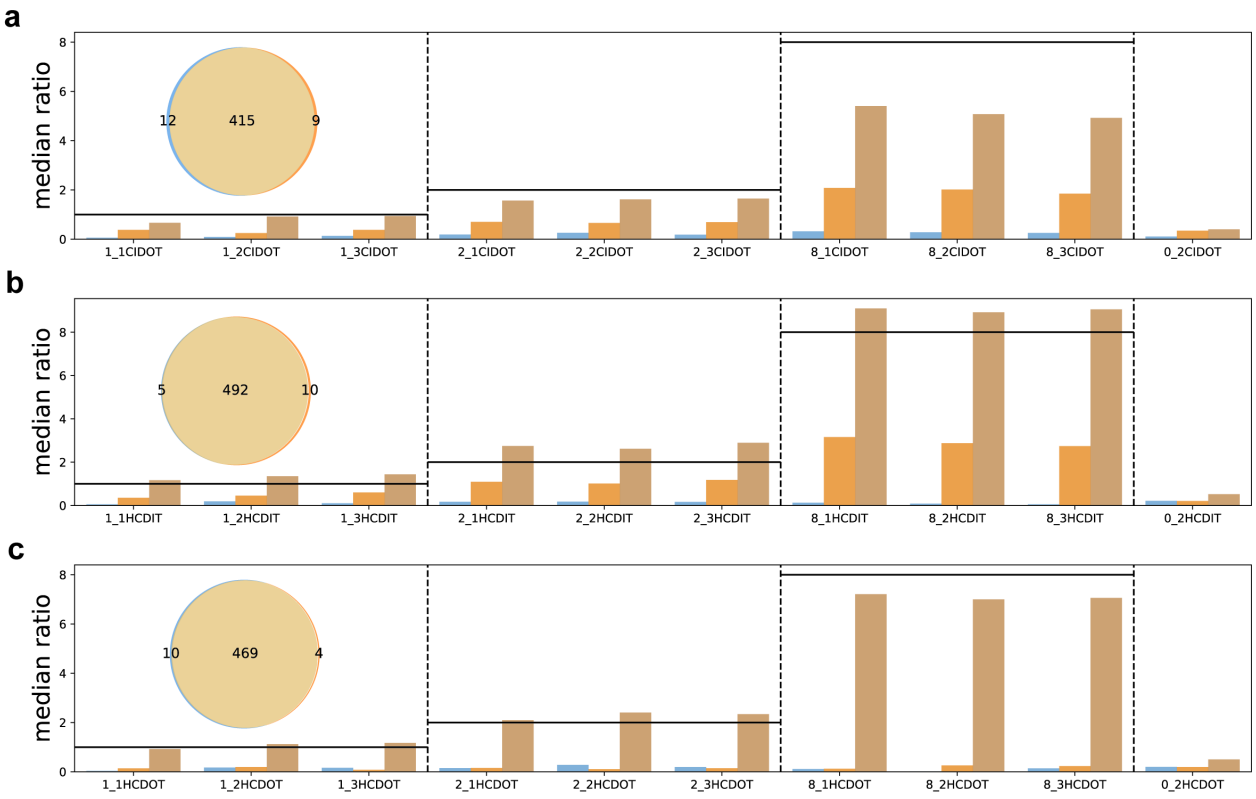
Appendix Figure B.3 Tryptic peptide IDs on a timsTOF

Swarmplot and box-and-whisker plot of the peptides identified in data collected on a timsTOF mass spectrometer from Meier et al., 2018.



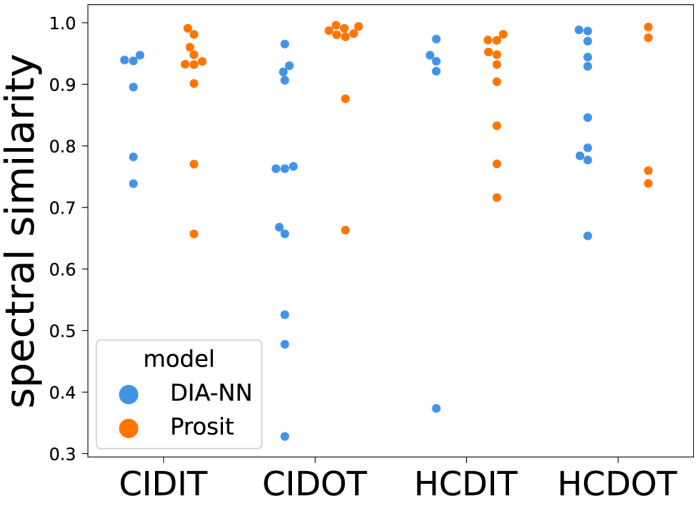
Appendix Figure B.4 AlphaPeptDeep transfer learning with HLA-I peptides on the timsTOF

Each eV value was tested three times using three different Percolator random seeds. The dashed lines for "lumos" and "timsTOF" indicate the average peptides identified by the PeptDeep and PeptDeep timsTOF models respectively in Figure 3.2f.



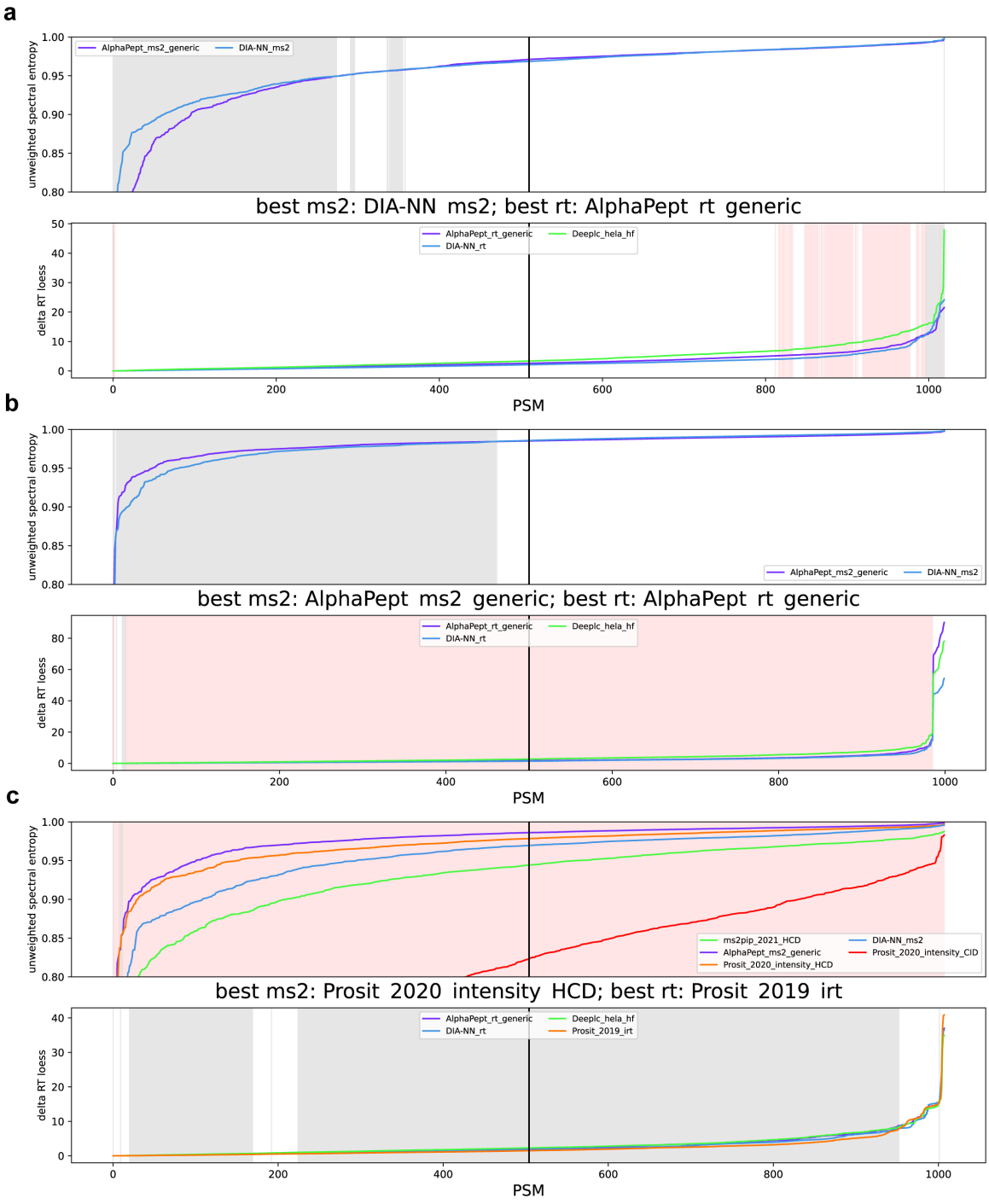
Appendix Figure B.5 Quantification accuracy in the human-yeast dataset

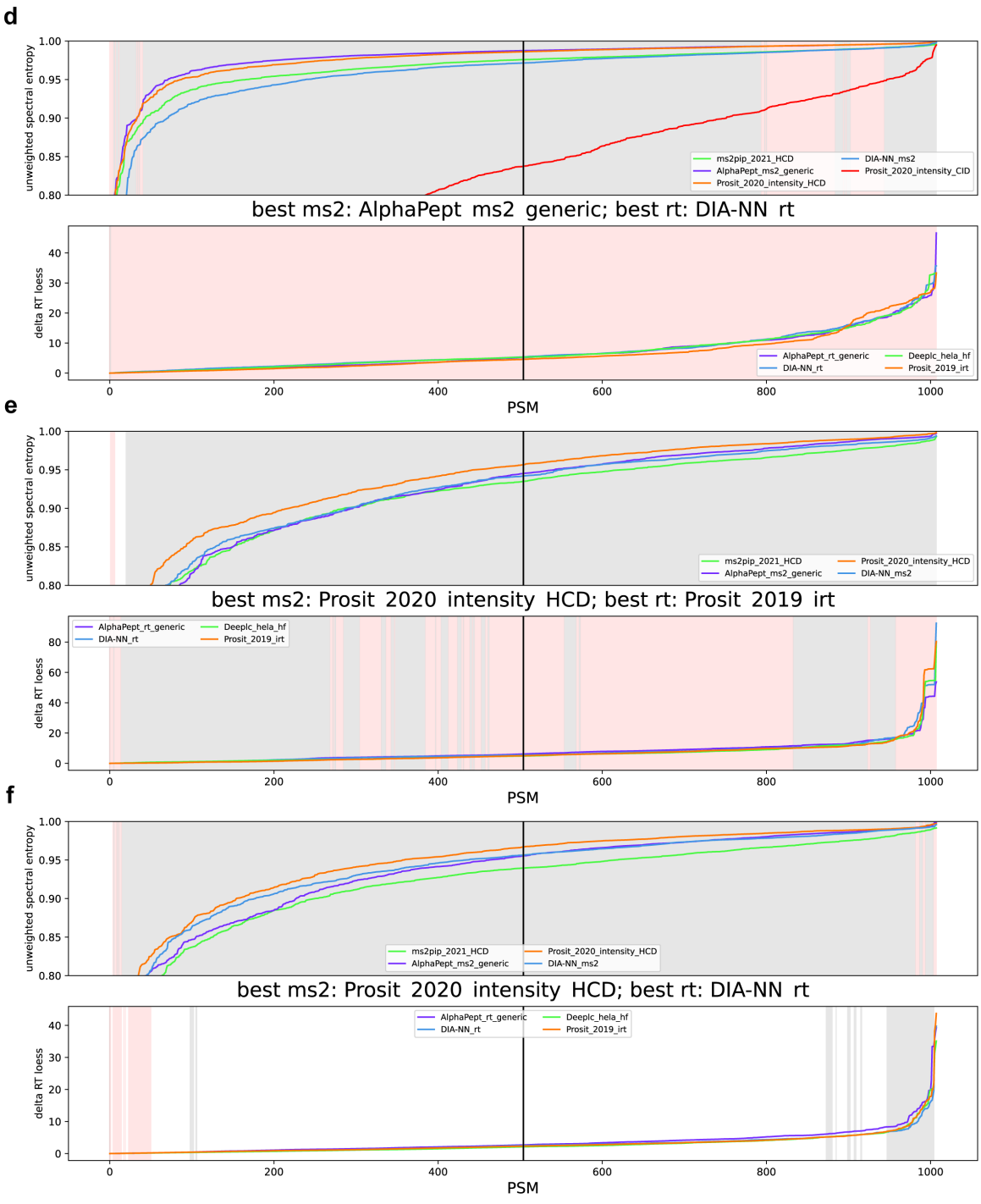
Quantified yeast peptides from DIA-NN and Prosit were compared via Venn diagram for CID OT (a), HCD IT (b), and HCD OT (c). For each subset of this Venn diagram, the median ratio of the peptides in each channel to the reference 126 TMT channel is visualized, with a horizontal black line indicating what the expected ratio is.

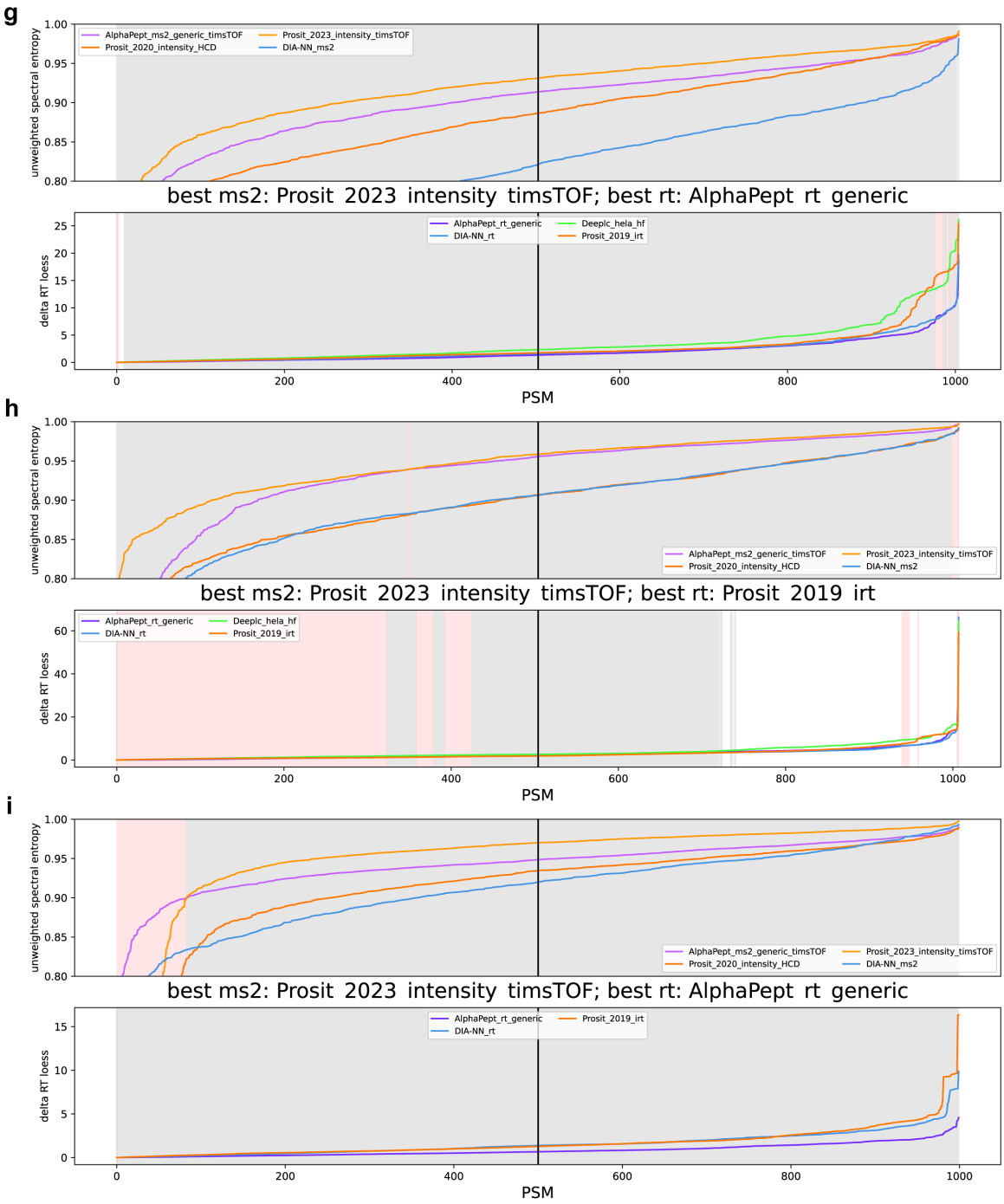


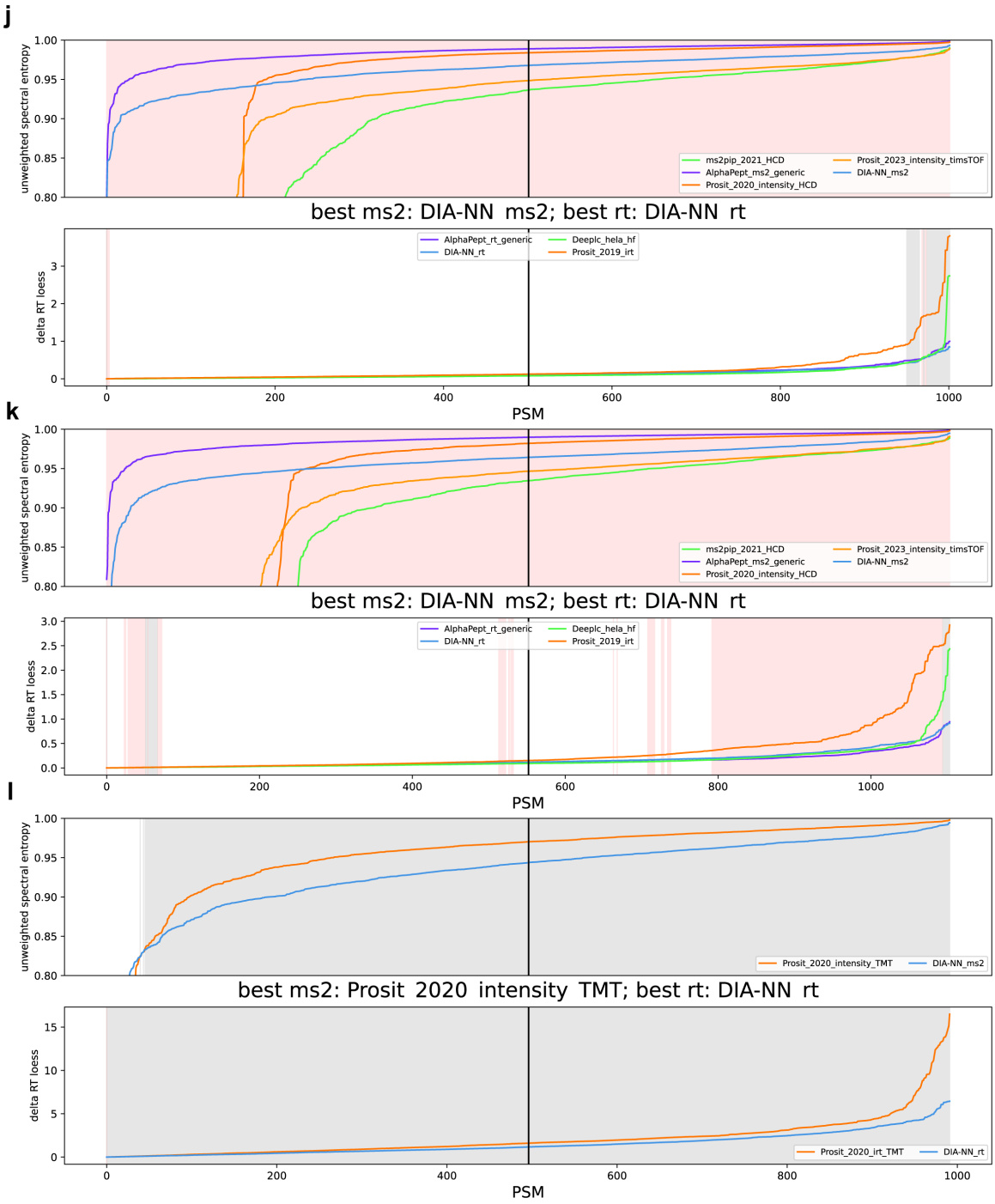
Appendix Figure B.6 Spectral similarity of unique peptides in the human-yeast dataset

For the unique peptides from each model, the highest spectral similarity for PSMs from each peptide are plotted.


a
unweighted spectral entropy
AlphaPept_ms2_generic
DIA-NN_ms2
best ms2: DIA-NN ms2; best rt: AlphaPept rt generic
delta RT loess
AlphaPept_rt_generic
DIA-NN_rt
Deeplc_hela_hf
PSM
b
unweighted spectral entropy
AlphaPept_ms2_generic
DIA-NN_ms2
best ms2: AlphaPept ms2 generic; best rt: AlphaPept rt generic
delta RT loess
AlphaPept_rt_generic
DIA-NN_rt
Deeplc_hela_hf
PSM
c
unweighted spectral entropy
ms2pip_2021_HCD
AlphaPept_ms2_generic
Prosit_2020_intensity_HCD
DIA-NN_ms2
Prosit_2020_intensity_CID
best ms2: Prosit 2020 intensity HCD; best rt: Prosit 2019 irt
delta RT loess
AlphaPept_rt_generic
DIA-NN_rt
Deeplc_hela_hf
Prosit_2019_irt
PSM

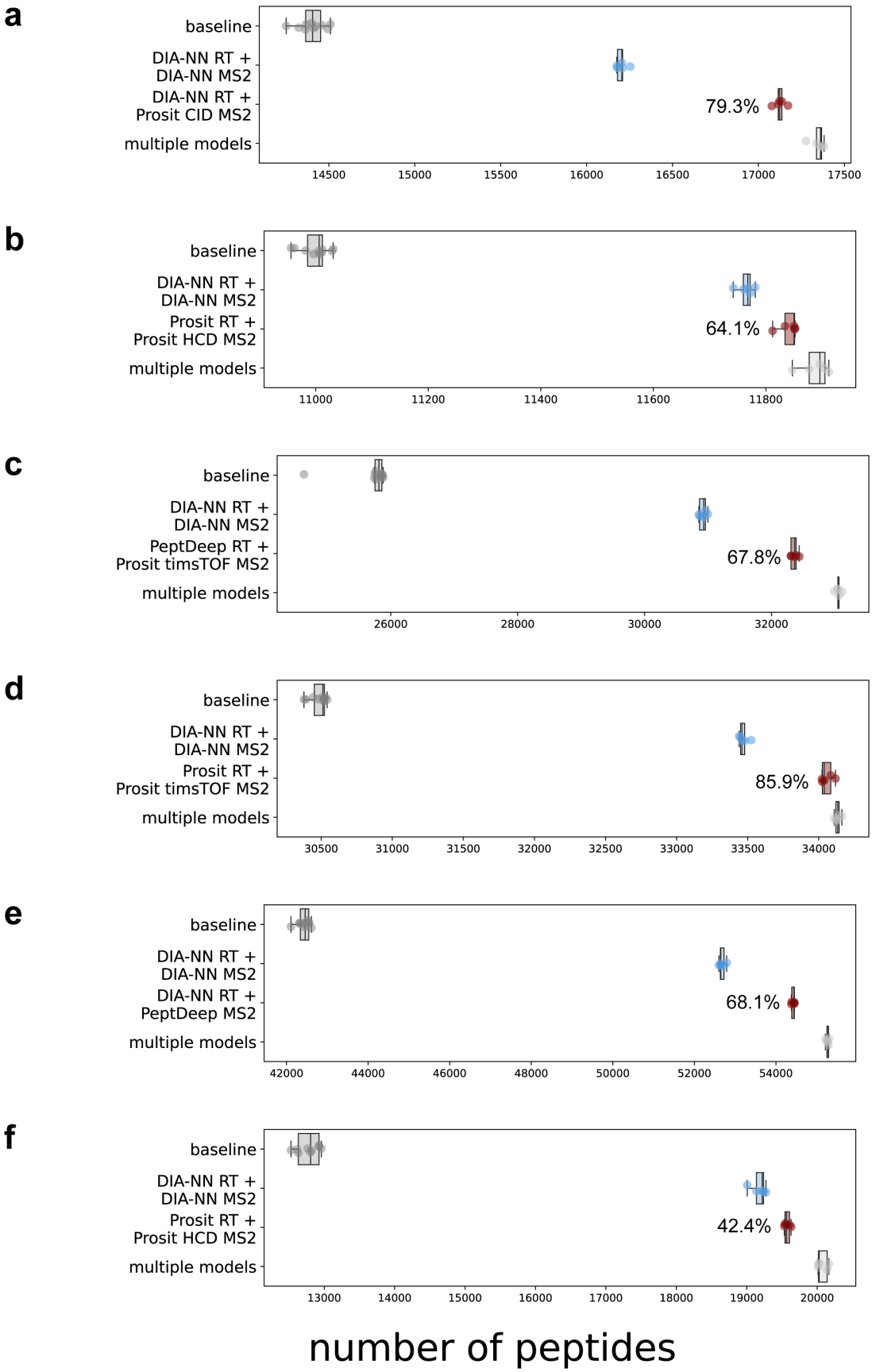

d
unweighted spectral entropy
ms2pip_2021_HCD
AlphaPept_ms2_generic
Prosit_2020_intensity_HCD
DIA-NN_ms2
Prosit_2020_intensity_CID
best ms2: AlphaPept_ms2_generic; best rt: DIA-NN_rt
delta RT loess
AlphaPept_rt_generic
DIA-NN_rt
Deeplc_hela_hf
Prosit_2019_irt
PSM
e
unweighted spectral entropy
ms2pip_2021_HCD
AlphaPept_ms2_generic
Prosit_2020_intensity_HCD
DIA-NN_ms2
best ms2: Prosit_2020_intensity_HCD; best rt: Prosit_2019_irt
delta RT loess
AlphaPept_rt_generic
DIA-NN_rt
Deeplc_hela_hf
Prosit_2019_irt
PSM
f
unweighted spectral entropy
ms2pip_2021_HCD
AlphaPept_ms2_generic
Prosit_2020_intensity_HCD
DIA-NN_ms2
best ms2: Prosit_2020_intensity_HCD; best rt: DIA-NN_rt
delta RT loess
AlphaPept_rt_generic
DIA-NN_rt
Deeplc_hela_hf
Prosit_2019_irt
PSM

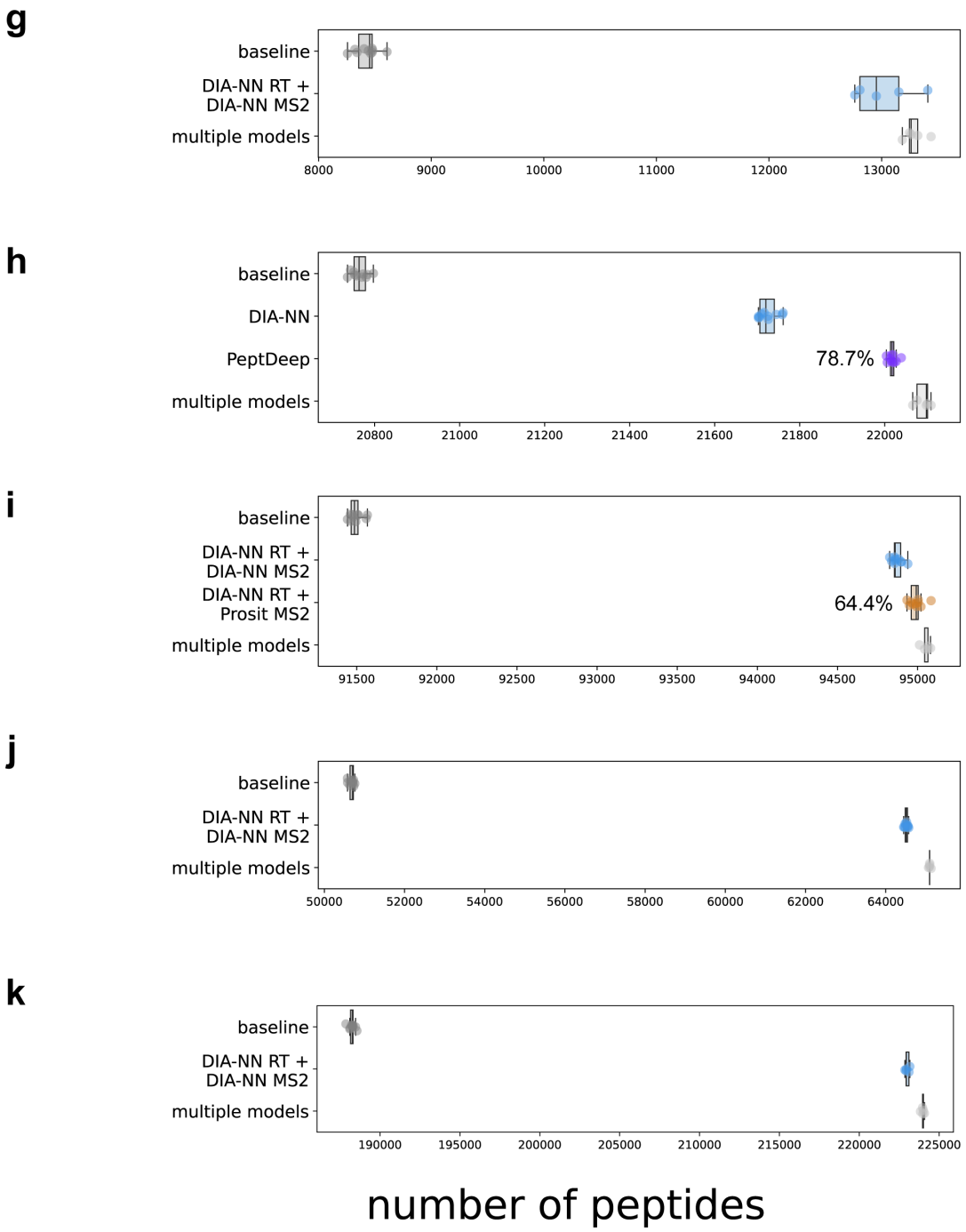

g
AlphaPept_ms2_generic_timsTOF
Prosit_2023_intensity_timsTOF
Prosit_2020_intensity_HCD
DIA-NN_ms2
unweighted spectral entropy
best ms2: Prosit 2023 intensity timsTOF; best rt: AlphaPept rt generic
AlphaPept_rt_generic
Deeplc_hela_hf
DIA-NN_rt
Prosit_2019_irt
delta RT loess
PSM
h
best ms2: Prosit 2023 intensity timsTOF; best rt: Prosit 2019 irt
i
best ms2: Prosit 2023 intensity timsTOF; best rt: AlphaPept rt generic



Appendix Figure B.7 Feature values for top PSMs for each model tested during the heuristic model search

The datasets covered here are Arabidopsis phosphoproteome (a), mouse phosphoproteome (b), HLA DDA class II from Marcu et al., 2021 (c), HLA DDA class I from Pak et al., 2021 (d), HLA DIA class I from Pak et al., 2021 (e) and from Ritz et al., 2017 (f), HLA DDA class I and II on timsTOF SCP (g-h), HeLa tryptic digest on timsTOF (i), Astral DIA from Serrano et al., 2024 (j) and Guzman et al. 2024 (k), and TMT11 LUAD (l).

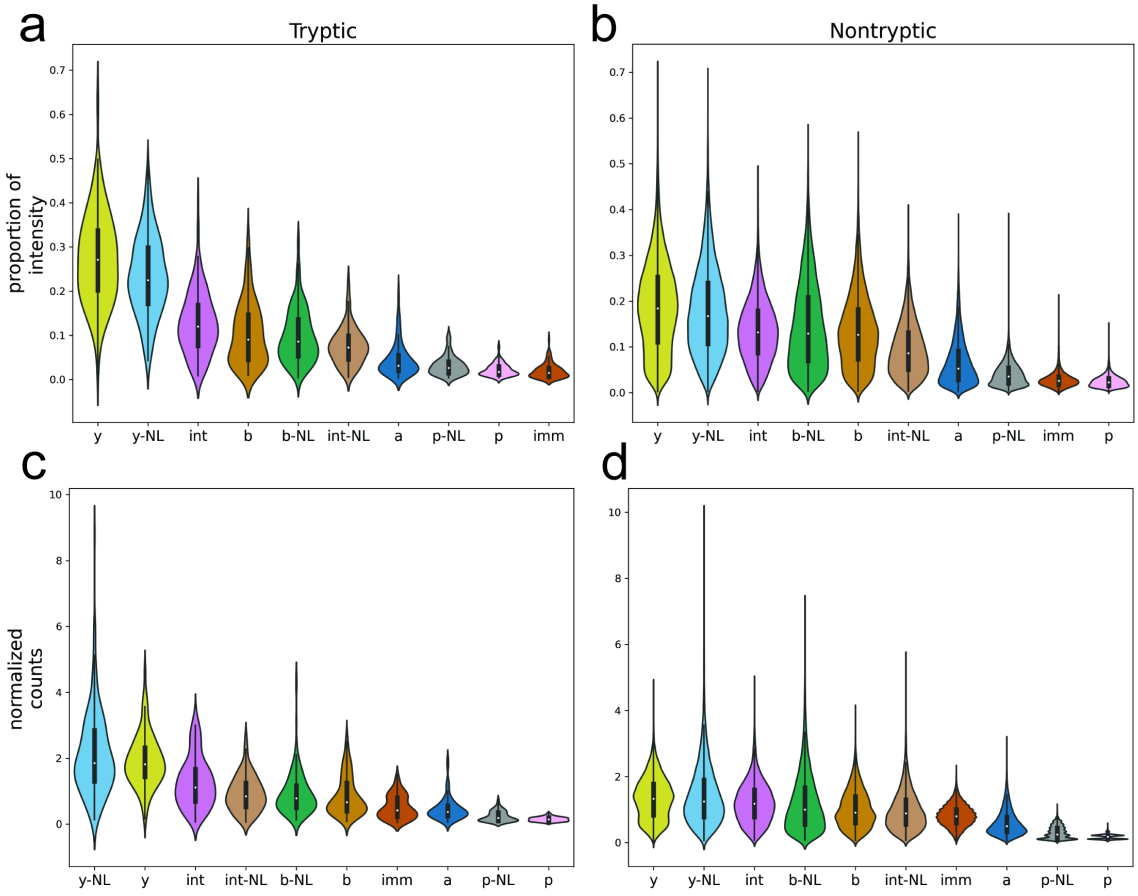

a
baseline
DIA-NN RT + DIA-NN MS2
DIA-NN RT + Prosit CID MS2
multiple models
79.3%
14500
15000
15500
16000
16500
17000
17500
b
baseline
DIA-NN RT + DIA-NN MS2
Prosit RT + Prosit HCD MS2
multiple models
64.1%
11000
11200
11400
11600
11800
c
baseline
DIA-NN RT + DIA-NN MS2
PeptDeep RT + Prosit timsTOF MS2
multiple models
67.8%
26000
28000
30000
32000
d
baseline
DIA-NN RT + DIA-NN MS2
Prosit RT + Prosit timsTOF MS2
multiple models
85.9%
30500
31000
31500
32000
32500
33000
33500
34000
e
baseline
DIA-NN RT + DIA-NN MS2
DIA-NN RT + PeptDeep MS2
multiple models
68.1%
42000
44000
46000
48000
50000
52000
54000
f
baseline
DIA-NN RT + DIA-NN MS2
Prosit RT + Prosit HCD MS2
multiple models
42.4%
13000
14000
15000
16000
17000
18000
19000
20000
number of peptides



Appendix Figure B.8 Peptides identified using multiple models

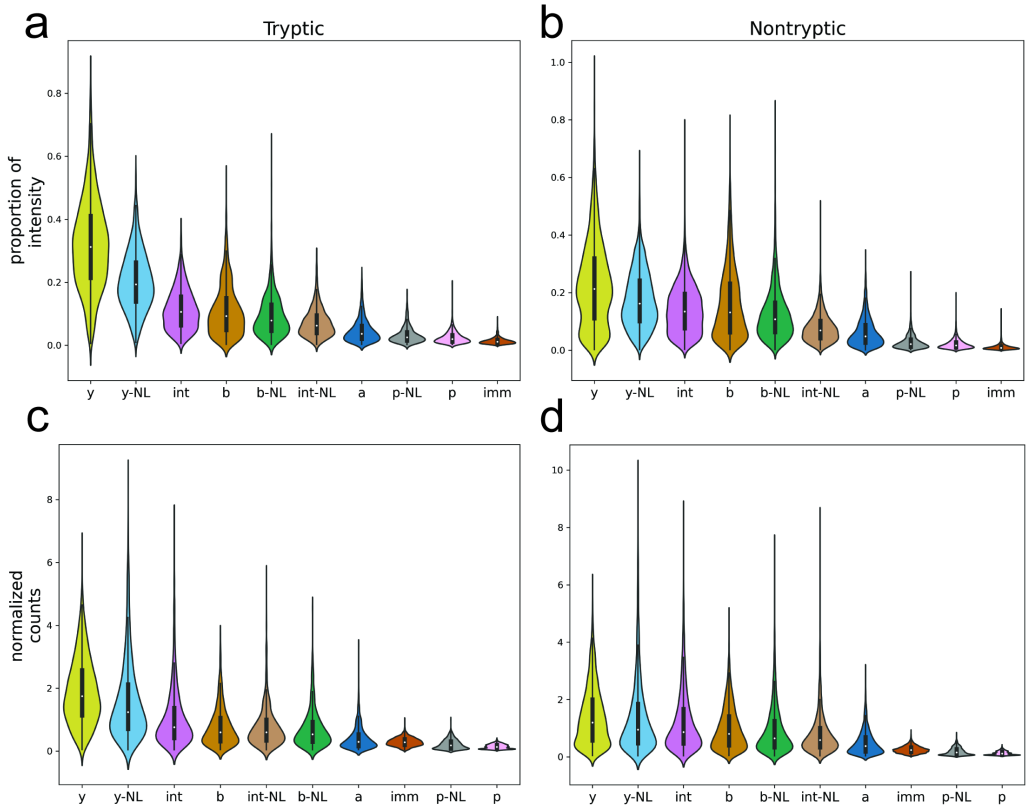
HLA DDA class I and II from Marcu et al., 2021 (a-b), HLA DDA class I and II on timsTOF SCP (c-d), HLA DDA and DIA class I from Pak et al., 2021 (e-f), HLA DIA class I from Ritz et al., 2017 (g), Arabidopsis phosphoproteome (h), TMT11 LUAD (i), and Astral DIA from Serrano et al., 2024 (j) and Guzman et al. 2024 (k). The percentage value represents the percent of peptides gained over DIA-NN by the best single models compared to the total percent gained by using multiple models. It was calculated from the median values of the test conditions: (“best single model combination” - “DIA-NN”) / (“multiple models” - “DIA-NN”). The percentage is not listed in cases where DIA-NN performs the best out of the single models.

Appendix C: Supplemental Material for Chapter 4



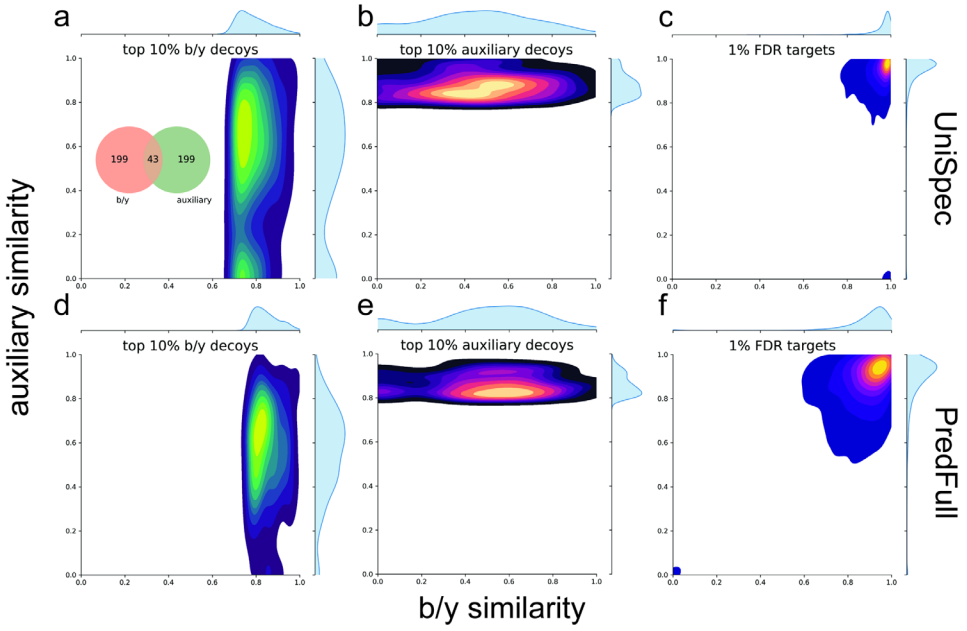
Appendix Figure C.1 Fragment ion annotation metrics in lung cancer data

One file belonging to a normal tissue sample (20180425_QE_HFX_LC2_HLAIp_CHC_SA_normC3N02671_10_R1.mzML[262]) was analyzed.

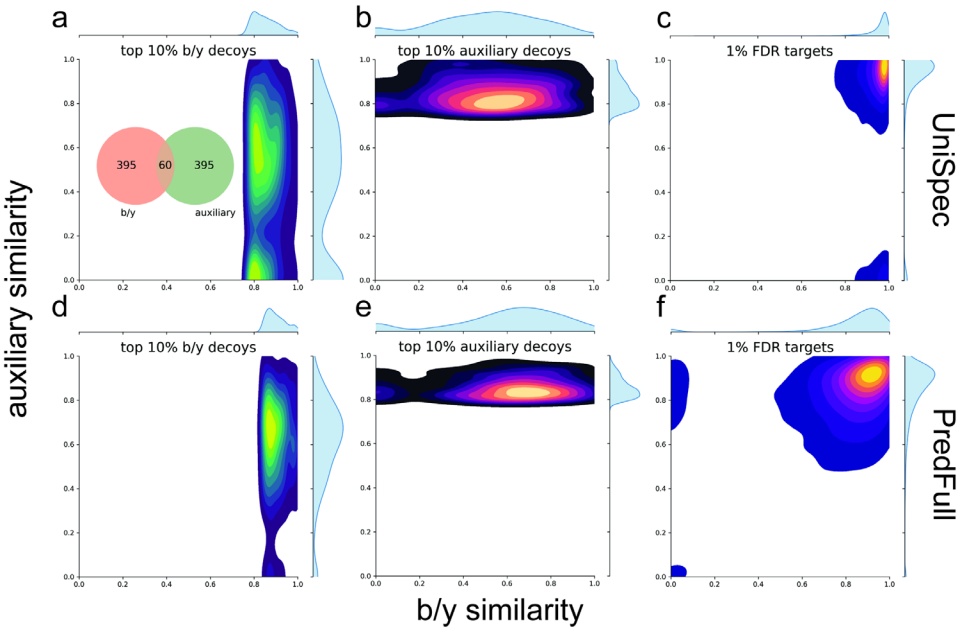


Appendix Figure C.2 Fragment ion annotation metrics in Atlas data

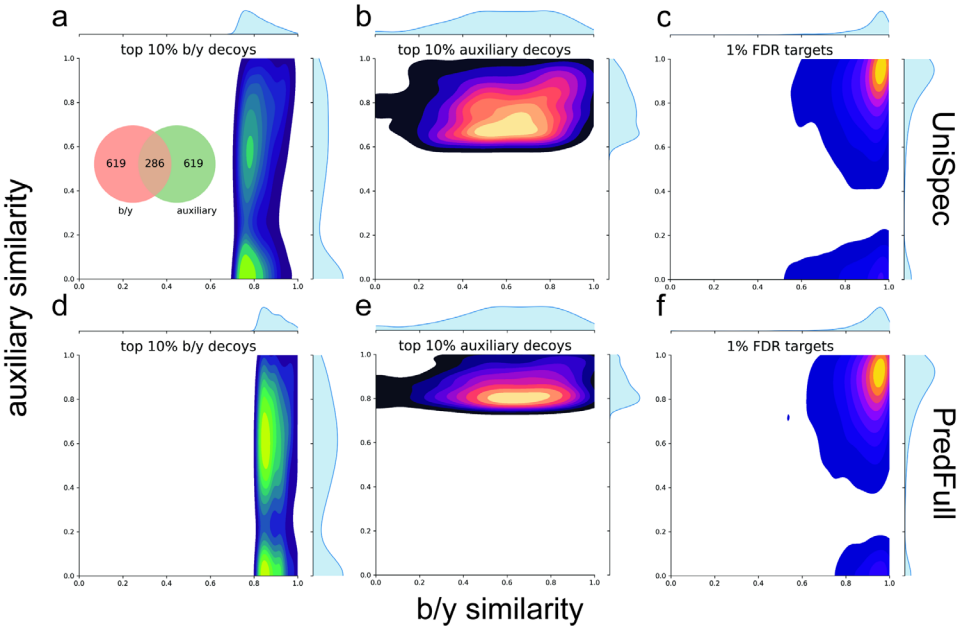
One file of the HLA class II data (QJWMX158AC_MS1_170627_AM_KK_GBMZH616_20_Tue39L243_DDA1_4001000mz_msms9.mzML[228]) was analyzed.



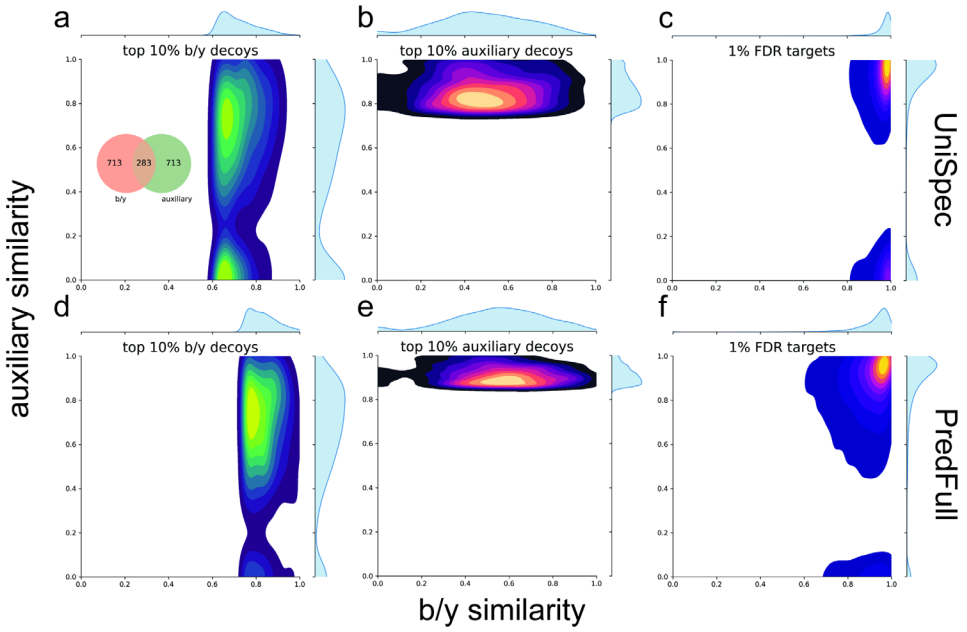
Appendix Figure C.3 The relationship between b/y and auxiliary ion similarity scores for class II data from Marcu et al., 2021



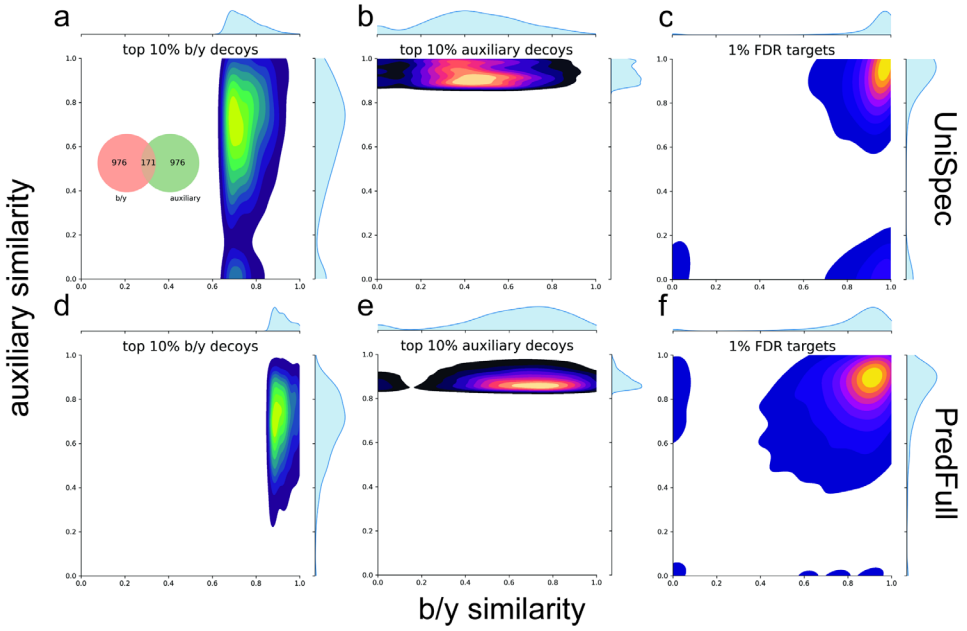
Appendix Figure C.4 The relationship between b/y and auxiliary ion similarity scores for class II data from Kraemer et al., 2023



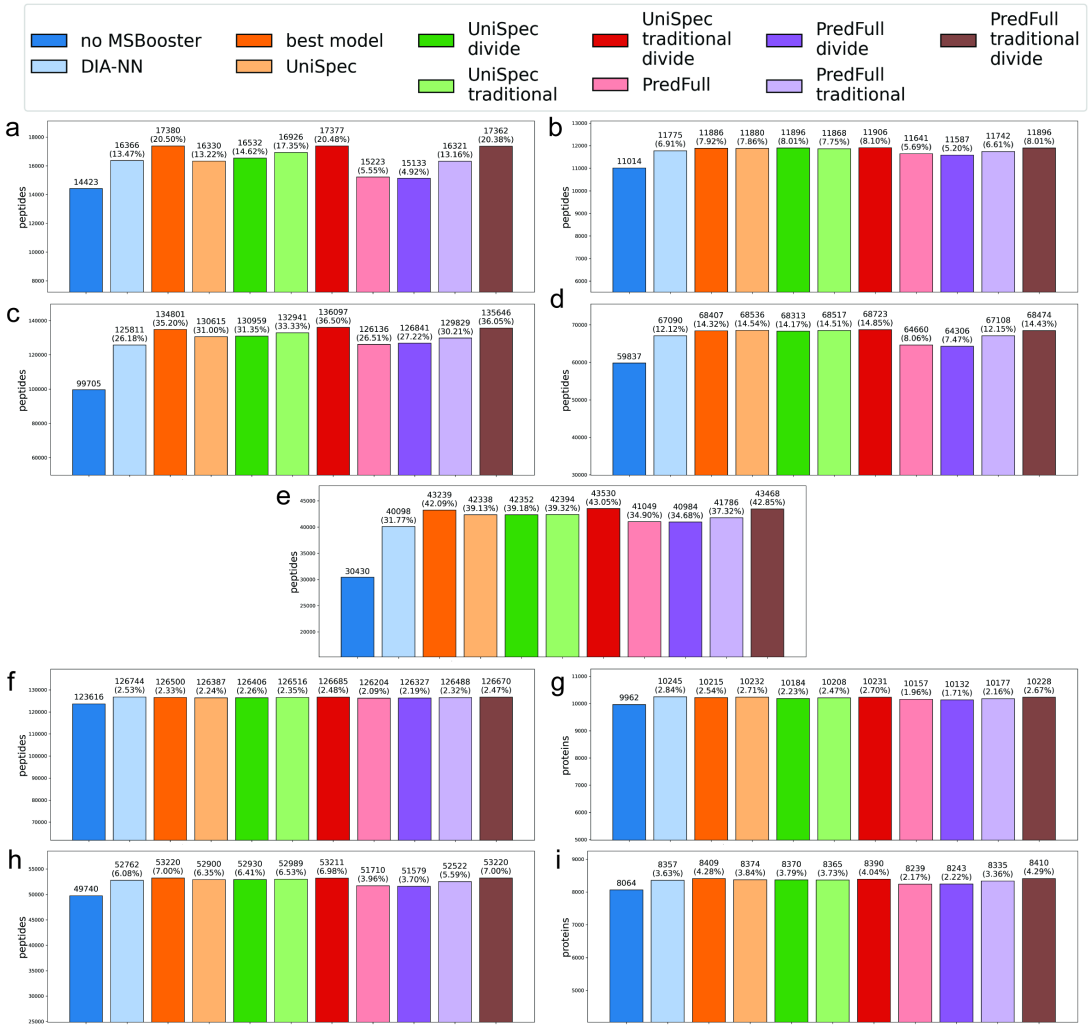
Appendix Figure C.5 The relationship between b/y and auxiliary ion similarity scores for class I DIA data from Pak et al., 2021



Appendix Figure C.6 The relationship between b/y and auxiliary ion similarity scores for tryptic data from Wang et al., 2019



Appendix Figure C.7 The relationship between b/y and auxiliary ion similarity scores for AspN data from Wang et al., 2019



Appendix Figure C.8 Identification numbers from full spectrum models (all other datasets)

IDs for each PSM rescoring mode are shown, with the exact number of peptides and percent gain over "no MSBooster" shown above each bar. The datasets analyzed are HLA class I and II from the Ligand Atlas (a-b), HLA class I and II from the lung cancer dataset (c-d), phospho HLA class I (e), tryptic peptides and proteins (f-g), and AspN peptides and proteins (h-i).

Appendix D: Relevant parameters in command line MSBooster

Below is a list of some of the parameters passed to command line MSBooster for the analyses presented in this work. A complete list of parameters is available at https://github.com/Nesvilab/MSBooster?tab=readme-ov-file#optional-parameters.

- useMultipleCorrelatedFeatures (Boolean): whether to use the multiple correlated features listed in Chapter 2.6.2. This feature is deprecated in the newest MSBooster. It is set to False by default in older versions.
- createPredFileOnly (Boolean): when set to True, MSBooster exits after creating input files for the specified spectra/auxiliary spectra/RT/IM models, without calling prediction models or calculating features. This is set to False by default. This is useful when a model is not supported by MSBooster or Koina, so an input file can be created for an external model to predict. Prediction files can be specified by the parameter spectraPredFile, etc.
- massesForLoessCalibration (String): a comma separated list of the PTM masses to be considered separately from other peptides during RT/IM calibration.
- spectraPredFile (String): path to a previously predicted spectra file. Passing this parameter will skip the prediction step. Similar parameters are auxSpectraPredFile, RTPredFile, and IMPredFile. Even if multiple properties are in the same file (e.g. DIA-NN predicted.spectra.bin file containing all predictions), each parameter must be specified if that property is desired.

- editedPinSuffix (String): the suffix to add onto edited pin files. By default, it is set to “edited.” For example, the original pin file A.pin will have scores added to a new file A_edited.pin.
- instrument (String): the type of mass spectrometer used. This information is used by AlphaPeptDeep and UniSpec as inputs for model predictions. MSBooster attempts to read instrument information from the mzML file, but sometimes the filter string is missing this information and must be specified manually.
- divideFragments (Integer): specifies the scheme by which to divide fragment ion types. The following numbers are accepted values:
    - 0: use all available ion types together
    - 1: ion types predicted by the main spectra model are used in one set of scores (Set 1). The other set of scores uses those ion types predicted by the auxiliary spectra model but not in Set 1
    - 2: b/y-ions are used in one set. All other ion types are used in the other set
    - 3: similar to option 2, but unknown ions are used separately in a third set of scores
    - 4: each detected ion type is used to calculate a separate set of scores (i.e. one set of scores for y-ions, one for b-ions, one for a-ions, etc.)
- ignoredFragmentIonTypes (String): comma separated list of fragment ion types to exclude from rescoring. They are still predicted, but filtered out before similarity scoring.

Bibliography

1 Roskoski, R. Nature's robots: A history of proteins: Tanford, C., Reynolds, J. *Biochemistry and Molecular Biology Education* **30**, 343-345 (2002). https://doi.org:10.1002/bmb.2002.494030059991

2 dos Santos, R., Carvalho, A. L. & Roque, A. C. A. Renaissance of protein crystallization and precipitation in biopharmaceuticals purification. *Biotechnology Advances* **35**, 41-50 (2017). https://doi.org:10.1016/j.biotechadv.2016.11.005

3 Blanco, G. & Mercer, R. W. Isozymes of the Na-K-ATPase: heterogeneity in structure, diversity in function. *American Journal of Physiology-Renal Physiology* **275**, F633-F650 (1998).

4 Kastenhuber, E. R. & Lowe, S. W. Putting p53 in Context. *Cell* **170**, 1062-1078 (2017). https://doi.org:10.1016/j.cell.2017.08.028

5 Wang, H. L., Guo, M., Wei, H. D. & Chen, Y. H. Targeting p53 pathways: mechanisms, structures, and advances in therapy. *Signal Transduction and Targeted Therapy* **8** (2023). https://doi.org:10.1038/s41392-023-01347-1

6 Salzberg, S. L. Open questions: How many genes do we have? *BMC Biology* **16** (2018). https://doi.org:10.1186/s12915-018-0564-x

7 Amaral, P. *et al.* The status of the human gene catalogue. *Nature* **622**, 41-47 (2023). https://doi.org:10.1038/s41586-023-06490-x

8 Liu, Q., Fang, L. & Wu, C. Alternative Splicing and Isoforms: From Mechanisms to Diseases. *Genes* **13**, 401 (2022). https://doi.org:10.3390/genes13030401

9 Carbonara, K., Andonovski, M. & Coorssen, J. R. Proteomes Are of Proteoforms: Embracing the Complexity. *Proteomes* **9**, 38 (2021). https://doi.org:10.3390/proteomes9030038

10 Aebersold, R. *et al.* How many human proteoforms are there? *Nature Chemical Biology* **14**, 206-214 (2018). https://doi.org:10.1038/nchembio.2576

11 Mudge, J. M. *et al.* Standardized annotation of translated open reading frames. *Nature Biotechnology* **40**, 994-999 (2022). https://doi.org:10.1038/s41587-022-01369-0

12 Ramazi, S. & Zahiri, J. Post-translational modifications in proteins: resources, tools and prediction methods. *Database* **2021** (2021). https://doi.org:10.1093/database/baab012

13 Yang, Z. *et al.* A mutation-induced drug resistance database (MdrDB). *Communications Chemistry* **6** (2023). https://doi.org:10.1038/s42004-023-00920-7

14 Santos, R. *et al.* A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery* **16**, 19-34 (2017). https://doi.org:10.1038/nrd.2016.230

15 Rock, K. L., Reits, E. & Neefjes, J. Present Yourself! By MHC Class I and MHC Class II Molecules. *Trends in Immunology* **37**, 724-737 (2016). https://doi.org:10.1016/j.it.2016.08.010
16 Admon, A. The biogenesis of the immunopeptidome. *Seminars in Immunology* **67** (2023). https://doi.org:10.1016/j.smim.2023.101766
17 Sánchez, A. & Vázquez, A. Bioactive peptides: A review. *Food Quality and Safety* **1**, 29-46 (2017). https://doi.org:10.1093/fqsafe/fyx006
18 Akbarian, M., Khani, A., Eghbalpour, S. & Uversky, V. N. Bioactive Peptides: Synthesis, Sources, Applications, and Proposed Mechanisms of Action. *International Journal of Molecular Sciences* **23**, 1445 (2022). https://doi.org:10.3390/ijms23031445
19 Kosti, I., Jain, N., Aran, D., Butte, A. J. & Sirota, M. Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues. *Scientific Reports* **6**, 24799 (2016). https://doi.org:10.1038/srep24799
20 Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics* **21**, 630-644 (2020). https://doi.org:10.1038/s41576-020-0258-4
21 Wang, D. *et al.* A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Molecular Systems Biology* **15**, e8503 (2019). https://doi.org:10.15252/msb.20188503
22 Sinha, A. & Mann, M. A beginner's guide to mass spectrometry–based proteomics. *The Biochemist* **42**, 64-69 (2020). https://doi.org:10.1042/bio20200057
23 Shuken, S. R. An Introduction to Mass Spectrometry-Based Proteomics. *Journal of Proteome Research* **22**, 2151-2171 (2023). https://doi.org:10.1021/acs.jproteome.2c00838
24 Ho, C. S. *et al.* Electrospray ionisation mass spectrometry: principles and clinical applications. *Clin Biochem Rev* **24**, 3-12 (2003).
25 Kaklamanos, G., Aprea, E. & Theodoridis, G. in *Chemical Analysis of Food (Second Edition)* (ed Yolanda Pico) 525-552 (Academic Press, 2020).
26 Scigelova, M. & Makarov, A. Orbitrap Mass Analyzer – Overview and Applications in Proteomics. *PROTEOMICS* **6**, 16-21 (2006). https://doi.org:10.1002/pmic.200600528
27 Stewart, H. *et al. Parallelized Acquisition of Orbitrap and Astral Analyzers Enables High-Throughput Quantitative Analysis* (Cold Spring Harbor Laboratory, 2023).
28 Heil, L. R. *et al.* Evaluating the Performance of the Astral Mass Analyzer for Quantitative Proteomics Using Data-Independent Acquisition. *Journal of Proteome Research* **22**, 3290-3300 (2023). https://doi.org:10.1021/acs.jproteome.3c00357
29 Kumar, S. in *Advances in Protein Molecular and Structural Biology Methods* (eds Timir Tripathi & Vikash Kumar Dubey) 327-356 (Academic Press, 2022).
30 Roepstorff, P. & Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom* **11**, 601 (1984). https://doi.org:10.1002/bms.1200111109

31 Johnson, R. S., Martin, S. A. & Biemann, K. Collision-induced fragmentation of (M + H)+ ions of peptides. Side chain specific sequence ions. *International Journal of Mass Spectrometry and Ion Processes* **86**, 137-154 (1988). https://doi.org:https://doi.org/10.1016/0168-1176(88)80060-0

32 Liu, X., Shan, B., Xin, L. & Ma, B. Better score function for peptide identification with ETD MS/MS spectra. *BMC Bioinformatics* **11**, S4 (2010). https://doi.org:10.1186/1471-2105-11-s1-s4

33 Wilburn, D. B., Richards, A. L., Swaney, D. L. & Searle, B. C. CIDer: A Statistical Framework for Interpreting Differences in CID and HCD Fragmentation. *J Proteome Res* **20**, 1951-1965 (2021). https://doi.org:10.1021/acs.jproteome.0c00964

34 Bereman, M. S. *et al.* Evaluation of front-end higher energy collision-induced dissociation on a benchtop dual-pressure linear ion trap mass spectrometer for shotgun proteomics. *Anal Chem* **84**, 1533-1539 (2012). https://doi.org:10.1021/ac203210a

35 Ichou, F. *et al.* Comparison of the activation time effects and the internal energy distributions for the CID, PQD and HCD excitation modes. *Journal of Mass Spectrometry* **49**, 498-508 (2014). https://doi.org:10.1002/jms.3365

36 Good, D. M., Wirtala, M., McAlister, G. C. & Coon, J. J. Performance Characteristics of Electron Transfer Dissociation Mass Spectrometry*. *Molecular & Cellular Proteomics* **6**, 1942-1951 (2007). https://doi.org:https://doi.org/10.1074/mcp.M700073-MCP200

37 Mechref, Y. Use of CID/ETD Mass Spectrometry to Analyze Glycopeptides. *Current Protocols in Protein Science* **68** (2012). https://doi.org:10.1002/0471140864.ps1211s68

38 Riley, N. M. & Coon, J. J. The Role of Electron Transfer Dissociation in Modern Proteomics. *Analytical Chemistry* **90**, 40-64 (2018). https://doi.org:10.1021/acs.analchem.7b04810

39 Kim, M. S. & Pandey, A. Electron transfer dissociation mass spectrometry in proteomics. *PROTEOMICS* **12**, 530-542 (2012). https://doi.org:10.1002/pmic.201100517

40 Frese, C. K. *et al.* Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. *Anal Chem* **84**, 9668-9673 (2012). https://doi.org:10.1021/ac3025366

41 Meyer, J. G. & Schilling, B. Clinical applications of quantitative proteomics using targeted and untargeted data-independent acquisition techniques. *Expert Review of Proteomics* **14**, 419-429 (2017). https://doi.org:10.1080/14789450.2017.1322904

42 Kreimer, S. *et al.* Advanced Precursor Ion Selection Algorithms for Increased Depth of Bottom-Up Proteomic Profiling. *Journal of Proteome Research* **15**, 3563-3573 (2016). https://doi.org:10.1021/acs.jproteome.6b00312

43 Adkins, J. N. *et al.* Toward a Human Blood Serum Proteome. *Molecular & Cellular Proteomics* **1**, 947-955 (2002). https://doi.org:10.1074/mcp.m200066-mcp200

44 Holewinski, R. J., Jin, Z., Powell, M. J., Maust, M. D. & Van Eyk, J. E. A fast and reproducible method for albumin isolation and depletion from serum and cerebrospinal fluid. *PROTEOMICS* **13**, 743-750 (2013). https://doi.org:10.1002/pmic.201200192

45 Webb-Robertson, B. J. M. *et al.* Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics. *Journal of Proteome Research* **14**, 1993-2001 (2015). https://doi.org:10.1021/pr501138h

46 Gillet, L. C. *et al.* Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular & Cellular Proteomics* **11** (2012). https://doi.org:10.1074/mcp.O111.016717

47 Barkovits, K. *et al.* Reproducibility, Specificity and Accuracy of Relative Quantification Using Spectral Library-based Data-independent Acquisition. *Molecular & Cellular Proteomics* **19**, 181-197 (2020). https://doi.org:10.1074/mcp.ra119.001714

48 Houel, S. *et al.* Quantifying the Impact of Chimera MS/MS Spectra on Peptide Identification in Large-Scale Proteomics Studies. *Journal of Proteome Research* **9**, 4152-4160 (2010). https://doi.org:10.1021/pr1003856

49 Frejno, M. *et al. Unifying the analysis of bottom-up proteomics data with CHIMERYS* (Cold Spring Harbor Laboratory, 2024).

50 Guzman, U. H. *et al.* Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition. *Nature Biotechnology* (2024). https://doi.org:10.1038/s41587-023-02099-7

51 Matzinger, M. *et al.* Micropillar arrays, wide window acquisition and AI-based data analysis improve comprehensiveness in multiple proteomic applications. *Nature Communications* **15** (2024). https://doi.org:10.1038/s41467-024-45391-z

52 Truong, T. *et al.* Data-Dependent Acquisition with Precursor Coisolation Improves Proteome Coverage and Measurement Throughput for Label-Free Single-Cell Proteomics**. *Angewandte Chemie International Edition* **62** (2023). https://doi.org:10.1002/anie.202303415

53 Li, J. *et al.* TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nature Methods* **17**, 399-404 (2020). https://doi.org:10.1038/s41592-020-0781-4

54 Zecha, J. *et al.* TMT Labeling for the Masses: A Robust and Cost-efficient, In-solution Labeling Approach. *Molecular & Cellular Proteomics* **18**, 1468-1478 (2019). https://doi.org:10.1074/mcp.TIR119.001385

55 Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature Methods* **8**, 937-940 (2011). https://doi.org:10.1038/nmeth.1714

56 Zuniga, N. R. *et al.* Achieving a 35-Plex Tandem Mass Tag Reagent Set through Deuterium Incorporation. *Journal of Proteome Research* **23**, 5153-5165 (2024). https://doi.org:10.1021/acs.jproteome.4c00668

57 Lee, P. Y., Saraygord-Afshari, N. & Low, T. Y. The evolution of two-dimensional gel electrophoresis - from proteomics to emerging alternative

applications. *Journal of Chromatography A* **1615** (2020). https://doi.org:10.1016/j.chroma.2019.460763
58 Magdeldin, S. *et al.* Basics and recent advances of two dimensional-polyacrylamide gel electrophoresis. *Clinical Proteomics* **11**, 16 (2014). https://doi.org:10.1186/1559-0275-11-16
59 Wittmann-Liebold, B., Graack, H. R. & Pohl, T. Two-dimensional gel electrophoresis as tool for proteomics studies in combination with protein identification by mass spectrometry. *PROTEOMICS* **6**, 4688-4703 (2006). https://doi.org:10.1002/pmic.200500874
60 Lenčo, J. *et al.* Reversed-Phase Liquid Chromatography of Peptides for Bottom-Up Proteomics: A Tutorial. *Journal of Proteome Research* **21**, 2846-2892 (2022). https://doi.org:10.1021/acs.jproteome.2c00407
61 Bekker-Jensen, D. B. *et al.* An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Systems* **4**, 587-+ (2017). https://doi.org:10.1016/j.cels.2017.05.009
62 Messner, C. B. *et al.* Ultra-High-Throughput Clinical Proteomics Reveals Classifiers of COVID-19 Infection. *Cell Systems* **11**, 11-+ (2020). https://doi.org:10.1016/j.cels.2020.05.012
63 Ridgeway, M., Park, M., Meier, J. & Silveira, J. Trapped ion mobility spectrometry. *Abstracts of Papers of the American Chemical Society* **249** (2015).
64 Zhou, Z. *et al.* Ion mobility collision cross-section atlas for known and unknown metabolite annotation in untargeted metabolomics. *Nature Communications* **11** (2020). https://doi.org:10.1038/s41467-020-18171-8
65 Meier, F. *et al.* Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer *. *Molecular & Cellular Proteomics* **17**, 2534-2545 (2018). https://doi.org:10.1074/MCP.TIR118.000900
66 Melby, J. A. *et al.* Novel Strategies to Address the Challenges in Top-Down Proteomics. *Journal of the American Society for Mass Spectrometry* **32**, 1278-1294 (2021). https://doi.org:10.1021/jasms.1c00099
67 Steen, H. & Mann, M. The abc's (and xyz's) of peptide sequencing. *Nature Reviews Molecular Cell Biology* **5**, 699-711 (2004). https://doi.org:10.1038/nrm1468
68 Nesvizhskii, A. I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods* **4**, 787-797 (2007). https://doi.org:10.1038/nmeth1088
69 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. *Nature Methods* **14**, 513-520 (2017). https://doi.org:10.1038/nmeth.4256
70 Cox, J. *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research* **10**, 1794-1805 (2011). https://doi.org:10.1021/PR101065J/SUPPL_FILE/PR101065J_SI_002.ZIP
71 Craig, R. & Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466-1467 (2004). https://doi.org:10.1093/bioinformatics/bth092

72 Eng, J. K., Mccormack, A. L. & Yates, J. R. 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). https://doi.org:10.1016/1044-0305(94)80016-2

73 Perkins, D. N., Pappin, D. J. C., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551-3567 (1999). https://doi.org:10.1002/(sici)1522-2683(19991201)20:18<3551::aid-elps3551>3.0.co;2-2

74 Lazear, M. R. Sage: An Open-Source Tool for Fast Proteomics Searching and Quantification at Scale. *J Proteome Res* **22**, 3652-3659 (2023). https://doi.org:10.1021/acs.jproteome.3c00486

75 Griss, J. Spectral library searching in proteomics. *Proteomics* **16**, 729-740 (2016). https://doi.org:10.1002/pmic.201500296

76 Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655-667 (2007). https://doi.org:10.1002/pmic.200600625

77 Deutsch, E. W. *et al.* Expanding the Use of Spectral Libraries in Proteomics. *Journal of Proteome Research* **17**, 4051-4060 (2018). https://doi.org:10.1021/acs.jproteome.8b00485

78 Hu, A., Noble, W. S. & Wolf-Yadlin, A. Technical advances in proteomics: new developments in data-independent acquisition. *F1000Res* **5** (2016). https://doi.org:10.12688/f1000research.7042.1

79 Heil, L. R. *et al.* Building Spectral Libraries from Narrow-Window Data-Independent Acquisition Mass Spectrometry Data. *J Proteome Res* **21**, 1382-1391 (2022). https://doi.org:10.1021/acs.jproteome.1c00895

80 Demichev, V. *et al.* dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts. *Nature Communications* **13** (2022). https://doi.org:10.1038/s41467-022-31492-0

81 Pathan, M., Samuel, M., Keerthikumar, S. & Mathivanan, S. in *Proteome Bioinformatics* Vol. 1549 *Methods in Molecular Biology* (eds S. Keerthikumar & S. Mathivanan) 67-74 (2017).

82 Noble, W. S. & Maccoss, M. J. Computational and Statistical Analysis of Protein Mass Spectrometry Data. *PLoS Computational Biology* **8**, e1002296 (2012). https://doi.org:10.1371/journal.pcbi.1002296

83 Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4**, 207-214 (2007). https://doi.org:10.1038/nmeth1019

84 Freestone, J., Short, T., Noble, W. S. & Keich, U. Group-walk: a rigorous approach to group-wise false discovery rate analysis by target-decoy competition. *Bioinformatics* **38**, ii82-ii88 (2022). https://doi.org:10.1093/bioinformatics/btac471

85 Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **74**, 5383-5392 (2002). https://doi.org:10.1021/ac025747h

86 Ma, K., Vitek, O. & Nesvizhskii, A. I. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinformatics* **13**, S1 (2012). https://doi.org:10.1186/1471-2105-13-s16-s1

87 Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & Maccoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods* **4**, 923-925 (2007). https://doi.org:10.1038/nmeth1113

88 Freestone, J., Käll, L., Noble, W. S. & Keich, U. *How to train a post-processor for tandem mass spectrometry proteomics database search while maintaining control of the false discovery rate* (Cold Spring Harbor Laboratory, 2023).

89 Granholm, V., Noble, W. S. & Käll, L. 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**, 2671-2678 (2011). https://doi.org:10.1021/pr1012619

90 Granholm, V., Navarro, J. F., Noble, W. S. & Käll, L. Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics. *Journal of Proteomics* **80**, 123-131 (2013). https://doi.org:10.1016/j.jprot.2012.12.007

91 Feng, X.-D. *et al.* Using the entrapment sequence method as a standard to evaluate key steps of proteomics data analysis process. *BMC Genomics* **18** (2017). https://doi.org:10.1186/s12864-017-3491-2

92 Wen, B. *et al. Assessment of false discovery rate control in tandem mass spectrometry analysis using entrapment* (Cold Spring Harbor Laboratory, 2024).

93 da Veiga Leprevost, F. *et al.* Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nature Methods 2020 17:9* **17**, 869-870 (2020). https://doi.org:10.1038/S41592-020-0912-Y

94 Potel, C. M., Lemeer, S. & Heck, A. J. R. Phosphopeptide Fragmentation and Site Localization by Mass Spectrometry: An Update. *Analytical Chemistry* **91**, 126-141 (2019). https://doi.org:10.1021/acs.analchem.8b04746

95 Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical chemistry* **75**, 4646-4658 (2003). https://doi.org:10.1021/AC0341261

96 Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018 (2016). https://doi.org:10.1038/sdata.2016.18

97 Jones, A. R., Deutsch, E. W. & Vizcaíno, J. A. Is DIA proteomics data FAIR? Current data sharing practices, available bioinformatics infrastructure and recommendations for the future. *PROTEOMICS* **23**, 2200014 (2023). https://doi.org:10.1002/pmic.202200014

98 Klein, M. *et al.* Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLoS ONE* **9**, e115253 (2014). https://doi.org:10.1371/journal.pone.0115253

99 Roche, D. G., Kruuk, L. E. B., Lanfear, R. & Binning, S. A. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLOS Biology* **13**, e1002295 (2015). https://doi.org:10.1371/journal.pbio.1002295

100 Federer, L. M. Long-term availability of data associated with articles in PLOS ONE. *PLOS ONE* **17**, e0272845 (2022). https://doi.org:10.1371/journal.pone.0272845
101 Briney, K. A. Measuring data rot: An analysis of the continued availability of shared data from a Single University. *PLOS ONE* **19**, e0304781 (2024). https://doi.org:10.1371/journal.pone.0304781
102 Martens, L. *et al.* PRIDE: The proteomics identifications database. *PROTEOMICS* **5**, 3537-3545 (2005). https://doi.org:10.1002/pmic.200401303
103 Okuda, S. *et al.* jPOSTrepo: an international standard data repository for proteomes. *Nucleic Acids Research* **45**, D1107-D1111 (2017). https://doi.org:10.1093/nar/gkw1080
104 Vizcaíno, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnology* **32**, 223-226 (2014). https://doi.org:10.1038/nbt.2839
105 Deutsch, E. W. *et al.* The ProteomeXchange consortium at 10 years: 2023 update. *Nucleic Acids Research* **51**, D1539-D1548 (2023). https://doi.org:10.1093/nar/gkac1040
106 Desiere, F. *et al.* Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biology* **6**, R9 (2004). https://doi.org:10.1186/gb-2004-6-1-r9
107 Omenn, G. S. *et al.* The 2023 Report on the Proteome from the HUPO Human Proteome Project. *Journal of Proteome Research* **23**, 532-549 (2024). https://doi.org:10.1021/acs.jproteome.3c00591
108 Choi, M. *et al.* MassIVE.quant: a community resource of quantitative mass spectrometry–based proteomics datasets. *Nature Methods* **17**, 981-984 (2020). https://doi.org:10.1038/s41592-020-0955-0
109 Deutsch, E. W. *et al.* Proteomics Standards Initiative at Twenty Years: Current Activities and Future Work. *Journal of Proteome Research* (2023). https://doi.org:10.1021/acs.jproteome.2c00637
110 Dai, C. *et al.* A proteomics sample metadata representation for multiomics integration and big data analysis. *Nature Communications* **12** (2021). https://doi.org:10.1038/s41467-021-26111-3
111 Claeys, T. *et al.* lesSDRF is more: maximizing the value of proteomics data through streamlined metadata annotation. *Nature Communications* **14** (2023). https://doi.org:10.1038/s41467-023-42543-5
112 LeDuc, R. D. *et al.* Proteomics Standards Initiative's ProForma 2.0: Unifying the Encoding of Proteoforms and Peptidoforms. *Journal of Proteome Research* **21**, 1189-1195 (2022). https://doi.org:10.1021/acs.jproteome.1c00771
113 Deutsch, E. W. *et al.* Universal Spectrum Identifier for mass spectra. *Nature Methods* **18**, 768-770 (2021). https://doi.org:10.1038/s41592-021-01184-6
114 Bateman, A. *et al.* UniProt: a hub for protein information. *Nucleic Acids Research* **43**, D204-D212 (2015). https://doi.org:10.1093/nar/gku989
115 Bateman, A. *et al.* UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **47**, D506-D515 (2019). https://doi.org:10.1093/nar/gky1049

116 Bateman, A. *et al.* UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51**, D523-D531 (2023). https://doi.org:10.1093/nar/gkac1052
117 Samaras, P. *et al.* ProteomicsDB: a multi-omics and multi-organism resource for life science research. *Nucleic Acids Research* (2019). https://doi.org:10.1093/nar/gkz974
118 Lautenbacher, L. *et al.* ProteomicsDB: toward a FAIR open-source resource for life-science research. *Nucleic Acids Research* **50**, D1541-D1552 (2022). https://doi.org:10.1093/nar/gkab1026
119 Navarro, P. *et al.* A multicenter study benchmarks software tools for label-free proteome quantification. *Nature Biotechnology* **34**, 1130-1136 (2016). https://doi.org:10.1038/nbt.3685
120 Ramus, C. *et al.* Benchmarking quantitative label-free LC–MS data processing workflows using a complex spiked proteomic standard dataset. *Journal of Proteomics* **132**, 51-62 (2016). https://doi.org:10.1016/j.jprot.2015.11.011
121 Fröhlich, K. *et al.* Benchmarking of analysis strategies for data-independent acquisition proteomics using a large-scale dataset comprising inter-patient heterogeneity. *Nature Communications* **13** (2022). https://doi.org:10.1038/s41467-022-30094-0
122 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021). https://doi.org:10.1038/s41586-021-03819-2
123 Wen, B. *et al.* Deep Learning in Proteomics. *PROTEOMICS* **20**, 1900335 (2020). https://doi.org:10.1002/pmic.201900335
124 Elabd, H. *et al.* Amino acid encoding for deep learning applications. *BMC Bioinformatics* **21** (2020). https://doi.org:10.1186/s12859-020-03546-x
125 Li, S. J., Arnold, R. J., Tang, H. X. & Radivojac, P. On the Accuracy and Limits of Peptide Fragmentation Spectrum Prediction. *Analytical Chemistry* **83**, 790-796 (2011). https://doi.org:10.1021/ac102272r
126 Bittremieux, W., May, D. H., Bilmes, J. & Noble, W. S. A learned embedding for efficient joint analysis of millions of mass spectra. *Nature Methods* **19**, 675-678 (2022). https://doi.org:10.1038/s41592-022-01496-1
127 Wu, L., Hoque, A. & Lam, H. Spectroscape enables real-time query and visualization of a spectral archive in proteomics. *Nature Communications* **14** (2023). https://doi.org:10.1038/s41467-023-42006-x
128 Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Research* **48**, W449-W454 (2020). https://doi.org:10.1093/NAR/GKAA379
129 Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. in *Numerical Computer Methods, Pt D* Vol. 383 *Methods in Enzymology* (eds L. Brand & M. L. Johnson) 66-+ (2004).
130 Wysocki, V. H., Tsaprailis, G., Smith, L. L. & Breci, L. A. Mobile and localized protons: a framework for understanding peptide dissociation. *Journal of Mass Spectrometry* **35**, 1399-1406 (2000). https://doi.org:10.1002/1096-9888(200012)35:12<1399::aid-jms86>3.0.co;2-r

131 Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Analytical chemistry* **76**, 3908-3922 (2004). https://doi.org:10.1021/ac049951b
132 Arnold, R. J., Jayasankar, N., Aggarwal, D., Tang, H. & Radivojac, P. A machine learning approach to predicting peptide fragmentation spectra. *Pac Symp Biocomput*, 219-230 (2006).
133 Degroeve, S., Martens, L. & Jurisica, I. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics* **29**, 3199-3199 (2013). https://doi.org:10.1093/BIOINFORMATICS/BTT544
134 Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods* **16**, 509-518 (2019). https://doi.org:10.1038/s41592-019-0426-7
135 Zeng, W.-F. *et al.* AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics. *Nature Communications* **13** (2022). https://doi.org:10.1038/s41467-022-34904-3
136 Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L. & Degroeve, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nature Methods 2021 18:11* **18**, 1363-1369 (2021). https://doi.org:10.1038/s41592-021-01301-5
137 Wilburn, D. B. *et al.* Deep learning from harmonized peptide libraries enables retention time prediction of diverse post translational modifications. *bioRxiv* (2023). https://doi.org:10.1101/2023.05.30.542978
138 Devreese, R. *et al.* Collisional cross-section prediction for multiconformational peptide ions with IM2Deep. *bioRxiv*, 2025.2002.2018.638865 (2025). https://doi.org:10.1101/2025.02.18.638865
139 Serrano, G., Guruceaga, E. & Segura, V. DeepMSPeptide: peptide detectability prediction using deep learning. *Bioinformatics* **36**, 1279-1280 (2020). https://doi.org:10.1093/BIOINFORMATICS/BTZ708
140 Abdul-Khalek, N. *et al.* To fly, or not to fly, that is the question: A deep learning model for peptide detectability prediction in mass spectrometry. *bioRxiv*, 2024.2010.2028.620610 (2024). https://doi.org:10.1101/2024.10.28.620610
141 Zolg, D. P. *et al.* Building ProteomeTools based on a complete synthetic human proteome. *Nature Methods 2017 14:3* **14**, 259-262 (2017). https://doi.org:10.1038/nmeth.4153
142 He, C., Wang, X., Lyu, M. & Bian, X. Research on the Upper Limit of Accuracy for Predicting Theoretical Tandem Mass Spectrometry. *Journal of Computer and Communications* **12**, 184-195 (2024). https://doi.org:10.4236/jcc.2024.123011
143 Castaño, J. D. & Beaudry, F. Comparative Analysis of Data-Driven Rescoring Platforms for Improved Peptide Identification in HeLa Digest Samples. *Proteomics* **25**, e202400225 (2025). https://doi.org:10.1002/pmic.202400225
144 Flender, D. *et al.* Exploring the dynamic landscape of immunopeptidomics: Unravelling posttranslational modifications and navigating bioinformatics terrain. *Mass Spectrom Rev* (2024). https://doi.org:10.1002/mas.21905
145 Kalhor, M., Lapin, J., Picciani, M. & Wilhelm, M. Rescoring Peptide Spectrum Matches: Boosting Proteomics Performance by Integrating Peptide Property

Predictors Into Peptide Identification. *Mol Cell Proteomics* **23**, 100798 (2024). https://doi.org:10.1016/j.mcpro.2024.100798
146 Declercq, A. *et al.* MS2Rescore: Data-Driven Rescoring Dramatically Boosts Immunopeptide Identification Rates. *Molecular & cellular proteomics* **21** (2022). https://doi.org:10.1016/j.mcpro.2022.100266
147 Picciani, M. *et al.* Oktoberfest: Open-source spectral library generation and rescoring pipeline based on Prosit. *PROTEOMICS* (2023). https://doi.org:10.1002/pmic.202300112
148 Andreatta, M. *et al.* MS-Rescue: A Computational Pipeline to Increase the Quality and Yield of Immunopeptidomics Experiments. *Proteomics* **19**, e1800357 (2019). https://doi.org:10.1002/pmic.201800357
149 Bichmann, L. *et al.* MHCquant: Automated and Reproducible Data Analysis for Immunopeptidomics. *J Proteome Res* **18**, 3876-3884 (2019). https://doi.org:10.1021/acs.jproteome.9b00313
150 Li, K., Jain, A., Malovannaya, A., Wen, B. & Zhang, B. DeepRescore: Leveraging Deep Learning to Improve Peptide Identification in Immunopeptidomics. *PROTEOMICS* **20**, 1900334-1900334 (2020). https://doi.org:10.1002/PMIC.201900334
151 Cormican, J. A., Horokhovskyi, Y., Soh, W. T., Mishto, M. & Liepe, J. inSPIRE: An Open-Source Tool for Increased Mass Spectrometry Identification Rates Using Prosit Spectral Prediction. *Mol Cell Proteomics* **21**, 100432 (2022). https://doi.org:10.1016/j.mcpro.2022.100432
152 Strauss, M. T. *et al.* AlphaPept: a modern and open framework for MS-based proteomics. *Nat Commun* **15**, 2168 (2024). https://doi.org:10.1038/s41467-024-46485-4
153 Zolg, D. P. *et al.* INFERYS Rescoring: boosting peptide identifications and scoring confidence of database search results. *Rapid Communications in Mass Spectrometry* (2021). https://doi.org:10.1002/rcm.9128
154 Dorfer, V., Maltsev, S., Winkler, S. & Mechtler, K. CharmeRT: Boosting Peptide Identifications by Chimeric Spectra Identification and Retention Time Prediction. *J Proteome Res* **17**, 2581-2589 (2018). https://doi.org:10.1021/acs.jproteome.7b00836
155 Frejno, M. *et al.* (American Society for Mass Spectrometry 2021 Conference, 2021).
156 Yang, K. L. *et al.* MSBooster: improving peptide identification rates using deep learning-based features. *Nature Communications* **14** (2023). https://doi.org:10.1038/s41467-023-40129-9
157 Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics* **73**, 2092-2123 (2010). https://doi.org:10.1016/j.jprot.2010.08.009
158 The, M., MacCoss, M. J., Noble, W. S. & Käll, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *Journal of the American Society for Mass Spectrometry* **27**, 1719-1727 (2016). https://doi.org:10.1007/s13361-016-1460-7

159 Griss, J., Foster, J. M., Hermjakob, H. & Vizcaíno, J. A. PRIDE Cluster: building a consensus of proteomics data. *Nature Methods* **10**, 95-96 (2013). https://doi.org:10.1038/nmeth.2343
160 Wang, M. *et al.* Assembling the Community-Scale Discoverable Human Proteome. *Cell systems* **7**, 412-421.e415 (2018). https://doi.org:10.1016/J.CELS.2018.08.004
161 Craig, R., Cortens, J. C., Fenyo, D. & Beavis, R. C. Using annotated peptide mass spectrum libraries for protein identification. *J Proteome Res* **5**, 1843-1849 (2006). https://doi.org:10.1021/pr0602085
162 Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S. & MacCoss, M. J. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Analytical chemistry* **78**, 5678-5684 (2006). https://doi.org:10.1021/ac060279n
163 Sun, S. *et al.* Improved validation of peptide MS/MS assignments using spectral intensity prediction. *Molecular & cellular proteomics : MCP* **6**, 1-17 (2007). https://doi.org:10.1074/mcp.M600320-MCP200
164 Yen, C. Y. *et al.* A simulated MS/MS library for spectrum-to-spectrum searching in large scale identification of proteins. *Molecular & cellular proteomics : MCP* **8**, 857-869 (2009). https://doi.org:10.1074/mcp.M800384-MCP200
165 Elias, J. E., Gibbons, F. D., King, O. D., Roth, F. P. & Gygi, S. P. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nature Biotechnology* **22**, 214-219 (2004). https://doi.org:10.1038/nbt930
166 Norbeck, A. D. *et al.* The utility of accurate mass and LC elution time information in the analysis of complex proteomes. *Journal of the American Society for Mass Spectrometry* **16**, 1239-1249 (2005). https://doi.org:10.1016/j.jasms.2005.05.009
167 Krokhin, O. Peptide retention prediction in reversed-phase chromatography: proteomic applications. *Expert review of proteomics* **9**, 1-4 (2012). https://doi.org:10.1586/epr.11.79
168 Palmblad, M. Retention time prediction and protein identification. *Methods in molecular biology (Clifton, N.J.)* **367**, 195-207 (2007). https://doi.org:10.1385/1-59745-275-0:195
169 Moruz, L., Tomazela, D. & Käll, L. Training, selection, and robust calibration of retention time models for targeted proteomics. *J Proteome Res* **9**, 5209-5216 (2010). https://doi.org:10.1021/pr1005058
170 Malmström, J. *et al.* Optimized peptide separation and identification for mass spectrometry based proteomics via free-flow electrophoresis. *J Proteome Res* **5**, 2241-2249 (2006). https://doi.org:10.1021/pr0600632
171 Meier, F. *et al.* Deep learning the collisional cross sections of the peptide universe from a million experimental values. *Nature Communications 2021 12:1* **12**, 1-12 (2021). https://doi.org:10.1038/s41467-021-21352-8
172 Tarn, C. & Zeng, W.-F. pDeep3: Toward More Accurate Spectrum Prediction with Fast Few-Shot Learning. *Analytical Chemistry* **93**, 5815-5822 (2021). https://doi.org:10.1021/acs.analchem.0c05427
173 Zeng, W.-F. *et al. AlphaPeptDeep: A modular deep learning framework to predict peptide properties for proteomics* (Cold Spring Harbor Laboratory, 2022).

174 Zeng, W.-F. *et al.* MS/MS Spectrum Prediction for Modified Peptides Using pDeep2 Trained by Transfer Learning. *Analytical Chemistry* **91**, 9724-9731 (2019). https://doi.org:10.1021/acs.analchem.9b01262
175 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. *Nature Methods* **17**, 41-44 (2020). https://doi.org:10.1038/s41592-019-0638-x
176 Demichev, V. *et al.* dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts. *Nature communications* **13**, 3944 (2022). https://doi.org:10.1038/s41467-022-31492-0
177 Yang, Y. *et al.* In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nat Commun* **11**, 146 (2020). https://doi.org:10.1038/s41467-019-13866-z
178 Lamanna, J. *et al.* Digital microfluidic isolation of single cells for -Omics. *Nature Communications* **11** (2020). https://doi.org:10.1038/s41467-020-19394-5
179 Wilhelm, M. *et al.* Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nature Communications 2021 12:1* **12**, 1-12 (2021). https://doi.org:10.1038/s41467-021-23713-9
180 Pak, H. *et al.* Sensitive Immunopeptidomics by Leveraging Available Large-Scale Multi-HLA Spectral Libraries, Data-Independent Acquisition, and MS/MS Prediction. *Mol Cell Proteomics* **20**, 100080 (2021). https://doi.org:10.1016/j.mcpro.2021.100080
181 Polasky, D. A., Yu, F., Teo, G. C. & Nesvizhskii, A. I. Fast and comprehensive N- and O-glycoproteomics analysis with MSFragger-Glyco. *Nat Methods* **17**, 1125-1132 (2020). https://doi.org:10.1038/s41592-020-0967-9
182 Yu, F. *et al.* Identification of modified peptides using localization-aware open search. *Nat Commun* **11**, 4065 (2020).
183 Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **75**, 4646-4658 (2003). https://doi.org:10.1021/ac0341261
184 Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature Methods* **12**, 258-264 (2015). https://doi.org:10.1038/nmeth.3255
185 Yu, F. *et al. One-stop analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform* (Cold Spring Harbor Laboratory, 2022).
186 Li, K., Vaudel, M., Zhang, B., Ren, Y. & Wen, B. PDV: an integrative proteomics data viewer. *Bioinformatics* **35**, 1249-1251 (2019). https://doi.org:10.1093/bioinformatics/bty770
187 Klaeger, S. *et al.* Optimized Liquid and Gas Phase Fractionation Increases HLA-Peptidome Coverage for Primary Cell and Tissue Samples. *Molecular & Cellular Proteomics* **20**, 100133-100133 (2021). https://doi.org:10.1016/J.MCPRO.2021.100133
188 Sarkizova, S. *et al.* A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nature Biotechnology* **38**, 199-209 (2020). https://doi.org:10.1038/s41587-019-0322-9

189 Parker, R. *et al.* The Choice of Search Engine Affects Sequencing Depth and HLA Class I Allele-Specific Peptide Repertoires. *Molecular & cellular proteomics : MCP* **20**, 100124 (2021). https://doi.org:10.1016/j.mcpro.2021.100124
190 M, A., B, A. & M, N. GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic acids research* **45**, W458-W463 (2017). https://doi.org:10.1093/NAR/GKX248
191 Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic Acids Research* **43**, D405-D412 (2015). https://doi.org:10.1093/NAR/GKU938
192 Meadows, L. *et al.* The HLA-A*0201-Restricted H-Y Antigen Contains a Posttranslationally Modified Cysteine That Significantly Affects T Cell Recognition. *Immunity* **6**, 273-281 (1997). https://doi.org:10.1016/s1074-7613(00)80330-1
193 Geiszler, D. J. *et al.* PTM-Shepherd: Analysis and Summarization of Post-Translational and Chemical Modifications From Open Search Results. *Mol Cell Proteomics* **20**, 100018 (2021). https://doi.org:10.1074/mcp.TIR120.002216
194 Bassani-Sternberg, M. *et al.* Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nature Communications* **7**, 13404 (2016). https://doi.org:10.1038/ncomms13404
195 Gao, E. *et al.* Data-independent acquisition-based proteome and phosphoproteome profiling across six melanoma cell lines reveals determinants of proteotypes. *Molecular Omics* **17**, 413-425 (2021). https://doi.org:10.1039/D0MO00188K
196 Williams, S. M. *et al.* Automated Coupling of Nanodroplet Sample Preparation with Liquid Chromatography-Mass Spectrometry for High-Throughput Single-Cell Proteomics. *Analytical Chemistry* **92**, 10588-10596 (2020). https://doi.org:10.1021/ACS.ANALCHEM.0C01551/SUPPL_FILE/AC0C01551_SI_001.PDF
197 Boekweg, H. *et al.* Features of Peptide Fragmentation Spectra in Single-Cell Proteomics. *Journal of Proteome Research* **21**, 182-188 (2022). https://doi.org:10.1021/acs.jproteome.1c00670
198 An, Y. H. *et al.* High-Efficient Production of Adipose-Derived Stem Cell (ADSC) Secretome Through Maturation Process and Its Non-scarring Wound Healing Applications. *Front Bioeng Biotechnol* **9**, 681501 (2021). https://doi.org:10.3389/fbioe.2021.681501
199 Guyon, I. & De, A. t. M. An Introduction to Variable and Feature Selection André Elisseeff. *Journal of Machine Learning Research* **3**, 1157-1182 (2003).
200 Polasky, D. A., Geiszler, D. J., Yu, F. & Nesvizhskii, A. I. Multiattribute Glycan Identification and FDR Control for Glycoproteomics. *Molecular & cellular proteomics : MCP* **21**, 100205 (2022). https://doi.org:10.1016/j.mcpro.2022.100205
201 Jin, C., Harvey, D. J., Struwe, W. B. & Karlsson, N. G. Separation of Isomeric O-Glycans by Ion Mobility and Liquid Chromatography–Mass Spectrometry. *Analytical Chemistry* **91**, 10604-10613 (2019). https://doi.org:10.1021/ACS.ANALCHEM.9B01772

202 Hofmann, J. & Pagel, K. Glycan Analysis by Ion Mobility–Mass Spectrometry. *Angewandte Chemie International Edition* **56**, 8342-8349 (2017). https://doi.org:10.1002/ANIE.201701309
203 Fondrie, W. E. & Noble, W. S. mokapot: Fast and Flexible Semisupervised Learning for Peptide Detection. *J Proteome Res* **20**, 1966-1971 (2021). https://doi.org:10.1021/acs.jproteome.0c01010
204 Degroeve, S. *et al.* ionbot: a novel, innovative and sensitive machine learning approach to LC-MS/MS peptide identification. *bioRxiv*, 2021.2007.2002.450686-452021.450607.450602.450686 (2021). https://doi.org:10.1101/2021.07.02.450686
205 Yu, F., Deng, Y. & Nesvizhskii, A. I. MSFragger-DDA+ enhances peptide identification sensitivity with full isolation window search. *Nature Communications* **16** (2025). https://doi.org:10.1038/s41467-025-58728-z
206 Muth, T. & Renard, B. Y. in *Briefings in Bioinformatics* Vol. 19 954-970 (Oxford Academic, 2018).
207 Yilmaz, M., Fondrie, W. E., Bittremieux, W., Oh, S. & Noble, W. S. 25514-25522 (PMLR, 2022).
208 Yilmaz, M. *et al. Sequence-to-sequence translation from mass spectra to peptides with a transformer model* (Cold Spring Harbor Laboratory, 2023).
209 Gonnelli, G. *et al.* A Decoy-Free Approach to the Identification of Peptides. *Journal of Proteome Research* **14**, 1792-1798 (2015). https://doi.org:10.1021/pr501164r
210 Wozniak, J. M. *et al.* Enhanced mapping of small-molecule binding sites in cells. *Nat Chem Biol* **20**, 823-834 (2024). https://doi.org:10.1038/s41589-023-01514-z
211 Fu, Y. & Qian, X. Transferred subgroup false discovery rate for rare post-translational modifications detected by mass spectrometry. *Mol Cell Proteomics* **13**, 1359-1368 (2014). https://doi.org:10.1074/mcp.O113.030189
212 Desai, H. *et al.* Multi-omic stratification of the missense variant cysteinome. *bioRxiv* (2023). https://doi.org:10.1101/2023.08.12.553095
213 Ferreira, H. J. *et al.* Immunopeptidomics-based identification of naturally presented non-canonical circRNA-derived peptides. *Nat Commun* **15**, 2357 (2024). https://doi.org:10.1038/s41467-024-46408-3
214 Toprak, U. H. *et al.* Conserved Peptide Fragmentation as a Benchmarking Tool for Mass Spectrometers and a Discriminating Feature for Targeted Proteomics. *Molecular & Cellular Proteomics : MCP* **13**, 2056-2056 (2014). https://doi.org:10.1074/MCP.O113.036475
215 Li, Y. *et al.* Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nature Methods* **18**, 1524-1531 (2021). https://doi.org:10.1038/s41592-021-01331-z
216 Wen, B., Li, K., Zhang, Y. & Zhang, B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nature Communications* **11** (2020). https://doi.org:10.1038/s41467-020-15456-w
217 Kacen, A. *et al.* Post-translational modifications reshape the antigenic landscape of the MHC I immunopeptidome in tumors. *Nat Biotechnol* (2022). https://doi.org:10.1038/s41587-022-01464-2

218 Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nature Methods* **11**, 1114-1125 (2014). https://doi.org:10.1038/nmeth.3144
219 Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Research* **47**, D442-D442 (2019). https://doi.org:10.1093/NAR/GKY1106
220 Lautenbacher, L. *et al. Koina: Democratizing machine learning for proteomics research* (Cold Spring Harbor Laboratory, 2024).
221 Gabriel, W. *et al.* Deep Learning Enhances Precision of Citrullination Identification in Human and Plant Tissue Proteomes. *Molecular & Cellular Proteomics* **24**, 100924 (2025). https://doi.org:https://doi.org/10.1016/j.mcpro.2025.100924
222 Declercq, A. *et al.* Updated MS²PIP web server supports cutting-edge proteomics applications. *Nucleic Acids Research* **51**, W338-W342 (2023). https://doi.org:10.1093/nar/gkad335
223 Liu, K., Li, S., Wang, L., Ye, Y. & Tang, H. Full-Spectrum Prediction of Peptides Tandem Mass Spectra using Deep Neural Network. *Analytical chemistry* **92**, 4275-4275 (2020). https://doi.org:10.1021/ACS.ANALCHEM.9B04867
224 Searle, B. C. *et al.* Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nature Communications 2018 9:1* **9**, 1-12 (2018). https://doi.org:10.1038/s41467-018-07454-w
225 MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966-968 (2010). https://doi.org:10.1093/bioinformatics/btq054
226 Mergner, J. *et al.* Mass-spectrometry-based draft of the Arabidopsis proteome. *Nature* **579**, 409-414 (2020). https://doi.org:10.1038/s41586-020-2094-2
227 Giansanti, P. *et al.* Mass spectrometry-based draft of the mouse proteome. *Nature Methods* **19**, 803-811 (2022). https://doi.org:10.1038/s41592-022-01526-y
228 Marcu, A. *et al.* HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. *Journal for ImmunoTherapy of Cancer* **9**, e002071 (2021). https://doi.org:10.1136/jitc-2020-002071
229 Ritz, D., Kinzi, J., Neri, D. & Fugmann, T. Data-Independent Acquisition of HLA Class I Peptidomes on the Q Exactive Mass Spectrometer Platform. *PROTEOMICS* **17**, 1700177 (2017). https://doi.org:10.1002/pmic.201700177
230 Phulphagar, K. M. *et al.* Sensitive, High-Throughput HLA-I and HLA-II Immunopeptidomics Using Parallel Accumulation-Serial Fragmentation Mass Spectrometry. *Molecular & Cellular Proteomics* **22**, 100563 (2023). https://doi.org:10.1016/j.mcpro.2023.100563
231 Serrano, L. R. *et al.* The one hour human proteome. *Molecular & Cellular Proteomics*, 100760 (2024). https://doi.org:10.1016/j.mcpro.2024.100760
232 Gabriel, W. *et al.* Prosit-TMT: Deep Learning Boosts Identification of TMT-Labeled Peptides. *Analytical Chemistry* **94**, 7181-7190 (2022). https://doi.org:10.1021/acs.analchem.1c05435

233 Soltis, A. R. *et al.* Proteogenomic analysis of lung adenocarcinoma reveals tumor heterogeneity, survival determinants, and therapeutically relevant pathways. *Cell Reports Medicine* **3**, 100819 (2022). https://doi.org:10.1016/j.xcrm.2022.100819
234 Adams, C. *et al. Fragment ion intensity prediction improves the identification rate of non-tryptic peptides in timsTOF* (Cold Spring Harbor Laboratory, 2023).
235 Adams, C., Laukens, K., Bittremieux, W. & Boonen, K. Machine learning-based peptide-spectrum match rescoring opens up the immunopeptidome. *PROTEOMICS* **24** (2024). https://doi.org:10.1002/pmic.202300336
236 Bouwmeester, R., Gabriels, R., Van Den Bossche, T., Martens, L. & Degroeve, S. The Age of Data-Driven Proteomics: How Machine Learning Enables Novel Workflows. *Proteomics* **20** (2020). https://doi.org:10.1002/pmic.201900351
237 Krensky, A. M. & Clayberger, C. HLA-DERIVED PEPTIDES AS NOVEL IMMUNOTHERAPEUTICS. *Clinical Immunology and Immunopathology* **75**, 112-116 (1995).
238 Shraibman, B., Kadosh, D. M., Barnea, E. & Admon, A. Human Leukocyte Antigen (HLA) Peptides Derived from Tumor Antigens Induced by Inhibition of DNA Methylation for Development of Drug-facilitated Immunotherapy. *Molecular & Cellular Proteomics* **15**, 3058-3070 (2016). https://doi.org:10.1074/mcp.M116.060350
239 Gomez-Zepeda, D. *et al.* Thunder-DDA-PASEF enables high-coverage immunopeptidomics and is boosted by MS2Rescore with MS2PIP timsTOF fragmentation prediction model. *Nature Communications* **15** (2024). https://doi.org:10.1038/s41467-024-46380-y
240 Shahbazy, M. *et al.* Benchmarking Bioinformatics Pipelines in Data-Independent Acquisition Mass Spectrometry for Immunopeptidomics. *Molecular & Cellular Proteomics* **22** (2023). https://doi.org:10.1016/j.mcpro.2023.100515
241 Van Puyvelde, B. *et al.* A comprehensive LFQ benchmark dataset on modern day acquisition strategies in proteomics. *Scientific Data* **9** (2022). https://doi.org:10.1038/s41597-022-01216-6
242 Dens, C., Adams, C., Laukens, K. & Bittremieux, W. Machine Learning Strategies to Tackle Data Challenges in Mass Spectrometry-Based Proteomics. *J Am Soc Mass Spectrom* **35**, 2143-2155 (2024). https://doi.org:10.1021/jasms.4c00180
243 Shteynberg, D. D. *et al.* PTMProphet: Fast and Accurate Mass Modification the Trans-Proteomic Pipeline. *Journal of Proteome Research* **18**, 4262-4272 (2019). https://doi.org:10.1021/acs.jproteome.9b00205
244 Yu, F., Haynes, S. E. & Nesvizhskii, A. I. IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs. *Molecular & Cellular Proteomics* **20**, 100077-100077 (2021). https://doi.org:10.1016/J.MCPRO.2021.100077
245 Djomehri, S. I. *et al.* Quantitative proteomic landscape of metaplastic breast carcinoma pathological subtypes and their relationship to triple-negative tumors. *Nature Communications* **11** (2020). https://doi.org:10.1038/s41467-020-15283-z
246 Neuhauser, N., Michalski, A., Cox, J. & Mann, M. Expert system for computer-assisted annotation of MS/MS spectra. *Mol Cell Proteomics* **11**, 1500-1509 (2012). https://doi.org:10.1074/mcp.M112.020271

247 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). https://doi.org:10.1021/acs.analchem.3c02321

248 Michalski, A., Neuhauser, N., Cox, J. & Mann, M. A systematic investigation into the nature of tryptic HCD spectra. *Journal of Proteome Research* **11**, 5479-5491 (2012). https://doi.org:10.1021/PR3007045/SUPPL_FILE/PR3007045_SI_001.PDF

249 Khatun, J., Ramkissoon, K. & Giddings, M. C. Fragmentation Characteristics of Collision-Induced Dissociation in MALDI TOF/TOF Mass Spectrometry. *Analytical chemistry* **79**, 3032-3032 (2007). https://doi.org:10.1021/AC061455V

250 Sun, S. *et al.* Deriving the probabilities of water loss and ammonia loss for amino acids from tandem mass spectra. *J Proteome Res* **7**, 202-208 (2008). https://doi.org:10.1021/pr070479v

251 Martin, D. B., Eng, J. K., Nesvizhskii, A. I., Gemmill, A. & Aebersold, R. Investigation of neutral loss during collision-induced dissociation of peptide ions. *Anal Chem* **77**, 4870-4882 (2005). https://doi.org:10.1021/ac050701k

252 Everley, R. A., Huttlin, E. L., Erickson, A. R., Beausoleil, S. A. & Gygi, S. P. Neutral Loss Is a Very Common Occurrence in Phosphotyrosine-Containing Peptides Labeled with Isobaric Tags. *J Proteome Res* **16**, 1069-1076 (2017). https://doi.org:10.1021/acs.jproteome.6b00487

253 Falick, A. M., Hines, W. M., Medzihradszky, K. F., Baldwin, M. A. & Gibson, B. W. Low-mass ions produced from peptides by high-energy collision-induced dissociation in tandem mass spectrometry. *J Am Soc Mass Spectrom* **4**, 882-893 (1993). https://doi.org:10.1016/1044-0305(93)87006-x

254 Hung, C. W., Schlosser, A., Wei, J. & Lehmann, W. D. Collision-induced reporter fragmentations for identification of covalently modified peptides. *Anal Bioanal Chem* **389**, 1003-1016 (2007). https://doi.org:10.1007/s00216-007-1449-y

255 Polasky, D. A. *et al.* MSFragger-Labile: A Flexible Method to Improve Labile PTM Analysis in Proteomics. *Mol Cell Proteomics* **22**, 100538 (2023). https://doi.org:10.1016/j.mcpro.2023.100538

256 Guan, Z., Yates, N. A. & Bakhtiar, R. Detection and characterization of methionine oxidation in peptides by collision-induced dissociation and electron capture dissociation. *Journal of the American Society for Mass Spectrometry* **14**, 605-613 (2003). https://doi.org:https://doi.org/10.1016/S1044-0305(03)00201-0

257 Armirotti, A., Millo, E. & Damonte, G. How to Discriminate Between Leucine and Isoleucine by Low Energy ESI-TRAP MSn. *Journal of the American Society for Mass Spectrometry* **18**, 57-63 (2007). https://doi.org:https://doi.org/10.1016/j.jasms.2006.08.011

258 Zhang, B. & Bassani-Sternberg, M. Current perspectives on mass spectrometry-based immunopeptidomics: the computational angle to tumor antigen discovery. *Journal for ImmunoTherapy of Cancer* **11**, e007073 (2023). https://doi.org:10.1136/jitc-2023-007073

259 Lyon, Y. A., Riggs, D., Fornelli, L., Compton, P. D. & Julian, R. R. The Ups and Downs of Repeated Cleavage and Internal Fragment Production in Top-Down Proteomics. *J Am Soc Mass Spectrom* **29**, 150-157 (2018). https://doi.org:10.1007/s13361-017-1823-8

260 Rolfs, Z. & Smith, L. M. Internal Fragment Ions Disambiguate and Increase Identifications in Top-Down Proteomics. *J Proteome Res* **20**, 5412-5418 (2021). https://doi.org:10.1021/acs.jproteome.1c00599
261 Durbin, K. R., Skinner, O. S., Fellers, R. T. & Kelleher, N. L. Analyzing internal fragmentation of electrosprayed ubiquitin ions during beam-type collisional dissociation. *J Am Soc Mass Spectrom* **26**, 782-787 (2015). https://doi.org:10.1007/s13361-015-1078-1
262 Kraemer, A. I. *et al.* The immunopeptidome landscape associated with T cell infiltration, inflammation and immune editing in lung cancer. *Nature Cancer* **4**, 608-628 (2023). https://doi.org:10.1038/s43018-023-00548-5
263 Solleder, M. *et al.* Mass Spectrometry Based Immunopeptidomics Leads to Robust Predictions of Phosphorylated HLA Class I Ligands. *Mol Cell Proteomics* **19**, 390-404 (2020). https://doi.org:10.1074/mcp.TIR119.001641
264 Valkenborg, D., Jansen, I. & Burzykowski, T. A Model-Based Method for the Prediction of the Isotopic Distribution of Peptides. *Journal of the American Society for Mass Spectrometry* **19**, 703-712 (2008). https://doi.org:https://doi.org/10.1016/j.jasms.2008.01.009
265 Neta, P., Pu, Q.-L., Yang, X. & Stein, S. E. Consecutive neutral losses of H2O and C2H4O from N-Terminal Thr–Thr and Thr–Ser in collision-induced dissociation of protonated peptides: Position dependent water loss from single Thr or Ser. *International Journal of Mass Spectrometry* **267**, 295-301 (2007). https://doi.org:https://doi.org/10.1016/j.ijms.2007.02.055
266 Harrison, A. G. Pathways for Water Loss from Doubly Protonated Peptides Containing Serine or Threonine. *Journal of The American Society for Mass Spectrometry* **23**, 116-123 (2012). https://doi.org:10.1007/s13361-011-0282-x
267 Yang, Y., Horvatovich, P. & Qiao, L. Fragment Mass Spectrum Prediction Facilitates Site Localization of Phosphorylation. *Journal of Proteome Research* **20**, 634-644 (2021). https://doi.org:10.1021/acs.jproteome.0c00580
268 Jiang, X. *et al.* Accelerating Proteomics Using Broad Specificity Proteases. *Journal of Proteome Research* **23**, 1360-1369 (2024). https://doi.org:10.1021/acs.jproteome.3c00852
269 Kohler, D., Staniak, M., Yu, F., Nesvizhskii, A. I. & Vitek, O. An MSstats workflow for detecting differentially abundant proteins in large-scale data-independent acquisition mass spectrometry experiments with FragPipe processing. *Nature Protocols* **19**, 2915-2938 (2024). https://doi.org:10.1038/s41596-024-01000-3
270 Mehta, S. *et al.* A modular, immunopeptidogenomic (iPepGen) analysis pipeline for discovery, verification, and prioritization of cancer peptide neoantigen candidates. *bioRxiv*, 2025.2004.2007.647596 (2025). https://doi.org:10.1101/2025.04.07.647596
271 Takechi, S. *et al. Silyl Ether Enables High Coverage Chemoproteomic Interaction Site Mapping* (American Chemical Society (ACS), 2024).
272 Tüshaus, J. *et al.* Towards routine proteome profiling of FFPE tissue: insights from a 1,220-case pan-cancer study. *The EMBO Journal* **44**, 304-329-329 (2025). https://doi.org:https://doi.org/10.1038/s44318-024-00289-w

273 Oliveira, J. *et al.* Skin mucus and blood plasma as non-lethal sources of malnutrition protein biomarkers in meagre (Argyrosomus regius). *Journal of Proteomics* **316**, 105432 (2025). https://doi.org:https://doi.org/10.1016/j.jprot.2025.105432
274 Pham, H. T. M. *et al.* High-Throughput Screening of 3-Dimensional Co-culture Hair Follicle Mimetic Tissue with an Enhanced Extracellular Matrix for the Screening of Hair Growth-Promoting Compounds. *Biomater Res* **28**, 0125 (2024). https://doi.org:10.34133/bmr.0125
275 Tretyachenko, V. *et al.* Encoded and non-genetic alternative protein variants expand human functional proteome. *bioRxiv*, 2025.2002.2017.638604 (2025). https://doi.org:10.1101/2025.02.17.638604
276 Vieira de Souza, E. *et al.* Rp3: Ribosome profiling-assisted proteogenomics improves coverage and confidence during microprotein discovery. *Nature Communications* **15**, 6839 (2024). https://doi.org:10.1038/s41467-024-50301-4
277 Tsai, Y.-C. *et al.* Dynamic Proximal Interactomics and Chemical Genetic Screening Reveal CCR4-NOT Sequestration in Stress Granules as a Mechanism for Transcript Stabilization. *bioRxiv*, 2025.2002.2004.636058 (2025). https://doi.org:10.1101/2025.02.04.636058
278 Cosenza-Contreras, M. *et al.* TermineR: Extracting information on endogenous proteolytic processing from shotgun proteomics data. *PROTEOMICS* **24**, 2300491 (2024). https://doi.org:https://doi.org/10.1002/pmic.202300491
279 Han, Y. *et al.* A Ratiometric Catalog of Protein Isoform Shifts in the Cardiac Fetal Gene Program. *bioRxiv* (2024). https://doi.org:10.1101/2024.04.09.588716
280 Willems, P., Staes, A., Demichev, V., Devos, S. & Impens, F. Data-independent immunopeptidomics discovery of low-abundant bacterial epitopes. *bioRxiv*, 2025.2005.2014.653952 (2025). https://doi.org:10.1101/2025.05.14.653952
281 Xu, C., Tang, H. & Wang, X. Systematic review on functional umami peptides: Physiological perceptions, bioinformatics and peptidomics approaches, developments and applications. *Trends in Food Science & Technology* **162**, 105068 (2025). https://doi.org:https://doi.org/10.1016/j.tifs.2025.105068
282 Heymans, M. W. & Twisk, J. W. R. Handling missing data in clinical research. *Journal of Clinical Epidemiology* **151**, 185-188 (2022). https://doi.org:10.1016/j.jclinepi.2022.08.016
283 Emmanuel, T. *et al.* A survey on missing data in machine learning. *Journal of Big Data* **8**, 140 (2021). https://doi.org:10.1186/s40537-021-00516-9
284 Zhang, Y., Gao, Z., Wittrup, E., Gryak, J. & Najarian, K. Increasing efficiency of SVMp+ for handling missing values in healthcare prediction. *PLOS Digital Health* **2**, e0000281 (2023). https://doi.org:10.1371/journal.pdig.0000281
285 Dens, C., Yeung, D., Krokhin, O., Laukens, K. & Bittremieux, W. Zero-shot retention time prediction for unseen post-translational modifications with molecular structure encodings. *bioRxiv*, 2024.2012.2018.629045 (2024). https://doi.org:10.1101/2024.12.18.629045
286 Pourpanah, F. *et al.* A Review of Generalized Zero-Shot Learning Methods. *IEEE Trans Pattern Anal Mach Intell* **45**, 4051-4070 (2023). https://doi.org:10.1109/tpami.2022.3191696

287 Wallmann, G. *et al.* AlphaDIA enables End-to-End Transfer Learning for Feature-Free Proteomics. *bioRxiv*, 2024.2005.2028.596182 (2024). https://doi.org:10.1101/2024.05.28.596182
288 Pan, S. J. & Yang, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* **22**, 1345-1359 (2010). https://doi.org:10.1109/TKDE.2009.191
289 Ctortecka, C. *et al.* Comparative Proteome Signatures of Trace Samples by Multiplexed Data-Independent Acquisition. *Molecular & Cellular Proteomics* **21**, 100177 (2022). https://doi.org:10.1016/j.mcpro.2021.100177
290 Cho, H. *et al.* Privacy-Enhancing Technologies in Biomedical Data Science. *Annu Rev Biomed Data Sci* **7**, 317-343 (2024). https://doi.org:10.1146/annurev-biodatasci-120423-120107
291 Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology* **32**, 219-223 (2014). https://doi.org:10.1038/nbt.2841
292 Pino, L. K. *et al.* The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. *Mass Spectrom Rev* **39**, 229-244 (2020). https://doi.org:10.1002/mas.21540