Structure-Aware Inference from Incomplete Mass Spectra: From *De Novo* Sequencing to PTM Discovery

by

Zeping Mao

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2026

© Zeping Mao 2026

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

Supervisor(s):	Ming Li University Professor, David R. Cheriton School of Computer Science, University of Waterloo
Internal Member:	Lila Kari Professor and Cheriton Faculty Fellow, David R. Cheriton School of Computer Science, University of Waterloo
	Forbes Burkowski Associate Professor Emeritus / Adjunct Associate Professor, David R. Cheriton School of Computer Science, University of Waterloo
Internal-External Member:	Brendan McConkey Associate Professor, Department of Biology, University of Waterloo
External Member:	Pavel Pevzner Ronald R. Taylor Chair and Distinguished Professor of Computer Science, University of California, San Diego

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

I was the sole author of Chapters 1, 2, 7, and 8, as well as the dissertation-level framing, chapter introductions, chapter transitions, and cross-chapter synthesis that connect Chapters 3–6. These sections were written under the supervision of Professor Ming Li and were not written for publication.

This thesis consists in part of four manuscripts written for publication. Material from these publications has been adapted, reorganized, paraphrased, and synthesized for the purposes of this dissertation. Exceptions to sole authorship of material are as follows:

Research presented in Chapter 3:

This research was conducted at the University of Waterloo by Zeping Mao under the supervision of Professor Ming Li, with collaboration from Ruixue Zhang and Lei Xin. Zeping Mao and Ruixue Zhang contributed equally to the published article. Zeping Mao conceived the initial idea and prototype of GraphNovo, developed the central computational formulation, co-implemented the proposed algorithm with Ruixue Zhang, and co-wrote the manuscript with Ruixue Zhang. Ruixue Zhang contributed to algorithm implementation, evaluation, results analysis, and manuscript writing. Lei Xin and Ming Li revised the manuscript, and Ming Li supervised the research project. Chapter 3 adapts and synthesizes material from this publication.

Citations:

Chapter 3: Mao Z, Zhang R, Xin L, Li M. Mitigating the missing-fragmentation problem in *de novo* peptide sequencing with a two-stage graph-based deep learning model. *Nature Machine Intelligence*. 2023;5(11):1250–1260. doi:10.1038/s42256-023-00738-x.

Research presented in Chapter 4:

This research was conducted through a collaboration between the University of Waterloo and Tsinghua University. Qianqiu Zhang, Zeping Mao, and Yuling Chen contributed equally to the manuscript. Qianqiu Zhang led the project. Qianqiu Zhang and Zeping Mao designed and implemented the model architecture and co-wrote the manuscript. Yuling Chen performed the wet-laboratory experiments and generated the EThcD mass-spectrometry data. Baozhen Shan revised the manuscript, and Professor Haiteng Deng and Professor Ming Li supervised the project. Chapter 4 adapts, reorganizes, and synthesizes material from this manuscript.

Citation:

Chapter 4: Zhang Q, Mao Z, Chen Y, Shan B, Deng H, Li M. High-accuracy glycan *de novo* prediction for N- and O-linked glycopeptides across multiple fragmentation techniques. Manuscript.

Research presented in Chapter 5:

This research was conducted through a collaboration involving the University of Waterloo, Bioinformatics Solutions Inc., and other collaborators. Ngoc Hieu Tran, Rui Qiao, Zeping Mao, and Shengying Pan contributed equally to the published article. Zeping Mao contributed to the conceptualization and methodology of the work, provided the GraphNovo code and related technical support, contributed ideas for validation experiments, and participated in reviewing and revising the manuscript. Ngoc Hieu Tran, Rui Qiao, Lei Xin, and Baozhen Shan led major aspects of the conceptualization and methodological design; Shengying Pan contributed software; Ngoc Hieu Tran led formal analysis; Qing Zhang and Wenting Li contributed validation and data curation; and Lei Xin, Ming Li, and Baozhen Shan provided supervision. Chapter 5 adapts and synthesizes material from this publication.

Citations:

Chapter 5: Tran NH, Qiao R, Mao Z, Pan S, Zhang Q, Li W, Xin L, Li M, Shan B. NovoBoard: A Comprehensive Framework for Evaluating the False Discovery Rate and Accuracy of *De Novo* Peptide Sequencing. *Molecular & Cellular Proteomics*. 2024;23(11):100849. doi:10.1016/j.mcpro.2024.100849.

Research presented in Chapter 6:

This research was conducted by Zeping Mao under the supervision of Professor Ming Li and Professor Haiteng Deng, with collaboration from Chao Peng, Yuling Chen, Ping Wu, Qianqiu Zhang, Yonghan Yu, Ruixue Zhang, Lei Xin, Baozhen Shan, and other co-authors. Zeping Mao conceptualized the study, contributed resources, developed the methodology with Qianqiu Zhang and Yonghan Yu, participated in the investigation with Chao Peng, Yuling Chen, Ping Wu, Qianqiu Zhang, Yonghan Yu, and Ruixue Zhang, created the visualizations, and co-drafted the manuscript with Qianqiu Zhang. Chao Peng, Yuling Chen, and Ping Wu contributed to data curation; Chao Peng and Zeping Mao contributed resources; Ming Li obtained funding; Ming Li and Haiteng Deng provided project administration and supervision; and all authors contributed to review and editing. Chapter 6 adapts and synthesizes material from this publication.

Citations:

Chapter 6: Mao Z, Peng C, Chen Y, Wu P, Zhang Q, Yu Y, Zhang R, Xin L, Shan B, Deng H, Li M. Zero-shot *de novo* peptide sequencing with open posttranslational modification discovery. *Nature Biotechnology*. Published online May 19, 2026. doi:10.1038/s41587-026-03116-1.

My roles differed across the projects adapted in Chapters 3–6. The chapter-specific statements above specify my contributions to each project and the corresponding responsibilities of my co-authors. My more focused contribution to the work adapted in Chapter 5 is likewise described there.

Abstract

Machine-learning systems are often developed under a convenient assumption: the set of possible answers is known before inference begins. Molecular discovery violates this assumption. A tandem mass spectrum is an incomplete and noisy record of a molecule, while the databases and molecular vocabularies used to train computational models contain only previously described molecules. A model forced to choose among known answers may therefore confuse missing evidence with a negative result or assign an unexpected molecule to the nearest familiar category.

This dissertation argues that computational inference should preserve both what a measurement supports and what it leaves unresolved until additional evidence justifies a more specific molecular claim. Rather than treating a database mismatch as the end of analysis, the methods developed here retain informative relationships among measured fragments, use them to formulate explicit hypotheses, and keep the evidence that generates a hypothesis distinct from the evidence used to accept it.

Experiments on peptide and glycopeptide inference demonstrate why this distinction matters. When peptide fragments are missing, postponing uncertain local sequence decisions prevents one ambiguous region from corrupting the remainder of the sequence. For branched glycans, relationships between measured fragments and partial molecular compositions recover evidence that cannot be represented by a single linear ordering. The accompanying evaluation work further shows that apparent accuracy and estimated error rates depend on the data distribution, scoring method, and construction of negative controls. Expanding the space of possible molecular explanations must therefore be accompanied by evidence appropriate to the claim being made.

These studies culminate in RNovA, which extends de novo sequencing from recognizing predefined modifications to discovering post-translational modifications beyond the training data and reference vocabulary. Without retraining for each new modification, RNovA can detect unexpected mass changes, reconstruct the corresponding modified peptides, and turn computational outputs into experimentally testable molecular hypotheses. Controlled experiments recovered modification types absent from training, while selected discoveries in biological samples were supported using synthetic reference peptides. Together, these results show that artificial intelligence need not treat an existing database as the boundary of molecular discovery. By preserving structural evidence in the measurement and making unresolved information explicit, it can propose and test molecular explanations that were not enumerated in advance.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Ming Li, for his guidance, support, and encouragement throughout my doctoral study. His broad vision, scientific insight, and willingness to pursue ambitious research questions have shaped the way I think about computational biology, machine learning, and scientific discovery. I am especially grateful for the freedom he gave me to explore difficult problems, as well as for his advice in turning early ideas into concrete research directions.

I am also deeply grateful to Professor Haiteng Deng for his guidance and support on the biological and experimental aspects of this work. His expertise in mass spectrometry, proteomics, and biological interpretation was essential for connecting the computational methods developed in this thesis to real biological questions. Many of the mass-spectrometry applications in this dissertation benefited from his scientific perspective and from the experimental resources and validation support provided by his group.

I would like to thank my committee members for their time, feedback, and encouragement. Their comments helped me think more carefully about the broader context of this work, its limitations, and its potential impact beyond *de novo* peptide sequencing.

I am grateful to all of my collaborators and co-authors on the GraphNovo, DeepGlycan, NovoBoard, and RNovA projects. These works would not have been possible without their contributions to experimental design, data generation, software development, biological validation, discussion, and manuscript revision. I have learned a great deal from working with collaborators across computer science, proteomics, and biology.

I would also like to thank the members of the laboratory and research community around me for many helpful discussions, technical suggestions, and moments of encouragement. The process of developing research ideas, debugging models, analyzing experiments, and writing papers is rarely solitary, and I am grateful for the people who made this process intellectually lively and personally sustainable.

Finally, and most importantly, I would like to thank my fiancée, Jingrui Chen. Her love, patience, and unwavering support carried me through many moments of failure, frustration, and self-doubt. During the most difficult parts of my doctoral journey, she encouraged me to keep going, helped me regain confidence, and reminded me why this work mattered. This thesis would not have been completed without her companionship, understanding, and belief in me.

Table of Contents

Examining Committee Membership ii

Author's Declaration iii

Statement of Contributions iv

Abstract vii

Acknowledgements viii

List of Figures xiii

List of Abbreviations xv

List of Symbols xvi

1 Introduction: From Molecular Identification to Discovery 1

1.1 Molecular discovery beyond the catalogue 1

1.2 What a tandem mass spectrum can reveal 2

1.3 The closed world of molecular identification 3

1.4 From unexplained mass to open PTM discovery 5

2 Tandem Mass Spectrometry and Peptide Inference **7**

2.1 Proteins become peptide ions . . . 7

2.2 From peptide mixtures to tandem mass spectra . . . 8

2.3 Reading sequence information from b and y ions . . . 10

2.4 Why the fragment ladder is incomplete . . . 12

2.5 Database search and de novo sequencing . . . 12

2.6 Modifications and the open chemical space . . . 13

3 GraphNovo: Mitigating Missing Fragmentation in *De Novo* Peptide Sequencing **16**

3.1 Missing Fragmentation in *De Novo* Peptide Sequencing . . . 17

3.2 Spectrum Graph Representation . . . 18

3.3 PathSearcher and SeqFiller . . . 21

3.4 Training Data and Evaluation . . . 24

3.5 Sequencing Accuracy and Path Recovery . . . 25

3.6 Error Propagation across Missing Regions . . . 27

3.7 Performance across Missing-Fragmentation and Spectrum Conditions . . . 29

3.8 Discussion . . . 32

4 DeepGlycan: *De Novo* Glycan Composition Prediction from Intact Glycopeptide Spectra **34**

4.1 Introduction . . . 34

4.2 Prediction Task and Data . . . 35

4.3 DeepGlycan Architecture . . . 37

4.3.1 Peak-to-Node Representation . . . 38

4.3.2 Relative-Mass Attention . . . 40

4.3.3 Spectrum-Conditioned Composition Decoding . . . 40

4.4 Dispersed Fragment Evidence and the Glyco-Gap . . . 42

4.5 Evaluation Across Fragmentation Methods and Glycan Classes . . . 43

4.5.1 sceHCD N-Glycopeptide Benchmark . . . 43
4.5.2 EThcD Evaluation and N-to-O Transfer . . . 45
4.6 Biological Applications . . . 46
4.6.1 Engineered Glycosylation Systems . . . 48
4.6.2 Expansion of an Alzheimer's Disease Glycoproteomics Analysis . . . 50
4.7 Discussion . . . 50
4.8 Conclusion . . . 52

5 NovoBoard: Evaluating Accuracy and False Discovery in *De Novo* Peptide Sequencing 53
5.1 Metrics at Multiple Levels . . . 54
5.2 Benchmarking beyond Familiar Spectra . . . 55
5.3 Target–Decoy Spectra for *De Novo* FDR . . . 56
5.4 Calibrating the Error Estimate . . . 59
5.5 Generalization across Species, Digestion, and Peptide Class . . . 62
5.6 Controlled Sequence Novelty and Training-Set Familiarity . . . 65
5.7 Discussion . . . 68

6 RNovA: Zero-Shot *De Novo* Peptide Sequencing with Open PTM Discovery 71
6.1 Introduction . . . 71
6.2 RNovA Architecture and Open PTM Workflow . . . 72
6.2.1 PathSearcher . . . 75
6.2.2 Mass-Tag Clustering and Candidate Generation . . . 75
6.2.3 Candidate-Conditioned Sequence Reconstruction . . . 76
6.2.4 Relative-Mass-Aware Learning . . . 79
6.3 Evaluation on Standard and Modified Peptides . . . 79
6.3.1 Standard Peptide Sequencing . . . 79
6.3.2 Zero-Shot Reconstruction of Modified Peptides . . . 82

6.4 False-Discovery Estimation and Isotope Validation 84
6.4.1 Target–Decoy Calibration 84
6.4.2 Validation with 15N-Labeled Yeast 87
6.5 End-to-End Discovery of Nonnatural Modifications 89
6.6 Open PTM Discovery in Biological Samples 93
6.6.1 Rheumatoid Arthritis and Selected Kynurenine Assignments 93
6.6.2 A1232E: Sequence and Modification Without a Reference Proteome 94
6.7 Generalization Across Acquisition Platforms 97
6.8 Discussion 99
6.9 Conclusion 100

7 Discussion: From Unexplained Spectra to Molecular Discovery 101
7.1 Unexplained spectra and the limits of closed identification 101
7.2 Constructing hypotheses from partial evidence 102
7.3 Reliability of open molecular inference 104
7.4 Scope, implications, and future work 106

8 Conclusions 109

References 111

Glossary 129

List of Figures

1.1 From closed identification to evidence-guided molecular discovery 4

2.1 Bottom-up LC–MS/MS workflow 9
2.2 Sequence information in b/y ions and a PTM mass shift 11
2.3 Candidate construction in database search and de novo sequencing 14

3.1 Overview of GraphNovo 20
3.2 GraphNovo model details 23
3.3 Overall GraphNovo evaluation 26
3.4 Evaluation of GraphNovo-PathSearcher 28
3.5 GraphNovo and missing fragmentation 30
3.6 Effect of missing-fragmentation severity 31

4.1 DeepGlycan architecture and spectrum-conditioned composition order . . . 39
4.2 DeepGlycan sceHCD benchmark and glyco-gap analysis 44
4.3 DeepGlycan EThcD N-to-O transfer 47
4.4 DeepGlycan profiles in glycosylation-engineered systems 49
4.5 DeepGlycan-assisted Alzheimer's disease analysis 51

5.1 Target–decoy spectrum workflow 58
5.2 Decoy-spectrum difficulty 60
5.3 FDR estimation and validation 61

5.4 ABRF and *A. thaliana* evaluation 64
5.5 Nontryptic and HLA evaluation 66
5.6 Evaluation on simulated mutated peptides 67
5.7 Peptide-sequence memorization analysis 69

6.1 RNovA architecture and open PTM workflow 74
6.2 Candidate-conditioned sequence reconstruction in RNovA 78
6.3 Conventional peptide sequencing with RNovA 81
6.4 Zero-shot reconstruction and localization of diverse PTMs 83
6.5 Calibration of RNovA decoy-spectrum corruption 85
6.6 Isotope-based evaluation of RNovA confidence thresholds 88
6.7 Controlled open discovery with nonnatural TestMods 90
6.8 Abundance dependence in the TestMod experiment 92
6.9 Real-sample hypotheses and selected synthetic-reference tests 96
6.10 Cross-platform generalization of RNovA 98

List of Abbreviations

DDA data-dependent acquisition

DIA data-independent acquisition

FDR false discovery rate

GPSM glycopeptide-spectrum match

HCD higher-energy collisional dissociation

EThcD electron-transfer/higher-energy collision dissociation

HDX-MS hydrogen-deuterium exchange mass spectrometry

LC-MS/MS liquid chromatography-tandem mass spectrometry

MDL minimum description length

MS mass spectrometry

MS/MS tandem mass spectrometry

PSM peptide-spectrum match

PTM post-translational modification

RoPE rotary positional embedding

sceHCD stepped collision energy higher-energy collisional dissociation

MGF Mascot generic format

XL-MS cross-linking mass spectrometry

List of Symbols

m/z Mass-to-charge ratio.

Δm Mass difference or mass shift.

Chapter 1

Introduction: From Molecular Identification to Discovery

1.1 Molecular discovery beyond the catalogue

Mass-spectrometry-based proteomics has made it possible to examine proteins and their chemical variation at a scale that would be difficult to achieve by targeted experiments alone. The molecules observed in a biological sample, however, are not limited to the unmodified protein sequences encoded by a genome. Proteolytic processing, sequence variation, chemical damage, and post-translational modification (PTM) can generate many molecular forms from the same gene product [1]. These changes are central to processes such as cellular signaling, immune regulation, and chromatin control [2, 3, 4]. Identifying them is therefore not merely an annotation problem. It is one route by which previously unrecognized molecular states become available for biological investigation.

Most computational proteomics pipelines begin from a finite catalogue of possible answers. A database-search engine considers peptides derived from reference proteins together with a specified set of modifications [5]. A learned sequence model uses an output vocabulary defined before training. These restrictions make inference efficient and make statistical evaluation possible, but they also determine which molecules the analysis is capable of reporting. If the true peptide or modification lies outside the candidate space, the closed system cannot return the correct identity; if forced to report a candidate, it may return the best available wrong answer.

Curated databases and modification catalogues are indispensable records of accumulated knowledge. The difficulty arises when the boundary of that knowledge is treated as the

boundary of molecular possibility. An unexpected spectrum may then be rejected as uninterpretable, assigned to a nearby known molecule, or discarded before the unexplained part of the measurement is examined [6]. In each case, such a closed system is effective at recognizing established categories but unable to formulate a candidate that those categories do not already contain.

This dissertation addresses a different problem: how can a computational method depart from known molecular labels while remaining constrained by physical evidence? The answer cannot be unconstrained generation. A discovery method must preserve what the measurement supports, distinguish that support from what the model merely prefers, and turn unexplained evidence into a candidate that can be tested. Tandem mass spectrometry provides a particularly revealing setting for this problem because its output is both structurally informative and inherently incomplete.

1.2 What a tandem mass spectrum can reveal

In bottom-up proteomics, proteins are digested into shorter peptides. A peptide precursor is isolated, fragmented, and measured as a tandem mass spectrum. The spectrum contains peaks at the mass-to-charge ratios of fragment ions, together with their measured intensities. Because many fragments retain one end of the peptide, mass differences between related peaks can provide evidence for the residues that lie between successive backbone cleavage sites [7, 8]. A collection of such relationships can support a peptide sequence even though the instrument never reads that sequence directly.

The evidence is rarely complete. Some bonds do not produce detectable fragments, and some ions fall below the detection threshold [9]. Many observed peaks arise from other ion types, neutral losses, co-isolated molecules [10], or noise. When fragments on both sides of a sequence interval are observed but fragments within the interval are missing, the spectrum may determine the interval's total mass without determining its internal residue order. An unexpected modification can produce a related pattern: the sequence remains largely compatible with the observed fragments, but one region carries a mass that is not explained by the unmodified residues.

This incompleteness changes the computational task. A decoder that must emit one residue at every step is forced to make a specific decision even when the spectrum supports only a cumulative mass or a sequence interval. If that decision is wrong, the decoder's running mass becomes incorrect and later predictions can fail even after direct fragment evidence returns. Retaining supported mass checkpoints before completing the unresolved

interval prevents a local absence of fragments from being immediately converted into an unsupported sequence claim [11].

The same principle extends beyond linear peptides. A glycan is assembled from monosaccharides and may be branched. Its fragments can support several partial compositions that do not lie along a single linear order. A model that receives only a serialized target can consequently hide useful relationships between the spectrum and those partial compositions. Representing the relevant mass relationships directly makes more of the measured evidence available to inference without confusing an arbitrary decoding order with molecular topology [12].

These examples reveal a common structure in the inverse problem. A spectrum may support different parts of a molecular description at different resolutions: an individual residue, an unresolved mass interval, or a partial composition. The computational representation determines whether those intermediate forms of evidence remain available or are replaced prematurely by a complete answer. Preserving them is not a refusal to make a prediction. It is what allows a later prediction to remain connected to the measurement that motivated it. Figure 1.1 summarizes the resulting distinction: closed identification ranks answers from a predefined catalogue, whereas evidence-guided discovery can preserve a localized, recurrent mass interval as a candidate for independent testing.

1.3 The closed world of molecular identification

Database search makes peptide identification tractable by restricting the candidate space. Reference proteins are digested computationally, candidate peptides compatible with the precursor mass are generated, and each candidate is scored against the observed fragments [5, 13]. Statistical procedures then estimate errors among the accepted peptide–spectrum matches [14, 15]. When the relevant peptide and modification are included in the search space, this approach is accurate, scalable, and biologically informative.

Its strength and its boundary arise from the same decision. A database search cannot identify a molecule that was never admitted as a candidate. Enlarging the search space can recover additional variants and modifications, but it also increases computation and the number of competing explanations [16, 17]. More fundamentally, a finite search specification cannot anticipate every rare or previously undescribed molecular change. The difficulty is especially acute for PTMs, because the modifications of greatest discovery interest may be precisely those for which no modification-specific training examples or search parameters are available.

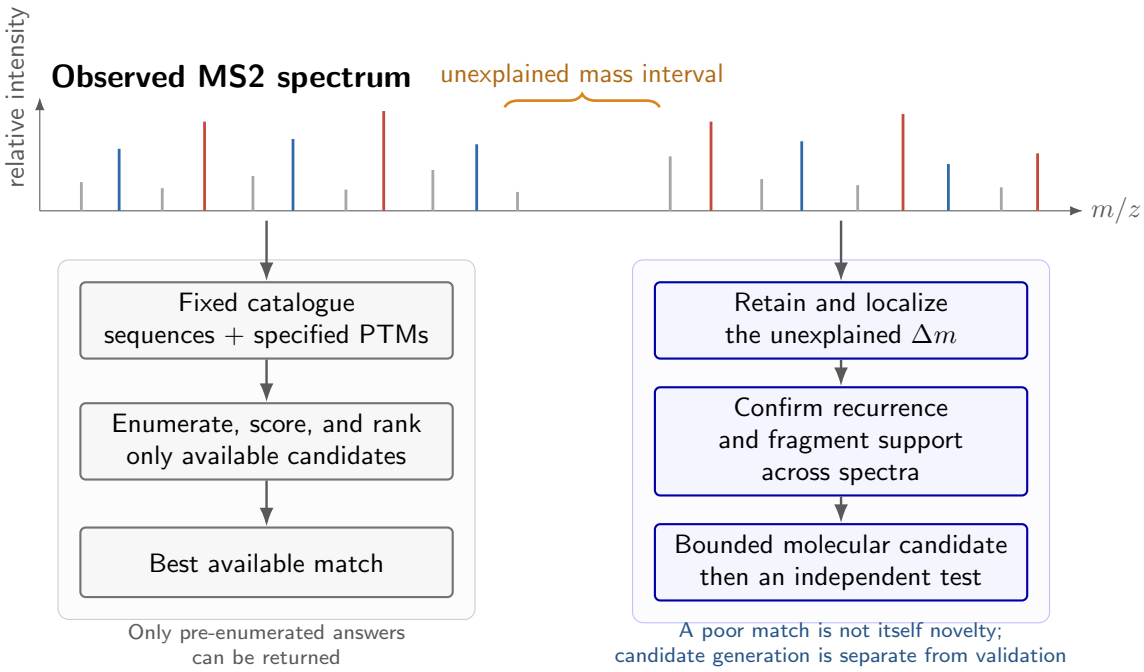


Figure 1.1: From closed identification to evidence-guided molecular discovery. The same tandem mass spectrum can be interpreted against a fixed catalogue of sequences and modifications, in which only pre-enumerated candidates can be returned, or by retaining a mass interval that the current vocabulary cannot explain. The latter route does not declare novelty from a poor match: recurrent, localized evidence is converted into a bounded molecular candidate, and candidate generation remains separate from subsequent validation. Original schematic; the spectrometric details are developed in Chapter 2.

De novo peptide sequencing relaxes one part of this restriction by inferring a sequence directly from the spectrum rather than selecting a complete peptide from a protein database. Classical spectrum-graph and dynamic-programming methods established this formulation [18, 19]. Later scoring systems improved its practical performance [20, 21], and deep learning has subsequently improved its accuracy [22, 23, 24]. Yet database independence does not by itself make the output space open. A neural decoder trained to emit a fixed set of residue and modified-residue tokens still cannot express a modification that is absent from that vocabulary. Supporting a new PTM commonly requires defining a new token, collecting labeled examples, and training the model again.

This limitation has long been recognized as a central computational challenge in protein mass spectrometry. Database-external sequence variation and PTMs become increasingly difficult to identify as the candidate space expands, particularly when reliable error control must be maintained [25]. Recent reviews likewise identify support for PTMs beyond small supervised alphabets, together with false-discovery control for *de novo* sequencing, as unresolved problems [26]. Open PTM discovery therefore requires more than a larger database or a more accurate closed-vocabulary decoder. It requires a mechanism for recognizing and describing evidence that the current vocabulary does not explain.

1.4 From unexplained mass to open PTM discovery

The central argument of this dissertation is that curated molecular knowledge should guide inference without predefining everything that the measurement is allowed to contain. When an existing label fails, the computation should retain the structural relationships that remain supported by the spectrum, use those relationships to construct an explicit molecular candidate, and expose that candidate to evidence beyond the score that proposed it.

The argument begins with incomplete peptide fragmentation. A mass path preserves reliable checkpoints across regions in which residue-level evidence is absent, allowing sequence completion to occur without propagating every local uncertainty through the remainder of the peptide. Branched glycans show why this idea cannot be reduced to one universal graph or decoding order: the representation must follow the relationships produced by the molecular structure and the fragmentation process. In both settings, the gain comes from delaying a more specific claim until the spectrum provides an appropriate basis for making it.

Opening the candidate space also changes what constitutes convincing evaluation. A higher sequence accuracy on database-derived test data does not by itself show that a model

can recognize evidence outside the training distribution. Scores can reflect familiar sequence priors, repeated labels, or properties of the benchmark rather than decisive fragment support. Evaluation across fragments, residues, and complete peptides, together with controlled distribution shifts and target–decoy spectra, shows that candidate generation and reliability assessment must be designed as related but distinct parts of the inference process [27].

These ideas culminate in RNovA. Instead of requiring every observed mass to correspond to a known modified-residue class, its discovery stage retains and localizes unexplained mass tags. Evidence that recurs across spectra is organized into residue–mass-shift candidates, after which candidate-conditioned sequence reconstruction produces complete modified-peptide hypotheses without PTM-specific retraining. Isotope-labeled data provide an external check on the structural confidence thresholds, while synthetic reference peptides test selected chemical assignments [28].

In this thesis, open PTM discovery denotes the conversion of an unexplained spectral mass into a localized and sequence-resolved modification candidate without requiring that modification to have been supplied as a trained class, followed by the advancement of selected candidates to testable chemical assignments. Under this operational definition, this dissertation solves the computational problem of open PTM discovery from tandem mass spectra [28].

The significance of this result extends beyond one proteomics workflow. Many applications of artificial intelligence in biology learn from labels produced by current databases, experimental conventions, and earlier analytical pipelines. Those labels capture valuable knowledge, but they also inherit the limits of the processes that created them. A discovery-oriented learning system should not discard that knowledge; neither should it mistake it for a complete account of the sample. By converting measurements that existing labels cannot explain into structured hypotheses that experiments can examine, artificial intelligence can help enlarge the space of biological questions that can be asked.

Chapter 2

Tandem Mass Spectrometry and Peptide Inference

2.1 Proteins become peptide ions

At the level of primary structure, a protein is a chain of amino-acid residues connected by peptide bonds. The chain has a chemically distinct N-terminus and C-terminus, and its sequence is conventionally written from the N-terminus to the C-terminus. A peptide is a shorter chain built by the same chemistry. In bottom-up proteomics, proteins extracted from a biological sample are first cleaved into peptides by a protease. Trypsin is widely used because it cleaves preferentially on the C-terminal side of lysine or arginine residues [29] and usually produces peptides that are well suited to liquid-chromatographic separation and tandem mass spectrometry [30, 2].

Let a peptide contain residues $r_1, \ldots, r_n$, and let $\mu(r_i)$ denote the monoisotopic residue mass of r_i. Its neutral unmodified mass is

$$M(P) = \sum_{i=1}^{n} \mu(r_i) + m_{\mathrm{H_2O}}, \tag{2.1}$$

where the terminal hydrogen and hydroxyl groups contribute one water molecule. This relation connects sequence with measurement: every proposed sequence implies a precursor mass and a collection of fragment masses. The mapping is not always unique. Leucine and isoleucine, for example, have the same elemental composition and residue mass, and

different combinations of residues can have equal or nearly equal total masses. Mass therefore constrains a peptide sequence without necessarily determining it uniquely. This ambiguity motivates using ordered fragment differences as sequence tags rather than relying on precursor mass alone [31].

2.2 From peptide mixtures to tandem mass spectra

A proteolytic digest contains many peptides whose concentrations can span several orders of magnitude; plasma is an extreme example of this dynamic range [32]. Liquid chromatography separates this mixture over time, although several peptides may still co-elute. As the column eluate enters the mass spectrometer, electrospray ionization transfers peptide molecules into the gas phase, commonly as ions carrying one or more protons [33]. The instrument records a peak at a mass-to-charge ratio rather than directly reporting the neutral molecular mass. For an ion with neutral mass M and charge z, a simplified relation is

$$\left(\frac{m}{z}\right)_{\text{obs}} = \frac{M + z m_{\text{p}}}{z}, \tag{2.2}$$

where m_{p} is the proton mass. The same peptide can therefore appear at several charge states. Terminology in this chapter follows the IUPAC distinction between an isolated precursor ion and the product ions formed after fragmentation [34].

An MS1 survey scan records the m/z and intensity of ions present at a particular chromatographic time. Natural isotopes produce a cluster of peaks for each ion rather than a single isolated peak. Neighboring isotope peaks are separated by approximately $1.0034/z$ in m/z, so the spacing of the isotope envelope helps estimate the charge state and recover a monoisotopic precursor mass [35].

In a data-dependent-acquisition (DDA) LC–MS/MS experiment, the instrument selects a precursor ion from the MS1 scan, isolates a narrow m/z window, and subjects the selected ions to fragmentation. Collision-based methods such as collision-induced dissociation (CID) or higher-energy collisional dissociation (HCD) [36] transfer energy to the peptide ions and cause a population of peptide molecules to break at different positions. A second scan records the resulting product ions and forms an MS2 spectrum. MS1 supplies a constraint on the mass of the intact peptide; MS2 supplies evidence about its internal sequence. Because an isolation window can contain more than one co-eluting ion, an MS2 spectrum can also contain fragments from multiple precursors [10]. Figure 2.1 follows the broader information flow from digestion to the mixed, charge-dependent MS2 peak list that becomes the input to molecular inference.

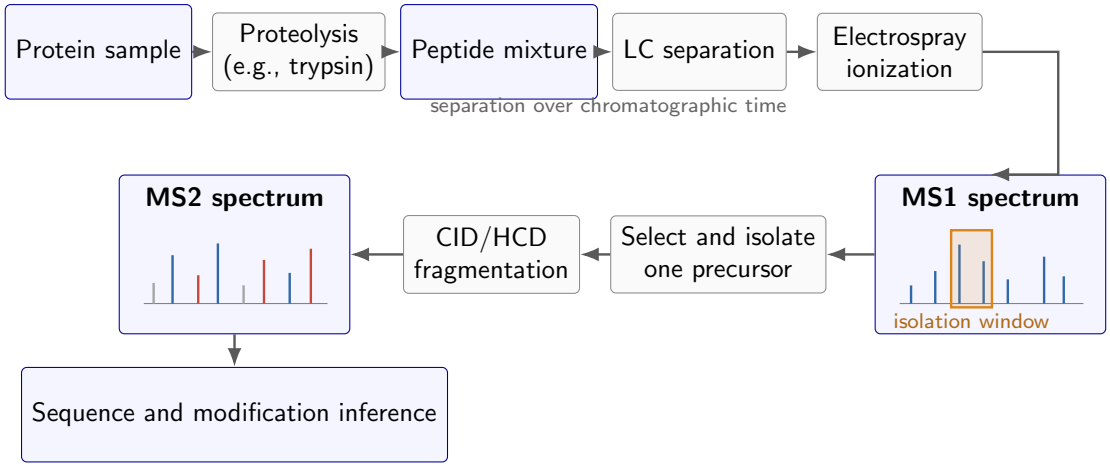


Figure 2.1: Bottom-up LC–MS/MS workflow. Proteins are digested, peptides are partially separated by liquid chromatography and ionized, and an MS1 scan records co-eluting precursor ions. Ions inside a selected isolation window are fragmented to produce an MS2 spectrum. The resulting peak list is a charge-dependent and potentially mixed observation from which molecular structure must be inferred. Original schematic based on standard bottom-up proteomics workflows [30, 2].

2.3 Reading sequence information from b and y ions

Collision-induced cleavage of a peptide bond commonly produces fragments that retain either the N-terminal or the C-terminal part of the peptide. The conventional b/y nomenclature indexes these fragments by the number of residues they contain [8, 37]. If the bond between r_i and r_{i+1} is cleaved, b_i contains residues $r_1, \ldots, r_i$, whereas y_{n-i} contains residues $r_{i+1}, \ldots, r_n$. Thus b ions are numbered from the N-terminus and y ions from the C-terminus.

For singly charged fragments, the theoretical monoisotopic mass-to-charge ratios are

$$\left(\frac{m}{z}\right)(b_i^+) = \sum_{k=1}^{i} \mu(r_k) + m_{\mathrm{p}}, \tag{2.3}$$

$$\left(\frac{m}{z}\right)(y_j^+) = \sum_{k=n-j+1}^{n} \mu(r_k) + m_{\mathrm{H_2O}} + m_{\mathrm{p}}. \tag{2.4}$$

Consequently, adjacent members of the same ion series differ by one residue mass:

$$\left(\frac{m}{z}\right)(b_i^+) - \left(\frac{m}{z}\right)(b_{i-1}^+) = \mu(r_i), \tag{2.5}$$

$$\left(\frac{m}{z}\right)(y_j^+) - \left(\frac{m}{z}\right)(y_{j-1}^+) = \mu(r_{n-j+1}). \tag{2.6}$$

A cleavage also defines a complementary pair. For singly charged products,

$$\left(\frac{m}{z}\right)(b_i^+) + \left(\frac{m}{z}\right)(y_{n-i}^+) = M(P) + 2m_{\mathrm{p}}. \tag{2.7}$$

These relations explain why a spectrum can be viewed as a set of partial cumulative sums. If a sufficiently complete sequence of b or y ions is observed, the mass differences between neighboring peaks reveal the residue sequence. Multiply charged fragments obey the same chemistry, but their mass differences are divided by the charge. Charge assignment and deconvolution are therefore part of interpreting the spectrum. Figure 2.2 combines these relations with two departures from the ideal ladder—missing fragments and a localized modification—to show which constraints remain directly supported by the spectrum.

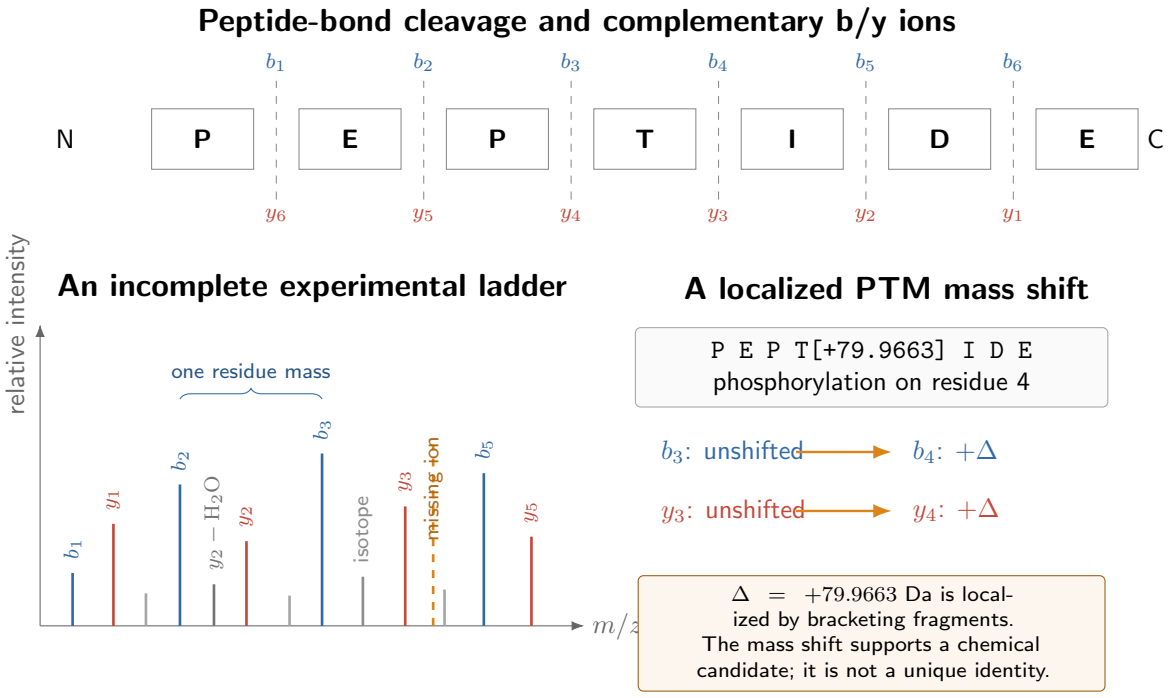


Figure 2.2: Sequence information in b/y ions and a PTM mass shift. Cleavage between r_i and r_{i+1} defines an N-terminal b_i ion and a C-terminal y_{n-i} ion; neighboring ions in either series differ by one residue mass. Experimental ladders are incomplete and contain additional peaks. In the lower-right example, phosphorylation of the fourth residue shifts only fragments that retain that residue by +79.9663 Da. Bracketing fragments can localize the shift, but the numerical shift alone does not establish a unique chemical identity. Original schematic based on standard fragment nomenclature and Unimod mass definitions [8, 37, 38].

2.4 Why the fragment ladder is incomplete

The idealized fragment ladder is a useful explanation, but it is not what an instrument returns. Fragmentation efficiency depends on peptide sequence, precursor charge, collision energy, and instrument configuration [39, 9]. Some peptide bonds produce abundant product ions, whereas others produce signals below the detection threshold. A cleavage may be represented by a b ion, a y ion, both, or neither. When several consecutive cleavage positions are unobserved, the peaks on either side constrain only the combined mass of the missing interval, not the order of the residues within it [40, 11].

Observed peaks also have more than one possible explanation. A fragment may appear in multiple charge states and as several isotope peaks. Product ions can lose neutral molecules after backbone cleavage; common examples include loss of water or ammonia, which shifts the observed m/z by the neutral loss divided by the fragment charge. Internal fragments, immonium ions, fragments produced by alternative cleavage pathways, and incompletely removed precursor ions add further peaks. Labile modifications can introduce additional neutral-loss channels [41].

It is therefore useful to distinguish an unassigned peak from pure instrumental noise. Some unassigned peaks are chemical signals whose ion type is not represented by a simple b/y model, while others arise from co-isolated peptides, contaminants, or measurement noise. Peak filtering, centroiding, deisotoping, and charge deconvolution determine which of these signals reaches the computational method [42, 43]. A processed spectrum is thus both incomplete and ambiguous: some expected evidence is absent, and some observed evidence admits several interpretations.

2.5 Database search and de novo sequencing

Database search identifies a spectrum by comparing it with peptides generated from a reference protein database. The search engine digests the protein sequences in silico, adds the modifications allowed by the search configuration, filters candidates using the precursor mass, predicts their fragment ions, and scores peptide–spectrum matches (PSMs) [13, 44, 45]. Accepted matches are often reported at a specified false discovery rate (FDR), the estimated proportion of incorrect assignments among accepted results [15]. This restriction makes large-scale identification efficient and supports established statistical procedures, but the reported answer must come from the candidate space that was constructed.

Spectral-library search changes the source of the reference pattern rather than the basic candidate assumption. An experimental spectrum is compared with previously observed

consensus spectra or spectra predicted for candidate peptides. Fragment intensities can provide discrimination beyond a simple theoretical peak list, but the relevant peptide still needs to be represented in the library or candidate index [46, 47].

De novo peptide sequencing instead constructs a sequence from precursor and fragment masses. Classical methods use spectrum graphs, dynamic programming, or probabilistic networks [18, 19, 48], whereas neural methods learn fragmentation and sequence regularities from labeled spectra [22, 24]. The resulting sequence need not occur in a reference proteome. De novo sequencing is nevertheless not assumption-free: most systems still decode from a fixed set of residue masses, and incomplete fragment ladders can support several mass-consistent sequences [49]. Database search and de novo sequencing therefore use the same physical evidence but place the principal restriction at different stages—before scoring through an enumerated candidate set, or during reconstruction through a residue vocabulary and a sequence model. Figure 2.3 makes this difference in candidate construction explicit and shows where a localized, unexplained mass can be retained instead of being forced into a fixed candidate space.

2.6 Modifications and the open chemical space

The molecular form of a protein is not determined by its unmodified amino-acid sequence alone. Post-translational modifications add or remove chemical groups after translation and can change both peptide mass and fragmentation behavior [50]. If a modification changes the peptide mass by Δ and occurs on residue r_k, fragments that do not contain r_k retain their unmodified masses, whereas fragments that contain r_k are shifted by Δ, provided that the modification is retained during fragmentation. For the b-ion series, b_i is unshifted when $i < k$ and shifted when $i \geq k$; the y-ion series provides the complementary pattern. Observing fragments on both sides of the transition can therefore localize a modification [51].

When the modification is specified in advance, a database search or sequence model can treat the modified residue as an allowed candidate. When it is not specified, the precursor and fragments can still support a localized mass difference even though no available token names it. Wide-tolerance database search expands the candidate space [16, 52], while modification-aware de novo methods approach the same boundary during sequence reconstruction [53, 28]. A recurrent, localized mass shift is the starting point of PTM discovery: spectral evidence establishes where additional mass is required, after which the shift can be compared with known chemistry and tested with further evidence [54].

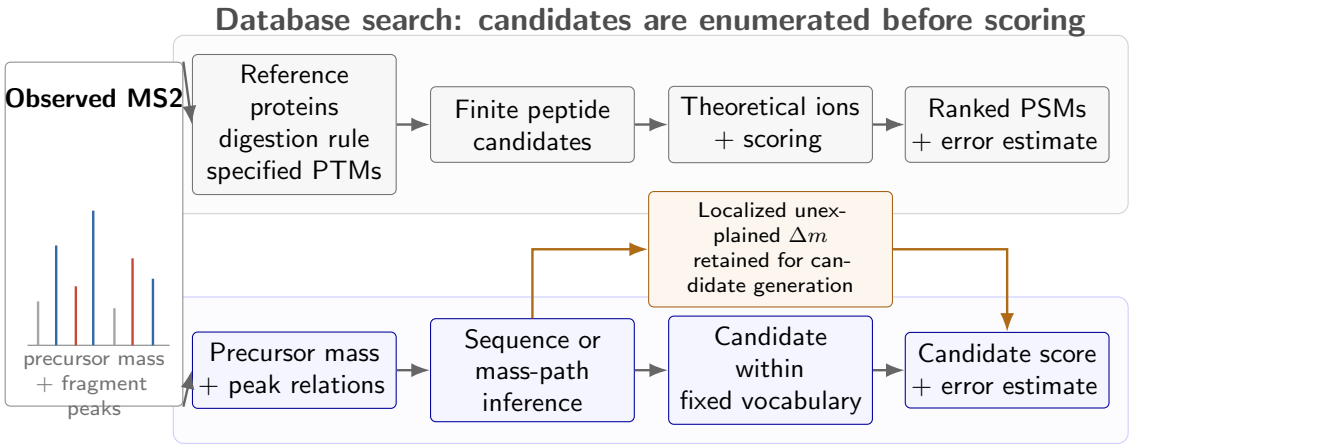


Figure 2.3: Candidate construction in database search and de novo sequencing. Database search enumerates peptides from reference proteins and specified modifications before scoring them against an observed spectrum. De novo sequencing builds a sequence or mass path from precursor and fragment relations, but a standard decoder may still be bounded by a fixed residue and PTM vocabulary. A localized unexplained mass can be retained outside that vocabulary as the starting point for a new candidate. Scoring and error assessment remain necessary in both routes. Original schematic based on established search and de novo formulations [13, 20, 26].

Glycosylation extends the same problem from a modified residue to a branched molecular structure. A glycan is assembled from monosaccharides and is covalently attached to a peptide. Its composition records the number of each monosaccharide class, whereas its topology specifies how those units are connected and branched. Distinct topologies can share the same composition and total mass [55]. Glycopeptide MS2 spectra distribute evidence among peptide backbone fragments, glycan-derived oxonium ions, glycosidic fragments, and peptide ions that retain part of the glycan. Collision-based and electron-transfer-based fragmentation emphasize different subsets of this evidence [56, 57, 58].

Across unmodified peptides, modified peptides, and glycopeptides, the computational task begins with the same physical fact: an MS2 spectrum is a partial collection of mass constraints rather than a direct molecular readout. The next chapter begins with the simplest version of this problem: a linear peptide for which several informative backbone cleavages are absent from the observed spectrum.

Chapter 3

GraphNovo: Mitigating Missing Fragmentation in *De Novo* Peptide Sequencing

De novo peptide sequencing reconstructs an amino-acid sequence from fragment ions rather than selecting a complete peptide from a protein database. In practice, the fragment series is often interrupted: one or more peptide bonds produce no detectable ions, leaving a mass interval whose total mass may be known even though its internal residue order is not. Spectrum-graph and probabilistic methods have long represented such intervals through mass relationships [20, 48]. Autoregressive neural sequencers must nevertheless choose a residue at every step. Once a choice shifts the running prefix mass, later peaks are interpreted from the wrong coordinate, and an ambiguity confined to one interval can corrupt the remainder of the sequence [22, 23].

GraphNovo was developed around this failure mode. It converts the spectrum into a directed graph of candidate cumulative masses and separates sequence reconstruction into two stages. PathSearcher first recovers a mass path through observed fragment evidence. Differences between adjacent path nodes are translated directly when they correspond to a single residue and retained as mass tags when the evidence supports only a larger interval. SeqFiller then resolves those tags under local mass constraints. The path therefore supplies checkpoints at which decoding can rejoin the measured evidence after an uncertain region [11].

The GraphNovo study evaluates this intermediate path independently, uses it to guide two existing neural decoders, and measures accuracy before, within, and after missing-

fragmentation regions. These experiments connect the architecture to the observed improvement: they test whether staged mass inference prevents an error from propagating beyond the gap, rather than treating a higher final peptide recall as sufficient explanation.

3.1 Missing Fragmentation in *De Novo* Peptide Sequencing

Suppose a peptide contains residues $a_1, \ldots, a_L$, and define the prefix mass after position j as

$$p_j = \sum_{k=1}^{j} m(a_k), \tag{3.1}$$

where $m(a_k)$ is the residue mass. If fragments corresponding to both p_j and p_{j+1} are observed, their difference provides direct evidence for a_{j+1}. If several consecutive cleavage sites are unobserved, the spectrum may constrain only a larger gap $p_{j+r} - p_j$. That gap can identify the total mass of a segment without uniquely identifying the order of its residues. Missing evidence is therefore not equivalent to evidence for a particular residue, nor is it necessarily evidence that the peptide is unidentifiable [40, 11].

Given a labeled peptide, its theoretical prefix masses define the cleavage positions that an ideal fragment series could reveal. A missing-fragmentation region is a consecutive peptide segment for which the corresponding internal prefix masses are not represented by matched nodes. A region of length two, for example, lacks the mass coordinate between two adjacent residues; the observed path crosses the combined mass of both residues in one step. Longer regions remove more intermediate coordinates and admit more possible residue combinations.

Two measures summarize severity. The number of missing cleavages is useful at the peptide level because one wrong residue makes an entire peptide prediction incorrect. At the residue level, GraphNovo defines the missing-fragmentation ratio

$$\text{MFR} = \frac{\text{number of missing fragmentation sites}}{\text{number of candidate peptide-bond fragmentation sites}}. \tag{3.2}$$

MFR permits comparisons among peptides of different lengths. It does not measure all aspects of spectrum quality: a spectrum may also contain noise, alternative ion types, or ambiguous peak assignments. Its purpose is to isolate the absence of sequential mass evidence.

An autoregressive neural decoder nevertheless has to emit one residue at every step. Its next decision is conditioned on the residues already emitted and often on their accumulated mass. A locally unsupported choice can therefore move the internal prefix mass away from the true peptide. Peaks that would have been informative under the correct prefix may then be queried in the wrong mass region, so the initial ambiguity affects decisions after direct fragmentation evidence returns. Missing fragmentation is thus a state-estimation problem as well as a local sequence-completion problem.

The spectrum still preserves the cumulative-mass relationship between fragments observed on either side of the gap. GraphNovo makes those surviving constraints explicit in a spectrum graph.

3.2 Spectrum Graph Representation

Let a centroided tandem mass spectrum be

$$S = \{(q_i, I_i)\}_{i=1}^{n}, \tag{3.3}$$

where q_i is a measured mass-to-charge ratio and I_i is its intensity. The precursor mass M supplies a global constraint: the neutral masses of the residues in a valid peptide, together with terminal groups, must sum to M. Fragment-ion hypotheses convert observed q_i values into candidate N-terminal residual masses. Peaks that may arise from complementary ion series can thus be expressed on one cumulative-mass axis. The spectrum-graph formulation and dynamic programming established mass-consistent paths as a basis for de novo sequence inference [18, 19].

GraphNovo constructs a directed acyclic graph $G = (V, E)$ on that axis. Two virtual nodes represent the beginning and end of the peptide. A nonterminal node represents a candidate cumulative residue mass derived from one or more observed peaks. An edge (u, v) is introduced when the positive mass difference

$$\Delta(u, v) = m(v) - m(u) \tag{3.4}$$

matches the mass of a feasible residue combination within instrument tolerance. A path from the beginning node to the end node is therefore a mass-conserving explanation of the precursor. If an edge mass matches one residue unambiguously, the edge can be translated directly. If it spans several possible residues, the edge is represented as a mass tag until sequence completion.

This representation distinguishes three objects that a direct sequence decoder tends to conflate. A *peak* is a measured event. A *node* is a hypothesis that one or more peaks mark a particular prefix mass. A *path* is a globally ordered set of those hypotheses consistent with the precursor. The path is intermediate structure: it is more informative than an unordered spectrum but less committed than a full peptide sequence.

Figure 3.1 connects these objects to the complete procedure. Panel (a) shows the selected mass path acting as a set of checkpoints between PathSearcher and SeqFiller, while panel (b) follows an observed peak through ion hypotheses, node construction, and the edge and path features supplied to the model.

Precursor mass provides a global conservation law but weak local supervision. Many residue strings can have the same total mass, particularly as peptide length increases. Fragment peaks make the constraint local by marking intermediate prefix or suffix masses. The spectrum graph combines these two forms of information: every full path must conserve the precursor mass, while each selected internal node asserts local fragment support.

This view explains why a graph is more than a convenient neural input format. If peaks were embedded as an unordered set and passed directly to a decoder, the model would have to learn mass conservation and reachability implicitly. In G, impossible orderings are excluded by edge direction, and residue-mass feasibility is encoded before learning. The neural network ranks structured alternatives rather than inventing the basic chemistry from data. Conversely, using a traditional graph score alone may underuse peak context and learned fragmentation regularities. GraphNovo assigns these complementary roles to the graph and the network: the graph defines physically feasible relations, while attention learns which feasible relations are supported in a particular spectrum.

An unresolved interval is consequently carried forward as an edge with a measured mass rather than converted immediately into a residue sequence. Immediate completion would force a locally preferred residue into the running prefix, while a generic unknown token would discard the quantitative distinction between, for example, a 128.09-Da gap and a 214.13-Da gap. A GraphNovo mass tag retains the supported coordinate before the gap, the supported coordinate after it, and their difference. It preserves the quantity supplied by the measurement and gives the decoder an explicit point at which to rejoin observed evidence.

The graph is constructed by representing each original peak under several fragment-ion hypotheses. For the richer *sequence spectrum*, GraphNovo considers 18 singly and doubly charged a, b, and y ion forms, including common neutral losses. Each hypothesis converts the measured m/z to an N-terminal residual-mass coordinate and carries peak-level descriptors such as normalized mass, relative and local intensity, global and local rank, and ion type.

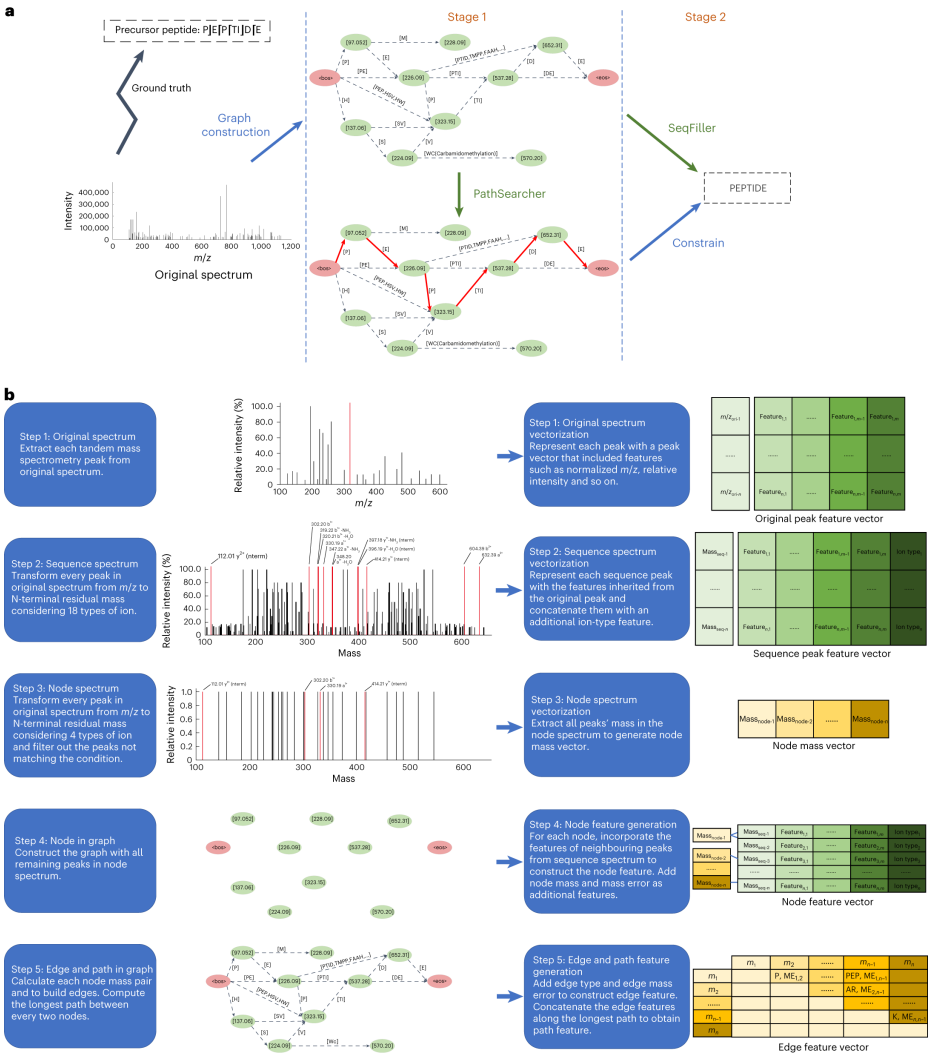


Figure 3.1: Overview of GraphNovo. (a) The complete GraphNovo workflow. GraphNovo-PathSearcher and GraphNovo-SeqFiller (green) are the two-component models. Red arrows emphasize the optimal path. Values in the green nodes are the N-terminal residual masses inferred from observed peak m/z values, and labels in square brackets are possible amino-acid sequences for the corresponding graph edges. (b) Step-by-step graph construction: spectrum-graph construction is shown on the left and generation of the input graph features on the right. One peak and its derivative peaks in the first three steps are highlighted in red. The indices n and m have different meanings across the steps. In step 5, m_i is the mass of node i; the edge type and mass error $\mathrm{ME}_{i,j}$ between nodes i and j form the edge feature vector (for example, PEP and $\mathrm{ME}_{1,n-1}$). Blank cells indicate that no edge connects the corresponding nodes. Reproduced from [11].

This expansion lets evidence from complementary ion series meet on a common coordinate system [11].

A smaller *node spectrum* is used to define graph vertices. To control graph size, node candidates are generated from b^+, y^+, a^+, and, for sufficiently charged precursors, y^{2+} hypotheses. The choice was informed by the empirical frequency and conditional co-occurrence of ion types in the training data. Candidates below 1,000 Da that cannot match any feasible peptide mass are removed, while larger candidates are retained if they fall below the precursor mass. Nearby sequence-spectrum peaks are then attached as features of each node rather than promoted to additional vertices.

Edges encode more than adjacency. For every ordered node pair, GraphNovo calculates the mass gap and records residue combinations of up to six amino acids that match within tolerance. Because several combinations can have the same or nearly the same mass, edge features include both candidate types and mass errors. Dynamic programming on the acyclic graph also computes a longest-path descriptor between node pairs. This descriptor tells the attention mechanism whether two nodes are connected by a plausible chain of intermediate residue masses, even when they are not directly adjacent.

The tolerance combines precursor and fragment measurement uncertainty. In the reported experiments, precursor tolerance was 5 ppm and fragment tolerance was 0.02 Da. These settings, together with the set of allowed residues and modifications, determine which graph edges exist. Consequently, graph construction is already a chemical prior: GraphNovo is database-independent at the protein-sequence level, but its original graph is not open to arbitrary modification masses.

The representation also supplies a graded notion of success. A model might recover the correct precursor-to-precursor-complement path while failing to name one internal residue combination, recover only a well-supported prefix and suffix, or select a wrong checkpoint that makes subsequent completion impossible. Path accuracy and final peptide accuracy can therefore be examined separately.

Within this graph, PathSearcher decides which cumulative-mass checkpoints are supported, whereas SeqFiller determines how the intervals between those checkpoints can be completed.

3.3 PathSearcher and SeqFiller

Both stages use the same high-level encoder design. A Graphormer-derived transformer represents nodes while injecting edge and path relations into attention [59, 60]. Node

features are aggregated from nearby sequence-spectrum hypotheses. Edge features and longest-path features are separately embedded and projected into the attention computation. A talking-heads transformation mixes information across attention heads to reduce the bottleneck created when each head communicates through a narrow scalar relation [61, 62].

The published architecture uses six encoder layers and 12 decoder layers. Its defining design choice is that the encoder receives a graph whose relations have physical meaning: node coordinates represent candidate prefix masses, direct edges represent feasible mass increments, and path features summarize intermediate connectivity. Attention is therefore biased toward relationships that a peptide explanation must respect.

The published paper calls the model-selected result the *optimal path*. Here and throughout this chapter, that term means the highest-scoring path selected by PathSearcher at inference time, not an oracle ground-truth path; ground-truth path nodes are introduced below only as evaluation targets. Figure 3.2 places this distinction in the architecture: panels (a) and (b) follow information from PathSearcher node indices to SeqFiller residue output, whereas panels (c)–(e) show the node, directed-edge and longest-path encoders, and relation-attention layer that represent the graph.

PathSearcher autoregressively selects graph nodes. Its decoder state includes the nodes already selected, and its next output is a distribution over candidate graph nodes. Training targets may assign probability to more than one node when multiple experimental peaks match the same ground-truth prefix mass. Kullback–Leibler divergence is therefore used to compare the predicted distribution with a target distribution that divides mass among equivalent matching nodes.

At inference time, the selected node sequence is sorted by its progression from the beginning to the precursor-complement end node. Consecutive differences are translated into residues when there is a unique one-residue match. Other differences become bracketed mass tags. For a conceptual peptide `PEPTIDE`, a missing pair of internal fragment coordinates might yield an intermediate output such as `PEP[214.13]DE`. This object is not an erroneous incomplete peptide. It is a statement that the path supports the surrounding prefix masses and the total missing mass, while the internal order remains unresolved.

SeqFiller uses a separate transformer decoder to replace mass tags with amino-acid sequences. It sees the spectrum graph and the PathSearcher output, but it is trained independently from scratch with residue-level cross-entropy. A knapsack dynamic-programming table enumerates residue combinations that can fill a remaining mass. The predicted path supplies several local mass constraints rather than only the full precursor mass, sharply reducing the completion space for each unresolved segment.

Beam search is modified so that hypotheses remain within the current path edge until

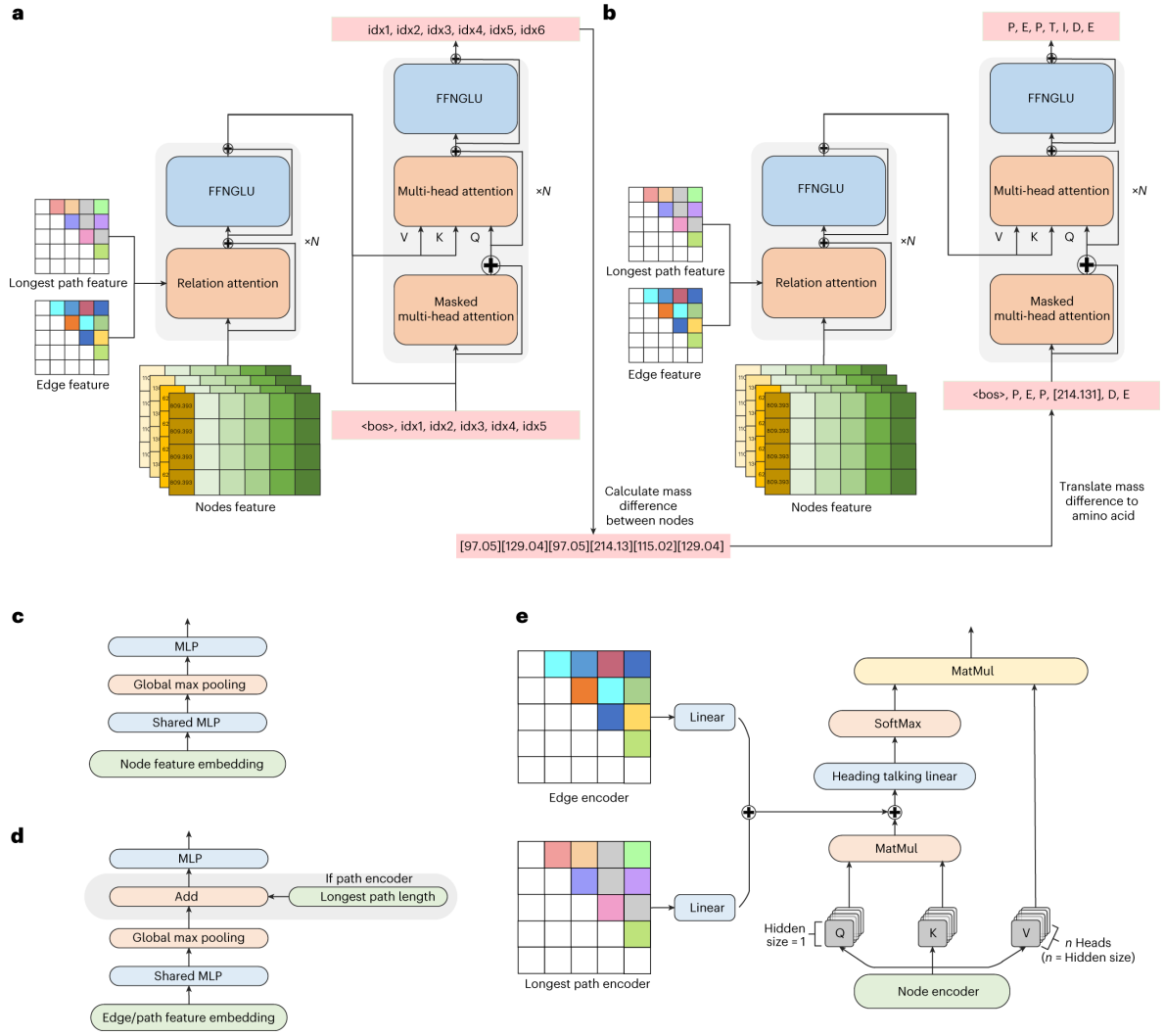


Figure 3.2: Model details of GraphNovo. (a) GraphNovo-PathSearcher generates a series of node indices (idx_1, idx_2, and so on) that define the optimal path; mass differences along that path are translated to amino acids when a unique match exists. (b) GraphNovo-SeqFiller takes the output of GraphNovo-PathSearcher and predicts an amino-acid sequence to replace each unresolved mass tag. Both models use an N-layer Transformer with attention and FFNGLU (a feed-forward network with a gated linear unit) modules; K, Q, and V denote key, query, and value vectors. (c) Node-encoder details; MLP denotes multilayer perceptron. (d) Directed-edge-encoder and longest-path-encoder details. (e) Relation-attention-layer details; $\oplus$ denotes element-wise addition. Reproduced from [11].

its mass has been filled. This matters because an ordinary beam search can discard a temporarily low-probability first residue even when that residue is necessary to complete a later gap. The reported beam size is 20. Once a tag is resolved, the next supported path coordinate resets the cumulative-mass state, preventing an error inside one gap from shifting the coordinate system for all subsequent regions.

This reset explains the intended mechanism. PathSearcher does not make missing fragmentation disappear, and SeqFiller may still choose the wrong internal sequence. The path instead provides a series of checkpoints. Between checkpoints, residue prediction is uncertain; at the next checkpoint, the model must satisfy the measured cumulative mass. The architecture converts unconstrained error propagation into locally bounded completion problems.

3.4 Training Data and Evaluation

The models were implemented in PyTorch and trained with teacher forcing and AdamW [63, 64]. Training used four 80-GB A100 GPUs and a fixed learning rate of 2×10^{-4}. The training set contained 410,366 database-search identifications from the human HeLa cell line and 1,249,397 from mouse cerebellum; 20,472 human-plasma identifications were used for validation. Peptides shorter than six residues or heavier than 4,500 Da were excluded; methionine oxidation and cysteine carbamidomethylation were the variable and fixed modifications, respectively. High-confidence search results were retained at a reported 1% peptide-level error threshold. Training graphs were limited to fewer than 512 nodes, although validation and test graphs were not subject to this cutoff [11].

The test sets contained spectra from *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Escherichia coli*. Peptide sequences shared with training data were removed, after which 12,500 spectra were initially sampled per species. Removing spectra for which any comparison method failed to return output left 12,222, 12,103, and 12,330 spectra, respectively. This intersection makes tool-to-tool comparison well defined, but it also means that the benchmark does not represent all acquired spectra. Source data, checkpoints, and code were released separately [65, 66].

Cross-species testing serves a specific purpose here. It weakens direct peptide-sequence overlap while preserving the same underlying residue chemistry and HCD measurement process. The experiment therefore probes whether the model learned transferable spectral structure, although it cannot by itself exclude every form of homologous subsequence leakage or instrument-specific adaptation.

The two-stage model creates an evaluation target that does not exist for an end-to-end sequence decoder. Let V^* be the nodes on the ground-truth prefix-mass path and $\widehat{V}$ the nodes selected by PathSearcher. Node recall and precision are

$$\text{NodeRecall} = \frac{|\widehat{V} \cap V^*|}{|V^*|}, \tag{3.5}$$

$$\text{NodePrecision} = \frac{|\widehat{V} \cap V^*|}{|\widehat{V}|}. \tag{3.6}$$

These metrics ask whether the mass path is correct before SeqFiller is considered. Final predictions are evaluated at peptide and residue levels using the established matching procedure employed by DeepNovo and PointNovo. A predicted residue is aligned using both residue-mass agreement and cumulative-mass agreement, allowing mass-spectrometric equivalences to be treated consistently [22, 23, 11].

A wrong final sequence can arise because PathSearcher omitted or added a mass coordinate, because SeqFiller chose the wrong residue combination inside a correct gap, or because the spectrum genuinely admits multiple explanations. Reporting only peptide recall merges these failure modes. The model instead exposes two quantities that can be evaluated separately: recovery of the mass path and recovery of the final peptide.

3.5 Sequencing Accuracy and Path Recovery

GraphNovo was compared with PEAKS, Novor, DeepNovo, PointNovo, and Casanovo on the three held-out species. The deep models were retrained on the common training set where required. GraphNovo obtained the highest peptide recall, residue recall, and residue precision across the reported species. The most informative absolute result is the variation across species rather than a single aggregate: peptide recall was 61.6%, 66.0%, and 71.2% for *A. thaliana*, *C. elegans*, and *E. coli*, respectively. These values are included to establish scale, not to imply that one benchmark number measures discovery performance in all settings [11].

Figure 3.3 presents the three levels together: panel (a) reports exact peptide recovery, while panels (b) and (c) separate amino-acid recall from precision. Their consistent ordering shows that the reported advantage is not confined to a single sequence-level metric.

PathSearcher recovered the intermediate representation with node recall between 93.2% and 96.3% and node precision between 92.6% and 95.9%. Its accuracy remained comparatively stable as peptide length and graph size increased, including graphs larger than the

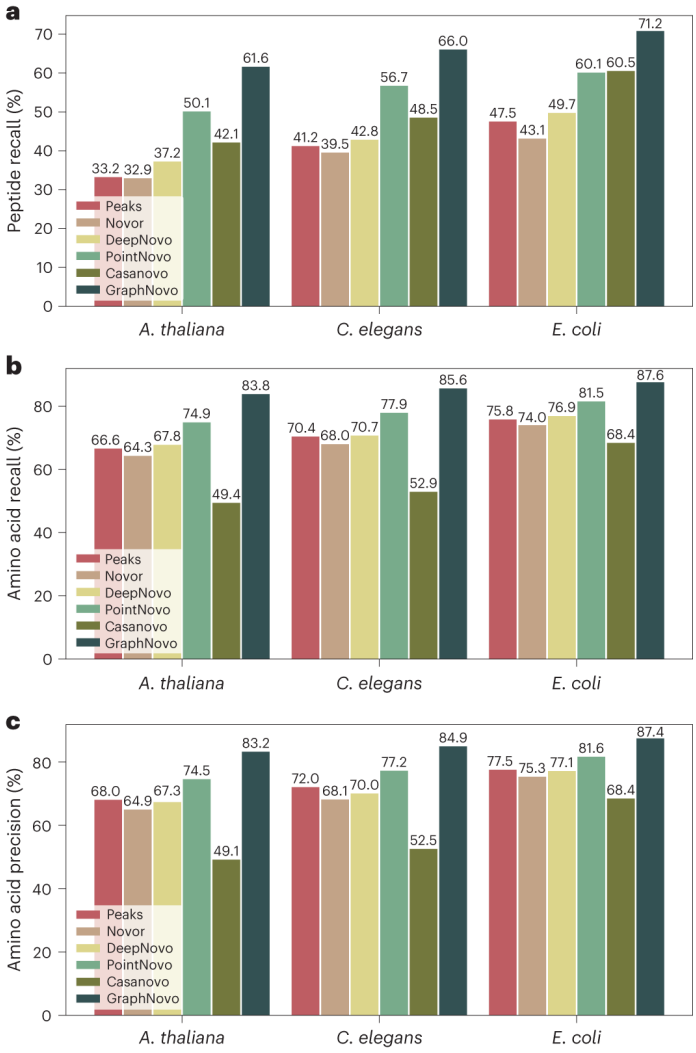


Figure 3.3: Overall evaluation. Performance of GraphNovo and the comparison algorithms on three held-out species, *A. thaliana*, *C. elegans*, and *E. coli*, measured by (a) peptide recall, (b) amino-acid recall, and (c) amino-acid precision. Reproduced from [11].

fewer-than-512-node training cutoff. These results matter because high final peptide recall alone would not show whether the proposed path abstraction is reliable. The node results establish that most evidence-supported prefix checkpoints are recovered before sequence completion.

Figure 3.4 links aggregate path recovery in panel (a) to robustness across peptide length in panel (b) and graph size in panel (c); the last comparison includes graphs beyond the cutoff used during training.

High path accuracy shows that the intermediate object is recoverable, but it does not establish that the path is responsible for the gain in final sequencing. The decoder-guidance experiment provides a more direct test.

3.6 Error Propagation across Missing Regions

The central mechanistic experiment asks whether the path itself helps, independent of GraphNovo's own residue decoder. The inference procedures of DeepNovo and PointNovo were modified to obey a PathSearcher-predicted path. A one-residue edge forces the corresponding residue; a mass-tag edge restricts the decoder through a local knapsack constraint. The underlying sequence models are otherwise retained.

Path guidance improved both models. Across the full test sets, the guided DeepNovo variants gained 9.1–12.9 percentage points in peptide recall, and the guided PointNovo variants gained 2.7–3.6 points. Gains at residue level were also observed. When evaluation was restricted to spectra containing missing fragmentation, the improvements became larger at residue level, and GraphNovo exceeded PointNovo in peptide recall by 12.4–15.0 points [11]. The exact ranges depend on species, but their direction supports the structural hypothesis: a mass path is useful even when paired with a decoder for which it was not originally designed.

This intervention distinguishes representation from capacity. If the advantage came only from a larger graph transformer, constraining DeepNovo or PointNovo with the predicted path should not systematically improve them. The result instead indicates that the path carries information that their original inference procedures fail to preserve.

The intervention establishes that the path can improve decoders for which it was not originally designed. Positional analysis then shows where that improvement occurs. In the combined test data, 55.06% of labeled peptides contained at least one missing-fragmentation region, and 19.23% of residues lay in such regions. Peptides were divided into the portion

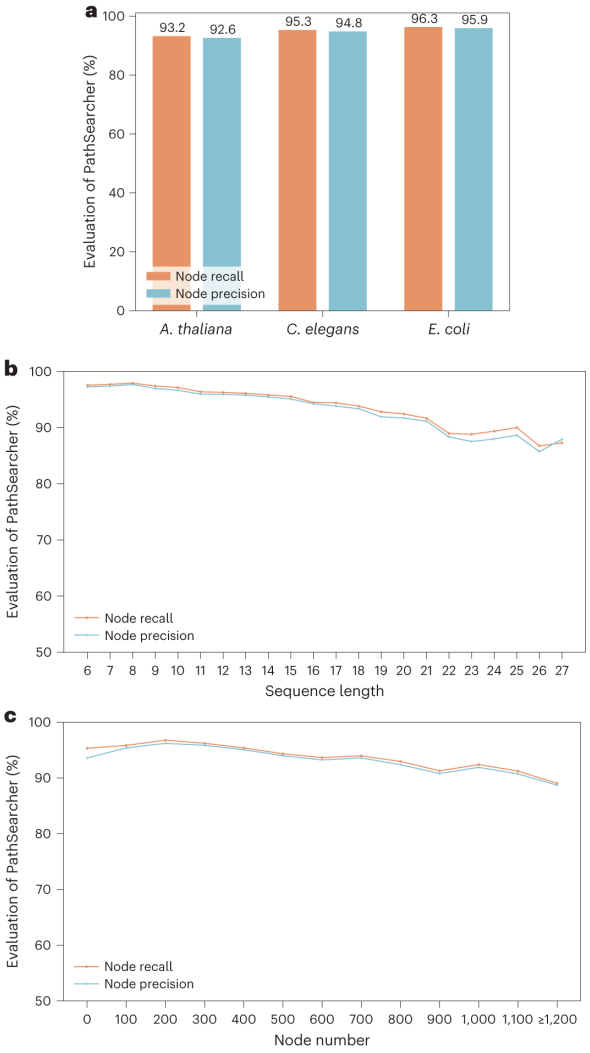


Figure 3.4: Evaluation of GraphNovo-PathSearcher. (a) Independent evaluation of node recall and node precision on *A. thaliana*, *C. elegans*, and *E. coli*. (b) Node recall and precision as a function of the sequence length of the labelled peptide. (c) Node recall and precision as a function of the number of nodes in the spectrum graph. Reproduced from [11].

before the first missing region, residues within all missing regions, and well-fragmented residues after the first missing region. This partition asks whether performance returns to its earlier level once direct fragment evidence resumes.

For the original DeepNovo and PointNovo decoders, it did not. Residue recall after the first missing region was 19.7 and 20.4 percentage points lower, respectively, than recall before it. Thus, even residues classified as well fragmented were harmed by an earlier unsupported decision. Adding PathSearcher guidance increased post-gap recall by 24.5 points for DeepNovo and 18.8 points for PointNovo. It also produced smaller improvements within missing regions by shrinking the knapsack search space [11].

Figure 3.5 organizes the evidence for this mechanism. Panels (a) and (b) show how common missing regions are and how their lengths are distributed, panel (c) defines the before/within/after partition on an example peptide, and panel (d) shows that path guidance restores much of the recall lost after a missing region.

Spectrum quality and peptide difficulty covary, so the experiment does not attribute every post-gap error to autoregressive state drift. The mechanistic interpretation instead rests on the agreement of three observations: post-gap performance drops despite renewed fragmentation, the path explicitly resets cumulative-mass constraints, and injecting that path into two different decoders restores much of the lost performance. Together, they provide stronger evidence than an end-to-end accuracy comparison alone.

3.7 Performance across Missing-Fragmentation and Spectrum Conditions

The regional analysis identifies the mechanism under typical missing-fragmentation conditions. The remaining experiments examine how far that mechanism holds as gaps become longer, spectra noisier, and peptides harder to sequence.

Performance declines as missing-fragmentation severity increases for every evaluated method. GraphNovo retains the best reported peptide and residue results across the observed range. Across the stratified data, 96.5% of spectra had MFR ≤ 0.5, including spectra with no missing-fragmentation sites. Within the nonzero subset $0 < \text{MFR} \leq 0.5$, GraphNovo achieved 82.6% residue recall. More important than that single value is the shape of the curves: a higher MFR systematically reduces the information available to all models, so the architecture mitigates rather than eliminates the measurement limitation [11].

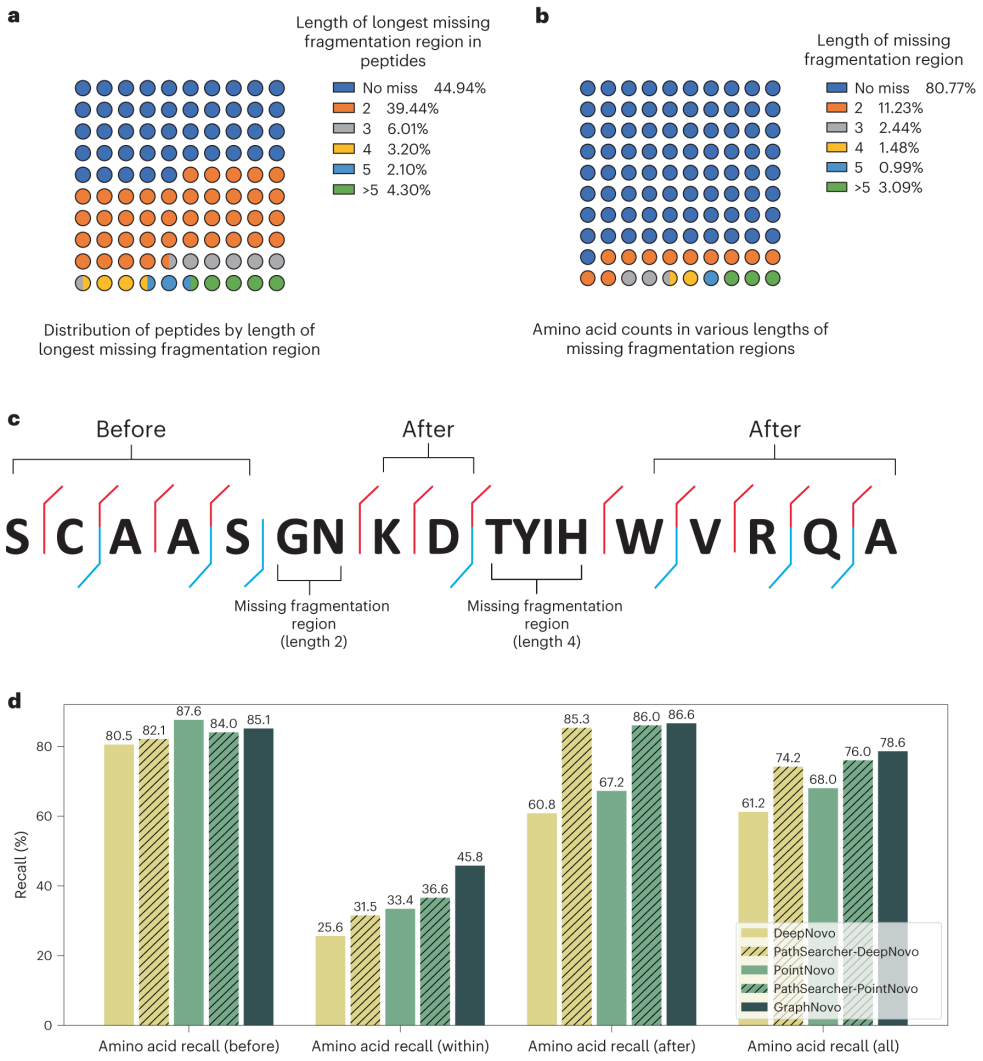


Figure 3.5: GraphNovo mitigates the missing-fragmentation problem. (a) Distribution of test peptides by the length of their longest missing-fragmentation region. For example, the peptide in panel (c) contains missing-fragmentation regions of lengths 2 and 4 and is therefore assigned a longest-region length of 4; "No miss" denotes peptides without missing fragmentation. (b) Distribution of amino acids by the length of the missing-fragmentation region containing them. (c) Example peptide containing two missing-fragmentation regions, together with the sequence segments defined as before, within, and after missing fragmentation. (d) Amino-acid recall before, within, and after missing-fragmentation regions, as well as over the complete peptide. PathSearcher-DeepNovo and PathSearcher-PointNovo are the modified DeepNovo and PointNovo models guided by the optimal path from GraphNovo-PathSearcher. Reproduced from [11].

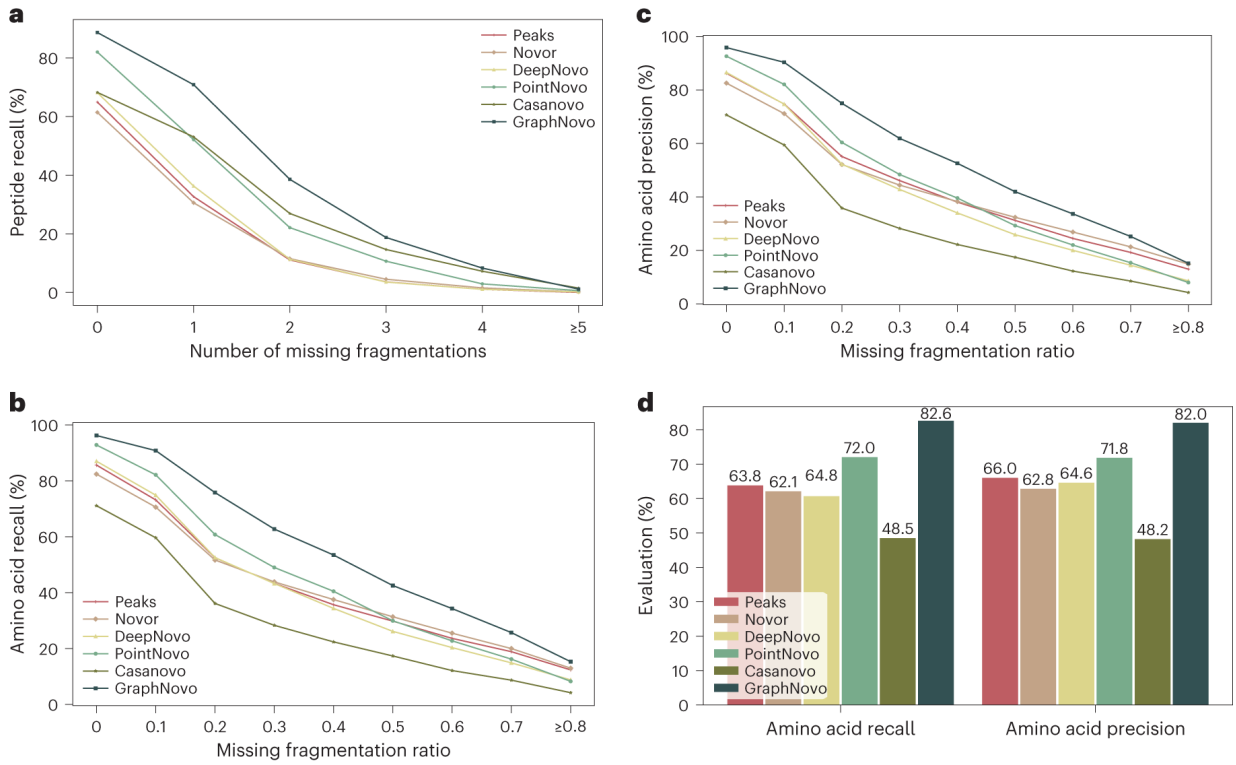


Figure 3.6: Effect of missing-fragmentation severity on model performance. Comparison of all evaluated models across (a) different numbers of missing-fragmentation sites in peptide recall, (b) different missing-fragmentation ratios in amino-acid recall, and (c) different missing-fragmentation ratios in amino-acid precision. (d) Amino-acid recall and precision for spectra with a missing-fragmentation ratio greater than 0 and no greater than 0.5. MFR denotes missing-fragmentation ratio. Reproduced from [11].

Figure 3.6 resolves this trend by endpoint: panel (a) relates peptide recall to the number of missing sites, panels (b) and (c) relate amino-acid recall and precision to MFR, and panel (d) summarizes the practically common nonzero range $0 < \text{MFR} \leq 0.5$.

Long gaps remain harder than short ones. Supplementary analysis shows larger degradation after regions longer than two residues, and the combinatorial explanation is straightforward: the number of residue combinations compatible with a mass gap grows rapidly with gap mass. A correct pair of path checkpoints can bound this search, but it cannot create the internal cleavage evidence needed to distinguish all isobaric or near-isobaric sequences.

Noise initially appeared to have a nonmonotonic relationship with accuracy: spectra with the lowest noise-to-signal ratio were not always easiest. The apparent paradox came

from confounding. Low-ratio spectra in the data often contained fewer signal peaks and more missing fragmentation. When comparisons were conditioned on missing-fragmentation level, increasing noise produced the expected decline. This analysis is methodologically important because a one-dimensional "spectrum quality" variable can hide distinct causes of difficulty. Missing evidence and distracting evidence should be analyzed separately [40, 11].

GraphNovo also declined with peptide length, as expected for an exact-sequence task, but the decline was smaller than for PointNovo and especially Casanovo in the reported stratification. An additional-training experiment added pig spectra to the training set. PathSearcher changed little and SeqFiller improved modestly, suggesting that the main reported gain did not result simply from adding more examples. This observation is not a general data-scaling law: it applies to the specific architecture, dataset, and augmentation used in the study.

Inspection of wrong paths revealed two recurring patterns. Because the training peptides were tryptic and usually terminated in lysine or arginine, atypical cleavage patterns could cause PathSearcher to prefer a terminal mass consistent with the training regularity instead of the observed peptide. Long or consecutive missing regions produced another class of errors: the model sometimes added an extra node, omitted a true node, or had too little evidence near the beginning of a peptide. In some such cases it still recovered later checkpoints, illustrating the same locality benefit without guaranteeing a correct complete path [11].

These failures show that an intermediate representation can itself encode dataset bias. A path is not raw measurement; it is a learned hypothesis over measurement. Its value is that the hypothesis is inspectable and evaluable at node level, not that it is immune to training distribution.

3.8 Discussion

GraphNovo shows that a missing stretch of fragment ions need not force a sequencer to lose its coordinate system. The model recovers supported cumulative-mass checkpoints before completing the residues between them. Node-level evaluation establishes that these checkpoints can be recovered accurately, while the decoder-intervention and before/within/after analyses show that they reduce downstream error in models that otherwise propagate an early mistake.

The checkpoints localize uncertainty rather than removing it. Long gaps admit many residue combinations, isobaric residues can remain indistinguishable by mass, and atypical

fragmentation can produce misleading node candidates. SeqFiller uses learned sequence regularity to rank compatible completions, but that regularity can reproduce the digestion bias seen in the wrong-path examples. Strong node precision therefore does not turn every mass tag into a uniquely determined local sequence.

Graph construction also defines a closed chemical vocabulary. Edges are created by matching mass gaps to combinations from an allowed residue and modification set. The original system includes the common fixed and variable modifications used during dataset construction, but a peptide containing an unmodeled modification can fail during graph construction before either neural stage is applied. Supporting an unrestricted collection of PTMs would expand the edge space substantially; the later RNovA chapter returns to this boundary by replacing predefined modified-residue identities with open mass differences.

The explicit graph carries a computational cost. Each peak generates multiple ion hypotheses, node pairs require relation features, and longest-path descriptors add memory. Training excluded graphs with 512 or more nodes even though evaluation examined larger graphs. The reported model was trained on high-resolution HCD data from bottom-up, tryptic workflows. Middle-down and top-down workflows interrogate substantially larger molecular units [67, 68], while data-independent acquisition (DIA) produces multiplexed fragment observations [69]; these and other fragmentation regimes remain proposed extensions rather than demonstrated applications.

The benchmark is likewise conditioned on database-labeled peptides and on spectra for which every comparison method returned a prediction. It measures sequence reconstruction under controlled labels, but it does not determine how many previously unidentified spectra can be reported at a calibrated false discovery rate. NovoBoard addresses that separate reliability question later in the dissertation.

GraphNovo ultimately assumes that the relevant molecular structure is a linear peptide backbone. A mass path is well suited to that object because cumulative mass orders the residues along one chain. The next chapter turns to glycopeptide spectra, where fragments can support partial compositions distributed across a branched glycan and no single linear path captures all of the useful evidence.

Chapter 4

DeepGlycan: *De Novo* Glycan Composition Prediction from Intact Glycopeptide Spectra

4.1 Introduction

Glycosylation covalently attaches carbohydrate groups to proteins and can affect protein folding and stability [70], cellular signalling [71], and immune recognition [72]. N-linked glycans attach to asparagine and commonly share a conserved core [73]. O-linked glycans attach to serine or threonine and exhibit a broader set of cores and contexts [74, 56]. Released-glycan analysis can characterize glycan populations, but it loses the direct association between a glycan, its peptide, and its protein site [55, 75]. Intact glycopeptide analysis retains that association by measuring the peptide and glycan together. Methods for released glycans can reconstruct candidate topologies or predict structures directly from glycan MS/MS, but they do not restore the lost peptide and site association [76, 77].

Retaining this context makes the spectrum more heterogeneous. Depending on the fragmentation method, an MS/MS scan can contain peptide fragments, low-mass oxonium ions, peptide-plus-partial-glycan fragments, and other glycan-derived ions [78, 79]. Database search engines manage this heterogeneity by scoring pre-enumerated peptide–glycan combinations [80, 81, 82]. Existing *de novo* methods use conserved cores, branch libraries, or dynamic programming over monosaccharide-sized mass differences [83, 84, 85]. These approaches provide important structural priors, but evidence that does not lie on the path or branch representation selected by the algorithm can be difficult to use.

Glycopeptide interpretation therefore presents a distinct representational problem. The peptide backbone is linear, whereas the attached glycan is branched. Evidence for its composition can be distributed among fragments associated with different branches, ion types, and fragmentation methods. A decoder may emit monosaccharides in a sequence for computational convenience, but that sequence is not the physical topology of the glycan. Useful fragments can be present yet remain disconnected under a single-path interpretation. DeepGlycan addresses this problem by exposing multiple mass relations to the model so that dispersed evidence can jointly constrain the final composition [12].

4.2 Prediction Task and Data

DeepGlycan accepts a preprocessed intact-glycopeptide spectrum together with precursor and peptide-mass information and produces a glycan composition. Its output vocabulary contains five monosaccharide classes used in the mammalian data: hexose (H), fucose (F), N-acetylhexosamine (N), N-acetylneuraminic acid (S), and N-glycolylneuraminic acid (G). These classes correspond to distinct residue masses and are generated autoregressively. A prediction such as $N4H5F1$ specifies counts; it does not specify which hexose is connected to which N-acetylhexosamine.

DeepGlycan operates over this fixed five-class vocabulary. It can generate new combinations of the supported monosaccharides, including compositions absent from the search library used by a downstream analysis. Reporting an additional monosaccharide chemistry would require extending the output space and obtaining suitable training support. The model's output is compositionally open within the five supported classes rather than chemically unrestricted.

The five classes also define what the decoder's symbols mean. H denotes a hexose, F a fucose, N an N-acetylhexosamine, S an N-acetylneuraminic acid, and G an N-glycolylneuraminic acid. These symbols identify residue classes by composition and mass; they do not distinguish every stereoisomer within a class. The decoder can therefore report that a composition contains four H units, but that output alone neither places the units in a branch nor specifies the linkage between them. This resolution is matched to the mass evidence used by the model.

Fragmentation method determines where that evidence appears. In stepped-collision-energy higher-energy collisional dissociation (sceHCD), collision energy is stepped during acquisition, allowing one scan to contain fragments formed under different collision strengths [86]. Glycan-related B/Y ions and oxonium ions can be abundant, while peptide fragments provide backbone context; the B/Y notation follows the standard nomenclature for

carbohydrate fragmentation [87]. Electron-transfer/ higher-energy collision dissociation (EThcD) combines electron transfer with supplemental collision energy. It can retain more of the glycan during peptide-backbone cleavage and provide c/z-type peptide ions useful for site localization [88, 89], although the electron-transfer reaction reduces scan throughput relative to a collision-only method [57, 58]. DeepGlycan does not assume that the two methods have the same intensity distribution. It asks whether both preserve enough mass relations for a shared representation.

The primary training resource was an EThcD collection of intact N-glycopeptide spectra from mouse, rat, and rabbit tissues. The original workflow searched these spectra with pGlyco 3.1 at 20-ppm precursor and fragment tolerances and retained glycopeptide-spectrum matches (GPSMs) passing the search engine's 1% filters at the peptide and glycan levels. The resulting N-glycopeptide set contained 478,483 GPSMs: 53.78% from mouse, 27.37% from rat, and 18.85% from rabbit samples. Data were partitioned 8:1:1 into training, validation, and test sets within species, with nonoverlapping glycan labels under the study's structure-based split key. O-glycopeptide spectra were not used for model training.

The EThcD spectra were produced from enriched glycopeptides in multiple tissues. The original experimental workflow reduced and alkylated extracted proteins, digested them with trypsin, enriched glycopeptides by anion-exchange solid-phase extraction, and analyzed them on an Orbitrap Ascend system in data-dependent mode. Enrichment choices are known to alter which intact N- and O-glycopeptides are observed [90, 91]. MS1 and MS2 resolving powers were 120,000 and 60,000, respectively; the isolation window width was 1 Th, and supplemental EThcD collision energy was 25%. These details matter because the learned peak distribution is tied to acquisition and preprocessing. Transfer to sceHCD changes the fragmentation distribution, but it does not test every instrument or enrichment protocol.

Raw files were converted to peak-list format before model input. Search parameters allowed up to three missed tryptic cleavages, fixed carbamidomethylation on cysteine, and variable methionine oxidation and N-terminal acetylation. Species-specific UniProtKB/Swiss-Prot databases and decoys supplied peptide candidates. The model therefore learns from spectra whose peptide mass and glycan composition have already passed several filters. Its training objective is not joint discovery of the peptide backbone, attachment site, and glycan from unconstrained raw data; it is composition inference given the intact-glycopeptide preprocessing context.

These labels make large-scale supervised learning possible, but they remain database-search-derived labels. A GPSM passing a nominal 1% threshold is a high-confidence reference under a specified peptide database, glycan database, score, and target–decoy procedure

[92]. It is not direct observation of every glycan bond. DeepGlycan is therefore evaluated against the composition assigned by the search workflow, not against an independently resolved topology.

Two output levels are reported. Glycan-level precision and recall require an exact composition match. Monosaccharide-level precision and recall compare the predicted and reference counts, so a near-complete composition can receive partial credit. The distinction parallels exact-peptide and amino-acid metrics in peptide *de novo* sequencing. It is particularly important here because a one-unit count error makes the whole composition incorrect even when most monosaccharides are recovered. Neither metric evaluates linkage or anomeric configuration.

More formally, let $\hat{\mathbf{c}}$ and $\mathbf{c}$ be predicted and reference count vectors over the five classes. Their shared monosaccharide count is

$$T_{\text{mono}} = \sum_{k} \min(\hat{c}_k, c_k). \tag{4.1}$$

Monosaccharide precision divides this count by $\sum_k \hat{c}_k$, and recall divides it by $\sum_k c_k$. Exact-composition correctness instead requires $\hat{\mathbf{c}} = \mathbf{c}$. Reporting both prevents a one-unit count error from being treated either as a completely unrelated prediction or as a fully correct composition.

The main generalization tests change one assumption at a time. An EThcD-trained N-glycan model is tested on held-out EThcD N-glycopeptides with nonoverlapping glycan labels under the study's split definition. The same learned representation is tested on mouse sceHCD N-glycopeptide spectra from an external study [93] after excluding glycan labels that overlap training under that definition. Finally, the EThcD N-glycan model is applied directly to EThcD O-glycopeptides without O-glycan training. The last experiment is termed zero-shot because O-glycopeptide spectra were excluded from training; the five monosaccharide actions and the spectrum-processing rules remain available.

4.3 DeepGlycan Architecture

The central design question is how to make fragment evidence from different ion types and partial glycan compositions comparable without first committing to one branch or one peak interpretation. DeepGlycan combines candidate-node expansion, relative-mass attention, and spectrum-conditioned composition decoding to organize this evidence in a shared coordinate system.

4.3.1 Peak-to-Node Representation

A raw spectrum represents each signal as an m/z position and an intensity. One peak, however, can admit several fragment interpretations. Charge may be uncertain, and the same observed position may be compatible with distinct ion types. Treating the peak as one anonymous token forces the model to infer these alternatives implicitly. DeepGlycan instead expands each peak into candidate nodes, following the general node-construction idea used in GraphNovo [11].

For EThcD spectra, a peak produces candidate glycan-retaining Y- and Z-type nodes over charge states below the precursor charge. For sceHCD spectra, the candidate expansion uses Y-type nodes. Every node retains the intensity rank of its source peak and records its candidate charge and ion semantics. The expansion does not assert that every node is correct. It enumerates plausible interpretations so that the encoder can compare them in a common neutral-mass space.

This expansion separates two uncertainties that a flat peak token would conflate. The first is whether the signal belongs to the glycopeptide at all; the second is which ion and charge interpretation makes it chemically comparable with other signals. Candidate nodes preserve both possibilities until contextual scoring. Nodes derived from one observed peak share its intensity evidence but differ in their implied neutral masses. Attention can favor the interpretation that participates in a coherent set of monosaccharide-sized differences while downweighting alternatives that do not.

Graph construction is deliberately less explicit than in GraphNovo. GraphNovo places edges between candidate peptide states and reasons about paths. The original DeepGlycan design explored predecessor information inspired by all-pairs graph representations, but the central model supplies node masses directly to attention rather than requiring one calculated glycan path. This choice is appropriate to a branched object: committing to one longest path before decoding could discard the very off-path fragments needed to reduce the remaining composition gap [59].

This distinction between observation and interpretation is visible in Fig. 4.1. The observed peak list is not replaced by a single manually chosen ion assignment. Instead, each peak contributes alternatives whose relative masses can be evaluated jointly. Noise peaks also create nodes, but their mass differences are less likely to participate consistently in monosaccharide-sized relations across the spectrum and the generated composition.

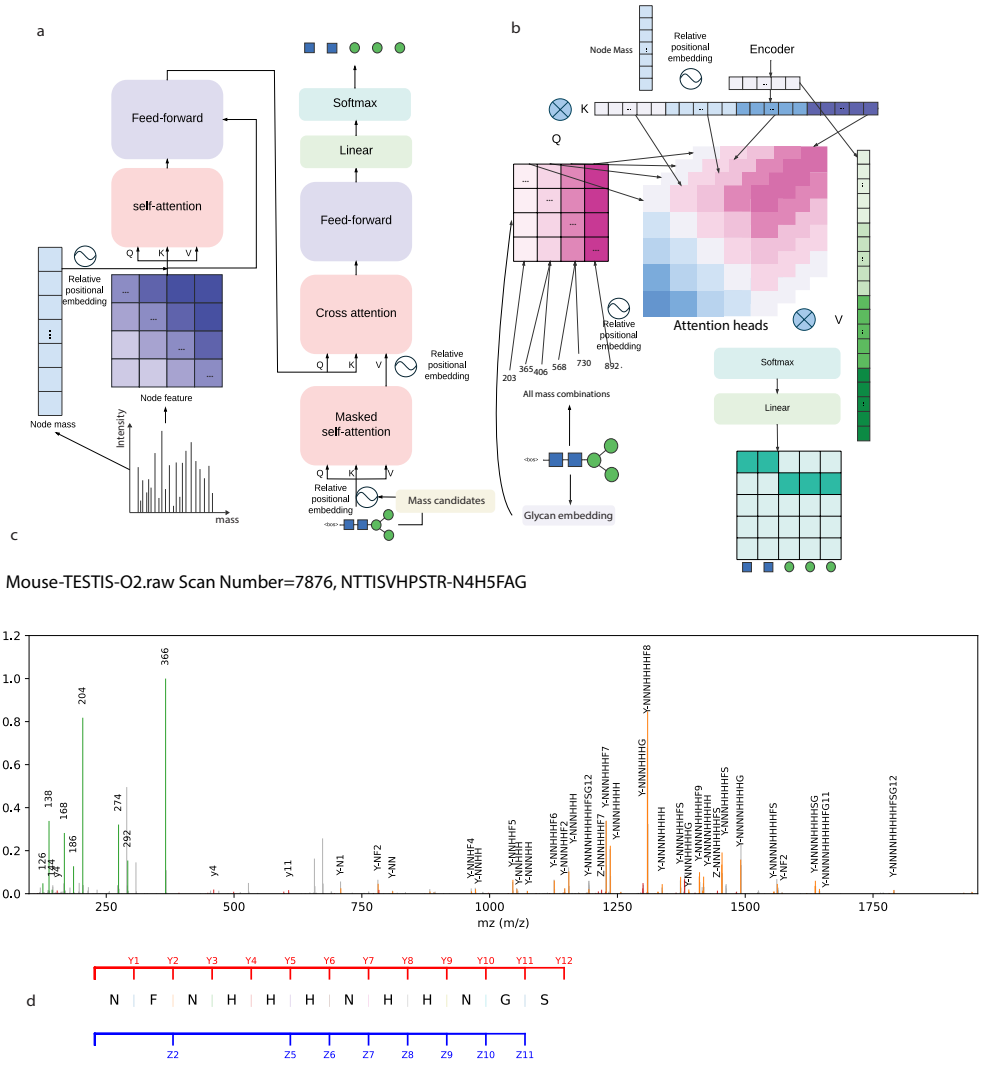


Figure 4.1: Workflow and model architecture of DeepGlycan. **a**, Preprocessed spectrum peaks are expanded into candidate fragment nodes following GraphNovo [11]. Relative-mass embeddings and node features enter the encoder, and an autoregressive decoder predicts one monosaccharide at a time. **b**, One decoder layer. Combination masses calculated from the current partial composition query the encoded spectral nodes through cross-attention to score the next monosaccharide. Blue squares and green circles denote N-acetylhexosamine (N) and hexose (H), respectively. **c**, Annotated intact-glycopeptide MS/MS spectrum showing oxonium ions, peptide fragments, and peptide-attached glycan Y- and Z-type ions. **d**, Spectrum-conditioned serialization of the reference composition in **c**. Monosaccharides are ordered by observed fragment evidence, so the same composition may have different target orders in different spectra. The order supervises composition decoding; it is not a unique glycan topology.

4.3.2 Relative-Mass Attention

Fragmentation methods alter which ions are abundant, but a hexose-sized mass difference remains a hexose-sized difference. DeepGlycan uses relative mass as the coordinate system shared by nodes, partial compositions, and output actions. Its encoder and decoder are based on the Transformer architecture [60], with six layers in each component. Rotary positional embeddings (RoPE) encode node masses so that the attention score depends on their difference rather than on two unrelated absolute positions [94].

For node representations x_m and x_n associated with masses m and n, the rotary construction gives an attention term of the form

$$q_m^\top k_n = x_m^\top W_q^\top \mathbf{R}_{n-m} W_k x_n, \tag{4.2}$$

where $\mathbf{R}_{n-m}$ is a rotation determined by the relative mass. The model can consequently learn that pairs separated by a supported monosaccharide mass are related even when their absolute m/z locations vary across precursors. Talking-head attention is used to exchange information across attention heads, while the relative-mass representation supplies the chemical scale on which those heads operate [61].

The encoder alone does not determine a glycan. It produces a contextual representation of alternative nodes. The decoder then generates a sequence of monosaccharide actions according to

$$P(m_i \mid m_0, \ldots, m_{i-1}, S), \tag{4.3}$$

where S is the encoded spectrum and $m_{<i}$ is the partial output. Generation terminates when the composition mass is compatible with the precursor constraint. The model was trained with cross-entropy loss for ten epochs and used beam search during inference. These engineering choices enable composition generation; the central representational choice lies in what the decoder is allowed to query at each step.

4.3.3 Spectrum-Conditioned Composition Decoding

If a glycan were a linear chain and every cleavage were observed, the decoder could look for a peak one monosaccharide beyond the last partial mass. A branched glycan does not provide that guarantee. After several monosaccharides have been predicted, fragments may

correspond to multiple subsets of that partial composition. Focusing only on the total accumulated mass would ignore those alternatives.

DeepGlycan explicitly computes the masses of all supported combinations within the current partial output. Suppose the decoder has generated two N units and three H units. In addition to the total mass, it can construct masses for combinations such as $N1H1$, $N1H2$, or $N2H1$, subject to the current multiset. These combination masses are embedded on the same relative-mass scale as the spectral nodes and passed to decoder cross-attention. Different attention heads evaluate different subsets, with pooling used to align the variable number of combinations to a fixed head structure.

The mechanism changes the conditional question posed to the model. Instead of asking only, "which monosaccharide follows the last emitted token?", it asks, "which candidate action makes one of the supported partial compositions consistent with additional spectrum nodes?" The output remains autoregressive, but the evidence consulted at a step is not restricted to a linear prefix. This is how a sequential decoder can represent nonlinear molecular evidence without claiming to decode a unique branch order.

Training labels are ordered to reinforce that interpretation. The database-derived composition is aligned to annotated fragment peaks, and monosaccharides are serialized according to the order in which supported subcompositions occur in the spectrum. When a next fragment is better explained by returning to an earlier partial composition, the ordering can traverse that earlier evidence rather than forcing one longest path. Monosaccharides not associated with a recognizable fragment are inserted by finding a compatible combination that fills the remaining mass. The same composition can therefore receive different valid serializations for different spectra. The target order is spectrum-conditioned supervision, not a canonical glycan topology.

Accumulated and remaining masses provide two additional constraints. The mass of the current partial composition tells the decoder how much of the precursor has already been explained, while precursor mass minus that value bounds what remains. All internal combination masses supplement these two endpoints. The resulting input at one step contains a set of chemically interpretable coordinates rather than only a token prefix. The relevant structure is the family of mass relations supported by a partial multiset rather than an asserted branch diagram.

4.4 Dispersed Fragment Evidence and the Glyco-Gap

The combination-mass mechanism was designed to recover evidence that appears disconnected when a spectrum is interpreted as a single path. Comparing the raw glyco-gap with the gap remaining after supported partial compositions are combined reveals how much apparently missing information can be recovered from fragments already present elsewhere in the spectrum.

The consequence of combination-mass decoding is most clearly seen in the sceHCD analysis. A *glyco-gap* measures the number of monosaccharide steps between interpretable glycan Y-ion evidence. The final glyco-gap lies between the largest interpretable glycan-retaining fragment and the intact precursor composition. In the test set, 95.5% of spectra placed their largest glyco-gap at this final position. An algorithm that starts at the precursor or follows only one fragment path must resolve the entire end gap at once.

Some of that apparent gap is not an absence of evidence. It is evidence located on another branch or partial composition. The example in Fig. 4.2e contains an observed fragment supporting $N3H4$ and earlier evidence supporting F. Combining them produces a computational mass state for $N3H4F1$. This state is called virtual because it is derived from observed pieces; it is not a newly measured peak. The *filled gap* is the composition still required to reach the precursor after all such supported combination states have been considered.

Panels f and g of Fig. 4.2 compare the raw end gap with the filled gap across the data. Only 17.9% of GPSMs retain the same gap after supported virtual states are incorporated. Many spectra with a raw gap larger than three monosaccharides have a filled gap smaller than two. Calling all of those spectra low quality would conflate an algorithm's path representation with the information content of the measurement. Their evidence is incomplete as a single ladder but informative as a set of nonlinear relations.

The histogram makes the scale of the representational change concrete. A reduction of one monosaccharide accounts for 60.9% of spectra, reductions of two and three account for 11.0% and 5.2%, and smaller groups receive still larger reductions. The common case is therefore not that every large gap disappears. It is that one or more already-supported units can be carried forward from a different partial composition, turning a difficult suffix into a shorter conditional decision. A one-unit reduction matters because the number of compatible multisets grows rapidly with every unresolved action.

The filled gap also exposes the remaining limitation. When it remains larger than three, DeepGlycan performance declines. The model cannot manufacture fragments that are absent from every compatible subcomposition. Its advantage is therefore not an unlimited

ability to guess long glycan suffixes. It reduces the unresolved problem by using evidence that a single-path formulation leaves disconnected. This mechanistic distinction is more informative than an aggregate gain in recall.

4.5 Evaluation Across Fragmentation Methods and Glycan Classes

The evaluation asks whether a relative-mass representation learned from EThcD spectra transfers to sceHCD spectra and to glycan compositions excluded from the training split. The sceHCD benchmark compares DeepGlycan with GlycoNovo, StrucGP, and pGlycoNovo on the same mouse tissue datasets; the EThcD experiments evaluate held-out N-glycopeptides and direct transfer from N-linked to O-linked glycopeptides.

4.5.1 sceHCD N-Glycopeptide Benchmark

The sceHCD comparison used pGlyco 3.1 GPSMs passing a nominal 1% peptide- and glycan-level FDR as reference labels. Glycan labels overlapping the training set under the study's structure-based split definition were excluded from the test data. StrucGP, GlycoNovo, and pGlycoNovo do not require neural training, so they were run directly on the same spectra. The comparison therefore tests whether an EThcD-trained relative-mass representation transfers to sceHCD and to held-out glycan labels under that split, not whether every algorithm has identical prior information.

Across the mouse heart, kidney, liver, and lung datasets, DeepGlycan reported an average exact-composition recall of 96.80% and precision of 98.53%. The reported average recalls for pGlycoNovo, StrucGP, and GlycoNovo were 64.30%, 67.06%, and 83.49%, respectively. In kidney, for example, DeepGlycan recall was 96.6%, compared with 62.74% for pGlycoNovo, 61.56% for StrucGP, and 79.27% for GlycoNovo. Precision was high for all methods; the principal separation was the number of reference compositions recovered [12].

Composition size helps interpret this separation. When only two monosaccharides beyond the N-linked core had to be predicted, DeepGlycan, GlycoNovo, and pGlycoNovo all had high recall. As the number of monosaccharides increased, recall for the comparison methods declined more sharply, while DeepGlycan retained an overall sensitivity of 95.89%. The raw-gap/filled-gap analysis links that result to a specific source of information: larger glycans create more opportunities for one path to miss useful branch-derived fragments, while combination masses allow those fragments to constrain later actions.

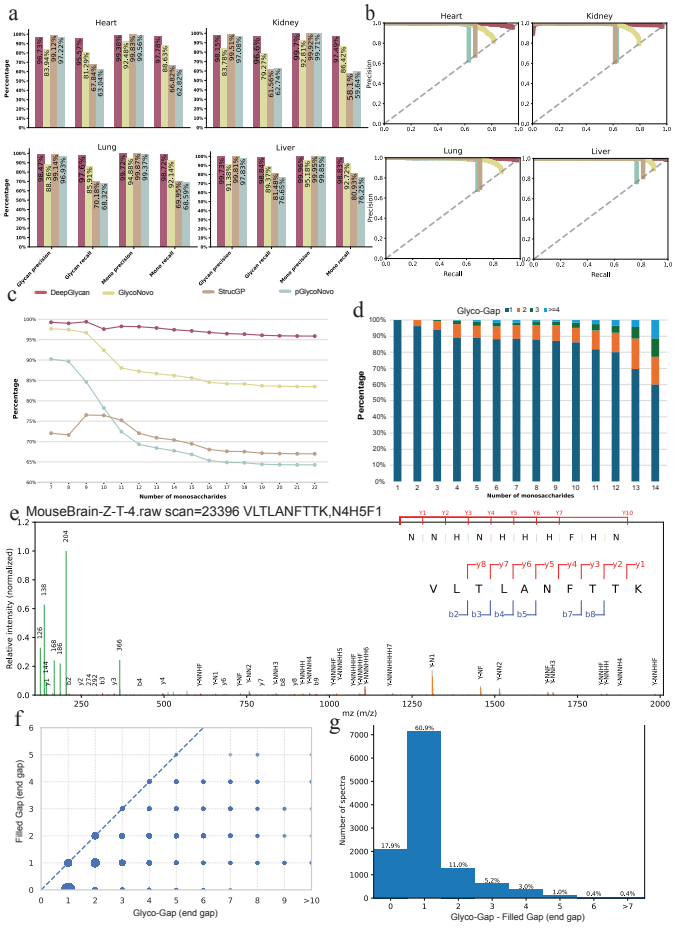


Figure 4.2: Performance on mouse N-linked glycopeptide sceHCD data and glyco-gap analysis. **a**, Glycan-composition and monosaccharide-level precision and recall for DeepGlycan, GlycoNovo, StrucGP, and pGlycoNovo in heart, kidney, lung, and liver. Complete composition matches define glycan correctness; values on bars are percentages. **b**, Precision–recall curves for the same methods and tissues; colours follow **a**. **c**, Glycan-composition recall versus the number of monosaccharides, pooled across tissues. **d**, Distribution of glyco-gaps of one, two, three, or at least four monosaccharide steps by glycan Y-ion size. A glyco-gap is the number of steps between adjacent interpretable Y-ion states. **e**, Annotated spectrum for VLTLANFTTK-$N4H5F1$. DeepGlycan agrees with the pGlyco 3.1 composition, whereas StrucGP and pGlycoNovo report no identification. **f**, Density-weighted comparison of the end glyco-gap (horizontal axis) and filled gap (vertical axis). The filled gap is the unresolved remainder after introducing virtual states supported by annotated partial compositions. Point area represents spectrum count, values > 10 are grouped, and the dashed line denotes equal gaps. **g**, Distribution of gap reduction, defined as end glyco-gap minus filled gap; percentages are printed above bars. Panels **a**–**d** show aggregate test-set proportions without replicate error bars. GPSM, glycopeptide-spectrum match; sceHCD, stepped collision energy higher-energy collisional dissociation.

The benchmark does not prove that DeepGlycan knows the correct topology. Exact agreement is computed on composition labels obtained by a database-search pipeline. Nor does the experiment isolate every difference among algorithms: StrucGP uses branch libraries, GlycoNovo uses dynamic programming, and pGlycoNovo uses a glycan-first search with a rainbow-table representation. What the experiment and gap analysis jointly support is narrower: under the specified labels and spectra, access to multiple partial-composition relations recovers compositions that a single-path end-gap formulation often misses.

The comparison methods clarify why this is a representation result rather than a generic deep-learning comparison. StrucGP uses conserved cores and predefined branch modules to construct N-glycan structures when the expected Y ions are available. GlycoNovo searches for monosaccharide mass differences with dynamic programming beginning from peptide-mass evidence. The pGlycoNovo method constructs a broad glycan table and follows a glycan-first strategy from the precursor. These priors are useful in their intended regimes. The observed recall decline on larger compositions indicates that the route by which dispersed evidence is assembled, not only the set of allowed monosaccharides, becomes limiting.

Precision–recall curves provide a complementary view to one operating point. The model score can be thresholded to trade coverage for precision, and DeepGlycan occupies a larger region of the curve on the four mouse tissues in the original analysis. This does not by itself make the score an FDR. A precision–recall curve uses labeled test spectra and reports retrospective correctness; an FDR procedure must estimate error for accepted predictions in a sample where the labels are not known.

4.5.2 EThcD Evaluation and N-to-O Transfer

EThcD and sceHCD emphasize different fragment families. EThcD combines electron-transfer and collision-based fragmentation and can preserve peptide backbone and glycosylation-site evidence while producing glycan-related ions [58]. sceHCD changes collision energy within the acquisition and often provides rich collision-derived glycan fragments. A model that memorized method-specific peak templates would be expected to degrade strongly when moved between them. DeepGlycan instead bases its main interactions on candidate ion semantics and relative mass.

The held-out EThcD N-glycopeptide results in Fig. 4.3a show exact-composition precision/recall of 98.35%/96.97% in mouse, 92.50%/90.85% in rat, and 91.82%/88.38% in rabbit. Monosaccharide-level metrics are higher, indicating that many composition-level errors are small count differences rather than unrelated predictions. Together with the

sceHCD result, these experiments show transfer across tissues, species, and a change in fragmentation mode, subject to the shared five-class vocabulary and preprocessing.

The more demanding test applies the N-trained EThcD model to O-glycopeptides without further training. Exact-composition precision/recall is 98.87%/95.79% in mouse, 93.05%/80.40% in rat, and 92.97%/79.18% in rabbit, according to the original study evaluation. The lower rat and rabbit recall shows that transfer is not uniform, but the high precision and partial-composition recovery indicate that the model can use conserved mass differences even when the N-linked core prior no longer describes the target class.

Panel c of Fig. 4.3 provides the mechanistic comparison. An N-glycopeptide and an O-glycopeptide with the same composition have visibly different peaks and intensities, yet selected peak pairs are separated by the same Hex-, HexNAc-, and Fuc-sized differences. Those differences are the invariant on which relative-mass attention operates. The result is thus a test of representational transfer, not evidence that N- and O-glycans have the same fragmentation template.

This transfer uses the same five monosaccharide classes available during training; the unseen element is the O-linked glycan class and its spectra, not the output alphabet. The reference O-GPSMs also come from a database-search procedure and inherit its candidate and FDR assumptions. Accordingly, the experiment establishes zero-shot composition prediction under the shared vocabulary. It does not assign a unique O-glycan topology or, in the absence of site-determining peptide fragments, a unique attachment site.

The original study also examined candidates not reported by several tools under their default settings. In one mouse-heart analysis, a precursor-mass-shift decoy procedure was used at a nominal 1% threshold, after which DeepGlycan reported 127 unique compositions compared with 86 from pGlycoNovo. An O-glycopeptide with backbone IRTTTSGVPR and composition $N3H3S2$ was supported by eight additional spectra in a spectral cluster [95], retention times between 2082 and 2153 seconds, an annotated glycoprotein context, and a related $N3H3S1$ composition. These checks make the candidate more testable than a single high model score, although they do not provide a prospective validation rate for all unique calls.

4.6 Biological Applications

The benchmark experiments measure recovery of reference compositions. We next examined whether the additional predictions preserve interpretable biological variation in a controlled glycosylation system and whether they can expand a downstream glycoproteomics analysis

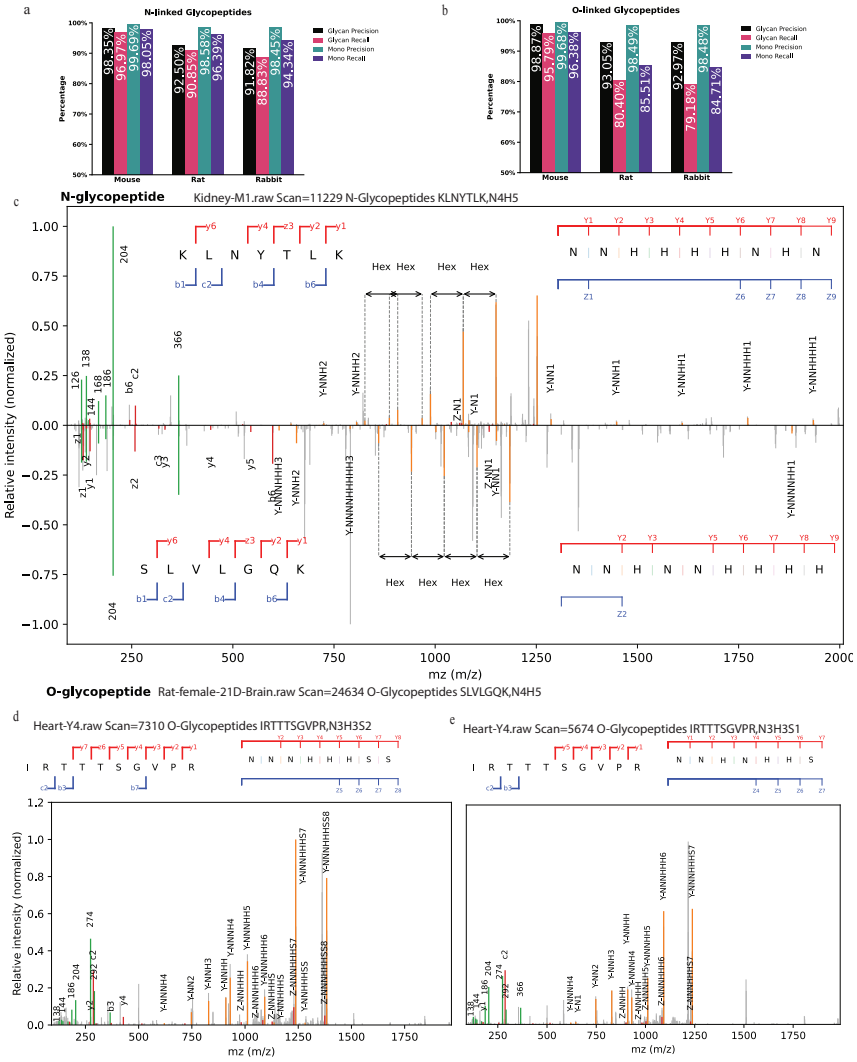


Figure 4.3: Performance on EThcD N-linked and O-linked glycopeptide data. **a**, Glycan-composition and monosaccharide-level precision and recall on held-out N-linked glycopeptides from mouse, rat, and rabbit; training and test glycans are disjoint under the study's split. **b**, The same metrics on O-linked glycopeptides. The N-glycan model was applied without O-glycopeptide training; bars show aggregate percentages without replicate error bars. **c**, Mirrored spectra of an N-glycopeptide (top; KLNYTLK-$N4H5$) and an O-glycopeptide (bottom; SLVLGQK-$N4H5$). Highlighted peak pairs retain hexose-sized mass differences despite different fragmentation patterns. **d** and **e**, Mouse-heart O-glycopeptide candidates IRTTTSGVPR-$N3H3S2$ and IRTTTSGVPR-$N3H3S1$, respectively, with annotated peptide and glycan-retaining ions. N, N-acetylhexosamine; H, hexose; S, N-acetylneuraminic acid; EThcD, electron-transfer/higher-energy collision dissociation. Predictions are compositions, not unique glycan topologies.

in a disease cohort. Site-specific intact-glycopeptide measurements have likewise supported quantitative disease comparisons while retaining protein and site context [96].

4.6.1 Engineered Glycosylation Systems

High benchmark recall is useful only if the additional compositions preserve biological signal. DeepGlycan was therefore applied to Fc gamma receptor data from glycosylation-engineered systems [97]. These experiments perturb enzymes with known roles and ask whether the predicted composition profiles change in the expected direction. The reported fucosylated, sialylated, and high-mannose categories are not mutually exclusive: one composition can carry more than one of these properties, so their percentages are not expected to sum to 100%. Because the original analysis used an empirical glycan-score cutoff greater than 5 rather than a calibrated *de novo* FDR, the results should be read as biological consistency tests, not as a universal error-rate validation.

In wild type, DeepGlycan reported 1,129 GPSMs. Fucosylated compositions accounted for 64.3%, sialylated compositions for 26.22%, and high-mannose compositions for 27.81%. In the ΔFUT8 sample, 603 GPSMs were reported; fucosylation fell to 28.36%, while the high-mannose fraction increased to 49.75%. FUT8 is the mammalian core fucosyltransferase, so the direction of the change agrees with the engineered perturbation [98]. Diagnostic core-fucosylated fragments were present for 89–100% of the seven most abundant fucosylated compositions in the wild-type analysis, adding fragment-level support to the composition summary.

The ΔMGAT1 condition shifted the profile toward oligomannose, which made up 56.11% of GPSMs, with Man5 accounting for 16.3%. The ΔMGAT2/3 + B4GALT1 condition had a different profile, including 47.99% sialylated and 67.9% fucosylated GPSMs. Figure 4.4 also connects Man5 to the peptide backbones on which it was observed. That peptide-level association is the information lost when glycans are released before analysis.

DeepGlycan reports intact-glycopeptide compositions whose aggregate shifts can be compared with the controlled biological intervention and released-glycan expectations. Agreement with the known perturbations shows that expanded composition coverage preserves the expected glycosylation signal while retaining the peptide context of each prediction.

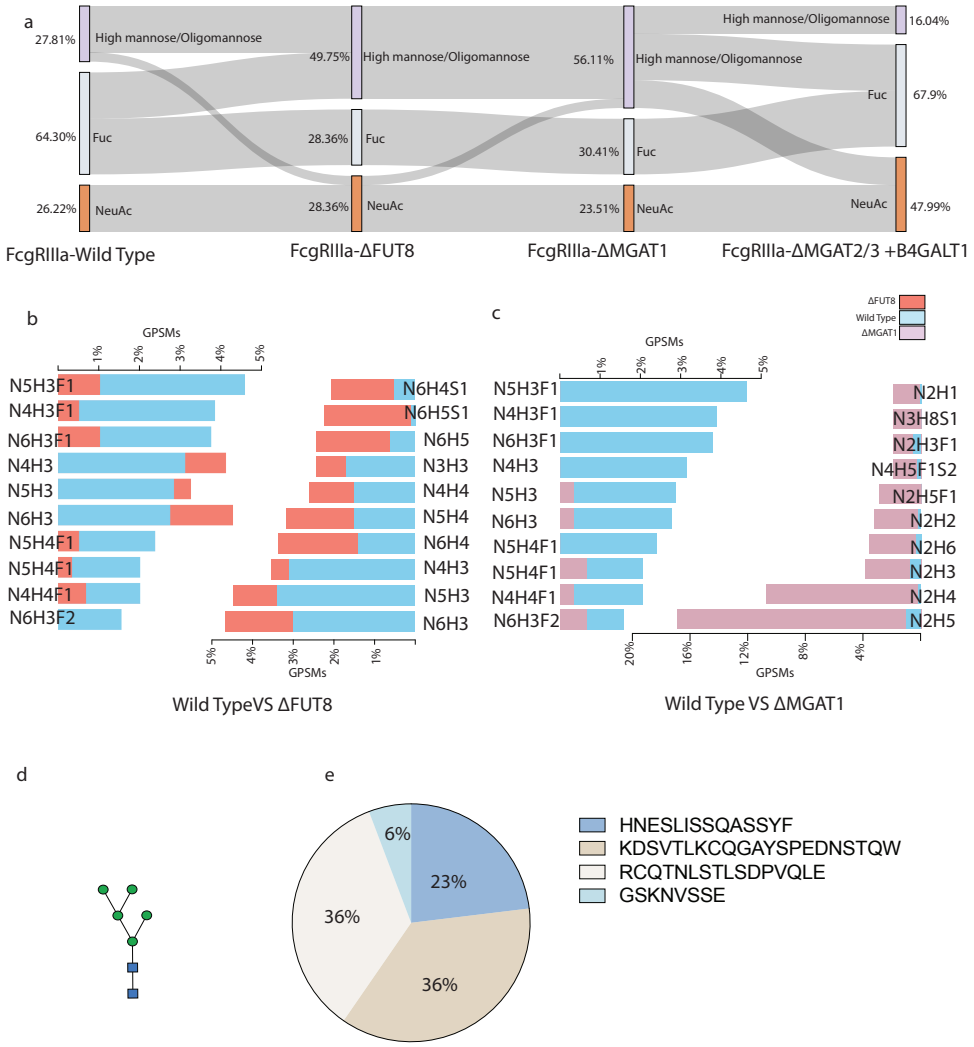


Figure 4.4: Intact glycopeptide profiles of wild-type and glycosylation-engineered FcγRIIIa samples. a, Fractions of GPSMs assigned to high-mannose/oligomannose, fucosylated (Fuc), and sialylated (NeuAc) compositions in wild type, ΔFUT8, ΔMGAT1, and ΔMGAT2/3 + B4GALT1. Categories are not mutually exclusive. **b**, Bidirectional comparison of the ten most abundant compositions in wild type (blue) and ΔFUT8 (red): wild-type-ranked compositions are shown on the left and treatment-ranked compositions on the right. **c**, Corresponding comparison of wild type (blue) and ΔMGAT1 (mauve). **d**, Man5 in Symbol Nomenclature for Glycans [99]; blue squares denote N-acetylhexosamine and green circles mannose. **e**, Distribution of Man5 GPSMs among the four labelled peptide backbones in the ΔMGAT2/3 + B4GALT1 sample. Panels show GPSM fractions without replicate error bars. GPSM, glycopeptide-spectrum match. DeepGlycan predicts composition rather than linkage topology.

4.6.2 Expansion of an Alzheimer's Disease Glycoproteomics Analysis

A second application used DeepGlycan to expand the glycan search space for an Alzheimer's disease cohort [100]. Previously reported peptide backbones were rematched to spectra, retaining spectra with fragment support for at least half of the peptide residues. Subtracting peptide mass from precursor mass produced pseudo-glycan masses; the 90 most frequent masses were advanced to *de novo* composition prediction. DeepGlycan produced 157 unique compositions, of which 65 were added to the glycan list used for downstream FragPipe label-free quantification [101].

The workflow used *de novo* inference to propose an expanded, finite search space and then returned to a matched quantitative analysis rather than using raw model scores as quantitative measurements. Figure 4.5 summarizes the resulting hypotheses. A composition labeled $N4H3F2$ was higher in the Alzheimer's group at adjusted $p = 0.017$ with a reported fold change of 3.662. One illustrated $N4H3F2$ GPSM was observed in seven of eight Alzheimer's samples and two controls.

The analysis then grouped 267 observed compositions using weighted correlation network analysis [102]. Ten co-regulated glycan modules were obtained, and their relationships were summarized by eigenglycan profiles [103]. $N4H4F4$, a composition reported uniquely by DeepGlycan in this workflow, occupied the hub position of module GM5. A network hub and a case–control association are hypothesis-generating results. They nominate compositions for targeted validation; they do not establish that the glycan causes disease or that its unique topology is known.

The exploratory statistics also have a compositional-data limitation. Glycoforms were treated as separate variables even though their abundances can be coupled by shared biosynthesis and finite total signal. More specialized compositional models could change the estimated associations [105]. The current result demonstrates how a *de novo* predictor can enlarge a downstream analysis; it is not an independent clinical replication.

4.7 Discussion

The results identify a specific source of DeepGlycan's gain. A large glyco-gap does not always indicate that the spectrum lacks glycan evidence; in many spectra, the relevant fragments are distributed across partial compositions that are not adjacent on one decoding path. Combination-mass attention makes those relations available, and the raw-gap/filled-gap

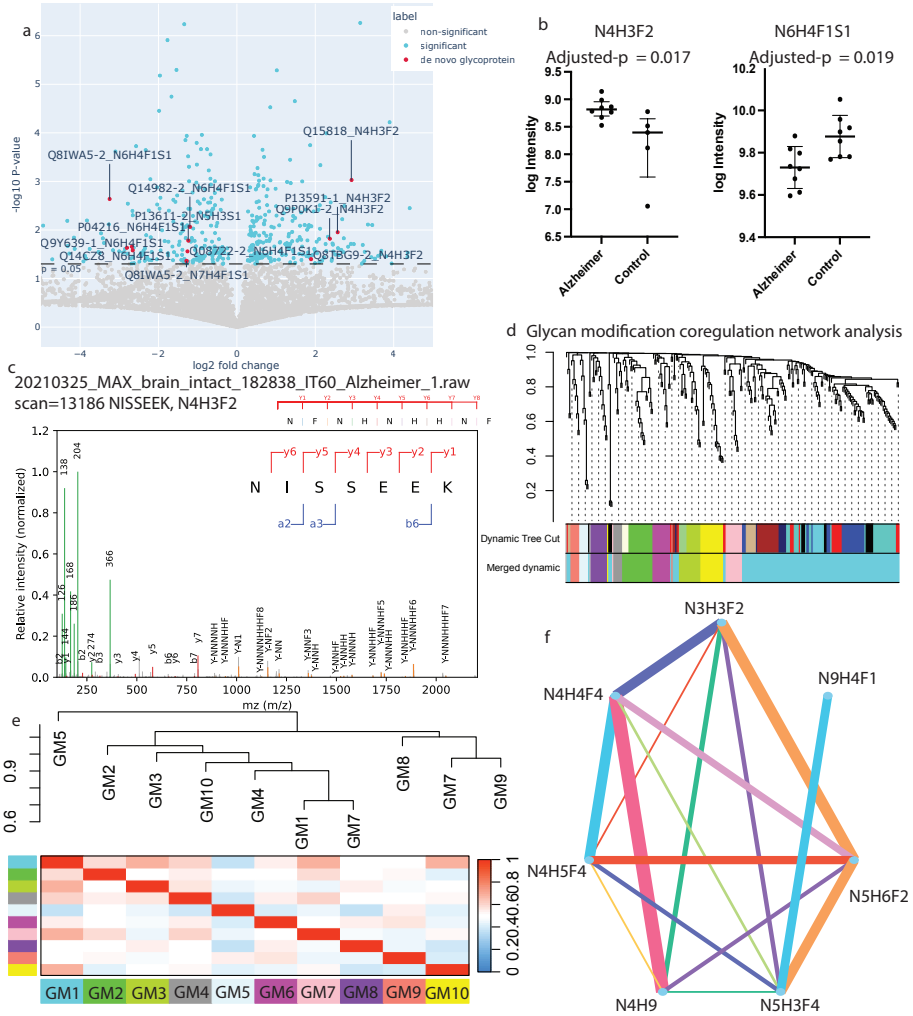


Figure 4.5: DeepGlycan-expanded glycoproteomic analysis of an Alzheimer's disease cohort. **a**, Volcano plot of glycoprotein abundance changes in Alzheimer's disease relative to control. Axes show $\log_2$ fold change and $-\log_{10} P$; grey points are nonsignificant, cyan points pass $P < 0.05$, and red labelled points contain compositions added by the DeepGlycan-expanded search. **b**, Sample log-intensities for $N4H3F2$ and $N6H4F1S1$. Points are individual samples. Adjusted P values use a two-sided unpaired Student's t-test with Benjamini–Hochberg correction [104]. **c**, Annotated NISSEEK-$N4H3F2$ GPSM observed in seven of eight Alzheimer's disease samples and two controls. **d**, WGCNA dendrogram of 267 glycan compositions; dynamic tree cutting and merging yield ten glycan modules. **e**, Adjacency heat map of module eigenglycans; red and blue indicate stronger and weaker inter-module connectivity. **f**, GM5 intramodular network. Edge width represents WGCNA connection strength, and the DeepGlycan-contributed composition $N4H4F4$ is the hub. GPSM, glycopeptide-spectrum match; GM, glycan module; WGCNA, weighted gene co-expression network analysis. Glycan labels denote compositions, not unique topologies or causal effects.

analysis links this representation directly to improved recovery of larger compositions. When compatible fragments are absent from every partial composition, however, no representation can recover them.

The resolution of the result is glycan composition. DeepGlycan predicts one count vector over five modeled monosaccharide classes and uses precursor and peptide-mass information supplied by preprocessing. It does not fully address peptides carrying multiple glycans, nor does exact-composition accuracy imply linkage, anomeric configuration, or unique topology. The present experiments concentrate on mammalian samples and EThcD, sceHCD, or related collision-based data; extension to other monosaccharides, organisms, instruments, and preprocessing pipelines requires corresponding data and model adaptation.

When fragmentation alone leaves an assignment ambiguous, retention behavior and biosynthetic glycome connectivity can provide complementary evidence. DeepGlycan's composition decoder does not use those additional signals [106].

Confidence is estimated differently in the four evaluation settings. The supervised benchmarks use pGlyco-derived labels filtered at a nominal 1% peptide- and glycan-level FDR, the mouse-heart analysis uses precursor-mass-shift decoys, the engineered system uses an empirical score cutoff, and the Alzheimer's workflow returns predicted compositions to a downstream search and quantitative analysis. These procedures answer different questions and do not assign one universal FDR to every DeepGlycan output. This distinction follows the broader evaluation principles developed for *de novo* sequencing [27].

4.8 Conclusion

DeepGlycan recovers glycan compositions by organizing intact-glycopeptide spectra around relative mass rather than a single fragment path. Peak-to-node expansion retains alternative ion interpretations, while partial-composition combination masses allow fragments associated with different branches to contribute to the same decoding decision. The glyco-gap analysis shows that this representation often converts a long unresolved suffix into a shorter composition problem.

Within its five-class composition space, the model transfers across EThcD and sceHCD data and from N-linked training spectra to O-linked glycopeptides without additional training. Its applications in engineered glycosylation systems and an Alzheimer's disease cohort further show how expanded composition recovery can support biological comparison and downstream hypothesis generation while preserving the peptide context of each glycan.

Chapter 5

NovoBoard: Evaluating Accuracy and False Discovery in *De Novo* Peptide Sequencing

De novo peptide sequencing is most valuable when the generating peptide is absent from a reference database. Its evaluation, however, commonly begins with spectra that a database search has already identified. This creates a mismatch between the benchmark and the intended use: a prediction that fails to match the reference can be a novel peptide, an incomplete sequence, an isobaric alternative, or a model error. Database absence alone does not distinguish among these possibilities [107, 49].

The mismatch becomes more consequential for learned sequencers. A model can extract fragmentation relationships from spectra while also learning peptide frequencies, enzyme-specific termini, and recurring sequences from its training corpus. It can therefore assign a high score to a plausible peptide whose fragment evidence is incomplete [22, 23]. Strong performance on familiar data does not by itself show that an unfamiliar prediction is supported by the measurement [24, 27].

The broader literature reaches the same task through learning-to-rank and sequence-mask-search hybrids [108, 109], and through fully convolutional sequence models [110]. A useful benchmark must therefore compare scoring behavior produced by materially different inductive biases, not only different implementations of one decoder.

Peptide and amino-acid accuracy on database-labeled spectra remain useful, but they describe only the subset that another identification procedure could already explain. They do not show how much of a prediction is supported by observed fragments, whether the

benchmark resembles the training distribution, or where a score threshold should be placed when no peptide label is available. Reliable false discovery rate (FDR) control consequently remains a central problem for *de novo* sequencing [26].

NovoBoard addresses this mismatch by making the evidence for a prediction part of the evaluation. It compares exact peptide recovery with residue- and fragment-level support, tests the same methods as the peptide population moves away from familiar training data, and calibrates model scores with decoy spectra whose sequence-specific fragment structure has been disrupted. Together, these experiments ask whether a reported sequence is supported by the spectrum and how many accepted predictions are expected to fail that criterion [27].

5.1 Metrics at Multiple Levels

Let spectrum S_i have database-derived reference peptide y_i and model prediction $\widehat{y}_i$. Exact peptide accuracy declares a prediction correct only when the full predicted sequence can be aligned to the reference under the study's mass-tolerant residue matching rule. Dataset-level amino-acid accuracy divides the total number of matched predicted residues by the total number of residues in the database-derived reference peptides across all labeled spectra. A predicted residue is treated as matching when its mass differs from the reference residue by less than 0.1 Da and the prefix masses immediately before the two residues differ by less than 0.5 Da. This accommodates mass-spectrometric ambiguities such as leucine and isoleucine without requiring literal token identity [27].

NovoBoard adds a fragment-ion metric. Let G_i be the set of observed peaks in S_i explained by at least one theoretical fragment ion from y_i, and let P_i be the observed peaks explained by at least one theoretical fragment ion from $\widehat{y}_i$. The shared evidence is $G_i \cap P_i$. Dataset-level fragment-ion accuracy is

$$\text{FragAcc} = \frac{\sum_i |G_i \cap P_i|}{\sum_i |G_i|}. \tag{5.1}$$

The implementation considers 12 theoretical ion categories: singly and doubly charged b and y ions, together with water and ammonia neutral-loss variants. A theoretical ion is matched to an observed peak within 0.02 Da [27].

Fragment accuracy does not replace sequence accuracy. A prediction can explain many major fragments while containing an unresolved local ordering; conversely, a correct sequence can be assigned from a sparse spectrum using a strong prior. The metrics reveal

different failure modes. Peptide accuracy measures exact reconstruction, residue accuracy localizes sequence error, and fragment accuracy asks whether the reported structure is visibly supported by the measurement. This fragment-level view is indispensable for a discovery claim because novelty should increase, rather than relax, the burden of spectral evidence.

Agreement with observed fragments adds an evidence-based view of correctness, but any labeled accuracy remains conditioned on how the labels were obtained. NovoBoard therefore treats dataset construction as part of the evaluation rather than as a neutral source of examples.

5.2 Benchmarking beyond Familiar Spectra

Database-derived labels are not missing at random. A search engine is most likely to identify spectra with recognizable fragment series, favorable signal-to-noise ratios, and generating peptides present in the reference. NovoBoard notes that strict database-search identifications often cover only a minority of all acquired spectra. The remaining spectra are not merely unlabeled copies of the same easy cases: they are enriched for weak fragmentation, unexpected sequences, mixtures, and other conditions under which *de novo* analysis is supposed to add value [27, 26].

Evaluating only the labeled subset therefore conditions on the success of another identification system. It can overestimate absolute performance, because easy spectra are selected, and distort relative performance, because a method whose inductive bias resembles the labeling search may be favored. It can also hide a circular relationship between training and testing: large public database-search collections supply neural training labels, then subsets generated by similar searches become test ground truth.

NovoBoard evaluates PEAKS, Novor, PointNovo, Casanovo, and GraphNovo. PEAKS and Novor provide dynamic-programming and decision-tree baselines [20, 21]. PointNovo and Casanovo represent learned sequence models [23, 24], while GraphNovo adds explicit fragment-graph structure [11]. The comparison is designed to observe how different inductive biases interact with controlled changes in data rather than to declare a permanent winner.

The benchmark begins with a human tryptic dataset from the Association of Biomolecular Resource Facilities (ABRF), a familiar setting for models trained largely on human tryptic spectra. An *A. thaliana* tryptic dataset shifts species while retaining the digestion regime [111]. The peptide population is changed more substantially in a nontryptic dataset combining six enzymes [112] and in a human leukocyte antigen (HLA) class I

immunopeptidomics dataset containing short, nonspecifically cleaved peptides [113]. These public datasets are distributed through the ProteomeXchange infrastructure, which supplies persistent project identifiers and repository provenance [114].

Two additional experiments control novelty more directly. A simulated mutation series changes one to ten residues in peptides drawn from MassIVE-KB, a large repository-derived peptide–spectrum knowledge base [115], and uses Prosit to generate corresponding major-ion spectra [47]. Matched human, yeast, and *E. coli* samples acquired under the same experimental conditions then separate peptide familiarity from obvious instrument and acquisition differences [27].

Species transfer changes the biochemical and sequence distribution, while nontryptic and HLA data change enzyme and length priors. Mutation simulation supplies a controlled distance from familiar training peptides, and the matched-species experiment examines familiarity after controlling acquisition conditions. NovoBoard does not eliminate the selection mechanism in database-derived labels. Its multi-level metrics and empirical FDR validation still require labeled spectra. The heterogeneous benchmarks instead reveal how strongly the conclusions depend on the subset for which those labels exist.

Heterogeneous labeled datasets can reveal distribution shift, but they still cannot label the unidentified spectra for which *de novo* sequencing is needed most. NovoBoard addresses that gap by moving target–decoy competition from peptide candidates to the spectra themselves.

5.3 Target–Decoy Spectra for *De Novo* FDR

Assume a sequencer returns a prediction and score $s(S)$ for each spectrum. At threshold τ, users need an estimate of the fraction of accepted predictions that are false. This set-level FDR is distinct from the posterior error probability of an individual match [116]. Database search obtains such an estimate by competing target peptide candidates with shuffled or reversed decoy peptides [15]. This construction does not transfer directly. A *de novo* method is optimized to find a peptide for a spectrum; a randomly generated peptide is unlikely to be a competitive alternative for that same measurement.

NovoBoard moves the decoy from output space to input space. For each target spectrum T_i, it generates a decoy spectrum D_i that preserves coarse properties such as peak count and the empirical distribution of mass-to-charge values and intensities, while disrupting part of the sequence-specific fragment structure. The same sequencer processes both. If

decoy predictions receive high scores, the score is responding to structure that can arise in a corrupted spectrum rather than uniquely to a valid peptide–spectrum relationship.

The random-removal strategy first pools peaks from all target spectra. For each target, it removes a proportion X of peaks uniformly and replaces them with the same number of peaks sampled from the pool. Thus, the decoy has the same peak count as its target, and replacement peaks are drawn from real experimental peaks rather than from an artificial uniform distribution. NovoBoard sweeps X from 10% to 90% to vary decoy difficulty.

A second strategy focuses corruption on intense, peptide-informative peaks. It estimates peptide length as precursor neutral mass divided by an average residue mass of approximately 123 Da, estimates the number of major b/y ions from that length, and removes an X-scaled number of the most intense peaks. Replacement peaks are sampled from the global pool of removed peaks. Because high-intensity signal is preferentially disrupted, this strategy can produce decoys that remain superficially realistic while changing the evidence used by the sequencer.

Figure 5.1 connects construction to competition. Panels (A)–(C) verify that 50% replacement preserves the target spectra's marginal m/z and intensity distributions, panel (D) shows target and decoy scores before pairing, panel (E) retains the scan-specific winners, and panel (F) compares the resulting whole-dataset and scan-specific FDR estimates across score thresholds.

For whole-dataset competition, target and decoy predictions are pooled. Let T_τ and D_τ denote the numbers of target and decoy predictions, respectively, with scores at least τ. Following the released NovoBoard implementation, the estimated FDR is the decoy-to-target ratio

$$\widehat{\mathrm{FDR}}(\tau) = \frac{D_\tau}{T_\tau}. \tag{5.2}$$

NovoBoard also studies scan-specific competition, which retains only the higher-scoring member of each target–decoy pair. The whole-dataset version is more stringent in the reported experiments because every decoy can compete with all target predictions rather than only with its source scan [27].

An estimated FDR is useful only if it tracks an empirical error rate on data where reference labels are available. This separation between an estimated confidence quantity and its calibration against an operational correctness criterion is a general requirement in shotgun-proteomics validation [117, 118]. NovoBoard therefore defines a validation target from fragment coverage. For a labeled spectrum, the released implementation counts a prediction as correct when its matched fragment set P_i covers at least a chosen proportion Y of the reference-supported set G_i. The default is $Y = 90\%$. Predictions below that

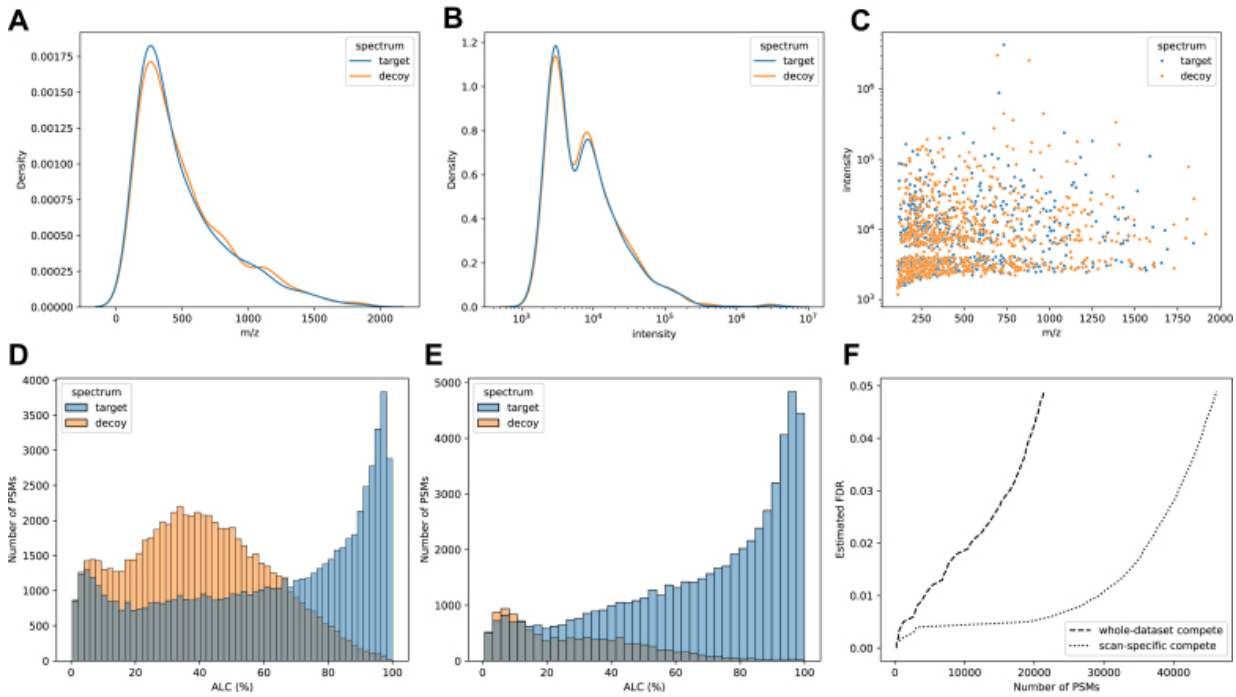


Figure 5.1: Generating decoy spectra and performing target–decoy competition for false discovery rate (FDR) estimation. (A) Peak m/z distributions, (B) peak-intensity distributions, and (C) the joint m/z–intensity distributions of target and decoy spectra from the Association of Biomolecular Resource Facilities (ABRF) dataset. For each scan ID, one decoy was generated by removing 50% of the target peaks and adding the same number of noise peaks drawn from the target-peak distribution. (D) Distributions of PEAKS *de novo* scores, reported as average local confidence (ALC), for 70,000 target and 70,000 decoy spectra. (E) Score distributions after scan-specific competition, which retains the higher-scoring peptide–spectrum match (PSM) from each target–decoy pair and leaves 70,000 *de novo* PSMs, one per scan ID. (F) FDR estimated as the proportion of decoy PSMs above each score threshold, using either the unpaired distributions in panel (D) for whole-dataset competition or the paired winners in panel (E) for scan-specific competition. Reproduced from [27].

coverage are counted as false, producing an empirical false-peptide-rate curve against which the decoy estimate can be compared.

The quantity labeled "true FDR" in the published plots is this empirical error rate, calculated on database-labeled spectra using the default 90% reference-fragment-coverage criterion. It is not an independently observed exact-sequence or chemical error rate.

This definition is intentionally not identical to exact peptide accuracy. A database search can label a spectrum that lacks enough fragments to determine every residue order uniquely. Requiring literal equality would call every alternative completion false even when the spectrum cannot discriminate it. Fragment coverage instead asks whether the prediction preserves nearly all observed evidence attributed to the reference peptide. The threshold remains a modeling choice, not a physical constant; NovoBoard exposes it so that the criterion can be tightened or relaxed for different applications.

The validation procedure uses database labels, but only to test calibration. Once a decoy construction has been chosen for a particular model and data regime, Equation 5.2 can be applied to unlabeled spectra. The labeled benchmark evaluates the error estimator; it is not a required candidate list during *de novo* inference.

A corrupted spectrum is only a useful decoy if it competes with false target predictions at an appropriate difficulty. Calibration therefore becomes an empirical part of the method rather than a fixed parameter copied between tools.

5.4 Calibrating the Error Estimate

Removing too few peaks leaves enough peptide-specific structure for a decoy to be legitimately sequenced, whereas removing too many creates an obviously random input that every model rejects. A useful decoy occupies the middle: it resembles the target in nuisance statistics but should not retain a valid high-scoring peptide explanation. The appropriate middle depends on the model's score.

Figure 5.2 makes this trade-off measurable. Panel (a) tracks residual peptide accuracy for four sequencers as random peak removal increases, while panel (b) shows for PEAKS that changing which peaks are removed can leave substantially different amounts of sequence-specific evidence at the same nominal removal percentage.

After sequencing targets and decoys, NovoBoard compares estimated and empirical FDR curves across corruption levels. For a given score threshold, an ideal setting lies near the diagonal: the fraction of accepted decoys agrees with the fragment-coverage-defined fraction

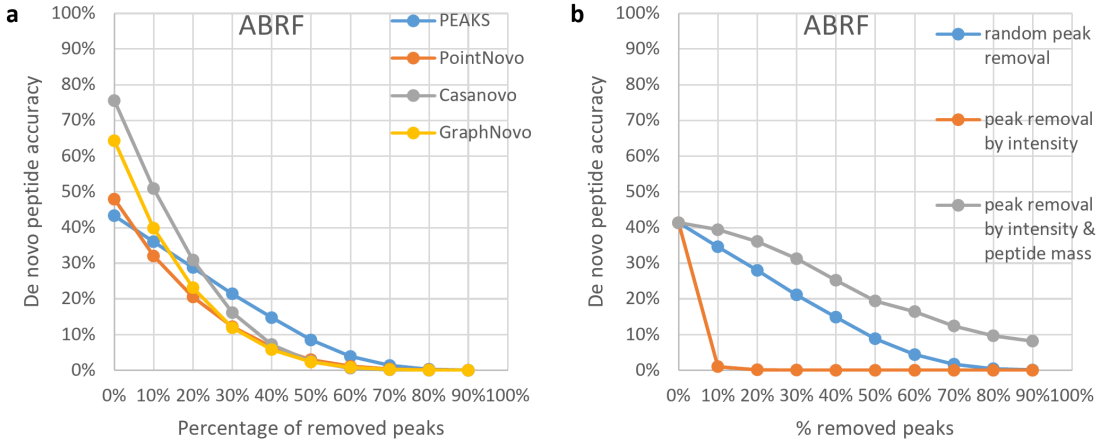


Figure 5.2: Decoy-spectrum difficulty on the ABRF dataset. (a) *De novo* peptide accuracy of PEAKS, PointNovo, Casanovo, and GraphNovo as a function of the percentage of peaks removed at random from each target spectrum when generating its decoy. (b) PEAKS *de novo* peptide accuracy as a function of the percentage removed under three strategies: random peak removal, removal by intensity, and removal by both intensity and peptide mass. Adapted from Supplementary Fig. S1 of [27].

of false target predictions. A decoy that is too easy yields too few high-scoring decoys and underestimates error. A decoy that is too hard can overestimate error by competing more strongly than false targets do.

The same 50% random-removal decoys produced approximately accurate estimation for one tool, underestimation for some tools, and overestimation for another. Suitable removal proportions therefore had to be selected separately for PEAKS, PointNovo, Casanovo, and GraphNovo. The intensity-aware strategy reduced some discrepancies but could overestimate error for algorithms that remain effective on its residual graph or dynamic-programming structure [27].

Figure 5.3 shows this calibration for GraphNovo. Panel (A) compares estimated FDR with the empirical reference curve across score cutoffs and removal proportions; panel (B) shows how the same choices change the number of accepted peptide–spectrum matches at a stated error threshold.

A decoy generator is therefore not an independent measuring device that can be attached unchanged to any sequencer. It is part of a calibration protocol involving the model, its score, the spectrum distribution, the corruption mechanism, and the definition of correctness. FDR becomes credible when these components are validated together.

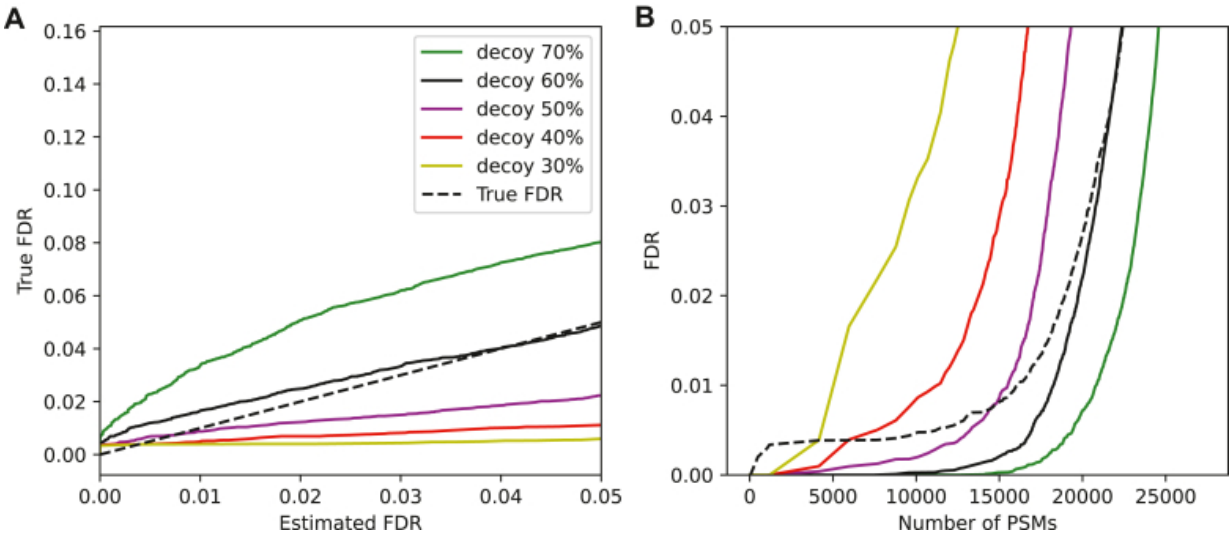


Figure 5.3: FDR estimation and validation for GraphNovo using decoy spectra generated from the ABRF dataset by random peak removal. Colors denote the percentage of target peaks removed and replaced with noise peaks; for example, "decoy 70%" (green) replaces 70% of the peaks. (A) Estimated FDR versus true FDR over different GraphNovo score cutoffs. (B) Number of *de novo* peptide–spectrum matches versus estimated and true FDR over the same score cutoffs. Reproduced from [27].

More generally, imperfect matching against an incomplete candidate space creates error modes that ordinary complete-null assumptions do not capture [119]. NovoBoard changes the object being decoyed, but the same lesson remains: the error model must match the identification procedure actually being evaluated.

Raw model scores are not comparable across tools. One system may report a percentage-like confidence, another an accumulated log probability, and another a graph-path score. Even within one tool, score distributions can move when peptide length or spectrum quality changes. A claim that one method has a "higher score" is therefore meaningless unless the score has been calibrated to an error criterion [120]. Semisupervised rescoring systems likewise learn a useful decision scale from the analysis context rather than treating a search engine's raw score as universal [121]; this precedent does not, by itself, calibrate a *de novo* score.

FDR curves provide a common decision axis. At a specified estimated or empirical FDR, the number of accepted target predictions measures identification power under the chosen error definition. On the ABRF data, for example, GraphNovo and Casanovo retain substantially more predictions than PEAKS and PointNovo over the reported FDR range. This does not imply that every retained peptide is correct or that the same ordering will hold under distribution shift. It means that, after tool-specific calibration on that dataset, their score rankings recover more targets at the selected aggregate error rate [27].

Accuracy on a labeled subset may improve by returning predictions only for easy spectra. A method intended to explore unidentified data must also decide how many candidates can be reported. Conversely, high coverage without calibrated error produces a long candidate list with unknown reliability. Reporting both accuracy and identification power prevents selective prediction from being confused with comprehensive discovery.

With the score and decoy construction evaluated together, the benchmark can be used to study a different source of unreliability: changes in the peptide population relative to the data on which a model was trained.

5.5 Generalization across Species, Digestion, and Peptide Class

On the human tryptic ABRF data, the deep-learning models operate close to their dominant training distribution. All tools achieve comparatively strong results, and Casanovo and GraphNovo report many predictions at controlled empirical error. A familiar benchmark is

useful for verifying that an implementation works, but it is a weak test of discovery because human tryptic spectra and peptide sequences are heavily represented in public training resources.

The *A. thaliana* dataset changes this interpretation. Accuracy declines for all tools, but not uniformly. Casanovo's peptide accuracy falls from 76% on ABRF to 43%, while its residue and fragment measures also lose their earlier advantage. PEAKS declines much less at peptide and residue levels, and GraphNovo leads the reported metrics and FDR comparison on the plant data [27]. These values should not be read as an immutable ordering of algorithms. They show that a model's apparent strength can depend strongly on proximity between training and test peptide distributions.

Figure 5.4 makes the distribution dependence visible. Panels (A) and (C) compare peptide, amino-acid, and fragment-ion accuracy on the familiar ABRF and cross-species *A. thaliana* data, while panels (B) and (D) compare how many predictions remain as the empirical error threshold is varied in each setting.

Fragment accuracy helps interpret the drop. If peptide accuracy decreases while fragment accuracy remains relatively strong, a method may still recover much of the observed structure but fail exact completion. If both fall, the predicted sequences have lost direct spectral support. Reporting both prevents exact-match sequence priors from being mistaken for universal spectrum interpretation.

The nontryptic experiment tests a different shortcut. Public training data overrepresent peptides produced by trypsin, which usually end in lysine or arginine. PointNovo and Casanovo were evaluated with models adapted to broader cleavage patterns, while the available GraphNovo model remained tryptic. Even so, GraphNovo achieved the strongest reported accuracy and FDR results. Removing nontryptic peptides that overlapped the MassIVE-KB training resource did not reverse the result. This is consistent with, but does not alone prove, the hypothesis that explicit fragment-graph structure transfers better than a stronger sequence prior [27, 11].

HLA class I peptides present another distribution. They are short and are not produced by a single proteolytic cleavage rule. Most tools except Novor obtained relatively high peptide accuracy, with several results in the 58–64% range. Their short length reduces the number of decisions required for exact reconstruction, so this favorable result should not be generalized to longer nontryptic peptides. Within ABRF, stratification by length confirms that all methods become less accurate as peptide length increases.

Figure 5.5 separates these two tests: panels (A) and (B) report accuracy and empirical FDR for the six-enzyme nontryptic data, whereas panels (C) and (D) report the corre-

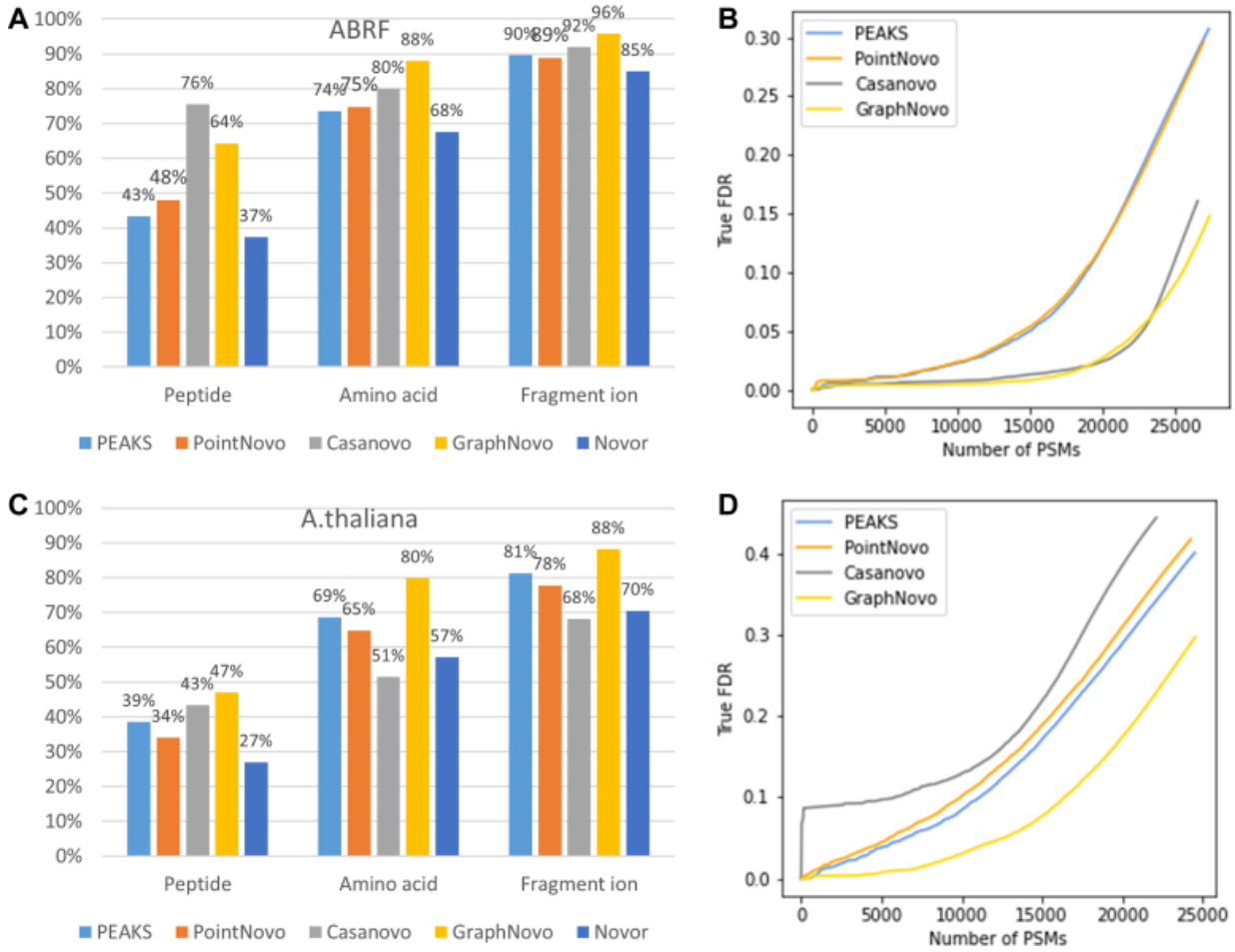


Figure 5.4: Evaluation of *de novo* sequencing results on the ABRF and *A. thaliana* datasets. (A) Peptide, amino-acid, and fragment-ion accuracies and (B) true FDR as a function of the number of accepted peptide–spectrum matches on the human tryptic ABRF dataset. (C) The same three accuracy measures and (D) true FDR as a function of the number of accepted peptide–spectrum matches on the *A. thaliana* tryptic dataset. Reproduced from [27].

sponding quantities for the shorter HLA peptides. Reading the panels together prevents the favorable HLA peptide accuracy from being attributed to cleavage-prior transfer alone.

Species and digestion shifts alter many variables at once. The mutation series and matched-acquisition experiment narrow the question by controlling, respectively, sequence distance and acquisition conditions.

5.6 Controlled Sequence Novelty and Training-Set Familiarity

NovoBoard next creates a controlled sequence shift. Ten thousand MassIVE-KB peptides are selected, one to ten residue substitutions are introduced, and Prosit predicts their major higher-energy collisional dissociation (HCD) fragment spectra. As mutation count increases, each peptide moves farther from a familiar training sequence while retaining a known answer. All tools deteriorate, but their curves differ. Casanovo shows the steepest peptide-accuracy decrease in the reported comparison, from 45% to 24%; GraphNovo and PointNovo show smaller declines at residue level [47, 27].

Position matters as well as mutation count. For singly mutated peptides, substitutions near the first or last three residues are recovered less accurately than substitutions in the middle. Inspection attributes part of this effect to missing terminal fragments, which leave adjacent swaps or equal-mass combinations unresolved. This finding connects NovoBoard's evaluation back to GraphNovo's mechanism: a nominally known mutation can still be ambiguous if the spectrum lacks the fragment coordinates needed to localize it.

Figure 5.6 separates the two effects. Panels (A) and (B) show peptide and amino-acid accuracy as the number of substitutions increases, while panels (C) and (D) localize the single-substitution result by exact position and by terminal-versus-middle region.

The simulation has a deliberate boundary. Prosit spectra contain predicted major ions and do not reproduce the full noise and alternative-fragment distribution of experimental data. The experiment isolates sequence change but cannot measure all aspects of real novel-peptide discovery. Its value lies in controlled interpolation from familiar to mutated sequences, not in replacing biological validation.

A naive comparison of peptides present versus absent in the training library confounds familiarity with spectrum quality: peptides absent from a large library may simply have weaker spectra. NovoBoard addresses this by acquiring human, yeast, and *E. coli* samples under the same LC–MS/MS conditions and selecting database identifications with similar

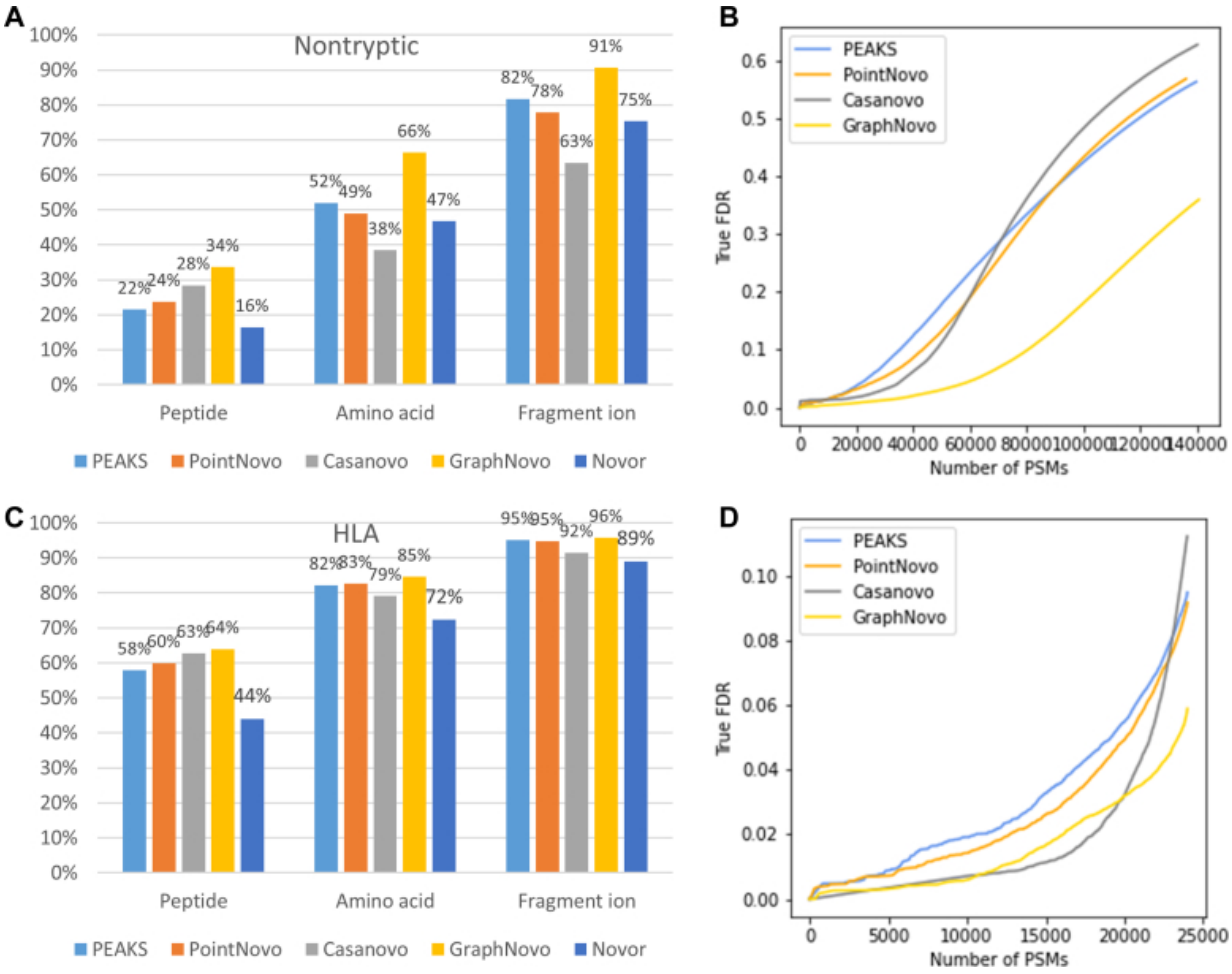


Figure 5.5: Evaluation of *de novo* sequencing results on the nontryptic and human leukocyte antigen (HLA) class I datasets. (A) Peptide, amino-acid, and fragment-ion accuracies and (B) true FDR as a function of the number of accepted peptide–spectrum matches on the six-enzyme nontryptic dataset. (C) The same three accuracy measures and (D) true FDR as a function of the number of accepted peptide–spectrum matches on the HLA dataset. Reproduced from [27].

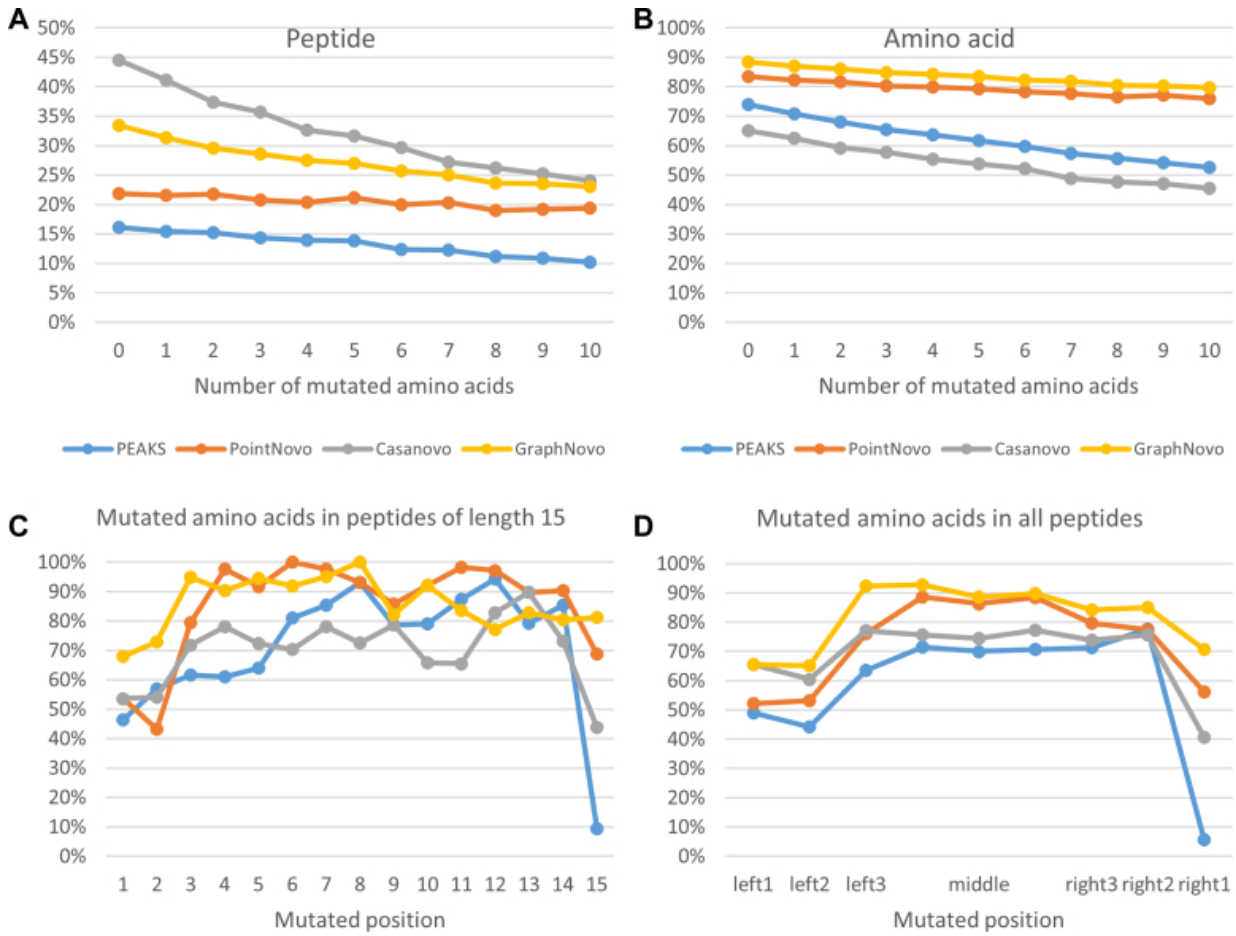


Figure 5.6: Evaluation of *de novo* results on the simulated dataset of 10,000 mutated peptides. (A) Peptide accuracy and (B) amino-acid accuracy as the number of substituted amino acids increases from 0 to 10. (C) Accuracy of the mutated amino acid at each sequence position for singly mutated peptides of length 15. (D) Accuracy of the mutated amino acid in the first three, middle, and last three positions of singly mutated peptides of all lengths. Reproduced from [27].

score and length distributions. Approximately 96% of human peptide–spectrum matches occur in MassIVE-KB, compared with only 0–2% for yeast and *E. coli*. Traditional tools and the evaluated PointNovo configuration remain nearly stable across species, whereas Casanovo and GraphNovo show a modest 3–4 percentage-point peptide-accuracy decrease on yeast and *E. coli* [27].

Figure 5.7 shows why the matched acquisition is necessary. Panels (a) and (b) expose the spectrum-quality confounding in a simple MassIVE-KB present/absent partition of ABRF, whereas panels (c) and (d) compare peptide and amino-acid accuracy across the matched human, yeast, and *E. coli* acquisitions.

The interpretation remains restrained. A small species-associated drop is compatible with memorization or a learned species-specific sequence prior, but it can also reflect residual biological differences not removed by acquisition matching. The experiment supplies a better control and identifies a warning signal; it does not reduce deep sequence models to lookup tables.

5.7 Discussion

NovoBoard shows that a *de novo* prediction cannot be assessed by database absence or raw model confidence alone. Peptide, residue, and fragment metrics describe exact completion, local recovery, and direct spectral support, and they can rank methods differently. The species, digestion, peptide-class, mutation, and length experiments likewise show that accuracy on one familiar distribution is not a sufficient proxy for performance elsewhere.

Target–decoy spectra add a way to estimate error without supplying a decoy peptide vocabulary. On labeled data, suitably chosen decoys produce estimated FDR curves close to empirical curves defined by fragment coverage; after calibration, the same procedure can set thresholds on unlabeled spectra. The result is conditional rather than universal. The model, score, spectrum distribution, corruption mechanism, and definition of correctness all participate in the calibration, so one removal proportion cannot be transferred unchanged among tools or datasets.

The reported experiments are concentrated on Orbitrap, data-dependent-acquisition spectra with HCD fragmentation. Scores and decoy behavior may differ under electron-transfer dissociation (ETD), EThcD, DIA, lower-resolution instruments, chimeric spectra, or other preprocessing pipelines. DIA *de novo* sequencing has required a dedicated model and data treatment [122], illustrating why extension of the framework to those settings requires new calibration rather than reuse of the reported removal percentage.

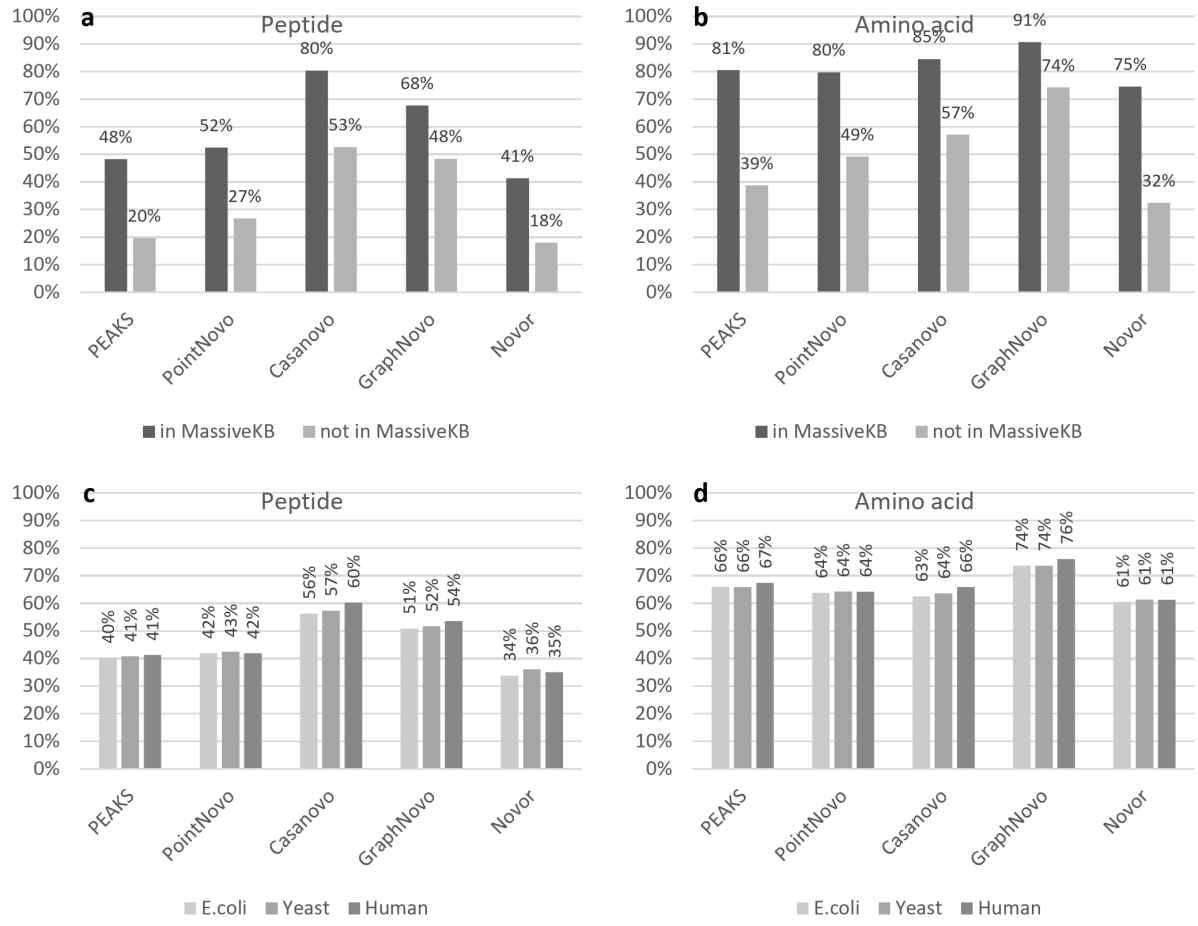


Figure 5.7: Analysis of peptide-sequence memorization by *de novo* sequencing tools. (a) Peptide accuracy and (b) amino-acid accuracy on ABRF peptide–spectrum matches partitioned by whether the reference peptide is present in MassIVE-KB. (c) Peptide accuracy and (d) amino-acid accuracy on the matched-acquisition *E. coli*, yeast, and human datasets generated in the study. Adapted from Supplementary Fig. S9 of [27].

The constructed decoys are controlled interventions, not samples from the unknown distribution of every false match. Random replacement preserves marginal peak distributions while changing joint fragment structure, whereas intensity-aware replacement makes a different intervention. A future model could exploit artifacts of either construction, so decoy difficulty and realism must be reevaluated as scoring functions change. The empirical "true FDR" has a related dependency: database-derived references are incomplete, and the 90% fragment-coverage criterion accepts some nonidentical sequences while rejecting predictions that explain less of the labeled evidence. Applications requiring exact residue identity may need a stricter target, whereas applications reporting sequence tags may choose another.

Other experimental conclusions retain their local qualifications. The mutation study uses predicted spectra without realistic experimental noise, and the matched-species study controls acquisition but not every biological covariate. Models were evaluated with available checkpoints and task-specific configurations. Novor's server could not process the generated decoy files and was therefore excluded from FDR comparisons, and the reported throughput check is approximate rather than a hardware-normalized resource benchmark.

An FDR threshold controls an aggregate rate, not the truth of an individual discovery. A set accepted at 1% estimated FDR may still contain false candidates, and a mass-supported novel peptide does not by itself establish a unique chemical modification. Especially for an unprecedented PTM, acceptance should lead to orthogonal validation rather than terminate the analysis.

These distinctions become essential when the relevant mass shift was not represented as a residue token during training. The next chapter applies target–decoy calibration within RNovA, checks its structural confidence thresholds with isotope-labeled data, and adds synthetic-peptide and biological evidence for selected PTM hypotheses [28]. The resulting workflow can evaluate an unfamiliar prediction without equating unfamiliarity with truth.

Chapter 6

RNovA: Zero-Shot *De Novo* Peptide Sequencing with Open PTM Discovery

6.1 Introduction

Post-translational modifications (PTMs) alter protein mass, activity, localization, and molecular interactions [123], yet most computational identification methods can report only modifications specified before analysis. A database search is restricted by the peptide sequences and variable modifications admitted to the search. Neural *de novo* sequencers remove the requirement for a complete database peptide, but their decoders commonly remain restricted to residue and modified-residue tokens available during training. A biologically relevant modification absent from both spaces cannot be expressed as the correct answer.

Expanding a fixed vocabulary is effective for known modifications but does not address discovery of an unexpected residue chemistry. Rare or newly characterized PTMs may have too few labeled spectra for modification-specific training, and an enlarged search space also increases the number of plausible false assignments. The same search-space and confidence trade-off appears in global PTM database searches [124, 52]. Noble and MacCoss identified out-of-database identification and false-positive control over expanding modification spaces as linked computational challenges; a later review by Bittremieux, Noble, and colleagues emphasized that most deep-learning *de novo* systems still support limited PTM vocabularies and usually need retraining to add a new one [25, 26].

Blind and mass-tolerant database searches progressively relaxed the modification list while retaining a peptide-sequence scaffold [125, 126]. Later algorithms made broad mass-shift searches more practical and recovered many spectra that narrow searches left unassigned [127, 16]. At larger scale, TagGraph used recurring evidence to map extensive modification landscapes, while open spectral-library search made broad modified-spectrum matching efficient [17, 128]. These methods established the power of open modification search, but openness over delta mass is not the same as reconstructing a peptide without a database or library sequence scaffold.

A tandem mass spectrum supplies a noisy set of constraints on the peptide that produced it. Fragmentation breaks peptide copies at different backbone positions, and the difference between two correctly assigned prefix-fragment masses often equals the mass of the intervening residue or residues. Real spectra contain missing and unrelated peaks, and a modified residue can introduce a mass not represented by any available output token. In that case, a closed decoder may substitute a canonical residue, split the mass across several residues, truncate the sequence, or select a familiar modification with a nearby mass. The correct action was excluded before inference began, so the failure need not appear as low model confidence.

RNovA changes the order of inference [28]. PathSearcher first recovers fragment-supported mass paths without requiring every interval to equal a known residue. Recurrent unresolved mass tags are then clustered and converted into residue–mass-shift candidates. SeqFiller uses those candidates to reconstruct complete peptide sequences without PTM-specific retraining. Chemical annotation occurs after spectral and sequence evidence has been assembled, and selected assignments can then be tested with synthetic reference peptides.

This organization extends the intermediate-path formulation introduced by GraphNovo [11]. GraphNovo preserves unresolved intervals caused by incomplete fragmentation but constructs transitions from a predefined residue vocabulary. RNovA retains the mass path while permitting intervals that are not explained by that vocabulary. Discovery is therefore open at the mass-tag stage, whereas complete sequence reconstruction remains explicitly candidate-conditioned.

6.2 RNovA Architecture and Open PTM Workflow

RNovA contains two independently trained transformer modules connected by a deterministic candidate-generation pipeline (Fig. 6.1). PathSearcher infers a fragmentation path and

exposes unexplained intervals. The intermediate pipeline clusters and aligns those intervals to generate recurrent residue-mass candidates, and SeqFiller constructs complete sequences from the resulting candidate list. The workflow needs neither a user-provided PTM list nor a protein database to generate candidates. Sequence reconstruction nevertheless operates over the explicit, data-derived set supplied to SeqFiller.

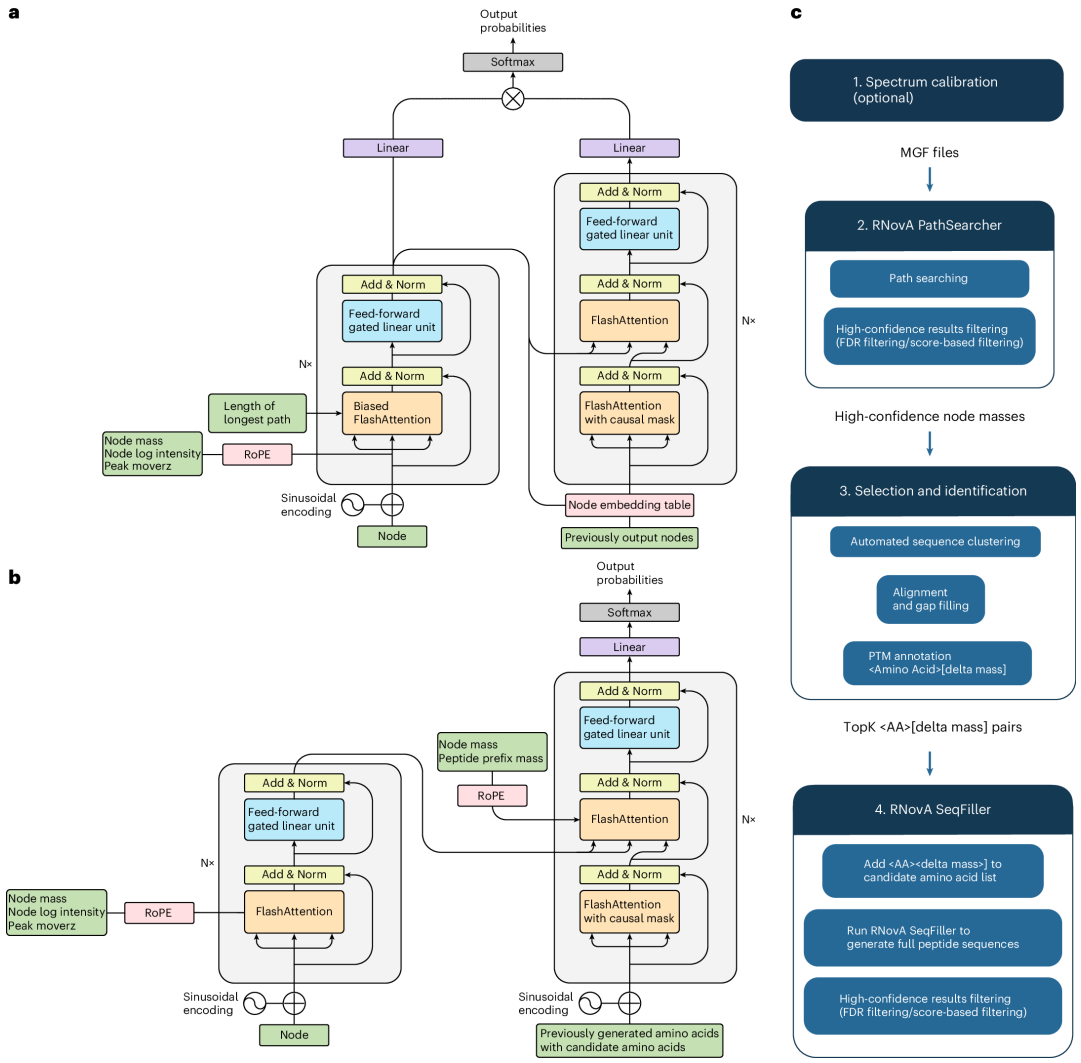


Figure 6.1: Overview of the RNovA architecture and workflow. Panel a shows RNovA PathSearcher, and panel b shows RNovA SeqFiller. Panel c shows the open post-translational modification (PTM) discovery workflow, which combines the two modules with confidence filtering, mass-tag clustering and alignment, residue–delta-mass candidate generation, and complete-sequence reconstruction. PathSearcher can retain intervals that do not match predefined residues, whereas SeqFiller remains conditioned on an explicit candidate list. MGF denotes Mascot Generic Format; optional chemical annotation follows sequencing. Reproduced from the RNovA article [28].

6.2.1 PathSearcher

RNovA represents a centroided MS/MS spectrum as a graph. Nodes are candidate fragment ions and carry exact mass, log-transformed intensity, observed m/z, and a node class corresponding to predicted fragment-ion type. Unlike a graph whose edges are created only when two nodes differ by a known residue mass, RNovA permits any inter-node mass difference. Connectivity is modeled through attention rather than through a hand-enumerated PTM edge list. The enlarged space is harder to search, but it allows a physically supported path to cross a residue mass that no training token describes.

PathSearcher begins at the origin, traverses spectrum nodes toward the precursor mass, and returns an ordered sequence of node masses. Adjacent differences form *mass tags*. A tag equal to a canonical residue mass can be translated directly. An unmatched tag is retained. It may be a modified single residue, the combined mass of two or more residues hidden by missing fragmentation, a substitution, or an incorrect path interval. PathSearcher does not need to choose among these meanings at this point. Its job is to localize the unresolved mass relative to other fragment-supported positions and to assign a structural score to the path.

This restraint is what makes open detection possible. If PathSearcher were required to name every transition during graph traversal, its search would again be closed by the names supplied in advance. If it accepted every arbitrary transition, the output would be dominated by noise. RNovA instead learns which node sequences are globally consistent with peptide fragmentation while leaving unfamiliar mass gaps symbolically unresolved. Confidence filtering and recurrence across spectra later provide the evidence needed to decide which gaps deserve further analysis.

6.2.2 Mass-Tag Clustering and Candidate Generation

An unusual interval in one spectrum is weak evidence for new chemistry. It can arise from a noise peak, a missing-fragment region, a co-isolated peptide, or a local model error. RNovA therefore extracts mass-tag sequences only from high-confidence PathSearcher outputs and looks for repeated patterns across the dataset. The automation between PathSearcher and SeqFiller is not an incidental post-processing script; it is the mechanism that converts open observations into a bounded hypothesis space.

Mass-tag sequences are grouped by a greedy incremental clustering procedure inspired by CD-HIT [129]. Pairwise comparison uses a mass-adapted Needleman–Wunsch global alignment [130]. The alignment permits block matches, so a contiguous group of tags on one

side can match a group on the other when their summed masses agree within the default 0.02-Da tolerance. This matters because the same peptide region may be resolved into individual residues in one spectrum and into a single missing-fragment interval in another.

Recurrence alone is not sufficient: flexible alignments can produce coincidental matches. RNovA constructs an empirical null distribution from 50,000 randomly generated tag-sequence pairs for each tested sequence length from 6 to 15. A real match must reach similarity of at least 0.7 and an empirical P value below 10^{-4}. Singleton clusters are removed. Within each retained cluster, sequences are aligned to a centroid and converted into a consensus. Canonical-mass tags become residue tokens; unmatched masses remain placeholders until residue context and delta mass can be inferred. This sequence of filters embodies the difference between detecting an anomaly and proposing a reproducible candidate [131, 28].

The candidate representation is a residue plus a mass shift, such as K[+42.011], rather than a mandatory PTM name. The system can optionally query Unimod or another resource for chemical interpretations, retaining multiple matches when the mass is ambiguous. Annotation does not alter SeqFiller inference: the decoder uses masses, so a raw delta-mass token and a named token with the same mass are computationally equivalent. By default, the automated workflow selects the three most abundant candidates and passes them to SeqFiller without requiring user annotation or selection. The number three is a pipeline setting, not a claim that a sample contains only three modifications.

Open-search postprocessors make a related distinction: an observed delta-mass distribution can be detected and summarized before its chemical causes are interpreted [132, 54]. Recurrence and accurate mass can prioritize an explanation, but neither makes a PTM name unique.

6.2.3 Candidate-Conditioned Sequence Reconstruction

SeqFiller answers a different question from PathSearcher. Given a spectrum, a generated peptide prefix, and a current list of possible residues, which candidate should extend the sequence? Its decoder concatenates the prefix with the candidate set and returns an action value for each candidate. Prefix masses and prefix-plus-candidate masses are compared with the spectrum graph through relative mass encoding. A newly discovered residue mass can therefore be inserted into the candidate list and evaluated immediately, without creating a new classifier output or collecting PTM-specific training spectra.

Figure 6.2 shows how SeqFiller evaluates an explicit candidate list rather than searching every continuous mass and chemical formula at each step. In the end-to-end workflow,

PathSearcher, clustering, alignment, and ranking generate this list automatically; in a benchmark, the evaluator can supply it directly. RNovA therefore discovers residue masses without a predefined PTM list, while SeqFiller reconstructs sequences conditionally on the candidates available for that analysis.

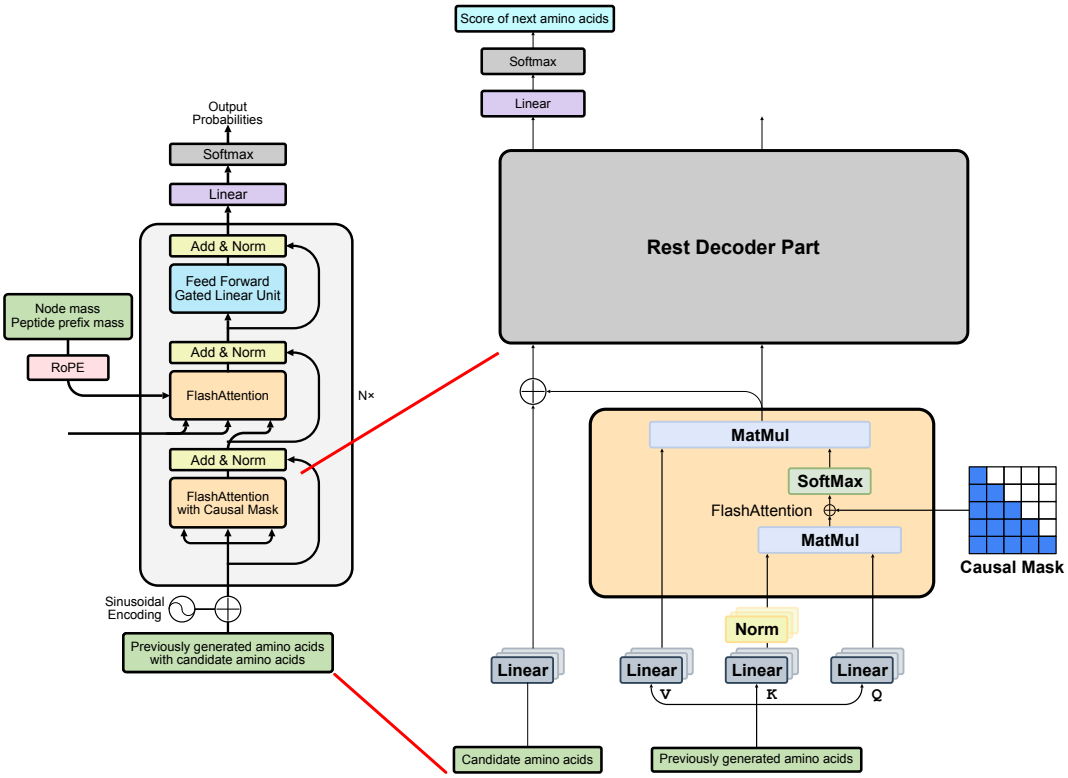


Figure 6.2: Candidate-conditioned decoding in RNovA SeqFiller. At each prediction step, the current list of candidate amino acids is appended to the autoregressively generated peptide prefix and the combined token sequence is passed to the decoder. Residue and prefix masses are encoded with rotary position encoding; the decoder stack contains FlashAttention layers with and without causal masks, and a linear projection plus softmax assigns a score to each candidate next residue. Thus a data-derived modified-residue mass can be evaluated without retraining, although decoding remains conditional on the explicit candidate set supplied for that analysis. Reproduced from Fig. S2 of the RNovA supplementary material [28].

6.2.4 Relative-Mass-Aware Learning

Both modules use six-layer transformers with hidden dimension 768 and 12 attention heads. FlashAttention-2 provides the attention implementation, and rotary positional embedding (RoPE) exposes mass differences to attention [60, 133, 94]. In language, relative position expresses how far tokens are separated. In RNovA, the analogous quantity is the difference between fragment masses or between a proposed prefix mass and an observed node. A modification changes an absolute residue mass, but the physical test remains relational: after adding the candidate mass, do the resulting prefixes align with fragment evidence?

RNovA is trained as a value-based, reinforcement-learning-style sequential decision problem. At a partial state, PathSearcher scores node actions and SeqFiller scores residue actions. Oracle action values indicate whether an action can still lead to a complete trajectory consistent with the known training target. After an incorrect decision makes exact completion impossible, subsequent actions on that trajectory receive zero oracle value. Parameters are optimized with binary cross-entropy on these action-value labels rather than ordinary token-level teacher forcing. At inference, the same free-running decision structure is used.

This formulation aligns learning with the global mass-consistency problem and avoids giving the decoder the correct prefix at every training step. Combined with relative mass encoding, it permits candidate actions to be judged by physical consistency even if a candidate's symbolic identity was never a training class.

6.3 Evaluation on Standard and Modified Peptides

We first evaluated whether permitting unrestricted mass intervals compromises ordinary peptide sequencing and whether SeqFiller can use modified-residue candidates that were absent from PTM-specific training. The standard benchmark measures path and sequence recovery on conventional peptide chemistry; the ProteomeTools experiment evaluates reconstruction and localization after modified residues are supplied as candidates.

6.3.1 Standard Peptide Sequencing

Figure 6.3 evaluates conventional peptide chemistry on the GraphNovo benchmarks. The peptides contain the 20 canonical residues, with fixed cysteine carbamidomethylation and variable methionine oxidation. The test sets are the GraphNovo benchmarks from

Arabidopsis thaliana, *Caenorhabditis elegans*, and *Escherichia coli*; the RNovA models were trained on human MassIVE-KB data and were not adapted to these organisms. The test therefore combines an across-species shift with a conventional chemical vocabulary.

SeqFiller achieved peptide recall of 79.17%, 79.25%, and 82.11% on the three species, respectively. Amino-acid recall was 87.47%, 87.97%, and 90.35%. These results establish that candidate conditioning and an open discovery interface do not prevent accurate prediction of ordinary peptides in the tested setting. The experiment measures conventional sequence recovery using database-derived ground truth; open PTM discovery is evaluated in the later TestMod and biological studies.

PathSearcher provides a more direct control on the cost of permitting arbitrary mass intervals. Its node recall was 92.41%, 94.56%, and 96.42%, compared with 93.20%, 95.30%, and 96.30% for GraphNovo PathSearcher. Node precision was 93.59%, 94.6%, and 96.41%, compared with 92.60%, 94.8%, and 95.90%. RNovA is therefore comparable rather than uniformly superior at the node level. That is the appropriate result for this control: enlarging the allowed transition space did not cause path recovery to collapse, while leaving unmatched intervals available for subsequent discovery.

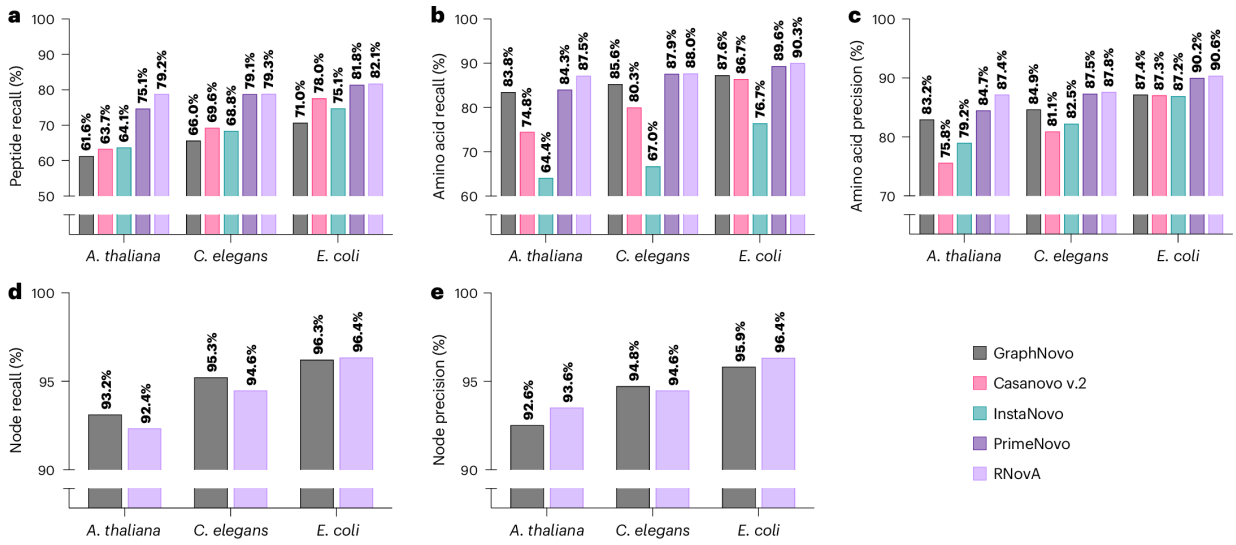


Figure 6.3: Benchmarking RNovA on datasets containing standard amino acids. Panels a–c report peptide recall (a), amino-acid recall (b), and amino-acid precision (c) for RNovA SeqFiller and compare it with GraphNovo, Casanovo v.2, InstaNovo, and PrimeNovo. Panels d and e report node recall (d) and node precision (e) for RNovA PathSearcher and GraphNovo PathSearcher on the same spectra. All evaluations use the GraphNovo held-out test data from *A. thaliana*, *C. elegans*, and *E. coli*. This is a conventional amino-acid control of path and sequence recovery, not an open-PTM discovery test. Reproduced from the RNovA article [28].

6.3.2 Zero-Shot Reconstruction of Modified Peptides

The next experiment changes the chemistry while keeping the candidate interpretation available to the evaluator. Nine modification types from the ProteomeTools synthetic peptide resource were tested [134]. Only acetylated lysine and phosphorylated tyrosine among these evaluated modified residues were included in the training data; the other tested modification types probe transfer beyond the trained PTM set. Candidate modified residues were supplied to SeqFiller for this benchmark. The experiment therefore tests *zero-shot reconstruction and localization*, not automatic mass-tag discovery.

RNovA outperformed PEAKS on the reported aggregate sequence metrics across the selected PTMs (Fig. 6.4) [20]. Several examples illustrate different aspects of the result. For phosphorylated tyrosine, amino-acid recall and precision were 94.97% and 95.44%, compared with 92.52% and 86.87% for PEAKS. Because phosphorylated tyrosine was represented in training, this example is a performance comparison, not evidence of an unseen PTM. Crotonylated lysine gives a cleaner transfer example: RNovA amino-acid precision was 95.67%, compared with 83.92% for PEAKS. For citrullinated arginine, peptide recall was 75.52% versus 27.18%. For glutarylated lysine, recovery of the modified residue itself was 91.57%, compared with 77.74%.

PathSearcher can be evaluated without giving the mass tag a chemical name. The site identification rate measures whether a partial path places the modified interval at the correct position. Reported rates included 88.46% for phosphorylated tyrosine, 93.63% for glutarylated lysine, 94.95% for hydroxyisobutyrylated lysine, and 72.35% for citrullinated arginine. These values show that localization is evidence-dependent and modification-dependent; 72.35% is useful performance, not unambiguous localization. More importantly, the path can localize an unusual mass before a symbolic PTM label is needed.

Established site-localization methods likewise treat the modified residue position as a separately scored hypothesis rather than an automatic consequence of peptide identification [135, 136]. Their phosphorylation-oriented scores are not transferred to RNovA here; the relevant principle is the separation of sequence identification from site evidence.

Together, Fig. 6.3 and Fig. 6.4 show that an unrestricted path-transition model can remain competitive on standard spectra, and that relative-mass decoding can use an evaluator-supplied residue-mass candidate without modification-specific retraining. These experiments do not yet test whether the candidate can be generated from the data. The nonnatural spike-in experiment below therefore evaluates the complete path from an unfamiliar mass tag to sequence reconstruction.

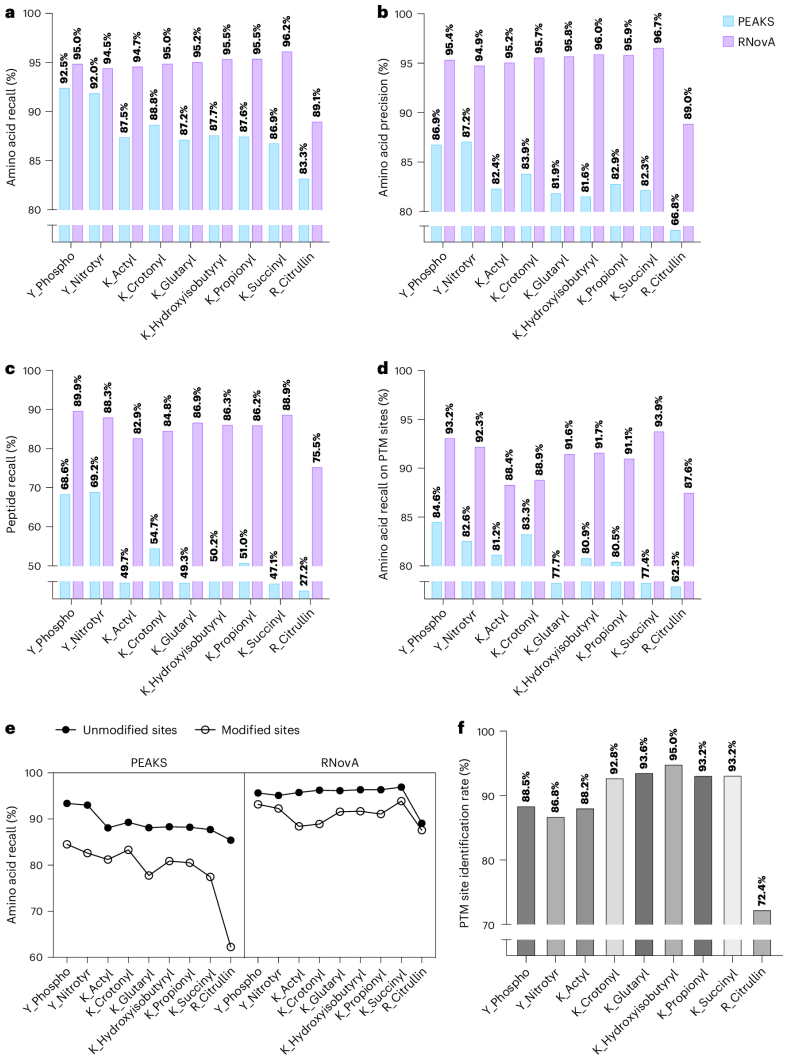


Figure 6.4: RNovA and PEAKS on synthetic peptides carrying nine PTMs from the ProteomeTools dataset PXD009449. Panels a–c compare amino-acid recall (a), amino-acid precision (b), and peptide recall (c). Panel d reports amino-acid recall at modified residues; panel e compares recall at unmodified and modified sites; and panel f reports the PathSearcher PTM-site identification rate from partial peptide predictions. Only phospho- and acetyl-modified residues occurred in RNovA training. Modified-residue candidates were supplied for this benchmark, so it tests zero-shot candidate use, sequence reconstruction, and localization rather than automatic candidate discovery. Reproduced from the RNovA article [28].

6.4 False-Discovery Estimation and Isotope Validation

PathSearcher and SeqFiller produce different statistical objects. PathSearcher accepts fragment-path nodes, whereas SeqFiller accepts amino-acid decisions. RNovA therefore applies NovoBoard's target–decoy spectrum principle separately to the two modules [27]. A node, a residue, a complete peptide, a recurring PTM candidate, and a named chemical structure are distinct hypotheses and do not inherit one another's error rate.

6.4.1 Target–Decoy Calibration

For every target MS/MS spectrum, RNovA creates a decoy by removing 40% of peaks and replacing them with noise peaks sampled from the same intensity distribution. The target and its decoy are processed independently, and only the higher-scoring result for a scan is retained in target–decoy competition. PathSearcher predictions are ranked by node-level output scores; SeqFiller residues are ranked by their action values. A separate threshold is chosen for each module so that decoy-derived predictions constitute no more than 1% of the accepted predictions at that level.

The 40% corruption level should not be mistaken for a universal physical constant. If too few peaks are replaced, a decoy can remain too similar to its target; if too many are replaced, the classification problem can become unrealistically easy. The supplementary removal experiment compares 40%, 50%, and 60% settings across replicates and motivates the selected value for RNovA (Fig. 6.5). This is model- and data-dependent calibration. Applying the same procedure to a new score, fragmentation regime, or acquisition distribution requires checking that decoys remain realistic and that target and decoy scores behave as intended.

Accepted PathSearcher output is described at *node level*; accepted SeqFiller output is described at *amino-acid level*. RNovA explicitly does not calculate peptide-level FDR. A single uncertain region can make an entire peptide string incorrect even when most positions are well supported. Rather than converting the residue threshold into an unjustified peptide claim, the pipeline masks low-confidence residues in contiguous segments defined by PathSearcher mass tags. This retains supported structure and exposes local uncertainty.

The same separation applies to downstream objects. A mass-tag cluster does not have 1% candidate FDR merely because its contributing nodes passed a 1% node threshold. A localized site does not automatically have 1% site FDR. An Unimod name assigned to a delta mass does not have 1% chemical-identity FDR. Throughout this analysis, "passing

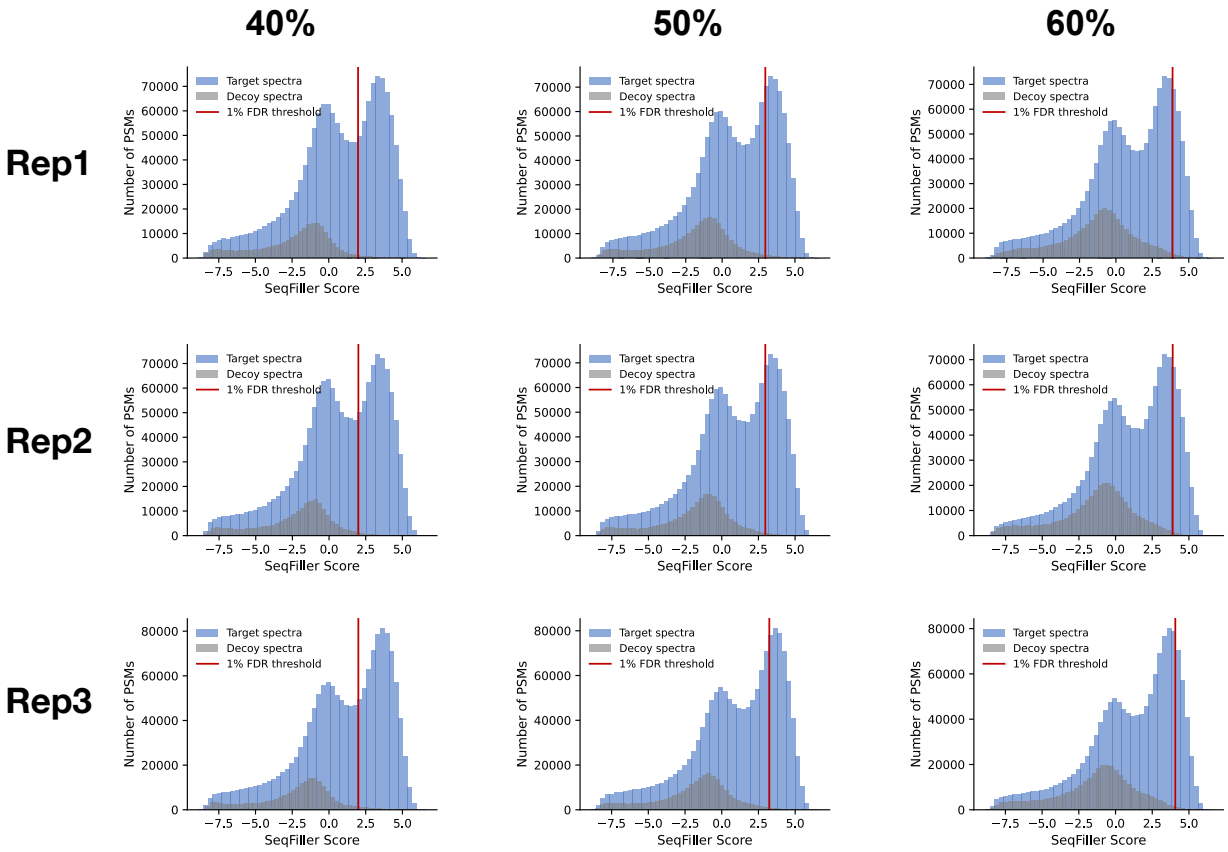


Figure 6.5: Optimization of the fragment-ion peak-removal ratio for target–decoy separation. Columns show SeqFiller score distributions when 40%, 50%, or 60% of fragment-ion peaks are removed in the NovoBoard decoy-spectrum construction; rows show replicates Rep1, Rep2, and Rep3. Blue and gray histograms are target and decoy spectra, respectively, and the red vertical line is the nominal 1% false discovery rate (FDR) score threshold. The 40% setting gives the clearest separation between high- and low-confidence target predictions, whereas the larger removal fractions compress the score distribution and reduce discrimination. This calibration is specific to the SeqFiller scoring distribution and its amino-acid-level acceptance rule; it does not estimate peptide-level or named-PTM FDR. Reproduced from Fig. S12 of the RNovA supplementary material [28].

1% FDR" refers only to the separately estimated structural thresholds unless the prediction level is stated explicitly.

6.4.2 Validation with ^{15}N-Labeled Yeast

Target–decoy calibration is useful only if its error estimates correspond to observable behavior. RNovA therefore evaluates accepted predictions on paired unlabeled and ^{15}N-labeled *Saccharomyces cerevisiae* data (Fig. 6.6) [137]. Incorporation of ^{15}N shifts a fragment according to the number of nitrogen atoms it contains. A prediction made from the unlabeled spectrum therefore implies a structured family of mass shifts in the labeled measurement. Matching the predicted labeled fragments to observed peaks supplies a check based on isotope physics rather than on agreement with the de novo model's original label.

For each accepted PathSearcher prediction, plausible nitrogen counts for residue mass gaps are constructed using an in silico yeast digest. Candidate labeled fragments are shifted by multiples of 0.9970 Da per nitrogen atom and compared with MS1 and MS2 evidence in the labeled run. Fragment matches within 0.02 Da provide a mismatch statistic. Because a correct prediction can fail this isotope check, the analysis first estimates the validation procedure's false-negative rate, α, using Comet identifications controlled at 1% sequence-level FDR. With the probability that an incorrect prediction is declared a mismatch conservatively set to $\beta = 1$, the correction can be written as

$$P^{\mathrm{DB}}_{\mathrm{mismatch}} = \alpha(1 - 0.01) + \beta(0.01), \tag{6.1}$$

$$P^{\mathrm{DN}}_{\mathrm{mismatch}} = \alpha(1 - P^{\mathrm{DN}}_{\mathrm{error}}) + \beta P^{\mathrm{DN}}_{\mathrm{error}}. \tag{6.2}$$

Solving the second relation after estimating α gives a corrected empirical error rate for the de novo predictions. Across three biological replicates, the reported values were $1.7\% \pm 1.6\%$ for PathSearcher and $0.6\% \pm 0.4\%$ for SeqFiller, consistent within the reported variation with the intended 1% thresholds. This is materially stronger than showing target and decoy score separation alone: it tests the accepted structural predictions against a new measurement channel.

The check is nevertheless not entirely database-free. A yeast proteome lookup is used to enumerate plausible nitrogen counts, and database-search identifications are used to estimate α. The correct conclusion is that isotope labeling provides an external empirical check of the target–decoy thresholds in this experiment, not that every element of the check is independent of prior sequence knowledge. Nor do the reported values establish calibration for whole peptides or named PTMs.

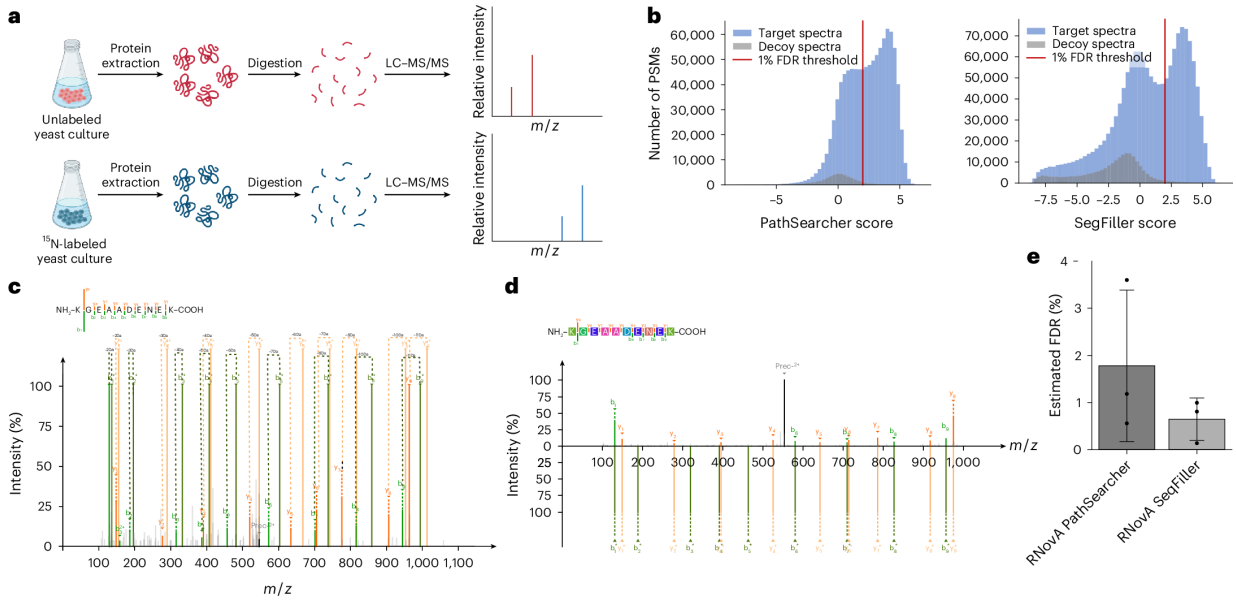


Figure 6.6: Isotopic FDR validation with a ^{15}N-labeled yeast proteome. Panel a summarizes the paired experiment: *Saccharomyces cerevisiae* was grown in unlabeled and ^{15}N-labeled media, then proteins from both conditions were extracted, digested, and analyzed by liquid chromatography–tandem mass spectrometry (LC–MS/MS) under identical conditions. Panel b shows target (blue) and decoy (gray) peptide–spectrum match (PSM) score distributions for PathSearcher (left) and SeqFiller (right); red vertical lines mark the nominal 1% FDR thresholds. Panel c shows an unlabeled MS/MS spectrum and its RNovA de novo prediction; theoretical b^* and y^* fragment ions for the labeled peptide are obtained by applying a 0.9970-Da shift for each nitrogen atom. Panel d matches these theoretical ^{15}N fragment ions to observed peaks in the paired labeled spectrum, and the matched-fragment support is used in the isotope-based empirical check. Panel e reports the resulting estimated FDRs for PathSearcher and SeqFiller; bars are means ± s.d., points are the three biological replicates ($n = 3$). Confidence thresholds and error estimates are defined separately for PathSearcher nodes and SeqFiller amino acids; the figure does not validate peptide-level or named-PTM FDR. Reproduced from the RNovA article [28].

6.5 End-to-End Discovery of Nonnatural Modifications

The ProteomeTools benchmark tests zero-shot reconstruction after candidate modified residues have been supplied. To evaluate the complete discovery workflow, RNovA uses four nonnatural TestMods that do not occur in natural proteomes. Twenty-four modified peptides were synthesized, eight per injection group, with the groups designed to avoid precursor-mass overlap. The peptides were measured alone to establish chromatographic references and were also spiked into a tryptic digest of 293T human cells. The mixed sample tests whether the unusual signals can be recovered in a realistic background rather than only in an isolated standard run.

This experiment exercises the whole computational chain. PathSearcher is not given TestMod residue tokens. It produces high-confidence paths, from which the pipeline extracts and clusters unusual mass tags. All four TestMod classes were detected under the reported structural confidence threshold. An MSFragger open search recovered 11 PSMs corresponding to one TestMod class and did not recover the other three in this experiment [45]. This comparison is informative because conventional open search begins with a database peptide scaffold, whereas RNovA can first recover the mass structure without requiring that scaffold. It compares the methods under this experimental configuration rather than across all possible open-search settings.

Once the four residue masses were available as candidates, SeqFiller was evaluated on complete reconstruction. A spectrum was eligible for evaluation when its precursor mass and retention time matched a known synthetic peptide; a correct prediction had to recover the full peptide sequence. Reported recall for the three groups was 59.3%, 93.8%, and 73.1%, while precision was 99.3%, 99.4%, and 99.3%, respectively (Fig. 6.7). The variation in recall is important. Open reconstruction can be highly specific while still missing spectra because the modified peptide was weakly sampled, fragmented poorly, or represented by too few informative peaks. Discovery performance cannot be summarized by precision alone.

The result provides unusually direct evidence for the division between the two modules. PathSearcher detects unfamiliar mass structure without a TestMod list; SeqFiller then uses the discovered candidates to recover complete sequences without changing its weights. The experiment also supplies controlled negatives: unrelated 293T spectra should not acquire TestMod residues merely because those residues were added to the candidate set. Greater than 99% reported precision shows that candidate expansion did not cause indiscriminate insertion of the new masses in this dataset.

The concentration series in Fig. 6.8 separates specificity from sampling. At 4, 40, and 400

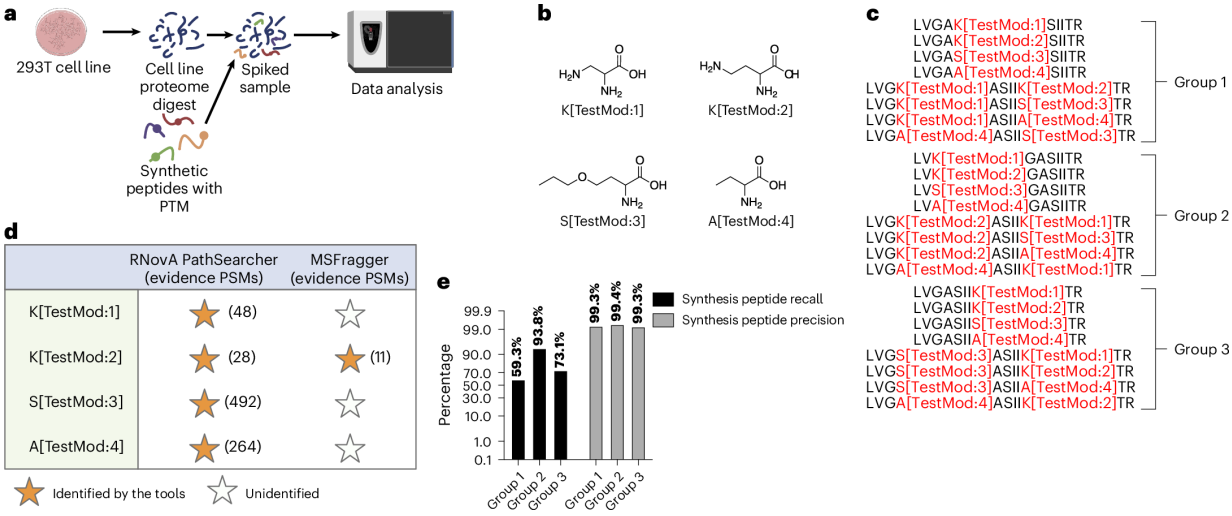


Figure 6.7: Benchmarking the RNovA modules for open PTM search with spiked-in synthetic modified peptides. Panel a shows the experiment: tryptic peptides from a 293T cell lysate were mixed with synthetic peptides carrying nonnatural PTMs and analyzed by LC–MS/MS; separate synthetic-only runs supplied retention-time references. Panel b gives the structures of K[TestMod:1], K[TestMod:2], S[TestMod:3], and A[TestMod:4], none of which occurs in a natural proteome. Panel c lists the 24 synthetic peptide sequences in three injection groups of eight; precursor masses do not overlap within a group, permitting unambiguous matching and recall evaluation. Panel d compares PTM evidence from RNovA PathSearcher and MSFragger. Filled and open stars denote identified and unidentified TestMods, and parenthetical values are evidence-PSM counts: PathSearcher recovered 48, 28, 492, and 264 PSMs for TestMods 1–4 under its reported node-level acceptance threshold, whereas MSFragger recovered 11 PSMs for TestMod:2 and did not recover the other three classes. Panel e evaluates candidate-conditioned SeqFiller on the same data. Synthesis-peptide recall is the fraction of retention-time- and precursor-mass-matched spectra for which the complete known synthetic sequence is recovered; synthesis-peptide precision is computed from unmatched background spectra by testing whether any TestMod residue is introduced. Recall for groups 1–3 is 59.3%, 93.8%, and 73.1%, and precision is 99.3%, 99.4%, and 99.3%, respectively. This controlled experiment uses four predesigned chemistries and known synthetic ground truth to test end-to-end mass discovery and reconstruction. Reproduced from the RNovA article [28].

fmol, reported precision remained above 99.89% across groups. Recall generally increased with abundance and with the number of precursor-mass- and retention-time-matched spectra. Each condition contained only 10–21 candidate spectra, so individual recall values are noisy and group-specific ionization contributes to the differences. The supported conclusion is that false TestMod insertion remained rare across the tested concentrations; the experiment does not show that sensitivity is independent of concentration.

The TestMod experiment demonstrates detection of masses absent from both the trained vocabulary and natural proteomes, localization through the fragmentation path, candidate-conditioned sequence reconstruction, and structural confidence filtering. Because the four chemical structures were designed in advance, the experiment begins with known chemical ground truth. The real-sample studies begin instead with recurrent mass evidence whose chemical interpretation is initially unknown.

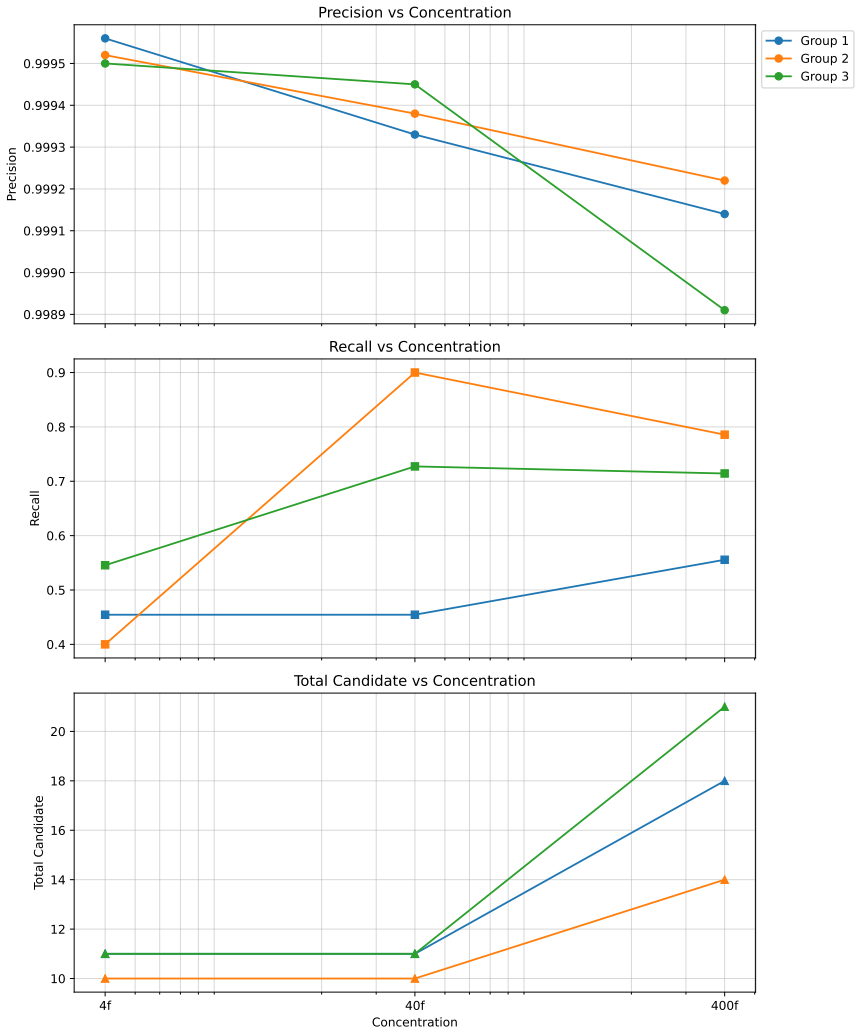


Figure 6.8: RNovA performance in the selected TestMod spike-in validation across 4, 40, and 400 fmol, spanning two orders of magnitude, for peptide groups 1–3. The top panel reports synthesis-peptide precision, which remains above 99.89% for every displayed group and concentration. The middle panel reports synthesis-peptide recall; recall generally increases with the amount of acquired evidence but differs among groups, consistent with group-specific ionization and sampling. The bottom panel gives the number of candidate spectra whose retention time and precursor mass match a synthetic peptide; only 10–21 spectra are available per group–concentration condition. The selected spike-in experiment therefore shows rare false TestMod insertion under these conditions, while the small candidate sets limit fine-grained conclusions about sensitivity as a function of concentration. Reproduced from Fig. S10 of the RNovA supplementary material [28].

6.6 Open PTM Discovery in Biological Samples

Real discovery reverses the controlled experiment. The system does not begin with a known TestMod structure. It begins with recurrent mass evidence and must decide which candidates merit chemical interpretation and experimental follow-up. RNovA demonstrates this process through a public human-proteomics discovery screen, independent rheumatoid arthritis (RA) synovial-tissue samples, and the bacterial strain A1232E. These applications remove different kinds of prior knowledge. The human studies have a reference proteome but seek modified states not enumerated by ordinary sequence databases. The bacterial study lacks a usable strain-specific reference proteome, so even the peptide sequence cannot be supplied by database search.

Proteogenomic workflows can enlarge sequence catalogues with genomic evidence, but their statistical search still operates over a constructed candidate space [138]. The A1232E analysis asks what sequence and modification evidence can be retained before such a strain-specific catalogue is available.

6.6.1 Rheumatoid Arthritis and Selected Kynurenine Assignments

PathSearcher mass tags from the public dataset PXD038702, which profiled several biological contexts including RA, were clustered and ranked by the number of high-confidence PSMs. The five leading cluster centers were 190.0732, 179.0590, 206.0741, 176.0320, and 83.0392 Da, supported by 9,419, 3,465, 1,510, 633, and 600 PSMs. Residue-context interpretation associated these masses with W+3.9938, Y+15.9957, W+19.9948, C+73.0228, and T−18.0085, respectively. The corresponding labels in Fig. 6.9a are explicitly putative: kynurenine, oxidation, hydroxykynurenine, an unresolved formula-level candidate, and dehydration. Cluster abundance is evidence of recurrence, not of chemical uniqueness.

The most abundant cluster in this discovery screen suggested conversion of tryptophan to kynurenine, a biologically interpretable +3.9938-Da change. Motivated by that result, RNovA was then applied separately to synovial tissue collected from patients with rheumatoid arthritis. Two high-confidence candidate sequences from those patient samples were selected:

```
NLQEAEEW[Kynurenine]YK
FQW[Kynurenine]YVDGVEVHNAK.
```

Synthetic versions were analyzed under matched LC–MS/MS conditions. For both peptides, endogenous and synthetic signals showed closely aligned retention times and near-complete fragment-spectrum overlap (Fig. 6.9e–h). A BLAST search connected the first predicted sequence exactly to a known protein and placed the second one residue away from a reported protein sequence [139].

These pieces of evidence play different roles. BLAST establishes plausible protein context, but cannot establish the PTM because the modified state is not encoded by an ordinary protein sequence. The recurrent mass and fragment path localize the change. The synthetic references test whether the proposed modified peptides behave like the endogenous signals chromatographically and spectrally. Together they support the selected kynurenine-containing peptide assignments. They do not convert the other thousands of cluster members, or every future W+3.9938 observation, into automatically confirmed kynurenine.

Retention time supplies an additional constraint that can flag inconsistent modified-peptide identifications, including modifications not seen during model training [140]. The use of measured reference peptides as a fit-for-purpose confirmation step also follows broader mass-spectrometric assay validation practice [141]. Neither form of agreement makes an accurate delta mass chemically unique, which is why the conclusion remains limited to the selected assignments.

6.6.2 A1232E: Sequence and Modification Without a Reference Proteome

The A1232E analysis tests the pipeline when a reference sequence catalogue is not available. High-confidence PathSearcher paths yielded an enriched mass tag of 172.08 Da. Residue context suggested that it could be glutamate, 129.04 Da, carrying an additional 43.04 Da. The pipeline represented this proposal as E[+43.0421] and supplied that mass to SeqFiller. Complete reconstructed sequences included `LDDKLNNKAEEE[+43.0421]GATK`.

At this stage, the result is a localized residue-mass hypothesis. Prior knowledge of bacterial PTMs made ethanolamine a plausible explanation for the +43.04-Da shift. The corresponding peptide, `LDDKLNNKAEEE[Eth]GATK`, was synthesized and measured under the same LC–MS/MS conditions. The synthetic peptide eluted at 36.96 min and the endogenous signal at 37.78 min, and their fragment spectra showed strong concordance. These observations support the reconstructed sequence, the modified glutamate site, and the proposed ethanolamine assignment in this selected example.

The sequence of commitments is essential. The model output at this point was a mass hypothesis, not a chemical name. RNovA found a recurrent mass, localized it in a partial structure, inferred a residue and delta mass, reconstructed a full peptide using that candidate, then used prior chemistry to choose a standard for experiment. A +43.04-Da change is not universally unique to ethanolamine; the synthetic comparison is what makes the assignment substantially stronger than a database mass match alone.

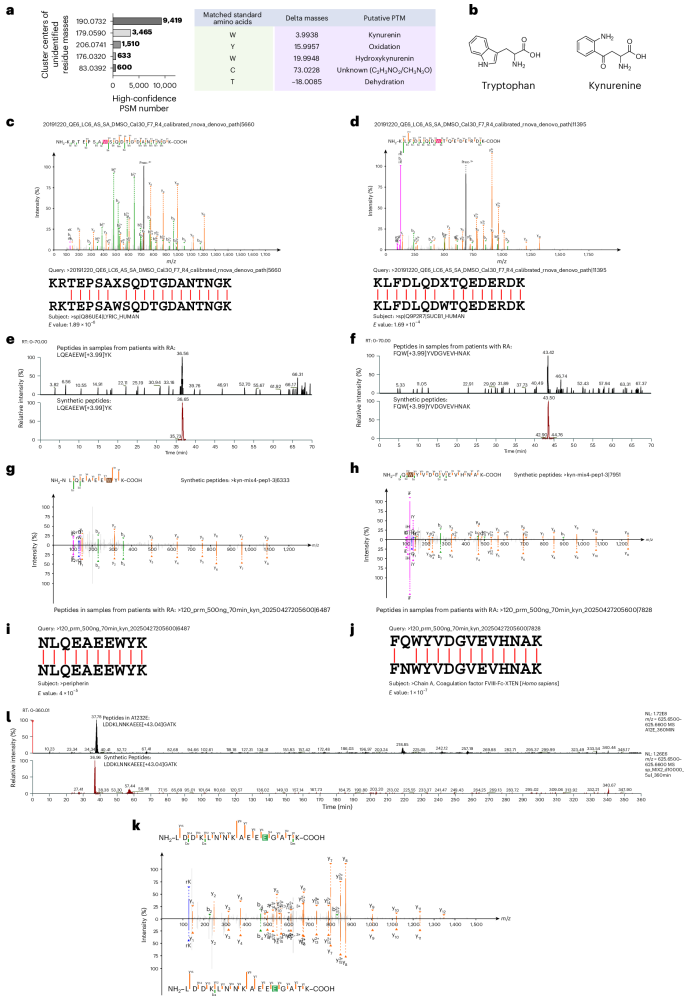


Figure 6.9: Open discovery and selected validation of PTM hypotheses. Panel a ranks the five leading mass-shift clusters in the public dataset PXD038702 by high-confidence PSM count and reports putative residue, delta-mass, and chemical interpretations. Panel b shows tryptophan and kynurenine structures; panels c and d show two high-confidence kynurenine-containing predictions from the discovery analysis. RNovA was then applied separately to independently collected rheumatoid arthritis (RA) synovial tissue, and two candidates were selected for synthetic-reference testing. Panels e and f compare extracted-ion chromatograms for endogenous NLQEAEEW[Kynurenine]YK and FQW[Kynurenine]YVDGVEVHNAK (top) with matched synthetic peptides (bottom); panels g and h compare their fragment-ion spectra. Panels i and j show BLAST alignments, with an exact protein match for the first sequence and one residue substitution in the second. Panels k and l compare retention time and MS/MS evidence, respectively, for a selected A1232E E[+43.04] peptide and its synthetic ethanolamine-containing counterpart. Panel a is the public-dataset screen, whereas panels e–j use separate RA patient samples; the standards support only the selected assignments. Reproduced from the RNovA article [28].

6.7 Generalization Across Acquisition Platforms

The principal evidence was obtained largely from collision-based, DDA-style data, with substantial use of Orbitrap HCD [36]. Within Orbitrap spectra collected at normalized collision energies 25, 30, and 35, reported metrics were stable over the tested range. Cross-platform transfer was less uniform. On *E. coli*, amino-acid precision/recall was 64.2%/64.1% on Thermo QE-HFX and 61.8%/62.4% on SCIEX TripleTOF 6600+, but 40.9%/40.8% on Bruker timsTOF Pro; peptide recall on timsTOF was 12.1%. The supplement attributes part of this gap to the different parallel accumulation–serial fragmentation (PASEF) acquisition distribution and ion-mobility separation [142], and to more frequent chimeric spectra, which are known to complicate peptide identification [10]. The result indicates that relative-mass structure transfers more effectively between the evaluated beam-type fragmentation systems than to the PASEF acquisition distribution and motivates instrument-specific training or domain adaptation (Fig. 6.10). Proteomics-specific spectrum prediction must account for both DDA and DIA regimes [143], while few-shot adaptation can recover part of an instrument or fragmentation mismatch [144]. This is motivation for adapting RNovA, not evidence that the adaptation has already been performed.

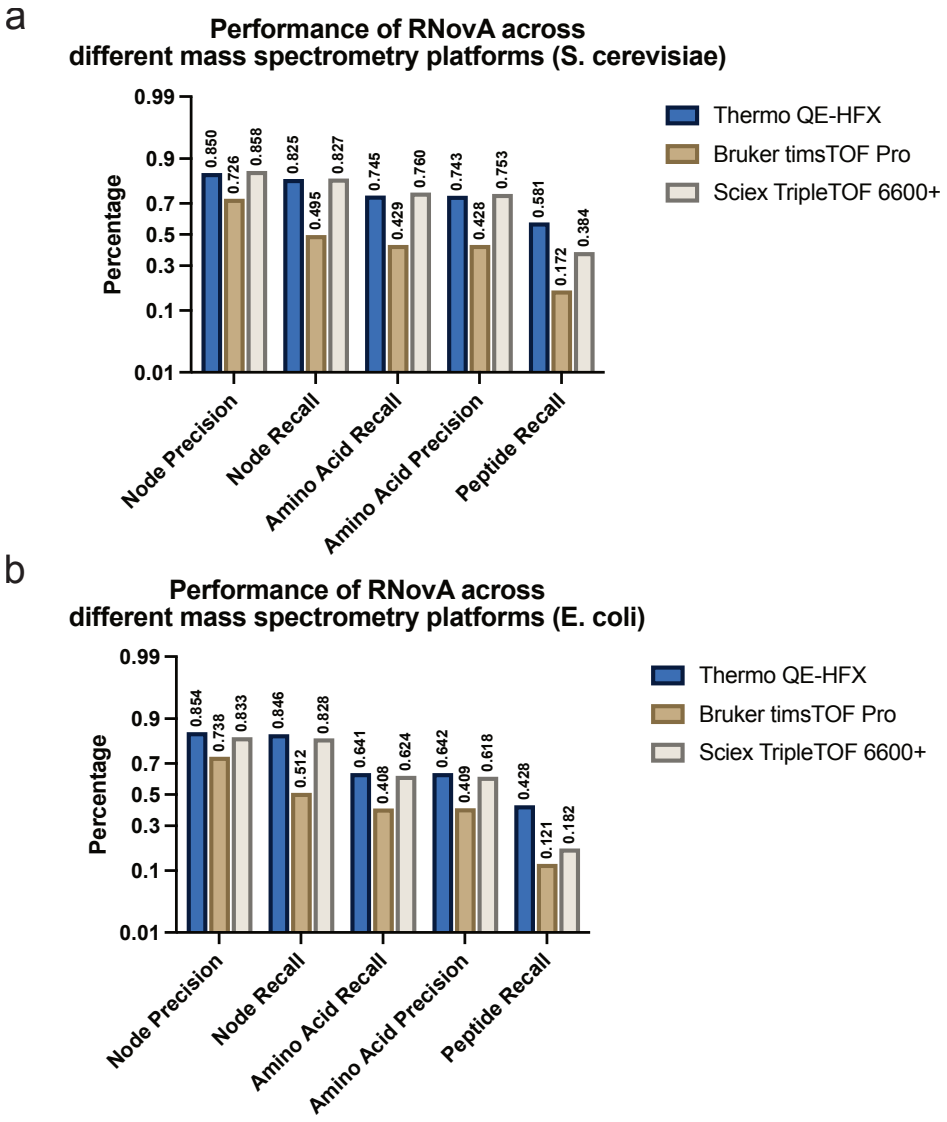


Figure 6.10: RNovA performance across mass spectrometry platforms without instrument-specific retraining. Panel a reports PathSearcher node precision and recall and SeqFiller amino-acid precision, amino-acid recall, and peptide recall for *S. cerevisiae* spectra from PXD028735; panel b reports the same metrics for *E. coli*. Both panels compare Thermo QE-HFX, Bruker timsTOF Pro, and SCIEX TripleTOF 6600+. The lower timsTOF results indicate acquisition-domain shift. Reproduced from Fig. S8 of the RNovA supplementary material [28].

6.8 Discussion

RNovA separates open candidate discovery from candidate-conditioned sequence reconstruction. PathSearcher can retain an interval whose mass is absent from the training vocabulary, and recurrence across spectra converts such intervals into a bounded candidate set. SeqFiller can score an unseen modified residue without retraining, but only after its mass candidate has been supplied. The automated pipeline normally advances the three most abundant candidates. A modification that is too rare to form a cluster, lies outside that shortlist, or is removed by the alignment and significance filters will not reach sequence reconstruction. Samples containing many concurrent PTM classes may therefore require broader or iterative candidate selection.

Localization also depends on the measured fragments. A mass tag may span several residues when fragmentation is missing, and noise or chimeric spectra can create an incorrect path. Recurrence and structural FDR reduce these risks but cannot replace unobserved evidence. The lower citrullination site-identification rate and the concentration-dependent TestMod recall show how localization and sequence recovery vary with the available spectrum information.

Confidence is attached to the structural decisions that RNovA directly scores. The 40% decoy peak-replacement setting was calibrated for the RNovA scores, and the ^{15}N analysis checks node- and amino-acid-level thresholds in the reported yeast experiment. These estimates are not peptide-, PTM-candidate-, or chemical-identity-level FDR, and calibration under a new score, spectrum-quality distribution, sample type, or instrument requires renewed empirical evaluation.

Accurate mass narrows chemical interpretation without making it unique. Isomers, residue substitutions, combinations of smaller changes, adducts, and sample-processing artifacts can share or approximate a delta mass. The RA and A1232E studies therefore combine recurrent mass evidence, residue context, complete sequence reconstruction, and synthetic-reference comparison for selected assignments. Retention-time and fragment-spectrum agreement support those tested peptides; other members of a mass cluster remain residue-mass candidates unless they receive corresponding evidence.

The complete workflow processes target and decoy spectra, searches an unrestricted mass-transition graph, clusters mass-tag sequences, and runs candidate-conditioned reconstruction. PathSearcher and SeqFiller inference dominate the reported H100 runtime. Public code and data make this computational trade-off inspectable [145, 146]. Together, the experiments establish a computational route from an unexpected spectral mass to a localized, sequence-resolved and experimentally testable PTM candidate, while leaving complete chemical-

structure determination and acquisition-independent calibration to the measurements and validation designed for those questions.

6.9 Conclusion

RNovA provides an end-to-end route from an unexplained fragment mass to a localized and sequence-resolved PTM candidate. PathSearcher recovers mass paths without restricting every transition to a trained residue vocabulary, recurrent tags are converted into residue–mass-shift candidates, and candidate-conditioned SeqFiller reconstructs complete modified peptides without PTM-specific retraining.

The evidence supporting these outputs is deliberately separated by resolution. Target–decoy spectra provide node- and amino-acid-level thresholds, paired ^{15}N yeast checks those thresholds under the reported conditions, and controlled TestMods test masses withheld from training. For selected RA and A1232E candidates, synthetic references add evidence about retention and fragmentation that the generating model score cannot supply.

Together, the conventional benchmarks, modified-peptide transfer experiments, target–decoy calibration, isotope-labeled yeast, nonnatural TestMods, clinical samples, and reference-free bacterial analysis show that previously unseen residue masses can remain expressible throughout sequencing and can be advanced to experimentally testable chemical hypotheses. RNovA thereby extends *de novo* sequencing from prediction over a fixed chemical alphabet to the discovery and evaluation of molecular candidates outside that alphabet.

Chapter 7

Discussion: From Unexplained Spectra to Molecular Discovery

7.1 Unexplained spectra and the limits of closed identification

An unexplained spectrum is not necessarily an uninformative one. It may lack the fragments required for a unique identification, but it may also contain coherent evidence for a molecule that a database search or a fixed output vocabulary cannot express. Conventional workflows often return the same outcome in both cases—no accepted match—even though the computational consequences are different. The first calls for an honest account of uncertainty; the second calls for a way to formulate a new hypothesis. The question running through this dissertation is what useful molecular evidence can be carried forward when a familiar label is unavailable.

This question changes the role of a database. In a closed identification task, the database defines the candidates that can be scored, and the absence of an accepted match ends the analysis. In a discovery task, the same database remains valuable as a record of known sequences and chemistry, but its boundary cannot be equated with the boundary of the sample. An unmatched observation may lie outside current knowledge rather than outside molecular reality. The difficulty is that a computational method cannot know this from novelty alone. Noise, poor fragmentation, interference, an incorrect precursor, and an unexpected molecule can all produce disagreement with a reference.

Open-world recognition provides a useful comparison because it distinguishes known

classes from inputs that do not belong to them [147, 148]. The molecular problem is harder than rejecting an image from an unfamiliar class. A spectrum is a partial consequence of a physical process, not a complete instance waiting to be assigned a hidden name. The unknown object must be reconstructed from fragment relationships, and several objects can be compatible with the same incomplete observation. A scientifically useful output must therefore express both a hypothesis and the remaining ambiguity.

This places the problem at the intersection of structured prediction and open-world recognition. Structured-prediction methods model dependencies within outputs such as sequences, trees, and sets, and latent-variable extensions allow useful intermediate structure to remain unobserved; both nevertheless normally define the output representation in advance [149, 150]. Here, the measurement and the usable vocabulary can both be incomplete. A spectrum may support an ordered set of mass checkpoints, a partial composition, or a localized mass shift without supporting a unique sequence or chemical name. Output resolution is therefore part of inference: the scientifically appropriate prediction may be a coarser structured statement whose unresolved degrees of freedom are retained for later evidence, rather than the most probable complete object.

Tandem spectra make this requirement concrete. Their peaks reveal mass relationships rather than amino-acid symbols directly. A missing pair of adjacent fragments does not imply that the corresponding residues are absent; it means that the spectrum does not determine their internal order at the usual resolution. Likewise, a glycopeptide spectrum may contain fragments that support a composition even though no single linear sequence of fragments captures the evidence. The distinction between an absent molecule and an absent computational representation begins here: evidence can be present in the measurement yet invisible to a model whose output or intermediate states cannot express it.

The most important negative conclusion is consequently that neither a database mismatch nor a confident generated answer constitutes a discovery. A mismatch states only that the current search failed to accept a known candidate. A neural score states only how the model ranks an output under its learned representation. Discovery requires an account of why the measurement remains informative after familiar explanations fail, what alternative is being proposed, and what observation could show that the proposal is wrong.

7.2 Constructing hypotheses from partial evidence

In a linear peptide, an unsupported interval can corrupt subsequent residues if an autoregressive decoder must commit to an exact local sequence. Treating the interval as

a mass-constrained gap preserves later checkpoints even while its internal order remains unknown. The intervention experiments reported with GraphNovo show why this matters: path guidance restores much of the amino-acid recall lost after a missing-fragmentation region in two otherwise independent decoders [11]. The path does not reconstruct ions that were never measured. It prevents an early guess from erasing constraints that the spectrum still provides elsewhere.

A branched glycan presents a different version of the same epistemic problem. Relevant fragments need not form one consecutive ladder, and the order used by an autoregressive model is not the molecular topology. Partial-composition masses allow evidence associated with several branches to contribute to the next decision. The contrast between raw and filled glycan gaps shows that some apparent absence is created by the representation: a gap that is large on one path can become small after fragment states implied by supported partial compositions are considered. When the filled gap remains large, performance still declines, confirming that reorganizing evidence cannot create information that the instrument did not measure [12].

These examples lead to a practical rule without implying a universal data structure. A useful representation should preserve the strongest statement supported by the measurement and leave finer distinctions unresolved. For a peptide this can be an ordered set of cumulative-mass checkpoints. For a glycan it can be a composition and the masses of its partial combinations. Both are abstractions: a mass gap omits internal residue order, and a composition omits linkage and topology. Their scientific value lies precisely in making those omissions visible rather than silently replacing them with a more specific answer.

The same approach becomes consequential when the expected molecule lies outside the current vocabulary. A transition unexplained by known residues need not be discarded or forced into the nearest familiar class. It can be retained as a localized residual mass, compared across spectra, and promoted to a modification candidate when recurrence and fragment evidence support it. Complete sequence reconstruction can then proceed conditionally on that candidate. RNovA uses this separation to handle residue–mass shifts absent from its trained PTM classes without asking the sequence decoder to search an unconstrained space of chemical structures [28].

Prior knowledge is not removed in this process. Precursor mass, fragment connectivity, residue context, mass tolerance, and learned fragmentation regularities continue to constrain the answer. What changes is their logical role. Known chemistry is used to interpret and bound hypotheses rather than to declare in advance that no other hypothesis can exist. This is why an unexplained mass can become computationally useful before it has a chemical name.

The distinction also clarifies what an open PTM workflow establishes. The computational result is a route from an unexplained spectral transition to a localized mass-shift hypothesis and then to a reconstructed modified peptide. It is not, by itself, a unique chemical-structure assignment. Isobaric modifications, alternative sites, and unsupported intervals can remain unresolved even when precursor mass and many fragment ions agree. The output is scientifically useful because it makes those alternatives explicit and gives later experiments something precise to test.

7.3 Reliability of open molecular inference

Once an unknown can be expressed, a second problem appears: a larger hypothesis space can turn noise into plausible answers. Confidence cannot simply be inherited from a conventional test set or read directly from a model score. Large peptide datasets are commonly labeled by database search, so they favor spectra that a database and search engine already explain. They also inherit choices about species, digestion, acquisition, filtering, and the candidate database. A model may perform well because its test data resemble its training distribution even when its apparent certainty weakens on the unfamiliar spectra for which *de novo* analysis is most valuable.

The NovoBoard experiments make this dependence visible. Peptide, amino-acid, and fragment-ion accuracy describe different properties, and comparisons change across species, digestion regimes, peptide classes, mutation counts, and proximity to the training data. Target–decoy spectra provide controlled negative examples for calibrating scores when complete sequence labels are unavailable, but their usefulness depends on the perturbation. A decoy that destroys too much fragment structure is easy to reject and can make error estimates optimistic; a decoy that remains too similar can make them conservative. The appropriate construction changes with the model, score, and data [27].

This result is more than a warning about benchmark design. It means that opening an output space also changes the null hypothesis against which predictions must be judged. Error control cannot be treated as a universal threshold transferred unchanged between instruments, datasets, or models. The negative examples must preserve enough nuisance structure to challenge the scoring function while disrupting the molecular relationships whose presence is supposed to justify acceptance.

This reliability problem is related to calibration and selective prediction. Conventional probability calibration asks whether reported confidence corresponds to empirical correctness [151]. Selective prediction instead trades coverage for lower risk by abstaining on uncertain

cases [152, 153]. Open molecular inference must additionally choose the resolution at which an answer remains defensible. The same spectrum may justify a fragment path or a localized mass shift without justifying a complete peptide or chemical identity. Risk must therefore be attached to the claim actually emitted, and the system should abstain from the unsupported part rather than discard every supported relation. The target–decoy calibration used here is not probability calibration: model-specific decoy spectra act as computational negative controls that test whether a score responds to the molecular structure claimed by the method rather than merely to familiar nuisance statistics [15, 27]. This objective has a limited analogy to knockoff procedures, which construct synthetic null controls that preserve the observed covariate correlation structure; NovoBoard decoys are not statistical knockoffs and do not inherit their finite-sample guarantees [154]. Because decoy difficulty depends on the model and data, calibration is part of defining the test, not a certificate appended after prediction.

RNovA applies this principle to the decisions directly scored by its two modules. Target–decoy spectra estimate error for PathSearcher nodes and SeqFiller amino acids, and paired ^{15}N-labeled yeast provides an external empirical check under the reported conditions [28]. These estimates support populations of nodes and residues. They should not be relabeled as peptide-level, PTM-candidate-level, or chemical-identity-level FDR. A calibrated structural decision narrows the space of explanations; it does not by implication prove one chemistry.

For an individual candidate, credibility increases when the kind of evidence changes, not merely when more scores from the same model are accumulated. Recurrence across spectra supports repeated detection. Site-determining fragments support localization. Residue context, known biochemical reactions, and exact mass can suggest chemical interpretations. Synthetic standards can test whether retention time and fragmentation agree with a selected assignment. These sources answer different questions and are not interchangeable.

The controlled TestMod study, the isotope-labeled experiment, the rheumatoid arthritis analysis, and the bacterial example illustrate this division of labor. Artificial modifications test whether masses withheld from training can be recovered. Isotope labeling tests structural error estimates. Reference-proteome-free analysis tests whether the workflow can proceed when a conventional sequence catalog is unavailable. Synthetic peptides provide strong evidence for selected sequence, site, and chemical assignments. None of these experiments estimates prospectively how often every generated candidate will be confirmed. Their combined force comes from stating which uncertainty each one addresses [28].

This is also why a population-level FDR and a chemical validation should remain separate in the scientific report. The former estimates how many accepted decisions are expected to be wrong under a defined competition procedure. The latter asks whether one

proposed molecule behaves like an independently prepared reference. A discovery claim should stop at the strongest level for which the relevant evidence has actually been obtained.

7.4 Scope, implications, and future work

The empirical conclusions are bounded by the experiments from which they were drawn. Most peptide results come from bottom-up, data-dependent spectra acquired with collision-based fragmentation, predominantly on Orbitrap platforms. Performance under substantially different instrument and acquisition conditions shows that transfer cannot be assumed. Graph construction, candidate-ion generation, mass tolerances, and score calibration may all need to change when the physical measurement changes.

The molecular output spaces remain bounded as well. GraphNovo uses a predefined residue and modification vocabulary. DeepGlycan predicts compositions within five modeled mammalian monosaccharide classes; it does not infer linkage, anomeric configuration, or complete topology, and its N-to-O transfer result depends on a shared alphabet and compatible preprocessing. RNovA permits unexplained masses during candidate discovery, while SeqFiller reconstructs sequences from a candidate set. Multiple simultaneous unknown modifications, long unsupported intervals, isobaric alternatives, and unrestricted chemical-structure inference remain difficult.

Ground truth is another limitation. Database-derived labels make large-scale training and comparison possible, but they are selected by the same kind of closed candidate system that discovery seeks to move beyond. Removing exact peptide overlap reduces one source of leakage without eliminating similar sequences, shared fragmentation regularities, or bias in which spectra receive confident labels. Negative controls and isotope experiments reduce dependence on those labels for specific questions; they do not create an exhaustive inventory of true unknown molecules.

The broader relevance to artificial intelligence for biology follows from these limits rather than from ignoring them. Many biological models are trained against curated entities even though physical measurements can contain phenomena outside those collections. A system capable only of selecting known labels treats the collection boundary as rejection. Scientific machine learning provides a broad precedent for embedding physical laws and constraints into learning rather than asking data alone to recover them [155]. More specifically, combining strong inductive biases with symbolic regression can distill compact, interpretable relations from learned predictors [156]. The case developed here makes a complementary point for molecular evidence: mass and fragmentation constraints do not

identify chemistry on their own, but they can keep an unknown expressible as a testable hypothesis and make explicit the observations needed to distinguish it from alternatives. This is a concrete role for AI in discovery without implying that model confidence is itself scientific validation.

The argument may extend to metabolite annotation, variant interpretation, and structural analysis, where several molecular explanations can fit a partial observation. Such transfer has not been demonstrated by the experiments in this dissertation. What transfers is a research question rather than a finished architecture: when the reference explanation fails, what structure remains in the measurement, how can that structure constrain a new hypothesis, and what independent observation would distinguish the hypothesis from its alternatives?

A decisive next step for open PTM discovery is prospective evaluation. Candidate generation, ranking, synthesis, and acceptance criteria should be specified before analysis and applied across laboratories and instruments. Such a study would measure not only precision among reported sequences but also the rate at which a high-ranking unexplained mass yields a reproducible chemical assignment. It would expose failures of candidate prioritization and experimental follow-up that retrospective selected validations cannot estimate.

Candidate prioritization itself is an open computational problem. Earlier work on automated science demonstrated a related hypothesis–experiment cycle, although the workflow in this dissertation remains human-directed rather than autonomous [157]. Spectral support could be combined with recurrence across samples, localization strength, biological context, feasibility of synthesis, and the expected information gained from a follow-up experiment [158]. A system could then choose not merely the candidate with the highest current score, but the experiment most likely to distinguish among competing explanations. This would connect machine learning with experimental design while keeping both model confidence and domain prior knowledge open to revision.

Future representations should also move beyond the present molecular alphabets without hiding uncertainty. Peptide models could incorporate elemental-composition constraints and fragmentation-aware candidate chemistry. Glycan inference should progress from composition toward topology and linkage only when the measurement supports those distinctions. Joint reasoning over sequence, modification, glycan, isotope pattern, retention time, and orthogonal measurements may constrain ambiguity more effectively than any single source, but it will also require new error models and new negative controls.

The broader contribution is therefore not a universal representation or decoder. It is the demonstration that a failed database match need not be the end of analysis. When

computation retains the structure still present in a spectrum, the unexplained remainder can be carried forward as an explicit molecular hypothesis and judged by evidence outside the model. This is the point at which prediction becomes useful for discovery.

Chapter 8

Conclusions

This dissertation asked how a learning system can contribute to molecular discovery when neither the measurement nor the available answer set is complete. Tandem mass spectra make the problem unusually explicit. They record fragments rather than molecules, leave some informative cleavages unobserved, and contain peaks that admit several interpretations. At the same time, databases and trained output vocabularies contain only molecular forms that have already been anticipated. A failed match can therefore mean that the evidence is insufficient, that the computation used the wrong representation, or that the molecule lies outside the current reference space.

The experiments support an open but constrained answer: the boundary of a database need not be the boundary of molecular inference, but an answer outside that boundary must remain accountable to the measurement. Across the peptide and glycopeptide studies, uncertainty was most useful when it remained visible. An unresolved interval, a dispersed set of partial-composition fragments, or a residual mass did not have to terminate inference, nor did it have to be hidden inside an unjustifiably complete answer. It could delimit what the spectrum still supported and become the starting point for a more specific hypothesis. The evaluation experiments add the complementary warning that a larger answer space also creates more plausible explanations for noise; accuracy and target–decoy calibration remain conditional on the data, score, negative-control construction, and definition of correctness [11, 12, 27].

The open PTM workflow is the empirical culmination of this idea. Mass tags absent from the trained PTM classes could be retained and localized, recurrent tags could be clustered into residue–mass-shift candidates, and complete modified peptides could be reconstructed conditionally on those candidates without PTM-specific retraining. FDR at

the PathSearcher node level and the SeqFiller amino-acid level was checked with isotope-labeled data, while selected chemical assignments were tested against synthetic reference peptides [28]. This demonstrates that molecular inference can move beyond a fixed reference space under the evaluated conditions. It does not imply unrestricted chemical identification, and it does not turn a mass shift or model score into a unique chemical identity.

The evidence remains bounded by bottom-up, predominantly data-dependent and collision-based experiments, by database-derived labels used for training and large-scale evaluation, and by the representations supported in each molecular setting. Long unsupported intervals, simultaneous unknown modifications, isobaric explanations, glycan topology, other instruments and fragmentation modes, and calibration across laboratories remain open problems. The selected biological validations also do not estimate how often an arbitrary generated candidate will be confirmed. A decisive prospective study would fix candidate generation, ranking, synthesis, and acceptance criteria before analysis and apply them across laboratories, directly measuring how often a high-ranking unexplained pattern becomes a reproducible chemical assignment.

The broader relevance to artificial intelligence in biology lies in the relationship between a model and the limits of its labels. Curated knowledge is indispensable, but it should constrain interpretation rather than serve as proof that every admissible answer is already known. When current labels fail, computation can keep an unknown scientifically accessible by preserving what the measurement supports, expressing what remains uncertain, and making the proposed explanation answerable to further evidence.

References

[1] Lloyd M. Smith, Neil L. Kelleher, and The Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. *Nature Methods*, 10(3):186–187, 2013.

[2] Ruedi Aebersold and Matthias Mann. Mass-spectrometric exploration of proteome structure and function. *Nature*, 537(7620):347–355, 2016.

[3] Mark Grimes, Benjamin Hall, Lauren Foltz, Tyler Levy, Klarisa Rikova, Jeremiah Gaiser, William Cook, Ekaterina Smirnova, Travis Wheeler, Neil R. Clark, Alexander Lachmann, Bin Zhang, Peter Hornbeck, Avi Ma'ayan, and Michael Comb. Integration of protein phosphorylation, acetylation, and methylation data sets to outline lung cancer signaling networks. *Science Signaling*, 11(531):eaaq1087, 2018.

[4] Juan Liu, Cheng Qian, and Xuetao Cao. Post-translational modification control of innate immunity. *Immunity*, 45(1):15–30, 2016.

[5] Jimmy K. Eng, Ashley L. McCormack, and John R. Yates, III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, 1994.

[6] Johannes Griss, Yasset Perez-Riverol, Steve Lewis, David L. Tabb, José A. Dianes, Noemi del Toro, Marc Rurik, Mathias Walzer, Oliver Kohlbacher, Henning Hermjakob, Rui Wang, and Juan Antonio Vizcaíno. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nature Methods*, 13(8):651–656, 2016.

[7] Donald F. Hunt, John R. Yates, III, Jeffrey Shabanowitz, Scott Winston, and Charles R. Hauer. Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences*, 83(17):6233–6237, 1986.

[8] Peter Roepstorff and Jesper Fohlman. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical Mass Spectrometry*, 11(11):601, 1984.

[9] David L. Tabb, Lori L. Smith, Linda A. Breci, Vicki H. Wysocki, Dayin Lin, and John R. Yates, III. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Analytical Chemistry*, 75(5):1155–1163, 2003.

[10] Stephane Houel, Robert Abernathy, Kutralanathan Renganathan, Karen Meyer-Arendt, Natalie G. Ahn, and William M. Old. Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *Journal of Proteome Research*, 9(8):4152–4160, 2010.

[11] Zeping Mao, Ruixue Zhang, Lei Xin, and Ming Li. Mitigating the missing-fragmentation problem in *de novo* peptide sequencing with a two-stage graph-based deep learning model. *Nature Machine Intelligence*, 5(11):1250–1260, 2023.

[12] Qianqiu Zhang, Zeping Mao, Yuling Chen, Baozhen Shan, Haiteng Deng, and Ming Li. High-accuracy glycan De Novo prediction for N- and O-linked glycopeptides across multiple fragmentation techniques. Manuscript, 2026.

[13] David N. Perkins, David J. C. Pappin, David M. Creasy, and John S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.

[14] Andrew Keller, Alexey I. Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry*, 74(20):5383–5392, 2002.

[15] Joshua E. Elias and Steven P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3):207–214, 2007.

[16] Joel M. Chick, Deepak Kolippakkam, David P. Nusinow, Bo Zhai, Ramin Rad, Edward L. Huttlin, and Steven P. Gygi. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature Biotechnology*, 33(7):743–749, 2015.

[17] Arun Devabhaktuni, Sarah Lin, Lichao Zhang, Kavya Swaminathan, Carlos G. Gonzalez, Niclas Olsson, Samuel M. Pearlman, Keith Rawson, and Joshua E. Elias.

TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. *Nature Biotechnology*, 37(4):469–479, 2019.

[18] Vlado Dancík, Theresa A. Addona, Karl R. Clauser, James E. Vath, and Pavel A. Pevzner. De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 6(3-4):327–342, 1999.

[19] Ting Chen, Ming-Yang Kao, Matthew Tepel, John Rush, and George M. Church. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 8(3):325–337, 2001.

[20] Bin Ma, Kaizhong Zhang, Chad Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 17(20):2337–2342, 2003.

[21] Bin Ma. Novor: real-time peptide *de novo* sequencing software. *Journal of the American Society for Mass Spectrometry*, 26(11):1885–1894, 2015.

[22] Ngoc Hieu Tran, Xianglilan Zhang, Lei Xin, Baozhen Shan, and Ming Li. *De novo* peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, 114(31):8247–8252, 2017.

[23] Rui Qiao, Ngoc Hieu Tran, Lei Xin, Xianglilan Chen, Ming Li, Baozhen Shan, and Ali Ghodsi. Computationally instrument-resolution-independent *de novo* peptide sequencing for high-resolution devices. *Nature Machine Intelligence*, 3(5):420–425, 2021.

[24] Melih Yilmaz, William E. Fondrie, Wout Bittremieux, Sewoong Oh, and William Stafford Noble. *De novo* mass spectrometry peptide sequencing with a transformer model. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25514–25522. PMLR, 2022.

[25] William Stafford Noble and Michael J. MacCoss. Computational and statistical analysis of protein mass spectrometry data. *PLOS Computational Biology*, 8(1):e1002296, 2012.

[26] Wout Bittremieux, Varun Ananth, William E. Fondrie, Carlo Melendez, Marina Pominova, Justin Sanders, Bo Wen, Melih Yilmaz, and William S. Noble. Deep

learning methods for de novo peptide sequencing. *Mass Spectrometry Reviews*, 45(3):507–526, 2026. Epub 2024 Nov 29.

[27] Ngoc Hieu Tran, Rui Qiao, Zeping Mao, Shengying Pan, Qing Zhang, Wenting Li, Lei Xin, Ming Li, and Baozhen Shan. Novoboard: A comprehensive framework for evaluating the false discovery rate and accuracy of *De Novo* peptide sequencing. *Molecular & Cellular Proteomics*, 23(11):100849, 2024.

[28] Zeping Mao, Chao Peng, Yuling Chen, Ping Wu, Qianqiu Zhang, Yonghan Yu, Ruixue Zhang, Lei Xin, Baozhen Shan, Haiteng Deng, and Ming Li. Zero-shot *de novo* peptide sequencing with open posttranslational modification discovery. *Nature Biotechnology*, 2026. Published online 19 May 2026.

[29] Jesper V. Olsen, Shao-En Ong, and Matthias Mann. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Molecular & Cellular Proteomics*, 3(6):608–614, 2004.

[30] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

[31] Matthias Mann and Matthias Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, 66(24):4390–4399, 1994.

[32] N. Leigh Anderson and Norman G. Anderson. The human plasma proteome: history, character, and diagnostic prospects. *Molecular & Cellular Proteomics*, 1(11):845–867, 2002.

[33] John B. Fenn, Matthias Mann, Chin Kai Meng, Shek Fu Wong, and Craig M. Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, 1989.

[34] Kermit K. Murray, Robert K. Boyd, Marcos N. Eberlin, G. John Langley, Liang Li, and Yasuhide Naito. Definitions of terms relating to mass spectrometry (IUPAC recommendations 2013). *Pure and Applied Chemistry*, 85(7):1515–1609, 2013.

[35] Michael W. Senko, Steven C. Beu, and Fred W. McLafferty. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, 6(4):229–233, 1995.

[36] Jesper V. Olsen, Boris Macek, Oliver Lange, Alexander Makarov, Stevan Horning, and Matthias Mann. Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods*, 4(9):709–712, 2007.

[37] Hanno Steen and Matthias Mann. The abc's (and xyz's) of peptide sequencing. *Nature Reviews Molecular Cell Biology*, 5(9):699–711, 2004.

[38] David M. Creasy and John S. Cottrell. Unimod: Protein modifications for mass spectrometry. *Proteomics*, 4(6):1534–1536, 2004.

[39] Vicki H. Wysocki, George Tsaprailis, Lori L. Smith, and Linda A. Breci. Mobile and localized protons: a framework for understanding peptide dissociation. *Journal of Mass Spectrometry*, 35(12):1399–1406, 2000.

[40] Kevin McDonnell, Enda Howley, and Florence Abram. The impact of noise and missing fragmentation cleavages on de novo peptide identification algorithms. *Computational and Structural Biotechnology Journal*, 20:1402–1412, 2022.

[41] Béla Paizs and Sándor Suhai. Fragmentation pathways of protonated peptides. *Mass Spectrometry Reviews*, 24(4):508–548, 2005.

[42] Nedim Mujezinovic, Günther Raidl, James R. A. Hutchins, Jan-Michael Peters, Karl Mechtler, and Frank Eisenhaber. Cleaning of raw peptide MS/MS spectra: improved protein identification following deconvolution of multiply charged peaks, isotope clusters, and removal of background noise. *Proteomics*, 6(19):5117–5131, 2006.

[43] David M. Horn, Roman A. Zubarev, and Fred W. McLafferty. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry*, 11(4):320–332, 2000.

[44] Sangtae Kim and Pavel A. Pevzner. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications*, 5:5277, 2014.

[45] Andy T. Kong, Felipe V. Leprevost, Dmitry M. Avtonomov, Dattatreya Mellacheruvu, and Alexey I. Nesvizhskii. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods*, 14(5):513–520, 2017.

[46] Henry Lam, Eric W. Deutsch, Jonathan S. Eddes, Jimmy K. Eng, Stephen E. Stein, and Ruedi Aebersold. Building consensus spectral libraries for peptide identification in proteomics. *Nature Methods*, 5(10):873–875, 2008.

[47] Siegfried Gessulat, Tobias Schmidt, Daniel P. Zolg, Patroklos Samaras, Karsten Schnatbaum, Julia Zerweck, Tobias Knaute, Johannes Rechenberger, Bernhard Delanghe, Andreas Huhmer, Ulf Reimer, Hans-Christian Ehrlich, Stephan Aiche, Bernhard Kuster, and Mathias Wilhelm. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods*, 16(6):509–518, 2019.

[48] Ari Frank and Pavel Pevzner. Pepnovo: *de novo* peptide sequencing via probabilistic network modeling. *Analytical Chemistry*, 77(4):964–973, 2005.

[49] Thilo Muth and Bernhard Y. Renard. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Briefings in Bioinformatics*, 19(5):954–970, 2018.

[50] Matthias Mann and Ole N. Jensen. Proteomic analysis of post-translational modifications. *Nature Biotechnology*, 21(3):255–261, 2003.

[51] Sean A. Beausoleil, Judit Villén, Scott A. Gerber, John Rush, and Steven P. Gygi. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature Biotechnology*, 24(10):1285–1292, 2006.

[52] Stefan K. Solntsev, Michael R. Shortreed, Brian L. Frey, and Lloyd M. Smith. Enhanced global post-translational modification discovery with MetaMorpheus. *Journal of Proteome Research*, 17(5):1844–1851, 2018.

[53] Hao Yang, Hao Chi, Wen-Feng Zeng, Wen-Jing Zhou, and Si-Min He. Open-pnovo: *de novo* peptide sequencing with thousands of protein modifications. *Journal of Proteome Research*, 16(2):645–654, 2017.

[54] Daniel J. Geiszler, Andy T. Kong, Dmitry M. Avtonomov, Fengchao Yu, Felipe da Veiga Leprevost, and Alexey I. Nesvizhskii. PTM-Shepherd: analysis and summarization of post-translational and chemical modifications from open search results. *Molecular & Cellular Proteomics*, 20:100018, 2021.

[55] Morten Thaysen-Andersen and Nicolle H. Packer. Advances in LC–MS/MS-based glycoproteomics: getting closer to system-wide site-specific mapping of the N- and O-glycoproteome. *Biochimica et Biophysica Acta (BBA) – Proteins and Proteomics*, 1844(9):1437–1452, 2014.

[56] Jiajia Li, Bo Guo, Wenqi Zhang, Shuang Yue, Shan Huang, Song Gao, Junfeng Ma, John F. Cipollo, and Shuang Yang. Recent advances in demystifying O-glycosylation in health and disease. *Proteomics*, 22(23–24):e2200156, 2022.

[57] Christian K. Frese, A. F. Maarten Altelaar, Henk van den Toorn, Dirk Nolting, Jens Griep-Raming, Albert J. R. Heck, and Shabaz Mohammed. Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. *Analytical Chemistry*, 84(22):9668–9673, 2012.

[58] Qing Yu, Bowen Wang, Zhengwei Chen, Go Urabe, Matthew S. Glover, Xudong Shi, Lian-Wang Guo, K. Craig Kent, and Lingjun Li. Electron-transfer/higher-energy collision dissociation (EThcD)-enabled intact glycopeptide/glycoproteome characterization. *Journal of the American Society for Mass Spectrometry*, 28(9):1751–1764, 2017.

[59] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems*, volume 34, pages 28877–28888, 2021.

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

[61] Noam Shazeer, Zhenzhong Lan, Youlong Cheng, Nan Ding, and Le Hou. Talking-heads attention, 2020.

[62] Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 864–873. PMLR, 2020.

[63] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8026–8037, 2019.

[64] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations*, 2019.

[65] Zeping Mao and Ruixue Zhang. GraphNovo dataset and checkpoint. Zenodo, 2023.

[66] Zeping Mao and Ruixue Zhang. AmadeusloveIris/GraphNovo: Nature machine intelligence original code. Zenodo, 2023.

[67] Cong Wu, John C. Tran, Leonid Zamdborg, Kenneth R. Durbin, Mingxi Li, Dorothy R. Ahlf, Bryan P. Early, Paul M. Thomas, Jonathan V. Sweedler, and Neil L. Kelleher. A protease for middle-down proteomics. *Nature Methods*, 9(8):822–824, 2012.

[68] Adam D. Catherman, Owen S. Skinner, and Neil L. Kelleher. Top down proteomics: facts and perspectives. *Biochemical and Biophysical Research Communications*, 445(4):683–693, 2014.

[69] Ludovic C. Gillet, Pedro Navarro, Stephen Tate, Hannes L. Röst, Nathalie Selevsek, Lukas Reiter, Robert Bonner, and Ruedi Aebersold. 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(6):O111.016717, 2012.

[70] Dalit Shental-Bechor and Yaakov Levy. Effect of glycosylation on protein folding: A close look at thermodynamic stabilization. *Proceedings of the National Academy of Sciences*, 105(24):8256–8261, 2008.

[71] Cécile Boscher, James W. Dennis, and Ivan R. Nabi. Glycosylation, galectins and cellular signaling. *Current Opinion in Cell Biology*, 23(4):383–392, 2011.

[72] Pauline M. Rudd, Tim Elliott, Peter Cresswell, Ian A. Wilson, and Raymond A. Dwek. Glycosylation and the immune system. *Science*, 291(5512):2370–2376, 2001.

[73] Katrine T. Schjoldager, Yoshiki Narimatsu, Hiren J. Joshi, and Henrik Clausen. Global view of human protein glycosylation pathways and functions. *Nature Reviews Molecular Cell Biology*, 21(12):729–749, 2020.

[74] Hayden Wilkinson and Radka Saldova. Current methods for the characterization of O-glycans. *Journal of Proteome Research*, 19(10):3890–3905, 2020.

[75] Christina M. Woo, Anthony T. Iavarone, David R. Spiciarich, Krishnan K. Palaniappan, and Carolyn R. Bertozzi. Isotope-targeted glycoproteomics (IsoTaG): a mass-independent platform for intact N- and O-glycopeptide discovery and analysis. *Nature Methods*, 12(6):561–567, 2015.

[76] Pengyu Hong, Hui Sun, Long Sha, Yi Pu, Kshitij Khatri, Xiang Yu, Yang Tang, and Cheng Lin. Glycodenovo—an efficient algorithm for accurate de novo glycan topology

reconstruction from tandem mass spectra. *Journal of the American Society for Mass Spectrometry*, 28(11):2288–2301, 2017.

[77] James Urban, Chunsheng Jin, Kristina A. Thomsson, Niclas G. Karlsson, Callum M. Ives, Elisa Fadda, and Daniel Bojar. Predicting glycan structure from tandem mass spectrometry via deep learning. *Nature Methods*, 21(7):1206–1215, 2024.

[78] Shadi Toghi Eshghi, Weiming Yang, Yingwei Hu, Punit Shah, Shisheng Sun, Xingde Li, and Hui Zhang. Classification of tandem mass spectra for identification of N- and O-linked glycopeptides. *Scientific Reports*, 6:37189, 2016.

[79] He Zhu, Luyao Liu, Zheng Fang, Hongqiang Qin, Mingming Dong, and Mingliang Ye. O-Search-Pattern: A searching tool utilizing the Y-ion pattern to enhance O-glycopeptide identification for the analysis of O-GalNAc glycoproteome. *Analytical Chemistry*, 95(26):9761–9768, 2023.

[80] Wen-Feng Zeng, Wei-Qian Cao, Ming-Qi Liu, Si-Min He, and Peng-Yuan Yang. Precise, fast and comprehensive analysis of intact glycopeptides and modified glycans with pGlyco3. *Nature Methods*, 18(12):1515–1523, 2021.

[81] Daniel A. Polasky, Fengchao Yu, Guo Ci Teo, and Alexey I. Nesvizhskii. Fast and comprehensive N- and O-glycoproteomics analysis with MSFragger-Glyco. *Nature Methods*, 17(11):1125–1132, 2020.

[82] Ling Y. Lee, Edward S. X. Moh, Benjamin L. Parker, Marshall Bern, Nicolle H. Packer, and Morten Thaysen-Andersen. Toward automated N-glycopeptide identification in glycoproteomics. *Journal of Proteome Research*, 15(10):3904–3915, 2016.

[83] Jiechen Shen, Li Jia, Liuyi Dang, Yuanjie Su, Jie Zhang, Yintai Xu, Bojing Zhu, Zexuan Chen, Jingyu Wu, Rongxia Lan, Zhifang Hao, Chen Ma, Ting Zhao, Ni Gao, Jieyun Bai, Yuan Zhi, Jun Li, Junying Zhang, and Shisheng Sun. StrucGP: de novo structural sequencing of site-specific N-glycan on glycoproteins using a modularization strategy. *Nature Methods*, 18(8):921–929, 2021.

[84] Weiping Sun, Qianqiu Zhang, Xiyue Zhang, Ngoc Hieu Tran, M. Ziaur Rahman, Zheng Chen, Chao Peng, Jun Ma, Ming Li, Lei Xin, and Baozhen Shan. Glycopeptide database search and de novo sequencing with PEAKS GlycanFinder enable highly sensitive glycoproteomics. *Nature Communications*, 14(1):4046, 2023.

[85] Wen-Feng Zeng, Guoquan Yan, Huan-huan Zhao, Chao Liu, and Weiqian Cao. Uncovering missing glycans and unexpected fragments with pGlycoNovo for site-specific glycosylation analysis across species. *Nature Communications*, 15(1):8055, 2024.

[86] Nicholas M. Riley, Stacy A. Malaker, Marc D. Driessen, and Carolyn R. Bertozzi. Optimal dissociation methods differ for N- and O-glycopeptides. *Journal of Proteome Research*, 19(8):3286–3301, 2020.

[87] Bruno Domon and Catherine E. Costello. A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate Journal*, 5(4):397–409, 1988.

[88] Yong Zhang, Xinfang Xie, Xinyuan Zhao, Fang Tian, Jicheng Lv, Wantao Ying, and Xiaohong Qian. Systems analysis of singly and multiply O-glycosylated peptides in the human serum glycoproteome via EThcD and HCD mass spectrometry. *Journal of Proteomics*, 170:14–27, 2018.

[89] Rui Zhang, Jianhui Zhu, David M. Lubman, Yehia Mechref, and Haixu Tang. GlycoHybridSeq: Automated identification of N-linked glycopeptides using electron transfer/higher-energy collision dissociation (EThcD). *Journal of Proteome Research*, 20(6):3345–3352, 2021.

[90] Weiming Yang, Punit Shah, Yingwei Hu, Shadi Toghi Eshghi, Shisheng Sun, Yang Liu, and Hui Zhang. Comparison of enrichment methods for intact N- and O-linked glycopeptides using strong anion exchange and hydrophilic interaction liquid chromatography. *Analytical Chemistry*, 89(21):11193–11197, 2017.

[91] Jialin Liu, Bo Cheng, Xinqi Fan, Xinyue Zhou, Jiankun Wang, Wen Zhou, Hengyu Li, Wenfeng Zeng, Pengyuan Yang, and Xing Chen. Click-iG: Simultaneous enrichment and profiling of intact N-linked, O-GalNAc, and O-GlcNAcylated glycopeptides. *Angewandte Chemie International Edition*, 62(36):e202303410, 2023.

[92] Daniel A. Polasky, Daniel J. Geiszler, Fengchao Yu, and Alexey I. Nesvizhskii. Multiattribute glycan identification and FDR control for glycoproteomics. *Molecular & Cellular Proteomics*, 21(3):100205, 2022.

[93] Ming-Qi Liu, Wen-Feng Zeng, Pan Fang, Wei-Qian Cao, Chao Liu, Guo-Quan Yan, Yang Zhang, Chao Peng, Jian-Qiang Wu, Xiao-Jin Zhang, Hui-Jun Tu, Hao Chi, Rui-Xiang Sun, Yong Cao, Meng-Qiu Dong, Bi-Yun Jiang, Jiang-Ming Huang, Hua-Li

Shen, Catherine C. L. Wong, Si-Min He, and Peng-Yuan Yang. pGlyco 2.0 enables precision N-glycoproteomics with comprehensive quality control and one-step mass spectrometry for intact glycopeptide identification. *Nature Communications*, 8(1):438, 2017.

[94] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[95] Weihong Xu, Jaeyoung Kang, Wout Bittremieux, Niema Moshiri, and Tajana Rosing. HyperSpec: Ultrafast mass spectra clustering in hyperdimensional space. *Journal of Proteome Research*, 22(6):1639–1648, 2023.

[96] Jianhui Zhu, Zhengwei Chen, Jie Zhang, Mingrui An, Jing Wu, Qing Yu, St. John Skilton, Marshall Bern, Ilker K. Sen, Lingjun Li, and David M. Lubman. Differential quantitative determination of site-specific intact N-glycopeptides in serum haptoglobin between hepatocellular carcinoma and cirrhosis using LC-EThcD-MS/MS. *Journal of Proteome Research*, 18(1):359–371, 2019.

[97] Julie Van Coillie, Morten A. Schulz, Arthur E. H. Bentlage, Noortje de Haan, Zilu Ye, Dionne M. Geerdes, Wim J. E. van Esch, Lise Hafkenscheid, Rebecca L. Miller, Yoshiki Narimatsu, Sergey Y. Vakhrushev, Zhang Yang, Gestur Vidarsson, and Henrik Clausen. Role of N-glycosylation in FcγRIIIa interaction with IgG. *Frontiers in Immunology*, 13:987151, 2022.

[98] Ana García-García, Laura Ceballos-Laita, Sonia Serna, Raik Artschwager, Niels C. Reichardt, Francisco Corzana, and Ramon Hurtado-Guerrero. Structural basis for substrate specificity and catalysis of α1,6-fucosyltransferase. *Nature Communications*, 11(1):973, 2020.

[99] Ajit Varki, Richard D. Cummings, Jeffrey D. Esko, Hudson H. Freeze, Pamela Stanley, Jamey D. Marth, Carolyn R. Bertozzi, Gerald W. Hart, and Marilynn E. Etzler. Symbol nomenclature for graphical representations of glycans. *Glycobiology*, 25(12):1323–1324, 2015.

[100] Qi Zhang, Cheng Ma, Lih-Shen Chin, Sheng Pan, and Lian Li. Human brain glycoform coregulation network and glycan modification alterations in Alzheimer's disease. *Science Advances*, 10(14):eadk6911, 2024.

[101] Fengchao Yu, Sarah E. Haynes, and Alexey I. Nesvizhskii. IonQuant enables accurate and sensitive label-free quantification with FDR-controlled match-between-runs. *Molecular & Cellular Proteomics*, 20:100077, 2021.

[102] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, 2008.

[103] Peter Langfelder and Steve Horvath. Eigengene networks for studying the relationships between co-expression modules. *BMC Systems Biology*, 1(1):54, 2007.

[104] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[105] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–177, 1982.

[106] Joshua Klein, Luis Carvalho, and Joseph Zaia. Expanding N-glycopeptide identifications by modeling fragmentation, elution, and glycome connectivity. *Nature Communications*, 15(1):6168, 2024.

[107] J. Alex Taylor and Richard S. Johnson. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 11(9):1067–1075, 1997.

[108] Hao Yang, Hao Chi, Wen-Feng Zeng, Wen-Jing Zhou, and Si-Min He. pNovo 3: Precise de novo peptide sequencing using a learning-to-rank framework. *Bioinformatics*, 35(14):i183–i190, 2019.

[109] Korrawe Karunratanakul, Hsin-Yao Tang, David W. Speicher, Ekapol Chuangsuwanich, and Sira Sriswasdi. Uncovering thousands of new peptides with sequence-mask-search hybrid de novo peptide sequencing framework. *Molecular & Cellular Proteomics*, 18(12):2478–2491, 2019.

[110] Kaiyuan Liu, Yuzhen Ye, Sujun Li, and Haixu Tang. Accurate de novo peptide sequencing using fully convolutional neural networks. *Nature Communications*, 14:7974, 2023.

[111] Jan Muntel, Tejas Gandhi, Lynn Verbeke, Oliver M. Bernhardt, Tobias Treiber, Roland Bruderer, and Lukas Reiter. Surpassing 10,000 identified and quantified proteins in a single run by optimizing current LC–MS instrumentation and data analysis strategy. *Molecular Omics*, 15(5):348–360, 2019.

[112] Dongxue Wang, Basak Eraslan, Thomas Wieland, Björn Hallström, Thomas Hopf, Daniel Paul Zolg, Jana Zecha, Anna Asplund, Li-Hua Li, Chen Meng, Martin Frejno, Tobias Schmidt, Karsten Schnatbaum, Mathias Wilhelm, Fredrik Ponten, Mathias Uhlén, Julien Gagneur, Hannes Hahne, and Bernhard Kuster. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Molecular Systems Biology*, 15(2):e8503, 2019.

[113] Mathias Wilhelm, Daniel P. Zolg, Michael Graber, Siegfried Gessulat, Tobias Schmidt, Karsten Schnatbaum, Celina Schwencke-Westphal, Philipp Seifert, Niklas de Andrade Krätzig, Johannes Zerweck, Tobias Knaute, Eva Bräunlein, Patroklos Samaras, Ludwig Lautenbacher, Susan Klaeger, Holger Wenschuh, Roland Rad, Bernard Delanghe, Andreas Huhmer, Steven A. Carr, Karl R. Clauser, Angela M. Krackhardt, Ulf Reimer, and Bernhard Kuster. Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nature Communications*, 12(1):3346, 2021.

[114] Eric W. Deutsch, Nuno Bandeira, Vagisha Sharma, Yasset Perez-Riverol, Jeremy J. Carver, Deepti J. Kundu, David García-Seisdedos, Andrew F. Jarnuczak, Suresh Hewapathirana, Benjamin S. Pullman, Julie Wertz, Zhi Sun, Shin Kawano, Shujiro Okuda, Yu Watanabe, Henning Hermjakob, Brendan MacLean, Michael J. MacCoss, Yunping Zhu, Yasushi Ishihama, and Juan A. Vizcaíno. The ProteomeXchange consortium in 2020: Enabling big data approaches in proteomics. *Nucleic Acids Research*, 48(D1):D1145–D1152, 2020.

[115] Mingxun Wang, Jian Wang, Jeremy Carver, Benjamin S. Pullman, Seong Won Cha, and Nuno Bandeira. Assembling the community-scale discoverable human proteome. *Cell Systems*, 7(4):412–421.e5, 2018.

[116] Lukas Käll, John D. Storey, Michael J. MacCoss, and William Stafford Noble. Posterior error probabilities and false discovery rates: Two sides of the same coin. *Journal of Proteome Research*, 7(1):40–44, 2008.

[117] Viktor Granholm and Lukas Käll. Quality assessments of peptide–spectrum matches in shotgun proteomics. *Proteomics*, 11(6):1086–1093, 2011.

[118] Viktor Granholm, José Fernández Navarro, William Stafford Noble, and Lukas Käll. Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics. *Journal of Proteomics*, 80:123–131, 2013.

[119] Uri Keich and William Stafford Noble. Controlling the FDR in imperfect matches to an incomplete database. *Journal of the American Statistical Association*, 113(523):973–982, 2018.

[120] Uri Keich and William Stafford Noble. On the importance of well-calibrated scores for identifying shotgun proteomics spectra. *Journal of Proteome Research*, 14(2):1147–1160, 2015.

[121] William E. Fondrie and William S. Noble. mokapot: Fast and flexible semisupervised learning for peptide detection. *Journal of Proteome Research*, 20(4):1966–1971, 2021.

[122] Ngoc Hieu Tran, Rui Qiao, Lei Xin, Xin Chen, Chuyi Liu, Xianglilan Zhang, Baozhen Shan, Ali Ghodsi, and Ming Li. Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nature Methods*, 16(1):63–66, 2019.

[123] E. Keith Keenan, Derek K. Zachman, and Matthew D. Hirschey. Discovering the landscape of protein modifications. *Molecular Cell*, 81(9):1868–1878, 2021.

[124] Michael R. Shortreed, Craig D. Wenger, Brian L. Frey, Gloria M. Sheynkman, Mark Scalf, Mark P. Keller, Alan D. Attie, and Lloyd M. Smith. Global identification of protein post-translational modifications in a single-pass database search. *Journal of Proteome Research*, 14(11):4714–4720, 2015.

[125] Dekel Tsur, Stephen Tanner, Ebrahim Zandi, Vineet Bafna, and Pavel A. Pevzner. Identification of post-translational modifications via blind search of mass spectra. In *2005 IEEE Computational Systems Bioinformatics Conference*, pages 157–166, 2005.

[126] Stephen Tanner, Hongjun Shu, Ari Frank, Ling-Chi Wang, Ebrahim Zandi, Marc Mumby, Pavel A. Pevzner, and Vineet Bafna. InsPecT: Identification of post-translationally modified peptides from tandem mass spectra. *Analytical Chemistry*, 77(14):4626–4639, 2005.

[127] Seungjin Na, Nuno Bandeira, and Eunok Paek. Fast multi-blind modification search through tandem mass spectrometry. *Molecular & Cellular Proteomics*, 11(4):M111.010199, 2012.

[128] Wout Bittremieux, Pieter Meysman, William Stafford Noble, and Kris Laukens. Fast open modification spectral library searching through approximate nearest neighbor indexing. *Journal of Proteome Research*, 17(10):3463–3474, 2018.

[129] Weizhong Li and Adam Godzik. CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

[130] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

[131] Daniel G. C. Treen, Mingxun Wang, Shipei Xing, Katherine B. Louie, Tao Huan, Pieter C. Dorrestein, Trent R. Northen, and Benjamin P. Bowen. SIMILE enables alignment of tandem mass spectra with statistical significance. *Nature Communications*, 13(1):2510, 2022.

[132] Dmitry M. Avtonomov, Andy Kong, and Alexey I. Nesvizhskii. DeltaMass: Automated detection and visualization of mass shifts in proteomic open-search results. *Journal of Proteome Research*, 18(2):715–720, 2019.

[133] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, volume 2024, pages 35549–35562, 2024.

[134] Daniel Paul Zolg, Mathias Wilhelm, Tobias Schmidt, Guillaume Médard, Johannes Zerweck, Tobias Knaute, Holger Wenschuh, Ulf Reimer, Karsten Schnatbaum, and Bernhard Kuster. ProteomeTools: Systematic characterization of 21 post-translational protein modifications by liquid chromatography tandem mass spectrometry (LC-MS/MS) using synthetic peptides. *Molecular & Cellular Proteomics*, 17(9):1850–1863, 2018.

[135] Robert J. Chalkley and Karl R. Clauser. Modification site localization scoring: Strategies and performance. *Molecular & Cellular Proteomics*, 11(5):3–14, 2012.

[136] Thomas Taus, Thomas Köcher, Peter Pichler, Carmen Paschke, Andreas Schmidt, Christoph Henrich, and Karl Mechtler. Universal and confident phosphorylation site localization using phosphoRS. *Journal of Proteome Research*, 10(12):5354–5362, 2011.

[137] Hongying Zhong, Sandra L. Marcus, and Liang Li. Two-dimensional mass spectra generated from the analysis of ^{15}N-labeled and unlabeled peptides for efficient protein identification and de novo peptide sequencing. *Journal of Proteome Research*, 3(6):1155–1163, 2004.

[138] Alexey I. Nesvizhskii. Proteogenomics: Concepts, applications and computational strategies. *Nature Methods*, 11(11):1114–1125, 2014.

[139] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped BLAST and PSI-BLAST:

a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[140] Robbin Bouwmeester, Ralf Gabriels, Niels Hulstaert, Lennart Martens, and Sven Degroeve. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nature Methods*, 18(11):1363–1369, 2021.

[141] Steven A. Carr, Susan E. Abbatiello, Bradley L. Ackermann, Christoph Borchers, Bruno Domon, Eric W. Deutsch, Russell P. Grant, Andrew N. Hoofnagle, Ruth Hüttenhain, John M. Koomen, Daniel C. Liebler, Tao Liu, Brendan MacLean, D. R. Mani, Elizabeth Mansfield, Hendrik Neubert, Amanda G. Paulovich, Lukas Reiter, Olga Vitek, Ruedi Aebersold, Leigh Anderson, Robert Bethem, Josip Blonder, Emily Boja, Julianne Botelho, Michael Boyne, Ralph A. Bradshaw, Alma L. Burlingame, Daniel Chan, Hasmik Keshishian, Eric Kuhn, Christopher Kinsinger, Jerry S. H. Lee, Sang-Won Lee, Robert Moritz, Juan Oses-Prieto, Nader Rifai, James Ritchie, Henry Rodriguez, Pothur R. Srinivas, R. Reid Townsend, Jennifer Van Eyk, Gordon Whiteley, Arun Wiita, and Susan Weintraub. Targeted peptide measurements in biology and medicine: Best practices for mass spectrometry-based assay development using a fit-for-purpose approach. *Molecular & Cellular Proteomics*, 13(3):907–917, 2014.

[142] Florian Meier, Scarlet Beck, Niklas Grassl, Markus Lubeck, Melvin A. Park, Oliver Raether, and Matthias Mann. Parallel accumulation–serial fragmentation (PASEF): Multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device. *Journal of Proteome Research*, 14(12):5378–5387, 2015.

[143] Shivani Tiwary, Roie Levy, Petra Gutenbrunner, Favio Salinas Soto, Krishnan K. Palaniappan, Laura Deming, Marc Berndl, Arthur Brant, Peter Cimermancic, and Jürgen Cox. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature Methods*, 16(6):519–525, 2019.

[144] Ching Tarn and Wen-Feng Zeng. pDeep3: Toward more accurate spectrum prediction with fast few-shot learning. *Analytical Chemistry*, 93(14):5815–5822, 2021.

[145] Zeping Mao. RNovA test datasets. Zenodo, 2026.

[146] Qianqiu Zhang. RNovA end-to-end workflow. GitHub repository, 2026. Accessed 2026-08-09.

[147] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013.

[148] Abhijit Bendale and Terrance E. Boult. Towards open world recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1893–1902. IEEE, 2015.

[149] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6(50):1453–1484, 2005.

[150] Chun-Nam John Yu and Thorsten Joachims. Learning structural SVMs with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1169–1176. ACM, 2009.

[151] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017.

[152] Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

[153] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[154] Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

[155] George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

[156] Miles Cranmer, Alvaro Sanchez-Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. Discovering symbolic models from deep learning with inductive biases. In *Advances in Neural Information Processing Systems*, volume 33, pages 17429–17442. Curran Associates, Inc., 2020.

[157] Ross D. King, Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N. Soldatova, Andrew Sparkes, Kenneth E. Whelan, and Amanda Clare. The automation of science. *Science*, 324(5923):85–89, 2009.

[158] David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

Glossary

Glycan composition The counts and identities of monosaccharide units associated with a glycan, without necessarily specifying linkage positions or stereochemistry.

Glycopeptide A peptide carrying one or more covalently attached glycans.

Fixed label space An output space whose classes or tokens are specified before training and cannot directly express an entity outside that predefined set.

Independent validation Evidence not reducible to the score that generated a candidate, such as a synthetic reference peptide, isotope experiment, orthogonal sequence measurement, or targeted follow-up assay.

Mass path A sequence of mass states through a spectrum graph that represents the fragment-level evidence for a peptide.

Mass tag An unresolved mass interval that may represent multiple residues, missing fragmentation, or an unknown modification.

Molecular candidate A peptide, modification, glycan composition, or other molecular explanation proposed from measurement evidence but not yet equivalent to a validated identification.

Open modification discovery Identification of residue mass shifts without restricting inference to a fixed list of modification types.

Open-world inference Inference in which observations may arise from entities absent from the training labels or reference database, requiring both candidate generation and explicit treatment of unknowns.

Structure-aware representation A representation that preserves chemically meaningful relations such as fragment connectivity, mass differences, or precursor constraints.

Staged structural inference A modeling strategy that recovers intermediate spectral structure before producing final biological interpretation.