WHAT COUNTS AS A FALSE DISCOVERY IN *De Novo* PEPTIDE SEQUENCING?

Anonymous authors
Paper under double-blind review

ABSTRACT

In *de novo* peptide sequencing, matching observed fragments does not necessarily establish sequence identity, complicating the interpretation of false-discovery estimates. We investigate this problem using isotope measurements, designed peptide libraries and public sequencing and error-estimation workflows. Isotope traces and fragment matches leave some sequence differences unresolved, whereas synthesis constraints distinguish alternatives under explicit assumptions about library composition. The distinction between fragment support and sequence agreement persists after confidence-based selection. Among spectra with identities supported by both synthesis design and database search, a Casanovo selection at a nominal 5% error rate yields 3.31% disagreement under a fragment-coverage criterion but 11.15% under a modified-sequence criterion. Calibrating confidence against these two criteria also produces different sequence-disagreement rates when the fitted models are applied to other biological studies. Even with the error criterion fixed, observed disagreement depends on the population used for calibration and evaluation, and on whether spectrum matches or distinct peptides are counted. Together, these results show why error estimates must be tied to the sequence identity being claimed, the evidence supporting its assessment and the predictions ultimately reported.

1 INTRODUCTION

A *de novo* peptide sequencer predicts a molecular sequence from a tandem mass spectrum. PEAKS used spectrum-based sequence optimization (Ma et al., 2003); DeepNovo learned sequence patterns from data (Tran et al., 2017), and recent systems use autoregressive or non-autoregressive architectures (Yilmaz et al., 2024; Eloff et al., 2025; Zhang et al., 2025). Yet the measured fragments can support more than one sequence. If a prediction differs from a database-search annotation but explains the observed peaks, what evidence would establish that it is wrong? The answer depends on whether the discovery concerns precursor composition, an ordered peptide sequence, or a particular modification site.

These identity claims require different evidence. Moving a phosphate between adjacent residues preserves the observed fragments when no measured cleavage distinguishes those sites; localization methods therefore use site-determining ions (Beausoleil et al., 2006). An isotope trace can support precursor composition while leaving sequence order ambiguous. Protein and translated-genome references add information about origin, with coverage determined by their completeness and translation conventions (Zhu et al., 2018). Synthetic peptide libraries constrain the candidate identities through their design (Marx et al., 2013; Zolg et al., 2018). Recent models for modified peptides and open modification discovery make these distinctions especially relevant (Chen et al., 2025; Lauridsen et al., 2026; Mao et al., 2026).

Error estimation adds a statistical question: how many accepted predictions fail the identity criterion? Target-decoy competition and learned rescoring are established approaches to estimating errors and selecting database-search identifications (Elias & Gygi, 2007; Käll et al., 2007). Confidence assessment also depends on the reported unit: assigning spectra and selecting unique peptides require different aggregation procedures (Granholm et al., 2013). Recent entrapment and decoy studies examine whether the complete selection procedure supports its reported error rate (Wen et al., 2025; Reznikov et al., 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

For *de novo* sequencing, spectrum decoys, calibrated confidence and learned counterpart comparisons provide distinct routes to error estimation (Tran et al., 2024; Mabona et al., 2025; Liang et al., 2026). Sequence ambiguity and alternative evaluation conventions are already documented (Frank et al., 2007; Devabhaktuni & Elias, 2016; van Puyenbroeck et al., 2025; ProteoBench contributors, 2026). Their interaction matters: an accepted set with few fragment-coverage disagreements can contain substantially more sequence disagreements. An estimator trained or evaluated on the first event answers a different question from one targeting the second.

This distinction also connects peptide identification to confidence calibration in machine learning. Calibration assesses whether predicted probabilities agree with observed correctness (Guo et al., 2017), while distribution shift can change that agreement after deployment (Ovadia et al., 2019). In peptide sequencing, the correctness event itself depends on the identity being evaluated. Separating a change in that event from a change in the evaluated population is therefore essential to interpreting transfer results.

We examine how measurement resolution, the definition of error and confidence-based selection interact in *de novo* sequencing. Isotope measurements and synthesis designs establish which sequence differences the available evidence can distinguish. With three sequencers on common inputs, we first hold accepted predictions fixed to isolate the effect of the error definition. We then hold the evaluation criterion fixed to examine calibration transfer, population restrictions, control construction and discovery units.

2 What is the error being counted?

Let $y_i^\star$ denote the peptide that generated the designated precursor for spectrum x_i, and $\hat{y}_i$ the reported sequence. An equivalence relation $\sim$ specifies the identity of interest, including whether leucine and isoleucine are distinguished and whether modification positions matter. For accepted set $A(t)$, the realized false discovery proportion and its expectation are (Benjamini & Hochberg, 1995)

$$V_\sim(t) = \sum_{i \in A(t)} \mathbf{1}\{\hat{y}_i \not\sim y_i^\star\}, \qquad \mathrm{FDP}_\sim(t) = \frac{V_\sim(t)}{|A(t)| \vee 1}, \qquad \mathrm{FDR}_\sim(t) = \mathbb{E}[\mathrm{FDP}_\sim(t)]. \tag{1}$$

Here $a \vee 1 = \max(a, 1)$. The discovery unit determines the assignments being counted: PSMs and separately selected peptide representatives have different denominators. Peptide presence asks a further question. A prediction can belong to the sample's molecular inventory $\mathcal{M}$ while being assigned to the wrong precursor. For mixed spectra, the identity claim must specify the designated precursor or any co-isolated component.

The generating peptide is generally unobserved. We compare predictions with a reference annotation ℓ_i using a binary agreement predicate $g(x_i, \hat{y}_i, \ell_i)$ and report

$$\mathrm{rFDP}_g(t) = \frac{\sum_{i \in A(t)} [1 - g(x_i, \hat{y}_i, \ell_i)]}{|A(t)| \vee 1}. \tag{2}$$

Sequence identity is the special case $g = \mathbf{1}\{\hat{y}_i \sim \ell_i\}$. Interpreting reference disagreement as physical error depends on the reference's accuracy and resolution.

Three comparisons recur. *Sequence/site* compares ordered modified sequences with I/L collapsed; *formula* compares complete elemental compositions. *Ion90* is a spectrum-conditioned predicate requiring recovery of at least 90% of the observed reference-annotated fragment peaks. Its implemented convention also passes spectra with no matched reference peaks. Published confidence and counterpart methods use *sequence-mass* matching under their residue and cumulative-mass tolerances. Table 6 defines each comparison.

Synthesis designs provide candidate identities. Let C_i contain the designed peptidoforms compatible with the measured precursor under a specified chemistry model. Conditional on $y_i^\star \in C_i$ and a nonempty accepted set, incompatible predictions B give a lower assignment-error bound $L = |B|/|A|$; adding compatible but unresolved assignments M and unparsed predictions Q gives $U = (|B| + |M| + |Q|)/|A|$. Appendix C defines the sets and chemical equivalences.

3 DATA AND EXPERIMENTAL COMPARISONS

Our evidence combines search-annotated biological spectra, isotope measurements and designed peptide libraries (Table 1). Human and yeast provide calibration tests from different studies than the *E. coli* source. The synthetic populations use progressively stronger identity constraints and share spectra.

Table 1: **Datasets and their roles.** Counts precede method-specific eligibility. Design anchors and supported anchors are nested subsets of the full synthetic population. Training-sequence overlap is reported in Appendix A.5.

Population	Spectra	Evidence and role
E. coli 30mM / 60mM	4,458 / 7,046	PXD008782; search references and isotope traces; source calibration and within-study transfer.
Human	36,049	PXD004424, 26 acquisitions; calibration transfer with substantial training-proxy sequence overlap.
Yeast	40,000	PXD003868, 27 acquisitions; external test with lower observed training-proxy overlap.
Synthetic full	131,036	PXD000138, 83 libraries; sets of precursor-compatible designed identities.
Design anchors	11,150	Unique design-compatible identity; 1,540 I/L-collapsed backbones.
Supported anchors	2,190	Robust design identity, agreeing fragment-search annotation and isotope-zero precursor; 440 backbones.
Additional data	pY: 45,180 PSMs; counterpart QC: 37,366 GraphNovo targets.	

Identity evidence. The labeled *E. coli* data connect candidate-specific isotope traces, observed fragments and protein/six-frame membership. A trace rule developed on 30mM is applied unchanged to 60mM. For PXD000138, the synthesis design specifies precursor-compatible candidate sets; the supported anchors additionally require agreement with the deposited fragment-search annotation. A separate phosphotyrosine library constrains phosphate location given the assigned backbone and modification inventory. Appendices B–D describe these measurements and identity conditions.

Decoy selection on common inputs. Casanovo 5.2 Orbitrap, InstaNovo 1.2 and the MassIVE-trained π-PrimeNovo model receive the same spectra without reference labels. NovoBoard spectrum decoys retain 50% of the peaks and use seed 99. Within each model, target and decoy inference settings are identical. Targets and decoys are decoded separately and pooled by score. At nominal 1% and 5%, we select the largest complete-score-tie prefix satisfying D/T and evaluate all criteria on the accepted targets. We also evaluate $(D+1)/T$. Post hoc comparisons restrict reference chemistry and isotope offsets to the models' common domain, or draw noise peaks within each synthetic library. Model and selection settings are in Appendix A.

Calibration comparisons. We fit scalar-score calibrators to the same *E. coli* rows using either sequence/site agreement or ion90 as supervision, then apply the fitted maps unchanged to the other studies. Each source partition uses identical scores and fitting rows for the two supervision criteria. Five rotations of the backbone-grouped source split assess fitting sensitivity. Pretrained Winnow uses its published 11 features and five-beam InstaNovo predictions; it provides a separate test of a deployed confidence model. For π-NovoQC, separate PSM and peptide heads use the published GraphNovo features, with each of three species held out in turn. Source and local calibration factors are compared on the same external scores (Appendix E).

Discovery units and uncertainty. PSM and peptide analyses select records separately. For spectrum decoys, each peptide is represented by its highest-scoring PSM within each arm, with I/L collapsed; synthetic grouping also includes library. Assignment to a precursor and membership in the intended library are evaluated separately on the selected representatives. Reported confidence intervals resample reference-backbone groups with fitted models, calibration and accepted sets fixed. Unknown chemistry gives lower/upper disagreement bounds on the full accepted denominator. Appendix A specifies aggregation and uncertainty for each method.

4 RESULTS

4.1 WHAT DO THE MEASUREMENTS DISTINGUISH?

Isotope and fragment support can leave sequence order unresolved. In labeled *E. coli*, different sequences can query the same isotope traces when they have the same precursor mass and label-eligible atom counts. After accounting for residue-level chemical aliases, 459 reference-discordant predictions in 30mM satisfy these conditions. Extracting traces separately for each prediction and reference yields identical traces because their query coordinates coincide. Of these pairs, 69 pass the RT30 rule for coelution and proximity to MS2. In the 60mM acquisition, the unchanged rule retains 78 of 539 such disagreements (Figure 1).

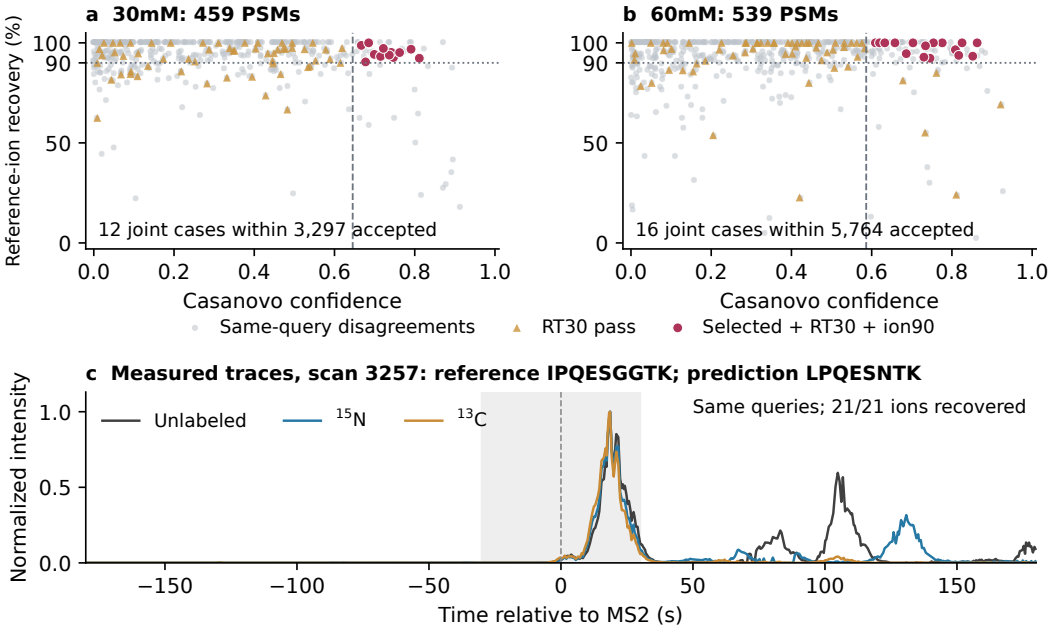

Figure 1: **Measured support can leave competing sequences unresolved.** (a,b) Nonalias reference disagreements with identical isotope queries in the two *E. coli* acquisitions. Triangles pass RT30; red points also pass ion90 and nominal-5% decoy selection. Vertical lines mark selection thresholds and horizontal lines mark 90% reference-ion recovery. (c) The highest-score joint case in 60mM. Both sequences query the same traces and recover all 21 observed reference-ion peaks. Each channel is normalized separately; shading marks ±30 seconds around MS2.

The nominal-5% decoy selections contain 12 and 16 disagreements passing both RT30 and ion90 in the two acquisitions. Their reference sequences all occur in the inspected proteome and six-frame translation, while their predictions do not. This origin evidence adds discrimination beyond the shared measurements. Translation itself has identifiable gaps: four annotated peptides, represented by six PSMs, require alternative initiation or the programmed *prfB* frameshift to be recovered (Appendix B).

Synthesis constraints distinguish alternatives with shared fragment support. Designed libraries provide another source of identity information. Before confidence-based selection, the supported-anchor predictions include formula-preserving sequence disagreements that recover every observed reference-ion peak: 100 for Casanovo, 70 for InstaNovo and 87 for PrimeNovo. These cases remain after excluding chemical aliases and have at least seven reference peaks each (Figure 2a).

In Figure 2, the prediction lies outside the intended library because the design fixes G at position three. Yet it recovers all 21 reference peaks in the 2+ scan, accounting for 37.7% of total ion intensity. A 3+ companion acquired 2.02 seconds earlier adds two reference-only matches, accounting for 0.8% of its intensity. The library design, precursor and fragment-search assignment jointly constrain identity; complete recovery of the reference peaks alone leaves this sequence difference unresolved.

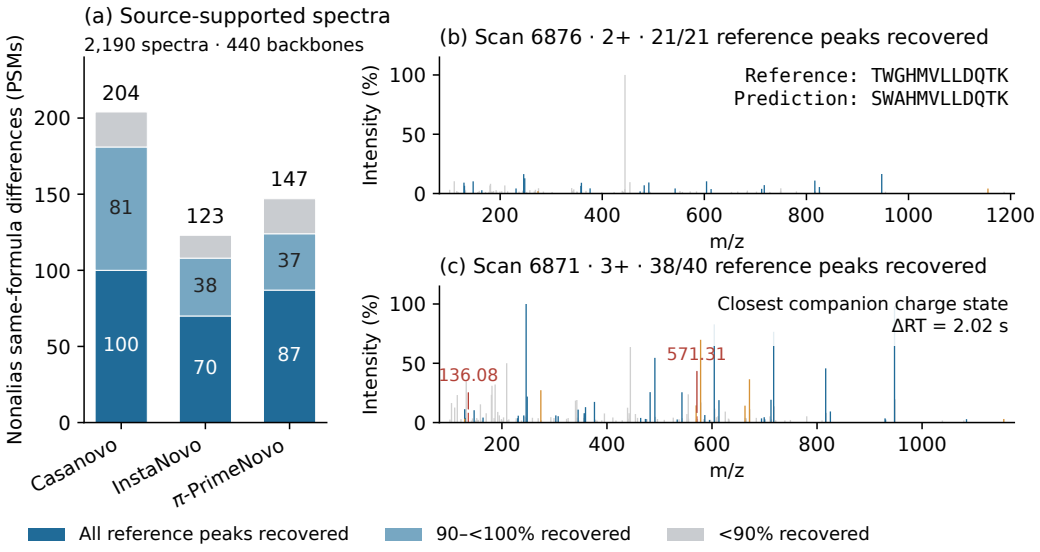

Figure 2: **A synthesis design distinguishes sequences with shared fragments.** (a) Formula-preserving, nonalias reference disagreements among supported anchors, grouped by reference-ion recovery. (b) The highest-score InstaNovo ion90 case changes `TWG` to `SWA`, preserving the formula but violating the design's fixed third-position G. All 21 reference peaks are recovered. (c) The nearest-RT companion with a different charge recovers 38 of 40. Blue peaks match both sequences, red only the reference, gold only the prediction and gray neither.

4.2 Does the error definition match the claimed identity?

Within the common reference-chemistry and isotope-zero domain, the choice of error definition places the same accepted set on opposite sides of the nominal rate. For Casanovo on supported anchors, nominal-5% selection gives 11.15% modified-sequence disagreement and 3.31% ion90 disagreement. Accounting for ordered residue-composition aliases leaves 10.74% sequence disagreement. On human spectra, the corresponding sequence and ion90 rates are 7.45% and 1.56%. Figure 3 shows all three models: within each selection, the predictions, scores and threshold are identical across comparisons.

Precursor compatibility changes the selected error rate. PrimeNovo's precursor-mass correction allows isotope offset zero. All 1,952 accepted isotope-one references in its unrestricted human 5% selection disagree with predictions. Restricting both arms to compatible reference chemistry and isotope offset zero, then reselecting the threshold, reduces sequence disagreement from 9.63% to 2.82%. This post hoc restriction uses reference annotations to define compatibility. Casanovo and InstaNovo retain a gap between sequence and ion90 disagreement within the common domain.

Modification location gives a further example. In the phosphotyrosine library, nine accepted OmniNovo predictions move phosphate from Y to S/T while preserving the assigned backbone and modification inventory. All nine pass ion90 and six recover every reference-ion peak. Their full selection has 2.15% sequence disagreement, showing that these site contradictions also occur in a population whose aggregate rate is below nominal (Appendix D).

Confidence inherits the error definition used in training. The error definition also affects calibration on new datasets. We fit two isotonic maps to the same *E. coli* scores and rows, using sequence/site agreement or ion90. For the primary source partition on human spectra, the ion-supervised maps give sequence disagreement of 10.27%, 14.66% and 11.88% for Casanovo, InstaNovo and PrimeNovo, with ion90 disagreement below 3.2%. Sequence supervision gives 3.35%, 4.08% and 2.28%, respectively (Figure 4).

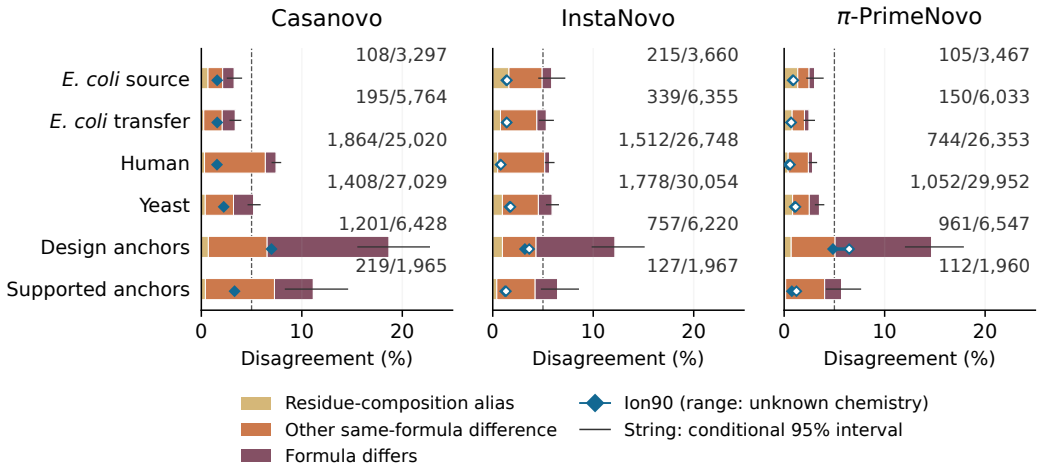

Figure 3: **Different error definitions on the same accepted predictions.** Nominal-5% selections restricted to common reference chemistry and isotope offset zero. Bars partition modified-sequence disagreements into chemical aliases, other same-formula differences and formula differences. Blue diamonds show ion90 disagreement; open endpoints include unknown ion chemistry. Black intervals resample reference backbones with the accepted set fixed. Labels give sequence disagreements/accepted targets. Design anchors and supported anchors are nested. Full results appear in Appendix F.

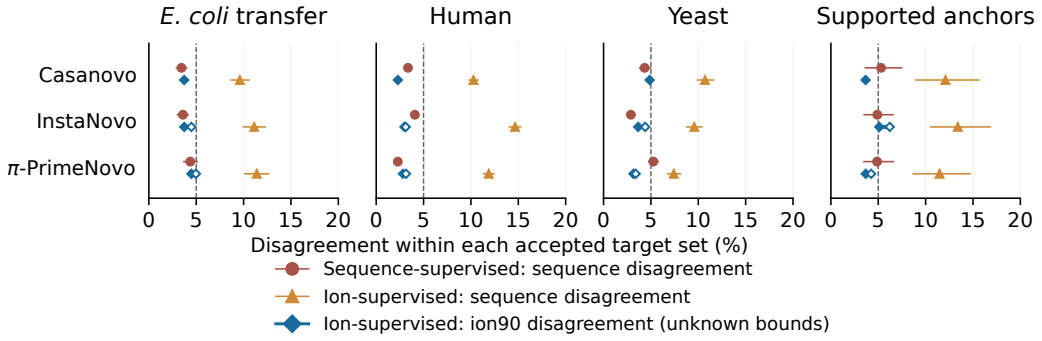

Figure 4: **Supervision changes what calibrated confidence predicts.** Source-fitted score calibrators applied unchanged at nominal 5% within the common domain. Circles evaluate sequence-supervised selections; triangles and diamonds evaluate the same ion-supervised selections against sequence and ion90 criteria. Intervals resample reference backbones with selections fixed; blue ranges include unknown ion chemistry. Both maps use identical fitting spectra and partitions. The vertical line marks 5%.

Human evaluates cross-study transfer under substantial training-proxy sequence exposure. The supervision gap also appears in the second *E. coli* acquisition and in yeast, whose sequences have much lower proxy overlap (Appendix A.5). Across five source partitions, all three models retain higher human sequence disagreement under ion supervision than under sequence supervision (Table 2). For InstaNovo, two sequence-supervised fits exceed nominal 5% (5.24% and 5.46%), while all five ion-supervised fits give 14.37–16.40% sequence disagreement. Across these fits, human ion90 disagreement under ion supervision ranges from 2.05% to 4.93%. The supervision contrast persists, while calibration within each criterion varies with the fitting partition.

Matching the error definition still leaves a transfer problem: on supported anchors, the ion-supervised InstaNovo map gives 5.11–6.22% ion90 disagreement at nominal 5%.

Table 2: **The supervision contrast persists across five source fits.** Human common-domain PSM selections at nominal 5%. Sequence columns give the minimum and maximum disagreement across five fitting partitions. The ion90 column spans their lower and upper disagreement bounds, including unknown chemistry.

	Sequence supervision	Ion90 supervision	
Model	Sequence (%)	Sequence (%)	Ion90 (%)
Casanovo	3.20–4.27	10.27–13.70	2.28–3.25
InstaNovo	4.08–5.46	14.37–16.40	2.78–4.93
π-PrimeNovo	2.01–2.82	10.29–13.96	2.05–4.42

4.3 With the error definition fixed, what changes after selection?

Winnow gives below-nominal sequence-mass disagreement on biological spectra. At nominal 5%, pretrained Winnow gives 1.08–1.72% sequence-mass disagreement across both *E. coli* acquisitions, human and yeast (Table 3). It uses five-beam InstaNovo predictions and all 11 published features. At nominal 1%, its selections are empty: every predicted error probability exceeds 1% (Appendix A.6).

Table 3: **Complete pretrained Winnow with sequence-mass evaluation.** The 5% column gives disagreements/accepted PSMs and 95% backbone intervals with selection fixed. Parentheses denote Prosit energy; 27 is the default and 40 the synthetic acquisition setting. Eligibility reflects feature requirements.

Population	Eligible	1% accepted	5% disagr./accepted	5% rate (%) [interval]
E. coli 30mM	4,404	0	25/2,323	1.08 [0.64, 1.55]
E. coli 60mM	7,001	0	47/4,143	1.13 [0.80, 1.49]
Human	35,581	0	393/22,870	1.72 [1.49, 1.96]
Yeast	39,692	0	350/22,013	1.59 [1.31, 1.91]
Design anchors (27)	9,776	0	739/5,367	13.77 [10.85, 17.27]
Design anchors (40)	9,776	0	722/5,333	13.54 [10.74, 17.01]
Supported anchors (27)	1,499	0	31/1,339	2.32 [1.36, 3.68]
Supported anchors (40)	1,499	0	27/1,335	2.02 [1.18, 3.24]

Winnow disagreement differs between design and supported anchors. Winnow's disagreement is 13.77% on design anchors and 2.32% on supported anchors at the default collision energy. With the design-anchor features and cutoff unchanged, restriction to supported anchors gives 28/1,646 disagreements (1.70%); reselecting the threshold with the same probabilities gives 37/1,722 (2.15%). The rate drops when selection is restricted to spectra with search-supported design identities. This restriction changes both reference support and spectral composition. Repeating the feature computation at acquisition energy 40 preserves the contrast (Table 3); paired feature and eligibility comparisons are given in Appendix C.4.

Unresolved identities yield conditional assignment bounds. The full synthetic population retains spectra with multiple compatible identities, including modified candidates. At nominal 5%, the design-based assignment-error lower bounds are 40.14–44.25% for the three decoy workflows and 26.65% for Winnow (Table 4). Expanding the chemistry lowers these bounds to 37.47–41.58% and 25.50%, respectively. The bounds assume that each generating peptide belongs to the specified candidate set. Unmodeled synthesis products or contaminants can contribute to apparent incompatibility; compatible alternatives account for the unresolved portion of each interval.

Peptide presence and precursor assignment answer different questions. Changing from PSMs to peptides also changes the selected population. We select peptide representatives separately, then compare assignment to the designated precursor with membership in the intended library (Figure 5). For Casanovo design anchors, 324 of 1,043 selected peptide representatives disagree with their precursor assignment, while 173 fail library membership. Both counts concern the same

Table 4: **Conditional assignment bounds on the full synthetic population.** Nominal-5% PSM selections, including modified candidate identities. Bounds assume that the generating peptide belongs to the intended library's precursor-compatible candidate set. The expanded model permits one N/Q deamidation and compares ordered residue compositions. Each row has its own accepted set and method eligibility; interval width reflects unresolved assignments.

Workflow	Accepted	Design bounds (%)	Expanded bounds (%)
Casanovo + decoys	75,524	43.42–92.96	41.41–93.27
InstaNovo + decoys	80,705	40.14–92.94	37.47–93.26
π-PrimeNovo + decoys	82,035	44.25–93.19	41.58–93.50
InstaNovo + Winnow	44,926	26.65–90.20	25.50–90.69

representatives. These 151 additional assignment disagreements concern peptides that belong to the intended library but are assigned to a different precursor. The direction of the PSM-to-peptide rate change varies across populations (Appendix F).

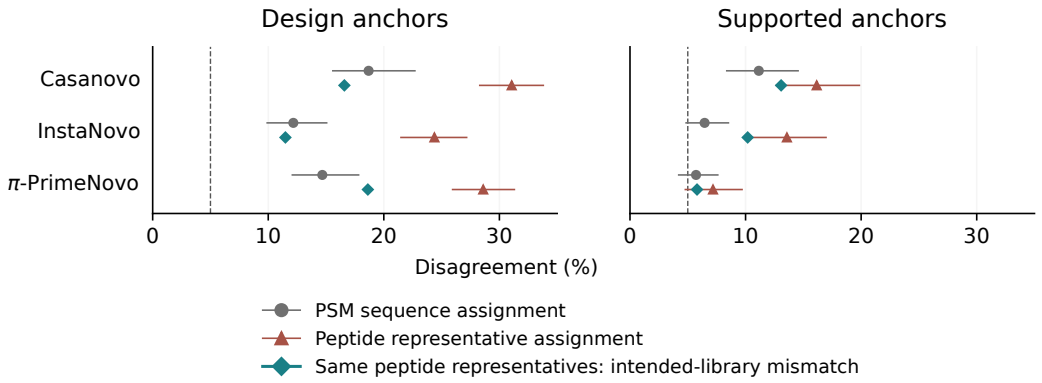

Figure 5: **Discovery unit and identity claim have separate effects.** Nominal-5% common-domain selections. Gray circles evaluate PSM assignments. Red triangles evaluate separately selected peptide representatives, and teal diamonds test membership in the intended library for those same representatives. Assignment intervals resample reference backbones; membership ranges include unknown chemistry. Both nested synthetic populations and all three models are shown.

Calibration source affects a fixed sequence criterion. On held-out *A. thaliana*, π-NovoQC source calibration selects 6,917 PSMs with 7.81% disagreement; independent local calibration selects 4,684 with 3.84% on the same external scores. The source-domain test already gives 6.23%, so the external excess reflects an imperfect source fit as well as possible domain change. Local calibration gives 4.92% disagreement on *C. elegans*; source calibration is undefined for *C. elegans*, and both factors are undefined for *E. coli*: no calibration point meets the published 10^{-4} FDP tolerance. Separate peptide heads give unadjusted external rates of 6.16%, 3.08% and 6.05% for *A. thaliana*, *C. elegans* and *E. coli* (Appendix E).

Changing the noise pool changes which predictions are accepted. Drawing added decoy peaks within each synthesis library yields smaller accepted sets and lower sequence disagreement for all three models on common-domain design anchors (Table 5). Target predictions and scores are fixed, so the change arises through each construction's selected threshold. The within-library rates remain above nominal 5%, at 7.19–12.64%.

Selection also changes the composition of the controls. In the *E. coli* decoy experiment, matches to the source reference rise from about 8% of all decoys to about 65% of the nominal-5% tail. Among reference disagreements, the relative fractions preserving the isotope query reverse between targets and decoys after selection. The selected controls consequently represent a different mixture from the full decoy population, even in settings with below-nominal target disagreement (Appendix B.5).

Table 5: **Control construction changes selection at a fixed sequence criterion.** Common-domain design-anchor PSMs at nominal 5%. Only the added-peak pool changes; target scores, retained-peak fraction and other model settings stay fixed. Each construction reselects its threshold. Counts and rates refer to modified-sequence disagreements.

	Cohort-wide pool		Within-library pool	
Model	Disagr./accepted	Rate (%)	Disagr./accepted	Rate (%)
Casanovo	1,201/6,428	18.68	728/5,759	12.64
InstaNovo	757/6,220	12.17	363/5,049	7.19
π-PrimeNovo	961/6,547	14.68	459/5,654	8.12

5 Discussion and conclusions

An error estimate is interpretable relative to the identity claim it evaluates. In the isotope and synthetic-library experiments, different sequences share fragment or isotope support, while origin and synthesis constraints distinguish alternatives. Evaluating the same accepted predictions against these criteria exposes the gap between explaining measured fragments and assigning a molecular sequence. The calibration experiments show its consequence for learned confidence: changing the supervision criterion changes sequence disagreement on external studies even when fitting spectra and scores are held fixed.

Matching the correctness event also requires attention to the selected population. The Winnow comparison links lower disagreement to search-supported spectra; the PrimeNovo comparison identifies sensitivity to precursor compatibility. Changing the decoy noise pool alters accepted counts and disagreement without changing target predictions. Together, these results place error assessment at the level of the complete selection procedure, including eligibility, calibration and the reporting unit.

The appropriate endpoint follows the scientific question. Modified-sequence assignment requires residue order and modification location, whereas peptide inventory asks whether a sequence is present in the sample. Their different counts on the same peptide representatives show why they need separate evaluation. In controlled libraries, candidate-set bounds retain unresolved assignments and quantify incompatibilities under an explicit chemistry model. Applying such bounds to open discovery requires a candidate space that covers the generating identities; here, that coverage is supplied by the synthesis assumptions.

For scientific inverse problems, the value of additional evidence depends on whether it distinguishes explanations already compatible with the observations. Complementary fragmentation, including mirror-protease measurements (Cao et al., 2026), offers one route in peptide sequencing. Benchmarks can connect such measurements to the correctness event used for supervision and to the population on which discoveries are reported.

6 Limitations

Biological references are database-search annotations, and synthetic references depend on library and chemistry coverage. Supported anchors further condition evaluation on search agreement. Training-proxy lists establish sequence overlap but leave exact checkpoint membership unknown. The evaluated chemistry and acquisitions do not cover non-tryptic or HLA discovery workflows. Confidence intervals condition on fitted models and selected sets, while repeated-study FDR also depends on fitting and selection variability. One counterpart configuration yields four peptide representatives at nominal 1%, providing little information about its error rate at that threshold.

Reproducibility statement

The appendices specify public datasets, model configurations, selection rules and fitting partitions. Analysis code, figure data and prediction-level results are organized in a companion package. Code and checkpoints will be released upon publication.

Ethics statement

This study reanalyzes previously collected, publicly available proteomics data and involves no new human-subject recruitment or intervention. Biological or clinical applications of revised peptide assignments require independent validation.

AI use statement

Generative AI tools assisted manuscript editing, experiment implementation, analysis and computational workflow management. The authors verified the reported results and retain responsibility for the final content.

REFERENCES

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. doi: 10.1038/nbt1240.

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. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Zixuan Cao, Xueli Peng, Di Zhang, Piyu Zhou, Li Kang, Hao Chi, et al. DiNovo enables high-coverage and high-confidence *de novo* peptide sequencing via mirror proteases and deep learning. *Nature Communications*, 17:2203, 2026. doi: 10.1038/s41467-026-70224-6.

Yuhan Chen, Shang Qu, Zhiqiang Gao, Yuejin Yang, Xiang Zhang, Sheng Xu, et al. Accurate *de novo* sequencing of the modified proteome with OmniNovo. *arXiv preprint arXiv:2512.12272*, 2025. URL https://arxiv.org/abs/2512.12272.

Hao Chi, Chao Liu, Hao Yang, Wen-Feng Zeng, Long Wu, Wen-Jing Zhou, et al. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nature Biotechnology*, 36:1059–1061, 2018. doi: 10.1038/nbt.4236. URL https://www.nature.com/articles/nbt.4236.

Arun Devabhaktuni and Joshua E. Elias. Application of *de Novo* Sequencing to Large-Scale Complex Proteomics Data Sets. *Journal of Proteome Research*, 15(3):732–742, 2016. doi: 10.1021/acs.jproteome.5b00861.

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. doi: 10.1038/nmeth1019.

Kevin Eloff, Konstantinos Kalogeropoulos, Amandla Mabona, Oliver Morell, Rachel Catzel, Esperanza Rivera-de Torre, et al. InstaNovo enables diffusion-powered *de novo* peptide sequencing in large-scale proteomics experiments. *Nature Machine Intelligence*, 7(4):565–579, 2025. doi: 10.1038/s42256-025-01019-5.

Ari M. Frank, Mikhail M. Savitski, Michael L. Nielsen, Roman A. Zubarev, and Pavel A. Pevzner. *De Novo* Peptide Sequencing and Identification with Precision Mass Spectrometry. *Journal of Proteome Research*, 6(1):114–123, 2007. doi: 10.1021/pr060271u.

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. doi: 10.1016/j.jprot.2012.12.007.

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*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

InstaDeep. Winnow general probability calibrator: model card. Hugging Face, 2026. URL https://huggingface.co/InstaDeepAI/winnow-general-model. Accessed 16 September 2026.

Lukas Käll, Jesse D. Canterbury, Jason Weston, William Stafford Noble, and Michael J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4(11):923–925, 2007. doi: 10.1038/nmeth1113.

Jesper Lauridsen, Vahap Canbay, Rachel Catzel, Pathmanaban Ramasamy, Amandla Mabona, Kevin Eloff, et al. InstaNovo-P: a *de novo* peptide sequencing model for phosphoproteomics. *Nature Communications*, 17(1):9277, 2026. doi: 10.1038/s41467-026-75138-x. URL https://doi.org/10.1038/s41467-026-75138-x.

Zhendong Liang, Chengxin Dai, Tianze Ling, Tingpeng Yang, Yun Yang, Yeye Leng, et al. False discovery rate control for trustworthy AI-based *de novo* peptide sequencing. *bioRxiv*, 2026. doi: 10.64898/2026.06.29.735174.

Bin Ma, Kaizhong Zhang, Christopher 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. doi: 10.1002/rcm.1196.

Amandla Mabona, Jemma Daniel, Henrik Servais Janssen Knudsen, Rachel Catzel, Kevin Michael Eloff, Erwin M. Schoof, et al. *De novo* peptide sequencing rescoring and FDR estimation with Winnow. *arXiv preprint arXiv:2509.24952*, 2025. URL https://arxiv.org/abs/2509.24952.

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. doi: 10.1038/s41587-026-03116-1. URL https://www.nature.com/articles/s41587-026-03116-1.

Harald Marx, Simone Lemeer, Jan Erik Schliep, Lucrece Matheron, Shabaz Mohammed, Jürgen Cox, Matthias Mann, Albert J. R. Heck, and Bernhard Kuster. A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nature Biotechnology*, 31(6):557–564, 2013. doi: 10.1038/nbt.2585. URL https://doi.org/10.1038/nbt.2585.

NCBI RefSeq. *Escherichia coli* K-12 MG1655 reference assembly GCF_000005845.2: genome, proteins and CDS annotations, 2026. URL https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_ASM584v2/. Files retrieved 15 September 2026; chromosome NC_000913.3.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

ProteoBench contributors. *De Novo*—DDA—HCD: evaluation modes and ambiguity conventions, 2026. URL https://proteobench.readthedocs.io/en/v0.18.0/available-modules/active-modules/11-denovo-dda-hcd/. Version 0.18.0 documentation; accessed 15 September 2026.

Grigory Reznikov, Fabrice Kusters, Majid Mohammadi, Henk W. P. van den Toorn, and Pavel Sinitcyn. Protein Language Model Decoys for Target Decoy Competition in Proteomics: Quality Assessment and Benchmarks. *Journal of Proteome Research*, 25(9):4772–4781, 2026. doi: 10.1021/acs.jproteome.6c00302.

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*, 114(31):8247–8252, 2017. doi: 10.1073/pnas.1705691114.

Ngoc Hieu Tran, Rui Qiao, Zeping Mao, Shengying Pan, Qing Zhang, Wenting Li, et al. NovoBoard: A Comprehensive Framework for Evaluating the False Discovery Rate and Accuracy of *De Novo* Peptide Sequencing. *Molecular & Cellular Proteomics*, 23(11):100849, 2024. doi: 10.1016/j.mcpro.2024.100849.

Sam van Puyenbroeck, Denis Beslic, Tomi Suomi, Tanja Holstein, Thilo Muth, Laura L. Elo, et al. Limitations of *de novo* sequencing in resolving sequence ambiguity. *bioRxiv*, 2025. doi: 10.1101/2025.08.19.671052.

Bo Wen, Jack Freestone, Michael Riffle, Michael J. MacCoss, William S. Noble, and Uri Keich. Assessment of false discovery rate control in tandem mass spectrometry analysis using entrapment. *Nature Methods*, 22(7):1454–1463, 2025. doi: 10.1038/s41592-025-02719-x.

Melih Yilmaz, William E. Fondrie, Wout Bittremieux, Carlo F. Melendez, Rowan Nelson, Varun Ananth, Sewoong Oh, and William Stafford Noble. Sequence-to-sequence translation from mass spectra to peptides with a transformer model. *Nature Communications*, 15(1), 2024. doi: 10.1038/s41467-024-49731-x.

Xiang Zhang, Tianze Ling, Zhi Jin, Sheng Xu, Zhiqiang Gao, Boyan Sun, Zijie Qiu, Jiaqi Wei, Nanqing Dong, Guangshuai Wang, Guibin Wang, Leyuan Li, Muhammad Abdul-Mageed, Laks V. S. Lakshmanan, Fuchu He, Wanli Ouyang, Cheng Chang, and Siqi Sun. π-PrimeNovo: an accurate and efficient non-autoregressive deep learning model for *de novo* peptide sequencing. *Nature Communications*, 16:267, 2025. doi: 10.1038/s41467-024-55021-3. URL https://doi.org/10.1038/s41467-024-55021-3.

Yafeng Zhu, Lukas M. Orre, Henrik J. Johansson, Mikael Huss, Jorrit Boekel, Mattias Vesterlund, et al. Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. *Nature Communications*, 9(1), 2018. doi: 10.1038/s41467-018-03311-y.

Daniel Paul Zolg, Mathias Wilhelm, Tobias Schmidt, Guillaume Médard, Jeannette 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. doi: 10.1074/mcp.TIR118.000783.

A Evaluation criteria and common experimental settings

Table 6 defines the comparisons used in the figures and operating tables. Each comparison is evaluated on the same accepted records within a model, population and selection rule.

Table 6: Definitions of the reported evaluation criteria. References are search annotations or identities derived from a specified synthesis design.

Criterion	Comparison
Sequence/site (String)	Ordered modified sequence, with I/L collapsed and modification positions retained.
Residue	Ordered per-residue elemental compositions and terminal composition; this groups chemical aliases such as deamidated N/D.
Formula	Full elemental composition, including modifications and terminal water.
Ion90 / ion100	Recovery of at least 90% / all observed reference-annotated fragment peaks.
Sequence-mass	A method's published cumulative- and individual-residue mass matching against its reference.
Isotope query	Trace coordinates determined by precursor mass, charge and label-eligible atom counts.
Library membership	Predicted peptide membership in the intended mixture, separately from assignment to the designated precursor.
Design-compatible set	Peptidoforms allowed by the synthesis library and measured precursor; used for conditional identity bounds.

A.1 Biological inputs

The *E. coli* 30mM and 60mM fractions come from PXD008782. Their reference and precursor filters are described in Appendix B. The human PXD004424 and yeast PXD003868 spectra come from the corrected MSV000090982 release. Human and yeast eligibility requires standard residues, carbamidomethylated cysteine and optionally oxidized methionine, length 7–30, charge 2–4, and precursor agreement within 20 ppm with isotope offset zero or one. Human contributes 36,049 of 44,604 source records across 26 acquisitions. Yeast contributes 40,000 of 477,649 eligible records across 27 acquisitions. We use a fixed, score-independent subsample of at most 40,000 spectra per cohort.

The *E. coli* inputs use their original reference filters, including two 30mM references with unmodified cysteine. A strict-chemistry sensitivity excludes these exceptions. The isotope-zero restriction, within-library peak pool, supported-anchor population and energy-40 Winnow comparison are post hoc diagnostics.

A.2 Sequencers and spectrum decoys

Table 7 gives the three-model configuration. Reference sequences are absent from model inputs. Each model uses the same settings for targets and decoys, and its own reported score determines the ranking.

Table 7: Sequencer configurations used for both target and decoy spectra.

Model	Decoding settings
Casanovo 5.2	Orbitrap model; beam 5; top prediction; 50 ppm precursor tolerance; isotope offsets 0/1; batch 128; length 6–100; maximum charge 4
InstaNovo 1.2	Beam 5; 50 ppm precursor tolerance; isotope offsets 0/1; length at most 40; batch 128; fp16; seed 42; no diffusion or knapsack
π-PrimeNovo	MassIVE model; released precursor correction at ± 0.1 Da; length at most 40; batch 64; seed 42

NovoBoard's random generator retains $\lfloor 0.5n \rfloor$ peaks from a spectrum containing n peaks and samples the remaining complete m/z–intensity pairs from the cohort's peak pool without replacement. Python and NumPy seed 99 restart for each spectrum, as in the published implementation. Decoys inherit

the source precursor m/z, charge and retention time. Noise sampling uses the full pool without a precursor-dependent m/z cutoff; sampled peaks can coincide with retained peaks. Each cohort has its own pool; the synthetic anchors are also processed as a separate population. The within-library sensitivity changes only the pool from which added peaks are drawn.

PrimeNovo's released mass correction allows no +1 isotope offset. The common-domain diagnostic restricts both target and decoy records to isotope-zero references and chemistry shared by the models, then repeats selection using the existing scores. Scalar calibrators are refitted under the same restriction and source partitions. This diagnostic uses reference annotations to assess input compatibility.

A.3 Selection and score calibration

At nominal level $\alpha \in \{0.01, 0.05\}$, independently decoded targets and decoys are pooled by native score. The primary rule selects the largest descending-score prefix satisfying $D/T \leq \alpha$; there is no pairwise target–decoy competition. Prefixes include complete equal-score groups. The plus-one sensitivity replaces D/T by $(D+1)/T$. An empty selection is reported when no prefix qualifies.

The score-calibration experiment uses *E. coli* 30mM as its source. Reference backbones, with I/L collapsed, are deterministically assigned to five groups. Three groups fit an isotonic map, one provides a source diagnostic set and one provides a source test. Sequence/site and ion90 maps use identical supported fitting rows and scores. The fitted maps are applied unchanged to other cohorts. Selection uses complete probability ties and the mean predicted error, $|A|^{-1}\sum_{i\in A}(1-p_i)$. Five rotations assess sensitivity to the source partition. A further analysis excludes external rows sharing source-fitting reference backbones.

Peptide-specific isotonic maps are fitted to source representatives. Source roles group identical predicted peptidoforms; a stricter sensitivity additionally removes test rows whose reference backbone appears in fitting.

Peptide assignment and inventory. For spectrum decoys, group eligible predictions within each arm by I/L-collapsed peptidoform (and library for synthetic data). For each group u, choose $r(u) = \arg\max_{i\in u} s_i$, using spectrum ID to break ties. The peptide threshold is then selected on these representatives, yielding accepted target groups $\mathcal{U}(t)$. The counterpart method groups literal sequences across arms and breaks ties by row ID (Appendix E). Assignment disagreement compares each representative with its precursor reference. Inventory error instead counts

$$\mathrm{FDP}_{\mathrm{presence}}(t) = \frac{\sum_{u\in\mathcal{U}(t)} \mathbf{1}\{\hat{y}_{r(u)} \notin \mathcal{M}\}}{|\mathcal{U}(t)| \vee 1}.$$

Our synthetic inventory comparison substitutes the intended library for $\mathcal{M}$, so it measures intended-library mismatch under the synthesis assumptions.

A.4 Fragment criteria and uncertainty

Observed reference peaks are matched to b/y ions at 0.02 Da, including water/ammonia losses and fragment charges one and two. Ion90 and ion100 measure the fraction of these observed peaks also explained by the prediction. Under the published rule, a spectrum with no matched reference peaks passes the coverage criterion. No accepted PSM has zero reference peaks in any of the three models' common-domain nominal-5% primary decoy selections, including the within-library pool sensitivity. Unsupported chemical tokens produce unknown outcomes. Lower and upper disagreement bounds assign those unknowns to agreement and disagreement, respectively, on the same accepted denominator.

Intervals for the common-input experiments use 2,000 fixed-seed resamples of I/L-collapsed reference-backbone groups. Models, calibration, thresholds and accepted sets remain fixed, so the intervals describe variation conditional on those choices. Acquisition/library resampling assesses grouping sensitivity. For zero-event sets, a one-sided binomial upper bound is also reported under an independent-unit assumption. The isotope and phosphotyrosine analyses use 3,000 backbone resamples where intervals are shown.

A.5 Training-sequence overlap

We compare reference backbones with MassIVE-KB v1 and the public InstaNovo AC-PT training-peptide list. MassIVE overlap is 35,694/36,049 human PSMs and 9,538/9,680 backbones; yeast overlap is 290/40,000 PSMs and 76/9,348 backbones; design-anchor overlap is 77/11,150 PSMs and 6/1,540 backbones. AC-PT PSM overlaps are 18,349, 216 and 22, respectively. Human consequently tests calibration transfer across studies under substantial sequence exposure. The proxy lists provide incomplete coverage of the evaluated models' training data.

A.6 Winnow features and selection

Winnow 2.0.0 uses its general-model weights and five-beam InstaNovo 1.2 outputs. Its 11 features are confidence, ion matches, matched intensity, iRT error, beam margin, median margin, entropy, z-score, minimum and standard deviation of token probability, and mass error. Prosit models `Prosit_2025_intensity_22PTM` and `Prosit_2025_irt_22PTM` provide intensity and retention predictions. The default collision energy is 27 with HCD. Within each acquisition, the top-confidence 10% of predictions fit the RT-to-iRT conversion; the neural confidence model stays fixed.

Feature eligibility requires the model's supported chemistry and at most 30 prediction tokens, counting a terminal modification as a separate token. All inputs have charge 2–4 and measured retention time. Sequence-mass agreement uses cumulative tolerance 0.5 Da and individual-residue tolerance 0.1 Da. Selection applies the published nonparametric rule: sort confidence p_i in descending order and retain scores at or above the largest-prefix cutoff satisfying the mean predicted error $|A|^{-1}\sum_{i\in A}(1-p_i) \le \alpha$. Complete-score-tie selection yields the same sets. Winnow's documented training-study accessions differ from the evaluation studies (InstaDeep, 2026).

For the synthetic data, HCD energy 40 is specified by the acquisition metadata. We evaluate this energy alongside the default on design anchors and supported anchors, holding weights and other feature settings fixed. Appendix C.4 separates restriction of the evaluated population from recomputation of the RT feature.

Confidence at the 1% threshold. Across the seven evaluated populations at energy 27 and the two anchor populations at energy 40, the maximum confidence ranges from 0.98497 to 0.98885. The corresponding minimum predicted error probabilities range from 1.115% to 1.503%. Every nonempty prefix therefore has mean predicted error above 1%, which accounts for the empty selections. At 5%, the same rule produces the selections in Tables 3 and 23.

A.7 Measurement sensitivity and selected control counts

Let $Y = 1$ denote an incorrect identity and $Z = 1$ a measurement flag, such as reference absence or a missing isotope counterpart. In a fixed selected population A, define $f = \mathbb{P}(Y = 1 \mid A)$, sensitivity $a = \mathbb{P}(Z = 1 \mid Y = 1, A)$ and false-positive probability $b = \mathbb{P}(Z = 1 \mid Y = 0, A)$. Then

$$\mathbb{P}(Z = 1 \mid A) = af + b(1 - f). \tag{3}$$

When a and b are known and unequal, $f = (\mathbb{P}(Z = 1 \mid A) - b)/(a - b)$. Open-pFind uses this mixture correction for isotope validation, estimating incorrect-match missingness from low-score tails and correct-match missingness from an eight-engine intersection (Chi et al., 2018, Supplementary Note 1). Its application to selected *de novo* predictions requires the corresponding sensitivities for their peptide classes, scores and reference coverage.

For a nonempty accepted set, let $D(t)$ be the selected decoy count, $R(t) = |A(t)|$, c the decoy-to-target scaling and δ a finite-sample correction. The estimate $\widehat{\mathrm{FDP}}(t) = (cD(t) + \delta)/R(t)$ differs from realized FDP by

$$\widehat{\mathrm{FDP}}(t) - \mathrm{FDP}_{\sim}(t) = \frac{cD(t) + \delta - V_{\sim}(t)}{R(t)}. \tag{4}$$

FDR control further depends on the procedure's exchangeability, dependence and threshold assumptions (Elias & Gygi, 2007; Wen et al., 2025; Reznikov et al., 2026). Our empirical comparisons substitute the specified reference disagreement for $V_{\sim}(t)$.

B ISOTOPE TRACES, FRAGMENTS AND GENOMIC REFERENCES

B.1 ACQUISITIONS AND REFERENCE ASSIGNMENTS

PXD008782 contains mixtures of unlabeled, ^{13}C-labeled and ^{15}N-labeled *E. coli* used in the OpenpFind study (Chi et al., 2018). The 30mM fraction supplies the source analysis and 60mM tests transfer of its trace rule. RAW spectra are converted with ThermoRawFileParser v2.0.0-dev, excluding reference and exception peaks. Retention times are expressed in seconds. The acquisitions contain 38,540 and 39,098 MS2 scans, respectively.

References require pFind target status, $q \leq 0.001$, unlabeled identity, charge 2–4 and only carbamidomethylation/oxidation. RAW precursor charge must agree, and mass agreement must be within 20 ppm. Scans with multiple surviving distinct identities are excluded. These conditions yield 4,458 PSMs with 2,295 I/L-collapsed backbones in 30mM and 7,046 PSMs in 60mM. Casanovo provides 4,447 and 7,035 target predictions, respectively. The two 30mM references with unmodified cysteine are excluded from the shared-vocabulary measurement analysis.

B.2 CHEMICAL IDENTITY AND TRACE EXTRACTION

Full CHNOSP formulas include residues, terminal water and each modification. Metabolic labels apply to residue C/N atoms; Carbamidomethylation (CAM) reagent atoms are unlabeled, oxidation adds oxygen and deamidation removes one label-eligible N. Thus query coordinates depend on both full neutral mass and label-eligible atom counts, at the observed charge and retention time. Five observed formula-equal pairs have different label-eligible counts because their CAM inventories differ.

For each query, centroid intensities are summed within 20 ppm over ± 180 seconds around MS2. Each labeled channel is compared with the unlabeled channel. Both comparisons require at least three jointly nonzero MS1 scans, full-window Pearson correlation ≥ 0.7 and global-apex separation ≤ 20 seconds. RT30 additionally requires all three apices within 30 seconds of MS2 and mean intensity in each first/last 15-second window segment at most half its channel maximum. RT15 and RT60 vary only the MS2 proximity requirement. RT30 was developed by inspecting 30mM traces and applied unchanged to 60mM.

The proximity and edge conditions remove pairs whose correlated peak lies far from MS2 or at the extraction boundary. In 30mM, 44, 69 and 98 of the 459 nonalias identical-query disagreements pass RT15, RT30 and RT60. The corresponding reference counts are 962, 1,605 and 1,814 out of 4,458. The global-apex criterion can exclude traces with multiple peaks, so these rules define the measured support used in the comparisons.

Table 8: Trace sensitivity among the same nominal-5% accepted Casanovo 5.2 predictions in 30mM. All rows require identical isotope queries and exclude residue-composition aliases.

Rule	PSMs	Backbones	Also ion90	Also ion100
Coelution	26	20	20	5
RT15	7	7	7	0
RT30	12	11	12	1
RT60	15	13	13	1

B.3 TRANSFER TO THE 60MM ACQUISITION

The 60mM predictions contain 593 formula-preserving reference disagreements, of which 587 preserve label-eligible C/N counts. Removing 48 residue-level aliases leaves 539 cases; 181 pass coelution and 78 pass RT30. Nominal-5% selection contains 16 cases passing both RT30 and ion90, with 16 distinct reference backbones. These backbones are disjoint from the 11 represented by the 12 joint 30mM cases. RT15 and RT60 give 11 and 19 joint cases. At nominal 1%, the only RT30 case fails ion90.

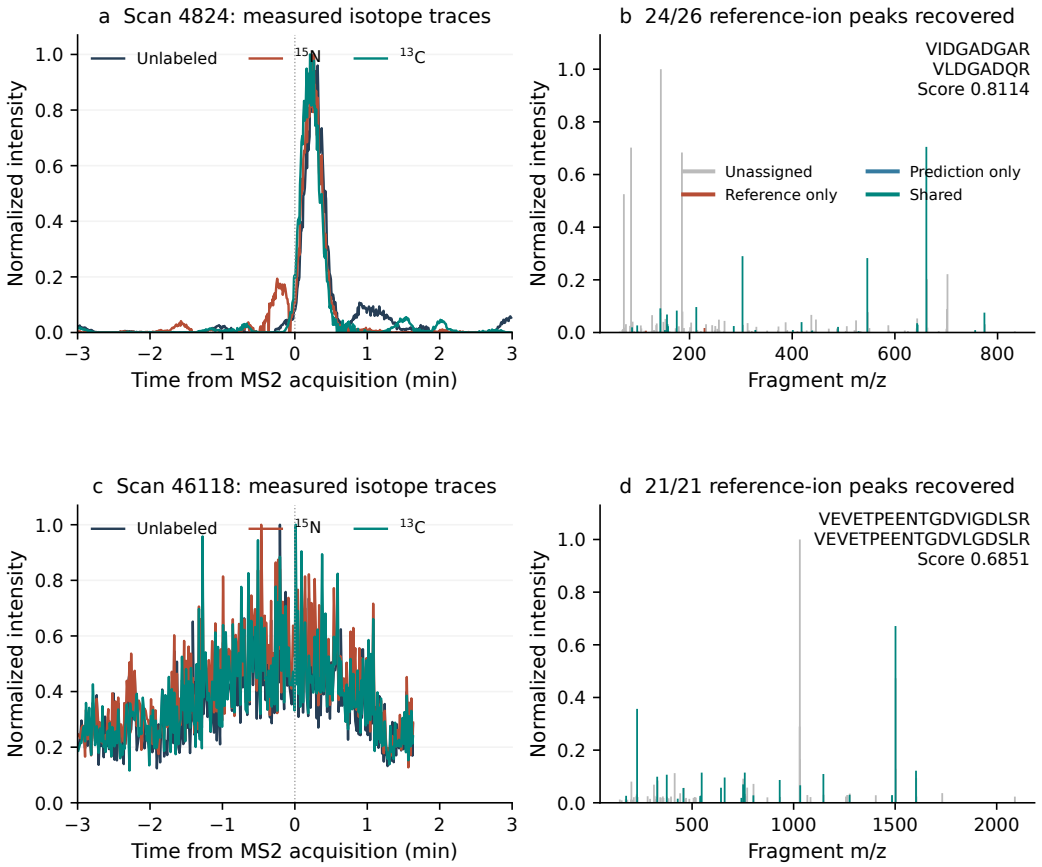

Figure 6: **Isotope traces and fragment matches in 30mM.** Top: the highest-score nonalias disagreement passing RT30. Bottom: the highest-score nominal-5% accepted case also recovering all observed reference-ion peaks. Both references occur in the inspected protein and genomic sequences; neither prediction does. MS2 colors distinguish observed peaks matched by the two candidates, with unassigned peaks visible.

Table 9: Casanovo 5.2 decoy selections in the 60mM acquisition. Rates are reference disagreement percentages.

Rule	Nominal	Targets	Decoys	Sequence	Ion90	Formula
D/T	1%	1,244	12	0.322	0.322	0.161
D/T	5%	5,764	288	3.383	1.579	1.301
Complete ties, $+1$	1%	1,242	11	0.322	0.322	0.161
Complete ties, $+1$	5%	5,763	287	3.384	1.579	1.301

B.4 Protein and translated-genome membership

References are NCBI RefSeq GCF_000005845.2, chromosome NC_000913.3 and its 4,300 annotated proteins (NCBI RefSeq, 2026). Exact I/L-collapsed backbones are searched within each protein and six continuous reading frames, with stops and record boundaries delimiting matches. These are reference-strain resources rather than sample-matched sequencing.

Four protein-supported peptides fail the continuous six-frame search: `MSTLEQK` (*rimP*, TTG start), `MDDLTAQALK` (*syd*, GTG), `MEIYENENDQVEAVK` (*yfgM*, GTG) and `GYLDYDAK` (*prfB*, programmed frameshift). Reconstruction from annotated CDS segments and initiation recovers all four proteins. These peptides are therefore absent from a continuous six-frame translation despite being encoded by the genome.

B.5 Composition of selected decoys

At nominal 5%, 108/164 selected 30mM decoys and 185/288 selected 60mM decoys predict their unperturbed spectrum's reference sequence. The remaining 56 and 103 decoy disagreements compare with 108 and 195 disagreements among accepted targets. Let V_{ref} count reference-discordant accepted targets. Writing the decoy count as reference matches plus disagreements, $D = C + E_D$, gives

$$D - V_{\mathrm{ref}} = C + (E_D - V_{\mathrm{ref}}) = 56 \text{ and } 93$$

for the two acquisitions. Reference-matching decoys make a positive contribution to $D - V_{\mathrm{ref}}$, while the remaining decoy disagreements are fewer than the target disagreements in both acquisitions.

The fraction preserving the source-relative isotope query also changes under selection. Among all reference disagreements, it is 25.9%/26.2% for decoys and 57.6%/58.3% for targets in 30mM/60mM. Among selected disagreements, those fractions become 82.1%/78.6% and 64.8%/61.5%, respectively. The ordering reverses in the selected tail. These are structural comparisons to each source spectrum's reference; the decoys have no acquired MS1 traces.

C Synthetic identities and population comparisons

C.1 Library design and precursor-compatible candidates

PXD000138 provides a synthesis workbook and experimental description (Marx et al., 2013). Its 96 plate libraries enumerate 101,520 backbone entries and 91,656 distinct I/L-collapsed backbones. The design specifies terminal K/R, variable positions and a designated S/T/Y phosphorylation site. Candidate peptidoforms include the unmodified or designated-site-phosphorylated peptide and all methionine oxidation combinations.

RAW plate-vial metadata and deposited search assignments link acquisitions to libraries. One acquisition has conflicting assignments (vial H10 and search library H9) and is excluded. Of the 95 linked acquisitions, 83 contribute eligible spectra. For each measured precursor, C_i contains the library peptidoforms within 5 ppm, allowing isotope offsets zero and one.

The full synthetic population contains 131,036 spectra with at least one unmodified, cysteine-free, length-7–30 candidate in C_i. All compatible competitors, including modified ones, contribute to identity uncertainty. The *design anchors* are the 11,150 spectra with one compatible identity after I/L normalization, spanning 1,540 backbones. They form a separate selection population with their own decoys. Of these, 6,339 remain singletons both at 10 ppm and after adding one-residue synthesis deletions at 5 ppm.

C.2 Identity bounds and chemical equivalence

The interpretation of C_i assumes the designated library, enumerated modifications and precursor tolerances, with the generating identity in C_i. For a nonempty accepted set A, let $Q \subseteq A$ contain records whose prediction chemistry cannot be parsed, and let $K = A \setminus Q$. Define the disjoint sets

$$B = \{i \in K : \hat{y}_i \notin C_i\}, \qquad M = \{i \in K : \hat{y}_i \in C_i,\ |C_i| > 1\}.$$

Here B contains known incompatible predictions and M contains compatible predictions with unresolved identity. The conditional assignment-error proportion is bounded by

$$L = \frac{|B|}{|A|}, \qquad U = \frac{|B| + |M| + |Q|}{|A|}.$$

Unparsed predictions contribute only to the upper bound; both denominators retain every accepted record. A sensitivity expands the candidates by at most one N/Q deamidation while preserving oxidation and phosphorylation states. Exact modified strings and ordered residue compositions are evaluated separately: deamidated N/D and Q/E can have the same residue composition despite different representations. These bounds condition on candidate-set coverage. A correctly sequenced synthesis deletion, truncation or contaminant outside C_i enters B; such species affect the interpretation of L as physical assignment error.

C.3 Supported anchors

Supported anchors add fragment-search evidence to the robust design identity. Deposited phosphoRS/Mascot assignments are linked by acquisition and scan, with matching charge, precursor m/z within 1 ppm and retention time within one second. The population requires one High-confidence rank-1 annotation agreeing with the robust singleton and an isotope-zero precursor. It contains 2,190 spectra, 440 reference backbones and 70 libraries. This post hoc subset conditions on search agreement within the synthesis library, enriching for spectra that support a search assignment.

Target predictions are shared with the design-anchor population, and decoys are generated from the supported population's own peak pool. Each model's formula-preserving sequence disagreements exclude residue-composition aliases. The example in Figure 2 is the highest-score InstaNovo ion90 case. Its companion scan is chosen by nearest retention time among scans with the same reference backbone and a different precursor charge, independently of fragment recovery.

C.4 Population restriction and Winnow features

The population comparison separates restricting spectra from recalculating confidence features. We compare three operations: intersecting the design-anchor selection with the supported population, reselecting the threshold using the design-anchor probabilities, and recomputing features on the supported population before selection. References and base predictions are identical for matched rows.

Table 10: Winnow population and feature comparisons at nominal 5%. All rows use the same references and base predictions. Fixed-cutoff rows intersect an existing selection; the other rows reselect thresholds in the stated population. Rates use sequence-mass disagreement.

Energy	Population / computation	Eligible	Disagr. / accepted	Rate (%)
27	Design anchors	9,776	739/5,367	13.77
27	Supported: fixed anchor cutoff	2,105	28/1,646	1.70
27	Supported: reselect anchor features	2,105	37/1,722	2.15
27	Supported: recompute features	1,499	31/1,339	2.32
27	Common eligible: anchor features	1,499	30/1,336	2.25
27	Common eligible: recomputed features	1,499	31/1,339	2.32
40	Design anchors	9,776	722/5,333	13.54
40	Supported: fixed anchor cutoff	2,105	25/1,652	1.51
40	Supported: reselect anchor features	2,105	33/1,728	1.91
40	Supported: recompute features	1,499	27/1,335	2.02
40	Common eligible: anchor features	1,499	27/1,332	2.03
40	Common eligible: recomputed features	1,499	27/1,335	2.02

At energy 27, restriction and reselection with the design-anchor features give 37/1,722 disagreements (2.15%), close to 31/1,339 (2.32%) after recomputation. Computing features on the design-anchor population yields scores for 2,105 supported spectra. Computing them within the supported population yields 1,499, because fewer spectra meet the acquisition-level RT-regression requirements. Among the 1,499 common eligible rows, only iRT error changes; paired selections give 30/1,336 and 31/1,339 disagreements. Table 10 gives the corresponding energy-40 comparisons.

C.5 Noise-pool sensitivity

The within-library decoy construction draws added peaks from each anchor spectrum's own synthesis library. The retained-peak fraction, randomization, target scores and model settings are held fixed. Appendix F shows both constructions at 1% and 5%, with PSM assignment, peptide-representative assignment and intended-library membership evaluated separately.

D PHOSPHORYLATION-SITE CONSTRAINTS

D.1 REFERENCE SUBSET AND INFERENCE

The PXD009449-pY acquisition contains 45,180 spectra from a tyrosine-phosphorylated peptide study (Zolg et al., 2018). The reference subset requires one phosphorylation on the unique Y, no reverse/contaminant flag, MaxQuant PEP ≤ 0.01, score ≥ 40, delta score ≥ 6, charge two or three and membership in the Ymod FASTA. Precursor m/z agreement within 0.01 Da gives 998 PSMs and 151 backbones. FASTA membership uses a substring search in the concatenated pool entries; the backbone assignment comes from the search reference.

OmniNovo epoch 80 uses seed 42, 300 input peaks, batch 2,048, bfloat16 and one beam with its configured mass control. NovoBoard's random generator retains 50% of source peaks for the primary comparison and 40% for sensitivity, with its published seed-99 behavior. Selection uses pooled D/T at 1% and 5%.

A site contradiction requires the same I/L-equivalent backbone and complete modification inventory, including carbamidomethylation, with the reference phosphate on the unique Y and the prediction's phosphate on S or T. The comparison therefore conditions on the assigned backbone and modification inventory.

D.2 ACCEPTED SITE CONTRADICTIONS

The primary nominal-5% selection contains nine site contradictions, all passing ion90 and six passing ion100. The 40%-retained sensitivity yields ten at nominal 1% and eleven at 5%. Table 11 reports the full selections and the reference-subset counts; Table 12 lists the individual cases.

Table 11: OmniNovo decoy selections and conditional site comparisons. Full-cohort rates are percentages. Ion90 ranges assign unsupported chemistry to agreement or disagreement. Subset counts describe records within the full accepted sets; Ref. disagr. denotes reference disagreements.

Retained / nominal	Targets	Decoys	Sequence/site	Ion90	Formula
50% / 1%	33,482	334	0.221	0.149	0.122
50% / 5%	40,091	2,004	2.150	0.384–0.397	0.192
40% / 1%	40,105	401	2.169	0.384–0.396	0.192
40% / 5%	41,982	2,099	4.035	1.022–1.210	0.705
	Subset	Ref. disagr.	Y→S/T	ion90 pass	ion100 pass
50% / 1%	857	0	0	0	0
50% / 5%	957	9	9	9	6
40% / 1%	958	10	10	10	7
40% / 5%	982	21	11	11	8

Unsupported prediction chemistry affects 0, 5, 5 and 79 records in the four accepted sets, respectively; every site-contradiction case has a defined ion criterion. For the primary 5% selection, backbone intervals are 1.44–3.00% for full-cohort sequence disagreement and 0.33–1.93% for the conditional site-contradiction fraction among 957 subset records. The 40%-retained 1% selection gives 1.46–3.02% and 0.36–2.19%, respectively. Complete-tie grouping leaves these primary counts unchanged. Plus-one selection changes the default 1% subset from 857 to 854 records, with no observed disagreement in either set.

Table 12: All 11 conditional site contradictions, ordered by confidence score. Positions count from the N terminus. Backbone notation follows the reference; prediction identity collapses I/L. Membership columns denote 50%-retained/5%, 40%-retained/1%, and 40%-retained/5%. No case enters the default nominal-1% set.

Scan	Reference backbone	Site move	Score	Ions	50/5	40/1	40/5
3421	GPQNATDSYVHK	Y9→S8	0.956415	20/20	+	+	+
40710	LQEVEAEVAATGTYQLR	Y14→T13	0.952161	27/27	+	+	+
43802	AAVEINVAVLHSYQLAK	Y13→S12	0.949059	23/23	+	+	+
19393	SYHLQIVTK	Y2→S1	0.947298	22/22	+	+	+
36179	LSITGTYDLK	Y7→T6	0.914507	28/29	+	+	+
18100	SYCAEIAHNVSSK	Y2→S1	0.908940	34/35	+	+	+
56465	VSFLSALEEYTK	Y10→T11	0.908232	21/21	+	+	+
1858	YTTNNSSEQR	Y1→T2	0.905736	21/21	+	+	+
24347	SYHLQIVTK	Y2→S1	0.886220	20/21	+	+	+
43008	AAVEINVAVLHSYQLAK	Y13→S12	0.882414	24/24	–	+	+
40035	LYHSEAFTVNFGDTEEAK	Y2→S4	0.753620	21/21	–	–	+

E Counterpart calibration at PSM and peptide levels

E.1 Features, fitting and independent calibration

π-NovoQC constructs learned counterparts in feature space and estimates a counting factor γ (Liang et al., 2026). We use its public GraphNovo features, which contain 37,366 targets and an equal number of counterparts from *A. thaliana*, *C. elegans* and *E. coli*. The 33 features and sequence-mass labels come from the published implementation. Sequence-mass matching uses cumulative tolerance 0.5 Da and residue tolerance 0.1 Da; exact modified sequence and ordered residue composition are evaluated separately.

Each species is held out in turn. Before assigning source roles, we remove source reference backbones that occur in the held-out species, with modifications removed and I/L collapsed. Reference backbones are deterministically assigned to ten groups. Six source groups are used for training, two for validation, one for source calibration and one for source testing. One group in the held-out species supplies local calibration; the remaining nine supply the external test set. Each target and its counterpart stay in the same role.

Strict and relaxed MLPs use the official architecture, dropout zero, AdamW learning rate 0.001, weight decay 10^{-5}, batch 1,024 and seed 42. Fitting lasts at most 50 epochs, with validation-AUROC early stopping after eight epochs without improvement. Nonfinite features become zero and scaling uses training rows. The strict head retains counterpart sequence-mass labels; the relaxed head labels every counterpart negative. Their mean probability is the combined score.

We estimate separate source and local calibration factors using the published procedure at nominal 5%. The test rule is $\gamma C/T \leq 0.05$, where C and T count counterparts and targets above the score threshold. Source and local factors apply to the same external scores. We use the published selection rule and also evaluate selection with equal scores grouped together.

E.2 PSM results and discrete calibration

Source and local calibration yield different external rates even though the predictions and scores are fixed (Table 13). For *A. thaliana*, source calibration gives 540/6,917 disagreements and local calibration gives 180/4,684. The source-domain test already gives 109/1,749 (6.232%), identifying imperfect source calibration as part of the external excess.

The published factor is undefined for three calibration sets because none has an observed FDP within 10^{-4} of 0.05. A nearest-point sensitivity replaces the tolerance condition for these cases, leaving defined factors and scores unchanged (Table 14). *C. elegans* remains below nominal under local calibration; *E. coli* gives 5.364% with a conditional interval of 4.678–6.054%.

Table 13: External PSM results with the published calibration rule. Cells show disagreement percentage and accepted targets. Undefined denotes a calibration factor that the procedure cannot estimate from the supplied calibration set.

Held-out species	$\gamma = 1$	Source calibration	Local calibration
A. thaliana	5.875 (5,991)	7.807 (6,917)	3.843 (4,684)
C. elegans	3.575 (5,091)	Undefined	4.921 (6,442)
E. coli	4.769 (7,465)	Undefined	Undefined

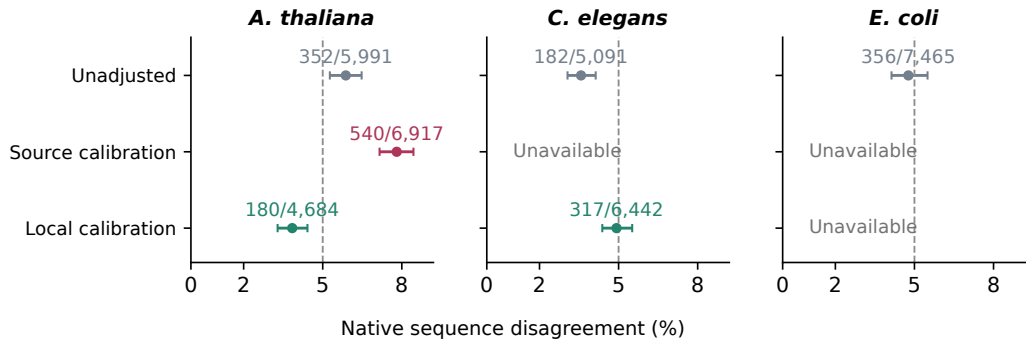

Figure 7: **Calibration source changes external PSM selection.** Three held-out species at nominal 5%, using the published sequence-mass criterion. Source and local factors use separate calibration populations. Labels give disagreements/accepted targets; intervals resample reference backbones with fitting and selection fixed.

Table 14: Nearest-point calibration sensitivity on the same external scores.

	Source calibration		Local calibration	
Species	γ	Rate (%), targets	γ	Rate (%), targets
A. thaliana	0.554839	7.807, 6,917	1.923077	3.843, 4,684
C. elegans	1.075758	3.388, 4,782	0.530120	4.921, 6,442
E. coli	1.129630	4.614, 7,304	0.777778	5.364, 7,849

E.3 Separate peptide heads after PSM selection

The peptide analysis follows the published two-stage workflow. A nominal-5% PSM stage supplies candidates under unadjusted and, where defined, source-adjusted factors. Peptides are aggregated by literal predicted sequence across both arms, keeping the highest-PSM-score representative and breaking ties by row ID. Thus a target/counterpart collision can change which arm represents the peptide.

Peptide fitting, validation, source calibration and source test use disjoint predicted sequences. Conflicts are resolved in the order calibration, test, validation and training. Sequences occurring in the external set are removed from all source peptide roles. The preceding PSM heads retain their reference-backbone partitions. Strict and relaxed peptide MLPs use the same 33 features and fitting settings described above. Their mean score is calibrated by a separate peptide factor. Complete-score-tie prefixes define the 1% and 5% peptide selections.

At nominal peptide level 5%, unadjusted stages give 308/5,002 disagreements for *A. thaliana*, 121/3,926 for *C. elegans* and 147/2,428 for *E. coli*. Adjusted factors are defined for both stages in *A. thaliana*, giving 387/5,301. At 1%, the source-adjusted PSM stage with an unadjusted peptide stage accepts only four *A. thaliana* representatives: zero disagree under sequence-mass matching and one differs in exact modified sequence. Figure 8 shows both rates and their uncertainty.

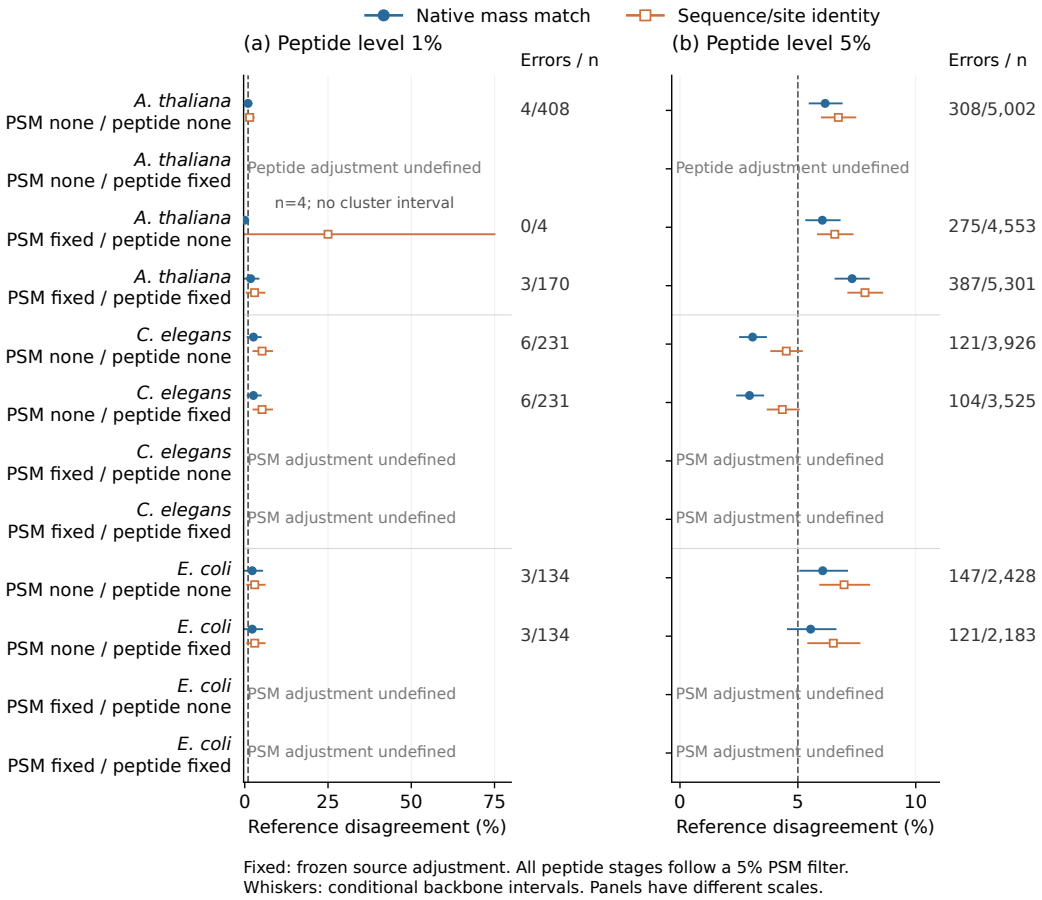

Figure 8: **Two-stage counterpart selection at both peptide levels.** The PSM stage is fixed at 5%; rows compare unadjusted and source-adjusted factors for each stage. Points evaluate sequence-mass and sequence/site disagreement on the same peptide representatives. Labels give sequence-mass disagreements/accepted representatives; intervals resample reference backbones. Panel ranges accommodate the wide interval for the four-representative selection.

F Complete operating results and sensitivities

The tables report every model, population, nominal level and discovery unit in the three-model decoy comparison. String, Residue, Formula and Ion90 use the definitions in Table 6. All rates are percentages. Ranges assign unknown comparisons to either outcome on the full accepted denominator; empty selections are labeled separately. Design and supported anchors are nested subsets of the full synthetic population.

F.1 Unrestricted selections

The synthetic-full rows use single-reference comparisons: a reference is assigned only when the precursor-compatible design set is a singleton. Multiple-candidate spectra have unknown sequence, composition and ion comparisons. Their ranges are $[e/N, (e+u)/N]$, with e known disagreements and u unknown comparisons among N accepted records. Candidate-set bounds in Table 4 also identify predictions outside non-singleton sets. Thus Casanovo's nominal-5% String range starts at 2.94%, whereas its candidate-set lower bound is 43.42% on the same 75,524 records.

Table 15: Unrestricted PSM selections at nominal 1%. Rates are percentages; ranges retain unknown endpoints on the full accepted denominator. Synthetic-full ranges use singleton-design references, with non-singleton sets treated as unknown (Appendix F).

Population	Model	Accepted	String	Residue	Formula	Ion90
E. coli 30mM	Casanovo	514	0.78	0.78	0.58	0.78
E. coli 30mM	InstaNovo	2,209	1.63	1.09	0.59	0.63–0.72
E. coli 30mM	π-PrimeNovo	1,826	0.66	0.27	0.16	0.22–0.27
E. coli 60mM	Casanovo	1,244	0.32	0.32	0.16	0.32
E. coli 60mM	InstaNovo	3,939	1.29	1.04	0.25	0.38–0.41
E. coli 60mM	π-PrimeNovo	3,302	0.27	0.27	0.12	0.15–0.21
Human	Casanovo	15,410	1.12	1.04	0.27	0.33
Human	InstaNovo	19,946	1.35	1.20	0.38	0.35–0.40
Human	π-PrimeNovo	20,615	5.55	5.52	5.23	4.97–5.18
Yeast	Casanovo	2,000	0.30	0.30	0.15	0.25
Yeast	InstaNovo	26,366	3.31	2.74	1.20	1.17–1.27
Yeast	π-PrimeNovo	32,060	17.54	17.33	16.90	16.34–16.81
Synthetic full	Casanovo	48,903	2.04–90.43	2.01–90.40	1.83–90.22	0.73–89.12
Synthetic full	InstaNovo	53,671	1.78–91.07	1.72–91.02	1.59–90.89	0.54–90.00
Synthetic full	π-PrimeNovo	51,340	2.00–90.42	1.99–90.40	1.86–90.28	0.80–89.45
Design anchors	Casanovo	5,945	18.96	18.64	16.75	6.96
Design anchors	InstaNovo	5,323	15.29	14.88	13.79	4.60–6.07
Design anchors	π-PrimeNovo	6,115	17.96	17.81	16.57	7.34–9.40
Anchors, local pool	Casanovo	2,661	14.77	14.77	13.60	5.71
Anchors, local pool	InstaNovo	4,396	11.81	11.58	10.83	3.55–4.78
Anchors, local pool	π-PrimeNovo	5,134	15.08	15.02	14.12	5.40–7.34
Supported anchors	Casanovo	1,740	4.89	4.60	2.13	1.61
Supported anchors	InstaNovo	1,728	2.49	2.31	1.04	0.46–0.52
Supported anchors	π-PrimeNovo	1,727	1.62	1.62	0.81	0.29–0.46

Table 16: Unrestricted PSM selections at nominal 5%. Rates are percentages; ranges retain unknown endpoints on the full accepted denominator. Synthetic-full ranges use singleton-design references, with non-singleton sets treated as unknown (Appendix F).

Population	Model	Accepted	String	Residue	Formula	Ion90
E. coli 30mM	Casanovo	3,297	3.28	2.61	1.15	1.58
E. coli 30mM	InstaNovo	3,660	5.87	4.26	0.96	1.34–1.42
E. coli 30mM	π-PrimeNovo	3,467	3.03	1.67	0.55	0.84–0.92
E. coli 60mM	Casanovo	5,764	3.38	3.14	1.30	1.58
E. coli 60mM	InstaNovo	6,355	5.33	4.56	0.94	1.37–1.40
E. coli 60mM	π-PrimeNovo	6,033	2.49	1.69	0.46	0.66–0.71
Human	Casanovo	26,897	7.92	7.60	1.25	1.74
Human	InstaNovo	29,017	6.22	5.71	0.61	0.88–0.94
Human	π-PrimeNovo	28,465	9.63	9.24	7.25	6.83–7.21
Yeast	Casanovo	32,766	6.71	6.24	2.61	2.83
Yeast	InstaNovo	37,026	7.53	6.53	1.96	2.24–2.39
Yeast	π-PrimeNovo	36,924	21.40	20.64	19.15	18.54–19.09
Synthetic full	Casanovo	75,524	2.94–92.96	2.88–92.89	2.36–92.38	1.20–91.21
Synthetic full	InstaNovo	80,705	2.39–92.94	2.31–92.86	1.95–92.50	0.79–91.51
Synthetic full	π-PrimeNovo	82,035	2.74–93.19	2.68–93.13	2.33–92.78	1.17–91.96
Design anchors	Casanovo	7,578	29.74	29.12	23.87	12.13
Design anchors	InstaNovo	7,293	23.09	22.21	19.16	7.38–9.12
Design anchors	π-PrimeNovo	7,908	29.22	28.63	24.80	12.43–16.14
Anchors, local pool	Casanovo	6,580	22.60	22.05	19.18	8.48
Anchors, local pool	InstaNovo	5,800	16.81	16.24	14.97	5.14–6.66
Anchors, local pool	π-PrimeNovo	6,508	19.91	19.67	18.07	8.25–10.49
Supported anchors	Casanovo	1,965	11.15	10.74	3.87	3.31
Supported anchors	InstaNovo	1,967	6.46	6.05	2.24	1.22–1.32

Table 16 (continued)

Population	Model	Accepted	String	Residue	Formula	Ion90
Supported anchors	π-PrimeNovo	1,960	5.71	5.56	1.68	0.77–1.22

Table 17: Unrestricted peptide-representative selections at nominal 1%. Rates are percentages; ranges retain unknown endpoints on the full accepted denominator. Synthetic-full ranges use singleton-design references, with non-singleton sets treated as unknown (Appendix F).

Population	Model	Accepted	String	Residue	Formula	Ion90
E. coli 30mM	Casanovo	364	0.55	0.55	0.27	0.55
E. coli 30mM	InstaNovo	1,339	1.27	0.82	0.37	0.45–0.52
E. coli 30mM	π-PrimeNovo	1,114	0.72	0.36	0.27	0.27–0.36
E. coli 60mM	Casanovo	835	0.36	0.36	0.12	0.36
E. coli 60mM	InstaNovo	2,468	1.42	1.13	0.16	0.28–0.32
E. coli 60mM	π-PrimeNovo	1,568	0.26	0.26	0.13	0.19–0.26
Human	Casanovo	472	0.85	0.64	0.21	0.21
Human	InstaNovo	5,049	1.25	1.09	0.57	0.48–0.57
Human	π-PrimeNovo	5,116	7.72	7.72	7.47	7.08–7.37
Yeast	Casanovo	515	0.58	0.58	0.19	0.58
Yeast	InstaNovo	5,801	3.38	2.91	1.81	1.65–1.90
Yeast	π-PrimeNovo	9,064	28.28	28.13	27.71	26.68–27.67
Synthetic full	Casanovo	10,861	1.72–95.85	1.72–95.85	1.50–95.63	0.71–94.83
Synthetic full	InstaNovo	20,521	1.89–96.76	1.84–96.71	1.64–96.52	0.65–95.74
Synthetic full	π-PrimeNovo	19,054	2.25–96.54	2.24–96.53	2.06–96.34	0.96–95.55
Design anchors	Casanovo	746	34.85	34.85	31.64	13.94
Design anchors	InstaNovo	786	31.42	31.30	29.01	10.56–14.12
Design anchors	π-PrimeNovo	1,058	40.64	40.45	37.52	15.31–20.42
Anchors, local pool	Casanovo	144	34.72	34.72	32.64	13.89
Anchors, local pool	InstaNovo	706	30.45	30.45	28.05	10.06–13.46
Anchors, local pool	π-PrimeNovo	841	35.79	35.67	33.06	12.25–17.24
Supported anchors	Casanovo	313	7.03	6.71	3.51	1.60
Supported anchors	InstaNovo	268	2.61	2.61	1.87	0.75–1.12
Supported anchors	π-PrimeNovo	311	3.54	3.54	2.57	0.64–0.96

Table 18: Unrestricted peptide-representative selections at nominal 5%. Rates are percentages; ranges retain unknown endpoints on the full accepted denominator. Synthetic-full ranges use singleton-design references, with non-singleton sets treated as unknown (Appendix F).

Population	Model	Accepted	String	Residue	Formula	Ion90
E. coli 30mM	Casanovo	1,925	1.71	1.19	0.52	0.94
E. coli 30mM	InstaNovo	2,009	4.63	3.38	0.65	0.80–0.95
E. coli 30mM	π-PrimeNovo	2,049	2.54	1.27	0.59	0.73–0.88
E. coli 60mM	Casanovo	3,537	2.37	2.18	0.71	0.99
E. coli 60mM	InstaNovo	3,747	4.83	4.16	0.53	0.91–0.93
E. coli 60mM	π-PrimeNovo	3,628	1.96	1.27	0.36	0.50–0.55
Human	Casanovo	7,520	6.13	5.90	1.01	1.46
Human	InstaNovo	7,752	5.90	5.47	0.90	1.12–1.26
Human	π-PrimeNovo	9,131	15.62	15.46	14.13	13.37–13.96
Yeast	Casanovo	7,332	4.54	4.34	2.71	2.95
Yeast	InstaNovo	8,624	8.50	7.58	2.84	3.08–3.41
Yeast	π-PrimeNovo	11,883	34.06	33.61	32.15	30.87–32.03
Synthetic full	Casanovo	25,643	2.67–97.23	2.60–97.16	2.10–96.67	1.06–95.62
Synthetic full	InstaNovo	31,899	2.56–97.63	2.48–97.55	2.00–97.07	0.87–96.17
Synthetic full	π-PrimeNovo	32,106	2.87–97.67	2.82–97.61	2.45–97.25	1.31–96.47
Design anchors	Casanovo	1,405	48.40	47.54	41.07	18.08

Table 18 (continued)

Population	Model	Accepted	String	Residue	Formula	Ion90
Design anchors	InstaNovo	1,308	45.34	44.27	39.91	15.90–19.95
Design anchors	π-PrimeNovo	1,580	52.22	51.71	46.33	22.41–28.35
Anchors, local pool	Casanovo	1,055	37.91	37.54	33.36	14.03
Anchors, local pool	InstaNovo	1,054	38.61	37.67	34.25	13.09–17.27
Anchors, local pool	π-PrimeNovo	1,251	44.36	43.96	40.93	17.83–23.10
Supported anchors	Casanovo	390	16.15	15.13	6.92	5.38
Supported anchors	InstaNovo	383	13.58	12.53	5.74	2.87–3.13
Supported anchors	π-PrimeNovo	362	7.18	6.91	3.04	1.10–1.93

F.2 Common chemistry and isotope-zero references

Both arms are restricted to reference chemistry shared by the models and isotope offset zero before threshold selection. The ambiguous full synthetic population has no unique reference-derived isotope offset and is excluded from this diagnostic.

Table 19: Common-domain PSM selections at nominal 1%. Rates are percentages; ranges retain unknown endpoints on the full accepted denominator. Synthetic identities use the composition model in Appendix C.

Population	Model	Accepted	String	Residue	Formula	Ion90
E. coli 30mM	Casanovo	514	0.78	0.78	0.58	0.78
E. coli 30mM	InstaNovo	2,209	1.63	1.09	0.59	0.63–0.72
E. coli 30mM	π-PrimeNovo	1,826	0.66	0.27	0.16	0.22–0.27
E. coli 60mM	Casanovo	1,244	0.32	0.32	0.16	0.32
E. coli 60mM	InstaNovo	3,939	1.29	1.04	0.25	0.38–0.41
E. coli 60mM	π-PrimeNovo	3,302	0.27	0.27	0.12	0.15–0.21
Human	Casanovo	14,469	1.00	0.93	0.21	0.28
Human	InstaNovo	18,515	1.22	1.06	0.30	0.28–0.32
Human	π-PrimeNovo	19,413	0.55	0.50	0.21	0.21–0.23
Yeast	Casanovo	2,307	0.35	0.35	0.22	0.30
Yeast	InstaNovo	21,345	2.49	1.99	0.83	0.85–0.92
Yeast	π-PrimeNovo	26,512	1.46	1.23	0.75	0.70–0.78
Design anchors	Casanovo	5,300	9.91	9.55	7.55	3.64
Design anchors	InstaNovo	4,750	6.57	6.11	4.91	1.52–1.89
Design anchors	π-PrimeNovo	5,405	7.18	7.01	5.61	2.42–3.07
Anchors, local pool	Casanovo	2,577	6.29	6.29	5.04	2.21
Anchors, local pool	InstaNovo	3,978	4.30	4.07	3.32	1.08–1.48
Anchors, local pool	π-PrimeNovo	4,547	5.61	5.54	4.55	1.78–2.27
Supported anchors	Casanovo	1,740	4.89	4.60	2.13	1.61
Supported anchors	InstaNovo	1,728	2.49	2.31	1.04	0.46–0.52
Supported anchors	π-PrimeNovo	1,727	1.62	1.62	0.81	0.29–0.46

Table 20: Common-domain PSM selections at nominal 5%. Rates are percentages; ranges retain unknown endpoints on the full accepted denominator. Synthetic identities use the composition model in Appendix C.

Population	Model	Accepted	String	Residue	Formula	Ion90
E. coli 30mM	Casanovo	3,297	3.28	2.61	1.15	1.58
E. coli 30mM	InstaNovo	3,660	5.87	4.26	0.96	1.34–1.42
E. coli 30mM	π-PrimeNovo	3,467	3.03	1.67	0.55	0.84–0.92
E. coli 60mM	Casanovo	5,764	3.38	3.14	1.30	1.58
E. coli 60mM	InstaNovo	6,355	5.33	4.56	0.94	1.37–1.40
E. coli 60mM	π-PrimeNovo	6,033	2.49	1.69	0.46	0.66–0.71
Human	Casanovo	25,020	7.45	7.13	1.08	1.56
Human	InstaNovo	26,748	5.65	5.16	0.48	0.76–0.81
Human	π-PrimeNovo	26,353	2.82	2.42	0.41	0.42–0.60

Table 20 (continued)

Population	Model	Accepted	String	Residue	Formula	Ion90
Yeast	Casanovo	27,029	5.21	4.82	2.01	2.23
Yeast	InstaNovo	30,054	5.92	4.95	1.37	1.66–1.77
Yeast	π-PrimeNovo	29,952	3.51	2.67	1.01	1.03–1.14
Design anchors	Casanovo	6,428	18.68	17.98	12.13	6.99
Design anchors	InstaNovo	6,220	12.17	11.21	7.86	3.20–3.63
Design anchors	π-PrimeNovo	6,547	14.68	13.98	9.61	4.86–6.48
Anchors, local pool	Casanovo	5,759	12.64	12.00	8.87	4.43
Anchors, local pool	InstaNovo	5,049	7.19	6.65	5.33	1.74–2.10
Anchors, local pool	π-PrimeNovo	5,654	8.12	7.84	6.12	2.78–3.48
Supported anchors	Casanovo	1,965	11.15	10.74	3.87	3.31
Supported anchors	InstaNovo	1,967	6.46	6.05	2.24	1.22–1.32
Supported anchors	π-PrimeNovo	1,960	5.71	5.56	1.68	0.77–1.22

Table 21: Common-domain peptide-representative selections at nominal 1%. Rates are percentages; ranges retain unknown endpoints on the full accepted denominator. Synthetic identities use the composition model in Appendix C.

Population	Model	Accepted	String	Residue	Formula	Ion90
E. coli 30mM	Casanovo	364	0.55	0.55	0.27	0.55
E. coli 30mM	InstaNovo	1,339	1.27	0.82	0.37	0.45–0.52
E. coli 30mM	π-PrimeNovo	1,114	0.72	0.36	0.27	0.27–0.36
E. coli 60mM	Casanovo	835	0.36	0.36	0.12	0.36
E. coli 60mM	InstaNovo	2,468	1.42	1.13	0.16	0.28–0.32
E. coli 60mM	π-PrimeNovo	1,568	0.26	0.26	0.13	0.19–0.26
Human	Casanovo	471	0.85	0.64	0.21	0.21
Human	InstaNovo	4,850	1.09	0.93	0.45	0.39–0.45
Human	π-PrimeNovo	4,647	0.49	0.49	0.22	0.22–0.26
Yeast	Casanovo	505	0.59	0.59	0.20	0.59
Yeast	InstaNovo	5,086	2.38	2.06	1.18	1.14–1.28
Yeast	π-PrimeNovo	6,235	2.18	2.02	1.52	1.35–1.57
Design anchors	Casanovo	444	16.89	16.89	12.16	4.50
Design anchors	InstaNovo	615	13.17	13.01	10.08	3.90–5.04
Design anchors	π-PrimeNovo	778	19.67	19.41	15.68	5.78–7.71
Anchors, local pool	Casanovo	146	14.38	14.38	11.64	2.74
Anchors, local pool	InstaNovo	562	12.63	12.63	9.79	3.56–4.80
Anchors, local pool	π-PrimeNovo	615	15.77	15.61	12.36	3.25–5.04
Supported anchors	Casanovo	313	7.03	6.71	3.51	1.60
Supported anchors	InstaNovo	268	2.61	2.61	1.87	0.75–1.12
Supported anchors	π-PrimeNovo	311	3.54	3.54	2.57	0.64–0.96

Table 22: Common-domain peptide-representative selections at nominal 5%. Rates are percentages; ranges retain unknown endpoints on the full accepted denominator. Synthetic identities use the composition model in Appendix C.

Population	Model	Accepted	String	Residue	Formula	Ion90
E. coli 30mM	Casanovo	1,925	1.71	1.19	0.52	0.94
E. coli 30mM	InstaNovo	2,009	4.63	3.38	0.65	0.80–0.95
E. coli 30mM	π-PrimeNovo	2,049	2.54	1.27	0.59	0.73–0.88
E. coli 60mM	Casanovo	3,537	2.37	2.18	0.71	0.99
E. coli 60mM	InstaNovo	3,747	4.83	4.16	0.53	0.91–0.93
E. coli 60mM	π-PrimeNovo	3,628	1.96	1.27	0.36	0.50–0.55
Human	Casanovo	7,247	5.71	5.48	0.83	1.26
Human	InstaNovo	7,425	5.29	4.88	0.66	0.89–1.00

Table 22 (continued)

Population	Model	Accepted	String	Residue	Formula	Ion90
Human	π-PrimeNovo	7,753	2.06	1.90	0.61	0.62–0.71
Yeast	Casanovo	6,644	3.67	3.48	2.11	2.32
Yeast	InstaNovo	7,591	6.63	5.78	1.90	2.13–2.37
Yeast	π-PrimeNovo	8,080	4.32	3.82	2.12	2.00–2.25
Design anchors	Casanovo	1,043	31.06	29.91	22.24	11.31
Design anchors	InstaNovo	923	24.38	23.19	17.88	6.83–8.45
Design anchors	π-PrimeNovo	1,042	28.60	27.93	21.79	9.79–12.67
Anchors, local pool	Casanovo	685	19.85	19.85	15.18	6.72
Anchors, local pool	InstaNovo	757	18.10	17.04	13.21	5.28–7.00
Anchors, local pool	π-PrimeNovo	862	22.27	22.04	17.87	7.42–9.51
Supported anchors	Casanovo	390	16.15	15.13	6.92	5.38
Supported anchors	InstaNovo	383	13.58	12.53	5.74	2.87–3.13
Supported anchors	π-PrimeNovo	362	7.18	6.91	3.04	1.10–1.93

F.3 Bounds for the full synthetic population

The full synthetic population includes spectra with both model-compatible and modified candidate identities. Its bounds describe assignment under the library and chemistry assumptions in Appendix C. Figure 9 compares the declared design, one-deamidation expansion and residue-composition interpretation. Table 23 reports the complete Winnow selection: non-singleton design sets contribute unknown reference assignments to its sequence-mass range, while candidate sets give design-based bounds on the same records. Figure 10 gives the corresponding 1% selections.

Table 23: **Winnow on the full synthetic population.** Native denotes sequence-mass matching. Its references are design singletons; the range assigns unknown comparisons to either outcome on the full accepted denominator. Design ranges use the candidate sets and composition assumptions in Appendix C. Expanded permits one deamidation and compares ordered residue compositions.

Nominal	Accepted	Native known	Native (%)	Design (%)	Expanded (%)
1%	0	0	empty	empty	empty
5%	44,926	5,115	1.50–90.12	26.65–90.20	25.50–90.69

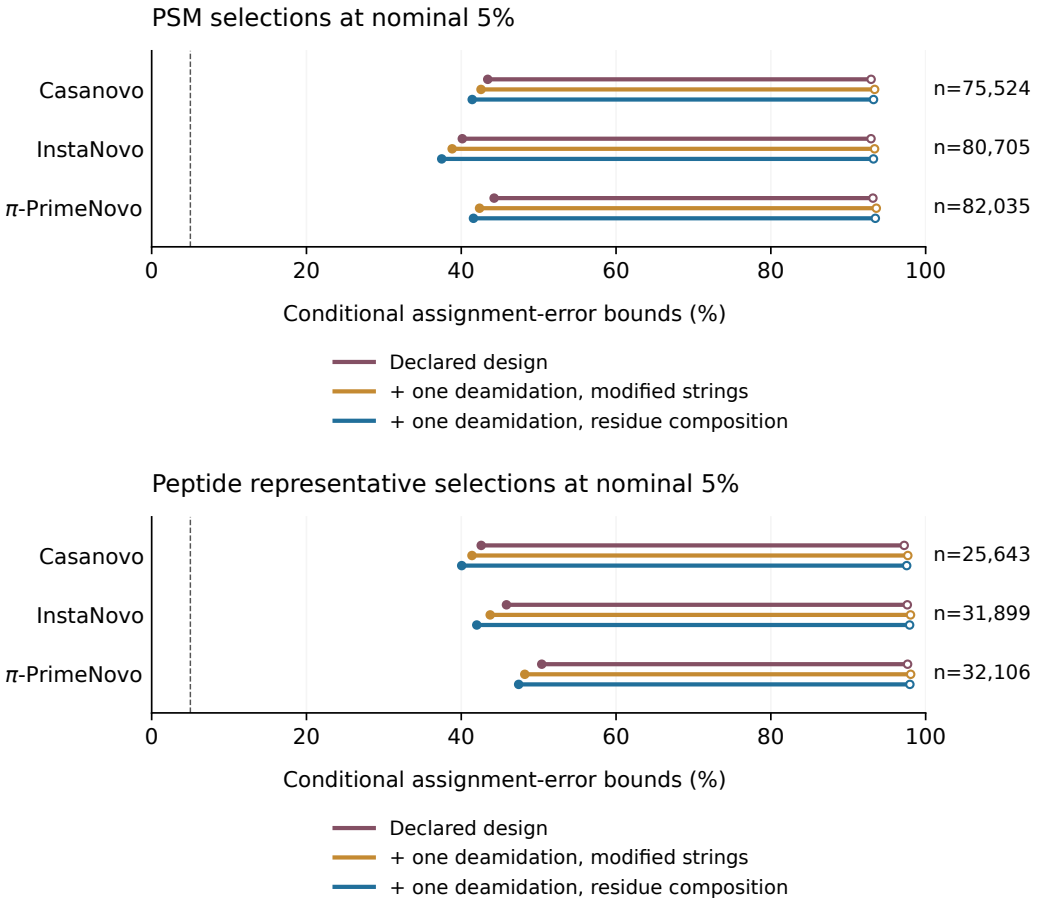

Figure 9: **Assignment bounds for the full synthetic population at 5%.** Top: PSMs; bottom: separately selected peptide representatives. Filled/open endpoints are lower/upper bounds under each composition model, evaluated on the same accepted records. These ranges express unresolved identity.

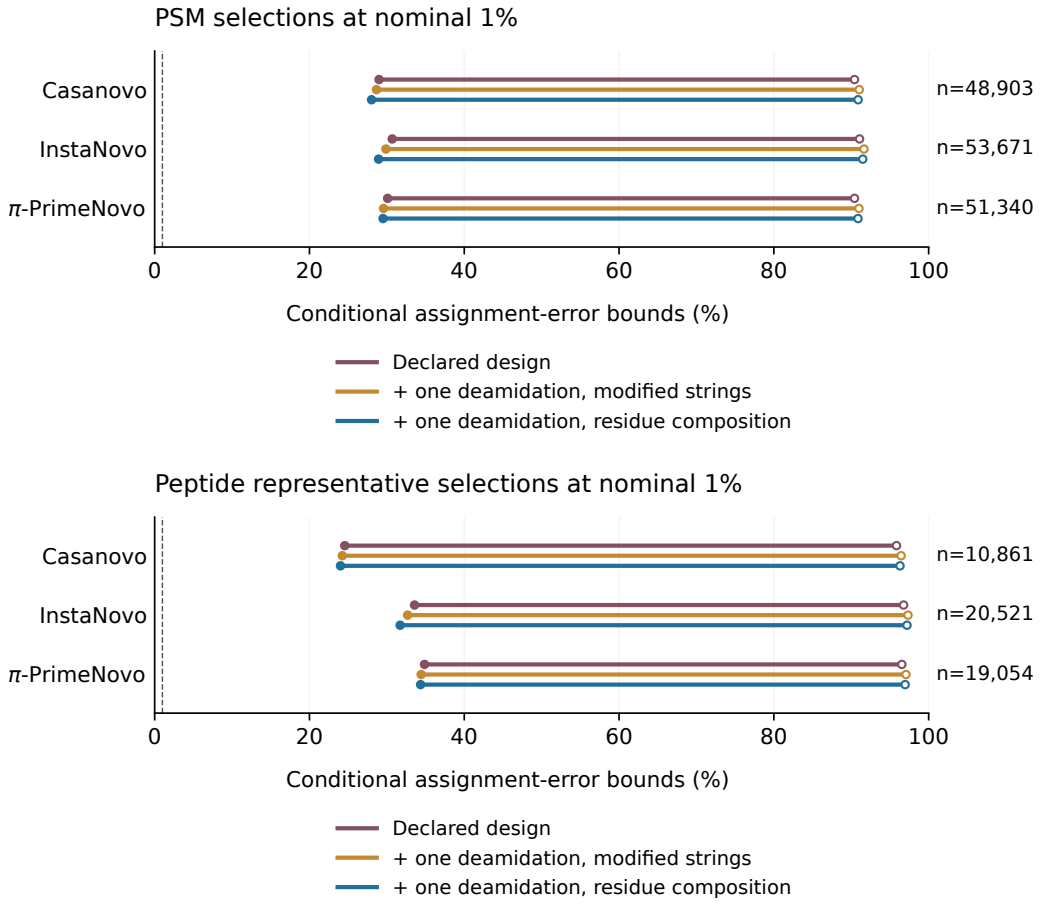

Figure 10: **Assignment bounds at 1%.** Thresholds are selected at nominal 1% using the same composition assumptions and discovery units as Figure 9.

F.4 Noise-pool sensitivity

Within-library noise pools yield lower assignment disagreement and smaller accepted sets for all three models on common-domain design anchors at nominal 5%. For Casanovo, acceptance changes from 6,428 to 5,759 PSMs and disagreement from 18.68% to 12.64%. Figure 11 shows both constructions and operating levels. The peptide panels compare precursor assignment with intended-library membership on identical representatives. PSM-to-peptide changes depend on the population and separate threshold selection: human Casanovo disagreement falls from 7.45% to 5.71%, whereas design-anchor disagreement rises from 18.68% to 31.06%.

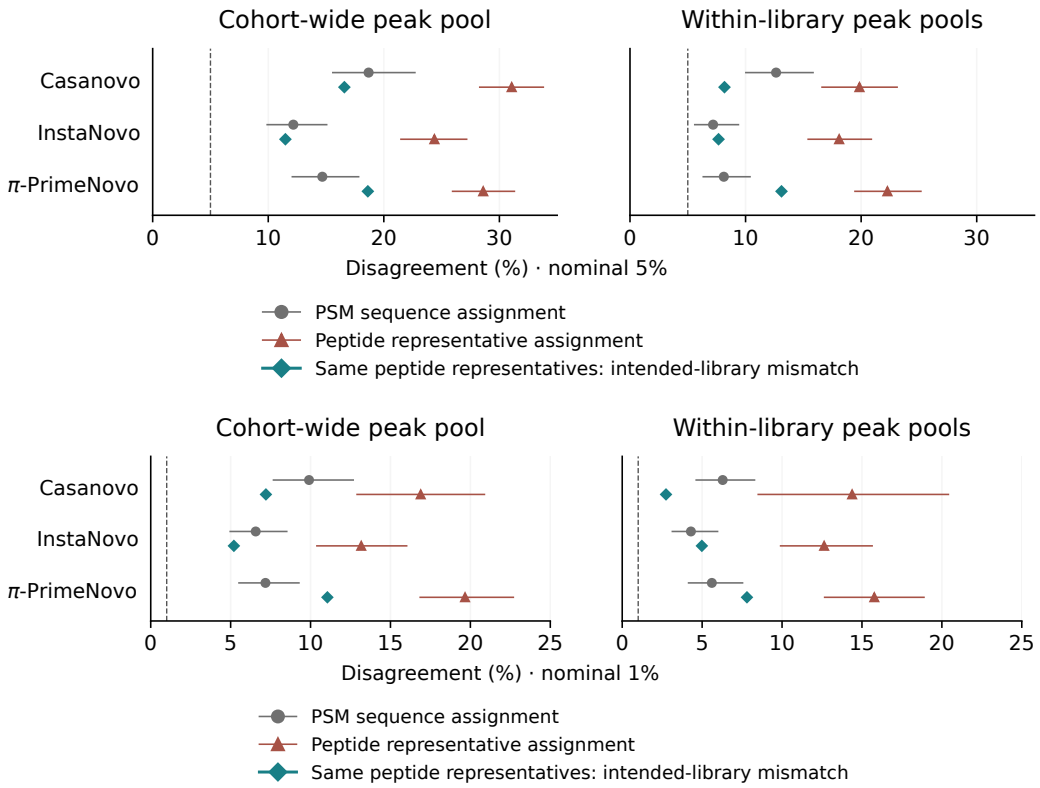

Figure 11: **Cohort-wide and within-library noise pools.** Top: nominal 5%; bottom: 1%. Target predictions, retained-peak fraction and other generator/model settings are identical. Gray points evaluate PSM assignments, red points peptide assignments and teal points library membership for those same peptide representatives. Each construction has its own selected threshold.

F.5 Sensitivity to the source partition

The human PSM contrast between supervision criteria persists across all five source fits. Sequence-supervised disagreement ranges from 3.20–4.27% for Casanovo, 4.08–5.46% for InstaNovo and 2.01–2.82% for π-PrimeNovo. Ion-supervised sequence disagreement ranges from 10.27–13.70%, 14.37–16.40% and 10.29–13.96%, respectively. Two sequence-supervised InstaNovo fits exceed nominal 5%, and peptide results vary further (Figure 12).

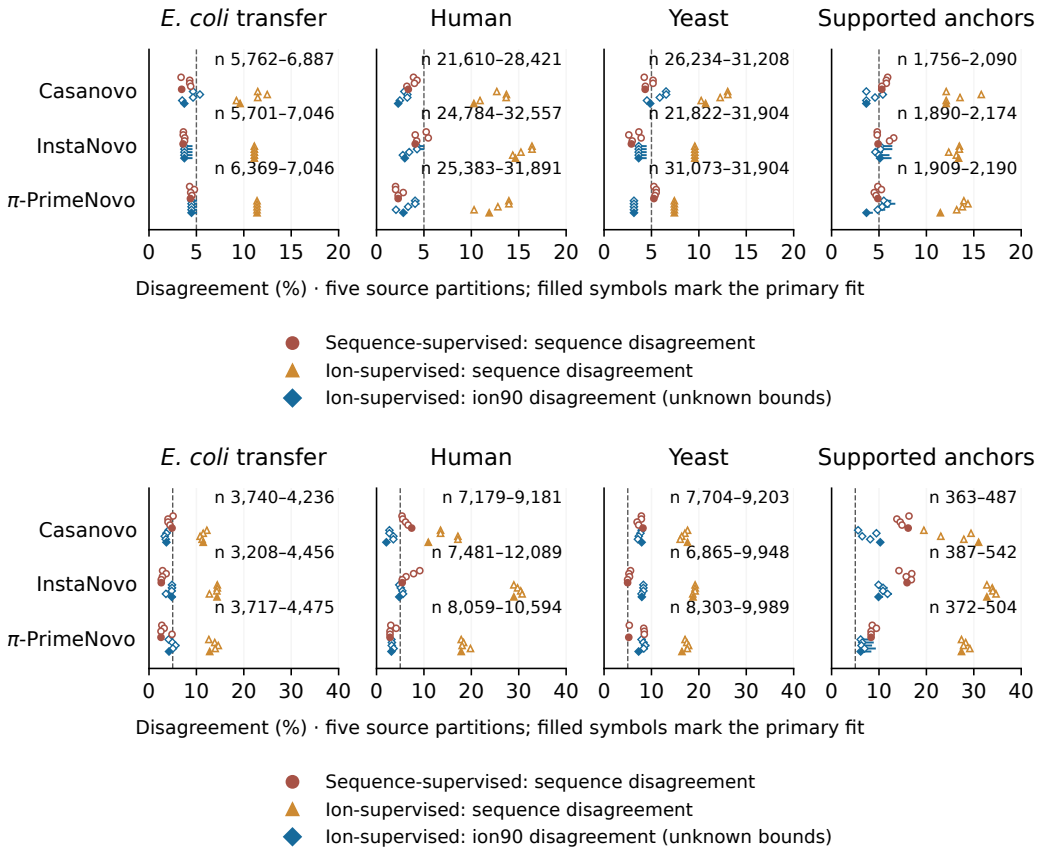

Figure 12: **Five source fits at nominal 5%.** Top: PSMs; bottom: peptide representatives. Each point is a source partition; filled markers identify the primary fit. Vertical offsets distinguish fits and supervision. Their spread describes fitting sensitivity. Blue ranges include unknown ion chemistry, and labels span accepted counts across both supervision criteria and all fits.