Accurate and ultra-fast *de novo* HLA-I immunopeptide sequencing with FoxNovo

Ze-Xuan Chen[1,2,3]; Chang-Rong You[1,2,3]; Ching Tarn[2,3]; Xie-Xuan Zhou[3]; Wen-Feng Zeng[2,3,*]

1. Zhejiang University, Hangzhou, China
2. Department of Artificial Intelligence, School of Engineering, Westlake University
3. Center for Infectious Disease Research, School of Medicine, Westlake University

* Email: zengwenfeng@westlake.edu.cn

We present FoxNovo, a hybrid deep learning-combinatorial framework for de novo sequencing of immunopeptides trained on a large-scale HLA-I immunopeptidomics dataset assembled and reprocessed from public mass spectrometry (MS) repositories. This integration achieved >90% peptide accuracy on the reported benchmarks while enabling repository-scale analysis at ~2,800 spectra per second—more than 100-fold faster than the evaluated beam-search baseline under the reported benchmark conditions. To mimic the heterogeneous spectral quality encountered in experimental MS analyses, we constructed controlled peak-removal stress tests, in which FoxNovo retained higher accuracy than the evaluated methods at different simulation levels. We subsequently re-analyzed 168 million spectra from all collected 4,423 MS raw files in only 18 hours on a single GPU, equivalent to ~245 raw files per hour. This repository-scale application yielded score-filtered canonical and putative ncORF-mapped peptide predictions and recovered 41 of 42 non-canonical HLA-I peptides previously validated by targeted MS. FoxNovo demonstrates the potential of integrating AI with combinatorial decoding for scalable immunopeptidomics. The source code is available at https://github.com/fennomix/fennomix.novo.

Introduction

Immunopeptidomics seeks to identify human leukocyte antigen (HLA)-bound peptides that drive antigen presentation and neoantigen discovery in immunotherapy, including peptides derived from non-canonical genomic sources[1–7]. However, many non-canonical peptides are absent from current reference protein databases, limiting the ability of conventional database-search approaches to identify those unexplored sequences. Therefore, de novo peptide sequencing, which directly infers peptide sequences from tandem mass spectra (MS/MS) without requiring reference databases, has emerged as a promising approach for immunopeptidomics analysis[8–10].

Recent advances in deep learning and the rapid accumulation of large-scale mass spectrometry (MS) datasets in public repositories such as PRIDE[11], iProX[12], and MassIVE[13] have accelerated the development of deep learning-based de novo peptide sequencing. Learning from large collections of annotated spectra, deep learning models has enabled direct translation of fragmented ion peaks into peptide sequences. Existing deep learning-based de novo sequencing methods generally fall into two categories. The first category consists of autoregressive models, such as DeepNovo[14], Casanovo[9], InstaNovo[15], and pUniFind[16], which predict amino acids sequentially based on previously predicted residues. The second category comprises non-autoregressive models, including PepNet[17] and π-PrimeNovo[18], which predict all amino acids simultaneously in a single decoding step. Although both strategies have improved peptide sequencing performance, their accuracy remains insufficient for many practical applications, particularly for immunopeptidomics.

Immunopeptidomics pose unique challenges for de novo sequencing, including the identification of non-canonical peptides and the analysis of highly diverse HLA-associated peptide repertoires. Although public MS repositories now contain large collections of immunopeptidomics datasets that enable the development of immunopeptide-specific AI models, few de novo sequencing models have been specifically designed and benchmarked for immunopeptidomics applications.

Beyond model specialization, a persistent limitation of deep learning-based de novo sequencing is the handling of precursor-mass information during decoding. Neural networks typically estimate amino acid probabilities and then generate one or more candidate sequences from those probabilities. Greedy or autoregressive decoding can consequently return sequences inconsistent with the measured precursor mass. Previous studies have introduced knapsack-based correction[14], diffusion-based refinement[15], and sequence-level decoding schemes[18] to improve mass consistency. However, the resulting performance depends on both the quality of the probability estimates and the extent of the candidate search. This motivates decoding strategies that efficiently explore multiple candidates and explicitly retain sequences satisfying the experimental precursor-mass constraint.

Before the advent of deep learning methods, peptide sequencing from MS spectra mainly relied on combinatorial algorithms. For example, pNovo[19,20] constructs a spectrum graph by connecting peaks whose mass differences correspond to one or two amino acids. The sequencing task is then formulated as finding the highest-scoring paths from zero mass to the precursor mass. Although combinatorial approaches generally achieve lower overall accuracy than modern deep learning models, they naturally encode physical

constraints such as precursor-mass consistency. We therefore hypothesized that a unified framework combining deep learning with combinatorial optimization could overcome this limitation by coupling the predictive power of neural networks with the rigorous constraint handling of exact optimization.

Beyond methodological advances, another largely overlooked challenge is how to evaluate deep learning-based de novo peptide sequencing models on real-world MS/MS data. Previous studies have primarily trained and benchmarked models by using confidently identified spectra that have passed stringent database-search filters, while spectra with reduced or heterogeneous signal quality remain largely unevaluated. In practical MS analyses, however, spectra span a broad range of quality levels. Given the high expressive capacity of modern AI models, predictions may still be generated from not high-quality spectra, yet the reliability of these predictions remains largely unexplored. Therefore, comprehensive evaluation under varying spectral quality is essential for assessing the practical applicability of deep learning-based de novo sequencing methods.

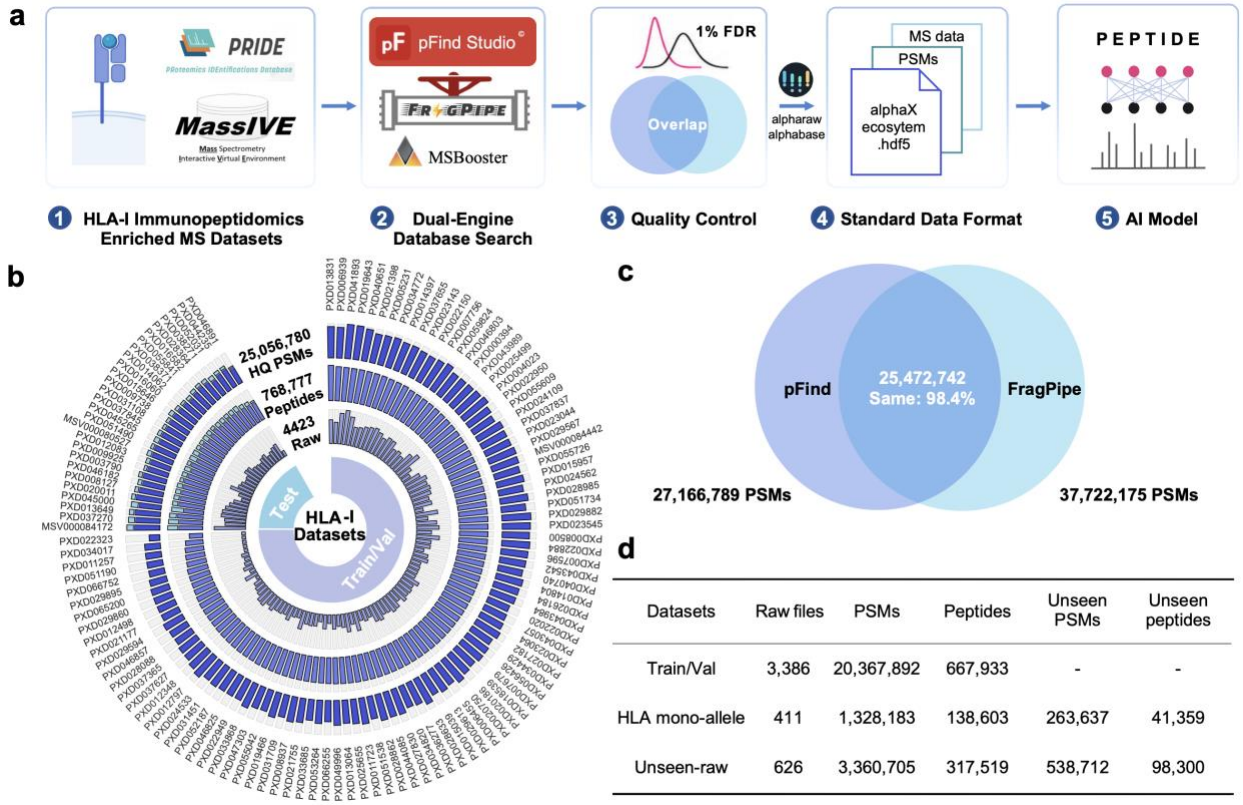
In this study, we collected 4,423 MS raw files from 124 immunopeptidomics datasets and constructed 25.1 million consensus peptide-spectrum matches (PSMs) identified concordantly by pFind[21] and FragPipe[22,23]. Based on these data, we developed FoxNovo, a non-autoregressive deep learning model featuring dual-token peak encoding for improved spectral representation and a dynamic programming-based decoding algorithm that enforces exact precursor-mass constraints. On the test datasets, FoxNovo achieved >90% peptide-level accuracy while processing ~2,800 spectra per second, more than 100-fold faster than the evaluated beam-search baseline under the reported benchmark conditions. To assess robustness to the heterogeneous spectral quality encountered in experimental MS analyses, we constructed controlled peak-removal stress tests, in which FoxNovo consistently outperformed the evaluated methods. Using FoxNovo, we identified putative non-canonical peptides supported by public ncORF annotations and recovered nearly all (41/42) targeted-MS-validated HLA-I neoantigen peptides from an independent melanoma proteogenomic dataset. FoxNovo provides a practical framework for large-scale immunopeptidomics de novo sequencing. The source code is available at: https://github.com/fennomix/fennomix.novo.

Results

Immunopeptidomics Dataset Preparation

To develop a de novo sequencing model tailored for HLA-I immunopeptidomics, we first collected large-scale immunopeptidomics datasets from PRIDE[11] and MassIVE[13], comprising 4,423 raw files, the detailed information is listed in **Supplementary Table 1**. Raw data were re-searched by using pFind[21] and FragPipe[22,23] at 1% false discovery rate (**Fig. 1a and 1b**). pFind and FragPipe identified 27.2M and 37.7M PSMs, respectively. Approximately 25.5M MS/MS spectra were commonly identified by both search engines, wherein 98.4% of these spectra shared the same peptide assignments (**Fig. 1c**). The resulting consensus dataset contained 25.1 million PSMs representing 769K unique peptide sequences. It was converted into an AlphaX-compatible HDF5 format[24,25], and used for subsequent AI model training and benchmarking.

The dataset was divided into training/validation and independent raw-file test sets. The training/validation set included 3,386 raw files, 20.4 million PSMs and 667,933 unique peptides (**Fig. 1b**). The mono-allelic test set ("mono-allele") comprised 411 raw files from MSV000084172[26] and MSV000080527[27], containing 1.3 million PSMs and 138,603 unique peptides; 41,359 of these peptides were absent from training (**Fig. 1d**). The unseen-raw set comprised 626 held-out raw files, 3.4 million PSMs and 317,519 unique peptides; 98,300 peptides were absent from training (**Fig. 1d**). Thus, the complete raw-file tests assess performance on new spectra that include both previously observed and previously unobserved peptide sequences. The unseen-peptide subsets provide the stricter evaluation of sequence-level generalization.



d

Datasets	Raw files	PSMs	Peptides	Unseen PSMs	Unseen peptides
Train/Val	3,386	20,367,892	667,933	-	-
HLA mono-allele	411	1,328,183	138,603	263,637	41,359
Unseen-raw	626	3,360,705	317,519	538,712	98,300

Figure 1. Large-scale HLA-I immunopeptidomics data. (a) Workflow for dataset curation, standardization, and storage for model training and evaluation. (b) Overview of the dataset composition and partitioning into training/validation and independent test sets. Dataset sizes are displayed on a logarithmic scale to enable visualization across datasets with different scales. (c) Overlap of PSMs identified by pFind and FragPipe, used to construct high-confidence training data for AI model development. (d) Summary of peptides and PSMs across dataset splits, including test subsets containing unseen peptides.

Architecture of FoxNovo

Leveraging this resource, we developed FoxNovo, a de novo peptide sequencing model that combines spectrum encoding with dynamic programming-based non-autoregressive decoding (**Fig. 2a, Methods**).

Conventional de novo sequencing models, such as Casanovo, typically use sinusoidal encoding of m/z values to represent MS/MS peaks. This representation aims to capture mass-difference information for de novo peptide sequencing. Similar to positional embeddings in large language models (LLMs), sinusoidal

m/z encoding provides continuous positional peak information; however, unlike LLMs, it lacks learnable embeddings for discrete peak tokens. A straightforward approach to discretizing MS/MS peaks is mass binning, such as using 0.001 m/z interval. However, for fragment ions extending to 3,000 m/z, coarse binning sacrifices the high mass accuracy enabled by modern MS instruments, whereas fine-grained binning preserves mass precision but results in an extremely large vocabulary space (~3 million bins at 0.001 m/z resolution). LLMs, like GPT series, employ subword tokenization strategies such as byte-pair encoding (BPE)[28], which decomposes a word into multiple reusable subword tokens, enabling a compact vocabulary while preserving semantic representations. Inspired by this idea, FoxNovo represents each peak using two separate tokens: an integer token and a decimal token. This dual-token representation avoids the need for embeddings of ~3 million fine-grained m/z bins by decomposing m/z values into compact integer (~3,000 tokens) and decimal (~1,000 tokens) vocabularies while preserving high mass resolution (**Fig. 2b**). FoxNovo combines dual-token representation with sinusoidal m/z encoding in the spectrum encoder, as shown in **Fig. 2a**.

Beyond spectral representation, effective model training requires addressing biases introduced by the highly uneven distribution of peptide observations. In the training dataset, we observed that the number of spectra per peptide followed a long-tail distribution (**Fig. 2c**), with a small subset of peptides accounting for a disproportionate fraction of spectra. Such spectral imbalance may bias model training toward frequently observed peptides and lead to overfitting. To address this issue, we introduced a peptide-charge-balanced weighted sampling strategy that assigns higher sampling weights to underrepresented peptide-charge pairs and lower weights to highly abundant peptide-charge pairs, thereby balancing the effective training distribution (**Fig. 2c, Methods and Supplementary Fig. 1**). Both the dual-token representation and peptide-charge-balanced sampling independently improved recall on the validation datasets, and their combination achieved the best performance (**Fig. 2d**). However, accurate spectrum encoding and balanced training alone do not guarantee globally optimal peptide identification, as the predicted amino acid probabilities still require decoding under strict precursor mass constraints.

To address this final challenge, FoxNovo uses a non-autoregressive decoder that predicts amino acid probabilities at all peptide positions in one forward pass (**Fig. 2a**). Greedy decoding (FoxNovo-Greedy) selects the highest-probability amino acid at each position and can produce a sequence inconsistent with the measured precursor mass. FoxNovo therefore applies dynamic programming-based top-K candidate generation to the non-autoregressive probability matrix and subsequently retains candidates whose theoretical masses agree with the experimental precursor mass (**Fig. 2e, Methods**). FoxNovo-DP1 reports the highest-ranked mass-consistent candidate, whereas FoxNovo-DP10 reports up to ten retained candidates.

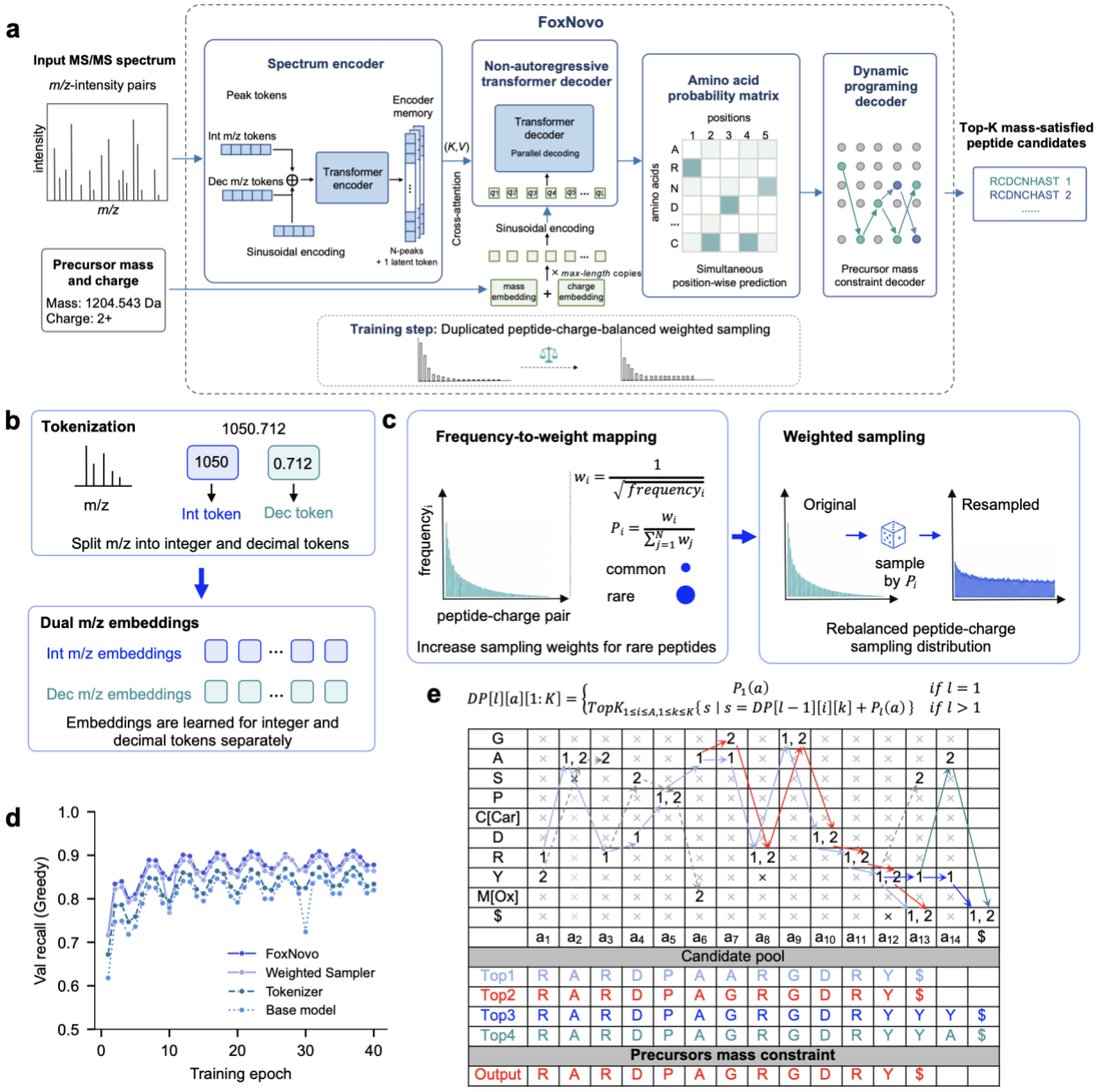


Figure 2. FoxNovo architecture: encoding, training and decoding. (a) Overview of the FoxNovo architecture. (b) Dual-token m/z representation used in the spectrum encoder. (c) Peptide-charge-weighted sampling strategy. (d) Ablation study of dual-token m/z representation and sampling using greedy decoding. (e) Dynamic programming-based top-K candidate generation followed by precursor-mass filtering.

Model Evaluation

To evaluate the performance of FoxNovo, we compared it with representative de novo peptide sequencing tools, including pNovo+[19], PEAKS[14,29], Casanovo-ne[9] (the Casanovo model trained for non-specific peptides), π-PrimeNovo[18] (PrimeNovo hereafter), InstaNovo+ (with diffusion model)[15], RNovA[30], and a Casanovo model we trained using the same training dataset in this study (denoted as Casanovo-HLA). To enable fair comparisons, beam search was applied in both InstaNovo+ and Casanovo-HLA (Casanovo-

HLA-Bm) to generate ten candidate peptide sequences. All models were evaluated on "unseen-raw" and "mono-allele" sets, both of which contained peptides absent from the training sets ("unseen-peptides") for a more stringent evaluation.

On the unseen-raw spectra, FoxNovo-Greedy achieved a mean top-1 sequence accuracy of 94.8% across raw files (**Fig. 3a; Methods**). FoxNovo-DP1 increased the mean accuracy to 96.0%, whereas FoxNovo-DP10 achieved a top-10 candidate recall of 97.2%. The latter represents an oracle candidate-recovery metric rather than final top-1 identification accuracy. InstaNovo+ achieved 92.6% top-1 accuracy, while Casanovo-HLA-Bm, Casanovo-HLA, PrimeNovo and RNovA achieved 89.8%, 86.1%, 84.4% and 82.1%, respectively, under the reported analysis settings. Similar trends were observed in the mono-allelic evaluation (**Fig. 3c**). Although PrimeNovo and RNovA were not specifically trained on HLA-I immunopeptidomics datasets, both retained appreciable performance, suggesting that models trained on large-scale general proteomics data can learn transferable relationships between peptide sequences and fragmentation spectra.

To further assess FoxNovo generalization to previously unobserved peptide sequences, we constructed an unseen-peptide benchmark from the unseen-raw and mono-allele test sets. **Fig. 3b** and **Fig. 3d** evaluated only spectra whose reference peptides were absent from the training set. In both scenarios, FoxNovo-Greedy achieved >90% median recall, which increased to >92% and >94% with FoxNovo-DP1 and FoxNovo-DP10, respectively. These results demonstrate that FoxNovo maintains high sequencing accuracy on peptides not encountered during training.

FoxNovo also showed high computational throughput under the reported benchmark conditions. Casanovo-HLA-Bm and InstaNovo+ processed 22.3 and 7.4 spectra/s, respectively, and PrimeNovo processed a median of 184 spectra/s on an RTX 4090 GPU. FoxNovo processed approximately 2,800 spectra/s on the same GPU and approximately 200 spectra/s on the evaluated CPU system (**Fig. 3e**). By reducing the computational cost of de novo peptide sequencing, FoxNovo enables practical large-scale immunopeptidomics analysis that was previously constrained by the inference speed of existing approaches.

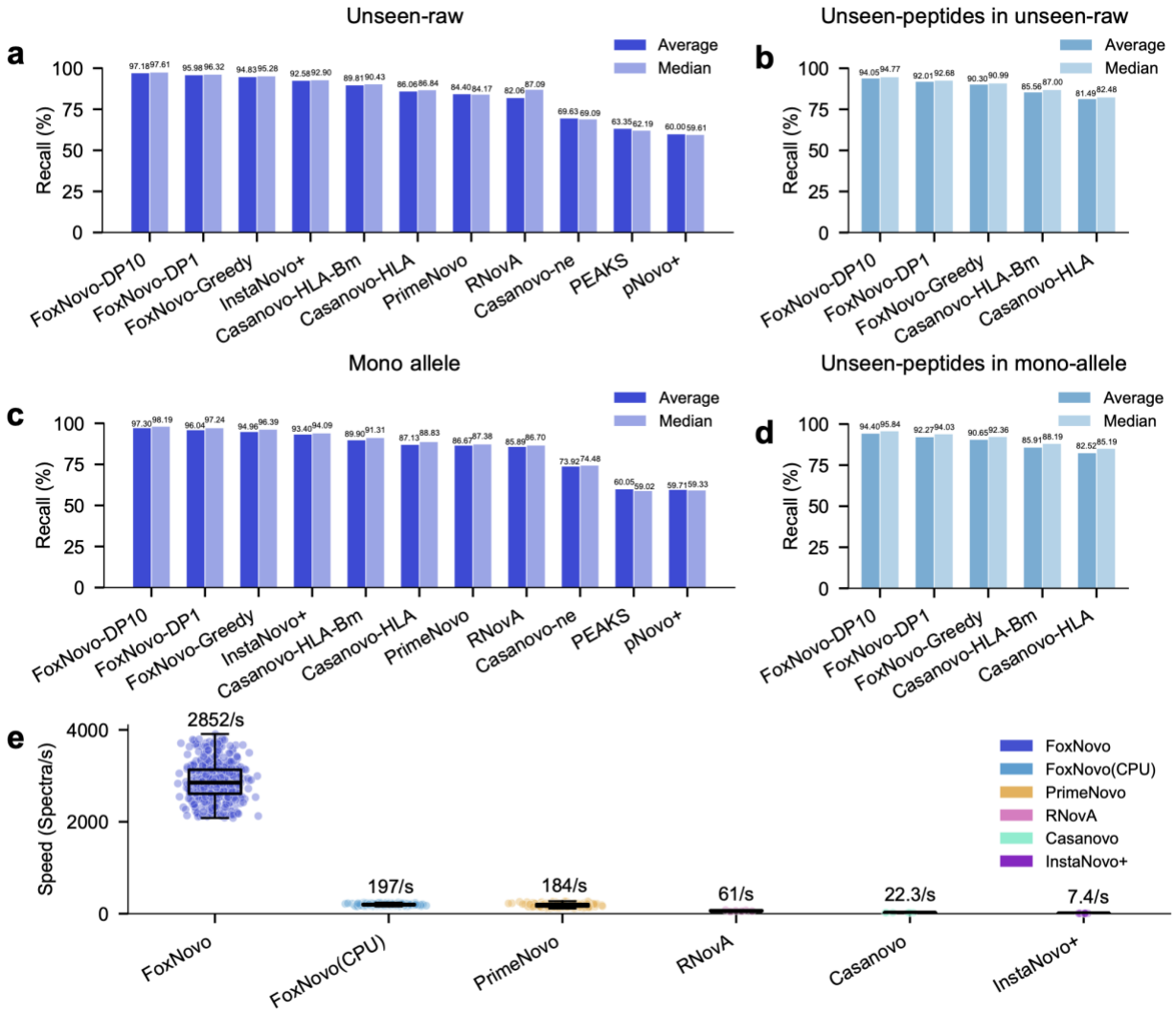


Figure 3. Performance benchmarking across different de novo sequencing tools. (a) Evaluation on the unseen-raw dataset. (b) Evaluation on unseen-peptide subsets in the unseen-raw dataset. (c) Evaluation on the mono-allele dataset. (d) Evaluation on unseen-peptide subsets in the mono-allele dataset. (e) Comparison of inference speed across different models. Beam search was enabled for InstaNovo+ and Casanovo; RNovA was evaluated using its default decoding mode. Box plots indicate the median (center line), interquartile range (box limits), and whiskers extending to 1.5× the interquartile range. Each dot represents an individual MS file.

Controlled Peak-Removal Stress Tests

Unlike database search approaches that employ false discovery rate (FDR)-based quality control to filter low-confidence identifications, de novo sequencing currently lacks comparable robust frameworks, leaving identifications derived from low- or medium-quality MS/MS spectra insufficiently controlled. Current AI models were predominantly trained on high-quality spectra, as only these spectra yield confident peptide annotations under stringent FDR filtering from database search. Consequently, spectra with reduced signal quality, which are common in real-world datasets, are largely absent from training and testing due to the lack of reliable ground-truth labels. Whether current de novo sequencing models can generalize from high-quality training data to such spectra remains unclear.

To measure sensitivity to missing spectral information, we established a controlled peak-removal stress benchmark using the unseen-raw test set. Starting from high-confidence reference PSMs, we removed 10-60% of low-abundance peaks and retained the original peptide annotations as reference labels (**Fig. 4a; Methods**). The perturbed spectra were mixed with an equal number of intact spectra and reanalyzed using pFind and the de novo sequencing tools. pFind retained high agreement with the original labels among accepted PSMs, while the number of recovered PSMs decreased as more peaks were removed (**Supplementary Fig. 2**). These results indicate that controlled peak removal reproduces an important property of low-information experimental spectra—reduced fragment-ion evidence and consequently reduced database-search sensitivity. Although this simulation does not capture every source of poor spectral quality, such as co-isolation, precursor-assignment errors or instrument-dependent background, it provides a reproducible framework for evaluating de novo sequencing under systematically reduced spectral information.

De novo model performance was evaluated on all test spectra ("Removal@All") and on the unseen-peptide test subsets from FoxNovo ("Removal@Unseen"). Removal@All denotes all degraded spectra at each peak-removal level, whereas Removal@Unseen denotes the subset of perturbed spectra whose peptide sequences were absent from the FoxNovo training set. On the Removal@Unseen tests, FoxNovo-DP1 showed remarkable robustness to spectral degradation, with only an 11.4% decrease in recall at 60% peak removal compared to original identifications. Even with a peak removal ratio at 50%, FoxNovo-DP1 still achieved 90.2% recall, outperformed competing methods by approximately 10 percentage points under comparable levels of spectral degradation. For example, as shown in **Fig. 4b**, InstaNovo+ recall progressively declined from 92% on original spectra to 88.6%, 86.1%, 82.4%, and 75.6% after 30%, 40%, 50%, and 60% peak removal, respectively. Similar degradation patterns were observed for Casanovo-HLA, PrimeNovo, and RNovA, each exhibiting approximately 20 percentage-point losses in recall at 60% peak removal. Notably, Casanovo-HLA-Bm achieved 74.1% recall, substantially higher than Casanovo-HLA (67.7%). The only difference between Casanovo-HLA and Casanovo-HLA-Bm is that the former greedily predicts a single locally optimal sequence for each spectrum, whereas the latter selects the highest-probability sequence from the top-10 beam candidates. This result further highlights the importance of considering multiple candidate sequences and motivates globally optimized decoding strategies for accurate peptide sequencing.

To further characterize prediction failures, we classified incorrect identifications into three categories (**Fig. 4c-e**). "Mass mismatch" refers to predictions whose peptide masses are inconsistent with the detected precursor mass. For predictions with matched precursor masses, "AA error ≤ 2" denotes sequences differing from the ground truth by at most two amino acids (edit distance ≤ 2; **Methods**), whereas "AA error > 2" denotes larger deviations (edit distance > 2). Casanovo-HLA exhibited the highest proportion of mass-mismatched identifications (**Fig. 4c**), likely because autoregressive decoding relies on learned amino-acid transition probabilities without explicitly enforcing precursor mass constraints. As a result, spectral degradation propagated into sequence generation, yielding peptides with incorrect masses. The use of top-10 beam search in Casanovo-HLA-Bm reduced mass-mismatched predictions by >10 percentage points, highlighting the benefit of exploring multiple candidate sequences during decoding (**Fig. 4c**). InstaNovo+ incorporates a knapsack-based heuristic decoding to enforce precursor mass consistency. However, this formulation does not guarantee global optimality, as amino-acid tokens are still selected in a largely greedy

manner, leading to suboptimal sequence assemblies with 8.4% AA error > 2 at 60% peak removal **(Fig. 4e)**. RNovA employs a graph-based path search strategy, ensuring that all candidate paths connect zero mass to the precursor mass, which effectively eliminates mass mismatches. Consequently, RNovA exhibits near-zero mass mismatches across conditions. However, strict precursor mass enforcement alone cannot ensure accurate peptide sequencing, as 26.9% of RNovA predictions still contained more than two amino acid errors at 60% peak removal (**Fig. 4e**). FoxNovo first identifies high-scoring candidate sequences from the predicted amino acid probability matrix and then filters them according to theoretical and experimental precursor mass agreement. As shown in **Fig. 4c-e**, FoxNovo-DP1 consistently obtains the lowest error rates across all categories, particularly for AA error > 2, which remains below 3%, indicating that globally optimal decoding closely matches ground-truth peptide sequences.

The unseen-peptide perturbation set provides a stringent evaluation of sequence-level generalization under progressively reduced spectral information. FoxNovo-DP1 achieved an average top-1 sequence accuracy of approximately 87% across all perturbation levels. Its accuracy was 87.5%, 84.1% and 78.3% at 40%, 50% and 60% peak removal, respectively. Under the most severe condition, Casanovo-HLA and Casanovo-HLA-Bm achieved accuracies of only 62.6% and 69.0%, respectively, compared with 78.3% for FoxNovo-DP1 (**Fig. 4f**). These results demonstrate that FoxNovo retains comparatively high sequencing accuracy for previously unseen peptide sequences even when a substantial proportion of spectral peaks is removed. To improve the reliability of predictions for practical applications, we evaluated a spectrum-matching score adapted from the fragment-ion evidence used by database search engines[31] (**Methods and Supplementary Fig. 3**). Applying the prespecified quality-control (QC) threshold of 30, together with precursor-mass filtering, increased the accuracy among retained FoxNovo predictions to 85% at 60% peak removal and to >90% when averaged across perturbation levels, with only a modest reduction in prediction coverage (**Fig. 4g**). Because this threshold is a descriptive score filter rather than formal FDR control, we report both accuracy among accepted predictions and the proportion of predictions retained. FoxNovo-DP10 achieved >90% top-10 candidate recalls even at 60% peak removal; this value represents candidate recovery and should not be interpreted as top-1 identification accuracy. Depending on the removal level, pFind assigned peptides different from the original reference annotations to 5-20% of the perturbed spectra (**Supplementary Fig. 2**). Among these spectra, QC-filtered FoxNovo predictions still achieved high sequence accuracy relative to the original reference peptides (**Supplementary Fig. 4**).

Similar trends were observed when removed peaks were replaced with intensity-matched background peaks (**Methods and Supplementary Fig. 5**). Collectively, these experiments show that FoxNovo is comparatively robust to the evaluated peak-removal-with-refill perturbations. Validation on naturally low-quality spectra with independently established peptide identities will be required to determine how these results transfer to unidentified experimental spectra.

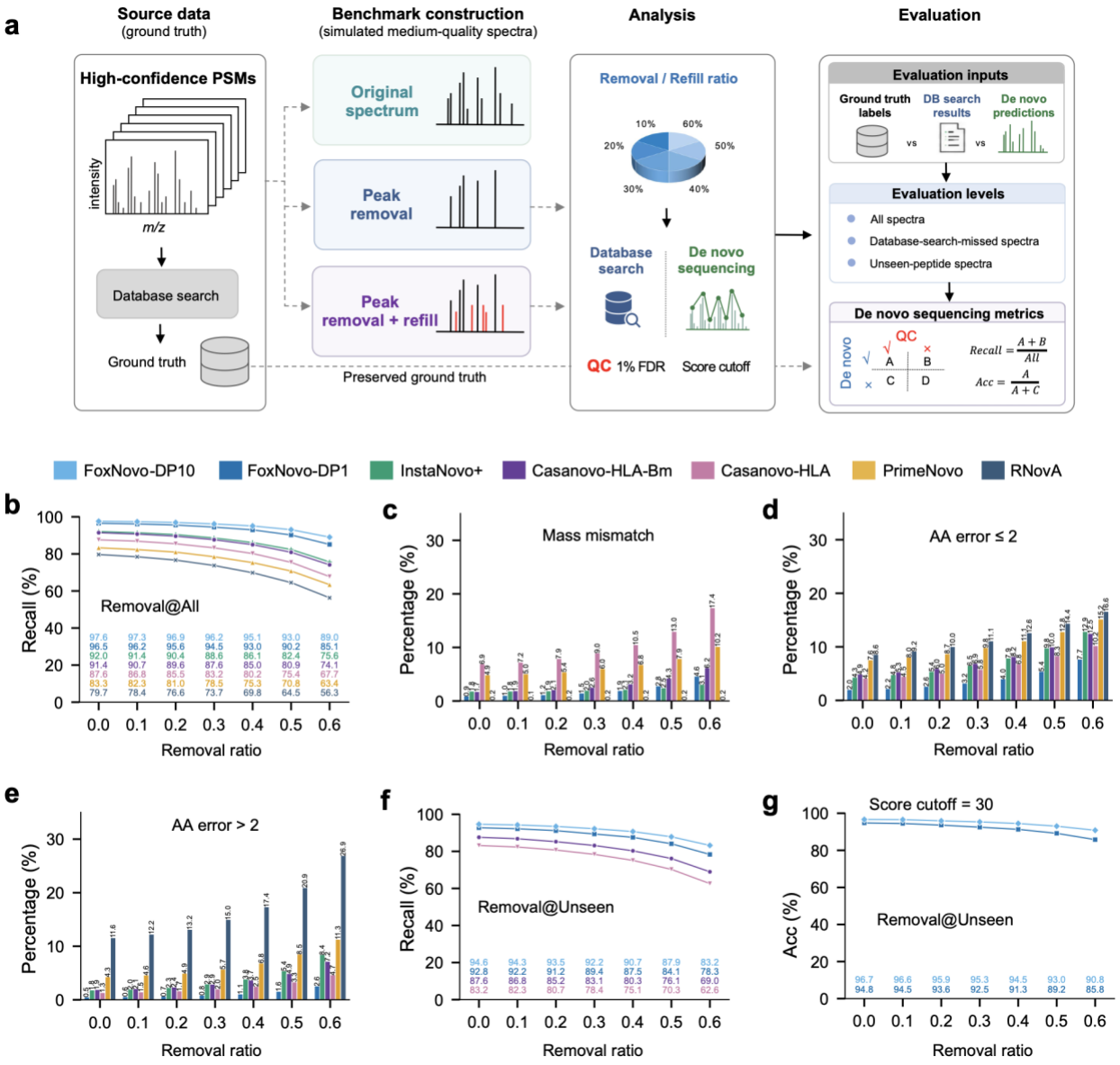


Figure 4. Controlled peak-removal stress tests. (a) Workflow for generating and analyzing peak-removed spectra. (b) Benchmarking on perturbed spectra from the unseen-raw set. (c–e) Distribution of incorrect predictions among mass mismatch, amino acid error ≤2 and amino acid error >2 categories. (f) Performance on the unseen-peptide subset. (g) Accuracy and retained prediction yield after applying the FoxNovo spectrum-matching score threshold.

Discovery of Non-canonical Peptides with FoxNovo

After benchmarking FoxNovo, we applied it to approximately 168 million spectra from 4,423 HLA-I raw files to prioritize canonical and putative non-canonical peptide candidates. The analysis required approximately 18 hours on one RTX 4090 GPU, equivalent to approximately 245 raw files per hour. Predictions were filtered using the spectrum-matching score threshold of 30 and the matched-ion criterion described in **Methods**. The resulting sequences were mapped to the reviewed human proteome and a public ncORF database[32–34] (**Fig. 5a**). The analysis produced 1,759,567 score-filtered peptide predictions, of which

728,580 mapped to reviewed proteins, 9,061 mapped only to ncORFs and 448 mapped to both databases. Predictions mapping to neither database were classified as other sequences (**Fig. 5b**).

Among the 9,061 FoxNovo-identified ncORF peptides, 2,647 showed overlap with previously reported non-canonical HLA-I immunopeptides from E. W. Deutsch et al. [35] (**Fig. 5c**). These ncORF peptides showed PSM scores comparable to canonical peptides mapped to reviewed proteins, showing that non-canonical identifications obtained similar fragment ion evidence (**Fig. 5d**). To further assess their potential HLA-I presentation, we used NetMHCpan 4.2[36] to predict peptide binding affinities across 138 common HLA class I alleles. For each peptide, the lowest predicted affinity rank among all alleles was selected as the representative binding score. Among 3,094 reference ncORF peptides from E. W. Deutsch et al., 2,674 (86.43%) achieved a best binding rank below 1%. Similarly, 5,549 of the 6,414 peptides newly identified by FoxNovo (86.51%) met the same threshold, with predicted HLA-I binding distributions comparable to those of the references (**Fig. 5e**). Lengths of the ncORF peptides displayed similar distribution to canonical HLA-I immunopeptides (**Fig. 5f**). The ncORF classification further showed that uORFs contributed the largest number of protein sources, consistent with previous report (**Fig. 5g**). Finally, peptide multiplicity analysis across ncORFs revealed substantial heterogeneous in translational activity among loci (**Fig. 5h**).

To further assess the ability of FoxNovo to recover experimentally validated non-canonical antigens, we re-analyzed the immunopeptidomic dataset from a previous proteogenomic study[5]. In that study, 42 lncRNA- or TE-derived non-canonical HLA-I peptides from the 0D5P melanoma sample were validated by targeted MS. FoxNovo recovered 41 of these 42 peptides, despite only 12 being represented in the ncORF reference database (**Fig. 5i and Supplementary Table 2**). This result indicates that even well-curated ncORF references may be insufficient to capture all non-canonical immunopeptides, highlighting the value of database-independent de novo sequencing and personalized referenc construction.

These analyses show that FoxNovo can prioritize canonical peptides and putative ncORF-mapped candidates supported by MS/MS evidence at repository scale. Individual ncORF-mapped candidates can subsequently undergo calibrated error-rate assessment and orthogonal validation to confirm their non-canonical origin.

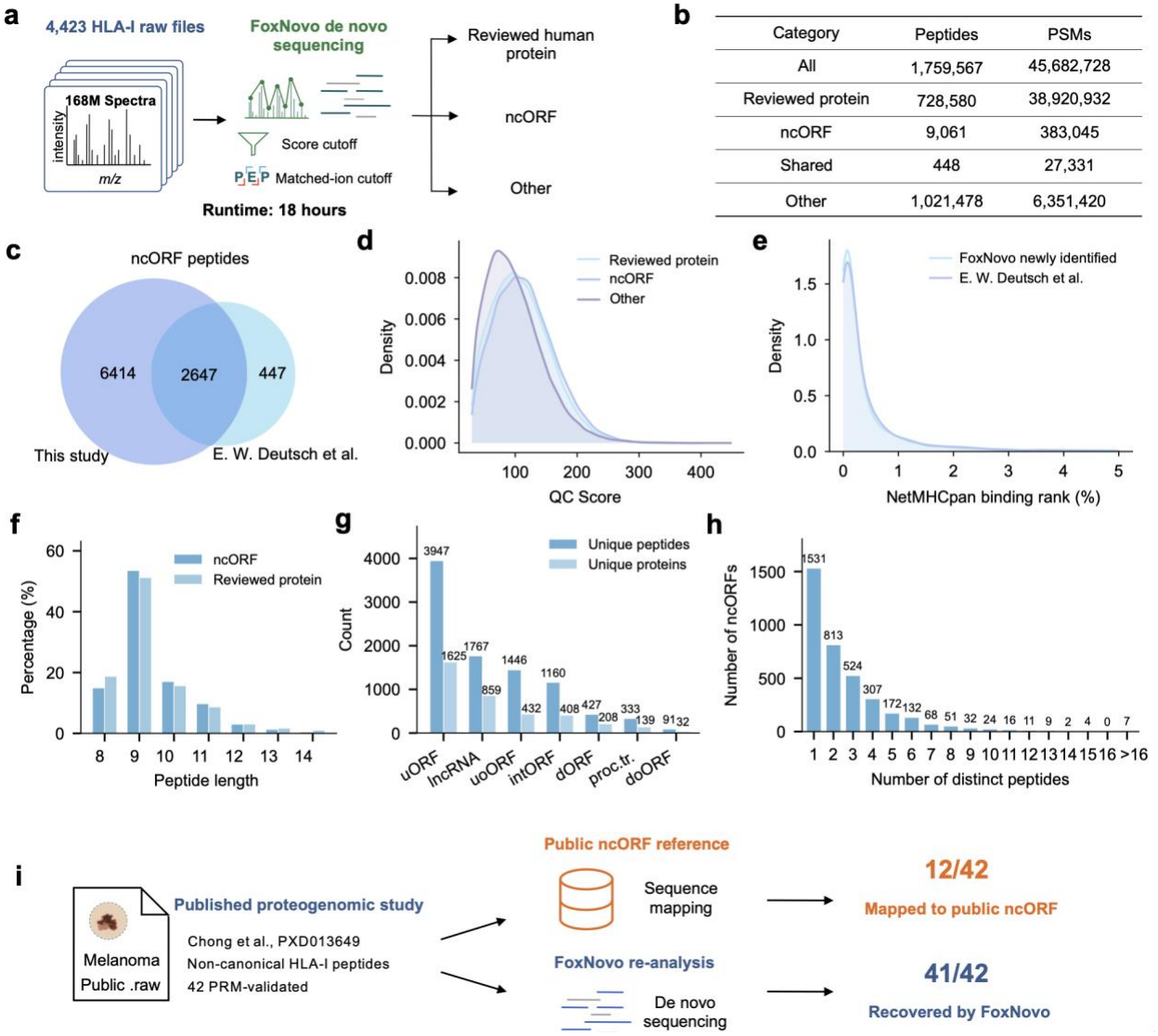


Figure 5. Large-scale de novo sequencing and mapping of HLA-I immunopeptides. (a) Pipeline for large-scale de novo sequencing and mapping of HLA-I immunopeptides. (b) Distribution of de novo sequenced peptides across reviewed proteins, ncORFs, and unannotated sequences. (c) Overlap of ncORF-only peptides with previously reported database-search-based identifications by E. W. Deutsch et al. (d) Distribution of QC scores across the three peptide categories. (e) Predicted HLA binding affinities of newly identified and previously reported ncORF-derived peptides, computed using NetMHCpan. (f) Length distribution of ncORF-mapped identified peptides. (g) Distribution of peptides across ncORF categories. (h) Number of distinct peptides detected per ncORF. (i) Validation of FoxNovo on externally reported non-canonical HLA-I peptides.

Discussion

The rapid accumulation of public immunopeptidomics datasets provides an unprecedented opportunity to develop AI models specifically optimized for immunopeptide analysis. By systematically collecting and reprocessing thousands of HLA-I immunopeptidomics datasets, we established a large-scale resource that can enhance existing approaches, such as Casanovo, for immunopeptide prediction, and facilitate the development of next-generation deep learning models.

Previous models have demonstrated a strong ability to estimate amino acid probabilities from large MS/MS datasets, but greedy decoding can return peptides inconsistent with precursor mass. FoxNovo combines non-autoregressive probability estimation with dynamic programming-based top-K candidate generation and precursor-mass filtering. This design improves mass consistency and supports prediction speed of approximately 2,800 spectra per second under the reported GPU benchmark conditions. In the current formulation, mass filtering is applied during each step of top-K candidate expansion; the results therefore establish efficient recovery of high-scoring mass-consistent candidates.

A major challenge in evaluating de novo sequencing on spectra with limited fragment-ion evidence—which are commonly encountered in MS datasets—is the absence of reliable reference labels, as many of these spectra cannot be assigned by conventional database searching. To address this benchmark gap, we developed a controlled peak-removal framework that progressively removes low-abundant peaks in confidently identified spectra while preserving their original peptide annotations as reference labels. This strategy simulates spectra with different degrees of information loss, enabling reproducible and quantitative evaluation of de novo sequencing under defined degradation conditions. We evaluated performance on both complete held-out raw-file sets and stricter unseen-peptide subsets to distinguish overall performance from sequence-level generalization. Across these simulated conditions, FoxNovo retained higher sequence accuracy than the evaluated baselines, demonstrating its robustness to progressive loss of spectral evidence. Although peak removal does not reproduce every source of poor experimental spectral quality, including co-isolation, precursor-assignment errors, altered fragmentation and instrument-dependent noise, it provides a novel and practical benchmark for evaluating commonly encountered spectra with incomplete fragment-ion evidence while retaining known peptide identities. Future studies using synthetic peptides, targeted MS or matched replicate acquisitions could complement this framework and further establish performance on naturally lower-quality spectra.

In the era of AI-enabled immunopeptidomics, computational scalability is essential for extracting biological knowledge from rapidly expanding public MS repositories. FoxNovo analyzed approximately 168 million spectra from 4,423 raw files in only 18 hours on a single consumer GPU, corresponding to approximately 245 raw files per hour. This throughput makes repository-scale de novo immunopeptidomics practically feasible and enabled comprehensive mapping of score-filtered peptide predictions to the reviewed human proteome and annotated ncORFs. Several observations supported the biological relevance of the putative ncORF-derived peptides, including their overlap with previously reported non-canonical immunopeptides, fragment-ion evidence comparable to that of canonical peptides, characteristic HLA-I peptide-length and binding patterns, and recovery of 41 of 42 non-canonical HLA-I peptides previously validated by targeted MS. Together, these findings demonstrate that scalable de novo sequencing can extend immunopeptide discovery beyond conventional reference proteomes and provide a focused set of non-canonical candidates for subsequent confirmation through independent error calibration, sample-specific HLA information, and orthogonal spectral or translational evidence.

One inherent limitation of MS-based de novo peptide sequencing is the inability to distinguish leucine (Leu) from isoleucine (Ile) with standard collision-based fragmentation. This ambiguity may be partially resolved by integrating matched transcriptomic data, which enables candidate peptide mapping against translated protein sequences and subsequent prioritization using HLA-binding prediction. If more electron-transfer dissociation (ETD) HLA immunopeptidomics datasets, such as those reported by Amy et al.[37], become

available for model training, the Leu/Ile ambiguity may be further reduced, as ETD provides complementary fragmentation patterns that encode residue-level differences between these two isomeric amino acids. Beyond residue discrimination, ETD-derived immunopeptides display distinct sequence motifs compared with collision-based HLA immunopeptidomics datasets, introducing orthogonal sequence-level information that could enable AI models to better learn the underlying grammar of HLA-presented peptides and improve de novo sequencing performance.

Additional limitations include the identification of modifications, potential study-level domain shifts, and the absence of a calibrated de novo false-discovery rate. The current score threshold should be interpreted as a ranking and filtering rule. Future work should calibrate posterior error probabilities on study-held-out data, validate them on an untouched external benchmark and report precision together with prediction coverage.

References

1. Chong, C., Coukos, G. & Bassani-Sternberg, M. Identification of tumor antigens with immunopeptidomics. *Nat. Biotechnol.* **40**, 175–188 (2022).

2. Apavaloaei, A. *et al.* Tumor antigens preferentially derive from unmutated genomic sequences in melanoma and non-small cell lung cancer. *Nat. Cancer* **6**, 1419–1437 (2025).

3. Huber, F. *et al.* A comprehensive proteogenomic pipeline for neoantigen discovery to advance personalized cancer immunotherapy. *Nat. Biotechnol.* **43**, 1360–1372 (2025).

4. Laumont, C. M. *et al.* Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, eaau5516 (2018).

5. Chong, C. *et al.* Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat. Commun.* **11**, 1293 (2020).

6. Tsour, S. *et al.* Alternate RNA decoding results in stable and abundant proteins in mammals. *Nature* (2026).

7. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science.* **348**, 69–74 (2015).

8. Liao, H. *et al.* MARS an improved de novo peptide candidate selection method for non-canonical antigen target discovery in cancer. *Nat. Commun.* **15**, 661 (2024).

9. Yilmaz, M. *et al.* Sequence-to-sequence translation from mass spectra to peptides with a transformer model. *Nat. Commun.* **15**, 6427 (2024).

10. shah, D. *et al.* Proteogenomics and de novo sequencing based approach for neoantigen discovery from the immunopeptidomes of patient CRC liver metastases using Mass Spectrometry. *J. Immunol.* **204**, 217.16-217.16 (2020).

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

12. Ma, J. *et al.* iProX: an integrated proteome resource. *Nucleic Acids Res.* **47**, D1211–D1217 (2019).

13. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).

14. Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proc. Natl. Acad. Sci. USA.* **114**, 8247–8252 (2017).

15. Eloff, K. *et al.* InstaNovo enables diffusion-powered de novo peptide sequencing in large-scale proteomics experiments. *Nat. Mach. Intell.* **7**, 565–579 (2025).

16. Zhao, J. *et al.* A large-scale unified deep learning model for peptide mass spectrum interpretation trained on multimodal data. *Nat. Mach. Intell.* **8**, 997–1011 (2026).

17. Liu, K., Ye, Y., Li, S. & Tang, H. Accurate de novo peptide sequencing using fully convolutional neural networks. *Nat. Commun.* **14**, 7974 (2023).

18. Zhang, X. *et al.* π-PrimeNovo: an accurate and efficient non-autoregressive deep learning model for de novo peptide sequencing. *Nat. Commun.* **16**, 267 (2025).

19. Chi, H. *et al.* pNovo+: De Novo Peptide Sequencing Using Complementary HCD and ETD Tandem Mass Spectra. *J. Proteome Res.* **12**, 615–625 (2013).

20. Chi, H. *et al.* pNovo: *De novo* Peptide Sequencing and Identification Using HCD Spectra. *J. Proteome Res.* **9**, 2713–2724 (2010).

21. Chi, H. *et al.* Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat. Biotechnol.* **36**, 1059–1061 (2018).

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

23. Yang, K. L. *et al.* MSBooster: improving peptide identification rates using deep learning-based features. *Nat. Commun.* **14**, 4539 (2023).

24. Wallmann, G. *et al.* AlphaDIA enables DIA transfer learning for feature-free proteomics. *Nat. Biotechnol.* **44**, 1168–1177 (2026).

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

26. Sarkizova, S. *et al.* A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* **38**, 199–209 (2020).

27. Abelin, J. G. *et al.* Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity* **46**, 315–326 (2017).

28. Wang, C., Cho, K. & Gu, J. Neural Machine Translation with Byte-Level Subwords. *Proc. AAAI Conf. Artif. Intell.* **34**, 9154–9160 (2020).

29. Zhang, J. *et al.* PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. *Mol. Cell. Proteomics* **11**, M111.010587 (2012).

30. Mao, Z. *et al.* Zero-shot de novo peptide sequencing with open posttranslational modification discovery. *Nat. Biotechnol.* (2026).

31. Liu, M.-Q. *et al.* pGlyco 2.0 enables precision N-glycoproteomics with comprehensive quality control and one-step mass spectrometry for intact glycopeptide identification. *Nat. Commun.* **8**, 438 (2017).

32. Bateman, A. *et al.* UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617 (2025).

33. Desiere, F. The PeptideAtlas project. *Nucleic Acids Res.* **34**, D655–D658 (2006).

34. Mudge, J. M. *et al.* Standardized annotation of translated open reading frames. *Nat. Biotechnol.* **40**, 994–999 (2022).

35. Deutsch, E. W. *et al.* Expanding the human proteome with microproteins and peptideins. *Nature* **654**, 813–825 (2026).

36. Nilsson, J. B., Greenbaum, J., Peters, B. & Nielsen, M. NetMHCpan-4.2: improved prediction of CD8+ epitopes by use of transfer learning and structural features. *Front. Immunol.* **16**, 1616113 (2025).

37. Kessler, A. L. *et al.* Increased EThcD Efficiency on the Hybrid Orbitrap Excedion Pro Mass Analyzer Extends the Depth in Identification and Sequence Coverage of HLA Class I Immunopeptidomes. *Mol. Cell. Proteomics* **24**, 101049 (2025).

Methods

Dataset collection and processing

All HLA-I-related raw files were downloaded from the PRIDE[11] and MassIVE[13] repositories. To select high-quality HLA-I-enriched immunopeptide data, we first performed database searching using pFind[21] with peptide length constraints of 8-45 amino acids. Raw files were retained only if they contained more than 1,000 PSMs, with 9-mer peptides accounting for >30% and 8-14-mer peptides accounting for >60% of total identifications. This step was used to filter out incorrectly annotated HLA raw files. To reduce search-engine-specific biases and improve peptide-level confidence, we further retained only consensus peptide-spectrum matches (PSMs) identified by both pFind[21] and FragPipe[22,23]. Specifically, only PSMs assigned to the same peptide sequence by both search engines were included in the final HLA-I dataset construction.

Detailed database search parameters: protein FASTA = human reviewed proteome (UP000005640_9606) downloaded from UniProt[32] on 22 April 2025 with 20,644 protein entries; MS1 matching mass tolerance = ±20 ppm; MS2 matching mass tolerance = ±20 ppm; enzyme = nonspecific; FDR = 1% at spectrum level. Other engine-specific parameters:

pFind Studio (v3.2.2): open-search mode = on; variable modifications = Oxidation[M]; peptide length = 8-45.

FragPipe (v22.0): MSFragger = v4.1; variable modifications = Oxidation (M, +15.9949 Da), Carbamidomethyl (C, +57.0214 Da), and Cysteinyl (C, +119.004 Da); peptide length = 8-14; MSBooster (v1.2.31) = enabled.

Other search parameters were set as default.

By using the above workflow, a total of 4,423 raw files were classified as HLA-I immunopeptidomics data. All the spectrum data and PSMs were saved as HDF5 files and processed using AlphaBase (v1.8.0) and AlphaRaw (v0.5.0)[24,25] Python packages. A PyTorch DataLoader was also designed based on HDF5 files as the data source for model training and evaluation.

FoxNovo model design

Tokenization of peak m/z

FoxNovo decomposes each peak m/z value m_i into an integer component token and a decimal component token:

$$q_i = \lfloor m_i \rfloor \ ,$$
$$r_i = \min(\lfloor 1000(m_i - q_i) + 0.5 \rfloor, 999),$$

where q_i represents the integer part of the m/z value and r_i represents the decimal part at 0.001 m/z resolution. Integer tokens encode integer m/z values from 0 to 3,000, whereas decimal tokens are discretized

into 1,000 bins from 0 to 999 at 0.001 m/z resolution. Thus, each peak m/z value is represented using an integer-decimal token pair, with approximately 3,000 integer-level tokens and 1,000 decimal tokens.

The integer and decimal tokens are independently mapped to learnable embeddings using torch.nn.Embedding. Both embedding layers project their input tokens into d_{model}-dimensional vectors, where d_{model}=512 in this work. These embedding vectors are then integrated with sinusoidal m/z positional embeddings and peak intensity representations.

Transformer encoder

Given an input MS/MS spectrum S, FoxNovo encodes the spectrum into a latent memory representation:

$$M = Encoder(S),$$

where $M \in \mathbb{R}^{B\times(N+1)\times d_{model}}$, B is the batch size, N is the number of retained peaks per spectrum, and d_{model} is the embedding dimension. In this work, $N = 150$ by default after spectrum preprocessing. Each spectrum is represented as a peak list $S \in \mathbb{R}^{B\times N\times 2}$, where each peak contains an m/z value and a 0–1 normalized intensity value. Before encoding, precursor-related peaks are removed, and spectra are truncated to the top $N = 150$ peaks, sorted by m/z, square root transformed and L2-normalized.

For each peak, the m/z value is encoded using a sinusoidal encoder, and the peak intensity is projected using a learnable linear layer, similar to the representations used in Casanovo. The two components are summed to obtain a continuous peak representation in $\mathbb{R}^{d_{model}}$. In parallel, the m/z value is further represented using the tokenization scheme described above. Tokenized embedding vectors are then integrated with sinusoidal m/z positional embeddings and peak intensity representations. The combined representations are projected from $4 \times d_{model}$ to d_{model} through a linear layer and yield a fused peak representation. A learnable latent spectrum token is prepended to the sequence of fused peak embeddings to capture global spectrum-level information. The resulting sequence of length $N + 1$ is processed by a 9-layer Transformer encoder with hidden dimension 512, 8 attention heads and feed-forward dimension of 1,024. The encoder of FoxNovo contains 26,067,137 trainable parameters. Padding masks are applied to prevent zero-padded peaks from contributing to self-attention. Under the default settings, the encoder output is:

$$M \in \mathbb{R}^{B\times 151\times 512}.$$

Non-autoregressive transformer decoder

FoxNovo adopts a 9-layer non-autoregressive (NAR) transformer decoder for parallel peptide sequence prediction, with hidden dimension 512, 8 attention heads and feed-forward dimension of 1,024. The FoxNovo decoder contains 28,402,713 trainable parameters. Precursor mass and charge information are incorporated into the decoder through a precursor-conditioned target representation. Specifically, precursor mass m_{pre} is encoded using the same sinusoidal float encoder used for continuous mass-related features, while precursor charge z is represented using a learnable charge embedding E_z:

$$c = \varphi(m_{\mathrm{pre}}) + \mathrm{E}_z(z).$$

The resulting precursor representation c is repeated across all decoding positions and combined with positional encodings to form the decoder target input:

$$H = \text{TransformerDecoder}(PE([c] \times L), M),$$

where M denotes the spectrum encoder output and L represents the maximum peptide length plus an additional position for the end-of-sequence (EOS) token. The output hidden states $H = [h_1, h_2, \ldots, h_L]$ are projected and normalized with a log-softmax layer to obtain amino acid log-probability distributions:

$$P(y_l = a \mid S, M_{\text{pre}}, Z) = LogSoftmax(W_{out} h_l)_a$$

where W_{out} denotes the output feed-forward layer.

As a result, the final decoder outputs a log-probability matrix P over 23 sequence tokens, denoted as A:

$$P = [P_1(A), P_2(A), \ldots, P_L(A)] \in \mathbb{R}^{L \times |A|}.$$

In this version, isoleucine and leucine are treated as equivalent by converting I to L. The decoder output vocabulary therefore contains 23 sequence tokens: 19 canonical amino acid residue symbols after I/L normalization, three modified residue tokens (M+Oxidation, C+Carbamidomethyl and C+Cysteinyl), and one EOS token.

Dynamic programming-based top-K peptide decoding and precursor-mass filtering

Given the NAR amino-acid probability matrix, the score of a peptide sequence $a_1 a_2 \ldots a_l$ is defined as the sum of position-wise log-probabilities:

$$S(a_1, \ldots, a_l) = \sum_{i=1}^{l} P_i(a_i),$$

where $P_i(a_i)$ denotes the predicted log-probability of amino acid a_i at position i.

To generate high-scoring peptide candidates, we used a dynamic programming-based top-K procedure. Let the decoder provide position-wise log-probabilities over the residue vocabulary and EOS token. For every permitted peptide length, the procedure retains the K highest-scoring candidate paths according to the sum of position-wise log-probabilities. States ending in EOS are not extended further. The implementation used K = 10 unless otherwise specified. Let $DP[l][a][k]$ denote the score of the k-th highest-scoring partial sequence of length l ending with amino acid a. Because every length-l sequence can be constructed by extending a length-$(l-1)$ sequence, the DP recurrence is defined as:

$$DP[l][a][1:K] = \begin{cases} P_1(a) & if\ l = 1 \\ TopK_{1 \le i \le A, 1 \le k \le K}\{ s \mid s = DP[l-1][i][k] + P_l(a) \} & if\ l > 1' \end{cases}$$

where A denotes the output vocabulary size, including residue tokens and the EOS token, and $TopK$ returns the K highest-scoring candidates. This recurrence guarantees that the optimal top-K sequences are

retained at every peptide length. To ensure valid peptide termination, states ending with the EOS token are excluded from subsequent state transitions.

Precursor-mass consistency was applied after candidate generation. For each candidate ending in EOS at an allowed peptide length, FoxNovo calculated the theoretical neutral peptide mass and retained the candidate only when it agreed with the precursor-derived neutral mass within ±20 ppm. The calculation included the residue and modification masses in the decoder vocabulary and the mass of water. Candidate scores were normalized by peptide length before final ranking.

Because summed log-probability scores are length-dependent, sequence scores are normalized by peptide length.

$$Score[l][a][k] = \frac{DP[l][a][k]}{l}.$$

After precursor-mass filtering, **FoxNovo-DP1** reports the highest-ranked retained candidate, whereas **FoxNovo-DP10** retains up to ten mass-consistent candidates. **FoxNovo-Greedy** reports one sequence without precursor-mass filtering.

Model training

Among all the 4,423 collected HLA-I raw files, 3,386 files were used for training and validation, and 1,037 were used for testing. The training and validation data were split at the unique modified-sequence level using a 9:1 ratio, rather than at the raw-file level, and each peptide was allowed to have at most 500 PSMs. All PSMs assigned to the same unique modified peptide sequence were kept in the same subset to prevent the same modified peptide label from appearing in both training and validation sets.

FoxNovo is trained as a non-autoregressive peptide sequence prediction model using a position-wise cross-entropy loss. The padding index is ignored in the loss function, and label smoothing is set to 0.01 during training. The model uses a transformer encoder-decoder architecture with a hidden dimension of 512, 8 attention heads and a feed-forward dimension of 1,024. The encoder consists of 9 transformer layers, and the non-autoregressive decoder also consists of 9 transformer layers. Dropout is set to 0.1. The final FoxNovo model contains 54,469,850 trainable parameters. The peptide length range is set to 8–14.

Training was performed using the AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-5}. The learning rate followed a warmup schedule with cosine decay. The batch size was 64 for training and 1,024 for evaluation, and the maximum number of training epochs was 40. The random seed was set to 454. Model checkpoints were saved after each epoch, and the checkpoint with the highest validation recall was selected for downstream evaluation. FoxNovo was trained on a single RTX 4090 GPU for ~12 days.

To mitigate the long-tailed distribution of peptide observations, training samples are drawn using the peptide-charge-pair weighted sampling strategy. The calculated sampling weights are passed to a weighted random sampler during training.

For sequence tokenization, isoleucine and leucine are treated as equivalent by converting I to L. An EOS token is appended to each peptide sequence, and shorter sequences are padded to the maximum target length. The maximum peptide length is set to 14, with one additional position used for the EOS token.

Peptide-charge-balanced weighted sampling

To mitigate the long-tailed distribution of detected peptides in the HLA-I dataset, we introduce a peptide-charge-based weighted sampling strategy during model training. For each training sample i, the modified peptide sequence and precursor charge are combined to define a peptide-charge pair. The frequency of the i-th peptide-charge pair in the training set is denoted as $frequency_i$. Each sample is then assigned a sampling weight inversely proportional to the square-root of its peptide-charge-pair frequency:

$$w_i = \frac{1}{\sqrt{frequency_i}},$$

and the probability of sampling sample i is defined as

$$Prob(i) = \frac{w_i}{\sum_{j=1}^{N} w_j},$$

where N is the total number of training pairs.

Spectrum-matching score for quality control of peak matches

FoxNovo calculates a spectrum-matching score for each decoded peptide sequence as an interpretable confidence-control metric based on fragment ion evidence.

For each decoded peptide candidate, theoretical fragment ions, including singly and doubly charged b and y ions, are generated according to the predicted amino acid sequence. The theoretical fragment ions are then matched against the observed MS/MS peaks within a predefined fragment mass tolerance. Following the peptide backbone scoring scheme used in pGlyco2[31], the spectrum-matching score is calculated as:

$$Score_{Seq} = \sum \log(inten_i)\left(1 - \left|\frac{merr_i}{tol_i}\right|^4\right)(ratio_{ion})^{\gamma},$$

where $inten_i$ is the intensity of the experimental peak matched to the i-th theoretical fragment ion, $merr_i$ is the matching mass error of the i-th matched peak, and tol_i is the corresponding fragment mass tolerance. $ratio_{ion}$ denotes the fraction of matched theoretical fragment ions. The parameter γ controls the contribution of fragment ion coverage and is fixed at 0.94, following the peptide backbone scoring parameter reported in pGlyco2.

Model evaluation

Among all collected raw files, 1,037 were used for independent testing, including 626 unseen-raw files and 411 raw files from HLA-I mono-allelic cell lines (mono-allele).

For all models, predicted sequences and reference labels were normalized using the same sequence-processing rules before evaluation. Isoleucine and leucine were treated as equivalent by converting I to L. A prediction is counted as correct only when the normalized predicted modified sequence exactly matches the normalized reference sequence.

For each raw file, model performance was evaluated using recall as the primary accuracy metric. Recall is defined as the fraction of spectra for which the predicted peptide sequence matches the reference peptide label:

$$Recall = \frac{N_{correct}}{N_{total}},$$

where $N_{correct}$ is the number of spectra with correctly predicted peptide sequences and N_{total} is the total number of evaluated spectra. Recall is calculated for each raw file, and both the mean and median recalls across raw files are reported to summarize model performance.

To evaluate FoxNovo performance, we tested three decoding strategies: FoxNovo-Greedy, FoxNovo-DP1, and FoxNovo-DP10. For comparison, InstaNovo+ was evaluated with top-10 beam search with diffusion module, and the diffusion-corrected sequence was used as the final prediction. PrimeNovo was evaluated with its precursor-mass correction module. Casanovo was retrained on our HLA-I training set, and two decoding strategies were evaluated: greedy decoding (**Casanovo-HLA**) and beam-search decoding, in which the highest-probability peptide among the top-10 beam candidates was selected as the final prediction (**Casanovo-HLA-Bm**). To ensure consistent prediction settings with FoxNovo, the "decoded max length" parameter for Casanovo was set to 14 for faster prediction speed. For RNovA, a subset of raw files exhibited abnormally low recall values, likely due to analysis failures. Therefore, raw files with recall below 50% were excluded from summary statistics. For PEAKS, we used the DeepNovo workflow with a precursor mass tolerance of 20 ppm and a fragment ion tolerance of 0.02 Da. Oxidation (M), Carbamidomethyl (C), and Cysteinyl (C) were specified as variable modifications, while all other parameters were kept at default settings. For pNovo+, Oxidation (M), Carbamidomethyl (C), and Cysteinyl (C) were specified as variable modifications and other parameters were kept at default settings.

To further evaluate generalization to unseen peptide sequences, we constructed a stricter unseen-peptide benchmark from the held-out raw files. Only spectra with reference peptide sequences not observed in the training set were retained for evaluation.

The inference speed of FoxNovo was evaluated by processing all spectra from each of the 626 unseen-raw files. Wall-clock time included spectrum loading, preprocessing, model inference, dynamic programming-based decoding, and output generation. Throughput was measured on both GPU and CPU with top-10 candidate decoding. CPU-based evaluations were performed on an AMD EPYC 9554 64-core processor. For comparison, PrimeNovo was evaluated on the full spectrum set with the precursor-mass correction module enabled. InstaNovo+ was evaluated using top-10 beam search with diffusion correction enabled, and Casanovo-ne was evaluated using greedy decoding. Inference speed for Casanovo-HLA, InstaNovo+, and RNovA was benchmarked on six raw files. HDF5-formatted MS files were used for FoxNovo testing,

while MGF-formatted MS files were used for testing the other models. GPU-based evaluations were performed on an NVIDIA RTX 4090 platform running Ubuntu 22.04.

Throughput was defined as the number of spectra processed per second. For each model, the distribution of per-file inference throughput across all evaluated MS files is presented with box plots.

Construction and evaluation of spectral degradation

To construct controlled spectral degradation benchmarks from the unseen-raw test set, high-confidence PSMs with known peptide annotations were used as source spectra. The original peptide annotations were retained as ground-truth labels throughout the benchmark construction.

Before constructing the controlled spectral degradation benchmarks, all spectra were first preprocessed to ensure that the peaks entering the encoder were effectively perturbed according to the designed degradation ratios. Specifically, multiple de novo sequencing models, including FoxNovo, apply spectrum preprocessing before encoder input, including m/z range filtering (50–2500), precursor peak removal, low-intensity peak filtering (1%), and retention of the top N = 150 peaks. Therefore, all MS/MS spectra were first processed using this preprocessing procedure before the subsequent low-abundance peak removal and background refill operations described below. This ensured that the designed fraction of peaks was actually removed or refilled from the final peak set provided to the models. This preprocessing strategy may have a minor influence on the performance of other models because different methods may use different spectrum preprocessing settings, such as the number of retained peaks.

For each benchmark construction, the selected spectra were split into two equal groups. One group was retained without modification and used as the intact control, corresponding to a degradation ratio of 0%. The other group was divided into six equally sized subsets, and peak-removal ratios of 10%, 20%, 30%, 40%, 50% and 60% were applied to the six subsets, respectively. In the peak-removal setting, for each spectrum, peaks were sorted by intensity, and the lowest 10%, 20%, 30%, 40%, 50% or 60% of peaks were removed according to the assigned degradation ratio.

In the peak-removal-with-refill setting, after removing the specified fraction of low-abundance peaks, we refilled the spectra with sampled peaks whose intensities were matched to the intensity distribution of the removed low-abundance peaks. This procedure preserved the original peptide label while introducing additional background interference, thereby simulating degraded spectra with both missing informative fragments and increased noise peaks.

After spectral degradation, the perturbed spectra were combined with the intact control spectra and written into reconstructed MGF files following the original file organization. The reconstructed MGF files were then reanalyzed by pFind database searching. Peptide annotations assigned before perturbation were retained as ground-truth labels, and the pFind reanalysis results were used to determine whether each degraded spectrum could still be matched to its original peptide annotation by database search.

We evaluated de novo sequencing performance at multiple levels. The “all-spectra” setting included all spectra in the constructed benchmark. “Unseen-peptide” spectra were defined as spectra whose ground-

truth peptide sequences are absent from the training set. All spectra were divided into two groups according to whether their original peptide annotations were recovered by database-search reanalysis. FoxNovo was also evaluated on the subset of perturbed spectra whose original peptide annotations were not recovered by database-search reanalysis.

Top-1 sequence accuracy was used as the primary metric. Incorrect predictions were divided into mass-mismatch errors and mass-consistent errors. Mass mismatch was defined as disagreement between theoretical and measured precursor mass beyond ±20 ppm. Mass-consistent errors were stratified by Levenshtein distance into amino acid error ≤2 and amino acid error >2. For descriptive score filtering, FoxNovo predictions were filtered using the spectrum-matching score described above with a fixed cutoff of 30. Accuracy was calculated among predictions passing the threshold, and retained prediction yield was calculated relative to all evaluated spectra:

$$Acc = \frac{N_{correct,score>cutoff}}{N_{score>cutoff}},$$

$$Loss\ rate = 1 - \frac{N_{correct,score>cutoff}}{N_{original}},$$

where $N_{score>cutoff}$ denotes the number of predictions passing the score cutoff, $N_{correct,score>cutoff}$ is the number of correct predictions among them. $N_{original}$ is the number of spectra before score cutoff.

We note that RNovA did not show abnormally low recall value or analysis failure on the spectral degradation benchmarks, potentially due to the standardized MS/MS spectrum preprocessing procedure applied before dataset construction. Therefore, all MS files were retained for downstream analysis.

De novo sequencing of HLA-I datasets and mapping

All raw files in the HLA-I dataset were analyzed using FoxNovo with top-10 dynamic programming decoding enabled. For each PSM, the highest-ranked mass-constrained candidate sequence, namely the DP1 output, was used as the unique peptide prediction for every spectrum. Predicted PSMs were first filtered using a FoxNovo spectrum-matching score cutoff of 30.

For peptide-level filtering, PSMs were collapsed by predicted peptide sequence, and the highest-scoring PSM for each unique peptide was used as the representative spectrum-level evidence. A peptide was retained only if this representative PSM had no more than one missing matched-ion position; otherwise, the peptide was discarded. Here, a missing matched-ion position refers to a peptide backbone cleavage position for which no corresponding fragment ion of any considered type was matched in the observed MS/MS spectrum.

The ncORF reference database and ncORF classification annotations were obtained from PeptideAtlas[33,34], comprising 7,264 protein or microprotein sequences. Filtered FoxNovo peptide sequences were subsequently mapped to the reviewed human proteome and the PeptideAtlas ncORF database for peptide-origin annotation. Before mapping, modification annotations were removed from peptide sequences, and I was converted to L in both peptide and reference sequences. Each peptide was independently searched

against the reviewed proteome and the ncORF database by exact substring matching. Peptides were classified according to their mapping status against the reviewed proteome and ncORF database into four categories: reviewed-protein, ncORF, shared, and other sequences. For ncORF peptides, the corresponding ncORF entries and PeptideAtlas ncORF classification annotations were retained for downstream characterization of ncORF categories and peptide-to-ncORF assignments. HLA-I binding prediction was performed using NetMHCpan 4.2 across 138 common HLA class I alleles. For each peptide, the lowest predicted percentile rank across all tested alleles was used as the representative binding score.

Competing models

Casanovo (v4.3.0) and its pretrained non-specific model were downloaded on 4 April 2025 from https://github.com/Noble-Lab/casanovo. π-PrimeNovo was downloaded on 2 February 2026 from https://github.com/PHOENIXcenter/pi-PrimeNovo. InstaNovo (v1.2.2) was downloaded on 28 February 2026 from https://github.com/instadeepai/InstaNovo. RNovA was downloaded on 20 May 2026 from https://github.com/zqq66/RNovA. PEAKS Studio 12.5 was tested on a server at the mass spectrometry facility at Westlake University. pNovo+ was downloaded from the pFind official website https://pfind.ict.ac.cn/se/pnovo/.

Data availability

All mass spectrometry (MS) datasets utilized in this study were downloaded from the PRIDE and MassIVE repositories. Accession identifiers and corresponding raw files numbers for all collected datasets are provided in Supplementary Table 1. The ncORF database was retrieved from the PeptideAtlas (https://peptideatlas.org).

Code availability

The complete source code and pre-trained models of FoxNovo are publicly accessible on GitHub at https://github.com/fennomix/fennomix.novo.

Acknowledgements

We acknowledge the support from the start-up package provided by Westlake University for this research.

Author contributions

Conceptualization, W.-F. Z.; Methodology, W.-F. Z and Z.-X. C.; Investigation, Z.-X. C.; Data curation and formal analysis, Z.-X. C.; Software, Z.-X. C., W.-F. Z., C.-R. Y. and C.T.; Writing – original draft, W.-F. Z. and Z.-X. C; Writing – review & editing X.-X. Z. Supervision, W.-F. Z.

Competing interests

The authors declare no competing interests.