Journal of proteome research

pubs.acs.org/jpr

Technical Note

NovoRank: Refinement for *De Novo* Peptide Sequencing Based on Spectral Clustering and Deep Learning

Published as part of Journal of Proteome Research special issue "Software Tools and Resources 2025".

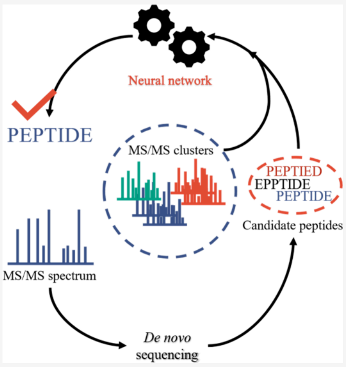
Jangho Seo, Seunghyuk Choi,* and Eunok Paek*

Cite This: https://doi.org/10.1021/acs.jproteome.4c00300

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: *De novo* peptide sequencing is a valuable technique in mass-spectrometry-based proteomics, as it deduces peptide sequences directly from tandem mass spectra without relying on sequence databases. This database-independent method, however, relies solely on imperfect scoring functions that often lead to erroneous peptide identifications. To boost correct identification, we present NovoRank, a postprocessing tool that employs spectral clustering and machine learning to assign more plausible peptide sequences to spectra. Prior to *de novo* peptide sequencing, spectral clustering is applied to group similar spectra under the assumption that they originated from the same peptide species. NovoRank then employs a deep learning model, incorporating both cluster-derived proteomic features and individual spectrum characteristics, to rerank the candidate peptides produced by *de novo* peptide sequencing. Our results show that NovoRank significantly enhances the performance of various *de novo* peptide sequencing tools, increasing both recall and precision by 0.020 to 0.080 at the peptide-spectrum match (PSM) level. Notably, NovoRank achieves a recall as high as 0.830 for Casanovo at the PSM level. The source code of NovoRank is freely available at https://github.com/HanyangBISLab/NovoRank and is licensed under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International.



KEYWORDS: *bioinformatics, proteomics, peptide identification, de novo peptide sequencing, spectral clustering, deep learning*

INTRODUCTION

Mass spectrometry (MS)-based proteomics has been widely used to identify protein expression and analyze their cellular functions. The cornerstone of MS-based proteomics is identifying peptides from tandem mass spectrometry (MS/MS) spectra. Typically, MS/MS spectra are searched against a protein sequence database, using software tools such as SEQUEST,[1] MaxQuant,[2] MS-GF+[3] and Comet.[4] This database search approach, however, is limited in its capacity to uncover novel peptides absent from existing databases. On the other hand, *de novo* peptide sequencing, unrestricted by predefined sequences, holds the promise of discovering new peptides. Various software tools such as PEAKS,[5] pNovo3[6] and Casanovo[7] have been developed; however, these are prone to inaccuracies due to noise and incomplete data (e.g., missing backbone fragment ion peaks) because they depends solely on MS/MS spectra to infer sequences.[8] For example, both N- and C-terminal fragment ions (i.e., b1 and y1 ions) tend to be missing in MS/MS spectra, leading to an ambiguous candidate sequence order near the peptide termini. When the peptide sequencing fails by such ambiguities, the correct identification can often be rescued by reordering ranks. For instance, pNovo3 trained SVM-rank based on the spectral similarity between experimental and predicted spectra and database statistics and reranked the top 10 candidate peptides for each spectrum obtained from pNovo.[6] It also utilized "spectrum merging" based on their precursor m/z and sequence tags to refine the top-ranking peptide. Compared to pNovo,[9] pNovo3 improved the recall of top-ranking peptide up to ~2 times, demonstrating the importance of reranking candidate peptides. However, it only used pNovo as a built-in peptide sequencing tool, making it difficult to apply the reranking approach for other tools such as PEAKS and Casanovo.

To enhance the accuracy of *de novo* peptide sequencing, we present NovoRank, a versatile postprocessing tool that can improve the precision and recall of *de novo* peptide sequencing results by reassigning the best peptide among possible candidates. NovoRank first clusters MS/MS spectra based on spectral similarity, precursor mass-to-charge ratio (m/z), and retention time (RT). On the premise that MS/MS spectra belonging to the same cluster originated from the same peptide, two candidate peptides with the highest cluster-score (C-score) are selected for each cluster. Then, it selects the

Received: April 11, 2024
Revised: November 22, 2024
Accepted: December 19, 2024

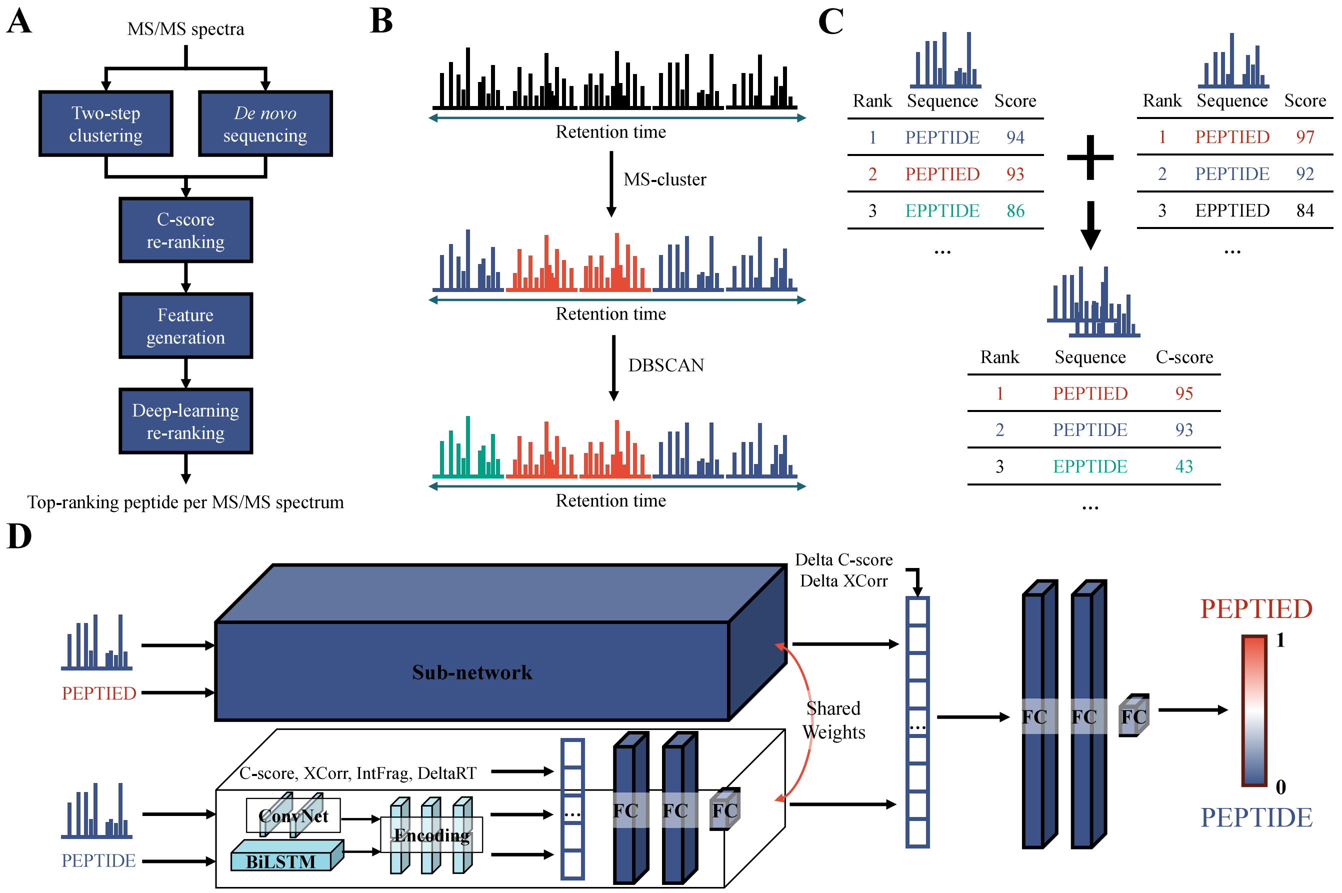


Figure 1. Workflow of NovoRank. (A) Analysis flow of NovoRank. (B) Two-step clustering method. Colors represent a cluster in each step. (C) Example cluster-score (C-score) calculation in a cluster of two MS/MS spectra. (D) Deep learning model to predict a more desirable peptide among the top two candidate peptides in a cluster. ConvNet, BiLSTM and FC stand for convolutional neural network, bidirectional long short-term memory, and fully connected layer.

more desirable of the two peptides by taking the learning-to-rank approach.[10] NovoRank's utility is demonstrated by its ability to improve both peptide precision and recall by an average of 0.046 and 0.045, respectively, when tested with results from various *de novo* peptide sequencing tools, thus confirming its adaptability and broad application potential in proteomics.

■ EXPERIMENTAL SECTION

Experimental Data Sets

Four data sets of MS/MS spectra were downloaded from the ProteomeXchange Consortium[11] via the PRIDE partner repository with the data set identifiers PXD004732, PXD014222, PXD002395 and PXD001468.[12−15] Briefly, the first data set, PXD004732, was acquired from the Orbitrap Fusion ETD with the HCD activation mode.[12] This data set contained 6,3590,460 MS/MS spectra derived from synthetic peptides in the ProteomeTools project. PXD014222 was acquired from Q-Exactive HF with the HCD activation mode.[13] This data set consisted of 761,549 MS/MS spectra from colorectal cancer tissues. PXD002395[14] and PXD001468[15] were acquired from the LTQ Orbitrap Velos and Q-Exactive with the HCD activation mode, respectively. The two data sets contained 576,114 and 1,121,149 MS/MS spectra from Hela and HEK293 cell lines, respectively. PXD004732, PXD014222, PXD002395 and PXD001468 are referred to herein as the ProteomeTools, colorectal cancer (CRC), Hela, and HEK293, respectively.

Training and Test

We performed *de novo* peptide sequencing on four data sets using three different tools: 1) PEAKS (version 10.6), an algorithm-based commercial tool, 2) Casanovo (v3.3.0), a deep learning-based tool using transformer, and 3) pNovo3 (v3.1.2), combining both algorithm and machine learning techniques to sequence peptides. For each data set, appropriate precursor tolerances (10 or 20 ppm) and fragment tolerances (0.02 or 0.025 Da) depending on data sets were used (Table S1). Carbamidomethylation on Cys was set as a fixed modification, and oxidation on Met was set as a variable modification. For Casanovo, we sequenced spectra using a pretrained model for tryptic/HCD (https://github.com/Noble-Lab/casanovo/releases/tag/v3.0.0) and subsequently discarded modified peptides with undesirable variable modifications except for oxidation on Met.

When training a deep learning model in NovoRank, we used the MaxQuant search results of the ProteomeTools data set (synthetic peptides). This low complexity MS/MS spectrum allowed us to learn a clear pattern of fragment ion peaks. To reduce false positives in the training data set, we only extracted peptide-spectrum matches (PSMs) with a posterior error probability (PEP) below 0.01 to obtain PSMs with high-confidence. As a result, we retrieved 3,506,774 PSMs corresponding to 134,615 peptides and subsequently used

90% of the PSMs for training and the remaining 10% for testing. By partitioning the training set, the validation set for evaluating the deep learning model was obtained, and a 5-fold cross validation was conducted. Train, validation, and test sets were separated so that the peptides never overlapped. The model was trained for a total of 10 epochs, optimizing based on validation loss. Finally, to evaluate the model, we retrained it using the entire ProteomeTools data set and tested it on three independent data sets.

After training the model using the ProteomeTools data sets, we evaluated the performance of the trained model using three independent data sets (HEK293, Hela and CRC). To define positives for the three data sets, we conducted a Comet (v2021.01) search against the human protein sequences (SwissProt v2021.04; 42,336 human protein sequences with isoforms) and 179 common laboratory contaminants, concatenated with pseudoreverse decoy sequences. The search parameters were the same as those for *de novo* peptide sequencing. The false discovery rate (FDR) was calculated as the ratio of decoy to target PSMs above the XCorr score threshold using the target-decoy approach.[16] We selected the score threshold that maximized the number of target PSMs while keeping the FDR below 0.01. As a result, we obtained 380,854 PSMs (138,644 peptides), 257,654 PSMs (76,536 peptides), and 298,485 PSMs (109,888 peptides) from HEK293, Hela and CRC data sets, respectively. Those data sets were used to evaluate the performance of NovoRank and deposited in Zenodo at 10.5281/zenodo.14046459.

NovoRank Algorithms

Once peptides were sequenced by any *de novo* peptide sequencing tool, NovoRank generated new candidate peptides from the initial identification results and selected the best peptide per MS/MS spectrum (Figure 1A). In the new candidate generation step, we used the top *n* candidate peptides for each spectrum obtained from the *de novo* peptide sequencing, along with the spectral clustering result obtained by sequentially applying two clustering methods, MS-Cluster[17] and DBSCAN[18] (Figure 1B). Since pNovo3 allows either one or ten candidate peptides per spectrum, we used the top ten candidate peptides to ensure a fair comparison in this study. Initially, spectra with similar patterns were grouped together using MS-Cluster with the following parameters: "--fragment-tolerance" set to 0.02 and "--mixture-prob" set to 0.01, forming preliminary clusters that might still contain some degree of heterogeneity. To refine these clusters, we applied the DBSCAN algorithm, further segmenting the initial clusters based on precursor m/z and RT in minutes, sequentially. RT was consistently set to 2 across all samples, and precursor m/z aligned with the *de novo* peptide sequencing parameter. It is not straightforward to determine the optimized parameter settings, but this configuration showed reasonably good performance (Figure S1). This two-step clustering approach resulted in more coherent clusters, where spectra in the same cluster were more likely to originate from the same peptide sequence.

After clustering, all of the original *de novo* peptide sequencing results, known as PSMs, were collected within the cluster (Figure 1C). To reflect the frequency of candidate peptides in the same cluster, we introduced the cluster-score (C-score) as follows:

$$C\text{-}score_{p,C} = \frac{\sum_{i\in C} score_{p,i}}{|C|}$$

For each candidate peptide p in the cluster C, its C-score was defined by dividing the sum of original *de novo* scores of a candidate peptide p for each spectrum $i \in C$ by the cluster size $|C|$. Thus, the C-score rewards peptides with higher *de novo* scores and penalizes less frequent peptides within a given cluster. In our example shown in Figure 1C, a peptide candidate EPPTIDE was assigned a C-score of 43, because it had only one *de novo* score of 86 and its cluster consisted of two spectra. In the reranking problem, the C-score serves as an important feature for selecting the most desirable candidate peptide within the same cluster; therefore, the average of *de novo* scores is functionally equivalent to their sum. However, we have mainly two reasons for using the average of *de novo* scores instead of their sum. First, it is not a simple average but a normalizing sum by its cluster size, which reflects the frequency of peptides within the cluster. Second, it allows for discrimination between clusters with the same total *de novo* scores but different sizes. For example, a small cluster with a total *de novo* score of 100 should not be treated the same as a large cluster with the same total *de novo* score.

In the reranking step, the two peptides with the highest C-score in a cluster were input to a deep learning model (Figure 1D). We designed the model to embed PSM information for each peptide and compare the two embedding vectors so that the model can predict a relative score representing the better peptide for a given MS/MS spectrum. To achieve this, we implemented the model with two parts: 1) PSM embedding and 2) comparing two PSMs.

For embedding a PSM, we used a concept equivalent to the Siamese Network[19] and RankNet,[20] featuring two identical subnetworks with shared weights. This ensures that each PSM is embedded in parallel without any bias toward the input order (i.e., preventing one subnetwork from learning only from positive data and the other from negative data). In the subnetwork, we used a convolutional neural network (CNN) which effectively extracted the localized features and patterns in the high-dimensional MS/MS spectra. For embedding a peptide sequence, we employed a bidirectional long short-term memory (BiLSTM) network, which captured the sequential features of the peptide sequences, preserving the order and contextual information of amino acids. The embeddings from the CNN (for the MS/MS spectrum) and BiLSTM (for peptide) were combined with additional PSM features: C-score, XCorr, internal fragment ions, and delta RT (Table S2), predicting RT using DeepLC[21] for delta RT calculation.

The two vectors emitted from each subnetwork were concatenated with delta C-score and delta XCorr, leveraging both the learned representations of the spectra and sequences as well as the relative scores to improve the accuracy of peptide reranking. This combined vector was transformed into a single value, ultimately indicating which identification is better. A detailed explanation of the deep learning model is provided in Supplementary Note 1.

NovoRank Score Threshold

The output value of the deep learning model is between 0 and 1. Given two PSM inputs, the first sequence is selected if the output value is greater than 0.7 and the second sequence is selected if the output is smaller than 0.3. If the value is closer to 0.50, the NovoRank result is rejected and the original *de*

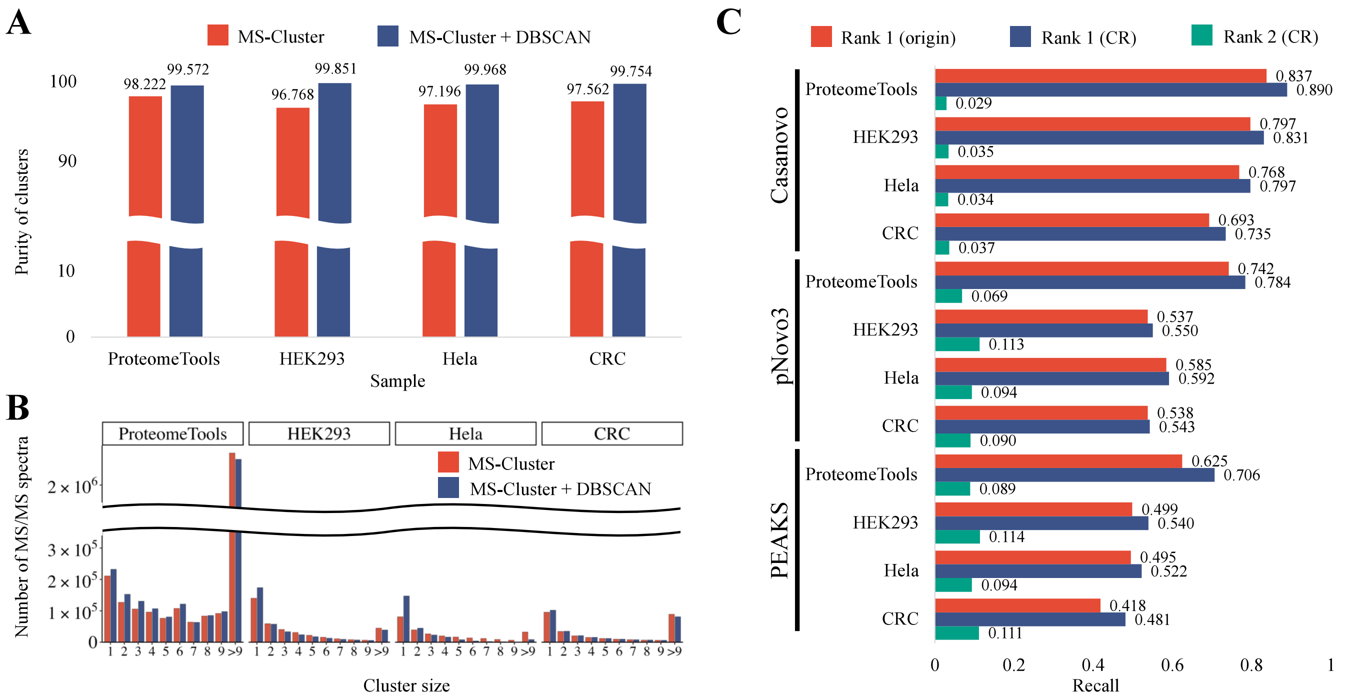


Figure 2. Performance evaluation of two-step clustering and C-score reranking (CR) methods. (A) Purity of clusters for each data set is displayed for MS-Cluster with/without DBSCAN. (B) Distributions of the number of tandem mass (MS/MS) spectra according to cluster sizes for MS-Cluster with/without DBSCAN are shown across samples. (C) Comparison of recall at a peptide-spectrum match level between original and CR results across tools and data sets.

novo peptide sequencing result is accepted. The boundary cutoff values were determined through empirical comparisons of six thresholds (Table S3).

RESULTS AND DISCUSSION

Higher Purity of Two-Step Clustering Method

As input features for machine learning, NovoRank utilized traditional peptide-spectrum match attributes but also employed multiple spectral features (i.e., those defined by spectral clusters). Assuming that similar spectra were likely to be derived from the same peptide, clusters could lead to more robust decisions. We employed MS-Cluster to aggregate spectra by spectral similarity. Unlike typical clustering for spectral library construction, we used MS-Cluster to gather similar spectra but did not generate a representative (consensus) spectrum. Based on confidently identified PSMs retrieved from the ProteomeTools, HEK293, Hela and CRC data sets' database search results, we evaluated the spectral clustering. If a cluster contains a single peptide, then it was defined as a unique cluster. The purity of clustering results was calculated as the fraction of unique clusters. MS-Cluster showed high purity across four data sets. However, its clustering was solely based on peak lists of spectra, resulting in clusters with spectra having distant retention times and/or precursor m/z, thus increasing the cluster impurity. To resolve the problem, a secondary refinement was performed on each preliminary cluster using DBSCAN, considering both precursor m/z and RT. This two-step approach achieved an average purity over 99.57% (less than 1% of the clusters contained mixed peptides), demonstrating that our two-step approach could cluster spectra effectively while rarely sacrificing sensitivity (Figure 2A and Figure S2).

Due to the trade-off between purity and cluster size, the two-step clustering generated smaller clusters compared to MS-Cluster (Figure 2B). For example, in the Hela data set, the two-step clustering generated 1.82 times more unique clusters by further dividing clusters of size five or more. While this approach might result in unnecessarily splitting clusters and potentially losing cluster information (Figure S2), it resulted in better accuracy by reducing impure clusters. Since these clusters were used to calculate more reliable scores of candidate peptides, we opted to use the higher purity method, *i.e.*, two-step clustering.

Notably, in the ProteomeTools data set, 72.51% and 69.45% of MS/MS spectra were grouped into clusters with a size above nine by MS-Cluster and the two-step clustering, respectively, largely due to the redundant scanning of synthetic peptides.

Improved Identification by C-Score Reranking

Given the two-step clustering results, we introduced a new feature, C-score, which represents agreement among multiple MS/MS spectra. To assess the effectiveness of the C-score, we retained each result of *de novo* peptide sequencing from Casanovo, pNovo3 and PEAKS across four data sets. Following the two-step clustering for each data set, we selected the top two candidate peptides per cluster based on C-score and assigned them to each spectrum, ensuring all spectra in the same cluster have the same top two candidate peptides. Using the confidently identified PSMs from database searches as positive data, we compared the top-ranking peptides between the original results and those refined through C-score (Figure 2C and Table S4). We observed a higher recall rate for the refined top-ranking peptides in all tools and data sets at the PSM level. This result implied that a higher portion of similar spectra were inconsistently sequenced in *de novo* peptide sequencing, indicating that scoring functions were too sensitive

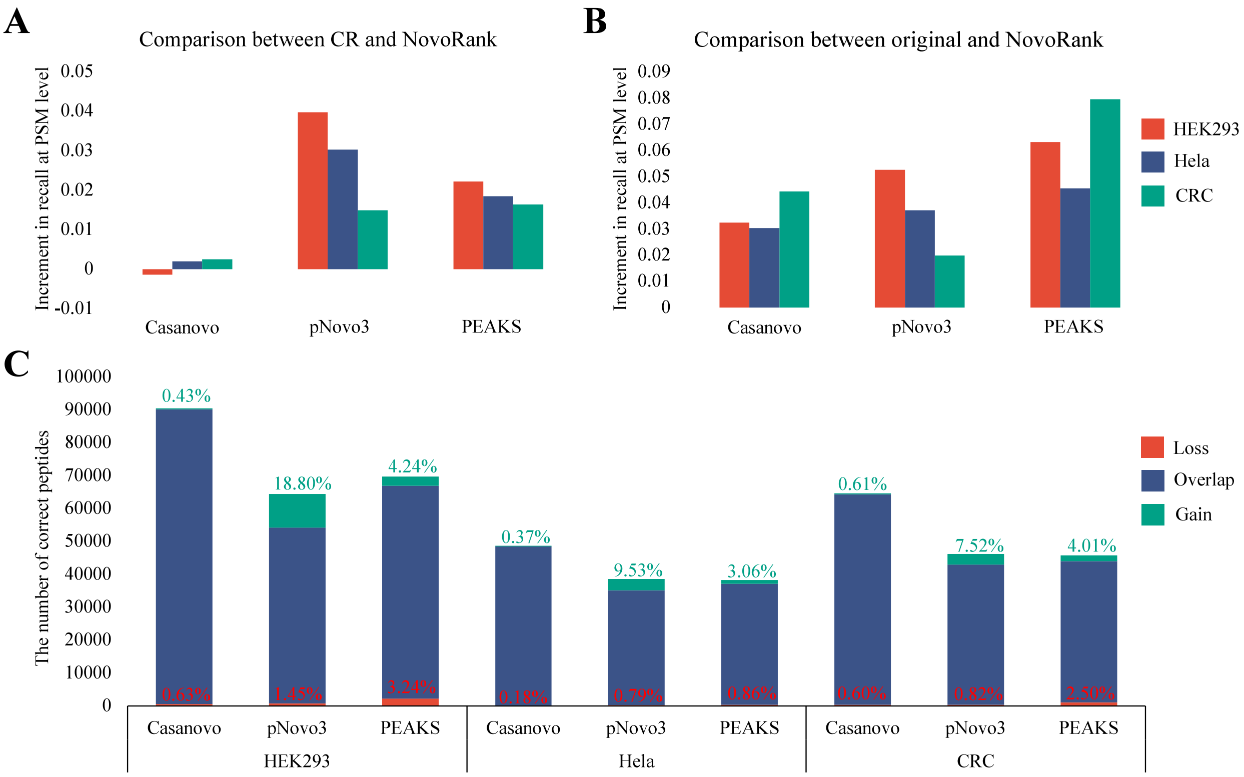


Figure 3. Performance of NovoRank - comparing C-score reranking (CR) and original results across three *de novo* peptide sequencing tools and three data sets. (A) Comparison of recall between CR and NovoRank at a peptide-spectrum match (PSM) level. (B) Comparison of recall between original results and NovoRank at a PSM level. (C) Bar chart visualizing the overall gains, overlaps and losses of correct identifications.

to individual spectral peaks. Interestingly, Casanovo exhibited more improvement compared with pNovo3, even though Casanovo had relatively smaller room for improvement. The ProteomeTools data set showed consistent increments across all three tools. This may be because the synthetic peptides were redundantly scanned by mass spectrometry, making the C-score more reliable. Notably, pNovo3 and PEAKS had a higher proportion of misassignment in rank 2 (an average of 0.092 and 0.102, respectively), indicating that a binary classifier can possibly rescue them. Collectively, reranking candidate peptides using C-score showed enhancement in identifications, even for Casanovo, a state-of-the-art deep learning-based *de novo* peptide sequencing tool, proving the effectiveness of C-score.

Enhanced Performance Using Deep-Learning Approach

We have further developed a deep learning model as a part of NovoRank to select a more plausible peptide between the first and second ranking peptides determined by C-score (details in the Experimental Section). Briefly, we trained a deep learning model for each tool using the ProteomeTools data set and evaluated its performance against the independent published data sets such as HEK293, Hela and CRC data.

We compared the final results of NovoRank with 1) the C-score reranking method and 2) the original search results. Compared to the C-score reranking method, there was no meaningful improvement in Casanovo (Figure 3A). Such results were expected because Casanovo had a lower proportion of correct peptides in rank 2 (less than 3% in the ProteomTools data set) (Figure 2C), making it difficult for the model to learn the properties of rank 2 PSMs. In practice, the simpler method of C-score reranking without deep learning optimization would be sufficient for Casanovo. On the other hand, further increments in recall rates were observed in both pNovo3 and PEAKS across all three data sets. In contrast to the lower increments by the C-score reranking method observed in pNovo3 (less than 1.4%) (Figure 2C), an average of 0.028 increment was achieved by NovoRank. When compared with the original search results, NovoRank achieved an average of 0.036, 0.037, and 0.063 increased recall at the PSM level in Casanovo, pNovo3 and PEAKS, respectively (Figure 3B).

Next, we examined whether the enhanced PSM identification yielded an increase in the peptide identification rate (Figure 3C). When compared with the original search results, NovoRank missed only 0.18%−0.63%, 0.79%−1.45% and 0.86%−3.24% of the correct peptides in Casanovo, pNovo3 and PEAKS, respectively. This indicated that after refining PSMs, 98.77% of peptides were consistently identified on average, demonstrating that NovoRank could robustly refine PSMs without comprising sensitivity. In terms of gains, NovoRank achieved 0.37%−0.61%, 7.52%−18.80% and 3.06%−4.24% increases in correct peptide identifications in Casanovo, pNovo3 and PEAKS, respectively. These gains were consistently higher than their losses in pNovo3 and PEAKS; however, Casanovo brought about no meaningful changes at a peptide level (less than 1% for both gains and losses), while an average of 0.036 improvement in recall at a PSM level was observed (Figure 3B). It is noteworthy that the original results of PEAKS identified more than the original results of pNovo3 at the PSM level in all samples; however, pNovo3 outperformed PEAKS in Hela and CRC after applying NovoRank. Moreover, a large improvement in pNovo3 was observed on HEK293, implying that NovoRank is more compatible with

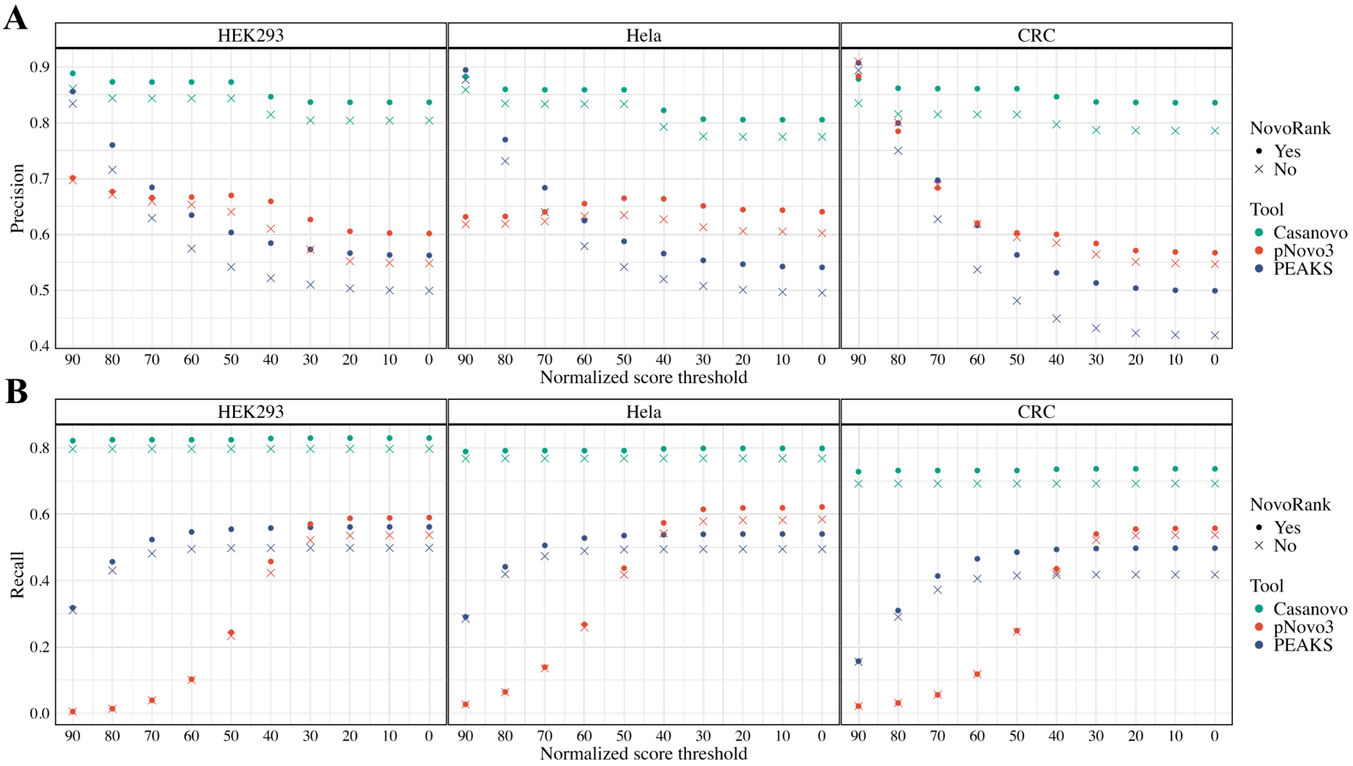


Figure 4. Precision and recall of NovoRank across three *de novo* peptide sequencing tools and three data sets. Precision and recall are depicted in (A) and (B), respectively, according to various score thresholds.

pNovo3 among the three tools. Taken together, these results demonstrate that NovoRank can robustly refine results of *de novo* peptide sequencing tools adopting the deep learning approach in combination with the C-score reranking.

Practical Assessment of NovoRank's Performance

Due to the lack of proper FDR estimation methods for *de novo* peptide sequencing, score thresholding has been widely used in practice. Therefore, traditional evaluations, such as precision and recall, according to each data point, hardly provide users with a sense of a proper score threshold. To make the evaluation more practically meaningful, the scores were normalized to range from 0 to 100. And then we calculated precision and recall for scores above specific thresholds (e.g., 90, 80, ···, 0). This approach allowed us to present cumulative estimates of precision and recall for scores exceeding these thresholds, thereby facilitating a more effective assessment of the method's performance.

We treated isoleucine as leucine and used only a positive data set (i.e., PSMs with an FDR below 1%) to evaluate the performance of NovoRank. We defined a true positive (TP) when the predicted peptide sequence above a score threshold exactly matched the ground truth sequence. While a false positive (FP) is defined as where the predicted peptide sequence above a score threshold does not match the ground truth, a false negative (FN) occurs if the *de novo* tool fails to predict a peptide sequence (*i.e.*, no result is returned for the MS/MS spectrum) or if the score is below the threshold. In this analysis, we focused on precision and recall to evaluate the performances. Based on these definitions, precision and recall are defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

After that, we further evaluated the performance by comparing the precision and recall between the original results and those refinements made by NovoRank according to the score thresholds at the PSM level (Figure 4 and Figure S3).

As expected, the precision and recall were consistently improved by NovoRank across the score thresholds in general. Interestingly, the three tools showed clearly different patterns regardless of the refinement by NovoRank. pNovo3 showed unstable precision, with scores exceeding 70; for example, it achieved precision of over 0.9 in CRC data, while the others ranged between 0.6 and 0.7. We could say that a sparsity of high-scoring PSMs in pNovo3 resulted in unstable performance estimation at the precision level. On the other hand, Casanovo showed consistent precision and recall across all score thresholds in terms of the original results as well as refinements by NovoRank, indicating that correct PSMs tended to have high-scores. When we used a score threshold of 90 for Casanovo after applying NovoRank, we could expect a high precision (~0.88) with a satisfactory recall greater than 0.7. This is particularly valuable for the identification of novel peptides. As for PEAKS, it showed the most typical curves across three samples in terms of precision and recall rate. Using score threshold 90 for PEAKS, we could expect a comparable precision to that of Casanovo. To sum up, these empirical evaluations demonstrated an enhanced performance achieved by NovoRank and provided valuable insights into the proper score thresholds for utilizing *de novo* peptide sequencing in practical applications.

CONCLUSIONS AND FUTURE WORK

In this work, we introduced NovoRank as a deep learning based postprocessing tool for *de novo* peptide sequencing, capable of improving both precision and recall at the same time. To demonstrate that NovoRank can be effective regardless of specific *de novo* sequencing tools employed, we widely categorized *de novo* peptide sequencing methods into 1) an algorithm-oriented (PEAKS), 2) a machine learning based (pNovo3) and 3) a deep learning based (Casanovo) approach and evaluated NovoRank performance when each method was adopted. It would be particularly interesting to compare between pNovo3 and pNovo with NovoRank, because pNovo3 is an upgraded version of pNovo with a machine learning-based reranking module of its own; however, the standalone software of pNovo has been deprecated. We have not performed comprehensive comparison with Spectralis, a most recently published postprocessing tool for *de novo* peptide sequencing.[22] The authors have applied Casanovo (v3.2.0) results to test the performance of Spectralis as a rescorer. Similarly with the NovoRank, it showed marginal improvement in recall at 90% precision, implying that both methods may have the potential to offer comparable performance.

We used four different MS/MS data sets—ProteomeTools, HEK293, Hela and CRC data sets—and evaluated the C-score reranking method. The results demonstrated that the simple two-step clustering and C-score reranking can improve results in the three *de novo* peptide sequencing tools, indicating a potential extension to any tools. To further enhance accuracy, we introduced a newly designed deep learning model for reranking, embedding both spectra and sequences to find hidden information. Using six features that assist in distinguishing accurate peptide spectrum matches, we achieved additional performance gains in the *de novo* sequencing results. Moreover, we provided valuable insights to select proper score thresholds for *de novo* peptide sequencing, which could be particularly useful for the applications oriented in identifying novel peptides such as neoantigen discovery. While the performance of NovoRank for Casanovo may be somewhat limited at the peptide level, the enhanced identification of PSMs across all three tools highlights its potential as a versatile reranking tool for *de novo* peptide sequencing.

Lastly, we designed NovoRank to rerank candidate peptides at the PSM level, while the identification is ultimately influenced by the level of evidence from multiple spectra in the same cluster. As future work, an improved algorithm that addresses the peptide level (or cluster level) identification probability (or score) rather than the PSM level could be beneficial to select the better candidate peptide.

ASSOCIATED CONTENT

Data Availability Statement

The source code of NovoRank is freely available at https://github.com/HanyangBISLab/NovoRank and is licensed under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International. Data used for training and inference as well as model predictions used to generate results are available on Zenodo at 10.5281/zenodo.14046459.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.4c00300.

Table S1. Data sets and data processing used in training and evaluation. Table S2. Description of six additional features for a deep learning model. Table S3. Peptide precision and recall according to the NovoRank score threshold. Table S4. Changes in top-ranking peptides based on C-score reranking. Figure S1. Purity and number of clusters according to DBSCAN parameters. Figure S2. Number of clusters and proportion of fully clustered MS/MS spectra. Figure S3. A conventional precision-recall curve. Supplementary Note 1. A neural network architecture of NovoRank. (PDF)

AUTHOR INFORMATION

Corresponding Authors

Seunghyuk Choi − *Department of Computer Science, Hanyang University, Seoul 04763, Republic of Korea;* orcid.org/0000-0001-6558-9379; Phone: +82 2-2220-4704; Email: ok30010@hanyang.ac.kr

Eunok Paek − *Department of Artificial Intelligence and Department of Computer Science, Hanyang University, Seoul 04763, Republic of Korea;* orcid.org/0000-0003-3655-9749; Phone: +82 2-2220-2377; Email: eunokpaek@hanyang.ac.kr; Fax: +82-2-2200-1723

Author

Jangho Seo − *Department of Artificial Intelligence, Hanyang University, Seoul 04763, Republic of Korea*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.4c00300

Author Contributions

J.S. performed proteomic data analysis; S.C. and E.P. designed this study; and J.S. developed and implemented the software tool. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) [2019M3E5D3073568] and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [No. 2020-0-01373, Artificial Intelligence Graduate School Program (Hanyang University), No. 2021-0-02068, Artificial Intelligence Innovation Hub].

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Eng, J. K.; McCormack, A. L.; Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976−989.

(2) Cox, J.; Mann, M. MaxQuant Enables High Peptide Identification Rates, Individualized P.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification. *Nat. Biotechnol.* **2008**, *26* (12), 1367−1372.

(3) Kim, S.; Pevzner, P. A. MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics. *Nat. Commun.* **2014**, *5*, 5277.

(4) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: An Open-Source MS/MS Sequence Database Search Tool. *Proteomics* **2013**, *13* (1), 22−24.

(5) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. PEAKS: Powerful Software for Peptide de Novo Sequencing by Tandem Mass Spectrometry. *Rapid Commun. Mass Spectrom.* **2003**, *17* (20), 2337−2342.

(6) Yang, H.; Chi, H.; Zeng, W.-F.; Zhou, W.-J.; He, S.-M. pNovo 3: Precise de Novo Peptide Sequencing Using a Learning-to-Rank Framework. *Bioinformatics* **2019**, *35* (14), i183−i190.

(7) Yilmaz, M.; Fondrie, W. E.; Bittremieux, W.; Melendez, C. F.; Nelson, R.; Ananth, V.; Oh, S.; Noble, W. S. Sequence-to-Sequence Translation from Mass Spectra to Peptides with a Transformer Model. *Nat. Commun.* **2024**, *15* (1), 6427.

(8) McDonnell, K.; Howley, E.; Abram, F. The Impact of Noise and Missing Fragmentation Cleavages on Peptide Identification Algorithms. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 1402−1412.

(9) Chi, H.; Sun, R.-X.; Yang, B.; Song, C.-Q.; Wang, L.-H.; Liu, C.; Fu, Y.; Yuan, Z.-F.; Wang, H.-P.; He, S.-M.; Dong, M.-Q. pNovo: De Novo Peptide Sequencing and Identification Using HCD Spectra. *J. Proteome Res.* **2010**, *9* (5), 2713−2724.

(10) Liu, T.-Y. Learning to Rank for Information Retrieval. *FNT in Information Retrieval* **2009**, *3* (3), 225−331.

(11) Deutsch, E. W.; Bandeira, N.; Perez-Riverol, Y.; Sharma, V.; Carver, J. J.; Mendoza, L.; Kundu, D. J.; Wang, S.; Bandla, C.; Kamatchinathan, S.; Hewapathirana, S.; Pullman, B. S.; Wertz, J.; Sun, Z.; Kawano, S.; Okuda, S.; Watanabe, Y.; MacLean, B.; MacCoss, M. J.; Zhu, Y.; Ishihama, Y.; Vizcaíno, J. A. The ProteomeXchange Consortium at 10 Years: 2023 Update. *Nucleic Acids Res.* **2023**, *51* (D1), D1539−D1548.

(12) Zolg, D. P.; Wilhelm, M.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Delanghe, B.; Bailey, D. J.; Gessulat, S.; Ehrlich, H.-C.; Weininger, M.; Yu, P.; Schlegl, J.; Kramer, K.; Schmidt, T.; Kusebauch, U.; Deutsch, E. W.; Aebersold, R.; Moritz, R. L.; Wenschuh, H.; Moehring, T.; Aiche, S.; Huhmer, A.; Reimer, U.; Kuster, B. Building ProteomeTools Based on a Complete Synthetic Human Proteome. *Nat. Methods* **2017**, *14* (3), 259−262.

(13) Blank-Landeshammer, B.; Richard, V. R.; Mitsa, G.; Marques, M.; LeBlanc, A.; Kollipara, L.; Feldmann, I.; Couetoux du Tertre, M.; Gambaro, K.; McNamara, S.; Spatz, A.; Zahedi, R. P.; Sickmann, A.; Batist, G.; Borchers, C. H. Proteogenomics of Colorectal Cancer Liver Metastases: Complementing Precision Oncology with Phenotypic Data. *Cancers* **2019**, *11* (12), 1907.

(14) Geiger, T.; Wehner, A.; Schaab, C.; Cox, J.; Mann, M. Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins. *Mol. Cell. Proteomics* **2012**, *11* (3), No. M111.014050.

(15) Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; Gygi, S. P. A Mass-Tolerant Database Search Identifies a Large Proportion of Unassigned Spectra in Shotgun Proteomics as Modified Peptides. *Nat. Biotechnol.* **2015**, *33* (7), 743−749.

(16) Elias, J. E.; Gygi, S. P. Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. *Nat. Methods* **2007**, *4* (3), 207−214.

(17) Frank, A. M.; Bandeira, N.; Shen, Z.; Tanner, S.; Briggs, S. P.; Smith, R. D.; Pevzner, P. A. Clustering Millions of Tandem Mass Spectra. *J. Proteome Res.* **2008**, *7* (1), 113−122.

(18) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Knowledge Discovery and Data Mining* **1996**, *96* (34), 226−231.

(19) Bromley, J.; Bentz, J. W.; Bottou, L.; Guyon, I.; Lecun, Y.; Moore, C.; Säckinger, E.; Shah, R. Signature Verification using a ″Siamese″ Time Delay Neural Network. *Int. J. Patt. Recogn. Artif. Intell.* **1993**, *07* (04), 669−688.

(20) Burges, C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N.; Hullender, G. Learning to Rank Using Gradient Descent. In *Proceedings of the 22nd International Conference on Machine Learning - ICML '05*; ACM Press: Bonn, Germany, 2005; pp 89−96.

(21) Bouwmeester, R.; Gabriels, R.; Hulstaert, N.; Martens, L.; Degroeve, S. DeepLC Can Predict Retention Times for Peptides That Carry as-yet Unseen Modifications. *Nat. Methods* **2021**, *18* (11), 1363−1369.

(22) Klaproth-Andrade, D.; Hingerl, J.; Bruns, Y.; Smith, N. H.; Träuble, J.; Wilhelm, M.; Gagneur, J. Deep Learning-Driven Fragment Ion Series Classification Enables Highly Precise and Sensitive de Novo Peptide Sequencing. *Nat. Commun.* **2024**, *15* (1), 151.