nature communications

Article

https://doi.org/10.1038/s41467-025-63977-z

Enhancing peptide identification in metaproteomics through curriculum learning in deep learning

Received: 8 February 2024

Accepted: 3 September 2025

Published online: 08 October 2025

Check for updates

Shichao Feng[1], Bailu Zhang[1], Huan Wang[2], Yi Xiong[3], Athena Tian[4], Xiaohui Yuan[1], Chongle Pan[3,5] & Xuan Guo[1]

Metaproteomics offers a powerful window into the active functions of microbial communities, but accurately identifying peptides remains challenging due to the size and incompleteness of protein databases derived from metagenomes. These databases often contain vastly more sequences than those from single organisms, creating a computational bottleneck in peptide-spectrum match (PSM) filtering. Here we present WinnowNet, a deep learning-based method for PSM filtering, available in two versions: one using transformers and the other convolutional neural networks. Both variants are designed to handle the unordered nature of PSM data and are trained using a curriculum learning strategy that moves from simple to complex examples. WinnowNet consistently achieves more true identifications at equivalent false discovery rates compared to leading tools, including Percolator, MS^{2}Rescore, and DeepFilter, and outperforms filters integrated into popular analysis pipelines. It also uncovers more gut microbiome biomarkers related to diet and health, highlighting its potential to support advances in personalized medicine.

Metaproteomics measures complex microbial communities in biological samples from natural environments, such as soil rhizosphere[1], ocean[2,3], and fecal microbiome[4–8]. Understanding the functional roles of microorganisms in an ecosystem is crucial for gaining insights into the interactions and dynamics of the ecosystem[9]. This can provide a deeper understanding of how microorganisms participate in processes, such as nutrient cycling[10], disease state, and supporting the digestive and immune system[11–14]. In shotgun MS-based metaproteomics, tandem mass spectrometry (MS/MS) data is generated as follows: proteins are first hydrolyzed into peptides through an in-solution digestion method, generating a large number of peptides. These peptides are then ionized, isolated, fragmented, and detected in a mass analyzer as they elute from high-performance liquid chromatography (HPLC). A key step in analyzing MS-based metaproteomics data is database searching, which involves comparing the measured mass spectra of the peptides to theoretical mass spectra of peptides in silico digested from protein databases. Each of these comparisons yields a peptide-spectrum match (PSM) score, which measures the similarity between the measured and theoretical mass spectra. The peptide with the highest PSM score is considered the top candidate for the query MS/MS. After the database searching, a filtering step is applied to eliminate false positive identifications by setting a score threshold to obtain a set of confident PSMs at a predefined false discovery rate (FDR).

The PSM scoring function in the database search pipeline serves two key purposes: ranking peptide candidates for a given spectrum to identify the most compatible match and ranking PSMs from a proteomics run to eliminate spurious matches. The challenge of constructing a well-calibrated scoring function has intensified with rapid advances in mass spectrometry and metagenomics sequencing

[1]Department of Computer Science and Engineering, University of North Texas, Denton, TX, USA. [2]College of Informatics, Huazhong Agricultural University, Wuhan, Hubei, China. [3]School of Biological Sciences, University of Oklahoma, Norman, OK, USA. [4]Department of Mathematics, Emory University, Atlanta, GA, USA. [5]School of Computer Science, University of Oklahoma, Norman, OK, USA. e-mail: cpan@ou.edu; xuan.guo@unt.edu

technologies, which have led to a substantial increase in the number of mass spectra and the size of protein databases. Ideally, an MS/MS spectrum should achieve a high score for its match with the correct peptide, while random matches typically follow a probabilistic distribution with a small tail of high scores. As peptide databases grow larger, the likelihood of an incorrect random match scoring higher than the correct match increases. Consequently, developing efficient and sophisticated PSM filtering algorithms to re-score PSMs for improved ranking has become imperative.

In recent years, various PSM filtering algorithms have been proposed. Statistical methods, such as PeptideProphet[15], Tailor[16], and H-Score[17], use approaches like Bayesian statistical assumptions, empirical observations, and confidence-based recalibrations, respectively. Machine learning (ML)-based algorithms, such as Percolator[18], CRanker[19], QRanker[20], and Gradient Boosting[21], identify confident PSMs to train models that classify remaining matches. Other methods leverage spectrum comparison features, as seen in MS^{2}Rescore[22], or integrate results from multiple search engines for comprehensive analysis, exemplified by iProphet[23] and IDPicker[24]. Our previous work, Sipros-Ensemble[25], employed logistic regression to calculate new scores based on three distinct scoring functions.

While these approaches effectively extract PSM features like charge states and mass errors, they may not fully exploit the information within measured and theoretical spectra. To address this, we proposed DeepFilter[26], a deep learning architecture that automatically learns matching patterns between measured and theoretical spectra, complemented by human-engineered features. Although DeepFilter achieved promising results, it has limitations: it was trained on data from a single database search engine, restricting its generalizability, and its input format–a large, sparse matrix constructed from ascending-order peak masses–slows inference compared to other widely used filtering tools.

Motivated by the benefits of leveraging MS/MS information and the need to accelerate peptide identification, we developed WinnowNet, a deep-learning-based architecture for re-scoring PSM candidates. WinnowNet utilizes experimentally verified PSM datasets from the ProteomeTools study[27,28] and PSM candidates generated by multiple search engines to construct large, diverse training datasets. The training process employs a curriculum learning strategy[29] to enhance model performance and accelerate convergence. By leveraging the order-invariant properties of CNN and transformer architectures, WinnowNet reduces the representation matrix size while effectively capturing complex matching patterns between measured MS/MS spectra and theoretical peptide spectra. Experimental results demonstrate that WinnowNet significantly improves identifications at the PSM, peptide, and protein levels, outperforming other widely used filters benchmarked in this study. Also, WinnowNet reduces the need for ad hoc training and can be applied to analyze different metaproteome samples without fine-tuning and still obtain substantial improvements over existing tools. WinnowNet is freely available under the GNU GPL license at https://github.com/Biocomputing-Research-Group/WinnowNet[30].

Results

Benchmark datasets and evaluation metrics

To provide a comprehensive performance assessment, WinnowNet was benchmarked on twelve metaproteome datasets. These datasets include those derived from a synthetic microbial mixture (Synthetic), an artificially assembled mock community (P1, P2, and P3), three distinct microbial communities (Marine 1-3, Soil 1-3, Human Gut, and Human Gut TimsTOF), each characterized by increasing complexity in mass spectra and protein databases (see Supplementary Table 1 and Supplementary Note 1). All metaproteome samples except Human Gut TimsTOF were analyzed using the Multidimensional Protein Identification Technology (MudPIT) approach[31] on a Thermo Scientific LTQ Orbitrap Elite mass spectrometer. Human Gut TimsTOF (HGT) dataset was obtained from fecal samples of human patients, analyzed using a trapped ion mobility spectrometry (timsTOF) + TOF mass spectrometry approach.

To ensure an accurate performance comparison and to mitigate overfitting during protein identification, we incorporated entrapment proteins into database search, following approaches proposed in many previous studies[32–35]. Entrapment proteins were generated by randomly shuffling target protein sequences to create false target sequences, which were then used alongside the target-decoy strategy[36]. The effective ratio of entrapment proteins to original target proteins in the database was set to 1:1. Identifications at the PSM, peptide, and protein levels were evaluated at a 1% false discovery rate (FDR), as shown in Eq. (1), where n_t and n_e denote the number of original target and entrapment identifications, respectively. This estimation follows the "combined" method in ref. 35, which provides a conservative upper bound on the FDR. In addition, we performed entrapment analysis using the paired estimation method proposed in ref. 35 (described in Supplementary Discussion), which is proved to yield a tighter bound. Only original target matches with FDR controlled at the predefined level (1% in this study) were reported for all benchmarked methods. It is worth noting that MS/MS spectrum data were extracted using MSConvert from the ProteoWizard release 3.0.11841[37], in contrast to our previous study[26], which utilized RawConverter Version 1.1.0.23[38].

$$Estimated\ FDR = \frac{2 \times n_e}{n_t + n_e} \tag{1}$$

For some experiments, we also used foreign proteins as an entrapment strategy. In this approach, proteins from foreign species were incorporated into the original target protein set to create an extended target database, which was then augmented with decoys generated by randomly shuffling its entries. For this entrapment setup, we estimated the FDR using Eq. (2), where n_t represents the combined set of original target proteins and foreign proteins, and n_d denotes the number of decoys. To further assess the reliability of the identifications, we computed the False Matching Rate (FMR)[33,34], defined as the proportion of false target identifications at a 1% FDR, following Eq. (3). In Eq. (3), n_f represents the number of matches to proteins from foreign species.

$$FDR_{td} = \frac{n_d}{n_t} \tag{2}$$

$$FMR = \frac{n_f}{n_t} \tag{3}$$

Performance comparison to state-of-the-art filtering algorithms

We evaluated WinnowNet against six leading filtering algorithms: Percolator[18], Q-ranker[20], PeptideProphet[39], iProphet[23], MS^{2}Rescore[22], and DeepFilter[26], all of which have been released or updated within the past six years. Unlike traditional filtering algorithms, WinnowNet eliminates the ad hoc training and can be applied to analyze different metaproteome samples without fine-tuning and still obtain substantial improvements over existing tools. The evaluation was conducted using PSM candidates derived from three standalone database search engines: Comet[40], Myrimatch[41], and MS-GF+[42]. Note that Percolator, Q-ranker, and PeptideProphet relied directly on the PSM scores from these search engines, whereas iProphet utilized scores generated by PeptideProphet. To ensure a fair comparison, iProphet was run without the addition of peptide- and protein-level features. While Percolator, Q-ranker, PeptideProphet, and iProphet employ traditional machine learning or statistical methods with human-engineered PSM

features, MS²Rescore enhances rescoring by incorporating predicted peptide fragmentation patterns and retention time information. In contrast, both DeepFilter and WinnowNet are deep-learning-based approaches that automatically learn discriminative features from PSMs.

For performance assessment, we applied the entrapment method described in section "Benchmark datasets and evaluation metrics." Protein identifications were reported only when supported by at least one unique peptide. Identification results for the marine, human gut, soil, and mock datasets at 1% FDR are summarized in Fig. 1 and detailed in Supplementary Tables 16–17. Both WinnowNet variants–the self-attention-based and the CNN-based architectures–achieved the highest numbers of identifications at the PSM, peptide, and protein levels across all datasets and three standalone database search engines. Among the baseline methods, either DeepFilter or MS²Rescore consistently provided the highest identification counts. In the following analysis, we focus on the self-attention-based WinnowNet, which demonstrated the best overall performance; the analysis of the CNN-based variant is provided in Supplementary Methods.

For the marine datasets (see Fig. 1), WinnowNet outperformed MS²Rescore by, on average, identifying 12.6% more PSMs, 12.4% more peptides, and 9.3% more proteins. The intersection bar plot in Fig. 2 shows the overlap of the unique identifications at 1% PSM/peptide/protein FDR levels across the benchmark datasets. Specifically, for the two marine datasets, WinnowNet uniquely identified an average of 7727 PSMs, 5088 peptides, and 1561 proteins, whereas MS²Rescore uniquely identified 2562 PSMs, 1793 peptides, and 875 proteins.

In the human gut dataset, which is the most complex metaproteome tested with extensive MS/MS spectra and comprehensive protein databases, WinnowNet yielded an average increase of 8.0% more PSMs, 6.8% more peptides, and 5.7% more proteins than MS²Rescore. Specifically, WinnowNet uniquely identified 19,463 PSMs, 10,683 peptides, and 4180 proteins, in contrast to 6892 PSMs, 2137 peptides, and 1794 proteins found uniquely by MS²Rescore.

For the soil datasets (see Supplementary Table 16), WinnowNet achieved on average 9.4% more identified PSMs, 11.6% more peptides, and 7.6% more proteins at 1% FDR compared to MS²Rescore. On average, WinnowNet uniquely identified 13,531 PSMs, 4841 peptides, and 1247 proteins, whereas MS²Rescore uniquely identified 5408 PSMs, 1813 peptides, and 836 proteins.

WinnowNet also demonstrated strong performance on the mock datasets, which consist of artificial microbial complexes containing 30 species with uniform protein content. On these datasets, WinnowNet achieved an average increase of 9.1% more identified PSMs, 9.3% more peptides, and 7.5% more proteins at a 1% FDR (see Supplementary Table 17). In addition, WinnowNet uniquely identified up to 14,063 PSMs, 3655 peptides, and 1081 proteins, compared to up to 5905 PSMs, 846 peptides, and 779 proteins uniquely identified by MS²Rescore on average. A detailed analysis of the gained and lost identifications between WinnowNet and MS²Rescore is provided in Supplementary Discussion.

When compared to our previous method, DeepFilter, WinnowNet demonstrated consistent improvements: in the marine datasets, it identified 11.9% more PSMs, 10.0% more peptides, and 6.9% more proteins; in the soil datasets, increases averaged 7.8% for PSMs, 7.7% for peptides, and 4.8% for proteins; and in the mock community, improvements were 4.3% for PSMs, 4.8% for peptides, and 2.9% for proteins. Even in the complex human gut dataset, WinnowNet yielded average gains of 3.4% in PSMs, 3.8% in peptides, and 4.1% in proteins relative to DeepFilter. These results underscore WinnowNet's enhanced ability to leverage spectral information, resulting in improved identification outcomes.

We also collected the number of identifications reported in the original publications for comparison (see Supplementary Table 18). Note that not all publications reported discoveries at the PSM, peptide, and protein levels, nor at the same FDR thresholds. WinnowNet consistently outperformed the original studies in terms of identification counts. For example, while the original publication reported 30,062 proteins, WinnowNet identified 36,143 proteins, representing a 20.2% increase. These findings underscore the potential of applying WinnowNet for the secondary analysis of existing datasets to uncover new biological insights.

Given WinnowNet's improved performance relative to MS²Rescore, we further analyzed the score distributions to evaluate their comparative efficacy. Figure 3 and Supplementary Fig. 14 present score distributions for top-ranked PSMs from the marine 2 and mock P2 datasets as generated by WinnowNet (self-attention-based) and MS²Rescore, respectively. In contrast to MS²Rescore, WinnowNet assigns higher scores to a larger proportion of target PSMs at a 1% PSM-level FDR, with scores predominantly concentrated in the lower-right quadrant delineated by the solid cutoff lines. Notably, a bimodal score distribution is apparent in Supplementary Fig. 14 for MS²Rescore, a pattern not observed in Fig. 3. This discrepancy stems from the differences between the datasets: the Mock dataset employs a well-annotated protein database, whereas the Marine 2 dataset is based on an incomplete protein database derived from annotated assembled genomes. In metaproteomics, incomplete metagenome assemblies and technical biases during sample extraction frequently result in protein databases that represent only a subset of the actual proteome. As a consequence, many spectra correspond to peptides absent from these databases, leading to high-scoring false PSMs and causing the target PSM distribution to approximate that of decoy/entrapment PSMs. To control the FDR and exclude such false positives, metaproteomics analyses often require higher score thresholds compared to those used in simple culture-based proteomics, albeit at the expense of some true PSMs. The results presented in Fig. 3 and Supplementary Fig. 14 clearly demonstrate that WinnowNet's incorporation of spectral information through auto-learned features yields a more robust filtering strategy compared to traditional methods, such as the SVM-based approach employed by MS²Rescore.

Performance evaluation of WinnowNet-integrated protein identification pipelines

We integrated the self-attention-based WinnowNet into four popular protein identification pipelines: our Sipros-Ensemble platform[25], FragPipe[43], Peaks Studio 12.5[44], and AlphaPept[45]. The evaluation was conducted on four benchmark datasets–Marine3, Soil3, P3, and Human Gut. For each pipeline, we compared the recommended workflow against an alternative in which the filtering step was replaced by WinnowNet. Due to the modular architecture of Peaks, directly substituting its built-in filter was not feasible. As a workaround, the alternative Peaks workflow involved performing the database search in Peaks, followed by WinnowNet filtering and protein inference using Philosopher within FragPipe. Unlike traditional filtering algorithms, WinnowNet eliminates the ad hoc training and can be applied to analyze different metaproteome samples without fine-tuning and still obtain substantial improvements over existing tools. The same entrapment method and FDR estimation described in section "Benchmark datasets and evaluation metrics" were applied.

Figure 4 and Supplementary Table 14 present the results. Across all four datasets and pipelines, integrating WinnowNet led to substantial improvements in PSM, peptide, and protein identification levels. For instance, at the PSM level, identifications increased from 61,190 to 66,432 (8.6% improvement) for Sipros-Ensemble, from 47,970 to 53,276 (11.1%) for FragPipe, from 46,727 to 52,789 (13.0%) for Peaks, and from 43,791 to 49,841 (13.8%) for AlphaPept. At the peptide level, improvements were observed from 40,519 to 43,071 (6.3%) for Sipros-Ensemble, from 25,658 to 31,769 (23.8%) for FragPipe, from 24,864 to 30,091 (21.0%) for Peaks, and from 23,895 to 29,857 (25.0%)

https://doi.org/10.1038/s41467-025-63977-z

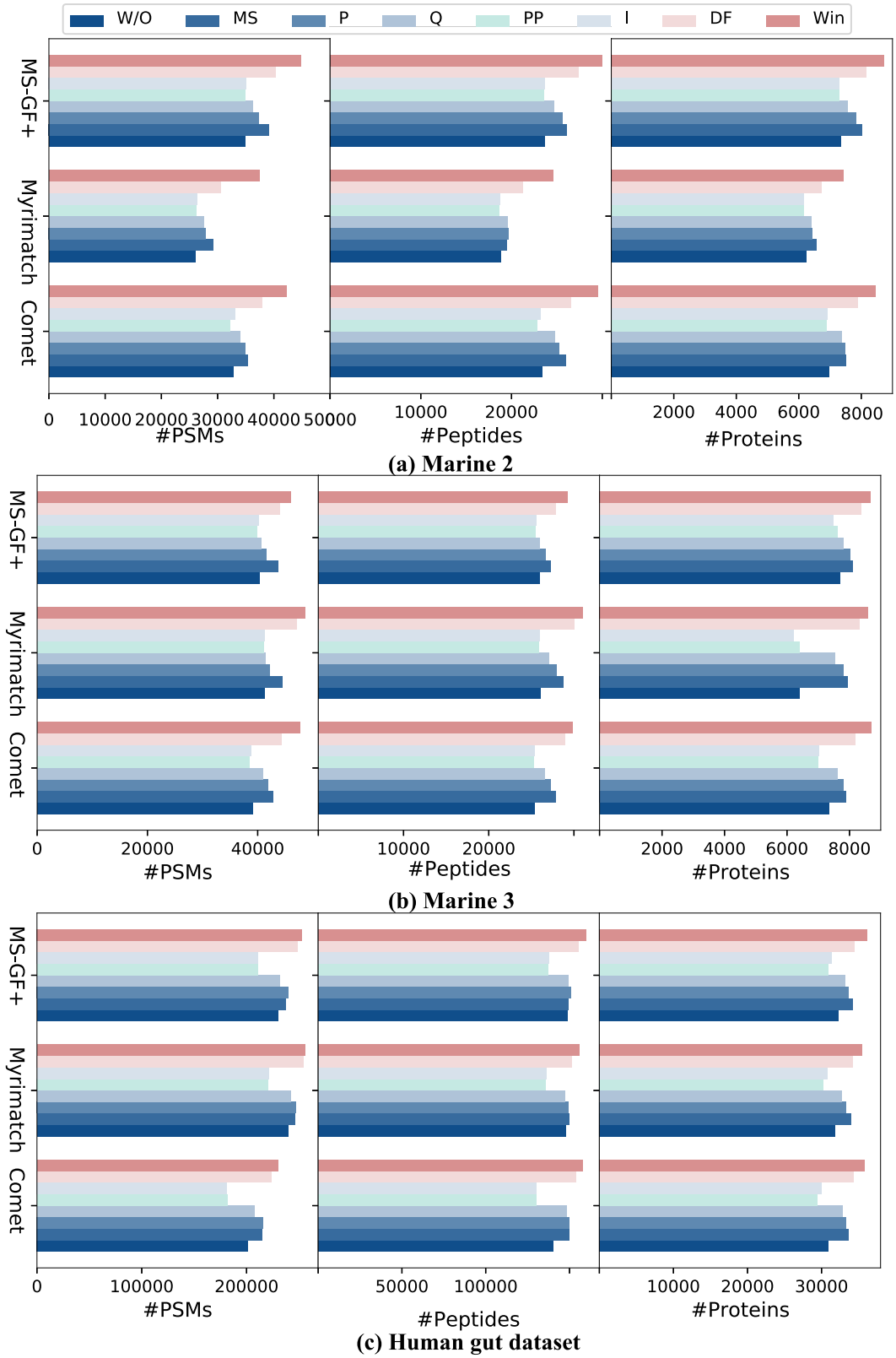

Fig. 1 | Identification results on marine and human gut datasets at 1% FDR using the entrapment method. a Results for the Marine 2 dataset. **b** Results for the Marine 3 dataset. **c** Results for the Human Gut dataset. W/O represents database search results without any filtering; MS refers to MS^{2}Rescore; P to Percolator; Q to Q-ranker; PP to PeptideProphet; I to iProphet; DF to DeepFilter; Win to the self-attention-based WinnowNet.

for AlphaPept. Significant gains were also evident at the protein level; for example, in Marine3, protein identifications increased from 9500 to 10,416 (9.6%) for Sipros-Ensemble, from 9909 to 10,277 (3.7%) for FragPipe, from 9001 to 9327 (3.6%) for Peaks, and from 8796 to 8426 (4.4%). Similar trends were observed in the Soil3, P3, and Human Gut datasets, with percentage gains ranging approximately from 2.5% to 11.1% at the PSM level, 4.4% to 14.5% at the peptide level, and 1.1% to 12.8% at the protein level. These consistent improvements across

https://doi.org/10.1038/s41467-025-63977-z

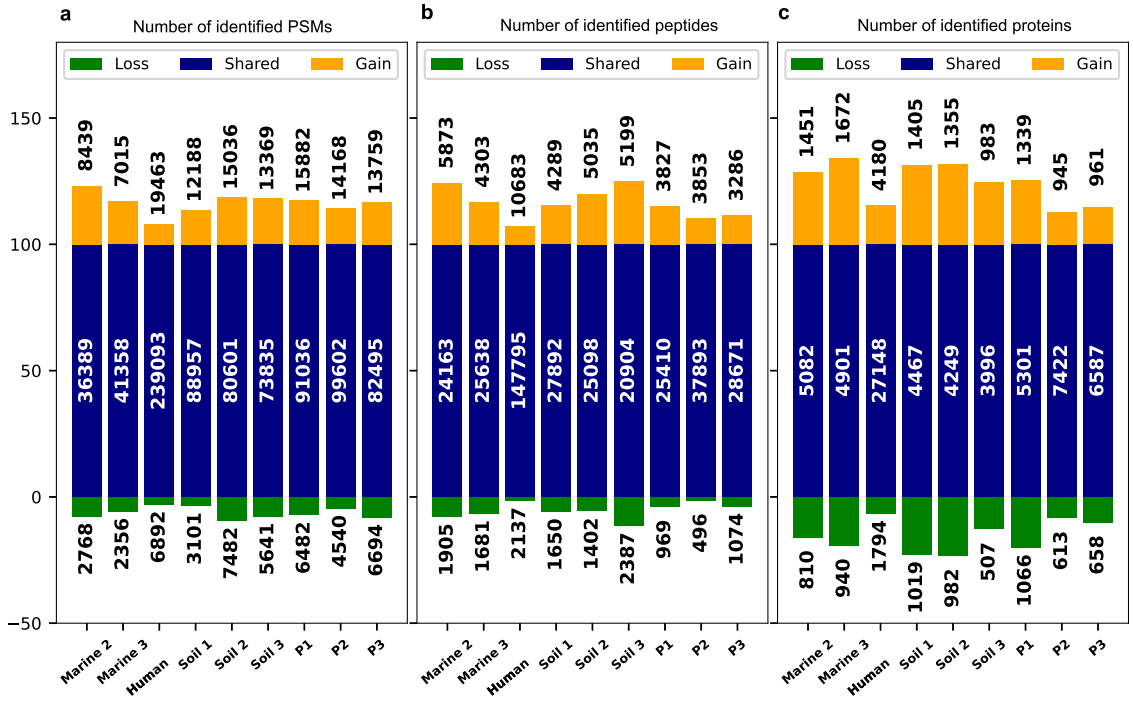

Fig. 2 | Comparison of PSM, peptide, and protein identification results between WinnowNet and MS²Rescore using MS-GF+ output. a PSM-level results. **b** Peptide-level results. **c** Protein-level results. The term Loss denotes identifications exclusively derived from either the MS²Rescore or WinnowNet, while Shared indicates identifications obtained by both methods. Protein groups were excluded from the comparison. Note that the numbers at the center, top, and bottom of each bar indicate the absolute number of identifications at the PSM, peptide, or protein level with a 1% FDR. The y-axis represents the relative percentages of identifications categorized as Shared, Loss, and Gain, with Shared identifications normalized to 100%.

diverse datasets and pipelines highlight the robustness and scalability of WinnowNet.

To simulate real-world analysis conditions for benchmarking, we constructed a composite protein database by combining proteins from mock microbial cultures (30 species) with entrapment proteins from 27 foreign species present in the human gut microbiome (a complete list of foreign species is provided in the Supplementary Table 2). The MS/MS dataset P1, generated from the mock microbial cultures, was searched against this database, which was further augmented with shuffling target sequences as decoys. PSMs corresponding to the mock microbial proteins were considered true identifications, whereas those mapping to the entrapment proteins were treated as false positives. The FDR was estimated using the target-decoy strategy[36] as in Eq. (2) and controlled at 1%. Additionally, we calculated the false matching rate (FMR), defined as the proportion of false target identifications among all accepted targets at 1% FDR (see Eq. (3)). For benchmarking, we employed the same protein identification pipelines as in previous analyses, namely Sipros Ensemble, FragPipe, Peaks, and AlphaPept.

The identification results and FMR values are presented in Fig. 5 and Supplementary Table 15. All original and WinnowNet-enhanced pipelines demonstrated robust performance, consistently maintaining FMRs below 1% at PSM and peptide levels. Notably, the integration of WinnowNet led to consistent improvements in identification accuracy across all pipelines. At the PSM level, WinnowNet increased identifications from 87,275 to 91,271 (+4.6%) for Sipros Ensemble, from 84,448 to 88,691 (+5.0%) for FragPipe, from 85,033 to 89,013 (+4.7%) for Peaks, and from 82,891 to 87,381 (+5.4%) for AlphaPept. Similarly, at the peptide level, the number of identifications increased from 24,633 to 25,827 (+4.8%) for Sipros Ensemble, from 20,235 to 21,138 (+4.5%) for FragPipe, from 21,155 to 22,849 (+8.0%) for Peaks, and from 19,243 to 20,195 (+4.9%) for AlphaPept. At the protein level, WinnowNet also led to notable gains: identifications increased from 7126 to 7389 (+3.7%) for Sipros Ensemble, from 7007 to 7272 (+3.8%) for FragPipe, from 7015 to 7267 (+3.6%) for Peaks, and from 6715 to 6942 (+3.4%) for AlphaPept. These consistent improvements across all levels–PSM, peptide, and protein–highlight WinnowNet's effectiveness in enhancing true identifications while maintaining a stringent FDR threshold.

An additional evaluation was performed using a dataset acquired with the timsTOF instrument. The corresponding results are provided in the Supplementary Discussion.

Analysis of the taxonomic profile of human gut metaproteome

To investigate the biological significance of the proteins identified exclusively by WinnowNet (CNN-based), we searched the human gut protein database against the NCBI public database using Protein-Protein BLAST version 2.11.0+[46]. Pathway annotations were performed using eggnog-mapper against the EggNOG database[47,48]. After excluding the protein groups that shared the same identified peptides, we found 1015 proteins only identified by WinnowNet. In Fig. 6, we present the species associated with proteins only identified by WinnowNet in the human gut metaproteome sample. The phylogenetic tree includes 50 taxa at the species level, providing a detailed taxonomic profile of proteins. Circular phylogenetic tree visualizations in Fig. 6 depict the number of genes and spectra for each species as blue and red bars, respectively. The percentage of genes for each species is represented as a decimal value between 0 and 1, calculated using the min-max normalization method. The number of spectra was determined by counting the total PSMs belonging to each species, considering only the PSMs associated with the unique peptides for each protein.

In our taxonomic profiling analysis, we observed that WinnowNet identified numerous gut microorganisms characterized by low gene counts. The mean normalized number of genes for each species stood at 24.27%, as annotated in the protein database. Notably, 33 species exhibited gene abundances lower than this average, with 4 out of these

https://doi.org/10.1038/s41467-025-63977-z

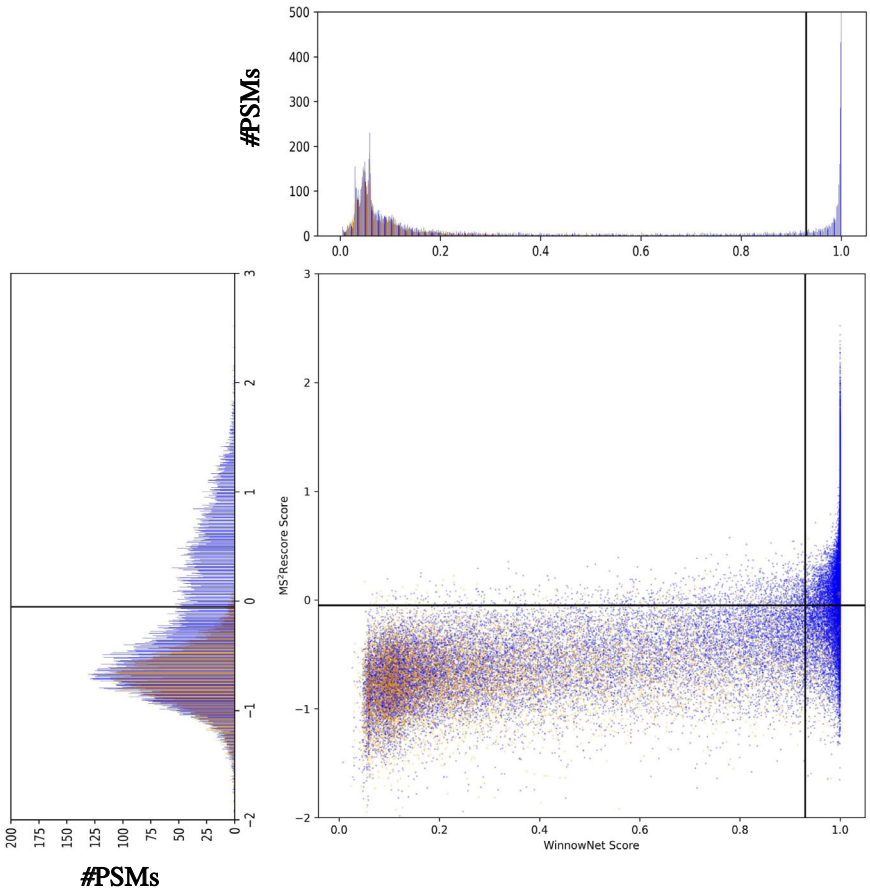

Fig. 3 | Score distributions for top-ranked PSMs in the Marine 2 dataset, derived from WinnowNet (self-attention-based) and MS²Rescore. Original target PSMs are shown in blue, and entrapment and decoy PSMs in yellow. The top panel displays PSM scores from WinnowNet, the left panel shows scores from MS²Rescore, and the central panel illustrates the correlation between WinnowNet scores (x-axis) and MS²Rescore scores (y-axis). Solid lines indicate the 1% PSM-level FDR cutoffs for WinnowNet (0.92) and MS²Rescore (−0.05).

33 species having gene abundances of less than 2%. Interestingly, these four species are recognized as common constituents of the human gut microbiome. For instance, *Dorea longicatena* contributes significantly to short-chain fatty acid (SCFA) production and plays a vital role in dietary carbohydrate metabolism[49,50]. *Hungatella hathewayi* is associated with various human infections[51], while *Sutterella sp. AM11-39* is linked to health conditions such as inflammatory bowel disease and autism[52]. These findings underscore WinnowNet's proficiency in identifying proteins from species characterized by low-abundance genes.

Shifting the focus to the pathway level, proteins exclusively detected by WinnowNet are associated with three distinct pathways identified in the KEGG database: map05231, map00622,and map00440. Notebly, the KEGG pathways Xylene degradation (map00622) and Phosphonate and phosphinate metabolism (map00440) play pivotal roles in probiotics, influencing infant health and human diet[53,54].

Computation time

Table 1 summarizes the computation time for CNN-based and self-attention-based WinnowNet models compared to other filtering algorithms across various datasets. The lightweight architecture of the CNN-based WinnowNet is evident from its significantly reduced number of parameters, containing only 22.2% and 31.5% of those in DeepFilter and self-attention-based WinnowNet, respectively. This results in faster training and inference times, making it an efficient solution for PSM rescoring tasks. The self-attention-based WinnowNet model, with 2.6 million parameters, incorporates advanced spectrum representations at the expense of increased computation time. Benchmarks were performed on a workstation equipped with 8 NVIDIA GeForce RTX 2080 Ti GPUs (12 GB memory each) for neural network models. Baseline algorithms were executed on a desktop computer with a 2.3 GHz Intel Xeon Gold 5118 CPU and 32 GB of memory. Under GPU acceleration, the CNN-based WinnowNet completed most rescoring tasks within 10 minutes, while the self-attention-based WinnowNet required under 30 minutes for the majority of datasets. These results demonstrate the adaptability of WinnowNet, providing users with a trade-off between computational efficiency and advanced feature representation depending on their specific requirements.

Discussion

Existing re-scoring algorithms rely primarily on human-engineered features derived from PSM properties and spectrum comparison attributes. Recently, fragment intensity prediction models using deep learning have emerged to improve re-scoring accuracy[55–57]. In this study, we introduced WinnowNet, a re-scoring framework featuring two neural network architectures: one optimized for accuracy to reduce false identifications (self-attention-based WinnowNet) and the other lightweight for enhanced inference speed (CNN-based WinnowNet). While WinnowNet successfully improves PSM identification, it currently operates as a post-processing tool dependent on pre-screened candidates generated by traditional database search engines. A natural question arises: can we improve database search engines to such an extent that machine learning-based rescoring becomes redundant? Database search engines remain essential for candidate

https://doi.org/10.1038/s41467-025-63977-z

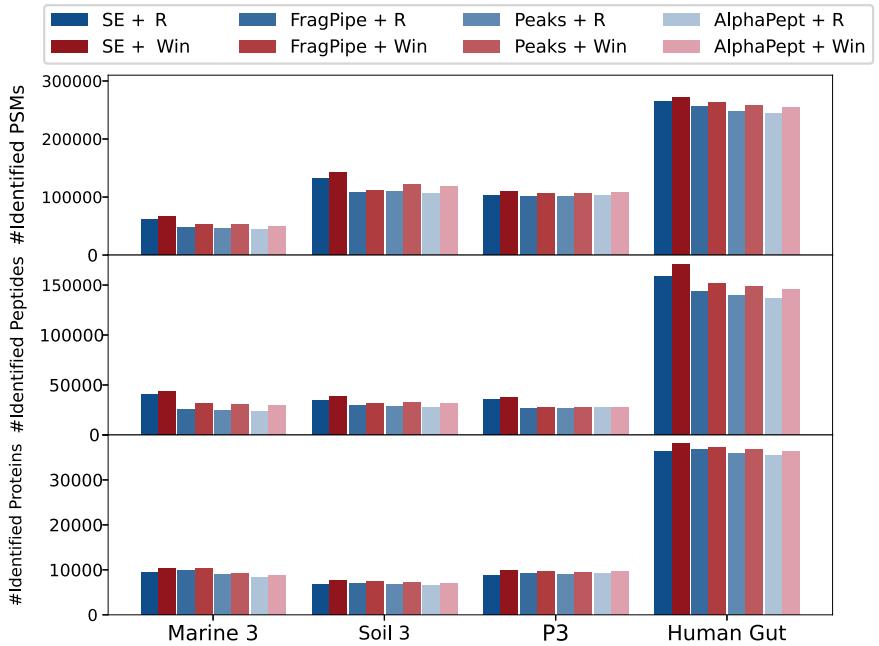

Fig. 4 | PSM, peptide, and protein identification results at 1% FDR on the Marine3, Soil3, P3, and human gut datasets, using four metaproteomics pipelines with either default filters or WinnowNet as an alternative filter. SE denotes Sipros-Ensemble, R refers to the default filters used in the pipelines, and Win represents the self-attention-based WinnowNet method.

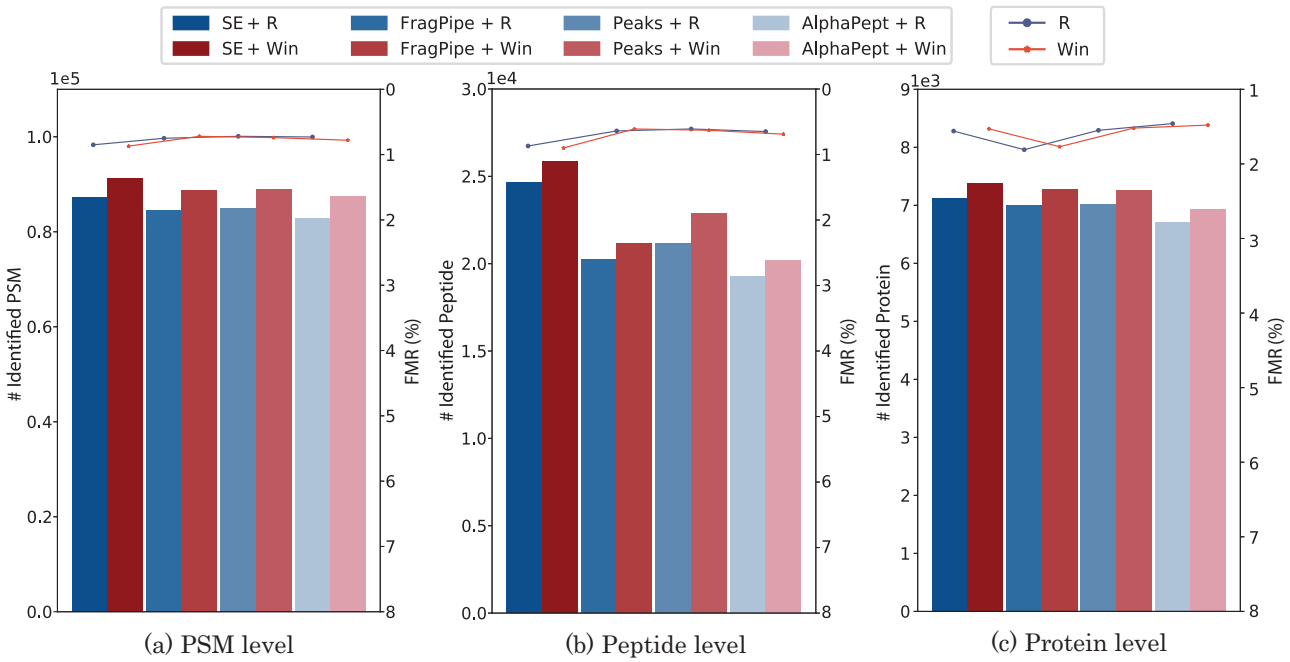

Fig. 5 | Comparison of identification results at multiple levels. a–c PSM, peptide, and protein identification results, along with the false matching rate (FMR), at 1% FDR on the mock P1 dataset mixed with foreign species, using four metaproteomics pipelines with either default filters or WinnowNet as an alternative filter. The blue line plot shows the FMR using default filters, while the red line plot shows the FMR when using WinnowNet. SE denotes Sipros-Ensemble, R refers to the default filters used in the pipelines, and Win represents the self-attention-based WinnowNet method.

generation, but their shortcomings, particularly in handling complex spectra from metaproteomes and highly homologous peptides, limit their effectiveness. WinnowNet bridges these gaps by leveraging deep learning to optimize the rescoring step, effectively mitigating false identifications that arise from limitations in the initial database search. That said, re-scoring tools alone cannot eliminate the need for robust search engines.

To address these limitations, we envision a future extension of WinnowNet into a comprehensive database search engine. Moving beyond its current reliance on pre-processed PSM candidates, we aim

https://doi.org/10.1038/s41467-025-63977-z

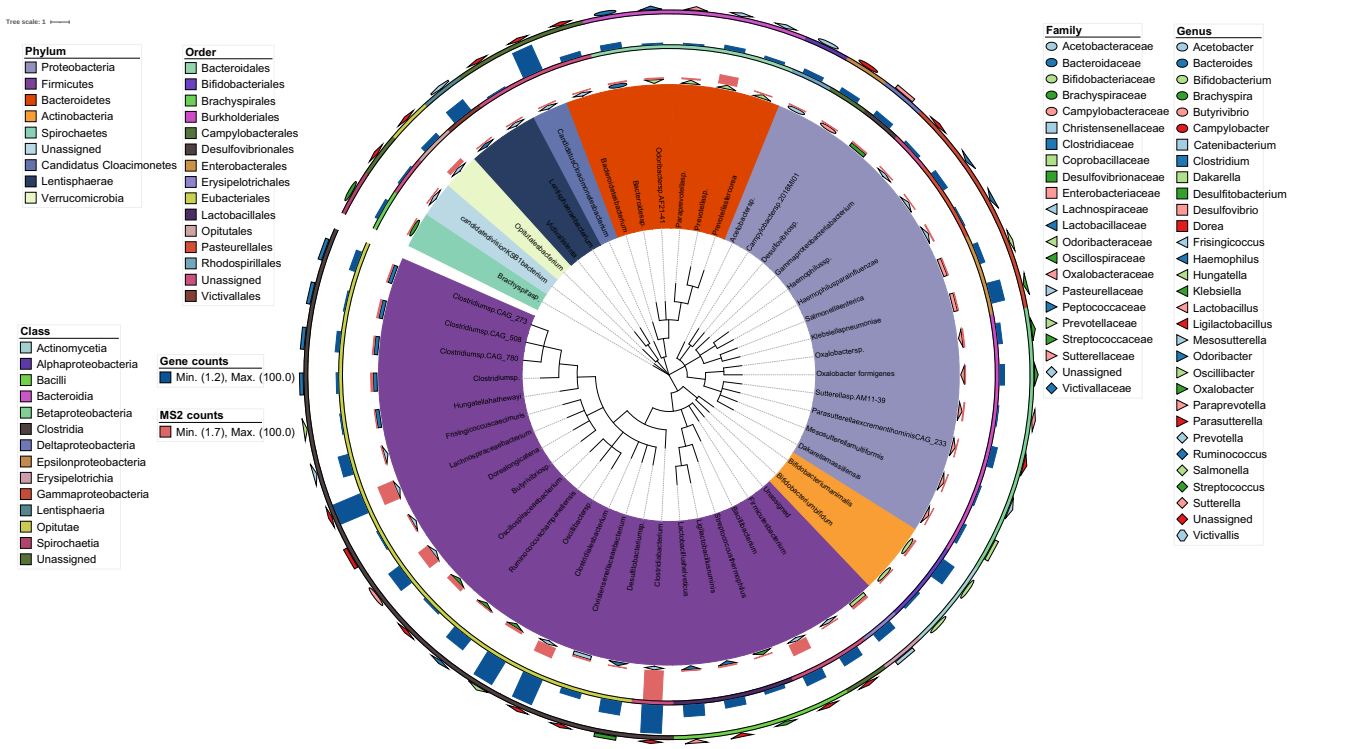

Fig. 6 | Phylogenetic tree of proteins uniquely identified by WinnowNet. Pholygenetic tree with corresponding gene and spectral abundance searched by the proteins only identified by WinnowNet*.

Table 1 | Computation time for benchmark datasets (precise to second)

#Params[c]	Filters[a]	P –	Q –	PP –	I –	DF 3.7M	W* 0.82M	W 2.6 M
Datasets[d]	Size[b]							
M2	0.91M	134	251	68	314	218	74	155
M3	0.89M	127	242	82	300	210	73	149
S1	2.9M	359	671	159	835	678	286	480
S2	2.5M	277	550	142	691	576	252	410
S3	1.8M	266	528	136	654	502	250	395
P1	1.4M	145	247	96	429	226	80	157
P2	1.4M	157	260	108	319	231	82	159
P3	1.5M	159	266	110	322	234	82	160
HG	2.6 M	902	1640	473	2217	2148	706	1525

[a]Filter: P, Percolator; Q, Q-ranker; PP, PeptideProphet; I, IProphet; DF, DeepFilter; W*, CNN-based WinnowNet; W, self-attention-based WinnowNet.
[b]Size: the number of PSMs used for inference (precise to million).
[c]Number of parameters for the models of three deep learning architectures, i.e., DeepFilter, WinnowNet*, and WinnowNet (precise to million).
[d]Datasets: M2 and M3 indicate the two marine metaproteomes; S1-3 indicate three soil metaproteomes; P1-3 indicate the three mock communities; HG indicates a human gut metaproteome.

to develop a pre-trained scoring model capable of ranking candidate peptides directly for a given spectrum and across multiple spectra. This transition will require innovations in handling the large search spaces inherent to MS-based proteomics. Computational efficiency, particularly inference speed, will be a key challenge. For instance, while the Marine 2 dataset search time using Sipros-Ensemble (a database search engine) on a 128-core node was 27 minutes, WinnowNet required 245 minutes for rescoring. Addressing this disparity will involve optimized CPU/GPU parallel implementations to accelerate inference.

Another critical consideration is the risk of overfitting caused by peptide sequence homology. As MS datasets grow larger, homologous peptide sequences between training and inference datasets may inflate identification performance. To evaluate this risk, we removed homologous peptides from our inference datasets using BLASTP[46]. Matches with E-values below 1e-10 were classified as homologous peptides. The results of this evaluation are presented in Supplementary Tables 10–12. Overall, the homology rate across datasets was low (0.426% to 2.464%), yet removal of homologous peptides led to slight declines in identification performance. For soil metaproteomes, WinnowNet exhibited decreases of 0.04%, 0.04%, and 0.05% at the PSM, peptide, and protein levels, respectively, comparable to or slightly higher than other filtering tools. For human gut datasets, similar patterns were observed, with declines ranging from 0.05% to 0.07%. For mock communities, average decreases increased slightly as the proportion of homologous peptides grew (1.04% to 1.21% for WinnowNet, compared to 0.85% to 1.09% for other tools). Despite these small decreases, WinnowNet consistently performed better than all other filtering tools, indicating that peptide homology-induced overfitting does not compromise its performance. Moving forward, careful management of homologous peptide sequences and larger, diverse datasets will be essential to further enhance WinnowNet's generalizability.

To gain insight into the features learned by WinnowNet, we examined the attention map weights generated by the experimental and theoretical spectrum encoders. These weights were aggregated, normalized, and projected onto the spectrum representations to assess the contribution of fragment ions in distinguishing true PSMs from false ones. For this analysis, three PSMs corresponding to the same experimental spectrum were selected: a top-ranked true positive, a second-ranked false negative, and a low-ranked false negative. The visualized feature maps were shown in Supplementary Fig. 6. True positive samples (Supplementary Fig. 6(a)) exhibit more matching ions with higher attention weights compared to the other two. Notably, the y12(+1) ion exhibited significant attention in both experimental and theoretical spectra, underscoring its importance in classification. This

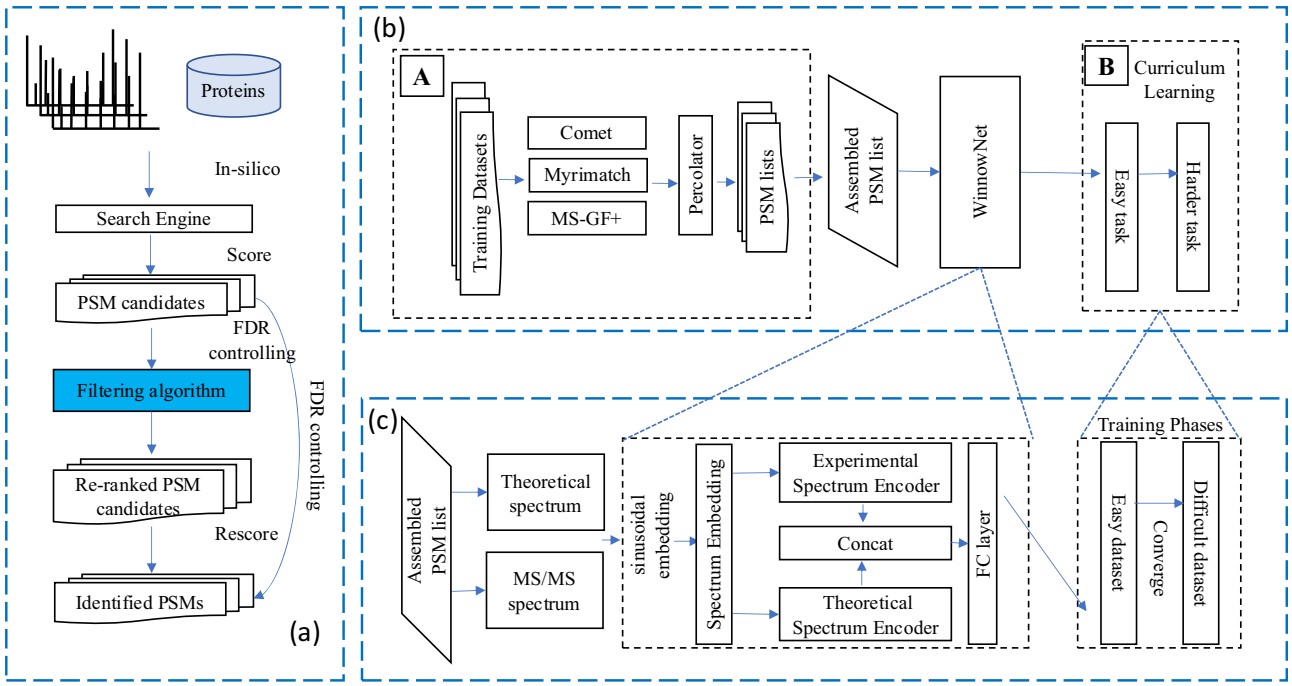

Fig. 7 | WinnowNet workflow and architecture. a Overview of peptide and protein identification using database search engines. **b** Construction of the training datasets and the training process. **c** Architecture of self-attention-based WinnowNet.

observation highlights WinnowNet's ability to successfully capture important patterns required to correlate peaks between experimental and theoretical spectra. In contrast, negative samples with high (Supplementary Fig. 6(b)) and low (Supplementary Fig. 6(c)) predictive scores show distinct patterns. Supplementary Fig. 5(b), with a high predictive score, contains approximately 15 matching ions but with relatively lower weights, while Supplementary Fig. 6(c), with a low predictive score, has fewer matching ions and lower weights, demonstrating a clear correlation between ion matching patterns and model confidence.

Methods

Our work aims to enhance and accelerate peptide identification by leveraging an intelligent learning strategy and efficient neural network architectures with reduced input dimensionality and parameter size. This section describes how these objectives were achieved by detailing the components of WinnowNet, including curriculum learning for peptide identification, the construction of training datasets, the architecture of WinnowNet, and its training procedures. Figure 7b provides an overview of the WinnowNet workflow, highlighting the training dataset construction (Part A) and the curriculum learning process (Part B). Figure 7c depicts the detailed architecture of WinnowNet.

In this study, we designed two neural network architectures: a self-attention-based model, referred to as WinnowNet, and a convolutional neural network (CNN)-based model, referred to as WinnowNet*. While the self-attention-based WinnowNet consistently demonstrated better performance generally, the CNN-based WinnowNet* is described in the Supplementary Methods for comparison. The trained WinnowNet model takes MS/MS spectra and PSM identifications reported by database search engines as input and predicts the probability that a given PSM is a true match.

Training dataset construction

Twelve datasets were used in this study, including two training datasets and nine benchmark datasets. A summary of the MS/MS spectra and protein databases for each dataset is provided in Supplementary Table 1. The ProteomeTools dataset originates from a synthetic peptide library of the human proteome provided by the ProteomeTools project, while the remaining datasets are metaproteomes from five distinct microbial communities. These include three marine microbial communities (Marine 1, Marine 2, Marine 3)[58], three soil microbial communities (Soil 1, Soil 2, Soil 3)[59], a mock microbial community (P1, P2, P3)[60], a human gut microbial community (HG)[61], and a synthetic dataset consisting of a quad-culture of four microorganisms[10]. The ProteomeTools and Marine 1 datasets were used to construct training datasets, while the remaining datasets were used for benchmarking. The raw files for the benchmarking datasets were sourced from various data repositories. Specifically, the marine and soil metaproteome datasets are available in the PRIDE repository under the identifier PXD007587, the datasets from mock community available at PXD006118, the human gut dataset can be retrieved from the iProX repository, available at IPX0001564000.

As illustrated in Fig. 7b, we constructed a high-quality training dataset by searching mass spectra from the respective datasets against their corresponding peptide libraries or protein databases (e.g., derived either from the original study of the ProteomeTools dataset or from the metagenome-assembled protein database of the Marine 1 data). This process employed three widely used database search engines: Comet[40], MyriMatch[41], and MS-GF+[42]. For each spectrum and search engine, the top five scoring PSM candidates were retained. Search results were filtered independently using Percolator to generate three lists of Percolator-scored PSMs. These lists were then merged, preserving all PSM candidates from the three search engines, resulting in up to fifteen candidates per MS/MS spectrum. In cases of duplicate PSMs, only the one with the lowest posterior error probability (PEP) assigned by Percolator[18] was retained. PSMs with PEP > 0.9 were excluded, while the top-ranked target PSMs were annotated as positive samples, and the remaining PSMs, regardless of target or decoy status, were annotated as negative samples. The ProteomeTools training dataset contained 240,000 total PSMs, including 150,785 positive samples and 89,215 negative samples. The Marine 1 training dataset contained 1,849,686 PSMs, of which 976,979 were positive samples.

Curriculum learning in WinnowNet

The data-level curriculum learning (CL) approach trains machine learning models by progressively increasing data complexity in an order similar to human curricula. CL can effectively enhance generalization ability and accelerate convergence for a wide range of applications, including image classification, object detection, scene classification, sentiment analysis, and sequence prediction[29,62–64]. The key components of CL are the difficulty measurer and the training scheduler. The former determines training sample easiness, while the latter decides the order of datasets during training based on the difficulty measurer[65,66]. For example, the learning difficulty could be measured by the noise of the data source[62], such as the images collected from Flickr (an image hosting service) were considered noisier, and thus harder, than those from Google images. The training scheduler could be used to fine-tune the model with the Flickr dataset once the model converged with the Google dataset.

The WinnowNet training was based on the CL strategy. The data difficulty level was measured by dataset complexity. Specifically, we distinguished between datasets from single-organism proteome and complex metaproteome, as well as between synthetic peptide libraries and real-world data. Two training datasets are listed in Supplementary Table 1. The easier dataset originates from the synthetic peptide library designed to cover the complete human proteome, referred as the ProteomeTools dataset. The harder dataset is from a real-world metaproteome (Marine 1) and presents noisier and more complex PSM samples. The ProteomeTools dataset is from the PRIDE Proteomics Identifications database, which includes Swiss-Prot annotated isoforms with post-translational modifications (PTMs) considered[28]. All peptides were individually synthesized with a purpose-built peptide synthesizer. The PSM quality was controlled by searching the MS/MS spectra against their synthetic peptide library with a stringent cutoff to preserve only the high-scored PSMs. Thus, it was believed that the PSMs in the ProteomeTools dataset are approximate ground-truth PSMs with relatively low noise. The ProteomeTools dataset was used to train WinnowNet as easier learning cases. Once WinnowNet converged with this dataset, we continued training it with the Marine 1 dataset. The construction of Marine 1 dataset is described in section "Benchmark datasets and evaluation metrics." The raw files are sourced from the following data repositories: the ProteomeTools dataset was retrieved from the PRIDE archive, available at PXD010595 and PXD004732; the Marine 1 dataset was retrieved from PXD007587. The details of all the datasets were described in Supplementary Note 1. The models and processed training datasets are available at https://figshare.com/articles/dataset/Models/25513531 and https://figshare.com/articles/dataset/Datasets/25511770.

Spectrum embedding

Each PSM candidate consists of a measured spectrum and a theoretical spectrum, with visualization examples provided in Supplementary Fig. 5. The theoretical spectrum was generated from the corresponding peptide sequences. To process the PSM data, we normalized the intensities and isotopic probabilities of both the experimental and theoretical spectra. These normalized tuples were then input into the spectrum embedding layer. The spectrum embedding is based on sinusoidal embedding projection, inspired by research in de novo peptide sequencing[67,68], as defined in Eq. (4), where $M/Z_{\max}$ is 7000, $M/Z_{\min}$ is 0.001, and d represents the embedding dimension (set to 256 in this study). The spectrum embedding transforms m/z values into a 256-dimensional vector, while the corresponding fragment ion intensity values are transformed through a linear layer. Finally, the transformed intensity values are concatenated with their respective m/z embeddings to form the complete representation. This spectrum encoding provides a unique representation for each peak in a spectrum, ensuring that the subsequent model captures the positional relationships among peaks. Unlike traditional approaches that rely on structured high-dimensional arrays to encode peak indices, our method leverages convolutional and self-attention mechanisms, which inherently operate independently of the input order. Specifically, the self-attention layers compute relationships between all peaks in two spectra without being constrained by their sequential arrangement, thereby achieving an order-invariant property. Here, order-invariance refers to the model's ability to identify and compare spectral features regardless of their original position in the input representation, similar to how convolutional models exhibit translation-invariance in image processing. This property is particularly advantageous for handling mass spectrometry data, where peaks do not have a strict ordering constraint, allowing for more flexible and computationally efficient spectrum analysis. This approach eliminates the need for a large matrix to represent peak indices, as required in many studies, such as DeepNovo[69] and our previous work[26].

$$\boldsymbol{Emb} = \begin{cases} \sin\left((m/z)/\left(\frac{(M/Z)_{\max}}{(M/Z)_{\min}}\left(\frac{(M/Z)_{\max}}{2\pi}\right)^{2i/d}\right)\right), & \text{for } i \le d/2 \\ \cos\left((m/z)/\left(\frac{(M/Z)_{\max}}{(M/Z)_{\min}}\left(\frac{(M/Z)_{\max}}{2\pi}\right)^{2i/d}\right)\right), & \text{for } i \ge d/2 \end{cases} \tag{4}$$

Spectrum encoders and loss function

The experimental and theoretical spectrum encoders utilize the self-attention mechanism to capture relationships between input features effectively. The process begins by embedding the fragment ions from both spectra to create spectrum embeddings. These embeddings are further contextualized using a transformer encoder comprising four self-attention layers, each with four attention heads. The outputs of the two encoders are concatenated, allowing the self-attention mechanism to model the similarity between experimental and theoretical fragment ions for each PSM. This combined representation is passed through a fully connected layer with 1024 hidden dimensions to produce the final feature vector.

The model is optimized using a loss function defined in Eq. (5), where q_i denotes the q-value of a PSM sample, and p_i represents the probability of a true PSM. To calculate the q-value of a PSM, the following steps are performed: Let f represent the score of a PSM reported by a search engine. The number of target and decoy identifications with scores better than f are denoted as t and d, respectively. The estimated false discovery rate (FDR) at a score threshold f is computed using Eq. (6), and the q-value for the PSM with score f is derived using Eq. (7).

$$Loss = -\sum [q_i \log(p_i) + q_i \log(p_i)] \tag{5}$$

$$FDR(f) = \frac{d}{t} \tag{6}$$

$$q(f) = \min_{f_i \le f}(FDR(f_i)) \tag{7}$$

WinnowNet training

WinnowNet was implemented using PyTorch version 1.4.0 and was trained on a workstation with 8 GeForce RTX 2080 Ti GPUs. The learning rate and weight decay were set to 1e-5 and the mini-batch size was set to 32.

For the easy task, we employed a ratio of 8:1:1 for training, validation, and testing using the ProteomeTools dataset. For the hard task of training on the Marine dataset, we adjusted the ratio between the training and validation datasets to 9:1. WinnowNet demonstrated convergence on the ProteomeTools dataset with training and validation accuracies of 99.32% and 99.05%, respectively. The training process for WinnowNet utilized an early stopping mechanism to prevent

overfitting and optimize model convergence. Specifically, the validation loss was monitored after each training epoch. If no improvement in validation loss was observed for 10 consecutive epochs, the training process was halted. This approach ensures that the model stops training once it reaches optimal performance on the validation set, minimizing unnecessary computation and mitigating overfitting. The maximum number of epochs was set to 200 as a safeguard, although early stopping typically occurred much earlier, as indicated by the convergence trends in Supplementary Fig. 8.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The filenames for the raw MS data and protein databases are listed in Supplementary Table 3 and are accessible via the PRIDE repository under the dataset identifiers PXD007587, PXD006118, PXD013386 [https://www.iprox.cn/page/project.html?id=IPX0001564000], PXD023217, and PXD035759. The proteomics datasets generated in this study and the protein databases with entrapment proteins have been deposited in the PRIDE repository under accession number [PXD067277] in the PRIDE database. The training datasets and learned models are available from the figshare repository at Datasets [https://figshare.com/articles/dataset/Datasets/25511770] and Models [https://figshare.com/articles/dataset/Models/25513531]. Source data are provided with this paper.

Code availability

The source code and scripts used to generate the study results are available on GitHub at https://github.com/Biocomputing-Research-Group/WinnowNet and Zenodo https://doi.org/10.5281/zenodo.16747713.

References

1. Li, Z. et al. Genome-resolved proteomic stable isotope probing of soil microbial communities using 13co2 and 13c-methanol. *Front. Microbiol.* **10**, 2706 (2019).
2. Kieft, B. et al. Phytoplankton exudates and lysates support distinct microbial consortia with specialized metabolic and ecophysiological traits. *Proc. Natl Acad. Sci.* **118**, e2101178118 (2021).
3. Kieft, B. et al. Microbial community structure–function relationships in yaquina bay estuary reveal spatially distinct carbon and nitrogen cycling capacities. *Front. Microbiol.* **9**, 1282 (2018).
4. Priya, P., Aneesh, B. & Harikrishnan, K. Genomics as a potential tool to unravel the rhizosphere microbiome interactions on plant health. *J. Microbiol Methods* **185**, 106215 (2021).
5. Mikan, M. P. et al. Metaproteomics reveal that rapid perturbations in organic matter prioritize functional restructuring over taxonomy in western arctic ocean microbiomes. *ISME J.* **14**, 39–52 (2020).
6. Thuy-Boun, P. S. et al. Quantitative metaproteomics and activity-based protein profiling of patient fecal microbiome identifies host and microbial serine-type endopeptidase activity associated with ulcerative colitis. *Mol. Cell. Proteomics* **21**, 100197 (2022).
7. Stamboulian, M., Canderan, J. & Ye, Y. Metaproteomics as a tool for studying the protein landscape of human-gut bacterial species. *PLoS Comput. Biol.* **18**, e1009397 (2022).
8. Feng, S. et al. Metalp: An integrative linear programming method for protein inference in metaproteomics. *PLOS Comput. Biol.* **18**, e1010603 (2022).
9. Shrestha, H. K. et al. Metaproteomics reveals insights into microbial structure, interactions, and dynamic regulation in defined communities as they respond to environmental disturbance. *BMC Microbiol.* **21**, 1–17 (2021).
10. Wang, D. et al. Cross-feedings, competition, and positive and negative synergies in a four-species synthetic community for anaerobic degradation of cellulose to methane. *Mbio* **14**, e03189–22 (2023).
11. Zheng, Q. et al. Metagenomic and metaproteomic insights into photoautotrophic and heterotrophic interactions in a synechococcus culture. *MBio* **11**, e03261–19 (2020).
12. Henry, C. et al. Modern metaproteomics: A unique tool to characterize the active microbiome in health and diseases, and pave the road towards new biomarkers–example of crohn's disease and ulcerative colitis flare-ups. *Cells* **11**, 1340 (2022).
13. Pan, S. & Chen, R. Metaproteomic analysis of human gut microbiome in digestive and metabolic diseases. *Adv. Clin. Chem.* **97**, 1–12 (2020).
14. Nyman, T. A., Lorey, M. B., Cypryk, W. & Matikainen, S. Mass spectrometry-based proteomic exploration of the human immune system: focus on the inflammasome, global protein secretion, and t cells. *Expert Rev. Proteom.* **14**, 395–407 (2017).
15. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. *Anal. Chem.* **74**, 5383–5392 (2002).
16. Sulimov, P. & Kertész-Farkas, A. Tailor: a nonparametric and rapid score calibration method for database search-based peptide identification in shotgun proteomics. *J. proteome Res.* **19**, 1481–1490 (2020).
17. Savitski, M. M., Mathieson, T., Becher, I. & Bantscheff, M. H-score, a mass accuracy driven rescoring approach for improved peptide identification in modification rich samples. *J. proteome Res.* **9**, 5511–5516 (2010).
18. Kall, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. methods* **4**, 923–925 (2007).
19. Liang, X., Xia, Z., Jian, L., Niu, X. & Link, A. An adaptive classification model for peptide identification. *BMC genomics* **16**, 1–9 (2015).
20. Spivak, M., Weston, J., Bottou, L., Kall, L. & Noble, W. S. Improvements to the percolator algorithm for peptide identification from shotgun proteomics data sets. *J. proteome Res.* **8**, 3737–3745 (2009).
21. Ivanov, M. V., Levitsky, L. I., Bubis, J. A. & Gorshkov, M. V. Scavager: a versatile postsearch validation algorithm for shotgun proteomics based on gradient boosting. *Proteomics* **19**, 1800280 (2019).
22. Silva, C., Bouwmeester,A.S.,,Martens,R. & Degroeve, S. Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions. *Bioinformatics* **35**, 5243–5248 (2019).
23. Shteynberg, D. et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **10**, M111.007690 (2011).
24. Ma, Z.-Q. et al. Idpicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J. Proteome Res.* **8**, 3872–3881 (2009).
25. Guo, X. et al. Sipros ensemble improves database searching and filtering for complex metaproteomics. *Bioinformatics* **34**, 795–802 (2018).
26. Feng, S., Sterzenbach, R. & Guo, X. Deep learning for peptide identification from metaproteomics datasets. *J. Proteom.* **247**, 104316 (2021).
27. Gessulat, S. et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. methods* **16**, 509–518 (2019).
28. Zolg, D. P. et al. Building proteometools based on a complete synthetic human proteome. *Nat. methods* **14**, 259–262 (2017).

29. Bengio, Y., Louradour, J., Collobert, R. & Weston, J. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48 (2009).
30. Feng, S. et al. Winnownet (2025). https://github.com/Biocomputing-Research-Group/WinnowNet.
31. Washburn, M. P., Wolters, D. & Yates, J. R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247 (2001).
32. Granholm, V., Navarro, J. F., Noble, W. S. & Käll, L. Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics. *J. Proteom.* **80**, 123–131 (2013).
33. Feng, X.-d et al. Using the entrapment sequence method as a standard to evaluate key steps of proteomics data analysis process. *BMC genomics* **18**, 1–9 (2017).
34. Lee, S., Park, H. & Kim, H. False discovery rate estimation using candidate peptides for each spectrum. *BMC Bioinforma.* **23**, 454 (2022).
35. Wen, B. et al. Assessment of false discovery rate control in tandem mass spectrometry analysis using entrapment. *bioRxiv* (2024).
36. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. methods* **4**, 207–214 (2007).
37. Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
38. He, L., Diedrich, J., Chu, Y.-Y. & Yates III, J. R. Extracting accurate precursor information for tandem mass spectra by rawconverter. *Anal. Chem.* **87**, 11361–11367 (2015).
39. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75**, 4646–4658 (2003).
40. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: an open-source ms/ms sequence database search tool. *Proteomics* **13**, 22–24 (2013).
41. Tabb, D. L., Fernando, C. G. & Chambers, M. C. Myrimatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. proteome Res.* **6**, 654–661 (2007).
42. Kim, S. & Pevzner, P. A. Ms-gf+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).
43. 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).
44. 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).
45. Strauss, M. T. et al. Alphapept: a modern and open framework for ms-based proteomics. *Nat. Commun.* **15**, 2168 (2024).
46. Camacho, C. et al. Blast+: architecture and applications. *BMC Bioinforma.* **10**, 1–9 (2009).
47. Huerta-Cepas, J. et al. Fast genome-wide functional annotation through orthology assignment by eggnog-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
48. Huerta-Cepas, J. et al. eggnog 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
49. Companys, J. et al. Gut microbiota profile and its association with clinical variables and dietary intake in overweight/obese and lean subjects: a cross-sectional study. *Nutrients* **13**, 2032 (2021).
50. Peterson, C. T. et al. Short-chain fatty acids modulate healthy gut microbiota composition and functional potential. *Curr. Microbiol.* **79**, 128 (2022).
51. Elsayed, S. & Zhang, K. Human infection caused by clostridium hathewayi. *Emerg. Infect. Dis.* **10**, 1950 (2004).
52. Hiippala, K., Kainulainen, V., Kalliomäki, M., Arkkila, P. & Satokari, R. Mucosal prevalence and interactions with the epithelium indicate commensalism of sutterella spp. *Front. Microbiol.* **7**, 1706 (2016).
53. Leonard, M. M. et al. Multi-omics analysis reveals the influence of genetic and environmental risk factors on developing gut microbiota in infants at risk of celiac disease. *Microbiome* **8**, 1–15 (2020).
54. Bautista-Gallego, J. et al. Probiotic potential of a lactobacillus rhamnosus cheese isolate and its effect on the fecal microbiota of healthy volunteers. *Food Res. Int.* **119**, 305–314 (2019).
55. Declercq, A. et al. Updated ms^{2}pip web server supports cutting-edge proteomics applications. *Nucleic Acids Res.* **51**, W338–W342 (2023).
56. Yang, K. L. et al. Msbooster: improving peptide identification rates using deep learning-based features. *Nat. Commun.* **14**, 4539 (2023).
57. Li, K., Jain, A., Malovannaya, A., Wen, B. & Zhang, B. Deeprescore: leveraging deep learning to improve peptide identification in immunopeptidomics. *Proteomics* **20**, 1900334 (2020).
58. Bryson, S. et al. Proteomic stable isotope probing reveals taxonomically distinct patterns in amino acid assimilation by coastal marine bacterioplankton. *Msystems* **1**, e00027–15 (2016).
59. Butterfield, C. N. et al. Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. *PeerJ* **4**, e2687 (2016).
60. Kleiner, M. et al. Assessing species biomass contributions in microbial communities via metaproteomics. *Nat. Commun.* **8**, 1–14 (2017).
61. Long, S. et al. Metaproteomics characterizes human gut microbiome function in colorectal cancer. *NPJ Biofilms Microbiomes* **6**, 14 (2020).
62. Chen, X. & Gupta, A. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 1431–1439 (2015).
63. Sarafianos, N., Giannakopoulos, T., Nikou, C. & Kakadiaris, I. A. Curriculum learning for multi-task classification of visual attributes. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2608–2615 (2017).
64. Cirik, V., Hovy, E. & Morency, L.-P. Visualizing and understanding curriculum learning for long short-term memory networks. arXiv preprint arXiv:1611.06204 (2016).
65. Wang, X., Chen, Y. & Zhu, W. A survey on curriculum learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 4555–4576 (2021).
66. Soviany, P., Ionescu, R. T., Rota, P. & Sebe, N. Curriculum learning: A survey. *Int. J. Computer Vis.* **130**, 1526–1565 (2022).
67. Yilmaz, M. et al. Sequence-to-sequence translation from mass spectra to peptides with a transformer model. *Nat. Commun.* **15**, 6427 (2024).
68. Jin, Z. et al. Contranovo: A contrastive learning approach to enhance de novo peptide sequencing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 144–152 (2024).
69. Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proc. Natl Acad. Sci.* **114**, 8247–8252 (2017).

Acknowledgements

This work was supported by the National Library of Medicine, the National Center for Complementary & Integrative Health, and the National Institute of General Medical Sciences of the National Institutes of Health [R15LM013460 (X.G.) and R01AT011618 (X.G. and C.P)]. The authors also acknowledge the Department of Research Computing Services at The University of North Texas for providing High-Performance Computing resources that have contributed to the research results reported within this paper, URL: https://research.unt.edu/research-services/research-computing. We thank Jiancheng Li for his assistance in verifying the codebase and reproducing the experimental results, which helped ensure the reliability of this work.

Author contributions

X.G. came up with this conceptualization. S.F., C.P., and X.G. developed the methodology and designed the experiments. S.F. implemented the algorithm. S.F., B.Z., Y.X., C.P., and X.G. curated the data, conducted formal analysis. S.F., B.Z., and X.G. worked on visualization. H.W., A.T., and X.Y. investigated past research data and verified the results. X.G. served as project administrator. X.Y., C.P., and X.G. acquired funding resources and provided supervision. S.F. wrote the original draft. S.F., B.Z., C.P., and X.G.reviewed and edited the draft.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-63977-z.

Correspondence and requests for materials should be addressed to Chongle Pan or Xuan Guo.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.