SUPPLEMENTARY NOTE: ENHANCING PEPTIDE IDENTIFICATION IN METAPROTEOMICS THROUGH CURRICULUM LEARNING AND ORDER-INVARIANT FILTERING

1 Performance measures of Prosit and Marine 2 dataset

We evaluated the performance of the Prosit dataset using 10-fold cross-validation in the easy task of curriculum learning. The mean values of various measures obtained from the 10-fold cross-validation are presented in Tab 1.

Real-world metaproteomics datasets present challenges in evaluation compared to synthetic datasets. Target PSMs can be incorrect matches, e.g., search engines will assign a random peptide to a spectrum when the true matching peptide sequence is not in the protein database. To address this issue, a stringent false discovery rate (FDR) is employed to select high-quality target PSMs. In order to provide a more comprehensive assessment of performance, the Marine 2 dataset is chosen to present the ROC curves of WinnowNet and the second-best filtering algorithm shown in Fig .1, which indicates that WinnowNet outperforms the second-best bechmarking method.

Table 1: Performance measures of Prosit dataset using 10-fold cross validation

Accurcy	Presicion	Recall	F1-Score
0.987	0.991	0.984	0.987

[1] The entires the means of the performance measures from 10-fold cross-validations

2 Result evaluation by PSM score distributions

In order to provide a clearer representation of the performance, we visualize the PSM score distributions for different datasets that were re-scored by WinnowNet and the second-best benchmarked filtering algorithm, as shown in Fig 2. We selected the Prosit dataset and Marine2 dataset to present. In Fig 2(a) and Fig 2(b), a distinct separation is observed between the target and decoy PSMs when using Percolator and WinnowNet for PSM rescoring. This strong separation is due to the fact that the Prosit dataset only retains high-quality ground-truth PSMs after processing. For the Marine 2 dataset, both algorithms present many target PSMs in the upper tail of the score distributions. However, there is a lot of overlap between target and decoy distributions of the Percolator scoring result (shown in Fig 2(c)). In contrast, the decision boundaries are more pronounced in the PSM score distributions generated by WinnowNet, e.g., in Fig 2(d), the target PSMs and decoy PSMs are almost completely separated around a score cutoff of 0.1.

3 Memory usage

We evaluated the memory usage for all benchmarking datasets and filtering algorithms, and the results are shown in Tab 2. WinnowNet, incorporating additional mass spectra information, was found to be the second-most memory-intensive algorithm, yet it presented improved memory usage compared to DeepFilter.

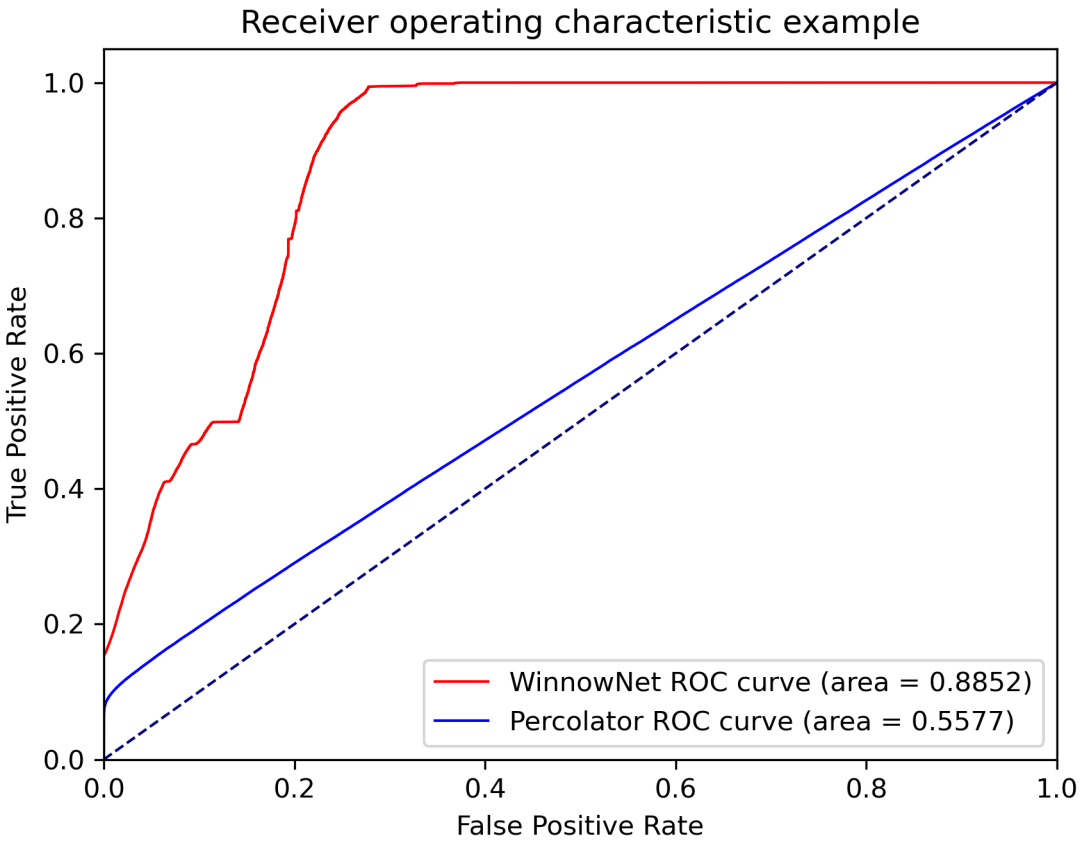


Figure 1: The ROC curves on the Marine 2 dataset

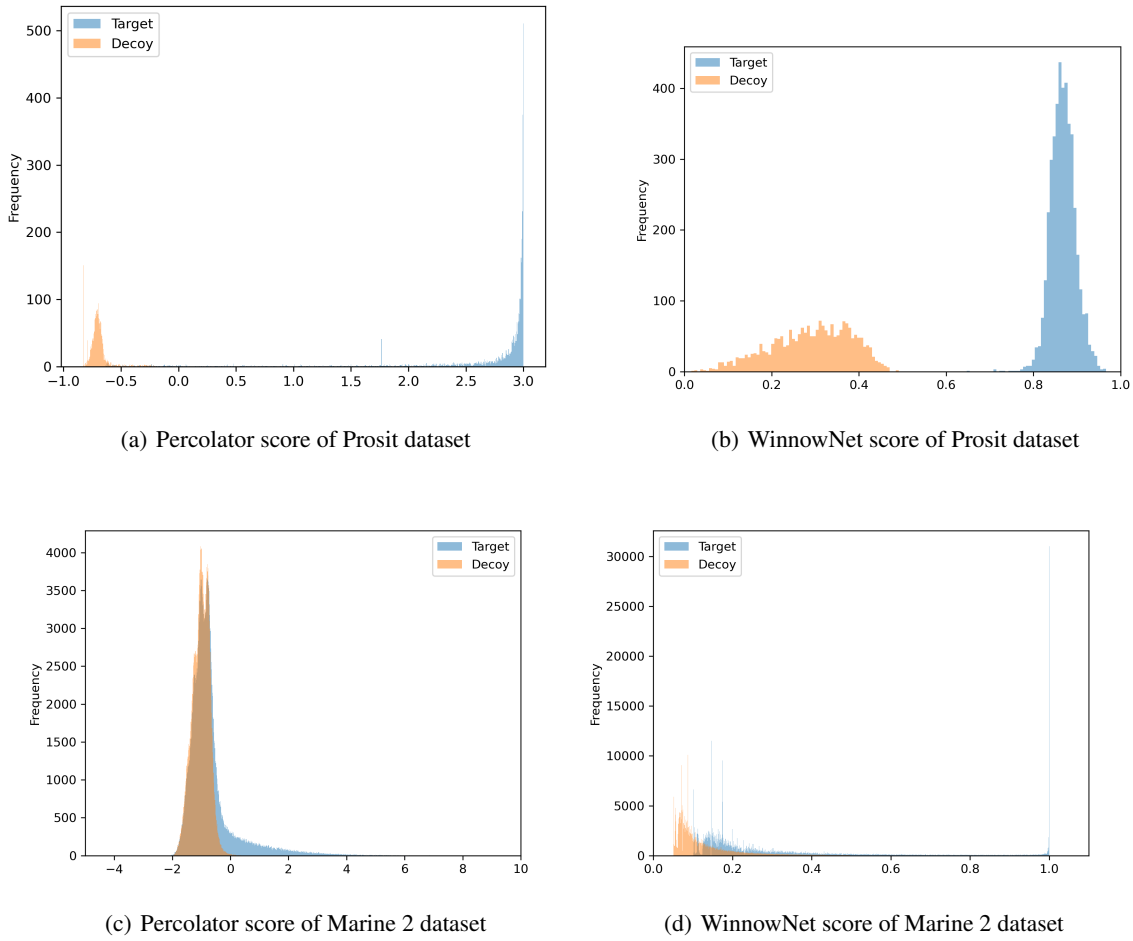
Table 2: Memory usage for benchmark datasets (precise to second).

	Filter[a]	P	Q	PP	I	DF	Win
#Params[c]		-	-	-	-	3.7M	0.82M
Datasets[d]	Size[b]						
M2	0.91M	0.51	0.89	0.21	0.42	10.2	2.8
M3	0.89M	0.50	0.87	0.21	0.41	10.1	2.8
P1	1.4M	1.11	1.70	0.62	0.93	12.59	4.27
P2	1.4M	1.10	1.66	0.62	0.93	12.55	4.25
P3	1.5M	1.12	1.68	0.65	1.02	12.81	4.52

[a] Filter: P, Percolator; Q, Q-ranker; PP, PeptideProphet; I, IProphet; DF, DeepFilter; Win, WinnowNet.
[b] Size: the number of PSMs used for inference (precise to million).
[c] Number of parameters for the models of two deep learning architectures, i.e., DeepFilter and WinnowNet (precise to million).
[d] Dataset: M2 and M3 indicate the two marine metaproteomes; P1-P3 indicate the three mock metaproteomes.

4 Performance comparison on the human gut microbial community

To evaluate WinnowNet in a more complex metaproteome, we benchmarked a dataset from the human gut microbial community [1] was benchmarked. This human gut metaproteome consisted of 670,418 spectra and was searched against a database containing 4,854,034 proteins. The identification results for the human gut microbial community at an FDR of 1% are presented in Tab. 3, demonstrating that WinnowNet outperformed other methods. In this complicated



(a) Percolator score of Prosit dataset

(b) WinnowNet score of Prosit dataset

(c) Percolator score of Marine 2 dataset

(d) WinnowNet score of Marine 2 dataset

Figure 2: PSM score distributions comparison between the Percolator and WinnowNet

metaproteome with extensive MS/MS spectra and protein databases, WinnowNet identified on average 8.5% more PSMs, 7.8% peptides, and 6.4% more proteins than the second-best filtering algorithm with our DeepFilter excluded. When compared with the Sipros-Ensemble framework shown in Tab. 4, WinnowNet using the assembled PSMs from three search engines could identify 13.7% more PSMs, 10.7% more peptides, and 8.1% more proteins. Moreover, incorporating WinnowNet with different search engines and utilizing the assembled PSMs from three search engines resulted in an average improvement of 6.9% in identified PSMs, 6.1% in identified peptides, and 4.7% in identified proteins compared to the Sipros-Ensemble search engine.

References

[1] Ngom Issa Isaac, Decloquement Philippe, Armstrong Nicholas, Didier Raoult, and Chabriere Eric. Metaproteomics of the human gut microbiota: Challenges and contributions to other omics. Clinical Mass Spectrometry, 14:18–30, 2019.

Table 3: Identification performance of the human gut metaproteome at FDR 1%

Level[a]	Search and Filter[b]	Comet	Myrimatch	MS-GF+
PSM	Without filters	206,152	246,556	236,558
	Percolator	221,274	254,212	245,824
	Q-ranker	213,012	248,634	237,584
	PeptideProphet	185,936	226,431	216,549
	IProphet	185,936	226,431	216,549
	DeepFilter	235,884	266,547	260,972
	WinnowNet	**243,131**	**272,576**	**266,408**
Peptide	Without filters	144,151	151,205	152,672
	Percolator	153,756	153,294	154,853
	Q-ranker	151,394	150,927	152,296
	PeptideProphet	133,492	139,547	140,923
	IProphet	133,492	139,547	140,923
	DeepFilter	161,327	159,693	162,562
	WinnowNet	**166,951**	**163,369**	**167,536**
Protein	Without filters	32,350	33,173	33,810
	Percolator	35,204	35,153	35,623
	Q-ranker	34,722	34,596	35,198
	PeptideProphet	30,742	31,592	32,267
	IProphet	30,742	31,592	32,267
	DeepFilter	36,124	35,901	36,346
	WinnowNet	**37,619**	**37,225**	**37,901**

[a] Level: Number of identifications at PSM/Peptide/Protein level within FDR 1%

[b] Search and Filter: Without filters represent the identification result of corresponding search engines without any filtering algorithm

[c] The best entry was bold, and the best entry with the traditional method was underlined.

Table 4: Identification performance comparison with the assembled PSMs at FDR 1%

		Level[a]		
Search[b]	Filter[c]	PSM	Peptide	Protein
SE-S	SE-F	264,753	158,865	264,753
SE-S	WinnowNet	282,939	168,564	282,939
Assemble	WinnowNet	**301,195**	**175,913**	**301,195**

[a] Level: Number of identifications at PSM/Peptide/Protein level within FDR 1%

[b] Search algorithms: SE-S, Sipros-Ensemble searching; Assembling, assembling three search engines

[c] Filtering algorithms: SE-F, Sipros-Ensemble filtering

[d] The best entry was bold, and the second best was underlined.