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. The ROC (Receiver Operating Characteristic Curve) for the test dataset is shown in Fig. 1, which shows that WinnowNet has already generalized well to the unseen dataset for the easy task and achieved the best AUC (Area under the ROC Curve)

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 .2, 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 3. We selected the Prosit dataset and Marine2 dataset to present. In Fig 3(a) and Fig 3(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 3(c)). In contrast, the decision boundaries are more pronounced in the PSM score distributions generated by WinnowNet, e.g., in Fig 3(d), the target PSMs and decoy PSMs are almost completely separated around a score cutoff of 0.1.

3 Performance comparison on mock communities

We also tested the WinnowNet using a mock community dataset to see if WinnowNet is effective for artificial microbial complexes with only a few species (30 species). We chose three "P" type communities from [1]. The "P" means the datasets have the same protein contents. Here, we labeled them as P1, P2 and P3. The identification results of three

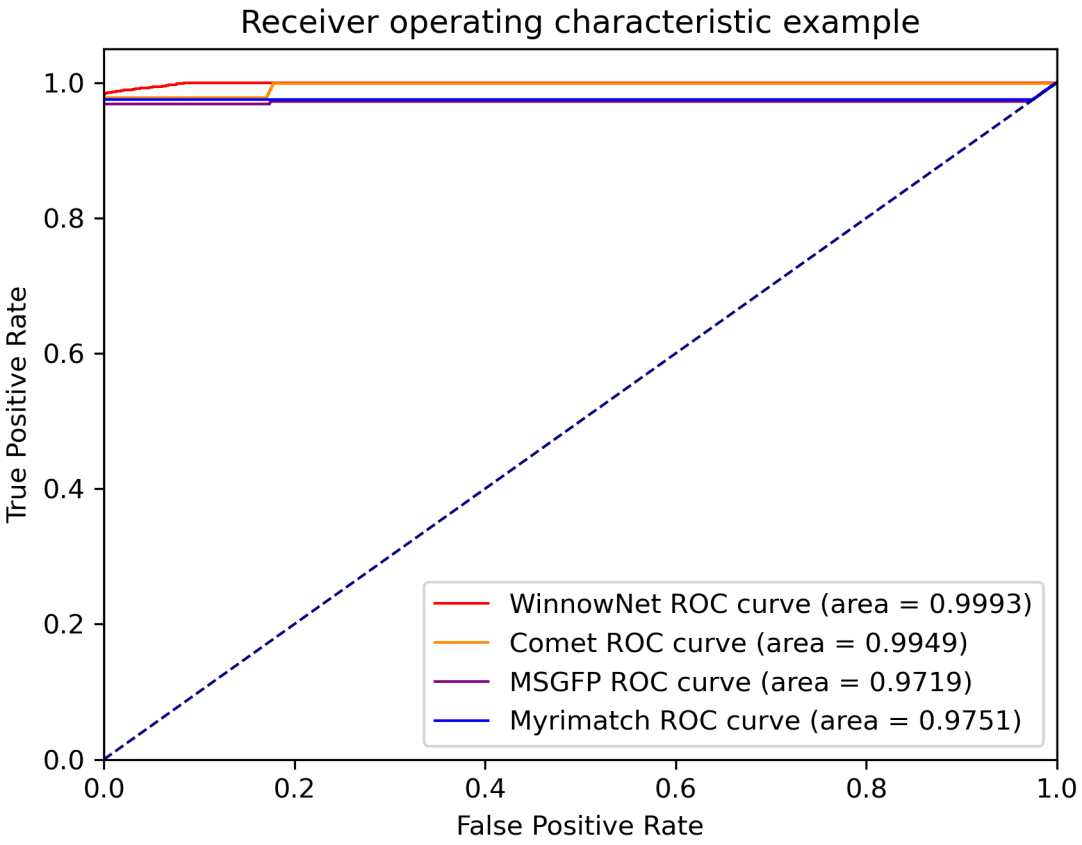


Figure 1: The ROC curves on the Marine 2 dataset

datasets are shown in Tab. 3 and Tab. 2. WinnowNet consistently achieves the highest numbers of identified PSMs, peptides, and proteins at 1% FDR. WinnowNet generated more identifications than our previous method, i.e., DeepFilter. Also, the improvements of WinnowNet over the second best filtering algorithm excluding our DeepFilter were 9.4% more PSMs, 8.1% more peptides, and 6.5% more proteins at 1% FDR on average. Compared to the Sipro-Ensemble framework (Tab. 3), WinnowNet using the assembled PSMs from three search engines could identify 7.5% more PSMs, 6.1% more peptides, and 19.2% more proteins. Additionally, when incorporating WinnowNet with different search engines, the utilization of assembled PSMs from three search engines resulted in an average improvement of 3.7% in identified PSMs, 2.9% in identified peptides, and 7.1% in identified proteins compared to the Sipros-Ensemble search engine.

4 Performance comparison on a synthetic dataset

To make the benchmarking a better simulation of real-world analysis, we combined the marine database containing ∼392,000 protein sequences and the database of the mock communities containing ∼123,100 protein sequences to generate a synthetic protein database. We searched the Marine 2 MS/MS dataset against this synthetic database with reverse decoys. The FDR was estimated as before, but we regarded identifications with peptide sequences from the marine database as true identifications and identifications with peptide sequences from the mock community database as false identifications after the FDR was controlled at 1%. The identification results from all benchmarked tools at different identification levels are shown in Tab. 4. The accuracy is defined as the ratio of the true identifications to the false identifications at 1% FDR. All the methods achieved more than 94.4% accuracy. WinnowNet obtained the largest number of true identifications among all the benchmarked filtering algorithms. Also, it yielded the second best accuracy for Comet and the highest accuracy for MyriMatch and MS-GF+ at the protein level.

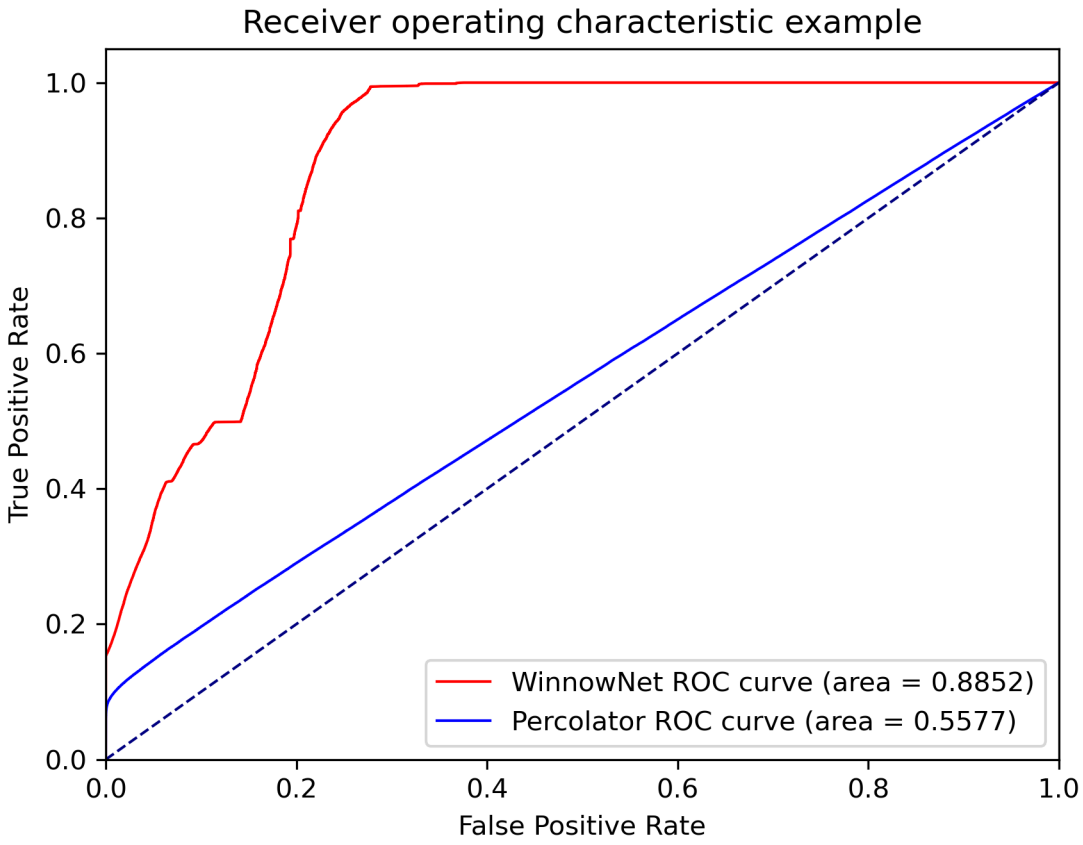
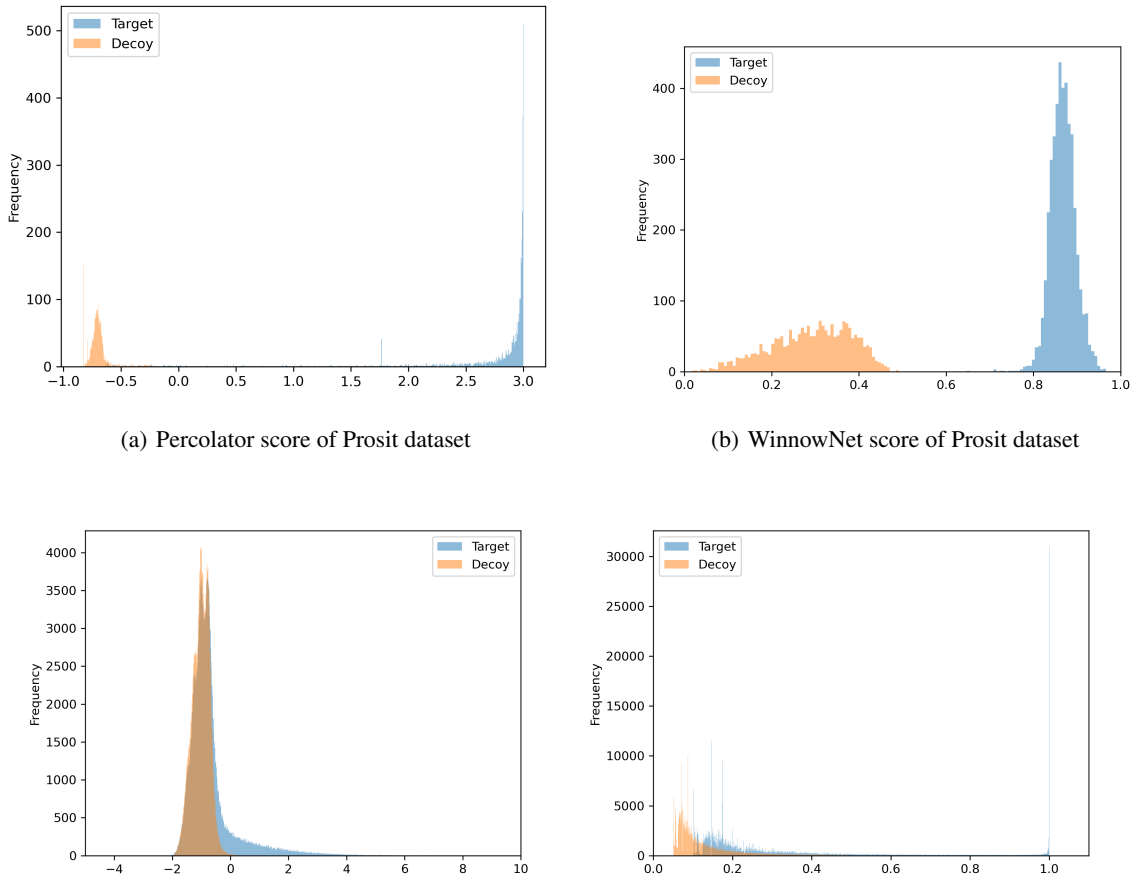


Figure 2: The ROC curves on the Marine 2 dataset

5 Analysis of the significance for the features learned by WinnowNet

To evaluate the significance of the spectrum encoder, we compared the performance of WinnowNet with the spectrum encoder disabled. We selected two metaproteome datasets, i.e., Marine 2 and P1. Fig. 4 shows the improvements of WinnowNet and its variants over the best traditional filtering algorithm at PSM/Peptide/Protein level with FDRs controlled at 1%. We can find that without the spectrum encoder, the PSM identifications of WinnowNet were still slightly better than the existing tools but dropped significantly compared to WinnowNet with the spectrum encoder enabled.

To visualize and analyze the features learned by WinnowNet, we applied the class activation mapping (CAM) ([2]) generation technique to interpret the learning decision of WinnowNet. We generated CAMs in the spectrum representation to visualize the patterns that help WinnowNet predict PSMs to be true or false. After extracting the CAMs, we tracked back to the original spectra data and reflected the weight from the CAMs to the actual spectra. We then visualized the extracted weight from CAMs of the measured and theoretical spectra for a PSM. Fig. 5 presents the CAM visualization for a target PSM and a decoy PSM. For the target PSMs, one CAM figure was from the Maine 1 model, and the other was from the Prosit Model. The right legend indicates that the red color regions make the PSMs more likely to be true (positive) PSMs if there are non-zero input values. In contrast, the blue color regions make the PSMs more likely to be false (negative) PSMs. Compared the target PSM with decoy PSM (Fig. 5(a) and Fig. 5(b)), it is obvious that there are more matching peaks in the target PSMs. Comparing the CAMs for the same PSM but generated by different models (Fig. 5(a) and Fig. 5(c)), we can observe that there are more matching peaks and their colors are more toward red after our WinnowNet was fine-tuned with the harder dataset. This demonstrates that curriculum learning did play a significant role in training a better model.



(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 3: PSM score distributions comparison between the Percolator and WinnowNet

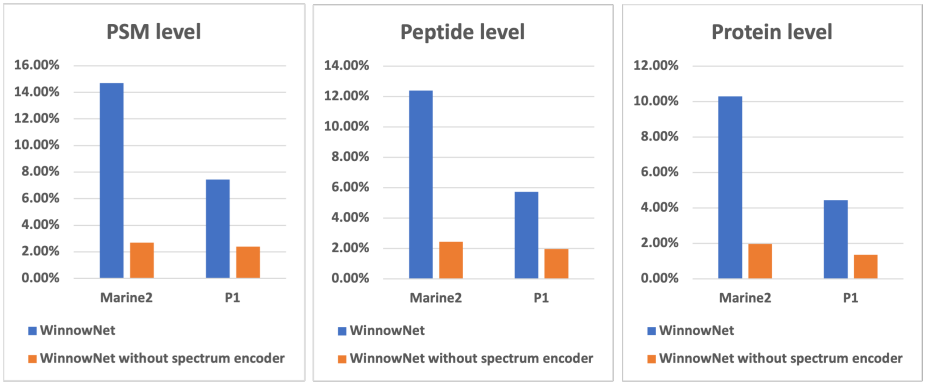
6 Memory usage

We evaluated the memory usage for all benchmarking datasets and filtering algorithms, and the results are shown in Tab 5. 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.

References

[1] Manuel Kleiner, Erin Thorson, Christine E Sharp, Xiaoli Dong, Dan Liu, Carmen Li, and Marc Strous. Assessing species biomass contributions in microbial communities via metaproteomics. Nature communications, 8(1):1–14, 2017.

[2] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2921–2929, 2016.



(a) PSM level (b) Peptide level (c) Protein level

Figure 4: Performance comparison of WinnowNet with/without the spectrum encoder.

Table 2: Identification performance of the mock communities at FDR 1%

Level[a]	Metaproteomes	P1			P2			P3		
	Search and Filter[b]	Comet	Myrimatch	MS-GF+	Comet	Myrimatch	MS-GF+	Comet	Myrimatch	MS-GF+
PSM	Without filters	94,154	84,745	99,664	102,381	97,307	109,684	80,552	74,603	89,047
	Percolator	100,559	95,666	101,977	109,920	106,684	112,041	90,777	85,356	91,900
	Q-ranker	97,342	93,127	99,233	104,591	101,036	109,848	89,342	93,127	91,233
	PeptideProphet	92,481	82,291	96,401	101,874	96,581	108,293	78,848	72,701	86,965
	IProphet	92,481	82,291	96,401	101,874	96,581	108,293	78,848	72,701	86,965
	DeepFilter	108,428	104,311	108,934	117,108	114,236	119,585	95,837	91,185	96,921
	WinnowNet	**110,258**	**107,765**	**110,696**	**119,533**	**117,783**	**120,937**	**98,418**	**94,462**	**99,188**
Peptide	Without filters	26,507	25,331	26,721	39,033	37,736	39,539	28,745	27,487	29,638
	Percolator	27,408	26,638	27,511	40,109	39,597	40,601	30,427	29,449	30,600
	Q-ranker	26,831	25,977	23,763	39,861	39,314	40,385	29,831	28,091	30,663
	PeptideProphet	25,519	24,251	25,679	38,820	37,501	39,148	27,429	26,014	27,679
	IProphet	25,519	24,251	25,679	38,820	37,501	39,148	27,429	26,014	27,679
	DeepFilter	29,012	28,841	29,139	42,319	41,372	42,866	31,769	31,026	31,901
	WinnowNet	**29,777**	**29,396**	**29,807**	**43,293**	**42,896**	**43,489**	**32,305**	**31,996**	**32,897**
Protein	Without filters	7,157	7,342	7,541	9,412	9,901	9,909	8,276	8,553	8,729
	Percolator	7,624	7,468	7,735	10,031	10,102	10,173	8,778	8,638	8,992
	Q-ranker	7,547	7,403	7,669	9,805	9,937	9,940	8,615	8,572	8,998
	PeptideProphet	6,890	7,021	7,214	9,357	9,786	9,792	8,013	8,325	8,586
	IProphet	6,890	7,021	7,214	9,357	9,786	9,792	8,013	8,325	8,586
	DeepFilter	8,017	7,961	8,065	10,413	10,484	10,601	9,113	9,014	9,288
	WinnowNet	**8,205**	**8,072**	**8,251**	**10,615**	**10,757**	**10,785**	**9,257**	**9,179**	**9,523**

[a] Level: Number of identifications at PSM/Peptide/Protein level within FDR 1%
[b] Search and Filter: "Without filters" represents the identification results of corresponding search engines without any filtering algorithm
[c] The best entry was bold, and the second best entry excluding our DeepFilter was underlined.

Table 3: Identification performance comparison of the mock communities with the assembled PSMs at FDR 1%

	Metaproteomes		P1	P2	P3
Level[a]	Search[b]	Filter[c]			
PSM	SE-S	SE-F	118,127	128,122	103,343
	SE-S	WinnowNet	122,459	132,998	107,197
	Assembled	WinnowNet	**126,668**	**137,584**	**111,727**
Peptide	SE-S	SE-F	32,937	47,331	35,414
	SE-S	WinnowNet	34,076	48,719	36,589
	Assembled	WinnowNet	**35,048**	**50,038**	**37,735**
Protein	SE-S	SE-F	7,490	10,054	8,674
	SE-S	WinnowNet	8,388	11,193	9,614
	Assembled	WinnowNet	**8,865**	**12,188**	**10,261**

[a] Level: Number of identifications at PSM/Peptide/Protein level within FDR 1%
[b] Search algorithms: SE-S, Sipros-Ensemble searching; Assembled: input PSMs assembled from three search engines
[c] Filtering algorithms: SE-F, Sipros-Ensemble filtering
[d] The best entry was bold, and the second best was underlined.

Table 4: Identification performance of the synthetic dataset at FDR 1%

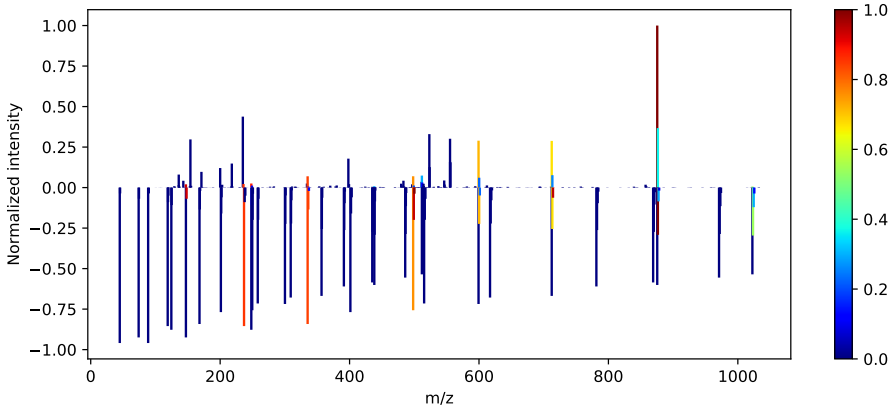
Level[a]	Search and Filter[b]	Comet			Myrimatch			MS-GF+		
		True[c]	False[d]	Acc[e]	True[c]	False[d]	Acc[e]	True[c]	False[d]	Acc[e]
PSM	Without filters	32,733	1,466	0.95713	26,379	1,573	0.94372	36,041	1,733	0.95412
	Percolator	35,492	1,568	0.95769	28,113	1,664	0.94412	38,259	1,802	0.95502
	Q-ranker	35,207	1,550	0.95783	27,530	1,626	**0.94423**	37,213	1,755	0.95496
	PeptideProphet	31,542	**1,411**	0.95718	26,026	**1,545**	0.94396	35,716	**1,681**	0.95505
	IProphet	31,903	1,441	0.95678	26,146	1,550	0.94404	35,928	1,694	0.95497
	DeepFilter	38,230	1,687	0.95774	30,589	1,811	0.94410	40,523	1,910	0.95499
	WinnowNet	**41,083**	1,802	**0.95798**	**35,310**	2,096	0.94397	**43,359**	2,035	**0.95517**
Peptide	Without filters	22,703	541	0.97673	18,663	687	0.96450	23,675	**601**	**0.97524**
	Percolator	24,916	590	0.97687	19,437	689	0.96577	25,144	645	0.97499
	Q-ranker	24,394	568	0.97725	19,340	709	0.96464	24,326	625	0.97495
	PeptideProphet	22,101	**502**	**0.97779**	18,454	**667**	0.96512	23,453	**601**	0.97501
	IProphet	22,393	523	0.97718	18,512	672	0.96497	23,507	616	0.97446
	DeepFilter	27,006	630	0.97720	21,973	769	0.96619	28,105	751	0.97397
	WinnowNet	**28,534**	672	0.97699	**24,236**	771	**0.96917**	**29,846**	797	0.97399
Protein	Without filters	7,033	150	0.97912	6,074	236	0.96260	7,413	197	0.97411
	Percolator	7,539	165	0.97858	6,373	248	0.96254	7,786	208	0.97398
	Q-ranker	7,525	158	0.97944	6,342	241	0.96339	7,786	208	0.97387
	PeptideProphet	6,942	**142**	**0.97995**	6,031	**229**	0.96342	7,376	**190**	0.97489
	IProphet	6,964	**142**	0.98002	6,031	**229**	0.96342	7,376	**190**	0.97489
	DeepFilter	8,136	169	0.97965	6,766	260	0.96299	8,290	215	0.97472
	WinnowNet	**8,378**	173	0.97977	**7,029**	262	**0.96407**	**8,662**	223	**0.97490**

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

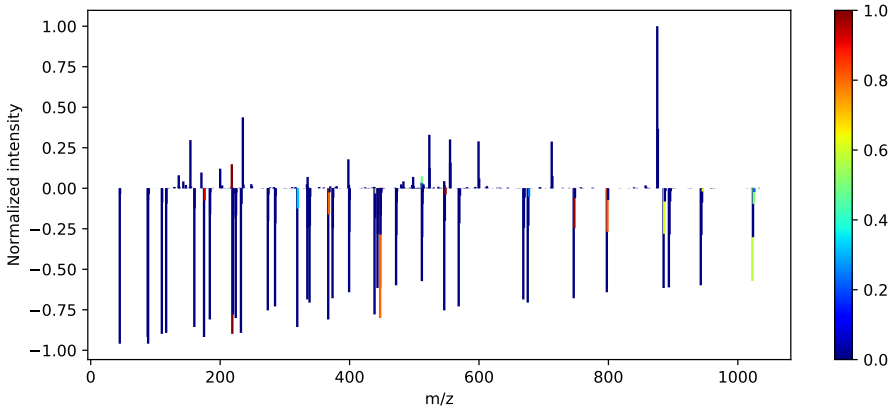
[b] Search and Filter: "Without filters" represents the identification results of corresponding search engines without any filtering algorithms

[c,d,e] True, true identifications with peptide sequences from the marine protein database; False, false identifications with peptide sequences from the mock protein proteins; Acc, accuracy defined as the ratio of the true identifications to the false identifications at 1% FDR.

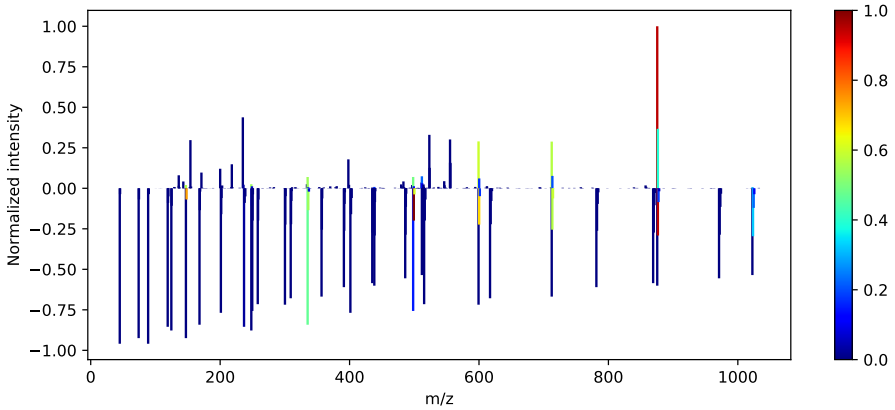
[f] The best entry was bold, and the second best was underlined.



(a) CAM for a target PSM from Marine 1 model. The peptide sequence is: SM∼YITYSTK. "∼" indicates the variable modification methionine oxidation



(b) CAM for a decoy PSM from by Marine 1 model. The peptide sequence is: SEFAEVQSGR



(c) CAM for a target PSM from Prosit model. The peptide sequence is the same as in (a), SM∼YITYSTK. "∼" indicates the variable modification (methionine oxidation)

Figure 5: Visualization of CAMs for PSMs

Table 5: 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.