

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	The data was processed with Comet(v201801.rev2), Myrimatch(v2.2.2036), MS-GF+(v2019.04.18), and post-processing with Percolator (v3.03.0) after acquisition. Training data was obtained using modified Sipros-Ensemble(v1.2) isotope extraction functions, implemented with C++ (version 11.4.0), custom Bash scripts (version 4.4.20), and Python script(version 3.7).
Data analysis	The raw mass spectrometry data was converted into ms2 format using the pre-processing tool MSconvert of ProteoWizard (release 3.0.11841). For benchmarking analysis, proteomics workflows including MaxQuant (v2.6.7), FragPipe (v22.0), PEAKS (v12.5), SpectroMine (v3), and AlphaPept (v0.5); database search engines including Comet (v201801.rev2), Myrimatch (v2.2.2036), MS-GF+ (v2019.04.18), and Sipros-Ensemble (v1.2) were utilized, along with post-processing tools such as Percolator (v3.03.0), Q-ranker (part of Crux v3.2), PeptideProphet (part of TPP v5.2.0), iProphet (part of TPP v5.2.0), and DeepFilter (v1.0). Further analysis was conducted using modified Python scripts of Sipros-Ensemble (v1.2). WinnowNet is freely available under the GNU GPL license at https://github.com/Biocomputing-Research-Group/WinnowNet

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

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 [<https://www.ebi.ac.uk/pride/archive/projects/PXD007587>], PXD006118 [<https://www.ebi.ac.uk/pride/archive/projects/PXD006118>], PXD013386 [<https://www.iprox.cn/page/project.html?id=IPX0001564000>], PXD023217 [<https://www.ebi.ac.uk/pride/archive/projects/PXD023217>], and PXD035759 [<https://www.ebi.ac.uk/pride/archive/projects/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.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

NA

Reporting on race, ethnicity, or other socially relevant groupings

NA

Population characteristics

NA

Recruitment

NA

Ethics oversight

NA

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

The number of datasets used in this study was chosen to ensure robust training and validation of our machine learning models. The ProteomeTools training dataset comprises 240,000 peptide-spectrum matches (PSMs), including 150,785 positive and 89,215 negative samples, providing a balanced and diverse set of examples. The Marine 1 training dataset contains 1,849,686 candidate PSMs, with 976,979 positive samples, offering a large volume of data for effective learning. The combined size and diversity of these datasets are sufficient to capture relevant patterns and enable the model to generalize well to unseen data, as supported by consistent validation performance metrics.

Data exclusions

No data were excluded

Replication

The performance of our models could be reproduced, and we also provide codes and tutorials for the key results. WinnowNet and the code used to reproduce the experiments in the paper are freely available under the GNU GPL license at <https://github.com/Biocomputing-Research-Group/WinnowNet>

Randomization

Samples in the ProteomeTools training dataset were allocated by selecting the peptide-spectrum matches (PSMs) with the best scores from the entire PSM result list generated by ProteomeTool study. Since this selection is based on score optimization rather than random allocation, covariates are controlled by consistently choosing the highest-quality PSMs, ensuring data reliability for the analysis.

Blinding

No, blinding is not applicable in our experiments. Blinding was not relevant to this study because all data analyses were performed using automated computational methods without subjective interpretation.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	NA
Novel plant genotypes	NA
Authentication	NA