

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 (n) 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
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ 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. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ 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 d , Pearson's r), 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

We downloaded spectra from MassBank.us (<https://mona.fiehnlab.ucdavis.edu/downloads>) and GNPS (https://gnps-external.ucsd.edu/gnpslibrary/ALL_GNPS.mgf) on March 3, 2023. Additionally, we downloaded MS/MS spectra from public repositories, including MassIVE/GNPS, MetabolomicWorkbench.org, MetaboLights, and the West Coast Metabolomics Center in December 2022. For public repositories data, over 939 million spectra were available, comprising 237,185,147 negative ESI and 701,996,947 positive ESI MS/MS spectra. To serve as library and query spectra, we randomly sampled between 100 to 1,000,000 spectra from these repositories.

Data analysis

We used version 1.0.0 of the code from <https://github.com/YuanyueLi/SpectralEntropy> for calculating the Native entropy similarity. For MatchMS, we used version 0.18.0. The code for BLINK was downloaded from <https://github.com/biorack/blink> on February 9, 2023. All the code used for benchmarking can be accessed at <https://github.com/YuanyueLi/FlashEntropySearch/tree/main/manuscript>.

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](#)

All mass spectra used were from public sources.

- (1) GNPS & MasST <https://masst.ucsd.edu/>
- (2) MetabolomicsWorkbench <https://www.metabolomicsworkbench.org/>
- (3) MetaboLights <https://www.ebi.ac.uk/metabolights/>
- (4) MassBank of North America <https://massbank.us/>

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender	n/a
Population characteristics	n/a
Recruitment	n/a
Ethics oversight	n/a

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

Life sciences study design

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

Sample size	Modern widely-used metabolism spectral libraries, such as MassBank.us and NIST, typically contain 1-2 million spectra. Some users may have their own spectral library with fewer spectra. To account for this variability, we sampled our benchmark spectra size between 100 to 1,000,000. Furthermore, since public spectral repositories like MassIVE/GNPS contain about 1 billion spectra, we also tested our methods' search time on more than 939 million spectra to simulate searching large, public spectral repositories.
Data exclusions	Spectra without any fragment ions were excluded from our analysis since they do not provide useful information for molecule identification.
Replication	As a replication, we measured the time used for searching 100 positive ESI and 100 negative ESI spectra, respectively.
Randomization	The query and library spectra were randomly selected using the Numpy function "np.random.choice".
Blinding	Blinding of results were no necessary as the results are calculated by automatic scripts.

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

Methods

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