

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
<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. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P 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 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 used serum NMR data from a total of 1842 individuals enrolled from 18 recruitment centers in the framework of three published studies. All spectra processed and used in this study are available on Figshare at https://figshare.com/s/2523b08fe8c2a23a341d . For each study, metadata and a selection of the acquired NMR spectra are available in the metabolomic repository Metabolights (https://www.ebi.ac.uk/metabolights/) with accession numbers MTBLS242, MTBLS395, and MTBLS424. NMR data were processed using Bruker Topspin 4.3.0 and metabolites assigned with Chenomx NMR suite 10.0.
Data analysis	Open source software "R" version 4.3.1. All original codes are available in the Supplementary Materials. A Rcpp/RcppArmadillo implementation of the fast correlation and the fast PLS is made available in the CRAN package "fastPLS" (https://cran.r-project.org/web/packages/fastPLS/index.html).

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](#)

Serum NMR data used for the present study are available at <https://figshare.com/s/2523b08fe8c2a23a341d>. For each study, metadata and a selection of the acquired NMR spectra are available in the metabolomic repository Metabolights (<https://www.ebi.ac.uk/metabolights/>) with accession numbers MTBLS242 (severe obesity), MTBLS395 (acute myocardial infarction), and MTBLS424 (breast cancer). Source data for figures and tables are provided with this paper as Source data files.

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	This information has not been collected in this study. NMR data were used in a fully anonymized manner, and no clinical or demographic data were collected, as they fell outside the scope of this methodological study.
Reporting on race, ethnicity, or other socially relevant groupings	This information has not been collected in this study. NMR data were used in a fully anonymized manner, and no clinical or demographic data were collected, as they fell outside the scope of this methodological study.
Population characteristics	This information has not been collected in this study. NMR data were used in a fully anonymized manner, and no clinical or demographic data were collected, as they fell outside the scope of this methodological study.
Recruitment	N.A. We reused NMR data already available in public repositories.
Ethics oversight	We reused NMR data already available in public repositories, for ethics we refer to the original publications.

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	The final dataset comprised 1753 serum spectra.
Data exclusions	Poor-quality spectra (e.g. large water signal, presence of contaminants) were manually inspected and removed, thus, the final dataset comprised 1753 serum spectra.
Replication	The dataset, comprising serum spectra of samples collected in 17 different recruitment centers, was divided into training and validation sets. To select training samples from such a large multivariate dataset the Kennard-Stone algorithm was applied. The function "kenStone" (from the R package "prospectr") was used with Euclidean distance as the chosen metric. As a result, 80% of the spectra (1200 samples) were included in the training set, while the remaining 20% (321 samples) were allocated to the validation set. Furthermore, the spectra of 233 samples from one independent recruitment center were used as the independent test set. Results were confirmed in all datasets.
Randomization	Not applicable since this is not a clinical cohort study.
Blinding	Not applicable since this is not a clinical cohort study.

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	Involvement 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	Involvement in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Not applicable

Novel plant genotypes

Not applicable

Authentication

Not applicable