

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	All data and scripts are available at https://github.com/spell00/BERNN_MSMS
Data analysis	https://github.com/spell00/BERNN_MSMS . Python packages used: numpy==1.23.5, scikit-learn==1.0.2, pandas==1.4.4, scikit-optimize==0.9.0, matplotlib==3.6.3, seaborn==0.12.2, tabulate==0.9.0, scipy==1.9.1, tqdm==4.64.1, joblib==1.0.0, ax-platform==0.2.10, torch==1.13.1, torchvision==0.14.1, tensorflow==2.11.0, psutil==5.9.4, mpmath==1.3.0, patsy==0.5.3, umap-learn==0.5.3, shapely==2.0.0, rpy2==3.5, mlflow==2.2.1. R packages used: harmony==1.2.0, sva==3.50.0, lisi== 1.0, WavelCA==1.0

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](#)

5.6 Data availability

All the data necessary to reproduce the experiments can be found in the repository https://github.com/spell00/BERNN_MSMS.

Alzheimer dataset, original data: proteomexchange PXD043216

Adenocarcinoma dataset, original data: https://github.com/dengkuistat/WaveICA_2.0/tree/master/data

Aging Mice dataset, original data: proteomexchange PXD009160

Benchmark dataset, original data: <https://doi.org/10.3929/ethz-b-000545373>

Mixed tissues dataset, original data: proteomexchange PXD015912

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

Five datasets with different levels of batch mixing heterogeneity and different numbers of batches were used to demonstrate how BERNN can apply to different scenarios.

- (1: Alzheimer) 839 samples, 889 features, 22 batches, 2 classes.
 (2: Adenocarcinoma) 642 samples, 6461 features, 3 batches, 2 classes.
 (3: Aging Mice) 372 samples, 17887 features, 7 batches, 2 classes.
 (4: Benchmark) 72 samples (1027 with replicates), 1235 features, 7 batches, 6 classes
 (5: Mixed tissues) 8 samples (1553 with replicates), 18114 features, 78 batches, 8 classes

Data exclusions

Some quality control samples not in relation with the classes to predict were not used for training

Replication	Model evaluation were performed with usual train/validation/test splits. Various samplings were performed to measure variability and ensure generalization of the models. For each dataset, each batch represent a replication of the data acquisition
Randomization	Samplings were randomized during the train/validation/test splits
Blinding	The investigators were not blinded to group allocation during data collection and analysis.

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

Antibodies

Antibodies used	NA
Validation	NA

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	NA
Authentication	NA
Mycoplasma contamination	NA
Commonly misidentified lines (See ICLAC register)	NA

Palaeontology and Archaeology

Specimen provenance	NA
Specimen deposition	NA
Dating methods	NA
☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.	
Ethics oversight	NA

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

Animals and other research organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals	NA
Wild animals	NA
Reporting on sex	NA
Field-collected samples	NA
Ethics oversight	NA

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

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	NA
Study protocol	NA
Data collection	NA
Outcomes	NA

Dual use research of concern

Policy information about [dual use research of concern](#)

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes
☒	☐ Public health
☒	☐ National security
☒	☐ Crops and/or livestock
☒	☐ Ecosystems
☒	☐ Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes
☒	☐ Demonstrate how to render a vaccine ineffective
☒	☐ Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐ Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐ Increase transmissibility of a pathogen
☒	☐ Alter the host range of a pathogen
☒	☐ Enable evasion of diagnostic/detection modalities
☒	☐ Enable the weaponization of a biological agent or toxin
☒	☐ Any other potentially harmful combination of experiments and agents

Plants

Seed stocks	NA
Novel plant genotypes	NA
Authentication	NA

ChIP-seq

Data deposition

- ☐ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- ☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links <i>May remain private before publication.</i>	NA
Files in database submission	NA
Genome browser session (e.g. UCSC)	NA

Methodology

Replicates	NA
Sequencing depth	NA
Antibodies	NA
Peak calling parameters	NA
Data quality	NA
Software	NA

Flow Cytometry

Plots

Confirm that:

- ☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☐ All plots are contour plots with outliers or pseudocolor plots.
- ☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	NA
Instrument	NA
Software	NA
Cell population abundance	NA
Gating strategy	NA

- ☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	NA
Design specifications	NA
Behavioral performance measures	NA

Acquisition

Imaging type(s)	NA
Field strength	NA
Sequence & imaging parameters	NA
Area of acquisition	NA
Diffusion MRI	☐ Used ☐ Not used

Preprocessing

Preprocessing software	NA
Normalization	NA
Normalization template	NA
Noise and artifact removal	NA
Volume censoring	NA

Statistical modeling & inference

Model type and settings	NA
Effect(s) tested	NA
Specify type of analysis:	☐ Whole brain ☐ ROI-based ☐ Both
Statistic type for inference	NA
(See Eklund et al. 2016)	
Correction	NA

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis

BERNN is based on autoencoders, Domain Adaptation Neural Networks, inverse triplet loss and variational autoencoder. These deep learning architectures take as input all the available data with batches, and handle the dimension reduction. Models are trained on the data of specific batches and tested on other batches. The models are trained to predict the available classes of the datasets proposed in the study.