nature portfolio

Corresponding author(s): Michele Vendruscolo

Last updated by author(s): Jan 17, 2021

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	The ZINC 15 database was used for initial similarity searches around the starting molecules which were carried out using RDKit (2020.09.1) on Python (3.9.7 |Anaconda)
Data analysis	The following software was used, implemented on Python (3.9.7 |Anaconda): MulticoreTSNE==0.1 fbpca==1.0 matplotlib==3.3.4 numpy==1.20.1 pandas==1.2.4 scikit-learn==0.24.1 scipy==1.6.2 seaborn==0.11.1 umap-learn==0.3.10 Note: Python 3.6.13 required for jtnnencoder functioning torch==1.10.0 rdkit==2020.09.1 scipy==1.5.2

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

A data subset and all of the code will be made publicly available in a GitHub repository.

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 maximum number of molecules within the ZINC 15 database with Tanimoto coefficient > 0.3 to the starting hit molecules were used as the molecule library (~8000 molecules). ~160 molecules from the initial docking simulation work and similarity searches comprised the initial training set.
Data exclusions	ZINC 15 database subset: 2D representations, clean reactivity, in stock, logP <=5, MW <=500 Da. These cutoffs were chosen given that these molecules are aimed at an eventual therapeutic context so low reactivity and a logP and MW that conformed to Lipinski Rules was considered of greater interest for this study. The 'in stock' cut off was to ensure the molecules could be procured without need for synthesis. The subset of top predicted molecules available from Mcule or MolPort was then purchased. No other exclusions.
Replication	All algorithmic work relying on initial random states was repeated from multiple different random states, and for experimental testing molecules were screened in duplicate at 25 uM during iterative cycles and strong hits (norm. half time > 4) were re-screened in triplicate at a concentration gradient beginning from 25 uM. Another screen was carried out for hits (norm. half time > 2) at a lower concentration (3.12 uM) to further validate compound potency. All other validatory experiments were repeated at least twice.
Randomization	During model testing on simulated data multiple different random starting samples were picked during training using pandas random sampling functions, and different random states for all relevant functions were used during training and prediction on the experimental data to provide an average output.
Blinding	It was not possible to blind this study as analytical and experimental work was carried out by the same individual, which required knowledge of the chemical matter involved. However, molecule purchase orders were purely based on the algorithm output and the algorithm settings were not altered in response to the results during the process.

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
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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