nature research

Corresponding author(s): Haiping Lu

Last updated by author(s): Dec 14, 2022

Reporting Summary

Nature Research 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 Research 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	Python 3.8, Pandas 1.2.4 and RDKit 2021.03.2
Data analysis	Python 3.8, PyTorch 1.7.1, DGL 0.7.1, DGLlife 0.2.8, Scikit-learn 1.0.2, NumPy 1.20.2, Pandas 1.2.4, RDKit 2021.03.2 and MOE 2020.09. We also make the source code of this study available at GitHub (https://github.com/peizhenbai/DrugBAN) and Zenodo (https://doi.org/10.5281/zenodo.7231657)

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 Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

The experimental data used in this work is available at our public repository https://github.com/peizhenbai/DrugBAN/tree/main/datasets. All datasets are from public resources. The BindingDB source is at https://www.bindingdb.org/bind/index.jsp. The BioSNAP source is at https://github.com/kexinhuang12345/MolTrans/tree/master/dataset/BIOSNAP/full_data. The Human source used in a previous study is at https://github.com/lifanchen-simm/transformerCPI/blob/master/Human%2CC.elegans/dataset/human_data.txt. The co-crystalized ligands from Protein Data Bank (PDB) are available at https://www.rcsb.org.

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	We studied three public datasets: BindingDB, BioSNAP and Human, with sample sizes 49k, 27k and 6.7k respectively. These sample sizes were determined by considering the following three aspects: i) we studied the development in this area and chose sample sizes to be of the same magnitude as those in most state-of-the-art works; ii) we chose highly reputable and widely used datasets that are publicly available; iii) all interaction data used was experimentally validated. In this way, our chosen sample sizes are sufficient to facilitate a fair performance evaluation against the state-of-the-art works.
Data exclusions	After datasets were chosen as described above, no data was excluded in this work.
Replication	To verify the reproducibility of our experimental findings, we conducted five independent runs in every experiment and reported the mean and standard deviation to provide quantitative assessment of the replications. The source code and data are available at our public GitHub repository for replication by other researchers.
Randomization	We allocated data samples into experimental groups (splits) randomly with two split strategies. The first split strategy was just random split, which randomly divided drug-target pairs (data samples) into training, validation, and test sets. The second split strategy was clustering-based pair split for evaluating prediction performance on out-of-distribution data. This second strategy firstly clustered original data into clusters and then randomly split the clusters into different sets. Thus, both strategies were based on random sample allocation.
Blinding	We were blinded to the group allocation during data collection and analysis. The group allocation process was performed by computer script without any manual intervention.

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