nature portfolio

Corresponding author(s): Professor Sheng Wang

Last updated by author(s): Nov 24, 2023

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 collected 8 datasets for our research: 1. CHEMBL(version 32): a large scale database of bioactive molecules with drug-like properties. (https://www.ebi.ac.uk/chembl/; 2. BindingDB(version 2022m7): a large scale database of measured binding affinities (https://www.bindingdb.org); 3. FS-MOL: a dataset for few-shot bioactivity prediction curated from CHEMBL27 (https://github.com/microsoft/FS-Mol); 4. pQSAR-ChEMBL: a dataset for bioactivity prediction curated from CHEMBL24, chemical scaffold split is supplied in this dataset (https://doi.org/10.1021/acs.jcim.9b00375); 5. KIBA : a dataset for kinase inhibitor with KIBA score (https://github.com/thinng/GraphDTA); 6. Davis: a dataset for kinase inhibitor with experimental binding affinities (https://github.com/thinng/GraphDTA); 7. FEP benchmarks: two FEP calculation benchmarks from Schrödinger and Merck (https://github.com/schrodinger/public_binding_free_energy_benchmark); 8. GDSC: a dataset for drug response of cancer cells. (https://www.cancerrxgene.org/).
Data analysis	We used the Rdkit package of version 2023.3.1 for compounds processing, and we used the open web service Proteins Plus (https://proteins.plus/) for creating the 2D interaction graph in our case study.

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 data are available on the figshare https://figshare.com/articles/dataset/ActFound_data/24452680, all codes are available in our GitHub project https://github.com/BFeng14/ActFound.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	No human object is involved
Population characteristics	No human object is involved
Recruitment	No human object is involved
Ethics oversight	No human object is involved

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	For most of our experiments, we selected 16 compounds for fine-tuning and the remaining compounds for testing when conducting few-shot experiments on each test assay. We chose this number because it's close to real-world drug discovery, as most projects have dozens of compounds with experimentally measured bioactivity in the lead optimization phase. We further compared all methods under different numbers of data for fine-tuning (ranging from 5 to 128) on some datasets and benchmarks.
Data exclusions	First, our training dataset were carefully curated from CHEMBL32 and BindingDB, we only retained assays with a required number of tested bioactivities (larger than 20) and desirable BAO (Bio-Assay-Ontology) format. We only retained bioactivies with molar, density, or dosing units. We further excluded all bioactivies that were repeatedly tested or have inaccurate value. Second, to solve the data leakage problem for cross-domain experiments between CHEMBL and Binding, we excluded all assays that involve data leakage from the test assays. Third, to solve the data leakage problem when testing on KIBA, Davis, and FEP benchmarks, we excluded all assays that involve data leakage from the training assays.
Replication	All codes and data necessary for the replication of our results are available through our GitHub project https://github.com/BFeng14/ActFound and figshare https://figshare.com/articles/dataset/ActFound_data/24452680.
Randomization	We ensured sufficient randomization when conducting few-shot experiments on each test assay. For most of our experiments, the fine-tuning and test compounds of each test assay were uniformly split at random, and the test experiments were repeated 10 times. For in-domain experiments on pQSAR-ChEMBL, the test compounds of each test assay were split by the chemical scaffold, which is a more challenging setting than the random split. To decrease the randomness of experimental results on FEP benchmarks, the test experiments were repeated 40 times.
Blinding	we ensured data blinding in all experiments. For the in-domain experiment on ChEMBL, BindingDB, FS-MOL, and pQSAR-ChEMBL datasets, we split the datasets into training, validation, and test assays, employing the optimal validation epoch for testing. For testing on other datasources including KIBA, Davis, FEP benchmarks, and GDSC, we ensure that all test assays are unseen during model training through our effort to eliminate the potential data leakage problem.

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