nature portfolio

Corresponding author(s): Bruno Correia

Last updated by author(s): Jan 16, 2024

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	Dataset processing was done in Python (v3.10.5) using RDKit (v2022.03.2) for generating molecular conformers and splitting them in fragments and linkers, scikit-learn (v1.0.1) for splitting datasets, BioPython (v1.79) for processing protein structures. MMPA-based algorithm and BRICS used for molecule fragmentation, as well as force field relaxation procedure MMFF, are components of RDKit package. Central packages used for writing DiffLinker as well as training and sampling scripts include NumPy (v1.22.3), PyTorch (v1.11.0), PyTorch Lightning (v1.6.3), WandB (v0.12.16), RDKit (v2022.03.2) and OpenBabel (v3.0.0). For sampling molecules with baseline methods, we used pre-trained models and sampling scripts available at the corresponding repositories: 3DLinker (https://github.com/YinanHuang/3DLinker), DeLinker (https://github.com/oxpig/DeLinker), DiffSBDD (https://github.com/arneschneuing/DiffSBDD), ResGen (https://github.com/HaotianZhangAI4Science/ResGen). None of these repositories provide version releases. All custom algorithms, scripts and dependencies are available on GitHub (https://github.com/igashov/DiffLinker).
Data analysis	Data analysis and vizualization was done in Python (v3.10.5) using RDKit (v2022.03.2), imageio (v2.19.2), NetworkX (v2.8.4), SciPy (v1.7.3), matplotlib (v3.5.2), seaborn (v0.11.2), and GNINA v1.0.3 (https://github.com/gnina/gnina). All custom algorithms, scripts and dependencies used for data analysis are available on GitHub (https://github.com/igashov/DiffLinker).

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 the processed datasets, as well as pre-trained models are available at Zenodo. Datasets: ZINC (https://doi.org/10.5281/zenodo.7121271), CASF (https://doi.org/10.5281/zenodo.7121264), GEOM (https://doi.org/10.5281/zenodo.7121278), Pockets (https://doi.org/10.5281/zenodo.7121280). Models: https://doi.org/10.5281/zenodo.7775568. Molecules used in ZINC dataset are available at ZINC database (https://zinc.docking.org/). Molecules used in CASF dataset were taken from the CASF-2016 benchmark package (http://www.pdbbind.org.cn/download/CASF-2016.tar.gz) of the PDBbind database (http://www.pdbbind.org.cn/). Molecules used in GEOM dataset are available at the repository of the original GEOM dataset (https://github.com/learningmatter-mit/geom). Molecules used in Pockets dataset were taken from Binding MOAD (http://www.bindingmoad.org/). Crystal structures of the Hsp90 inhibitor and initially bound fragments are available at Protein Data Bank under the access codes PDB-3HZ5 and PDB-3HZ1 respectively. Molecules inactive to Hsp90 were collected from three binding assays reported in PubChem under identifiers 754 (657 molecules), 687006 (81 molecules), and 1803875 (18 molecules). Crystal structures of the most potent IMDPH inhibitor and initially bound fragments are available at Protein Data Bank under the access codes PDB-5OU3 and PDB-5OU2 respectively. Crystal structures of JNK inhibitors with indazole and aminopyrazole scaffolds are available at Protein Data Bank under the access codes PDB-3FI3 and PDB-3FI2 respectively.

Human research participants

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

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

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 ZINC and CASF datasets, we used the same train/validation/test splits and sizes as in previous works. Sizes of our own GEOM and Pocket datasets are the maximum possible given the available data and chosen fragmentation procedures. The number of samples for evaluation was chosen according to the previous works for ZINC, CASF and GEOM datasets. For Pockets dataset we used smaller sample size due to a higher computational complexity of the conditioned model. The number of samples in case studies was chosen as a trade-off between high sample diversity and speed (to be feasible in the real-world applications).
Data exclusions	Box-and-whisker plots in Figure 2 do not include outlier data points falling beyond the interval of ±1.5xIQR. Figure 3f does not include outlier measurements (higher than 2) of Vina scores for unconditioned samples.
Replication	By design of the algorithm, it contains non-deterministic elements that can be however fixed via random seed. Weights of all the trained models along with data splits, as well as evaluation scripts for replication of the paper results are available in the DiffLinker repository.
Randomization	GEOM and Pocket datasets were randomly split in train/validation/test sets under the conditions preventing data leakage. DiffLinker by design contains non-deterministic elements.
Blinding	The data splits were performed in the blind fashion using randomized split procedure.

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