nature portfolio

Corresponding author(s): Dr Gaurav Ahuja

Last updated by author(s): 30-06-2026

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.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

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	Molecular datasets were collected from public and curated chemical resources, including ChEMBL v35, benchmark molecular property datasets, specialized external datasets, and external chemical libraries. Molecular structures were standardized as canonical SMILES before feature generation and downstream analysis. Yeast experimental data were collected under controlled laboratory conditions using Saccharomyces cerevisiae BY4741 and BY4880 strains. Growth kinetics were measured as OD600 using a BioTek Synergy HTX multi-mode reader. DNA-damage imaging was performed using a Nikon Eclipse Ci-L fluorescence microscope with phase-contrast and FITC channels for quantification of total cells and Rad52-GFP foci-positive cells.
Data analysis	Data analyses were performed using Python v3.10. Statistical analyses were performed using SciPy v1.15.3 and StatsModels v0.14.2. Machine-learning analyses used scikit-learn, XGBoost, LightGBM and CatBoost, where applicable. Molecular processing and featurization used RDKit, Mordred, ChemBERTa, GROVER, ImageMol, Signaturizer, MOPAC, PyTorch and Mamba SSM-based CDI models. CDI software was implemented using PyTorch and deployed using FastAPI, Docker, Uvicorn and NGINX. Data visualization was performed using Matplotlib and Seaborn. Exact software versions and dependencies are provided in Supplementary Table 1.

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

The molecular datasets used in this study are publicly available. The ChEMBL v35 molecular dataset used for CDI training, containing approximately 2.23 million molecules after processing, is available through Zenodo at https://zenodo.org/records/17582661.Molecular property prediction datasets used for classification and regression benchmarking are summarized in Supplementary Data 2 and 3, respectively. External benchmark and screening datasets are described in the Methods and Supplementary Data. Source data are provided with this paper.The source code version associated with this study has also been archived on Zenodo https://doi.org/10.5281/zenodo.21222967.

The CDI software is available as a Python package through PyPI at https://pypi.org/project/ChemicalDice/. The Python package and R library are available through GitHub at https://github.com/the-ahuja-lab/ChemicalDice. The trained model is available through the Hugging Face model card at https://huggingface.co/the-ahuja-lab/ChemicalDice. Documentation is available at https://the-ahuja-lab.github.io/ChemicalDice/. The source code version associated with this study has been archived on Zenodo: https://doi.org/10.5281/zenodo.21222967. CDI is free for academic institutions; a commercial license key is required for commercial use.

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	This study did not involve human participants, human data, or human biological material. Therefore, reporting on sex and gender is not relevant to this study.
Reporting on race, ethnicity, or other socially relevant groupings	This study did not involve human participants, human data, or human biological material. Therefore, reporting on race, ethnicity, or other socially relevant groupings is not relevant to this study.
Population characteristics	This study did not involve human participants or human-derived data.
Recruitment	This study did not involve recruitment of human participants.
Ethics oversight	This study did not involve human participants, human data, or human biological material; therefore, human-subject ethics approval was not required.

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	Sample sizes were not predetermined using a formal statistical power calculation. For computational analyses, sample sizes were determined by the available curated datasets and predefined benchmark libraries used in the study, including public classification and regression datasets, external chemical libraries, ChEMBL-based datasets, and fixed train/test or cross-validation splits. Robustness was assessed through repeated evaluation across datasets, model backbones, random seeds, folds, and statistical comparisons rather than by prospective power calculation. For yeast-based experimental validation, sample sizes followed standard practice for yeast stress-response and rescue assays. Hydroxyurea optimization, compound concentration screening, and rescue assays were performed using independent experimental/biological replicates. The growth assay was performed across three independent biological replicates with three technical replicates per condition. Imaging-based rescue assays were performed across independent biological replicates, and damage proportion per image was used as the unit of analysis. No formal power calculation was performed because these experiments were targeted validation assays rather than powered clinical or animal studies. The selected sample sizes were considered sufficient because treatment effects were reproducible across independent replicates/days and were evaluated using predefined statistical tests.
Data exclusions	No data were excluded from the analyses unless predefined quality-control criteria were not met. For molecular datasets, empty, invalid, non-standardizable or duplicate SMILES were removed before feature generation and analysis. For yeast experiments, exclusion criteria included technical failure, contamination, lack of measurable growth, imaging failure or inability to quantify cells/foci reliably. All exclusion criteria were

established before analysis and applied consistently across conditions.

Replication	Computational analyses were replicated across benchmark datasets, model classes, random seeds, folds and evaluation settings, including random split, scaffold split, low-data, ablation, replacement, k-nearest-neighbor and external benchmark analyses. Yeast experiments, including hydroxyurea-induced DNA-damage assays, growth-based concentration optimization and metabolite rescue assays, were independently repeated across biological replicates. Growth assays included three biological replicates with technical replicates per condition. Consistent direction of effect across independent replicates supported reproducibility of the findings.
Randomization	Formal randomization was not applied to yeast culture experiments because cultures were grown under controlled laboratory conditions and assigned uniformly to predefined treatment conditions. For computational analyses, train/test splits, cross-validation folds, random seeds and scaffold-based splits were generated programmatically according to predefined analysis protocols, minimizing subjective allocation bias.
Blinding	Blinding was not performed. Computational analyses were automated using predefined scripts, datasets, splits and statistical workflows. Experimental readouts, including OD600 growth measurements and image-based Rad52-GFP foci quantification, were objective and instrument- or assay-based, reducing potential observer bias.

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

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	No vertebrate laboratory animals were used. The study used Saccharomyces cerevisiae yeast strains BY4741 and BY4880 for growth and DNA-damage rescue assays. BY4741 was used for the yeast outgrowth assay, and BY4880, expressing the Rad52-GFP DNA-damage reporter, was used for imaging-based DNA-damage and rescue assays. Strain details and culture conditions are provided in the Methods.
Wild animals	No wild animals were used.
Reporting on sex	Not Applicable
Field-collected samples	No field-collected samples were used.
Ethics oversight	No vertebrate animals, human participants, human data, or human biological materials were used. The experimental work was limited to Saccharomyces cerevisiae laboratory strains; therefore, animal or human ethics approval was not required.

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

Plants

Seed stocks	This study did not use plants or seed stocks.
Novel plant genotypes	This study did not generate or use novel plant genotypes.
Authentication	This study did not use plant material; therefore, plant authentication is not relevant.