nature portfolio

Corresponding author(s): Patrick Schwab

Last updated by author(s): Aug 7, 2025

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	This study did not contain any primary data collection.
Data analysis	A full list of software dependencies and installation instructions is included in the accompanying source code release at https://github.com/perturblib/perturblib

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

Perturbation data used in this study is from publicly available sources, including GSE133344 (link: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE133344), Replogle et al. (Link: https://doi.org/10.25452/figshare.plus.20022944), GSE116198 (Link: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE116198), and Subramanian et al. (Link: https://clue.io/). The Optum de-identified Electronic Health Record database used to validate in silico findings in

real-world data is available for accredited researchers from Optum, Inc. but third-party restrictions apply to the availability of these data.
The data were used under license for this study with restrictions that do not allow for the data to be redistributed or made publicly available.
Data access to the Optum de-identified Electronic Health Record database may require a data sharing agreement and may incur data access fees. Source data for panels in Figures 2, 3, 4b and 6b are available with this manuscript.

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	Sex was included as a covariate in the retrospective matched cohort study to validate in-silico predicted therapeutic candidates in comparable populations. The data were collected as part of the individuals' electronic health records (EHRs) as recorded by their healthcare provider organisation.
Reporting on race, ethnicity, or other socially relevant groupings	Ethnicity was included as a covariate in the retrospective matched cohort study to validate in-silico predicted therapeutic candidates in comparable populations. The data were collected as part of the individuals' electronic health records (EHRs) as recorded by their healthcare provider organisation.
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	All available data were used. The study was re-using existing data and therefore no sample size calculations were performed.
Data exclusions	No data were excluded in training the large perturbation model. Individuals that had received other statins predicted to upregulate PKD1 were excluded in the retrospective cohort study validating the efficacy of predicted therapeutic candidates for polycystic kidney disease to ensure the control cohort did not receive another statin treatment that was predicted to influence PKD1 expression.
Replication	The respective methodology used to derive uncertainty estimates is described in the caption of the accompanying Figures.
Randomization	For the performance evaluation of LPM, data were randomly split into training, validation and test folds to estimate the generalisability of the performance metrics in held-out data. Validation fold data were used for hyperparameter optimisation of models.
Blinding	n/a - the study was conducted in-silico.

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

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.