

## 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 No software was used for data collection

Data analysis We built custom code using Python 3.8.12 and PyTorch 1.10.1 (Python implementation of PDGrapher is available at <https://github.com/ims-harvard/PDGrapher>); We used GENIE3 for Python obtained from <https://github.com/vahuyinh/GENIE3/tree/856416bacd93a76328dfdaf7df7a33070306a1ea>

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](#)

Processed data used in this paper, including the cell line gene expression dataset, protein-protein interaction network, drug targets, and disease-associated genes, are available via the project website at <https://zitniklab.hms.harvard.edu/projects/PDGrapher> or directly at <https://figshare.com/articles/dataset/>

Combinatorial\_prediction\_of\_therapeutic\_targets\_using\_a\_causally-inspired\_neural\_network/24798855. The raw protein-protein interaction network data was obtained from <https://downloads.thebiogrid.org/File/BioGRID/Release-Archive/BIOGRID-3.5.186/BIOGRID-MV-Physical-3.5.186.tab3.zip>, [https://www.science.org/doi/suppl/10.1126/science.1257601/suppl\\_file/datasets\\_sl-s4.zip](https://www.science.org/doi/suppl/10.1126/science.1257601/suppl_file/datasets_sl-s4.zip) and <http://www.interactome-atlas.org/data/HuRI.tsv> Raw gene expression datasets were obtained from <https://clue.io/releases/data-dashboard>. Disease-associated genes were obtained from COSMIC at [https://cancer.sanger.ac.uk/cell\\_lines/archivedownload#:~:text=Complete%20mutation%20data](https://cancer.sanger.ac.uk/cell_lines/archivedownload#:~:text=Complete%20mutation%20data) and <https://cancer.sanger.ac.uk/cosmic/curation>. Drug targets were extracted from DrugBank at <https://go.drugbank.com/releases/5-1-9> and a list of cancer drugs was obtained from NCI at <https://www.cancer.gov/about-cancer/treatment/types/targeted-therapies/approved-drug-list>. STRING (v12.0) was downloaded from <https://stringdb-downloads.org/download/protein.physical.links.detailed.v12.0.txt.gz>

## 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

N/A.

Reporting on race, ethnicity, or other socially relevant groupings

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](https://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

No formal statistical method was used to predetermine sample size. Instead, we included all samples from the LINCS dataset that passed a set of stringent filtering criteria designed to ensure biological relevance and data quality (see the information below). As a result, the sample size reflects the maximum number of high-quality, biologically relevant perturbations available for each model. Tables S9–S10 detail the cell line selection process, gene filtering criteria, and rationale for inclusion/exclusion. Sample sizes can be found in Tables S11–12

Data exclusions

1. Cell line filtering: Cell lines were filtered to keep those with sufficient perturbational coverage and the inclusion of healthy cell line counterparts. Figures S9 and S10 contain a description of the cell line selection criteria together with a list of cell lines with the largest number of perturbed samples and a reason for inclusion/exclusion.
2. Healthy counterpart selection is described in Tables S11 and S12 in the supplementary material.
3. Disease-associated genes: We extracted disease-associated genes from COSMIC (Accessed in September 2022) in addition to expert-curated genes available at <https://cancer.sanger.ac.uk/cosmic/curation>. Genes were represented using the HUGO Gene Nomenclature Committee ID. For each cell line in our dataset that has disease intervention data, we extracted cancer-causing mutations as the list of genes with "Verified" Mutation verification status in COSMIC and present in the list of genes curated by experts. Mapping the resulting genes to our list of genes in the PPI resulted in disease-associated genes. We excluded cell lines for which there were no disease-associated genes in COSMIC.
4. Gene matching: Treated samples were excluded if the targeted genes were not included in the protein-protein interaction (PPI) network. Genes in the gene expression dataset (LINCS) were matched to proteins in the PPI using the HUGO Gene Nomenclature Committee ID, which identified 10,716 overlapping genes.
5. Data level selection: Only level 3 gene expression data, which is quantile-normalized and can be compared across plates, was used.
6. Treatment and measurement specifics: Chemical interventions were included at all dose levels and time points.

Replication

We used 5-fold cross-validation to assess the stability of the model across different subsets of the data

Randomization

The datasets were split randomly into training and test sets to measure model performance. Additionally, a leave-cell-line-out setting was used for some tests to assess performance on unseen cell lines

Blinding

Blinding was not applicable because the study exclusively involved computational analysis of publicly available, pre-existing datasets (e.g., LINCS, COSMIC) where no new data collection or subjective labeling was performed. All labels (e.g., cell line identity, perturbation type, gene

expression readouts) were defined prior to analysis and were not influenced by the investigators, eliminating potential for observer or experimenter 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		Methods	
n/a	Involved in the study	n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies	<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines	<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology	<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants		

### Plants

Seed stocks	N/A.
Novel plant genotypes	N/A.
Authentication	N/A.