nature portfolio

Corresponding author(s): NMED-A147809, Zhongze Gu, Tao Wang, Lili Xu, Zaozao Chen

Last updated by author(s): 2025.12.11

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

Sample collection, processing, assays and imaging are processed in-house, details specified in Methods section. Zinc20 database involved in the deep learning pre-training phase, which is publicly accessible. No custom software was required for dataset acquisition. Molecular docking and virtual screening data were generated using Discovery Studio 2019 (Accelrys). Microscopy images were acquired using IX83P2ZF software accompanying the Olympus IX83 microscopes. Organoid morphology data were collected using the automated segmentation algorithm on the Avatarget high-content analysis system software AI-Driven organoid informatics system. Raw sequencing data were generated using the [NovaSeq Control Software (NVCS)] on the Illumina NovaSeq 6000 platform. Luminescence and TR-FRET data for drug sensitivity and kinase assays were collected using the software accompanying the plate reader BioTek of Synergy H1 microplate reader.

Data analysis

Data analysis involved with next generation sequencing data was processed with fastp (v0.20.0) for quality control, bwa for alignment to GRCh38/hg38, GATK Mutect2 (v2.2) for variant calling, GENCODE, dbSNP, COSMIC databases utilized for annotation step. RNA-seq processed with FASTQC (v0.11.5) for quality control, STAR (v2.5.3) for alignment, all integrated into multiple self-assembled bioinformatics pipelines executed via the Linux command line. Downstream analysis implemented using R (v4.2.2) within the Rstudio enviroment, including edgeR (v3.24.3) for differentially expressed gene detection and RPKM, with FDR and log2FC cutoffs applied. Enrichment analysis with clusterProfilter (v4.6.0) using hypergeometric test and BH adjustment. Heatmaps generated using ComplexHeatmap (v2.14.0) with Euclidean distance and Ward.D2 method, unless specified otherwise. Survival analysis using survival (v3.5.7/ v3.4.0) and survminer (v0.4.9) packages with log-rank test method, unless specified otherwise. Deep learning model implemented in PyTorch with CUDA 11.6 for molecular embedding pre-training, CUDA 11.7.1 for fine-tuning with organoids datasets.
The rest statistical analyses were performed using GraphPad Prism software (version 9; GraphPad Software, USA). For comparisons between two groups, an unpaired, two-tailed Student's t-test was applied. Continuous data were analyzed using either the Wilcoxon rank-sum test (for non-parametric distributions) or the Student's t-test (for parametric distributions), as appropriate. For comparisons across multiple groups,

one-way analysis of variance (ANOVA) was used for parametric variables, while the Kruskal-Wallis test was employed for non-parametric variables. A two-sided P-value of less than 0.05 was considered statistically significant for all tests, unless specified otherwise. All experiments under each condition were performed using biologically independent organoid samples. A biologically independent sample was defined as an organoid line derived from the same donor but established and cultured independently in separate batches and distinct wells.

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 trained deep learning model, along with the associated code for model invocation and implementation, is available on GitHub at https://github.com/js-ish/tsme. The whole-exome sequencing (WES) data from the primary tumor tissues used for generating the organoids in model training have been deposited in the National Genomics Data Center (NGDC) under BioProject accession PRJCA053170 at https://ngdc.cncb.ac.cn/?lang=zh. These data are under controlled access; non-commercial users can apply online for data access permissions. Detailed information on patients and compounds is provided in the Supplementary Tables of the 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	1. Determination: Sex was determined based on clinical records assigned at hospital admission. 2. Demographics: The study included both male and female participants. Specifically, the organoid cohort comprised 30 male and 11 female patients. The clinical validation cohort included 12 males and 12 females. 3. Analyses: Sex (biological attribute) was considered as a variable in the analysis of drug sensitivity profiles. As shown in the hierarchical clustering analysis , patient metadata, including sex, were observed to have no preferential influence on compound inhibition rates in these ex vivo models. 4. Gender: Gender (shaped by social/cultural circumstances) was not explicitly analyzed or disaggregated, as the study focused on biological drug responses in patient-derived organoids where sociocultural factors are not considered primary drivers of ex vivo sensitivity.
Reporting on race, ethnicity, or other socially relevant groupings	1. Variables Used: Race, ethnicity, and other socially constructed groupings were not explicitly analyzed as categorization variables in this study. 2. Rationale: The study was conducted at a single center (Nanjing Drum Tower Hospital, China). Given the geographical localization, the cohort was considered ethnically homogenous for the purpose of this genomic and pharmacological study. The primary focus was on biological determinants (somatic mutations and drug structures) rather than social constructs. 3. Data Source: Basic demographic data (age, sex) were obtained from hospital administrative records and medical charts by the researchers. No self-reported race/ethnicity data were collected for analysis. 4. Confounding Variables: As race and ethnicity were not variable within the single-center cohort, they were not treated as confounding variables. For other potential confounders (e.g., age, sex), we performed hierarchical clustering analysis on drug sensitivity profiles to assess their impact. The analysis indicated that drug response was primarily driven by the compound's mechanism of action and the patient's genomic profile (e.g., EGFR mutation status), rather than demographic or social characteristics.
Population characteristics	The study cohort comprised adult patients with lung cancer recruited for organoid establishment (n=75) and clinical validation (n=24). 1. Diagnosis Categories: The organoid cohort included 45 lung adenocarcinomas (ADCs), 12 squamous cell carcinomas (SCCs), 2 small cell lung carcinomas (SCLCs), 1 adenosquamous carcinoma, and 1 pleomorphic carcinoma. 2. Genotypic Information: Genomic profiling (WES) identified covariate-relevant somatic mutations, including EGFR (e.g., Exon 19 Del, L858R, T790M), TP53, which were used to group patients for drug sensitivity analysis. 3. Treatment Categories: Participants included both treatment-naive patients and those with acquired resistance to tyrosine kinase inhibitors (TKIs) (e.g., osimertinib resistance). The clinical validation cohort included patients receiving targeted therapy or chemotherapy (paclitaxel, carboplatin). 4. Demographics: Age and sex were recorded as covariates and analyzed for their influence on drug inhibition profiles.
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

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	No statistical methods were used to predetermine sample size. Sample size was determined based on the availability of high-quality surgical tissues and the success rate of organoid establishment (81.3%). Rationale for sufficiency: The final cohort of 61 organoid lines and 36 lines for comprehensive drug screening provided a sufficient dataset to train the deep learning model and achieve statistically significant correlations (Spearman's ρ>0.75) between predicted and experimental results. This sample size is comparable to or exceeds similar studies in the field of organoid pharmacogenomics.
Data exclusions	Yes, data exclusions were applied based on pre-established criteria standard in the field: Genomic Data: Germline variants (>0.1% frequency in dbSNP) and sequencing artifacts were excluded. Deep Learning: 1% of excessively long SMILES sequences were excluded to manage computational resources. Clinical Cohort: Patients without successful organoid establishment or comparable clinical drug-response data were excluded. RNA-seq: Low-expression genes were filtered prior to differential expression analysis.
Replication	All attempts at replication were successful. Technical Replication: Drug sensitivity assays were performed in triplicates, with high reproducibility confirmed (Spearman correlation > 0.970). Biological Replication: Genetic stability was verified by sequencing paired organoids at different passages (P1 vs P4), confirming that results are reproducible over time. Validation: The predictive performance of the model was successfully replicated in an independent external validation cohort.
Randomization	Randomization was not relevant to this study because it involved an observational collection of patient samples rather than a prospective randomized trial. Patients received standard-of-care treatments prescribed by physicians. Covariate Control: Although allocation was not random, potential covariates (e.g., age, sex) were analyzed using hierarchical clustering to ensure they did not introduce bias into the drug sensitivity results. For the clinical survival analysis, patients were stratified based on model predictions rather than randomized assignment.
Blinding	Blinding was applied during the model validation phase. Double-blind testing was conducted for the drug sensitivity screening of the external validation cohort (4 patients), where experimentalists were blinded to the deep learning model's predictions to prevent bias. For the retrospective analysis of the clinical cohort (24 patients), blinding to treatment allocation was not applicable as the data were collected historically; however, model predictions were generated independently of clinical outcomes.

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

Antibodies

Antibodies used	Primary antibodies used for IHC: TTF-1 (Abcam, Cat# ab227652), Cytokeratin 7 (CK7,Abcam, Cat# ab68459), Napsin A(Abcam, Cat# ab133249), Cytokeratin 5/6 (CK5/6, Proteintech, Cat# 68295-1), P63(Abcam, Cat# ab124762), Synaptophysin (Syn, Abcam, Cat# ab32127), Chromogranin A (CgA, Abcam, Cat# ab283265), and CD56 ((Abcam, Cat# ab133345). Antibodies used for IF: Anti-EGFR

(Abcam, Cat# ab52894). Secondary antibodies: Goat Anti-Rabbit IgG H&L (ABflo® 488) (Abclonal, Cat# AS053).

Validation

All antibodies were validated for specificity in the context of lung cancer histopathology. IHC validation: Antibodies were used to stain patient-derived organoids (LCOs) and matched primary tumor tissues. The staining patterns confirmed that LCOs retained the histological subtypes (ADC, SCC, SCLC) of the parental tumors . IF validation: The anti-EGFR antibody was validated by observing the specific downregulation of fluorescence signal in organoids treated with the EGFR inhibitor AZD5363 compared to controls.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)

No commercial cell lines were used. A total of 61 patient-derived lung cancer organoid (LCO) lines were established in-house from fresh surgical tumor tissues collected from patients at the Department of Thoracic Surgery, Nanjing Drum Tower Hospital. Sex of cell lines: The established lines were derived from 42male and 19 female participants. LCOs of different types were generated, including 45 lung adenocarcinomas (ADCs), 12 squamous cell carcinomas (SCCs), 2 small cell lung carcinomas (SCLCs), one adenosquamous carcinoma (ASC), and one pleomorphic carcinoma.

Authentication

The identity of the LCO lines was authenticated by comparing them with matched parental tumor tissues using multimodal methods: 1. Histopathology: H&E and IHC staining (markers: TTF-1, Napsin A, P63, CK5/6, CD56, synaptophysin, chromogranin A) confirmed that organoids retained the histological subtype of the original tumors . 2. Genomics: Whole-exome sequencing (WES) confirmed the concordance of somatic driver mutations (e.g., EGFR, TP53) between organoids and parent tissues . 3. Transcriptomics: RNA-seq analysis demonstrated high correlation in global gene expression profiles .

Mycoplasma contamination

All patient-derived organoid lines tested negative for mycoplasma contamination. Routine screening was performed using Luciferase-based assay. Additionally, cell cultures were maintained in medium supplemented with penicillin, streptomycin, and amphotericin B to prevent microbial contamination.

Commonly misidentified lines
(See ICLAC register)

No commonly misidentified cell lines (as listed by the International Cell Line Authentication Committee, ICLAC) were used in this study. All models were patient-derived organoids established specifically for this research.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration

Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.

Study protocol

Note where the full trial protocol can be accessed OR if not available, explain why.

Data collection

Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.

Outcomes

Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.

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.