nature portfolio

Corresponding author(s): Faisal Mahmood

Last updated by author(s): Aug 4, 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	The study is based on in-house pathology repositories and publicly available whole slide images (see Data availability statement). To curate the datasets that we introduce in the scope of this study, TCGA-UT-8K, TCGA-OT, TCGA-Slide-Reports, Rare-Cancers, and Rare-Cancers-Public, we used Python (3.9.16) and Pandas (2.2.3). For processing the reports in TCGA-Slide-Reports, GPT4o-mini was used.
Data analysis	We used Python (version 3.9.16) for all experiments and analyses in the study, which can be replicated using open-source libraries as outlined below. We used PyTorch (version 2.0.1, CUDA 11.8) for deep learning model training and inference. To train TITAN-V and TITAN, we modified the public implementation of iBOT (github.com/bytedance/ibot) and CoCa (github.com/mlfoundations/open_clip). We used four and eight x 80GB NVIDIA A100 GPUs configured for multi-GPU training using distributed data-parallel (DDP) for TITAN-V and TITAN training, respectively. All downstream experiments were conducted on single 24GB NVIDIA 3090 GPUs. All WSI processing was supported by OpenSlide (version 4.3.1), openslide-python (version 1.2.0), and CLAM (github.com/mahmoodlab/CLAM). We used Scikit-learn (version 1.2.2) for its implementation of K-Nearest Neighbors, and the logistic regression implementation and SimpleShot implementation provided by the LGSSL codebase (github.com/mbanani/lgssl). For survival tasks, we used scikit-survival (Version 0.23.1). Implementations of other slide encoders benchmarked in the study are found at the following links: GigaPath (github.com/prov-gigapath/prov-gigapath), PRISM (huggingface.coco/paige-ai/Prism), and CHIEF (github.com/hms-dbmi/CHIEF). For training weakly-supervised ABMIL models, we adapted the training scaffold code from the CLAM codebase (github.com/mahmoodlab/CLAM). Matplotlib (version 3.8.4) and Seaborn (version 0.13.2) were used to create plots and figures.

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

GTEx data used in pretraining can be accessed through the GTEx portal (https://www.gtexportal.org/home/). For benchmarks, TCGA and CPTAC data can be accessed through the NIH Genomic Data Commons (https://portal.gdc.cancer.gov) and the Proteomics Data Commons (https://proteomic.datacommons.cancer.gov), respectively. Coordinates and labels of the TCGA-UniformTumor-8K dataset are publicly available in the GitHub repository for our project.

All other publicly available datasets benchmarked in this work can be accessed through their respective data portals:

EBRAINS: https://doi.org/10.25493/WQ48-ZGX
DHMC-RCC: https://bmirds.github.io/KidneyCancer/
DHMC-LUAD: https://bmirds.github.io/LungCancer/
BRACS: https://www.bracs.icar.cnr.it/
PANDA: https://panda.grand-challenge.org/data/
IMP: https://rdm.inesctec.pt/dataset/nis-2023-008
BCNB: https://bupt-ai-cz.github.io/BCNB/
MUT-HET-RCC: https://aacrjournals.org/cancerres/article/82/15/2792/707325/Intratumoral-Resolution-of-Driver-Gene-Mutation

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	No covariates relating to sex or gender were collected, used or analyzed in the study.
Reporting on race, ethnicity, or other socially relevant groupings	No covariates regarding race, ethnicity, and other social groupings were collected, used or analyzed in the study.
Population characteristics	No covariates relating to population characteristics were collected, used or analyzed in the study.
Recruitment	No patient recruitment was necessary for using histology whole slide images retrospectively.
Ethics oversight	Brigham and Women's Hospital IRB committee approved the study, approval: 2020P000233.

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	With 335,645 whole slide images, our pretraining dataset is the largest currently published dataset to the best of our knowledge. We use public datasets as evaluations of our model on a diverse set of downstream tasks
Data exclusions	For pretraining data, data filtering was performed on the clinical reports for each data source individually to ensure privacy-presentation and relevance for training a pathology-specific vision language model. We processed the reports with QWEN2 to remove identifiers such as patient names, doctor names, and hospital names, complemented by regex pattern matching (e.g., "dr.", "hospital", etc.). We further processed the reports to include only relevant information corresponding to the slide under consideration and excluded mentions of morphological features that are not present.
Replication	Attempts at replication were successful for the calculation of model test results reported.
Randomization	For downstream evaluation that required creating train, validation, test splits, we either used official splits created by the original investigators of each dataset when available, or created them randomly. In general, we created random splits straitfied by class (ensuring that

the proportions of each class are similar across splits) and at the patient level if possible (ensuring that slides from the same patient are only in the same split). For TCGA, we further stratify on the slide submission site to make the splits site-preserved.

Blinding: N/A

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.*