Nucleic Acids Research, 2025, **53**, W143–W150
https://doi.org/10.1093/nar/gkaf414
Advance access publication date: 16 May 2025
Web Server issue

CellHit: a web server to predict and analyze cancer patients' drug responsiveness

Francesco Carli[1,2,3,*,†], Natalia De Oliveira Rosa[1,*,†], Simon Blotas[1], Pierluigi Di Chiaro[4], Luisa Bisceglia[1], Mariangela Morelli[5], Francesca Lessi[5], Anna Luisa Di Stefano[6], Chiara Maria Mazzanti[5], Gioacchino Natoli[4], Francesco Raimondi[1,*]

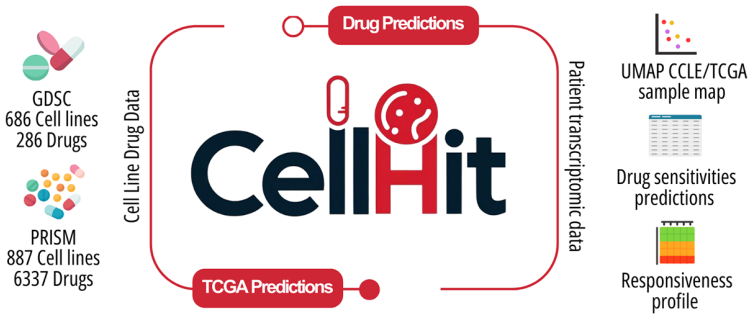
[1]Laboratorio di Biologia Bio@SNS, Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy
[2]Department of Computer Science, Univerisity of Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italy
[3]Present address: European Molecular Biology Laboratory,European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton CB10 1SD, UK
[4]Department of Experimental Oncology, IEO, European Institute of Oncology IRCCS, Via Ripamonti 435, 20141 Milano, Italy
[5]Fondazione Pisana per la Scienza, Pisa, Via F. Giovannini 13, 56017 Pisa, Italy
[6]Neurosurgical Department of Spedali Riuniti di Livorno, Via V. Alfieri 36, 57124 Livorno, Italy

[*]To whom correspondence should be addressed. Email: francesco.raimondi@sns.it
Correspondence may also be addressed to Natalia De Oliveira Rosa. Email: natalia.deoliveirarosa@sns.it
Correspondence may also be addressed to Francesco Carli. Email: francesco.carli@sns.it
[†]Joint Authors.

Abstract

We present the CellHit web server (https://cellhit.bioinfolab.sns.it/), a web-based platform designed to predict and analyze cancer patients' responsiveness to drugs using transcriptomic data. By leveraging extensive pharmacogenomics datasets from the Genomics of Drug Sensitivity in Cancer v1 and v2 (GDSC) and Profiling Relative Inhibition Simultaneously in Mixtures (PRISM) and transcriptomic data from the Cancer Cell Line Encyclopedia (CCLE) and The Cancer Genome Atlas Program (TCGA). CellHit integrates a computational pipeline for preprocessing, gene imputation, and robust alignment between patient and cell line transcriptomic data with pre-trained SOTA models for drug sensitivity prediction. The pipeline employs batch correction, enhanced Celligner methodology, and Parametric UMAP for stable and actionable alignment. The intuitive interface requires no programming expertise, offering interactive visualizations, including low-dimensional embeddings and drug sensitivity heatmaps for the input transcriptomic samples. Results feature contextual metadata, SHAP-based feature importance, and transcriptomic neighbors from reference datasets, simplifying interpretation and hypothesis generation. CellHit provides precomputed predictions across TCGA samples and offers the ability to run custom analyses online on input samples, democratizing precision oncology by enabling rapid, interpretable predictions accessible the research community.

Graphical abstract



Introduction

Multi-modal, molecular tumor profiling [1] is transforming cancer treatment by enabling the development of targeted mono- and combination therapies for genetically defined patient subgroups. However, many patients still lack effective treatments, and resistance limits therapeutic efficacy. The rise of large-scale pharmacogenomics databases, including the Cancer Cell Line Encyclopedia (CCLE) [2], Genomics of Drug Sensitivity in Cancer v1 and v2 (GDSC) [3], Cancer Therapeutics Response Portal v2 (CTRPv2) [4], Profiling Relative

Received: March 24, 2025. **Revised:** April 17, 2025. **Editorial Decision:** April 29, 2025. **Accepted:** May 2, 2025
© The Author(s) 2025. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Inhibition Simultaneously in Mixtures (PRISM) [5], and DepMap [6], has accelerated personalized oncology research. Iorio *et al.* [3] first demonstrated how genomic alterations influence drug sensitivity and how diverse genomic data types contribute to predictive modeling. Since then, multiple computational approaches have leveraged CCLE and GDSC to predict drug response in cancer cell lines (reviewed in [7–9]), with best practices for model development established [10]. Sensitivity signatures obtained from such large scale cancer cell drug screenings have also been combined with single-cell (sc) RNAseq datasets to improve drugs prioritization [11, 12].

Existing methodologies generally lack the possibility to mechanistically interpret sensitivity predictions, as well as to carry out inference on patient's samples. To address these challenges, we recently introduced a framework, i.e. CellHit [13], designed to enhance the interpretability and transferability to patients data of drug sensitivity prediction over existing baselines [14–20]. The CellHit's pipeline requires multiple steps, including data preprocessing, cell line–tumor sample alignment, model training, inference and results interpretation, which may hinder broader adoption. To overcome these limitations and democratize access for researchers lacking computational expertise, we present an intuitive web server integrating sample alignment, model inference, and interactive visualization within a unified workflow that requires no programming knowledge. This platform enables rapid exploration of drug sensitivity profiles across tumor types, facilitating the identification of potential therapeutic vulnerabilities without specialized computational training. In parallel, to develop this integrated pipeline, we created a suite of modular tools that can be independently leveraged and are released alongside the web server. This website is free and open to all users and there is no login requirement.

Materials and methods

Datasets

We obtained the data model's training from two comprehensive high-throughput studies: GDSC and PRISM. We downloaded the GDSC2 dataset, release version from 24 July 2022, from its official website (https://www.cancerrxgene.org/). We also incorporated the PRISM Repurposing Public 23Q2 dataset from the Dependency Map (DepMap, Broad Institute) portal, considering log-fold change (LFC) in cell viability measurements. We obtained RNAseq data from CCLE, which is available from DepMap. We considered log2(TPM + 1) data for protein-coding genes. We integrated RNA bulk transcriptomic data from TCGA into our pipeline. Specifically, our study employs the Tumor Compendium v11 Public PolyA dataset, released in April 2020, which amalgamates data from various publicly accessible repositories, including the TCGA and Therapeutically Applicable Research to Generate Effective Treatments (TARGET) projects. These data, sourced from the UCSC Treehouse Public Data platform, are downloaded in the RSEM log2(TPM + 1) normalized format. For the drug prediction in Pancreatic Ductal Adenocarcinoma (PDAC) subtypes (Glandular, GL; Transitional, TR; Undifferentiated, UN), data were obtained from the transcriptional profiles of multiple morphologically distinguishable tumor areas isolated by laser micro-dissection (LMD) in primary PDACs of treatment-naïve patients [21]. In conclusion, we leverage the patient-derived primary cell culture dataset of Glioblastoma Multiforme (GBM) introduced in our previous work [13].

Pre-alignment

To ensure consistency between incoming patient data and data already present in TCGA, a batch correction procedure is performed using the pyComBat package [22] (0.3.3). Notably, if the user assigns a TCGA tumor label (e.g. BRCA) to the samples in a dedicated column along with transcriptomics data, this metadata is used to achieve improved harmonization with TCGA (through the "mod" argument of pyComBat). The complete list of TCGA acronyms is available at https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations. Input data must be derived from bulk transcriptomics of patient tissue raw count format rather than from commercial cell lines. This step ensures the harmonization of new patient bulk RNA-seq data with TCGA bulk RNA-seq data.

Gene imputer

A machine learning pipeline was developed to handle missing values in incoming transcriptomic data. The imputation step leverages an XGBoost [23] (2.1.3) regressor trained on synthetically corrupted TCGA RNA-seq data. To simulate realistic missing data scenarios, a random proportion (between 10% and 20%) of the genes was randomly masked for each sample. Optimal hyperparameters for the XGBoost model were identified using Optuna [24] (4.1.0), which employed a multivariate Tree-structured Parzen Estimator (TPE) sampler [25]. Hyperparameter optimization was performed through three-fold cross-validation on random subsets of 400 genes per fold. The optimization procedure runs for a fixed budget of 300 trials. The folds were stratified by tumor type to ensure robust generalization across different cancer subtypes. The imputer receives the corrupted RNA-seq profiles along with tumor type information as input and outputs the gene expression profile suitable for downstream analysis with CellHit models [13].

Cell-line patients samples alignment

An enhanced Celligner [26] process was created to match transcriptional profiles from TCGA and CCLE. The new pipeline introduces significant changes in contrast to the prior implementation. To prevent data leakage, dataset-specific standardization statistics were computed separately for CCLE and TCGA, ensuring independent mean and standard deviation calculations before transforming new samples. These statistics are then used to process new incoming samples. The contrastive principal component analysis (cPCA) [27] procedure was also reimplemented from scratch following its original publication. Notably, Celligner original implementation relies on a mix of scikit-learn [28] and limma R [29] package functions (eBayes and F-statistics metric calculations) for differential gene expression analysis and cPCA. For better reproducibility, the entire process was rewritten in Python using PyTorch [30] (2.5.1), enabling GPU acceleration. This reimplementation also reinstates the α hyperparameter, originally fixed in Celligner. Lastly, a comprehensive hyperparameter optimization was conducted for the new alignment procedure, introducing an objective function based on a "Neighbour-

hood Consistency" filter to obtain a more robust assessment of the alignment procedure (see Supplementary Information).

Parametric UMAP

We implemented a ParametricUMAP [31] model in PyTorch to visualize user-provided transcriptomic data within a fixed reference space aligned to CCLE and TCGA datasets. Unlike conventional UMAP [32], which requires recalculating embeddings for new data, our approach enables seamless integration without recomputation. The model is a three-layer feedforward neural network that maps high-dimensional transcriptomic data to a two-dimensional embedding. This architecture maintains a stable reference space, facilitating the direct alignment of new samples.

Pipeline

The pipeline expects a CSV, ZIP, or GZ format file with bulk transcriptomic data of cancer cells, formatted as raw counts. Gene identifiers must be in either Ensembl [33] or HGNC Gene [34] Symbol format. If genes are provided in Ensembl format, they are automatically mapped to HGNC Gene Symbols using the MyGene.info (v3.2.2) Python library, which ensures consistency and simplifies downstream analyses. An optional column indicating the closest TCGA tumor category allows sample pre-alignment to TCGA data using ComBat [35] as described in the pre-alignment chapter. The server generates three outputs. The first consists of aligned RNA-seq data obtained through the pipeline described in the Cell-line patients sample alignment chapter, visualized using a parametric-UMAP low-dimensional projection that maps input samples onto the transcriptomic space of TCGA and CCLE. The second output is a tabular dataset containing sample and assay-specific predictions obtained by running pretrained models from the predictive pipeline introduced in our previous work [13]. Outputs from this computational pipeline are further processed using a post-processing pipeline that extracts a wide range of features, detailed in the "Results" section. Among these features, the post-processing pipeline identifies transcriptomic and response neighbors from CCLE and TCGA (using FAISS [36] (1.9.0post1)) and computes sample-specific SHAP importances with TreeExplainer method of the Python shap [37] (0.46.0) library. The third output is a Plotly-generated clustermap displaying drug sensitivity profiles, with samples as rows and drugs as columns. To refine the extensive drug set in GDSC and PRISM, the user can select from a dropdown menu between 15 and 50 top drugs, based on the highest variance within predicted samples. All drugs featuring an average ln(IC50) (Half-maximal inhibitory concentration) value smaller than −1 are included. Responses are computed by subtracting each drug's median IC50 (GDSC) or LFC (PRISM). Alternatively, users can standardize results per sample by subtracting the mean and dividing by the standard deviation of the predictions. The web server is implemented using ReactJS (v18.2.0) for the front end, with Plotly.js (v2.6.0) handling 2D visualizations. PrimeReact (v10.5.0) and React-Bootstrap (v2.10.0) are used for UI components. The backend is built on FastAPI (v0.104.1) to process HTTP requests and interact with a GraphQL server via Strawberry GraphQL (v0.217.1), which queries a MySQL database. Task management is executed through Celery (v5.4.0), integrated with the CellHit machine learning pipeline, while Redis (v3.5.3) is used as a task queue broker to manage concurrent processes.

Results

Using the web server

The web server's landing page offers access to two sections: one for exploring precomputed results on TCGA data for drugs included in both PRISM and GDSC (detailed in the corresponding subchapter) accessible through the "Explore now" button and another for running new analyses via the "Run Cell Hit" button. In the inference interface, users can upload RNASeq data from cell lines or patient samples by clicking "Upload Dataset." The required format includes samples as columns and genes as rows, with two mandatory columns, "TCGA_CODE" and "TISSUE," for pre-aligning data to CCLE (cell lines) or TCGA (patient samples). If no TCGA category fits, users can select "MISC." A sample file is available for reference. Users can specify whether inference should be performed on GDSC or PRISM. Those only interested in aligning data to the CCLE–TCGA reference space can enable "No Cell Hit." Upon submission, inference begins, and a notification appears when the results are ready. A graphical representation of the steps taken by the computational pipeline is reported in Fig. 1. Results can be filtered, exported, and shared via a unique URL, enhancing collaboration. The web server also includes an "About" page (explaining methodology), a "Help" page (with usage instructions), and an "FAQs" section.

Precomputed results on TCGA for both GDSC and PRISM

We provide precomputed predictions for all TCGA samples across both GDSC and PRISM drugs, totaling 4 060 342 and 17 958 038 predictions, respectively, along with a comprehensive set of contextual information. These predictions include ln(IC50) values for GDSC, LFC values for PRISM, standard deviation-based uncertainty estimates, and the Quantile Score [11]. Additionally, empirical drug statistics such as minimum, median, and maximum values are available. We include transcriptomic neighbors from CCLE and TCGA, response-matched samples, and relevant metadata, including OncoTree classifications and tissue of origin, to facilitate interpretation. Furthermore, we report the 15 most important genes identified via SHAP analysis and putative drug-associated genes where applicable.

Online aligner

Our online aligner tool extends the original Celligner tool from DepMap, incorporating methodological improvements detailed in the methods section. It maps new samples into the aligned TCGA–CCLE space, enabling users to quickly contextualize alignment results and inspect the closest transcriptomic neighbors via an interactive scatterplot. The tool supports two coloring options: by tissue and by OncoTree code, offering insights at different granularities of sample stratification (coarser for tissue and finer for OncoTree). Using ParametricUMAP, we provide a fixed reference space via a web server, ensuring consistent comparisons across alignment runs while eliminating the need for repeated dimensionality reductions. In Fig. 2, we illustrate how patient-derived primary cell culture data from a GBM dataset (diamond marks) align correctly within the neighborhood of TCGA (cross marks) and CCLE GBM samples (circular marks) after the aligner procedure.

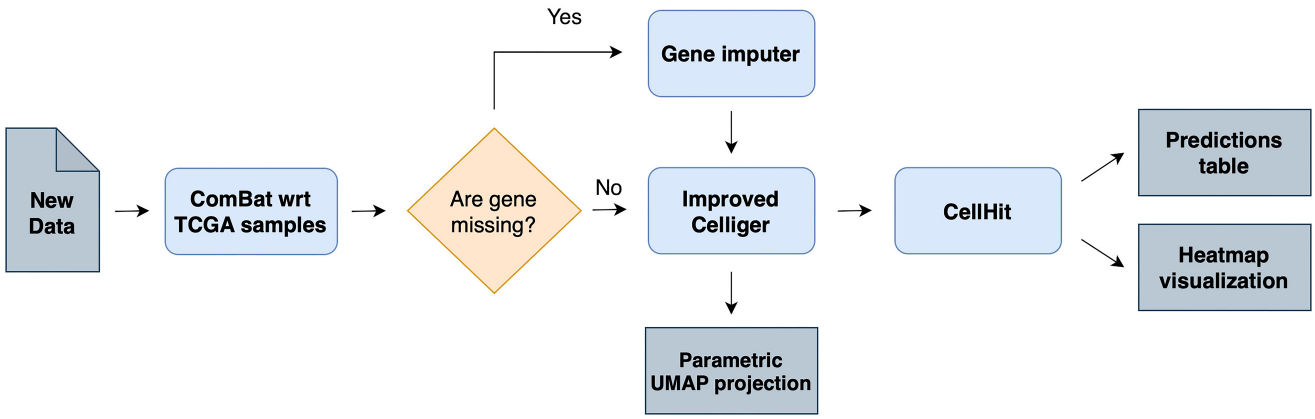


Figure 1. Overview of the computational pipeline: bulk RNA-seq data are pre-aligned to TCGA reference samples using ComBat. If necessary, missing genes are imputed prior to processing with Improved Celliger and the CellHit pipeline. The resulting outputs include a parametric UMAP projection of the alligned data, assay-specific predictions of drug sensitivity, provided as a tabular dataset, and a clustermap visualization highlighting drug response profiles across samples and drugs.

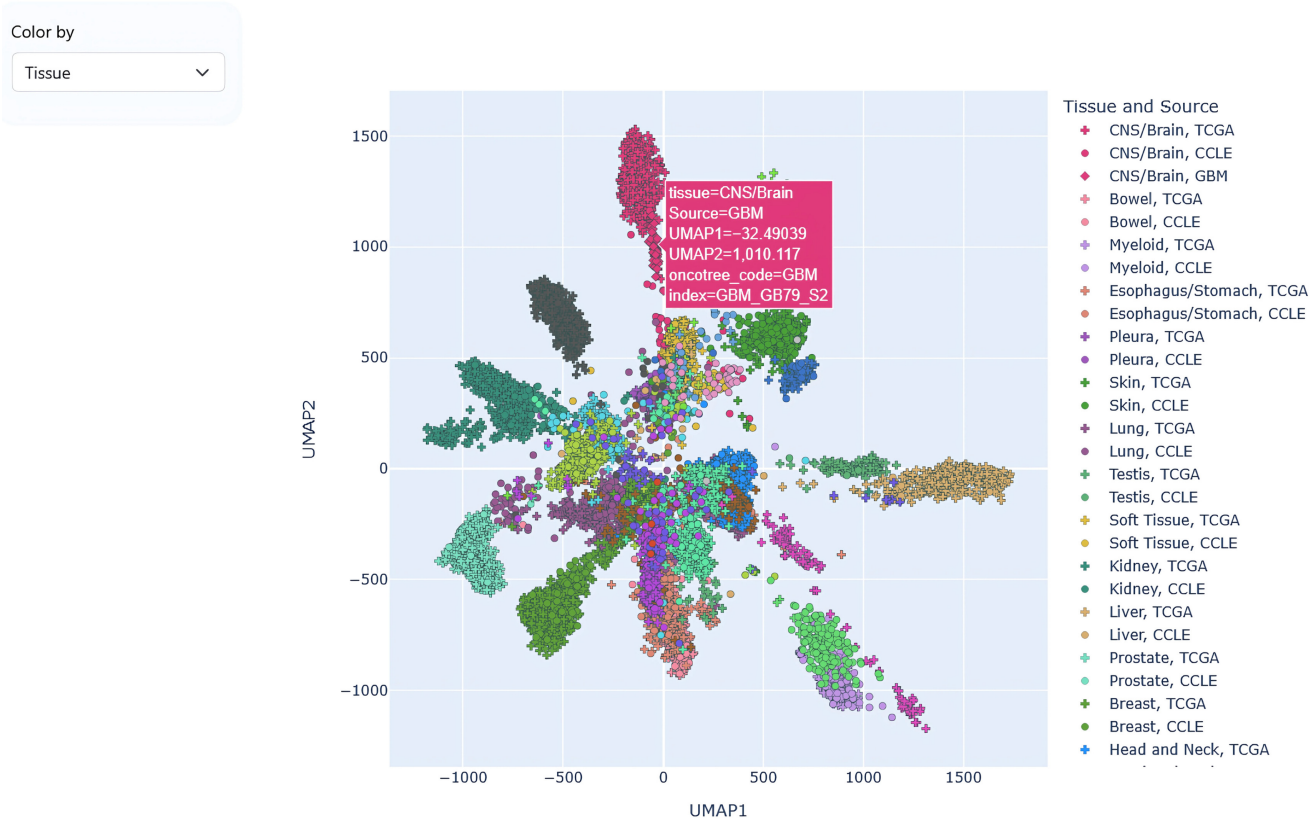


Figure 2. ParametricUMAP alignment embedding demonstrating the accurate positioning of patient-derived primary GBM cell culture samples (diamonds) within the transcriptomic neighborhood of reference glioblastoma (GBM) samples from TCGA (crosses) and CCLE (circles). Colors represent tissue origins, facilitating intuitive interpretation of alignment quality.

Prediction interface

Through the predictive interface, the user can run the entire predictive pipeline and compute on the fly the same feature already introduced in the "Precomputed" results on TCGA for both GDSC and PRISM chapter. Some columns are hidden to provide a more streamlined output but can be re-enabled to access a richer feature space. Additionally, by clicking on a sample-specific prediction (Fig. 3A), the user can toggle additional information: a graph of the top 15 genes ranked by absolute SHAP importance (Fig. 3B). We recall that, SHAP values sum up to the model's final prediction (in this case, the predicted IC50). Therefore, positive SHAP values (represented by red bars in our figures) contribute towards a higher predicted IC50 value, indicating increased resistance to the drug. Conversely, negative SHAP values (blue bars) contribute towards a lower IC50 and thus higher sensitivity. Furthermore, we visualize two orthogonal distributional features of the predictions leveraging Kernel Density Estimation (KDE) plots. The

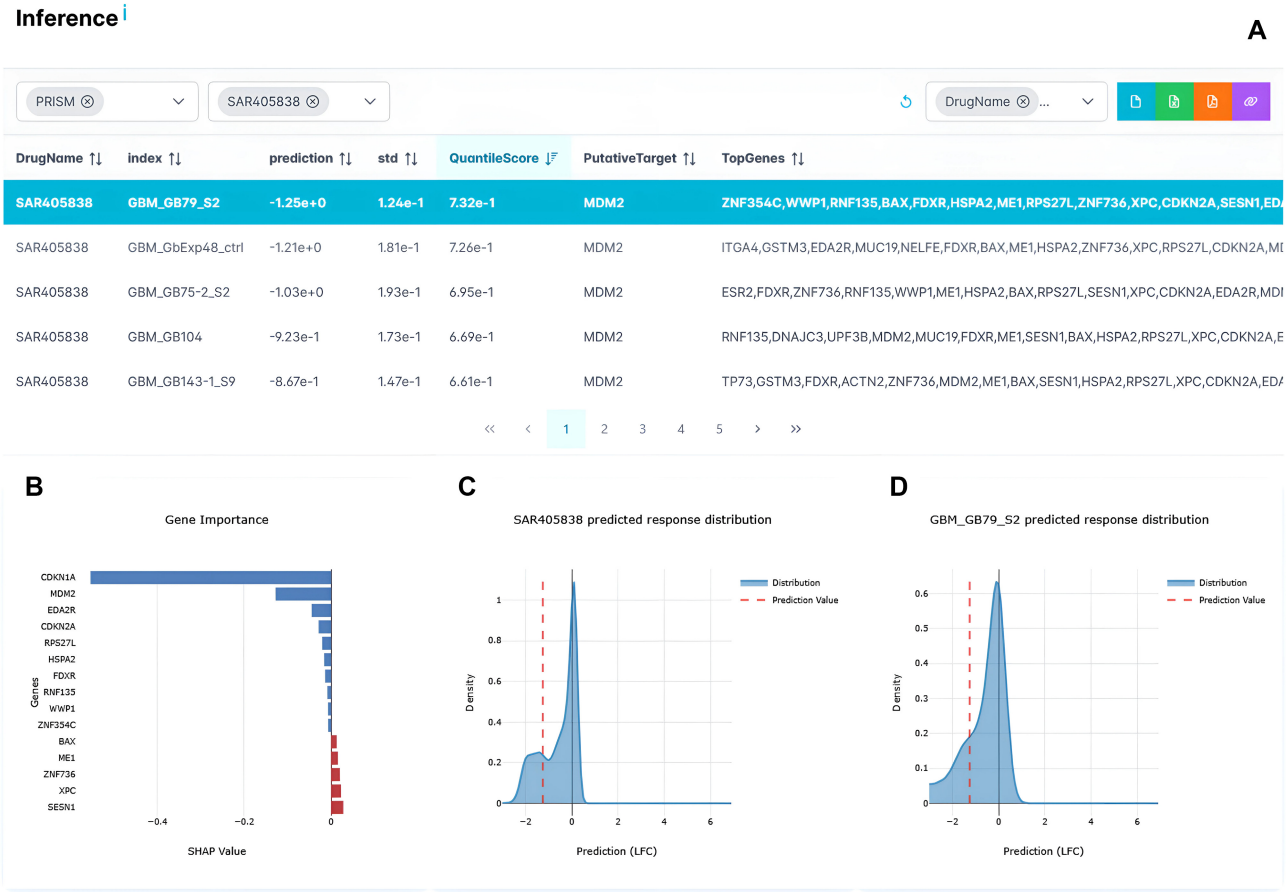


Figure 3. Predictive interface showcasing SAR405838 sensitivity prediction for a GBM sample obtained after clicking on a specific prediction. (**A**) Interactive prediction summary with putative target and key genes. (**B**) SHAP-based gene importance highlighting MDM2 relevance. (**C** and **D**) KDE plots illustrating drug efficacy relative to other drugs on the same cell line (**C**) and SAR405838 across multiple cell lines (**D**), indicating enhanced sensitivity for this specific GBM sample.

first plot (Fig. 3C) compares a drug's predicted response to other drugs tested on the same cell line, where a red line on the left side of the distribution indicates higher efficacy. The second (Fig. 3D) examines the drug's behavior across multiple cell lines, helping identify cases where consistently low IC50 values suggest general toxicity rather than selective effectiveness. Figure 3B–D presents an example of SAR405838 drug prediction on a GBM sample. The model correctly identifies MDM2, the drug's putative target, as one of the most important genes. Additionally, SAR405838 is predicted to perform better than the average drug for this sample (lower IC50, Fig. 3D). Furthermore, the predicted IC50 for this sample is lower than typical values for SAR405838 across other samples, suggesting increased sensitivity to the drug (Fig. 3C).

Heatmap visualization of sample responsiveness

The cluster map in Fig. 4 provides an overview of drug response predictions under different scaling methods. Users can adjust the number of top columns displayed using the "Top" dropdown. The "Scaling" option offers two normalization methods: "Median" Subtraction, where predictions are adjusted by subtracting the median computed on either GDSC or PRISM data. This highlights whether a sample is predicted to have higher sensitivity (blue, below median) or lower sensitivity (red, above median), helping to assess drug specificity. "Standardization," which scales values using the mean and standard deviation computed on the processed sample. This scaling emphasizes relative differences across given samples, making it useful for identifying potential tumor subtypes and precision medicine opportunities. The response information, in conjunction with information coming from the hierarchical clustering of rows, further simplifies grouping patients with similar response patterns. In Fig. 4, we show an example of such an interface on PDAC patients. We highlight the upper red box, enriched with TR-type samples, which are expected to exhibit greater resistance compared to those from the GL group.

Discussion

This work presents a user-friendly web platform for cancer drug response modeling based on transcriptomics data, designed to be accessible to users with minimal programming expertise. The platform simplifies transcriptomic-based drug sensitivity predictions by integrating data alignment, model inference, and interactive visualizations, facilitating broader adoption. Beyond raw predictions, the system provides interpretability and contextualization of the results. Outputs include SHAP values to identify transcriptional drivers of sensitivity and distribution-based visualizations to compare predicted responses across drugs and patient samples. These insights might help researchers refine hypotheses on poten-

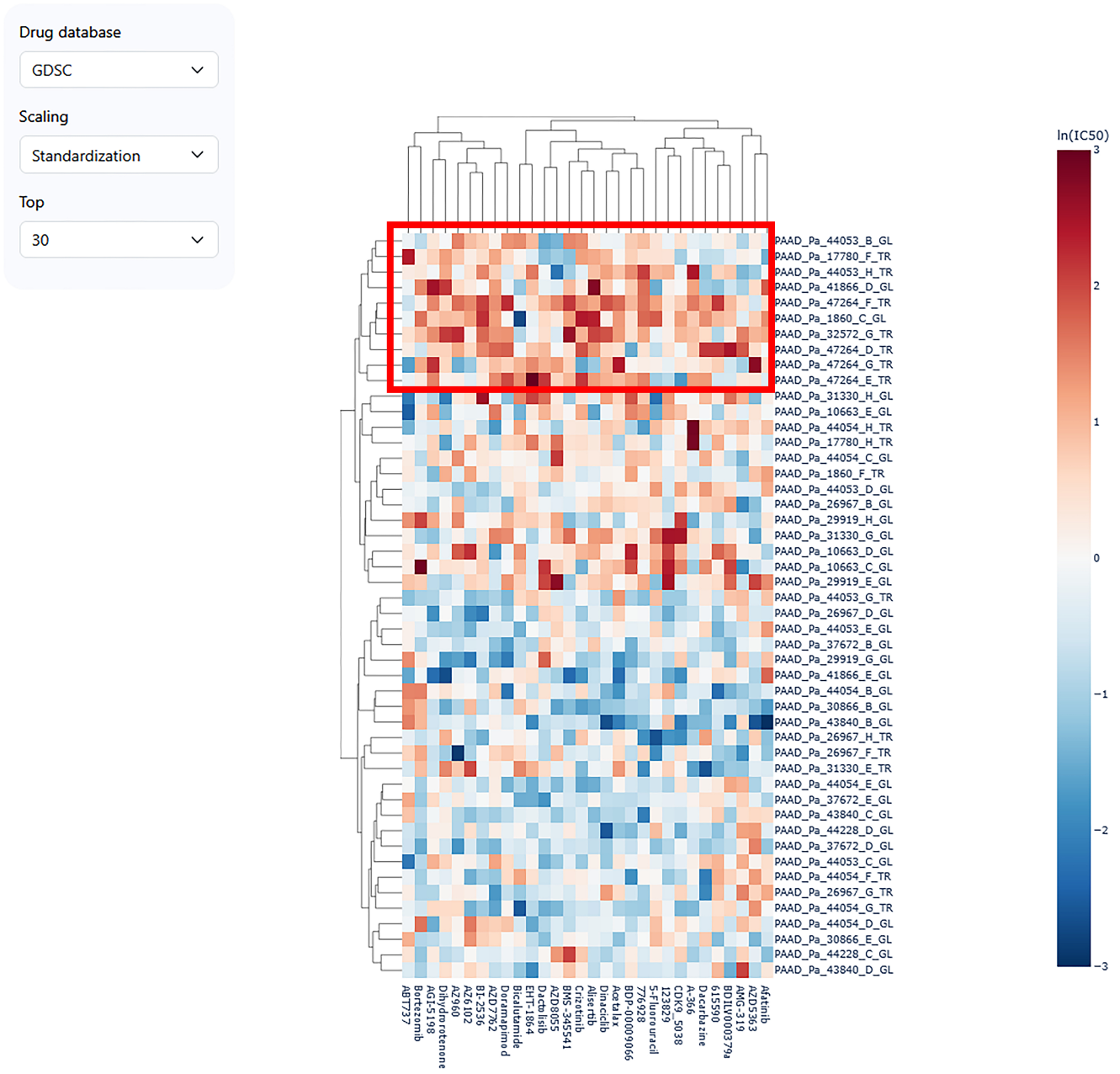


Figure 4. Clustermap displaying drug response predictions for PDAC patients. The color scale represents the scaled IC50 values, with values at the bottom of the scale indicating higher sensitivity and values at the top of the scale indicating greater resistance. The highlighted region (rectangular box) marks a cluster enriched with TR-type samples, which are predicted to be more resistant compared to the GL group. Users can interactively adjust scaling methods and the number of top drugs displayed to explore response patterns.

tial mechanisms of action and prioritize validation efforts. Furthermore, the heatmap-based grouping of samples can guide patient stratification and reveal specific therapeutic opportunities.

Approaches to predict drug sensitivity from scRNAseq perturbation experiments are also emerging in the literature (e.g. [11, 12, 38]). Benchmark datasets to facilitate the training of sensitivity prediction models on these datasets have also been established and released to the community (i.e. [39] and [40]). Such datasets are expected to provide an unprecedented view on the cell-type dependent mechanisms governing the response to anti-cancer drugs. However, the covered drugs and cancer cell lines are still small compared to the datasets used to train the CellHit model. Indeed, the largest single-cell perturbation atlas, i.e. Tahoe-100M [40] comprises the perturbation data of 1100 drugs to "cell villages" derived from 50 different cancer cell lines. Another critical consideration for scRNAseq perturbation models is their actual translatability to clinical settings. Indeed, the costs associated with scRNAseq sequencing are still relatively high and prevent its wider adoption to

clinical settings, where bulk transcriptomics still represents the most cost-effective solution for personalized medicine programs.

Currently, the methodology focuses on monotherapy predictions. This web server is intended to facilitate the exploration of predicted sensitivities for hundreds of drugs, including many non-oncological candidates modulating protein families not typically targeted in cancer (e.g. G protein-coupled receptors [36, 37]), with the potential for repurposing them for anti-cancer strategies. A promising future direction lies in extending the framework to combination therapies, where interactions among drugs can yield synergistic effects that are not captured by single-drug therapies.

Acknowledgements

We gratefully acknowledge the CINECA award, in collaboration with AIRC, for the availability of high-performance computing resources and generous support, as well as the computational resources of the Center for High-Performance Computing (CHPC) at Scuola Normale Superiore.

Author contributions: Francesco Carli (Conceptualization, Formal analysis, Methodology, Validation, Writing—original draft, Software, Visualization), Natalia De Oliveira Rosa (Conceptualization, Software, Validation, Visualization, Writing—original draft), Simon Blotas (Formal analysis, Methodology), Pierluigi Di Chiaro (Formal analysis, Validation, Investigation), Luisa Bisceglia (Investigation, Validation), Mariangela Morelli (Investigation, Validation), Francesca Lessi (Investigation, Validation), Anna Luisa Di Stefano (Conceptualization, Investigation, Validation), Chiara Maria Mazzanti (Conceptualization, Investigation, Validation), Gioacchino Natoli (Conceptualization, Investigation, Validation), Francesco Raimondi (Conceptualization, Methodology, Funding acquisition, Supervision, Writing—original draft).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

F.R. was supported through the Italian Association for Cancer Research (AIRC) under My First AIRC Grant (MFAG) 2020 – ID. 24317. The research leading to these results also received funding from project Department of Excellence "Faculty of Sciences" of Scuola Normale Superiore, as well as from the Project granted by Next Generation EU – National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, NRRP) – Mission 4 Component 2 Investment 1.4 – Ministry of University and Research (MUR) Call N. 3277 Project Code ECS_00000017 MUR Directoral Decree n.1055, 23 June 2022, CUP B83C22003930001, project title "Tuscany Health Ecosystem – THE", Spoke 8 (to F.R.). P.D.C. was partially supported by an AIRC fellowship for Italy, grant #25542. G.N. was supported by AIRC (Investigator Grant #27555 and AIRC 5 × 1000 Grant #21147) and Fondazione IEO - Monzino. This work was also partially supported by the Italian Ministry of Health with the "Ricerca Corrente" and" 5 × 1000" funds to the IEO IRCCS. Funding to pay the Open Access publication charges for this article was provided by the Italian Association for Cancer Research (AIRC) and from the Open Access Publishing Fund of the Scuola Normale Superiore.

Data availability

We used the following dataset:

DepMap (Metadata, RNA Seq, mutations,PRISM): https://depmap.org/portal/data_page/?tab=allData

GDSC2: https://cog.sanger.ac.uk/cancerrxgene/GDSC_release8.5/GDSC2_fitted_dose_response_27Oct23.xlsx

TCGA: https://xenabrowser.net/datapages/?dataset=TumorCompendium_v11_PolyA_hugo_log2tpm_58581genes_2020-04-09.tsv&host=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

PRISM Drug metadata: https://depmap.org/repurposing/

OncoKB: https://www.oncokb.org/actionable-genes#sections=Tx

Reactome: https://reactome.org/download-data.

The fastaq and TPM files of the GBM RNAseq data are available under accession number GSE287932. The reimplemented Celligner and the Parametric UMAP codebases are available on Github at https://github.com/fcarli/celligner and https://github.com/fcarli/parametric_umap (also available through pypi), respectively. The code has also been archived on Zenodo: https://doi.org/10.5281/zenodo.15309741.

References

1. Weinstein JN, Collisson EA, Mills GB *et al.* The cancer genome atlas pan-cancer analysis project. *Mol Syst Biol* 2013;**45**:D1003–9. https://doi.org/10.1038/ng.2764
2. Barretina J, Caponigro G, Stransky N *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;**483**:603–7. https://doi.org/10.1038/nature11003
3. Iorio F, Knijnenburg TA, Vis DJ *et al.* A landscape of pharmacogenomic interactions in cancer. *Cell* 2016;**166**:740–54. https://doi.org/10.1016/J.CELL.2016.06.017/ATTACHMENT/2690C3CA-4270-4D2A-8984-FDDCBFA738C7/MMC9.PDF
4. Rees MG, Seashore-Ludlow B, Cheah JH *et al.* Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol* 2015;**12**:109–16. https://doi.org/10.1038/nchembio.1986
5. Corsello SM, Nagari RT, Spangler RD *et al.* Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nat Cancer* 2020;**1**:235–48. https://doi.org/10.1038/s43018-019-0018-6
6. Tsherniak A, Vazquez F, Montgomery PG *et al.* Defining a cancer dependency map. *Cell* 2017;**170**:564–76. https://doi.org/10.1016/j.cell.2017.06.010
7. Chen J, Zhang L. A survey and systematic assessment of computational methods for drug response prediction. *Brief Bioinform* 2021;**22**:232–46. https://doi.org/10.1093/bib/bbz164
8. Firoozbakht F, Yousefi B, Schwikowski B. An overview of machine learning methods for monotherapy drug response prediction. *Brief Bioinform* 2022;**23**:bbab408. https://doi.org/10.1093/bib/bbab408
9. Xia F, Allen J, Balaprakash P *et al.* A cross-study analysis of drug response prediction in cancer cell lines. *Brief Bioinform* 2022;**23**:bbab356. https://doi.org/10.1093/bib/bbab356

10. Sharifi-Noghabi H, Jahangiri-Tazehkand S, Smirnov P *et al.* Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models. *Brief Bioinform* 2021;**22**:bbab294. https://doi.org/10.1093/BIB/BBAB294
11. Fustero-Torre C, Jiménez-Santos MJ, García-Martín S *et al.* Beyondcell: targeting cancer therapeutic heterogeneity in single-cell RNA-seq data. *Genome Med* 2021;**13**:187. https://doi.org/10.1186/S13073-021-01001-X
12. Pellecchia S, Viscido G, Franchini M *et al.* Correction: predicting drug response from single-cell expression profiles of tumours. *BMC Med* 2024;**22**:70. https://doi.org/10.1186/S12916-024-03289-Z
13. Carli F, Di Chiaro P, Morelli M *et al.* Learning and actioning general principles of cancer cell drug sensitivity. *Nat Commun* 2025;**16**:1654. https://doi.org/10.1038/s41467-025-56827-5
14. Kuenzi BM, Park J, Fong SH *et al.* Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* 2020;**38**:672–84. https://doi.org/10.1016/j.ccell.2020.09.014
15. Chawla S, Rockstroh A, Lehman M *et al.* Gene expression based inference of cancer drug sensitivity. *Nat Commun* 2022;**13**:5680. https://doi.org/10.1038/s41467-022-33291-z
16. Ferraro L, Scala G, Cerulo L *et al.* MOViDA: multiomics visible drug activity prediction with a biologically informed neural network model. *Bioinformatics* 2023;**39**:btad432. https://doi.org/10.1093/bioinformatics/btad432
17. Wang X, Sun Z, Zimmermann MT *et al.* Predict drug sensitivity of cancer cells with pathway activity inference. *BMC Med Genomics* 2019;**12**:15. https://doi.org/10.1186/s12920-018-0449-4
18. Tang Y-C, Gottlieb A. Explainable drug sensitivity prediction through cancer pathway enrichment. *Sci Rep* 2021;**11**:3128. https://doi.org/10.1038/s41598-021-82612-7
19. Samal BR, Loers JU, Vermeirssen V *et al.* Opportunities and challenges in interpretable deep learning for drug sensitivity prediction of cancer cells. *Front Bioinform* 2022;**2**:1036963. https://doi.org/10.3389/fbinf.2022.1036963
20. Ma J, Fong SH, Luo Y *et al.* Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat Cancer* 2021;**2**:233–44. https://doi.org/10.1038/s43018-020-00169-2
21. Chiaro PDi, Nacci L, Arco F *et al.* Mapping functional to morphological variation reveals the basis of regional extracellular matrix subversion and nerve invasion in pancreatic cancer. *Cancer Cell* 2024;**42**:662–81. https://doi.org/10.1016/j.ccell.2024.02.017
22. Behdenna A, Colange M, Haziza J *et al.* pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods. *BMC Bioinf* 2023;**24**:459. https://doi.org/10.1186/s12859-023-05578-5
23. Chen T, Guestrin C 2016; XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 785–94. https://doi.org/10.1145/2939672.2939785
24. Akiba T, Sano S, Yanase T *et al.* 2019; Optuna: a next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 2623–31. https://doi.org/10.1145/3292500.3330701
25. Watanabe S Tree-structured parzen estimator: understanding its algorithm components and their roles for better empirical performance. arXiv, https://arxiv.org/abs/2304.11127, 26 May 2023, preprint: not peer reviewed.
26. Warren A, Chen Y, Jones A *et al.* Global computational alignment of tumor and cell line transcriptional profiles. *Nat Commun* 2021;**12**:22. https://doi.org/10.1038/s41467-020-20294-x
27. Abid A, Zhang MJ, Bagaria VK *et al.* Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nat Commun* 2018;**9**:2134. https://doi.org/10.1038/s41467-018-04608-8
28. Pedregosa F, Varoquaux G, Gramfort A *et al.* Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;**12**:2825–30. https://doi.org/10.48550/arXiv.1201.0490
29. Ritchie ME, Phipson B, Wu D *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47. https://doi.org/10.1093/nar/gkv007
30. Paszke A, Gross S, Massa F *et al.* 2019; PyTorch: an imperative style, high-performance deep learning library. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 8026–37.
31. Sainburg T, McInnes L, Gentner TQ. Parametric UMAP embeddings for representation and semi-supervised learning. *Neural Comput* 2021;**33**:2881–907. https://doi.org/10.1162/neco_a_01434
32. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. *JOSS* 2020;**3**:861. https://doi.org/10.21105/joss.00861
33. Dyer SC, Austine-Orimoloye O, Azov AG *et al.* Ensembl 2025. *Nucleic Acids Res* 2025;**53**:D948–57. https://doi.org/10.1093/nar/gkae1071
34. Seal RL, Braschi B, Gray K *et al.* Genenames.Org: the HGNC resources in 2023. *Nucleic Acids Res* 2023;**51**:D1003–9. https://doi.org/10.1093/nar/gkac888
35. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform* 2020;**2**:lqaa078. https://doi.org/10.1093/nargab/lqaa078
36. Douze M, Guzhva A, Deng C *et al.* . The Faiss library. arXiv, https://doi.org/10.48550/arXiv.2401.08281, 11 February 2025, preprint: not peer reviewed.
37. Lundberg SM, Lee S-I . 2017; A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 4768–77.
38. Lotfollahi M, Klimovskaia Susmelj A, De Donno C *et al.* Predicting cellular responses to complex perturbations in high-throughput screens. *Mol Syst Biol* 2023;**19**:e11517. https://doi.org/10.15252/MSB.202211517
39. NeurIPS Poster A benchmark for prediction of transcriptomic responses to chemical perturbations across cell types. https://neurips.cc/virtual/2024/poster/97655 (12 December 2024, date last accessed).
40. Zhang J, Ubas AA, de Borja R *et al.* Tahoe-100M: a giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. bioRxiv, https://doi.org/10.1101/2025.02.20.639398, 23 April 2025, preprint: not peer reviewed.

Received: March 24, 2025. **Revised:** April 17, 2025. **Editorial Decision:** April 29, 2025. **Accepted:** May 2, 2025
© The Author(s) 2025. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.