nature communications

Article

https://doi.org/10.1038/s41467-025-56827-5

Learning and actioning general principles of cancer cell drug sensitivity

Received: 22 April 2024

Accepted: 3 February 2025

Published online: 15 February 2025

Check for updates

Francesco Carli[1,2] ✉, Pierluigi Di Chiaro[3], Mariangela Morelli[4], Chakit Arora[1], Luisa Bisceglia[1], Natalia De Oliveira Rosa[1], Alice Cortesi[3], Sara Franceschi[4], Francesca Lessi[4], Anna Luisa Di Stefano[5], Orazio Santo Santonocito[5], Francesco Pasqualetti[6], Paolo Aretini[4], Pasquale Miglionico[1], Giuseppe R. Diaferia[3,10], Fosca Giannotti[7], Pietro Liò[8], Miquel Duran-Frigola[9], Chiara Maria Mazzanti[4], Gioacchino Natoli[3] & Francesco Raimondi[1] ✉

High-throughput screening of drug sensitivity of cancer cell lines (CCLs) holds the potential to unlock anti-tumor therapies. In this study, we leverage such datasets to predict drug response using cell line transcriptomics, focusing on models' interpretability and deployment on patients' data. We use large language models (LLMs) to match drug to mechanisms of action (MOA)-related pathways. Genes crucial for prediction are enriched in drug-MOAs, suggesting that our models learn the molecular determinants of response. Furthermore, by using only LLM-curated, MOA-genes, we enhance the predictive accuracy of our models. To enhance translatability, we align RNAseq data from CCLs, used for training, to those from patient samples, used for inference. We validated our approach on TCGA samples, where patients' best scoring drugs match those prescribed for their cancer type. We further predict and experimentally validate effective drugs for the patients of two highly lethal solid tumors, i.e., pancreatic cancer and glioblastoma.

Landmark cancer genomic projects such as The Cancer Genome Atlas (TCGA) have provided an unprecedented, multimodal picture of the complex genetic and molecular landscape characterizing major tumor types[1]. The possibility to define tumors at a molecular level is changing cancer treatment through the development of single- or combinatorial-targeted therapies for patients characterized by specific genetic makeups. Unfortunately, effective treatments are still lacking for many patients and the emergence of resistance further limits the clinical benefit of therapies.

In recent years, the availability of large-scale pharmacogenomics databases, including the Cancer Cell Line Encyclopedia (CCLE)[2], the Genomics of Drug Sensitivity in Cancer v1 and v2 (GDSC)[3], the Cancer Therapeutics Response Portal v2 (CTRPv2)[4], the Profiling relative inhibition simultaneously in mixtures (PRISM)[5] and DepMap[6], has fostered the development of personalized oncology strategies.

Initial landmark work by Iorio, Garnett & coworkers highlighted the genomic alterations that sensitize to drugs as well as the contribution of different genome sequencing data types for drug sensitivity predictions[3]. Since then, several computational forecast methodologies have been proposed to leverage the information in CCLE and GDSC to predict cancer cell lines drug-sensitivity (reviewed in refs. 7–9) and best practices to develop predictive models have been

[1]Laboratorio di Biologia Bio@SNS, Scuola Normale Superiore, Pisa, Italy. [2]Department of Computer Science, Univerisity of Pisa, Pisa, Italy. [3]Department of Experimental Oncology, IEO, European Institute of Oncology IRCCS, Milano, Italy. [4]Fondazione Pisana per la Scienza ONLUS, Pisa, Italy. [5]Neurosurgical Department of Spedali Riuniti di Livorno, Livorno, Italy. [6]Radiotherapy Department, Azienda Ospedaliera Universitaria Pisana, Pisa, Italy. [7]Scuola Normale Superiore, Pisa, Italy. [8]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK. [9]Ersilia, Can Sutirà, Porqueres, Girona, Spain. [10]Present address: Botton-Champalimaud Pancreatic Cancer Center, Champalimaud Foundation, Lisbon, Portugal. ✉e-mail: francesco.carli@sns.it; francesco.raimondi@sns.it

https://doi.org/10.1038/s41467-025-56827-5

issued [10]. Models "primed" with prior knowledge have also been developed. In general, these models exploit representations of biological processes instead of individual genes as features to reduce the number of input, independent variables, while leading to an increase of model performance and interpretability[11–16]. However, a systematic investigation of whether drug sensitivity models are able to learn, in an unbiased fashion, the biological processes associated with drugs' mechanism of action (MOA) is still missing from the literature. Moreover, to what extent models trained on expression data in cancer cell lines can be exploited to interpret patients' bulk RNAseq data is still a largely unsolved issue, although earlier studies tried to tackle this fundamental point[11,17].

In this work, to forecast drug sensitivity in cancer cell lines we developed a machine learning algorithm trained on large cancer cell lines drug sensitivity datasets, i.e., GDSC and PRISM, which represent the largest screening datasets of oncological and non-oncological drugs, respectively. We focused on the interpretability and deployment of the predictive models, which are both critical to generate new

Fig. 1 | Training a machine learning framework for predicting cancer cell sensitivity to drugs. A Schematic workflow of the pipeline from dataset acquisition and model training through to benchmarking and external validation. Leveraging large-scale transcriptomics datasets, several machine learning models are trained on Genomics of Drug Sensitivity in Cancer (GDSC) and Profiling Relative Inhibition Simultaneously in Mixtures (PRISM) viability screening datasets. A Large Language Model (LLM) assists the curation of drug Mechanisms of Action (MOA), enhancing model interpretability and facilitating feature selection. Three model types are trained: one using cell transcriptomics and drug features, another using only transcriptomics, and a third using a selected subset of transcriptomic data informed by pathways identified by the LLM. These models undergo benchmarking across 20 distinct train/validation/test splits. The best-performing model is then applied in inference on The Cancer Genome Atlas (TCGA) bulk RNA-seq data and on external patient datasets for pancreatic ductal adenocarcinoma (PDAC) and glioblastoma multiforme (GBM). Predictions are externally validated on TCGA data using National Cancer Institute (NCI) cancer drug indications to assess the recovery of known information, as well as experimentally on primary (GBM dataset) and commercial (PDAC dataset) cancer cell lines; **B** Bar plot comparing the performance of different model architectures (MLP, XGBoost and literature baselines) and input feature representations (cell features and drug features) in terms of Pearson correlation with observed drug sensitivities. Different colors denote different learning algorithms (e.g., light blue XGBoost and purple MLP). Etched bars highlight models using only transcriptomic data (no drug featurizations). Results are obtained by averaging results across 20 distinct test splits. Error bars represent SD of the results; **C** Bar plot depicting Mean Squared Error (MSE) for the same models and features as in (**B**). Also in this case results are obtained by averaging results across 20 distinct test splits. Error bars represent SD of the results; **D** Histogram of the distribution of Pearson correlation coefficients for drug-specific models using all genes, indicating the median, mean, and standard deviation; **E** Box plots illustrating the variability in the distribution of Pearson correlation coefficients across 20 different random training/testing splits. Each box plot represents a specific model and displays the median correlation (central line), interquartile range (box edges), and variability outside the upper and lower quartiles (whiskers). Source data are provided as a Source Data file.

testable hypotheses (Fig. 1A). We created a pipeline, that we called *CellHit*, by combining the predictive models with Celligner[18], an unsupervised alignment strategy that allows to identify cell lines whose transcriptomics profile most closely match the bulk RNAseq from patient tumors. We employed our pipeline to infer the best-scoring drugs for the entire TCGA cohort based on patient transcriptomics profile, as well as on samples from PDAC and GBM patients, which were experimentally validated.

Results

An interpretable model for cell line drug response predictions with or without drug representations

We developed an interpretable machine learning framework for the prediction of drug sensitivity of cancer cell lines by using the latest version of GDSC and PRISM datasets. Given the lower dimensionality of GDSC, we tested alternative modeling strategies on this dataset and then transferred optimal settings to PRISM. After GDSC data pre-processing, we obtained the profiling of 286 unique drugs in 686 cell lines. We first implemented an integrated model by jointly considering representations of both drugs and cell lines to predict IC50 values as target variables (Fig. 1A). We employed different strategies to numerically represent drugs chemical structures and cell line expression profiles to provide inputs to ML algorithms (Methods). As for cell lines, we first transformed the RNAseq data using Celligner[18] along with TCGA bulk RNAseq samples. This preprocessing step is required to match the cell line closest to TCGA samples by aligning their RNAseq data, enabling the application of our ML model, which was trained on cancer cell lines, to patient samples (see below). We employed drug and cellular representations as features to train a supervised regression model to predict IC50 values (i.e., "Joint feature" models, see Fig. 1A). We used a tissue-dependent, cell line stratification strategy to generate training, validation and testing splits to avoid data leakages (see Methods). We found that XGBoost, in combination with all-gene expression vectors and one-hot encodings of the molecules, achieved the best performance (Pearson correlation coefficient $\rho = 0.89$, Mean Square Error MSE = 1.55; Fig. 1B, C), being able to outperform other architectures, such as Multi-Layer Perceptron (MLP), as well as other competitive methods available from the literature (Chen & Zhang, 2021) (Fig. 1B, C; Supplementary Fig. 1; Supplementary data 1).

The fact that the simplest representation of the drug, i.e., one-hot encoding, achieved the best results, suggested that the model leverages drugs identifiers, disregarding molecular properties of individual drugs. Since our main goal was to learn the transcriptional programs responsible for drug responses for post-hoc interpretation, we generated drug-specific models by employing only gene expression as input features (i.e., "All-genes" models, see Fig. 1A). We trained the models by repeating the same procedure used for the joint representation. Overall, aggregated IC50 predictions have correlations with experimental values very close to the joint model ($\rho = 0.88$, MSE = 1.73; Fig. 1B, C; Supplementary data 1). By evaluating performances on a held-out testing set for each of the 286 drug-specific models, we obtained a median $\rho = 0.40$ (Fig. 1D). The best performance was achieved by Venetoclax ($\rho = 0.72$, Fig. 1E), a small molecule that increases apoptosis by inhibiting *BCL2*[19]. A total of 73 drug models (25%) had correlation $\rho > 0.5$ (Supplementary data 2). In summary, we developed reliable models for drug sensitivity based on just cell lines transcriptomics data.

Models' interpretation reveals convergence between important genes and known drug-targets

For each drug-specific model, we inspected the genes most important for prediction and checked whether they were either known targets, or members of pathways associated with the known MOA of that drug. Gene importance, defined as the contribution of individual gene expression to model predictiveness, was evaluated in two ways. First, we inspected whether genes gave a positive or negative contribution to the final IC50 prediction via a game theory approach (i.e., Shapley Additive exPlanations[20]). Second, we used an importance permutation method, randomly shuffling genes and evaluating the effect of the perturbation on test set metrics. We considered only those genes deemed important with both approaches.

We found that 39% of the drug-specific models trained on GDSC identified the known target among important genes in at least one model out of 20 random splits (Fig. 2A, Supplementary data 2). Remarkably, models for *BCL2* inhibitors, such as Venetoclax, Navitoclax and ABT737, consistently recovered their target in the majority of the trained models (Fig. 2A, B). Several other drug models recovered the corresponding targets in more than 50% of the splits (e.g., Gefitinib-*EGFR*, Nutlin-3a(-)-*MDM2*, or Linsitinib-*IGF1R*; Fig. 2A). We assessed the statistical validity of these results by estimating the drug-specific recovery rate of a random gene across the 20 splits (see Methods). We found that 70% of the targets are found at or above the 90th percentiles of the background distributions, suggesting a significant positive recovery rate for most models (Fig. 2A, Supplementary data 2). Analysis of the importance scores, for example for the top performing model (i.e., Venetoclax), showed that higher values are attributed to the corresponding target, i.e., *BCL2*. This connection is seen both as a strong negative contribution to the predicted IC50 value (i.e., SHAP importance; Fig. 2B teal) and in test metrics (i.e., Correlation delta; Fig. 2B orange). By plotting for each cell line the target gene expression, Venetoclax's experimental IC50s, as well as Venetoclax's model SHAP values, it is possible to understand how the Venetoclax model leverages the expression levels of its target (i.e., *BLC2*) to successfully predict IC50 (Fig. 2C). Indeed, cell lines having a higher expression of

https://doi.org/10.1038/s41467-025-56827-5

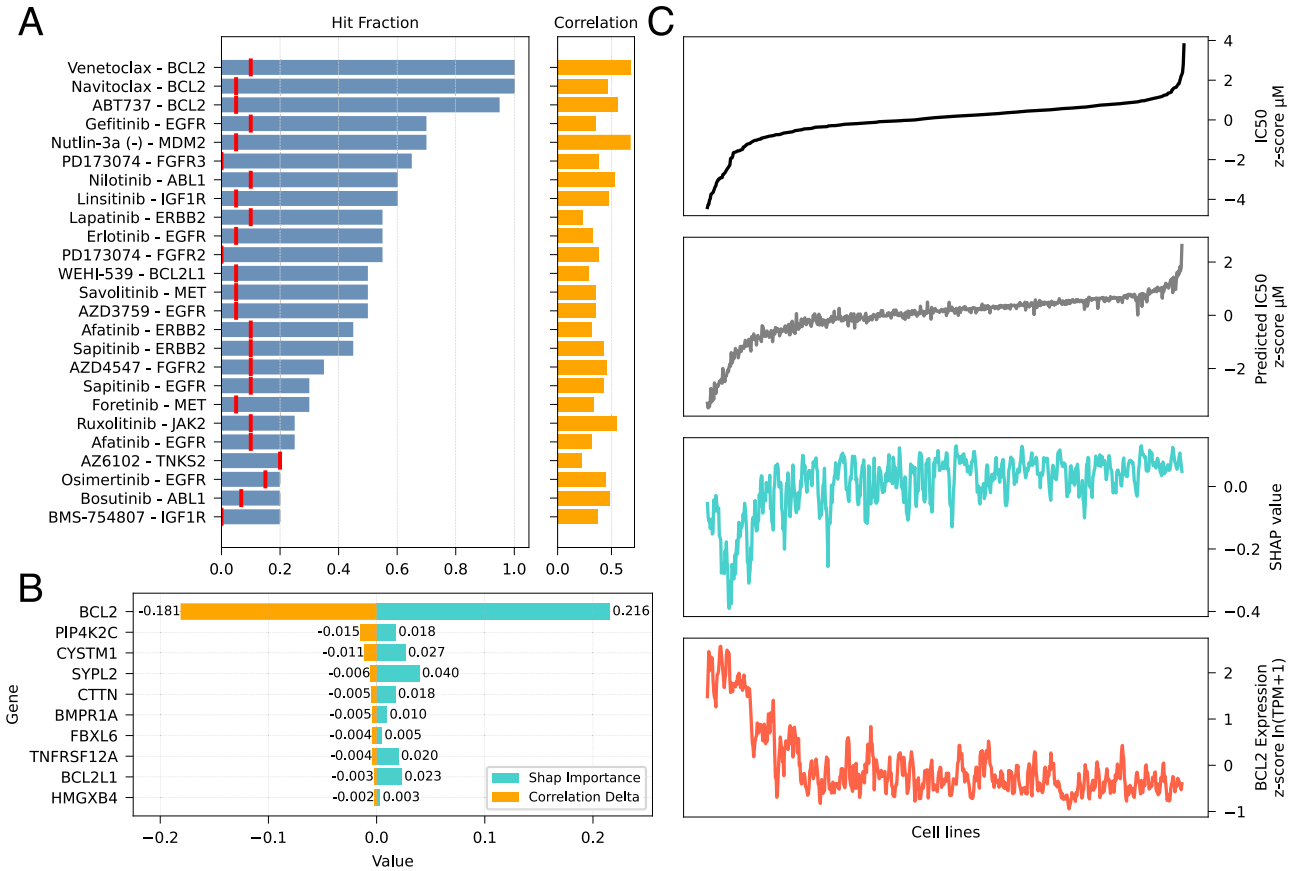


Fig. 2 | GDSC model interpretability. A Target recovery for the top 25 ligand-target pairs. The length of the bars on the left indicates the fraction of recovery (# of times the drug-specific model identifies the putative gene as important out of 20 train/test splits). Red lines represent the 95th percentile of the Hit Fraction distribution across all genes for a given drug. The length of the bars on the right shows the median pearson correlation for each drug-specific model; **B** SHAP (teal) and correlation delta (orange) importances for the Venetoclax drug. Permutation importance reflects the decrease in the model's prediction accuracy when a feature's values are shuffled, indicating its importance (greater drops signify higher importance). SHAP importance represents a feature's contribution to the model's prediction, with larger absolute values indicating greater importance; **C** An integrated assessment of the Venetoclax model across various cell lines (X axis). The top plot (in black) shows the experimental IC50 z-scores, while the second plot (in gray) depicts the predicted IC50 values, providing a comparison of model performance against experimental data. The third and fourth plots (in teal and red) respectively represent the SHAP values and expression levels of *BCL2*. Overall, the figure shows how lower IC50 values (higher drug efficacy) are associated with higher *BCL2* expression levels and correctly identified impact (negative SHAP value) of the gene on predicted IC50. Source data are provided as a Source Data file.

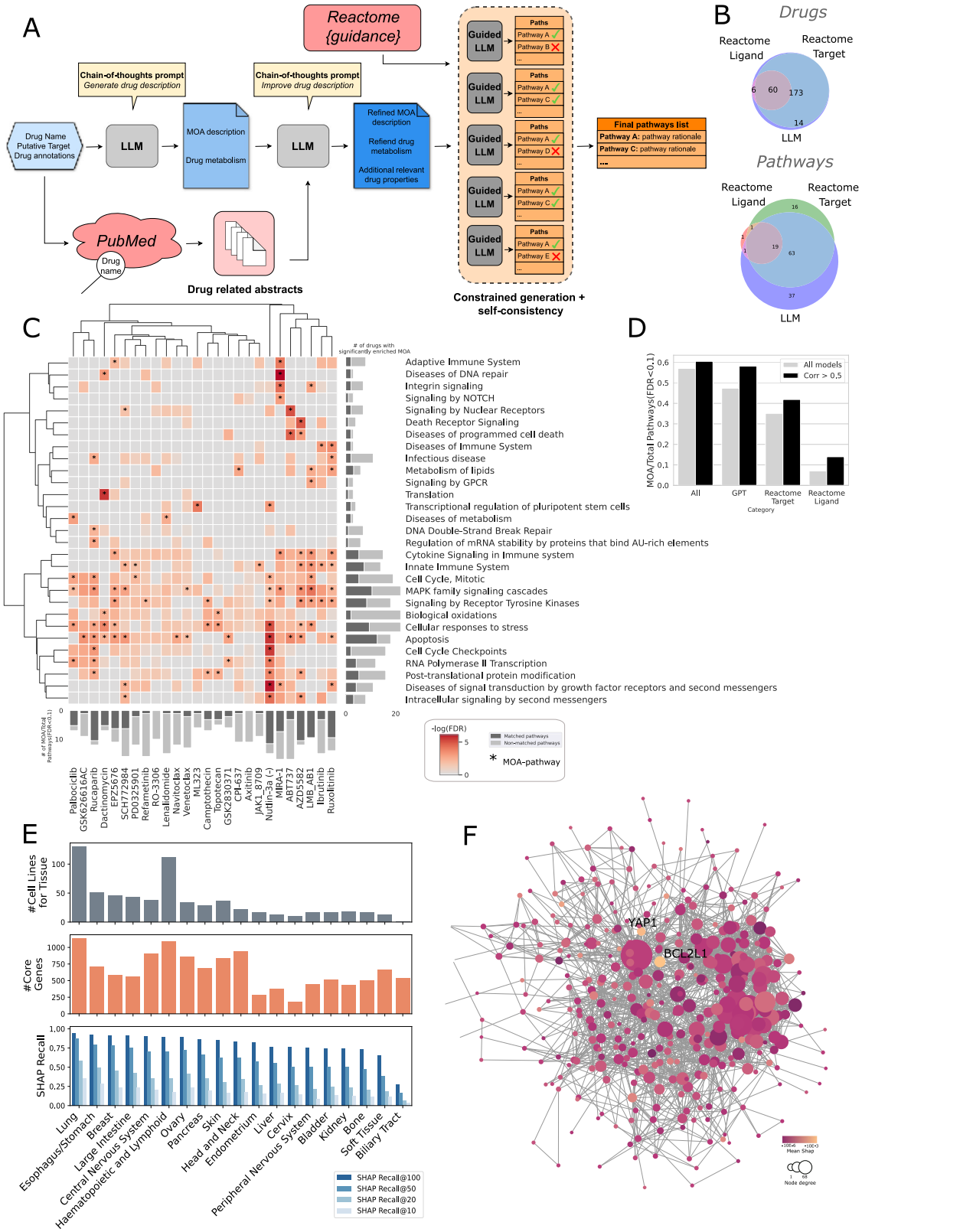
BCL2 are characterized by lower IC50 as well as strongly negative SHAP values (i.e., higher efficacy of Venetoclax as a *BCL2* inhibitor) (Fig. 2C).

MOA processes and gene dependencies are learnt by the drug sensitivity models

We further inspected whether the models learned about MOA-related biological processes and pathways. The information of biological pathways associated to the action of a certain drug is sporadically annotated in chemical bioactivity or biological pathway knowledge bases. We therefore systematically curated the association of GDSC drugs to the pathways of a reference knowledgebase, Reactome[21]. We first retrieved a total of 66 GDSC drugs that are annotated to Reactome pathways through Guide to Pharmacology/Pubchem[22] mappings. We then considered all pathways containing the drugs' putative targets, finding matches for 233/287 (81%) unique GDSC drugs. To further extend the drug-MOAs' pathway coverage, we leveraged an open-source Large Language Model (LLMs), i.e., Mixtral Instruct 8x7b, to generate detailed descriptions of drug-related processes by inputting the generic name of the drug (or available synonyms) and available annotations from the GDSC compound information table. The resulting annotations informed a second prompt designed to identify from Reactome the semantically closest pathways for each drug (Fig. 3A; see Methods). Through this approach, we were able to retrieve detailed information about MOA and associated pathways for 253/287 (88%; Supplementary data 3) GDSC drugs (Fig. 3B, top), therefore increasing the coverage of annotated drugs (Fig. 3B, bottom). We finally combined the three sets of drug-pathway associations, for a total of 5662 instances, 253 unique drugs and 138 unique pathways (Supplementary data 3).

We performed pathway enrichment analysis on important genes from All-genes drug specific models to check whether known MOAs for the corresponding drug were identified. We found a total of 114 GDSC drug models with at least one enriched pathway (FDR < 0.1). Out of these, 65 (57%) unique drugs have at least one MOA-pathway significantly enriched when considering the aggregated list of MOA-pathways (Fig. 3D; Supplementary data 4), with the LLM-derived ones being the single list providing the highest recovery of enriched pathways (Fig. 3D; Supplementary data 4). The fraction of drug models with at least one significantly enriched MOA-pathway increased upon considering models with a correlation $\rho > 0.5$ (Fig. 3D; Supplementary data 4). We then clustered pathway enrichments of the best-performing drug models ($\rho > 0.5$) having at least one significant MOA-pathway (Fig. 3C). We found that certain processes, such as "Apoptosis", "Cellular responses to stress", "Biological Oxidations", "MAPK family signaling cascades", were widely enriched across drugs, often consistently with drugs' MOAs, and clustered apart (Fig. 3C).

https://doi.org/10.1038/s41467-025-56827-5



We evaluated whether important genes also recovered information about gene dependencies of cancer cell lines from different tissues. To this end, we retained drug-cell line instances yielding the most significant predictions, ranked the top k most important genes based by SHAP values, and pooled them on the basis of the tissue of origin of the cell lines. We then evaluated the recall of the top k important genes to identify core essential genes from an updated dependency map across 27 cancer tissues[23]. Remarkably, when aggregating the top 100 genes by SHAP importance, we identified core essential genes with a recall greater than 0.9 in several tissues (Fig. 3E, Supplementary data 5). These essential, prediction-important genes are often found in highly connected protein-protein interaction networks (e.g., STRING network of important, essential genes in lung, Fig. 3F). By ranking genes based on their average SHAP importance

Fig. 3 | Analysis of Drug MOAs and Gene Essentiality via GDSC Models. **A** Workflow depicting the use of a large language model (LLM) for generating drug MOAs and identifying semantically relevant pathways. Starting from the drug's available metadata, an LLM is repeatedly tasked with specialized prompts to generate a drug textual description. In parallel, PubMed is queried programmatically with the drug name to retrieve abstracts related to the drug. The information is integrated in a final textual description. The obtained drug description is used by a "Guided" LLM to choose which are the Reactome pathways which are most likely to modulate drug efficacy. This last procedure is repeated 5 different times and only pathways selected at least two times are retained; **B** Venn diagram showing the different drugs(top) and pathways (bottom) recovered using the LLM procedure as compared to pathway match based on drugs' and target's names; **C** Heatmap of significant MOA-pathways for various drug models, filtered by a correlation threshold $\rho > 0.5$. Drug names and involved pathways are labeled along the x-axis and y-axis, respectively. Starred squares highlight pathways linked to drugs via at least one annotation criterion. Adjacent bar plots show the count of significantly enriched elements per row/column in light gray, with those annotated by the pipeline in dark gray. A vertical dashed line highlights the presence of a group of drugs that most frequently recover pathways and known MOAs; **D** number of significantly enriched MOA-pathways obtained from different annotation criteria; **E** tissue-wise statistics of number of cell lines (top), number of core essential genes (middle), recall of essential genes at different important genes (SHAP) stringencies (top k 10, 20, 50, 100); **F** STRING PPI network of lung core essential genes recovered by SHAP importances. Nodes' have diameters proportional to node degree and are colored according to SHAP values (the brighter the more important). Source data are provided as a Source Data file.

across drug models, we found *BCL2L1* (Bcl2-like 1) and *YAP1* (Yes1 Associatd Transcriptional Regulator) as the top 2 most important genes (brighter nodes, Fig. 3F; Supplementary data 6). Overall, this analysis suggests that drug sensitivity is also realized through the modulation of genes that are essential for cell survival and proliferation.

Extending explainable drug sensitivity predictions to the PRISM dataset

We employed a similar strategy to train drug-specific, All-genes models for 6337 drugs and 887 cell lines available in the PRISM database[5]. We obtained a total of 762 drug models with a correlation $\rho = 0.2$ (Fig. 4A; Supplementary data 7; see Methods). The aggregated IC50 predictions of these models with experimental values achieved a $\rho = 0.80$ and MSE = 1.18 (Fig. 4B). Drugs targeting kinases are by far the category with the highest number of models with a correlation $\rho > 0.2$, being also more effective in recovering the corresponding targets among important genes (Fig. 4C). Other recurrent drug target categories are generic Enzymes, followed by Epigenetic enzymes and GPCR targets (Fig. 4C). Kinases also stand out when normalizing the number of models for the total number of drugs considered in PRISM (Supplementary Fig. 2A, B), or considering the number of drugs achieving a certain LFC threshold (Supplementary Fig. 2C).

Inspection of the most important genes revealed that 62% (339 out of 547) of the models for drugs with annotated target information identified one corresponding target among the important genes across at least one train/test split (Fig. 4D; Supplementary data 8). Also for the PRISM dataset, we successfully identified 73.7% of drug targets at or above the 90th percentile threshold of the background recovery distribution (Fig. 4D). STF-31 and CGM097 were the two drug models that recovered their corresponding targets (i.e., the regulator of intracellular NAD+ pool Nicotinamide phosphoribosyltransferase, NAMPT; and the p53 ubiquitin ligase MDM2, respectively) in 100% of the training/testing splits (Fig. 4D). Among the most supported models, we also found a few non-oncological drugs, such as PARDOPRUNOX (targeting the Serotonin 5-HT7 receptor HTR7), which is approved as a treatment for Parkinson.

Similarly to the GDSC drug datasets, we curated MOA-pathways for 6305 PRISM drugs (Supplementary data 9) and performed pathway enrichment on the most important genes of All-genes models. To determine whether the LLM was identifying drug-specific biological pathways, and not those broadly associated with cancer, we leveraged available PRISM drugs' categorization into chemotherapeutic agents, targeted therapies, and non-oncological drugs. We analyzed the proportion of biological pathways assigned to each drug type and found that different classes of drugs are associated with distinct sets of pathways (Supplementary Fig. 3). We also performed pathway analysis on important genes for each drug model and similarly displayed enrichments for the best-performing instances (i.e., models with $\rho > 0.5$). Certain pathways, e.g., "Cell Cycle, Mitotic", "Cellular responses to stress", "Apoptosis", "Cell Cycle Checkpoints", "Cytokine Signaling in Immune System" and "Signaling by Receptor Tyrosine Kinases", were significantly enriched in a larger set of drugs, consistently with their MOAs (Fig. 4E, Supplementary data 10). Notably, these pathways are largely overlapping with recurrently enriched pathways in GDSC drug models (Fig. 3D). Intriguingly, while several enriched MOA-pathways also match processes that are recurrently enriched in GDSC oncology drugs, we found a bigger proportion of pathways exclusively enriched only for PRISM drugs, suggesting a broader diversity of mechanisms of actions (Fig. 4F, Supplementary Fig. 3A).

MOA-primed models improve drug-sensitivity predictions

We leveraged the information of curated MOA-pathways to select the most informative variables (i.e., gene expression) for a given drug to develop knowledge-driven models (hereinafter referred to as MOA-primed models).

For model priming, we considered MOA-pathways generated via GPT-4 as well as through the freely available Mixtral Instruct 8x7b model. Models obtained through the latter approach achieved higher performances (Fig. 1B, C) and are discussed further below.

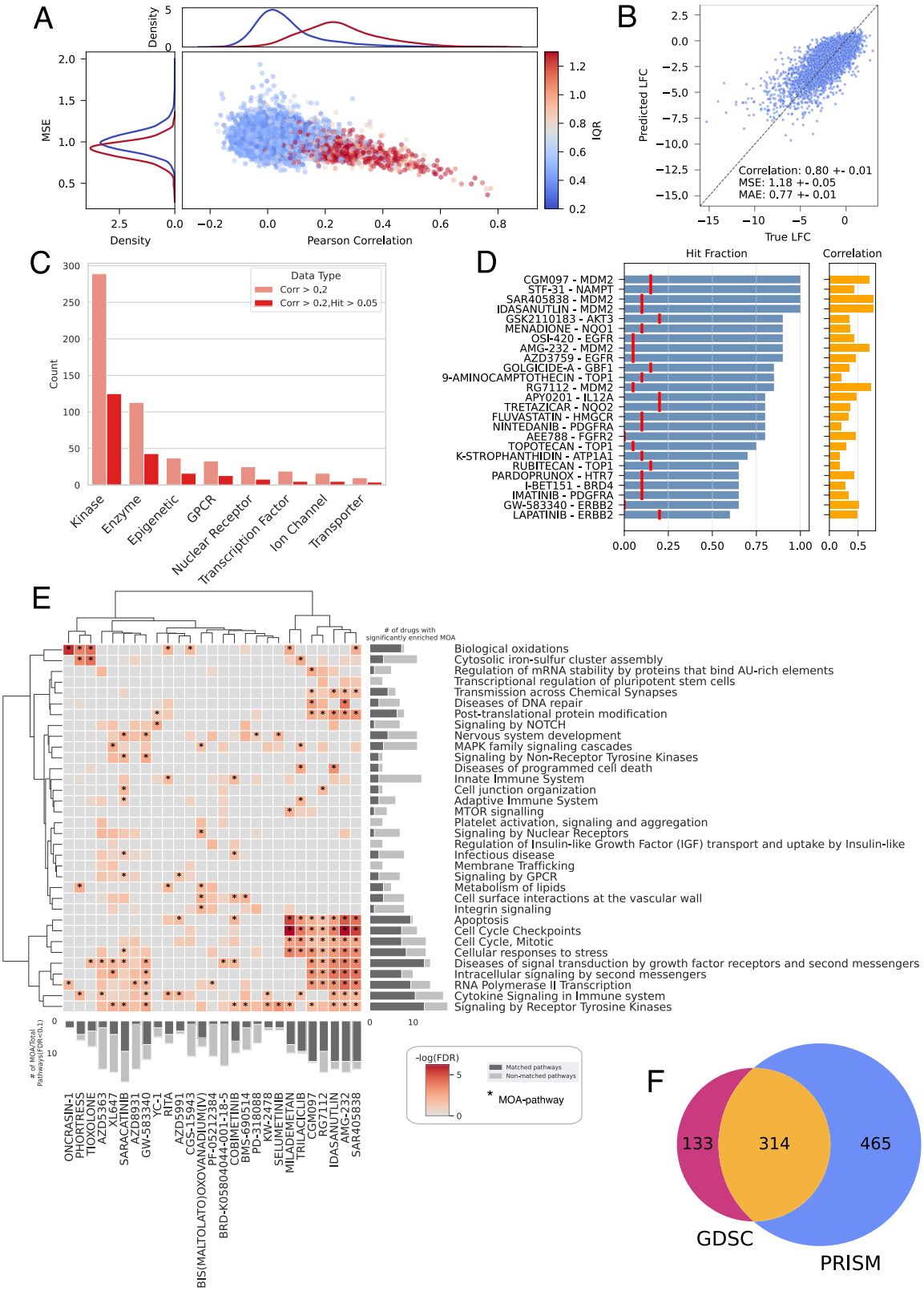
On GDSC, MOA-primed models performed overall better than their All-genes counterparts, with median $\rho = 0.5$ (Fig. 5A). Overall, aggregated IC50 predictions have a high correlation with experimental values ($\rho = 0.89$) and the lowest MSE (1.52) (Fig. 5B). Certain drug models, such as Bicalutamide, showed a remarkable increase in their performance, by almost doubling their correlations when considering only MOA-pathway genes as dependent variables (Fig. 5E). On the other hand, a few drug models, such as BX795, Gemcitabine or Savolitinib displayed a reduction in performance of the MOA-primed models compared to All-genes models (Supplementary Fig. 4).

Also for PRISM, we employed MOA-pathways to select genes to train MOA-primed models, which outperformed All-genes models. Notably, we obtained almost twice the models with $\rho > 0.2$ compared to All-genes counterparts (1254 vs 762; Supplementary data 7). When considering only drugs with IQR > 1, we obtained a median $\rho = 0.32$ (Fig. 5C) and a correlation of aggregated IC50 predictions with experimental values $\rho = 0.93$ (Fig. 5D). Also in this case, several drugs (e.g., Rolapitant) displayed great improvement in the performance of the MOA-primed model with respect to the All-genes counterpart (Fig. 5F).

CellHit inference on aligned bulk RNAseq from TCGA patients data recovers cancer type-specific mono and combination therapies

We deployed our model to forecast effective drug treatments for TCGA tumors based on their transcriptomic profiles. We first compiled a list of FDA drugs approved for specific cancer types (i.e., https://www.cancer.gov/about-cancer/treatment/drugs/cancer-type) and matched them to the corresponding cancers in TCGA (Supplementary data 11). This yielded a total of 41 GDSC drugs approved for 23 cancer types. For each drug, we ranked the top *k*

https://doi.org/10.1038/s41467-025-56827-5



predicted clinical samples according to two criteria: either predicted log IC50 or quantile score (Fig. 6A; Supplementary data 12). The latter is a quantitative measure that trades-off between efficacy and selectivity of a drug, i.e., how much a given drug is predicted to be potent for a particular sample relative to all the other samples inferred (see Methods). We determined the number of top k samples to consider (i.e., 600) as the one that optimized metrics related to the binary classification of drug prediction matching cancer type prescriptions (see Methods; Supplementary Fig. 5). In general, both metrics worked well in prioritizing patients with matching cancer types (Fig. 6B). For certain drugs, such as Cytarabine, Venetoclax and 5-azacytidine, we achieved excellent recall statistics for the cancer type for which the drug is prescribed (Fig. 6B). Overall, we found that 37 out 41 (90%) GDSC drugs' models found, among the top 600 ranked patient

https://doi.org/10.1038/s41467-025-56827-5

Fig. 4 | PRISM model performance and interpretation. A scatter plot of PRISM drug-specific model correlations and MSEs. Dots are colored according to IQR values, ranging from blue to red for increasing values of IQR. The density plots located at the top and left of figure compares the distribution of correlation and MSE values between all models (in blue) and models with an IQR greater than 1 (in red); **B** scatter plot of predicted vs experimental IC50 values from models with correlation ρ > 0.2; **C** barplot statistics of PRISM models with corr ρ > 0.2 (salmon) and corr ρ > 0.2 and target recovered, stratified by putative target protein families (red); **D** fraction of target recovery for the top 25 ligand-target pairs (same plot as 2 A but for PRISM). Red lines represent the 95th percentile of the Hit Fraction distribution across all genes for a given drug. The length of the bars on the right shows the median pearson correlation for each drug-specific model; **E** heatmap showing significant MOA-pathways (rows) for drug models (columns) with corr ρ > 0.5 (same plot as 3E but for PRISM); **F** Venn diagram comparing MOA-pathway significantly enriched on PRISM's and GDSC's drug model important genes. Source data are provided as a Source Data file.

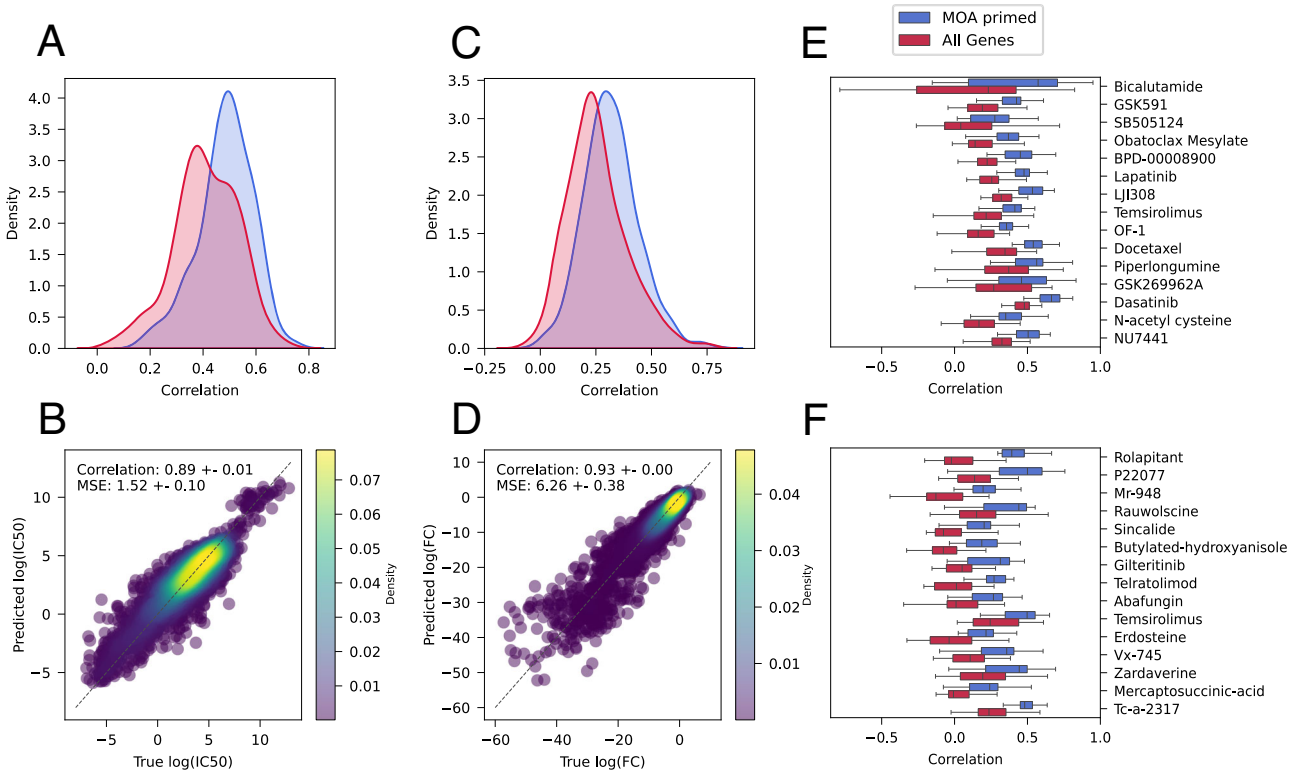
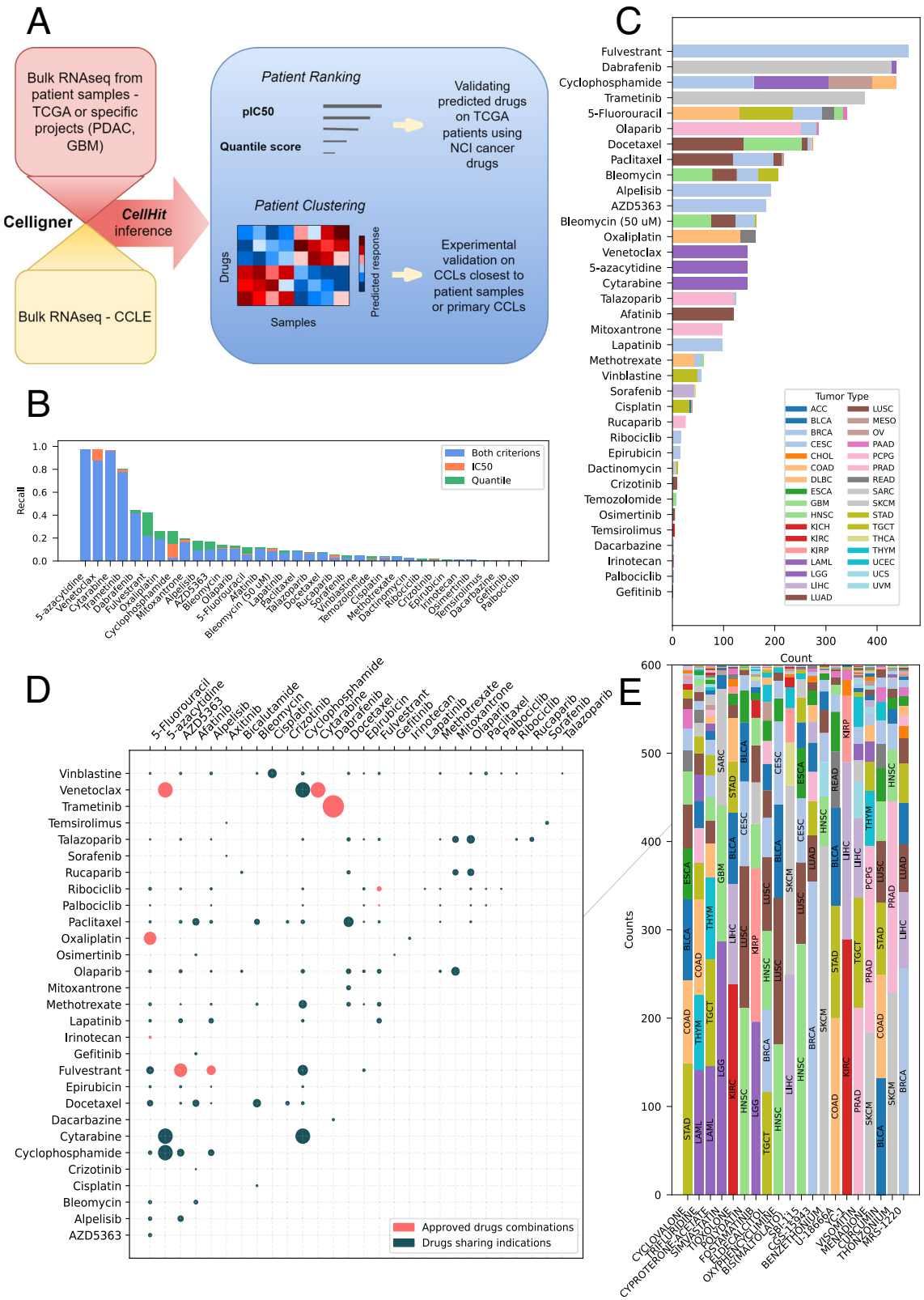


Fig. 5 | MOA-driven models. A comparison of the correlation distribution of all-genes (red) vs MOA-primed (blue) GDSC models; **B** scatter plot of predicted vs experimental IC50 values from GDSC MOA-primed models. The coloring of the dots on the scatterplot indicates the density of points around a particular area; **C** comparison of the correlation distributions of all-genes (red) vs MOA-primed (blue) PRISM models (considering only models with IQR > 1); **D** scatter plot of predicted vs experimental IC50 values from PRISM MOA-primed models with correlation ρ > 0.2. The coloring of the dots on the scatterplot indicates the density of points around a particular area; **E** boxplot of correlation distributions for GDSC models showing the greatest correlation improvement of the MOA-primed vs all-genes models. Each box plot represents a specific drug model and displays the median correlation (central line), interquartile range (box edges), and variability outside the upper and lower quartiles (whiskers); **F** boxplot of correlation distributions for PRISM models showing the greatest correlation improvement of the MOA-primed vs all-genes models. Boxplots follow the structure of (5E). Source data are provided as a Source Data file.

samples, at least one from a cancer type for which the corresponding drug is approved (Fig. 6C). For several drugs, most of predicted samples are derived from cancer types for which that drug has been developed. For instance, Fulvestrant (an anti-estrogen) in Breast Cancer (BRCA), BCL2 inhibitors and Cyclophosphamide in Breast Cancer or Acute Myeloid Leukemia (LAML), the mutated BRAF inhibitor Dabrafenib and the MEK1/2 inhibitor Trametinib in skin cutaneous melanoma (SKCM) (Fig. 6C). Since it is known that Dabrafenib is only effective against BRAF V600E mutated tumors, we looked if genes involved in the drug's MOA were deemed important by the model. However, consistent with the notion that is the mutation in BRAF and not its absolute expression level to determine sensitivity to Dabrafenib, we didn't find BRAF among the important genes. Moreover, none of the BRAF-associated pathways were found among the significantly enriched ones. On the other hand, we found that BRAF mutations are the most recurrent among the top 600 patients scored by the Dabrafenib model (Supplementary Fig. 6A), confirming that the model can infer the mutational status from gene expression signatures. While the majority of the best scored samples with BRAF mutations are affected by melanoma, as expected, we also ranked several BRAF mutated samples from other tumors, such as thyroid carcinoma (THCA) or Diffuse Large B-cell Lymphoma (DLBC) (Supplementary Fig. 6B), confirming the repurposing potential of Dabrafenib in more rarely BRAF-mutated tumors[24]. We also observed in most other drug models that among the top 600 ranked samples, several came from cancer types for which the drug is not currently prescribed (Supplementary Fig. 7), suggesting additional possible repurposing candidates.

A total of 10.5k samples across 33 different cancer types were ranked among the top 600 scoring ones by multiple drug models, suggesting the potential for combination therapies (Fig. 6D; Supplementary data 13). We inspected the predicted drug combinations and ranked them according to the number of predicted samples. This analysis showed that many of the top-ranking combinations are already approved, such as Trametinib and Dabrafenib in SKCM, Venetoclax and Cytarabine or 5-azacytidine in LAML, Fulvestrant and AZD5363, Alpelisib, Ribociclib or Palbociclib in BRCA, as well as Oxaliplatin and 5-Fluorouracil in colon adenocarcinoma (COAD) (Fig. 6D; Supplementary data 14). We found additional predicted combinations

https://doi.org/10.1038/s41467-025-56827-5



characterized by shared indications, highlighting the potential for combination therapy (Fig. 6D; Supplementary data 14).

Likewise, we used the models trained on the PRISM dataset to infer drug sensitivities for each TCGA sample by considering the top 600 samples ranked by each drug model. We inlcuded only the top 20 best performing models for non-oncological drugs, including 8 for Enzymes and 6 for GPCRs (Fig. 6E). The analysis revealed cancer-type specific patterns among the best scoring samples for each drug. For instance, two Adenosine Receptors antagonists, i.e., CGS-15943 and MRS-1220, which have been recently proposed as effective therapies against multiple cancers[25,26], are indeed predicted for multiple samples of breast (BRCA), liver (LIHC), prostate (PRAD) as well as gastric (STAD) cancers (Fig. 6E).

Hence, we demonstrated that the CellHit models can reliably predict anti-cancer therapies based on patients' transcriptomic signatures.

Fig. 6 | Drug sensitivity predictions on TCGA samples. A Schema of a drug response prediction workflow leveraging Bulk RNA sequencing data from TCGA patients, as well from new PDAC and GBM cohorts. The process begins with data harmonization using Celligner, followed by drug inhibitory concentration (IC50) predictions through CellHit. Patients are then ranked by their predicted predIC50 values and quantile score to assess drug efficacy. Validation involves comparing TCGA predictions with NCI cancer drug metadata and refining tumor-specific predictions by clustering patient responses within cancer subtypes for experimental validation.; **B** recall of the recovered drug indications, from NCI cancer drugs, for the TCGA best ranked samples (top 600), according to either predicted IC50 (predIC50) or quantile score metrics; **C** stacked barplot of the GDSC drugs scoring among the top 600 samples patients with cancer types matching the prescription according to NCI drugs. The height of the barplot's stacks corresponds to the number of unique samples and the color of the specific cancer type; **D** circle plot showing drugs predicted for the same pool of patients, i.e., suggesting combination therapies. Circle diameter is proportional to the number of unique samples, among the top 600, best scoring for both drug models. Colors indicate the level of support for that combination, i.e., approved (red) or sharing indication for the same cancer type (dark green); **E** Inference on TCGA data for the 20 best performing non-oncological drug models in the PRISM dataset. The height of the barplot's stacks corresponds to the number of unique samples and the color of the specific cancer type (highlighting potential drug repurposing opportunities). Colors are shared with panel C. Source data are provided as a Source Data file.

CellHit predictions identified drugs selective for distinct PDAC subtypes

Next, we determined whether CellHit could infer possible drugs with specific effects against selected pancreatic ductal adenocarcinoma (PDAC) subtypes recently identified using a laser microdissection-based spatial transcriptomics approach[27].

We first verified the projection of the PDAC samples with the cancer cell lines from CCLE database. As expected, PDAC samples were mapped to the tissue of origin at the transcriptome level (Supplementary Fig. 8) while they displaying different profiles in terms of cancer cell lines responsiveness to drugs (Fig. 7A). PDAC samples assigned to the well-differentiated Glandular (GL) subtype were mainly associated with the esophagogastric adenocarcinoma cell lines and only in a minor fraction to the pancreatic adenocarcinoma ones. Conversely, samples of the Transitional (TR) subtype, which is characterized by gene expression programs suggestive of epithelial-mesenchymal transition, were mainly associated with invasive breast carcinoma and head and neck squamous carcinoma. This result suggests that PDAC cancer cell lines may exhibit heterogenous responses to drugs[28] and that the closest cancer cell lines are rarely associated with the pancreatic lineage.

We next applied the GDSC-trained version of CellHit to the PDAC samples to predict drugs for which PDAC subtypes exhibited differential sensitivities. Hierarchical clustering based on the predicted IC50 of the available drugs showed that PDAC samples segregated into two main groups according to their subtype of origin (Fig. 7B; Supplementary data 15). Moreover, it also revealed two main clusters of drugs to which PDAC samples showed different sensitivities. In cluster 1, TR samples were associated with resistance to the treatments compared to the GL ones. This finding aligns with the behavior and the poor prognosis of PDACs, in which the Transitional subtype is the most abundant tumor component[27]. Cluster 2 was instead enriched in drugs with different effects on both subtypes likely due to the co-existence of endodermal gene programs in these two PDAC variants[27].

Standard-of-care chemotherapeutic drugs, such as Gemcitabine (cluster 1), were previously described to act mainly on classical epithelial (glandular) cells[28]. As expected, CellHit predicted that the GL subtype was more sensitive to Gemcitabine, since these cells have epithelial gene expression programs, compared with the TR subtype which displays quasi-mesenchymal phenotype (Fig. 7C). This analysis confirms the subtype-specific responses to this drug[28].

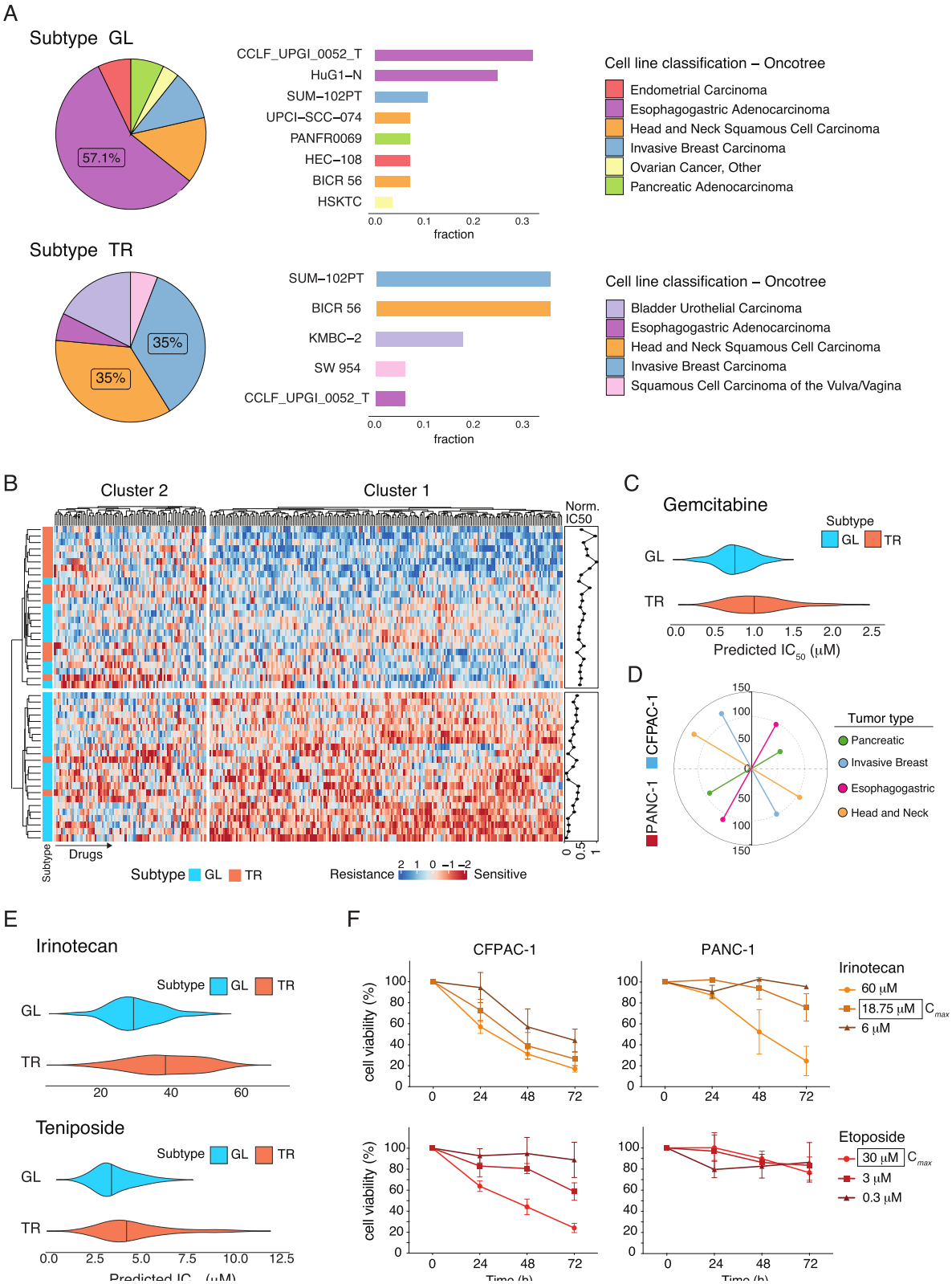
To test these drugs for their repurposing potential, we sought the PDAC cell lines that most closely resembled the cancer cell lines shown by CellHit to have similar drug response profiles to those inferred from patients' samples (Fig. 7A). Celligner revealed that CFPAC-1 was transcriptionally closest to both the pancreatic and the esophagogastric adenocarcinomas while PANC-1 was associated to the invasive breast and the head and neck squamous carcinomas (Fig. 7D), lineages that were previously associated by CellHit to GL and TR subtypes, respectively. Among the drugs enriched in cluster 1 and predicted to be more effective in the GL subtype compared to the TR one (Fig. 7C, E), we also found two topoisomerases inhibitors, such as Irinotecan and Teposide (or its analogous Etoposide), that are approved and used in clinics for cancer therapy[29,30]. To validate our predictions, we treated CFPAC-1 and PANC-1 cells with these drugs at different drug concentrations to test the cell viability over three days (Fig. 7F). Both treatments were preferentially active in the CFPAC-1 cell line with respect to PANC-1 at the dose corresponding to its Cmax, the maximal concentration that can be reached in patients' blood. Overall, the experimental validation was in concordance with CellHit predictions confirming the robustness of the model to infer drug responses in different tumor subtypes and underlining the capacity to repurpose FDA-approved drugs for one of the most lethal solid tumors for which there are no efficient drugs available.

Validation on primary cells from GBM patient's samples the specific response profiles predicted by CellHit

To further assess *CellHit*'s capabilities in a real case scenario, we employed the model trained on the GDSC dataset to infer the drug sensitivity profiles for 64 samples obtained from patients with Glioblastoma Multiforme (GBM). We employed the inferred drug sensitivity profiles to cluster patients, which identified two main groups displaying different sensitivity profiles to drugs (Supplementary Fig. 9). We chose for experimental validation primary cell cultures obtained from the samples Gb130 and Gb107, considered as representatives of the two groups (Supplementary Fig. 9). We tested two compounds, the Mcl-1-specific inhibitor AZD5991 and the E3 ubiquitin-protein ligase XIAP inhibitor, AZD5582, as we found them to be respectively more and less sensitive relative to the median values across samples. When considering only the predicted lnIC50, we predicted for both samples Gb130 and Gb107 higher sensitivities for AZD5582 than AZD5991 (Fig. 8A), which we experimentally confirmed (Fig. 8C, E). Indeed, experimental lnIC50 are in line with the predicted ones, particularly for the Gb130 sample. Predictions of the Gb107 sample differ more compared to experimental results, which anyway confirmed that AZD5582 is more effective in inhibiting the growth of the cancer cell line. Such discrepancy might likely be due to highly specific transcriptional program characterizing this sample and its derived primary cell culture, which is closest to a Leiomyosarcoma cell line based on its response patterns (Supplementary data 16).

By considering the deviation of the predicted lnIC50 values from the median of the distribution of experimental lnIC50 across cell lines, it is possible to observe that AZD5991 had a predicted lnIC50 lower than the median lnIC50 (i.e., 4.591) for Gb130 and higher for Gb107, while AZD5582 had a predicted lnIC50 higher than the median lnIC50 (i.e., 2.014) for both Gb130 and Gb107 samples (Fig. 8B). Such different extents of deviation from the median are also associated to differences in quantile scores for these drugs for the inferred samples. Indeed, while AZD5991 has a bigger difference of quantile scores among cell lines, with higher values for Gb130, AZD5582 is characterized by more comparable values. We therefore proceeded to test the difference in response of the same drug on the two primary cultures simultaneously (see Methods). We observed a greater reduction in viability in Gb130 compared to Gb107 in response to the treatment with AZD5991 (Fig. 8D). This result was consistent with the model, which predicted a lower lnIC50 for GB130

https://doi.org/10.1038/s41467-025-56827-5



(3.05) than for GB107 (5.57). For AZD5582, we did not observe a clear difference in cell viability reduction following treatment (Fig. 8F).

Discussion

In this study, we have developed a ML framework to predict and interpret the sensitivity of cancer cell lines to drug treatments and proposed a strategy to perform robust inference with this model on bulk RNAseq obtained from patients' samples.

Our analysis confirmed that transcriptomics data are a critical component to predict cell line drug sensitivity. Indeed, drug-specific models, developed by considering as features only the cell line expression data, showed similar performances to top performing predictors entailing a joint representation of drug and cell lines. The usage of drug-specific models provided several advantages: it is less memory intensive during training, a crucial factor when dealing with big datasets such as PRISM and can be easily adapted to new datasets

https://doi.org/10.1038/s41467-025-56827-5

Fig. 7 | CellHit predictions on distinct PDAC subtypes and experimental validation. **A** Enrichment of predicted cancer cell lines that react most similarly to the available drugs for the Glandular (GL) and Transitional (TR) subtypes. Classification of the tissue types of cell lines (oncotree system) is shown; **B** Heatmap of predicted IC50 (predIC50) of GDSC drugs derived by CellHit prediction in PDAC samples. K-means clustering (n_clusters=2, method=Euclidean distance) was performed. Subtype annotations are shown for each sample; **C** Violin plot showing the predicted IC50 (predIC50) of Gemcitabine for the Glandular (GL) and Transitional (TR) subtypes; **D** Euclidean distance derived from Celligner between each PDAC cell line (CFPAC-1, PANC-1) and each selected tumor type (Pancreatic, Esophagogastric, Invasive Breast, Head and Neck); **E** Violin plot showing the predicted IC50 (predIC50) of Irinotecan and Etoposide for the Glandular (GL) and Transitional (TR) subtypes; **F** percentage of cell viability of CFPAC-1 and PANC-1 cells treated with increasing concentrations of Irinotecan or Etoposide at 24, 48 and 72 h. Individual values represent the average of three independent experiments ±SD. Source data are provided as a Source Data file.

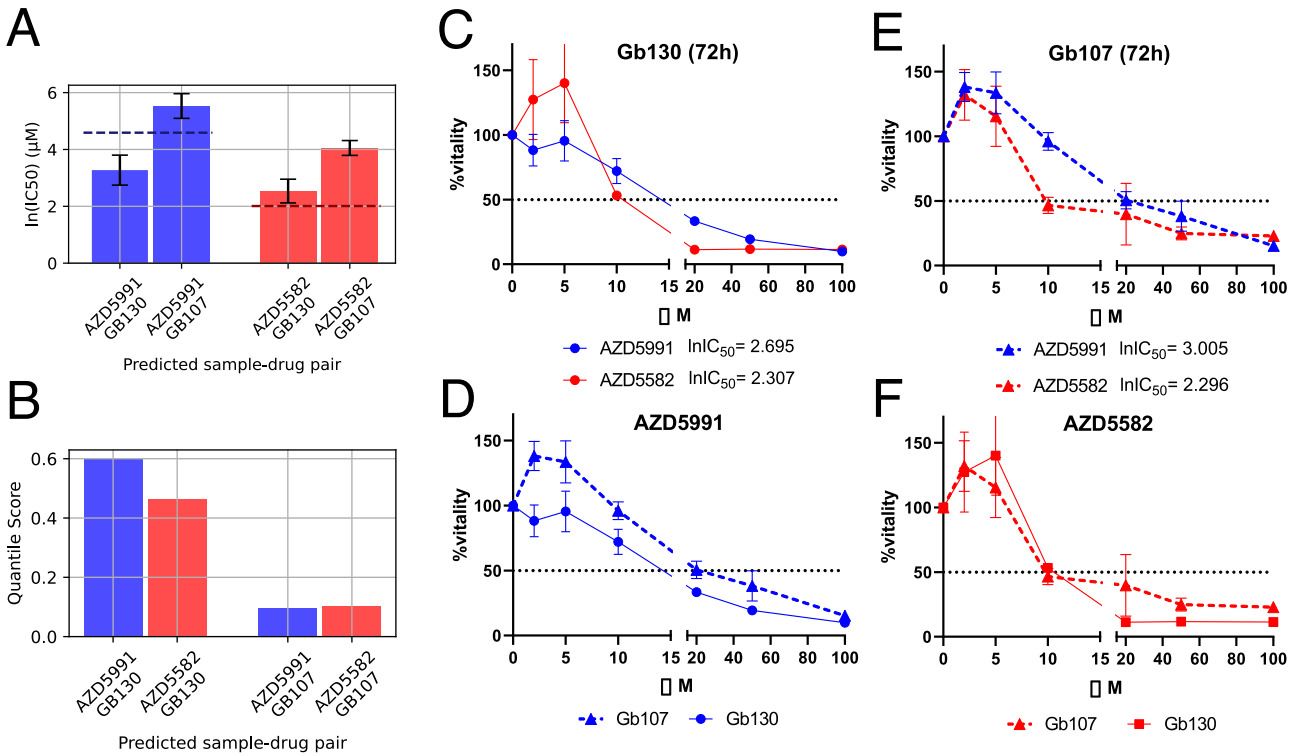


Fig. 8 | GBM patient samples inference and experimental validation. **A** predicted lnIC50 of drugs AZD5991 (blue) and AZD5582 (red) for samples Gb130 and Gb107. Horizontal dashed lines indicate the median of the experimental lnIC50 of each drug on GDSC cell lines. Error bars are obtained by computing SD across models in the ensemble to estimate the overall model uncertainty; **B** predicted quantile scores of drugs AZD5991 (blue) and AZD5582 (red) for samples Gb130 and Gb107; Dose-response curves for the two patient-derived primary cultures of GBM, **C** Gb130 and **E** Gb107 using the Crystal Violet assay. The curves illustrate the response to treatment in terms of reduction of cell viability with AZD5991 and AZD5582, measured 72 h post-treatment for Gb107 (**C**) and Gb130 (**E**). The IC50 values for both drugs are provided. The dose-response curves are also compared across the two cell lines for AZD5991 (**D**) and AZD5582 (**F**). The AZD5991 curve is represented by blue lines, while red lines are used for AZD5582. The profiles for Gb107 are depicted with dashed lines and triangles marking the points, while those for Gb130 are represented with solid lines and circles marking the points. The threshold for a 50% reduction in cell viability is indicated by a dark dashed line. Error bars on plots (**C**, **F**) are computing considering data from assys performed three times in triplicates. Source data are provided as a Source Data file.

and drugs. Most importantly, our XGBoost-based model for individual drugs provides an easy mean to interpret the predictions and explain which genetic expression features are leveraged by the model. As highlighted by ref. 5, numerous drugs display selective activity profiles. Variations in the efficacy of drugs might reveal key molecular mechanisms that underpin specific cancer types. Insights gained from the interpretability of our models shed light on these nuances, enabling a deeper understanding of drug interactions with genetic expression profiles. This knowledge can be pivotal in the development of targeted therapies within a personalized medicine framework, where understanding the particularities of drug responses is essential for tailoring treatments to individual patient profiles. The dual criterion that we have employed for determining feature importance, i.e., SHAP and permutation importance, sets a very stringent requirement for the identification of important expression signatures. Despite this, we found that many drug models are indeed learning the mechanisms responsible for cell-line drug sensitivity solely based on cell-line basal transcriptomics data.

We employed a strategy based on a freely available LLM, i.e., Mixtral Instruct 8x7b, to improve the MOA description for each drug and use it to associate semantically closest pathways from a reference knowledge base. Through this resource, we assessed that many of the All-genes models are characterized by significantly enriched pathways matching the MOA for the corresponding drug, in addition to frequently identifying the nominal targets. Overall, 135 GDSC drug models out of 253 can recover either the target or MOA-related pathways among important genes, suggesting that drug-specific models leverage known biological mechanisms to carry out their prediction task. We also demonstrated that the important genes of the model successfully recapitulated cancer tissue essential genes from a recently published compendium[23]. Our models' interpretation clearly suggests that drug sensitivity emerges from the interplay of drug MOAs and genes essential for cell survival. We speculate that the higher the degree of interaction between MOAs and essential genes, the more effective a drug is in inhibiting cancer cell growth. Our purely data-driven approach to explain drug sensitivity mechanisms can be considered as a complementation of previous approaches, either based on

simple correlation of sensitivity and basal gene expression[4], or integrated with PPI network and pathway analysis[31].

The proposed LLM pipeline extracts features in a human-readable and interpretable manner, and represents an example of how to use powerful LLMs by leveraging relevant domain knowledge. The growing 'reasoning' capabilities of future models could further improve the capabilities of the proposed approach by leveraging multi-modal contents, such as images and knowledge graphs and similar applications of LLMs are already appearing in the literature (e.g., https://biochatter.org/[32]). The usage of novel strategies to mitigate LLMs' hallucination (e.g., ref. 33) will be critical to systematically assess the predicted outputs and reliably incorporate them in knowledgebase annotation procedures.

In our study, we first tailored the training and interpretability strategies to model the GDSCv2 dataset, which has also been extensively tackled by a multitude of previous methods[7–9]. We then applied the same strategy to the PRISM dataset, which consists of a much larger panel of drugs screened against cancer cell lines. Although the PRISM dataset poses specific challenges, such as many drugs showing little or no effect at all on cancer cells, we showed that several models achieved good performances and recovered the information of targets and associated MOAs, which encompass a broader range of biological processes compared to GDSC drugs. This ML model tackled drug sensitivity predictions on both GDSC and PRISM through an explainable framework. However, it is possible that the kind of regression model that we are employing here might not be best suited for many of the drug candidates that we have observed in PRISM, characterized by highly specific activity profiles (in other words, showing activity on a limited number of cells). Considering the high anti-cancer potential of many non-oncological drugs against certain targets (e.g., GPCR drugs)[26,27,34], it will be interesting to explore in the future alternative ML frameworks to model the sensitivity of those drugs for which we obtained low performances due to high specificity and low variability of the response.

To deploy the models in real world scenarios, i.e., on bulk RNAseq data obtained from patients, we employed Cellligner[18] to align bulk RNAseq from patients to those from cell lines upon which the model has been trained. We used Celligner's transformed RNAseq data both to train the model (i.e., CCLE) as well as to perform inference on more than 10k patients' samples from TCGA to predict an IC50 for each drug. Best predicted drugs for each patient often matched mono- and combination-therapies approved for the corresponding cancer type, such as Venetoclax AML, Fulvestrant in BRCA and Dabrafenib in SKCM, along with multiple combination drugs whose usage has been already approved. These results support the high potential for translation of our model predictions, as we find many more predicted mono- or combination-therapies for many TCGA cancer types, with some evidence of indications for combined usage, which might represent new repositioning opportunities.

To further validate our strategy, we transformed RNAseq data from different morpho-biotypes recently identified for PDAC[27] and inferred the most likely drugs for the samples of distinct subtypes. We showed that Irinotecan and Etoposide, predicted to be more effective against the "GL" than "TR" biotypes, indeed displayed differential sensitivity on cell lines more closely resembling the different tumor types reacting most similarly to the two PDAC subtypes, corroborating our predictions. The high intratumor heterogeneity, namely the coexistence in the same patient of heterogeneous groups of tumor cells with distinct morphological and transcriptional profiles, may lead to the adaptation of these cells to therapy. Thus, this approach could be important to design ad hoc combinatorial therapies based on the tumor's subtype composition. We also demonstrated on GBM patients' tumor samples the capability of our model to exploit predicted drug sensitivity profiles to cluster samples based on their similarities in drug sensitivity profiles. Through this approach we have identified representative samples with specific sensitivities, which we experimentally validated using match patient-derived tumor primary cell lines. These additional validations not only strengthen the reliability of our model but also highlights its potential translatability in clinical practice providing the therapeutic field of glioblastoma, which has remained stagnant for years, with a broad spectrum of potential new therapeutic possibilities.

These results pave the way for future exploitation of the CellHit pipeline for inference using larger sets of drugs (i.e., PRISM) on alternative patient cohorts. In this respect, it will be critical to develop faster algorithms to align bulk RNAseq for inference with the model. The current strategy, based on Celligner, requires an initial alignment between CCLE, TCGA and any additional input RNAseq dataset, and subsequent retraining of the model on the transformed CCLE data. We plan to employ deep learning architectures, such as Variational Auto Encoders (e.g., Mober[35]), to improve this preliminary alignment step which is of critical importance to effectively deploy cell line-based models on patient samples. We will provide our models via a webapp for fast drug-sensitivity inference of bulk RNAseq data from patient samples provided as input. This will allow to analyze and compare inputted samples based on the similarity of their "responsiveness" profile, in addition to the one based on transcriptomic profile, which will speed up the hypothesis generation process to find new personalized treatments against cancer.

Methods

The research was conducted in compliance with the principles outlined in the Declaration of Helsinki, and the protocol for sample collection received approval from the Ethics Committee of the University Hospital of Pisa (787/2015).

Datasets

Transcriptomics, IC50s and LFCs. We obtained data model's training from two comprehensive high-throughput studies: the Genomics of Drug Sensitivity in Cancer (GDSC)[3] and the Profiling Relative Inhibition Simultaneously in Mixtures (PRISM)[5]. We downloaded the GDSC2 dataset, release version from 24 July 2022, from its official website (https://www.cancerrxgene.org/). This dataset consists of 969 cancer cell lines profiled for their responses to a panel of 286 drugs. Our primary focus within this dataset are the half-maximal inhibitory concentrations (IC50) values, which serve as target values for predictive modeling. 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 and consisting of 919 cell lines and 6.415 compounds. We also retrieved drugs metadata available both in GDSC and PRISM regarding drug putative targets and mechanism of action. We obtained RNA sequencing (RNASeq) data from the Cancer Cell Lines Encyclopaedia (CCLE)[2] as available from DepMap. We considered log2-transformed transcripts per million plus one (TPM + 1) data for protein-coding genes. We mapped CCLE cell lines data to GDSC by using available COSMIC to DepMap identifier mappings which allowed us to link cell lines to additional resources available on the DepMap portal such as mutations, gene essentiality, and additional metadata. We employed DepMap IDs to sort cell lines into different tissues and disease categories via the OncoTree classification system[36].

TCGA data. We integrated into our pipeline RNA bulk transcriptomic data from TCGA[1] which we used for initial validation of the designed models on actual patient-derived samples. 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. This

data, sourced from the UCSC Treehouse Public Data platform, is downloaded in the RSEM log2(TPM + 1) normalized format.

Gene essentiality data. We retrieved data from the 23Q4 CRISPR Gene Dependency dataset from the DepMap portal. This dataset contains synthetic lethality experimental data across 1100 unique cell lines and 18,444 genes. We refined our analysis to a subset of 17,425 genes, which are also represented in both the CCLE and TCGA databases. Matching and integration of this dataset with cell lines from GDSC and PRISM is made through DepMapID.

Mutations data. We obtained somatic mutations data from the 23Q4 release of the DepMap portal. This dataset includes mutation data for 1750 unique cell lines. We considered the mutations of 693 high-consensus oncogenes and tumor suppressor genes as curated in the OncoKB database[37]. We binarized oncodriver genes as mutated or not mutated regardless of the specific mutation type. We cross-referenced this dataset with cell lines from GDSC and PRISM via DepMapID.

Reactome. Pathway data from the Reactome database[21] is incorporated into our analysis. Notably, we leveraged the directed acyclic graph representing the hierarchical structure of the pathways alongside a comprehensive list of pathways and their associated genes. Data extraction from Reactome is executed through two main methods: file dumps from Reactome, which provide us with the list of pathways and their hierarchical organization, and the REST API, which is used to obtain information on pathway-associated genes and the drugs that have been manually annotated to these pathways. The API is queried programmatically using the "get" function from Python's built-in "requests" library. We employed topological sorting to organize the nodes (i.e., pathways) of the Reactome hierarchy, considered as a directed acyclic graph (DAG). We carried out the following steps: 1. We performed a topological sort on the Reactome graph using NetworkX's 'topological_sort' function; 2. we iterated through the sorted nodes, and assigned a layer number to each node based on its position relative to its predecessors; 3. The layer assignment follows this logic:

- If a node has no incoming edges (no predecessors), it's assigned to layer 0.
- Otherwise, the node is assigned to the layer immediately following the minimum layer of its predecessors.

Through this process, we systematically categorized nodes into hierarchical "layers", designed to only have incoming edges from the layer above and outgoing edges to the layer below, hence maintaining the hierarchical structure of the Reactome pathways. Crucially, the use of topological sorting guarantees that when we move to lower level of the DAG, nodes predecessors already have an assigned layer (since assignment happens recursively by looking at them). We focused on pathways within the "1st Layer," selecting a subset of 169 unique pathways from Reactome. Processing of this data is carried out by leveraging Python's networkx library, specifically its "topological_sort" function.

NCI Cancer drugs. We developed a natural language processing pipeline to programmatically identify drugs' clinical indications for the different types of cancer as defined in TCGA. In more depth, this pipeline works by extracting information from textual data. Utilizing the Beautiful Soup (https://www.crummy.com/software/BeautifulSoup/bs4/doc/) package, we retrieved links to drugs listed on the National Cancer Institute's website (https://www.cancer.gov/about-cancer/treatment/drugs). We then combined Python's requests package with Beautiful Soup to extract textual descriptions of each drug. The extracted text is used as input to create a custom textual prompt. This prompt is then fed to an LLM with generation constrained by the Guidance package (see "Mixtral pipeline" section). The goal was to determine whether the text provided any evidence that the drug was prescribed for any of the 37 cancer categories from TCGA. Given the potential discrepancies in drug naming conventions between NCI and GDSC, we recover PubChem IDs from free-text drug names using PubChemPy, a tool developed by the curators of PubChem[38] to programmatically retrieve information about chemical compounds. This step was vital in aligning these drugs with their counterparts in the GDSC, ultimately allowing us to obtain clinical indications for a total of 41 drugs. All data obtained through this pipeline is reported in Supplementary data 11.

PRISM drugs' metadata and IQR analysis. We obtained PRISM drugs categories, i.e., Chemotherapeutics, Targeted Oncology Drugs, and Non-oncology Drugs, from the secondary screening metadata provided by the 2019 PRISM project data release, available through the DepMap portal (https://depmap.org/repurposing/). We transferred the classification to the drugs in the 2023 release of PRISM by matching their BROAD IDs. This yielded a final set of 835 drugs, including 435 targeted therapies, 346 general oncology treatments, and 54 chemotherapeutics. We calculated the Interquartile Range (IQR) for the Log Fold Change (LFC) data across all 6337 drugs represented in the PRISM dataset. It is important to note that, due to the properties of the logarithm in base 2, an IQR greater than 1 implies that the viability counts of the cell line at the 75th percentile after a given drug are at least twice as high as those observed for the cell line at the 25th percentile.

PDAC subtype data. For the drug prediction in PDAC subtypes (Glandular, GL; Transitional, TR; Undifferentiated, UN), data were obtained from the transcriptional profiles of multiple morphological distinguishable tumor areas isolated by laser micro-dissection (LMD) in primary PDACs of treatment-naïve patients[27].

Data pre-processing

Standardization and filtering. We first selected a subset of common genes among the datasets, using the Human Gene Nomenclature Committee (HGNC) (https://www.genenames.org/) system. We removed genes exhibiting zero standard deviation in either of the datasets (CCLE or TCGA), yielding a final set of 18.174 genes. We standardized transcriptomic data by subtracting the mean and dividing by the standard deviation across the whole dataset. This normalization is not particularly important for tree-based models but is important for the stability of neural networks and other models (used as baselines in this work). For the response values (Ys), we conducted a drug-by-drug standardization, by removing for each drug the mean and standard deviation computed specifically for that drug. This standardization ensured that the model's predictions were not skewed by the inherent differences in the drug response scales but were instead sensitive to the nuances in response patterns specific to each drug, considering that different drugs can have varying ranges and distributions of response values. The standardization was performed only on the training set. The splitting strategy used throughout the analysis is described in the "Model testing" section. At the end of the preprocessing step, the GDSC dataset comprised 686 unique cell lines, 286 unique drugs and a total of 169.208 drug-cell line pairs (IC50 values). The PRISM dataset comprised 887 unique cell lines, 6337 unique drugs and a total of 3.810.028 drug-cell line pairs (LFC values).

Alignment of patients' bulk RNAseq with Celligner. To harmonize representations of bulk RNAseq transcriptomic profiles from cancer cell lines and patient-derived samples, we adopted the pipeline proposed by the Celligner method[18]. We employed the data from the CCLE and the Tumor Compendium v11 Public PolyA (see "Datasets" section). Unlike the Celligner publication, in our study we also integrated, together with TCGA, bulk RNASeq data obtained from two additional

patient cohorts with pancreatic ductal adenocarcinoma (PDAC) and glioblastoma multiforme (GBM)(see below). This integration is achieved by subsetting all data based on a common set of available genes, concatenating experimental samples with TCGA data, and then executing the cell alignment. We used the Python version of Celligner, available through its GitHub repository (https://github.com/broadinstitute/celligner/tree/master).

Drugs and cells featurization

We featurized the drugs using their two-dimensional structures. Since the GDSC dataset does not provide the structural identifiers for the tested drugs, we used PubChemPy to retrieve the SMILES representations for each tested compound, for a total of 229 unique drugs. We employed the Extended-Connectivity Fingerprints (ECFP)[39] and ChemBerta fingerprints[40] for drug featurization. Both ECFP and ChemBerta fingerprints were computed using the MolFeat (https://molfeat.datamol.io/) library. As a control, we also introduced a OneHot embedding technique. This approach produced a 229-dimensional vector for each drug, with the ith position marked as '1' for the *i*th drug and '0' in all other positions. In parallel, we also explored alternative featurization strategies for the cell lines. On one hand, we considered the log2 + 1 normalized expression values of a total of 18.174 genes. We also explored a dimensionality reduction technique, specifically Principal Component Analysis (PCA), retaining eigenvectors describing 90% of the total variance (395 components). PCA analysis is performed through the "PCA" function from Scikit-learn library (ref. 41 and fitted on the log2 + 1 normalized expression values.

MOA-pathway annotation

To identify pathways potentially involved in the mechanism of action (MOA) of a certain drug, we implemented a pipeline with three different attribution criteria: an LLM attribution criterion, a pathway membership criterion and a manual Reactome pathway annotation criterion. Below we provide detailed explanations for each of these three criteria.

Extending drug MOA with LLM

We employed two distinct Large Language Models (LLMs) for the intelligent extraction and interpretation of drug-related information: GPT-4, a proprietary model developed by OpenAI (https://openai.com/chatgpt), and Mistral Instruct, which is freely available and developed by MistralAI (https://mistral.ai/). Information extracted from the LLMs was leveraged to identify biological pathways likely influencing drug efficacy, which can then be used to select on which genes a model should be trained. The full list of curated pathways for each criterion is reported in Supplementary data 3 and S9 for GDSC and PRISM respectively.

GPT-4 based pipeline. The GPT-4 pipeline begins with the extraction of GDSC's drug metadata. These primarily include short textual labels describing the mechanism of action or the putative target of each drug (253 drugs out of the 286 available in GDSC). Utilizing a specialized prompt, we engaged GPT-4 to expand this basic metadata into a comprehensive textual description, which elucidates the drug's mechanism of action and its metabolic pathways in greater detail. We then used the expanded drug description as a second specialized prompt to task GPT-4 with the identification of the top 15 biological pathways likely to modulate the drug's efficacy as well as to provide a reasoned explanation for the selection of each pathway. The model's ability to elucidate its reasoning offers valuable insights into the complex interplay between drugs and biological systems, significantly augmenting our understanding of drug responses and offering an entry point to possibly debug the pipeline. Both of these steps are implemented using the OpenAI API, with a specific emphasis on the "function calling" feature introduced in July 2023. This feature is pivotal as it enables us to receive responses in a structured JSON format, adhering to a pre-specified schema and allowing integration in the data pipeline. The selection of pathways by GPT-4 is constrained and informed by the Reactome knowledge base, specifically it is forced to select pathways only among "Level 1" pathways (see "Reactome preprocessing"). This hierarchical approach not only provided a comprehensive overview of biological interactions at various levels of complexity but also allowed us to effectively control the scope of pathways among which GPT-4 operates, balancing the complexity and operative cost of our pipeline.

Mixtral pipeline. We introduce an AI curation strategy employing Mixtral Instruct[42], a freely available instruction-tuned 8×7 billion parameter mixture of experts LLM. Notably, the usage of a freely available architecture allowed us to devise a cost-effective and reproducible methodology, addressing computational and accessibility challenges. Our approach guarantees predictive accuracy on par with GPT-4 (or even exceeding) by leveraging three core strategies: structured Chain-of-Thought prompts[43] for detailed reasoning, Self-Consistency[44] procedures for the minimization false positives, and Retrieval Augmented Generation (RAG)[45] for the integration of validated scientific literature.

Operatively the pipeline unfolds across three phases: initial drug description generation based on metadata, refinement of these descriptions with RAG leveraging PubMed abstracts, and the selection of biological pathways through a self-consistency approach. Initially, drug descriptions are generated using detailed prompts that elicit step-by-step reasoning from Mixtral Instruct, mirroring the GPT-4 pipeline but with enhanced specificity. To further diminish the likelihood of model hallucinations, descriptions are refined by integrating a RAG strategy through the Entrez module of the biopython python library[46]. This involves systematically retrieving and synthesizing information from the top 10 PubMed abstracts (sorted by relevance) related to each drug. Subsequently, the LLM is tasked with refining these initial descriptions, specifically prioritizing information from the abstracts in the event of conflicting statements and seamlessly integrating any relevant, missing knowledge. In the pathway selection phase, we introduce a self-consistency methodology where Mixtral Instruct is tasked multiple times (with different random seeds) to identify the most relevant biological pathways influencing drug efficacy. By considering pathways identified at least twice across multiple iterations, we significantly diminished the risk of false positives. Due to the absence of function calling capabilities in Mixtral Instruct, we interleave "generate" and "choice" functions from the Guidance library (https://github.com/guidance-ai/guidance) to impose constraints on the LLM's generative process and obtain structured outputs from the LLM. The use of Guidance not only allows for a controlled selection among predefined pathways but also ensures that the rationale behind each choice is generated before the pathway itself. This enhances the causal coherence of the model's output and leverages its autoregressive nature.

Operationally, we deploy a GPTQ 4-bit quantized version of Mixtral Instruct, sourced from the Huggingface model Hub (https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.1-GPTQ). This quantized model configuration strikes a balance between model precision and inference speed, significantly reducing computational costs. Crucially, its compatibility with NVIDIA V100 GPU cards, requiring less than 30 GB of VRAM, enabled us to conduct extensive deployment within our High-Performance Computing (HPC) cluster efficiently. Furthermore, the use of the vLLM[47] library facilitates continuous batching and maximizes GPU utilization during inference for drug description generation and refinement. For the task of pathway selection, we integrated the Huggingface transformers library[48] with the Guidance framework to adhere to Guidance requirements. All prompts utilized throughout the Mixtral Instruct pipeline, alongside the project's code, will be publicly released.

Extending drug MOA with target pathways and drug annotations in Reactome

We also linked drugs with potential Reactome pathways based on their putative targets. Utilizing the Reactome API, specifically the "referenceEntities" endpoint, we extracted entities classified under the ReferenceGeneProduct class for each pathway of "Level 1" (see Reactome Preprocessing). This process yielded a comprehensive list of gene products associated with each identified pathway. A pathway was considered relevant to a drug if its putative target was among the pathway's gene products. We extracted known ligand associations to Reactome's pathways, by using the Reactome API's "referenceEntities" endpoint. Specifically, we retrieved for each of the "Level 1" pathways the objects assigned to the ReferenceTherapeutic class. This step provided a nested list of compounds along with their common names and corresponding identifiers from either PubChem or the Guide to Pharmacology for each pathway.

To merge annotations, we re-employed the PubChemPy pipeline introduced in the "Drugs and cells featurization" chapter. For those instances where the Guide to Pharmacology was the initial source, this pipeline converted identifiers to PubChem format. Since we had previously acquired PubChem IDs for molecules within the GDSC dataset, this streamlined the integration process. Through this approach, we successfully associated pathway information with 66 distinct drugs present in the GDSC. We compared retrieved pathways using the three distinct attribution criteria using the Python matplotlib_venn library (https://github.com/konstantint/matplotlib-venn).

LLM pathway recovery probability. We assessed the specificity of the LLM in identifying distinct pathways for various drug categories by utilizing metadata from the PRISM 2019 release (refer to "Metadata on PRISM Drug Categorization"). Our examination spanned different drug categories, focusing on the pathways highlighted by the LLM. For each drug category, we calculated the probability of an L1 pathway (see Data pre-processing) to be selected as relevant by dividing the number of times that pathways appear in that drug category by the total number of drugs within that category. To consolidate and visually represent our findings, we compiled the 25 most commonly identified pathways across all categories, presenting them in an annotated heatmap depicted in Supplementary Fig. S3.

Model training

We employed a dual-strategy training methodology for our predictive models, each addressing unique aspects of the drug-cell line interaction landscape. The first strategy involved a joint drug and cell-line featurization approach under the hypothesis that a dual representation of both drug molecules and cells would lead to a more nuanced and accurate prediction of how various cancer cell lines respond to different drugs. As a second strategy, we employed a drug-by-drug modeling approach, where we focused on examining the effects of individual drugs on cancer cell lines, with particular attention to the transcriptional responses of the cells. This method allowed us to delve into the specific mechanisms through which each drug influences cell behavior, thereby offering insights into drug-specific interactions and responses.

Validation and hyperparameter optimization

We describe a rigorous approach for validation and hyperparameter optimization essential for extracting maximum performance from deployed models. To obtain a fair comparison between methods, this process is uniformly applied across various models in our study.

The proposed model-based hyperparameter search procedure leverages the Optuna framework[49]. Optuna is a cutting-edge tool for automating hyperparameter optimization and, specifically, deploys a Multi-Objective Tree Parzen Estimator (MO TPE)[50]. This estimator simultaneously optimizes two key metrics: correlation and mean squared error, both evaluated on the validation dataset. By optimizing these metrics concurrently, we ensure a balanced approach to model performance, focusing on both absolute predictive accuracy and the strength of the relationship between predicted and observed values. MO TPE leverages a Bayesian-inspired strategy that is initialized with the evaluation of a specified budget of random hyperparameters.

For the models optimized using Optuna in our study, we adhered to a structured budget for evaluations. This included 100 initial random evaluations, serving as a broad exploration of the hyperparameter space. We then conducted an additional 200 evaluations, which were more focused or "greedy." This approach, totaling 300 trials, is designed to strike a balance between exploring a wide range of possibilities and honing in on the most effective hyperparameters. We ensure transparency and reproducibility of our methodology by providing detailed information about the prior spaces for the tuned hyperparameters of each model.

Baselines from the literature

Our investigation includes a benchmarking of various models that have been recognized as state-of-the-art in the literature[7].

Kernel Ridge Regression (KRR). KRR emerges as a pivotal machine learning technique, offering a sophisticated blend of ridge regression's regularization capabilities with the kernel trick's ability to operate in higher-dimensional spaces. Fundamentally, KRR extends linear ridge regression by incorporating a kernel function, thus enabling the modeling of non-linear relationships without explicitly transforming data into a high-dimensional space. KRR stands as a benchmark alternative for feature selection, distinct from the approaches proposed via Large Language Models (LLMs). This method solely utilizes transcriptomic data, constructing a separate model for each drug without incorporating drug featurizations. We implemented KRR through the Scikit-Learn library.

Similarity-Regularized Matrix Factorization (SRMF). The Similarity-Regularized Matrix Factorization (SRMF) method[7] is an innovative approach for predicting anticancer drug responses in cell lines. It leverages the inherent similarities between drugs and cell lines to enhance prediction accuracy. Specifically, SRMF incorporates chemical structure similarities of drugs and gene expression profile similarities of cell lines as regularization terms in the matrix factorization model. We implemented this model by modifying the original Matlab code available at (https://github.com/linwang1982/SRMF).

Full joint models

Following methodologies that are widely recognized in the literature, we developed multiple predictive pipelines intended to effectively extrapolate and harness meaningful information from both the drugs and the cell lines. Crucially, models in this chapter make use of the featurization described in the "Drugs and cells featurization" chapter.

Multi-layer perceptron. We implemented a customizable multi-layer perceptron (MLP) model architecture. The model dynamically constructs its architecture based on specified hyperparameters, including the number of input features, the number of hidden layers, the number of neurons in each hidden layer, and the dropout rate to mitigate overfitting. Each hidden layer is normalized using batch normalization to enhance stability and employs the ReLU activation function to introduce non-linearity, with an optional dropout applied based on the specified rate. The MLP is optimized using the AdamW optimizer, leveraging hyperparameters such as learning rate and weight decay for regularization. Training involves a customizable loss criterion, defaulting to Mean Squared Error Loss for regression tasks, with early stopping implemented based on a patience parameter to prevent overfitting by halting training if the validation metric does not improve

for a specified number of epochs. Optimal hyperparameters for this model are fixed following the procedure described in the "Validation and hyperparameter optimization" chapter. The full list of hyperparameters and the prior space of the hyperparameters is provided in the supplementary material. The MLP is implemented using the PyTorch library[51].

XGBoost. XGBoost[52] is a highly efficient machine learning algorithm based on gradient boosting, using decision trees to iteratively improve predictions. It includes innovations like regularization to prevent overfitting, efficient tree pruning, scalable handling of missing data, and optimized splitting algorithms for large-scale datasets. Crucially, given the high dimensionality of our dataset, we leverage the GPU acceleration included in XGBoost from version 2.1.0. Optimal hyperparameters for this model are fixed following the procedure described in the "Validation and hyperparameter optimization" chapter. XGBoost is implemented through the official library (https://xgboost.readthedocs.io/en/stable/).

Drug-by-drug models

We delineate the methodology employed for developing a model to predict cancer cell line responses to drugs, focusing exclusively on transcriptomic data. The following methodologies were applied to both the GDSC and the PRISM dataset, resulting in the development of 286 models for GDSC and 6337 models for PRISM. Crucially, we develop two different types of drug-by-drug models: all genes models and MOA-primed models.

All genes models. Given the consistently superior performance of the XGBoost algorithm, as evidenced by our results and a multitude of studies in the literature on tabular data, we have selected it as our primary tool for developing drug-specific models. Details about this model type can be seen in the dedicated "XGBoost" chapter. For the PRSIM dataset, we found that the majority of trained models were characterized by overall low performances (Fig. 4A, blue curve and dots; median correlation $\rho = 0.04$), likely due to moderate and tissue-specific responses of cancer cell lines to the many non-oncological drugs present in PRISM. On the other hand, the drug models characterized by dispersed experimental log-fold changes (LFCs) in their dataset, i.e., with InterQuartile Range (IQR) greater than 1, are characterized by a substantially higher average performance (median $\rho = 0.24$, Fig. 4A, red curve), yielding a total of 713 out of 6337 drug models. Based on this result, we decided to restrict the downstream analysis only to those models with $\rho > 0.2$, as they represent most of those with IQR > 1 (Fig. 4A). A cutoff of $\rho > 0.2$ is moreover consistent with earlier modeling efforts on a previous release of the PRISM dataset[5].

MOA-Primed models. MOA-primed models differ from regular full-gene models in two key aspects: firstly, they are trained exclusively on a subset of genes identified as relevant to the drug's MOA by the three criteria described in the chapter "MOA pathway annotation." Reducing model training to this subset of genes leads to a significant reduction in the number of genes from 18.174 to an average of 4.117, achieving a factor of reduction of 4.4. Secondly, this narrower gene focus allows for a more comprehensive optimization process within a reasonable computation time. In more detail, the MOA-primed model consists of an ensemble of three XGBoost models (each comprising five parallel trees) trained on different subsets of the training data. This approach, along with the reduced number of genes used, greatly decreases the overfitting problem both during hyperparameter optimization and training. As for other models, optimal hyperparameters are fixed following the procedure described in the "Validation and hyperparameter optimization" chapter. Differently from other models, we leverage a 3-fold cross-validation strategy during hyperparameter search. Crucially, each XGBoost model belonging to the ensemble is trained on a different train-validation split.

Model testing

Data splitting procedure. To guarantee a robust evaluation of the predictive performance of our models downstream, we have implemented a strategic approach to partitioning the dataset. This partitioning is guided by the OncoTree classification system, which offers a detailed categorization of cell lines across different tissues involved in cancer. We selected two distinct cell lines from each tissue type specified in the OncoTree classification. This diverse selection ensures a representative cross-section of cancer types, aiding in the generalizability of the model. Subsequently, all drug-cell line pairs related to these selected cell lines were systematically assigned to either the training, validation, or test sets. Such an approach ensures that each set reflects a broad spectrum of tissue types and their corresponding responses. To ensure a reliable estimate of our trained models' performance, we repeated the train-validation-test procedure described earlier 20 times, each with a different random seed. We use random seeds from 0 to 19. The median values of the metrics obtained, along with their standard deviations, are reported.

Aggregated evaluation. We conducted a comprehensive comparison of all trained models by calculating the Mean Squared Error (MSE) and Pearson correlation coefficient for the predicted values across all drug-cell line pairs (encompassing all drugs and cell lines) in both the GDSC and PRISM datasets. It is important to note that baselines, full joint models, and drug-by-drug models adhere to the same standardization schema outlined in the section "Data Pre-processing." As such, before evaluating the metrics on the test set, we adjusted both the experimental and predicted values (Ys) to their original scales on a drug-by-drug basis. The cumulative predictions of the drug-by-drug models resulted from pooling forecasts made by individual models.

Drug specific evaluation. We assessed the models' local performances by computing the median and standard deviation of both MSE and Pearson correlation coefficient across 20 random splits for each drug-specific model. This evaluation is conducted without reverting the standardized data to its original scale. Crucially, this approach assessed a model's predictive power around its mean activity value (either IC50 or LFC). Additionally, we derived a final summary metric by calculating the median and standard deviation of all these single-model performances, providing a concise overview of model efficacy in a drug-centric context.

Model interpretability

Importance computations. We have adopted a stringent criterion to identify genes that are pivotal in our models' predictions of cancer cell line responses to drugs. To determine the importance of genes, we utilized two distinct but complementary methods[53]: SHAP (SHapley Additive exPlanations)[20] importance and Permutation Importance[54]. We imposed that a gene is relevant if both methods yielded importance values greater than zero. Notably, SHAP values provide insights into how each gene contributes to the model's final prediction. This method effectively quantifies the impact of each variable (gene) in the context of the model's decision-making process. On the other hand, Permutation Importance evaluates the influence of each gene on the overall performance of the model. This is achieved by measuring the degradation in model performance when the values of a gene are randomly shuffled, thereby disrupting the gene's original relationship with the response variable. In the results discussed above we chose as reference metric for permutation importance a drop in correlation (we provide in the released code also computations for MSE). TreeSHAP, as implemented in the Python SHAP library, was employed to ascertain the importance of each gene (https://shap.readthedocs.io/). Notably,

this method offers a fast and exact computation of otherwise costly SHAP importances for tree-based models. Given the high dimensionality of our dataset, encompassing almost 19k genes, we designed a computational pipeline to expedite the computation of Permutation Importance. Initially, we employed XGBoost built-in feature importance methods based on impurity decrease, a standard computation in tree-based methods, to significantly narrow down the list of potential genes of interest. This approach allowed us to reduce the number of genes to be evaluated for importance by at least a factor of ten. To further accelerate the process, we batched multiple variable computations by designing a Numba-enhanced routine and then performed joint fast inference leveraging the GPU-accelerated implementation of XGBoost. Feature importance computations were conducted individually for each of the 20 random seeds. While the SHAP mean importance values were computed exactly, the permutation feature importances were derived by randomly shuffling a gene's values and observing the resultant drop in model performance. A gene's permutation importance, for a given drug and a fixed seed, is obtained by averaging importance values obtained across three independent repetitions of the shuffling procedure. This methodology allowed us to mitigate aleatoric uncertainties inherent in the shuffling process, ensuring a more reliable identification of genes that are genuinely influential in predicting drug responses in cancer cell lines.

Putative and pathway gene recovery. In evaluating the soundness of our models, we investigated their ability to identify known putative targets and associated genes (i.e., those within the same pathway as the putative targets). For each random seed, we selected genes with non-zero importance value (as outlined in the "Importance Computation" section) and then calculated the frequency with which each putative gene appears in this subset of significant genes. Additionally, we conducted a parallel analysis involving genes functionally linked to putative targets through pathways. For each random seed and corresponding putative target, we aggregated all genes from the pathways involving the targets. We then determined the overlap between this aggregated set and the significant genes, documenting both the quantity and specific identities of the genes identified. Similar to the approach for putative target identification, the rigor of the selection criteria can be adjusted by focusing on pathway genes that consistently emerged across various random seeds. We refined our evaluation by calculating the empirical distribution of recovery rates for each drug, starting with the importation and binarization of gene importance scores—SHAP values (converted to 1 if greater than 0) and permutation importances (set to 1 if less than 0). By intersecting genes with both metrics equal to one, we identified those of dual significance. A baseline distribution of recovery frequencies across splits was established by averaging these intersected, binarized importances. This drug-specific procedure allowed us to derive critical quantiles (90th, 95th) as benchmarks for significant recovery. We then compared the putative targets' recovery rates to these percentiles.

Pathway enrichment. We performed pathway enrichment on the list of important genes (see above section) identified by each All-Genes model using the GSEAPY (v 1.0.6) package[55]. We performed an over-representation analysis using the *enrichr* function, considering as categories the "Reactome_2022" pathways from the MsigDB library[56], version 2023.1.Hs. We considered all the pathways with FDR < 0.1. To check if an enriched pathway represents a mechanism of action of the corresponding drug, we checked if it is connected to the MOA-pathways that we compiled for each drug by traversing the graph of Reactome pathway hierarchy using the *python-igraph* (v 0.9.10) library. If the enriched pathway is either a parent or a child of any MOA-pathway for that drug, it is considered to match the MOA for that drug. Enrichment analysis was performed for both GDSC and PRISM. Pathway enrichments obtained through this analysis are reported in Supplementary datas 4 and 10 for GDSC and PRISM respectively. Enrichment data obtained is also used to assess the proportion of MOA pathways that are overrepresented across the three criteria (refer to the "MOA Pathway Annotation" section). To generate Fig. 3D, we only considered the models exhibiting a correlation exceeding $\rho > 0.5$.

We compared significantly enriched pathways on GDSC and PRISM models using the Python matplotlib_venn library (https://github.com/konstantint/matplotlib-venn).

Clustering analysis of drug and MOA-related pathways. To elucidate the intricate relationships between drug MOA and their biological pathways, we clustered pathway enrichments and applied stringent statistical filters to identify and visualize significant drug-pathway associations. We focused on drug models with a correlation greater than 0.5 and the pool of MOA pathways found enriched, with a FDR below 0.1, in at least one drug model (see "Pathway enrichment"). For the clustering approach, we employed hierarchical clustering to organize both the drugs and the MOA pathways based on their patterns of association. This method allows for the grouping of drugs with similar pathway profiles and, conversely, pathways commonly influenced by a cohort of drugs. The clustering was conducted using the Ward method within the *clustermap* function from the seaborn python library (https://seaborn.pydata.org/, v0.12.2). The Ward method minimizes variance within clusters, thereby enhancing the interpretability of complex relationships by showcasing clusters of drugs and pathways with similar biological effects.

Correlating IC50s, predicted IC50s, Gene Expressions and SHAP values. We aimed to elucidate the relationship between IC50 values, predicted IC50s, gene expressions, and SHAP values to better understand the dynamics of drug response in cell lines. For a given drug, we sorted the IC50 values from the lowest to the highest. This sorting strategy enabled simultaneous comparison of actual versus predicted IC50 values, gene expression levels, and SHAP values corresponding to the gene expression across cell lines. The result of this analysis is depicted in Fig. 2C, specifically focusing on the interaction between the Venetoclax drug and the *BCL2* gene. This figure serves to visually convey the complex interplay and correlations among the variables of interest, providing insights into the predictive model's predictions and the gene's influence on drug sensitivity. To improve the graphical representation and clarity of trends within the gene expression and SHAP value data, both datasets were subjected to smoothing using a 5-lag moving window.

Gene essentiality analysis. We assessed the potential of local SHAP values, representing gene impacts on drug-cell line pair predictions, to identify genes involved in gene essentiality, as evidenced by CRISPR knock-out screenings. Utilizing a dataset from Pacini et al.[23], which lists essential genes across tissues and CRISPR Gene dependencies (see "Datasets"), we tailored this analysis to include only those cell lines and tissues mentioned therein. This required harmonizing tissue nomenclature differences between our dataset and Pacini et al., notably merging Esophagus and Stomach categories into a single group and mapping other tissues' names to match our dataset's structure (the mapping will be released together with the code to reproduce the analysis). Our analysis was further refined to drug-cell line pairs exhibiting the highest sensitivity, based on IC50 values and quantile score (see "Quantile Score"). This results in retaining 266 unique drugs and selecting 43,174 "top pairs." For these pairs, we examined the top k most negative SHAP values to identify influential genes, with k set at 10, 20, 50, and 100 thresholds. We remark on how negative SHAP values indicate genes whose inhibition likely enhances drug efficacy, aligning with the biological outcomes expected from CRISPR knockouts that lead to reduced cell viability. By aggregating these key genes and comparing them against the catalog of essential genes from the

reference genes, we aimed to ascertain the extent to which local importance estimates could recover genes critical for cell survival and implicated in gene essentiality. We constructed protein-protein interaction networks using STRING[57] within the Cytoscape platform[58]. Nodes within the network were annotated with degree information and a cumulative SHAP value pooled from the set SHAP values over 'top pairs.'

Responsiveness similarity analysis. After GDSC model training, we characterize cell lines through a 286-dimensional vector representation derived from the aggregated predictions of these models. This defines a "response space" which complements transcriptomic similarity analyses and can be used to compare samples based on their drug responsiveness similarity. This approach enables two key applications: clustering of cell lines to identify groups with similar drug sensitivities and the use of the Python library *faiss* for efficient nearest neighbor searches in the response space.

Model inference on TCGA

Quantile score. IC50 values are a commonly used measure of drug potency. However, raw values may not fully capture the nuances of drug efficacy and specificity, particularly for cytotoxic drugs, which could feature systematically lower IC50 levels. We propose a scoring system aimed at providing a more comprehensive assessment of a drug's potential.

Our approach integrates two dimensions: drug efficacy and drug specificity. Drug efficacy is defined as the likelihood of identifying an alternative drug within the dataset that possesses a higher IC50 value than the current drug in question. This measure reflects the drug's ability to inhibit or kill cancer cells at lower concentrations, thereby indicating its potency. On the other hand, drug specificity evaluates the probability of finding another cell line upon which the given drug exhibits a higher IC50 value. This dimension captures the drug's targeted action, ensuring that it effectively inhibits the growth of specific cancer cells with minimal effects on others, thereby reducing potential side effects.

To synthesize these dimensions into a singular, coherent score, we employed the harmonic mean of the two probabilities. The harmonic mean, a type of average, is particularly suited for our purpose because it tends to skew towards the lower of its inputs, ensuring that both high efficacy and high specificity are necessary for a drug to achieve a high score. As such, our metric ranges between 0 and 1, where a score closer to 1 indicates a drug that is both effective and specific. This property of the harmonic mean serves to penalize drugs that are highly effective against a broad range of cell lines (potentially indicating high toxicity) or highly specific but with limited overall efficacy.

$$Specificity = P\left(Y_d > y_d^c\right) \tag{1}$$

$$Efficacy = P(Y_c > y_d^c) \tag{2}$$

$$QS = \frac{2 * (Specificity * Efficacy)}{Specificity + Efficacy} \tag{3}$$

TCGA inference and cancer-specific drug prescription validation. We aimed to validate our models using the TCGA dataset. This allowed us to assess whether the proposed pipeline can accurately identify tumor-specific clinical drug prescriptions annotated in the GDSC through the use of processed NCI indications (refer to the "Dataset" section). A key step in our approach was aligning TCGA data with CCLE using the Celligner method (detailed in "Alignment of patients' bulk RNAseq with Celligner"), which enabled model training on CCLE data and subsequent deployment to TCGA. Moreover, an important part of this validation was to stratify TCGA patients by cancer type, utilizing the extensive metadata available within the TCGA dataset. Our analysis encompassed 9,805 TCGA samples for 41 clinically relevant drugs, with models' predictions ranked by logIC50 values and quantile scores.

We have performed additional analyses to validate the statistical significance of our model's predictions on TCGA dataset and defined an optimal number of k sample to consider for downstream analysis. First, we assessed for different top k samples the proportions of TCGA samples predicted for a drug prescribed for a specific cancer type, relative to the baseline frequency of the corresponding cancer type on TCGA. We assessed the extent of the model to predict samples from prescribed cancer types via classical machine learning metrics, considering as negatives all the cancer types for which a given drug is not prescribed. We proceeded as follows:

1. We fixed a number k of TCGA samples to prioritize starting from our model's predictions. Specifically, we selected the union of the k predictions with the lowest predicted IC50 values and the k predictions with the highest quantile scores.
2. We computed the relative proportions of tumor types in the entire TCGA dataset to obtain a baseline proportions vector.
3. For each drug, we calculated the relative proportions of tumor types within the extracted samples.
4. For each drug, we compared the predicted proportions with the baseline distribution. Given that we have ground truth indications for several drugs (considering as negatives all the cancer types for which a given drug is not prescribed):
   - True Positives (TP): Tumor types where the drug is actually prescribed and are relatively more abundant in the predicted proportions.
   - False Positives (FP): Tumor types where the drug is not prescribed but are predicted as more abundant.
   - True Negatives (TN): Tumor types where the drug is not prescribed and are not predicted as more abundant.
   - False Negatives (FN): Tumor types where the drug is prescribed but are not predicted as more abundant.

   This classification allows us to construct a 2×2 confusion matrix for each drug.
5. Using the confusion matrix, we compute standard performance metrics including accuracy, precision, recall, specificity, and F1-score.

We repeated the analysis over a range of k values, from 100 to 2500 in increments of 100. As expected, we found that higher k Values led to higher recall rates in both analyses due to the inclusion of more samples, but resulted in lower precision. On the other hand, lower k values offered higher precision by focusing on the most confident predictions, but reduced recall. We determined that $k = 600$ is the best tradeoff, since it maximized the F1-score, effectively balancing precision and recall.

A similar analysis is carried out also on the PRISM dataset. We first ranked the models by performance, then curated a list of the top 20 non-oncological drug models. For each model, predictions were ordered by Log Fold Change (LFC) from lowest to highest, and the initial 600 samples were selected. These samples were stratified by tumor type within the TCGA patient dataset to identify potential drug repurposing opportunities.

In conclusion, utilizing the same pipeline, we identified and quantified patients within the TCGA dataset where specific drug pairs are concurrently recommended, indicating potential synergies. We corroborated our findings by cross-referencing manually with the approved drug combinations in DrugBank[59], successfully identifying both known and approved combinations.

Experimental validation on PDAC samples

Ranking PDAC cell lines for selected tumor types with Cellligner. To determine which PDAC cell lines most closely resemble the tumor types (Pancreatic adenocarcinoma, Esophagogastric adenocarcinoma, Invasive breast carcinoma and head and neck squamous carcinoma) derived from Fig. 7A, Cellligner (https://depmap.org/portal/cellligner/) was used. For each of these tumors, we ranked and selected PDAC cell lines based on their Euclidean distance from that tumor type. The Euclidean distance was calculated as the average distance between a cell line and all TCGA tumor samples of that tumor type.

Cell lines, culture and treatments. The following human PDAC cell lines were used: CFPAC1 (KRAS G12V; ATCC CRL-1918) and PANC1 (KRAS G12D; ATCC CRL-1469), were used for the following experiments. CFPAC1 cells were maintained in Iscove's Modified Dulbecco's Medium (IMDM; Euroclone) + 10% fetal bovine serum (FBS) while PANC1 were maintained in Dulbecco's Modified Eagle Medium (DMEM; Euroclone) + 10% fetal bovine serum (FBS). Media were all supplemented with 2 mM L-glutamine. The cell line was authenticated by the IEO Tissue Culture Facility using the GenePrint10 System (Promega) and was routinely screened for Mycoplasma contamination.Irinotecan (Sigma-Aldrich, I1406) and Etoposide (Sigma-Aldrich, E1383) were diluted in DMSO and used at 3 different concentrate ions for the cell viability experiments. No commonly misidentified cell lines were used in the study.

Cell viability assays. 5–10,000 cells per well were plated in 96-well plates. One day after, cells were treated with the drug (Irinotecan or Etoposide). The cell viability was measured after 72 h of treatment using CellTiter-Glo Luminescent Cell Viability Assay (Promega, G9242) and GloMax® (Promega). The assay was performed three times in triplicates.

Experimental validation on GBM samples

Human Subjects. The research was conducted in compliance with the principles outlined in the Declaration of Helsinki, and the protocol for sample collection received approval from the Ethics Committee of the University Hospital of Pisa (787/2015).

Tumor specimens were obtained from 64 patients who underwent surgical resection of histologically confirmed GBM after providing informed consent. Samples were acquired from the Unit of Neurosurgery of Livorno Civil Hospital. All patients had a GBM diagnosis without a prior history of brain neoplasia and did not exhibit R132 IDH1 or R172 IDH2 mutations. Resected tumors were preserved in MACS tissue storage solution (Miltenyi Biotec, Bergisch Gladbach, Germany, cod. 130-100-008) at 4 °C for 2–4 h. Each tumor specimen was rinsed with Dulbecco's phosphate-buffered saline (DPBS) (Gibco, Carlsbad, CA, USA, cod. 14190094) within a sterile dish and divided into ~0.5–1 mm^2 pieces under a biological hood. Biopsies were cryopreserved in 90% fetal bovine serum (FBS, Gibco, Carlsbad, CA, USA, cod. 26140079) and 1% dimethyl sulfoxide (DMSO, Invitrogen, Waltham, USA, cod. D12345) at −140 °C until further preparations. We replaced the exact age of the patients with a categorical variable indicating whether the patient's age is above or below the threshold of 55 years. This threshold is widely used in clinical diagnostics for glioblastoma and maintains the clinical relevance of the data while eliminating the possibility of identifying individuals. The GBM samples were collected after the participants signed the informed consent form. The sex of the participants was determined based on self-report All clinical and molecular data of GBM samples are provided in Supplementary data 17.

RNA Isolation. RNA extraction from GBM tissues was performed using the Maxwell 16 LEV Simply RNA Tissue Kit (Promega, Madison, WI, USA, cod. AS1270) according to the manufacturer's protocol. The concentration of RNA was assessed using the Qubit Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA, cod. Q10210), and quality was evaluated using the Agilent 2200 Tapestation (Agilent Technologies, Santa Clara, CA, G2964AA) system.

Whole Transcriptome RNA Analysis Libraries. RNA-Seq was performed on the NextSeq 500 platform (Illumina, San Diego, CA, USA). Libraries were prepared using the Illumina Stranded Total RNA Prep with Ribo-Zero Plus kit (Illumina, cod. 20040525), starting from 200 ng of total RNA, following the manufacturer's protocol. The libraries were quantified using Qubit reagents (Thermo Fisher Scientific, Waltham, MA, USA, cod. Q10210) and analyzed for validation using TapeStation (Agilent Technologies, Santa Clara, CA, G2964AA). Up to 10 libraries were loaded onto the NextSeq High Output cartridge (150 cycles) (Illumina, San Diego, CA, USA, cod. 20024908).

Human Subjects for GBM Primary Cell Cultures. Tumor specimens were obtained from 2 patients who underwent surgical resection of histologically confirmed GBM after providing informed consent. Samples were acquired from the Unit of Neurosurgery of Livorno Civil Hospital. All patients had a GBM diagnosis without a prior history of brain neoplasia and did not exhibit R132 IDH1 or R172 IDH2 mutations. The patients included one female and one male, aged 37 and 76 years, respectively. Resected tumors were preserved in MACS tissue storage solution (Miltenyi Biotec, Bergisch Gladbach, Germany, cod. 130-100-008) at 4 °C for 2–4 hours. Each tumor specimen was rinsed with Dulbecco's phosphate-buffered saline (DPBS) (Gibco, Carlsbad, CA, USA, cod. 14190094) within a sterile dish and divided into ~ 0.5–1 mm^2 pieces under a biological hood. Biopsies were cryopreserved in 90% fetal bovine serum (FBS; Gibco, Carlsbad, CA, USA, cod. 26140079) and 1% dimethyl sulfoxide (DMSO; Invitrogen, Waltham, USA, cod. D12345) at −140 °C until further preparations.

GBM primary cell cultures. Patient-derived GB living tissues were dissociated into single-cell suspensions using mechanical dissociation combined with enzymatic degradation, utilizing the Brain Tumor Dissociation Kit (Myltenyi Biotech, Bergisch Gladbach, Germany, cod. 130-095-942), following the manufacturer's instructions. The culture medium consisted of DMEM/F12 medium (Gibco, Carlsbad, CA, USA, cod. 12634010), supplemented with 10% FBS (Gibco, Carlsbad, CA, USA, cod. 26140079), 100 U/mL penicillin, and 0.1 mg/mL streptomycin (Gibco, Carlsbad, CA, USA, cod. 15070063), 1% Amphotericin B (Gibco, Carlsbad, CA, USA, cod. 15290026), 1% G-5 supplement (Gibco, Carlsbad, CA, USA, cod. 17503012), and 1% Glutamax (Gibco, Carlsbad, CA, USA, cod. 35050061). Cultures were maintained in a 5% CO_2 atmosphere at 37 °C. The medium was replaced the day after culture. Upon reaching confluency, cells were seeded for viability assays, WST-1, and Crystal Violet, as described below. No commonly misidentified cell lines were used in the study.

Treatments of GBM primary cell cultures. AZD5591 (Aurogene S.r.l, Rome, Italy, cod. S8643) and AZD5582 (Sigma, St. Louis, MI, USA, cod. SML2900) were dissolved in DMSO (Invitrogen, Waltham, USA, cod. D12345) and water, respectively, to create stock concentrations of 200 mM and 4.8 mM, respectively. Dilutions of these drugs for treatment were prepared with the cell medium. We used concentrations of 2, 5, 10, 20, 50, and 100 uM of AZD5991 or AZD5582, or an equivalent volume of vehicle (DMSO or water) for control groups.

Viability and cytotoxicity assays. We used the Crystal Violet (CV) assay to evaluate cytotoxicity of the same chemicals on the 2 different cell types. For this assay, 2500 cells were seeded in 48-well plates (3 wells per experimental condition). After 72 h post-treatment, cells were fixed with 4% PFA (Thermo Fisher Scientific, Waltham, MA, USA, cod. Q10210cod. J19943.K2) and stained using a crystal violet solution

(0.1% crystal violet, 20% methanol, in water). Following staining, the excess crystal violet was washed with tap water, and plates were dried. Cells were de-stained using a 10% acetic acid solution, and the absorbance of the solution was then measured at 590 nm. IC50 values were computed using python's SciPy, library reimplementing the procedure of Corsello (2020).

Software

We employed customized scripts in python (version 3.8.11), using matplotlib (v3.6.0), seaborn (v0.11.1), and biopython (v1.78) libraries.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Transcriptomic and cell line drug response data were sourced from publicly available repositories as detailed in the "Datasets" subsection of the Methods. Key intermediate results and small datasets required for reproducing findings are available in the GitHub repository https://github.com/raimondilab/CellHit. Larger outputs, including predictions, trained MOA models, LLM drug annotations, and Cellligner-aligned TCGA-CCLE data, are hosted on Zenodo https://doi.org/10.5281/zenodo.14356698. Additional insights referenced in the main text are provided in the Supplementary Information. Source data for all figures and Supplementary Figs. are included in the Source Data file. 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 generated datasets can also be accessed at: https://cellhit.bioinfolab.sns.it Source data are provided with this paper.

Code availability

The code of the model and pipelines employed in this study are available at: https://github.com/raimondilab/CellHit. A snapshot of the code used to ontain the results showed in this manuscript is made available on Zenodo https://doi.org/10.5281/zenodo.14356698[60]. The code employed is under MIT license, BSD-3-Clause license or Unlicensed.

References

1. Chang, K. et al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
2. Barretina, J. et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
3. Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
4. Rees, M. G. et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat. Chem. Biol.* **12**, 109–116 (2016).
5. Corsello, S. M. et al. Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. *Nat. Cancer* **1**, 235–248 (2020).
6. Tsherniak, A. et al. Defining a cancer dependency map. *Cell* **170**, 564–576.e16 (2017).
7. Chen, J. & Zhang, L. A survey and systematic assessment of computational methods for drug response prediction. *Brief. Bioinform.* **22**, 232–246 (2021).
8. Xia, F. et al. A cross-study analysis of drug response prediction in cancer cell lines. *Brief. Bioinform.* **23**, bbab356 (2022).
9. Firoozbakht, F., Yousefi, B. & Schwikowski, B. An overview of machine learning methods for monotherapy drug response prediction. *Brief. Bioinform.* **23,** bbab408 (2022).
10. Sharifi-Noghabi, H. et al. Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models. *Brief. Bioinform.* **22**, bbab294 (2021).
11. Kuenzi, B. M. et al. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* **38**, 672–684.e6 (2020).
12. Chawla, S. et al. Gene expression based inference of cancer drug sensitivity. *Nat. Commun.* **13**, 5680 (2022).
13. Ferraro, L., Scala, G., Cerulo, L., Carosati, E. & Ceccarelli, M. MOViDA: multiomics visible drug activity prediction with a biologically informed neural network model. *Bioinformatics* **39**, btad432 (2023).
14. Wang, X., Sun, Z., Zimmermann, M. T., Bugrim, A. & Kocher, J. P. Predict drug sensitivity of cancer cells with pathway activity inference. *BMC Med Genom.* **12**, 15 (2019).
15. Tang, Y. C. & Gottlieb, A. Explainable drug sensitivity prediction through cancer pathway enrichment. *Sci. Rep.* **11**, 1–10 (2021).
16. Samal, B. R., Loers, J. U., Vermeirssen, V. & De Preter, K. Opportunities and challenges in interpretable deep learning for drug sensitivity prediction of cancer cells. *Front. Bioinform.* **2**, 1036963 (2022).
17. Ma, J. et al. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat. Cancer* **2**, 233–244 (2021).
18. Warren, A. et al. Global computational alignment of tumor and cell line transcriptional profiles. *Nat. Commun.* **12**, 1–12 (2021).
19. Kapoor, I., Bodo, J., Hill, B. T., Hsi, E. D. & Almasan, A. Targeting BCL-2 in B-cell malignancies and overcoming therapeutic resistance. *Cell Death Dis.* **11**, 941 (2020).
20. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process Syst.* **2017-December**, 4766–4775 (2017).
21. Jassal, B. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).
22. Harding, S. D. et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. *Nucleic Acids Res.* **46**, D1091–D1106 (2018).
23. Pacini, C. et al. A comprehensive clinically informed map of dependencies in cancer cells and framework for target prioritization. *Cancer Cell* https://doi.org/10.1016/j.ccell.2023.12.016 (2024).
24. Subbiah, V. et al. Dabrafenib plus trametinib in BRAFV600E-mutated rare cancers: the phase 2 ROAR trial. *Nat. Med.* **29**, 1103–1112 (2023).
25. Corsello, S. M. et al. Abstract 3400: adenosine receptor antagonists exhibit potent and selective off-target killing of FOXA1-high cancers. *Cancer Res.* **80**, 3400–3400 (2020).
26. Arora, C. et al. The landscape of cancer-rewired GPCR signaling axes. Cell Genomics **4**, https://doi.org/10.1016/j.xgen.2024.100557 (2024).
27. Di Chiaro, P. et al. Mapping functional to morphological variation reveals the basis of regional extracellular matrix subversion and nerve invasion in pancreatic cancer. *Cancer Cell* **42**, 662–681.e10 (2024).
28. Collisson, E. A. et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat. Med.* **17**, 500–503 (2011).
29. Fujita, K. I., Kubota, Y., Ishida, H. & Sasaki, Y. Irinotecan, a key chemotherapeutic drug for metastatic colorectal cancer. *World J. Gastroenterol.* **21**, 12234–12248 (2015).

30. Zhang, W., Gou, P., Dupret, J. M., Chomienne, C. & Rodrigues-Lima, F. Etoposide, an anticancer drug involved in therapy-related secondary leukemia: enzymes at play. *Transl Oncol.* **14**, 101169 (2021).
31. Fernández-Torras, A., Duran-Frigola, M. & Aloy, P. Encircling the regions of the pharmacogenomic landscape that determine drug response. *Genome Med.* **11**, 1–15 (2019).
32. Lobentanzer, S. et al. A platform for the biomedical application of large language models. *Nat Biotechnol* https://doi.org/10.1038/s41587-024-02534-3 (2025).
33. Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **630**, 625–630 (2024).
34. Wu, V. H. et al. The GPCR–Gαs–PKA signaling axis promotes T cell dysfunction and cancer immunotherapy failure. *Nat. Immunol.* **24**, 1318–1330 (2023).
35. Dimitrieva, S. et al. Biologically relevant integration of transcriptomics profiles from cancer cell lines, patient-derived xenografts, and clinical tumors using deep learning. *Sci. Adv.* **11**, 5596 (2025).
36. Kundra, R. et al. OncoTree: a cancer classification system for precision oncology. *JCO Clin. Cancer Inform.* **5**, 221–230 (2021).
37. Chakravarty, D. et al. OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol.* 1–16 https://doi.org/10.1200/PO.17.00011/SUPPL_FILE/DS_17.00011.DOCX (2017).
38. Kim, S. et al. PubChem substance and compound databases. *Nucleic Acids Res.* **44**, D1202–D1213 (2016).
39. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model* **50**, 742–754 (2010).
40. Ahmad, W., Simon, E., Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa-2: towards chemical foundation models. *arXiv preprint arXiv:2209.01712* (2022).
41. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Machine Learn. Res.* **12**, 2825–2830 (2011).
42. Jiang, A. Q. et al. Mixtral of Experts. *arXiv preprint arXiv:2401.04088* (2024).
43. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **35**, 24824–24837 (2022).
44. Wang, X. et al. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).
45. Lewis, P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Adv. Neural Inf. Process. Syst.* **33**, 9459–9474 (2020).
46. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
47. Kwon, W. et al. Efficient memory management for large language model serving with pagedattention. *Proceedings 29th ACM Symposium Operating Systems Principles* **1**, 611–626 (2023).
48. Wolf, T. et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. (2019).
49. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: a next-generation hyperparameter optimization framework. *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2623–2631 (2019) https://doi.org/10.1145/3292500.3330701.
50. Ozaki, Y., Tanigaki, Y., Watanabe, S. & Onishi, M. Multiobjective tree-structured parzen estimator for computationally expensive optimization problems. *GECCO 2020 - Proc. 2020 Genetic and Evolutionary Computation Conference* 533–541 https://doi.org/10.1145/3377930.3389817 (2020).
51. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, (2019).
52. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. *Proc. ACM SIGKDD International Conference Knowledge Discovery Data Mining*. 785–794 (2016)..
53. Molnar, C. et al. General Pitfalls of model-agnostic interpretation methods for machine learning models. in *xxAI - Beyond Explainable AI*, 39–68 (2020)..
54. Breiman, L. Random forests. *Mach Learn* **45**, 5–32 (2001).
55. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39**, btac757 (2023).
56. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).
57. Szklarczyk, D. et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
58. Lotia, S., Montojo, J., Dong, Y., Bader, G. D. & Pico, A. R. Cytoscape app store. *Bioinformatics* **29**, 1350–1351 (2013).
59. Knox, C. et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **39**, D1035–D1041 (2011).
60. Carli, F. & RAIMONDI, F. Learning and actioning general principles of cancer cell drug sensitivity. Zenodo, https://doi.org/10.5281/ZENODO.14356698 (2024).

Acknowledgements

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.). 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. 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.

Author contributions

Training and inference of the machine learning models: F.C.; Data analysis: F.C., P.D.C., C.A., L.B., P.A., P.M., and F.R.; Experimental validation: P.D.C., M.M., A.C., S.F., F.L., and C.M.M.; Webapp development: NDOR; Samples acquisition: A.L.D.S., O.S.S., F.P., G.R.D., C.M.M., and G.N.; Results interpretation: F.C., P.D.C., M.D.F., A.L.D.S., C.M.M., G.N., and F.R.; Conceptualization: F.C., F.G., P.L., M.D.F., and F.R.; Manuscript writing: F.C., P.D.C., M.D.F., C.M.M., G.N., and F.R.; Funding acquisition and management: G.N. and F.R.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-56827-5.

Correspondence and requests for materials should be addressed to Francesco Carli or Francesco Raimondi.

Peer review information *Nature Communications* thanks the anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.