



OPEN Deep learning-driven drug response prediction and mechanistic insights in cancer genomics

Guili Yu^{1,3} & Qiangqiang Fan^{2,3}✉

In the field of cancer therapy, the diversity and heterogeneity of cancer genomes in clinical patients complicate and challenge the effective use of non-targeted drugs, as these drugs often fail to address specific genetic events. Recent advancements in large-scale in vitro drug screening assays have generated extensive drug testing and genomic data, providing valuable resources to explore the relationship between genomic features and drug responses. In this study, we developed a deep neural network model, DrugS (Drug Response prediction Utilizing Genomic features Screening), utilizing gene expression and drug testing data from human-derived cancer cell lines to predict cellular responses to drugs. Leveraging gene expression and mutation data, we elucidated potential molecular mechanisms underlying SN-38 resistance. Additionally, we used DrugS to evaluate the effects of drugs on cancer cell proliferation in patient-derived xenograft models. In in vitro combination drug experiments, DrugS revealed that CDK inhibitors, mTOR inhibitors, and apoptosis inhibitors effectively reverse Ibrutinib resistance, providing new therapeutic strategies to overcome drug resistance. Furthermore, we assessed the applicability of the DrugS model in drug screening and patient prognosis evaluation using drug information and gene expression data from The Cancer Genome Atlas. In summary, our study offers a novel approach for drug response prediction and mechanism research in cancer therapy from a genomic perspective and demonstrates the potential applications of the DrugS model in personalized therapy and resistance mechanism elucidation.

Keywords Oncogenomic profiling, Pharmacogenomics, Neural network modeling, Therapeutic resistance, Precision oncology

The emergence of next-generation sequencing (NGS) has ushered in a new era of genomics, with an expanding array of cancer genomic data from cell lines and patients now accessible to the public^{1,2}. This wealth of information has facilitated a deeper understanding of the intricate relationship between cellular metabolism and its external environment at a molecular level, as well as the crucial link between phenotypes and their underlying genetic bases³. By delving into genomic data, including gene expression profiles, mutational spectra, gene fusion events, and copy number variations, we can identify key genetic anomalies that correspond to specific phenotypic changes^{4–6}. This approach sheds light on the molecular mechanisms at play and enhances our grasp of the pathogenesis, which in turn aids in the formulation of targeted therapeutic interventions^{7,8}. For instance, comparative genomic analyses between cancer tissues and their adjacent normal tissues can reveal genetic mutations that are critical to the process of carcinogenesis^{9,10}. A prime example is the BCR-ABL gene fusion, which is strongly associated with chronic myeloid leukemia (CML). The targeted use of tyrosine kinase inhibitors has been shown to markedly improve the prognosis for CML patients presenting this gene fusion¹¹. Similarly, elevated HER2 expression in gastric cancer tissues represents a promising target for therapeutic intervention¹². Furthermore, the presence of the KRAS G12C mutation, prevalent in lung and gastric cancers, offers a targetable molecular aberration that, when addressed, can significantly enhance patient survival outcomes¹³.

In recent years, the pharmaceutical screening of cancer cells has been extensively pursued by the scientific community, yielding a trove of valuable data^{14–17}. Renowned institutions such as the Broad Institute have harnessed the power of NGS platforms and sophisticated bioinformatics to curate an expansive genomic dataset from over a thousand cancer cell lines. This dataset, which includes gene expression profiles, mutational landscapes, gene fusions, and copy number variations, has been made publicly accessible through the Dependency Map (DepMap) project database¹⁸. Pioneering efforts by Yang et al.¹⁹ led to the establishment of the first large-scale, public cell line drug response repository, the Genomics of Drug Sensitivity (GDSC). This database encompasses

¹Pujiang Community Health Service Center, Shanghai, China. ²Pujiang People Institute, Shanghai, China. ³Guili Yu and Qiangqiang Fan contributed equally to this work. ✉email: dongliulou@126.com

sensitivity data for 138 drugs across 700 cancer cell lines, spotlighting the genomic factors that dictate drug responsiveness. Building upon this groundwork, Iorio et al.²⁰ extended the scope of drug response research to include 11,289 cancer samples and 1001 human cancer cell lines, uncovering that cell lines are reliable proxies for the oncogenic variations. This underscores the potential of cell line-based research to significantly inform clinical drug development. Seashore-Ludlow et al.²¹ adopted annotated cluster multidimensional enrichment analysis to scrutinize drug responses across more than 800 human cancer cell lines, culminating in the Cancer Therapeutic Response Portal (CTRP v2), a publicly accessible resource. The National Cancer Institute's NCI60 anticancer drug screen, initiated in the late 1980s, has evolved to include drug response profiles for 162 human cancer cell lines and is now publicly accessible²². Beyond these foundational datasets, numerous research teams have ventured into the application of diverse computational algorithms, including machine learning, deep neural networks (DNNs), and convolutional neural networks (CNNs), to predict cellular drug responses using genomic features as predictive variables^{23–25}. Iorio et al.²⁰ employed elastic net regression and random forest models, incorporating copy number variations, methylation, and gene expression data, to construct predictive models for 265 compounds across various cancer types. Their findings highlighted the predictive prowess of gene expression data. The NCI-DREAM initiative²³ evaluated 44 algorithms through a collaborative challenge, exploring kernel methods, nonlinear regression, and ensemble modeling techniques. Utilizing data from 53 human cancer cell lines, the project demonstrated that Bayesian models excelled in predicting IC₅₀ values, with gene expression data being a significant contributor to predictive accuracy. Jia et al.²⁴ leveraged variational autoencoders (VAE) for dimensionality reduction of highly variable genes, subsequently employing elastic net regression for drug response modeling. Sakellaropoulos et al.²⁵ mirrored this approach, developing DNN, elastic net, and random forest models to forecast drug responses for individual drugs. The CaDRReS-SC model²⁶ harnesses latent space algorithms to decipher the interplay between drug action and cellular transcriptomic profiles, enabling the prediction of drug responses at the single-cell level based on transcriptomic similarities. Precily model, developed by Chawla et al.²⁷, performs dimensionality reduction on expression data using the Gene Set Variation Analysis (GSVA) method. This model integrates drug data from the GDSC dataset to construct a feature-rich DNN predictive model for drug response, further exemplifying the convergence of bioinformatics and pharmacology in advancing personalized medicine.

While existing models often rely on highly variable genes or predefined gene sets for dimensionality reduction of gene expression data and sometimes develop drug-specific models for individual compounds, they frequently encounter challenges in integrating gene expression datasets from diverse sources due to limited compatibility. To address these limitations, we developed the DrugS model, an advanced analytical framework that incorporates 20,000 protein-coding genes. We employed the following strategies to enhance the robustness of the model: (1) Log Transformation and Scaling: To mitigate the influence of outliers and ensure cross-dataset comparability, gene expression values were log-transformed and scaled to a uniform range. (2) Dimensionality Reduction via Autoencoder: The autoencoder reduces data dimensionality by extracting key gene expression features, enabling the model to capture the intrinsic structure of the data while minimizing variability due to data source-specific differences. (3) Dropout Layers in the Neural Network (NN) Model: Dropout layers were integrated into the neural network architecture to prevent overfitting and enhance generalizability. These layers improve the model's robustness and reduce sensitivity to variations in data originating from different sources. These preprocessing and modeling strategies collectively improve the compatibility and integrative capacity of the DrugS model, enabling it to effectively handle gene expression data from multiple sources. Moreover, the DrugS model transcends the limitations of drug-specific mathematical models by integrating compound fingerprint information as an additional input layer. This feature provides a more nuanced understanding of drug action and response mechanisms at the molecular level. It demonstrates robust performance across different normalization methods, making it suitable for a wide range of experimental data. This flexibility enhances its applicability across diverse datasets and research contexts²⁸.

In this study, we have crafted a sophisticated DNN model designed to predict compound responses predicated on gene expression profiles and SMILES (Simplified Molecular-Input Line-Entry System) strings of pharmaceuticals. Our model acknowledges that gene expression data stands as the most pivotal variable in forecasting drug reactions^{29–31}. Beyond gene expression, other genomic features—including gene methylation³², mutations, copy number variations³³, and fusions³⁴—also modulate drug responses, especially within the realm of targeted therapies. Given that the pronounced effects of targeted drugs might manifest as outliers in machine learning or neural network modeling, potentially skewing model accuracy, we implemented a strategic approach to counteract their influence. Initially, we conducted TSNE clustering on cellular data based on gene expression, subsequently excluding assay data that exhibited considerable variability within homogeneous clusters for identical drugs. This step ensured a more refined input for our model. We then employed an autoencoder model to distill the complexity of over 20,000 protein-coding genes into a concise set of 30 features. From the SMILES strings of the drugs, we extracted 2048 features, culminating in a robust input matrix of 2078 features. The natural logarithm of the inhibitory concentration 50 (LN IC₅₀) served as the outcome variable for our DNN model training. To substantiate the model's predictive efficacy, we subjected it to rigorous testing using the CTRPv2 and NCI-60 datasets. Moreover, we correlated the model's predictions with drug response data and viability scores derived from PDX models, enriching our analysis with a deeper integrative perspective. In the context of combination therapy for drug-resistant cell lines, our model facilitated the identification of compounds with the potential to override Ibrutinib resistance by leveraging cell line expression data. Finally, we gauged the clinical relevance of our model by harmonizing TCGA patient expression data with details on clinical drug administration. In essence, our study advances the frontiers of previous research, introducing a robust analytical strategy and a DNN model capable of forecasting drug responses in a multitude of scenarios. This model stands as a testament to the potential of integrating diverse genomic and pharmacological data to unravel the complexities of drug action and resistance.

Results

Gene expression-based clustering of cancer cell lines

For our analysis, we obtained gene expression data for 20,000 protein-coding genes across 1450 cell lines from the DepMap database, encompassing 29 cancer types with a focus on lung cancer and lymphoma (Fig. 1A). Using TSNE clustering at a resolution of 0.2, we identified nine distinct clusters (Fig. 1B). Differential analysis revealed the top 10 overexpressed genes for each cluster, showcasing unique expression patterns (Fig. 1C). Functional annotation highlighted pathways enriched in specific clusters, such as leukocyte activation in cluster 3 (implicating immune response, Fig. 1D), myeloid leukocyte activation in cluster 6 (relevant to myeloid leukemia, Fig. 1E), pigment biosynthesis in cluster 5 (suggesting pigmentation anomalies), and hormone response in cluster 8 (indicating roles in hormonal signaling, Supplementary Fig. 1C). The TSNE clustering showed strong correlations with cancer type classifications. For instance, lymphomas were predominantly associated with cluster 3, skin cancers and melanomas with cluster 5, myeloid tumors with cluster 6, and breast cancers with cluster 8 (Fig. 1F). Notably, Cluster 0 demonstrated a high degree of heterogeneity, encompassing genes with elevated expression levels linked to tumorigenesis and progression-related pathways, such as the PI3K-Akt signaling pathway, TNF signaling pathway, and NF- κ B signaling pathway (Supplementary Fig. 1D). The molecular characteristics of Cluster 0 likely account for its inclusion of multiple cancer types and its heterogeneous composition (Fig. 1F). These findings, further supported by Gene Set Enrichment Analysis (GSEA), confirm that the identified clusters not only align with known cancer classifications but also capture biologically significant pathways.

Drug response analysis across cancer cell clusters

We obtained drug assay data for 378 and 286 compounds from the GDSC1 and GDSC2 datasets, respectively, across 968 and 967 cancer cell lines. Among these, 637 cell lines had matching gene expression data (Fig. 2A), representing assays for 542 compounds targeting 23 biological pathways (Fig. 2B). Drug response analysis revealed that cluster 3, predominantly comprising lymphoma cell lines, was highly sensitive to Cytarabine, with significantly lower IC50 values compared to other clusters (Fig. 2C,D). Cytarabine, a clinically used agent for treating acute leukemia and lymphoma, inhibits DNA synthesis and cancer cell proliferation³⁵. Similar trends

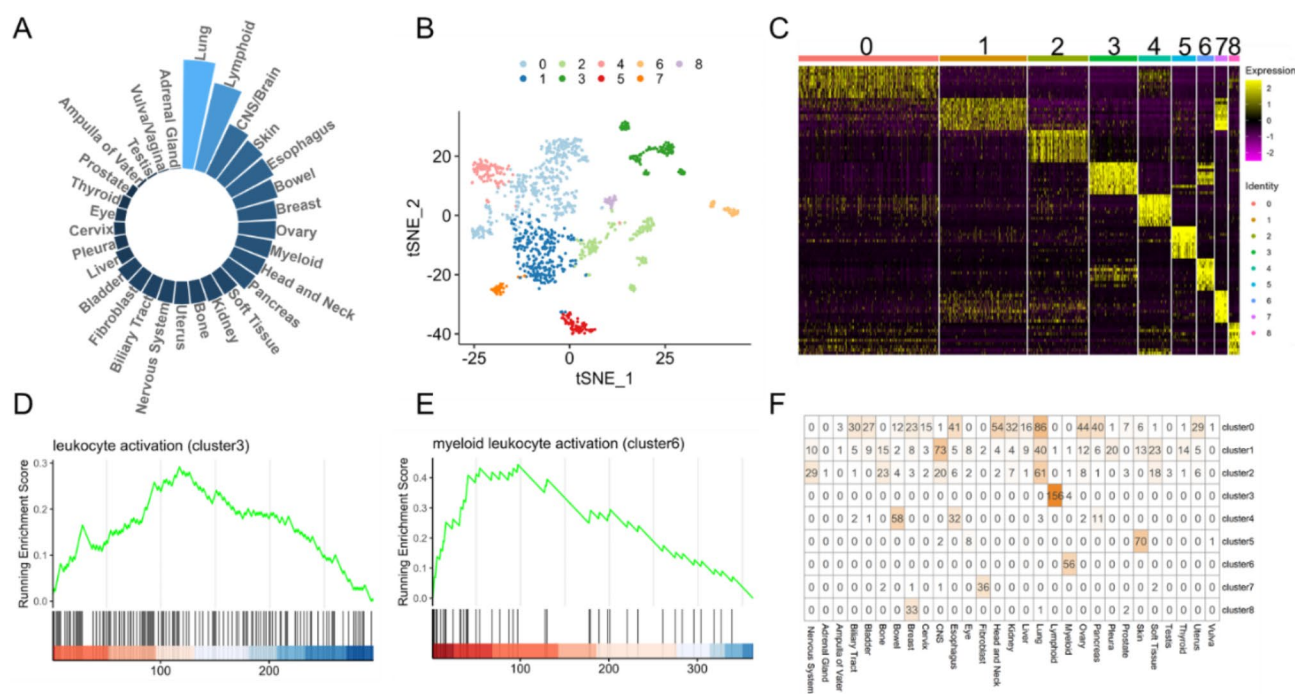


Fig. 1. Integrative genomic analysis and clustering of cancer cell lines. (A) Overview of the comprehensive gene expression dataset obtained from the DepMap database, comprising 20,000 protein-coding genes across 1450 cell lines representing 29 cancer types, with a specific focus on lung cancer and lymphoma. (B) Results of TSNE clustering analysis applied to the gene expression data, identifying nine distinct cell clusters. (C) Differential analysis conducted within each cluster, pinpointing the top 10 genes that are specifically overexpressed and exhibit distinctive expression patterns across the cell lines. (D,E) GSEA annotation of the specifically overexpressed genes within each cluster, indicating an enrichment in biological pathways associated with leukocyte activation for cluster 3 and myeloid leukocyte activation for cluster 6. (F) Alignment of the TSNE clustering results with the known cancer types of the cells, with cluster 3 mainly representing lymphomas, cluster 5 skin cancers and melanomas, cluster 6 myeloid tumors, and cluster 8 predominantly breast cancer. The consistency with GSEA functional analysis underscores the biological significance of the clustering outcomes.

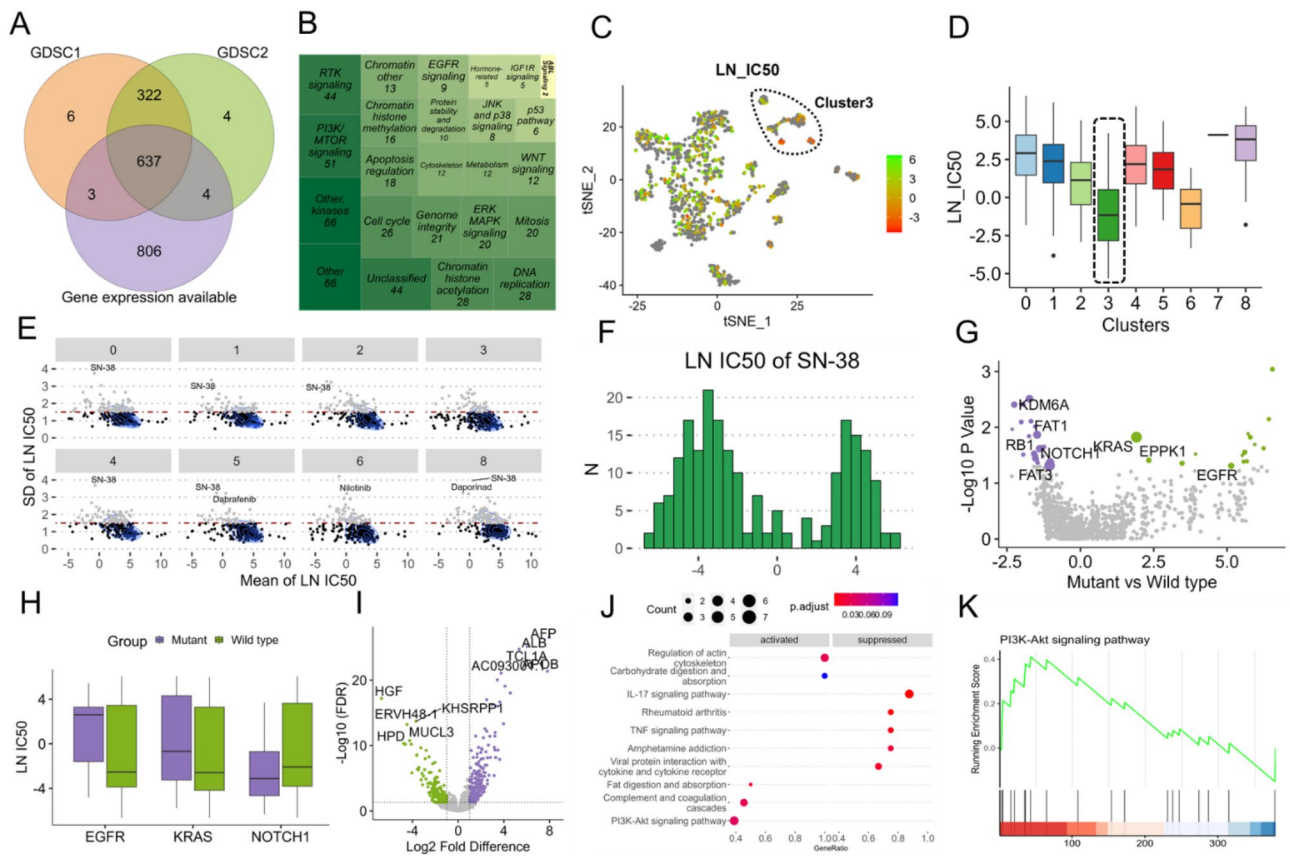


Fig. 2. Drug sensitivity profiling and mechanistic insights in cancer cell lines. **(A)** Compilation of gene expression data from 644 cell lines within the GDSC1 and GDSC2 datasets, highlighting the intersection of cell lines with available expression profiles. **(B)** Overview of drug assay data for 542 compounds, targeting 23 distinct biological pathways. **(C,D)** Analysis of drug response data revealing increased sensitivity of cluster 3, primarily Lymphoma cell lines, to Cytarabine, with significantly lower IC₅₀ values compared to other clusters. **(E)** Identification of pronounced anomalies in response to SN-38 across multiple clusters, particularly in cluster 0, demonstrating a dichotomy between high sensitivity and resistance. **(F)** Mutation analysis implicating KRAS and EGFR mutations in SN-38 resistance, and NOTCH1 mutations in increased sensitivity to SN-38. The left panel displays the distribution of LN IC₅₀ values in cluster 0, illustrating the variance in drug response. **(G)** Results of Wilcoxon tests conducted to compare the LN IC₅₀ values between mutant and wild-type genotypes, providing statistical evidence of the impact of these mutations on drug sensitivity. **(H)** The differences in LN IC₅₀ values between gene mutant and wild-type groups, further emphasizing the biological significance of these genetic alterations on SN-38 drug response. **(I)** Comparative IC₅₀-based analysis showing aberrant overexpression of genes in the PI3K-AKT signaling pathway associated with SN-38 resistance. The differential expression of genes between resistant (LN IC₅₀ > 0) and sensitive (LN IC₅₀ < 0) cell lines, identifying key genetic markers that may contribute to the phenotype of drug resistance. **(J)** Gene Ontology (GO) enrichment results for the two groups, providing insights into the biological processes and molecular functions that are significantly altered in the context of SN-38 resistance. **(K)** GSEA outcomes for the PI3K-AKT signaling pathway, demonstrating its significant overrepresentation in resistant cell lines, which suggests a potential role in mediating resistance mechanisms to SN-38.

were observed for drugs such as Dactinomycin, Gemcitabine, and AZD7762 (Supplementary Fig. 2A–C). Gene expression-based clustering demonstrated consistent drug responses within clusters, with deviations likely influenced by experimental variability or genetic factors like copy number variations, fusions, and mutations. To minimize these effects, cell lines with high variability in drug response ($SD_{ln_ic50} > 1.5$) were excluded from modeling (Fig. 2E and Supplementary Fig. 2D). Anomalies were particularly notable for SN-38, a DNA topoisomerase inhibitor³⁶, which showed both high sensitivity and resistance within cluster 0 (Fig. 2F). Further analysis suggested that resistance to SN-38 was associated with KRAS and EGFR mutations, while NOTCH1 mutations increased sensitivity (Fig. 2G,H). Resistant cell lines also exhibited overexpression of genes linked to the PI3K-AKT signaling pathway (Fig. 2I,K), potentially contributing to SN-38 resistance.

Development and validation of the drugs model

To develop and train the deep neural network model DrugS, we integrated LN IC₅₀ data from the GDSC1 and GDSC2 datasets with gene expression profiles from the DepMap database. SMILES strings for the compounds

were sourced from PubChem, and molecular fingerprint information was generated using RDKit, resulting in 2048-dimensional sparse matrices for each compound. Gene expression data for 20,000 protein-coding genes across 642 cell lines were reduced to 30 features using an autoencoder neural network. Each compound-cell assay was represented by 2078 independent variables, with LN IC₅₀ serving as the dependent variable. The DrugS model was trained to predict LN IC₅₀ based on these features (Fig. 3A). The GDSC2 dataset was utilized for model training and validation, with 70% of the compound-cell assays used as the training set and 30% as the validation set. The GDSC1 dataset was employed as the test set for evaluating the DrugS model's performance. The predictions of the DrugS model showed good agreement with experimental results ($R_{\text{drugs}} = 0.67$, Fig. 3B). Comparatively, we assessed two other models, Precily and CaDRReS_SC, using the GDSC1 dataset as the test set. DrugS and CaDRReS_SC exhibited similar predictive correlations, while Precily demonstrated slightly lower correlation performance (Supplementary Fig. 3A-B). When evaluating the models' predictive accuracy for drug response across different compounds, DrugS and Precily showed comparable efficacy, while CaDRReS_SC outperformed both (Fig. 3C). However, when mean squared error (MSE) was used as the evaluation metric, no significant differences were observed among the three models (Fig. 3D). To further explore the influence of cancer type on drug response prediction, we grouped data by cancer type and assessed the correlation and MSE of predicted versus observed values. DrugS exhibited the highest median Pearson correlation coefficient ($\text{PCC}_{\text{median}} = 0.7$), followed by CaDRReS_SC ($\text{PCC}_{\text{median}} = 0.68$) and Precily ($\text{PCC}_{\text{median}} = 0.62$). In terms of MSE, CaDRReS_SC showed the best performance ($\text{MSE}_{\text{median}} = 3.0$), with DrugS and Precily yielding comparable results (Fig. 3E). Notably, CaDRReS_SC is limited in that it cannot incorporate compound-specific features as input variables, making it unsuitable for predicting drug response for compounds not included in the training set. In summary, the DrugS model demonstrated superior predictive performance in both pooled and individual correlation analyses compared to Precily and CaDRReS_SC. While MSE differences among the models were negligible, all models showed some degree of variability in predictive efficacy depending on cancer type. The ability of DrugS to integrate compound-specific features further underscores its versatility and robustness for drug response prediction.

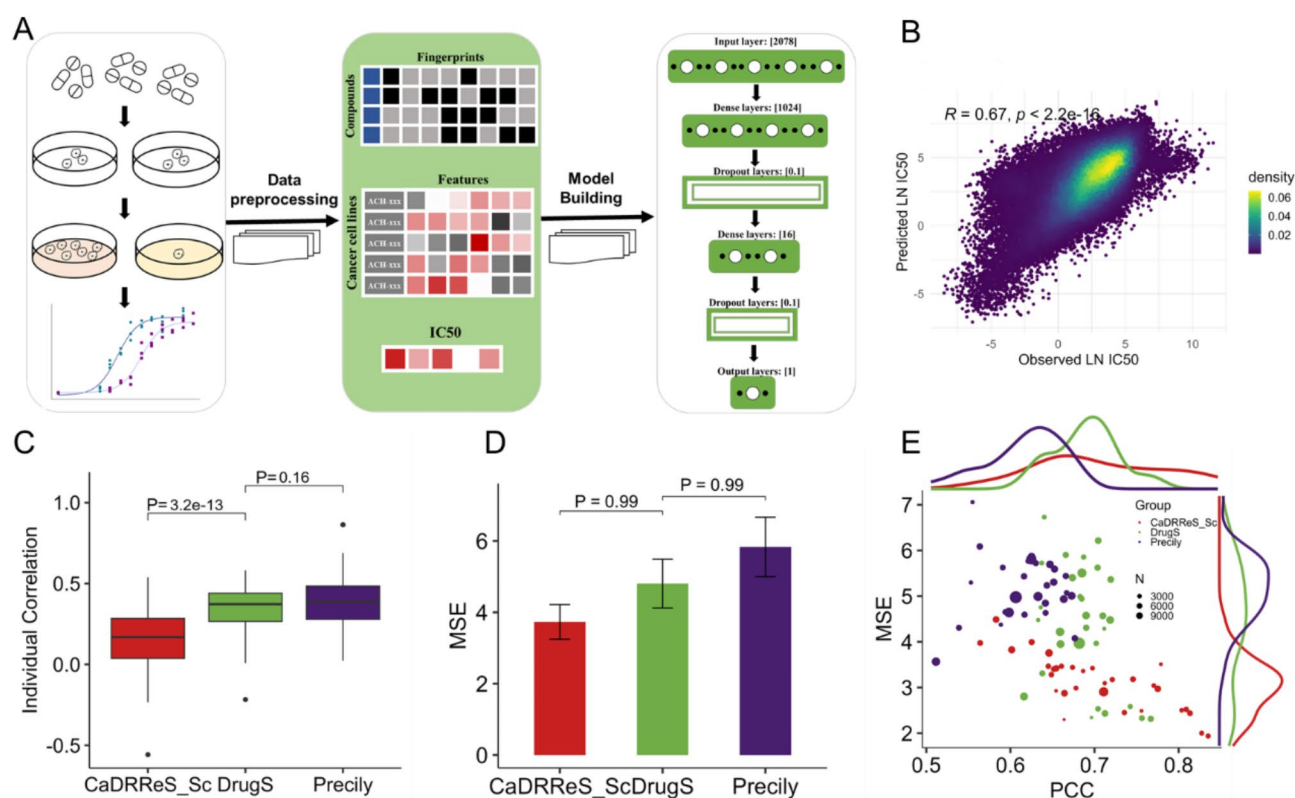


Fig. 3. DrugS model development and predictive performance. (A) Schematic workflow illustrates the development process of the DrugS model, detailing the integration of gene expression data for 20,000 protein-coding genes from 642 cell lines and compound fingerprint information derived from SMILES, culminating in a DNN with 2078 inputs and one output. (B) Predictive analysis of LN IC₅₀ values by the DrugS model using the GDSC1 dataset. (C,D) Benchmarking analysis of the DrugS model against published models, highlighting its performance in predicting LN IC₅₀ values with improved correlation. (E) MSE and PCC analyses between predicted and observed LN IC₅₀ values, with each data point representing a distinct cancer type. The size of each point reflects the number of compound-cell assays associated with that cancer type.

Evaluating drugs model for predicting drug response and unveiling mechanisms of action

The CTRPv2 dataset includes 804 cell lines and 461 compounds, 60 of which overlap with the GDSC2 database (Fig. 4A). Using the DrugS model, we predicted drug responses by integrating gene expression data with compound fingerprints. The predicted LN IC₅₀ values for the overlapping compounds correlated strongly with observed data ($R=0.75$, $p<0.05$), demonstrating the model's robustness (Fig. 4B and Supplementary Fig. 4A). For drugs absent from the training set, such as Bendamustine and Decitabine, the model still achieved high prediction accuracy (Fig. 4C), suggesting that these drugs may share mechanisms of action with compounds in the training set. Bendamustine, a DNA topoisomerase I inhibitor, binds to DNA and halts replication³⁷. By combining IC₅₀ data for Bendamustine from CTRPv2 and GDSC, we identified drugs highly correlated with Bendamustine, including Oxaliplatin³⁸, Irinotecan³⁹, Niraparib⁴⁰, and Topotecan⁴¹ (Fig. 4D). These drugs also disrupt DNA replication, highlighting a shared mechanism of action. To further explore the model's potential in revealing unknown drug mechanisms, we predicted Decitabine's IC₅₀ across CTRPv2 cell lines and analyzed its correlations with over 200 drugs in the GDSC database (Fig. 4E). High-correlation drugs were mainly involved in chromatin modification pathways, consistent with Decitabine's known mechanism of disrupting DNA synthesis and inhibiting cell proliferation by altering chromatin structures⁴² (Fig. 4F). These results suggest that the DrugS model can provide valuable insights into the mechanisms of action for poorly characterized drugs.

Applications in PDX samples and drug resistance studies

To substantiate the applicability of the DrugS model to PDX samples, we engaged gene expression microarray data from the GSE151343 dataset, comprising 20 Medulloblastoma patient-derived xenografts⁴³. We applied the DrugS model to predict drug responses in these PDX models and performed clustering analysis. The results indicated that the G3/G4 subtypes were more sensitive to drug treatments compared to the SHH subtype, aligning with prior research⁴³ (Fig. 5A). Furthermore, an in vitro drug screening assay was conducted to evaluate the impact of 7729 compounds on cell proliferation across the 20 PDX models. This screen identified 375 compounds that significantly curtailed cell proliferation, with 15 small molecules found in both the screen and the GDSC dataset (Fig. 5B). By integrating the viability score data, we discovered a substantial correlation between the LN IC₅₀ values predicted by the DrugS model for these 15 small molecules and the observed viability scores ($R=0.66$, $P<0.05$) (Fig. 5C). This finding suggests that the DrugS model, despite being trained on RNA-Seq data, can effectively utilize microarray-based gene expression data to anticipate drug responses in PDX models. The validation at the in vitro experimental level underscores the model's prospective utility in the

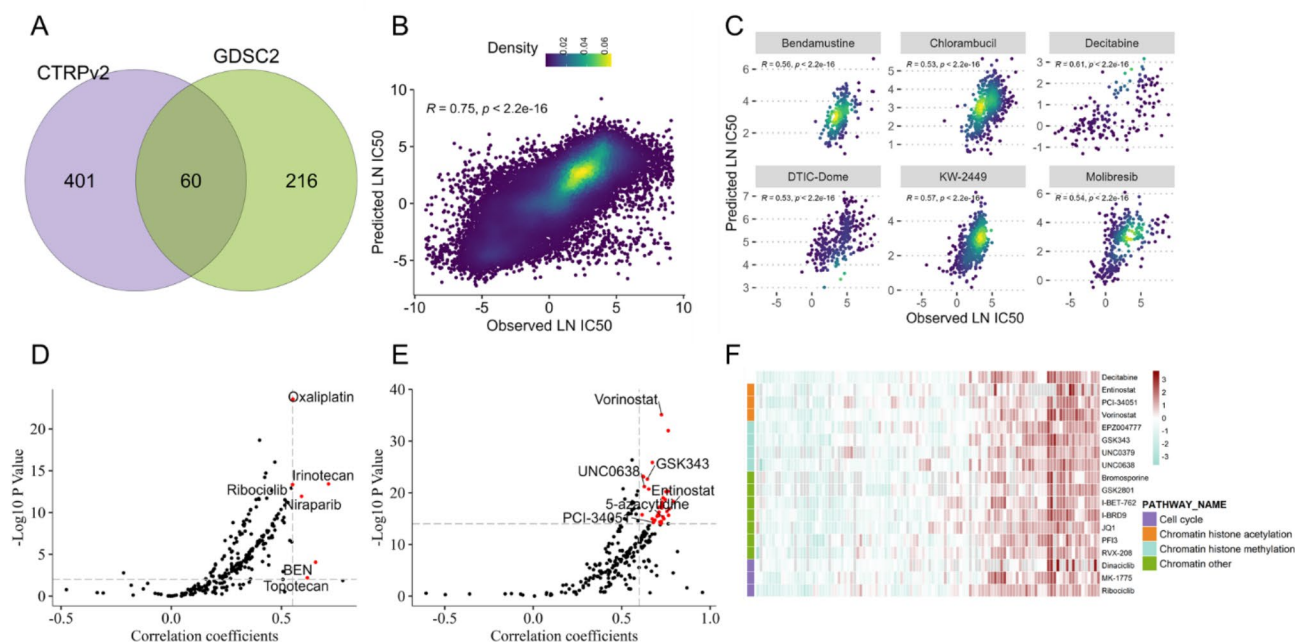


Fig. 4. DrugS model predicts mechanisms in CTRPv2. (A) Overview of the CTRPv2 dataset, highlighting the 804 cell lines and 461 compounds with a focus on the 60 compounds overlapping with the GDSC database.

(B) Demonstration of high correlation between predicted and observed LN IC₅₀ values for the overlapping compounds using the DrugS model, with statistical significance indicated. (C) Presentation of high correlation in predicted results for Bendamustine and Decitabine, etc., drugs not included in the training set, suggesting the model's robustness in predicting unknown compounds. (D) Identification of drugs highly correlated with Bendamustine, including Oxaliplatin, Irinotecan, Niraparib, and Topotecan, based on integrated IC₅₀ data from CTRPv2 and GDSC. (E) Prediction of Decitabine IC₅₀ using gene expression data from the CTRPv2 dataset and subsequent analysis of correlation with over 200 drugs in the GDSC database. (F) Suggestion of Decitabine's mechanism of action through chromatin modification, inferred from the correlation analysis with drugs known to affect similar pathways.

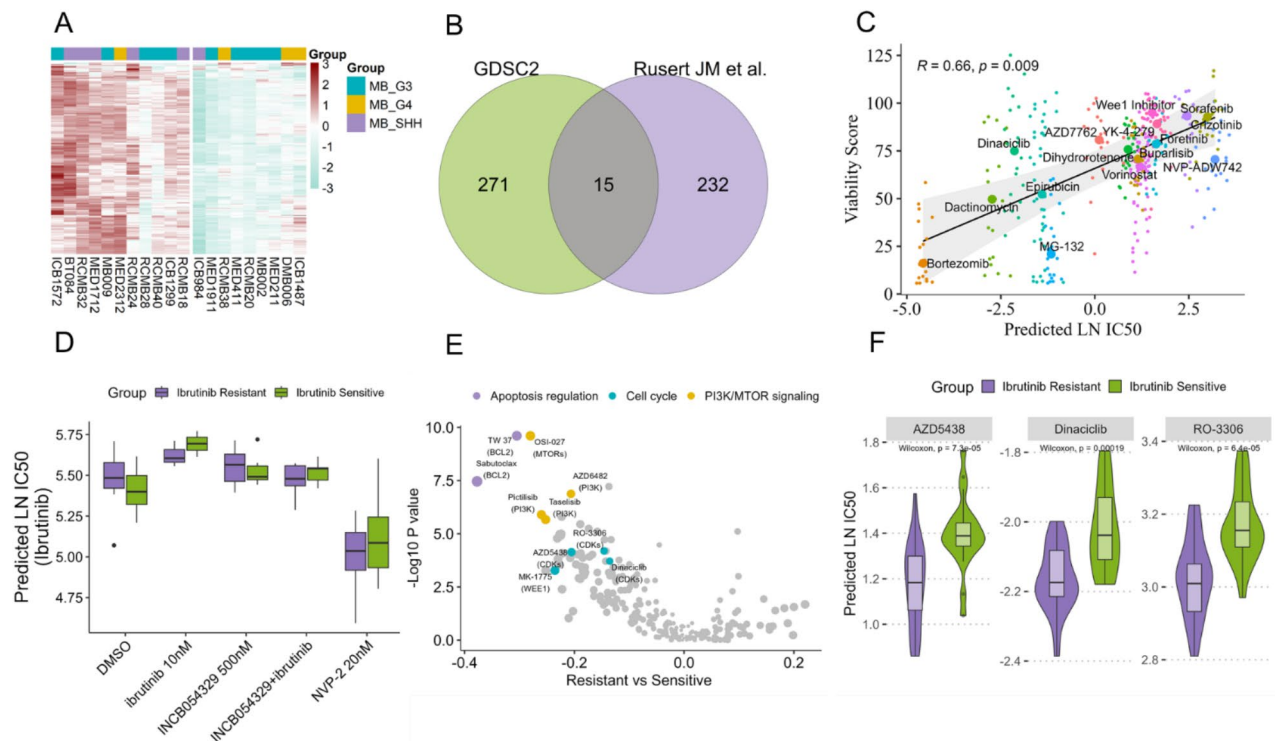


Fig. 5. DrugS model application in drug screening and resistance. **(A)** Clustering analysis following the application of the DrugS model to predict drug responses in 20 Medulloblastoma PDX samples. The analysis reveals that the G3/G4 subtypes are prone to be sensitive to drugs compared to the SHH subtype. **(B)** Results of in vitro drug screening with 7729 compounds on 20 PDX models, identifying 375 proliferation-inhibiting compounds, with 15 small molecules matching the GDSC dataset. **(C)** Correlation analysis between the LN IC50 values predicted by the DrugS model for the 15 overlapping small molecules and their corresponding viability scores. The analysis shows a significant correlation ($R = 0.66$, $P < 0.05$), validating the DrugS model's predictive power using microarray data and its potential application in drug screening endeavors. **(D)** DrugS model's validation through its prediction that NVP-2 can reverse Ibrutinib resistance in cell lines, a finding consistent with previous in vitro studies, thereby confirming the model's predictive accuracy. **(E)** Comparative analysis of drug response predictions for Ibrutinib-resistant and sensitive cell lines, indicating that, aside from cell cycle inhibitors like CDK inhibitors, drugs that target apoptosis regulation may also play a role in sensitizing resistant cell lines to Ibrutinib. **F** CDK inhibitors, AZD5438, Dinaciclib, RO-3306, could significantly enhance sensitivity to Ibrutinib in resistant cell lines.

realm of drug screening. Ibrutinib, a BTK inhibitor, is extensively utilized in the clinical management of chronic lymphocytic leukemia⁴⁴. However, resistance to Ibrutinib is a significant contributor to unfavorable prognoses in patients⁴⁵. Zhao et al. reported that NVP-2, a CDK inhibitor, has the potential to reverse Ibrutinib resistance in vitro⁴⁶. Utilizing the DrugS model and gene expression data from the same in vitro cell line, our prediction analysis corroborated these findings, indicating that CDK inhibitors like NVP-2 could notably increase cellular sensitivity to Ibrutinib, aligning with the existing literature (Fig. 5D). Further analysis comparing drug response predictions between Ibrutinib-resistant and sensitive cell lines unveiled that, aside from cell cycle-related inhibitors such as CDK inhibitors, compounds targeting apoptosis regulation and the PI3K/MTOR signaling pathways might also heighten sensitivity in Ibrutinib-resistant cell lines (Fig. 5E,F). These findings offer novel perspectives for investigating resistance mechanisms and for the development of combination therapy strategies.

Predicting drug responses and prognosis in TCGA patients

Using TCGAbiolinks, we downloaded clinical information and drug administration records from 32 TCGA projects, and procured patient gene expression data via the gdc-client tool. Cisplatin, is a platinum-based chemotherapy drug used to treat various types of cancers⁴⁷, had the most extensive record data available, involving 394 patients from TCGA (Fig. 6A,B). Based on the gene expression profiles of these patients and cisplatin's fingerprint data, we applied the DrugS model for prediction and found that the predicted LN IC50 values were significantly lower in responsive patients compared to non-responsive patients, indicating that the model's proficiency in identifying drug-sensitive patients (Fig. 6C).

From the TCGA dataset, we meticulously identified 602 patients with detailed follow-up data and a documented history of Cisplatin treatment. Using these patients' gene expression profiles, we predicted the LN IC50 values for cisplatin and stratified the patients into two groups based on the median predicted LN IC50

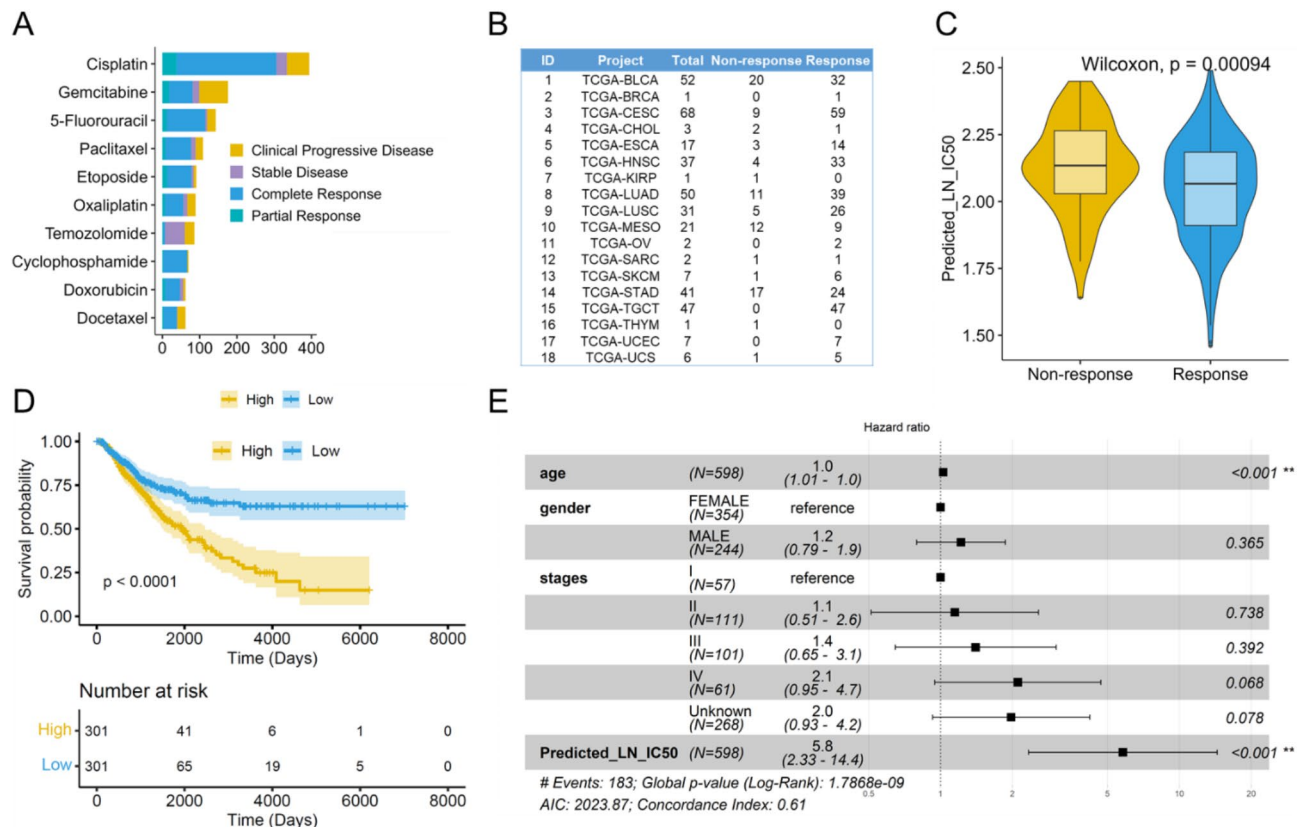


Fig. 6. DrugS model predicts cisplatin sensitivity and prognosis in TCGA patients. **(A)** Compilation of clinical information and drug administration data from 32 TCGA projects using TCGAbiolinks, emphasizing the extensive data available for Cisplatin, a widely used chemotherapy drug. **(B)** Analysis of gene expression data from 394 patients with Cisplatin treatment records, encompassing 18 different types of cancer. The Response group, comprising patients with Partial Response and Complete Response, and the Non-response group, which includes patients with Clinical Progressive Disease and Stable Disease. **(C)** DrugS model predictions of LN IC50 values for Cisplatin, showing significant differences between responsive and non-responsive patients, demonstrating the model's ability to identify drug sensitivity. **(D)** Stratification of 602 Cisplatin-treated patients based on median predicted LN IC50 values and subsequent survival analysis, revealing a correlation between lower LN IC50 values and better prognosis. **(E)** Multivariate COX regression analysis illustrating the close relationship between predicted IC50 values and patient prognosis, underscoring the predictive power of the DrugS model.

value. Survival analysis revealed a significant association between the predicted LN IC50 values and patient prognosis: patients with lower LN IC50 values demonstrated significantly better outcomes compared to those with higher values (Fig. 6D). Multivariate Cox regression analysis confirmed this association, demonstrating that predicted LN IC50 values were independently and significantly correlated with patient survival outcomes (Fig. 6E). Encouraged by these findings, we extended our analysis to other drugs using the DrugS model and conducted corresponding survival analyses. Consistently, patients with higher predicted LN IC50 values tended to have poorer prognoses (Supplementary Fig. 5). These results underscore the utility of the DrugS model not only in drug sensitivity prediction but also in prognostic assessment. This demonstrates the model's potential to provide valuable insights when applied to patient data from large-scale genomic initiatives like TCGA.

Discussion

In the realm of precision medicine, the clinical selection of targeted therapies is often predicated on the identification of specific genomic markers in cancer patients, such as KRAS mutations, BCR-ABL fusions, and EGFR mutations, etc. While targeted therapies have made significant strides, the existence of a vast array of non-targeted chemotherapeutic agents^{48,49} presents a complex landscape for clinicians. The accurate selection of suitable drugs and the execution of effective clinical treatments continue to pose formidable challenges^{50–52}. A substantial body of research has underscored the pivotal role of gene expression in dictating drug responses^{30,31}. In light of this, our study is poised to delve into the analysis of gene expression data, meticulously excluding outliers that may arise from genetic variations or experimental discrepancies. By scrutinizing drug response events within cellular or sample contexts, we endeavor to cultivate a DNN model with broad applicability, offering a novel methodology for drug screening. This approach is anticipated to markedly amplify our predictive

capabilities regarding drug responses in cancer patients, enabling the customization of therapeutic strategies that are aligned with individual genomic profiles.

We analyzed gene expression data from over 1000 cancer cell lines in the DepMap database using TSNE clustering, identifying nine clusters consistent with established cancer cell types, which validated our clustering method. To minimize the effects of genetic variability and outliers, we excluded data with significant response variability to the same drug within identical clusters. During this process, we noted anomalies with the drug SN-38 in several clusters. A detailed analysis of cluster 0 revealed that mutations in KRAS, EGFR, and NOTCH1, along with dysregulated PI3K-AKT signaling, likely contributed to observed drug response differences. These revelations highlight the intricate nature of drug sensitivity in cancer cells and emphasize the critical role of genomic data in deciphering the mechanisms that contribute to treatment variability. Our methodical approach lays a foundational framework for further dissection of the genetic underpinnings of drug resistance and sensitivity.

To ascertain the broader applicability of our model, we conducted predictive analyses leveraging the CTRPv2 and NCI-60 datasets. The model showed strong performance in CTRPv2, accurately predicting responses to known compounds and generating reasonable predictions for untested compounds, especially those with mechanisms of action similar to the training dataset. This predictive ability for previously uncharacterized compounds could help elucidate the mechanisms of action for drugs with unknown targets. Additionally, testing with the NCI-60 dataset also demonstrated high accuracy in predicting LN IC50 values. These results highlight the model's capacity to extend beyond the GDSC data, showcasing its adaptability to diverse datasets. The consistent performance across datasets underscores the DrugS model's value in pharmacological research and its potential to advance precision medicine by facilitating drug response prediction and personalized treatment planning.

To extend the application of the DrugS model beyond cell lines, we evaluated its utility in PDX models for drug response prediction and combination therapy screening. Using gene expression data from 20 PDX models, we predicted responses to 15 compounds, finding a strong correlation between predicted LN IC50 values and viability scores. For Ibrutinib-resistant cell lines, the model identified increased sensitivity to CDK inhibitors, suggesting that their combination with Ibrutinib could enhance therapeutic efficacy. Additionally, inhibitors targeting apoptosis and the mTOR pathway were found to improve cell sensitivity to Ibrutinib, potentially overcoming resistance. These findings highlight the DrugS model's ability to uncover synergistic drug combinations and its utility in developing personalized therapeutic strategies based on cancer-specific genomic features.

We further substantiated the applicability of our DrugS model using the expansive TCGA dataset. By harnessing gene expression data from TCGA samples, we discerned the model's capability to distinguish between potential cisplatin-sensitive and cisplatin-resistant patients. This triage was accomplished by leveraging the predictive power of DrugS, which aligned with the clinical outcomes of these patients. Integrating the follow-up data from TCGA, our analysis revealed a correlation between the predicted LN IC50 values and patient prognosis, where higher LN IC50 values were indicative of a poorer prognosis. The COX regression analysis provided a significant hazard ratio (HR) for the predicted LN IC50 values, which was greater than one, thereby classifying it as a potential risk factor for patient outcomes in the TCGA cohort. These analyses collectively suggest that the DrugS model extends its utility beyond preclinical settings, offering a viable tool for drug screening and for prognostication of drug treatment outcomes in clinical practice. However, acknowledging the scarcity of comprehensive clinical drug administration data within the TCGA dataset, the clinical relevance of our findings necessitates further validation. Future clinical studies will be pivotal in affirming the model's applicability and reliability in real-world clinical scenarios.

While the model achieves high predictive correlation with observed values, discrepancies for certain drugs highlight limitations. Specifically, its accuracy depends on the diversity and representativeness of drug information in the training dataset, with reduced performance for drugs or mechanisms underrepresented during training. Additionally, the model's reliance on gene expression as its sole biological input may bias predictions towards targeted therapies, potentially overlooking drugs whose effects stem from specific mutation sites or other genomic features not captured in expression profiles. Further refinement is needed to incorporate additional biological data, such as mutations or structural variations, to enhance predictive performance for such drugs.

In conclusion, the DrugS model marks a notable advancement in computational drug screening by providing a reliable and versatile platform for drug response prediction. Through comprehensive transcriptomic input and the integration of compound fingerprint data, the model offers a robust framework for preclinical drug screening, combination therapy design, and precision oncology. While challenges remain, including limited mechanistic interpretability and the need for clinical validation, this study highlights the transformative potential of the DrugS model in pharmacological research. Ongoing refinements and applications in clinical studies will be crucial to bridging preclinical insights with clinical practice, advancing the field of personalized medicine.

Methods

Data preprocessing for CCLE and GDSC

We employed the Seurat package⁵³ to preprocess CCLE data and perform clustering analysis. Initially, a data object was created using the CreateSeuratObject function, and cellular subgroups were identified through TSNE clustering with the RunTSNE function. To capture representative information, we focused on highly variable genes (HVGs) within the dataset. HVGs, critical for unsupervised clustering and dimensionality reduction, were identified using the FindVariableFeatures function, which applies a mean-variance model to detect genes with significant variability. The top 2000 HVGs were selected for clustering analysis. Subgroup-specific overexpressed genes were further identified using the FindAllMarkers function, followed by Gene Set Enrichment Analysis

(GSEA) through the ClusterProfiler package⁵⁴ to investigate their associated biological functions and pathways. The GDSC1 and GDSC2 datasets were downloaded from the GDSC website. For each compound, the corresponding SMILES strings were obtained from PubChem. Using these SMILES strings, Morgan fingerprints were generated via the RDKit package in Python. These fingerprints served as the input features for the DrugS model, enabling compound characterization and subsequent drug response predictions.

Data pruning to address outliers in drug response

To refine the dataset for model development, we leveraged t-SNE clustering results to assess drug response variability within each cluster. Specifically, for cells treated with the same drug, we calculated the standard deviation (SD) and mean of LN IC50 values within each cluster. Drug-cell assay pairs exhibiting an SD greater than 1.5 were identified as outliers and excluded from further analysis. This pruning step was essential to mitigate the impact of experimental anomalies or biological heterogeneity, thereby enhancing the predictive accuracy of the model. The SD of a drug's LN IC50 values within a cluster was calculated across all cell lines belonging to that cluster. High SD values suggested that factors beyond gene expression, such as genomic alterations (e.g., mutations or fusions), might be influencing drug response. Since cell lines within a cluster are characterized by similar gene expression patterns, their drug responses were expected to be relatively consistent. Drugs with highly variable LN IC50 values within a cluster were flagged as outliers and excluded. The threshold of SD > 1.5 was determined empirically based on the distribution of IC50 variability across clusters. As illustrated in Supplementary Fig. 2D, 75% of drugs exhibited SD values below this threshold, ensuring that the filtering process retained the majority of the data while removing highly dispersed responses. By applying this threshold, we preserved data reflecting consistent biological patterns and excluded highly variable points that could compromise model performance.

Drug response retrospective analysis in cancer cell lines

We extracted gene expression and mutation data for 212 cell lines in Cluster 0 from DepMap. Cell lines were grouped into mutant and wildtype categories based on mutations, and Wilcoxon tests identified mutations associated with sensitivity or resistance to SN-38. Cell lines were further classified into resistant (LN IC50 > 0) and sensitive groups for additional statistical analysis. Statistical analysis was conducted using edgeR⁵⁵ to discern gene expression alterations and pathway activity changes correlated with SN-38 sensitivity or resistance.

CTRPv2 and NCI-60 data preprocessing and testing

Utilizing the PharmacoGx⁵⁶ package, we downloaded the CTRPv2 and NCI-60 datasets. For the 461 compounds within the CTRPv2 dataset, SMILES were retrieved from PubChem and converted to Morgan fingerprints using RDKit for IC50 predictions across 804 cell lines. Pearson correlation was used to evaluate prediction relevance. Similarly, IC50 data for 122 drugs from NCI-60 were used to test the DrugS model.

Development of the DNN model

Autoencoder for dimensionality reduction

To process gene expression data, an autoencoder with two layers (2048 and 30 neurons) was constructed using the TensorFlow library in Python. The exponential linear unit (ELU) activation function was utilized to introduce non-linearity. Transcript-per-million (TPM) values were log2-transformed and z-score normalized to ensure consistency across samples. This autoencoder effectively reduced the high-dimensional gene expression data into 30 essential features, facilitating downstream analysis.

DNN model architecture and training

The pharmacological response prediction model was developed using a DNN framework implemented in TensorFlow. The input consisted of 2078 features, combining 30 gene expression features distilled from the autoencoder and 2048 compound features derived from Morgan fingerprints. The DNN architecture included:

- Input Layer: 2078 neurons.
- Hidden Layer 1: 1024 neurons with ReLU activation and a 0.1 dropout rate.
- Hidden Layer 2: 16 neurons with ReLU activation and a 0.1 dropout rate.
- Output Layer: A single neuron for predicting LN IC50 values.

The model was optimized using the Adam optimizer with a learning rate of 0.001 and evaluated using the mean square error (MSE) metric. A 7:3 data split strategy was adopted, allocating 70% of the data for training and 30% for validation. Model testing was performed on external datasets, including GDSC1, NCI-60, and CTRPv2.

Model selection and hyperparameter optimization

To identify the optimal model architecture, we first leveraged the H2O AutoML framework for preliminary evaluation. This involved pre-training 30 models, including XGBoost, Distributed Random Forest (DRF), Generalized Linear Model (GLM), and DNNs, using 5-fold cross-validation. The DNN model was selected based on superior performance metrics, particularly the lowest MSE. Further refinement of the DNN model was conducted using the Keras library. The Keras Tuner was employed for hyperparameter optimization, performing a grid search over dropout rates, the number of layers, nodes per layer, and learning rates. This iterative process ensured the final DNN model was fine-tuned for optimal performance in predicting drug responses.

Model benchmarking

The performance of the DrugS model was benchmarked against two established models, Precily and CaDRReS-SC, using data from the GDSC2 and GDSC1 datasets. The CaDRReS-SC model was retrained using the GDSC2

dataset and subsequently tested on the GDSC1 dataset. Predicted LN IC50 values were compared with observed values, and model performance was evaluated using MSE and PCC. For the Precily model, cell line GSVA scores and drug-related features were utilized as inputs to predict LN IC50 values, with MSE and PCC calculated in the same manner. Additionally, for each model, PCC and MSE values were computed across different cancer types to assess the predictive performance in various cancer backgrounds.

TCGA data prediction and prognosis analysis

Gene expression data were downloaded using gdc-client.exe, and clinical/drug information was retrieved via TCGAAbiolinks⁵⁷. TPM values were log2-transformed for compatibility with the DrugS model. LN IC50 predictions for Cisplatin were used to stratify 602 patients into High and Low groups based on the median value. Survival analysis was performed with survfit and cph functions from the survival package⁵⁸, and a nomogram was constructed using the rms package to estimate 1-year and 5-year survival probabilities.

Data availability

CCLE expression data were available from <https://depmap.org/portal/>. GDSC data were available from <https://www.cancerrxgene.org/>. TCGA expression data were available from: <https://portal.gdc.cancer.gov/>. PDX data were available in the NCBI GEO database under accession code GSE151343. Ibrutinib resistant cancer cell lines data were available in the NCBI GEO database under accession code GSE141333. All relevant data are available from Qiangqiang Fan (dongliulou@126.com). Source data are provided with this paper.

Received: 24 September 2024; Accepted: 21 February 2025

Published online: 01 July 2025

References

- Gibbs, S. N. et al. Comprehensive review on the clinical impact of next-generation sequencing tests for the management of advanced cancer. *JCO Precis. Oncol.* **7**, e2200715 (2023).
- Berger, M. F. & Mardis, E. R. The emerging clinical relevance of genomics in cancer medicine. *Nat. Rev. Clin. Oncol.* **15**, 353–365 (2018).
- Morganti, S. et al. Next-generation sequencing (NGS): A revolutionary technology in pharmacogenomics and personalized medicine in cancer. *Adv. Exp. Med. Biol.* **1168**, 9–30 (2019).
- Webber, J. T., Kaushik, S. & Bandyopadhyay, S. Integration of tumor genomic data with cell lines using multi-dimensional network modules improves cancer pharmacogenomics. *Cell. Syst.* **7**, 526–536 (2018).
- Heo, Y. J., Hwa, C., Lee, G. H., Park, J. M. & An, J. Y. Integrative multi-omics approaches in cancer research: from biological networks to clinical subtypes. *Mol. Cells.* **44**, 433–443 (2021).
- Yu, K. et al. Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nat. Commun.* **10**, 3574 (2019).
- Dong, H. & Wang, S. Exploring the cancer genome in the era of next-generation sequencing. *Front. Med.* **6**, 48–55 (2012).
- Abdalla, M. et al. Mapping genomic and transcriptomic alterations spatially in epithelial cells adjacent to human breast carcinoma. *Nat. Commun.* **8**, 1245 (2017).
- Troester, M. A. et al. DNA defects, epigenetics, and gene expression in cancer-adjacent breast: A study from the Cancer genome atlas. *NPJ Breast Cancer.* **2**, 16007 (2016).
- Kim, K. et al. Spatial and clonality-resolved 3D cancer genome alterations reveal enhancer-hijacking as a potential prognostic marker for colorectal cancer. *Cell. Rep.* **42**, 112778 (2023).
- Braun, T. P., Eide, C. A. & Druker, B. J. Response and resistance to BCR-ABL1-targeted therapies. *Cancer Cell.* **37**, 530–542 (2020).
- Shitara, K. & Bang, Y. J. Trastuzumab Deruxtecan in previously treated HER2-positive gastric cancer. *N. Engl. J. Med.* **382**, 2419–2430 (2020).
- Punekar, S. R., Velcheti, V., Neel, B. G. & Wong, K. K. The current state of the Art and future trends in RAS-targeted cancer therapies. *Nat. Rev. Clin. Oncol.* **19**, 637–655 (2022).
- Zhao, W. et al. Large-scale characterization of drug responses of clinically relevant proteins in cancer cell lines. *Cancer Cell.* **38**, 829–843 (2020).
- Garnett, M. J. et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
- Rees, M. G. et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat. Chem. Biol.* **12**, 109–116 (2016).
- Pacini, C. et al. A comprehensive clinically informed map of dependencies in cancer cells and framework for target prioritization. *Cancer Cell.* **42**, 301–316 (2024).
- Barretina, J. et al. The Cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
- Yang, W. et al. Genomics of drug sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–D961 (2013).
- Iorio, F. et al. A landscape of Pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
- Seashore-Ludlow, B. et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov.* **5**, 1210–1223 (2015).
- Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer.* **6**, 813–823 (2006).
- Costello, J. C. et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **32**, 1202–1212 (2014).
- Jia, P. et al. Deep generative neural network for accurate drug response imputation. *Nat. Commun.* **12**, 1740 (2021).
- Sakellaropoulos, T. et al. A deep learning framework for predicting response to therapy in cancer. *Cell. Rep.* **29**, 3367–3373 (2019).
- Suphailai, C. et al. Predicting heterogeneity in clone-specific therapeutic vulnerabilities using single-cell transcriptomic signatures. *Genome Med.* **13**, 189 (2021).
- Chawla, S. et al. Gene expression based inference of cancer drug sensitivity. *Nat. Commun.* **13**, 5680 (2022).
- Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).
- Li, Y. et al. Predicting tumor response to drugs based on gene-expression biomarkers of sensitivity learned from cancer cell lines. *BMC Genom.* **22**, 272 (2021).
- Geeleher, P., Cox, N. J. & Huang, R. S. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol.* **15**, R47 (2014).

31. Plana, D., Palmer, A. C. & Sorger, P. K. Independent drug action in combination therapy: implications for precision oncology. *Cancer Discov.* **12**, 606–624 (2022).
32. Cheng, Y. et al. Targeting epigenetic regulators for cancer therapy: mechanisms and advances in clinical trials. *Signal. Transduct. Target. Ther.* **4**, 62 (2019).
33. Ippolito, M. R. et al. Gene copy-number changes and chromosomal instability induced by aneuploidy confer resistance to chemotherapy. *Dev. Cell.* **56**, 2440–2454 (2021).
34. Schram, A. M., Chang, M. T., Jonsson, P. & Drilon, A. Fusions in solid tumours: diagnostic strategies, targeted therapy, and acquired resistance. *Nat. Rev. Clin. Oncol.* **14**, 735–748 (2017).
35. Wu, D., Duan, C., Chen, L. & Chen, S. Efficacy and safety of different doses of cytarabine in consolidation therapy for adult acute myeloid leukemia patients: a network meta-analysis. *Sci. Rep.* **7**, 9509 (2017).
36. Nguyen, F. et al. Enhanced intratumoral delivery of SN38 as a Tocopherol oxyacetate prodrug using nanoparticles in a neuroblastoma xenograft model. *Clin. Cancer Res.* **24**, 2585–2593 (2018).
37. Lalic, H. et al. Bendamustine: A review of pharmacology, clinical use and immunological effects. *Oncol. Rep.* **47**, 114 (2022).
38. Mauri, G. et al. Oxaliplatin retreatment in metastatic colorectal cancer: systematic review and future research opportunities. *Cancer Treat. Rev.* **91**, 102112 (2020).
39. Kciuk, M., Marciniak, B. & Kontek, R. Irinotecan—Still an important player in cancer chemotherapy: A comprehensive overview. *Int. J. Mol. Sci.* **21**, 4919 (2020).
40. González-Martín, A., Pothuri, B. & Vergote, I. Niraparib in patients with newly diagnosed advanced ovarian cancer. *N Engl. J. Med.* **381**, 2391–2402 (2019).
41. Horita, N. et al. Topotecan for relapsed small-cell lung cancer: systematic review and meta-analysis of 1347 patients. *Sci. Rep.* **5**, 15437 (2015).
42. Xie, J. et al. Venetoclax with decitabine as frontline treatment in younger adults with newly diagnosed ELN adverse-risk AML. *Blood* **142**, 1323–1327 (2023).
43. Ruser, J. M. et al. Functional precision medicine identifies new therapeutic candidates for Medulloblastoma. *Cancer Res.* **80**, 5393–5407 (2020).
44. Zhou, H., Hu, P., Yan, X., Zhang, Y. & Shi, W. Ibrutinib in chronic lymphocytic leukemia: clinical applications, drug resistance, and prospects. *Onco Targets Ther.* **13**, 4877–4892 (2020).
45. Wang, E. et al. Mechanisms of resistance to noncovalent Bruton's tyrosine kinase inhibitors. *N. Engl. J. Med.* **386**, 735–743 (2022).
46. Zhao, X. et al. Transcriptional programming drives ibrutinib-resistance evolution in mantle cell lymphoma. *Cell. Rep.* **34**, 108870 (2021).
47. Amable, L. Cisplatin resistance and opportunities for precision medicine. *Pharmacol. Res.* **106**, 27–36 (2016).
48. Anand, U. et al. Cancer chemotherapy and beyond: current status, drug candidates, associated risks and progress in targeted therapeutics. *Genes Dis.* **10**, 1367–1401 (2022).
49. Shafagati, N. et al. Comparative efficacy and tolerability of novel agents vs chemotherapy in relapsed and refractory T-cell lymphomas: A meta-analysis. *Blood Adv.* **6**, 4740–4762 (2022).
50. Goetz, L. H. & Schork, N. J. Personalized medicine: motivation, challenges, and progress. *Fertil. Steril.* **109**, 952–963 (2018).
51. Shin, S. H., Bode, A. M. & Dong, Z. Addressing the challenges of applying precision oncology. *NPJ Precis. Oncol.* **1**, 28 (2017).
52. Masucci, M., Karlsson, C., Blomqvist, L. & Ernberg, I. Bridging the divide: A review on the implementation of personalized cancer medicine. *J. Pers. Med.* **14**, 561 (2024).
53. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
54. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. ClusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
55. Robinson, M. D., McCarthy, D. J., Smyth, G. K. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
56. Smirnov, P. et al. PharmacGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics* **32**, 1244–1246 (2016).
57. Colaprico, A. et al. TCGAAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71 (2016).
58. Therneau, T. A Package for Survival Analysis in R. R package version 3.7-0. <https://CRAN.R-project.org/package=survival> (2024).

Acknowledgements

I am deeply grateful to my family for their patience, understanding, and constant encouragement, which made this work possible.

Author contributions

G.Y. and Q.F. contributed equally to this work. G.Y. responsible for data collection, manuscript writing, and submission. Q.F. responsible for data analysis and manuscript formatting.

Declarations

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/s41598-025-91571-2>.

Correspondence and requests for materials should be addressed to Q.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025