An interpretable deep learning framework for genome-informed precision oncology

Shuangxia Ren [1], Gregory F Cooper [1,2], Lujia Chen[2], Xinghua Lu[1,2*,]

1 Intelligent Systems Program, School of Computing and Information, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America

2 Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America

∗ Correspond to: xinghua@pitt.edu

Abstract

Cancers result from aberrations in cellular signaling systems, typically resulting from driver somatic genome alterations (SGAs) in individual tumors. Precision oncology requires understanding the cellular state and selecting medications that induce vulnerability in cancer cells under such conditions. To this end, we developed a computational framework consisting of two components: 1) A representation-learning component, which learns a representation of the cellular signaling systems when perturbed by SGAs, using a biologically-motivated and interpretable deep learning model. 2) A drug-response-prediction component, which predicts the response to drugs by leveraging the information of the cellular state of the cancer cells derived by the first component. Our cell-state-oriented framework significantly enhances the accuracy of genome-informed prediction of drug responses in comparison to models that directly use SGAs as inputs. Importantly, our framework enables the prediction of response to chemotherapy agents based on SGAs, thus expanding genome-informed precision oncology beyond molecularly targeted drugs.

Introduction

Precision medicine utilizes genomic and other advanced technologies to define diseases at a more detailed level than before, enabling tailored therapies for individuals[1,2]. This approach largely relies on understanding the impact of genomic alterations within cells and prescribing medications to counteract aberrant signals caused by these alterations. The common practice of genome-informed precision oncology is to examine the somatic genome alterations (SGAs) and match patients with targetable SGAs to corresponding targeted drugs [1,3,4]. While of clinical value, this approach is applicable to a relatively small number of molecularly targetable drugs, patient coverage is relatively low, and prediction accuracy (positive predictive value) remains modest [5-7]. Marquart et al [5] reported that as of 2018, the percentage of patients who receive genomic screening and could be matched with targeted therapies was only about 15%; the median overall response rate to all genome-informed therapies was 54%; and the percentage of all cancer patients estimated to benefit was about 7%. Thus, the current practice is insufficient to meet the needs of precision oncology for the general cancer population.

Although chemotherapies remain the backbone of general oncology, their application is largely not guided by genomic information. Recently, Liu et al [8] systematically studied mutation-treatment interactions based on real-world patient data and discovered that certain mutations are associated with responses to certain chemotherapy agents. Generally speaking, a "mutation-to-treatment" rule for guiding molecularly targeted or chemotherapeutic agents fails to consider that multiple SGAs in a cancer cell may influence the cellular state and, thereby, drug responses, which may contribute to the observed low accuracy [5] of the current genome-

45 informed precision oncology. Thus, there is an urgent, unmet need to develop comprehensive
46 clinical decision support systems (CDSSs) capable of utilizing genome-scale omics profiles of
47 tumors to guide the selection of effective anticancer drugs from the entire pool of FDA-
48 approved agents.
49
50 Developing a CDSS for guiding all anticancer drugs in pan-cancer patients using real-patient data
51 remains challenging because it would require large-scale randomized trials testing many drugs
52 in all cancer types, which is not feasible. To address the challenge, large-scale pre-clinical
53 models screening anticancer-drug sensitivity have been developed by the Genomics of Drug
54 Sensitivity in Cancer (GDSC) [9,10] and the Cancer Cell Line Encyclopedia [11]. The GDSC project has
55 examined multi-omics profiles of close to a thousand cancer cell lines and recorded their
56 response to hundreds of drugs. This dataset fills the gaps for developing artificial intelligence
57 (AI) models for pan-cancer and pan-drug precision oncology. GDSC studies indicate that
58 transcriptomes of cell lines are more informative features than SGAS in predicting cell line drug
59 sensitivity. However, in clinical practice, genomic data are more readily available, and thus
60 effectively utilizing such information would be of high clinical value. Therefore, we set out to
61 develop a computational framework to predict drug sensitivity based on SGA data of cell lines.
62
63 Developing a genome-based CDSS faces several challenges: 1) Drug responses are usually
64 determined by the state of multiple signaling pathways in a cancer cell. Therefore, the genomic
65 status of individual genes considered in isolation is insufficient to predict drug sensitivity; 2) A
66 signaling pathway can be perturbed by SGAs affecting different member genes in the pathway

that bear similar consequences on drug responses; and, 3) the SGAs perturbing a common signaling pathway tend to be mutually exclusive in individual tumors [12,13]. As such, the signal of one SGA on a drug response may become noise when training a model learning the signal of another SGA on the same drug.

To overcome the above challenges, we developed an AI system that first transforms the SGA data of cancer cells into a representation of cellular signaling systems and then learns to predict the drug responses of the cells based on the inferred cellular states. The framework consists of two main modules: 1) A representation-learning module using the Residual Genome Impact Transformer (ResGit) model (**Fig. 1C**), which infers the cellular states based on the SGAs of a cancer cell line, and 2) a drug-response-prediction module (**Fig. 1D**), which predicts the cells' responses to drugs based on the inferred cellular states. The combined system is referred to as the ResGit-based Drug Response Prediction (ResGitDR) model (**Fig. 1A**). We show that by more closely mimicking the cellular signaling systems, the ResGit model can learn interpretable and biologically sensible representations of the impact of SGAs on cellular signaling systems. We also show that by considering cellular states, the ResGitDR performs better in predicting drug response to both molecularly targeted and chemotherapy agents than the models that only use SGAs as inputs. Finally, we show that ResGitDR indeed takes advantage of the cellular states learned within our framework and performs state-oriented predictions. The results presented below support that the ResGitDR framework provides a new and promising direction for developing biologically motivated and interpretable systems for predicting drug responses.

Result

Overview of the ResGitDR model

Heterogeneous responses to a drug by different cancer cells can be attributed to the heterogeneity of cellular states, which are driven by distinct causal SGAs that perturb cellular signaling systems. Thus, the capability of inferring cell states of cancer cells based on their SGAs lays a foundation for predicting drug responses. Based on the assumption that driver SGAs eventually influence gene expression, we designed ResGit (**Fig. 1C**) to model the relationships between SGAs and gene expression. It uses hierarchically organized latent variables to represent the cellular signaling system of cells and encode the impact of SGAs[14]. It then transforms the encoded information to predict gene expression.

Specifically, for each tumor, a binary vector indicating which genes are perturbed by SGA events is fed into ResGit to predict gene expression. Then four distinct embedding layers are applied to convert the binary vector into four hidden-layer-specific SGA embedding matrices, which represent the impact of SGAs in a tumor on the signal-encoding hierarchy. Each SGA embedding matrix is fed through a multi-head self-attention component to derive tumor-specific signal embedding (e_i), representing the integrated impact of SGAs in a tumor on the signaling systems. The state of an internal hidden layer (H_i) is a function of signal embedding (e_i) and the state of the previous layer (H_{i-1}). To incorporate the knowledge of transcription factors (TFs) on gene expression, we instantiated the final hidden layer based on prior knowledge following the example by Tao et al [15], such that the parameters associated with known TF-gene edges are updated during training, and the rest is set to 0. ResGit is trained with SGA and expression data

of TCGA tumors and GDSC cell lines. To predict the drug response, we trained an elastic network model [16] for each drug. We combined the inferred state of the latent variables (reflecting cellular states) from ResGit and SGAs of cell lines as inputs and binarized drug sensitivity as the target (**Fig. 1D**). In the testing phase, as shown in **Fig. 1B**, the trained ResGit model is firstly used to obtain hidden representations by taking SGAs and cancer type as input, no gene expression data is needed during this process. Then these hidden representations are then combined with SGAs to predict drug response.

ResGit learns to encode the impact of SGAs and transforms it into the gene expression of tumors and cancer cell lines.

We collected SGA and gene expression data from 8,586 TCGA tumors and 976 cancer cell lines studied by GDSC. We trained the ResGit model using this combined dataset through a series of experiments. We evaluated model performance using the Spearman correlation coefficients between predicted and observed gene expression values of a gene as the performance metric. The distributions for the coefficients in different cancer types are shown as box plots in **Fig. 2A&B**. The mean correlation in TCGA is 0.8, while in GDSC is 0.72. The results indicate that ResGit can accurately map SGA input data to gene expression predictions. The results support that the latent variables in the model encode the impact of SGAs on the cellular signaling system and translate the information of SGAs to gene expression. Interestingly, when modeled separately, the GDSC dataset exhibited lower Spearman correlations than the TCGA dataset, which suggests that the larger sample size in the TCGA dataset made the prediction more

132 robust, resulting in higher correlation values. From here on, we report the results of ResGit
133 trained with pooled TCGA and GDSC data.
134
135 **ResGit captures biologically sensible representations of SGAs.**
136 In the ResGit model, an SGA is designed to be connected to every latent variable in the signaling
137 hierarchy, and SGA embeddings represent the impact of the SGA on the system, and the model
138 learns "optimal" connections between SGA and hidden nodes that would predict gene
139 expression well. If two SGAs affect distinct members of a common pathway, their impact on the
140 cellular signaling system should be similar, i.e., their embedding should be similar. We
141 examined all pairwise similarities of SGA embeddings using cosine similarity. We identified the
142 top 10 neighbor SGAs for each SGA and examined whether they perturb a common signaling
143 pathway according to existing knowledge.
144
145 Sanchez-Vega *et al*.[17] had reported SGAs perturbing ten major cancer pathways, which was used
146 as ground truth for evaluating our results. We constructed a connectivity graph among 64 SGAs
147 gene found in both our dataset and the reported cancer pathways gene by Sanchez-Vega *et al*.,
148 where an edge was added between a pair of SGAs if one (or both) of them was among the
149 neighbors of the other. We colored the edges with a pseudo-color corresponding to a pathway
150 if the connected SGAs were in a pathway (**Fig. 2C**). The learned embeddings of the members of
151 the PI3K pathway *PIK3CA, PIK3R1, PTEN,* and *AKT1* are among the closest neighbors to each
152 other. The graph also shows similar results for other cancer pathways. The results indicate that

ResGit has learned embeddings of SGAs reflecting their similar impact on cell signaling systems, conforming to established knowledge.

Self-attention mechanism revealed the impact of SGAs in cancers.

ResGit employs self-attention mechanisms and assigns a tumor-specific attention weight to an SGA observed in a cell line to reflect its relative importance. Collective attention assigned to an SGA reflects its importance in influencing gene expression in cancers (**Fig. 2D**) or in different cancer types (**Fig. 2E**). As shown in **Fig. 2D**, ResGit assigned high attention values to well-known cancer drivers[18], such as *TP53*, *PTEN*, *KRAS*, *BRAF*, etc. Interestingly, some genes encoding signaling proteins, such as G-proteins *GNAQ* and *GNA11,* are not well-known as "cancer drivers" but were assigned with high attention weight, despite their relatively low frequencies. The results suggest ResGit captures their impact on gene expression of cells and potential role in cancers, which is supported by recent research indicating they may play an essential role in the tumorigenesis [19]. Our analysis also revealed the importance of SGAs in different cancer types (**Fig. 2E**). For example, the results show that SGA events in *GATA3* play a significant role in breast cancer (BRCA), as confirmed by Takaku *et al*. [20]; SGAs in *DHX9 and KEAP1* appear to play a significant role in lung cancer (LUAD), aligning with previous studes[21,22]; alterations in *TP53* are universally involved in most cancers, as demonstrated by earlier research[23].

The latent representation of the cellular system is informative of drug sensitivity.

The results above indicate that ResGit can encode the signals perturbed by the SGAs using the latent variables in the deep learning model. We then set out to test whether the information

represented by the latent variables can be used to predict cancer cell responses to anticancer drugs.

As a baseline, we used SGAs and cancer-type labels as input to train an elastic network model (EN, **Supplementary Fig. S1A**) and an end-to-end feedforward neural network (NN, **Supplementary Fig. S1B**) model to predict cell sensitivity to each drug tested by GDSC. We evaluated the performance of each model in 10-fold cross-validation experiments. The EN and NN models for 367 drugs achieved moderate performance in terms of area under the receiver operating curve (AUROC) (**Fig. 3A**), with median AUROC at 0.595 and 0.619 for the NN and EN, respectively. We arbitrarily set the threshold that an AUROC of 0.7 indicates a potentially useful model in the clinical setting. The total number of models with AUROC above 0.7 is 7 and 32 for NN and EN, respectively. Interestingly, in this setting, the elastic network outperforms the neural network model, suggesting it is more robust in a setting with a small training sample size.

We then examined whether the latent representation learned by ResGit is informative with respect to drug sensitivity. In a 10-fold cross-validation experiment, we trained ResGit and retrieved the estimated states of latent variables (H_1 – H_3, *and TF*, **Fig. 1C**) for the GDSC cell line in the training dataset. We concatenated the states of the latent variables with the original SGAs of each cell line as input features and trained an elastic network model for each drug (**Fig. 1D**). We called these models the ResGit-based Drug Response prediction model (ResGitDR). To examine the value of self-attention and other unique approaches of ResGit, we also trained a conventional neural network to model the relationship between SGAs and gene expression

without direct connections from SGAs to internal latent nodes or self-attention. We extract the estimated hidden-node states to train an elastic net model, and we call this model the neural-network-based drug response prediction model (NNDR, as shown in **Supplementary Fig.S1C**). The median AUROCs of the models are 0.667 and 0.633 for ResGitDR and NNDR, respectively (**Fig. 3B**), which are significantly higher than EN and NN (ResGitDR vs. each of the rest, $p < 0.01$).

The numbers of models with AUROC greater than 0.7 are 117 and 63 for the ResGitDR and NNDR, respectively, and the detailed information about these drugs are listed in **Supplementary Table. S1**. Compared to the EN model, which only uses the original SGAs and cancer type as features, including the states of latent variables in ResGitDR and NNDR led to 3.7 and 2-fold increases in the number of models with AUROC greater than 0.7. We further examined models' performances for targeted therapy and chemotherapy drugs by the four methods as an indication of what information is provided by input features and captured by the models (**Fig. 3C**). The number of ResGitDR models for targeted therapy agents with an AUROC larger than 0.7 is 72, which is 1.7-fold that of NNDR and 2.7-fold that of the EN model. Importantly, the results show that for many chemotherapy drugs, ResGitDR achieved comparable performance in terms of AUROC when compared with molecularly targeted drugs. The number of ResGitDR models for chemotherapy drugs with AUROC above 0.7 is 45, which is 2.1-fold of NNDR and 9-fold of the EN model. The results indicate that it is possible to perform genome-informed precision chemotherapy, beyond molecularly targeted drugs.

To examine the potential clinical utility of ResGitDR, we performed a simulated clinical decision experiment of assigning FDA-approved drugs to cell lines based on FDA guidelines and compared it with decisions by ResGitDR. There are 61 FDA-approved drugs (different drug_id in GDSC), 39 are for targeted drugs, and 22 are for chemotherapy agents. We applied the FDA guidelines based on cancer types and genomic biomarkers, with a preference for targeted therapy over chemotherapy. For example, the targeted therapy lapatinib is assigned to LUAD cell lines hosting SGAs in *EGFR*. If multiple drugs are eligible for a cell line, we select the one with the highest response rate among cell lines of a given cancer type, with a preference for targeted drugs over chemotherapy ones. We compared the positive predictive values (PPVs) of simulated FDA-guideline-based decisions and ResGitDR decisions.

As shown in **Fig. 3D**, in the majority of cancer types, such as MM, SKCM, LUAD, SCLC, NB, BRCA, HNSC, KIRC, LAML, PAAD, PRAD, and OV, ResGitDR predictions would make better recommendations on average. The FDA rules perform better than ResGitDR in a few cancer types, such as CESE, LUSC, ESCA, LGG, THCA, LCML, and MESO. The average PPV across all cancer types for ResGitDR and FDA rules are 0.761 and 0.549, separately. Interestingly, all OV cell lines have *BRCA1* and/or *BRCA2* mutations, and rucaparib was assigned to these cell lines per FDA rules, but these cell lines didn't respond to this drug, leading to a PPV of zero. Similarly, cell lines in STAD were assigned with sunitinib according to the above rules and got zero positive predicted value.

To illustrate the utility of our two-component framework of first learning representation of cellular systems using gene expression as objectives and then performing cell-state-oriented drug-response prediction, we also trained a model with the same architecture as ResGit to predict drug sensitivity directly, referred to as SGA2DR model (**Fig. 4A**). The performance of SGA2DR model was worse (mean AUROC 0.602) (**Fig. 4C**) than that of ResGitDR, indicating that learning relationships between SGA and gene expression led to a better representation of cellular states that enhanced the performance of downstream drug sensitivity prediction. Further, we trained a multi-task learning model, which aimed to predict gene expression and drug response simultaneously (**Fig. 4B**). Interestingly, this model performs better (mean AUROC 0.635) than the aforementioned SGA2DR model, indicating that including gene expression as an object led to a better representation that enhanced drug response prediction. However, the multi-task model's performance was inferior in predicting drug sensitivity compared to the two-stage approach of ResGitDR (**Fig. 4C**). This could be due to the limited size of our dataset, which consisted of only around 1000 samples. With its increased number of parameters, the multi-task model is prone to overfitting.

Finally, as a control, we shuffled SGAs and cancer-type data and re-trained a ResGitDR to predict drug sensitivity. As anticipated, the average AUROC dropped to 0.5, indicating that ResGitDR captures the “true” impact of SGAs and cancer type, which is required for predicting drug response (**Supplementary Fig. S2)**.

ResGitDR predicts responses to molecularly targeted drugs in a cell-state-oriented fashion.

Contemporary genome-informed precision oncology assigns treatment based on the genomic status of targeted signaling proteins. We evaluated the utility of genomic biomarkers for drugs targeting the PI3K/mTOR pathway, more specifically, PIK-93 and AKT inhibitor VIII, by examining whether cell lines carrying SGAs in these member genes are more sensitive (lower IC50s) than general cell lines (**Fig. 5A&B**). The results show that none of the SGAs in the pathway is informative of the sensitivity of the drugs when measured by IC50, whereas the cell lines predicted to be sensitive to the drugs by the ResGitDR models exhibit significantly lower IC50 (more sensitive). The results suggest that by considering the inferred cellular states, ResGitDR performed better in predicting molecularly targeted drugs than the conventional genomic biomarkers.

We then investigated whether ResGitDR utilized certain characteristic cellular states to predict responses to drugs that share similar mechanisms of action (MOA), e.g., drugs targeting the PI3K/mTOR pathway. We extracted the parameters from the models for three drugs, *ATK inhibitor VIII.1*, *PIK-93*, and *GSK690693*, and we identified a union of the top 50 features based on the absolute weights of drugs targeting on PI3K/mTOR pathway in the elastic net model, which reflect the importance of a feature, including both hidden representations and SGAs. We extracted the values of these features from GDSC cell lines and grouped them using clustering analysis (**Fig. 5C**). The cell lines' mutation status of genes in the PI3K/mTOR signaling pathway is shown to illustrate whether they carry information with respect to drug sensitivity as biomarkers. The figure shows that inferred cell states underlie cell line clusters consisting of cells from diverse cancer types, and certain clusters (e.g., clusters 3, 10, and 11) are enriched

283 with responders to the three drugs, supporting the notion that cell states influence the
284 response to drugs. The AUROCs for the three models are 0.81, 0.78, and 0.76 for ATK inhibitor
285 VIII.1, PIK-93, and GSK690693, respectively. Similar results were observed for other molecularly
286 targeted drugs, such as anti-EGFR drugs (**Supplementary Fig. S3**). The results indicate that
287 ResGitDR learns to predict drug response in a cell-state-oriented manner instead of relying on
288 the genomic status of the biomarker genes. **Table. 1** shows the important SGAs gene in top 50
289 features in different pathways when predicting the drug response. For instance, in the
290 PI3K/MTOR pathway, PIK3CA and PTEN are identified as important genes. On the other hand, in
291 the ERK MAPK pathway, BRAF is recognized as a significant gene.
292
293 We further investigated the cell-state-oriented nature of ResGitDR from another perspective. If
294 a family of drugs shares a common MOA, it is expected that they will have a similar impact on
295 cells sharing similar cell states. For each drug, we extracted the parameter vectors of the elastic
296 net in ResGitDR model, which reflect the relative importance of features used by the model. We
297 call this representation "drug embeddings", and we performed pairwise cosine similarity
298 analysis of the drug embeddings. For each drug, we identified five drugs with the closest
299 embeddings and visualized the relationships among the drugs (**Fig. 5D**). The results show that
300 drugs targeting a common pathway share similar embeddings, supporting our assumption that
301 ResGitDR identified the features reflecting the cell states indicative of sensitivity to drugs
302 sharing MOAs.
303
304 **Discussion**

In this study, we presented a novel framework for genome-informed precision oncology. Our approach overcomes the limitations of the current rule-based precision oncology [5,8] or simple machine learning approaches of directly using SGAs as inputs to predict drug responses [9]. Instead, we designed the biologically-motivated ResGit model that learns to encode the information of SGAs with respect to gene expression using hierarchically organized latent variables, which mimic the cellular signaling systems of cancer cells. Hence, by transforming genomic data into features reflecting the functional state of cellular signaling systems, the integrated ResGitDR achieved significantly enhanced performance in predicting drug response.

Several novel designs in ResGitDR contribute to its utility. First, ResGit closely mimics the processes by which SGAs perturb cellular signaling systems, eventually leading to cancer. The cellular signaling system consists of hierarchically organized signaling proteins, and genomic perturbation at the different levels of the hierarchy exert distinct effects on cellular systems. By connecting SGAs to all latent variables, ResGit can learn the direct impact of an SGA on the specific components of the signaling system and allow the neural network to transmit such impact through the system. This makes the system transparent and interpretable, enabling ResGit to capture more efficiently the shared functional implications of different SGAs that perturb a common pathway in cells. Second, the self-attention mechanism enables the ResGit to capture the instance-specific impact of SGAs on the cellular signaling system, enabling the model to detect different roles of SGAs in individual tumors. Finally, explicitly including the hidden representation in ResGitDR makes the state of latent variables transparent, which enables ResGitDR to perform drug response prediction in a cell-state-oriented fashion.

Conventional genome-informed precision oncology mainly uses genomic biomarkers to guide the application of molecularly targeted drugs. As pointed out in previous studies [5,9] and our experiments, the accuracy of the rule-based or simple "black box" neural net models for guiding molecularly targeted drugs has room to be improved. Here, we show that by learning a representation of the cell signaling system, ResGitDR significantly outperforms simple models such as elastic networks and feed-forward neural networks. Although the current model has limited clinical utility because it is trained with pre-clinical data and not tested in real-world patient data, we anticipate that our framework has the potential to improve the accuracy of genome-informed targeted therapy in clinical settings if trained with large real-world data. Moreover, our framework can be expanded to guide chemotherapies as demonstrated by our results and other studies [24-26], which will significantly expand the scope of precision oncology beyond the genome-informed application of molecularly targeted drugs.

Materials and methods

Somatic genomic alterations (SGAs) pre-processing

The mutation data of GDSC was downloaded from Iorio *et al.* [9] and the CNV data and cancer type data were downloaded from Cell Model Passports (https://cellmodelpassports.sanger.ac.uk). The mutation data of TCGA were downloaded from the TCGA website (https://portal.gdc.cancer.gov), and the CNV data and cancer type data were downloaded from the Xena portal (http://xena.ucsc.edu). We represent an SGA event in a gene in a tumor as a binary variable, such that genes with mutations or somatic copy number

alteration (deletion or amplification) were given a value of 1 and otherwise were given a value of 0. Since the majority of SGAs observed in tumors are likely passenger events, we take the union of 527 driver genes defined by the Cell Model Passports, 634 genes that are found to causally influence gene expression in cancers identified by Cai et al [27], and 324 mutation genes used in Foundation Medicine (https://www.foundationmedicineasia.com) to obtain the final set of 1,084 SGAs.

Gene expression and TF-target gene matrix pre-processing

To take advantage of existing cancer big data, we combined both TCGA and GDSC RNA-Seq data. The RNAseq data of GDSC was obtained from Garcia-Alonso *et al.* [28] and of TCGA from the Xena portal. We selected the genes using the gene set described in Ding *et al.*[24] with the selection rule that genes with high variances were identified by medium variance analysis, bimodal mixture fitting, and statistical significance of modes. We obtained the processed TF-gene connectivity matrix from Tao *et al* [15]. If a TF is known to regulate a gene, the corresponding element in the connectivity matrix is 1; otherwise, it is 0. The final set contained 320 TFs and 1,613 genes and had 105,224 connections.

Drug sensitivity data pre-processing

Drug sensitivity data were downloaded from the GDSC website (https://www.cancerrxgene.org), and activity area (AA) was used to evaluate drug responses. In the GDSC1 dataset, there are a total of 367 drugs. Within this dataset, there are multiple drugs that share the same name but have different drug IDs. We considered these drugs as distinct

371 entities. To facilitate future application in clinical practice, we discretized the drug response of a
372 cell line with respect to a drug into two categories, sensitive (1) and resistant (0), by applying
373 the waterfall method to each drug which was described in Ding *et al.*[24]. Specifically, the drug
374 sensitivity measurements of all cell lines to a specific drug are sorted to generate a waterfall
375 distribution. A linear regression is fitted to this distribution, and a Pearson correlation
376 determines the goodness of fit. If the correlation coefficient is <0.95, the major inflection point
377 is estimated as the point with maximal distance from a line drawn between the start and end
378 points. If the correlation coefficient is >0.95, the median value is used. This value serves as the
379 cutoff to separate sensitive and resistant cell lines to this drug.
380
381 ***ResGitDR architecture***
382 The overall architecture of ResGitDR is shown in **Fig. 1**. The model has two modules: 1) The
383 Representation Learning Module (ResGit), which is a deep learning model that aims to encode
384 the impact of SGAs on cellular signaling system by performing the task of predicting gene
385 expression using SGAs and cancer type data as input. When trained, the model can be used to
386 infer the state of the cellular signaling system by feeding SGAs and cancer type into the model.
387 2) The Drug Response Prediction Module, which utilizes elastic net to predict drug sensitivity by
388 taking the hidden features learned in the first module and SGAs as input.
389
390 **Representation Learning Module in ResGitDR**
391 The residual genomic impact transformer (ResGit) is similar to the genomic impact transformer
392 (GIT) model developed by Tao *et al.*[29] with several modifications of the architecture and

393 procedures. Compared with GIT model, ResGit has more than one hidden layer and allows the
394 connection of the SGAs to both the first hidden layer and each additional hidden layer (**Fig. 1C**).
395 Through a series of hyperparameter tuning experiments, we set the number of hidden layers in
396 ResGit to 4 (H_1, H_2, H_3, and *TF*) and the number of hidden nodes number in H_1, H_2, H_3, and *TF*
397 layers to 200, 200, 200, and 320, respectively.
398
399 Input to the model consists of the cancer type label and *m* SGAs observed in a tumor. The inputs
400 is firstly converted into embeddings using the "torch.nn.Embedding" class in PyTorch. The
401 cancer type of the sample is transformed into a cancer-type embedding (e_c) through an
402 embedding layer. To capture the diverse impacts of a specific gene *m* on different hidden nodes,
403 four distinct embedding layers are employed to convert the SGA gene *m* into four embedding
404 vectors ($e_m^1, e_m^2, e_m^3, e_m^4$) . Additionally, instead of randomly initializing the SGA embeddings, we
405 applied the Word2Vec[30] algorithm to the SGA data to “pre-train” the SGA embedding.
406 Embeddings learned in this fashion can capture the co-occurrence patterns of SGAs, so that the
407 SGAs affecting a common pathway share a similar embedding. After initializing the SGA
408 embedding with the pre-training gene embedding, the SGA embedding will further update with
409 the supervision of gene expression data in ResGit.
410
411 After obtaining SGAs embedding, we employed a multi-head self-attention mechanism, which
412 could distribute importance weights to SGAs in the training phase. Given a specific sample with
413 cancer type (*C*) and a set of SGAs events (*M*), we obtained the first signal embedding layer (e_1)

by the Equation (1), then applied a Relu activation function to get the first hidden representation (H_1) through Equation (2):

$$e_1 = e_c + \alpha_1^1 * e_1^1 + \alpha_2^1 * e_2^1 + \cdots + \alpha_m^1 * e_m^1 \quad (1)$$

$$H_1 = Relu(e_1) \quad (2)$$

Where $\alpha_1^1, \alpha_2^1, \ldots \alpha_m^1$ are the attention weights for the first hidden layer.

The attention weights in our experiment were calculated using the method described in Tao *et al*[29]. In brief, we calculated the attention weights ($\alpha_1^i, \alpha_2^i, \ldots \alpha_m^i$) for hidden layer H_i by following steps. First, the single-head (h) attention weights were calculated by Equation (3):

$$\alpha_{1,h}^i, \alpha_{2,h}^i, \ldots, \alpha_{m,h}^i = softmax\left((\theta_h^i)^T \tanh\left(W_0^i \cdot e_1^i\right), (\theta_h^i)^T \tanh\left(W_0^i \cdot e_2^i\right), \ldots, (\theta_h^i)^T \tanh\left(W_0^i \cdot e_m^i\right)\right) \quad (3)$$

Where $(\theta_h^i)^T$ is the single-head parameter for head h and W_0^i is the parameter matrix, both of them are for hidden layer H_i. Then we calculated the multi-head attention weights by adding all the single head weights:

$$\alpha_m^i = \alpha_{m,1}^i + \alpha_{m,2}^i + \cdots + \alpha_{m,h}^i \quad (4)$$

To obtain the subsequent signal embedding layer (e_2-e_4), only SGAs were used:

$$e_i = \alpha_1^i * e_1^i + \alpha_2^i * e_2^i + \cdots + \alpha_m^i * e_m^i \quad (5)$$

In order to obtain the second hidden representation layer (H_2), we performed an addition operation to combine the initial hidden representation layer (H_1) with the signal embedding layer (e_2). Subsequently, we applied a ReLu layer:

$$H_2 = Relu(H_1 + e_2) \quad (6)$$

Similarly, we obtained the third hidden representation layer (H_3):

$$H_3 = Relu(H_2 + e_3) \quad (7)$$

To obtain the last hidden layer, the transcription factor layer (TF layer), we used sigmoid function instead:

$$TF = Sigmoid(H_3 + e_4) \tag{8}$$

We used the TF layer to represent the state of transcription factors (TFs) explicitly, and the last linear layer learns the relationships between TFs and their target genes. This learning process was guided by a sparse matrix of prior knowledge derived from a TF-gene connectivity matrix ($P \in {}^{24k \times l}$), where *k* is number of TF and *l* is the number of gene. To predict the gene expression values, the Equation (9) was used:

$$\hat{y}_{exp} = W_{TF-gene} * TF \tag{9}$$

Where $\hat{y}_{exp}$ is the predicted gene expression value, $W_{TF-gene}$ share the same shape with prior matrix P, and $W_{TF-gene,i,j}$ is allowed to be nonzero and updated during learning only when $P_{i,j}$ = 1. The gene expression is a continuous value, and mean square loss was used as the loss function:

$$\sum_{i=1}^{n} \left(y_{exp} - \hat{y}_{exp}\right)^2 \tag{10}$$

Where *n* is the number of samples, and y_{exp} is the observed gene expression value.

To avoid overfitting and increase robustness, we applied the pruning technique on both hidden layers and gene embeddings. For the weights matrix of the first three hidden layers, 90% of low-ranking weights are removed. For the last TF-gene expression weights matrix, we used prior knowledge, the TF-to-gene matrix, to regulate the weights, and the connections in this matrix are about 20%. For the embedding pruning, for each layer, every gene has its own gene

457 embedding (the dimension of the first three layers is 200, and of the last one is 320). We first
458 train the ResGit model without pruning any element in embedding, and after the model
459 converges, we rank the nodes of each embedding, only nodes with the top 60% high value will
460 be kept, and other elements will be changed into zero. Then, we re-train the ResGit module
461 again till it converges. We used 10-fold cross-validation to evaluate the performance.
462
463 **Drug Response Prediction Module in ResGitDR**
464 We used the elastic network model as the classifier for ResGitDR, which is a form of logistic
465 regression with a hybrid regularization term that combines lasso and ridge regularization. We
466 concatenated the original SGAs and latent variables derived by ResGit (H_1, H_2, H_3, TF) of cell
467 lines as the input features for the classifier, and binary drug-sensitivity label as targets. We used
468 class sklearn.linear_model.LogisticRegression with penalty of elastic net. It contains two
469 hyperparameters, L1_ratio and *C*. L1_ratio defines the relative weight of the lasso and ridge
470 penalization terms, and *C* determines the regularization strength. We used grid search to select
471 L1_ratio and *C* for each drug. The elastic net was performed with 10-fold cross-validation. Since
472 ResGitDR model contains two modules, to avoid data leakage, we performed the cross-
473 validation experiment simultaneously, using the same training/testing dataset for ResGit and
474 elastic net.
475
476 To predict drug sensitivity, $SGAs, H_1, H_2, H_3, TF$ were firstly concatenated together, then elastic
477 net was used:
478 $$\hat{y}_{drug} = W_{drug} * (SGAs, H_1, H_2, H_3, TF) \quad (11)$$

Where $\hat{y}_{drug}$ is the predicted drug sensitivity value and W_{drug} is the weight matrix of elastic net.

Since drug response is binarized, the cross-entropy loss was used as loss function:

$$-(y_{drug}\log(\hat{y}_{drug}) + (1 - y_{drug})\log(1 - \hat{y}_{drug}) \quad (12)$$

Where y_{drug} is the observed drug sensitivity value.

Abbreviations

ACC	Adrenocortical carcinoma
ALL	Acute lymphoblastic leukemia
BLCA	Bladder Urothelial Carcinoma
LGG	Brain Lower Grade Glioma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
CLL	Chronic Lymphocytic Leukemia
COAD/ READ	Colon adenocarcinoma/Rectum adenocarcinoma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
ESCA	Esophageal carcinoma
GBM	Glioblastoma multiforme
HNSC	Head and Neck squamous cell carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LAML	Acute Myeloid Leukemia
LCML	Chronic Myelogenous Leukemia
LGG	Brain Lower Grade Glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MB	Medulloblastoma
MESO	Mesothelioma
MM	Multiple Myeloma
NB	Neuroblastoma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate adenocarcinoma

SARC	Sarcoma
SCLC	Small Cell Lung Cancer
SKCM	Skin Cutaneous Melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular Germ Cell Tumors
THYM	Thymoma
THCA	Thyroid carcinoma
UCS	Uterine Carcinosarcoma
UCEC	Uterine Corpus Endometrial Carcinoma
UVM	Uveal Melanoma

Acknowledgments

This study is supported by the NIH grant 5R01LM01201.

Reference

1 Ashley, E. A. Towards precision medicine. *Nature Reviews Genetics* **17**, 507-522, doi:10.1038/nrg.2016.86 (2016).

2 Tsimberidou, A. M., Fountzilas, E., Nikanjam, M. & Kurzrock, R. Review of precision cancer medicine: Evolution of the treatment paradigm. *Cancer Treat Rev* **86**, 102019, doi:10.1016/j.ctrv.2020.102019 (2020).

3 Milbury, C. A. *et al.* Clinical and analytical validation of FoundationOne(R)CDx, a comprehensive genomic profiling assay for solid tumors. *PLoS One* **17**, e0264138, doi:10.1371/journal.pone.0264138 (2022).

4 Malone, E. R., Oliva, M., Sabatini, P. J. B., Stockley, T. L. & Siu, L. L. Molecular profiling for precision cancer therapies. *Genome Med* **12**, 8, doi:10.1186/s13073-019-0703-1 (2020).

5 Marquart, J., Chen, E. Y. & Prasad, V. Estimation of the Percentage of US Patients With Cancer Who Benefit From Genome-Driven Oncology. *JAMA Oncol* **4**, 1093-1098, doi:10.1001/jamaoncol.2018.1660 (2018).

6 Prasad, V. Perspective: The precision-oncology illusion. *Nature* **537**, S63, doi:10.1038/537S63a (2016).

7 Flaherty, K. T. *et al.* Molecular Landscape and Actionable Alterations in a Genomically Guided Cancer Clinical Trial: National Cancer Institute Molecular Analysis for Therapy Choice (NCI-MATCH). *J Clin Oncol* **38**, 3883-3894, doi:10.1200/JCO.19.03010 (2020).

8 Liu, R. *et al.* Systematic pan-cancer analysis of mutation-treatment interactions using large real-world clinicogenomics data. *Nat Med* **28**, 1656-1661, doi:10.1038/s41591-022-01873-5 (2022).

9 Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740-754, doi:10.1016/j.cell.2016.06.017 (2016).

10 Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570-575, doi:10.1038/nature11005 (2012).

11 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607, doi:10.1038/nature11003 (2012).

12 Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res* **22**, 398-406, doi:10.1101/gr.125567.111 (2012).

13 Vandin, F., Upfal, E. & Raphael, B. J. Finding driver pathways in cancer: models and algorithms. *Algorithms Mol Biol* **7**, 23, doi:10.1186/1748-7188-7-23 (2012).

14 Chen, L., Cai, C., Chen, V. & Lu, X. Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. *BMC Bioinformatics* **17 Suppl 1**, 9, doi:10.1186/s12859-015-0852-1 (2016).

15 Tao, Y. *et al.* Interpretable deep learning for chromatin-informed inference of transcriptional programs driven by somatic alterations across cancers. *Nucleic Acids Res* **50**, 10869-10881, doi:10.1093/nar/gkac881 (2022).

16 Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**, 1-22 (2010).

17 Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321-337 e310, doi:10.1016/j.cell.2018.03.035 (2018).

18 Bailey, M. H. *et al.* Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371-385. e318 (2018).

19 Larribere, L. & Utikal, J. Update on GNA Alterations in Cancer: Implications for Uveal Melanoma Treatment. *Cancers (Basel)* **12**, doi:10.3390/cancers12061524 (2020).

20 Takaku, M., Grimm, S. A. & Wade, P. A. GATA3 in Breast Cancer: Tumor Suppressor or Oncogene? *Gene Expr* **16**, 163-168, doi:10.3727/105221615X14399878166113 (2015).

21 Singh, A. *et al.* Dysfunctional KEAP1-NRF2 interaction in non-small-cell lung cancer. *PLoS Med* **3**, e420, doi:10.1371/journal.pmed.0030420 (2006).

22 Yan, X. *et al.* DHX9 inhibits epithelial-mesenchymal transition in human lung adenocarcinoma cells by regulating STAT3. *Am J Transl Res* **11**, 4881-4894 (2019).

23 Olivier, M., Hollstein, M. & Hainaut, P. TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb Perspect Biol* **2**, a001008, doi:10.1101/cshperspect.a001008 (2010).

24 Ding, M. Q., Chen, L., Cooper, G. F., Young, J. D. & Lu, X. Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics. *Mol Cancer Res* **16**, 269-278, doi:10.1158/1541-7786.MCR-17-0378 (2018).

25 Tao, Y., Ren, S., Ding, M. Q., Schwartz, R. & Lu, X. in *Machine Learning for Healthcare Conference.* 660-684 (PMLR).

26 Ren, S. *et al.* De novo Prediction of Cell-Drug Sensitivities Using Deep Learning-based Graph Regularized Matrix Factorization. *Pac Symp Biocomput* **27**, 278-289 (2022).

27 Cai, C. *et al.* Systematic discovery of the functional impact of somatic genome alterations in individual tumors through tumor-specific causal inference. *PLoS Comput Biol* **15**, e1007088, doi:10.1371/journal.pcbi.1007088 (2019).

28 Garcia-Alonso, L. *et al.* Transcription Factor Activities Enhance Markers of Drug Sensitivity in Cancer. *Cancer Res* **78**, 769-780, doi:10.1158/0008-5472.CAN-17-1679 (2018).

29 Tao, Y., Cai, C., Cohen, W. W. & Lu, X. in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020.* 79-90 (World Scientific).

30 Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

Table

Table 1. The top important SGA genes that were included as features when predicting the response to specific signaling pathways drugs.

	Top important SGA genes to predict drug response
PI3K/MTOR signaling	PIK3CA, PTEN, ZFHX4, VPS13B, RELN, USH2A
ERK MAPK signaling	BRAF, TTN
WNT signaling	FLG, MUC16, ZNF208, VCAN, ATM, HRNR, CSMD3, RSPH10B2, APOBEC3B
JNK signaling	CSMD1, ROS1, VPS13B, TET1, FAT3
p53 pathway	TP53, CDH10, SYNE1, TCHH, APC, PTPRC, DMBT1, VCAN
EGFR signaling	KRAS, ERBB2, RELN, HRNR, LRP1B, EGFR
IGF1R signaling	PTEN, RYR2, GLI1, XIRP2, MYH2, MUC16, RYR1, FANCM, CSMD3, FAT1, DNAH14, IKZF3, IL7R

Figures

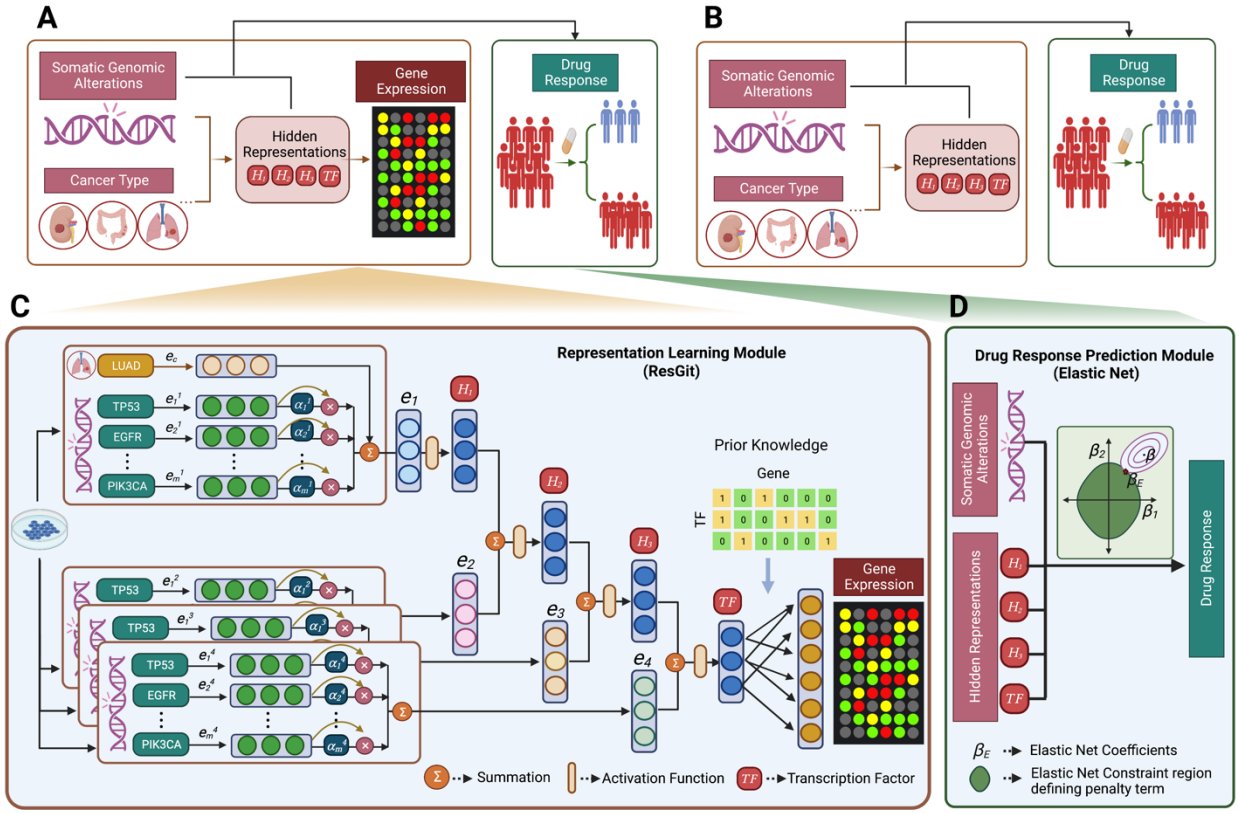


Fig. 1 Flowchart of overall drug sensitivity prediction framework. **(A).** ResGitDR comprises two modules: the Representation Learning Module, which employs the Residual Genome Impact Transformer (ResGit) model, and the Drug Response Prediction Module, which utilizes an elastic net. In the training phase, the Representation Learning Module uses SGAs and cancer types to predict gene expression, and the Drug Response Prediction Module incorporates the hidden representations learned in the Representation Learning Module and SGAs as input to predict drug sensitivity. **(B).** In the testing phase, the trained ResGit model is used to obtain hidden representations using SGAs and cancer type as input. These hidden representations are then combined with SGAs as inputs to predict drug response. **(C).** The detailed diagram of the Representation Learning Module. **(D).** The detailed diagram of the Drug Response Prediction Module.

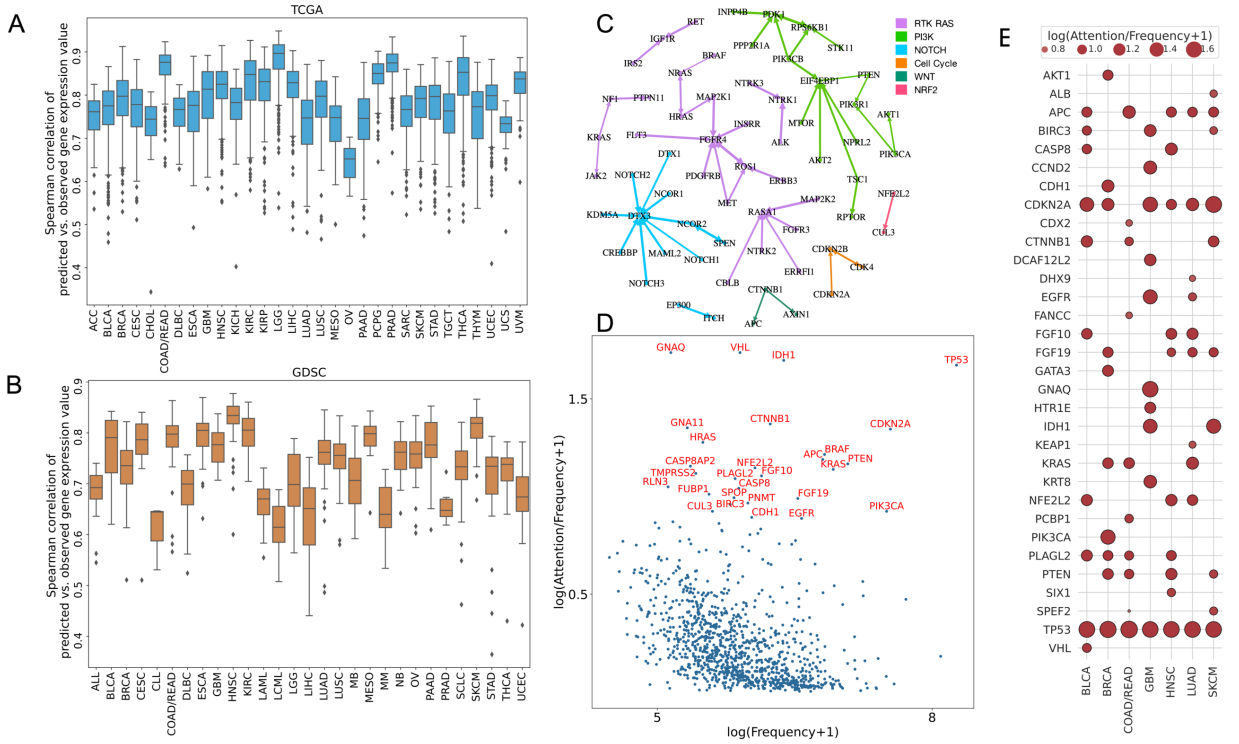


Fig. 2 Evaluation of the performance of ResGit. The distribution of Spearman correlation coefficients between predicted and observed gene expression values **(A)** in the TCGA dataset and **(B)** in the GDSC datasets, respectively. (**C**). The connectivity map shows the similarity of SGA embeddings among the SGAs perturbing common pathways. The weight vector connecting an SGA to hidden nodes is used as an embedding of the SGA, and similarity between a pair of SGAs is calculated with cosine similarity. If gene *A* is a neighbor of gene *B*, the arrow direction points from gene *B* to gene *A*; a double-arrowed edge indicates that two SGAs are mutually among the top 10 neighbors. The thickness of an arrow represents the degree of similarity. **(D).** The attention weights of SGAs gene in a pan-cancer analysis. Genes with high overall attention weights are shown in red font. **(E).** The attention weights of SGAs gene across different cancer types.

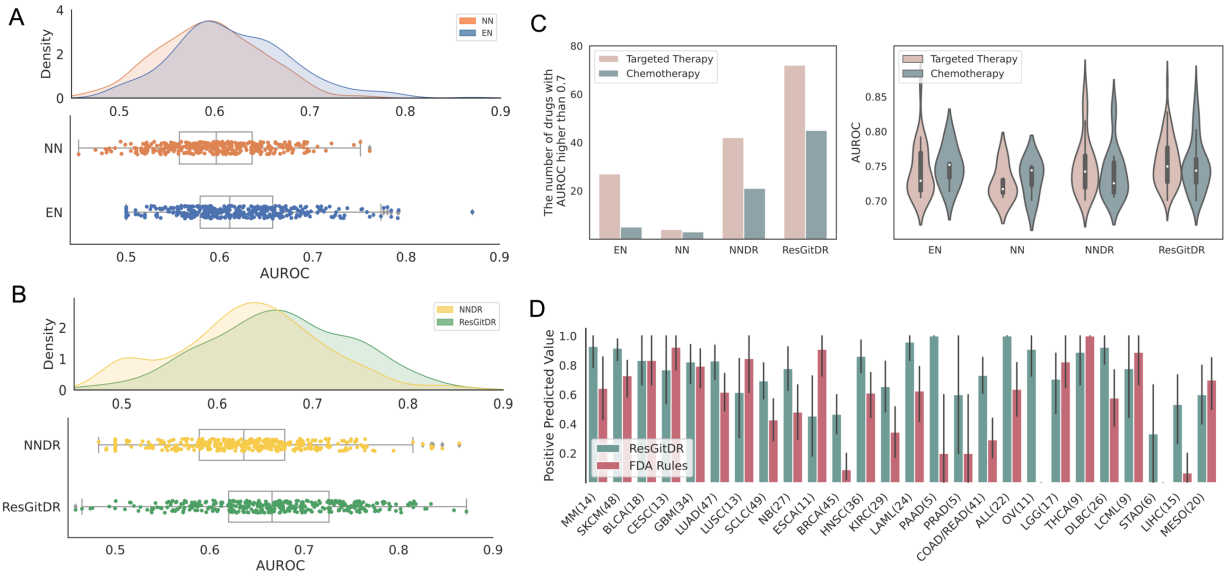


Fig. 3 The performance comparison in drug response prediction. **(A).** The performance of two baseline models (EN and NN) which both use SGA and cancer type to predict drug sensitivity directly. **(B).** The performance of two models (NNDR and ResGitDR), which both firstly use the SGAs and cancer type to predict gene expression and obtain the hidden representations, then concatenate SGA and hidden representations to predict drug sensitivity. **(C).** The number and AUROC distribution of Targeted Therapy and Chemotherapy drugs with AUROC higher than 0.7 across EN, NN, NNDR and ResGitDR. **(D).** The Positive Predicted Value of ResGitDR and FDA rules methods with error bar representing 95% confidence interval. The numbers in parentheses indicate the corresponding cell line counts for each cancer type.

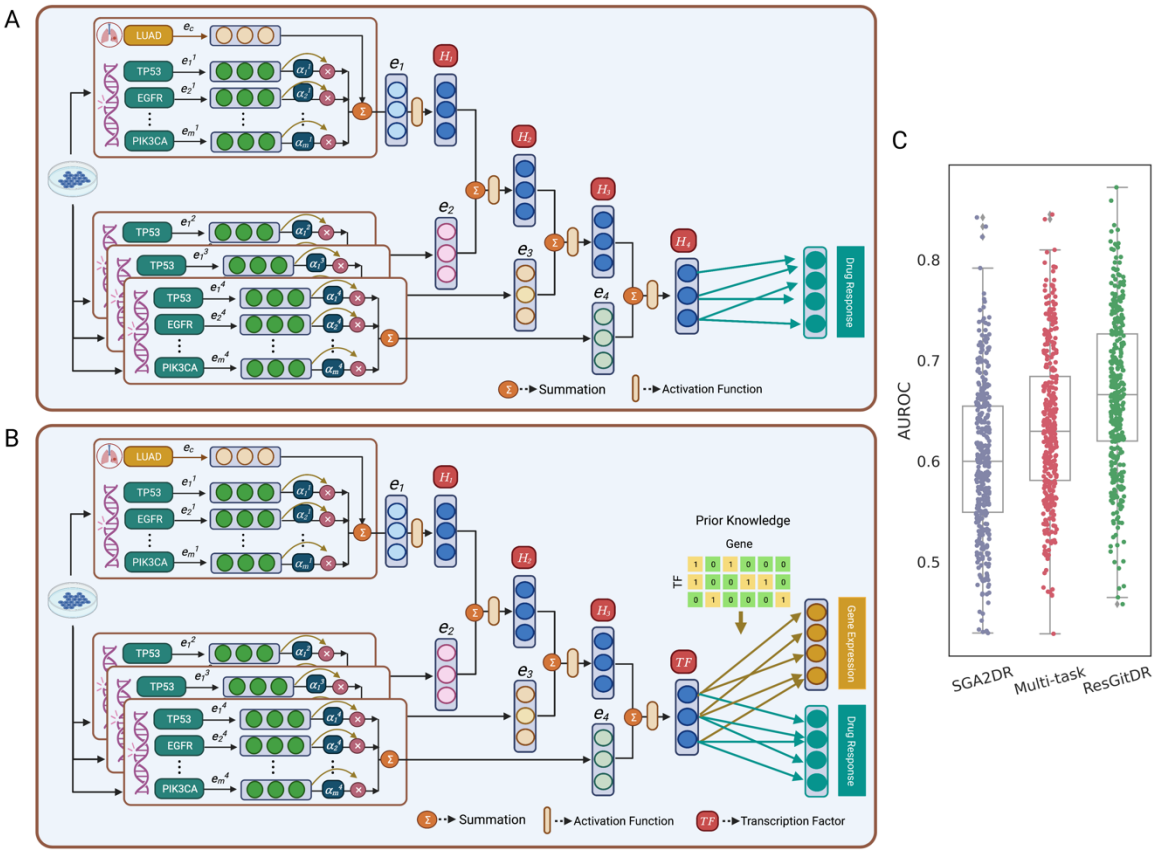


Fig. 4 (A). The architecture of the SGA2DR model. It predicts drug sensitivity directly using the same architecture of ResGit by taking the cancer type and SGAs as input. **(B).** The architecture of the multi-task learning model. It aims to predict drug sensitivity and gene expression simultaneously using the same architecture of ResGit by taking the cancer type and SGAs as inputs. **(C).** The performance comparison of SGA2DR, multi-task learning, and ResGitDR models.

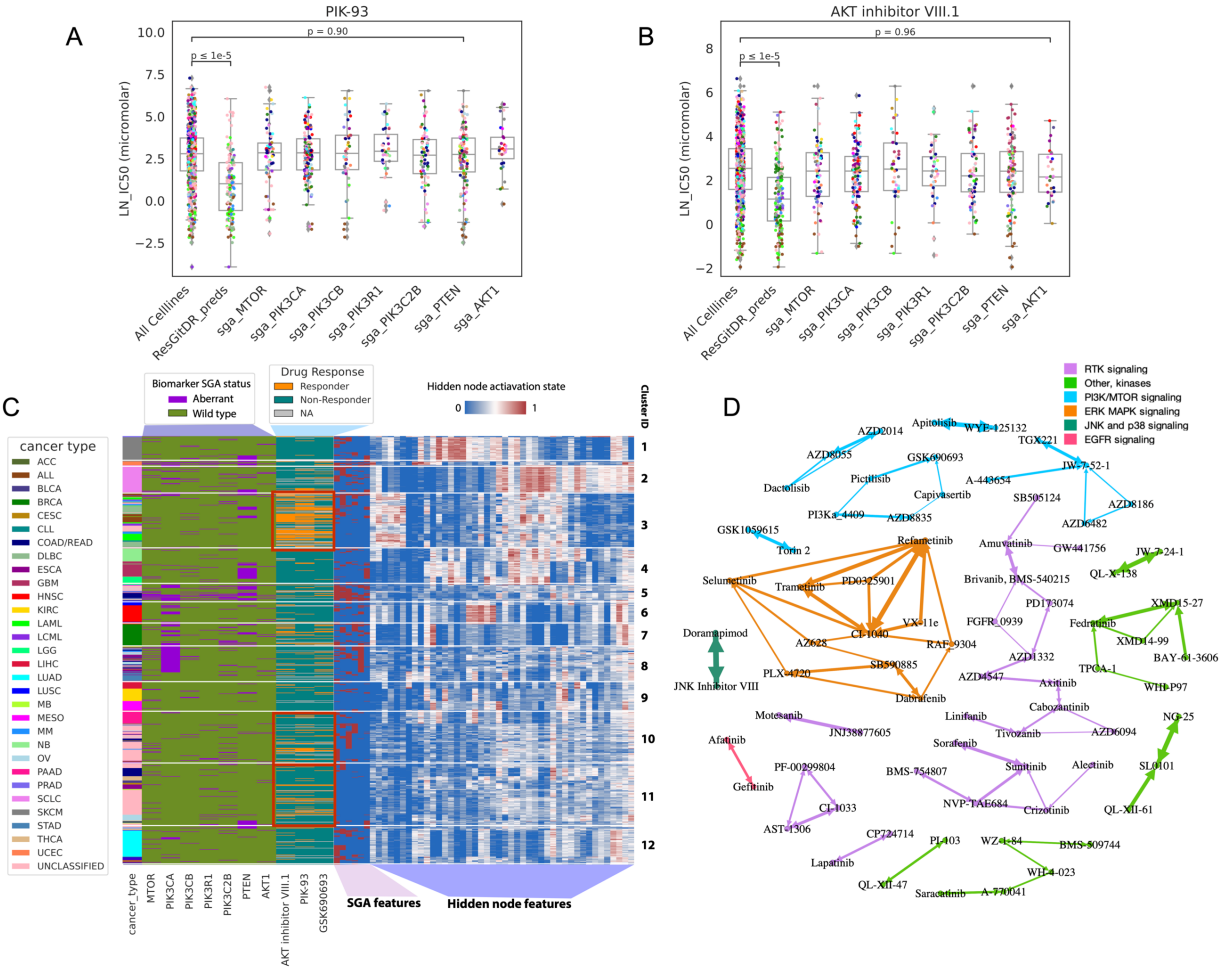
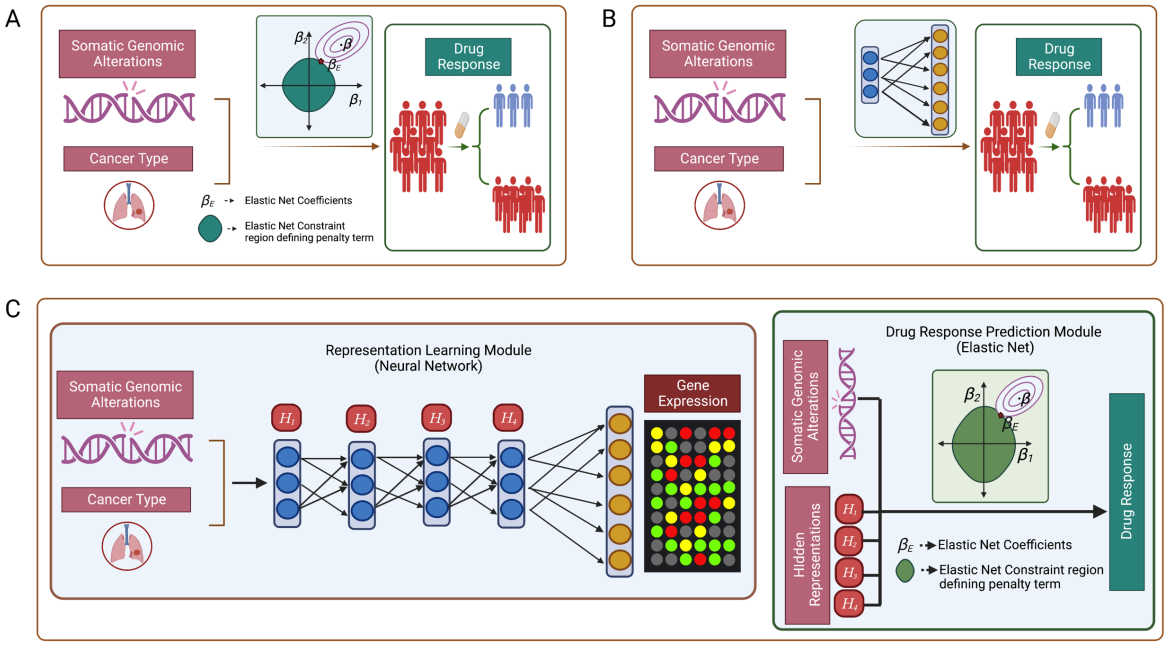
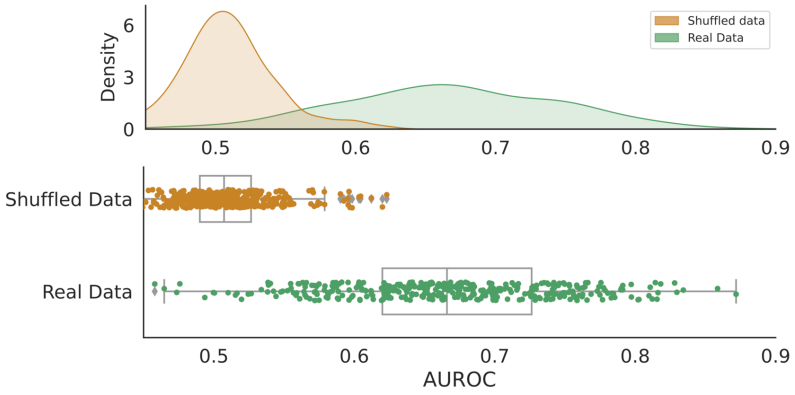


Fig. 5 The distributions of drug sensitivity (represented as log IC50s) to **(A)** *PIK-93* and **(B)** *AKT inhibitor VIII* by cancer cell lines grouped according to the mutation status of genes involved in the PI3K pathway. The distribution of drug sensitivity by the cell lines predicted by ResGitDR to be sensitive to the drugs is also shown. **(C).** Cancer cell lines were clustered using the based on the selected top 50 predictive features from ResGitDR models for 3 anti-PI3K PI3K/MTOR drugs: *AKT inhibitor VIII*, *PIK-93*, and *GSK690693*. The features consist of hidden representations and individual SGAs. The SGAs are represented as binary values. The hidden node values are standardized within the range of 0 to 1. The binary drug responses to each of the three drugs by cell lines are shown. Three red boxes highlight the clusters with enriched responder cell lines (clusters 3, 10, and 11). The mutation status of genes in the PI3K/mTOR signaling pathway is shown to illustrate their relationship with respect to drug sensitivity. **(D).** The connectivity map shows the similarity of the embedding of drugs targeting common pathways. The top 50 important features of the ResGitDR for a drug are used as its embedding. The similarity of embeddings of two drugs is measured with cosine similarity. Molecularly targeted drugs are shown as nodes; an edge is added between a pair of drugs whose embeddings are among the top 5 highest cosine similarities of each other. If drug A is a neighbor of drug B, the arrow direction points from drug B to drug A; a double-arrowed edge indicates that a pair of drugs are mutually among the top 5 neighbors of each other. The thickness of an arrow is proportional to cosine similarity.

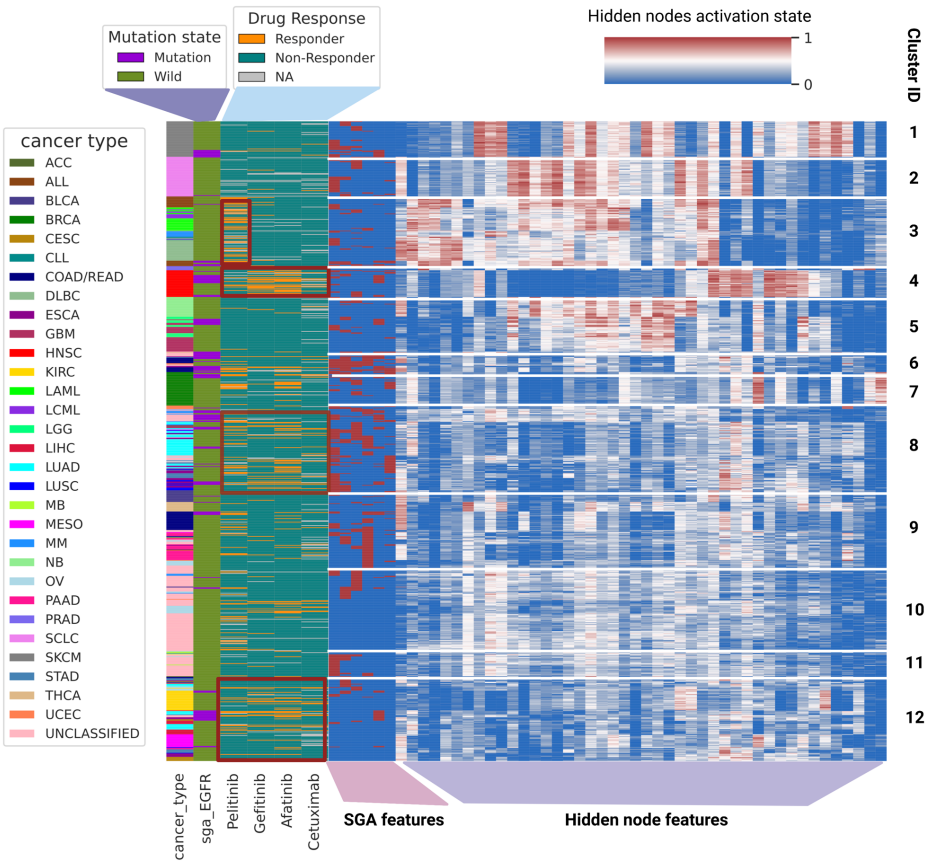
Supplementary Figures



Supplementary Fig. S1 The model architectures of **(A)** the elastic net (EN) and **(B)** neural network (NN) models. Both models take SGAs and cancer type as inputs to directly predict drug response. **(C).** The architecture of the NNDR model involves a four-layer neural network (NN) that predicts gene expression using cancer type and SGAs as input. In the drug prediction phase, the NN is used to infer the state of hidden nodes, which are further used as inputs for the drug response prediction model.



Supplementary Fig. S2 Using ResGitDR to predict drug sensitivity with shuffled data and Real Data



Supplementary Fig. S3 Cell-state-oriented prediction of sensitivity to anti-EGFR drugs. Annotations are the same as Fig. 5 in the main text.

Supplementary Table

Supplementary Table S1. The targeted therapy drugs and chemotherapy drugs with AUROC higher than 0.7 when using ResGitDR, NNDR and EN to predict drug response (please see the excel file).