Cancer Mutations Converge on a Collection of Protein Assemblies to Predict Resistance to Replication Stress

Xiaoyu Zhao[1], Akshat Singhal[2], Sungjoon Park[1], JungHo Kong[1,3], Robin Bachelder[1], and Trey Ideker[1,2,3,4]

ABSTRACT   Rapid proliferation is a hallmark of cancer associated with sensitivity to therapeutics that cause DNA replication stress (RS). Many tumors exhibit drug resistance, however, via molecular pathways that are incompletely understood. Here, we develop an ensemble of predictive models that elucidate how cancer mutations impact the response to common RS-inducing (RSi) agents. The models implement recent advances in deep learning to facilitate multidrug prediction and mechanistic interpretation. Initial studies in tumor cells identify 41 molecular assemblies that integrate alterations in hundreds of genes for accurate drug response prediction. These cover roles in transcription, repair, cell-cycle checkpoints, and growth signaling, of which 30 are shown by loss-of-function genetic screens to regulate drug sensitivity or replication restart. The model translates to cisplatin-treated cervical cancer patients, highlighting an RTK–JAK–STAT assembly governing resistance. This study defines a compendium of mechanisms by which mutations affect therapeutic responses, with implications for precision medicine.

SIGNIFICANCE: Zhao and colleagues use recent advances in machine learning to study the effects of tumor mutations on the response to common therapeutics that cause RS. The resulting predictive models integrate numerous genetic alterations distributed across a constellation of molecular assemblies, facilitating a quantitative and interpretable assessment of drug response.

INTRODUCTION

DNA replication is central to cell division and proliferation, involving closely orchestrated functions among hundreds of proteins (1, 2). Although the replication machinery is highly accurate, it faces challenges from both extrinsic and intrinsic factors (3). These challenges can result in stalled replication forks, occurrence of DNA breaks, reduced replication precision, and other factors collectively known as RS (4). Accordingly, cells have evolved a robust RS response that activates DNA damage repair signaling or, alternatively, induces cell death, to maintain genome integrity within the cell population (5–9). Due to sustained proliferative signaling and/or defective DNA repair, cancer cells undergo persistent replication stress (10, 11), making them strongly RS response dependent. A consequence of this dependency is that replication stress becomes an exploitable therapeutic vulnerability in cancer treatment (12, 13).

A multitude of cancer therapeutics leverage replication stress to eliminate cancer cells, using a diversity of RSi mechanisms (Supplementary Fig. S1). Classic chemotherapeutic agents induce RS by directly affecting DNA integrity. Examples include DNA crosslinkers and alkylating agents (e.g., cisplatin and methotrexate), topoisomerase I and II inhibitors (e.g., camptothecin, etoposide, and doxorubicin), ribonucleotide reductase inhibitors (gemcitabine and hydroxyurea), and nucleotide analogues (5-fluorouracil; refs. 12, 14, 15). Other agents elevate RS by targeting DNA polymerases or DNA damage response proteins. Some of these targeted agents have progressed to advanced stages of preclinical and clinical development, including inhibitors of the DNA polymerase (e.g., CD437), the base-excision repair factors PARP1 and PARP2 (e.g., olaparib), ATR–CHK1 signaling (e.g., ceralasertib), or the WEE1 checkpoint kinase (e.g., MK-1775; refs. 13, 16). Treatment outcomes vary widely due to significant differences in drug sensitivity and resistance across tumors, motivating efforts to better understand response mechanisms. These efforts have led to an expanding catalog of genetic alterations associated with sensitivity or resistance to RSi drugs, including mutations in *BRCA1*, *BRCA2*, *ATM*, or *ATR* (Fig. 1A; refs. 17–26). This list is almost certainly not exhaustive, because the vast majority of genetic alterations identified are rare rather than common in tumors. Moreover, it is unclear how these multiple single-gene effects integrate to constitute an overall drug response.

Currently, two machine learning (ML) methodologies are emerging as particularly useful in understanding the effects of genetic alterations in precision medicine applications. The first methodology, known as "interpretable" ML, includes a wide array of approaches that attempt to gain insight into the mechanisms or rationale underlying a model's predictions (27). In this regard, we and others have developed a series of "visible" neural network (VNN) models that guide the internal architecture of the model using knowledge maps of biological components and functions (28–34). In contrast to conventional "black-box" neural networks (35), a prime advantage of VNNs is that they not only learn to predict biomedical outcomes from sparse genetic feature sets, but their predictions can be mapped to internal state changes in molecular

[1]Division of Human Genomics and Precision Medicine, Department of Medicine, University of California, San Diego, La Jolla, California. [2]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California. [3]Moores Cancer Center, School of Medicine, University of California, San Diego, La Jolla, California. [4]Department of Bioengineering, University of California, San Diego, La Jolla, California.

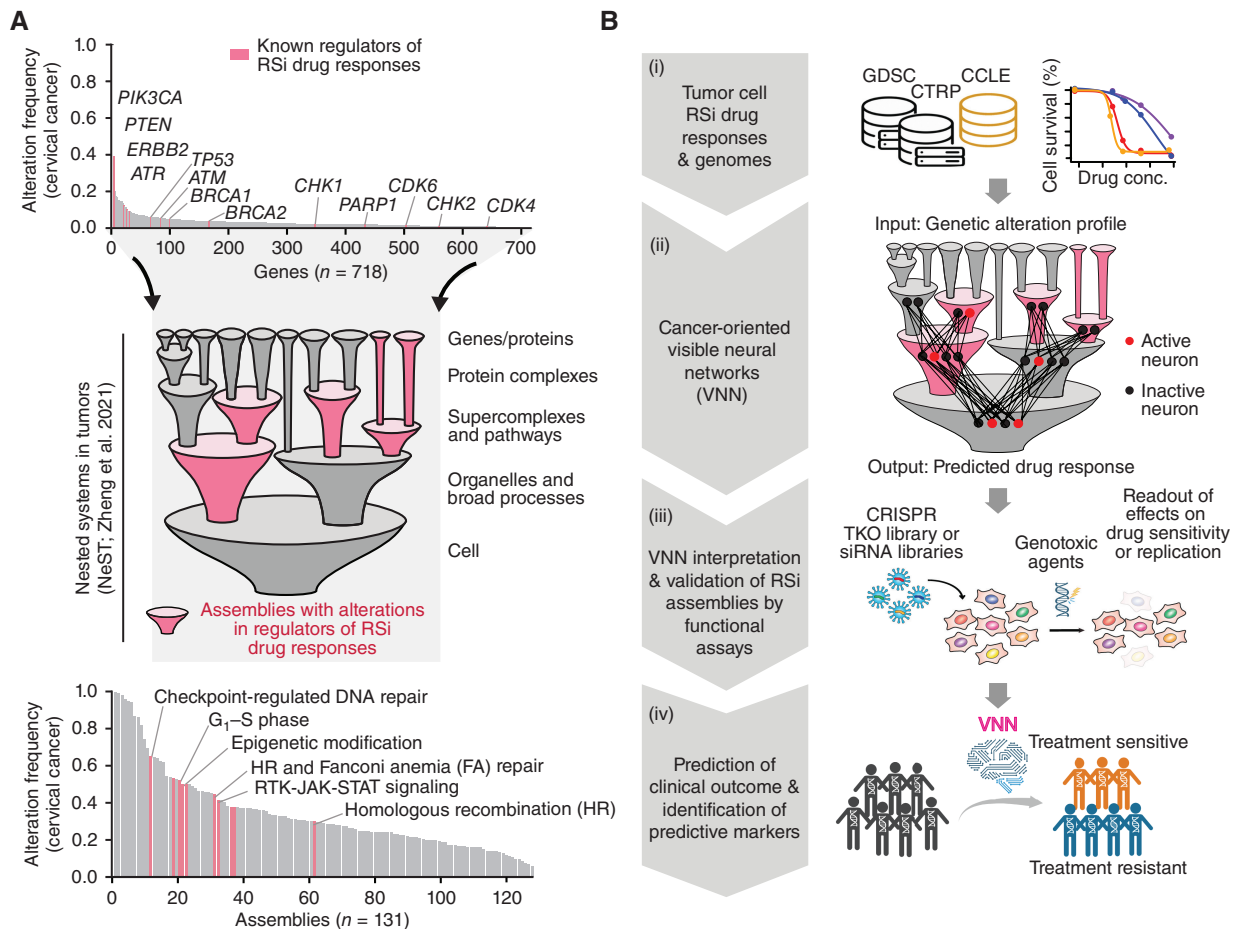X. Zhao and A. Singhal contributed equally to this article.

Corresponding Author: Trey Ideker, UCSD, 9500 Gilman Drive, La Jolla, CA 92093-0688. E-mail: tideker@ucsd.edu

Figure 1. Study overview. **A,** Genetic alterations that modulate replication stress–inducing (RSi) drug responses are typically rare but converge on multiscale protein assemblies in cancer. Top, waterfall plot of gene alteration frequencies in TCGA cervical squamous carcinoma cohort, showing that most genetic alterations are rare. Alterations include somatic mutations and copy-number aberrations of genes covered by clinical cancer gene panels. Middle, genetic alterations converge on specific multiprotein assemblies in the NeST hierarchy. Bottom, waterfall plot showing genetic alteration frequencies of NeST protein assemblies, which, by nature, tend to be higher than for individual genes. **B,** Pipeline for drug response prediction. (i) Acquisition of data sets used in this study. RSi drug response data from GDSC and CTRP databases. Tumor cell line genome sequencing data from CCLE. (ii) Construction of VNN models with architecture guided by NeST. (iii) Interpretation of VNN models to recognize assemblies for which genetic alterations are most important for RSi drug responses. Validation by genome-wide loss-of-function screens. (iv) Translation of VNNs to predict chemotherapy responses in cancer patients.

mechanisms and pathways, allowing for model interpretation. To initiate a proteomics-driven resource of molecular mechanisms in cancer, we recently used affinity purification mass spectrometry to delineate the physical interactions of a broad set of cancer proteins. We integrated these interactions with other "omics" data sets to create a large cancer protein–protein association network, which was iteratively clustered to generate a map of frequently altered protein assemblies (NeST, Nested Systems in Tumors; Supplementary Fig. S2A and S2B; ref. 36). This map exhibits a hierarchical structure, whereby individual cancer proteins converge to form protein complexes and larger assemblies that, in turn, incorporate into broad processes and organelles (Fig. 1A; Supplementary Fig. S2B). The second ML methodology is multitask learning, whereby several learning tasks are solved at the same time while exploiting commonalities and differences across these tasks (37). Multitask learning has been previously used to

boost the efficacy of drug response and mechanism of action prediction (38, 39), although it has not yet been combined with knowledge bases of cancer pathways like NeST or (to our knowledge) applied to understand the response to RSi drugs.

Motivated by these challenges and advancements, here we describe and evaluate interpretable deep learning models of RSi drug response (Fig. 1B). Starting from the genetic alterations detected in a tumor sample, the models predict the sensitivity or resistance to specific RSi drugs. The architecture of these models is based on the knowledge of cancer protein complexes in NeST, meaning their predictions can be interpreted mechanistically, and both single-drug and multitask (multidrug) models are evaluated. These models identify a constellation of protein assemblies that integrate genetic alterations to predict drug sensitivity or resistance, which we show can be validated by multiple types of genome-wide functional assays and by their utility in predicting responses
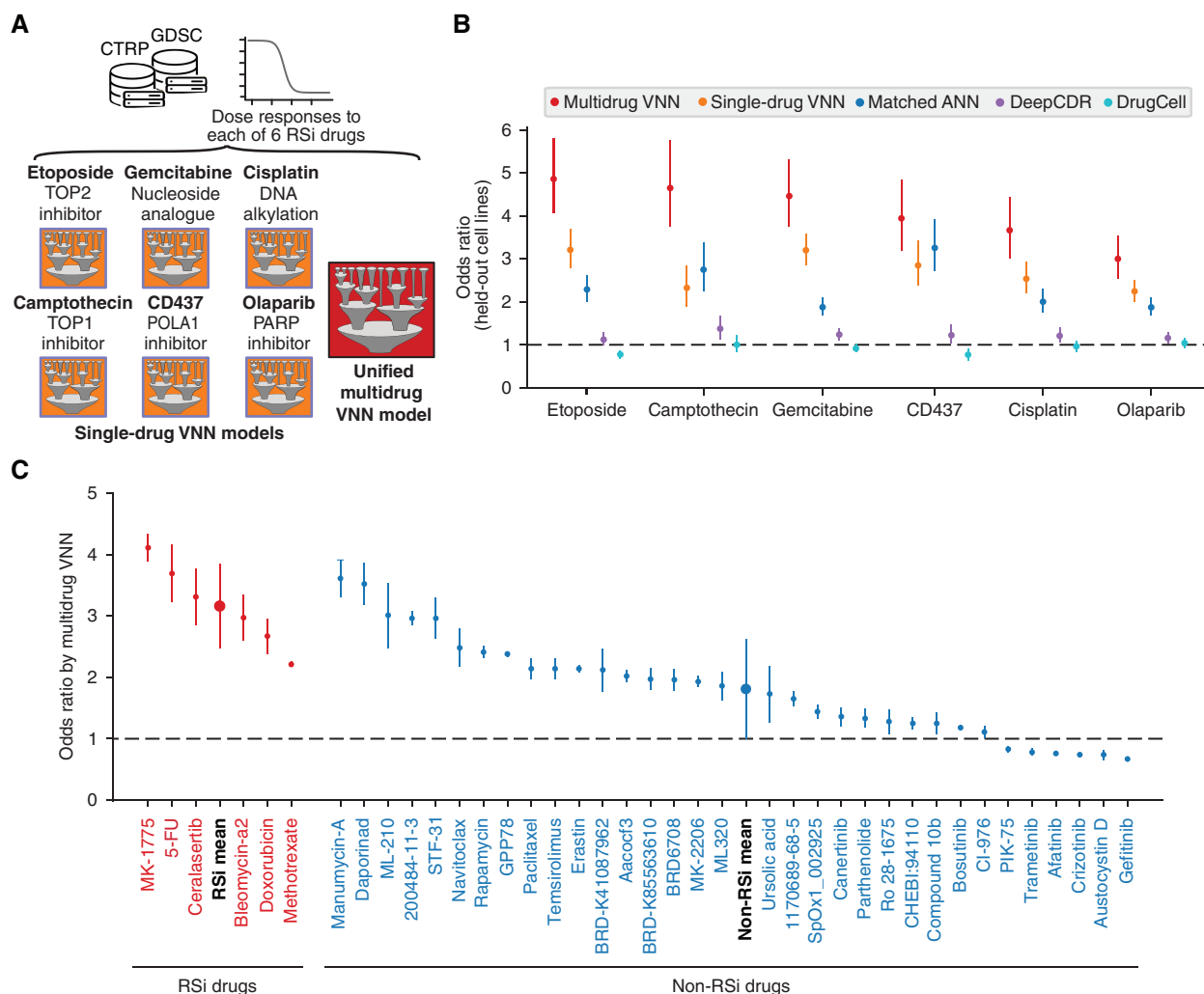
Figure 2. Performance of RSi drug response models. **A,** Schematic representation of the VNN models implemented in this study for single-drug and multidrug scenarios. **B,** Accuracy (odds ratio) of drug response prediction for each of the initial six RSi agents. Multidrug and single-drug VNNs (red, orange) are compared against black-box artificial neural networks with a matched number of parameters and layers (ANN, blue), DeepCDR (purple), and DrugCell (cyan). Error bars show the confidence interval of the odds ratio for each model. Performance was assessed on drug responses of held-out cell lines not seen during training. **C,** Accuracy (odds ratio) of drug response prediction by the multidrug model, evaluated for additional RSi drugs (n = 6, red) and non-RSi drugs (n = 33, blue), none of which were seen during training. Error bars show standard deviation of the odds ratio for each drug. Performance was assessed using held-out cell lines. Statistical comparisons were conducted with the Mann–Whitney U test.

to previously unseen RSi agents as well as in chemotherapy-treated patients (Fig. 1B).

RESULTS

A Cancer-Oriented Interpretable Neural Network

To model responses to RSi drugs, we focused on the set of 718 genes assessed by current clinical cancer gene panels, including one or more of FoundationOne CDx (40), Tempus xT (41), and Project GENIE (42). The genomic alteration status of these genes in a tumor sample, including the presence/absence of mutation and copy-number aberration, was used as input (Supplementary Fig. S3A). Models were trained using drug response data for genomically characterized tumor cell lines harmonized from the Cancer Therapeutics Response Portal (CTRP; refs. 43, 44) and the Genomics of

Drug Sensitivity in Cancer (GDSC; 45, 46) databases. These databases included the measured responses to many RSi drugs targeting DNA replication or DNA damage response. In another recent study, Olivieri and colleagues used genome-wide CRISPR chemogenetic screens to investigate DNA damage response against genotoxic agents (47). Therefore, we initially focused on six RSi agents characterized by Olivieri (cisplatin, gemcitabine, camptothecin, etoposide, olaparib, and CD437; Fig. 2A), which had also been examined in CTRP or GDSC.

Instead of associating genetic alterations with drug responses via classic "black-box" ML, we implemented VNN in which the layers of artificial neurons are designed to propagate the effects of individual gene mutations over the NeST map of cancer protein assemblies (36). In constructing the VNN, each protein assembly encoded by genes on the clinical

panel (131 assemblies; Supplementary Table S1) was assigned a bank of artificial neurons to represent the *in silico* activity of that system (Supplementary Fig. S3A; Methods). The use of multiple neurons allowed protein assemblies to be multi-functional, with the ability to adopt a range of values along several dimensions. For the neurons assigned to an assembly, weighted input connections were permitted from the neurons assigned to each of its subassemblies in the preceding layers of the NeST hierarchy. This design enabled genetic information to flow from the input gene alterations to small protein complexes (e.g., "CDK holoenzyme complex"), then to larger-scale assemblies (e.g., "checkpoint-regulated DNA repair"), and finally to the root assembly, representing the whole cell (Supplementary Fig. S3B). Connection weights were learned during training so that the integrated activity of the root neurons represented the predicted drug response of a tumor sample given its genetic alteration profile. In the multidrug configuration, this architecture was extended so that the activity of the root was provided as input to an additional (final) neuronal processing layer that was optimized separately for each drug response (Supplementary Fig. S3C; Methods).

Training and Performance Analysis of the Models

We first trained single-drug VNN models for predicting the response to each of the six RSi agents (Fig. 2A). Training was conducted by minimizing the mean squared error (MSE) between the predicted and observed drug responses using standard back-propagation techniques (Methods). The accuracy of drug response prediction was assessed for each model using nested cross-validation, in which 64% of cell lines were randomly selected for model training, 16% for model validation and hyperparameter tuning, and 20% as held-out cell lines not yet seen during training or validation (this entire procedure was repeated over five folds; Methods). This assessment produced predictive odds ratios (OR) in the range of 2.2 to 3.2 across the six RSi drug models (Fig. 2B, orange points). These models generally demonstrated performance that was comparable to, or better than, matched black-box neural networks (Fig. 2B, blue points; ORs of 1.8 to 3.3; $P = 0.04$). Their performance was significantly better than that of two general pan-drug models (DrugCell, DeepCDR) that had not been specifically developed for RSi drug prediction (refs. 30, 48; Fig. 2B, cyan and purple points; with P values of 2.5×10^{-8} and 1.2×10^{-10}, respectively).

We also used multitask learning to train a unified multi-drug VNN with six outputs, predicting the response to each of the six RSi drugs (Supplementary Fig. S3C; Methods). This unified model yielded ORs of 3.0 to 4.9 (Fig. 2B, red points), significantly outperforming both the single-drug models ($P = 4.3 \times 10^{-5}$) and the matched black-box models ($P = 2.1 \times 10^{-7}$). To further test the generality of this multi-drug model, we evaluated its accuracy in predicting cell line responses to additional RSi agents not yet seen in this study (predictions based on the average of model outputs; Methods). The new agents included RS-inducing chemotherapies such as doxorubicin, 5-fluorouracil, methotrexate, and bleomycin-a2 as well as drugs targeting DNA damage response factors such as ceralasertib (ATR inhibitor) and MK-1775 (WEE1 inhibitor). The predictive ORs were significant for all new RSi drugs (mean OR 3.2; Fig. 2C, red points; Supplementary Table S2), and they were significantly higher than ORs obtained when predicting responses to a broad panel of 33 agents not associated with replication stress (mean OR 1.8, $P = 3.1 \times 10^{-4}$; Fig. 2C, blue points). Furthermore, the performance of the multidrug model decreased considerably when, during training, some of the RSi drug responses were substituted with unrelated drug responses (Supplementary Fig. S4A). These results were consistent with the expectation that RSi drugs share common response pathways, a finding that was also reflected by varying degrees of similarity in their drug response profiles across tumor cell lines (Supplementary Fig. S4B).
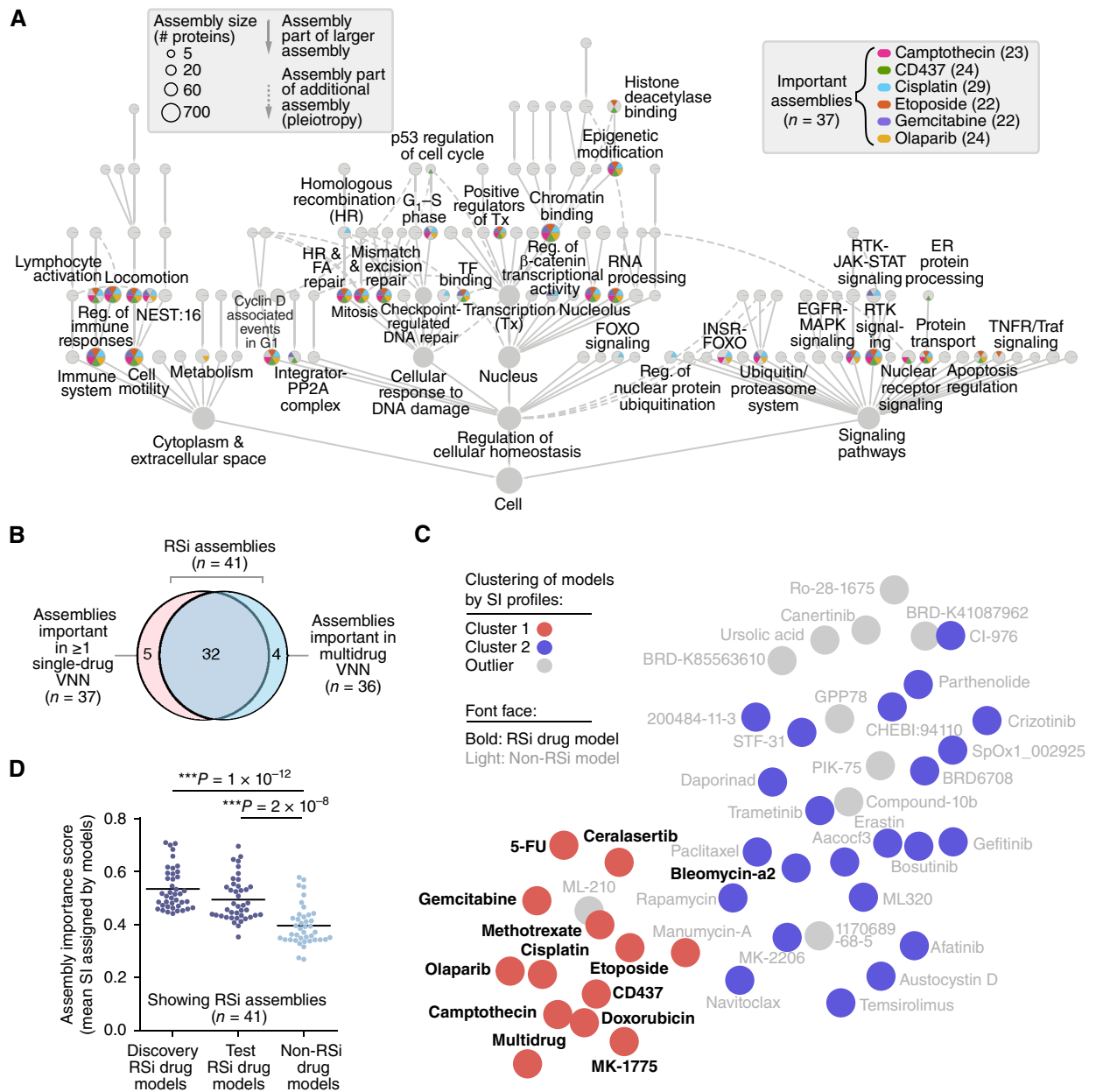
Molecular Assemblies Important for RSi Drug Response Prediction

We next scored all protein assemblies to rank their importance in the prediction of RSi drug responses (Supplementary Table S3). For this purpose, we computed a system importance (SI) score, which measures the dependence of the drug response prediction on genetic alterations within an assembly, captured by changes in the *in silico* activity of that assembly's neurons (Methods). We created an importance profile for each RSi agent using the single-drug models and then mapped these to the NeST protein assembly hierarchy (Fig. 3A); SI profiles were also computed for the multidrug model, with qualitatively consistent results (Supplementary Fig. S4C; Supplementary Table S3).

Due to the progressive integration of genetic information, we noted that importance scores for all models tended to increase with assembly size and depth in the hierarchy. To thus highlight distinct protein assemblies important to RSi drugs, we focused our examination on small-to-medium-scale assemblies (124 assemblies with fewer than 100 genes), identifying 41 that were high scoring in the single-drug or multidrug RSi models (Fig. 3A and B; Supplementary Fig. S4C; Methods). These "RSi assemblies" recovered proteins associated with known drug indications or mechanisms of action where applicable (camptothecin: TOP1; etoposide: TOP2A; olaparib: BRCA1, BRCA2, PARP1; the CD437 target POLA1 was absent from clinical gene panels). For example, in predicting etoposide response, we identified "Checkpoint-regulated DNA repair" and "G_1–S phase" assemblies, which contained the etoposide target topoisomerase 2α (TOP2A; Supplementary Table S3). Turning from positive to negative controls, we found that the various RSi drug models yielded assembly importance scores that were distinct from those of non-RSi drugs but similar to one another (Fig. 3C). Moreover, these RSi assemblies were specifically important for the prediction of the cellular response to RSi drugs, as opposed to Non-RSi drugs (Fig. 3D).

Validation of Specific Protein Assemblies Using Genome-wide Functional Assays

For the assemblies identified as important to the RSi drug models, we investigated whether engineered genetic perturbations impacting each of these assemblies were able to predictably affect the drug response. For this purpose, we turned to the recent work of Olivieri and colleagues (47), who had conducted genome-wide drug-sensitivity screens in the presence

Figure 3. Identification of molecular mechanisms important for RSi drug responses. **A,** System importance (SI) profiles for prediction of initial six RSi drug responses. Nodes indicate protein assemblies; node sizes denote assembly sizes in numbers of proteins; colors mark assemblies of above-threshold importance (Methods) for each drug response. The number of important assemblies is indicated in parentheses for each single-drug model, with a total number of 37 assemblies important for at least one model. **B,** Venn diagram showing numbers of important assemblies identified by single-drug ($n = 37$, left) vs. multidrug VNNs ($n = 36$, right). The union set is defined as RSi assemblies ($n = 41$). **C,** UMAP projection of SI profiles of each drug model constructed in this study (points). The HDBSCAN algorithm (hierarchical density-based spatial clustering of applications with noise) is used to group drug models into clusters (colors) by similarity of important assemblies (SI profiles); gray models fall outside of stable clusters. Bold font face indicates RSi drug models, light font face indicates non-RSi drug models. **D,** RSi assemblies (defined in **B**) are plotted by their average importance scores for models trained on the initial RSi discovery agents (left), the later RSi test agents (middle), or non-RSi drug models (right). Horizontal line in each swarm represents the mean. ***, $P < 0.01$; P values by one-sided Mann–Whitney U test.

of each of 27 genotoxic agents, including the 6 RSi drugs used to train the models in our study (Supplementary Fig. S5A). For each drug, genes were scored by the degree to which loss of function by CRISPR/Cas9 modulates drug sensitivity in either direction (i.e., absolute z-score, capturing both increases and decreases in sensitivity). Using these scores,

we found that 24 of the 41 RSi assemblies were significantly enriched for genes affecting sensitivity to one or more RSi drugs, a rate higher than for non-RSi assemblies (59% vs. 45%, $P < 0.001$; Supplementary Fig. S5B). Such enrichment was particularly strong in the drug-sensitivity screens for etoposide, camptothecin, and CD437 (Supplementary Fig. S5C).

We then explored a second genome-wide loss-of-function screen, directly measuring immunofluorescent readouts of DNA damage and replication restart after an RS challenge (49). This screen measured the effect of siRNA gene knockdowns on the phosphorylation of histone variant H2AX (γH2AX), a marker of stalled replication forks, and on incorporation of EdU in DNA, a marker of active genome replication (Fig. 4A). The ratio of these two readouts, the "replication restart score" (RRS), was used to score all gene knockdowns for their importance in RS response. Genes that affected replication restart in either direction (absolute z-score) were enriched in 21 of the 41 RSi assemblies, a rate significantly higher than for non-RSi assemblies (51% vs. 30%, $P < 0.001$; Fig. 4B). In addition, we found that the converse enrichment relationship was also true, in that NeST assemblies enriched in genes affecting drug sensitivity or replication restart were significantly more important to RSi drug models than to non-RSi models (Fig. 4C and D; Supplementary Figs. S5D–S5J; with P values of 1×10^{-9} and 3×10^{-6}, respectively).

Summarizing the functional screening results, the 41 RSi assemblies included 30 with support from either the drug sensitivity or replication restart readouts (Fig. 4E). Among these, 15 assemblies had support from both, including assemblies involved in mismatch and excision repair, homologous recombination, Fanconi anemia repair, checkpoint-regulated DNA repair, locomotion, ubiquitin/proteasomal assemblies, and the RTK–JAK–STAT pathway (Fig. 4F; Supplementary Table S4). The set of 30 assemblies and the subset of 15 assemblies were equally important for predicting RSi drug response, surpassing non-RSi drug response (Supplementary Fig. S5K).

Prediction of Clinical Responses to Cisplatin

Among adult solid tumors cataloged by The Cancer Genome Atlas (TCGA), five cancer subtypes commonly receive cisplatin therapy, with substantial numbers of associated samples (>30 cisplatin-treated subjects per subtype; Fig. 5A). Among these subtypes, we noted that patients with cervical and lung squamous carcinomas (CESC and LUSC) currently show cisplatin benefit, with long progression-free survival times (PFS; Fig. 5B) that are significantly improved over patients not receiving cisplatin (Fig. 5C and D; contrast to other types in Supplementary Fig. S6A–S6C). Despite this benefit, approximately 35% of cervical and lung tumors continue to progress after treatment (Fig. 5B). Accordingly, we investigated whether these variable outcomes could be predicted by the cisplatin VNN (Fig. 5E; Methods). Indeed, we found that cisplatin-treated patients with CESC or LUSC, who were predicted as sensitive to treatment, had significantly better PFS outcomes than those who were predicted as resistant [CESC hazard ratio (HR) = 2.2 at $P = 0.02$, LUSC HR = 3.4 at $P = 0.03$; Fig. 5F–G]. This performance was significantly better than baseline random forest or elastic net models (Fig. 5H; with P values of 9.1×10^{-5} and 1.0×10^{-2} for comparison with random forest and elastic net, respectively). Among the remaining cancer cohorts that failed to show cisplatin benefit—head-and-neck squamous cell carcinoma (HNSC), lung adenocarcinoma (LUAD), and ovarian carcinoma (OV)—we observed that the vast majority of

these patients were predicted as resistant by the cisplatin VNN model, as expected (75%, Supplementary Fig. S6D; compared with 25% for CESC and LUSC, Supplementary Fig. S6E).

We next investigated which protein assemblies were responsible for the clinical predictions. Starting from the 29 assemblies found to be important to cisplatin response during training in cell lines (Fig. 6A), we computed the *in silico* activity of each of these assemblies (per patient) as the first principal component of its neuron values (Methods). This activity was then investigated, separately for each assembly, as a quantitative marker for the prediction of cisplatin response in cervical cancer patients. We found that these 29 assembly activities were generally predictive of cisplatin treatment outcomes and tended to outperform those drawn from other protein assemblies in the NeST map (Fig. 6B; Supplementary Table S5). This observation also applied to the 41 RSi assemblies as well as to the sets of assemblies that were validated in functional screens (Supplementary Fig. S6F). Notably, the 15 assemblies validated by both screening modes exhibited the highest predictive performance (Supplementary Fig. S6F). They also surpassed the predictive power of single-gene biomarkers based on the presence/absence of coding alterations in individual genes (Fig. 6B). Some of the most predictive assemblies included roles in the regulation of β-catenin transcriptional activity, DNA repair, epigenetic modification, and RTK–JAK–STAT signaling (Fig. 6B), most of which had also been validated in the earlier drug sensitivity or replication restart screens (Supplementary Table S4).

For example, the top assembly with predictive power in patients was NeST:126 (regulation of β-catenin transcriptional activity; Figs. 4F and 6C), capturing activation of gene transcription by β-catenin via multiple mechanisms, including β-catenin coactivation (EP300 and CREBBP; refs. 50, 51), negative regulation of β-catenin activity (HDAC2; ref. 52), and regulation of β-catenin ubiquitination (TP53; ref. 53). High *in silico* activity of this assembly was associated with significantly worse PFS outcomes (Fig. 6D), reflecting that genetic alteration of this complex predicts cisplatin resistance. Notably, single proteins in this assembly with frequent genetic alterations in cancer patients, such as TP53, were not individually predictive of drug response (Fig. 6E; *TP53* mutation frequency 14% in TCGA cervical cancer). *BARD1* and *TOP2A*, two DNA repair genes that do not encode members of NeST:126, exhibited relatively high predictive accuracy but had limited predictive value due to the rare occurrence of alterations in these genes in cervical cancer patients, making them inapplicable to a large portion of the population (Fig. 6E). The assembly as a whole, however, had both high alteration frequency and relatively high predictive accuracy (Fig. 6E; 42% alteration frequency, C-index: 0.62). Beyond the cisplatin VNN, we also evaluated the performance of the other RSi drug models on prediction of cisplatin survival outcomes, including the single-drug and multidrug models. Although other single-drug VNNs were less accurate than the cisplatin VNN, their C-index generally outperformed the non-RSi VNNs (Supplementary Fig. S6G). The multidrug VNN was also able to stratify cisplatin-treated patients into resistant and sensitive groups with comparable accuracy to the cisplatin VNN (Supplementary Fig. S6G).
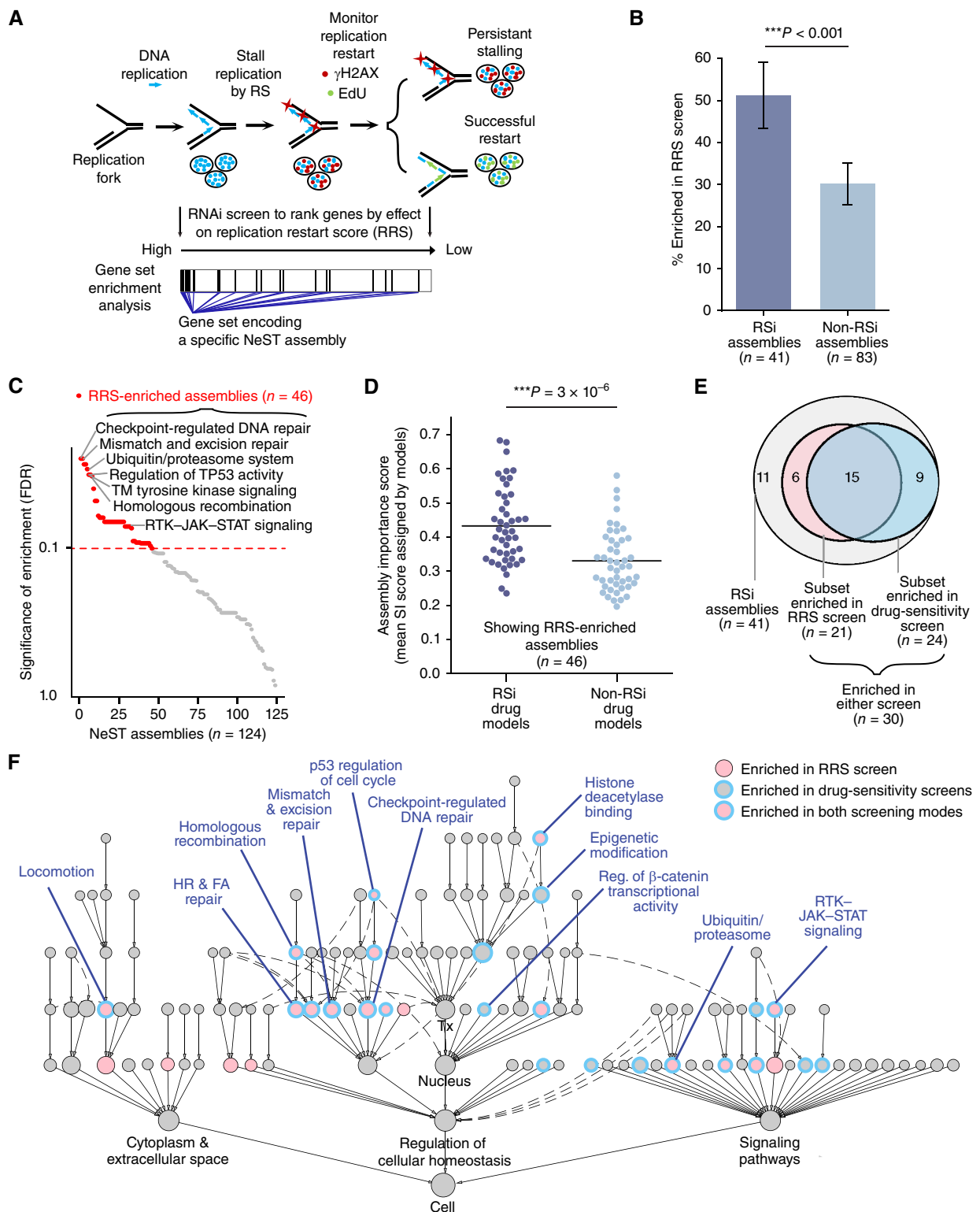
Figure 4. Evaluation of important assemblies by systematic siRNA screening. **A,** Assemblies are analyzed for their enrichment for gene knockdowns that affect RS response via the absolute value of the RRS, defined as the ratio of γH2AX to EdU intensity readouts. RRS screening data by Kavanaugh et al. (49). **B,** Percent of assemblies validated by enrichment in RRS assay, shown separately for those important to RSi drug models (left) vs. all others (right). Error bars display the standard error of the proportion. **C,** NeST assemblies ranked by their degree of enrichment, from most to least significant. Assemblies with FDR < 0.1 are highlighted in red (n = 46). Selected RSi assemblies were labeled. **D,** Mean importance scores for the 46 RRS-enriched assemblies (highlighted in **C**), shown separately for RSi (left) and non-RSi drug models (right). **E,** Venn diagram showing numbers of RSi assemblies validated by enrichment in the RRS assay or a separate chemogenetic drug-sensitivity analysis (Supplementary Fig. S5). **F,** Validated assemblies in the two functional assays are indicated on the NeST assemblies map. ***, P < 0.01; P value by two-sample z-test and Mann–Whitney U test.
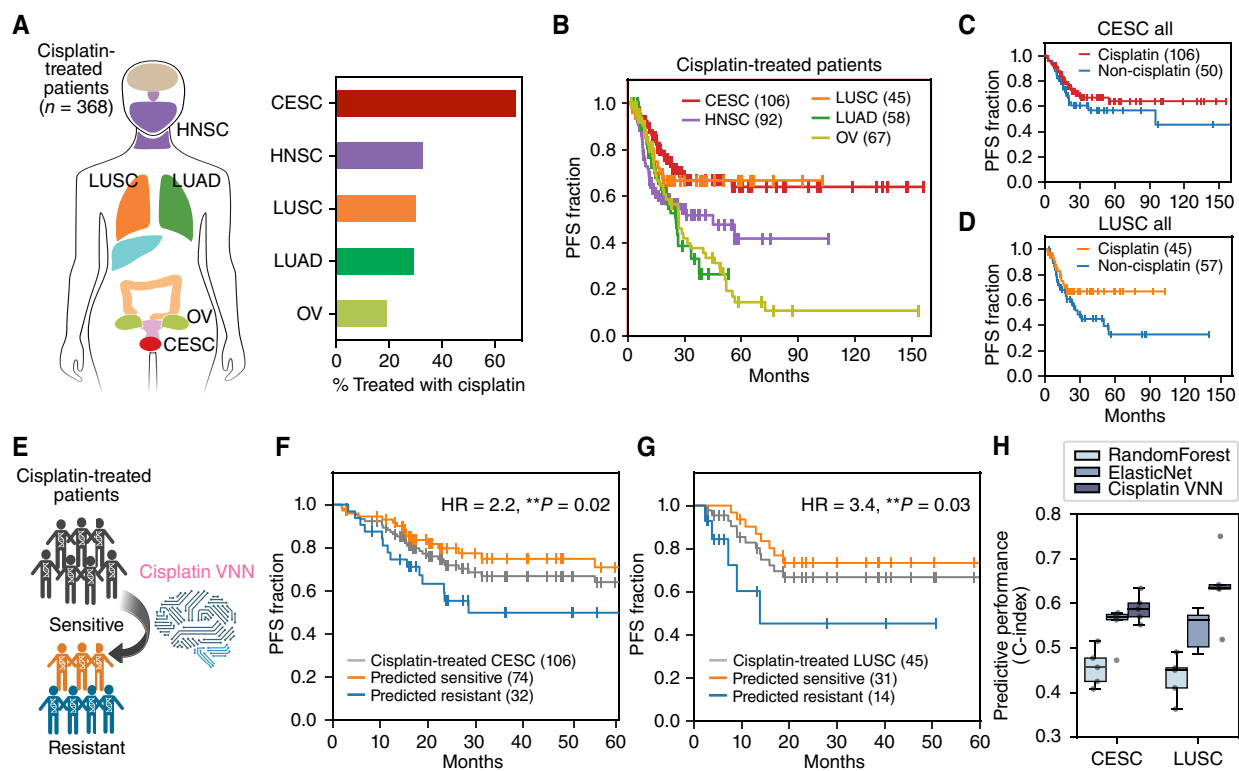
Figure 5. Model translation to predict clinical outcomes in cervical and lung cancers. **A,** Top five cancer types in TCGA by the largest number of patients treated with cisplatin: CESC, LUSC, HNSC, LUAD, and OV. Right barplot shows the fraction of patients of each type that were treated with cisplatin. **B,** Kaplan–Meier survival curves of PFS for TCGA patients with various cancer types that are indicated by different colors. Patients with CESC or LUSC were chosen for the cisplatin VNN model validation because they showed the best outcome following cisplatin treatment. **C,** Kaplan–Meier survival curves of PFS for cisplatin-treated and non–cisplatin-treated patients with CESC. **D,** Kaplan–Meier survival curves of PFS for cisplatin-treated and non–cisplatin-treated patients with LUSC. **E,** Application of the cisplatin VNN to predict patient responses to cisplatin treatment. **F,** Kaplan–Meier survival curves of PFS for patients with CESC stratified by the cisplatin VNN prediction status (sensitive vs. resistant, shown by orange vs. blue). **G,** Kaplan–Meier survival curves of PFS for patients with LUSC stratified by the cisplatin VNN prediction status (sensitive vs. resistant, shown by orange vs. blue). HR, hazard ratio. **H,** Predictive performance in response to cisplatin of RandomForest, ElasticNet, and the cisplatin VNN models (boxplots of different colors) in patients with CESC (left) or LUSC (right). Performance measured using the concordance index (C-index). Boxplot midline is median, box boundaries show upper and lower quartile, and dots show C-indices measured by five respective cross-validation modes of RandomForest, ElasticNet, and the cisplatin VNN models. **, $P < 0.05$; P value by log-rank test.

## The RTK–JAK–STAT Assembly as a Predictive Marker of Cisplatin Response

The RTK–JAK–STAT assembly (NeST:89) provides an excellent illustration of how predictive accuracy is driven by the integration of multiple rare alterations in a protein assembly (Fig. 7A). It consists of dense interactions of upstream growth factors (EGF) and receptor tyrosine kinases (RTKs, including EGFR and ERBB2/3/4) with downstream signaling messengers and effectors (e.g., JAK1/2, STAT3/5A; Fig. 7B). Genetic alterations in this assembly had been initially identified as important for predicting cisplatin and gemcitabine responses in cell lines (Fig. 3A). The impact of such alterations was subsequently validated in both the drug sensitivity and replication restart screens (Fig. 4C and 4F; Supplementary Fig S5I; Supplementary Table S4), and high activity of this assembly was shown to be predictive of drug resistance in cisplatin-treated cervical cancer patients (Figs. 6B and 7C, HR = 2.8, $P = 0.004$).

Notably, none of the individual proteins in this complex were frequently genetically altered in cervical cancer, with all frequencies less than 10% (Fig. 7D). As an integrator of these rare alterations, however, the complex as a whole showed a relatively high frequency of alteration (44.3% in TCGA cervical cancer) and predictive accuracy (C-index: 0.58; Fig. 7D). Perturbation of this complex (*in silico* activity > 0) was associated with drug resistance, driven predominantly by alterations spread over eight genes. They included copy-number amplifications in *ERBB2*, *JAK2*, and *EGFR*; copy-number deletions in *ERBB4*, *CBL*, and *IRS1*; and point mutations in *CBLB* and *PDGFRA* (Fig. 7E, several of these genes had multiple alteration types). In general, the presence of any of these alterations was able to signal high assembly activity and thus cisplatin resistance, following approximate OR-type Boolean logic (Fig. 7F).

DISCUSSION

Here we have advanced a set of interpretable deep learning models aimed at understanding genetic mechanisms of susceptibility and resistance to drugs that induce replication stress. Instead of associating single-gene alterations with drug responses directly, the strategy is to project these alterations on a map of protein complexes and larger molecular assemblies associated with cancer. This approach is prompted
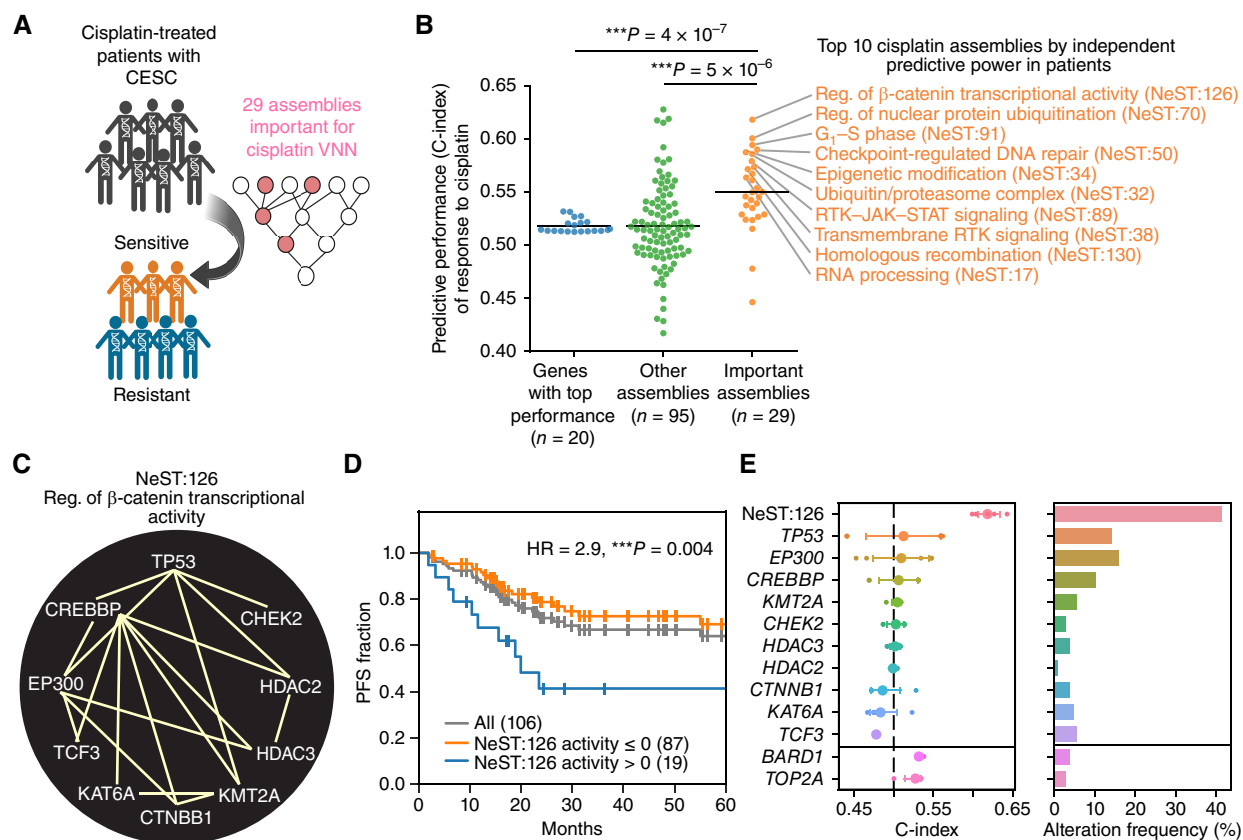
Figure 6. Protein assemblies as biomarkers of cisplatin outcome. **A,** Evaluation of individual assemblies for prediction of responses to cisplatin in patients with CESC. **B,** Swarmplot showing the predictive performance (C-index) of genes and assemblies in response to cisplatin in patients with CESC. Assemblies are grouped into the 29 important for cisplatin response in cell lines (right) vs. all other assemblies (middle). The top 10 predictive assemblies in patients with CESC are listed in orange. **C,** Interaction network of proteins in NeST:126. Edges denote Integrated Association Stringency (IAS) scores used to assemble the NeST map (36). **D,** Kaplan–Meier survival curves of PFS for cisplatin-treated patients with CESC stratified by *in silico* activity of NeST:126. **E,** Left, predictive performance (C-index) of NeST:126 assembly (top row) and its genes (all other rows except BARD1 and TOP2A) for prediction of CESC cisplatin response. Right, corresponding frequency at which the assembly or its genes are genetically altered in patients with CESC. Large circles indicate the mean C-index over five models from cross-validation. Error bars denote 95% confidence intervals (CI). C-index and alteration frequency of the top two predictive genes BARD1 and TOP2A that are not members of NeST:126 are shown at the bottom, separately. ***, $P < 0.01$; P value by Mann–Whitney U test and log-rank test.

and supported by the concept that cancer is a network-based disease arising from the action of hallmark cancer pathways (11, 54). According to this concept, a particular driving mutation may occur rarely in a tumor population, but such rare events can sometimes be better understood and modeled by their impacts on common subcellular components. In this regard, our model identifies 41 protein assemblies in which genetic alterations modulate RSi drug responses, of which many could be corroborated by systematic gene loss-of-function assays (Fig. 4 and Supplementary Fig. S5) and/or predictive power in clinical cohorts (Figs. 5–7).

Predictive Assemblies Identified by the Models

It is widely recognized that cancer cells acquire resistance to the damage caused by cisplatin and other RSi agents through multiple mechanisms, with some of the key factors being the activation of DNA damage response pathways as well as rewiring of antiapoptotic and prosurvival signaling cascades (24, 55). In agreement with these previous studies, our models highlight the significant roles of alterations in DNA repair–related assemblies, including mismatch and

excision repair (NeST:30) as well as homologous recombination and Fanconi anemia repair (NeST:79; Fig. 3A and Supplementary Table S3). Such assemblies also capture the previously observed interactions of multiple DNA polymerases, including pol δ, ε, and η (POLD1, POLE, and POLH; refs. 56–58), with DNA repair proteins in regulating cisplatin resistance. Beyond DNA repair, an established component of cisplatin resistance is the dysregulation of cell growth and survival signaling, including the insulin receptor PI3K–AKT–FOXO (INSR-FOXO, NeST:52; refs. 59, 60) and RAS–MAPK (NeST:73; ref. 61) cascades, aspects that our models also capture (Fig. 3A; Supplementary Table S3).

In addition to these well-characterized response pathways, a growing body of evidence has begun to implicate aberrant epigenetic regulation in cisplatin resistance, but the mechanistic details have remained unclear (62, 63). Here, all RSi drug models (including the specific agent models and the multidrug model) identify an assembly of epigenetic regulators in which alterations to multiple genes converge to modulate drug resistance in both cell lines and patients (epigenetic modification, NeST:34; cell lines: Fig. 3A and
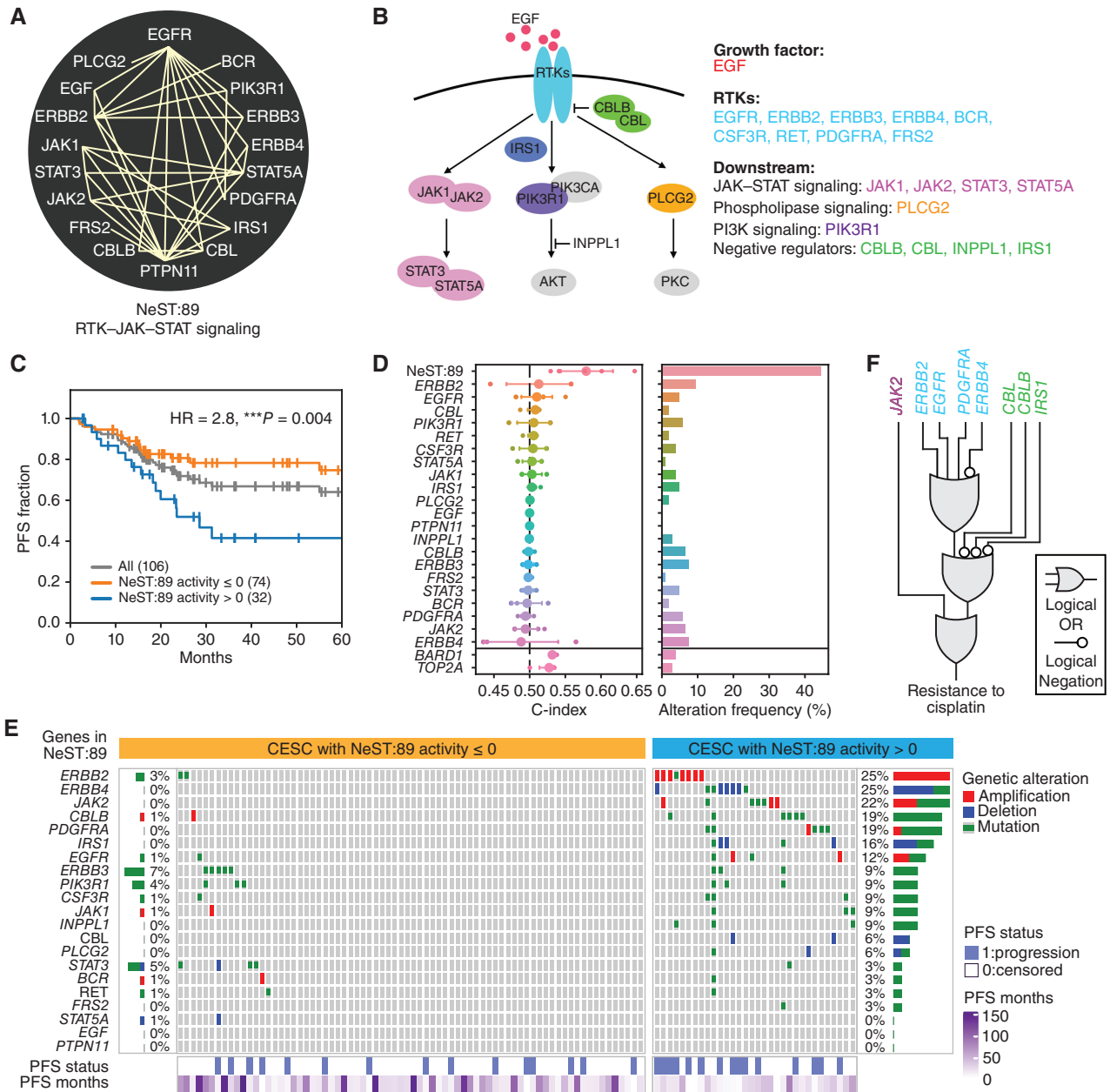
Figure 7. RTK–JAK–STAT assembly as an example predicting cisplatin resistance. **A,** Protein network defining NeST:89 (RTK–JAK–STAT signaling). Edges denote Integrated Association Stringency (IAS) scores used to assemble the NeST map. **B,** Oncogenic signaling pathways captured by NeST:89. **C,** Kaplan–Meier survival curves of PFS for cisplatin-treated patients with CESC stratified by *in silico* activity of NeST:89. HR, hazard ratio. ***, *P* < 0.01; *P* value by log-rank test. **D,** Left, predictive performance (C-index) of the NeST:89 assembly (top row) and its genes (all other rows except BARD1 and TOP2A) for prediction of CESC cisplatin response. Right, corresponding frequency at which the assembly or its genes are genetically altered in patients with CESC. Large circles represent the mean C-index over five models from cross-validation. Error bars denote 95% confidence intervals. C-index and alteration frequency of the top two predictive genes *BARD1* and *TOP2A* that are not members of NeST:89 are shown at the bottom, separately. **E,** OncoPrint showing NeST:89 genes (rows) by cisplatin-treated patients with CESC (columns). Patients were stratified into binary response subtypes based on NeST:89 status (low activity at left, enriched for cisplatin-sensitive patients; high activity at right, enriched for cisplatin-resistant patients). Genes sorted from top-to-bottom based on alteration frequency in the cisplatin-resistant subtype. **F,** Boolean circuit diagram representing the approximate logic function by which the genetic alteration status of eight NeST:89 genes predicts resistance to cisplatin in the CESC cohort. Inputs from activating or upregulating alterations are directly connected to the OR gates, whereas inputs from inactivating or downregulating alterations are first negated to represent an opposite signal.

Supplementary Fig. S4C; patients: Fig. 6B). This assembly integrates genetic alterations across DNA methyltransferases (*DNMT1, DNMT3A, DNMT3B*), histone methyltransferases (*EZH2, PRDM1*), histone demethylase (*KDM5A*), histone acetyltransferase (*CREBBP*), histone deacetylases (*HDAC1–3*),

chromatin regulators (*SMARCA1, CHD4, MTA1*), and transcription factors (*RUNX1, ZBTB2, ZNF217, BRD4, TP53*). Among these, overexpression of the histone methyltransferase *EZH2* has been shown to promote, and its inhibitor to effectively reverse, cisplatin resistance (64–66). One possibility

is that this and other epigenetic factors can affect cisplatin response via interactions with transcription factors in this assembly, such as RUNX1 and the zinc-finger proteins ZBTB2 and ZNF217. The impact of alterations to other proteins in NeST:34 on cisplatin response is less well-studied. Beyond this particular assembly, our model also highlights the importance of assemblies engaged in protein transport (NeST:12), locomotion (NeST:14), and regulation of immune responses (NeST:18). These predictive assemblies present promising avenues for further research and therapeutics development.

The Significance of Integrating across Many Genetic Alterations

Historically, drug response prediction has been pursued predominantly through the search for single-gene markers. Here, a key contribution of our work is to demonstrate how individual genetic alterations may be integrated into a unified quantitative assessment of drug resistance, calling attention to both well-known and understudied effects (Fig. 1A). For example, the predictive NeST:89 assembly (RTK–JAK–STAT signaling, Fig. 7) integrates and quantitatively weighs the effects of well-documented genetic markers of cisplatin resistance in cervical cancer, such as *EGFR*, *JAK2*, *STAT3,* and *STAT5* (67–70). It also integrates additional predictive features that have been less well-appreciated, including copy-number deletion of *CBL* and *ERBB4* as well as mutations of *PDGFRA*. Negative regulators like ubiquitin ligases (CBL and CBLB) may counterbalance the activation of EGFR and other RTKs, influencing the sensitivity to RSi agents like cisplatin. Unlike other oncogenic ERBB family proteins in this assembly, ERBB4 can form tumor-suppressing homodimers and negatively regulate STAT5A (71). This unique aspect of the ERBB4 paralog may explain why in our models *ERBB4* deletions are predictive of cisplatin resistance in cervical cancer patients (Fig. 7E and 7F). More generally, the extensive interactions in this assembly interconnecting RTKs to JAK–STAT factors (Fig. 7A) suggest that alterations of RTKs may trigger downstream JAK–STAT signaling to modulate the cellular response to cisplatin and potentially other RSi agents. Notably, the integration of all of these effects yields an integrated marker of cisplatin resistance with both high alteration frequency and predictive power (Fig. 7D).

Comparison with Previous ML Models

Our work extends previous RSi drug response models along several lines. A first group of models representing the majority of ML drug modeling efforts (72) has focused on mRNA expression features for prediction. For example, Jin and colleagues used expression features very effectively in their state-of-the-art pan-drug-response model, which was also interpretable (73). Here the VNN models are complementary, as they focus on the integration of rare and common genetic alterations, including somatic mutations and copy-number aberrations. A second group of models has sought to identify gene signatures associated with responses to single RSi drugs (74–78), such as with Sui and colleagues, who focused on predicting cisplatin resistance in non–small cell lung cancer (77). Conversely, at the opposite end of the spectrum are models trained across very large numbers of drugs, including Deep-CDR and DrugCell (30, 48). Although such studies do not

focus on RSi agents in particular, they nonetheless include representative RSi agents among hundreds of others, and some of these models also provide mechanistic pathway interpretation. Likely due to their generality, these models appear to sacrifice precision in RSi drug response prediction (e.g., see comparisons in Fig. 2B). In this respect, the multitask implementation presented here appears to strike a useful balance between single RSi drug models and pan-drug models in that it can capture both drug-specific features alongside general mechanisms underlying multiple RSi responses (Supplementary Fig. S3C).

Systematic Validation of Assemblies via Complementary Functional Assays

Our exploration of the predictive protein assemblies invoked two distinct types of genome-wide functional assay, with complementary insights. We first sought to confirm assemblies in which loss of gene function affects tumor cell fitness during drug treatment (47). This assay provides a direct test of the prediction that genetic alterations in an assembly affect drug response, and it scales easily to multiple drugs using a pooled gene knockout library. On the other hand, drug sensitivity provides less information about the specific functions of an assembly, as it integrates the outputs of replication stress pathways with numerous other cellular responses such as growth signaling, apoptosis, immune response, drug export, and so on. In contrast to this first screen, the second genome-wide assay sought to identify assemblies that relate specifically to mechanisms of DNA replication and replication restart (49). This second screen offers richer phenotypic readouts than cell fitness, using dual markers for replication damage and restart, respectively, and thus is a more direct probe of RS responses. On the other hand, not all genes underlying a replication phenotype necessarily affect the ultimate drug response. Furthermore, the direct replication phenotypes were screened in an arrayed format one gene at a time and thus were only available for a single RSi agent in our study. Regardless, 24 assemblies important to the RSi drug models could be supported by drug-sensitivity screens and 21 by replication restart screens, covering 30 assemblies total (Fig. 4E).

Limitations and Future Directions

Although we have mainly explored protein assemblies for their power as integrated genetic markers of drug response, a complementary direction will be to incorporate other layers of omics data, such as gene expression. Another compelling direction will be to explore the important assemblies as a source of drug targets. Targeted therapies that induce tumor-specific replication stress have emerged as a promising avenue for cancer treatment (12), as have synthetic combinations of RSi drugs with second points of intervention (79). In this regard, we identified as many as 41 important protein assemblies that mediate RSi drug responses, some of which may cause further replication stress when targeted. These assemblies might thus be used to design and conduct a systematic screen to disrupt and overexpress proteins from each RSi assembly in combination with a broad panel of FDA-approved drugs. These assemblies also serve as a valuable resource for elucidating the biology of DNA replication stress and its associated response pathways. Furthermore, the presence of common RSi drug regulatory mechanisms

observed in non-RSi drugs (e.g., manumycin and ML210; Figs. 2C and 3C) suggests the possibility that these drugs induce indirect or secondary effects on DNA replication or establish cross-talk between their primary effects and DNA damage responses. Although these drugs are not originally designed to target RS, our study underscores the importance of a systematic elucidation of drug response mechanisms.

METHODS

Preparation of Therapeutics Response Data

Drug response data were retrieved from the GDSC database (with separate data sets GDSC1 and GDSC2) and the CTRP (with separate data sets CTRP1 and CTRP2) in July 2022 (43–46). Together, these repositories covered a total of 692,859 cell line–drug pairs, comprising 1,244 cell lines and 888 drugs with some pairs missing. The numbers of cell lines with response data informing each RSi drug were as follows: cisplatin (947), gemcitabine (1,210), camptothecin (799), etoposide (1,181), olaparib (1,213), and CD437 (825). *Drug information*: Each molecule's published name, synonym, or SMILES string was queried using PubChemPy, and the corresponding associated InChiKey was extracted and stored. Duplicate drugs (within or between repositories) were then matched with one another using InChiKeys, and PubChemPy was used to extract isomeric SMILES strings. Some compounds with no matches were manually annotated. *Cell viability data:* For CTRP, the average percent viability files, which have been normalized to vehicle control, were used. For GDSC1, data were normalized to "cells-only" controls on a per-plate basis. For GDSC2, data were normalized to DMSO control wells on a per-plate basis. Data were then averaged across all replicates for each data set separately. For drug response measurement, we used area under the dose–response curve (AUC), in which AUC = 0 corresponds to complete cell killing and AUC = 1 corresponds to no cell killing; AUC > 1 represents a growth advantage conferred by the drug. The harmonized AUCs calculated in this study were in agreement with AUCs reported by the original consortia (Pearson correlations of 0.92, 0.83, 0.91, and 0.91 for CTRP1, CTRP2, GDSC1, and GDSC2, respectively).

Selection of Gene Features

A set of 718 clinical genes was assembled from the union of those in which mutations and/or copy-number aberrations are assessed by one or more of the following clinical panels: FoundationOne CDx (40), Tempus xT (41), PALOMA-3 trial (80), or Project GENIE (42). To compile genotypes for all cell lines, we extracted nonsynonymous coding mutations and copy-number alterations for the clinical panel genes from the Cancer Cell Line Encyclopedia (release 22Q1; ref. 81). Gene mutations were marked as either present ("1") or absent ("0"), with mutations filtered to recognize the following types: missense, nonsense and nonstop mutations, frame-shift insertions and deletions, splice-site and region variations, and in-frame insertions and deletions. Gene copy-number deletions or amplifications were marked separately, also using binary (1/0) indications. Together, mutations, copy-number deletions, and copy-number amplifications served as features for each of the clinical panel genes.

Model Architecture

We queried the NeST hierarchy of cancer protein assemblies (36) to identify those that contained clinical panel genes. We define a hierarchy of protein assemblies as a multilayered directed acyclic graph in which individual genes recursively cluster together to form assemblies that are connected to other assemblies in parent–child relationships all the way up to the root of the structure that represents a cancer cell (Supplementary Fig. S2). In our models, protein assemblies were filtered to require ≥5 clinical panel genes or >1 child assembly, producing a final hierarchy consisting of 131 assemblies distributed over 7 layers. An assembly in NeST can have both single proteins and other assemblies as its children (Supplementary Fig. S3A).

We trained six single-drug VNNs (29) and one multidrug VNN (Fig. 2A), in which the VNN architecture followed the connections of the 718 genes and 131 assemblies in NeST (Supplementary Fig. S3). Gene alterations (mutations, copy-number deletions, and copy-number amplifications) form the input feature layer, which connect together to form the gene layer (Supplementary Fig. S3A). For any gene g_i, we denote the input features as a vector I and the output as g_i, in which $i \epsilon$ [1, 718], $I_i \epsilon$ [0, 1]^3 and $g_i \epsilon R$. Hence, a gene-layer equation is given by

$$g_i = BatchNorm(Tanh(Linear(I_i)))$$

BatchNorm indicates batch normalization (82), Tan*h* indicates a hyperbolic tangent function, and *Linear* indicates a weighted linear transformation.

The remaining layers of the model represent the 131 NeST assemblies, in which each assembly is represented by n neurons, and every parent–child connection follows the edges in the hierarchical map. An assembly–gene pair is connected through $n \times 1$ connections and an assembly–assembly pair through $n \times n$ connections. The number of neurons is a hyperparameter. Dropout (83) with probability = 0.3 was added to layers 5 through 8 after hyperparameter optimization. Given an assembly A_j connected to k child assemblies and m genes, its input is denoted by a vector I_j of dimension $n*(n*k + m)$ and output by a vector A_j of dimension n, in which $A_j \epsilon R^n$ and $j \epsilon$ [1, 131] . Thus, the equation for an assembly is given by

$$A_j = BatchNorm(Tanh(Linear(Dropout(I_j))))$$

The state of an assembly, which we refer to as the "*in silico* activity," is defined as a function of the states of its k child assemblies and m genes. To reduce *in silico* activity to a single dimension, we compute the first principal component resulting from principal component analysis (PCA; ref. 84) of the set of neuron values for each assembly. The final output for a single-drug model is a linear transformation of the output of the root, resulting in a real number in the range [0,1]. For the multidrug model, the output of the root node is connected to an additional layer of neurons, one for each of d drugs, resulting in an output vector $O \epsilon R^d$. Hence, the output-layer equation is given by

$$O = Tanh(Linear(A_{root}))$$

The objective function (*Loss*) of the VNN aggregates the MSE across every assembly in the hierarchy. The loss function for a single-drug model can be defined as

$$Loss = MSE(O, y) + \alpha \sum_{j \neq root} MSE(Linear(A_j), y) + \beta \|W\|$$

The parameter α was set to 0.3, whereas β was tuned by hyperparameter optimization. *Linear* denotes the linear function used for transforming the vector A_j to a scalar. Note that for the single-drug models, O is already a scalar. W denotes the weights of the neural network. AdamW (arXiv 1711.05101) was used to optimize weights. For the multidrug model, the final loss is the sum of loss for each drug, given by

$$Loss = \sum_d \left(MSE(O^d, y^d) + \alpha \sum_{j \neq root} MSE(Linear(A_j), y^d) + \beta \|W\| \right)$$

NB: Assembly states and weights remain the same because they are shared across all drugs.

Model Training

All models were trained using a five-fold nested cross-validation procedure. For each fold setting, 64% of cell lines were split as a

training set, 16% as a validation set (used in hyperparameter tuning), and 20% as a test set, ensuring that cell line replicate measurements (e.g., from different GDSC or CTRP data sets) were not split between test and training sets. Hyperparameter optimization was performed using Optuna (85). All VNN models were implemented in PyTorch and trained using five GPU servers containing four Nvidia Tesla V100s each with 5120 CUDA cores and 32GB GDDR6 RAM.

Alternate Models for Performance Comparison

We assessed three models for benchmarking the performance of VNN models. We constructed a black-box artificial neural network (ANN; ref. 35) using the Python scikit-learn library (86), which was allotted the same number of neurons and layers as the VNN model. The ANN was trained and optimized using the exact same procedure as the VNN models. We also evaluated two previously published models, DeepCDR (48) and DrugCell (30). Both models were retrained against the exact mutation calls and drug responses used for the VNN models using five-fold cross-validation. We assessed the performance of these models on held-out cell lines, as compared with the held-out drug, cell line pairs used originally in these studies.

Predicting Drug Response Using the Multidrug Model

To predict cellular response to each drug (6 discovery and 39 test drugs), first we divided the cell lines into five held-out sets, such that no cell line was used in training the model. The model predicted six responses per cell line, one for each RSi drug (*O*, Supplementary Fig. S3C). As each fold predicted response to a distinct set of cell lines, the responses from all the folds were simply aggregated to calculate one odds ratio per discovery drug and six odds ratio values per test drug. The odds ratio and its confidence interval were reported for the discovery drugs (Fig. 2B). The mean and standard error of the six ORs were used to compare the predictions across the test drugs (Fig. 2C).

SI Score

To determine the importance scores of protein assemblies for drug response prediction, we adopted a variation of "relative local improvement in predictive power" previously reported by Ma and colleagues (29). For each of the five pretrained models corresponding to each drug, we used Ridge Regression, an L2-norm regularized linear regression method trained on assembly neuron values (*A*, Supplementary Fig. S3A), to model the predicted drug response of an assembly (system). SI was calculated using the Spearman correlation between the prediction of Ridge Regression and VNN drug response for the single-drug models or the *in silico* activity of the root node for the multidrug model. Throughout the paper, SI scores were averaged across all five models. A higher SI score indicated an assembly for which the neurons had a strong contribution to VNN predictions and could therefore be considered important; conversely, a low SI score indicated an assembly for which the neurons were weak predictors of VNN predictions and could therefore be considered of low importance. Assemblies with SI scores of ≥0.5 were referred to as important assemblies (Fig. 3A).

Enrichment of Important Assemblies in Functional Screens

The function "gost" from the R package "g:Profiler" (87) was used to identify the functional enrichment of genes in the NeST assemblies. An assembly-to-genes annotation file (.gmt; Supplementary Table S1) was created using the function "CreatePathwayCollection" in the R package "pathwayPCA." Genes were first rank-ordered by the degree (absolute z-score) to which their loss-of-function modulates the response to an RSi drug in the CRISPR/Cas9 screens or RNAi screen (47, 49). The ranks in each of the chemogenetic screens or the RNAi screen were then loaded into "gost" to find the

overrepresentation of assembles from NeST. The Benjamini–Hochberg procedure was utilized to adjust *P* values for multiple comparisons. Overrepresentation with FDR < 0.1 was considered enrichment in the CRISPR/Cas9 screens. (Supplementary Fig. S5). Note that the models identify assemblies important for sensitivity or resistance, by integrating the effects of both gain- and loss-of-function genetic events. However, the models currently only have binary annotations for mutations, without distinguishing gain or loss of function. Given the integrated bidirectional complexity of mutation functions and response outputs, we evaluated CRISPR knockouts or RNAi knockdowns as signless modifiers for systematic validations. A more detailed evaluation can be performed for a specific assembly by carefully examining individual patient mutations.

Preparation of Cisplatin Clinical Data

Genotypes (mutations, copy-number variations) for patients treated with cisplatin were queried from TCGA (PanCancer Atlas) via cBioPortal (https://www.cbioportal.org/datasets). Patients treated with multiple RSi drugs (cisplatin and etoposide; cisplatin and gemcitabine) were excluded, resulting in 368 cisplatin-treated patients with genomic profiles and clinical information. The patient populations in which cisplatin is the standard of care were selected for a focused model evaluation. We further selected cisplatin-treated cancer types with at least 30 patients. This resulted in five types, including 106 patients with CESC, 45 patients with lung squamous carcinoma (LUSC), 92 patients with head-and-neck squamous cell carcinoma (HNSC), 58 patients with LUAD, and 67 patients with ovarian serous cystadenocarcinoma (OV; Fig. 5).

Model Evaluation for Cisplatin-Treated Patients

We evaluated the prediction performance of the cisplatin VNN model using the concordance index (C-index) and HR (88), which were calculated using "Lifelines," a Python survival analysis library. C-index quantifies the fraction of concordant pairs in patients, comparing predicted AUC with actual survival time, out of all possible pairs. Those pairs were discarded if the earlier time was censored. For a given test set, its risk scores with event time and event status were used to calculate the C-index. A higher C-index value indicates a better time-to-event prediction model, with a C-index of 1.0 indicating perfect prediction and a C-index of 0.5 indicating random prediction. C-index <0.5 indicates an anticoncordance relation. HR is a statistical measure that compares the relative hazard rates between predicted resistant and sensitive patients, using the Cox proportional hazards regression method (89). An HR >1 indicates a higher risk in the predicted resistant group compared with the sensitive group (Figs. 5–7 and Supplementary Fig. S6).

Data Availability

The data sets used in this study are all publicly available: GDSC1 and GDSC2: https://www.cancerrxgene.org/; CTRP1: https://portals.broadinstitute.org/ctrp.v1/; CTRP2: https://portals.broadinstitute.org/ctrp.v2.1/; DepMap 22Q2: https://doi.org/10.6084/m9.figshare.19700056.v2; cBioPortal (https://www.cbioportal.org/datasets). The pretrained models are available on GitHub in their respective repositories.

Code Availability

The source codes is available at GitHub: VNN https://github.com/idekerlab/nest_vnn; multidrug VNN https://github.com/idekerlab/multitask_vnn.

Authors' Disclosures

Authors' Contributions

X. Zhao: Conceptualization, data curation, formal analysis, validation, investigation, visualization, methodology, writing–original draft, writing–review and editing. **A. Singhal:** Conceptualization, data curation, software, formal analysis, validation, investigation, visualization, methodology, writing–review and editing. **S. Park:** Investigation. **J. Kong:** Data curation, investigation. **R. Bachelder:** Investigation. **T. Ideker:** Conceptualization, resources, supervision, funding acquisition, writing–original draft, project administration, writing–review and editing.

Acknowledgments

Note

REFERENCES

1. Burgers PMJ, Kunkel TA. Eukaryotic DNA replication fork. Annu Rev Biochem 2017;86:417–38.
2. Fragkos M, Ganier O, Coulombe P, Méchali M. DNA replication origin activation in space and time. Nat Rev Mol Cell Biol 2015;16:360–74.
3. Zeman MK, Cimprich KA. Causes and consequences of replication stress. Nat Cell Biol 2014;16:2–9.
4. Saxena S, Zou L. Hallmarks of DNA replication stress. Mol Cell 2022;82:2298–314.
5. Byun TS, Pacek M, Yee M-C, Walter JC, Cimprich KA. Functional uncoupling of MCM helicase and DNA polymerase activities activates the ATR-dependent checkpoint. Genes Dev 2005;19:1040–52.
6. Zellweger R, Dalcher D, Mutreja K, Berti M, Schmid JA, Herrador R, et al. Rad51-mediated replication fork reversal is a global response to genotoxic treatments in human cells. J Cell Biol 2015;208:563–79.
7. Saldivar JC, Cortez D, Cimprich KA. The essential kinase ATR: ensuring faithful duplication of a challenging genome. Nat Rev Mol Cell Biol 2017;18:622–36.
8. Ge XQ, Blow JJ. Chk1 inhibits replication factory activation but allows dormant origin firing in existing factories. J Cell Biol 2010;191: 1285–97.
9. Toledo LI, Altmeyer M, Rask M-B, Lukas C, Larsen DH, Povlsen LK, et al. ATR prohibits replication catastrophe by preventing global exhaustion of RPA. Cell 2013;155:1088–103.
10. Macheret M, Halazonetis TD. DNA replication stress as a hallmark of cancer. Annu Rev Pathol 2015;10:425–48.
11. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646–74.
12. Ubhi T, Brown GW. Exploiting DNA replication stress for cancer treatment. Cancer Res 2019;79:1730–9.
13. Dobbelstein M, Sørensen CS. Exploiting replicative stress to treat cancer. Nat Rev Drug Discov 2015;14:405–23.
14. Zhang J, Dai Q, Park D, Deng X. Targeting DNA replication stress for cancer therapy. Genes 2016;7:51.
15. Vesela E, Chroma K, Turi Z, Mistrik M. Common chemical inductors of replication stress: focus on cell-based studies. Biomolecules 2017; 7:19.
16. Zhang J, Chan DW, Lin S-Y. Exploiting DNA replication stress as a therapeutic strategy for breast cancer. Biomedicines 2022;10:2775.
17. Shah MA, Schwartz GK. Cell cycle-mediated drug resistance: an emerging concept in cancer therapy. Clin Cancer Res 2001;7:2168–81.
18. Beretta GL, Gatti L, Perego P, Zaffaroni N. Camptothecin resistance in cancer: insights into the molecular mechanisms of a DNA-damaging drug. Curr Med Chem 2013;20:1541–65.
19. Dias MP, Moser SC, Ganesan S, Jonkers J. Understanding and overcoming resistance to PARP inhibitors in cancer therapy. Nat Rev Clin Oncol 2021;18:773–91.
20. Cao X, Hou J, An Q, Assaraf YG, Wang X. Towards the overcoming of anticancer drug resistance mediated by p53 mutations. Drug Resist Updat 2020;49:100671.
21. Li L-Y, Guan Y-D, Chen X-S, Yang J-M, Cheng Y. DNA repair pathways in cancer therapy and resistance. Front Pharmacol 2020;11:629266.
22. Jia Y, Xie J. Promising molecular mechanisms responsible for gemcitabine resistance in cancer. Genes Dis 2015;2:299–306.
23. Galluzzi L, Senovilla L, Vitale I, Michels J, Martins I, Kepp O, et al. Molecular mechanisms of cisplatin resistance. Oncogene 2012;31: 1869–83.
24. Bhattacharjee R, Dey T, Kumar L, Kar S, Sarkar R, Ghorai M, et al. Cellular landscaping of cisplatin resistance in cervical cancer. Biomed Pharmacother 2022;153:113345.
25. Navaei ZN, Khalili-Tanha G, Zangouei AS, Abbaszadegan MR, Moghbeli M. PI3K/AKT signaling pathway as a critical regulator of Cisplatin response in tumor cells. Oncol Res 2021;29:235–50.
26. Kiss RC, Xia F, Acklin S. Targeting DNA damage response and repair to enhance therapeutic index in cisplatin-based cancer treatment. Int J Mol Sci 2021;22:8199.
27. Li Y, Hostallero DE, Emad A. Interpretable deep learning architectures for improving drug response prediction performance: myth or reality? Bioinformatics 2023;39:btad390.
28. Deng L, Cai Y, Zhang W, Yang W, Gao B, Liu H. Pathway-guided deep neural network toward interpretable and predictive modeling of drug sensitivity. J Chem Inf Model 2020;60:4497–505.
29. Ma J, Yu MK, Fong S, Ono K, Sage E, Demchak B, et al. Using deep learning to model the hierarchical structure and function of a cell. Nat Methods 2018;15:290–8.
30. Kuenzi BM, Park J, Fong SH, Sanchez KS, Lee J, Kreisberg JF, et al. Predicting drug response and synergy using a deep learning model of human cancer cells. Cancer Cell 2020;38:672–84.
31. Elmarakeby HA, Hwang J, Arafeh R, Crowdis J, Gang S, Liu D, et al. Biologically informed deep neural network for prostate cancer discovery. Nature 2021;598:348–52.
32. Chen H-IH, Chiu Y-C, Zhang T, Zhang S, Huang Y, Chen Y. GSAE: an autoencoder with embedded gene-set nodes for genomics functional characterization. BMC Syst Biol 2018;12:142.
33. Hao J, Kim Y, Kim T-K, Kang M. PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data. BMC Bioinf 2018;19:510.
34. Huang X, Huang K, Johnson T, Radovich M, Zhang J, Ma J, et al. ParsVNN: parsimony visible neural networks for uncovering cancer-specific and drug-sensitive genes and pathways. NAR Genom Bioinform 2021;3:lqab097.
35. Hinton GE. Connectionist learning procedures. Artif Intell 1989;40: 185–234.
36. Zheng F, Kelly MR, Ramms DJ, Heintschel ML, Tao K, Tutuncuoglu B, et al. Interpretation of cancer mutations using a multiscale map of protein systems. Science 2021;374:eabf3067.
37. Caruana R. Multitask learning. Mach Learn 1997;28:41–75.
38. Douglass EF Jr, Allaway RJ, Szalai B, Wang W, Tian T, Fernández-Torras A, et al. A community challenge for a pancancer drug mechanism of action inference from perturbational profile data. Cell Rep Med 2022; 3:100492.
39. Yuan H, Paskov I, Paskov H, González AJ, Leslie CS. Multitask learning improves prediction of cancer drug sensitivity. Sci Rep 2016;6: 31619.
40. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol 2013;31:1023–31.

41. Beaubier N, Tell R, Lau D, Parsons JR, Bush S, Perera J, et al. Clinical validation of the tempus xT next-generation targeted oncology sequencing assay. Oncotarget 2019;10:2384–96.

42. Smyth LM, Zhou Q, Nguyen B, Yu C, Lepisto EM, Arnedos M, et al. Characteristics and outcome of AKT1E17K-mutant breast cancer defined through AACR Project GENIE, a clinicogenomic registry. Cancer Discov 2020;10:526–35.

43. Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell 2013;154:1151–61.

44. Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. Cancer Discov 2015;5:1210–23.

45. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 2012;483:570–5.

46. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A landscape of pharmacogenomic interactions in cancer. Cell. 2016;166:740–54.

47. Olivieri M, Cho T, Álvarez-Quilón A, Li K, Schellenberg MJ, Zimmermann M, et al. A genetic map of the response to DNA damage in human cells. Cell 2020;182:481–96.

48. Liu Q, Hu Z, Jiang R, Zhou M. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. Bioinformatics 2020;36:i911–8.

49. Kavanaugh G, Ye F, Mohni KN, Luzwick JW, Glick G, Cortez D. A whole genome RNAi screen identifies replication stress response genes. DNA Repair (Amst) 2015;35:55–62.

50. Takemaru KI, Moon RT. The transcriptional coactivator CBP interacts with beta-catenin to activate gene expression. J Cell Biol 2000;149:249–54.

51. Hecht A, Vleminckx K, Stemmler MP, van Roy F, Kemler R. The p300/CBP acetyltransferases function as transcriptional coactivators of beta-catenin in vertebrates. EMBO J 2000;19:1839–50.

52. Shen Y, Xu L, Ning Z, Liu L, Lin J, Chen H, et al. ARHGAP4 regulates the cell migration and invasion of pancreatic cancer by the HDAC2/β-catenin signaling pathway. Carcinogenesis 2019;40:1405–14.

53. Sadot E, Geiger B, Oren M, Ben-Ze'ev A. Down-regulation of beta-catenin by activated p53. Mol Cell Biol 2001;21:6768–81.

54. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med 2004;10:789–99.

55. Lugones Y, Loren P, Salazar LA. Cisplatin resistance: genetic and epigenetic factors involved. Biomolecules 2022;12:1365.

56. Fuchs J, Cheblal A, Gasser SM. Underappreciated roles of DNA polymerase δ in replication stress survival. Trends Genet 2021;37:476–87.

57. Rocha CRR, Silva MM, Quinet A, Cabral-Neto JB, Menck CFM. DNA repair pathways and cisplatin resistance: an intimate relationship. Clinics 2018;73:e478s.

58. Tian W, Ji Z, Wang J, Meng J, Bi R, Ren Y, et al. Characterization of hotspot exonuclease domain mutations in the DNA polymerase ε gene in endometrial cancer. Front Oncol 2022;12:1018034.

59. Park J, Ko YS, Yoon J, Kim MA, Park J-W, Kim WH, et al. The forkhead transcription factor FOXO1 mediates cisplatin resistance in gastric cancer cells by activating phosphoinositide 3-kinase/Akt pathway. Gastric Cancer 2014;17:423–30.

60. Beretta GL, Corno C, Zaffaroni N, Perego P. Role of FoxO proteins in cellular response to antitumor agents. Cancers 2019;11:90.

61. Achkar IW, Abdulrahman N, Al-Sulaiti H, Joseph JM, Uddin S, Mraiche F. Cisplatin based therapy: the role of the mitogen activated protein kinase signaling pathway. J Transl Med 2018;16:96.

62. Wang N, Ma T, Yu B. Targeting epigenetic regulators to overcome drug resistance in cancers. Signal Transduct Target Ther 2023;8:69.

63. Yang C, Zhang J, Ma Y, Wu C, Cui W, Wang L. Histone methyltransferase and drug resistance in cancers. J Exp Clin Cancer Res 2020;39:173.

64. Sun J, Cai X, Yung MM, Zhou W, Li J, Zhang Y, et al. miR-137 mediates the functional link between c-Myc and EZH2 that regulates cisplatin resistance in ovarian cancer. Oncogene 2019;38:564–80.

65. Sun S, Zhao S, Yang Q, Wang W, Cai E, Wen Y, et al. Enhancer of zeste homolog 2 promotes cisplatin resistance by reducing cellular platinum accumulation. Cancer Sci 2018;109:1853–64.

66. Qiu Z, Zhu W, Meng H, Tong L, Li X, Luo P, et al. CDYL promotes the chemoresistance of small cell lung cancer by regulating H3K27 trimethylation at the CDKN1C promoter. Theranostics 2019;9:4717–29.

67. Hugo de Almeida V, Guimarães IDS, Almendra LR, Rondon AMR, Tilli TM, de Melo AC, et al. Positive crosstalk between EGFR and the TF-PAR2 pathway mediates resistance to cisplatin and poor survival in cervical cancer. Oncotarget 2018;9:30594–609.

68. Morgan EL, Macdonald A. JAK2 inhibition impairs proliferation and sensitises cervical cancer cells to cisplatin-induced cell death. Cancers 2019;11:1934.

69. Gutiérrez-Hoya A, Soto-Cruz I. Role of the JAK/STAT pathway in cervical cancer: Its relationship with HPV E6/E7 oncoproteins. Cells 2020;9:2297.

70. Huang L-L, Rao W. SiRNA interfering STAT3 enhances DDP sensitivity in cervical cancer cells. Eur Rev Med Pharmacol Sci 2018;22:4098–106.

71. Lucas LM, Dwivedi V, Senfeld JI, Cullum RL, Mill CP, Piazza JT, et al. The yin and yang of ERBB4: tumor suppressor and oncoprotein. Pharmacol Rev 2022;74:18–47.

72. Partin A, Brettin TS, Zhu Y, Narykov O, Clyde A, Overbeek J, et al. Deep learning methods for drug response prediction in cancer: predominant and emerging trends. Front Med 2023;10:1086097.

73. Jin I, Nam H. HiDRA: hierarchical network for drug response prediction with attention. J Chem Inf Model 2021;61:3858–67.

74. Mucaki EJ, Zhao JZL, Lizotte DJ, Rogan PK. Predicting responses to platin chemotherapy agents with biochemically-inspired machine learning. Signal Transduct Target Ther 2019;4:1.

75. Gao Y, Lyu Q, Luo P, Li M, Zhou R, Zhang J, et al. Applications of machine learning to predict cisplatin resistance in lung cancer. Int J Gen Med 2021;14:5911–25.

76. Scarborough JA, Eschrich SA, Torres-Roca J, Dhawan A, Scott JG. Exploiting convergent phenotypes to derive a pan-cancer cisplatin response gene expression signature. NPJ Precis Oncol 2023;7:38.

77. Sui Q, Chen Z, Hu Z, Huang Y, Liang J, Bi G, et al. Cisplatin resistance-related multi-omics differences and the establishment of machine learning models. J Transl Med 2022;20:171.

78. Berthelet J, Foroutan M, Bhuva DD, Whitfield HJ, El-Saafin F, Cursons J, et al. Computational screening of anti-cancer drugs identifies a new BRCA independent gene expression signature to predict breast cancer sensitivity to cisplatin. Cancers 2022;14:2404.

79. Ngoi NYL, Pham MM, Tan DSP, Yap TA. Targeting the replication stress response through synthetic lethal strategies in cancer medicine. Trends Cancer Res 2021;7:930–57.

80. Lira ME, Xie T, Deng S, Kinong J, Gao J, Zhu Z, et al. Abstract 2749: Liquid biopsy testing allows highly-sensitive detection of plasma cfDNA mutations in 87 breast cancer-related genes. Cancer Res 77, 2017 (suppl 13; abstr 2749).

81. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer cell line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483:603–7.

82. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Bach F, Blei D, editors. Proceedings of the 32nd International Conference on Machine Learning. Lille, France: PMLR; 7–9 Jul 2015:448–56.

83. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–58.

84. Jolliffe IT. Principal component analysis. NY: Springer New York; 2013.

85. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA: Association for Computing Machinery; 2019:2623–31.

86. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. JMLR. org; 2011;12:2825–30.

87. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update), Nucleic Acids Res 2019;47:W191–8.

88. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA 1982;247:2543–6.

89. Cox DR. Regression models and life-tables. J R Stat Soc 1972;34:187–202.