Predicting Drug Responses by Machine Learning

by

Hanrui Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2024

Doctoral committee:

Professor Yuanfang Guan, Chair
Professor Lana Gamire
Professor Jie Liu
Professor Kayvan Najarian
Professor Duxin Sun
Professor Joshua Welch

Hanrui Zhang

rayezh@umich.edu

ORCID: 0000-0001-9924-0319

DEDICATION

To my family and especially my beloved partner, Adam.

ACKNOWLEDGMENTS

I would express my most sincere gratitude to my mentor, Dr. Yuanfang Guan. Her decision to take me in as her student during my most difficult time has changed my life. For a Ph.D. student like me, who was still in the very early stages of exploring life and science, Yuanfang is not only a good mentor on research, but she unveiled a new perspective of this world in front of me. There are many smart people in this world, and Yuanfang, who is undoubtedly one of them, showed me her unique wisdom beyond the knowledge of sciences: being honest, being true, and being kind; consistently working hard and never giving up, even during the most difficult times. Many people considered gradation school and research as laborious, difficult, and hard. However, working with Yuanfang made me realize research is so joyful, beautiful, and meaningful, that it become something I want to do for the rest of my life. Before meeting Yuanfang, I had no idea what my life direction could be. After all these years, I'm grateful she made me realize my dreams and goals, which once were just a vague silhouette, have become much clearer.

I'm extremely grateful to my partner, Adam, and his family. From now and then, I constantly reflected on that snowy, cold evening, when I had met Adam for the first time, and the most wonderful conversation we had on the 3rd floor of Palmer Commons. Meeting Adam is a miracle, another gift Ann Arbor has given me, except for meeting my excellent mentor. He made me realize I'm finally complete. My immense gratitude also extends to my dear family in Michigan– Karen, Steve, Ryan, and Zach– for their unconditional love and support during my PhD. They are my family far away from China, and now Michigan has become my second

hometown because of them. Lastly, I am eternally grateful to my mother, the strongest woman I've ever met, whose enduring support has shaped me into the person I am today.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACT: Artemisinin-based Combination Therapy

ADME: Absorption, Distribution, Metabolism, and Excretion

ADORA2A: Adenosine A2A Receptor

AI: Artificial Intelligence

ALMANAC: A Large Matrix of Anti-Neoplastic Agent Combinations

API: Application Programming Interface

ART: ARTemisinin

ATCC: American Type Culture Collection

ATM: Ataxia Telangiectasia Mutated

ATR: Ataxia Telangiectasia And Rad3 Related

AUROC: the Area Under the Receiver Operating Curve

AUPRC: the Area Under the Precision-Recall Curve

AZ: AstraZeneca

BCL: B-cell Lymphoma

BET: Bromodomain and Extra-Terminal domain

BRCA1: BReast CAncer gene 1

BRCA2: BReast CAncer gene 2

BTB: Bric-a-brac, Tramtrack, and Broad Complex

CASP3: CASPase 3

CCLE: Cancer Cell Line Encyclopedia

CEP76: Centrosomal Protein 76

CHEK1i: Checkpoint Kinase 1 inhibitor

CHEK2i: Checkpoint Kinase 2 inhibitor

CI: Confidence Interval

CLS: Cell Lines Service

CNV: Copy Number Variation

COL5A1: Collagen Type V Alpha 1 Chain

COPII: Coat Protein Complex II

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats

CSS: Change-Sensitive Score

DDR: DNA Damage Response

DIGRE: Drug-Induced Genomic Residual Effect

DHA: DiHydroArtemisinin

DMSO: DiMethyl SulfOxide

DNA: DeoxyriboNucleic Acid

DNA-PK: DNA-dependent protein kinase

DREAM: Dialogue on Reverse Engineering Assessment and Method

DSB: Double Strand Breaks

DSMZ: Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures GmbH

DGIdb3.0: Drug Gene Interaction Database 3.0

DTP: Development Therapeutics Program

E_{OBS}: Observed Effect

EHR: Electric Health Record

FC: Fold-Change

FCS: Fetal Bovine Serum

FDR: False Discovery Rate

FP2: Finger Print 2

FP3: Finger Print 3

FP4: Finger Print 4

GEO: Gene Expression Omnibus

GI_{50}: concentration for 50% of maximal inhibition of cell proliferation

GO: Gene Ontology

GPR: Gaussian Process Regression

HR: Homologous Recombination

HRD: Homologous Recombination Deficiency

HSA: Highest Single Agent

HSP90AA1: Heat Shock Protein 90 Alpha Family Class A Member 1

HTS: High-Throughput Screening

IC_{50}: Half-maximal Inhibitory Concentration

IMPACT: Integrated Mutation Profiling of Actionable Cancer Targets

IR: IRradiation

KP: Kelch Protein

LIG4: DNA Ligase 4

LINCS: Library of Integrated Network-based Cellular Signatures

LOF: Loss-Of-Function

MACCS: Molecular ACCess Systems keys

MSigDB: Molecular Signatures Database

ML: Machine Learning

MSI: MicroSatellite Instable

MYD88: Myeloid Differentiation Primary Response Protein MyD88

NCI: National Cancer Institute

ORF: Open Reading Frame

PARP1: Poly [ADP-ribose] Polymerase 1

PCR: Polymerase Chain Reaction

PEA15: Proliferation And Apoptosis Adaptor Protein 15

PHIST: Poly-Helical Interspersed Sub-Telomeric protein family

PKA: Protein Kinase A

PKIB: CAMP-Dependent Protein Kinase Inhibitor Beta

PLK1i: Polo Like Kinase 1 inhibitor

PPP4R1: Protein Phosphatase 4 Regulatory Subunit 1

PR: Performance Ratios

PRKDC: PRotein Kinase, DNA-activated, Catalytic subunit

PTPN14: Protein Tyrosine Phosphatase Non-Receptor Type 14

QC: Quality Control

RBBP8: RB Binding Protein 8, Endonuclease

RI: Relative Inhibition

RMSE: Root Mean Square Error

RNA: Ribonucleic acid

RNS: small-subunit ribosomal RNA

RRM2Bi: Ribonucleotide Reductase regulatory TP53 inducible subunit M2B inhibitor

RWD: Real-World Data

SESN1: Sestrin 1

SHAP: SHapley Additive exPlanations

SIFT: Sorting Intolerant From Tolerant

SMILES: Simplified Molecular Input Line Entry System

SNV: Single Nucleotide Variation

SRB: SulfoRhodamine B

TEAD1: TEA Domain Transcription Factor 1

TLR9: Toll Like Receptor 9

TOP1i: topoisomerase 1 inhibitor

TPM: Transcripts Per Million

t-SNE: t-distributed Stochastic Neighbor Embedding

TP53: Tumor Protein P53

UPR: Unfolded Protein Response

VEP: Variant Impact Predictor

XPA: Xeroderma Pigmentosum Group A-Complementing Protein

XRCC1: X-Ray Repair Cross Complementing 1

XRCC6: X-Ray Repair Cross Complementing 6

YWHAZ: Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein Zeta

ZIP: Zero Interaction Potency

ABSTRACT

Machine learning (ML) has revolutionized the pharmaceutical industry in recent decades, influencing molecule design, drug target identification, biomarker discovery, and various stages of drug development. This transformation, driven by the synergy between ML and high-throughput drug screening technologies, has broadened the scope for novel treatments and therapeutic indications. This dissertation explores the application of ML algorithms in surmounting fundamental challenges in drug development, including stabilizing high-throughput screening outcomes and transforming initial discoveries into clinical practices.

The first part of the dissertation enhances the generalizability of drug-based experimental results. Our first project in this part assesses the reproducibility across experimental batches in vitro, using data from DrugComb, the most extensive public portal for combination treatment currently available. A critical experimental variable identified is the concentration selection for dose-response matrices. To address this, a concentration imputation method is implemented during feature preparation, markedly improving the predictive transferability of ML algorithms across datasets. The next project shifts focus to the transferability of results between different biological contexts (*in vivo* and *in vitro*). I present the winning algorithm from the Malarian DREAM Challenge, which predicts artemisinin resistance in laboratory isolates using models trained on transcriptome and response data from *Plasmodium falciparum* strains. This project tackles challenges arising from different microarray platforms, response evaluation methods, and biological backgrounds. A rank normalization method is employed to mitigate platform

discrepancies, and model visualization highlights key genes and pathways indicative of artemisinin resistance in both *in vivo* and i*n vitro* settings.

The second part discusses ML's role in discovering new treatments, using DNA damage response (DDR) targeted combination therapy as a case study. An original high-throughput screening dataset featuring 87 anti-cancer drugs and 12 cancer tissues is introduced for DDR combination therapy. Effective and synergistic treatments were identified in combination with ATM, ATR, or DNAPK inhibitors. An ML model is developed, incorporating molecular readouts, synthetic lethality, drug-target interaction, biological networks, chemical structure, and drugs' modes of action, to predict DDR combination treatment responses in new biological contexts. This model shows promise in prescribing optimal DDR treatments based on the patient's biological characteristics, enhancing treatment responses. Furthermore, a core gene panel of only 40 genes was found to be more efficient in predicting DDR combination treatment responses than using full genomic or transcriptomic profiles, leading to the development of a rapid-selection interface for DDR combination treatments in pharmaceutical and clinical applications.

CHAPTER I: Introduction

Drug Development and Current Challenges

Modern drug development is an intricate journey of multi-stage process (Office of the Commissioner, 2020), commencing with the initial stage of drug discovery. This phase is characterized by rigorous target identification and validation, a crucial step in pinpointing biological markers—typically proteins or genes—associated with specific diseases (Mohs & Greig, 2017; Morgan et al., 2018). Post-validation, the procedure employs high-throughput screening (HTS), a method involving the testing of numerous compounds against the identified target to unearth potential drug candidates (Wildey et al., 2017). Following HTS, the journey progresses to lead identification and optimization. This critical phase refines initial HTS 'hits' through a series of chemical modifications, enhancing their pharmacological properties, such as efficacy and safety. These enhanced compounds are then prepared for the pivotal preclinical testing stage, involving both in vitro and *in vivo* studies to ascertain their safety and biological activity (Van Norman, 2016a). Successful preclinical evaluation paves the way for clinical trials, systematically divided into three phases. Phase I trials primarily focus on assessing the drug's safety and dosage in a small cohort. Phase II trials broaden the scope to evaluate efficacy and side effects. Phase III trials, which are more extensive, confirm the drug’s effectiveness, monitor side effects, and compare it to existing standard treatments (Van Norman, 2016b). Post-clinical trials, the drug undergoes a stringent review and approval process by regulatory authorities such as the FDA, ensuring its safety and efficacy for the intended use (Sherman et al., 2016). Once approved, the drug is launched into the market, subject to continuous post-market surveillance to

monitor its long-term effects and maintain ongoing safety and effectiveness (Eichler et al., 2012). The extensive nature of this process often leads to significant investments in research, with the average cost for developing a new drug often exceeding one billion dollars (DiMasi et al., 2016), and the timeline extending over a decade (*Research and Development Policy Framework*, n.d.). A key challenge in this process is the high rate of failure, particularly during clinical trials, with many compounds failing to demonstrate efficacy or safety in human trials. This issue is especially pronounced in the development of treatments for complex diseases like Alzheimer's or cancer (Vamathevan et al., 2019), presenting opportunities for the application of machine learning to optimize various aspects of drug development.

Reproducibility and Transferability of Pharmaceutical Knowledge

The reproducibility between differential experimental batches, i.e. high-throughput drug screenings, is crucial for the identification of reliable drug candidates. However, due to variations in experimental conditions, such as differences in cell lines, reagents, or assay protocols, results from one HTS experiment may not always be directly applicable to another, and conclusions and methods that are developed from data within a single study can be not generalizable to different studies (Ding et al., 2017; Larsson et al., 2020; Xia et al., 2022; H. Zhang, Wang, et al., 2023).

Another great challenge in the transferability of pharmaceutical knowledge is between the preclinical *in vitro* to *in vivo* models, primarily due to the fundamental differences between the simplified, controlled environments of in vitro systems and the complex, multifaceted nature of living organisms in *in vivo* studies. *In vitro* experiments, often conducted using cell lines or tissue cultures, provide valuable initial insights into the biological activity of compounds. However, these settings lack the intricate interplay of systems found in an organism, such as

immune responses, metabolic processes, and organ interactions. Consequently, a drug candidate that appears effective and safe in a controlled *in vitro* environment may not exhibit the same properties *in vivo*, where metabolism, bioavailability, and potential toxicity present significant hurdles (J. Yadav et al., 2021).

Another challenge is the physiological relevance of the in vitro models. While these models are instrumental for initial screenings, they may not accurately mimic the disease state or the tissue-specific context in an organism. This discrepancy can lead to misleading results regarding a drug's efficacy or mechanism of action (Horvath et al., 2016). Moreover, the pharmacokinetic and pharmacodynamic profiles of compounds can differ markedly between *in vitro* systems and living organisms. *In vitro* studies do not account for factors such as drug absorption, distribution, metabolism, and excretion (ADME), which are critical for determining a drug's effectiveness and safety *in vivo* (Markossian et al., n.d.).

These challenges highlight the necessity for advancing methodological approaches and developing more representative models, aiming to harmonize the outcomes of *in vitro* and *in vivo* studies, thereby enhancing the translational success rate in drug development.

Machine Learning in the Drug Development Process

Addressing the aforementioned challenges, machine learning methods have emerged as pivotal in expediting the drug development pipeline, enhancing both efficiency and reliability in the discovery of new pharmaceuticals. These methods adeptly manage vast data quantities and facilitate information compression. Notably, machine learning techniques such as transfer learning, few-shot learning, and reinforcement learning are increasingly utilized to augment the transference of insights from preclinical studies to clinical trials within the pharmaceutical domain.

Transfer learning, in particular, has demonstrated potential in harnessing data from analogous fields to address data paucity in specific drug research areas. This technique allows for the application of models trained on extensive datasets to smaller, similar datasets, thereby enhancing drug response and toxicity predictions (Cai et al., 2020). In the context of rare diseases, where data are often scarce, few-shot learning proves invaluable. This approach, capable of making precise predictions from a minimal dataset, substantially accelerates the preclinical research phase and hastens the transition to clinical trials (Ma et al., 2021). Concurrently, reinforcement learning is being employed to refine drug dosing regimens and treatment strategies. Through simulating various clinical scenarios, it aids in identifying optimal treatment plans, thus diminishing the duration and resources required for clinical trials (Korshunova et al., 2022).

Thesis Outline

In this dissertation, I focus on four projects that address the above problems in the pharmaceutical development process. In Chapter II, I introduced an imputation method that normalizes the differences in dose range in the *in vitro* high-throughput drug screening to increase the cross-dataset reproducibility of compound selection. In Chapter III, I describe a machine learning solution to infer *in vitro* treatment resistance from *in vivo* population-based studies. In Chapters IV and V, I describe the preclinical selection of an emerging type of cancer treatment: DNA damage response targeted combination therapy, and how machine learning can play a role in accelerating primary research. In Chapter VI, I summarize my work and propose future directions for these studies.

CHAPTER II: Harmonizing across Datasets to Improve the Transferability of Drug Combination Prediction

Abstract

Combination treatment has multiple advantages over traditional monotherapy in clinics, thus becoming a target of interest for many high-throughput screening (HTS) studies, which enables the development of machine learning models predicting the response of new drug combinations. However, most existing models have been tested only within a single study, and these models cannot generalize across different datasets due to significantly variable experimental settings. Here, we thoroughly assessed the transferability issue of single-study-derived models on new datasets. More importantly, we propose a method to overcome the experimental variability by harmonizing dose-response curves of different studies. Our method improves the prediction performance of machine learning models by 184% and 1367% compared to the baseline models in intra-study and inter-study predictions, respectively, and shows consistent improvement in multiple cross-validation settings. Our study addresses the crucial question of the transferability in drug combination predictions, which is fundamental for such models to be extrapolated to new drug combination discovery and clinical applications that are de facto different datasets.

Introduction

Combining multiple therapeutic agents has become an emerging strategy in cancer treatment. While the monotherapy approach is often the standard of care, the combination of multiple treatments has become inevitable as multiple comorbid conditions occur in cancer patients (Fowler et al., 2020; Ketcher et al., 2019). Moreover, drug combinations have shown advantages

over monotherapy by overcoming drug resistance, and increasing efficacy by synergistic interactions (Bayat Mokhtari et al., 2017). To accelerate the development of new combination therapies, a large number of studies on high-throughput screening of drug combinations have been launched (Forcina et al., 2017; Holbeck et al., 2017; O'Neil et al., 2016), and thereafter have been made comparable in large-scale databases such as DrugComb (Zagidullin et al., 2019; Zheng et al., 2021), DrugCombDB (H. Liu et al., 2020), and SYNERGxDB (Seo et al., 2020). These databases provide abundant resources for training a powerful model to predict new potent combination treatments. For example, multiple machine learning tools have been developed, by hundreds of international participants in the NCI-DREAM Drug Sensitivity and Drug Synergy Challenge, and the AstraZeneca-Sanger Drug Combination Prediction (AZ-DREAM) Challenge (Bansal et al., 2014; Menden et al., 2019).

However, most existing drug combination prediction models have been trained and tested using the same datasets (Julkunen et al., 2020; J. Li et al., 2020; Shim et al., 2022; Sidorov et al., 2019; Torkamannia et al., 2022; Zagidullin et al., 2021; T. Zhang et al., 2021). Cross-dataset prediction remains a significant challenge due to experimental variability between independent studies (Larsson et al., 2020; Xia et al., 2022). For example, when determining the drugs' efficacy, different dosing regimens are used. The O'Neil study used 5×5 dose-response matrices to determine the drug combination response (O'Neil et al., 2016), while the ALMANAC drug combinations were tested by 4×4 or 6×4 dose-response matrices (Holbeck et al., 2017). While different dosages may not have a huge impact on summary monotherapy measurements, such as Hill coefficient (slope of the dose-response curve), IC_{50} (dose at 50% of maximum response), GR_{AoC} (area over the dose-response curve), and RI (relative inhibition normalized by positive control) (Hafner et al., 2017; Malyutina et al., 2019), they may

easily result in different interpolations of the dose-response curves, thus are often not used as features by machine learning models for cross-study drug combination prediction (Güvenç Paltun et al., 2021).

Due to the above challenges from different experimental settings, previous drug combination machine learning models only considered the summary monotherapy measurements as their dose-response features (Menden et al., 2019; Torkamannia et al., 2022). The complete dose-response curves of monotherapies, which contain the full spectrum of pharmacodynamics under different doses, cannot be fully captured by a single summary metric (Calabrese, 2014, 2016). Therefore, a method for harmonizing different dose settings is crucial for cross-study drug combination machine learning models.

In this study, we propose to explore drug combination prediction across different studies with variable dose settings. In particular, we develop a method to standardize the dose-response curves across different studies. We show that such a method enables more efficient utilization of pharmacodynamics profiles of monotherapies in machine learning models, hence improving the prediction accuracy when transferring to new datasets. Our modeling strategy is of particular importance to solve the replicability issue of machine learning for drug combination discovery.

Results

A framework of intra- and inter-study machine learning prediction

Our goal is to test the capability of machine learning models in predicting combination treatment response, not only within a single study but also between different studies and on unseen drug combinations. To achieve this goal, we first explore the publicly available high-throughput screening datasets for anti-cancer combination treatments, to build a gold standard for our experiment. We explore the current latest version of the DrugComb portal

(https://drugcomb.org/), which contains the most comprehensive publicly-available drug combination high-throughput screening datasets, including 24 independent studies. Among them, we select four major datasets: ALMANAC, O'Neil, FORCINA, and Mathews, as they are of the biggest sizes and therefore are commonly used in machine learning prediction of combination responses (Fan et al., 2021; Güvenç Paltun et al., 2021; Preuer et al., 2018; Shim et al., 2022; Sidorov et al., 2019; Xia et al., 2018; T. Zhang et al., 2021). These four studies contain a total of 406,479 drug combination experiments, 9,163 drugs, and 92 cell lines, while the size, drug, and cell line composition, as well as experimental settings, vary significantly among them (Supplementary Table 1). Of the four datasets, ALMANAC is the largest dataset with the most drug-cell line combinations, and FORCINA has the largest number of drugs screened. O'Neil has the best quality, where all the combinations are tested with four replicates, whereas ALMANAC tested at most three replicates for each combination and Mathews tested two replicates for each combination. In contrast, the FORCINA dataset contains no replicates.

We carry out a two-step cross-study validation strategy (**Figure 2.1a**). First, we train dataset-specific models and carry out intra-study cross-validation. The training and testing sets in this step do not share the same treatment-cell line combinations. Therefore, we aim to test the performance of machine learning models in predicting unseen combination treatments within the same study. Next, during the inter-study cross-validation step, we test these dataset-specific models on new individual datasets, which are denoted as "1 vs 1" in **Figure 2.1a**. Furthermore, to explore more versatile inter-study scenarios, we design a "3 vs 1" cross-validation strategy by combining three of the four datasets as the training set and the remaining one as the test set.

To analyze the potential of transferability, we determine the overlap of the drugs, cancer cell lines, and treatment-cell line combinations between the four studies (**Figure 2.1b**). While drugs

are overlapped between all the studies, no overlap of cell lines exists between FORCINA and Mathews with the other datasets, since both FORCINA and Mathews include only one unique cancer cell line. Overall, only 612 treatment-cell line combinations exist between ALMANAC and O'Neil, providing reference data for evaluating the performance of cross-dataset prediction.

Using the replicates within each dataset and the overlapping treatment-cell line combinations between the datasets, we analyze the reproducibility of a drug combination sensitivity score called CSS (Malyutina et al., 2019), as well as multiple drug combination synergy scores, including S, Bliss, HSA, Loewe, and ZIP (Malyutina et al., 2019; B. Yadav et al., 2015). The intra- and inter-study reproducibility can be used as a benchmark for the drug combination prediction model we build in the next step (**Supplementary Figure 2.1**). While no replicates exist in the FORCINA dataset, the O'Neil dataset shows the best intra-study replicability (0.93 Pearson's r for CSS, 0.929 Pearson's r for S, 0.778 for Bliss, 0.777 for HSA, 0.938 for Loewe, and 0.752 for ZIP), possibly due to the relatively more abundant replicates in this study (**Supplementary Figure 2.1**). When testing the overlapping treatment-cell line combinations between ALMANAC and O'Neil, as expected, all the drug combination synergy scores show significant drops of replicability (0.2 Pearson's r for S, 0.12 for Bliss, 0.18 for HSA, 0.25 for Loewe, and 0.09 for ZIP), while the CSS score still maintains a higher correlation (0.342 Pearson's r). The higher reproducibility of the CSS score, both within and across the studies, suggests that drug combination sensitivity is more reproducible than synergy, which may justify why most of the clinically approved drug combinations rely on their combinatorial efficacy rather than synergy (Palmer & Sorger, 2017; Plana et al., 2022).

The above result highlights the challenges of predicting cross-dataset drug combinations including 1) the scarcity of overlapped compounds and cell lines between studies, and 2) the

variability in the assay and experimental settings, such as the total number and ranges of doses. To combat these challenges, we propose a machine learning model using the following features (**Figure 2.1c**): 1) for both drugs, we use chemical structure-derived fingerprints, which can be transferred to chemicals that may not be present in the training set; 2) we use pharmacodynamic properties, such as monotherapy efficacy scores and dose-response curves of the drugs. The dose-response curves will be normalized; 3) we use the expression of 273 essential cancer genes (Cheng et al., 2015) to represent the molecular states of the cell lines. The above features will be fed into a lightGBM boosting model, as it has shown higher efficiency than other tree-based algorithms such as XGboost and Random Forest when training on large datasets (Ke et al., 2017). We will evaluate the accuracy of predicting the six types of drug combination response scores (i.e. CSS, S, Bliss, HSA, Loewe, and ZIP).

Combating inter-study variability by integrating monotherapy efficacy and imputation of dose-response curves

We observe that experimental settings differ not only between different studies but also within the same study (**Supplementary Table 2.1 and Supplementary Figure 2.2**). For example, the dose-response matrix ranges from 2×2 (FORCINA) to 10×10 (Mathews), and within the O'Neil dataset, both 4×4 and 4×6 dose-response matrices are used. Meanwhile, the dose ranges differ significantly within and between studies (Supplementary Table 1). For example, within the ALMANAC study, more than 40 different doses were used (**Supplementary Figure 2.2**), and the maximum doses tested for each drug were different due to their distinctive pharmacodynamic properties(O'Neil et al., 2016). Therefore, we precalculate the replicability of monotherapy efficacy scores, in terms of IC_{50}, RI, and the distribution statistics (maximum, minimum, mean, and median of all inhibitions in the dose-response curves) within and between

different datasets (**Supplementary Figure 2.3**). We notice that RI and IC_{50} show comparable reproducibility within datasets, with Pearson's r of RI ranging from 0.363 (within Mathews) to 1 (within O'Neil), while Pearson's r of IC_{50} ranges from 0.537 (within ALMANAC) to 1 (within O'Neil). However, the replicability of IC_{50} is much lower than that of RI in the cross-dataset analysis, (Pearson's r = 0.084 for IC_{50} versus r = 0.451 for RI between ALMANAC and O'Neil). Most dose-response curve shape statistics show Pearson's r better than or comparable with IC_{50} and RI, either within or between studies, suggesting potential in cross-study prediction (**Supplementary Figure 2.3**).

We start exploring the drug combination response prediction based on the monotherapy responses such as efficacy and dose-response curves (M1-M12, **Figure 2.2, Supplementary Figure. 2.4-10, and Supplementary Tables 2.2-5**). Three types of features based on monotherapy responses are constructed, denoted as "drc_baseline", "drc_imputation" and "monotherapy_efficacy", where the former two features are based on the exact dose-response relationships, and the efficacy is summarized score of the curve (IC_{50} or RI) (**Figure 2.2a**). Since the total number of doses varies significantly, we interpolate all the dose-response curves to the same length for all the datasets (**Figure 2.2b**). We test linear, Lagrange, 4-parameter log-logistic regression (LL4) interpolation (M2-M4, **Supplementary Figure 2.5 and 2.6**). Among the three interpolation methods, linear interpolation performs the best in the intra-study cross-validation while LL4 performs the best in the intra-study cross-validation. Furthermore, combining all three methods shows better performances in both scenarios and thus is used in the final "imputation" model (M5, **Supplementary Figure 2.5 b and d**). Also, since using IC_{50} and RI together is generally better than them alone in the intra- and inter-study cross-validations, the final monotherapy efficacy feature contains both measurements (M7-M9, **Supplementary Figures**

2.7 and 2.8). Five models using different combinations of the monotherapy response-based features mentioned above are shown in **Figure 2.2**. We notice that M12, which is a combination of all three types of monotherapy features, performs slightly better in the intra-study cross-validation (101~102% fold change compared to the other models), while M5, which is the pure imputation model, performs the best in the inter-study cross validations (107%~115% fold change compared to the other models), and this advantage is especially significant in the prediction of Bliss (112% ~ 113% fold change compared to the other models) and Loewe scores (119%~138% fold change compared to the other models) (**Supplementary Figure 2.4**). It is expected that M12 performs the best in the intra-study validation since the un-imputed dose-response baseline features contain the doses for dose-response evaluation. These doses chosen for monotherapy response evaluation can be significantly different (**Supplementary Table 2.2**), thus causing biases in the cross-study prediction. However, the monotherapy doses can still be effective for within-study prediction since they contain unique experimental information for each drug. The imputation method, on the other hand, indeed alleviates the biases in the experimental settings and is more universally transferable between different experimental settings, thus M5, which only imputes dose-response information, outperforms all other monotherapy-based models.

When comparing the monotherapy efficacy directly with dose-response curve-based models, interestingly, the efficacy model shows the best performance in inter-study prediction while the worst in intra-study prediction (**Supplementary Figures 2.9 and 2.10**). We notice that the efficacy model performs especially well when trained or tested on the FORCINA dataset, which adopts a 2×2 dose-response matrix design (**Supplementary Figure 2.9a**). We reckon that the

coarse dose-response relationship may not be as good as the total efficacy in this case, as the imputation becomes unreliable with only two doses.

The imputation methods improve the benchmark model's performance in the cross-study prediction

Previously, the DrugComb study provided a benchmark model using the O'Neil dataset, by integrating one-hot encoding of drugs and cell lines as well as drug chemical fingerprints, drug doses, and cell line gene expressions in the model construction (Zheng et al., 2021). In this study, we construct a reference model based on their schemes, by encoding the chemical structure properties and molecular profiles of drugs and cell lines in the feature set, and explore if the imputation method of the dose-response curve can further improve the prediction accuracy across different individual datasets (**Figure 2.3** and **Supplementary Figure 2.11**).

We construct five models step-by-step, from the label information (categorical encoding of both drugs and cell lines) to adding the chemical structure of both drugs encoded by molecular fingerprints and cell line cancer gene expression, to adding monotherapy efficacy, and adding the dose-response curve baseline feature and imputation feature, respectively. The performances of all models are listed as M13-M20 (**Supplementary Tables 2.2-5**). And five models, including M13-16, and M20, are listed for the main comparison (**Figure 2.3a**).

We notice that, while the benchmark models with only information directly from drugs and cell lines (M13 and M14) still achieve decent performances around the experimental reproducibility levels in intra-study cross-validation (Supplementary Table 2), neither of these models achieve better-than-random performances in the cross-study predictions, due to a lack of shared drugs and cell lines across different studies (**Figure 2.3b** and **Supplementary Table 2.3**). Incorporating pharmacological properties such as monotherapy activity on the same cell lines

(M15) improves both the intra-study and inter-study prediction performances to 178% and 1299% compared to the reference model (M13), showing the robustness of monotherapy efficacy information between studies (**Figure 2.3c**). Adding the monotherapy baseline information (M6) further improves the inter-study performance but not the intra-study, possibly due to the same reason we mentioned in the previous section, that the baseline information contains the dose settings, which is a dataset-exclusive artifact. Furthermore, adding the imputed information (M20) further improves the performances in both intra- and inter-study cross-validation, to 184% in the intra-study cross-validation and 1367% in the inter-study validation (**Figure 2.3c**). This improvement is consistent in terms of all the drug combination sensitivity and synergy scores, with 1187% in CSS, 2141% in Bliss, 949% in HSA, 2257% in Loewe, 723% in ZIP, and 2019% in S score, respectively (**Supplementary Figure 2.11b**). Notably, the models achieve better performances than experimental replicates within and between studies (Supplementary Tables 2-5). We conclude that the imputed dose-response curve contains orthogonal information to the monotherapy efficacy, which can be effectively used to improve the prediction of combination treatment response by overcoming the variability between different experimental settings.

To understand which information plays the most important role in the inter-study prediction, we carry out SHAP (SHapley Additive exPlanations) analysis to visualize the contribution of all the features in the best-performing model (M20, **Figure 2.3**). As expected, the dose-response curve-derived feature shows significant SHAP importance and remains the top feature for all the drug combination response score predictions, while the monotherapy efficacy score also shows significant importance in the S score prediction (**Supplementary Figure 2.12**). We then analyze the contributions of the dose-response imputation features specifically and noticed that the imputed responses at the beginning and end of the curve show significant importance in the

prediction, suggesting that the minimum and the maximum response of the monotherapies are informative for predicting the drug combination response (**Supplementary Figure 2.13**).

To demonstrate the robustness of our models in broader inter-dataset validation settings, we carry out 3 vs. 1 cross-validation experiments based on the four datasets we use in this study (**Figure 2.4 and Supplementary Figure 2.14-17**). For each training and test setting, we combine three datasets and use the combination as the training set, then test the model on the remaining datasets. We expect that using a multi-sourced training set can lead to improved model performances, by including more types of drugs and cell lines in the training instances. Thus, the training datasets can potentially contain more transferable information to new datasets. As expected, the optimal model in 1 vs. 1 inter-study cross-validation settings, M20, which is the baseline model plus dose-response curve imputation feature, shows the same advantages compared to the other models, with 910% performance compared to M1 and 1544% performance compared to M2 (**Figure 2.4b**).

Discussion

How to tackle the replicability in results between different studies to draw meaningful conclusions has been a critical issue in drug discovery (Bailey, 1987). During cancer treatment, resistance is frequently developed against monotherapies, and a combination usage of multiple drugs targeting parallel pathways is needed to overcome this issue. While the application of high-throughput screening on cancer cells accelerates the rational design of drug combinations toward clinical trials (Bush et al., 2018; He et al., 2018; Ling & Huang, 2020), the inconsistency between currently available datasets has been a major concern, posing a challenge to translate these in-vitro studies into an in-vivo setting (Blucher & McWeeney, 2014; Caraus et al., 2015; Chan et al., 2016; Szymański et al., 2012; Xia et al., 2022). As the experimental replicability

between independent combination screening datasets can be quite low (0.089~0.342 Pearson's r between ALMANAC and O'Neil) (**Supplementary Figure 2.1**), which is much lower than that for monotherapy screening (0.194~0.683 R^2) (Xia et al., 2022), a robust machine learning strategy is urgently needed for meaningful clinical applications.

Our study, for the first time, addresses the inter-study transferability issue in large-scale screening. We identify a major cause of variability between different studies, which is the experimental setting of drug dosage. The total number of doses, and the dose ranges, can be significantly different between studies, and even between replicates within single studies (**Supplementary Figure 2.3**). Based on the above observation, we consider the dose-response relationship as part of the features in our machine learning model for drug combination sensitivity and synergy prediction and find out that such a modeling strategy significantly improves the transferability of machine learning models between datasets, with an accuracy that is comparable with in-study replicabilities (**Supplementary Tables 2.2-5**).

Our study focuses on the transfer learning between in vitro high throughput drug combination screening studies (Kim et al., 2021), however, future work is needed to further improve the clinical translation of drug combination predictions. For example, it remains unknown whether the top drug combinations from the in vitro studies are transferable to clinical treatment (Plana et al., 2022), and whether the response of monotherapy treatment can help infer clinical efficacious combinations (Jafari et al., 2022; Narayan et al., 2020) Furthermore, a mechanistic model on signaling pathways is needed to validate that the predicted drug combination biomarkers can be used for patient stratification in clinical trials (Boshuizen & Peeper, 2020; Tan et al., 2021). Future modeling of transferability should be carried out between

in vitro and preclinical studies, such as patient-derived ex-vivo and mouse models, as well as multiple clinical trial meta-analyses (Kim et al., 2021; Ma et al., 2021).

Methods

Data collection

Currently, DrugComb has been the largest public data portal for in vitro high-throughput combination treatment screening studies. We selected the four largest datasets (ALMANAC, O'Neil, FORCINA, and Mathews) from DrugComb (https://drugcomb.org/) for the inter and cross-study analysis in this paper, where the detailed comparisons for the four datasets are shown in **Supplementary Table 2.1**.

DrugComb provides six metrics (CSS, S, Bliss, HSA, ZIP, Loewe) for the responses of combination treatments, and two metrics (IC_{50} and RI (relative inhibition)) for the response of single drug treatments. The details of the formula of these metrics have been described in Zheng et al. (Zheng et al., 2021). Briefly, CSS analyzes the overall drug efficacy for the combination treatment, while S, Bliss, HSA, ZIP, and Loewe evaluate the synergy or the degree of interaction between the two drugs used in a combination treatment. Besides the efficacy and synergy metrics for monotherapy/combination therapy, DrugComb also provides the SMILES (Simplified molecular input line entry system) format chemical structure of drugs, which is used for structural encoding in this study.

The transcriptomic profiles of all the cancer cell lines used in this study were obtained from CCLE (Cancer Cell Line Encyclopedia) (https://sites.broadinstitute.org/ccle/datasets). We obtained 279 cancer-associated genes from the IMPACT (Integrated Mutation Profiling of Actionable Cancer Targets) project (Cheng et al., 2015), 273 of which were found to be

overlapped with the CCLE transcriptomic profiles. Therefore, these 273 genes were used for combination treatment response prediction in this study.

Hyperparameters of machine learning models

We chose the lightGBM gradient boosting model as the base learner used in the experiment. The hyperparameters of the lightGBM models were set as follows:

```
param = {'boosting_type': 'gbdt',
'objective': 'regression',
'num_leaves': 20,
'max_depth': 8,
'force_col_wise': 'true',
'learning_rate': 0.05,
'verbose': 0,
'n_estimators': 1000,
'reg_alpha': 2.0,}
```

where the total number of leaves was set to 20 and the maximum depth was set to 8 to avoid overfitting on the training dataset. 'num_boost_round' was set as 500 for boosting iterations.

Training and cross-validation of models within and between studies

For cross-validation of the models, we carried out model training in the following steps:

1) *intra-study training and cross-validation*: in this step, we carried out five-fold cross-validation for model training and testing. We split the training dataset by combination treatment-cell line, therefore the model can be tested on unseen examples to predict new combination treatment synergy and efficacy. As a result, for each of the four datasets, five

models were generated by training on different combination treatment-cell line splits. Since the two drugs in the combination should be considered equally, during the training steps, the first and second drugs were switched and put in the training set again to adjust for the possible bias by order of the two drugs.

2) *1 vs. 1 inter-study validation*: in this step, no extra models need to be trained. The models trained within each study from step (1) were used for prediction in other datasets except for the training dataset. In this step, the final prediction results from the five intra-study models generated from step (1) are ensembled by averaging. The ensemble method can reduce the prediction variance thus improving the stability of inter-study prediction performance (Hashem, 1997).

3) *3 vs. 1 inter-study validation*: To explore the generalization of *1 vs. 1 inter-study validation* in step (2), we tested the same feature settings on datasets with different compositions. In this step, we combined 3 of the 4 datasets as the training set and tested it on the remaining dataset. The training process is still carried out by inter-study five-fold cross-validation as step (1) and tested on the remaining dataset as step (2).

Feature preprocessing and construction

We applied the following types of information to generate an inter-study-transferable model. The chemical and pharmacological properties of both drugs and the biological characteristics of the treated cell lines were used to construct the feature space.

Firstly, we defined a reference model by applying the following types of information:

1) Categorical encoding of the names of both chemical agents in the treatment (denoted as "drug_categorical"), and categorical encoding of the cancer cell line (denoted as

"cell_line_categorical"). Both features were implemented as categorical features during the training of lightGBM models.

2) To provide information in terms of the drugs' chemical properties, we generate molecular fingerprints from the chemical structure of both chemical agents (denoted as "chemical_structure"). 166 MACCS, 1024 Morgan, and 2048 RDK molecular fingerprints were generated based on the SMILES format of the chemical structure of drugs, using openbabel and rdkit modules from Python. The three types of fingerprints were concatenated together directly for the chemical structure encoding.

3) To provide a meaningful biological background of the treated cell lines, we used the gene expression levels of 273 cancer-associated genes obtained from CCLE as the representation of the cell line features (denoted as "cancer_gene_expression"). The gene expression levels for each cell line were quantile normalized before implementation.

4) To provide pharmacological properties of the single drugs, we used two efficacy metrics of each of the cancer drugs on the same cell line: IC_{50} (denoted as "monotherapy_ic50") and RI (denoted as "monotherapy_ri"), where IC_{50} represents the dose of the drug achieving 50% of the maximum response, and RI is the normalized area under the log10-transformed dose-response curve.

5) For more detailed pharmacological properties, and also to evaluate the variability of experimental settings in different studies, we used the information from the dose-response dose of the single drugs on the same cell lines, which is also provided by the DrugComb datasets. We encoded the dose-response curves using different methods as follows:

a) dose-response curve baseline encoding (denoted as "drc_baseline"): the doses of and corresponding responses were flattened as a vector and concatenated together. Since in

different experiments, the total number of doses measured could be different, ranging from two to ten, the total number of doses is padded to ten by -1 from the right. For example, for the monotherapy MK-5108 tested on ES2 cell line, the response was measured at five different doses (μm): [0, 0.075, 0.225, 0.675, 2], and the corresponding response is [0, -0.48, -0.47,4.32, 20.72], then both doses and responses will be padded to [0,0.075,0.225,0.675,2,-1,-1,-1,-1,-1] and [0,-0.48, -0.47,4.32, 20.72,-1,-1,-1,-1,-1], and concatenated together for feature input.

b) dose-response curve imputation encoding (denoted as "drc_imputation"): Instead of directly taking dose-response curve information as the baseline encoding, we normalized the dose-response relationship by interpolation since the dose-response curves within and between different studies are measured by significantly different dose numbers and ranges (**Supplementary Figure 2.4**), the total number of responses on the curve can be different, introducing a significant challenge for applying this information in inter-study validation. Therefore, interpolating the dose-response curves to the same length can help them to be interpreted at the same magnitude. While all dose-response relationships were measured at logarithmic dose scales, the maximum length of the dose-response curve ranges from 2-10. Therefore, all dose-response curves are first log10-transformed and then interpolated to the length of 10. We carried out the following commonly-used interpolation methods and tested the difference between them:

i) Linear interpolation (denoted as "drc_intp_linear"): We use the Numpy Python package to generate the linear interpolated dose-response curve. The linear interpolation is computed using the equation (1):

$$y = y_0 + \left(x - x_0\right)\frac{y_1 - y_0}{x_1 - x_0} \quad \ldots\ldots \text{Eq(1)}$$

where (x, y) is the coordinate for the interpolated point between (x_0, y_0) and (x_1, y_1).

ii) Lagrange interpolation (denoted as "drc_intp_lagrange"): We used the Scipy Python package to compute the Lagrange interpolation of the dose-response curve. The formula for computing Lagrange interpolation is equation (2):

$$y = P(x) = \sum_{j=1}^{n} P_j(x) \quad \ldots\ldots \text{Eq(2)}$$

Where

$$P_j(x) = y_j \prod_{k=1, k\neq j}^{n} \frac{x - x_k}{x_j - x_k},$$

n: total number of doses before interpolation.

iii) Four-parameter log-logistic (LL4) regression interpolation (denoted as "drc_intp_4PL"): As dose-response curves are often fitted by a four-parameter logistic regression function in the standard analysis, we implemented a Python version of the drc R package using the same parameter implementation (Ritz et al., 2015). The LL4 interpolated curve is computed by equation (3):

$$y = b + \frac{c - b}{(1 + exp(a(\log(x) - \log(IC50))))} \quad \ldots\ldots \text{Eq(3)}$$

where,

$$a = \frac{y_n - y_1}{x_n - x_1},$$

$$b = y_{max},$$

$$c = y_{min},$$

$$d = IC50.$$

In total, 20 different combinations of the above features are tested in this paper. For details of all the models, please refer to **Supplementary Table 2.5**, and the corresponding performances are summarized in **Supplementary Tables 2.2-4**.

Visualization of feature importance in machine learning models

To visualize the feature importance during cross-study validation, we carried out SHAP (SHapley Additive exPlanations) analysis, a game-theory-based AI visualization method, on both individual features and grouped features, by taking advantage of the addictive nature of Shapley values (S. M. Lundberg et al., 2020; Shapley, 1983). The SHAP analysis is carried out and plots are generated by using the Python shap package (S. M. Lundberg et al., 2018).

Statistical quantification and evaluation metrics

The model's performances, as well as the replicability of drug response measurements, are evaluated by Pearson's correlation coefficient (r). Pearson's correlation coefficient is defined by equation (4):

$$r = \frac{\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum \left(x_i - \bar{x}\right)^2 \sum \left(y_i - \bar{y}\right)^2}} \quad \ldots\ldots \text{Eq(4)}$$

Where x is the gold standard and y is the prediction value when evaluating the machine learning model performances. When evaluating the intra- and inter-study experimental replicability, we selected all possible paired permutations from the replicate experiments with the same treatment-cell line combinations and computed the Pearson's r between the two replicates in these permutations. This step demonstrates the variability of experiments and provides a reference for the upper bound for the machine learning model prediction.

As the distribution of each dataset deviates significantly we didn't use RMSE as the main evaluation metric in this study. Since RMSE can be significantly decreased by approaching the average values of all responses, but not as sensitive by distinguishing higher and lower responses in the test dataset. Thus, the models failed to generate meaningful predictions to differentiate combination experiments with different responses that can have lower RMSE. This drawback can be overcome by using a relativity-based metric, such as Pearson's correlation coefficient, instead.

The confidence of evaluation metrics, of the 95% confidence interval, is generated by bootstrapping the predictions from the total datasets. We randomly sampled the prediction results from the test set without replacement 100 times to generate the 95% confidence interval.

Since all models were tested in different training and testing dataset combinations, to evaluate the consistency of model performances in the intra- and inter-study cross-validation, we carried out two-sided paired t-tests to evaluate the significance of differences between each pair of models. The fold-change (FC) and significance of the p-value were used to show the magnitude of differences between the two models.

Code Availability

The source code of the analysis and models are available on GitHub: https://github.com/GuanLab/DrugComb-cross-study-prediction

Figures

Figure 2.1. Overview of the framework on intra- and inter-study drug combination predictions. a, The cross-validation strategy. We carry out the cross-validation in two steps: intra-study, which is five-fold cross-validation carried out within a single dataset, where the training and test sets are split by drug combination and cell lines, and inter-study, which is carried out between different datasets. The models used in the 1 vs. 1 inter-study **cross-validation** are the models generated from the inter-study training step. For the 3 vs. 1 inter-study cross-validation, three of the four datasets are combined and used as the training set to generate five models by five-fold cross-validation and then tested on the remaining dataset. **b**, The overlapped information (drug, cell line, and treatment-cell line combination) between the four datasets used in this study. **c**, The schematic of model construction in this study. We use four different data sources to generate the machine learning model used in this study. For drug-related features, we used chemical structure, monotherapy efficacy score, and their corresponding dose-response relationship. For the treated cancer cell lines, we used the transcription levels of 293 cancer-related genes. The constructed features are input into a lightGBM learner to generate models predicting the six different response metrics of the combination treatment: CSS, which is the sensitivity score representing the efficacy of the combination, and five synergy scores (S, Bliss, HSA, Loewe, and ZIP) representing the degree of interaction between the two drugs.

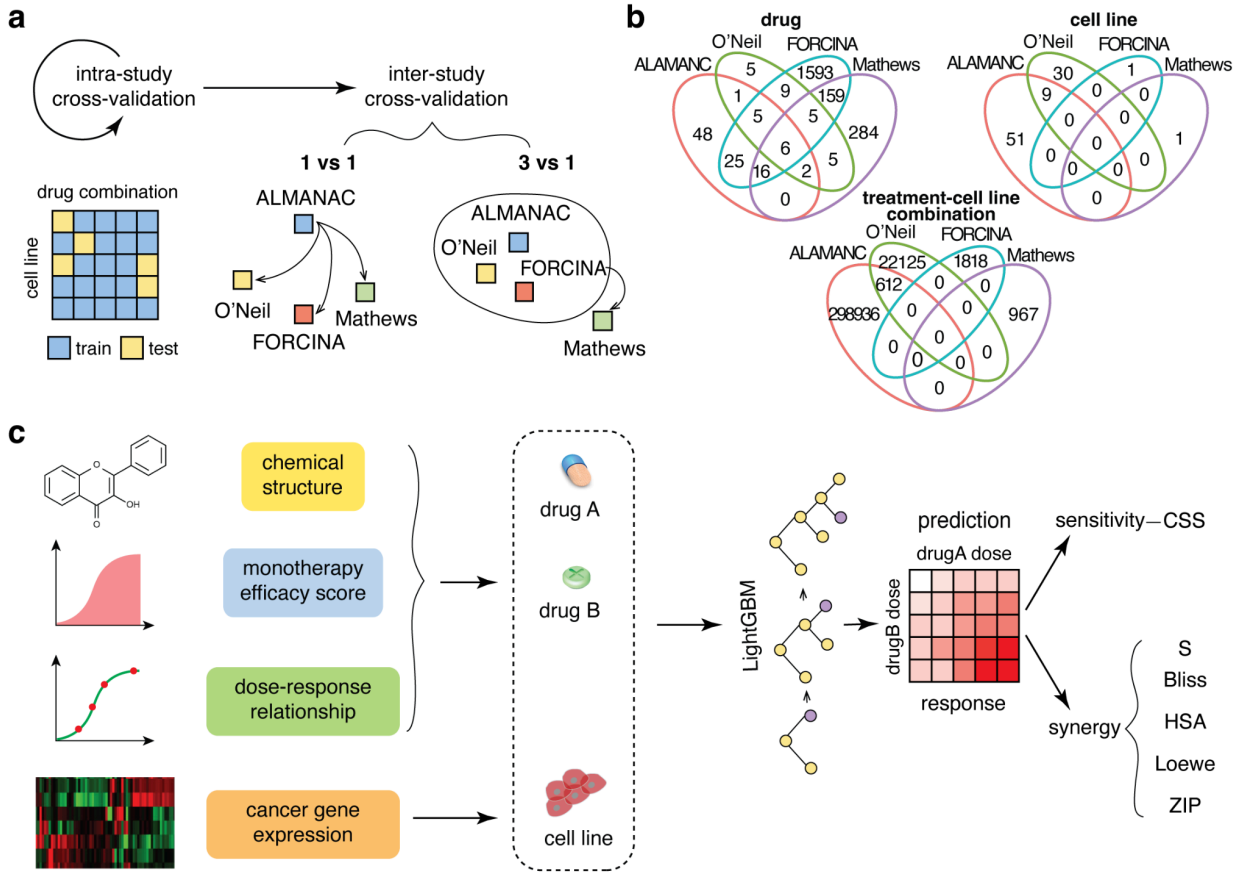


Figure 2.2. Strategy to normalize the differences in inter-study experimental settings. a, Demonstration of different dose-response curves (drc) feature construction schemes. The drc baseline feature is defined as the direct concatenation of doses and corresponding responses, where the total number of doses and responses will be padded by "-1" to the same length for different experimental settings. The drc imputation feature is the concatenation of imputed responses by different interpolation methods (see **Methods** for details). The monotherapy efficacy feature is the IC_{50} and RI of both drugs on the same cell line. **b**, Schematics of inter-study interpolation normalization in experimental settings. For experimental settings A, B, and C, which are tested using a different total number of doses, N_1, N_2, and N_3, we pull out the largest number of doses across all the studies, denoted as N_{max}. Then, the dose-response information of each setting is interpolated to the same size as N_{max}. **c**, Performances are evaluated by Pearson's r for all models, which are models with different combinations of the three features. The top performance in each training set (top) and testing set (right) is denoted by "*". **d**, Heatmap shows the results from paired t-test between the performances of five models in intra- and inter-study cross-validation. The color in the heatmap shows the fold change (FCs) of the average performances between each model pair.

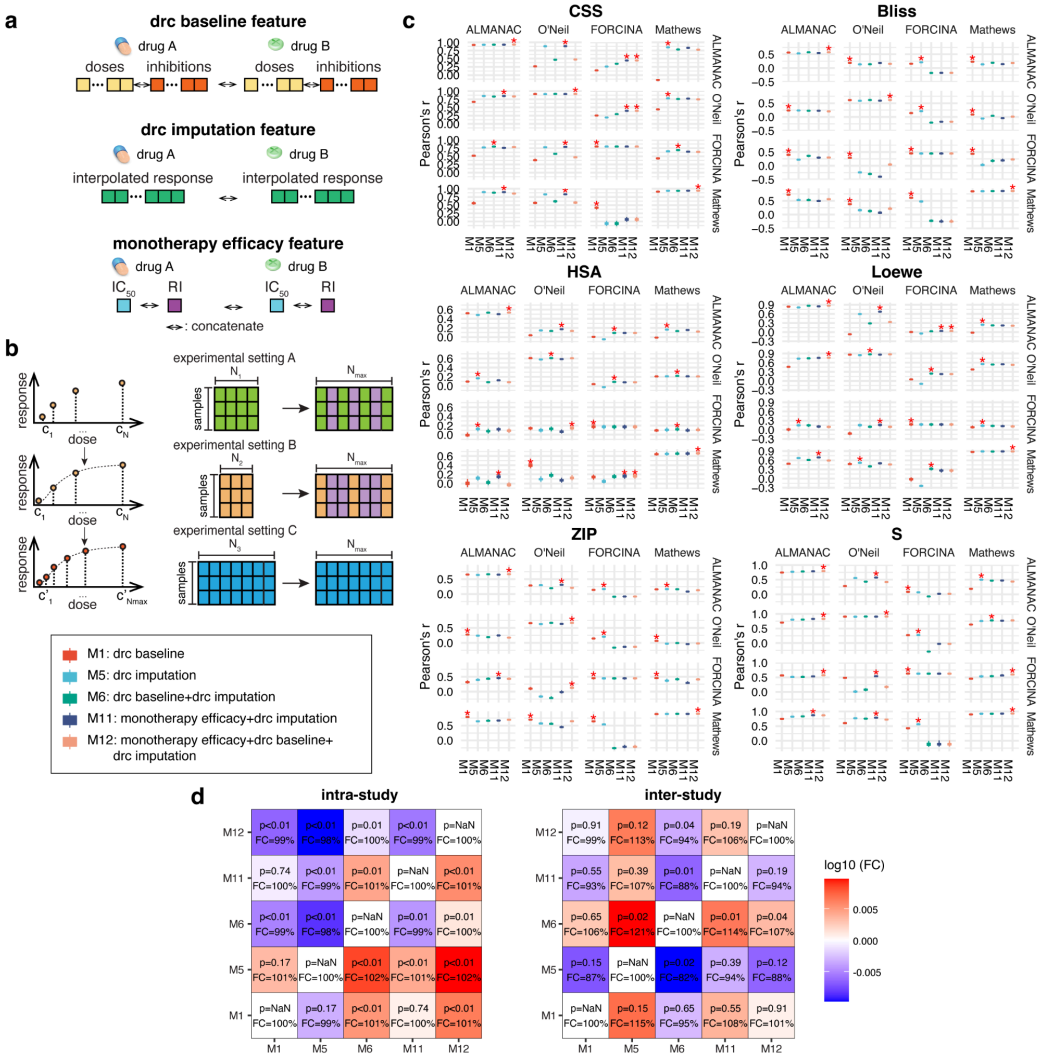


Figure 2.3. Normalized dose-response information improves the intra- and inter-study prediction performances of benchmark models. a, Schematic of the step-by-step feature construction strategy from the benchmark models (M13-M15) to the dose-response-curve-incorporated models (M16 and M20). **b**, Performances in all the training and testing scenarios for M1-M5. The best-performing models were denoted by "*" **c**, Comparisons of performances of M1-M5 from paired t-test. The fold changes (FCs) between model pairs are shown in different colors.

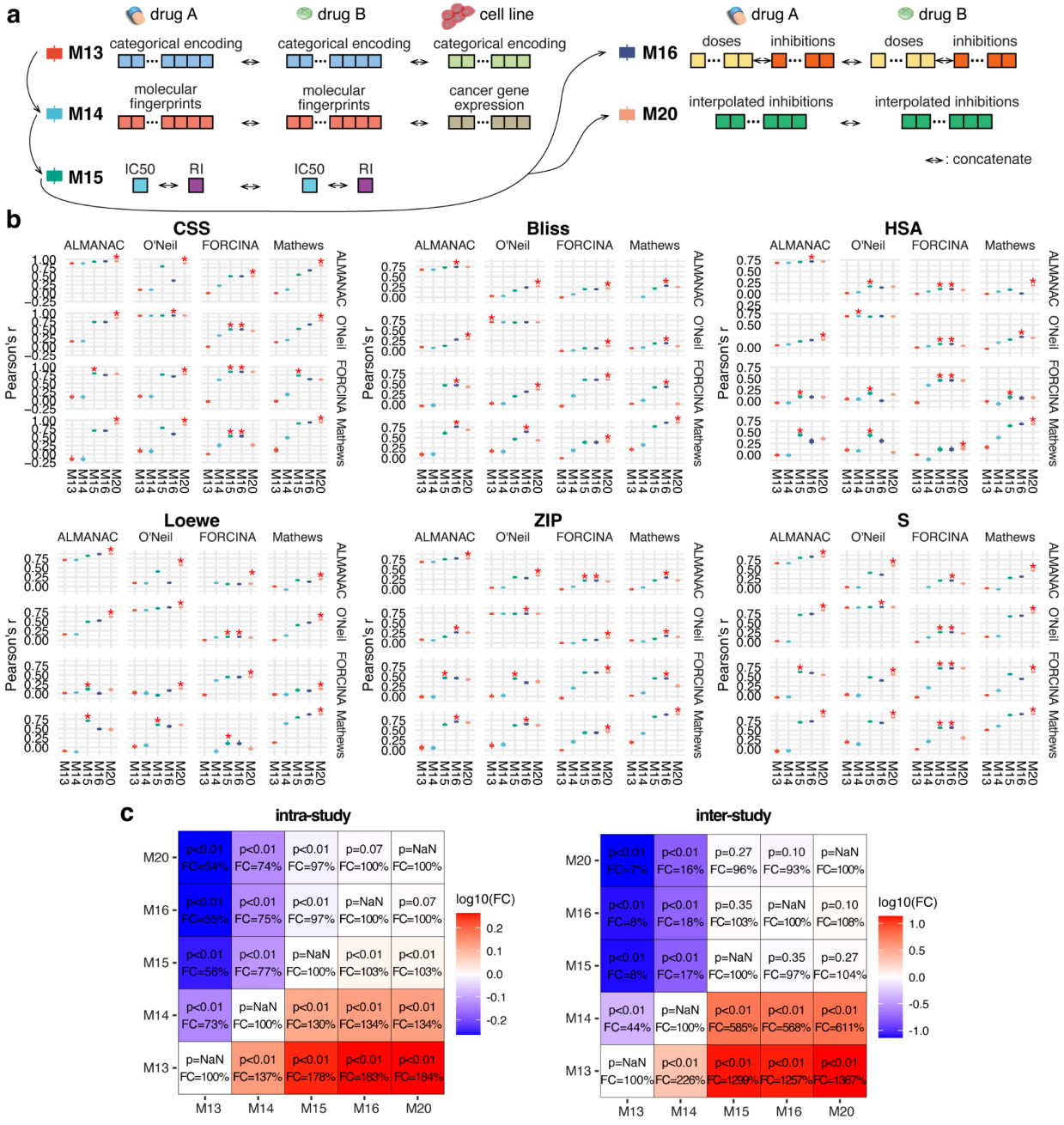
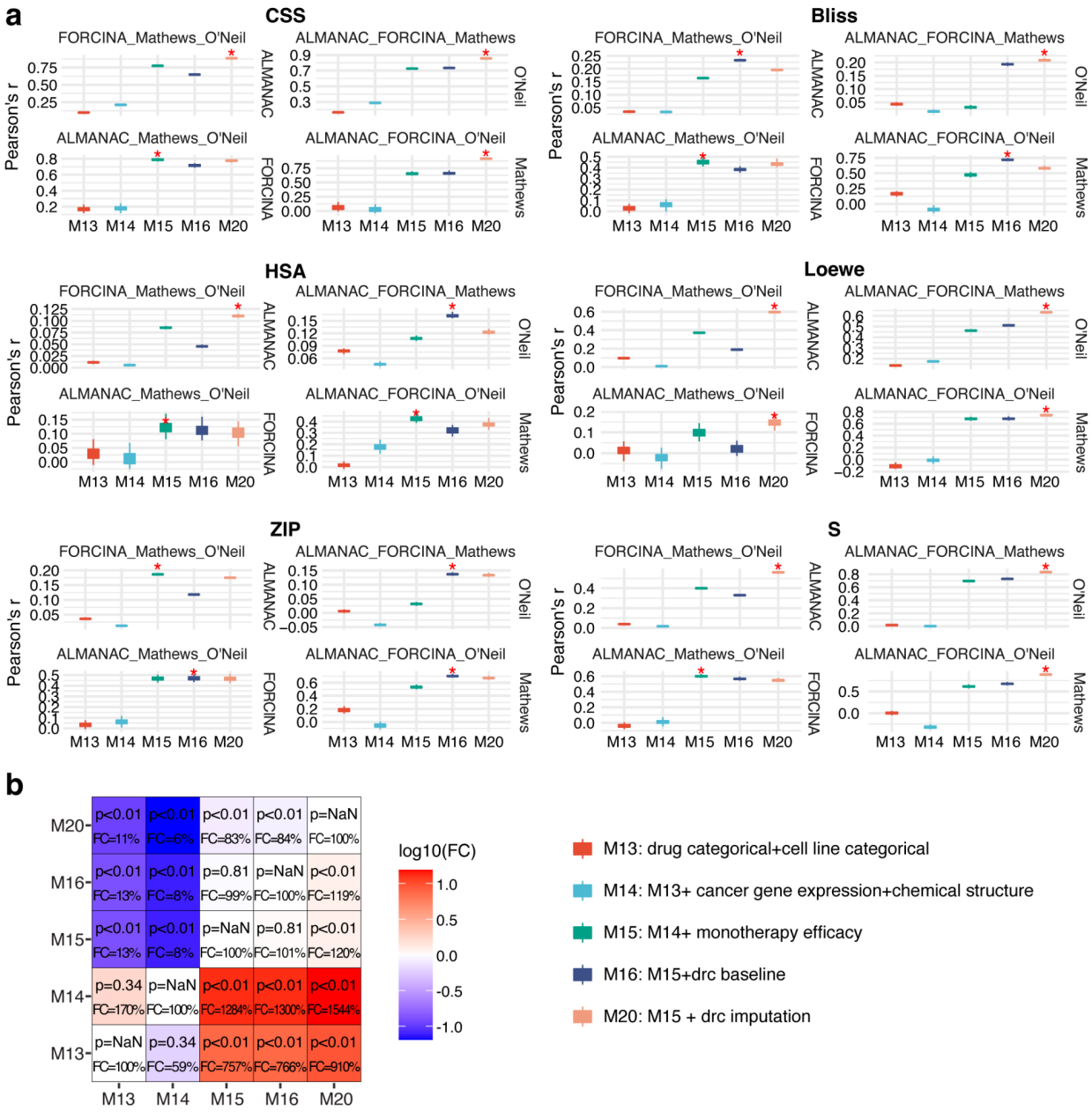


Figure 2.4. Comparison of performances before and after incorporating dose-response curve into the baseline model in inter-study predictions. Models were trained using three datasets and then tested on the remaining dataset. The models refer to the same model definition in **Figure 2.3. a**, Performances of machine learning models in the 3 vs 1 training-testing setting. For each comparison, the training set includes three studies shown on the top, while the test set contains one study shown on the right. Top performances are marked by "*". **b**, The Pairwise comparison of performances of five models, showing the fold-changes (FCs) and their p-values (paired t-test).



Supplementary Tables

Supplementary Table 2.1. Summary basic information of all datasets used in this study obtained from DrugComb. The numbers of total experiments, monotherapy drugs, drug pair combinations, cell lines, cell line-treatment combinations, and experimental settings (dose-response matrices and dose ranges) are shown in the table below.

Study	#experiment	# drug	#drug combination	#cell line	#cell line/-treatment combination	dose-response matrix	dose range (μm)
ALMANAC	311,604	103	5142	60	299,548	4 × 4 or 4 × 6	0~250
O'Neil	92,208	38	583	39	22,737	5 × 5	0~20
FORCINA	1,818	1818	1818	1	1,818	2 × 2	0~400
Mathews	1,119	477	967	1	967	6 × 6 or 10 × 10	0~1000

Supplementary Table 2.2. Performances (Pearson's r) of all models tested in this study in intra-study cross-validation. The best performances for each dataset, each score, were marked as red. The features for all models are listed in **Supplementary Table 2.5.**

ALMANAC

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.7453	0.7396	0.7371	0.7222	0.772	0.7816	0.2574	0.4743	0.6048	0.771	0.7879	0.7949	0.6615	0.6608	0.8051	0.8353	0.8352	0.8357	0.8326	0.8413
ZIP	0.6527	0.6243	0.623	0.6016	0.6402	0.6627	0.1733	0.5051	0.5687	0.6413	0.6485	0.665	0.6833	0.6847	0.7522	0.7741	0.7751	0.776	0.7654	0.7752
Loewe	0.8369	0.6387	0.6377	0.7597	0.8095	0.8659	0.2718	0.4918	0.6943	0.8142	0.8463	0.8698	0.7085	0.709	0.8205	0.8651	0.8615	0.8615	0.8521	0.8753
HSA	0.5213	0.4519	0.4583	0.4154	0.4835	0.5326	0.1421	0.2727	0.4003	0.4905	0.5015	0.535	0.6872	0.6889	0.7075	0.7194	0.7186	0.7194	0.7046	0.7183
Bliss	0.5643	0.5071	0.5087	0.4728	0.5339	0.5737	0.1296	0.334	0.4459	0.5367	0.548	0.5758	0.6707	0.672	0.7256	0.7418	0.7415	0.7426	0.7297	0.7401
CSS	0.9205	0.9181	0.9177	0.9112	0.9268	0.9301	0.2712	0.8457	0.8756	0.9234	0.9285	0.9306	0.8813	0.8815	0.9314	0.9436	0.9435	0.9436	0.9422	0.9463

O'Neil

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.9063	0.9018	0.8995	0.8892	0.9035	0.9087	0.5895	0.7756	0.8673	0.9047	0.9049	0.9091	0.9069	0.9071	0.9202	0.9235	0.9233	0.9233	0.9213	0.9218
ZIP	0.6608	0.6175	0.6165	0.5956	0.6307	0.6677	0.2478	0.3342	0.5654	0.6257	0.6347	0.6677	0.7297	0.7298	0.728	0.7343	0.7361	0.7322	0.7275	0.7318
Loewe	0.8618	0.7964	0.802	0.8271	0.8464	0.8727	0.4022	0.5958	0.7663	0.8489	0.863	0.8721	0.7988	0.7996	0.8489	0.8765	0.8752	0.8717	0.8748	0.8794
HSA	0.606	0.5655	0.5685	0.5442	0.577	0.6075	0.2478	0.3113	0.5224	0.5757	0.5851	0.6041	0.6916	0.6956	0.684	0.6942	0.6938	0.6849	0.688	0.6881
Bliss	0.6125	0.5745	0.5723	0.556	0.585	0.6134	0.2352	0.3191	0.5287	0.5864	0.5906	0.6159	0.703	0.7009	0.6922	0.6997	0.697	0.6985	0.698	0.7
CSS	0.9109	0.9055	0.9053	0.8987	0.9089	0.9128	0.5115	0.7838	0.8755	0.908	0.9099	0.9131	0.9193	0.9194	0.923	0.9263	0.9257	0.9257	0.9254	0.9258

FORCINA

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.6301	0.6272	0.6272	0.6221	0.6274	0.6274	−0.0428	0.6297	0.6289	0.6267	0.6268	0.6268	−0.0432	0.2074	0.7218	0.7218	0.7185	0.7185	0.7214	0.7208
ZIP	0.4503	0.4462	0.4462	0.4384	0.4461	0.4461	−0.0395	0.4504	0.4506	0.4462	0.4467	0.4467	−0.038	0.2166	0.6042	0.6042	0.6115	0.6115	0.6027	0.606
Loewe	0.178	0.1673	0.1673	0.1594	0.1742	0.1742	−0.0347	0.1769	0.1782	0.1677	0.1742	0.1742	−0.0329	0.3603	0.4502	0.4502	0.4527	0.4527	0.4518	0.4563
HSA	0.1773	0.1675	0.1675	0.1582	0.174	0.174	−0.0347	0.1778	0.1768	0.167	0.174	0.174	−0.0329	0.3559	0.4637	0.4637	0.4569	0.4569	0.4604	0.4611
Bliss	0.4503	0.4462	0.4462	0.4384	0.4461	0.4461	−0.0395	0.4504	0.4506	0.4462	0.4467	0.4467	−0.038	0.2166	0.6042	0.6042	0.6115	0.6115	0.6027	0.606
CSS	0.7999	0.7986	0.7986	0.7956	0.7988	0.7988	−0.031	0.7999	0.8	0.7987	0.7987	0.7987	−0.0313	0.6104	0.8478	0.8478	0.8459	0.8459	0.8471	0.8459

Mathews

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.9018	0.8938	0.8066	0.8712	0.9235	0.9234	0.64	0.818	0.8934	0.9261	0.9283	0.9294	0.5083	0.6129	0.9028	0.927	0.9284	0.923	0.92	0.9287
ZIP	0.8858	0.8842	0.8551	0.8237	0.8894	0.8931	0.3739	0.7934	0.8302	0.891	0.8909	0.8946	0.1955	0.4186	0.8376	0.8933	0.9018	0.8883	0.8848	0.9022
Loewe	0.8753	0.8829	0.853	0.8281	0.8891	0.8876	0.2481	0.6977	0.7584	0.8909	0.8901	0.8901	0.1307	0.6391	0.8004	0.8799	0.8888	0.8721	0.8473	0.8882
HSA	0.6383	0.6464	0.5585	0.5008	0.646	0.6517	0.1429	0.4661	0.5703	0.6566	0.6515	0.6591	0.1667	0.3793	0.6403	0.682	0.6939	0.6528	0.6325	0.6826
Bliss	0.8434	0.845	0.8138	0.7472	0.8456	0.8495	0.3008	0.7379	0.7685	0.8491	0.8525	0.8532	0.2109	0.3207	0.7731	0.8646	0.8613	0.8557	0.8213	0.865
CSS	0.911	0.9158	0.8846	0.9088	0.9366	0.9358	0.4411	0.8261	0.8987	0.9378	0.9393	0.9397	0.1171	0.4796	0.9066	0.9372	0.9422	0.9375	0.9318	0.9412

FORCINA_Mathews_O'Neil

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.9024	0.8986	0.8957	0.8862	0.9015	0.9058	0.5722	0.7765	0.8651	0.9017	0.9024	0.9067	0.8968	0.8752	0.915	0.9195	0.9192	0.919	0.917	0.9193
ZIP	0.6652	0.6295	0.6296	0.6089	0.6401	0.6723	0.2918	0.4135	0.5824	0.64	0.6462	0.6724	0.7034	0.6533	0.7151	0.7247	0.7259	0.7262	0.7209	0.7273
Loewe	0.8574	0.795	0.7983	0.8223	0.8427	0.8692	0.4107	0.602	0.7595	0.8456	0.8584	0.8685	0.7888	0.7746	0.8408	0.8671	0.8684	0.8646	0.8659	0.8739
HSA	0.5876	0.5531	0.5574	0.53	0.5688	0.5957	0.255	0.3112	0.5107	0.5644	0.5745	0.5954	0.6697	0.6418	0.6578	0.6594	0.6657	0.6656	0.6694	0.6651
Bliss	0.6177	0.5868	0.5919	0.5681	0.6016	0.627	0.2654	0.3729	0.5449	0.5957	0.6058	0.6253	0.6832	0.6444	0.6819	0.6908	0.6902	0.6939	0.6903	0.6936
CSS	0.9086	0.9039	0.9038	0.898	0.9079	0.9116	0.5048	0.7875	0.8734	0.9069	0.9086	0.9121	0.9037	0.8877	0.9191	0.9231	0.9232	0.9229	0.9228	0.9234

ALMANAC_FORCINA_Mathews

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.7464	0.7397	0.737	0.7234	0.7729	0.7824	0.2573	0.4795	0.6062	0.7711	0.7887	0.7952	0.6584	0.6406	0.8032	0.8345	0.8337	0.8352	0.8308	0.8399
ZIP	0.6646	0.638	0.6383	0.6189	0.6524	0.6728	0.2344	0.5212	0.5838	0.6525	0.6594	0.6737	0.6809	0.647	0.7466	0.7697	0.7716	0.7713	0.7633	0.7717
Loewe	0.8353	0.6421	0.6419	0.7593	0.8092	0.8648	0.2859	0.4936	0.6958	0.813	0.8458	0.8685	0.7088	0.7006	0.8179	0.8632	0.8605	0.8593	0.85	0.8735
HSA	0.5222	0.4578	0.461	0.4247	0.4869	0.5344	0.1835	0.2763	0.4076	0.4931	0.5039	0.5356	0.6807	0.6614	0.691	0.7115	0.7107	0.7093	0.6955	0.7094
Bliss	0.579	0.5259	0.5272	0.4962	0.5491	0.5877	0.1829	0.3619	0.4658	0.5504	0.5606	0.5884	0.6655	0.6398	0.7212	0.7382	0.737	0.7365	0.7268	0.74
CSS	0.9196	0.9173	0.9168	0.9103	0.9261	0.9296	0.2732	0.8446	0.8743	0.9226	0.9278	0.9299	0.8772	0.8686	0.9303	0.9424	0.9423	0.9427	0.9413	0.9452

ALMANAC_Mathews_O'Neil

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.8085	0.8026	0.7983	0.7796	0.8257	0.8335	0.3357	0.5999	0.6928	0.824	0.8357	0.8414	0.761	0.76	0.849	0.8692	0.8699	0.8698	0.8667	0.8731
ZIP	0.658	0.6211	0.6203	0.5731	0.6353	0.6646	0.1787	0.4098	0.5181	0.6317	0.6397	0.6646	0.698	0.6966	0.7461	0.7619	0.7616	0.7611	0.7517	0.762
Loewe	0.8292	0.6645	0.6673	0.7621	0.8055	0.8583	0.2689	0.5139	0.6861	0.8077	0.8392	0.8617	0.7201	0.7189	0.8168	0.8508	0.8529	0.8512	0.8461	0.8654
HSA	0.5809	0.533	0.5309	0.4849	0.5451	0.5868	0.2012	0.3248	0.4482	0.5494	0.5567	0.5887	0.7014	0.7018	0.717	0.7262	0.7191	0.721	0.7121	0.7189
Bliss	0.5613	0.5109	0.51	0.4578	0.5276	0.5686	0.1321	0.2688	0.4111	0.5267	0.5366	0.5681	0.6626	0.6636	0.6995	0.7141	0.712	0.7121	0.7044	0.7138
CSS	0.9326	0.931	0.9303	0.9232	0.9375	0.9405	0.3473	0.8624	0.89	0.9344	0.9384	0.9407	0.9063	0.9053	0.9419	0.9505	0.9502	0.9503	0.9494	0.9525

ALMANAC_FORCINA_O'Neil

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.8067	0.8016	0.7966	0.7787	0.8241	0.8318	0.3393	0.5995	0.6929	0.8224	0.8339	0.8399	0.759	0.7561	0.8471	0.8678	0.8678	0.8681	0.8653	0.8713
ZIP	0.6545	0.6176	0.6176	0.5734	0.6321	0.6607	0.2133	0.4152	0.524	0.627	0.637	0.6616	0.6898	0.6817	0.7377	0.7537	0.7541	0.754	0.7455	0.7551
Loewe	0.829	0.6653	0.668	0.7625	0.806	0.8579	0.2794	0.5128	0.6872	0.807	0.8382	0.8614	0.7211	0.7203	0.8155	0.851	0.8525	0.85	0.8462	0.864
HSA	0.5805	0.5322	0.5315	0.4856	0.5485	0.5871	0.216	0.3287	0.4547	0.5492	0.5577	0.5888	0.697	0.6947	0.7112	0.7227	0.7174	0.7171	0.7095	0.7143
Bliss	0.5589	0.5108	0.512	0.4614	0.5309	0.5698	0.1656	0.2819	0.4223	0.5255	0.5388	0.5711	0.658	0.6538	0.6969	0.7098	0.7067	0.7091	0.7008	0.7105
CSS	0.9322	0.9303	0.9297	0.9228	0.9371	0.9401	0.347	0.8628	0.8903	0.9339	0.938	0.9403	0.9052	0.9031	0.9416	0.9499	0.9497	0.9496	0.949	0.952

Supplementary Table 2.3. Performances (Pearson's r) of all models tested in this study in 1 vs. 1 inter-study cross-validation. The best performances for each train-test setting, each score, were marked as red. The features for all models are listed in **Supplementary Table 5.**

Train

Test

ALMANAC → O'Neil

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.7034	0.7721	0.6561	0.8102	0.803	0.8052	0.3277	0.7297	0.7334	0.8142	0.8284	0.8284	0.023	0.0122	0.7138	0.7383	0.8095	0.7806	0.8273	0.8349
ZIP	0.2765	0.2574	0.1867	0.2376	0.2407	0.1915	−0.0545	0.147	0.1125	0.2608	0.2492	0.1734	0.0828	0.064	0.1558	0.2653	0.2684	0.2244	0.2343	0.2591
Loewe	0.4629	0.5142	0.4371	0.7459	0.7348	0.7424	0.1542	0.5438	0.5941	0.7058	0.7558	0.7598	0.1513	0.1555	0.488	0.5218	0.5853	0.5722	0.634	0.6281
HSA	0.0939	0.1608	0.0579	0.1925	0.1664	0.0821	−0.0145	0.1875	0.0745	0.1579	0.1235	0.081	0.0472	0.0696	0.1348	0.165	0.154	0.1516	0.1553	0.1765
Bliss	0.2352	0.2428	0.1973	0.2365	0.2322	0.2223	−0.0299	0.1933	0.126	0.2349	0.2228	0.2011	0.0922	0.0775	0.1257	0.2785	0.2836	0.2237	0.2251	0.2846
CSS	0.6725	0.8384	0.8138	0.8369	0.8509	0.8359	0.2542	0.7424	0.7487	0.8435	0.8511	0.8392	0.1631	0.1679	0.7336	0.7324	0.8404	0.8263	0.8404	0.8579

ALMANAC → FORCINA

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.566	0.1449	0.2316	0.5254	0.5129	0.5147	0	0.5229	0.6063	0.5541	0.5311	0.5711	−0.0506	−0.0345	0.6251	0.597	0.6103	0.566	0.6161	0.5475
ZIP	0.3164	0.3992	0.3887	0.3862	0.3933	0.4335		0.4056	0.4138	0.4494	0.4536	0.432	−0.01	−0.0084	0.4671	0.4625	0.4211	0.4499	0.4558	0.4285
Loewe	0.0069	0.1689	0.1588	0.1122	0.1477	0.146	0	0.0929	0.1006	0.0534	0.1042	0.0948	0.0187	0.0309	0.1209	0.006	0.0663	0.0957	0.1025	0.1102
HSA	−0.0074	0.1126	0.1274	0.0932	0.1216	0.0752	0	0.1013	0.0807	0.0735	0.1178	0.0883	−0.0381	−0.0399	0.095	0.0893	0.0873	0.0896	0.069	0.0836
Bliss	0.3973	0.4588	0.4455	−0.0582	0.2205	0.3562		0.2261	0.3224	0.391	0.2932	0.2962	−0.0349	−0.0163	0.4688	0.4715	0.4317	0.45	0.4617	0.4322
CSS	0.5171	0.7939	0.7735	0.7621	0.7689	0.7975	0	0.7405	0.7799	0.7412	0.7562	0.7807	0.1001	0.0989	0.8014	0.7517	0.7741	0.7657	0.797	0.7865

ALMANAC → Mathews

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.7485	0.6928	0.4333	0.799	0.8216	0.8331	0.4336	0.6494	0.6984	0.8467	0.8719	0.8686	−0.0602	−0.0464	0.7107	0.735	0.8381	0.792	0.8336	0.87
ZIP	0.785	0.6645	0.4054	0.69	0.641	0.6616	0.0226	0.6303	0.6154	0.6358	0.6655	0.6935	0.0649	0.0609	0.6403	0.7141	0.7287	0.6757	0.7123	0.6901
Loewe	0.4872	0.6994	0.7178	0.5646	0.6743	0.6107	0.0967	0.6927	0.6249	0.5335	0.6948	0.5932	−0.1017	−0.1235	0.7101	0.4885	0.402	0.7239	0.7478	0.4776
HSA	−0.0022	−0.1322	0.0548	−0.0217	0.1161	0.0202	0.0791	0.3745	0.3582	−0.1061	0.1448	−0.0309	−0.0044	−0.007	0.436	0.2965	0.4293	0.3481	0.3415	0.3531
Bliss	0.7186	0.694	0.3346	0.5691	0.5263	0.5261	−0.0295	0.5163	0.5128	0.6707	0.4889	0.5551	0.091	0.0978	0.6089	0.7671	0.7512	0.6648	0.7129	0.6898
CSS	0.5578	0.7878	0.7264	0.7248	0.8969	0.878	0.0485	0.7104	0.7272	0.8479	0.8974	0.8558	−0.1436	−0.1465	0.695	0.6935	0.8731	0.8433	0.8177	0.9053

O'Neil → ALMANAC

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.2818	0.4659	0.3928	0.5591	0.5575	0.4333	0.1091	0.3753	0.3666	0.5107	0.5728	0.4202	0.0317	0.0258	0.4076	0.3597	0.4984	0.5152	0.5904	0.6003
ZIP	0.2869	0.3427	0.2677	0.3243	0.2967	0.1941	−0.0202	0.3517	0.323	0.3186	0.3077	0.209	0.0379	0.0473	0.3036	0.2811	0.323	0.3131	0.318	0.3486
Loewe	−0.0687	0.5025	0.4338	0.6247	0.5944	0.2871	0.056	0.4254	0.524	0.6666	0.6709	0.3332	0.0942	0.0871	0.3989	0.0978	0.5362	0.5168	0.582	0.5723
HSA	0.0394	0.2123	0.1179	0.1627	0.1459	0.133	0.0266	0.1769	0.1432	0.1968	0.1713	0.134	0.0159	0.0347	0.1661	0.1403	0.169	0.1425	0.1577	0.1594
Bliss	0.1913	0.2099	0.0316	0.1815	0.1349	0.141	−0.0116	0.2208	0.1977	0.203	0.187	0.1553	0.032	0.0338	0.1635	0.2402	0.2405	0.2296	0.1983	0.2694
CSS	0.2676	0.8812	0.8577	0.8669	0.8802	0.4692	0.1286	0.8146	0.8088	0.8823	0.8839	0.476	0.1005	0.0995	0.795	0.3719	0.8779	0.8707	0.8763	0.886

O'Neil → FORCINA

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.4837	−0.4228	−0.4979	0.5472	0.0088	0.0775		0.4379	0.5287	−0.4143	0.5367	0.1703	0.0278	0.0088	0.4807	0.3998	−0.4137	0.5405	0.5839	0.5575
ZIP	0.1022	0.1743	0.1699	0.3747	−0.1659	−0.2152	0	0.4407	0.4168	0.1891	−0.0159	0.1291	0.0136	0.0278	0.451	0.3453	0.4112	0.442	0.4532	0.3745
Loewe	−0.121	0.0117	0.1483	0.1548	0.1484	0.098	0	0.1066	0.1097	0.1335	0.1562	0.0972	0.0233	0.0124	−0.0379	0.0824	0.0661	0.0436	0.144	0.1403
HSA	0.1388	0.0766	0.085	0.1102	0.0947	0.1254	0	0.1412	0.1417	0.0663	0.0642	0.1421	0.0452	0.0432	0.1709	0.0088	0.1832	0.1144	0.1461	0.1464
Bliss	0.2867	−0.2602	0.1662	−0.0842	−0.2438	−0.3018	0	0.3777	0.4316	−0.3174	−0.3978	0.041	0.0306	0.0273	0.2048	0.3127	0.0517	0.3697	0.3098	0.3667
CSS	0.3831	0.7948	0.783	0.7711	0.7766	0.5763	0	0.7387	0.7913	0.7696	0.7869	0.4692	0.1197	0.1178	0.7664	0.7117	0.8036	0.7961	0.7653	0.7775

O'Neil → Mathews

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.6031	0.7899	0.1961	0.7848	0.7689	0.7555	0.3114	0.6133	0.5672	0.8168	0.7901	0.7259	0.1908	0.1396	0.7284	0.6964	0.7953	0.704	0.8212	0.8399
ZIP	0.6959	0.5616	0.2523	0.5702	0.5682	0.5567	0.0371	0.5718	0.5825	0.6597	0.4472	0.643	0.1182	0.1428	0.6232	0.6518	0.6222	0.4758	0.4892	0.624
Loewe	0.4762	0.6692	0.4325	0.5971	0.5172	0.3994	0.0238	0.6188	0.5896	0.7454	0.5017	0.4627	0.0222	0.0542	0.6042	0.5581	0.7624	0.6348	0.5172	0.6029
HSA	0.389	0.0596	0.0783	0.0481	0.0937	0.1741	−0.0274	0.327	0.3371	0.1639	0.0726	0.119	0.1032	0.0863	0.4274	0.3044	0.2709	0.1126	0.1515	0.0529
Bliss	0.3746	0.4699	−0.105	0.2949	0.1568	0.1243	0.0711	0.4261	0.4304	0.4973	0.0637	0.2113	0.1783	0.1657	0.4663	0.6427	0.5203	0.3271	0.2183	0.4367
CSS	0.5677	0.8604	0.6666	0.7261	0.8281	0.6171	−0.0219	0.6921	0.687	0.8586	0.8308	0.5784	0.0887	0.0798	0.7766	0.5911	0.8826	0.8111	0.8329	0.8731

FORCINA → ALMANAC

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.0846	0.0759	0.0574	0.2092	0.0685	−0.0736	0.0465	0.142	0.1444	0.0396	0.0198	0.0198		0.0347	0.2041	0.2041	0.1183	0.1001	0.1584	0.1268
ZIP	0.1377	0.1547	0.1499	0.1283	0.1678	−0.0912	−0.0326	0.182	0.1835	−0.0673	−0.0832	−0.0832	0	0.0648	0.2176	0.2176	0.2061	0.2002	0.2024	0.1976
Loewe	0.0051	−0.0146	−0.0027	0.1078	−0.0421	0.0275	0.0243	−0.0609	−0.0587	−0.005	0.038	0.038		0.0952	0.0646	0.0646	0.0625	0.0626	0.0617	0.0741
HSA	0.0064	−0.0082	−0.0248	−0.0781	−0.0477	0.091	−0.0224	0.0329	0.0344	0.0851	0.0899	0.0899	0	0.0524	0.1065	0.1065	0.0917	0.0917	0.1061	0.0847
Bliss	0.1582	0.1996	0.1938	0.138	0.2131	−0.1747	−0.0023	0.1101	0.1129	−0.1787	−0.1737	−0.1737	0	0.0722	0.1964	0.1964	0.2345	0.2273	0.1876	0.219
CSS	0.1412	0.253	0.2386	0.1834	0.2614	0.3449	−0.0441	0.6511	0.6501	0.4219	0.445	0.445	0	0.2224	0.5022	0.5022	0.4379	0.4296	0.4288	0.5072

FORCINA → O'Neil

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.2738	0.3086	0.2666	0.2991	0.2758	−0.2887	−0.0077	0.149	0.1527	0.0131	−0.025	−0.025	0	0.1221	0.2579	0.2579	0.2569	0.2323	0.1579	0.2262
ZIP	0.1445	0.2059	0.193	0.0647	0.2041	−0.1598	0.0011	0.0293	0.0316	−0.1264	−0.1352	−0.1352	0	−0.006	0.0756	0.0756	0.1441	0.1454	0.0669	0.1283
Loewe	0.051	−0.0445	−0.0529	−0.1323	−0.0945	0.2444	−0.0679	0.0318	0.0366	0.1554	0.2276	0.2276	0	0.0723	0.0873	0.0873	0.0634	0.0644	0.0893	0.0665
HSA	0.0382	0.011	−0.0192	−0.0544	−0.0268	0.0871	−0.0445	0.0304	0.0287	0.063	0.0791	0.0791	0	0.0267	0.0777	0.0777	0.047	0.0462	0.0682	0.0374
Bliss	0.1393	0.218	0.2006	0.1131	0.208	−0.2075	0.0087	−0.0023	0.0026	−0.1864	−0.1744	−0.1744	0	0.0126	0.0657	0.0657	0.1336	0.1335	0.0591	0.1188
CSS	0.2552	0.1802	0.1618	0.2955	0.1929	0.2968	−0.0872	0.5667	0.5657	0.3789	0.3978	0.3978	0	0.3331	0.5135	0.5135	0.406	0.395	0.4667	0.4697

FORCINA → Mathews

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.4269	0.4979	0.4813	0.5746	0.5644	−0.1097	0.0537	0.4944	0.4988	−0.1697	−0.1101	−0.1101	0	0.2066	0.5699	0.5699	0.438	0.416	0.4706	0.2989
ZIP	0.6413	0.5308	0.4581	0.5014	0.5376	−0.2575	−0.0186	0.5144	0.5104	−0.2313	−0.2108	−0.2108	0	0.2081	0.433	0.433	0.5234	0.5143	0.4584	0.4658
Loewe	−0.0224	−0.1323	−0.1122	−0.1945	−0.2389	0.3173	−0.0391	0.0052	0.0081	0.3137	0.2617	0.2617		−0.111	0.1004	0.1004	−0.0331	−0.0296	0.1104	−0.0395
HSA	0.1439	−0.0061	−0.0327	−0.1851	0.0476	0.1489	−0.0323	0.0916	0.0863	0.1334	0.1537	0.1537	0	−0.104	0.1152	0.1152	0.1028	0.0964	0.1273	0.14
Bliss	0.6072	0.4847	0.453	0.467	0.4744	−0.2234	−0.0034	0.4274	0.4236	−0.266	−0.2508	−0.2508	0	0.19	0.3837	0.3837	0.4728	0.4672	0.4035	0.4113
CSS	0.4221	−0.0593	−0.0182	0.4442	−0.067	−0.066	−0.0401	0.5689	0.5682	0.0289	0.0572	0.0572	0	0.2604	0.5333	0.5333	0.188	0.2141	0.4818	0.2635

Mathews → ALMANAC

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.1877	0.3506	0.2066	0.3965	0.5009	0.4702	0.0978	0.2555	0.2621	0.3113	0.473	0.4369	0	−0.0236	0.2772	0.3078	0.3174	0.3388	0.4401	0.4668
ZIP	0.1574	0.1747	0.1145	0.3093	0.1656	0.161	0.0352	0.2423	0.063	0.1244	0.0829	0.1277	0	0.0452	0.2223	0.2944	0.2165	0.2063	0.2797	0.2211
Loewe	−0.0175	0.271	0.2683	0.2138	0.2427	0.2234	0.0271	0.1913	0.1819	0.2394	0.2211	0.2096	0	−0.0801	0.1699	0.1286	0.2169	0.2024	0.159	0.1958
HSA	−0.0109	0.1475	0.1793	0.1041	0.1628	0.131	0.0165	0.092	0.1107	0.1454	0.1551	0.12	0	0.0498	0.094	0.0081	0.1915	0.1802	0.1558	0.1868
Bliss	0.2326	0.1545	0.1472	0.1243	0.1378	0.1935	0.0146	0.1903	0.1552	0.1646	0.1434	0.1852		0.0107	0.2134	0.2832	0.2396	0.2744	0.2336	0.2519
CSS	−0.1503	0.8333	0.8335	0.6378	0.8478	0.7817	−0.0171	0.6025	0.4511	0.8277	0.831	0.7761	0	0.107	0.5474	0.6752	0.8233	0.8229	0.7559	0.8183

Mathews → O'Neil

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.6361	0.7441	0.6788	0.715	0.7558	0.778	0.3001	0.6939	0.6715	0.7093	0.7677	0.7776	0.1356	0.1565	0.6743	0.6935	0.7164	0.7165	0.7675	0.7733
ZIP	0.0641	−0.0537	0.0518	0.0439	−0.0289	−0.0092	−0.015	0.0982	0.0663	−0.0313	−0.0409	−0.0221	0.0638	0.0355	0.1012	0.177	0.1339	0.1636	0.1236	0.1493
Loewe	0.3769	0.5527	0.564	0.5019	0.5544	0.5383	−0.0429	0.4258	0.3837	0.5496	0.5472	0.5281	0	0.1233	0.3983	0.4695	0.5552	0.5436	0.506	0.547
HSA	0.2078	0.1972	0.2268	0.1599	0.21	0.2172	−0.0367	0.1276	0.1254	0.1991	0.2086	0.2063	−0.0295	0.11	0.1768	0.2366	0.215	0.1885	0.2062	0.216
Bliss	0.0866	−0.0353	0.0656	−0.0038	−0.0568	0.0384	0.0025	0.1108	0.1301	−0.0299	−0.044	0.01	0.0686	0.0851	0.1813	0.1872	0.1502	0.2029	0.1408	0.116
CSS	0.4416	0.7645	0.792	0.6471	0.78	0.7601	−0.0867	0.6579	0.5386	0.7652	0.7734	0.7493	0.1356	0.1996	0.5285	0.6609	0.7838	0.7869	0.7216	0.7824

Mathews → FORCINA

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
S	0.4522	−0.4267	−0.3899	0.6324	0.5313	0.5394		0.4828	0.5515	0.2437	0.5638	0.5981	0.0095	0.1311	0.5647	0.4388	0.3931	0.4866	0.6323	0.626
ZIP	0.4545	0.4508	0.2807	0.4522	0.3691	0.4161	0	0.4021	0.4178	0.1387	0.2924	0.4101	0.0229	0.0885	0.4547	0.4547	0.0067	0.0942	0.4448	0.2624
Loewe	0.0824	0.1195	0.1323	0.1363	0.1304	0.1256	0	0.0876	0.0761	0.1095	0.1187	0.1091	−0.0268	−0.0221	0.0952	0.0753	0.1108	0.1162	0.0991	0.1301
HSA	0.0961	0.1193	0.1424	0.1638	0.1042	0.1153		0.1201	0.1321	0.1102	0.0935	0.1035	−0.0027	−0.0581	0.0899	0.0636	0.0745	0.1344	0.1452	0.0777
Bliss	0.4317	−0.1625	−0.2661	0.4446	0.0261	0.1767	0	0.361	0.4379	0.0164	0.1932	0.2288	0.0216	0.0989	0.4179	0.4346	−0.0026	0.0812	0.4196	0.0156
CSS	0.438	0.6462	0.624	0.6456	0.6527	0.6899		0.6422	0.7514	0.6733	0.643	0.6352	−0.0235	0.1756	0.7394	0.6291	0.6395	0.7238	0.7483	0.6154

Supplementary Table 2.4. Performances (Pearson's r) of all models tested in this study in 3 vs. 1 inter-study cross-validation. The best performances for each train-test setting, each score, were marked as red. The features for all models are listed in **Supplementary Table 5.**

Train

FORCINA_Mathews_O'Neil

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	
S	0.2997	0.4638	0.4707	0.5584	0.582	0.5668	0.1135	0.3875	0.3689	0.5389	0.6035	0.5768	0.0381	0.0167	0.4007	0.3311	0.4942	0.5055	0.5595	0.5614	ALMANAC
ZIP	0.2845	0.3556	0.343	0.3356	0.3327	0.282	−0.0114	0.3646	0.34	0.3463	0.3419	0.2625	0.0361	0.0127	0.1863	0.1182	0.2525	0.217	0.1804	0.1751	
Loewe	−0.1386	0.5231	0.4892	0.6326	0.6458	0.6623	0.049	0.4266	0.5186	0.6658	0.69	0.6756	0.0957	0.0095	0.3719	0.1885	0.5299	0.5115	0.5816	0.5959	
HSA	0.0054	0.2189	0.1633	0.1408	0.2129	0.1325	0.0223	0.1571	0.1377	0.2071	0.2113	0.1491	0.0115	0.0056	0.0853	0.0456	0.1143	0.0947	0.088	0.1104	
Bliss	0.2114	0.2316	0.2196	0.2069	0.2791	0.2791	−0.0136	0.2421	0.2157	0.2611	0.2671	0.2217	0.0344	0.0336	0.1639	0.233	0.1937	0.1872	0.1798	0.1953	
CSS	0.2465	0.8803	0.8705	0.8671	0.8918	0.8735	0.1275	0.8157	0.8007	0.882	0.8922	0.8706	0.0992	0.2102	0.7755	0.6486	0.8645	0.867	0.8659	0.8832	

ALMANAC_FORCINA_Mathews

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	
S	0.6791	0.7738	0.6255	0.8077	0.8019	0.8134	0.327	0.7333	0.7338	0.8227	0.8283	0.8297	0.0203	0.0032	0.6958	0.7284	0.8116	0.7905	0.8193	0.8328	O'Neil
ZIP	0.2367	0.2725	0.1916	0.2341	0.2341	0.2137	−0.0484	0.1563	0.1213	0.2657	0.2424	0.2273	0.0061	−0.042	0.0319	0.1366	0.1618	0.1207	0.1154	0.1333	
Loewe	0.5188	0.5448	0.5178	0.746	0.749	0.7612	0.1394	0.5406	0.5977	0.7239	0.7634	0.7758	0.136	0.1743	0.4629	0.5119	0.5741	0.5533	0.6315	0.6335	
HSA	0.1426	0.1643	0.099	0.1895	0.1863	0.1053	0.0173	0.184	0.1006	0.1446	0.1614	0.14	0.0779	0.0461	0.1081	0.1628	0.1146	0.1146	0.1143	0.123	
Bliss	0.2428	0.2309	0.1546	0.2306	0.2383	0.2336	−0.0278	0.1912	0.1346	0.2341	0.2277	0.2304	0.0434	0.0158	0.0314	0.1933	0.1764	0.1389	0.0963	0.2083	
CSS	0.6885	0.8419	0.8184	0.8374	0.8585	0.8465	0.265	0.7438	0.7484	0.843	0.854	0.8518	0.1674	0.2874	0.7256	0.734	0.8385	0.8314	0.8317	0.8555	

ALMANAC_Mathews_O'Neil

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	
S	0.579	−0.1093	−0.2428	0.512	0.5187	0.2213	0	0.5357	0.5987	0.5262	0.5473	0.5623	−0.0331	0.0124	0.6003	0.5665	0.6079	0.5707	0.593	0.5487	FORCINA
ZIP	0.4066	0.3845	0.3801	0.3553	0.3555	0.4704		0.419	0.4457	0.4359	0.4466	0.4696	0.0344	0.0583	0.4667	0.4704	0.4548	0.4535	0.4728	0.4663	
Loewe	−0.0083	0.1712	0.1622	0.1305	0.1712	0.1252	0	0.1052	0.1109	0.1134	0.1648	0.1411	0.0092	−0.0242	0.0975	0.0165	0.1115	0.1372	0.1464	0.1459	
HSA	0.0402	0.1175	0.1129	0.1121	0.1365	0.1794	0	0.1316	0.1249	0.0868	0.1254	0.0886	0.0304	0.0111	0.1188	0.1119	0.0972	0.0927	0.1154	0.1012	
Bliss	0.3919	0.4546	0.446	−0.056	0.3246	−0.1195	0	0.4584	0.4274	0.2895	0.4115	0.4482	0.0265	0.0571	0.4519	0.3835	0.4322	0.4511	0.4556	0.4328	
CSS	−0.2484	0.7893	0.78	0.7723	0.7885	0.7746	0	0.7484	0.7924	0.7296	0.7315	0.758	0.1625	0.1715	0.7882	0.716	0.7581	0.7819	0.7894	0.7752	

ALMANAC_FORCINA_O'Neil

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	
S	0.7225	0.7719	0.5404	0.8007	0.8672	0.8799	0.4347	0.6708	0.7061	0.8541	0.8669	0.8749	−1e−04	−0.3197	0.6143	0.6756	0.8471	0.8238	0.8398	0.8889	Mathews
ZIP	0.8249	0.6433	0.5666	0.691	0.6508	0.7363	0.0198	0.6351	0.6263	0.6653	0.6778	0.7479	0.183	−0.0612	0.5315	0.7018	0.6414	0.602	0.6359	0.6712	
Loewe	0.5858	0.7249	0.6705	0.614	0.7012	0.7036	0.0902	0.6662	0.6425	0.7896	0.7266	0.6665	−0.1099	−0.0093	0.6784	0.6819	0.7567	0.6882	0.7164	0.741	
HSA	0.3641	0.0354	0.0645	0.0134	0.0853	0.2315	0.0785	0.4105	0.4303	0.0068	0.0351	0.1826	0.0157	0.1717	0.4222	0.3174	0.4012	0.3077	0.2784	0.3692	
Bliss	0.7622	0.6021	0.4286	0.5778	0.5985	0.6569	0.0384	0.5265	0.5397	0.6328	0.5947	0.6599	0.1636	−0.0944	0.473	0.7193	0.5895	0.5274	0.6838	0.5797	
CSS	0.5521	0.8443	0.8114	0.7596	0.8807	0.8987	−0.0406	0.7185	0.7323	0.8716	0.886	0.9012	0.0651	0.0273	0.6546	0.6626	0.8751	0.8707	0.8155	0.917	

Test

Supplementary Table 2.5. Table legend for all models in Supplementary Table 2.2-4.

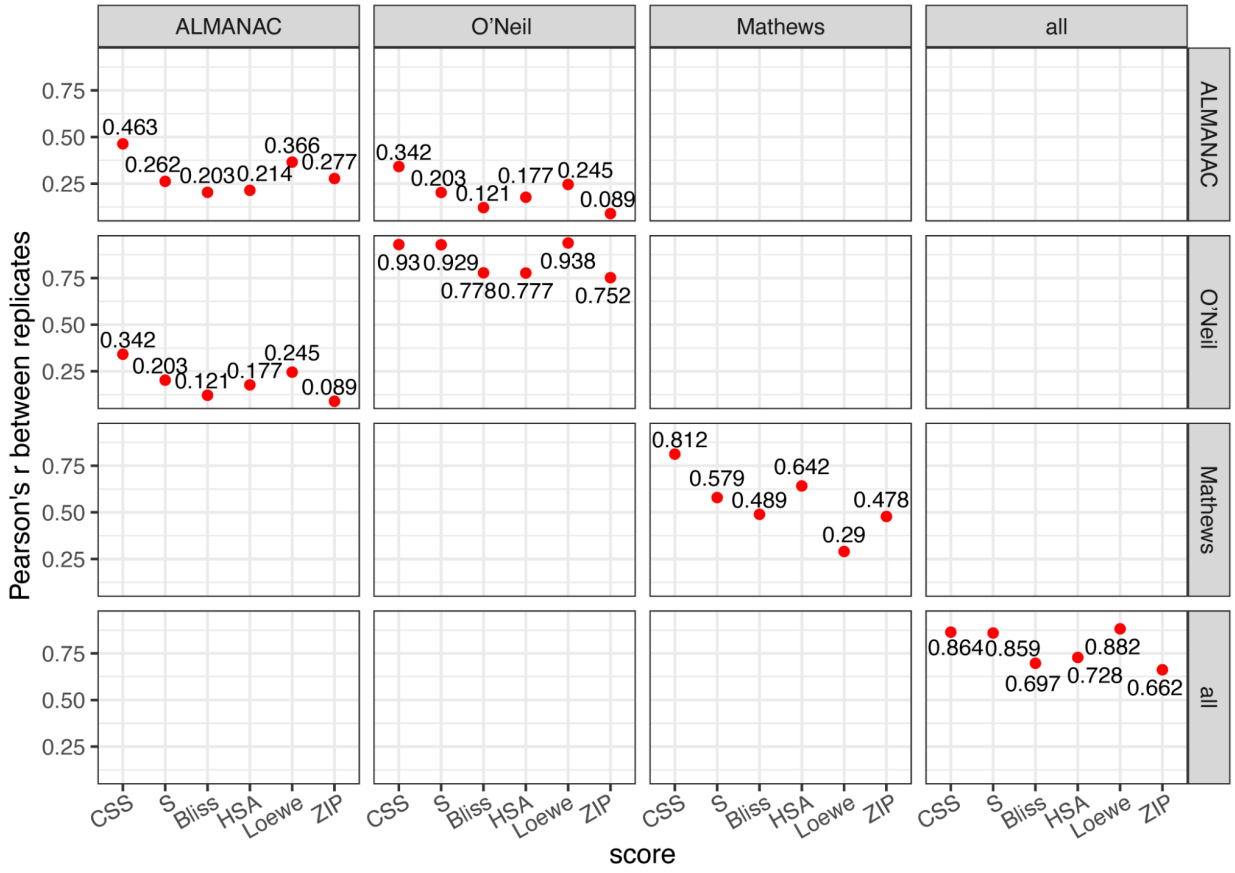
model	feature
M1	drc_baseline
M2	drc_intp_linear
M3	drc_intp_lagrange
M4	drc_intp_4PL
M5	drc_intp_linear+drc_intp_lagrange+drc_intp_4PL
M6	drc_baseline+drc_intp_linear+drc_intp_lagrange+drc_intp_4PL
M7	monotherapy_ic50
M8	monotherapy_ri
M9	monotherapy_ri+monotherapy_ic50
M10	monotherapy_ic50+monotherapy_ri+drc_intp_linear
M11	monotherapy_ic50+monotherapy_ri+drc_intp_linear+drc_intp_lagrange+drc_intp_4PL
M12	monotherapy_ic50+monotherapy_ri+drc_baseline+drc_intp_linear+drc_intp_lagrange+drc_intp_4PL
M13	drug_categorical+cell_line_categorical
M14	drug_categorical+cell_line_categorical+cancer_gene_expression+chemical_structure
M15	drug_categorical+cell_line_categorical+cancer_gene_expression+chemical_structure+monotherapy_ic50+monotherapy_r
M16	drug_categorical+cell_line_categorical+cancer_gene_expression+chemical_structure+monotherapy_ic50+monotherapy_ri+drc_baseline
M17	drug_categorical+cell_line_categorical+cancer_gene_expression+chemical_structure+monotherapy_ic50+monotherapy_ri+drc_intp_linear
M18	drug_categorical+cell_line_categorical+cancer_gene_expression+chemical_structure+monotherapy_ic50+monotherapy_ri+drc_intp_lagrange
M19	drug_categorical+cell_line_categorical+cancer_gene_expression+chemical_structure+monotherapy_ic50+monotherapy_ri+drc_intp_4PL
M20	drug_categorical+cell_line_categorical+cancer_gene_expression+chemical_structure+monotherapy_ic50+monotherapy_ri+drc_intp_linear+drc_intp_lagrange+drc_intp_4PL

Supplementary Table 2.6. Example of the drug combination dataset provided by DrugComb. For each combination, two drugs (drug_row and drug_col) and the treated cell line (cell_line_name) are shown. Five synergy scores (ZIP, Bliss, Loewe, HSA, and S) and sensitivity scores (CSS) are shown for each experiment. For each drug combination, there could be more than one replicated experiment. The source of each experiment can be traced using the block ID.

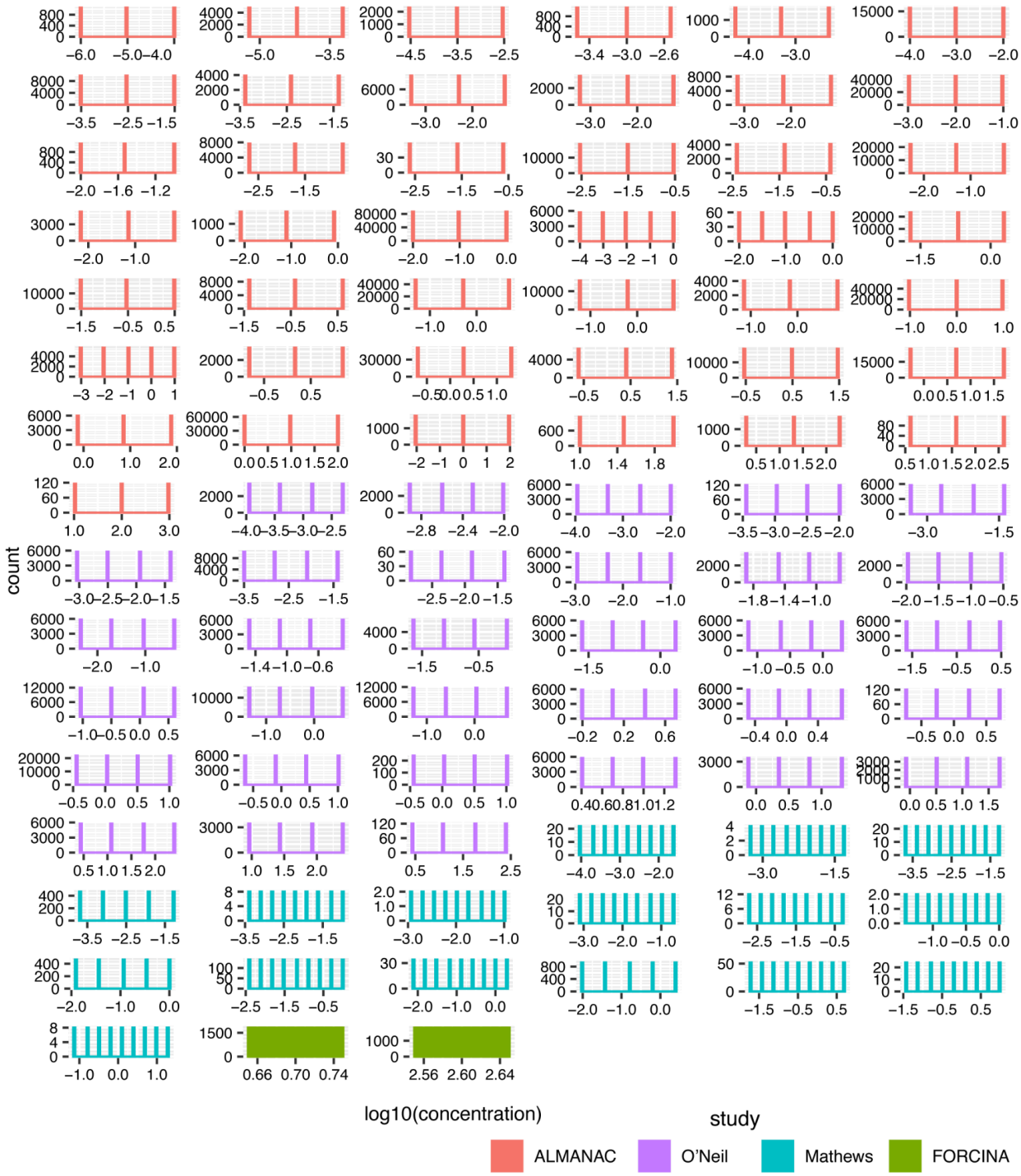
block id	drug_row	drug_col	cell_line_name	CSS	ZIP	Bliss	Loewe	HSA	S
1	5-FU	ABT-888	A2058	30.869	3.865	6.256	-2.951	5.537	19.839
2	5-FU	ABT-888	A2058	27.46	8.247	12.334	3.126	11.614	16.43
3	5-FU	ABT-888	A2058	29.901	6.063	11.660	2.452	10.941	18.871
4	5-FU	ABT-888	A2058	24.016	-4.280	5.145	-4.063	4.426	12.986

Supplementary Figures

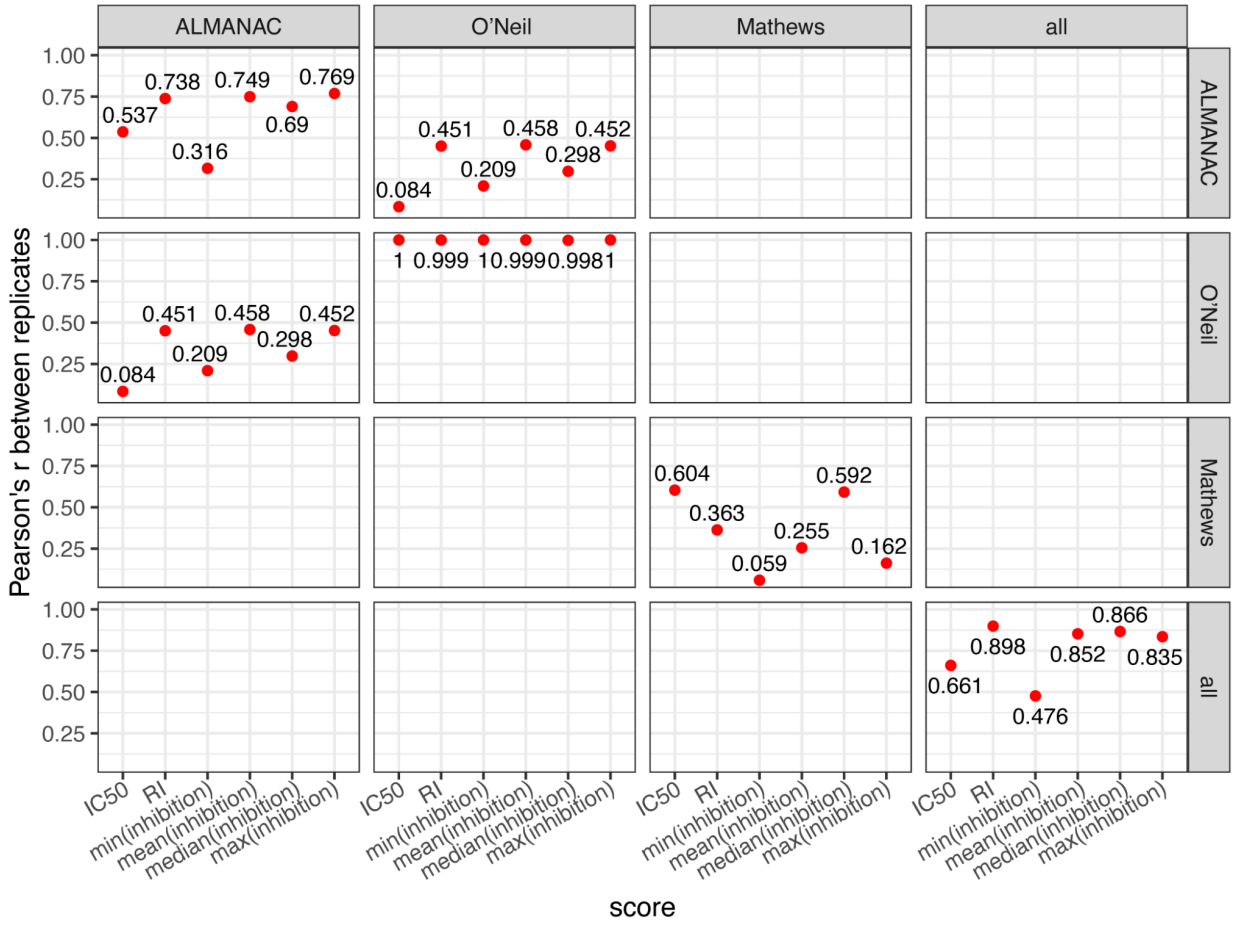
Supplementary Figure 2.1. Reproducibility of the six drug combination activity measurements within and between studies used in this dataset. The reproducibility is measured by Pearson's correlation between the response scores (S, CSS, Bliss, HSA, Loewe, and ZIP) of the same treatment-cell line combinations. The reproducibility can be used as a standard to evaluate the prediction performance in study and cross-study. For cross-study, only studies with overlapped treatment-cell line combinations were evaluated. FORCINA is not included in this figure since there are neither replicates within this dataset nor between this dataset and the others.



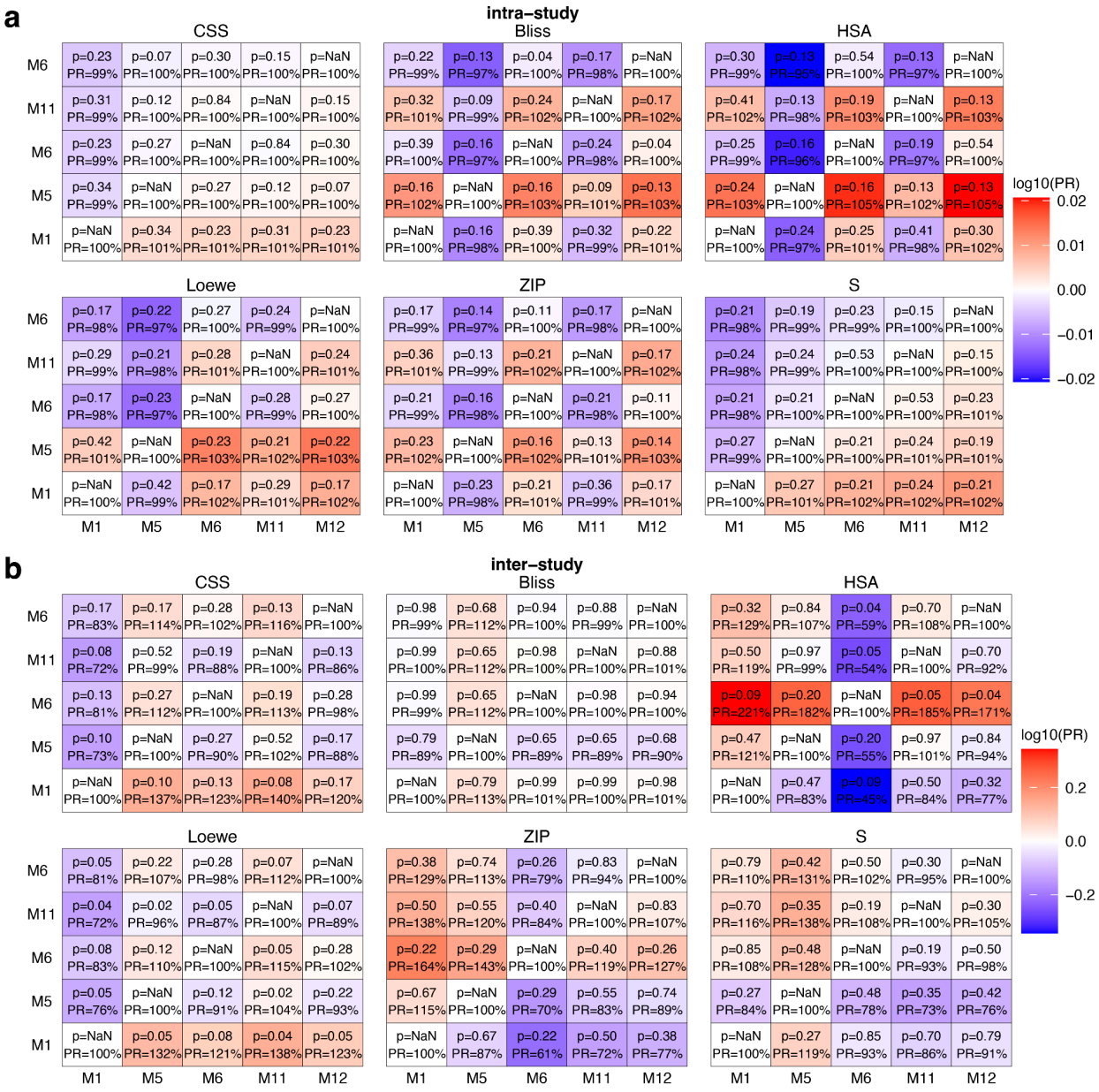
Supplementary Figure 2.2. Histograms show the concentration ranges (log10) for single drug dose-response measurements adopted in the four high-throughput screening studies. Different colors (red, purple, blue, and green) denote the HTS study (ALMANAC, O'Neil, Mathews, and FORCINA) the monotherapy dose setting is used for each dose-response curve. Since all doses start from 0, the log10 of the first concentration (-∞) is not shown in this graph.



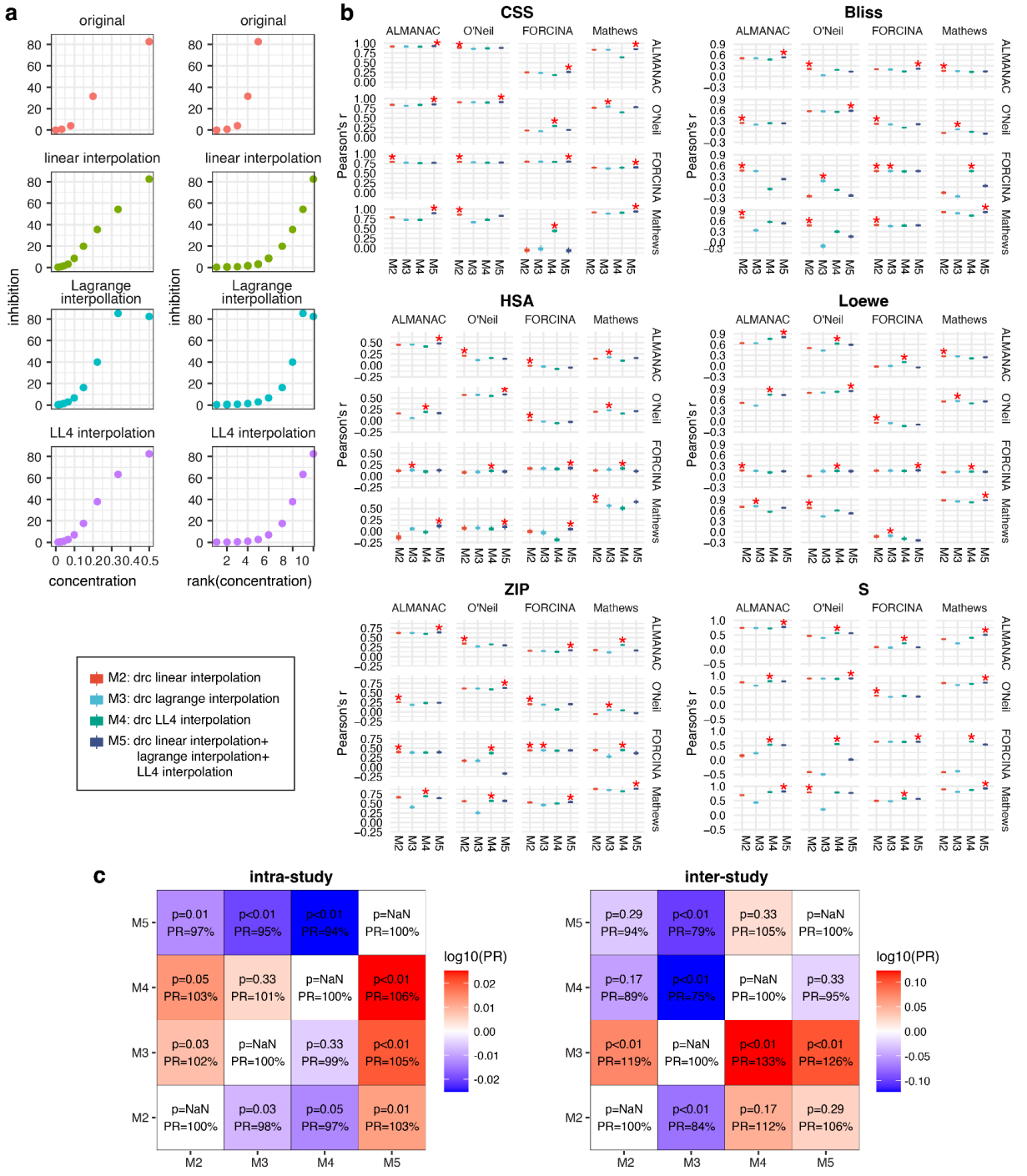
Supplementary Figure 2.3. Reproducibility of monotherapy response measurements in different studies used in this dataset. The reproducibility is measured by Pearson's correlation between the response score (IC50 and RI (relative inhibition)), and statistics of the dose-response curve (min, mean, median, and max of inhibition) of the same monotherapy treatment-cell line combinations. The reproducibility can be used as a reference for the prediction performance in-study and cross-study. For cross-study, only studies with overlapped monotherapy treatment-cell line combinations were evaluated. FORCINA is not included in this figure since there are no replicates in this dataset. "all" refers to the overall replicability across all studies.



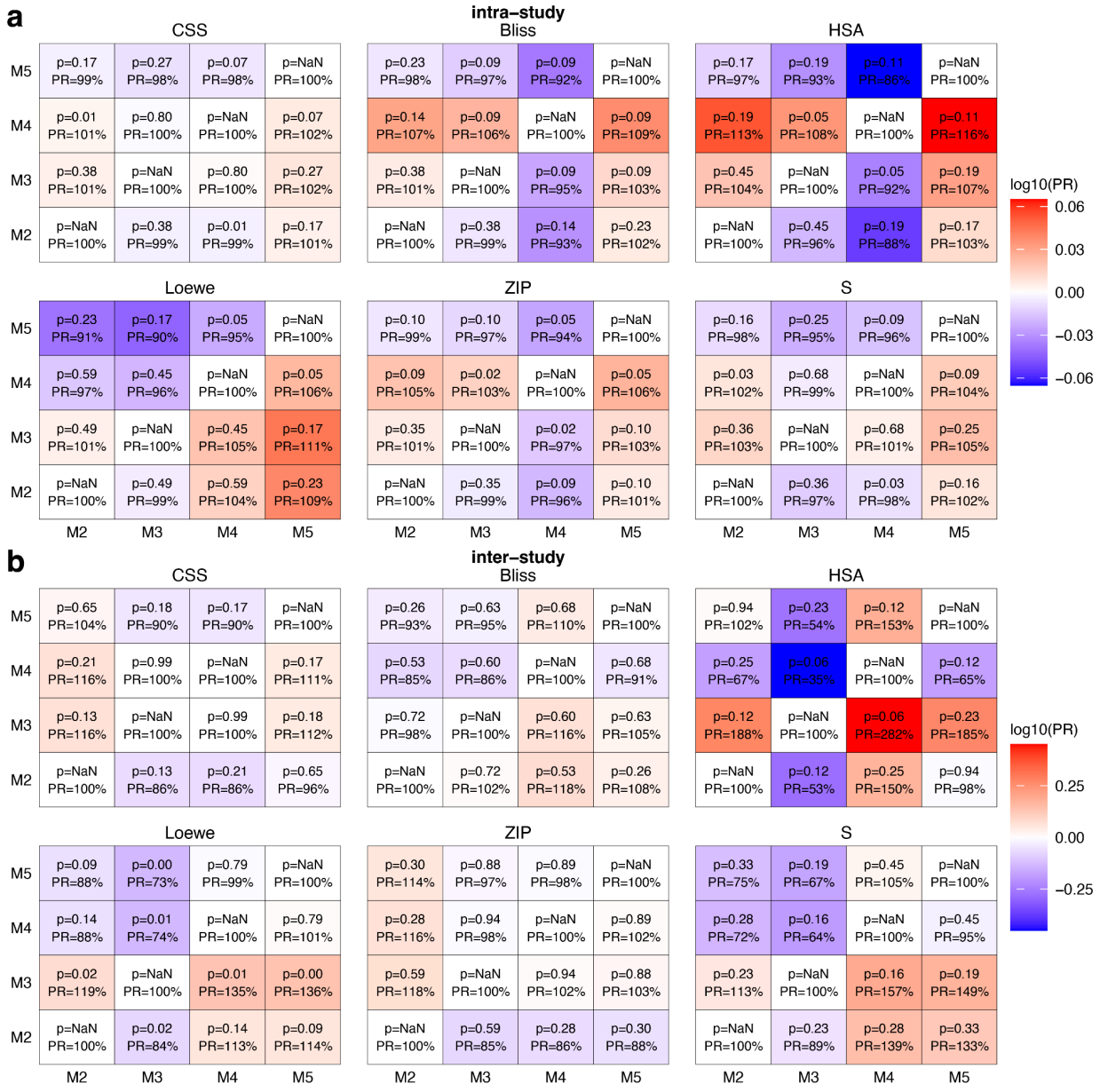
Supplementary Figure 2.4. Pair-wise comparison of model performances over each combination treatment response score (CSS, Bliss, HSA, Loewe, ZIP, S) using p-values from paired t-test and performance ratios (PR). The models in this figure correspond to **Figure 2.2 d**. **a.** intra-study cross-validation. **b.** inter-study cross-validation.



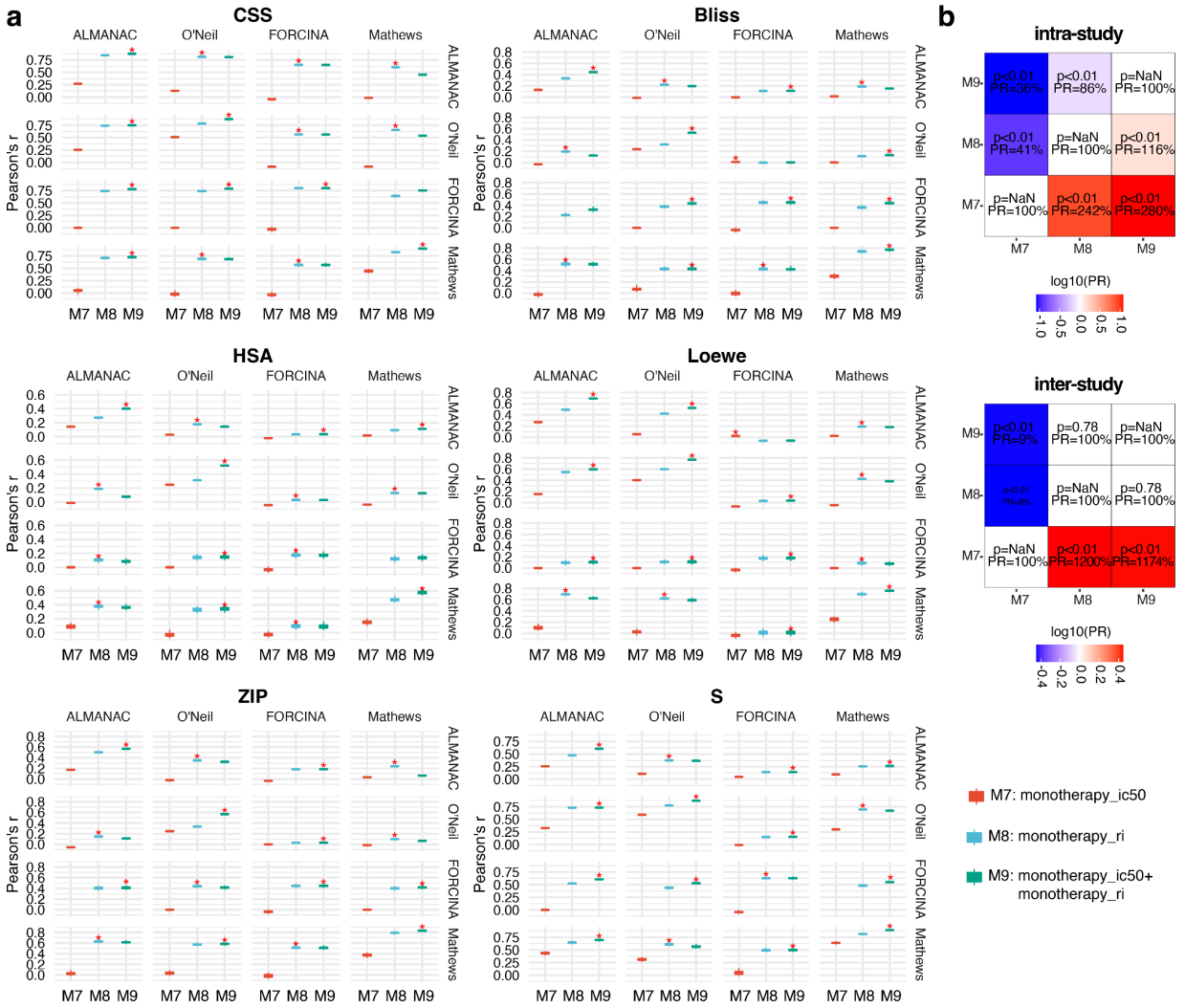
Supplementary Figure 2.5. The comparison between different interpolation methods for dose-response curves. a. example of original drc and different interpolation methods. The original dose-response curve contains only five doses and is interpolated to the maximum length, which is ten doses. Also, for the interpolation models, we used the magnitude of interpolated inhibition as the final feature. **b.** Performances of all interpolation models in different training (horizontal) and testing (vertical) settings. **c.** Comparison of performances between models by paired t-test and performance ratio (PR) on the average.



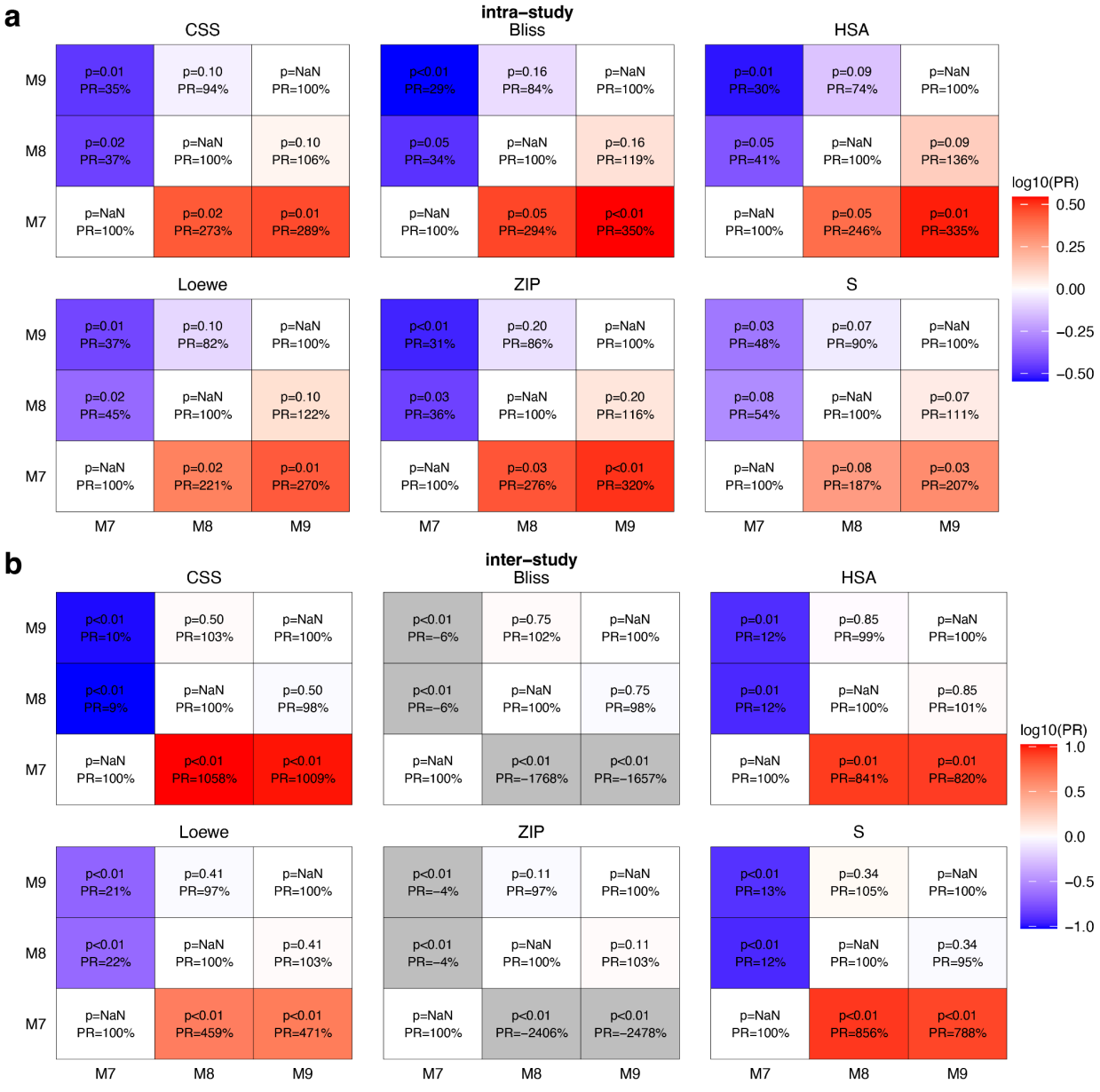
Supplementary Figure 2.6. Pair-wise comparison of model performances over each combination treatment response score (CSS, Bliss, HSA, Loewe, ZIP, S) using p-values from paired t-test and performance ratios (PR). The models in this figure are corresponding to **Supplementary Figure 2.5 c. a.** intra-study cross-validation. **b**. inter-study cross-validation.



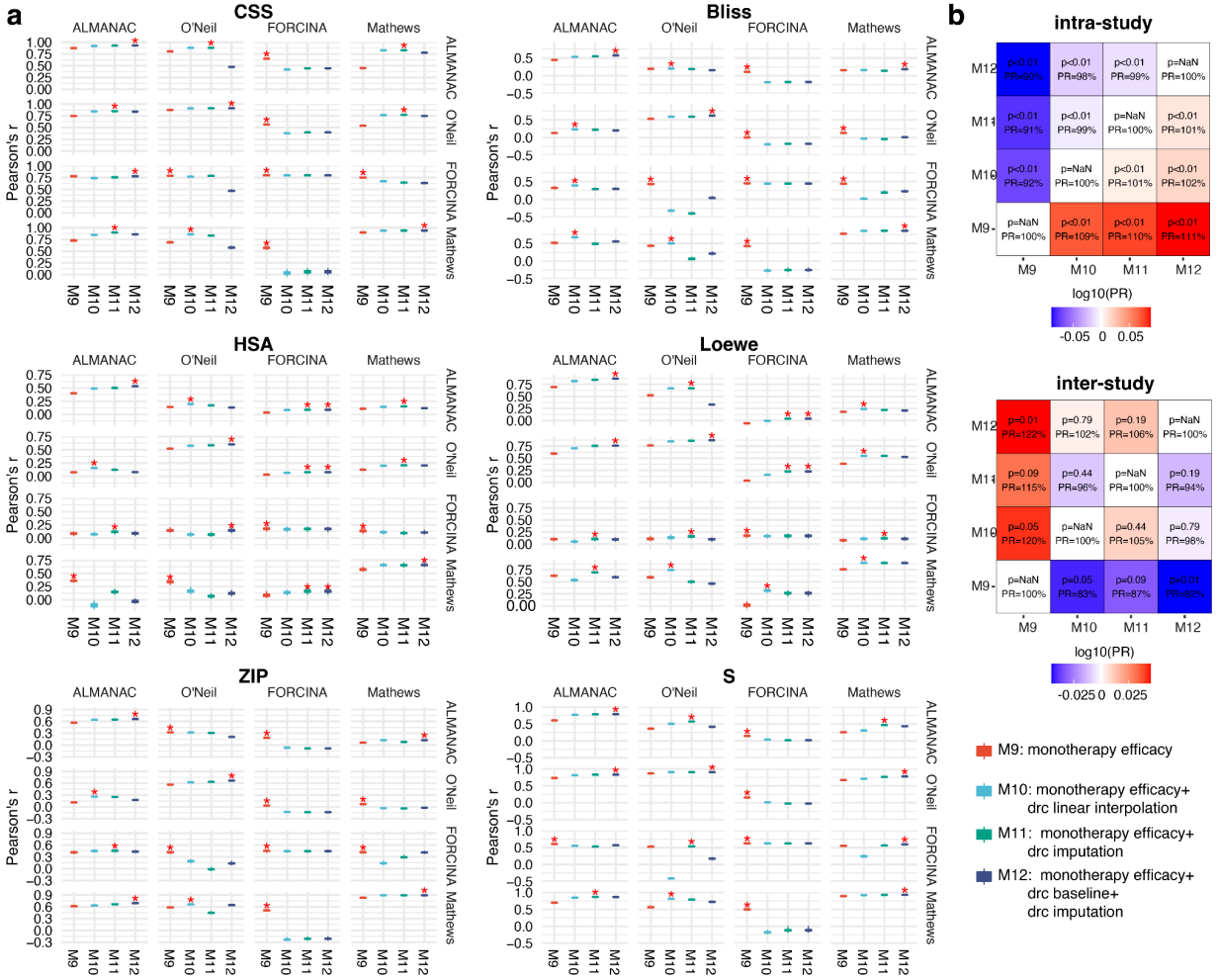
Supplementary Figure 2.7. The comparison between using different monotherapy efficacy scores as features. Both IC50 and RI (relative inhibition) are used to measure the monotherapy efficacy. We tested the performances by using either or both in intra- and inter-study predictions. **a.** Performances of all interpolation models in different training (top) and testing (right) settings. **b.** Comparison of performances between models by paired t-test and performance ratio (PR) on the average.



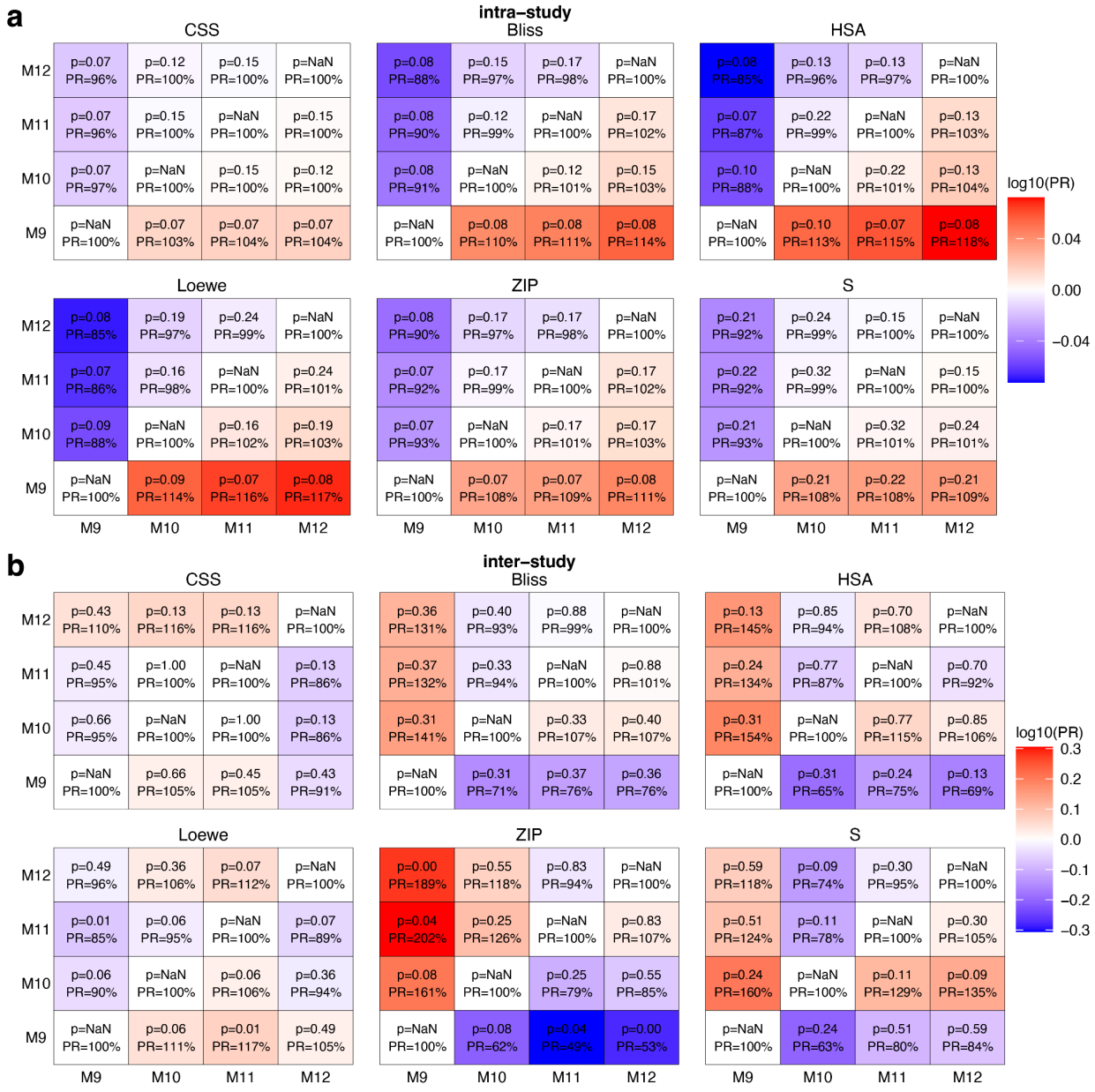
Supplementary Figure 2.8. Pair-wise comparison of performances of models from Supplementary Figure 2.7b over each combination treatment response score (CSS, Bliss, HSA, Loewe, ZIP, S) using p-values from paired t-test and performance ratios (PR). a. intra-study cross-validation. **b.** inter-study cross-validation.



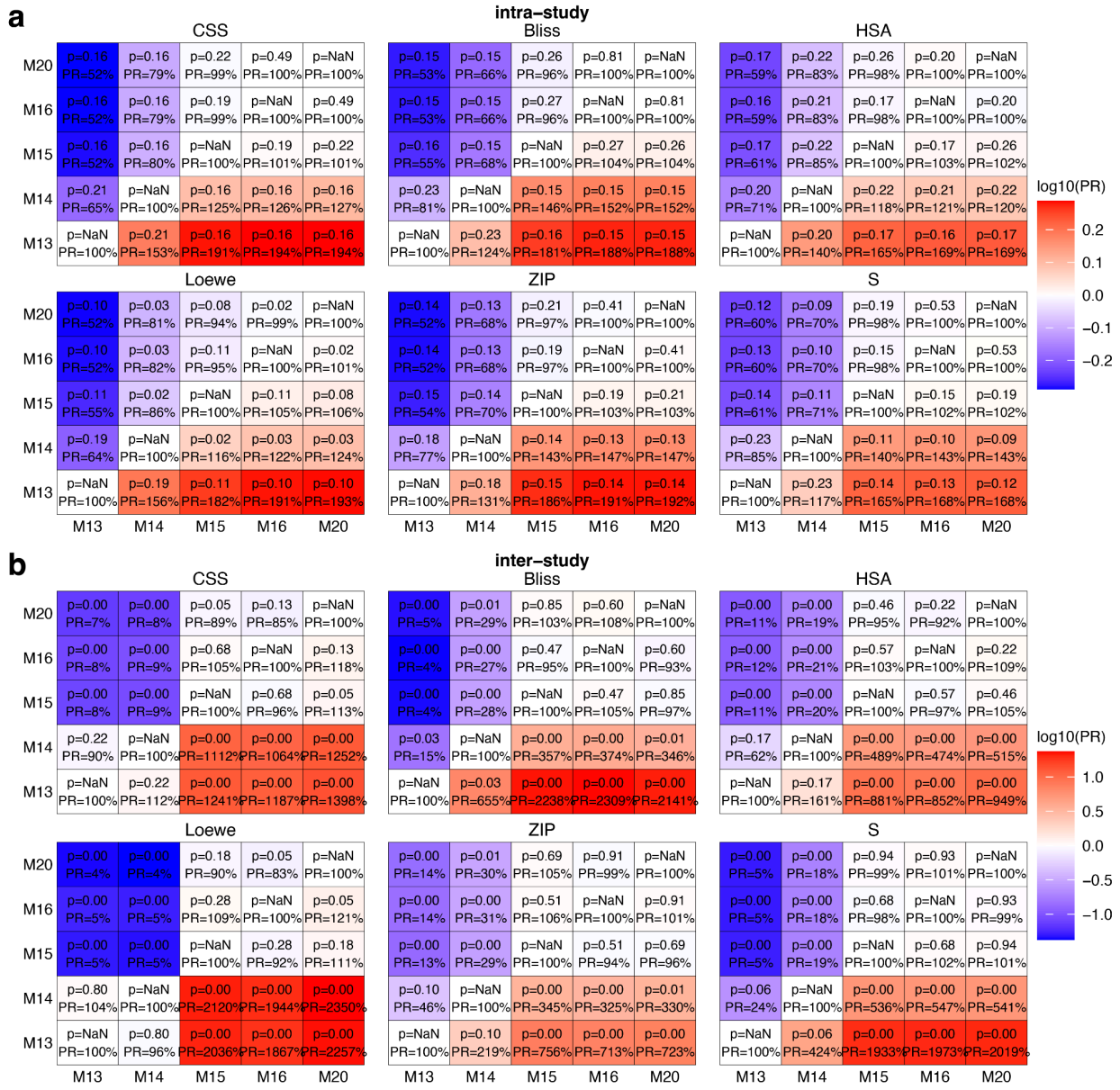
Supplementary Figure 2.9. Comparison between the monotherapy efficacy model as a baseline and baseline model added with different drc models. a. Performances of all interpolation models in different training (top) and testing (right) settings. **b.** Comparison of performances between models by paired t-test and performance ratio (PR) on the average.



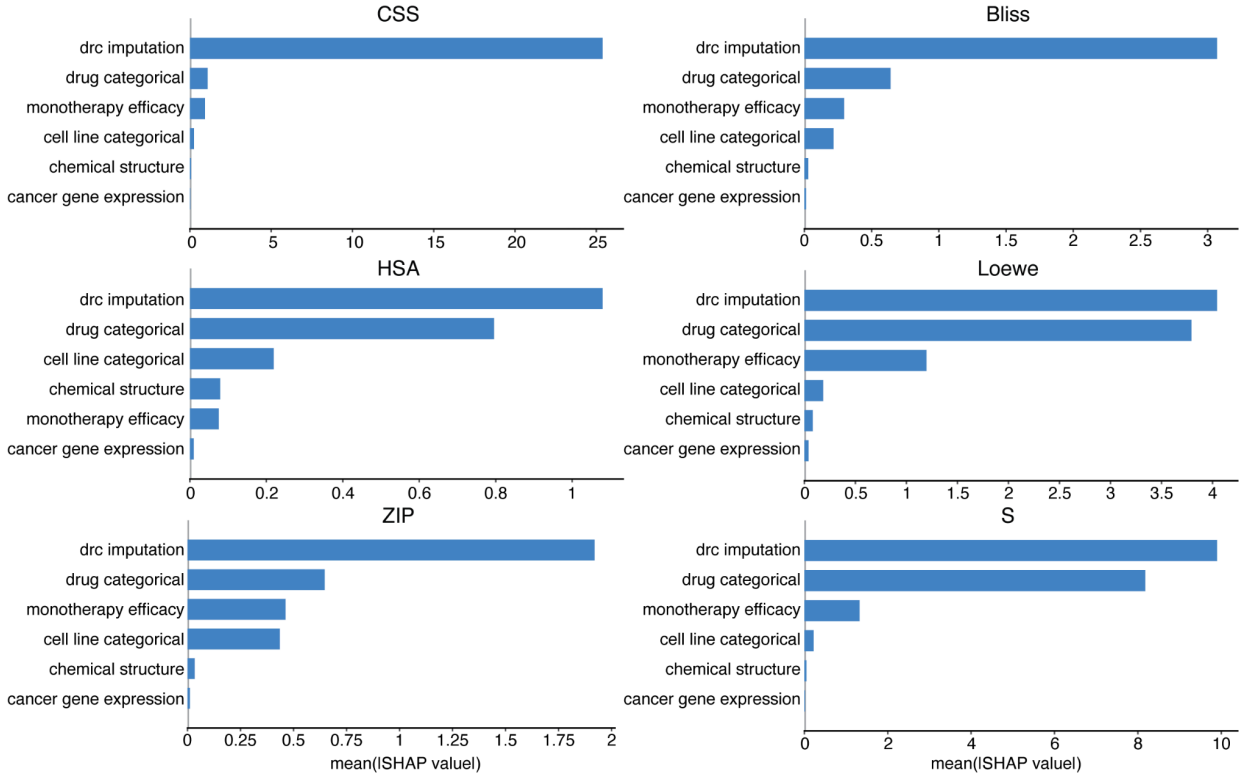
Supplementary Figure 2.10. Pair-wise comparison of performances of models from Supplementary Figure 2.9b over each combination treatment response score (CSS, Bliss, HSA, Loewe, ZIP, S) using p-values from paired t-test and performance ratios (PR). a. intra-study cross-validation. **b.** inter-study cross-validation.



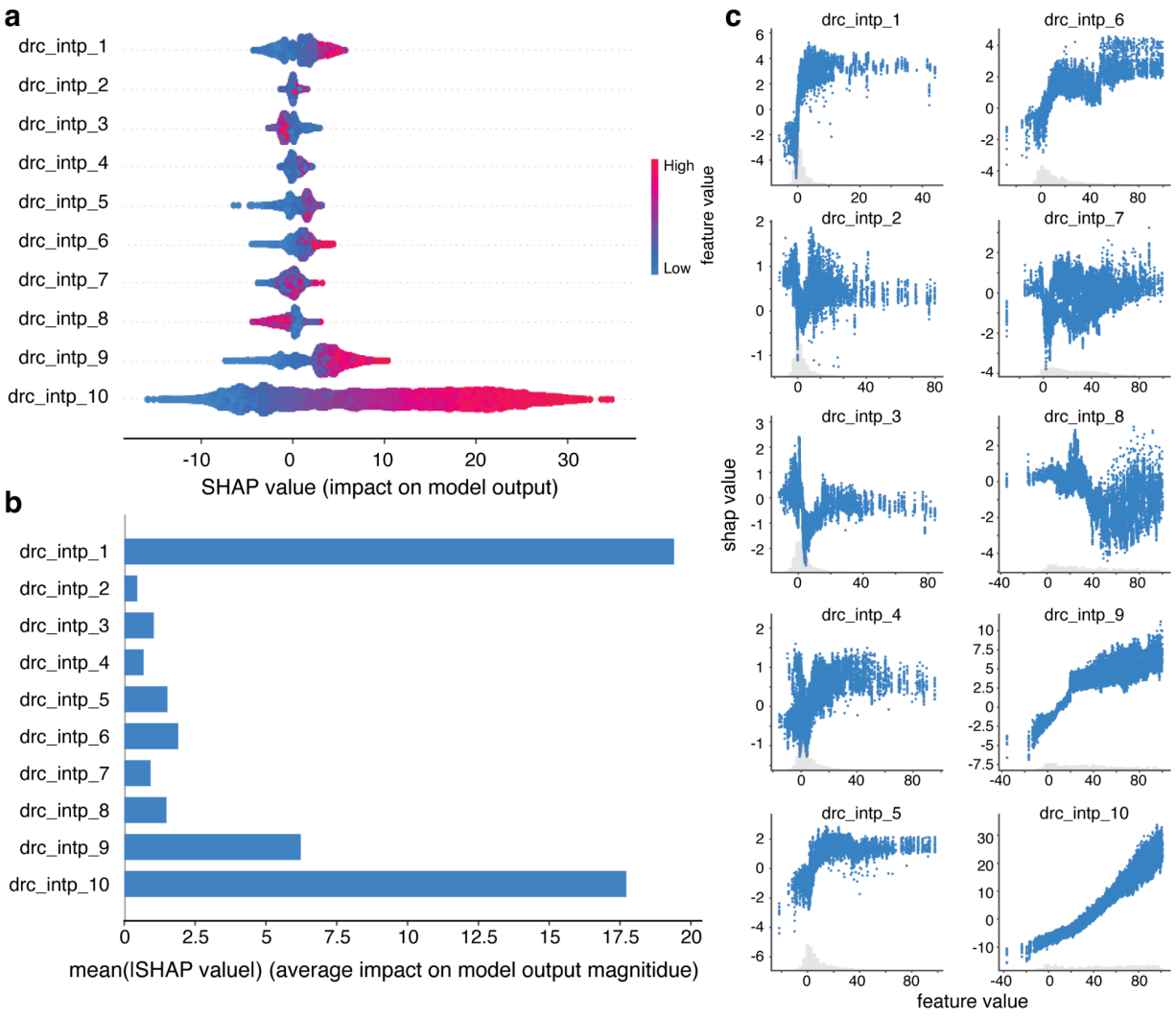
Supplementary Figure 2.11. Pair-wise comparison of performances of models from Figure 2.3c over each combination treatment response score (CSS, Bliss, HSA, Loewe, ZIP, S) using p-values from paired t-test and performance ratios (PR). a. intra-study cross-validation. **b.** inter-study cross-validation.



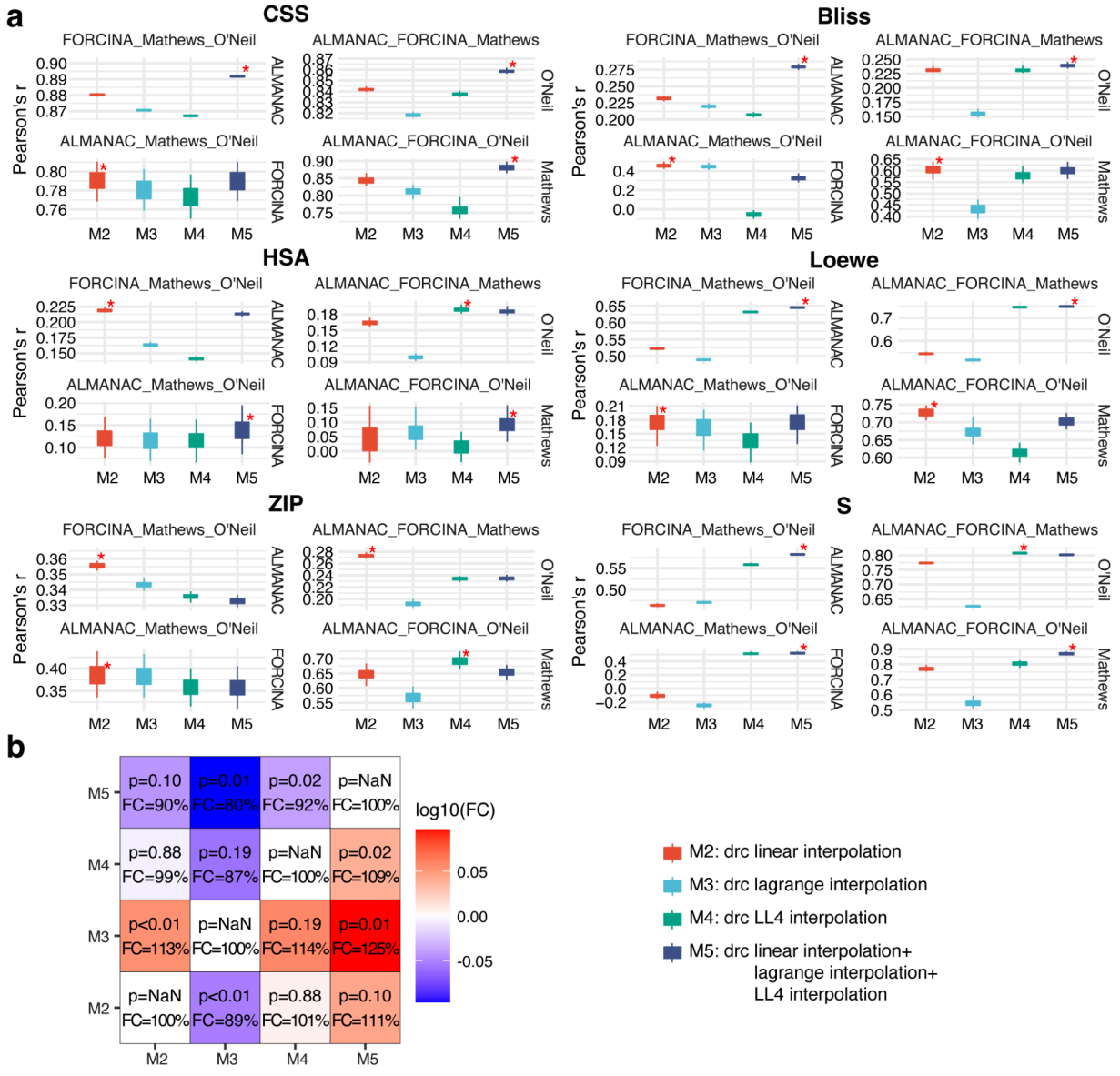
Supplementary Figure 2.12. Feature contribution of the best-performing model (M20 in Figure 2.3) in inter-study prediction when trained on ALMANAC and tested on the O'Neil study. The importance when predicting all six response scores is shown below.



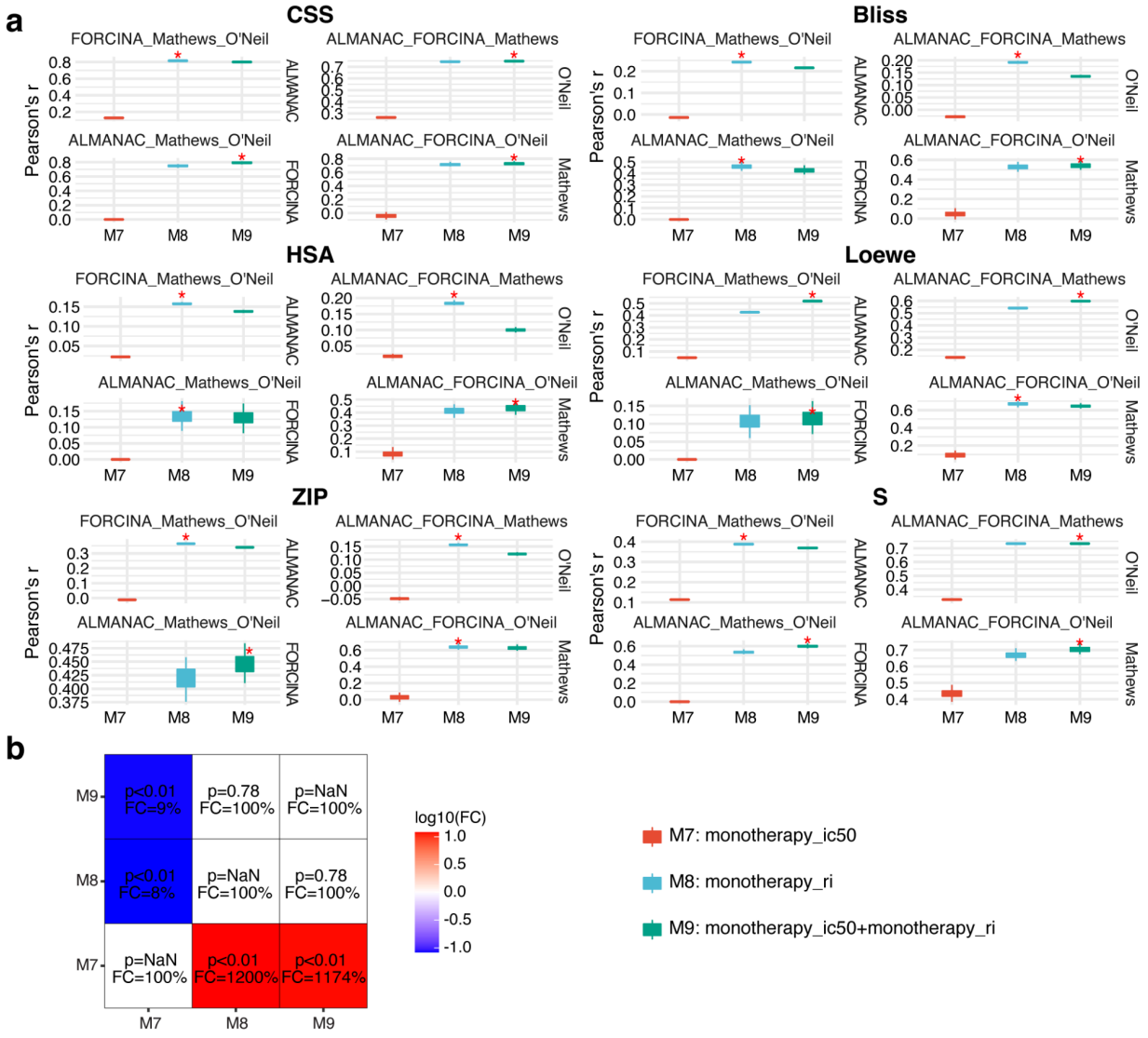
Supplementary Figure 2.13. Feature contribution of dose-response curve imputation features (M12) in inter-study CSS score prediction when trained on ALMANAC and tested on O'Neil study. a. summary plot of ten imputation features. **b.** The bar plot shows the contribution (average impact in model output magnitude) of ten imputation features. **c.** scatter plot shows the relationship between feature value and contribution (SHAP value) of 10 imputation features.



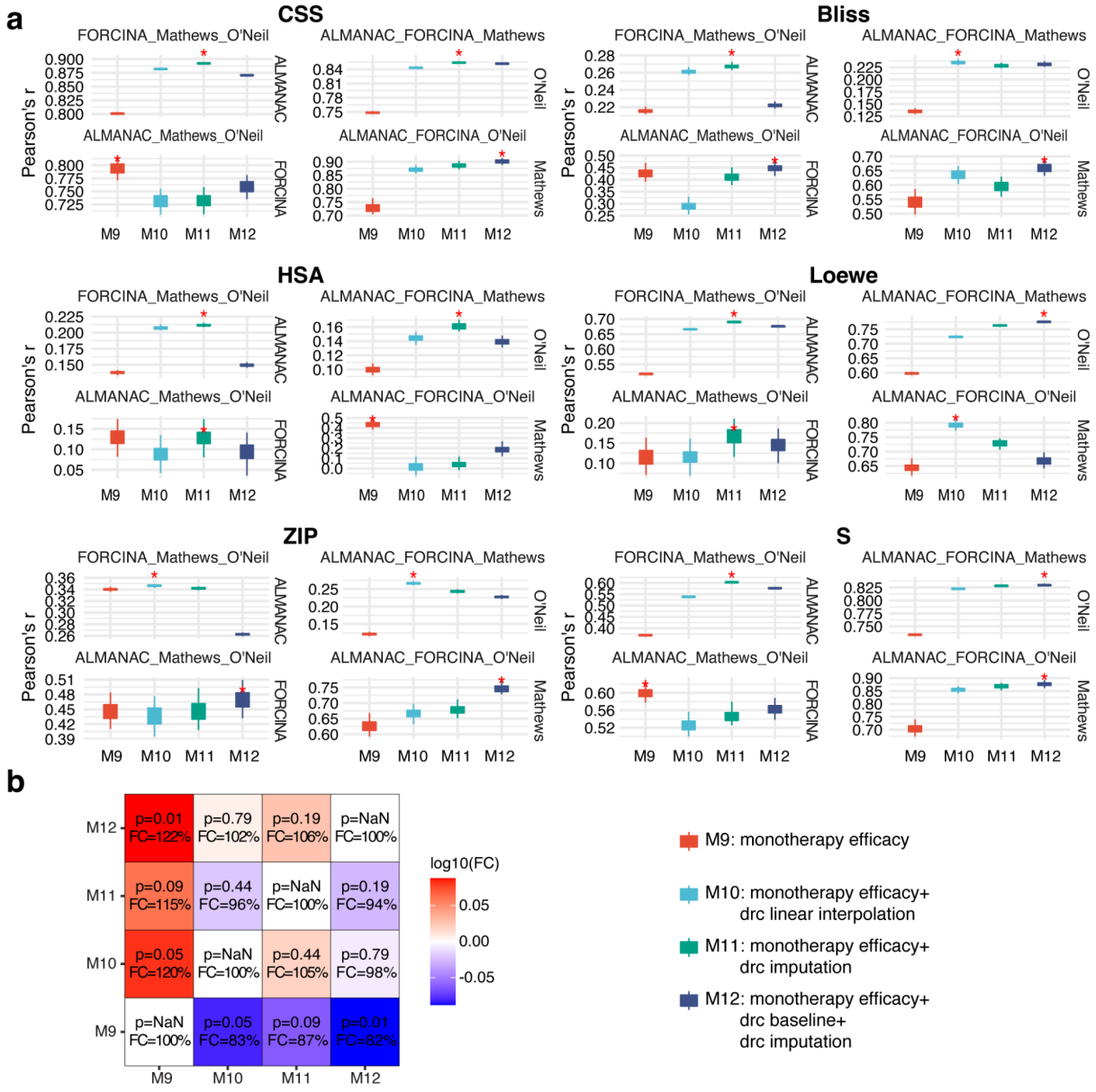
Supplementary Figure 2.14. Comparison between different dose-response curve interpolation methods in 3 vs. 1 inter-study cross-validation. a. Performances of all interpolation models in different training (top) and testing (right) settings. **b.** Comparison of performances between models by paired t-test and performance ratio (PR) on the average. The models in this figure are the same as in **Supplementary Figure 2.5.**



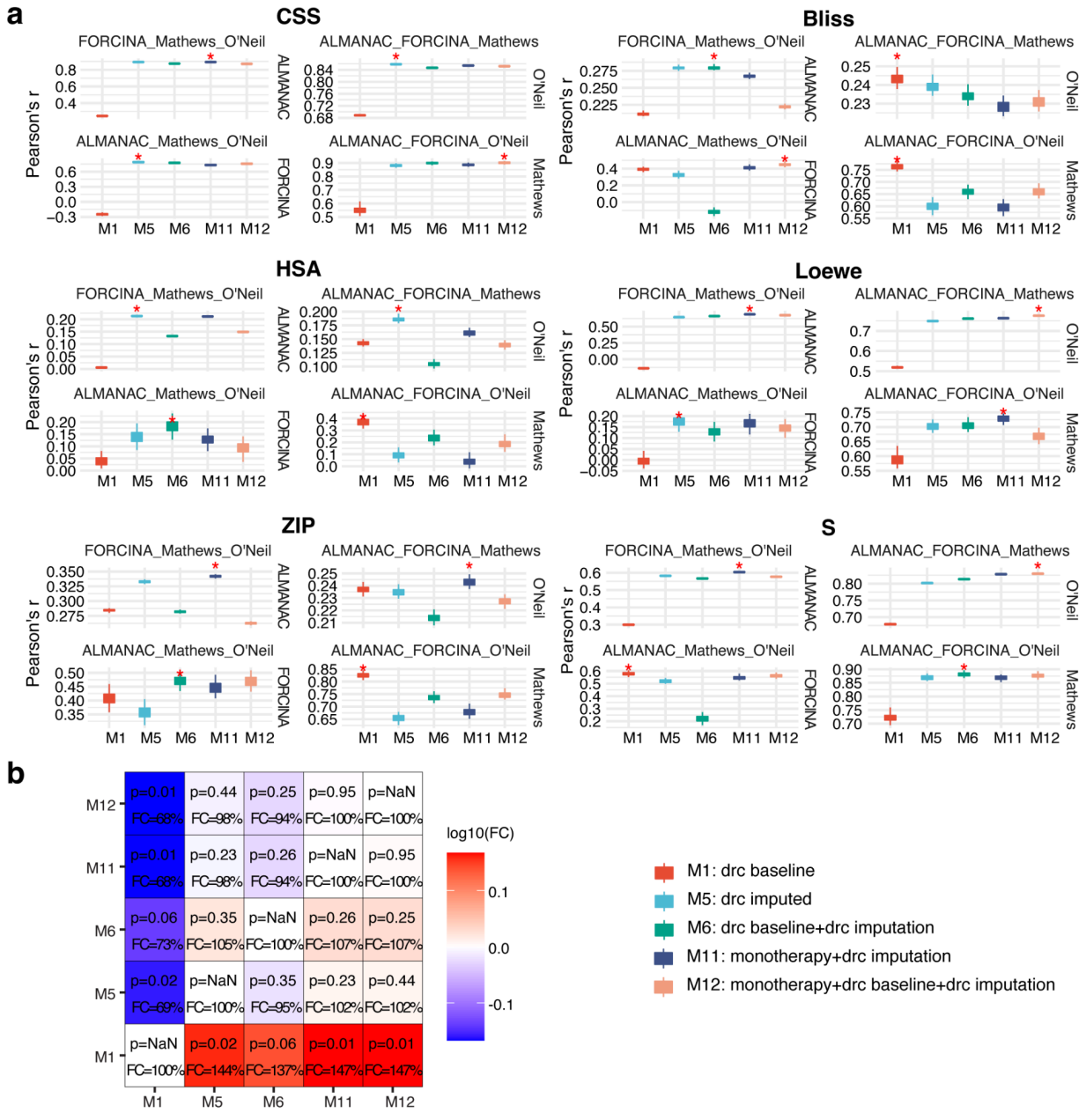
Supplementary Figure 2.15. Comparison between different monotherapy efficacy features in 3 vs. 1 inter-study cross-validation. a. Performances of all monotherapy efficacy models in different training (top) and testing (right) settings. **b.** Comparison of performances between models by paired t-test and performance ratio (PR) on the average. The models in this figure are the same as in **Supplementary Figure 2.7.**



Supplementary Figure 2.16. Comparison between monotherapy efficacy features in addition to different dose-response curve features in 3 vs. 1 inter-study cross-validation. a. Performances of all models in different training (top) and testing (right) settings. **b.** Comparison of performances between models by paired t-test and performance ratio (PR) on the average. The models in this figure are the same as in **Supplementary Figure 2.9.**



Supplementary Figure 2.17. Comparison between different combinations of pharmacological features in 3 vs. 1 inter-study cross-validation. a. Performances of all monotherapy efficacy models in different training (top) and testing (right) settings. **b.** Comparison of performances between models by paired t-test and performance ratio (PR) on the average. The models in this figure are the same as in **Figure 2.2.**



CHAPTER III: Machine Learning for Artemisinin Resistance in Malaria Treatment across *In Vivo-In Vitro* Platforms

Abstract

Drug resistance has been rapidly evolving with regard to the first-line malaria treatment, artemisinin-based combination therapies. It has been an open question whether predictive models for this drug resistance status can be generalized across *in vivo-in vitro* transcriptomic measurements. In this study, we present a model that predicts artemisinin treatment resistance developed with transcriptomic information of *Plasmodium falciparum.* We demonstrated the robustness of this model across *in vivo* clearance rate and *in vitro* IC50 measurement, and based on different microarray and data processing modalities. The validity of the algorithm is further supported by its first placement in the DREAM Malaria Challenge. We identified transcription biomarkers to artemisinin treatment resistance that can predict artemisinin resistance and are conserved in their expression modules. This is a critical step in the research of malaria treatment as it demonstrated the potential of a platform-robust, personalized model for artemisinin resistance using molecular biomarkers.

Introduction

Malaria raises major public health concerns in southeastern Asia and Africa (Asenso-Okyere et al., 2011; Conn et al., 2018; Dhiman, 2019; Mbacham et al., 2019; Organization & Others, 2020; Sachs & Malaney, 2002; Tabbabi et al., 2020; World Health Organization, 2020). *Plasmodium falciparum*, one of the five Plasmodium species leading to malaria, is the main cause of mortality, resulting in 400,000 deaths each year (*Fact Sheet about Malaria*, n.d.,

"Malaria: Biology and Disease," 2016; Talapko et al., 2019). The most effective treatment is artemisinin-based combination therapies, which have been used as the first-line treatment for malaria since the late 1990s (Miller & Su, 2011). Today, malaria remains a global health threat and drug resistance is a major contributor (Dhiman, 2019; Dondorp et al., 2009; Mok et al., 2015). After being transmitted from mosquitoes into the human body, P. falciparum experiences the rest of its life cycle in the peripheral bloodstream and liver. In the blood stage, they propagate asexually in red blood cells in the form of ring, trophozoite, and schizont developmental stages in 48 hours, resulting in daughter cells released in the peripheral bloodstream. The artemisinin (ART) resistance of P. falciparum happens specifically at the ring stage, when the parasites lose their apical complex and de-differentiate into round immature trophozoites, pushing their nuclei to one side of the cell, making the cell morphologically resemble rings under the microscope (Dondorp et al., 2009; "Mechanisms of Artemisinin Resistance in Plasmodium Falciparum Malaria," 2018).

In the past years, the research field has been tirelessly searching for the genomic and transcriptomic traits associated with artemisinin resistance (Ariey et al., 2014; Ashley et al., 2014; Cheeseman et al., 2012; Hunt et al., 2010; Mok et al., 2015; Takala-Harrison et al., 2013). For instance, it has been reported that a point mutation in the gene ubp1 confers artemisinin resistance in a *P. chabaudi* mouse malaria model (Hunt et al., 2010). This gene encodes a de-ubiquitinating enzyme and the missense mutation reduces de-ubiquitinating activity and alters the associated protein degradation pathways (Hunt et al., 2007). Additionally, multiple loci on chromosomes 10, 13, and 14 have been identified to be associated with the heritable trait of artemisinin resistance (Cheeseman et al., 2012; Takala-Harrison et al., 2013). Particularly, mutations in the gene kelch PF3D7_1343700 ('K13-propeller') on chromosome 13 have been

reported to be a significant molecular marker associated with artemisinin resistance (Ariey et al., 2014; Ashley et al., 2014; L. Zhu et al., 2018). Beyond mutations, changes in the expression of genes involved in the unfolded protein response (UPR) pathways have been linked to human artemisinin resistance (Mok et al., 2015).

Although many studies have focused on the relationship between individual gene mutation and expression and drug resistance in malaria, a systematic evaluation of the value of these biomarkers in clinical or pre-clinical applications remains needed. The recent Malaria DREAM Challenge, which blindly evaluated algorithms for predicting ART resistance addressed this need (Bionetworks, n.d.-a). The Malaria DREAM challenge leveraged an important dataset previously published (Mok et al., 2015), in which transcriptome profiles of *P. falciparum* isolates from 1,043 patients were measured *in vivo* without treatment and the resistance status was reported. The participants of the challenge were asked to predict the *in vitro* drug response of independent isolates with expression data obtained before and after perturbations with dihydroartemisinin (DHA).

We are presenting here the top-performing algorithm ranked by accuracy to the above-described question, a machine-learning model for predicting artemisinin resistance based on the transcriptomic profile of the parasite. This model addresses several key challenges in malaria genomics and drug research: how to build models that can deliver across *in vivo* and *in vitro* datasets? Most of the *P. falciparum* experiments are cultured with human blood and carried out *in vitro*, while clinical applications require the model to be robust for *in vivo* datasets. How to make models deliverable from one measurement platform to another and thus allow wide application and generalization of the models? Of note, the training dataset of the DREAM challenge comes from a customized, two-color expression panel, while the test dataset came

from one-color Agilent HD Exon Array with many more probes for each gene. How to identify the biomarkers and create the minimal panel of genes that both reveal the biological insights/pathways related to ART resistance and are capable of making good predictions? We address the above challenges by developing a cross-platform, *in vivo-in vitro* generalizable model for ART resistance prediction, and analyzing independent contributions of gene expression signatures. We identified four molecular signatures important to the model: PF3D7_0523000 (pfmdr1), PF3D7_1245300, PF3D7_1372000, PF3D7_0805000, creating a panel that almost matched the entire transcriptome in performance when predicting the cross-*in vivo-in vitro* drug resistance. Examination of co-expression modules reveals stable co-regulation modules of the top molecular features related to ART resistance.

Results

Study design to investigate the transferability of models for in vivo-in vitro and cross-platform generalization

The overall study design intends to construct a model that is transferable across microarray platforms and across i*n vivo-in vitro* conditions. The training dataset comes from Mok et al., which is a large cohort (1,043 isolates) of transcriptomic data of *P. falciparum* collected from southeast Asia during 2012-2014 (Mok et al., 2015). The parasite samples were directly taken from the peripheral blood of patients with acute falciparum malaria. The customized, printed expression panel measured 4978 genes out of ~5591 genes of the *P. falciparum* genome. ART-resistance phenotype was identified by the rate of clearance of parasites in the patient's peripheral blood, which is quantified by the clearance half-life upon ACT treatment. In this study, the samples with clearance half-life>5 hours are considered as ART-resistant, and labeled with “Slow” clearance rate. On the other hand, The samples with <=5 hours of clearance half-life

are labeled as "Fast" in terms of clearance rate, and considered as non-ART-resistant samples (**Figure 3.1**).

This study, as shown below, starts with cross-validation with the above-described dataset. Additionally, the design of the test set differs from the training set in its sampling geographic site and timing of sample collection, synchronization status, microarray platform, and measurement target, introducing new challenges to the prediction models. The *in vitro* test set consisted of unpublished data from 32 isolates collected from the Thai-Mayanmar border (**Figure 3.1a**). The isolates are synchronized *in vitro*. Each isolate was examined twice, once without treatment and once with ART (DHA) treatment. The expression level was taken separately at 6 hours and 24 hours post-infection (hpi). The test data was measured using an Agilent HD Exon Array with many more probes (on average 12 per gene) than the printed array in the training data (on average 2 per gene) (**Figure 3.1b-c**). This test set was the test set for sub-challenge 1 of the Malaria DREAM challenge in which the task was to predict ART IC50 given a training set consisting of transcriptomes of parasites with known IC50. Additionally, the training data used a two-color array and the test set used the Bozdech one-color array, which is expected to introduce challenges in data analytics (Patterson et al., 2006). Due to the differences in the array platforms, the methods used to pre-process the arrays also differ (see Methods). The test set panel included non-coding RNAs, which are excluded in the training set. This results in a total of 5,540 genes in the test set data. For the test set, a continuous value of IC50 upon artemisinin treatment is given as the testing target. The direct test data on this challenge remains a hidden set for future model refinement by the scientific community. However, an independent test set of 30 isolates collected in the exact manner and cohort was available through sub-challenge 1 of this challenge (Bionetworks, n.d.-a), which is used as the test set to evaluate model transferability in this study.

Besides the DREAM challenge dataset, we also collected four independent public P. falciparum transcriptome datasets, of which two were sampled *ex vivo* and two *in vitro*, to further validate the robustness of transferability of the cross-platform model in this study. All transcriptomes used in this study were analyzed by t-SNE to show the differences between ART-resistance/sensitive samples, sampled conditions (*in vivo*, *in vitro,* or *ex vivo*), independent studies, and treatment type (**Supplementary Figure 3.4**).

Excellent performance for within-cohort prediction of artemisinin clearance rate

The large collection of the Mok et al. data allows us to evaluate the models by two approaches. First, we can evaluate the model performance by cross-validation within the 1,043 isolates. Cross-validation is a commonly used scheme to evaluate model performance by holding out part of the data as the testing set and using the other part as the training set. Second, we can evaluate the model performance by training a model on the Mok et al. data and test on the *in vitro* data as described above. In this section, we describe the behavior of the model in the within-cohort cross-validation using the Mok et al. data. Clearance half-life was labeled "fast" or "slow" according to whether the parasite clearance half-life is longer than 5 hours. We labeled 'slow' as 1, and 'fast' as 0 in the following experiments.

We carried out ten-fold cross-validation by including all genes as features (**Figure 3.2**). Specifically, in each round, 10% of the isolates were held out as the test set, and 90% were used as the training set. We tested a selection of base learners, including LightGBM, XGboost, random forest, Gaussian Process Regression (GPR), and linear regression (see Methods). Because an important goal of this study is to develop a model transferable to transcriptome data collected using different platforms, which can be of drastically different distribution, we also tested if rank normalization of the expression data changed performance.

LightGBM, a tree-based gradient boosting method, marginally excelled in performance for both Area Under the Receiver Operating Curve AUROC and AUPRC measurements (**Figure 3.2c**) compared to other alternatives. It achieved a mean AUROC [95% confidence interval] of 0.8384 [0.8121, 0.8705], compared to XGboost (0.7669 [0.7262, 0.7910]), random forest (0.7782 [0.7441, 0.8099]), GPR (0.8456 [0.8212, 0.8673]) and linear regression (0.8448 [0.8206, 0.8668]). For AUPRC [95% confidence interval], LightGBM performed at 0.6983 [0.6438, 0.7522], compared to XGboost (0.6613 [0.5994, 0.7234]), random forest (0.5752 [0.5049, 0.6387]), GPR (0.6742 [0.6198, 0.7280]) and linear regression (0.6717 [0.6176, 0.7252]). Rank normalization does not present substantial changes in performance (**Figure 3.2e, Supplementary Table 3.1**), we chose to maintain this operation to support cross-platform robustness.

Transferring models across platforms

The test data differs from the above examined *in vivo* data in that it was collected from laboratory-cultured P. falciparum strains. This allows synchronization, and thus the gene expression levels were sampled under four different conditions: 1) 6 hours post-invasion (hpi), 2) 24 hpi, 3) 6 hpi and treated with dihydroartemisinin (DHA) (6 hpi-p), 4) 24 hpi and treated with DHA (24 hpi-p). We evaluated the models based on different base learners as described above for each of the expression data. Because the test target is IC50, we labeled ‘slow’ as 1, and ‘fast’ as 0 in our training.

As expected, 6 hpi without treatment demonstrated the strongest performance, as the original training data was pre-treatment as well (**Figure 3.3d**). Additionally, LightGBM maintains to be the strongest base learner. In this case, rank normalization does not change the performance substantially, so we retained it in the pre-processing steps (**Figure 3.3e, Supplementary Table**

2). This combination achieved a Pearson correlation [95% confidence interval] of 0.2318 [0.1379, 0.5306], Spearman's correlation of 0.2467 [0.1457, 0.3548], and a C-index of 0.5837 [0.5474, 0.6216] between the predicted clearance rate and IC50. Of note, the gold standard used in training is non-granular values but rather a binary value of 'fast' and 'slow'. Yet, we still received meaningful predictions using a different microarray platform and data collection status ($p < 1e\text{-}6$) compared to random prediction.

We further evaluate the best-performing *in vivo* LightGBM model to four other public datasets for ART resistance prediction, where the ART resistance for each sample was available (**Supplementary Table 3**) (Mok et al., 2011, 2015, 2021; Shaw et al., 2015; L. Zhu et al., 2018), and results were shown in **Supplementary Figure 3.5**. We noticed that on ev vivo data, the model achieved better cross-platform accuracy than *in vitro* data overall. The *in vivo* model achieved 0.75[0.6431, 0.9773] and 0.6894[0.6065,0.8060] AUROC[95% confidence interval] on the GSE25878 and GSE59098 dataset, respectively. While on the *in vitro* dataset GSE151189, the model only achieved 0.5355[0.4530, 0.6416] overall AUROC[95% confidence interval]. One possible reason could be the *ex vivo* transcriptomes show more similarity to the *in vivo* data the model was trained on (**Supplementary Figure 3.4 b and c**). Interestingly, we also noticed the model prediction heavily relies on the *ex vivo* cultured time, treatment by DHA, and developmental stages (hpi), indicating these factors may change the expression levels of effector genes related to ART resistance.

Robustness in molecular features across in vivo and in vitro environments

It was very encouraging that a model can be developed and carried across such different *in vivo* and *in vitro* scenarios, and across experimental platforms, which prompted us to examine the top molecular features that contributed to this prediction. We first used SHapley Additive

exPlanations analysis (SHAP) to find out which genes played important roles in the *in vivo* ART-resistance prediction (S. Lundberg & Lee, 2017). SHAP analysis is a feature importance analysis method that recently gained popularity, in which the importance of one feature is considered in the context of all other features. This approach has the advantage of delineating gene features that are important for predicting ART resistance versus the ones that happened to be correlated to an important feature. Table 1 shows the top genes during the ten-fold cross-validation. Among them, there were five genes recognized by all ten models, showing consistent importance (**Supplementary Figure 3.2**). The SHAP analysis is test set-dependent. This unique feature allows us to test the robustness of these features further in the *in vitro* data. We found the same set of top genes still showed significant contribution in *in vitro* prediction (**Figure 3.3a-b, Supplementary Figure 3.3**). Of note, about ~70% of top genes (four out of top five, seven out of top ten, 14 out of top 20, and 22 out of top 30) were found to be shared by both *in vivo* and *in vitro* datasets, showing coherence in top-ranked features across platforms (**Figure 3.5a**). Pfmdr1 is among the most significant contributors in both *in vitro* prediction and *in vivo* prediction. This result supports the robustness of the identified molecular features.

We further investigated the functions of top contributing genes considering both *in vivo* and *in vitro* predictions of ART resistance in malaria **(Figure 3.4a and Table 3.1**). Among them, pfmdr1 (PF3D7_0523000), *Plasmodium falciparum* multidrug drug resistance gene 1, has been reported to play an essential role in response to a broad range of ACT antimalarials (Gil & Krishna, 2017; Koenderink et al., 2010; Sidhu et al., 2006). Mutants and polymorphisms of this protein have been widely reported to be associated with antimalarial drug resistance, and the increase of pfmdr1's expression will increase susceptibility to artemisinin (Chavchich et al., 2010; Dahlström et al., 2009; Eastman et al., 2016; Gupta et al., 2014; Holmgren et al., 2006,

2007; Imwong et al., 2010; Ngalah et al., 2015; Ould Ahmedou Salem et al., 2017; Sidhu et al., 2006; Sisowath et al., 2007; Ursing et al., 2006). The identification of this gene at the top of the list and its positive contribution to both IC50 and clearance rate corroborates the validity of the approach (**Supplementary Figures 3.2 and 3.3**).

We found other interesting genes in this list. First, PF3D7_1372000 is a *Plasmodium* exported protein of the Poly-Helical Interspersed Sub-Telomeric (PHIST) protein family (Tarr et al., 2014; Warncke et al., 2016), also known as the PRESAN family (Oakley et al., 2007; Sargeant et al., 2006). Although detailed functions of most *Plasmodium* exported proteins are yet to be revealed, in general, the parasite-exported proteins are pivotal for parasite survival by interacting and interfering activities of the infected cells (Maier et al., 2008). A recent study has suggested that the expression level of PF3D7_1372000 is associated with mutations of kelch PF3D7_1343700 ('K13-propeller') (Siddiqui et al., 2020), whose mutations have been reported to be a significant molecular marker associated with ART resistance (Ariey et al., 2014; L. Zhu et al., 2018). Second, PF3D7_1245300 is a Nedd8-conjugating enzyme UBC12, which has a central role in the cell cycle and DNA damage repair (Karpiyevich et al., 2019). Since the malaria parasite has a unique and unusual life cycle, the molecular machines in cell replication processes are specially designed for its survival. As Plasmodium responds to artemisinin-induced stress by delaying their cell cycle progression and inducing a state of dormancy during early ring-stage development (van Biljon et al., 2018), UBC12 likely presents as an important feature through this mechanism. Leave-one-out feature selection strategy based on the top ten genes shows that taking PF3D7_1245300 away will undermine *in vitro* prediction performance (**Figure 3.3d**), indicating this gene is crucial for *P. falciparum*'s survival in both laboratory environments and in the human body. Two other genes, PF3D7_0805000, a putative member of the alpha/beta

serine hydrolase superfamily that mediates a variety of metabolic reactions of ester hydrolysis, and PF3D7_1038700, another *Plasmodium* exported protein with unknown function, appeared in the top list. The association between these 2 genes with ART resistance is currently unknown.

We further investigated other proteins related to these top contributing genes based on the protein-protein interactome generated from blue native-polyacrylamide electrophoresis with quantitative mass spectrometry (Hillier et al., 2019). We first extracted interacting proteins with pfmdr1 and PF3D7_1245300 and found 20 and 37 interacting proteins, respectively. The other three proteins of the top genes were not observed in the interactome. Then we performed GO functional enrichment analysis of these proteins and identified the significantly enriched protein clusters with FDR p-value cutoff of 0.05 (**Figure 3.4b**). For the multidrug resistance gene pfmdr1, the interacting proteins are associated with RNA processing, COPII-coated vesicle budding, and the formation of the translation preinitiation complex. For the Nedd8-conjugating enzyme UBC12 (PF3D7_1245300), as expected, the interacting proteins are associated with protein ubiquitination, a process previously found to be important for treatment resistance in malaria (Dogovski et al., 2015; Tilley et al., 2016).

We went on to construct models only based on the top genes identified by SHAP analysis (**Figure 3.3c**). We found that for within-*in vivo* cross-validation, 30 genes can completely recover the performance of the model using the entire transcriptome. Additionally, the top genes identified in the above analysis successfully reached the performance of the entire gene panel when delivering the model to the *in vitro* test set. We acknowledge the existence of fluctuation in performance after the sixth top genes. The likely reason is that SHAP identifies independent features, and as we increase the number of features beyond six, the ones that are comparably weaker yet orthogonal to the top features are included. Despite this limitation, this result supports

the validity of the top features we identified in this study as potential biomarkers for ART resistance.

While kelch13 genetic mutations are significantly correlated with ART resistance phenotype in the *in vivo* population study (Pearson's r = 0.6143, p<1e-6), no significant correlation of kelch13 transcription with ATR-resistance phenotype has been found (Mok et al., 2015). This result is concordant with our SHAP analysis results, as the kelch13 transcription level turned out with no contribution to ART resistance prediction. Machine learning models with feature sets excluding kelch13 transcription level still maintained similar performances (**Supplementary Tables 3.1 and 3.3**). We also evaluated the ART-resistance model performances in different genetic variation cohorts, including K13 KP/BTB mutations, crt-N326S, crt-I356T, fd-D193Y and mdr2-T484I (**Supplementary Table 3.4 and Supplementary Figure 3.1**). The ART-resistance model is still quite predictive within K13 subgroups, with mutations (group 2) and heterozygous alleles (group 3) (**Supplementary Figure 3.1**).

Conserved co-expression patterns of top-ranking features

We next examined if the top-ranking features in the *in vivo* test and in the *in vitro* test share similar expression patterns or regulatory modules. We took the top 30 features for each and calculated the Pearson correlation of expression values across all samples separately for the *in vivo* and *in vitro* datasets. This step created coexpression networks (**Figure 3.5**). Among the top 30 genes, 22 are shared between *in vivo* and *in vitro* tests, a piece of supporting evidence for the robustness of the features (**Figure 3.5a and Table 3.1**).

We then examined if the co-expression networks of the top features share similarities between the *in vivo* and *in vitro* datasets. We identified many co-expression relationships maintained across the *in vivo* and *in vitro* datasets. For example, the correlation between

PF3D7_0523000 and PF3D7_1466400 is 0.46 ($p < 2.2e-16$, the smallest value storable in the computer) in the *in vivo* dataset and 0.42 ($p < 2.2e-16$) in *in vitro* dataset. Therefore, we calculated the correlation values of the network weights (i.e., the correlation between genes) for the 22 shared genes. The correlation is 0.55 ($p < 2.2e-16$) indicating strong and conserved co-expression modules involved in ART resistance.

Discussion

In this study, we presented a model that is transferable between *in vivo* measured clearance rate and *in vitro* measured IC50 for ART in malaria treatment and across expression measurement platforms. This is a meaningful step in the research of malaria treatment as the work demonstrated the potential and robustness of a personalized model for ART resistance, which has not been achieved before. Some studies addressed the prediction in either *in vivo* or *in vitro* studies but did not generalize the model across different conditions (Ford & Janies, 2020; D. Li et al., 2021; Sastry et al., 2021). In fact, previous studies reported that generating predictive models for ART resistance has been challenging since the *in vitro* IC50 of P. falciparum in standard drug susceptibility assay correlates poorly with its clearance rate *in vivo (Chotivanich et al., 2014; Fairhurst & Dondorp, 2016)*. Thus the ability of this model to deliver across drastically different sceneria makes this model favorable.

Delivering models between platforms and *in vivo-in vitro* environments has always been a challenge for many medical problems. Several techniques developed in this study may be instructive to other problems. For example, rank normalization of the shared genes in the transcriptomic profiles can potentially help to match two different sets of data and address batch effects. Tree-based algorithms may help interrogate the interactions and overlaps between genes and construct robust models.

We discovered important biomarkers that can be used to create a simplified model for predicting ART resistance. Among them, interesting molecular biomarkers were identified. Pfmdr1 (PF3D7_0523000), *Plasmodium falciparum* multidrug drug resistance gene 1, was identified among the shared top genes by both *in vivo* and *in vitro* datasets, consistent with previous reports stating that it plays an essential role in the response processes of a broad range of ACT antimalarials (Chavchich et al., 2010; Dahlström et al., 2009; Eastman et al., 2016; Gupta et al., 2014; Holmgren et al., 2006, 2007; Imwong et al., 2010; Ngalah et al., 2015; Ould Ahmedou Salem et al., 2017; Sidhu et al., 2006; Sisowath et al., 2007; Ursing et al., 2006). PF3D7_1372000, a Plasmodium exported protein of the Poly-Helical Interspersed Sub-Telomeric (PHIST) protein family (Tarr et al., 2014; Warncke et al., 2016), was also identified among the shared top genes. Literature has reported that the parasite-exported proteins are pivotal for parasite survival by interacting and interfering activities of the infected cells (Maier et al., 2008). Additionally, UBC12, which plays a central role in cell cycle and DNA damage repair (Karpiyevich et al., 2019), was identified, possibly reflecting the mechanism that Plasmodium responds to artemisinin-induced stress by delaying their cell cycle progression and inducing a state of dormancy during early ring-stage development (van Biljon et al., 2018). Other important features whose molecular mechanisms are yet unclear were also identified, pointing to future studies that follow up and validate these new molecular markers for ART resistance.

While our model has achieved satisfying performances on the same population study, we noticed that during the cross-platform prediction, the performance has been impacted severely by the condition of samples in the target datasets, i.e. *in vivo*, *ex viv*o, or *in vitro*, whether treated by DHA, developmental stage (hpi). These observations imply that genes related to ART resistance are expressed differently under different conditions. While many studies have addressed the

dependency between artemisinin resistance with developmental stages (Intharabut et al., 2019; Mok et al., 2011, 2015), *in vitro* environments may also impact the artemisinin resistance phenotype, which needs more experimental assessments in the future.

Furthermore, while top genes were identified in this study, further experimental evidence is still needed to elucidate their roles in artemisinin resistance. For further verification of these biomarkers, gene function perturbations could be carried out on the ART-resistant strains in both *in vivo* and *in vitro* conditions. For example, translation and ubiquitin-activating enzyme inhibitors were found to antagonize the activity of DHA *in vivo* and *in vitro* on *Plasmodium falciparum* strains (Bridgford et al., 2018). Moreover, atovaquone, a mitochondrial electron transport chain inhibitor, could reverse the ART resistance in Cambodian Cam3.II line *in vitro* (Mok et al., 2021). Instead of broad inhibitors that deactivate certain pathways, more targeted gene silencing methods, such as RNAi or CRISPR, would be recommended to inhibit certain top biomarkers, to elucidate the mechanisms of ART resistance.

Materials and Methods

Data and code availability

- All *Plasmodium falciparum* transcriptome data used in this paper have been deposited in GEO and Synapse storage, and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- All original code has been deposited at GitHub and is publicly available as of the date of publication at: https://github.com/GuanLab/Predict-Malaria-ART-Resistance.

Data preprocessing

The *in vivo* prediction model was built based on clinical population data from the published paper by Mok et al. (Mok et al., 2015) and provided by the Malaria DREAM challenge. The P. falciparum isolates were collected ~18 hours post-invasion from 1,043 acute patients under varying treatment and health conditions mainly from Southeast Asia. The parasite isolates transcriptome was analyzed by the Bozdech two-color microarray platform, with 10,159 unique probes covering 5363 genes(Bionetworks, n.d.-a). The Artemisia resistance status of the P. falciparum isolates was labeled as 'Fast' or 'Slow', indicating the clearance rate of P. falciparum after ART treatment.

The test transcriptome data was generated by Agilent HD Exon one-color microarray platform from 30 P. falciparum isolates collected from Thai-Myanar border from 2007 to 2012, as provided by the Malaria DREAM Challenge, which includes 63,976 unique probes covering 5440 genes including non-coding RNS (Bionetworks, n.d.-b). The isolates were cultured in blood cells and treated by artemisinin 6 and 24 hours post-invasion (hpi). The IC50 of P. falciparum culture, i.e., the drug concentration that 50% of parasites die was recorded as an indicator of ART resistance. Higher IC50 means stronger ART resistance, therefore corresponds to a slow clearance rate.

The training and testing microarray data were then processed and normalized by different pipelines with respect to their own microarray platforms (Bionetworks, n.d.-a). The two-color *in vivo* microarray data were processed by GenePix Pro v6.0 software, where features of each array were extracted with foreground intensity > 1.5 fold background intensity for either channel and went through background correction and lowess normalization using the limma R package. Then the arrays were log normalized against co-hybridized 3D7 control, and the gene expression levels

were acquired by averaging their ORF Probe intensities. The *in vitro* single-color microarray data were processed by Agilent Feature extraction and QC pipeline, then quantile normalized by the preprocessCore R package. Then samples were log normalized against NF54 control and batch corrected by the sva R package. Then the gene expression levels were obtained by the reshape R package.

The microarray data usually contains missing values due to artifacts and technical failures. If the expression level of gene i of sample j is missing, we fill in the average gene expression level of gene i, based on the data from the rest of the samples.

$$\overline{x}_i = \frac{1}{N} \sum_{j=0}^{j=N} x_{ij}$$

In order to make a robust cross-platform model, we used rank normalization to process the raw gene expression data, specifically,

N: the total number of samples

m: the total number of genes

x_{ij}: the expression level of gene j in sample i

x_{ir}: the expression level of rth ranked gene in sample i

R_i: the expected expression level of ranked i gene in a sample

$$X_i = \left\{x_{j_1}, x_{j_2}, x_{j_3}, \ldots, x_{j_m}\right\}$$

$$= \left\{x_{r_1}, x_{r_2}, x_{r_3}, \ldots, x_{r_m}\right\} \; where \; x_{r_1} < x_{r_2} < x_{r_3} < \ldots < x_{r_m}$$

$$R_i = \frac{1}{N} \sum_{1}^{N} x_{r_i}$$

$$X_i' = \left\{R_1, R_2, R_3, \ldots, R_m\right\}$$

The microarray record of sample i is transformed from X_i to X_i' . The preprocessed *in vivo* and *in vitro* data was then used in the model training and prediction.

Model training

We tested five types of base learners, including LightGBM, XGboost, random forest, GPR, and linear regression. The first three base learners are tree-based and the later two are kernel-based algorithms. For LightGBM, we used gradient-boosted decision trees, with 5 as the number of leaves, a learning rate of 0.05, and a total of 800 estimators, and 1000 boosting rounds. For random forest, we used a maximal depth of 2 and 100 estimators. For GPR we used dot products and a white kernel. For all other base learners, we used the default parameters. ten-fold cross-validation was used to evaluate the performance of models. The ten *in vivo* models were transferred to *in vitro* data to make predictions of the ART resistance of P. falciparum.

For cross-platform prediction, the shared genes were used in model construction. Each P. falciparum strain was sampled under four different conditions (6 hpi, 24 hpi, with or without ART perturbation), and each sample carried two biological replicates. We conducted cross-platform prediction on the 4 conditions, respectively. For each condition, the average prediction values of the two biological replicates are used as the final prediction.

SHAP feature importance analysis

We conducted SHAP (SHapley Additive exPlanations) analysis to evaluate the contributions of different genes in ART resistance prediction. The SHAP value describes the average marginal contribution of a feature across all instances (S. Lundberg & Lee, 2017). We summed up the absolute values of SHAP values of all samples for each feature. The summary plot sorting features by the sum of the absolute SHAP values over all samples are included in **Supplementary Figure 3.2-3**.

Coexpression and functional analysis of top genes

We conducted co-expression analysis on rank normalized gene expression level among the top-ranked genes by SHAP analysis, for both *in vivo* and *in vitro* datasets. The co-expression significance between two genes is defined as Pearson's correlation of their normalized expression level across all samples. For example, for gene i and j in all N samples, X_i and X_j refer to the rank normalized expression level of both genes, respectively. Then,

$$X_i = \left\{ x_{i1,} x_{i2}, \ \ldots x_{in} \right\}$$
$$X_j = \left\{ x_{j1,} x_{j2}, \ \ldots x_{jn} \right\}$$

Where n refers to the total number of samples in the dataset. The co-expression level $r_{i,j}$ between two genes is:

$$r_{i,j} = cor\left(X_i, X_j \right)$$

Where $r_{i,j}$ is the Pearson's correlation between gene i and gene j. The co-expression networks of both *in vivo* and *in vitro* datasets were constructed based on the significantly correlated genes ($r_{i,j}$>0.4) and visualized using *ggraph*.

Quantification and statistical analysis

Because *in vivo* data and bore binary labels, we used AUROC (Area under the Receiver Operating Curve) and AUPRC (Area under the Precision Recall Curve). For the *in vitro* data, because the evaluation is a real value, we used Spearman and Pearson's correlations and C-index, as clearance rate and IC50 do not share the same distribution (**Figure 3.2a and b**). The C-index is calculated as the following:

$$C-index = \frac{\sum_{i,j} 1_{p_i<p_j} \cdot 1_{IC50_i<IC50_j}}{\sum_{i,j} 1_{p_i<p_j}}$$

p_i: the predicted value of sample i, ranges from 0 to 1.

$IC50$: the IC50 of sample i.

$1_{p_i<p_j} = 1$ if $p_i < p_j$, else 0

$1_{IC50_i<IC50_j} = 1$ if $IC50_i < IC50$, else 0.

C-index is equivalent to AUROC when predicting binary labels.

For external validation datasets, the labels were also binary, thus we use AUROC for performance evaluation. AUPRC was not used for horizontal comparison since the baseline for each dataset is different. 95% confidence intervals of all performances were calculated by bootstrapping.

Pearson and Spearman's correlation coefficient, AUROC, and AURPC were calculated using the Python Sklearn module. The code implementing the c-index was provided in the GitHub repository (see **Data and Code Availability** in the **Resource Availability** section).

Figures

Figure 3.1. Study design. a. Demonstration of the training data given by the DREAM Challenge. b. Strategy of training *in vivo* malaria ART prediction models, and transferring the model to *in vitro* malaria transcriptome datasets. First, we imputed missing values and rank-normalized the expression data. Second, we cross-validated models of different base learners. We then selected the base learner and sample conditions with the best performance by cross-validations and reverse tests. Lastly, important predictive biomarkers are prioritized by SHAP analysis.

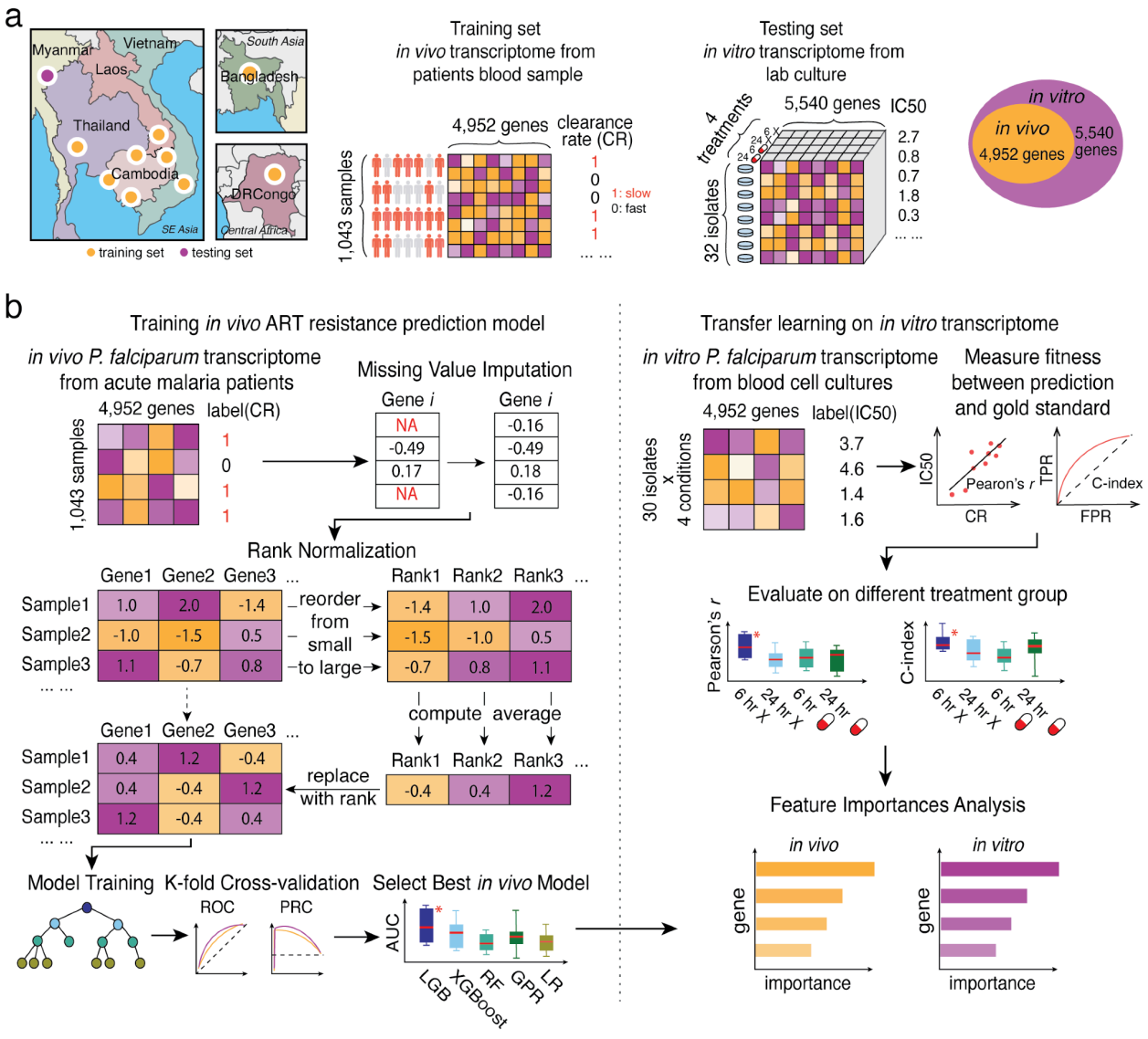


Figure 3.2. Model performances across platforms. a and b: Distribution of ART resistance measurement labels in the *in vivo* and the *in vitro* datasets. c. Cross-validation performance in the *in vivo* dataset. d. Performance of transferring the model trained on the *in vivo* dataset to the *in vitro* dataset, presented as the correlation between prediction and gold-standard (IC50)) under 4 conditions. e. Performance of transfer learning with/without rank normalization.

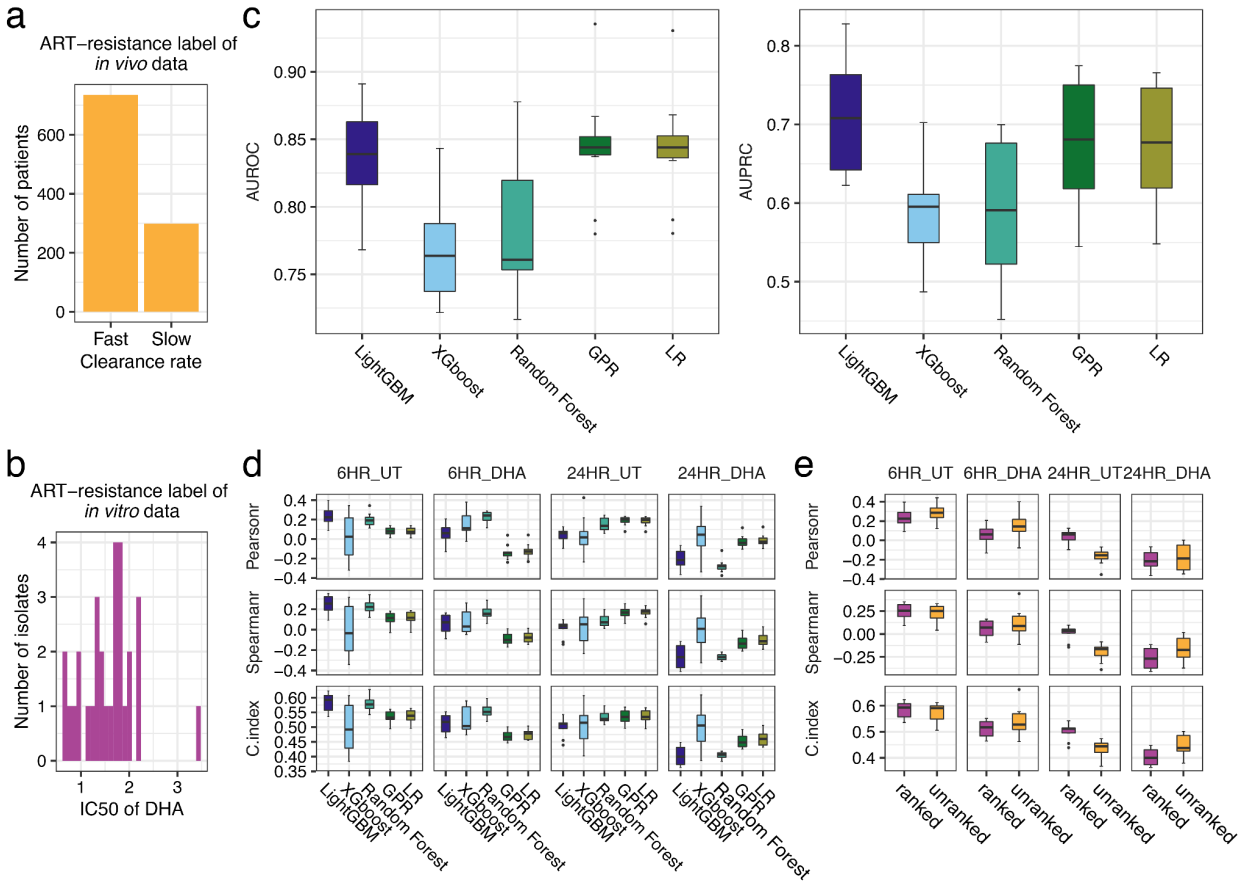


Figure 3.3. Top genes related to malaria ART resistance as identified by SHAP feature importance analysis, and performances of machine learning model after feature selection using the top-ranked genes. a. Top 30 genes of ART resistance prediction model visualized by SHAP analysis based on *in vivo P. falciparum* transcriptome. b. Top 30 genes of ART resistance prediction model visualized by SHAP analysis based on *in vitro P. falciparum* transcriptome. Genes were ordered by mean SHAP contributions across all test examples in a ten-fold cross-validation. c. Model performances for *in vivo* and *in vitro* predictions when including only top genes selected by SHAP analysis, as evaluated by AUROC (for binary labels) and C-index (for continuous labels). 'All genes' shows prediction performance without feature selection. d. Comparison of *in vitro* prediction performances between using all top ten genes as features and leaving one gene out at a time.

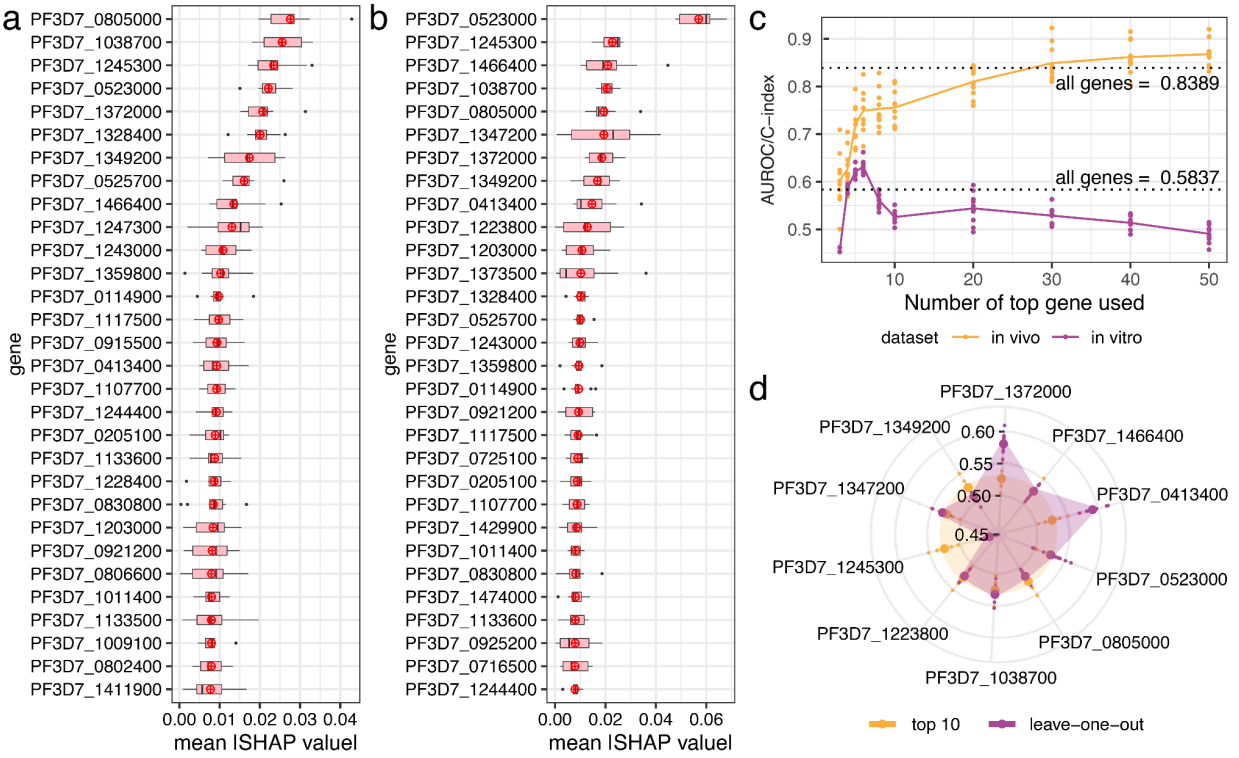


Figure 3.4. Cellular functions of top contributing genes in predicting ART resistance. a. The functions of top contributing genes and their relationship with ART resistance. b. The functionally enriched protein clusters that interact with PF3D7_052300 and PF3D7_1245300. The prefix "PF3D7_" of these gene IDs is omitted and only the numbers are shown for simplicity.

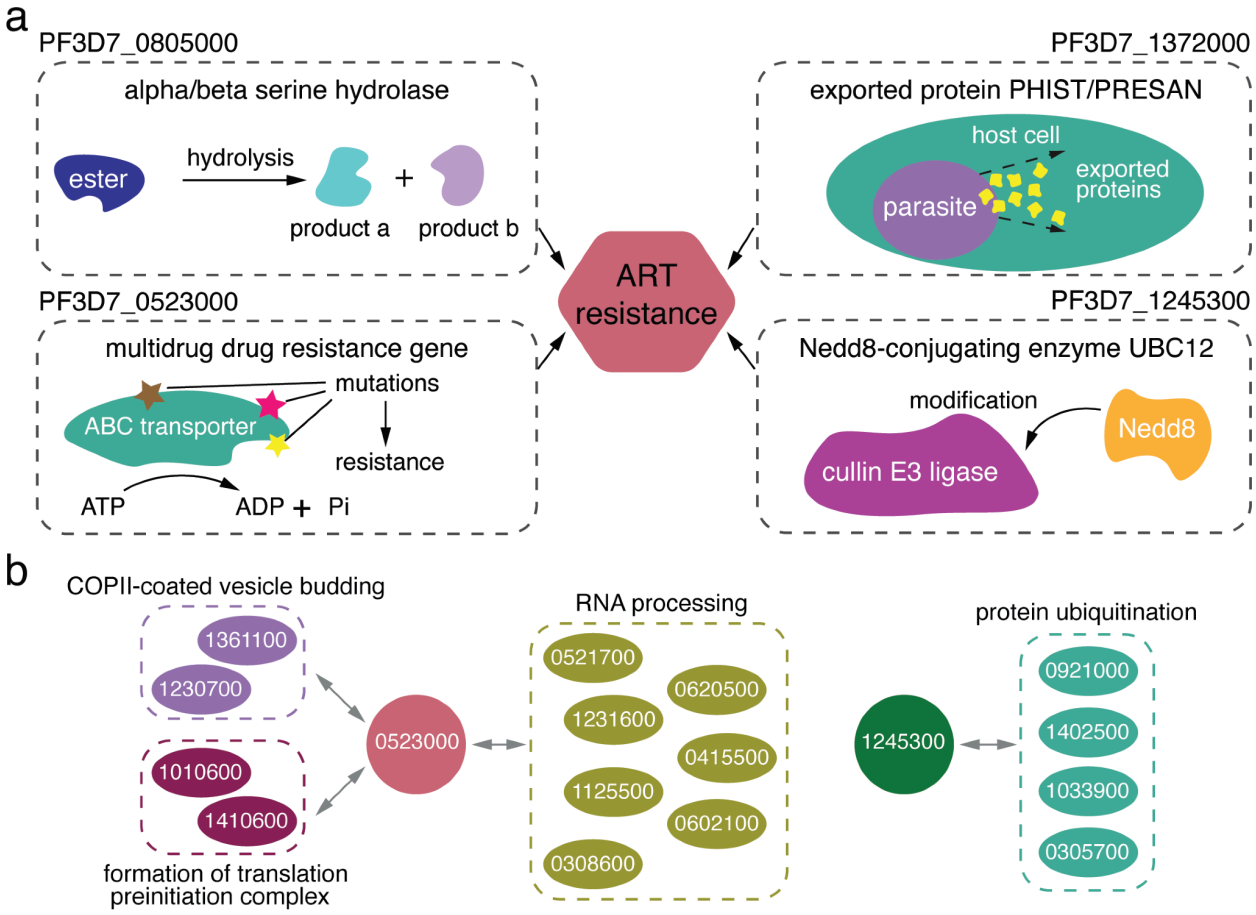
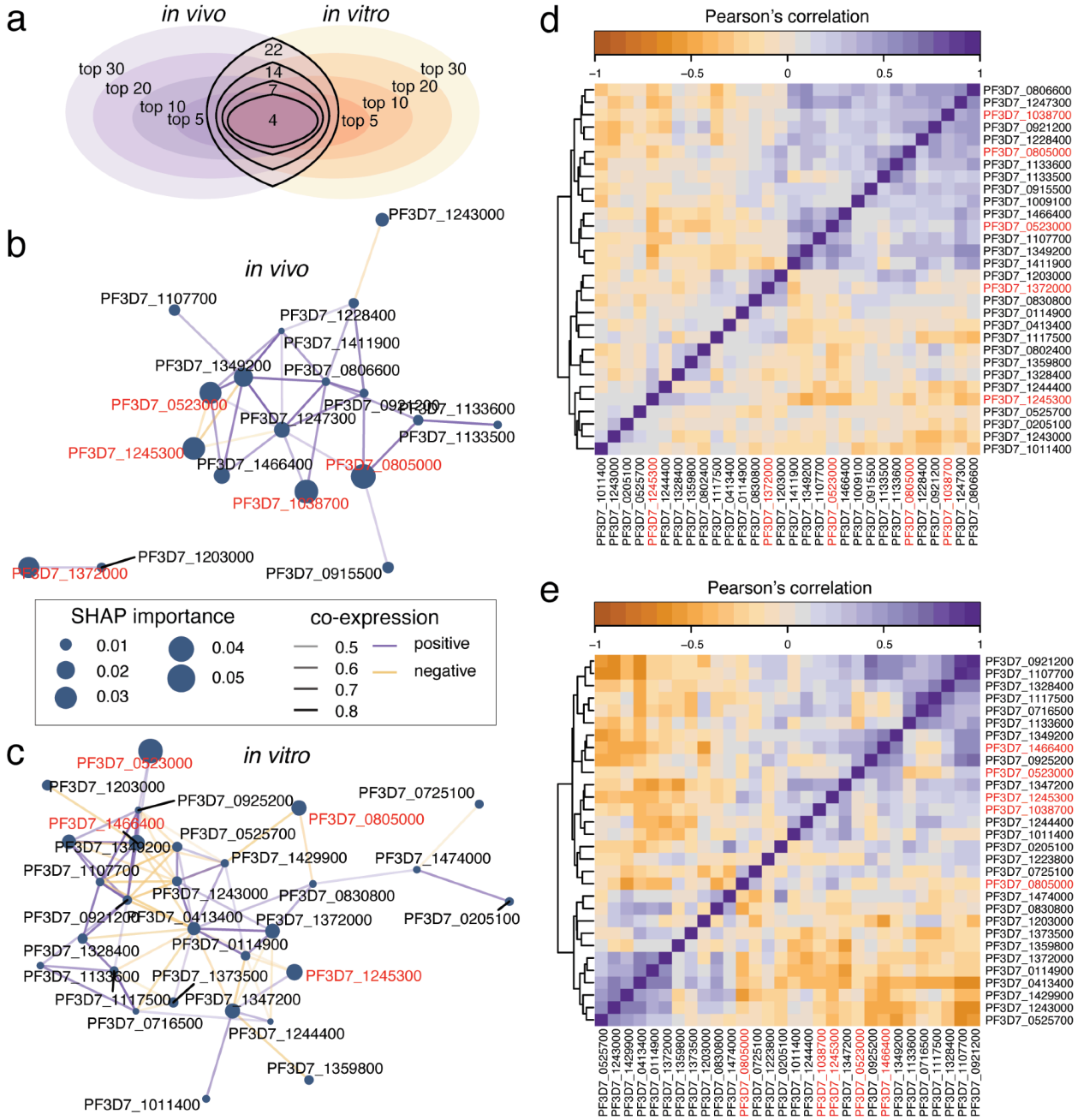


Figure 3.5. Coexpression networks of top genes in *in vivo* and *in vitro* datasets. a. Sharing of top genes across in vivo/in vitro datasets. b and d. The co-expression network and the co-expression matrix of the top 30 genes in the *in vivo* dataset. c and e. The co-expression network and co-expression matrix of the top 30 genes in the *in vitro* dataset. We retained all co-expression relationships with an absolute correlation value > 0.4 in the plot. The five most important genes in either *in vivo* or *in vitro* dataset were marked as red (except PF3D4_1038700, which was not shown in c, since there were no other genes that shared a significant correlation (|r| > 0.4) with this gene).



Tables

Table 3.1. 22 shared features (among the top 30) between the *in vivo* and the *in vitro* datasets.

gene id	*in vivo* SHAP importance	*in vitro* SHAP importance	Annotations
PF3D7_0805000	0.027577556	0.019392157	alpha/beta hydrolase, putative
PF3D7_1038700	0.025549987	0.020552675	Plasmodium exported protein, unknown function
PF3D7_1245300	0.023449244	0.022728070	NEDD8-conjugating enzyme UBC12, putative
PF3D7_0523000	0.022131747	0.056972396	multidrug resistance protein 1
PF3D7_1372000	0.020715635	0.018587311	Plasmodium exported protein (PHISTa), unknown function
PF3D7_1328400	0.020094142	0.010121155	conserved protein, unknown function
PF3D7_1349200	0.017402912	0.016821882	glutamate--tRNA ligase, putative
PF3D7_0525700	0.016087702	0.010102856	conserved protein, unknown function
PF3D7_1466400	0.013496112	0.020929573	AP2 domain transcription factor AP2-EXP
PF3D7_1243000	0.01083543	0.009874289	syntaxin-16, putative
PF3D7_1359800	0.010260771	0.009438231	ADP-ribosylation factor, putative
PF3D7_0114900	0.00976088	0.009389257	Plasmodium exported protein, unknown function, pseudogene
PF3D7_1117500	0.009725304	0.009184379	tyrosine--tRNA ligase
PF3D7_0413400	0.009276499	0.014652890	erythrocyte membrane protein 1 (PfEMP1), exon 1, pseudogene
PF3D7_1107700	0.009240973	0.008681170	pescadillo homolog
PF3D7_1244400	0.009191554	0.007895322	RNA-binding protein, putative
PF3D7_0205100	0.008908881	0.008790011	conserved Plasmodium protein, unknown function
PF3D7_1133600	0.0087618	0.008062268	conserved Plasmodium protein, unknown function
PF3D7_0830800	0.00846947	0.008136531	surface-associated interspersed protein 8.2 (SURFIN 8.2)
PF3D7_1203000	0.008352643	0.010673660	origin recognition complex subunit 1
PF3D7_0921200	0.008144911	0.009305749	conserved Plasmodium membrane protein, unknown function
PF3D7_1011400	0.007971128	0.008214462	proteasome subunit beta type-5

Supplementary Tables

Supplementary Table 3.1. Performance of ART-resistance prediction models on *in vivo* data in ten-fold cross-validation, related to Figure 3.2.

Model	Pearson's r mean[95CI]	Spearman's r mean[95CI]	AUROC mean[95CI]	AUPRC mean[95CI]
LightGBM	**0.5616[0.5136, 0.6153]**	0.5315[0.4845, 0.5753]	0.8384[0.8121, 0.8705]	**0.6983[0.6438, 0.7522]**
LightGBM (without kelch 13)	0.5484[0.5008, 0.6019]	0.5217[0.4659, 0.5680]	0.8322[0.8027, 0.8658]	0.6813[0.6264, 0.7374]
LightGBM (no normalization)	0.5293[0.4804, 0.5723]	0.5084[0.4631, 0.5485]	0.8237[0.7977, 0.8483]	0.6613[0.5994, 0.7234]
XGboost	0.4377[0.3703, 0.4837]	0.4191[0.3557, 0.4592]	0.7669[0.7262, 0.7910]	0.5752[0.5049, 0.6387]
Random Forest	0.4528[0.3909, 0.5063]	0.4369[0.3777, 0.4812]	0.7782[0.7441, 0.8099]	0.5797[0.5088, 0.6547]
GPR	0.5406[0.4986, 0.5753]	**0.5428[0.5029, 0.5818]**	**0.8456[0.8212, 0.8673]**	0.6742[0.6198, 0.7280]
Linear Regression	0.5390[0.4968, 0.5739]	0.5415[0.5015, 0.5809]	0.8448[0.8206, 0.8668]	0.6717[0.6176, 0.7252]

Supplementary Table 3.2. Prediction performance of ART-resistance during transfer validation on *in vitro* data, related to Figure 3.2.

Model	*in vitro* Test Data	Pearson's r mean[95CI]	Spearman's r mean[95CI]	C-index mean[95CI]
LightGBM	24HR_DHA	-0.1947[-0.2872, -0.0612]	-0.2502[-0.3388, -0.0935]	0.4069[0.3761, 0.4613]
LightGBM	24HR_UT	0.0411[-0.0451, 0.1872]	0.0139[-0.0821, 0.1518]	0.5060[0.4746, 0.5517]
LightGBM	6HR_DHA	0.0623[-0.0299, 0.1860]	0.0625[-0.0232, 0.1912]	0.5146[0.4856, 0.5590]
LightGBM	6HR_UT	0.2319[0.1379, 0.3206]	**0.2467[0.1457, 0.3548]**	**0.5837[0.5474, 0.6216]**
LightGBM (no kelch 13)	24HR_DHA	-0.1912[-0.2754, -0.0635]	-0.2380[-0.3450, -0.1143]	0.4148[0.3799, 0.4591]
LightGBM (no kelch 13)	24HR_UT	0.0406[-0.0341, 0.1458]	0.0241[-0.0727, 0.1589]	0.5121[0.4806, 0.5582]
LightGBM (no kelch 13)	6HR_DHA	0.0418[-0.0557, 0.1840]	0.0463[-0.0485, 0.1650]	0.5143[0.4815, 0.5516]
LightGBM (no kelch 13)	6HR_UT	0.2010[0.0968, 0.3197]	0.1836[0.0695, 0.2841]	0.5636[0.5237, 0.5980]
LightGBM(no normalization)	24HR_DHA	-0.1445[-0.2405, -0.0103]	-0.1507[-0.2306, -0.0080]	0.4488[0.4205, 0.4983]
LightGBM(no normalization)	24HR_UT	-0.1400[-0.2404, -0.0356]	-0.1678[-0.2901, -0.0566]	0.4446[0.4073, 0.4809]
LightGBM(no normalization)	6HR_DHA	0.1165[0.0135, 0.2762]	0.0898[-0.0068, 0.2405]	0.5307[0.4972, 0.5838]
LightGBM(no normalization)	6HR_UT	**0.2449[0.1419, 0.3727]**	0.2040[0.0961, 0.3316]	0.5690[0.5310, 0.6148]
XGboost	24HR_DHA	0.0075[-0.0942, 0.1283]	0.0026[-0.0977, 0.1421]	0.5009[0.4669, 0.5488]
XGboost	24HR_UT	0.0258[-0.0563, 0.1668]	0.0302[-0.0572, 0.1736]	0.5110[0.4809, 0.5615]
XGboost	6HR_DHA	0.1178[-0.0205, 0.2543]	0.0703[-0.0456, 0.1931]	0.5241[0.4815, 0.5671]
XGboost	6HR_UT	0.0364[-0.0622, 0.1839]	0.0018[-0.1044, 0.1425]	0.5003[0.4647, 0.5491]
Random Forest	24HR_DHA	-0.2676[-0.3329, -0.1839]	-0.2634[-0.3351, -0.1458]	0.4068[0.3794, 0.4473]
Random Forest	24HR_UT	0.1499[0.0657, 0.2480]	0.0830[-0.0028, 0.1931]	0.5329[0.5038, 0.5683]
Random Forest	6HR_DHA	0.2201[0.1059, 0.3638]	0.1536[0.0423, 0.3109]	0.5519[0.5132, 0.6063]
Random Forest	6HR_UT	0.1849[0.0740, 0.2982]	0.2120[0.1199, 0.3321]	0.5754[0.5427, 0.6208]
Gaussian process regression	24HR_DHA	-0.0225[-0.1508, 0.1557]	-0.1160[-0.2300, 0.0503]	0.4576[0.4180, 0.5147]
Gaussian process regression	24HR_UT	0.1669[0.0846, 0.2526]	0.1462[0.0464, 0.2693]	0.5440[0.5060, 0.5886]
Gaussian process regression	6HR_DHA	-0.1243[-0.2086, -0.0308]	-0.0893[-0.1749, 0.0192]	0.4709[0.4403, 0.5088]
Gaussian process regression	6HR_UT	0.0691[-0.0030, 0.1363]	0.0831[-0.0063, 0.1852]	0.5283[0.4986, 0.5578]
Linear Regression	24HR_DHA	-0.0138[-0.1486, 0.1590]	-0.0910[-0.2037, 0.0751]	0.4681[0.4268, 0.5258]
Linear Regression	24HR_UT	0.1618[0.0717, 0.2481]	0.1326[0.0323, 0.2260]	0.5377[0.5016, 0.5734]
Linear Regression	6HR_DHA	-0.1117[-0.2032, -0.0138]	-0.0707[-0.1720, 0.0403]	0.4774[0.4436, 0.5135]
Linear Regression	6HR_UT	0.0667[-0.0139, 0.1581]	0.0845[-0.0159, 0.1741]	0.5277[0.4961, 0.5605]

Supplementary Table 3.3. ART-resistance datasets used in this study, related to STAR methods.

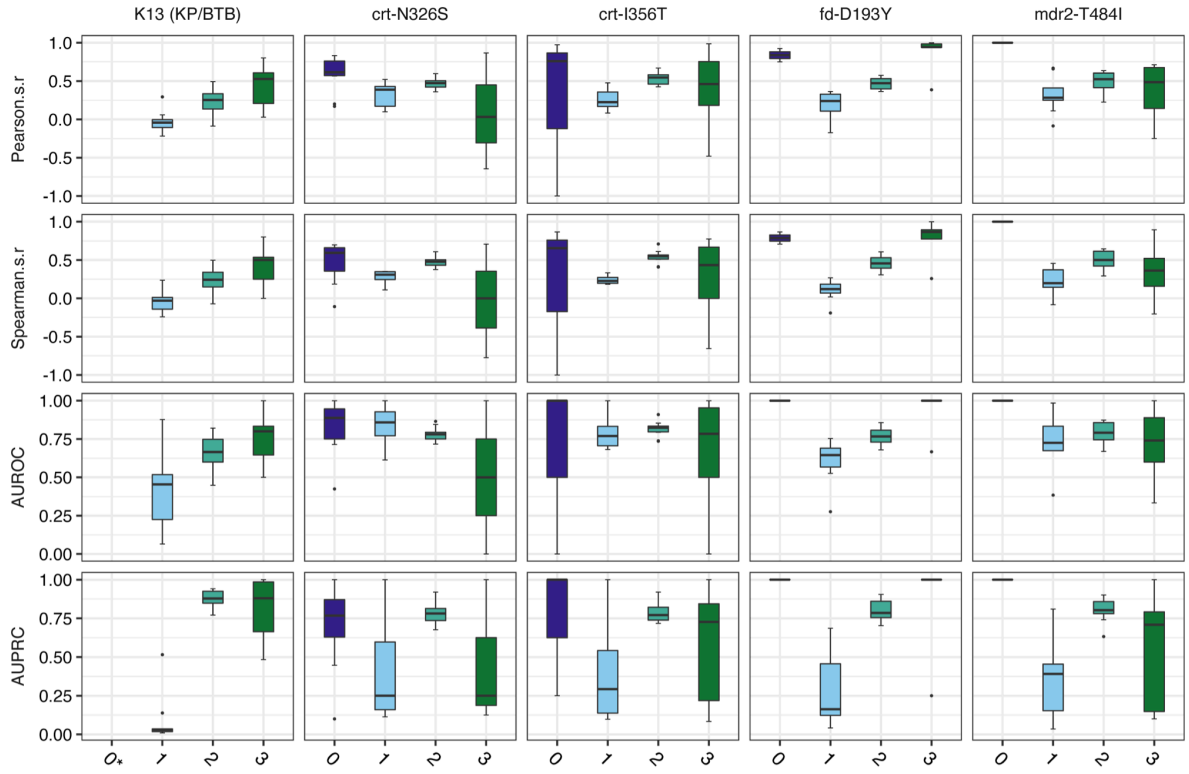
reference	#isolates	#samples	condition	GEO
datasets for original training and testing				
(Mok et al., 2015)	1,043	1,043	*in vivo*	GSE59099
(Bionetworks, n.d.-a)	32	128	*in vitro*	N/A
datasets for external validation				
(Mok et al., 2011)	11	91	*ex vivo*	GSE25878
(Mok et al., 2015)	19	110	*ex vivo*	GSE59098
(Shaw et al., 2015)	4	29	*in vitro*	GSE61536
(Mok et al., 2021)	5	156	*in vitro*	GSE151189

Supplementary Table 3.4. Performance of ART-resistance prediction models on *in vivo* data within different genetic variation cohorts. genotype: 0: information missing; 1: reference sequence of PF3D7; 2: mutation 3: heterozygous., related to Figure 3.2.

gene	genotype	Pearson's r[95CI]	Spearman's r[95CI]	AUROC[95CI]	AUPRC[95CI]
K13 (KP/BTB)	1	-0.0289[-0.0943, 0.1004]	-0.0364[-0.1071, 0.0865]	0.4382[0.3021, 0.6690]	0.0272[0.0118, 0.0725]
K13 (KP/BTB)	2	0.2354[0.1348, 0.3715]	0.2373[0.1446, 0.3677]	0.6629[0.6025, 0.7580]	0.8648[0.8196, 0.9177]
K13 (KP/BTB)	3	0.4495[0.3084, 0.6174]	0.4521[0.2997, 0.6265]	0.7621[0.6731, 0.8664]	0.7615[0.6573, 0.8849]
crt-N326S	0	0.6238[0.4936, 0.7817]	0.4990[0.3423, 0.6744]	0.8437[0.7503, 0.9553]	0.7358[0.5860, 0.8872]
crt-N326S	1	0.2467[0.1411, 0.3887]	0.2120[0.1308, 0.3147]	0.7835[0.6874, 0.9165]	0.1749[0.0962, 0.4198]
crt-N326S	2	0.4696[0.3978, 0.5313]	0.4747[0.3995, 0.5417]	0.7741[0.7307, 0.8128]	0.7766[0.7147, 0.8273]
crt-N326S	3	0.4392[-0.1846, 0.9313]	0.2584[-0.2757, 0.7249]	0.7000[0.2875, 1.0000]	0.5190[0.0769, 1.0000]
crt-I356T	0	0.5467[0.1409, 0.9012]	0.2866[0.0000, 0.7184]	0.7500[0.5000, 1.0000]	0.7777[0.3098, 1.0000]
crt-I356T	1	0.2104[0.1135, 0.3576]	0.1931[0.1168, 0.3024]	0.7642[0.6741, 0.8615]	0.1361[0.0622, 0.3907]
crt-I356T	2	0.5328[0.4844, 0.6142]	0.5388[0.4948, 0.6241]	0.8120[0.7862, 0.8615]	0.7793[0.7370, 0.8439]
crt-I356T	3	0.3995[0.0687, 0.7087]	0.3053[-0.0114, 0.6133]	0.7188[0.4896, 0.9154]	0.5486[0.2440, 0.8988]
fd-D193Y	0	0.6794[0.4418, 0.8443]	0.4809[0.2932, 0.7223]	0.9892[0.9586, 1.0000]	0.9028[0.3750, 1.0000]
fd-D193Y	1	0.2126[0.0760, 0.3697]	0.1095[0.0188, 0.2428]	0.6233[0.5204, 0.7775]	0.2832[0.1293, 0.4755]
fd-D193Y	2	0.4557[0.3592, 0.5437]	0.4563[0.3572, 0.5465]	0.7662[0.7094, 0.8180]	0.7949[0.7389, 0.8486]
fd-D193Y	3	0.5023[0.1882, 0.7961]	0.4143[0.0982, 0.7543]	0.7760[0.5778, 1.0000]	0.6882[0.3842, 1.0000]
mdr2-T484I	0	0.4845[0.3873, 1.0000]	0.7746[0.7746, 1.0000]	1.0000[1.0000, 1.0000]	1.0000[1.0000, 1.0000]
mdr2-T484I	1	0.3203[0.1684, 0.4385]	0.2173[0.1082, 0.3397]	0.7483[0.6322, 0.9107]	0.3293[0.1734, 0.5015]
mdr2-T484I	2	0.4941[0.4170, 0.5805]	0.5048[0.4246, 0.5829]	0.7921[0.7460, 0.8383]	0.7986[0.7566, 0.8617]
mdr2-T484I	3	0.4223[0.1895, 0.5858]	0.3449[0.1198, 0.5081]	0.7533[0.5861, 0.9069]	0.5443[0.3642, 0.7420]

Supplementary Figures

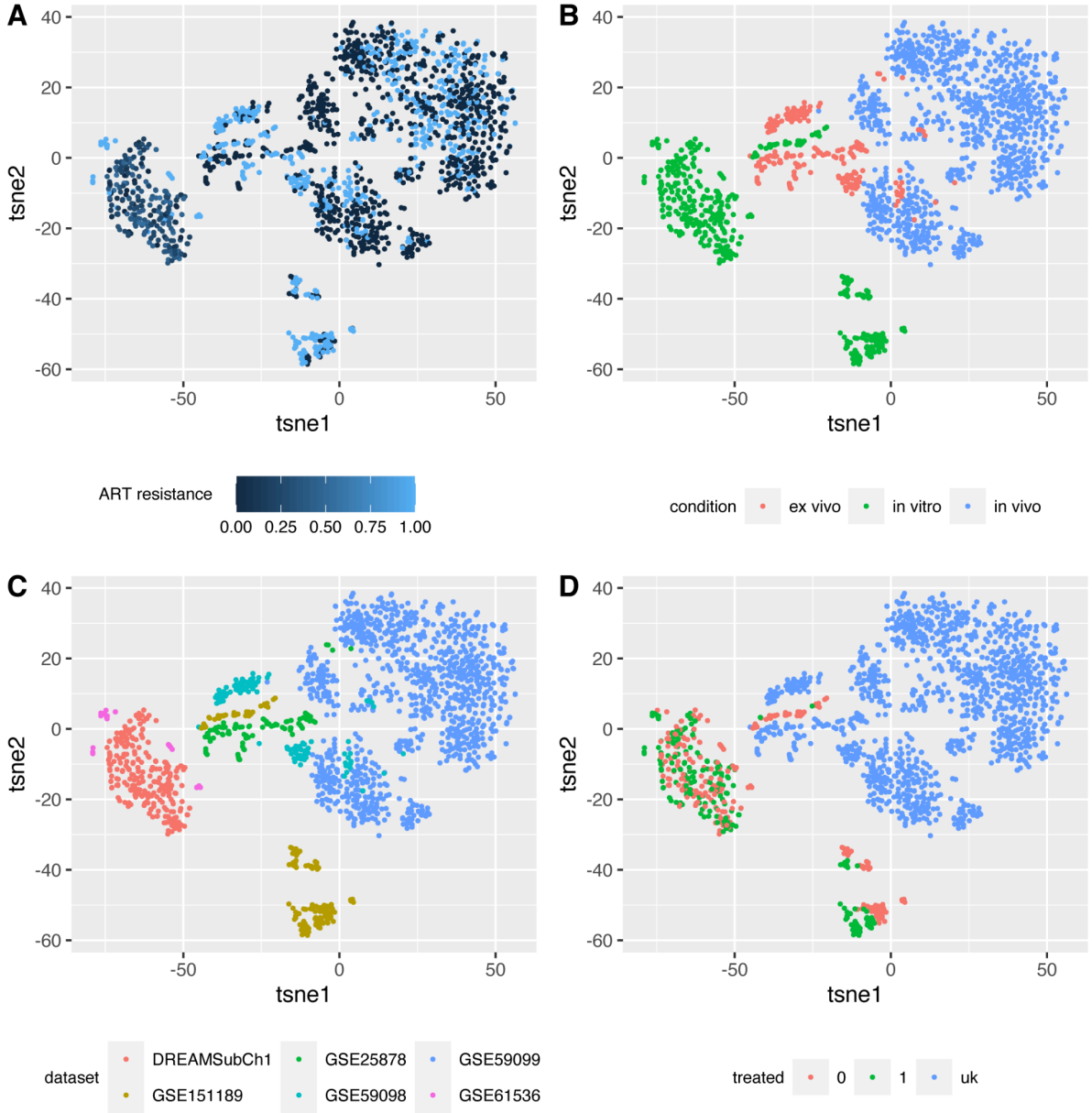
Supplementary Figure 3.1. ART-resistance prediction performances across different genetic variation cohorts. genotype: 0: information missing; 1: reference sequence of PF3D7; 2: mutation 3: heterozygous. *for K13, there's only one sample in the "0" genotype subgroup, therefore there are no evaluation results, related to **Figure 3.2**.



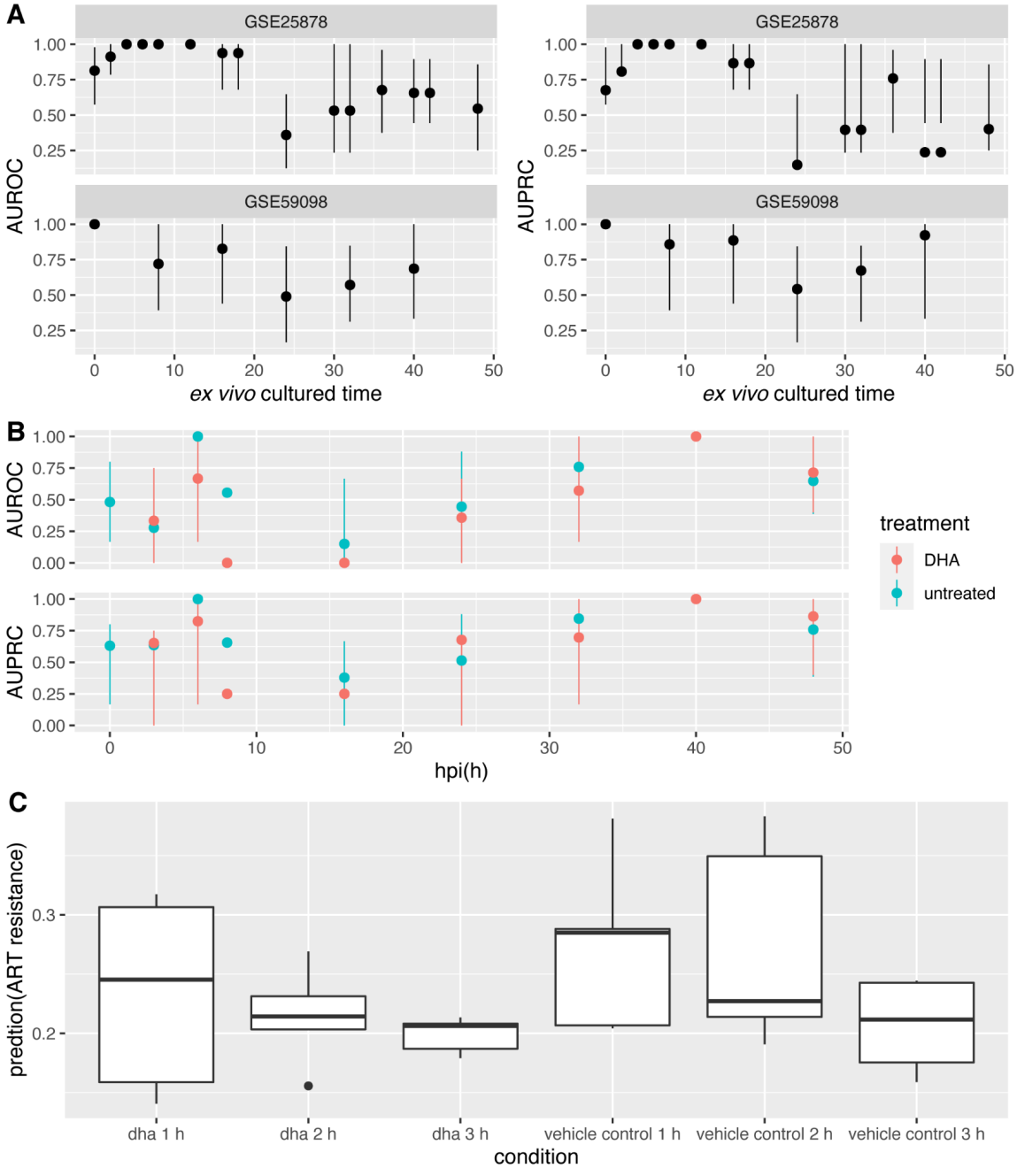
Supplementary Figure 3.2. SHAP Summary plot of *in vivo* data based on ten-fold cross-validation, related to Figure 3.3.

Supplementary Figure 3.3. SHAP Summary plot of *in vitro* data based on ten-fold cross-validation, related to Figure 3.3.

Supplementary Figure 3.4. t-Distributed stochastic neighbor embedding(t-SNE) analysis on the transcriptome data of all datasets used in this study, related to STAR Methods. The first two dimensions were shown in the scatter plots. Each data point, or transcriptome profile of a sample, is denoted by a different color showing their: **A.** ART resistance; **B.** sample condition (*in vivo*, *ex vivo*, or *in vitro*) **C.** dataset name. Except for the DREAM challenge dataset (denoted by DREAMSubCh1), all other datasets are denoted by their GEO accession number. **D.** DHA treatment status (treated, untreated, or unknown), respectively.



Supplementary Figure 3.5. Prediction performance of the *in vivo* model on publicly available datasets besides DREAM Challenge, related to Figure 3.2. A. cross-platform prediction on *ex vivo* datasets GSE25878 and GSE95098, where the prediction performances (AUROC and AUPRC) of each *ex vivo* cultured time point were shown. B. cross-platform prediction results on *in vitro* dataset GSE151189, and prediction performances of each developmental stage (hpi) were shown. DHA treatment status was denoted by red and green. C. cross-platform prediction results on *in vitro* dataset GSE61536. While all samples in this dataset were collected from the K1 strain, which is known for chloroquine resistance instead of artemisinin, we show the predicted ART resistance instead of AUROC or AUPRC. The predicted results at different treatment statuses (DHA/vehicle control for 1-3 hours) were shown separately.



CHAPTER IV: Mapping Combinatorial Drug Effects to DNA Damage Response Kinase Inhibitors

Abstract

One fundamental principle that underlies various cancer treatments, such as traditional chemotherapy and radiotherapy, involves the induction of catastrophic DNA damage, leading to the apoptosis of cancer cells. In our study, we conduct a comprehensive dose-response combination screening focused on inhibitors that target key kinases involved in the DNA damage response (DDR): ATR, ATM, and DNA-PK. This screening involves 87 anti-cancer agents, including six DDR inhibitors, and encompasses 62 different cell lines spanning 12 types of tumors, resulting in a total of 17,912 combination treatment experiments. Within these combinations, we analyze the most effective and synergistic drug pairs across all tested cell lines, considering the variations among cancers originating from different tissues. Our analysis reveals inhibitors of five DDR-related pathways (DNA topoisomerase, PLK1 kinase, p53-inducible ribonucleotide reductase, PARP, and cell cycle checkpoint proteins) that exhibit strong combinatorial efficacy and synergy when used alongside ATM/ATR/DNA-PK inhibitors.

Introduction

Cancers are aggressive, invasive diseases characterized by uncontrolled growth. Many cancers exhibit genome instability resulting from tumor-specific DNA repair defects and increased replication stress, making them more susceptible than normal tissues to DNA damage, such as single and double-strand breaks (SSBs and DSBs, respectively) (Hanahan & Weinberg,

2011; Jackson & Bartek, 2009). Taking advantage of this vulnerability, DNA-damaging treatments such as ionizing radiation and platinum-based antineoplastic have long been used as anti-cancer treatments (Baskar et al., 2012; Cohen & Lippard, 2001). More recently, a suite of therapeutic agents targeting DNA damage response (DDR) pathways has been developed that specifically exploits this susceptibility, promising reduced side effects compared to non-targeted treatments (Kelley & Fishel, 2016; Lord & Ashworth, 2012; McLaren et al., 2016; O'Connor, 2015). In this context, it is hypothesized that the simultaneous deactivation of multiple DDR pathways could lead to improved treatment efficacy by addressing both acquired treatment resistance and buffering by parallel DDR pathways (Jackson & Bartek, 2009).

A set of 450 proteins involved in different pathways of the DNA damage response has recently been mapped (Pearl et al., 2015). While it is commonly assumed that specific pathways exist that address different types of DNA damage, e.g., for SSBs, DSBs, or mismatch repair, loss of function of a DDR pathway can be compensated by parallel repair pathways (Ciccia & Elledge, 2010; Jackson & Bartek, 2009). The simultaneous inhibition of multiple complementary DDR pathways by somatic mutation in the tumor and/or one or more targeted treatments, such as the synthetic lethality between PARP1 inhibition and BRCA1 loss of function (Bryant et al., 2005; Farmer et al., 2005), was therefore identified as a promising therapeutic strategy in clinical cancer treatments. This strategy also inspired the development of combination treatments of multiple DDR inhibitors to overcome resistance to single drugs, achieve synergistic effects, and expand DDR drugs' usage to other indications beyond BRCA-deficient cancers (O'Connor, 2015; Pilié et al., 2019).

Three canonical DNA damage-sensing kinases that are central to the human DDR are ataxia telangiectasia mutated (ATM), ataxia telangiectasia and Rad3-related (ATR), and

DNA-dependent protein kinase (DNA-PK), which is also referred to as protein kinase, DNA-activated, catalytic subunit (PRKDC) (Balmus et al., n.d.; Blackford & Jackson, 2017). So far, studies that comprehensively map the synergistic effects between small molecule inhibitors of these key DDR kinases and other anti-cancer drugs are lacking in both coverages across tumor types and the number of combination therapy partners. In this study, we generated cancer cell line drug combination screens of six kinase inhibitors, including two ATM inhibitors (M3541 and M4076 (Zimmermann, Zenke, et al., 2022)), three ATR inhibitors (berzosertib (Hall et al., 2014; Reaper et al., 2011), gartisertib (Jo et al., 2021), M1774 (Zimmermann, Dahmen, et al., 2022)), and one DNA-PK inhibitor (peposertib (van Bussel et al., 2021; Zenke et al., 2020)) against 87 anti-cancer drugs of a wide range of mode-of-actions on 22~62 cancer cell lines across 12 tissues (or tumor types), forming a total of 17,912 combination treatment experiments.

In order to characterize tissue-specific patterns of DDR inhibitor combination treatments, we carried out full-genome and transcriptomic profiling of all 62 cell lines and statistically associated dose responses with genomic and transcriptomic readouts. This screen represents a large DDR inhibitor combination study and allowed us to identify a small set of inhibitors to proteins involved in five pathways that displayed strong co-therapeutic efficacy and synergy with ATM/ATR/DNA-PK inhibition globally: the DNA topoisomerase pathway, the serine/threonine-protein kinase PLK1 pathway, the p53-inducible ribonucleotide reductase pathway, the PARP pathway, and the cell cycle checkpoint proteins.

Results

The experimental dose-response screen of three DDR inhibitors across a wide range of anti-cancer combination treatments

The goal of this study was to comprehensively analyze the synergistic relationship between the inhibitors of canonical DDR kinases (ATM, ATR, and DNA-PK) and a panel of anti-cancer drugs. In total, we combined six kinase inhibitors, including two ATM inhibitors (M3541 and M4076), three ATR inhibitors (berzosertib, gartisertib, M1774), and one DNA-PK inhibitor (peposertib) with 87 anti-cancer drugs, on 62 cancer cell lines covering 12 tissues or tumor types (**Figure 4.1 a, Supplementary Table 4.1 and Data Availablity**). For each of the cell lines, we carried out RNA- and whole-genome DNA sequencing, and derived genome-wide readouts covering gene expression, copy-number profiling, and loss-of-function mutation both for single genes as well as biological pathways.

In vitro combination treatment responses were quantified on the level of both efficacy and synergy. The efficacy of treatment was estimated by the area above the parametric dose-response curve divided by the sum of the areas above and below this curve, a quantity that we denote as relative *AoC score*. The synergy between two combination partners within treatment was measured by the *Bliss score*, which reflects the additional effect of two drugs over the expected response if the two drugs were to act independently (see **Methods** Section for detailed discussions of the dose-response experimental setup, cell line sequencing, and computation of response measures).

In total, we generated 17,912 combination treatment experiments and 7,081 monotherapy experiments, with reproducibility of Pearson's correlation = 0.8380 (p < 1e-22) in AoC score for monotherapy and 0.7611 (p < 1e-22) in Bliss score for combination treatment, which is

comparable with previously reported combination treatment screening datasets including DREAM (Menden et al., 2019), ALMANAC (Holbeck et al., 2017; Menden et al., 2019), and O'Neil (O'Neil et al., 2016). While various DDR inhibitor combinations were used in our screens, we report results on the level of *mode-of-action combination* (e.g., ATMi-PARPi) for conciseness and generalizability. However, all analyses were conducted using and are supported by *individual drug combinations* (such as M3541-olaparib).

Mapping the global interaction relationships between DDR inhibitors and combination treatment partners

In anti-cancer treatment, ideal drug combinations are not only safe and effective but also complement each other in a synergistic manner (O'Connor, 2015). Due to the complex relationships between DDR pathways (Jackson & Bartek, 2009), finding optimal drug combinations that show broad efficacy across multiple tumor types and genomic contexts of tumors is particularly challenging. This large-scale screen, therefore, provides a unique opportunity to map the overall global efficacy and synergy relationships between DDR inhibitors and other anti-cancer agents.

To visualize these global relationships, we generated comprehensive heatmaps showing the efficacy and synergy responses of all 87 anti-cancer drugs screened in combination with six ATM/ATR/DNA-PK inhibitors across all 62 cell lines and 12 tissues (**Figure 4.1b, Supplementary Figure 4.1-3**). By visual and numerical analysis, we identified several drugs that result in high efficacy when combined with ATM, ATR, and DNA-PK inhibitors. In general, ATR inhibitors have stronger synergy and efficacy compared to other DDR inhibitors in all combinations tested. In terms of the combination partners, tubulin inhibitors achieved high efficacy but low synergy with DDR inhibitors, possibly due to the high cytotoxicity of tubulin

inhibitors alone (Lu et al., 2012) that may result in a plateau effect in cell growth inhibition which could not be further increased by combination with DDR inhibitors. Combination treatments with PARP inhibitors, such as veliparib, talazoparib, rucaparib, olaparib, and niraparib, which, with the exception of veliparib, are approved as targeted drugs for BRCA-mutated cancer treatment (Minchom et al., 2018; O'Connor, 2015), demonstrated the highest synergy with ATM and ATR inhibitors across multiple cancer types. The TOP1/2 (DNA topoisomerase 1/2) inhibitors SN-38 (the active metabolite of irinotecan), topotecan, etoposide, and doxorubicin, also display high efficacy and synergy with ATM/ATR/DNA-PK inhibitors (DNA-PK>ATR>ATM), as previously reported in preclinical studies (Fok et al., 2019; Jo et al., 2021; "Therapeutic Targeting of ATR Yields Durable Regressions in Small Cell Lung Cancers with High Replication Stress," 2021). Last, selected chemotherapeutics such as gemcitabine, an antimetabolite that inhibits DNA synthesis, also achieved high efficacy and synergy when combined with ATR and ATM inhibitors (**Figure 4.2a and b**). While the synergistic relationship between ATRi and gemcitabine has been reported before (Konstantinopoulos et al., 2021), we note that similar relationships between gemcitabine and either DNA-PKi or ATMi have not been reported before, to our knowledge. Overall, the dataset shows a low Pearson's correlation of 0.2 ($p<1e-22$) between efficacy and synergy, which, while well within the range of values observed in previous studies (Ianevski et al., 2020; Sen et al., 2019), highlights the need of analyzing both measures of response independently.

In addition to analyzing results on the level of individual drugs, we further characterized the most efficacious and synergistic combination treatments identified in our screen by their mode-of-actions. Hierarchical clustering based on responses in different cell lines shows treatments with the same mode of actions tend to cluster together (**Supplementary Figure**

4.3-7). For example, for monotherapy, ATM inhibitors (M3541 and M4076), CHK1 inhibitors (GDC0425 and LY2603818), and BET inhibitors (IBET151, CPI0610, and GSK525762A) are located adjacent to each other (**Supplementary Figure 4.3**). The same pattern, i.e., combinations with the same or similar mode-of-actions are more likely to cluster together, also appears in combination response in terms of efficacy (**Supplementary Figures 4.4 and 4.5**) and synergy (**Supplementary Figures 4.6 and 7**). When combined with ATM, ATR, and DNA-PK, several modes of action consistently showed high efficacy and synergy (**Figure 4.2c,** also see **Supplementary Table 4.2)**, in particular, TOP1i (Subhash et al., 2016), RRM2Bi (the small subunit of p53-inducible ribonucleotide reductase) (Sagawa et al., 2017; Xu et al., 2008), PLK1i (polo-like kinase 1) (Ragland et al., 2013), and checkpoint inhibitors CHEK1i and CHEK2i, suggesting that targeting cell cycle checkpoint may confer a significant benefit in the combination setting as has recently been suggested for ATRi-CHEK1i (Smith et al., 2010).

Drug mode-of-actions identified from synergy analyses alone partly overlapped with those for efficacy scores; inhibiting RRM1/2 and TOP pathways seems to be broadly effective in combination with ATR/ATM/DNA-PK inhibition. The inhibition of RRM1/2 pathway is only synergistic in combination with ATR, but not ATM and DNA-PK inhibition, while inhibiting TOP pathway is synergistic with all ATR, ATM, and DNA-PK inhibition. Lastly, PARP inhibitors appeared to be strongly and broadly synergistic in combination with ATRi/ATMi, but not DNA-PKi (**Figure 4.2d** and **Supplementary Table 4.3**).

Four monotherapy and two DDR inhibitor combinations show significant variability in response between different cancer types

To investigate whether general biological backgrounds, such as cancer or tissue types, influence treatment response, we carried out statistical comparisons of the efficacy and synergy

responses between different cancer types covering the 87 monotherapy agents and 465 combination treatments screened in our study.

As the number of cell lines covering each of the 12 cancer types varies, we chose the non-parametric Kruskal-Wallis test to analyze the variance of treatment response of each treatment across all cancer types in this study. After multiple testing corrections, only four out of the 87 monotherapy agents showed significant variance in efficacy across different cancer types ($p < 0.01$), including doxorubicin (p = 2.8e-08), M3541 (p = 2.2e-06); peposertib (p = 1.3e-05), and oxaliplatin (p = 3.4e-05)) (**Supplementary Figure 4.1**). Analogously, only two combination treatments out of the 465 combinations we tested showed significant variation in response across different cancer types: peposertib-gamma-ionizing-radiation (a DNA-PKi-IR combination showing significant cross-cancer type variance in terms of both efficacy (p = 3.38e-3) and synergy (p = 7.82e-5)), and M4076-berzosertib (an ATMi-ATRi combination showing variance only in terms of synergy (p = 2.39e-05)) (**Figure 4.3a and b**). As in the results on the raw efficacy and synergy values (see previous sections), also no correlation of cross-cancer variance significance values between efficacy and synergy scores was detected (Pearson's r = -0.028, p = 0.54) (**Figure 4.3c**), indicating again that the two scores are measurements of different pharmaceutical properties.

For all monotherapy and combination therapies that showed significant differences in responses across cancer types, we carried out statistical *post-hoc* analysis including Dunn's test, to identify individual cancer types with variable responses to individual drugs and drug combinations (**Figure 4.3d-f and Supplementary Figure 4.2**). Of the four significantly variable mono-therapeutic agents, doxorubicin showed significantly higher efficacy in hematological cancers than other cancer types, while M3541 demonstrated lower efficacy in both pancreas and

melanoma cancers than other cancer types (**Supplementary Figure 4.2b**). For peposertib and oxaliplatin, the difference in efficacy was only significant between bladder and ovary/hematological cancers, as well as between sarcoma and hematological cancers (**Supplementary Figure 4.2b**). For the combination treatments, the peposertib-gamma-ionizing-radiation combination displayed significantly higher efficacy in hematological cancers compared to bladder cancers (**Figures 4.3d and e**). Last, the case of M4076-berzosertib, shows a significantly lower synergy in hematological cancers compared to the pancreas, prostate, melanoma, and sarcoma cancers were observed (**Figure 4.3f**). Interestingly, no significant correlation between average treatment efficacy or synergy and the significance of variance in different cancer types (across monotherapies (Pearson's $r = -0.01$, $p = 0.92$ and Spearman's $r = -0.075$, $p = 0.478$) and combination therapies, as well as for both efficacy (Pearson's $r = 0.01$, $p = 0.835$ and Spearman's $r = 0.01$, $p = 0.8$) and synergy (Pearson's $r = -0.04$, $p = 0.37$ and Spearman's $r = -0.021$, $p = 0.64$)) could be identified, indicating that the cancer type specificity and overall average treatment response are independent pharmaceutical characteristics.

Discussion

We present a comprehensive combination treatment screening dataset focusing on DDR inhibitors, which allows us to identify interactions between DDR inhibitors and a broad range of anti-cancer drugs and map the molecular dependencies of their relationships. DDR inhibitors are an increasingly important class of targeted therapies explored for the treatment of cancer, and the results will help inform and recommend effective treatments depending on available genomic information. In our data, both the sequencing as well as combination treatment response data

were generated from the same cell culture lines, avoiding potential issues resulting from differing molecular backgrounds between screened and sequenced cell lines that may bias the analysis.

We identified inhibitors to four biological pathways that achieve strong combination efficacy in the screened cell lines when combined with any of the investigated DDR kinase inhibitors: the DNA topoisomerase pathway (TOP1 and TOP2 inhibitors), the serine/threonine-protein kinase PLK1 pathway (PLK1 inhibitors), the p53-inducible ribonucleotide reductase pathway (gemcitabine and cytarabine) and cell cycle checkpoints (in particular, CHK1 inhibitors). In addition, we found that PARP inhibitors achieve strong synergistic effects in combination with the ATR and ATM inhibitors, a finding that is currently being investigated for ATRi in ongoing clinical trials (*Study of M1774 in Combination With DNA Damage Response Inhibitor or Immune Checkpoint Inhibitor (DDRiver Solid Tumors 320)*, n.d.; Yap et al., 2022).

Concerning drug combination synergy, we identified peposertib-gamma-ionizing-radiation (ionizing radiation) (DNA-PKi-IR) and M4076-berzosertib (ATMi-ATRi) as combination treatments that show cross-cancer type variability in efficacy and synergy. Peposertib-gamma-ionizing-radiation is a DDR inhibitor combination that has been actively under preclinical evaluation (Romesser et al., 2021; Van Triest et al., 2018; Zenke et al., 2020) and shows robust response in cervical cancer xenograft model (Gordhandas et al., 2022) and enhances the response of immunotherapy (Carr et al., n.d.). Meanwhile, ATM and ATR loss-of-function have been proposed as being in a synthetically lethal relationship (Weber & Ryan, 2015), and ATM has been identified as a predictive biomarker of single-agent ATRi in multiple tumor types (Dunlop et al., 2020; Kwok et al., 2016; Min et al., 2017). Both combinations show synergy *in vitro* (0.14 bliss score for the peposertib-gamma-ionizing-radiation combination and 0.11 bliss score for the

M4076-berzosertib combination), indicating the potential for further investigation of the proper indication of both combinations in clinical use.

Our investigation has yielded crucial evidence shedding light on the potential of DDR-targeted combination therapies, highlighting their significant clinical prospects. However, it is essential for future studies to meticulously evaluate the toxicity and adverse events linked to such combined treatment approaches, ensuring patient safety and precise dosage calibration. The concept of synthetic lethality, which forms the foundation of DDR-targeted combination therapy, inherently enhances efficacy while concurrently increasing the risk of toxicity and adverse events (Martorana et al., 2022; Mullard, 2022). For example, PARP inhibitors, both as monotherapies and as components of combination regimens, have been extensively researched due to their pioneering role in DDR-targeted therapy, with a clinical history spanning over a decade (Coleman et al., 2019; LaFargue et al., 2019; Madariaga et al., 2020; C. Wang & Li, 2021). The simultaneous administration of the PARP inhibitor olaparib with the ATR inhibitor ceralasertib, for instance, has been correlated with the onset of anemia, neutropenia, and thrombocytopenia (Mahdi et al., 2021; Shah et al., 2021). Furthermore, certain combinations elucidated in our current study have previously been reported to increase the incidence of toxicity and adverse events. The ATR inhibitor berzosertib, usually well-tolerated as a single-agent therapy, has shown an increased prevalence of adverse events and hematological toxicities, including anemia, nausea, and neutropenia, when combined with carboplatin (Yap et al., 2020), gemcitabine (Konstantinopoulos, Cheng, Wahner Hendrickson, Penson, Schumer, Doyle, et al., 2020; Middleton et al., 2021), or topotecan (Thomas et al., 2018) in early-phase clinical trials. Despite the progress we have made in our research, we acknowledge that our efforts are still limited to the preliminary phase of *in vitro* high-throughput screening. Therefore, a comprehensive

exploration of *in vivo* toxicity associated with all the synergistic combinations unveiled in this study awaits future clinical trials.

Methods

Cell culture and drug response detection

This study is carried out on cell lines only and complies with all relevant ethical regulations of Merck Healthcare KGaA and the University of Michigan. All dose-response experiments were conducted at Oncolead GmbH & Co. KG (Karlsfeld, Germany). Cell lines were purchased directly from the ATCC, NCI, CLS, and DSMZ cell line collections. The cell lines were grown in the media recommended by the suppliers in the presence of 100 U/ml penicillin G and 100 µg/ml streptomycin supplied with 10% FCS.

Cells were grown in a 5% CO_2 atmosphere. Cell growth and treatment were performed in 96-well microtiter plates CELLSTAR® (Greiner Bio-One, Germany). Cells harvested from exponential phase cultures by trypsinization or by splitting (in the case of suspension growing cells) were plated in 90 µl of media at optimal seeding densities. The optimal seeding density for each cell line was determined to ensure exponential growth for the duration of the experiment. All cells growing without anticancer agents were sub-confluent by the end of the treatment, as determined by visual inspection.

Cells were allowed to stay for another 48 hours prior to compound treatment. The treatment was performed for 120 hours and stopped by the addition of trichloroacetic acid followed by using a total protein staining protocol (Sulforhodamine B (SRB) staining) (Vichai & Kirtikara, 2006). The bound SRB was solubilized with 100 µl of 10 mM Tris base. Optical density was measured at 492, 520, and 560 nm. Compound dilutions were performed in DMSO and diluted 1:100 in the RPMI medium. Combined treatment has been performed simultaneously. 90 µl of

cells were treated by mixing with 10 µl of the compound-containing media (resulting in a final DMSO concentration of 0.1%). In the case of combination, both agents were mixed together in DMSO at equal volumes so that the final concentration of DMSO was 0.2%. In addition, all experiments contained a few plates with cells that were analyzed immediately after the 48 hours recovery period. These plates contained information about the cell number, T_z, at time zero, i.e., before treatment, and served to calculate the cytotoxicity.

The calculation nomenclature used was introduced by DTP of the NCI (Shoemaker, 2006). The first step in data processing was calculating an average background value for each plate, derived from plates and wells containing mediums without cells. The average background optical density was then subtracted from the appropriate control values (containing cells without the addition of a drug), from values representing the cells treated with an anticancer agent, and from values of wells containing cells at time zero. Thus, the following values were obtained for each experiment: control cell growth, C; cells in the presence of an anticancer agent T_i and cells prior to compound treatment at time zero, T_z (or T_0, in some publications).

The selection of the concentration range for all agents was based on previous experiments using a panel of 62 cell lines. A 4-fold dilution and 5 data points were sufficient to cover the complete activity range for most of the agents (**Supplementary Figure 4.8 and 4.9**).

Dose-response evaluation measures

The non-linear curve fitting calculations were performed using algorithms and visualization tools using four-parameter log-logistic regression (DeLean et al., 1978; Ritz et al., 2015).

To obtain an estimate of treatment efficacy that encompasses both potency and maximum effect, the relative area over the curve (AoC) was computed by estimating the area under the fitted dose-response curve by the trapezoidal rule within ranges of relative growth rates

compared to untreated controls between 0% and 100%, and within ranges of drug concentrations between 1 nM and 1 mM, and dividing the estimated area by the sum of areas below and above the curve. The relative AoC measure used in this work thus captures both the potency of a compound combination (usually measured by IC_{50} or GI_{50}) as well as the maximum effect on cellular growth (as measured by the minimum of the curve); the relative AoC is of particular usefulness for capturing the efficacy of DDR inhibitors, many of which often have a comparatively low maximum effect less than 50% growth inhibition at realistic concentrations, which makes IC_{50} and GI_{50} less practically relevant.

Combination effects for the different compound combinations are calculated using the Bliss independence model (Berenbaum, 1989; Greco et al., 1995) under the assumption of independent modes of action of the combination partners. Bliss excess was calculated as the average excess of the observed effect E_{OBS} (*i.e.*, the relative reduction of growth rate compared to untreated controls) over the calculated linear combination of the monotherapy treatments effects ($E_{1+2} = E_1 + E_2 – E_1 E_2$) for all concentrations used:

$$Bliss_{excess} = \frac{1}{n}\Sigma_{i=1}^{n} E_{OBS_i} - E_{1+2_i} \quad \ldots\ldots \text{Eq. (1)}$$

In this formulation, the $Bliss_{excess}$ is a continuous value between -1 and 1 where values higher than about 0.2 are usually considered synergistic, and values below about -0.2 are usually considered antagonistic.

Statistics & Reproducibility

The reproducibility of measured response (*i.e.* AoC and Bliss score) are measured by Pearson's correlation within the replicated experiments. No data were excluded from the analyses.

Quantification and statistical analysis for drug response variance test

For hierarchical clustering based on drug responses, we used heatmap.2 function of gplots module (3.1.3) from R (4.2.3) for hierarchical clustering using Euclidean as the distance function and ward.D2 as the cluster function.

We used Python (>=3.8) module *scipy* (1.11.3) to carry out the Kruskal-Wallis test to test if a drug has different responses between different cancer types. The Kruskal-Wallis test is especially suitable for this situation as a non-parametric test, so it won't be affected by the different sample sizes of the subsets. For the significantly tissue-specific drugs ($p<0.01$), we also used *scipy* to carry out posthoc tests, including Dunn's test, Mann-Whitney Pairwise test, Conover-Iman test, and bootstrapping 10,000 times to locate the significantly different tissue types. Bonferroni correction was performed to adjust the above multiple comparisons.

Data Availability

The DDR combination *in vitro* screening data collected in this study are shared at and can be freely downloaded from: https://osf.io/8hbsx/. Source data are provided with this paper.

Code Availability

The source code of all statistical analyses is available from GitHub: https://github.com/GuanLab/DDR_combination_analysis.

Figures

Figure 4.1. Overview of combination treatment synergy screening experiments. (**a**) Dose-response curves were used to calculate drug pairs' efficacy and synergy scores. Inhibitors to DDR kinases ATM, ATR, and DNA-PK (ATMi/ATRi/DNAPKi) were tested against 62 cell lines across 12 tissues. (**b**) DDR inhibitor combination treatment screens show strong interactions between drugs targeting different DDR factors. The efficacy (left panel, by area over the curve (AoC) score) and synergistic (right panel, by Bliss score) responses of all combination treatments across the 12 tissue types tested in this study are shown. Six DDR inhibitors of interest of three mode-of-actions (ATMi, ATRi, and DNA-PKi, shown on the *y*-axis) combined with 87 drugs (*x*-axis), form 546 different combinations, which are facetted by the 12 different cancer cell line tissues of origin. Some drugs (and their mode-of-actions) with significant synergistic effects, when combined with the six DDR inhibitors of interest, are marked and shown in pop-out tables. More detailed information on all drug/mode-of-action combinations is shown in **Supplementary Figures 4.3 and 4.4.**

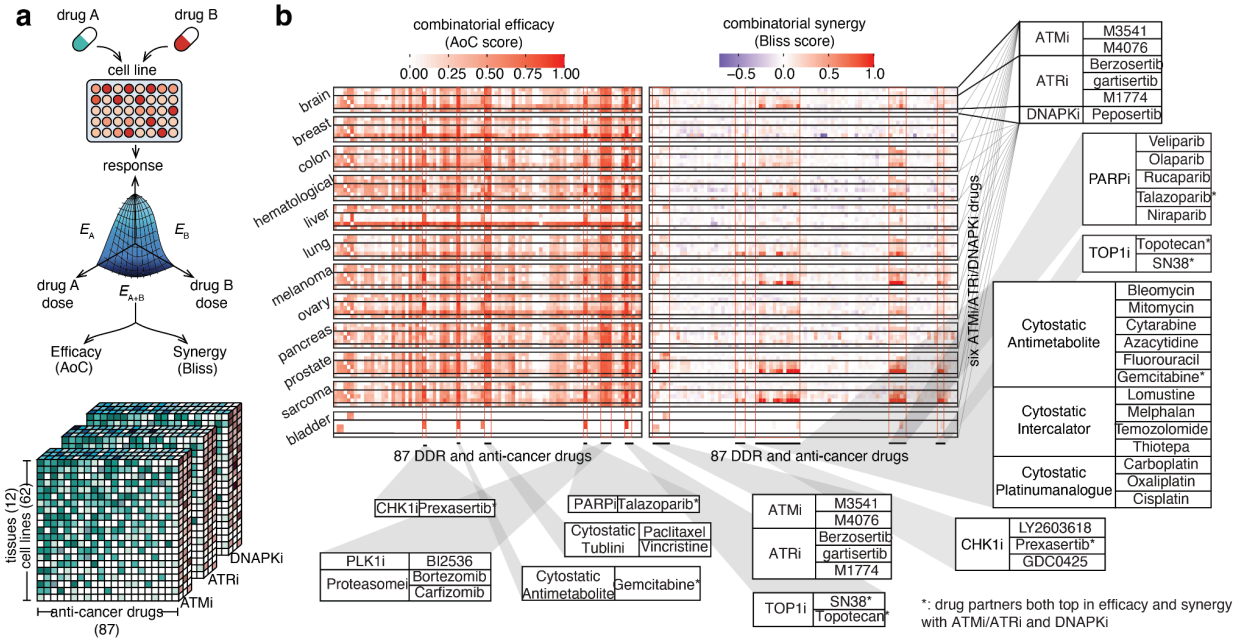


Figure 4.2. Top DDR inhibitor combination treatments that achieve the highest efficacy and synergy across all cell lines in the high-throughput treatment screening in this study. (**a** and **b**) Boxplots showing the treatment responses of drug combinations with the top 50 averaged (**a**) efficacy and (**b**) synergy responses in all 62 cancer cell lines (n = 62). Drug combinations are shown on the left side. Mode-of-actions of the DDR inhibitors are denoted by red (ATR inhibitor), blue (ATM inhibitor), green (DNA-PK inhibitor), and yellow (ATR inhibitor-ATM inhibitor combination) in the box plot, while mode-of-actions of the partner drugs are shown at the right side. The interquartile range (25th to 75th percentile) and median lines are shown, with whiskers extending to 1.5 times the interquartile range. (**c** and **d**) show the top 10 target genes with the highest average (**C**) efficacy and (**D**) synergy in combination with ATR, ATM, and DNA-PK (PRKDC) inhibitors. Each target gene of a partner drug is denoted by a node in the diagram, and the combination response (efficacy or synergy) is denoted by the relative strength of the connection.

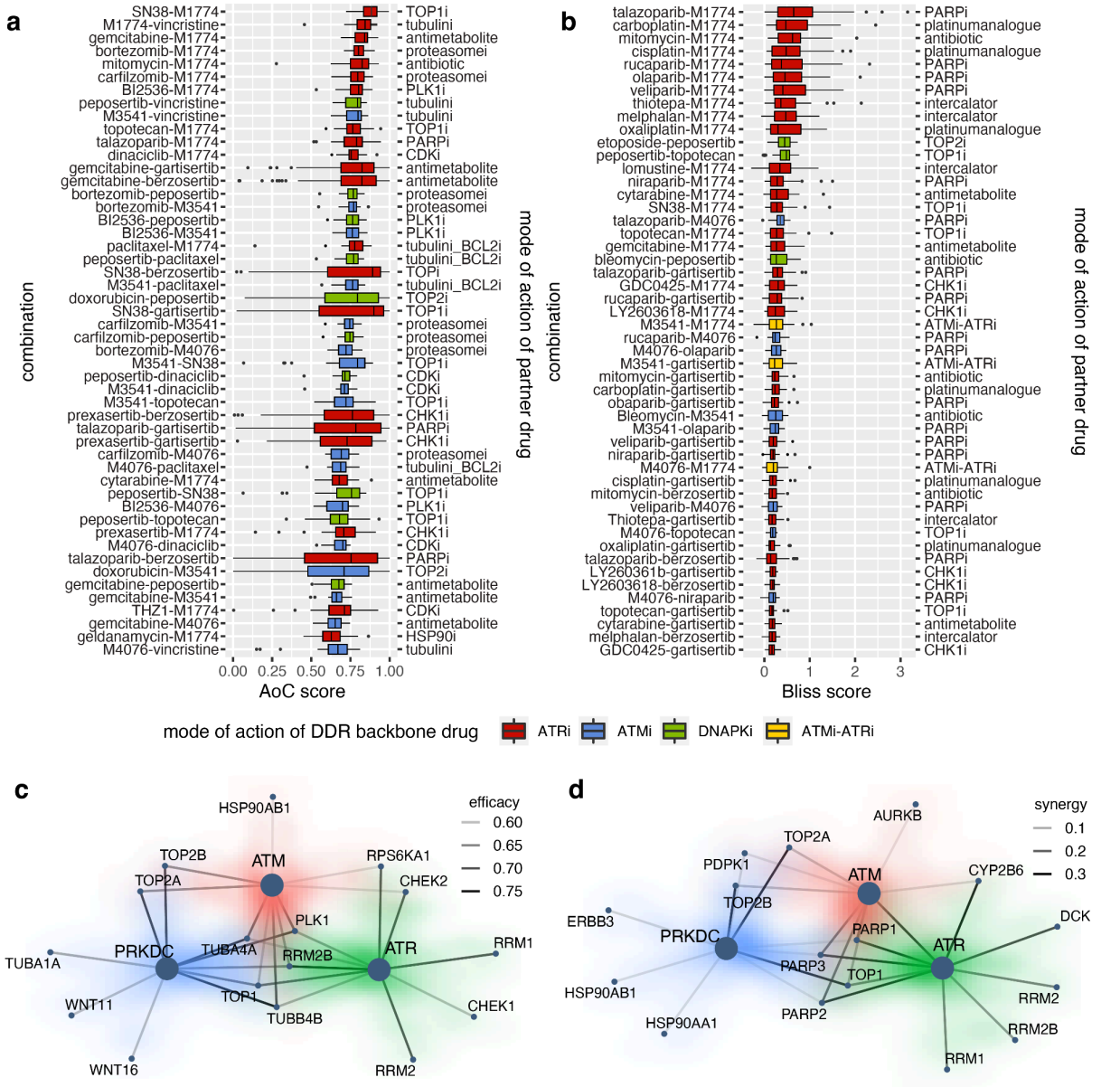
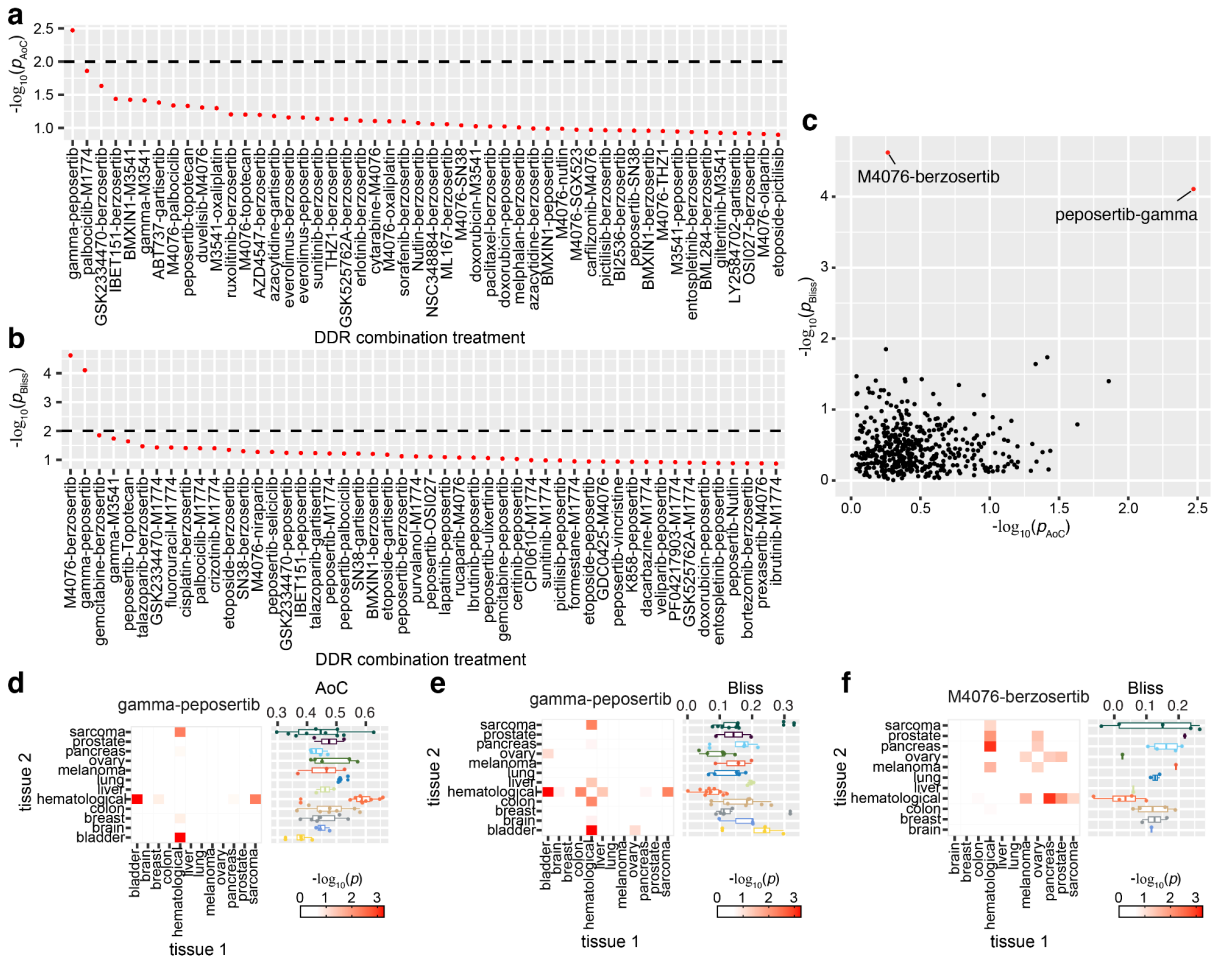


Figure 4.3. Results from cross-cancer type variance test of DDR inhibitor combination treatment response. (**a** and **b**) Kruskal-Wallis test shows the significance of cross-cancer type variance of DDR inhibitor combinations tested in this study. -log10(p) from the cross-tissue variance test for (**a**) efficacy (AoC score) and (**b**) synergy (Bliss) of the top 50 combinations are shown, and the significance threshold (p = 0.01) is marked by a dashed line. (**c**) shows the correlation between cross-cancer-type variance significance in AoC score and Bliss score for all combination treatments tested in this study. Each dot in (**c**) denotes a combination treatment. (**d-f**) Heatmap shows the results from post-hoc analysis by Dunn's test on the significantly variant combination treatments (peposertib-gamma-ionizing-radiation and M4076-berzosertib) from the Kruskal-Wallis test, and the right lane shows the distribution of responses (AoC or Bliss scores) in different cancer types (boxplots show the 25, 50 and 75 percentiles with whiskers extending to 1.5 times the interquartile range; for each cancer types the total numbers of cell lines are: bladder=4; brain=3; breast=6; colon=8; hematological=10; liver=2; lung=5; melanoma=3; ovary=5; pancreas=4; prostate=2; sarcoma=10). As M4076-berzosertib only shows the cross-cancer-type variance in the Bliss score, only the post hoc test result on the Bliss score is shown for this combination. All statistically significant values from the variance test are two-sided.



Supplementary Tables

Supplementary Table 4.1. Target gene and mode-of-action of all anti-cancer drugs tested in this study.

drug_name	mode-of-action	drug target	reference
Ceritinib	ALKi	ABCB1,ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2B,AKT1,AKT2,AKT3,ALK,ARAF,ATR,AURKA,AURKB,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,CAMK2G,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK9,CHD4,CHEK1,CIT,CLK2,CSNK1D,CSNK1E,CSNK1G3,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,EGFR,EIF2AK1,EML4,EPHA2,EPHA4,EPHA5,EPHA7,EPHB6,ERCC2,ETV6,FES,FGFR1,FGFR2,FGFR3,FGFR4,FLT3,GSK3B,IGF1R,IKBKE,INSR,IRAK1,IRAK4,JAK1,JAK2,JAK3,KDR,KIT,LATS1,LCK,LYN,MAP2K1,MAP2K2,MAP3K1,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MCM4,MET,NAT10,NEK2,NEK7,NLK,NPM1,NTRK1,NTRK2,PAK4,PDGFRA,PDGFRB,PEBP1,PIM1,PIM2,PKMYT1,PLK1,PLK4,PRKAA1,PRKAG2,PRKAR2A,PRKCD,PRKCI,PRKCQ,PRKCZ,PRKD2,PTK2B,PTK6,RAN,RET,ROCK1,ROCK2,ROS1,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SMC1A,SMC2,SRC,STAT3,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TNIK,TOP2A,TOP2B,TTK,TYK2,WEE1,YES1	CHEMBL; DGIdb3.0; Drugbank; LINCS
Formestane	Aromatasei	ESR1,TDP1,YES1	CHEMBL; DGIdb3.0; Drugbank; LINCS
M3541	ATMi	ATM	CHEMBL; DGIdb3.0; Drugbank; LINCS
M4076	ATMi	ATM	CHEMBL; DGIdb3.0; Drugbank; LINCS
AZD6738	ATRi	ATM,ATR,CYP2D6,MTOR,PIK3C2G,PIK3CA,PRKDC	CHEMBL; DGIdb3.0; Drugbank; LINCS
BAY1895344	ATRi	ATM,ATR,CYP2D6,MTOR,PIK3CB,PRKDC	CHEMBL; DGIdb3.0; Drugbank; LINCS
M1774	ATRi	ATR	CHEMBL; DGIdb3.0; Drugbank; LINCS
Gartisertib	ATRi	ATR	CHEMBL; DGIdb3.0; Drugbank; LINCS
Berzosertib	ATRi	ABL1,ATM,ATR,DYRK2,FLT3,FLT4,GSK3B,JAK2,KIT,MTOR,PIK3CA,PIK3CG,PIK3R1,PRKDC,SYK	CHEMBL; DGIdb3.0; Drugbank; LINCS
Tozasertib	Aurorai	ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2A,ACVR2B,ACVRL1,AKT1,AKT2,AKT3,ALK,ARAF,AURKA,AURKB,AXL,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,BRSK2,CAMK2A,CAMK2G,CAMKK1,CASK,CCNA1,CCNA2,CCNB1,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,CDKL2,CHD4,CHEK1,CHEK2,CIT,CLK2,CLK3,CSF1R,CSNK1A1L,CSNK1D,CSNK1E,CSNK1G3,CSNK2A1,DAPK1,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,DLK1,DMPK,DSTYK,DYRK2,EGFR,EIF2AK1,EIF2AK4,EPHA2,EPHA3,EPHA4,EPHA5,EPHA6,EPHA7,EPHB1,EPHB6,ERBB2,ERBB3,ERBB4,ERCC2,FES,FGFR1,FGFR2,FGFR3,FGFR4,FLG,FLT1,FLT3,FLT4,GSK3B,HIPK2,IGF1R,IKBKE,INSR,INSRR,IRAK1,IRAK4,ITK,JAK1,JAK2,JAK3,KDM4A,KDR,KIT,LATS1,LATS2,LCK,LRRK2,LY	CHEMBL; DGIdb3.0; Drugbank; LINCS

		N,MAP2K1,MAP2K2,MAP2K4,MAP2K7,MAP3K1,MAP3K13,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP3K7,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MAST1,MATK,MCM4,MET,MST1,MTOR,MYO3B,NAT10,NEK2,NEK4,NEK6,NEK7,NFE2L2,NLK,NTRK1,NTRK2,NTRK3,PAK1,PAK3,PAK4,PBK,PDGFRA,PDGFRB,PDK1,PDPK1,PEBP1,PIK3C2B,PIK3C2G,PIK3CA,PIK3CB,PIK3CD,PIK3CG,PIM1,PIM2,PIM3,PIP4K2B,PIP5K1A,PKMYT1,PLK1,PLK4,PRKAA1,PRKAA2,PRKAG2,PRKAR2A,PRKCD,PRKCG,PRKCI,PRKCQ,PRKCZ,PRKD1,PRKD2,PTK2B,PTK6,RAF1,RAN,RAPGEF3,RET,RIOK3,RIPK1,ROCK1,ROCK2,ROS1,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SGK1,SGK2,SGK3,SMAD3,SMC1A,SMC2,SRC,SRPK2,STK11,STK16,STK3,STK4,STRADA,SUFU,SYK,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TLK1,TLK2,TNIK,TOP2A,TOP2B,TTK,TYK2,TYRO3,WEE1,YES1	
ABT737	BCL2i	BAD,BAK1,BAX,BBC3,BCL2,BCL2L1,BCL2L11,BCL2L2,BID,MCL1,MDM2	CHEMBL; DGIdb3.0; Drugbank; LINCS
Imatinib	BCRi_ABLi	ABCB1,ABCB11,ABCG2,ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2A,ACVR2B,ACVRL1,ADORA2A,AGTR2,AKT1,AKT2,AKT3,ALK,APEX1,ARAF,AURKA,AURKB,AXL,BCR,BDKRB2,BMP2K,BMPR1A,BMPR1B,BRAF,BRSK2,CAMK2A,CAMK2G,CAMKK1,CASK,CCNA1,CCNA2,CCNB1,CCNE1,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,CDKL2,CHD4,CHEK1,CHEK2,CIT,CLK2,CLK3,CSF1R,CSNK1A1L,CSNK1D,CSNK1E,CSNK1G3,CSNK2A1,CYP2B6,CYP2D6,DAPK1,DCK,DDR1,DDR2,DDX3X,DDX6,DMPK,DSTYK,DYRK2,EGFR,EIF2AK1,EIF2AK4,ELANE,EPHA2,EPHA3,EPHA4,EPHA5,EPHA6,EPHA7,EPHB1,EPHB6,ERBB2,ERBB3,ERBB4,ERCC2,ESR1,ESR2,FEN1,FES,FGFR1,FGFR2,FGFR3,FGFR4,FLT1,FLT3,FLT4,GSK3B,HDAC1,HDAC2,HDAC7,HDAC8,HIPK2,HMGCR,IDH1,IGF1R,IKBKE,INSR,INSRR,IRAK1,IRAK4,ITK,JAK1,JAK2,JAK3,KDR,KIT,LATS1,LATS2,LCK,LMNA,LRRK2,LYN,MAP2K1,MAP2K2,MAP2K4,MAP2K7,MAP3K1,MAP3K13,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP3K7,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MAST1,MATK,MCM4,MET,MTOR,MYO3B,MYT1,NAT10,NEK2,NEK4,NEK6,NEK7,NLK,NPM1,NTRK1,NTRK2,NTRK3,OPRK1,PAK1,PAK3,PAK4,PBK,PDGFRA,PDGFRB,PDPK1,PEBP1,PIK3C2B,PIK3C2G,PIK3CA,PIK3CB,PIK3CD,PIK3CG,PIM1,PIM2,PIM3,PIP4K2B,PIP5K1A,PKMYT1,PLK1,PLK4,POLI,POLK,PRKAA1,PRKAA2,PRKAB1,PRKAG2,PRKAR2A,PRKCD,PRKCG,PRKCI,PRKCQ,PRKCZ,PRKD1,PRKD2,PTK2B,PTK6,PTPRC,RAD52,RAF1,RAN,RET,RIOK3,RIPK1,ROCK1,ROCK2,ROS1,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SFN,SGK1,SGK2,SGK3,SMAD3,SMC2,SRC,SRPK2,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TDP1,TGFBR1,TGFBR2,TLK1,TLK2,TNIK,TOP2A,TOP2B,TTK,TYK2,TYRO3,VRK1,WEE1,YES1	CHEMBL; DGIdb3.0; Drugbank; LINCS
CPI0610	BETi	BRD4	CHEMBL; DGIdb3.0; Drugbank; LINCS
GSK525762A	BETi	BRD2,BRD3,BRD4,CREBBP,CYP2D6,SMARCA4	CHEMBL; DGIdb3.0; Drugbank; LINCS
IBET151	BETi	BRD2,BRD3,BRD4,CREBBP,CYP2D6,PDE4B,SMARCA4	CHEMBL; DGIdb3.0; Drugbank; LINCS

BMXIN1	BMXi	JAK3	CHEMBL; DGIdb3.0; Drugbank; LINCS
Vemurafenib	BRAFi	ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2B,AKT1,AKT2,AKT3,ARAF,ATR,AURKA,AURKB,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,CAMK2G,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK9,CHD4,CHEK1,CIT,CLK2,CSNK1D,CSNK1E,CSNK1G3,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,EGFR,EIF2AK1,EPHA2,EPHA4,EPHA5,EPHB6,ERCC2,FES,FGFR1,FLT3,GSK3B,IGF1R,IKBKE,INSR,IRAK1,IRAK4,JAK1,JAK2,KDR,KRAS,LATS1,LCK,LYN,MAP2K1,MAP2K2,MAP2K4,MAP3K1,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MCM4,MET,NAT10,NEK2,NEK7,NLK,NTRK1,PAK4,PDGFRB,PEBP1,PIM1,PIM2,PKMYT1,PLK1,PLK4,PRKAA1,PRKAG2,PRKAR2A,PRKCD,PRKCI,PRKCQ,PRKD2,PTK2B,PTK6,RAF1,RAN,RET,ROCK1,ROCK2,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SMC1A,SMC2,SRC,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TNIK,TOP2A,TOP2B,TYK2,WEE1,YES1	CHEMBL; DGIdb3.0; Drugbank; LINCS
Ibrutinib	BTKi	ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2B,AKT1,AKT2,AKT3,ARAF,AURKA,AURKB,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,CAMK2G,CCNE1,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK9,CHD4,CHEK1,CIT,CLK2,CSF1R,CSNK1D,CSNK1E,CSNK1G3,CYP2B6,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,EGFR,EIF2AK1,EIF4EBP1,EPHA2,EPHA4,EPHA5,EPHA7,EPHB6,ERBB2,ERBB3,ERBB4,ERCC2,FES,FGFR1,FGFR2,FLT1,FLT3,GSK3B,IGF1R,IKBKE,INSR,INSRR,IRAK1,IRAK4,ITK,JAK1,JAK2,JAK3,LATS1,LCK,LYN,MAP2K1,MAP2K2,MAP2K7,MAP3K1,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MCM4,MET,NAT10,NEK2,NEK7,NLK,NTRK1,PAK4,PDGFRB,PEBP1,PIM1,PKMYT1,PLK1,PLK4,PRKAA1,PRKAG2,PRKAR2A,PRKCD,PRKCI,PRKCQ,PRKD2,PTK2B,PTK6,RAN,RET,ROCK1,ROCK2,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SMAD3,SMC1A,SMC2,SRC,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TDP1,TGFBR1,TGFBR2,TNIK,TOP2A,TOP2B,TTK,TYK2,WEE1,YES1	CHEMBL; DGIdb3.0; Drugbank; LINCS
LY2857785	CDK9i	CCNC,CCNH,CDK7,CDK8,CDK9,MNAT1	CHEMBL; DGIdb3.0; Drugbank; LINCS
Palbociclib	CDKi	ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2B,AKT1,AKT2,AKT3,ALK,ARAF,AURKA,AURKB,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,CAMK2A,CAMK2G,CCNA1,CCNA2,CCNB1,CCND1,CCND2,CCND3,CCNE1,CCNH,CDC7,CDK1,CDK12,CDK2,CDK20,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,CHEK1,CIT,CLK2,CLK3,CSF1R,CSNK1D,CSNK1E,CSNK1G3,CSNK2A1,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,EGFR,EIF2AK1,EPHA2,EPHA4,EPHA5,EPHA7,EPHB6,ERCC2,FES,FGFR1,FGFR2,FGFR3,FGFR4,FLG,FLT1,FLT3,GSK3B,IGF1R,IKBKE,INSR,IRAK1,IRAK4,JAK1,JAK2,JAK3,KDR,LATS1,LCK,LRRK2,LYN,MAP2K1,MAP2K2,MAP2K4,MAP3K1,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MCM4,MET,MNAT1,NAT10,NEK2,NLK,NTRK1,NTRK2,NTRK3,PAK4,PDGFRA,PDGFRB,PIK3CD,PIM1,PIM2,PKMYT1,PLK4,PRKAA1,PRKAG2,PRKAR2A,PRKCD,PRKCI,PRKCQ,PRKD2,PTK2B,PTK6,RAN,RET,ROCK1,ROCK2,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SMC	CHEMBL; DGIdb3.0; Drugbank; LINCS

		2,SRC,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TNIK,TOP2A,TOP2B,TYK2,TYRO3,WEE1,YES1	
Purvalanol	CDKi	ABL1,ABL2,ACVR1B,AKT1,AKT2,AKT3,ALK,APEX1,AURKA,AURKB,AXL,BRSK2,CAMK2G,CAMKK1,CCNA1,CCNA2,CCNB1,CCNB2,CCNB3,CCND1,CCND3,CCNE1,CCNE2,CCNH,CDK1,CDK12,CDK2,CDK20,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,CHEK1,CHEK2,CLK2,CLK3,CSF1R,CSNK1D,CSNK1G3,CSNK2A1,CYP2D6,DAPK1,DDR2,DMPK,DYRK2,EGFR,EPHA2,EPHA3,EPHA4,EPHA5,EPHA7,EPHB1,ERBB4,FES,FGFR1,FGFR2,FGFR3,FGFR4,FLT1,FLT3,FLT4,GBA,GSK3B,HIF1A,HIPK2,IGF1R,INSR,INSRR,IRAK1,IRAK4,ITK,JAK2,JAK3,KDR,KIT,KMT2A,LCK,LYN,MAP2K1,MAP2K7,MAP3K5,MAP3K7,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MEN1,MET,MTOR,NEK2,NEK6,NEK7,NLK,NTRK1,NTRK2,PAK3,PAK4,PDGFRA,PDGFRB,PDPK1,PIM1,PIM2,PIM3,PLK1,PRKAA1,PRKAA2,PRKAB1,PRKAG2,PRKCD,PRKCG,PRKCI,PRKCQ,PRKCZ,PRKD1,PRKD2,PTK2B,PTK6,RAF1,RET,ROCK1,ROCK2,ROS1,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SGK1,SGK2,SGK3,SRC,SRPK2,STAT6,STK11,STK3,STK4,SYK,TAOK1,TAOK2,TAOK3,TDP1,TGFBR1,TLK2,TP53,TSHR,TYRO3,USP1,YES1	CHEMBL; Drugbank; LINCS DGIdb3.0;
RO3306	CDKi	AKT1,AKT2,AKT3,AURKA,AURKB,CCNA2,CCNB1,CCND1,CCNE1,CCNH,CDK1,CDK12,CDK2,CDK20,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,GSK3B,MAPK1,MNAT1,PRKCD,RPS6KA3	CHEMBL; Drugbank; LINCS DGIdb3.0;
Seliciclib	CDKi	ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2A,ACVR2B,ACVRL1,AKT1,AKT2,AKT3,ALK,ARAF,ATAD5,AURKA,AURKB,AXL,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,BRSK2,CAMK2A,CAMK2G,CAMKK1,CCNA1,CCNA2,CCNB1,CCNB2,CCNB3,CCND1,CCND3,CCNE1,CCNE2,CCNH,CDK1,CDK12,CDK2,CDK20,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,CHD4,CHEK1,CHEK2,CIT,CLK2,CLK3,CSF1R,CSNK1A1L,CSNK1D,CSNK1E,CSNK1G3,CSNK2A1,CYP2D6,DAPK1,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,DMPK,DYRK2,EGFR,EIF2AK1,EIF2AK4,EPHA2,EPHA3,EPHA4,EPHA5,EPHA6,EPHA7,EPHB1,EPHB6,ERBB2,ERBB4,ERCC2,FES,FGFR1,FGFR2,FGFR3,FGFR4,FLT1,FLT3,FLT4,GSK3B,HDAC1,HIF1A,HIPK2,IGF1R,IKBKE,INSR,INSRR,IRAK1,IRAK4,ITK,JAK1,JAK2,JAK3,KAT2A,KDR,KIT,LATS1,LATS2,LCK,LMNA,LYN,MAP2K1,MAP2K2,MAP2K4,MAP3K1,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MATK,MCM4,MET,MNAT1,MYO3B,NAT10,NEK2,NEK4,NEK6,NEK7,NFE2L2,NFKB1,NLK,NTRK1,NTRK2,NTRK3,PAK1,PAK3,PAK4,PBK,PDGFRA,PDGFRB,PDPK1,PEBP1,PIK3CA,PIK3R1,PIM1,PIM2,PIM3,PIP4K2B,PIP5K1A,PKMYT1,PLK1,PLK4,PRKAA1,PRKAA2,PRKAG2,PRKAR2A,PRKCD,PRKCG,PRKCI,PRKCQ,PRKCZ,PRKD1,PRKD2,PTK2B,PTK6,RAF1,RAN,RB1,RET,RIOK3,RIPK1,ROCK1,ROCK2,ROS1,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SGK1,SGK2,SMC1A,SMC2,SRC,SRPK2,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TDP1,TGFBR1,TGFBR2,TLK1,TLK2,TNIK,TOP2A,TOP2B,TP53,TSHR,TTK,TYK2,TYRO3,VRK1,WEE1,YES1	CHEMBL; Drugbank; LINCS DGIdb3.0;
Dinaciclib	CDKi	ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2B,AKT1,AKT2,AKT3,ARAF,ATR,AURKA,AURKB,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,BRD4,CAMK2G,CCNA2,CCNB1,CCND1,CCNE1,CCNH,CDK1,CDK12,CDK2,CDK20,CDK4,CDK5,CD	CHEMBL; Drugbank; LINCS DGIdb3.0;

		K6,CDK7,CDK8,CDK9,CHD4,CHEK1,CIT,CLK2,CSNK1D,CSNK1E,CSNK1G3,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,EGFR,EIF2AK1,EPHA2,EPHA4,EPHA5,EPHA7,EPHB6,ERCC2,FES,FGFR1,FLT3,GSK3B,IGF1R,IKBKE,INSR,IRAK1,IRAK4,JAK1,LATS1,LCK,LYN,MAP2K1,MAP2K2,MAP2K4,MAP3K1,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MCM4,MET,MNAT1,NAT10,NEK2,NEK7,NLK,NTRK1,PAK4,PDGFRB,PEBP1,PIM1,PIM2,PKMYT1,PLK4,PRKAA1,PRKAG2,PRKAR2A,PRKCD,PRKCI,PRKCQ,PRKCZ,PRKD2,PTK2B,PTK6,RAN,RET,ROCK1,ROCK2,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SMC1A,SMC2,SRC,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TNIK,TOP2A,TOP2B,TYK2,WEE1,YES1	
THZ1	CDKi	CCNH,CDK1,CDK12,CDK2,CDK20,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,MNAT1	CHEMBL; DGIdb3.0; Drugbank; LINCS
ML167	CDKi_CLK4i	CDK1,CDK12,CDK2,CDK20,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,CLK2,CLK3	CHEMBL; DGIdb3.0; Drugbank; LINCS
GSK923295	CENPEi	CENPE	CHEMBL; DGIdb3.0; Drugbank; LINCS
GDC0425	CHK1i	CHEK1	CHEMBL; DGIdb3.0; Drugbank; LINCS
LY2603618	CHK1i	ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2B,AKT1,AKT2,AKT3,ARAF,ATR,AURKA,AURKB,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,CAMK2G,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK9,CHD4,CHEK1,CIT,CLK2,CLK3,CSNK1D,CSNK1E,CSNK1G3,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,EGFR,EIF2AK1,EPHA2,EPHA4,EPHA5,EPHA7,EPHB6,ERCC2,FES,FGFR1,FLT3,GSK3B,IGF1R,IKBKE,INSR,IRAK1,IRAK4,JAK1,JAK2,LATS1,LCK,LYN,MAP2K1,MAP2K2,MAP2K4,MAP3K1,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MCM4,MET,NAT10,NEK2,NEK7,NLK,NTRK1,PAK4,PDGFRB,PEBP1,PIM1,PIM2,PKMYT1,PLK1,PLK4,PRKAA1,PRKAG2,PRKAR2A,PRKCD,PRKCI,PRKCQ,PRKCZ,PRKD2,PTK2B,PTK6,RAN,RET,ROCK1,ROCK2,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SMC1A,SMC2,SRC,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TNIK,TOP2A,TOP2B,TYK2,WEE1,YES1	CHEMBL; DGIdb3.0; Drugbank; LINCS
Prexasertib	CHK1i	CDK1,CDK2,CHEK1,CHEK2,FLT3,LCK,LYN,PIM1,PIM3,PRKAA1,PRKAB1,RPS6KA1,RPS6KB1	CHEMBL; DGIdb3.0; Drugbank; LINCS
Pomalidomide	CRBNi	ABCB11,BRD4,CDK6,CUL4A,DDB1,IKZF1,IKZF3,NFKB1,NFKB2,PDE4B,RBX1,RELA	CHEMBL; DGIdb3.0; Drugbank; LINCS
Bleomycin	Cytostatic_Antibiotic	ADORA2A,AGTR2,BDKRB2,CYP2D6,EGFR,ELANE,ERBB2,ESR1,ESR2,FLT1,HMGCR,LCK,MAPK1,MAPK14,MAPK3,OPRK1,PTPRC	CHEMBL; DGIdb3.0; Drugbank; LINCS
Mitomycin	Cytostatic_Antibiotic	ABCB1,ADORA2A,AGTR2,APEX1,BDKRB2,CYP2D6,EGFR,ELANE,ERBB2,ESR1,ESR2,FLT1,GNAS,HMGCR,IDH1,IDO1,LCK,MAPK1,MAPK14,MAPK3,NFE2L2,OPRK1,POLI,POLK,PTPRC,SMAD3,TDP1	CHEMBL; DGIdb3.0; Drugbank; LINCS
Cytarabine	Cytostatic_Antimetabolite	ABCB1,ABCB11,ADORA2A,AGTR2,APEX1,ATAD5,BDKRB2,BLM,CBFB,CYP2D6,DCK,DNMT1,EGFR,ELANE,ERBB2,ESR1,ESR2,FLT1,HMGCR,KMT2A,LCK,LMNA,MAPK1,MAPK14,	CHEMBL; DGIdb3.0; Drugbank; LINCS

		MAPK3,MDM2,MEN1,MTOR,NCOA1,NCOA3,NFE2L2,NFKB1,OPRK1,PIN1,POLA1,POLB,POLD1,POLE,POLI,PTPRC,RRM1,RUNX1,SIRT1,SRC,TDP1,TK1,USP1	
Azacytidine	Cytostatic_Antimetabolite	ADORA2A,AGTR2,APEX1,AR,ATAD5,BDKRB2,BLM,CYP2D6,DNMT1,DNMT3A,EGFR,ELANE,ERBB2,ESR1,ESR2,FLT1,FLT3,HMGCR,IDH1,KAT2A,LCK,LMNA,MAPK1,MAPK14,MAPK3,MTOR,NFE2L2,NFKB1,OPRK1,PPARG,PTPRC,RORC,RXRA,SMAD3,TDP1,TP53,UHRF1,VDR	CHEMBL; Drugbank; LINCS DGIdb3.0;
Fluorouracil	Cytostatic_Antimetabolite	ABCB11,ADORA2A,AGTR2,APEX1,ATAD5,BDKRB2,CTNNB1,CYP2D6,DTYMK,EGFR,ELANE,ERBB2,ESR1,ESR2,FEN1,FLT1,HMGCR,IDH1,LCK,LMNA,MAPK1,MAPK14,MAPK3,MBNL1,MTOR,NFE2L2,OPRK1,PIK3CA,PIK3R1,PTPRC,TDP1,TP53,TSHR,TYMS	CHEMBL; Drugbank; LINCS DGIdb3.0;
Gemcitabine	Cytostatic_Antimetabolite	ABCB1,ABCB11,ADORA2A,AGTR2,BDKRB2,CYP2D6,DCK,EGFR,ELANE,ERBB2,ESR1,ESR2,FLT1,HMGCR,LCK,MAPK1,MAPK14,MAPK3,OPRK1,PTPRC,RRM1,RRM2,RRM2B,SIRT1,TDP1	CHEMBL; Drugbank; LINCS DGIdb3.0;
Dacarbazine	Cytostatic_Intercalator	ABCB11,ADORA2A,AGTR2,BDKRB2,CYP2D6,EGFR,ELANE,ERBB2,ESR1,ESR2,FLT1,HMGCR,IDH1,LCK,LMNA,MAPK1,MAPK14,MAPK3,OPRK1,PAX8,PTPRC,TSHR	CHEMBL; Drugbank; LINCS DGIdb3.0;
Lomustine	Cytostatic_Intercalator	ABCB11,ADORA2A,AGTR2,BDKRB2,CYP2D6,EGFR,ELANE,ERBB2,ESR1,ESR2,FLT1,HMGCR,LCK,MAPK1,MAPK14,MAPK3,OPRK1,PTPRC,TDP1	CHEMBL; Drugbank; LINCS DGIdb3.0;
Melphalan	Cytostatic_Intercalator	ABCB11,ADORA2A,AGTR2,AR,BDKRB2,CYP2D6,EGFR,ELANE,ERBB2,ESR1,ESR2,FLT1,GBA,HIF1A,HMGCR,LCK,MAPK1,MAPK14,MAPK3,NFE2L2,NFKB1,OPRK1,PPARG,PTPRC,RORC,RXRA,SMAD3,TDP1,TP53,USP1,VDR	CHEMBL; Drugbank; LINCS DGIdb3.0;
Temozolomide	Cytostatic_Intercalator	ABCB1,ABCB11,ABL1,ABL2,KMT2A,MBNL1,MEN1,PABPC1,POLK	CHEMBL; Drugbank; LINCS DGIdb3.0;
Thiotepa	Cytostatic_Intercalator	ABCB11,CYP2B6,LMNA,PLK1,POLI,TDP1,VDR	CHEMBL; Drugbank; LINCS DGIdb3.0;
Carboplatin	Cytostatic_Platinumanalogue	ADORA2A,AGTR2,APEX1,BDKRB2,CASP7,CBFB,CYP2D6,EGFR,ELANE,ERBB2,ESR1,ESR2,FLT1,HMGCR,KAT2A,KMT2A,LCK,MAPK1,MAPK14,MAPK3,MEN1,OPRK1,POLK,PTPRC,RUNX1,TDP1,USP1,VDR	CHEMBL; Drugbank; LINCS DGIdb3.0;
Cisplatin	Cytostatic_Platinumanalogue	ABCB1,ADORA2A,AGTR2,APEX1,AR,BDKRB2,CBFB,CDK2,CDK4,CYP2D6,EGFR,ELANE,ERBB2,ESR1,ESR2,FEN1,FLT1,HMGCR,KAT2A,LCK,MAPK1,MAPK14,MAPK3,NFE2L2,OPRK1,PIN1,POLI,POLK,PTPRC,RECQL,RUNX1,RXRA,TDP1,USP1,VDR	CHEMBL; Drugbank; LINCS DGIdb3.0;
Oxaliplatin	Cytostatic_Platinumanalogue	ADORA2A,AGTR2,BDKRB2,CYP2D6,EGFR,ELANE,ERBB2,ESR1,ESR2,FLT1,HMGCR,LCK,MAPK1,MAPK14,MAPK3,OPRK1,PTPRC,TDP1	CHEMBL; Drugbank; LINCS DGIdb3.0;
Vincristine	Cytostatic_Tubulini	ABCB1,ABCB11,ADORA2A,AGTR2,ATAD5,BDKRB2,CYP2D6,EGFR,ELANE,ERBB2,ESR1,ESR2,FLT1,HIF1A,HMGCR,IDH1,KAT2A,LCK,LMNA,MAPK1,MAPK14,MAPK3,NFE2L2,OPRK1,PTPRC,SMAD3,TDP1,TUBA1A,TUBA4A,TUBB4B	CHEMBL; Drugbank; LINCS DGIdb3.0;
Paclitaxel	Cytostatic_Tubulini_BCL2i	ABCB1,ABCB11,ABL1,ADORA2A,AGTR2,AR,ATAD5,BCL2,BCL2L1,BDKRB2,BLM,CYP2D6,EGFR,ELANE,ERBB2,ESR1,ESR2,FGFR1,FLT1,HIF1A,HMGCR,HSP90AA1,HSPE1,IDH1,KMT2A,LCK,LMNA,MAPK1,MAPK14,MAPK3,MEN1,MTOR,NFE2L2,OPRK1,PIN1,PPARG,PTPRC,RECQL,RXRA,SMAD3,S	CHEMBL; Drugbank; LINCS DGIdb3.0;

		NCA,SRC,TDP1,TK1,TP53,TUBA1A,TUBA4A,TUBB4B,USP1,VDR	
Methotrexate	DHFRi	ABCB1,ABCB11,ABCG2,ADORA2A,AGTR2,APEX1,ATAD5,ATIC,BDKRB2,BLM,CASP7,CBFB,CYP2D6,DHFR,EGFR,ELANE,ERBB2,ERG,ESR1,ESR2,FEN1,FLT1,HMGB1,HMGCR,IDH1,KAT2A,KDM4A,KMT2A,LCK,LMNA,MAPK1,MAPK14,MAPK3,MBNL1,MEN1,MMP2,MMP7,NCOA1,NCOA3,NFE2L2,OPRK1,PABPC1,PIN1,POLB,POLH,POLI,POLK,PPARG,PTPRC,RFC1,RUNX1,RXRA,TDP1,TP53,TSHR,TYMS,USP1,VDR,WRN	CHEMBL; Drugbank; LINCS DGIdb3.0;
Peposertib	DNAPKi	PRKDC	CHEMBL; Drugbank; LINCS DGIdb3.0;
Erlotinib	EGFRi	ABCB1,ABCB11,ABCG2,ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2A,ACVR2B,ACVRL1,AKT1,AKT2,AKT3,ALK,ARAF,ATR,AURKA,AURKB,AXL,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,BRSK2,CAMK2A,CAMK2G,CAMKK1,CASK,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,CDKL2,CHD4,CHEK1,CHEK2,CIT,CLK2,CLK3,CSF1R,CSNK1A1L,CSNK1D,CSNK1E,CSNK1G3,CSNK2A1,DAPK1,DCK,DDR1,DDR2,DDX3X,DDX6,DMPK,DSTYK,DYRK2,EGFR,EIF2AK1,EIF2AK4,EPHA2,EPHA3,EPHA4,EPHA5,EPHA6,EPHA7,EPHB1,EPHB6,ERBB2,ERBB3,ERBB4,ERCC2,FES,FGFR1,FGFR2,FGFR3,FGFR4,FLT1,FLT3,FLT4,GSK3B,HDAC1,HDAC2,HDAC7,HDAC8,HIPK2,IGF1R,IKBKE,INSR,INSRR,IRAK1,IRAK4,ITK,JAK1,JAK2,JAK3,KDR,KIT,LATS1,LATS2,LCK,LMNA,LRRK2,LYN,MAP2K1,MAP2K2,MAP2K4,MAP2K7,MAP3K1,MAP3K13,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP3K7,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MAST1,MATK,MCM4,MET,MTOR,MYO3B,MYT1,NAT10,NEK2,NEK4,NEK6,NEK7,NLK,NTRK1,NTRK2,NTRK3,PAK1,PAK3,PAK4,PBK,PDGFRA,PDGFRB,PDPK1,PEBP1,PIK3C2B,PIK3C2G,PIK3CA,PIK3CB,PIK3CD,PIK3CG,PIM1,PIM2,PIM3,PIP4K2B,PIP5K1A,PKMYT1,PLK1,PLK4,PRKAA1,PRKAA2,PRKAG2,PRKAR2A,PRKCD,PRKCG,PRKCI,PRKCQ,PRKCZ,PRKD1,PRKD2,PTK2B,PTK6,RAF1,RAN,RET,RIOK3,RIPK1,ROCK1,ROCK2,ROS1,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SGK2,SGK3,SMC1A,SMC2,SRC,SRPK2,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TDP1,TGFBR1,TGFBR2,TLK1,TLK2,TNIK,TOP2A,TOP2B,TTK,TYK2,TYRO3,VRK1,WEE1,YES1	CHEMBL; Drugbank; LINCS DGIdb3.0;
Lapatinib	ERBB1i_ERBB2i	ABCB11,ABL1,ABL2,ACTR2,ACTR3,ACVR1B,ACVR2A,ACVR2B,ACVRL1,AKT1,AKT2,AKT3,ALK,ARAF,ATR,AURKA,AURKB,AXL,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,BRSK2,CAMK2A,CAMK2G,CAMKK1,CASK,CCNE1,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,CDKL2,CHD4,CHEK1,CHEK2,CIT,CLK2,CLK3,CSF1R,CSNK1A1L,CSNK1D,CSNK1E,CSNK1G3,CSNK2A1,CYP2D6,DAPK1,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,DMPK,DSTYK,DYRK2,EGFR,EIF2AK1,EIF2AK4,EPHA2,EPHA3,EPHA4,EPHA5,EPHA6,EPHA7,EPHB1,EPHB6,ERBB2,ERBB3,ERBB4,ERCC2,FES,FGFR1,FGFR2,FGFR3,FGFR4,FLT1,FLT3,FLT4,GSK3B,HDAC1,HDAC2,HDAC7,HDAC8,HIPK2,IGF1R,IKBKE,INSR,INSRR,IRAK1,IRAK4,ITK,JAK1,JAK2,JAK3,KDR,KIT,KMT2A,LATS1,LATS2,LCK,LRRK2,LYN,MAP2K1,MAP2K2,MAP2K4,MAP2K7,MAP3K1,MAP3K13,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP3K7,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MAST1,MATK,MCM4,MEN1,MET,MTOR,MYO3B,MYT1,NAT10,NEK2,NEK4,NEK6,NEK7,NLK,NTRK1,NTRK2,NTRK3,PAK1,PAK3,	CHEMBL; Drugbank; LINCS DGIdb3.0;

		PAK4,PBK,PDGFRA,PDGFRB,PDPK1,PEBP1,PIK3C2B,PIK3C2G,PIK3CA,PIK3CB,PIK3CD,PIK3CG,PIM1,PIM2,PIM3,PIP4K2B,PIP5K1A,PKMYT1,PLK1,PLK4,PRKAA1,PRKAA2,PRKAG2,PRKAR2A,PRKCD,PRKCG,PRKCI,PRKCQ,PRKCZ,PRKD1,PRKD2,PTK2B,PTK6,RAF1,RAN,RET,RIOK3,RIPK1,ROCK1,ROCK2,ROS1,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SGK2,SGK3,SMC1A,SMC2,SRC,SRPK2,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TLK1,TLK2,TNIK,TOP2A,TOP2B,TTK,TYK2,TYRO3,WEE1,YES1	
Ulixertinib	ERKi	ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2B,AKT1,AKT2,AKT3,ARAF,AURKA,AURKB,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,CAMK2G,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK9,CHD4,CHEK1,CIT,CSNK1D,CSNK1E,CSNK1G3,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,EGFR,EIF2AK1,EPHA2,EPHA4,EPHA5,EPHB6,ERCC2,FES,FGFR1,FLT3,GSK3B,IGF1R,IKBKE,INSR,IRAK1,IRAK4,JAK1,LATS1,LCK,LYN,MAP2K1,MAP2K2,MAP2K4,MAP3K1,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MCM4,MET,NAT10,NEK2,NEK7,NLK,NTRK1,PAK4,PDGFRB,PIM1,PKMYT1,PLK1,PLK4,PRKAA1,PRKAG2,PRKAR2A,PRKCD,PRKCI,PRKCQ,PRKD2,PTK2B,PTK6,RAN,RET,ROCK1,ROCK2,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SMC1A,SMC2,SRC,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TNIK,TOP2A,TOP2B,TYK2,WEE1,YES1	CHEMBL; Drugbank; LINCS DGIdb3.0;
AZD4547	FGFRi	ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2B,ACVRL1,AKT1,AKT2,AKT3,ALK,ARAF,ATR,AURKA,AURKB,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,CAMK2G,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK9,CHD4,CHEK1,CIT,CLK2,CLK3,CSF1R,CSNK1D,CSNK1E,CSNK1G3,DCK,DDR1,DDR2,DDX3X,DDX6,EGFR,EIF2AK1,EPHA2,EPHA4,EPHA5,EPHA7,EPHB6,ERCC2,FES,FGFR1,FGFR2,FGFR3,FGFR4,FLT1,FLT3,GSK3B,IGF1R,IKBKE,INSR,IRAK1,IRAK4,JAK1,JAK3,KDR,LATS1,LCK,LYN,MAP2K1,MAP2K2,MAP3K1,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MCM4,MET,NAT10,NEK2,NLK,NTRK1,NTRK2,NTRK3,PAK4,PDGFRB,PIM1,PIM2,PKMYT1,PLK1,PLK4,PRKAA1,PRKAG2,PRKAR2A,PRKCD,PRKCI,PRKCQ,PRKCZ,PRKD2,PTK2B,PTK6,RAN,RET,RIPK1,ROCK1,ROCK2,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SMC1A,SMC2,SRC,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TNIK,TOP2A,TOP2B,TYK2,WEE1,YES1	CHEMBL; Drugbank; LINCS DGIdb3.0;
Gilteritinib	FLT3i_AXLi	ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2B,AKT1,AKT2,AKT3,ALK,ARAF,AURKA,AURKB,AXL,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,CAMK2G,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK9,CHD4,CHEK1,CIT,CSNK1D,CSNK1E,CSNK1G3,DCK,DDR1,DDR2,DDX3X,DDX6,EGFR,EIF2AK1,EML4,EPHA2,EPHA4,EPHA5,EPHA7,EPHB6,ERCC2,ETV6,FES,FGFR1,FLT3,GSK3B,IGF1R,IKBKE,INSR,IRAK1,IRAK4,JAK1,LATS1,LCK,LYN,MAP2K1,MAP2K2,MAP3K1,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MCM4,MET,NAT10,NEK2,NLK,NTRK1,PAK4,PDGFRB,PIM1,PKMYT1,PLK4,PRKAA1,PRKAG2,PRKAR2A,PRKCD,PRKCI,PRKD2,PTK2B,PTK6,RAC1,RAN,RET	CHEMBL; Drugbank; LINCS DGIdb3.0;

		,ROCK1,ROCK2,ROS1,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SMC1A,SMC2,SRC,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TNIK,TOP2A,TOP2B,TTK,TYK2,WEE1,YES1	
PLX647	FMSi_KITi	CSF1R,KIT	CHEMBL; DGIdb3.0; Drugbank; LINCS
Geldanamycin	HSP90i	ABL1,ATAD5,BCR,CAMK2A,CAMK2G,CBFB,CDK2,CDK6,CHEK2,CLK2,CLK3,CSNK1G3,CYP2D6,DMPK,ERBB2,FES,GSK3B,HDAC1,HDAC2,HDAC7,HDAC8,HIF1A,HSP90AA1,HSP90AB1,JAK1,LMNA,MAP2K2,MAP3K5,MAPK3,MTOR,NEK2,NEK6,PAK4,PBK,PDPK1,PIM1,PIM2,PIM3,PLK1,PLK4,PRKAA2,PRKDC,RIPK1,RPS6KA3,RUNX1,STK16,STK3,STK4,TDP1,TNIK,USP1,VDR,VRK1	CHEMBL; DGIdb3.0; Drugbank; LINCS
GAMMA	IR	ABCB1,ABCB11,ADORA2A,AGTR2,AKT1,APEX1,APH1A,APH1B,AR,ATAD5,ATM,ATR,BDKRB2,BLM,BRCA1,CASP6,CBFB,CDK2,CREBBP,CYP2D6,DHFR,EGFR,ELANE,ERBB2,ERG,ESR1,ESR2,FDPS,FEN1,FLT1,GABBR2,GABRA6,GBA,GNAS,GSK3B,HDAC1,HDAC2,HDAC7,HDAC8,HMGCR,HRH4,HSD11B2,HSP90AA1,IDH1,KAT2A,KDM4A,KIF11,KMT2A,LCK,LMNA,MAPK1,MAPK14,MAPK3,MBNL1,MDM2,MEN1,MGMT,MMP2,MMP7,MTOR,NCSTN,NFE2L2,NFKB1,NOS2,NOS3,NR1H2,NTRK1,NUDT1,OPRK1,PDPK1,PHLPP1,PIN1,PLK1,POLB,POLH,POLI,POLK,PPARG,PRKAA1,PRKAA2,PRKAB1,PRKAG2,PRKDC,PSEN1,PSEN2,PSENEN,PTPRC,RELA,RET,RUNX1,RXRA,SMAD3,TDP1,TERT,TOP2A,TSHR,TUBA1A,TUBA4A,TUBB4B,USP1,VDR,WRN	CHEMBL; DGIdb3.0; Drugbank; LINCS
Ruxolitinib	JAKi	ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2A,ACVR2B,ACVRL1,AKT1,AKT2,AKT3,ALK,ARAF,AURKA,AURKB,AXL,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,BRSK2,CAMK2A,CAMK2G,CAMKK1,CASK,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,CDKL2,CHD4,CHEK1,CHEK2,CIT,CLK2,CLK3,CSF1R,CSNK1A1L,CSNK1D,CSNK1E,CSNK1G3,CSNK2A1,DAPK1,DCK,DDR1,DDR2,DDX3X,DDX6,DMPK,DSTYK,DYRK2,EGFR,EIF2AK1,EIF2AK4,EPHA2,EPHA3,EPHA4,EPHA5,EPHA6,EPHA7,EPHB1,EPHB6,ERBB2,ERBB3,ERBB4,ERCC2,FES,FGFR1,FGFR2,FGFR3,FGFR4,FLG,FLT1,FLT3,FLT4,GSK3B,HDAC1,HDAC2,HIPK2,IGF1R,IKBKE,INSR,INSRR,IRAK1,IRAK4,ITK,JAK1,JAK2,JAK3,KDR,KIT,LATS1,LATS2,LCK,LRRK2,LYN,MAP2K1,MAP2K2,MAP2K4,MAP2K7,MAP3K1,MAP3K13,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP3K7,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MAST1,MATK,MCM4,MET,MTOR,MYO3B,NAT10,NEK2,NEK4,NEK6,NEK7,NLK,NR1H2,NTRK1,NTRK2,NTRK3,PAK1,PAK3,PAK4,PDGFRA,PDGFRB,PDPK1,PEBP1,PIK3C2B,PIK3C2G,PIK3CA,PIK3CB,PIK3CD,PIK3CG,PIM1,PIM2,PIM3,PIP4K2B,PIP5K1A,PKMYT1,PLK1,PLK4,PRKAA1,PRKAA2,PRKAG2,PRKAR2A,PRKCD,PRKCI,PRKCQ,PRKD1,PRKD2,PTK2B,PTK6,RAF1,RAN,RET,RIOK3,RIPK1,ROCK1,ROCK2,ROS1,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SGK3,SMC1A,SMC2,SRC,SRPK2,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TLK1,TLK2,TNIK,TOP2B,TTK,TYK2,TYRO3,WEE1,YES1	CHEMBL; DGIdb3.0; Drugbank; LINCS
K858	KIF11i	APEX1,BLM,CBFB,IDH1,KIF11,LMNA,MBNL1,POLB,RECQL,RUNX1,TDP1	CHEMBL; DGIdb3.0; Drugbank; LINCS
Ralimetinib	MAPKi	MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9	CHEMBL; DGIdb3.0; Drugbank; LINCS

Nutlin	MDM2i	BCL2,DAPK1,HDAC1,HIF1A,MCL1,MDM2,MDM4,TP53,VDR,YES1	CHEMBL; DGIdb3.0; Drugbank; LINCS
PF04217903	METi	ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2B,AKT1,AKT2,AKT3,ALK,ARAF,ATR,AURKA,AURKB,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,CAMK2A,CAMK2G,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,CHD4,CHEK1,CIT,CLK2,CLK3,CSF1R,CSNK1D,CSNK1E,CSNK1G3,CYP2D6,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,EGFR,EIF2AK1,EPHA2,EPHA4,EPHA5,EPHA7,EPHB6,ERCC2,FES,FGFR1,FLG,FLT1,FLT3,GSK3B,IGF1R,IKBKE,INSR,IRAK1,IRAK4,JAK1,JAK2,JAK3,KDR,LATS1,LCK,LRRK2,LYN,MAP2K1,MAP2K2,MAP2K4,MAP3K1,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MCM4,MET,NAT10,NEK2,NEK7,NLK,NTRK1,NTRK2,NTRK3,PAK4,PDE10A,PDE1C,PDE4B,PDGFRA,PDGFRB,PEBP1,PIM1,PIM2,PKMYT1,PLK1,PLK4,PRKAA1,PRKAG2,PRKAR2A,PRKCD,PRKCI,PRKCQ,PRKCZ,PRKD2,PTK2B,PTK6,RAN,RET,ROCK1,ROCK2,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SMC2,SRC,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TNIK,TOP2A,TOP2B,TTK,TYK2,TYRO3,WEE1,YES1	CHEMBL; DGIdb3.0; Drugbank; LINCS
SGX523	METi	ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2A,ACVR2B,ACVRL1,AKT1,AKT2,AKT3,ALK,ARAF,AURKA,AURKB,AXL,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,BRSK2,CAMK2A,CAMK2G,CAMKK1,CASK,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,CDKL2,CHD4,CHEK1,CHEK2,CIT,CLK2,CLK3,CSF1R,CSNK1A1L,CSNK1D,CSNK1E,CSNK1G3,CSNK2A1,DAPK1,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,DMPK,DSTYK,DYRK2,EGFR,EIF2AK1,EIF2AK4,EPHA2,EPHA3,EPHA4,EPHA5,EPHA6,EPHA7,EPHB1,EPHB6,ERBB2,ERBB3,ERBB4,ERCC2,FES,FGFR1,FGFR2,FGFR3,FGFR4,FLT1,FLT3,FLT4,GSK3B,HIPK2,IGF1R,IKBKE,INSR,INSRR,IRAK1,IRAK4,ITK,JAK1,JAK2,JAK3,KDR,KIT,LATS1,LATS2,LCK,LRRK2,LYN,MAP2K1,MAP2K2,MAP2K4,MAP2K7,MAP3K1,MAP3K13,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP3K7,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MAST1,MATK,MCM4,MET,MTOR,MYO3B,NAT10,NEK2,NEK4,NEK6,NEK7,NLK,NTRK1,NTRK2,NTRK3,PAK1,PAK3,PAK4,PDGFRA,PDGFRB,PDPK1,PEBP1,PIK3C2B,PIK3C2G,PIK3CA,PIK3CB,PIK3CD,PIK3CG,PIM1,PIM2,PIM3,PIP4K2B,PIP5K1A,PKMYT1,PLK1,PLK4,PRKAA1,PRKAA2,PRKAG2,PRKAR2A,PRKCD,PRKCI,PRKCQ,PRKD1,PRKD2,PTK2B,PTK6,RAF1,RAN,RET,RIOK3,RIPK1,ROCK1,ROCK2,ROS1,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SGK3,SMC1A,SMC2,SRC,SRPK2,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TLK1,TLK2,TNIK,TOP2A,TOP2B,TTK,TYK2,TYRO3,WEE1,YES1	CHEMBL; DGIdb3.0; Drugbank; LINCS
Cabozantinib	METi	ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2B,AKT1,AKT2,AKT3,ALK,ARAF,AURKA,AURKB,AXL,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,CAMK2G,CCDC6,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK9,CHD4,CHEK1,CIT,CLK2,CLK3,CSNK1D,CSNK1E,CSNK1G3,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,EGFR,EIF2AK1,EPHA2,EPHA4,EPHA5,EPHA7,EPHB6,ERCC2,FES,FGFR1,FLT1,FLT3,FLT4,GSK3B,IGF1R,IKBKE,INSR,IRAK1,IRAK4,JAK1,JAK2,KDR,KIF5B,KIT,LATS1,LCK,LYN,MAP2K1,MAP2K2,MAP3K1,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP4K3,MAP4K4,MAPK1,MAPK1	CHEMBL; DGIdb3.0; Drugbank; LINCS

		0,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MCM4,MET,NAT10,NEK2,NEK7,NLK,NTRK1,NTRK2,PAK4,PDGFRA,PDGFRB,PEBP1,PIM1,PKMYT1,PLK1,PLK4,PRKAA1,PRKAG2,PRKAR2A,PRKCD,PRKCI,PRKCQ,PRKCZ,PRKD2,PTK2B,PTK6,RAN,RET,ROCK1,ROCK2,ROS1,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SMC1A,SMC2,SRC,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TNIK,TOP2B,TTK,TYK2,TYRO3,WEE1,YES1	
Crizotinib	METi_ALKi	ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2A,ACVR2B,ACVRL1,AKT1,AKT2,AKT3,ALK,ARAF,AURKA,AURKB,AXL,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,BRSK2,CAMK2A,CAMK2G,CAMKK1,CASK,CCNB1,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,CDKL2,CHD4,CHEK1,CHEK2,CIT,CLK2,CLK3,CSF1R,CSNK1A1L,CSNK1D,CSNK1E,CSNK1G3,CSNK2A1,DAPK1,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,DMPK,DSTYK,DYRK2,EGFR,EIF2AK1,EIF2AK4,EML4,EPHA2,EPHA3,EPHA4,EPHA5,EPHA6,EPHA7,EPHB1,EPHB6,ERBB2,ERBB3,ERBB4,ERCC2,FES,FGFR1,FGFR2,FGFR3,FGFR4,FLT1,FLT3,FLT4,GSK3B,HIPK2,IGF1R,IKBKE,INPPL1,INSR,INSRR,IRAK1,IRAK4,ITK,JAK1,JAK2,JAK3,KDR,KIT,LATS1,LATS2,LCK,LRRK2,LYN,MAP2K1,MAP2K2,MAP2K4,MAP2K7,MAP3K1,MAP3K13,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP3K7,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MAST1,MATK,MCM4,MET,MLKL,MST1,MTOR,MYO3B,NAT10,NEK2,NEK4,NEK6,NEK7,NEK8,NLK,NPM1,NTRK1,NTRK2,NTRK3,NUDT1,PAK1,PAK3,PAK4,PDGFRA,PDGFRB,PDPK1,PEBP1,PIK3C2B,PIK3C2G,PIK3CA,PIK3CB,PIK3CD,PIK3CG,PIM1,PIM2,PIM3,PIP4K2B,PIP5K1A,PKMYT1,PLK1,PLK4,PRKAA1,PRKAA2,PRKAG2,PRKAR2A,PRKCD,PRKCI,PRKCQ,PRKD1,PRKD2,PTK2B,PTK6,RAF1,RAN,RET,RIOK3,RIPK1,ROCK1,ROCK2,ROS1,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SGK3,SMC1A,SMC2,SRC,SRPK2,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TLK1,TLK2,TNIK,TOP2A,TOP2B,TTK,TYK2,TYRO3,WEE1,YES1	CHEMBL; Drugbank; LINCS DGIdb3.0;
Everolimus	mTORi	ABCB11,ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2B,AKT1,AKT2,AKT3,ARAF,AURKA,AURKB,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,CAMK2G,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK9,CHD4,CHEK1,CIT,CLK2,CLK3,CSNK1D,CSNK1E,CSNK1G3,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,EGFR,EIF2AK1,EPHA2,EPHA4,EPHA5,EPHA7,EPHB6,ERCC2,FES,FGFR1,FKBP1A,FLT3,GSK3B,IGF1R,IKBKE,INSR,IRAK1,IRAK4,JAK1,LATS1,LCK,LYN,MAP2K1,MAP2K2,MAP3K1,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MARK3,MARK4,MCM4,MET,MLST8,MTOR,NAT10,NEK2,NEK7,NLK,NTRK1,PAK4,PDGFRB,PEBP1,PIM1,PKMYT1,PLK1,PLK4,PRKAA1,PRKAG2,PRKAR2A,PRKCD,PRKCI,PRKCQ,PRKD2,PTK2B,PTK6,RAN,RET,ROCK1,ROCK2,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,RPTOR,SMC1A,SMC2,SRC,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TDP1,TGFBR1,TGFBR2,TNIK,TOP2A,TOP2B,TTK,TYK2,WEE1,YES1	CHEMBL; Drugbank; LINCS DGIdb3.0;
OSI027	mTORi	ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2B,AKT1,AKT2,AKT3,ARAF,AURKA,AURKB,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,CAMK2G,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK9,CHD4,CHEK1,CIT,CLK2,CSNK1D,CSNK1	CHEMBL; Drugbank; LINCS DGIdb3.0;

		E,CSNK1G3,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,EGFR,EIF2AK1,EPHA2,EPHA4,EPHA5,EPHA7,EPHB6,ERCC2,FES,FGFR1,FLT3,GSK3B,IGF1R,IKBKE,INSR,IRAK1,IRAK4,JAK1,LATS1,LCK,LYN,MAP2K1,MAP2K2,MAP3K1,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MCM4,NAT10,NEK2,NLK,NTRK1,PAK4,PDGFRB,PEBP1,PIM1,PIM2,PKMYT1,PLK4,PRKAA1,PRKAG2,PRKAR2A,PRKCD,PRKCI,PRKCQ,PRKCZ,PRKD2,PTK2B,PTK6,RAN,RET,ROCK1,ROCK2,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SMC1A,SMC2,SRC,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TNIK,TOP2A,TOP2B,TYK2,WEE1,YES1	
NSC348884	NPMi	NPM1	CHEMBL; DGIdb3.0; Drugbank; LINCS
LY2584702	p70S6Ki	ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2B,AKT1,AKT2,AKT3,ARAF,AURKA,AURKB,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,CAMK2G,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK9,CHD4,CHEK1,CIT,CSNK1D,CSNK1E,CSNK1G3,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,EGFR,EIF2AK1,EPHA2,EPHA4,EPHA5,EPHA7,EPHB6,ERCC2,FES,FGFR1,FLT3,GSK3B,IGF1R,IKBKE,INSR,IRAK1,IRAK4,JAK1,LATS1,LCK,LYN,MAP2K1,MAP2K2,MAP3K1,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MET,NAT10,NEK2,NLK,NTRK1,PAK4,PDGFRB,PEBP1,PIM1,PIM2,PKMYT1,PLK4,PRKAA1,PRKAG2,PRKCD,PRKCI,PRKCQ,PRKD2,PTK2B,PTK6,RAN,RET,ROCK1,ROCK2,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,RPS6KB2,SMC2,SRC,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TNIK,TOP2B,TYK2,WEE1,YES1	CHEMBL; DGIdb3.0; Drugbank; LINCS
Veliparib	PARPi	CYP2D6,PARP1,PARP2,PARP3,PARP4,TNKS,TNKS2	CHEMBL; DGIdb3.0; Drugbank; LINCS
Rucaparib	PARPi	AKT1,AKT2,AKT3,AURKA,AURKB,CDK1,CDK2,CDK5,CHEK1,CHEK2,CSNK1D,CSNK2A1,EPHA2,FGFR1,FLG,FLT4,GSK3B,INSR,IRAK4,KDR,MAPK1,MARK3,NEK2,PAK1,PAK4,PARP1,PARP2,PARP3,PARP4,PIM1,PLK1,PRKAA1,PRKCD,PRKCG,PRKCZ,PRKD2,ROCK1,ROCK2,RPS6KA3,RPS6KB1,SGK2,STK3,TNKS,TNKS2,YES1	CHEMBL; DGIdb3.0; Drugbank; LINCS
Talazoparib	PARPi	CYP2D6,PARP1,PARP2,PARP3,PARP4,TNKS,TNKS2	CHEMBL; DGIdb3.0; Drugbank; LINCS
Niraparib	PARPi	PARP1,PARP2,PARP3,PARP4,TNKS,TNKS2	CHEMBL; DGIdb3.0; Drugbank; LINCS
Olaparib	PARPi	HDAC1,PARP1,PARP2,PARP3,PARP4,TNKS,TNKS2	CHEMBL; DGIdb3.0; Drugbank; LINCS
GSK2334470	PDK1i	AKT1,AKT3,AURKA,AURKB,BRSK2,CHEK2,EGFR,GSK3B,KDR,MAP3K5,MET,NLK,PDK1,PDPK1,PIK3CG,PRKCQ,RET,ROCK1,ROCK2,RPS6KA1,RPS6KA3,RPS6KA6,SGK1,SGK2,SYK,TGFBR1	CHEMBL; DGIdb3.0; Drugbank; LINCS
Duvelisib	PI3Ki	PIK3CA,PIK3CB,PIK3CD,PIK3CG,PIK3R1	CHEMBL; DGIdb3.0; Drugbank; LINCS
Pictilisib	PI3Ki	ABCB11,ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2A,ACVR2B,ACVRL1,AKT1,AKT2,AKT3,ALK,ARAF,AURKA,AURKB,AXL,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,BRSK	CHEMBL; DGIdb3.0; Drugbank; LINCS

		2,CAMK2A,CAMK2G,CAMKK1,CASK,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,CDKL2,CHD4,CHEK1,CHEK2,CIT,CLK2,CLK3,CSF1R,CSNK1A1L,CSNK1D,CSNK1E,CSNK1G3,CSNK2A1,CYP2D6,DAPK1,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,DMPK,DSTYK,DYRK2,EGFR,EIF2AK1,EIF2AK4,EPHA2,EPHA3,EPHA4,EPHA5,EPHA6,EPHA7,EPHB1,EPHB6,ERBB2,ERBB3,ERBB4,ERCC2,FES,FGFR1,FGFR2,FGFR3,FGFR4,FLT1,FLT3,FLT4,GSK3B,HIPK2,IGF1R,IKBKE,INSR,INSRR,IRAK1,IRAK4,ITK,JAK1,JAK2,JAK3,KDR,KIT,LATS1,LATS2,LCK,LRRK2,LYN,MAP2K1,MAP2K2,MAP2K4,MAP2K7,MAP3K1,MAP3K13,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP3K7,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MAST1,MATK,MCM4,MET,MTOR,MYO3B,NAT10,NEK2,NEK4,NEK6,NEK7,NLK,NTRK1,NTRK2,NTRK3,PAK1,PAK3,PAK4,PDGFRA,PDGFRB,PDPK1,PEBP1,PIK3C2B,PIK3C2G,PIK3C3,PIK3CA,PIK3CB,PIK3CD,PIK3CG,PIK3R1,PIK3R2,PIK3R3,PIM1,PIM2,PIM3,PIP4K2B,PIP5K1A,PKMYT1,PLK1,PLK4,PRKAA1,PRKAA2,PRKAG2,PRKAR2A,PRKCD,PRKCI,PRKCQ,PRKCZ,PRKD1,PRKD2,PRKDC,PTK2B,PTK6,RAF1,RAN,RET,RIOK3,RIPK1,ROCK1,ROCK2,ROS1,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SGK3,SMC1A,SMC2,SRC,SRPK2,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TDP1,TGFBR1,TGFBR2,TLK1,TLK2,TNIK,TOP2A,TOP2B,TTK,TYK2,TYRO3,WEE1,YES1	
BI2536	PLK1i	ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2A,ACVR2B,ACVRL1,AKT1,AKT2,AKT3,ALK,ARAF,AURKA,AURKB,AXL,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,BRD4,BRSK2,CAMK2A,CAMK2G,CAMKK1,CASK,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,CDKL2,CHD4,CHEK1,CHEK2,CIT,CLK2,CLK3,CSF1R,CSNK1A1L,CSNK1D,CSNK1E,CSNK1G3,CSNK2A1,DAPK1,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,DMPK,DSTYK,DYRK2,EGFR,EIF2AK1,EIF2AK4,EPHA2,EPHA3,EPHA4,EPHA5,EPHA6,EPHA7,EPHB1,EPHB6,ERBB2,ERBB3,ERBB4,ERCC2,FES,FGFR1,FGFR2,FGFR3,FGFR4,FLT1,FLT3,FLT4,GSK3B,HIPK2,IGF1R,IKBKE,INSR,INSRR,IRAK1,IRAK4,ITK,JAK1,JAK2,JAK3,KDR,KIT,LATS1,LATS2,LCK,LRRK2,LYN,MAP2K1,MAP2K2,MAP2K4,MAP2K7,MAP3K1,MAP3K13,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP3K7,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MAST1,MATK,MCM4,MET,MTOR,MYO3B,NAT10,NEK2,NEK4,NEK6,NEK7,NLK,NTRK1,NTRK2,NTRK3,PAK1,PAK3,PAK4,PBK,PDGFRA,PDGFRB,PDPK1,PIK3C2B,PIK3C2G,PIK3CA,PIK3CB,PIK3CD,PIK3CG,PIM1,PIM2,PIM3,PIP4K2B,PIP5K1A,PKMYT1,PLK1,PLK4,PRKAA1,PRKAA2,PRKAG2,PRKAR2A,PRKCD,PRKCG,PRKCI,PRKCQ,PRKCZ,PRKD1,PRKD2,PTK2B,PTK6,RAF1,RAN,RET,RIOK3,RIPK1,ROCK1,ROCK2,ROS1,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SGK2,SGK3,SMC1A,SMC2,SRC,SRPK2,STK11,STK16,STK3,STK4,STRADA,SYK,TAF1,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TLK1,TLK2,TNIK,TOP2A,TOP2B,TTK,TYK2,TYRO3,WEE1,YES1	CHEMBL; DGIdb3.0; Drugbank; LINCS
Bortezomib	Proteasomei	ABCB11,ADAM17,APH1A,APH1B,BAX,BCL2,CASP2,CASP3,CASP7,CASP8,CASP9,CTSC,ELANE,EPAS1,FLT3,HIF1A,MME,MMP2,MMP7,NCSTN,NFKB1,NFKB2,PLAT,PRSS1,PSEN1,PSEN2,PSENEN,PSMB3,PSMD11,PSMD9,RELA,TDP1,TP53	CHEMBL; DGIdb3.0; Drugbank; LINCS
Carfilzomib	Proteasomei	CASP3,CYP2D6,ESR1,PRSS1,PSMB3,PSMD11,PSMD9,TDP1,VHL	CHEMBL; DGIdb3.0; Drugbank; LINCS

GSK429286A	ROCK1i	ROCK1,ROCK2,RPS6KA1,RPS6KB1	CHEMBL; DGIdb3.0; Drugbank; LINCS
Sunitinib	RTKi	ABCB11,ABL1,ABL2,ACTR2,ACTR3,ACVR1B,ACVR2A,ACVR2B,ACVRL1,AKT1,AKT2,AKT3,ALK,APH1A,APH1B,ARAF,ATR,AURKA,AURKB,AXL,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,BRD7,BRD9,BRSK2,CAMK2A,CAMK2G,CAMKK1,CASK,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,CDKL2,CHD4,CHEK1,CHEK2,CIT,CLK2,CLK3,CSF1R,CSNK1A1L,CSNK1D,CSNK1E,CSNK1G3,CSNK2A1,DAPK1,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,DLK1,DMPK,DSTYK,DYRK2,EGFR,EIF2AK1,EIF2AK4,EPHA2,EPHA3,EPHA4,EPHA5,EPHA6,EPHA7,EPHB1,EPHB6,ERBB2,ERBB3,ERBB4,ERCC2,FES,FGFR1,FGFR2,FGFR3,FGFR4,FLG,FLT1,FLT3,FLT4,GSK3B,HIPK2,IGF1R,IKBKE,INSR,INSRR,IRAK1,IRAK4,ITK,JAK1,JAK2,JAK3,KDR,KIT,KMT2A,LATS1,LATS2,LCK,LMNA,LRRK2,LYN,MAP2K1,MAP2K2,MAP2K4,MAP2K7,MAP3K1,MAP3K13,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP3K7,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MAST1,MATK,MCM4,MEN1,MET,MST1,MTOR,MYO3B,MYT1,NAT10,NCSTN,NEK2,NEK4,NEK6,NEK7,NLK,NOS3,NR2C2,NTRK1,NTRK2,NTRK3,PAK1,PAK3,PAK4,PBK,PDGFB,PDGFC,PDGFD,PDGFRA,PDGFRB,PDK1,PDPK1,PEBP1,PIK3C2B,PIK3C2G,PIK3CA,PIK3CB,PIK3CD,PIK3CG,PIM1,PIM2,PIM3,PIP4K2B,PIP5K1A,PKMYT1,PLK1,PLK4,POLK,PRKAA1,PRKAA2,PRKAB1,PRKAG2,PRKAR2A,PRKCD,PRKCG,PRKCI,PRKCQ,PRKCZ,PRKD1,PRKD2,PRKDC,PSEN1,PSEN2,PSENEN,PTK2B,PTK6,RAF1,RAN,RET,RIOK3,RIPK1,ROCK1,ROCK2,ROS1,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SGK1,SGK2,SGK3,SIRT1,SMC1A,SMC2,SRC,SRPK2,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TGFBR1,TGFBR2,TLK1,TLK2,TNIK,TNKS2,TOP2A,TOP2B,TTK,TYK2,TYRO3,USP1,VRK1,WEE1,YES1	CHEMBL; DGIdb3.0; Drugbank; LINCS
Entospletinib	SYKi	CYP2D6,FLT3,JAK2,KDR,RET,SYK	CHEMBL; DGIdb3.0; Drugbank; LINCS
SN38	TOP1i	ABCG2,HDAC1,HDAC2,HDAC7,HDAC8,TOP1	CHEMBL; DGIdb3.0; Drugbank; LINCS
Topotecan	TOP1i	ABCB1,ABCG2,ADORA2A,AGTR2,AKT1,BDKRB2,BRCA1,CASP7,CYP2D6,EGFR,ELANE,EPAS1,ERBB2,ESR1,ESR2,FLT1,HIF1A,HMGCR,IDH1,KAT2A,KEAP1,KMT2A,LCK,LMNA,MAPK1,MAPK14,MAPK3,MEN1,NFE2L2,OPRK1,PAX8,PLK1,POLI,PTPRC,RAPGEF3,SMAD3,TDP1,TOP1,TOP2A,TOP2B,USP1	CHEMBL; DGIdb3.0; Drugbank; LINCS
Doxorubicin	TOP2i	ABCB1,ABCB11,ABCG2,ADORA2A,AGTR2,APAF1,APEX1,AR,ATAD5,AURKA,BCL2,BCL2L1,BDKRB2,BLM,BRCA1,CYP2D6,DHFR,EGFR,ELANE,EPAS1,ERBB2,ERG,ESR1,ESR2,FANCD2,FEN1,FLT1,HDAC1,HDAC2,HDAC7,HDAC8,HIF1A,HMGCR,HSP90AA1,HSP90AB1,IDH1,KAT2A,KDR,KEAP1,KMT2A,LCK,MAPK1,MAPK14,MAPK3,MBNL1,MEN1,MMP2,NFE2L2,NFKB1,OPRK1,PAX8,PDGFRB,PIK3CA,PIK3R1,PLK1,POLB,POLH,POLI,POLK,PPARG,PPM1D,PRSS1,PTPRC,RAPGEF3,RECQL,RORC,RXRA,SMAD3,SNCA,STAT6,TDP1,TERT,TNF,TOP1,TOP2A,TOP2B,TOP3A,TP53,TUBA1A,TUBA4A,TUBB4B,USP1,VDR,WRN	CHEMBL; DGIdb3.0; Drugbank; LINCS
Etoposide	TOP2i	ABCB1,ABCB11,ABCG2,ADORA2A,AGTR2,ATM,BDKRB2,CASP3,CHEK1,CYP2D6,EGFR,ELANE,ERBB2,ESR1,ESR2,FLT1,HDAC1,HMGCR,LCK,MAPK1,MAPK14,MAPK3,MTOR,NC	CHEMBL; DGIdb3.0; Drugbank; LINCS

		OA1,NCOA3,NFE2L2,OPRK1,POLI,PTPRC,TDP1,TOP1,TOP2A,TOP2B,TUBA1A	
Sorafenib	VEGFRi	ABCB11,ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2A,ACVR2B,ACVRL1,AKT1,AKT2,AKT3,ALK,ARAF,ATAD5,ATR,AURKA,AURKB,AXL,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,BRSK2,CAMK2A,CAMK2G,CAMKK1,CASK,CCNA1,CCNA2,CCNC,CCND1,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,CDKL2,CHD4,CHEK1,CHEK2,CIT,CLK2,CLK3,CSF1R,CSNK1A1L,CSNK1D,CSNK1E,CSNK1G3,CSNK2A1,DAPK1,DCK,DDR1,DDR2,DDX3X,DDX6,DLK1,DMPK,DSTYK,DYRK2,EGFR,EIF2AK1,EIF2AK4,EPHA2,EPHA3,EPHA4,EPHA5,EPHA6,EPHA7,EPHB1,EPHB6,ERBB2,ERBB3,ERBB4,ERCC2,FES,FGFR1,FGFR2,FGFR3,FGFR4,FLG,FLT1,FLT3,FLT4,GNAS,GSK3B,HDAC1,HDAC8,HIPK2,IDH1,IGF1R,IKBKE,INSR,INSRR,IRAK1,IRAK4,ITK,JAK1,JAK2,JAK3,KDM4A,KDR,KIT,KMT2A,LATS1,LATS2,LCK,LRRK2,LYN,MAP2K1,MAP2K2,MAP2K4,MAP2K7,MAP3K1,MAP3K13,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP3K7,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MAST1,MATK,MCM4,MEN1,MET,MTOR,MYO3B,NAT10,NEK2,NEK4,NEK6,NEK7,NFE2L2,NLK,NTRK1,NTRK2,NTRK3,PAK1,PAK3,PAK4,PAX8,PBK,PDGFRA,PDGFRB,PDPK1,PIK3C2B,PIK3C2G,PIK3CA,PIK3CB,PIK3CD,PIK3CG,PIM1,PIM2,PIM3,PIP4K2B,PIP5K1A,PKMYT1,PLK1,PLK4,POLI,PRKAA1,PRKAA2,PRKAG2,PRKAR2A,PRKCD,PRKCG,PRKCI,PRKCQ,PRKCZ,PRKD1,PRKD2,PRKDC,PTK2B,PTK6,PTPN6,RAF1,RAN,RET,RIOK3,RIPK1,ROCK1,ROCK2,ROS1,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SGK1,SGK2,SGK3,SMAD3,SMC1A,SMC2,SNCA,SRC,SRPK2,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TDP1,TGFBR1,TGFBR2,TLK1,TLK2,TNIK,TNKS2,TOP2A,TOP2B,TTK,TYK2,TYRO3,USP1,WEE1,YES1	CHEMBL; DGIdb3.0; Drugbank; LINCS
Vatalanib	VEGFRi	ABCB11,ABL1,ABL2,ACAD10,ACTR2,ACTR3,ACVR1B,ACVR2A,ACVR2B,ACVRL1,AKT1,AKT2,AKT3,ALK,ARAF,AURKA,AURKB,AXL,BCR,BMP2K,BMPR1A,BMPR1B,BRAF,BRSK2,CAMK2A,CAMK2G,CAMKK1,CASK,CDC7,CDK1,CDK12,CDK2,CDK4,CDK5,CDK6,CDK7,CDK8,CDK9,CDKL2,CHD4,CHEK1,CHEK2,CIT,CLK2,CLK3,CSF1R,CSNK1A1L,CSNK1D,CSNK1E,CSNK1G3,CSNK2A1,DAPK1,DCK,DDR1,DDR2,DDX3X,DDX42,DDX6,DMPK,DSTYK,DYRK2,EGFR,EIF2AK1,EIF2AK4,EPHA2,EPHA3,EPHA4,EPHA5,EPHA6,EPHA7,EPHB1,EPHB6,ERBB2,ERBB3,ERBB4,ERCC2,FES,FGFR1,FGFR2,FGFR3,FGFR4,FLT1,FLT3,FLT4,GSK3B,HIPK2,IGF1R,IKBKE,INSR,INSRR,IRAK1,IRAK4,ITK,JAK1,JAK2,JAK3,KDR,KIT,LATS1,LATS2,LCK,LRRK2,LYN,MAP2K1,MAP2K2,MAP2K4,MAP2K7,MAP3K1,MAP3K13,MAP3K2,MAP3K3,MAP3K4,MAP3K5,MAP3K7,MAP4K3,MAP4K4,MAPK1,MAPK10,MAPK14,MAPK15,MAPK3,MAPK8,MAPK9,MAPKAPK5,MARK3,MARK4,MAST1,MATK,MCM4,MET,MTOR,MYO3B,MYT1,NAT10,NEK2,NEK4,NEK6,NEK7,NLK,NTRK1,NTRK2,NTRK3,PAK1,PAK3,PAK4,PBK,PDGFRA,PDGFRB,PDPK1,PIK3C2B,PIK3C2G,PIK3CA,PIK3CB,PIK3CD,PIK3CG,PIM1,PIM2,PIM3,PIP4K2B,PIP5K1A,PKMYT1,PLK1,PLK4,PRKAA1,PRKAA2,PRKAG2,PRKAR2A,PRKCD,PRKCG,PRKCI,PRKCQ,PRKCZ,PRKD1,PRKD2,PTK2B,PTK6,RAF1,RAN,RET,RIOK3,RIPK1,ROCK1,ROCK2,ROS1,RPS6KA1,RPS6KA3,RPS6KA4,RPS6KA5,RPS6KA6,RPS6KB1,SGK2,SGK3,SMC1A,SMC2,SRC,SRPK2,STK11,STK16,STK3,STK4,STRADA,SYK,TAOK1,TAOK2,TAOK3,TDP1,	CHEMBL; DGIdb3.0; Drugbank; LINCS

		TGFBR1,TGFBR2,TLK1,TLK2,TNIK,TOP2A,TOP2B,TTK,TYK2,TYRO3,VRK1,WEE1,YES1	
BML284	WNTi	WNT1,WNT11,WNT16,WNT2,WNT3,WNT4,WNT6	CHEMBL; DGIdb3.0; Drugbank; LINCS

Supplementary Table 4.2. The top ten directly targeted genes (among the 272 genes directly targeted by all drugs in this study) that achieved the highest efficacy (AoC score) across all cell lines in combination with the inhibition of drug targets ATM, ATR, or DNA-PK (PRKDC) using the best model predicting across cell lines. **: genes that occur in the top ten in combination with all three drug targets. *: genes that occur in the top ten in combination with two out of three drug targets.

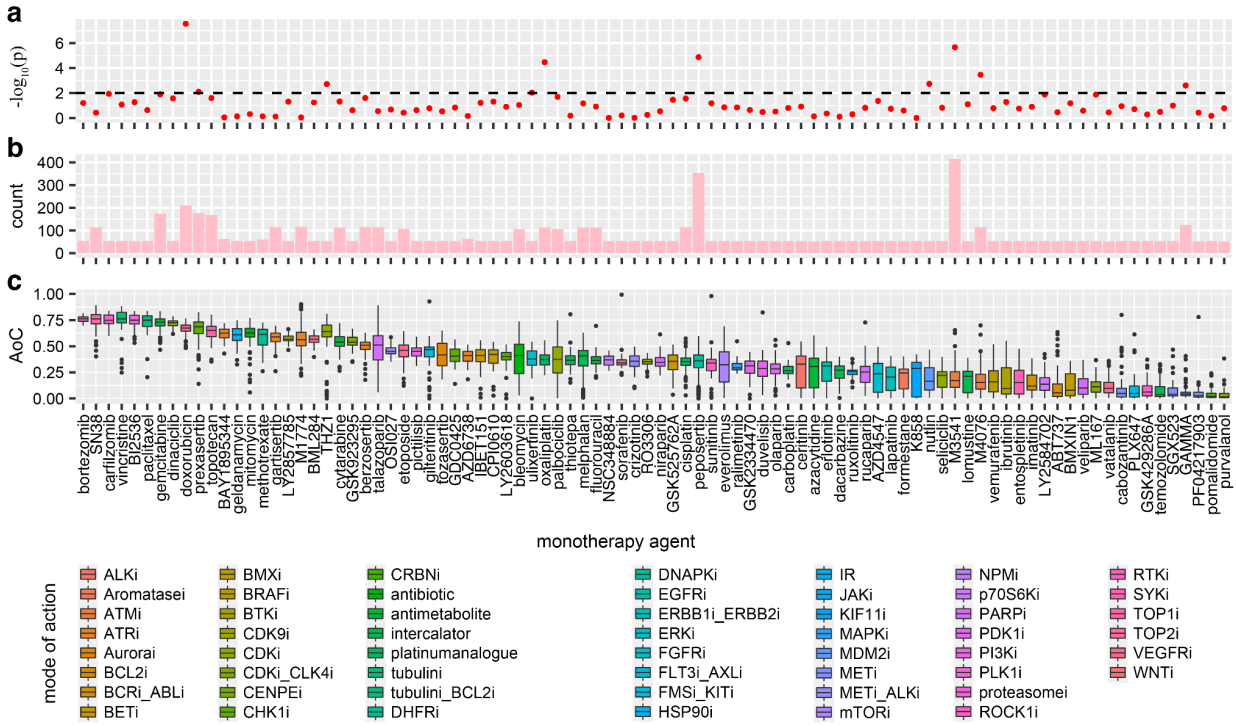
Gene1	Gene2	Alias of gene2	Average efficacy across all cell lines
ATM	PLK1**	Polo-like kinase 1 (PLK-1)	0.7185
	TUBA4A**	Tubulin Alpha 4a	0.711
	TUBB4B**	Tubulin Beta 4B Class IVb	0.711
	RRM2B**	Small subunit of p53 (191170)-inducible ribonucleotide reductase	0.652
	TOP2A*	DNA Topoisomerase II Alpha	0.6444
	TOP2B*	DNA Topoisomerase II Beta	0.6444
	TOP1**	DNA topoisomerase 1	0.6137
	RPS6KA1*	Ribosomal Protein S6 Kinase A1	0.5758
	CHEK2*	Checkpoint Kinase 2	0.5758
	HSP90AB1	Heat Shock Protein 90 Alpha Family Class B Member 1	0.5684
ATR	RRM2B**	Small subunit of p53 (191170)-inducible ribonucleotide reductase	0.7703
	TOP1**	DNA topoisomerase 1	0.6959
	RPS6KA1*	Ribosomal Protein S6 Kinase A1	0.6924
	CHEK2*	Checkpoint Kinase 2	0.6924
	RRM2	Ribonucleotide Reductase Catalytic Subunit M2	0.6915
	RRM1	Ribonucleotide Reductase Catalytic Subunit M1	0.6915
	PLK1**	Serine/threonine-protein kinase PLK1, also known as polo-like kinase 1 (PLK-1)	0.6278
	TUBB4B**	Tubulin Beta 4B Class IVb	0.6142
	TUBA4A**	Tubulin Alpha 4a	0.6142
	CHEK1	Checkpoint Kinase 1	0.602
DNA-PK (PRKDC)	TUBA4A**	Tubulin Alpha 4a	0.7634
	TUBB4B**	Tubulin Beta 4B Class IVb	0.7634
	PLK1**	Serine/threonine-protein kinase PLK1, also known as polo-like kinase 1 (PLK-1)	0.7573
	TOP2B*	DNA Topoisomerase II Beta	0.7226
	TOP2A*	DNA Topoisomerase II Alpha	0.7226
	TOP1**	DNA topoisomerase 1	0.6805
	RRM2B*	Small subunit of p53 (191170)-inducible ribonucleotide reductase	0.6667
	TUBA1A	Tubulin Alpha 1a	0.6199
	WNT11	Wnt family member 11	0.6026
	WNT16	Wnt family member 16	0.6026

Supplementary Table 4.3. The top ten directly targeted genes (among the 272 genes directly targeted by all drugs in this study) that achieved the highest synergy (Bliss score) across all cell lines in combination with the inhibition of drug targets ATM, ATR, or DNA-PK (PRKDC). **: genes that occur in the top ten in combination with all three drug targets. *: genes that occur in the top ten in combination with two out of three drug targets.

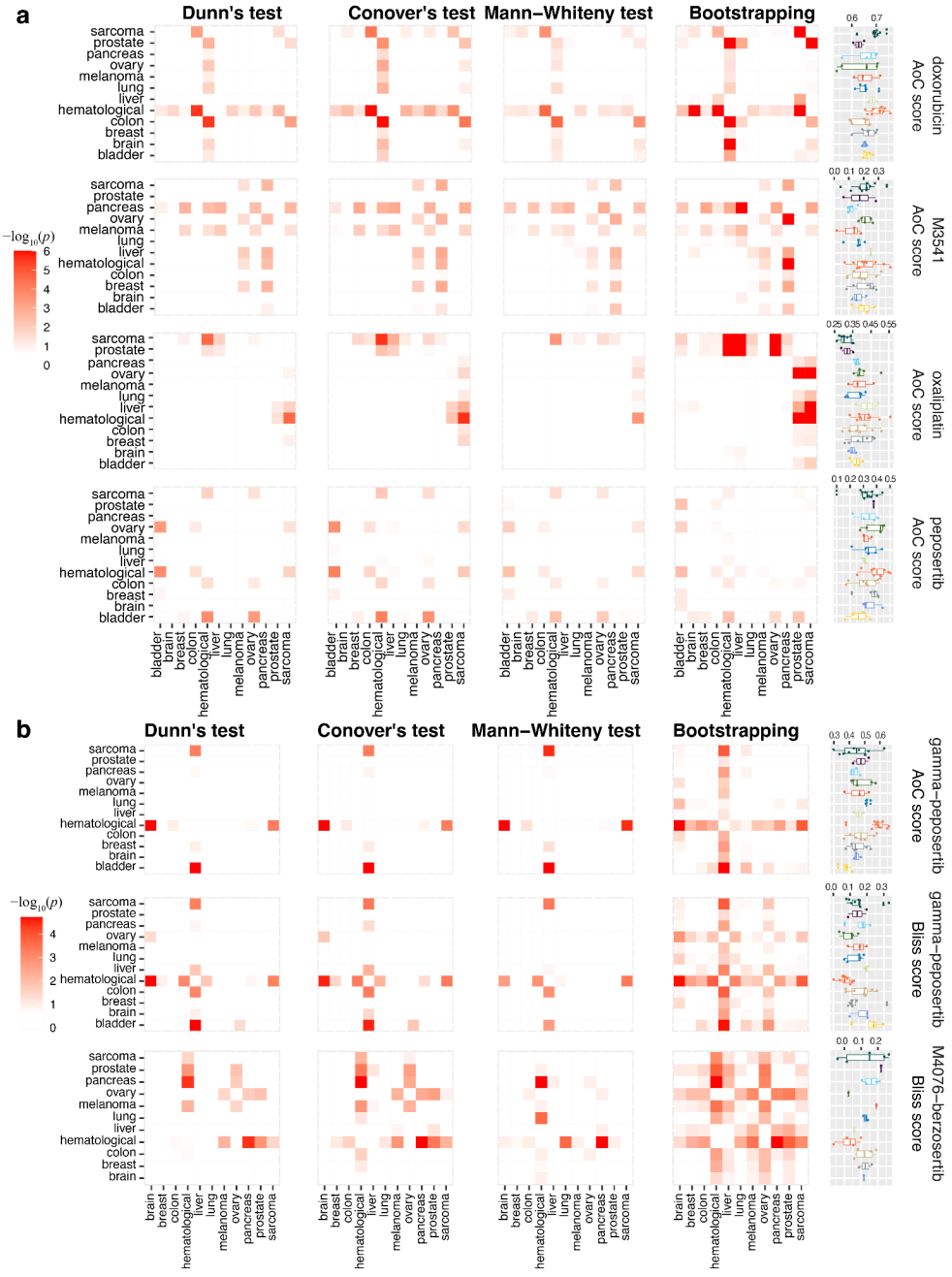
Gene1	Gene2	Alias of gene 2	Average synergy across all cell lines
ATM	PARP1**	Poly(ADP-Ribose) Polymerase 1	0.2454
	PARP3**	Poly(ADP-Ribose) Polymerase 3	0.2199
	PARP2**	Poly(ADP-Ribose) Polymerase 2	0.2127
	ATR	ATR serine/threonine kinase	0.184
	TOP1**	DNA topoisomerase 1	0.1093
	TOP2B*	DNA Topoisomerase II Beta	0.0743
	TOP2A*	DNA Topoisomerase II Alpha	0.0743
	CYP2B6*	Cytochrome P450 family 2 subfamily B member 6	0.0377
	PDPK1*	3-phosphoinositide dependent protein kinase 1	0.0296
	AURKB	Aurora kinase B	0.0216
ATR	PARP3**	Poly(ADP-Ribose) Polymerase 3	0.3165
	PARP1**	Poly(ADP-Ribose) Polymerase 1	0.3109
	CYP2B6*	Cytochrome P450 family 2 subfamily B member 6	0.2959
	PARP2**	Poly(ADP-Ribose) Polymerase 2	0.2919
	DCK	Deoxycytidine kinase	0.2237
	ATM	ATM serine/threonine kinase	0.1955
	TOP1**	DNA topoisomerase 1	0.166
	RRM1	Ribonucleotide reductase catalytic subunit M1	0.152
	RRM2	Ribonucleotide reductase catalytic subunit M2	0.152
	RRM2B	Ribonucleotide reductase regulatory TP53 inducible subunit M2B	0.1237
DNA-PK (PRKDC)	TOP1**	DNA topoisomerase 1	0.3321
	TOP2A*	DNA Topoisomerase II Alpha	0.3165
	TOP2B*	DNA Topoisomerase II Beta	0.3165
	ERBB3	Erb-b2 receptor tyrosine kinase 3	0.0412
	PDPK1*	3-phosphoinositide dependent protein kinase 1	0.037
	HSP90AB1	Heat shock protein 90 alpha family class B member 1	0.0313
	HSP90AA1	heat shock protein 90 alpha family class A member 1	0.0313
	PARP3**	Poly(ADP-Ribose) Polymerase 3	0.0231
	PARP2**	Poly(ADP-Ribose) Polymerase 2	0.0231
	PARP1**	Poly(ADP-Ribose) Polymerase 1	0.0231

Supplementary Figures

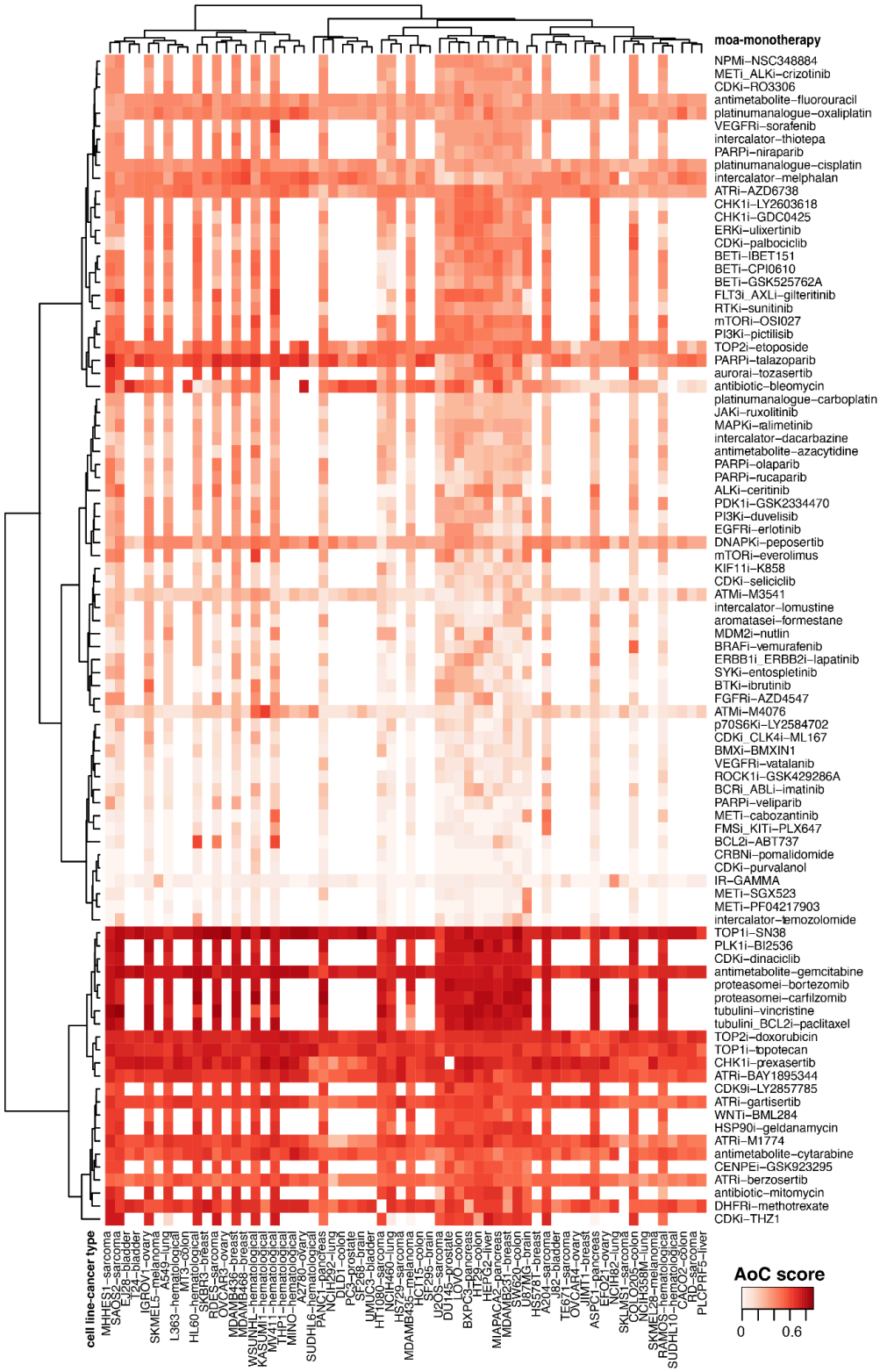
Supplementary Figure 4.1. Overview of all monotherapies used in this study. (**a**) the significance ($-\log_{10}(p)$) from the Kruscal-Wallis variance test across all cancer types for each monotherapy. A dashed line marks the significance threshold (p=0.01, two-sided). (**b**) The total count of experiments of monotherapy. (**c**) boxplot shows the efficacy of all anti-cancer drugs used in this study. The color of the boxplot indicated the mode of action. Drugs were ordered by average efficacy in all experiments in descending order.



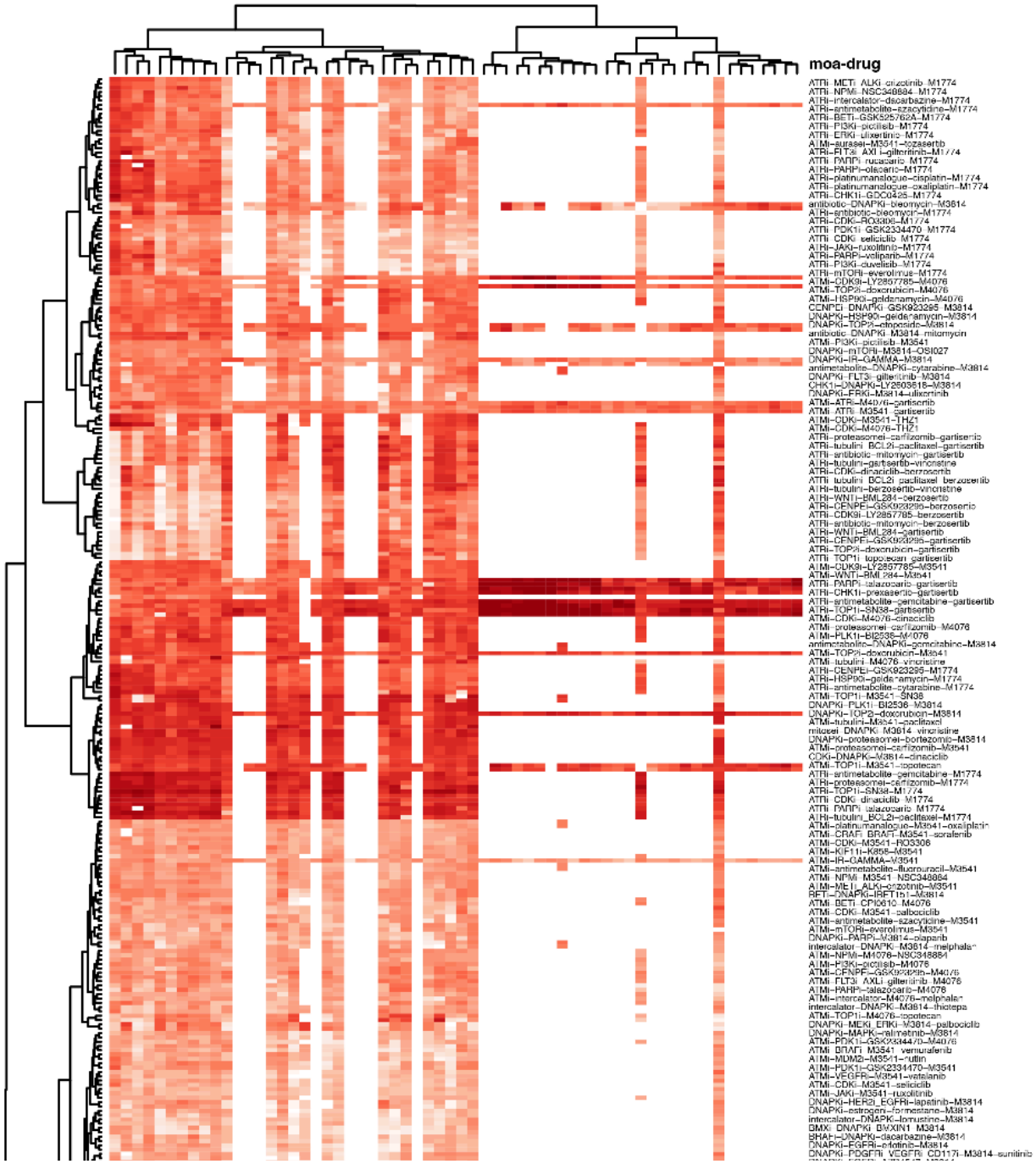
Supplementary Figure 4.2. The heatmap shows the results from post-hoc analysis on the significantly variant monotherapy and combination treatments from the Kruscal-Wallis test, and the right lane shows the distribution of response scores (AoC or Bliss) in different cancer types. (**a**) shows post-hoc analysis results of monotherapy doxorubicin, M3541, peposertib, and oxaliplatin, respectively.**, and** (**b**) shows post-hoc analysis results of combination therapy peposertib-gamma-ionizing-radiation (AoC and Bliss score) and M4076-berzosertib (Bliss score). Boxplots show the 25, 50, and 75 percentiles with whiskers extending to 1.5 times the interquartile range; for each cancer types the total numbers of cell lines are: bladder=4; brain=3; breast=6; colon=8; hematological=10; liver=2; lung=5; melanoma=3; ovary=5; pancreas=4; prostate=2; sarcoma=10. All statistically significant values from the variance test are two-sided.



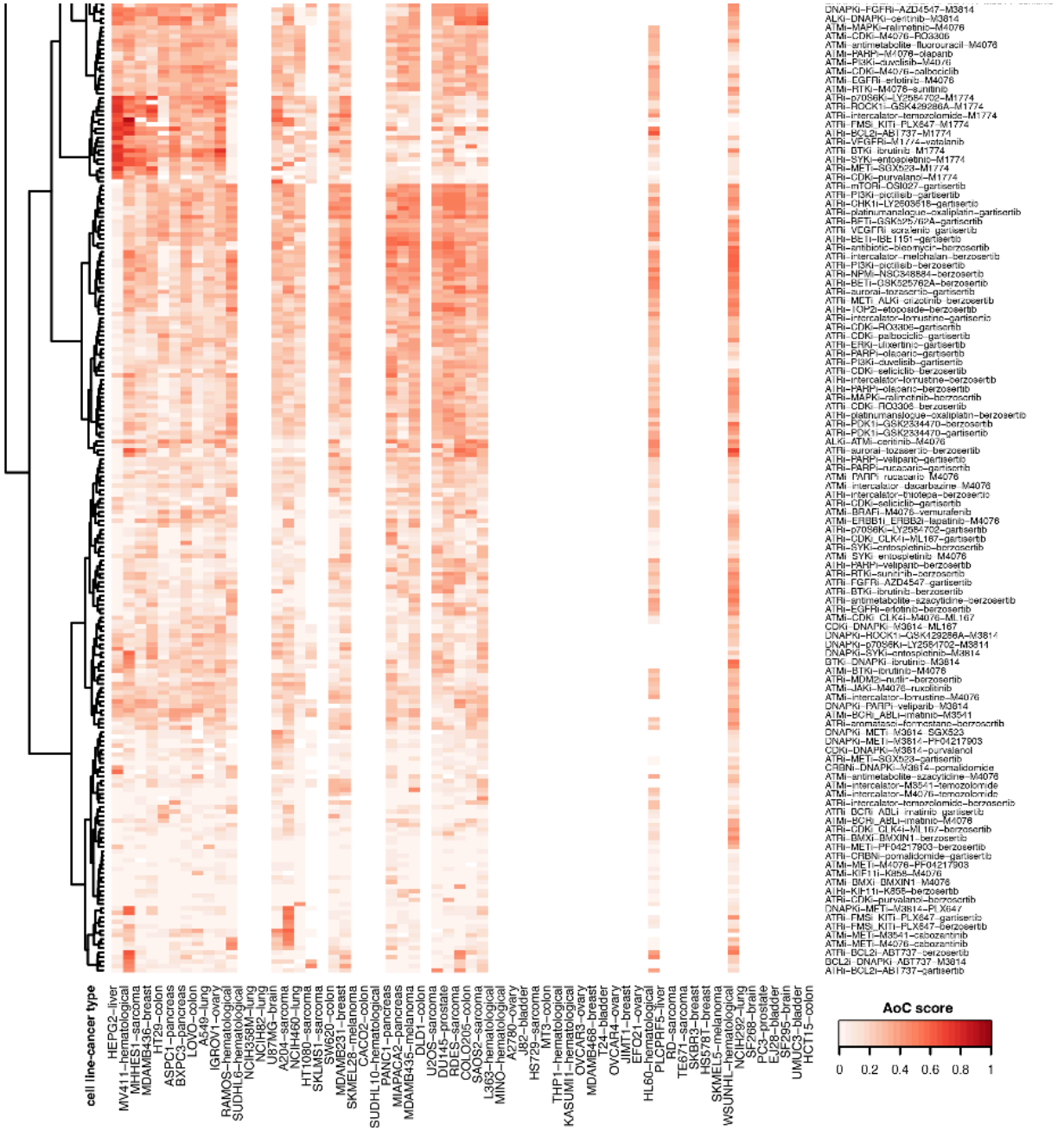
Supplementary Figure 4.3. Hierarchy clustering of monotherapy from responses (efficacy) on different cell lines.



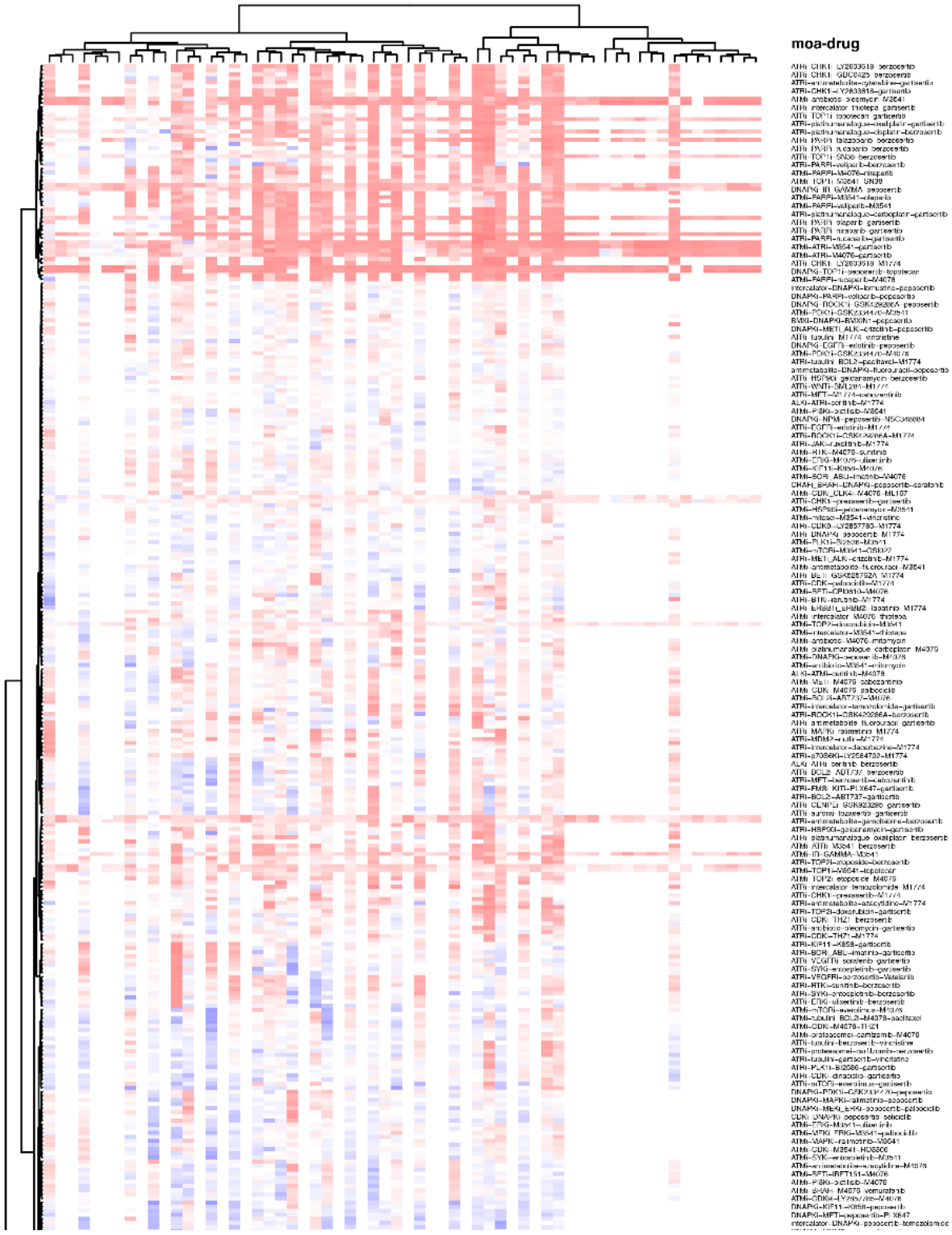
Supplementary Figure 4.4. Hierarchy clustering of combinations from responses (efficacy) on different cell lines.



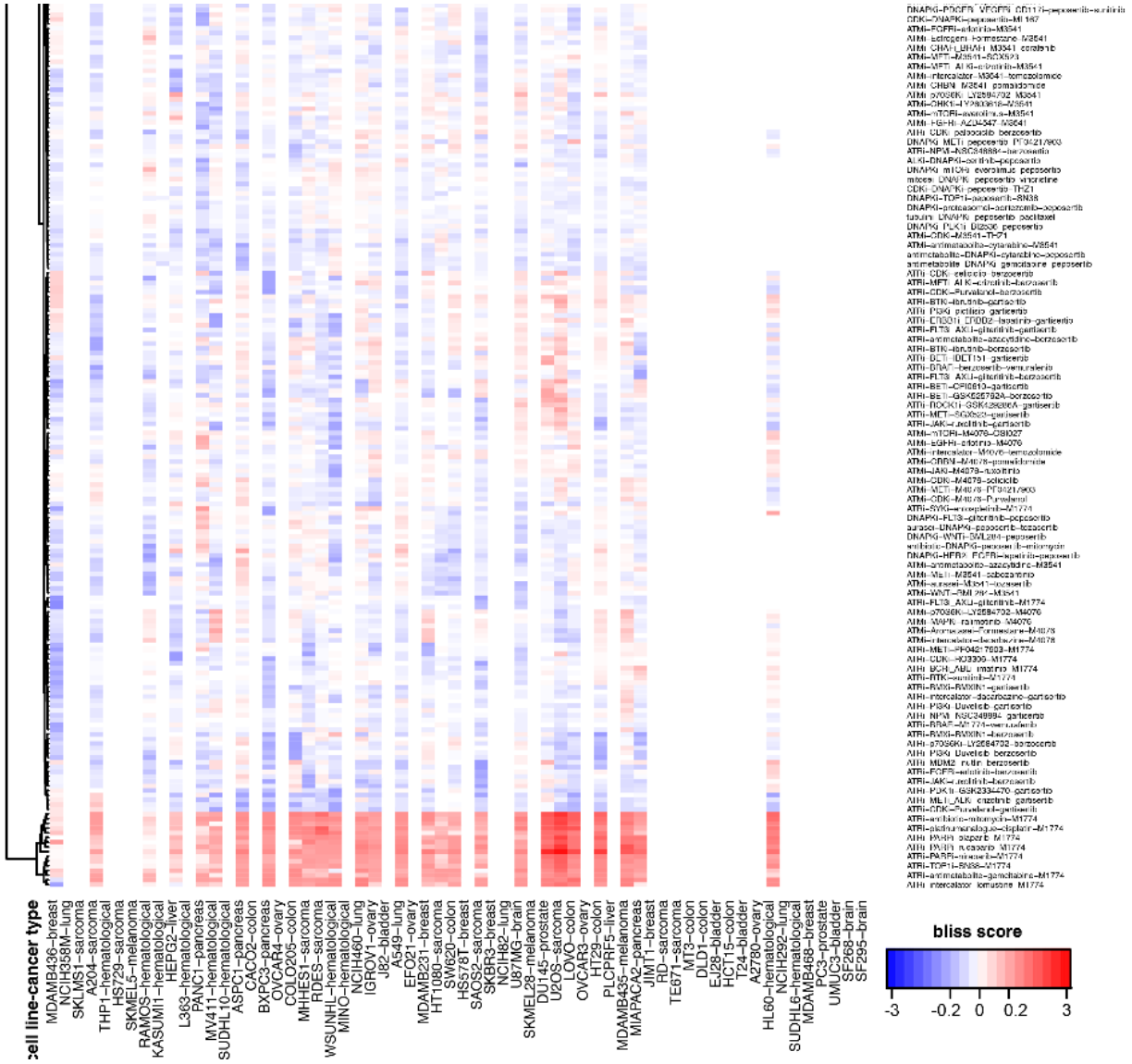
Supplementary Figure 4.5. Hierarchy clustering of combinations from responses (efficacy) on different cell lines.



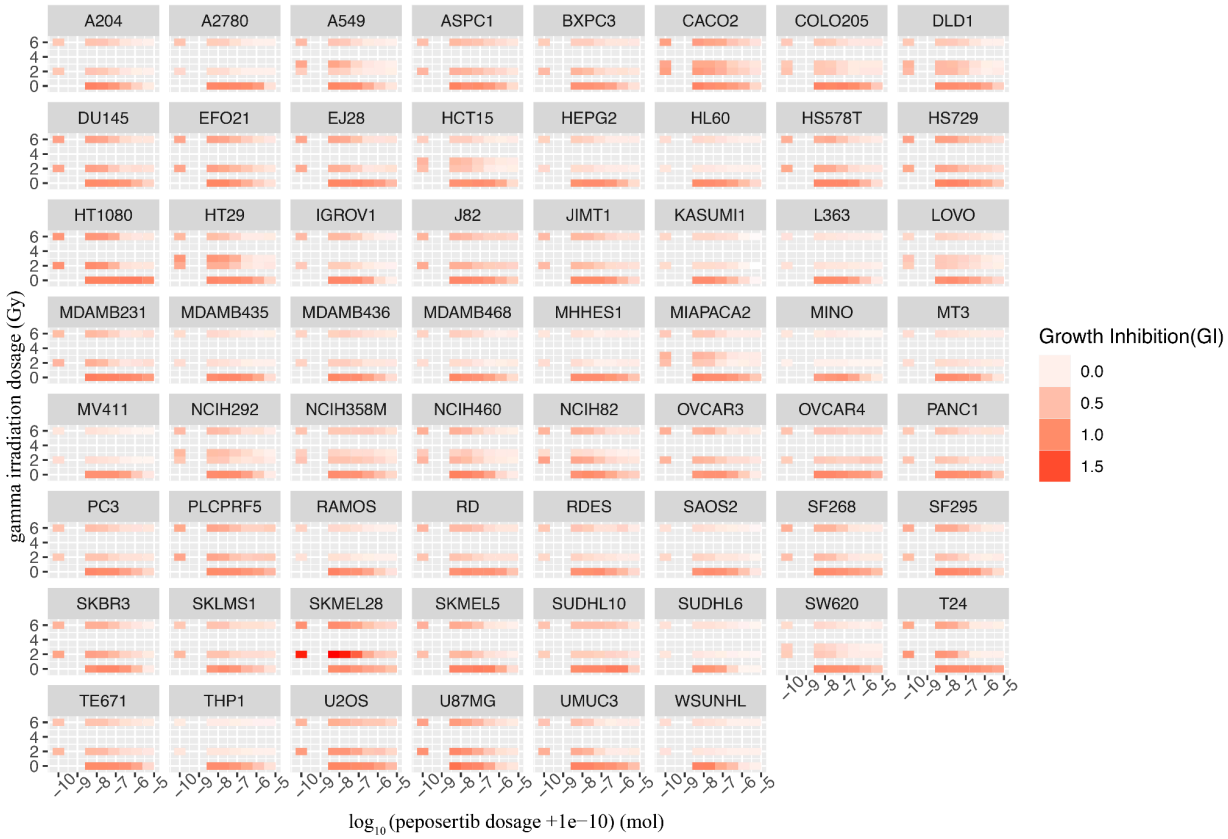
Supplementary Figure 4.6. Hierarchy clustering of combinations from responses (synergy) on different cell lines.



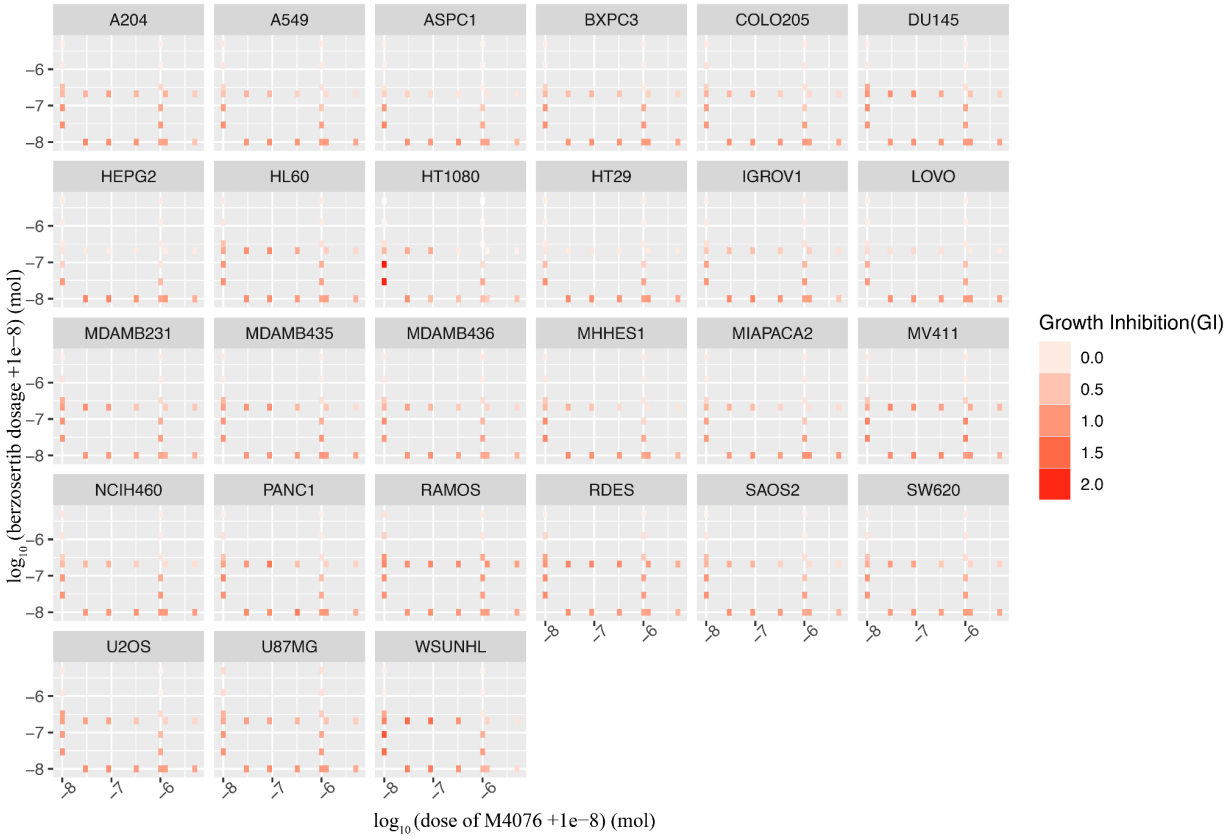
Supplementary Figure 4.7. Hierarchy clustering of combinations from responses (synergy) on different cell lines.



Supplementary Figure 4.8. Demonstration of the dose-response matrices of peposertib-gamma-ionizing-radiation combination treatment. The responses (growth inhibition rate, GI) in cell lines at different doses of peposertib (mol) and gamma ionizing-radiation (Gy) were shown by heatmaps.



Supplementary Figure 4.9. Demonstration of the dose-response matrices of M4076-berzosertib combination treatment. The responses (growth inhibition rate, GI) in cell lines at different doses of M4076 (mol) and berzosertib (mol) were shown by heatmaps.



CHAPTER V: Machine Learning Predicts and Interprets the Synergistic DNA Damage Response Combination Treatments in Variable Biological Contexts

Abstract

The combination therapy of DNA-damage sensing kinase inhibitors with other anti-cancer therapies is a promising strategy in DNA damage response (DDR) targeted clinical cancer treatment. It remains an open question to choose an optimal partner agent with the DNA-damage sensing kinase inhibitors. In this study, we employed state-of-the-art algorithms of drug response prediction on combination therapies with DDR kinase inhibitors by utilization of prior knowledge of gene clusters and synthetic lethality and simulation of post-treatment gene expression through network propagation. Based on feature importance visualization from SHAP analysis, we selected a core set of global and tissue-specific molecular markers to create a surrogate feature set, enabling us to build an optimal gene panel that is predictive for DDR targeted treatment response. This method was further validated on a hold-out dataset with cell lines and cancer tissue types that are not previously included in existing public datasets, showing improvement of treatment efficacy in 100% of cases when selecting appropriate DDR combinatorial therapy.

Introduction

DNA-damage response (DDR) pathways, often referred to as the “Achilles’ heel” of cancer, have garnered major interest as a therapeutic target in the research and development of the pharmaceutical industry in the past few years (O’Connor, 2015). A preferred strategy in clinical

anti-cancer therapy involves the combination of multiple DDR-targeted agents. This approach accounts for the complexity and interconnected nature of various parallel DDR pathways. Such treatment strategy was exemplified by previous successful applications, including the combination of PARP inhibitors with cancer patients with BRCA1/BRCA2 mutation (Bryant et al., 2005; Farmer et al., 2005). This strategy was further extended by using targeted therapy for the homologous recombination (HR) repair process to simulate "BRCAness" on BRCA-proficient cancer cells, which resensitizes them to PARP inhibitors in clinical settings (Lord & Ashworth, 2016). ATM, ATR, and DNA-PK, which are the core DDR regulators by sensing double-strand break (DSBs), and transducing the DSB signal (Blackford & Jackson, 2017), have been proposed as the target of the backbone agents for DDR-targeted combination treatments. The inhibitors of ATM, ATR, or DNA-PK have shown significant synergism in combination with other anticancer treatments, such as radiotherapy, chemotherapy, or immunotherapy in many previous preclinical studies (Barnieh et al., 2021; Brandsma et al., 2017; Weber & Ryan, 2015). Consequently, they have been included as main agents of interest in high-throughput *in vitro* combination treatment screening.

While the above experimental evidence has shown some DDR-targeted treatment combinations may be more synergistic than others (H. Zhang, Kreis, et al., 2023), there still lacks a unified approach to comprehensively analyze the biological context and select a suitable DDR combination therapy for variable treatment subjects. Machine learning becomes a handy tool to untangle the convoluted problem to choose an optimal DDR-targeted therapy for different biological contexts. Benchmarks in combination synergy prediction include international community challenges such as NCI-DREAM (Costello et al., 2014) and AstraZeneca-Sanger DREAM (Menden et al., 2019), which have been launched to call for optimal machine learning

algorithms to predict combination treatment response-based on high-throughput drug screening. A variety of machine learning algorithms, such as TAIJI (H. Li et al., 2019) and DIGRE (Yang et al., 2015), showed top performance in those competitions. Pharmaceutical properties and molecular structure of drugs, as well as the molecular characterization of the treated cell lines, such as genomic, epigenomic, and transcriptomic biomarkers, have been utilized for response prediction. The top-performing method integrated prior knowledge such as drug-target interactions and biological networks to simulate post-treatment molecular profiles and use these simulated features for machine learning models, which achieved experimental replicate level accuracy (H. Li et al., 2018).

In this study, we adapted the state-of-the-art machine learning strategy to the prediction of DDR-targeted combination therapy, based on the pan-cancer DDR-targeted combination therapy high throughput screening dataset we curated previously (H. Zhang, Kreis, et al., 2023). Our strategies included introducing prior knowledge, such as drug target information, mode of action of the drugs, geneset cluster information, and synthetic lethality gene pairs. Network propagation (H. Li et al., 2018) on gene expression profiles using tissue-specific gene-gene networks was also integrated into our model to improve the model's transferability between different cancer tissues (Greene et al., 2015; Guan et al., 2012; Wong et al., 2018). These approaches are critical to our study and distinguish tumors sensitive to particular combination treatments and for proposing putative response biomarkers in a genome-wide fashion. Our machine learning model was further validated on a hold-out high throughput screening dataset on DDR combination therapies, which comprises 24 cell lines and two tissue types that are not covered by the training set, to demonstrate the model's generalizability to unseen biological contexts. Our model improved treatment efficacy in over 100% of the cases over the baseline

treatments (the median efficacy of all treatments tested on the same cell line), thus demonstrating the value of optimizing the selection of combining DDR therapeutics. In addition, by employing global and local AI interpretation methods on the machine learning model, we identified molecular biomarkers particularly associated with co-therapeutic efficacy and synergy of general DDR combination therapy, and some particular combinations of interests, such as ATMi-ATRi, ATRi-PARPi, ATRi-TOP1i, and DNA-PKi-IR. These biomarkers opened new avenues for patient stratification by the genetic setup of a tumor or drug development efforts aimed at novel DDR targets. The AI interpretation strategy also allows us to perform feature selection for a simplified model, which showed improved performances compared to using the full set of genes and also generated a minimal gene panel for accurate response prediction. Based on the simplified model, we created a surrogate machine learning model to select optimal DDR-targeted combination treatment, using the minimal gene panel readouts from the biological backgrounds. This interface suggested the potential for future research and clinical usage in DDR-targeted therapeutics. The hold-out validation dataset published in this paper will also enhance the current scope of DDR combination therapy screening.

Results

Building machine learning models using simulated features that improve combination treatment assignment by leveraging molecular features and drug information

To demonstrate the generalizability of our machine learning model across various biological contexts, we adopted a two-step training and validation approach in this study (**Figure 5.1a**). Initially, our machine learning dataset was trained using a high-throughput DDR combination treatment screening dataset from a prior study (H. Zhang, Kreis, et al., 2023). This dataset includes 17,912 combination experiments with 87 anti-cancer drugs tested on 62 unique cancer

cell lines across 12 tissue types. We assessed model performance using k-fold cross-validation, where the training and testing data were stratified by cell line and tissue type, respectively. Subsequently, we performed additional validation on an independent dataset comprising 4,915 combination treatment experiments generated during this study. This dataset featured cell lines and tissues not included in the initial training set and utilized the same set of anti-cancer drugs (see **Figure 5.1b**). Genomic and transcriptomic analyses were conducted on cell lines from both the previous and newly generated datasets (refer to **Methods**). t-SNE analysis revealed distinct clustering patterns of cells from both datasets across various tissue types (**Figure 5.1c and Supplementary Figure 5.1**).

The most critical aspect is the simulated molecular profiles that assume drug-targeted genes turning their expression values to zero. Besides yielding information about which features (*i.e.*, the putative biomarkers) are most important for the machine learning model in prediction, this approach also provides a testing scenario for automatically assigning "optimal" (as in maximizing synergy or efficacy) treatment to cell lines. To investigate the relative importance of different kinds of molecules (*i.e.*, cancer cell-derived) and drug-related features for prediction, we trained multiple machine learning models on different groups of features as displayed in **Figure 5.1d and Supplementary Figure 5.2**. In the initial naive model, we only used the molecular characteristics for the treated cell lines, which encompassed genetic markers (such as single nucleotide variants (denoted as "*snv*"), copy number variations ("*cnv*"), loss-of-function of the gene ("*lof*"), and mRNA-based gene expression ("*exp*")), derived gene cluster-based features measuring expression ("*coh_pat*") and loss-of-function ("*lof_pat*") of cancer pathways, as well as specialized DDR-related readouts ("*ddr*") such as the cell doubling time of the cell line, homologous recombination deficiency (HRD) scores, microsatellite instable (MSI) scores, and

tumor mutational burden. In the second model, we incorporated publicly available drug target information was included to provide additional pharmacological context. In the third model, tissue-specific networks are further added upon the molecular signatures by network propagation, in a fashion we have introduced in previous studies of TAIJI and AstraZeneca DREAM Challenge, by tuning down the expression level of genes that are not directly targeted by the drugs by its proximity with the targeted genes in the network. Synthetic lethality information was further added on, by calculating the combined expression level of each synthetic lethal gene pair on the post-treatment expression profiles we have generated earlier. After that, the drug names, chemical structure of the drugs, the monotherapy efficacy responses for each drug in a given combination, the gene set annotations, and the mode-of-actions for both drugs are included to form our final model.

We implemented the above features in the following manner:

a) A total of 10,462 molecular biomarkers, including genomic (denoted as "*snv*", "*cnv*", "*lof*" and "*lof_pat*"), transcriptomic ("*exp*" and "*coh_pat*") and DNA damage response readouts ("*ddr*") were obtained as molecular features (see **Methods**). The expression level of individual genes was first quantile normalized before being used as features. Besides, the target gene information was added to the molecular information to provide additional information on the drug mechanism in the cellular context. For example, for a combination treatment drug A-drug B in cell line C, we have the expression level readouts of the 2,725 genes potentially involved in DDR pathways, and m genes that are functionally perturbed by these two drugs. We alter the expression levels by setting the perturbed genes to zero, as follows:

$$exp_{gene_i} = 0 \text{ (for } gene_i \text{ in all perturbed } gene_i \ldots gene_n) \ldots\ldots \text{Eq. (1)}$$

This operation simulates the target gene's activation/inactivation by the administered drugs inside each cell and provides more variable information for context-based combination treatment response prediction.

b) Tissue-specific network information was added upon the expression profiles by network propagation, a method we have introduced in our previous studies (H. Li et al., 2018, 2019) and the winning solution in AstraZeneca DREAM challenge (Menden et al., 2019). Simplicity, we alter the expression level of genes that are not directly perturbed by the drugs by the interaction probability of this gene to the target genes. For example, for $gene_j$, its closeness to all *n* perturbed genes in the network is:

$$\{p_{1j},\ p_{2j},\ p_{3j},\ \ldots..\ p_{nj}\} \ldots\ldots \text{Eq. (2)}$$

The altered expression levels of $gene_j$ is:

$$\widehat{exp}_{gene_j} = exp_{gene_j} \times (1 - max\{p_{1j},\ p_{2j},\ p_{3j},\ \ldots..\ p_{nj}\}) \ldots\ldots \text{Eq. (3)}$$

c) Synthetic lethality information is integrated to the altered gene expression profile (which is the simulation of post-treatment expression). The definition of synthetic lethality is, only the co-occurrence of down-regulation of both gene *a* and gene *b* will lead to cell death (Kaelin, 2005). Therefore, we constructed the synthetic lethal feature for a synthetic lethal gene pair *a* and *b* as:

$$synleth_{gene_a,\ gene_b} = max(exp_{gene_a}, exp_{gene_b}) \ldots\ldots \text{Eq. (4)}$$

Where the expression levels of gene *a* and *b* were also normalized by the deduction of the mean of this gene in all different cell lines and divided by standard deviation. We also implemented the maximum, mean, and minimum of all synthetic lethality features in the feature set.

d) Drug names and mode-of-actions were converted to indexes and implemented as one-hot encodings. To note, the order of the features of the two drugs is randomly exchanged during training since we want to avoid information bias due to the order of the drugs.

e) The SMILE format of the chemical structure of both drugs was translated into bit strings using six types of fingerprints, including MACCS, Morgan, RDK, FP2, FP3, and FP4, using rdkit, pubchempy and openbabel Python API. This will generate binary strings (e.g., 101000111) that embed the topological, substructural, and chemical characteristics of the drugs.

f) Monotherapy responses of the two drugs on the same cell line in the combinations were regarded as two independent features and added to the feature set respectively. For example, in combination treatment M1774-berzosertib on cell line HEPG2, the monotherapy efficacy of M1774 and berzosertib on HEPG2, respectively, are used as input in the model in the feature set. To note, we have obtained 7,326 single-drug tests in total (compared to $87\ (drug)\times 62\ (cell\ line) = 5,394$ different types of monotherapy tests), and some of the tests have been replicated multiple times for monotherapy. For these tests, we take the average value of all replicates as the experimental monotherapy responses.

g) Gene set annotations of the target genes were also implemented as independent features. We extracted 18,681 gene set annotations covering 9,450 genes in the human genome from MSigDB (Subramanian et al., 2005) and literature. Each gene set could contain from one to up to 15 genes. We incorporated the gene set annotation information by counting how many genes of that gene set were targeted by the drugs used in the combination treatments.

The final machine learning model was built by stepwise addition of each of the above feature sets into LightGBM learners. Synergy and efficacy predictions were evaluated against the experimental gold standard. In cross-validation, we partitioned the training data either by cell

lines or, to be more challenging for the machine learning model, by tissue type (cancer indications), *i.e.*, each cell line or, alternatively, each tissue type was either present in the training or the testing partition during training, but not in both **(Figure 5.1a**). This approach has been demonstrated previously to excel in integrating diverse feature types (Spiro et al., 2019; Y. Zhang et al., 2019; Y. Zhu et al., 2020).

For cross-validation between cell lines, molecular profiles alone do not show any signals in predicting efficacy and synergy responses (Pearson's correlation of -0.0162[-0.032, 0.0074] (mean[95%CI]) and 0.0145[-0.0027, 0.0316] (mean[95%CI])). However, integrating the target gene information of both treatments in the combination improved the performances significantly to Pearson's correlation of 0.6087[0.6018, 0.6206] and 0.4095[0.3905, 0.4425] in predicting efficacy and synergy, reflecting the importance of introducing of treatment-specific data. Integration of tissue-specific networks further improved the performances to 0.6141[0.6067, 0.6264] and 0.433[0.4171,0.4647], for efficacy and synergy, respectively. The same pattern of improvement by integrating target information and network propagation is also observed by comparing to the synthetic lethality baseline model (**Supplementary Figures 5.3 and 5.4**). Adding synthetic lethality does not further improve the efficacy prediction, but improves the synergy prediction to 0.4791[0.4621, 0.5083]. Adding treatment-specific features, including drug name one-hot encoding, chemical structure fingerprints, monotherapy responses, and mode-of-actions, further improved the model performances step by step. The best model that was thus achievable produced 0.7834[0.7766, 0.79] Pearson's correlation for AoC score and 0.679[0.6634, 0.6995] Pearson's correlation for Bliss score, including all features mentioned above (**Figure 5.2a, left panel**). The same trend is also observed in the validation of the above models on the hold-out dataset with new cell lines from the same tissue types as the training set,

achieving 0.7522[0.748, 0.756] Pearson's correlation for AoC score and 0.7709 [0.7643, 0.7774] Pearson's correlation for Bliss score (**Figure 5.2b, left panel**).

We next took on a more challenging setup, *i.e.*, examining the cross-validation model performance across tissue types. We found similar contributions from each type of feature (**Figure 5.2a, right panel**). The best model achieved 0.7765 [0.7697, 0.785] Pearson's correlation for AoC score and 0.7028[0.6901, 0.7156] Pearson's correlation for Bliss score, values practically identical to the cell-line cross-validation scenario. When validating the above models on the two new tissues in the external validation sets, the features exhibit the same additivity trend, and achieved 0.6611[0.641, 0.6818] Pearson's correlation in AoC score and 0.7822[0.7627, 0.8064] Pearson's correlation in Bliss score on cervix tissue, and 0.5454[0.5271, 0.5739] Pearson's correlation and 0.8429[0.8278, 0.8585] Pearson's correlation in Bliss score on kidney tissue, respectively (**Figure 5.2b, middle and right panel**).

We also assessed the cross-tissue type model's validation performance across 14 different tissue types (**Figure 5.2c**). The model demonstrated superior performance in both efficacy and synergy predictions for prostate, melanoma, and brain tissues compared to the overall baseline performance across all tissues. In contrast, liver tissues showed lower performance in these metrics, indicating limited generalizability of the model to this tissue type. Bladder, cervix, and kidney tissues exhibited lower generalizability in efficacy, yet outperformed the baseline in synergy. Interestingly, for breast cancer, while the model was more effective in predicting efficacy, its performance in predicting synergy was notably weaker.

To demonstrate the ability of the machine learning model to optimize combination treatments, we compared, for each cell line, the most effective DDR combination therapy selected by the machine learning model by the baseline administration (the combination achieved

the median efficacy in all combinations tested). We noticed in all of cell lines, the machine learning model could select combinations with improved treatment efficacy compared to the baseline treatment, with an improvement of efficacy from 0% to 280% (**Figure 5.2d).**

Identifying global determinants for combination treatment response to DNA damage response kinase inhibitors

To gain deeper insight into the machine learning models and prioritize molecular features for biomarker and drug target development, we utilized the improved SHapley Additive exPlanations (SHAP) analysis (S. M. Lundberg & Lee, 2017) to identify the most important predictive features for combination treatment's efficacy and synergy. Being broadly used in machine learning, SHAP values estimate the contribution of each feature to the final prediction given the values of all other features. We carried out the SHAP analysis using the best performing model in cross-cell line validation (**Figure 5.3, Supplementary Figure 5.5-13**), within each mode-of-action combination, and for each type of molecular feature, and ranked SHAP values by relative importance to gain insight into the most predictive features (higher SHAP values indicate higher predictiveness of features). Based on this analysis, monotherapy responses (AoC score) of both treatments in the combination on the same cell lines, and the enrichment of pathways of the target genes (geneset annotation), proved to be the most important predictors for efficacy and synergy, respectively (**Figure 5.3a and f**). The next important predictors were chemical structure for both efficacy and synergy. Geneset annotations ranked third in efficacy prediction, followed by molecular biomarkers and synthetic lethality of genes on the cell lines. For synergy prediction, these two kinds of information also played important roles, while monotherapy responses are considered much less important than efficacy prediction.

Among the molecular features, gene expression levels are more informative for predicting combination response in terms of both efficacy and synergy scores compared to gene loss-of-function mutations and gene copy number alterations, as the target gene information and tissue-specific networks are incorporated into the expression profile to simulate the post-treatment effect (**Figure 5.3b and g**). The fact that genes in core DDR pathways are relatively rarely mutated compared to known cancer drivers is the likely reason why gene loss-of-function mutations and copy number readouts are not as predictive as expression features, which encode the activity of DDR pathways and are thus more usable by the machine learning model (Knijnenburg et al., 2018). Highly mutated outliers such as TP53 exist, of course, but seem not to be predictive for most DDR-targeted combination treatments we investigated in our data. This result is also consistent with the experiments we carried out earlier in this study: when features are progressively added into the model, expression data incorporating target gene and network information outperformed all other types of molecular features (**Supplementary Figure 5.4**).

SHAP values are additive, *i.e.*, the combined contribution of a set of features is simply the sum of all individual SHAP values (Shapley, 1983). This property allowed us to explore the per-gene contribution by summarizing the contributions from expression, loss-of-function, single nucleotide variation, and copy number variation features relating to the same gene. The top genes contributing most to efficacy and synergy prediction are shown in **Figures 5.3d and i (Supplementary Figures 5.7 and 5.8)**. At the same time, the top synthetic lethality gene pairs that are contributing most to efficacy and synergy prediction are shown in **Figures 5.3e and j (Supplementary Figure 5.9).** As expected, when ranking across all combination treatments, the drug targets of the key DNA damage response kinase inhibitors screened in this study, ATM and

ATR and PRKDC (DNA-PK) received high SHAP values for both efficacy and synergy prediction. In the same global view, the PARP family, including PARP1 and PARP2 were also among the top ten genes and synthetic lethal gene pairs relevant for synergy prediction, indicating the broad importance of the molecular status, predominantly on the level of mRNA expression, of these genes across multiple combination treatments.

Next, we analyzed the SHAP values of molecular features for the three major modes of action investigated in this study, ATMi, ATRi, and DNA-PKi, across all combination partners. First, we extracted the most important genes by SHAP values for ATM/ATR/DNA-PK inhibitor combination treatments and combined the results into a global inhibitor-gene interaction network (**Figure 5.3k and l**).

To internally validate the SHAP approach, we also compared the top genes and gene pairs in SHAP analysis with directly targeted genes in the experimental screen that showed the highest efficacy and synergy (H. Zhang, Kreis, et al., 2023). As expected, for both efficacy and synergy responses, top-ranked drug targets in combination with ATM/ATR/DNA-PK inhibitors based on experimental screening, such as TOP1 and PARP1, also appeared among the top-ranked genes of the SHAP-based interaction network, as did the combination drug targets ATR, ATM, and DNA-PK, which were also identified as important biomarkers in our analysis (**Figure 5.3d-l)**. Moreover, as the machine learning model also employs fine-grained molecular biomarker information from the cellular context such as the expression level of DDR-related genes, we observed, more interestingly, molecular features that were not used as drug targets in the experimental screen but that nevertheless appear to be as important as the direct target genes in almost all combination treatments.

We further searched the biomarkers from functional reports and summarized the biological pathways that may be involved in regulating the DDR process to influence the treatment responses (**Figure 5.4**). While the minority of the prioritized genes, such as GUSB and COL5A1, are involved in general housekeeping (Iyer et al., 2017; S. Lee & Greenspan, 1995), all other identified factors have clear relations to DNA damage response: HSP90AA1 (HSPC1, Heat shock protein HSP 90-alpha, a stabilizer of CHK1, MSH2, and XRCC1 (Sottile & Nadin, 2018)), PKIB (cAMP-dependent protein kinase inhibitor beta, a potent competitive inhibitor of cAMP-dependent protein kinase (PKA) activity), SESN1 (a target of TP53 regulation upregulated in DNA damage (M. Wang et al., 2017)), and RBBP8 (a key modulator between ATM and ATR (S. Li et al., 2000) physically associated with CTBP and BRCA1 that is involved in cell proliferation (Yu et al., 2020)) are all broadly associated with combination efficacy (**Figure 5.3d and Supplementary Figure 5.7)**. Similarly, TEAD1 (involved in YAP/Hippo regulation and thus proliferation, anti-apoptosis, and epithelial-to-mesenchymal transition (Huh et al., 2019)), ZMYND8 (related to DNA repair activities at DNA double-strand breaks (Gong & Miller, 2018)), and YWHAZ (a DNA-PK substrate and involved in several aspects of cell cycle control and associated with TP53 (Anisenko et al., 2020; S. Liu et al., 2016)) were top hits for synergy prediction (**Figure 5.3i and Supplementary Figure 5.8)**. Other genes of interest associated with response are XRCC6, which directly binds to the DNA-PK complex in the NHEJ process in DSB repair (H. Liu et al., 2010; Roberts et al., 2010; West et al., 1998), and a suite of other regulators of different aspects of the DNA damage response, such as PEA15 (an ATM substrate (Nagarajan et al., 2014)), as well as CASP3 and BCL (both central regulators of apoptosis but also involved in regulation of Fanconi Anemia pathway (CASP3 (Sakai & Sugasawa, 2014)), and suppressor of DSB repair (BCL (Q. Wang et al., 2008))).

Identifying molecular determinants of efficacy and synergy in clinically relevant combination treatments

To conduct focused investigations of specific mode-of-action combinations that are of particular clinical or pharmacological interest at the moment, such as the combination treatments ATMi-ATRi, ATRi-PARPi, ATRi-TOPi, ATRi-Cytostatic Antimetabolites (including Gemcitabine), and DNA-PKi-IR (irradiation), we extracted genes with highest SHAP values separately for each of these mode-of-action combinations (**Supplementary Figure 5.14-23**). Most salient in these analyses was the presence of LIG4, ADORA2A and PKIB gene expression status as being highly predictive for combination efficacy in the top 10 ranked features (among 2,725 ranked genes) of all five highlighted combinations. The role of ATR in these combinations was expected due to it being a drug target in these combinations and its general importance as a DDR factor, and LIG4 was also previously highlighted as a biomarker for DNA damage response targeted treatment by playing a role in DSB repairing process (Buchbinder et al., 2018; Felgentreff et al., 2016), ADORA2A and PKIB were not yet highlighted as a potential biomarker in these settings. Previous studies indicated lower ADORA2A expression levels occurring in cancer patients with DNA damage repair deficiency (Chang et al., 2022). We tentatively hypothesize that PKIB's function as a competitive inhibitor of PKA activity, and thus of PKA phosphorylation of ATR which actively recruits the key NER protein XPA (C. H. Lee et al., 2001), may relate to its recurring importance in our analysis results.

While PARP1 expression status was highly predictive for synergy across all five combination treatments, this is less surprising since synergistic relationships between PARP1, ATM, and ATR (Lloyd et al., 2020) as well as between PARP1 and DNA-PK (C. Wang et al., 2020) had been shown before.

Additional, highly ranked predictive factors of combination synergy in ATRi-ATMi, ATRi-TOP1i, as well as ATRi-Cytostatic antimetabolites, were the expression status of DNA-PK, CEP76 (which inhibits centriole amplification after DNA damages detected by PLK1) and MYD88 (which induces inflammatory genes as a result of single-stranded DNA detection by TLR9 (Nakad & Schumacher, 2016)), which were predictive for combination synergy in the ATRi-ATMi combinations, as well as YWHAZ, which was predictive for ATRi-Cytostatic Antimetabolites synergy. Lastly, SESN1 expression was found to be predictive of combination efficacy. Interestingly, none of the aforementioned genes are strongly correlated with ATRi monotherapy **(Supplementary Table 5.1 and 5.2)**, so these findings are specific to ATRi *combination* treatments.

Pruning the full machine learning model to a portable, highly accurate surrogate model

Based on the molecular determinants identified above (**Supplementary Figure 5.7 and 5.8**), we further constructed a simpler surrogate machine learning model that only includes the top biomarkers we identified from model interpretations of the full machine learning model (**Figure 5.5 a and b)**. We noticed that when only using the top 40 genes with highest SHAP values, the resulting surrogate model runs significantly faster, requires fewer input features, and even slightly outperforms the full model (0.7834[0.7766, 0.79], mean[95% confidence interval]) by a small margin (0.8035[0.7973, 0.8100], mean [95% confidence interval] Pearson's correlation for AoC score prediction, 0.679[0.6634, 0.6995] to 0.7853[0.7701,0.8062] Pearson's correlation for Bliss score prediction). Since the surrogate model displayed better runtime performance and fewer input data requirements while also demonstrating nearly identical accuracy, we decided to publish the surrogate model as part of a public R Shiny app, SynDDR, to allow researchers to predict drug combination efficacy and synergy on their own data using our approach (**Figure**

5.5c) (https://github.com/GuanLab/DDR-drug-synergy-prediction-Shiny). Defining gene expression signatures for prioritizing clinically relevant combination treatments.

Since accurate but complex machine learning models such as those mentioned here may be less practical for everyday use by clinicians, we aimed to derive succinct gene expression signatures according to methods presented by Staub et al. (Staub, 2012) that are both robust and predictive of response. Once derived, these gene signatures can be easily applied to new datasets without major computational knowledge to rank predicted drug combination efficacy or synergy. For the previously highlighted mode-of-action combinations ATMi-ATRi, ATRi-PARPi, ATRi-TOPi, ATRi-Cytostatic Antimetabolites (including Gemcitabine), and DNA-PKi-IR (irradiation), we identified three robust gene expression signatures with high coherence scores (see **Methods**). Among these, the signature for the synergy of ATRi-Cytostatic Antimetabolites (YWHAZ, ANLN, PPP4R1) and ATRi-TOP1i (YWHAZ, PPP4R1, CD9, HUS2, ANLN, TEAD1, XRCC6) treatment combinations achieved 0.28[0.0292, 0.4926] ([95% confidence interval]) and 0.55[0.3471, 0.7026] ([95% confidence interval]) Pearson's correlation. The signature score for the efficacy of DNA-PKi-IR (PEA15, PTPN14, COL5A1) achieved -0.57[-0.7170, -0.3723] Pearson's correlation. For the remaining mode-of-action combinations and readouts, no gene expression signature with high coherence could be identified. We note that further evaluation of these signatures on independent datasets may be required, a task that is outside of the scope of this work.

Discussion

DNA damage response has attracted a lot of research focus as recently both academia and industry have turned their focus on developing new DDR-targeted therapies. First-line DDR targeted agents have been developed, including a series of PARP, ATM, ATR, and DNA-PK

inhibitors (Keung et al., 2019; Mohiuddin & Kang, 2019; Nam et al., 2019; Vecchio & Frosina, 2016). The applications of these new DDR-targeted therapies need to be urgently evaluated, in terms of combination use with other drugs, applicable indications, and biological context. In this study, we present a comprehensive analysis of DDR-targeted combination treatment, by incorporating a novel feature construction strategy into state-of-the-art machine learning models and generating satisfying results. Our study shows the strong potential of machine models to be applied to DDR pathway-targeted clinical treatment strategies. With a selection of a core gene panel (40 genes) as input, our model still maintains highly accurate performances compared to all genes.

Meanwhile, we identified molecular features that are predictive for the synergy and efficacy of DDR-targeted combination therapy both globally (i.e., across all combination partners of ATRi, ATMi, and DNA-PKi) and specific to particular combination treatments that are of high interest to drug development. We took particular care to control the overfitting of our machine learning approach and identified features that are highly predictive for estimating the efficacy and synergy of the screened combination treatments. Among the more than 10,000 cancer-related features investigated, DDR-specific genes were highly enriched in the top hits **(Figures 5.3 and 5.4)**. In particular, we note that in the global analysis, nearly all top predictive molecular features were located in clearly interpretable cellular pathways of DNA damage repair or DNA synthesis, with additional hits in apoptosis, cell survival, and proliferation that also have known molecular relations to DDR. In addition to the direct drug targets ATR, ATM, and DNA-PK (PRKDC) whose mRNA expression states are particularly predictive biomarkers of treatment efficacy and synergy, PARP1 and PKIB seem to be of particular interest since they also frequently appear as important biomarkers in particular combination treatments that are of high clinical interest such

as ATMi-ATRi, ATRi-PARPi, ATRi-TOPi ("Therapeutic Targeting of ATR Yields Durable Regressions in Small Cell Lung Cancers with High Replication Stress," 2021), ATRi-Cytostatic Antimetabolites (including Gemcitabine (Konstantinopoulos, Cheng, Wahner Hendrickson, Penson, Schumer, Austin Doyle, et al., 2020)), and DNA-PKi-IR (irradiation). To facilitate applications of our results by researchers with varying levels of computational expertise, we have derived both a performant and accurate R Shiny app for conducting synergy and accuracy predictions on new data, as well as succinct gene expression signatures for a subset of mode-of-action combinations.

Methods

High throughput screening of DDR combination treatment dataset for training and hold-out dataset

The DDR combination treatment training dataset used in this study was obtained from Open Science Framework (OSF): https://osf.io/8hbsx/, which consists of 17,912 combination treatment tests performed on 62 cell lines from 12 different tissue types. For the hold-out validation dataset, high throughput screening of DDR combination treatment was carried out using the same method as the previous study (H. Zhang, Kreis, et al., 2023). All cell lines used in this study were purchased from ATCC, NCI, CLS GmbH, and Leibniz-Institute DSMZ–German Collection and the dose-response experiments were performed at Oncolead GmbH & Co. KG (Karlsfeld, Germany), resulting in 4,915 combination treatment experiments on 24 cancer cell lines from 14 different tissue types, which are not encompassed by the training set. The relative Area over Curve (AoC) and Bliss score were computed from the fitted dose-response curve to measure the efficacy and synergy of the DDR combination treatment, respectively, following the same fashion we reported earlier (H. Zhang, Kreis, et al., 2023).The experimental reproducibility

of is measured by Pearson's correlation coefficient between replicates, which is 0.8429 ($p < 1e-22$) for monotherapy efficacy score (AoC) and 0.7419 ($p < 1e-22$) for combination synergy score (Bliss), and comparable to previously published datasets (H. Zhang, Kreis, et al., 2023).

Characterization of molecular readouts on all cell lines

For short nucleotide (SNV) and copy number variation (CNV) calling, the qualified genomic DNA of the cell line samples wee fragmented by an ultrasonicator (Covaris). By adjusting shearing parameters, DNA fragments were concentrated in 500bp peaks for each sample. These fragments were purified, end blunted, 'A' tailed, and adaptor ligated. DNA templates with adapters were then selectively enriched using PCR in order to obtain a sufficient amount for the DNA library. The concentration of the libraries was quantified by a bioanalyser (Agilent Technologies) and real-time PCR method. Each qualified DNA library was sequenced on the Illumina HiSeq platform using paired-end reads according to the Illumina manufacturer's instructions. Sequencing-derived raw image files were processed by Illumina base calling Software for base-calling with default parameters and the sequence data of each cell line was generated using an Illumina HiSeq 2000 instrument in paired-end mode at 2 × 100 bp read length.

Subsequently, short nucleotide variations and copy number variations were computed using VarDict (Lai et al., 2016) and CNVkit (Lai et al., 2016; Talevich et al., 2016), respectively, in the bcbio workflow system (Chapman et al., 2020) using default parameters against the human reference genome hg19 with Ensembl 75 gene annotations. Variant calling by VarDict was conducted by requiring a minor allele frequency of at least 10% and minimal support for four de-deduplicated reads for each variant call; in addition, calibrated filters for strand bias, mean

position of variant in read, minimum mean base quality, NM/MQ mapping qualities, and DP/QUAL variant qualities were applied as per the bcbio default configuration.

For identifying functionally relevant mutations and indels, only SNVs with Variant Impact Predictor (VEP) (McLaren et al., 2016) assessment of "HIGH" or variants that were deemed to be at least likely pathogenic in SIFT, Polyphen, or Clinvar were retained. In addition, variants with at least 1% prevalence in normal populations according to gnomAD (Karczewski et al., 2020) were excluded. Note that for reasons of confidence and coverage, SNV analyzed in this study are either homozygous or heterozygous (i.e., they affect at least one allele).

For the "*snv*" features used in downstream analyses, filtered SNVs were summarized on the gene level by coding "1" for genes with at least one detected SNV and "0" for genes without any detected variants. Integer allele calls from CNVkit were used directly as "*cnv*" features. For the loss-of-function ("*lof*") features, loss-of-function events were also summarized on the gene level by coding "1" for genes with at least one detected SNV or an integer copy number call <2, while coding "0" for all genes without any such variants.

For gene expression analysis, after total RNA extraction and DNase I treatment, magnetic beads with Oligo (dT) were used to isolate mRNA (for eukaryotes) or by removing rRNAs from the total RNA (for prokaryotes). Mixed with the fragmentation buffer, the mRNA was fragmented into short fragments. Then cDNA was synthesized using the mRNA fragments as templates. Short fragments were purified and resolved with EB buffer for end reparation and single nucleotide A (adenine) addition. After that, the short fragments were connected with adapters. After agarose gel electrophoresis, the suitable fragments were selected for PCR amplification as templates. During the QC steps, Agilent 2100 Bioanalyzer and ABI StepOnePlus Real-Time PCR System were used in quantification and qualification of the sample

library before sequencing of the library using an Illumina HiSeq 2000 instrument in paired-end mode at 2 × 100 bp read length.

Subsequently, transcript-based and gene-based quantitations in transcripts per million (TPM) were computed using kallisto (Bray et al., 2016) in the bcbio workflow system (Chapman et al., 2020) using default parameters against the human reference genome hg19 with Ensembl 75 transcript annotations. The TPM expression quantitations were summarized on the gene level and directly used as "*exp*" features for downstream analyses.

For all subsequent analyses, the "*snv*", "*cnv*", "*lof*", and "*exp*" markers were subset to a list of 2,725 genes involved in cancer and/or known DDR pathways as derived from literature curation in order to decrease issues resulting from multiple testing and increase biological interpretability of the results.

In addition to the aforementioned single-gene features, also derived gene set-based molecular features were used in this study to capture the molecular status of signaling pathways and protein complexes. For this purpose, 252 "*coh_pat*" and 451 "*lof_pat*" markers, which stand for the expression level for coherently expression genesets and loss-of-function patterns for gene sets, respectively, were generated in the following manner:

For the "*coh_pat*" set of markers, the gene set collection was scored to identify coherently expressed gene sets similar to the method described in Staub et al. (Staub, 2012). Briefly, first the median pairwise Kendall correlation of the TPM expression of all genes in an individual gene set was computed. For gene sets with median pairwise correlation $\tau >= 0.5$, TPM values for all genes in an individual gene set were percentile-normalized and for each cell line the median percentile value across all genes in the gene set was selected as "*coh_pat*" feature value. The

features so derived summarize the expression status of the gene set as a single number for each cell line.

For the "*lof_pat*" set of markers, the same gene set collection was scored to identify gene sets enriched in loss-of-function mutations and copy number deletions. This was achieved by identifying gene sets whose number of loss-of-function events across all cell lines (as described above in the "*lof*" features) exceeded the expected number of such loss-of-function events under a hypergeometric null model (using number of loss-of-function events across all gene as background). If for a specific gene set the null hypothesis could be refuted with $\alpha <= 0.1$ (without multiple testing corrections), then the "*lof_pat*" feature for all cell lines with at least one loss-of-function gene in that gene set was coded as "1", and "0" else. The features so derived summarize the loss-of-function status of the gene set as a single number for each cell line.

Meanwhile, gene sets annotations used for above "*coh_pat*" and "*lof_pat*" were obtained from DDR-related literature curation, msigDB (Liberzon et al., n.d., 2011; Subramanian et al., 2005), Corum (Ruepp et al., 2010), KEGG (Kanehisa et al., 2020), Pathway Commons (Cerami et al., 2011), Pathway Interaction Database (PID, http://pid.nci.nih.gov). The molecular characteristics of all 86 cell lines (62 from the training set, 24 from the hold-out set) we obtained in this study are uploaded to the public data repository OSF: https://osf.io/8mxgj/.

Data resources for constructing feature sets

The chemical structure fingerprints, including MACCS, Morgan, RDK, FP2, FP3, and FP4, were generated by python OpenBabel (O'Boyle et al., 2011), PubChemPy (*PubChemPy Documentation — PubChemPy 1.0.4 Documentation*, n.d.), Pybel (*Pybel*, n.d.) and RDKit (Landrum, 2006) packages. Drugs-gene interactions were pulled from DGIdb3.0 (Cotto et al., 2018), LINCS (*HMS LINCS Project*, n.d.), DrugBank 3.0 (Knox et al., 2011), and CHEMBL

(Mendez et al., 2019). Synthetic lethality gene pairs were obtained from SynLethDB2.0 (J. Wang et al., 2022) with a cut-off with a confidence score over 0.8. Tissue-specific networks were obtained from HumanBase (https://hb.flatironinstitute.org/) (Wong et al., 2018).

The machine learning model construction

Based on the DDR combination treatment tests we obtained from the above, we constructed a machine-learning model using LightGBM, the gradient boosting machine, to predict both efficacy and synergy of combination treatments on new cell lines and tissue types (Ke et al., 2017). Gradient boosting methods have achieved the top performances in many recent data science challenges (Abel, n.d.; *ALASKA2 Image Steganalysis*, n.d., *IEEE-CIS Fraud Detection*, n.d., *iMaterialist Challenge (Fashion) at FGVC5*, n.d., *RecSys 2020 Challenge*, n.d.). Compared to traditional tree-based methods such as random forest and XGboost, LightGBM is especially suitable for tackling large datasets with its leave-wise tree growth characteristics, which shrinks training time without sacrificing prediction accuracy(Ke et al., 2017).

The machine learning model employed five types of feature sets as we mentioned in **Figure 5.1c**, including basic information (name and mode-of-action of the drugs in the combination), monotherapy responses, the chemical structure of drugs, gene set annotations, and molecular biomarkers, including genomic, expression and ddr readouts. The detailed numbers of each type of feature respectively were shown in **Supplementary Figure 5.2.**

Evaluation of model prediction performances

To evaluate the prediction performances on new cell lines/tissues, we implemented five-fold cross-validation by splitting the training and test dataset by cell line and tissue. Each time the model was trained on $4/5$ of cell lines or tissues and tested on the remaining $1/5$.

The average and 95% confidence interval of overall model predictions during all five folds were computed by bootstrapping. We combined predictions of all five folds together and calculated the prediction performance by bootstrapping. The accuracy of both combination efficacy (AoC score) and synergy (Bliss score) were evaluated by Pearson's correlation between prediction and gold standard.

To evaluate how the administration of combination treatments can be improved based on the prediction of machine learning models, we compared the machine learning model's choice of combination treatment with the combination treatment that achieves median efficacy on the cell lines (**Figure 5.2d**). The experimental efficacy of the machine learning model's choice of combination treatment, most of the time (95.2%), exceeds the median efficacy on the same cell line.

Interpretation of machine learning models by SHAP analysis

SHAP (SHapley Additive exPlanations) analysis evaluates the contributions of features on the predicted datasets using the SHapley value, which describes the average marginal contribution of a feature across all tested instances (Shapley, 1983). The average of the absolute values of all Shapley values of samples for each feature can be used to describe the contribution of the feature during prediction. We visualized the importance of all features during predictions in the k-fold cross-validation and selected the most important ones, as shown in **Supplementary Figures 5.5 and 5.6**.

Based on the additivity of SHAP values, we can carry out the following analysis to elucidate the relationship between features. First, the overall importance of a group of features can be computed as the sum of all SHAP values belonging to the feature set as the following:

$$SHAP_{a\ set\ of\ features} = \sum_{i=0}^{n} SHAP_{feature\ i} \quad \ldots\ldots \text{Eq. (5)}$$

Where n is the number of features in that feature set. For example, as for the molecular feature that comprises 10,462 molecular biomarker readouts, n = 10,462.

And the contribution of the feature set is the absolute value, or the magnitude of the SHAP value, as the feature could positively or negatively influence the prediction. Overall, the contribution of this set of features on the overall test set is computed as the average contribution, which is:

$$SHAP\ contribution\ of\ \ a\ set\ of\ \ features = \frac{1}{m}\sum_{j=1}^{m}\left|SHAP_{set\ of\ \ features}\right| \quad \ldots\ldots \text{Eq. (6)}$$

Where m is the total number of test sets.

Also, the importance of a single gene can be computed in a similar manner. As we have four types of molecular readouts (exp, snv, cnv, and lof) for single genes, the contribution of a single gene *i* when predicting from a single record is calculated as the sum of all types of molecular features of this gene.

$$SHAP_{gene\ i} = SHAP_{gene\ i_{exp}} + SHAP_{gene\ i_{snv}} + SHAP_{gene\ i_{cnv}} + SHAP_{gene\ i_{lof}} \quad \ldots\ldots \text{Eq. (7)}$$

The overall importance of this gene on the test set is the sum of the magnitude of all importance of this gene.

$$SHAP\ contribution\ of\ gene\ i = \frac{1}{m}\sum_{j=1}^{m}\left|SHAP_{gene\ i}\right| \quad \ldots\ldots \text{Eq. (8)}$$

Furthermore, we can also explore the fluctuation of SHAP importance between different genes by considering the synchronization of fluctuations of SHAP values of different genes. In this way, we would know that two genes are considered simultaneously important (or unimportant) in a drug combination effect prediction situation. Therefore, we constructed a correlation heatmap based on the SHAP values of all genes across all test sets.

Meanwhile, as the SHAP contribution is calculated based on the test sets, we can also analyze the mode-of-action specificity and tissue specificity of each feature by carrying out SHAP analysis on each tissue or mode-of-action subset. For instance, the SHAP contribution of feature i in subset S is:

$$SHAP\ contribution\ of\ feature_i = \frac{1}{m_s}\sum_{j=1}^{m_s}\left|SHAP_{feature_i}\right| \quad \ldots\ldots \text{Eq. (9)}$$

Where m_s is the total number of datasets in subset S. The distribution of SHAP contribution of top features on each different dataset, or the difference between top features in different datasets, can be used to elucidate how the machine learning model solves problems under different circumstances.

Shiny app for model application and evaluation

For feasible access to the presented model, we implemented a shiny app. The app provides a series of tools to analyze the above-described input data, predicted scores (efficacy and synergy), and derived SHAP values of the test data. The user can visualize the different input data groups or compare predicted efficacy and synergy scores for the prioritization of drug combinations. Various groupings of SHAP values provide vital insights into the contribution of individual genes or feature groups. Lastly, users can upload model input data to evaluate the efficacy and synergy of their own drug combinations. For further details, please see https://github.com/GuanLab/DDR-drug-synergy-prediction-ShinyApp.

Procedure for gene expression signature definition

Based on the SHAP analysis of gene expression features, we define gene signatures for clinically relevant treatment combinations ATMi-ATRi, ATRi-PARPi, ATRi-TOPi, ATRi-Cytostatic Antimetabolites, and DNA-PKi-IR. In particular, we extracted the top 40 gene

expression features (**Supplementary Figure 5.14-23**) for each combination. Next, we split these features into separate signatures that were either positively or negatively correlated with combination efficacy (AoC) or synergy (Bliss). For deriving succinct gene signatures, we further split these signatures into smaller subsets of genes based on their contribution as defined by the SHAP analysis (resulting in 180 signatures). Finally, we select the smallest gene signature for each treatment combination with coherent gene expression scores larger than 0.1. The coherent gene expression score is computed by a method similar to the method described in Staub et al. (Staub, 2012). Briefly, the median pairwise Kendall correlation of the TPM expression of all genes within a gene set was computed. The signature score for each signature then was computed as the mean z-scaled log2 transformed expression values of all genes in the signature, across all cell lines.

Data Availability

The molecular readouts for all 87 cell lines involved in this study were downloaded from https://osf.io/8mxgj/**.** Source data are provided with this paper.

Code Availability

The source code of the analysis containing the surrogate machine learning model is available from GitHub: https://github.com/GuanLab/DDR-Drug-Synergy-Prediction. The R Shiny app in this study is available at; https://github.com/GuanLab/DDR-drug-synergy-prediction-ShinyApp.

Figures

Figure 5.1. Overview of the Combination Treatment Synergy Screening Dataset and Construction of the Machine Learning Synergy Prediction Model. (a). Training and external validation of the DDR combination treatment prediction model. Cross-validation was conducted with splits based on cell lines/tissue types for the training set, and then validation was performed on a hold-out dataset with new cell lines and tissue types, independent of the training set. **(b).** Distribution of cell lines from the training set and hold-out sets across different tissue types. **(c).** t-SNE clustering based on molecular marker read-outs of all cell lines from various tissues and training/hold-out sets. **(d).** Strategy for building the machine learning model in this study. Nine types of information were integrated into the input features, including basic drug information such as drug names and modes of action, chemical structure, monotherapy response, drug-target interaction, gene set annotations, tissue-specific biological networks, synthetic lethality, and molecular signatures. These feature sets were then input into the LightGBM gradient boosting machine to predict efficacy and synergy for DDR combination treatments. Feature visualization was performed on the model generated by SHAP analysis, and top features were selected as candidate biomarkers.

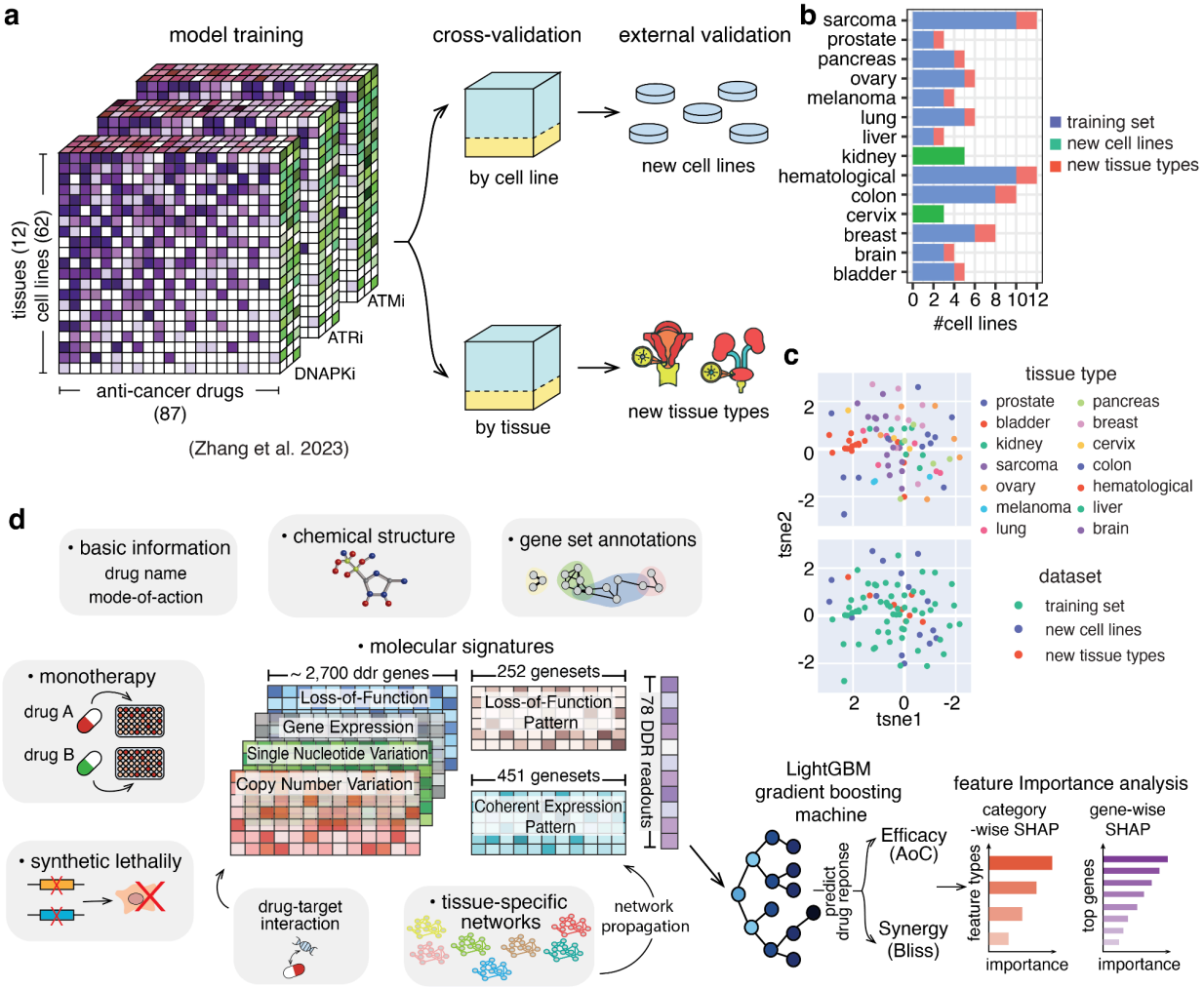


Figure 5.2. Performances of machine learning models in predicting DDR combination treatment responses. (a) and **(b)** show the performances of machine learning models using the information mentioned in **Figure 5.1a** in **(a)** cross-validation and **(b)** external validation, respectively. Boxplots display median lines and 25th to 75th interquartile ranges, with whiskers extending to 1.5 times the interquartile range. The best-performing model was marked by asterisks. **(c).** Performance of the cross-tissue model when validated on different tissue types, for both efficacy and synergy. The dashed line marks the global performance of all tissue types combined. Error bars indicate the 95% confidence interval. **(d).** Improvement of the efficacy of the prioritized combination treatments from the baseline treatment (the treatment achieved median efficacy of all combination treatments on the same cell line) on different cancer cell lines by the machine learning model in this study.

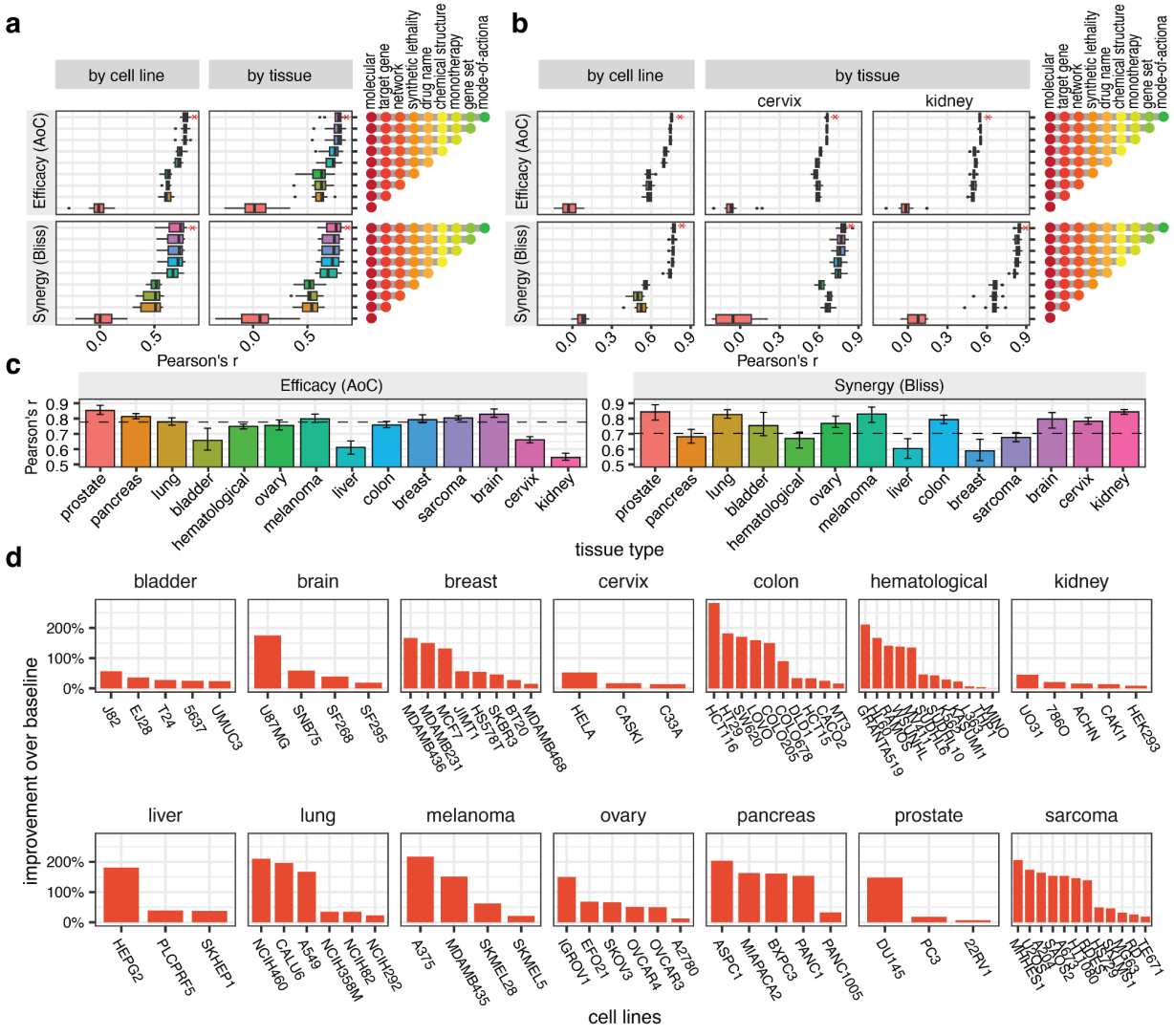


Figure 5.3. Contribution of predictive features. (a) and (f). Contribution of broad categories of features in predicting **(a)** efficacy and **(f)** synergy. **(b) and (g).** Contribution of different types of molecular features in predicting **(b)** efficacy and **(g)** synergy. **(c) and (h).** Contribution of different types of fingerprinting methods in predicting **(c)** efficacy and **(h)** synergy. **(d) and (i)**. Contribution of top synthetic lethality in predicting **(d)** efficacy and **(i)** synergy. DDR kinases ATM and ATR are marked in red, and direct drug targets are marked by asterisks. **(e) and (j)**. Contribution of top molecular features in predicting **(e)** efficacy and **(j)** synergy. **(k) and (l)**. Top predictive features of ATM, ATR, and PRKDC (DNA-PK)co-therapies for the **(k)** efficacy and **(l)** synergy**.**

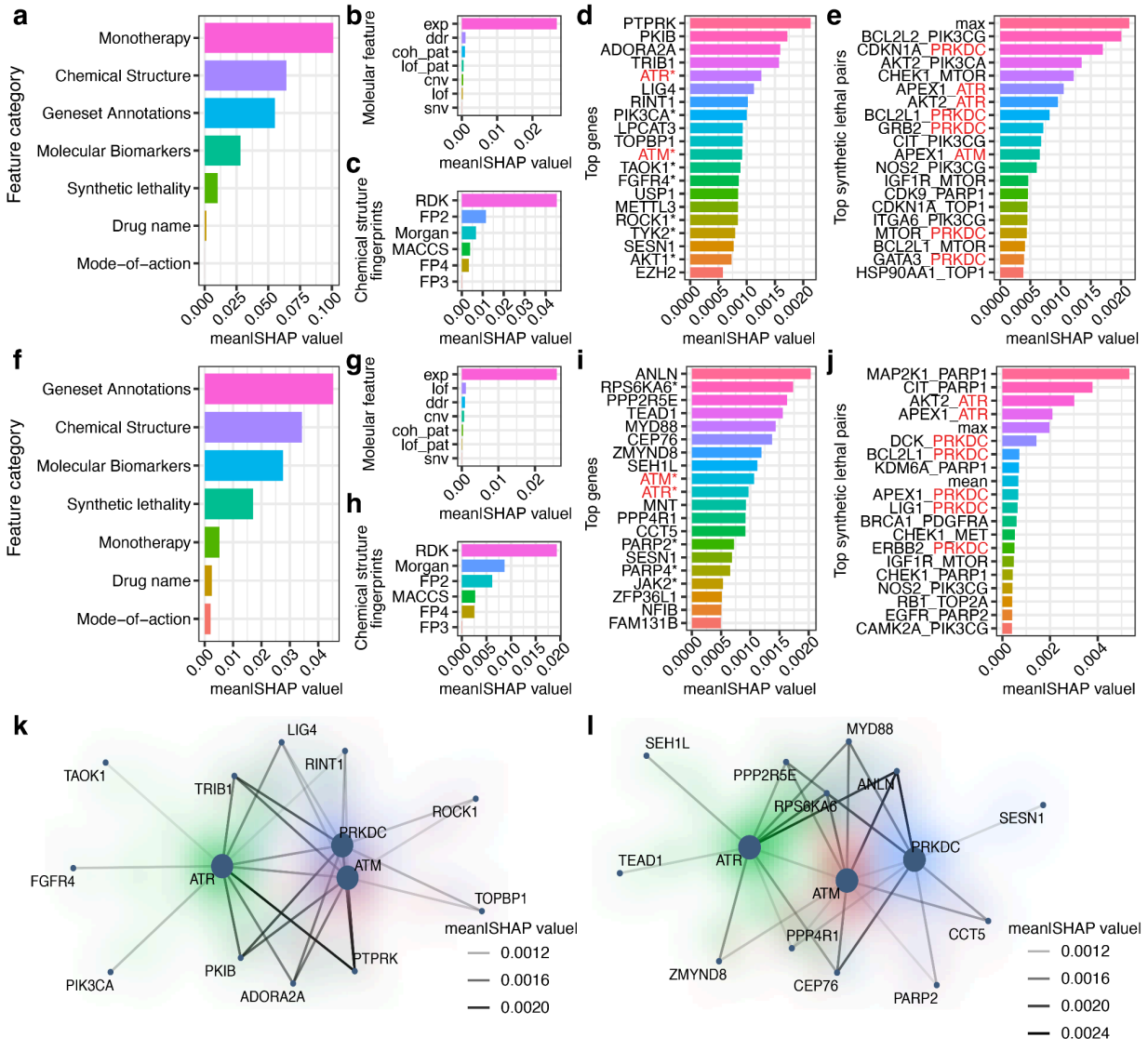


Figure 5.4. Important genes and pathways and genes identified in this study are highly correlated with ATM/ATR/DNA-PK inhibitor combination treatment.

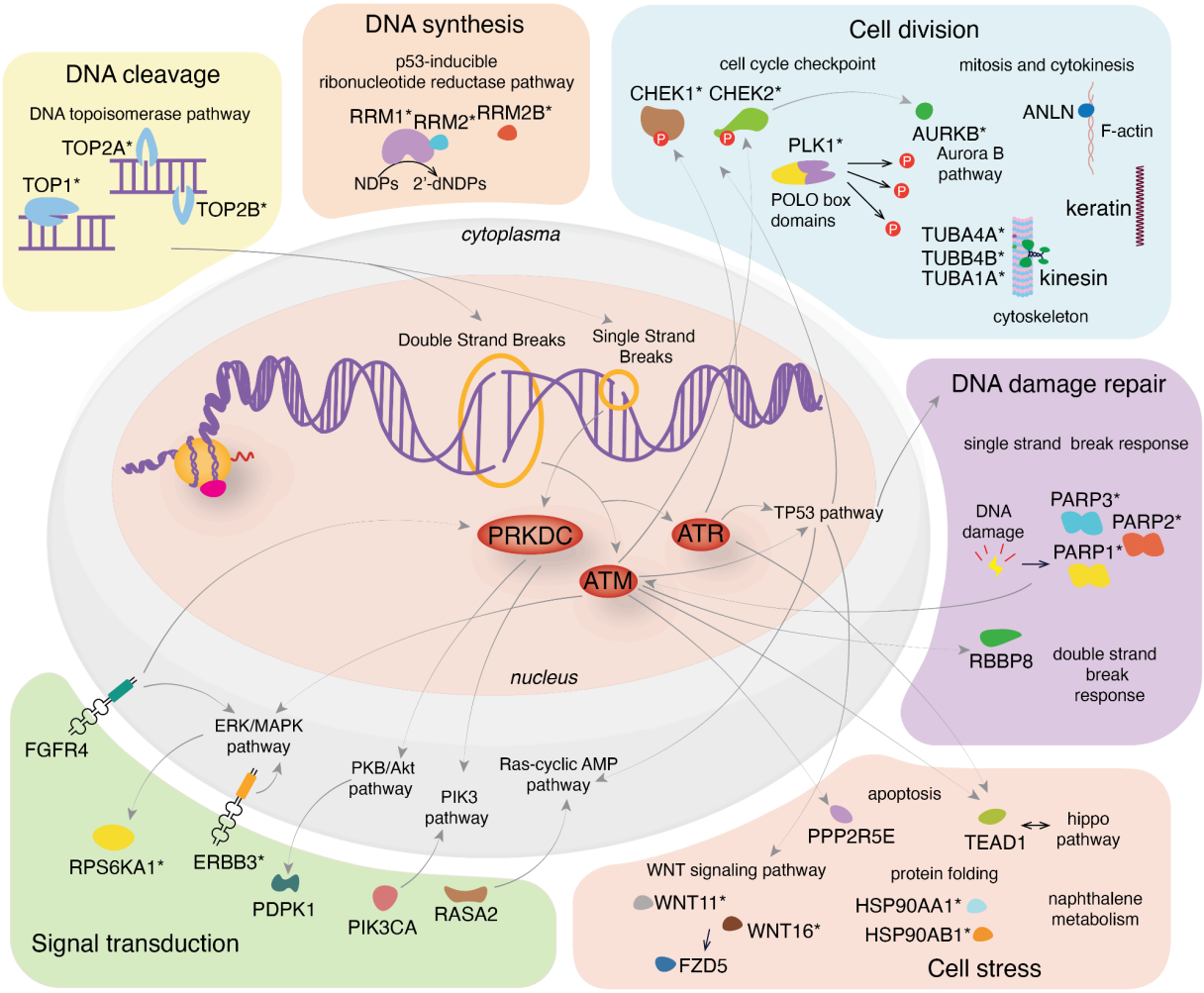
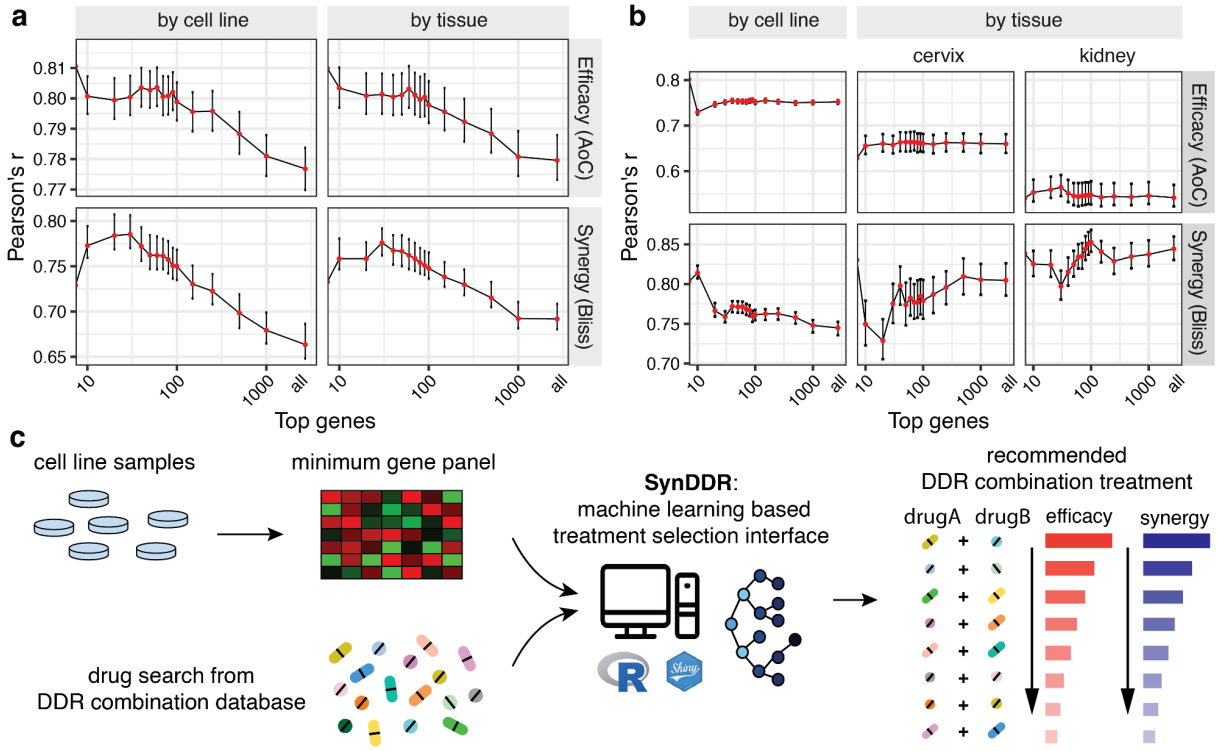


Figure 5.5. Constructing a surrogate model using a minimal gene panel. (a) and **(b)** shows the model performances by using only top contributing genes in the machine learning models in **(a)** cross-validation and **(b)** external validation on the hold-out dataset. **(c).** The diagram shows the process of collecting samples, drug search, and treatment recommendations by SynDDR Shiny App. Only the expression level of a minimum gene panel of 40 genes is required for treatment prediction. The optimal treatment can be selected based on both synergy and efficacy.



Supplementary Tables

Supplementary Table 5.1. Top 50 genes positively correlated with efficacy in ATRi monotherapy.

gene	Pearson's r	P value
CEBPA	0.2188828747526469	3.1528514498582526e-08
BRD3	0.2051399970863431	2.248731999656275e-07
PDSS1	0.20214827878897496	3.389593673651752e-07
ETV6	0.1975317723035595	6.308338387463387e-07
ZFP36L2	0.19093541704531555	1.49432455678798e-06
ERF	0.18721938117653691	2.397804230069868e-06
MT1G	0.18522706494745508	3.0779526985667635e-06
POLR2B	0.18165777665032803	4.78264938978935e-06
APEX1	0.18021272487010326	5.70312283089032e-06
FBXO18	0.17632106001753828	9.098987620026203e-06
TFAP4	0.17042589410802944	1.8115711357122857e-05
LRP6	0.16956621069903582	1.9991236013571346e-05
KEAP1	0.16360220353168275	3.9074086935800896e-05
LTBP4	0.16272471081054996	4.303919535654521e-05
ZNF384	0.15962168362884444	6.0334305303904244e-05
PRPF19	0.1589317449862145	6.498535289195255e-05
TRAF7	0.15783761927941475	7.30613139943933e-05
MPG	0.1573200846744955	7.72029810869797e-05
FES	0.1567661208178487	8.188106091310689e-05
BAG4	0.1564466255672635	8.469933733831009e-05
MARK4	0.1535317734037663	0.00011499
GFI1	0.15345822128707548	0.00011587
E2F4	0.1529751543258198	0.00012182
PRR12	0.15064728510333836	0.00015478
MXD4	0.15008871128712464	0.00016385
CCNB1IP1	0.14928373878745077	0.0001778
CD4	0.14810257441833954	0.00020029
CCT7	0.1480566798688799	0.00020122
DKK4	0.1473384966761277	0.00021624
TGFB1	0.14657856644939982	0.00023327
INSR	0.14569368189223536	0.00025468
GTF2H3	0.14491036	0.00027516
TTBK2	0.14468615556688785	0.0002813

CPSF6	0.14457189606434545	0.00028448
DOT1L	0.14417774965522515	0.0002957
LGR4	0.1438418518378297	0.00030558
RASGRP4	0.1436957302101679	0.00030998
FUS	0.14314331748002912	0.00032714
RUVBL2	0.14291464749676236	0.0003345
MCMBP	0.14291330581263992	0.00033454
RALGDS	0.14279739693842491	0.00033833
CAT	0.1425477311017998	0.00034663
FOXP2	0.14231076676022308	0.00035468
TRIB1	0.14124455220303933	0.00039312
PTPN6	0.1406827765676939	0.0004149
IKBKAP	0.1404735634041649	0.0004233
LYZ	0.14039486680518065	0.0004265
LPCAT3	0.1401497593812305	0.0004366
FCGR1A	0.1389510736670722	0.00048932
HSPE1	0.13882158204091127	0.00049536

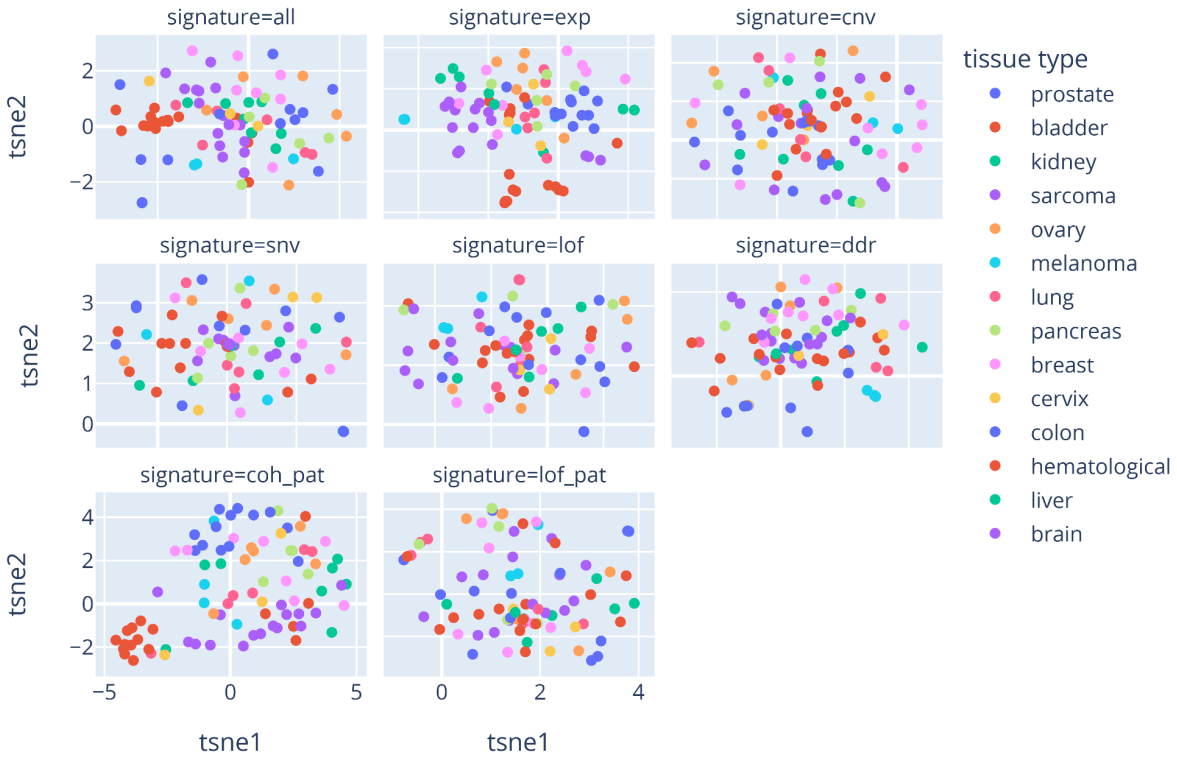
Supplementary Table 5.2. Top 100 genes negatively correlated with efficacy in ATRi monotherapy.

gene	Pearson's r	P value
PLAT	-0.2799987	9.69546521613379e-13
ALPP	-0.23851	1.5125877274690443e-09
FUT8	-0.237553	1.7653024027001322e-09
TNFRSF10C	-0.2180376	3.571407173519536e-08
TTC37	-0.2176899	3.758740911067073e-08
SNX19	-0.217207	4.034783460523463e-08
RRM1	-0.2167851	4.291856589209018e-08
CD9	-0.2151778	5.424621940301407e-08
LASP1	-0.2149607	5.598176883114944e-08
GATA2	-0.2135997	6.815166586603542e-08
CLIC5	-0.2132912	7.124603686491893e-08
NF1	-0.2112306	9.56855240645544e-08
ARSK	-0.2074292	1.635941437114444e-07
DCUN1D1	-0.1983852	5.630184837466466e-07
ETS1	-0.1977406	6.135469079253801e-07
HMCES	-0.1880753	2.1521402471594824e-06
RASA1	-0.1879021	2.199819548434763e-06
RPA4	-0.1861071	2.75741552095053e-06
FOSL1	-0.1858553	2.845694753552529e-06
CDH1	-0.1844147	3.405281309676142e-06
CAST	-0.1828027	4.155979828190317e-06
WNT9A	-0.1785731	6.952174968124344e-06
SYTL2	-0.1780671	7.38765992084861e-06
NABP1	-0.1779763	7.468493564512224e-06
GAD1	-0.1774478	7.956119137805754e-06
COPS6	-0.176799	8.596306646716824e-06
PICALM	-0.1752939	1.027574956357842e-05
PLAU	-0.1749289	1.072791603378244e-05
WEE1	-0.1726088	1.407571439257513e-05
PEA15	-0.1718813	1.5316054115582397e-05
TEAD1	-0.1706769	1.7600362672490354e-05
KAT2B	-0.1705629	1.783274185599646e-05
TOLLIP	-0.1681692	2.3437955449662482e-05
LCK	-0.1655348	3.152656559622629e-05
MAP3K13	-0.1641321	3.684993456526911e-05

ABL2	-0.16361	3.9040644696342785e-05
SPAG9	-0.1623612	4.479071636284836e-05
EIF5A2	-0.1616369	4.8483922923373674e-05
RASGRP1	-0.1615469	4.89624046186599e-05
CCDC127	-0.1604579	5.5118394540816674e-05
RAB25	-0.1604332	5.526625515690283e-05
SP100	-0.1589243	6.503699868149149e-05
YWHAZ	-0.158072	7.125508552234728e-05
BIRC3	-0.1569782	8.005913250122755e-05
SP140L	-0.1564441	8.472239085603003e-05
IFI6	-0.1553763	9.482196364963771e-05
PPM1H	-0.1544828	0.00010413
GOSR1	-0.1540279	0.0001092
MICALCL	-0.152422	0.000129
PDGFB	-0.1518864	0.00013632

Supplementary Figures

Supplementary Figure 5.1. t-SNE clustering based on all different types of molecular marker read-outs of all cell lines from various tissues in this study. all: all biomarker combined; exp: mRNA-based gene expression;cnv: copy number variations; snv: single nucleotide variants; lof: loss-of-function of the gene; ddr: DDR-related readouts; coh-pat: derived gene cluster-based features measuring expression; lof_pat: loss-of-function of cancer pathways)



Supplementary Figure 5.2. The composition of all types of features used in the machine learning model in this study. (a) Molecular biomarkers. **(b)** Geneset annotations. **(c)** Synthetic lethality gene pairs. **(d)** General drug information. **(e)** Chemical structure fingerprints. **(f)** Overlapped genes between four types of single-gene molecular markers.

a single gene molecular biomarker

type of biomarker	# genes
single nucleotide variation (snv)	1,740
copy number variation (cnv)	2,569
gene expression level (exp)	2,725
loss of function (lof)	2,647

gene cluster molecular biomarker

type of biomarker	# cluster
coherently expressed gene set (coh_pat)	252
loss-of-function gene set (lof_pat)	451

special DDR readouts

type of biomarker	# readout
DDR-related signatures (ddr)	78

b geneset annotations

# geneset annotations	# genes
18,681	9,450

c synthetic lethality

type of biomarker	# gene pairs
synthetic lethal gene pair	1157

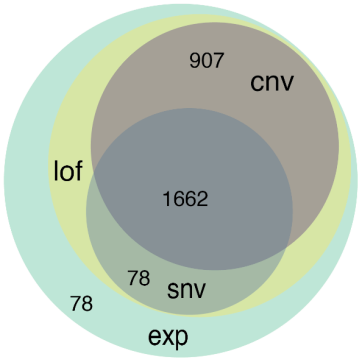
d general drug information

type of drug information	# names
drug name	87
mode-of-action	61

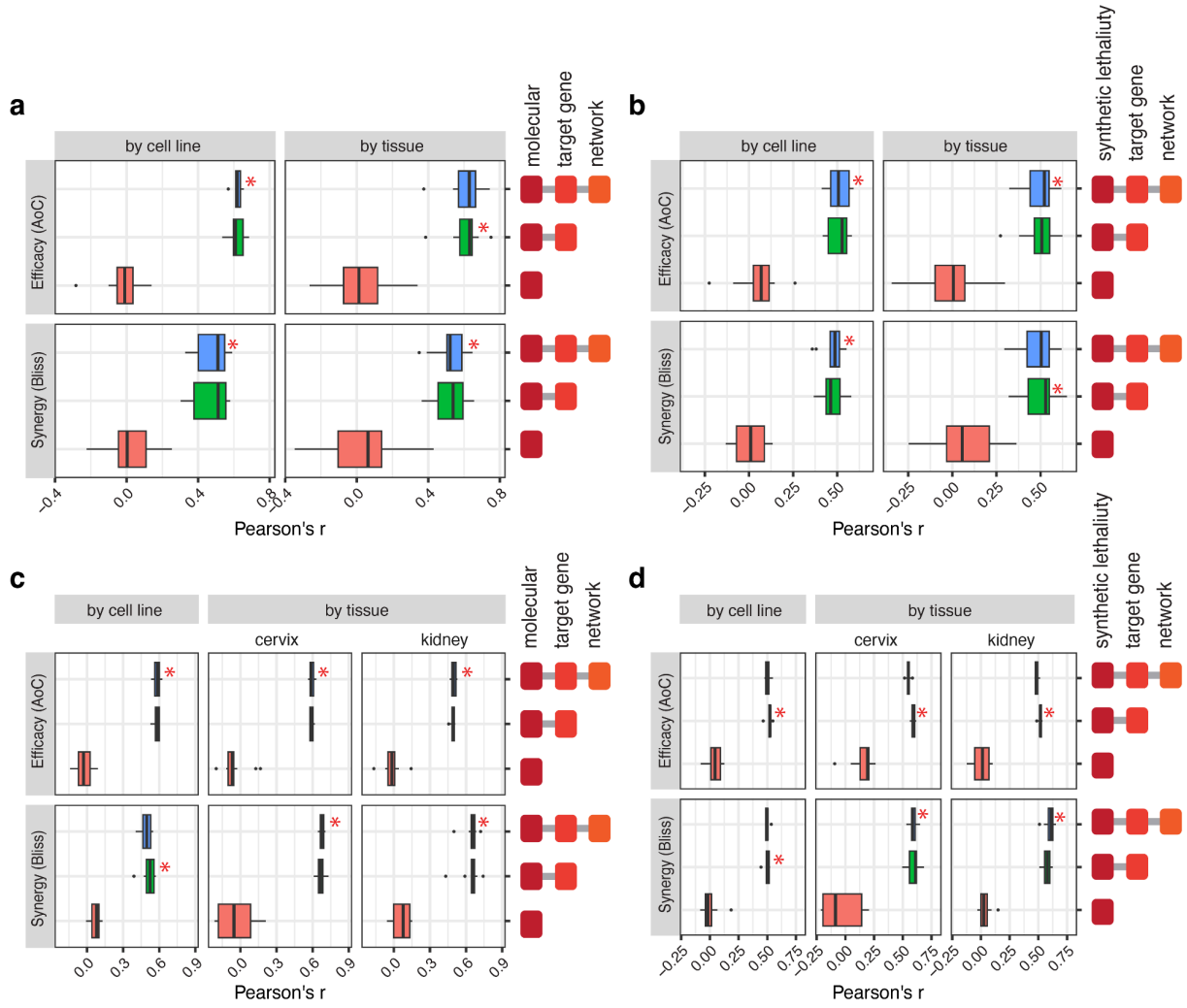
e chemical structure fingerprints

fingerprint type	# fingerprints
MACCS	167
Morgan	1,024
RDK	2,048
FP2	1,024
FP3	56
FP4	308

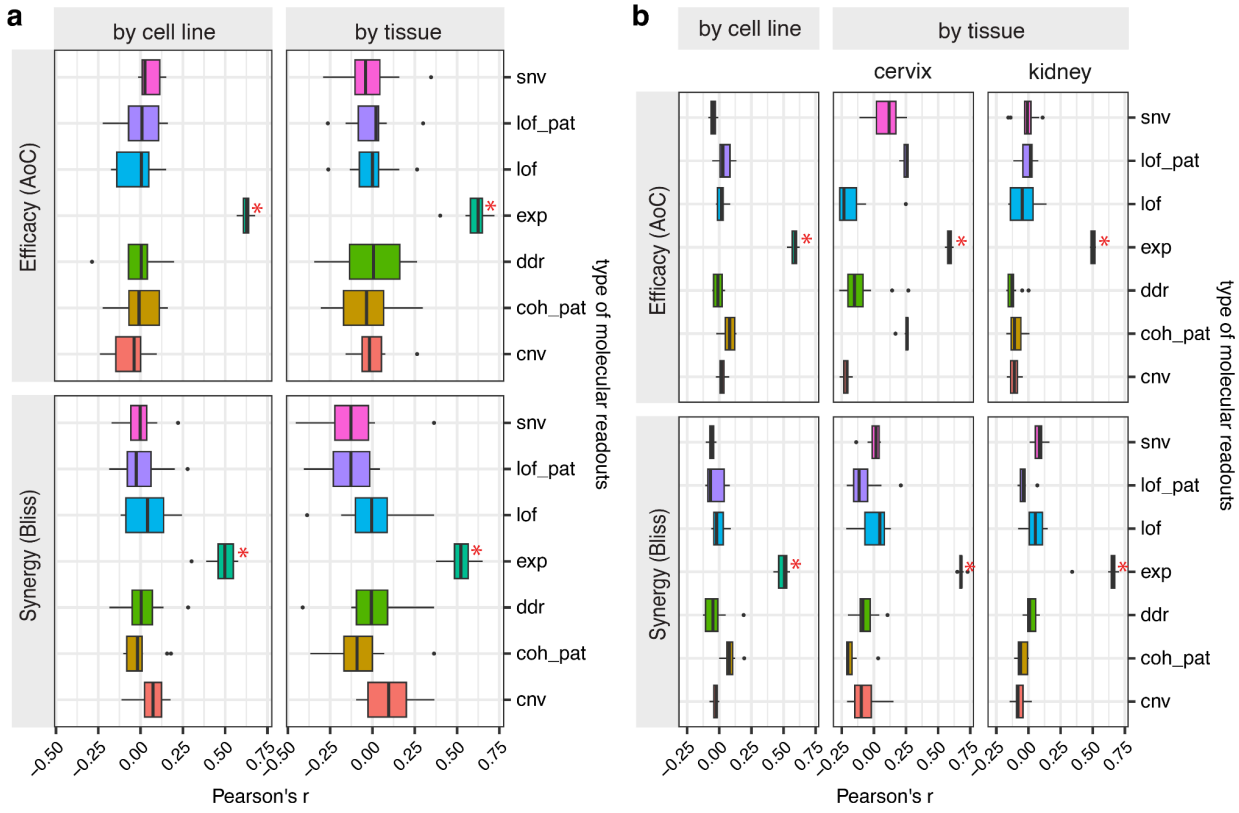
f



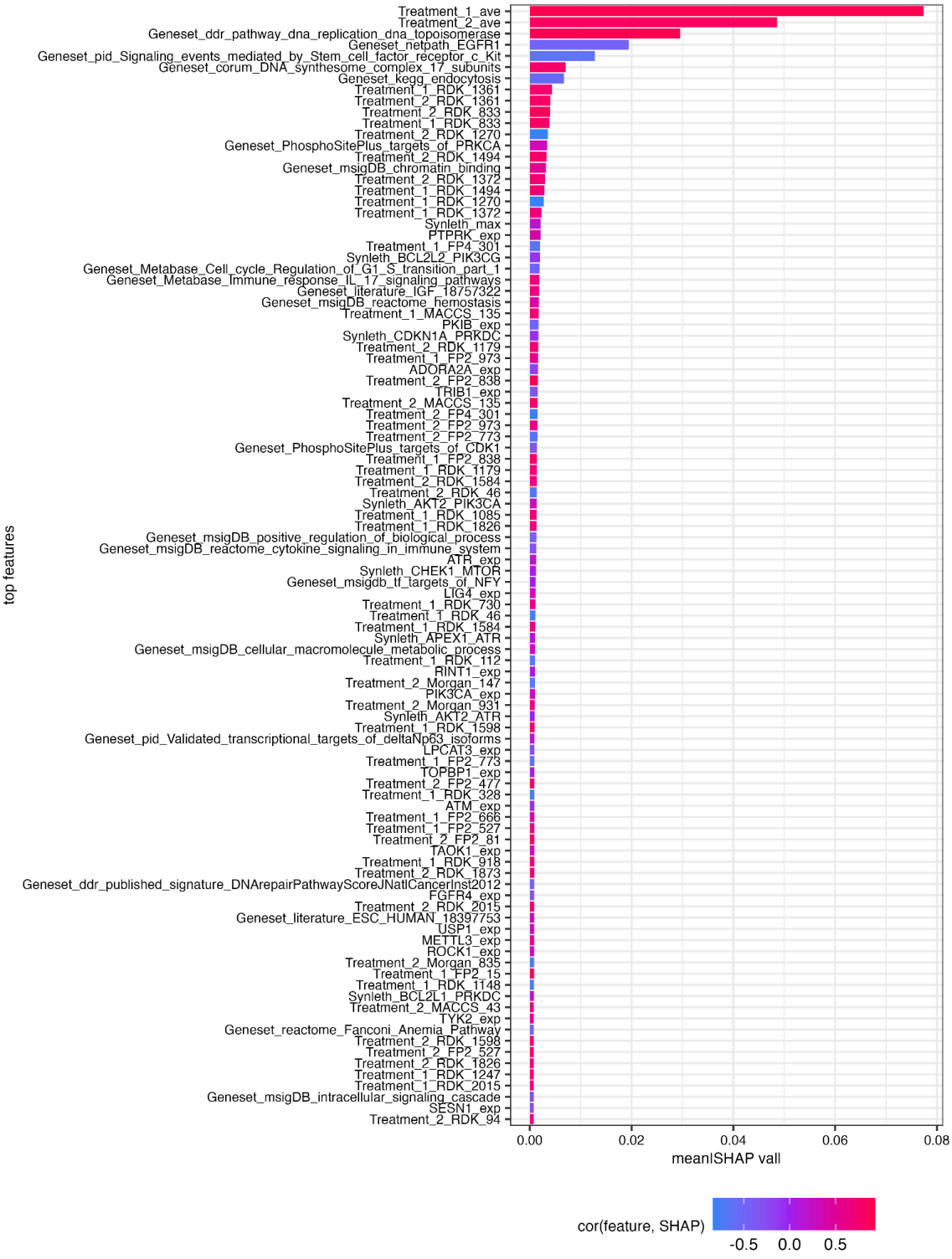
Supplementary Figure 5.3. Performances of machine learning models using molecular/synthetic lethality information in combination with target genes and tissue-specific network information. **(a)** and **(b)** show the performances of machine learning models using molecular profiles and **(c)** and **(d)** show the performances of machine learning models using synthetic lethality as a baseline. **(a) and (c)** shows results from cross-validation, and **(b) and (d)** shows results from external validation, respectively. Boxplots display median lines and 25th to 75th interquartile ranges, with whiskers extending to 1.5 times the interquartile range. Asterisks marked the best-performing models.



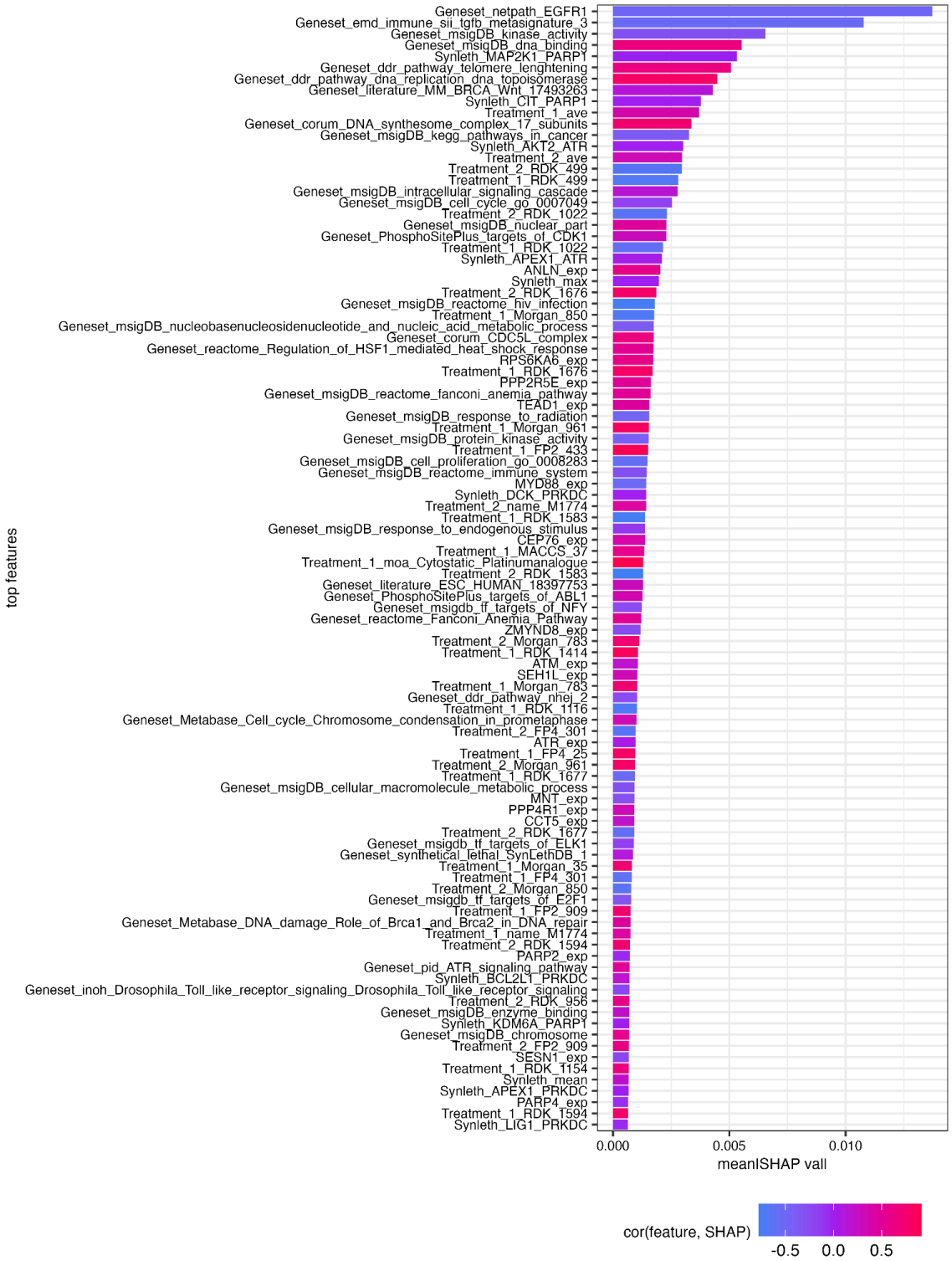
Supplementary Figure 5.4. Performances of machine learning models using different types of molecular information in combination with target genes and tissue-specific network information. **(a)** and **(b)** show the performances of machine learning models using different kinds of molecular profiles of results from **(a)** cross-validation, and **(b)** external validation, respectively. Boxplots display median lines and 25th to 75th interquartile ranges, with whiskers extending to 1.5 times the interquartile range. Asterisks marked the best-performing models.



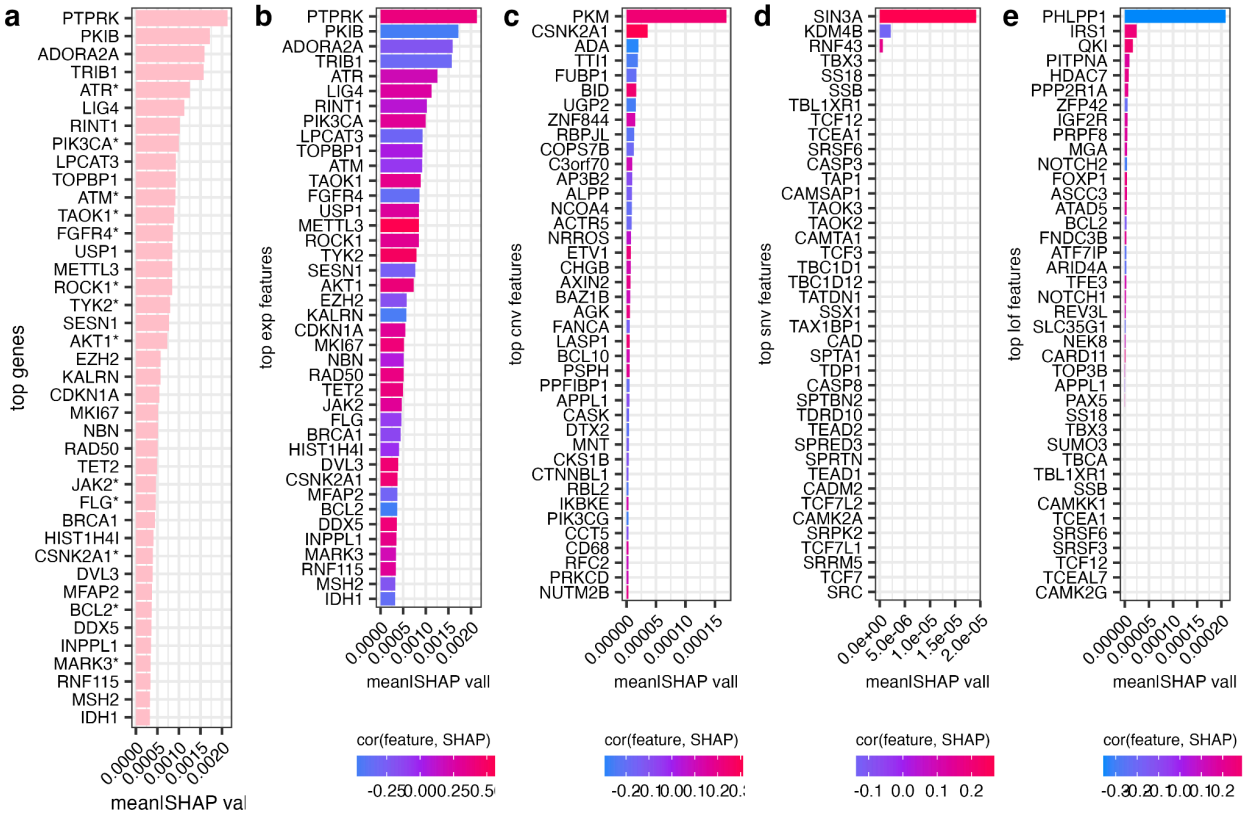
Supplementary Figure 5.5. SHAP contributions of top 50 features in drug efficacy prediction model.



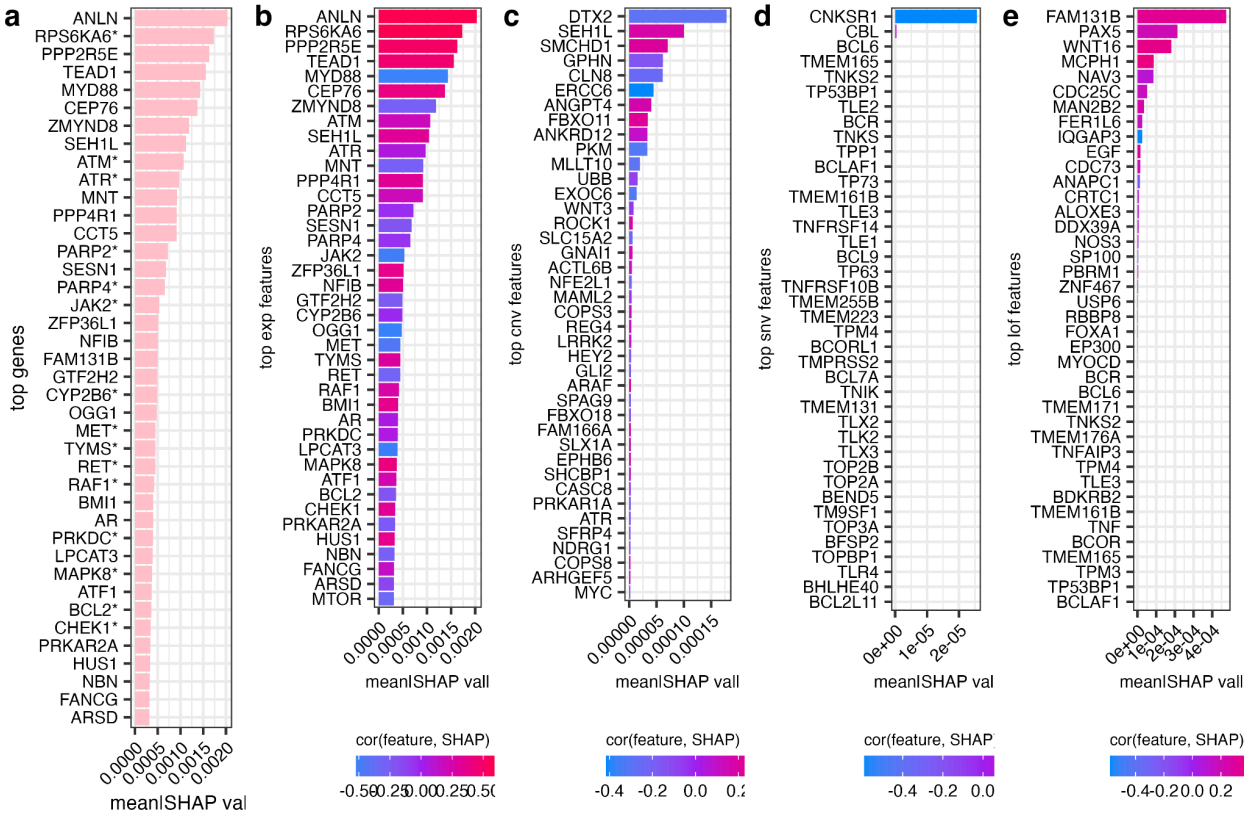
Supplementary Figure 5.6. SHAP contributions of top 50 features in drug synergy prediction model.



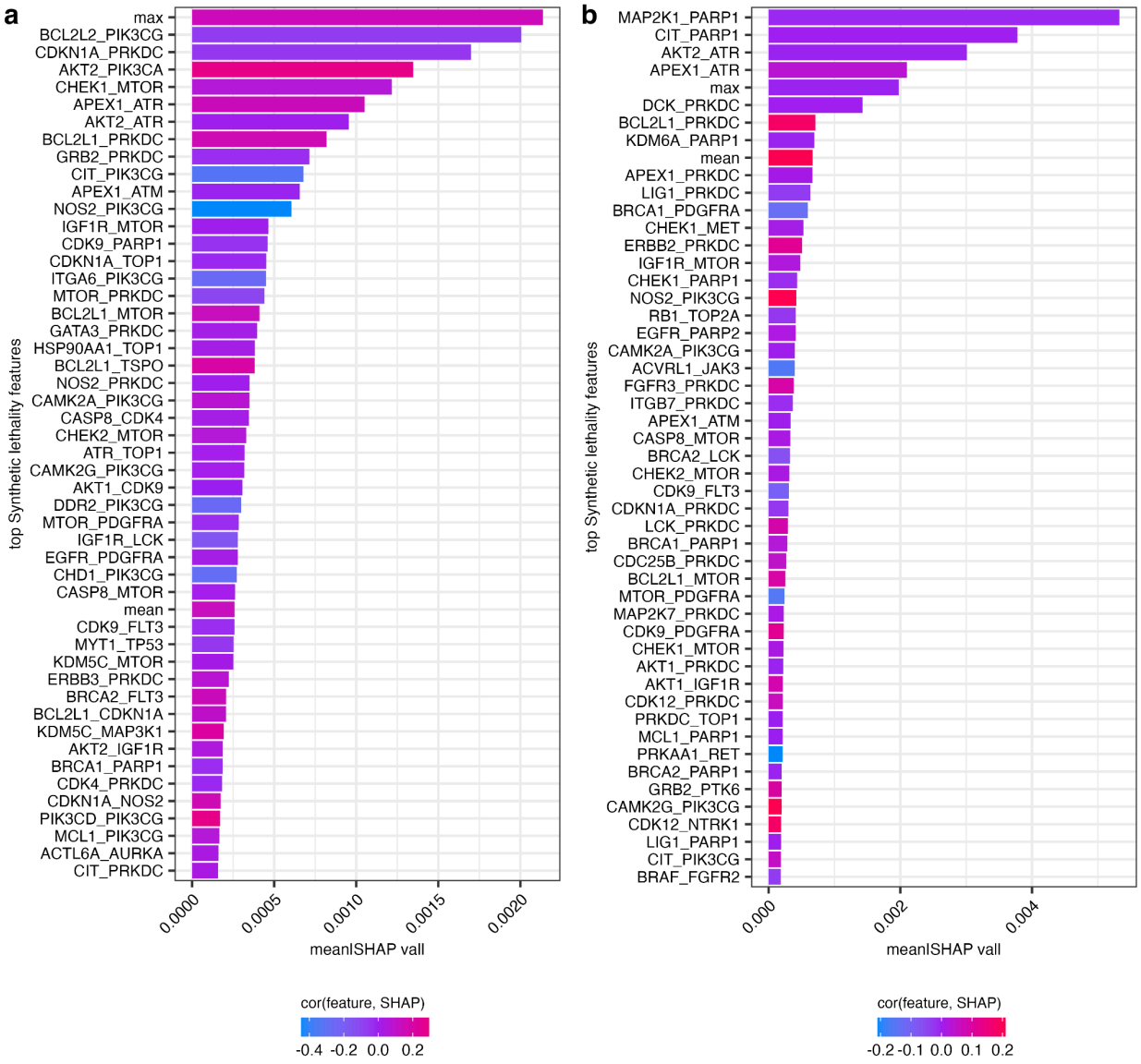
Supplementary Figure 5.7. Top genes in drug efficacy prediction as evaluated by each type of molecular biomarkers ("*exp*", "*cnv*", "*snv*" and "*lof*"). **(a).** The top 40 genes after combining all types of molecular markers. **(b), (c), (d), and (e)** show the top genes when considering expression levels ("*exp*"), copy number variation ("*cnv*"), loss-of-function of the gene ("*lof*"), and single nucleotide variation ("*snv*"), respectively. *: direct drug targets.



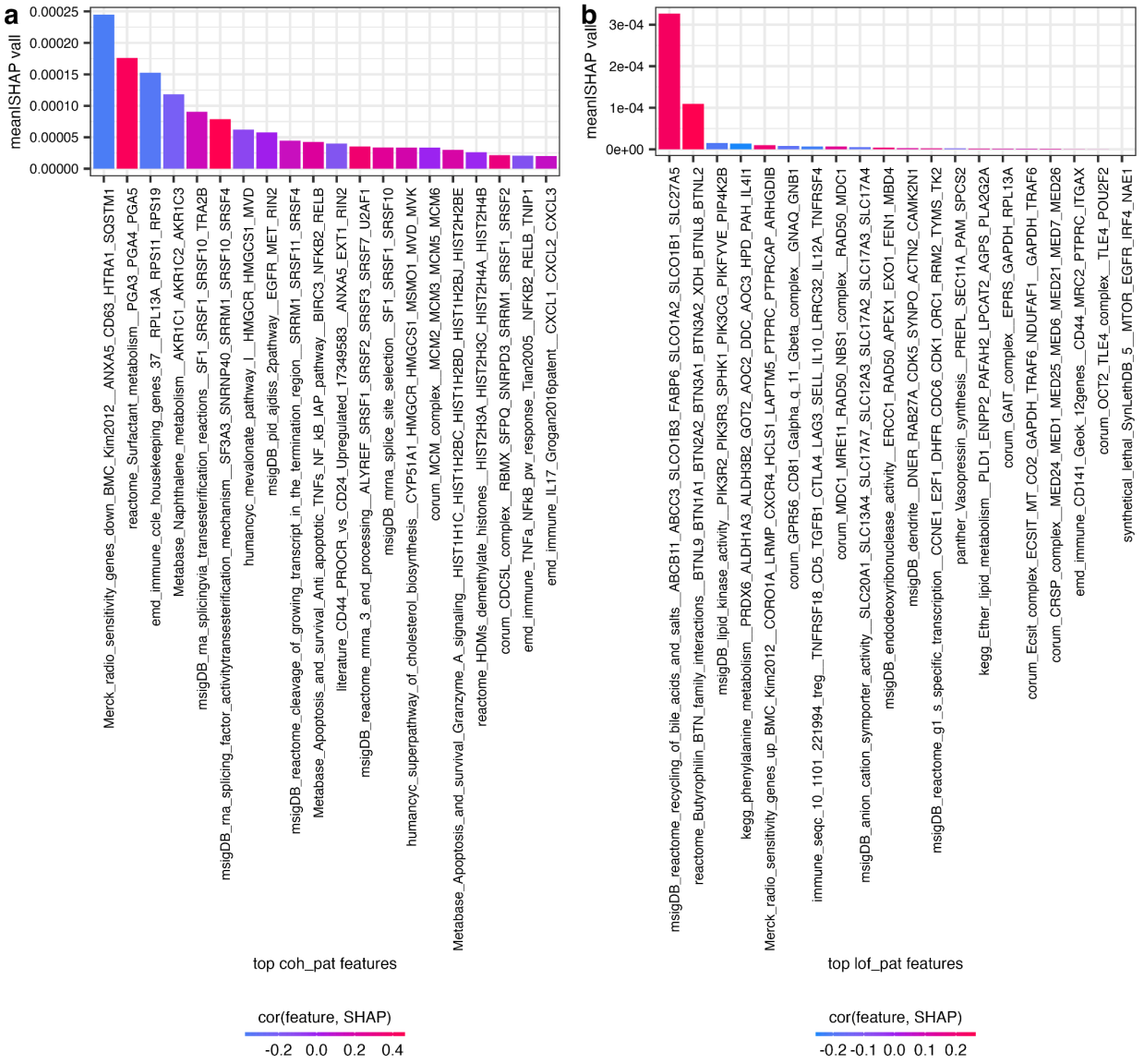
Supplementary Figure 5.8. Top genes in drug synergy prediction as evaluated by each type of molecular biomarkers ("*exp*", "*cnv*", "*snv*" and "*lof*"). **(a).** The top 40 genes after combining all types of molecular markers. **(b), (c), (d), and (e)** show the top genes when considering expression levels ("*exp*"), copy number variation ("*cnv*"), loss-of-function of the gene ("*lof*"), and single nucleotide variation ("*snv*"), respectively. *: direct drug targets.



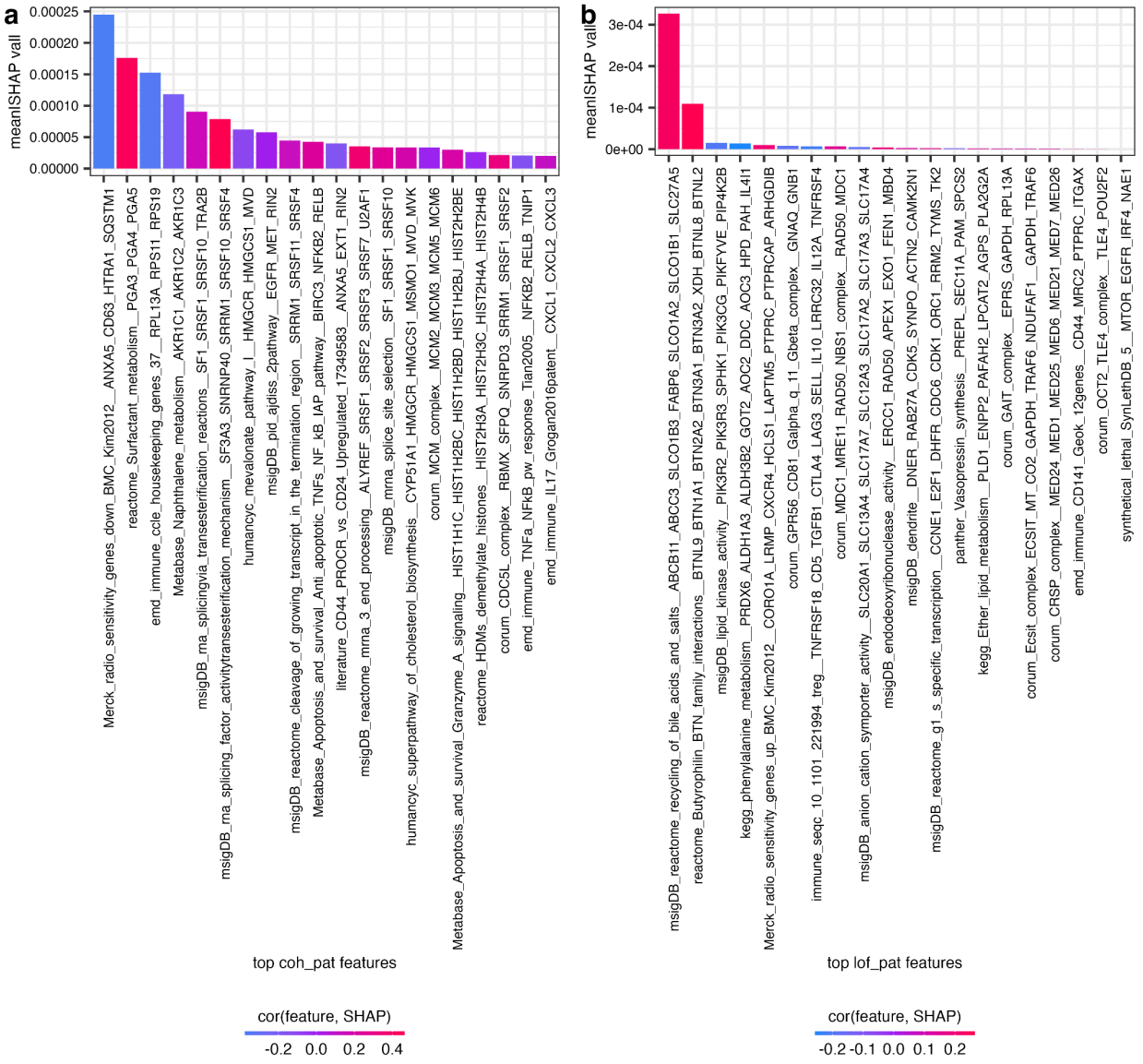
Supplementary Figure 5.9. Top synthetic lethality features in drug combination prediction. **(a) and (b)** show top synthetic lethal gene pair features in **(a)** efficacy and **(b)** synergy prediction, respectively.



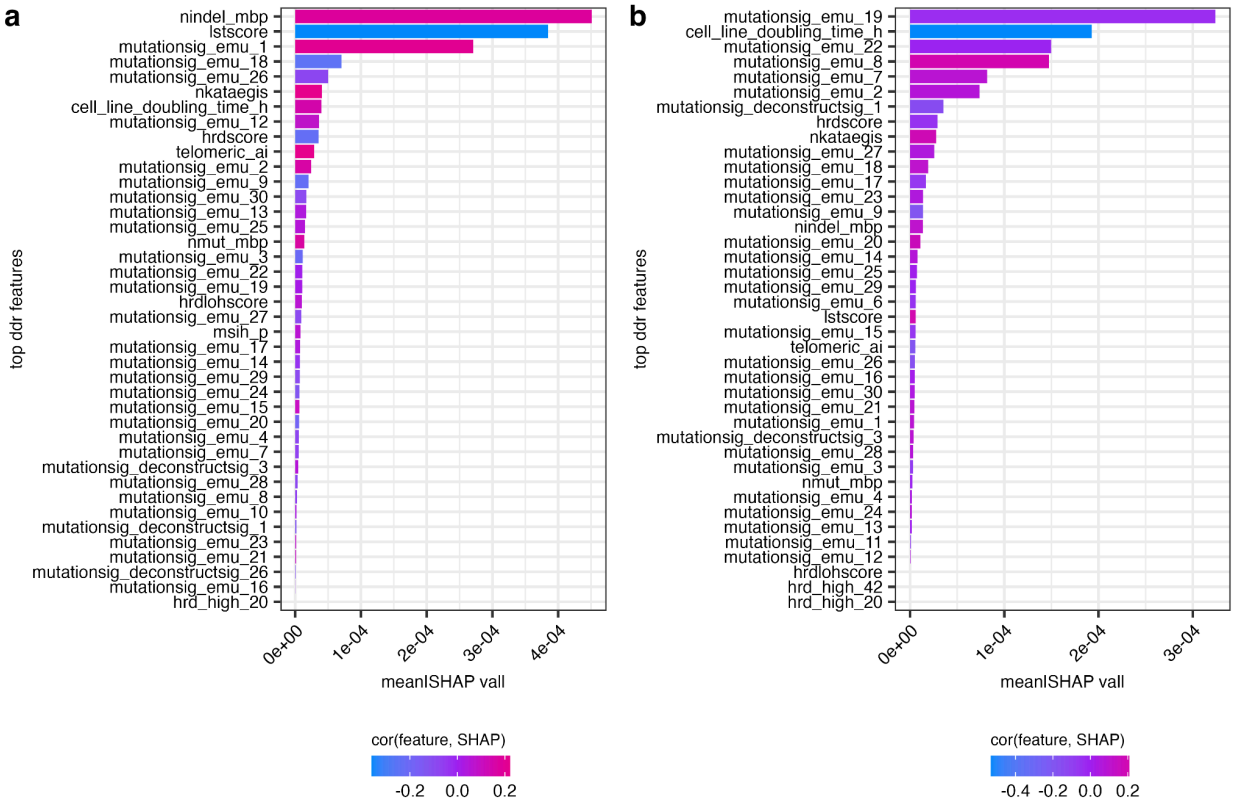
Supplementary Figure 5.10. Top geneset cluster molecular biomarkers in drug combination prediction. **(a)** and **(b)** show the top gene clusters in terms of **(a)** coherent expression patterns ("*coh_pat*") and **(b)** loss-of-function patterns ("*lof_pat*") for efficacy prediction.



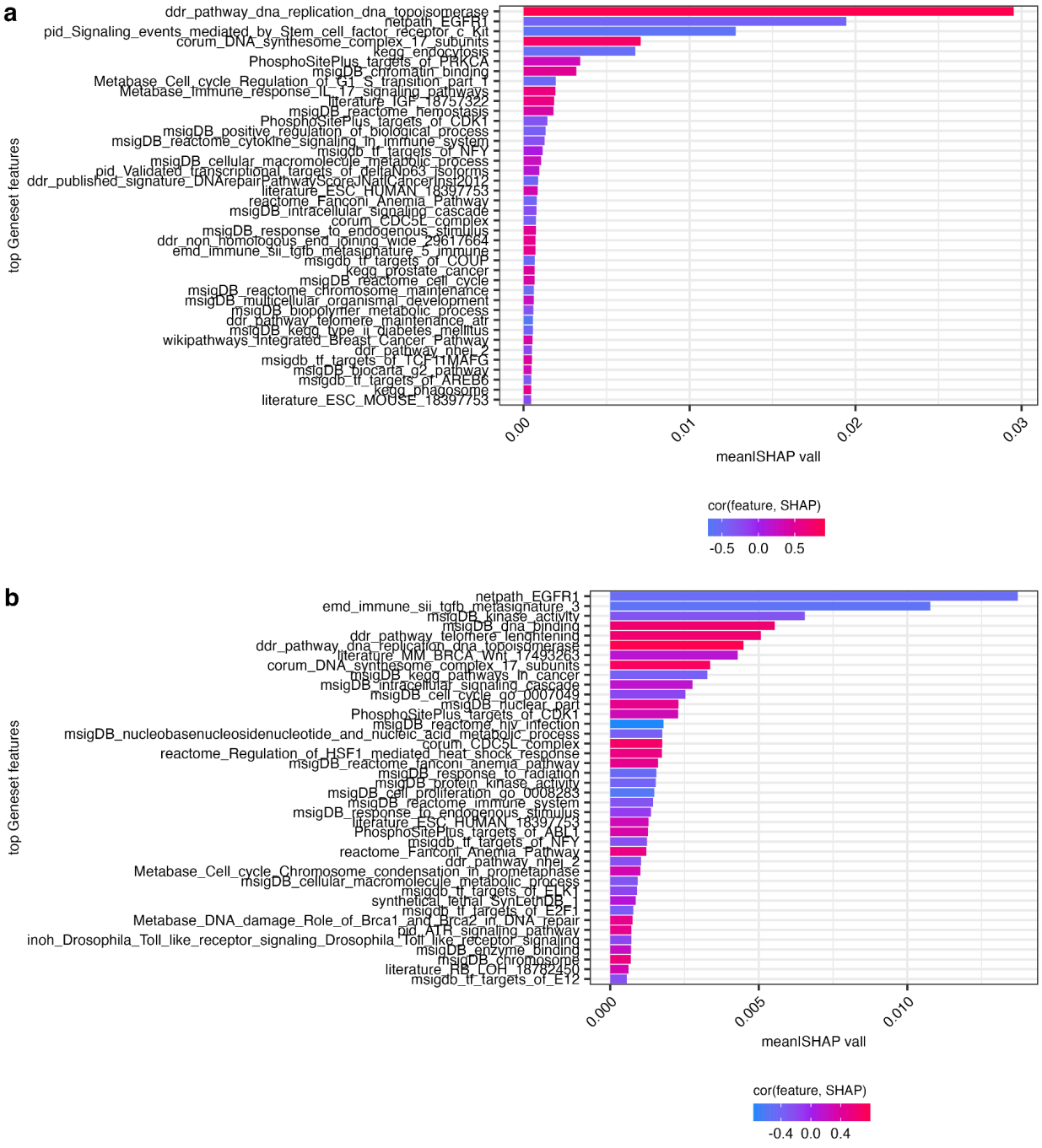
Supplementary Figure 5.11. Top geneset cluster molecular biomarkers in drug combination prediction. **(a)** and **(b)** show the top gene clusters in terms of **(a)** coherent expression patterns ("*coh_pat*") and **(b)** loss-of-function patterns ("*lof_pat*") for synergy prediction.



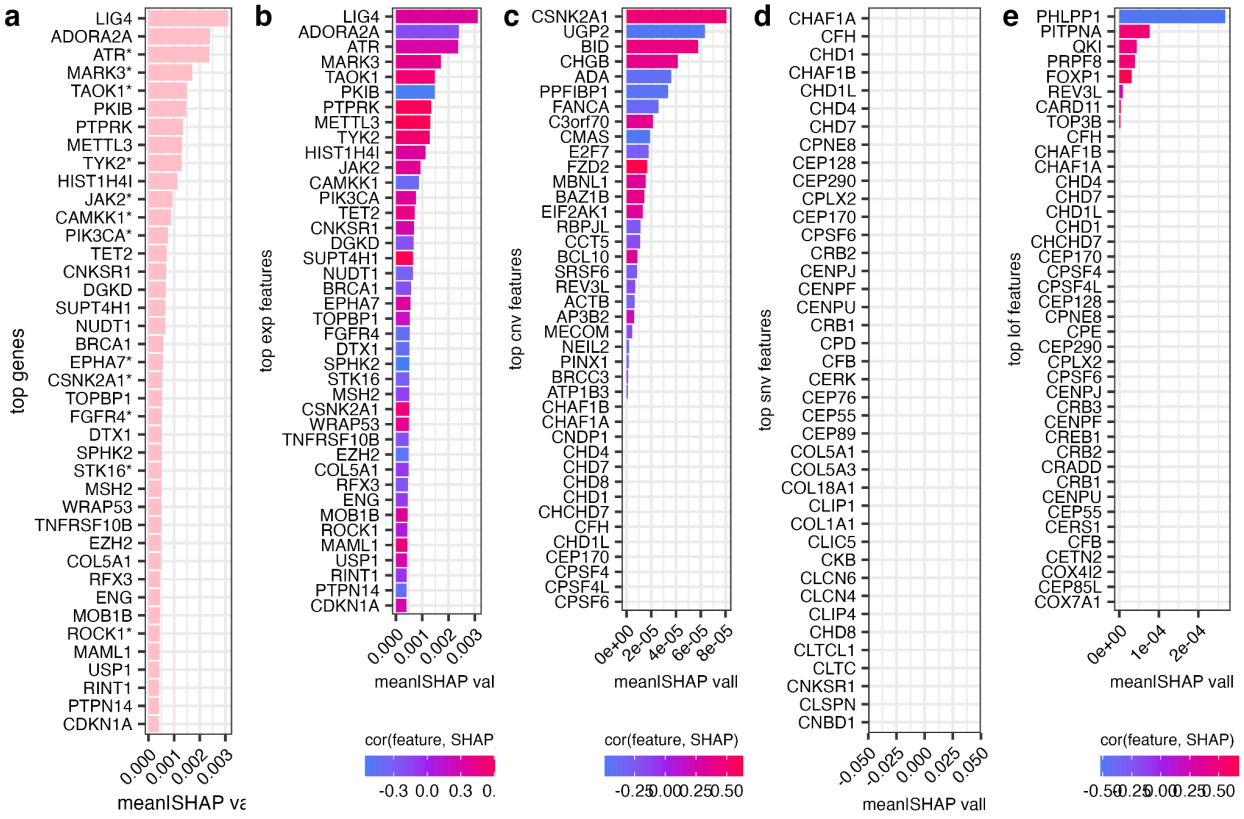
Supplementary Figure 5.12. Top DNA damage response readouts ("*ddr*") biomarkers for efficacy prediction. (a) and (b) show top ddr features in **(a)** efficacy and **(b)** synergy prediction, respectively.



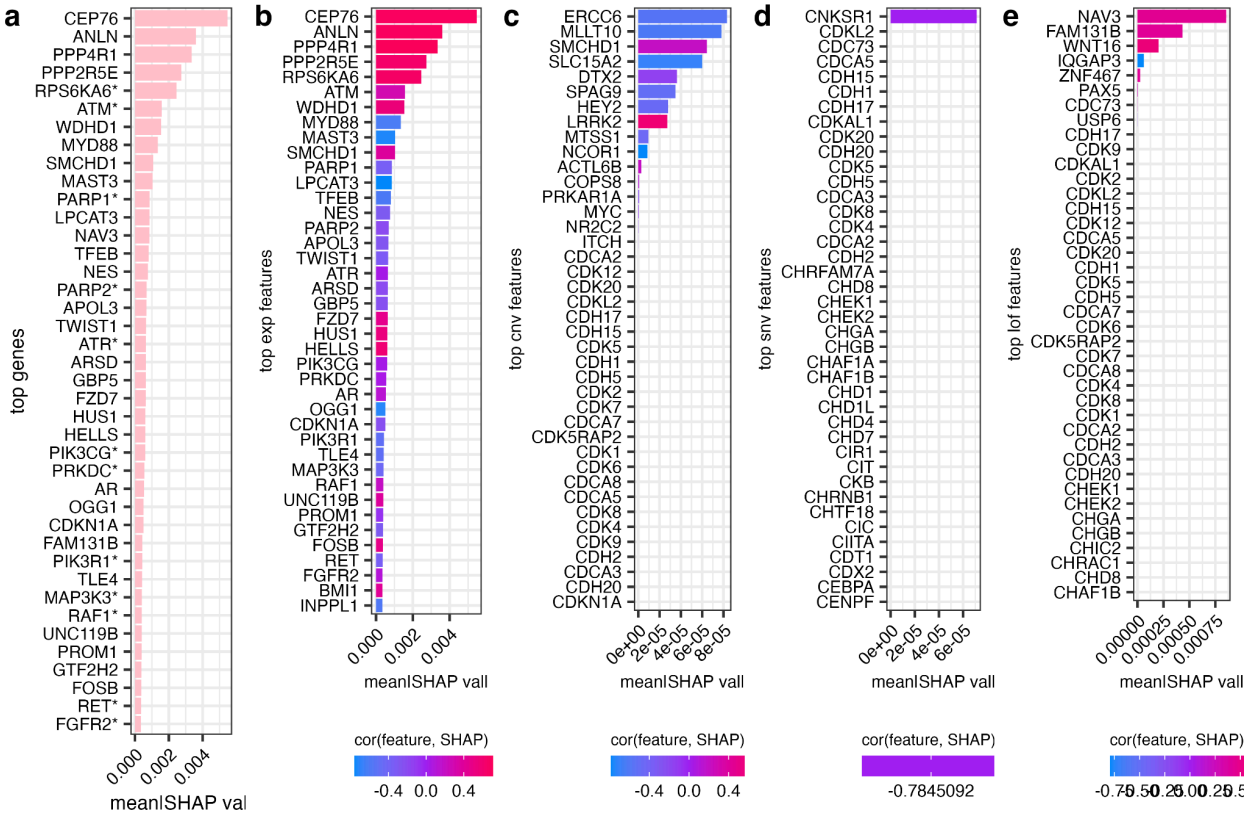
Supplementary Figure 5.13. Top geneset annotation features for drug target encrichment for in DDR drug combination response prediction. (a) and (b) show top geneset annotations in **(a)** efficacy and **(b)** synergy prediction, respectively.



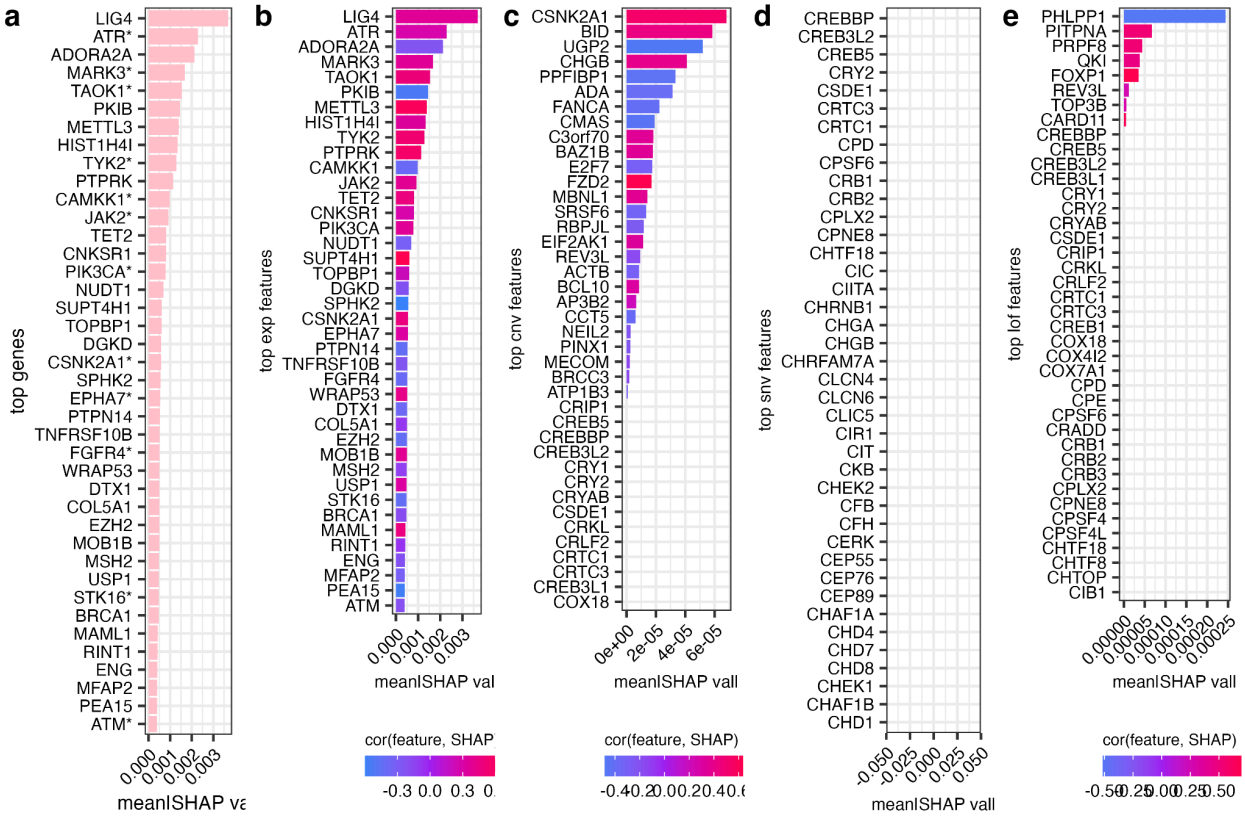
Supplementary Figure 5.14. Top genes in drug efficacy prediction for ATMi-ATRi combination treatments. (a). The top 40 genes after combining all types of molecular markers. **(b), (c), (d) and (e)** show the top genes when considering expression levels ("*exp*"), copy number variation ("*cnv*"), loss-of-function of the gene ("*lof*"), and single nucleotide variation ("*snv*"), respectively. *: direct drug targets.



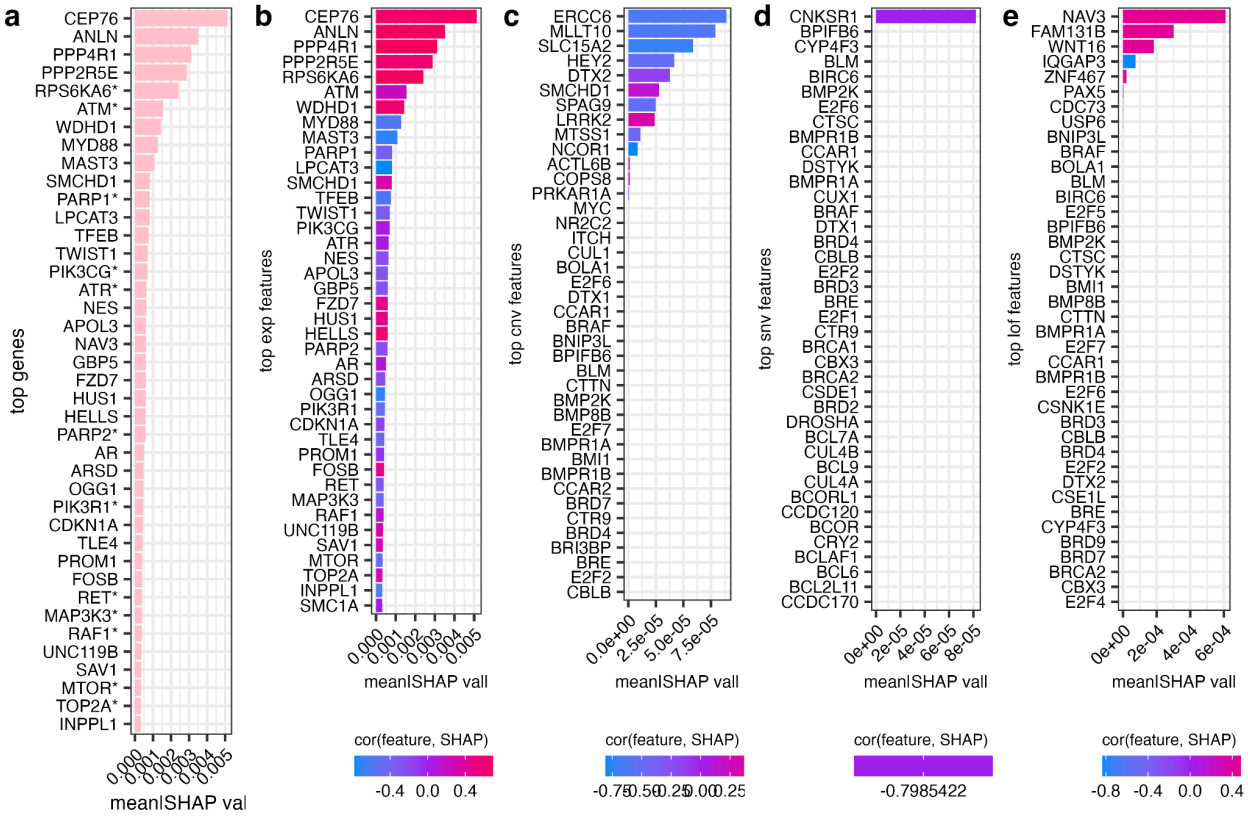
Supplementary Figure 5.15. Top genes in drug synergy prediction for ATMi-ATRi combination treatments. (a). The top 40 genes after combining all types of molecular markers. **(b), (c), (d) and (e)** show the top genes when considering expression levels ("*exp*"), copy number variation ("*cnv*"), loss-of-function of the gene ("*lof*"), and single nucleotide variation ("*snv*"), respectively. *: direct drug targets.



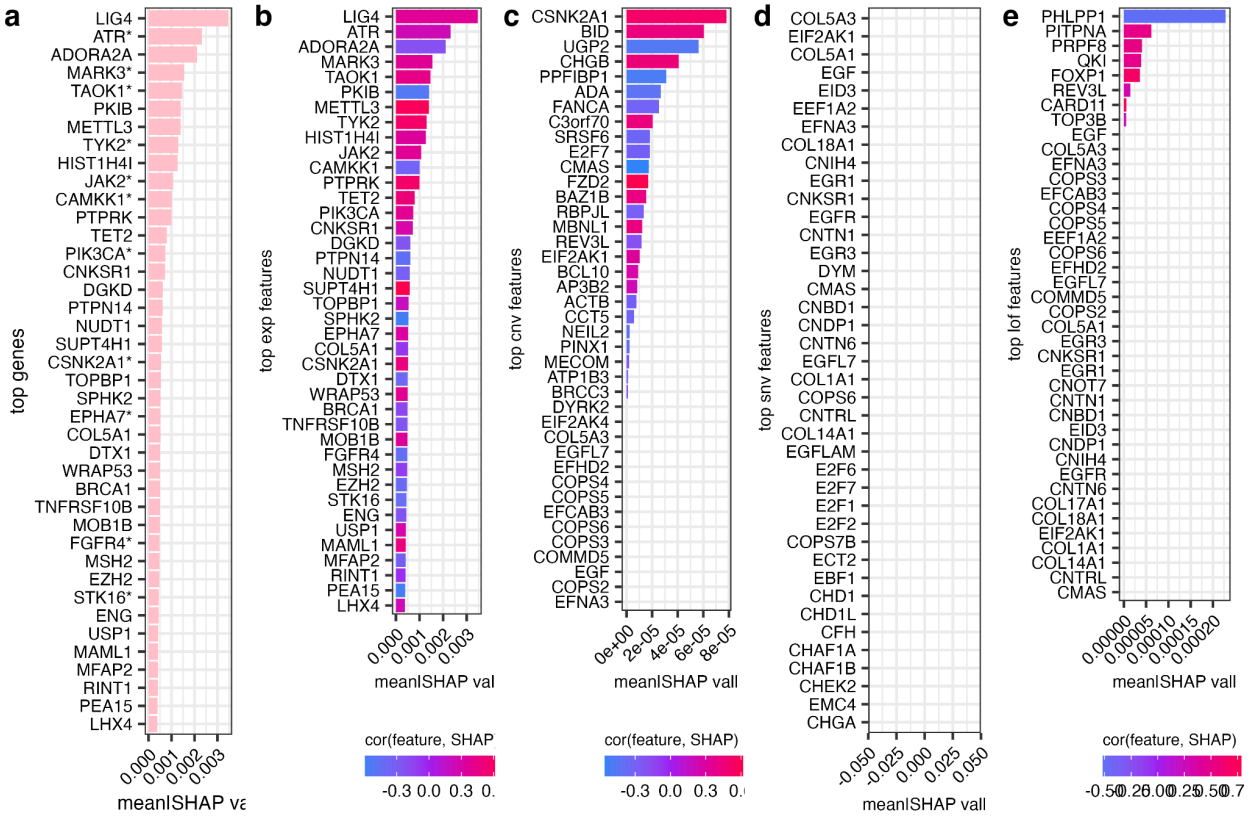
Supplementary Figure 5.16. Top genes in drug efficacy prediction for ATRi-PARPi combination treatments. (a). The top 40 genes after combining all types of molecular markers. **(b), (c), (d) and (e)** show the top genes when considering expression levels ("*exp*"), copy number variation ("*cnv*"), loss-of-function of the gene ("*lof*"), and single nucleotide variation ("*snv*"), respectively. *: direct drug targets.



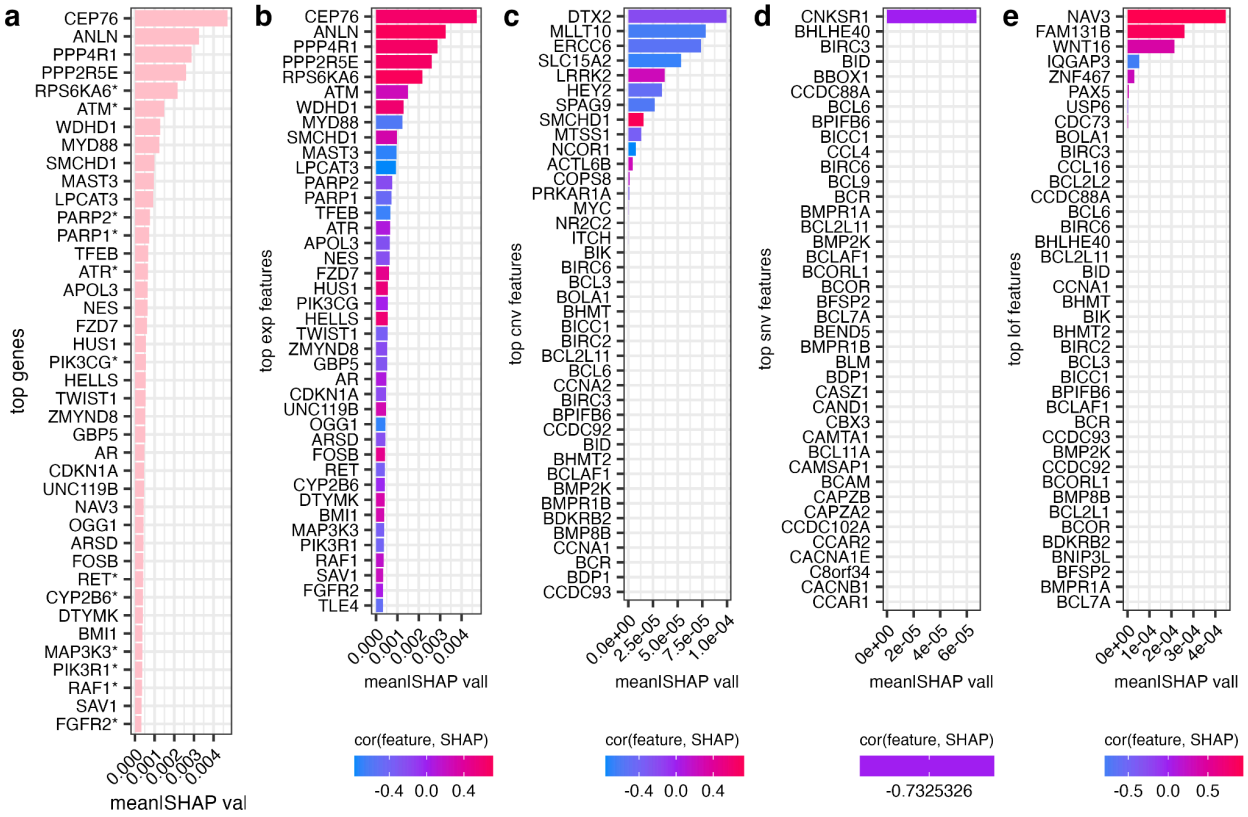
Supplementary Figure 5.17. Top genes in drug synergy prediction for ATRi-PARPi combination treatments. (a). The top 40 genes after combining all types of molecular markers. **(b), (c), (d), and (e)** show the top genes when considering expression levels ("*exp*"), copy number variation ("*cnv*"), loss-of-function of the gene ("*lof*"), and single nucleotide variation ("*snv*"), respectively. *: direct drug targets.



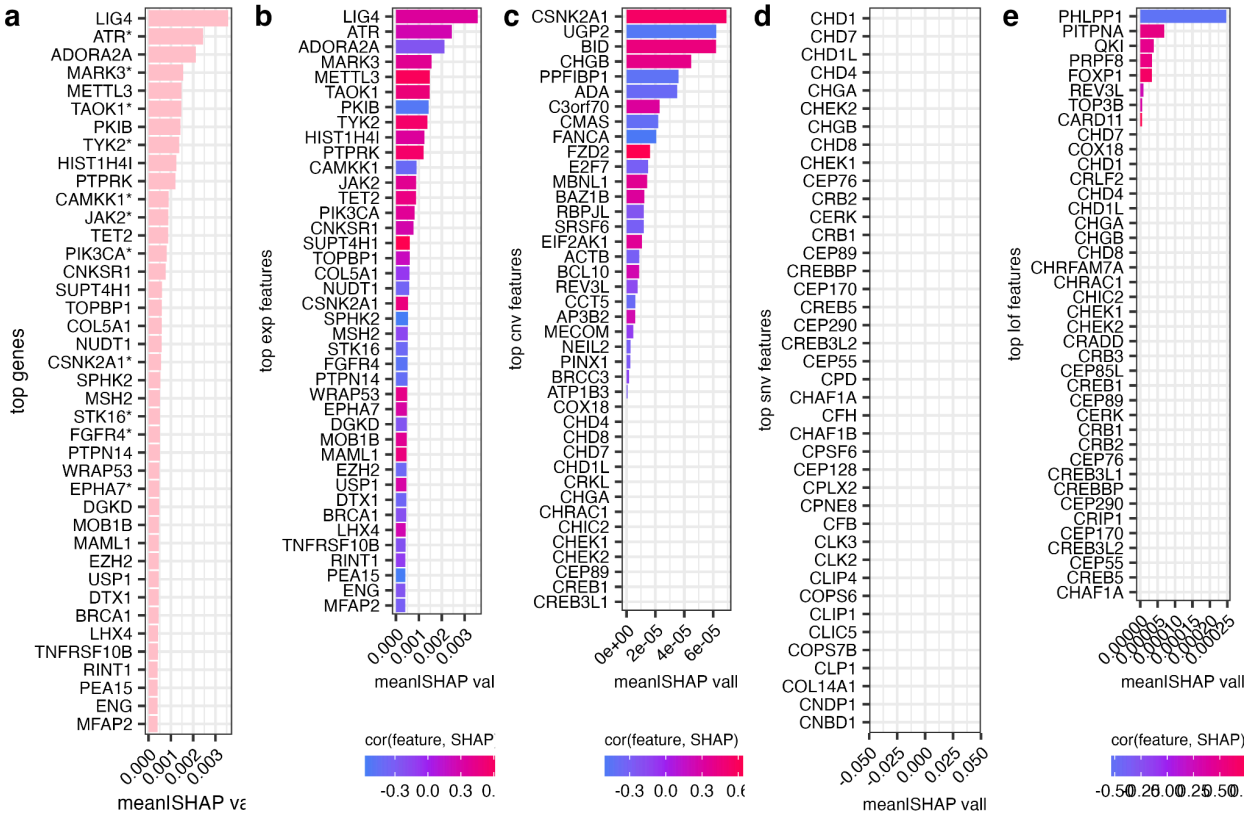
Supplementary Figure 5.18. Top genes in drug efficacy prediction for ATRi-TOP1i combination treatments. (a). The top 40 genes after combining all types of molecular markers. **(b), (c), (d), and (e)** show the top genes when considering expression levels ("*exp*"), copy number variation ("*cnv*"), loss-of-function of the gene ("*lof*"), and single nucleotide variation ("*snv*"), respectively. *: direct drug targets.



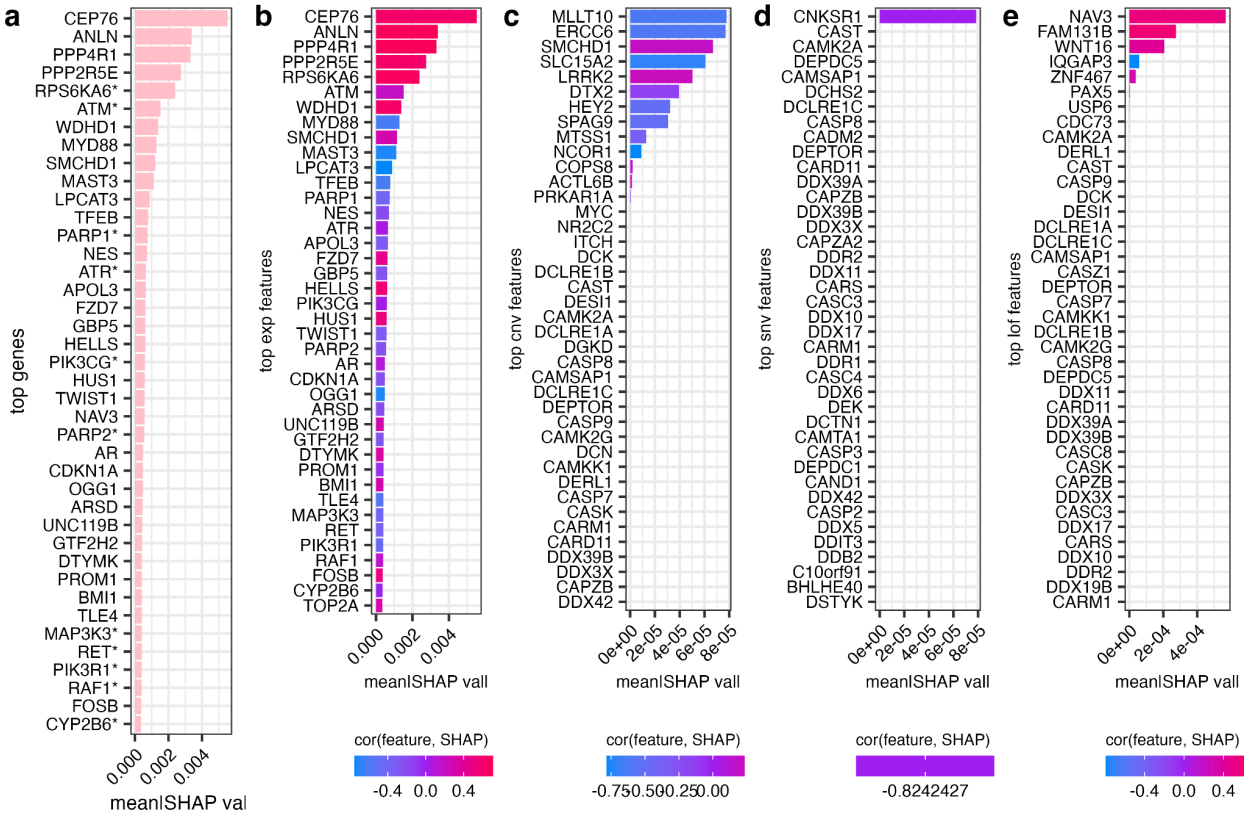
Supplementary Figure 5.19. Top genes in drug synergy prediction for ATRi-TOP1i combination treatments. (a). The top 40 genes after combining all types of molecular markers. **(b), (c), (d) and (e)** show the top genes when considering expression levels ("*exp*"), copy number variation ("*cnv*"), loss-of-function of the gene ("*lof*"), and single nucleotide variation ("*snv*"), respectively. *: direct drug targets.



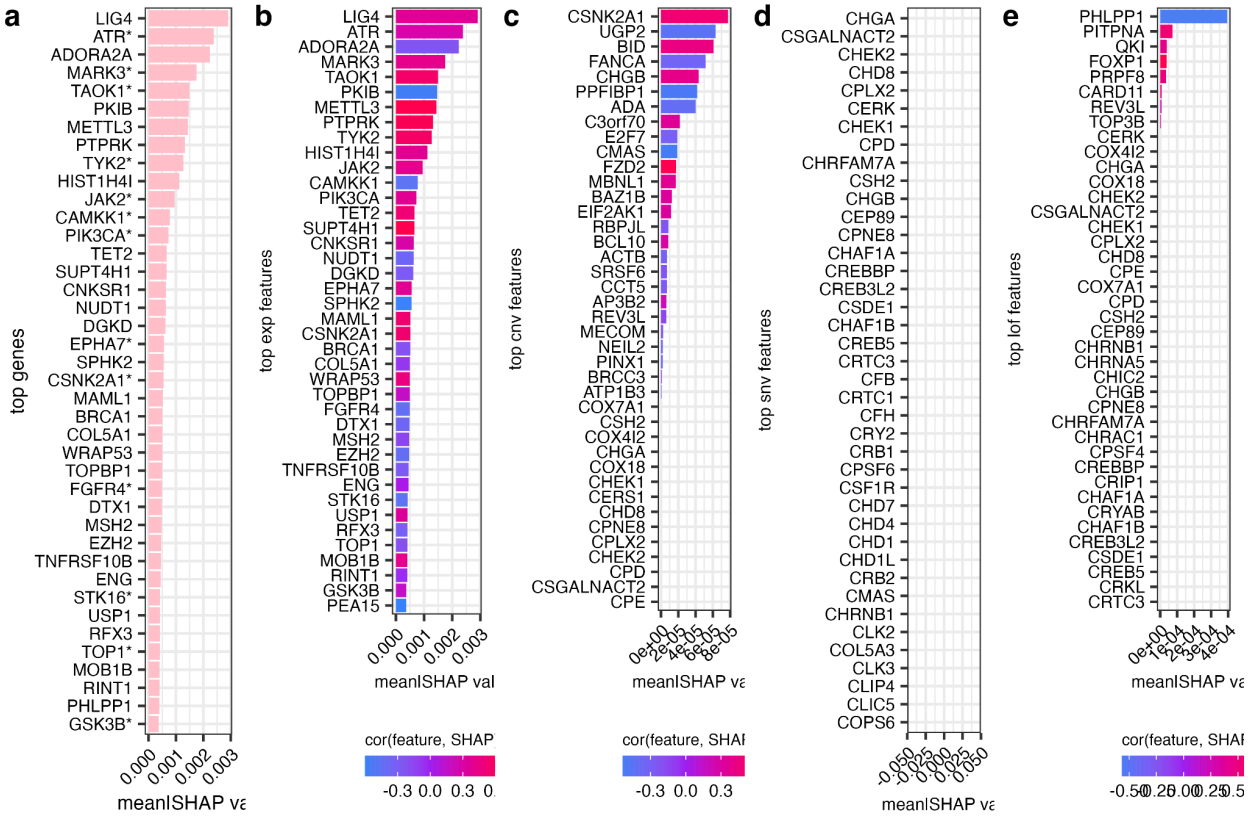
Supplementary Figure 5.20. Top genes in drug efficacy prediction for ATRi-Cytostatic Antimetabolite combination treatments. (a). The top 40 genes after combining all types of molecular markers. **(b), (c), (d) and (e)** show the top genes when considering expression levels ("*exp*"), copy number variation ("*cnv*"), loss-of-function of the gene ("*lof*"), and single nucleotide variation ("*snv*"), respectively. *: direct drug targets.



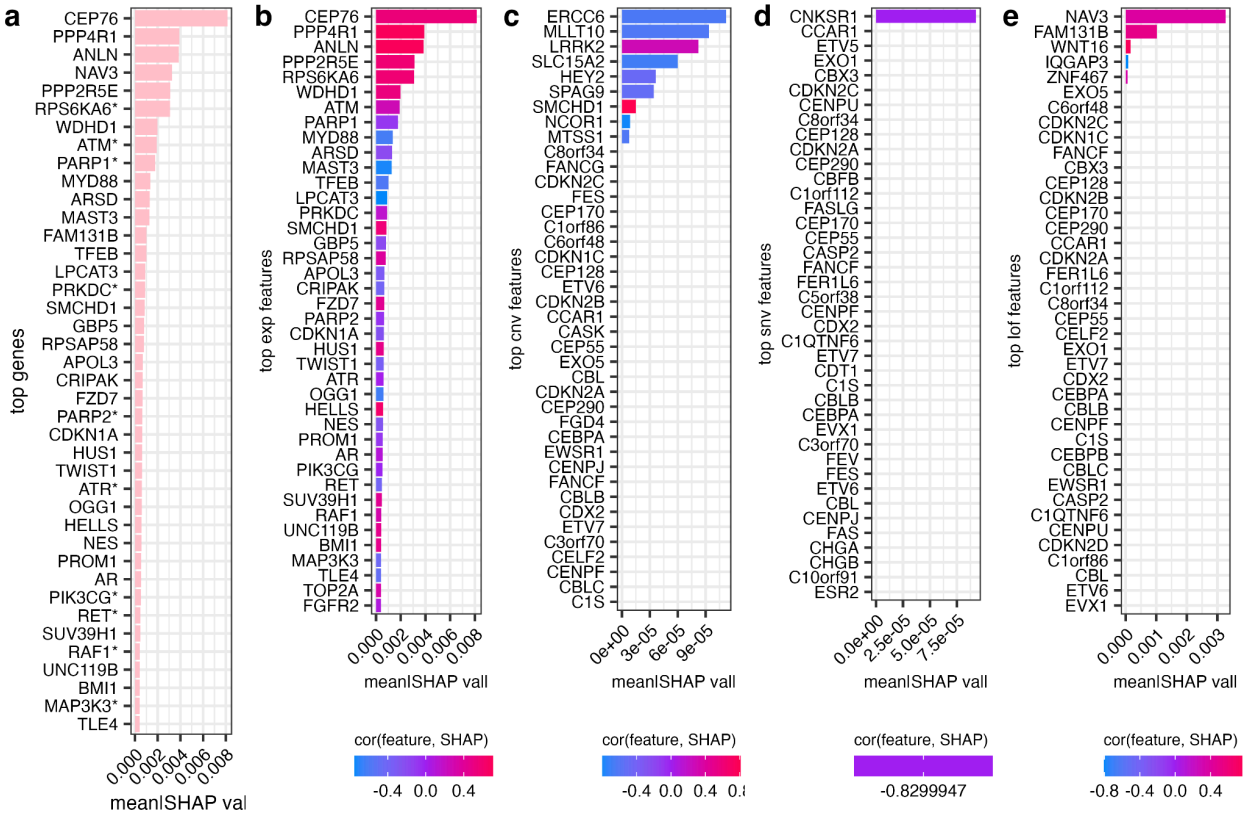
Supplementary Figure 5.21. Top genes in drug synergy prediction for ATRi-Cytostatic Antimetabolite combination treatments. (a). The top 40 genes after combining all types of molecular markers. **(b), (c), (d) and (e)** show the top genes when considering expression levels ("*exp*"), copy number variation ("*cnv*"), loss-of-function of the gene ("*lof*"), and single nucleotide variation ("*snv*"), respectively. *: direct drug targets.



Supplementary Figure 5.22. Top genes in drug efficacy prediction for DNA-PKi-IR combination treatments. (a). The top 40 genes after combining all types of molecular markers. **(b), (c), (d) and (e)** show the top genes when considering expression levels ("*exp*"), copy number variation ("*cnv*"), loss-of-function of the gene ("*lof*"), and single nucleotide variation ("*snv*"), respectively. *: direct drug targets.



Supplementary Figure 5.23. Top genes in drug synergy prediction for DNA-PKi-IR combination treatments. (a). The top 40 genes after combining all types of molecular markers. **(b), (c), (d) and (e)** show the top genes when considering expression levels ("*exp*"), copy number variation ("*cnv*"), loss-of-function of the gene ("*lof*"), and single nucleotide variation ("*snv*"), respectively. *: direct drug targets.



CHAPTER VI: Summary, Conclusions, and Future Works

Summary and Conclusions

In recent years, the pharmaceutical sector has witnessed a significant increase in the application of machine learning (ML) technologies. This trend is evident in the growing number of submissions to the FDA that incorporate AI/ML methodologies in drug and biologic product development (Center for Drug Evaluation & Research, 2023). These submissions cover a wide range of drug development activities, including the identification of novel drug-target interactions, refinement of treatment modalities, augmentation of drug safety measures, mitigation of drug toxicity, and enhancements in both drug manufacturing processes and the personalization of treatment plans.

Central to these advancements is the synergy between ML models and their training data. The practical effectiveness of these models heavily relies on this interplay, particularly given the diverse and complex nature of datasets derived from experimental procedures. These datasets, which include single-cell RNA sequencing, bulk RNA sequencing, medical imaging, and high-throughput screenings, often contain confounding factors such as batch effects, experimental artifacts, equipment variability, and biological variances among samples. To address these issues, various methodologies have been developed and implemented to detect, adjust, or harmonize batch effects within training datasets, thus improving the performance of ML models (Hu et al., 2023; X. Li et al., 2020; Trabucco et al., 18--24 Jul 2021).

Additionally, the potential of ML models to process real-world data (RWD) — information obtained from healthcare settings beyond traditional, controlled clinical research, like electronic health records (EHR) and medical claims (Sherman et al., 2016) — is gaining more attention. This shift is crucial for translating ML models into clinical practice and real-world scenarios involving end-users. The unstructured nature of RWD, characterized by variability in data sources, quality, formats, and collection methodologies (Bakouny & Patt, 2021), underscores the need for ML models to demonstrate robust generalizability and reliability. Consequently, regulatory bodies like the FDA and the pharmaceutical industry are increasingly requiring ML models to be adaptable and reliable across a broad spectrum of real-world applications.

In Chapters II and III, I delve into solutions for the generalizability issues of ML models in various treatment-prediction scenarios. Chapter II addresses the crucial question of batch effect in *in vitro* high-throughput drug screening. The reproducibility between four publicly available benchmark high-throughput screening (HTS) combination treatment datasets was calculated, revealing a significant drop in inter-dataset reproducibility, from 0.3~0.9 to 0.09~0.2 Pearson's correlation (**Supplementary Figure 2.1**). The limited shared information between datasets, such as cell lines and drugs, posed significant challenges for cross-dataset prediction (**Supplementary Table 2.1**). However, after optimizing the ML model by incorporating the imputed dose-response relationship into the features, the treatment response prediction performances improved significantly, achieving 0.6~0.8 Pearson's correlation (**Supplementary Table 2.3**). This study demonstrates that major measurement artifacts, often considered inevitable due to technical setting variations, can be rectified in the downstream in silico analysis using ML model predictions. Notably, the shape of the dose-response curve, including the start and end points, and slope, proved highly informative for predicting combination treatment responses. This could be used to predict responses to unseen drugs or those with unknown modes of action, a

phenomenon previously noted but not comprehensively analyzed (B. Yadav et al., 2015). The ML model developed in this study effectively processed and utilized this characteristic by abstracting knowledge from the extensive training data.

Chapter III explores the generalizability of ML models in treatment response prediction between real-world data and laboratory tests. This work was part of the Malaria DREAM Challenge, which sought computational solutions to the rapidly growing resistance to artemisinin in *Plasmodium falciparum*, the malaria pathogen, observed in Southeast Asia and Africa (Achan et al., 2011; Miller & Su, 2011). Artemisinin, a transformative antimalarial drug replacing quinine, and artemisinin-based combination treatments (ACT) have historically shown over 95% efficacy in clinics (Nosten & White, 2007). However, artemisinin resistance has increasingly been observed worldwide, with unclear mechanisms (Ariey et al., 2014; Mok et al., 2015; L. Zhu et al., 2022). The challenge's subchallenge 2 required participants to train an ML model using transcriptomes of 1,043 isolates collected from population blood samples labeled by *in vivo* artemisinin resistance (clearance rate) and predict the resistance of 32 laboratory-cultured isolates measured by IC50. Technical inconsistencies due to different microarray platforms, differences between clearance rate and IC50 as measurements for *in vivo* or *in vitro* resistance, and inherent biological gaps between *in vivo* and *in vitro* experiments were significant obstacles. The transferability of primary screening results from preclinical research to clinical trials remains a major issue; only about one in 10,000 candidates from the compound screening stage passes the third phase of clinical trials (Sun et al., 2022). The true drug-target interactions *in vivo* can differ from *in vitro* experiments, leading to off-target effects (Moffat et al., 2017). Similarly, a drug's pharmacological effects, such as toxicity, can be affected by unknown molecular targets *in vivo* (Lin et al., 2019). Cell lines and animal models, although established to mimic human disease conditions, cannot perfectly recapitulate disease phenotypes or pathophysiology (*New*

Approaches to Drug Discovery, n.d.). In our case of malaria ART-resistance prediction, the transcriptomic patterns of P. falciparum strains collected from blood samples and those cultured in laboratory conditions differed significantly, influenced by factors including culture temperature, the presence of the human immune system, and the pathogen's developmental stage. Our model achieved first-place performance in the Malaria DREAM Challenge but also highlighted the challenges of applying real-world knowledge to controlled laboratory systems. The differences between *in vivo* and *in vitro* systems resulted in distinct co-expression gene set patterns and top biomarkers for these conditions (**Figure 3.5**). We also discussed the transferability between *in vivo* and *ex vivo* states, an intermediate state between *in vivo* and *in vitro*, showing that models based on *in vivo* data have greater predictive ability for ART resistance in *ex vivo P. falciparum* strains than *in vitro* conditions. The prediction performance declined as the *ex vivo* culture time exceeded 10 hours (**Supplementary Figures 3.4 and 3.5**). This finding suggests that the stress response of *P. falciparum* to laboratory environments could be a significant impediment to studying ART resistance mechanisms, despite the necessity of controlled environments for quantification.

In Chapters IV and V, I showcase the role of machine learning in developing novel treatments and drug repurposing. DNA damage response (DDR) targeted treatments have garnered immense interest as a promising future direction in cancer treatment, supported by extensive mode-of-action and preclinical studies, and some successful FDA-approved clinical treatments like PARP inhibitors (Bryant et al., 2005; O'Connor, 2015). However, large-scale screening and target identification for these treatments are still limited. Through industrial collaboration with Merck, we initiated a project to comprehensively analyze DDR-targeted treatments with combinations of two different anticancer drugs. We selected a treatment strategy

using one DDR-targeted treatment (ATR, ATM, or DNAPK inhibitor) as the backbone drug in combination with another anticancer drug, targeting multiple factors in the interlocked DDR pathways to induce synthetic lethality in cancer cells. This resulted in an *in vitro* high-throughput screening of 17,912 experiments, covering approximately 450 different treatment combinations on 62 cell lines. The most efficacious/synergistic treatments were selected from the direct screening output, and the cancer-type variations of these treatments were also analyzed (**Figures 4.2 and 4.3**). Since most drugs used in the screening dataset were already on the market and had known modes of action and target interactions, we extracted drug target information from external sources, including DrugBank (Knox et al., 2011), CHEMBL (Mendez et al., 2019), and LINCS (*HMS LINCS Project*, n.d.). We identified targets that achieved the highest efficacy and synergy when simultaneously targeted with core DDR sensing kinases (**Figure 4.2c**). We then built a machine learning model to predict DDR-targeted combination therapy using the above HTS dataset as the training set. Other information, such as transcriptomic and genomic profiles of the treated cell lines, chemical structure information, tissue-specific networks, and synthetic lethality, was integrated into the model. We adapted our first-place method from the AstraZeneca-Sanger DREAM Challenge, which integrated drug-target information and network propagation into molecular profiles to simulate post-treatment transcriptomic and epigenomic profiles (H. Li et al., 2018; Menden et al., 2019). Our model was tested across different cancer types and cell lines from the training set during cross-validation and further validated on hold-out validation datasets with different cell lines and cancer types (**Figure 5.2**). It achieved accuracy on par with experimental replicability and demonstrated clinical potential by optimizing treatments for all cancer types in this study. An AI interpretation method enabled our machine learning model to identify the genomic/transcriptomic phenotypes of cancer cells relevant for treatment selection for DDR-targeted combination therapy, and top biomarkers were identified

(**Figures 5.3 and 5.4**). We also showed that using a minimal gene panel of approximately 40 genes could further optimize the machine learning model's performance and alleviate potential overfitting. This demonstrates further applications of machine learning in the industry to accelerate the screening and development process of DDR-targeted combination treatments (**Figure 5.5**)

Future works

In the dynamic realm of pharmaceutical research, the advent of machine learning (ML) and artificial intelligence (AI) is catalyzing a paradigm shift, heralding an era marked by unprecedented innovation and efficiency. Far from being mere supplements to existing methodologies, these technologies are pivotal in reshaping the industry's approach to drug discovery and development. Particularly noteworthy is their impact in the field of biomarker development, where AI's application in genomics and immunology has expedited progress in targeted therapies and immunotherapy, extending its influence beyond oncology into other medical domains such as cardiology and pulmonology (Subbiah, 2023).

One of the most promising aspects of this technological integration is dimensionality reduction—a technique critical in distilling vast biological datasets into their most informative components. Through advanced methods like manifold learning and autoencoders, researchers are uncovering latent patterns that elucidate complex relationships between genes, proteins, and disease phenotypes, a breakthrough that has implications for novel drug target identification and compound optimization (Moon et al., 2018).

Equally important is the refinement of drug representation—the lexicon through which molecular structures are interpreted by ML models. Here, generative models such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) are instrumental in

proposing potential candidate structures, thus guiding chemists towards more promising leads (Y. Wang et al., 2022). However, the challenge lies in ensuring that these representations accurately reflect the intricate molecular-biological interactions. An integration of quantum mechanics with ML frameworks may offer profound insights into these interactions, potentially revolutionizing the design of tailored pharmaceuticals (Dral, 2022).

Another innovative frontier is meta-learning—an approach that equips models with the capability to learn and generalize across diverse datasets. This adaptability is particularly valuable in streamlining clinical trial designs, enhancing patient response predictions, and addressing the variability inherent in real-world disease presentations (Finn et al., 2017). By leveraging insights from previous trials and integrating real-world evidence from electronic health records and wearable sensors, meta-learning algorithms are poised to refine treatment strategies in real time, optimizing patient-specific outcomes (Tanwar et al., 2020).

In summary, as the pharmaceutical industry confronts a rapidly evolving landscape, the integration of ML and AI stands as a beacon, guiding its progression. Embracing these advanced methodologies will not only unlock a treasure trove of scientific insights but also pave the way for personalized, efficient, and transformative medical innovations—a true testament to the revolutionary impact of machine learning in the realm of pharmaceutical research.

REFERENCES

Abel, F. (n.d.). *ACM RecSys Challenge 2017*. Retrieved October 8, 2020, from http://www.recsyschallenge.com/2017/

Achan, J., Talisuna, A. O., Erhart, A., Yeka, A., Tibenderana, J. K., Baliraine, F. N., Rosenthal, P. J., & D'Alessandro, U. (2011). Quinine, an old anti-malarial drug in a modern world: role in the treatment of malaria. *Malaria Journal*, *10*, 144.

ALASKA2 Image Steganalysis. (n.d.). Retrieved October 8, 2020, from https://www.kaggle.com/c/alaska2-image-steganalysis/overview

Anisenko, A., Kan, M., Shadrina, O., Brattseva, A., & Gottikh, M. (2020). Phosphorylation Targets of DNA-PK and Their Role in HIV-1 Replication. *Cells* , *9*(8). https://doi.org/10.3390/cells9081907

Ariey, F., Witkowski, B., Amaratunga, C., Beghain, J., Langlois, A.-C., Khim, N., Kim, S., Duru, V., Bouchier, C., Ma, L., Lim, P., Leang, R., Duong, S., Sreng, S., Suon, S., Chuor, C. M., Bout, D. M., Ménard, S., Rogers, W. O., … Ménard, D. (2014). A molecular marker of artemisinin-resistant Plasmodium falciparum malaria. *Nature*, *505*(7481), 50–55.

Asenso-Okyere, K., Asante, F. A., Tarekegn, J., & Andam, K. S. (2011). A review of the economic impact of malaria in agricultural development. *Agricultural Economics* , *42*(3), 293–304.

Ashley, E. A., Dhorda, M., Fairhurst, R. M., Amaratunga, C., Lim, P., Suon, S., Sreng, S., Anderson, J. M., Mao, S., Sam, B., Sopha, C., Chuor, C. M., Nguon, C., Sovannaroth, S., Pukrittayakamee, S., Jittamala, P., Chotivanich, K., Chutasmit, K., Suchatsoonthorn, C., … Tracking Resistance to Artemisinin Collaboration (TRAC). (2014). Spread of artemisinin resistance in Plasmodium falciparum malaria. *The New England Journal of Medicine*, *371*(5), 411–423.

Bailey, K. R. (1987). Inter-study differences: how should they influence the interpretation and analysis of

results? *Statistics in Medicine*, *6*(3), 351–360.

Bakouny, Z., & Patt, D. A. (2021). Machine Learning and Real-World Data: More than Just Buzzwords [Review of *Machine Learning and Real-World Data: More than Just Buzzwords*]. *JCO Clinical Cancer Informatics*, *5*, 811–813.

Balmus, G., Pilger, D., Coates, J., Demir, M., Sczaniecka-Clift, M., Barros, A., Woods, M., Fu, B., Yang, F., Chen, E., Ostermaier, M., Stankovic, T., Ponstingl, H., Herzog, M., Yusa, K., Martinez, F. M., Durant, S. T., Galanty, Y., Beli, P., … Jackson, S. P. (n.d.). *ATM orchestrates the DNA-damage response to counter toxic non-homologous end-joining at broken replication forks*. https://doi.org/10.1101/330043

Bansal, M., NCI-DREAM Community, Yang, J., Karan, C., Menden, M. P., Costello, J. C., Tang, H., Xiao, G., Li, Y., Allen, J., Zhong, R., Chen, B., Kim, M., Wang, T., Heiser, L. M., Realubit, R., Mattioli, M., Alvarez, M. J., Shen, Y., … Califano, A. (2014). A community computational challenge to predict the activity of pairs of compounds. In *Nature Biotechnology* (Vol. 32, Issue 12, pp. 1213–1222). https://doi.org/10.1038/nbt.3052

Barnieh, F. M., Loadman, P. M., & Falconer, R. A. (2021). Progress towards a clinically-successful ATR inhibitor for cancer therapy. *Current Research in Pharmacology and Drug Discovery*, *2*, 100017.

Baskar, R., Lee, K. A., Yeo, R., & Yeoh, K.-W. (2012). Cancer and Radiation Therapy: Current Advances and Future Directions. In *International Journal of Medical Sciences* (Vol. 9, Issue 3, pp. 193–199). https://doi.org/10.7150/ijms.3635

Bayat Mokhtari, R., Homayouni, T. S., Baluch, N., Morgatskaya, E., Kumar, S., Das, B., & Yeger, H. (2017). Combination therapy in combating cancer. *Oncotarget*, *8*(23), 38022–38043.

Berenbaum, M. C. (1989). What is synergy? *Pharmacological Reviews*, *41*(2), 93–141.

Bionetworks, S. (n.d.-a). *Synapse*. Retrieved August 18, 2020, from https://www.synapse.org/#!Synapse:syn16924919/wiki/

Bionetworks, S. (n.d.-b). *Synapse | Sage Bionetworks*. Retrieved January 31, 2020, from https://www.synapse.org/#!Synapse:syn16924919/wiki/590948

Blackford, A. N., & Jackson, S. P. (2017). ATM, ATR, and DNA-PK: The Trinity at the Heart of the DNA Damage Response. *Molecular Cell*, *66*(6), 801–817.

Blucher, A. S., & McWeeney, S. K. (2014). Challenges in secondary analysis of high throughput screening data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 114–124.

Boshuizen, J., & Peeper, D. S. (2020). Rational Cancer Treatment Combinations: An Urgent Clinical Need. *Molecular Cell*, *78*(6), 1002–1018.

Brandsma, I., Fleuren, E. D. G., Williamson, C. T., & Lord, C. J. (2017). Directing the use of DDR kinase inhibitors in cancer treatment. *Expert Opinion on Investigational Drugs*, *26*(12), 1341–1355.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527.

Bridgford, J. L., Xie, S. C., Cobbold, S. A., Pasaje, C. F. A., Herrmann, S., Yang, T., Gillett, D. L., Dick, L. R., Ralph, S. A., Dogovski, C., Spillman, N. J., & Tilley, L. (2018). Artemisinin kills malaria parasites by damaging proteins and inhibiting the proteasome. In *Nature Communications* (Vol. 9, Issue 1). https://doi.org/10.1038/s41467-018-06221-1

Bryant, H. E., Schultz, N., Thomas, H. D., Parker, K. M., Flower, D., Lopez, E., Kyle, S., Meuth, M., Curtin, N. J., & Helleday, T. (2005). Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature*, *434*(7035), 913–917.

Buchbinder, D., Smith, M. J., Kawahara, M., Cowan, M. J., Buzby, J. S., & Abraham, R. S. (2018). Application of a radiosensitivity flow assay in a patient with DNA ligase 4 deficiency. *Blood Advances*, *2*(15), 1828–1832.

Bush, K. T., Boichard, A., & Tsigelny, I. F. (2018). In Vitro Elucidation of Drug Combination Synergy in Treatment of Pancreatic Ductal Adenocarcinoma. *Anticancer Research*, *38*(4), 1967–1977.

Cai, C., Wang, S., Xu, Y., Zhang, W., Tang, K., Ouyang, Q., Lai, L., & Pei, J. (2020). Transfer Learning for Drug Discovery. *Journal of Medicinal Chemistry*, *63*(16), 8683–8694.

Calabrese, E. J. (2014). Dose–Response Relationship. In P. Wexler (Ed.), *Encyclopedia of Toxicology (Third Edition)* (pp. 224–226). Academic Press.

Calabrese, E. J. (2016). The Emergence of the Dose–Response Concept in Biology and Medicine. *International Journal of Molecular Sciences*, *17*(12), 2034.

Caraus, I., Alsuwailem, A. A., Nadon, R., & Makarenkov, V. (2015). Detecting and overcoming systematic bias in high-throughput screening technologies: a comprehensive review of practical issues and methodological solutions. *Briefings in Bioinformatics*, *16*(6), 974–986.

Carr, Chiu, Guo, Xu, Lazorchak, & Yu. (n.d.). DNA-PK Inhibitor Peposertib Amplifies Radiation-Induced Inflammatory Micronucleation and Enhances TGFb/PD-L1 Targeted Cancer Immunotherapy. *Scholar.archive.org*. https://scholar.archive.org/work/vsw3isfnmfbwxd7gnkqkvk4o44/access/wayback/https://mcr.aacrjournals.org/content/molcanres/early/2022/01/21/1541-7786.MCR-21-0612.full.pdf

Center for Drug Evaluation, & Research. (2023, May 16). *Artificial intelligence and machine learning (AI/ML) for drug development*. U.S. Food and Drug Administration; FDA. https://www.fda.gov/science-research/science-and-research-special-topics/artificial-intelligence-and-machine-learning-aiml-drug-development

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G. D., & Sander, C. (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, *39*(Database issue), D685–D690.

Chan, G. K. Y., Wilson, S., Schmidt, S., & Moffat, J. G. (2016). Unlocking the Potential of High-Throughput Drug Combination Assays Using Acoustic Dispensing. *Journal of Laboratory Automation*, *21*(1), 125–132.

Chang, Y., Huang, Z., Quan, H., Li, H., Yang, S., Song, Y., Wang, J., Yuan, J., & Wu, C. (2022). Construction of a DNA damage repair gene signature for predicting prognosis and immune response in breast cancer. *Frontiers in Oncology*, *12*, 1085632.

Chapman, B., Kirchner, R., Pantano, L., Naumenko, S., De Smet, M., Beltrame, L., Khotiainsteva, T., Saveliev, V., Sytchev, I., Guimera, R. V., Kern, J., Brueffer, C., Carrasco, G., Giovacchini, M., Ahdesmaki, M., Tang, P., Kanwal, S., Porter, J. J., Le, V., … Turner, S. (2020). *bcbio/bcbio-nextgen:*

v1.2.4. https://doi.org/10.5281/zenodo.4041990

Chavchich, M., Gerena, L., Peters, J., Chen, N., Cheng, Q., & Kyle, D. E. (2010). Role of pfmdr1 amplification and expression in induction of resistance to artemisinin derivatives in Plasmodium falciparum. *Antimicrobial Agents and Chemotherapy*, *54*(6), 2455–2464.

Cheeseman, I. H., Miller, B. A., Nair, S., Nkhoma, S., Tan, A., Tan, J. C., Al Saai, S., Phyo, A. P., Moo, C. L., Lwin, K. M., McGready, R., Ashley, E., Imwong, M., Stepniewska, K., Yi, P., Dondorp, A. M., Mayxay, M., Newton, P. N., White, N. J., … Anderson, T. J. C. (2012). A major genome region underlying artemisinin resistance in malaria. *Science*, *336*(6077), 79–82.

Cheng, D. T., Mitchell, T. N., Zehir, A., Shah, R. H., Benayed, R., Syed, A., Chandramohan, R., Liu, Z. Y., Won, H. H., Scott, S. N., Brannon, A. R., O'Reilly, C., Sadowska, J., Casanova, J., Yannes, A., Hechtman, J. F., Yao, J., Song, W., Ross, D. S., … Berger, M. F. (2015). Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *The Journal of Molecular Diagnostics: JMD*, *17*(3), 251–264.

Chotivanich, K., Tripura, R., Das, D., Yi, P., Day, N. P. J., Pukrittayakamee, S., Chuor, C. M., Socheat, D., Dondorp, A. M., & White, N. J. (2014). Laboratory Detection of Artemisinin-Resistant Plasmodium falciparum. In *Antimicrobial Agents and Chemotherapy* (Vol. 58, Issue 6, pp. 3157–3161). https://doi.org/10.1128/aac.01924-13

Ciccia, A., & Elledge, S. J. (2010). The DNA damage response: making it safe to play with knives. *Molecular Cell*, *40*(2), 179–204.

Cohen, S. M., & Lippard, S. J. (2001). Cisplatin: From DNA damage to cancer chemotherapy. In *Progress in Nucleic Acid Research and Molecular Biology* (pp. 93–130). https://doi.org/10.1016/s0079-6603(01)67026-0

Coleman, R. L., Fleming, G. F., Brady, M. F., Swisher, E. M., Steffensen, K. D., Friedlander, M., Okamoto, A., Moore, K. N., Efrat Ben-Baruch, N., Werner, T. L., Cloven, N. G., Oaknin, A., DiSilvestro, P. A., Morgan, M. A., Nam, J.-H., Leath, C. A., 3rd, Nicum, S., Hagemann, A. R.,

Littell, R. D., … Bookman, M. A. (2019). Veliparib with First-Line Chemotherapy and as Maintenance Therapy in Ovarian Cancer. *The New England Journal of Medicine*, *381*(25), 2403–2415.

Conn, J. E., Grillet, M. E., Correa, M., & Sallum, M. A. M. (2018). Malaria Transmission in South America—Present Status and Prospects for Elimination. In *Towards Malaria Elimination - A Leap Forward*. https://doi.org/10.5772/intechopen.76964

Costello, J. C., NCI DREAM Community, Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., Bansal, M., Ammad-ud-din, M., Hintsanen, P., Khan, S. A., Mpindi, J.-P., Kallioniemi, O., Honkela, A., Aittokallio, T., Wennerberg, K., Collins, J. J., Gallahan, D., Singer, D., … Stolovitzky, G. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. In *Nature Biotechnology* (Vol. 32, Issue 12, pp. 1202–1212). https://doi.org/10.1038/nbt.2877

Cotto, K. C., Wagner, A. H., Feng, Y.-Y., Kiwala, S., Coffman, A. C., Spies, G., Wollam, A., Spies, N. C., Griffith, O. L., & Griffith, M. (2018). DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. *Nucleic Acids Research*, *46*(D1), D1068–D1073.

Dahlström, S., Ferreira, P. E., Veiga, M. I., Sedighi, N., Wiklund, L., Mårtensson, A., Färnert, A., Sisowath, C., Osório, L., Darban, H., Andersson, B., Kaneko, A., Conseil, G., Björkman, A., & Gil, J. P. (2009). Plasmodium falciparum multidrug resistance protein 1 and artemisinin-based combination therapy in Africa. *The Journal of Infectious Diseases*, *200*(9), 1456–1464.

DeLean, A., Munson, P. J., & Rodbard, D. (1978). Simultaneous analysis of families of sigmoidal curves: application to bioassay, radioligand assay, and physiological dose-response curves. *The American Journal of Physiology*, *235*(2), E97–E102.

Dhiman, S. (2019). Are malaria elimination efforts on right track? An analysis of gains achieved and challenges ahead. *Infectious Diseases of Poverty*, *8*(1), 14.

DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*, *47*, 20–33.

Ding, K.-F., Finlay, D., Yin, H., Hendricks, W. P. D., Sereduk, C., Kiefer, J., Sekulic, A., LoRusso, P. M.,

Vuori, K., Trent, J. M., & Schork, N. J. (2017). Analysis of variability in high throughput screening data: applications to melanoma cell lines and drug responses. *Oncotarget*, *8*(17), 27786–27799.

Dogovski, C., Xie, S. C., Burgio, G., Bridgford, J., Mok, S., McCaw, J. M., Chotivanich, K., Kenny, S., Gnädig, N., Straimer, J., Bozdech, Z., Fidock, D. A., Simpson, J. A., Dondorp, A. M., Foote, S., Klonis, N., & Tilley, L. (2015). Targeting the cell stress response of Plasmodium falciparum to overcome artemisinin resistance. *PLoS Biology*, *13*(4), e1002132.

Dondorp, A. M., Nosten, F., Yi, P., Das, D., Phyo, A. P., Tarning, J., Lwin, K. M., Ariey, F., Hanpithakpong, W., Lee, S. J., Ringwald, P., Silamut, K., Imwong, M., Chotivanich, K., Lim, P., Herdman, T., An, S. S., Yeung, S., Singhasivanon, P., … White, N. J. (2009). Artemisinin resistance in Plasmodium falciparum malaria. *The New England Journal of Medicine*, *361*(5), 455–467.

Dral, P. O. (2022). *Quantum Chemistry in the Age of Machine Learning*. Elsevier.

Dunlop, C. R., Wallez, Y., Johnson, T. I., Bernaldo de Quirós Fernández, S., Durant, S. T., Cadogan, E. B., Lau, A., Richards, F. M., & Jodrell, D. I. (2020). Complete loss of ATM function augments replication catastrophe induced by ATR inhibition and gemcitabine in pancreatic cancer models. *British Journal of Cancer*, *123*(9), 1424–1436.

Eastman, R. T., Khine, P., Huang, R., Thomas, C. J., & Su, X.-Z. (2016). PfCRT and PfMDR1 modulate interactions of artemisinin derivatives and ion channel blockers. *Scientific Reports*, *6*, 25379.

Eichler, H.-G., Oye, K., Baird, L. G., Abadie, E., Brown, J., Drum, C. L., Ferguson, J., Garner, S., Honig, P., Hukkelhoven, M., Lim, J. C. W., Lim, R., Lumpkin, M. M., Neil, G., O'Rourke, B., Pezalla, E., Shoda, D., Seyfert-Margolis, V., Sigal, E. V., … Hirsch, G. (2012). Adaptive licensing: taking the next step in the evolution of drug approval. *Clinical Pharmacology and Therapeutics*, *91*(3), 426–437.

Fact sheet about Malaria. (n.d.). Retrieved February 10, 2020, from https://www.who.int/news-room/fact-sheets/detail/malaria

Fairhurst, R. M., & Dondorp, A. M. (2016). Artemisinin-Resistant Plasmodium falciparum Malaria. *Microbiology Spectrum*, *4*(3). https://doi.org/10.1128/microbiolspec.EI10-0013-2016

Fan, K., Cheng, L., & Li, L. (2021). Artificial intelligence and machine learning methods in predicting anti-cancer drug combination effects. *Briefings in Bioinformatics*, *22*(6). https://doi.org/10.1093/bib/bbab271

Farmer, H., McCabe, N., Lord, C. J., Tutt, A. N. J., Johnson, D. A., Richardson, T. B., Santarosa, M., Dillon, K. J., Hickson, I., Knights, C., Martin, N. M. B., Jackson, S. P., Smith, G. C. M., & Ashworth, A. (2005). Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature*, *434*(7035), 917–921.

Felgentreff, K., Baxi, S. N., Lee, Y. N., Dobbs, K., Henderson, L. A., Csomos, K., Tsitsikov, E. N., Armanios, M., Walter, J. E., & Notarangelo, L. D. (2016). Ligase-4 Deficiency Causes Distinctive Immune Abnormalities in Asymptomatic Individuals. *Journal of Clinical Immunology*, *36*(4), 341–353.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1703.03400

Fok, J. H. L., Ramos-Montoya, A., Vazquez-Chantada, M., Wijnhoven, P. W. G., Follia, V., James, N., Farrington, P. M., Karmokar, A., Willis, S. E., Cairns, J., Nikkilä, J., Beattie, D., Lamont, G. M., Finlay, M. R. V., Wilson, J., Smith, A., O'Connor, L. O., Ling, S., Fawell, S. E., … Cadogan, E. B. (2019). AZD7648 is a potent and selective DNA-PK inhibitor that enhances radiation, chemotherapy and olaparib activity. *Nature Communications*, *10*(1), 5065.

Forcina, G. C., Conlon, M., Wells, A., Cao, J. Y., & Dixon, S. J. (2017). Systematic Quantification of Population Cell Death Kinetics in Mammalian Cells. *Cell Systems*, *4*(6), 600–610.e6.

Ford, C. T., & Janies, D. (2020). Ensemble machine learning modeling for the prediction of artemisinin resistance in malaria. In *F1000Research* (Vol. 9, p. 62). https://doi.org/10.12688/f1000research.21539.1

Fowler, H., Belot, A., Ellis, L., Maringe, C., Luque-Fernandez, M. A., Njagi, E. N., Navani, N., Sarfati, D., & Rachet, B. (2020). Comorbidity prevalence among cancer patients: a population-based cohort study of four cancers. *BMC Cancer*, *20*(1), 2.

Gil, J. P., & Krishna, S. (2017). pfmdr1 (Plasmodium falciparum multidrug drug resistance gene 1): a pivotal factor in malaria resistance to artemisinin combination therapies. *Expert Review of Anti-Infective Therapy*, *15*(6), 527–543.

Gong, F., & Miller, K. M. (2018). Double duty: ZMYND8 in the DNA damage response and cancer. *Cell Cycle* , *17*(4), 414–420.

Gordhandas, S. B., Manning-Geist, B., Henson, C., Iyer, G., Gardner, G. J., Sonoda, Y., Moore, K. N., Aghajanian, C., Chui, M. H., & Grisham, R. N. (2022). Pre-clinical activity of the oral DNA-PK inhibitor, peposertib (M3814), combined with radiation in xenograft models of cervical cancer. *Scientific Reports*, *12*(1), 974.

Greco, W. R., Bravo, G., & Parsons, J. C. (1995). The search for synergy: a critical review from a response surface perspective. *Pharmacological Reviews*, *47*(2), 331–385.

Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., Zhang, R., Hartmann, B. M., Zaslavsky, E., Sealfon, S. C., Chasman, D. I., FitzGerald, G. A., Dolinski, K., Grosser, T., & Troyanskaya, O. G. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, *47*(6), 569–576.

Guan, Y., Gorenshteyn, D., Burmeister, M., Wong, A. K., Schimenti, J. C., Handel, M. A., Bult, C. J., Hibbs, M. A., & Troyanskaya, O. G. (2012). Tissue-specific functional networks for prioritizing phenotype and disease genes. *PLoS Computational Biology*, *8*(9), e1002694.

Gupta, B., Xu, S., Wang, Z., Sun, L., Miao, J., Cui, L., & Yang, Z. (2014). Plasmodium falciparum multidrug resistance protein 1 (pfmrp1) gene and its association with in vitro drug susceptibility of parasite isolates from north-east Myanmar. *The Journal of Antimicrobial Chemotherapy*, *69*(8), 2110–2117.

Güvenç Paltun, B., Kaski, S., & Mamitsuka, H. (2021). Machine learning approaches for drug combination therapies. *Briefings in Bioinformatics*, *22*(6). https://doi.org/10.1093/bib/bbab293

Hafner, M., Heiser, L. M., Williams, E. H., Niepel, M., Wang, N. J., Korkola, J. E., Gray, J. W., & Sorger, P. K. (2017). Quantification of sensitivity and resistance of breast cancer cell lines to anti-cancer

drugs using GR metrics. In *Scientific Data* (Vol. 4, Issue 1). https://doi.org/10.1038/sdata.2017.166

Hall, A. B., Newsome, D., Wang, Y., Boucher, D. M., Eustace, B., Gu, Y., Hare, B., Johnson, M. A., Milton, S., Murphy, C. E., Takemoto, D., Tolman, C., Wood, M., Charlton, P., Charrier, J.-D., Furey, B., Golec, J., Reaper, P. M., & Pollard, J. R. (2014). Potentiation of tumor responses to DNA damaging therapy by the selective ATR inhibitor VX-970. *Oncotarget*, *5*(14), 5674–5685.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646–674.

Hashem, S. (1997). Optimal Linear Combinations of Neural Networks. *Neural Networks: The Official Journal of the International Neural Network Society*, *10*(4), 599–614.

He, L., Kulesskiy, E., Saarela, J., Turunen, L., Wennerberg, K., Aittokallio, T., & Tang, J. (2018). Methods for High-throughput Drug Combination Screening and Synergy Scoring. *Methods in Molecular Biology* , *1711*, 351–398.

Hillier, C., Pardo, M., Yu, L., Bushell, E., Sanderson, T., Metcalf, T., Herd, C., Anar, B., Rayner, J. C., Billker, O., & Choudhary, J. S. (2019). Landscape of the Plasmodium Interactome Reveals Both Conserved and Species-Specific Functionality. *Cell Reports*, *28*(6), 1635–1647.e5.

HMS LINCS Project. (n.d.). Retrieved September 29, 2020, from https://lincs.hms.harvard.edu/db

Holbeck, S. L., Camalier, R., Crowell, J. A., Govindharajulu, J. P., Hollingshead, M., Anderson, L. W., Polley, E., Rubinstein, L., Srivastava, A., Wilsker, D., Collins, J. M., & Doroshow, J. H. (2017). The National Cancer Institute ALMANAC: A Comprehensive Screening Resource for the Detection of Anticancer Drug Pairs with Enhanced Therapeutic Activity. *Cancer Research*, *77*(13), 3564–3576.

Holmgren, G., Gil, J. P., Ferreira, P. M., Veiga, M. I., Obonyo, C. O., & Björkman, A. (2006). Amodiaquine resistant Plasmodium falciparum malaria in vivo is associated with selection of pfcrt 76T and pfmdr1 86Y. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, *6*(4), 309–314.

Holmgren, G., Hamrin, J., Svärd, J., Mårtensson, A., Gil, J. P., & Björkman, A. (2007). Selection of pfmdr1 mutations after amodiaquine monotherapy and amodiaquine plus artemisinin combination therapy in East Africa. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and*

Evolutionary Genetics in Infectious Diseases, *7*(5), 562–569.

Horvath, P., Aulner, N., Bickle, M., Davies, A. M., Nery, E. D., Ebner, D., Montoya, M. C., Östling, P., Pietiäinen, V., Price, L. S., Shorte, S. L., Turcatti, G., von Schantz, C., & Carragher, N. O. (2016). Screening out irrelevant cell-based models of disease. *Nature Reviews. Drug Discovery*, *15*(11), 751–769.

Hu, F., Chen, A. A., Horng, H., Bashyam, V., Davatzikos, C., Alexander-Bloch, A., Li, M., Shou, H., Satterthwaite, T. D., Yu, M., & Shinohara, R. T. (2023). Image harmonization: A review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization. *NeuroImage*, *274*, 120125.

Huh, H. D., Kim, D. H., Jeong, H.-S., & Park, H. W. (2019). Regulation of TEAD Transcription Factors in Cancer Biology. *Cells* , *8*(6). https://doi.org/10.3390/cells8060600

Hunt, P., Afonso, A., Creasey, A., Culleton, R., Sidhu, A. B. S., Logan, J., Valderramos, S. G., McNae, I., Cheesman, S., do Rosario, V., Carter, R., Fidock, D. A., & Cravo, P. (2007). Gene encoding a deubiquitinating enzyme is mutated in artesunate- and chloroquine-resistant rodent malaria parasites. *Molecular Microbiology*, *65*(1), 27–40.

Hunt, P., Martinelli, A., Modrzynska, K., Borges, S., Creasey, A., Rodrigues, L., Beraldi, D., Loewe, L., Fawcett, R., Kumar, S., Thomson, M., Trivedi, U., Otto, T. D., Pain, A., Blaxter, M., & Cravo, P. (2010). Experimental evolution, genetic analysis and genome re-sequencing reveal the mutation conferring artemisinin resistance in an isogenic lineage of malaria parasites. *BMC Genomics*, *11*, 499.

Ianevski, A., Timonen, S., Kononov, A., Aittokallio, T., & Giri, A. K. (2020). SynToxProfiler: An interactive analysis of drug combination synergy, toxicity and efficacy. *PLoS Computational Biology*, *16*(2), e1007604.

IEEE-CIS Fraud Detection. (n.d.). Retrieved October 8, 2020, from https://www.kaggle.com/c/ieee-fraud-detection/overview

iMaterialist Challenge (Fashion) at FGVC5. (n.d.). Retrieved October 8, 2020, from

https://www.kaggle.com/c/imaterialist-challenge-fashion-2018/overview

Imwong, M., Dondorp, A. M., Nosten, F., Yi, P., Mungthin, M., Hanchana, S., Das, D., Phyo, A. P., Lwin, K. M., Pukrittayakamee, S., Lee, S. J., Saisung, S., Koecharoen, K., Nguon, C., Day, N. P. J., Socheat, D., & White, N. J. (2010). Exploring the contribution of candidate genes to artemisinin resistance in Plasmodium falciparum. *Antimicrobial Agents and Chemotherapy*, *54*(7), 2886–2892.

Intharabut, B., Kingston, H. W., Srinamon, K., Ashley, E. A., Imwong, M., Dhorda, M., Woodrow, C., Stepniewska, K., Silamut, K., Day, N. P. J., Dondorp, A. M., White, N. J., & Tracking Resistance to Artemisinin Collaboration. (2019). Artemisinin Resistance and Stage Dependency of Parasite Clearance in Falciparum Malaria. *The Journal of Infectious Diseases*, *219*(9), 1483–1489.

Iyer, G., Wang, A. R., Brennan, S. R., Bourgeois, S., Armstrong, E., Shah, P., & Harari, P. M. (2017). Identification of stable housekeeping genes in response to ionizing radiation in cancer research. *Scientific Reports*, *7*, 43763.

Jackson, S. P., & Bartek, J. (2009). The DNA-damage response in human biology and disease. *Nature*, *461*(7267), 1071–1078.

Jafari, M., Mirzaie, M., Bao, J., Barneh, F., Zheng, S., Eriksson, J., Heckman, C. A., & Tang, J. (2022). Bipartite network models to design combination therapies in acute myeloid leukaemia. *Nature Communications*, *13*(1), 2128.

Jo, U., Senatorov, I. S., Zimmermann, A., Saha, L. K., Murai, Y., Kim, S. H., Rajapakse, V. N., Elloumi, F., Takahashi, N., Schultz, C. W., Thomas, A., Zenke, F. T., & Pommier, Y. (2021). Novel and Highly Potent ATR Inhibitor M4344 Kills Cancer Cells With Replication Stress, and Enhances the Chemotherapeutic Activity of Widely Used DNA Damaging Agents. *Molecular Cancer Therapeutics*, *20*(8), 1431–1441.

Julkunen, H., Cichonska, A., Gautam, P., Szedmak, S., Douat, J., Pahikkala, T., Aittokallio, T., & Rousu, J. (2020). Leveraging multi-way interactions for systematic prediction of pre-clinical drug combination effects. In *Nature Communications* (Vol. 11, Issue 1). https://doi.org/10.1038/s41467-020-19950-z

Kaelin, W. G., Jr. (2005). The concept of synthetic lethality in the context of anticancer therapy. *Nature Reviews. Cancer*, *5*(9), 689–698.

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., & Tanabe, M. (2020). KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkaa970

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443.

Karpiyevich, M., Adjalley, S., Mol, M., Ascher, D. B., Mason, B., van der Heden van Noort, G. J., Laman, H., Ovaa, H., Lee, M. C. S., & Artavanis-Tsakonas, K. (2019). Nedd8 hydrolysis by UCH proteases in Plasmodium parasites. *PLoS Pathogens*, *15*(10), e1008086.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 3146–3154). Curran Associates, Inc.

Kelley, M. R., & Fishel, M. L. (2016). *DNA Repair in Cancer Therapy: Molecular Targets and Clinical Applications*. Academic Press.

Ketcher, D., Otto, A., & Reblin, M. (2019). Chronic conditions among advanced cancer patients and their spouse caregivers. *Journal of Clinical Orthodontics: JCO*, *37*(31_suppl), 20–20.

Keung, M. Y. T., Wu, Y., & Vadgama, J. V. (2019). PARP Inhibitors as a Therapeutic Agent for Homologous Recombination Deficiency in Breast Cancers. *Journal of Clinical Medicine Research*, *8*(4). https://doi.org/10.3390/jcm8040435

Kim, Y., Zheng, S., Tang, J., Jim Zheng, W., Li, Z., & Jiang, X. (2021). Anticancer drug synergy prediction in understudied tissues using transfer learning. *Journal of the American Medical Informatics Association: JAMIA*, *28*(1), 42–51.

Knijnenburg, T. A., Wang, L., Zimmermann, M. T., Chambwe, N., Gao, G. F., Cherniack, A. D., Fan, H., Shen, H., Way, G. P., Greene, C. S., Liu, Y., Akbani, R., Feng, B., Donehower, L. A., Miller, C., Shen, Y., Karimi, M., Chen, H., Kim, P., … Wang, C. (2018). Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Reports*, *23*(1), 239–254.e6.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A. C., & Wishart, D. S. (2011). DrugBank 3.0: a comprehensive resource for “omics” research on drugs. *Nucleic Acids Research*, *39*(Database issue), D1035–D1041.

Koenderink, J. B., Kavishe, R. A., Rijpma, S. R., & Russel, F. G. M. (2010). The ABCs of multidrug resistance in malaria. *Trends in Parasitology*, *26*(9), 440–446.

Konstantinopoulos, P. A., Cheng, S.-C., Wahner Hendrickson, A. E., Penson, R. T., Schumer, S. T., Austin Doyle, L., Lee, E. K., Kohn, E. C., Duska, L. R., Crispens, M. A., Olawaiye, A. B., Winer, I. S., Barroilhet, L. M., Fu, S., McHale, M. T., Schilder, R. J., Färkkilä, A., Chowdhury, D., Curtis, J., … Matulonis, U. A. (2020). Berzosertib plus gemcitabine versus gemcitabine alone in platinum-resistant high-grade serous ovarian cancer: a multicentre, open-label, randomised, phase 2 trial. In *The Lancet Oncology* (Vol. 21, Issue 7, pp. 957–968). https://doi.org/10.1016/s1470-2045(20)30180-7

Konstantinopoulos, P. A., Cheng, S.-C., Wahner Hendrickson, A. E., Penson, R. T., Schumer, S. T., Doyle, L. A., Lee, E. K., Kohn, E. C., Duska, L. R., Crispens, M. A., Olawaiye, A. B., Winer, I. S., Barroilhet, L. M., Fu, S., McHale, M. T., Schilder, R. J., Färkkilä, A., Chowdhury, D., Curtis, J., … Matulonis, U. A. (2020). Berzosertib plus gemcitabine versus gemcitabine alone in platinum-resistant high-grade serous ovarian cancer: a multicentre, open-label, randomised, phase 2 trial. *The Lancet Oncology*, *21*(7), 957–968.

Konstantinopoulos, P. A., da Costa, A. A. B. A., Gulhan, D., Lee, E. K., Cheng, S.-C., Hendrickson, A. E. W., Kochupurakkal, B., Kolin, D. L., Kohn, E. C., Liu, J. F., Stover, E. H., Curtis, J., Tayob, N., Polak, M., Chowdhury, D., Matulonis, U. A., Färkkilä, A., D’Andrea, A. D., & Shapiro, G. I. (2021). A Replication stress biomarker is associated with response to gemcitabine versus combined

gemcitabine and ATR inhibitor therapy in ovarian cancer. *Nature Communications*, *12*(1), 5574.

Korshunova, M., Huang, N., Capuzzi, S., Radchenko, D. S., Savych, O., Moroz, Y. S., Wells, C. I., Willson, T. M., Tropsha, A., & Isayev, O. (2022). Generative and reinforcement learning approaches for the automated de novo design of bioactive compounds. *Communications Chemistry*, *5*(1), 129.

Kwok, M., Davies, N., Agathanggelou, A., Smith, E., Oldreive, C., Petermann, E., Stewart, G., Brown, J., Lau, A., Pratt, G., Parry, H., Taylor, M., Moss, P., Hillmen, P., & Stankovic, T. (2016). ATR inhibition induces synthetic lethality and overcomes chemoresistance in TP53- or ATM-defective chronic lymphocytic leukemia cells. *Blood*, *127*(5), 582–595.

LaFargue, C. J., Dal Molin, G. Z., Sood, A. K., & Coleman, R. L. (2019). Exploring and comparing adverse events between PARP inhibitors. *The Lancet Oncology*, *20*(1), e15–e28.

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., Johnson, J., Dougherty, B., Barrett, J. C., & Dry, J. R. (2016). VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Research*, *44*(11), e108.

Landrum, G. (2006). RDKit: Open-source cheminformatics. 2006. *Google Scholar*. https://cir.nii.ac.jp/crid/1370004237630036224

Larsson, P., Engqvist, H., Biermann, J., Werner Rönnerman, E., Forssell-Aronsson, E., Kovács, A., Karlsson, P., Helou, K., & Parris, T. Z. (2020). Optimization of cell viability assays to improve replicability and reproducibility of cancer drug sensitivity screens. *Scientific Reports*, *10*(1), 5798.

Lee, C. H., Sidik, K., & Chin, K. V. (2001). Role of cAMP-dependent protein kinase in the regulation of DNA repair. *Cancer Letters*, *169*(1), 51–58.

Lee, S., & Greenspan, D. S. (1995). Transcriptional promoter of the human alpha 1(V) collagen gene (COL5A1). *Biochemical Journal*, *310 (Pt 1)*, 15–22.

Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (n.d.). *The molecular signatures database (MSigDB) hallmark gene set collection. Cell Syst. 2015; 1 (6): 417--25*. Epub 2016/01/16. https://doi. org/10.1016/j. cels. 2015.12. 004 PMID: 26771021.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011).

Molecular signatures database (MSigDB) 3.0. *Bioinformatics* , *27*(12), 1739–1740.

Li, D., Wang, Y., Hu, W., Chen, F., Zhao, J., Chen, X., & Han, L. (2021). Application of Machine Learning Classifier to Drug Resistance Analysis. *Frontiers in Cellular and Infection Microbiology*, *11*, 742062.

Li, H., Hu, S., Neamati, N., & Guan, Y. (2019). TAIJI: approaching experimental replicates-level accuracy for drug synergy prediction. *Bioinformatics* , *35*(13), 2338–2339.

Li, H., Li, T., Quang, D., & Guan, Y. (2018). Network Propagation Predicts Drug Synergy in Cancers. *Cancer Research*, *78*(18), 5446–5457.

Li, J., Tong, X.-Y., Zhu, L.-D., & Zhang, H.-Y. (2020). A Machine Learning Method for Drug Combination Prediction. *Frontiers in Genetics*, *11*, 1000.

Lin, A., Giuliano, C. J., Palladino, A., John, K. M., Abramowicz, C., Yuan, M. L., Sausville, E. L., Lukow, D. A., Liu, L., Chait, A. R., Galluzzo, Z. C., Tucker, C., & Sheltzer, J. M. (2019). Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. *Science Translational Medicine*, *11*(509). https://doi.org/10.1126/scitranslmed.aaw8412

Ling, A., & Huang, R. S. (2020). Computationally predicting clinical drug combination efficacy with cancer cell line screens and independent drug action. *Nature Communications*, *11*(1), 5848.

Li, S., Ting, N. S., Zheng, L., Chen, P. L., Ziv, Y., Shiloh, Y., Lee, E. Y., & Lee, W. H. (2000). Functional link of BRCA1 and ataxia telangiectasia gene product in DNA damage response. *Nature*, *406*(6792), 210–215.

Liu, H., Herrmann, C. H., Chiang, K., Sung, T.-L., Moon, S.-H., Donehower, L. A., & Rice, A. P. (2010). 55K isoform of CDK9 associates with Ku70 and is involved in DNA repair. *Biochemical and Biophysical Research Communications*, *397*(2), 245–250.

Liu, H., Zhang, W., Zou, B., Wang, J., Deng, Y., & Deng, L. (2020). DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. *Nucleic Acids Research*, *48*(D1), D871–D881.

Liu, S., Wu, Y., Yang, T., Feng, C., & Jiang, H. (2016). Coexistence of YWHAZ amplification predicts

better prognosis in muscle-invasive bladder cancer with CDKN2A or TP53 loss. *Oncotarget*, *7*(23), 34752–34758.

Li, X., Wang, K., Lyu, Y., Pan, H., Zhang, J., Stambolian, D., Susztak, K., Reilly, M. P., Hu, G., & Li, M. (2020). Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nature Communications*, *11*(1), 2338.

Lloyd, R. L., Wijnhoven, P. W. G., Ramos-Montoya, A., Wilson, Z., Illuzzi, G., Falenta, K., Jones, G. N., James, N., Chabbert, C. D., Stott, J., Dean, E., Lau, A., & Young, L. A. (2020). Combined PARP and ATR inhibition potentiates genome instability and cell death in ATM-deficient cancer cells. *Oncogene*, *39*(25), 4869–4883.

Lord, C. J., & Ashworth, A. (2012). The DNA damage response and cancer therapy. *Nature*, *481*(7381), 287–294.

Lord, C. J., & Ashworth, A. (2016). BRCAness revisited. *Nature Reviews. Cancer*, *16*(2), 110–120.

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. http://arxiv.org/abs/1705.07874

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, *2*(1), 56–67.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates, Inc.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., & Lee, S.-I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, *2*(10), 749–760.

Lu, Y., Chen, J., Xiao, M., Li, W., & Miller, D. D. (2012). An overview of tubulin inhibitors that interact with the colchicine binding site. *Pharmaceutical Research*, *29*(11), 2943–2971.

Madariaga, A., Bowering, V., Ahrari, S., Oza, A. M., & Lheureux, S. (2020). Manage wisely: poly

(ADP-ribose) polymerase inhibitor (PARPi) treatment and adverse events. *International Journal of Gynecological Cancer: Official Journal of the International Gynecological Cancer Society*, *30*(7), 903–915.

Mahdi, H., Hafez, N., Doroshow, D., Sohal, D., Keedy, V., Do, K. T., LoRusso, P., Jürgensmeier, J., Avedissian, M., Sklar, J., Glover, C., Felicetti, B., Dean, E., Mortimer, P., Shapiro, G. I., & Eder, J. P. (2021). Ceralasertib-mediated ATR inhibition combined with olaparib in advanced cancers harboring DNA damage response and repair alterations (Olaparib Combinations). *JCO Precision Oncology*, *5*(5), 1432–1442.

Maier, A. G., Rug, M., O'Neill, M. T., Brown, M., Chakravorty, S., Szestak, T., Chesson, J., Wu, Y., Hughes, K., Coppel, R. L., Newbold, C., Beeson, J. G., Craig, A., Crabb, B. S., & Cowman, A. F. (2008). Exported proteins required for virulence and rigidity of Plasmodium falciparum-infected human erythrocytes. *Cell*, *134*(1), 48–61.

Ma, J., Fong, S. H., Luo, Y., Bakkenist, C. J., Shen, J. P., Mourragui, S., Wessels, L. F. A., Hafner, M., Sharan, R., Peng, J., & Ideker, T. (2021). Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nature Cancer*, *2*(2), 233–244.

Malaria: Biology and Disease. (2016). *Cell*, *167*(3), 610–624.

Malyutina, A., Majumder, M. M., Wang, W., Pessia, A., Heckman, C. A., & Tang, J. (2019). Drug combination sensitivity scoring facilitates the discovery of synergistic and efficacious drug combinations in cancer. *PLoS Computational Biology*, *15*(5), e1006752.

Markossian, S., Grossman, A., Brimacombe, K., Arkin, M., Auld, D., Austin, C., Baell, J., Chung, T. D. Y., Coussens, N. P., Dahlin, J. L., Devanarayan, V., Foley, T. L., Glicksman, M., Gorshkov, K., Haas, J. V., Hall, M. D., Hoare, S., Inglese, J., Iversen, P. W., … Xu, X. (Eds.). (n.d.). *Assay Guidance Manual*. Eli Lilly & Company and the National Center for Advancing Translational Sciences.

Martorana, F., Da Silva, L. A., Sessa, C., & Colombo, I. (2022). Everything Comes with a Price: The Toxicity Profile of DNA-Damage Response Targeting Agents. *Cancers*, *14*(4).

https://doi.org/10.3390/cancers14040953

Mbacham, W. F., Ayong, L., Guewo-Fokeng, M., & Makoge, V. (2019). Current Situation of Malaria in Africa. *Methods in Molecular Biology* , *2013*, 29–44.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122.

Mechanisms of artemisinin resistance in Plasmodium falciparum malaria. (2018). *Current Opinion in Pharmacology*, *42*, 46–54.

Menden, M. P., AstraZeneca-Sanger Drug Combination DREAM Consortium, Wang, D., Mason, M. J., Szalai, B., Bulusu, K. C., Guan, Y., Yu, T., Kang, J., Jeon, M., Wolfinger, R., Nguyen, T., Zaslavskiy, M., Jang, I. S., Ghazoui, Z., Ahsen, M. E., Vogel, R., Neto, E. C., Norman, T., … Saez-Rodriguez, J. (2019). Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. In *Nature Communications* (Vol. 10, Issue 1). https://doi.org/10.1038/s41467-019-09799-2

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C. J., Segura-Cabrera, A., … Leach, A. R. (2019). ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, *47*(D1), D930–D940.

Middleton, M. R., Dean, E., Evans, T. R. J., Shapiro, G. I., Pollard, J., Hendriks, B. S., Falk, M., Diaz-Padilla, I., & Plummer, R. (2021). Phase 1 study of the ATR inhibitor berzosertib (formerly M6620, VX-970) combined with gemcitabine±cisplatin in patients with advanced solid tumours. *British Journal of Cancer*, *125*(4), 510–519.

Miller, L. H., & Su, X. (2011). Artemisinin: Discovery from the Chinese Herbal Garden. *Cell*, *146*(6), 855–858.

Min, A., Im, S.-A., Jang, H., Kim, S., Lee, M., Kim, D. K., Yang, Y., Kim, H.-J., Lee, K.-H., Kim, J. W., Kim, T.-Y., Oh, D.-Y., Brown, J., Lau, A., O'Connor, M. J., & Bang, Y.-J. (2017). AZD6738, A

Novel Oral Inhibitor of ATR, Induces Synthetic Lethality with ATM Deficiency in Gastric Cancer Cells. *Molecular Cancer Therapeutics*, *16*(4), 566–577.

Minchom, A., Aversa, C., & Lopez, J. (2018). Dancing with the DNA damage response: next-generation anti-cancer therapeutic strategies. *Therapeutic Advances in Medical Oncology*, *10*, 1758835918786658.

Moffat, J. G., Vincent, F., Lee, J. A., Eder, J., & Prunotto, M. (2017). Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nature Reviews. Drug Discovery*, *16*(8), 531–543.

Mohiuddin, I. S., & Kang, M. H. (2019). DNA-PK as an Emerging Therapeutic Target in Cancer. *Frontiers in Oncology*, *9*, 635.

Mohs, R. C., & Greig, N. H. (2017). Drug discovery and development: Role of basic biological research. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *3*(4), 651–657.

Mok, S., Ashley, E. A., Ferreira, P. E., Zhu, L., Lin, Z., Yeo, T., Chotivanich, K., Imwong, M., Pukrittayakamee, S., Dhorda, M., Nguon, C., Lim, P., Amaratunga, C., Suon, S., Hien, T. T., Htut, Y., Faiz, M. A., Onyamboko, M. A., Mayxay, M., … Bozdech, Z. (2015). Drug resistance. Population transcriptomics of human malaria parasites reveals the mechanism of artemisinin resistance. *Science*, *347*(6220), 431–435.

Mok, S., Imwong, M., Mackinnon, M. J., Sim, J., Ramadoss, R., Yi, P., Mayxay, M., Chotivanich, K., Liong, K.-Y., Russell, B., Socheat, D., Newton, P. N., Day, N. P. J., White, N. J., Preiser, P. R., Nosten, F., Dondorp, A. M., & Bozdech, Z. (2011). Artemisinin resistance in Plasmodium falciparum is associated with an altered temporal pattern of transcription. *BMC Genomics*, *12*, 391.

Mok, S., Stokes, B. H., Gnädig, N. F., Ross, L. S., Yeo, T., Amaratunga, C., Allman, E., Solyakov, L., Bottrill, A. R., Tripathi, J., Fairhurst, R. M., Llinás, M., Bozdech, Z., Tobin, A. B., & Fidock, D. A. (2021). Artemisinin-resistant K13 mutations rewire Plasmodium falciparum's intra-erythrocytic metabolic program to enhance survival. *Nature Communications*, *12*(1), 530.

Moon, K. R., Stanley, J. S., Burkhardt, D., van Dijk, D., Wolf, G., & Krishnaswamy, S. (2018). Manifold

learning-based methods for analyzing single-cell RNA-sequencing data. *Current Opinion in Systems Biology*, *7*, 36–46.

Morgan, P., Brown, D. G., Lennard, S., Anderton, M. J., Barrett, J. C., Eriksson, U., Fidock, M., Hamrén, B., Johnson, A., March, R. E., Matcham, J., Mettetal, J., Nicholls, D. J., Platz, S., Rees, S., Snowden, M. A., & Pangalos, M. N. (2018). Impact of a five-dimensional framework on R&D productivity at AstraZeneca. *Nature Reviews. Drug Discovery*, *17*(3), 167–181.

Mullard, A. (2022). DNA damage response drugs for cancer yield continued synthetic lethality learnings. *Nature Reviews. Drug Discovery*, *21*(6), 403–405.

Nagarajan, A., Dogra, S. K., Liu, A. Y., Green, M. R., & Wajapeyee, N. (2014). PEA15 regulates the DNA damage-induced cell cycle checkpoint and oncogene-directed transformation. *Molecular and Cellular Biology*, *34*(12), 2264–2282.

Nakad, R., & Schumacher, B. (2016). DNA Damage Response and Immune Defense: Links and Mechanisms. *Frontiers in Genetics*, *7*, 147.

Nam, A.-R., Jin, M. H., Park, J. E., Bang, J.-H., Oh, D.-Y., & Bang, Y.-J. (2019). Therapeutic Targeting of the DNA Damage Response Using an ATR Inhibitor in Biliary Tract Cancer. *Cancer Research and Treatment: Official Journal of Korean Cancer Association*, *51*(3), 1167–1179.

Narayan, R. S., Molenaar, P., Teng, J., Cornelissen, F. M. G., Roelofs, I., Menezes, R., Dik, R., Lagerweij, T., Broersma, Y., Petersen, N., Marin Soto, J. A., Brands, E., van Kuiken, P., Lecca, M. C., Lenos, K. J., In ’t Veld, S. G. J. G., van Wieringen, W., Lang, F. F., Sulman, E., … Westerman, B. A. (2020). A cancer drug atlas enables synergistic targeting of independent drug vulnerabilities. *Nature Communications*, *11*(1), 2935.

New Approaches to Drug Discovery. (n.d.). Springer International Publishing.

Ngalah, B. S., Ingasia, L. A., Cheruiyot, A. C., Chebon, L. J., Juma, D. W., Muiruri, P., Onyango, I., Ogony, J., Yeda, R. A., Cheruiyot, J., Mbuba, E., Mwangoka, G., Achieng, A. O., Ng’ang'a, Z., Andagalu, B., Akala, H. M., & Kamau, E. (2015). Analysis of Major Genome Loci Underlying Artemisinin Resistance and pfmdr1 Copy Number in pre- and post-ACTs in Western Kenya. In

Scientific Reports (Vol. 5, Issue 1). https://doi.org/10.1038/srep08308

Nosten, F., & White, N. J. (2007). *Artemisinin-Based Combination Treatment of Falciparum Malaria*. American Society of Tropical Medicine and Hygiene.

Oakley, M. S. M., Kumar, S., Anantharaman, V., Zheng, H., Mahajan, B., Haynes, J. D., Moch, J. K., Fairhurst, R., McCutchan, T. F., & Aravind, L. (2007). Molecular factors and biochemical pathways induced by febrile temperature in intraerythrocytic Plasmodium falciparum parasites. *Infection and Immunity*, *75*(4), 2012–2025.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, *3*, 33.

O'Connor, M. J. (2015). Targeting the DNA Damage Response in Cancer. *Molecular Cell*, *60*(4), 547–560.

Office of the Commissioner. (2020, February 20). *The Drug Development Process*. U.S. Food and Drug Administration; FDA. https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process

O'Neil, J., Benita, Y., Feldman, I., Chenard, M., Roberts, B., Liu, Y., Li, J., Kral, A., Lejnine, S., Loboda, A., Arthur, W., Cristescu, R., Haines, B. B., Winter, C., Zhang, T., Bloecher, A., & Shumway, S. D. (2016). An Unbiased Oncology Compound Screen to Identify Novel Combination Strategies. *Molecular Cancer Therapeutics*, *15*(6), 1155–1162.

Organization, W. H., & Others. (2020). *World malaria report 2020: 20 years of global progress and challenges*. https://apps.who.int/iris/bitstream/handle/10665/337660/9789240015791-eng.pdf

Ould Ahmedou Salem, M. S., Mint Lekweiry, K., Bouchiba, H., Pascual, A., Pradines, B., Ould Mohamed Salem Boukhary, A., Briolant, S., Basco, L. K., & Bogreau, H. (2017). Characterization of Plasmodium falciparum genes associated with drug resistance in Hodh Elgharbi, a malaria hotspot near Malian-Mauritanian border. *Malaria Journal*, *16*(1), 140.

Palmer, A. C., & Sorger, P. K. (2017). Combination Cancer Therapy Can Confer Benefit via Patient-to-Patient Variability without Drug Additivity or Synergy. *Cell*, *171*(7), 1678–1691.e13.

Patterson, T. A., Lobenhofer, E. K., Fulmer-Smentek, S. B., Collins, P. J., Chu, T.-M., Bao, W., Fang, H., Kawasaki, E. S., Hager, J., Tikhonova, I. R., Walker, S. J., Zhang, L., Hurban, P., de Longueville, F., Fuscoe, J. C., Tong, W., Shi, L., & Wolfinger, R. D. (2006). Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. *Nature Biotechnology*, *24*(9), 1140–1150.

Pearl, L. H., Schierz, A. C., Ward, S. E., Al-Lazikani, B., & Pearl, F. M. G. (2015). Therapeutic opportunities within the DNA damage response. *Nature Reviews. Cancer*, *15*(3), 166–180.

Pilié, P. G., Tang, C., Mills, G. B., & Yap, T. A. (2019). State-of-the-art strategies for targeting the DNA damage response in cancer. *Nature Reviews. Clinical Oncology*, *16*(2), 81–104.

Plana, D., Palmer, A. C., & Sorger, P. K. (2022). Independent Drug Action in Combination Therapy: Implications for Precision Oncology. *Cancer Discovery*, *12*(3), 606–624.

Preuer, K., Lewis, R. P. I., Hochreiter, S., Bender, A., Bulusu, K. C., & Klambauer, G. (2018). DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics* , *34*(9), 1538–1546.

PubChemPy documentation — PubChemPy 1.0.4 documentation. (n.d.). Retrieved November 30, 2023, from https://pubchempy.readthedocs.io/en/latest/

pybel. (n.d.). PyPI. Retrieved November 29, 2023, from https://pypi.org/project/pybel/

Ragland, R. L., Patel, S., Rivard, R. S., Smith, K., Peters, A. A., Bielinsky, A.-K., & Brown, E. J. (2013). RNF4 and PLK1 are required for replication fork collapse in ATR-deficient cells. *Genes & Development*, *27*(20), 2259–2273.

Reaper, P. M., Griffiths, M. R., Long, J. M., Charrier, J.-D., Maccormick, S., Charlton, P. A., Golec, J. M. C., & Pollard, J. R. (2011). Selective killing of ATM- or p53-deficient cancer cells through inhibition of ATR. *Nature Chemical Biology*, *7*(7), 428–430.

RecSys 2020 Challenge. (n.d.).

Research and development policy framework. (n.d.). Retrieved December 21, 2023, from https://phrma.org/policy-issues/Research-and-Development-Policy-Framework

Ritz, C., Baty, F., Streibig, J. C., & Gerhard, D. (2015). Dose-Response Analysis Using R. *PloS One*, *10*(12), e0146021.

Roberts, S. A., Strande, N., Burkhalter, M. D., Strom, C., Havener, J. M., Hasty, P., & Ramsden, D. A. (2010). Ku is a 5'-dRP/AP lyase that excises nucleotide damage near broken ends. *Nature*, *464*(7292), 1214–1217.

Romesser, P. B., Holliday, E. B., Philip, T., Garcia-Carbonero, R., Capdevila, J., Tuli, R., Sarholz, B., Kuipers, M., Rodriguez, A., Diaz-Padilla, I., & Miller, E. D. (2021). A multicenter phase Ib/II study of DNA-PK inhibitor peposertib (M3814) in combination with capecitabine and radiotherapy in patients with locally advanced rectal cancer. In *Journal of Clinical Oncology* (Vol. 39, Issue 3_suppl, pp. TPS144–TPS144). https://doi.org/10.1200/jco.2021.39.3_suppl.tps144

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., & Mewes, H.-W. (2010). CORUM: the comprehensive resource of mammalian protein complexes--2009. *Nucleic Acids Research*, *38*(Database issue), D497–D501.

Sachs, J., & Malaney, P. (2002). The economic and social burden of malaria. *Nature*, *415*(6872), 680–685.

Sagawa, M., Ohguchi, H., Harada, T., Samur, M. K., Tai, Y.-T., Munshi, N. C., Kizaki, M., Hideshima, T., & Anderson, K. C. (2017). Ribonucleotide Reductase Catalytic Subunit M1 (RRM1) as a Novel Therapeutic Target in Multiple Myeloma. In *Clinical Cancer Research* (Vol. 23, Issue 17, pp. 5225–5237). https://doi.org/10.1158/1078-0432.ccr-17-0263

Sakai, W., & Sugasawa, K. (2014). FANCD2 is a target for caspase 3 during DNA damage-induced apoptosis. *FEBS Letters*, *588*(20), 3778–3785.

Sargeant, T. J., Marti, M., Caler, E., Carlton, J. M., Simpson, K., Speed, T. P., & Cowman, A. F. (2006). Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. *Genome Biology*, *7*(2), R12.

Sastry, A. V., Dillon, N., Anand, A., Poudel, S., Hefner, Y., Xu, S., Szubin, R., Feist, A. M., Nizet, V., & Palsson, B. (2021). Machine Learning of Bacterial Transcriptomes Reveals Responses Underlying Differential Antibiotic Susceptibility. *mSphere*, *6*(4), e0044321.

Sen, P., Saha, A., & Dixit, N. M. (2019). You Cannot Have Your Synergy and Efficacy Too. *Trends in Pharmacological Sciences*, *40*(11), 811–817.

Seo, H., Tkachuk, D., Ho, C., Mammoliti, A., Rezaie, A., Madani Tonekaboni, S. A., & Haibe-Kains, B. (2020). SYNERGxDB: an integrative pharmacogenomic portal to identify synergistic drug combinations for precision oncology. *Nucleic Acids Research*, *48*(W1), W494–W501.

Shah, P. D., Wethington, S. L., Pagan, C., Latif, N., Tanyi, J., Martin, L. P., Morgan, M., Burger, R. A., Haggerty, A., Zarrin, H., Rodriguez, D., Domchek, S., Drapkin, R., Shih, I.-M., Smith, S. A., Dean, E., Gaillard, S., Armstrong, D., Torigian, D. A., … Simpkins, F. (2021). Combination ATR and PARP Inhibitor (CAPRI): A phase 2 study of ceralasertib plus olaparib in patients with recurrent, platinum-resistant epithelial ovarian cancer. *Gynecologic Oncology*, *163*(2), 246–253.

Shapley, L. S. (1983). *Additive and Non-additive Set Functions*.

Shaw, P. J., Chaotheing, S., Kaewprommal, P., Piriyapongsa, J., Wongsombat, C., Suwannakitti, N., Koonyosying, P., Uthaipibull, C., Yuthavong, Y., & Kamchonwongpaisan, S. (2015). Plasmodium parasites mount an arrest response to dihydroartemisinin, as revealed by whole transcriptome shotgun sequencing (RNA-seq) and microarray study. *BMC Genomics*, *16*, 830.

Sherman, R. E., Anderson, S. A., Dal Pan, G. J., Gray, G. W., Gross, T., Hunter, N. L., LaVange, L., Marinac-Dabic, D., Marks, P. W., Robb, M. A., Shuren, J., Temple, R., Woodcock, J., Yue, L. Q., & Califf, R. M. (2016). Real-World Evidence - What Is It and What Can It Tell Us? *The New England Journal of Medicine*, *375*(23), 2293–2297.

Shim, Y., Lee, M., Kim, P.-J., & Kim, H.-G. (2022). A novel approach to predicting the synergy of anti-cancer drug combinations using document-based feature extraction. *BMC Bioinformatics*, *23*(1), 163.

Shoemaker, R. H. (2006). The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews. Cancer*, *6*(10), 813–823.

Siddiqui, F. A., Boonhok, R., Cabrera, M., Mbenda, H. G. N., Wang, M., Min, H., Liang, X., Qin, J., Zhu, X., Miao, J., Cao, Y., & Cui, L. (2020). Role of Plasmodium falciparum Kelch 13 Protein Mutations

in P. falciparum Populations from Northeastern Myanmar in Mediating Artemisinin Resistance. *mBio*, *11*(1). https://doi.org/10.1128/mBio.01134-19

Sidhu, A. B. S., Uhlemann, A.-C., Valderramos, S. G., Valderramos, J.-C., Krishna, S., & Fidock, D. A. (2006). Decreasing pfmdr1 copy number in plasmodium falciparum malaria heightens susceptibility to mefloquine, lumefantrine, halofantrine, quinine, and artemisinin. *The Journal of Infectious Diseases*, *194*(4), 528–535.

Sidorov, P., Naulaerts, S., Ariey-Bonnet, J., Pasquier, E., & Ballester, P. J. (2019). Predicting Synergism of Cancer Drug Combinations Using NCI-ALMANAC Data. In *Frontiers in Chemistry* (Vol. 7). https://doi.org/10.3389/fchem.2019.00509

Sisowath, C., Ferreira, P. E., Bustamante, L. Y., Dahlström, S., Mårtensson, A., Björkman, A., Krishna, S., & Gil, J. P. (2007). The role of pfmdr1 in Plasmodium falciparum tolerance to artemether-lumefantrine in Africa. *Tropical Medicine & International Health: TM & IH*, *12*(6), 736–742.

Smith, J., Tho, L. M., Xu, N., & Gillespie, D. A. (2010). The ATM-Chk2 and ATR-Chk1 pathways in DNA damage signaling and cancer. *Advances in Cancer Research*, *108*, 73–112.

Sottile, M. L., & Nadin, S. B. (2018). Heat shock proteins and DNA repair mechanisms: an updated overview. *Cell Stress & Chaperones*, *23*(3), 303–315.

Spiro, A., Fernández García, J., & Yanover, C. (2019). Inferring new relations between medical entities using literature curated term co-occurrences. *JAMIA Open*, *2*(3), 378–385.

Staub, E. (2012). An interferon response gene expression signature is activated in a subset of medulloblastomas. *Translational Oncology*, *5*(4), 297–304.

Study of M1774 in Combination With DNA Damage Response Inhibitor or Immune Checkpoint Inhibitor (DDRiver Solid Tumors 320). (n.d.). Retrieved February 7, 2023, from https://clinicaltrials.gov/ct2/show/NCT05396833

Subbiah, V. (2023). The next generation of evidence-based medicine. *Nature Medicine*, *29*(1), 49–58.

Subhash, V. V., Tan, S. H., Yeo, M. S., Yan, F. L., Peethala, P. C., Liem, N., Krishnan, V., & Yong, W. P.

(2016). ATM Expression Predicts Veliparib and Irinotecan Sensitivity in Gastric Cancer by Mediating P53-Independent Regulation of Cell Cycle and Apoptosis. *Molecular Cancer Therapeutics*, *15*(12), 3087–3096.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. In *Proceedings of the National Academy of Sciences* (Vol. 102, Issue 43, pp. 15545–15550). https://doi.org/10.1073/pnas.0506580102

Sun, D., Gao, W., Hu, H., & Zhou, S. (2022). Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica. B*, *12*(7), 3049–3062.

Szymański, P., Markowicz, M., & Mikiciuk-Olasik, E. (2012). Adaptation of high-throughput screening in drug discovery-toxicological screening tests. *International Journal of Molecular Sciences*, *13*(1), 427–452.

Tabbabi, A., Alkishe, A. A., Samy, A. M., Rhim, A., & Peterson, A. T. (2020). Malaria in North Africa: A Review of the Status of Vectors and Parasites. *Journal of Entomological Science*, *55*(1), 25–37.

Takala-Harrison, S., Clark, T. G., Jacob, C. G., Cummings, M. P., Miotto, O., Dondorp, A. M., Fukuda, M. M., Nosten, F., Noedl, H., Imwong, M., Bethell, D., Se, Y., Lon, C., Tyner, S. D., Saunders, D. L., Socheat, D., Ariey, F., Phyo, A. P., Starzengruber, P., … Plowe, C. V. (2013). Genetic loci associated with delayed clearance of Plasmodium falciparum following artemisinin treatment in Southeast Asia. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(1), 240–245.

Talapko, J., Škrlec, I., Alebić, T., Jukić, M., & Včev, A. (2019). Malaria: The Past and the Present. *Microorganisms*, *7*(6). https://doi.org/10.3390/microorganisms7060179

Talevich, E., Shain, A. H., Botton, T., & Bastian, B. C. (2016). CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Computational Biology*, *12*(4), e1004873.

Tan, A. C., Bagley, S. J., Wen, P. Y., Lim, M., Platten, M., Colman, H., Ashley, D. M., Wick, W., Chang,

S. M., Galanis, E., Mansouri, A., Khagi, S., Mehta, M. P., Heimberger, A. B., Puduvalli, V. K., Reardon, D. A., Sahebjam, S., Simes, J., Antonia, S. J., … Khasraw, M. (2021). Systematic review of combinations of targeted or immunotherapy in advanced solid tumors. *Journal for Immunotherapy of Cancer*, *9*(7). https://doi.org/10.1136/jitc-2021-002459

Tanwar, P., Jain, V., Liu, C.-M., & Goyal, V. (2020). *Big Data Analytics and Intelligence: A Perspective for Health Care*. Emerald Group Publishing.

Tarr, S. J., Moon, R. W., Hardege, I., & Osborne, A. R. (2014). A conserved domain targets exported PHISTb family proteins to the periphery of Plasmodium infected erythrocytes. *Molecular and Biochemical Parasitology*, *196*(1), 29–40.

Therapeutic targeting of ATR yields durable regressions in small cell lung cancers with high replication stress. (2021). *Cancer Cell*, *39*(4), 566–579.e7.

Thomas, A., Redon, C. E., Sciuto, L., Padiernos, E., Ji, J., Lee, M.-J., Yuno, A., Lee, S., Zhang, Y., Tran, L., Yutzy, W., Rajan, A., Guha, U., Chen, H., Hassan, R., Alewine, C. C., Szabo, E., Bates, S. E., Kinders, R. J., … Pommier, Y. (2018). Phase I Study of ATR Inhibitor M6620 in Combination With Topotecan in Patients With Advanced Solid Tumors. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, *36*(16), 1594–1602.

Tilley, L., Straimer, J., Gnädig, N. F., Ralph, S. A., & Fidock, D. A. (2016). Artemisinin Action and Resistance in Plasmodium falciparum. *Trends in Parasitology*, *32*(9), 682–696.

Torkamannia, A., Omidi, Y., & Ferdousi, R. (2022). A review of machine learning approaches for drug synergy prediction in cancer. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbac075

Trabucco, B., Kumar, A., Geng, X., & Levine, S. (18--24 Jul 2021). Conservative Objective Models for Effective Offline Model-Based Optimization. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 10358–10368). PMLR.

Ursing, J., Zakeri, S., Gil, J. P., & Björkman, A. (2006). Quinoline resistance associated polymorphisms in the pfcrt, pfmdr1 and pfmrp genes of Plasmodium falciparum in Iran. *Acta Tropica*, *97*(3), 352–356.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews. Drug Discovery*, *18*(6), 463–477.

van Biljon, R., Niemand, J., van Wyk, R., Clark, K., Verlinden, B., Abrie, C., von Grüning, H., Smidt, W., Smit, A., Reader, J., Painter, H., Llinás, M., Doerig, C., & Birkholtz, L.-M. (2018). Inducing controlled cell cycle arrest and re-entry during asexual proliferation of Plasmodium falciparum malaria parasites. *Scientific Reports*, *8*(1), 16581.

van Bussel, M. T. J., Awada, A., de Jonge, M. J. A., Mau-Sørensen, M., Nielsen, D., Schöffski, P., Verheul, H. M. W., Sarholz, B., Berghoff, K., El Bawab, S., Kuipers, M., Damstrup, L., Diaz-Padilla, I., & Schellens, J. H. M. (2021). A first-in-man phase 1 study of the DNA-dependent protein kinase inhibitor peposertib (formerly M3814) in patients with advanced solid tumours. *British Journal of Cancer*, *124*(4), 728–735.

Van Norman, G. A. (2016a). Drugs, Devices, and the FDA: Part 1: An Overview of Approval Processes for Drugs. *JACC. Basic to Translational Science*, *1*(3), 170–179.

Van Norman, G. A. (2016b). Drugs, Devices, and the FDA: Part 2: An Overview of Approval Processes: FDA Approval of Medical Devices. *JACC. Basic to Translational Science*, *1*(4), 277–287.

Van Triest, B., Damstrup, L., Falkenius, J., Budach, V., Troost, E., Samuels, M., Debus, J., Sørensen, M. M., Berghoff, K., Strotman, R., van Bussel, M., Goel, S., & Geertsen, P. F. (2018). A phase Ia/Ib trial of the DNA-PK inhibitor M3814 in combination with radiotherapy (RT) in patients (pts) with advanced solid tumors: Dose-escalation results. In *Journal of Clinical Oncology* (Vol. 36, Issue 15_suppl, pp. 2518–2518). https://doi.org/10.1200/jco.2018.36.15_suppl.2518

Vecchio, D., & Frosina, G. (2016). Targeting the Ataxia Telangiectasia Mutated Protein in Cancer Therapy. *Current Drug Targets*, *17*(2), 139–153.

Vichai, V., & Kirtikara, K. (2006). Sulforhodamine B colorimetric assay for cytotoxicity screening. *Nature Protocols*, *1*(3), 1112–1116.

Wang, C., & Li, J. (2021). Haematologic toxicities with PARP inhibitors in cancer patients: an up-to-date

meta-analysis of 29 randomized controlled trials. *Journal of Clinical Pharmacy and Therapeutics*, *46*(3), 571–584.

Wang, C., Tang, H., Geng, A., Dai, B., Zhang, H., Sun, X., Chen, Y., Qiao, Z., Zhu, H., Yang, J., Chen, J., He, Q., Qin, N., Xie, J., Tan, R., Wan, X., Gao, S., Jiang, Y., Sun, F.-L., & Mao, Z. (2020). Rational combination therapy for hepatocellular carcinoma with PARP1 and DNA-PK inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(42), 26356–26365.

Wang, J., Wu, M., Huang, X., Wang, L., Zhang, S., Liu, H., & Zheng, J. (2022). SynLethDB 2.0: a web-based knowledge graph database on synthetic lethality for novel anticancer drug discovery. *Database: The Journal of Biological Databases and Curation*, *2022*. https://doi.org/10.1093/database/baac030

Wang, M., Xu, Y., Liu, J., Ye, J., Yuan, W., Jiang, H., Wang, Z., Jiang, H., & Wan, J. (2017). Recent Insights into the Biological Functions of Sestrins in Health and Disease. *Cellular Physiology and Biochemistry: International Journal of Experimental Cellular Physiology, Biochemistry, and Pharmacology*, *43*(5), 1731–1741.

Wang, Q., Gao, F., May, W. S., Zhang, Y., Flagg, T., & Deng, X. (2008). Bcl2 negatively regulates DNA double-strand-break repair through a nonhomologous end-joining pathway. *Molecular Cell*, *29*(4), 488–498.

Wang, Y., Wang, J., Cao, Z., & Barati Farimani, A. (2022). Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, *4*(3), 279–287.

Warncke, J. D., Vakonakis, I., & Beck, H.-P. (2016). Plasmodium Helical Interspersed Subtelomeric (PHIST) Proteins, at the Center of Host Cell Remodeling. *Microbiology and Molecular Biology Reviews: MMBR*, *80*(4), 905–927.

Weber, A. M., & Ryan, A. J. (2015). ATM and ATR as therapeutic targets in cancer. *Pharmacology & Therapeutics*, *149*, 124–138.

West, R. B., Yaneva, M., & Lieber, M. R. (1998). Productive and nonproductive complexes of Ku and DNA-dependent protein kinase at DNA termini. *Molecular and Cellular Biology*, *18*(10),

5908–5920.

Wildey, M. J., Haunso, A., Tudor, M., Webb, M., & Connick, J. H. (2017). Chapter Five - High-Throughput Screening. In R. A. Goodnow (Ed.), *Annual Reports in Medicinal Chemistry* (Vol. 50, pp. 149–195). Academic Press.

Wong, A. K., Krishnan, A., & Troyanskaya, O. G. (2018). GIANT 2.0: genome-scale integrated analysis of gene networks in tissues. *Nucleic Acids Research*, *46*(W1), W65–W70.

World Health Organization. (2020). *World Malaria Report 2019*. World Health Organization.

Xia, F., Allen, J., Balaprakash, P., Brettin, T., Garcia-Cardona, C., Clyde, A., Cohn, J., Doroshow, J., Duan, X., Dubinkina, V., Evrard, Y., Fan, Y. J., Gans, J., He, S., Lu, P., Maslov, S., Partin, A., Shukla, M., Stahlberg, E., … Stevens, R. (2022). A cross-study analysis of drug response prediction in cancer cell lines. *Briefings in Bioinformatics*, *23*(1). https://doi.org/10.1093/bib/bbab356

Xia, F., Shukla, M., Brettin, T., Garcia-Cardona, C., Cohn, J., Allen, J. E., Maslov, S., Holbeck, S. L., Doroshow, J. H., Evrard, Y. A., Stahlberg, E. A., & Stevens, R. L. (2018). Predicting tumor cell line response to drug pairs with deep learning. *BMC Bioinformatics*, *19*(Suppl 18), 486.

Xu, X., Page, J. L., Surtees, J. A., Liu, H., Lagedrost, S., Lu, Y., Bronson, R., Alani, E., Nikitin, A. Y., & Weiss, R. S. (2008). Broad overexpression of ribonucleotide reductase genes in mice specifically induces lung neoplasms. *Cancer Research*, *68*(8), 2652–2660.

Yadav, B., Wennerberg, K., Aittokallio, T., & Tang, J. (2015). Searching for Drug Synergy in Complex Dose-Response Landscapes Using an Interaction Potency Model. *Computational and Structural Biotechnology Journal*, *13*, 504–513.

Yadav, J., El Hassani, M., Sodhi, J., Lauschke, V. M., Hartman, J. H., & Russell, L. E. (2021). Recent developments in in vitro and in vivo models for improved translation of preclinical pharmacokinetics and pharmacodynamics data. *Drug Metabolism Reviews*, *53*(2), 207–233.

Yang, J., Tang, H., Li, Y., Zhong, R., Wang, T., Wong, S., Xiao, G., & Xie, Y. (2015). DIGRE: Drug-Induced Genomic Residual Effect Model for Successful Prediction of Multidrug Effects. *CPT: Pharmacometrics & Systems Pharmacology*, *4*(2), e1.

Yap, T. A., O'Carrigan, B., Penney, M. S., Lim, J. S., Brown, J. S., de Miguel Luken, M. J., Tunariu, N., Perez-Lopez, R., Rodrigues, D. N., Riisnaes, R., Figueiredo, I., Carreira, S., Hare, B., McDermott, K., Khalique, S., Williamson, C. T., Natrajan, R., Pettitt, S. J., Lord, C. J., … de Bono, J. S. (2020). Phase I Trial of First-in-Class ATR Inhibitor M6620 (VX-970) as Monotherapy or in Combination With Carboplatin in Patients With Advanced Solid Tumors. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, *38*(27), 3195–3204.

Yap, T. A., Tolcher, A. W., Plummer, R., & Mukker, J. (2022). 457MO A phase I study of ATR inhibitor M1774 in patients with solid tumours (DDRiver Solid Tumours 301): Part A1 results. *Annals of*. https://www.annalsofoncology.org/article/S0923-7534(22)02437-1/abstract

Yu, Y., Chen, L., Zhao, G., Li, H., Guo, Q., Zhu, S., Li, P., Min, L., & Zhang, S. (2020). RBBP8/CtIP suppresses P21 expression by interacting with CtBP and BRCA1 in gastric cancer. *Oncogene*, *39*(6), 1273–1289.

Zagidullin, B., Aldahdooh, J., Zheng, S., Wang, W., Wang, Y., Saad, J., Malyutina, A., Jafari, M., Tanoli, Z., Pessia, A., & Tang, J. (2019). DrugComb: an integrative cancer drug combination data portal. *Nucleic Acids Research*, *47*(W1), W43–W51.

Zagidullin, B., Wang, Z., Guan, Y., Pitkänen, E., & Tang, J. (2021). Comparative analysis of molecular fingerprints in prediction of drug combination effects. *Briefings in Bioinformatics*, *22*(6). https://doi.org/10.1093/bib/bbab291

Zenke, F. T., Zimmermann, A., Sirrenberg, C., Dahmen, H., Kirkin, V., Pehl, U., Grombacher, T., Wilm, C., Fuchss, T., Amendt, C., Vassilev, L. T., & Blaukat, A. (2020). Pharmacologic Inhibitor of DNA-PK, M3814, Potentiates Radiotherapy and Regresses Human Tumors in Mouse Models. *Molecular Cancer Therapeutics*, *19*(5), 1091–1101.

Zhang, H., Kreis, J., Schelhorn, S.-E., Dahmen, H., Grombacher, T., Zühlsdorf, M., Zenke, F. T., & Guan, Y. (2023). Mapping combinatorial drug effects to DNA damage response kinase inhibitors. *Nature Communications*, *14*(1), 1–8.

Zhang, H., Wang, Z., Nan, Y., Zagidullin, B., Yi, D., Tang, J., & Guan, Y. (2023). Harmonizing across

datasets to improve the transferability of drug combination prediction. *Communications Biology*, *6*(1), 397.

Zhang, T., Zhang, L., Payne, P. R. O., & Li, F. (2021). Synergistic Drug Combination Prediction by Integrating Multiomics Data in Deep Learning Models. *Methods in Molecular Biology* , *2194*, 223–238.

Zhang, Y., Wang, Y., Zhou, W., Fan, Y., Zhao, J., Zhu, L., Lu, S., Lu, T., Chen, Y., & Liu, H. (2019). A combined drug discovery strategy based on machine learning and molecular docking. *Chemical Biology & Drug Design*, *93*(5), 685–699.

Zheng, S., Aldahdooh, J., Shadbahr, T., Wang, Y., Aldahdooh, D., Bao, J., Wang, W., & Tang, J. (2021). DrugComb update: a more comprehensive drug sensitivity data repository and analysis portal. *Nucleic Acids Research*, *49*(W1), W174–W184.

Zhu, L., Tripathi, J., Rocamora, F. M., Miotto, O., van der Pluijm, R., Voss, T. S., Mok, S., Kwiatkowski, D. P., Nosten, F., Day, N. P. J., White, N. J., Dondorp, A. M., Bozdech, Z., & Tracking Resistance to Artemisinin Collaboration I. (2018). The origins of malaria artemisinin resistance defined by a genetic and transcriptomic background. *Nature Communications*, *9*(1), 5158.

Zhu, L., van der Pluijm, R. W., Kucharski, M., Nayak, S., Tripathi, J., White, N. J., Day, N. P. J., Faiz, A., Phyo, A. P., Amaratunga, C., Lek, D., Ashley, E. A., Nosten, F., Smithuis, F., Ginsburg, H., von Seidlein, L., Lin, K., Imwong, M., Chotivanich, K., … Bozdech, Z. (2022). Artemisinin resistance in the malaria parasite, Plasmodium falciparum, originates from its initial transcriptional response. *Communications Biology*, *5*(1), 274.

Zhu, Y., Brettin, T., Evrard, Y. A., Partin, A., Xia, F., Shukla, M., Yoo, H., Doroshow, J. H., & Stevens, R. L. (2020). Ensemble transfer learning for the prediction of anti-cancer drug response. *Scientific Reports*, *10*(1), 18040.

Zimmermann, A., Dahmen, H., Grombacher, T., Pehl, U., Blaukat, A., & Zenke, F. T. (2022). Abstract 2588: M1774, a novel potent and selective ATR inhibitor, shows antitumor effects as monotherapy and in combination. *Cancer Research*, *82*(12_Supplement), 2588–2588.

Zimmermann, A., Zenke, F. T., Chiu, L.-Y., Dahmen, H., Pehl, U., Fuchss, T., Grombacher, T., Blume, B., Vassilev, L. T., & Blaukat, A. (2022). A New Class of Selective ATM Inhibitors as Combination Partners of DNA Double-Strand Break Inducing Cancer Therapies. *Molecular Cancer Therapeutics*, *21*(6), 859–870.