TOWARDS LEVERAGING INHIBITION STATE OF THE KINOME FOR PRECISION ONCOLOGY

Chinmaya U. Joisa

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the department of Biomedical Engineering in the School of Medicine.

Chapel Hill
2023

Approved by:

Imran Rizvi

Jen Jen Yeh

Lee M. Graves

Gary L. Johnson

Shawn M. Gomez

ABSTRACT

Chinmaya U. Joisa: Towards Leveraging Inhibition State of the Kinome for Precision Oncology
(Under the direction of Shawn M. Gomez)

Protein phosphorylation forms the most common method of regulation in eukaryotes, and kinases are enzymes that chiefly enable its application. Due to their central role in physiology, dysregulation of the kinome is implicated in a myriad of diseases, particularly cancer. This dissertation demonstrates that the measured inhibition of the kinome (the kinome inhibition state) by cancer targeted therapies can be predictive of cell line and patient-derived xenograft (PDX) tumor responses to treatment by that therapy using interpretable machine learning models. The predictive capability of kinome inhibition states with currently used baseline genomics for monotherapy cancer cell line responses across diverse cancer types is demonstrated first using multi-dose kinome inhibition states, and second using multi-assay single-dose data. Then, the predictive value of kinome inhibition states is extended to kinase inhibitor combination therapies, demonstrating that combined kinome inhibition states can accurately predict cancer cell line sensitivity and synergy to combination treatments, providing the basis for rational kinome-informed drug combination selection. Finally, the predictive capacity of kinome inhibition states is demonstrated for PDX tumor responses in five common solid tumor types, confirming the

generalizability of kinome inhibition-based prediction models in a preclinical setting, and emphasizing their potential for clinical translation and application in precision oncology. Overall, this dissertation provides compelling evidence that integrating kinome inhibition states in

machine learning models can enhance the prediction of cancer cell line and PDX tumor responses. This work shows that kinome inhibition data has potential to be included in precision oncology platforms alongside baseline genomic profiling, aiding in the identification of effective therapeutic strategies and ultimately improving patient outcomes.

To Mrs. Raman, Ms. Bhat, Mrs. Venkatesh, and Mr. Manjunath who showed me that in science there is beauty.

ACKNOWLEDGEMENTS

None of the work presented in this dissertation would have been possible without the immense help offered by members of the Gomez and Johnson labs. Particularly, I would like t0 thank Matthew Berginski and Michael East for all their training and advice over the years which has chiefly made me into the scientist I am today. I would also like to thank Shawn for being an incredible advisor over the past five years, allowing me room to fail time and time again, staying true to the nature of scientific pursuit. Additionally, I'd like to thank Kevin Chen, Madison Jenner, and Felix Olivares for their help and companionship with the more technical aspects of the work presented here.

Finally, on a more personal note, I would like to thank my family, my friends, and my partner for their immense sacrifices over the years, supporting me all the way through this arduous PhD route, and especially through the pandemic. I find myself incredibly grateful and fortunate to be able to conduct this work freely and without burdens, and the support of the people around me is what makes any of this possible.

Chinmaya Joisa

University of North Carolina at Chapel Hill

May 2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial intelligence
AML	Acute myeloid leukemia
AUC	Area under the curve
BRCA	Breast cancer
BRET	Bioluminescent resonance energy transfer
CCLE	Cancer cell line encyclopedia
CM	Cutaneous melanoma
CML	Chronic myeloid leukemia
CNV	Copy-number variation
CRC	Colorectal cancer
CRISPR	Clustered regularly interspersed short palindromic repeats
CRorPR	Complete or partial response
DREAM	Dialogue on Reverse Engineering Assessment and Methods
FDA	Food and drug administration
GEO	Gene expression omnibus
GIST	Gastrointestinal stromal tumors
GPCR	G-protein coupled receptor
HDAC	Histone-deacetylase
HGNC	HUGO gene nomenclature committee
HSA	Highest single agent
IC50	Half-maximal inhibitory concentration
KCGS	The kinase chemogenomic set
KO	Knockout

LASSO	Least absolute shrinkage and selection operator
LINCS	The Library of Integrated Network-Based Cellular Signatures
MS	Mass spectrometry
NCI	National cancer institue
NSCLC	Non-small cell lung cancer
PDAC	Pancreatic ductal adenocarcinoma
PDO	Patient-derived organoid
PDX	Patient-derived xenograft
PKIS	Published kinase inhibitor set
PR	Precision recall
PRISM	Profiling Relative Inhibition Simultaneously in Mixtures
QSAR	Quantitative structure-activity relationships
RECIST	Response evaluation criteria in solid tumors
RMSE	Root mean squared error
ROC	Receiver-operating curve
SGC	Structural genomics consortium
STRING	Search tool for the retrieval of interacting genes/proteins
TNBC	Triple-negative breast cancer
TPM	Transcripts per million
UMAP	Uniform manifold approximation and projection
ZIP	Zero interaction potency

CHAPTER 1: INTRODUCTION

PROTEIN KINASES IN CANCER

The cover of the May 2001 TIME magazine read "There is new ammunition in the war against cancer, and these are the bullets"[1], stirring up the same hope that had been relentlessly fanned since President Nixon officially declared his "war" on the disease[2]. The drug in question, imatinib (Gleevec ©) was titled "A New Hope for Cancer", and was rationally designed against the BCR-ABL1 fusion protein, a result of the infamous "Philadelphia Chromosome"[3] mutation. This drug as a first / second line treatment for Philadelphia chromosome-positive chronic myelogenous leukemia (CML) was a transformative discovery for "targeted" cancer therapies, differing from traditional chemotherapies which were generally broadly toxic in contrast to being designed against a particular target. Gleevec drastically increased patient survival from being nearly always lethal to not significantly different from survival of the average population, the first of its kind for leukemia[4]. The target of this drug, BCR-ABL1 was a mutant protein kinase, kicking off a frenzy of drug discovery surrounding this protein family that lasts to this day[5]. Today, drugs that inhibit the activity of kinases or "Kinase Inhibitors" (KIs) are a central drug class in cancer, with ~40+ approved for clinical oncological use worldwide[6], and have transformed the standard of clinical care for multiple cancer types[5].

Kinases Are the Primary Mode of Cellular Information Transfer

The transfer of a phosphate group from ATP onto proteins forms one of the fundamental methods of regulation for nearly all aspects of cell life in eukaryotes[7]. This simple modification can alter the function of a protein in almost every conceivable way, including activating, deactivating, stabilizing, marking for degradation, or controlling protein-protein interactions. Kinases are a protein enzyme family that catalyze this transfer of phosphate groups for nearly all organisms, numbering around 500 in humans[8]. Originally discovered for their control of metabolism, it is no surprise that kinases now find themselves in the center of all cell physiology, as the primary mediator of information transfer from extracellular stimuli to intracellular events[9]. This mediation is largely achieved through large chains of successive phosphorylation events by individual kinases, often called a "kinase cascade"(**Fig 1**), and usually ending by altering the activity of a transcription factor. These pathways are further parts of large complex interconnected networks that respond dynamically to perturbation[10], activating compensatory pathways in response to single nodes being deactivated. Altogether, this interaction network of kinases that transmit extracellular stimuli intracellularly is commonly referred to as the "Kinome"[11], and it is estimated that about a third of all expressed proteins are phosphorylated by this network at any given time[12]. Phosphorylation also plays a pivotal role in regulating the enzyme activity of kinases, with their activation state (on or off) generally dependent on upstream kinase activities(**Fig 2A**). The collective activation state of all kinases within a cell, referred to as the "Kinome State"[13], is determined by the degree of activation of each individual kinase(**Fig 2B**). As such, phosphorylation acts as a crucial molecular switch in regulating kinase activity and orchestrating cellular signaling pathways.

Kinase Targeting in Cancer

Given the central role of kinases in cell physiology, aberrant kinase activity or "dysregulation" has major implications in immunity, diabetes, neurological disorders, and most cancers[14]. Specifically relevant to cancer, kinases have important roles in the cell cycle, differentiation, motility, apoptosis, DNA repair, and proliferation. For example, mutations in receptor tyrosine kinases (RTK's) that transmit extracellular signals by starting phosphorylation cascades like EGFR, FGFR, ERBB2, and VEGF are biomarkers for various cancer types[15], while dysregulation in intracellular kinase families like the cell cycle and proliferation regulators MAPK, CDK[16], and PI3K are all frequent drivers in tumorigenesis[17].

Drug targeting strategies for these dysregulated kinases have improved steadily over the past two decades since the approval of Imatinib in 2001[5]. While kinases are known to be "highly druggable" through the ATP-binding site[18], structure-guided discovery has led to improvements in drug design, such as the FDA approval of the extremely selective allosteric MEK1/2 inhibitor trametinib[19]. Despite the broad clinical impact of kinase inhibitors, high rates of eventual resistance acquisition continue to prevent robust patient response, even in those who initially respond to treatment[20]. This is in part due to the dynamic nature of the kinome network[10], which is able to respond to perturbation through epigenetic upregulation of alternate pathways that drive cell proliferation. From a kinome network view, this necessitates rational targeting of the kinome to prevent resistance, For example, designing ideal inhibitors or combinations that target complimentary kinases either in the same pathway ("vertical inhibition") or alternate pathways ("horizontal inhibition")[21]. This idea of rational kinome targeting first requires a detailed understanding of the target landscape of kinase targeting drugs, and a kinome-wide view of the effect of each drug.

KINOME STATE PROFILING

The clinical kinase drug crizotinib is a potent inhibitor of ALK, forming the first-line treatment for Non-Small Cell Lung Cancer (NSCLC) that expresses the kinase[22]. However, this drug was originally designed for c-MET, and the optimal mechanism of action, off-target inhibition of ALK, was discovered a year after clinical trials, allowing pre-screening of patients and increased response rates. Similar discoveries were made for the drugs Imatinib for unexpected efficacy in Gastrointestinal Stromal Tumors (GIST) through off-target inhibition of KIT, and midostaurin in lung cancer through off-target EGFR inhibition[23]. Along with off-target induced toxicity[24], These examples illustrate the inherent promiscuity in kinase inhibitor design, since conservation of the ATP-binding site across most kinases will almost always causes large off-target effect effects in designed inhibitors[25]. As a result, there has been increasing interest recently in the profiling of all targets of known kinase inhibitors, both for repurposing potential and general understanding of mechanisms. This en masse view of all kinases and the interaction of drugs with the kinome network is termed "Kinomics", and has been spearheaded by technological improvements in high-throughput measurement of proteomics and phosphoproteomics[26].

Kinome State Measurement

Recent advances in drug profiling technologies have made high-throughput kinome profiling practically feasible through multiple assay types. A landmark paper in 2017 used a mass spectrometry-based assay that used promiscuous kinase-binding compounds immobilized on beads to measure the binding competition between any given inhibitor and any given kinase (henceforth called the "kinobeads" assay)[27](**Fig 3**). Using this assay, the kinome-wide binding profiles for ~230 clinical kinase inhibitors at eight doses each were elucidated using cancer cell lysates, forming the largest in-cell drug-target binding database collected. Similarly, the

"KINOMEscan©" (DiscoverX) assay is a commercially available alternative that performs the same bead-based competition kinome profiling *in vitro*, using recombinantly expressed kinases with a qPCR readout[28]. The data generated from these assays allow interrogation of how clinical and investigational drugs interact with the entire kinome on an unprecedented scale. By analyzing the degree of inhibition of all kinases simultaneously for a given inhibitor, we can treat this as a departure from the "baseline kinome state", thus moving through drug-induced alteration of multiple kinase activities to a "kinome inhibition state". Since we know that the central role of the kinome directs cell behavior on almost all levels, these baseline kinome states and kinome inhibition states can be directly connected to various measured cellular phenotypes[13], and potentially even disease phenotypes.

TOWARDS PRECISION ONCOLOGY

Cancer Heterogeneity Demands Precision Medicine

Cancer is a highly complex disease that displays molecular heterogeneity on two main levels: intratumoral heterogeneity (difference within individual tumors in a patient), and intertumoral heterogeneity(difference between patients with the same tumor type)[29]. Both types result in frequent treatment response failure for chemotherapies, and resistance development for targeted therapies[30]. To increase long-term response rates, recent advances have incorporated patient-specific genomic signatures derived from DNA sequencing or microarrays[31] to prescribe individualized treatment regimens. With technological improvements in tumor molecular profiling, it is now possible to simultaneously profile a given tumor's genome, transcriptome, proteome, gene copy number variation, single-cell genomics[32], and use these global measurements to inform a given patient's treatment plan. This strategy of data-driven, patient-specific treatment is termed "precision medicine", and has been described as moving cancer medicine into a new "Precision Oncology Era"[33]. The

success of BCR-ABL1 mutant based Imatinib treatment in transforming CML into a manageable chronic disease is frequently cited as an early success of this strategy, but only considers a singular genetic mutation that was common knowledge for ~40 years[34]. The exponential increase in data available for each patient, going from considering single genes to now considering multiple layers of measurement across the entire genome, demands the use of cutting-edge computational tools, much like the heterogeneity in cancer demands precision medicine.

Predicting Cancer Treatment Responses

Various computational modeling strategies have been used to predict cancer treatment responses, with most commercially available tools using simple univariate statistical methods that rely on specific biomarkers to predict responses to treatment. With recent advancements in machine learning and AI, many studies have used large public datasets of cancer drug responses to fit complex multivariate models, ranging from linear regression to deep neural networks, attempting to predict the responses of cancer cell lines, organoids, patient-derived xenografts(PDX), and patients themselves. The majority of these approaches use cancer-specific genomic modalities such as baseline gene expression, mutation, and copy number variation to predict treatment responses.

To consolidate the landscape of drug response prediction strategies, the National Cancer Institute (NCI) and the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project held a challenge to accurately predict drug responses for 60 breast cancer cell lines[35]. All participating teams used a combination of genomic, epigenetic, and proteomic data to fit various non-linear regression models, along with pre-filtering the ~100k possible model variables or "features" using "feature selection" techniques to only include informative features, or dimensionality reduction to condense the high-dimensional space[36]. This trimming of all

available features to remove noise is necessary to ensure feasibility of any modeling effort, and has been shown to increase performance of drug response models[37]. The methods employed by the challenge participants achieved reasonable prediction accuracy (modified concordance index ~0.5) but did not include any experimental validation to test if the model predictions could extend into real-world experiments, emphasizing another limitation of current prediction efforts that lack rigorous testing of models on new real-world experiments. Further, these modeling strategies of multi omics-based response prediction have not been restricted to cancer cell lines, as demonstrated in predicting patient-derived organoid (PDO) responses[38], PDX responses[39], and patient survival[40].

Drug-Target Interactions for Response Prediction

More recent attempts to predict drug response have relied on complex neural network architectures to predict responses to both monotherapies and combination therapies, along with increased diversity in data types considered. Notably, newer strategies place increasing emphasis on drug-specific features like molecular fingerprints and drug-target interactions, which serves to add the most relevant layer of information, since drug interactions with proteins are fundamentally what change phenotypes in cells; but on the informatics level it serves an additional function to increase the ratio of measurement to samples, improving the performance of the predictive models[41]. For example, an integrated multi comics approach including compound molecular fingerprints and target protein-protein interaction networks performed relatively well in predicting monotherapy response data[42] (Overall R^2 ~ 0.75), while a high-dimensional tensor-based modeling strategy used similar data and achieved impressive accuracy (Overall R^2 ~ 0.8) in predicting response to combination therapies, validated experimentally[43]. Despite increasing data diversity for drug response prediction, drug-specific data has been largely limited to chemical structure-based descriptors. Incorporating drug-target information has the

potential to further advance prediction by providing insights into the complex interactions between drugs and their targets, particularly if the biological relevance of the targets is well known, like with the 500-member protein kinome. In part, the minimal amount of drug-target information leveraged in current response prediction efforts is because of the sheer amount of data generated by genomics and molecular fingerprinting, generating thousands of features for each measurement, while drug target data has been generally sparse with only a few annotated targets per drug. However, recent advances in technology to profile the interactions of clinical drugs with all the members of the kinome represent an unprecedented ability to measure drug-target information across ~500 proteins simultaneously in a quantitative manner[44]. The breadth, density, and ease of acquisition of this data, often measured at multiple dose points, is ideal for integration into machine learning models that can leverage diverse data types for drug response prediction.

Model Explainability

The ability to explain how and why machine learning models arrive at predictions from the features they consider is termed "Explainability". For linear regression, the importance of a given feature in predicting the outcome is simply its coefficient in the equation[45], and similar metrics exist for tree-based models (decision trees, random forest, gradient boosting). It is important to note that as precision oncology seeks to affect actual clinical decision making, explainability to clinicians is critical for successful integration into medical systems[46]. As the complexity of computational models increases due to efforts to include larger amounts of data, model explainability decreases drastically. As such, traditional regression and tree-based models are still preferred by commercial clinical analysis[47], unlike the most cutting edge neural network and transformer-based models which still suffer from a lack of interpretability, where

potentially predictive features in data are subject to multiple layers of abstraction, earning the "black box" label.

CONCLUSION

The kinome is at the center of human physiology and disease, working as the primary method of information transfer from stimulus to cell response. As cancer therapy is forced into a new era of patient-specific precision by heterogeneity, treatments that target kinases based on cancer-specific genotypes have been some of the key successes in targeted therapeutics. Recent complementary advances in cancer multiomics and machine learning have enabled robust prediction of responses to specific cancer therapeutics, but mostly rely on baseline genomics and molecular fingerprints to make predictions. As the primary means of drug action, protein-drug interactions have clear biological relevance but are underutilized in response prediction efforts partly due to the lack of high-throughput assays that generate data of sufficient breadth. Kinome profiling assays like kinobeads and kinomescan offer an opportunity to measure the effect of a given drug simultaneously on all 500 members of the kinome quantitatively for each dose of a potential drug, generating rich drug-target interactions. The research presented in this dissertation bridges the gap in bioinformatics between cancer-specific multiomics and drug-specific interactions by using the kinome to build machine learning models that predict cancer responses to kinase inhibitor monotherapies and combination therapies. Together, these chapters support the development of kinome-informed computational drug screening platforms in cancer cell lines, and precision oncology platforms in PDX tumors.

FIGURES

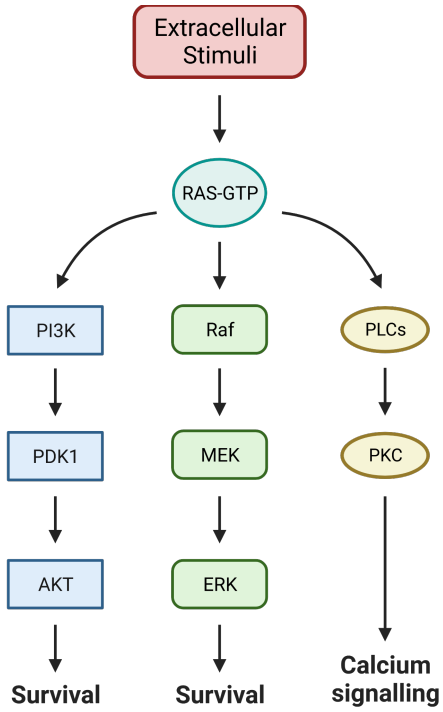


Figure 1.1: Schematic of Example Kinase Cascades. Illustration showing the information transfer chain from extracellular stimuli to a kinase cascade beginning with PI3K, RAF, and PLC respectively, via the mediator RAS. Each arrow thereafter represents a phosphorylation event until reaching the final cellular phenotype.

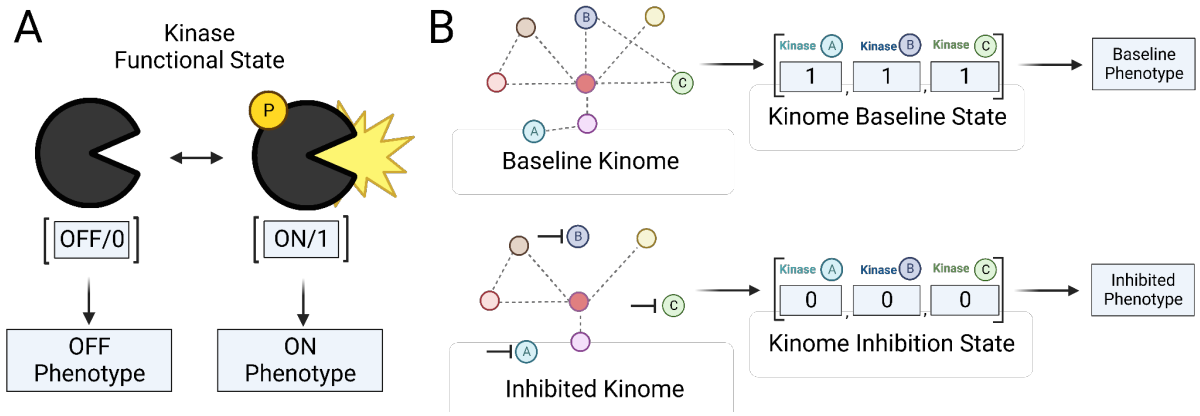


Figure 1.2: Schematic of Single Kinase Activation and Kinome Inhibition **A**) Simplified illustration showing a single kinase and its two possible states ("ON" or "OFF"), connected to a related downstream phenotype. **B**) Illustration showing the kinome network at its baseline state (upper) and inhibited state (lower) in response to drug treatment. Each node in the network represents a kinase, and each kinase has an activation state (1 = ON, 0 = OFF). After drug treatment, kinases A, B, C are shown to be inhibited, resulting in the "Inhibited Phenotype" shown downstream.

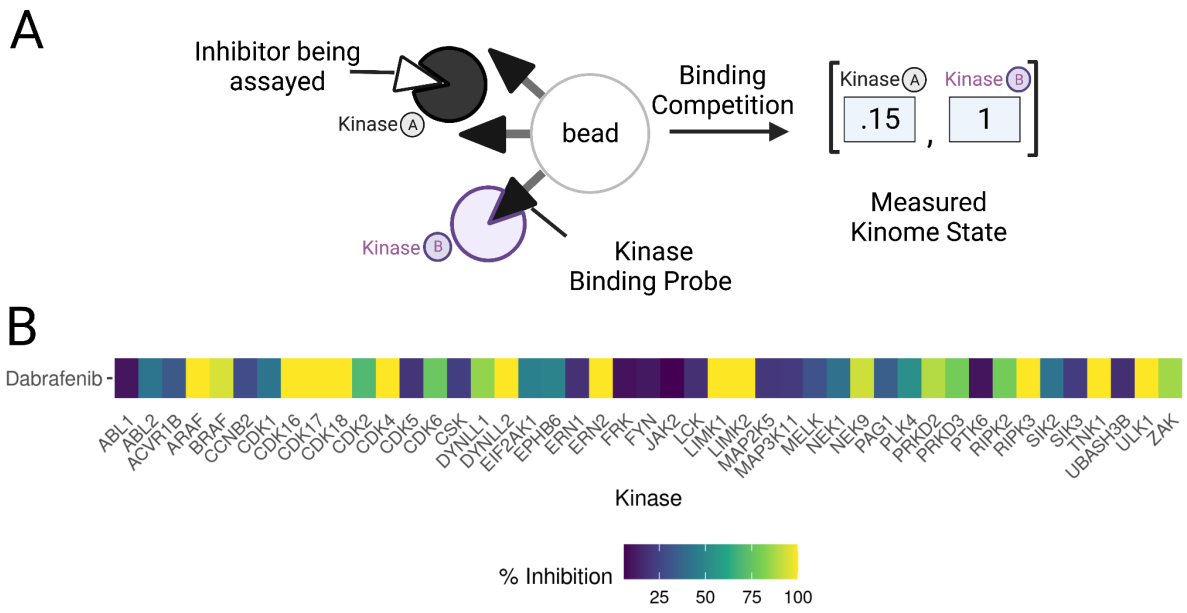


Figure 1.3: Kinome Profiling Through the Kinobeads Assay. A) Illustration of the kinobeads assay: binding strength is measured through competition of assayed inhibitor and kinase binding probes. **B**) Visualization of the actual kinome profile of RAF inhibitor Dabrafenib, measured through the kinobeads assay. Kinases inhibited are represented on the x axis, with color representing the degree of inhibition (values in % Inhibition).

REFERENCES

1. New Hope For Cancer. Time. Available: https://content.time.com/time/magazine/article/0,9171,999978,00.html. Accessed 20 Mar 2023.

2. Sporn MB. The war on cancer: a review. Ann N Y Acad Sci. 1997;833: 137–146.

3. Koretzky GA. The legacy of the Philadelphia chromosome. J Clin Invest. 2007;117: 2030–2032.

4. Gambacorti-Passerini C, Antolini L, Mahon F-X, Guilhot F, Deininger M, Fava C, et al. Multicenter independent assessment of outcomes in chronic myeloid leukemia patients treated with imatinib. J Natl Cancer Inst. 2011;103: 553–561.

5. Cohen P, Cross D, Jänne PA. Kinase drug discovery 20 years after imatinib: progress and future directions. Nat Rev Drug Discov. 2021;20: 551–569.

6. Essegian D, Khurana R, Stathias V, Schürer SC. The Clinical Kinase Index: A Method to Prioritize Understudied Kinases as Drug Targets for the Treatment of Cancer. Cell Rep Med. 2020;1: 100128.

7. Cohen P. The origins of protein phosphorylation. Nat Cell Biol. 2002;4: E127-30.

8. Yu J-S. From discovery of tyrosine phosphorylation to targeted cancer therapies: The 2018 Tang Prize in Biopharmaceutical Science. Biomed J. 2019;42: 80–83.

9. Cohen P. The role of protein phosphorylation in neural and hormonal control of cellular activity. Nature. 1982;296: 613–620.

10. Duncan JS, Whittle MC, Nakamura K, Abell AN, Midland AA, Zawistowski JS, et al. Dynamic reprogramming of the kinome in response to targeted MEK inhibition in triple-negative breast cancer. Cell. 2012;149: 307–321.

11. Zhang H, Cao X, Tang M, Zhong G, Si Y, Li H, et al. A subcellular map of the human kinome. Elife. 2021;10. doi:10.7554/eLife.64943

12. Hogan JM, Pitteri SJ, McLuckey SA. Phosphorylation site identification via ion trap tandem mass spectrometry of whole protein and peptide ions: bovine alpha-crystallin A chain. Anal Chem. 2003;75: 6509–6516.

13. Berginski ME, Joisa CU, Golitz BT, Gomez SM. Kinome inhibition states and multiomics data enable prediction of cell viability in diverse cancer types. PLoS Comput Biol. 2023;19: e1010888.

14. Albanese SK, Parton DL, Işık M, Rodríguez-Laureano L, Hanson SM, Behr JM, et al. An Open Library of Human Kinase Domain Constructs for Automated Bacterial Expression. Biochemistry. 2018;57: 4675–4689.

15. Pottier C, Fresnais M, Gilon M, Jérusalem G, Longuespée R, Sounni NE. Tyrosine Kinase Inhibitors in Cancer: Breakthrough and Challenges of Targeted Therapy. Cancers . 2020;12. doi:10.3390/cancers12030731

16. Scheiblecker L, Kollmann K, Sexl V. CDK4/6 and MAPK-Crosstalk as Opportunity for Cancer Treatment. Pharmaceuticals . 2020;13. doi:10.3390/ph13120418

17. Yang J, Nie J, Ma X, Wei Y, Peng Y, Wei X. Targeting PI3K in cancer: mechanisms and advances in clinical trials. Mol Cancer. 2019;18: 26.

18. Xu D, Jalal SI, Sledge GW, Meroueh SO. Small-molecule binding sites to explore protein-protein interactions in the cancer proteome. Mol Biosyst. 2016;12: 3067–3087.

19. Hoffner B, Benchich K. Trametinib: A Targeted Therapy in Metastatic Melanoma. J Adv Pract Oncol. 2018;9: 741–745.

20. Gillis NK, McLeod HL. The pharmacogenomics of drug resistance to protein kinase inhibitors. Drug Resist Updat. 2016;28: 28–42.

21. Yesilkanal AE, Johnson GL, Ramos AF, Rosner MR. New strategies for targeting kinase networks in cancer. J Biol Chem. 2021;297: 101128.

22. Blackhall F, Cappuzzo F. Crizotinib: from discovery to accelerated development to front-line treatment. Ann Oncol. 2016;27 Suppl 3: iii35–iii41.

23. Red Brewer M, Pao W. Maximizing the benefits of off-target kinase inhibitor activity. Cancer discovery. 2013. pp. 138–140.

24. Grossman M, Adler E. Protein Kinase Inhibitors - Selectivity or Toxicity? In: Singh RK, editor. Protein Kinases. Rijeka: IntechOpen; 2021.

25. Nicholson GC, Holloway RA, Leaker BR, Kilty I, Salganik M, Tan L, et al. A novel flow cytometric-based method to measure kinase inhibition in sputum from COPD subjects. BMJ Open Respir Res. 2016;3: e000140.
26. Graves LM, Duncan JS, Whittle MC, Johnson GL. The dynamic nature of the kinome. Biochem J. 2013;450: 1–8.

27. Klaeger S, Heinzlmeir S, Wilhelm M, Polzer H, Vick B, Koenig P-A, et al. The target landscape of clinical kinase drugs. Science. 2017;358. doi:10.1126/science.aan4368

28. Fabian MA, Biggs WH 3rd, Treiber DK, Atteridge CE, Azimioara MD, Benedetti MG, et al. A small molecule-kinase interaction map for clinical kinase inhibitors. Nat Biotechnol. 2005;23: 329–336.

29. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol. 2018;15: 81–94.

30. Bhang H-EC, Ruddy DA, Krishnamurthy Radhakrishna V, Caushi JX, Zhao R, Hims MM, et al. Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. Nat Med. 2015;21: 440–448.

31. The Lancet. 20 years of precision medicine in oncology. Lancet. 2021;397: 1781.

32. Heo YJ, Hwa C, Lee G-H, Park J-M, An J-Y. Integrative Multi-Omics Approaches in Cancer Research: From Biological Networks to Clinical Subtypes. Mol Cells. 2021;44: 433–443.

33. Wahida A, Buschhorn L, Fröhling S, Jost PJ, Schneeweiss A, Lichter P, et al. The coming decade in precision oncology: six riddles. Nat Rev Cancer. 2023;23: 43–54.

34. Pc N. A minute chromosome in human chronic granulocytic leukemia. Science. 1960;132: 1497.

35. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol. 2014;32: 1202–1212.

36. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. J Open Source Softw. 2018;3: 861.

37. Koras K, Juraeva D, Kreis J, Mazur J, Staub E, Szczurek E. Feature selection strategies for drug sensitivity prediction. Sci Rep. 2020;10: 9377.

38. Lee J-K, Liu Z, Sa JK, Shin S, Wang J, Bordyuh M, et al. Pharmacogenomic landscape of patient-derived tumor cells informs precision oncology therapy. Nat Genet. 2018;50: 1399–1411.

39. Zwep LB, Duisters KLW, Jansen M, Guo T, Meulman JJ, Upadhyay PJ, et al. Identification of high-dimensional omics-derived predictors for tumor growth dynamics using machine learning and pharmacometric modeling. CPT Pharmacometrics Syst Pharmacol. 2021;10: 350–361.

40. Sousa A, Dugourd A, Memon D, Petursson B, Petsalaki E, Saez-Rodriguez J, et al. Pan-Cancer landscape of protein activities identifies drivers of signalling dysregulation and patient survival. Mol Syst Biol. 2023;19: e10631.

41. Adam G, Rampášek L, Safikhani Z, Smirnov P, Haibe-Kains B, Goldenberg A. Machine learning approaches to drug response prediction: challenges and recent progress. NPJ Precis Oncol. 2020;4: 19.

42. Cheng X, Dai C, Wen Y, Wang X, Bo X, He S, et al. NeRD: a multichannel neural network to predict cellular response of drugs by integrating multidimensional data. BMC Med. 2022;20: 368.

43. Julkunen H, Cichonska A, Gautam P, Szedmak S, Douat J, Pahikkala T, et al. Leveraging multi-way interactions for systematic prediction of pre-clinical drug combination effects. Nat Commun. 2020;11: 1–11.

44. Cann ML, McDonald IM, East MP, Johnson GL, Graves LM. Measuring Kinase Activity-A Global Challenge. J Cell Biochem. 2017;118: 3595–3606.

45. Bazdaric K, Sverko D, Salaric I, Martinovic A, Lucijanic M. The ABC of linear regression analysis: What every author and editor should know. Esencia Odontol. 2021;47: e63780.

46. Yoon CH, Torrance R, Scheinerman N. Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? J Med Ethics. 2022;48: 581–585.

47. Kundu S. AI in medicine must be explainable. Nat Med. 2021;27: 1328.

CHAPTER 2: KINOME INHIBITION STATES AND MULTIOMICS DATA ENABLE PREDICTION OF CELL VIABILITY IN DIVERSE CANCER TYPES

INTRODUCTION

While chemotherapy remains a mainstay in cancer treatment, the use of targeted therapies clearly holds significant promise, with their use leading to improved outcomes in a variety of cancers [1,2]. Examples include the use of imatinib (Gleevec) for chronic myelogenous leukemia, crizotinib and other anaplastic lymphoma kinase (ALK) inhibitors for non-small-cell lung cancers, and trastuzumab and lapatinib for ERBB2/HER2 amplified breast cancers [3–8]. Together with the potential to reduce toxicity and associated side effects, the development of targeted therapies has gained increasing momentum over the last two decades [9,10].

Since the development of imatinib, protein kinases have emerged as a primary focus for targeted therapy development [11–14]. Kinases are a ~500-member enzyme family that catalyzes the transfer of phosphate groups from ATP to specific substrates. Integrated into a complex network of interactions defined as the kinome, kinases regulate information transfer across a myriad of cellular processes including growth, proliferation, differentiation, motility, and apoptosis [15]. Linked to its role in this wide array of functions, dysregulation of one or more members of the kinome is directly implicated in numerous pathologies, especially cancer [16]. Modulation of kinase activity through targeted inhibition has been the primary therapeutic approach to date and as of 2021, over 85 kinase inhibitors have been clinically approved worldwide, though only targeting 42 kinases from the 21 kinase families[17], highlighting the opportunity for further advancement of this large family of druggable targets.

Recent work characterizing kinome behavior in response to targeted kinase inhibitor therapies has established that the kinome is a highly dynamic system, with significant ramifications in our understanding of drug resistance, adaptive reprogramming and the broader design of effective therapies [18–23]. Underlying these investigations of kinome dynamics are the advancement of proteomic approaches that enable the characterization of protein kinome behavior in response to perturbation en masse, allowing characterization of changes not just to the kinase to which the inhibitor was designed, but also across the entire kinome [24,25]. However, while providing transformative insight into how these targeted therapies interact with and modify cellular systems, our understanding of kinome changes and the resulting downstream cellular changes is still lacking.

Given the potential of targeted therapies and the potential to quantitatively assess their effect on the protein kinome, in this work we sought to establish a predictive framework that links the behavior of the kinome as defined by "kinase inhibition states" with a downstream phenotype - in this instance, cell viability. Enabling this effort is recent work by Klaeger et al., who conducted a comprehensive investigation using a proteomic kinobead approach, establishing a target landscape for 229 kinase inhibitors across a wide range of compound concentrations [26]. This work was conducted using a lysate mixture derived from four cell lines which provided a broad representation of the kinome. The results from Klaeger et al. show that many kinase inhibitors have broad target promiscuity and that the kinases targeted by each inhibitor also varies on the basis of the specific compound concentration. Throughout the rest of this paper, we will use the phrase kinase inhibition state to indicate the specific set of kinases targeted by a given compound and to what degree each kinase is inhibited at each concentration. In addition, we utilized the extensive data available via the Broad Institute's Cancer Cell Line Encyclopedia (CCLE) [27], including the PRISM (Profiling Relative Inhibition Simultaneously in

Mixtures) highly multiplexed cell viability assay, along with accompanying multi-omics data (gene expression, copy number variation, proteomics and gene essentiality) from the Cancer Dependency Map. These data consist of cell viability measurements for 499 cell lines across 1448 drugs, transcriptomic profiles for 1389 cell lines, whole proteomic profiles for 375 cell lines, whole genome copy number variation for 1750 cell lines and CRISPR-KO genetic dependency scores for 1054 cell lines. While predictive models for drug-induced cell viability have been built using various strategies [28–31], most have focused on using baseline and drug-perturbed transcriptomic data to make predictions on the sensitivity of cancer cell lines to drugs. Drug-target interaction data like kinome profiles are relatively underutilized in these approaches, but have been shown to have predictive power in smaller datasets[32].

Here, we describe a framework that integrates kinome profiling data with general multi-omics and build tree-based regression models to predict cell viability for 480 cancer cell lines across 230 kinase inhibitors with high accuracy ($R^2 = 0.79$). Integrating nearly half a million data points, we find that kinome inhibition profiles have by far the greatest predictive power of any single data set. While not highly predictive on its own, baseline transcriptomic data does significantly enhance prediction accuracy, "tuning" the model to individual cell lines. Remarkably, adding in other multi-omics data does not significantly increase the quality of predictions. As the model enables prediction of complete dose-response curves, we experimentally validate predictions for over two dozen compounds on two breast cancer cell lines and find strong agreement for most compounds tested. These results suggest that the link between kinotype and phenotype is significant, with sufficient power to enable the prediction of cell viability and likely other cellular phenotypes as well. Along with integration of transcriptional data, these predictive models can greatly enhance our understanding of adaptive kinome

reprogramming and drug resistance while facilitating the development of future targeted therapy regimes.

RESULTS

This work is divided into three parts. We start by describing how we processed and organized the data sets used to build predictive models of cell viability related to a set of kinase inhibitors. Next, we describe the methods we used to select which features and data sets to include in these models and apply a set of modeling methods to the organized data. Finally, we make a set of cell viability predictions and then experimentally test these predictions in a panel of breast cancer cell lines.

Linking Kinome Inhibiton States with Cancer Cell Viability

There are two primary data sources that we needed to process and combine in order to link kinotype with phenotype and build a model to predict the cell viability effects of kinase inhibitors. The first of these data sources is the large-scale PRISM cell viability screening effort. The PRISM data collection consists of a set of cell line viability measurements following exposure to a wide range of compounds[33] (Figure 1A). These compounds span multiple different target classes, but in this work we have focused on a specific subset of kinase inhibitors that have been independently assayed using the kinobead/MS-based method. This approach determines the precise kinase targets as well as the magnitude of their inhibition in response to different concentrations of the inhibitors[26]. Given that the compounds used in Klaeger et al. are all well known kinase inhibitors, most of the proteins that appear in the assay results are either known kinases or closely associated proteins. As such, we'll refer to the data originating from the Klaeger et al. result as "kinase inhibition states."

The primary challenge with combining these data sets is a lack of overlap between some of the concentrations used in the PRISM assay and those used by Klaeger et al. To overcome

this problem, we used the viability curve fits provided by the PRISM database and imputed cell viability values for all of the concentrations used by Klaeger et al (Figure 1B). These cell viability results are represented as a value from 0-1, with 0 indicating complete cell death and 1 indicating no effect on cell viability. As expected, a majority of the treatments yielded little change in cell viability (Figure 1C). The distribution of cell viability values within each individual compound shows that while many of the compounds have minimal effects on cell viability, some compounds show a much wider range of viability effects (Figure 1D).

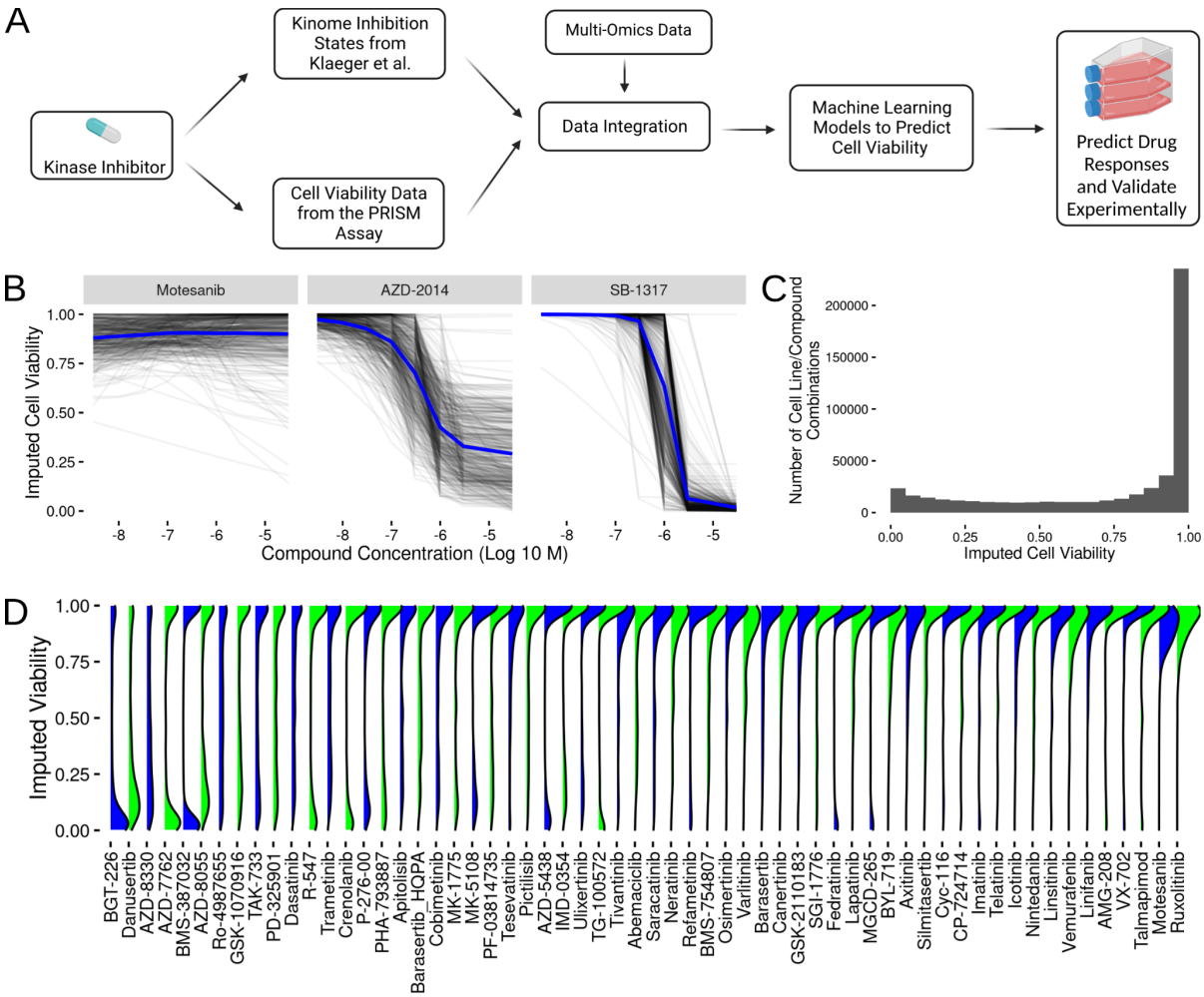


Figure 2.1: Study Design Overview and Imputation of Cell Viability from PRISM. (A) Flow chart showing data source collection, integration and modeling strategy. (B) Sample imputed cell viability curves for all assayed cell lines (gray underlying lines) and corresponding average imputed cell viability response (blue line) for three compounds showing low changes (motesanib), medium level changes (AZD-2014) and high changes (SB-1317) in cell viability. (C) Overall distribution of cell viability values imputed at Klaeger et al compound concentrations. (D) Distribution of imputed cell viability across all concentrations for a selection (60 out of 168) sampled evenly across the average imputed cell viability effect of the compounds present in both PRISM and the Klaeger et al set. The blue and green color scheme does not indicate anything about the underlying data and is meant to act as a visual aid for differentiating between adjacent curves. (Panel A was created with BioRender.com)

After combining the PRISM and Klaeger et al. data sets, we have 168 compounds which have been assayed across 480 cell lines. We imputed the cell viabilities at each of the 8 concentrations used in the Klaeger et al. work, yielding about half a million treatment combinations across combinations of cell line, compound and concentration. With this data set, we also integrated the gene expression data available through the Cancer Cell Line Encyclopedia[34]. These gene expression values (log2 TPM values with a pseudocount of 1) were available in all but four of the 480 lines used in the PRISM compound screens. Following the integration of gene expression, we next examined how well single kinase inhibition and gene expression values were correlated with cell viability.

Cell Viability After Treatment with Kinase Inhibitors Shows Mild Correlation with Kinase Inhibition State

We investigated the relationship between kinase inhibition states (~520 proteins) and gene expression values with inhibitor-induced cell viability. To do this, we took each individual kinase inhibition state and gene expression value (~21,000 TPM values) and calculated the Pearson's correlation coefficient with the imputed cell viabilities (Figure 2A,B). The kinase inhibition states from Klaeger et al. are represented as a value lying mostly between zero and one, where zero indicates a fully inhibited kinase and values of one or above indicate that a kinase isn't inhibited. These correlations were in general significantly lower for the gene expression values, while the kinase inhibition state values showed both a higher average correlation and greater variance (Figure 2C). This was not unexpected as the gene expression values are all characterized in unperturbed cell lines. Thus, as cell viability changes the gene expression values remain fixed, and any variation across gene expression must be correlated with broad changes in drug response between the cell lines. The examination of single correlation values gives a picture of how well single expression or inhibition states are related to the cell viability phenotype.

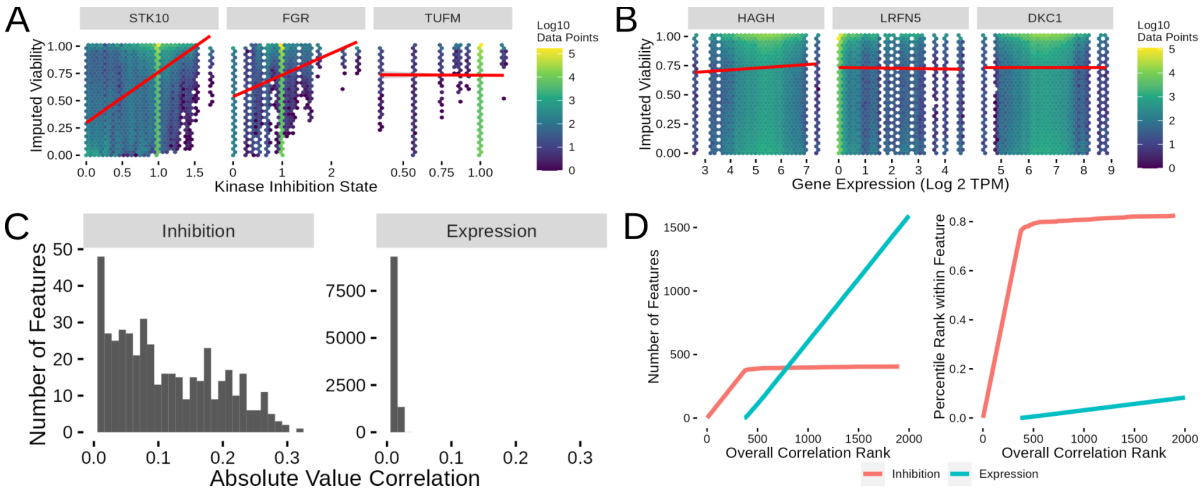


Figure 2.2: Single Feature Correlations Across Kinase inhibition and Gene Expression.

(A) Sample kinase inhibition state versus imputed viability heatmap plots showing inhibition states with high (STK10), medium (FGR) and low (TUFM) correlation values. (B) Sample gene expression versus imputed viability heatmap plots showing genes with high (*HAGH*), medium (*LRFN5*) and low (*DKC1*) correlation values. (C) Overall distribution of correlations between kinase inhibition states and gene expression levels. (D) Plots showing what order classes of features are selected from the inhibition and expression correlations. The number of features from each class (left) selected at a given rank value and the percentage of the possible features (right) selected at a given feature selection rank cutoff.

While single features with correlation coefficient values in the ~0.3 range (the highest value observed in the kinase inhibition state data) will not produce sufficiently predictive models, the integration of multiple features may provide greater power. As such, we next sought to use the correlation values for feature selection. The most obvious way to use the correlation values is to put all the potential features (in this case, kinase inhibition state and gene expression) in correlation rank order and then select the top-X number of features for model inclusion. This produces differing sets of feature class counts and ratios depending on the number of features selected (Figure 2D right). Interestingly, the top ~350 features all come from the kinase inhibition states, with gene expression then starting to be included into the list after the first 350 features. As an alternative method to visualize the same selection process, we plotted what percent of a given feature class is included in the top list for the top 2000 features (Figure 2D left). This alternative view of the feature selection process shows that ~80% of the inhibition states are included in the model before gene expression starts to be included. This indicates that nearly all of the inhibition states are more highly correlated than the gene inhibition states and will thus be the sole factor utilized in lower feature count models. Extending the feature list visualization to include lists greater than 2000 show that remaining inhibition states are slowly included as the top feature list expands (Figure S1A). This analysis of the structure of the single feature correlation results lays the groundwork for working with more sophisticated computational models to predict cell viability.

Computational Models Can Predict Cell Viability from a Combination of Kinase Inhibition State and Gene Expression

With our initial analysis of the predictive power of single features from the Klaeger and gene expression data sets completed, we next moved to the development of models that integrated more than one feature with the end goal of predicting cell viability. To do this, we tested four types of models: linear regression, random forest, TabNet and XGBoost. For our initial tests with these models, we used the default settings for all four model types and varied the number of features (either kinase inhibition states or gene expression values) provided to the model. Our cross-validation strategy sought to balance our eventual goals of using the resulting models to make predictions about the cell viability effects in new cell lines and in untested compounds. As such, we choose a 10-fold cross validation strategy that randomized fold exclusion across the cell line-compound treatments (63767 total combinations) to improve the likelihood that our model testing results would be similar to downstream experiments. After producing the cross-validation splits, we selected a specific number of features and built corresponding models (Figure 3A).

For benchmarking model performance, we built a naive model that simply used the average cell viability at each of the tested concentrations as a baseline prediction that can be used for comparison (gray dotted lines in Figure 3A), and also compared results to previously run models on similar datasets[31]. Initially, we tested each model type with 100, 200, 300, 400, 500, 1000, 1500 and 2000 features. These preliminary tests showed that the random forest method performed the best at all of these feature counts and that performance (R^2 and RMSE) peaked at 500 features and out-performed our baseline dose-concentration-only model. To ensure that we had indeed found the peak in feature performance, we then tested 600, 700, 800 and 900 feature models and found that the 500-feature model was the peak (although all of these models

performed very similarly). To better understand this model, we also looked more closely at the predicted versus actual imputed viability of the 500-feature random forest model (Figure 3B). This examination of the cross-validation model results showed that the average model performance was best at higher imputed viability values, while the predictions at lower imputed viabilities were not as accurate. In addition to examining the global model performance we also subsetted the results along compound and cell line results and re-calculated R^2 and RMSE (Figure S2A). This result showed that the compound results showed greater variability in R^2 as compared to the cell line results, but the RMSE values were similarly distributed.

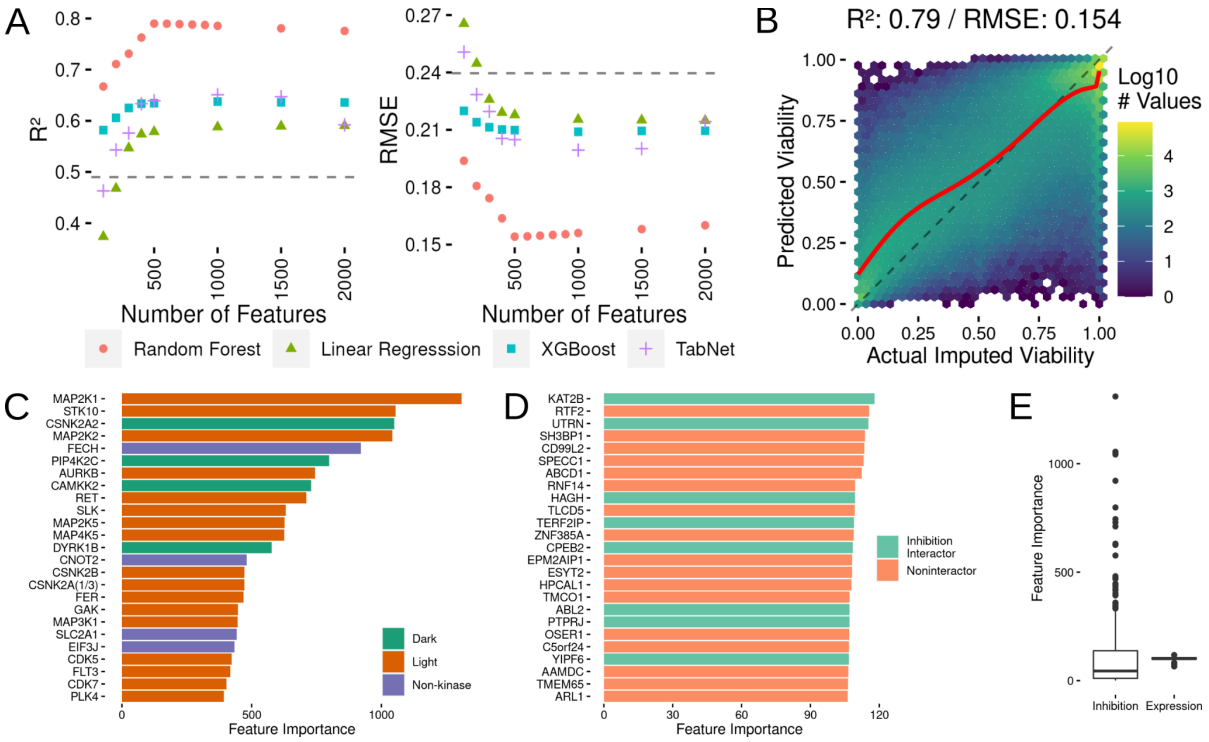


Figure 2.3: Development of a Regression Model to Predict Cell Viability and Assessment of Which Features Contribute to Model Predictions. (A) Comparison of R^2 and RMSE values from linear regression, random forest, XGBoost and TabNet models. The gray dotted line shows the performance level of a dose-only model performance. (B) Actual imputed viability versus cross validated model predictions for the random forest model. The dashed line indicates where a perfect set of predictions would appear, while the red line shows a loess fit through the actual results. (C) Variable importance plot for the top 25 features in the final regression model. Each feature is prefixed with act or exp representing either kinase inhibition or gene expression respectively. (D) The top 25 most important expression features in the final model. (E) The overall distributions of feature importance values for the inhibition and expression features.

With random forest using 500 features selected as our best modeling strategy, we moved on to examining feature selection in the cross-validation models as well as parameter tuning. One concern with doing feature selection in each cross-validation set was that there would be a large amount of volatility in feature selection between each cross-validation model run. We found that in each of the cross validation runs, at least 75% of the features are included in all of the feature selection sets (Figure S2B). To ensure that the default random forest parameter models were near the optimal tuning, we also tested cross-validated models with 1000, 1500 and 2000 trees (500 trees is the default value). Increasing the tree count had little effect on model quality (Figure S2C), so we opted to use the default value of 500 trees. In addition, we also tested the effect of modifying the minimal leaf node size and the number of predictors selected at each branch (Figure S2D) and found minimal effects on R^2 and RMSE, so we once again decided to keep the default parameter values.

Our first step in building the final kinase inhibition and gene expression model was to first select the 500 features that would be included in the model. Using the same correlation ranking scheme used in our cross validated models, 390 out of 520 kinase inhibition states and 110 out 19177 gene expression features were selected for model inclusion. We next built the final random forest model with the full data set and collected variable importance metrics for each of the included features. In order to understand the kinase and non-kinases included in the selected inhibition states, we classified each protein as either a non-kinase or as a well-studied (Light) or understudied (Dark) kinase (Figure 3C)[35]. Several of these genes have well-known roles in cell viability and cancer, including MAP2K1 (MEK1), AURKB and CDK7. Interestingly, the model also identifies several understudied kinases, CSNK2A2, PIP4K2C, CAMKK2 and DYRK1B, as being influential in the model's cell viability predictions. To better contextualize the expression values included in the model, we used the STRING database to see how many of the selected genes interacted with the proteins included in the inhibition features (Figure 4D). Of the 110 genes included in the expression values, 40 interact with at least one protein in the inhibition set and the average expression gene interacts with 1.7 inhibition state genes. In comparison to 10,000 randomly drawn expression gene sets of size 110, 84% interact with fewer than 40 inhibition states and 80% have a lower average inhibition gene interactor count below 1.7. The global view of the variable importance metrics also shows that nearly all of the expression features have similar importance values in the final random forest model (Figure 3E). We next attempted to better understand how the interaction between inhibition states and gene expression levels affected model performance.

The Combination of Kinase Inhibition States and Baseline Gene Expression Produces the Best Predictions

After thoroughly examining the results of the inhibition state and gene expression combined model, we next wanted to investigate how the model would perform when we excluded certain parts of the full data set. Using the same 10-fold feature selection cross validation strategy and the same cross validation fold splits described above, we rebuilt the model using only inhibition state or only gene expression (Figure 4A). The gene-expression-only models performed very poorly (R^2 of ~0.01 and RMSE of 0.33), which was expected due to the fact that the gene expression values are fixed and do not vary with the compound concentrations. These model performance differences were also reflected in direct comparisons between individual cell line and compounds, where none of the expression-only models outperformed the inhibition state only models. When we built models using the inhibition states alone, we observed identical performance for feature counts 300 and below. This was also expected as the correlational feature selection methods always select inhibition features for the first ~350 features. With feature counts of 400 and 500, we observed that the additional information provided by the gene expression features began to improve the model (0.05 improvement in R^2 and a 0.02 decrease in RMSE). Thus, while the expression features alone are not sufficient to predict cell viability, they do provide an appreciable improvement in the model performance in combination with inhibition features.

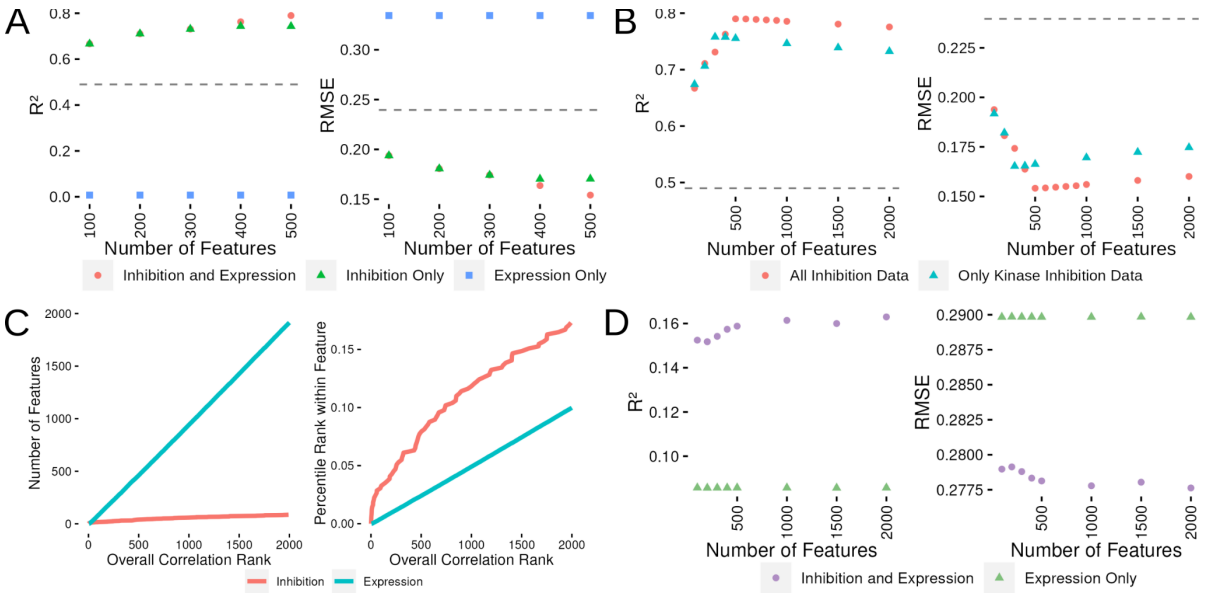


Figure 2.4: Model Performance is Best with Access to All Inhibition States and Gene Expression Values. (A) Comparison of R^2 and RMSE performance for models using only expression, only inhibition or inhibition and expression features. (B) Comparison of R^2 and RMSE performance for models using gene expression and all inhibition data or only the kinase subset. (C) Plots showing the order of feature selection for the single dose model. (D) Single dose model performance comparison across a range of feature count and with either kinase inhibition state and expression or expression alone.

Having established that both inhibition and expression data are needed for the best model performance, we next investigated how the non-kinases in the inhibition data set affected model performance. This question is an interesting avenue to explore as, while the Klaeger et al. study was confined to kinase inhibitors, the presence of ~50% non-kinase proteins inspired us to assess how the model would perform without the non-kinases. We rebuilt the inhibition data set and ran the same modeling methodology including the gene expression values to allow us to compare to our previous models (Figure 4B). The optimum kinase-only inhibition data model had a maximum R^2 of 0.76 and a RMSE of 0.17 (compared to R^2 of 0.79 and RMSE of 0.15 for the full set). These results indicate that the non-kinases are providing some additional information that the model is able to use, which is in agreement with the presence of non-kinases in the top 25 of the variable importance metrics (Figure 3C).

To further investigate whether the kinase inhibition states are more informative than gene expression values alone, we subset our data to only include the dose for each compound with the highest variation in viability. This is following from a previous publication [36] which built a wide range of models covering chemical and genetic perturbations. By subsetting the data in this fashion, we can more easily compare the relative contributions of kinase inhibition state and gene expression without the variation induced by multiple doses. We used the same feature selection methodology as in the previous section and the shift to only a single dose for each compound generally decreased the kinase inhibition correlations. This allowed more expression values to be included in the model (Figure 4C). After conducting feature selection, we built a set of random forest models with differing numbers of features and found that the kinase inhibition state and expression models outperformed models built with expression data alone. This result demonstrates that even in a more constrained modeling environment, the availability of proteomic based inhibition profiles improves model quality for kinase inhibitors and that the

additional information provided by multiple doses can improve modeling results. This is also a limitation though as data comparable to kinase inhibition state does not exist for many classes of compounds, so we view this work as complementary to the broader modeling efforts of Dempster et al. Having fully examined the kinase inhibition state and expression model, we next investigated if any of the other multiomics data sets available could improve upon these models.

Models Only Show Mild Improvement from Inclusion of a Broad Spectrum of Omics Data

Gene expression is only one of several different types of comprehensive data that has been collected for many of the cell lines used in the PRISM assay. These additional data sets include:

- DepMap CRISPR-KO screening: genome-wide gene knockout viability measurements (DepMap Score)
- Copy-number-variation: gene level copy number variation (CNV)
- Whole Genome Proteomics: mass spectroscopy-based measurement of relative protein abundance (proteomics)

Given the broad and complementary nature of these data sets, we investigated whether we could integrate these data sets to improve upon the kinase inhibition and gene expression models we described above. The Depmap, CNV and proteomics data sets all overlap with a different number of cell lines present in the PRISM data set (Figure S3A). All of the data sets are available for 212 cell lines (gene expression is available for 476 cell lines represented in PRISM). We focused our modeling efforts on these 212 cell lines to ensure that a complete collection of data was available. We followed the same strategy as in the above modeling effort where we first

investigated the correlation between single features and cell viability. The 212 cell line subset showed very similar correlation distributions between kinase inhibition and gene expression (Figure S3B). The newly added feature (CNV, DepMap scores and proteomics) correlations, had correlation distributions very similar to gene expression (Figure S3B). Using the correlation feature ranking, we also determined which features would be included in models of various sizes (Figure 5A and Figure S3C). With these data sets organized and our feature selection techniques specified, we tested how inclusion of these data sets affected model quality.

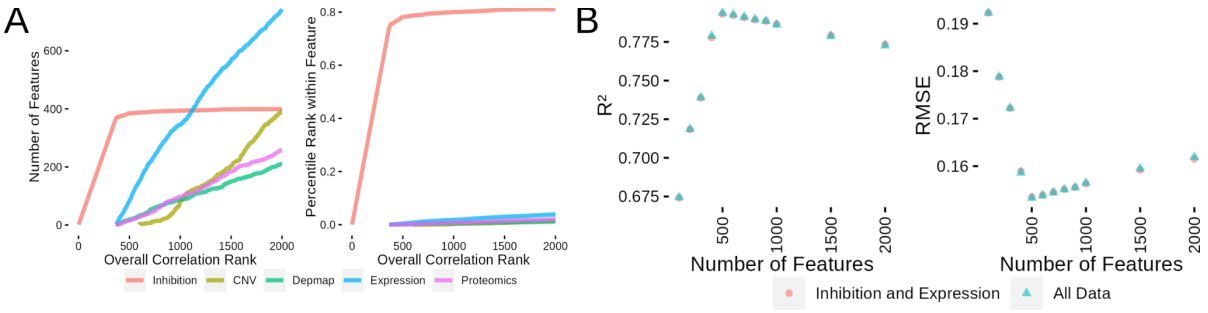


Figure 2.5: Regression Models using Additional Data Sets Don't Dramatically Outperform inhibition and Expression Models. (A) Plot showing the order features are selected for model inclusion (left) and the percentile rank within each feature class as features are selected for inclusion in the model (right). (B) Comparison between models built with inhibition data and expression or all available data sets by R^2 (left) and by RMSE (right).

Based on our previous experience with building the kinase inhibition and expression models, we decided to only test the best-performing random forest method. We also used the same 10-fold cross validation across the cell line/compound combinations. This resulted in higher instability in feature inclusion across the cross-validation folds (Figure S3D). As shown in Figure 5B, integration of these other data sets led to performance that was nearly identical to the model with only kinase inhibition and gene expression. The peak performance was achieved at

500 features in both model variants with R^2 values of 0.794 (0.153 RMSE) and 0.793 (0.154 RMSE) for the all data and inhibition/expression models respectively. This indicates that gene expression values alone contain substantially similar information as the remaining set of multiomics data. Given our desire to build a model which uses the most easily reproducible data sets and only minor improvements were observed with the full data collection, we decided to move forward with the integrated model combining kinase inhibition states and gene expression values.

Validating the Models was Successful within Our Ability to Replicate Previous PRISM Results

With the model production decisions finalized, we then applied this model to the untested cell line and compound combinations. The final model was produced using the 63189 cell line and compound combinations with interpolated viability values (Figure 6A). Of the data that went into model production, 476 cell lines and 168 compounds were represented. This left 903 cell lines in the CCLE gene expression data set and 61 Kleager kinase inhibitors that have not been tested in the PRISM viability assays (in addition to a few other untested combinations) where we were able to apply our model to predict cell viability at each of the compound concentrations used in the Klaeger assay. Ultimately, this resulted in us producing predictions for about 250,000 cell line and compound combinations (Sup Data 1). We hope that providing these prediction results will enable other researchers to find interesting or unexpected compounds that target specific cancer types. For the work presented here, we focused our validation efforts on a subset of breast cancer cell lines.

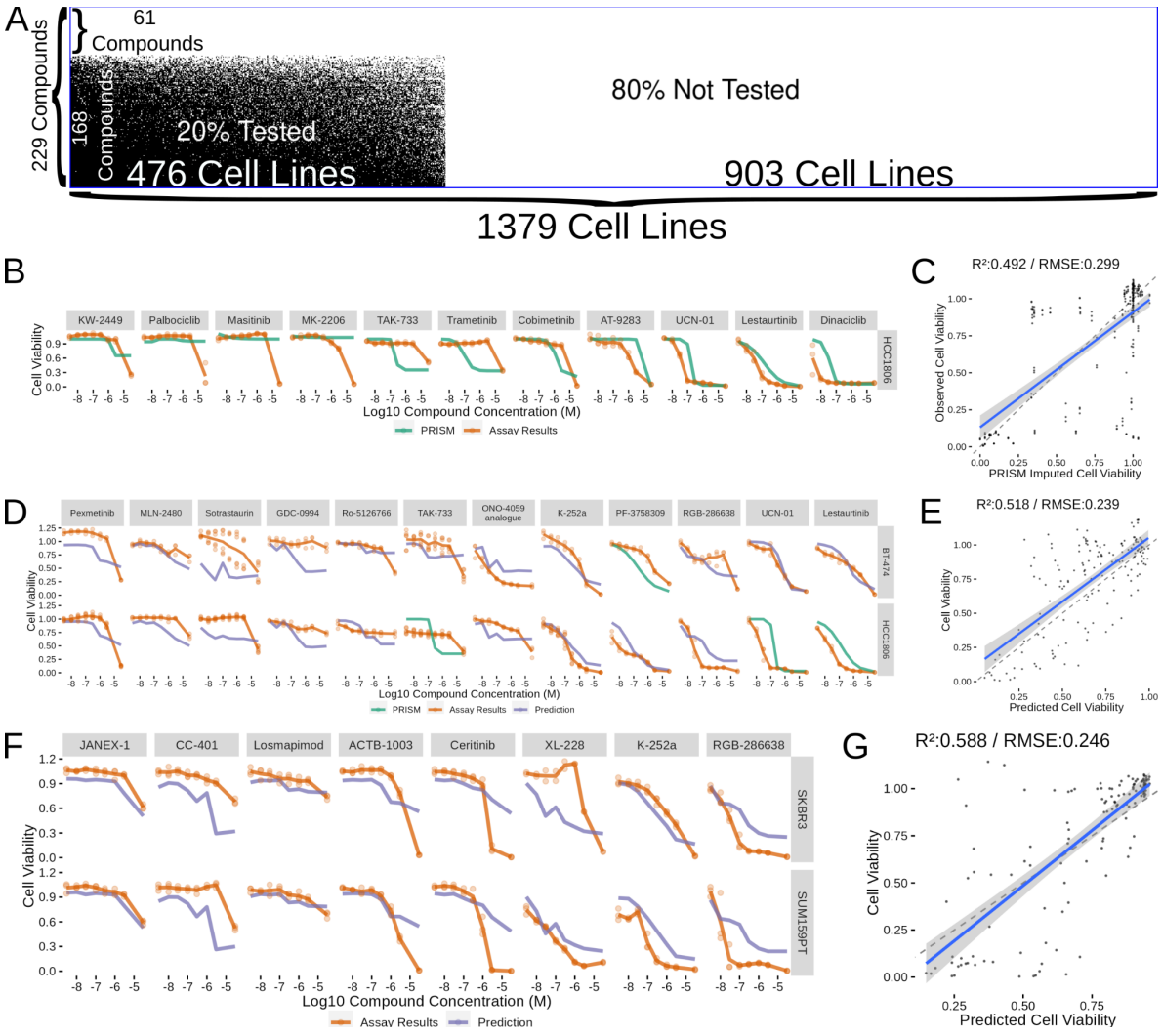

A
61 Compounds
229 Compounds
168 Compounds
20% Tested
476 Cell Lines
80% Not Tested
903 Cell Lines
1379 Cell Lines
B
KW-2449
Palbociclib
Masitinib
MK-2206
TAK-733
Trametinib
Cobimetinib
AT-9283
UCN-01
Lestaurtinib
Dinaciclib
HCC1806
Cell Viability
Log10 Compound Concentration (M)
PRISM
Assay Results
C
R²:0.492 / RMSE:0.299
Observed Cell Viability
PRISM Imputed Cell Viability
D
Pexmetinib
MLN-2480
Sotrastaurin
GDC-0994
Ro-5126766
TAK-733
ONO-4059 analogue
K-252a
PF-3758309
RGB-286638
UCN-01
Lestaurtinib
BT-474
HCC1806
Cell Viability
Log10 Compound Concentration (M)
PRISM
Assay Results
Prediction
E
R²:0.518 / RMSE:0.239
Cell Viability
Predicted Cell Viability
F
JANEX-1
CC-401
Losmapimod
ACTB-1003
Ceritinib
XL-228
K-252a
RGB-286638
SKBR3
SUM159PT
Cell Viability
Log10 Compound Concentration (M)
Assay Results
Prediction
G
R²:0.588 / RMSE:0.246
Cell Viability
Predicted Cell Viability

Figure 6: Validating a Subset of Compound Predictions in Breast Cancer Cell Lines. (A) Visualization of the space of compound (Y-axis) and cell line (X-axis) combinations that have been tested (white) and non tested (black) with a blue box surrounding the entire visualization. (B) Cell viability results from testing a set of compounds (labeled above each curve) and a cell line (HCC1806) already tested in the PRISM collection. (C) Scatterplot summarizing all the results from part B into a single plot with a linear best fit line showing in blue. (D) Cell viability results and corresponding predictions or PRISM results from a set of cell lines included in PRISM (BT-474 and HCC1806) and a selection of compounds which were mostly not included in PRISM. (E) Scatterplot summarizing all the prediction results from part D into a single plot. (F) Cell viability results and corresponding predictions for a set of cell lines and compounds not included in PRISM. (G) Scatterplot summarizing all the results from part F into a single plot.

Our first goal when beginning to validate a subset of model predictions was to see how well we could replicate the results from the PRISM assay. We selected the well characterized triple negative breast cancer (TNBC) cell line HCC1806 and a set of compounds that displayed a range of viability effects from the 134 Klaeger kinase inhibitors that had been used in the PRISM assay with the HCC1806 cell line (Figure 6B). Several of these compounds performed very similarly in our assay as compared with the imputed viability PRISM values, notably cobimetinib, UCN-01, AT-9283, lestauritinib and dinaciclib. However, several of the compounds that showed high viability effects at high concentrations were not reflected in the imputed viability results, which lowered the replication R^2 to 0.492 and the RMSE to 0.299 (Figure 6C). To put these replication efforts in context, we looked for experimental cell viability results from the NCI-60 [37]and PRISM results where the same compound and cell line were assayed. In order to gain a broader understanding of cell line viability replicability, we included

every compound match we could find between the two data sets. We found 172 compounds and 32 cell line matches between these data sets and found an overall R^2 of 0.444 and an RMSE of 0.296 (Figure S4). These results were in agreement with our much smaller PRISM replication effort, indicating that the variance between model predictions and experiments is no worse than the variance observed between experiments replicated by different groups. With the inherent limitations identified by the replication effort acknowledged, we next moved into testing new cell line and compound combinations.

We started testing new cell line and compound combinations by continuing with the HCC1806 line and adding in the HER2 positive breast cancer cell line BT-474. We selected a set of compounds predicted to have a range of effects on the two cell lines and then conducted a viability screen with each of these compounds (Figure 6D). Much like the replication attempt, we observed several compounds where the predicted viabilities were close to the measured viability (K-252a, UCN-01, PF-3758309 and lesauritinib). Overall, the R^2 (0.518) and RMSE (0.239) values were comparable with replication effort, indicating that the model was performing well on new compounds (Figure 6E). As our most challenging final test, we decided to test two cell lines that are not present in the PRISM data set (HER2+ line SKBR3 and TNBC line SUM159PT) against a set of compounds that weren't included in the PRISM compound set. Once again, with this "double-untested" experiment, we selected a set of compounds predicted to have varying effects across concentrations and observed a combination of compounds with strong and weak correlation between predictions and results (Figure 6F). Notable among the better results were JANEX-1, losmapimod and K-252a, while the model struggled with CC-401 and parts of the RGB-286638, ACTB-1003 and ceritinib curves. The overall performance of the model (R^2 of 0.588 and RMSE of 0.246) were comparable to the other model validation results

(Figure 6G). These independent validation efforts demonstrate that the model predictions are able to generalize into previously untested cell lines and compounds.

DISCUSSION

Given the potential of targeted kinase inhibitor therapies, the ability to predict how a given treatment may alter kinome state and lead to a given phenotype is fundamentally enabling. In this work, we developed a set of computational models that predict cell viability after treatment with a set of small molecule kinase inhibitors. To accomplish this, we used several publicly available data sets that provided information concerning the untreated gene expression of the cell lines used in the viability screen and another that gave detailed information about the proteins targeted by small molecule kinase inhibitors. We examined how single gene expression and kinome state values were related to cell viability and how models with various numbers of gene expression and kinome state values varied in quality. In addition to gene expression, we also tested a set of models which included a broader range of baseline measurements (CNV, proteomics and gene essentiality) and concluded that these additional data sets were not able to significantly improve model performance. Finally, we tested some of the model predictions in several triple negative and HER2 positive breast cancer lines and found acceptable agreement between the model predictions and experimental results.

This work demonstrates how knowledge of the inhibition state of the kinome, derived from a proteomic assay based on a four-cell lysate mixture, can predict a cellular process as fundamental as viability. Importantly, the models achieved these surprising results by using a "generic" or "general" kinase inhibition profile measured with proteomic kinobead profiling of a four cell line lysate exposed to an extensive library of kinase inhibitors at multiple doses[26]. Thus, the models learned by linking non-cell line specific kinome inhibition state information with that of specific drug-cell line relationships.

We acknowledge several limitations of this work. First, all of the results in this paper rest on the availability of kinome profiling data specific to a given kinase inhibitor, so the methods here are not applicable to prediction of cell viability effects in any other class of compound. We believe that a similar strategy could be used to build models in compound classes where the spectrum of targets were as comprehensively identified. The universe of small molecule kinase inhibitors is substantially larger than those that were surveyed by Klaeger et al., but since our modeling methodology depends on the comprehensive nature of their work, we're limited in the number of compounds where we can make predictions. One of our next goals is to attempt to broaden the scope of compounds through integration of other high-content kinome profiling techniques such as KINOMEscan and Nano-BRET. In addition, while the models described in this paper do make somewhat accurate predictions, these results point to a degree of missing predictability in cell viability for which new methods and data will need to be developed and collected. Also, since this work has targeted building a single comprehensive model, it is likely that subtle cancer type specific relationships are not captured such as the relationship between *RXRG* expression and melanoma [36]. This can be addressed by subsetting the model to make predictions about specific cancer types/subtypes. There is also an extensive set of alternative hyperparameter settings and potential modeling methodologies that we did not explore in this work. We also hope that by providing a full set of viability predictions for the broad range of cancer cell lines covered by the CCLE that this work can act as a resource for other researchers to find unexpected or interesting kinase inhibitors that affect their most used cell line model systems.

This work also suggests several extensions that would broaden or improve the model. Given recent interest in finding new compound combinations computationally, we are beginning to examine how best to combine the information from multiple compound kinome inhibition

states to predict the resulting cell viability effects. This would allow us to run computational drug combination screens. In addition, the methods outlined here will also likely work for any phenotype that can be measured after treatment with small molecule inhibitors and with sufficient throughput to gather a large enough data set. Finally, while we have made all of the code and data necessary to reuse our models available to the public on github, we also acknowledge that this is not the most user-friendly method for allowing non-computationally minded users to access the model. Thus, we also plan on developing a web-based system for allowing non-computationally minded users to submit a gene expression profile and receive a set of predictions concerning how their cellular system is expected to respond to the Klaeger set of kinase inhibitors.

Overall, we hope that this chapter makes a contribution to our understanding of how the overall state of kinome in response to small molecule inhibitors contributes to cell viability phenotypes. Our findings demonstrate that while individual kinase inhibition states and other single gene or protein readings are not very predictive of cell viability, machine learning approaches are able to combine sets of measurements related to the small molecule kinase inhibitors and gene expression data to make cell viability predictions. The results presented here show how a thorough understanding of kinase activity levels in conjunction with baseline omics data can be used to gain a better understanding of phenotypes such as cell viability.

METHODS

Our methods can be divided into two parts describing the computational aspects of this work and the experimental methods used to test the output of the computational components.

Data Sources

We used two primary data sources for this paper: the supplemental data section from Klaeger et al.[26] and the cell viability screening results from the PRISM lab. Specifically, we collected and organized the kinase inhibition states from supplemental table 2 of Klaeger et al, focusing on the Kinobeads subsheet. As for the PRISM data, we used the data from 2019 Q4 (labeled 19Q4 in the depmap portal), specifically the secondary screening data. In addition to these two data sets, we used supplemental data sets from the CCLE[34] and DepMap[38]. These data included results from baseline RNAseq (CCLE_expression.csv), copy number variation (CNV, CCLE_gene_cn.csv) and CRISPR-KO viability screening (CRISPR_gene_effect.csv). The 2021Q3 versions of these files were used. The proteomics data was downloaded from the Gygi lab website (https://gygi.hms.harvard.edu/publications/ccle.html), specifically Table S2

[39]. We also used version 11.5 of the STRING [40] protein network database (9606.protein.links.v11.5.txt.gz).

Data Preprocessing

The scripts implementing these descriptions are all available through github.

Klaeger et al. Kinase Inhibition Profiles: We read the values from the supplemental data table into R and produced a list of all proteins observed in any of the kinase inhibitor treatments. Since this table only contains the proteins affected by each compound, we filled in the relative intensity values for genes not associated with a given inhibitor with the default value of 1. There was a small (1.8%) number of single concentration values missing from the listed affected proteins, so we filled these values as the average of two nearest concentrations. Finally, a smaller set (0.01%) of likely outlier relative intensity readings were truncated to the 99.99 percentile (3.43).

PRISM Cell Viability: Since relatively few of the concentrations used in the PRISM assay match those used by Klaeger et al., we opted to use the response curve parameters provided through the depmap portal to interpolate the cell viability values. We interpolated these values at 30 μM, 3 μM, 1 μM, 300 nM, 100 nM, 30 nM, 10 nM and 3 nM to match those used by Klaeger et al. We applied a filter to remove any response curve parameter set that indicated that a given cell line and compound combination produced enhanced cell growth with increasing compound concentration. To perform the viability extrapolation, we used the four-parameter log-logistic formula described in the drc R package[41].

Gene Expression, CNV, CRISPR-KO and Proteomics: The files provided by the depmap portal for gene expression, CNV and CRISPR-KO values required very little modification to work in our machine learning pipelines. The primary modification was to add identifiers to each gene label, to ensure that omics data related to the same gene weren't accidentally combined.

The CRISPR-KO data also required an additional filter to remove 10 cell lines with missing data. The proteomics data processing was slightly more complicated, as there were substantially more protein readings missing from many more lines. In the cases of missing protein readings, we imputed these values to the minimum value for the overall distribution of that protein minus one standard deviation.

String: The STRING database[40] also required only mild preprocessing to extract the proteins that interacted with the components of our models. We filtered the interaction list to the high confidence (above 0.7) set and used bioMart[42] to convert the Ensembl protein identifiers to HGNC identifiers for matching with the other data sets.

Modeling Techniques and Types

To assess our models, we used a 10-fold cross validation strategy which randomized training and test set inclusion across the cell line and compound combinations. Thus, for any given viability curve resulting from treatment of a cell line with a compound, all of the results from the assay were considered as one unit for cross validation purposes. All steps of feature selection were also conducted under this cross-validation framework as well. For every fold of our data, we recalculated the correlation coefficient between cell viability and the features available to the model (kinase inhibition state, gene expression, etc) using only the data in the training set. The number of features was varied as specified in the results section. We used the entire data set to build the final model used to make the predictions in Supplemental Data File 1 and the results displayed in Figure 6.

We used random forest, XGBoost, TabNet and linear regression for all of our modeling efforts. All of our models are implemented using the tidymodels framework in R. We used the ranger random forest engine[43], the default XGBoost engine[44] and the default ordinary least squares linear regression engine. For all of our initial testing of these models we used the default

single set of hyperparameter settings to narrow our search for an acceptable model. This search indicated that the random forest model performed the best, so we attempted to further tune three additional parameters, the number of trees, the number of selected predictors and the minimal node size across the following ranges:

- Number of Trees: 500 (default), 1000, 1500 and 2000
- Number of Predictors: 11, 22 (default for 500 tree model), 33 and 44
- Minimal Node Size: 3, 5 (default) and 10

Compound Testing

BT-474, HCC1806, SUM-159 and SKBR-3 cells were grown in ATCC recommended media and seeded at 4000, 2000, 4000 and 500 cells per well respectively, in white flat-bottom 96-well plates (Corning). 24 hours after seeding, cells were treated with the respective drugs prepared in DMSO. All drugs were dosed at the same eight concentrations used in the Klaeger study: 30 μM, 3 μM, 1 μM, 300 nM, 100 nM, 30 nM, 10 nM and 3 nM. Seventy-two hours post-treatment, cells were lysed with CellTiter-Glo (Promega) per the manufacturer's protocol. Luminescence was read using the PHERAstar FS microplate reader (BMG Labtech) and gain adjustments were conducted for each cell line. Data were normalized row-wise to the DMSO-only (0.5% on cells) control samples on each plate to calculate relative viability. Quality checks were performed to look at the data distribution and the presence of spatial bias on a plate. A quality control metric of <120% of DMSO was applied to all rows analyzed. Across all >150 rows analyzed, only one row of XL-228 treated SKBR-3 cells failed to meet this criteria and was removed from analysis.

Software and Data Availability

All of the code written to support this paper is available through github (https://github.com/gomezlab/kinotype_viability) along with a walkthrough explaining where to find the code relevant to each part of the paper. We have also made all of the model validation results available through zenodo (https://doi.org/10.5281/zenodo.6323686).

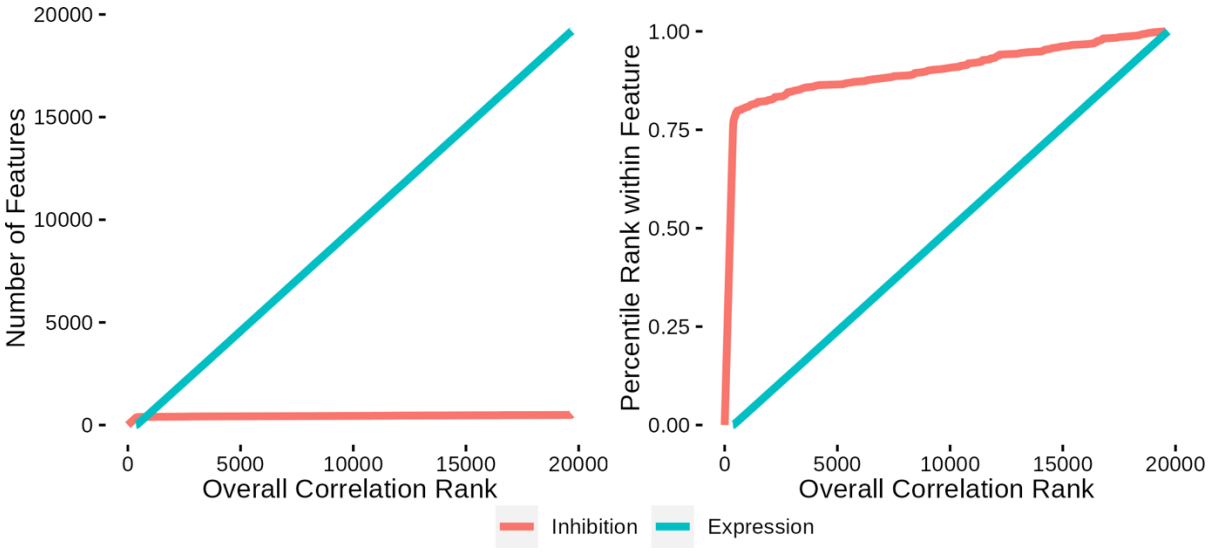
ACKNOWLEDGEMENTS

We would like to thank Donglin Zeng for helpful discussions concerning feature selection techniques and UNC Research Computing for access to the computational resources necessary for this work. We would also like to thank Madison Jenner and Jen Jen Yeh for providing a sample of JANEX-1.

Funding

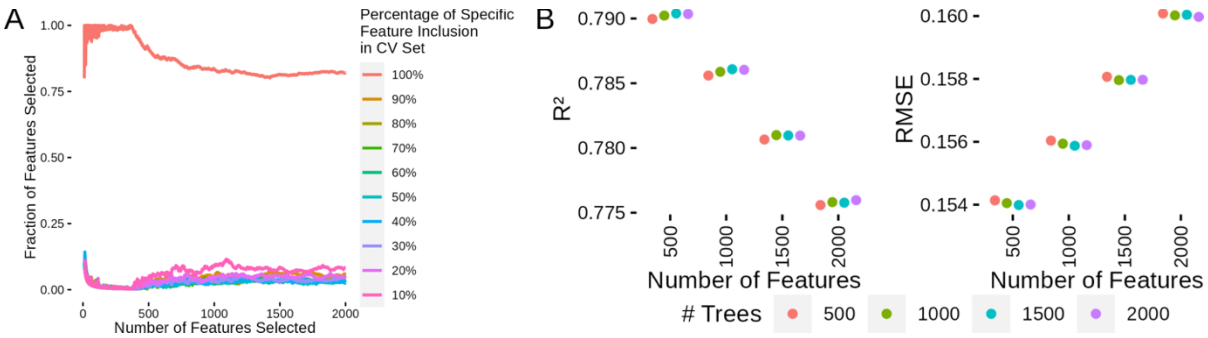
This work was supported by grants from the National Institutes of Health (grant numbers CA233811, CA238475 and DK116204)

SUPPLEMENTARY MATERIAL

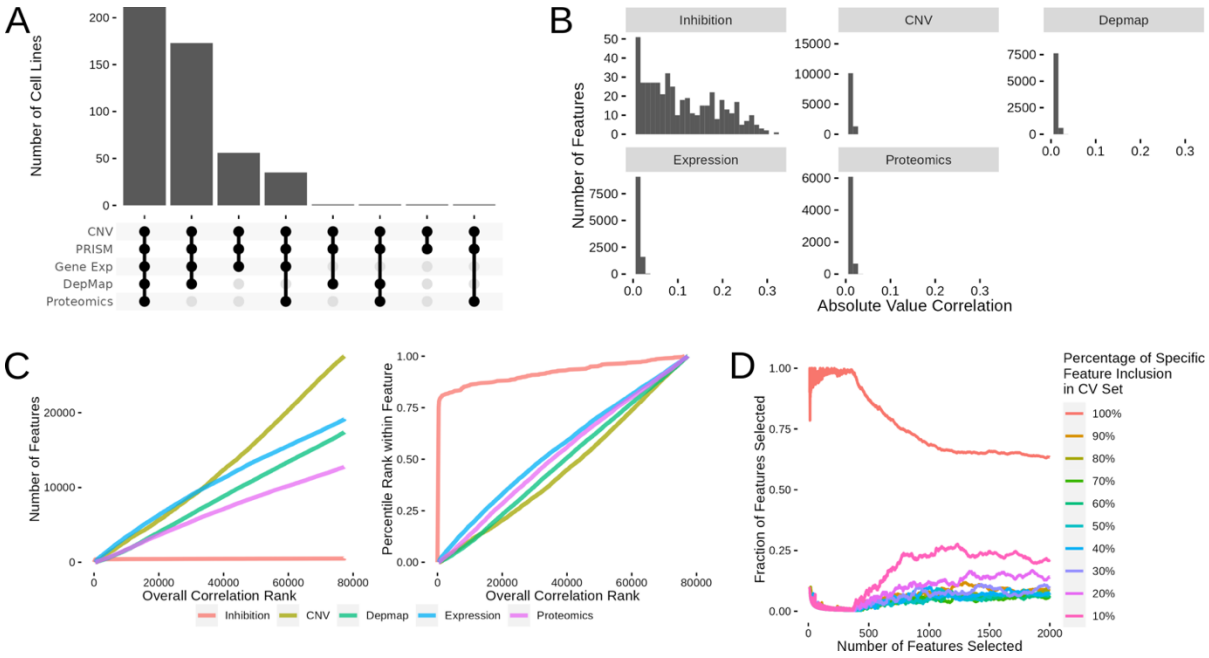


Supplemental Figure 2.1: Expanded Correlation Rankings. (Associated with Figure 2).

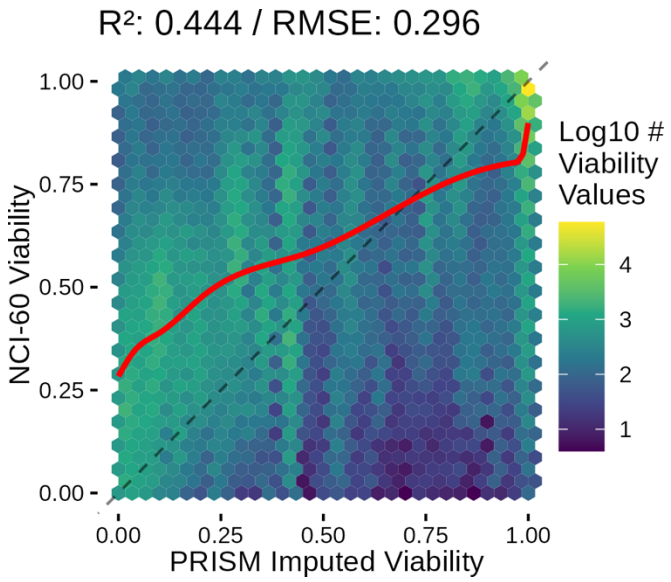
Extended version of Figure 2D covering all correlation ranks.



Supplemental Figure 2.2: Feature Selection with Cross Validation and Assessment of Increasing Random Forest Trees. (Associated with Figure 3). (A) The effect of cross validation data division on which features are selected for model inclusion. (B) The effect on R^2 and RMSE of increasing the number of trees used in the random forest algorithm. (C) The effect on R^2 and RMSE of modifying the selected predictor count and the minimal node size used in the random forest algorithm. (D) The distribution of R^2 and RMSE for single compound or cell line cross validation results.



Supplemental Figure 2.3: Expanded Correlation Rankings and Effect of Cross Validation Subsetting on Feature Selection. (Associated with Figure 5) (A) Upset plot showing the overlap between data sets across cell lines in the PRISM assay. (B) Small multiples plot showing the correlation of individual features to imputed cell viability for each of the feature types considered in this model. (C) Full feature correlation rankings for all data set types considered for Figure 5. (D) Effect of random 10-fold cross validation subsetting on which features are included in what percentage of the cross-validation data sets.



Supplementary Figure 2.4: Comparison of Cell Line and Compound Matched PRISM and NCI-60 Viability Results. The red line shows a loess fit through the data set.

REFERENCES

1. Falzone L, Salomone S, Libra M. Evolution of Cancer Pharmacological Treatments at the Turn of the Third Millennium. Front Pharmacol. 2018;9: 1300.

2. Keefe DMK, Bateman EH. Potential Successes and Challenges of Targeted Cancer Therapies. J Natl Cancer Inst Monogr. 2019;2019. doi:10.1093/jncimonographs/lgz008

3. O'Brien SG, Guilhot F, Larson RA, Gathmann I, Baccarani M, Cervantes F, et al. Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. N Engl J Med. 2003;348: 994–1004.

4. Deininger M, O'Brien SG, Guilhot F, Goldman JM, Hochhaus A, Hughes TP, et al. International Randomized Study of Interferon Vs STI571 (IRIS) 8-Year Follow up: Sustained Survival and Low Risk for Progression or Events in Patients with Newly Diagnosed Chronic Myeloid Leukemia in Chronic Phase (CML-CP) Treated with Imatinib. Blood. 2009;114: 1126.

5. Solomon BJ, Mok T, Kim D-W, Wu Y-L, Nakagawa K, Mekhail T, et al. First-line crizotinib versus chemotherapy in ALK-positive lung cancer. N Engl J Med. 2014;371: 2167–2177.

6. Geyer CE, Forster J, Lindquist D, Chan S, Romieu CG, Pienkowski T, et al. Lapatinib plus capecitabine for HER2-positive advanced breast cancer. N Engl J Med. 2006;355: 2733–2743.

7. Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med. 2001;344: 783–792.

8. Yuan M, Huang L-L, Chen J-H, Wu J, Xu Q. The emerging treatment landscape of targeted therapy in non-small-cell lung cancer. Signal Transduct Target Ther. 2019;4: 61.

9. Zhong L, Li Y, Xiong L, Wang W, Wu M, Yuan T, et al. Small molecules in targeted cancer therapy: advances, challenges, and future perspectives. Signal Transduct Target Ther. 2021;6: 201.

10. Seebacher NA, Stacy AE, Porter GM, Merlot AM. Clinical development of targeted and immune based anti-cancer therapies. J Exp Clin Cancer Res. 2019;38: 156.

11. Attwood MM, Fabbro D, Sokolov AV, Knapp S, Schiöth HB. Trends in kinase drug discovery: targets, indications and inhibitor design. Nat Rev Drug Discov. 2021;20: 839–861.

12. Bhullar KS, Lagarón NO, McGowan EM, Parmar I, Jha A, Hubbard BP, et al. Kinase-targeted cancer therapies: progress, challenges and future directions. Mol Cancer. 2018;17: 48.

13. Pottier C, Fresnais M, Gilon M, Jérusalem G, Longuespée R, Sounni NE. Tyrosine Kinase Inhibitors in Cancer: Breakthrough and Challenges of Targeted Therapy. Cancers . 2020;12. doi:10.3390/cancers12030731

14. Cohen P, Cross D, Jänne PA. Kinase drug discovery 20 years after imatinib: progress and future directions. Nat Rev Drug Discov. 2021;20: 551–569.

15. Shapiro P. Next Generation Kinase Inhibitors: Moving Beyond the ATP Binding/Catalytic Sites. Shapiro P, editor. Springer, Cham; 2020.

16. Lahiry P, Torkamani A, Schork NJ, Hegele RA. Kinase mutations in human disease: interpreting genotype–phenotype relationships. Nat Rev Genet. 2010;11: 60–74.

17. Laufer S, Bajorath J. New Horizons in Drug Discovery - Understanding and Advancing Different Types of Kinase Inhibitors: Seven Years in Kinase Inhibitor Research with Impressive Achievements and New Future Prospects. J Med Chem. 2022;65: 891–892.

18. Golkowski M, Lau H-T, Chan M, Kenerson H, Vidadala VN, Shoemaker A, et al. Pharmacoproteomics Identifies Kinase Pathways that Drive the Epithelial-Mesenchymal Transition and Drug Resistance in Hepatocellular Carcinoma. Cell Syst. 2020. doi:10.1016/j.cels.2020.07.006

19. Collins KAL, Stuhlmiller TJ, Zawistowski JS, East MP, Pham TT, Hall CR, et al. Proteomic analysis defines kinase taxonomies specific for subtypes of breast cancer. Oncotarget. 2018;9: 15480–15497.

20. Duncan JS, Whittle MC, Nakamura K, Abell AN, Midland AA, Zawistowski JS, et al. Dynamic reprogramming of the kinome in response to targeted MEK inhibition in triple-negative breast cancer. Cell. 2012;149: 307–321.

21. Frejno M, Zenezini Chiozzi R, Wilhelm M, Koch H, Zheng R, Klaeger S, et al. Pharmacoproteomic characterisation of human colon and rectal cancer. Mol Syst Biol. 2017;13: 951.

22. Yesilkanal AE, Yang D, Valdespino A, Tiwari P, Sabino AU, Nguyen LC, et al. Limited inhibition of multiple nodes in a driver network blocks metastasis. Elife. 2021;10. doi:10.7554/eLife.59696

23. Zawistowski JS, Bevill SM, Goulet DR, Stuhlmiller TJ, Beltran AS, Olivares-Quintero JF, et al. Enhancer Remodeling during Adaptive Bypass to MEK Inhibition Is Attenuated by Pharmacologic Targeting of the P-TEFb Complex. Cancer Discov. 2017;7: 302–321.

24. Bantscheff M, Eberhard D, Abraham Y, Bastuck S, Boesche M, Hobson S, et al. Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. Nat Biotechnol. 2007;25: 1035–1044.

25. Plowright AT, editor. Target Discovery and Validation: Methods and Strategies for Drug Discovery. Wiley; 2019. pp. 97–130.

26. Klaeger S, Heinzlmeir S, Wilhelm M, Polzer H, Vick B, Koenig P-A, et al. The target landscape of clinical kinase drugs. Science. 2017;358. doi:10.1126/science.aan4368

27. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER 3rd, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569: 503–508.

28. Lu J, Chen M, Qin Y. Drug-induced cell viability prediction from LINCS-L1000 through WRFEN-XGBoost algorithm. BMC Bioinformatics. 2021;22: 13.

29. Lu J, Chen M, Qin Y, Yu X. [Prediction of drug-induced cell viability by SAE-XGBoost algorithm based on LINCS-L1000 perturbation signal]. Sheng Wu Gong Cheng Xue Bao. 2021;37: 1346–1359.

30. Daemen A, Griffith OL, Heiser LM, Wang NJ. Modeling precision treatment of breast cancer. Genome Biol. 2013. Available: https://link.springer.com/article/10.1186/gb-2013-14-10-r110

31. Corsello SM, Nagari RT, Spangler RD, Rossen J, Kocak M, Bryan JG, et al. Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. Nat Cancer. 2020;1: 235–248.

32. Vidović D, Koleti A, Schürer SC. Large-scale integration of small molecule-induced genome-wide transcriptional responses, Kinome-wide binding affinities and cell-growth inhibition profiles reveal global trends characterizing systems-level drug action. Front Genet. 2014;5: 342.

33. Yu C, Mannan AM, Yvone GM, Ross KN, Zhang Y-L, Marton MA, et al. High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. Nat Biotechnol. 2016;34: 419–423.

34. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483: 603–607.

35. Berginski ME, Moret N, Liu C, Goldfarb D, Sorger PK, Gomez SM. The Dark Kinase Knowledgebase: an online compendium of knowledge and experimental results of understudied kinases. Nucleic Acids Res. 2020. doi:10.1093/nar/gkaa853

36. Dempster JM, Krill-Burger JM, McFarland JM, Warren A, Boehm JS, Vazquez F, et al. Gene expression has more power for predicting in vitro cancer cell vulnerabilities than genomics. bioRxiv. 2020. p. 2020.02.21.959627. doi:10.1101/2020.02.21.959627

37. Holbeck SL, Camalier R, Crowell JA, Govindharajulu JP, Hollingshead M, Anderson LW, et al. The National Cancer Institute ALMANAC: A Comprehensive Screening Resource for the Detection of Anticancer Drug Pairs with Enhanced Therapeutic Activity. Cancer Res. 2017;77: 3564–3576.

38. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, et al. Defining a Cancer Dependency Map. Cell. 2017;170: 564-576.e16.

39. Nusinow DP, Szpyt J, Ghandi M, Rose CM, McDonald ER 3rd, Kalocsay M, et al. Quantitative Proteomics of the Cancer Cell Line Encyclopedia. Cell. 2020;180: 387-402.e16.

40. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021;49: D605–D612.

41. Ritz C, Baty F, Streibig JC, Gerhard D. Dose-Response Analysis Using R. PLoS One. 2015;10: e0146021.

42. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4: 1184–1191.

43. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw. 2017;77: 1–17.

44. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery; 2016. pp. 785–794.

CHAPTER 3: INTEGRATED SINGLE-DOSE KINOME PROFILING DATA IS PREDICTIVE OF CANCER CELL LINE SENSITIVITY TO KINASE INHIBITORS

INTRODUCTION

Computational drug screening has recently emerged as a powerful approach to integrate vast amounts of cancer cell line multi-omics data into predictive models, with the goal of predicting downstream phenotypic responses such as growth and viability. Advancing rapidly with recent developments in machine learning, these methods have the potential to predict outcomes for large drug libraries with minimal experimental cost, reducing the number of drug candidates fed into downstream validation efforts. Most current approaches to the prediction of drug response use baseline cell-line multi-omics data (e.g., mutation status, gene expression, copy number variation, etc.) and Quantitative Structure-Activity Relationships (QSAR) to map drug structure characteristics onto their biological phenotypes, but information describing drug-target interactions, especially at the protein level, remain underutilized because of the unique nature of associated data acquisition methods. For example, the recent DREAM challenge[1] hosted by the National Cancer Institute (NCI) for drug-response predictions saw the winning team utilize high throughput drug screening data along with baseline gene expression features[2] and achieved at most 80% accuracy in predicting cell line responses in a binary fashion.

As one of the foundations of cellular information transfer, protein kinases are enzymes that have also shown promise as therapeutic targets, with initial success being found through the development of Imatinib (Gleevec). Drugs that inhibit kinases ("kinase inhibitors") are now one of the fastest growing clinical drug classes (74 FDA approved as of 2022), but around 1/3rd of

all known kinases still have relatively unknown functions and few chemical tools exist to interrogate and expand this knowledge. To explore the potential of the kinome as a therapeutic target, recent work has focused on profiling the full breadth of targets for kinase inhibitors, especially since many inhibitors have significant off target effects as a result of targeting the conserved ATP-binding pocket. Continued improvements in high-throughput assays such as Kinobead/MS[3], KINOMEscan(© DiscoverX), and KiNativ[4] now enable measurement of a given inhibitor's interactions across 250-500 kinases, providing a snapshot of its effect on the physiological kinome. We refer to this kinome-wide profiling data as the "kinome inhibition state" of a given inhibitor. This ability to generate drug-target interaction data on a large scale for a compound class is relatively unique, providing a novel means to leverage knowledge of off-target effects for drug response prediction.

The DepMap portal database[5] contains thorough multi-omic characterization of ~1000 cancer cell lines of all types, and cell viability measurements for about ~1500 repurposed compounds, ~250 of which are kinase inhibitors. Using this data, we can connect kinase inhibitor phenotypes of cell viability to their "kinome inhibition states" and build models to predict the cellular responses to treatment with different kinase inhibitors. We have previously shown that these kinome states obtained through the Kinobeads assay for clinical inhibitors are predictive of cancer cell viability, and also validated these predictions experimentally[6]. However, the kinobeads assay is unique and requires dedicated lab personnel to run, restricting its use to relatively few labs. In contrast, the KINOMEscan assay is a popular and easily accessible commercial alternative that assays a panel of ~500 native and mutant kinases recombinantly. Large amounts of KINOMEscan data have been deposited online by various groups[7,8], including data for inhibitors developed against understudied kinases. These altogether account for four times as many inhibitors profiled (~800) when compared to the data

available from the kinobeads assay, representing a massive expansion of publicly available inhibitor state data. However, due to the uncharacterized nature of the inhibitors in the large KINOMEscan data set, only a small number of them (~40) have been tested in the DepMap screening database, compared to ~200 inhibitors from the kinobeads set.

In this work, we describe a framework to create an integrated kinome inhibition state data set by combining kinobeads and KINOMEscan data, and then leverage the breadth of this data into predictive models. This combined set contains single-dose inhibitor profiling data for a total of ~800 kinases and kinase interacting proteins, spanning almost 1000 kinase inhibitors that target a diverse section of the overall kinome space. When leveraged within a machine learning framework, and supplemented with baseline gene expression data, we are able to predict the sensitivity of ~450 cancer cell lines in the DepMap screening dataset, with a reasonable R^2 of ~0.7. Using this model, we were able to generate sensitivity predictions for 1.2 million inhibitor-cell line combinations, many of them targeted towards understudied kinases. We then experimentally validated these predictions in well characterized breast cancer cell lines seen by the model, as well as primary derived pancreatic cancer cell lines. We find reasonable agreement between predicted and observed outcomes in most compounds, seeing an expected drop in performance for understudied compounds and unique patient-derived cell lines. Together, these results show that there is a strong and predictive relationship between the state of the kinome (its "kinotype") and downstream cellular phenotypes, while further suggesting potential opportunities for leveraging computational models in inhibitor therapy design.

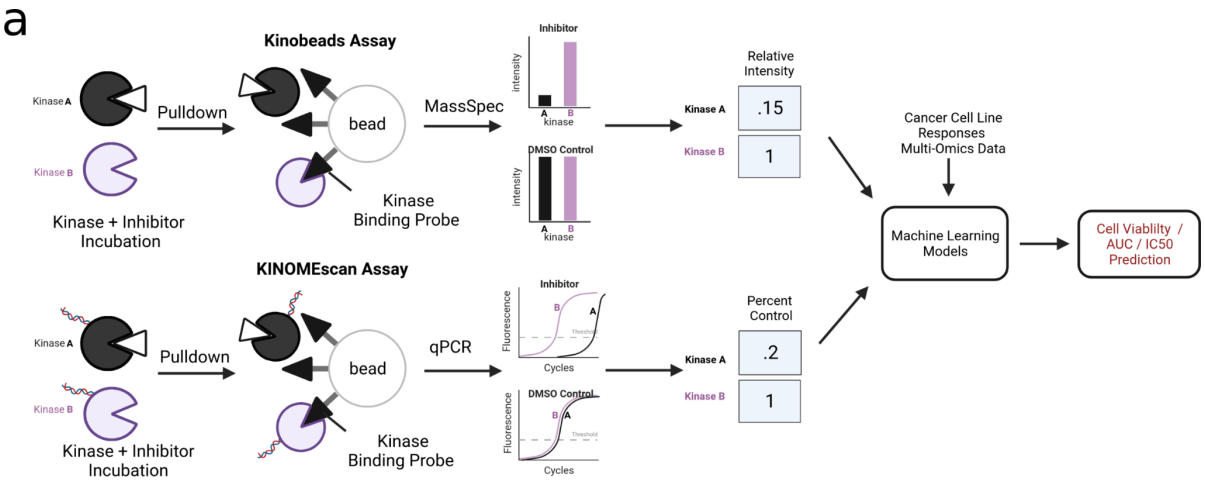
RESULTS

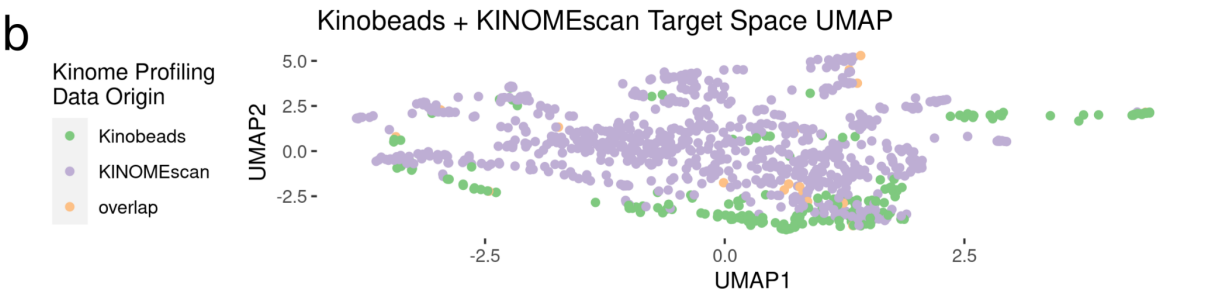
Creating an Integrated Set of Kinome Profiling Data Across a Wide Chemical Space

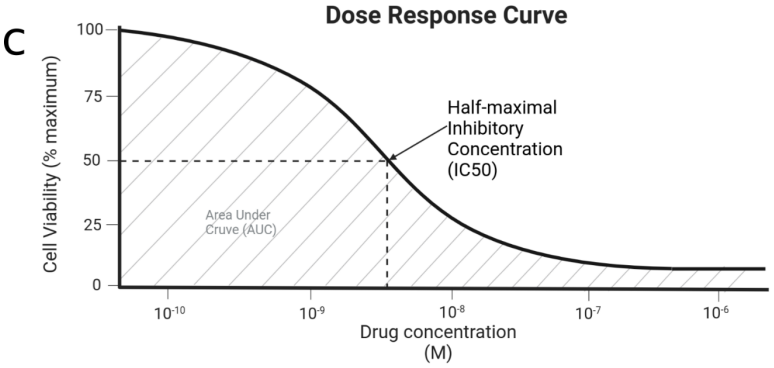
Kinase inhibitors have been profiled using a number of assays, but for this study we have focused on a specific subset of kinase inhibitors that have been assayed using the kinobead/MS-based method [9] or the KINOMEscan® (DiscoverX) method. These methods assess their specific kinase targets as well as the magnitude of inhibition of each kinase in response to different inhibitor concentrations[9]. We combined kinome profiling datasets from Klaeger et al (Kinobeads), LINCS[7] (KINOMEscan), and UNC[10] (KINOMEscan), filtering down to profiles measured only at 1uM. For the small amount of overlap between datasets, the mean inhibition value was taken across drug-kinase combinations. Given that both assays measure the engagement of inhibitors to kinases, most of the proteins that appear in assay results are either known kinases or closely associated proteins. Specifically, kinase inhibitor profiles include measurements on all wild-type and phosphorylated kinases (~500), along with a set of associated proteins (~300). As such, we will refer to this data as "kinase inhibition states", and the profile of each individual drug as its "kinome inhibition state" (fig 1a).

After integration, we were left with a final set of ~1000 compounds with corresponding information on their inhibitor-induced kinome states, describing changes in ~800 kinases and kinase interactors. We then performed a UMAP dimensionality reduction [11] on the dataset for visualization(fig 1b). The UMAP coordinates represent the aggregate effect of each inhibitor on the kinome, i.e. it is a representation of the uniqueness of its kinome inhibition state. Inhibitors that have similar effects on the kinome will have similar coordinates, while disparate inhibitors will have coordinates that are far apart. Using this, we can examine the diversity of kinome space targeting in our dataset, based on the origin of kinome profiling data. Our analysis shows that

integrating KINOMEscan data for 800 inhibitors vastly increases the kinome space targeted (fig 1b), compared to just kinobeads alone.







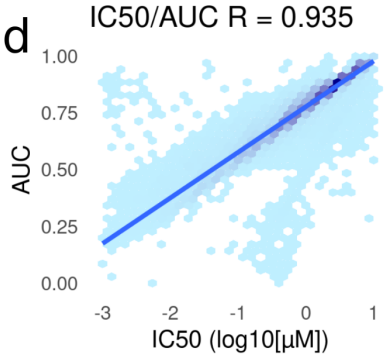


Figure 3.1. Modeling Pipeline and Target Variable Overview. (a) Schematic of Kinobead/MS (upper) and KINOMEscan assay data integration into machine learning models predicting IC50 and AUC. (b) Visualization of UMAP dimensionality reduction on the combined kinome profiling data set, each point represents a single compound's position in the target space, and colors representing the origin of kinome profiling data. Target Variables for Modeling: (c) Extraction of IC50 and AUC from a drug's dose response curve in a given cell line. (d) Correlation and Scales of IC50 vs AUC values. Blue line indicates a linear model fit through the data.

Connecting Inhibited Kinome States to Cancer Cell Line Sensitivities from the DepMap Repurposing Screen

To connect these kinase inhibitors and their induced inhibition states with their corresponding phenotypes in cancer cell lines, we make use of the DepMap repurposing screen, which uses the PRISM assay[12] to run highly multiplexed cell viability assays. This dataset contains cell viability measurements for over 1500 drugs profiled in 450 cell lines. From within this data, we found ~200 drugs for which we also have corresponding profiling data as described above.

The DepMap repurposing dataset provides cell viability measurements across multiple drug doses, but since our dataset of kinome states is restricted to single-dose measurements, we extracted two single summary statistics for describing cell line sensitivity to kinase inhibitors: Dose-response Area Under the Curve (AUC) and half-maximal Inhibitory Concentration (IC50). These properties are highly correlated with each other, having a Pearson's correlation coefficient ~ 0.9 (fig 1d). We extracted these properties from DepMap and matched them to our kinome states (fig 1c). The final integrated dataset has ~250 drugs tested across ~450 cell lines, representing ~70,000 inhibitor-cell line combinations representing nearly all cancer types.

Examining Bivariate Association of Features to Cell Line Sensitivities Provides a Means for Feature Selection

The 450 cell lines tested in the DepMap dataset also have accompanying baseline RNAseq gene expression data, so we integrated the ~20,000 TPM values for each cell line into the kinome-state and cell line sensitivity dataset. This adds baseline cell line-specific gene expression information to our cell line agnostic inhibitor-induced kinome states.

We examined bivariate associations of each of the ~21000 features against the outcome variables (dose response AUC and IC50) using Pearson's correlation coefficient, and ranked them from largest to smallest absolute value of association. We found that the most correlated feature is the drug-induced kinase inhibition state of TP53RK (fig 2a) with a correlation coefficient R ~0.3, while the most correlated baseline gene expression value was OGFRL1 with a correlation coefficient R ~ 0.05. Overall, inhibitor-induced kinome states showed stronger correlation with cell line sensitivity metrics (Fig 2c) despite there being 40x more baseline gene expression features than kinome states.

After exploring the relationship between each feature and cell line sensitivity, we sought to use machine learning models to combine these features to predict cell line sensitivities to kinase inhibitor treatment. Using the ranked list of feature associations, we utilized a feature selection scheme where we tested discrete increments of the ranked features included in each model to find the best performers (fig 2d).

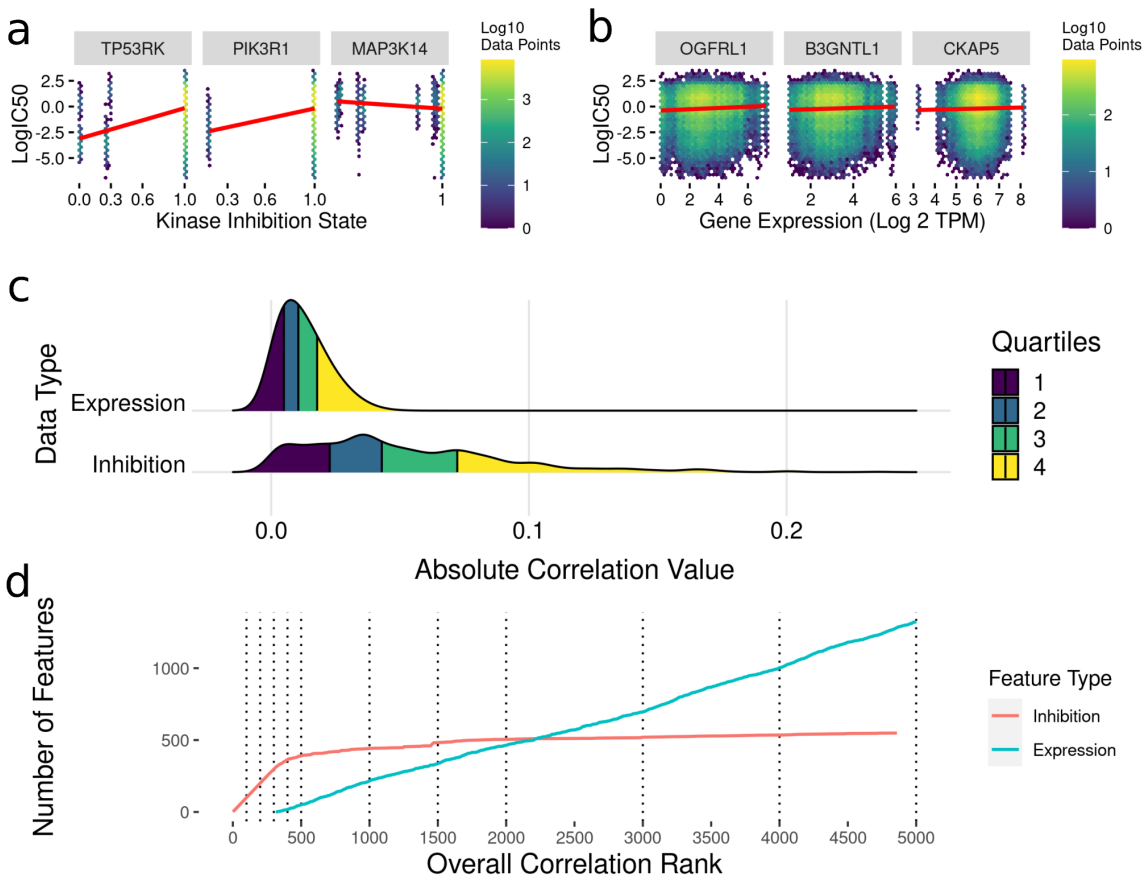


Figure 3.2. Feature Selection by Bivariate Association with Cancer Cell Line Sensitivity. (a) Sample kinase inhibition state versus LogIC50 heatmap plots showing inhibition states with high (TP53RK), medium (PIK3R1) and low (MAPK14) correlation values. (b) Sample gene expression versus LogIC50 heatmap plots showing genes with high (*OGFRL1*), medium (*B3GNTL1*) and low (*CKAP5*) correlation values. (c) Ridgeline Plot showing distributions of correlations with drug IC50s and AUC values across the data types included in analysis. (d) Plots showing what order classes of features are selected from the ranked set of inhibition states and baseline gene expression values. The dotted lines indicate the discrete increments of feature rank cutoffs at which model performance was tested. Kinase inhibition states were the most informative feature within the first ~300, after which gene expression features started to show predictive value.

Machine Learning Models Can Predict Cancer Cell Line Sensitivity from a Combination of Kinome Inhibition States and Baseline Transcriptomics

To build machine learning models to predict cancer cell line AUC and IC50 in response to treatment with kinase inhibitors, the highest ranked 100-5000 features were selected from the dataset linking drug-induced kinome states to cancer cell line responses (fig 2d).

We compared three model types: LASSO regression, random forest and XGBoost. All models were trained with 10-fold cross validation to minimize overfitting on the training data, ensuring comparable accuracy of the model predictions on new kinase inhibitors and cell lines. For each feature number from 100-5000 we tuned sets of 30 hyperparameters for all model types (fig 3a). The R-squared value between predicted and actual value was utilized as the metric for model comparison. Overall, the 5000 feature XGBoost model performed the best with a cross-validation R-squared of ~0.7 (fig 3b).

Since tree-based machine learning models like XGBoost offer in-built explainability, it is possible to interrogate and explain which features were most important in predicting the outcome of cell line sensitivities. These importances generated via shapley values[13] show kinase inhibition states to be overwhelmingly more important for predicting cell line responses when compared to baseline gene expression. Kinases involved in cell cycle and proliferation are overrepresented in the top 25 features (MAP2K, MEK2, CDKL5 etc.), but interestingly six kinase interactor proteins are included as well, suggesting that interactions between inhibitors and non-kinases (off-target effects) have important consequences for cell viability. Baseline gene expression features show much lower model importances, but 40% of the top 25 genes have known interactions with kinases whose inhibition states are used in the model.

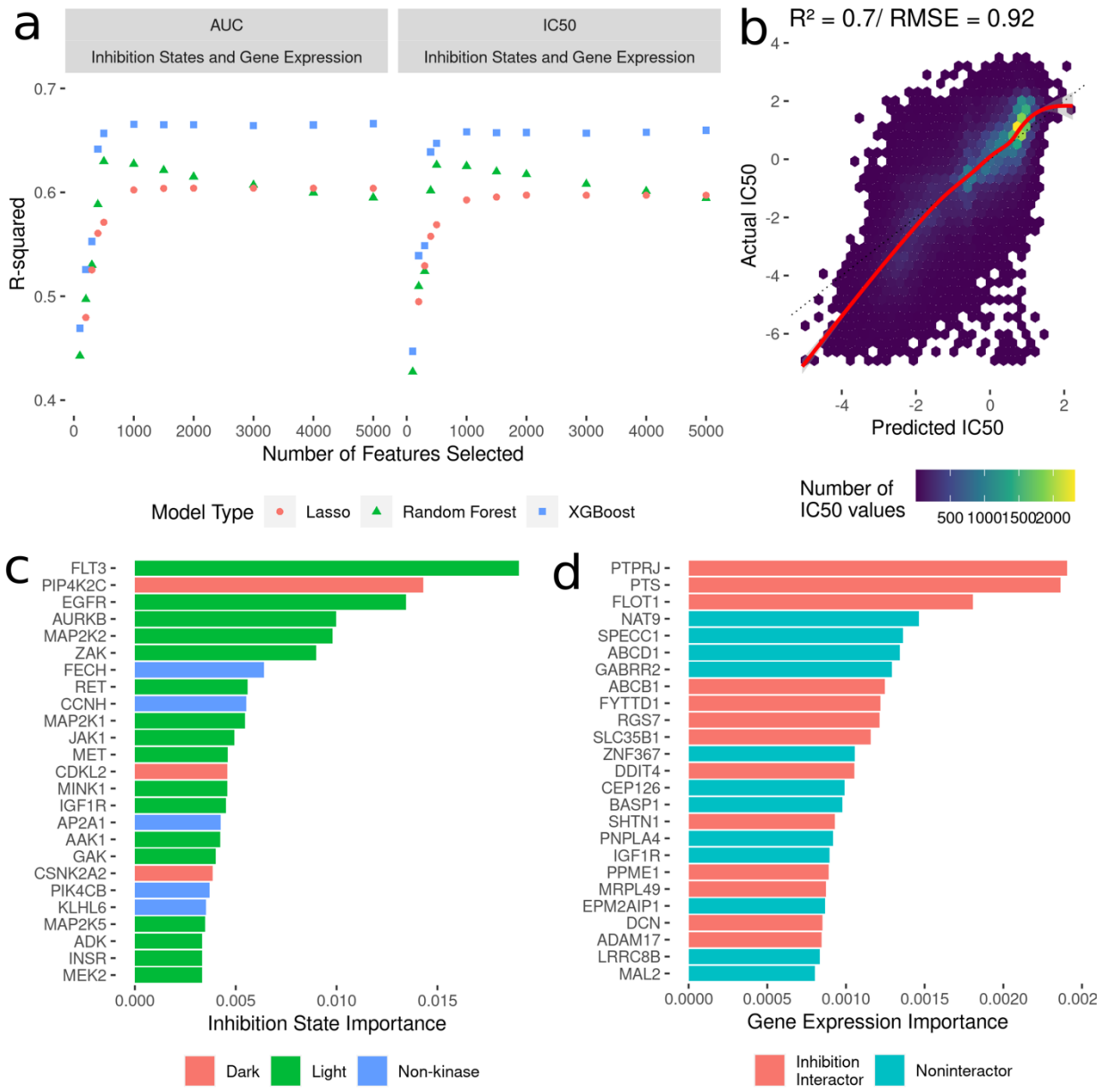

a
AUC
Inhibition States and Gene Expression
IC50
Inhibition States and Gene Expression
R-squared
Number of Features Selected
Model Type
Lasso
Random Forest
XGBoost
b
R² = 0.7/ RMSE = 0.92
Actual IC50
Predicted IC50
Number of IC50 values
c
FLT3
PIP4K2C
EGFR
AURKB
MAP2K2
ZAK
FECH
RET
CCNH
MAP2K1
JAK1
MET
CDKL2
MINK1
IGF1R
AP2A1
AAK1
GAK
CSNK2A2
PIK4CB
KLHL6
MAP2K5
ADK
INSR
MEK2
Inhibition State Importance
Dark
Light
Non-kinase
d
PTPRJ
PTS
FLOT1
NAT9
SPECC1
ABCD1
GABRR2
ABCB1
FYTTD1
RGS7
SLC35B1
ZNF367
DDIT4
CEP126
BASP1
SHTN1
PNPLA4
IGF1R
PPME1
MRPL49
EPM2AIP1
DCN
ADAM17
LRRC8B
MAL2
Gene Expression Importance
Inhibition Interactor
Noninteractor

Figure 3.3. Development of Models to Predict Cancer Cell Line Sensitivities to Kinase Inhibitors by Integrating Single-Dose Kinome Profiling Data (a) Model performance metrics (R-squared) for LASSO (orange dot), Random Forest (green triangle) and XGBoost (blue square). (b) Scatterplot of predicted IC50 values from the best-performing model vs actual IC50 values. The red line indicated a smooth fit through the data points. (c) Horizontal bar plot showing model importance of individual kinase inhibition states by shapley values. (d) Horizontal bar plot showing model importance of individual baseline gene expression by shapley values.

Inclusion of Various Multi-Omics Data with Kinome Inhibition States and Gene Expression Did Not Improve Model Predictive Performance

In addition to the baseline gene expression data, all the cell lines in the DepMap database have three other profiling data types available: copy number variation, gene essentiality from CRISPR/KO, and baseline proteomics. To see if inclusion of these data into models would improve predictions, we integrated these with the modeling dataset of kinome inhibition states and gene expression, and used identical modeling strategies described above to select correlated features, build, and evaluate LASSO, random forest, and XGBoost models (Supp. Fig 1). We found that adding in the various multi-omic data types did not significantly outperform the models limited to kinase inhibition states and baseline gene expression (R-squared of ~0.69 for predicting IC50).

Experimental Validation of Model Predictions were Successful in Characterized and Novel Cell Lines

After fitting the model on 70,000 cell line-drug combinations, predictions were made on 1.2 million unseen (not seen by the model) drug-cell line combinations. Approximately 90% of the untested inhibitors were associated with KINOMEscan datasets. As an initial validation, we tested a subset of the predictions in well-characterized breast cancer cell lines (HER2 positive: SK-BR-3, BT-474 and two triple negative: SUM159, HCC1806). We analyzed the performance of the model on experimental data for unseen drug-cell line combinations, arriving at an R value ~ 0.6 for all but one (SKBR3) breast cancer cell line(fig 4b). Notably, all the drugs tested had kinome profiling data from the Kinobeads assay.

We then further validated the model by predicting inhibitor effects from collected RNAseq data in tumor (two) and stroma (one) derived cell lines from PDAC patients [14,15]. Importantly, these patient-derived cell lines were profiled for baseline gene expression in-house, and represent a novel and highly heterogeneous transcriptional landscape which the model has not seen before. Dose-response AUC predictions were made by the model for 58 drugs with kinome profiling data from the Kinobeads assay and 18 drugs with kinome profiling data from the KINOMEscan assay. The model predicted AUC was compared to experimentally generated AUC, revealing an average R ~ 0.5 for drugs with kinome profiling data from the Kinobeads assay tested in patient stroma-derived cell lines (fig 4c), and R ~ 0.4 for drugs with kinome profiling data from the KINOMEscan assay. On the other hand, in patient tumor-derived cell lines, drugs from kinobeads had model accuracy R ~0.49, and drugs from KINOMEscan had R ~ 0.3 (fig. 4c).

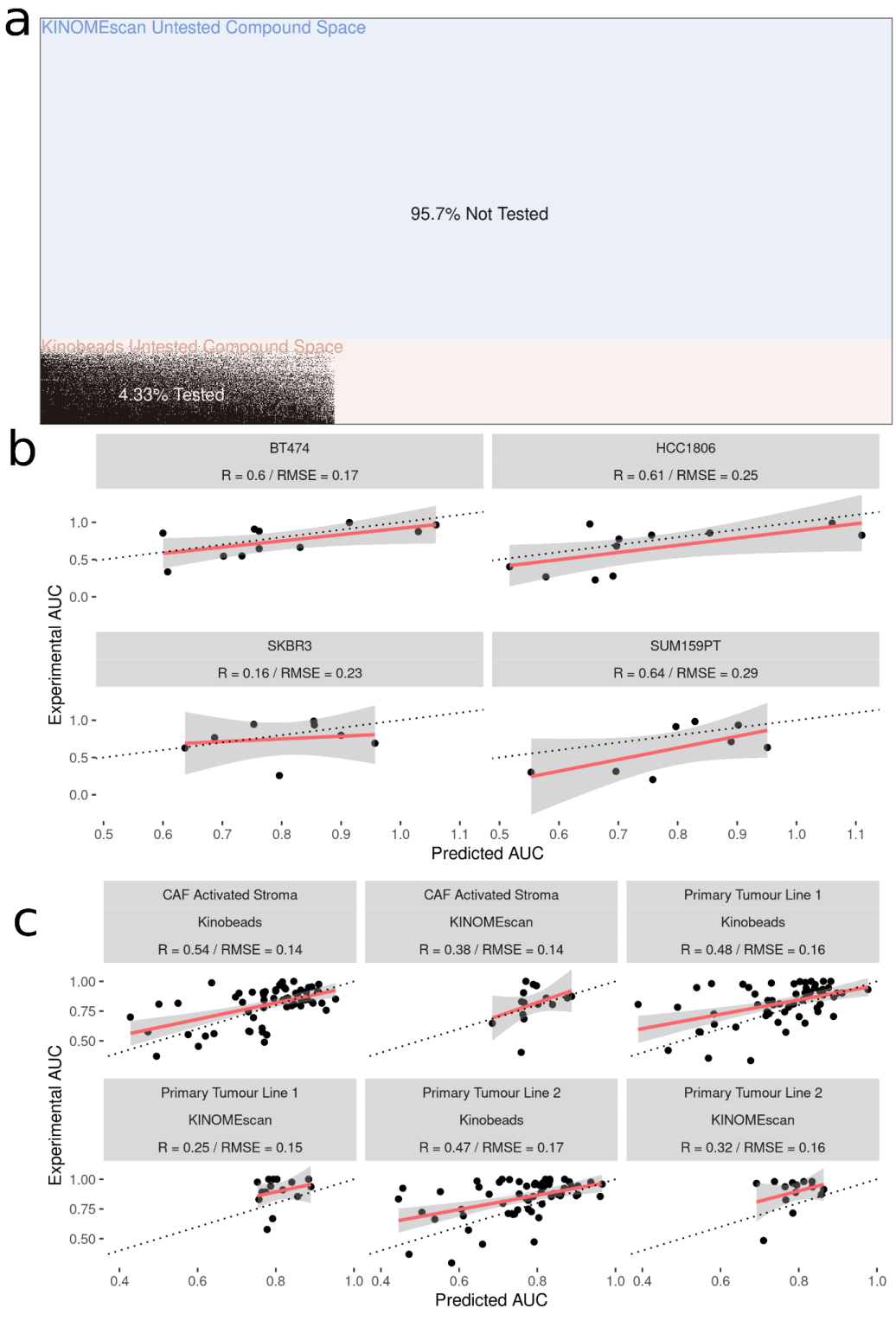

a
KINOMEscan Untested Compound Space
95.7% Not Tested
Kinobeads Untested Compound Space
4.33% Tested
b
BT474
R = 0.6 / RMSE = 0.17
HCC1806
R = 0.61 / RMSE = 0.25
SKBR3
R = 0.16 / RMSE = 0.23
SUM159PT
R = 0.64 / RMSE = 0.29
Experimental AUC
Predicted AUC
c
CAF Activated Stroma
Kinobeads
R = 0.54 / RMSE = 0.14
CAF Activated Stroma
KINOMEscan
R = 0.38 / RMSE = 0.14
Primary Tumour Line 1
Kinobeads
R = 0.48 / RMSE = 0.16
Primary Tumour Line 1
KINOMEscan
R = 0.25 / RMSE = 0.15
Primary Tumour Line 2
Kinobeads
R = 0.47 / RMSE = 0.17
Primary Tumour Line 2
KINOMEscan
R = 0.32 / RMSE = 0.16
Experimental AUC
Predicted AUC

Figure 3.4. Experimental Validation of Model in Breast Cancer Cell and Patient-Derived PDAC Cell Lines (a) Visualization of the space of compound (Y-axis) and cell line (X-axis) combinations that have been tested (white) and not tested (black) with colors denoting the origin of the drug kinome profiling data as Kinobeads (Pink) and KINOMEscan (blue). (b) Scatter plot showing relationship between AUC's predicted by model and experimentally generated AUC's for drugs not yet tested by PRISM in Breast Cancer Cell Lines (c) Scatter plot showing relationship between AUC's predicted by model and experimentally generated AUC's for drugs from Kinobeads and KINOMEscan tested in primary-tumor and stroma PDAC cell lines.

DISCUSSION

Kinase inhibitors are one of the fastest growing classes of targeted cancer therapies, but only a small fraction of the druggable kinome has been explored to date[16]. We have previously used data describing kinobeads-derived kinome inhibition states to predict cell viability in cancer cell lines in response to clinical kinase inhibitors and shown high prediction accuracy. However, public databases of kinome inhibition states derived through the more accessible KINOMEscan assay cover a large number of uncharacterized compounds and vastly widen the kinome space that can be targeted. In this work, we created a large integrated set of inhibitor-altered kinome states across the kinobeads and KINOMEscan assays, representing a broad space of kinome targets and including a host of tool compounds targeting understudied kinases. We linked these kinome inhibition states to cancer cell line responses to kinase inhibitors (dose response AUC and IC50). We then built machine learning models that integrate these kinome states with cell line baseline gene expression values to predict cell line response to kinase inhibitors. Finally, we predicted cell line sensitivity to previously untested kinase inhibitors in characterized breast cancer and patient derived PDAC cell lines and validated them experimentally.

Prediction of therapy response for cancer cell lines has been demonstrated through various methods, mostly utilizing chemical structure information, baseline gene expression and gene mutation status. Drug-target interaction data is relatively under-utilized for phenotype prediction, but offers opportunities for biological hypothesis building, especially for compounds with uncharacterized mechanisms of action. Kinome profiling data provides an exciting opportunity to use a wide array of functionally relevant drug-target interactions and is almost unique (except for GPCR inhibitors and HDAC inhibitors) in terms of ability to assay potential off target interactions. While we have previously shown that kinome profiling data generated from the kinobeads assay is informative for cell viability prediction, KINOMEscan data is much

more easily accessible, available publicly for uncharacterized compounds and easier to generate. In addition, it vastly broadens the kinome space capable of being targeted and increases the number of inhibitors that can be virtually screened as potential therapeutics.

It is important to note that this model linking kinome inhibition states to cell line response is generalizable to any human cancer sample, provided the sample has baseline transcriptomic data available. We have shown in this work and previously that the model can reasonably extrapolate to kinase inhibitors that have not been tested before in well-characterized cell lines. Significantly, in this work we have extended the scope of the model by using novel RNAseq data from patient derived PDAC cell lines, and testing kinase inhibitors against both tumor and stroma cell lines and achieving reasonable prediction accuracy.

Using tree-based models like gradient boosting lends us the ability to explain to some degree which features most affected cell viability. Shapley importance values generated from the best-performing model show that the inhibition states of kinases had overwhelmingly more predictive power compared to baseline gene expression values, with FLT3 as the most important feature. FLT3 mutations are observed in 30% of acute myeloid leukemia (AML) patients, and various FLT3 inhibitors are commonly prescribed for treatment[17]. However, the study dataset contains a majority of cell lines from non-small cell lung cancer (NSCLC), and FLT3 inhibitors have recently shown promise in preclinical studies by abrogating DNA damage[18]. Although the baseline gene expression features had considerably lower predictive power in the model, it is important to note that they provided crucial cell line specific context to the model. Interestingly, 40% of all the top 50 gene expression features were annotated as known kinase interactors in the STRING database. An additional strength of modeling cancer response from kinase-drug interactions is the generation of new hypotheses for understudied kinases. For example, the understudied[19] "Dark" kinases PIP4K2C and CSNK2A2 appear in the top 25 features of the

best-performing model, suggesting possible functional roles in cancer cell viability. Interestingly, PIP4K2C expression has been associated with outcomes in Acute Myeloid Leukemia (AML)[20], while CSNK2A2 has been associated significantly with prognoses of 14 different cancer types[21].

There are still many limitations to the results reported in this study. The creation of a combined kinome profiling dataset involves gluing together results from different assay types. Although the assays produce the same output (ratio of kinase in treatment sample to kinase in control), there may be numerous methodological artifacts that add noise to the data. We have attempted to address this by analyzing model performance on each assay type individually, and we can see that responses to inhibitors with data originating from KINOMEscan are noisier and more difficult to predict than inhibitors with data from kinobeads. This discrepancy is potentially due to the imbalance in training data availability for the KINOMEscan inhibitors, with only 15 inhibitors having annotated cell line sensitivity data available. In the future, as more cell line sensitivity testing is performed for compounds in the KINOMEscan dataset, model performance for this assay may improve. Additionally, model performance also decreases when shifting from gene expression data from well-characterized cancer cell lines to that of novel patient-derived cell lines. This is potentially due to the innate and significant heterogeneity that exists in such samples and because the models have been trained on baseline transcriptomics set of the given ~450 cell lines.

It should be possible to extend these models to incorporate multiple kinase inhibitors in combination. This is significant, given the frequency of resistance to targeted cancer monotherapies [22] and the potential to escape kinome reprogramming through multi-inhibitor combinations[23]. Thus, another area of future work is to combine the kinome inhibition states of multiple inhibitors to gain an understanding of their dual effect on the kinome, connecting

them to biological phenotypes that arise in response to inhibitor combinations to eventually build models and predict effective kinome-targeting combination therapies.

While targeted therapies such as kinase inhibitors have had significant clinical impact, much work remains to better understand how their modulation of the kinome leads to both desirable and undesirable phenotypic effects. The results presented here provide one approach where knowledge of the inhibition state of the kinome can be linked to downstream phenotypes through predictive models, greatly expanding our ability to predict the effects of existing targeted therapies as well as facilitating the design of novel ones.

METHODS

Data Sources

The primary data sources we used can be split into two categories: the integrated kinome profiling data set and the cancer cell line set. The following were downloaded from the respective supplementary materials to create the integrated set of kinome profiling data:

1. Kinome profiling data from the kinobeads assay
    a. Klaeger et. al 2017
2. Kinome profiling data from KINOMEscan assay
    a. LINCS: kinome profiling datasets for individual compounds downloaded programmatically from http://lincs.hms.harvard.edu/db/datasets/ [7]
    b. Kinome profiling data for the PKIS drug set was downloaded from the supplementary data of https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181585 [10]
    c. Kinome profiling data for the KCGS drug set was downloaded from the supplementary data of https://www.mdpi.com/1422-0067/22/2/566 [8] and from internal data provided by SGC-UNC.

The following were downloaded from the DepMap portal (https://depmap.org/portal/download/all/) to create the set of cancer cell line sensitivities and their gene expression characteristics:

1. DepMap secondary repurposing screen ("secondary-screen-dose-response-curve-parameters.csv")
2. CCLE gene expression set ("CCLE_expression.csv")

The high-throughput screening data gathered in PDAC patient-derived cell lines was gathered from Lipner et al. [14] with methods as described in Berginski et al 2021 [15].

Data Preprocessing

The scripts implementing these descriptions are all available through github.

Klaeger et al. Kinobead Kinase Inhibition Profiles: As previously described [6], we read the values from the supplemental data table into R and produced a filtered list of kinase and kinase interactor relative intensity values. We imputed missing values with the default "no interaction" value of 1, and truncated likely outlier values to the 99.99 percentile (3.43).

KINOMEscan Inhibition Profiles: We read in the three datasets mentioned above into R and concatenated them into a single combined data set. All the individual data sets contain identical protein lists because of the same assay type. Values are reported as "Percent Control", a ratio of protein pulled down in experimental condition (with inhibitor) vs control condition (without inhibitor). These were divided by 100 to convert the scale to 0-1 to match the Kinobeads relative intensity data.

Creating the Combined Kinome Inhibition Profiling Set: We took the kinobeads dataset and the KINOMEscan dataset and concatenated them into one large set containing inhibitor-kinase interaction states for ~800 total kinases and kinase interactors. We left out assays that included recombinantly mutated kinases, but left those with naturally occurring post-translational modifications. The vast majority (99.95%) of the inhibitor-kinase pairs represented was unique for either assay type, but for the 0.05% inhibitor-kinase pairs, we took the mean value of the measurements across the two assay types. Additionally, any missing values were imputed with the default "no interaction" value of 1. In the end we were left with kinome inhibition states for ~1000 kinase inhibitors.

Dataset of Cancer Cell Line Sensitivity to Drugs from DepMap: The DepMap repurposing dataset contains cell viability measurements across multiple doses, but since our dataset of kinome states is restricted to single-dose measurements, we extracted single summary statistics of cell line sensitivity to kinase inhibitors: Dose-response Area Under the Curve (AUC) and half-maximal Inhibitory Concentration (IC50). We extracted these by reading in the "secondary-screen-dose-response-curve-parameters" dataset into R, which contains curve parameters for a log-logisitic curve fit to the cell viability dose response curve and filtered it down to cell line name, IC50, AUC and other associated metadata.

Matching of Kinase Inhibitors between Profiling Dataset and Cell-Line Sensitivity Dataset: The compound names from each dataset were read into R, and the package Webchem [24] was used to retrieve PubChem compound IDs. The two sets of compound names were then matched based on these reference IDs. There were 252 matches between the two sets, forming a final set of ~70,000 inhibitor-cell line combinations.

Baseline Gene Expression: As described before[6] the RNAseq data provided in the "CCLE_expression.csv" file needed no modifications while preprocessing. Our only modification was to add identifiers to each gene label ("exp_"), to ensure that kinome inhibition data and expression data related to the same gene weren't accidentally combined.

String: The STRING database[25] was processed as described previously[6] to annotate kinases and kinase interacting genes.

Modeling Techniques

To assess our models we used a random 10-fold cross validation strategy. The number of features was varied as specified by the feature selection scheme described in the results section. We compared the performance of three model types using this strategy: LASSO (Least Absolute Shrinkage and Selection Operator) regression using the glmnet engine[26], random forest using the ranger engine[27] and gradient boosting using the XGBoost (eXtreme Gradient Boosting) engine[28]. Model performance was assessed by the R-squared value between predicted and actual outcome within the cross validation scheme. For each model type, we tuned sets of 30 hyperparameters to find the best possible performer as follows:

1. LASSO
    a. Penalty (1E-10 - 0.9)
2. Random Forest
    a. Trees (100 - 2000)
3. XGBoost
    a. Trees (100 - 1000)
    b. Tree Depth (4 - 30)

After final model selection, we fit the model on the entire dataset and then made predictions on inhibitor-cell line pairs not found in the original DepMap screening data.

Compound Testing

BT-474, HCC1806, SUM-159 and SKBR-3 cells were assayed as described previously [6]. Briefly, cells were grown in ATCC recommended media and seeded at 4000, 2000, 4000 and 500 cells per well respectively. 24 hours after seeding, cells were treated with inhibitors at 30 μM, 3 μM, 1 μM, 300 nM, 100 nM, 30 nM, 10 nM, and 3 nM, along with the appropropriate DMSO controls. seeded at, in white flat-bottom 96-well plates (Corning). Seventy-two hours post-treatment, cells were lysed with CellTiter-Glo (Promega) and luminescence was read using the PHERAstar FS microplate reader (BMG Labtech) and gain adjustments were conducted for each cell line. Data were normalized row-wise to the DMSO-only (0.1% on cells) control samples on each plate to calculate relative viability. Quality checks were performed to look at the data distribution and the presence of spatial bias on a plate. A quality control metric of <120% of DMSO was applied to all rows analyzed.

The functions "ComputeAUC" and "ComputeIC50" from the R package dr4pl [29] was used to fit a four-parameter log-logistic curve to the cell viability data, and extract AUC and IC50 values from the 9-point cell viability curves.

Software Availability

All of the code written to support this paper is available through github (https://github.com/gomezlab/kinomescan_viability_prediction) along with a walkthrough explaining where to find the code relevant to each part of the paper.

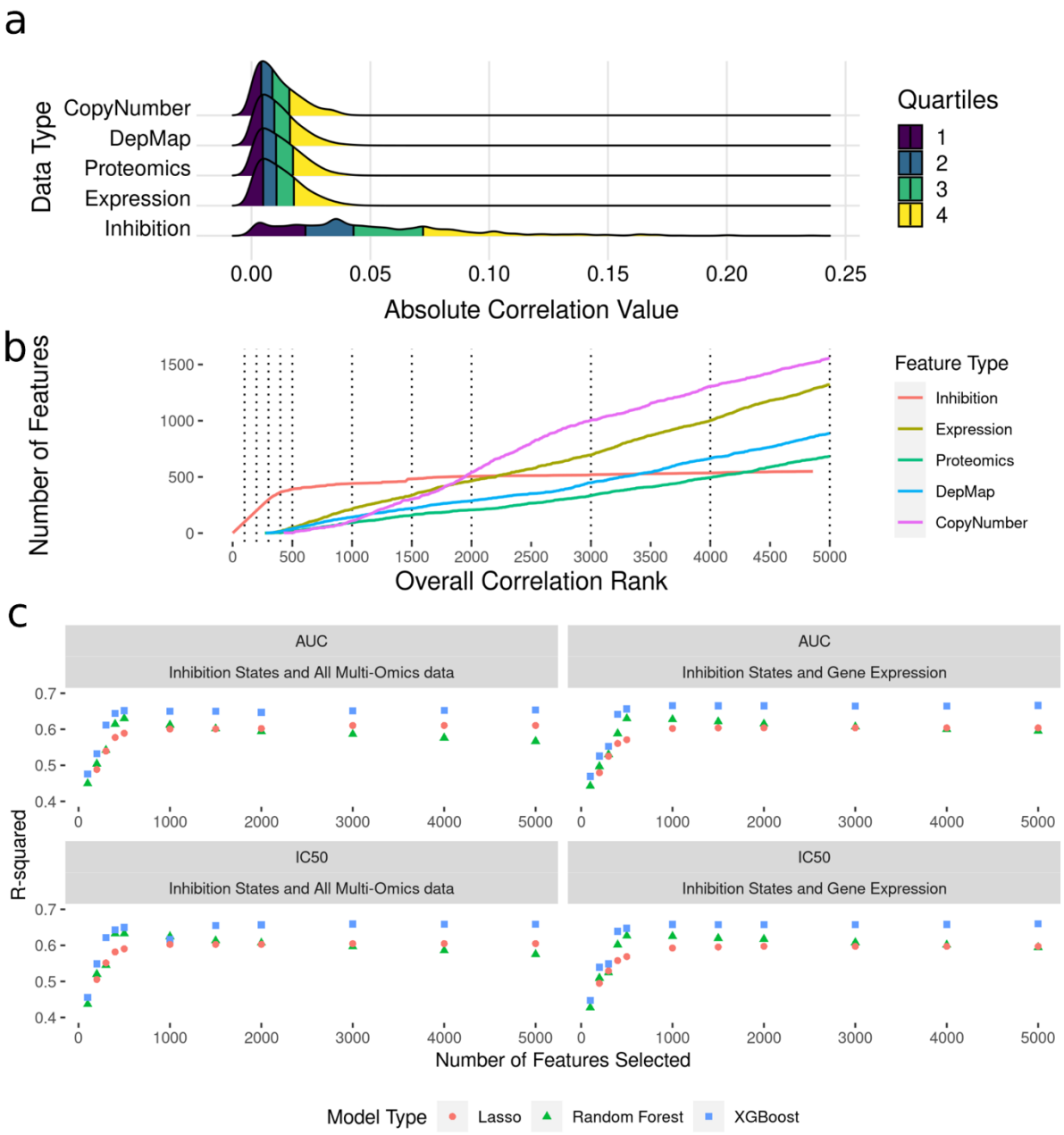
ACKNOWLEDGEMENTS

We would like to thank Madison Jenner for providing data access for our PDAC validation. We would like to thank UNC Research Computing for access to the computational resources necessary for this work.

Funding

This work was supported by grants through the National Institutes of Health (Grant #s CA274298, CA233811, CA238475, DK116204).

SUPPLEMENTARY MATERIAL



Supplementary Figure 3.1. Modeling Results from Integrating Multi-Omics Data with Kinome Inhibition States and Baseline Gene Expression to Predict Cell Line Sensitivity. (a) Ridgeline Plot showing distributions of correlations with drug IC50s and AUC values across the data types included in analysis. (b) Plots showing what order classes of features are selected from the ranked set of inhibition states and baseline gene expression values. The dotted lines indicate the discrete increments of feature rank cutoffs at which model performance was tested. (c) Model performance metrics (R-squared) for LASSO (orange dot), Random Forest (green triangle) and XGBoost (blue square).

REFERENCES

1. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol. 2014;32: 1202–1212.

2. Gönen M, Margolin AA. Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning. Bioinformatics. 2014;30: i556-63.

3. Reinecke M, Heinzlmeir S, Wilhelm M, Médard G, Klaeger S, Kuster B. Kinobeads: A chemical proteomic approach for kinase inhibitor selectivity profiling and target discovery. Methods and Principles in Medicinal Chemistry. Wiley; 2019. pp. 97–130. doi:10.1002/9783527818242.ch4

4. Patricelli MP, Nomanbhoy TK, Wu J, Brown H, Zhou D, Zhang J, et al. In situ kinase profiling reveals functionally relevant properties of native kinases. Chem Biol. 2011;18: 699–710.

5. Corsello SM, Nagari RT, Spangler RD, Rossen J, Kocak M, Bryan JG, et al. Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. Nat Cancer. 2020;1: 235–248.

6. Berginski ME, Joisa CU, Golitz BT, Gomez SM. Kinome Inhibition States and Multiomics Data Enable Prediction of Cell Viability in Diverse Cancer Types. bioRxiv. 2022. p. 2022.04.08.487646. doi:10.1101/2022.04.08.487646

7. Koleti A, Terryn R, Stathias V, Chung C, Cooper DJ, Turner JP, et al. Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. Nucleic Acids Res. 2018;46: D558–D566.

8. Wells CI, Al-Ali H, Andrews DM, Asquith CRM, Axtman AD, Dikic I, et al. The Kinase Chemogenomic Set (KCGS): An Open Science Resource for Kinase Vulnerability Identification. Int J Mol Sci. 2021;22. doi:10.3390/ijms22020566

9. Klaeger S, Heinzlmeir S, Wilhelm M, Polzer H, Vick B, Koenig P-A, et al. The target landscape of clinical kinase drugs. Science. 2017;358. doi:10.1126/science.aan4368

10. Drewry DH, Wells CI, Andrews DM, Angell R, Al-Ali H, Axtman AD, et al. Progress towards a public chemogenomic set for protein kinases and a call for contributions. PLoS One. 2017;12: e0181585.

11. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. J Open Source Softw. 2018;3: 861.

12. Yu C, Mannan AM, Yvone GM, Ross KN, Zhang Y-L, Marton MA, et al. High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. Nat Biotechnol. 2016;34: 419–423.

13. Rozemberczki B, Watson L, Bayer P, Yang H-T, Kiss O, Nilsson S, et al. The Shapley Value in Machine Learning. arXiv [cs.LG]. 2022. Available: http://arxiv.org/abs/2202.05594

14. Lipner MB, Peng XL, Jin C, Xu Y, Gao Y, East MP, et al. Irreversible JNK1-JUN inhibition by JNK-IN-8 sensitizes pancreatic cancer to 5-FU/FOLFOX chemotherapy. JCI Insight. 2020;5. doi:10.1172/jci.insight.129905

15. Berginski ME, Jenner MR, Joisa CU, Herrera Loeza SG, Golitz BT, Lipner MB, et al. Kinome state is predictive of cell viability in pancreatic cancer tumor and stroma cell lines. bioRxiv. 2021. p. 2021.07.21.451515. doi:10.1101/2021.07.21.451515

16. Laufer S, Bajorath J. New Horizons in Drug Discovery - Understanding and Advancing Different Types of Kinase Inhibitors: Seven Years in Kinase Inhibitor Research with Impressive Achievements and New Future Prospects. J Med Chem. 2022;65: 891–892.

17. Antar AI, Otrock ZK, Jabbour E, Mohty M, Bazarbachi A. FLT3 inhibitors in acute myeloid leukemia: ten frequently asked questions. Leukemia. 2020;34: 682–696.

18. Ryu H, Choi H-K, Kim HJ, Kim A-Y, Song J-Y, Hwang S-G, et al. Antitumor Activity of a Novel Tyrosine Kinase Inhibitor AIU2001 Due to Abrogation of the DNA Damage Repair in Non-Small Cell Lung Cancer Cells. International Journal of Molecular Sciences. 2019. doi:10.3390/ijms20194728

19. Essegian D, Khurana R, Stathias V, Schürer SC. The Clinical Kinase Index: A Method to Prioritize Understudied Kinases as Drug Targets for the Treatment of Cancer. Cell Rep Med. 2020;1: 100128.

20. Lima K, Coelho-Silva JL, Kinker GS, Pereira-Martins DA, Traina F, Fernandes PACM, et al. PIP4K2A and PIP4K2C transcript levels are associated with cytogenetic risk and survival outcomes in acute myeloid leukemia. Cancer Genet. 2019;233–234: 56–66.

21. Strum SW, Gyenis L, Litchfield DW. CSNK2 in cancer: pathophysiology and translational applications. Br J Cancer. 2022;126: 994–1003.

22. Lovly CM, Shaw AT. Molecular pathways: resistance to kinase inhibitors and implications for therapeutic strategies. Clin Cancer Res. 2014;20: 2249–2256.

23. Yesilkanal AE, Johnson GL, Ramos AF, Rosner MR. New strategies for targeting kinase networks in cancer. J Biol Chem. 2021;297: 101128.

24. Szöcs E, Stirling T, Scott ER, Scharmüller A, Schäfer RB. webchem: An R Package to Retrieve Chemical Information from the Web. J Stat Softw. 2020;93: 1–17.

25. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021;49: D605–D612.

26. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33: 1–22.

27. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw. 2017;77: 1–17.

28. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. arXiv [cs.LG]. 2016. Available: http://arxiv.org/abs/1603.02754

29. Gadagkar SR, Call GB. Computational tools for fitting the Hill equation to dose-response curves. J Pharmacol Toxicol Methods. 2015;71: 68–76.

CHAPTER 4: COMBINED KINOME INHIBITION STATES ARE PREDICTIVE OF CANCER CELL LINE SENSITIVITY TO KINASE INHIBITOR COMBINATION THERAPIES

INTRODUCTION

Protein kinases, which serve as the primary conduits for information transfer within cells, are often implicated as key drivers in cancer development and have thus become a cornerstone in current targeted therapies[1]. The rapid expansion of kinase inhibitor therapies as an oncology drug class is exemplified by the FDA's approval of nearly 60 such therapies over the past 20 years[2]. Despite their initial promise, kinase-targeting monotherapies frequently give rise to resistance[3], in part due to the dynamic nature of the kinase network, i.e. the "kinome," which has been shown to reprogram and respond to the inhibition of single kinases by upregulating expression of partner pathways[4–6]. This necessitates the development of novel strategies to effectively target the kinome and harness the vast array of potential drug targets it offers.

One emerging strategy to counteract the dynamic acquisition of kinase inhibitor resistance involves the design of combination therapies, which perturb multiple targets with two or more drugs. These targets may be either known compensatory pathway partners, referred to as "horizontal pathway inhibition," or multiple targets within the same pathway, known as "vertical pathway inhibition"[7]. This approach has recently gained traction with the FDA approval of the combination of trametinib and dabrafenib for treating advanced melanoma[8]. This combination therapy "vertically" targets both BRAF and MEK within the RAF-MEK-ERK (MAPK) pathway, demonstrating the potential effectiveness of this strategy. However, this

method of empirical design of combination therapies is not feasible for less characterized kinase pathways, and the sheer number of possible combinations of potential kinase targets ($2^{\sim 500}$) prevents comprehensive screening or drug design.

To circumvent this issue, computational screening offers an appealing alternative, enabling the prediction of effective drug combinations *in-silico* prior to testing a reduced pool of potential combinations *in-vitro*. This method not only potentially streamlines the drug development process but, when combined with patient-specific genomic profiling, can also enable personalized drug combination selection to potentially achieve resistance-proof responses in patients.

In recent years, a variety of computational approaches have been developed to predict combination therapy responses for cancer drug screening [9,10]. The majority of these methods primarily rely on drug structure characteristics and cancer-specific baseline genomic profiling to predict effective drug combinations, spurred by advancements in the high-throughput acquisition of these data types. For example, a high-dimensional tensor-based modeling strategy used similar data and achieved impressive accuracy (Overall $R^2 \sim 0.8$) in predicting response to combination therapies, validated experimentally[11]. This approach and others employ intricate neural network architectures that, while capable of producing high performing models, can be challenging to interpret, posing a barrier to the broader adoption and understanding of their underlying mechanisms. Tree-based machine learning models on the other hand, although simpler and sometimes less powerful, are generally considered interpretable depending on the type of data fed to them[12]. Notably, drug-protein interactions, which are intuitively central to the process of phenotype reversal, have been relatively underexplored in these computational approaches. In part, the minimal amount of drug-target information leveraged in current response prediction efforts is because of the sheer amount of data generated by genomics and

molecular fingerprinting, generating thousands of features for each measurement, while drug target data has been generally sparse with only a few annotated targets per drug. However, recent advances in technology to profile the interactions of clinical drugs with all the members of the kinome represent an unprecedented ability to measure drug-target information across ~500 proteins simultaneously in a quantitative manner[13,14]. The breadth, density, and ease of acquisition of this data, often measured at multiple dose points, is ideal for integration into machine learning models that can leverage diverse data types for drug response prediction.

Specifically, recent advances in proteomics techniques have facilitated the large-scale characterization of drug-kinase interactions, providing valuable information on the extent to which the entire kinome is inhibited by specific drugs or drug combinations. A landmark paper in 2017 used a mass spectrometry-based assay that used promiscuous kinase-binding compounds immobilized on beads to measure the binding competition between any given inhibitor and any given kinase (henceforth called the "kinobeads" assay)[15]. Using this assay, the kinome-wide binding profiles for ~230 clinical kinase inhibitors at eight doses each were elucidated using cancer cell lysates, forming the largest in-cell drug-target binding database publicly available at this time. The data generated from these assays allow interrogation of how clinical and investigational drugs interact with the entire kinome on an unprecedented scale. By analyzing the degree of inhibition of all kinases simultaneously for a given inhibitor, we can treat this as characterizing the degree of departure from the "baseline kinome state", thus moving through drug-induced alteration of multiple kinase activities to a new "kinome inhibition state". Given the degree to which modulation of the kinome alters cellular state and downstream behavior, these baseline kinome states and kinome inhibition states can be directly connected to various measured cellular phenotypes. We have recently demonstrated this idea by showing that kinome inhibition state is significantly predictive of cancer cell responses to kinase inhibitor

monotherapies when integrated with cancer-specific information, such as baseline transcriptomics, using tree-based machine learning models[16].

In this work, we show that by combining the inhibition states of two kinase inhibitors, we can generate a hypothetical "combined" inhibition state for an untested inhibitor combination. In this manner, we can rationally use all combinatorial kinome inhibition states to sample all possible kinase target combinations, hypothetically including all pathway partners. By integrating these inhibition states with cancer-specific baseline transcriptomics, we demonstrate that the combined inhibition state can predict the sensitivity of cancer cell lines to inhibitor combination treatments from the NCI-ALMANAC dataset using interpretable machine learning models. We further validate these models experimentally by examining novel inhibitor combinations in a PDX-derived triple-negative breast cancer (TNBC) cell line. By focusing on dual-inhibitor drug-kinase interactions combined with cancer-specific baseline genomic profiling, we can enhance computation combination drug screening pipelines with combinatorial kinase targeting. Furthermore, this approach lays the foundation for the rational design and a priori prediction of combination kinase inhibitor treatments for patients with the potential to ultimately reduce single kinase inhibitor resistance acquisition by prior rational targeting of partner pathways and associated kinases.

RESULTS

Creating a Set of Combined Kinome Inhibition States Representing Current and Potential Kinase Inhibitor Combination Therapies

In this work, we have focused on a specific set of 200 kinase inhibitors characterized using the kinobeads assay[15]. These inhibitors were profiled in-cell for their interactions with ~500 kinases and kinase-interacting proteins, across eight doses. From this data, as described previously (insert citation), we extracted monotherapy "kinome inhibition states", denoting the degree to which they inhibit each kinase in the kinome at eight doses on a scale of 0-1 (0 is complete inhibition and 1 is no inhibition of a given kinase).

We next tested different methods to approximate the kinome inhibition state of a kinase inhibitor combination. Intuitively, this can be thought of as simply superimposing two individual monotherapy inhibition states, but for the few cases where different inhibitors target the same kinase, we have to find ways to accurately reflect the resulting effect on the kinome. Here, we tested combining monotherapy kinome inhibition state vectors through addition, multiplication, truncated multiplication (excluding kinase inhibition values >1). All three methods were compared for downstream model performance.

After combining the individual inhibition states, we were left with a dataset describing all possible pairwise combinations of ~220 kinase inhibitors. These ~45,000 combinations represent the kinome inhibition states of existing clinical therapies (example), therapies currently in clinical trials (example), as well as potential therapies. Together, they interrogate a search space that includes nearly every known kinase on the phylogenetic tree (Fig S1).

Connecting Inhibited Kinome States with Cancer Cell Line Combination Sensitivities

Next, we linked the data set describing kinase inhibitor combinations to their cell sensitivity phenotypes in the large-scale ALMANAC drug combination screen. The ALMANAC screen contains cell sensitivity data for 53 kinase inhibitor combinations, over ~200 unique dose combinations for 45 cell lines across 9 cancer types. Additionally, previous high-throughput combination screens conducted in our lab in breast cancer offered data for 56 inhibitor combinations in four cell lines. Ideally, we would like exact matches between the dose at which kinome inhibition state is profiled and the dose at which cell sensitivity was measured. However, there are very few exact matches between the datasets. To overcome this, we found the nearest dose (at maximum differing by 1uM) at which kinome inhibition was profiled for each cell sensitivity measurement and connected the two datasets using these dose matches.

Additionally, we added cell line specific information to the dataset to complement the drug-specific kinome inhibition states. The CCLE database contains baseline transcriptomics data for ~1500 cancer cell lines, and almost all of the cell lines included in our data set were represented. Using this, we further added baseline gene expression into the dataset, now containing kinase inhibitor combinations, their inhibition state of the kinome, the cell line sensitivity to their treatment, as well as that cell line's baseline gene expression. In this way, the dataset connects the kinome inhibition states of inhibitor combinations to their cell sensitivity phenotypes.

The collected dataset represents a total of eight major cancer types, with the majority having ~7 cell lines represented each, while breast cancer had the most representation (11 cell lines). To ensure that the machine learning model downstream could find cancer-specific linkages between the kinome and cell sensitivity, we split the dataset into eight individual cancer type datasets, and conducted all modeling on each data split in parallel.

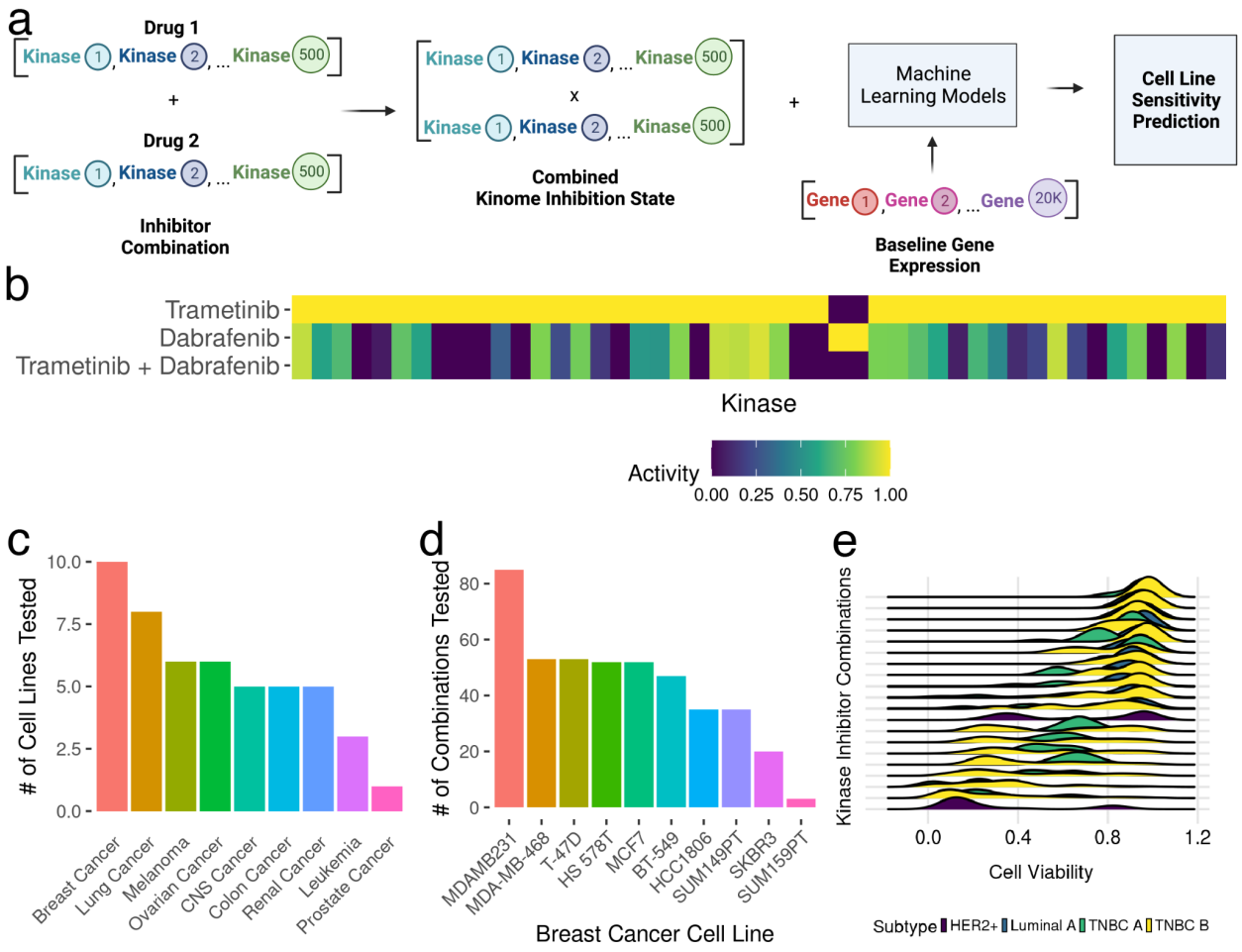


Figure 4.1. Kinome inhibition State Combination Modeling and Data Overview. (a) Schematic of modeling pipeline. (b) Heatmap showing the inhibition state of individual kinase inhibitors (row 1 and 2), and the hypothetical "combined" inhibition state for the two inhibitors (row 3) (c) Bar plot showing number of cell lines tested per cancer type in training data set (d) Bar plot showing number of unique combinations tested per cell line for the breast cancer subset of the training data set (e) Ridge plots showing cell viability (x-axis) variation for a random subset of different kinase inhibitor combinations (y-axis) in the NCI-ALMANAC data for breast cancer cell lines. Different breast cancer subtypes are represented with differing colors.

Elastic-Net Feature Selection Reveals Kinome Inhibition States to be Most Informative

In our collected dataset, kinome inhibition states and baseline gene expression together represent ~20,000 variables or "features" that could affect the phenotype of cell sensitivity to kinase inhibitors. It is both practically prohibitive and ineffective to build models using all available features, and so keeping in mind computational efficiency we sought to filter down the dataset to include only the most informative features. To accomplish this "feature selection", we built our machine learning pipeline starting with an elastic-net regression[17] model built against the outcome of cell sensitivity. This generated coefficients for each feature, with the absolute value of the feature coefficient directly proportional to its predictive value for the outcome. We ensured non-informative features were not included in modeling by only considering features with non-zero coefficients. We fit the model on the entire dataset to visualize a snapshot of the feature coefficients globally. This revealed overwhelmingly larger coefficients for kinome inhibition states compared to baseline gene expression (Fig 2a), thus indicating that kinome inhibition states were globally more informative for cell sensitivity prediction compared to baseline gene expression.

For downstream model building, the data set was split into a training and testing set five times (five-fold cross validation). For the training set data to not have any influence on the test set (to prevent data leakage), the elastic net model is fit on only the training data, and features are selected within each fold. Parameters for the elastic net model and hyperparameters for the tested model types were also tuned this way.

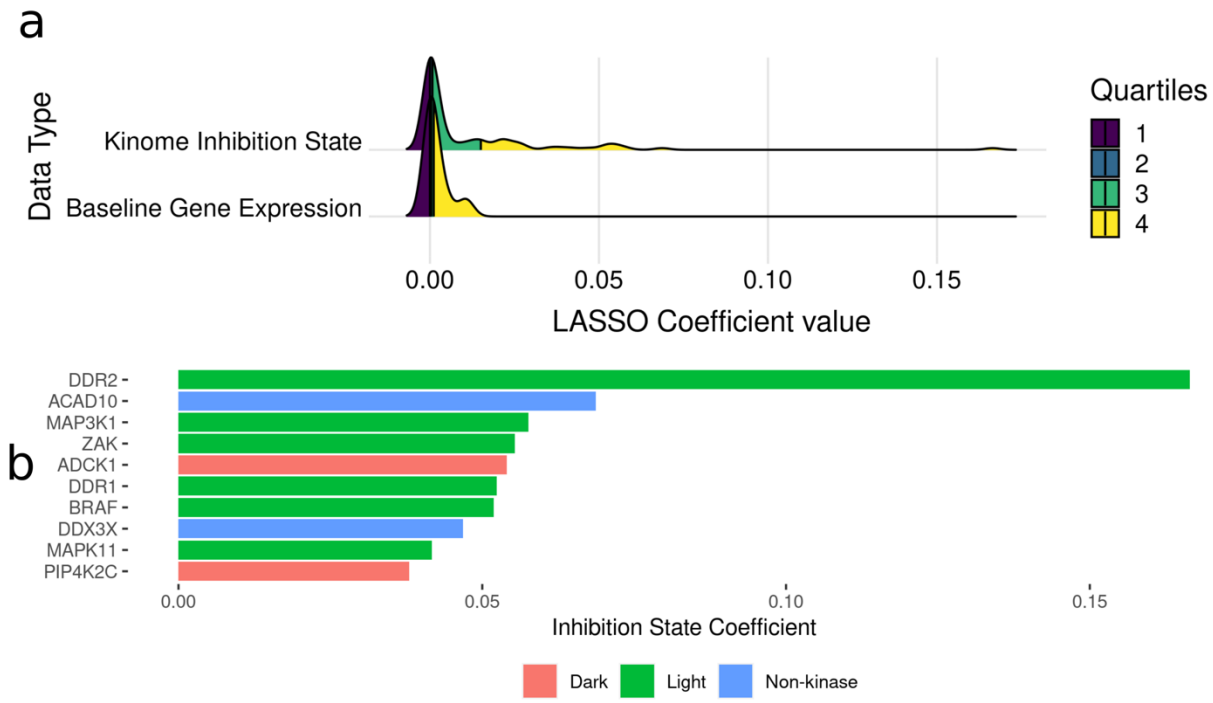


Figure 4.2. Feature Selection using an Elastic-net Regression Model against Cancer Cell Line Sensitivity. (a) Ridge plot showing the distribution of LASSO coefficient sizes as a metric for feature importance, for each feature type (b) Horizontal bar plot showing kinases with the largest elastic-net coefficient values, coloured by whether they are defined as "understudied" (Dark) or "well-characterized" (Light).

Machine Learning Models Can Predict Cancer Cell Line Sensitivity to Combination Therapies by Integrating Kinome Inhibition States and Baseline Transcriptomics

After data set preparation and feature selection, we built machine learning models that can predict cell sensitivity to kinase inhibitor combinations. For each cancer type, three machine learning model types were tested: random forest, boosted trees (xgboost) and deep neural networks. Xgboost performed the best for all cancer types, with type-specific performance largely dependent on abundance of data in the training set (Fig 3b). The most abundant cancer type (breast) had the best performing model with an R-squared score of 0.93 (Fig 3b) while the

lowest performing model was prostate cancer with R-squared = 0.73. Given our previous lab experience with breast cancer, we chose the breast cancer model for downstream experiments and validation.

Additionally, since the best-performing model was tree-based gradient boosting, we were able to further analyze the model using computed feature importance to find the most informative features in the data set based on the feature importance metric. Similar to the feature selection output, we saw much higher feature importance scores overall for kinome inhibition states when compared to baseline gene expression, and several kinases implicated in breast cancer dysfunction had high importance scores, such as MAP2K1/2 and EGFR(Fig. 3c).

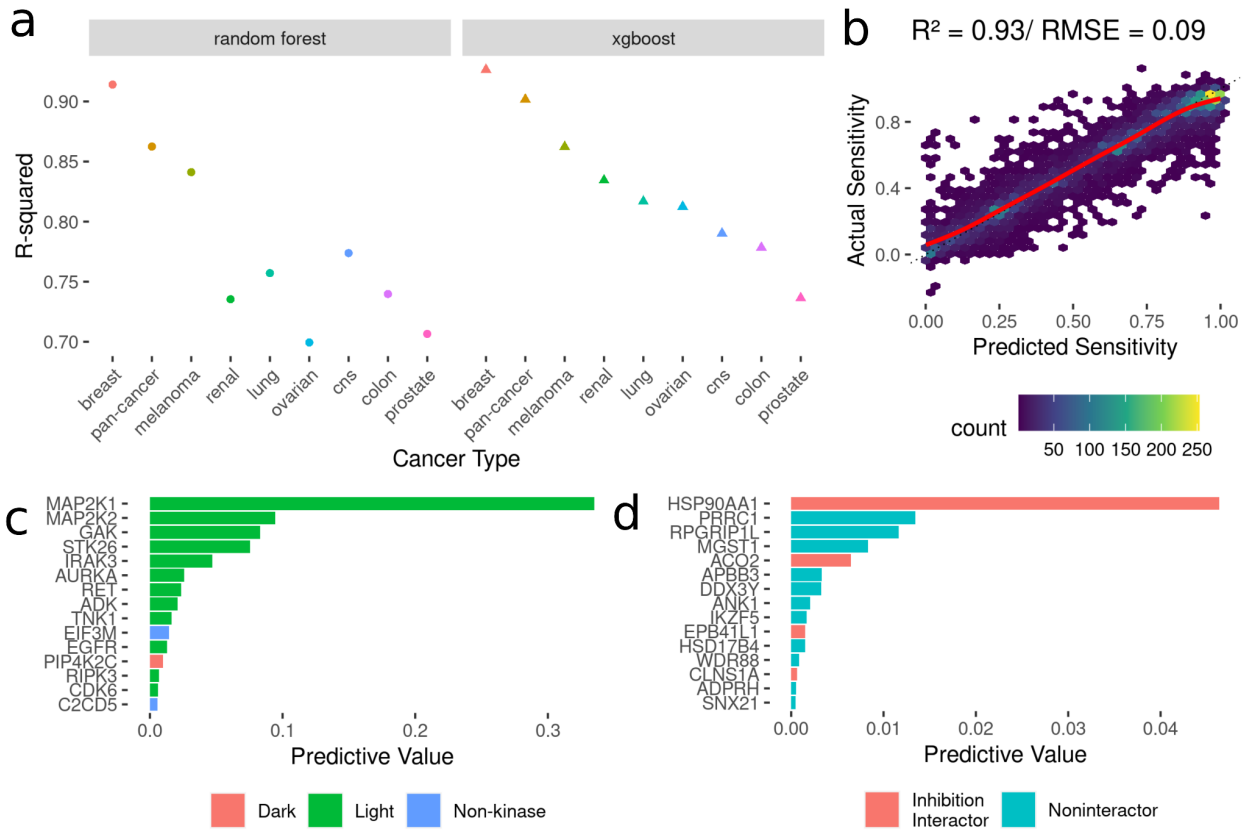


Figure 4.3. Development of Models to Predict Cancer Cell Line Sensitivities to Kinase Inhibitor Combination Therapies from Kinome Inhibition States (a) Model performance metrics (R-squared) for Random Forest (dots) and XGBoost (triangles). (b) Scatter Plot of predicted sensitivity values from the best-performing model vs actual sensitivity values. The red line indicates a smooth fit through the data points. (c) Horizontal bar plot showing model importance of individual kinase inhibition states by importance values. (d) Horizontal bar plot showing model importance of individual baseline gene expression by importance values.

Experimental Validation of Model Predictions in a PDX-Derived Triple Negative Breast Cancer Cell Line was Successful

We demonstrated that machine learning models using the kinome inhibition states of combination therapies along with cell-specific baseline gene expression could robustly predict cell sensitivity in multiple cancer types. However, to see if these predictive models could extend to real-world experiments, we experimentally validated 35 kinase inhibitor combinations in a PDX-tumor derived cell line(Fig 4A).

High-throughput cell line drug screens have been widely documented to suffer from a lack of reproducibility and poor translation to more complex samples like patient tumours. We sought to test whether our model of cell sensitivity in breast cancer, trained on 11 well-characterized immortalized cell lines, could effectively predict cell sensitivity in a PDX (Patient-Derived Xenograft) derived cell line. We chose the WHIM12 PDX-derived cell line, which was generated from a highly chemo-resistant TNBC tumor [18]. Previous experiments in the lab had conducted a drug combination screen in the WHIM12 cell line, out of which 35 kinase inhibitors were tested in combination with trametinib. Complementary baseline gene expression data was also generated through RNAseq. Using these in-house data, we were able to input the unseen

WHIM12 gene expression into the trained model, and predict the cell sensitivity outcomes of the conducted drug combination screen. We achieved robust prediction accuracy (Global R-squared = 0.74 / RMSE = 0.14) in predicting exact cell viability in response to treatment with 35 kinase inhibitor combinations, across 64 dose combinations (Fig 4c, d).

Model Predictions Reveal Known Synergy in trametinib/omipalisib Combination

The model predictions in the WHIM12 cell line were further interrogated for potential synergy. We generated synergy scores for all 35 combinations at each of the 64 dose points using the R package SynergyFinder[19] based on four different metrics: Zero-Interaction Potency[10] (ZIP), Bliss Independence[20], Highest Single-Agent (HSA), and Loewe Additivity[21]. Additionally, we generated similar synergy scores using the actual experimental data generated for validation as a comparison. We found a high degree of similarity (Global R-squared ~ 0.94/ RMSE ~ 0.5) between predicted and actual synergy, with trametinib + omipalisib as our most synergistic predicted combination, with a ZIP score of ~8 at certain dose combinations(Fig 4e, f). This is significant as the model predictions were in a TNBC PDX-derived line, and the trametinib/omipalisib combination represents the popular strategy of simultaneously targeting the MAPK and PI3K pathways[22].

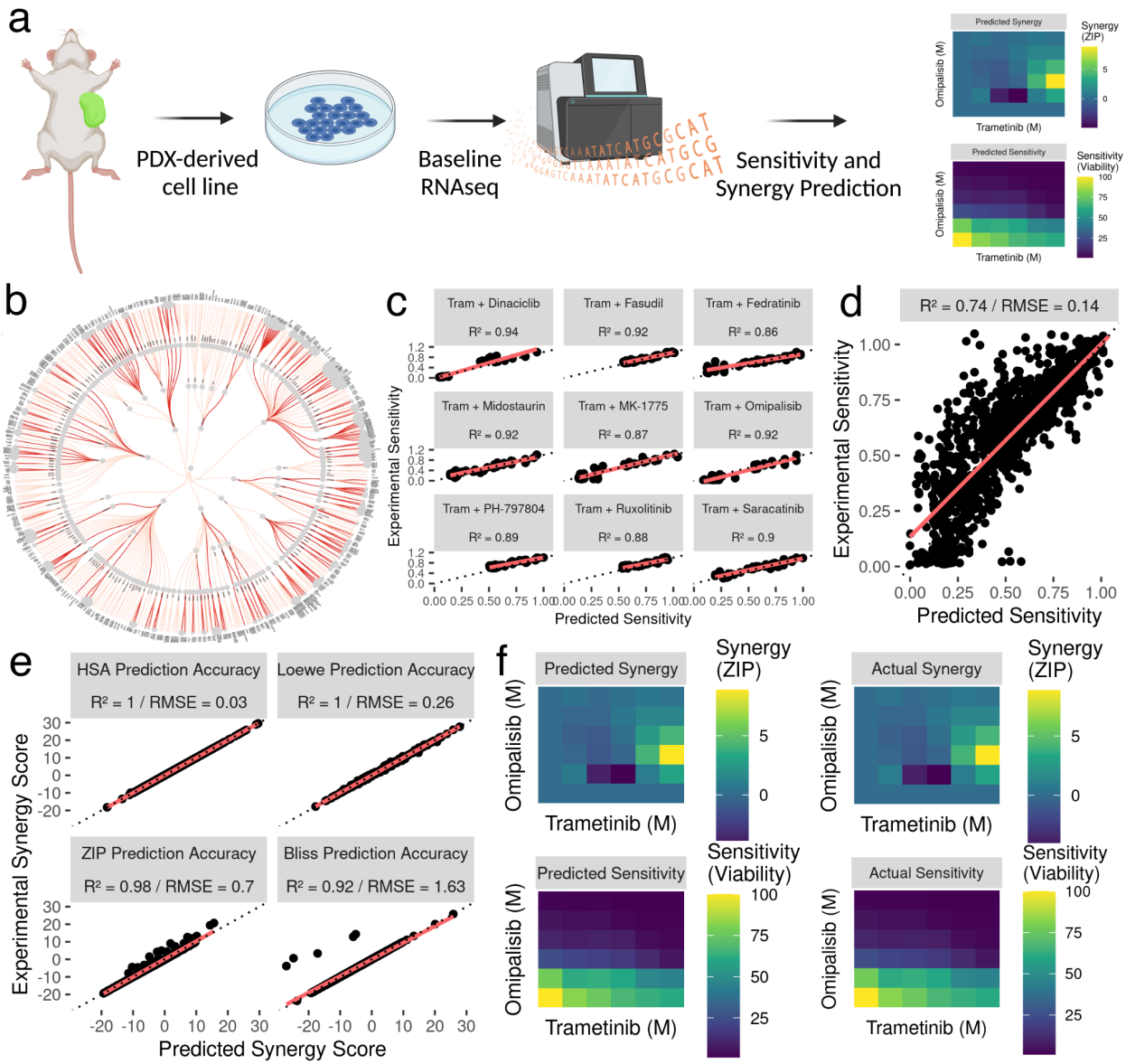

a
PDX-derived cell line
Baseline RNAseq
Sensitivity and Synergy Prediction
Predicted Synergy
Synergy (ZIP)
Omipalisib (M)
Trametinib (M)
Predicted Sensitivity
Sensitivity (Viability)
b
c
Tram + Dinaciclib R² = 0.94
Tram + Fasudil R² = 0.92
Tram + Fedratinib R² = 0.86
Tram + Midostaurin R² = 0.92
Tram + MK-1775 R² = 0.87
Tram + Omipalisib R² = 0.92
Tram + PH-797804 R² = 0.89
Tram + Ruxolitinib R² = 0.88
Tram + Saracatinib R² = 0.9
Experimental Sensitivity
Predicted Sensitivity
d
R² = 0.74 / RMSE = 0.14
Experimental Sensitivity
Predicted Sensitivity
e
HSA Prediction Accuracy R² = 1 / RMSE = 0.03
Loewe Prediction Accuracy R² = 1 / RMSE = 0.26
ZIP Prediction Accuracy R² = 0.98 / RMSE = 0.7
Bliss Prediction Accuracy R² = 0.92 / RMSE = 1.63
Experimental Synergy Score
Predicted Synergy Score
f
Predicted Synergy
Actual Synergy
Predicted Sensitivity
Actual Sensitivity
Synergy (ZIP)
Sensitivity (Viability)
Omipalisib (M)
Trametinib (M)

Figure 4.4. Experimental Validation of Model through a Trametinib Combination Screen in the WHIM12 Patient-Derived TNBC Cell Line. (a) Schematic showing experimental validation pipeline for the WHIM12 PDX-derived cell line. (b) Kinome phylogenetic map showing diversity of kinome space targeted (red = inhibited by a validated kinase inhibitor combination). (c) Grid of scatter plots showing accuracy of predicted vs experimental sensitivity to the top 9 tested combinations. For all scatter plots, the dashed line indicates where perfect predictions would lie and the red line shows a linear fit through the data. Quantitative accuracy is represented by the R-squared score. (d) Scatter plot showing the global accuracy of model predicted sensitivity compared to experimental sensitivity. (e) Grid of scatter plots showing accuracy of model predicted synergy scores compared to experimentally measured synergy scores across four metric types (HSA, Loewe, ZIP, Bliss). (f) Grid of heatmap plots showing comparison of predicted vs experimentally measured sensitivity and synergy for the highly synergistic Trametinib / Omipalisib combination.

DISCUSSION

Kinase inhibitors are one of the fastest growing drug classes for cancer therapy, with ~62 FDA approved in total against neoplasms[2]. With 500 potential druggable targets, there is significant interest in decoding the spectrum of kinases targeted by current inhibitors, and streamlining the kinase inhibitor screening process. We have previously introduced[16,23,24] the idea that the full spectrum of a given inhibitor's effect on the kinome as measured by recent advances in kinobead-competition/MS technology[15] can be represented as a "kinome inhibition state", i.e. a vector representing the effect of a given inhibitor on the kinome as a whole.

In this work, we have extended this idea to represent the kinome inhibition state of a combination of inhibitors, using a simple multiplicative probability model to "combine" the inhibition states of two given kinase inhibitors. By generating these "combined" inhibition states, we can vastly expand the search space targeted by inhibitor monotherapies, sampling all possible combinations of currently available therapies. To accomplish this, we used publicly available drug-kinome interaction data to generate snapshots of the combined effect of a combination therapy on the protein kinome. We then linked these kinome inhibition states of inhibitor combinations to cancer cell sensitivity phenotypes to combination treatment, creating a framework for predicting the efficacy of combination therapies in different cancer types.

We fit tree-based machine learning models as well as neural networks on this linked data set to robustly predict precise cancer cell line sensitivity and synergy for untested kinase inhibitor combinations therapies, and validate those predictions in complex patient derived samples. gradient-boosted tree models were highly accurate across cancer types (R-squared 0.75-0.93), comparable to two recent neural-network driven attempts to predict cell line response to drug combinations[9,11]. We chose to move forward with the highest performing breast cancer

model for further validation. We chose to validate our model predictions in the PDX-dervied WHIM12 line, reasoning that PDX-derived cell lines retain many of the molecular and genetic features of the xenografted original tumors and offer a closer representation of the disease state compared to traditional cell lines. We were able to show that the models performed robustly on novel baseline gene expression data (Global Accurcacy R-squared ~0.74) , representing its ability to extend to complex and clinical-adjacent samples compared to well-characterized cell line data.

One of the strengths of tree-based models compared to deep neural networks is that they are generally considered to be interpretable through feature importance computation[12,25]. Using this, we were able to investigate the "black box" and query which specific kinase inhibition states and baseline genes were most predictive of cell sensitivity. We found that for the best performing breast cancer model, the inhibition of the kinases MAP2K1/2 were the most informative by far. This is intuitive considering the most abundant kinase inhibitor in the dataset is the allosteric MEK inhibitor trametinib, but it must be noted that MEK inhibition is always only just one half of the kinome targeting in the combination. There has been increasing clinical interest recently in targeting the PI3K and MAPK pathways[22], and our lab has shown before that MEK1/2 inhibition in TNBC by trametinib induces widespread transcriptional adaptation, and that there is potential for clinical efficacy for complementary kinome targeting with trametinib[26]. Since our model's sensitivity predictions can effectively simultaneously predict synergy, our top synergy prediction for breast cancer according to the ZIP metric was trametinib and omipalisib, which we were able to validate experimentally in the WHIM12 line. This indicates that from the training breast cancer screening data, the model was able to learn that targeting the complementary PI3K and MAPK pathways is effective and synergistic in TNBC cell lines.

Interestingly, the predicted high-synergy combination of trametinib/omipalisib was recently in phase I clinical trials for advanced solid tumors but failed due to patient intolerability [27]. This highlights some limitations of our modeling approach. Ideally, kinome inhibition state would be one of many different drug modalities included for response prediction, and we plan to further expand these models in the future by considering toxicity, drug structure and cancer-describing multi-omic data types not limited to baseline gene expression. Additionally, in this proof-of-concept study we utilized a simple multiplicative probability model to generate the "combined" inhibition state of two inhibitors on the kinome, by assuming that the inhibition of a given kinase is mutually exclusive from that of other kinases. We know that kinases function physiologically as part of complex signaling networks, and their inhibition may have downstream effects on other kinases and signaling pathways. To address this limitation, future models will incorporate more biologically representative schemes to hypothesize combined kinome inhibition states.

In summary, through this work we demonstrate the development of a framework for predicting the efficacy of combination therapies in different cancer types using just kinome-drug interactions and baseline gene expression. We use a multiplicative probability model to generate the "kinome inhibition state" of a combination of kinase inhibitors and link these states to cancer cell sensitivity phenotypes. First, we were able to show that a given combination therapy's cancer-agnostic interaction with the kinome was far more informative than baseline genomics in predicting downstream response. This is intuitive fundamentally, as drug-protein interactions are the primary means of drug effect on physiology, but this type of data is still underutilized in computational screening approaches. We then used machine learning models to predict cell line sensitivity and synergy for untested kinase inhibitor combination therapies and validate those predictions in complex patient derived samples. We found that the inhibition of the kinases

MAP2K1/2 was the most informative for predicting breast cancer cell sensitivity, and the predicted high-synergy combination of trametinib/omipalisib was validated experimentally in a PDX-derived TNBC cell line.

METHODS

Data Sources

The primary data sources we used can be split into three categories: the kinome profiling data set, the combination-treated cell line sensitivity set, and the cancer cell line transcriptomics set:

The kinome profiling data set from the kinobeads assay was downloaded from the supplementary materials of Klaeger et al. 2017[15].

For cancer cell line sensitivity to kinase inhibitor combinations, data was downloaded from the following:

1. NCI-ALMANAC cell sensitivity data was downloaded from the NCI wiki database (https://wiki.nci.nih.gov/display/NCIDTPdata/NCI-ALMANAC).
2. Supplementary materials of previous lab combination screens published in Beville et al. 2019 [28] and Stuhlmiller et al. 2015 [29].

The CCLE gene expression set ("CCLE_expression.csv") was downloaded from the DepMap portal (https://depmap.org/portal/download/all/) to create the set of cancer cell lines and their gene expression characteristics.

In-house baseline gene expression data for the PDX-derived WHIM12 line was downloaded from the GEO repository for the Zawitowski et al. paper[26] (GSE87424).

Data Preprocessing

The scripts implementing these descriptions are all available through github.

Klaeger et al. Kinobead Kinase Inhibition Profiles: As previously described[16], we read the values from the supplemental data table into R and produced a filtered list of kinase and kinase interactor relative intensity values. We imputed missing values with the default "no interaction" value of 1 and truncated likely outlier values to the 99.99 percentile (3.43).

Creating the Combination Inhibition State Data Set: To create a "combined" inhibition state of a given kinase inhibitor combination, we sought to superimpose the inhibition states of two individual states at specific doses. There were eight doses measured for each individual inhibitor, thus there were 64 possible combinations for each combination. We took the monotherapy kinome inhibition states from the Klaeger et al. set and computed a "combined" inhibition state for each kinase, based on three different combination schemes:

1. Simple Multiplicative: The simple conditional probability rule assumes two independent events (Eq. 1). Since the default "no interaction" inhibition value is 1, for kinases that are not targeted by both inhibitors simultaneously, the "combined" inhibition state value is simply either one in monotherapy.
2. Truncated Multiplicative: A minority of measured kinase inhibition states (~1%) have values > 1 in the Klaeger et al. dataset, a possible artifact from the mass spectrometry measuring process. To avoid inflating those values, all >1 values were truncated at 1 and simple multiplication was performed as described above.
3. Addition: All kinase inhibition states were inverted into "Percent Inhibition" values, where 0 denotes no inhibition and 100 denotes complete inhibition. Then, when two inhibition states were combined they were simply added together (truncated at a max value of 100).

$$Equation\ 1. P(AB) = P(A) * P(B)\ \{If\ P(A|B) = P(A)\}$$

All three methods were tested in downstream modeling, resulting in minor variation. Truncated multiplied vectors were slightly more predictive (R-squared score of ~0.01 greater) so we used that scheme for all downstream modeling. In this way, we were able to compute hypothetical "combined" inhibition states for all possible combinations of ~220 inhibitors, altogether comprising ~2,000,000 combined inhibition states.

Dataset of Cancer Cell Line Sensitivity to Kinase Inhibitor Combinations: The cell sensitivity dataset from NCI-ALMANAC and previous lab publications were filtered to contain only kinase inhibitor small molecules, then summarized over replicates and converted to cell viability (1 = fully viable cell and 0 = full cell death). Relevant cancer types were annotated and individual cancer type datasets were subsetted for downstream cancer type-specific modeling.

Matching of Kinase Inhibitors between Inhibition State Dataset and Cell Line Sensitivity Dataset: The drug names from each dataset were read into R, and the package Webchem [30] was used to retrieve PubChem compound IDs (cid's). The two sets of drug names were then matched based on these reference IDs, with a total of ~100 matches between the two sets.

Baseline Gene Expression from CCLE: Data was preprocessed as described before[31] from the "CCLE_expression.csv" file. Cell line names were matched manually between CCLE and the NCI naming scheme. All cell lines represented in NCI-ALMANAC had a match in the CCLE database.

String: The STRING database[32] was processed as described previously[31] to annotate kinases and kinase interacting genes.

Modeling Techniques

To assess our models we used a random 5-fold cross validation strategy. The features included in each fold were selected as specified by the feature selection scheme described in the results section. We implemented Elastic-net regression using the glmnet engine[33] for the feature selection scheme[17], We compared the performance of three model types using this strategy: random forest using the ranger engine[34] and gradient boosting using the XGBoost (eXtreme Gradient Boosting) engine[35]. Model performance was assessed by the R-squared value between predicted and actual outcome within the cross-validation scheme. For each model type and for the feature selection model, we tuned sets of 20 hyperparameters to find the best possible performer as follows:

1. Elastic-net
    a. Penalty (0 - 0.1)
    b. Regularization (0.1-1)
2. Random Forest
    a. Trees (100 - 2000)
3. XGBoost
    a. Trees (100 - 1000)
    b. Tree Depth (4 - 30)

After final model selection, we fit the model on the entire dataset and then made predictions on the experimental validation data.

Experimental Validation

6x6 dose combination screens were performed in the WHIM12 cell line as described in Beville et al. 2019[28]. Briefly, cells were seeded in 384-well plates and dosed with drug after 24h. The screening library was tested for growth inhibition alone or in combination with Trametinib across 6 doses: 10 nmol/L, 100 nmol/L, 300 nmol/L, 1 μmol/L, 3 μmol/L, and 10 μmol/L. 0.1% DMSO was included as the control for growth inhibition on each plate. Plates were incubated at 37°C for 96 hours and lysed using CellTiter-Glo Reagent (Promega, catalog. no. G7570). Luminescence was measured using a PHERAstar FS instrument and growth inhibition was calculated relative to DMSO-treated wells.

Software Availability

All of the code written to support this paper is available through github (https://github.com/gomezlab/kinotype_combination_prediction) along with a brief walkthrough explaining where to find the code relevant to each part of the paper.

ACKNOWLEDGEMENTS

We would like to thank UNC Research Computing for access to the computational resources necessary for this work. We would like to thank Michael P. East for his help with data compilation.

Funding

This work was supported by grants through the National Institutes of Health (Grant #s CA274298, CA233811, CA238475, DK116204)

REFERENCES

1. Kothari V, Wei I, Shankar S, Kalyana-Sundaram S, Wang L, Ma LW, et al. Outlier kinase expression by RNA sequencing as targets for precision therapy. Cancer Discov. 2013;3: 280–293.

2. Roskoski R Jr. Properties of FDA-approved small molecule protein kinase inhibitors: A 2023 update. Pharmacol Res. 2023;187: 106552.

3. Jiang L, Li L, Liu Y, Lu L, Zhan M, Yuan S, et al. Drug resistance mechanism of kinase inhibitors in the treatment of hepatocellular carcinoma. Front Pharmacol. 2023;14: 1097277.

4. Duncan JS, Whittle MC, Nakamura K, Abell AN, Midland AA, Zawistowski JS, et al. Dynamic reprogramming of the kinome in response to targeted MEK inhibition in triple-negative breast cancer. Cell. 2012;149: 307–321.

5. Engelman JA, Zejnullahu K, Mitsudomi T, Song Y, Hyland C, Park JO, et al. MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. Science. 2007;316: 1039–1043.

6. Chandarlapaty S. Negative feedback and adaptive resistance to the targeted therapy of cancer. Cancer Discov. 2012;2: 311–319.

7. Yesilkanal AE, Johnson GL, Ramos AF, Rosner MR. New strategies for targeting kinase networks in cancer. J Biol Chem. 2021;297: 101128.

8. Atkinson V, Sandhu S, Hospers G, Long GV, Aglietta M, Ferrucci PF, et al. Dabrafenib plus trametinib is effective in the treatment of BRAF V600-mutated metastatic melanoma patients: analysis of patients from the dabrafenib plus trametinib Named Patient Program (DESCRIBE II). Melanoma Res. 2020;30: 261–267.

9. Kuenzi BM, Park J, Fong SH, Sanchez KS, Lee J, Kreisberg JF, et al. Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. Cancer Cell. 2020;38: 672-684.e6.

10. Yadav B, Wennerberg K, Aittokallio T, Tang J. Searching for Drug Synergy in Complex Dose-Response Landscapes Using an Interaction Potency Model. Comput Struct Biotechnol J. 2015;13: 504–513.

11. Julkunen H, Cichonska A, Gautam P, Szedmak S, Douat J, Pahikkala T, et al. Leveraging multi-way interactions for systematic prediction of pre-clinical drug combination effects. Nat Commun. 2020;11: 1–11.

12. Izza Y, Ignatiev A, Marques-Silva J. On Tackling Explanation Redundancy in Decision Trees. jair. 2022;75: 261–321bussy.

13. Cann ML, McDonald IM, East MP, Johnson GL, Graves LM. Measuring Kinase Activity-A Global Challenge. J Cell Biochem. 2017;118: 3595–3606.

14. Koleti A, Terryn R, Stathias V, Chung C, Cooper DJ, Turner JP, et al. Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. Nucleic Acids Res. 2018;46: D558–D566.

15. Klaeger S, Heinzlmeir S, Wilhelm M, Polzer H, Vick B, Koenig P-A, et al. The target landscape of clinical kinase drugs. Science. 2017;358. doi:10.1126/science.aan4368

16. Berginski ME, Joisa CU, Golitz BT, Gomez SM. Kinome inhibition states and multiomics data enable prediction of cell viability in diverse cancer types. PLoS Comput Biol. 2023;19: e1010888.

17. Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. J R Stat Soc Series B Stat Methodol. 2005;67: 301–320.

18. Li S, Shen D, Shao J, Crowder R, Liu W, Prat A, et al. Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts. Cell Rep. 2013;4: 1116–1130.

19. Ianevski A, Giri AK, Aittokallio T. SynergyFinder 2.0: visual analytics of multi-drug combination synergies. Nucleic Acids Res. 2020;48: W488–W493.

20. Bliss CI. THE TOXICITY OF POISONS APPLIED JOINTLY1. Ann Appl Biol. 1939;26: 585–615.

21. Chou TC, Talalay P. Quantitative analysis of dose-effect relationships: the combined effects of multiple drugs or enzyme inhibitors. Adv Enzyme Regul. 1984;22: 27–55.

22. Lee J, Liu H, Pearson T, Iwase T, Fuson J, Lalani AS, et al. PI3K and MAPK Pathways as Targets for Combination with the Pan-HER Irreversible Inhibitor Neratinib in HER2-Positive Breast Cancer and TNBC by Kinome RNAi Screening. Biomedicines. 2021;9. doi:10.3390/biomedicines9070740

23. Berginski ME, Jenner MR, Joisa CU, Herrera Loeza SG, Golitz BT, Lipner MB, et al. Kinome state is predictive of cell viability in pancreatic cancer tumor and stroma cell lines. bioRxiv. 2021. p. 2021.07.21.451515. doi:10.1101/2021.07.21.451515

24. Joisa CU, Chen KA, Berginski ME, Golitz BT, Jenner MR, Herrera Loeza SG, et al. Integrated Single-Dose Kinome Profiling Data is Predictive of Cancer Cell Line Sensitivity to Kinase Inhibitors. bioRxiv. 2022. p. 2022.12.06.519165. doi:10.1101/2022.12.06.519165

25. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. Explainable AI for Trees: From Local Explanations to Global Understanding. arXiv [cs.LG]. 2019. Available: http://arxiv.org/abs/1905.04610

26. Zawistowski JS, Bevill SM, Goulet DR, Stuhlmiller TJ, Beltran AS, Olivares-Quintero JF, et al. Enhancer Remodeling during Adaptive Bypass to MEK Inhibition Is Attenuated by Pharmacologic Targeting of the P-TEFb Complex. Cancer Discov. 2017;7: 302–321.

27. Grilley-Olson JE, Bedard PL, Fasolo A, Cornfeld M, Cartee L, Razak ARA, et al. A phase Ib dose-escalation study of the MEK inhibitor trametinib in combination with the PI3K/mTOR inhibitor GSK2126458 in patients with advanced solid tumors. Invest New Drugs. 2016;34: 740–749.

28. Bevill SM, Olivares-Quintero JF, Sciaky N, Golitz BT, Singh D, Beltran AS, et al. GSK2801, a BAZ2/BRD9 Bromodomain Inhibitor, Synergizes with BET Inhibitors to Induce Apoptosis in Triple-Negative Breast Cancer. Mol Cancer Res. 2019;17: 1503–1518.

29. Stuhlmiller TJ, Miller SM, Zawistowski JS, Nakamura K, Beltran AS, Duncan JS, et al. Inhibition of Lapatinib-Induced Kinome Reprogramming in ERBB2-Positive Breast Cancer by Targeting BET Family Bromodomains. Cell Rep. 2015;11: 390–404.

30. Szöcs E, Stirling T, Scott ER, Scharmüller A, Schäfer RB. webchem: An R Package to Retrieve Chemical Information from the Web. J Stat Softw. 2020;93: 1–17.

31. Berginski ME, Joisa CU, Golitz BT, Gomez SM. Kinome Inhibition States and Multiomics Data Enable Prediction of Cell Viability in Diverse Cancer Types. bioRxiv. 2022. p. 2022.04.08.487646. doi:10.1101/2022.04.08.487646

32. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021;49: D605–D612.

33. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33: 1–22.

34. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw. 2017;77: 1–17.

35. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. arXiv [cs.LG]. 2016. Available: http://arxiv.org/abs/1603.02754

CHAPTER 5: KINOME INHIBITION STATES WITH BASELINE GENOMIC PROFILING ARE PREDICTIVE OF PDX TUMOR RESPONSE IN FIVE COMMON SOLID CANCER TYPES

INTRODUCTION

The premise of precision oncology is to specifically target drivers of tumor growth in individual patients[1]. Since the approval of the Imatinib as the first "precision" therapy for Philadelphia chromosome positive CML, protein kinases have been a major focus in precision cancer therapy for the past 20 years[2]. As of today, kinase inhibitors are one of the largest classes of FDA-approved drugs and have transformed the standard of clinical care for multiple malignancies[3]. As a result, there has been widespread interest in developing similar strategies for individual patient focused treatment in oncology, complemented by advances in the high-throughput acquisition of baseline genomics[4]. However, resistance to kinase inhibitors is a major hurdle for effective pre-clinical and clinical drug development[5].

Patient-derived xenograft (PDX) models have emerged as important tools in precision medicine to aid effective development of targeted therapies to treat cancer patients[6]. Compared to traditional screening methods involving immortalized cell lines, PDX models retain many of the molecular characteristics of their original patients, providing a more realistic platform for pre-clinical testing[7]. In particular, PDX models have been shown to be useful in evaluating the efficacy of kinase inhibitors and predicting clinical responses to kinase inhibitors in patients[8]. For example, in the development of trametinib, a MEK1/2 inhibitor, PDX models derived from human melanoma tumors were used to evaluate the efficacy and safety in preclinical studies. Along with cell lines, PDX models enabled researchers to identify predictive

biomarkers for patient selection, such as the presence of *BRAF* V600E mutations, which helped inform the design of subsequent clinical trials[9]. Ultimately, trametinib was approved by the FDA in 2013 for the treatment of patients with unresectable or metastatic melanoma with *BRAF* V600E/K mutations, the first MEK inhibitor to get clinical approval[10].

Although categorizing patients according to individual biomarkers can be effective, information about key mutations is not always easily accessible, and even when biomarkers are identified, they may not reliably predict complete responses because of inter-patient and intra-tumor heterogeneity. Recently, machine learning models have leveraged the broad spectrum of baseline genomic profiling data to predict PDX tumor responses to targeted therapies by identifying complex patterns and correlations in baseline genomic profiling and imaging not immediately visible to the human eye[11–14]. However, despite drug-protein interactions being the primary and intuitive route for reversing phenotypes, the utilization of drug-target information in these precision oncology pipelines has been relatively limited, partly due to the lack of high-throughput drug-target screening technologies similar to genomic profiling. Recent advances in kinome profiling assays offer a way to leverage drug-specific information generated in high throughput for the 500-member kinase enzyme family simultaneously[15–17]. This approach allows us to capture a snapshot of kinase activity changes across the kinome influenced by a specific drug, henceforth referred to as a "kinome inhibition state." Further, recent landmark studies have profiled the target landscape of pre-clinical and clinical kinase inhibitors using assays such as Kinobeads and KINOMEscan (© DiscoverX), providing the opportunity to integrate this drug-specific knowledge into precision oncology pipelines. We have shown previously that adding in this information in an assay-agnostic fashion can improve the performance of predictive models for cell line responses across cancer types[18,19], but their predictive power in more complex preclinical models like PDX tumors remains unproven.

In this work, we demonstrate the integration of kinome inhibition states with baseline genomics using machine learning models that predict PDX tumor complete or partial response (CRorPR). We use response and baseline genomic profiling data from a high-throughput PDX screen conducted by Novartis[20], which tested a variety of monotherapies and combination therapies in five main cancer types: breast cancer (BRCA), cutaneous melanoma (CM), non-small cell lung cancer (NSCLC), colorectal cancer (CRC), and pancreatic ductal adenocarcinoma (PDAC). We then connect this data to our collected database of kinome inhibition states across assay types, while also using multimodal baseline genomic profiling to add cancer specific characteristics (gene expression, exome mutations, and copy number variation). We then perform simple correlation analysis to compare the predictive value of all the assay types, and then fit tree-based machine learning models to predict to a given treatment. This way we demonstrate the expansion of precision oncology models to include drug-specific data in the form of kinome inhibition states and improve both their performance and interpretability.

RESULTS

Connecting Kinome Inhibition States to PDX tumor responses

In this study, we aimed to connect the kinome inhibition states to the response of patient-derived xenograft (PDX) tumors to kinase-targeting therapies. From a precision oncology standpoint, our goal was to generate a rigorously tested probability of given PDX's response to any given kinase-targeted treatment. We utilized the Novartis high throughput PDX screen, which tested ~200 models across five cancer types (Fig 1B) against ~20 kinase-targeted therapies (Fig 1C), in addition to conducting multi-modal baseline genomic profiling. The response outcome of interest was a binary measure of complete response or partial response (CRorPR), determined by adapting the RECIST criteria[21]. To connect these responses to the treatment induced kinome inhibition states, we utilized kinome profiling data from the

kinobeads and kinomescan assays, complied as described previously. We then matched drugs between datasets for the monotherapy treatments and generated combined kinome inhibition states for combination therapies as described previously. The resulting dataset included the response of PDX tumors to kinase-targeted therapies, their kinome inhibition states, and baseline genomic profiling of three types (baseline gene expression, copy number variation. And exome mutations). Notably, the data was inherently unbalanced, as only 10% of PDX tumors fell into the "response" category. We split the data into five cancer type-specific sets for greater interpretability and specificity of downstream models. Our final dataset included 35 PDX models and ~10 kinase-targeting therapies per cancer type, and with the aim of building machine learning models to predict the response of PDX tumors, we first examined the relationships of all the multiomics data present in the dataset to the response outcome of CRorPR.

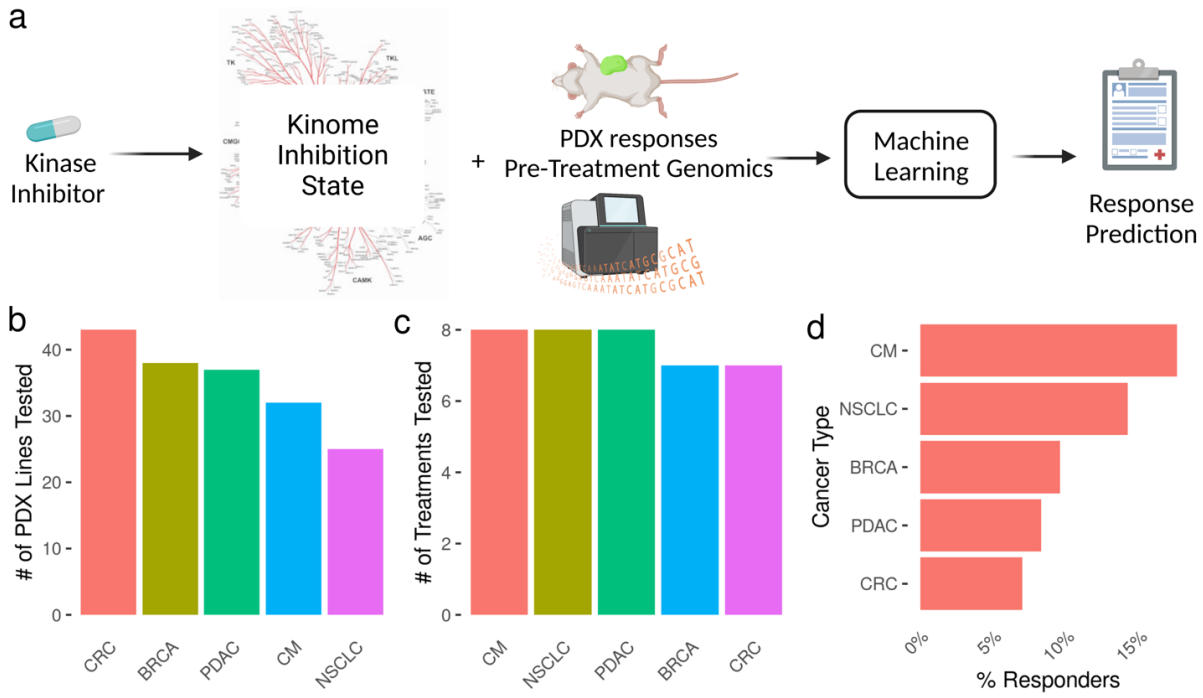


Figure 5.1. PDX modeling workflow overview and data description. A) Schematic describing data curation, integration and modeling strategy. B) Bar plot showing PDX models tested per cancer type. C) Bar plot showing kinase targeting therapies tested per cancer type. D) Visualization of number of responders per cancer type.

Examination of Bi-serial Correlations with Partial or Complete Tumor Response Reveals Kinome Inhibition States are Marginally More Informative than Genomic Data

We first investigated the predictive power of various individual data types in relation to the outcome of partial or complete tumor response (CRorPR). To achieve this, we used bi-serial correlation analysis, a statistical method that measures the strength of the association between a binary variable (CRorPR) and a continuous variable (all data types except exome mutations). For analysis of the relationship between CRorPR and exome mutations, we used tetrachoric correlation analysis[22]. We generated correlation coefficients for all different feature types against CRorPR for comparison: kinome inhibition states, copy number variation, baseline gene

expression, and exome mutations, allowing us to compare the predictive power of each feature type for CRorPR.

Overall, we observed that kinome inhibition states and single exome mutations showed the highest median correlation to CRorPR(Fig 3A). Kinome inhibition states were found to be the most informative feature type in 3 out of 5 cancer types, including in the pan-cancer dataset. Moreover, we identified a biologically relevant high correlation features for each cancer type, showing that the correlation analysis was able to recapitulate known relationships with tumor response. Notably, the central cell cycle kinases MAP2K1/2 were found to be the highest correlated features for PDAC(Fig 3B), which is known to be a crucial regulator in its progression and a potential candidate for targeted therapy[23]. Overall, these findings suggest that kinome inhibition states could be a valuable feature type in predicting CRorPR, by having marginally higher correlation overall to CRorPR compared to baseline genomic measurements, further highlighting the importance of incorporating feature types beyond genomic data in precision medicine pipelines. To further explore the predictive power of these features, we then built more complex machine learning models using gradient boosting, aiming to leverage the multi-variable correlations with response in a coordinated manner.

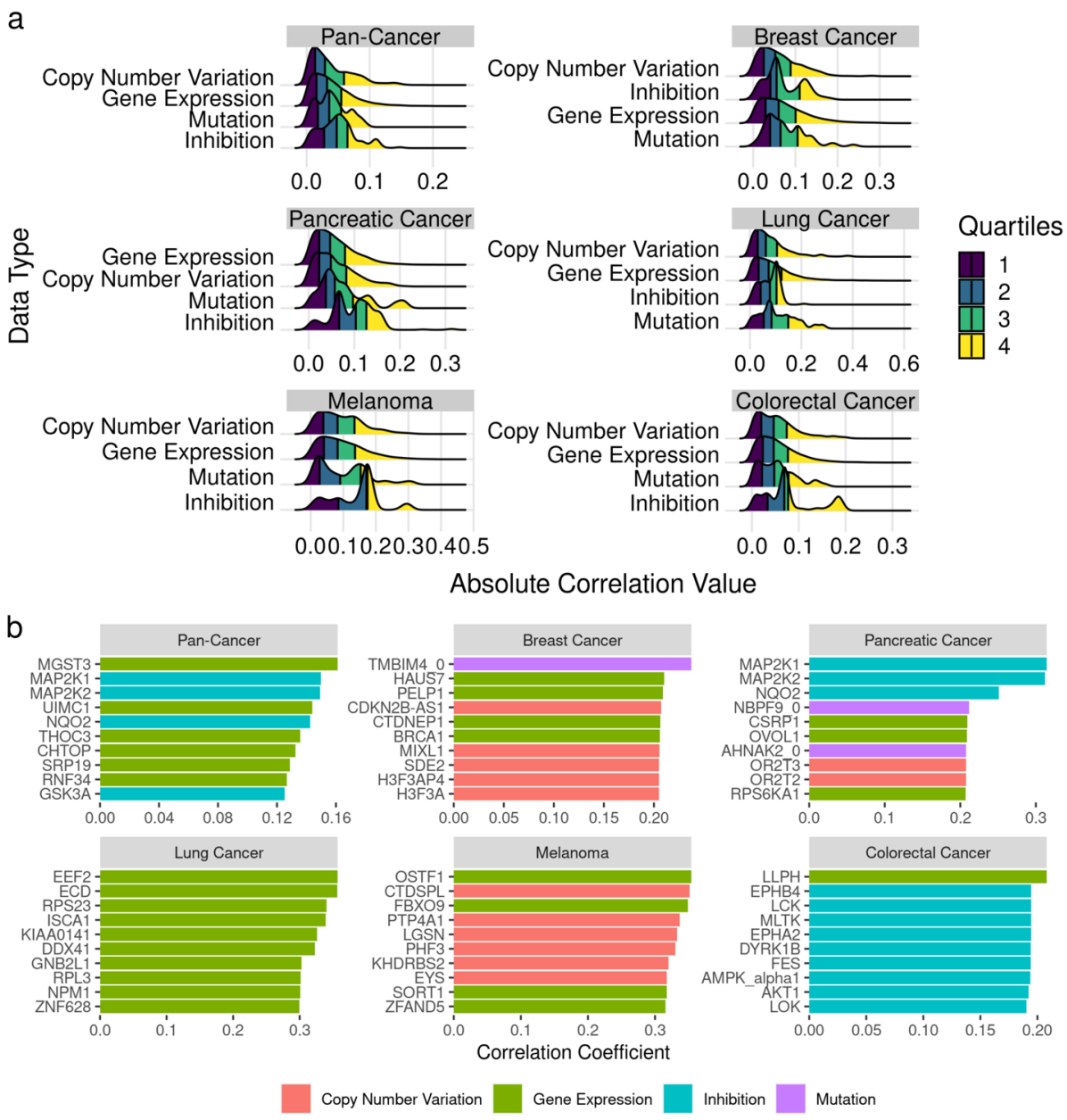

a
Pan-Cancer
Breast Cancer
Pancreatic Cancer
Lung Cancer
Melanoma
Colorectal Cancer
Copy Number Variation
Gene Expression
Mutation
Inhibition
Data Type
Quartiles
1
2
3
4
Absolute Correlation Value
b
MGST3
MAP2K1
MAP2K2
UIMC1
NQO2
THOC3
CHTOP
SRP19
RNF34
GSK3A
TMBIM4_0
HAUS7
PELP1
CDKN2B-AS1
CTDNEP1
BRCA1
MIXL1
SDE2
H3F3AP4
H3F3A
MAP2K1
MAP2K2
NQO2
NBPF9_0
CSRP1
OVOL1
AHNAK2_0
OR2T3
OR2T2
RPS6KA1
EEF2
ECD
RPS23
ISCA1
KIAA0141
DDX41
GNB2L1
RPL3
NPM1
ZNF628
OSTF1
CTDSPL
FBXO9
PTP4A1
LGSN
PHF3
KHDRBS2
EYS
SORT1
ZFAND5
LLPH
EPHB4
LCK
MLTK
EPHA2
DYRK1B
FES
AMPK_alpha1
AKT1
LOK
Correlation Coefficient
Copy Number Variation
Gene Expression
Inhibition
Mutation

Figure 5.2. Examination of bi-serial borrelations between all data types and PDX CRorPR A) Combined ridge plots showing distribution of individual feature correlation distributions per cancer type, with feature types arranged on Y axis from low to high median correlation value top to bottom. Colours and bars within ridge plots indicate quartiles. B) Combined horizontal bar plots showing the top 10 highest correlated features per cancer type, with colour indicating feature type. Order of horizontal bars indicates high to low correlation coefficient value from top to bottom.

Machine Learning Models Can Predict CRorPR of PDX Tumors from a Combination of Kinome Inhibition States and Baseline Genomics

After examining individual feature correlations with response, we sought to leverage the predictive power of all feature types including kinome inhibition states using machine learning to find combinations of features that are together informative.

In order to emphasize explainability of the fitted models, we tested gradient boosting (XGboost) in favor of deep neural networks and transformers, with 5-fold random cross validation stratified because of unbalanced response outcome (~10% mean response rate across cancer types). The collected features for modelling together in the final dataset number ~40,000. To ensure computational feasibility, reduce noise, and minimize collinearity, we performed feature selection on the dataset prior to fitting and tuning the xgboost models. Logistic regression against CRorPR was used for feature selection within each cross-validation fold to prevent data leakage. Additionally, hyperparameters for each model type and the elastic-net feature selector were also tuned within each cross-validation fold.

The results from the cross validation revealed robust model performance across cancer types, with the best-performing pan-cancer model demonstrating an area under receiver operating curve ~0.85 and area under precision recall curve ~0.5 (Fig 4A). Although no direct comparisons in data used exist for comparison with the literature, this outperforms similar models built on this dataset previously by 5-10%[12]. Notably, the low precision recall AUC is somewhat expected, since only 10% of samples belong to the responder class in the original dataset. Overall, there was some variance in overall cancer-type performance (ROC-AUC 0.75 – 0.85) with melanoma response being the easiest to predict and breast cancer the most difficult.

We further split the modelling effort into data-type specific sets to compare the predictive value of kinome inhibition states and baseline genomics. We tested kinome inhibition states alone against CRorPR, as well as each baseline genomics alone respectively. Expectedly, kinome inhibition states alone had little predictive power (Fig 4B dark blue bars), since they are drug-specific, and without any baseline characteristics specific to the individual cancer they will generate the same prediction for all tumors treated with the same therapy. However, by comparing the performance of all baseline genomics together against including kinome inhibition states along with baseline genomics, we see a significant increase in model performance across cancer types (Fig 4B orange bars vs other colors). This increase represents an improvement in ~5-10% depending on the cancer type.

Further, having confirmed the predictive power of kinome inhibition states in ensemble with baseline genomics, we can take advantage of the inherent interpretability in gradient boosting to determine which features were most informative for predicting CRorPR. We found that kinome inhibition states had higher feature importance scores overall, seeing a departure from the marginal increase we observed in the simple correlation analysis. This indicates that a

more complex model like gradient boosting was able to find associations between kinase inhibition states and other baseline genomics to extract more information overall from the kinome. Specifically, we saw high feature importance scores assigned in BRCA, PDAC, CM, and CRC for known cancer driving kinases involved in the MAPK and AKT pathways, as well as particular importance for the kinases ZAK and RIPK2.

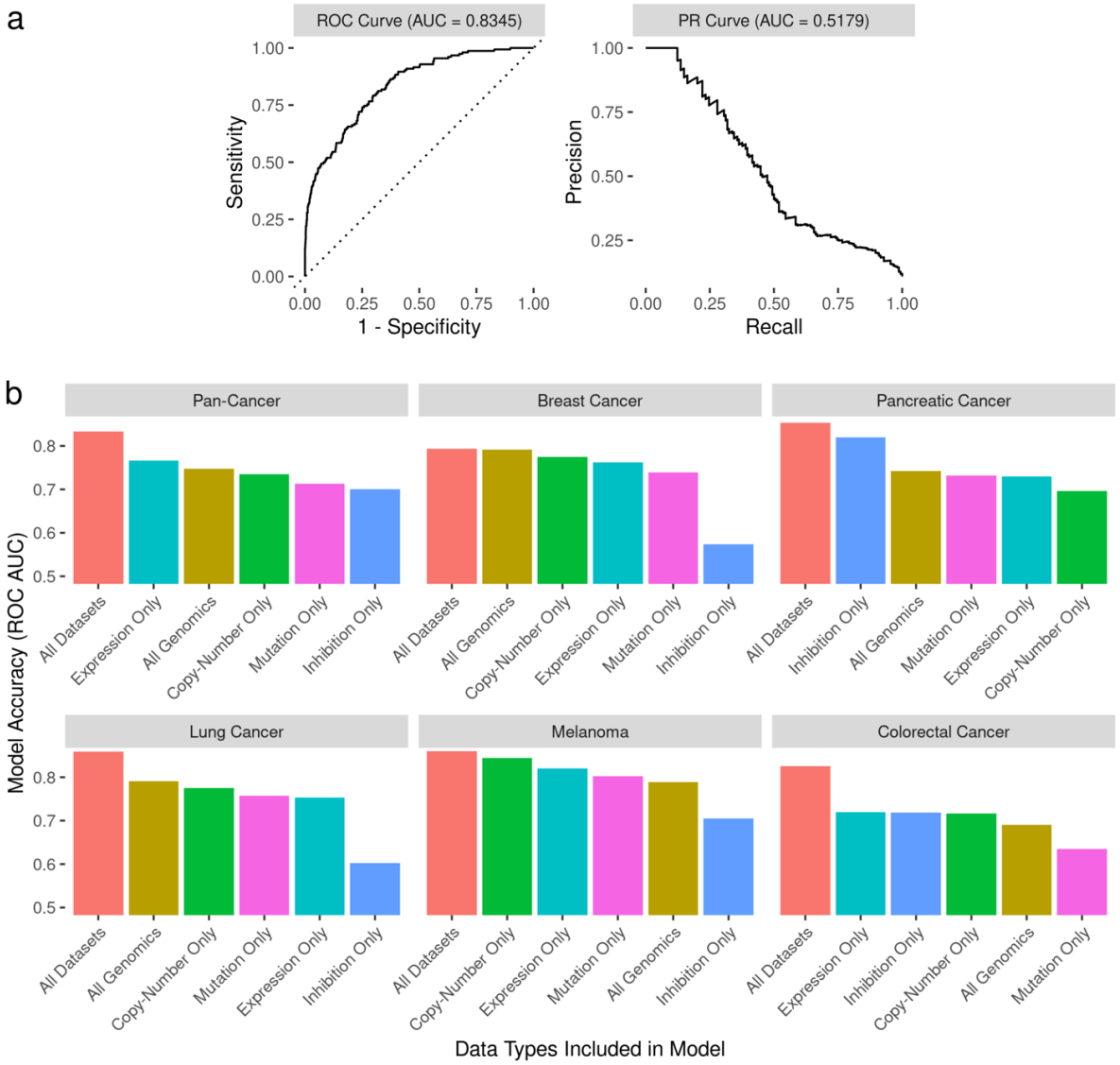


Figure 5.3. Model performance metrics across cancer types and data types. A) Left plot shows Receiver-Operating Curve (ROC) for the best performing pan-cancer model, with title indicating the Area Under the Curve (AUC) metric. Right plot shows Precision-Recall (PR) curve for the same model, with AUC indicated in the title. B) Six bar plots showing the accuracy of all models via ROC AUC per cancer type, with the data types included in the models being varied across the X axis. Colors also indicate data types.

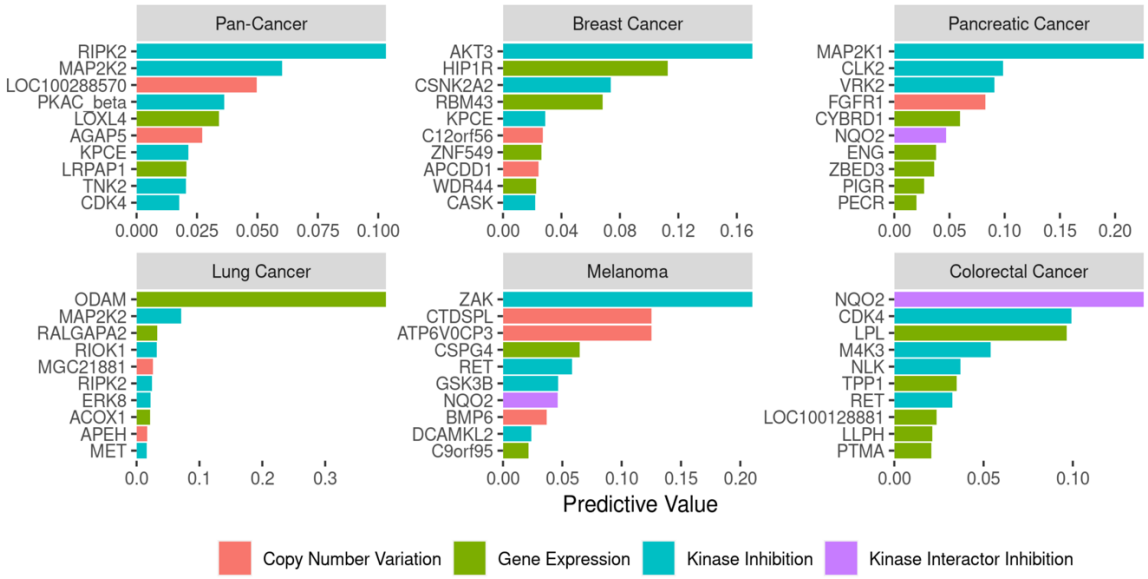


Figure 5.4. Interpretation of Best-Performing Models. Combined plot of six horizontal bar-plots, showing the top 10 most informative features for each best-performing model of a cancer type. Bars are colored according to their data type.

DISCUSSION

In this chapter, we leverage the interactions of PDX treatments with the kinome by using "kinome inhibition states" derived from kinome profiling assays and connect these to baseline cancer profiling and PDX CrorPR.

A recent high-throughput PDX screen conducted by Novartis tested ~200 models with a variety of kinase-targeting therapies across five common solid tumor types[20]. In this chapter, we used this data to connect the binary response (complete or partial) of these tumors to the kinome inhibition states induced by these treatments. In addition, we took advantage of the comprehensive multimodal baseline genomic profiling of each PDX model to combine model-specific characteristics with kinome inhibition states, aiming to create a full precision medicine platform. The collected dataset maps both drug-specific and cancer-specific omics data to the downstream responses of the tumors, and our first step was to analyze the base correlations of each individual variable (henceforth termed "feature") with the binary outcome of complete or partial response (CrorPR). Using bi-serial and tetrachoric correlation analysis, we found that a significant number of features across assay types had meaningful correlations with CrorPR. Overall, kinome inhibition states had the largest mean correlation in 3/5 cancer types, with exome mutations having the next-largest mean correlation. Notably, the inhibition state of the kinases MAP2K1/MAP2K2 were the most correlated feature by far for the response of PDAC tumors, recapitulating known information about the treatment efficacy of MAPK targeting in PDAC. These initial findings indicate that there is significant simple statistical association between inhibition state of the kinome and CrorPR of diverse PDX tumors, and there is potential value in including this data in precision oncology pipelines alongside baseline genomics.

To confirm the predictive value of kinome inhibition states, we built machine models on the collected data to leverage all the associations between the diverse features and CrorPR in a coordinated fashion. We showed that gradient boosting models fit on the data performed robustly, with an overall ROC AUC of ~0.85 for the pan cancer model, and ~0.75-0.85 for cancer-type specific models. By including data from each assay type individually and in tandem, we found that kinome inhibition states alone were not very predictive of CrorPR (ROC AUC ~ 0.6) but added significant accuracy to models including all baseline genomics data. This is expected, since the kinome inhibition states represent a drug-specific and tumor agnostic data type, which without any baseline genomics predicts the same response for the same treatment, regardless of the tumor to which the treatment is being applied. However, we see a 5-10% increase in accuracy of models fit with both baseline genomics and kinome inhibition states, compared to models with only baseline genomics alone. This suggests that the baseline genomics "tunes" the kinome inhibition states to the specific tumor for prediction of response, while the baseline genomics alone lacks any drug-specific information to enhance predictions.

Importantly, our decision to fit gradient boosting models in favor of deep neural networks or transformer-based networks was based on the critical nature of model explainability in clinical settings[24]. By probing the best-performing models we can generate feature importance scores for each model that are generally accepted to be indicative of a given feature's impact on prediction of a desired outcome[25]. After generating these scores for each best-performing model per cancer type, we were able to distinguish the top informative features for each of five solid tumor types in the dataset. Particularly, we confirmed our earlier finding from the correlation analyses that the inhibition state of MAP2K1 was overwhelmingly the most informative feature for predicting CrorPR in PDAC PDX tumors. Since PDAC is generally considered an aggressive cancer type with few treatment options, any potential therapeutic

strategies have a high impact on disease progression for patients[26]. Specifically, dysregulated MAP2K1 is known to be involved in pancreatic cancer progression, and its inhibition by small molecules has been shown to reduce pancreatic cancer growth and progression[27]. Moreover, targeting of this kinase by the inhibitor cobimetinib in combination with the EGFR inhibitor erlotinib is now in phase II clinical trials for pancreatic cancer (NCT03193190). This finding indicates that the PDAC-specific model was able to learn the significance of the kinome inhibition state of MAP2K1 in combination with baseline genomics to effectively predict CrorPR.

There are many limitations to the results presented in this work. First, the size of the available data restricts the analysis to include only ~20 kinase inhibitor monotherapies and combination therapies, with an average of ~10 per cancer type. Future improvements in model performance and scope of usable drugs can be expected with an increase in PDX response data gathered experimentally. Further, regardless of complexity and explainability of our fit models, they only provide a correlative link between kinome inhibition states, baseline genomics, and CrorPR, and obtaining the top 10 informative features per model is no substitute for a mechanistic study of treatment efficacy. Additionally, the responses to kinase-targeting therapies presented and predicted are very much subject to the heterogeneity present in the Novartis set of PDX response data. In future work we aim to complete thorough experimental validation in independent PDX treatments and in-house baseline genomic profiling to test the model's performance in real-world PDX-based precision oncology. Finally, although PDX models provide certain improvements over immortalized cell lines, they are still limited in their ability to replicate the complexity and tumor microenvironment of human cancers. In future work we hope to extend this modelling approach to generate treatment response probabilities directly in clinical settings.

In summary, through this work we demonstrate that the treatment-induced kinome inhibition state is powerfully predictive of PDX tumor response, in combination with multimodal baseline genomic profiling. Using both simple statistical correlation analysis as well as machine learning models, we were able to show robust prediction of PDX CrorPR (ROC AUC 0.85) for five common solid tumor types and interpret these models to find a specific use-case for MAP2K1 inhibition as a potential treatment strategy for PDAC tumors. Overall, this analysis will encourage the inclusion of kinome inhibition states along with baseline genomic profiling in precision medicine pipelines and improve prediction of individual patient responses to treatments.

METHODS

Data Sources

The primary data sources we used can be split into two categories: the kinome profiling data set of kinome inhibition states and the data set of PDX tumor baseline genomics with annotated responses to kinase-targeting therapies.

The kinome profiling data sets from the kinobeads and KINOMEscan assays were downloaded as described previously in chapter 3[19].

For PDX tumor response and baseline genomic profiling, data was downloaded from the supplementary material of Rashid et al. 2021[12].

Data Preprocessing

The scripts implementing these descriptions are all available through github.

Kinase Inhibition Profiles: As previously descrived, kinome profiles for ~200 drugs from the kinobeads assay and ~800 drugs from the KINOMEscan assay were integrated using the 1uM measurement, providing a dataset of ~500 kinase inhibition values ranging from 0 (full inhibition) to 1 (no inhibition). Additionally, for combination therapies, hypothetical "combined" inhibition states were generated as described previously in chapter 4.

Dataset PDX Tumor Response to Kinase-Targeting Therapies: The "split.cm.data.rda" file from the Rashid et. Al 2017 paper contains the full data for PDX responses to various therapies, as well as multimodal baseline genomic profiling (copy number variation, exome mutations, and gene expression). The data is presented for each cancer type individually and we retained all variables that were common across all of the cancer types (95% variables retained).

Matching of Kinase Inhibitor Mono- and Combination Therapies between Inhibition State Dataset and PDX response Dataset: The drug names from each dataset were read into R, and the package Webchem[28] was used to retrieve PubChem compound IDs (cid's). The two sets of drug names

were then matched based on these reference IDs, with a total of ~20 matches between the two sets.

Modeling Techniques

To assess our models we used a random 5-fold cross validation strategy. The features included in each fold were selected as specified by the feature selection scheme described in the results section. We implemented logistic regression using the glmnet engine[29] for the feature selection scheme using the "colino" package in R. We fit gradient boosting models using the XGBoost (eXtreme Gradient Boosting) engine[30]. Model performance was assessed by the ROC AUC value within the cross validation scheme. For Xgboost and for the feature selection logistic regression model, we tuned sets of 20 hyperparameters to find the best possible performer as follows:

1. Logistic Regression
    a. Penalty (0 – 0.1)
    b. Regularization (0.1-1)
2. XGBoost
    a. Trees (100 – 1000)
    b. Tree Depth (4 – 30)

Software Availability

All the code written to support this paper is available through github (https://github.com/gomezlab/PDX_response_prediction) along with a brief walkthrough explaining where to find the code relevant to each part of the paper.

ACKNOWLEDGEMENTS

We would like to thank UNC Research Computing for access to the computational resources necessary for this work. We would like to thank Novartis and the authors of Rashid et al. 2017 for making their code and processed data available online.

Funding

This work was supported by grants through the National Institutes of Health (Grant #s CA274298, CA233811, CA238475, DK116204)

REFERENCES

1. Heppner DE, Eck MJ. A structural perspective on targeting the RTK/Ras/MAP kinase pathway in cancer. Protein Sci. 2021;30: 1535–1553.

2. Yegnasubramanian S, Maitra A. Aiming for the outliers: cancer precision medicine through targeting kinases with extreme expression. Cancer discovery. 2013. pp. 252–254.

3. Li W, Croce K, Steensma DP, McDermott DF, Ben-Yehuda O, Moslehi J. Vascular and Metabolic Implications of Novel Targeted Cancer Therapies: Focus on Kinase Inhibitors. J Am Coll Cardiol. 2015;66: 1160–1178.

4. Darzynkiewicz Z. Novel strategies of protecting non-cancer cells during chemotherapy: are they ready for clinical testing? Oncotarget. 2011. pp. 107–108.

5. Denisova OV, Merisaari J, Kaur A, Yetukuri L, Jumppanen M, von Schantz-Fant C, et al. Development of actionable targets of multi-kinase inhibitors (AToMI) screening platform to dissect kinase targets of staurosporines in glioblastoma cells. Sci Rep. 2022;12: 13796.

6. Oh D-Y, Jung K, Song J-Y, Kim S, Shin S, Kwon Y-J, et al. Precision medicine approaches to lung adenocarcinoma with concomitant MET and HER2 amplification. BMC Cancer. 2017;17: 535.

7. Jung H-Y, Kim TH, Lee J-E, Kim HK, Cho JH, Choi YS, et al. PDX models of human lung squamous cell carcinoma: consideration of factors in preclinical and co-clinical applications. J Transl Med. 2020;18: 307.

8. Koga Y, Ochiai A. Systematic Review of Patient-Derived Xenograft Models for Preclinical Studies of Anti-Cancer Drugs in Solid Tumors. Cells. 2019;8. doi:10.3390/cells8050418

9. Gilmartin AG, Bleam MR, Groy A, Moss KG, Minthorn EA, Kulkarni SG, et al. GSK1120212 (JTP-74057) is an inhibitor of MEK activity and activation with favorable pharmacokinetic properties for sustained in vivo pathway inhibition. Clin Cancer Res. 2011;17: 989–1000.

10. Banks M, Crowell K, Proctor A, Jensen BC. Cardiovascular Effects of the MEK Inhibitor, Trametinib: A Case Report, Literature Review, and Consideration of Mechanism. Cardiovasc Toxicol. 2017;17: 487–493.

11. Mer AS, Ba-Alawi W, Smirnov P, Wang YX, Brew B, Ortmann J, et al. Integrative Pharmacogenomics Analysis of Patient-Derived Xenografts. Cancer Res. 2019;79: 4539–4550.

12. Rashid NU, Luckett DJ, Chen J, Lawson MT, Wang L, Zhang Y, et al. High-Dimensional Precision Medicine From Patient-Derived Xenografts. J Am Stat Assoc. 2021;116: 1140–1154.

13. Partin A, Brettin T, Zhu Y, Dolezal JM, Kochanny S, Pearson AT, et al. Data augmentation and multimodal learning for predicting drug response in patient-derived xenografts from gene expressions and histology images. Front Med. 2023;10: 1058919.

14. O'Farrell AC, Jarzabek MA, Lindner AU, Carberry S, Conroy E, Miller IS, et al. Implementing Systems Modelling and Molecular Imaging to Predict the Efficacy of BCL-2 Inhibition in Colorectal Cancer Patient-Derived Xenograft Models. Cancers . 2020;12. doi:10.3390/cancers12102978

15. Zhao Z, Bourne PE. Using the structural kinome to systematize kinase drug discovery. arXiv [q-bio.BM]. 2021. Available: http://arxiv.org/abs/2104.13146

16. Golkowski M, Lius A, Sapre T, Lau H-T, Moreno T, Maly DJ, et al. Multiplexed kinase interactome profiling quantifies cellular network activity and plasticity. Mol Cell. 2023;0. doi:10.1016/j.molcel.2023.01.015

17. Reinecke M, Heinzlmeir S, Wilhelm M, Médard G, Klaeger S, Kuster B. Kinobeads: A chemical proteomic approach for kinase inhibitor selectivity profiling and target discovery. Methods and Principles in Medicinal Chemistry. Wiley; 2019. pp. 97–130. doi:10.1002/9783527818242.ch4

18. Berginski ME, Joisa CU, Golitz BT, Gomez SM. Kinome inhibition states and multiomics data enable prediction of cell viability in diverse cancer types. PLoS Comput Biol. 2023;19: e1010888.

19. Joisa CU, Chen KA, Berginski ME, Golitz BT, Jenner MR, Herrera Loeza SG, et al. Integrated Single-Dose Kinome Profiling Data is Predictive of Cancer Cell Line Sensitivity to Kinase Inhibitors. bioRxiv. 2022. p. 2022.12.06.519165. doi:10.1101/2022.12.06.519165

20. Gao H, Korn JM, Ferretti S, Monahan JE, Wang Y, Singh M, et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. Nat Med. 2015;21: 1318–1325.

21. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, et al. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst. 2000;92: 205–216.

22. Muthén B. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika. 1984;49: 115–132.

23. Furukawa T, Kanai N, Shiwaku HO, Soga N, Uehara A, Horii A. AURKA is one of the downstream targets of MAPK1/ERK2 in pancreatic cancer. Oncogene. 2006;25: 4831–4839.

24. Yoon CH, Torrance R, Scheinerman N. Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? J Med Ethics. 2022;48: 581–585.

25. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. Explainable AI for Trees: From Local Explanations to Global Understanding. arXiv [cs.LG]. 2019. Available: http://arxiv.org/abs/1905.04610

26. Hu C, Dadon T, Chenna V, Yabuuchi S, Bannerji R, Booher R, et al. Combined Inhibition of Cyclin-Dependent Kinases (Dinaciclib) and AKT (MK-2206) Blocks Pancreatic Tumor Growth and Metastases in Patient-Derived Xenograft Models. Mol Cancer Ther. 2015;14: 1532–1539.

27. Mizuno S, Ikegami M, Koyama T, Sunami K, Ogata D, Kage H, et al. High-Throughput Functional Evaluation of MAP2K1 Variants in Cancer. Mol Cancer Ther. 2023;22: 227–239.

28. Szöcs E, Stirling T, Scott ER, Scharmüller A, Schäfer RB. webchem: An R Package to Retrieve Chemical Information from the Web. J Stat Softw. 2020;93: 1–17.

29. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33: 1–22.

30. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. arXiv [cs.LG]. 2016. Available: http://arxiv.org/abs/1603.02754

CHAPTER 6: CONCLUSIONS AND FUTURE DIRECTIONS

The work presented in this dissertation demonstrates the powerful potential in considering a novel data type for improving drug discovery and precision medicine in both investigational and clinical settings. For the first time, we have shown that inhibition state of the kinome can accurately predict drug responses in cancer cell lines and solid tumors when combined with current state of the art genomic profiling. We showed comparable or higher predictive performance of interpretable computational models including kinome inhibition states relative to current models considering baseline genomic profiling, drug structure, and complex modalities such as histology slide images. Importantly, we have also shown that this high accuracy extends to strict experimental validation in unseen samples, including patient-adjacent samples like PDX-derived cells. Together, this shows the immense potential in including novel data types like kinome inhibition states in the patient-specific design of targeted therapeutics across cancer types, laying the foundation for the design of mono and combination therapies using rational principles instead of empirical principles. Together, this work will improve drug development pipelines for kinase inhibitors starting from discovery to clinical testing and increase the chance of finding new transformative targeted therapies like Gleevec.

In the future, the main limitations for extension of this work are the lack of high-throughput interrogations of the kinome. The thorough characterizations used in this work literature recently are relatively rare and unique and cover only a small percentage of known kinase inhibitors. However, given the drug-specific and cancer-agnostic nature of kinome

profiling, each acquisition assay for a given inhibitor needs to be run only once, after which it can be applied as shown along the entire drug discovery and treatment pipelines. Ongoing projects show promise in characterizing more optimized kinase inhibitor libraries that are engineered to target all families of the kinome equally, increasing targeting diversity and potentially novel therapeutic promise. The other important limitation for future work is the lack of high-throughput data annotating response of cancer samples to kinase inhibitors. Even in the most expansive dataset covering immortalized cancer cell line responses to monotherapy kinase inhibitors, we are currently limited to data characterizing ~120 inhibitors, compared to our collection of inhibition state data for ~1000 inhibitors. Moving forward, the mining of cell line, tumor, and patient phenotypes in response to kinase inhibitor treatments is a crucial focus for extensions of this work and represents exciting prospects for predicting not only direct response phenotypes, but also specific imaging and clinical phenotypes.

Finally, the long-term goal for this work is eventual inclusion in patient-facing settings, designing precision therapies based on both tumor-specific characteristics and drug-specific effect on the kinome. While we have demonstrated initial limited success in predicting response of cancer-type specific PDX tumors with a high level of accuracy, we need more extensive data and in-house experimental validation to verify the validity of this approach in patient samples. Another opportunity for patient-facing applications is the modelling of patient responses to clinical trial compounds. We in the lab have put significant effort in curating publicly available clinical trial data with annotated patient responses as well as baseline genomic profiling, and we hope to extend the work present in this dissertation to directly stratify predicted patient responses.

Overall, the work presented in this dissertation shows the powerful predictive value of the kinome and lays a solid foundation for the inclusion of kinome inhibition information in the rational design of targeted therapies for cancer.