Predicting drug (combination) response through data integration

The whole is greater than the sum of its parts

Aben, Nanne

DOI
10.4233/uuid:765e7d74-7d60-407a-91fe-3d585724cb96

Publication date
2019

Document Version
Final published version

Citation (APA)
Aben, N. (2019). *Predicting drug (combination) response through data integration: The whole is greater than the sum of its parts*. [Dissertation (TU Delft), Delft University of Technology].
https://doi.org/10.4233/uuid:765e7d74-7d60-407a-91fe-3d585724cb96

PREDICTING DRUG (COMBINATION) RESPONSE THROUGH DATA INTEGRATION

THE WHOLE IS GREATER THAN THE SUM OF ITS PARTS

PREDICTING DRUG (COMBINATION) RESPONSE THROUGH DATA INTEGRATION

THE WHOLE IS GREATER THAN THE SUM OF ITS PARTS

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op donderdag 31 oktober 2019 om 15:00 uur

door

Nanne Nicolaas ABEN

Master of Science in Computer Science,
TU Delft, Delft, Nederland,
geboren te Schiedam, Nederland.

Dit proefschrift is goedgekeurd door de promotor.

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. L.F.A. Wessels	Nederlands Kanker Instituut Amsterdam & Technische Universiteit Delft, promotor
Onafhankelijke leden:	
Prof. dr. A. Berns	Nederlands Kanker Instituut Amsterdam & Universiteit van Amsterdam
Dr. ir. J. de Ridder	UMC Utrecht
Prof. dr. A.B. Houtsmuller	Erasmus MC Rotterdam
Prof.dr. ir. E.P.J.G. Cuppen	UMC Utrecht
Prof. dr. ir. D. de Ridder	Wageningen University & Research
Prof. dr. ir. M.J.T. Reinders	Technische Universiteit Delft
Prof. dr. ir. R.L. Lagendijk	Technische Universiteit Delft, reservelid

Printed by: Ipskamp Printing

Front: Microscope picture of A549 tumor cells from one of the anti-cancer drug screens performed at the Sanger institute. Special thanks to Patricia Jaaks and Syd Barthorpe for providing the picture.

ISBN 000-00-0000-000-0

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

CONTENTS

SUMMARY

In order to improve anti-cancer treatment, we need to better understand why some patients respond to a given anti-cancer treatment, while others do not. To this end, several large-scale drug response screens have been performed in recent years, in which hundreds of tumor cell lines have been characterized for many molecular features (e.g. mutations, CNAs, methylation and gene expression), as well as for response to hundreds of anti-cancer drugs. By statistically associating these molecular features with the drug response, we can identify biomarkers of drug response: markers that (after thorough testing) can ultimately be used to help identify which treatment should be given to which patient.

While performing such statistical analyses, we found that there are strong relationships between the different molecular datasets (e.g. mutations, CNAs, methylation and gene expression) and that these relationships can negatively affect our ability to identify biomarkers. Following these results, we have developed TANDEM, a method to identify biomarkers while taking into account these relationships between datasets, and iTOP, a method to infer how different datasets are related to each other.

For difficult cases where the number of cell lines is very small, we have developed a method that predicts drug response simultaneously for all drugs in the screen, thereby gaining statistical power. We based this method on a machine learning methodology called multi-task learning. In contrast to other multi-task learning methods, our approach provides insight into which features are important for a given treatment, thereby allowing us to identify biomarkers from these models.

Finally, we analyzed a screen of 54 drug combinations across 765 cell lines. We report which combinations show synergy (i.e. where the effect of the combination was larger than one would expect based on the individual drug effects) most frequently, hence making them broadly applicable. In addition, for each drug combination, we statistically associated molecular features (i.e. mutations, copy number aberrations, gene expression and proteomics) with the synergy, from which the strongest associations may be good candidate biomarkers.

SAMENVATTING

Om kankerbehandeling te verbeteren moeten we beter begrijpen waarom een gegeven medicijn sommige patiënten helpt, maar anderen niet. Hiertoe zijn de afgelopen jaren verscheidene grootschalige experimenten uitgevoerd, waarbij honderden tumorcellijnen zijn gekarakteriseerd voor vele moleculaire kenmerken (bijv. mutaties, CNA's, methylatie en genexpressie), evenals voor respons op honderden kankermedicijnen. Door deze moleculaire kenmerken statistisch te associëren met de medicijnrespons kunnen we biomarkers identificeren: markers die (na grondig testen) kunnen worden gebruikt om te helpen identificeren welke behandeling aan welke patiënt moet worden gegeven.

Bij het uitvoeren van dergelijke statistische analyses ontdekten we dat er sterke relaties zijn tussen de verschillende moleculaire datasets (bijv. mutaties, CNA's, methylatie en genexpressie) en dat deze relaties ons vermogen om biomarkers te identificeren negatief kunnen beïnvloeden. Naar aanleiding van deze resultaten hebben we TANDEM ontwikkeld, een methode die bij het identificeren van biomarkers corrigeert voor de relaties tussen datasets, en iTOP, een methode om te bepalen hoe verschillende datasets gerelateerd zijn aan elkaar.

Voor moeilijke gevallen waarbij het aantal cellijnen erg klein is hebben we een methode ontwikkeld die de statistische kracht vergroot door tegelijkertijd de medicijnrespons voor alle geteste medicijnen te voorspellen. We hebben deze methode gebaseerd op een machinaal leren methodologie die multi-task learning wordt genoemd. In tegenstelling tot andere multi-task learning methoden geeft onze aanpak inzicht in welke moleculaire kenmerken belangrijk zijn voor een bepaalde behandeling, waardoor we deze methode kunnen gebruiken om biomarkers te identificeren.

Ten slotte hebben we een experiment van 54 medicijncombinaties in 765 cellijnen geanalyseerd. We rapporteren welke combinaties het vaakst synergie vertonen (d.w.z. waar de respons op de combinatie groter was dan verwacht op basis van de individuele medicijnen) en dus potentieel breed toepasbaar zijn. Daarnaast hebben we voor iedere medicijncombinatie moleculaire kenmerken (mutaties, CNA's, genexpressie en proteomica) statistisch geassocieerd met de synergie, waarvan de sterkste associaties goede kandidaat-biomarkers kunnen zijn.

1

Introduction

Nanne Aben, Lodewyk F.A. Wessels

1

1.1. CANCER AND ITS TREATMENT

Cancer is a disease that arises due to genetic aberrations, such as mutations. Certain genetic aberrations allow a cell to start proliferating (i.e. dividing uncontrollably) and the resulting mass of cells then forms a tumor. Anti-cancer drugs aim to specifically kill these cells, either by targeting rapidly dividing cells, as is done in chemotherapy; or by specifically targeting the genetic processes that these cells depend on.

Each of these anti-cancer drugs only works in subsets of patients. Unfortunately, it is often not known which patients will benefit most from a given drug. Therefore, in order to select the best treatment for each patient, we aim to identify biomarkers of drug response: genetic features (e.g. mutations) that can predict whether a patient will respond to a given drug.

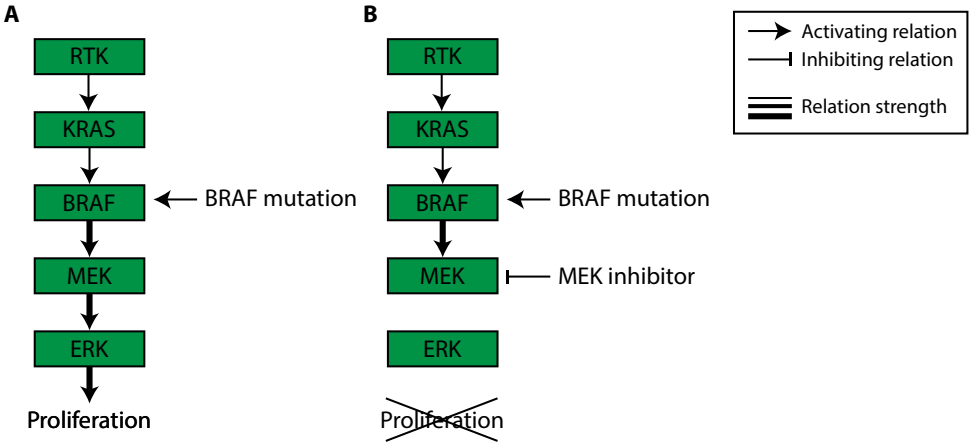


Figure 1.1: Illustration of a biomarker in a well-studied signaling pathway. A mutation in BRAF leading to increased signalling, which eventually results in cell proliferation. B The increased signalling from BRAF is stopped using a MEK inhibitor, which in turn stops the proliferation.

An example of a biomarker is the BRAF mutation, which usually indicates sensitivity to a class of drugs called MEK inhibitors. BRAF mutations activate a signaling pathway through MEK (Figure 1.1A), which ultimately results in cell proliferation. By using a drug to inhibit MEK, the signalling coming from BRAF is blocked and the proliferation typically gets halted (Figure 1.1B). Therefore, patients whose tumor harbors a BRAF mutation would be eligible for treatment with a MEK inhibitor.

While the BRAF mutation - MEK inhibitor provides a clear example of a biomarker, we do not always have such a good understanding the relevant signaling pathway. This means that for many drugs no (good) biomarkers are known, and hence there is a great need for new biomarkers.

1.2. CELL LINES AS A MODEL SYSTEM FOR ANTI-CANCER DRUG RESPONSE

One approach to identifying new biomarkers of anti-cancer drug response is to use cell lines as model system (Figure 1.2). These cell lines are tumor cells that have been derived from (patient) tumors and have subsequently been cultured in a dish in the lab. They make great model systems for studying cancer: even though cell lines may not perfectly capture all of the biology of a patient tumor (e.g. since they are not in a patient, they do not capture interactions with immune cells or surrounding non-tumor tissue), they are very useful for experiments that would be impossible in patients. For example, since cell lines will keep on growing in the lab, we can always produce more of a certain cell line, which allows us to test many different treatments on the same tumor material. While such studies can also be done in other model systems (e.g. in mice), the main advantage of cell lines is that they can be used on a very large scale (which would otherwise be unethical and/or impractical).

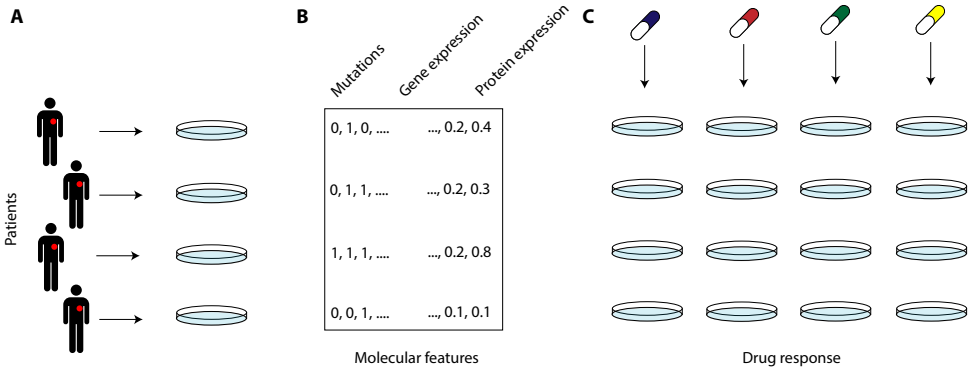


Figure 1.2: Simplified example of a drug response screen. (A) Derivation of cell lines (tumor cell that can be cultivated in a dish in the lab) from patient tumors. (B) These cell lines are characterized for molecular features, such as mutations, gene expression or protein expression. (C) In addition, these cell lines are profiled for their sensitivity to several anti-cancer drugs.

In recent years, several groups have performed large-scale drug response screens using cell lines [1, 3, 4, 6]. The cell lines in these studies have been profiled for molecular features (e.g. mutations, gene expression, protein expression), as well as response to many anti-cancer drugs. By identifying statistical associations between the molecular features and the drug response, new biomarkers of drug response can be identified.

1.3. DRUG RESPONSE PREDICTION

Drug response prediction, a subfield of computational biology, aims to identify biomarkers of response by analyzing the drug response screens using machine learning algorithms. These machine learning algorithms identify putative biomarkers by modeling the relationship between the molecular features and the drug response as follows:

$$y = f(\mathbf{X})$$

1

Where y represents the drug response (e.g. the concentration at which the cell line show response to the drug), $\mathbf{X}$ represents the molecular features (e.g. mutations, gene expression and proteins expression in the cell line), and $f()$ represents a function that can predict the drug response y from the molecular data $\mathbf{X}$.

We can use machine learning models to find this function $f()$. The challenge in identifying this function is that it should not only fit the available data, but it should also generalize to unseen data: given molecular data of a new, unseen cell line, it should be able to accurately predict the drug response for this cell line. We note that this is a hard problem and that the function $f()$ is never perfectly identified.

Since the function $f()$ is trained on cell lines rather than patient tumors, we cannot directly apply this model to patients to decide which therapy they should receive. Instead, we aim to use these models to understand why some cell lines respond to the drug, while others do not. However, using these models to improve our understanding of why some cell lines respond is by no means a simple task. The main challenge here lies in the number of molecular features: we often have tens of thousands of these, and most machine learning models will use all of these to explain the drug response. This results in models that are so large that they do not really give us a clue about which processes in the tumor cells are associated with drug response.

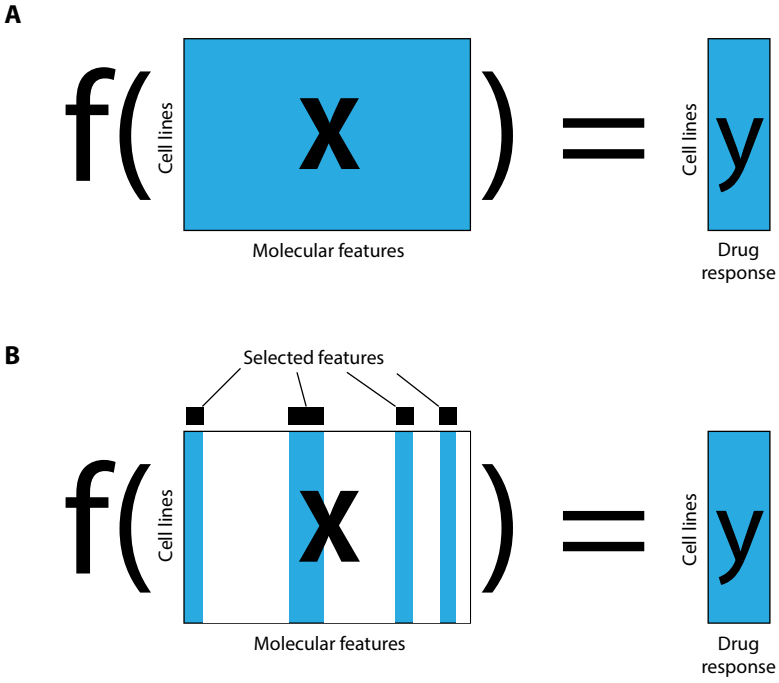


Figure 1.3: Graphical representation of drug response prediction. (A) The goal in drug response prediction is to find a function f() that can map the molecular features of each cell line to their drug response. (B) In feature selection we optimize both the function f() and the subset of features that is used in the model.

To reduce the size of these models, a machine learning technique called feature se-

lection is typically used to identify the subset molecular features that is most predictive of the drug response (Figure 1.3). This subset can then be further investigated to gain insight into why some cell lines respond to the drug, and to assess whether any of the associations in the subset could be used as a biomarker. In drug response prediction, the most commonly used feature selection algorithm is Elastic Net regression [1–5].

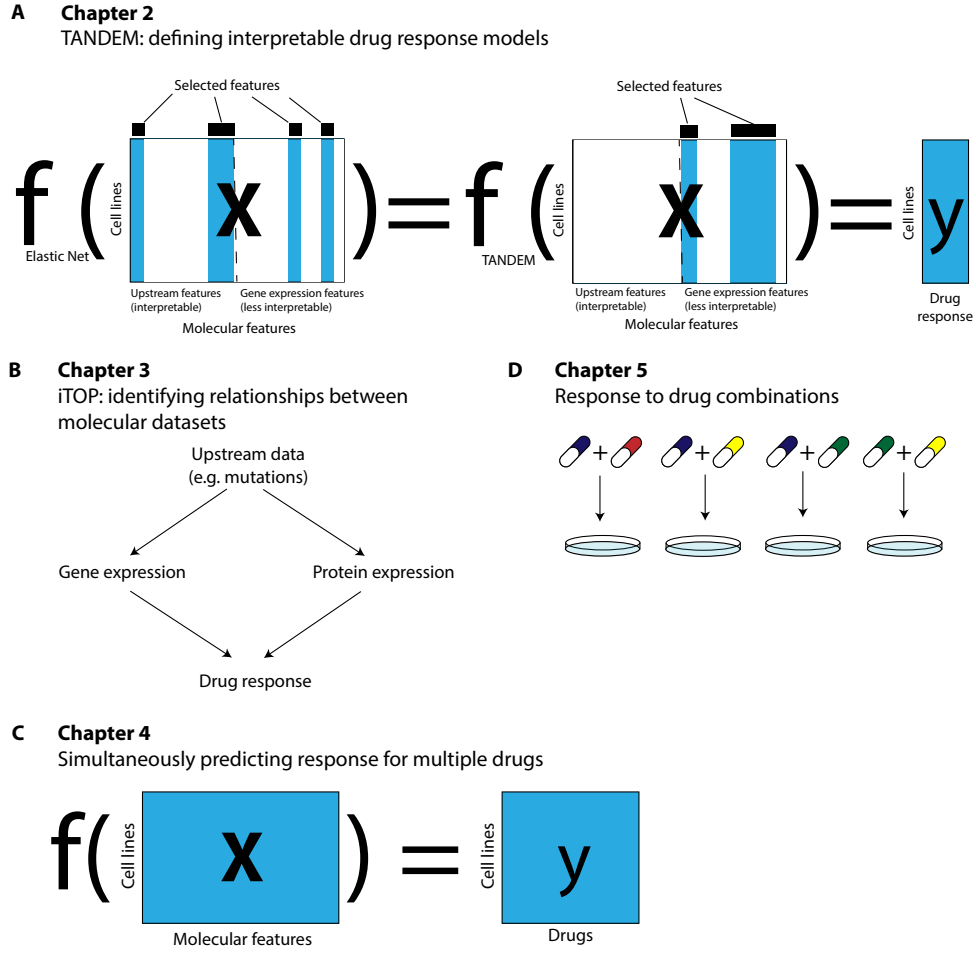


Figure 1.4: Illustration of the different chapters in this thesis.

1.4. THESIS OUTLINE

In this thesis, we focus on the prediction of drug (combination) response through data integration. We split this in three parts:

1. Data integration (Chapter 2 & 3)
2. Prediction (Chapter 2 & 4)

3. Drug combination response (Chapter 5)

1.4.1. CHAPTER 2

The data from drug response screens typically consist of multiple parts (to which we will refer as datasets), which each characterize different molecular features (e.g. mutations or gene expression). We typically want to integrate all these molecular datasets in our model, such that we get the most complete description of the drug response. But to our surprise, when we applied Elastic Net regression to all these datasets combined, we found that the resulting models that were almost completely based on one dataset: gene expression.

This result surprised us, because drug response is typically interpreted in the context of the other datasets (e.g. mutations), for instance in the BRAF mutation - MEK inhibitor example from Figure 1.1. Hence, we prefer to have models that make use of all datasets (and not only gene expression), as this closer to our current understanding of cancer biology and hence enhances the interpretability of these models.

To better understand the above result, we had a look at the overlap in information between datasets. We found that the information that was shared between any of the datasets and the drug response was fully contained in gene expression. This explains why an Elastic Net regression model can be based on only gene expression: in terms of predictive performance there is no added value (no extra information) in the other datasets. However, as argued above, in terms of interpretability of the model, there is added value in using the other datasets.

To this end, we developed TANDEM: a method that preferentially uses the upstream datasets (i.e. datasets other than gene expression) to predict drug response. We show that in TANDEM upstream data contributes to a larger extent to the response prediction, thereby enhancing the interpretability of these models, while maintaining the same predictive performance as Elastic Net regression.

1.4.2. CHAPTER 3

Our work in Chapter 2 shows how the different datasets in drug response screens have complex relationships between them, and how these relationships can affect statistical analyses performed on them (e.g. Elastic Net regression models being almost completely based on gene expression). In addition, Chapter 2 showed that if we have a way of identifying these relationships between datasets, we can adapt our methods (e.g. as we have done in TANDEM).

However, the approach we used in Chapter 2 to identify the relationships between the datasets becomes increasingly more difficult as the number of datasets increases. Therefore, in the third chapter of this thesis, we propose iTOP, a much faster and much simpler approach to identify these relationships.

In iTOP we combined matrix correlations (which quantify the overlap in information between two datasets, such as mutation data and gene expression data) with partial correlations (which quantify the correlation between x and y corrected for z, or $cor(x, y \mid z)$) into partial matrix correlations. These allow us, for example, to quantify how much information is shared between mutation data and drug response data, but is not present in gene expression data. In other words: cor(mutation data, drug response | gene expres-

sion). We then use the PC algorithm, a method typically used to reconstruct networks based on partial correlations, to infer the topology between datasets, such as the one depicted in Figure 1.4C.

1.4.3. CHAPTER 4

When the number of screened cell lines is extremely small, it becomes hard to accurately predict drug response, i.e. to accurately identify the underlying function $f()$. However, when the number of screened drugs is big, this problem can be addressed using multi-task learning [7, 8], a machine learning approach that leverages its statistical power by simultaneously learning over multiple tasks (or in this case: over multiple drugs).

Multi-task learning can most easily be explained using an analogy. Suppose we want to learn Spanish and suppose we would need thousands of examples to properly learn Spanish. However, we might only have hundreds of examples available. But if we also have hundreds examples from Italian, Portuguese and French, we might be able to leverage those examples to better learn Spanish as we can exploit some of the commonalities in vocabulary and grammar between the languages. Similarly, with drug response prediction, we borrow information from other (similar) drugs to predict the response to a specific drug of interest.

A drawback of applying multi-task learning to drug response prediction is that previously available methods do not provide insight into which features are important for a given drug and hence are not suitable for biomarker identification. To overcome this limitation, we developed the DVI, a modified version of the Random Forest variable importance score, in which we can compute the importance for each variable and drug separately, even when the underlying Random Forest is jointly trained on all drugs.

1.4.4. CHAPTER 5

Many tumors are dependent on a signalling pathway, such as the one from Figure 1.1, and hence using drugs to inhibit that pathway is a good therapeutic strategy. Unfortunately, there are many examples where using one drug by itself is not effective. For example, the tumor may activate a second pathway as a backup mechanism when the first pathway is inhibited. In these cases the addition of a second drug that inhibits the backup pathway can increase the treatment efficacy.

Drug combinations like the one described above have been extensively researched in recent years. Many of them have been described as showing synergy: an effect on cell kill greater than you would expect based on the sum of the effects of the individual drugs. However, the efficacy of these combinations has usually only been shown in a small subset of tumor cell lines, hence it is not clear to what extent these synergies generalize to a larger set of tumors. In addition, many combinations lack a biomarker that allows us to identify patients that are likely to respond to the combination, and those that should preferably receive a different treatment.

To help address these problems, we performed a screen of 54 drug combinations across 765 cell lines. We report which combinations show synergy most frequently, hence generalizing to a large set of tumors. In addition, we statistically associated molecular data (i.e. mutations, copy number aberrations, gene expression and proteomics) with synergy to a given drug combination, to identify candidate biomarkers.

1

REFERENCES

[1] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603, 2012.

[2] James C Costello, Laura M Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P Menden, Nicholas J Wang, Mukesh Bansal, Petteri Hintsanen, Suleiman A Khan, John-Patrick Mpindi, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*, 32(12):1202, 2014.

[3] Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, Chris D Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I Richard Thompson, Xi Luo, Jorge Soares, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570, 2012.

[4] Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, 2016.

[5] In Sock Jang, Elias Chaibub Neto, Justin Guinney, Stephen H Friend, and Adam A Margolin. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. In *Biocomputing 2014*, pages 63–74. World Scientific, 2014.

[6] Brinton Seashore-Ludlow, Matthew G Rees, Jaime H Cheah, Murat Cokol, Edmund V Price, Matthew E Coletti, Victor Jones, Nicole E Bodycombe, Christian K Soule, Joshua Gould, et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer discovery*, 2015.

[7] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[8] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM, 2004.

2

TANDEM: A TWO-STAGE APPROACH TO MAXIMIZE INTERPRETABILITY OF DRUG RESPONSE MODELS BASED ON MULTIPLE MOLECULAR DATA TYPES

NANNE ABEN, DANIEL J. VIS, MAGALI MICHAUT, LODEWYK F.A. WESSELS

Parts of this chapter have been published in Bioinformatics 17:i413–i420 (2016) [1].

ABSTRACT

Motivation: *As patient response to anti-cancer treatment is highly variable, predictive biomarkers are required to identify which patients are likely to benefit. To aid biomarker identification, the DREAM challenge consortium recently released data from a screen containing 85 cell lines and 167 drug combinations. The main challenge of these data is the low sample size: a median of 14 cell lines have been screened per drug combination.*

Results: *To utilize all data types in a more balanced way, we developed TANDEM, a two-stage approach in which the first stage explains response using upstream features (mutations, copy number, methylation and cancer type) and the second stage explains the remainder using downstream features (gene expression). Applying TANDEM to 934 cell lines profiled across 265 drugs (GDSC1000), we show that the resulting models are more interpretable, while retaining the same predictive performance as the classic approach. Using the more balanced contributions per data type as determined with TANDEM, we find that response to MAPK pathway inhibitors is largely predicted by mutation data, while predicting response to DNA damaging agents requires gene expression data, in particular SLFN11 expression.*

Availability: *TANDEM is available as an R package on CRAN (for more information, see https://cran.r-project.org/package=TANDEM).*

Contact: *m.michaut@nki.nl and l.wessels@nki.nl*

2.1. INTRODUCTION

Large-scale pharmacogenomics screens provide a wealth of information about potential mechanisms of drug response. In these screens, cell lines of different cancer types have been profiled molecularly (mutations, copy number alterations, DNA methylation and gene expression), as well as pharmacologically (response to anti-cancer drugs) [3, 15]. Using drug response prediction models, statistical associations can be identified between the drug response and the molecular data. For example, the presence of a BRAF mutation predicts sensitivity to Vemurafenib in melanoma cell lines and a mutation in TP53 predicts resistance to Nutlin-3a [11]. By combining various data types in an integrative analysis, all molecular data can be employed to explain drug response. This is commonly achieved by performing Elastic Net regression [28] on all molecular data types simultaneously [3, 5, 11, 15, 16]. Throughout this work, we will refer to this approach as the 'classic approach' (Fig. 2.1). While this approach could, in theory, use information from all molecular data types, we find that it typically leads to models that are mostly based on gene expression data. For instance, a BRAF mutation activates, via a cascade of signaling events, the transcription of many genes. As a result, the expression of these genes is tightly linked to the mutation status of the BRAF gene, and thus also predictive of response to Vemurafenib. When all molecular data are combined to build a predictive model for response to Vemurafenib, expression of these genes may be selected instead of the BRAF mutation, which would make the resulting model more difficult to interpret. Instead, selecting the BRAF mutation as a feature in the model would be more informative about the mechanism of the drug and thus lead to a more interpretable model.

We propose TANDEM, an approach that employs a two-stage analysis to improve

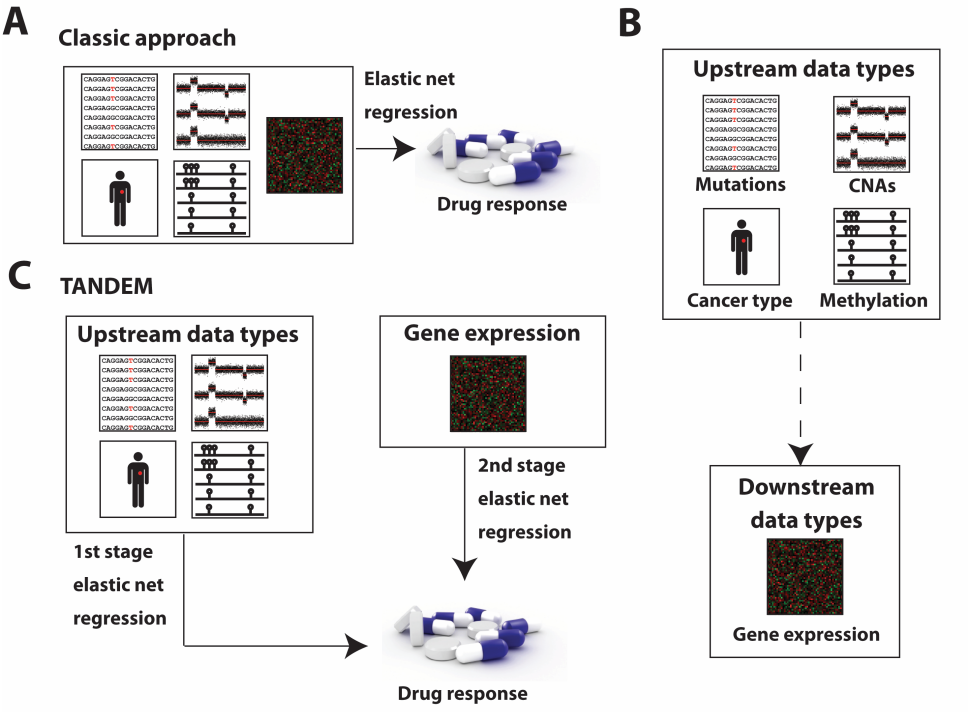


Figure 2.1: Illustration of TANDEM and the classic approach. (A) The classic approach: an Elastic Net regression trained on all data types simultaneously. (B) The information predictive of drug response contained in the upstream data types is also present in the gene expression data. (C) TANDEM: our two-stage approach, which first uses the upstream data types to explain as much of the drug response as possible, and then uses the gene expression to explain the remainder.

the interpretability of prediction models by preferentially using the data types upstream of gene expression. To this end, we first split the molecular data types into 'upstream data' (somatic mutation, copy number alteration (CNA), methylation and cancer type) and 'downstream data' (gene expression) (Fig. 2.1B). This separation is based on the idea that mutation status, for example, affects the transcription of genes downstream of the pathway in which the mutation resides. TANDEM analyzes the upstream and downstream data 'in tandem': it first explains as much of the drug response as possible using the upstream (more interpretable) data and then explains the remainder using gene expression data (Fig. 2.1C). Applying TANDEM to a panel of 934 cell lines profiled across 265 drugs [15], we find that the upstream data types contribute more to the prediction than in the classic approach. At the same time, TANDEM retains the same predictive performance as the classic approach. The features selected by TANDEM result in twice as many significant pathway enrichments compared to the classic approach, implying that the selected features are more informative about the mechanisms of drug response. Additionally, using the more balanced contributions of the various data types, we find that response to MAPK targeting drugs is mostly explained by mutation data, while pre-

dicting response to DNA damaging agents requires gene expression data.

2.2. METHODS

2.2.1. DATA SET

The Genomic Determinants of Sensitivity in Cancer 1000 (GDSC1000) data comprises a panel of cell lines screened for 265 anti-cancer drugs [15]. This panel contains 926 cell lines that are fully characterized for point mutations, copy number alterations (CNAs), methylation status and gene expression profiles. Based on human tumor data from The Cancer Gene Atlas (TCGA) [26], Iorio et al. [15] have performed feature selection resulting in a set of 305 mutation, 409 CNA and 312 methylation features, all of which are binary. Additionally, we considered 29 binary features indicating the cancer type and 17,737 continuous gene expression features. The drug response was summarized by the IC50 (concentration that inhibits 50% of the target).

2.2.2. DRUG RESPONSE PREDICTION USING THE CLASSIC APPROACH

For drug response prediction models based on the classic approach, we used linear Elastic Net regression [28] implemented in the R package glmnet [8]. The hyper-parameter λ was optimized using 10-fold cross-validation and α was set to 0.5. Predictive performance estimates were made using double-loop cross-validation.

2.2.3. PREDICTING THE BINARY VALUE OF UPSTREAM FEATURES FROM GENE EXPRESSION

e first identified upstream features that are associated with drug response using a Mann-Whitney U test, and only selecting features significantly associated with response to at least one drug (Benjamini-Hochberg corrected $p < 0.05$). For each of the identified upstream features, we then predicted its binary value using logistic regression of the gene expression data. Again, we used the implementation from the R package glmnet (Friedman and Hastie, 2009), optimized λ using 10-fold cross-validation and set α to 0.5. The classification performance (Area Under the ROC, AUROC) was determined using double loop cross-validation. Because the classes are often highly unbalanced (i.e., a mutation typically only occurs in tens of samples out of 926), we used stratified cross-validation for the outer loop. This way, we ensured that each outer loop contains at least one sample per class. For the same reason, we omitted all upstream features that appear in fewer than ten samples in total.

2.2.4. RELATIVE CONTRIBUTION OF EACH DATA TYPE TO THE PREDICTION

In order to determine the relative contribution of each data source, we created a prediction per data source. We determined the relative contribution RC_i for each data source by dividing the sum-of-squares of a prediction from a certain data type by the sum-of-squares of the overall prediction (see Supplementary Materials and Methods). We only took into account drugs for which we achieved a predictive performance $r > 0.4$. This prevents models with poor predictive performance from confounding the analysis.

2.2.5. THE TANDEM ALGORITHM

We used a two-stage approach to predict drug response: i) Fit an Elastic Net model to predict the drug response using the upstream data types; ii) Fit an Elastic Net model to predict the residuals from the first stage using the gene expression data. Like in the classic approach, λ was optimized using cross-validation and α was set to 0.5. We used the same separation in cross-validation folds for both stages. Similar to the classic approach, we used a double-loop cross-validation to estimate performance.

2.2.6. FEATURE IMPORTANCE SCORE

The feature importance FI for feature j was determined as follows:

$$FI = \frac{\left\| X_j \beta_j \right\|_2^2}{\left\| \hat{y} \right\|_2^2}$$

Where X_j is the column j of $\mathbf{X}$. Without loss of generality, we assume that all columns of $\mathbf{X}$ and the prediction $\hat{y}$ are mean-centered.

2.2.7. PATHWAY ENRICHMENT

We downloaded version 5 of the KEGG pathways from MSigDB [23] and used a hypergeometric test to quantify the enrichment of selected features within a pathway. The p-values were controlled for FDR by applying Benjamini-Hochberg correction per drug. For more details, see Supplementary Materials and Methods.

2.3. RESULTS

2.3.1. THE INFORMATION IN ALL DATA TYPES IS CAPTURED IN THE GENE EXPRESSION DATA

For each of the 265 drugs of the GDSC1000 pharmacogenomics panel, we first built drug response models for each drug and each data type separately using Elastic Net regression. We assessed the predictive performance of these models using the Pearson correlation coefficient between the observed and the predicted IC50s (Supplementary Table S2.1). The most predictive data type was found to be gene expression data: the median predictive performance of these models is higher compared to models based on other data types (Fig. 2.2A). This finding is consistent with previous work by Costello et al. [5] and Jang et al. [16]. Subsequently, we built drug response models using all data types simultaneously, referred to as the 'classic approach'. We found that the predictive performance of models based on only gene expression and models based on the classic approach was nearly identical (median difference across drugs: 0.001, Fig. 2.2A). The predictive performance of these two methods is not only comparable at the median, but it is also highly correlated across all drugs in the panel (Pearson correlation coefficient across drugs: 0.99, Supplementary Fig. S2.1A), indicating that both methods achieve similar performance for the same drugs. Altogether, we found that adding upstream data does not improve a model based on gene expression only, implying that the information from the upstream data types is already contained in the gene expression data.

To investigate the possible redundancy between the upstream and the downstream data, we attempted to predict the upstream features (e.g. aberration status or cancer

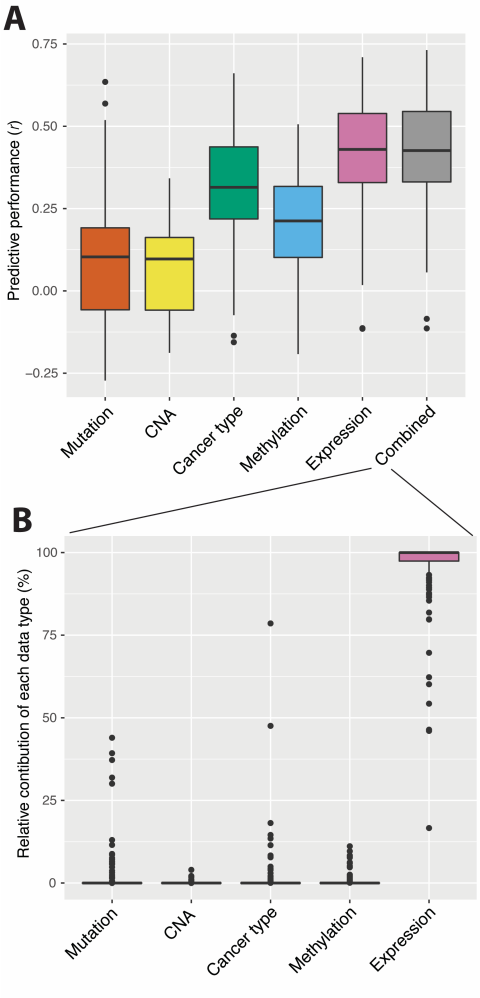


Figure 2.2: Predictive performance of individual molecular data types. (A) Predictive performance (Pearson correlation between measured IC50s and predictions from the classic approach) across 265 drugs using individual data types (mutation, CNA, tissue of origin, methylation, gene expression) or a combination of all data types (combined) with the classic approach. (B) Relative contribution of each data type in the combined models, across all drugs for which we achieved a predictive performance $r > 0.4$.

type) from downstream data (gene expression). For the 503 upstream features associated with drug response, predicting the aberration status or cancer type from gene expression resulted in a median AUROC (Area Under the ROC curve) of 0.88 (Supplementary Fig. S2.1B). Hence, we found that it is indeed possible to predict the upstream features with high accuracy from downstream data, which further corroborates that the information in the upstream features is also present in the gene expression data. Finally, we investigated the relative contribution of each data type to models based on the classic

approach. To assess the relative contribution of a given data type, we determined what fraction of the prediction using all data types is explained by that particular data type (Methods) (Supplementary Table S2.1). Despite the redundancy between the upstream and the downstream data, the models preferentially select gene expression features (Fig. 2.2B). For 89% of the drugs, more than 90% of the variation in the prediction was attributed to gene expression. To investigate whether the high dimensionality and the continuous nature of the gene expression data had an effect on this result, we reduced the number of features and discretized the gene expression (Supplementary Methods). In both cases, we still observed the domination of the gene expression in the models (Supplementary Fig. S2.1C). We concluded that neither the dimensionality nor the continuous nature of the data explain the high relative contribution of gene expression in the models based on the classic approach.

Altogether, we have shown that, in the context of drug response prediction, gene expression recapitulates the information contained in upstream data. Thus, we set out to exploit the redundancy between the upstream and downstream data to create more interpretable models.

2.3.2. TANDEM PRODUCES A MORE BALANCED CONTRIBUTION OF DIFFERENT DATA TYPES WHILE MAINTAINING THE SAME PERFORMANCE

To utilize the information from gene expression data, without allowing it to completely dominate the models, we propose a two-stage approach to predict drug sensitivity. In the first stage, TANDEM constructs a model to predict as much of the variation in the drug response as possible using the – more interpretable – upstream data types only. In the second stage, TANDEM explains the remainder of the variation in the drug response using gene expression data. We illustrate the results of our method and its differences with the classic approach using three well-characterized drugs: Trametinib (a MEK inhibitor), Nutlin-3a (an MDM2 inhibitor) and Nilotinib (a BCR-ABL inhibitor). Using the classic approach, gene expression accounts for most of the prediction (Fig. 2.3A). For Trametinib 94% of the prediction is attributed to gene expression data and only 6% is attributed to the upstream data types. In contrast, using TANDEM, we obtain a model where 32% of the prediction is attributed to gene expression and 68% to the upstream data types (Fig. 2.3B). The same holds for Nutlin-3a and Nilotinib: when employing TANDEM, the contribution of upstream data types increases dramatically, albeit in different proportions, while maintaining the same level of predictive performance (Fig. 2.3A, 3B).

Across all drugs for which we obtained a predictive performance $r > 0.4$, the median percentage of variation attributed to gene expression was 100% when using the classic approach, while it dropped to 52% when using TANDEM (Fig. 2.3C). In the latter case, the median percentage of variation explained by mutations, CNAs, methylation status and cancer type was 3%, 2%, 20% and 11%, respectively (Fig. 2.3C). In addition, TANDEM obtains virtually the same predictive performance as the classic approach (Fig. 2.3D) (Pearson correlation: 0.99, median difference: 0.002). In summary, TANDEM results in models that use all data types in a more balanced fashion, while retaining the same predictive performance as the classic approach.

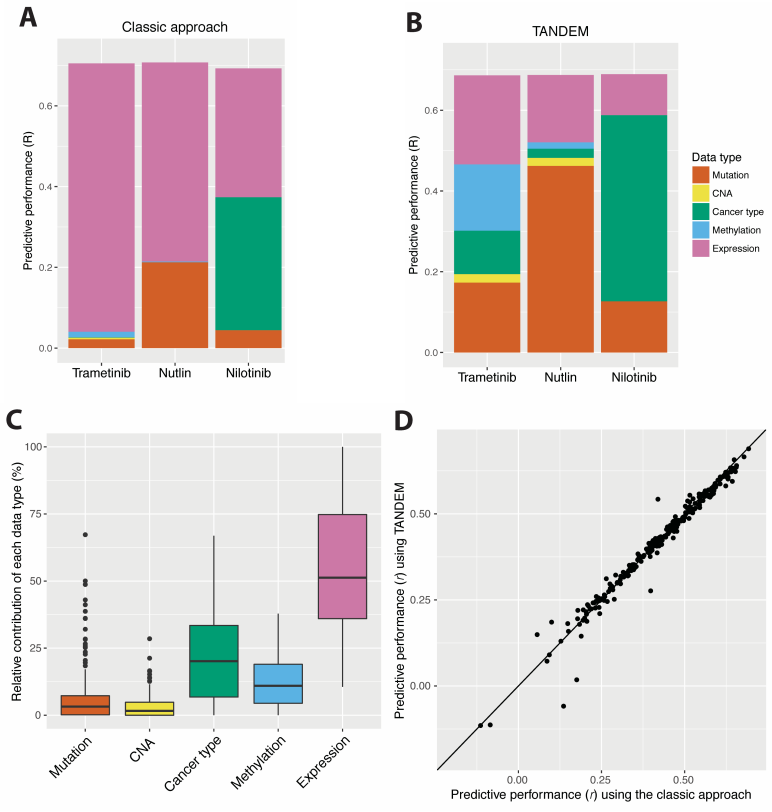


Figure 2.3: Data type contribution and predictive performance. Relative contribution of each data type (indicated by the colors) and predictive performance (r, the Pearson correlation between observed and predicted IC50s) for three example drugs, using (A) the classic approach for data integration and (B) TANDEM. (C) Relative contribution of each data type in TANDEM, across 265 drugs, across all drugs for which we achieve a predictive performance $r > 0.4$. (D) Predictive performance of the classical approach vs. TANDEM.

2.3.3. TANDEM PRODUCES MORE INTERPRETABLE MODELS

TANDEM produces models that are mostly based on upstream data features. As these upstream features are more likely causally related to drug response, the resulting models are easier to interpret. To demonstrate the improved interpretability, we performed a pathway enrichment analysis of the genes identified by TANDEM as being associated with drug response. Using the KEGG pathways [18, 19], we tested all drug-pathway pairs for enrichment of predictive genes (i.e. genes associated with response to the drug in our model) amongst the genes annotated to this pathway. Since TANDEM preferentially uses the upstream data, which is enriched for well-studied genes, we were concerned with selection bias when testing for pathway enrichment against a genome-wide background distribution. To account for this bias, we instead defined the background distribution using only genes present in at least one KEGG pathway (Methods). After correcting for

multiple testing, TANDEM yielded more than twice (164 versus 64) the number of significant enrichments as compared to the classic approach (Supplementary Fig. S2.2A & S2.2B). The features selected by TANDEM can thus be related to existing knowledge (pathways) more easily than those selected by the classic approach, implying that the resulting models are more easily interpreted.

We illustrate these results using two significant enrichments from TANDEM: the features in the MAPK pathway associated with response to the MEK inhibitor Trametinib (Benjamini-Hochberg FDR corrected p: 1.0e-3, Fig. 2.4A) and the features in the B cell receptor signaling pathway associated with the HDAC6 inhibitor Tubastatin (Benjamini-Hochberg FDR corrected p: 5.3e-5, Fig. 2.4B). In both examples, the features selected by TANDEM resulted in a significant enrichment, whereas the features selected by the classic approach did not.

For Trametinib (a MEK inhibitor), both methods identified $KRAS$, $NRAS$ and $BRAF$ mutations to be associated with sensitivity (Supplementary Fig. S2.3A, S2.3B & S2.3C). This is expected as these mutations all activate MAPK signaling through MEK, and inhibition of MEK shuts down the pathway, thereby mitigating their effect and rendering mutated cell lines sensitive to Trametinib. TANDEM selected two additional mutations in the pathway: $HRAS$ and MYC (Supplementary Fig. S2.3D & S2.3E). Like the aforementioned mutations, $HRAS$ signals through MEK and hence $HRAS$ mutations are associated with sensitivity. Myc proteins can harbor a mutation in their regulatory phosphorylation site, which allows them to escape ubiquitin/proteasome-mediated turnover and leads to accumulation of Myc protein [2]. Because the mutated Myc proteins activate the downstream targets of the pathway independently of MEK, mutated cell lines are insensitive to the MEK inhibitor. Thus, this mutation is associated with resistance to MEK inhibition. In addition, both methods identified DUSP6 as a predictive feature (Supplementary Fig. S2.3F). $DUSP6$ transcription is induced by ERK activation [9]. Hence, by proxy, high DUSP6 expression is an indication of high phospho-ERK levels. Since phospho-ERK can be attenuated by MEK inhibition, high $DUSP6$ expression is associated with sensitivity to MEK inhibition [17]. $DUSP6$ is an example of a gene expression feature whose selection not only increases the predictive performance, but also benefits the interpretability.

Our second example models the response to the HDAC6 inhibitor Tubastatin (Fig. 2.4B), an anti-inflammatory drug [4, 24] that has shown anti-cancer potential [20, 22]. Unlike other members of the HDAC family, HDAC6 is exclusively localized in the cytoplasm and hence does not have a histone deacetylase function [10, 13]. Instead, Gao, et al. [10] have proposed that HDAC6 is required for efficient Rac1 activation. Interestingly, TANDEM identifies $RAC1$ amplifications to be associated with resistance to Tubastatin (Supplementary Fig. S2.3G), whereas the classic approach does not. This could mean that when Rac1 is available in abundant levels, efficient activation of Rac1 by HDAC6 is not required anymore and hence HDAC6 inhibition has little effect, causing resistance. Both methods associated $PKC\beta$ expression with sensitivity to Tubastatin (Supplementary Fig. S2.3H). One additional gene expression feature was uniquely identified using TANDEM: the expression of $Ig\beta$ (Supplementary Fig. S2.3I). As $PKC\beta$ and $Ig\beta$ both reside in the B cell receptor signaling pathway, their selection could mean that Tubastatin is especially potent in B cell derived lymphoid cancers with active B cell receptor

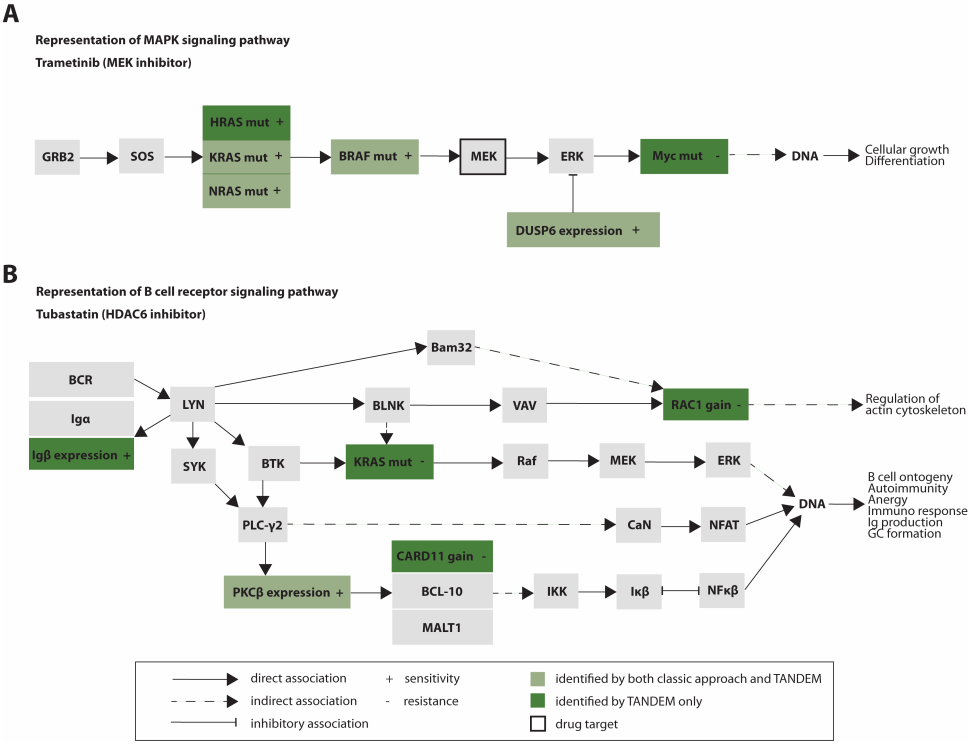


Figure 2.4: Features selected by TANDEM in the context of two pathways. Representation of (A) the MAPK signaling pathway and (B) the B cell receptor signaling pathway from KEGG. Indicated in color are the genes associated with response to (A) Trametinib or (B) Tubastatin by TANDEM (dark green) or by both approaches (light green).

signaling. This is further supported by a negative correlation between the expression of $Ig\beta$ (a component of the B cell receptor) and response to Tubastatin within the 68 B cell derived lymphoid cell lines in the GDSC1000 data set (Pearson correlation coefficient: -0.49, Supplementary Fig. S2.3J).

Altogether, we found that the features identified by TANDEM can be interpreted in the context of pathways. Due to the more balanced contributions of upstream and downstream data types, we show that our method leads to improved interpretability of the drug response models, while achieving the same predictive performance.

2.3.4. DIFFERENT DATA TYPES PREDICT RESPONSE TO DIFFERENT DRUG CLASSES

To test if certain data types better predicted response to certain classes of drugs, we used the drug classification provided with the GSDC1000 data [15], where all 265 drugs are categorized into 21 classes, based on either the mechanism of action (e.g. DNA damaging agents) or the pathway in which the drug target resides (e.g. MAPK pathway). For a given drug class, we considered the relative contribution each data type makes to the

prediction using TANDEM, using only the drugs for which a model could be built with predictive performance r > 0.4. Using these relative contributions, we tested each drug class for association with each data type (Supplementary Fig. S2.4). We further investigated two associations: the most significant association using upstream data (MAPK pathway inhibitors and mutation data) and the most significant association using downstream data (DNA damaging agents (DDAs) and gene expression). For these drug classes, we determined the top 10 most important features using both the classic approach and TANDEM. The feature importance was assessed based on the size of the regression coefficient, corrected for the variance of the corresponding feature (Supplementary Table S2.1).

2.3.5. GENE EXPRESSION DATA IS THE BEST PREDICTOR OF RESPONSE TO DNA DAMAGING AGENTS

For the 10 drugs from the DDA drug class, the response models produced by TANDEM had a higher contribution of gene expression compared to other drug classes (Benjamini-Hochberg corrected p: 0.046, one-tailed Mann Whitney test, Supplementary Fig. S2.5A). Given that our method preferentially uses upstream features, we found it intriguing that gene expression still accounts for a median 76% of the explained variation. In fact, the contribution of gene expression is mostly due to the expression of *SLFN*11, which is the most important predictor of response to DDAs in both the classic approach and TANDEM (Fig. 2.5A & 2.5B). Part of the information contained in the expression of *SLFN*11 is also present in some upstream features, which results in a lower feature importance for *SLFN*11 when using TANDEM. For example, *SLFN*11 expression is significantly higher in the ALL (p-value: $5.2e-9$, Supplementary Fig. S2.5B). However, as TANDEM selects SLFN11 expression after the acute lymphoid leukemia (ALL) cancer type has been selected, we can rule out that *SLFN*11 is merely selected as a proxy for ALL. Altogether, this points to an important role for SLFN11 in DDA response. Indeed, Zoppoli et al. [27] have found that knockdown of *SLFN*11 leads to increased resistance to many DDAs, indicating a causative role for *SLFN*11 expression.

2.3.6. MUTATIONS ARE THE BEST PREDICTORS OF RESPONSE TO MAPK PATHWAY INHIBITORS

For the 16 drugs from the MAPK pathway inhibition class, the response models produced by TANDEM had a significantly higher contribution of mutation data compared to other drug classes (Benjamini-Hochberg corrected p: 1.1e-5, one-tailed Mann Whitney test, Supplementary Fig. S2.5C). Investigating the most important features obtained using both methods (Fig. 2.5C & 2.5D), we found that they both identified the *BRAF* mutation as the strongest predictor of response, as expected (Downward, 2003). The remaining part of the top 10 features is completely different between the two methods: for the classic approach, it solely consists of gene expression features, whereas for TANDEM it consists of upstream features. TANDEM identifies *KRAS* and *NRAS*, two canonical mutations known to modulate response to MAPK pathway inhibitors (Downward, 2003), while the gene expression features identified by the classic approach do not give clear insight into the mechanisms of drug response. Consistent with the literature, TANDEM also associates a number of cancer types with response to MAPK inhibition: melanoma (SKCM), acute myeloid leukemia (LAML) and chronic myeloid leukemia (LCML) are as-

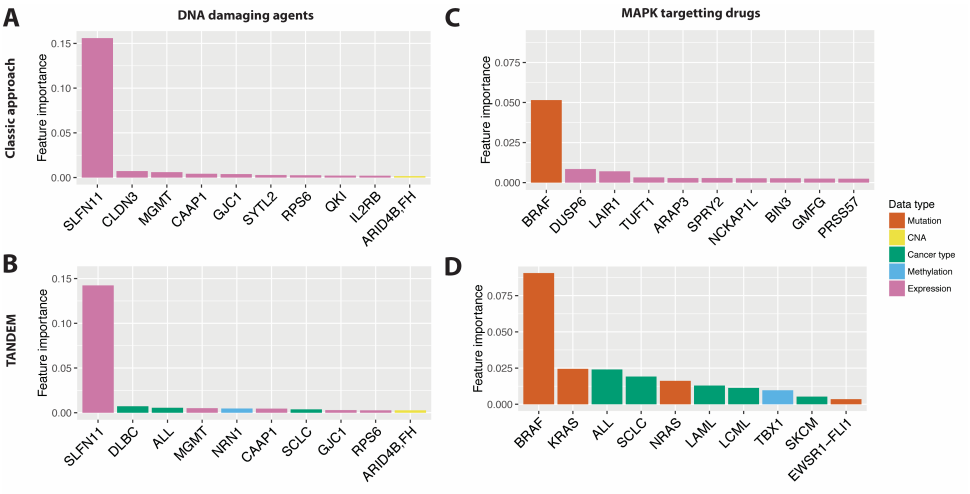


Figure 2.5: Most important features for predicting response to DNA Damaging Agents and MAPK pathway inhibitors. Top ten most important features (based on their average feature importance score) for predicting response to MAPK-targeting drugs (A, B) or DNA damaging agents (C, D) using the classic approach (A, C) or TANDEM (B, D).

sociated with sensitivity [12, 14], whereas small cell lung cancer (SCLC) is associated with resistance [7, 21].

2.3.7. TANDEM PREVENTS CANCER TYPE SPECIFIC EXPRESSION FROM CONFOUNDING THE RESULTS

Using cancer type as an upstream feature, TANDEM avoids the selection of genes whose expression is specific to one cancer type. In the MAPK inhibitors example above, the classic approach selects LAIR1 and PRSS57 as important features (positions 3 and 10 in the top 15 classic approach features). However, these genes are preferentially expressed in LAML and LCML ($p < 2.2e-16$, Mann Whitney U test, Supplementary Fig. S2.6A & S2.6B). Thus, the selection of LAML and LCML cancer types as important features by TANDEM is much more informative. Similarly, the classic approach selects BIN3 expression, but BIN3 is preferentially expressed in SKCM ($p < 2.2e-16$, Mann Whitney U test, Supplementary Fig. S2.6C). The selection of SKCM by TANDEM is therefore more informative.

To further look for a possible link between expression of these genes and drug response as identified by the classic approach, we investigated whether these three genes are involved in the resistance mechanism in the cell lines of the corresponding cancer type. To do this, we tested the correlation between these genes and response to MAPK pathway inhibitors within the respective cancer type. None of these genes showed a significant correlation with the drug response (Supplementary Fig. S2.6D, S2.6E & S2.6F) (Benjamini-Hochberg corrected $p > 0.05$, Pearson correlation). Unless this is due to small sample size and multiple testing correction, this supports the conclusion that

these gene expression features are selected as a proxy for cancer type and are not directly associated with drug response. Hence, TANDEM more accurately indicates the cancer type as a predictive feature. Altogether, we have shown that by using the different data types in a more balanced fashion, TANDEM replaces part of the gene expression signatures by various upstream features, such as mutations and cancer type features (MAPK pathway inhibitors). At the same time, for the gene expression features that are selected by TANDEM, such as *DUSP*6 (Trametinib) and *SLFN*11 (DDAs), we can rule out that they are merely selected as a proxy for a specific cancer type.

2.4. DISCUSSION

Large-scale pharmacogenomics screens can offer insights into relations between molecular data and drug response. By integrating the various data types, the molecular data can be comprehensively associated to drug response. However, we have shown that the classic approach for data integration (Elastic Net regression on all molecular data types simultaneously) results in models that are largely based on gene expression. This can be attributed to the redundancy in information between the upstream and downstream data. Here, we introduced TANDEM, an approach that preferentially uses the upstream data types, and only adds gene expression when necessary. The resulting models have a much larger contribution of upstream data types, while retaining the same predictive performance as the classic approach.

The main advantage of TANDEM is that the resulting models are more interpretable. By focusing on the upstream data types first, the analysis is prevented from being confounded by the expression of genes that are either specific to the cancer type or serve as 'signatures' of the aberration status of upstream genes. Yet, because the model uses gene expression in the second stage, our method also identifies relevant genes, such as SLFN11 (DNA damaging agents) or DUSP6 (Trametinib), based on their gene expression patterns. De Bin et al. [6] have investigated additional strategies to combine redundant data, in particular clinical and molecular data. In their 'favoring' strategy, they remove the regularization penalty from the clinical data to 'favor' clinical data over the rest. This approach was not feasible in our setting, as the upstream data is high-dimensional and removing the regularization would result in the inversion of a singular matrix. Similar to their 'dimension reduction' strategy, we reduced the dimensionality of the gene expression data, but we found that this still leads to models that are dominated by gene expression data (Supplementary Fig. S2.1C). For the combination of multiple molecular data types, we found that a two-stage approach (in their terminology: a 'residuals strategy') works well to combine upstream and downstream data types.

Redundancy between molecular data types has been explored before. Wang et al. [25] have shown that the information from methylation status is captured in gene expression profiles. Although they did not study drug response prediction in cell lines, but rather investigated clinical outcome in patients, their results support our idea of redundancy captured by upstream and downstream data types. In the model by Wang et al. [25], the gene expression is decomposed in two parts, based on whether it can be modulated by methylation. This can provide insight in relations between methylation and gene expression features. Explicitly modeling the relations between gene expression and upstream data could be an interesting extension for TANDEM.

Similar to the redundancy between methylation and gene expression, Iorio, et al. [15] observed that, in GDSC1000, the gene expression data captures a large fraction of the information regarding the cancer type. In agreement with the observations made by Iorio et al. [15], we found that the cancer type features show the strongest redundancy with gene expression. We extended these ideas by considering not only the redundancy between gene expression and either methylation or cancer type, but by jointly considering all other data types. In the future, it would be interesting to assess whether gene expression also captures information from other molecular effects, such as miRNAs. In this work, we have introduced TANDEM, a two-stage approach that improves the interpretability of the resulting drug response models by focusing on upstream features, while retaining good predictive performance. We believe that advances in the integrated analysis of multiple molecular data types will lead to a better understanding of the mechanisms of drug response and ultimately to improved treatments in the clinic.

ACKNOWLEDGEMENTS

We thank Ultan McDermott and Mathew Garnett for early pre-publication access to the GDSC1000 pharmacogenomics data set. We also thank Gergana Bounova and Remco Nagel for critically reading the manuscript and providing feedback.

FUNDING

This work was funded by the ERC Synergy Project CombatCancer.

2.5. SUPPLEMENTARY METHODS & MATERIALS

2.5.1. RELATIVE CONTRIBUTION OF EACH DATA TYPE TO THE PREDICTION

Throughout, all columns of $\mathbf{X}$ and y are assumed to be of zero mean. Consider a feature matrix $\mathbf{X}$ that consists out of q concatenated data types.

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_q]$$

Consequently, β can be interpreted as a concatenation of q data type specific coefficient vectors.

$$\beta = [\beta_1, \beta_2, \cdots, \beta_q]$$

A prediction from a linear regression model can be made by:

$$\hat{y} = \mathbf{X}\beta$$

In order to determine the relative contribution of each data type, we created a predictive model per data type.

$$\begin{aligned}
\hat{y}_1 &= \mathbf{X}[\beta_1, 0, \cdots, 0]^T \\
\hat{y}_2 &= \mathbf{X}[0, \beta_2, \cdots, 0]^T \\
&\vdots \\
\hat{y}_q &= \mathbf{X}[0, 0, \cdots, \beta_q]^T
\end{aligned}$$

Meaning that:

$$\hat{y} = \hat{y}_1 + \hat{y}_2 + \cdots + \hat{y}_q$$
$$\|\hat{y}\|_2^2 = \|\hat{y}_1 + \hat{y}_2 + \cdots + \hat{y}_q\|_2^2$$

Next, we quantified the relative contribution RC_i of each data type ($i = 1 \cdots q$) to the prediction by taking the sum-of-squares of a prediction from a certain data type ($\hat{y}_i$) over the sum-of-squares of the prediction ($\hat{y}$). This is analogous to the coefficient of determination (R^2), in which the sum-of-squares of a prediction is divided by the sum-of-squares of the original.

$$RC_i = \frac{\|\hat{y}_i\|_2^2}{\|\hat{y}\|_2^2}$$

When all projections $\hat{y}_i$ are uncorrelated to each other, the angle between them is 90 degrees. Using Pythagoras' Theorem, we can then show that the relative contributions RC_i sum to 1 (Supplementary Fig. S2.7A).

$$\begin{aligned}\|\hat{y}\|_2^2 &= \|\hat{y}_1 + \hat{y}_2 + \cdots + \hat{y}_q\|_2^2 \\ &= \|\hat{y}_1\|_2^2 + \|\hat{y}_2\|_2^2 + \cdots + \|\hat{y}_q\|_2^2\end{aligned}$$

Hence

$$\frac{\|\hat{y}_1\|_2^2}{\|\hat{y}\|_2^2} + \frac{\|\hat{y}_2\|_2^2}{\|\hat{y}\|_2^2} + \cdots + \frac{\|\hat{y}_q\|_2^2}{\|\hat{y}\|_2^2} = 1$$

However, in most practical cases, $\hat{y}_i$ are correlated and the relative contributions as determined above do not sum to one. We used the Law of Cosines, a generalization of Pythagoras' Theorem for vectors whose angle is not 90 degrees, to adjust for these correlations (Supplementary Fig. S2.7B).

$$RC_1 = \frac{\|\hat{y}_1\|_2^2 + \hat{y}_1^T\hat{y}_2 + \hat{y}_1^T\hat{y}_3 + \cdots + \hat{y}_1^T\hat{y}_q}{\|\hat{y}\|_2^2}$$
$$RC_2 = \frac{\|\hat{y}_2\|_2^2 + \hat{y}_2^T\hat{y}_1 + \hat{y}_2^T\hat{y}_3 + \cdots + \hat{y}_2^T\hat{y}_q}{\|\hat{y}\|_2^2}$$
$$\vdots$$
$$RC_q = \frac{\|\hat{y}_q\|_2^2 + \hat{y}_q^T\hat{y}_1 + \hat{y}_q^T\hat{y}_2 + \cdots + \hat{y}_q^T\hat{y}_{q-1}}{\|\hat{y}\|_2^2}$$

When $\|\hat{y}\|_2^2$ is small and the inner product between two predictions $\hat{y}_i^T\hat{y}_j$ ($i \neq j$) is negative, the relative contribution RC_i can be negative. This is infrequently observed ($n = 4/265$ for the classic approach, $n = 0/265$ for TANDEM) and if observed the magnitude is

small ($< 0.1\%$ of $\sum RC_i$). To maintain consistency, we use the absolute value of the RC_i and normalize it such that the RC_i sum to one.

$$RC_i^* = \frac{|RC_i|}{\sum_{j=1}^{q} |RC_j|}$$

For all results where we consider the relative contribution per data type, we only took into account drugs for which we achieved a predictive performance $r > 0.4$. This prevents models with poor predictive performance from confounding the analysis.

2.5.2. RULING OUT DIMENSIONALITY, CONTINUOUS AND SCALING AS A CAUSE FOR THE DOMINATION OF GENE EXPRESSION

We aimed to rule out three factors that could influence the contribution of gene expression. Since the gene expression data has a much higher dimensionality than the other data types, we tested whether reducing the number of features would affect the contribution of gene expression data. To this end, we selected the top m features with the highest variance (m = 16,000, 8,000, 4,000, 2,000 and 1,000). Subsequently, we predicted drug response using the reduced gene expression data and the upstream data and assessed the relative contribution of each data type.

Likewise, to rule out that the gene expression dominates the analysis because its features are continuous (whereas all upstream data types are binary), we binarized the gene expression features. For each gene, we replaced the corresponding feature with two binary features: one indicating low expression and one indicating high expression. We have observed that across upstream features, the average percentage of cell lines encoded with a one was four percent. Therefore, for each gene we determined the 4th and the 96th percentile of the expression values and used those values as thresholds for the low and high expression features of that respective gene.

Finally, the binary features and the continuous features have very different ranges of values, which could lead to differences in scaling and thereby affect the relative contribution of the data types. However, these differences in scaling are corrected for by auto-scaling all features. Doing so is common practice in machine learning and is used per default in the glmnet package.

2.5.3. PATHWAY ENRICHMENT

We have downloaded version 5 of the KEGG pathways from MSigDB [23]. From these data, we excluded all pathways that are directly related to a disease, such as 'Pathways in Cancer' and 'Type 1 Diabetes Mellitus' (full list can be found in the Supplementary Data of Aben et al. (2016) [1]).

We used a hypergeometric test to test for enrichment of selected features within a pathway. Since both the upstream features and the KEGG pathways contain many well-studied genes, preferentially using the upstream features for drug response prediction could introduce a bias in the pathway enrichment. To correct for this bias, we created a background distribution using genes that are present in at least one KEGG pathway. We controlled the p-values for FDR by applying Benjamini-Hochberg correction per drug.

2.6. SUPPLEMENTARY TABLES

Measure	Defined as	Scale	Example of scale
Predictive performance (R)	The Pearson correlation coefficient between the observed and the predicted IC50s	Per drug	Trametinib
Relative contribution per data type	The portion of the prediction using all data types that is explained by a given data type	Per drug & per data type	Trametinib & mutation data
Feature importance	The size of the regression coefficient, corrected for the variance of the corresponding feature	Per drug & per feature	Trametinib BRAF mutation

Table S2.1: Throughout this work, we use the measures listed in this table to compare TANDEM with the classic approach.

2.7. SUPPLEMENTARY FIGURES

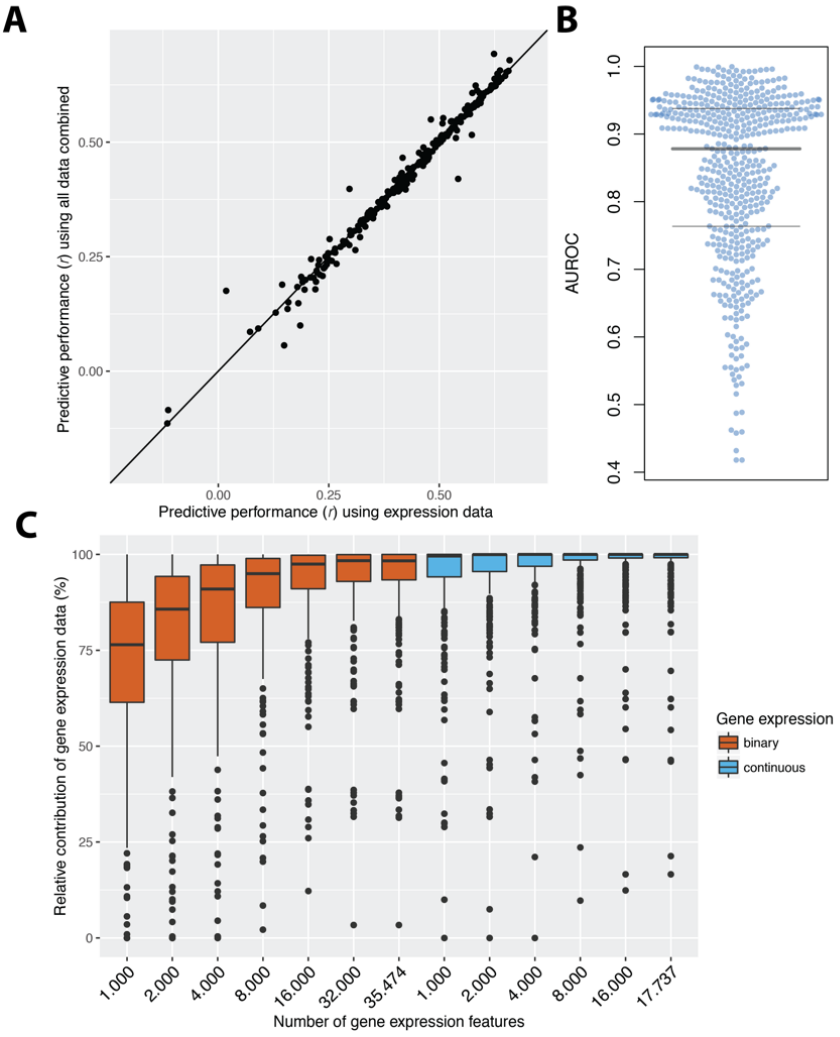


Figure S2.1: (A) Predictive performance (Pearson correlation between measured IC50s and predictions) across 265 drugs of models based on gene expression only versus models that combine all data types. (B) Predictive performance (AUROC) obtained when using gene expression data to predict upstream features related to drug response. (C) Relative contribution of gene expression under different scenarios. We reduced the number of features, to test the influence of the high dimensionality of gene expression data. Likewise, we converted the gene expression to binary format, to rule out effects from mixing binary and continuous features. For the binarized gene expression, each gene was encoded using two binary features (indicating high and low expression respectively), hence the total number of features was twice as high compared to the continuous setting.

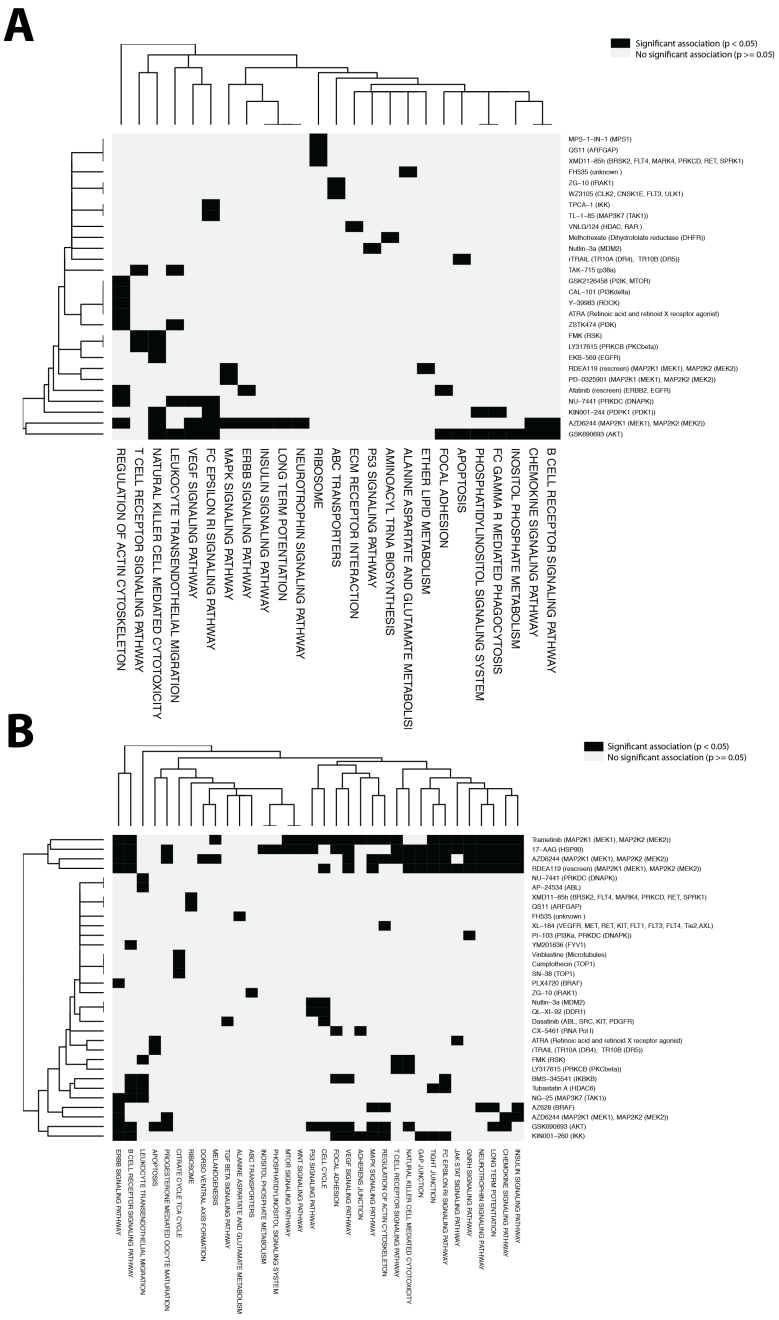


Figure S2.2: Heatmaps indicating significant pathway enrichments of the genes associated with response to a drug in (A) the classic approach and (B) TANDEM. Drugs and pathways without significant enrichments were left out of the figure.

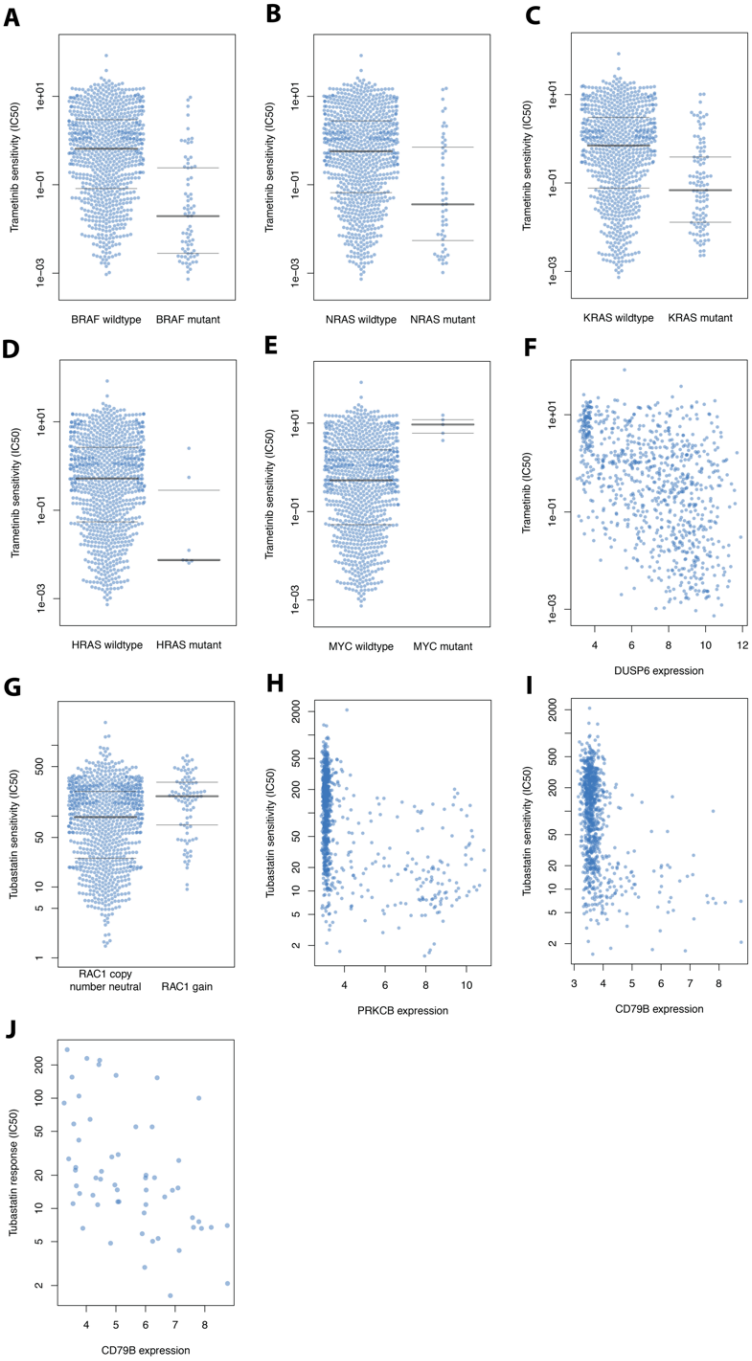


Figure S2.3: (A-E) Sensitivity to Trametinib (IC50) of mutant versus w.t. cell lines for (A) *BRAF*, (B) *NRAS*, (C) *KRAS*, (D) *HRAS* and (E) *MYC*. (F) Sensitivity to Trametinib (IC50) plotted against *DUSP6* expression. (G) Sensitivity to Tubastatin (IC50) of *RAC1* copy number neutral versus *RAC1* copy number gain cell lines. (H-J) Sensitivity to Tubastatin (IC50) plotted against expression of (H) *PRKCB*, (I) *BTK* and (J) *CD79B*.

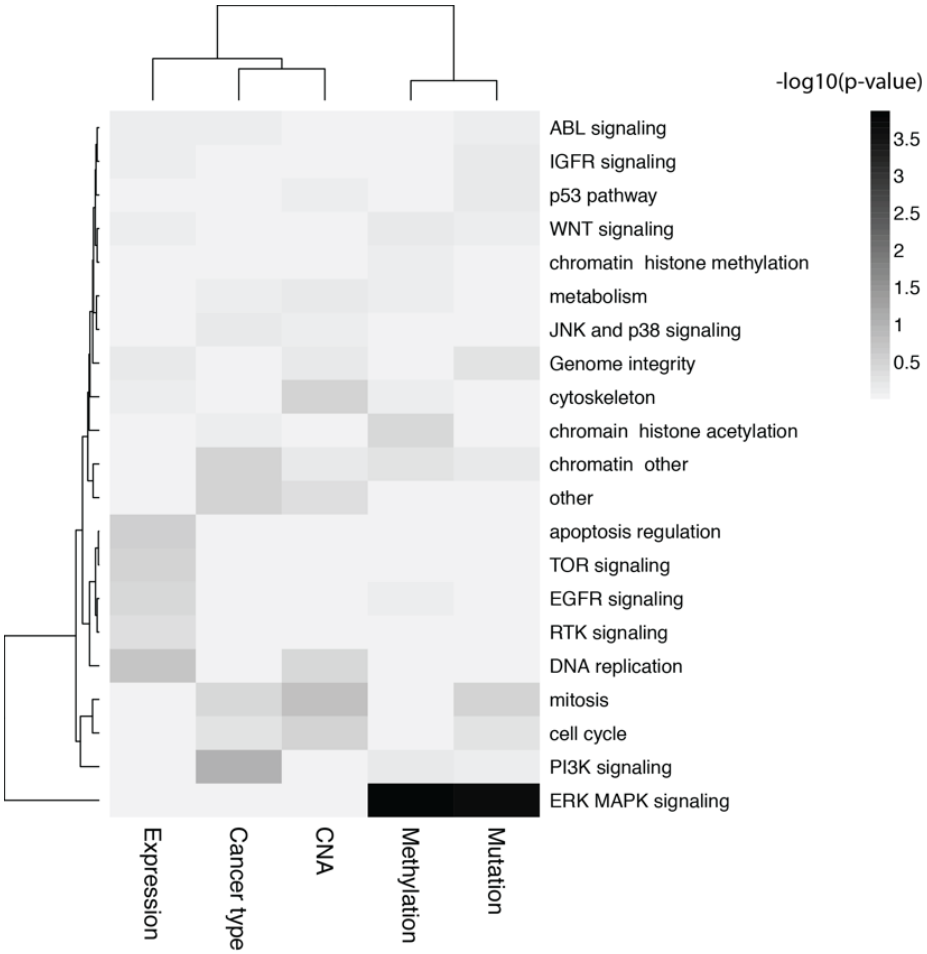


Figure S2.4: Heatmap of significant enrichments between drug classes and data types. For each drug class, TANDEM was used to determine the relative contribution that each data type makes to the prediction, using only the drugs for which a model could be build with predictive performance $r > 0.4$.

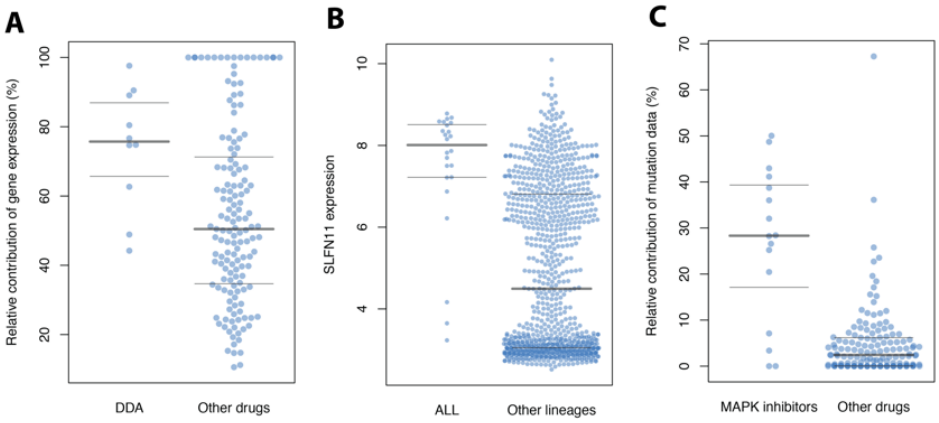
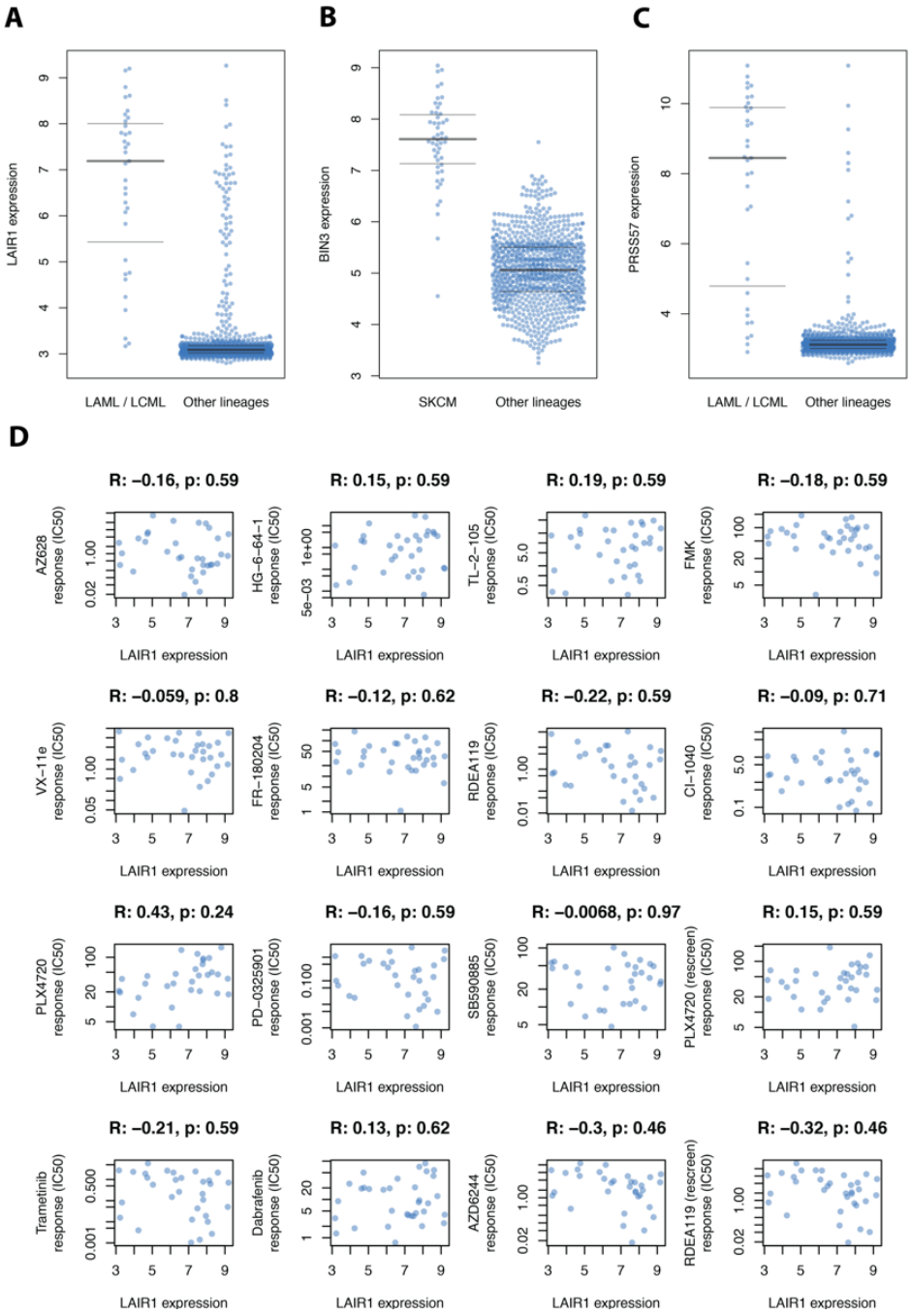
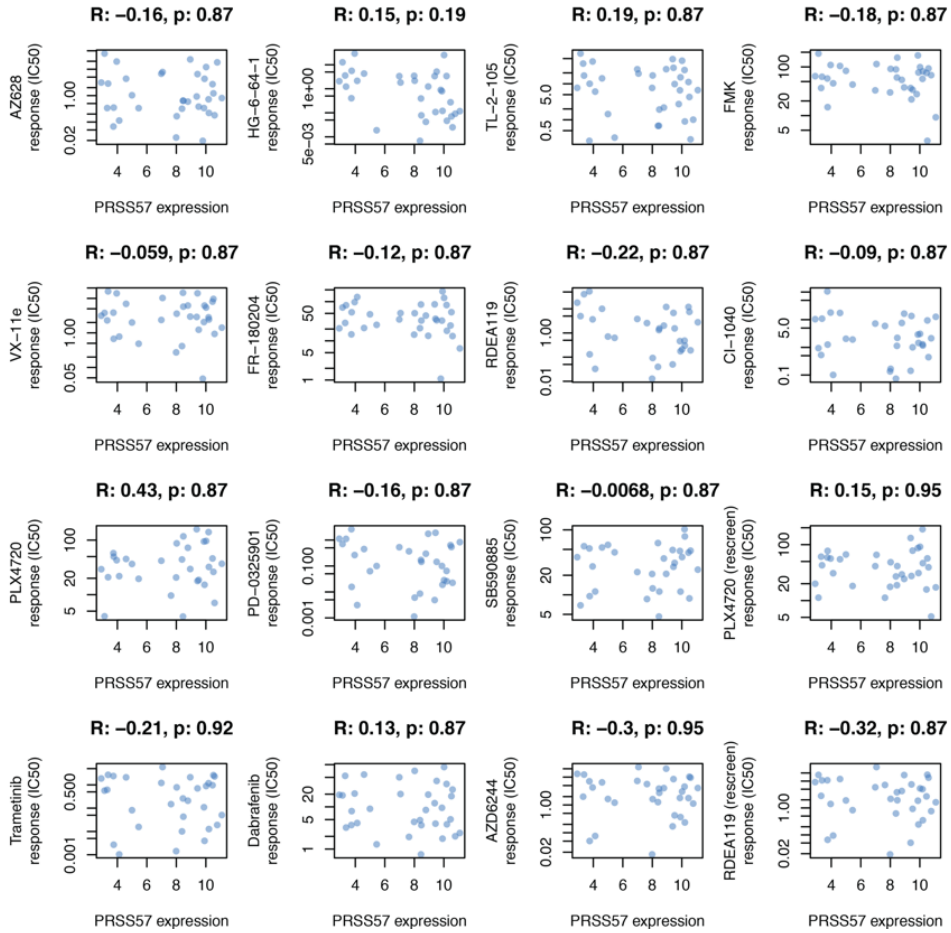


Figure S2.5: (A) Relative contribution of gene expression data, using the two-stage approach, for DNA damaging agents (DDA) versus all other drugs. (B) *SLFN*11 expression of ALL (acute lymphoblastic leukemia) cell lines versus those of other tissue of origin. (C) Relative contribution of mutation data, using the two-stage approach, for MAPK pathway inhibitors versus all other drugs. For (A) and (C), only compounds for which we obtained a predictive performance of $r > 0.4$ were taken into account.


A
LAIR1 expression
LAML / LCML
Other lineages
B
BIN3 expression
SKCM
Other lineages
C
PRSS57 expression
LAML / LCML
Other lineages
D
R: −0.16, p: 0.59
AZ628 response (IC50)
R: 0.15, p: 0.59
HG−6−64−1 response (IC50)
R: 0.19, p: 0.59
TL−2−105 response (IC50)
R: −0.18, p: 0.59
FMK response (IC50)
R: −0.059, p: 0.8
VX−11e response (IC50)
R: −0.12, p: 0.62
FR−180204 response (IC50)
R: −0.22, p: 0.59
RDEA119 response (IC50)
R: −0.09, p: 0.71
CI−1040 response (IC50)
R: 0.43, p: 0.24
PLX4720 response (IC50)
R: −0.16, p: 0.59
PD−0325901 response (IC50)
R: −0.0068, p: 0.97
SB590885 response (IC50)
R: 0.15, p: 0.59
PLX4720 (rescreen) response (IC50)
R: −0.21, p: 0.59
Trametinib response (IC50)
R: 0.13, p: 0.62
Dabrafenib response (IC50)
R: −0.3, p: 0.46
AZD6244 response (IC50)
R: −0.32, p: 0.46
RDEA119 (rescreen) response (IC50)
LAIR1 expression

E


R: −0.16, p: 0.87
AZ628 response (IC50)
R: 0.15, p: 0.19
HG−6−64−1 response (IC50)
R: 0.19, p: 0.87
TL−2−105 response (IC50)
R: −0.18, p: 0.87
FMK response (IC50)
R: −0.059, p: 0.87
VX−11e response (IC50)
R: −0.12, p: 0.87
FR−180204 response (IC50)
R: −0.22, p: 0.87
RDEA119 response (IC50)
R: −0.09, p: 0.87
CI−1040 response (IC50)
R: 0.43, p: 0.87
PLX4720 response (IC50)
R: −0.16, p: 0.87
PD−0325901 response (IC50)
R: −0.0068, p: 0.87
SB590885 response (IC50)
R: 0.15, p: 0.95
PLX4720 (rescreen) response (IC50)
R: −0.21, p: 0.92
Trametinib response (IC50)
R: 0.13, p: 0.87
Dabrafenib response (IC50)
R: −0.3, p: 0.95
AZD6244 response (IC50)
R: −0.32, p: 0.87
RDEA119 (rescreen) response (IC50)
PRSS57 expression

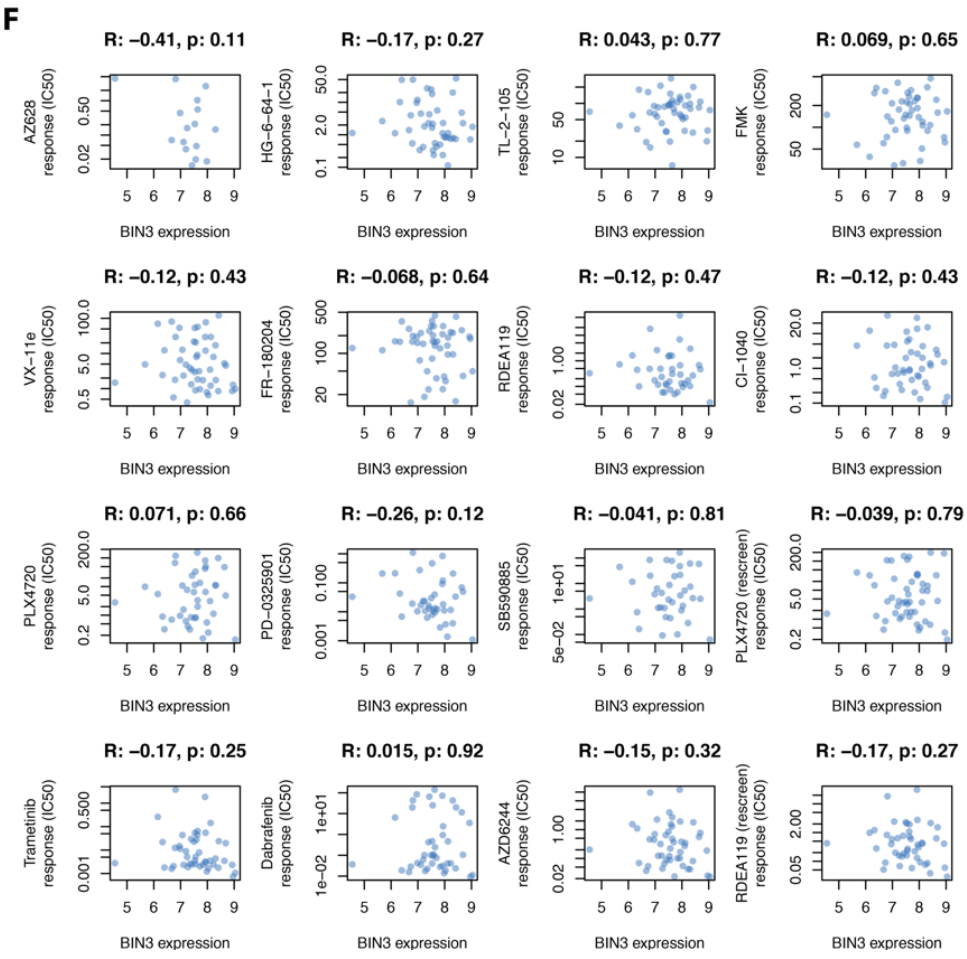


Figure S2.6: (A) *LAIR*1 expression for LAML and LCML cell lines versus cell lines of other cancer types. (B) *PRSS*57 expression for LAML and LCML cell lines versus cell lines of other cancer types. (C) *BIN*3 expression for SKCM cell lines versus cell lines of other cancer types. (D & E) Pearson correlation between the drug response, for each of the 16 MAPK pathway inhibitors, and the expression of (D) *LAIR*1 and (E) *PRSS*57 in LAML and LCML cell lines,. (F) Pearson correlation between the drug response, for each of the 16 MAPK pathway inhibitors, and the expression of *BIN*3 in SKCM cell lines.

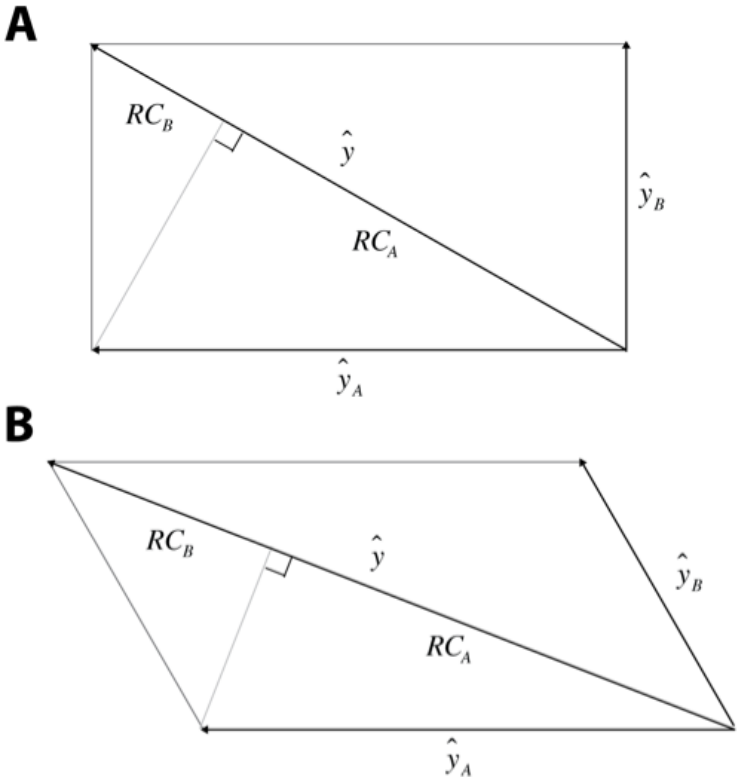


Figure S2.7: Geometric interpretation of the relative contribution per data type, for the 2D case (i.e. for two data types). We distinguish between two cases: the types are either (A) uncorrelated or (B) positively correlated.

REFERENCES

[1] Nanne Aben, Daniel J Vis, Magali Michaut, and Lodewyk Fa Wessels. Tandem: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics*, 32(17):i413–i420, 2016.

[2] Fuad Bahram, Natalie von der Lehr, Cihan Cetinkaya, and Lars-Gunnar Larsson. c-myc hot spot mutations in lymphomas result in inefficient ubiquitination and decreased proteasome-mediated turnover. *Blood*, 95(6):2104–2110, 2000.

[3] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603, 2012.

[4] Kyle V Butler, Jay Kalin, Camille Brochier, Guilio Vistoli, Brett Langley, and Alan P Kozikowski. Rational design and simple chemistry yield a superior, neuroprotective hdac6 inhibitor, tubastatin a. *Journal of the American Chemical Society*, 132(31):10842–10846, 2010.

[5] James C Costello, Laura M Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P Menden, Nicholas J Wang, Mukesh Bansal, Petteri Hintsanen, Suleiman A Khan, John-Patrick Mpindi, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*, 32(12):1202, 2014.

[6] Riccardo De Bin, Willi Sauerbrei, and Anne-Laure Boulesteix. Investigating the prediction ability of survival models based on both clinical and omics data: two case studies. *Statistics in medicine*, 33(30):5310–5329, 2014.

[7] James DeGregori. Surprising dependency for retinoblastoma protein in ras-mediated tumorigenesis. *Molecular and cellular biology*, 26(4):1165–1169, 2006.

[8] Jerome Friedman and Trevor Hastie. glmnet: Lasso and elastic-net regularized generalized linear models. 2009.

[9] Toru Furukawa, Etsuko Tanji, Shanhai Xu, and Akira Horii. Feedback regulation of dusp6 transcription responding to mapk1 via ets2 in human cells. *Biochemical and biophysical research communications*, 377(1):317–320, 2008.

[10] Ya-sheng Gao, Charlotte C Hubbert, Jianrong Lu, Yi-Shan Lee, Joo-Yong Lee, and Tso-Pang Yao. Histone deacetylase 6 regulates growth factor-induced actin remodeling and endocytosis. *Molecular and cellular biology*, 27(24):8637–8647, 2007.

[11] Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, Chris D Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I Richard Thompson, Xi Luo, Jorge Soares, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570, 2012.

[12] Christian R Geest and Paul J Coffer. Mapk signaling pathways in the regulation of hematopoiesis. *Journal of leukocyte biology*, 86(2):237–250, 2009.

[13] Charlotte Hubbert, Amaris Guardiola, Rong Shao, Yoshiharu Kawaguchi, Akihiro Ito, Andrew Nixon, Minoru Yoshida, Xiao-Fan Wang, and Tso-Pang Yao. Hdac6 is a microtubule-associated deacetylase. *Nature*, 417(6887):455, 2002.

[14] Gajanan S Inamdar, SubbaRao V Madhunapantula, and Gavin P Robertson. Targeting the mapk pathway in melanoma: why some approaches succeed and other fail. *Biochemical pharmacology*, 80(5):624–637, 2010.

[15] Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, 2016.

[16] In Sock Jang, Elias Chaibub Neto, Justin Guinney, Stephen H Friend, and Adam A Margolin. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. In *Biocomputing 2014*, pages 63–74. World Scientific, 2014.

[17] Junping Jing, Joel Greshock, Joanna Dawn Holbrook, Aidan Gilmartin, Xiping Zhang, Elizabeth McNeil, Theresa Conway, Christopher Moy, Sylvie Laquerre, Kurt Bachman, et al. Comprehensive predictive biomarker analysis for mek inhibitor gsk1120212. *Molecular cancer therapeutics*, 11(3):720–729, 2012.

[18] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[19] Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: back to metabolism in kegg. *Nucleic acids research*, 42(D1):D199–D205, 2013.

[20] Saverio Minucci and Pier Giuseppe Pelicci. Histone deacetylase inhibitors and the promise of epigenetic (and more) treatments for cancer. *Nature Reviews Cancer*, 6(1):38, 2006.

[21] Rajani K Ravi, Erich Weber, Martin McMahon, Jerry R Williams, Stephen Baylin, Asoke Mal, Marian L Harter, Larry E Dillehay, Pier Paolo Claudio, Antonio Giordano, et al. Activated raf-1 causes growth arrest in human small cell lung cancer cells. *The Journal of clinical investigation*, 101(1):153–159, 1998.

[22] Loredana Santo, Teru Hideshima, Andrew L Kung, Jen-Chieh Tseng, David Tamang, Min Yang, Matthew Jarpe, John H van Duzer, Ralph Mazitschek, Walter C Ogier, et al. Preclinical activity, pharmacodynamic and pharmacokinetic properties of a selective hdac6 inhibitor, acy-1215, in combination with bortezomib in multiple myeloma. *Blood*, pages blood–2011, 2012.

[23] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[24] Santosh Vishwakarma, Lakshmi R Iyer, Milind Muley, Pankaj Kumar Singh, Arun Shastry, Ambrish Saxena, Jayanarayan Kulathingal, G Vijaykanth, J Raghul, Navin Rajesh, et al. Tubastatin, a selective histone deacetylase 6 inhibitor shows anti-inflammatory and anti-rheumatic effects. *International immunopharmacology*, 16(1):72–78, 2013.

[25] Wenting Wang, Veerabhadran Baladandayuthapani, Jeffrey S Morris, Bradley M Broom, Ganiraju Manyam, and Kim-Anh Do. ibag: integrative bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, 29(2):149–159, 2012.

[26] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013.

[27] Gabriele Zoppoli, Marie Regairaz, Elisabetta Leo, William C Reinhold, Sudhir Varma, Alberto Ballestrero, James H Doroshow, and Yves Pommier. Putative dna/rna helicase schlafen-11 (slfn11) sensitizes cancer cells to dna-damaging agents. *Proceedings of the National Academy of Sciences*, 109(37):15030–15035, 2012.

[28] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

3

iTOP: Inferring the Topology of Omics Data

Nanne Aben, Johan A. Westerhuis, Yipeng Song, Henk A.L. Kiers, Magali Michaut, Age K. Smilde, Lodewyk F.A. Wessels

Parts of this chapter have been published in Bioinformatics 34:i988–i996 (2018) [2].

ABSTRACT

Motivation: *In biology, we are often faced with multiple datasets recorded on the same set of objects, such as multi-omics and phenotypic data of the same tumors. These datasets are typically not independent from each other. For example, methylation may influence gene expression, which may, in turn, influence drug response. Such relationships can strongly affect analyses performed on the data, as we have previously shown for the identification of biomarkers of drug response. Therefore, it is important to be able to chart the relationships between datasets.*

Results: *We present iTOP, a methodology to infer a topology of relationships between datasets. We base this methodology on the RV coefficient, a measure of matrix correlation, which can be used to determine how much information is shared between two datasets. We extended the RV coefficient for partial matrix correlations, which allows the use of graph reconstruction algorithms, such as the PC algorithm, to infer the topologies. In addition, since multi-omics data often contain binary data (e.g. mutations), we also extended the RV coefficient for binary data. Applying iTOP to pharmacogenomics data, we found that gene expression acts as a mediator between most other datasets and drug response: only proteomics clearly shares information with drug response that is not present in gene expression. Based on this result, we used TANDEM, a method for drug response prediction, to identify which variables predictive of drug response were distinct to either gene expression or proteomics.*

Availability: *An implementation of our methodology is available in the R package iTOP on CRAN. Additionally, an R Markdown document with code to reproduce all figures is provided as Supplementary Material.*

Contact: *a.k.smilde@uva.nl and l.wessels@nki.nl*

3.1. INTRODUCTION

Rapid developments in high throughput measurement techniques together with rapid reduction in profiling costs have, for many biological problems, endowed us with multiple molecular datasets recorded on the same set of objects. For example, pharmacogenomics data contain, in addition to cancer type and drug response, various omics datasets (mutation, copy number aberration (CNA), methylation, gene expression and proteomics) recorded on the same set of tumor cell lines [5, 7]. While this provides an unprecedented view on the underlying biological problem, it also comes with some unique challenges. Specifically, the recorded datasets are not independent of each other, but are characterized by specific relationships. For example, copy number alterations and methylation changes may influence gene expression, which may, in turn, influence drug response. As we have demonstrated earlier [1], these relationships can have profound effects on further integrative analyses, especially biomarker discovery. It is therefore imperative to obtain a full quantitative characterization of these relationships, such as the illustrative topology of relationships between datasets depicted in Figure 3.1A.

Here we set out to characterize the relationships between datasets in terms of the amount of information that is shared between a pair of datasets, and, more importantly, how this shared information manifests itself in the relationship of a pair of datasets to a third dataset. For example, suppose we have two datasets, $\mathbf{X}_1$ and $\mathbf{X}_2$. Suppose we

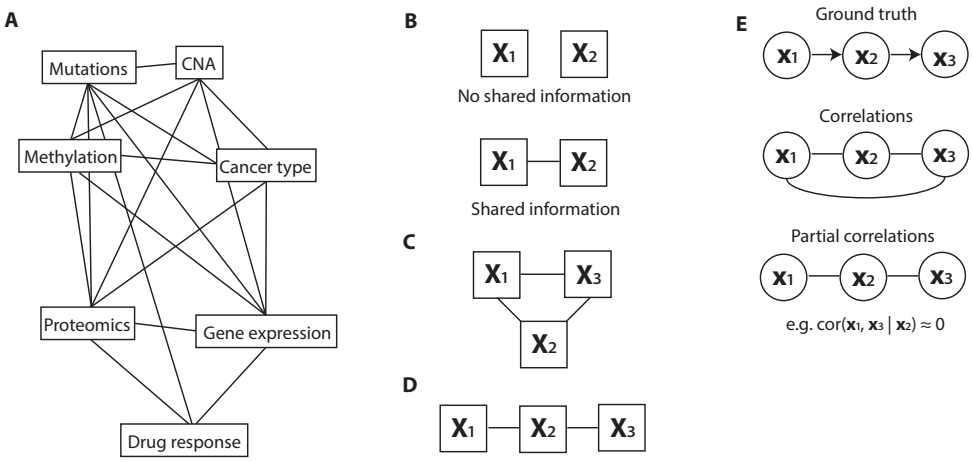


Figure 3.1: High-level overview of this work. (A) The goal of this work is to infer a topology of relationships between pharmacogenomics datasets (an example topology is illustrated here). (B) When two datasets share information (i.e. when their RV coefficient is non-zero), we will indicate them as connected in a topology. (C) A topology of three datasets that all share information. We will convert this topology to the one depicted in (D) if the shared information between $\mathbf{X}_1$ and $\mathbf{X}_3$ is fully contained in $\mathbf{X}_2$. (E) To create these topologies we will draw on methods for inferring a topology between single variables using partial correlations. Top: the original causality graph. Middle: the topology as inferred using correlations. Bottom: the inferred topology using partial correlations.

can characterize the amount of shared information between $\mathbf{X}_1$ and $\mathbf{X}_2$ by a number between 0 and 1, with 0 being no shared information and 1 representing maximal overlap in information (Figure 3.1B). This characterization of pairwise relationships can be informative as such, as it can reveal whether, for example, there is any shared information between gene expression and mutation data. If we now introduce a third dataset, $\mathbf{X}_3$, we can also quantify the amount of information shared between $\mathbf{X}_1$ and $\mathbf{X}_3$ and $\mathbf{X}_2$ and $\mathbf{X}_3$. Assuming that these relationships are non-zero, we obtain the graph in Figure 3.1C. Now it becomes particularly interesting to know whether the shared information between $\mathbf{X}_1$ and $\mathbf{X}_3$ depends on $\mathbf{X}_2$. Specifically, is the shared information between $\mathbf{X}_1$ and $\mathbf{X}_3$ contained in the information in $\mathbf{X}_2$? In other words, does $\mathbf{X}_2$ mediate the effect between $\mathbf{X}_1$ and $\mathbf{X}_3$? When these questions can be answered for all datasets at hand, it reveals the minimal graph that represents the conditional relationships between all datasets. As the number of datasets grows, such a graph not only gives a very concise overview of the relationships, but it is also an important guide in structuring the analyses aimed at finding biomarkers of a given phenotype. More specifically, suppose that $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ represent mutation, gene expression and drug response data for a cell line panel, and that our goal is to extract molecular biomarkers of drug response. Assume that, from our analyses, it emerged that all the information shared between mutation ($\mathbf{X}_1$) and drug response ($\mathbf{X}_3$) is contained in the gene expression data ($\mathbf{X}_2$) (Figure 3.1D). This implies that we only need to employ gene expression data to find biomarkers of drug response.

To infer dataset topologies, we draw upon the approaches employed to infer topologies between single variables (instead of matrices). Specifically, for our earlier example,

we can employ partial correlation, e.g. $cor(\mathbf{x}_1, \mathbf{x}_3|\mathbf{x}_2)$, to quantify the amount of information that is shared between two variables ($\mathbf{x}_1$ and $\mathbf{x}_3$) that is not present in the other variable ($\mathbf{x}_2$). If the effect of $\mathbf{x}_1$ on $\mathbf{x}_3$ is (almost fully) mediated through $\mathbf{x}_2$, it follows that $cor(\mathbf{x}_1, \mathbf{x}_3|\mathbf{x}_2) \approx 0$, which implies that we can remove the direct link between $\mathbf{x}_1$ and $\mathbf{x}_3$ (Figure 3.1E). Graph reconstruction algorithms, such as the PC algorithm [3, 13], use this property to infer the topology between multiple variables.

Here, we propose iTOP, a methodology for inferring topologies between datasets. As with topology inference for single variables, this methodology consists of two components: 1) a measure of (conditional) similarity between datasets and 2) the PC algorithm that employs the (conditional) similarity measure to perform structure learning, i.e. to infer the topology. As similarity measure we employ the RV coefficient [10], a measure of matrix correlation. The basic idea of the RV coefficient is that datasets are correlated when they have a similar configuration (e.g. similar clustering) of the objects. We extend the RV coefficient to be applicable to binary data by using Jaccard similarity to determine the configuration of objects. This allows us to measure the shared information between any of the molecular datasets, including intrinsically binary datasets such as mutation data. In addition, to measure conditional matrix similarity, we extend the RV coefficient for partial matrix correlations. This allows us to quantify the amount of information that is shared between two *datasets* (matrices), but not present in the other dataset, analogous to single variables.

We employ iTOP, i.e. partial matrix correlation in conjunction with the PC algorithm, to infer a topology of relationships between datasets. First, we will demonstrate the RV coefficient with both extensions (i.e. for partial matrix correlations and for binary data) on artificial data. Subsequently, we will use this to infer the topology of relationships between the pharmacogenomics datasets. We show that gene expression acts as a mediator between most other datasets and the drug response, and that only proteomics clearly shares information with drug response that is not present in gene expression. Based on this result, we will employ TANDEM, a method for drug response prediction from multiple datasets [1], to identify markers predictive of drug response that are distinct for proteomics and gene expression.

3.2. Methods and Materials

3.2.1. Matrix correlation using the RV coefficient

For dataset i, consider $\mathbf{X}_i$ the $n \times p_i$ data matrix with objects in the rows and variables in the columns. Here, we assume $\mathbf{X}_i$ to be column-centered (of note, there is no need to scale the columns of $\mathbf{X}_i$). We define the corresponding $n \times n$ configuration matrix $\mathbf{S}_i$ as follows:

$$\mathbf{S}_i = \mathbf{X}_i \mathbf{X}_i^T$$

Now consider a second dataset j, whose data matrix $\mathbf{X}_j$ has the same objects on the same rows as $\mathbf{X}_i$, but has a different set of variables. Hence, $\mathbf{X}_j$ is of size $n \times p_j$. Analogous to $\mathbf{X}_i$, we will define a configuration matrix $\mathbf{S}_j$ for $\mathbf{X}_j$.

$$\mathbf{S}_j = \mathbf{X}_j \mathbf{X}_j^T$$

Using the configuration matrices $\mathbf{S}_i$ and $\mathbf{S}_j$, we can then determine the matrix correlation between these matrices using the RV coefficient:

$$RV(\mathbf{S}_i, \mathbf{S}_j) = \frac{vec(\mathbf{S}_i)^T vec(\mathbf{S}_j)}{\sqrt{vec(\mathbf{S}_i)^T vec(\mathbf{S}_i) \times vec(\mathbf{S}_j)^T vec(\mathbf{S}_j)}}$$

Where $vec(\mathbf{S})$ is the $n^2 \times 1$ vector in which the columns of $\mathbf{S}$ are stacked on top of each other. When $\mathbf{X}_i$ and $\mathbf{X}_j$ are column-centered, then $mean(vec(\mathbf{S}_i)) = 0$ and $mean(vec(\mathbf{S}_j)) = 0$, which means we can interpret the above as a Pearson correlation coefficient.

$$RV(\mathbf{S}_i, \mathbf{S}_j) = cor(vec(\mathbf{S}_i), vec(\mathbf{S}_j))$$

3.2.2. THE MODIFIED RV COEFFICIENT

For data matrices $\mathbf{X}$ where the number of variables is much greater than the number of objects (i.e. $p \gg n$), the RV coefficient is known to be biased upwards [9, 11]. To account for this bias, we subtract the diagonal from the configuration matrix, as in the modified RV coefficient [11].

$$\tilde{\mathbf{S}}_i = \mathbf{S}_i - diag(\mathbf{S}_i)$$
$$\tilde{\mathbf{S}}_j = \mathbf{S}_j - diag(\mathbf{S}_j)$$
$$RV(\tilde{\mathbf{S}}_i, \tilde{\mathbf{S}}_j) = \frac{vec(\tilde{\mathbf{S}}_i)^T vec(\tilde{\mathbf{S}}_j)}{\sqrt{vec(\tilde{\mathbf{S}}_i)^T vec(\tilde{\mathbf{S}}_i) \times vec(\tilde{\mathbf{S}}_j)^T vec(\tilde{\mathbf{S}}_j)}}$$

For a more complete discussion of the modified RV coefficient, as well as our rationale for not using the adjusted RV coefficient [9] instead, we refer to the Supplementary Material.

3.2.3. PARTIAL MATRIX CORRELATIONS

We extend the above matrix correlation formulation to partial matrix correlations. Consider a third dataset, the $n \times p_k$ matrix $\mathbf{X}_k$, that will be processed as above.

$$\mathbf{S}_k = \mathbf{X}_k \mathbf{X}_k^T$$
$$\tilde{\mathbf{S}}_k = \mathbf{S}_k - diag(\mathbf{S}_k)$$

We can then compute the partial matrix correlation between dataset i and j, corrected for dataset k, as

$$RV(\tilde{\mathbf{S}}_i, \tilde{\mathbf{S}}_j | \tilde{\mathbf{S}}_k) = cor(vec(\tilde{\mathbf{S}}_i), vec(\tilde{\mathbf{S}}_j) | vec(\tilde{\mathbf{S}}_k))$$

Of note, the concept of partial matrix correlations has been explored previously by Smouse et al. (1986) [12], who based their measure on the Mantel Test [8]. For a discussion of the Mantel Test and why we prefer to base our measure of partial matrix correlation on the RV coefficient, we refer to the Supplementary Materials.

3.2.4. STATISTICAL INFERENCE FOR PARTIAL MATRIX CORRELATIONS

We provide two methods for statistical inference for partial matrix correlations: significance estimates and confidence intervals. We note that these cannot be determined analytically (e.g. using Fisher Transformation, which is commonly used to derive a p-value for Pearson correlations), as the entries in $vec(\mathbf{S})$ are not i.i.d.: multiple entries in $vec(\mathbf{S})$ correspond to the same object in $\mathbf{S}$. Instead, we will discuss a permutation test for significance estimates and a bootstrapping procedure for calculating confidence intervals.

We used a permutation test to assess significance of a (partial) matrix correlation. In every permutation, the objects of every dataset were independently shuffled and the (partial) matrix correlation was computed on the shuffled data. Subsequently, the observed (partial) matrix correlation was compared to the permuted values, and the p-value was set to

$$p = \begin{cases} \frac{\sum_{i=1}^{nperm} \mathbb{1}_{RV_{obs} < RV_i}}{nperm}, & \text{for } RV_{obs} \geq 0 \\ \frac{\sum_{i=1}^{nperm} \mathbb{1}_{RV_{obs} > RV_i}}{nperm}, & \text{for } RV_{obs} < 0 \end{cases}$$

Where $\mathbb{1}_A$ is the indicator function that equals 1 when A is true, RV_{obs} is the observed (partial) matrix correlation, RV_i the permutated (partial) matrix correlation from the ith permutation and $nperm$ the number of permutations. Throughout the manuscript, we used $nperm = 1000$.

We used a percentile bootstrap procedure to calculate confidence intervals. In each bootstrap, objects were obtained by drawing complete cases randomly (with replacement) from the dataset, after which the (partial) matrix correlation was calculated as defined above. The 99% percentile interval of the obtained (partial) matrix correlations was then used as a confidence interval. Throughout the manuscript, we used 1000 bootstraps to determine a confidence interval.

We note that row-wise permutation of the data matrices ($\mathbf{X}[ind,]$, with ind the indices of the objects after permutation) is equivalent to permutation of both the rows and the columns of the configuration matrices ($\mathbf{S}[ind, ind]$). Using this property, we decided to permute the configuration matrices, as this prevents having to calculate the configuration matrix in each permutation and hence greatly speeds up the calculations. A similar approach was used for bootstrapping.

3.2.5. BINARY SIMILARITY MEASURES

An advantage of converting the data matrices $\mathbf{X}$ to configuration matrices $\mathbf{S}$ is that it allows us to use different similarity measures for different data types. For example, for continuous data, we use:

$$\mathbf{S} = \mathbf{X}\mathbf{X}^T$$

Note that each entry of $\mathbf{S}$ corresponds to an inner product between different objects in $\mathbf{X}$, i.e.

$$\mathbf{S} = \mathbf{X}\mathbf{X}^T = \begin{pmatrix} \mathbf{x}_1^T\mathbf{x}_1 & \mathbf{x}_1^T\mathbf{x}_2 & \cdots & \mathbf{x}_1^T\mathbf{x}_n \\ \mathbf{x}_2^T\mathbf{x}_1 & \mathbf{x}_2^T\mathbf{x}_2 & \cdots & \mathbf{x}_2^T\mathbf{x}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n^T\mathbf{x}_1 & \mathbf{x}_n^T\mathbf{x}_2 & \cdots & \mathbf{x}_n^T\mathbf{x}_n \end{pmatrix}$$

Where $\mathbf{x}_i$ is the i'th row in $\mathbf{X}$ and n is the number of rows in $\mathbf{X}$. We will refer to this similarity measure as 'inner product similarity'.

JACCARD SIMILARITY

For binary data, we use Jaccard similarity. Jaccard similarity is defined as the ratio of the number of elements where these vectors have ones in common and the total number of positions where ones occur in any of these two vectors. Consider the following contingency table.

	$y=0$	$y=1$
$x=0$	a	c
$x=1$	b	d

Where a is the number of elements where $x=0$ and $y=0$, b is the number of elements where $x=1$ and $y=0$, etc. The Jaccard Similarity can then be written as:

$$Jaccard(x,y) = \frac{d}{b+c+d}$$

When all $x=0$ and all $y=0$, then $b=c=d=0$, which would result in $Jaccard(x,y) = {}^0/_0$. In these cases, we define the Jaccard similarity as $Jaccard(x,y)=0$.

Note that the Jaccard similarity is based on the number of positive matches (d) and not at all on the number of negative matches (a). This is in line with our intuition of similarity in the binary data at hand (mutation, CNA and cancer type). For example, when two objects share the same mutations, we think this should contribute more to their similarity than the number of mutations that both objects lack.

We define configuration matrices using the Jaccard similarity in the following way:

$$\begin{aligned}\mathbf{S} &= Jaccard\_config(\mathbf{X}) \\ &= \begin{pmatrix} Jaccard(\mathbf{x}_1,\mathbf{x}_1) & Jaccard(\mathbf{x}_1,\mathbf{x}_2) & \cdots & Jaccard(\mathbf{x}_1,\mathbf{x}_n) \\ Jaccard(\mathbf{x}_2,\mathbf{x}_1) & Jaccard(\mathbf{x}_2,\mathbf{x}_2) & \cdots & Jaccard(\mathbf{x}_2,\mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ Jaccard(\mathbf{x}_n,\mathbf{x}_1) & Jaccard(\mathbf{x}_n,\mathbf{x}_2) & \cdots & Jaccard(\mathbf{x}_n,\mathbf{x}_n) \end{pmatrix}\end{aligned}$$

KERNEL CENTERING

We used kernel centering to center the configuration matrix $\mathbf{S}$ rather than the underlying data matrix $\mathbf{X}$. Essentially, kernel centering is double centering (i.e. column- and row-wise centering) of the configuration matrix $\mathbf{S}$ (or in other words: the kernel), which we will show to be equal to first column-centering the data matrix $\mathbf{X}$ and then computing $\mathbf{S} = \mathbf{X}\mathbf{X}^T$. Consider $\mathbf{X}$ the original data matrix and $\overline{\mathbf{X}}$ the column-centered data matrix. Likewise, consider $\mathbf{S}$ the original configuration matrix and $\overline{\mathbf{S}}$ the centered configuration matrix. Finally, consider $\mathbf{m}$ the column-wise means of $\mathbf{X}$ and n the number of rows in $\mathbf{X}$.

We will first consider an example using inner products as a similarity measure.

$$
\begin{aligned}
\mathbf{S} &= \mathbf{X}\mathbf{X}^T \\
\overline{\mathbf{S}} &= \overline{\mathbf{X}}\overline{\mathbf{X}}^T \\
&= \left(\mathbf{X} - \mathbf{1}\mathbf{m}^T\right)\left(\mathbf{X} - \mathbf{1}\mathbf{m}^T\right)^T \\
&= \left(\mathbf{X} - \frac{\mathbf{1}\mathbf{1}^T\mathbf{X}}{n}\right)\left(\mathbf{X} - \frac{\mathbf{1}\mathbf{1}^T\mathbf{X}}{n}\right)^T \\
&= \mathbf{X}\mathbf{X}^T - \frac{\mathbf{1}\mathbf{1}^T\mathbf{X}\mathbf{X}^T}{n} - \frac{\mathbf{X}\mathbf{X}^T\mathbf{1}\mathbf{1}^T}{n} + \frac{\mathbf{1}\mathbf{1}^T\mathbf{X}\mathbf{X}^T\mathbf{1}\mathbf{1}^T}{n^2} \\
&= \mathbf{S} - \frac{\mathbf{1}\mathbf{1}^T\mathbf{S}}{n} - \frac{\mathbf{S}\mathbf{1}\mathbf{1}^T}{n} + \frac{\mathbf{1}\mathbf{1}^T\mathbf{S}\mathbf{1}\mathbf{1}^T}{n^2}
\end{aligned}
$$

Interestingly, the final term expresses the kernel centered $\overline{\mathbf{S}}$ in terms of the non-centered $\mathbf{S}$. This allows us to center configuration matrices that are not based on inner-product similarity, such as $\mathbf{S} = Jaccard\_config(\mathbf{X})$. Column-centering $\mathbf{X}$ (the input space) makes no sense here, as the resulting matrix would not consist of 0s and 1s anymore and hence $Jaccard\_config(\overline{\mathbf{X}})$ is not defined. However, we can use kernel centering here to center the so-called kernel space corresponding to $\mathbf{S}$.

$$
\begin{aligned}
\mathbf{S} &= Jaccard\_config(\mathbf{X}) \\
\overline{\mathbf{S}} &= \mathbf{S} - \frac{\mathbf{1}\mathbf{1}^T\mathbf{S}}{n} - \frac{\mathbf{S}\mathbf{1}\mathbf{1}^T}{n} + \frac{\mathbf{1}\mathbf{1}^T\mathbf{S}\mathbf{1}\mathbf{1}^T}{n^2}
\end{aligned}
$$

3.2.6. PHARMACOGENOMICS DATA

The mutation, copy number aberration (CNA), methylation, cancer type, gene expression and drug response data were sourced from GDSC1000 [5], and the proteomics data were sourced from MCLP [7] (Table 3.1). For the mutation and CNA data, we used the reduced set of Cancer Functional Events (CFEs) [5], resulting in 300 and 425 binary variables respectively. For the methylation data, we used the CpG-island summarized data,

	Dimensionality	Source	Type	Missing values
Mutation	300	GDSC1000	Binary	No
CNA	425	GDSC1000	Binary	No
Methylation	14,429	GDSC1000	Continuous	No
Cancer type	31	GDSC1000	Binary	No
Gene expression	17,419	GDSC1000	Continuous	No
Proteomics	452	MCLP	Continuous	Yes
Drug response	265	GDSC1000	Continuous	Yes

Table 3.1: Overview of the pharmacogenomics datasets used in this manuscript.

resulting in 14,426 continuous variables. For the cancer type data, we used the classification into 30 TCGA cancer types or 'OTHER', resulting in 31 binary variables [5]. For gene expression data, we used the gene level summarized data, resulting in 17,419 continuous variables. The proteomics data consist of 452 variables, of which 108 represent phospho-protein levels and the remaining 344 represent protein abundance levels. For the drug response data, we used the IC50-values (concentration at which half of the cells are killed) for all 265 drugs.

Of the 282 cell lines that were profiled in both GDSC1000 and MCLP, 266 cell lines were characterized across all seven datasets. This number was further reduced due to missing values in the proteomics and drugs response data. For the proteomics data, after removing all variables with >30% missing values, we retained 186 variables. Subsequently, after removing all objects with >30% missing values, we retained 221 objects. We then intersected all datasets with these 221 objects and applied the same two steps to the drug response data, where we retained 206 objects and 217 variables. These 206 objects cover 27 of the 31 cancer types in the GDSC1000 data. The remaining missing values (1% for the proteomics and 5% for the drug response) were imputed using SVD imputation [14] as implemented in the R package bcv.

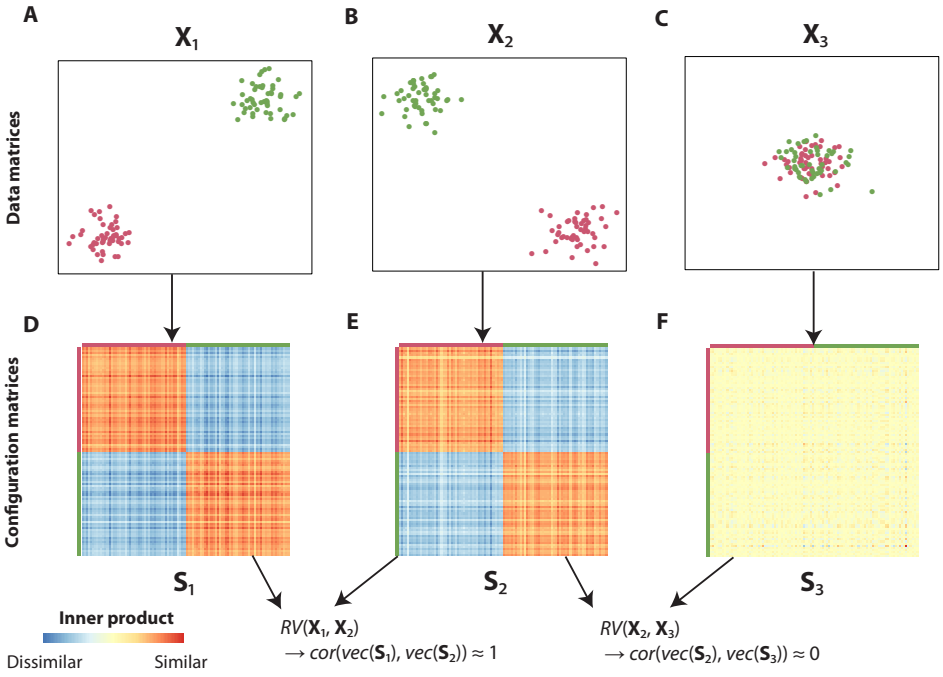


Figure 3.2: The RV coefficient explained using three simple example datasets. The data matrices $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ (represented in A-C) are converted to configuration matrices $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_3$ respectively (D-F). Using the configuration matrices, it can be readily seen that $RV(\mathbf{X}_1,\mathbf{X}_2) \approx 1$ and $RV(\mathbf{X}_2,\mathbf{X}_3) \approx 0$.

3.3. Results

3.3.1. The RV coefficient

To illustrate the RV coefficient, consider the following example. Figure 3.2A represents data matrix $\mathbf{X}_1$, a dataset with two variables and 100 objects, where the first 50 objects form the green cluster and the second 50 objects form the purple cluster. The second data matrix, $\mathbf{X}_2$ (Figure 3.2B), also consists of two variables and the same 100 objects with the same clustering as in $\mathbf{X}_1$. The third data matrix, $\mathbf{X}_3$ (Figure 3.2C), is again a dataset with two variables and the same objects as before, but now without any apparent clustering. When converting these data matrices to configuration matrices (similarity matrices), which indicate the configuration of the different objects with respect to each other, it can be readily observed that $\mathbf{X}_1$ and $\mathbf{X}_2$ contain the same information in terms of clustering (Figure 3.2D & E). Indeed, when computing the RV coefficient between $\mathbf{X}_1$ and $\mathbf{X}_2$ (by computing the Pearson correlation of the vectorized forms of the corresponding configuration matrices, see Methods and Materials), we obtain an RV coefficient close to one, indicating a strong relationship. Conversely, when computing the RV coefficient between $\mathbf{X}_2$ and $\mathbf{X}_3$, where the latter contains no clustering information, we see that the configuration matrices are very different and $RV(\mathbf{X}_2, \mathbf{X}_3) \approx 0$ (Figure 3.2C & F).

3.3.2. Extending the RV coefficient for partial matrix correlations

We illustrate partial matrix correlations using the following example. Consider three datasets: $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$. Let $\mathbf{X}_1$ affect $\mathbf{X}_2$, and let $\mathbf{X}_2$ affect $\mathbf{X}_3$ (Figure 3.3). Observe that, consistent with the proposed causality, $\mathbf{X}_1$ is most similar to $\mathbf{X}_2$ (only the purple cluster

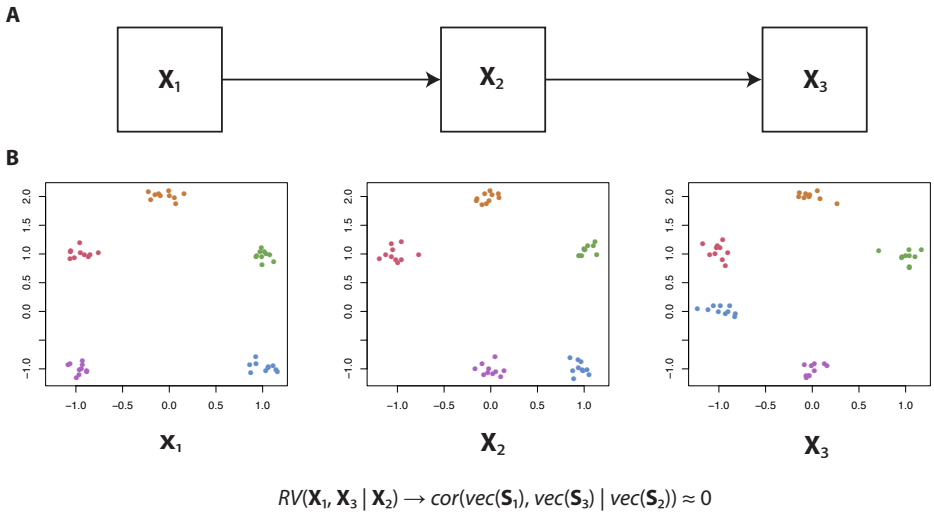


Figure 3.3: Illustration of the partial matrix correlation. (A) We will create artificial data such that $\mathbf{X}_1$ influences $\mathbf{X}_2$, which in turn influences $\mathbf{X}_3$. (B) Artificial data consistent with the abovementioned causality, resulting in $RV(\mathbf{X}_1, \mathbf{X}_3 | \mathbf{X}_2) \approx 0$.

in the bottom-left has been moved) and $\mathbf{X}_3$ is most similar to $\mathbf{X}_2$ (only the blue cluster in the bottom-right has been moved). This of course means that $RV(\mathbf{X}_1,\mathbf{X}_2)$ and $RV(\mathbf{X}_2,\mathbf{X}_3)$ will be non-zero. However, note that also $RV(\mathbf{X}_1,\mathbf{X}_3)$ will be non-zero, as $\mathbf{X}_1$ and $\mathbf{X}_3$ do share information: the top three clusters have the same configuration in both datasets. Therefore, if we were to infer a topology based on the matrix correlations, we cannot rule out a direct link from $\mathbf{X}_1$ to $\mathbf{X}_3$.

Using the partial matrix correlation $RV(\mathbf{X}_1,\mathbf{X}_3|\mathbf{X}_2)$, we can rule out a direct link from $\mathbf{X}_1$ to $\mathbf{X}_3$. As $\mathbf{X}_2$ has the same configuration in the top three clusters, correcting for $\mathbf{X}_2$ results in $RV(\mathbf{X}_1,\mathbf{X}_3|\mathbf{X}_2) = 0.005$, which is not significantly different from zero (p-value: 0.354, 99% confidence interval: -0.27 – 0.28). Therefore, using partial matrix correlations, we can indeed reconstruct the original topology.

3.3.3. EXTENDING THE RV COEFFICIENT FOR BINARY DATA

The RV coefficient has been proposed for comparing data matrices containing continuous values. Specifically, in the original formulation of the RV coefficient, the configuration matrices are determined using the inner product between objects (Methods and Materials), which is tailored to comparing continuous values. To determine (partial) matrix correlations for datasets containing binary values, we propose to create the configuration matrices using Jaccard similarity, which determines similarity between binary variables (Methods and Materials). We assessed the performance of this approach using a simulation study.

First, to establish a reference, we performed a simulation study in which two continuous valued matrices were compared. In this simulation, the values in $\mathbf{X}_1$ and $\mathbf{X}_2$ were randomly drawn from $N(10,1)$ and $N(0,1)$ respectively, where $N(\mu,\sigma)$ represents a Gaussian distribution with mean μ and standard deviation σ. Subsequently, we defined a third matrix as $\mathbf{X}_3 = (1-\alpha)\mathbf{X}_1 + \alpha\mathbf{X}_2$. We compared $RV(\mathbf{X}_1,\mathbf{X}_3)$ for different values of α, and both with and without column-wise centering of the data matrices (Figure 3.4A). Regardless of centering, we found that $RV(\mathbf{X}_1,\mathbf{X}_3) = 1$ for $\alpha = 0$ and $RV(\mathbf{X}_1,\mathbf{X}_3) \approx 0$ for

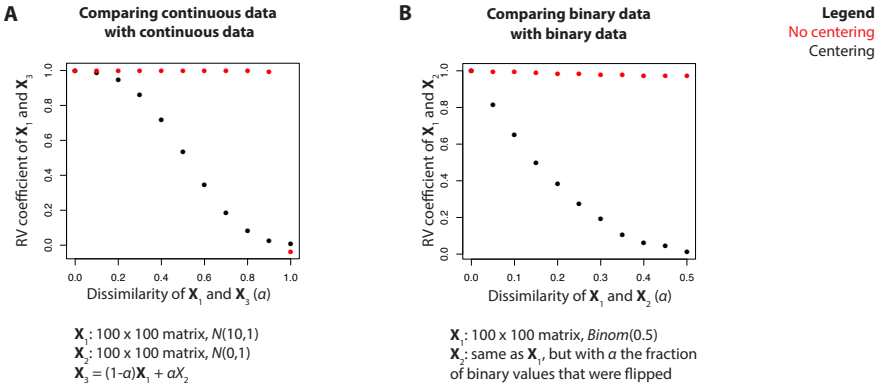


Figure 3.4: Artificial data experiment in which the RV coefficient (y-axis) is measured at different levels of similarity (α, x-axis), both with and without centering, for (A) two continuous datasets and (B) two binary datasets.

$\alpha = 1$, as expected. For intermediate values of α however, we see big differences between the approach using centering and the one without centering. Without centering, $RV(\mathbf{X}_1, \mathbf{X}_3)$ remains very close to 1 for values of α approaching 1, which is counterintuitive. With centering, $RV(\mathbf{X}_1, \mathbf{X}_3)$ slowly decreases to 0 as α increases, which is according to expectation. These differences can be attributed to the fact that inner product distance is dependent on the relative position of the objects with respect to the origin: in the uncentered case, for $\alpha \leq 0.9$, the vectors representing the objects in $\mathbf{X}_1$ and $\mathbf{X}_3$ will be highly collinear, resulting in an RV coefficient close to one (Supplementary Figure 3.6). This experiment emphasizes the importance of centering the data prior to applying the RV coefficient.

We then performed a simulation in which two binary valued matrices were compared. Values in $\mathbf{X}_1$ were randomly drawn from $Binom(0.5)$ (Binomial distribution with $p = 0.5$). $\mathbf{X}_2$ was set equal to $\mathbf{X}_1$, but with α the fraction of binary values that were flipped. We varied α only up to 0.5, as this is the point at which the configuration of objects in $\mathbf{X}_1$ and $\mathbf{X}_2$ is maximally apart (at $\alpha = 1$, $\mathbf{X}_1$ and $\mathbf{X}_2$ are simply inverted and, given that the RV coefficient is rotation independent, the resulting RV coefficient will be 1 again). Again $RV(\mathbf{X}_1, \mathbf{X}_2)$ was compared for different values of α and both with and without centering (Figure 3.4B). As binary data cannot be column centered (it would not be binary anymore after centering), we instead used kernel centering to center the configuration matrix obtained using the Jaccard similarity (Methods and Materials). For $\alpha = 0$, $RV(\mathbf{X}_1, \mathbf{X}_2) = 1$, both with and without centering, as the two matrices are exactly the same. However, for α in (0,50], $RV(\mathbf{X}_1, \mathbf{X}_2)$ remained very close to 1 in the uncentered case, while it slowly decreased to 0 in the centered case. Hence, as at $\alpha = 0.5$ the configuration of $\mathbf{X}_1$ and $\mathbf{X}_2$ is maximally apart, the centered case is preferable.

Using these simulation experiments, we have shown that the Jaccard similarity can be used to construct configuration matrices for binary data. Additionally, we have shown the importance of centering and that kernel centering can be used for the binary case.

3.3.4. APPLICATION TO PHARMACOGENOMICS DATA

We applied the RV coefficient with both extensions to a collection of pharmacogenomics data (a combination of GDSC1000 [5] and MCLP [7], see Methods and Materials) to infer how the different datasets in this collection are related to each other. This collection consists of 3 binary datasets (mutation, CNA and cancer type) and 4 continuous datasets (methylation, gene expression, proteomics and drug response). Intersecting the objects that are present in all datasets resulted in data for 206 objects.

We used the PC algorithm [3, 13] (Supplementary Materials) to study the relationships between datasets. Briefly put, this algorithm starts out with a fully connected graph, where each node corresponds to a dataset, and removes the edge between two datasets $\mathbf{X}_1$ and $\mathbf{X}_2$ when $RV(\mathbf{X}_1, \mathbf{X}_2|\mathrm{C}) \approx 0$ (i.e. when it is not significantly different from 0). This step is repeated for increasingly larger sets of C, from $\mathrm{C} = \emptyset$ (no datasets) to $\mathrm{C} = \mathrm{U} \setminus \{\mathbf{X}_1, \mathbf{X}_2\}$ (all datasets except $\mathbf{X}_1$ and $\mathbf{X}_2$), until either the edge is removed or all possible sets have been assessed. Finally, the PC algorithm attempts to, under certain assumptions, determine the directionality of the edges (Supplementary Materials). However, for the pharmacogenomics data, the algorithm was unable to infer the directionality of any edge in the graph.

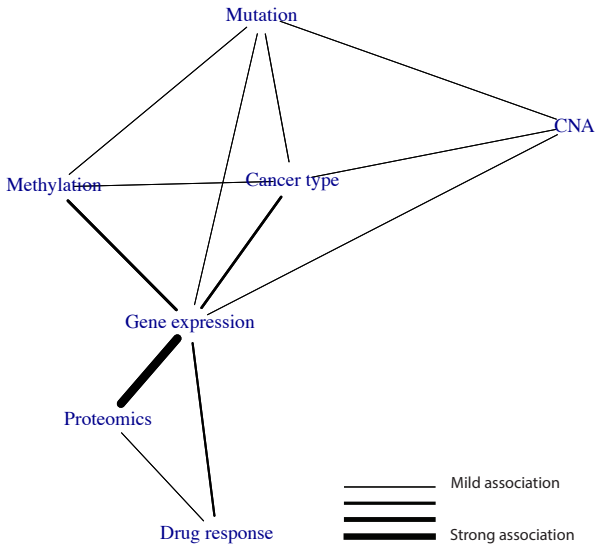


Figure 3.5: Relationships between datasets in the pharmacogenomics data, as determined using the PC algorithm run on the partial matrix correlations. An edge indicates that two datasets share information that is not present in any of the other datasets.

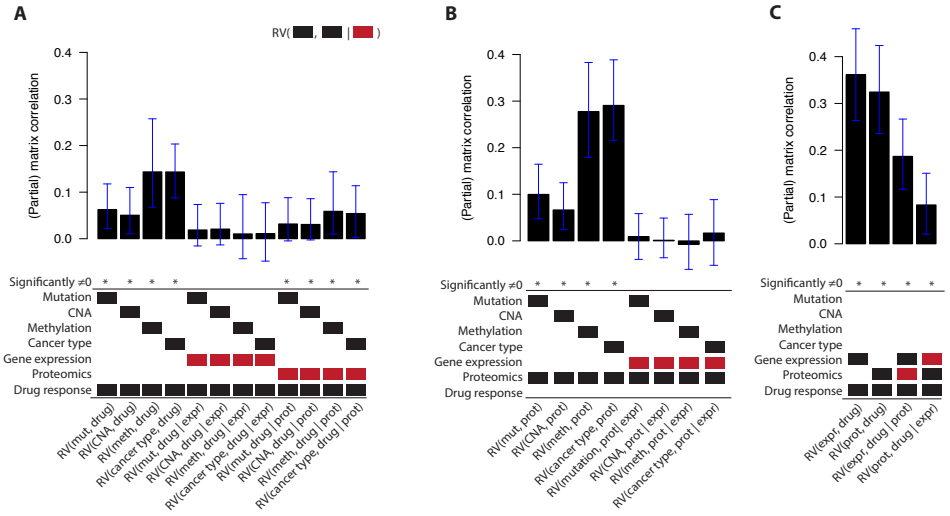


Figure 3.6: The (partial) matrix correlations for different $RV(\mathbf{X}_1, \mathbf{X}_2 | \mathbf{X}_3)$ in the pharmacogenomics data. For each bar in the barplot, $\mathbf{X}_1$ and $\mathbf{X}_2$ are indicated by the black blocks, and $\mathbf{X}_3$ is indicated by the red block. A (partial) matrix correlation was significant when $p < 0.01$. The error bars indicate the 99% confidence interval. Abbreviations: mut, mutation; meth, methylation; expr, gene expression; prot, proteomics; drug, drug response.

Using the approach outlined above, the PC algorithm essentially summarizes the set of all 560 partial matrix correlations in a topology. An important caveat of this approach is that it uses the absence of a significant association to determine the absence of a relation between two datasets. As this may not always be true (there may be such a relation, but we may not have enough objects to detect it), we will also inspect the underlying (partial) matrix correlations and their confidence intervals for the most important hypotheses generated from the topology.

Figure 3.5 shows the topology resulting from the PC algorithm. Gene expression takes up a strikingly central position in the graph, being connected to all other data types. Using the underlying partial correlations and their confidence intervals, we verify that gene expression acts as a mediator between the 'upstream data' (mutation, CNA, methylation and cancer type) on the one hand and the drug response data on the other hand: the partial matrix correlations between these datasets and the drug response drop to nearly zero when correcting for gene expression (Figure 3.6A).

Proteomics also takes up an interesting position in the graph. The proteomics data shows a very strong relationship with gene expression ($RV = 0.76$). Interestingly, using the underlying partial matrix correlations, we see that this relationship fully contains the information shared between the upstream data and proteomics: $RV(\mathbf{X}_i, \text{proteomics} \mid \text{expression}) \approx 0$, for each dataset $\mathbf{X}_i$ in the upstream datasets (Figure 3.6B). Finally, gene expression and proteomics share information with drug response that is not present in the other dataset: $RV(\text{expression}, \text{drug response} \mid \text{proteomics}) > 0$ and $RV(\text{proteomics}, \text{drug response} \mid \text{expression}) > 0$ (Figure 3.6C). Hence, even though gene expression and proteomics share a large amount of information, they both contain unique information with respect to drug response.

Overall, we have shown here that our methodology can be used to infer how different datasets are related to each other.

3.3.5. Identifying which variables predictive of drug response are distinct to either gene expression or proteomics

The topology that we have inferred suggests that for accurate prediction of drug response we only need gene expression and proteomics. Indeed, when we train Elastic Net models [18] (Supplementary Materials) to predict the drug response from either all datasets (other than drug response) or from only gene expression and proteomics, we found that they result in virtually identical predictive performance (Supplementary Figure S3.2A).

We then asked which variables are both predictive of drug response and distinct to either gene expression or proteomics. To answer this question, we used TANDEM [1] (Supplementary Materials). Briefly, given a response vector y (e.g. drug response of a single drug) and two datasets $\mathbf{X}_1$ and $\mathbf{X}_2$ (e.g. gene expression and proteomics), TANDEM uses two stages of Elastic Net regression to first identify all variables in $\mathbf{X}_1$ that are associated with y, and then identify all variables in $\mathbf{X}_2$ that are associated with y but whose information is not present in $\mathbf{X}_1$.

For each drug, we trained two TANDEM models:

- $\text{GEX}_{\text{unique}}$: a model that uses proteomics in the first stage and gene expression in the second stage, thereby identifying variables with information that is unique to the gene expression data.

- $\text{PROT}_{\text{unique}}$: the counterpart of $\text{GEX}_{\text{unique}}$, with gene expression in the first stage and proteomics in the second stage.

We found that $\text{GEX}_{\text{unique}}$ mostly uses proteomics data and $\text{PROT}_{\text{unique}}$ mostly uses gene expression data, while both achieve similar predictive performance (Supplementary Figure S3.2B-D. This is of course not very surprising, as we have already seen using the RV coefficient that a lot of information is shared between the gene expression and proteomics data.

For each drug and for both TANDEM models, we then determined variable importance scores (Supplementary Materials) and averaged these over drugs to identify variables that made the largest overall contribution to the prediction of drug response. For $\text{GEX}_{\text{unique}}$, the most important gene expression variable was ABCB1 expression. ABCB1 is a protein in the cell membrane that pumps foreign substances (including drugs) out of the cell. As such, it is known to be associated with resistance to a wide range of drugs [4]. The proteomics data we considered here did not contain ABCB1, hence it is not unexpected that this information is not present in the proteomics data.

For $\text{PROT}_{\text{unique}}$, the most important variable was MEK1 S217/S221 phosphorylation (pMEK1). The phosphorylation of MEK1 indicates MAPK pathway activation and is hence associated to sensitivity to MAPK pathway inhibitors, such as BRAF, MEK and ERK inhibitors. As the proteomics data contains both phosphorylation and protein abundance variables, we wondered whether one of these classes might be enriched in the distinct proteomics – drug response part. However, we found no significant difference between the variable importance scores in the $\text{PROT}_{\text{unique}}$ models for these two classes ($p = 0.68$, Mann-Whitney U Test) (Supplementary Figure S3.2E).

Altogether, we have shown here that, informed by the topology of the datasets we inferred with iTOP, we can identify which variables correspond to distinct gene expression – drug response and proteomics – drug response relationships.

3.4. DISCUSSION

In this work, we have introduced iTOP, a methodology to infer a topology of relationships between datasets. To this end, we have extended the RV coefficient for partial matrix correlations, allowing one to identify how much information is shared between two datasets, but not present in other datasets. In addition, we have also extended the partial RV coefficient for binary data, using the Jaccard coefficient. We have tested both extensions using artificial data and used them to infer a topology of the pharmacogenomics data. Finally, we have zoomed in on part of the topology and have identified variables predictive of drug response that are distinct to either gene expression or proteomics using TANDEM.

An important caveat of the PC algorithm used in our approach is that the absence of a significant p-value does not necessarily mean the absence of a relationship between two datasets: it can also mean this relationship is present, but that we did not have enough power to detect it. Of note, this also means that the inferred topology can change as the number of objects increases, simply because this enhances our ability to detect very small effects. For these reasons, we suggest to not solely rely on p-values to determine the absence or presence of these links. Instead, we suggest using the PC algorithm as

a tool to summarize the results from the numerous possible partial matrix correlations into a topology, after which the hypotheses generated from this topology should also be assessed by inspecting the relevant (partial) matrix correlations and their confidence intervals. These values will give an indication of both the strength of the associations and how well we can estimate these, and may hence suggest the inclusion of an association that is strong but uncertain, or the exclusion of a certain – but weak – association.

We note that there are other options for binary similarity measures besides the Jaccard coefficient. For example, we have considered the phi coefficient, which is the Pearson correlation applied to binary measurements [16, 17]. The main benefit of the phi coefficient is that it is a centered measure and hence kernel centering of the resulting configuration is not required. A minor disadvantage of the phi coefficient is that it is not defined in cases where objects consist of only zeroes or only ones. This can be easily circumvented however, for example by defining phi(x,y) = 0 in these cases. The main disadvantage of the phi coefficient lies in its definition of similarity: for the phi coefficient, both coinciding zeroes and ones contribute towards similarity, whereas for the Jaccard similarity only coinciding ones do. We believe objects are similar when they share the same mutations (rather than the absence of mutations) and hence prefer the Jaccard similarity here.

In future work, the RV coefficient could be further extended for other types of data. For example, a matrix with ordinal data could be converted into a configuration matrix using the Spearman rank correlation or the r_{OZ} coefficient similarity [15, 17]. Additionally, other semi-positive definite kernels that describe the similarity between objects could be used as a configuration matrix. For example, if we were to consider a dataset that is represented as a graph (where each node corresponds to an object), then a configuration matrix could be constructed using a graph diffusion kernel [6]. Finally, as many multi-omics data contain patient survival data, defining a configuration matrix for survival data opens up interesting avenues for future research. For each of these extensions, careful assessment of the need of kernel centering will be required.

We believe that iTOP can be applied to a broad range of data, beyond the pharmacogenomics data analyzed here. Essentially, for all data in which the same objects have been characterized in multiple modalities, this methodology can be used to infer a topology of relationships between the resulting datasets. Hence, as multi-omics and phenotypic data is collected for increasingly more experiments, we believe our methodology will be highly relevant and widely applicable.

Funding

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC synergy grant agreement n° 319661 COMBATCANCER.

3.5. Supplementary Materials

3.5.1. The modified RV coefficient

For data matrices $\mathbf{X}$ where the number of variables is much greater than the number of objects (i.e. $p \gg n$), the RV coefficient is known to be biased upwards [9, 11]. To account

for this bias, we remove the diagonal of the configuration matrix, as in the modified RV coefficient [11].

$$\tilde{\mathbf{S}}_i = \mathbf{S}_i - diag(\mathbf{S}_i)$$
$$\tilde{\mathbf{S}}_j = \mathbf{S}_j - diag(\mathbf{S}_j)$$
$$RV(\tilde{\mathbf{S}}_i, \tilde{\mathbf{S}}_j) = \frac{Vec(\tilde{\mathbf{S}}_i)^T Vec(\tilde{\mathbf{S}}_j)}{\sqrt{Vec(\tilde{\mathbf{S}}_i)^T Vec(\tilde{\mathbf{S}}_i) \times Vec(\tilde{\mathbf{S}}_j)^T Vec(\tilde{\mathbf{S}}_j)}}$$

We note that for the modified RV coefficient, the average of $Vec(\tilde{\mathbf{S}})$ is not zero. This means that $RV(\tilde{\mathbf{S}}_i, \tilde{\mathbf{S}}_j)$ is actually not equal to the correlation (but rather to the congruence) between $Vec(\tilde{\mathbf{S}}_i)$ and $Vec(\tilde{\mathbf{S}}_j)$. Regardless, for simplicity, we do describe the RV coefficient in terms of the correlation between $Vec(\tilde{\mathbf{S}}_i)$ and $Vec(\tilde{\mathbf{S}}_j)$ in the introduction and the first results subsection.

Mayer et al. (2011) [9] have reported that the modified RV coefficient does not correct all of the abovementioned $p \gg n$ bias. They propose the adjusted RV coefficient, based on the adjusted r^2 measure. However, the adjusted RV coefficient requires the data to be column-wise centered and autoscaled (i.e. scaled such that each column has a standard deviation of one). As we have shown in the Methods and Materials of the main text, binary datasets can be centered by kernel centering the configuration matrix (essentially using a set of linear transformation to center the kernel space (corresponding to $\mathbf{S}$) rather than the input space ($\mathbf{X}$)). However, a similar approach cannot be taken with autoscaling, because determining the standard deviation (by which each column needs to be scaled) is a non-linear operation and hence cannot be performed in kernel space. Similarly, the adjusted RV coefficient requires one to take the adjusted r^2 between columns in the input space, which is also a non-linear operation that hence cannot be performed in kernel space. Finally, the benefit of the adjusted RV coefficient over de modified RV coefficient is extremely small when using a sufficient number of objects (e.g. $n > 50$) [9]. Therefore, we prefer to use the modified RV coefficient, which does not have the aforementioned limitations, while practically correcting the same amount of bias.

3.5.2. PARTIAL MANTEL TEST

The concept of partial matrix correlations has been explored previously by Smouse et al. (1986) [12], who based their measure on the Mantel Test [8]. The Mantel test essentially measures the correlation on the vectorized form of the distance matrices (rather than configuration matrices) corresponding to $\mathbf{X}_1$ and $\mathbf{X}_2$. We prefer to base the partial matrix correlation on the RV coefficient instead because of two disadvantages of the Mantel Test. First, the Mantel Test does not necessarily result in a correlation close to zero for orthogonal data, while the RV coefficient does. Second, the Mantel Test always results in high matrix correlations when applied to high-dimensional matrices. While the original RV coefficient also suffers from the second limitation, the modified RV coefficient [11] alleviates this problem. Notably, this modification does not alleviate the problem for the Mantel Test. While both issues do not affect significance estimates resulting from a permutation test, they greatly affect the interpretation of the coefficients. Hence, we prefer to base our work on the RV coefficient rather than the Mantel Test.

3.5.3. PC ALGORITHM

We used the order-independent PC algorithm proposed by Colombo and Maathuis (2014) [3], that was implemented in the R package pcalg. This algorithm uses partial correlations to infer a topology between variables (or in our work: partial matrix correlations to infer a topology between datasets). After inferring the topology, the PC algorithm can also attempt to infer causality between nodes in the topology, using two additional assumptions: 1) the causality graph underlying the data is a DAG (Directed Acyclic Graph); and 2) all variables are observed (or in our work: there are no hidden / unobserved datasets). It is important to keep these assumptions in mind when interpreting causality inferred by the PC algorithm.

3.5.4. ELASTIC NET REGRESSION

We used Elastic Net regression [18] as implemented in the R package glmnet, with λ set to λ_{min} and α set to 0.5. Predictive performance was assessed by using nested cross-validation, as implemented in the R package TANDEM, where the inner cross-validation loop was used to optimize the λ parameters for each stage, and the outer cross-validation loop was used to determine the predictive performance.

3.5.5. TANDEM

TANDEM [1] is a variable selection method that prioritizes variables selection from certain datasets over others. Consider a response vector $\mathbf{y}$ (e.g. drug response of a single drug) and two datasets $\mathbf{X}_1$ and $\mathbf{X}_2$. TANDEM performs the variable selection in two stages. In the first stage, Elastic Net regression [18] is used to explain as much of $\mathbf{y}$ as possible using $\mathbf{X}_1$. In the second stage, Elastic Net regression is used to explain the residuals from the first stage (i.e. the part of $\mathbf{y}$ that could not be explained using $\mathbf{X}_1$) using $\mathbf{X}_2$.

We used the implementation from the R package TANDEM, with λ set to λ_{min} for both stages and α set to 0.5. Predictive performance was assessed by using nested cross-validation, where the inner cross-validation loop was used to optimize the λ parameters for each stage, and the outer cross-validation loop was used to determine the predictive performance.

The relative contribution of a dataset was determined by dividing the sum-of-squares of the prediction from one dataset divided by the sum-of-squares of the overall prediction. For more information, we refer to Aben et al. (2016) [1].

We determined the variable importance VI of variable j in the same way as in our previous work on TANDEM [1], using:

$$VI = \frac{\left\|\mathbf{x}_j \boldsymbol{\beta}\right\|_2^2}{\left\|\mathbf{X}\boldsymbol{\beta}\right\|_2^2}$$

Where $\mathbf{X}$ is the input matrix for TANDEM, defined as $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$; $\mathbf{x}_j$ is the j'th variable of $\mathbf{X}$; and $\boldsymbol{\beta}$ is the regression coefficients estimated by TANDEM.

3.6. SUPPLEMENTARY FIGURES

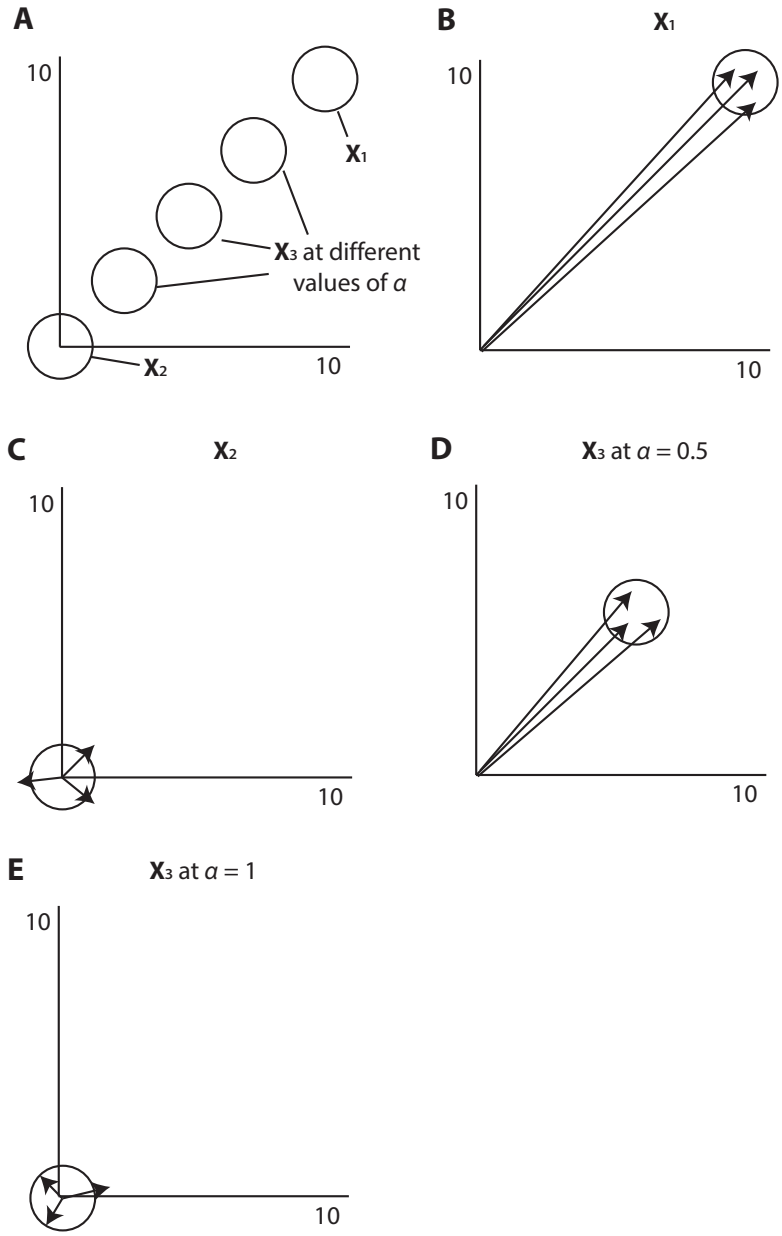
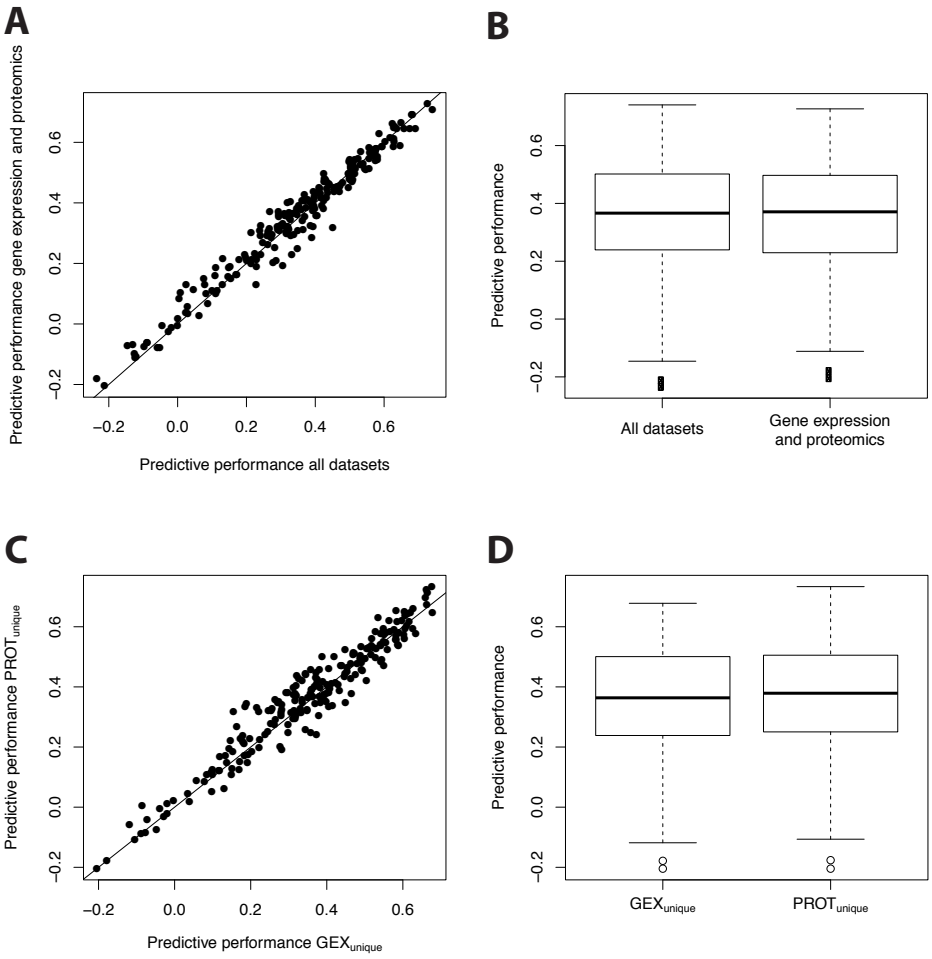


Figure S3.1: Illustration accompanying Figure 4A. (A) Cartoon of the densities of $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ in a two-dimensional space. (B-E) Cartoon of the directions of the inner products between objects from (B) $\mathbf{X}_1$, (C) $\mathbf{X}_2$, (D) $\mathbf{X}_3$ at $\alpha = 0.5$, and (E) $\mathbf{X}_3$ at $\alpha = 1$.


A
Predictive performance gene expression and proteomics
−0.2
0.0
0.2
0.4
0.6
Predictive performance all datasets
B
Predictive performance
All datasets
Gene expression and proteomics
C
Predictive performance PROT$_{unique}$
Predictive performance GEX$_{unique}$
D
Predictive performance
GEX$_{unique}$
PROT$_{unique}$

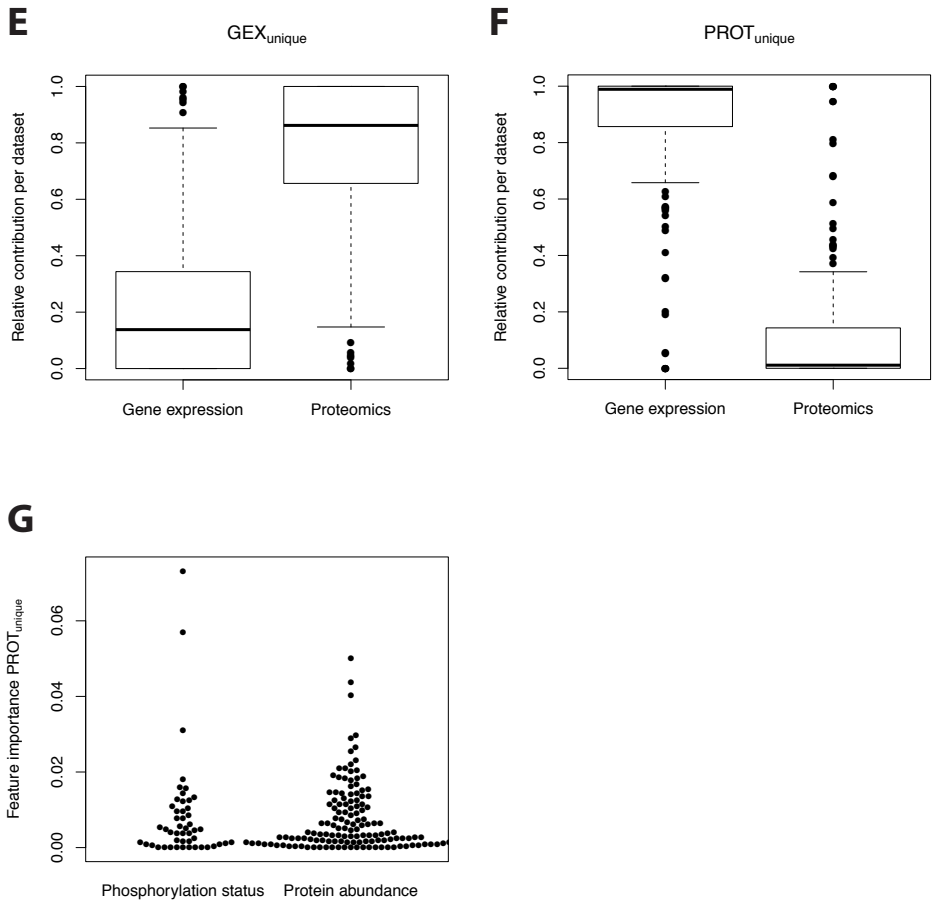


Figure S3.2: Drug response prediction models. (A) Predictive performance (Pearson correlation between observed and predicted drug response) of either a model trained on all datasets except drug response (i.e. mutation, CNA, methylation, cancer type, gene expression and proteomics), or a model trained on on gene expression and proteomics only, for each of the 217 drugs. (B) Predictive performance (Pearson correlation between observed and predicted drug response) of GEX_{unique} vs. $PROT_{unique}$ models for each of the 217 drugs. (C&D) Distribution of relative contributions of gene expression and proteomics in GEX_{unique} and $PROT_{unique}$ models respectively, across all 217 drugs. (E) variable importance for $PROT_{unique}$ models (averaged across drugs) for two classes of variables in the proteomics data: phosphorylation status and protein abundance.

REFERENCES

[1] Nanne Aben, Daniel J Vis, Magali Michaut, and Lodewyk FA Wessels. Tandem: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics*, 32(17):i413–i420, 2016.

[2] Nanne Aben, Johan A Westerhuis, Yipeng Song, Henk AL Kiers, Magali Michaut, Age K Smilde, and Lodewyk FA Wessels. itop: Inferring the topology of omics data. *bioRxiv*, page 293993, 2018.

[3] Diego Colombo and Marloes H Maathuis. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

[4] Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, Chris D Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I Richard Thompson, Xi Luo, Jorge Soares, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570, 2012.

[5] Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, 2016.

[6] Risi Imre Kondor and John Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *ICML*, volume 2, pages 315–322, 2002.

[7] Jun Li, Wei Zhao, Rehan Akbani, Wenbin Liu, Zhenlin Ju, Shiyun Ling, Christopher P Vellano, Paul Roebuck, Qinghua Yu, A Karina Eterovic, et al. Characterization of human cancer cell lines by reverse-phase protein arrays. *Cancer cell*, 31(2):225–239, 2017.

[8] Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1):209–220, 1967.

[9] Claus-Dieter Mayer, Julie Lorent, and Graham W Horgan. Exploratory analysis of multiple omics datasets using the adjusted rv coefficient. *Statistical applications in genetics and molecular biology*, 10(1), 2011.

[10] Paul Robert and Yves Escoufier. A unifying tool for linear multivariate statistical methods: the rv-coefficient. *Applied statistics*, pages 257–265, 1976.

[11] Age K Smilde, Henk AL Kiers, S Bijlsma, CM Rubingh, and MJ Van Erk. Matrix correlations for high-dimensional data: the modified rv-coefficient. *Bioinformatics*, 25(3):401–405, 2008.

[12] Peter E Smouse, Jeffrey C Long, and Robert R Sokal. Multiple regression and correlation extensions of the mantel test of matrix correspondence. *Systematic zoology*, 35(4):627–632, 1986.

[13] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[14] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[15] Jan Vegelius. On generalizations of the g index. *Educational and Psychological Measurement*, 36(3):595–600, 1976.

[16] G Udny Yule. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, 75(6):579–652, 1912.

[17] Frits Eduard Zegers. *A general family of association coefficients*. Boekhandel Boomker, 1986.

[18] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–320.

4

IDENTIFYING BIOMARKERS OF ANTI-CANCER DRUG SYNERGY USING MULTI-TASK LEARNING

NANNE ABEN, JULIAN R. DE RUITER, EVERT BOSDRIESZ, YONGSOO KIM, GERGANA BOUNOVA, DANIEL J. VIS, LODEWYK F.A. WESSELS, MAGALI MICHAUT

Parts of this chapter have been submitted to bioRxiv (https://doi.org/10.1101/243568) [1].

ABSTRACT

Motivation: *In biology, we are often faced with multiple datasets recorded on the same set of objects, such as multi-omics and phenotypic data of the same tumors. These datasets are typically not independent from each other. For example, methylation may influence gene expression, which may, in turn, influence drug response. Such relationships can strongly affect analyses performed on the data, as we have previously shown for the identification of biomarkers of drug response. Therefore, it is important to be able to chart the relationships between datasets.*

Results: *We found that methods widely used in single drug response prediction, such as Elastic Net regression per drug, are not predictive in this setting. Instead, we propose a multi-task learning approach: training a single model simultaneously on all drug combinations, which we show results in increased predictive performance. In contrast to other multi-task learning approaches, our approach allows for the identification of biomarkers, by using a modified random forest variable importance score, which we illustrate using artificial and DREAM challenge data.*

Availability: *A Python implementation of our approach is available on Github. (https://github.com/NKI-CCB/multitask_vi/)*

Contact: *l.wessels@nki.nl and m.michaut@nki.nl*

4.1. INTRODUCTION

Combining drugs is a promising strategy for cancer treatment, as drug combinations can increase the efficacy of treatment. For example, Prahallad *et al.* (2012) [21] have shown that combining a BRAF inhibitor with an EGFR inhibitor shows synergy in *BRAF* mutant colorectal cancer. However, for most drug combinations it is not known what subset of patients will respond. By identifying biomarkers (e.g. mutations in the tumor's DNA that are associated with a favorable response to the drug combination), the selection of a given patient's treatment can be improved. To facilitate biomarker identification, data from a large-scale drug combinations screen were recently released as part of the AstraZeneca-Sanger DREAM challenge [18], containing 85 cell lines with their response to 167 drug combinations.

While the data from this screen can provide information on potential biomarkers of synergy, it is not yet clear what is the best way to identify them. In the context of single drug response prediction, the default approach is to fit 'individual models' that are trained separately per drug. We applied a similar approach here in the context of drug combinations, training 'individual models' for each drug combination separately (Fig. 4.1A). However, we show that such an approach is unsuitable for the dataset at hand due to the extremely low sample size: a median number of 14 cell lines have been screened per drug combination.

We propose to alleviate the problem of low sample size by training 'joint models' that use information from all drug combinations simultaneously (Fig. 4.1B). In the literature, this is known as multi-task learning [5, 19]. This approach has been employed before in single drug response prediction by Gönen *et al.* (2014) [10], Menden *et al.* (2013) [17] and Yuan *et al.* (2016) [26], and in synergy prediction by other participants in

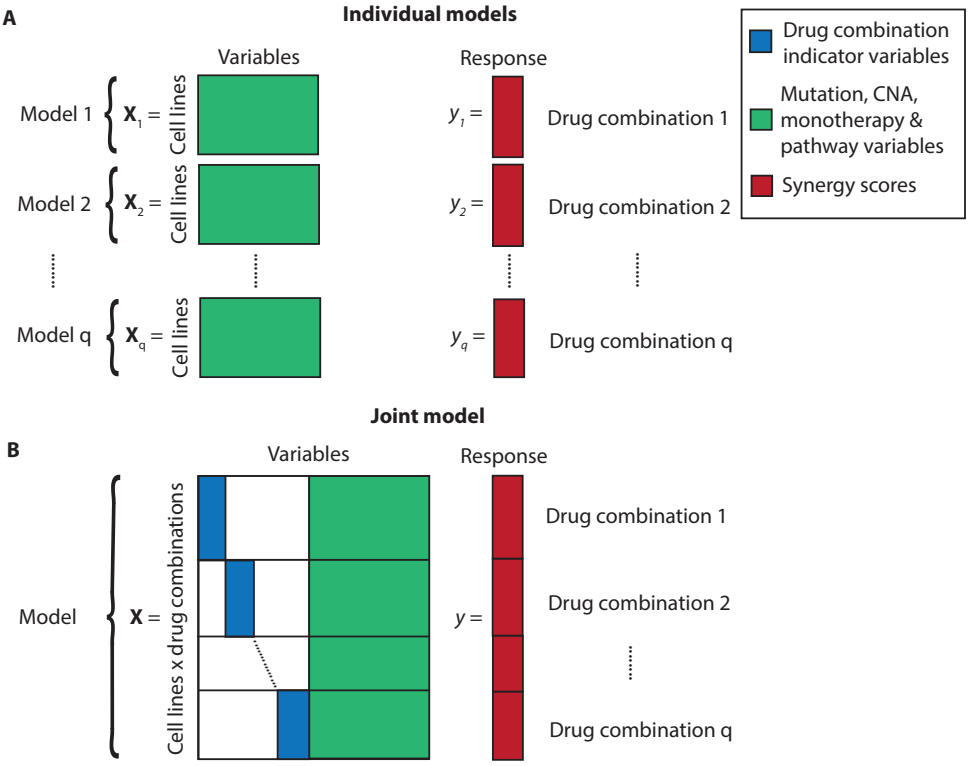


Figure 4.1: Overview of the models used in this work. (A) Graphical representation of q individual models, in which a different model is trained independently for each of the q drug combinations. (B) Graphical representation of the joint model, in which a single model is jointly trained on all q drug combinations simultaneously.

the AstraZeneca-Sanger DREAM challenge [18]. What distinguishes our approach from other multi-task learning approaches is that we are able to identify biomarkers, whereas others have proposed black-box models. Specifically, the Joint Random Forest model we propose is simultaneously trained on all drug combinations, after which we apply our Drug-combination-specific Variable Importance (DVI) score to the trained Joint Random Forest to identify biomarkers of synergy.

We show that the joint model outperforms individual models in terms of predictive performance. Using the joint model together with the DVI, we are able to identify biomarkers of response on both simulated and real data. Finally, we found that *MYO15A* mutations associate with synergy between an ALK / IGFR dual inhibitor and PI3K pathway inhibitors in triple-negative breast cancer.

4.2. METHODS & MATERIALS

4.2.1. INDIVIDUAL AND JOINT PREDICTION MODELS

Predictive models are typically trained per drug combination (Fig. 4.1A). We refer to these models as 'individual models'. In this work, we propose a 'joint model', which is si-

multaneously trained on data from all drug combinations (Fig. 4.1B). Such an approach can be viewed as multi-task learning, where each drug combination represents a task.

The joint model takes an augmented matrix **X** as input, in which each sample represents a cell line, drug combination pair. We used an indicator variable to code the different drug combinations. The remaining variables are either:

1. Private to a cell line, drug combination pair (i.e. monotherapy and pathway rules), and hence unique to every sample in **X**.

2. Private to a cell line only (i.e. mutation and CNA data), and hence repeated across drug combinations.

These two categories are visualized in Supplementary Fig. 4.6 in purple and green respectively.

The response vector y was defined as the concatenation of the response values, such that each sample corresponds to a cell line, drug combination pair. The resulting input data **X** can be fitted onto y using standard machine learning algorithms. In this work, we have compared three different algorithms:

- Elastic Net [27], as implemented in the R package glmnet [7], with α set to 0.5 and λ optimized in a nested cross-validation loop.

- SVM [4] with RBF kernels, as implemented in the Python package scikit-learn [20], optimizing the hyper-parameters C over [50, 100, 200, 300] and γ over [0.001, 0.0001, 0.0005, 0.00005, 0.00001] in a nested cross-validation loop.

- Random Forest [12], as implemented in the Python package scikit-learn, using default parameters.

We compared these joint models to 'individual models', which are trained per drug combination and contain the same variables, except the drug combination indicator variables (which are constant within a given drug combination). Predictive performance was assessed using 2-fold cross-validation with the 'primary score' (a weighted average of the correlation between the observed and predicted synergy scores) as defined in the AstraZeneca-Sanger DREAM challenge [18] as endpoint. The exact definition of the primary score is

$$\frac{\sum_{i=1}^{q} \sqrt{n_i - 1}\, r(y_i, \hat{y}_i)}{\sum_{i=1}^{q} \sqrt{n_i - 1}} \tag{4.1}$$

where q is the number of drug combinations, n_i the number of cell lines in drug combination i, r a function that computes the Pearson correlation, y_i the synergy scores for drug combination i and $\hat{y}_i$ the predicted synergy scores for drug combination i. For a fair comparison, all different models (individual and joint; Elastic Net, SVM and Random Forest) were tested using the same cross-validation folds.

4.2.2. VARIABLE IMPORTANCE MEASURES

We found that the Random Forest obtained the best predictive performance and hence decided to use this model for biomarker identification. To this end, we compared three Random Forest variable importance measures (Fig. 4.2).

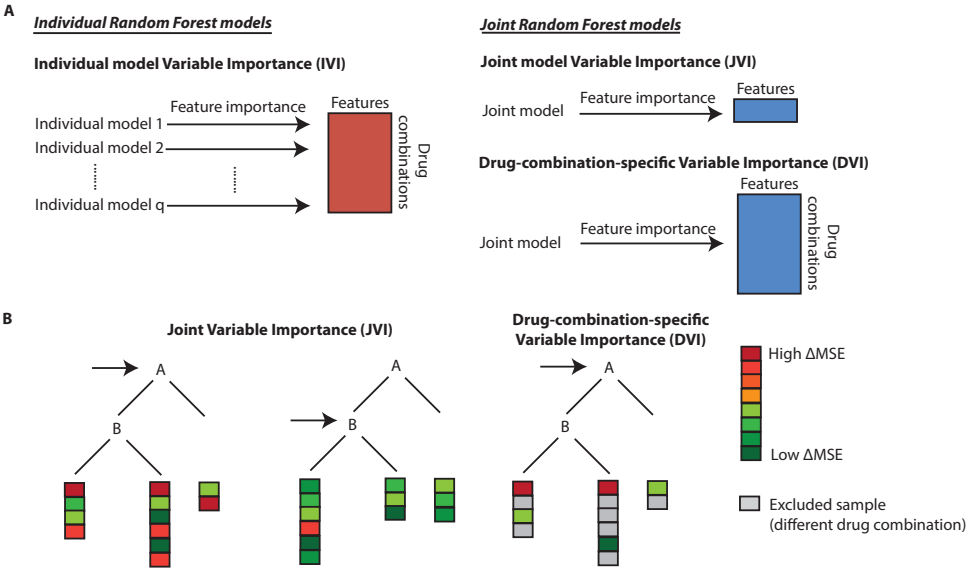


Figure 4.2: Overview of variable importance scores used in this work. A: Illustration of variable importance scores. Applying the variable importance score normally used with Random Forests to the Joint Random Forest model yields the Joint model Variable Importance (JVI), a measure of variable importance across all drug combinations. Applying the variable importance score normally used with Random Forests to Individual Random Forest models yields the Individual model Variable Importance (IVI), a measure of variable importance per variable and per drug combination. In order to obtain a variable importance for each drug combination using the Joint Random Forest model, we propose the Drug-combination-specific Variable Importance (DVI). B: Illustration of the JVI and the DVI in a single decision tree from the random forest. For both variable importance scores, the importance is assessed by permuting the values of the given variable (a permuted variable is indicated by a horizontal arrow here) and then calculating for each sample (a sample is indicated by a box at the bottom of the tree) the difference between the permuted and unpermuted errors. In the given example, variable A is more important than variable B, as indicated by the higher difference in error (ΔMSE) when permuting variable A.

In all three variable importance measures, a variable is considered important if it has a positive effect on the predictive performance. More specifically, the importance of a variable $\mathbf{X}_j$ for a given tree in the forest is evaluated by calculating the prediction accuracy of the tree on out-of-bag (OOB) samples before and after permuting the values of variable $\mathbf{X}_j$. The absolute difference between the two accuracy values is then the permuted variable importance for the given tree. These steps are repeated for every tree in the Random Forest, after which the resulting scores are averaged, resulting in the permuted variable importance for the whole Random Forest.

Applying the permutated variable importance to individual Random Forest models results in the Individual model Variable Importance (IVI), which gives us variable importance scores per variable and per drug combination. Applying the permutated variable

importance joint Random Forest models results in the Joint model Variable Importance (JVI), which gives us variable importance scores per variable, but not per drug combination. Though useful for comparisons, these two variable importance scores have two important caveats: i) the IVI is based on individual models, which we show have lower predictive performance than the joint models; and ii) the JVI uses the joint model, but the resulting variable importances cannot be traced back to a specific drug combination.

For these reasons, we propose the Drug-combination-specific Variable Importance (DVI), which leverages the performance of the joint models while maintaining drug combination specificity. The DVI uses a modified version of the permuted variable importance, in which the prediction accuracy is calculated using only the samples that correspond to the combination of interest. More specifically, this accuracy is calculated using a weighted mean square error, in which OOB samples belonging to the combination have a weight of 1 and all other samples have a weight of 0. A Python implementation of the DVI is available at https://github.com/NKI-CCB/multitask_vi/.

4.2.3. THE ASTRAZENECA-SANGER DREAM CHALLENGE DATA

In order to predict synergy from molecular data, we have used the data from the AstraZeneca-Sanger DREAM challenge (from here on referred to as: DREAM data) [18]. The goal of this community challenge was to create models that predict whether a given drug combination will show synergy in certain cell lines. The DREAM data include 85 cell lines and 167 drug combinations, with a median of 14 cell lines screened per drug combination. The dataset consists of three parts: synergy scores, monotherapy response data and molecular data of the cell lines (e.g. mutations and copy number alteration data).

As the response variable for our model, we used the synergy scores as provided in the DREAM data, which were based on a Loewe additivity model [6, 9]. For each cell line, drug combination pair, monotherapy data were available, quantifying the response of a cell line to each individual drug in the drug combination by the 50% Inhibitory Concentration (IC50) or the Area Under the dose-response Curve (AUC). For each cell line, molecular data were provided in the form of mutation, copy number alteration (CNA), methylation and gene expression data. Because of the high dimensionality and the low sample size, we restricted mutations and CNAs to a reduced set of potential driver genes. Finally, we defined 'pathway rules' that integrate the mutation and CNA data with information from KEGG [14, 15]. More information on how these data were processed is provided in the Supplementary Materials.

We used the monotherapy and the molecular data of the cell lines to predict drug synergy (Fig. 4.1). More formally, we defined the input matrix $\mathbf{X}$ using 382 mutation, 76 copy number, 23 monotherapy, and 16 pathway rule variables. The response vector y was defined using the synergy scores. Each of the input data types explain a part of the synergy and are therefore useful to include in a predictive model. For biomarker identification we focused on genomic variables only, as monotherapy data are unlikely to be useful as clinical biomarkers (this information is typically not available for most drugs for a given patient).

4.3. RESULTS

4.3.1. PER-COMBINATION INDIVIDUAL MODELS PERFORM POORLY

For our initial approach, we used the DREAM data to create 'individual models' that are trained separately per drug combination (Fig. 4.1A). To test the variability across different prediction methods, the individual models were trained using either Elastic Net, SVM (with RBF kernels) or Random Forest. For each method, predictive performance was assessed using cross-validation with the 'primary score' (a weighted average of the correlation between the observed and predicted synergy scores) defined in the DREAM challenge [18] as endpoint.

Overall, the predictive performance of the individual models was low for all methods (0.04 on average) (Fig. 4.3A), most likely due to the extremely low sample size (median of 14 cell lines per combination). We also observed that the predictions from the individual SVM models resulted in negative correlations between the observed and predicted synergy scores (Fig. 4.3A). This is due to a cross-validation artifact that leads to negative correlations when the model is unable to detect structure in the data (Supplementary Materials), which likely mostly affected the SVM due to the high complexity of the RBF kernels.

4.3.2. SIMULTANEOUSLY LEARNING ACROSS DRUG COMBINATIONS IMPROVES PREDICTIVE PERFORMANCE

To alleviate the low sample size problem, we created 'joint models', which are trained on all drug combinations simultaneously, thereby leveraging the information from the entire dataset. Using cross-validation, we found that the joint models achieve higher predictive performance compared to individual models (Fig. 4.3A), regardless of the underlying method (Elastic Net, SVM, Random Forest).

A drawback of the standard cross-validation scheme is that the same cell line can be in different cross-validation folds (but for different drug combinations), which could bias the predictive performance. To test for this, we also performed leave-one-cell-line-out cross-validation, in which all data associated with a given cell line were left out from the training step of a given fold. Overall, we found that joint models were more predictive than individual ones using leave-one-cell-line-out cross-validation too (Fig. 4.3B), ruling out this bias. We also observed that the individual Random Forest models resulted in negative predictive performance in this setting (Fig. 4.3B), whereas the predictive performance was positive using regular cross-validation (Fig. 4.3A). This too can be attributed to the aforementioned cross-validation artifact (Supplementary Materials).

To determine whether the joint models were predictive for specific classes of drug, we grouped the 119 drugs into 19 drug classes and checked whether the difference in predictive performance between the individual or joint models was associated with any of the drug classes. This showed that drug combinations containing IGFR inhibitors are significantly better predicted using the joint model (Mann-Whitney U test, FDR-corrected $p = 0.047$) (Fig. 4.3C). Furthermore, for drug combinations containing DNA damaging agents (DDA), the joint model showed on average no increase in predictive performance (Fig. 4.3D, bottom panel). Compared to the overall increase in predictive performance between individual and joint models, this effect was significant (Mann-Whitney U test,

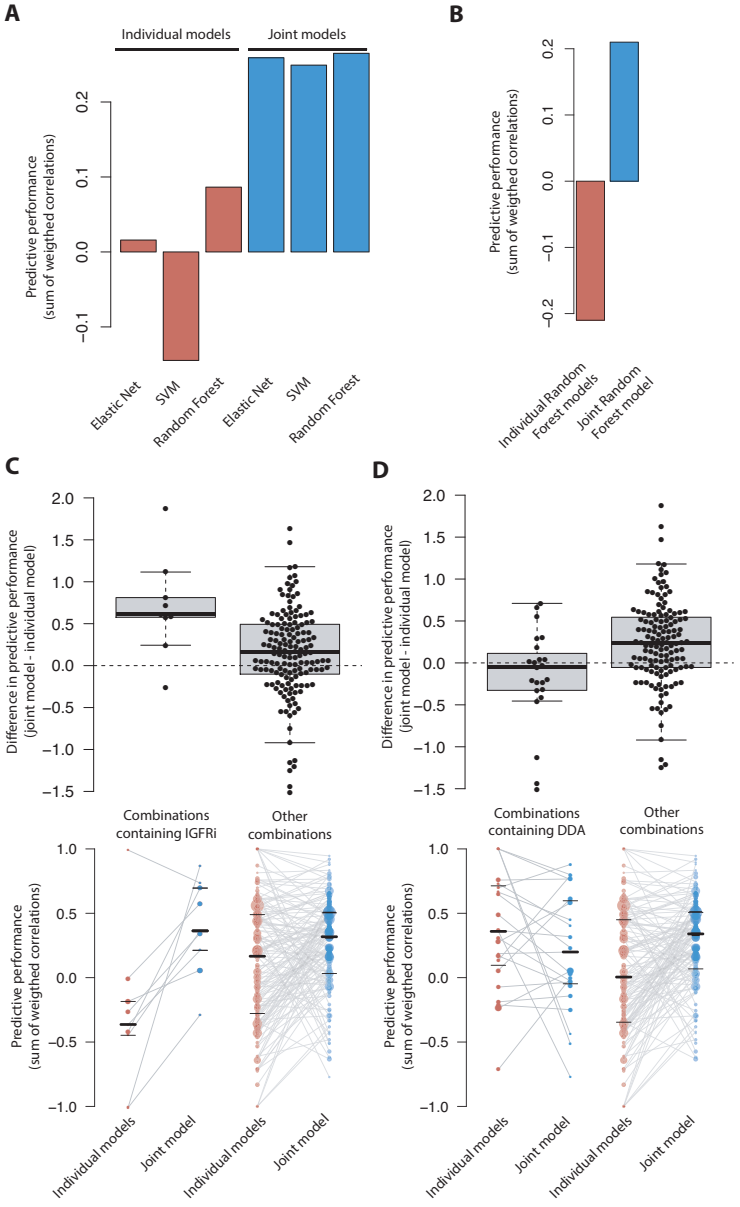


Figure 4.3: Predictive performance of the joint and individual models. Performance was measured using the 'primary score' defined in the DREAM challenge [18], i.e. the weighted sum of correlations between the observed and predicted synergy scores. (A) Predictive performance stratified by method (Elastic Net, SVM, Random Forest) and model (individual, joint). (B) Predictive performance for individual and Joint Random Forest models, assessed using leave-one-cell-line-out cross-validation. (C&D) Association between a specific drug class and the difference in predictive performance between individual and joint models (top panel). How the predictive performance changed between the individual and joint models is illustrated in the bottom panel. Each dot represents the predictions for a given drug combination. Predictions for the same drug combination are connected between the individual and the joint models. The size of the dot is proportional to the number of cell lines the model was trained on. From left to right: combinations containing IGFR inhibitors (IGFRi), all other drug combinations (not containing IGFRi), combinations containing DNA damaging agents (DDA), and all other combinations (not containing DDA). Note that the extreme correlations (i.e. correlations close to 1 or -1) can be attributed to small sample size (indicated here by the size of the dot in the bottom panel).

FDR-corrected $p = 0.036$) (Fig. 4.3D, top panel).

To further characterize the joint model predictive performance improvement, we used a simulated dataset and assessed under which conditions joint models outperform individual models using Random Forests. In this simulation, we created a data set of similar size as the DREAM data and then varied the sample size or the number of features (Supplementary Materials). We found that simultaneously learning across drug combinations was most beneficial in highly underdetermined cases, i.e. when the sample size was low or the number of variables was high (Supplementary Fig. S4.2). Interestingly, when the number of samples was sufficiently high (e.g. $n = 100$), the individual and joint Random Forest models achieved virtually identical predictive performance.

Altogether, our results show that, for most combinations, joint models obtain a higher predictive performance compared to individual models by simultaneously learning across drug combinations. As the joint Random Forest model obtained the highest predictive performance, we decided to further use this joint model for biomarker identification.

4.3.3. JOINT MODEL VARIABLE IMPORTANCE SCORES ARE NOT SUFFICIENT TO IDENTIFY BIOMARKERS OF SYNERGY

In an initial attempt to identify biomarkers of synergy using the joint model, we first computed the Random Forest's variable importance score (VI), referred to as the joint model VI score (JVI) (Fig. 4.2A). Ranking the variables by their JVI, we identified variables that had a large impact on the prediction of many different drug combinations. We found that the monotherapy variables were the most important variables overall (highest JVI scores in Supplementary Fig. 4.6) (one-tailed Mann-Whitney U test, $p = 2.474e\text{-}16$), followed by pathway rules (one-tailed Mann-Whitney U test, $p = 6.248e\text{-}06$) (Supplementary Materials).

A major drawback of the JVI is that the associations cannot be traced back to specific drug combinations, limiting its use for finding biomarkers of synergy for specific drug combinations. For example, the highest-ranked molecular data variable was mutations in *ATAD5*, which we are unable to link to a specific drug combination using the JVI. To identify biomarkers, we needed a drug-combination-specific variable importance score. Although this could be achieved by computing Random Forest VI scores for the individual models, referred to as Individual model VI score (IVI) (Fig. 4.2A), we preferred to do this using the Joint Random Forest model because of its superior predictive performance. Thus, we needed to define a measure of variable importance per drug combination and per variable using the Joint Random Forest model.

4.3.4. DRUG-COMBINATION-SPECIFIC VARIABLE IMPORTANCE IDENTIFIES BIOMARKERS OF SYNERGY

To identify biomarkers for a specific drug combination using the Joint Random Forest model, we developed a Drug-combination-specific Variable Importance score (DVI) (Fig. 4.2A). The DVI determines the contribution of each variable to the prediction in the same way as the original Random Forest VI score, but only considers the samples from one drug combination at a time (Fig. 4.2B). To evaluate the DVI, we created a simulated dataset in which we engineered a biomarker with two parameters: 1) e: the effect size of the association of the biomarker with synergy; and 2) d: the number of drug combinations for which this biomarker was engineered to be associated with synergy (Supple-

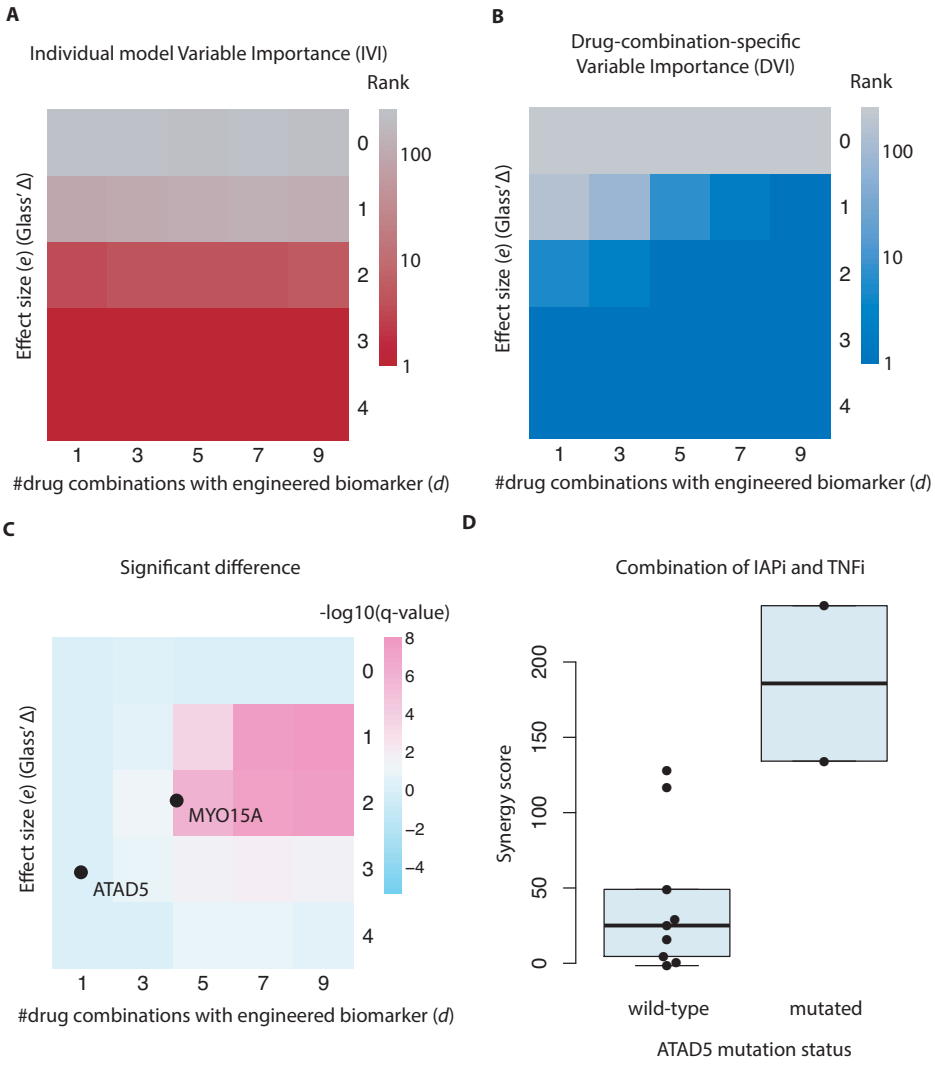


Figure 4.4: Associations using the IVI and the DVI. Associations using the Individual model Variable Importance (IVI) and the Drug-combination-specific Variable Importance (DVI) (A&B) Heatmap showing the median rank of the engineered biomarker in the simulated dataset, stratified by effect size (e) and number of drug combinations for which the biomarker was engineered to be associated with synergy (d), using either (A) the IVI or (B) the DVI. (C) Heatmap showing for which e (effect size) and which d (number of drug combinations for which the biomarker was engineered to be associated with synergy) the DVI is significantly better (indicated in pink) than the IVI at retrieving the association in a simulated dataset. Examples used in this paper from the DREAM data (associations with *MYO15A* and *ATAD5*) are indicated in this plot based on their effect size and the number of drug combinations in which we observe them. (D) Synergy score for a combination of an IAPi and an TNFi, stratified by *ATAD5* mutation status.

mentary Materials). As expected, increasing either one of these enhances the ability of the DVI to identify the biomarker (Fig. 4.4B).

We then used the simulated dataset to compare the DVI to the Individual model Variable Importance (IVI, Random Forest VI score on individual models). This showed that the ability of the IVI to identify the engineered biomarker is correlated with the effect size (e), but not with the number of drug combinations (d) (Fig. 4.4A). This is expected, since the Individual Random Forest models (underlying the IVI scores) do not share information across different drug combinations. Hence, increasing d has no effect on the model's ability to recover the biomarker. Interestingly, we found that biomarkers with a sufficiently large effect size are identified by both the IVI and the DVI. We found that the DVI is significantly better than the IVI at identifying the biomarker in scenarios where e is small and d is high (Fig. 4.4C).

These findings were reflected in the DREAM data. For example, ranking the associations by their DVI, the highest-ranking molecular data variable was the association of *ATAD5* mutation status with synergy between IAP inhibitors and TNF inhibitors (Fig. 4.4D). As *ATAD5*, IAP and TNF are all part of the apoptosis pathway, this illustrates that the DVI is able to identify interesting associations. Given the large effect size, it is not surprising that this association is ranked high for this drug combination by both the DVI (ranked #3) and the IVI (ranked #1). Using the effect size from this association (Glass' $\Delta = 2.8$) and assuming that the biomarker is not shared with any other drug combinations, we related this example to the simulated data and found that this example falls in the region where the DVI has little added value (Fig. 4.4C).

In addition, we identified a biomarker (*MYO15A* mutations) that is exclusively identified by the DVI. Given that we observed this association in four related drug combinations and an average Glass' Δ of 0.87, this example indeed falls in the region where the DVI improves over the IVI (Fig. 4.4C).

4.3.5. MYO15A MUTATIONS ASSOCIATE WITH SYNERGY BETWEEN AN ALK / IGFR DUAL INHIBITOR AND PI3K PATHWAY INHIBITORS IN TRIPLE-NEGATIVE BREAST CANCER

We set out to identify biomarkers that were exclusively identified by the DVI. To this end, we decided to focus on the drug combinations containing IGFR inhibitors, for which the joint model obtained the largest increase in predictive performance over the individual models (Fig. 4.3C). Using the DVI to rank all non-monotherapy variables for each of these drug combinations, we found that *MYO15A* mutations had the highest average rank. The association between *MYO15A* mutations and synergy was strongest in combinations of an ALK / IGFR dual inhibitor with PI3K pathway inhibitors (two AKT inhibitors, one PIK3CB / PIK3CD inhibitor and one mTOR inhibitor), hence we decided to further focus on these. These combinations were tested in 23 breast cancer cell lines, of which 20 were triple negative.

The association between *MYO15A* and the synergy score was strongest in the combination containing the PIK3CB / PIK3CD inhibitor (Fig. 4.5A). For the other combinations, the effect was in the same direction and hence in support of this association. Even though these effects would not have been considered significant individually, the model leverages the information across the drug combinations.

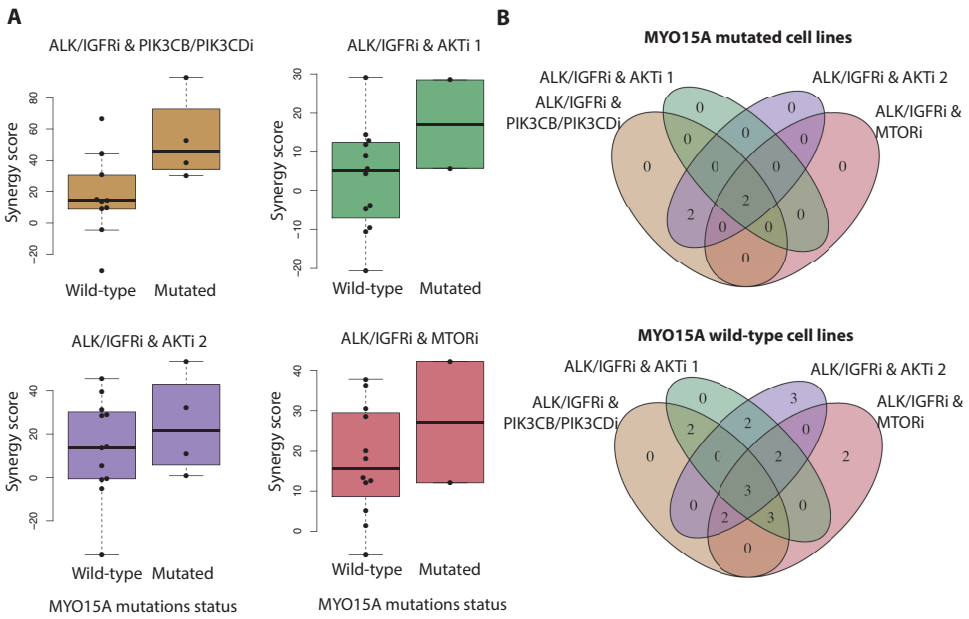


Figure 4.5: MYO15A mutations associate with synergy between ALK / IGFR dual inhibitors and PI3K pathway inhibitors. (A) Synergy scores for combinations of ALK / IGFR dual inhibitors and PI3K pathway inhibitors, stratified by *MYO15A* mutation status. (B) Venn diagrams illustrating the overlap in cell lines screened for these four drug combinations. Top panel: *MYO15A* mutant cell lines, bottom panel: *MYO15A* wild-type cell lines.

Of the cell lines in which these combinations have been screened, only five cell lines (two *MYO15A* mutant lines, three wild-type lines) were screened in all combinations; the remaining 18 cell lines (two mutant, 16 wild-type) were not (Fig. 4.5B). Hence, by combining these different combinations, the sample size is effectively increased to 23 cell lines. Altogether, this illustrates how the DVI can be used to identify biomarkers of synergy using the joint Random Forest model.

4.4. DISCUSSION

Drug combinations are of great interest in cancer care, as they can increase treatment efficacy. However, without specific biomarkers, it is difficult to predict which drug combinations will have a synergistic effect in a given patient. Most current approaches for identifying biomarkers of single drug response fit a separate model for each drug. We have shown that such an approach does not obtain good prediction performance for predicting synergy in the DREAM data, likely due to the low sample size. To alleviate this limitation, we used multi-task learning to leverage the information contained in several drug combinations. Compared to previous work [10, 17, 18, 26], our model has the advantage that it is not a 'black-box method' and hence can identify biomarkers.

In our models, we found that monotherapy data are important for predicting synergy. Recently, Gayvert *et al.* (2017) [8] have analyzed a similar drug synergy screen, in

which they report the same, but do not offer a rationale. We believe that the link between monotherapy and synergy could be attributed to both biological and technical reasons. A biological explanation may be that a small reduction in viability using monotherapy can be evidence of target engagement by the drug, which is required for synergy. On the other hand, the high variable importance of monotherapy can also be technical. When one of the drugs is very potent (e.g. kills 80% of the cells by itself), the expectation is that the combination will kill most cells even if the effect is only additive. Hence, detecting the difference between synergy and additivity would become very difficult in this scenario, as this difference may not exceed the noise level. We note that both scenarios are supported by the data: some drug combinations show positive correlation between monotherapy sensitivity and synergy (corresponding to the first scenario), while others show a negative correlation (supporting the second scenario) (Supplementary Materials).

Another interesting observation is that, on average, using a joint or individual Random Forest model leads to similar predictive performance for drug combinations containing DNA damaging agents. This may reflect previous observations that monotherapy response to DNA damaging agents is notoriously hard to predict [2, 22], which might also apply to synergy prediction.

An interesting extension of our method would be the inclusion of variables specific to single drugs or drug combinations, such as chemical structures. These variables could be encoded for each drug or drug combination, similar to the way information is currently encoded specifically for cell lines. While such variables would be correlated with the drug combination indicator variables used in the current model, they could contain additional information, for example when two drug combinations are chemically similar to each other. As chemical information was only available for a subset of the drugs in the DREAM data, we were unable to use this information efficiently. This could be investigated in future work.

In summary, we have presented a method that circumvents the problem of low sample sizes by combining information across drug combinations. In contrast to previous work, our method allows for the identification of biomarkers. With the large number of possible drug combinations, many future drug combination screens are likely to be performed in a small number of cell lines. We believe that our approach can aid to identify biomarkers specifically in such screens.

ACKNOWLEDGEMENTS

We would like to thank the AstraZeneca-Sanger DREAM challenge consortium, for organising the DREAM challenge and making the data available. We would like to thank Marlous Hoogstraat for subtyping the breast cancer cell lines.

FUNDING

This work was supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC synergy grant agreement n° 319661 COMBATCANCER; the Dutch Cancer Society and AACR (SU2C) Dream Team Award (SU2C-AACR-DT1213); and the Netherlands Organization for Scientific Research (NWO, grant Zenith 93512009).

4.5. SUPPLEMENTARY METHODS & MATERIALS

4.5.1. SUPPLEMENTARY METHODS

DATA DESCRIPTION

We used the data from the AstraZeneca-Sanger DREAM challenge [18], which were released as part of a community challenge to create models that predict whether certain drug combinations show synergy in some cell lines. These data were split over two challenges (challenge 1 and challenge 2) and were then split again in so-called training, leaderboard and test data. An extensive description of how these parts are defined is beyond the scope of this work, for more details we refer to [18]. For this work, it suffices to state that we used the training and leaderboard data (at the time of writing, the test data had not been publically released yet) from challenge 1. Of note, the drug names were not disclosed in the DREAM data. Instead, each drug is indicated by its target.

The DREAM data consists of 85 cell lines and 167 drug combinations. For 2790 cell line, drug combination pairs, synergy scores have been given. We removed data that failed QC, which left us with 2386 cell line, drug combination pairs. From these data, we defined mutation, copy number aberration (CNA), pathway and monotherapy variables. For each of the 2386 cell line, drug combination pairs, for both drugs in the combination, we used the following monotherapy data:

- Relative cell viabilities at five different concentrations;
- Curve fitting parameters (EC50, Emax and slope) as provided in the DREAM data;
- Additional curve fitting parameters (IC50, AUC) as determined using the algorithm by Vis *et al.* (2016) [24]; and
- Maximum concentration at which a drug was given (which was not always constant for a given drug).

Finally, we defined one variable as the maximum viability at the three highest drug concentrations of both drugs. The underlying rationale is that when both drugs are very potent (i.e. kill many cells by themselves), it becomes technically very difficult to detect synergy.

Additionally, for each of the cell lines, molecular data (mutation, copy number aberration, methylation and gene expression) were available. Due to the high dimensionality of these data and the low sample size, we decided to only use mutations and copy number aberrations as direct input variables. We converted the mutations data provided in the DREAM data to Annovar's input format and subsequently re-annotated the mutations using Annovar (version 2015-06-17) [25]. For the re-annotation we used the following databases: refGene, ljb26all, avsnp142, cosmic70 and clinvar20150629 (genome build hg38). These extra annotations were then merged with original mutation annotations from the DREAM data. Using these re-annotated variants we created two binary mutation matrices (of cell lines by genes): a 'strict' variant containing high confidence disease-related/drug-response-related mutations and a 'deleterious' variant containing a larger list of deleterious mutations. For the strict matrix, we only selected mutations that occurred at least three times in the COSMIC database and any mutations that were

annotated with the following keywords in ClinVar: drug-response, inhibitor, cancer, carcinoma, leukaemia, melanoma, tumor, neoplasm. For the deleterious matrix we also (in addition to the strict mutations) selected mutations that were missense mutations according to the AZ-Sanger DREAM annotation and were deleterious according to the FATHMM [23] prediction in the Annovar annotation, and mutations that resulted in frameshifts or nonsense mediated decay. The strict mutation matrix was used for defining the pathways rules (see below). A filtered version of the deleterious matrix, which only considered genes mutated in at least 4 cell lines, was used as a direct input variable in the predictive models.

We considered a list of regions with recurrent copy number aberrations identified by Iorio *et al.* (2016) [13], from which we focused on the region's annotated putative driver genes. The resulting list contained 76 genes (Supplementary Table 1 of Aben et al. (2018) [1]). We defined a binary matrix indicating the presence of a gain or loss in that gene in that cell line, using the PICNIC [11] predicted copy number of ≥ 6 for amplifications and < 2 for deletions.

Finally, we defined a set of pathway rules, which were used as input variables. We selected 39 KEGG [14, 15] pathways related to cell growth and signaling. We mapped the driver mutations (using the 'strict mutation matrix') and drivers CNAs onto these pathways, defining a binary matrix of cell lines by pathways, where 1 indicates that at least one gene from the given pathway is altered (mutation or CNA) in that given cell line. In addition, we mapped the suggested drug targets to official gene symbols and used these to map the drugs onto the same selected pathways, defining a binary matrix of 119 drugs by 39 pathways, where a 1 indicates that at least one of the targets of the given drug is part of the given pathway. All these data were evaluated on the training and leaderboard sets. We defined the following six criteria:

- Does the pathway harbor a mutation and is it targeted by both drugs in the combination?
- Does the pathway harbor a mutation and/or CNA and is it targeted by both drugs in the combination?
- Does the pathway harbor a mutation and is it targeted by at least one drug in the combination?
- Does the pathway harbor a mutation and/or CNA and is it targeted by at least one drug in the combination?
- Is the pathway targeted by both drugs in the combination?
- Is the pathway targeted by at least one drug in the combination?

For each pathway and each criterion, we compared different groups of cell lines and drug combinations (criterion true vs. criterion false) and tested for difference in the mean of the synergy scores using a t-test. We defined a given pathway and a given criterion as a pathway rule when the difference in synergy score exceeded 10 and the false discovery rate was below 0.05. Using this approach on the training and leaderboard sets, we selected 29 pathway rules as a new set of variables.

SIMULATION STUDY TO COMPARE THE PREDICTIVE PERFORMANCE OF INDIVIDUAL AND JOINT RANDOM FOREST MODELS

To assess in which situations the joint model outperformed the individual one, we simulated a dataset with 14 cell lines, 497 variables and 10 drug combinations. This dataset closely follows the DREAM dataset in terms of number of samples (median number of cell lines per drug combination is 14) and number of variables (497). For the number of drug combinations, we have limited ourselves to 10, to reduce the computational burden.

For each entry in the input matrix $\mathbf{X}$, a value was drawn from a Bernoulli distribution with $p = 0.5$. For each entry in the response vector y, a value was drawn from a standard-normal distribution. We then engineered an association between the synergy scores (for all drug combinations in the simulation) and variable j by setting $y = y + 2\mathbf{X}_j$, resulting in an average effect size of two.

Subsequently, we trained an individual and a joint Random Forest on these data. We evaluated the predictive performance by generating a separate test set, using the same characteristics as the ones used to create the training data. The performance was measured using the 'primary score' described above.

To study the effect of sample size on the predictive performance, we varied the sample size between 14, 25, 50 and 100. Likewise, to study the effect of the number of variables, we varied the number of variables between 125, 250, 497 and 1000.

SIMULATION STUDY TO COMPARE THE INDIVIDUAL VARIABLE IMPORTANCE AND THE DRUG-COMBINATION-SPECIFIC VARIABLE IMPORTANCE

We generated a simulated dataset as above (14 cell lines, 497 variables and 10 drug combinations; values in $\mathbf{X}$ drawn from a Bernoulli distribution with $p = 0.5$; and values in y drawn from a standard-normal distribution). We then engineered an association between drug combination i and variable j by setting $y_i = y_i + e\mathbf{X}_j$, where e is the effect size. To compare the individual and joint models in different configurations, we varied the effect size of the association between $e = [0,1,2,3,4]$ and we varied the number of drug combinations in which the association occurred between $d = [1,3,5,7,9]$, leading to a total of 25 configurations. To estimate the variability in each configuration, we repeated the aforementioned process 50 times for each configuration, leading to a total of 1250 datasets. Each of the 1250 datasets was analyzed using:

- A Joint Random Forest, followed by DVI to rank the variables per drug combination.
- 10 Individual Random Forest models, followed by IVI to rank the variables per drug combination.

For each of the parametrizations (e and d), we determined:

- The median rank of the engineered association using a Joint Random Forest.
- The median rank of the engineered association using an Individual Random Forest.

- The significance of the difference between these two medians, using a Wilcoxon signed-rank test.

When $d = 1$, we determine median rank of the engineered association using the 50 repeats. When $d = 3$, we use the ranks for the 50 repeats in each of the 3 drug combinations in which the association was engineered, essentially yielding 150 repeats. In general, for each parameterization we determine the median rank of the engineered association using the $50d$ repeats.

We determined the significance of the difference between the individual and the joint model as follows. For each repeat and for each parametrization of e and d, we determined the median rank of the association across the d drug combinations in which the association was engineered. For each parametrization of e and d, we then determined the significance using a Wilcoxon signed-rank test (across the 50 repeats). The resulting p-values were corrected for multiple testing using a Benjamini-Hochberg correction.

4.5.2. NEGATIVE CORRELATIONS IN CROSS-VALIDATION RESULTS

We observed negative correlations between the observed and predicted synergy scores in the main text. These can be attributed to an artifact in cross-validation, which can be illustrated using the following example. Suppose we have a predictive model that simply always returns the average of its training data as a prediction (i.e. an intercept-only model). This model can be viewed as giving the best possible prediction in case there is no detectable structure in the training data. We now apply this model in a cross-validation to dataset with two samples, where $y = [1, 2]$ (note that the values in $\mathbf{X}$ do not matter using this predictive model). In the first cross-validation fold, we train on sample 1 and test on sample 2, hence $y_{train} = 1$ and $\hat{y}_2 = 1$. In the second cross-validation fold, we train on sample 2 and test on sample 1, hence $y_{train} = 2$ and $\hat{y}_1 = 2$. We now have $\hat{y} = [2, 1]$ and correlation$(y, \hat{y}) = -1$. Informally speaking, by taking the average of the training data, the predictive model overshoots in the first fold and undershoots in the second fold.

To show that this also holds in cases with more than two samples, consider the following experiment. Define y as 100 samples drawn from N(0,1). Again, we use the same intercept-only model in a cross-validation setting. Fig. 4.6A shows that, when using leave-one-out cross-validation, correlation$(y, \hat{y}) = -1$. In addition, we observe that these negative correlations are dampened when the the number of folds is decreased (Fig. 4.6B&C).

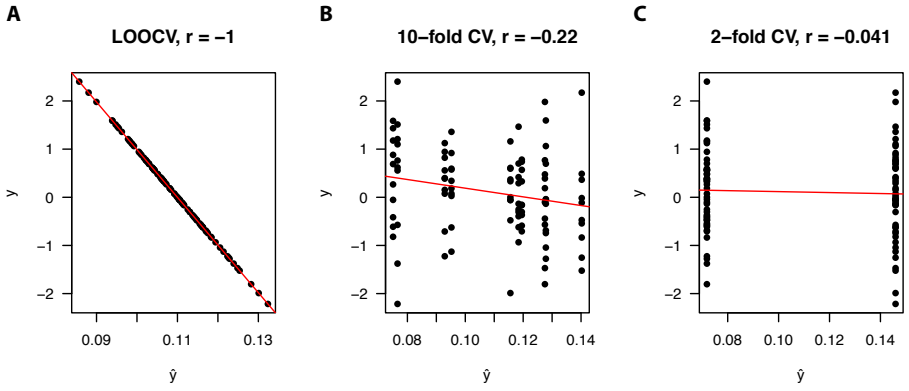


Figure 4.6: Correlation between the observed y and the predicted $\hat{y}$, using an intercept-only model with either (A) leave-one-out cross-validation (LOOCV), (B) 10-fold cross validation (CV), or (C) 2-fold CV.

4.5.3. ASSOCIATION OF MONOTHERAPY VARIABLES WITH SYNERGY

Ranking the variables by their JVI, we found that the monotherapy variables were the most important variables overall (one-tailed Mann-Whitney U test, $p = 2.474e\text{-}16$) (Supplementary Fig. 4.6). The top 21 variables contain only monotherapy variables, with the remaining two monotherapy variables in the top 27. Interestingly, the direction of the association varied per drug combination: for some drug combinations monotherapy sensitivity associated with synergy while for others it associated with antagonism (Fig. 4.7).

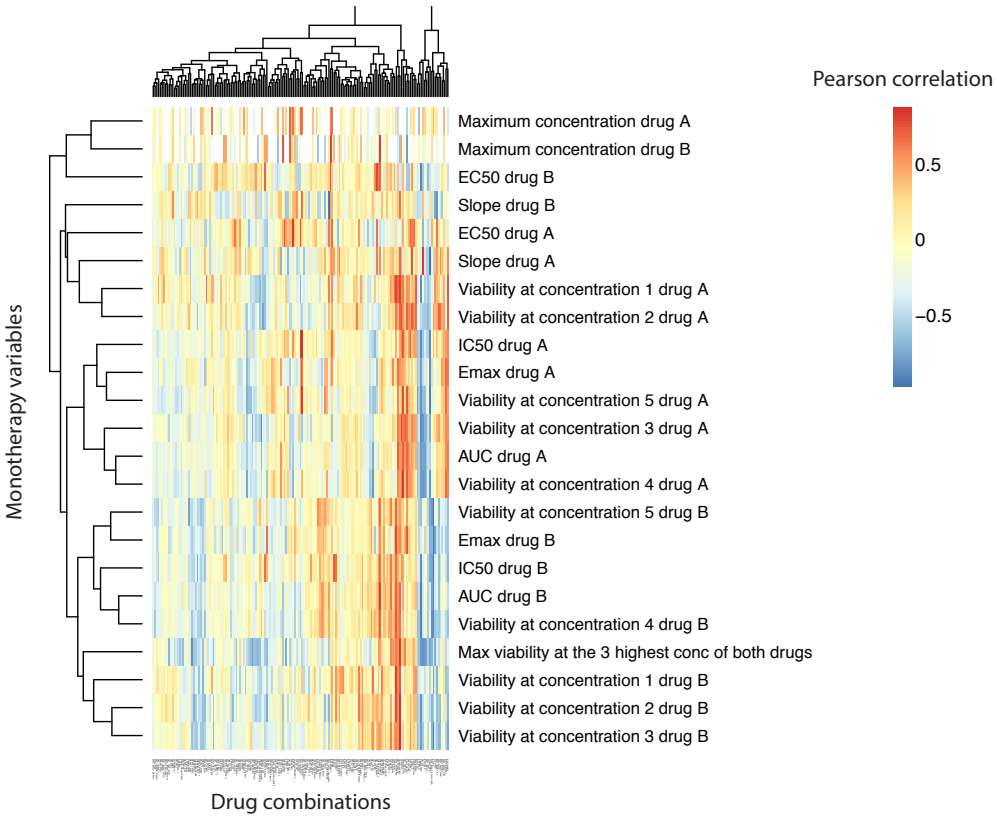


Figure 4.7: Pearson correlation between monotherapy variables and synergy scores for each of the drug combinations.

We note that the variation in the direction of the association is in line with the two hypotheses (the biological and the technical explanation) raised in the discussion in the main text. In the biological explanation, a small reduction in monotherapy viability can mean that the target is being engaged by the drug, which would be required for synergy

to occur. Hence, in this scenario, sensitivity would be associated with synergy. Conversely, in the technical explanation, when one drug kills most cells by itself (e.g. 80% of cells), it becomes very difficult to determine synergy, as most cells will already be dead when the two drugs are additive. Hence, in this scenario, resistance will be associated with resistance.

4.5.4. DESCRIPTION OF THE PATHWAY RULES

The pathway rules uncover some interesting global patterns about the contexts in which certain kinds of drugs tend to be synergistic. For instance, we find that pairs of drugs that both target the MAPK pathway are often synergistic, but only in cell lines that have mutations or copy number aberrations in the MAPK pathway. This is in line with expectation, as MAPK targeting drugs are typically only effective if the MAPK pathway is activated and many of the most prominent examples of synergy are between MAPK targeting drugs, e.g. BRAF and EGFR inhibitors in colorectal cancer [16, 21].

There are several other potentially interesting examples. For instance, if one of the drugs in the combination targets p53 signaling, the drug combination is likely to be antagonistic. Conversely, when both drugs target the p53-signaling pathway, the combination tends to be synergistic. Both of these associations are significant regardless of whether the cell line has any alterations in the p53 signaling pathway. Finally, the strongest association, in terms of effect size, is the synergy observed for TGFβ targeting drugs in cell lines that have an alteration in this pathway. The full list of pathway rules is given in Supplementary Table 2 of Aben et al. (2018) [1].

4.6. SUPPLEMENTARY FIGURES

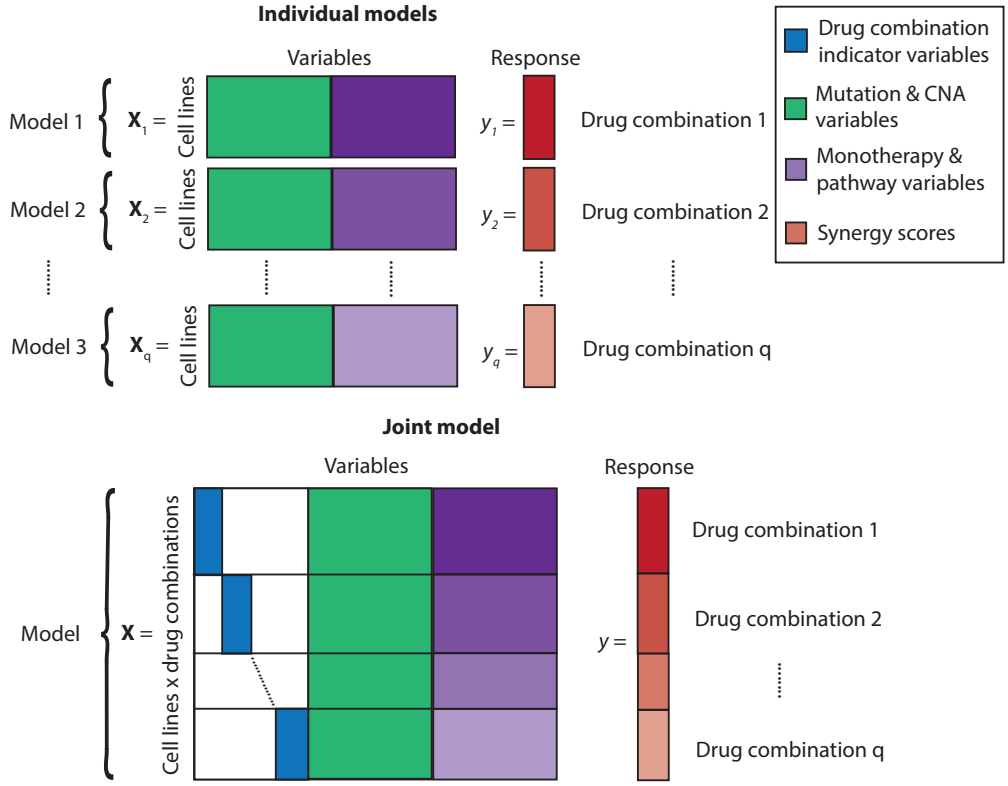


Figure S4.1: Overview of models used in this work. This figure is similar to Fig. 4.1, but with the added value of showing which variables are private to a cell line, drug combination pair (purple), private to a cell line (green) or private to a drug combination (blue). A: Individual models B: Joint model.

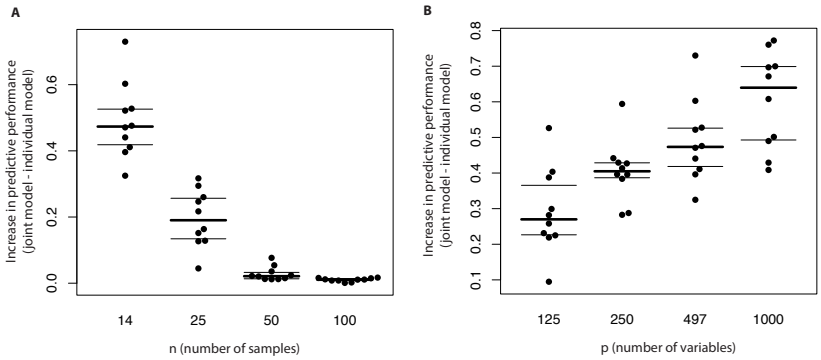


Figure S4.2: Simulation study to assess under what conditions the joint model outperforms the individual model using Random Forests. A: Difference in predictive performance between the joint and individual model for different values of n (number of samples). B: Difference in predictive performance between the joint and individual model for different values of p (number of variables).

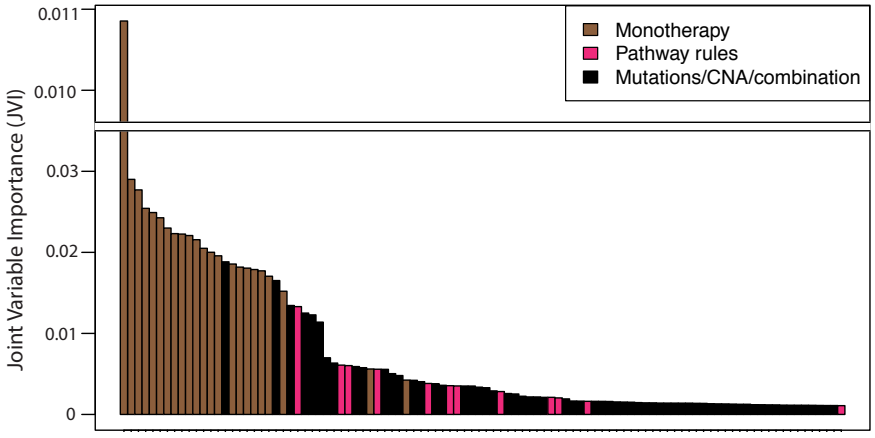


Figure S4.3: Top 100 highest JVI scores. Joint model Variable Importance (JVI) of the top 100 (out of 664) overall most important variables (for all drug combinations together).

REFERENCES

[1] Nanne Aben, Julian de Ruiter, Evert Bosdriesz, Yongsoo Kim, Gergana Bounova, Daniel Vis, Lodewyk Wessels, and Magali Michaut. Identifying biomarkers of anti-cancer drug synergy using multi-task learning. *bioRxiv*, page 243568, 2018.

[2] Piet Borst and Lodewyk Wessels. Do predictive signatures really predict response to cancer chemotherapy? *Cell Cycle*, 9(24):4836–4840, 2010.

[3] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[4] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[5] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM, 2004.

[6] Jonathan B Fitzgerald, Birgit Schoeberl, Ulrik B Nielsen, and Peter K Sorger. Systems biology and combination therapy in the quest for clinical efficacy. *Nature chemical biology*, 2(9):458, 2006.

[7] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4), 2009.

[8] Kaitlyn M Gayvert, Omar Aly, James Platt, Marcus W Bosenberg, David F Stern, and Olivier Elemento. A computational approach for identifying synergistic drug combinations. *PLoS computational biology*, 13(1):e1005308, 2017.

[9] Nori Geary. Understanding synergy. *American Journal of Physiology-Endocrinology and Metabolism*, 304(3):E237–E253, 2013.

[10] Mehmet Gönen and Adam A Margolin. Drug susceptibility prediction against a panel of drugs using kernelized bayesian multitask learning. *Bioinformatics*, 30(17):i556–i563, 2014.

[11] Chris D Greenman, Graham Bignell, Adam Butler, Sarah Edkins, Jon Hinton, Dave Beare, Sajani Swamy, Thomas Santarius, Lina Chen, Sara Widaa, et al. Picnic: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics*, 11(1):164–175, 2009.

[12] Tin Kam Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.

[13] Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, 2016.

[14] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[15] Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: back to metabolism in kegg. *Nucleic acids research*, 42(D1):D199–D205, 2013.

[16] Bertram Klinger, Anja Sieber, Raphaela Fritsche-Guenther, Franziska Witzel, Leanne Berry, Dirk Schumacher, Yibing Yan, Pawel Durek, Mark Merchant, Reinhold Schäfer, et al. Network quantification of egfr signaling unveils potential for targeted combination therapy. *Molecular systems biology*, 9(1):673, 2013.

[17] Michael P Menden, Francesco Iorio, Mathew Garnett, Ultan McDermott, Cyril H Benes, Pedro J Ballester, and Julio Saez-Rodriguez. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS one*, 8(4):e61318, 2013.

[18] Michael Patrick Menden, Dennis Wang, Yuanfang Guan, Michael Mason, Bence Szalai, Krishna C Bulusu, Thomas Yu, Jaewoo Kang, Minji Jeon, Russ Wolfinger, et al. Community assessment of cancer drug combination screens identifies strategies for synergy prediction. *bioRxiv*, page 200451, 2017.

[19] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[20] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[21] Anirudh Prahallad, Chong Sun, Sidong Huang, Federica Di Nicolantonio, Ramon Salazar, Davide Zecchin, Roderick L Beijersbergen, Alberto Bardelli, and René Bernards. Unresponsiveness of colon cancer to braf (v600e) inhibition through feedback activation of egfr. *Nature*, 483(7387):100, 2012.

[22] Sven Rottenberg, Marieke A Vollebergh, Bas de Hoon, Jorma de Ronde, Philip C Schouten, Ariena Kersbergen, Serge AL Zander, Marina Pajic, Janneke E Jaspers, Martijn Jonkers, et al. Impact of intertumoral heterogeneity on predicting chemotherapy response of brca1-deficient mammary tumors. *Cancer research*, 72(9):2350–2361, 2012.

[23] Hashem A Shihab, Julian Gough, David N Cooper, Peter D Stenson, Gary LA Barker, Keith J Edwards, Ian NM Day, and Tom R Gaunt. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models. *Human mutation*, 34(1):57–65, 2013.

[24] Daniel J Vis, Lorenzo Bombardelli, Howard Lightfoot, Francesco Iorio, Mathew J Garnett, and Lodewyk FA Wessels. Multilevel models improve precision and speed of ic50 estimates. *Pharmacogenomics*, 17(7):691–700, 2016.

[25] Kai Wang, Mingyao Li, and Hakon Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010.

[26] Han Yuan, Ivan Paskov, Hristo Paskov, Alvaro J González, and Christina S Leslie. Multitask learning improves prediction of cancer drug sensitivity. *Scientific reports*, 6, 2016.

[27] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

5

A SCREEN OF 765 CELL LINES AND 54 DRUG COMBINATIONS TO STUDY SYNERGISTIC DRUG INTERACTIONS IN CANCER

NANNE ABEN, PATRICIA JAAKS, DANIEL J. VIS, SYD BARTHORPE, HOWARD LIGHTFOOT, DIEUDONNE VAN DER VEER, RUBAYTE RAHMAN, ELIZABETH COKER, MAGALI MICHAUT, MATHEW J. GARNETT, LODEWYK F.A. WESSELS

ABSTRACT

Single-agent anti-cancer treatments are often ineffective by themselves, for example due to the tumor bypassing the drug blockade by activating an alternative signaling pathway. Simultaneously inhibiting the alternative pathway using a second drug can result in a more effective treatment, a concept known as synergy. Previous reports on synergistic drug combinations were often limited by the small number of combinations and tumor cell lines tested. Additionally, many combinations lack a biomarker that allows for identification of responders. We performed a large-scale drug combination screen of 54 combinations across 765 cell lines. We found that synergies are rare events and generally not cancer type specific. Rescreening 20 combination in 736 cell lines showed reproducibility is associated with the strength of the synergy. Five combinations showed synergy in >10% of the cell lines, including several combinations of AZD7762 (CHEK1/2) with a DNA damaging agent (DDA) and the combination of Olaparib (PARP) with Temozolomide (DDA). We found high SLFN11 and low NQO1 expression to be associated with synergy to Olaparib+Temozolomide, and TP53 mutations with synergy to AZD7762+DDA.

5

5.1. INTRODUCTION

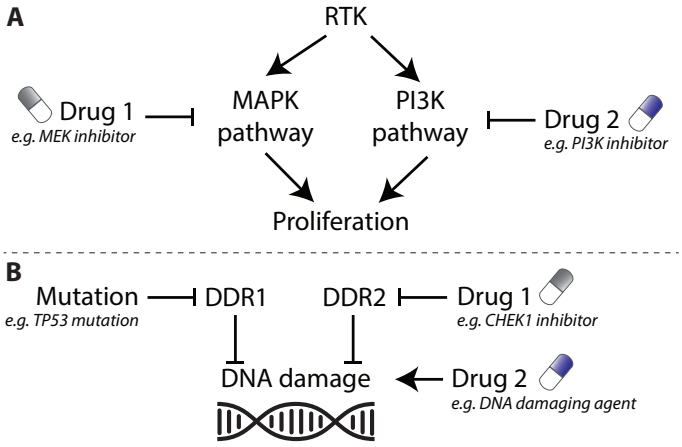


Figure 5.1: Illustration of synergistic interactions between anti-cancer drugs. (A) Example using parallel signaling pathways. (B) Example using DNA damage response (DDR) pathways.

Anti-cancer therapies aim to kill tumor cells by targeting their vulnerabilities. For example, many tumors depend on the MAPK pathway, hence inhibiting this pathway has the potential to eradicate them [21] (Figure 5.1A). Similarly, tumors often lose the functionality of a particular DNA damage response (DDR) pathway, resulting in a stronger dependency on the remaining DDR pathways. Hence inhibiting these pathways becomes an interesting treatment option [16] (Figure 5.1B).

Unfortunately, there are many examples were these treatments are not effective by themselves. For instance, MAPK pathway inhibition may be rendered ineffective by the activation of a parallel pathway, such as the PI3K pathway [17] (Figure 5.1A). Similarly,

while the tumor may have developed an increased sensitivity to DNA damage upon inhibition of a DDR pathway, it may not accumulate enough DNA damage to die [16].

In these cases, the treatment efficacy can be increased by adding a second drug. In the first example we may add a PI3K inhibitor, thereby inhibiting both the dependency (MAPK) and the backup (PI3K), hence killing the cancer cells [17] (Figure 5.1A). Similarly, for the DDR pathway inhibition example, we may add a DNA damaging agent (DDA), enforcing accumulation of DNA damage to the point of no return and subsequent cell death [16] (Figure 5.1B).

Drug combinations like these are often referred to as synergistic combinations, where the combination response is greater than one would expect based on the response to the individual drugs. The concept of synergy is clearly illustrated in the first example, where neither the MAPK nor the PI3K pathway inhibitor were effective by themselves, but the combination was effective. In the second example, the DDA may have been effective by itself, but by inhibiting a DDR pathway, the DDA can be used at a much lower dose, thereby limiting the toxicity to non-cancer cells. Hence, in this case, we observe synergy between the DDR pathway inhibitor and a low dose of the DDA.

Many synergistic combinations of anti-cancer drugs have previously been reported (e.g. [5, 7, 13, 19, 22]). However, these combinations have often only been tested in small subsets of tumor cell lines, hence it is not clear to what extent these synergies generalize to a larger set of tumors and across a larger amount of genetic variation. In addition, many combinations lack a biomarker that allows us to stratify which patients are likely to respond to the combination. Even when we have an intuition for the mechanism by which a given drug combination works (like in the examples above), this mechanistic insight may not readily function as a biomarker, as these backup pathways are often unknown until their activation upon first treatment.

To help address these problems, we have performed a large-scale drug combination screen, in which we screened 54 drug combinations across 765 tumor cell lines, representing 42 cancer types. In addition, to assess the reproducibility of our screen, we have rescreened 20 combinations and 736 cell lines. We report which combinations showed synergy most frequently, thus indicating their relevance to a large and diverse set of tumors. Finally, we statistically associated molecular data (i.e. mutations, copy number aberrations, gene expression and proteomics) with synergy to a given drug combination, from which the strongest associations may be good candidate biomarkers.

5.2. RESULTS

5.2.1. THE SCREENING APPROACH

We have screened 54 drug combinations across 765 cell lines, resulting in 40,850 unique cell line - drug combination pairs (Figure 5.2A) (Supplementary Data 1-6). These cell lines span 42 different cancer types, of which the largest sets include breast carcinoma (47 cell lines), lung adenocarcinoma (47), colorectal carcinoma (45), small-cell lung cancer (40) and melanoma (31) (Figure 5.2B). Drug combinations were chosen based on key cancer pathways, with a preference for using drugs that have high target specificity and are either approved or in late-stage clinical trials. Screening was performed using 2,972 1536-well plates, from which drug response was determined using a CellTiter-Glo (CTG)

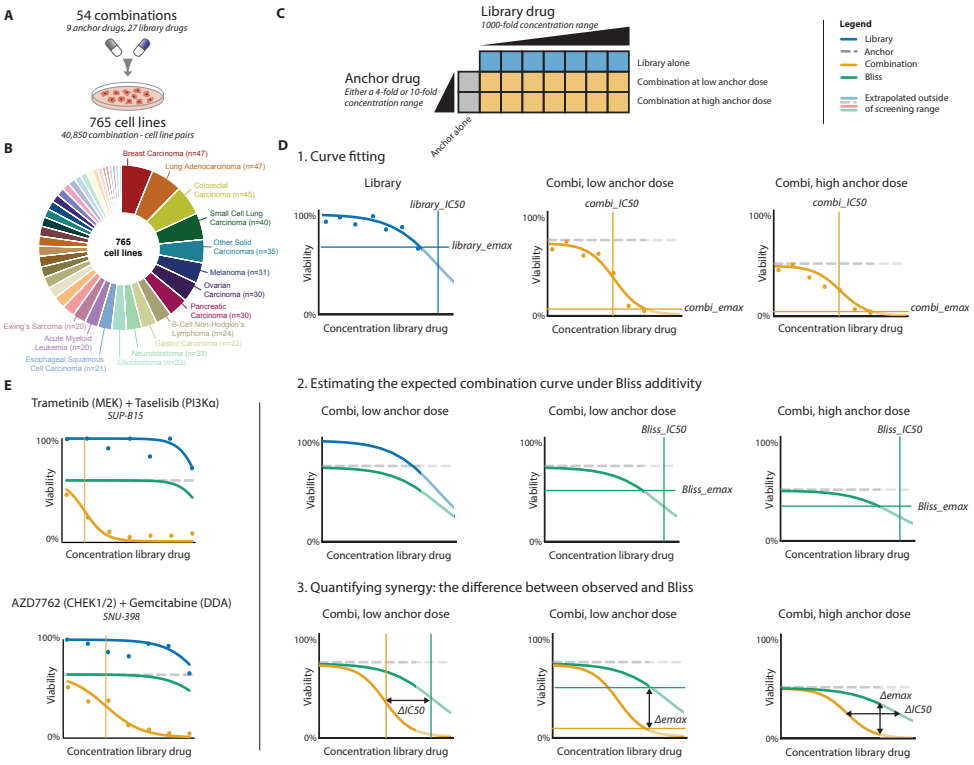


Figure 5.2: Overview of the screening approach. (A) Number of cell lines and drug combinations screened. (B) Number of cell lines per cancer type. (C) Schematic overview of the anchor - library approach. (D) Cartoon representing curve fits and synergy estimation (E) Examples of synergy from our data.

readout 72h after drug treatment (Methods).

To enable screening at this scale, we used an 'anchor - library' approach, in which the 'anchor drug' is screened at two concentrations and the 'library drug' is screened at seven concentrations (Figure 5.2C). For each cell line - drug combination pair, this results in three different seven-point dose response experiments: one for the library alone, one for the combination at the low anchor dose, and one for the combination at the high anchor dose (Figure 5.2D.1). Each of these dose-response experiments was summarized by fitting a sigmoidal curve, from which we derived two parameters: the *IC50*, which is the concentration that leads to a 50% reduction in cell viability (relative to 100% for the library or to the anchor viability for the combination; concentration is expressed in fold-changes, see Methods); and the *emax*, the viability corresponding to the highest concentration of the library drug. For the library curves, we call these parameters *library_IC50* and *library_emax*. For the combination curves, we call these parameters *combi_IC50* and *combi_emax*.

Rigorous quality checks were performed throughout, for example using viability and curve fit quality statistics (Supplementary Figure 5.5A-C). In addition, we used the *library_IC50* and $anchor\_viability$ parameters to verify that we recover known associ-

ations between molecular data and monotherapy response, such as the association between BRAF mutation and Dabrafenib (BRAF) response (Supplementary Figure 5.5D). Finally, we checked the concordance of $anchor\_viability$, *IC50* and *emax* parameters between biological replicates (median correlation: 0.84) and technical replicates (median correlation: 0.96) (Supplementary Figure 5.5E&F).

To use the *IC50* and *emax* parameters to detect synergy, we first determined what the combination curves would look like under additivity (i.e. in the absence of synergy). To this end, we used Bliss independence [3] of the response to the anchor and the library drug alone. From the resulting curves we determined the *IC50* and *emax* parameters, to which we will refer as *Bliss_IC50* and *Bliss_emax*, respectively (Figure 5.2D.2).

We quantified synergy by considering the difference between the observed parameters and the parameters derived from the Bliss combination curve. Specifically, we used two measures: the Δ*IC50*, which is defined as the difference between the *Bliss_IC50* and the *combi_IC50*; and the Δ*emax*, which is defined as the difference between the *Bliss_emax* and the *combi_emax* (Figure 5.2D.3).

To focus on strong synergies, we set thresholds for Δ*IC50* and Δ*emax* at three standard deviations above the population mean (Methods) (Supplementary Figure 5.5). For the Δ*IC50*, this corresponds to a difference of 5.2 fold-changes between the *Bliss_IC50* and *combi_IC50*, or equivalently, a $2^{5.2}$ = 37-fold difference in concentration (when the *library_IC50* is outside of the screening range, this threshold was set to a $2^{5.6}$ = 49-fold difference in concentration, see Methods). For the Δ*emax*, the threshold corresponds to a difference of 33 percentage points between the *Bliss_emax* and the *combi_emax*. We considered a cell line - drug combination pair as synergistic when, for at least one of the two anchor concentrations, either the Δ*IC50* or the Δ*emax* was higher than these thresholds.

We found that 23% of the synergistic cases passed both the Δ*IC50* and the Δ*emax* thresholds, whereas 12% passed the Δ*IC50* threshold only and 65% passed the Δ*emax* threshold only, indicating that combining these two measures captured different types of synergies (Supplementary Figure 5.5D-E).

The anchor doses were optimized such that the anchor by itself would result in roughly 20% cell kill (Supplementary Figure 5.5A). This way, the anchor drug is likely to be given at a functional concentration at which it engages the target, while not killing most cells by itself (in which case it becomes very hard to reliably detect synergy). Indeed, we find that synergy was most frequently observed when the anchor viability was between 50% and 80% (Supplementary Figure 5.5B). Interestingly, the majority of synergistic cases (80%) were based on a single anchor dose only, with 22% being based on the low anchor dose only and 58% being based on the high anchor dose only. This further underlines the importance of optimizing the anchor doses.

Figure 5.2E shows two examples of synergistic cell line - drug combination pairs that pass the Δ*IC50* / Δ*emax* thresholds. Plots similar to these ones can be made using CombiXplore, a web interface that we have developed to efficiently browse and visualize our anchor - library combination data. CombiXplore allows the user to filter (e.g. on cell line, anchor and library names, as well as on curve fit statistics such as Δ*IC50*); create summary statistics; visualize dose response curves; inspect replicates; and perform simple biomarker identification.

5.2.2. THE SCREENING APPROACH

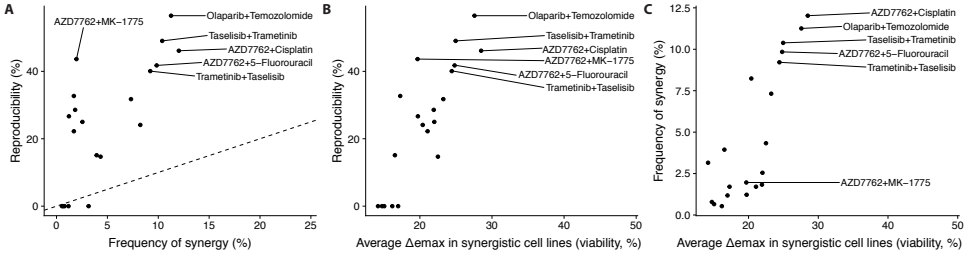


Figure 5.3: Drug combinations that are more frequently synergistic are also more reproducible. (A) Relation between reproducibility and the frequency at which synergy is observed. The dashed line indicates the reproducibility rate expected by chance (Methods). (B) Relation between reproducibility and the effect size of the synergy (in this case $\Delta emax$). (C) Relation between effect size of the synergy (in this case $\Delta emax$) and the frequency at which the synergy is observed.

5

In order to assess the reproducibility of our screen, 20 combinations and 736 cell lines were rescreened (Supplementary Figure 5.5), from which we determined the reproducibility rate by considering how many of the cell line - drug combination pairs that were synergistic in the original set were also synergistic in the validation screen (Methods). While the overall reproducibility rate was 36%, we found that combinations that were more frequently synergistic also showed a higher reproducibility than expected by chance (e.g. Olaparib+Temozolomide: 58%) (Figure 5.3A) (Methods). This appeared to be related to the effect size of the synergy (e.g. $\Delta emax$): larger effect sizes are more likely to validate (Figure 5.3B), and in combinations that frequently showed synergy, the effect size of the synergy also tended to be larger (Figure 5.3C). Given these reproducibility rates, we decided to focus on the top 10 most frequently synergistic combinations.

5.2.3. LANDSCAPE OF SYNERGY

Using the combined data from the original and the validation set, we determined how frequently each combination resulted in synergy (Figure 5.4A&B). The most frequently synergistic combination was AZD7762+Gemcitabine, a combination of a CHEK1/2 inhibitor with a DDA, which showed synergy in 14% of the cell lines. CHEK1 controls the G2/M checkpoint, hence inhibition of CHEK1 reduces the cell's ability to detect DNA damage, especially when the G1 checkpoint has also been lost, for example due to inactivation of TP53 [8, 11, 23]. Interestingly, other combinations of AZD7762+DDA (AZD7762+Cisplatin and AZD7762+5-Fluorouracil) also frequently showed synergy (in second and fifth place, at 12% and 10% of the cell lines respectively), further corroborating that CHEK1/2 inhibition in combination with a DDA is a potent strategy.

In third place, at 11% of the cell lines, is Olaparib+Temozolomide, a combination of a PARP inhibitor and a DDA. Many tumors have lost their ability to perform homologous recombination, a pathway for error-free DNA damage repair, making them dependent on an alternative pathway that uses PARP to perform error-free DNA repair. Hence, if PARP is inhibited, these cells have to resort to inaccurate forms of DNA repair, making them more sensitive to DDAs [16].

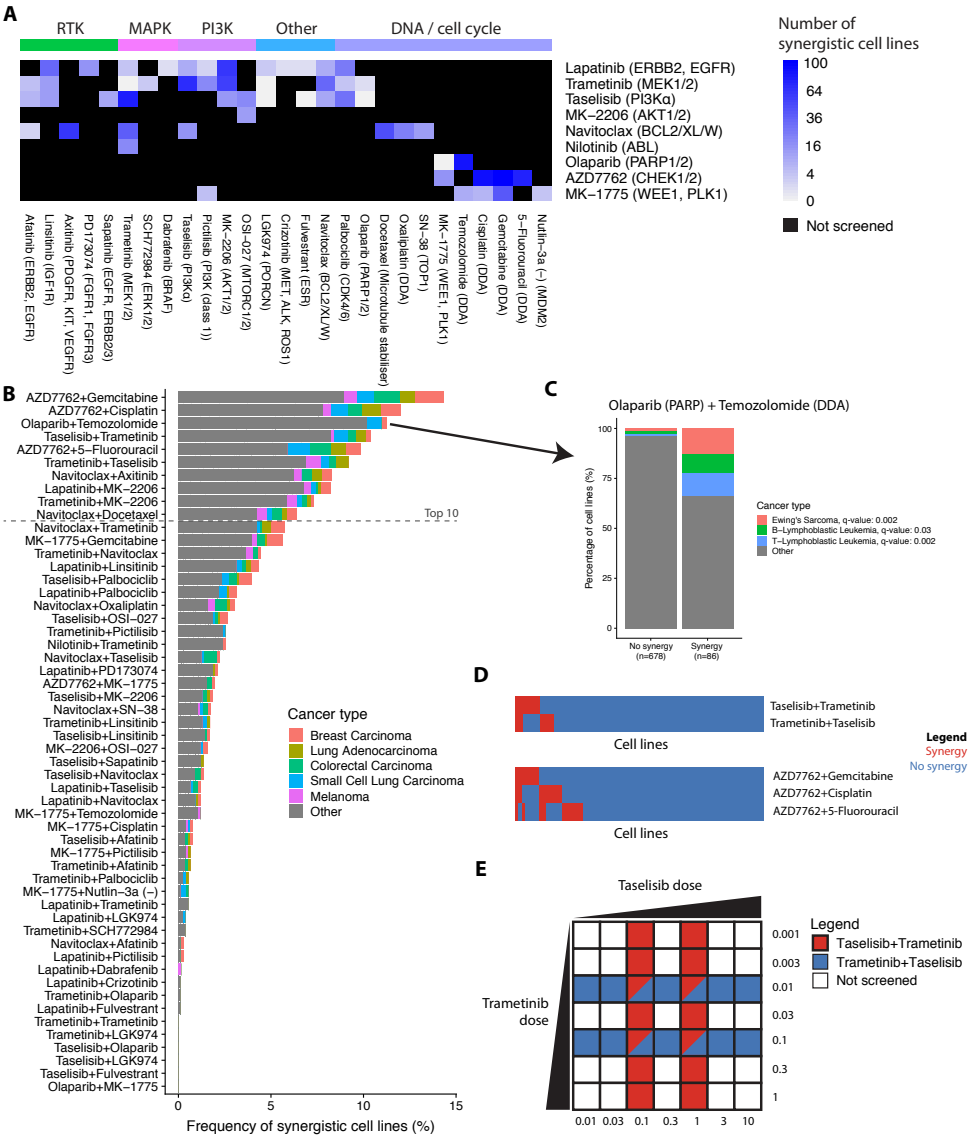


Figure 5.4: Overview of synergy in the screen. (A) Heatmap with anchor drugs on the rows and library drugs on the columns, where the color indicates whether a given combination has been screened, and if so, how often it resulted in synergy. (B) Barplot indicating for each combination how frequently it showed synergy. The colors represent the five biggest cancer types. (C) The statistically significant associations between cancer type and synergy for the Olaparib (PARP) + Temozolomide (DDA) combination. (D) Overlap in synergistic cell lines between similar drug combinations. (E) Dose-range screened for Taselisib+Trametinib and Trametinib+Taselisib combinations.

The combination Taselisib+Trametinib, a combination of a PI3Kα inhibitor with a MEK inhibitor (targeting the MAPK pathway), ranked in fourth place, showing synergy

in 10% of the cell lines. This combination is an instance of the example used in the introduction where two parallel pathways (MAPK and PI3K) are inhibited (Figure 5.1A). Interestingly, unlike the other combinations, this combination was also screened with the anchor and the library swapped, as Trametinib+Taselisib, which was also frequently synergistic (6th place, 9% of the cell lines). In addition, the combination Trametinib+MK2206, combining a MEK inhibitor with an AKT inhibitor (another PI3K pathway inhibitor) was also frequently synergistic (9th place, 7%). Interestingly, Trametinib+Pictilisib (a PI3K Class I inhibitor) was not frequently synergistic (19th place, 3%). These results suggest that inhibition of the MAPK pathway and the PI3K pathway is a potent strategy, but that the exact drugs with which the pathways are inhibited (which may have different specificity, potency and targets) make a big difference.

Though many of the frequently synergistic combinations were very similar to each other (e.g. the AZD7762+DDA combinations), they often did not show synergy in the same cell lines (Figure 5.4D). For example, for the cell lines that showed synergy to at least one of the two combinations of Trametinib and Taselisib (with the library and anchor roles of the drugs swapped), only 22% showed synergy to both combinations, whereas 34% of the synergies was specific to Trametinib+Taselisib, and 44% of the synergies was specific to Taselisib+Trametinib. We believe this is due to differences in the concentrations at which the drugs were screened (Figure 5.4E), similar to how we often observe synergy at only one of the anchor doses. We see a similar pattern for the AZD7762+DDA combinations, and while these DDA drugs may result in synergy through different modes of action, we think the lack of overlap is also largely due to differences in the effective dose range (i.e. at the administered doses, one DDA leads to more cell kill by itself than another DDA).

Interestingly, we observed that combinations were not highly specific to a given cancer type, but showed synergy across cancer types. Figure 5.4B illustrates this for the five biggest cancer types. More systematically, we tested, for each drug combination and each cancer type, whether there are more synergistic cell lines in a given cancer type than one would expect by chance. Focusing on the top 10 most frequently synergistic drug combinations, we found three statistically significant enrichments (FDR-corrected p-value (q-value) < 0.05, Fisher exact test), all with Olaparib+Temozolomide (Figure 5.4C). Of these, the enrichment with Ewing's sarcoma corroborates what has previously been found [4], whereas the enrichments with lymphoblastic leukemia may be novel. If we extend our scope to all 54 drug combinations, we find one more enrichment (Lapatinib (EGFR, ERBB2) + Linsitinib (IGF1R) in Oral Cavity Carcinoma, Supplementary Figure 5.5), though given the reduced reproducibility rate we urge caution with the interpretation thereof. Overall, since these enrichments are small in number and in effect (a single cancer type accounts for only up to 21% of the synergistic cell lines for a drug combination (Figure 5.4C, Supplementary Figure 5.5)), we conclude that synergy is not strongly associated with cancer type.

5.2.4. BIOMARKER IDENTIFICATION

To identify putative biomarkers, we evaluated the association between the molecular data (mutations, CNA, gene expression and proteomics) and the synergy. In light of the reproducibility rates, we again focused on the top 10 most frequently synergistic combi-

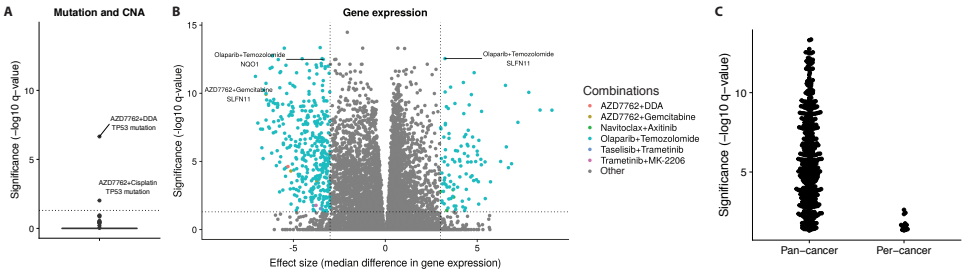


Figure 5.5: Associations between molecular data and synergy using only the top 10 most frequently synergistic drug combinations. (A) Pan-cancer associations for mutation and CNA data. (B) Pan-cancer associations for gene expression data. (C) Comparison of statistically significant pan-cancer and per-cancer associations.

nations, though results were very similar when we extended the analysis to all 54 combinations (Supplementary Figure 5.5). For each combination and each molecular feature, we evaluated the association between the molecular feature and the synergy status using either a Mann-Whitney U test (for continuous features, such as gene expression) or a Fisher exact test (for binary features, such as mutations). We performed this analysis both pan-cancer (across all cancer types) and per-cancer (per cancer type). This resulted in 9,312,756 statistical tests, of which 499 were significant (FDR-corrected p-value (q-value) < 0.05, absolute effect size > 3) (Supplementary Data 8).

The strongest associations between molecular data and synergy (q-value < 10^{-4}) were found for Olaparib+Temozolomide and AZD7762+DDA combinations (Figure 5.5). This is in line with expectation, as 1) these combinations frequently showed synergy (Figure 5.4B), resulting in more balanced classes (synergy vs. no synergy) and hence greater statistical power; and 2) the reproducibility is higher for these combinations (Figure 5.3).

Of note, we detected many more associations between monotherapy response (defined as *library_emax* < 50%) and molecular features than between synergy and molecular features. While we found an average of 31 associations per drug combination for synergy, we found 343 associations per drug for monotherapy response. In part this is likely also due to a larger class imbalance for synergy: synergy is rarer than monotherapy response (median: 78.5 synergies per drug combination – for the top 10 most frequently synergistic combinations – vs. 198.5 monotherapy responses per drug). Additionally, the added complexity of a second drug may simply make the identification of biomarkers of synergy more challenging.

Most associations with synergy (>99%) were found using gene expression data, with only 2 associations using mutation data and 4 associations using the proteomics data (no associations were found using the CNA data). For a large part, this can be attributed to the number of features that we tested for each of these datasets: out of the 19,688 molecular features used in the analysis, 18,672 (95%) were gene expression features. The lack of strong associations with proteomics data is surprising (none have a q-value < 0.01), but may be attributed to proteomics data being available for only 291 cell lines, as well as being the only dataset that was sourced from a different project (i.e. the MCLP [10]).

For biomarkers of monotherapy response, it has previously been reported that many

associations with gene expression are actually a proxy for cancer type [1, 6]. For example, when a given cancer type is generally more sensitive to a given drug, all genes that are highly expressed specifically in that cancer type will also be associated with response. We note that for synergy we found only few enrichments with cancer type, hence this proxy effect is more limited. For the associations between molecular data and Olaparib+Temozolomide, a combination for which we found three cancer type enrichments, we will further discuss the effect of cancer type in the corresponding section. For AZD7762+DDA, we note that we found no enrichments with cancer type, hence for this combination the associations with molecular data are not a proxy for cancer type.

Finally, pan-cancer biomarker analyses resulted in many more associations than per-cancer analyses (485 vs. 14), and these associations were on average also much stronger (Figure 5.5C). This increase in the detected number of associations is likely due to the increased statistical power stemming from being able to employ a much larger number of cell lines than can be used in the per-cancer analyses.

OLAPARIB (PARP) + TEMOZOLOMIDE (DDA)

Most of the significant associations (469 out of 499) were found for the combination Olaparib (PARP) +Temozolomide (DDA). The majority of these genes appear to contribute to a single molecular process, as a large amount of their variance can be explained by a single principal component (PC1) (Figure 5.6A). Of the genes that are most strongly associated with both PC1 and synergy (Figure 5.6B), *NQO1* may be most closely related to PARP biology, as it is known to regulate NAD+ [15], a cofactor of PARP [2]. Hence, low *NQO1* can reduce PARP activity, which in turn can make the cancer cells more prone to PARP inhibition (Figure 5.6C).

Interestingly, Figure 5.6B also shows that *SLFN11* is strongly associated with synergy, but hardly with PC1. This suggests that *SLFN11* is associated with synergy through an independent molecular process. Indeed, while *NQO1* may modulate the synergy via NAD+ depletion, *SLFN11* is known to do so by inducing prolonged S-phase arrest, during which the sister chromatids are not available, thereby indirectly reducing homologous recombination [14] (Figure 5.6C). In this way, *SLFN11* sensitizes cancer cells to PARP inhibition, many different DDAs, and hence also to the combination of PARP inhibition with a DDA [14, 24].

Our hypothesis of two independent processes is further corroborated by the lack of correlation between *SLFN11* and *NQO1* ($r = -0.13$) (Figure 5.6E). In addition, $SLFN11_{\text{high}}$ & $NQO1_{\text{low}}$ predicts synergy better than each of the genes individually (Fisher exact test, $p = 10^{-26}$, $p = 10^{-16}$ and $p = 10^{-18}$ respectively), indicating that both processes contribute to the synergy.

We note that Ewing's sarcoma and lymphoblastic leukemia cell lines were almost invariably $SLFN11_{\text{high}}$ & $NQO1_{\text{low}}$ (Supplementary Figure 5.5B-D), which may explain why these cancer types show synergy to Olaparib+Temozolomide so frequently (Figure 5.4C). Of note, cell lines from these cancer types were not the only $SLFN11_{\text{high}}$ & $NQO1_{\text{low}}$ cell lines to show synergy (Supplementary Figure 5.5A), hence these two genes do not solely function as a proxy for these cancer types.

Interestingly, both *SLFN11* and *NQO1* are associated not only with synergy, but also with monotherapy response of both Olaparib and Temozolomide (Supplementary Figure 5.5E-H). We wondered whether a monotherapy response to one of the drugs was

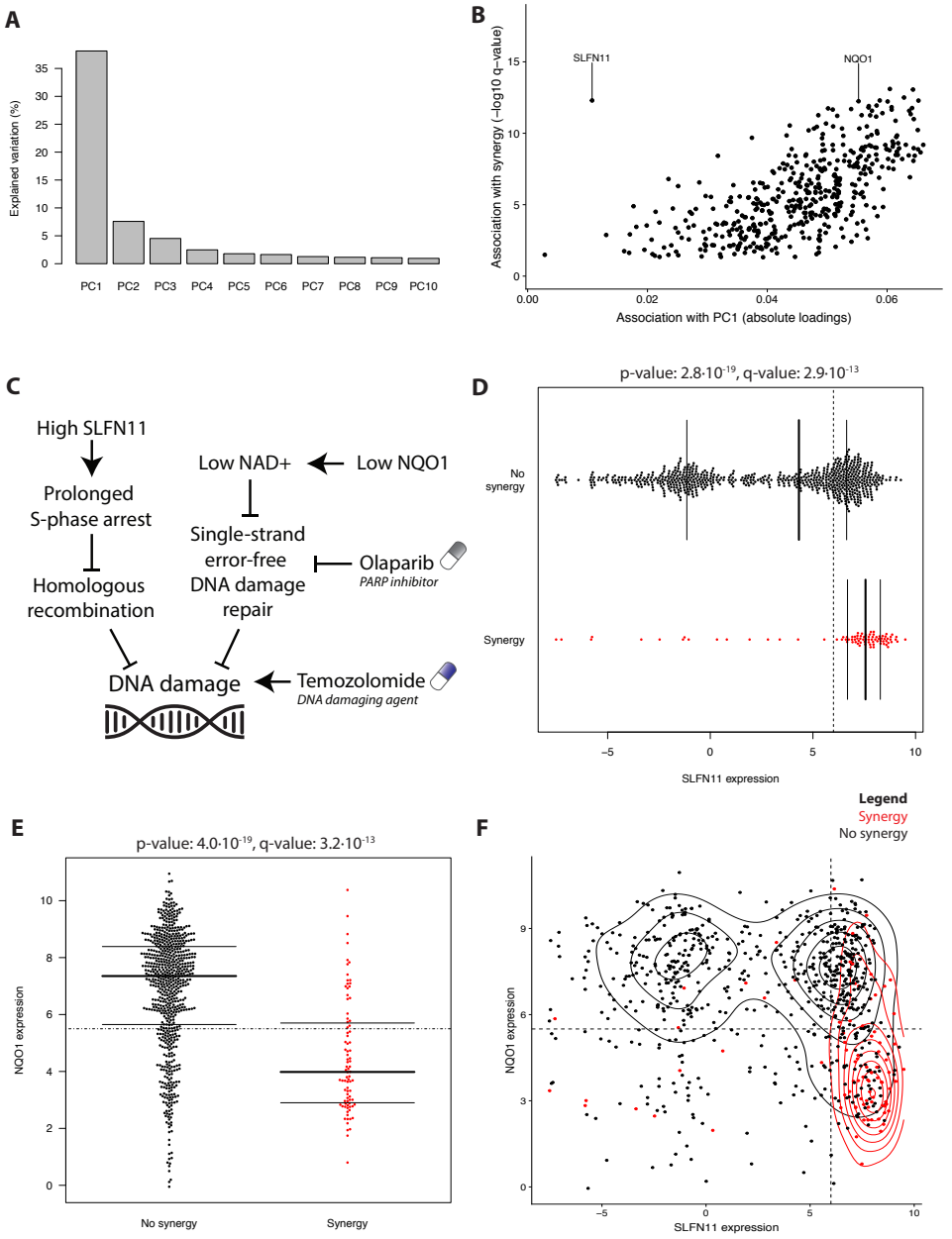


Figure 5.6: Associations between molecular features and synergy to the combination Olaparib (PARP) + Temozolomide (DDA). (A&B) Principal Component Analysis (PCA) of 459 gene expression features that were significantly associated with synergy. Panel A shows the amount of variation explained by the first 10 principal components (PCs). Panel B shows the relation between the absolute loadings of PC1 and the association with synergy. (C) Cartoon of proposed mechanism by which these two genes may modulate the synergy. (D-F) Relationships between *SLFN11* expression, *NQO1* expression and synergy to Olaparib+Temozolomide. Dashed lines indicate thresholds used to define $SLFN11_{\text{high}}$ and $NQO1_{\text{low}}$ cell lines.

more likely to result in synergy. Indeed, we found that cell lines that responded to Olaparib were more often synergistic for Olaparib+Temozolomide, while this was not the case for Temozolomide (Supplementary Figure 5.5I&J). We believe this is likely because the response to Olaparib indicates a dependency of the tumor cell on PARP, which is then amplified by the addition of a DDA.

AZD7762 (CHEK1/2) + DDA

Following Olaparib+Temozolomide, the next most significant associations were found for AZD7762 (CHEK1/2) + DDA. For example, for all three AZD7762+DDA combinations, nearly all synergistic cell lines were TP53 mutants (Figure 5.7A, Supplementary Figure 5.5A). As mentioned above, this is in line with previous reports: a mutation in TP53 leads to loss of the G1 checkpoint, making the tumor cells dependent on the G2/M checkpoint. This checkpoint is controlled by CHEK1, hence by inhibiting CHEK1, the tumor cells are more sensitive to DNA damage and the addition of a DDA then leads to subsequent cell death [8, 11, 23] (Figure 5.7B).

Interestingly, even though almost all synergistic cell lines are TP53 mutants, only 35% of the TP53 mutants show synergy to at least one of the AZD7762+DDA combinations. For the remaining 65%, there may be an additional molecular event that causes these tumors to not respond synergistically, for example the reactivation of one of these checkpoints or the use of drug transporters to transport the DDA out of the tumor cell. To investigate this, we tried to identify molecular features that can predict synergistic cases within the TP53 mutant cell lines. However, we found no strong associations that were specific to this scenario (Figure 5.7C).

In addition to TP53 mutations, we also found *SLFN11* to be associated with AZD7762+Gemcitabine, in this case with low *SLFN11* being associated with synergy (Figure 5.7E). This association can best be understood by considering the response to Gemcitabine alone, for which we see that *SLFN11* high cells are especially sensitive (Figure 5.7D). Since in most *SLFN11* high lines Gemcitabine alone is sufficient to kill the tumor cells, the addition of a second drug has little added value. On the other hand, in *SLFN11* low cell lines Gemcitabine alone is typically less potent, leaving room for synergy when AZD7762 is added.

The above may suggest that when treating *SLFN11* high tumors with the combination AZD7762+Gemcitabine, the dose of Gemcitabine can be further reduced, which could clinically mean a great reduction of toxicity. Indeed, if we consider other combinations of AZD7762 with a DDA (e.g. Cisplatin or 5-Fluorouracil), where the effective dose of the DDA was much lower (i.e. these DDAs were much less potent by themselves) (Supplementary Figure 5.5B), we observe more synergy in *SLFN11* high lines (Supplementary Figure 5.5C&D).

TRAMETINIB (MEK) + TASELISIB (PI3Kα)

Despite frequently showing synergy, we observed only one borderline significant association with Taselisib+Trametinib (ZNF486 expression, q-value: 0.04) and no associations with Trametinib+Taselisib. Combining these two configurations into one (annotating a cell line as synergistic when it was synergistic in at least one of the two configurations) also did not lead to any significant associations.

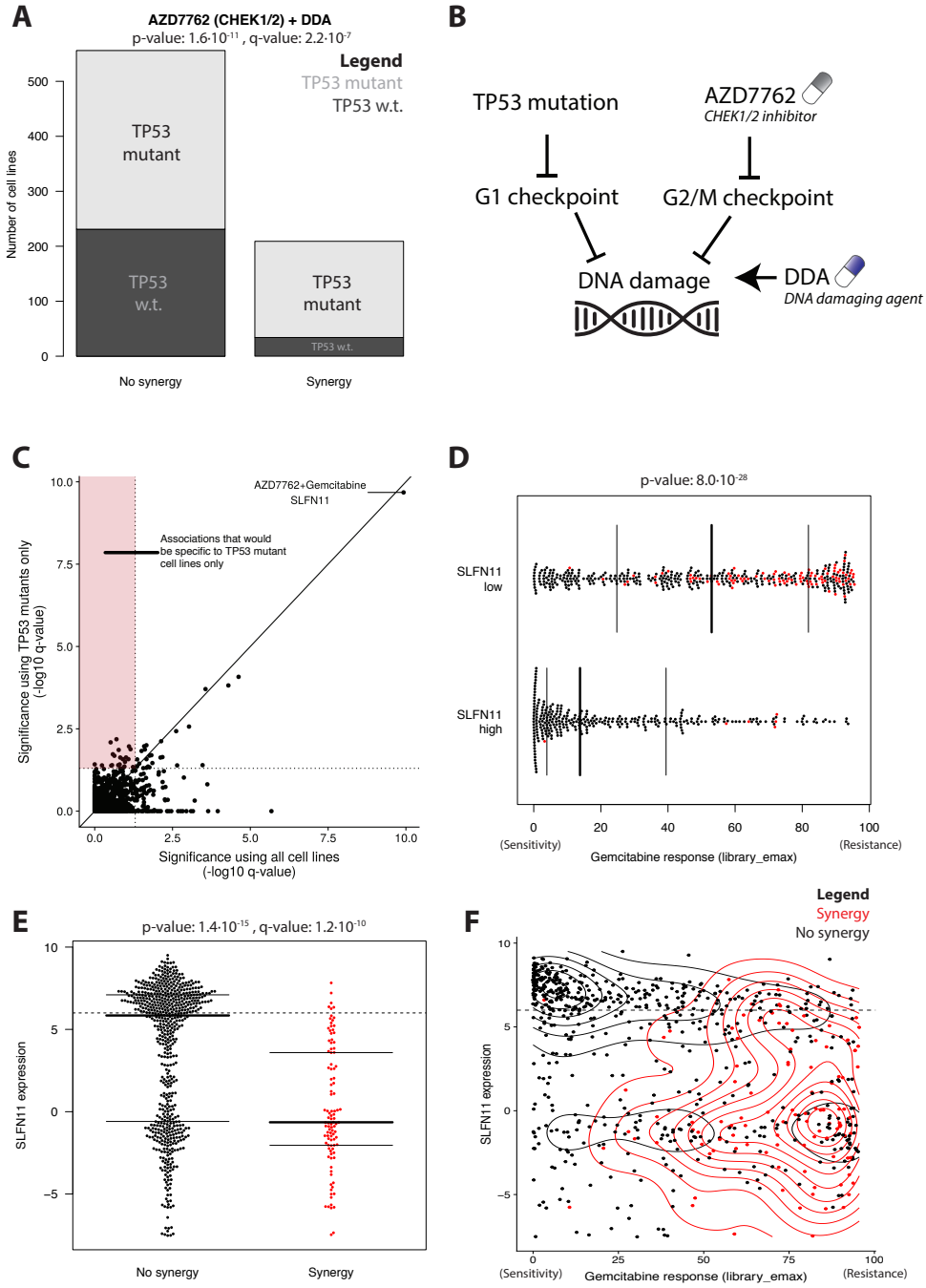


Figure 5.7: Associations between molecular features and AZD7762+DDA. (A) Association between TP53 mutation and synergy to AZD7762+DDA. (B) Mechanism through which TP53 may modulate the synergy. (C) Associations with synergy to AZD7762+DDA combinations using either all cell lines or TP53 mutants only. The dashed lines indicate the significance threshold at q=0.05. (D-F) Association between *SLFN11* expression, Gemcitabine response and synergy to the AZD7762+Gemcitabine combination. Red dots indicate synergistic cell lines. The dashed line in panel E and F indicates the threshold used for *SLFN11* in panel D.

Most strikingly, neither aberrations in the MAPK pathway (BRAF or RAS mutations), nor aberrations in the PI3K pathway (PI3K, AKT, or MTOR mutations; PTEN mutations or loss) were associated with synergy. Of the 74 cell lines that had a driver aberration in both pathways, only 10 showed synergy to at least one of the two combinations, hence combining markers between pathways does not appear to be predictive of synergy either. Altogether, this may indicate that while these mutations are essential in driving the MAPK and PI3K pathways, they do not modulate synergy.

Instead, we should likely expect genes that modulate the interaction between these two pathways to be good biomarkers of synergy. There are many places in which cross-pathway signaling can occur between these pathways [12], and this synergy may hence not be modulated at a single location. Such heterogeneity of mechanisms would make it much harder to find biomarkers and may explain why we do not find strong associations for this combination.

5.3. DISCUSSION

In the associations that we have highlighted, we observe different relationships between monotherapy and synergy. For example, for Olaparib (PARP) + Temozolomide (DDA), *SLFN11* and *NQO1* were associated with both monotherapy response and synergy, and the effect of the monotherapy response was amplified in the synergy. On the other hand, for AZD7762 (CHEK1/2) + Gemcitabine (DDA) in *SLFN11* high lines, the synergy was masked by the effect of monotherapy. Finally, TP53 mutations were associated with AZD7762+DDA synergy, but not to monotherapy response of either drug. These examples show that monotherapy response is related to synergy, but that it is not a priori clear what the nature of that association is.

Recently, Palmer and Sorger (2017) reported that most drug combinations in the clinic were effective not due to synergy, but due to independent action: in part of the patient population one drug was effective, while in another part of the population another drug is effective. They suggest that this is likely due to a lack of good biomarkers, making it hard to identify which patients should receive a given treatment, in which case it can be an effective treatment strategy to give both drugs in combination. We think this is a very interesting approach for in vivo studies. For in vitro studies on the other hand, finding combinations that maximize cell kill, without a clear biomarker or mechanistic intuition, may very well lead to treatments that are highly toxic in patients. Therefore, in our view, in vitro studies should focus on identifying combination treatments that are specific, preferably with a biomarker and an understanding of the mechanism.

In order to facilitate the large scale of this screen, we have chosen for a relatively short assay duration: the cell lines in this screen have been exposed to the drug combinations for 72 hours. This means that we are limited to observing synergies that manifest themselves on a relatively short timescale, and that a combination that did not show synergy in our screen may still show synergy in a longer assay setup. For example, this is likely the reason that we only see a modest synergy for the combination of an EGFR inhibitor with a BRAF inhibitor (e.g. Lapatinib+Dabrafenib), which was previously reported to be synergistic in BRAF mutant colorectal and thyroid cancer [18]. Specifically, Prahallad et al. [18] show that upon inhibition of BRAF in these cancer types, the MAPK pathway is reactivated due to a feedback loop to EGFR. When an EGFR inhibitor is added to BRAF

inhibitor, the feedback loop is abrogated, resulting in synergy. This synergy was much stronger in longer-term experiments (Chong Sun, personal communication), suggesting that it takes time for this feedback loop to be fully activated. We postulate that such mechanisms may also affect other drug combinations in our screen.

Similarly, the scale of this screen was facilitated by using cell lines as a model system. Cell lines capture a very large part of tumor biology [6], but they cannot perfectly model patient tumors. Therefore, putative biomarkers of synergy should always be further tested, for example in patient-derived xenografts (PDXs), prior to being used in the clinic. In addition, we may be missing synergies that only occur in vivo, for example due to interactions with the immune system or the microenvironment.

While the associations of TP53 mutations with AZD7762+DDA and *SLFN11* with Olaparib+Temozolomide have been reported before [8, 11, 14, 23], the novelty of our results is that these associations generalize to a large set of tumor cell lines. *NQO1* is known to regulate NAD+ [15], which is a cofactor of PARP [2]. However, to the best of our knowledge, the link between *NQO1* and synergy to Olaparib+Temozolomide has not been described before. Importantly, we also show for this association that it generalizes to a large set of tumor cell lines.

Overall, we have performed a large-scale drug combination screen of 54 combinations and 765 cell lines. This allowed us to assess for which combinations synergy generalizes to a large collection of tumor cell lines. Indeed, we found that five combinations result in synergy in more than 10% of the cell lines, including combinations of AZD7762 (CHEK1/2) + DDA, Olaparib (PARP) + Temozolomide (DDA) and Trametinib (MEK) + Taselisib (PI3Kα). In addition, we have associated molecular data to the synergy, from which the strongest associations may be good candidate biomarkers. We find that high *SLFN11* and low *NQO1* expression are associated with synergy to Olaparib+Temozolomide, and that TP53 mutations are associated with synergy to AZD7762+DDA.

5.4. METHODS

5.4.1. CELL LINES

Cell lines were acquired from commercial cell banks. All lines were grown in RPMI (supplemented with 10% FBS, 1% Penicillin/Streptomycin, 1% Glucose, 1 mM Sodium Pyruvate) or DMEM/F12 media (supplemented with 10% FBS, 1% Penicillin/Streptomycin) at 37°C in a humidified atmosphere at 5% CO2. To prevent cross-contamination or misidentification, all lines were profiled using a panel of 95 SNPs (Fluidigm, 96.96 Dynamic Array IFC). Short tandem repeat (STR) analysis was also performed and cell line profiles were matched to those generated by the cell line repository.

5.4.2. COMPOUNDS

Compounds were sourced from commercial vendors. DMSO-solubilized compounds were stored at room temperature in a low humidity (<12%), low oxygen (<2.5%) environment using Storage Pods (Roylan Developments). Water-solubilized compounds were maintained at 4°C.

5.4.3. SCREENING

Cells were transferred into 1536-well plates in 7.5 μl of their respective growth medium using XRD384 (FluidX) dispensers. The seeding density was optimized to ensure that each cell line was in growth phase at the end of the assay. Assay plates were incubated at 37°C in a humidified atmosphere at 5% CO2 for 24 hours then dosed with the test compounds using an Echo555 (Labcyte), final DMSO concentration was typically 0.2%. Following dosing, plates were incubated and the drug treatment duration was 72 hours. To measure cell viability, 2.5 μl of CellTiter-Glo 2.0 (Promega) were added to each well and incubated at room temperature for 10 minutes, quantification of luminescence was performed using a Paradigm (Molecular Devices) plate reader.

Seeding density optimization was carried out prior to screening by preparing a serial dilution of cells across six seeding densities with a two-fold dilution step. The maximum density tested varied based on cell type; typically 5000 cells/well for suspension cells and 1250 cells/well for adherent cells. Each density was dispensed into 224 wells of a single 1536-well assay plate using a XRD384 (FluidX) dispenser. The plate was then incubated at 37°C in a humidified atmosphere at 5% CO2 for 96 hours. Cell number was quantified using CellTiter-Glo 2.0 (Promega). Optimal densities selected were required to be within the linear range of the assay.

5.4.4. ASSAY PLATE QUALITY CONTROL

All screening plates contained several negative control wells (untreated wells, $n = 6$; DMSO-treated wells, $n = 126$) and positive control wells (blanks, i.e. medium only wells, $n = 28$; Staurosporin treated wells, $n = 20$; and MG-132 treated wells, $n = 20$) distributed across the plate. We used these positive and negative control wells to test whether the plates meet defined quality control criteria. For instance, using the DMSO-treated negative controls and all three positive controls, we determined z-factors as follow:

$$zfactor = 1 - \frac{3(\sigma_P + \sigma_N)}{|\mu_P - \mu_N|}$$

With σ_N and σ_P the standard deviation of the negative and positive controls, and μ_N and μ_P the mean of the negative and positive controls, respectively. The z-factors were required to exceed a minimum threshold of 0.2. In addition, a maximum threshold of 0.18 was applied to the coefficient of variation (CV) of the DMSO-treated negative controls ($CV = \frac{\sigma_N}{\mu_N}$, with σ_N the standard deviation of negative controls and μ_N the mean of the negative controls). Plates that did not meet these requirements were excluded from the study.

5.4.5. CURVE FITTING

For each plate, the raw fluorescent intensity values were normalized to a relative viability scale (ranging from 0 to 1) using the positive and negative control values. For the positive control the blank wells ($n = 28$) were used, and for negative control both wells with medium only ($n = 6$) and wells with medium and DMSO ($n = 126$) were used. The following equation shows this normalization:

$$viability = \frac{intensity - PC}{NC - PC}$$

Where *intensity* represents the intensity of a well containing cancer cells treated with a particular compound, *PC* represents the mean intensity of the positive control wells, and *NC* represents the mean intensity of the negative control wells.

AThe viability data is derived from seven-point dose-response assays. To obtain dose-response curves from these, the seven concentrations were first mapped to a relative scale (x), in which each unit difference translates to a 2-fold concentration difference and in which 9 represents the maximum test concentration. All library drug dose-response curves were fitted as a 2-parameter sigmoid function:

$$f(x_{pos_{ij}}, x_{shape_i}, x) = \frac{1}{1 + e^{-\frac{x - x_{pos_{ij}}}{x_{shape_i}}}}$$

The inference of the parameters $x_{pos_{ij}}$ and x_{shape_i} was performed in a hierarchical mixed model using restricted maximum likelihood (REML) [20], with a cell line effect (i), a library drug effect (j) and a replicate effect (k). Following standard nomenclature, β refers to fixed effects and the b refer to random effects.

$$x_{pos_{ijk}} = \beta_1 + b_{1i} + b_{1ij} + b_{1ijk}$$
$$x_{shape_i} = \beta_2 + b_{2i}$$

On the cell line level, both x_{pos} and x_{shape} are estimated, while on the library drug and replicate level only x_{pos} was estimated. This model is an adapted version of the IC50 model used in Iorio et al. [6], which was found to improve the precision of the inferred IC50 values by borrowing strength across the entire data set [20].

The dose-response curves for the combinations were fitted in a similar way, but with two notable differences: 1) the cell line parameters (i.e. b_{1i} and b_{2i}) were obtained from the library drug fits; and 2) $f()$ was scaled to go from 0 up to the anchor viability (rather than from 0 to 1).

To assess the quality of the fits, we computed the root mean square error (RMSE).

$$RMSE = \sqrt{\frac{1}{7}\sum_{l=0}^{7}(f(x_{pos}, x_{shape}, x_l) - viability_l)^2}$$

Where l refers to a point on the dose-response curve. Curves with an $RMSE > 0.2$ were excluded from further analysis.

Using the fitted models, we determined the *IC50* and *emax* parameters for each dose response curve. The *IC50* was set equal to the x_{pos} (we note that here we determined the IC50 on a fold-change scale, rather than the concentration scale). The *emax* values

were computed using $f(x_{pos}, x_{shape}, x)$ (or the rescaled version for drug combinations) with the estimated xpos and xshape parameters and with x set to the maximum test concentration ($x = 9$). For the library curves, we call these parameters *library_IC50* and *library_emax*. For the combination curves, we call these parameters *combi_IC50* and *combi_emax*.

To use the *IC50* and *emax* parameters to detect synergy, we first determined what the combination curves would look like under additivity (i.e. in the absence of synergy). To this end, we used Bliss independence [3] of the response to the anchor and the library drug alone. Conceptually, every point on the Bliss dose-response curve is defined as the product between the anchor viability and the corresponding point on the library dose-response curve. More specifically, the *Bliss_IC50* equals the *library_IC50*, and the *Bliss_emax* equals the product between the anchor viability and the *library_emax*.

For each cell line, drug combination and replicate, synergy was quantified using three scores:

$$\Delta IC50 = Bliss\_IC50 - combi\_IC50$$
$$\Delta emax = Bliss\_emax - combi\_emax$$
$$zscore = \frac{Bliss\_IC50 - combi\_IC50}{\sigma}$$

Where σ is the variance estimate for b_{1ijk} (i.e. the replicate level x_pos parameter) as estimated in the mixed-effect model.

5.4.6. SYNERGY CLASSIFICATION

We considered a given cell line, drug combination, anchor concentration and replicate as synergistic when:

- The *zscore* was greater than 3, roughly corresponding to a p-value < 0.001;
- The anchor viability was greater than 50%, since it is much harder to accurately determine the combination IC50 parameter (and hence the $\Delta IC50$ and *zscore* parameters) when the anchor viability is below this threshold; and
- Either the $\Delta IC50$ or the $\Delta emax$ was above a specific threshold.

The $\Delta emax$ threshold was determined by considering all $\Delta emax$ observations (where each observation corresponds to a particular cell line, drug combination and anchor concentration, replicates were averaged), from which we determined the mean $\mu_{\Delta emax}$ and the standard deviation $\sigma_{\Delta emax}$. The $\Delta emax$ threshold was set to $\mu_{\Delta emax} + 3\sigma_{\Delta emax}$, corresponding to a 33% reduction in viability (Supplementary Figure 5.5A).

The $\Delta IC50$ threshold was defined in a similar way, but we separated two cases. In the first case, when the *library_IC50* was observed within the screened range, we have more certainty about the strength of a given synergy. In the second case, when the *library_IC50* was outside of the screening range and hence had to be extrapolated, we want to be more conservative in calling synergy. To this end, we defined two $\Delta IC50$ thresholds: a low $\Delta IC50$ threshold and a high $\Delta IC50$ threshold. To determine the low $\Delta IC50$ threshold,

we first determined the mean $\mu_{\Delta IC50,low}$ and the standard deviation $\sigma_{\Delta IC50}, low$ using all observations (i.e. corresponding to a cell line, drug combination and anchor concentration) for which the *library_IC50* was within the screening range. We then set the low $\Delta IC50$ threshold to $\mu_{\Delta IC50,low} + 3\sigma_{\Delta IC50,low}$, corresponding to a shift of 5.2 fold-changes in *IC50* (or equivalently, a $2^{5.2}$ = 37-fold difference in concentration) (Supplementary Figure 5.5B). Similarly, using all observations for which the *library_IC50* was outside of the screening range, we defined the high $\Delta IC50$ threshold as $\mu_{\Delta IC50,high} + 3\sigma_{\Delta IC50,high}$, corresponding to a shift of 5.6 fold-changes in IC50 (or equivalently, a $2^{5.6}$ = 49-fold difference in concentration) (Supplementary Figure 5.5C).

We considered a cell line, drug combination and replicate synergistic when we the above approach resulted in synergy for at least one of the two anchor concentrations. Finally, when at least half of the replicates resulted in synergy, we considered a given cell line - drug combination pair as synergistic.

5.4.7. REPRODUCIBILITY

To assess the reproducibility of the screen, we rescreened 20 combinations and 736 cell lines. For the cell line - drug combination pairs that were screened in both cases, we determined synergy as described above. Of note, to ensure comparability, we used the same $\Delta IC50$ and $\Delta emax$ thresholds for both the original and the validation data set, by defining them using the entire dataset (i.e. the concatenation of the original and the validation data).

We determined the reproducibility by considering how many of the cell line - drug combination pairs that were synergistic in the original set were also synergistic in the validation screen. More specifically, we divide the number of cell line - drug combination pairs that showed synergy in both the original and the validation screen by the number of cell line - drug combination pairs that showed synergy in the original screen.

To increase the stability of our estimate, we also performed the analysis with the original and the validation set swapped. The final reproducibility rate was the average of the two analyses.

To determine the reproducibility rate expected by chance, consider poriginal the frequency at which we observe synergy for a given combination in the original screen and pvalidation the frequency at which we observe synergy for this combination in the validation screen. In addition, let us assume that the synergies in the original and the validation sets have been determined completely randomly. Then our expected reproducibility rate is:

$$\frac{p_{original} \cdot p_{validation}}{p_{original}} = p_{validation}$$

Hence, this would mean that for a drug combination that shows synergy in 11% of the cell lines, we would also expect at least a 11% reproducibility rate. We note that we observe reproducibility rates well above this expected rate (e.g. Olaparib+Temozolomide shows synergy in 11% of the cell lines and has a reproducibility rate of 58%) (Figure 5.3).

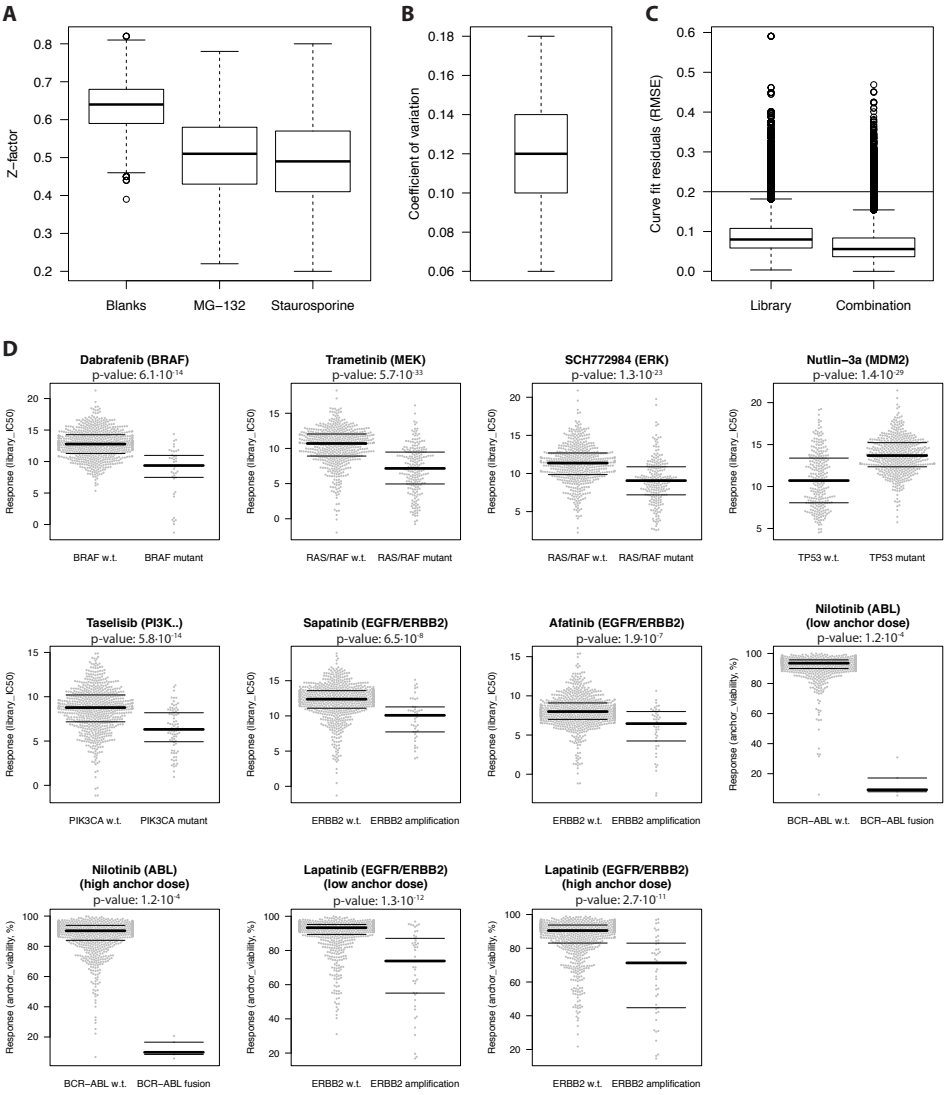
5.4.8. BIOMARKER ANALYSES

To identify putative biomarkers, we associated molecular features (mutations, CNA, gene expression and proteomics) with synergy. Binary synergy calls were obtained using all available data, i.e. original and validation screen combined. For the mutation and CNA data we used the Cancer Functional Events (CFEs), a list of putative driver events, as defined in Iorio et al. [6]. This resulted in 300 mutation features for 765 cell lines and 425 CNA features for 761 cell lines. For the gene expression data, we used Voom-transformed [9] RNAseq counts from 765 cell lines and 18,672 genes. Finally, we sourced RPPA-based proteomics data for 291 cell lines and 452 proteins from MCLP [10].

In addition to the 54 screened drug combinations, we defined a number of pseudo-combinations. For instance, we combined the response of Trametinib+Taselisib and Taselisib+Trametinib into the pseudo-combination MEK+PI3Kα, for which we considered a cell line as synergistic when it showed synergy to at least one of the underlying drug combinations. Similarly, we defined the pseudo-combination AZD7762+DDA as the combined response of AZD7762+Gemcitabine, AZD7762+Cisplatin and AZD7762+5-Fluorouracil. Finally, for each of the AZD7762+DDA combinations, we defined additional pseudo-combinations in which we only included TP53 mutant cell lines.

For each combination (or pseudo-combination) and each molecular feature, we then evaluated association between the molecular feature and the synergy status using either a Mann-Whitney U test (for binary features, such as mutations) or a Fisher exact test (for continuous features, such as gene expression). We performed this analysis both pan-cancer (across all cancer types) and per-cancer (per cancer type). The resulting 16,512,528 p-values were FDR corrected using a Benjamini-Hochberg correction.

5.5. SUPPLEMENTARY FIGURES



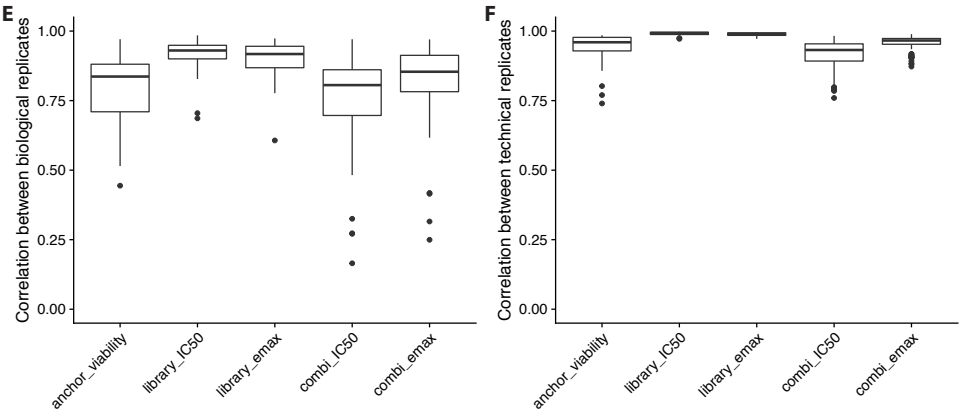


Figure S5.1: Quality assessment of the data. (A) Z-factor distributions across all 2,972 assay plates calculated using three positive control sets: media only (blanks), MG-132 and Staurosporine (positive controls). (B) Distribution of the coefficient of variation for negative control (DMSO) wells across all assay plates. (C) Curve fit residuals (RMSE) across all 167,213 dose-response experiments, of which 2% had an RMSE > 0.2 and were hence filtered out from subsequent analyses. (D) Expected monotherapy associations. The p-values were determined using a Mann-Whitney U test. (E&F) Correlation between (E) biological replicates and (F) technical replicates.

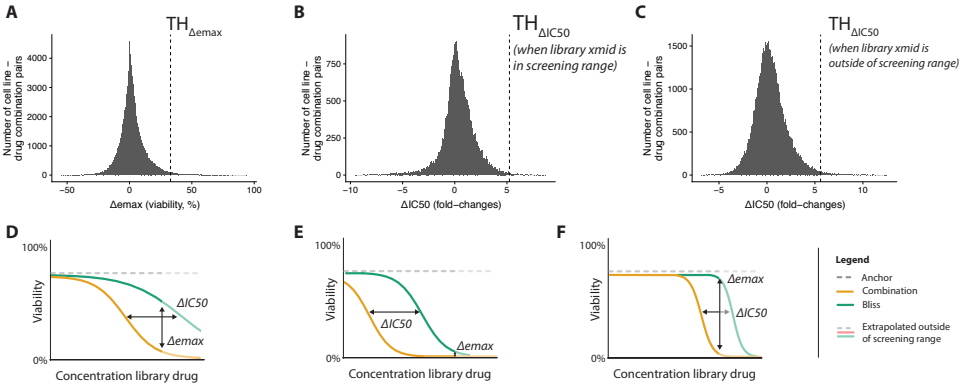


Figure S5.2: The Δ*IC50* and the Δ*emax*. (A-C) Distributions of Δ*IC50* and Δ*emax*, along with the thresholds used to determine synergy. (A) Distribution of Δ*emax* values. (B) Distribution of Δ*IC50* values when the *library_IC50* is within the screening range. (C) Distribution of Δ*IC50* values when the *library_IC50* is outside of the screening range. (D-E) Cartoons of different types of synergy: (D) one that passes both the Δ*IC50* and the Δ*emax* threshold, (E) one that passes only the Δ*IC50* threshold, and (F) one that passes only the Δ*emax* threshold.

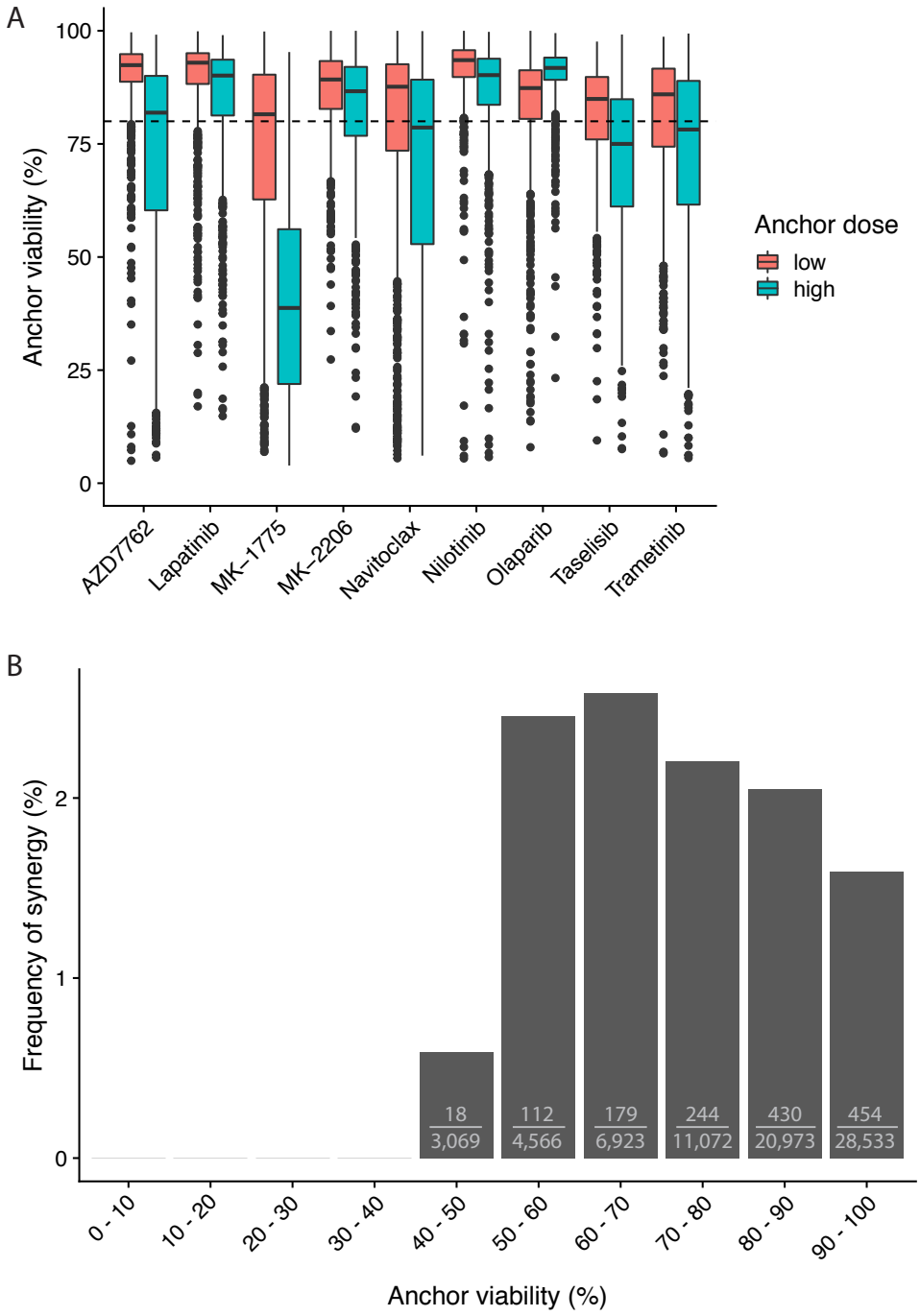


Figure S5.3: Anchor viabilities. (A) Distributions of obtained anchor viability, stratified by anchor drugs, as well as low- and high-dose. (B) Frequency of observing synergy stratified by different intervals of anchor viability. In each bin, this percentage is also represented as a fraction, showing the number of synergistic observations over the total number of observations. An observation corresponds to a given cell line, drug combination and anchor concentration.

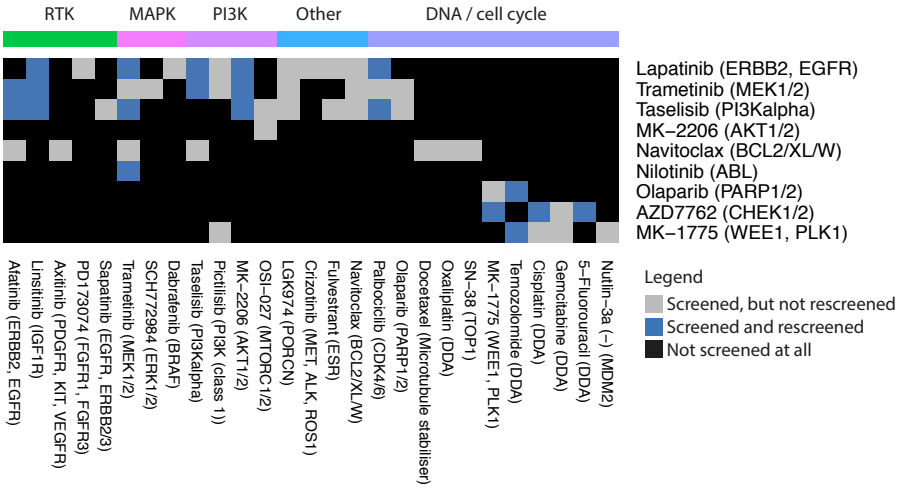


Figure S5.4: Heatmap with anchor drugs on the rows and library drugs on the columns, where the color indicates whether a given combination has been screened and whether it has been re-screened in the validation screen.

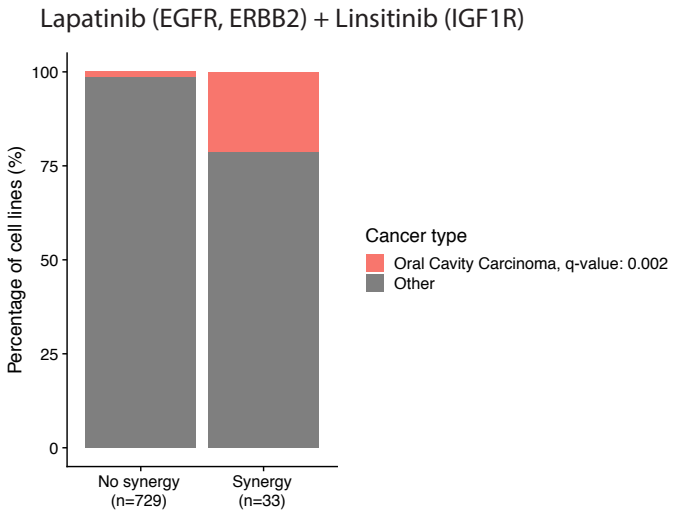


Figure S5.5: The statistically significant association between Oral Cavity Carcinoma and synergy to the Lapatinib (EGFR, ERBB2) + Linsitinib (IGF1R) combination.

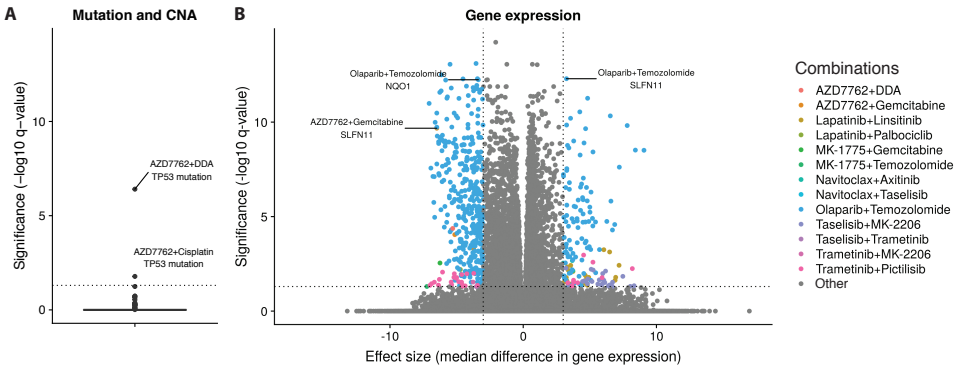
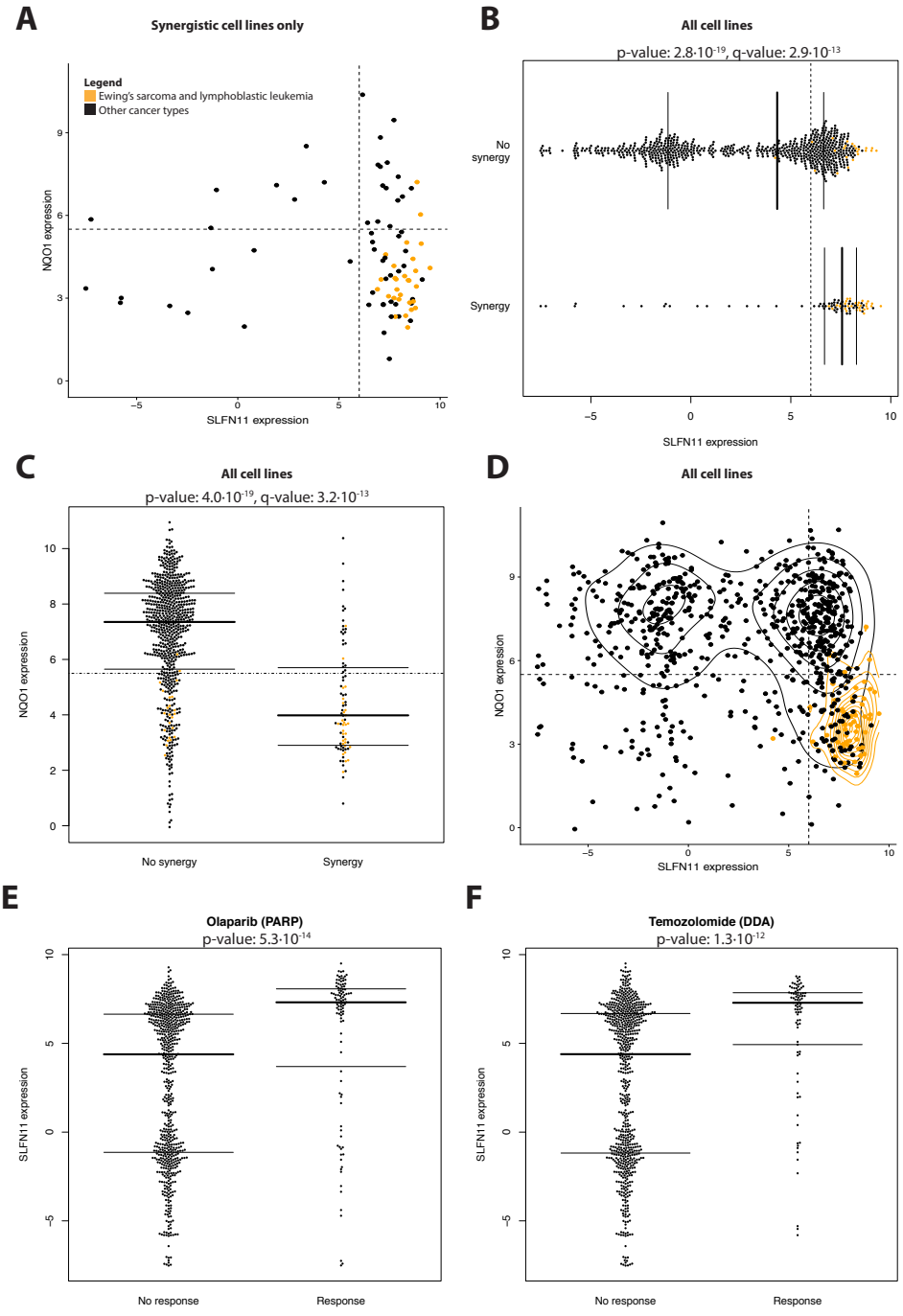


Figure S5.6: Associations between molecular data and synergy using all drug combinations. (A) Pan-cancer associations for mutation and CNA data. (B) Pan-cancer associations for gene expression data.


A
Synergistic cell lines only
Legend
Ewing's sarcoma and lymphoblastic leukemia
Other cancer types
NQO1 expression
SLFN11 expression
B
All cell lines
p-value: 2.8·10-19, q-value: 2.9·10-13
No synergy
Synergy
SLFN11 expression
C
All cell lines
p-value: 4.0·10-19, q-value: 3.2·10-13
NQO1 expression
No synergy
Synergy
D
All cell lines
NQO1 expression
SLFN11 expression
E
Olaparib (PARP)
p-value: 5.3·10-14
SLFN11 expression
No response
Response
F
Temozolomide (DDA)
p-value: 1.3·10-12
SLFN11 expression
No response
Response

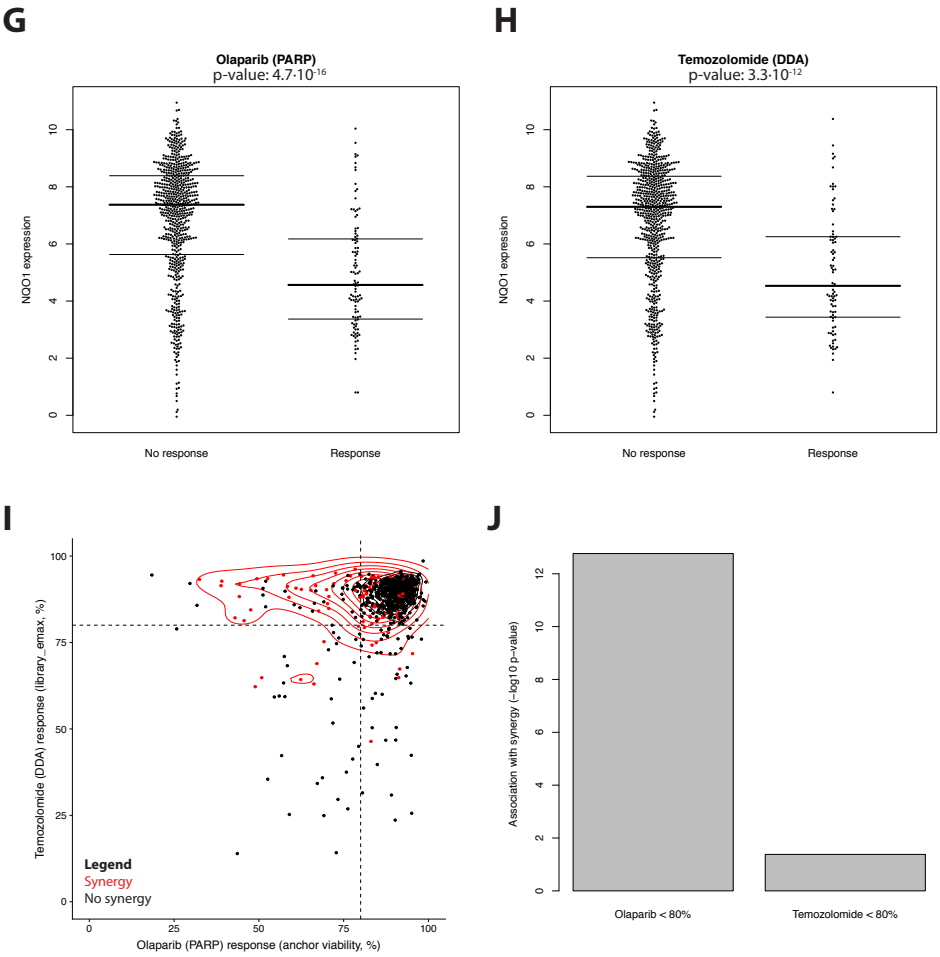


Figure S5.7: Associations between molecular features and synergy to the combination Olaparib (PARP) + Temozolomide (DDA). (A-D) Relationships between *SLFN11* expression, *NQO1* expression, cancer types and synergy to Olaparib+Temozolomide. In all four panels, orange dots represent Ewing's sarcoma and lymphoblastic leukemia cell lines. Panel A specifically shows synergistic cell lines, while Panel B-D show all cell lines. (E-H) Expression of either *SLFN11* or *NQO1* vs. response to either Olaparib or Temozolomide. Response was defined as anchor viability < 80% for Olaparib and *library_emax* < 80% for Temozolomide. (I) Olaparib response vs. Temozolomide response. Olaparib response was determined by taking the mean anchor viability across both anchor doses. The dashed lines indicate the 80% response threshold that is also used in panel E-H. (J) Association between synergy and either response to Olaparib or response to Temozolomide. P-values were determined using a Fisher exact test.

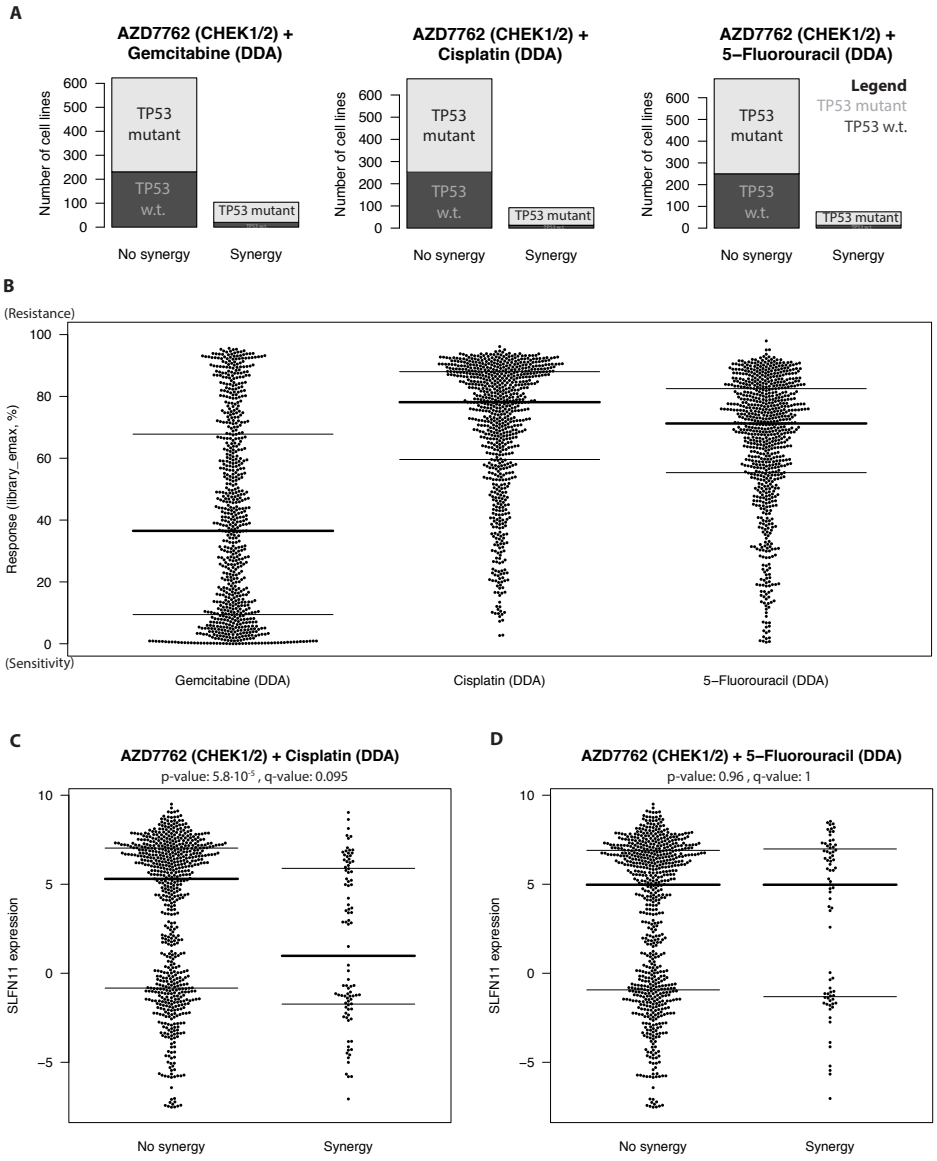


Figure S5.8: Associations between molecular features and synergy to the combination AZD7762+DDA. (A) Association between TP53 mutation and synergy to each of the AZD7762+DDA combinations. (B) Distributions of monotherapy response (*library_emax*) for the DDA in each of the AZD7762+DDA combinations. (C&D) *SLFN11* expression stratified by synergistic response to (C) AZD7762+Cisplatin and (D) AZD7762+5-Fluorouracil.

REFERENCES

[1] Nanne Aben, Daniel J Vis, Magali Michaut, and Lodewyk Fa Wessels. Tandem: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics*, 32(17):i413–i420, 2016.

[2] Nathan A Berger. Poly (adp-ribose) in the cellular response to dna damage. *Radiation research*, 101(1):4–15, 1985.

[3] CI Bliss. The toxicity of poisons applied jointly 1. *Annals of applied biology*, 26(3):585–615, 1939.

[4] Sonja J Gill, Jon Travers, Irina Pshenichnaya, Fiona A Kogera, Syd Barthorpe, Tatiana Mironenko, Laura Richardson, Cyril H Benes, Michael R Stratton, Ultan McDermott, et al. Combinations of parp inhibitors with temozolomide drive parp1 trapping and apoptosis in ewing's sarcoma. *PLoS One*, 10(10):e0140988, 2015.

[5] Emma J Haagensen, Huw D Thomas, Wolfgang A Schmalix, Andrew C Payne, Lara Kevorkian, Rodger A Allen, Paul Bevan, Ross J Maxwell, and David R Newell. Enhanced anti-tumour activity of the combination of the novel mek inhibitor wx-554 and the novel pi3k inhibitor wx-037. *Cancer chemotherapy and pharmacology*, 78(6):1269–1281, 2016.

[6] Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, 2016.

[7] Xinhui Kou, Yonghua Yang, Xiaoxiao Jiang, Huijuan Liu, Fanghui Sun, Xuan Wang, Longkai Liu, Hongrui Liu, Zhaohu Lin, and Lan Jiang. Vorinostat and simvastatin have synergistic effects on triple-negative breast cancer cells via abrogating rab7 prenylation. *European journal of pharmacology*, 813:161–171, 2017.

[8] Heather J Landau, Samuel C McNeely, Jayasree S Nair, Raymond L Comenzo, Takashi Asai, Hillel Friedman, Suresh C Jhanwar, Stephen D Nimer, and Gary K Schwartz. The checkpoint kinase inhibitor azd7762 potentiates chemotherapy-induced apoptosis of p53-mutated multiple myeloma cells. *Molecular cancer therapeutics*, 2012.

[9] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2):R29, 2014.

[10] Jun Li, Wei Zhao, Rehan Akbani, Wenbin Liu, Zhenlin Ju, Shiyun Ling, Christopher P Vellano, Paul Roebuck, Qinghua Yu, A Karina Eterovic, et al. Characterization of human cancer cell lines by reverse-phase protein arrays. *Cancer Cell*, 31(2):225–239, 2017.

[11] Zhikun Ma, Guoliang Yao, Bo Zhou, Yonggang Fan, Shegan Gao, and Xiaoshan Feng. The chk1 inhibitor azd7762 sensitises p53 mutant breast cancer cells to radiation in vitro and in vivo. *Molecular medicine reports*, 6(4):897–903, 2012.

[12] Michelle C Mendoza, E Emrah Er, and John Blenis. The ras-erk and pi3k-mtor pathways: cross-talk and compensation. *Trends in biochemical sciences*, 36(6):320–328, 2011.

[13] Wei Mo, Qingxin Liu, Curtis Chun-Jen Lin, Hui Dai, Yang Peng, Yulong Liang, Guang Peng, Funda Meric-Bernstam, Gordon B Mills, Kaiyi Li, et al. mtor inhibitors suppress homologous recombination repair and synergize with parp inhibitors via regulating suv39h1 in brca-proficient triple-negative breast cancer. *Clinical Cancer Research*, 22(7):1699–1712, 2016.

[14] Junko Murai, Ying Feng, K Yu Guoying, Yuanbin Ru, Sai-Wen Tang, Yuqiao Shen, and Yves Pommier. Resistance to parp inhibitors by slfn11 inactivation can be overcome by atr inhibition. *Oncotarget*, 7(47):76534, 2016.

[15] Gi-Su Oh, Hyung-Jin Kim, Jae-Hyuck Choi, AiHua Shen, Seong-Kyu Choe, Anzani Karna, Seung Hoon Lee, Hyang-Jeong Jo, Sei-Hoon Yang, Tae Hwan Kwak, et al. Pharmacological activation of nqo1 increases nad+ levels and attenuates cisplatin-mediated acute kidney injury in mice. *Kidney international*, 85(3):547–560, 2014.

[16] Laurence H Pearl, Amanda C Schierz, Simon E Ward, Bissan Al-Lazikani, and Frances MG Pearl. Therapeutic opportunities within the dna damage response. *Nature Reviews Cancer*, 15(3):166, 2015.

[17] Christian Posch, Homayoun Moslehi, Luzviminda Feeney, Gary A Green, Anoosheh Ebaee, Valentin Feichtenschlager, Kim Chong, Lily Peng, Michelle T Dimon, Thomas Phillips, et al. Combined targeting of mek and pi3k/mtor effector pathways is necessary to effectively inhibit nras mutant melanoma in vitro and in vivo. *Proceedings of the National Academy of Sciences*, 110(10):4015–4020, 2013.

[18] Anirudh Prahallad, Chong Sun, Sidong Huang, Federica Di Nicolantonio, Ramon Salazar, Davide Zecchin, Roderick L Beijersbergen, Alberto Bardelli, and René Bernards. Unresponsiveness of colon cancer to braf (v600e) inhibition through feedback activation of egfr. *Nature*, 483(7388):100, 2012.

[19] M Puglisi, P Thavasu, A Stewart, JS de Bono, MER O'Brien, S Popat, J Bhosle, and U Banerji. Akt inhibition synergistically enhances growth-inhibitory effects of gefitinib and increases apoptosis in non-small cell lung cancer cell lines. *Lung Cancer*, 85(2):141–146, 2014.

[20] Daniel J Vis, Lorenzo Bombardelli, Howard Lightfoot, Francesco Iorio, Mathew J Garnett, and Lodewyk FA Wessels. Multilevel models improve precision and speed of ic50 estimates. *Pharmacogenomics*, 17(7):691–700, 2016.

[21] Terence M Williams, Athena R Flecha, Paul Keller, Ashwin Ram, David Karnak, Stefanie Galbán, Craig J Galbán, Brian D Ross, Theodore S Lawrence, Alnawaz Rehemtulla, et al. Co-targeting mapk and pi3k signaling with concurrent radiotherapy as a strategy for the treatment of pancreatic cancer. *Molecular cancer therapeutics*, pages molcanther–0098, 2012.

[22] Masafumi Yamashita, Hiroshi Wada, Hidetoshi Eguchi, Hisataka Ogawa, Daisaku Yamada, Takehiro Noda, Tadafumi Asaoka, Koichi Kawamoto, Kunihito Gotoh, Koji Umeshita, et al. A cd13 inhibitor, ubenimex, synergistically enhances the effects of anticancer drugs in hepatocellular carcinoma. *International journal of oncology*, 49(1):89–98, 2016.

[23] Sonya D Zabludoff, Chun Deng, Michael R Grondine, Adam M Sheehy, Susan Ashwell, Benjamin L Caleb, Stephen Green, Heather R Haye, Candice L Horn, James W Janetka, et al. Azd7762, a novel checkpoint kinase inhibitor, drives checkpoint abrogation and potentiates dna-targeted therapies. *Molecular cancer therapeutics*, 7(9):2955–2966, 2008.

[24] Gabriele Zoppoli, Marie Regairaz, Elisabetta Leo, William C Reinhold, Sudhir Varma, Alberto Ballestrero, James H Doroshow, and Yves Pommier. Putative dna/rna helicase schlafen-11 (slfn11) sensitizes cancer cells to dna-damaging agents. *Proceedings of the National Academy of Sciences*, 109(37):15030–15035, 2012.

6

DISCUSSION

6.1. REFLECTIONS ON TANDEM

6.1.1. WHICH TYPES OF DATA SHOULD WE MEASURE TO PREDICT DRUG RESPONSE?

In Chapter 2 we have shown that gene expression is the most predictive dataset for the prediction of drug response. Moreover, we have shown that adding other datasets (e.g. mutation, CNA, methylation and cancer type data), does not further increase the predictive performance, as this information is already contained in the gene expression data. Given this result, one might be tempted to think that gene expression profiles are sufficient, and that a significant cost reduction can be achieved by only recording gene expression profiles.

If one is only interested in predictive performance, this could be a viable option. However, as also shown in Chapter 2, the other datasets (mutation, CNA, methylation and cancer type) are often easier to interpret in relation to the drug response. Hence, if we want to get insight into why certain cell lines respond to the drug, while others do not, collecting all of these datasets is very useful.

6.1.2. EXTENDING TANDEM BEYOND THE GAUSSIAN LINK FUNCTION

We initially formulated TANDEM for linear regression, i.e. regression with a Gaussian link function (predicting continuous values). Recall that in this setting we first explained as much as possible of the response using the upstream data, and then we explained the residuals from the first stage using the downstream data. We can easily show that the resulting TANDEM model is then also a linear model.

Specifically, consider $\mathbf{X}_1$ as the upstream data matrix, $\mathbf{X}_2$ as the downstream data matrix and y as the response vector. For ease of notation, let $\mathbf{X}_1$, $\mathbf{X}_2$ and y be (column-wise) mean-centered. The result of the first stage of TANDEM, employing a Gaussian link function, is then $\mathbf{X}_1\beta_1 \approx y$. Subsequently, in the second stage, we try to find β_2 such that $\mathbf{X}_2\beta_2 = y - \mathbf{X}_1\beta_1 + \epsilon$. We can rewrite this to the equivalent $\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 = y + \epsilon$, which clearly shows that the resulting TANDEM model is, in fact, a linear model.

Such a residual-based approach is much less elegant for other link functions. As an example, consider regression with a binomial link function, i.e. logistic regression (predicting binary values). For a binomial TANDEM model, in the first step, we find β_1 such that $\frac{1}{(1+e^{-\mathbf{X}_1\beta_1})} \approx y$. In the second step, we would then predict the residuals $y - \frac{1}{(1+e^{-\mathbf{X}_1\beta_1})}$. However, these residuals are continuous-valued, hence the second step would require a Gaussian link function. Our binomial TANDEM model would then be a concatenation of a binomial model and a Gaussian model. This is not a very elegant solution, and importantly, reduces the interpretability of the resulting regression coefficients.

Fortunately, there is a straightforward solution to this problem. Let us reconsider the earlier TANDEM example where we employed a Gaussian link function to predict continuous-valued outputs. In the second stage, instead of predicting the residuals from the first stage ($y - \mathbf{X}_1\beta_1$), we can set the offset (a constant value to be added per sample) to $\mathbf{X}_1\beta_1$ in glmnet (the community standard for Elastic Net regression). This ensures that in the second stage, we optimize β_2 in $\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 = y + \epsilon$, while keeping β_1 fixed. This formulation is equivalent to the formulation using residuals.

Now let us return again to the example with the binomial TANDEM model. In the first stage, we will again search for β_1 such that $\frac{1}{(1+e^{-\mathbf{X}_1\beta_1})} \approx y$. Then, fixing the offset parameter in glmnet to $\mathbf{X}_1\beta_1$, we can optimize $\beta_2$such that $\frac{1}{(1+e^{-\mathbf{X}_1\beta_1-\mathbf{X}_2\beta_2)}} = y + \epsilon$. The resulting model is clearly a binomial model and the resulting coefficients can hence be interpreted as such.

In a similar fashion, TANDEM can be extended for all other link functions available in the glmnet package: Poisson, Cox, multinomial (i.e. multi-class logistic regression) and mgaussian (i.e. multi-task learning, where y is a matrix of gaussian responses). These extensions have been implemented in the current version of the R package TANDEM.

6.1.3. SURVIVAL ANALYSIS

Extending TANDEM for Cox regression opens the door to the analysis of drug response in patient data, where response is frequently measured in survival time. This gives rise to several interesting questions.

TRANSFER LEARNING

For example, we could ask whether the predictive models trained on cell lines have any predictive power in patient tumors. This question can be addressed using transfer learning, in which a predictive model is transferred from one domain to a different (but related) domain.

A simple example of transfer learning is a setting in which we have labelled speech data for one speaker, but where we want to label the speech data from another speaker. Of course, we could could obtain labelled data for this new speaker, but this is very laborious, especially since we would have to do this every time we want to apply this model to a new speaker. Using transfer learning, we circumvent this problem by transferring the model that we have obtained based on the first speaker (the source domain) to our second speaker (the target domain). This is typically done by determining which part of the feature space is shared between both speakers, after which we can create a predictive model based on this shared space.

These concepts can also be applied to transfer predictive models from cell lines to patient tumors. In relation to the work in this thesis, Webber *et al.* [7] found that mutation and CNA data are more similar between cell lines and patient tumors than gene expression data, hence is would be very interesting to extend TANDEM models (in which these datasets are prioritized) for transfer learning.

ANALYZING TREATMENT RESPONSE IN PATIENTS WITH TANDEM

The features selected in a penalized Cox regression can be biomarkers for the treatment that the patients received, but they may also be general prognostic factors: factors that would have affected the survival regardless of whether or not the patient had received treatment. In other words, these prognostic factors would also predict survival in another set of patients that received a completely different therapy. Hence, they cannot be used as biomarkers to determine whether a patient should receive a given treatment. Therefore, we need to separate the prognostic factors from the potential biomarkers.

In the literature, this has been achieved in several ways [1, 4]. The approach from Li *et al.* [1] essentially fits a straightforward Cox regression model in which the survival is predicted from the gene expression, but with the addition of an interaction term for treatment.

$$\mathbf{X}_{ij}\beta_j + t_i\alpha + \mathbf{X}_{ij}t_i\gamma_j + b = y_i$$

Where i is the sample index, j is the feature index, $\mathbf{X}$ is the gene expression data, t is the treatment variable (encoded as 0 for no treatment and 1 for treatment), y is the response, β is the coefficient for the expression data, α is the coefficient for the treatment, γ is the coefficient for the interaction, and b is the intercept term. This way, the prognostic factors are captured by the main effects (i.e. the coefficients corresponding to the gene expression), whereas the putative biomarkers are captured by the interaction effects (i.e. the coefficients corresponding to the interaction between treatment and gene expression).

While simple, the above approach has one major caveat: in penalized regression, different definitions of the interaction term t lead to different results. For example, instead of encoding no treatment as 0 and treatment as 1, we may use -1 and 1, or 0 and 1. While in classical regression these are identical to each other, in penalized regression there are marked differences. It is currently an open question what the best way is to encode the interaction term.

That said, the above approach is quite elegant and it can easily be extended to a multi-omics approach (by simply including the other datasets, as well as their interactions with treatment) and, subsequently, to a TANDEM approach.

PATIENT-DERIVED XENOGRAFTS

Patient-Derived Xenografts (PDXs) are model systems where patient tumor material has been implanted into an immunocompromised mouse. Conceptually, these PDXs lie somewhere between cell lines and patient tumors, as, on the one hand, they more closely resemble human tumors as the transplanted tumors actually grow in a (modified) organism allowing the development of blood vessels and a tumor micro-environment, while on the other hand allowing systematic testing of multiple treatments (albeit at a smaller

scale than cell lines) in an in vivo setting. For the two scenarios described above (i.e. transfer learning and identifying biomarkers of treatment response), PDXs could potentially form a stepping stone between cell lines and patient tumors.

For the transfer learning, we could aim to transfer our models from cell lines to PDXs and then use the treatment response data to benchmark our approach. Such an approach would be almost impossible with patient tumor data, because of the very limited availability of treatment data, as well as confounding factors such patient age and overall health.

For the prediction of biomarkers of treatment response, we can use PDX data to test whether our stratification into prognostic factors and putative biomarkers of the treatment is correct. To this end, we would first fit our model, a Cox regression (TANDEM) model with an interaction term as described above, for a treatment of interest. We could then use data from untreated PDXs (something that would not be possible in patients) to assess whether the prognostic factors are associated with survival, while the putative biomarkers are not, thereby benchmarking our approach.

6.1.4. SELECTING THE PENALIZATION LEVEL

In glmnet, one is able to choose whether to use λ_{min} (i.e. the optimal lambda, corresponding to the lowest cross-validation error) or the λ_{1se} (i.e. a more sparse model, selecting fewer features, whose predictive performance is within one standard error of the minimum and is hence not statistically distinguishable from the λ_{min} model).

The TANDEM R package uses glmnet to optimize the Elastic Net regression in each step. Hence, in the TANDEM R package the user is able to specify whether he wants to use λ_{min} or λ_{1se} for the upstream data, and which one he would like to use for the downstream data. Which one should be chosen and how to choose these largely depend on the way the upstream data is related to the downstream data.

When the upstream data contains information (related to the drug response) that is not present in the downstream data (something that can be easily checked using iTOP), we recommend using λ_{min} for both the upstream and the downstream data. When doing so, the resulting TANDEM model will achieve the same predictive performance as an Elastic Net model using all datasets (both upstream and downstream) fitted with λ_{min}. Using λ_{1se} for both upstream and downstream data would lead to extra penalization twice, resulting in predictive performance that is lower than an Elastic Net model using λ_{1se}. Likewise, when using λ_{1se} either the upstream data or the downstream data (and λ_{min} for the other), no guarantees can be given on the predictive performance.

When the information relevant to the drug response in the upstream data is fully contained in the downstream data, one has more freedom in this choice of lambda. Like before, a model using λ_{min} for both the upstream and the downstream data results in the same predictive performance as an Elastic Net model using λ_{min}. But in this scenario, the same predictive performance can be achieved using λ_{1se} for the upstream data (to find the most relevant upstream features) and λ_{min} for the downstream data. The reason for this is that the variation that we leave unexplained by choosing λ_{1se} for the upstream data is fully contained in the downstream data. Hence, by using λ_{min} for the downstream data, we can still explain this variation, thereby achieving the same predictive performance as before.

6.1.5. ON SAMPLE SIZES

One important factor when choosing your model is the number of samples available for training. For example, as shown in Supplementary Figure 4.1, multi-task learning obtains the biggest increase in performance (compared to regular, single-task learning) when the number of samples is small.

On the other hand, TANDEM performs best when sample sizes are larger. When applied to hundreds of samples (934 samples in Chapter 2 and 206 samples in Chapter 3), we clearly see that the information shared between the upstream data and the drug response is captured in the gene expression. However, in a lower sample size setting we do not always see this pattern. For instance, in a study of 45 colorectal cancer cell lines [5], we found that we could predict some drugs well with mutation data, but that we could not always create predictive models for these drugs using gene expression. We suspect that this is due to the curse of dimensionality: when the sample size is limited, it may be easy to predict drug response using 38 mutation features, but hard to extract its corresponding gene expression signature from the ~17,000 dimensional gene expression dataset. As the sample size increases , this problem is alleviated.

6.2. REFLECTIONS ON iTOP

6.2.1. WHY IS GENE EXPRESSION SO DOMINANT IN CLASSIC APPROACH MODELS?

We have long wondered why the Elastic Net models from Chapter 2 were almost completely based on gene expression. Of course, we have seen that the information that is shared between the upstream data and the drug response is contained in the gene expression. This explains why gene expression alone is sufficient: adding the other datasets does not increase predictive performance. However, it does not explain why we almost never see models using upstream data. If such models achieve the same predictive performance, then what is the reason that we never see them?

Note that in Chapter 2 we have ruled out dimensionality and binary vs. continuous data as possible reasons (Supplementary Figure 2.1C). To further investigate the above question, let us consider a far more simple example. Consider the topology of variables in S6.1. Suppose we would perform Elastic Net regression on $\mathbf{X}$ and y, which feature(s) would be selected first?

$\mathbf{X} = [x_1, x_2, x_3]$
$x_1 = \varepsilon_1$
$x_2 = x_1 + \varepsilon_2$
$x_3 = x_1 + \varepsilon_3$
$y = x_2 + x_3 + \varepsilon_4$

With $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$ drawn from N(0,1)

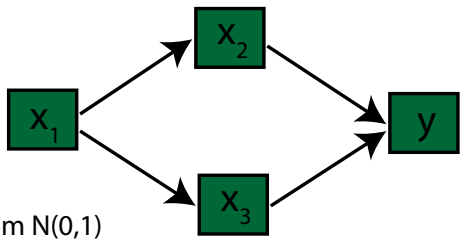


Figure S6.1: Thought experiment on causality and feature selection.

The answer to this question is that the two mediator variables (the ones in the mid-

dle) would be selected first. The reason for this is that Elastic Net works very much like forward feature selection and first selects variables that are most correlated to the response y.

Now let us reconsider Chapter 3, in which we have discussed iTOP, a methodology to infer topologies of relationships between datasets. Specifically, we have seen that gene expression is a mediator dataset in between the upstream datasets and the drug response. If we combine this observation with the observation that feature selection algorithms favor mediator variables, it now makes perfect sense that the Elastic Net models were almost completely based on gene expression, as these mediator variables are more correlated to the response.

Interestingly, the preferential selection of mediator variables is not limited to Elastic Net regression. Almost all feature selection algorithms (Elastic Net, forward feature selection, Random Forest, etc.) select variables that correlate most strongly to the response first.

6.2.2. ON CAUSAL RELATIONSHIPS BETWEEN DATASETS

In the previous section, we have seen that in the example in Figure 1 the mediator variables are selected first, for example using Elastic Net regression. Interestingly, in this example we also had information on the causal relationships between the variables, and we observed that the Elastic Net ended up selecting the causal variant last.

The observation that Elastic Net regression tends to not select causal variables has previously been made on a much larger scale by Maathuis *et al.* [2]. In their work, they compare their method IDA [3] (a method for causal inference on high-dimensional data) to Elastic Net on yeast data with transcriptomics of 234 single-gene deletion mutant strains as well as 63 wild-type strains. Specifically, for a given gene X, they use the wild-type strains to identify which other genes are associated with the natural variation in gene X (using either Elastic Net regression or IDA). They then validate whether these associated genes are functionally related to gene X by checking whether their expression changes in the strain where gene X is deleted. This lead them to conclude that IDA identifies more causally related genes than Elastic Net regression.

While IDA is clearly better at identifying causal relations, we had trouble applying it to drug response data due to its long running time. During the inference of the causality network, IDA uses stringent cut-offs on the partial correlations. If these cutoffs are too lenient, the number of partial correlations that need to be checked grows exponentially in the number of features and the problem becomes intractable. On the other hand, if these cutoffs are too stringent, (almost) all gene - drug relations will be filtered out. Searching for the right cut-off can be a very lengthy process, as we found that the algorithm often needs thousands of cpu hours per drug and per parameters setting (as opposed to roughly one minute for an Elastic Net regression).

As an alternative approach, we can consider causality between datasets rather than individual features. Specifically, if we have a good idea about which datasets are more causally related to the drug response, we can prioritize these datasets using TANDEM. For example, we can prioritize mutation data, because mutations are more likely to affect gene expression than the other way around.

In some cases we can also use a data-driven approach to determine which datasets

are more causally related to the drug response and could hence be prioritized using TANDEM. In Chapter 3, we have described how iTOP uses the PC algorithm (which is also employed in IDA to infer the feature causality network) to infer the topology between datasets. We have also shown how - under certain assumptions, namely that the causality graph is directed and acyclic and that all variables are observed - this algorithm can be used to infer causality between datasets. While for the data in Chapter 3 we were unable to infer any causal relationships, this is an interesting direction that can be further explored on other data.

We could take a similar approach to infer causality within a dataset. For example, we could run iTOP on subsets of the gene expression data, where each subset is based on a geneset representing a pathway or molecular process. This may allow us to search for causal relations between these pathways and/or molecular processes.

Another interesting future direction would be to incorporate prior knowledge into the causal inference, for example using the NPC algorithm [6]. This algorithm uses the same two steps as the PC algorithm (first inferring an undirected skeleton using partial correlations, after which it infers causality using that skeleton), but adds an intermediate step in which the user can indicate the directionality of some of the edges. For example, the user might indicate that gene expression affects drug response rather than the other way around, because in this case gene expression was measured before drug response. In many cases, this can make the causal inference of the remaining edges much easier.

REFERENCES

[1] L Li, T Guennel, S Marshall, and L WK Cheung. A multi-marker molecular signature approach for treatment-specific subgroup identification with survival outcomes. *The pharmacogenomics journal*, 14(5):439, 2014.

[2] Marloes H Maathuis, Diego Colombo, Markus Kalisch, and Peter Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247, 2010.

[3] Marloes H Maathuis, Markus Kalisch, Peter Bühlmann, et al. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

[4] Joske Ubels, Pieter Sonneveld, Erik H van Beers, Annemiek Broijl, Martin H van Vliet, and Jeroen de Ridder. Predicting treatment benefit in multiple myeloma through simulation of alternative treatment effects. *Nature communications*, 9, 2018.

[5] Theodoros I Roumeliotis, Steven P Williams, Emanuel Gonçalves, Clara Alsinet, Martin Del Castillo Velasco-Herrera, Nanne Aben, Fatemeh Zamanzad Ghavidel, Magali Michaut, Michael Schubert, Stacey Price, et al. Genomic determinants of protein abundance variation in colorectal cancer cells. *Cell reports*, 20(9):2201–2214, 2017.

[6] Harald Steck. *Constraint-based structural learning in Bayesian networks using finite data sets*. PhD thesis, Technische Universität München, 2001.

[7] James T Webber, Swati Kaushik, and Sourav Bandyopadhyay. Integration of tumor genomic data with cell lines using multi-dimensional network modules improves cancer pharmacogenomics. *Cell systems*, 7(5):526–536, 2018.

ACKNOWLEDGEMENTS

Lodewyk, you provided me with freedom to specify my own projects, while at the same time being very invested in the details of anything I chose to work on. This mix allowed me to develop my qualities as an independent researcher, something that will be very useful even now that I have chosen to work on a topic that is well outside of the bioinformatics field. Thank you for all the discussions, insights and suggestions!

Magali, I cannot imagine what this thesis would have looked like without you. You and Lodewyk made a great supervisory team, complementing each other in areas where one was not enough. In addition to our work-related meetings, I also fondly remember our conversations about music and creativity. Thank you for all of this!

Daniel, thank you for our extensive discussions on drug response, synergy, growth rates, office temperatures, the differences in color perception between men and women, and many, many more topics. I greatly appreciate your (scientific) insights, as well as your sense of humor.

Daniel, Magali and Lodewyk, I just thanked you individually, but I would like to thank you again for working with me on my first paper: TANDEM. Like many before me, the first paper was not the easiest to make. But judging by the reactions I have seen at conferences, I think we did an amazing job!

Julian, Evert, Yongsoo, Gergana, Daniel and Magali, together we participated the DREAM challenge for predicting anti-cancer drug synergy from molecular data. Of course, we expected to do well, given that this subject was very much in line with our own research (and especially with my PhD topic). And indeed, in the first four rounds, we did amazingly well, consistently scoring as the number one. Unfortunately, for reasons that are still unknown, we did not do well in the final round. Despite that, I really enjoyed working with you guys on this. In our field it is not common to take on a project with so many people, which is a pity, because it was a lot of fun to do!

Age and Johan, I met the two of you at BioSB. After my talk on TANDEM, we discussed ideas on data integration, and you invited me to further discuss these at the UvA. These discussions, where Yipeng and Henk later joined us, turned into a fruitful collaboration, where we jointly wrote quite a few papers. It's really great to have these technical, in-depth discussions with people outside of your own group, as it sparks all sorts of new ideas. iTOP is a great example where our joint expertise led to an excellent idea that is completely different from everything else out there.

Patricia, Lizzie, Syd, Howard, Donny, Mathew, for years we have been calling each other every second Wednesday, as part of a collaboration that lead to numerous papers, including Chapter 5 in this thesis. Besides these TCs, I also fondly remember the ERC Synergy CombatCancer meetings: every year we would go to Cambridge, and half a year later you would come to Amsterdam. This was always a lot of fun, with a strong scientific program and many informal discussions afterwards. Thank you for the many insightful scientific discussions, as well as the relaxed ones over beer afterwards.

Daniel, Jelle and Patty, it was great fun sharing an office with you during the first few years of my PhD. We had numerous discussions on countless subjects. Unfortunately, it was also a very busy office, with sometimes multiple phonecalls happening simultaneously and people constantly walking in to ask where Patty was (trust me, I don't keep track of that). So with pain in my heart, I decided to usurp Ewald's desk when he left, which brings me to...

Gergana, Tesa, Marlous and Marie, it was great fun sharing an office with you! Also thanks to the people who sat in the office 'part-time': Julian, Roebi, Maarten and Magali. It was quite a bit quieter than my first office, but by no means was it really quiet, as we too would often discuss and make jokes. I really laughed my ass off with you guys. Just a random selection (without replacement) out of my memories: you guys giving me a personalized coffee muck, really lame jokes about Tesa's back, naming the plants in the office (the only ones I remember, Bruce and Gandalf the Green, are the ones that died btw; is there a pattern in that?). Thank you for all of this, I will miss you!

Gergana, Tesa, Marlous, Evert, Maarten, Magali, Daniel Jelle, Patty, Julian, Marie, Roebi, Sander, Silvana, Kathy, Bram, Tycho, Ewald, Yongsoo, Andi, Jinhyuk, Joana, Kat, Abhishek, Joris, Soufiane, Elmer, Katja, Fredrick, Sergio, Nicola, Lodewyk, you guys made the Computational Cancer Biology an awesome place to be. Thank you for all the jokes and the stories, as well as the scientific discussions. Many of you taught me so much of what I know now, and I have also had the pleasure of passing on some of that knowledge to some of you. Thank you!

Chiara, Giusi, Ewald, Lorenzooitje, Jacobien, Philip, Marcelo, Monique, Melanie, Maarten, Max, Rocco, Tess, Lisanne, Anke, thank you for conversations that often had nothing to do with science, but could cover everything else, including:

- Where are we gonna have dinner?
- Shouldn't we form a band?
- Wanna grab a coffee?
- Champaign?
- Would you like to brew beer?

Sebas and Gerran, thanks for providing me with the necessary distractions from my PhD.

I will even forive you guys for naming our first album Synergy ;)

Weiyi, Gerard, Nicas, Thomas, Steve, Daniel, Maurits, Bart, Arnout, Barbara, John, Lianne, Emma, Willemijn, Rolf, Paul, Sanne, friends and family, we hung out for a huge variaty of reasons, ranging from parties to Dungeons & Dragons. All of this had, of course, nothing to do with the PhD, which was exactly the point. Thank you for all of that!

Diana, mi amor, Donaji, I'm so happy that, on the night that I was supposed to finish my ISMB poster, I instead met you. I could not imagine a better partner than you. Thank you for loving and supporting me while I finished this book! Te amo!!

MaM, jouw bijdrage aan dit werk is misschien wel de meest essentiële: zonder jou geen mij! Maar uiteraard gaat jouw rol veel verder dan dat, want je hebt me door dik en dun gesteund in mijn PhD. Ik hou van je, mam!

CURRICULUM VITÆ

Nanne Nicolaas ABEN

06-12-1988	Born in Schiedam, The Netherlands.

EDUCATION

2001–2007	Gymnasium Groen van Prinsterer, Vlaardingen
2007–2011	B.Sc. Computer Science Delft University of Technology
2010–2011	B.M.A. Pop Music (not completed) Conservatory of Rotterdam
2011–2014	M.Sc. Computer Science - track Bioinformatics Delft University of Technology

PROFESSIONAL CAREER

2011–2013	Research assistant (part-time, one day a week) VUmc / DZNE Tübingen
2014–2019	Ph.D. Computational Cancer Biology Netherlands Cancer Institute & Delft University of Technology

LIST OF PUBLICATIONS

16. **Aben, N.**, Jaaks, P., Vis, D.J., ..., Wessels, L.F.A. A screen of 765 cell lines and 54 drug combinations to study synergistic drug interactions in cancer. Manuscript in preparation.

15. Vis, D.J., Jaaks, P., **Aben, N.**, ..., Wessels, L.F.A. A critical evaluation of novel growth rate corrections. Manuscript in preparation.

14. Smilde, A.K., Song, Y., Westerhuis, J.A., Kiers, H.A.L., **Aben N.**, Wessels, L.F.A. Heterofusion: fusing genomics data of different measurement scales. Manuscript in preparation.

13. **Aben, N.**, de Ruiter, J., Bosdriesz, E., Kim, Y., Bounova, G., Vis, D., ... & Michaut, M. Identifying biomarkers of anti-cancer drug synergy using multi-task learning. Under review.

12. Brammeld, J., Price, S., Ranzani, M., Coker, E., Roumeliotis, T., ..., **Aben, N.**, McDermott, U. An integrated genetic screens approach identifies aberrant activation of Src signalling as a resistance bypass pathway in BRAF mutant colon cancer. Under review.

11. Menden, M. P., Wang, D., Guan, Y., Mason, M., Szalai, B., Bulusu, K. C., ... & Nguyen, T. (2019). A cancer pharmacogenomic screen powering crowd-sourced advancement of drug combination prediction. Nature Communications. (As part of the AstraZeneca-Sanger Drug Combination DREAM Consortium.)

10. Nagel, R., Avelar, A.T., **Aben, N.**, Proost, N., van de Ven, M., ..., Berns, A. (2019). Inhibition of the replication stress response is a synthetic vulnerability in SCLC that acts synergistically in combination with cisplatin. Molecular Cancer Therapeutics.

9. Song, Y., Westerhuis, J. A., **Aben, N.**, Wessels, L. F., Groenen, P. J., & Smilde, A. K. (2018). Generalized Simultaneous Component Analysis of Binary and Quantitative data. arXiv preprint arXiv:1807.04982.

8. **Aben, N.**, Westerhuis, J. A., Song, Y., Kiers, H. A., Michaut, M., Smilde, A. K., & Wessels, L. F. (2018). iTOP: Inferring the Topology of Omics Data. Bioinformatics.

7. Song, Y., Westerhuis, J. A., **Aben, N.**, Michaut, M., Wessels, L. F., & Smilde, A. K. (2017). Principal component analysis of binary genomics data. Briefings in Bioinformatics.

6. Ranzani, M., Kemper, K., Michaut, M., Krijgsman, O., **Aben, N.**, Iyer, V., ... & Turner, G. (2017). A screen for combination therapies in BRAF/NRAS wild type melanoma identifies nilotinib plus MEK inhibitor as a synergistic combination. bioRxiv, 195354.

5. Roumeliotis, T. I., Williams, S. P., Gonçalves, E., Alsinet, C., Velasco-Herrera, M. D. C., **Aben, N.**, ... & Wright, J. C. (2017). Genomic determinants of protein abundance variation in colorectal cancer cells. Cell reports, 20(9), 2201-2214.

4. Janssen, S., Schutz, C., Ward, A., Nemes, E., Wilkinson, K. A., Scriven, J., Huson, M. A., **Aben, N.**, ... & Meintjes, G. (2017). Mortality in severe human immunodeficiency virus-tuberculosis associates with innate immune activation and dysfunction of monocytes. Clinical Infectious Diseases, 65(1), 73-82.

3. Cuypers, B., Jacobsen, A., Siranosian, B., Schwahn, K., Conard, A. M., **Aben, N.**, ... & Meysman, P. (2016). Highlights from the ISCB Student Council Symposia in 2016. F1000Research, 5.

2. **Aben, N.**, Vis, D. J., Michaut, M., & Wessels, L. F. (2016). TANDEM: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. Bioinformatics, 32(17), i413-i420.

1. Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., **Aben, N.**, ... & Garnett, M. J. (2016). A landscape of pharmacogenomic interactions in cancer. Cell, 166(3), 740-754.