Dissertation
submitted to the
Combined Faculty of Natural Sciences and Mathematics
of the Ruperto Carola University Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Presented by
MI YANG, PharmD, M.Sc.
Born in Fuxin (China)
Oral examination: 12.11.2018

Applying machine learning to derive actionable insights in precision oncology

Referees: Prof. Dr. Benedikt Brors
Prof. Dr. Julio Saez-Rodriguez

Applying machine learning to derive actionable insights in precision oncology

MI YANG

Fakultät für Biowissenschaften

Universität Heidelberg

Abstract

Cancer drugs have among the lowest response rates across all diseases. Combining the wealth of omics data and machine learning is a promising way to reach this goal. In this thesis, we addressed the following aspects of precision oncology: (i) We used Macau, a bayesian multitask multi-relational algorithm to explore the associations between the drugs' targets and signaling pathways' activation. We applied this methodology to drug synergy prediction and stratification. (ii) We leveraged through a collaborative machine learning competition to understand the association between genome, transcriptome and proteome in tumors. The main focus of this thesis is to use machine learning to generate actionable insights, for more personalized therapies.

Zusammenfassung

Die Ansprechrate bei Krebstherapeutika ist im Vergleich zu anderen Arzneimitteln niedrig. Die Kombination aus Omics-Daten und maschinellem Lernen ist ein vielversprechender Weg um eine höhere Ansprechrate zu erlangen. In dieser Arbeit haben wir uns mit den folgenden Aspekten der Präzisions-Onkologie befasst: (i) Wir verwendeten Macau, einen multi-relationalen bayesianischen Multitasking-Algorithmus, um die Assoziation zwischen den Zielproteinen und der Aktivierung von Signalwegen zu untersuchen. Diese Methode haben wir zur Vorhersage und Stratifizierung von Synergien zwischen Medikamenten angewendet. (ii) Wir nutzten einen kollaborativen Wettbewerb zum maschinellen Lernen, um die Assoziation zwischen Genom, Transkriptom und Proteom in Tumoren zu verstehen. Der Schwerpunkt dieser Arbeit liegt auf dem Gewinnung wertvoller Erkenntnisse für personalisierte Therapien mit Hilfe von maschinellem Lernen.

Acknowledgment

4 years ago, I decided to become a scientist. It was at the end of my last year of Pharmacy school at the University of Paris Sud 11, that I made the decision to switch into the emerging field of systems biology. I learnt about this field through a friend of mine: Amel Camelia Bencherif, whose team just won the Grand Prize of the international Genetically Engineered Machine competition. It was her enthusiasm and absolute confidence that awakened my passion for scientific discoveries. In the same year, I met Prof. Andre Tartar, from the University of Lille. It was the first time I've seen such extent of knowledge and skills, which ultimately awakened my interest for research in drug development. These were the key people leading to my decision to switch career.

I decided to have my first research experience at the University of California San Francisco, where the biopharmaceutical industry originated from. My supervisor Professor C. Anthony Hunt was doing extremely innovative research, which put him against the whole scientific community. It was the first time I've witnessed such courage and independence of mind. Although we did not agree on everything, I was truly impressed and grateful that this was my first initiation to research in systems biology.

In the Bay area, I took some valuable advice from Prof. Atul J. Butte, from Stanford/UCSF. This short mentorship was the foundation of my mid term career plan, which combines my thirst for knowledge, social responsibility as a healthcare professional and the desire to bring discoveries to the patients.

I started my PhD with Prof. Julio Saez-Rodriguez in a strange state of mind. My vision was clear and objectives perfectly defined for the next few years. I took advantage of this big and diverse group to strengthen my research and bioinformatics skills. I enjoyed absolute freedom of subjects, which made this PhD extremely productive, although only a small part is presented in this thesis. I can never be grateful enough for being part of this lab.

At the beginning of my PhD, I started working on machine learning with a fellow PhD student, Michael P. Menden. He has been a great technical and scientific mentor and he still is. During my PhD, I collaborated with Jaak Simm, from KU Leuven, on a tensor factorization project. I learnt a lot from him. I am also grateful for the numerous technical and scientific discussions with Luis Tobalina and Bence Szalai. My interest for proteogenomics was initiated by Emanuel Gonçalves. I learnt about functional genomics with Francesco Iorio, Luz Garcia-Alonso and Aurelien Dugourd. The number and diversity of projects in this group were able to satisfy my ardent desire for self improvement for my whole PhD.

Finally, many thanks to Anila Liu and Christian Holland who helped me translate the abstract in german, and the 4 members of this thesis committee from Heidelberg University: Benedikt Brors, Julio Saez-Rodriguez, Ursula Klingmüller and Amir Abdollahi, for making this defense possible.

Table of content

General introduction

Precision medicine, also called "personalized medicine" is based on the customization of diagnosis and treatment according to molecular, genetic, transcriptomic and epigenetic information of the patient. This implies that patients are fundamentally different and therefore should be treated differently. Instead of the “one size for all” paradigm that ruled the pharmaceutical industry for decades, we are switching to the development of drugs which are efficient for only a subset of patients. With increasing spending in drug development industry due to adverse events of lack of efficacy, targeted therapies are evermore needed. Indeed, a drug that is toxic for one may be safe for another. Similarly, a drug that is ineffective for one may be effective for another. Different classes of targeted therapies are available in oncology (https://www.cancer.gov): hormone therapies, signal transduction inhibitors, gene expression modulators, apoptosis inducers, angiogenesis inhibitors, immunotherapies, and toxin delivery molecules.

In precision oncology, the therapeutic success for one patient can be used in a accurate way for many other patients whose tumors have a similar genetic profile. For instance, in breast cancer, a patient receives a standard chemotherapy, followed by remnography (MRI) to measure the efficacy. We now aim at including a genomic analysis of the breast tumor biopsy to choose a suitable combination therapy tailored to her genetic profile. This new approach adds several levels of information: a unique genetic profile, early MRI tracking, real-time assessment of the effectiveness of treatment, and the ability to respond quickly to new omics data.

The major technological advances of recent years have undeniably contributed in the rise of this new therapeutic approach. The decrease in the costs of sequencing of the human genome and custom analyzes such as DNA microarrays are examples of such advances. There are at least three essential implications in this metamorphosis of medical practices: (i) The first relates to the technological challenges necessary for this practice, such as bioinformatics infrastructures, in order to process genomic information in real time to guide the patient's treatment, as well as the power of the genome sequencing machines. (ii) The second is the training of the staff. Nowadays, it is essential to promote interdisciplinarity and a transversal approach to knowledge, *e.g.* to train highly qualified people in two or three fields of expertise. (iii) Finally, fundamental application of bioinformatics and machine learning in knowledge discovery. In this thesis, we will focus on the application of machine learning in deriving therapeutic insights for cancer.

We will present different projects to illustrate the role of machine learning and bioinformatics in precision oncology. The first project involves using matrix factorization to explore the underlying associations in cancer drug screenings (Yang et al. 2018a). We applied this concept in predicting drug combination synergy in breast and colorectal cancer. In this case, machine learning is used to generate hypothesis instead of purely prediction. In a second project, we

organized the NCI-CPTAC Proteogenomics challenge to understand the interplays between mRNA and protein level in breast and ovarian tumors.

Cancer cell lines have been the workhorse of preclinical study in oncology. Machine learning is widely used to predict drug response on the treated cell lines. However, insights derived from such studies usually refer to single drug-gene association. If the drug is a MEK inhibitor and the gene belongs to the MAPK pathway, researchers could report a relationship between MEK and MAPK. But it is not a direct and quantifiable association between protein MEK and MAPK pathway with respect to drug response. There is currently no such analysis regarding association between the features of the drug and the features of the cell lines. In order to directly capture the interaction between a protein target and a gene/pathway, we used Macau (Simm et al. 2017a), a matrix factorization type algorithm especially suited for cancer drug screening data. In a real life scenario, such interactions could answer the question: “which type of person likes which type of movie ?”, instead of a simpler high level association such as “which person likes which movies ?”. We analyzed the Genomics of Drug Sensitivity in Cancer (GDSC) (Iorio et al. 2016a) data, on 16 different cancer types and explored the interactions between drug targets and signaling pathways’ activations.

We applied this concept to drug synergy prediction and stratification. Cancer monotherapies are hampered by the ability of tumor cells to escape inhibition through rewiring or alternative pathways. Therefore, smart drug combination approaches are essential in controlling cancer proliferation and survival. We present two complementary workflows: One for prioritising drug synergy enrichment in high-throughput screens, and a consecutive workflow to predict hypothesis-driven patient stratification. Both workflows rely on bayesian matrix factorization to explore mechanistic relations between pathway activations derived from gene expression profiles and putative drug targets. We introduce the notion of Target functional similarity between 2 protein targets, which reflects how similarly effective drugs are as a function of targeted signaling pathway activities. Our synergy prediction workflow revealed that two drugs targeting the same or functionally opposite pathways are more likely to be synergistic, enabling experimental prioritisation in high-throughput screens and furthermore supporting the notion that synergy can be achieved by either redundant pathway inhibition or targeting independent compensatory mechanisms. We tested our synergy stratification workflow on a drug combination dataset for 7 pairs of protein targets (AKT/ALK, AKT/MTOR, AKT/EGFR, BCL2/MTOR, EGFR/MTOR and AKT/BCL2) applied to 33 breast cancer cell lines. For performance metric, we used the Pearson’s correlation of observed versus predicted synergy scores. We were able to reach an average drug-wise correlation of 0.27. We next experimentally validated our synergy stratification workflow with a BRAF/Insulin Receptor combination (Dabrafenib/BMS−754807) in 48 colorectal cancer cell lines. The performance is 0.31 for all 48 cell lines and 0.4 by taking into account KRAS status. The synergy prediction workflow can be a powerful framework for compound prioritization in large scale drug screenings. For instance, only testing drugs targeting two functionally very similar or very distinct proteins could significantly reduce the search space. The synergy stratification workflow could potentially maximize the drug efficacy of drugs already known for inducing synergy.

Signaling molecules, as well as most cancer drugs bind to protein receptors. If a protein is the most predictive of patient outcome, then making therapeutic decision based on the corresponding mRNA may be a mistake. Therefore, it is essential to characterize protein level to the best of our abilities. For this purpose, we launched a community-based collaborative competition: The NCI-CPTAC DREAM Proteogenomics Challenge. The challenge used public and novel proteogenomic data generated by the CPTAC to answer fundamental questions about how different levels of biological signal relate to one another. In particular, in Proteomics subchallenge we focused on the question: Can one predict abundance of any given protein from mRNA and genetic data? We predict the protein abundance based on mRNA and/or other molecular data. Proteins being the product of mRNA translation, there should be correlation between mRNA level and protein abundance. In cases where mRNA expression does not correlate with protein level, we explored through machine learning the potential post translational modifications and protein regulations *e.g.* the effect of other proteins.

This thesis is structured as follow: beside the abstract and the general introduction, two main chapters are presenting the main projects of the PhD, as described previously. Each research project is structured as follow: Introduction, Result, Methods, Discussion and Supplementary information in the annexe. Figures and tables are named separately for each chapter. Finally, a general conclusion and future perspectives.

Chapter 1: Target functional similarity based workflows for drug synergy prediction and stratification

1.1 Introduction

In the quest for clinical efficacy, drug combinations have been widely used in cancer therapies (Dry, Yang, and Saez-Rodriguez 2016a; Al-Lazikani, Banerji, and Workman 2012). Targeting a signaling pathway at one step may not be sufficient for reaching maximal effects on pathway inhibition. Resistance mechanisms to monotherapy can occur by activation of compensatory signaling, for example the activation of ERK signaling in melanoma when treated with BRAF inhibitors may lead to paradoxical activation of CRAF (Montagut et al. 2008). Targeting BRAF and downstream MEK at the same time proved to be beneficial for overall patient survival (Lopez and Banerji 2017), by inhibiting the initial BRAF driver mutation and paradox CRAF activation. Alternatively to inhibiting two key proteins within the same pathway, a common strategy is to parallel inhibit two separate cancer pathways to maximise drug efficacy. For example, parallel inhibition of ERK and AKT could be beneficial as those pathways may be connected through cross talks and feedback loops in breast cancer (Saini et al. 2013).

Many methods predict drug synergy using chemical structure and genomic information (Bulusu et al. 2016; Bansal et al. 2014; Preuer et al. 2017). Drug chemical structure does not reflect the drug's mode of actions as well as the putative drug targets (Yang et al. 2018b). Preuer *et al.* used deep learning to predict synergy within the space of explored drugs and cell lines, but still underperformed in predicting untested drugs on untested cell lines (Preuer et al. 2017), with mean square errors (MSEs) of 255 versus 414. One common bottleneck for the application of all those methods is the limited publicly available training set. Jaeger *et al* identified new drug combinations using network topology of pathway cross-talk (Jaeger et al. 2017), however, they did not consider gene mutation, which could be highly relevant in cancer treatment. Synergy is not a universal property of the drugs' chemical structures but also highly context dependent (Sun et al. 2015).

In the recent Dialogue on Reverse-Engineering Assessment and Methods (DREAM) drug combination challenge (Menden et al. 2017a), led by AstraZeneca, the best performing team used a mouse protein-protein interaction network to augment the genomic features based on their network distance from drug targets. Whilst the best performer achieved outstanding predictability on level of experimental replicates, synergy was predicted based on supervised machine learning algorithms. Sparsity in the DREAM training dataset was expert-knowledge driven, and therefore may bias towards biological known and ultimately bias the performance of supervised learning. In practice, the combinatorial explosion of drug pairs is the limiting factor to both the number of experimentally tested drugs, as for the number of tested cell lines. Finally, knowledge based cancer gene sets have been used to enhance predictive models. However such methods have been demonstrated to be less informative than data derived gene sets (Cantini et al. 2018; Schubert et al. 2018).

We here propose a methodology for identifying increased synergy likelihood based on the notion of target functional similarity and being independent from combinatorial experiments. This entity reflects how similarly effective drugs with common target are, as a function of signaling

pathways' activities. Two target proteins that are functionally very similar are likely to belong to the same signaling pathway. We argue that functional similarities between protein targets shed light on different synergy mechanisms. We then use the resulting mechanisms to build models to predict synergy. Thus, our prediction model is hypothesis derived and does not originate from a training set of drug synergy. We applied our methodology to the AstraZeneca breast dataset for drug synergy prediction (Menden et al. 2017a) and experimentally validated on predicted drug combination in colorectal cancer cell lines. In the process of synergy prediction, we revealed different synergy mechanisms depending on the cancer type.

1.2 Results

1.2.1 Synergy prediction workflow

We propose a workflow for highlighting synergy enrichment (**Figure 1**), based on multitask learning including the following steps: 1) We compute the pathway activity from gene expression using Pathway RespOnsive GENes (PROGENy) (Schubert et al. 2018). 2) Next, we apply the Macau algorithm (Simm et al. 2017b) to find interactions between the drugs' nominal targets and pathway activities i.e. how targeting a protein may affect different signaling pathways(Yang et al. 2018b) (**Methods**). 3) We then use the previously determined interactions to compute the functional similarity between two protein targets i.e how similar are the system's responses when targeting those two proteins. 4) Finally, functional similarity between protein targets pointing to different synergy mechanisms is estimated. This synergy prediction workflow allows for any given pair of protein targets, to estimate the likelihood of inducing synergy when targeting those 2 proteins. Our method returns a ranking of experimentally untested drug combinations from being likely to unlikely synergistic, which ultimately enables a prioritization for future experiments.

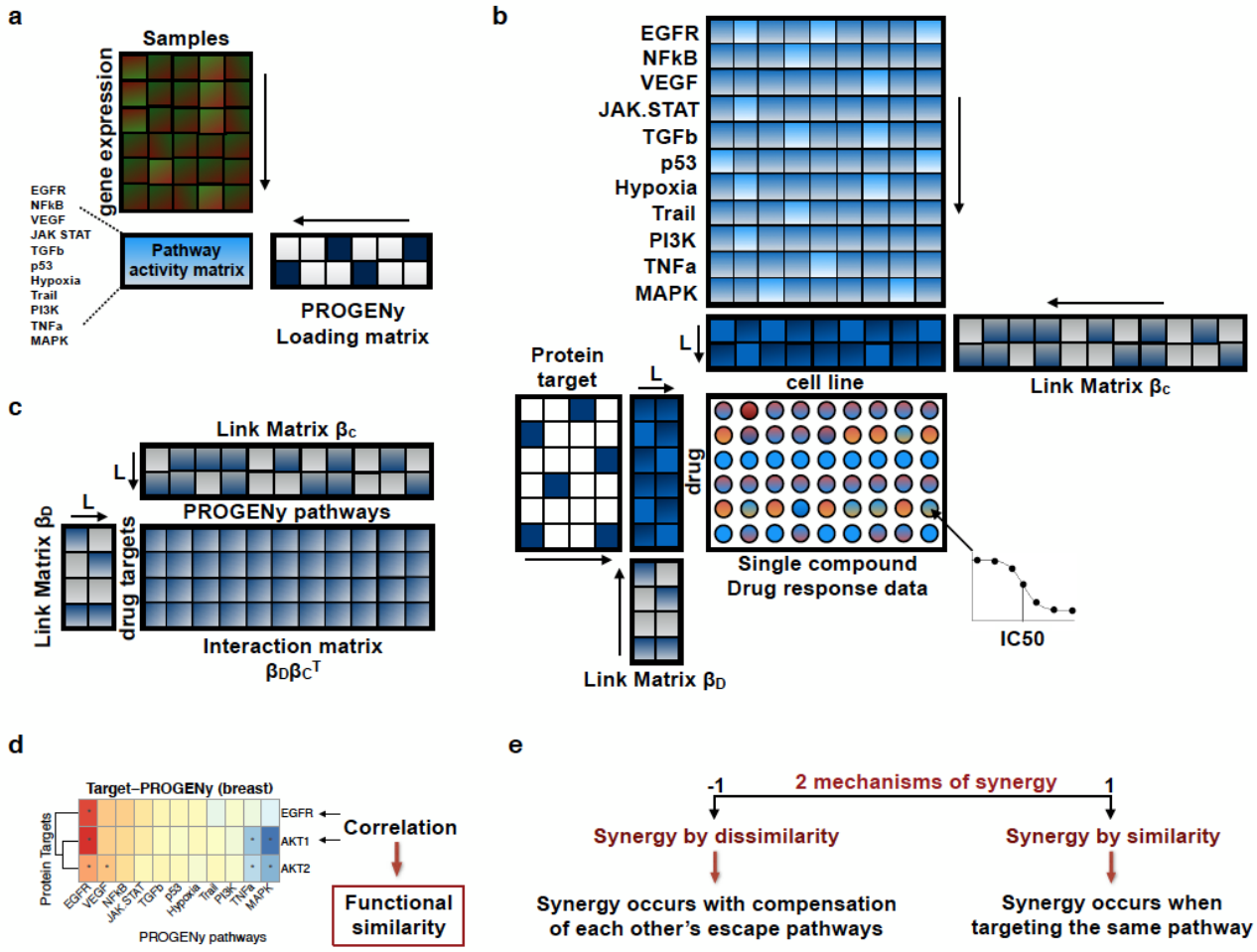


Figure 1: Methodology for drug synergy prediction and stratification.

(a) First, we reduce gene expression of cancer cell lines of single compound drug screening, into a small subset of pathway activities. It consists in multiplying the transcriptomics data by a loading matrix, as described in Schubert *et al.*

(b) We then use Macau algorithm(Simm et al. 2017c) to predict multiple drugs' responses simultaneously by uncovering the common (latent) features that can benefit each individual learning task. We use the previously derived pathway scores as input features (side information) for cell lines, and nominal target for drugs. Each side information matrix is transformed into a matrix of L latent dimensions by a link matrix. Drug response is then computed by a matrix multiplication of the 2 latent matrices.

(c) Concurrently to drug response prediction, we derive the interactions between drug features (targets) and cell line features (pathway activity), by multiplying the 2 link matrices. An association between protein X and pathway Y means that activation of pathway Y correlates with drug sensitivity when targeting protein X. In case of causality, we can say that activation of pathway Y confers sensitivity to any drug targeting protein X.

(d) These interactions allow us to define the functional similarity between two protein targets. In this example of breast tissue, The functional similarity between proteins EGFR and AKT1 is the correlation of their interaction values with the 11 PROGENy pathways. As final step of the

synergy prediction workflow, the derived target functional similarity inform us about the likelihood of synergy.
(e) For synergy stratification workflow, we start with target pairs already known to be synergistic. The value of the functional similarity between the protein targets reflects different synergy mechanisms. If the similarity is close to 1, synergy occurs by targeting the same signaling pathways. A similarity close to -1 suggests a synergy induced by compensation of escape mechanism. We build specific synergy model for each case to predict synergy scores of cancer cell lines.

1.2.2 Pathway activities

We transformed the transcriptomics data into pathway activity scores using PROGENy (Schubert et al. 2018) (**Figure 1a**), a more recent version of the Signaling Pathway Enrichment using Experimental Data sets (SPEED) signatures (Parikh et al. 2010). PROGENy is a data driven pathway method aiming at summarizing high dimensional transcriptomics data into a small set of pathway activities. PROGENy derives pathway signatures from the genes that are altered when perturbing a pathway instead of solely from the genes within the pathway as other methods do. This improves the estimation of pathway activities (Schubert et al. 2018). The 11 PROGENy pathways are EGFR, NFkB, TGFb, MAPK, p53, TNFa, PI3K, VEGF, Hypoxia, Trail and JAK-STAT.

1.2.3 Interactions between drug target and pathway scores

We then computed the interactions between protein targets and signaling pathway activation status with respect to drug response (IC50) using matrix factorization (**Figure 1b, 1c, Methods**). This interaction can be defined as the importance for those two entities to be simultaneously involved in order to have an impact on drug response (Yang et al. 2018b), e.g. how the simultaneous activation of a certain pathway and targeting a certain protein can be associated with drug response. For instance, a strong interaction between protein MEK1/MEK2 and pathway EGFR in pancreatic cancer is interpreted as follows: Activation of the EGFR pathway correlates with sensitivity when targeting MEK1/MEK2. If this were a causal relationship, it could mean that EGFR pathway activation confers sensitivity to any drug targeting protein MEK1/MEK2.

We used the Genomics of Drug Sensitivity in Cancer (Iorio et al. 2016b) (GDSC) cell line panel that contains drug response (IC50) data of 265 drugs on 990 cell lines. For each of the 16 tissues (with more than 20 cell lines), we computed the interaction matrix between drug targets and pathway activities using the multitask learning algorithm Macau (Simm et al. 2017b; Yang et al. 2018b). Our algorithm tries to learn multiple tasks (predicting multiple drugs) simultaneously and uncovers the common (latent) features that can benefit each individual learning task (Pan and Yang 2010). We used manually curated protein targets for the drug (**Supplementary**

Table 3), and gene expression derived pathway scores for the cell lines. The interaction matrix gives hints about the drug's mode of action, by uncovering in which condition (pathway status) targeting a certain protein correlates with higher drug sensitivity.

1.2.4 Target functional similarity

Next, we studied how similar two targets are in terms of pathway level impact on drug response. We reasoned that we can use the correlation as an estimate of the functional similarity between two protein targets: high correlation being the most similar pair and high anti-correlation being the most opposite pair. A pathway contains more information than a single gene's expression level. Therefore, functional similarity based on a small subset of essential pathways is likely to be more robust than using thousands of genes, of which the vast majority are not involved in drug response. We considered the target pathway interaction matrix for breast tissue and 102 protein targets which are targeted by at least two drugs in the GDSC dataset. We selected the 25 protein targets from GDSC that are also part of the AstraZeneca drug combination challenge data (**Figure 1d**, **Figure 2a**). There are $\frac{(n^2 - n)}{2}$= 300 pairwise combinations from the n=25 proteins. We then kept 99 target pairs where the two proteins are targeted by two different drugs in the GDSC panel. For each combination of targets, we computed the Pearson's correlation of the interaction score with the 11 PROGENy pathways. The target combinations were then ranked from the most correlated pair to the most anti-correlated pair. For instance, proteins BRAF and MEK have a functional similarity of 0.74 (p=0.0088) in skin cell lines, which illustrates the synergy mechanism of inhibiting two key proteins within the same pathway. We consider a similarity greater than 0.7 to reflect similar effects upon perturbation of those targets that are closely related in the signaling cascade (**Supplementary Figure 7** for distribution of similarity values).

Synergy scores in AstraZeneca breast dataset is derived from a dose-response surface of two drugs at different concentrations (**Methods**). A score of 50 is equivalent to an extra synergistic effect of 50% compared to the expected effect derived from the Loewe's additivity model (Fitzgerald et al. 2006). To ascertain if target functional similarity can influence drug synergy, for each target pair, we plotted the observed average synergy scores of the top three ranked synergistic drug1-drug2-cell triplets, against it's target functional similarity (**Figure 2b**). We observed that synergy arises in both highly correlated and highly anti-correlated target groups (**Figure 2b**). Very few synergistic target pairs were found with a functional similarity close to zero (lowly correlated target group).

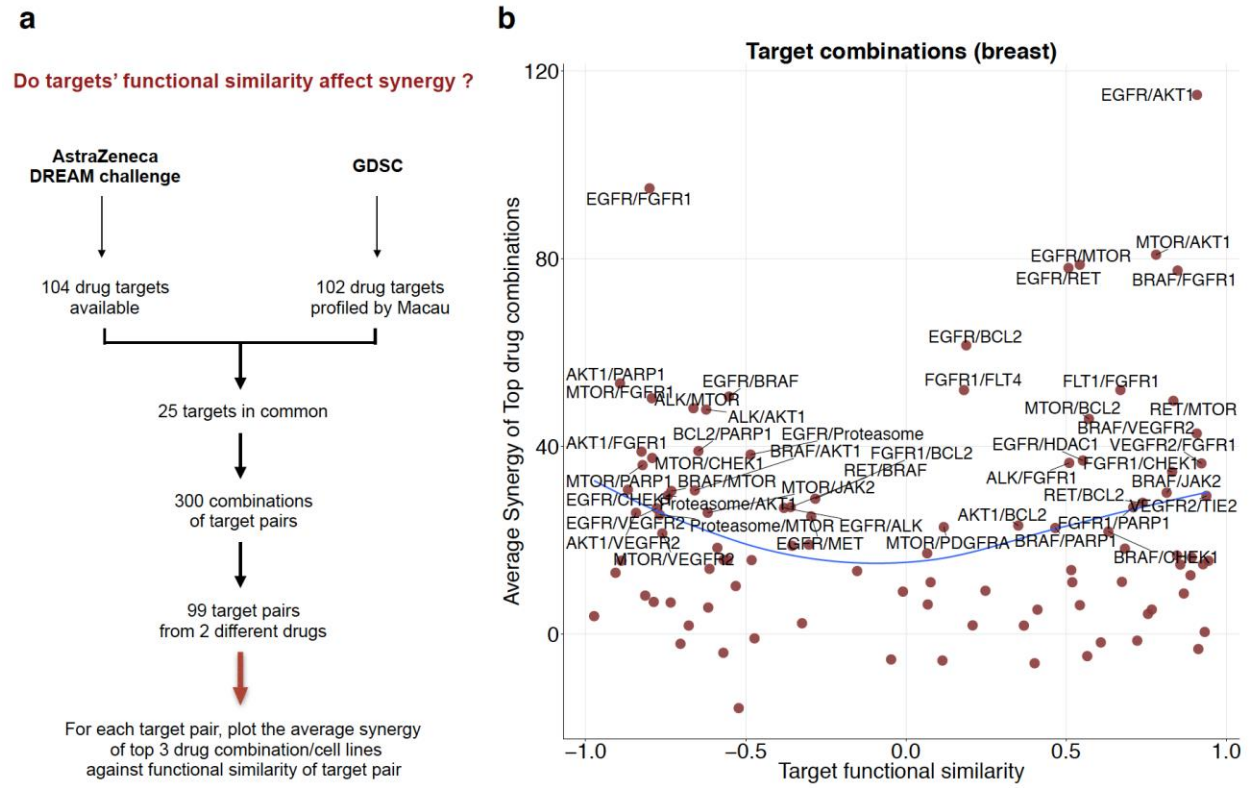


Figure 2: Influence of the similarity between protein targets on drug synergy. (a) We selected common targets from AstraZeneca and GDSC data sets. **(b)** The target functional similarity is the correlation between 2 targets by their interactions with the PROGENy pathways. A correlation of 1 implies that the activities of pathways correlates in the same way with drug efficacy on those proteins. A correlation of -1 implies opposite effects. The average synergy is computed for each target pair, as the mean of the top three synergistic drug-cell line pairs. We chose a threshold of 20 as synergistic effect, and a score lower than -20 as antagonistic effect, as in Menden *et al (Menden et al. 2017b)*.

We tested the significance of our observation on breast (33 cell lines), colon (12 cell lines) and NSCLC (22 cell lines), by computing the correlation between the top synergistic combinations and the absolute value of the target functional similarity (**Supplementary Figure 1**). For breast, colon and lung tissues, the Pearson's correlations are r=0.25 (p=0.014), r=0.45 (p=0.27) and r=0.14 (p=0.56), respectively. Although not optimal, the trend is stronger for colon than for lung, which was the reason guiding our choice for colorectal cancer cell lines in experimental validation.

Target functional similarity is therefore a metric that can be used for compound prioritization. For any given target pair, along with single drug response, we can increase the likelihood of synergy. This is, the more functional similar or opposite two proteins are, the mostly likely synergy will arise. We reason that this could be due to complementary mechanisms of synergy that take place:

Mechanism 1: When two drugs have similar interaction profiles, they are most likely targeting some common mechanism. In this case, synergy may be achieved by double hit of the same pathway. **Mechanism 2**: Conversely, targeting one protein may lead to resistance by an escape pathway or feedback loop. Targeting another protein which has opposite functional similarity may act on the escape pathway. Here, synergy reflects a compensation of escape mechanisms.

1.2.5 Synergy stratification workflow

As an addition to the synergy prediction workflow, we propose a consecutive step, which enables patient stratification. For this, we use the inferred synergy mechanism and pathway activities of new samples to build specific models to predict synergy for new drug combinations. The synergy stratification workflow predicts the actual synergy scores on samples for a given target pair for which synergy has been described (either through experiments or from literature).

Synergy stratification for each model

For each of the previously described synergy mechanisms (**Figure 1e**), we built models to predict synergy scores on cancer cell lines. We only consider drug combination known to induce synergy, for several reasons: (i) Our method relies heavily on literature/pathway knowledge, therefore difficult to test on all drug pairs. (ii) Synergy stratification when there is no synergy can be difficult to interpret. (iii) In practice, it is more likely to decide about stratification after knowledge of synergy potential.

Synergy Model 1: For functionally similar target pairs (Mechanism 1), we rank the pathways based on their sensitive or resistant interaction profile with respect to the drug targets (**Supplementary Figure 2**). We postulate that synergy is maximized under a pathway condition where both drugs' effects are maximized. The optimal condition for synergy is therefore when pathways associated with drug sensitivity are upregulated, and pathways associated with drug resistance are downregulated. As a consequence, if two protein targets have strong functional similarity e.g. high correlation between their interaction profile with pathway activities, synergy is maximized by maximizing the sensitizing pathways and minimizing the pathways conferring resistance. We predict synergy by taking the average of the N top sensitive pathway scores, subtracted by the average of the M top resistant pathway scores. Therefore, for each cell line, we introduce the concept of Delta Pathway Activity (Delta PA) to predict synergy:

$$Delta\ PA\ = \frac{\sum_1^N\ (Top\ sensitive\ pathways)}{N} - \frac{\sum_1^M\ (Top\ resistant\ pathways)}{M} \pm genomics$$

We compute the average pathway score for both sensitive and resistant groups. Each group should include a minimum of one to a maximum of three pathways. We select the top pathways with cross validation of group membership thresholds (**Methods**). If applicable, we include in the formula the genomic information which can be mutation (SNP) or copy number variation (CNV).

For instance if protein EGFR is targeted, we include CNV_{EGFR}. Group membership parameters are defined using cross validation (**Methods**).

Synergy Model 2: For functionally opposite target pairs (Mechanism 2), when a pathway's activation is associated with resistance for one protein target, it is also associated with sensitivity for the other protein target, to compensate. Two drugs can be individually ineffective, but more effective when combined. Therefore, synergy may arise in a situation of drug resistance. This could be explained by the fact that if a cell line is resistant for one (or both) of the drugs, there is "more opportunities" to be synergistic. When both drugs kill a given cell very efficiently, there is no synergy, as both drug A alone, drug B alone and combination A + B can kill all the cells. Unsurprisingly, resistance biomarkers were found to be predictive of synergy in the recent AstraZeneca DREAM challenge(Menden et al. 2017a). Therefore Delta PA should maximize the pathways conferring resistance and minimize the sensitizing pathways. The formula becomes:

$$Delta\ PA\ = \frac{\sum_1^M\ (Top\ resistant\ pathways)}{M} - \frac{\sum_1^N\ (Top\ sensitive\ pathways)}{N} \pm genomics$$

Model 2 is less likely to suit functionally similar pairs (Mechanism 1). If the two drugs have similar functional profile, maximizing the resistance scenario equals increasing the dose of the same inefficient drug, thus, unlikely to improve the outcome. Likewise, Model 1 is less suitable for Mechanism 2, as maximizing the sensitizing pathways is the same as prioritizing a situation where drug 1's sensitive effect outweighs drug 2's resistant effect. Thus, Mechanism 2's core idea would become obsolete, as by definition, the resistance scenario must prevail in case of escape mechanism. Of note, having an opposite functional profile does not imply Mechanism 2. An opposite pathway-response profile for 2 targets, offers the "functional scenario" for the cell to escape the damage induced by one drug. Yet, there could still be a scenario which maximizes the sensitizing pathways. This corresponds to 2 drugs targeting completely independent pathways, which is more due to independent actions rather than additivity or synergy(Palmer and Sorger 2017).

Our general framework to predict synergy scores follows several key steps and we emphasize on the notion of "target combination" which represents the dual inhibition of two protein targets, regardless of the drugs that are used (**Methods**).

Motivated by these general trends across all samples, we developed models to predict synergy in individual samples. We predicted synergy as a linear combination of pathway activation scores and built one model for functionally similar target pairs and one model for opposite pairs. We then applied our methodology on AstraZeneca drug combination data for breast tissue and experimentally validated a predicted synergistic drug combination for colorectal cancer cell lines.

Application to the AstraZeneca breast data set (**Supplementary text 1**)

We tested our synergy models on different target pairs by computing the Pearson's correlation of observed versus predicted average synergy on all available cell lines. The observed average synergy includes all drug combinations targeting the target pair of interest, computed as described in the Methods section. Therefore, for each cell line, the observed average synergy may be computed for different drug combinations since the matrix of drug-cell line synergy is very sparse.

We selected target pairs that fulfilled the following conditions: 1) Observed CombeneFit synergy score(Di Veroli et al. 2016) of top hits must be greater than 20 (**Figure 2b**), considered as a clear threshold for synergy(Menden et al. 2017a). 2) Drug combinations have had to be tested in at least 10 cell lines, owing to the limitations of measuring performance by Pearson's correlation. 3) At least two different drug combinations for the target pair were tested in each cell line, otherwise we excluded the cell line. We focused on the target pairs rather than specific drug pairs, in order to derive more robust insights.

This leaves us with the following 7 target pairs: AKT/ALK, AKT/MTOR, AKT/PARP1, AKT/EGFR, BCL2/MTOR, EGFR/MTOR, and AKT/BCL2, each representing several distinct drug combinations (3, 5, 3, 4, 4, 6 and 4, respectively). We applied our methodology on those target pairs (**Methods, Supplementary text 1**), and the prediction performances defined as correlations of observed versus predicted synergies are as follow: AKT/ALK (r=0.33), AKT/MTOR (r=0.086), AKT/PARP1 (r=0.50), AKT/EGFR (r=0.15), BCL2/MTOR (r=0.20), EGFR/MTOR (r=0.43) and AKT/BCL2 (r=0.19) (**Table 1, Figure 3**), with an average performance of 0.27 using Leave One Out Cross Validation (**Methods**, **Table 1**).

	Target pairs	Target functional similarity	Synergy model	Top sensitive pathways	Top resistant pathways	Delta Pathway Activity	Prediction performance from LOOCV
AstraZeneca breast data	AKT/ALK (3 combinations)	-0.4	Model 2	EGFR, VEGF PI3K	MAPK, TNFa	$\frac{MAPK+TNFa}{2}-\frac{EGFR+VEGF+PI3K}{3}$	0.33
	AKT/MTOR (5 combinations)	0.8	Model 1	EGFR, VEGF, PI3K	MAPK, TNFa	$\frac{EGFR+VEGF+PI3K}{3}-\frac{MAPK+TNFa}{2}$	0.086
	AKT/PARP1 (3 combinations)	-0.8	Model 2	EGFR, VEGF PI3K	MAPK, TNFa	$\frac{MAPK+TNFa}{2}-\frac{EGFR+VEGF+PI3K}{3}$	0,50
	AKT/EGFR (4 combinations)	0.9	Model 1	EGFR, NFkB, PI3K	MAPK	$\frac{EGFR+NFkB+PI3K}{3}-MAPK+CNV_{EGFR}$	0.15
	BCL2/MTOR (4 combinations)	0.7	Model 1	VEGF, NFkB, Trail	MAPK, TNFa	$\frac{VEGF+NFkB+Trail}{3}-\frac{MAPK+TNFa}{2}$	0.2
	EGFR/MTOR (6 combinations)	0.6	Model 1	EGFR, NFkB, VEGF	MAPK, TNFa	$\frac{EGFR+NFkB+VEGF}{3}-\frac{MAPK+TNFa}{2}+CNV_{EGFR}$	0.43
	AKT/BCL2 (4 combinations)	0.5	Model 1	EGFR, VEGF, PI3K	MAPK	$\frac{EGFR+VEGF+PI3K}{3}-MAPK$	0.19
Sanger colon data	BRAF/IR Dabrafenib/ BMS−754807	0.8	Model 1	Hypoxia, p53, MAPK	PI3K, VEGF, Trail	$\frac{Hypoxia+p53+MAPK}{3}-\frac{Trail+VEGF+PI3K}{3}$	All 48 cells: 0.31
						$\frac{Hypoxia+p53+MAPK}{3}-\frac{Trail+VEGF+PI3K}{3}+SNP_{BRAF}$	All 48 cells: 0.22 KRAS_mut: 0.5
						$\frac{Hypoxia+p53+MAPK}{3}-\frac{Trail+VEGF+PI3K}{3}+SNP_{KRAS}$	All 48 cells: 0.4

Table 1: Drug synergy prediction for breast and colorectal cancer cell lines.

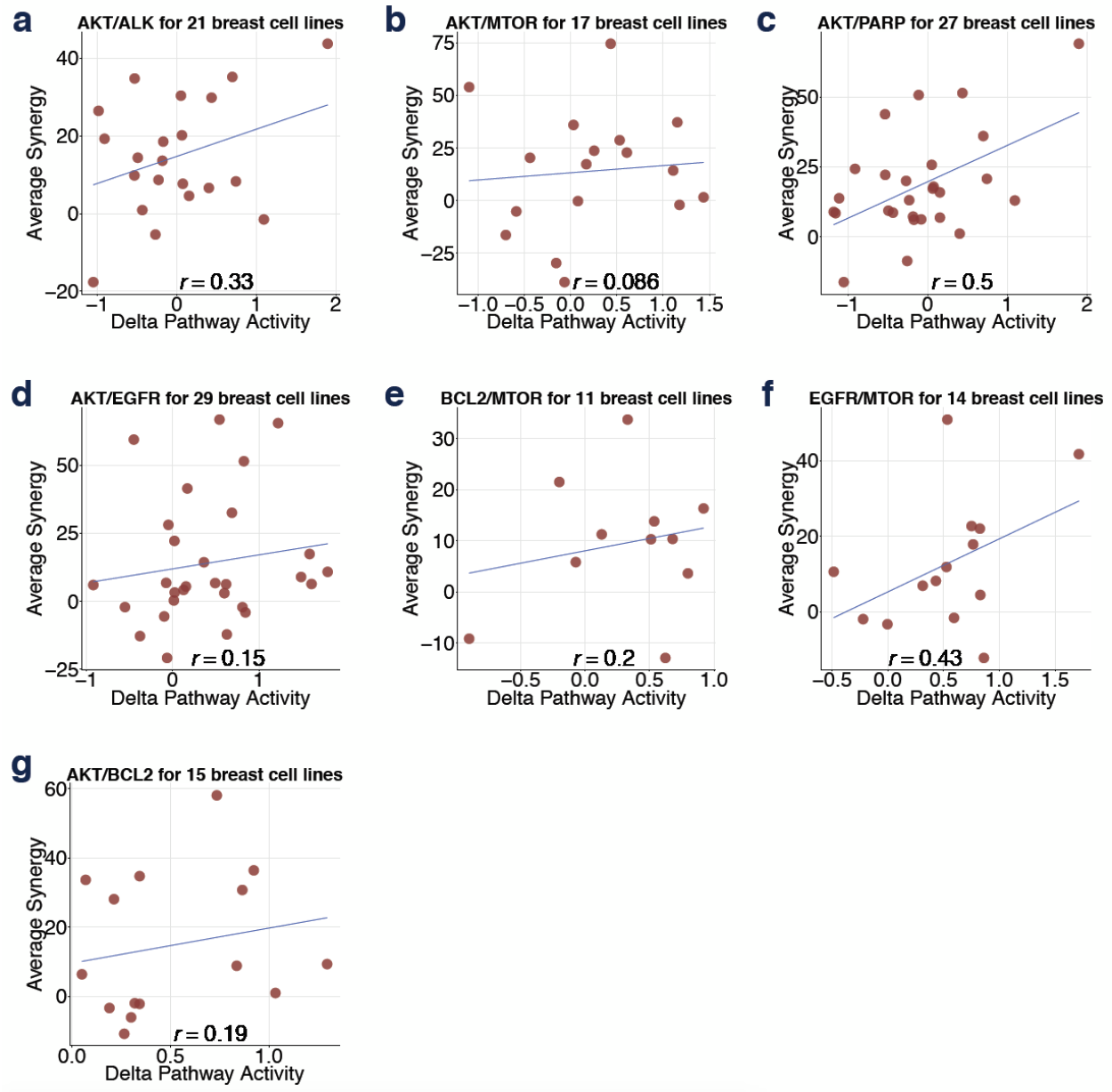


Figure 3: Prediction of drug synergy on breast tissue. (a), **(b)**, **(c)**, **(d)**, **(e), (f) and (g)** show the prediction result for AKT/ALK, AKT/MTOR, AKT/PARP1, AKT/EGFR, BCL2/MTOR, EGFR/MTOR and AKT/BCL2 targets pairs on breast tissue, respectively (from AstraZeneca DREAM challenge).

Literature supporting the drug combinations

AKT/ALK:
(i) It has been shown that synergy arises by targeting ALK and a downstream signaling pathway such as PI3K/AKT/MTOR in neuroblastoma (Moore et al. 2014), but not yet known in breast.
(ii) MAPK pathway is the top predictive feature of the Delta PA formula (**Table 1**), and it is known that MAPK is a critical downstream pathway necessary for ALK+ tumor cell survival (Hrustanovic and Bivona 2015), in agreement with the fact that Model 2 (synergy by maximizing resistance) was used as predictive model.

AKT/MTOR: Since AKT and MTOR are from the same pathway, dual targeting of those proteins is an obvious choice since breast cancer growth is often dependant on the PI3K/AKT/MTOR cascade (Cidado and Park 2012).

AKT/PARP1: PI3K pathway is among the top predictive features in the Delta PA formula (**Table 1**). Unsurprisingly, PARP inhibitor and PI3K inhibitor were described as an effective combination therapy for breast and ovarian cancer (Condorelli and André 2017; D. Wang et al. 2016; Rehman, Lord, and Ashworth 2012). We haven't found any literature evidence for dual targeting of AKT and PARP1 in breast cancer, but the efficacy would not be surprising since PI3K and AKT are closely related in the same pathway.

AKT/EGFR: Inhibition of the PI3K/AKT pathway potentiates cytotoxicity of EGFR inhibitors in triple-negative breast cancer cells (Yi et al. 2013).

BCL2/MTOR: There is a strong synergy between BCL2 and MTOR inhibitors(Hamunyela, Serafin, and Akudugu 2017). A cross talk between VEGF and BCL2(Bufalo et al. 2004) and the potential of targeting VEGF/MTOR(Chen et al. 2012) could explain that VEGF pathway score was the top predictive feature of synergy (**Table 1**).

EGFR/MTOR: Dual inhibition of EGFR and MTOR has been described for small cell lung cancer(Schmid et al. 2010) and for breast cancer (Glaysher et al. 2014).

AKT/BCL2: AKT regulates BCL2 expression in breast cancer(Bratton et al. 2010) but dual targeting of AKT and BCL2 has not been described.

Overall, among all target pairs, AKT/EGFR, AKT/MTOR, BCL2/MTOR, EGFR/MTOR and AKT/BCL2 were predicted with Model 1 (synergy by similarity). AKT/ALK and AKT/PARP1 were predicted using Model 2 (synergy by dissimilarity).

1.2.6 Validation on colorectal cancer cell lines

Within our presented study, we chose to validate our target-pathway interaction metric, synergy prediction and synergy stratification workflows in colon cancer. In order to ascertain our method's capability in detecting synergy, we chose a drug combination in the following way:

(i) We focused on drug combination involving protein BRAF which is an important protein in colorectal cancer as a mutation can result in uncontrolled, non–EGFR-dependent cellular proliferation (Nazemalhosseini Mojarad et al. 2013). About 10% of TCGA patients have this mutation for colorectal tissue.
(ii) From the 101 target combinations involving BRAF, we computed their functional similarities with BRAF. Insulin Receptor ranked first with a target functional similarity of 0.8 for BRAF/IR pair. We next computed the Delta PA formula and MAPK pathway ranked in the top three sensitizing pathways for BRAF/IR. Therefore, we chose BRAF/IR as a candidate for validation. We used Dabrafenib as a BRAF inhibitor and BMS−754807 as a selective inhibitor of IR/IGF1R (Carboni et al. 2009).
(iii) We chose the proteins BRAF, IR, and IGF1R as drug targets and used the target pathway interaction matrix to derive the Delta PA formula. Hypoxia, p53 and MAPK pathways belong to the sensitive group. The top resistant pathways are Trail, VEGF and PI3K. The synergy formula for Model 1 is therefore:

$$Delta\ PA\ (BRAF/IR)_{colon} = \frac{Hypoxia + p53 + MAPK}{3} - \frac{Trail + VEGF + PI3K}{3}$$

We validated our methodology on 48 colorectal cancer cell lines from the GDSC panel. Synergy score is computed with DeltaXMID (**Methods**). The Pearson's correlation of observed versus predicted synergy score is 0.31 for all 48 cell lines (**Methods**, **Table 1, Figure 4a**). We further reasoned that inclusion of additional information of top predictive pathway should increase the predictive power. In this case, the most predictive pathway is Hypoxia. KRAS mutation has been shown to differentially regulate the hypoxic induction of HIF-1α and HIF-2α in colon cancer (Kikuchi et al. 2009). Hence, we added kRAS status in the Delta PA formula and the prediction performance rose to 0.4 (**Figure 4c**). The performance rose to 0.5 by including BRAF status in the Delta PA formula and by only considering the subset of 26 KRAS mutant cell lines (**Figure 4d**).

We applied our synergy stratification workflow on breast and colon tissues. As real world use case, we envision that for any drug combination described as synergistic, this method could potentially inform about the subset of patients most likely to benefit, based on their transcriptomics profiles.

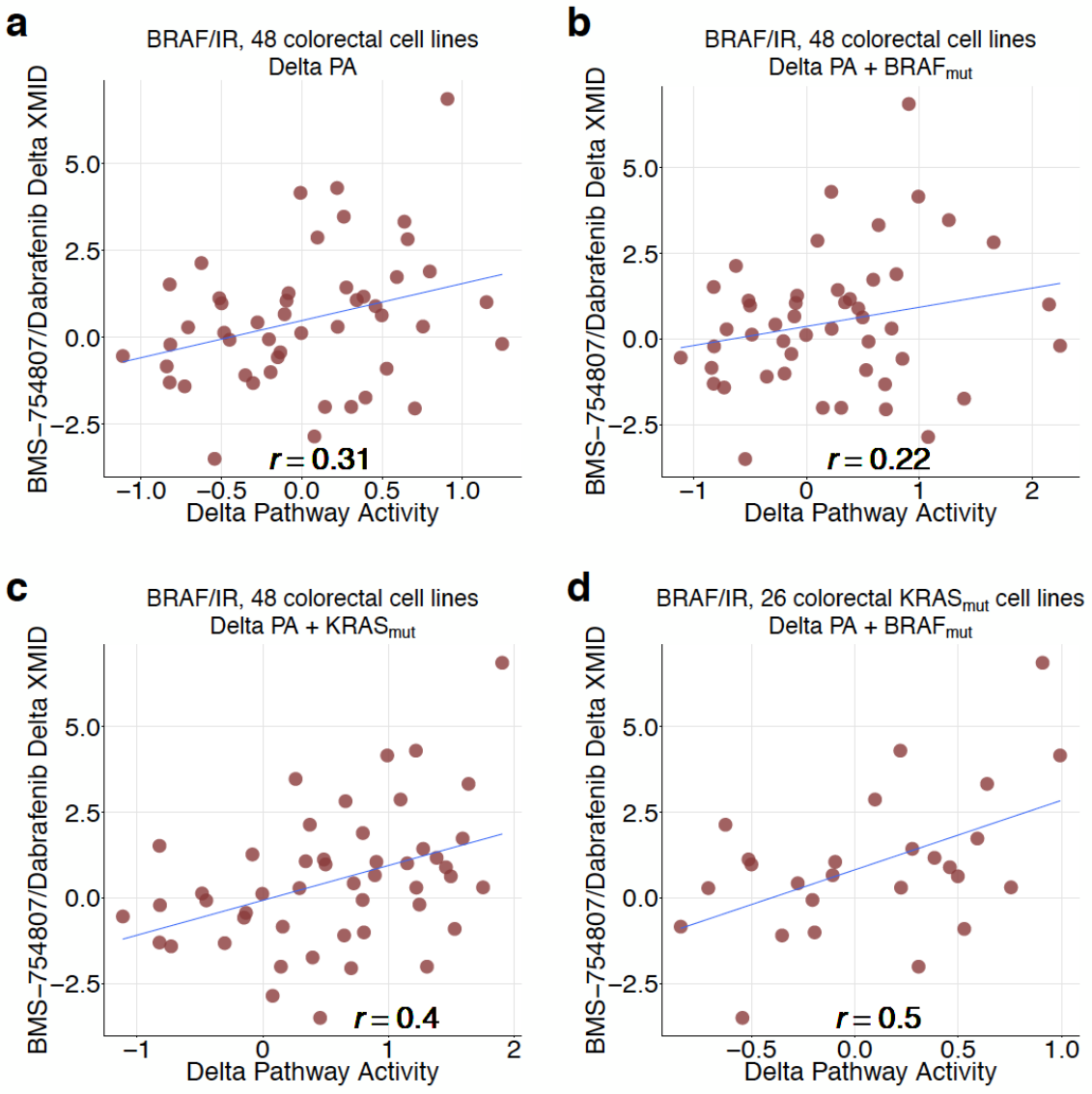


Figure 4: Prediction of BRAF/IR synergy on colorectal tissue. (a) shows the prediction result of BRAF/IR (BMS−754807/Dabrafenib) on all 48 colorectal cancer cell lines. **(b)** shows the result with BRAF status included in Delta PA formula. **(c)** and **(d)** show the result on $KRAS_{mut}$ colorectal cancer cell lines. **Drug screens were performed by the Translational Cancer Genomics drug screening team of the Wellcome Sanger Institute.**

1.2.7 Validation of synergy mechanism on external dataset

We explored the O'Neil et al Merck drug combination dataset(O'Neil et al. 2016) for colon, NSCLC lung and ovarian tissues, comprising data for 6, 6, and 5 cell lines, respectively. Given the number of samples per tissue, assessment of synergy is not possible on a drug wise setting. Nevertheless, we plotted the top synergistic pairs against the target functional similarity (**Supplementary Figure 4**). For colon tissue, the correlation between the average top synergistic pairs and the absolute value of target functional similarity is r=0.13 (p=0.11), versus r=0.45 (p=0.27) for AstraZeneca (**Supplementary Figure 1**). For NSCLC lung, there is no correlation with r=0.04 (p=0.62), versus r=0.14 (p=0.56) for AstraZeneca (**Supplementary Figure 1**). For ovarian tissue where the correlation is -0.16 (p=0.044), synergy seems to occur in case of low correlation between the drug targets. For colon and lung tissues, there is a slight agreement between the AstraZeneca and Merck datasets, supporting the methodology based on targeting functionally very similar or very opposite proteins.

1.2.8 Comparison with supervised learning approach

In the AstraZeneca DREAM challenge, an ensemble of best performing models was trained on the AstraZeneca DREAM combinatorial data, and consecutively tested on an independent combinatorial screen from Merck (O'Neil et al. 2016), which achieved a weighted mean correlation of 0.15-0.17. We considered this setting for predicting synergy of new drugs and new cell lines (**Supplementary Figure 5**). In comparison, our synergy stratification workflow uses the GDSC panel for hypothesis generation and the AstraZeneca dataset for testing. For the 7 target pairs (29 drug combinations) from breast tissue and one pair of drug combination validated on colorectal cancer cell lines, we were able to reach an average drug-wise correlation of 0.27. Of note, the two methods are of very different nature and have very different applications. Therefore, prediction performances should not be compared directly. In the DREAM challenge, synergy scores of drugs/samples are predicted without any prior knowledge of a drug combination leading to synergy or not. In contrast, in our synergy stratification workflow, we assume that at least one drugs/samples is synergistic and consecutively predict the stratification based on pathway activity and mutational profiles. We highlight the pros and cons for each methodology in **Supplementary Table 2**:

(i) Naive supervised learning approaches are easy to implement, do not require extensive domain expertise, and can be used for all possible prediction settings, drug wise and cell line wise (**Supplementary Figure 5**). On the other hand, it requires an extensive set of drug combination drug response data as training set.
(ii) For our synergy stratification methodology, linear combination of pathway activities is well suited for biological interpretation. However, it can only be used in drug wise setting and requires significant domain knowledge and literature evidence.

1.3 Methods

1.3.1 Matrix factorization with Macau

Macau trains a Bayesian model for collaborative filtering by also incorporating side information on rows and/or columns to improve the accuracy of the predictions (Simm et al. 2017b) (**Figure 1b**). Drug response matrix (IC50) can be predicted using side information from both drugs and cell lines. We use protein target as drug side information and transcriptomics/pathway as cell line side information. Each side information matrix is then transformed into a matrix of L latent dimension by a link matrix. Drug response is then computed by a matrix multiplication of the 2 latent matrices. Macau employs Gibbs sampling to sample both the latent vectors and the link matrix, which connects the side information to the latent vectors. It supports high dimensional side information (e.g. millions of features) by using conjugate gradient based noise injection sampler.

1.3.2 Drug synergy metrics

For AstraZeneca dataset, drug effects on cancer cell lines are measured at several concentrations for each drug. Therefore, the effect is described by a dose-response surface rather than a curve. The benefit of a drug combination can be partly assessed by the extra effect obtained when combining the drugs. Drug combinations are classified as synergistic, additive or antagonistic, based on the deviation of the observed drug combination response from the expected response. The expected response is quantified with the Loewe additivity model (Loewe 1953; Berenbaum 1989; Loewe 1928; Fitzgerald et al. 2006). Loewe additivity assumes the two drugs act on a protein through a similar mechanism. Synergy score is quantified with Combenefit (Di Veroli et al. 2016).

"In colorectal cancer we tested the drug combination of BMS-754807 and dabrafenib in 48 colorectal cancer cell lines. BMS-754807 (S2807, Selleckchem) was screened at 0.5 μM against a 7 point dose response of dabrafenib (S1124, Selleckchem), ranging from 10 nM- 10 μM. The XMID, which is akin to an IC50, of dabrafenib alone and dabrafenib in combination with BMS-754807 were calculated and the ΔXMID=XMID(dabrafenib)-XMID(dabrafenib+BMS-754807) calculated. The fold difference in XMID can be calculated by y-fold=2^ΔXMID, hence a ΔXMID of 3.32 corresponds to a 10-fold lower XMID for dabrafenib + BMS-754807 compared to dabrafenib alone." (**Text written by Patricia Jaaks, Wellcome Sanger institute**)

1.3.3 General framework for predicting synergy score

Step 1: For two given protein targets T1 and T2, find their interactions with the PROGENy pathways using Macau (**Supplementary Figure 3**).

Step 2A: If available, use literature to guide the choice of Model e.g if we know that a drug combination is synergistic when a pathway X is activated, the model would be the one which gives a positive sign for pathway X. Otherwise go to **Step 2B**.

Step 2B: Compute the functional similarity between T1 and T2 (pearson correlation between T1 and T2's interactions with the pathways).

- If the correlation is close to 1, use **Model 1** to define the Delta PA formula.
- If the correlation is close to -1, use **Model 2** to define the Delta PA formula.
- If the correlation is between -0.3 and 0.3, it is an undetermined case.

Step 3: Find top sensitive and top resistant pathways (as previously described in synergy models). Take into account literature evidence in choice of pathways (for known drugs or targets). If a pathway is described as important in literature but does not appear in top 3 of a group, we include it, as well as any pathway separating the first from the one of interest, while respecting the limit of three pathways per group.

Step 4: In case of multiple drugs representing the same target pair, as in the AstraZeneca data set, remove as many drugs as possible to reduce off target effects (at least 3 drug pairs left). An alternative is to take the off target into account if some drugs are targeting one of PROGENy pathways.

Step 5: Use the Delta PA formula to predict synergy of a drug combination targeting T1 and T2. The pathway activities of the formula are computed by PROGENy on the cell lines of interest.

1.3.4 Cross validation of group membership thresholds

Prediction using a fixed threshold for both groups

For computing the Delta Pathway Activity formula, we chose the group membership to be the same in the top sensitive and top resistant groups. We only included pathways producing more than 70% of the effect of the first selected pathway. This choice of parameter seems to be a reasonable choice between including no additional pathway (threshold close to 100%) and including too many potentially irrelevant pathways (50%). We did a sensitivity analysis for this parameter and used different thresholds for the 2 different groups (sensitive and resistant). We fixed one group's threshold while varying the other group's threshold. We observe that for many target pairs, the sensitive group's threshold is rather stable (**Supplementary Figure 6**). For the sensitive group, the model is quite robust to variation of this parameter, whereas for the

resistant group, a lower threshold (including more pathways) seems to yield better result. This could be explained by the fact that GDSC drug screening adjusts the drug concentration so that only a few cell lines respond while most are resistants. Therefore, more information are reflected from the resistant side than from the sensitive side.

Prediction using cross validated and different thresholds for each group

We next predicted each target pairs of AstraZeneca breast data with a Leave One Out Cross Validation (LOOCV) to optimize the group membership thresholds. For each target pair, we used as thresholds the best values all target pair of the training set. The average prediction of the 7 target pairs is 0.27. Finally, we used the average parameters for the breast data to predict the BRAF/IR in colon data (**Table 1**).

1.3.5 Data

GDSC data were downloaded from: http://www.cancerrxgene.org/

Drug IC50 version 17a

Basal gene expression 12/06/2013 version 2

Drug target version March 2017

DREAM drug combination challenge data were acquired through an AstraZeneca Open Innovation Proposal.

Merck drug combination data is downloaded from the publication O'Neil et al(O'Neil et al. 2016)

1.4 Discussion

In this project, we presented two workflows for drug synergy prediction and patient stratification. The synergy prediction workflow can be a powerful framework for compound prioritization in large scale drug screenings. For instance, only testing drugs targeting two functional very similar or very opposite proteins (|correlation| > 0.7) could significantly reduce the search space, therefore decreasing the cost of drug combination research. The synergy stratification workflow could potentially be used to maximize the drug efficacy of drugs already known for inducing synergy. Indeed, knowing that a pair of compounds is synergistic does not tell us on whom to use it. We modeled the genomic context with linear combination of pathway activities.

We introduced the notion of functional similarity between two protein targets. This metric shed lights on two scenarios where drug synergy occurs: when drugs are targeting functionally similar proteins (AKT/EGFR, AKT/MTOR, BCL2/MTOR, EGFR/MTOR and AKT/BCL2) and when they are targeting functionally opposite proteins (AKT/ALK and AKT/PARP1). Our results support that synergy occurs and is much easier predicted when the targets are functionally very similar or very anti similar. Portraying the interaction between protein targets and pathway activities allowed us to recognize the different synergy cases. Based on that, we predicted synergies of 7 target pairs (AKT/ALK, AKT/MTOR, AKT/EGFR, BCL2/MTOR, EGFR/MTOR and AKT/BCL2, for 29 drug combinations) in breast cancer cell lines. We validated the synergy hypothesis for colon and lung tissues in an independent dataset. Finally, we predicted and validated a drug combination synergy (Dabrafenib/BMS−754807) on 48 colorectal cancer cell lines.

There are several limitations to this study that can be the focus of future work: (i) Better synergy models are needed, such as one which takes into account non-linear effects of pathways; we could envision adding coefficients to each pathway and including AND/OR gates. But this would require an extensive training set. (ii) In this present work, target functional similarity is defined with respect to 11 PROGENy pathways, which do not necessarily capture all cancer mechanisms. Therefore, the need to expand this geneset to include more cancer relevant pathways. (iii) In order to predict synergy of new compounds, drug targets have to be profiled by large scale monotherapy drug screening experiments across hundreds of cell lines. Thus, the need to expand single drug response data, while this is at a complexity cost of O(n), running drug combinations are $O(n^2)$.

Our study findings are aligned to those of the DREAM drug combination challenge(Bansal et al. 2014), where synergy was found to be highly context dependent. In our case, we predicted synergy with a linear combination of pathway activities. Bansal et al. predicted synergy from single-compound perturbation data e.g. synergy occurs for drug pairs which induce very similar or very opposite gene perturbation statuses. We used single-compound drug response data and the Macau algorithm to compute the target functional similarity, which reflects the similarity of drug response changes for different pathways after targeting a specific protein. We found that in breast and colorectal cancer, compounds which have very similar or very opposite functional profiles tend to be more synergistic. We used the inferred synergy mechanism and pathway

activities to predict synergy of new compounds. We found that whether synergy arises in case of similarity or dissimilarity was also tissue specific. Hence adding more complexity in drug synergy predictions.

Palmer et *al* stated that successful drug combination in tumour shrinkage are mostly due to targeting unrelated pathways, without any real synergy(Palmer and Sorger 2017). Drug action similarity is defined by the correlation of single drug response data, which resembles our use of target - pathway based similarity score. We used synergy and not additivity as response variable, thus in line with the lack of real synergy in lowly correlated group both in our analysis and as described by Palmer et *al*. They also concluded that drug interaction (synergy and not additivity) can explain the majority of combination clinical trial only if the drugs have strong cross-resistance (i.e. highly correlated independent drug responses), whereas low cross-resistance (i.e. lowly correlated independent drug responses) makes independent action of drugs the dominant mechanism in clinical populations. While the assessment of synergy is different in our case, as we used cell line data, the overall conclusions are in agreement.

In summary, exploring the interactions between drug targets and signaling pathways in a tissue specific manner can provide a novel in-depth view of cellular mechanisms and drug modes of action, which can ultimately rationalize drug combination strategies in cancer. Target functional similarity could be used as a metric for compound prioritization. Synergy by similarity hypothesis could be a rational for first line treatment, while synergy by opposite effect could potentially fit patients having acquired resistance.

Chapter 2: Quantitative prediction of proteome for large scale proteogenomics characterization of tumor samples

2.1 Introduction

DNA sequence information is transcribed into mRNA, which is then translated into protein. Such process is known as the central dogma of molecular biology, and characterizes a complex series of events allowing the flow of genetic information to phenotype. Since proteins are the product of mRNA translation, their abundance is likely to correlate with mRNA level. In steady-state conditions, protein abundances are largely determined by mRNA levels (Liu, Beyer, and Aebersold 2016). But during highly dynamic phases, such as cellular differentiation or stress response, post-transcriptional modifications (PTM) may weaken the protein/mRNA correlation (Liu, Beyer, and Aebersold 2016). This could explain the correlation between protein/mRNA of 0.36-0.5 for the majority of human tissues (Kosti et al. 2016).

Characterization and analyses of alterations in the proteome hold the promise to revolutionize cancer research, through understanding the association between genome, transcriptome and proteome in tumors. Signaling molecules bind to protein receptors, which are the targets of many cancer therapies. Therefore, it is essential to characterize them to the best of our abilities. For this purpose, we launched a community-based collaborative competition: The NCI-CPTAC DREAM Proteogenomics Challenge. The challenge used public and novel proteogenomic data generated by the CPTAC to answer fundamental questions about how different levels of biological signal relate to one another. In particular, we focused on the following questions: (i) Can one predict abundance of any given protein from mRNA and genetic data ? (ii) Can one predict phosphoprotein abundances from protein abundance ?

We explore through machine learning the role of mRNA, potential PTMs, protein regulations and degradation (*e.g.* the effect of other proteins), in predicting protein abundance. PTMs of proteins play essential roles in a large number of biological processes. However, technical difficulties and high costs of PTMs profiling greatly hamper the abilities of scientists to study and understand these important molecules in biological systems. Thus, it is of great interest to assess the ability of predicting PTMs activities based on more easily accessible molecular data (*e.g.* mRNA).

We also construct prediction models for phospho-protein abundances based on global protein abundances, RNAseq data and copy number variation (CNV) data from CPTAC breast and ovarian studies. Moreover, it is of interest to assess which data type can better predict phosphorylation activities and whether an integrative framework could outperform analysis based on single-data type.

As a result of this competition, we used the winning method for downstream analysis of protein translation and regulation. We then assessed the utility of of predicted protein abundance using: (i) TCGA samples of breast tissue which were not used in the challenge, for survival analysis. (ii) Cancer cell line for drug response prediction.

2.2 Results

2.2.1 Challenge design

The NCI-CPTAC Proteogenomics DREAM challenge is divided into different subchallenges: (1) imputing missing proteomics data (not described in this thesis). (2) The Proteomics subchallenge (sc2, led by myself) consists in predicting protein abundance based on mRNA and CNA. (3) The Phosphoproteomics subchallenge (sc3, led by Francesca Petralia, Icahn Institute) consists in predicting phosphosite abundance based on protein, mRNA and CNA (**Figure 1**). Each subchallenge is composed of multiple rounds: 2 leaderboard rounds, a validation round and a collaborative round. To cope with data confidentiality, participants used a docker container to store their pre trained models (infrastructure implemented by Thomas Yu, Sage Bionetworks). Submission consists of applying the pre trained model to the test data. Participants were given a leaderboard dataset to test their model and generate one prediction file and one confidence file per leaderboard round. Scores were returned to participants so that they can improve their model throughout these rounds for their final round submission which was scored against a held-out dataset. In this thesis, we will mainly focus on the Proteomics subchallenge.

2.2.2 Challenge data

As training data, we used TCGA retrospective collection of 77 breast and 174 ovarian tumor samples measured at four biological level along with their measured numbers in breast and ovarian tissue: proteomics (10005, 7061), phosphoproteomics (31981, 10057), transcriptomics (15107, 15121) and copy number alterations (16884, 11859) (**Methods, Figure 1**). Retrospective proteomics and phosphoproteomics data were downloaded from CPTAC data portal and processed by the common data analysis pipeline from CPTAC. For both tissues, proteome and phosphoproteome data were acquired using iTRAQ (isobaric Tags for Relative and Absolute Quantification) protein quantification methods. As testing data, we used 108 prospective samples of breast tissue and 82 samples of ovarian tissue, for all four level of measurements (**Methods, Figure 1**).

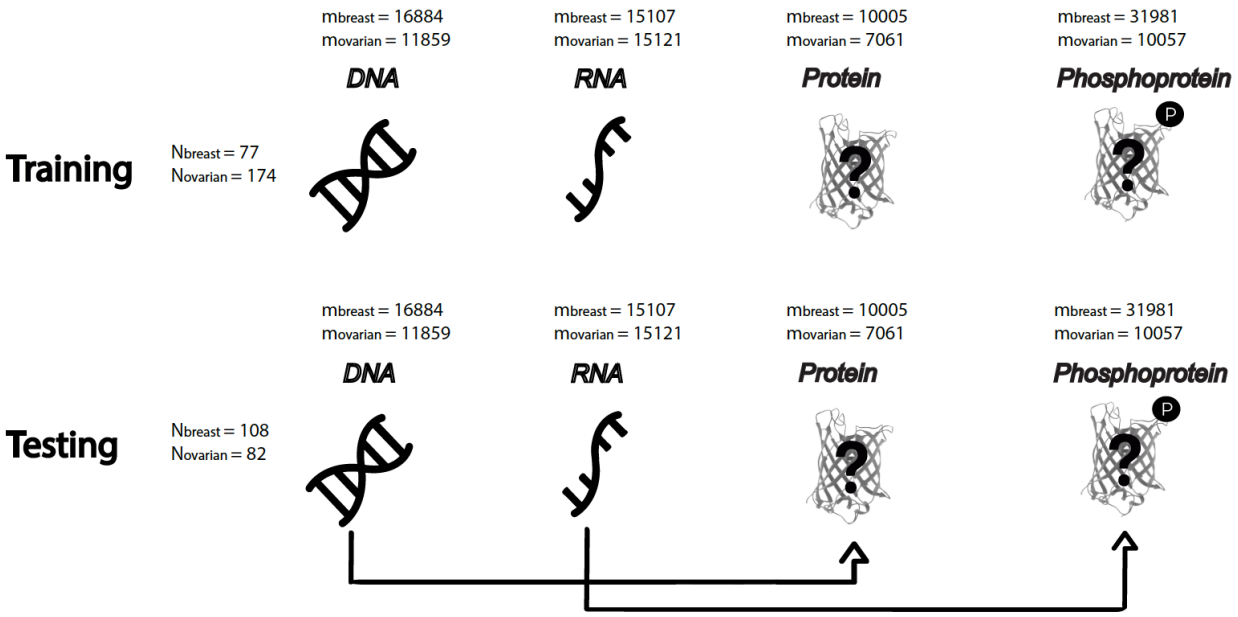


Figure 1: Challenge design and data. For both training and test sets, we have four layers of omics data (Copy Number Alteration, mRNA, Proteomics and Phosphoproteomics). CNA and mRNA can be used to predict Protein level (Proteomics subchallenge). CNA, mRNA and Protein level can be used to predict Phosphosites abundances (Phosphoproteomics subchallenge). **Data prepared by Zhi Li, New York University**.

2.2.3 General outcome of the challenge

A total of 29 teams submitted for Proteomics subchallenge breast. Prediction were evaluated based on Pearson's correlation of observed versus predicted protein abundance across new patient samples (**Methods, Figure 2**). The average prediction performance is r=0.26 (sd=0.17), and the best performance r=0.51. For Proteomics subchallenge ovarian, 32 teams submitted, with average performance of 0.29 (sd=0.18) and best performance r=0.53. If we consider the subset of proteins for which the corresponding mRNA is available, the winning team reached an average correlation of 0.55 and 0.53 for ovarian and breast tissues, respectively. The improvement are 17% (ovarian) and 15% (breast) compared to the naive correlation between mRNA and protein level.

For Phosphoproteomics subchallenge breast, 16 teams submitted, with average performance of r=0.17 (sd=0.16) and best performance r=0.42. For Phosphoproteomics subchallenge ovarian, 22 teams submitted, with average performance of r=0.11 (sd=0.11) and best performance r=0.33.

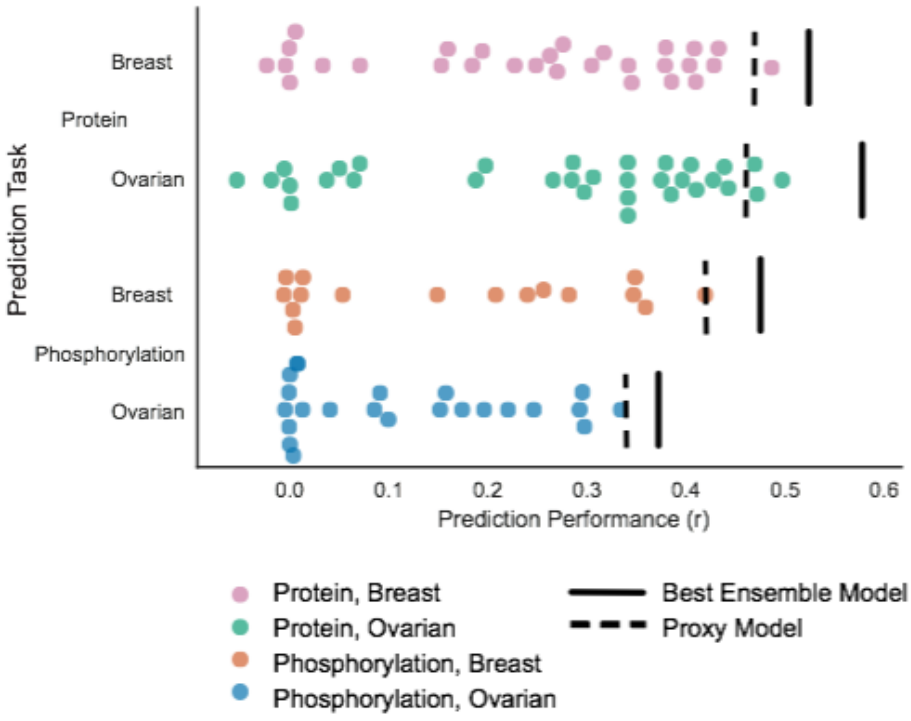


Figure 2: Overall performances in the challenges. For each subchallenge, we plot the performances of all participating teams in breast and ovarian tissues. The random distribution is generated by permutation of protein/phosphoprotein abundances across each patient. **Figure made by Zhi Li, New York University**.

2.2.4 Global insights

Data preprocessing and algorithms

The best performing methods in this challenge were tree and ensemble based. Missing value were excluded from predictor variable in training phase. External information such as KEGG pathways, Interaction networks, CODON count, GC percentage, protein folding energy and transcription factors were also used. Training data were scaled or feature wise standardized. The best performing team performed quantile normalization on the training and testing data altogether.

The best performing teams only used mRNA to predict protein abundance for Proteomics subchallenge, and found no benefit in including CNA data. For Phosphoproteomics subchallenge, protein abundance was the chosen omics layer to predict phosphosite abundance. mRNA and CNV were not used as no additional improvement compared to proteomics only.

Factors influencing protein prediction performance

We observed that proteins for which the corresponding mRNA is present were better predicted compared to those where it is not measured (**Figure 3**, $p<2.2e-16$). Similarly, proteins that are “free” were better predicted than those belonging to a protein complex from CORUM database (**Figure 3**, $p<2.2e-16$). This could be explained by the fact that “free ”proteins are more likely to follow the corresponding mRNA level. On the other hand, proteins inside a complex are co-regulated by other proteins and the structure is more robust to transient variation of mRNA level. The best predicted proteins are those which are directly influenced and accessible by mRNA. Another factor could be protein size, as complex proteins are bigger, thus mass spectrometry is less able to fractionalize the peptides.

One could hypothesize that more abundant proteins are better predicted, as easily measured by mass spectrometry. But we found no evidence of positive correlation between protein abundance and predictability. For breast and ovarian tissues, the correlation between prediction performance and average abundance across samples are respectively -0.077 ($p=7.6e-13$) and -0.25 ($p=7.2e-75$). The slightly negative correlation could be explained by the fact that proteins in complex are generally more abundant than protein out of complex ($p<2.2e-16$ and $p<4.7e-07$, for ovarian and breast, respectively).

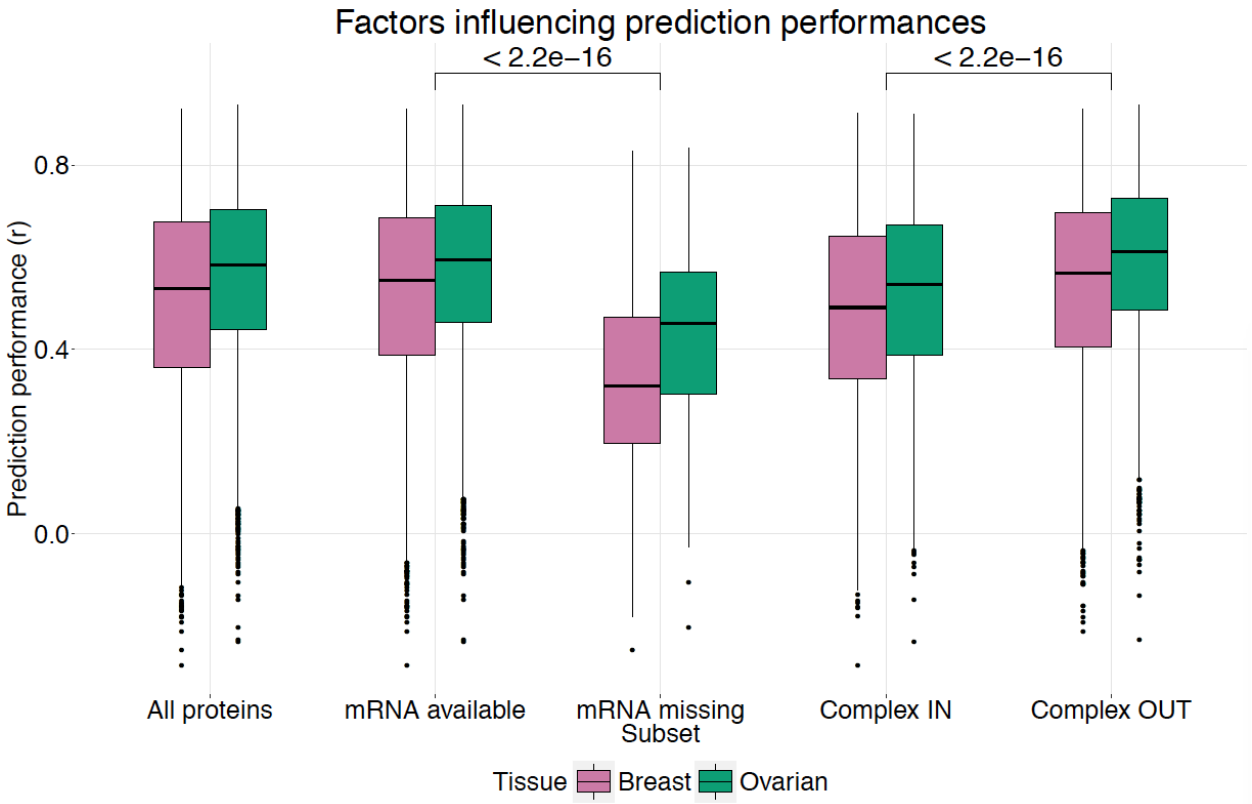


Figure 3: Factors influencing predictability. We present for each tissue, the winning team's prediction performance for: (i) all proteins; (ii) subset of proteins for which the corresponding mRNA is measured; (iii) subset of proteins for which the corresponding mRNA is missing; (iv) subset of proteins belonging to a protein complexe (CORUM); (v) subset of proteins not belonging to a protein complexe.

We further grouped the genes into HGNC families, then plotted the average prediction performance of each family for breast and ovarian (**Figure 4**). The prediction performance do not vary a lot between breast and ovarian tissues (r=0.88, p=6.4e-33). The best predicted families were Aldehyde dehydrogenase, Acyl-CoA synthetase, Integrin alpha subunits and Glutathione S-transferase. The least predicted families were histones and ribosomal proteins, in both cases functioning as complexes.

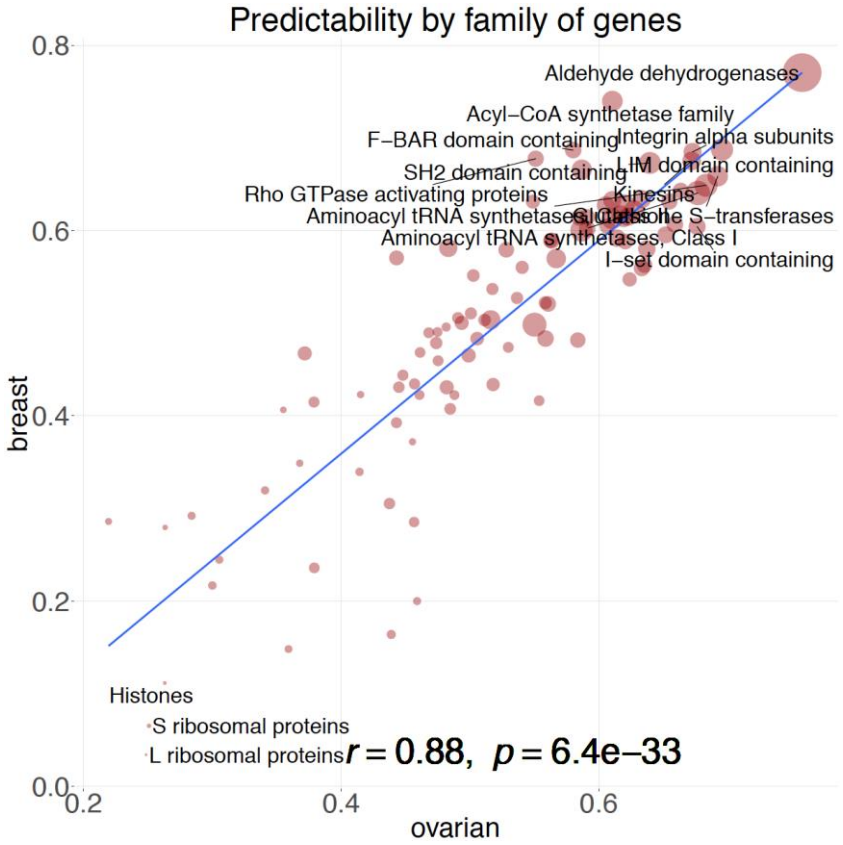


Figure 4: predictability by family of genes. We define the stability of each family as the inverse of the coefficient of variation of the predictability across the family. The more stable a family is, the less variation in predictive performances across all proteins of the family.

2.2.5 Common protein regulators predict survival

We used the best performing method to explore the feature importance and considered only the predictor genes that are top ranked across all proteins. Those genes are likely to be the common protein regulators. We took the top 10 regulators with corresponding protein prediction performance greater than 0.7, for ovarian: NNMT, SFRP4, FAP, SULF1, MYLK, MCM7, PLS3, KRT17, HBB and CILP. And for breast: GBP5, MMP11, DOPEY2, LASP1, CXCL9, ICAM1, DENND2D, WDR12, PLCD1 and HOPX.

We want to assess if the best predicted common regulators can also be predictive of patient survival. We chose breast cancer as it has the biggest sample size of patients and used the predicted proteomics (901 samples) based on the winning model. After log rank tests of the 10 predictors after correcting for age, gender, and false discovery rate, proteins PLCD1 and WDR12 are the top predictive biomarker of survival with q-values of 0.039 and 0.013, respectively. Similarly, we used mRNA of those two proteins for the same survival prediction. mRNA PLCD1 and WDR12 have an q-value of 0.079. Protein Phospholipase C Delta 1 (PLCD1) functions as a tumor suppressor in several types of cancer(Xiang et al. 2010), and it comes with no surprise that a higher level of PLCD1 is associated with a better outcome (**Figure 5**). Protein WDR12 is required for maturation of ribosomal protein (Lewinska et al. 2017), therefore an increase of this protein may result in cell proliferation, thus associated to poor clinical outcome.

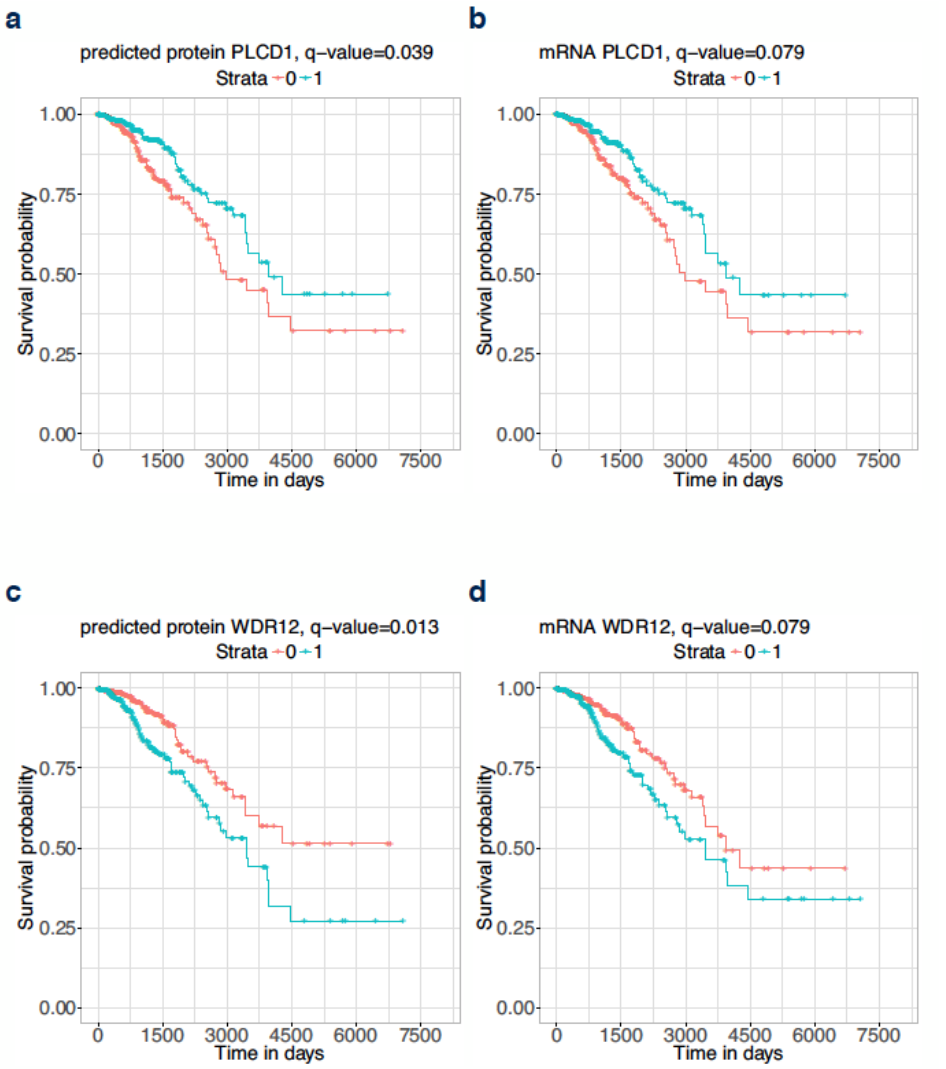


Figure 5: Patient stratification from common regulators. Kaplan Meier plot with proteins PLCD1 and WDR12 as the top predictive protein regulators, using predicted proteomics data of 901 breast cancer samples.

2.2.6 Proteomics insights from patient stratification

Patient stratification is an important goal in oncology research. However, it is difficult to do any survival analysis using real proteomics data due to the small sample size (77 samples for breast tissue). Therefore, we applied the winning method to 901 new breast samples from TCGA to generate predicted protein abundances from mRNA. Another issue in patient stratification is that testing all predictors for survival (log rank test) is highly inefficient, as a simple multiple hypothesis correction may invalidate all selected biomarkers. Therefore, we used the state of the art Multi Omics Factor Analysis (Argelaguet et al. 2017) to reduce an omics dataset into a subset of hidden variables (Factors) that capture the biological/technical variability of the dataset (**Methods**). We applied MOFA on proteomics, mRNA and a

combination of proteomics and mRNA. The derived factors were then tested as prognostic biomarker of patient survival (**Supplementary Table 1**).

We used the top predictive factors to stratify patients. Predicted protein alone identified 2 predictive factors with respective q values of 0.0014 and 0.00071 of log rank test (**Supplementary Table 1**). One factor was identified from mRNA with comparable q value of 0.0025 (**Supplementary Table 1**). Using the combined proteomics and mRNA, we identified 3 predictive factors with q values of 0.0050, 0.044, 0.0071.

The top predictive factor of predicted protein is enriched in Peptide chain elongation pathways (**Supplementary Table 1**). Top contributing genes of those factors includes RPL26, RPL29, RPL34 and RPL37. A factor analysis on predicted protein abundance was able to identify this class of proteins which play an important role in breast cancer (Goudarzi and Lindström 2016; Belin et al. 2009; Van Long et al. 2016). Another predictive factor includes TMEM26 and VGLL1 as to predictive genes. TMEM26 is highly expressed in triple negative breast cancer, and is associated with higher risk of recurrence particularly in ERα-negative cases (Mitra 2017; Nass et al. 2016). VGLL1 expression is associated with a triple-negative basal-like phenotype in breast cancer and correlates with poor survival (Castilla et al. 2014).

By combining predicted protein with mRNA, we identified the same factor enriched in Peptide chain elongation pathway. Another important predictor of survival is Factor 4, with top weighted genes in the protein view: ARGLU1, SULT1E1, CEACAM5, AKR1B10 and ING4. And the top enriched pathway is Extracellular matrix organization. ARGLU1 has been described as new MED1-interacting protein for breast cancer cell growth (Zhang et al. 2011). Genetic polymorphisms of SULT1E1 were found to be associated with increased risk and a disease free survival of breast cancer (Choi et al. 2005). CEACAM6 plays a role in tumor cell migration, invasion and adhesion, and formation of distant metastases (Choi et al. 2005; Blumenthal, Hansen, and Goldenberg 2005). AKR1B10 promotes breast cancer metastasis through FAK/Src/Rac1 signaling pathway(Huang et al. 2016). ING4 inhibits estrogen receptor activity in breast cancer cells (Keenen and Kim 2016). For most of the factors, protein view and mRNA view are in agreement for top enriched pathways.

In overall, predicted proteomics identified more predictive factors than mRNA, with comparable predictive performance. It was also able to identify new insights as predictor of survival, which were not found using mRNA.

2.2.7 Validation on colorectal cancer cell line

We applied the best performing model (from ovarian) on 47 colorectal cancer cell lines(Roumeliotis et al. 2017). The input feature is microarray mRNA and the real proteomics data was acquired by isobaric peptide labeling (TMT-10plex) and MS3 quantification. There are 3039 proteins in common between the DREAM CPTAC predicted and the ones measured in the reference paper. The average correlation between the predicted protein and the real protein abundance is 0.36 (from -0.37 to 0.88), which is an encouraging result considering that the model was: (i) trained on RNAseq and tested on microarray. (ii) trained on proteomics acquired using iTRAQ and tested on proteomics acquired by isobaric peptide labeling (TMT-10plex) and MS3 quantification.

2.2.8 Application to drug response prediction

We applied the winning model (from ovarian) to 990 cancer cell lines from the Genomics of Drug Sensitivity in Cancer (GDSC) screening(Iorio et al. 2016c). Each cell line is treated by 265 drugs. We used Elastic net algorithm to predict drug response IC50 (Concentration of Inhibition at 50% viability) using 10 fold cross validation, repeated 100 times. The average predictive performance for the 265 drugs is 0.41 using mRNA and 0.40 using predicted proteomics. We found no significant difference between the two groups in term of performance ($p=0.37$).

2.3 Methods

2.3.1 Challenge data

TCGA retrospective collection of breast and ovarian tumor samples quantitatively measured at four biological levels (proteomics, phosphoproteomics, transcriptomics (mRNA) and copy number alterations (CNA) were used as training data for CPTAC-NCI dream challenge. Prospective samples of the same cancer types with all four level measurement was generated and used as testing data for performance evaluation. Sample size varies between different platforms due to the availability and quality of original tumor samples at the time of the study. Mass-Spectrometry based proteomic and phosphoproteomic characterization of these tumor samples yield more than hundred of thousand protein and phosphosite identifications combined, which will serve as the target to be predicted in the sub-challenges.

Training Data

Breast cancer:

- Proteome: 10005 proteins for 105 patients
- Phosphoproteome: 31981 phospho-sites for 105 patients
- CNA: 16884 genes for 77 patients
- mRNA: 15107 genes for 77 patients

Ovarian cancer:

- Proteome from PNNL: 7061 proteins for 84 patients
- Proteome from JHU: 7061 proteins for 122 patients
- Phosphoproteome: 10057 phosphosites for 69 patients
- CNA: 11859 genes for 559 patients
- mRNA(Array): 15121 genes for 569 patients
- mRNA(RNA-seq): 15121 genes for 294 patients

Training proteomics and phosphoproteomics data of breast and ovarian tumors were downloaded from CPTAC data portal and processed by the common data analysis pipeline from CPTAC. For both tissues, global proteome and phosphoproteome data were acquired using iTRAQ (isobaric Tags for Relative and Absolute Quantification) protein quantification methods as described previously. For breast proteome, 105 (77 passed QC) tumors from different patients were analyzed at the Broad Institute. The protein log ratios of the protein abundance were calculated including only peptides that map unambiguously to the protein. Breast tumor samples ('TCGA-AO-A12B', 'TCGA-AO-A12D', 'TCGA-C8-A131' assayed in duplicate for quality control purposes) were mean aggregated in the uploaded training data (https://cptac-data-portal.georgetown.edu/cptac/s/S015). For ovarian proteome, there are 206 samples from 174 unique patients (84 from Pacific Northwest National Laboratory (PNNL), 122 from Johns Hopkins University and 32 measured by both centers). We provided participants with both proteome collections for training to cover the maximum number of samples for Proteomics subchallenge.

CNA data were directly downloaded from two CPTAC publications. Non unique gene IDs were median aggregated. Transcriptomics data for both cancer types were downloaded from TCGA firehose. RNA-seq (RSEM z-score, median aggregated) were chosen for breast cancer. Microarray data and RNA-seq data were both downloaded for participants to use for ovarian cancer. The main reasons of providing participants with both datasets are sample coverage is greater between microarray and proteome, however only RNA-seq was performed for prospective collection.

Testing Data

CNA, RNA-seq, Proteome from prospectively collected patients were provided as testing data.

First and second round:
Ovarian cancer:

- Proteome: 7061 proteins for 20 patients
- Phosphoproteome: 10057 phosphosites for 20 patients
- CNA: 11859 genes for 20 patients
- mRNA(RNA-seq): 15121 genes for 20 patients

Final round:

Breast cancer:

- Proteome: 10005 proteins for 108 patients
- Phosphoproteome: 31981 phospho-sites for 108 patients
- CNA: 16884 genes for 108 patients
- mRNA: 15107 genes for 108 patients

Ovarian cancer:

- Proteome: 7061 proteins for 62 patients
- Phosphoproteome: 10057 phosphosites for 62 patients
- CNA: 11859 genes for 62 patients
- mRNA(RNA-seq): 15121 genes for 62 patients

The training data has been prepared by both Zhi Li (New York University) and I. The testing data has been exclusively prepared by Zhi Li.

2.3.2 Scoring

Proteomics subchallenge: prediction of protein abundance based on mRNA

These models are evaluated in two novel, unpublished held-out datasets: ovarian and breast tissue. We will only focus on 5220 proteins for ovarian and 8649 proteins for breast with less than 30% missing values in both training and testing data. We first compute the pearson correlation between observed and predicted abundances across all samples for each protein. We then take the mean correlations of proteins in the test data set as the final evaluation score. If there is a tie, we will further use NRMSE for all proteins to select the winner.

Phosphoproteomics subchallenge: prediction of phosphosites abundance based on protein abundance

We will only focus on 1318 phosphosites for ovarian and 4907 phosphosites for breast with less than 30% missing values in both training and testing data. We first compute the pearson correlation between observed and predicted phosphosite abundances across all samples for each phosphoprotein. We then take the mean correlations of phosphoproteins in the test data set as the final evaluation score. If there is a tie, we will further use NRMSE for all phosphoproteins to select the winner.

2.3.3 Winning method

The winning team (Hongyang Li, Yuanfang Guan, University of Michigan) used a weighted average of four major models for prediction (**Figure 6**):

Protein proxy model: This model is based on the observations that protein and transcript levels are correlated, and simply uses the transcript level of a given gene as a proxy for its protein level. Missing values positions are replaced with the gene average across non-missing samples. This model has several limitations including that it assumes no differential translational regulation and degradation, and it disregards interaction between genes.

Interaction model: Since different genes are regulated differently, individual models were built for each gene using random forests with maximum depth of 5 and 100 trees. The response variables for training are the non-missing observations across all samples, and as features the values for all genes as training features to take into account gene interactions.

Pan-cancer model: The performance of individual model is limited by the sample size. The training data only contain 77 and 174 tumors for breast and ovarian, respectively. This is a relatively small sample size, but when combining all the samples, a better performance was

achieved as the majority of genes have similar regulation across different tissues (Wang, Zhang, and Du 2013).

For the phosphorylation prediction task, the proxy model was changes to use protein levels instead of transcript levels and a fourth model was added:

Phosphorylation proxy model: This model is based on the observations that protein and phosphorylation levels are correlated albeit only modestly, and simply uses the protein level of a given gene as a proxy for its phosphorylation level. This model assumes that for any given gene a constant fraction of the proteins are phosphorylated.

Phosphosite correlation model: The levels of multiple phosphorylation sites from the same protein are not independent. The biological rationale behind this model is that if a protein is phosphorylated, it is likely that multiple phosphosites are phosphorylated co-regulated. In addition, for technical limitations it is sometimes not possible to distinguish two phosphosites that are very close in the linear sequence so that they are in the same peptide after digestion and no fragment peaks are observed from fragmentation between them. Therefore, a phosphorylation site is correlated with other phosphorylation sites on the same protein, the winning team utilized this and calculated the weighted average prediction from all phosphorylation sites of the same gene as the multi-site prediction.

2.3.4 Ensemble method from top performers

We next sought to further improve the prediction by the organizing a collaboration round between the top 4 ranked teams (led by Francesca Petralia, Icahn Institute). Each team has for objectives to improve their own methods and to incorporate other teams' specificities (**Figure 6**).

Team **hyu** used Random Forest algorithm and mRNA as proxy. mRNA features were filtered based on KEGG signaling pathway and human PPI network information, as well as their correlation with the responses. When applicable, all neighbors within a distance of 2 from a protein, in either KEGG pathway or PPI network, were selected. This set is then expanded by all genes belonging to pathways of mRNA surveillance, RNA degradation, RNA polymerase, basal transcription factors, cell cycle, protein processing in endoplasmic reticulum and microRNAs in cancer, because these pathways are assumed to play important roles in regulating translation.

Team **DEARGENpg** predicted protein and phosphoprotein abundances, by grouping genes into signalling pathways, selected 300 Protein, CNA, mRNA features using Pearson correlation scores and an ensemble model of XGboost, Extra Tree and Random Forest, using stacking. PAM50 breast subtypes (basal-like&HER2 and LuminalA&LuminalB) have been added to each patient (0 or 1) using PCA in 2 dimensions and then K-means clustering. Gene meta information were added considering Codon count, GC count and Folding energy.

Team **DMIS_PTG** trained models for each protein using LASSO regression. Features were selected based on PPI networks (BioGRID, BTNET and CORUM) and biological pathways. The genes which are 1-hop from the target protein gene in the union network explained above. The

average of the number of these features for all protein models is approximately 20. Gene sets were selected using MSigDB where all the component genes are included in the training gene expression data. Then, a median value of the expression value for each gene set were used as feature. Around 700 median values of gene sets were included. Both microarray and RNAseq data were used for training. Although the testing data only provides RNAseq, microarray training data contain more sample than RNAseq.

"To improve the prediction performance, we assembled the models of the top 4 teams from the challenge. By analysing the 5-fold cross validation results of these models on the training data, the prediction correlation of each protein was calculated. For each protein, the correlation scores were used as the stacking weights of these top 4 models (hereafter referred to as the individual ensemble model). To estimate the overall performance, the average correlation of all proteins was calculated and used as the weights for all proteins (hereafter referred to as the global ensemble model). For the ovarian cancer, we observed a significant improvement (0.5605) of the global ensemble model, compared with the best performer in the challenge final round (0.5284), $p<2.2e-16$. However, the improvement of the global ensemble model is very marginal in breast cancer, only from 0.5052 to 0.5063, $p=1$. We further calculated the normalized root mean square error (NRMSE) of these model and found that the global ensemble model reduced the error from 0.1863 to 0.1750 in the ovarian cancer. Similar correlation scores were observed for predictions of the individual ensemble model in both breast and ovarian cancers." (**Analysis performed and text written by Hongyang Li, University of Michigan**)

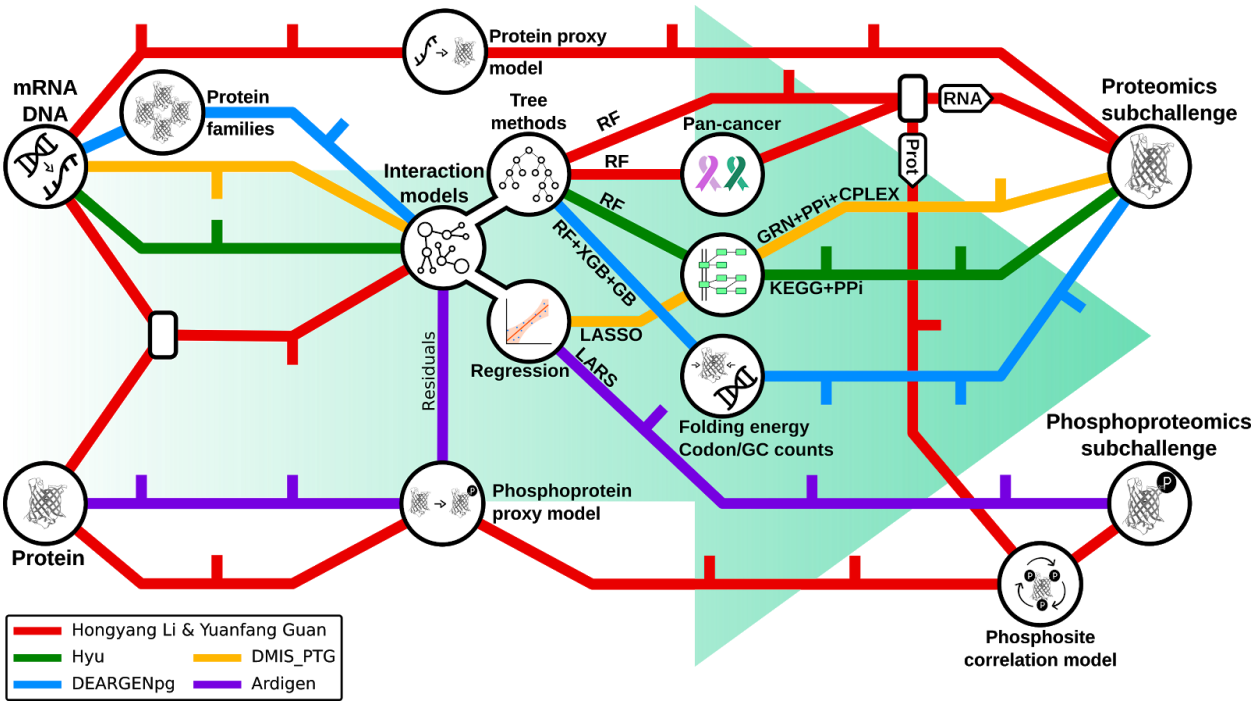


Figure 6: Prediction models of the best performers.

The input data are on the left (mRNA, DNA, proteomics) and the prediction output are on the right (Proteomics and Phosphoproteomics subchallenges). Each team starts with the input data on the left and navigates to the right side, the stations they crossed representing the methods they used.

Paths taken by each team for the Proteomics Subchallenge:
1) Team Hongyang Li and Yuanfang Guan used Protein proxy model, Random forest and Pan cancer model.
2) Team Hyu used Random forest. Features were selected by KEGG pathway and PPI (Human Protein Reference Database).
3) Team DEARGENpg built models on a groups of proteins, used ensemble of Random Forest+XGboost+Gradient Boost, and additional features such as gene metadata (codon bias, GC count and folding energy of each protein).
4) Team DMIS_PTG used LASSO. Features were selected based on Gene Regulatory Network, CORUM protein complexes, PPI, and LASSO.

Paths taken by each team for the Phosphoproteomics Subchallenge:

1) Team Hongyang Li and Yuanfang Guan used Phosphoprotein proxy model, Random forest, Pan cancer model and Phosphoprotein correlation model.
2) Team Ardigen used Phosphoprotein proxy model and used algorithm LARS.

Figure made by Nicolàs Palacio and designed by both of us.

2.3.5 Multi Omics Factor Analysis

MOFA is a statistical model which can identify the principal sources of variation in multi omics datasets (Argelaguet et al. 2017). It infers a set of hidden variables (Factors) that capture the biological/technical variability of the dataset. MOFA takes an arbitrary number of data matrices (omics layers) (Y^1,..., Y^M) with co-occurrent samples but possibly differing number of features. MOFA decomposes these matrices into a matrix of factors, Z, for each sample and M weight matrices, one for each view (loadings W^1,.., W^M). MOFA approximates the true posterior using a variational distribution in a factorized form, which is optimized by minimizing the lower bound of the marginal likelihood (also called evidence lower bound, ELBO). Unlike the standard principal component analysis, each Factor can be defined by several layers of information (proteomics AND/OR mRNA AND/OR mutation...etc). Kernel and graph based methods also allow to combine multiple omics data, but those approaches suffer from the lack interpretability. Our motivation for using MOFA is the panoply of downstream analysis to biologically define a factor of interest and associate it with clinical phenotype (such as patient survival). In addition to that, MOFA is highly scalable and handle missing values and non-gaussian omics layers (such as mutation data). Gaussian distribution was used to model protein and mRNA's likelihood. For binary mutation, we used Bernoulli. We iterated until convergence.

2.4 Discussion

The winning method of this competition is an ensemble of four models, which consist in: (i) Using input feature as proxy of the response variable (Generic model). (ii) Modeling each protein based on mRNA expression of other genes, with Random Forest. (iii) Including another tissue in training phase (Trans-tissue model). (iv) Modeling phosphosite abundance based on the biology and the mass spectrometry technology.

Unsurprisingly, proteins were better predicted when the corresponding mRNA is available. Proteins outside a protein complex were also better predicted than those belonging to a complex. The best predicted families were Aldehyde dehydrogenase, Acyl-CoA synthetase, Integrin alpha subunits and Glutathione S-transferase. The least predicted families were histones and ribosomal proteins. Analysis of commonly protein regulators revealed key proteins predictive of patient survival *e.g.* WDR12 and PLCD1.

We then assessed the utility of the predicted proteins using: (i) TCGA samples of breast tissue which were not used in the challenge for survival analysis and (ii) Cancer cell line for drug response prediction.

We used Multi Omics factor Analysis to reduce the proteomics/mRNA dimension and used breast tissue to assess the predicted proteomics' performance in patient stratification. Predicted protein's performance is comparable to mRNA, and revealed more predictive biomarker.

Predicted proteomics identified L ribosomal proteins contributing to patient stratification, which were not found using mRNA. Another group of proteins identified by combining proteomics and mRNA, were ARGLU1, SULT1E1, CEACAM5, AKR1B10 and ING4, many of them involved in cell migration and Extracellular matrix organization. Those results suggest the use of predicted protein abundance to explore biological insights under a new angle.

We applied the winning model to predict protein level on 48 colorectal cancer cell lines from the GDSC screening, and reached an average correlation of 0.36 between predicted protein level and real protein level, despite the difference of technology used in training and testing. Drug response prediction based on predicted proteomics is comparable to using mRNA, which is encouraging considering that the model was built on mRNA. These results suggest the potential use of predicted proteomics to facilitate drug mode of action elucidation and improve therapeutic decisions.

Regarding prediction of protein level based on mRNA, Wilhelm *et al*. reported correlations of approximately 0.9 between observed versus predicted protein levels (Wilhelm et al. 2014) and concluded that protein abundance can be predicted with good accuracy from the corresponding gene's mRNA levels. Fortelny *et al*. replied that the model was built within genes, whereas the assessment of performance was done across genes (Fortelny et al. 2017). This is due to a discrepancy between model building and model assessment. The reported performance score is not generalizable to new experiments.

Cancer drugs are mostly targeting proteins. Therefore, protein targets with low/anti correlated protein/mRNA could lead to therapeutic mistakes. If the measured protein abundance reflects better the biology, then using mRNA to make clinical decision would be a mistake. On the other hand, if mRNA reflects the biology better than the measured proteomics, this could point to potential directions of improvement of the mass spectrometry technology or data processing. To determine which situation prevails, domain specific gold standards are needed.

General conclusions and outlook

We presented in this thesis different applications of machine learning in knowledge discovery for systems pharmacology and cancer biology. We mined the largest public available databases of cancer drug screenings (GDSC, CTRPv2 and CCLE), and drug combination data from AstraZeneca and Merck. We then worked with the largest primary tumor databases: TCGA for mRNA and CPTAC for proteomics. Mining public available databases has several advantages over generating your own data: (i) The sample size for a given omics layer is much bigger, thus allowing more exploration, discovery, stronger statistical power and use of more sophisticated algorithms. (ii) The overlapping samples between different omics layers is also bigger, allowing more exploration of the underlying associations between omics layers. We argue that biomedical scientists should always start with an initial step of data driven hypothesis generation or confirmation, and then experimentally validate the hypothesis. This could significantly increase the success rate and better allocation of research time and fund.

Multitask learning and matrix factorization have been successfully applied to preclinical drug response prediction on cancer cell lines. Since the response data is in a matrix format (cell lines treated by drugs), this class of algorithm can easily capture the underlying associations between the descriptors of the cell lines and the descriptors of the drugs. Such association is on a deeper level than a simple drug-gene association, and could potentially be applied to target discovery, drug repurposing and patient stratification.

Machine learning has been widely used for prediction purpose. In the drug combination project, we used Macau, a multitask learning algorithm, in a unsupervised way for hypothesis generation. Based on the generated hypothesis, we built specific models for prediction. We applied this workflow to drug synergy discovery and prediction. This method could: (i) Predict whether a pair of compounds could be synergistic for a given tissue, therefore used as a compound prioritization framework. (ii) Predict synergy of new drug combinations on new cell lines, without performing any experiments.

In this last project, we explored through a collaborative machine learning competition, the relations between protein level and mRNA expression. The best performing algorithm can accurately predict protein level on tumor samples from mRNA. Since cancer drugs mostly target proteins, the possibility to explore the protein level of tumor samples, could potentially reveal new facets of cancer biology.

In this thesis, we successfully applied machine learning to preclinical drug development and cancer biology in real patient tumors. Due to the availability of data, matrix factorization is especially suited for in vitro drug screenings. 3D Tensor factorization could also be used, provided the availability of a third mode, which could be drug concentration or treatment by an additional drug (combination). In clinical scenario, such class of algorithm is unlikely to be applied for response prediction (patient survival or drug response), as this would require treating the same patient with hundreds of drugs. Therefore, the most commonly used algorithm for

clinical data are single task Cox regression, linear regression and Random Forest. Nevertheless, Tensor factorization and Factor Analysis could be applied to reduce the dimensions of multi omics data (Transcriptomics, Proteomics, Metabolomics...etc). The resulting Factors or latent matrix could be used as input features for prediction purpose.

Cancer drug screenings most often only focus on drugs targeting intracellular processes of the cancer cells (Dry, Yang, and Saez-Rodriguez 2016b). It is essential to also consider primary tumors and other systems, as cancer cell lines, although the best existing model to study drugs' mode of action, do not take into consideration the immune system nor the 3D structure of the in vivo tumor. In the future, ex vivo tumor culture (such as organoids and patient-derived xenografts) could be used to reproduce the drug response matrix as in preclinical drug screenings. Organoids could mimic in vivo architecture of the tumor within an organ (Dutta, Heo, and Clevers 2017), therefore more realistic than in vitro cell line experiments. Patient-derived xenografts are immunodeficient mice implanted with patients' tumors and are currently the best in vivo system beside the patient (Lai et al. 2017).

Clinical trial data and electronic health record are likely to play an important role in precision medicine. However, clinical trial data are difficult to obtain and electronic health record could be sparse and noisy. Biosensors and smart wearables are promising ways for real-time monitoring of patient response, health, and adverse events (Dry, Yang, and Saez-Rodriguez 2016b; Cleeland et al. 2012). Such technologies are more efficient in collecting data than from a hospital (expensive and slow).

Big data approaches could significantly improve the drug development process, decrease the cost of making a new drug, and potentially reducing the need for animal experiments. However, such perspectives are often perceived as risky by the pharmaceutical industry which most often prefers to do what has always been done. In this context of drug development attrition, industry-academia collaborations are evermore needed.

In this thesis, we extensively used machine learning for cancer bioinformatics, from preclinical drug screenings to patient outcome prediction. We hope that those results could be used for new drug developments, repositioning of old drugs, synergy and biomarker discovery.

References

Al-Lazikani, Bissan, Udai Banerji, and Paul Workman. 2012. “Combinatorial Drug Therapy for Cancer in the Post-Genomic Era.” *Nature Biotechnology* 30 (7): 679–92.

Argelaguet, Ricard, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C. Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. 2017. “Multi-Omics Factor Analysis Disentangles Heterogeneity in Blood Cancer.” *bioRxiv*. https://doi.org/10.1101/217554.

Bansal, Mukesh, Jichen Yang, Charles Karan, Michael P. Menden, James C. Costello, Hao Tang, Guanghua Xiao, et al. 2014. “A Community Computational Challenge to Predict the Activity of Pairs of Compounds.” *Nature Biotechnology* 32 (12): 1213–22.

Belin, Stéphane, Anne Beghin, Eduardo Solano-Gonzàlez, Laurent Bezin, Stéphanie Brunet-Manquat, Julien Textoris, Anne-Catherine Prats, Hichem C. Mertani, Charles Dumontet, and Jean-Jacques Diaz. 2009. “Dysregulation of Ribosome Biogenesis and Translational Capacity Is Associated with Tumor Progression of Human Breast Cancer Cells.” *PloS One* 4 (9): e7147.

Berenbaum, M. C. 1989. “What Is Synergy?” *Pharmacological Reviews* 41 (2): 93–141.

Blumenthal, Rosalyn D., Hans J. Hansen, and David M. Goldenberg. 2005. “Inhibition of Adhesion, Invasion, and Metastasis by Antibodies Targeting CEACAM6 (NCA-90) and CEACAM5 (Carcinoembryonic Antigen).” *Cancer Research* 65 (19): 8809–17.

Bratton, Melyssa R., Bich N. Duong, Steven Elliott, Christopher B. Weldon, Barbara S. Beckman, John A. McLachlan, and Matthew E. Burow. 2010. “Regulation of ERalpha-Mediated Transcription of Bcl-2 by PI3K-AKT Crosstalk: Implications for Breast Cancer Cell Survival.” *International Journal of Oncology* 37 (3): 541–50.

Bufalo, Donatella Del, Donatella Del Bufalo, Daniela Trisciuoglio, and Michele Milella. 2004. “Crosstalk between VEGF and Bcl-2 in Tumor Progression and Angiogenesis.” In *VEGF and Cancer*, 26–39.

Bulusu, Krishna C., Rajarshi Guha, Daniel J. Mason, Richard P. I. Lewis, Eugene Muratov, Yasaman Kalantar Motamedi, Murat Cokol, and Andreas Bender. 2016. “Modelling of Compound Combination Effects and Applications to Efficacy and Toxicity: State-of-the-Art, Challenges and Perspectives.” *Drug Discovery Today* 21 (2): 225–38.

Cantini, Laura, Laurence Calzone, Loredana Martignetti, Mattias Rydenfelt, Nils Blüthgen, Emmanuel Barillot, and Andrei Zinovyev. 2018. “Classification of Gene Signatures for Their Information Value and Functional Redundancy.” *NPJ Systems Biology and Applications* 4: 2.

Carboni, Joan M., Mark Wittman, Zheng Yang, Francis Lee, Ann Greer, Warren Hurlburt, Stephen Hillerman, et al. 2009. “BMS-754807, a Small Molecule Inhibitor of Insulin-like Growth Factor-1R/IR.” *Molecular Cancer Therapeutics* 8 (12): 3341–49.

Castilla, María Ángeles, María Ángeles López-García, María Reina Atienza, Juan Manuel Rosa-Rosa, Juan Díaz-Martín, María Luisa Pecero, Begoña Vieites, et al. 2014. “VGLL1 Expression Is Associated with a Triple-Negative Basal-like Phenotype in Breast Cancer.” *Endocrine-Related Cancer* 21 (4): 587–99.

Chen, Chun-Te, Yi Du, Hirohito Yamaguchi, Jung-Mao Hsu, Hsu-Ping Kuo, Gabriel N. Hortobagyi, and Mien-Chie Hung. 2012. “Targeting the IKKβ/mTOR/VEGF Signaling Pathway as a Potential Therapeutic Strategy for Obesity-Related Breast Cancer.” *Molecular Cancer Therapeutics* 11 (10): 2212–21.

Choi, Ji-Yeob, Kyoung-Mu Lee, Sue Kyung Park, Dong-Young Noh, Sei-Hyun Ahn, Hye-Won Chung, Wonshik Han, et al. 2005. “Genetic Polymorphisms of SULT1A1 and SULT1E1 and the Risk and Survival of Breast Cancer.” *Cancer Epidemiology, Biomarkers & Prevention: A*

Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology 14 (5): 1090–95.
Cidado, Justin, and Ben Ho Park. 2012. “Targeting the PI3K/Akt/mTOR Pathway for Breast Cancer Therapy.” *Journal of Mammary Gland Biology and Neoplasia* 17 (3-4): 205–16.
Cleeland, Charles S., Jeff D. Allen, Samantha A. Roberts, Joanna M. Brell, Sergio A. Giralt, Aarif Y. Khakoo, Rebecca A. Kirch, Virginia E. Kwitkowski, Zhongxing Liao, and Jamey Skillings. 2012. “Reducing the Toxicity of Cancer Therapy: Recognizing Needs, Taking Action.” *Nature Reviews. Clinical Oncology* 9 (8): 471–78.
Condorelli, R., and F. André. 2017. “Combining PI3K and PARP Inhibitors for Breast and Ovarian Cancer Treatment.” *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 28 (6): 1167–68.
Di Veroli, Giovanni Y., Chiara Fornari, Dennis Wang, Séverine Mollard, Jo L. Bramhall, Frances M. Richards, and Duncan I. Jodrell. 2016. “Combenefit: An Interactive Platform for the Analysis and Visualization of Drug Combinations.” *Bioinformatics* 32 (18): 2866–68.
Dry, Jonathan R., Mi Yang, and Julio Saez-Rodriguez. 2016a. “Looking beyond the Cancer Cell for Effective Drug Combinations.” *Genome Medicine* 8 (1): 125.
———. 2016b. “Looking beyond the Cancer Cell for Effective Drug Combinations.” *Genome Medicine* 8 (1): 125.
Dutta, Devanjali, Inha Heo, and Hans Clevers. 2017. “Disease Modeling in Stem Cell-Derived 3D Organoid Systems.” *Trends in Molecular Medicine* 23 (5): 393–410.
Fitzgerald, Jonathan B., Birgit Schoeberl, Ulrik B. Nielsen, and Peter K. Sorger. 2006. “Systems Biology and Combination Therapy in the Quest for Clinical Efficacy.” *Nature Chemical Biology* 2 (9): 458–66.
Fortelny, Nikolaus, Christopher M. Overall, Paul Pavlidis, and Gabriela V. Cohen Freue. 2017. “Can We Predict Protein from mRNA Levels?” *Nature* 547 (July): E19.
Glaysher, Sharon, Louise M. Bolton, Penny Johnson, Christopher Torrance, and Ian A. Cree. 2014. “Activity of EGFR, mTOR and PI3K Inhibitors in an Isogenic Breast Cell Line Model.” *BMC Research Notes* 7 (June): 397.
Goudarzi, Kaveh M., and Mikael S. Lindström. 2016. “Role of Ribosomal Protein Mutations in Tumor Development (Review).” *International Journal of Oncology* 48 (4): 1313–24.
Hamunyela, Roswita H., Antonio M. Serafin, and John M. Akudugu. 2017. “Strong Synergism between Small Molecule Inhibitors of HER2, PI3K, mTOR and Bcl-2 in Human Breast Cancer Cells.” *Toxicology in Vitro: An International Journal Published in Association with BIBRA* 38 (February): 117–23.
Hrustanovic, Gorjan, and Trever G. Bivona. 2015. “RAS-MAPK in ALK Targeted Therapy Resistance.” *Cell Cycle* 14 (23): 3661–62.
Huang, Chenfei, Steven Verhulst, Yi Shen, Yiwen Bu, Yu Cao, Yingchun He, Yuhong Wang, et al. 2016. “AKR1B10 Promotes Breast Cancer Metastasis through Integrin α5/δ-Catenin Mediated FAK/Src/Rac1 Signaling Pathway.” *Oncotarget* 7 (28): 43779–91.
Iorio, Francesco, Theo A. Knijnenburg, Daniel J. Vis, Graham R. Bignell, Michael P. Menden, Michael Schubert, Nanne Aben, et al. 2016a. “A Landscape of Pharmacogenomic Interactions in Cancer.” *Cell* 166 (3): 740–54.
———. 2016b. “A Landscape of Pharmacogenomic Interactions in Cancer.” *Cell* 166 (3): 740–54.
———. 2016c. “A Landscape of Pharmacogenomic Interactions in Cancer.” *Cell* 166 (3): 740–54.
Jaeger, Samira, Ana Igea, Rodrigo Arroyo, Victor Alcalde, Begoña Canovas, Modesto Orozco, Angel R. Nebreda, and Patrick Aloy. 2017. “Quantification of Pathway Cross-Talk Reveals Novel Synergistic Drug Combinations for Breast Cancer.” *Cancer Research* 77 (2): 459–69.
Keenen, Madeline M., and Suwon Kim. 2016. “Tumor Suppressor ING4 Inhibits Estrogen Receptor Activity in Breast Cancer Cells.” *Breast Cancer* 8 (November): 211–21.

Kikuchi, Hirotoshi, Maria S. Pino, Min Zeng, Senji Shirasawa, and Daniel C. Chung. 2009. “Oncogenic KRAS and BRAF Differentially Regulate Hypoxia-Inducible Factor-1alpha and -2alpha in Colon Cancer.” *Cancer Research* 69 (21): 8499–8506.

Kosti, Idit, Nishant Jain, Dvir Aran, Atul J. Butte, and Marina Sirota. 2016. “Cross-Tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues.” *Scientific Reports* 6 (May): 24799.

Lai, Yunxin, Xinru Wei, Shouheng Lin, Le Qin, Lin Cheng, and Peng Li. 2017. “Current Status and Perspectives of Patient-Derived Xenograft Models in Cancer Research.” *Journal of Hematology & Oncology* 10 (1): 106.

Lewinska, Anna, Diana Bednarz, Jagoda Adamczyk-Grochala, and Maciej Wnuk. 2017. “Phytochemical-Induced Nucleolar Stress Results in the Inhibition of Breast Cancer Cell Proliferation.” *Redox Biology* 12 (August): 469–82.

Liu, Yansheng, Andreas Beyer, and Ruedi Aebersold. 2016. “On the Dependency of Cellular Protein Levels on mRNA Abundance.” *Cell* 165 (3): 535–50.

Loewe, S. 1928. “Die Quantitativen Probleme Der Pharmakologie.” *Ergebnisse Der Physiologie, Biologischen Chemie Und Experimentellen Pharmakologie* 27 (1): 47–187.

———. 1953. “The Problem of Synergism and Antagonism of Combined Drugs.” *Arzneimittel-Forschung* 3 (6): 285–90.

Lopez, Juanita S., and Udai Banerji. 2017. “Combine and Conquer: Challenges for Targeted Therapy Combinations in Early Phase Trials.” *Nature Reviews. Clinical Oncology* 14 (1): 57–66.

Menden, Michael Patrick, Dennis Wang, Yuanfang Guan, Michael Mason, Bence Szalai, Krishna C. Bulusu, Thomas Yu, et al. 2017a. “Community Assessment of Cancer Drug Combination Screens Identifies Strategies for Synergy Prediction.” https://doi.org/10.1101/200451.

———. 2017b. “Community Assessment of Cancer Drug Combination Screens Identifies Strategies for Synergy Prediction.” https://doi.org/10.1101/200451.

Mitra, S. 2017. “MicroRNA Therapeutics in Triple Negative Breast Cancer.” https://www.researchgate.net/profile/Sarmistha_Mitra/publication/319529911_Archives_of_Pathology_and_Clinical_Research_MicroRNA_Therapeutics_in_Triple_Negative_Breast_Cancer/links/59b171e20f7e9b37434ab783/Archives-of-Pathology-and-Clinical-Research-MicroRNA-Therapeutics-in-Triple-Negative-Breast-Cancer.pdf.

Montagut, Clara, Sreenath V. Sharma, Toshi Shioda, Ultan McDermott, Matthew Ulman, Lindsey E. Ulkus, Dora Dias-Santagata, et al. 2008. “Elevated CRAF as a Potential Mechanism of Acquired Resistance to BRAF Inhibition in Melanoma.” *Cancer Research* 68 (12): 4853–61.

Moore, Nathan F., Anna M. Azarova, Namrata Bhatnagar, Kenneth N. Ross, Lauren E. Drake, Stacey Frumm, Qinsong S. Liu, et al. 2014. “Molecular Rationale for the Use of PI3K/AKT/mTOR Pathway Inhibitors in Combination with Crizotinib in ALK-Mutated Neuroblastoma.” *Oncotarget* 5 (18): 8737–49.

Nass, Norbert, Angela Dittmer, Vicky Hellwig, Theresia Lange, Johanna Mirjam Beyer, Benjamin Leyh, Atanas Ignatov, et al. 2016. “Expression of Transmembrane Protein 26 (TMEM26) in Breast Cancer and Its Association with Drug Response.” *Oncotarget* 7 (25): 38408–26.

Nazemalhosseini Mojarad, Ehsan, Roya Kishani Farahani, Mahdi Montazer Haghighi, Hamid Asadzadeh Aghdaei, Peter Jk Kuppen, and Mohammad Reza Zali. 2013. “Clinical Implications of BRAF Mutation Test in Colorectal Cancer.” *Gastroenterology and Hepatology from Bed to Bench* 6 (1): 6–13.

O’Neil, Jennifer, Yair Benita, Igor Feldman, Melissa Chenard, Brian Roberts, Yaping Liu, Jing Li, et al. 2016. “An Unbiased Oncology Compound Screen to Identify Novel Combination Strategies.” *Molecular Cancer Therapeutics* 15 (6): 1155–62.

Palmer, Adam C., and Peter K. Sorger. 2017. “Combination Cancer Therapy Can Confer Benefit via Patient-to-Patient Variability without Drug Additivity or Synergy.” *Cell* 171 (7): 1678–91.e13.
Pan, Sinno Jialin, and Qiang Yang. 2010. “A Survey on Transfer Learning.” *IEEE Transactions on Knowledge and Data Engineering* 22 (10): 1345–59.
Paplomata, Elisavet, and Ruth O’Regan. 2014. “The PI3K/AKT/mTOR Pathway in Breast Cancer: Targets, Trials and Biomarkers.” *Therapeutic Advances in Medical Oncology* 6 (4): 154–66.
Parikh, Jignesh R., Bertram Klinger, Yu Xia, Jarrod A. Marto, and Nils Blüthgen. 2010. “Discovering Causal Signaling Pathways through Gene-Expression Patterns.” *Nucleic Acids Research* 38 (Web Server issue): W109–17.
Preuer, Kristina, Richard P. I. Lewis, Sepp Hochreiter, Andreas Bender, Krishna C. Bulusu, and Günter Klambauer. 2017. “DeepSynergy: Predicting Anti-Cancer Drug Synergy with Deep Learning.” *Bioinformatics* , December. https://doi.org/10.1093/bioinformatics/btx806.
Rehman, Farah L., Christopher J. Lord, and Alan Ashworth. 2012. “The Promise of Combining Inhibition of PI3K and PARP as Cancer Therapy.” *Cancer Discovery* 2 (11): 982–84.
Roumeliotis, Theodoros I., Steven P. Williams, Emanuel Gonçalves, Clara Alsinet, Martin Del Castillo Velasco-Herrera, Nanne Aben, Fatemeh Zamanzad Ghavidel, et al. 2017. “Genomic Determinants of Protein Abundance Variation in Colorectal Cancer Cells.” *Cell Reports* 20 (9): 2201–14.
Saini, Kamal S., Sherene Loi, Evandro de Azambuja, Otto Metzger-Filho, Monika Lamba Saini, Michail Ignatiadis, Janet E. Dancey, and Martine J. Piccart-Gebhart. 2013. “Targeting the PI3K/AKT/mTOR and Raf/MEK/ERK Pathways in the Treatment of Breast Cancer.” *Cancer Treatment Reviews* 39 (8): 935–46.
Schmid, K., Z. Bago-Horvath, W. Berger, A. Haitel, D. Cejka, J. Werzowa, M. Filipits, B. Herberger, H. Hayden, and W. Sieghart. 2010. “Dual Inhibition of EGFR and mTOR Pathways in Small Cell Lung Cancer.” *British Journal of Cancer* 103 (5): 622–28.
Schubert, Michael, Bertram Klinger, Martina Klünemann, Anja Sieber, Florian Uhlitz, Sascha Sauer, Mathew J. Garnett, Nils Blüthgen, and Julio Saez-Rodriguez. 2018. “Perturbation-Response Genes Reveal Signaling Footprints in Cancer Gene Expression.” *Nature Communications* 9 (1): 20.
Simm, J., A. Arany, P. Zakeri, T. Haber, J. K. Wegner, V. Chupakhin, H. Ceulemans, and Y. Moreau. 2017a. “Macau: Scalable Bayesian Factorization with High-Dimensional Side Information Using MCMC.” In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*. https://doi.org/10.1109/mlsp.2017.8168143.
———. 2017b. “Macau: Scalable Bayesian Factorization with High-Dimensional Side Information Using MCMC.” In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*. https://doi.org/10.1109/mlsp.2017.8168143.
———. 2017c. “Macau: Scalable Bayesian Factorization with High-Dimensional Side Information Using MCMC.” In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*. https://doi.org/10.1109/mlsp.2017.8168143.
Sun, Yi, Zhen Sheng, Chao Ma, Kailin Tang, Ruixin Zhu, Zhuanbin Wu, Ruling Shen, et al. 2015. “Combining Genomic and Network Characteristics for Extended Capability in Predicting Synergistic Drugs for Cancer.” *Nature Communications* 6 (September): 8481.
Van Long, F. Nguyen, N. Pion, A. Lardy-Cleaud, E. Lavergne, J. C. Bourdon, S. Chabaud, I. Treilleux, F. Catez, J. J. Diaz, and V. Marcel. 2016. “Ribosome Biogenesis Factors: Novel Clinical Markers of Breast Cancer Outcome.” *European Journal of Cancer* 61: S201.
Wang, Dong, Chengbo Li, Yuan Zhang, Min Wang, Nan Jiang, Lin Xiang, Ting Li, et al. 2016. “Combined Inhibition of PI3K and PARP Is Effective in the Treatment of Ovarian Cancer Cells with Wild-Type PIK3CA Genes.” *Gynecologic Oncology* 142 (3): 548–56.
Wang, Haiwei, Yuxing Zhang, and Yanzhi Du. 2013. “Ovarian and Breast Cancer Spheres Are

Similar in Transcriptomic Features and Sensitive to Fenretinide." *BioMed Research International* 2013 (October): 510905.
Wilhelm, Mathias, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M. Savitski, Emanuel Ziegler, et al. 2014. "Mass-Spectrometry-Based Draft of the Human Proteome." *Nature* 509 (7502): 582–87.
Xiang, Tingxiu, Lili Li, Yichao Fan, Yanyan Jiang, Ying Ying, Thomas Choudary Putti, Qian Tao, and Guosheng Ren. 2010. "PLCD1 Is a Functional Tumor Suppressor Inducing G2/M Arrest and Frequently Methylated in Breast Cancer." *Cancer Biology & Therapy* 10 (5): 520–27.
Yang, Mi, Jaak Simm, Chi Chung Lam, Pooya Zakeri, Gerard J. P. van Westen, Yves Moreau, and Julio Saez-Rodriguez. 2018a. "Linking Drug Target and Pathway Activation for Effective Therapy Using Multi-Task Learning." *bioRxiv*. https://doi.org/10.1101/225573.
———. 2018b. "Linking Drug Target and Pathway Activation for Effective Therapy Using Multi-Task Learning." *Scientific Reports* 8 (1): 8322.
———. 2018c. "Linking Drug Target and Pathway Activation for Effective Therapy Using Multi-Task Learning." *Scientific Reports* 8 (1): 8322.
Yi, Yong Weon, Wooyoung Hong, Hyo Jin Kang, Hee Jeong Kim, Wenjing Zhao, Antai Wang, Yeon-Sun Seong, and Insoo Bae. 2013. "Inhibition of the PI3K/AKT Pathway Potentiates Cytotoxicity of EGFR Kinase Inhibitors in Triple-Negative Breast Cancer Cells." *Journal of Cellular and Molecular Medicine* 17 (5): 648–56.
Zhang, Dingxiao, Pingping Jiang, Qinqin Xu, and Xiaoting Zhang. 2011. "Arginine and Glutamate-Rich 1 (ARGLU1) Interacts with Mediator Subunit 1 (MED1) and Is Required for Estrogen Receptor-Mediated Gene Transcription and Breast Cancer Cell Growth." *The Journal of Biological Chemistry* 286 (20): 17746–54.

APPENDIX

A Supplementary information to chapter 1

A.1 Supplementary text 1: Methodology applied to breast tissue

We explained through target-pathway interactions, two mechanisms of drug synergy. In order to validate our synergy models, we first looked at public data, using the DREAM AstraZeneca drug combination challenge(Menden et al. 2017a), which experimentally tested >120 folds drug combinations compared to the previous Bansal et al. challenge. Furthermore, the AstraZeneca challenge expanded the number of tested cell lines including their deep molecular characterisation enabling for the first time identification of synergy biomarkers. We tested our model on 7 target pairs (29 drug combinations) from the AstraZeneca DREAM challenge(Menden et al. 2017a), and chose breast as the most represented tissue with 33 cell lines.

We applied our general framework to predict synergy scores. The first step was to determine the top sensitive and top resistant pathways for a certain target - pathway pair (**Supplementary Figure 3**). We then derived the formula of Delta Pathway Activity and predicted the drug synergy (**Table 1**). When choosing between Model 1 and 2 for the synergy model, the target functional similarity was the main criteria. If the similarity is close to 1, we use Model 1. If the similarity is close to -1, we use Model 2.

PI3K/AKT/MTOR pathway plays a significant role in treatment resistance in breast cancer(Paplomata and O'Regan 2014). Therefore, we hypothesized that the PI3K pathway will be informative of the synergy if AKT is targeted. Therefore, each time AKT is targeted, we included PI3K pathway as well as any pathway between the first one and PI3K, while respecting the limit of maximum 3 pathways per group.

When grouping pathways in the top sensitive and top resistant groups, we consider only those that have at least one significant interaction with the drug targets. If not significant, we discard the pathway. Exceptions are made when only one pathway is included (the top sensitive or top resistant one) and when the pathway has a stronger interaction than a pathway included by prior knowledge (literature).

For AKT/ALK (**Supplementary Figure 3a, Figure 3a**): the top sensitive pathway is EGFR and the top resistant pathways are MAPK and TNFa. The target functional similarity between AKT1/2 and ALK is -0.4 (**Table 1**). Therefore, we used synergy Model 2:

$$Delta\ PA\ (AKT/ALK)_{breast} = \frac{MAPK + TNFa}{2} - \frac{EGFR + VEGF + PI3K}{3}$$

For AKT/MTOR (**Supplementary Figure 3b, Figure 3b**): the top sensitive pathways are EGFR and VEGF. The top resistant pathways are MAPK and TNFa. The target functional similarity between AKT1/2 and MTOR is 0.8 (**Table 1**). Therefore, we used synergy Model 1:

$$Delta\ PA\ (AKT/MTOR)_{breast} = \frac{EGFR + VEGF + PI3K}{3} - \frac{MAPK + TNFa}{2}$$

For AKT/PARP1 (**Supplementary Figure 3c, Figure 3c**): the top sensitive pathway is EGFR and the top resistant pathways are MAPK and TNFa. The correlation between AKT1/2 and PARP1 is -0.8 (**Table 1**). In this case, we used Model 2:

$$Delta\ PA\ (AKT/PARP1)_{breast} = \frac{MAPK + TNFa}{2} - \frac{EGFR + VEGF + PI3K}{3}$$

For AKT/EGFR (**Supplementary Figure 3d, Figure 3d**): the top sensitive pathway is EGFR and the top resistant are MAPK. The target functional similarity between AKT1/2 and EGFR is 0.9 (**Table 1**). Therefore, we used synergy Model 1. Since protein EGFR is targeted, we also added CNV information:

$$Delta\ PA\ (AKT/EGFR)_{breast} = \frac{EGFR + NFkB + PI3K}{3} - MAPK + CNV_{EGFR}$$

For BCL2/MTOR (**Supplementary Figure 3e, Figure 3e**): the top sensitive pathways are VEGF, NFkB and Trail and the top resistant pathways are MAPK and TNFa. The target functional similarity between BCL2 and MTOR is 0.7 (**Table 1**). Therefore, we used synergy Model 1:

$$Delta\ PA\ (BCL2/MTOR)_{breast} = \frac{VEGF + NFkB + Trail}{3} + \frac{MAPK + TNFa}{2}$$

For EGFR/MTOR (**Supplementary Figure 3f, Figure 3f**): the top sensitive pathways are EGFR and NFkB. The top resistant are MAPK and TNFa. The target functional similarity between EGFR and MTOR is 0.6 (**Table 1**). Therefore, we used synergy Model 1. Since protein EGFR is targeted, we also added CNV information:

$$Delta\ PA\ (EGFR/MTOR)_{breast} = \frac{EGFR + NFkB}{2} - \frac{MAPK + TNFa}{2} + CNV_{EGFR}$$

For AKT/BCL2 (**Supplementary Figure 3g, Figure 3g**): the top sensitive pathway is EGFR and the top resistant pathway is MAPK. The correlation between AKT1/2 and BCL2 is 0.5 (**Table 1**). In this case, we used Model 1:

$$Delta\ PA\ (AKT/BCL2)_{breast} = \frac{EGFR + VEGF + PI3K}{3} - MAPK$$

A.2 Supplementary tables

	Setting 1 predicting new cell lines	**Setting 2** predicting new drugs	**Setting 3** predicting existing drugs on existing cell lines	**Setting 4** predicting new drugs on new cell lines
use case	- Personalized medicine	- Drug repositioning	- prioritization for new experiments - Interaction matrix generation	- Personalized medicine with previously untested drugs - Quality control of the interaction matrix
drug features	optional	**required**	optional	**required**
cell line features	**required**	optional	optional	**required**
cross validation	10 fold CV	10 fold CV	NA	2 x 10 fold CV
prediction metrics	For each drug, correlation of observed versus predicted IC50 across all cell lines.	For each cell line, correlation of observed versus predicted IC50 across all drugs.	correlation of observed versus predicted IC50 for all drug-cell line pairs.	correlation of observed versus predicted IC50 for all drug-cell line pairs.

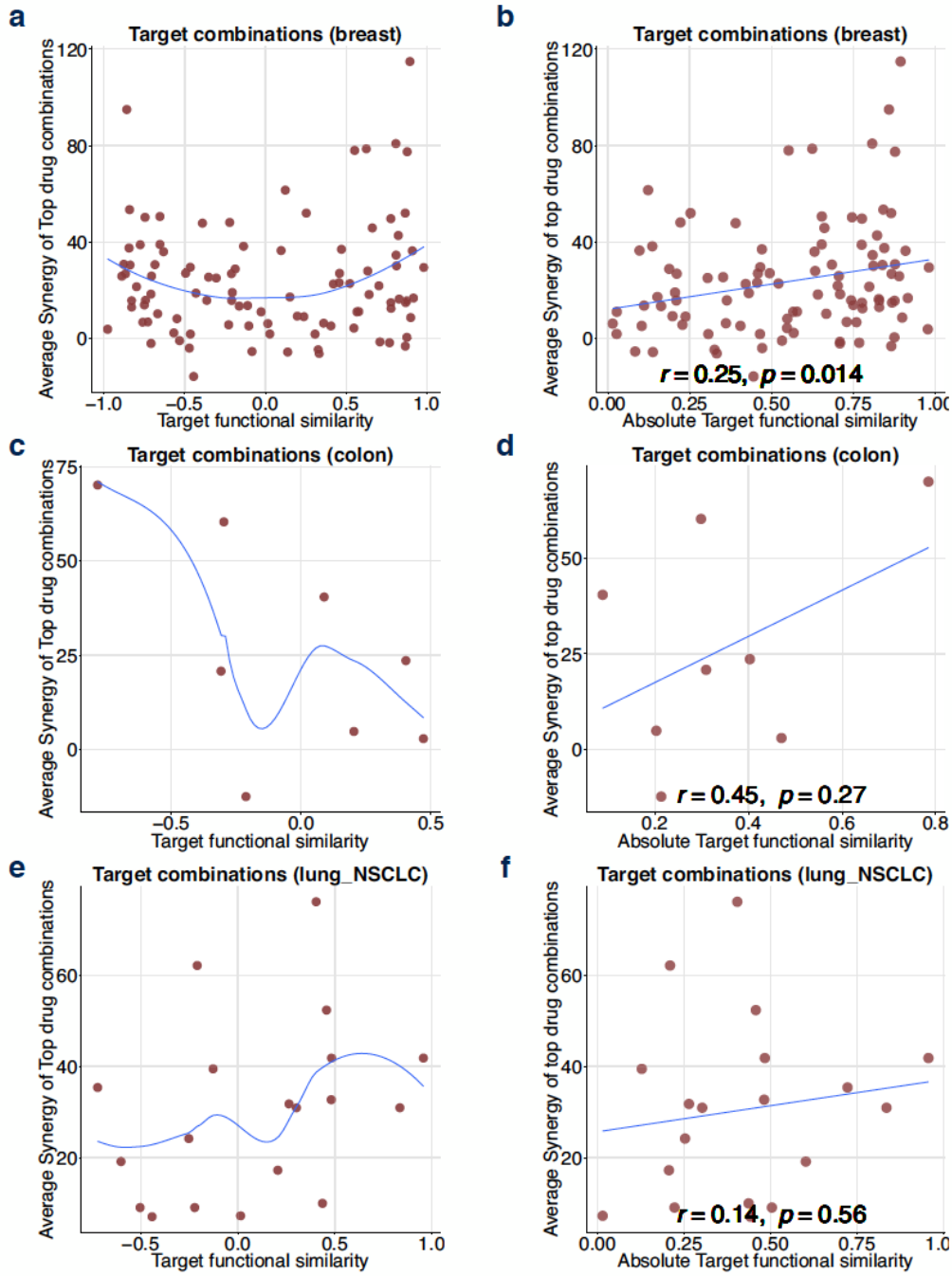
Supplementary Table 1: Different settings for drug response prediction

	Supervised learning	Hypothesis driven synergy stratification
Data source	Drug combination drug response on cancer cell lines	Single agent drug response on cancer cell lines
Input features	Gene expression and drug target	Gene expression and drug target
Additional information	mutation, CNV, cancer subtypes	mutation, CNV, cancer subtypes
Synergy prediction algorithm	Supervised learning algorithms such as tree based algorithms (Random Forest, XGBOOST) and matrix factorization.	Linear combination of gene expression derived pathway scores (from PROGENy)
Prediction settings (Supplementary Figure 5, Supplementary Table 1)	**Setting 1**: prediction of new cell lines for existing drugs **Setting 2**: prediction of new drugs for existing cell lines **Setting 3**: prediction of existing drugs for existing cell lines **Setting 4**: prediction of new drugs on new cell lines	**Setting 1**: prediction of new cell lines for existing drugs **Setting 4**: prediction of new drugs on new cell lines
Strength	(1) General purpose usage in drug wise and cell line wise settings (2) Does not require domain expertise (3) Easy to implement	(1) Does not require many drug combination experiments as prior knowledge (2) Linear combination of pathway activation is suited for biological interpretation
Weakness	Requires an extensive set of drug combination drug response data	(1) Relies heavily on domain knowledge and literature evidence, making automated processing challenging (2) Can only be used in a drug wise setting

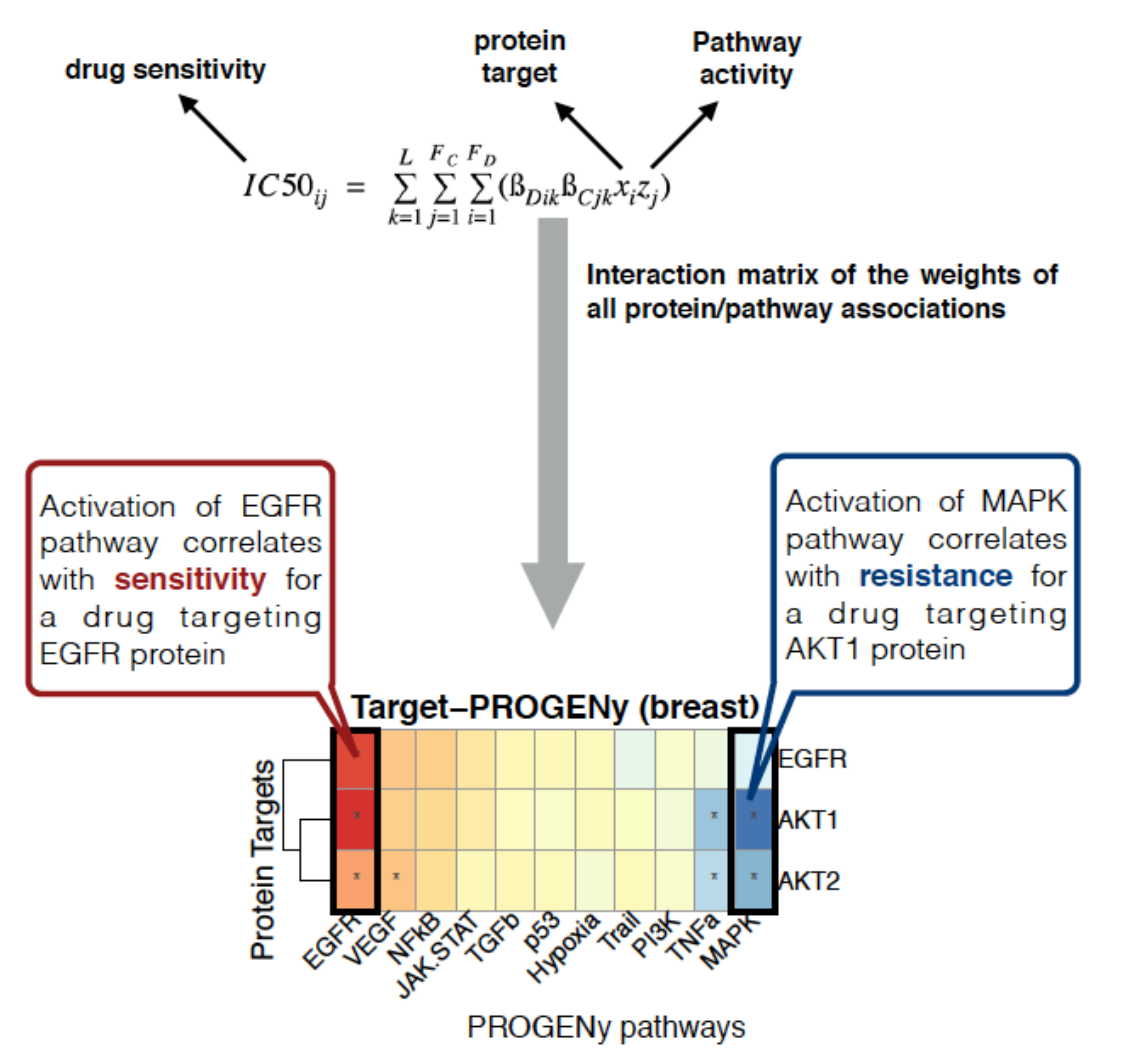
Supplementary Table 2: Comparison of synergy stratification workflow with supervised learning.

Supplementary Table 3: drug target information downloaded from https://www.cancerrxgene.org on March 2017.

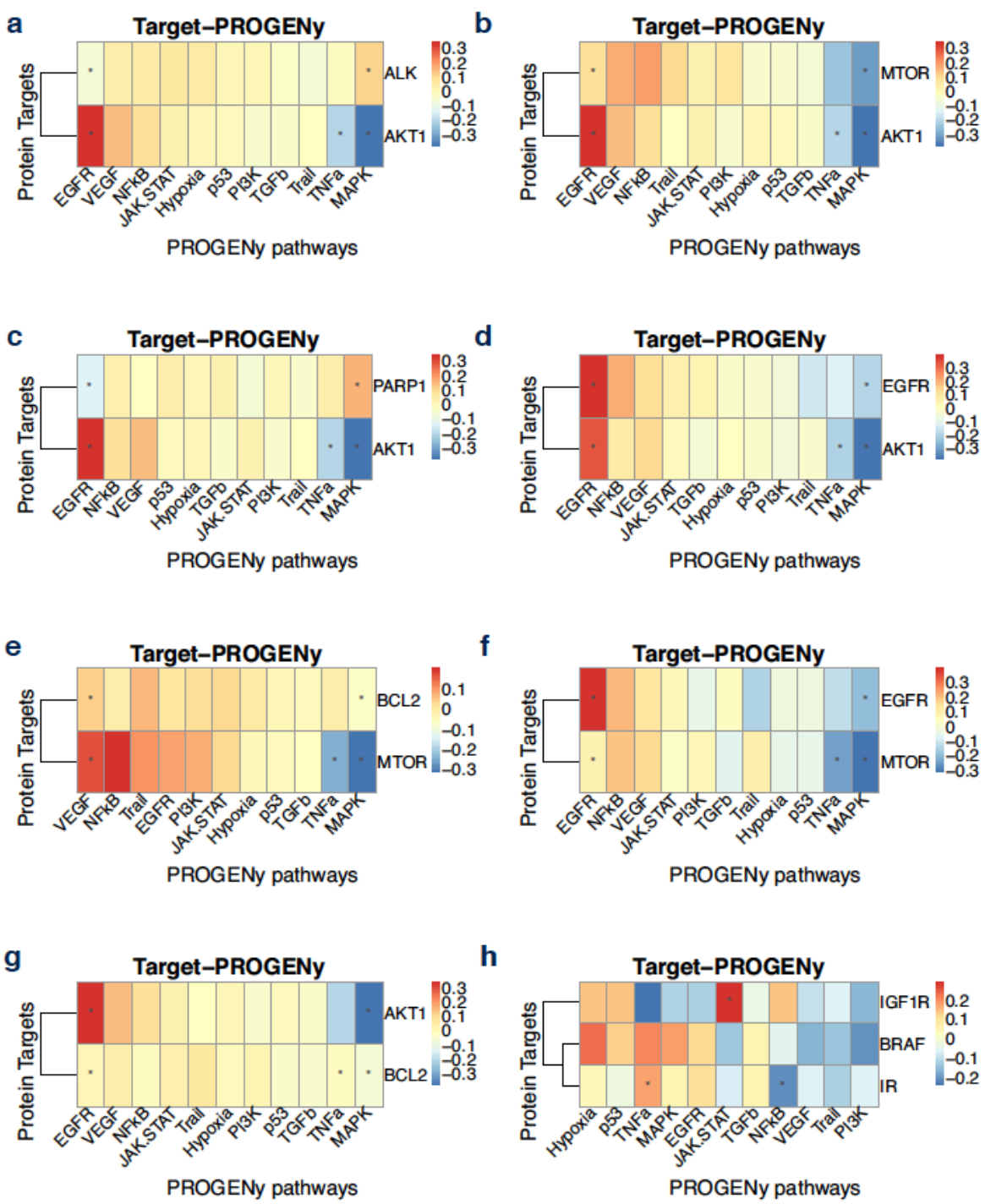
A.3 Supplementary figures



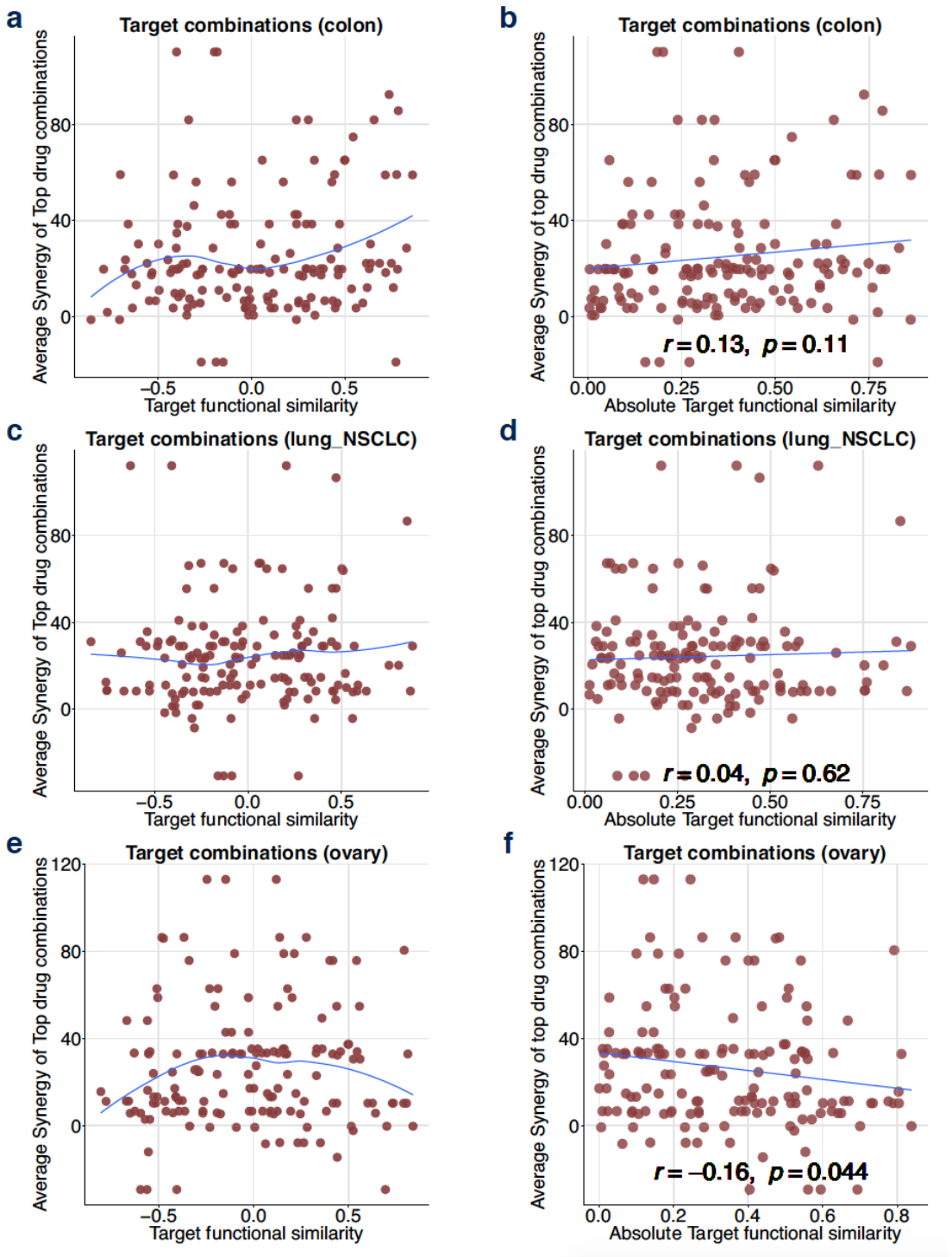
Supplementary Figure 1: Influence of target functional similarity in drug synergy for AstraZeneca DREAM dataset. The target functional similarity is the correlation between two protein targets by their interactions with the PROGENy pathways. For each tissue, we plot the synergy against the target functional similarity and its absolute value. **(a)** and **(b)** for breast tissue. **(c)** and **(d)** for colon tissue. **(e)** and **(f)** for NSCLC lung tissue.



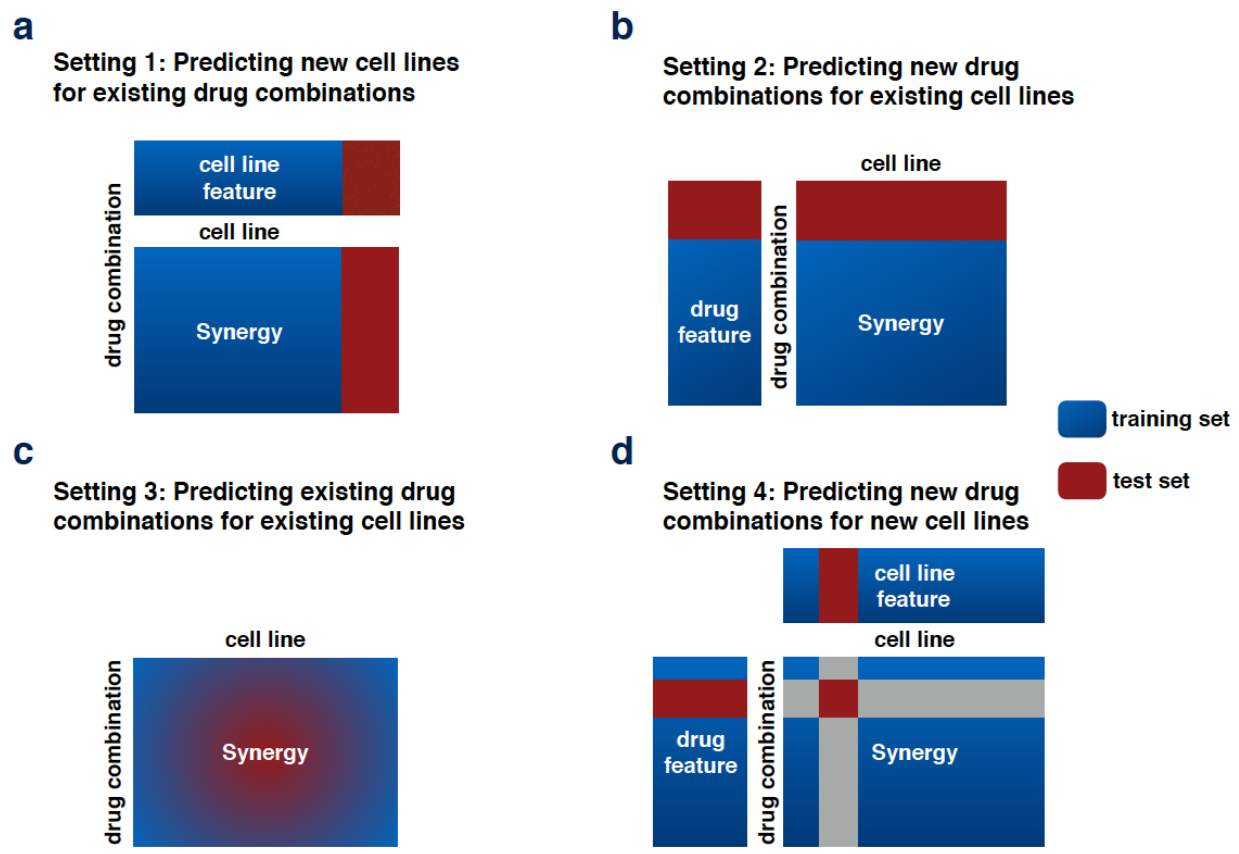
Supplementary Figure 2: Interpretation of the interaction matrix. Enhanced sensitivity occurs when targeting several proteins involved in drug response under the activation of the right pathway. The same rule applies to resistance.



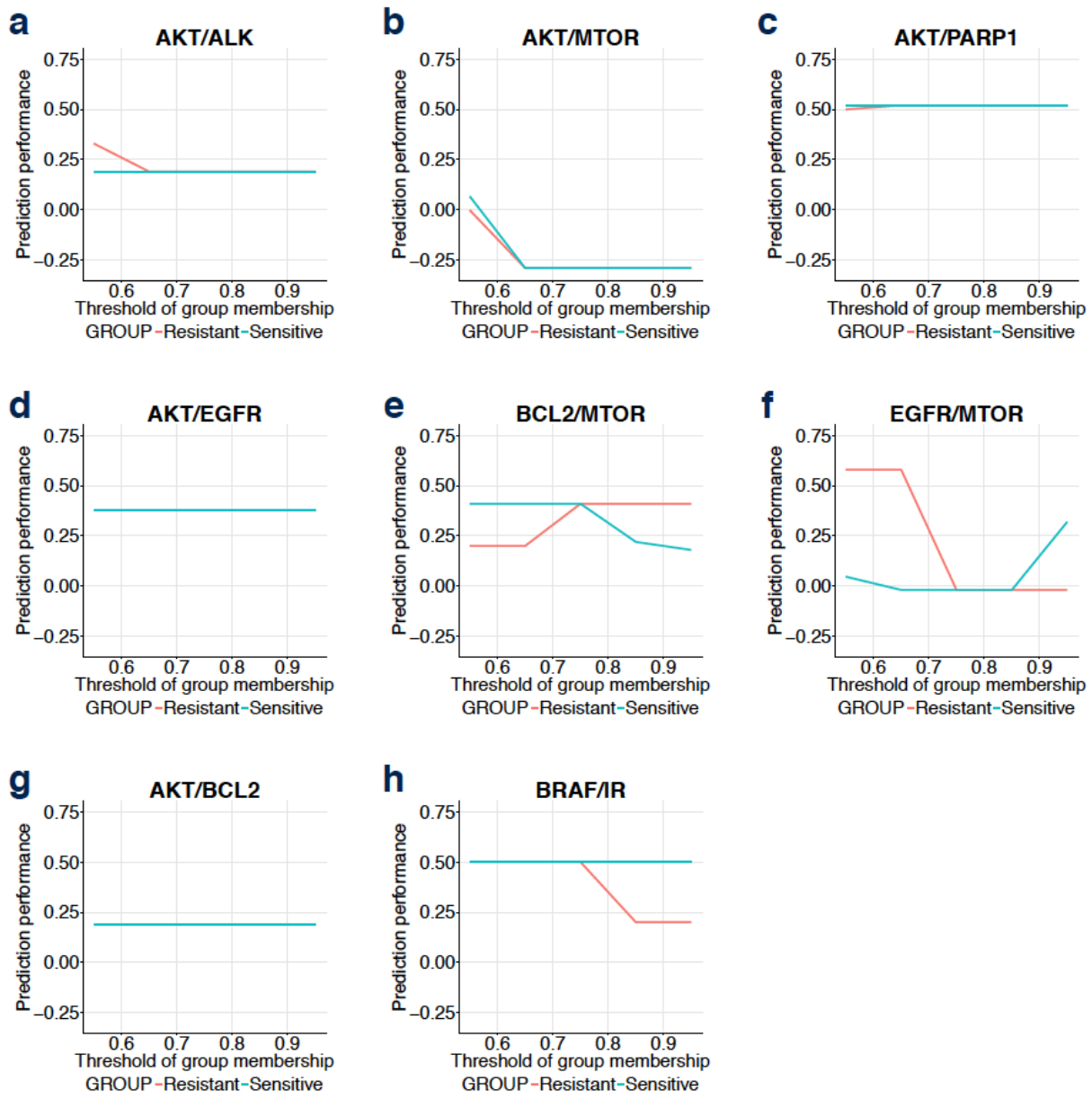
Supplementary Figure 3: Functional profile of protein targets in breast and colorectal tissues. (a), **(b)**, **(c)**, **(d)**, **(e), (f) and (g)** describe the functional profile of AKT/ALK, AKT/MTOR, AKT/PARP1, AKT/EGFR, BCL2/MTOR, EGFR/MTOR and AKT/BCL2 pairs in breast tissue. **(h)** describes BRAF/IR's functional profile in colorectal tissue. The functional profile is a subset of the target pathway interaction in the Macau model. Pathways are ordered from the most sensitizing to the least. Significance of the interaction values is corrected according to Benjamini & Yekutieli procedure (20% FDR) as described in Yang *et al*. (Yang et al. 2018c).



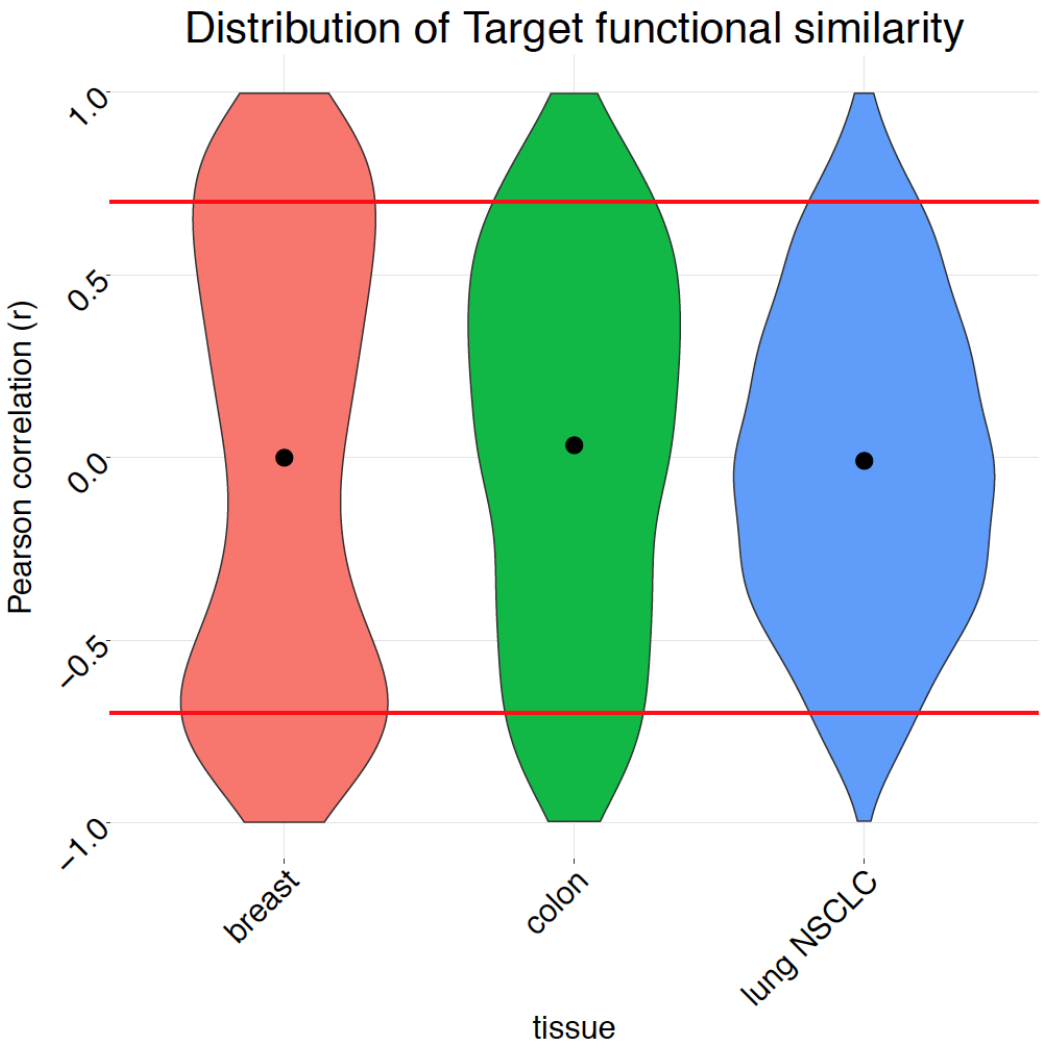
Supplementary Figure 4: Influence of target functional similarity in drug synergy for O'Neil et al Merck dataset. The target functional similarity is the correlation between two protein targets by their interactions with the PROGENy pathways. For each tissue, we plot the synergy against the target functional similarity and its absolute value. Only tissues with at least 5 cell lines were chosen. **(a)** and **(b)** for colon tissue. **(c)** and **(d)** for NSCLC lung tissue. **(e)** and **(f)** for ovary tissue.



Supplementary Figure 5: Different settings in drug synergy prediction. (**a)** Predicting new cell lines for existing drugs. For each drug pair, we compute the pearson correlation of observed versus predicted synergy across all cell lines of the test set. **(b)** Predicting new drug synergy for existing cell lines. For each cell line we compute the pearson correlation of observed versus predicted synergy across all drug pairs of the test set. **(c)** Predicting existing drug synergy for existing cell lines. This is a missing value imputation setting where side information of drug and cell lines are not required, but can be used to improve the result. The test data is defined by a percentage of the whole data set. We compute the pearson correlation of observed versus predicted synergy for all randomly chosen drugs - cell line triplets of the test set. **(d)** Predicting new drug synergy for new cell lines. We do 2 simultaneous cross validation on both drug and cell line sides. The test data is defined by association of the test set of the drug side with the test set of the cell lines side. We compute the pearson correlation of observed versus predicted synergy for all drug - cell line pairs of the test set.



Supplementary Figure 6: Sensitivity analysis for group membership parameters. In the determination of Delta Pathway Equation, we explored the prediction performance for each target pair in AstraZeneca breast data and colorectal validation data, based on the following parameters: threshold for group membership of the top sensitive pathways and top resistant pathways.



Supplementary Figure 7: Distribution of similarity values across tissues. For each tissue, we plotted the target functional similarities of the profiled protein targets and set the cut off of high similarity and high dissimilarity at 0.7 and -0.7.

B Supplementary information to chapter 2

B.1 Supplementary analysis 1: MOFA, Robustness assessment

As we use MOFA factor as prognostic biomarker, one essential condition is to be able to recover the relevant factor in a new dataset. We ran MOFA in a multi omics setting (predicted protein + mRNA + mutation) on the breast dataset with 2 fold cross validation, with 1000 MOFA iterations, repeated 50 times. At each run, we identified the most predictive Factor on the training set, using log rank test on binarized Factor after correcting for age, and gender, then false discovery rate. We then retrieved the corresponding Factor on the test set using correlation of the weights, as a factor's weights is really what defines it. The average absolute correlation between the training Factor weight and the testing's identified Factor weights is 0.76 (sd=0.14). In addition to that, the numbering of factor is conserved 33 times out of 50. We can conclude that it is possible to retrieve a biologically relevant Factor from external dataset.

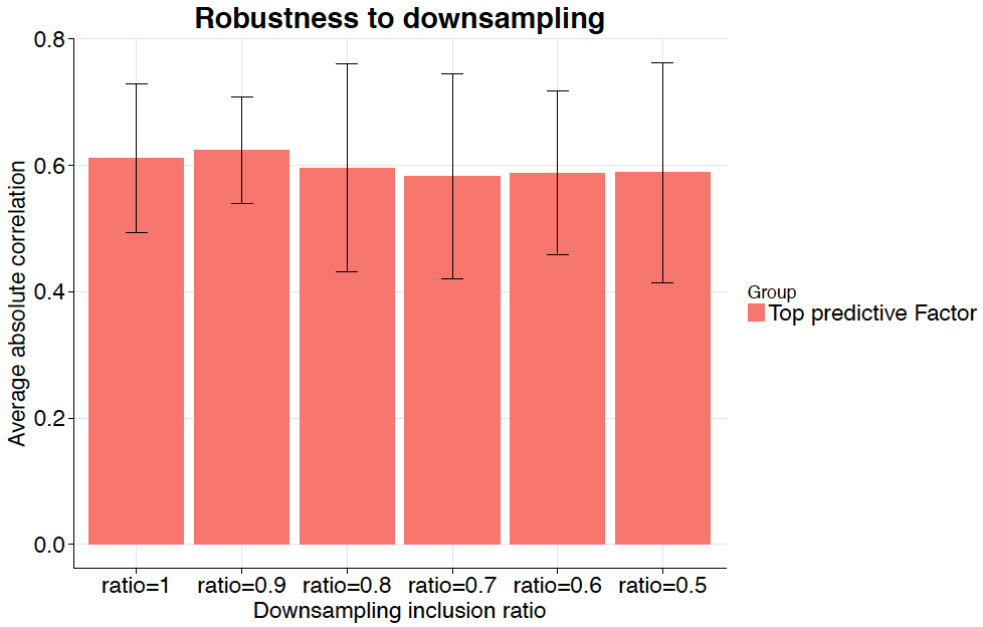
We tested the robustness to downsampling of the top predictive factor for different inclusion ratios (1, 0.9, 0.8, 0.7, 0.6, 0.5), in a 2 fold cross validation setting, with 100 MOFA iterations, repeated 30 times. In overall, the correlation between the top predictive factor of the training set and the corresponding factor in test set is conserved (**Supplementary Figure 1**).

B.2 Supplementary tables

Omics layer	Top Factors	q-value	Top genes contributing to each Factor	Top enriched pathways using PCGSEA with Reactome
Predicted Protein^{10006}	Factor 1	0.0014	MT1X, FABP7, SFT2D2, **TMEM26**, **VGLL1**, MT1F, MAEL	Processing of Capped Intron−Containing Pre−mRNA Antimicrobial peptides
	Factor 6	0.00071	**RPL26**, **RPL29**, **RPL34**, **RPL37**, DPYSL5, BRI3BP, ABCA2, GABARAP	Peptide chain elongation
mRNA^{15107}	Factor 2	0.0025	FOXA1, LBR, CDCA7, GATA3, PRKX, B3GNT5, MSN	Cell Cycle, Mitotic Prometaphase, RHO GTPases Activate Formins
Predicted Protein^{10006} + mRNA^{15107}	Factor 2	0.0050	**Protein view**: **TMEM26**, TMEM259, SFT2D2 **mRNA view**: TTK, SUV39H2, SRPK1, WDR43, CDCA8	**Protein view**: Mitotic Prometaphase **mRNA view**: Cell Cycle, Mitotic Prometaphase
	Factor 4	0.044	**Protein view**: **SULT1E1**, **ARGLU1**, **AKR1B10**, **ING4**, **CEACAM5** **mRNA view**: NRF1, RALB, DFFB	**Protein view**: Extracellular matrix organization **mRNA view**: Degradation of DVL, Hh mutants abrogate ligand secretion
	Factor 6	0.0071	**Protein view**: **RPL34, RPL37**, GABARAP **mRNA view**: HCFC1	**Protein view**: Peptide chain elongation **mRNA view**: Peptide chain elongation

Supplementary Table 1: Top predictive Factors and their functional characterization. We applied MOFA algorithm on protein, mRNA and a combination of both. The algorithm is run 20 times and the best model is chosen based on highest Evidence Lower Bound (ELBO). For each predictive factor of survival, we showed relevant pathways ranked in top 3. The top contributing genes of each factor are those ranked in top 10 and with at least one Pubmed association with the keywords “breast cancer”. The genes in bold are those described in the result section. For the combination of protein and mRNA, we described for each view (protein and mRNA), the enriched pathways and top weighted genes.

B.3 Supplementary figures



Supplementary Figure 1: MOFA robustness to downsampling. For each ratio of the total number of samples, we identify the top predictive Factor in the training set and retrieve the corresponding factors in the test set.