Mining therapeutic insights from large scale drug screenings with transfer learning

Mi Yang[1], Jaak Simm[3], Pooya Zakeri[3], Yves Moreau[3], Julio Saez-Rodriguez[1,2]

[1]RWTH Aachen University, Faculty of Medicine, Joint Research Center for Computational Biomedicine, Aachen, Germany.

[2]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, UK.

[3]ESAT-STADIUS, KU Leuven B-3001 Heverlee, Belgium.

ABSTRACT

Despite the abundance of large-scale molecular and drug-response data, our ability to extract the underlying mechanisms of diseases and treatment efficacy has been in general limited. Machine learning algorithms applied to those data sets most often are used to provide predictions without interpretation, or reveal single drug-gene association and fail to derive robust insights. We propose to use Macau, a bayesian multitask multi-relational algorithm to generalize from individual drugs and genes and explore the association between the drug targets and signaling pathways' activation. A typical insight would be: "Activation of pathway Y will confer sensitivity to any drug targeting protein X". We applied our methodology to the Genomics of Drug Sensitivity in Cancer (GDSC) screening, using signaling pathways' activities as cell line input and nominal targets as drug input. The interactions between the drug target and the pathway activity can guide a tissue specific treatment strategy by for example suggesting how to modulate a certain protein to maximize the drug response for a given tissue. We confirmed in literature drug combination strategies derived from our result for brain, skin and stomach tissues. Such an analysis of interactions across tissues might help drug repurposing and patient stratification strategies.

Availability and Implementation: The source code of the method is available at https://github.com/saezlab/Macau_project_1

1. INTRODUCTION 3
2. RESULTS 4
2.1 Multitask learning with Macau 4
2.2 Assessment of drug response prediction with Macau 4
2.3 Feature interactions: Tissue specific analysis 6
2.4 Therapeutic applications of the interactions 8
3. DISCUSSION 9
4. METHODS 11
4.1 Macau: Algorithm 11
4.2 Macau: Feature interaction 11
4.3 Macau: Parameter setting 12
4.4 PROGENy 13
BIBLIOGRAPHY 14
FIGURES 17
TABLES 28

1. INTRODUCTION

Translating preclinical models into actionable insight is essential for more personalized treatments. Despite the wealth of omics data since a few decades, our ability to decipher the underlying mechanisms of diseases has been much less effective (Alyass, Turcotte, and Meyre 2015). This is particularly apparent in cancer. Large scale drug screenings involving dozens to hundreds of drugs applied to cell lines have been the main driver of in silico drug discovery. Public drug screening projects such as the Genomics of Drug Sensitivity in Cancer (GDSC) (Iorio et al. 2016a), the Cancer Therapeutics Response Portal (CTRP) (Seashore-Ludlow et al. 2015) and the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al. 2012) generated drug response data for hundreds of drugs and around one thousand cell lines. The main objective of these datasets is to shed light on the molecular mechanisms regulating drug response.

Machine learning is widely used to predict drug response on the treated cell lines. Most of the analyses consist of building a model for one drug at a time, which is of limited power given the relative low number of samples. If we can bring together all drugs in a single model, we can learn common patterns reflecting the underlying mechanisms. Towards this end, transfer-learning type algorithms which use information gained in one task for another task are a promising approach.

Multitask frameworks have been recently used to demonstrate the preclinical feasibility of drug sensitivity prediction from large scale drug screening experiments (Yuan et al. 2016; Menden et al. 2013); (Cortés-Ciriano, Mervin, and Bender 2016). Methods ranging from standard random forest (Menden et al. 2013) to Kernelized bayesian matrix factorization (Ammad-ud-din et al. 2014) and trace norm multitask learning (Yuan et al. 2016) have been used to predict drug response by integrating genomic features for cell lines, as well as target and chemical information for drugs. While many algorithms perform better than standard methods, interpretability is often challenging, especially for kernel based learning algorithms (Ammad-ud-din et al. 2014); (Wang, Fang, and Chen 2016); (Gönen and Margolin 2014), although they can be used to identify biologically relevant genes (Nikolova et al. 2017) and derive meaningful predictive features (Ammad-ud-din et al. 2017).

The motivation of our work was to leverage the power of transfer learning to provide novel insights into the molecular underpinnings of drug response. Towards this end, we applied a multitask learning strategy for drug response prediction and feature interaction, using the tool Macau (Simm et al. 2015). 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 focused on gene expression as molecular input data, using it to estimate activities of signaling pathways, along genetic aberrations. For the drugs, we chose their nominal target as the key feature. We applied our methodology to the Genomics of Drug Sensitivity in Cancer (Iorio et al. 2016b) (GDSC) cell line panel with drug response (IC50) of 265 drugs on 990 cell lines. The interactions between protein targets and signalling pathways' activities supports a personalized treatment strategy, as it, for example, can determine how to

modulate a certain pathway to maximize the drug response. To portray tissue specificity in cancer treatments, we explored the differences of interactions across tissues with different compounds. Analyzing those interactions across many tissues can enable patient stratification, drug repositioning, and drug combination selection.

2. RESULTS

2.1 Multitask learning with Macau

Macau is a Scalable Bayesian multi-task learning algorithm which can incorporate millions of features and hundred millions of observations (Simm et al. 2015). In traditional machine learning analysis, we predict the response variable based on descriptive features of the samples. For instance, in drug screening experiments where cell lines are treated by drugs, the effect of a certain drug X is predicted from the mRNA expression of a gene Y via regression. With Macau, we unveil the interaction matrix of the drugs' feature (for example, protein target) with the cell lines' feature (e.g. transcriptomics, pathway activity). A typical insight would be: "Upregulation of gene Y will confer sensitivity to any drug targeting protein X" (**see Fig. 1A and Methods**). We refer to this from now on as feature interaction analysis. Such analysis gives hints about the drug's mode of action, by uncovering how acting on one protein affect the drug response and in which conditions (gene/pathway status).

In our feature interaction analysis, we use manually curated protein targets for the drug side. For the cell line side, we transform the transcriptomics data into pathway activity using PROGENy (Schubert et al. 2016). PROGENy is a data driven pathway method aiming at summarizing high dimensional transcriptomics data into a small set of pathway activities (**see Methods**). The 11 PROGENy pathways currently available are EGFR, NFkB, TGFb, MAPK, p53, TNFa, PI3K, VEGF, Hypoxia, Trail and JAK STAT. We obtain the interaction matrix with features of the drugs on the rows and features of cell lines on the column (**see Methods**).

2.2 Assessment of drug response prediction with Macau

When building models to predict drug response taking into account multiple drugs and cell lines, one can define four differents settings which mirror different use cases (**Supp Fig. S1**). We will describe for each setting the interpretation and prediction performance of Macau compared to standard linear regression.

Setting 1: Prediction of new cell lines for existing drugs (**Fig. S1A, Table 1**)

The meaning of this framework is to start with a subset of drugs and assign them to the right patient, e.g. new patients based on their genomic information.

We tested three different input data sets for the cell lines: (i) complete gene expression, (ii) PROGENy scores, and (iii) a combination of Single Nucleotide Polymorphism (SNP) and Copy Number Variation (CNV) (**Supp Fig. S2A**). Gene expression performed best (r=0.40), not surprisingly as it uses all 17419 genes. PROGENy also has good performance (r=0.30), specially considering its low dimension (only 11 pathways). SNP/CNV performs the worst (r=0.21) despite a dimension of 735. These results supports the use of gene expression derivative methods as predictive input features, in agreement with previous studies (Iorio et al. 2016a); (Costello et al. 2014).

We then compared multitask Macau with standard single task linear regression. Using gene expression and drug target, there is no significant difference between Macau (r = 0.40) and LASSO (r = 0.41), p = 0.39. For PROGENy scores, there is no significant difference between Macau and Ridge (p = 0.92). Finally for SNP/CNV, Macau performed significantly better than Ridge (p = 0.00051).

In single task, the binary sparse features (SNP/CNV) may not be present (value=1) in both training and test set. The multi-task effect of sparse features by latent dimension is getting more information than a single task algorithm could. In addition to that, with Macau, the response does not depend on one side, but also on the other side: where even without features, there are still latent variables.

Setting 2: Prediction of new drugs on existing cell lines (**Supp Fig. S1B, Table 1**)

A second important scenario is to predict the effect of a new drug on a set of patients based on the side information of the drug. If the new drug is predicted to be better than the existing ones, then a therapeutical switch can be considered. The concept of “new drug” is relative to the patient, it can concern existing drugs which have never been used for a patient group.

As a benchmark, we compared Macau with standard Ridge regression (**Supp Fig. S2B**). To be able to predict the effect of new drugs, we considered as additional side features ECFP4 chemical fingerprints (Rogers and Hahn 2010). The average correlation with Macau for the cell lines is 0.42 with drug target and 0.28 with ECFP4, in both cases significantly better than Ridge regression (r=0.12; $p < 2.2e^{-16}$, and r= 0.05; $p < 2.2e^{-16}$, respectively). The performance gap can as in the previous setting be explained by the effect of sparse features, where multitask has the advantage.

Setting 3: Prediction of existing drugs and existing cell lines (**Supp Fig. S1C, Table 1**)

In this setting we solve an imputation problem, where the test set is randomly chosen from the drug response matrix. We can use side information from both sides to improve the result. We tested setting 3 on GDSC (**Table S1**) datasets. In overall, we were able to get an excellent prediction: mean r=0.932 with 90% of the data as training set and 10% and even r=0.834 with 99% as test set.

Setting 4: Prediction of new drugs on new cell lines (**Supp Fig. S1D, Table 1**)

This setting aims at predicting a new drug's effect on a new cell line solely based on drug target information and whole transcriptomics, hence a very challenging task. We use 2 simultaneous 10-fold cross validation of drugs and cell lines, obtaining a correlation of r=0.45, which is only marginally lower when replacing transcriptomics with PROGENy scores (r=0.42).

We use this setting as quality control for the feature interaction analysis. Predicting new drugs' responses on new cell lines allows us to evaluate the quality of the interactions between features of the drugs and features of the cell lines.

For setting 4, we also perform a benchmark in a tissue-specific setup. We compared for 16 tissues, the prediction using GEX, PROGENy and SNP/CNV. For most of the tissues, the pearson correlation of observed versus predicted IC50 is close to 0.4. Gene expression does not perform significantly better than PROGENy in most tissues, except for colon (p-value = 0.02), liver (p-value = 0.01), soft tissue (p-value = 0.03) and stomach (p-value = 0.002). Compared to SNP and CNV, gene expression performs significantly better for 14 out of the 16 tissues.

In summary, our multitask learning achieves a similar or better predictability performance than standard methods across all possible settings. Confirmed this, we moved on to the focus of our work, use the models to gain insight on the interactions between pathway activities and drug targets.

2.3 Feature interactions: Tissue specific analysis

Using features on both sides of the drug response matrix, we can measure the association between features of drugs and features of cell lines, by taking into account all drugs and all cell lines in a generalized model. From the available options (**Table S2**), and based on the results of the previous sections, we chose to use PROGENy pathway activity for cell lines due to performance and interpretability reasons, and protein targets for drug side.

We performed a feature interaction analysis drug target - PROGENy pathways for all 16 tissues (**Supp Fig. S3**). In order to assess the quality of the interaction matrix between drug targets and cell lines features, we used setting 4 (**Supp Fig. S1D**) for each tissue type (**Table S3**), since setting four describes the generalization to new cases, which is what we want to obtain with the interaction analysis. We will detail in the following for 4 tissues, the evidences from literature of the top hits (**Fig. 2**).

Bone (Fig. 2A): From the heatmap result, we observe that MEK1/MEK2 inhibition confers sensitivity when MAPK pathway is activated. Indeed, MEK inhibition induces apoptosis in osteosarcoma cells with constitutive ERK1/2 phosphorylation (Baranski et al. 2015).

Brain (Fig. 2B): EGFR activates mTORC2-NF-κB pathway which renders glioblastoma cells and tumors resistant to chemotherapy in a manner independent of Akt (Tanaka et al. 2011). As expected, EGFR pathway activation confers sensitivity if associated with MTORC2 inhibitors in our results (**Supp Fig. S4.3)**. But at the same time, targeting PLK1 confers resistance. We can assume that blocking EGFR pathway while targeting PLK1 could lead to synergistic effect. PLK1 and EGFR inhibitor were described as orthogonal therapeutic agents in glioblastoma, with enhanced tumoricidal activity when combined (Tanaka et al. 2011; Shen et al. 2015).

Skin (Fig. 2C): Activation of TNFa pathway confers sensitivity when associated with anti TOP1. TNF-alpha increases human melanoma cell invasion and migration in vitro (Katerinaki et al. 2003) and Topoisomerase I amplification in melanoma is associated with more advanced tumours and poor prognosis (Ryan et al. 2010). Repression of TOP1 activity inhibited IFN-β- and TNFα-induced gene expression and protects against lethal inflammation in vivo (Rialdi et al. 2016).

We observe that MAPK activation confers sensitivity when targeting BRAF (**Supp Fig. S4.14**). Indeed, BRAF activates MAPK pathway and a key target in this signaling cascade. Therapies targeting $BRAF^{V600E}$ have significant potential to halt the progression of malignant tumors (Inamdar, Madhunapantula, and Robertson 2010). activation of VEGF pathway confers resistance with targeting BRAF. We can reasonably think that blocking VEGF can have a synergistic effect with targeting BRAF. Dual $BRAF^{V600E}$ and VEGF targeting has been shown to provide a combinatorial benefit against $BRAF^{V600E}$ mutants tumor growth *in vivo* (Comunanza et al. 2017).

Stomach (Fig. 2D): One striking example is increased sensitivity by targeting MET, EGFR and ERBB2 when activation of EGFR pathway (**Supp Fig. S4.16**). MET protein overexpression was associated with tumor progression and survival in gastric cancer (Inokuchi 2015). It has been shown that combination of ERBB2-inhibitor (lapatinib) and MET-inhibitor offered a more profound inhibition in the ERK/AKT pathway and cell proliferation than lapatinib alone (Ha et al. 2015).

In summary, we could find literature support for the results of the tissue specific analysis, suggesting that insights generated from feature interaction analysis could have clinical impact. We will now focus on the clinically relevant applications.

2.4 Therapeutic applications of the interactions

Deriving pathway biomarkers from target/pathway associations

All PROGENy pathways are defined by perturbation experiments. Therefore, we can activate or inhibit a pathway by the compounds used to produce the perturbations. In both cases, activating or inhibiting a pathway could improve drug sensitivity and decrease resistance. We illustrate this point with MDM2-p53 pair in ovarian cancer (**Fig. 3A, Supp Fig. S3.12**). We found that higher expression of p53 pathway leads to attenuation of resistance to anti MDM2 drugs. MDM2 binds to and inhibits p53 (Shi and Gu 2012). Coexpression of p53 and MDM2 in ovarian tumor biopsy specimens from 82 patients was also related to poor outcome (Dogan et al. 2005), which supports the rationale of targeting MDM2.

Deriving drug combination strategy

If the association between a pathway activity and the drug efficacy is causal and not just a correlation, modulating the pathway would affect the drugs' effect.

For example, in lymphoma, decreased activity of the NFkB pathway, which is constitutively deregulated in lymphoma development (Jost and Ruland 2007), confers sensitivity to antimetabolites, a common type of chemotherapy (**Fig. 3B, Supp Fig. S3.11**). Thus, blocking NFkB may restore sensitivity to antimetabolite drugs. Interestingly, in Non Hodgkin lymphoma, antimetabolites are used together with corticosteroids (protocol CVAD + Methotrexate and Cytarabine). As corticosteroids inhibit NFkB (Auphan et al. 1995), this could explain the combination.

Harnessing tissue variability of interactions

In order to explore dissimilarities across tissues is for each pathway-target pair, to analyse the tissue where it has the highest interaction weight and the tissue with the lowest weight. Then, we keep the pairs with smallest difference in absolute value between maximal and minimal weight. The objective is to find target-pathway pairs which have the greatest and most antagonistic effect for two different tissues (**Table S4**). For instance, NFkB confers high sensitivity in breast but resistance in stomach to drugs targeting ERBB2 (**Supp Fig. S5**). In most cases we could discern an antagonistic behavior from one tissue to the other, except for EGFR-DNA damage pairs.

Another way to explore similarities between tissues is to vectorize for each tissue all the pathway-target interaction values. We start with a matrix of dimension 16 tissues x 1122 pathway-target pairs. We then subset the associations by taking only into account the pathway-target pairs for which at least one tissue appears in the top 5% absolute value. Finally, we rank the remaining pairs by the variance of their associations across the 16 tissues and keep the lowest 30. In this highest interaction heatmap (**Supp Fig. S6A**), we highlight the pathway-target pairs which confer drug sensitivity for many tissues. This allows the use of the same drug in the

same condition, but on a different tissue. To find the dissimilarities between tissues, we followed the previous steps, but instead keeping the top 30 pairs with the highest variance of interactions across tissue. This divergent interaction heatmap (**Supp Fig. S6B**) displays the pathway-target pairs which have a huge variance across tissues.

We also explored how mutations (SNP) and copy number variations (CNV) interact with drug targets (**Supp Fig. S7**). It should be noted that the prediction performance (quality control) using SNP/CNV is generally lower than using PROGENy and that not all SNP/CNV are present in every tissue. For instance BCR-ABL mutation appears only for leukemia tissue (**Supp Fig. S7**), which makes this biomarker difficult to generalize to other cancer types. In this cross tissue analysis, we explored the triplet target-pathway/SNP/CNV-tissue, highlighting the similarities and dissimilarities of interactions.

3. DISCUSSION

In this paper, we provide a powerful machine learning framework for large scale drug screenings to find associations between the drugs' and the cell lines' characteristics. We focused on exploring how pathway activities modulate response to drugs targeting specific proteins.

In traditional analyses, findings are typically about the association between a drug and a gene. Such approach has the limitations that a gene alone may not capture the entire complexity of the signaling landscape, and the drug may not be very relevant and not used after the publication, therefore the insight is lost and more generalizable insights are desirable.

To overcome these issues, we introduced the feature interaction analysis in cancer specific settings. We rely on a data driven pathway method (using perturbation experiments) that has proven to be efficient at estimating pathway status (Schubert et al. 2016) from gene expression. We explored the tissue specificity of target - pathway pairs, which may ultimately improves clinical decision and therapeutical switch. We were able to confirm literature supported gold standards regarding the effect of targeting a specific protein in presence of a pathway's activation for a certain cancer type. This would not have been possible without an efficient way to reduce high dimensional omics data into a small and interpretable subset of pathways. Our results show that multitask learning can handle large scale experiments and derive interpretable insights.

There are several limitations to this study: First, the quality of the insights depends on the quality of the target pathway interaction. The performance (in setting 4) is ~0.4 for breast and colon cell lines, and up to 0.45 for skin and aerodigestive tract (**Table S3**). Although this is an encouraging result, it is still far from perfect. A significant part of mechanism are not explained by those pathways. We could address this issue by, for example, expanding the progeny pathways and including tissue specific pathways for each cancer type. Second, one limitation of

the GDSC panel for our analysis is that it adjusts the drug concentration range for each compound individually, to have a few cell lines responding, while the large bulk of cell lines does not respond, which makes the drug sensitivity in cell lines a relative concept. Therefore, we have good resolution to identify sensitive associations, but not necessarily resistance. Third, unknown off-target effects can not be taken into account. Nevertheless, we took precautions in considering only protein targets aimed by at least two drugs. Finally, our analysis we had less than 50 samples for some tissues and used only 102 protein targets for the interaction matrix. Having more cell lines and more drugs should lead to more findings.

Multitask learning framework can handle very diverse prediction settings (**Supp Fig. S1**), and can be a useful tool for the advance of precision medicine. Depending on the availability of the data and objectives, it allows us to find genomically defined patients for existing drugs and ideal drugs for existing patients, as well as giving existing drugs to existing patients and test new drugs on new patients. Although our results are based on cell lines and hence unlikely to be directly suited for predicting clinical outcome, they can still be used for exploring mechanism of action of drugs and their contribution to the overall outcome.

Exploring the interactions between drug targets and signaling pathways can provide novel in-depth view of cellular mechanism and drug mode of action, which will ultimately rationalize tissue specific therapies. In cross tissue analysis (**Supp Fig. S6**), the triplet pathway/target/tissue allow drug repositioning and patient stratification strategies. It highlights cases of interaction that can provide useful biomarkers on one cancer type but potentially provide the inverse stratification for another cancer type, thus leading to treating the wrong patients. Knowing the variation of those interactions across tissues may be informative for drug repurposing, drug combination design and patient stratification.

4. METHODS

4.1 Macau: Algorithm

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 (**Fig. 1A**). 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 N 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. For more information, see Supp methods.

4.2 Macau: Feature interaction

Concept

We would like to know the interactions between the features of the drugs and the features of the cell lines. In our analysis, we used protein target to describe the drugs and gene expression/PROGENy pathways to describe the cell lines. Let IC50 be the matrix of drug response, D be the latent matrix of the drugs and C be the latent matrix of the cell lines (**Fig. 1A**):

$$IC50 = D^T C$$

if side information (feature) are available on both sides:

$$IC50 = (\beta_D{}^T x)^T \beta_C{}^T z$$

$$IC50 = x^T \boldsymbol{\beta_D \beta_C}^T z$$

The matrix $\boldsymbol{\beta_D \beta_C}^T$ is the interaction term or the interaction matrix through which the 2 feature sets interact in order to produce the response variable IC50 (**Fig. 1B**). We generate the interaction matrix between features of the drugs and features of the cell lines by multiplying the 2 link matrices $\boldsymbol{\beta_D}$ and $\boldsymbol{\beta_C}$ and averaging across 600 MCMC samplings. We use setting 3 (**Supp Fig. S1C, Table 1**) to compute the interaction matrix for the feature interaction analysis. This setting allows the use of the whole data set, without cross validation. MCMC sampling is also less prone to overfitting than optimization methods.

Each drug response observation can be written as a linear combination of all the possible

interactions between the protein targets and the pathways' activity, across all latent dimensions.

$$IC50_{ij} = \sum_{k=1}^{L} \sum_{j=1}^{F_C} \sum_{i=1}^{F_D} (ß_{Dik} ß_{Cjk} x_i z_j)$$

L: number of latent dimensions

F_D: length of drug feature (number of protein targets)

F_C: length of cell line feature (number of PROGENy pathways)

x: drug feature (protein target)

z: cell line feature (PROGENy pathway activity)

β_D: link matrix which projects drug feature X_D into latent matrix D

β_C: link matrix which projects cell line feature X_C into latent matrix C

Interpretation

The interaction matrix $\boldsymbol{\beta_D \beta_C^T}$ has as dimension the number of protein targets multiplied by the number of pathways. We then multiply the matrix by -1 so that the interpretation would be:

In case of a positive value in the matrix, the association of the corresponding protein target and the corresponding pathway confers sensitivity upon drug treatment. If the value is negative, it would be resistance.

Another way to say it would be: If the value is positive, activation of this specific pathway confers sensitivity to any drug targeting this specific protein.

Quality control

We use setting 4 (**Supp Fig. S1D, Table 1**) as quality control for the feature interaction analysis. Predicting new drugs' responses on new cell lines allows us to evaluate the quality of the interactions between features of the drugs and features of the cell lines.

4.3 Macau: Parameter setting

When predicting drug response on new cell lines (**Supp Fig. S1A**), we set the number of latent dimension L to 10 if we only use cell line feature. In case of adding drug feature, we set L to 30. Smaller L could lead to overcrowded latent space and decrease performance. In MCMC sampling, we choose a burn in of 400 samples, then we collect 600 samples. At each of those collected samples, we make the prediction and average across all 600 samples. In quality control of both sides of features, we use setting 4 (**Supp Fig. S1D**) and 2 simultaneous 10 fold cross validation and 30 latent dimensions. In feature interaction analysis, we use setting 3 (**Supp Fig. S1C**), predicting existing drug for existing cell line) with 30 latent dimensions.

4.4 PROGENy

PROGENy (Schubert et al. 2016) is a data driven dimension reduction method for gene expression data. It reduces high dimensional gene expression into a small number of pathway activity scores by a matrix multiplication with a weight matrix. PROGENy leverages hundreds of perturbation experiments. For each experiment, we assign a manually curated pathway activation status. The chosen experiments have been treated by a perturbation agent which activates or inhibits one of the PROGENy pathways.

We compute the gene expression z-scores of the $\text{Microarray}_{\text{perturbed}}$ **-** $\text{Microarray}_{\text{control}}$. Then, we fit a multiple linear model of the z-scores in function of the pathway status (**Supp Fig. S8A**). The z-scores representing the change in gene expression, we aim at determining the role of the pathway activation statuses in this change.

$$Z_{\text{gene}} = f(\text{pathways}) = \beta_0 + \beta_1 \text{EGFR} + \ldots\ldots + \beta_n \text{PI3K}$$

We obtain a pathway weight matrix from the fitted model. And for each pathway we select the 100 smallest p-values and keep those genes while setting the other genes' weights to zeros (**Supp Fig. S8B**).

For new gene expression data where we would like to know the pathway information, we compute the pathway scores by multiplying the gene expression matrix with the pathway weight matrix. If we take the example of EGFR, the pathway activity of EGFR on sample 1 (s1) is defined as:

$$PA_{EGFR,s1} = \sum_{k=1}^{100} \left(gene_{k,s1} \; x \; \beta_{EGFR,k} \right)$$

The pathway activity is defined as the product of a gene' expression by the contribution of a pathway's activation to the change in expression of this gene. From this formula, the higher the gene expression, the higher the pathway activity. Similarly, the higher the contribution of EGFR's activation to the change of gene expression, the higher the pathway activity.

The result is the pathway scores matrix with new experiments on the rows and pathways on the columns (**Supp Fig. S8C**). In practice, for any transcriptomics dataset, we can determine which pathway is up regulated or down regulated for a certain cell line relative to other cell lines. In this paper, we are using the pathway scores as features to predict drug response on cell lines. Therefore, PROGENy is used as a data driven dimension reduction method.

BIBLIOGRAPHY

Alyass, Akram, Michelle Turcotte, and David Meyre. 2015. “From Big Data Analysis to Personalized Medicine for All: Challenges and Opportunities.” *BMC Medical Genomics* 8 (June): 33.

Ammad-ud-din, Muhammad, Elisabeth Georgii, Mehmet Gönen, Tuomo Laitinen, Olli Kallioniemi, Krister Wennerberg, Antti Poso, and Samuel Kaski. 2014. “Integrative and Personalized QSAR Analysis in Cancer by Kernelized Bayesian Matrix Factorization.” *Journal of Chemical Information and Modeling* 54 (8): 2347–59.

Ammad-ud-din, Muhammad, Suleiman A. Khan, Krister Wennerberg, and Tero Aittokallio. 2017. “Systematic Identification of Feature Combinations for Predicting Drug Response with Bayesian Multi-View Multi-Task Linear Regression.” *Bioinformatics* 33 (14): i359–68.

Auphan, N., J. A. DiDonato, C. Rosette, A. Helmberg, and M. Karin. 1995. “Immunosuppression by Glucocorticoids: Inhibition of NF-Kappa B Activity through Induction of I Kappa B Synthesis.” *Science* 270 (5234): 286–90.

Baranski, Zuzanna, Tijmen H. Booij, Marieke L. Kuijjer, Yvonne de Jong, Anne-Marie Cleton-Jansen, Leo S. Price, Bob van de Water, Judith V. M. G. Bovée, Pancras C. W. Hogendoorn, and Erik H. J. Danen. 2015. “MEK Inhibition Induces Apoptosis in Osteosarcoma Cells with Constitutive ERK1/2 Phosphorylation.” *Genes & Cancer* 6 (11-12): 503–12.

Barretina, Jordi, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, et al. 2012. “The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity.” *Nature* 483 (7391): 603–7.

Comunanza, Valentina, Davide Corà, Francesca Orso, Francesca Maria Consonni, Emanuele Middonti, Federica Di Nicolantonio, Anton Buzdin, et al. 2017. “VEGF Blockade Enhances the Antitumor Effect of BRAFV600E Inhibition.” *EMBO Molecular Medicine* 9 (2): 219–37.

Cortés-Ciriano, Isidro, Lewis H. Mervin, and Andreas Bender. 2016. “Current Trends in Drug Sensitivity Prediction.” *Current Pharmaceutical Design*, October. https://www.ncbi.nlm.nih.gov/pubmed/27784247.

Costello, James C., Laura M. Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P. Menden, Nicholas J. Wang, Mukesh Bansal, et al. 2014. “A Community Effort to Assess and Improve Drug Sensitivity Prediction Algorithms.” *Nature Biotechnology* 32 (12): 1202–12.

Deng, Zhantao, Xiaozhou Liu, Jiewen Jin, Haidong Xu, Qian Gao, Yong Wang, and Jianning Zhao. 2016. “Histone Deacetylase Inhibitor Trichostatin a Promotes the Apoptosis of Osteosarcoma Cells through p53 Signaling Pathway Activation.” *International Journal of Biological Sciences* 12 (11): 1298–1308.

Dogan, Erbil, Ugur Saygili, Burçin Tuna, Mert Gol, Duygu Gürel, Berrin Acar, and Meral Koyuncuoğlu. 2005. “p53 and mdm2 as Prognostic Indicators in Patients with Epithelial Ovarian Cancer: A Multivariate Analysis.” *Gynecologic Oncology* 97 (1): 46–52.

Gajria, Devika, and Sarat Chandarlapaty. 2011. “HER2-Amplified Breast Cancer: Mechanisms of Trastuzumab Resistance and Novel Targeted Therapies.” *Expert Review of Anticancer Therapy* 11 (2): 263–75.

Gönen, Mehmet, and Adam A. Margolin. 2014. “Drug Susceptibility Prediction against a Panel of Drugs Using Kernelized Bayesian Multitask Learning.” *Bioinformatics* 30 (17): i556–63.

Ha, Sang Yun, Jeeyun Lee, Jiryeon Jang, Jung Yong Hong, In-Gu Do, Se Hoon Park, Joon Oh Park, et al. 2015. “HER2-Positive Gastric Cancer with Concomitant MET And/or EGFR Overexpression: A Distinct Subset of Patients for Dual Inhibition Therapy.” *International Journal of Cancer. Journal International Du Cancer* 136 (7): 1629–35.

Inamdar, Gajanan S., Subbarao V. Madhunapantula, and Gavin P. Robertson. 2010. “Targeting the MAPK Pathway in Melanoma: Why Some Approaches Succeed and Other Fail.” *Biochemical Pharmacology* 80 (5): 624–37.

Inokuchi, Mikito. 2015. “Clinical Significance of MET in Gastric Cancer.” *World Journal of Gastrointestinal Oncology* 7 (11): 317.

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.

Jost, Philipp J., and Jürgen Ruland. 2007. “Aberrant NF-kappaB Signaling in Lymphoma: Mechanisms, Consequences, and Therapeutic Implications.” *Blood* 109 (7): 2700–2707.

Katerinaki, E., G. S. Evans, P. C. Lorigan, and S. MacNeil. 2003. “TNF-Alpha Increases Human Melanoma Cell Invasion and Migration in Vitro: The Role of Proteolytic Enzymes.” *British Journal of Cancer* 89 (6): 1123–29.

Liu, Xianzhe, Edwin Choy, David Harmon, Shuhua Yang, Cao Yang, Henry Mankin, Francis J. Hornicek, and Zhenfeng Duan. 2011. “Inhibition of Polo-like Kinase 1 Leads to the Suppression of Osteosarcoma Cell Growth in Vitro and in Vivo.” *Anti-Cancer Drugs* 22 (5): 444–53.

Menden, Michael P., Francesco Iorio, Mathew Garnett, Ultan McDermott, Cyril H. Benes, Pedro J. Ballester, and Julio Saez-Rodriguez. 2013. “Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties.” *PloS One* 8 (4): e61318.

Nikolova, Olga, Russell Moser, Christopher Kemp, Mehmet Gönen, and Adam A. Margolin. 2017. “Modeling Gene-Wise Dependencies Improves the Identification of Drug Response Biomarkers in Cancer Studies.” *Bioinformatics* 33 (9): 1362–69.

Pan, Sinno Jialin, and Qiang Yang. 2010. “A Survey on Transfer Learning.” *IEEE Transactions on Knowledge and Data Engineering* 22 (10): 1345–59.

Rialdi, Alex, Laura Campisi, Nan Zhao, Arvin Cesar Lagda, Colette Pietzsch, Jessica Sook Yuin Ho, Luis Martinez-Gil, et al. 2016. “Topoisomerase 1 Inhibition Suppresses Inflammatory Genes and Protects from Death by Inflammation.” *Science* 352 (6289): aad7993.

Rogers, David, and Mathew Hahn. 2010. “Extended-Connectivity Fingerprints.” *Journal of Chemical Information and Modeling* 50 (5): 742–54.

Ryan, Denise, Mairin Rafferty, Shauna Hegarty, Patrick O’Leary, William Faller, Gabriela Gremel, Michael Bergqvist, et al. 2010. “Topoisomerase I Amplification in Melanoma Is Associated with More Advanced Tumours and Poor Prognosis.” *Pigment Cell & Melanoma Research* 23 (4): 542–53.

Schubert, Michael, Bertram Klinger, Martina Klünemann, Mathew J. Garnett, Nils Blüthgen, and Julio Saez-Rodriguez. 2016. “Perturbation-Response Genes Reveal Signaling Footprints in Cancer Gene Expression.” doi:10.1101/065672.

Seashore-Ludlow, B., M. G. Rees, J. H. Cheah, M. Cokol, E. V. Price, M. E. Coletti, V. Jones, et al. 2015. “Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset.” *Cancer Discovery* 5 (11): 1210–23.

Shen, Ying, Jie Li, Masayuki Nitta, Diahnn Futalan, Tyler Steed, Jeffrey M. Treiber, Zack Taich, et al. 2015. “Orthogonal Targeting of EGFRvIII Expressing Glioblastomas through Simultaneous EGFR and PLK1 Inhibition.” *Oncotarget* 6 (14): 11751–67.

Shi, Dingding, and Wei Gu. 2012. “Dual Roles of MDM2 in the Regulation of p53: Ubiquitination Dependent and Ubiquitination Independent Mechanisms of MDM2 Repression of p53 Activity.” *Genes & Cancer* 3 (3-4): 240–48.

Simm, Jaak, Adam Arany, Pooya Zakeri, Tom Haber, Jörg K. Wegner, Vladimir Chupakhin, Hugo Ceulemans, and Yves Moreau. 2015. “Macau: Scalable Bayesian Multi-Relational Factorization with Side Information Using MCMC.” *arXiv [stat.ML]*. arXiv.

http://arxiv.org/abs/1509.04610.
Tanaka, Kazuhiro, Ivan Babic, David Nathanson, David Akhavan, Deliang Guo, Beatrice Gini, Julie Dang, et al. 2011. “Oncogenic EGFR Signaling Activates an mTORC2-NF-κB Pathway That Promotes Chemotherapy Resistance.” *Cancer Discovery* 1 (6): 524–38.
Wang, Yongcui, Jianwen Fang, and Shilong Chen. 2016. “Inferences of Drug Responses in Cancer Cells from Cancer Genomic Features and Compound Chemical and Therapeutic Properties.” *Scientific Reports* 6 (September): 32679.
Yuan, Han, Ivan Paskov, Hristo Paskov, Alvaro J. González, and Christina S. Leslie. 2016. “Multitask Learning Improves Prediction of Cancer Drug Sensitivity.” *Scientific Reports* 6 (August): 31619.

FIGURES

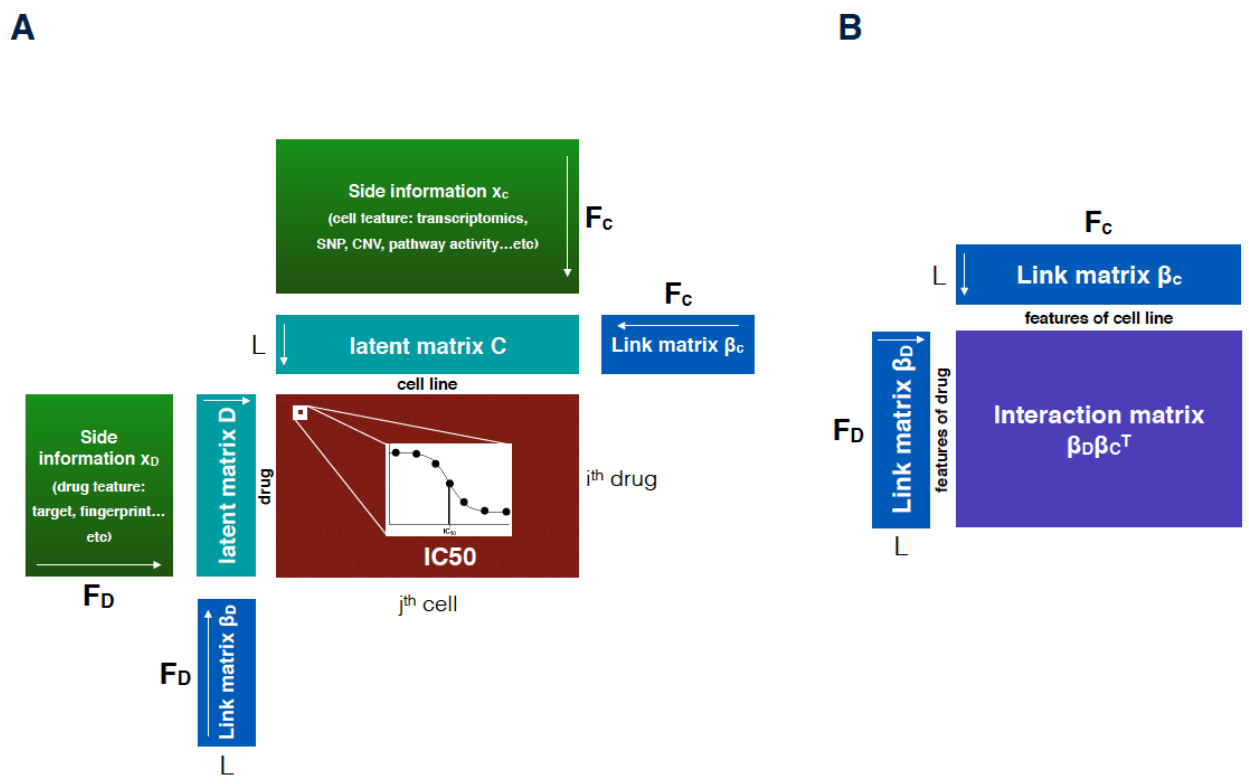


Fig. 1: **A)** Macau's factorization model. The drug response (IC50) is computed by 2 latent matrices. Each of them is being sampled by a Gibbs sampler. In presence of additional information (side information), the latent matrix is predicted by a multiplication of a link matrix and the side information matrix. Arrows in this Fig. indicate the matrix multiplication. **B)** By multiplying the 2 link matrices, we obtain the interaction matrix, which is the interaction between the features of the drugs with the features of the cell lines.

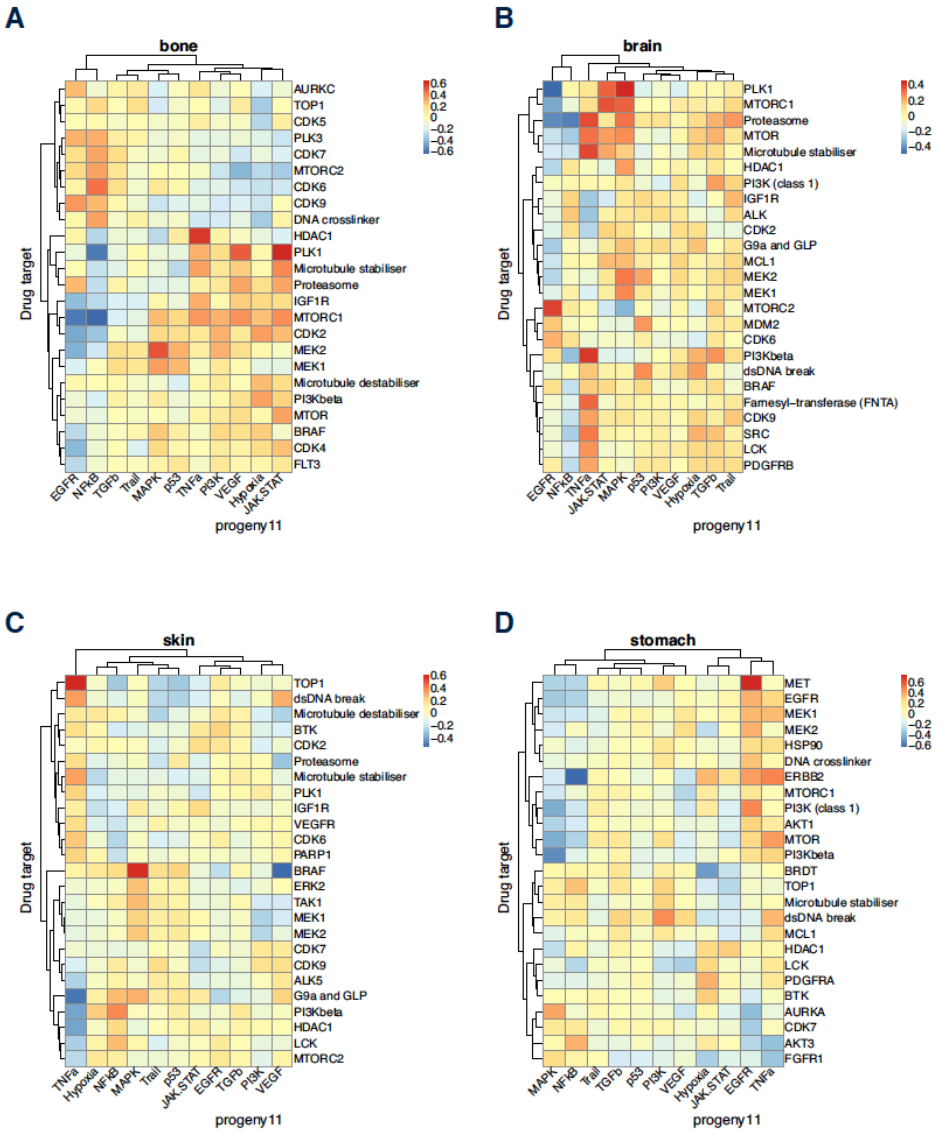


Fig. 2: Tissue specific analysis of interaction matrix with target on drug side and pathway on cell line side. We analyzed all tissues in the GDSC panel with at least 20 samples, and display the targets which have an association for at least 1 pathway in the top 5% absolute value. We subset the targets a second time by keeping the top 25 targets with the highest variance across the pathways in term of associations. Here, we highlight 4 representative tissues: **A)** Bone. **B)** Brain. **C)** Skin. **D)** Stomach.

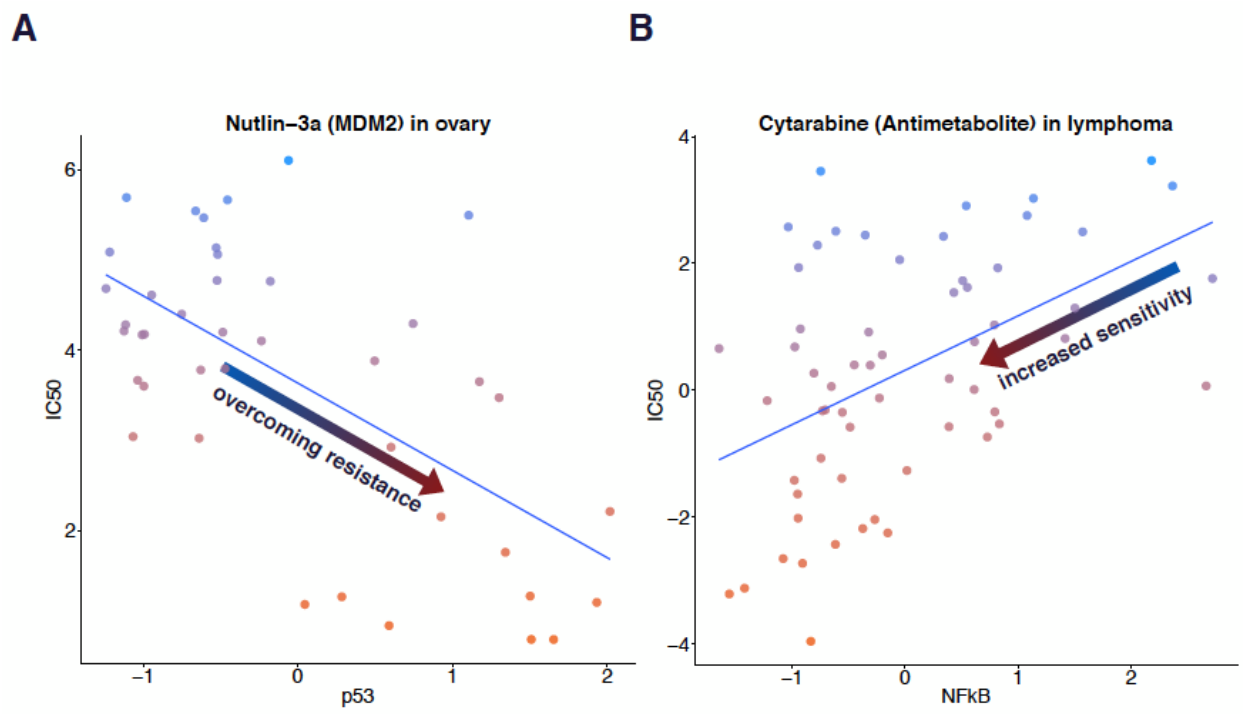
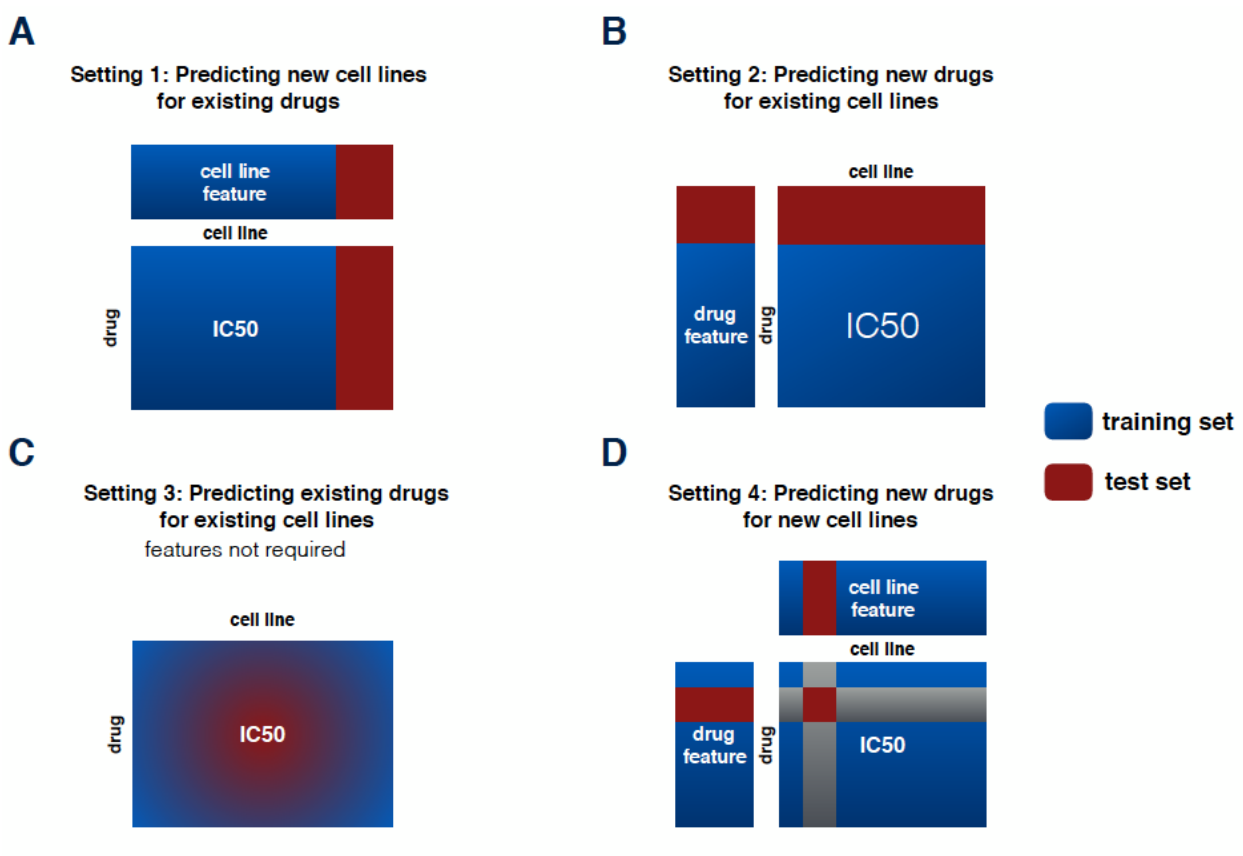
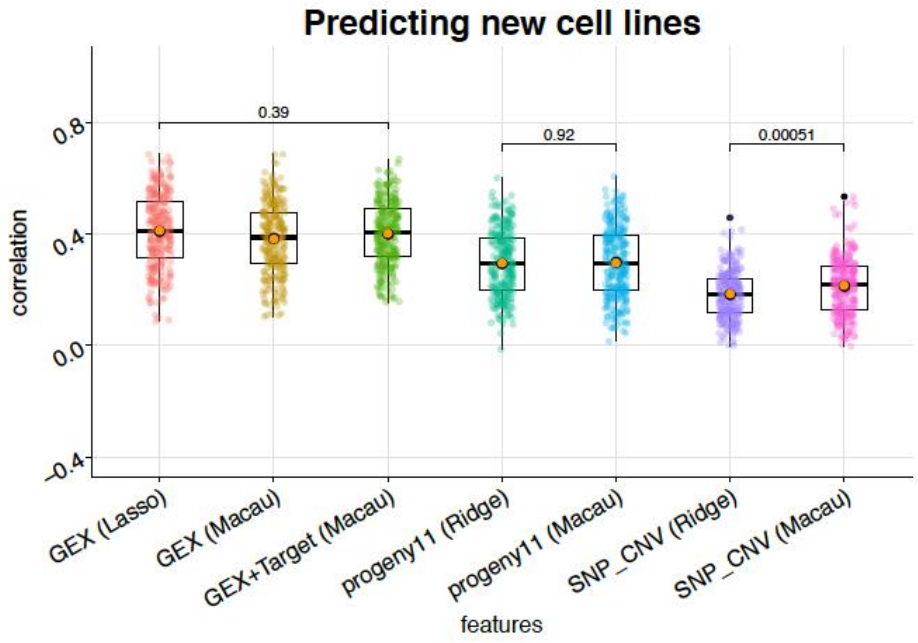


Fig. 3: Increasing sensitivity and overcoming resistance. We chose 2 examples to illustrate the power of PROGENy pathway as a biomarker. **A)** From tissue specific interaction matrix of ovarian cancer, we chose the top hits MDM2 - p53 (as target - pathway pairs). We plot the IC50 (in log scale) of drug Nutlin-3a which targets MDM2 against p53 pathway's activity. **B)** From tissue specific interaction matrix of lymphoma, we chose the top hits Antimetabolite - NFkB (as target - pathway pairs). We plot the IC50 of drug Cytarabine against NFkB pathway's activity.

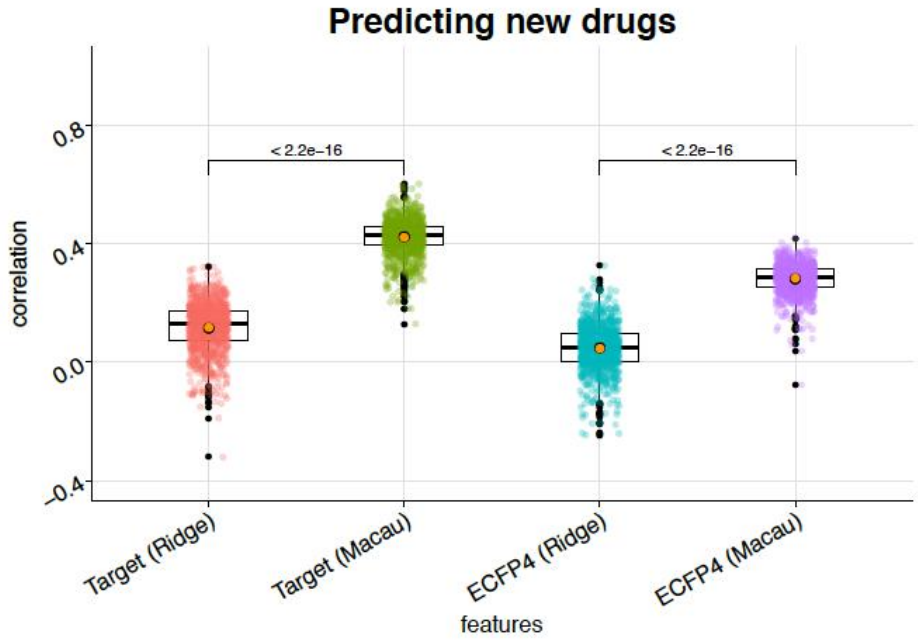


Supp Fig. S1: Different settings in drug response prediction. **A)** Predicting new cell lines for existing drugs. For each drug, we compute the pearson correlation of observed versus predicted IC50 across all cell lines of the test set. **B)** Predicting new drugs for existing cell lines. For each cell line we compute the pearson correlation of observed versus predicted IC50 across all drugs of the test set. **C)** Predicting existing drugs 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 IC50 for all randomly chosen drug - cell line pairs of the test set. **D)** Predicting new drugs 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 IC50 for all drug - cell line pairs of the test set.

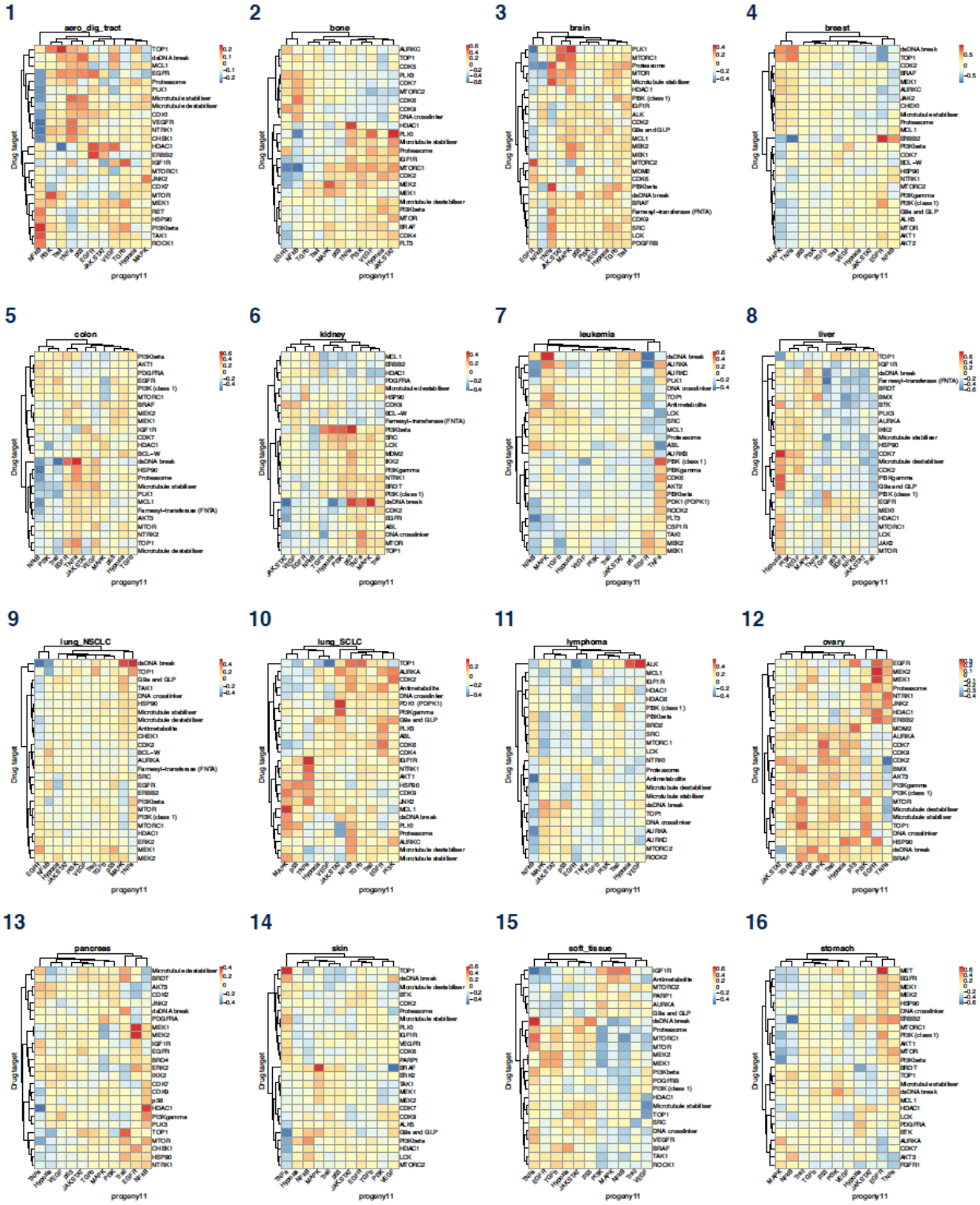
A



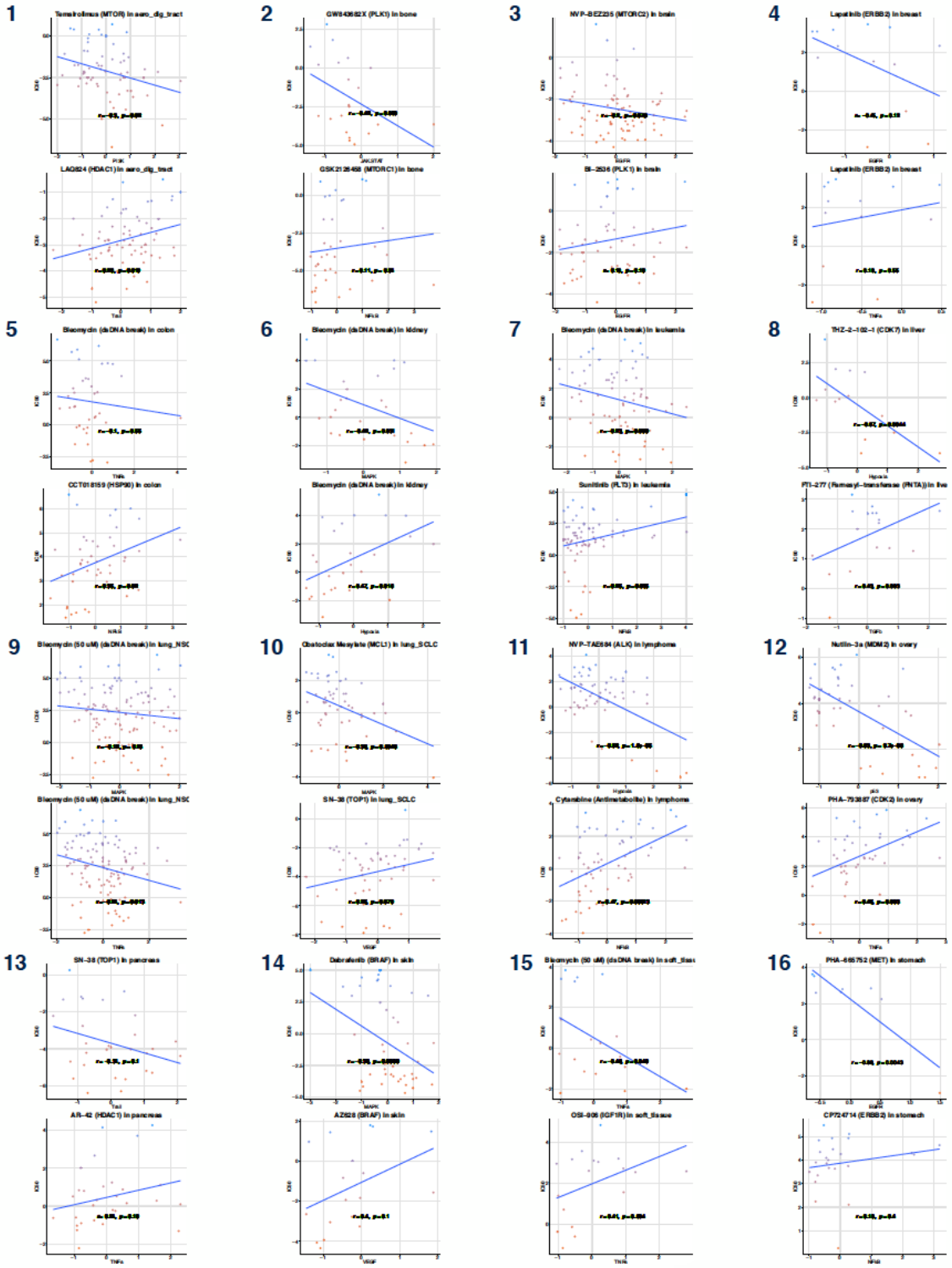
B



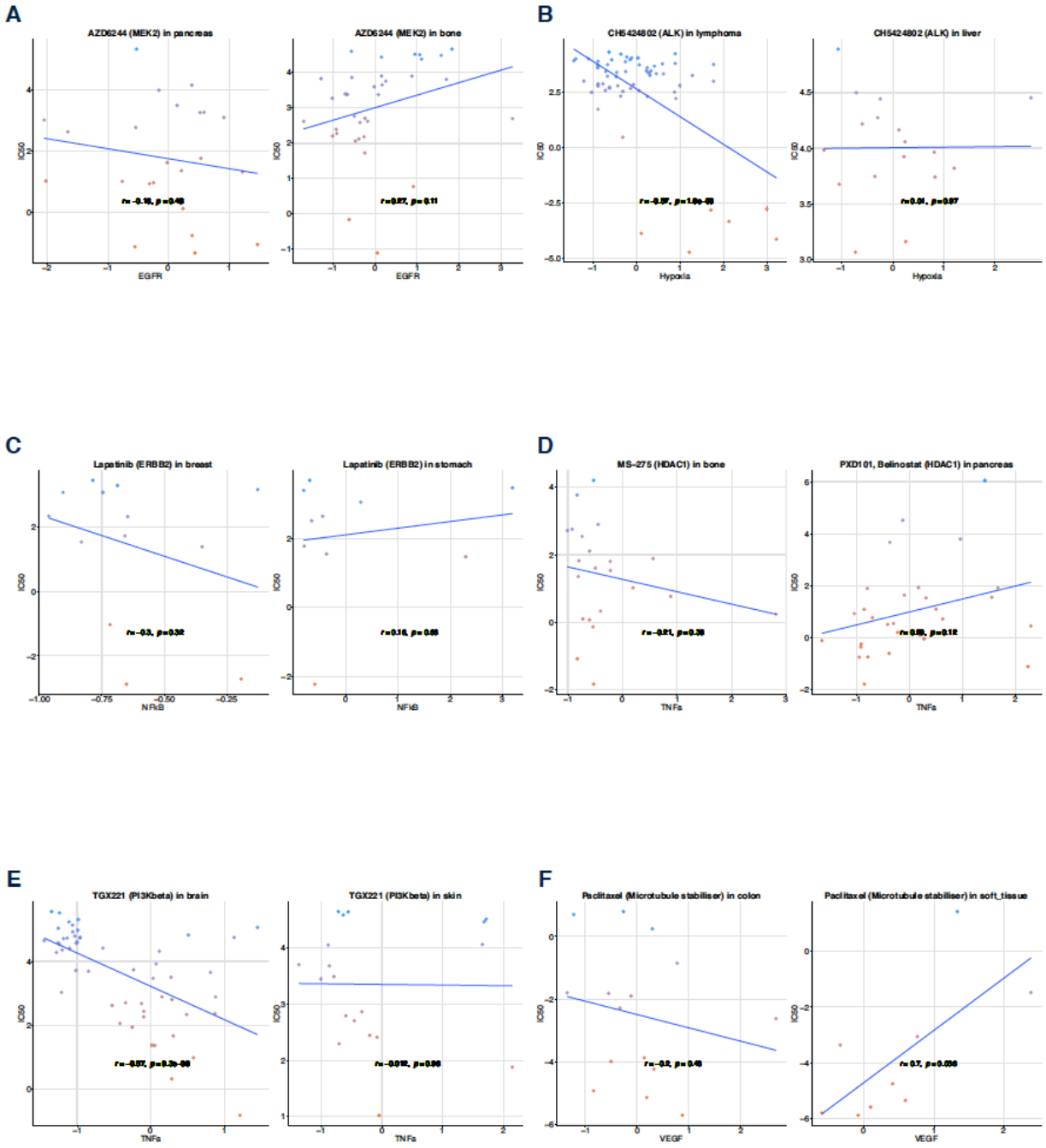
Supp Fig. S2: Drug response prediction performance. **A)** We compare prediction performance (correlation of observed versus predicted IC50) of existing drugs on new cell lines. We use Macau and standard ridge/lasso regression. The features are gene expression, pathway activity, mutation (SNP) and copy number variation (CNV). **B)** We compare prediction performance of existing cell lines on new drugs. The features are drug protein targets and ECFP4 fingerprint.



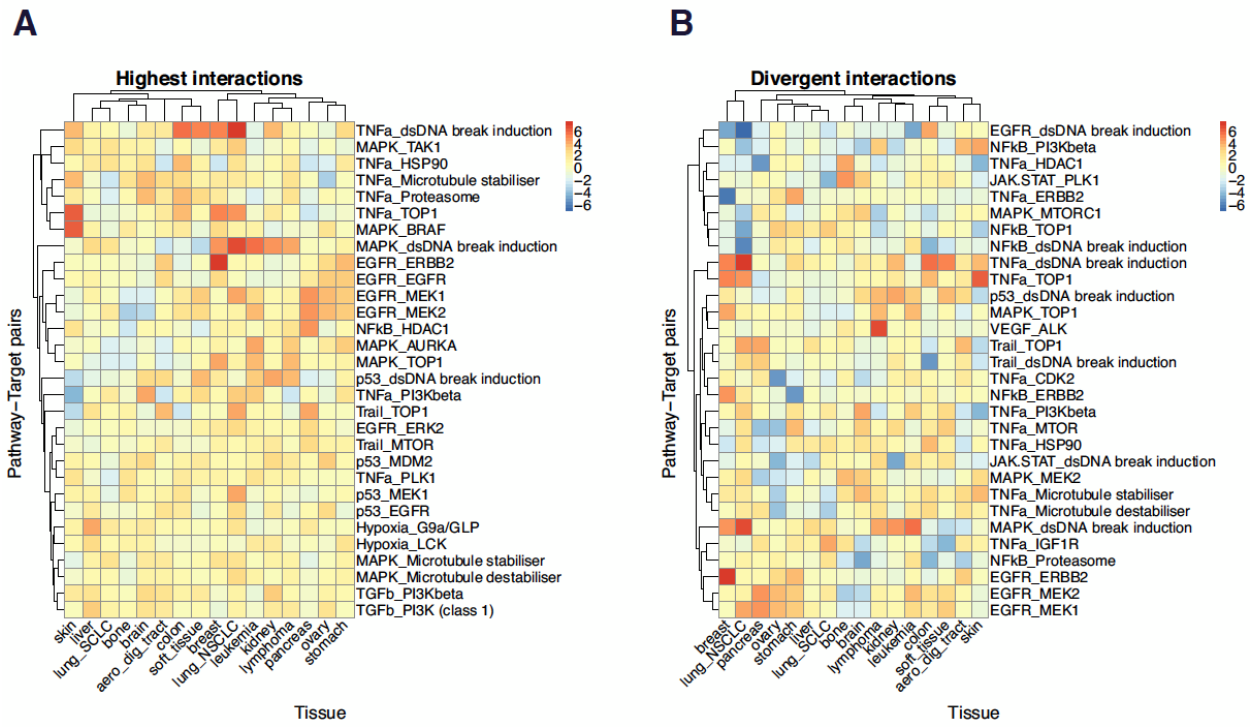
Supp Fig. S3: Tissue specific analysis of interaction matrix. We chose 16 tissues in the GDSC panel with at least 20 samples. We keep the targets which have an association for at least 1 pathway in the top 5% absolute value. We subset a second time by keeping the top 25 targets with the highest variance across the pathways in term of associations.



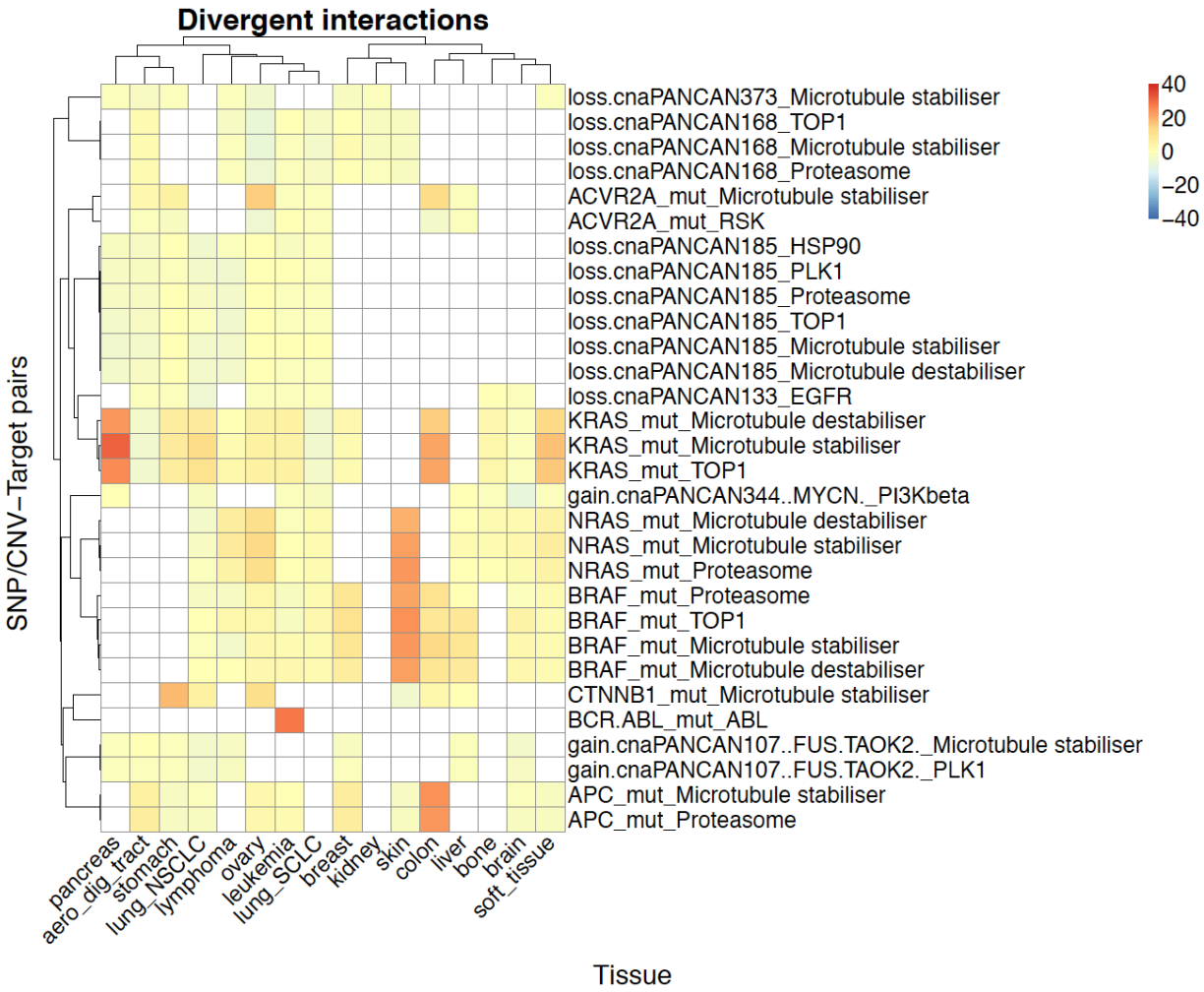
Supp Fig. S4: PROGENy as biomarker. For each tissue specific interaction matrix, we select a top positive association and a top negative association. For both target - pathway pairs, we then find a drug which targets this protein (as described in the manually curated list) and plot its IC50 (log scale) against the corresponding pathway's activity in the specific tissue.



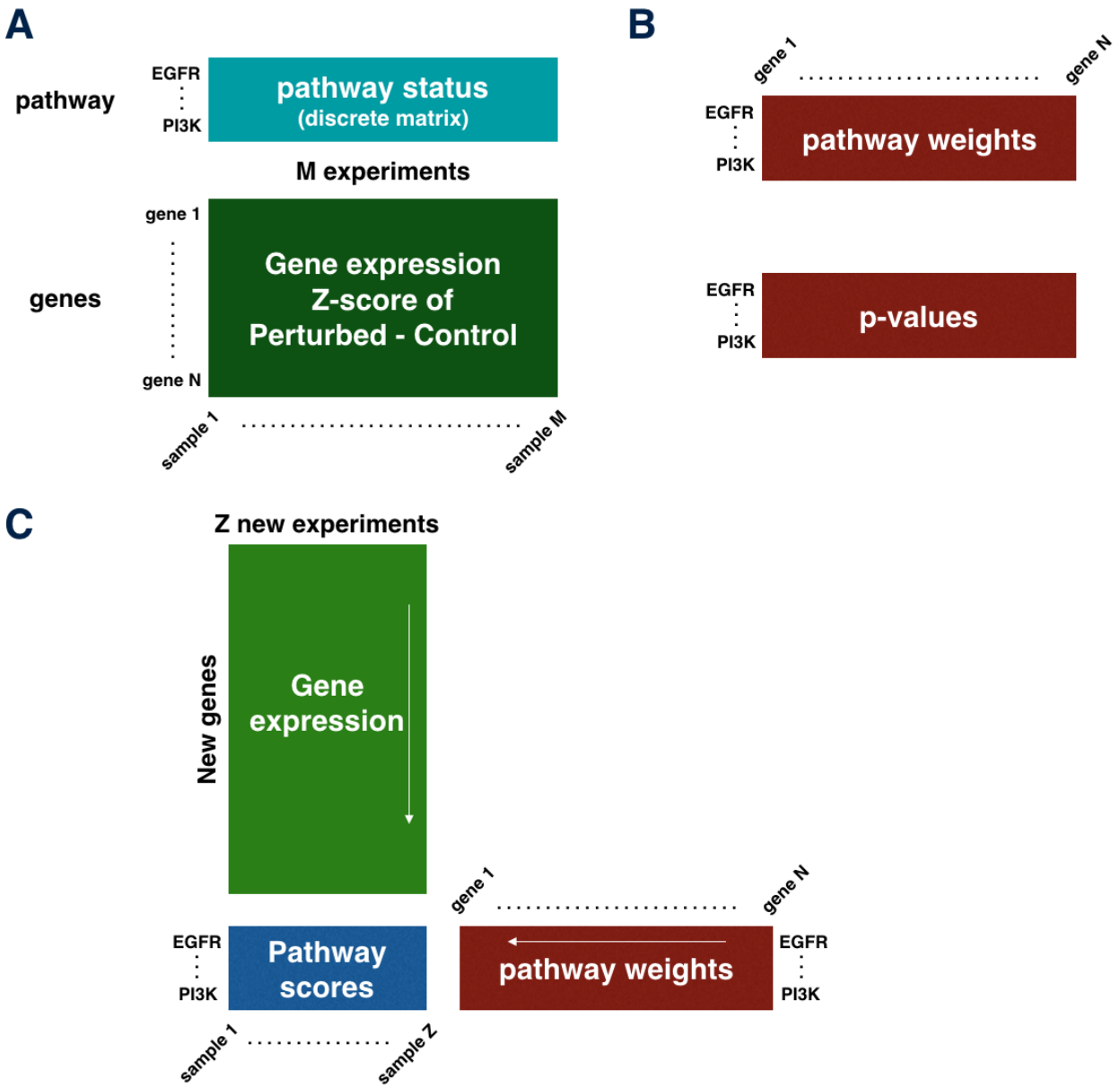
Supp Fig. S5: Antagonistic tissues based on target pathway interaction. For all target - pathway pairs which have opposite effect from one tissue to another, we select a drug which specifically targets the protein and plot the drug's IC50 (log scale) as function of PROGENy activity for the corresponding tissues.



Supp Fig. S6: Feature interaction analysis across tissues. **A)** Highest interactions. We vectorize all cancer specific interaction matrices between target and PROGENy pathways and obtain a matrix of dimension (number of tissues x number of pathway-target pairs). We do a first subsetting by taking only into account the pairs for which at least one pathway appears in the top 5% absolute value. We then keep the 30 pathway-target pairs with the highest mean value across tissues in term of association. **B)** Divergent interaction. Same as in A, except that we keep the top 30 pairs with highest variance across tissues.



Supp Fig. S7: Feature interaction analysis across tissues for SNP/CNV. We vectorize all cancer specific interaction matrices between target and SNP/CNV and obtain a matrix of dimension (number of tissues x number of SNP/CNV-target pairs). We do a first subsetting by taking the pairs for which at least one pathway appears in the top 1% highest value, and chose 15 SNP/CNV-target pairs with highest variance of interaction across tissues. We then subset by taking the pairs for which at least one pathway appears in the top 1% lowest value, and chose 15 SNP/CNV-target pairs with highest variance of interaction across tissues. We combine the top hits and then keep the 30 pathway-target pairs. White color indicates when the mutation or CNV is not present.



Supp Fig. S8: Workflow to produce PROGENy scores. **A)** We fit a linear model for each z-score of the perturbation in function of the pathway status. **B)** We select for each pathway, the top 100 genes with smallest p-values. **C)** We compute pathway scores for new gene expression dataset by a matrix multiplication with the weight matrix.

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, pearson correlation of observed versus predicted IC50 across all cell lines.	For each cell line, pearson correlation of observed versus predicted IC50 across all drugs.	Pearson correlation of observed versus predicted IC50 for all drug-cell line pairs.	Pearson correlation of observed versus predicted IC50 for all drug-cell line pairs.

Table 1: Different settings for drug response prediction

response	repetition	test set	latent	N samples	Pearson correlation	RMSE
IC50	10	10%	10	600	0.932 (sd=0.0011)	0.966 (sd=0.015)
IC50	10	20%	10	600	0.931 (sd=0.00053)	0.982 (sd=0.007)
IC50	10	30%	10	600	0.929 (sd=0.00062)	0.996 (sd=0.005)
IC50	10	50%	10	600	0.927 (sd=0.00036)	1.038 (sd=0.005)
IC50	10	70%	10	600	0.919 (sd=0.0002)	1.135 (sd=0.003)
IC50	10	99%	10	600	0.834 (sd=0.004)	2.239 (sd=0.044)

Supp Table S1: Prediction performance for missing value imputation

Drug	Cell line
target	Transcriptomics
pathway	Proteomics (only for colorectal cancer)
Fingerprint (ECFP4)	TF activity
	SNP, CNV
	Pathway activity (PROGENy)
	Phosphoproteomics (only for colorectal cancer)
	Kinase activity (only for colorectal cancer)

Supp Table S2: Available features for drug and cell line sides

	target_GEX	target_PROGENy11	target_SNP_CNV
aero_dig_tract	0.456 (sd=0.168)	0.441 (sd=0.189)	0.418 (sd=0.174)
bone	0.412 (sd=0.18)	0.401 (sd=0.177)	0.376 (sd=0.168)
brain	0.433 (sd=0.175)	0.426 (sd=0.166)	0.394 (sd=0.171)
breast	0.426 (sd=0.171)	0.407 (sd=0.17)	0.372 (sd=0.176)
colon	0.433 (sd=0.172)	0.395 (sd=0.164)	0.396 (sd=0.172)
kidney	0.38 (sd=0.189)	0.392 (sd=0.191)	0.36 (sd=0.184)
leukemia	0.393 (sd=0.148)	0.388 (sd=0.155)	0.35 (sd=0.156)
liver	0.388 (sd=0.202)	0.334 (sd=0.223)	0.323 (sd=0.216)
lung_NSCLC	0.421 (sd=0.171)	0.406 (sd=0.181)	0.351 (sd=0.163)
lung_SCLC	0.419 (sd=0.15)	0.415 (sd=0.153)	0.381 (sd=0.176)
lymphoma	0.417 (sd=0.164)	0.408 (sd=0.153)	0.332 (sd=0.168)
ovary	0.441 (sd=0.175)	0.418 (sd=0.178)	0.385 (sd=0.189)
pancreas	0.367 (sd=0.217)	0.337 (sd=0.213)	0.379 (sd=0.207)
skin	0.448 (sd=0.177)	0.454 (sd=0.158)	0.409 (sd=0.169)
soft_tissue	0.402 (sd=0.194)	0.359 (sd=0.207)	0.361 (sd=0.219)
stomach	0.437 (sd=0.181)	0.376 (sd=0.199)	0.383 (sd=0.187)

Supp Table S3: Tissue specific prediction performance for setting 4: prediction of new drugs on new cell lines

pathway	target	Max tissue	Min tissue	Max value	Min value	Max - Min	absolute mean
VEGF	Microtubule stabiliser	colon	Soft tissue	0.351	-0.474	0.825	0.123
Hypoxia	ALK	lymphoma	liver	0.519	-0.369	0.888	0.15
TNFa	HDAC1	bone	pancreas	0.592	-0.427	1.02	0.165
TNFa	PI3Kbeta	brain	skin	0.414	-0.39	0.804	0.024
NFkB	ERBB2	breast	stomach	0.632	-0.637	1.27	0.005
EGFR	dsDNA break	colon	breast	0.478	-0.667	1.15	0.189
EGFR	MEK2	pancreas	bone	0.424	-0.426	0.85	0.002

Supp Table S4: Top antagonistic pathway - target pairs across tissues