Multi-task learning

of transcriptomic signatures underlying cancer gene dependencies

B. Rentroia Pacheco

TU Delft

Multi-task learning

of transcriptomic signatures underlying cancer gene dependencies

by

B. Rentroia Pacheco

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday August 29, 2019 at 13:00.

Student number:	4718399	
Master programme:	Computer Science, Data Science and Technology track	
Faculty:	Electrical Engineering, Mathematics and Computer Science	
Project duration:	November 12, 2018 – August 29, 2019	
Thesis committee:	Prof. dr. ir. Marcel Reinders,	TU Delft
	Dr. Joana Gonçalves,	TU Delft, supervisor
	Dr. Elvin Isufi,	TU Delft
	Prof.dr. Roeland van Ham,	TU Delft

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

TU Delft

Preface

This MSc thesis is the final result of my graduation project in the Pattern Recognition and Bioinformatics research group at the Technical University of Delft. Ten months ago, my supervisor Joana Gonçalves introduced me to a recently released dataset focused on cancer genetic dependencies. I immediately felt drawn to this topic. It is well known that genetic alterations in cancer cells contribute to their malignant properties. Even so, I had never thought that such alterations could actually make cancer cells more dependent on specific genes for their survival. Therapies based on these genetic dependencies are therefore quite promising because healthy cells might not share the same dependencies. However, to develop these therapies it's important to understand the relationships between alterations in cancer cells and their dependencies. To do so, we explored the use of a multi-task learning method that can learn hidden (latent) patterns in the data. Previous research on predicting cancer dependencies has used this type of algorithms, but with a focus on their performance. Instead, we switched the problem from only predicting genetic dependencies to also interpreting the latent patterns learned by the multi-task learning method. This brings additional challenges but at the same time it unlocks other types of insights. Investigating what this method was learning ended up being an excellent learning experience for me as well.

The results of this investigation are reported in the thesis main article. Namely, we show that the latent patterns learned by the multi-task algorithm capture relevant biological information in the cancer dependency dataset. The supplementary material complements the main article with extended background and methods as well as supplementary results.

This work would not have been possible without the support of many people. First, I would like to thank my supervisor Joana Gonçalves, for her guidance and enthusiasm during this project. I always looked forward to our weekly meetings. Thank you for always providing me critical feedback and help me become a better researcher. I would like to thank Luisa and Daniel for introducing me to the world of bioinformatics. My career path would have been very different (and less fun) if it wasn't for you two. I am also grateful to my friends for their support during my master's studies at TU Delft. A special thanks to Nirmal, Milagros, Lucas, Laura, Ombretta, Anna, Julia, Karol, Daniel and Leonor for being there for me in several occasions. Lastly, I would like to thank my family for their unconditional support and encouragement.

Bárbara Rentroia Pacheco
Delft, August 2019

Article

Multi-task learning of transcriptomic signatures underlying cancer gene dependencies

Bárbara Rentroia Pacheco [1] and Joana Gonçalves [1,*]

[1] Department of Intelligent Systems, Faculty EEMCS, Delft University of Technology, The Netherlands.

*Supervisor

Abstract

Motivation: Due to their altered genetic context, cancer cells can become dependent on specific genes for their survival. Such cancer-specific dependencies may represent promising therapeutic targets. However, knowledge on which molecular features of cancer cells induce specific dependencies is still limited and hampers the development of effective targeted therapies. Several large scale studies have systematically measured the dependency of hundreds of known cancer cell lines on thousands of genes using gene silencing. These data have enabled the learning of supervised models to predict dependencies of cancer cells on each gene based on molecular features of the cells. In particular, linear regression with regularization, such as Elastic Net, has been used to select molecular features associated with such dependencies. Since these approaches model dependencies for each gene independently, the selected features provide limited insight into common mechanisms underlying gene dependency. Moreover, they may fail to identify robust associations with gene dependency due to the small size of the available training data.

Results: In this work, we apply a multi-task learning approach (Macau) to learn the relationship between transcriptome and gene dependency in cancer cell lines for multiple genes simultaneously. To do so, Macau projects genes, cancer cell lines and their features into a shared latent space. We explore this latent space to go beyond linking individual transcriptomic features with dependencies, and further associate pathway changes with functionally related genes without enforcing prior knowledge on pathway structure. Although Macau and Elastic Net yield similar predictive performance, they find different kinds of associations. First, Macau favors features that are relevant for predicting dependency across multiple genes. Second, Macau captures inherent functional relationships between genes and leverages these to predict cancer gene dependencies. Additionally, Macau can recover similarities between cancer cell lines belonging to the same cancer type based on their dependencies only. In summary, modelling cancer dependencies simultaneously for multiple genes can reveal underlying mechanisms shared by functionally related genes, which would be missed when learning models independently per gene.

1 Introduction

Progression from normal to cancer cells is characterized by the accumulation of genetic alterations, that is, mutations or changes to their DNA. This altered genetic context is responsible for malignant hallmarks of cancer enabling uncontrolled cell proliferation (1). Notably, genetic alterations can also give rise to new vulnerabilities in cancer cells, which may become more dependent on specific genes for their survival. These genes are interesting candidates for targeted therapy, as their inhibition can affect cancer cell viability without being harmful to healthy cells (2).

Characterizing genetic dependencies of cancer cells is not trivial, since they are influenced by both environmental and genetic contexts. For example, cancer cells may become dependent on a gene only when exposed to certain environmental conditions or when the function of another gene is altered (3). Therefore, identifying molecular features of cancer cells (e.g. DNA mutations or gene expression levels), that induce specific genetic dependencies is key to better understand the disease and develop effective targeted therapies.

To promote the systematic discovery of cancer dependencies and their molecular drivers, four research groups have recently conducted large-scale loss-of-function screens (4; 5; 6; 7). These screens measure the dependency of cancer cells on a given gene by quantifying how cell proliferation is affected when the gene is silenced using RNA interference or CRISPR-based technology (see Supplementary Background 1.1.1 for details). The four studies have profiled dependencies of hundreds of cancer cell lines on thousands of individual genes. Cell lines are cancer models whose molecular features are well characterized.

Many studies have used large-scale cancer dependency data to validate their own identified associations between specific cell lines and a handful of genes, such as (8; 9) (see Supplementary Background 1.2 for a review of studies on cancer dependency). However, more extensive analysis of these data as a whole is needed in order to generate new insights into molecular mechanisms underlying cancer dependencies. In this context, other recent studies showed that gene dependency profiles across cell lines can be used to infer gene function (10; 11; 12), providing evidence of functional relationships between cancer genetic dependencies. Two other studies used different feature selection techniques to identify molecular features that are predictive of dependency, one using regression models (4) and another

using clustering in combination with statistical tests (5). By modelling dependencies for one gene at a time, these approaches identify molecular features that contribute to dependency of cancer cell lines on one gene only. Consequently, they can easily miss relevant features whose predictive value may possibly be weaker, but that overall still associate strongly with dependency across multiple genes. Shared associations might not only be more robust, they are also likely to expose relevant functional relationships between genes in cancer that would not be apparent otherwise. Since genes do not act in isolation, these relationships can ultimately lead to a better understanding of the mechanisms enabling cancer cell survival. In addition, learning individual models per gene reduces the number of samples available for training from $\sim$1 million to a few hundred. This translates into lower statistical power to find associations between the tens of thousands of molecular features and dependencies of cancer cell lines.

Multi-task learning can overcome these challenges by learning related tasks concurrently - that is, predicting dependency on multiple genes simultaneously (see Supplementary Background 1.3 for a summary and (13) for a survey on multi-task learning). The idea is that exploiting patterns of similarity and difference across the learning tasks can lead to improved prediction performance compared to learning independent models. Multi-task learning approaches are widely used to predict drug response in cancer cell lines (14; 15) and have also been applied to cancer dependency data (16). However, the exploration of these approaches in gene dependency prediction has been mainly focused on improving prediction accuracy and associating individual features with gene dependency (16).

We reason that this additional power and robustness to discover features that contribute to multiple gene dependency profiles can be further explored. Amongst multi-task feature learning algorithms, we identify matrix factorization based approaches as most suitable for our purpose (17; 18; 19; 20). Such methods capture common patterns or signatures in cancer cell lines through latent spaces associated with dependency profiles across multiple genes. Latent spaces are appealing because they exploit inherent relationships in the data to reduce dimensionality and expose higher-order organization which can provide increased interpretability.

With these considerations in mind, we explore the use of Macau (21), a multi-task learning algorithm based on bayesian matrix factorization, to predict gene dependency. Macau has the particularity of finding latent patterns both in the cell line space and in the cancer gene dependency space. These latent patterns are predictive of gene dependency and are associated with cancer cell line features. We hypothesize that these latent patterns naturally capture the functional organization of both genes and cell line features into biological processes or pathways in cancer cells. As a result, Macau could enable the discovery of functional relationships in cancer cells, without relying on general prior knowledge of pathway structure.

Using Macau, we aim to move beyond identifying associations between individual cell line features and individual cancer gene dependencies. Specifically, we focus on analyzing associations at a higher level of abstraction: the functional or pathway level. In practice, this translates into finding groups of functionally related cell line features whose signatures associate with multiple (functionally related) gene dependency profiles in cancer cell lines. Analyzing pathways instead of individual features or genes provides: (i) additional explanatory value on the mechanisms underlying cancer gene dependencies, (ii) increased statistical power and robustness to detect relevant associations between molecular signatures and gene dependency in cancer cell lines.

In this work, we assess the performance of Macau to predict cancer gene dependencies compared to a baseline single-task learning approach (Elastic Net (22)). We systematically test the ability of Macau to capture biologically relevant functional relationships through its latent spaces. Finally, we investigate the biological interpretation of the latent spaces produced by Macau.

2 Methods

We applied Macau to identify molecular signatures of cancer cell lines modulating the dependency of such cell lines on different genes. In this section, we describe the data and methodology used to discover patterns of association between molecular signatures and gene dependencies, as well as the strategies used to evaluate the results.

2.1 Data and Preprocessing

We analyzed two types of data to find molecular signatures associated with gene dependency in cancer cell lines: dependency scores measuring cancer cell line dependency on different genes (23), and transcriptomic features of cancer cell lines (24). These data were made available by the Cancer Dependency Map Consortium (4).

2.1.1 Gene dependencies of cancer cell lines

We used D2 dependency scores (23) integrating measurements from three large-scale screens (4; 5; 6). A dependency score denotes the sensitivity of a given cancer cell line to the suppression of a particular gene. In loss-of-function shRNA screens, a dependency score is based on the depletion of the shRNA(s) used to target the gene after the perturbation. Of note, shRNA depletion scores cannot be interpreted directly, due to off-target effects. Therefore, the DEMETER2 (D2) method was applied to isolate the gene suppression effect in a given cell line from off-target effects and screen quality variations across different cell lines (see Supplementary Background 1.1.2 for details). D2 scores can be interpreted as follows: zero means that the cell line is not dependent on a given gene, whereas a negative or positive value indicates that the cell line growth was respectively impaired or promoted after suppression of such gene. We used the combinedRNAi DepMap release 18Q4 dataset, hereafter referred to as dependency data. This dataset contains D2 dependency scores for 17212 genes in 712 cancer cell lines. Since not all gene-cancer cell line pairs were screened, around 20% of the data corresponded to missing values.

2.1.2 Transcriptomic features of cancer cell lines

To characterize cancer cell lines at the molecular level, we used transcriptomic feature data from the Cancer Cell Line Encyclopedia (CCLE) (24). Transcriptomic features quantify the expression level of gene transcripts in a given cancer cell line. In other words, transcriptomic features measure the "activity" level of a gene in the cells. We used data containing RNA-seq expression measurements for $\sim$50k transcripts in hundreds of cancer cell lines (expression DepMap release 18Q4). The expression levels available in this release correspond to TPM (transcripts per million) normalized data followed by log2 transformation, using a pseudo-count of 1. Although other types of molecular features were available, such as mutation and copy number, we decided to focus on expression features only. The reasons for this were two-fold. First, mutation and copy number are associated with lower level features, the genetic code (DNA), while expression indicates function by measuring which parts of the code are transcribed to generate the proteins needed to perform biological tasks in the cell. Therefore, they capture underlying functional interactions between genes and expose downstream effects of genetic changes such as mutations. Second, expression features have shown to be more informative to predict cancer dependency (4). This was corroborated by us in a preliminary analysis (see Supplementary Results 3.3). In addition to expression features, we used metadata on cancer cell lines, such as cancer type.

2.1.3 Biological pathway data

We used three databases to obtain functional information about genes and provide biological interpretation of our results: CORUM core protein complexes (25), KEGG pathways (26) and Reactome (27) pathways. CORUM database contains thousands of protein complexes, which correspond to groups of proteins that interact with each other to perform a given function. KEGG and Reactome databases contain hundreds of gene sets, which correspond to groups of genes that are involved in the same biological process (i.e. pathway). We used these databases because they offer complementary information. Proteins in the same protein complex have a physical interaction, while genes involved in the same pathway coordinate functionally to perform a given biological task. In addition, Reactome pathways typically correspond to smaller gene sets than KEGG pathways. Therefore, analysis of Reactome pathways may provide more specific insight than KEGG pathways, while the latter can give a better overview of the functional interpretation of the results.

2.1.4 Data preprocessing

We analyzed only cancer cell lines present in both the dependency and expression data, or 647 cell lines in total. Transcriptomic features with 0 standard deviation were removed. In addition, we kept only transcripts corresponding to protein coding genes, since these are better characterized and therefore more easily interpretable from a biological perspective. We used the Biomart R package (28)(v2.38.0) to identify these transcripts. We then averaged the expression of all transcripts for the same gene. After preprocessing, each cell line was characterized by the expression levels of 18718 genes. We hereafter refer to these as transcriptomic features.

2.2 Learning Modulators of Cancer Gene Dependency

Our goal is to identify transcriptomic signatures that modulate the dependency of cancer cells on functionally related genes. We translate this goal into a supervised multi-task feature learning problem (Figure 1A). The supervised multi-task learning part consists in jointly learning regression models for multiple genes simultaneously (output data, comprising multiple continuous dependent variables) based on the transcriptomic features characterizing the cancer cell lines (input data, comprising a range of continuous independent variables). The feature learning part of the task consists in determining the importance or contribution of those features for predicting dependency scores. We use the term feature learning rather than feature selection or ranking, since we are not interested in selecting individual features. We aim to find signatures or meta-features that capture patterns shared by multiple cell line features and gene dependency profiles, under the assumption that these can expose functional relationships of biological interest. With this in mind, we used Macau, a multi-task bayesian matrix factorization method that allows for the incorporation of side information (21). Macau models dependencies on all genes simultaneously and projects both cancer cell line features and genes into a lower dimensional latent space. Of note, this method is particularly suited for this task because it is able to deal with high-dimensional feature spaces and with the missing values present in the dependency scores data, due to its probabilistic nature. To evaluate the performance of Macau, we compare it against a baseline single-task feature selection approach. Namely, we use Elastic Net to learn one regression model per gene separately and select relevant individual cell line features that are predictive of dependency (see Figure 1B).

2.2.1 Feature learning problem formulation

For the regression tasks, we define an input matrix $X \in \mathbb{R}^{N \times M}$, where N corresponds to the number of cancer cell lines and M to the number of transcriptomic features (independent variables). Each $(n, m)^{th}$ entry contains the expression level of gene m in cell line n. We also define

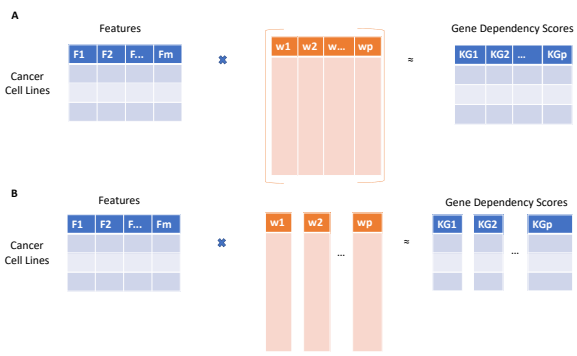

Fig. 1. Schematic representation of (A) multi-task learning and (B) single-task learning approaches to predict dependency scores across cancer cell lines for multiple genes, based on cancer cell line features. The prediction of dependency scores for each gene is a regression task. In multi-task learning, all regression tasks are learned simultaneously. In single-task learning, each regression task is learned independently from the other tasks.

an output or target matrix $Y \in \mathbb{R}^{N \times P}$, where each $(n, p)^{th}$ entry corresponds to the dependency score of cell line n on gene p. We are interested in the association of the transcriptomic features in the columns of matrix X with the dependency scores matrix Y (dependent variables). Specifically, we consider the prediction of the dependency scores for each gene p as a regression task, with the dependency scores vector $Y_p \in \mathbb{R}^N$ as the target variable and feature matrix X as the independent variables. Note that feature matrix X is the same for all regression tasks. The goal is then to obtain a vector of regression weights $w_p \in \mathbb{R}^M$, for each of these tasks. This weight vector reflects the association of each transcriptomic feature with dependency on gene p.

2.2.2 Macau multi-task feature learning

The Macau method learns the regression models jointly for all genes, allowing for information to be shared across these tasks. This is carried out by finding a lower dimensional latent space with K dimensions. In this latent space, genes and cancer cell lines are projected such that the dependency score of a cancer cell line on a gene is obtained by the dot product of the corresponding cancer cell line and gene latent vectors. Macau estimates a cell line latent matrix $U \in \mathbb{R}^{N \times K}$ and a gene latent matrix $V \in \mathbb{R}^{K \times P}$ in order to approximate the dependency scores matrix $Y \in \mathbb{R}^{N \times P}$:

$$Y \approx UV \tag{1}$$

where each row U_n in matrix U is the lower dimensional vector that represents cancer cell line n in the latent space and each column V_p in matrix V is the latent vector for gene p. The accuracy of the prediction can potentially be improved by incorporating side information that characterizes the cancer cell lines (i.e. feature matrix X) in the estimation of the latent spaces. In our application this means that Macau learns a link matrix $\beta_U \in \mathbb{R}^{M \times K}$ to predict the cancer cell line latent matrix U from the feature matrix X:

$$Y \approx UV \approx X\beta_U V \tag{2}$$

The latent and link matrices are estimated using Gibbs sampling (see Supplementary Methods 2.1 for more details). Additionally, we refer to matrix $\beta_U V \in \mathbb{R}^{M \times P}$ as the interaction matrix. This is essentially a matrix of regression coefficients associating every cancer cell line feature with dependency on every gene (see Figure 2). For example, the $(i, j)^{th}$

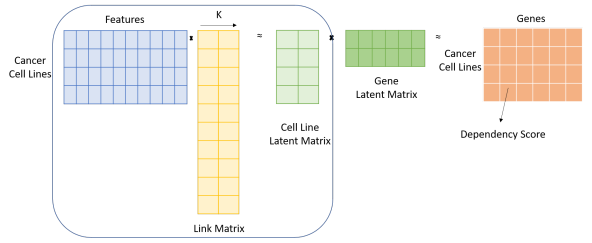

Fig. 2. Schematic representation of the Macau factorization model. The matrix of dependency scores is decomposed into two latent matrices, which aggregate cancer cell lines and genes into lower dimensional spaces. The cell line latent matrix results from multiplying the side information (cancer cell line features) and the link matrix. Multiplying the link matrix by the gene latent matrix yields the Interaction Matrix, which is essentially a regression weights matrix that associates every cancer cell line feature with dependency on every gene.

entry corresponds to the strength of the association between transcriptomic feature i and dependency on gene j, across all latent dimensions.

We used the Macau python package (v0.5.2) to apply this method to our data. Dependency scores were normalized per gene to zero mean and unit standard deviation, which was shown to improve performance (median increase 0.025, paired Wilcoxon test p-value < 0.01). Moreover, it avoided that the optimization focused on differences in dependency score means rather than variation across cell lines. Transcriptomic features were mean-centered. Other normalization methods were investigated but did not substantially improve prediction performance (see Supplementary Table 1). As for the remaining parameters, we relied on the same number of burn-in and collected samples as in (17) (400 burn-in, 600 collected samples). Each entry of the estimated matrices corresponds to its average across all the collected samples. We confirmed that the optimization converged (see Supplementary Figure 15). The number of latent dimensions K was set to 30, which was approximately the number of cancer types found in the dependency data.

2.2.3 Elastic Net single-task feature selection (baseline)

As a baseline for prediction performance, we fitted an Elastic Net regression model (22) for each gene. We chose this technique since it relies on a linear model and has the ability to select a relatively small set of predictive features, even in high dimensional feature spaces. The Elastic Net model is a linear regression model with two additional regularization terms based on the L1-norm and L2-norm. The former promotes sparsity in the coefficients assigned to the features, while the latter allows for correlated features to be selected simultaneously. The influence of the regularization terms is controlled by the λ parameter and the relative importance of the L1 and L2 penalties is controlled by another parameter α. For each gene p, we removed the missing values in Y_p and fitted the Elastic Net regression model based on the normalized X matrix. Each feature vector was normalized to zero mean and unit standard deviation. The optimal values for α and λ were chosen by minimizing the root mean squared error. We tried 10 values for $\alpha \in [0.1, 1]$ and noticed that prediction performance did not depend significantly on the value of α, provided that λ was properly tuned. Therefore, we set α to 0.2 and used a nested 10-fold cross validation to evaluate the model. In each outer loop, we performed an inner 10-fold cross-validation on the outer fold training data to tune λ. To avoid overfitting (see Supplementary Results 3.3), the optimal λ was set to the largest value of λ that led to an error between the minimum error and minimum error + 1 standard error in inner test folds. The estimated optimal value was then used to train an Elastic Net model on the outer fold training data. The resulting model was then tested in the outer fold test data. Models were fitted using the glmnet R package (v2.0.16)(29).

2.3 Selection of cancer dependencies for the learning task

Our initial exploration of the dependency data indicated that most of the genes were not essential for any cell line (see Supplementary Results 3.2). This means they had a dependency score close to 0 for all cell lines, within some experimental variation. These genes are of little biological and technical interest. We therefore sought to select genes whose dependency profiles showed interesting variation and potential relationships of dependency. In particular, we selected genes for which at least two cell lines showed dependency scores that were at least 6 median absolute deviation (MAD) units away from the median gene dependency across all cell lines. We used this deviation as an indication that the cell lines with more extreme values could be dependent on that gene. We used a similar approach to identify genes with more extreme values of dependency relative to other genes in the context of every cell line. Previous work focused on the former selection approach of comparing dependency scores for a given gene across cell lines with distinct molecular features (see Supplementary Methods 2.2). Here, we also employ the latter selection to be able to select genes that show more extreme dependencies within a cell line, therefore within the same molecular context. These could represent potential dependencies for such cell line, without necessarily being flagged as such using the gene-centric selection approach. Using these two selection procedures, we identified 2101 genes for which we then built learning models.

2.4 Evaluation

2.4.1 Dependency prediction performance

We used a 10-fold cross validation setup to evaluate the prediction performance of Macau and Elastic Net in section 3.1. In each fold, the features and dependency scores belonging to the cell lines in the test set were held-out and the models were trained using the remaining cell lines. We measured performance by calculating the Spearman's correlation between predicted and observed dependency scores. This measure of performance has been previously used to evaluate other dependency score prediction models (16). Of note, the remaining sections refer to Macau and Elastic Net trained using all of the cell lines.

2.4.2 Enrichment of latent factors with biological pathways

The latent factors (i.e. latent dimensions) learned by Macau capture hidden sources of variation in dependency scores. These factors can be interpreted by linking each factor back to the original data and using the associated coefficients to reflect the relevance of each gene/cancer cell line feature within each factor.

To investigate whether latent factors were associated with known biological processes or pathways, we performed pathway enrichment using Gene Set Enrichment Analysis (GSEA) (30) and Overrepresentation Analysis (OR). The idea of these methods is to investigate whether the most relevant genes or features in a factor belong to the same pathways, more than what would be expected by chance. GSEA takes a ranked gene list as input and calculates an enrichment score for each pathway. This score assesses to what extent genes belonging to such pathway consistently fall in the extreme values of the ranked list. This score is further normalized to account for differences in pathway sizes and its statistical significance is computed using permutation tests. OR takes as input a gene list of interest and uses a hypergeometric test to assess whether the ratio of genes that belong to a given pathway is significantly higher in the gene list compared to a background population.

In the case of the latent gene factors, we used GSEA to compute enrichment with Reactome and KEGG pathways. For each factor, we

ranked the genes based on the Macau-learned coefficients. The ranked list was used as input in GSEA to calculate pathway enrichment significance with 1000 permutations. For feature factors in the link matrix, GSEA delivered only one enriched factor, possibly due to the high dimensionality of the feature space. Therefore, we performed OR analysis on the list of genes corresponding to the 100 most relevant transcriptomic features for each factor, that is, features with the top 50 and bottom 50 link coefficients.

We used the R packages clusterProfiler (v3.10.1) (31) and ReactomePA (v1.26.0) (27) to perform both GSEA and OR analyses using KEGG and Reactome databases respectively. All computed p-values were corrected for multiple testing using the Benjamini-Hochberg procedure. Adjusted p-values smaller than 0.05 were considered significant. The background used for OR analysis comprised all protein coding genes corresponding to the cancer cell line features. Gene names were mapped from HGNC and ENSEMBL to Entrez identifiers using a combination of the mapping functions in clusterProfiler and the R package BiomaRt (28)(v2.38.0).

3 Results and Discussion

We used the Macau multi-task feature learning method to identify latent factors of genes and cell lines contributing to dependency of cancer cell lines on multiple genes. Macau models were built for a selection of genes showing stronger relationships of dependency for at least some of the cancer cell lines, as described in subsection 2.3.

3.1 Predictive performance of Macau shows less variance across genes

We first sought to validate the prediction performance of Macau against baseline approaches. In this context, Elastic Net was used to perform conventional single-task feature selection separately for each gene in the aforementioned gene selection. We compared the performances of Macau and Elastic Net based on the Spearman's correlation between predicted and observed dependency scores obtained in a 10-fold cross-validation setting.

Elastic Net did not find any significant predictors for approximately 40% of the genes. In these cases, the predictions of Elastic Net models defaulted to the mean of the gene dependency scores. On the other hand, Macau was able to predict these genes significantly better than genes with random dependency scores (Wilcoxon test p-value <0.01), with a median performance accuracy of 0.15 (Figure 3 A). This shows that Macau is able to predict dependencies on more genes than Elastic Net, due to its multi-task nature. For the remaining genes, the performance of Macau when trained on all genes was slightly worse than that of Elastic Net (median difference 0.01, paired Wilcoxon test p-value <0.01). However this is not a completely fair comparison, as Macau is simultaneously modelling genes with significant Elastic Net models and genes that are very difficult to predict. Indeed, the performance of Macau is not significantly different from that of Elastic Net (median difference 0.005, paired Wilcoxon test pval$>$0.01), when it is trained on the subset composed of the genes with significant Elastic Net models only.

Of note, the prediction performance of Macau showed smaller variance across genes than that of Elastic Net. This could be due to the ability of Macau to share information across regression tasks to yield reasonable predictions for multiple genes simultaneously. The shared context can benefit regression tasks for genes whose dependencies are more difficult to predict with Elastic Net. On the other hand, Macau is less suitable to identify associations that are specific only to a very small portion of the regression tasks, which could explain the decrease in performance for genes that were well predicted by Elastic Net (Figure 3 B). This suggests that borrowing information across tasks in multi-task learning can improve prediction. However, some negative transfer of information across non-sufficiently related tasks can occur, an issue that has been frequently reported in multi-task learning literature (32). Increasing the number of latent dimensions when training Macau could combat this issue, as information would be compressed in a larger latent space, allowing for more specific associations to be identified. We decided not to perform extensive parameter tuning of Macau, since our main goal was to identify and assess the relevance of latent factors rather than maximally improving prediction performance. In this context, we found that the performance of Macau using 30 latent factors was sufficiently comparable to that of Elastic Net.

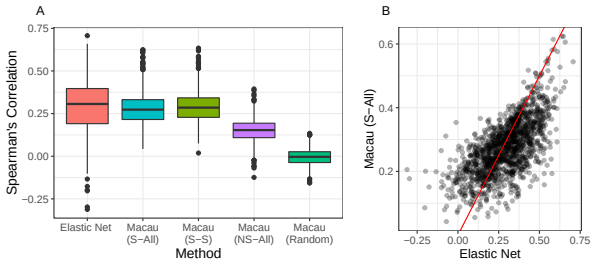

Fig. 3. Comparison of the prediction performance between Elastic Net and Macau, using average Spearman's correlation between predicted and observed dependency scores across 10 cross-validation folds. Note that correlation is only defined for genes for which Elastic Net was able to find significant predictors. We denote this subset of genes by 'S', and the remaining genes by 'NS'. (A) Overall performance for S genes obtained by Elastic Net, Macau trained on all genes (Macau (S-All)) and Macau trained on significant genes only (Macau (S-S)). For comparison, we show predictive accuracy for non significant genes, obtained by Macau trained on all genes (Macau (NS-All)). Macau (Random) corresponds to training Macau after breaking the relationship between dependency scores and cell lines i.e randomly permuting the dependency scores across cell lines (100 repetitions). (B) Comparison of per-gene predictive performance between Elastic Net and Macau trained on all genes.

3.2 Macau identifies features based on shared patterns

We investigated to what extent the cancer cell line features in the latent space captured by Macau differ from the features selected by Elastic Net. For this purpose, we compared the features showing the strongest associations with gene dependency, based on the regression/interaction weights yielded by Elastic Net and Macau. Of note, Macau does not enforce sparsity on the weights like Elastic Net. Therefore, for Macau we considered only a subset of features per gene, namely those with the 10 highest and 10 lowest interaction weights.

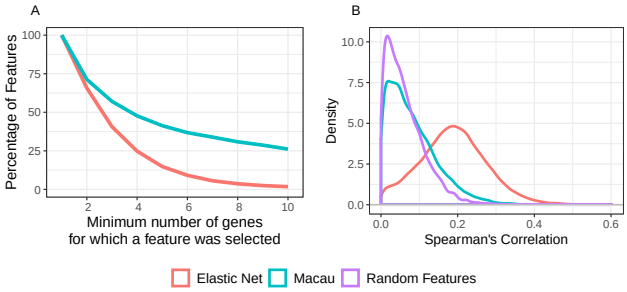

Fig. 4. Comparison of top features in gene dependency models built using Macau and Elastic Net. (A) Percentage of top features selected for several genes simultaneously. (B) Absolute Spearman's correlation between the expression profile of the top selected features and the dependency profile of the gene for which the feature was highly ranked. The curve for random features was obtained by randomly sampling 20 features for each gene dependency profile.

The transcriptomic features that were more strongly associated with gene dependency differed substantially between Macau and Elastic Net. In fact, the two methods found at least one common feature for only approximately 40% of the genes. One reason for this could be that Macau finds models and features that explain multiple gene dependencies, while Elastic Net focuses on finding the best feature associations to predict dependencies for each gene independently. Our analyses indeed showed that approximately 25% of the top features found by Macau across all genes were shared by more than 10 of corresponding models, while this percentage was close to 0% for features selected by Elastic Net (Figure 4 A). Another consequence of the multi-task approach to learn models jointly for multiple genes is that the top features selected by Macau for each gene were less globally correlated with the corresponding gene dependency profile than the top features selected by Elastic Net (Figure 4 B). This additionally suggests that Macau explores local associations during the factorization of the dependency data to support its gene dependency predictions. Detecting these kinds of associations is also made possible as a result of the increased statistical power gained when modelling dependencies for multiple genes simultaneously.

3.3 Macau latent spaces expose biological relationships

Macau projects genes, cancer cell lines and transcriptomic features into lower dimensional spaces, which could be informative to predict gene dependency. We investigated whether the sources of variability captured by these latent spaces can reveal relevant biological insight.

3.3.1 Latent gene space captures functional gene similarity

Previous work has shown that two functionally related genes, which are for instance involved in the same biological pathway or process, are likely to yield more similar dependency profiles than two functionally unrelated genes (10; 11). Here, we wondered to what extent Macau would be able to capture relationships of functional similarity through the coefficients learned for genes in either the latent gene space, or the interaction matrix. To investigate this, we calculated all pairwise Pearson's correlations between gene vectors. We then analyzed the distributions of pairwise correlations obtained for genes known or not known to be involved in a similar biological function, using data from the CORUM complex database. We visualized them in Figure 5 and quantitatively assessed the difference between the two distributions using Wilcoxon tests (p-value<0.01 in all cases, see Supplementary Table 2 for effect sizes).

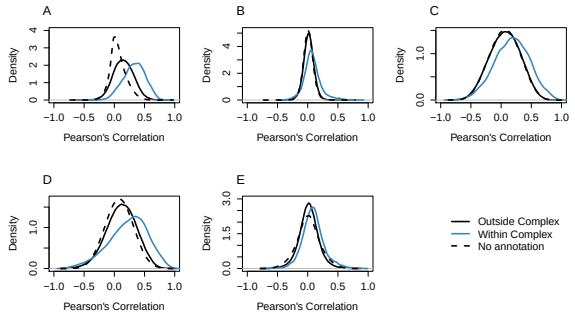

Fig. 5. Comparison of Pearson Correlation between pairs of genes within the same group (i.e. CORUM protein complex) and from different groups, using: (A) gene expression profiles, (B) gene dependency scores profiles, (C) gene latent weights learned by Macau trained using dependency scores only (no cell line features), (D) interaction weights learned by Macau using both dependency scores and cell line features as input (E) Elastic Net coefficients without sparsity regularization. No annotation corresponds to genes that could not be matched to any protein complex in the CORUM database.

We first looked into the functional information inherently contained in the dependency scores and cancer cell line features used to build Macau and Elastic Net models. For consistency, we focused on dependency scores and features associated with predicted genes only. Our results showed that cancer cell line features, measuring gene expression, clearly denote functional information between genes (Figure 5 A). On the other hand, dependency scores showed only very small differences between functionally related versus unrelated pairs of genes (Figure 5 B). To assess how Macau leverages cell line feature data to learn the latent spaces, we trained Macau using only dependency scores and no cell line feature data. Interestingly, the latent coefficients learned by Macau in this setting already improved the ability to capture functional relationships relative to dependency scores (Figure 5 C). This shows that Macau can deal with noise in the dependency profiles to find patterns of similarity underlying cancer genetic dependencies. The distinction between functionally related and unrelated genes was improved further based on the interaction weights learned by Macau using both dependency and cell line feature data (Figure 5D). This provides additional evidence that Macau is able to use cancer cell line data to learn better latent spaces. Lastly, regression coefficients learned by Elastic Net do not reflect functional relationships between genes substantially better than dependency scores (Figure 5E). This is not surprising as detecting these relationships is difficult if no information is shared across regression tasks. Of note, results were similar using KEGG and Reactome pathways instead of CORUM protein complexes (see Supplementary Figures 16 and 17). Taken together, these observations show that the latent spaces found by Macau can effectively extract functional relationships between genes based on the dependency scores and cell line feature data. What is more, the relationships are discovered automatically and are therefore latent in the dependency data, rather than imposed by prior knowledge.

3.3.2 Latent cell line space exposes tissue-based cell similarity

Several studies have identified cancer type-specific dependencies (4; 5). We set out to investigate to what extent the latent spaces learned by Macau would be able to capture similarity between cancer types. To do so, we used t-SNE (33) to reduce the 30 latent factors into 2 dimensions. This technique aims at preserving local neighborhood, so that similar cell lines in the original high-dimensional space correspond to neighboring points in the 2-dimensional representation and distant cell lines remain distant. Figure 6 shows how cell lines belonging to the same cancer type group together in this low-dimensional representation, based on the data given as input to Macau (gene expression feature matrix) and the latent spaces estimated by Macau trained with and without the feature matrix.

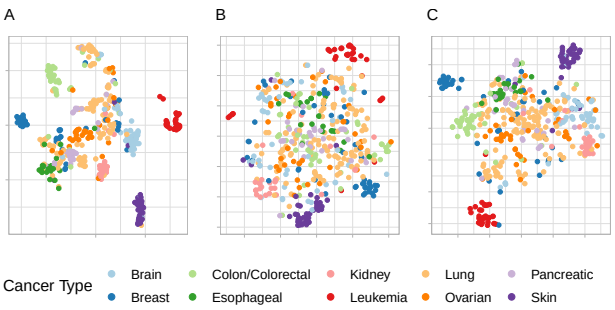

Fig. 6. t-SNE representations of cell lines based on: (A) PCA of cancer cell line gene expression profiles, (B) cell line latent matrix obtained with Macau trained on dependency scores only (no cell line features), and (C) cell line latent matrix obtained with Macau trained on both dependency scores and cell line features. In these plots, we used a perplexity value=15, but qualitative observations were similar for higher perplexity (=50).

Tissue-specific groupings were much stronger in cell line features (expression) compared to cell line dependency scores (Figures 6A and 6B). This was not surprising, since gene expression is frequently used to classify cancer types (34). Using the cell line latent space obtained via Macau with dependency scores alone (no cell line features), only four tissue-specific types of cell lines clearly formed separate clusters: leukemia, breast, skin and kidney (Figure 6B). Similarly to what we observed with functional similarities between genes, the tissue-specific clusters became more defined when the Macau model was trained using both dependency scores and cell line features (Figure 6C). We also observed that some cancer types that are scattered across the latent space are actually clustered according to cancer sub type (see Supplementary Figure 18). These observations indicate that the incorporation of cell line molecular features as side information produces more meaningful latent spaces, indicative of tissue type in this case. The use of such latent spaces by Macau could potentially lead to the identification of relationships between cell line features and dependency mainly driven by tissue type rather than specific biological processes or functions, as we would like. In that case, the strongest associations could correspond to cell line features that were a proxy of cancer type rather than related with dependency on those genes. To verify this hypothesis, we considered performing tissue-specific analyses, but abandoned the idea since this would drastically reduce the number of samples available for training the Macau models.

3.3.3 Latent factors are associated with relevant biological processes

We investigated whether the latent factors learned by Macau could be attributed a biological interpretation. To achieve this, we performed enrichment analysis based on the weights of genes and cancer cell line features in each factor, as described in subsection 2.4.2.

Nineteen out of the thirty latent gene factors were significantly enriched. The biological processes associated with those latent factors were mainly related with translation, DNA replication or cell cycle, and mitochondrial processes (see Figure 7). Aberrant translation and cell cycle activity are known cancer hallmarks, and mitochondrial metabolism has also been linked to several oncogenic processes, such as regulation of reactive oxygen species and cell death. Therapies specifically targeting these processes are currently being explored (35; 36; 37). Other cancer-related processes were also found enriched in the latent factors, even if not included in the top pathways shown here (see Supplementary Figure 19). Namely, immune system response and cancer-related signalling pathways such as TCR, NOTCH4, TGF-beta, WNT and MAPK signalling, whose role in sustaining cancer growth has been extensively studied (1). This suggests that the latent factors denote relevant biological processes which cancer cells are dependent on.

In the cell line feature space, twenty out of the thirty latent factors were significantly enriched. The main biological processes captured in those factors were related with cellular senescence, cell cycle, cancer signalling pathways and extra-cellular matrix, including glycosaminoglycan metabolism and collagen related processes (Figure 8). These processes are known key drivers of cancer development and proliferation across several tissues (38; 1; 39).

We also investigated whether latent factors learned by Macau captured general gene dependency trends. We observed that factor 23 significantly correlates with cell line mean dependency scores (Pearson correlation coefficient 0.46, see Supplementary Figure 20A). This suggests that this factor captures general sensitivity of cell lines to gene perturbations. A similar phenomenon has also been observed when analyzing drug response in cancer cells using matrix decomposition (40). In line with this, factor 23 is enriched in Reactome pathways related to Insulin-like Growth Factor and KEGG pathways associated with extracellular matrix organization. Both of these processes are involved in cancer growth and development (38; 41). In addition, a technical confounder in shRNA-based screens (AGO2 gene expression) is correlated with factor 7 (Pearson correlation coefficient -0.29, see Supplementary Figure 20B). Capturing general gene dependency trends in some factors can help to expose more specific trends in other factors.

3.3.4 Latent factors relate pathway alterations to cancer dependencies

The separate analysis of the projections of genes and cancer cell line features shows that the latent factors are capturing relevant biological processes. Therefore, we investigated next whether these latent factors could also uncover associations between pathway alterations and dependency on functionally related genes. We inspected each of the latent

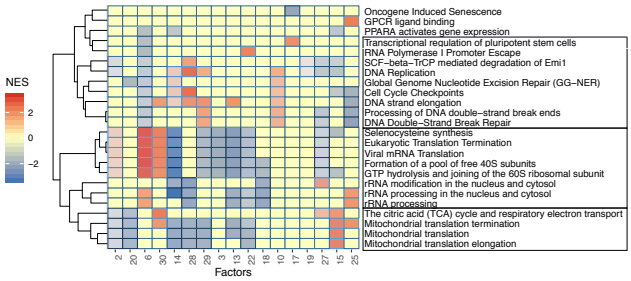

Fig. 7. Normalized Enrichment Score (NES) for the most highly enriched Reactome Pathways in the latent factors, based on the gene weights in each factor. For this heatmap, we focused on Reactome pathways only and selected 2 pathways per factor, corresponding to the ones with the most consistent behavior in the positive and negative tail of the latent factor. This corresponds to the pathways with the highest (positive) NES and lowest (negative) NES. Yellow cells indicate that pathways were not significantly enriched in the corresponding factor. Higher transparency indicates higher p-value. Factors and pathways are ordered based on hierarchical clustering. The boxes indicate the three main groups of pathways that are enriched in these factors: Cell Cycle, Translation and Mitochondria related, respectively.

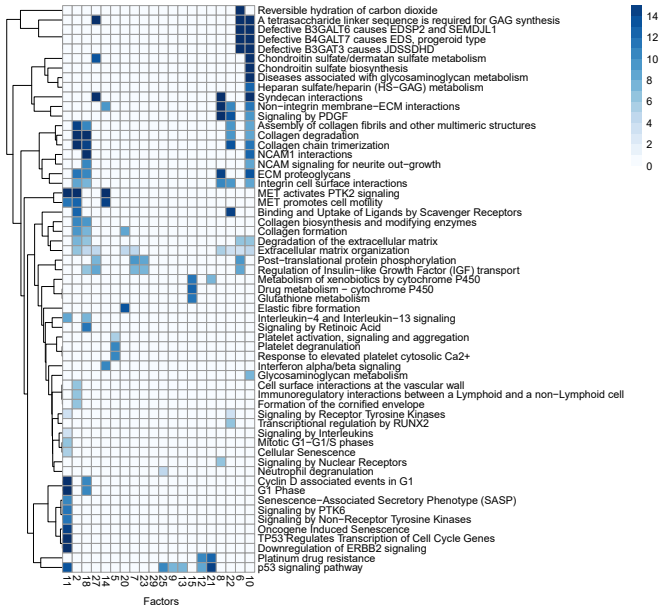

Fig. 8. Enriched Reactome pathways in the latent factors, based on the cell line feature weights. Colors indicate the effect size of the enrichment, which corresponds to the ratio of cell line features in the 100 most extreme weights to the background ratio. For factors that were not enriched in any Reactome pathways, we display the top 3 KEGG pathways with the highest effect size. White cells indicate pathways that were not significantly enriched in the corresponding factor. Factors and pathways are ordered based on hierarchical clustering.

factors that were enriched in both genes and cell line features (total of 14). Due to time constraints, we focused on the top enriched pathways and performed a biological literature search to find evidence of their relationship. For several latent factors, the top enriched pathways represent quite large and broad pathways, such as extracellular matrix organization or rRNA processing, which makes it harder to assess the biological relevance of the associations found. Even so, factor 15 shows a source of variation across genes that is likely related with differences in metabolic needs across different cell lines. Namely, the top enriched pathways for latent genes "PPARA activates gene expression" and "TCA cycle" are related with lipid metabolism regulation and mitochondrial energy production, respectively. The enriched pathways based on cancer cell line features are "drug metabolism by Cytochrome P450" and "Glutathione metabolism". Indeed, cytochrome P450 enzymes play a role in fatty acid oxidation (42) and the glutathione redox system has been found to be essential when fatty oxidation is inhibited in renal carcinomas (43). Interestingly, this latent factor also clearly distinguishes blood-related cancers (Leukemia, Lymphoma and Myeloma) from breast, uterine and esophageal cancers (see Supplementary Figures 21 and 22). In line with this, there is considerable evidence of metabolic reprogramming in leukemia cells that makes them more dependent on fatty acid oxidation for their proliferation (44). While fatty acid oxidation has also been linked to increased proliferation and chemoresistance of breast cancer cells (45), a recent study also showed that fatty acid oxidation was downregulated in breast cancers and that its activation could lead to lower cancer proliferation (46).

This example shows how Macau latent factors can be explored to suggest relationships between pathway alterations in cancer cells and increased dependency on functionally related genes.

4 Conclusions

We showed that Macau, a multi-task matrix factorization approach, leverages similarities between cancer cell lines and between genes to learn latent spaces that explain dependency of cancer cell lines on functionally related genes. These spaces are further linked to cancer cell line transcriptomic features, which allows the identification of biological processes associated with cancer genetic dependencies. We showed preliminary evidence that this latent space can capture relevant biological processes without the need of enforcing prior knowledge of pathway structure. However, there are some limitations in our work.

Namely, the number of latent dimensions should be more comprehensively explored. We initially decided to use 30 latent dimensions, because this approximately matches the number of cancer types present in the dependency data and yields comparable performance to Elastic Net. However, 30 latent dimensions might be limited to capture relevant processes and functional relationships especially amongst $\sim$19k cancer cell line features. Increasing the number of latent dimensions in the feature space could disentangle biological processes that are currently being conflated into the same latent dimensions. In addition, compressing the data over multiple latent spaces of different dimensions could complement our current analysis, by revealing latent characteristics at several levels of abstraction (47).

Other aspects could improve this work, if additional data were available. For example, more therapeutically relevant insights could be gained if the dependency data also contained data for normal cell lines. This could help to identify dependencies that are specific to cancer cells, potentially leading to the development of cancer therapies that are less toxic for healthy cells. In addition, having more cancer cell lines per tissue could potentially allow a tissue-specific analysis. This type of analysis could help controlling for the risk of finding predictive features that are mere proxies for cancer type. Of note, we experimented training Macau on breast and leukemia cell lines only, but the massive reduction of the number of cell lines reduces prediction performance and the ability of Macau to leverage functional relationships between genes. Lastly, Macau directly decomposes the dependency data but not the cell line feature matrix. This ensures that the latent factors learned by Macau are indeed associated with gene dependency. On the other hand, this also means that Macau does not fully leverage the intrinsic relationships between cell line transcriptomic features. Both goals could be achieved if both gene dependency and transcriptomic information were encoded in the matrix being decomposed. Datasets with transcriptomic measurements before and after shRNA perturbations exist (48), but not at large scale yet. Macau could possibly be used to learn meaningful latent spaces in this type of data, especially once they become available at a larger scale.

Despite these limitations, our work shows potential in the application of Macau to analyze cancer genetic dependencies. For example, similarities between functionally related genes are more evident in the latent space learned by Macau than in the original dependency data. Therefore, clustering in this latent space can potentially produce more biologically meaningful clusters of genetic dependencies than previous studies (10; 11). More importantly, the latent space learned by Macau can directly associate gene clusters to cancer cell line transcriptomic signatures that are predictive of dependency. In contrast, previous studies have to rely on thousands of statistical tests to associate the clusters with cell line features.

A general limitation of gene dependency prediction is that the number of cell lines screened is very low compared to the size of the feature space. Although there are several publicly available loss-of-function datasets, their integration is difficult due to technical effects. One interesting ability of Macau is that it can perform tensor factorization. That is, it can factorize multiple matrices that relate the same entities - in our case genes and cell lines. Therefore, we could extend our approach to simultaneously factorize dependency measurements from multiple loss-of-function screens. This could lead to finding more robust latent factors associated with gene dependency and indirectly increase the number of samples available for training.

References

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

[2]Beijersbergen, R. L., Wessels, L. F. & Bernards, R. Synthetic lethality in cancer therapeutics. *Annual Review of Cancer Biology* **1**, 141–161 (2017).

[3]Rancati, G., Moffat, J., Typas, A. & Pavelka, N. Emerging and evolving concepts in gene essentiality. *Nature Reviews Genetics* **19**, 34–49 (2017).

[4]Tsherniak, A. *et al.* Defining a cancer dependency map. *Cell* **170**, 564–576 (2017).

[5]McDonald, E. R. *et al.* Project drive: A compendium of cancer dependencies and synthetic lethal relationships uncovered by large-scale, deep rnai screening. *Cell* **170**, 577 – 592.e10 (2017).

[6]Marcotte, R. *et al.* Functional genomic landscape of human breast cancer drivers, vulnerabilities, and resistance. *Cell* **164**, 293–309 (2016).

[7]Behan, F. M. *et al.* Prioritization of cancer therapeutic targets using crispr–cas9 screens. *Nature* **568**, 511 (2019).

[8]Yamauchi, T. *et al.* Genome-wide crispr-cas9 screen identifies leukemia-specific dependence on a pre-mrna metabolic pathway regulated by dcps. *Cancer Cell* **33**, 386 – 400.e5 (2018).

[9]Tang, Y. C. *et al.* Functional genomics identifies specific vulnerabilities in pten-deficient breast cancer. *Breast Cancer Research*

20, 22 (2018).

[10]Kim, E., Dede, M., Lenoir, W. F., Wang, G., Srinivasan, S., Colic, M. & Hart, T. A network of human functional gene interactions from knockout fitness screens in cancer cells. *Life Science Alliance* **2** (2019).

[11]Pan, J. *et al.* Interrogation of mammalian protein complex structure, function, and membership using genome-scale fitness screens. *Cell Systems* **6**, 555–568 (2018).

[12]Boyle, E. A., Pritchard, J. K. & Greenleaf, W. J. High-resolution mapping of cancer cell networks using co-functional interactions. *Molecular Systems Biology* **14** (2018).

[13]Zhang, Y. & Yang, Q. A survey on multi-task learning. *CoRR* **abs/1707.08114** (2017).

[14]Costello, J. C. *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology* **32**, 1202 (2014).

[15]Yuan, H., Paskov, I., Paskov, H., González, A. J. & Leslie, C. S. Multitask learning improves prediction of cancer drug sensitivity. *Scientific Reports* **6**, 31619 (2016).

[16]Gönen, M. *et al.* A community challenge for inferring genetic predictors of gene essentialities through analysis of a functional screen of cancer cell lines. *Cell Systems* **5**, 485 – 497.e3 (2017).

[17]Yang, M., Simm, J., Lam, C. C., Zakeri, P., van Westen, G. J., Moreau, Y. & Saez-Rodriguez, J. Linking drug target and pathway activation for effective therapy using multi-task learning. *Scientific Reports* **8** (2018).

[18]Suphavilai, C., Bertrand, D. & Nagarajan, N. Predicting Cancer Drug Response using a Recommender System. *Bioinformatics* **34**, 3907–3914 (2018).

[19]Ammad-ud din, M., Khan, S. A., Malani, D., Murumägi, A., Kallioniemi, O., Aittokallio, T. & Kaski, S. Drug response prediction by inferring pathway-response associations with kernelized bayesian matrix factorization. *Bioinformatics* **32**, i455–i463 (2016).

[20]Knowles, D. A., Bouchard, G. & Plevritis, S. Sparse discriminative latent characteristics for predicting cancer drug sensitivity from genomic features. *PLoS Computational Biology* **15**, e1006743 (2019).

[21]Simm, J. *et al.* Macau: Scalable bayesian factorization with high-dimensional side information using mcmc. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6 (IEEE, 2017).

[22]Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: series B (statistical methodology)* **67**, 301–320 (2005).

[23]McFarland, J. M. *et al.* Improved estimation of cancer dependencies from large-scale rnai screens using model-based normalization and data integration. *Nature Communications* **9** (2018).

[24]Barretina, J. *et al.* The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603 (2012).

[25]Giurgiu, M. *et al.* Corum: the comprehensive resource of mammalian protein complexes 2019. *Nucleic Acids Research* **47**, D559–D563 (2018).

[26]Kanehisa, M. & Goto, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**, 27–30 (2000).

[27]Yu, G. & He, Q.-Y. Reactomepa: An r/bioconductor package for reactome pathway analysis and visualization. *Mol. BioSyst.* **12** (2015).

[28]Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A. & Huber, W. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).

[29]Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22 (2010).

[30]Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).

[31]Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics : a Journal of Integrative Biology* **16**, 284–7 (2012).

[32]Mishra, M. & Huan, J. Multitask learning with feature selection for groups of related tasks. In *2013 IEEE 13th International Conference on Data Mining*, 1157–1162 (2013).

[33]Maaten, L. v. d. & Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).

[34]Roychowdhury, S. & Chinnaiyan, A. M. Translating cancer genomes and transcriptomes for precision oncology. *CA: a Cancer Journal for Clinicians* **66**, 75–88 (2016).

[35]Bhat, M., Robichaud, N., Hulea, L., Sonenberg, N., Pelletier, J. & Topisirovic, I. Targeting the translation machinery in cancer. *Nature Reviews Drug Discovery* **14**, 261 (2015).

[36]Otto, T. & Sicinski, P. Cell cycle proteins as promising targets in cancer therapy. *Nature Reviews Cancer* **17**, 93 (2017).

[37]Porporato, P. E., Filigheddu, N., Bravo-San Pedro, J. M., Kroemer, G. & Galluzzi, L. Mitochondrial metabolism and cancer. *Cell Research* **28**, 265 (2018).

[38]Walker, C., Mojares, E. & del Río Hernández, A. Role of extracellular matrix in development and cancer progression. *International Journal of Molecular Sciences* **19**, 3028 (2018).

[39]Afratis, N. *et al.* Glycosaminoglycans: key players in cancer cell biology and treatment. *The FEBS Journal* **279**, 1177–1197 (2012).

[40]Geeleher, P., Cox, N. J. & Huang, R. S. Cancer biomarker discovery is improved by accounting for variability in general levels of drug sensitivity in pre-clinical models. *Genome Biology* **17**, 190 (2016).

[41]Weroha, S. J. & Haluska, P. The insulin-like growth factor system in cancer. *Endocrinology and Metabolism Clinics* **41**, 335–350 (2012).

[42]Johnson, A. L., Edson, K. Z., Totah, R. A. & Rettie, A. E. Cytochrome p450 ω-hydroxylases in inflammation and cancer. In *Advances in Pharmacology*, vol. 74, 223–262 (Elsevier, 2015).

[43]Miess, H. *et al.* The glutathione redox system is essential to prevent ferroptosis caused by impaired lipid metabolism in clear cell renal cell carcinoma. *Oncogene* **37**, 5435 (2018).

[44]Testa, U., Labbaye, C., Castelli, G. & Pelosi, E. Oxidative stress and hypoxia in normal and leukemic stem cells. *Experimental Hematology* **44**, 540 – 560 (2016).

[45]Kuo, C.-Y. & Ann, D. K. When fats commit crimes: fatty acid metabolism, cancer stemness and therapeutic resistance. *Cancer Communications* **38**, 47 (2018).

[46]Aiderus, A., Black, M. A. & Dunbier, A. K. Fatty acid oxidation is associated with proliferation and prognosis in breast and other cancers. *BMC Cancer* **18**, 805 (2018).

[47]Way, G. P., Zietz, M., Himmelstein, D. S. & Greene, C. S. Sequential compression across latent space dimensions enhances gene expression signatures. *BioRxiv* 573782 (2019).

[48]Smith, I. *et al.* Evaluation of rnai and crispr technologies by large-scale gene expression profiling in the connectivity map. *PLoS Biology* **15**, e2003213 (2017).

Supplementary Material

This supplementary material complements the thesis article. It is composed of three main sections: extended background, methods and results. In the extended background, we provide additional information about how loss-of-function screens are performed. We then describe studies that carried out large-scale screens to study cancer genetic dependencies and how these screens have been analysed by other research groups. We conclude with some additional information about the main types of multi-task learning algorithms. In the extended methods section, we provide more details on how the Macau method works and how it compares to other related machine learning approaches. We also explain in more detail our approach to select potential cancer dependencies for the learning task. The extended results contain an initial exploration of the dependency data using Elastic Net models. We highlighted some limitations of this initial analysis, which motivated us to use a multi-task learning method to analyse the dependency data. The last section of the extended results contains supplementary figures and tables that support the main thesis article.

1 Extended Background

1.1 Loss-of-function screens

In this subsection, we describe how RNA interference (RNAi) and CRISPR-Cas9 loss-of-function screens perform targeted gene silencing. We also highlight off-target effects that can occur and key differences between these two types of screens. In addition, we focus on the method to remove shRNA off-target effects developed by McFarland et al [1], which was used to generate the gene dependency measurements that we analysed in this thesis.

1.1.1 RNA interference and CRISPR-Cas9 based screens

Cancer dependencies can be identified in an unbiased manner through large-scale loss-of-function screens. These screens assess the dependency of cell lines on thousands of genes by measuring how cell line's proliferation is affected after individually silencing each gene in each cell line. Gene silencing is commonly known as gene knockout or knockdown, depending on whether the expression of the gene is completely or only partially suppressed, respectively.

Different experimental techniques exist for perturbing individual gene expression in a large-scale. Currently, the most widely-used are RNA interference (RNAi) and CRISPR/Cas9-based screens.

RNAi screens make use of the RNAi machinery of the cells, which regulates gene expression, to reduce the abundance of mRNA of a target gene. These screens can be performed using different RNAi reagents, such as shRNA (short-hairpin RNA) or siRNAs (short-interfering RNAs). We focus here on shRNA-based screens, as these provide long-term gene silencing. Gene knockdown in siRNA-based screens, on the other hand, is only achieved during a short period of time [2].

In large-scale shRNA screens, thousands of shRNA plasmid vectors or lentiviral particles are used to infect cell lines, with each of them targeting an individual gene (see Figure 1 A). These plasmids/lentiviral particles contain shRNA-coding sequences that are integrated into the cell genome, leading to a stable and continued production of shRNAs. As seen in Figure 1 B, these shRNAs are then processed into siRNAs by the Dicer enzyme and loaded into a multiprotein complex known as RISC (RNA-induced silencing complex). Within this complex, the siRNA binds to complementary target mRNAs so that a nuclease in RISC can cleave and destruct the target mRNA. This ultimately results in the silencing of the expression of the corresponding target gene [2]. After infection, cell lines are left to grow for a given time period and at the end of this period, the abundance of each shRNA is measured using Next Generation Sequencing. In order to measure the effect of shRNA knockdown on cell proliferation, the final shRNA abundance is compared to its abundance in the initial conditions. shRNA depletion indicates that the inactivation of the target gene hampers cell proliferation, whereas shRNA enrichment indicates that the target gene inactivation actually promotes cell proliferation.

One of the challenges with RNAi-based loss-of-function screens is the generation of accurate quantitative measures of the phenotypic effect of a target gene knockdown [5, 1]. This is because although gene silencing

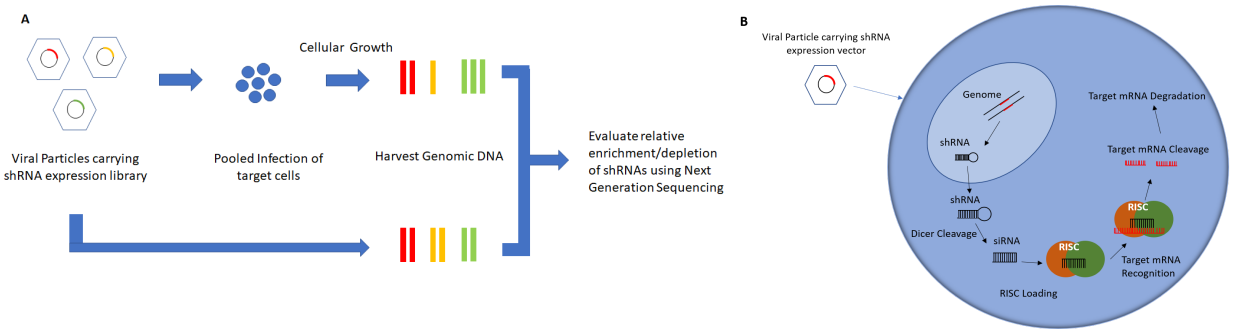

Figure 1: (A) Schematic representation of a pooled shRNA screen using lentiviral particles. The relative comparison of shRNA at the end of the experiment is interpreted as follows: silencing target gene of red shRNA does not affect cellular growth, while silencing target genes of orange and green shRNAs respectively impairs and promotes cellular growth. (B) Schematic representation of the mechanism that induces silencing of a target gene, after infection with a shRNA expression vector. Adapted from [3, 2, 4]

in shRNA screens is relatively sequence-specific, shRNAs can lead to the degradation of non-target mRNAs when their sequences are partly complementary. The extent of the repression of non-target transcripts will depend on the actual shRNA sequence. These off-target effects can be partially mitigated by using multiple shRNAs targeting the same gene [6]. In addition to these sequence-related effects, off-target effects can also occur due to the artificial introduction of shRNAs into the cell, as they can displace endogenous miRNAs present in the RISC complexes. This can also lead to an off-target phenotype, because it prevents endogenous miRNAs from repressing their transcripts [5].

CRISPR-Cas9-based screens are conceptually similar to RNAi screens, but gene silencing is achieved in a different manner. Namely, genes are silenced by targeted loss-of-function mutations that are introduced via the combined action of Cas9 nucleases and short single guide RNAs (sgRNAs). Put simply, both Cas9 and sgRNA are delivered into cells using a carrier (e.g a lentiviral vector). After their stable integration in the genome, sgRNAs recognize the target gene so that a Cas9 nuclease can induce site-specific DNA double-stranded breaks (DSBs). This activates a repair process that frequently leads to frame shift insertion/deletion (indel) mutations and consequently to the knockout of the target gene [7].

CRISPR-based screens are advocated by many to be superior to shRNA screens because they do not suffer from off-target effects so severely. Moreover, they cause complete gene activation, whereas shRNA-based gene knockdown is variable and typically incomplete. However, others argue that the approaches are actually complementary [8, 6]. This is because although CRISPR-based screens are expected to be more robust [9], the cell proliferation phenotype can also be altered due to sgRNAs matching off-target sites and other extrinsic factors. For example, DNA accessibility, expression of Cas9 nuclease and alterations in gene copy number [10, 11]. What is more, the incomplete knockdown in shRNA-based screens reflects more accurately the effects of drug inhibition, which can be explored to find more effective drug targets. This is because these screens allow the identification of genes whose partial inhibition can already lead to loss of viability of cancer cells [8].

1.1.2 Quantifying gene knockdown effects in shRNA screens - The DEMETER 2 (D2) method

Depletion measurements resulting from shRNA screens suffer from off-target, batch and other technical effects. Therefore, these measurements have to be processed in order to obtain more accurate estimates of the on-target knockdown effects in shRNA screens.

Several methods exist for isolating on-target effects in shRNA screens. However, we are only going to focus on the DEMETER2 method (D2), since this was the method that McFarland et al[1] employed to generate the dependency scores we used in this thesis. This method was shown to provide better estimates of on-target effects compared to previous methods [1]. Moreover, it enables analysis of shRNA depletion measurements obtained in different RNAi screens.

As shown in Figure 2, D2 models the depletion score of a shRNA in a given cell line, measured in a specific screen, mainly as a linear combination of two unobserved quantities: gene suppression and seed effects. The former corresponds to on-target effects, shared by shRNAs designed to target the same gene. The latter consist of off-target effects resulting from shRNAs having 7-mer seed sequences that are complementary to non-target genes. For each shRNA, only two seed sequences are considered, corresponding to positions 1-7 and 2-8 on the antisense strand. These sequences were chosen because grouping all of the shRNAs based on these sequences maximised the shRNA depletion intra-group correlation.

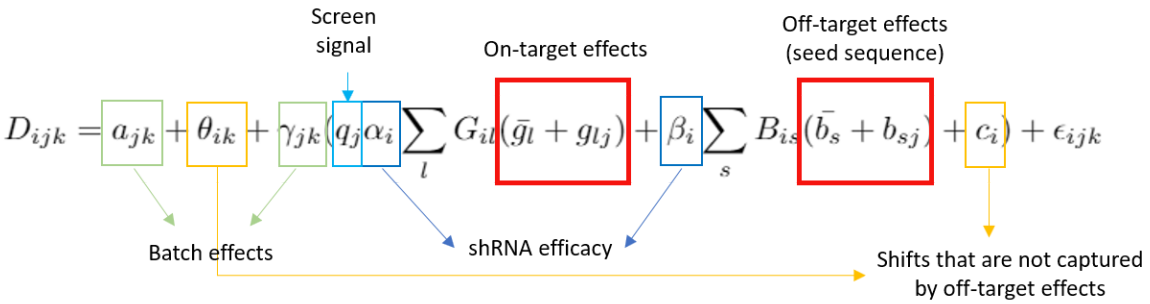

Figure 2: Modelling of the depletion score (D_{ijk}) of shRNA i in cell line j in dataset k, using the D2 method. l and s refer to the set of genes and seeds targeted by a given shRNA, respectively. G_{il} and B_{is} are the entries of fixed binary matrices, which indicate shRNA-gene and shRNA-seed mappings. See main text for the description of the remaining parameters.

Each of the gene and seed effects are in turn decomposed into an across cell-line effect component ($\bar{g}_l$ and $\bar{b}_s$) and a cell line specific component (g_{lj} and b_{sj}), for gene l and seed s. This allows for information sharing across cell lines as well as modelling cell-specific gene effects. In addition, differences in shRNA efficacy in eliciting on and off-target effects are modelled by parameters α_i and β_i, respectively.

D2 also addresses differences in screen quality between cell lines by including a cell-specific parameter (q_j). This parameter scales gene effects according to how well they can be captured using shRNA depletion scores. Addressing screen quality differences is important, as they can greatly bias downstream analyses. In order to integrate datasets from different screens, batch effects are also modelled explicitly using an offset parameter (α_{jk}) and a scale parameter (γ_{jk}).

A distinctive feature of D2 is that it produces absolute dependency scores, which makes it possible to directly compare gene knockdown effects across genes and across cell lines. In order to produce an absolute scale, D2 explicitly models systematic errors that affect shRNA depletion scores. This is accomplished by using a offset term (θ_{ik}) to account for errors in the initial measurement of shRNA abundance (screen and shRNA-specific) and another term (c_i) to account for non-seed-based off-target effects (shRNA-specific). As a result of these model choices, a zero score in D2 means that the cell line is not dependent on a particular gene.

Parameter estimation is carried out by alternating between estimating posterior distributions for intercept terms, gene and seed effects using Bayesian inference. The remaining terms are estimated using the maximum a posteriori (MAP) method (see [1] for more details). By employing bayesian inference, uncertainty estimates can be provided for gene effects, which can potentially improve the statistical power of downstream analyses.

D2 was shown to greatly improve the accuracy of gene dependency scores compared to other methods. Removal of screen effects was supported by solid evidence, namely through greater agreement with CRISPR screens [1]. Of note, the incorporation of screen normalization parameters led to accuracy improvements comparable to modelling off-target effects, although previous approaches have been mainly focusing on the latter.

1.2 Overview of large-scale cancer dependency datasets

The last 5 years have witnessed an increased interest in profiling cancer dependencies, with several groups performing CRISPR and shRNA-based screens. In this thesis, we analysed the dependency dataset generated by McFarland et al [1], which is currently the largest resource of gene dependency measurements obtained from shRNA-based screens. This dataset is the result of integrating three shRNA-based screens. In this subsection, we start by reviewing the three studies that generated each of these screens, as well as the main findings of their initial analyses on their data. In addition, we also review a recently released large scale CRISPR-based screen. Then we focus on relevant works that reflect how other research groups used these available data to characterize cancer genetic dependencies and to build predictive models of gene dependency.

Marcotte et al. [12] used a hierarchical mixed-effects regression model to analyse shRNA depletion scores on 77 breast cancer cell lines and identified both general and breast cancer subtype-specific dependencies. They also showed the potential of integrating gene dependency profiles with drug response profiles to find drug resistance mechanisms and suggest possible drug combinations.

In a larger scale, Project DRIVE [6] (Deep RNAi Interrogation of Viability Effects in cancer) analysed the dependency of cancer cell lines on approximately 8000 genes. They addressed the off-target effects characteristic of shRNA screens by substantially increasing the number of shRNAs per gene to 20. Moreover, they also

screened 398 cell lines belonging to different cancer types in order to have greater statistical power to detect cancer dependencies and their respective molecular drivers. They found several types of cancer dependencies, ranging from oncogenes and lineage-specific transcription factors to synthetic lethal relationships. The latter correspond to pairs of genes whose individual suppression is not lethal by itself, but only in combination.

The on-going project Achilles [13] developed the DEMETER metric to estimate gene dependency scores based on shRNA depletion scores and currently contains genome-scale shRNA screens of more than 500 cell lines. In the last version of the project, they identified more than 700 genetic dependencies and built a regression model per gene to find the molecular features that most contribute to dependency on each gene. Interestingly, they found that gene expression was overall a better indicator of gene dependency than DNA mutation or copy number. What is more, both the DRIVE and Achilles projects found that cancer dependencies can be grouped based on the type of features that are most relevant to predict dependency. Namely, they found that a large group of dependencies are well predicted by their own gene expression or copy number. Other dependencies are better predicted by the expression of other genes, such as synthetic lethal pairs. Collectively, these studies showed that common patterns across several genetic dependencies can be found by analysing hundreds of dependencies together.

CRISPR/Cas9 studies have also been performed to characterize cancer dependencies. Namely, a recent study performed a genome-wide CRISPR-Cas9 screen in 324 cell lines and developed a methodology to prioritise new drug targets [14]. Some of the found targets are known targets of anti-cancer drugs. However, the majority represent new targets, of which around half have literature support. This shows the potential of these screens to expand the arsenal of candidate drug targets, since anti-cancer drug development has been mainly focused on cancer mutated genes. However, the functional impact of genes on cancer viability, which is measured in loss-of-function screens, can be more informative of their therapeutic potential.

These screens represent powerful data resources that can contribute to a deeper understanding of the disease and potentiate the discovery of new targets. However, the use of these large scale datasets by other research groups has been rather limited. In fact, most studies only queried these datasets to validate the dependency of a given group of cell lines on a specific gene of interest. For example, Yamauchi et al [15] indicated the mRNA decapping enzyme DCPS as a potential target to treat Acute Myeloid Leukemia (AML), and corroborated their results with the dependency scores in AML cell lines in the Achilles cancer dependency dataset [13]. Tang et al [16] identified potential synthetic lethal genes in PTEN-deficient breast cancer cell lines based on their own loss-of-function screen and confirmed the reproducibility of their results using the Achilles dataset.

However, a more global analysis of these data can lead to new insights. Namely, some studies used large scale loss-of-function screens to infer gene function or protein interactions, based on the assumption that similarities in gene dependency profiles across cell lines can be a proxy for similarity in gene function or protein interaction. One of the first studies to do this was Wang et al [17]. In this study, they carried out a CRISPR/Cas9 screen on 14 acute myeloid leukemia cell lines and inferred the functions of two previously uncharacterized genes (C17orf89 and C1orf27). These functions were then experimentally validated. Building on this work, recent studies [18, 19, 20] generated gene networks based on pairwise correlations between gene dependency profiles obtained in large scale loss-of-function screens using hundreds of cell lines. Pan et al [18] derived functional gene networks from multiple loss-of-function screens in parallel and compared them with protein interaction and gene co-expression networks. They showed that structural and functional protein complexes can be recovered from these functional networks and, in some cases, these complexes cannot be recovered from gene co-expression networks. Two other studies [19, 20] independently built a gene network based on large scale CRISPR screens and showed that gene clusters in the network displayed functional coherence. Moreover, they investigated which cancer cell line features were driving dependency on members of the same gene cluster, by performing statistical tests per feature and per cluster. However, this approach can fail to identify weaker associations or feature combinations that are predictive of gene dependency.

Focusing on building better predictive models to find molecular drivers of cancer dependencies, a DREAM challenge was also carried out [21]. DREAM challenges aim at evaluating predictive models submitted by different research groups to identify which modelling strategies are the most effective for a biological problem. Over 3000 models were submitted by different research groups and systematically evaluated. The comparative analysis of the submitted models has led to several interesting observations. For example, using ensemble models and incorporating prior knowledge in the feature selection process increased gene dependency prediction accuracy. The challenge also showed that gene expression data was the most informative type of data for the predictive models. Furthermore, they found that prediction accuracy depends more on the gene being predicted than on the quality of the model. That is, some genes are considerably easier to predict, regardless of the used predictive model. Of particular relevance to this thesis, they showed that multi-task learning, i.e. modelling multiple gene dependencies simultaneously, was found to be better than training a separate model for each gene. They argue that the success of multi-task approaches is likely because they can provide additional support for finding patterns shared across genes involved in similar cellular processes. However,

their analysis was mostly focused on predictive performance. Therefore, they did not extensively investigate additional insights that can be gained by analysing large scale dependency data using multi-task approaches.

In this thesis, we decided to use a multi-task learning approach largely motivated by their potential to improve gene dependency prediction in the DREAM challenge and the evidence of functional relationships between genetic dependencies showed in the studies that built gene networks based on dependency profiles [17, 19, 18, 20].

1.3 Multi-Task Learning

Machine learning algorithms aim to extract useful information contained in available data samples in order to perform a given task. For example, in our application, we are interested in a regression task, where the goal is to predict a variable of interest (dependency on a given gene) for a given sample (cancer cell line), based on the cancer cell line features. A large amount of data is typically needed so that machine learning models can successfully generalise from training to unseen samples. In our case, large-scale loss-of-function screens do not contain many samples (≈ 600) compared to the number of features (≈ 18000). However, they contain data to perform multiple regression tasks, since they measured dependency of the same cell lines on thousands of genes.

Multi-task learning (MTL) algorithms precisely aim at combating this data scarcity problem in the presence of multiple related machine learning tasks. The main principle behind MTL is that performance on each individual task can be improved by sharing information from multiple learning tasks. This is based on the assumption that at least a subset of the tasks shares some similarities and therefore it can be beneficial to jointly learn them.

The two main ingredients that differentiate multi-task algorithms are the type of tasks that they are designed for (e.g. supervised or unsupervised tasks) and the way task relatedness is encoded in the model. In this thesis, we are only going to focus on multi-task algorithms in a supervised setting, since we are interested in improving multiple regression tasks. For a more comprehensive overview of multi-task learning, see [22].

Zhang et al [22] divide supervised MTL models into two main groups depending on how they encode task relatedness: feature or parameter-based approaches. Within these groups, methods can also differ in some important aspects. Namely, (i) whether task relatedness is inferred from the data or encoded based on prior information and (ii) whether information is shared by all tasks simultaneously or only within subsets of more related tasks. Feature-based methods assume that tasks share a feature representation that can be used to predict the outputs of all tasks. For example, deep learning networks or even a network with three layers (input, hidden and output), can be seen as a MTL algorithm. Other methods include feature selection methods that try to identify the feature subset that is relevant for all tasks simultaneously. Parameter-based approaches explore similarities between tasks in order to learn common model parameters. For example, task clustering methods cluster tasks and learn model parameters that are shared among tasks from the same cluster.

Multi-task learning methods have been frequently shown to improve performance over single task methods [21, 23, 24]. This improved performance is due to several reasons that help to combat overfitting, as reviewed in [25]. Namely, learning multiple tasks concurrently leads to an increased effective sample size during training, as more data is available for each task. This can help in discerning features that are consistently important across several tasks, which is particularly beneficial in high-dimensional feature spaces. What is more, multi-task algorithms are less prone to model noise, as they have to learn feature representations or parameters that are useful for multiple tasks simultaneously. Despite these considerations, predicting whether MTL will actually improve performance on individual tasks is difficult and problem dependent. In fact, this is still considered an open issue in MTL [22].

2 Extended Methods

In this section, we describe in more detail the Macau method as well as our approach to select cancer dependencies for the learning task. To provide more context to the reader, we also highlight how they compare to other methods or approches in the literature.

2.1 Macau

Macau belongs to the familiy of Bayesian matrix factorization (BMF), with the added property that side information can be incorporated. In essence, BMF approaches aim to decompose a matrix $Y \in \mathbb{R}^{N \times P}$ into matrices $U \in \mathbb{R}^{N \times K}$ and $V \in \mathbb{R}^{K \times P}$ such that

$$Y \approx UV \tag{1}$$

In our application, Y corresponds to the dependency scores matrix, with N cell lines and P genes. K corresponds to the number of chosen latent dimensions.

Non-probabilistic approaches typically find matrices U and V by solving a minimisation problem, with a cost function such as the total square error of predictions. Probabilistic methods estimate these matrices from a different point of view, where the goal is to infer the distributions over latent variables, after observing the data Y. For continuous data, a gaussian noise model is assumed for the observed values Y_{ij}:

$$p(Y|U,V,\alpha_R) = \prod_{(i,j)\in I_R} \mathcal{N}(Y_{ij}|U_i V_j, \alpha_R^{-1}) \tag{2}$$

where $\alpha_R > 0$ is a known precision parameter, U_i and V_j correspond to the ith row of U and jth column of V. I_R denotes the set of observed matrix entries, in this case, the measured dependency scores. In addition, Bayesian approaches introduce prior distributions over U and V, which are the equivalent of penalty terms on the latent vectors in the non-probabilistic approach. As with other BMF approaches, Macau places multivariate Gaussian priors on the latent variables and it also places Normal-Wishart priors over the respective hyperparameters: $\Theta_U = \{\mu_U, \Theta_U\}$ and $\Theta_V = \{\mu_V, \Theta_V\}$:

$$p(U, \mu_U, \Lambda_U|\Theta_0) = \prod_{i=1} \mathcal{N}(U_i|\mu_U, \Lambda_U^{-1})\mathcal{NW}(\mu_U, \Lambda_U|\Theta_0) \tag{3}$$

$$p(V, \mu_V, \Lambda_V|\Theta_0) = \prod_{j=1} \mathcal{N}(V_j|\mu_V, \Lambda_V^{-1})\mathcal{NW}(\mu_V, \Lambda_V|\Theta_0) \tag{4}$$

where μ_U and Θ_U (μ_V and Θ_V) correspond to the mean and precision matrix for the Gaussian prior of U and V , respectively. Θ_0 corresponds to the fixed hyperparameters of the Normal-Wishart hyperprior.

The key property of Macau is the ability to incorporate side information, such as row features $x_i \in \mathbb{R}^{F_{row}}$ and/or column features $z_j \in \mathbb{R}^{F_{column}}$. In order to do this, Macau incorporates an extra term ($\beta_U^T x_i$ or $\beta_V^T z_j$) into the Gaussian mean of each prior that is placed on the latent variables U_i and V_j. In other words, the prior mean of U_i and V_j becomes dependent on the side information. This means that equations 3 and 4 are modified in Macau as follows:

$$p(U, \mu_U, \Lambda_U|\Theta_0) = \prod_{i=1} \mathcal{N}(U_i|\mu_U + \beta_U^T x_i, \Lambda_U^{-1})\mathcal{NW}(\mu_U, \Lambda_U|\Theta_0) \tag{5}$$

$$p(V, \mu_V, \Lambda_V|\Theta_0) = \prod_{j=1} \mathcal{N}(V_j|\mu_V + \beta_V^T z_j, \Lambda_V^{-1})\mathcal{NW}(\mu_V, \Lambda_V|\Theta_0) \tag{6}$$

where $\beta_U \in \mathbb{R}^{F_{row}\times K}$ ($\beta_V \in \mathbb{R}^{F_{column}\times K}$) are the weight matrices for row/column features and F_{row} and F_{column} are the number of row/column features, respectively. Since these matrices link the side information to the latent matrices, they are called link matrices. In essence, these link matrices act as regression weight matrices with fixed inputs x_i and moving targets U_i (these targets are changing during the estimation with Gibbs Sampling). In our application, we do not have column side information, only on the rows. The row side information corresponds to the cancer cell line feature matrix $X \in \mathbb{R}^{N\times M}$, where $x_i \in R^M$ corresponds to the features that characterise cell line i. Note that the effect of the side information depends on the number of observations of each row/column. If a row has few observations, then the distribution of its respective latent vector will be more influenced by its features. On the other hand, this influence will be smaller if the row has many observations.

For a full bayesian treatment, a zero mean multivariate normal prior is placed on β_U:

$$p(\beta_U|\Theta_U, \lambda_{\beta_U}) = \mathcal{N}(vec(\beta_U)|0, \Lambda_U^{-1}(\lambda_{\beta_U} I)^{-1}) \tag{7}$$

where vec denotes vectorization and $\lambda_{\beta_U} \geq 0$ corresponds to the diagonal element of the precision matrix. A gamma distribution hyperprior is placed on λ_{β_U}, with fixed μ and υ:

$$p(\lambda_{\beta_U}|\mu, \upsilon) = gamma(\lambda_{\beta_U}|\mu, \upsilon) \tag{8}$$

The equations are similar for β_V and were omitted for the sake of brevity.

The estimation of the latent and link vectors can be performed via Gibbs Sampling because their conditional distributions are easily sampled from. Put simply, in each iteration, Gibbs sampling cycles through each model variable and collects a sample from each conditional distribution, while keeping all the other variables fixed. Samples collected in the beginning of the estimation (burn-in period) may not accurately represent the desired conditional distributions. Therefore, burn-in samples are usually discarded to improve the quality of the estimations. The remaining collected samples should correctly approximate the conditional distributions, since the sampler is guaranteed to obtain asymptotically correct distributions.

Of note, it is relatively fast to sample all model variables from their conditional distributions except for β_U and β_V. This is because it quickly becomes expensive to sample from multivariate Gaussian distributions when F_{row} or F_{column} are large. Therefore, Macau employs a trick: samples of β_U or β_V are obtained by solving a linear system whose solution corresponds to the posterior mean of each link variable. This step is crucial for the scalability of Macau to thousands of side information features and is proven in Macau original paper [26].

From a methodological perspective, Macau can be seen as a collaborative filtering technique and multi-task algorithm. Collaborative filtering (CF) techniques make use of similarities between users (in our case cell lines) and between items (genes) to predict the interest of a user in a new item. These preferences are summarised in a user-item matrix, which corresponds to our dependency scores matrix. Among CF techniques, Macau falls into the category of model-based techniques that make use of side information (see [27] for a extensive review of CF techniques). Specifically, as a collaborative filtering method, Macau incorporates side information about the user or items into the factorization of the user-item matrix to improve the estimation of the latent factors and consequently the user preference predictions. As a multi-task learning method, Macau falls into the category of a feature-based approach. This is because it finds a common feature representation of the side information (cancer cell line features) that is useful for the prediction of the dependency on several genes (tasks).

Macau has been used for predicting gene-disease associations [28], drug-protein binding [26] and drug response in cancer cell lines [29]. The first two studies mostly deal with a matrix completeness problem. In particular, their goal is to use biologically relevant side information either to improve the prediction of unmeasured gene-disease associations or of drug-protein binding. They show that Macau can greatly improve the predictions compared with previous methods. In [29], the authors use Macau to decompose a drug response matrix using side information about the drugs (target information) and the cancer cell lines (signalling pathway activation scores, obtained from gene expression data). They showed the method recovers several known associations between alterations in cancer related pathways and sensitivity or resistance to specific drug targets. However, they are quite limited by the number of pathways analysed (11 in total). Lastly, [30] have also analysed Macau, but from a more technical perspective. Specifically, they investigated the influence of the percentage of missing data in the predictive performance of the method. Interestingly, the performance of Macau was similar to that of a deep neural network, despite using considerably less time and computation resources. However, none of these studies systematically investigated how Macau leverages the inherent relationships between rows and columns to perform its predictions and how the latent space it finds can be biologically interpreted. These are main focuses in our work.

2.2 Selection of cancer dependencies for the learning task

Prioritizing cancer dependencies based on dependency scores from shRNA screens is not trivial due to the technical limitations of these screens, as mentioned in section 1.1.1. For example, applying a universal cutoff on the dependency scores is a poor strategy, due to screen quality differences between cell lines and between genes.

The most common approach in the literature for identifying cancer dependencies has been to investigate the dependency scores of the same gene across cell lines (i.e. gene dependency profile) and detect whether there are cell lines with unusually low scores for that gene. The idea behind this is that such cell lines are likely dependent on that gene. To select genes with promising dependency profiles, Tsherniak et al [13] use a threshold of 6 standard deviations away from the mean of the gene profile. McDonald et al [6] use a test statistic that investigates how divergent from a normal distribution the gene profile is. The main issue with the former approach is that it aims to find outlying observations, while using a central tendency and a scale indicators that are in turn quite sensitive to outliers. The latter approach makes strong assumptions regarding the non-normality of essential gene profiles, which is not guaranteed to be true. What is more, analysing gene profiles does not directly point to the most likely genetic dependencies in a given cell line, which is ultimately more informative to find new drug targets for that cell line.

Therefore, we propose a different approach to select cancer dependencies, which aims to tackle the aforementioned issues. We follow the same rationale as [13] but with key differences. First, we detect dependency scores with large deviations from the median of the dependency profile. Deviations are measured in median absolute deviation (MAD) units. Similarly to mean and standard deviation, median and MAD are indicators of central tendency and scale, respectively. However, these indicators are not as sensitive to the presence of outliers in the distribution under analysis [31]. Of note, an underlying assumption of normality of the dependency profile still exists, but it is not affected by the abnormality caused by potential outliers. Second, we analyse both gene and cell line dependency profiles. We consider that potential cancer dependencies should also be identified using cell line profiles because genes corresponding to the tail of the cell line dependency profile are more essential compared to other genes in the same molecular context. These are potentially good

targets to treat those cells without necessarily being flagged as such if only the gene dependency profiles are considered. Figure 3 exemplifies how genes are selected using the dependency profiles of the genes and the cell lines. In the gene-based approach, we analyse the gene dependency profile and select the gene if there are at least two cell lines whose dependency scores for that gene are lower than the threshold. We set this threshold as 6 MAD units away from the median of the dependency profile. In the cell line based approach, we select all the genes for which the dependency scores in that cell line are lower than the threshold. These genes should also be independently detected in the profiles of at least two cell lines.

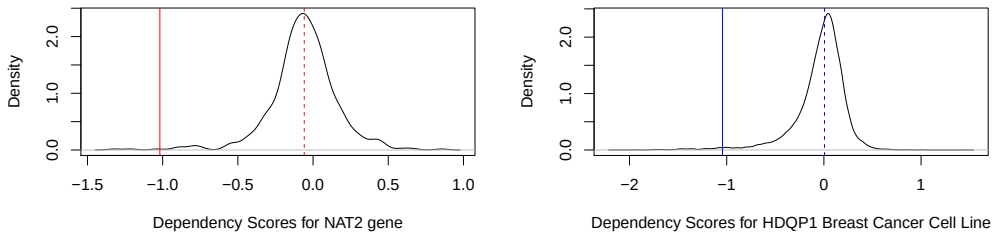

Figure 3: Procedure to select genes for the learning task, using two complementary approaches. Left: Gene-based approach. Right: Cell line-based approach. Dashed lines represent the median of the gene or cancer cell line profiles, while the solid lines represent the applied threshold (6 MAD from the median of the dependency profile).

Of note, a recent study [14] also focused on cell line dependency profiles to select cancer dependencies. However, unlike our approach, they use a reference set of essential and non essential genes to establish a cell line specific threshold on dependency scores. This approach has the limitation of relying on a reference gene set that was generated in a different experiment with a limited number of cell lines (< 20). This can bias the analysis and potentially lead to unwanted consequences, such as leaving out unknown context specific dependencies. We consider that a data-driven approach is more adequate for our purposes, since we are not aiming at defining a very restrictive set of cancer dependencies. Instead, we want to find genes whose dependency scores indicate potential cancer cell line dependency and are therefore worth keeping in our supervised model.

3 Extended Results

In this section, we show additional analyses that are not present in the thesis article. In the first subsections, we report results of our initial exploration of the dependency data. Namely, we highlight some limitations of analysing the dependency data using Elastic Net models. In the last subsection we provide supplementary figures and tables that support the main article.

3.1 Cancer cell lines belonging to the same cancer type have similar dependencies

The dependency data contains a diverse collection of cancer cell lines. In total, 30 cancer types are represented, of which 22 have at least 5 cell lines (Figure 4). We observed that correlations between the dependency profiles of cancer cell lines within the same cancer type are significantly higher than correlations between cancer cell lines from different cancer types (Wilcoxon test, pval$<$0.01). This shows that cell lines belonging to the same cancer type have more similar genetic dependencies. Indeed, if we focus on the genes with the 5 % most variant dependency scores, cancer cell lines cluster according to cancer type (Figure 5).

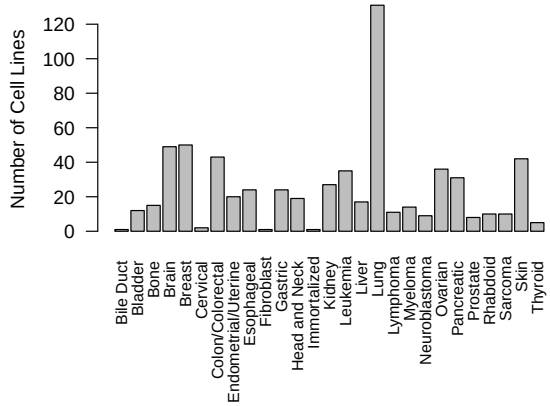

Figure 4: Number of cell lines in the dependency data, grouped by primary cancer type.

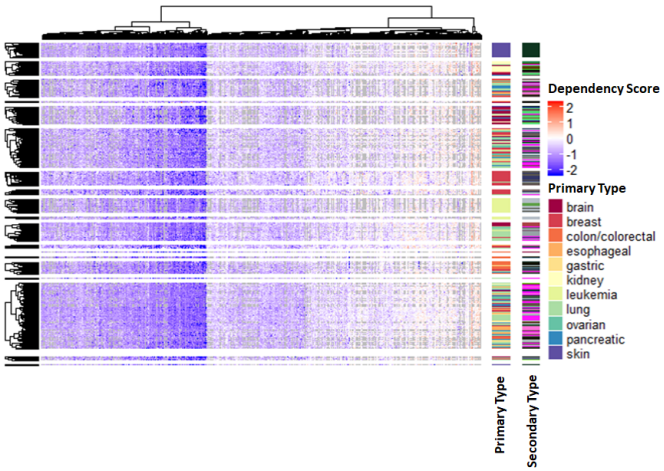

Figure 5: Heatmap showing the hierarchical clustering of cancer cell lines using pair-wise Pearson correlations based on the 5% most variant genes. Primary/Secondary Type denotes the cancer cell line primary type (tissue-related) and sub-type, respectively.

3.2 Cancer cell line and gene dependency profiles show different patterns

We analysed in more detail the cancer cell line and gene dependency profiles in the dependency data, where a cancer cell line dependency profile corresponds to the dependency scores of all genes measured in a given cell line and a gene dependency profile to the dependency scores measured for the same gene across all screened cancer cell lines. Figure 6 shows some trends in these profiles. Overall, cancer cell line dependency profiles are negatively skewed, with a large left tail, indicating that every cell line has a subset of genes which they are dependent on. On the contrary, gene dependency profiles are quite symmetric and typically centered around 0. This indicates that individual gene suppression of most genes does not have a substantial effect on cell line survival. This is expected, as some genes are only required in specific environmental conditions and cells have typically several redundancy mechanisms to cope with perturbations on a specific biological function [32]. Of note, Figure 6 shows that a considerable percentage of genes have very low mean dependency scores, which indicates that multiple cell lines are dependent on them. Not surprisingly, these genes are related with core cellular complexes, namely ribosome, proteasome and spliceosome.

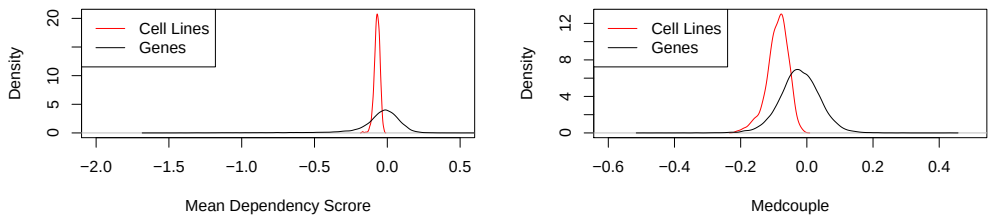

Figure 6: Means (left) and Medcouples (right) of the distributions of dependency scores per cell line and per gene. Medcouple is a measure of distribution skewness: 0 represents no skewness, while a positive/negative medcouple represents positive/ or negative skewness, repectively.

3.3 Elastic Net has some limitations when modelling gene dependency scores

In this subsection, we describe our analysis of the dependency data by fitting one Elastic Net model to each gene selected in 2.2. As mentioned in the main article, these models predict gene dependency scores based on cancer cell line features. Note that the results in this subsection refer to models trained with mutation, copy number and transcriptomic features of cancer cell lines, even though we only use transcriptomic features in the main paper. Mutation and copy number data were also obtained from the Cancer Dependency Map Consortium.

3.3.1 Alpha and Lambda Parameter Tuning

When fitting Elastic Net models, two parameters have to be tuned. However, cross-validation for finding the optimal pair for each gene is computationally exhaustive and was not directly implemented in the R package glmnet [33]. In order to explore the effect of these two parameters in model performance, we focused on a random sample of 100 genes and fitted 10 elastic net models per gene. Each of these models was trained with a different α value. The λ parameter was obtained through cross validation. Figure 7 shows that λ values depend on the set α value. However, there is no overall optimal α value across all modelled genes (Figure 8).

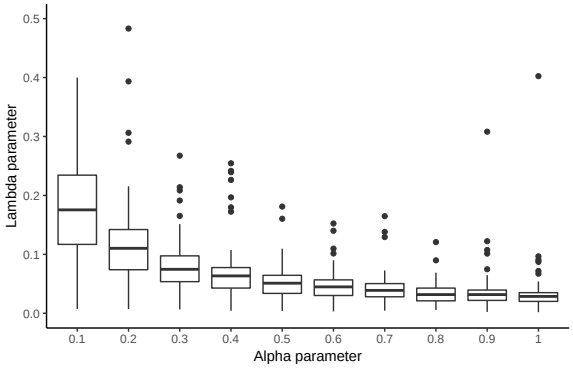

Figure 7: Optimal λ values for 10 different α values, obtained for elastic net models fitted to a random sample of 100 promising genes

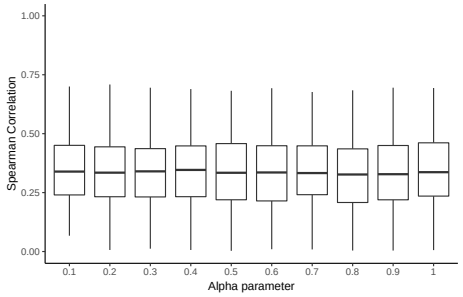

Figure 8: Performance measured using Spearman's correlation for elastic net models fitted to a random sample of 100 promising genes, for 10 different α values

Since the α parameter ultimately controls the number of significant features returned by each model, we decided to choose the value $\alpha = 0.2$, as it performs a substantial selection of genomic features (see Figure 9), without hurting overall performance. Of note, we could potentially slightly improve the regression *per gene* by performing cross-validation of α parameter, however it is not uncommon to fix this parameter (see, for example [34, 35], [1]), as its main influence is in the interpretability of the model, provided λ is cross-validated.

Having selected the α parameter, another choice had to be made regarding λ. Using λ_{min} typically leads to the selection of a higher number of cancer cell line features without having a better performance on an independent (test) dataset, with unseen cell lines. Indeed, a worse performance is achieved in terms of Spearman correlation for models with at least one non-zero coefficient (Figure 10). This indicates that choosing λ_{1se} prevents overfitting and therefore favours the selection of more relevant features. Hence we trained our elastic net models using λ_{1se}.

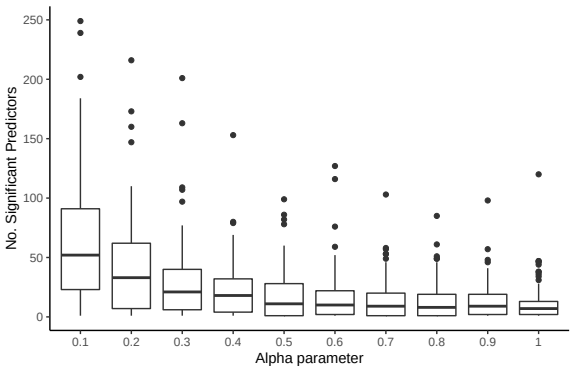

Figure 9: Number of significant predictors for 10 different α values, obtained for elastic net models fitted to a random sample of 100 promising genes. Significant predictors refers to features whose estimated coefficient is different from 0.

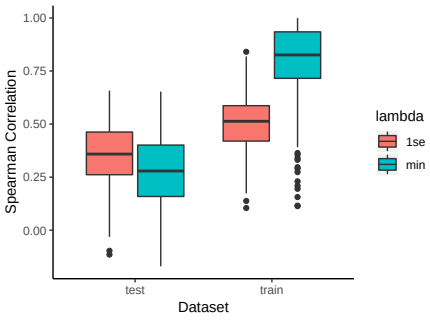

Figure 10: Comparison of Spearman correlation on train and test datasets for two different λ: min corresponds to λ that led to lowest MSE on inner cross validation in the training set and 1se to the largest value of λ which leads to an error between the minimum error and minimum error + 1 standard error. Data refers to 300 randomly sampled genes in a 80/20 % train test split of the original dependency data.

[1] https://web.stanford.edu/ hastie/glmnet/glmnet_alpha.html

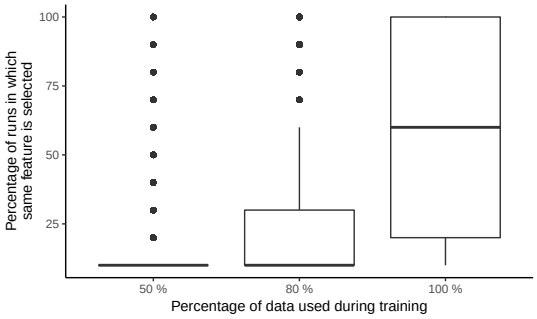

Figure 11: Percentage of different runs in which the same feature is selected by the Elastic Net models fitted to the same gene. Three different settings are shown, where we varied the percentage of data available during training.

3.3.2 Features selected by Elastic Net are not very robust across different subsamples of the cell lines in the dependency data

Next we investigated how stable the feature selection is. That is, how frequently we would select the same features if we were to train the elastic net model on i) different subsamples of the cell lines available in the dataset ii) different runs on the entire dependency dataset. To do so, we focused on a random sample of 100 genes. We fitted Elastic Net models to each gene, in three different settings: using 50, 80 and 100 % of the available cell lines. Ten elastic net models were fitted for each gene in each setting.

Figure 11 shows that the Elastic Net models are very sensitive to the dataset used during training. Most of the selected features are only selected once when drawing 10 random sets of cancer cell lines corresponding to 80% of the original dataset. The relatively low stability of the feature selection might be due to the imbalance between dependent and non-dependent cell lines for each gene, the high dimensionality of the feature space and the high correlation between features. Regarding the stability of different runs on the entire dataset, we observed that Elastic Net is relatively stable - half of the features are selected in at least half of the runs.

3.3.3 Elastic Net cannot model dependency for most of the genes

We next fitted an Elastic Net Model to each of the genes selected in the section 2.3 of the thesis article. In order to increase robustness, we fitted Elastic Net 10 times to each gene. For 817 genes, all the independent runs of Elastic Net selected at least one significant feature (i.e. at least one feature had a non-zero coefficient). We refer to these models as significant models. Of note, the percentage of significant models (39%) is similar to the one reported in [13] (37%), even though the Elastic Net model is much simpler than the non linear method they used. This is likely due to the increase in cell lines screened and also perhaps due to the filtering procedure to identify the subset of genes with unusually low dependency scores.

We also noted that prediction accuracy was quite different for genes selected based on cell line or on gene dependency profiles. Figure 12 shows the Pearson and Spearman's correlations for significant and non-significant models, grouped according to the approach that flagged the corresponding gene. As expected, the correlation between predicted and actual values is higher for significant models. More interestingly, the percentage of significant models is higher for genes selected using the approach based on cancer cell line dependency profiles than the gene-based approach (64 % vs 9% models). What is more, models corresponding to genes identified using the cell line-based approach have a higher predictive power than those identified using the gene approach (in terms of Spearman's correlation). This suggests that the cell line approach is as adequate as the gene approach in selecting genes of interest, even though the latter approach is the most frequently used in the literature. Indeed, gene dependency profiles of genes selected by the cell line-based approach have higher standard deviations than profiles of genes selected by the gene-based approach. Of note, [21] compared the performance of several gene dependency models on several genes and showed that genes whose dependency profiles have higher standard deviations are easier to predict, regardless of the model used. In line with this, we verified that genes selected using the cell line approach have larger standard deviations than those predicted with the gene-based approach (Wilcoxon test, p-value<0.01).

Additionally, we investigated the difference in relative performance of the approaches regarding Pearson and Spearman's correlations. We considered this is most likely due to outliers. While models of genes flagged through the cell line approach show very similar Pearson and Spearman correlations, models of genes flagged through the gene approach have in general higher Pearson correlation than Spearman's. This is a typical indicator of outliers influencing the regression: since Spearman's correlaton is more robust to outliers, its values are lower. However, Pearson correlation increases, as the outliers push the regression towards

themselves to decrease the mean squared error. This makes sense, as the gene approach is exactly trying to flag genes whose distribution contains outliers.

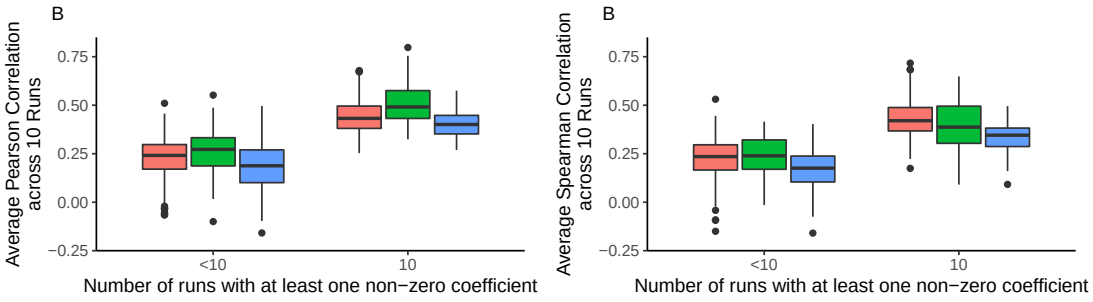

Figure 12: Comparison of (A) Pearson and (B) Spearman correlations for fitted Elastic Net models. Approach refers to the approach that selected the corresponding gene (either based on cell line or gene dependency profiles). Common refers to genes that were flagged independently by the cell line and gene- based approaches. Correlations refer to the correlations between predicted and actual dependency scores computed on an independent test set (80 20% split).

3.3.4 Biological processes underlying dependency on multiple genes might be hard to find with Elastic Net

Lastly, we analyzed the cancer cell line features that were considered predictive of gene dependency by the Elastic Net models. For our analysis, we decided to consider all of the features that were extracted for each gene in at least one of the 10 runs of Elastic Net. This is because features that were selected less frequently are not necessarily less relevant - they can simply be correlated with other relevant features.

The top 10 most frequently selected features across all genes are represented in Figure 13. The role of some of these genes have been extensively studied in cancer, namely TP53 as a tumour suppressor [36], with EDA2R and FDXR also being involved in p53-dependent apoptosis [37, 38]. Of note, as observed in [13], gene expression features were also the most informative type in our analysis, corresponding to 59% of the predictive features obtained from the models with significant predictors. Copy number and mutation features represented 32% and 8% of the selected features, respectively.

To obtain an overview of the biological processes that are related with gene dependency, we focused on the set of features selected by the Elastic Net models for each gene and performed a pathway overrepresentation analysis (see section 2.4.2 in the main article). We found enriched pathways for around 30 % of the genes with significant models. A median of 2 and 3 for KEGG and Reactome pathways were enriched per gene, respectively. Figure 14 shows an overview of the most frequently enriched pathways identified.

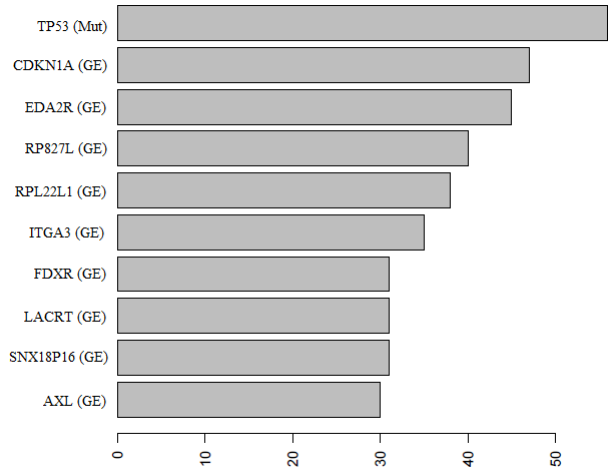

Figure 13: Top 10 most frequently selected features across all genes. GE indicates that the feature corresponds to a gene expression feature, whereas Mut to a mutation feature.

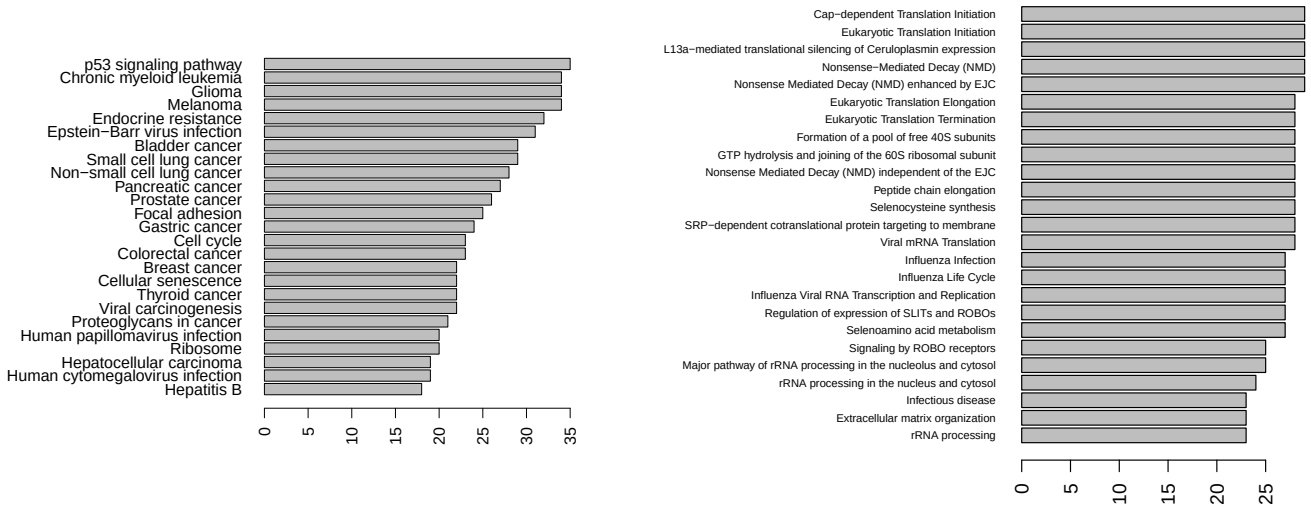

Figure 14: Overview of the 25 most frequently enriched pathways in the features selected by Elastic Net models fitted to each gene. (KEGG on the left, Reactome on the right)

The enriched KEGG pathways are mainly broad cancer-related pathways. This is expected as they represent alterations in cancer cells that can make them more dependent on the selected genes. On the other hand, most frequent Reactome pathways are mostly related with translation. This is likely because a relatively large percentage of the selected genes (approximately 7%) are related with translation, a core process for cells. Therefore, it is not surprising that cancer cell line features associated with dependency on several genes are frequently related with translation processes, as a gene's own expression is frequently predictive of its dependency scores [13], as well the expression of genes that interact with it [39]. It is also worth noting the difference between these two databases. KEGG pathways represent much more broad pathways, while Reactome pathways are more specific. Therefore, using the two databases allows to have a complementary view on the enriched pathways.

The number of cancer cell line features and enriched pathways shared across all of the Elastic Net models is quite low. This is unexpected because the genes selected for the learning task contain several genes involved in the same biological processes. Some studies have shown that functionally related genes have similar dependency profiles [18, 19]. Hence, one could hypothesize that dependency on functionally related genes is associated with similar cancer cell line features. The inability of Elastic Net to identify common features across different gene models can be explained by several reasons. First, Elastic Net is unable to predict dependency on most of the genes. This is likely because relationships between cancer cell line features and dependency are quite weak for most genes and therefore hard to capture with the available data. Second, Elastic Net does not necessarily select multiple highly correlated features even if they are all associated with dependency on a given gene. Lastly, Elastic Net models each gene independently from all the others. Collectively, these observations limit the power of Elastic Net models to identify common patterns behind cancer dependency on several genes. This motivated the use of a multi-task approach in our work.

3.4 Supplementary Figures and Tables

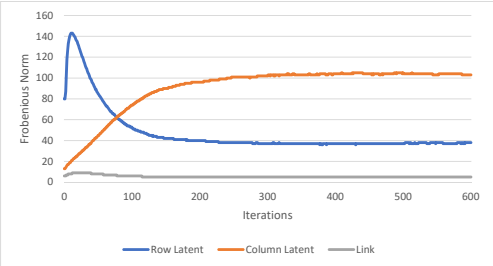

Figure 15: Frobenious Norm of the Row and Column Latent Matrices and of the Link Matrix for the first 600 iterations of Macau's optimization procedure (Gibbs Sampling). Stabilization is achieved after 400 iterations.

Table 1: Effect of feature normalization on the predictive performance of Macau and Elastic Net. Reported values are the average performance across all genes with significant Elastic Net models. Predictive performance for each individual gene corresponds to average Spearman correlation obtained across 10 cross-validation folds. (*) represents significant paired Wilcoxon tests (pval<0.01) for pairwise comparisons with our final model (in bold). Z-normalization corresponds to normalizing features to have 0 mean and unit standard deviation.

	Elastic Net	Macau
Without Transformation	0.221 (*)	0.288
Mean Centered	0.220 (*)	**0.288**
Z-normalization	0.292	0.292 (*)

Table 2: Effect size for the Wilcoxon test assessing the difference between within-group and between-group pairwise Pearson correlations. Groups correspond to CORUM protein complexes, KEGG and Reactome Pathways. Correlations were computed for different types of gene vectors. Tests for all gene vectors were significant (p-value < 0.01). For KEGG and Reactome pathways two values are reported: the first corresponds to smaller pathways (less than 50 genes) and the second to the remaining pathways.

Type of gene vector	CORUM	KEGG	Reactome
Dependency Scores	0.051	0.051/0.024	0.039/0.007
Elastic Net Coefficients	0.061	0.060/0.038	0.043/0.006
Macau - Latent weights (learned without cell line feature data)	0.117	0.105/0.07	0.081/0.018
Macau - Interaction weights	0.124	0.120/0.062	0.097/0.025
Gene Expression	0.182	0.185/0.131	0.158/0.078

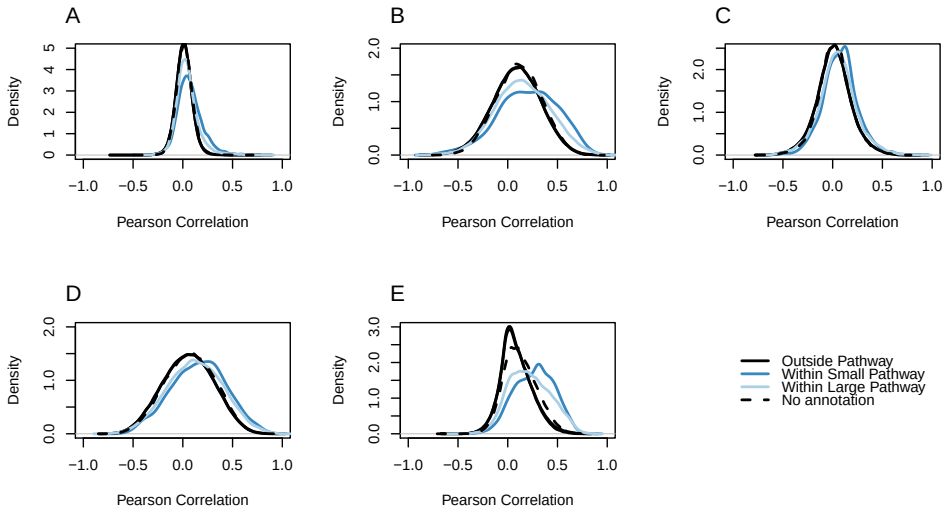

Figure 16: Pairwise Pearson's correlations between genes within the same KEGG pathway and between genes from different pathways. Correlations were computed using different types of gene vectors: (A) gene dependency scores profiles, (B) Macau interaction weights, (C) Elastic Net coefficients without sparsity regularization, (D) gene latent weights learned by Macau trained without cell line feature data and (E) gene expression profiles. No annotation corresponds to genes that could not be matched to any pathway in the KEGG database. Small pathways correspond to pathways with less than 50 genes. Large pathways correspond to the remaining KEGG pathways.

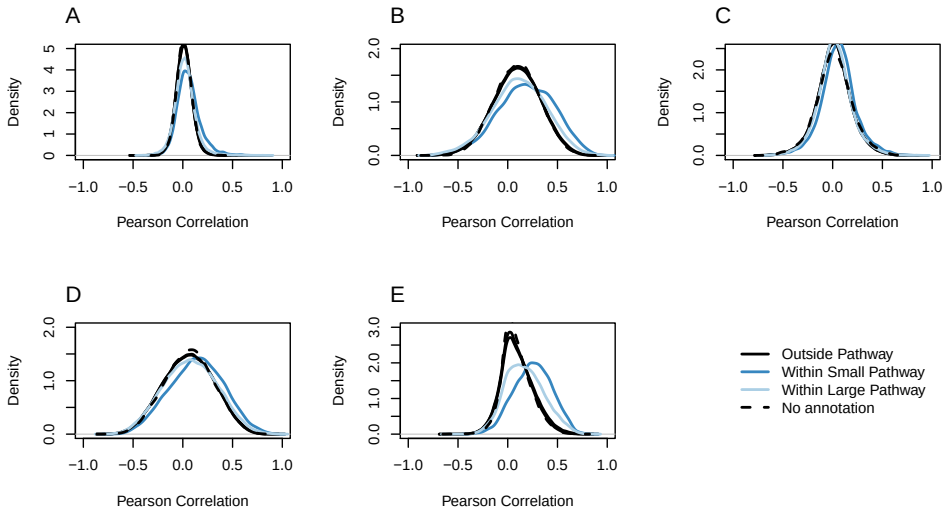

Figure 17: Pairwise Pearson's Correlations between genes within the same Reactome pathway and between genes from different pathways. Correlations were computed using different types of gene vectors: (A) gene dependency scores profiles, (B) Macau interaction weights, (C) Elastic Net coefficients without sparsity regularization, (D) gene latent weights learned by Macau trained without cell line feature data and (E) gene expression profiles. No annotation corresponds to genes that could not be matched to any pathway in the Reactome database. Small pathways correspond to pathways with less than 50 genes. Large pathways correspond to the remaining Reactome pathways.

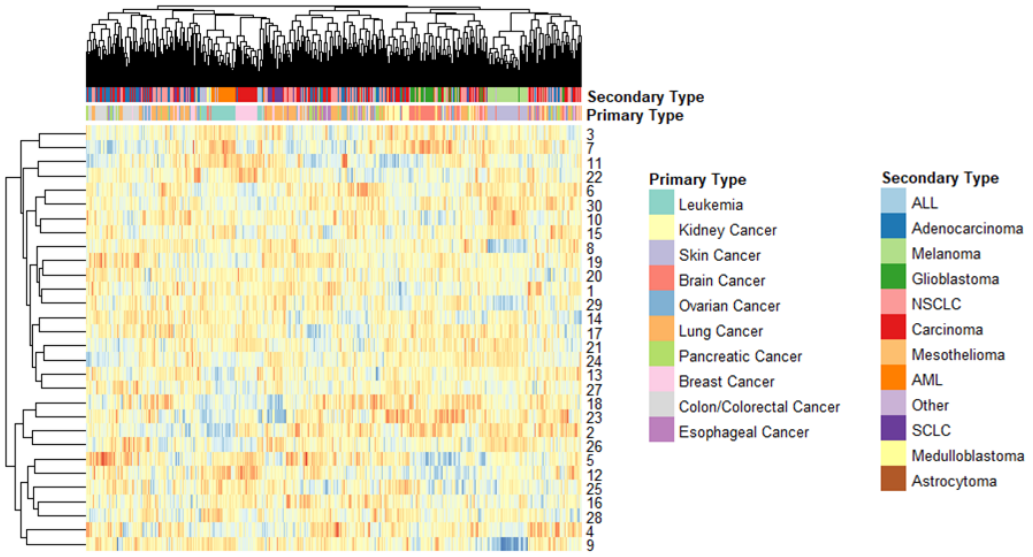

Figure 18: Latent Matrix coefficients for the cancer cell lines belonging to the top 10 most common cancer types in the dependency data. Cancer cell lines are clustered based on hierarchical clustering with euclidean distance and complete linkage. Some cancer types are clearly clustered in the latent space, namely Leukemia, Skin, Brain, Breast and Colon/Colorectal. Some separations within the same cancer type can be explained due to cancer subtype. For example, differences in Medulloblastoma and Glioblastoma subtypes in brain cancer, non small versus small cell (NSCLC/SCLC) in lung cancer and acute lymphoblastic leukemia (ALL) versus acute myelogenous leukemia (AML)

Figure 19: Most highly enriched Reactome pathways in the gene latent factors, together with a set of enriched Reactome signalling pathways. Color indicates NES (Normalized Enrichment Score). The most highly enriched pathways correspond to the ones with the most consistent behavior in the positive and negative tail of the latent factor. That is, the pathways with the highest (positive) NES and lowest (negative) NES. Yellow cells indicate the pathway was non-significantly enriched in the corresponding factor. Factors and pathways are ordered based on hierarchical clustering (euclidean distance, complete linkage). Based on their enrichment in the latent factors, pathways are roughly clustered in 4 main groups: Cell cycle, Signalling, Translation and Mitochondria related.

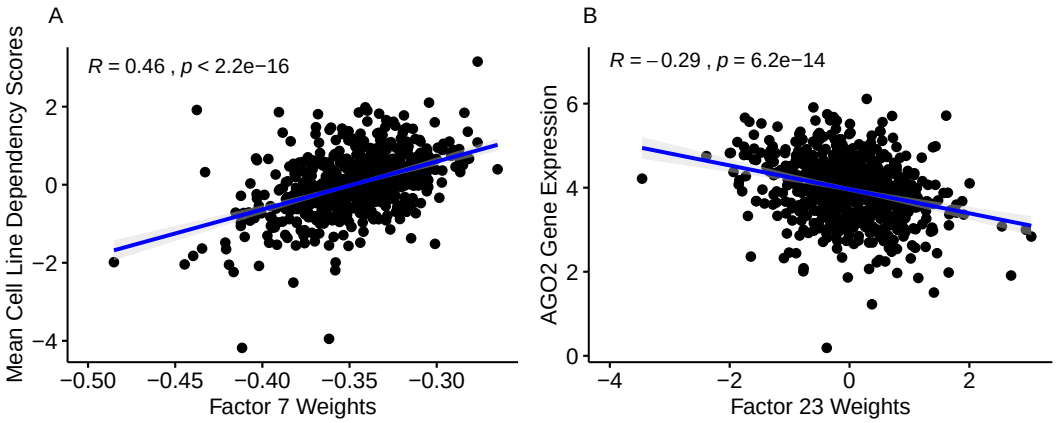

Figure 20: Correlation between covariates and factor weights. A) Correlation between factor 7 and a general level of cell line sensitivity to perturbations, computed by the mean cell line dependency scores. B) Correlation between factor 23 and AGO2 gene expression, a known technical confounder in shRNA-based screens. In both cases, we show the factor with the highest correlation with the covariate.

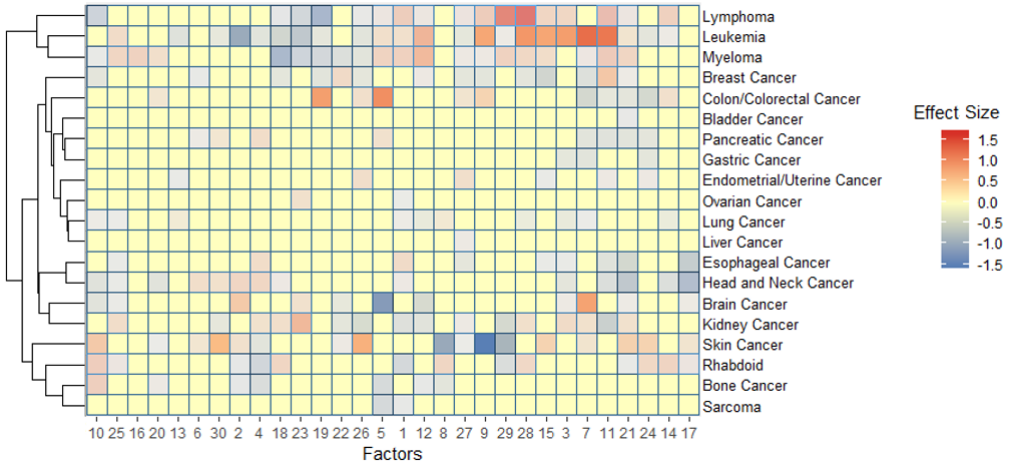

Figure 21: Specificity of each latent factor regarding cancer types. For each factor, a Wilcoxon test was performed for each cancer type. The tests assessed whether the latent weights distribution of cell lines within each cancer type was significantly different from that of other cell lines. Colors represent the test effect size. More specifically, the difference in location between the two distributions. Transparency represents the pvalue (the higher transparency, the higher the pvalue). Some latent factors clearly distinguish specific cancer types (e.g. skin cancer in factor 9). As expected, cancer types cluster according to their similarity (e.g. lymphoma, leukemia and myeloma, which are all blood related cancers).

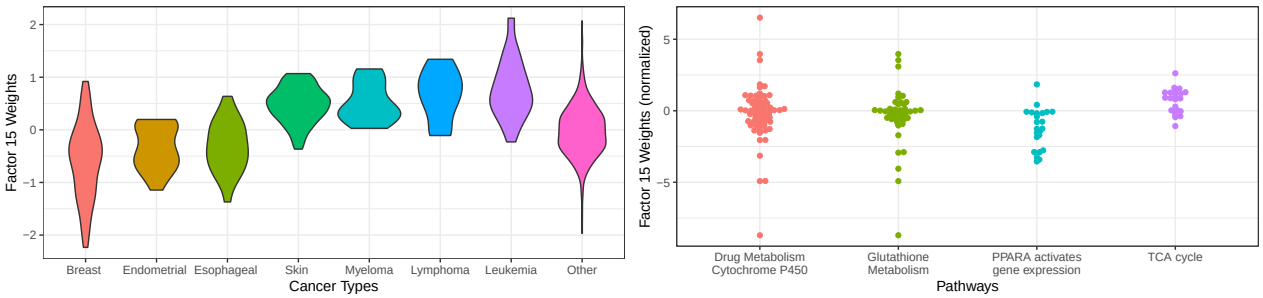

Figure 22: Inspection of Factor 15 Weights. Left: Weights of cancer types that are significantly different from other cancer types in this factor. Right: Weights of genes belonging to the enriched pathways in this factor. "KEGG Drug Metabolism Cytochrome P450" and "KEGG Glutathione Metabolism" pathways are enriched based on cell line features. "PPARA activates gene expression" and "TCA cycle" correspond to the top enriched Reactome pathways based on gene dependencies. Weights were normalized to have 0 mean and unit standard deviation.

References

[1] McFarland, J. M. et al. Improved estimation of cancer dependencies from large-scale rnai screens using model-based normalization and data integration. Nature Communications **9** (2018).

[2] Lord, C. J., Martin, S. A. & Ashworth, A. Rna interference screening demystified. Journal of Clinical Pathology **62**, 195–200 (2009).

[3] Cowley, G. S. et al. Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. Scientific Data **1**, 140035 (2014).

[4] Boettcher, M. & Hoheisel, J. Pooled rnai screens-technical and biological aspects. Current Genomics **11**, 162–167 (2010).

[5] Boettcher, M. & McManus, M. Choosing the right tool for the job: Rnai, talen, or crispr. Molecular Cell **58**, 575 – 585 (2015).

[6] McDonald, E. R. et al. Project drive: A compendium of cancer dependencies and synthetic lethal relationships uncovered by large-scale, deep rnai screening. Cell **170**, 577 – 592.e10 (2017).

[7] Shalem, O. et al. Genome-scale crispr-cas9 knockout screening in human cells. Science **343**, 84–87 (2014).

[8] Beijersbergen, R. L., Wessels, L. F. & Bernards, R. Synthetic lethality in cancer therapeutics. Annual Review of Cancer Biology **1**, 141–161 (2017).

[9] Evers, B., Jastrzebski, K., Heijmans, J. P. M., Grernrum, W., Beijersbergen, R. L. & Bernards, R. Crispr knockout screening outperforms shrna and crispri in identifying essential genes. Nature Biotechnology **34**, 631–633 (2016).

[10] Ling, A., Gruener, R. F., Fessler, J. & Huang, R. S. More than fishing for a cure: The promises and pitfalls of high throughput cancer cell line screens. Pharmacology & Therapeutics **191**, 178 – 189 (2018).

[11] Howard, T. P. et al. Functional genomic characterization of cancer genomes. In Cold Spring Harbor symposia on quantitative biology, 031070 (Cold Spring Harbor Laboratory Press, 2016).

[12] Marcotte, R. et al. Functional genomic landscape of human breast cancer drivers, vulnerabilities, and resistance. Cell **164**, 293–309 (2016).

[13] Tsherniak, A. et al. Defining a cancer dependency map. Cell **170**, 564–576 (2017).

[14] Behan, F. M. et al. Prioritization of cancer therapeutic targets using crispr–cas9 screens. Nature **568**, 511 (2019).

[15] Yamauchi, T. et al. Genome-wide crispr-cas9 screen identifies leukemia-specific dependence on a pre-mrna metabolic pathway regulated by dcps. Cancer Cell **33**, 386 – 400.e5 (2018).

[16] Tang, Y. C. et al. Functional genomics identifies specific vulnerabilities in pten-deficient breast cancer. Breast Cancer Research **20**, 22 (2018).

[17] Wang, T. et al. Gene essentiality profiling reveals gene networks and synthetic lethal interactions with oncogenic ras. Cell **168**, 890 – 903.e15 (2017).

[18] Pan, J. et al. Interrogation of mammalian protein complex structure, function, and membership using genome-scale fitness screens. Cell Systems **6**, 555–568 (2018).

[19] Kim, E., Dede, M., Lenoir, W. F., Wang, G., Srinivasan, S., Colic, M. & Hart, T. A network of human functional gene interactions from knockout fitness screens in cancer cells. Life Science Alliance **2** (2019).

[20] Boyle, E. A., Pritchard, J. K. & Greenleaf, W. J. High-resolution mapping of cancer cell networks using co-functional interactions. Molecular Systems Biology **14** (2018).

[21] Gönen, M. et al. A community challenge for inferring genetic predictors of gene essentialities through analysis of a functional screen of cancer cell lines. Cell Systems **5**, 485 – 497.e3 (2017).

[22] Zhang, Y. & Yang, Q. A survey on multi-task learning. CoRR **abs/1707.08114** (2017).

[23] Yuan, H., Paskov, I., Paskov, H., González, A. J. & Leslie, C. S. Multitask learning improves prediction of cancer drug sensitivity. Scientific Reports **6**, 31619 (2016).

[24] Costello, J. C. et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nature Biotechnology **32**, 1202 (2014).

[25] Ruder, S. An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017).

[26] Simm, J. et al. Macau: Scalable bayesian factorization with high-dimensional side information using mcmc. In 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP), 1–6 (IEEE, 2017).

[27] Shi, Y., Larson, M. & Hanjalic, A. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. ACM Comput. Surv. **47**, 3:1–3:45 (2014).

[28] Zakeri, P., Simm, J., Arany, A., ElShal, S. & Moreau, Y. Gene prioritization using bayesian matrix factorization with genomic and phenotypic side information. Bioinformatics **34**, i447–i456 (2018).

[29] Yang, M., Simm, J., Lam, C. C., Zakeri, P., van Westen, G. J., Moreau, Y. & Saez-Rodriguez, J. Linking drug target and pathway activation for effective therapy using multi-task learning. Scientific Reports **8** (2018).

[30] de León, A. d. l. V., Chen, B. & Gillet, V. J. Effect of missing data on multitask prediction methods. Journal of Chemoinformatics **10**, 26 (2018).

[31] Leys, C., Ley, C., Klein, O., Bernard, P. & Licata, L. Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology **49**, 764–766 (2013).

[32] Rancati, G., Moffat, J., Typas, A. & Pavelka, N. Emerging and evolving concepts in gene essentiality. Nature Reviews Genetics **19**, 34–49 (2017).

[33] Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software **33**, 1–22 (2010).

[34] Basu, A., Mitra, R., Liu, H., Schreiber, S. L. & Clemons, P. A. RWEN: response-weighted elastic net for prediction of chemosensitivity of cancer cell lines. Bioinformatics **34**, 3332–3339 (2018).

[35] Knowles, D. A., Bouchard, G. & Plevritis, S. Sparse discriminative latent characteristics for predicting cancer drug sensitivity from genomic features. PLoS Computational Biology **15**, e1006743 (2019).

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

[37] Brosh, R. et al. p53-dependent transcriptional regulation of eda2r and its involvement in chemotherapy-induced hair loss. FEBS letters **584**, 2473–2477 (2010).

[38] Zhang, Y. et al. Ferredoxin reductase is critical for p53-dependent tumor suppression via iron regulatory protein 2. Genes & Development **31**, 1243–1256 (2017).

[39] Lagziel, S., Lee, W. D. & Shlomi, T. Inferring cancer dependencies on metabolic genes from large-scale genetic screens. BMC Biology **17**, 37 (2019).