Machine Learning Methods for Learning Genetic Dependencies

by

Cathy Cai

B.S., Computer Science and Engineering and Mathematics, MIT, 2023

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Authored by: Cathy Cai
Department of Electrical Engineering and Computer Science
May 17, 2024

Certified by: Caroline Uhler
Professor of Electrical Engineering and Computer Science, Thesis Supervisor

Accepted by: Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Machine Learning Methods for Learning Genetic Dependencies

by

Cathy Cai

Submitted to the Department of Electrical Engineering and Computer Science
on May 17, 2024 in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

ABSTRACT

Synthetic lethality refers to a genetic interaction where the simultaneous perturbation of gene pairs leads to cell death. Synthetically lethal gene pairs (SL pairs) provide a potential avenue for selectively targeting cancer cells based on genetic vulnerabilities. The rise of large-scale gene perturbation screens such as the Cancer Dependency Map (DepMap) offers the opportunity to identify SL pairs automatically using machine learning. We build on a recently developed class of feature learning kernel machines known as Recursive Feature Machines (RFMs) to develop a pipeline for identifying SL pairs based on CRISPR viability data from DepMap. In particular, we first train RFMs to predict viability scores for a given CRISPR gene knockout from cell line embeddings consisting of gene expression and mutation features. After training, RFMs use a statistical operator known as average gradient outer product to provide weights for each feature indicating the importance of each feature in predicting cellular viability. We subsequently apply correlation-based filters to re-weight RFM feature importances and identify those features that are most indicative of low cellular viability. Our resulting pipeline is computationally efficient, taking under 3 minutes for analyzing all $17,453$ knockouts from DepMap for candidate SL pairs. We show that our pipeline more accurately recovers experimentally verified SL pairs than prior approaches. Moreover, our pipeline finds new candidate SL pairs, thereby opening novel avenues for identifying genetic vulnerabilities in cancer.

Thesis supervisor: Caroline Uhler
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

First, I would like to thank Professor Caroline Uhler for giving me the opportunity to conduct research under her guidance for the past three years. I learned everything I now know about research in her group, and I am tremendously grateful to her for her generosity and dedication to mentorship. Caroline entrusted me with a collaboration with Janssen, helped us finish our preprint before my PhD applications, and gave me the opportunity to TA her class to fund my MEng. I am incredibly thankful for her unconditional support, and I can only hope to pay a fraction of it forward in the future.

Second, I would like to thank Adityanarayanan (Adit) Radhakrishnan for being my research mentor and meeting with me week after week since my junior year of undergrad. Adit patiently taught me machine learning from the ground up, and I deeply cherish the opportunity to have been his student. I am sure that I would not be where I am today without Adit's teachings and mentorship.

I would also like to thank Chandler Squires and Jiaqi Zhang for teaching me causal inference and mentoring me on the causal representation learning project. I am also grateful to my collaborators, Christopher Moy and Barbara Weir, for supporting the synthetic lethality project and Narsis Attar for working me on the aneuploidy project. Thank you also to the other members of the Uhler Lab for making me feel welcome and keeping me company everyday.

I would also like to thank my family and friends for their support, particularly my parents and grandparents for giving me a wonderful home, my sister Angela for being my best friend in the whole world, and Allen for his unconditional emotional support. This thesis is dedicated to my family.

Contents

List of Figures

Chapter 1

Introduction

Synthetic lethality refers to the concept that simultaneous perturbation of gene pairs leads to cell death but individual perturbation does not [1]; see Fig. 1.1A for a schematic. The identification of synthetically lethal gene pairs provides a potential avenue for selective targeting of cancer cells based on genetic vulnerabilities and has already led to the development of therapies for specific patient subpopulations [2], [3]. The recent rise of large-scale perturbation screens such as the Cancer Dependency Map (DepMap) [4] offers an opportunity to identify novel SL pairs using machine learning. DepMap consists of a matrix of (real-valued) viability scores for each combination of 1078 cell lines and $17,453$ CRISPR gene knockouts; a subset of DepMap is visualized in Fig. 1.1B. To identify candidate SL pairs from such data, the goal is to find gene pairs (A,B), such that knockout of gene A induces a low viability score in cells that show a particular expression or mutation pattern for gene B. The difficulty in identifying SL pairs stems from the fact that the number of combinations of different expression and mutation patterns is huge.

Various computational approaches have been developed for identifying SL pairs. Motivated by the observation that SL pairs frequently arise in paralogs (i.e., genes arising as a result of duplication) [5]–[8], recent work trained a random forest model to predict whether a pair of paralog genes form an SL pair based on 22 hand-crafted features [9]. As such, the trained model is limited by the number of available features for prediction. Another line of work formulated SL pair identification as a link-prediction problem with links existing between SL gene pairs [10]–[13]. This approach requires SL pairs as training data, but only few experimentally validated SL pairs are known. To overcome this limitation, these works rely on non-experimentally verified SL pairs from SynLethDB [14] or protein-protein interaction databases to augment the training set, which may lead to many false positive results. An alternative approach [15] built a predictor for the confidence of SL interactions using knockdown data from the Connectivity Map [16]; namely, it trained a neural network to map from a pair of expression vectors, corresponding to expression after knockdown of each gene in a pair, to a confidence score quantifying whether the genes form an SL pair. Labeled data for these confidence scores were obtained using GEMINI [17], a computational model for gene interactions. This again could result in many false positive results. Alternatively, statistical hypothesis testing-based pipelines were developed to characterize SL pairs, and the test results were corrected and filtered through hand-crafted criteria to balance the number of false positives and false negatives [18], [19]. Overcoming the need for any labeled SL pair data or

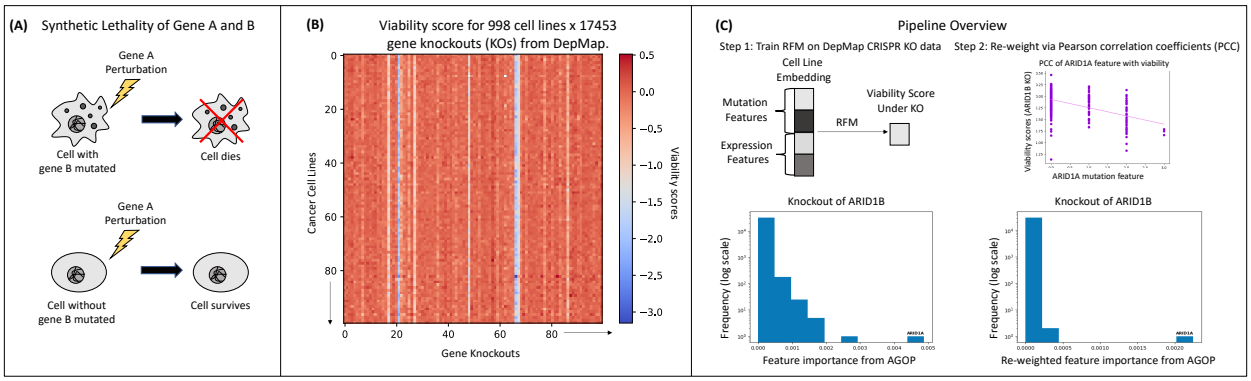


Figure 1.1: **(A)** Schematic describing the concept of synthetic lethality. Two genes (denoted A and B) form a synthetic lethality pair if the simultaneous perturbation of both genes leads to cell death but the individual perturbations do not. **(B)** A visualization of the DepMap CRISPR gene knockout data used in our pipeline. It consists of cellular viability scores for $17,453$ gene knockouts across 998 cell lines. **(C)** An overview of our pipeline. For each knockout, we train a Recursive Feature Machine (RFM) to predict the viability score from the gene expression and mutation features of a cell line. Feature importances are obtained from the trained RFM using a statistical operator known as average gradient outer product (AGOP). Given that each feature corresponds to (the expression pattern or mutation of) a gene, the genes can be ranked based on their feature importances. For example, for the knockout of ARID1B, the feature with the highest importance corresponds to the mutation of ARID1A. The second step of our pipeline re-weights the feature importances based on their Pearson correlation coefficient (PCC) with viability, thereby selecting features that are indicative of low cellular viability. The effect of such re-weighting is shown on the example of ARID1B knockout.

hand-crafted statistical criteria, recent work [20] formulated the problem of SL pair screening as a feature learning task where the goal is to identify the genomic features that are most predictive of low viability under a gene knockout. This feature learning approach is the one used by the "best model" listed in the DepMap portal. Current feature learning approaches have been limited to utilizing random forests, since these simple machine learning models have been the only non-linear models that could output feature importances. If instead we could identify the features learned by state-of-the-art machine learning models, we may be better powered to find SL pairs.

In this work, we present a computationally efficient pipeline for SL pair screening by leveraging a recently developed class of feature learning methods known as Recursive Feature Machines (RFMs). RFMs were introduced in [21] and identify features learned by kernel machines, a class of machine learning algorithms that have received renewed interest in machine learning due to their connection to infinitely wide neural networks [22]. RFMs identify task-relevant features using a statistical operator, known as average gradient outer product (AGOP). AGOP has been studied in the context of task-relevant dimensionality reduction [23]–[25], since it identifies and amplifies the directions in data for which predictions vary the most. Given the effectiveness of kernel machines on related tasks such as virtual drug screening [26], [27] and the ability of AGOP to identify task-relevant features, we

propose RFMs as a natural model for SL pair screening. Our approach is as follows: for each CRIPSR gene knockout from DepMap, we train an RFM to predict the viability score for each cell line from its mutation and gene expression features, and we use the AGOP to obtain the most predictive features. Since we are interested in pairs that reduce viability, we apply Pearson correlation-based filters to re-weight the predictive features and identify those that are indicative of low viability. Fig. 1.1C provides an overview of our pipeline. We will show that our pipeline recovers the experimentally verified SL pairs from [9] more accurately than previous random forest based approaches including PARIS [20] and the "best model" provided by the DepMap portal. Furthermore, by analyzing the top candidate pairs identified by our model, we obtain new candidate SL pairs that were not found using prior approaches.

Chapter 2

Results

2.1 Formulating SL pair screening as a single index problem.

We start by providing a novel formulation of the SL pair screening problem, which will motivate our computational approach to this problem. Recall that our goal is to find candidate SL gene pairs (A, B) such that the knockout of gene A induces a low viability score in cell lines exhibiting a particular expression or mutation pattern of gene B. We formulate this mathematically as follows. Let $(X, y_g) \in \mathbb{R}^{d\times n} \times \mathbb{R}^n$ denote training data where X denotes the embedding of n cell lines into d expression and mutation features and y_g denotes the viability scores when knocking out gene g. If g is part of an SL pair, then there exists a one-hot vector $u^* \in \mathbb{R}^d$ such that $y_g = h(X^T u^*)$ for some function $h : \mathbb{R}^d \to \mathbb{R}$, and the non-zero entry in u^* indicates the gene that forms an SL pair with g. This mathematical model is known as a *single-index model*, which has been extensively studied in the statistical literature [24], [25], [28] and received renewed interest in machine learning [21], [29]–[31]. Given that the work introducing RFMs empirically demonstrated higher sample efficiency of RFMs over other models including deep neural networks in solving such single-index problems [21], RFMs are a natural approach for SL pair screening.

2.2 Overview of our pipeline (SL-RFM).

Our pipeline is built using the publicly available cellular viability scores for all combinations of $17,453$ CRISPR gene knockouts on 998 cell lines provided by DepMap (see Data Availability). We dropped 80 cell lines for the downstream analysis to include only the ones present in The Cancer Genome Atlas (TCGA) [32]. The first step in our pipeline is to train an RFM-based predictor to map from mutation and genomic embeddings of a cell line to the corresponding viability score under a particular knockout. To build such a predictor, we encoded each cell line as a $32,629$ dimensional vector, where the first $16,568$ real-valued features reflect gene expression and the remaining $16,061$ features represent the presence of either a damaging or hotspot mutation (see Data Availability and Methods for a detailed description of the gene expression and mutation features). We note that we used only the

expression and mutation features provided by DepMap that are present in TCGA. Expression features are log transformed transcripts per million (TPM) (see Methods) and mutation features are binary (either 0 or 1) representing whether the gene has a mutation (either damaging or hotspot). After training the RFM, the model returns a list of weights for each feature denoting its importance for predicting viability for a given knockout. To ensure that the top ranked features, i.e., those with highest feature importance, are indicative of *decreased* viability, we re-weight the features by multiplying them with the Pearson correlation coefficient (PCC) between the value of the feature across all cell lines and the viability scores. Lastly, motivated by the formulation of SL pair screening as a single-index model, we identify candidate SL pairs by selecting those knockouts that contain a distribution of feature importances where the top ranked feature is well separated from the remaining features. We quantify such distributions via the difference between the maximum feature importance and the mean feature importance for a given knockout (see Methods). Training details for our pipeline are presented in Methods. In particular, the RFM framework requires the choice of a kernel. We tested the commonly used Gaussian and Laplace kernels and found that the Laplace kernel consistently outperformed the Gaussian kernel in terms of predictive performance on the DepMap data (see Methods and SI Fig. B.1). Thus, all results are shown using the Laplace kernel.

2.3 Our pipeline more accurately recovers known experimentally verified SL pairs.

We first analyze the performance of our pipeline on the experimentally verified SL pairs from [9] and compare the results to the following three models: (1) a Pearson correlation coefficient (PCC) baseline, which ranks features by their correlation with viability and corresponds to the last filtration step in our pipeline; (2) a state-of-the-art random forest model for SL screening (PARIS) [20]; (3) the "best model" from the DepMap portal. The models were evaluated as follows. For each SL pair (A, B) from [9], we applied each method to predict viability under knockout of gene A and to predict viability under knockout of gene B. For each prediction task, each method outputs an importance weight for each feature, which translates into a rank based on the importance weight for the feature corresponding to gene B in the first prediction task and gene A in the second prediction task. Fig. 2.1A shows the minimum of these ranks for each method; SI Fig. B.2 contains the ranks for each gene separately. Since the DepMap portal only provides the top 10 most important genes per knockout, for fair comparison across models the rank is denoted as > 10 if the pair was not identified within the top 10 most important features. We observe that SL-RFM recovers all experimentally validated SL pairs from [9] as the top ranked feature with the exception of ME2/ME3, which no model was able to identify accurately. In Fig. 2.1A, we also indicate whether the SL pair arises as a result of expression or mutation based features. In Fig. 2.1B, we present the feature importances for our method for specific knockouts from the table in Fig. 2.1A (the feature importance distributions for all pairs are provided in SI Fig. B.3). In particular, we observe that the knockouts for which SL-RFM accurately recovered the SL pairs contain distinct feature importance patterns consistent with those expected from a

single index model. Namely, there is usually one important feature corresponding to the gene that forms a synthetically lethal pair with the knockout. Note that in the one case for which SL-RFM does not identify the SL pair (ME2/ME3), there is no clear separation between the top feature and the remaining features. Lastly we emphasize the computational efficiency of our method: it involves solving a least squares problem with $32,629$ features and 998 samples for each knockout, which can be parallelized for all knockouts in a straightforward manner. As a result, the time to extract the feature importances for each knockout from SL-RFM is under 3 minutes, whereas that for PARIS is roughly 13 days.

2.4 RFM uncovers novel candidate SL pairs.

While we have thus far shown that our pipeline is able to accurately predict given experimentally verified SL pairs, the ultimate goal is to enable *de novo* identification of candidate SL pairs. Given that feature importances for known SL pairs followed our hypothesized single-index model by exhibiting a gap between the most important feature and the remaining features as shown in Fig. 2.1B, we used this to define a score (difference between maximum feature importance and the mean feature importance per knockout) to sort all knockouts. We selected a threshold to identify candidate SL pairs based on the elbow in the score distribution shown in SI Fig. B.4. This results in 92 candidate SL pairs given by top scoring knockouts and the gene corresponding to the top feature for each knockout (see Fig. 2.2A as well as SI Figs. B.5, B.6, B.7, B.8 for the feature importance distributions for these SL pairs). The 4 highest scoring pairs, DDX3X/DDX3Y, EIF1AX/EIF1AY, FAM50A/FAM50B, and RPP25L/RPP25, are all experimentally verified SL pairs [9], [35], [36]. Other high scoring (and experimentally verified [35]) paralog SL pairs include CDK4/CDK6, EAF1/EAF2, and COPG1/COPG2. Notably, SL-RFM also identifies non-paralog SL pairs including MTAP/PRMT5, MTAP/WDR77, GPX4/AIFM2, and MASTL/PPP2R2A, which all have been experimentally verified [37]–[39]. Furthermore, we identify 327 genes that are paired with themselves, including the prominent oncogenes KRAS, BRAF, PIK3CA, which were previously identified to build SL pairs with themselves [19]. Such genes may be related to the concept of oncogene addiction [40]. In Fig. 2.2B, we use OncoKB [34] to validate that the majority of high scoring self-pairs are indeed oncogenes. In contrast, SI Fig. B.9A shows that the majority of non-self gene pairs identified by the DepMap model that were not found by our model have not been experimentally verified. Similarly, SI Fig. B.9B shows that the DepMap model proposes far more self-pairs than our model with the majority of these self-pairs not appearing in OncoKB, suggesting that the DepMap model may be producing false positives.

To propose new candidate SL pairs, we filtered the top 92 scoring pairs to those that did not appear using the best model from DepMap. This resulted in 13 SL pairs (see Fig. 2.2C). Upon further investigation, seven of these pairs identified via our pipeline have already been experimentally verified [9], [35], [38], [39], [41], [42]. The remaining six pairs identified by our pipeline, RPS4X/ZFY, CHMP4B/CHMP4A, PELO/KLHL9, SCAP/MVK, SOX10/CDH19, and MED1/MED31 all present clear characteristics of SL pairs; namely, these knockouts have a select few top features, which are associated with reduced viability (see Fig. 2.2D). Of these pairs, we note that only one of these knockouts (RPS4X) formed an

experimentally verified SL pair when considering the top feature from the DepMap model. Indeed, DepMap identified the pair RPS4X/RPS4Y, which is a pair between an X chromosome encoded ribosomal protein gene and its Y chromosome encoded paralog [36]. SL-RFM identified RPS4Y as the second most important feature. Additionally, we note that for this knockout, nine out of the top ten ranking genes from our model are Y-linked genes, which upon their loss, are known to increase sensitivity of X-linked genes [36]. We also observe that for CHMP4B knockout and MED1 knockout, while the top features of DepMap and our model differ, the top two features of both models are CHMP4A and CHMP4C for CHMP4B knockout and MED31 and MED10 for MED1 knockout. Thus, we focus the following discussion on SCAP/MVK, PELO/KLHL9, and SOX10/CDH19 as potential novel candidate SL pairs.

MVK, or mevalonate kinase, is an essential enzyme in the mevalonate pathway, an important metabolic pathway that synthesizes isoprenoids for cellular processes [43]. Upregulation of mevalonate metabolism enhances both cancer development and the training of immunity cells [44]; thus a therapeutic challenge is to target the cancer cells without impairing the immunity cells. SCAP, or sterol regulatory element binding protein (SREBP) cleavage activating protein, regulates and chaperones genes SREBP-1 and SREBP-2, which regulate triglyceride and cholesterol levels in the body [45]. SREBP-2 appears to stimulate mevalonate metabolism [44]. β-catenin may bind with SREBP-2 to activate mevalonate genes and promote EMT towards an invasive cancer phenotype. On the other hand, SREBP2-dependent activation of mevalonate production seems to play a role in memory T cells, but the exact relationship is unclear [44]. Further research is needed to elucidate whether SCAP inhibition could decrease proliferation of cancer cells while enhancing immune cell response when mevalonate metabolism is upregulated.

Pelota mRNA surveillance and ribosome rescue factor (PELO) is a key gene in the cell meiotic division process. Recent work has shown that PELO and PLK1, an oncogene, bind to and degrade SMAD4, a tumor suppressor, in prostate cancer (PCa) [46]. PELO is highly expressed in PCa tissues and knockdown of PELO inhibits prostate cancer cell growth. On the other hand, kelch-like family member 9 (KLHL9) deletion is considered a driver of the mesenchymal subtype of Glioblastoma (GBM). There is a causal link between KLHL9 deletion and aberrant coactivation of transcription factors C/EBPβ, C/EBPα, and STAT3, which are key regulators of this subtype of GBM [47]. While they are individually implicated in the proliferation of tumors, existing literature has not yet determined the relationship between PELO and KLHL9 in tumors. Fig. 2.2D shows that both the RFM features and the PCC plot indicate that knockout of PELO under low expression of KLHL9 is correlated with low cellular viability.

SOX10, or SRY-box transcription factor 10, regulates the migration of neural crest (NC)-derived cells, which are essential in embryonic development [48], [49]. Recent work has shown that CDH19, or Cadherin 19, is a direct target of SOX10, and they work together to help develop NC cells into the enteric nervous system (ENS) during embryonic development [48]. However, NC cells can also develop into melanoma. SOX10 is considered an oncogene that is heterogeneously expressed in melanoma, and its deficiency is linked to a low proliferative/high invasive phenotype [49]. While less is known about the interaction between SOX10 and CDH19 in melanoma, there is evidence that CDH19 may be in a similar oncogenic pathway as SOX10. CDH19 is widely overexpressed in melanoma, and antibodies

targeting CDH19 cause tumor growth inhibition in melanoma cancer cell lines [50].

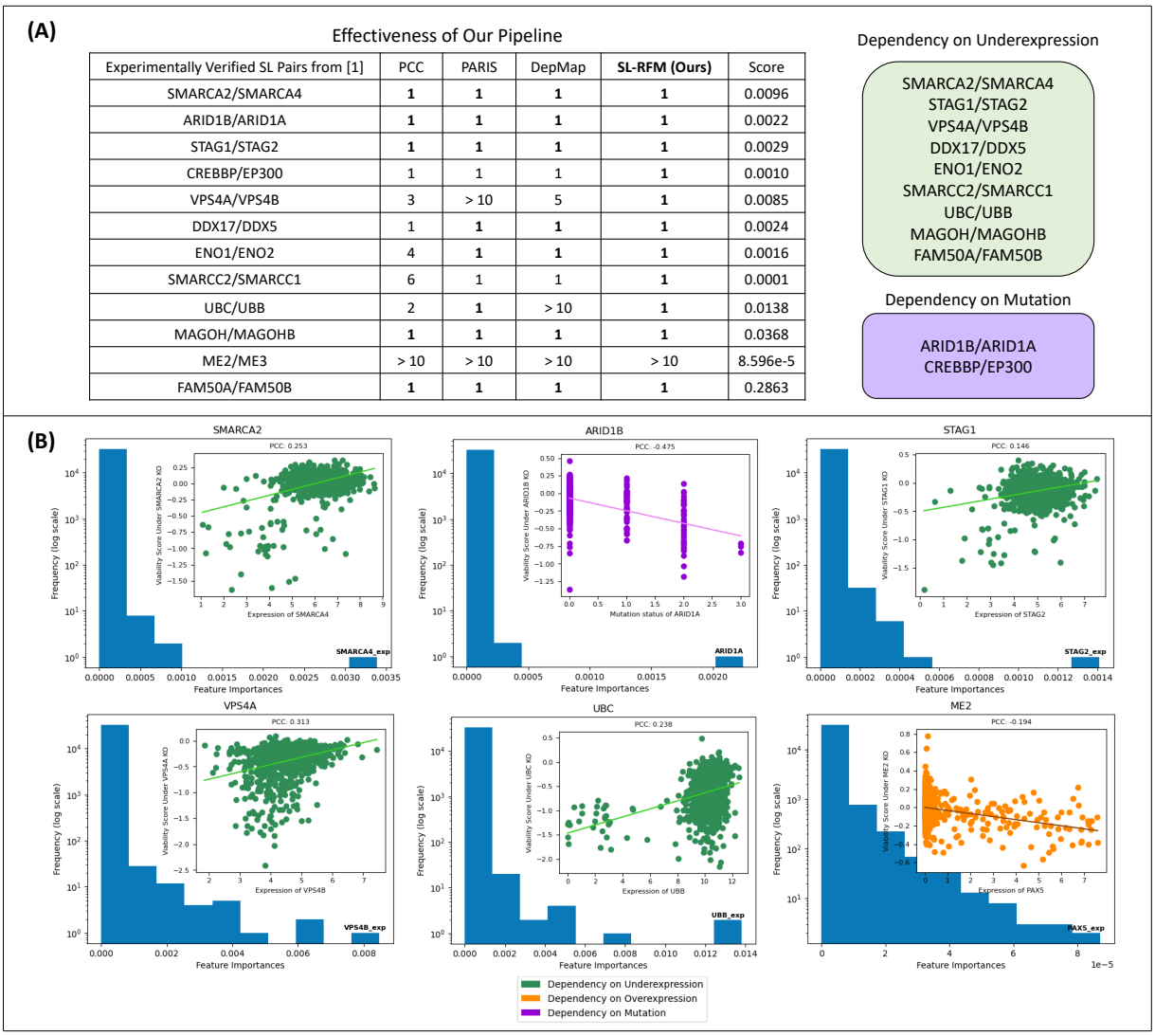
Interestingly, while the DepMap model identifies SOX10 as a self pair, SL-RFM identifies the following six genes as top features for predicting viability under SOX10 knockout: CDH19, ROPN1, EXTL1, SOX10, FGFBP2, and MIA. SI Fig. B.10 shows the expression of these genes against viability under SOX10 knockout as well as the joint expression of these genes with SOX10 in TCGA. We observe a clear relationship between these genes and low viability scores for melanoma cell lines. Moreover, the oncogene SOX10 is consistently overexpressed in TCGA melanoma samples; see SI Fig. B.10B. ROPN1, or Rhophilin Associated Tail Protein 1, produces a protein located in the fibrous sheath of sperm flagella [51]. Similar to CDH19, ROPN1 is an embryonic cell migration and neuronal development gene that is widely overexpressed in melanoma [52], [53]. MIA, or melanoma inhibitory activity, is used as a marker of melanoma and is a direct target gene of SOX10 [54]. Previous work has shown that SOX10 inhibition reduces MIA expression levels, which may be responsible for melanoma cell invasion [54]. There are some connections between SOX10, EXTL1, and FGFBP2 in the conjunctival epithelium. EXTL1 is a tumor suppressor that is among the top 5 most significantly upregulated factors in conjunctival melanoma [55]. FGFPB2, or Fibroblast Growth Factor Binding Protein 2, is part of the Fibroblast growth factor (FGF) system, and has been shown to be upstream of SOX9, which regulates SOX10, during the formation of ocular glands in embryonic development [56]. Reducing FGF signaling eliminated SOX10 expression, and knockout experiments in mice showed that SOX10 is essential for lacrimal gland development [56]. Another work [57] has shown that a different member of the EXT-family, EXTL2, controls FGF signaling in heparan sulfate biosynthesis. It is plausible that EXTL1 and FGFBP2 are in a pathway upstream of SOX10 that is implicated in conjunctival embryonic development and conjunctival melanoma. Overall, these experimental results suggest SOX10 as a highly attractive therapeutic target in melanoma.

2.5 Experimental data further validates selectivity of candidate SL pairs identified by our pipeline.

We first use DepMap data to demonstrate that the identified candidate SL pairs are selective across cancers. For example, if a knockout forms an SL pair based on under-expression of a gene, we expect to observe low viability in the cell lines that have low expression of the gene. On the other hand, if a knockout forms an SL pair based on over-expression of a gene, we expect to observe low viability in the cell lines that have high expression of the gene. To this end, we computed the product between the viability scores for a knockout and the expression of its suggested SL partner gene averaged across cell lines for a given cancer type. Fig. 2.3A shows the results for the top 92 SL pairs from our pipeline upon grouping cell lines by cancer type. This visualization highlights cancers that are predicted to be most susceptible to a given SL pair. In line with these results, somatic mutations of AXIN2, which cause over-expression of AXIN2, are associated with a higher risk of colorectal cancer and elimination of mutant CTNNB1 decreases clonogenicity of colorectal cancer cells [58], [59]. For the pair PELO/KLHL9 identified by SL-RFM in Fig. 2.2, which exhibited dependency on under-expression, the predicted susceptible cancers are Leukemia, Lung Cancer, and

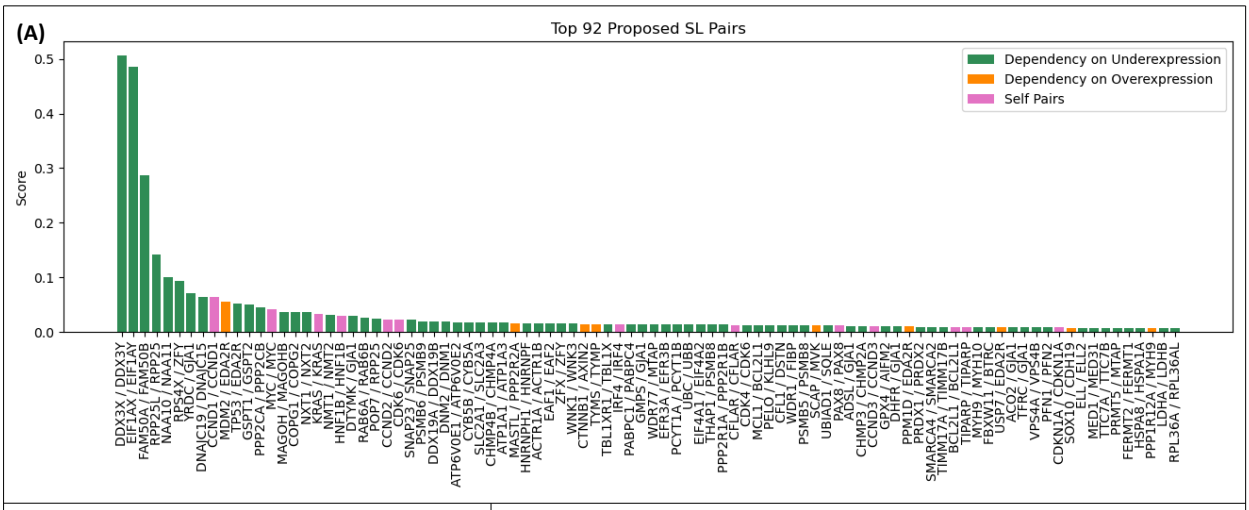
Pancreatic Cancer. For SCAP/MVK and SOX10/CDH19, which exhibited dependency on over-expression, the predicted susceptible cancers are Lung Cancer and Melanoma.

To further assess the relevance of the identified SL pairs, we used the 10,667 samples from TCGA to confirm that the simultaneous perturbation of genes in the predicted SL pairs does not occur in patients (see Methods). Analyzing first the 67 suggested SL pairs with dependency on under-expression, Fig. 2.3B, compares the percentage of TCGA samples exhibiting expression below a fixed level for these SL pairs as compared to 67 randomly sampled gene pairs (averaged over 10 random draws of pairs). By definition, both curves are monotonically increasing and obtain a maximum value of 100%. Yet, we observe that early on in the graph, there is a sizable nearly 40% gap between the two curves, indicating that patient samples less often exhibit simultaneous under-expression of SL gene pairs as compared to random gene pairs. Repeating this analysis for the 9 SL pairs with dependency on over-expression (i.e., comparing the percentage of TCGA samples exhibiting low expression of gene A and high expression of gene B for these predicted SL pairs as compared to 9 randomly sampled gene pairs (averaged over 10 random draws of pairs), we observe up to 10% gap between the two curves, indicating that there are fewer patient tumor samples with the simultaneous under-expression of gene A and over-expression of gene B in the candidate SL pairs compared to random gene pairs (see SI Fig. B.11). In particular, for the suggested candidate SL pairs SOX10/CDH19 and SCAP/MVK, where the dependency is on the knockout of the first gene and over-expression of the second gene, we observe that there are almost no samples with over-expression of CDH19/MVK and under-expression of SOX10/SCAP (see Fig. 2.3C and SI Fig. B.12). Similar plots based on GTEx data [60] are provided in SI Fig. B.13 and further corroborate these results. Overall, this analysis provides additional evidence that our pipeline can be used to automatically identifying potential SL pairs.



Experimentally Verified SL Pairs from [1]	PCC	PARIS	DepMap	**SL-RFM (Ours)**	Score
SMARCA2/SMARCA4	**1**	**1**	**1**	**1**	0.0096
ARID1B/ARID1A	**1**	**1**	**1**	**1**	0.0022
STAG1/STAG2	**1**	**1**	**1**	**1**	0.0029
CREBBP/EP300	1	1	1	**1**	0.0010
VPS4A/VPS4B	3	> 10	5	**1**	0.0085
DDX17/DDX5	1	**1**	**1**	**1**	0.0024
ENO1/ENO2	4	**1**	**1**	**1**	0.0016
SMARCC2/SMARCC1	6	1	1	**1**	0.0001
UBC/UBB	2	**1**	> 10	**1**	0.0138
MAGOH/MAGOHB	**1**	**1**	**1**	**1**	0.0368
ME2/ME3	> 10	> 10	> 10	> 10	8.596e-5
FAM50A/FAM50B	**1**	**1**	**1**	**1**	0.2863

Figure 2.1: Our pipeline (denoted SL-RFM) accurately recovers experimentally verified paralog SL pairs from [9], and the feature importances for the identified SL pairs are consistent with those expected from a single-index model. **(A)** SL-RFM outperforms (1) a Pearson correlation coefficient (PCC) baseline; (2) the PARIS random forest approach from [20]; and (3) the best model from the DepMap portal described in [9]. Values in the table indicate the rank of the SL pair out of $17,755$ possible gene pairs (lower is better with a minimum value of 1). The score column quantifies the difference between maximum feature importance and mean feature importance provided by our method for the knocked out gene in the SL pair. On the right, we group SL pairs by their dependence on expression or mutation features. **(B)** Plots of the feature importance distributions for six SL pairs from [9], with the top feature labelled. Each distribution of feature importances corresponding to an SL pair identified by SL-RFM indicates a top feature that is separated from the remaining features, which is consistent with that of a single-index model. On the other hand, for the pair ME2/ME3 not identified by SL-RFM, we do not observe a clear separation between the top feature and the remaining features. The insets show how the top feature varies with viability under the given knockout.



(C)

SL Pairs Within Our Top 92 that Differ From DepMap

Knockout	Our Top Feature	DepMap Rank of Our Feature	Score	Experimentally Verified
RPS4X	ZFY_exp	3	0.0936	
MAGOH	MAGOHB_exp	2	0.0368	✓
CHMP4B	CHMP4A_exp	2	0.0179	
MASTL	PPP2R2A_exp	2	0.0164	✓
UBC	UBB_exp	> 10	0.0138	✓
PPP2R1A	PPP2R1B_exp	5	0.0134	✓
PELO	KLHL9_exp	> 10	0.0123	
SCAP	MVK_exp	> 10	0.0118	
GPX4	AIFM2_exp	> 10	0.0104	✓
VPS4A	VPS4B_exp	> 10	0.0087	✓
SOX10	CDH19_exp	> 10	0.0079	
MED1	MED31_exp	2	0.0078	
FERMT2	FERMT1_exp	2	0.0075	✓

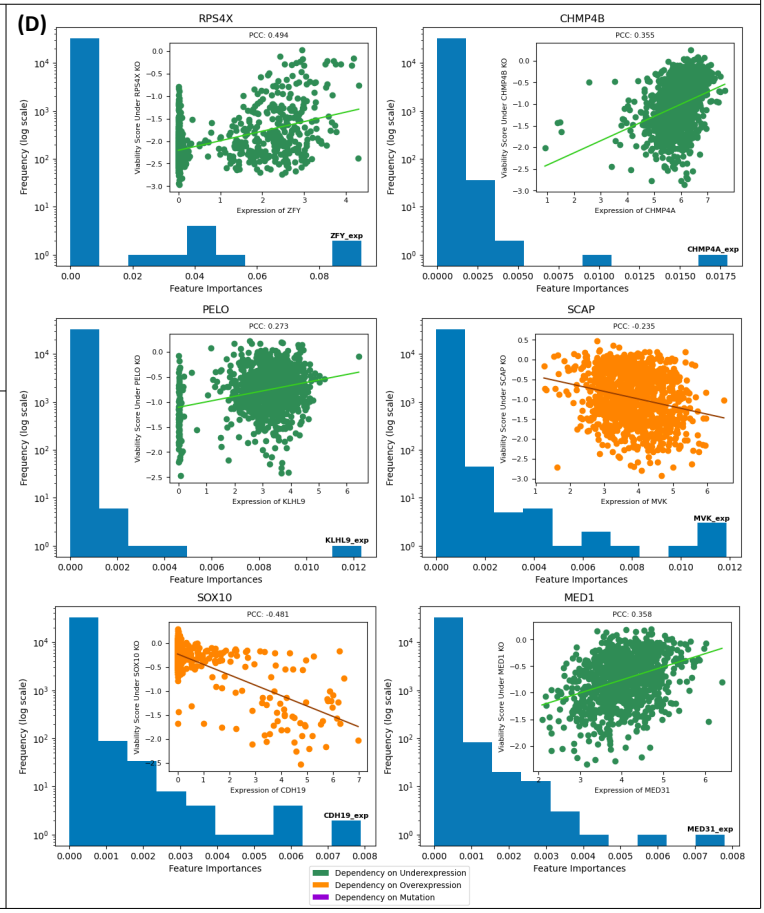


Figure 2.2: Analysis of top scoring SL pairs returned by our pipeline. **(A)** Visualization of the top 92 highest scoring SL pairs from SL-RFM and their corresponding top features, categorized by (1) if under-expression induces sensitivity to the knockout, (2) if over-expression induces sensitivity to the knockout, and (3) whether the two genes form a self pair. We note that expression features generally had more signal than mutation features, which is consistent with the findings of prior work [33]. **(B)** Validation that the majority of top scoring self-pairs identified by SL-RFM are oncogenes, based on OncoKB [34]. **(C)** A list of the SL pairs found among the top 92 highest scoring pairs from (A) for which the top feature differs from the top feature found using the best DepMap model. We observe that seven of these have been experimentally verified in prior work. (D) Distribution of feature importances for the remaining six candidate SL pairs identified from our analysis in (C) along with corresponding insets illustrating how the top feature varies with viability.

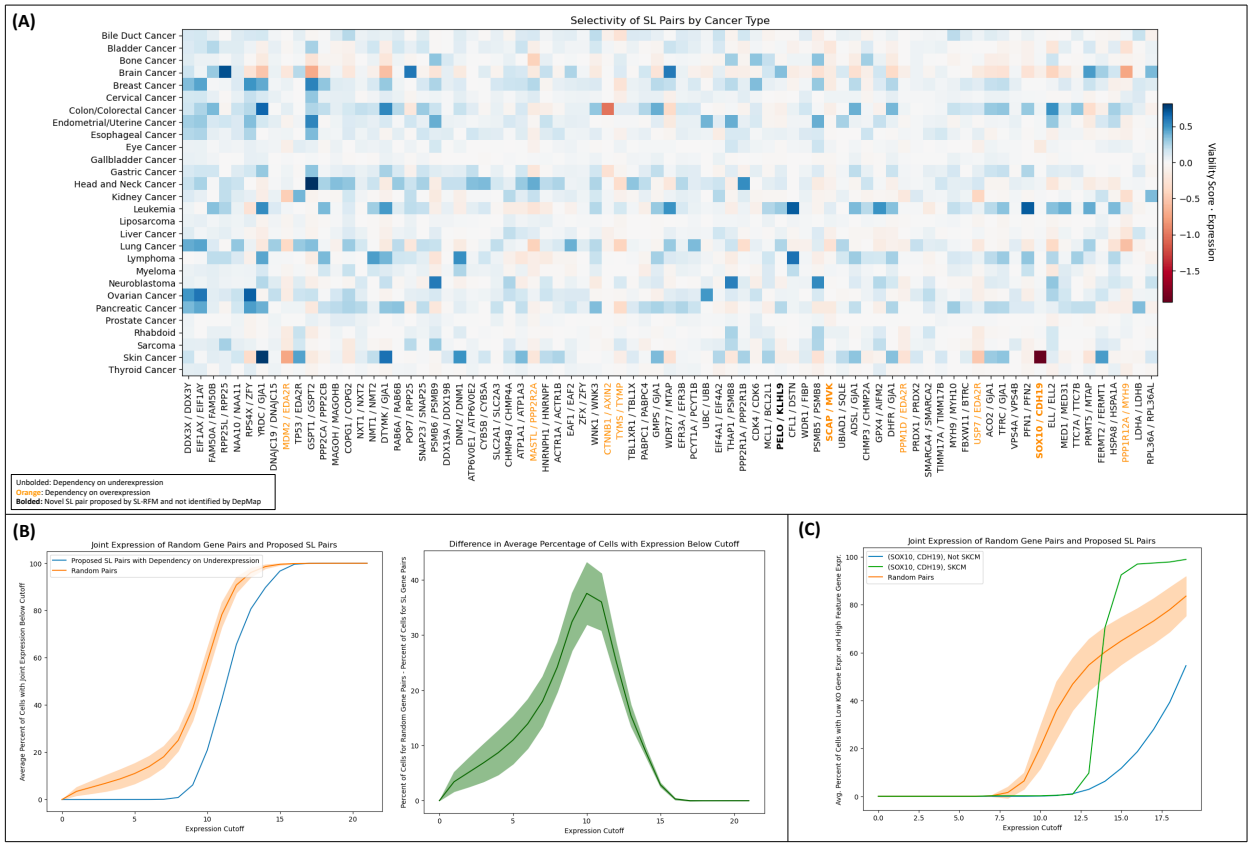


Figure 2.3: Existing experimental data further corroborates SL pairs suggested by our pipeline. **(A)** Heatmap of the predicted SL pairs and the average product between viability score of a knockout and the z-score of the expression of the top feature across DepMap cell lines aggregated by cancer type. For SL pairs with dependency on under-expression, we look for positive (blue) values. For SL pairs with dependency on over-expression, we look for negative (red) values. **(B)** Visualization validating that SL pairs with a dependency on under-expression are not simultaneously under-expressed in patient data from TCGA. Comparing the 67 out of 92 SL pairs with a dependency on under-expression that have data in TCGA (we omitted RPP25L/RPP25, COPG1/COPG2, and CHMP3/CHMP2A for this reason) to 67 randomly sampled gene pairs (sampled 10 times), we plot the average percentage of TCGA samples for which both genes have an expression below the given x-axis coordinate (error bars indicate 1 standard deviation). The curve on the right corresponds to the difference between the curves on the left. Overall, we observe that a substantial percentage of identified SL candidate pairs are never simultaneously under-expressed in patient samples. Analogous plots based on GTEx data are provided in SI Fig. B.13. (C) For the proposed pair SOX10/CDH19, we plot the average percentage of TCGA samples that have expression of SOX10 below the expression cutoff, c, and CDH19 expression above $20 - c$ and compare with the curve for randomly sampled gene pairs (sampled 10 times) and SOX10/CDH19 in melanoma (SCKM). These curves show that there is no simultaneous under-expression of SOX10 and over-expression of CDH19 in patient samples from TCGA. Gene expression for TCGA data used in plots (B, C) was transformed via log2(normalized count + 1).

Chapter 3

Summary and Discussion

The identification of synthetically lethal gene pairs is a promising approach for developing targeted treatments for cancer. Large-scale perturbation screens, such as the Cancer Dependency Map (DepMap), present exciting opportunities for screening SL pairs using machine learning. Despite the availability of such data, it has been difficult to identify clinically relevant SL pairs. A promising approach has been to frame SL pair screening as a feature learning problem, where the goal is to identify, for a given knockout, the corresponding expression or mutation features most indicative of low viability. Given that random forests have been the only (nonlinear) machine learning models that provide explicit feature importances, prior work, including the "best model" from DepMap and PARIS from [20], used random forests for SL screening. In recent work, we showed that kernel machines provide a more powerful approach for related screening tasks [26], [27], suggesting that such models may provide a more powerful alternative to random forests for SL pair screening. In this work, we built a computationally efficient and effective pipeline for SL pair screening by leveraging a recently developed class of feature learning kernel machines known as Recursive Feature Machines (RFMs). The construction of our pipeline is motivated by formulating the problem of SL pair screening as the prominent single-index model studied in the statistical literature, which can be solved using an RFM. We demonstrated that our pipeline more accurately recovers experimentally verified SL pairs than the "best model" from DepMap and previous state-of-the-art approaches based on random forests. Moreover, we applied our pipeline to identify novel candidate SL pairs. We identified as top candidates PELO/KLHL9, MVK/SCAP, and SOX10/CDH19, which were not found using prior approaches but have strong supporting evidence based on experimental data from DepMap and The Cancer Genome Atlas (TCGA). In the following, we discuss implications of our results and future extensions.

3.1 Identifying synthetically lethal gene groups.

In this work, we focused on the problem of identifying gene pairs, which when simultaneously perturbed lead to cell death, by formulating it as a single-index problem. An interesting next direction is to investigate whether our pipeline can also be used to identify patterns of groups of genes that are associated with low cellular viability under a given knockout; this represents a natural extension of single-index to multi-index modeling. Another interesting extension

is to consider combinatorial screens in which more than one gene is perturbed. While such settings are challenging since there is limited data on the effects of combinatorial perturbations on cells, we envision that our pipeline could serve as a useful step in determining candidate experiments for identifying novel SL gene interactions.

3.2 Flexible and scalable approach for screening.

The flexibility and computational efficiency of our pipeline, which takes less than 3 minutes to run on the full DepMap dataset on a single GPU, opens novel avenues for rapidly obtaining novel biological hypotheses of key features. In particular, while our current pipeline explores expression and mutation features, there are a number of other features including copy number, ploidy, and features from multimodal data that we could integrate to identify novel candidates of cancer vulnerabilities. We envision that integrating such features could provide novel insights into biological mechanisms associated with cancer and support the development of novel targeted treatments.

Appendix A

Methods

A.1 Overview of datasets and pre-processing

Below, we provide an outline of all datasets and data pre-processing utilized in this work.

- DepMap
  - The **viability** dataset is a 1078 cell lines $\times$ 17,453 knockout matrix where every entry is a number denoting the viability of the cell after the gene is knocked out.
  - The **gene expression** dataset contains a 19,194-dimensional gene expression (log2(TPM + 1)) vector for 1,408 cancer cell lines.
  - The **damaging mutations** dataset contains a 17,257-dimensional vector for 1,702 cancer cell lines with values in $[0, 1, 2]$, where 0 indicates no mutation, 1 indicates at least one heterozygous damaging mutation, and 2 indicates at least one homozygous damaging mutation.
  - The **hotspot mutations** dataset contains a 450-dimensional vector for 1,702 cancer cell lines with values in $[0, 1, 2]$, where 0 indicates no mutation, 1 indicates at least one heterozygous hotspot mutation, and 2 indicates at least one homozygous hotspot mutation.
- TCGA
  - The TCGA dataset contains 10667 samples from patients with 36 different cancers: Acute Myeloid Leukemia (LAML), Adrenocortical Cancer (ACC), Bile Duct Cancer (CHOL), Bladder Cancer (BLCA), Breast Cancer (BRCA), Cervical Cancer (CESC), Colon Cancer (COAD), Colon and Rectal Cancer (COADREAD), TCGA Endometrioid Cancer (UCEC), Esophageal Cancer (ESCA), Glioblastoma (GBM), Head and Neck Cancer (HNSC), Kidney Chromophobe (KICH), Kidney Clear Cell Carcinoma (KIRC), Kidney Papillary Cell Carcinoma (KIRP), Large B-cell Lymphoma (DLBC), Liver Cancer (LIHC), Lower Grade Glioma (LGG), Lower Grade Glioma and Glioblastoma (GBMLGG), Lung Adenocarcinoma (LUAD), Lung Cancer (LUNG), Lung Squamous Cell Carcinoma (LUSC),

Melanoma (SKCM), Mesothelioma (MESO), Ocular Melanomas (UVM), Ovarian Cancer (OV), Pancreatic Cancer (PAAD), Pheochromocytoma and Paraganglioma (PCPG), Prostate Cancer (PRAD), Rectal Cancer (READ), Sarcoma (SARC), Stomach Cancer (STAD), Testicular Cancer (TGCT), TCGA Thymoma (THYM), Thyroid Cancer (THCA), and Uterine Carcinosarcoma (UCS). These are all the cancers available on the Xena portal except for Pan-Cancer (PANCAN) and Formalin Fixed Paraffin-Embedded Pilot Phase II (FPPP). The gene expression dataset contains a 40,543-dimensional gene expression (log2(normalized count + 1)) vector and mutation dataset contains a 40,543 -dimensional binary vector indicating the presence of non-silent somatic mutation.

Each cell embedding is a concatenation of the DepMap gene expression vector and DepMap mutation vector, where the mutation vector contains a 1 if the cell has a damaging or hotspot mutation and 0 otherwise. The gene expression vector is z-scored across cell lines, so the mean gene expression for each gene is 0. The cell embedding is normalized to norm 1 under the ℓ_2 norm for each cell. For ease of downstream analysis, we only kept cell features that are present in both the DepMap and TCGA datasets. The final dimensions of the gene expression features and mutation features are 16,568 and 16,061, respectively. We also only kept cell lines that are present in both the viability and gene expression/mutation datasets. This leaves 998 cells. The final cell embedding is a $998 \times 32,629$-dimensional matrix.

A.2 Training details

We trained a RFM for each knockout that is trained to map cell line embeddings to viability scores for each knockout. Since the cell embeddings are the same per knockout, we trained RFMs efficiently by modeling this problem as a multi-output regression problem where the number of outputs are equal to the number of knockouts. When trained RFMs using a Laplace kernel as the base predictor, i.e., the kernel function is $K(x, \tilde{x}) = \exp(-L||x - \tilde{x}||_2)$. We directly solved the kernel regression problem with $L = 1$ and performed one iteration of feature-learning through the average gradient outer product. We used ridge regularization with a coefficient of 10^{-6} to avoid numerical issues with solving exactly. If $Y \in \mathbb{R}^{998\times 17453}$ is the viability matrix and $X \in \mathbb{R}^{998\times 32629}$ is the cell embedding, then the solution to our kernel regression problem is

$$\alpha = Y^\top (K(X, X) + 10^{-6} I)^{-1};$$

where $K(X, X)_{ij} = K(x_i, x_j)$ with x_i, x_j denoting embeddings of cell lines i, j. We then utilize the AGOP of the trained kernel machine to extract relevant features. In particular, for a cell line x and KO k, let the trained kernel machine be given by

$$f_k(x) = K(x, X)\alpha_k = \exp(-L||x - X||_2)\alpha_k \ .$$

Since we are primarily interested in feature importances, we compute the average magnitude of gradient entries of f_k. These are given by the diagonal of the AGOP as follows:

$$\tilde{g}_k = \frac{1}{n} \sum_{i=1}^{n} \nabla f_k(x_i) \odot \nabla f_k(x_i);$$

where $\odot$ denotes elementwise multiplication (Hadamard product). Note that feature importance alone is insufficient for identifying SL pairs. In particular, we must ensure that the top feature induces low cellular viability. To this end, we use Pearson correlation coefficient (PCC) to additionally re-weight features based on their association with low viability. Namely, for mutation features, the presence of a mutation should be negatively correlated with the viability score. For expression features, a strong positive correlation with viability indicates that the knockout is synthetically lethal with under-expression of the feature and a strong negative correlation indicates that the knockout is synthetically lethal with over-expression of the feature. To incorporate these relationships into our pipeline, we re-weight the feature importances according to

$$g_k = \tilde{g}_k \odot |c_k| \ .$$

where c_k is a $32,629$-dimensional vector where each entry is the PCC of the feature gene with viability under the knockout and coordinate i of c_k is 0 if coordinate i represents a mutation feature and the PCC of the feature is positive. Moreover, for entries of c_k corresponding to expression features, we only utilize features within 3 standard deviations of the mean to omit outliers.

A.3 Metrics for evaluating performance

To find the best kernel to use, we benchmarked the performance of Laplacian and Gaussian kernels to predict the viability scores for each knockout. The Laplacian kernel is defined as $K(x, \tilde{x}) = \exp(-L||x - \tilde{x}||_2)$ and Gaussian kernel is defined as $K(x, \tilde{x}) = \exp(-L||x - \tilde{x}||_2^2)$ for $L > 0$. For our experiments, we used $L = 1$ for both kernels. We also benchmark predicting the mean for each knockout over cell lines. We used 5-fold cross validation for the target task and reported the metrics computed across all folds in SI. Fig. B.1.

Let $C = 998$ denote the number of cell lines and $K = 17453$ denote the number of knockouts. Let $\hat{Y} \in \mathbb{R}^{C \times K}$ denote the predicted viabilities for cells under each knockout generated through 5-fold cross validation. Let $Y \in \mathbb{R}^{C \times K}$ denote the ground truth. Let $\bar{Y}^{(c)} = \frac{1}{K} \sum_{k=1}^{K} Y_k^{(c)}$ denote the average viability for cell line i. Let $\hat{\mathbf{y}}, \mathbf{y} \in \mathbb{R}^{CK}$ denote the vectorized versions of $\hat{Y}$ and Y respectively, and $\bar{\hat{\mathbf{y}}}, \bar{\mathbf{y}}$ their respective means. We use the same three metrics as those considered in [26], [27]. All evaluation metrics have a maximum value of 1 and are defined below.

1. Pearson R:
$$r = \frac{\langle \hat{\mathbf{y}} - \bar{\hat{\mathbf{y}}}, \mathbf{y} - \bar{\mathbf{y}} \rangle}{\|\hat{\mathbf{y}} - \bar{\hat{\mathbf{y}}}\|_2 \|\mathbf{y} - \bar{\mathbf{y}}\|_2} \ ;$$
2. Mean R^2:
$$R^2 = \frac{1}{C} \sum_{c=1}^{C} \left(1 - \frac{\|\hat{Y}^{(c)} - Y^{(c)}\|_2^2}{\|Y^{(c)} - \bar{Y}^{(c)}\|_2^2}\right) \ ;$$
3. Mean Cosine Similarity:
$$c = \frac{1}{C} \sum_{c=1}^{C} \frac{\langle \hat{Y}^{(c)}, Y^{(c)} \rangle}{\|\hat{Y}^{(c)}\| \|Y^{(c)}\|} \ .$$

In SI Fig. B.1B, we evaluate the prediction performance for computational simplicity in two specific cell lines, A549 and HUH7, using 5-fold cross validation. For each cell line, we obtain a list of knockouts sorted from most lethal to least lethal for the cell. We generate the same sorted list using predicted viability scores from the Laplacian kernel, mean over cell lines, and a random shuffle. We then plot the percentage of knockouts that overlap between these lists, indexed by the number of the most lethal knockouts. Laplacian kernel out-performs the other two benchmarks, demonstrating that SL-RFM effectively identifies the top most lethal knockouts for held out cell lines.

A.4 Score computation from feature importances

To propose candidate SL pairs, we look for knockouts for which feature importance distributions exhibit a gap between the most important feature and other features consistent with a single-index model. We thus define the following score to automatically identify such SL pairs. Let v_k be a vector of feature importances for the predictor f_k trained to predict viability scores of knockout k. We define

$$\text{score}(k) = \max(v_k) - \text{mean}(v_k) \ .$$

We then sort the knockouts by their scores. The knockouts with the highest scores and the genes associated with the top feature for these knockouts are proposed as candidate SL pairs.

A.5 Data Availability

All datasets considered in this work are publicly available. The CRISPR-Cas9 viability screens
(`CRISPRGeneEffects.csv`), cell line expression (`OmicsExpressionProteinCodingGenesTPMLogp1.csv`), cell line mutation data (`OmicsSomaticMutations.csv`), cell line (`Model.csv`), and DepMap feature importances (`Chronos_Combined_predictability_results.csv`) were downloaded from DepMap version 22Q4. The TCGA data was downloaded from the UCSC Xena datahub (tcga.xenahubs.net), and the GTEx data is Gene TPM data from GTEx Analysis V8 (`GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz`). A list of oncogenes was obtained from OncoKB at https://www.oncokb.org/cancer-genes.

A.6 Code Availability

The code is available at https://github.com/uhlerlab/synthetic_lethality.

Appendix B

Supporting Information

(A)

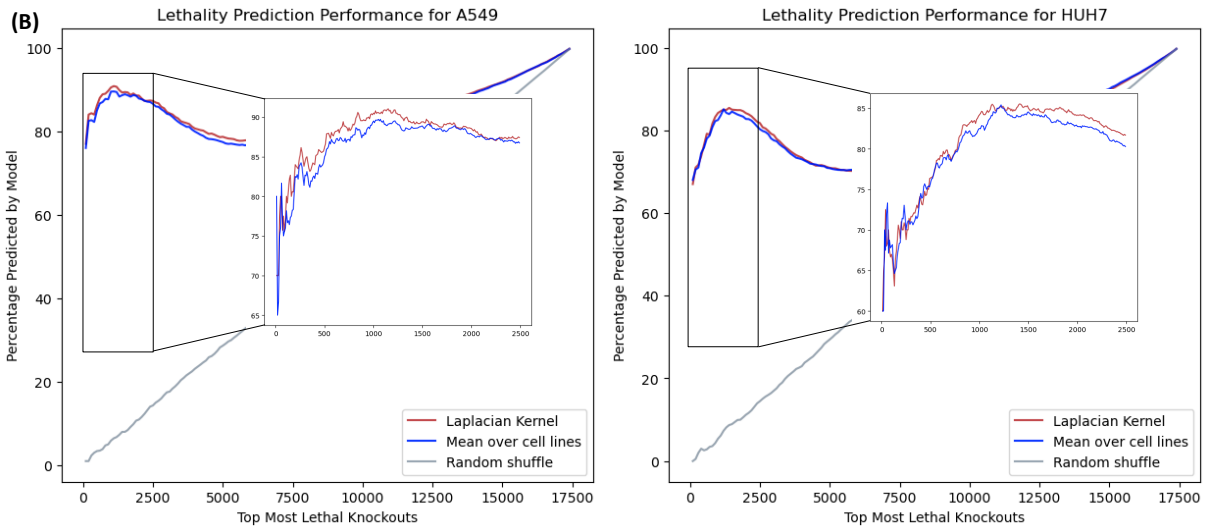
Comparing Performance in Predicting Cell Viability, All Genes

	Laplacian Kernel	Gaussian Kernel	Mean Over Cell Type
Pearson R	**0.935**	0.933	0.929
R^2	**0.875 ± 0.040**	0.872 ± 0.041	0.865 ± 0.041
Cosine Similarity	**0.943 ± 0.019**	0.942 ± 0.020	0.938 ± 0.020

Comparing Performance in Predicting Cell Viability, SL Genes

	Laplacian Kernel	Gaussian Kernel	Mean Over Cell Type
Pearson R	0.742	**0.744**	0.664
R^2	**0.522 ± 0.209**	0.522 ± 0.216	0.413 ± 0.227
Cosine Similarity	0.840 ± 0.072	**0.841 ± 0.072**	0.797 ± 0.083

(B)

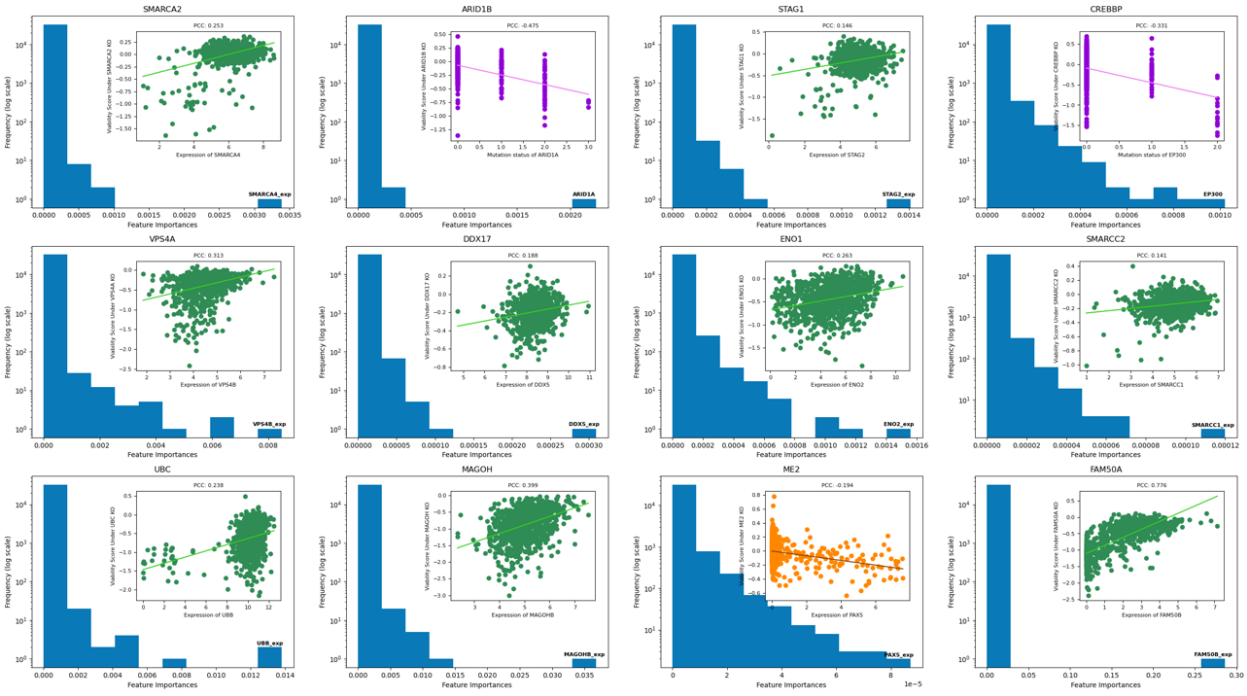


SI Figure B.1: Comparison of model performance in predicting viability data. All metrics are described in Methods. **(A)** Comparison of RFM with varying base kernels (Laplace and Gaussian kernel) on predicting cell viability, where test cell lines are held out using 5-fold cross validation. **(B)** Analyzing predictions for individual cell lines from five-fold cross-validation. We compare the list of predicted knockouts and the list of ground truth knockouts sorted by viability score for sample cell lines (A549 and HUH7). We observe that RFM with the Laplace kernel base predictor outperforms both a mean over cell line benchmark and a random baseline, illustrating the effectiveness of our selected model on held-out test data.

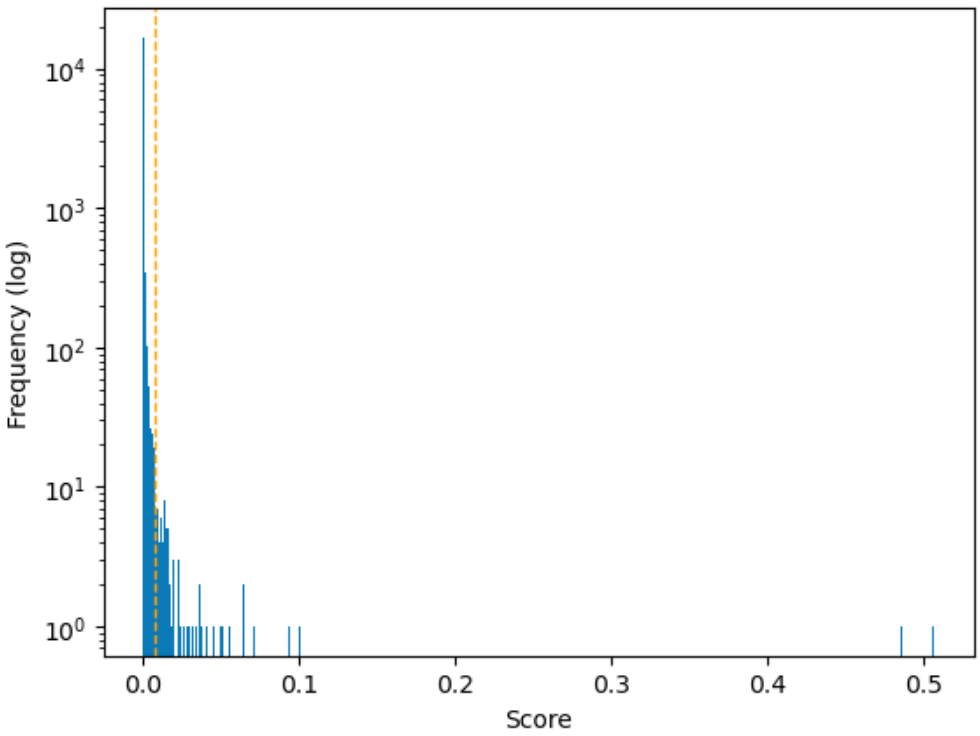
Effectiveness of Our Pipeline

Experimentally Verified SL Pairs from [1]	PCC	PCC	PARIS	PARIS	DepMap	DepMap	**SL-RFM (Ours)**	**SL-RFM (Ours)**
SMARCA2/SMARCA4	**1**	**1**	**1**	647	**1**	**1**	**1**	**1**
ARID1B/ARID1A	**1**	7837	**1**	419	**1**	> 10	**1**	228
STAG1/STAG2	**1**	**1**	**1**	3133	**1**	**1**	**1**	**1**
CREBBP/EP300	**1**	5077	**1**	435	**1**	> 10	**1**	14578
VPS4A/VPS4B	3	10	3471	4150	5	5	**1**	**1**
DDX17/DDX5	**1**	**1**	**1**	1071	**1**	**1**	**1**	**1**
ENO1/ENO2	4	1690	**1**	5070	**1**	> 10	**1**	8378
SMARCC2/SMARCC1	6	822	**1**	1577	**1**	> 10	**1**	1791
UBC/UBB	2	N/A	**1**	N/A	> 10	> 10	**1**	N/A
MAGOH/MAGOHB	**1**	15	**1**	5200	**1**	> 10	**1**	2
ME2/ME3	14850	15640	3307	4944	> 10	> 10	15938	16959
FAM50A/FAM50B	**1**	3	**1**	530	**1**	**1**	**1**	**1**

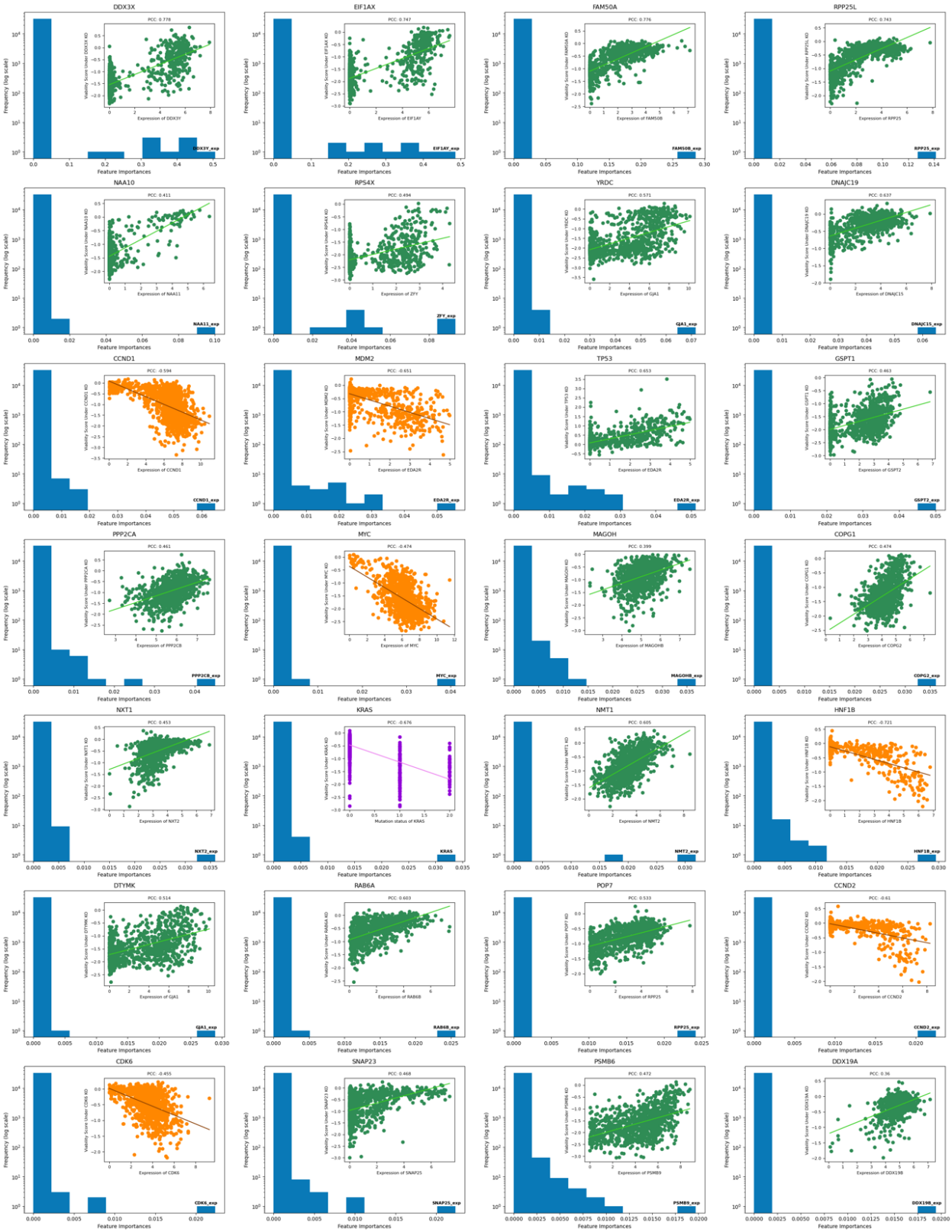
SI Figure B.2: Ranks of all models from Fig. 2.1 when considering the first gene in the pair as the knockout (white) or the second gene in the pair as the knockout (gray). The minimum rank between white and gray columns is reported in Fig. 2.1. We note that DepMap only provides the top 10 most important features for prediction. If the SL pair gene does not appear among the top 10, we denote the rank as > 10. All other models return the full set of feature importances.



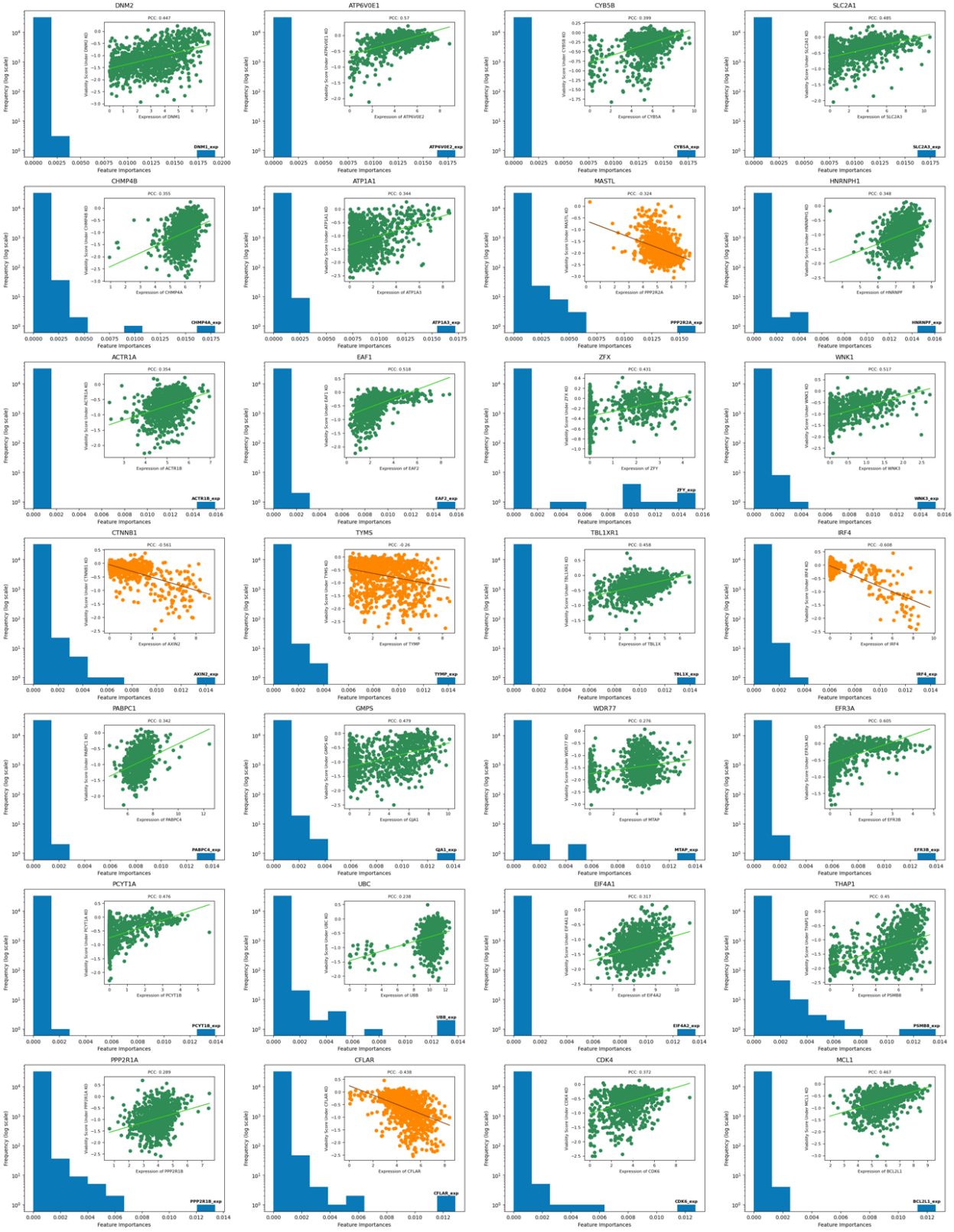
SI Figure B.3: SL-RFM feature importance plots for knockouts part of SL pairs from [9].



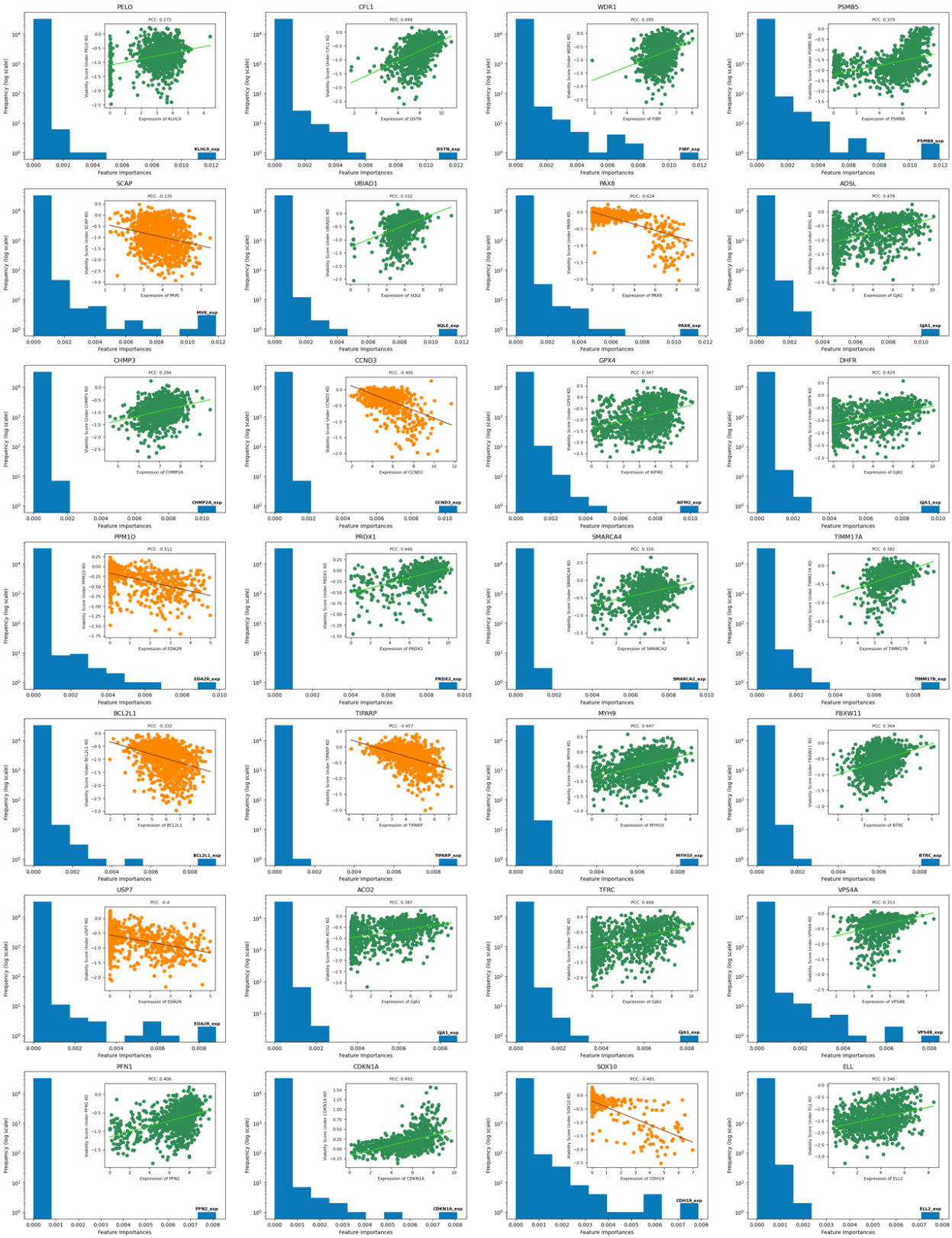
SI Figure B.4: The distribution of scores for the knockouts. To suggest candidate SL pairs, we utilize a score cutoff of 0.0071 (shown as a dashed vertical line). This results in 92 knockouts.



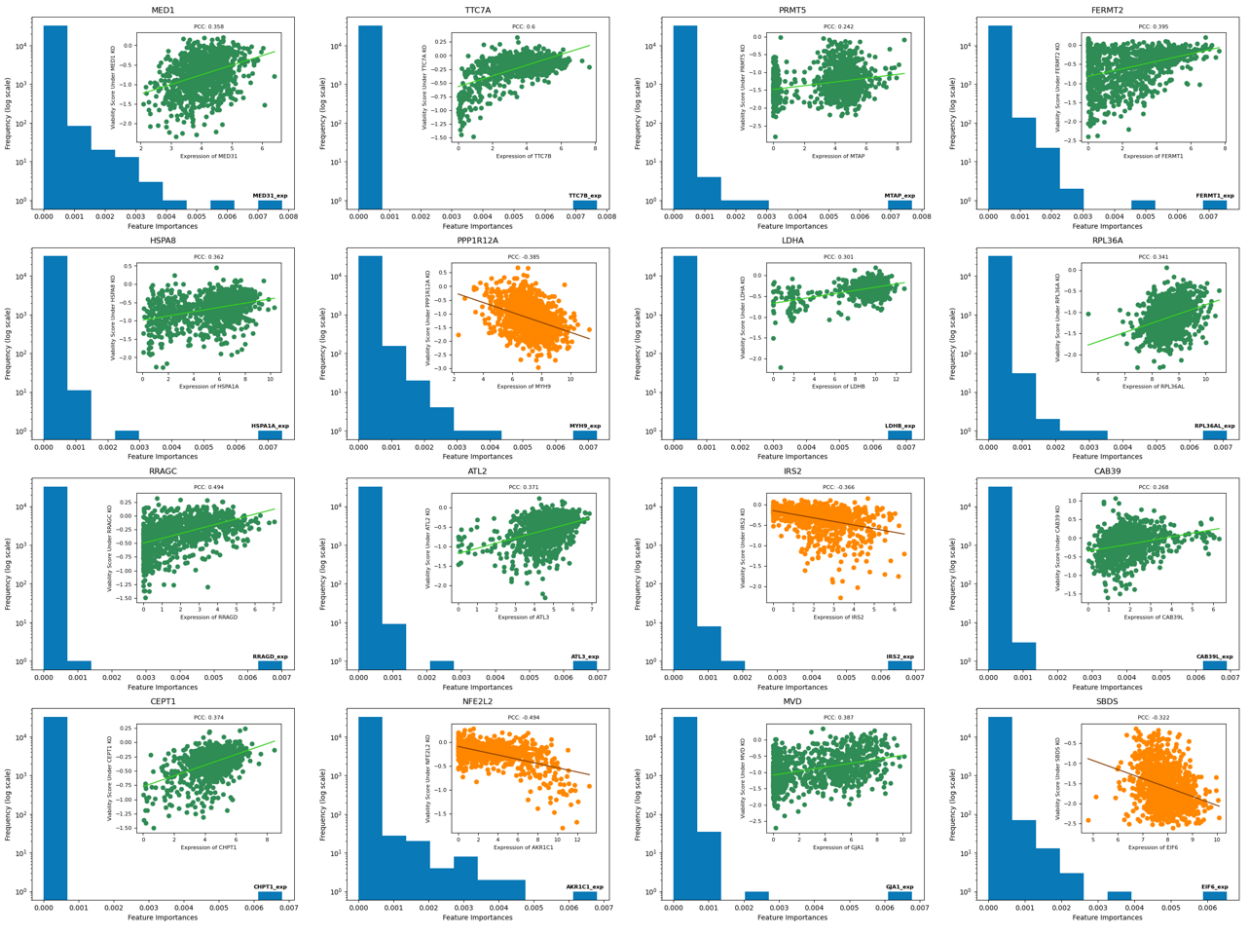
SI Figure B.5: SL-RFM feature importance plots 1-28 of our proposed SL pairs.



SI Figure B.6: SL-RFM feature importance plots 29-56 of our proposed SL pairs.



SI Figure B.7: SL-RFM feature importance plots 57-84 of our proposed SL pairs.

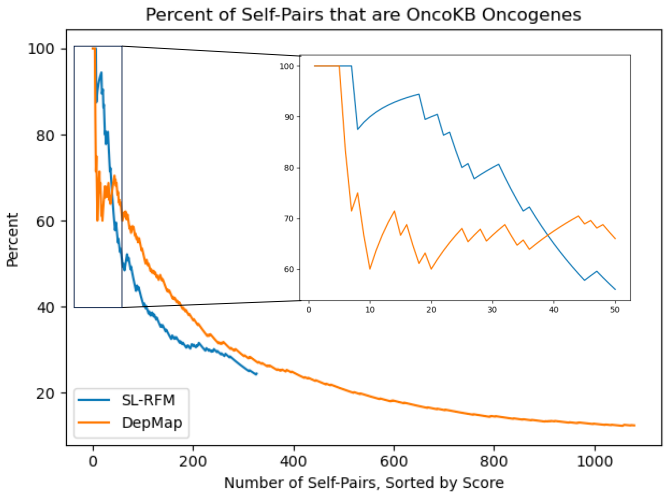


SI Figure B.8: SL-RFM feature importance plots 85-100 of our proposed SL pairs.

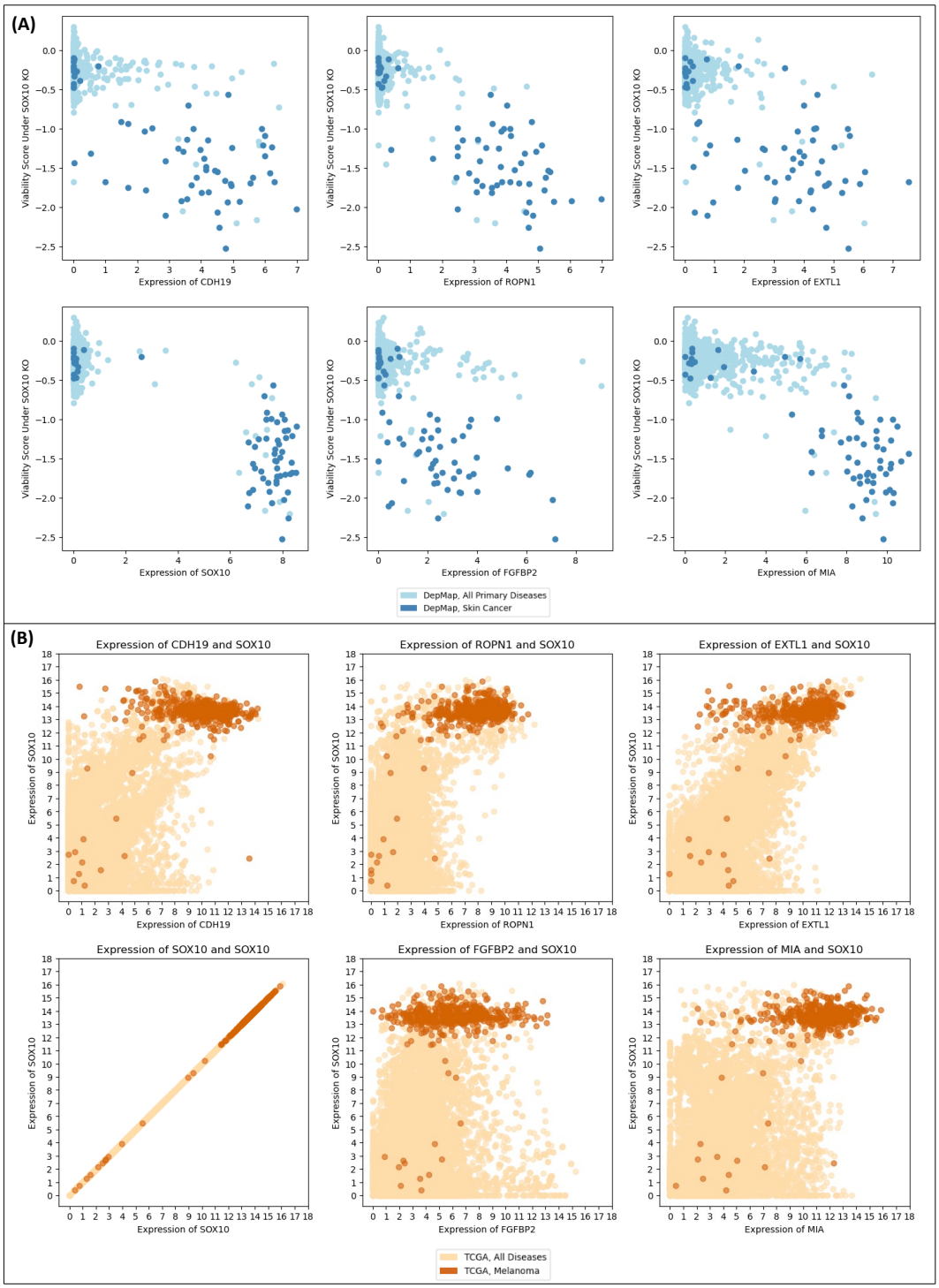
(A) Non-Self SL Pairs Within DepMap Top 100 that Differ From Our Predictions

Knockout	DepMap Top Feature	Rank of DepMap feature using our Method	Our Top Feature	Rank of Our Top Feature in DepMap	Rank Among Top 100	Experimentally Verified?
POU2AF2	TAS1R3	> 10	TRPM5	> 10	24	
SLF1	SIMC1	> 10	PNLIP	> 10	51	
FERMT2	SYDE1	> 10	FERMT1	2	75	
TCF7L2	RNF43	5	AXIN2	4	78	
CHMP4B	CHMP4C	2	CHMP4A	2	81	
EXOC6	EXOC6B	4	OR2T11	> 10	89	✓
PAX5	CD19	> 10	SH2D4B	> 10	96	

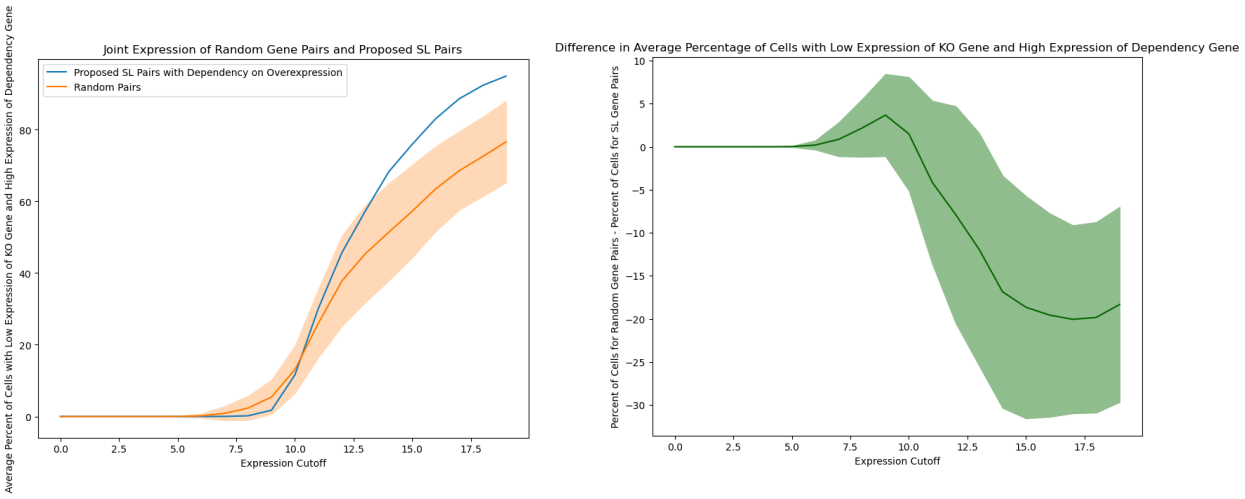
(B)



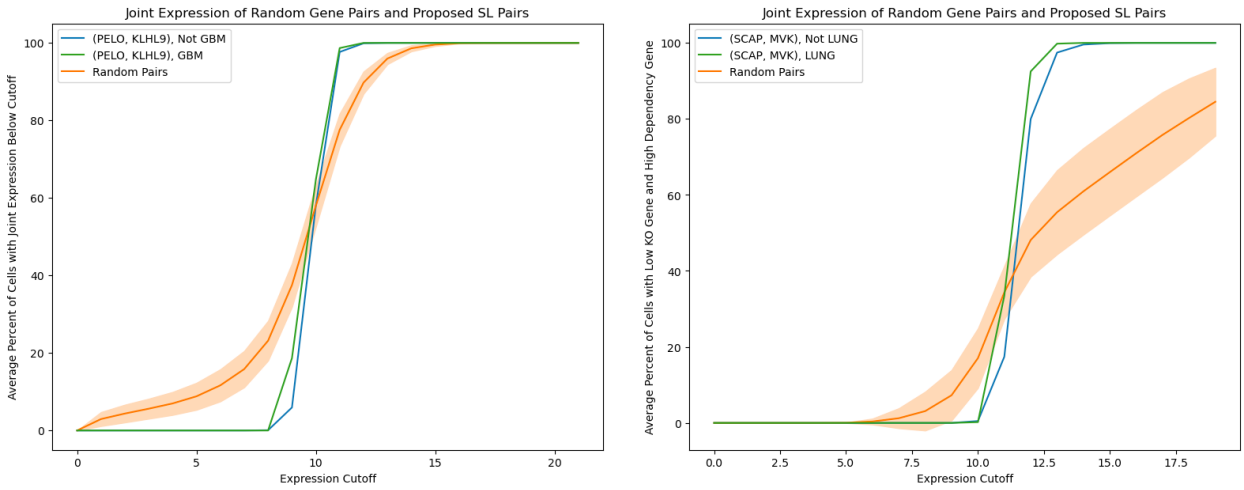
SI Figure B.9: Comparison of the SL pairs suggested by our pipeline to those suggested by DepMap. **(A)** A list of the non-self SL pairs out of the top 100 pairs proposed by the best model from DepMap that differ from pairs proposed by our pipeline. We sort DepMap pairs by score, as defined by difference between max feature importance and mean feature importance for a given knockout. Since DepMap only published the feature importances of the top 10 features for each knockout, the score was calculated using the 10 provided features. Note that only one out of ten of the top pairs of DepMap not found using our method are experimentally verified. **(B)** Comparison of the percent of self-pairs proposed by our method and DepMap that are verified oncogenes in OncoKB. We observe that the model from DepMap proposes many more self-pairs than our method, but that most of these pairs are not verified oncogenes in OncoKB.



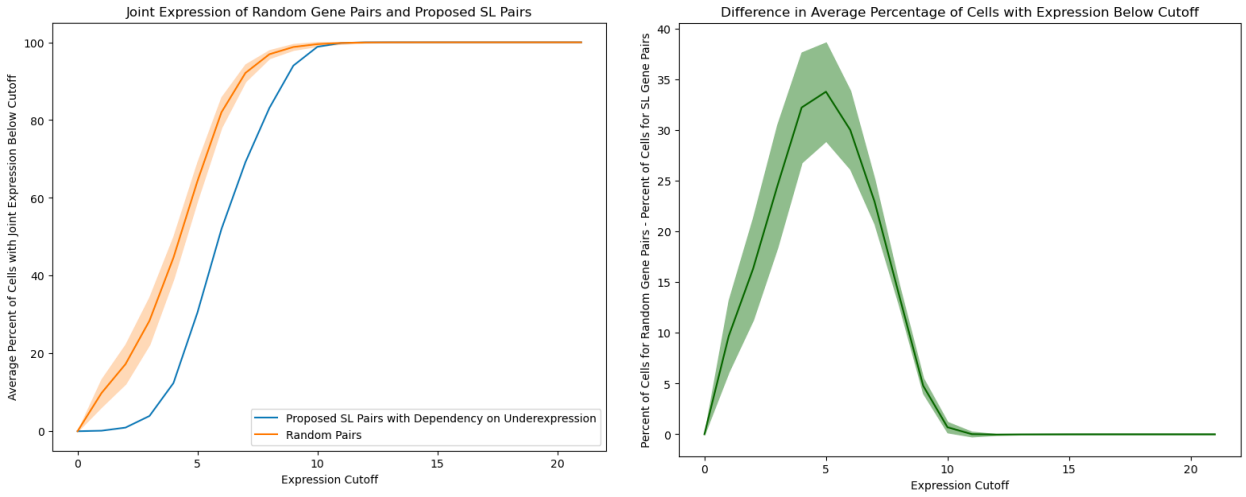
SI Figure B.10: **(A)** A visualization of the relationship between expression of genes associated with the top six most important features for predicting viability scores under knockout of SOX10 using SL-RFM. The majority of cell lines with over-expression of the targets are derived from skin cancer and have low viability under the SOX10 knockout. **(B)** Joint expression of these genes and SOX10 in TCGA. The majority of melanoma samples have SOX10 overexpressed. These analyses suggest that knocking out SOX10 causes low viability for cells with over-expression of the genes associated with these top features, which occurs frequently in melanoma.



SI Figure B.11: Visualization showing that for SL pairs of the form (A, B) with dependency on knockout of gene A and over-expression of gene B, there are few TCGA samples exhibiting low expression of gene A and high expression of gene B. For 9 randomly sampled gene pairs (sampled 10 times) and the 9 SL pairs with dependency on over-expression of the form (A, B), we plot the average percentage of TCGA samples that have expression of the gene A below the expression cutoff, c, and the expression of the gene B above $20 - c$. In the figure on the right, we plot the difference between the two curves. For $c < 10$, there are up to 10% fewer TCGA samples with low expression of gene A and high expression of gene B than random samples. As c increases, this relationship flips since there can be many samples for which gene B is under-expressed.



SI Figure B.12: Visualization showing that for the SL pairs we propose, there are fewer TCGA samples with expression patterns corresponding to proposed synthetically lethal gene interactions than random pairs. The dependency of PELO/KLHL9 is on under-expression, so we validate that there are fewer TCGA samples with the under-expression of both genes than random pairs. The dependencies of SCAP/MVK and SOX10/CDH19 are on over-expression, so we validate that for expression cutoff below 10, there are fewer TCGA samples with the under-expression of SCAP/SOX10 and over-expression of MVK/CDH19. For SCAP/MVK and SOX10/CDH19, we plot the average percentage of TCGA samples that have expression of SCAP/SOX10 below the expression cutoff, c, and the expression of MVK/CDH19 above $20 - c$. We also plot each of these curves upon stratifying the TCGA samples by cancer type based on our analysis in Fig. 2.3. Namely, the susceptible cancer type for PELO/KLHL9 is glioblastoma (GBM), the type for SCAP/MVK is lung cancer (LUNG), and the type for SOX10/CDH19 is skin cancer (SCKM).



SI Figure B.13: Visualization showing that SL genes with a dependency on under-expression are not simultaneously under-expressed in GTEx data. The GTEx TPM expression data was normalized as $\log_2(\text{TPM}+1)$. For 67 randomly sampled gene pairs (sampled 10 times) and the 67 out of 92 SL pairs with a dependency on under-expression and with data in TCGA (we omitted RPP25L/RPP25, COPG1/COPG2, and CHMP3/CHMP2A for this reason), we plot the percentage of GTEx samples for which both genes have an expression below the given x-axis coordinate. On the right, we plot the difference of the percentage of GTEx samples with expression of randomly sampled gene pairs below the cutoff and the percentage of GTEx samples with expression of SL pairs with expression below the cutoff. Overall, we observe up to a 40% difference, indicating that our candidate pairs are almost never simultaneously under-expressed in GTEx samples.

References

[1] S. M. Nijman, "Synthetic lethality: General principles, utility and detection using genetic screens in human cells," *FEBS letters*, vol. 585, no. 1, pp. 1–6, 2011.

[2] N. J. O'Neil, M. L. Bailey, and P. Hieter, "Synthetic lethality and cancer," *Nature Reviews Genetics*, vol. 18, no. 10, pp. 613–623, 2017.

[3] C. J. Lord and A. Ashworth, "PARP inhibitors: Synthetic lethality in the clinic," *Science*, vol. 355, no. 6330, pp. 1152–1158, 2017.

[4] Broad DepMap, "DepMap 22Q4 Public," Dec. 2022. DOI: 10.6084/m9.figshare.21637199.v2. URL: https://figshare.com/articles/dataset/DepMap_22Q4_Public/21637199.

[5] M. Costanzo, B. VanderSluis, E. N. Koch, A. Baryshnikova, C. Pons, G. Tan, W. Wang, M. Usaj, J. Hanchard, S. D. Lee, *et al.*, "A global genetic interaction network maps a wiring diagram of cellular function," *Science*, vol. 353, no. 6306, aaf1420, 2016.

[6] A. DeLuna, K. Vetsigian, N. Shoresh, M. Hegreness, M. Colón-González, S. Chao, and R. Kishony, "Exposing the fitness contribution of duplicated genes," *Nature genetics*, vol. 40, no. 5, pp. 676–681, 2008.

[7] E. J. Dean, J. C. Davis, R. W. Davis, and D. A. Petrov, "Pervasive and persistent redundancy among duplicated genes in yeast," *PLoS genetics*, vol. 4, no. 7, e1000113, 2008.

[8] G. Musso, M. Costanzo, M. Huangfu, A. M. Smith, J. Paw, B.-J. San Luis, C. Boone, G. Giaever, C. Nislow, A. Emili, *et al.*, "The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast," *Genome research*, vol. 18, no. 7, pp. 1092–1099, 2008.

[9] B. De Kegel, N. Quinn, N. A. Thompson, D. J. Adams, and C. J. Ryan, "Comprehensive prediction of robust synthetic lethality between paralog pairs in cancer cell lines," *Cell Systems*, vol. 12, no. 12, pp. 1144–1159, 2021.

[10] J. Huang, M. Wu, F. Lu, L. Ou-Yang, and Z. Zhu, "Predicting synthetic lethal interactions in human cancers using graph regularized self-representative matrix factorization," *BMC bioinformatics*, vol. 20, no. 19, pp. 1–8, 2019.

[11] R. Cai, X. Chen, Y. Fang, M. Wu, and Y. Hao, "Dual-dropout graph convolutional network for predicting synthetic lethality in human cancers," *Bioinformatics*, vol. 36, no. 16, pp. 4458–4465, 2020.

[12] G. Benstead-Hume, X. Chen, S. R. Hopkins, K. A. Lane, J. A. Downs, and F. M. Pearl, "Predicting synthetic lethal interactions using conserved patterns in protein interaction networks," *PLoS computational biology*, vol. 15, no. 4, e1006888, 2019.

[13] A. Jacunski, S. J. Dixon, and N. P. Tatonetti, "Connectivity homology enables inter-species network models of synthetic lethality," *PLoS computational biology*, vol. 11, no. 10, e1004506, 2015.

[14] J. Guo, H. Liu, and J. Zheng, "SynLethDB: Synthetic lethality database toward discovery of selective and sensitive anticancer drug targets," *Nucleic acids research*, vol. 44, no. D1, pp. D1011–D1017, 2016.

[15] F. Wan, S. Li, T. Tian, Y. Lei, D. Zhao, and J. Zeng, "Exp2sl: A machine learning framework for cell-line-specific synthetic lethality prediction," *Frontiers in pharmacology*, vol. 11, p. 112, 2020.

[16] A. Subramanian, R. Narayan, S. M. Corsello, *et al.*, "A next generation connectivity map: L1000 platform and the first 1,000,000 profiles," *Cell*, vol. 171, no. 6, pp. 1437–1452, 2017.

[17] M. Zamanighomi, S. S. Jain, T. Ito, D. Pal, T. P. Daley, and W. R. Sellers, "GEMINI: A variational bayesian approach to identify genetic interactions from combinatorial CRISPR screens," *Genome biology*, vol. 20, pp. 1–10, 2019.

[18] L. Jerby-Arnon, N. Pfetzer, Y. Y. Waldman, L. McGarry, D. James, E. Shanks, B. Seashore-Ludlow, A. Weinstock, T. Geiger, P. A. Clemons, *et al.*, "Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality," *Cell*, vol. 158, no. 5, pp. 1199–1209, 2014.

[19] S. Srivatsa, H. Montazeri, G. Bianco, M. Coto-Llerena, M. Marinucci, C. K. Ng, S. Piscuoglio, and N. Beerenwinkel, "Discovery of synthetic lethal interactions from large-scale pan-cancer perturbation screens," *Nature communications*, vol. 13, no. 1, p. 7748, 2022.

[20] S. Benfatto, Ö. Serçin, F. R. Dejure, A. Abdollahi, F. T. Zenke, and B. R. Mardin, "Uncovering cancer vulnerabilities by machine learning prediction of synthetic lethality," *Molecular Cancer*, vol. 20, no. 1, p. 111, 2021.

[21] A. Radhakrishnan, D. Beaglehole, P. Pandit, and M. Belkin, "Feature learning in neural networks and kernel machines that recursively learn features," *arXiv preprint arXiv:2212.13881*, 2022.

[22] A. Jacot, F. Gabriel, and C. Hongler, "Neural Tangent Kernel: Convergence and generalization in neural networks," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Curran Associates, Inc., 2018.

[23] S. Trivedi, J. Wang, S. Kpotufe, and G. Shakhnarovich, "A consistent estimator of the expected gradient outerproduct.," in *UAI*, 2014, pp. 819–828.

[24] W. Härdle and T. M. Stoker, "Investigating smooth multiple regression by the method of average derivatives," *Journal of the American statistical Association*, vol. 84, no. 408, pp. 986–995, 1989.

[25] Y. Xia, H. Tong, W. K. Li, and L.-X. Zhu, "An adaptive estimation of dimension reduction space," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 3, pp. 363–410, 2002.

[26] A. Radhakrishnan, G. Stefanakis, M. Belkin, and C. Uhler, "Simple, fast, and flexible framework for matrix completion with infinite width neural networks," *Proceedings of the National Academy of Sciences*, vol. 119, no. 16, e2115064119, 2022.

[27] A. Radhakrishnan, M. Ruiz Luyten, N. Prasad, and C. Uhler, "Transfer learning with kernel methods," *Nature Communications*, vol. 14, no. 1, p. 5570, 2023.

[28] M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny, "Structure adaptive approach for dimension reduction," *Annals of Statistics*, pp. 1537–1566, 2001.

[29] A. Damian, J. Lee, and M. Soltanolkotabi, "Neural networks can learn representations with gradient descent," in *Conference on Learning Theory*, PMLR, 2022, pp. 5413–5452.

[30] J. Ba, M. A. Erdogdu, T. Suzuki, Z. Wang, D. Wu, and G. Yang, "High-dimensional asymptotics of feature learning: How one gradient step improves the representation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 37 932–37 946, 2022.

[31] A. Bietti, J. Bruna, C. Sanford, and M. J. Song, "Learning single-index models with shallow neural networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9768–9783, 2022.

[32] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The cancer genome atlas pan-cancer analysis project," *Nature genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

[33] J. M. Dempster, J. M. Krill-Burger, J. M. McFarland, A. Warren, J. S. Boehm, F. Vazquez, W. C. Hahn, T. R. Golub, and A. Tsherniak, "Gene expression has more power for predicting in vitro cancer cell vulnerabilities than genomics," *BioRxiv*, pp. 2020–02, 2020.

[34] D. Chakravarty, J. Gao, S. Phillips, R. Kundra, H. Zhang, J. Wang, J. E. Rudolph, R. Yaeger, T. Soumerai, M. H. Nissan, *et al.*, "OncoKB: A precision oncology knowledge base," *JCO precision oncology*, vol. 1, pp. 1–16, 2017.

[35] P. C. Parrish, J. D. Thomas, A. M. Gabel, S. Kamlapurkar, R. K. Bradley, and A. H. Berger, "Discovery of synthetic lethal and tumor suppressor paralog pairs in the human genome," *Cell reports*, vol. 36, no. 9, 2021.

[36] A. Köferle, A. Schlattl, A. Hörmann, V. Thatikonda, A. Popa, F. Spreitzer, M. C. Ravichandran, V. Supper, S. Oberndorfer, T. Puchner, *et al.*, "Interrogation of cancer gene dependencies reveals paralog interactions of autosome and sex chromosome-encoded genes," *Cell reports*, vol. 39, no. 2, 2022.

[37] G. V. Kryukov, F. H. Wilson, J. R. Ruth, J. Paulk, A. Tsherniak, S. E. Marlow, F. Vazquez, B. A. Weir, M. E. Fitzgerald, M. Tanaka, *et al.*, "MTAP deletion confers enhanced dependency on the PRMT5 arginine methyltransferase in cancer cells," *Science*, vol. 351, no. 6278, pp. 1214–1218, 2016.

[38] K. Bersuker, J. M. Hendricks, Z. Li, L. Magtanong, B. Ford, P. H. Tang, M. A. Roberts, B. Tong, T. J. Maimone, R. Zoncu, *et al.*, "The CoQ oxidoreductase FSP1 acts parallel to GPX4 to inhibit ferroptosis," *Nature*, vol. 575, no. 7784, pp. 688–692, 2019.

[39] M. Álvarez-Fernández, M. Sanz-Flores, B. Sanz-Castillo, M. Salazar-Roa, D. Partida, E. Zapatero-Solana, H. R. Ali, E. Manchado, S. Lowe, T. VanArsdale, *et al.*, "Therapeutic relevance of the PP2A-B55 inhibitory kinase MASTL/Greatwall in breast cancer," *Cell Death & Differentiation*, pp. 1–13, 2017.

[40] I. B. Weinstein, "Addiction to oncogenes–the achilles heal of cancer," *Science*, vol. 297, no. 5578, pp. 63–64, 2002.

[41] A. Tsherniak, F. Vazquez, P. G. Montgomery, B. A. Weir, G. Kryukov, G. S. Cowley, S. Gill, W. F. Harrington, S. Pantel, J. M. Krill-Burger, *et al.*, "Defining a cancer dependency map," *Cell*, vol. 170, no. 3, pp. 564–576, 2017.

[42] P. Jaako, A. Faille, S. Tan, C. C. Wong, N. Escudero-Urquijo, P. Castro-Hartmann, P. Wright, C. Hilcenko, D. J. Adams, and A. J. Warren, "eIF6 rebinding dynamically couples ribosome maturation and translation," *Nature Communications*, vol. 13, no. 1, p. 1562, 2022.

[43] I. Buhaescu and H. Izzedine, "Mevalonate pathway: A review of clinical and therapeutical implications," *Clinical biochemistry*, vol. 40, no. 9-10, pp. 575–584, 2007.

[44] G. Gruenbacher and M. Thurnher, "Mevalonate metabolism in cancer stemness and trained immunity," *Frontiers in Oncology*, vol. 8, p. 394, 2018.

[45] S. H. Lee, J.-H. Lee, and S.-S. Im, "The cellular function of SCAP in metabolic signaling," *Experimental & molecular medicine*, vol. 52, no. 5, pp. 724–729, 2020.

[46] P. Gao, J.-L. Hao, Q.-W. Xie, G.-Q. Han, B.-B. Xu, H. Hu, N.-E. Sa, X.-W. Du, H.-L. Tang, J. Yan, *et al.*, "PELO facilitates PLK1-induced the ubiquitination and degradation of Smad4 and promotes the progression of prostate cancer," *Oncogene*, vol. 41, no. 21, pp. 2945–2957, 2022.

[47] J. C. Chen, M. J. Alvarez, F. Talos, H. Dhruv, G. E. Rieckhof, A. Iyer, K. L. Diefes, K. Aldape, M. Berens, M. M. Shen, *et al.*, "Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks," *Cell*, vol. 159, no. 2, pp. 402–414, 2014.

[48] T. Huang, Y. Hou, X. Wang, L. Wang, C. Yi, C. Wang, X. Sun, P. K. Tam, S. M. Ngai, M. H. Sham, *et al.*, "Direct interaction of Sox10 with cadherin-19 mediates early sacral neural crest cell migration: Implications for enteric nervous system development defects," *Gastroenterology*, vol. 162, no. 1, pp. 179–192, 2022.

[49] C. Capparelli, T. J. Purwin, M. Glasheen, S. Caksa, M. Tiago, N. Wilski, D. Pomante, S. Rosenbaum, M. Q. Nguyen, W. Cai, *et al.*, "Targeting SOX10-deficient cells to reduce the dormant-invasive phenotype state in melanoma," *Nature communications*, vol. 13, no. 1, p. 1381, 2022.

[50] G. E. Moody, J. Moriguchi, S. Li, F. Lee, B. Frank, A. Gilbert, R. Case, K. Dang, B. Hinkle, S. Coberly, *et al.*, "A novel bispecific CD3/CDH19 antibody construct (CDH19 BiTE) directs potent killing of melanoma cells in vitro and in vivo and is enhanced by blockade of PD-L1," *Cancer Research*, vol. 76, no. 14_Supplement, pp. 2968–2968, 2016.

[51] A. Fujita, K.-i. Nakamura, T. Kato, N. Watanabe, T. Ishizaki, K. Kimura, A. Mizoguchi, and S. Narumiya, "Ropporin, a sperm-specific binding protein of rhophilin, that is localized in the fibrous sheath of sperm flagella," *Journal of cell science*, vol. 113, no. 1, pp. 103–112, 2000.

[52] J. M. Saunus, X. M. De Luca, K. Northwood, A. Raghavendra, A. Hasson, A. E. McCart Reed, M. Lim, S. Lal, A. C. Vargas, J. R. Kutasovic, *et al.*, "Epigenome erosion and SOX10 drive neural crest phenotypic mimicry in triple-negative breast cancer," *NPJ Breast Cancer*, vol. 8, no. 1, p. 57, 2022.

[53] J. Da Gama Duarte, K. Woods, L. T. Quigley, C. Deceneux, C. Tutuka, T. Witkowski, S. Ostrouska, C. Hudson, S. C.-H. Tsao, A. Pasam, *et al.*, "Ropporin-1 and 1b are widely expressed in human melanoma and evoke strong humoral immune responses," *Cancers*, vol. 13, no. 8, p. 1805, 2021.

[54] S. A. Graf, C. Busch, A.-K. Bosserhoff, R. Besch, and C. Berking, "SOX10 promotes melanoma cell invasion by regulating melanoma inhibitory activity," *Journal of Investigative Dermatology*, vol. 134, no. 8, pp. 2212–2220, 2014.

[55] J. Wolf, C. Auw-Haedrich, A. Schlecht, S. Boneva, H. Mittelviefhaus, T. Lapp, H. Agostini, T. Reinhard, G. Schlunck, and C. A. Lange, "Transcriptional characterization of conjunctival melanoma identifies the cellular tumor microenvironment and prognostic gene signatures," *Scientific reports*, vol. 10, no. 1, p. 17 022, 2020.

[56] Z. Chen, J. Huang, Y. Liu, L. K. Dattilo, S.-H. Huh, D. Ornitz, and D. C. Beebe, "FGF signaling activates a Sox9-Sox10 pathway for the formation and branching morphogenesis of mouse ocular glands," *Development*, vol. 141, no. 13, pp. 2691–2701, 2014.

[57] S. Nadanaka and H. Kitagawa, "Exostosin-like 2 regulates FGF2 signaling by controlling the endocytosis of FGF2," *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1862, no. 4, pp. 791–799, 2018.

[58] L. Otero, E. Lacunza, V. Vasquez, V. Arbelaez, F. Cardier, and F. González, "Variations in AXIN2 predict risk and prognosis of colorectal cancer," *BDJ open*, vol. 5, no. 1, p. 13, 2019.

[59] T. A. Chan, Z. Wang, L. H. Dang, B. Vogelstein, and K. W. Kinzler, "Targeted inactivation of CTNNB1 reveals unexpected effects of β-catenin mutation," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 8265–8270, 2002.

[60] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, *et al.*, "The genotype-tissue expression (GTEx) project," *Nature genetics*, vol. 45, no. 6, pp. 580–585, 2013.