Machine Learning of Synthetic Lethality

Data Integration, Generalisation, and Selection Bias

Colm F. Seale

Technische Universiteit Delft

TU Delft

Machine Learning of Synthetic Lethality

Data Integration, Generalisation, and Selection Bias

by

Colm F. Seale

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Monday August 31st, 2020 at 3:00 PM.

Student number:	4772474	
Masters Programme:	Computer Science, Data Science and Technology Track	
Faculty:	Electrical Engineering, Mathematics and Computer Science	
Project duration:	November 10, 2019 – August 31, 2020	
Thesis committee:	Prof. dr. ir. Marcel Reinders,	TU Delft
	Dr. Joana Gonçalves,	TU Delft, supervisor
	Dr. Claudia Hauff	TU Delft

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

TU Delft

Preface

This MSc thesis is the culmination of a ten-month research project on machine learning of synthetic lethality. Almost two years ago, I quit my work as a software engineer to pursue a career in data science, with an eye on specializing in bioinformatics. Last year, I began my journey into the world of bioinformatics in earnest by conducting a literature survey on *in silico* approaches to identifying and prioritising drug targets. During this time, I came across the topic of synthetic lethality. I was immediately enthralled by this subject, and it's implications for the field of personalised medicine. I was filled with excitement and anticipation for my future work, and I knew I had made the right choice. This thesis marks the first milestone in this, hopefully, long and fruitful journey. I have written this thesis for my fellow computer scientists, especially those who also possess the same curiosity that I did for how computer science may be leveraged in the realms of biology and medicine.

This work would not have been possible without the contributions of many people. First, I would like to thank my supervisor Joana Gonçalves for her enthusiasm and patience. This daunting task of switching careers into research has not been an easy one, and your feedback and encouragement have helped me to improve and progress along this path. I would also like to thank my friends Bianca, Francesca, Gabriele, Ines, Matteo, Mila, Miranda, and Vinny for their support, their knowledge, their kind words, and inspiration throughout my studies at TU Delft. Given the unique global backdrop under which this work was performed, I would like to give a special thanks to Bianca, Francesca, Matteo, and Mila for their daily company, either digitally or in-person, through what was a trying time for all of us. Finally, I would like to thank my family for their support, love, and encouragement throughout the entirety of my studies.

Colm F. Seale
Mountrath, Ireland, August 2020

Glossary

Term	Definition
Cell proliferation	The process by which cells increase in number, as is defined by the balance between cell division and cell loss (by either cell death or cell differentiation).
CRISPR Screening	CRISPR screening is a large-scale, genome-wide, *in vitro* experimental technique which aims to determine gene function by silencing specific genes and measuring a particular phenotype. Gene silencing is achieved using the CRISPR-cas9 gene-editing system, which cleaves genes from the genome to prevent expression.
Endogeneous	Substances or processes that originate from within a biological system such as an organism, tissue, or cell. In this paper, this generally refers to something which originates from within the cell.
Gene expression	The process by which genes, encoded in DNA, are converted into a gene product, such as proteins, which carry out functions within the cell. A gene's expression can be measured by the abundance of RNA for that gene within the cell.
Gene silencing	The regulation of gene expression to prevent the expression of a certain gene.
Genotype	An organisms complete set of genes.
Inactive gene	A gene which is currently silenced.
In vitro	Performed or taking place in a test tube, culture dish, or elsewhere outside a living organism
In silico	Conducted or produced using computer modelling or computer simulation.
Knockdown	Partial suppression of a gene's expression through RNAi screening technology.
Knockout	Complete suppression of a gene's expression through CRISPR screening technology.
Missense mutation	A single nucleotide change in the DNA sequence of a gene which changes the amino acid sequence. May result in a change in protein function.
Phenotype	An observable physical property of an organism.
RNA interference (RNAi)	Similar to CRISPR screening (see above) except utilises RNAi technology to partially suppress gene expression. It does so by interrupting the genetic information flow in the translation phase from mRNA to protein.
Silent mutation	A change in the DNA sequence of a gene without a subsequent change in the amino acid or the protein function

Master's Thesis

Machine Learning of Synthetic Lethality: Data Integration, Generalisation, and Selection Bias

Colm Seale

Department of Pattern Recognition and Bioinformatics, Faculty EEMCS, Delft University of Technology, The Netherlands

Abstract

Motivation: Synthetic lethality (SL) arises between two genes when loss of function of both genes would lead cells to become inviable. This can be exploited for therapy, where a drug is used to selectively kill diseased cells by perturbing one gene of an SL pair where the other gene is inactive (e.g. through naturally occurring mutation). Computational prediction of SL relationships is very appealing as it can help reduce cost- and labour-intensive experimental testing to the most promising candidate pairs. Even though machine learning models have shown promising results for SL prediction compared to traditional statistical approaches, crucial questions remain. First, which sources of molecular data are most useful for SL prediction? Many approaches rely on either cell line or patient tumour data separately, and ignore data from healthy tissue. We argue these should be combined to leverage relevant data sources that are exclusively available for cancer cell models and patient tumours, and to enable the transfer of knowledge between models and actual patient tumours. Likewise, changes in the relationship of gene pairs between healthy and tumour tissue may be informative for SL prediction. We assess several machine learning techniques to best leverage molecular profiles for cancer-specific or pan-cancer SL prediction. Second, what are the effects of selection bias on SL prediction methods and which techniques are most robust? This has been insufficiently addressed, as models in the literature are often tested using data from one or two cancer types or datasets. We investigate robustness to cancer representation and gene selection biases, which are inherent to most SL datasets. We hypothesise that approaches based on matrix factorisation will be especially sensitive to the latter, as they are dependent on an *a priori* SL network structure, which also determines the scope of the prediction space.

Results: In this work, we used a dataset of over 10,000 experimentally validated SL and non-SL gene pairs spanning four cancers (breast, lung, ovarian, colon) to train logistic regression and random forest models. Our models achieved the best precision amongst the highest-ranking predictions compared to our selected baseline methods (DAISY, DiscoverSL, PCA-gCMF). Model feature importance scores showed that gene dependency and mutual exclusivity data contained useful information for SL prediction. Our results also show that our random forest models were more resilient to cancer representation bias in the context of pan-cancer SL classification than were our logistic regression models. We demonstrated that the impact of this bias on predictive performance can be mitigated through balancing cancer representation. Finally, our logistic regression models exhibited a superior ability to generalise to unseen cancer types and genes. Our selected matrix factorisation baseline, PCA-gCMF, was the most sensitive to gene-selection bias. Given the prevalence of gene-selection bias in the literature, we speculate that the real-world effectiveness of many SL prediction techniques may be over-estimated.

1 Introduction

Synthetic lethality (SL) describes a relationship between a pair of genes where the inactivation of a single gene within the pair allows the cell to retain viability, but the simultaneous inactivation of both genes results in the cell death (Boone *et al.*, 2007) (Fig 1). SL is an interesting concept in the field of personalised medicine (Fig 2). SL provides a mechanism for personalised therapeutics whereby drugs can be developed to target the SL neighbour of a naturally inactive gene (e.g. due to mutation). For instance, the PARP1 gene was found to be SL with both BRCA1 and BRCA2 genes in breast cancer. PARP1 inhibitors were later approved for use to selectively kill breast cancer cells with mutations in either of the BRCA1 or BRCA2 genes (Fong *et al.*, 2009) (see Huang *et al.* (2019) for a review of SL in cancer therapy). The first step to developing new SL therapies like these is to identify SL interactors which might later become viable drug targets.

Modern identification of SL typically occurs experimentally through *in vitro* loss-of-function (LoF) screening (Shen *et al.*, 2017; McDonald *et al.*, 2017). These screens identify SL pairs by quantifying the change in cell proliferation after the silencing of both queried genes within cells. Alternatively, one can measure this change by silencing one gene and comparing contexts where the second gene is or is not endogenously inactive. This silencing is achieved using CRISPR or RNAi technology (for a review of current LoF technologies, refer to Schuster *et al.* (2019); for applying these technologies to screen for SL pairs, refer Nijman (2011)). These screens are both laborious and expensive. Additionally, the number of combinations of candidate synthetic lethal pairs for validation is tremendous. To visualise the scale of this problem, imagine a search space encompassing every gene-gene interaction for every cell type of interest.

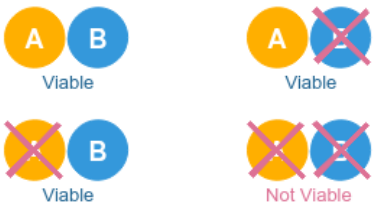


Fig. 1: Synthetic lethality. Circles A and B represent a synthetic lethal pair of genes within a cell. Inactivation of either gene A or B leaves the cell viable, whereas inactivation of both results in cell death.

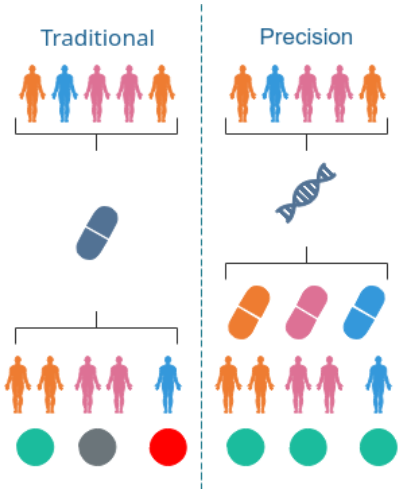


Fig. 2: Personalised medicine. (Left) Traditional approaches prescribe treatment based on population averages, potentially leading to mixed results, having either a positive (green circle), negative (red circle) or no effect (grey). (Right) Personalised medicine takes individual characteristics into account, such as genetic background, to stratify patients into certain treatment categories to improve the likelihood that treatments will have a positive effect.

Methods have been proposed to alleviate this issue by predicting SL interactions *in silico* between queried genes. In turn, such approaches can help inform the construction of new experimental studies by reducing the list of candidate gene pairs to only the most promising (O'Neil *et al.*, 2017). Numerous approaches for SL prediction have been discussed in the literature. These methodologies fall into three categories: statistics-, machine learning-, and matrix factorisation-based methods.

Statistics-based methods utilise hypotheses about SL relationships to make inferences about new SL interactors. For example, DAISY infers SL interactions between genes based on how either the cell line dependency, expression, or mutation of one gene changes with respect to mutation or expression in another (Jerby-Arnon *et al.*, 2014). BiSEp takes advantage of genes which exhibit bimodally distributed gene expression profiles. It outputs ranked lists of candidate SL gene pairs which exhibit mutually exclusive low expressivity, or exhibit high or low expressivity in one gene that is mutually exclusive with mutations in the other (Whalen *et al.*, 2016). However, reliance on expert knowledge of genomic data means these methods can miss underlying relationships between known SL interactors. These relationships could be advantageous for SL prediction.

Machine learning (ML) methods attempt to predict SL by applying general-purpose algorithms to discover these latent relationships in genomic data. DiscoverSL is a random forest pan-cancer SL classification model which has been trained on breast and lung cancer data. It uses features derived from patient tumour and biological pathway data (Das *et al.*, 2019). EXP2SL use neural networks to learn low-dimensional representations of gene expression profiles. These are then used as features in a logistic regression model to predict cell line-specific SL interactions (Wan *et al.*, 2020).

Matrix factorisation is a branch of ML methods which aim to factorise a large matrix into a product of matrices. In this paper, we treat them as entirely separate to other ML models for simplicity. These methods account for network structure between individual entities (e.g. SL interactions between genes). SL2MF uses logistic matrix factorisation to predict gene-gene SL interactions while incorporating neighbourhood regularization using gene ontology semantic similarities and protein-protein interaction topological similarities (Liu *et al.*, 2019). PCA-gCMF utilises group sparse-collective matrix factorisation to relate information between gene pairs across an arbitrary number of input matrices (Liany *et al.*, 2019).

Common to statistical-, ML-, and matrix factorisation-based approaches are several recurring weaknesses. Many utilise only a few genomic data types sourced from a limited number of biological contexts. Thus, they are not taking advantage of the full plethora of publicly available data genomic data. Furthermore, these studies tend to overlook model generalisation. The absence of such analysis in the literature means that their real-world applicability may be over-estimated. As part of our work, we aim to address both these problems.

First, we introduce several features into an ML context for SL prediction, and train both linear and non-linear ML models. These features broaden the types and sources of data under consideration by current ML approaches. We integrate new data points: gene dependency scores; mutual exclusivity scores; and clinical information such as a patient's age, race, sex, and survival. Gene dependency scores indicate the sensitivity of an immortalised cell line (cancer models whose molecular characteristics are well understood) to the inactivation of a single gene (McDonald *et al.*, 2017; Behan *et al.*, 2019). Incorporating mutation data with these gene dependency scores may indicate SL in cell-lines. Mutual exclusivity scores describe the likelihood that mutations in two or more given genes do not co-occur (Babur *et al.*, 2015; Canisius *et al.*, 2016). Non-co-occurrence in cells may indicate that these genes are SL. Coupling mutation, clinical, and survival data may reveal SL pairs which provide a therapeutic advantage (Lee *et al.*, 2018).

Moreover, we hypothesise that data extracted from other biological contexts, such as healthy tissue (donated from a cancer-free doner) and paired-normal tissue (healthy tissue extracted from a region adjacent to a tumour site) may contain useful information for SL prediction. Such data has not been previously considered in the literature. We incorporate gene expression data from these tissues, which may reveal functional relationships. The change in those relationships between healthy and diseased tissue from the same region may highlight SL interactions.

Second, we identify two forms of bias in SL datasets which affect the ability of a model to generalise: "cancer representation" and "gene selection" biases. Cancer representation bias occurs due to the popularity of some cancers in the literature and the rarity in the occurrence of others. This leads to an imbalance of cancer representation in current available datasets. Similarly, gene selection bias occurs because certain genes are better studied than others. Typical SL datasets utilised in many studies feature only a fraction of all human genes. Certain genes are over-represented due to historical or academic reasons and others are under-studied despite their possible importance to certain cancers (Stoeger *et al.*, 2018).

We construct several experiments to study the effect these biases have on model generalisability. We explore the capability of models to predict SL pan-cancer despite been trained on a very limited number of cancers, as assumed by DiscoverSL (Das *et al.*, 2019). We also assess the impacts of gene selection bias on the training and evaluation of different machine learning techniques. We argue that matrix factorisation methods are the most affected by this bias, considering their dependence on an *a priori* labelled gene-gene SL network to make further link predictions. Other SL prediction models do not mandate this network.

To summarise, the main aims of this research are twofold: (i) to introduce novel data sources and features into a machine learning context and propose machine learning models for *in silico* SL prediction; (ii) to investigate the generalisability of SL machine learning models, and the impact of cancer representation and gene selection bias. We evaluate these models in the context of cancer-specific and pan-cancer SL classification using linear and non-linear machine learning models and selected baselines from the literature.

2 Methods

In this section, we describe the methodology and data used to build and evaluate models for predicting SL. First, we formalize the SL prediction problem. Then, we describe the data collection and preprocessing, and the generation of features used in the prediction models. Next, we describe the different models, our selected baselines, and explain how they are trained and evaluated. Finally, we describe variations in the construction of the train and test sets. These variations are employed in each of our different experiments to assess generalisation across cancers and to unseen genes.

2.1 Problem Formulation

Throughout this paper, we refer to a "gene pair" as a pair of genes for which the presence of a synthetic lethal interaction in a specific biological context is either known or being predicted. The biological context is determined by a set of biological samples of a given tissue or cancer type, e.g. breast cancer, which is characterized by gene-wise measurements and pairwise gene relationships at the molecular level. We define the synthetic lethality prediction problem as follows. Let $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$, where $\boldsymbol{x}_i \in \mathbb{R}^m$ and $i \in [1..n]$. Here, $\boldsymbol{x}_i$ is an m-dimensional, real-valued vector representing a single example for a specific gene pair and biological samples context. Each of the m features denotes an individual gene measurement or a relationship between measurements obtained for the gene pair in the given biological sample. As a result, the set $\mathcal{X}$ is a set of n examples between n unique gene pairs (with context), where index i denotes the i-th example in the set $\mathcal{X}$. We let $y_i \in \{0, 1\}$ be the binary class label indicating the existence of a synthetic lethal relationship between the gene pair (in context) associated with the i-th example in $\mathcal{X}$. We also define $f(\boldsymbol{x}_i) = \bar{y}_i$ as a function that maps an example or feature vector $\boldsymbol{x}_i$ to $\bar{y}_i$, where $\bar{y}_i \in \mathbb{R}$. Additionally, let $\hat{y}_i \in \{0, 1\}$ be the predicted value for y_i, and let t be some threshold such that when $\bar{y}_i < t$, $\hat{y}_i = 0$ and when $\bar{y}_i \geq t$, $\hat{y}_i = 1$. For a given set of gene pairs $\mathcal{X}$, we ideally wish to learn a mapping function $f(\boldsymbol{x}_i)$ and select a threshold t such that for any gene pair in $\mathcal{X}$, $P(y_i = \hat{y}_i)$ is close as possible to 1. We treat this as a binary classification problem, and employ machine learning techniques to determine this mapping function $f(\boldsymbol{x}_i)$ from some dataset of examples D, where D is a $n \times m$ matrix in the form:

$$D = \left[\begin{array}{c|ccc} y_{11} & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & x_{n1} & \cdots & x_{nm} \end{array}\right].$$

2.2 Data Collection and Preprocessing

To be able to assess generalisation and bias of synthetic lethality prediction models, we sought to analyse data for multiple cancer types and datasets. We collected SL relationships from two studies and four cancer types, which we use as y_i class labels. We also derived gene pair features per cancer type $\boldsymbol{x}_i$ based on molecular profiles of both cell lines and patients, as well as of healthy tissues related to the four cancer types under study.

Synthetic lethality gold standard. In order to train and test ML models, we require a labelled dataset of *in vitro* experimentally validated gene-gene SL interactions. These labels will act as the ground truths during learning and evaluation. In this paper, we will refer to such labels as our "gold standard" class labels.

We obtained cancer-specific SL relationships between pairs of genes by combining data from two previous studies, ISLE and DiscoverSL (Lee *et al.*, 2018; Das *et al.*, 2019). Together, the ISLE and DiscoverSL datasets included both positive and negative pairwise SL gene relationships derived experimentally by 26 other studies. They were validated through either double gene knockdown/knockout or the targeting of one gene in contexts where the other gene is endogenously inactive (e.g. by naturally occurring mutations) using CRISPR or RNA interference. We conducted quality analysis and curation on these datasets, and subsequently, we combined these gold standards to maximise the number of examples available for training and testing.

First, we assessed the level of disagreement between duplicated gene pairs in each gold standard, separately, by calculating the Jaccard distance per cancer type. The Jaccard distance is defined as follows:

$$d_J = 1 - \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

where A and B are the sets of duplicate gene pairs which have positive and negative labels, respectively. When $d_J = 0$, no common gene pairs with opposing labels exist. When $d_J = 1$, every duplicate gene pair has an entry with at least 1 opposing label. The Jaccard distance for DiscoverSL was 0 for BRCA and LUAD, while the Jaccard distances were generally higher within the ISLE (Table 1). This indicated that ISLE contained more conflicting information than DiscoverSL. To clean our data, we reduced the number of duplicate gene pair entries within each dataset to zero. We retained an entry of a duplicate pair if all duplicate entries for a gene pair had identical class labels. Otherwise, we removed all duplicate entries if there was any disagreement on the label.

After we had curated ISLE or DiscoverSL to include no duplicate gene pair entries within the separate sets, we combined the two gold standards into one. To combine these datasets, we reduced 63 duplicate gene pair entries that existed between the gold standards to a single entry. When

	Number of Duplicates	Jaccard Distance (d_J)
BRCA	300	.663
COAD	7	0
LUAD	568	.112
OV	10	.8

(a) ISLE

	Number of Duplicates	Jaccard Distance (d_J)
BRCA	3	0
LUAD	0	0

(b) DiscoverSL

Table 1. Breakdown of the number of duplicates and the level of disagreement within each gold standard source. The level of disagreement is indicated by the Jaccard distance.

	ISLE		DiscoverSL		Combined	
	+	-	+	-	+	-
BRCA	713	1168	835	72	1548	1240
COAD	859	806	0	0	859	806
LUAD	202	5155	347	312	549	5467
OV	223	449	0	0	223	449
All	1997	7578	1182	384	3179	7962

Table 2. Breakdown of numbers of examples with positive and negative labels per cancer type for each of the gold standard synthetic lethality datasets, after quality analysis and curation.

the gold standards agreed on a gene pair's label, one entry was retained. When the gold standards disagreed on the label between them, the label from DiscoverSL was chosen as the ground truth. We choose the label from the DiscoverSL gold standard due to the lower level of disagreement within DiscoverSL as compared to ISLE (Table 1). Both studies included SL labels for different cancer types. We included in our dataset only those cancer types for which we had at least 200 positive and negative samples after quality analysis and curation. Specifically, we included breast, colon, lung, and ovarian cancer, which we denote respectively as BRCA, COAD, LUAD, and OV. Note that the ISLE, DiscoverSL, and combined gold standards have differing cancer type representation and class imbalances (Table 2). In all of our experiments, we use the combined dataset except where specified otherwise.

Cancer cell line data. Cell lines are often used as a simple model to aid in the study of more complex biological systems, including cancer. A cell line is an homogeneous, immortalised population of cells from a multicellular organism that has acquired the ability to propagate indefinitely through genetic mutation. Their immortal nature simplifies analysis of cell biology, as they are easier to control, manipulate, and publish, which allows for repeatable scientific experiments. Recently, cell lines have been used to measure gene dependencies. A cell is considered dependant on a given gene if the cell loses viability when that gene is inactivated. Gene dependencies for cancer cell lines are made available by the Cancer Dependency Map in the form of real-valued scores (Dempster *et al.*, 2019; Meyers *et al.*, 2017; McFarland *et al.*, 2018; Behan *et al.*, 2019). We downloaded dependency scores derived by CRISPR (CERES) and RNA interference (DEMETER2) screens from the 19Q3 DepMap public release. We also downloaded corresponding mutation data for these cancer cell lines (Ghandi *et al.*, 2019). This mutation data included which genes within a cell line were mutated and what was the nature of this mutation, such as a silent mutation, missense mutation, etc. We use cell line gene dependency and mutation data as we can derive relationships within a gene pair under specific cellular conditions, such as when one gene in a pair is mutated, as detailed in Section 2.3.

Cancer patient tumour and survival data. One of the main disadvantages of using cell lines to analyse cell biology is that, due to their immortalising mutations, these cells are less biologically relevant than cells isolated directly from multicellular organisms (also called primary cells). This means that the functions and relationships derived from cell lines may not be consistent with those found in their non-immortal cells of origin. For this reason, scientists also study primary cells. Primary cells, in contrast to cell lines, retain the true characteristics of the original tissue at the expense of possessing a finite lifespan. Cancer patient tumour samples are an example of primary cells. Studying patient tumours also presents an opportunity to study the donors themselves, allowing for the inclusion of factors such as age, race, sex, survival rates, etc. These data may allow a machine-learning algorithm to analyse characteristics of SL interactions which are biologically relevant to patient tumours. For instance, if mutations in an SL pair provided a therapeutic effect, we would expect patients harbouring such mutations to have a survival advantage. The Cancer Genome Atlas Research (TCGA) is a cancer genomics program, and vast source of primary sample data, which has molecularly characterised over 20,000 patient tumour samples and matched normal samples across 33 cancer types. We downloaded the following data from TCGA, which was generated on the 28-Jan-2016:

- Mutation data detailing which genes are mutated in which samples and the nature of these mutations, similar to our cancer cell line data above.
- Copy-number variation (CNV) data in the form of real- and integer-valued scores obtained via GISTIC (Mermel *et al.*, 2011). CNV describes a phenomenon in which the number of copies of a particular gene varies between individuals. Increases in the number of copies are termed amplifications, while reductions are referred to as deletions. The integer-valued scores are derived by applying low- and high-level thresholds to the real-valued data. Entries with +1 or -1 only exceed the respective low-level thresholds for amplifications and deletions, while entries with +2 or -2 exceed the high-level thresholds.
- Clinical data such as race, age, sex, and right-censored survival data indicating the number of days until the death of a patient.

We also downloaded the patient tumour sample gene expression data in the form of transcript read counts aggregated per gene, sourced from the TCGA RNA-seq data and preprocessed using the RSubread package. This preprocessing produced fewer zero-expression genes and less variance across replicate biological samples when compared to the TCGA preprocessing pipeline. These data can be found under accession number GSM1536837 in the Gene Expression Omnibus (GEO) (Rahman *et al.*, 2015).

Healthy tissue data. In the previous section, we described how *in vitro* patient tumour sample data is the most biologically relevant as it maintains many important biomarkers and characteristics seen *in vivo*. Nonetheless, healthy tissue data from the same region may still possess useful information, and the preservation or abandonment of characteristics and functions between patient tumour samples and data from the equivalent healthy tissue could provide SL indicators. We examine two different sources for healthy tissue data: tissue samples taken from healthy test subjects and matched-normal tissue taken from cancer patients. The latter are samples extracted from cancer patients from a region of healthy tissue adjacent to the cancerous tumour. Matched normal samples are often used as controls for understanding disease mechanisms. We downloaded the gene expression data for breast, lung, colon, and ovarian tissue extracted from healthy test subjects, provided as gene-aggregated transcript per million (TPM) values, from the GTEx Portal (dbGaP Accession phs000424.v8.p2). We downloaded the gene expression data for matched normal BRCA samples from GEO (accession number GSM1536837) as described previously for cancer patient data. No matched normal data existed for COAD or OV.

Biological pathway data. Genes which tend to co-participate in the same biological pathways may be more co-dependent on one another than genes that do not and therefore might also have a higher chance of being SL. To derive measures of co-participation, we downloaded canonical pathways gene sets derived from the KEGG, PID and Reactome pathway databases from the molecular signatures database v7.1 (Subramanian *et al.*, 2005; Liberzon *et al.*, 2011). These sets detail lists of genes which are known to be involved in different biological pathways within the cell.

2.3 Feature Generation

Every example $\boldsymbol{x}_i$ is characterized by 26 features (Table 3). Each feature, or element of $\boldsymbol{x}_i$, is calculated based on one or more data types. They

Symbol	Description	Biological sample	Data type
CRISPR_dep_stat	Change in CRISPR dependency score of one gene based on non-silent mutations in the other (Wilcoxon)	Cancer cell lines	Gene dependency, mutation
CRISPR_dep_pvalue	Significance of change in CRISPR dependency score in one gene based on non-silent mutations in the other (Wilcoxon)	Cancer cell lines	Gene dependency, mutation
CRISPR_cor_stat	Correlation of gene-wise CRISPR dependency scores (Pearson's)	Cancer cell lines	Gene dependency
CRISPR_cor_pvalue	Significance of correlation of gene-wise CRISPR dependencies (t-test)	Cancer cell lines	Gene dependency
CRISPR_avg	Average of gene-wise means of CRISPR dependency scores	Cancer cell lines	Gene dependency
RNAi_dep_stat	See CRISPR equivalent	-	-
RNAi_dep_pvalue	See CRISPR equivalent	-	-
RNAi_cor_stat	See CRISPR equivalent	-	-
RNAi_cor_pvalue	See CRISPR equivalent	-	-
RNAi_avg	See CRISPR equivalent	-	-
DISCOVER	Mutual exclusivity score	Patient tumour	CNV, mutation
discoversl_mutex_amp	Significance of non-co-occurrence of amplifications (hypergeom.)	Patient tumour	CNV, mutation
discoversl_mutex_del	Significance of non-co-occurrence of deletions (hypergeom.)	Patient tumour	CNV, mutation
discoversl_mutex_mut	Significance of non-co-occurrence of non-silent mutations (hypergeom.)	Patient tumour	CNV, mutation
discoversl_mutex	Combined p-value of previous three scores using Fisher's method	Patient tumour	CNV, mutation
mutex_alt	Significance of non-co-occurrence of amplifications, deletions or non-silent mutations	Patient tumour	CNV, mutation
logrank_pval	Significance of change in survival time between patients with and without aberrant expression or copy number in both genes	Patient tumour, patient clinical data	CNV, expression, mutation, patient age, race, sex, days until death
diff_exp_logFC	Differential expression of one gene when the other is mutated (log fold-change)	Patient tumour	CNV, mutation, expression
diff_exp_pvalue	Significance of differential expression of one gene when the other is mutated (edgeR test p-value)	Patient tumour	CNV, mutation, expression
gtex_corr	Co-expression measure (Pearson's correlation)	Healthy tissue	CNV, mutation, expression
gtex_corr._pvalue	Significance of co-expression (t-test)	Healthy tissue	CNV, mutation, expression
tumour_corr	See GTEx equivalent	Patient tumour	-
tumour_corr._pvalue	See GTEx equivalent	Patient tumour	-
normal_corr	See GTEx equivalent	Patient matched-normal tissue	-
normal_corr._pvalue	See GTEx equivalent	Patient matched-normal tissue	-
pathway_coparticipation	Significance of co-occurrence in pathways (hypergeom.)	Pathway databases	Pathway gene sets

Table 3. Symbols and descriptions of all the generated features used within our machine learning models. The biological sample column indicates where the measurements were taken from. The data type column indicates what measures were taken.

summarise individual gene measurements or a relationship between the pair of candidate genes across multiple samples of a given cancer type or corresponding tissue type. We rely on cancer type for patient tumour and cancer cell line data, or the corresponding tissue type for healthy and matched-normal tissue. The only exception to this is the pathway co-participation feature, which is calculated based on biological pathway data.

Gene dependencies. For each candidate pair of genes A and B, we calculated features based on cancer cell line gene dependencies. We calculated five features for the CRISPR-based dependency scores, and the same five features for the RNAi-based dependency scores, leading to 10 features in total. We determined the first two features by performing two two-tailed Wilcoxon rank-sum tests (Mann and Whitney, 1947), one for the pair (A, B) and another for the same pair in reverse order (B, A). Each Wilcoxon rank-sum test quantifies the change in cell line dependency on the first gene between two groups of cell lines: those which have a non-silent mutation in the second gene, and those which do not. We take the statistic and p-value from the tested pair (A,B) or (B,A) yielding the smallest p-value as our features. Smaller p-values indicate a higher likelihood that a mutation in one gene leads to a change in cell line dependency on the other gene. We expect this to occur between truly SL gene pairs.

The second two features were the Pearson's correlation and corresponding two-tailed t-test p-value of the dependency scores of genes A and B across the cell lines for a given cancer type. Correlations closer to +1 or -1, together with p-values closer to 0, may indicate some sort of interaction between genes. The fifth feature was the average of mean dependency score, defined as the average of the means of the dependency scores over the cell lines on genes A and B. Both genes of an SL gene pair should have low individual gene dependency scores across the cancer cell lines, and thus a low average gene dependency score.

Mutual exclusivity. A possible explanation for mutual exclusivity between a gene pair is that there might exist an SL interaction between the two. If a gene pair is SL, then we would expect to find less examples of co-mutations of these genes in cells. Mutual exclusivity p-values express the probability that the co-occurrence of mutations in a pair of genes A and B is lower than expected by chance. We calculated mutual exclusivity p-values based on patient tumour mutation data using three different methods:

DiscoverSL, DISCOVER, and MUTEX (Das *et al.*, 2019; Canisius *et al.*, 2016; Babur *et al.*, 2015). DiscoverSL tests three null hypotheses that non-silent mutations, amplifications and deletions are not mutually exclusive using Eq. 2 and Eq. 3. Eq. 2 is a hypergeometic test which represents the p-value for the co-occurrence of mutations. Eq. 3 then calculates the mutual exclusivity p-value for the i-th hypothesis test.

$$h_i = \sum_{j=n_{A,B}}^{min(n_A,n_B)} \frac{\binom{n_A}{j}\binom{n_T-n_A}{n_B-j}}{\binom{n_T}{n_B}} \tag{2}$$

$$p_i = 1 - h_i \tag{3}$$

Here, n_A and n_B are the numbers of tumour samples with mutations of the given type in genes A and B, respectively (Das *et al.*, 2019). Additionally, $n_{A,B}$ is the number of samples with mutations of the given type in both genes, and n_T is the total number of samples for that cancer type. DiscoverSL then combines the three p-values using Fisher's method, as follows:

$$X^2_{2k} \sim -2\sum_{i=1}^{k} ln(p_i) \tag{4}$$

where p_i is the p-value of the i-th of k hypothesis tests, and X^2_{2k} has a chi-squared distribution with $2k$ degrees of freedom. Using this fact, we can determine the combined p-value (Mosteller and Fisher, 1948).

Some p-values resulting from the above hypothesis tests had a value of 0 when a given alteration never co-occurred between two genes. This caused an issue since Fisher's method is a summation of log values and $ln(0)$ is undefined. Therefore, we only combined p-values when at least one of the individual p-values was larger than 0. If all p-values were 0, we took the combined p-value to be 0 as well. We used the p-values from the individual hypothesis tests and the combined p-value as features.

We defined an additional mutual exclusivity feature by treating every non-silent mutation, amplification and deletion as an "alteration" event, and then calculating a single p-value based on alteration events using the mutual exclusivity test as defined in Eq 2 and Eq. 3.

The MUTEX algorithm also uses a hypergeometric test to determine mutual exclusivity between a gene pair, but excludes tumour samples that have more alterations than Q3+(1.5×IQR), where Q3 is the third quartile in the distribution and IQR is the interquartile range (Babur *et al.*, 2015). The MUTEX software produced near-constant results and thus was not included in the final set of features. DISCOVER works by using a null model that takes into account the overall tumour-specific alteration rates when deciding whether alterations co-occur more or less often than expected by chance (refer to Canisius *et al.* (2016) for details). The software for these methods was downloaded from their respective repositories.

Survival analysis. SL interactions between gene pairs in cancer patient tumours could lend a survival advantage to patients harbouring deleterious mutations in both those genes. This may be due to SL interactions which make cancer cells less viable. We used Cox proportional hazard models to check for a significant difference in survival times between patients possessing tumour samples with and without copy-number amplifications or deletions, non-silent mutations, or aberrant expression in both genes. Amplifications and deletions were defined as having thresholded GISTIC values of 2 and -2, respectively. Aberrant expression was defined as having an expression level within the upper or lower 5th percentile of expression levels for that gene across all patients. We also controlled for age, race, and sex as follows:

$$ln\, h(t) \sim ln\, h_0 + \beta_1 inactiveOrAberrant(A,B) + \beta_2 sex + \beta_3 age + \beta_4 race \tag{5}$$

where $h(t)$ is the hazard function defined as the conditional probability of a patient dying at time t given that the patient has survived to time t (Bewick *et al.*, 2004). The indicator variable $inactiveOrAberrant(A,B)$ expresses the copy-number, mutation, and expression status of both genes in gene pair (A, B) in a patient tumour sample as described above. The β values are the regression coefficients for the explanatory variables, estimated using the R package "Survival" (Therneau, 2020). We calculated whether the β_1 coefficient of the indicator variable $inactiveOrAberrant(A,B)$ is statistically significantly different from 0 using the Wald statistic (Bangdiwala, 1989). Smaller p-values could be an indicator that an SL relationship is giving a survival advantage to patients harbouring mutations in the gene pair (A, B). We used this p-value as our feature.

Gene expression. Co-expression is commonly used as another indicator of interactions between two genes. Here we did similarly, but we determined co-expression between a gene pair for each of three types of biological samples: tumour tissue samples from TCGA, paired-normal tissue samples from TCGA (available for BRCA only), and tissue samples coming from healthy doners from GTEx. Our aim is to determine if extra information can be gained using the similarities or differences in interactions between cancerous, and healthy or matched normal tissues. This was done by calculating pairwise Pearson's correlations and the corresponding two-tailed t-test p-values, similar to the calculations performed for gene dependencies. We use the Pearson's correlation statistic and corresponding p-values per biological sample type as features, yielding six features in total.

In addition to co-expression, we used differential expression analysis on the tumour tissue samples to quantify the variation in the expression of a gene given the presence or absence of non-silent mutations in another gene. Differential expression values were calculated using the edgeR package based on the gene-aggregated expression count data obtained from GEO (Rahman *et al.*, 2015; Robinson *et al.*, 2010). EdgeR first normalises the gene expression counts to account for RNA library size and composition using the Trimmed Mean of M-values normalisation (TMM). Non-transcribed genes are ignored, and a reference sample is selected to determine scaling factors for each of the other samples. As part of deciding the scaling factor for each sample with respect to the reference sample, TMM ignores biased genes and genes which are highly or lowly transcribed in both samples (Robinson and Oshlack, 2010). Differential expression of a given gene between two groups of samples of interest is then determined using an exact test adapted for over-dispersed data (Robinson *et al.*, 2010), where gene-wise dispersion is calculated by quantile-adjusted conditional maximum likelihood (qCML) conditioning on the total counts for that gene. We perform two differential expression tests per gene pair (A, B): one for gene A where the two groups are defined as those containing patient tumour samples with and without a non-silent mutation in gene B; another for gene B between groups of patient tumour samples with and without non-silent mutations in gene A. The minimum p-value from the two differential expression tests and the corresponding log fold-change values are used as features for the gene pair (see Robinson and Smyth (2007) and Robinson and Smyth (2008) for more in-depth details on the edgeR differential expression methodology).

Pathway co-participation. We calculated p-values for the significance of the co-occurrence of genes in a set of gene pathways using the hypergeometric test defined in Eq. 2, where n_A, n_B is the number of occurrences of gene A and B in all gene pathways, respectively. $n_{A,B}$ is the number of occurrences of both genes in the same pathway, and n_T is the total number of pathways. The set of pathways was defined as the union of the KEGG, PID and Reactome gene sets.

2.4 Our synthetic lethality prediction models

Our aim is to produce a function $f(\boldsymbol{x}_i)$, which maps an example denoted by its feature vector to some real value expressing the predicted SL level for a gene pair in a specific biological context. We employed machine learning techniques to determine the function or model $f(\boldsymbol{x}_i)$. We elected to build both linear and non-linear models. Linear models were chosen to investigate whether SL gene pairs would be linearly separable from non-SL gene pairs, and because their simplicity could result in better generalisation. We utilised non-linear models to leverage eventual non-linear interactions between features to predict SL. Moreover, our selected models should be reasonably interpretable. This is important for understanding how our features contribute to the overall performance of the SL prediction models. For these purposes, we chose to use logistic regression for our linear models and random forests as our non-linear models, both of which we elaborate on in the following sections.

Logistic regression models. Logistic regression is a reasonable choice for a linear model when facing a linearly separable binary classification problem and interpretability is a desirable property. Consider a logistic regression equation as defined as follows:

$$\bar{y} = \frac{e^{\beta X}}{1 + e^{\beta X}}, \tag{6}$$

where X denotes an $n \times m$ matrix with n examples (or gene pairs) and m features, β denotes the coefficients of each of the m features in X, and $\bar{y}$ is model output, understood as the probability of an SL interaction within each of the n gene pairs. Note that both matrix X and β include the bias terms for simplicity. These models are interpretable because the model derives coefficients β for each feature, which describe how those features influence the predicted value $\bar{y}$. To control overfitting, we trained logistic regression models with both L0 and L2 regularisation as implemented in the L0Learn R package, or L1 and L2 regularisation as implemented in the glmnet package (Hazimeh and Mazumder, 2018; Friedman *et al.*, 2010). We refer to the former as "L0L2" and to the latter as "Elastic Net". L0L2 and Elastic Net attempt to find the optimal values for β which minimise the objective functions Eq. 7 and Eq. 8, respectively.

$$\arg\min_{\beta} \mathcal{L}(y, \bar{y}) + \lambda_0 ||\beta||_0 + \lambda_2 ||\beta||_2^2 \tag{7}$$

$$\arg\min_{\beta} \mathcal{L}(y, \bar{y}) + \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||_2^2 \tag{8}$$

Here, y denotes the actual value for the response variable. $\mathcal{L}$ is the logistic loss function between y and $\bar{y}$. The $||\beta||_0$ term is the L0 norm, which is the number of non-zero coefficient values in β. Higher values of λ_0 favour sparser models with more zero coefficients. The $||\beta||_2$ term is the euclidean distance of the vector of feature coefficients β from the origin. Higher values of λ_2 results in the shrinkage of coefficients and reduce the effect that multicollinearity has on the interpretability of the trained model (see Section 2.7, "Multicollinearity"). The $||\beta||_1$ term represents the L1 norm and acts as a middle ground between the L0 and L2 norms by both shrinking the coefficients and encouraging sparsity in the model by reducing small coefficients to zero.

Random forest models. Random forests are ensembles of decision tree classifiers capable of capturing non-linear relationships between the predictor and outcome variables. They are trained using bootstrap aggregation of individual decision tree models built on subsets of the training set chosen randomly with replacement and combined with random feature selection for each split in a tree. Random forests make no assumption about the underlying distribution of the data and are reasonably interpretable by their nature as rule-based classifiers. The relative importance of a feature can be measured by the average decrease of the weighted Gini impurity over every node that chooses that feature for node splitting, across all trees. The Gini impurity at node v is defined as follows for binary classification:

$$Gini(v) = \rho_+^v(1 - \rho_+^v) + \rho_-^v(1 - \rho_-^v) \tag{9}$$

where ρ_+^v and ρ_-^v are the proportion of positive and negative class labels at node v, respectively.

We used two implementations of random forests: Multivariate methods with Unbiased Variable selection in R (MUVR) (Shi *et al.*, 2019) and Regularised Random Forests (RRF) (Deng and Runger, 2012). Both methods attempt to control overfitting through different algorithms for feature selection.

MUVR combines the standard implementation as proposed by Breiman (2001) with a feature selection algorithm. This feature selection algorithm performs repeated, nested, cross-fold validation combined with backwards-elimination on the train set to select a reduced set of features. This assists in identifying redundant features and reduces overfitting. We also investigated Boruta (Kursa *et al.*, 2010) for the same purposes. Since the MUVR and Boruta algorithms produced similar results, we decided to only include MUVR in our final results because it has built-in cross-validation. We used the MUVR R package made publicly available by Shi *et al.* (2019).

RRF is a modified version of the classic random forest implementation as proposed by Breiman (2001). RRF has two parameters, $mtry$, which controls how many features are randomly sampled at each new node, and $coefReg$, which controls how much predictive information a feature needs to contribute before being added to an node. To understand the difference between the standard random forest and RRF models, consider the Gini impurity at node v split on feature j, $Gini(j, v)$. We calculate the information gain of a feature j for splitting node v as the difference in impurity at node v and the weighted average of the impurities at each child node v^L and v^R, as follows:

$$\begin{gathered} Gain(j, v) = \\ Gini(v) - (w_L \cdot Gini(j, v^L) + w_R \cdot Gini(j, v^R)) \end{gathered} \tag{10}$$

where v^L, v^R are the left and right child nodes of v, respectively. The w_L and w_R terms are the proportions of instances assigned to the left and right child nodes, respectively. In a classical random forest implementation, at every newly created node, each of the $mtry$ randomly selected features is evaluated and the feature j which maximises $Gain(j, v)$ is selected to split the tree at node v. In an RRF model, $Gain(j, v)$ is replaced by the regularised information gain $Gain_R(j, v)$ defined as follows:

$$Gain_R(j, v) = \begin{cases} Gain(j, v), & j \in F \\ coefReg \cdot Gain(j, v), & j \notin F \end{cases} \tag{11}$$

where F is the set of features used by any node previous to the creation of node v in the forest, and $coefReg \in (0, 1]$. When $j \notin$ F, $coefReg$ penalises the gain for using that feature to split the tree at node v. This means that new features have a larger hurdle to overcome if they are to be added to the feature set used to split the tree at node v. We used the "RRFglobal" model as provided by the caret package in R.

2.5 Training

Train and test sets. For our experiments, we constructed different train and test sets with varying constraints depending on the research question under study. Here we describe the train and test set creation assuming that we are working with a single cancer type from our combined dataset. In Section

2.8, we describe the other train and test set variants utilised. To deal with class imbalances highlighted earlier on the curation of our gold standard labels, we downsampled the dataset using uniform random sampling to ensure an equal number of synthetic lethal and non-synthetic lethal pairs. We then divided it into train and test sets with a 70/30 split via uniform sampling. We centred each feature of the training data by subtracting the mean and then scaled each feature by dividing it by the standard deviation. This standardisation is required due to the use of regularisation with the logistic regression models, as scaling a feature will affect the size of the coefficient and thus alter how that feature is penalised. The feature means and standard deviations found on the train set were used to standardise features in the test set. We removed near-constant features, which we defined as those where less than 95% of the values within the feature vector are constant.

Repeated k-fold cross-validation. To select the best hyperparameters for each model, we carried out repeated, stratified, and nested grid-search cross-validation as described in Krstajic *et al.* (2014). We defined a grid for the hyperparameters for each model, and repeated k-fold cross-validation was conducted on the train set with $k = 10$ and 5 repeats. This means that one-tenth of the training data was held out for evaluation of the selected parameters, while the model was trained on the other nine-tenths. The purpose of repeating the cross-validation is to reduce the variance in the chosen parameters. Where possible, we evaluated the parameters' performance at each element of the grid using the area-under-the-receiver-operating characteristic curve (AUROC, described in Section 2.7). However, the L0Learn package is limited to the use of logistic loss as its fitness function for tuning the L0L2 model. The logistic loss $\mathcal{L}$ is defined as follows:

$$\mathcal{L}(y, \bar{y}) = \sum_{i=1}^{n} -y\ log(\bar{y}_i)) - (1-y) log(1-\bar{y}_i), \qquad (12)$$

where n is the number of examples in the validation set, $y_i \in \{0, 1\}$ is the actual class label and $\bar{y}_i \in [0, 1]$ is the logistic regression model output for the i^{th} example. Similarly, the MUVR package is limited to the use of misclassification error as its fitness function. The misclassification error is simply the number of misclassified class labels on the validation set. These metrics were averaged across the repeats and the parameters with the best average performance were selected as the parameters for the final model.

2.6 Baseline synthetic lethality prediction methods

We compared our models against three other published methods for predicting synthetic lethality: DAISY, DiscoverSL and PCA-gCMF (Jerby-Arnon *et al.*, 2014; Das *et al.*, 2019; Liany *et al.*, 2019). Here, we provide a brief description and motivate our choice of these baselines.

DAISY. DAISY was selected for being one of the earliest, and most widely cited, approaches to *in-silico* SL prediction Jerby-Arnon *et al.* (2014). DAISY is a statistics-based approach which utilises cancer cell-line and patient tumour data. DAISY relies on statistical hypothesis testing, consisting of three tests. The first test is termed "Genomic Survival of the Fittest" (GSOF) and is defined as the p-value of a two-tailed Wilcoxon rank-sum test for the change in somatic copy number alteration (SCNA) score for gene A in a pair between cell lines with and without inactivations in gene B. Jerby-Arnon *et al.* (2014) define the inactivation of a gene in a sample as when that gene possesses non-silent mutations or SCNA score of < -0.3. As Wilcoxon rank-sum test is employed to determine the significance of the change in the SCNA score of gene B between cell lines with and without inactivations in gene A, and then select the most significant p-value. The second test is similar to the first but calculates the significance in the difference in RNAi-based gene dependency scores instead of SCNA scores. The final test determines the significance of the Pearson's correlation between the expression of the two genes. This results in three p-values which are again combined using Fisher's method (for further details, refer to the "Extended Experimental Procedures" provided in the supplementary materials of Jerby-Arnon *et al.* (2014)). We implemented this method using by conducting the GSOF and Pearson's correlation tests across cancer cell line and patient tumour datasets and combining the p-values using Fisher's method. The RNAi-based gene dependency tests are limited to only cancer cell line data. We implemented DAISY in R.

DiscoverSL. DiscoverSL is a recently published SL classifier which uses a random forest model trained on breast and lung cancer data (Das *et al.*, 2019). The DiscoverSL model incorporates four features, calculated as previously described in Section 2.3: differential expression and expression correlation in tumour samples, pathway co-participation, and combined mutual exclusivity p-value. The pre-trained model was made public as an R package for classification of SL pairs. DiscoverSL outputs probabilities that gene pairs have SL relationships. We used this pre-trained model in our experiments as an example of a pan-cancer or "general" SL interaction prediction model.

PCA-gCMF. PCA-gCMF is another recently published machine learning approach which uses group-sparse collective matrix factorisation (gCMF) and principal component analysis to predict SL relationships (Liany *et al.*, 2019). It was selected as an example of matrix factorisation approaches to SL prediction.

PCA-gCMF uses a combination of principal component analysis (PCA) and group-sparse collective matrix factorisation (gCMF) to predict SL interactions. Low-rank matrix factorisation decomposes a single matrix $X \in \mathbb{R}^{p \times q}$ into a product of matrices such that $X \approx U_1 \cdot U_2^T$, where $U_1 \in p \times K$, $U_2 \in q \times K$, and $K < min(p, q)$. Group-space collective matrix factorisation generalises this idea to where we can decompose an arbitrary collection of M matrices which describe the relationships between E entities, $\{X_1, \cdots, X_M\}$, into E low-rank matrices, $\{U_1, \cdots, U_E\}$, while imposing a group-sparse penalty on these decomposed low-rank matrices (Klami *et al.*, 2013). Liany *et al.* (2019) use PCA transformations as a method to overcome a gCMF limitation whereby unique representations for each entity cannot be learned when multiple input matrices contain identical row and column entity-types. We used PCA-gCMF with four feature matrices as presented by Liany *et al.* (2019):

- *Gene dependency profiles.* An $p \times p$ matrix with p genes. For any gene pair (A, B), the p-value of a Wilcoxon rank sum test is calculated which quantifies the change dependency scores for gene A for cell-lines with and without a mutation in gene B. This test is conducted in reverse on (B,A) and the smallest p-value is used in the resulting matrix.
- *mRNA Expression profiles:* An $p \times q$ matrix with p genes and q patient tumour samples, where each element of the matrice is the mRNA expression count.
- *Co-expression:* An $p \times q$ matrice with p genes where the p-value for the Spearman correlation coefficient is measured between gene pairs across all patient tumour samples.
- *CNV profile:* An $p \times q$ matrice with p genes and q patient tumour samples, where each element of the matrice are the CNV real-valued scores produced by the GISTIC algorithm.

The original publication reported high prediction performance for PCA-gCMF on their included test sets ($AUROC \approx 0.9$). We downloaded the publicly available software for this method and used it directly. It should be noted that the solution space for matrix factorisation methods is a lot smaller than for classical machine learning techniques, such as logistic regression. This is because matrix factorisation methods can only make

predictions on relationships between entities, such as genes, provided in the input. For many practical applications, machine learning models need to be able to make predictions about relationships between previously unseen genes, also in new biological contexts. Ideally, to make comparisons fair between approaches we could initialise the input matrix defining gene-wise relationships to include an entry for every gene in the human genome. For our experiments, we found it sufficient to initialise the input matrices to include all unique genes represented in our combined dataset across all cancer types. PCA-gCMF outputs a continuous score per queried gene pair relationship.

2.7 Evaluation

We evaluated each model $f(\boldsymbol{x}_i)$ by measuring prediction performance against the ground truth y_i in different experimental scenarios. For our SL prediction models, we also determined the contribution of each feature, or element in $\boldsymbol{x}_i$, in the context of each model.

Prediction performance. One of the goals of SL prediction models is to reduce the number of candidate gene pairs for *in-vitro* experimentation to one with a higher probability of producing true-positive results. In practice, this could mean choosing the top k results from a list of n candidate gene-pairs output by a prediction model for further *in-vitro* experimentation, where $k << n$. Therefore, we are more interested in the true-positive predictions ranked higher in our results rather than the entire results space. More formally, a classifier $f(\boldsymbol{x}_i)$ should have the following property: $P(y_i = 1|\bar{y}_i) \propto \bar{y}_i$. This means that as our model output $\bar{y}_i$ increases, the probability that the actual class label y_i is truely positive should also increase.

We used receiver-operating characteristic (ROC) and precision-recall (PR) curves as a graphical representation of the ability of the binary classifier, $f(\boldsymbol{x}_i)$, to correctly predict the class label, y_i, by varying a threshold t over the real-valued range of output values $\bar{y}_i$. ROC curves are a standard form of evaluation of models in machine-learning literature. ROC curves are plotted with the false positive rate (FPR) and true positive rate (TPR) on the x and y-axes, respectively. A line is generated from the bottom left corner of the plot to the top right by varying the threshold, t, from the most stringent possible value where all examples are classified negatively, $t > max(\bar{y}_i)$, to the most lenient where all examples are classified as positive, $t < min(\bar{y}_i)$, and plotting the TPR and FPR values. Curves which come closer to the top-right corner of the plot are indicative of better performance, while curves which form a straight line from the origin to the point (1, 1) are indicative of random classification.

Given the aforementioned goal of our model, we were interested in observing the precision of our classifier as we predict more positive labels from our test set. PR curves achieve this by representing the trade-off between recall and precision, which are plotted respectively on the x and y-axes. This produces a curve from the top-right point (0,1) to the point (1, $minprec$), where $minprec$ is the precision when all points are classified positively. Curves which come closer to the top-left corner of the plot are indicative of better performance. We averaged the ROC and PR curves across test set evaluations using the vertical-averaging method (see Fawcett (2006) for details).

We determined the area-under-the-curve (AUC) to summarise the performance of these classifiers, denoted AUROC and AUPRC for the corresponding ROC and PR curves, respectively. For both curves, an AUC closer to 1 indicates better performance. In the case of AUROC, a value of 0.5 indicates random performance. The AUPRC value which indicates random performance in a binary classification problem is equal to that of the class balance ratio for the test set. Our test sets are always class balanced, and thus this value is 0.5 for our experiments. However, determining the AUROC and AUPRC over the entire result space may not be the most informative measure of performance. As mentioned, one of the main practical applications of these methods would be to choose the top k results for *in vitro* experimentation based on their likelihood to produce true positive results.

Algorithm 1: Model-Agnostic Permutation Feature Importance

Result: List of 4-tuples of variable name and median, 5th quantile, and 95th quantile feature importance score, ordered by median score

$T \leftarrow Test\ dataset$;
$M \leftarrow Some\ model$;
$L \leftarrow Some\ loss\ function$;
$l_{orig} \leftarrow L(M, T)$;
$importances \leftarrow \{\}$;
for *var in M.variables* **do**
  $j \leftarrow \{\}$;
  for *i in 1:100* **do**
    $T_{perm} \leftarrow Permute(T, var)$;
    $l_{perm} \leftarrow L(M, T_{perm})$;
    $j.insert(l_{perm}/l_{orig})$
  end
  $importances.insert(\{var, median(j),$
  $05quantile(j), 95quantile(j)\})$
end
$importances \leftarrow order(importances)$;

To assess the performance at the top of the ranking, we also determined the Average Precision at rank k, AP@k, defined as the average precision over the top k ranked results as follows:

$$AP@k = \frac{1}{k}\sum_{i=1}^{k}\frac{TP(i)}{i}. \tag{13}$$

We took $k = floor(n/3)$, where n is the number of examples in the test set. $TP(i)$ is a function which returns the number of true positives in the first i results. The results are ordered by the model output $\bar{y}_i$ for these test examples as described in the problem formulation. We will refer to this value in later sections as AP@n/3. We evaluated the performance of every final model against the test set. To examine the variance in model performance we repeated the process, from downsampling to training to evaluation against the test set, 10 times. We calculated averages and standard deviations for all metrics across all 10 runs.

Feature importance. We measured feature importance (FI) using a model-agnostic permutation approach as proposed by Fisher *et al.* (2019) (Algorithm 1). We chose this to be able to generate feature importance scores in a consistent manner and improve interpretability across models. This algorithm provides an estimate per feature of the median, 5th, and 95th quantile of the changes in the prediction error, defined as $1 - AUROC$. This is done for the model when the values of each of the features are permuted. A median increase in error which is generally > 1 with a low variance indicates that a feature is "important". Features which are not important will cause no change in the error and have a value ≈ 1. A median decrease in the error of < 1 can happen when the permutation of an unimportant feature randomly leads to a reduction in error.

This metric can be calculated using the train or test set for T, as defined in Algorithm 1. As noted by Molnar (2020), the choice of training or test set comes down to what we mean when we describe the feature importance of our models. Feature importance (FI) is an ill-defined concept in the literature. To define feature importance, we need to pose the question of what we are trying to observe. If our question is: “what features are the most important to my model for predicting correctly on unseen data?”; then

we should use the test data. This is because our model could be positively biased towards certain features which it may have overfit on during the training process. Conversely, if our question was "what features has my model learned to use to make predictions?", then we would use the training data (refer to Molnar (2020), Chapter 5.5 for more details). In our case, we are interested in the former, and thus we will calculate feature importance on the test data.

Multicollinearity. A confounding factor to the interpretability of many machine learning models is multicollinearity. Multicollinearity is a phenomenon where one feature in a model may be linearly predicted using some combination of the other features. Multicollinearity can make the interpretation of feature coefficients or the relative ranking of feature importance difficult. For example, from a set of multicollinear features, a model may randomly select only one of these features as "important". Interpreting the feature importances from such a model directly may lead to the false impression that any features that are collinear with the chosen feature do not hold any predictive value. Multicollinearity can be assessed on a per-feature basis using variance inflation factors (VIF). The VIF is calculated using linear regression, where the feature j is the response variable, and every other feature in the model are explanatory variables. From this linear model, we can determine the multiple R^2 value, which measures the amount of variation in the response variable that can be explained by the predictor variables (Draper and Smith, 1998). The R^2 value for this model is used to calculate the VIF for feature j as follows:

$$VIF_j = \frac{1}{1 - R_j^2}. \tag{14}$$

Large multiple R^2 values imply that much of the variation in a feature can be predicted by other features in the model. In return, this will result in a large VIF value and indicates the presence of multicollinearity. Conversely, a multiple R^2 value close to 0 will result in a VIF close to 1, indicating the absence of multicollinearity. A high VIF which might highlight a problem is often considered to be > 5 or 10, although the ideal threshold is specific to each problem (Becker *et al.*, 2015).

2.8 Model generalisation and effect of selection bias

We wish to quantify the resilience of a learned mapping function, $f(\boldsymbol{x}_i)$, to biases within the dataset used to train the model. Recall that a dataset D encompasses both the gold-standard labels y and the feature vector $\boldsymbol{x}$ for each gene-pair as described in Section 2.1. Consider two datasets D_{source} and D_{target} where $P(\boldsymbol{x}_i|D_{source}) \neq P(\boldsymbol{x}_i|D_{target})$. The model trained on dataset D_{source}, $f_{D_{source}}(\boldsymbol{x}_i)$, should be learned such that $P(y_i = \hat{y_i}|\boldsymbol{x}_i \in D_{source}) \approx P(y_j = \hat{y_j}|\boldsymbol{x}_j \in D_{target})$ where $\hat{y_i} = f_{D_{source}}(\boldsymbol{x}_i) > t$, where t is a constant threshold. Ideally, this means that as the underlying distributions of the feature vectors $\mathcal{X}$ change, the learned relationship $f(\boldsymbol{x}_i)$ to the outcome variable y_i should not change. To examine how these models are affected by the aforementioned biases, we train and test said models against datasets D_{source} and D_{target}, respectively, where $P(\boldsymbol{x}|D_{source}) \neq P(\boldsymbol{x}|D_{target})$. We then evaluate and compare the predictive performance of these models to assess their ability to generalise to different datasets of different distributions. In this section, we enumerate the experiment variations conducted to perform this analysis. Each variation involves the construction of datasets, D_{source} and D_{target}, and is informed by which research question we are aiming to answer. We define the set of cancer types as $\mathcal{C} = \{\text{BRCA}, \text{COAD}, \text{LUAD}, \text{OV}\}$ and the datasets used for constructing our training and test sets as D, D_{isle}, and D_{dsl}, for the combined, isle, and discoverSL datasets, respectively.

We also investigate the effects of sampling biases. Sampling bias describes a situation where a population is sampled in such a way that some members have a higher probability of being sampled than others and, in turn, the distribution of these sampled data does not reflect the true distribution of the underlying population. We are interested in how two sources of sampling bias in the curation of a gold standard SL labels can affect the training and evaluation of SL classifiers. We term these biases cancer representation bias and gene selection bias. We will describe these biases later as we detail each of the experiments.

Cancer-specific models

The features we have detailed previously, in Section 2.3, describe different relationships between gene pairs within the biological context from which the raw measurements were taken, such as a tumour of a particular cancer type. These differing contexts could give rise to differing biological markers for SL identification. Accordingly, our initial aim was to determine if the knowledge a model gained through learning on one cancer type could generalise to effectively predict SL in a different cancer type.

Cancer-specific model performance. To establish a baseline for comparison in later experiments, we analysed the performance of cancer-specific models when trained and tested on disjoint sets of examples from the same cancer type. To achieve this we split D into four datasets, one per cancer type, D^c. Train and test sets D^c_{source} and D^c_{target} were then created separately from each cancer type dataset, D^c, using a 70/30 split as described in Section 2.5, "Training and test sets".

Cross-cancer generalisation. To evaluate the ability of the cancer-specific models to transfer knowledge across cancer types, we measured the prediction performance of models trained on each individual cancer type against three different test sets, each containing samples from one of the other three cancer types. We constructed the training and test datasets as per the per-cancer experiment above. Then for each model trained on a dataset D^c_{source}, we tested that model against every other cancer type d test set, i.e. $D^d_{target} \forall c$, where $d \neq c$.

Pan-cancer models

Despite inherent differences in biological tissue and molecular landscape, using data from different cancers could provide complementary information that is useful to learn more general rules that determine when a pair of genes is synthetic lethal. We, therefore, set to assess if incorporating knowledge from other cancers together with that of a select cancer type could improve prediction on this same cancer type. Additionally, we also consider cancer representation bias. This is a form of sampling bias which can occur due to the popularity of certain cancer types studied in the literature or the rarity of other cancers, leading to an over- or under-representation of available data per cancer type. We examine if such biases can have an impact on pan-cancer learning and if their effects can be mitigated through the balancing of cancer representation. Finally, we inspect how pan-cancer models generalise to previously unseen cancer types.

Pan-cancer model performance. To determine if predictive performance on a particular cancer type could be improved by incorporating knowledge from other cancers, we needed to construct a pan-cancer dataset. To accomplish this, we constructed four datasets (one per cancer) and split them into training and test, D^c_{source} and D^c_{target}, as described in the per-cancer experiment above. We created the pan-cancer training set as the union of the four training sets, $D^u_{source} = \{D^c_{source} : c \in C\}$. The trained models were then evaluated against every one of the four cancer-specific test sets D^c_{target}.

Resilience to cancer selection bias. We also investigated the influence of cancer type representation within a dataset on model prediction performance. For this, the union training set D^u_{source} was downsampled

to generate a new training set D^b_{source} containing an equal number of positive and negative samples per cancer type. As before, the trained models were then evaluated against each cancer-specific test sets D^c_{target}, and the performance was compared to that of the models trained on D^u_{source}.

Cross-cancer generalisation (leave-one-cancer-out). Subsequently, we examined how models leveraging knowledge gained from multiple cancer types could generalise to an unseen cancer type. This was done by holding out one cancer type c from the pan-cancer training set described above and testing the trained model against the held-out cancer type. Again, beginning with our per-cancer datasets as constructed in the per-cancer experiment, the training set is constructed such that for each held-out cancer c, $D^u_{source} = \{D^{nc}_{source} : nc \in \mathcal{C}, \forall nc \neq c\}$. D^u_{source} is then downsampled as in the pan-cancer experiment to produce D^b_{source}. Models are trained on D^b_{source} and are evaluated against D^c_{target}.

Effects of gene selection bias

Gene selection bias occurs when gene pairs are sampled non-uniformly from all available genes in the human genome. This typically occurs due to the popularity of certain genes in the literature or their likelihood to be involved in cell processes. In turn, as with cancer selection bias, this leads to an unequal representation of genes across gold standard datasets. We created dot plots to investigate gene selection bias in the ISLE and DiscoverSL gold standard datasets. In these plots, the x-axis and the y-axis both contain entries for every unique gene included in either dataset. Dots appear on the plot where either a positive or negative label exists between each pair of genes, coloured by the SL gold standard dataset. Obvious patterns in these plots, such as long sequences of consecutive dots, may highlight gene-selection biases. Essentially, we are visualising the adjacency matrix for gene pairs in either SL gold standard. Our hypothesis was that models that account for the structure of such a matrix, such as matrix factorisation models, will be more susceptible to effects of gene selection bias than other models.

Cross-gold standard generalisation. Since the ISLE and DiscoverSL gold standard SL datasets showed different gene selection biases, we did a preliminary analysis to investigate model generalisation ability across SL datasets. For this experiment, we first selected a single cancer type $c \in \{\text{BRCA, LUAD}\}$. We then chose one dataset, either D_{isle} or D_{dsl}, to act as the source of the training data, and the other dataset was used for testing. For example, when $c = BRCA$, we randomly sampled $BRCA$ instances from D_{isle} to construct a class-balanced training set $D^c_{isle,source}$. Recall from Section 2.2 that when curating our datasets, we remove any pairs that may be duplicated between these sets. The same was done to construct the test set $D^c_{dsl,target}$ from D_{dsl}. When $c = LUAD$, we constructed the training and test sets from D_{dsl} and D_{isle}, respectively.

Resilience to gene selection bias (gene holdout). We then investigated the effect of gene selection bias in the gold standard SL datasets on the prediction performance of the best performing machine learning models under two different scenarios. In the first scenario, training and test sets were constructed such that for every gene pair (A, B) in the test set, either A or B was present in the training set, while the other was not. In the second scenario, the sets were created such that for every gene pair (A, B) in the test set, neither gene A or B appeared in the training set. Data from the first cancer-specific experiment was used as a baseline to compare model performances against these two scenarios. We used only BRCA data for this experiment since it was the only cancer type with enough diversity in its gene pairs to construct reasonably sized training and test sets under the given constraints.

2.9 Other experiments

Gene dependency-based feature reliance analysis

We conducted an experiment to examine how the predictive power of our models was affected when we removed gene dependency-based features from the feature set. We trained and tested our models on two pan-cancer datasets, one which included all features, and one which excluded any feature which was generated using gene-dependency scores. Both datasets were constructed from the combined dataset, which was downsampled through uniform sampling to include an equal number of positive and negative labels across all cancer types. The datasets were split 70/30, also through uniform sampling. Identical indices were used to create the training and test splits in both datasets, i.e. the only difference between the training and test sets was the number of included features. Training and evaluation were performed as described earlier in this section.

3 Results and Discussion

In this section, we describe and discuss the results of our experiments designed to analyse the predictive performances of our cancer-specific and pan-cancer SL prediction models. We detail their generalisation and the effects of representation and selection biases on training and evaluation. Our initial set of experiments analysed predictive performances in a cancer-specific context. We compared the performances of our models with those of our selected baselines: DAISY, DiscoverSL, and PCA-gCMF. We then examined the effect of cancer representation bias on SL prediction. This included: examining how models trained on a single cancer type performed when tested against the other cancer types; investigating whether training models on datasets containing all available cancer types could help improve the performances in cancer-specific test sets; assessing whether pan-cancer models would be able to leverage knowledge from multiple cancers to generalise to unseen cancer types. Following these, we examined gene selection bias in gold standard data and quantified the sensitivities of different models to such biases. Finally, we analysed the feature importances for each model.

3.1 Linear models show best precision at the top

We first evaluated the prediction performance of the regularised logistic regression (L0L2, Elastic Net) and random forest (MUVR, RRF) models trained using all 26 features in a cancer-specific context. Four models were trained and tested separately using each algorithm, one per cancer type: BRCA, COAD, LUAD, and OV. We compared these models against the selected baseline methods DAISY, DiscoverSL and PCA-gCMF based on AUROC, AUPRC and AP@$(n/3)$ values. This experimental setup represents the simplest scenario for an SL prediction task, where the train and test data comes from the same distribution. Both the DAISY and DiscoverSL methods performed close to random across all cancer types with respect to both AUROC and AUPRC, and will not feature prominently in further analysis (Tables 4a and 4b).

Our results show that our cancer-specific models and PCA-gCMF performed well on the BRCA and LUAD datasets. On BRCA data, our regularised logistic regression (Elastic Net, L0L2) and random forests (MUVR, RRF) achieved average AUROCs of 0.84 to 0.86, and average AUPRCs of 0.88 to 0.89. PCA-gCMF performed slightly better with an average AUROC of 0.92. On LUAD data, regularised logistic regression and random forests showed consistently high AUROC (0.85 to 0.87) and AUPRC (0.87), while PCA-gCMF had an AUROC of 0.87 and a lower AUPRC of 0.81 (Tables 4a and 4b). Precision-recall curves showed that logistic regression and random forest models exhibited high precision among higher-ranking positively predicted gene pairs (recall between 0 and 0.25 or 0.3), whereas the precision of PCA-gCMF remained significantly

	BRCA	COAD	LUAD	OV
DAISY	.61 ± .02	.38 ± .02	.44 ± .03	.41 ± .04
DiscoverSL	.54 ± .02	.54 ± .02	.54 ± .03	.45 ± .04
Elastic Net	.84 ± .01	.6 ± .02	.85 ± .02	.59 ± .03
L0L2	.84 ± .01	.6 ± .02	.85 ± .02	.59 ± .03
MUVR	.86 ± .01	.64 ± .01	.86 ± .01	.54 ± .07
pca-gCMF	**.92 ± .01**	.54 ± .03	.87 ± .03	**.94 ± .02**
RRF	.86 ± .01	**.63 ± .02**	**.87 ± .02**	.57 ± .07

(a) AUROC

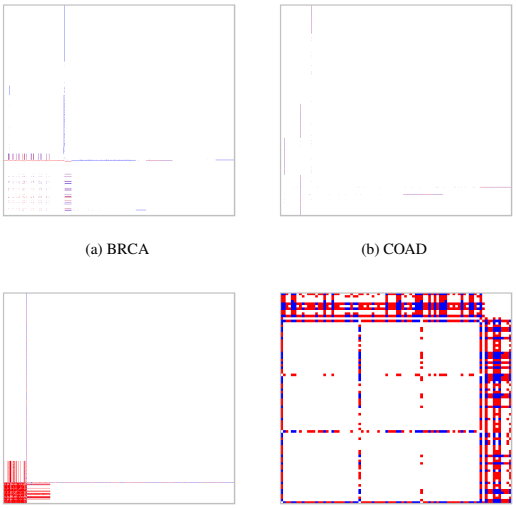
	BRCA	COAD	LUAD	OV
DAISY	.58 ± .02	.42 ± .01	.46 ± .03	.47 ± .04
DiscoverSL	.55 ± .02	.53 ± .02	.54 ± .03	.48 ± .04
Elastic Net	.87 ± .01	.59 ± .01	.87 ± .02	.58 ± .04
L0L2	.88 ± .01	.59 ± .02	.87 ± .02	.58 ± .05
MUVR	.89 ± .01	.62 ± .01	**.87 ± .01**	.51 ± .05
pca-gCMF	**.92 ± .01**	.56 ± .03	.81 ± .06	**.92 ± .04**
RRF	.89 ± .01	**.63 ± .02**	.87 ± .02	.54 ± .05

(b) AUPRC

	BRCA	COAD	LUAD	OV
DAISY	.6 ± .04	**.77 ± .04**	.42 ± .07	.52 ± .1
DiscoverSL	.6 ± .05	.54 ± .03	.58 ± .07	.5 ± .12
Elastic Net	.99 ± .01	.67 ± .04	**.97 ± .01**	.62 ± .11
L0L2	**.99 ± 0**	.66 ± .05	**.97 ± .01**	.61 ± .12
MUVR	**.99 ± 0**	.68 ± .03	.95 ± .02	.44 ± .11
pca-gCMF	.97 ± .02	.66 ± .04	.79 ± .11	**.93 ± .06**
RRF	.99 ± .01	.72 ± .04	.95 ± .02	.51 ± .12

(c) Average precision over the first third of ranked predictions.

Table 4. AUROC, AUPRC, and AP@$(n/3)$ estimates for cancer-specific models tested on the same cancer type. Mean and standard deviation of 10 repetitions. AP@$(n/3)$ denotes average precision over the first third of ranked predictions.



(a) BRCA (b) COAD (c) LUAD (d) OV

Fig. 3: Adjacency matrix dot plot showing labelled gene pairs from BRCA, COAD, LUAD, and OV cancer types in the combined dataset. The x-axis and y-axes are the unique genes per cancer type gene pairs. Each dot represents a gene pair where a positive (blue) or negative (red) label exists. Whitespace represents gene pairs with no labels. These matrices are symmetric along the diagonal.

lower (Fig. 4). This was quantified by the average precision calculated over the first third of the ranked results, where all logistic regression and random forest models scored higher than PCA-gCMF (Table 4c). The difference between these models was particularly significant for LUAD models.

On COAD data, our logistic regression and random forest models performed worse in comparison to their performance on BRCA and LUAD data, with AUROCs between 0.60 and 0.64 (Table 4a). We hypothesize that different characteristics of cancer datasets could reduce the effectiveness of certain predictors. For example, COAD has a higher prevalence of microsatellite instabilities (MSI) (Bonneville *et al.*, 2017). MSI are regions of hypermutability in the genome caused by the loss of DNA mismatch repair activity. Hypermutability may add noise to some mutation-based predictors, such as $CRISPR\_dep\_stat$, as it becomes harder to determine if changes in a measured phenotype, like gene dependency, are correlated with a particular mutational event (Behan *et al.*, 2019). As our later results in Section 3.5 demonstrate, mutation-based features like $CRISPR\_dep\_stat$ are of high importance to our BRCA and LUAD models. On OV, our logistic regression models outperformed our random forest models slightly, with AUROCs of 0.59 for both L0L2 and Elastic Net, while MUVR and RRF had AUROCs of 0.54 and 0.57, respectively (Table 4a). We advance two potential reasons for the lower performance of our models on OV data. First, the number of samples available for training and testing of OV models was approximately one-third of that available for BRCA and LUAD (Table 2). This is a good example of cancer representation bias. The second reason for this might be that the OV cell lines contain a much lower number of mutations per gene pair than the other cancer types (OV: 1.6 mutations per gene pair on average, BRCA: 4.33, LUAD: 11.35, COAD: 5.97). Lack of mutation data may cause similar issues with mutation-based features as hypothesised for COAD.

Our results on COAD and OV data demonstrated interesting behaviour on behalf of our baseline matrix factorisation method, PCA-gCMF. For COAD, PCA-gCMF performed close to random (AUROC: 0.54), while obtaining very high performance on OV data (AUROC: 0.94). We noted that COAD and OV gold standard class labels display very different characteristics. The COAD gold standard labels are very sparse, and only 105 of the 1560 unique genes are featured in more than one gene pair (Fig. 3b). Only five genes appear in more than three labelled pairs, with roughly even priors of being involved in SL interactions (BLM: 0.57, KRAS: 0.5, MUS81: 0.5, PTEN: 0.54, PTTG1: 0.48). Conversely, the number of unique genes in the SL gold standard for OV is at most 10% that of other datasets (OV: 83 unique genes, BRCA: 1072, LUAD: 804, COAD: 1560) and all genes except two are featured more than twice as members of a gene pair within the dataset (Fig. 3d). In comparison to COAD, it is very likely for individual genes to be featured in both the train and test sets. The performance of PCA-gCMF in this context is in line with our hypothesis that matrix factorisation techniques will be less effective when predicting on previously genes with no *a priori* class label information. We provide further evidence for this in Section 3.4.

Overall, these observations show the potential of our models in classifying SL interactions on BRCA and LUAD data while displaying the most desirable property of ranking the most promising candidates more consistently at the top. However, our COAD and OV results illustrate that, while our models might demonstrate high efficacy on certain cancer types such as BRCA and LUAD, our models are not as effective across all cancer types.

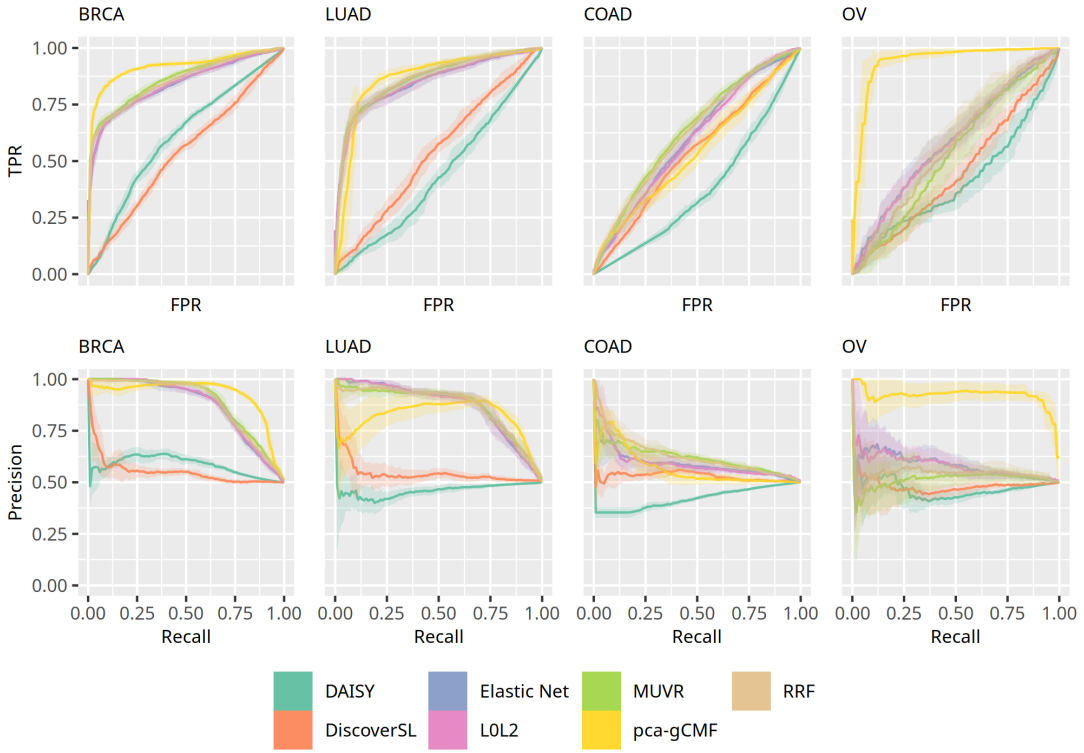


Fig. 4: Receiver-operating characteristic (ROC) curves and precision-recall (PR) curves for each cancer-specific model tested against samples from the same cancer type that it was trained on. The top plots show ROC curves, the bottom plots show PR curves, and each column corresponds to a different cancer type.

3.2 Cancer representation affects pan-cancer learning

We experimented to ascertain whether our models could improve their predictive performances on a single cancer type by leveraging information from other cancers. We also investigated whether pan-cancer model learning is affected by cancer representation bias. In this context, we trained separate models on two training sets: one with and one without equal cancer representation. We then evaluated these models against each of our four cancer types in turn.

Our results show that the pan-cancer random forest models outperformed the pan-cancer logistic regression models with respect to average AUROC in every case except one (Table 5). They also illustrate that the pan-cancer models performed worse than their cancer-specific counterparts and the difference between the two was larger in general for our logistic regression models compared to our random forest models. We also noted that the random forest models were less affected by training on a dataset with unbalanced cancer representation. The random forest models suffered a drop of between 0.05 and 0.06 in AUROC when trained on a dataset with unbalanced instead of a balanced cancer representation, whereas the predictive performance of the logistic regression models dropped by 0.11, to 0.12.

From these results, it is difficult to conclude that knowledge from other cancers cannot be leveraged to improve predictive performance on a target cancer. Our results suggest that training on multiple cancers could dilute our models ability to predict SL on individual cancer. This seems especially true for logistic regression models. However, the fact that our cancer-specific models performed poorly on COAD and OV data may also partially explain the drop in performance when we use this data to train our pan-cancer models. Nonetheless, our results do demonstrate that cancer representation is an issue which must be considered when constructing datasets for pan-cancer learning. As we examine the columns in Table 5 from right to left, we note that the AUROCs of our logistic regression models drop more between columns than do the AUROCs of our random forest models. A possible explanation for this is that the logistic regression models average out the effects across cancers, and the impact of this worsens as the cancer representation becomes more unbalanced. Conversely, the non-linearity in the random forest models enable context-specific rules which may empower its resilience when exposed to data from multiple cancers at different ratios. Finally, we demonstrate that balancing cancer representation can lead to an improvement in pan-cancer model learning, highlighting that future research on machine learning modelling of pan-cancer SL prediction should take cancer representation into consideration.

	Pan-cancer (all cancers)		
	Unbalanced	Balanced	Single cancer
BRCA	.64 ± .02	.75 ± .01	**.83 ± .01**
COAD	.52 ± .02	.51 ± .02	**.6 ± .02**
LUAD	.73 ± .03	.79 ± .02	**.83 ± .02**
OV	.4 ± .04	.53 ± .04	**.58 ± .03**

(a) L0L2

	Pan-cancer (all cancers)		
	Unbalanced	Balanced	Single cancer
BRCA	.65 ± .02	.77 ± .02	**.84 ± .01**
COAD	.52 ± .02	.53 ± .02	**.6 ± .02**
LUAD	.74 ± .02	.8 ± .02	**.85 ± .02**
OV	.4 ± .04	.5 ± .04	**.59 ± .03**

(b) Elastic Net

	Unbalanced	Balanced	Single cancer
BRCA	.76 ± .01	.82 ± .02	**.86 ± .01**
COAD	.62 ± .02	.6 ± .01	**.64 ± .01**
LUAD	.81 ± .02	.83 ± .02	**.86 ± .01**
OV	**.55 ± .06**	.52 ± .04	.54 ± .07

(c) MUVR

	Unbalanced	Balanced	Single cancer
BRCA	.75 ± .02	.8 ± .02	**.86 ± .01**
COAD	.62 ± .02	.61 ± .02	**.63 ± .02**
LUAD	.8 ± .02	.83 ± .02	**.87 ± .02**
OV	.55 ± .04	.53 ± .05	**.57 ± .07**

(d) RRF

Table 5. AUROC estimates for pan-cancer models (with unbalanced or balanced cancer representation) and cancer-specific models tested on each individual cancer type. Mean and standard deviation of 10 repetitions. Models: logistic regression models with L0 and L2 regularisation (L0L2) or with L1 and L2 regularisation (Elastic Net), and random forest models MUVR and RRF.

3.3 Our models can generalise between certain cancers

Next, we investigated how our models generalised to unseen cancer types. In the case of our cancer-specific models, we tested these models on the test sets for each of our four cancers, individually. For our pan-cancer models, we did this by holding out one cancer type, training models using samples from the other three, and then testing on the samples from the held out cancer type. We dub this experiment as "leave-one-cancer-out". For this section, our conclusions are the same regardless of our choice of linear or non-linear models. For brevity, we primarily focus on L0L2 and MUVR models. The interested reader can optionally refer to the supplementary materials where indicated for more details.

First, we look at the results for our cancer-specific models. We see that for the L0L2 models, when trained on BRCA and tested against LUAD, and vice versa, the models generalised well with average AUROCs of 0.79 and 0.69, respectively (Fig. 5, top) . Elastic Net behaved similarly (Supplementary Figure S5, top-right). We hypothesize that the logistic regression models can generalise between BRCA and LUAD because the BRCA and LUAD datasets possess similar linear relationships between predictors and class labels. We will explore this further in Section 3.5. MUVR generalised well when trained on BRCA and tested against LUAD (AUROC: 0.73), but not when trained on LUAD and tested on BRCA (AUROC: 0.53) (Fig. 5, bottom). This behaviour is consistent with that of the RRF model (Supplementary Figure S5, bottom-right). All models struggled to generalise when trained on OV or COAD to other cancers, with predictive performance varying from poor to random when tested against other cancer types (Fig. 5 and Supplementary Figure S5).

The results of our leave-one-cancer-out experiments for pan-cancer L0L2 and MUVR models are shown in Fig. 6, bottom. Both the L0L2 and MUVR models performed well on LUAD (AUROC: 0.78 and 0.72), possibly owing to the inclusion of BRCA data in the training set (Fig. 6). As observed in earlier cross-cancer experiments, all our cancer-specific models were able to predict well on LUAD data when they had been trained on BRCA data (Fig. 5). On BRCA, the L0L2 model had a modest AUROC of 0.67 and the MUVR model exhibited near-random performance with an AUROC of 0.52 (Fig. 6). This is also in accordance with the cross-cancer experiments showing that cancer-specific MUVR models trained on COAD, LUAD or OV did not generalise to BRCA at all (with near-random AUROCs), while the L0L2 model was able to leverage the LUAD data to predict on BRCA with an AUROC of 0.71 (Fig. 5). The Elastic Net and RRF models performed similarly to the L0L2 and MUVR models, respectively (Supplementary Figure S6). We also note that the leave-one-cancer-out models performed similarly to the best generalising

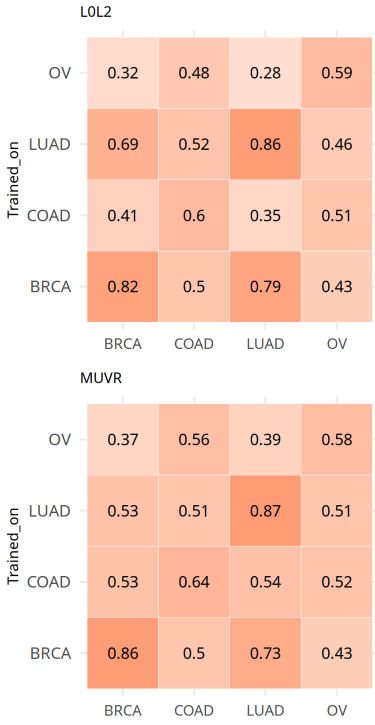


Fig. 5: Heatmaps of average AUROC performances for the L0L2 (top) and MUVR models (bottom) over 10 repetitions. On the y-axis is the cancer type that each model was trained on. On the x-axis is the cancer that each model was tested against.

cancer-specific models from the cross-cancer experiment. For example, pan-cancer models trained on COAD, LUAD and OV generalised equally as well to BRCA as did their cancer-specific counterparts trained only on LUAD (Fig. 5 and Fig. 6, and Supplementary Figures S5 and S6). This might indicate that training on multiple cancer types does not have a strong

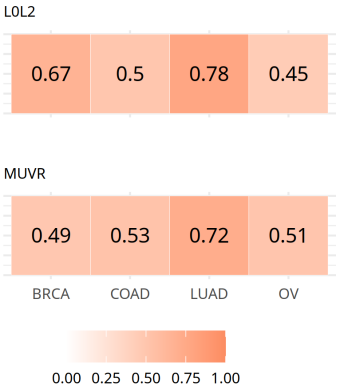
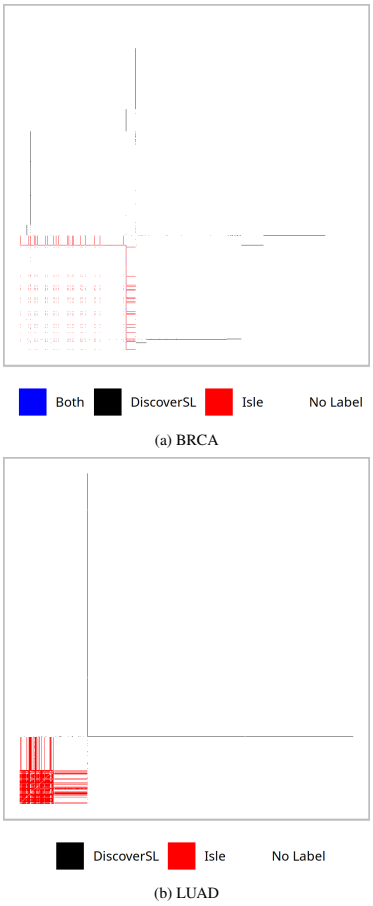


Fig. 6: Heatmaps of average AUROC performances for L0L2 models (top) and MUVR random forest models (bottom) over 10 repetitions. On the x-axis is each cancer type that was held out for testing. The models were trained on a balanced dataset of all other cancers.

negative effect on a models predictive power when generalising to unseen cancers. All models exhibited poor-to-random performance when OV and COAD were held out for testing, similar to the cross-cancer generalisation experiment.

Overall, our results demonstrated that our cancer-specific and pan-cancer logistic regression models could generalise between BRCA and LUAD datasets. Random forest models were shown to generalise from BRCA to LUAD, but not vice-versa. Training models on multiple cancer types did not improve a models' ability to predict on samples from a previously seen cancer type or generalise to unseen cancer data. Nonetheless, it is also encouraging to note that the models' abilities to generalise to unseen cancers only marginally decreased when trained pan-cancer. This is despite being trained on additional data in COAD and OV which have proved to be difficult problems for our models to solve in our previous experiments (Table 4a). This suggests that our models may still be able to learn and handle multiple contexts, which could form a basis for pan-cancer SL modelling.



(a) BRCA

(b) LUAD

Fig. 7: Adjacency matrix dot plot showing BRCA (a) and LUAD (b) gene pairs in the ISLE and DiscoverSL gold standard datasets. Elements along the x-axis and y-axis represent unique genes in the combined dataset. Each dot represents a gene pair with a label in the ISLE (red), DiscoverSL (black), or both (blue) gold standards. Whitespace denotes gene pairs unknown to either gold standard dataset.

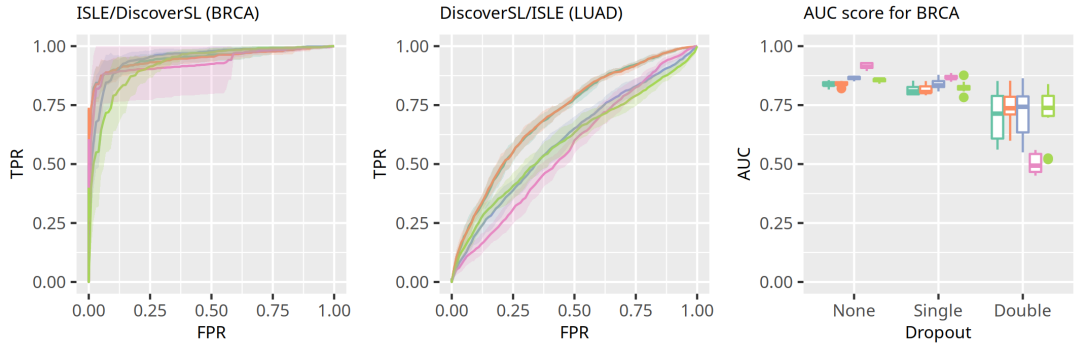
3.4 Complex models are sensitive to gene selection bias

Our next goal was to understand the impact of gene selection biases on classifier generalisability. We observed some gene selection bias in the gold standard gene pairs by visualising the adjacency matrix dot plots of the ISLE and DiscoverSL gold standard SL datasets. We saw distinct lines of labelled pairs when we stratified the combined dataset by cancer type and gold standard SL dataset (DiscoverSL or ISLE). Each horizontal or vertical line represents experimentally validated SL labels that exist between a single gene and other genes in the dataset. One striking pattern for DiscoverSL in LUAD was that all pairs involved one particular gene. In this case, the labels were originally obtained from a single double knockdown gene experiment which focused on KRAS as the cancer driver gene. The ISLE dataset featured more diverse pairwise gene combinations but covered a far smaller set of unique genes (Fig. 7b). Both datasets very clearly had different structural characteristics due to individual gene and gene pair selection biases (Fig. 7).

Linear models generalise better across SL gold standards

We first sought to assess which models would be most susceptible to gene-selection biases in SL gold standard class labels. We trained our models and PCA-gCMF on one cancer type based on one gold standard SL dataset and tested on the same cancer type based on the other gold standard. We first trained models based on the pairs in the ISLE gold standard and tested them against the pairs in the DiscoverSL gold standard, using BRCA data. Secondly, we trained the opposite scenario, this time using LUAD data. The cancer types chosen were due to data restrictions and an intent to have an adequate and equal number of class labels for training the models, consistent with all the other experiments.

Our results show that logistic regression methods generalised better than the random forest and PCA-gCMF methods in both cases (Fig. 8a and 8b, Table 6). Our random forest models displayed a large difference in performance when generalising between the two BRCA gold standards and



(a) BRCA: Trained on ISLE, Tested on DiscoverSL

(b) LUAD: Trained on DiscoverSL, Tested on ISLE

(c) Gene Holdout

Fig. 8: ROC curves of models trained using one gold standard dataset and tested using the other gold standard dataset for BRCA (a) and LUAD (b). The BRCA models was trained on ISLE and tested on DiscoverSL gold standards. The LUAD models were trained on DiscoverSL and tested on ISLE. The curves are averaged over 10 runs. (c) Boxplots of AUROC for the gene holdout experiment (10 runs per boxplot). The y-axis denotes the AUROC value. Categories on the x-axis denote one of three scenarios: for "None", we only guaranteed that the training and test sets of gene pairs were disjoint; for "Single", only one of the genes out of every gene pair in the test set was present in the training set; and for "Double" neither gene of a gene pair in the test set appeared in the training set.

the two LUAD gold standards. One possible explanation for our logistic regression models' superior abilities to generalise is their simpler linear nature. By comparison, the complexity of the random forest model may inform it's tendency to overfit.

PCA-gCMF showed the largest difference in AUROC between the BRCA and LUAD experiments (BRCA: 0.93, LUAD: 0.56), and performed close to random on LUAD with a large standard deviation on BRCA (0.07) across 10 runs. Interestingly for PCA-gCMF, the ISLE and DiscoverSL gold standards possess similar characteristics to those of the COAD and OV gold standards as described in Section 3.1. For the BRCA data, 522 gene pairs out of the 907 in DiscoverSL contained genes which also appeared in ISLE. This means that uniformly sampled train and test sets would have a high probability of sharing genes, comparable to OV. Conversely, DiscoverSL and ISLE LUAD gold standards only share 19 genes, each partaking in a single labelled gene pair. This would produce train and test sets with little gene overlap, akin to COAD. Likewise, PCA-gCMF displayed similar predictive patterns on these datasets: performing very well on train and test sets with high gene overlap, and almost randomly on sets where little overlap exists (Tables 4a and 6, Fig. 4, 8a and 8b). These results supported our hypothesis that matrix factorisation techniques will be less effective when predicting on previously genes with no *a priori* class label information. We investigated this hypothesis using our gene holdout experiments, which we detail in the following section.

PCA-gCMF fails to predict on gene-pairs with no *a priori* class labels

We conducted gene holdout experiments to investigate the effect of gene selection bias on SL prediction performance. For this experiment, we trained and tested our models and PCA-gCMF using the combined gold standard pairs under three different scenarios. The first scenario involved no explicit gene holdout, as per our original setup, where we only made sure that there was no overlap between gene pairs in the training and test sets. This scenario was our baseline for this experiment. For the second scenario, referred to as single gene holdout or "Single", we constructed training and test sets such that every gene pair in the test set featured only one of its genes in the training set. For the third scenario, referred to as double gene holdout or "Double", we created training and test sets such that all individual genes in the test set were absent from the training set. These models were trained and tested on BRCA data only since all three machine learning techniques performed well on BRCA under the first scenario of no explicit gene holdout and the size of the BRCA datasets allowed us to construct reasonably sized training and test sets under the given constraints (see Section 2.8).

	Trained on ISLE, Tested on DiscoverSL (BRCA)	Trained on DiscoverSL, Tested on ISLE (LUAD)
LR w/ L0 L2	**.95 ± .02**	**.71 ± .02**
LR w/ L1 L2	**.95 ± .02**	**.71 ± .02**
MUVR	**.95 ± .02**	.6 ± .03
PCA-gCMF	.93 ± .07	.56 ± .03
RRF	.92 ± .03	.59 ± .02

Table 6. Average and standard-deviations of AUROC scores for each model as follows: trained based on ISLE gold standard gene pairs and tested on DiscoverSL gene pairs for BRCA; trained based on DiscoverSL gold standard gene pairs and tested on ISLE gene pairs for LUAD.

All models exhibited a baseline average AUROC in the range of 0.84 to 0.92 on BRCA, where PCA-gCMF achieved the highest average performance, followed by our random forest models and then our logistic regression models. Under the second scenario, we observed only a small drop in performance and the ranked performance of the models was similar to the first scenario with the exception of RRF. However, under the third scenario, the average performance of PCA-gCMF dropped to that of a random classifier, while other models still achieved an average AUROC $\approx$ 0.72 (Fig. 8c). Recall our hypothesis that a matrix factorisation methods would predict less effectively than classical machine learning models on genes with no *a priori* class label information. The results from the gene-holdout experiment partially agreed with this argument. The predictive performance of PCA-gCMF remained higher than the other models in the "Single" holdout scenario when the class labels for a single gene from each pair in the test set was held out of the training data. Conversely,

Experiment	Training Set		Test Set	
	Samples	Genes	Samples	Genes
None	1252	799.3	536	395.8
Single	710	221.3	381.6	181.3
Double	707.2	121.5	160.4	65.3

Table 7. Average number of samples and unique genes in the training and test sets used in the gene holdout experiment.

it fell to that of a random classifier for the 'Double" holdout scenario, when it predicted on pairs of genes where neither gene had class label featured in the training data. The poor performance suggests that this matrix factorisation approach goes beyond considering the structure of the SL gold standard pairs to heavily relying on this structure for making SL predictions. Our result could also indicate that the collective matrix factorisation approach in PCA-gCMF is insufficiently leveraging relevant information contained in the other four distinct feature matrices provided as input.

Interestingly, we also noted that between the "Single" and "Double" holdout scenarios, the performance and stability of all our models dropped. This was despite both scenarios possessing a similar number of training samples. However, the number of unique training genes is almost halved between the two cases. A confounding factor here is the reduction in test samples between the two scenarios, but this may indicate that gene diversity may also have a part to play in generalisability of models to new data (Table 7). These results should give researchers some pause for thought to consider if their reported SL predictive performances are overly optimistic, especially with regards to methods which employ matrix factorisation.

3.5 Gene dependency-based features are most important

In our final analyses, we aimed to quantify the contribution of our features to the overall predictive performance for our models. First, we discuss the potential for multicollinearity to confound any resulting analyses. We inspect our dataset for the presence of multicollinearity by calculating variance inflation factors (VIF, refer to Section 2.7). We follow our discussion on multicollinearity with a description of the FI scores for the BRCA and LUAD cancer-specific models. To get meaningful insights on feature importance (FI), the trained models should prove to be reasonably accurate. Attempting to run permutation FI algorithms (see Section 2.7) on inaccurate models could lead to results which vary greatly and can not be treated as significant (Parr *et al.*, 2018). Thus, OV and COAD models are excluded due to their poor performances (Table 4a). Finally, we assess the reliance of our models on gene dependency-based features. We do this by training separate models on datasets which either include or exclude these gene-dependency based features and comparing their predictive performances. Finally, we analyse the feature importances for these gene dependency-free models.

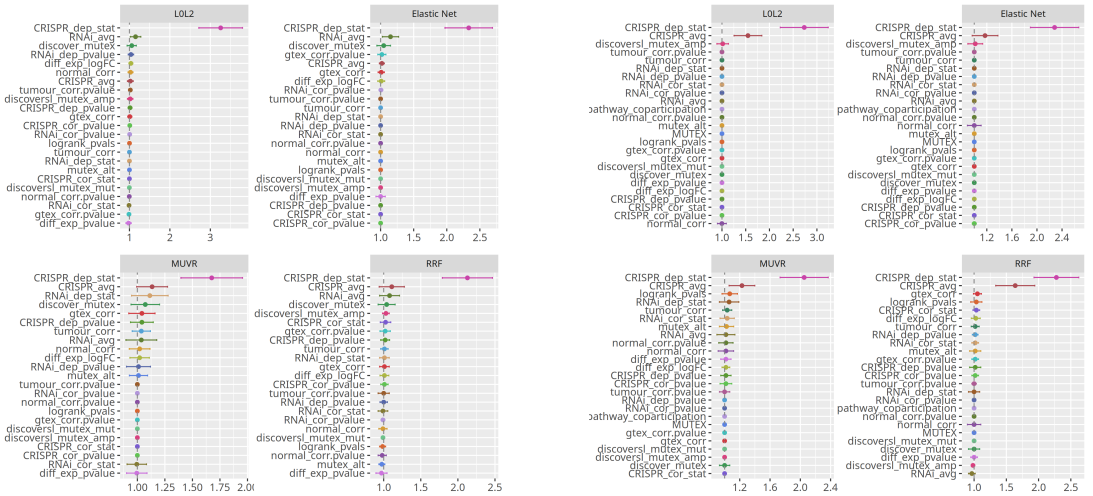
Multicollinearity may reduce the reliability of any FI metrics. For example, random forest models can inflate the FI scores of correlated variables when measured through permutation methods (Molnar, 2020). To quantify its presence, we calculated VIF values for each feature in our combined dataset (Table 8). We found that both the $RNAi\_dep\_stat$ and $CRISPR\_dep\_stat$ features had VIF values of approximately five. These features displayed a high correlation with one another, but the linear models that produced these VIFs also suggested that these high-VIF features were significantly multicollinear with other features. To account for this, we first conducted FI analyses across the full feature set. Afterwards, we removed these features and reassessed the models.

$CRISPR\_dep\_stat$, a feature quantifying the change in cell line dependency on one gene given a mutation in the other, ranks highest in all models for both BRCA and LUAD (Fig. 9, see Supplementary Figures S8 and S9 for higher resolution). $CRISPR\_avg$ ranks second in most models, consistently scoring above one in all of the LUAD models and the BRCA random forest models. $RNAi\_avg$, which is correlated with $CRISPR\_avg$, ranks second in the BRCA logistic regression models. The rankings for the BRCA and LUAD logistic regression models seem to suggest the data are linearly separable along the $CRISPR\_dep\_stat$ and $RNAi\_avg$ or $CRISPR\_avg$ features. The commonality between these sparse BRCA and LUAD logistic regression models in ranking $CRISPR\_dep\_stat$ at the top may partially explain their ability to generalise between these two cancer types (Fig. 5, top). These three top features are all based on gene dependency scores and exhibit the highest VIF values (Table 8). Conversely, metrics based on the correlations between dependency scores are consistently shown to be unimportant across models (Fig. 9). Considering our previous concerns regarding multicollinearity, interpretability, and score inflation, we conducted an additional experiment to assess the reliance of our models on gene dependency-based features.

Feature	VIF	Feature	VIF
discoversl_mutex_amp	1.39	**CRISPR_dep_stat**	**4.95**
discover_mutex	1.04	gtex_corr	1.33
mutex_alt	1.38	gtex_corr.pvalue	1.07
RNAi_avg	**2.22**	tumour_corr	1.17
RNAi_cor_pvalue	1.01	tumour_corr.pvalue	1.05
RNAi_cor_stat	1.01	normal_corr	1.29
RNAi_dep_pvalue	1.11	normal_corr.pvalue	1.43
RNAi_dep_stat	**4.92**	diff_exp_logFC	1.02
CRISPR_avg	**2.25**	diff_exp_pvalue	1.07
CRISPR_cor_pvalue	1.00	pathway_coparticipation	1.01
CRISPR_cor_stat	1.00	logrank_pvals	1.35
CRISPR_dep_pvalue	1.14		

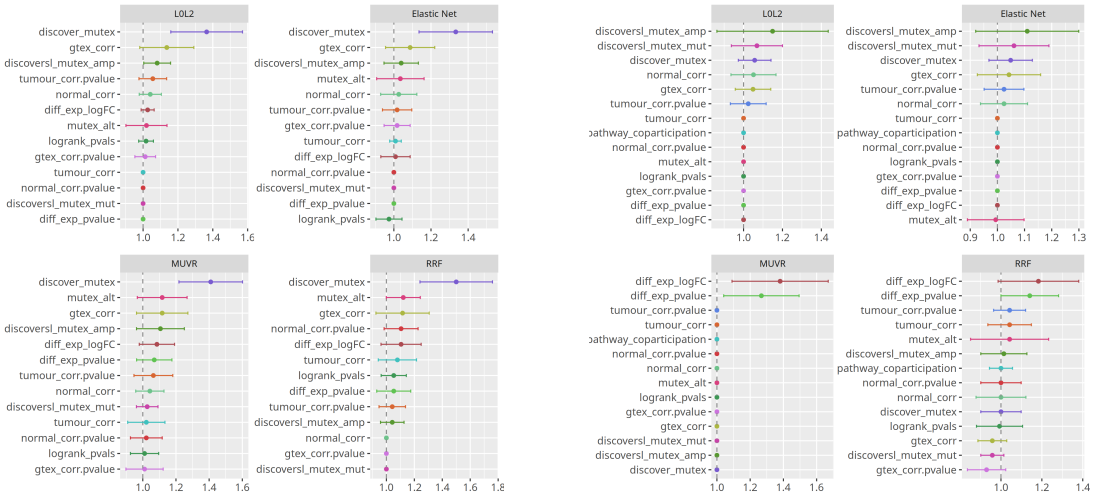
Table 8. Variance inflation factors for each feature of the combined dataset. The features in bold exhibited high multicollinearity.

To evaluate the impact of excluding gene dependency-based features, we trained and tested our models on BRCA or LUAD datasets which either included the entire feature set or excluded the gene dependency-based features. Overall, we find that excluding the gene dependency-based features has a significant negative impact on predictive performance across all models (Fig. 11). However, we also find that all models still perform significantly better than that of a random classifier i.e. AUROC > 0.5. Analysing the FI of these models using the permutation-based approach leads to results with higher variability (Parr *et al.* (2018)), but a few clear patterns do emerge. A measure of mutual exclusivity, $discover\_mutex$, appears to be important across all BRCA models and LUAD logistic regression models (Fig. 10). Our mutual exclusivity score, $mutex\_alt$ seems important to random forest BRCA models (Fig. 10a). Differential expression measures also rank highly across random forest models, especially for LUAD, where they rank at the top and the distributions of their scores are greater than one (Fig. 10). Finally, gene coexpression in healthy tissue samples, $gtex\_corr$, ranked highly for the BRCA models, with almost the entire distributions of their feature scores greater than one (Fig. 10a). Our results demonstrate that some features, other than those based on gene dependency, do contain useful information. They also reinforce our previous findings that gene dependency-based features are the most important, as we see that all gene dependency-free models perform worse than any model which includes these features (Fig. 11).



(a) BRCA (b) LUAD

Fig. 9: Median FI scores for the BRCA (top) and LUAD (bottom) cancer-specific L0L2 and MUVR models trained. These were scored using 100 repetitions of the model agnostic permutation FI algorithm described in Section 2.7. The bars represent the distribution of scores between the lower and upper 5% quantiles of importance values from the repetitions. See Supplementary Figures S8 and S9 for higher resolution images.



(a) BRCA (b) LUAD

Fig. 10: Median FI scores for the BRCA (top) and LUAD (bottom) cancer-specific L0L2 and MUVR models trained without gene dependency-based features. These were scored using 100 repetitions of the model agnostic permutation FI algorithm described in Section 2.7. The bars represent the distribution of scores between the lower and upper 5% quantiles of importance values from the repetitions. See Supplementary Figures S10 and S11 for higher resolution images.

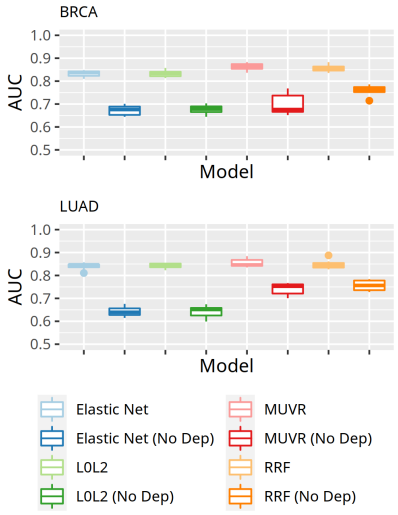


Fig. 11: Distribution of AUROCs of the logistic regression and random forest models trained on data sets with or without features based on gene dependency scores over 10 runs (top: BRCA, bottom: LUAD). The y-axis denotes the AUROC value. The models without gene dependency score features are labelled "No Dep".

4 Conclusion

Synthetic lethality is a promising concept in the field of personalised medicine. SL provides researchers with a mechanism by which they can selectively target tumour cells. However, *in vitro* identification of SL interactions is laborious and expensive. Consequently, this impedes the development of novel SL anti-cancer therapies. *In silico* prediction of SL interactions could help to substantially improve the efficiency of these efforts by focusing experimentation on the most promising gene pairs. Machine learning provides an avenue by which we can tackle this *in silico* SL prediction problem.

The aims of this research were twofold: (i) to introduce novel data sources and features into a machine learning context and propose machine learning models for *in silico* SL prediction in cancer-specific and pan-cancer contexts; (ii) to examine the impact of cancer representation and gene selection bias on SL model generalisation. Firstly, our research illustrates the clear potential for including gene dependency- and mutation-based features in SL prediction models for breast and lung cancer. Our models were especially effective at ranking positive SL interactions highly in their predictions compared to our selected baselines. Our work also suggests that other proposed features, like those based on molecular data from healthy tissue, contain useful information for SL prediction. Secondly, based on our quantitative analysis of cross-cancer model generalisation we can conclude that it is possible for models to generalise cross-cancer but this ability to generalise cannot be assumed. This is contrary to the pan-cancer approach proposed by Das *et al.* (2019). From our analysis of biases in SL gold standard datasets, we can conclude that: cancer representation bias does impact the predictive performance of pan-cancer models, and can be mitigated through cancer representation balancing; gene-selection bias in SL gold standards can lead to an optimistic estimation of model performance. We can also conclude that simpler logistic regression models are more sensitive to cancer representation bias than their random forest counterparts, but are less affected by gene selection biases and can generalise better across gold standard datasets.

Compared with previous works on SL prediction, the evaluation of our SL machine learning models offers several key differences. Uniquely among studies in the field, we consider quantifying the predictive performance at the top of our ranked results instead of summarising over the whole results list using the "Average Precision@k" metric. This intentionally mimics the intended use of these models in practical applications as a precursor to *in vitro* experimentation. We argue that this allows for a more realistic comparison between approaches in a single study. However, as this value is dependent on the cut-off parameter k and the proportion of positive classes in the test set, it is difficult to use to compare across studies. We also put a much greater emphasis on the generalisability of our methods to new data, lending greater fidelity to our results and avoiding the pitfalls of over-estimating our models' predictive capabilities.

Limiting the generalisability of our results is the fact that we simply demonstrate one example of both a linear and non-linear method when conducting our experiments. Our choices, therefore, do not comprehensively cover the families of linear and non-linear methods to draw any broader conclusions about either family of classifiers. Similarly, it is difficult to conclude that all methods based on matrix factorisation techniques will suffer the same weakness with respect to gene-selection bias as demonstrated by the matrix factorisation technique PCA-gCMF that we used in our experiments. Therefore, expanding on the number of approaches under examination per classification technique may permit us to draw wider conclusions. Also, considering our pan- and cross-cancer experiments, the fact that our cancer-specific models performed poorly on colon and ovarian cancer data may also explain the drop in performance when we train our pan-cancer models. It may be better to revisit these pan-cancer experiments if further studies suggest differing SL predictors correlate to good performances in different cancers. This could allow us to more accurately gauge a pan-cancer model's ability to learn cancer-specific SL rules. Despite these limitations, we provide clear evidence for the effectiveness of our breast and lung cancer models and the potential impacts of cancer representation and gene selection bias on classifier performance and evaluation. Based on our conclusions, we advocate that practitioners should take care to consider cancer representation and gene selection bias during the curation and construction of their train and test sets as to avoid over-estimation of predictive performance.

Considering the broad scope of our work, there are numerous paths left unexplored and areas that deserve deeper analysis. Future work could focus on the gene dependency-based features which demonstrated high importance for breast and lung cancer-specific models but were shown to be ineffective for colon and ovarian. Further research is needed to investigate the sensitivity of these features to the individual characteristics of particular cancers, such as the prevalence of microsatellite instabilities in colon cancer. Another interesting avenue could be to explore the potential use of the raw gene dependency values as predictors in cancer-specific models. Furthermore, our demonstration of the impact of cancer representation bias could be valuable for future research into pan-cancer SL predictive modelling. The successful development of such pan-cancer models could allow for SL predictive capabilities to be extended to rarer cancer types with insufficient available data for model training and testing. Continuing our theme of SL model generalisation, a meta-analysis of the current literature might indicate if the problem of cancer representation and gene selection bias is widespread among SL predictive research. Finally, we also identified gene diversity (the number of unique genes) within SL gold standard datasets as a possible issue which could affect model generalisation and deserves further examination.

References

Babur, Ö. *et al.* (2015). Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biology*, **16**(1), 45.

Bangdiwala, S. I. (1989). The wald statistic in proportional hazards hypothesis testing. *Biometrical Journal*, **31**(2), 203–211.

Becker, J.-M. *et al.* (2015). How collinearity affects mixture regression results. *Marketing Letters*, **26**(4), 643–659.

Behan, F. M. *et al.* (2019). Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature*, **568**(7753), 511–516.

Bewick, V. *et al.* (2004). Statistics review 12: survival analysis. *Critical care (London, England)*, **8**(5), 389–394. 15469602[pmid].

Bonneville, R. *et al.* (2017). Landscape of microsatellite instability across 39 cancer types. *JCO precision oncology*, **1**, 1–15.

Boone, C. *et al.* (2007). Exploring genetic interactions and networks with yeast. *Nature Reviews Genetics*, **8**(6), 437–449.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**(1), 5–32.

Canisius, S. *et al.* (2016). A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome Biology*.

Das, S. *et al.* (2019). DiscoverSL: an R package for multi-omic data driven prediction of synthetic lethality in cancers. *Bioinformatics*, **35**(4), 701–702.

Dempster, J. M. *et al.* (2019). Extracting biological insights from the project achilles genome-scale crispr screens in cancer cell lines. *bioRxiv*.

Deng, H. and Runger, G. (2012). Feature selection via regularized trees. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Draper, N. R. and Smith, H. (1998). *Applied regression analysis*, volume 326. John Wiley & Sons.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, **27**(8), 861 – 874. ROC Analysis in Pattern Recognition.

Fisher, A. *et al.* (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, **20**(177), 1–81.

Fong, P. C. *et al.* (2009). Inhibition of poly(adp-ribose) polymerase in tumors from brca mutation carriers. *New England Journal of Medicine*, **361**(2), 123–134. PMID: 19553641.

Friedman, J. *et al.* (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**(1), 1–22.

Ghandi, M. *et al.* (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*, **569**(7757), 503–508.

Hazimeh, H. and Mazumder, R. (2018). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms.

Huang, A. *et al.* (2019). Synthetic lethality as an engine for cancer drug target discovery. *Nature Reviews Drug Discovery*, pages 1–16.

Jerby-Arnon, L. *et al.* (2014). Predicting Cancer-Specific Vulnerability via Data-Driven Detection of Synthetic Lethality. *Cell*, **158**(5), 1199–1209.

Klami, A. *et al.* (2013). Group-sparse embeddings in collective matrix factorization. *arXiv preprint arXiv:1312.5921*.

Krstajic, D. *et al.* (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, **6**(1), 10.

Kursa, M. B. *et al.* (2010). Boruta–a system for feature selection. *Fundamenta Informaticae*, **101**(4), 271–285.

Lee, J. S. *et al.* (2018). Harnessing synthetic lethality to predict the response to cancer treatment. *Nature Communications*, **9**(1), 2546.

Liany, H. *et al.* (2019). Predicting Synthetic Lethal Interactions using Heterogeneous Data Sources. *bioRxiv*, page 660092.

Liberzon, A. *et al.* (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**(12), 1739–1740.

Liu, Y. *et al.* (2019). Sl 2 mf: Predicting synthetic lethality in human cancers via logistic matrix factorization. *IEEE/ACM transactions on computational biology and bioinformatics*.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, **18**(1), 50–60.

McDonald, E. R. *et al.* (2017). Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. *Cell*, **170**(3), 577–592.e10.

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

Mermel, C. H. *et al.* (2011). Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology*, **12**(4), R41.

Meyers, R. M. *et al.* (2017). Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nature Genetics*, **49**(12), 1779–1784.

Molnar, C. (2020). Limitations of interpretable machine learning methods. Accessed: 2020-02-08.

Mosteller, F. and Fisher, R. A. (1948). Questions and answers. *The American Statistician*, **2**(5), 30–31.

Nijman, S. M. (2011). Synthetic lethality: general principles, utility and detection using genetic screens in human cells. *FEBS letters*, **585**(1), 1–6.

O'Neil, N. J. *et al.* (2017). Synthetic lethality and cancer. *Nature Reviews Genetics*, **18**(10), 613–623.

Parr, T. *et al.* (2018). Beware default random forest importances. *March*, **26**, 2018.

Rahman, M. *et al.* (2015). Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. *Bioinformatics*, **31**(22), 3666–3672.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, **11**(3), R25.

Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**(21), 2881–2887.

Robinson, M. D. and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, **9**(2), 321–332.

Robinson, M. D. *et al.* (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.

Schuster, A. *et al.* (2019). Rnai/crispr screens: from a pool to a valid hit. *Trends in biotechnology*, **37**(1), 38–55.

Shen, J. P. *et al.* (2017). Combinatorial crispr–cas9 screens for de novo mapping of genetic interactions. *Nature methods*, **14**(6), 573.

Shi, L. *et al.* (2019). Variable selection and validation in multivariate modelling. *Bioinformatics*, **35**(6), 972–980.

Stoeger, T. *et al.* (2018). Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS biology*, **16**(9).

Subramanian, A. *et al.* (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**(43), 15545–15550.

Therneau, T. M. (2020). *A Package for Survival Analysis in R*. R package version 3.1-12.

Wan, F. *et al.* (2020). Exp2sl: A machine learning framework for cell-line-specific synthetic lethality prediction. *Frontiers in Pharmacology*,

11, 112.

Whalen, S. *et al.* (2016). Predicting protein function and other biomedical characteristics with heterogeneous ensembles. *Methods*, **93**, 92–102.

Machine Learning of Synthetic Lethality: Data Integration, Generalisation, and Selection Bias - Supplementary Material

Contents

1 Supplementary Figures

1.1 Counts of gene occurrence per cancer type in combined gold standard

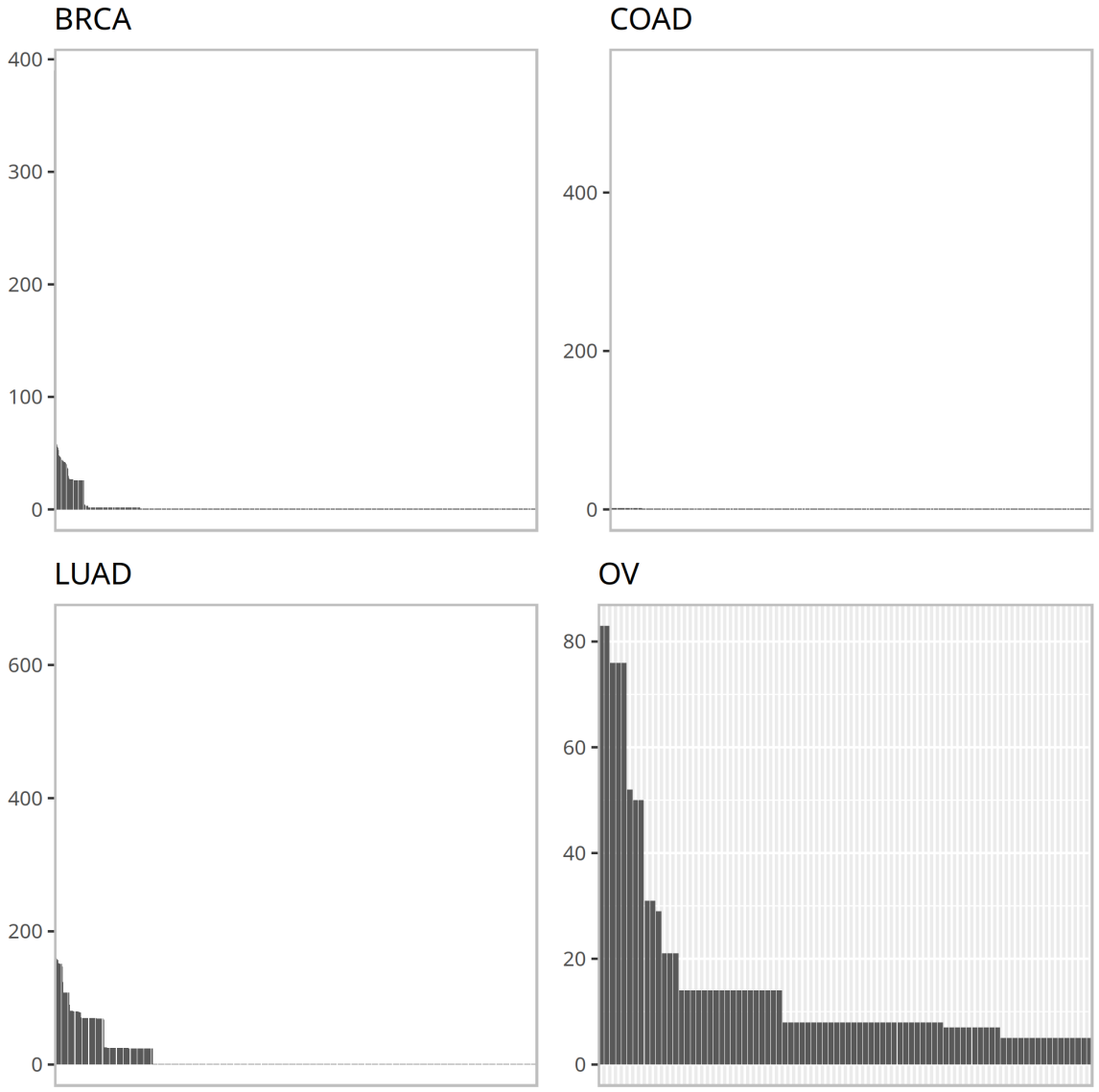


Figure S1: Counts of the number of occurrences of each gene in a labelled synthetic lethal pair in our combined gold standard dataset, separated per cancer type. BRCA has 1072 unique genes, LUAD has 804, COAD has 1560 and OV has 83. Due to resolution, COAD is not clear, but only 5 genes are featured more than 3 times, and only 105 genes out of 1560 genes are featured more than 1 time. For our OV gold standard labels, we can see that every gene is featured in multiple labelled gene pairs. Note the different scales for the y-axis.

1.2 Number of mutations per gene-pair per cancer type in combined gold standard

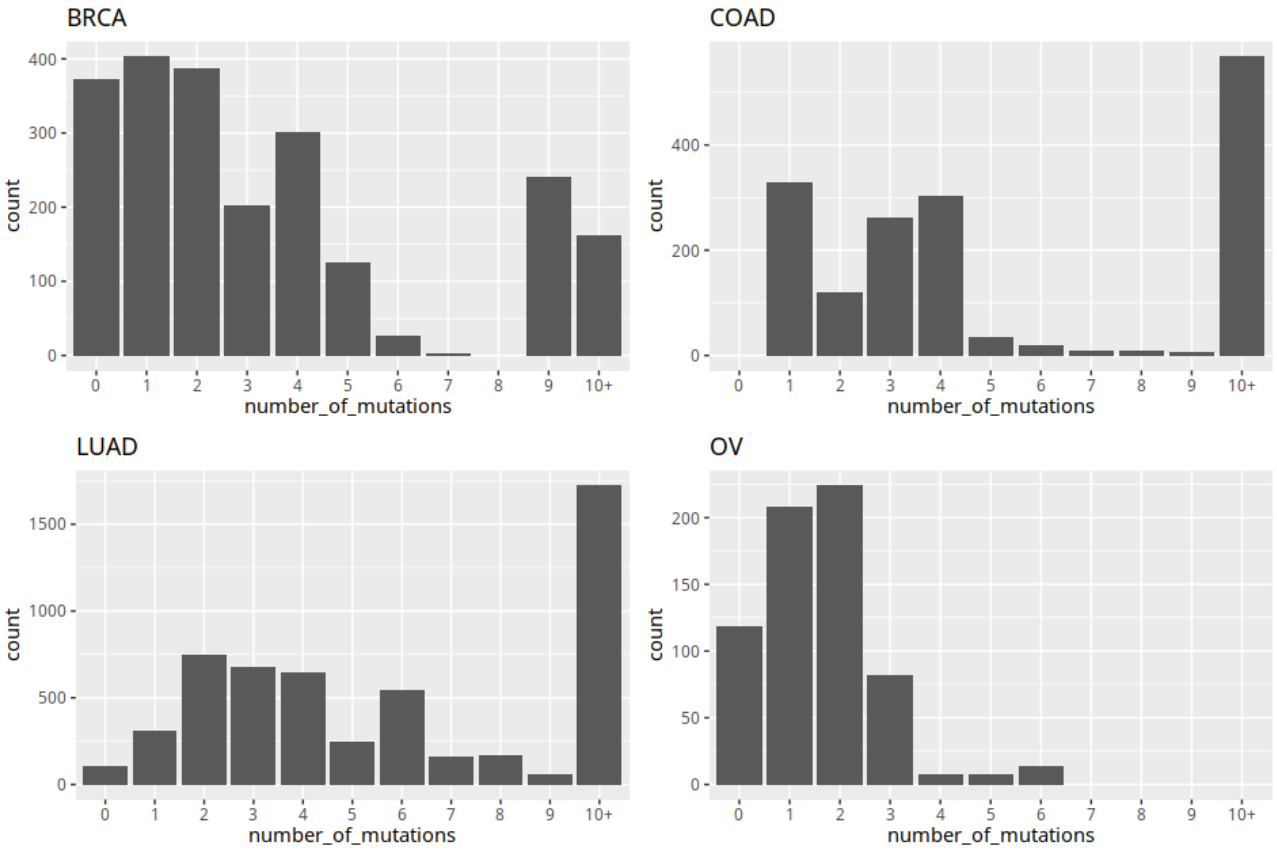
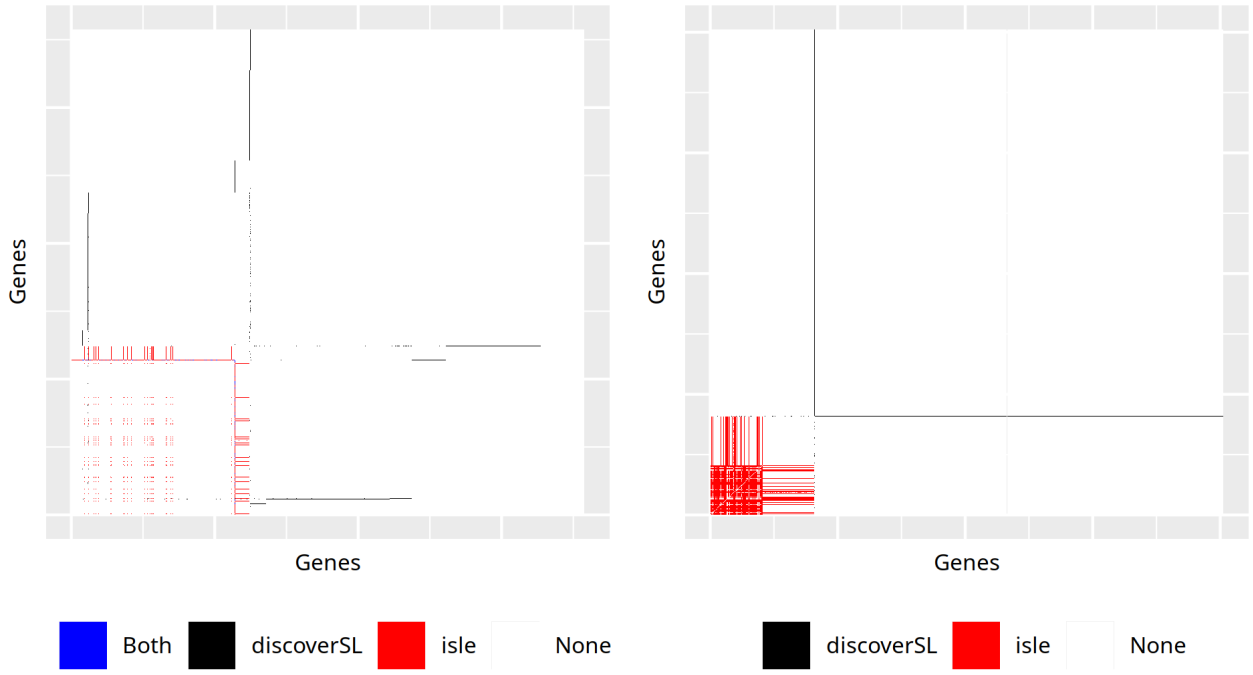


Figure S2: Counts of the maximum number of mutations between two genes of any gene pair in the CCLE cell-lines in our labelled datasets. The x-axis is the maximum number of mutations present between either gene in a pair, ranging from 0 to 10+. The y-axis is the counts for the number of gene pairs. Note the different scales for the y-axis.

1.3 Adjacency matrix dot plots stratified per dataset



(a) BRCA

(b) LUAD

Figure S3: Dot plot showing labelled gene pairs from BRCA(a) and LUAD(b) data from both the ISLE and DiscoverSL dataset. The x-axis and y-axis are the unique genes. Each dot represents a gene pair where a label exists in either the ISLE (red), DiscoverSL (black), or both (blue) datasets. Whitespace represents gene pairs which are unknown to either dataset.

1.4 Adjacency matrix dot plots stratified per cancer type

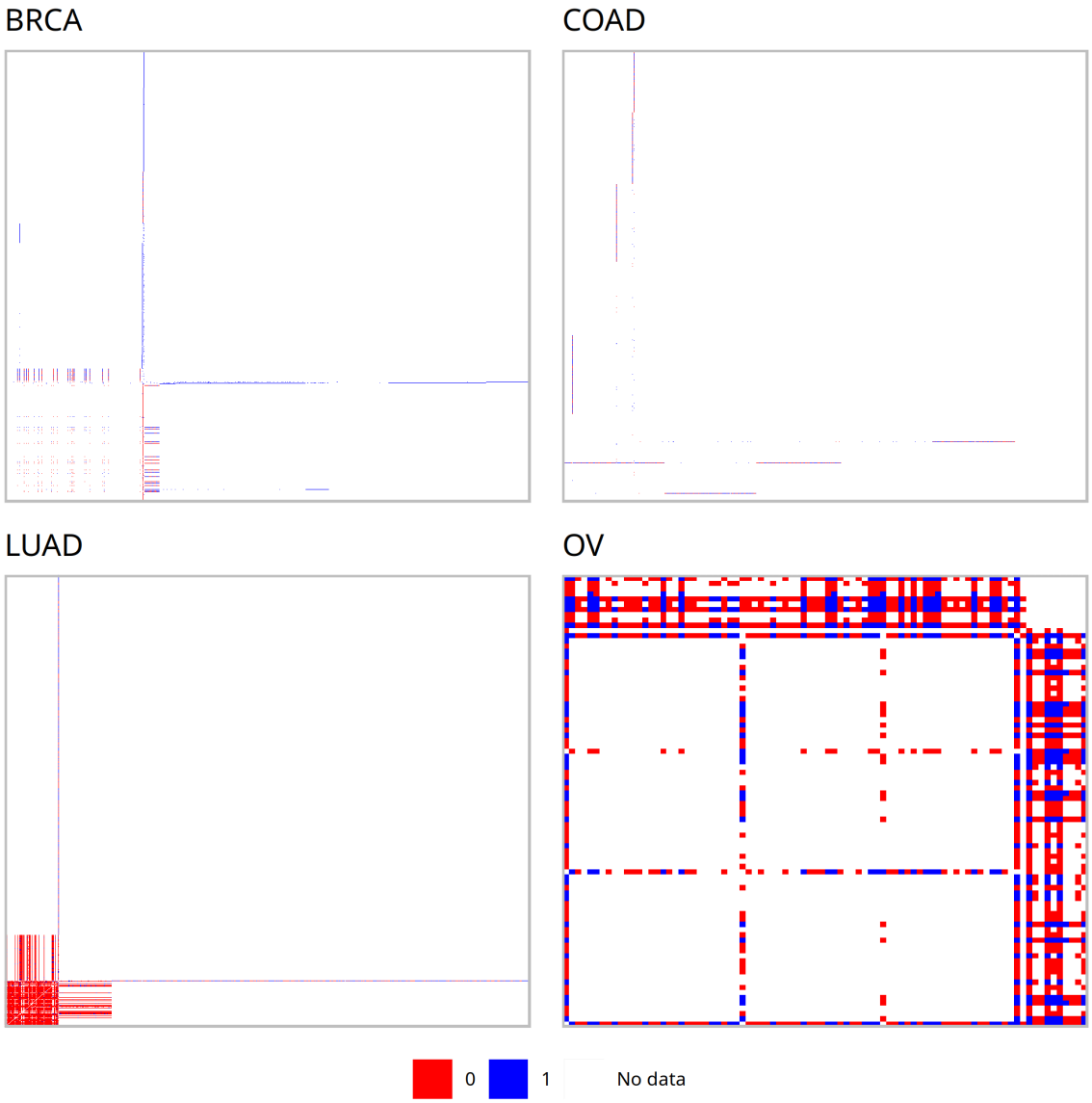


Figure S4: Dot plot showing labelled gene pairs from BRCA, COAD, LUAD, and OV cancer types in the combined dataset. The x-axis and y-axes are the unique genes per cancer type gene pairs. Each dot represents a gene pair where a positive (blue) or negative (red) label exists. Whitespace represents gene pairs with no labels.

1.5 Cross-cancer heatmap of predictive performance

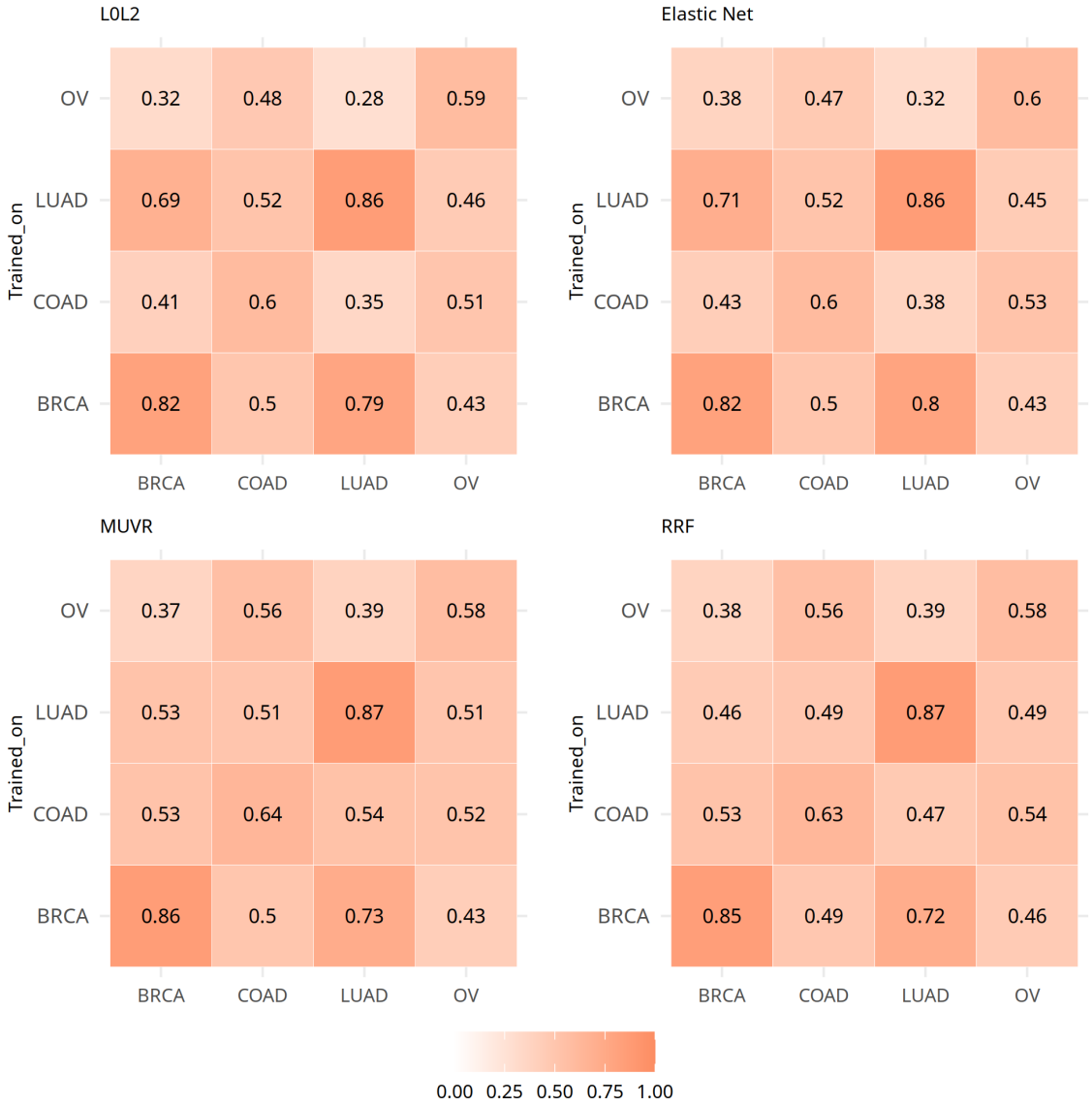


Figure S5: Heatmaps of average AUROC performances for the L0L2 (top-left), Elastic Net (top-right), MUVR (bottom-left), and RRF models (bottom-right) over 10 runs. On the y-axis is the cancer type that each model was trained on. On the x-axis is the cancer that each model was tested against.

1.6 Leave-one-cancer-out heatmap of predictive performance

Figure S6: Heatmaps of average AUROC performances for L0L2 (top-left), Elastic Net (top-right), MUVR (bottom-left), and RRF models (bottom-right) over 10 runs. On the x-axis is each cancer type that was held out for testing. The models were trained on a balanced dataset of all other cancers.

1.7 Pearson's correlation between features in combined dataset

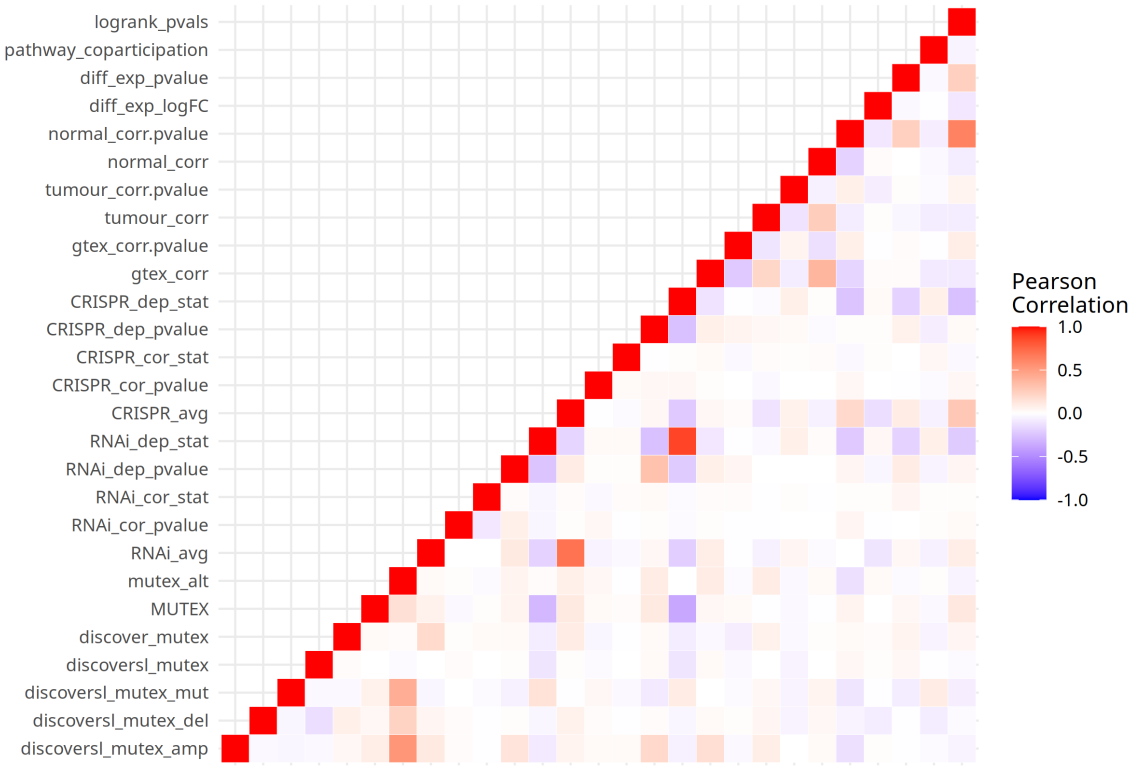


Figure S7: Heatmap of Pearson's correlation coefficient between features from the combined dataset.

1.8 Median feature importance scores for the BRCA one-cancer models

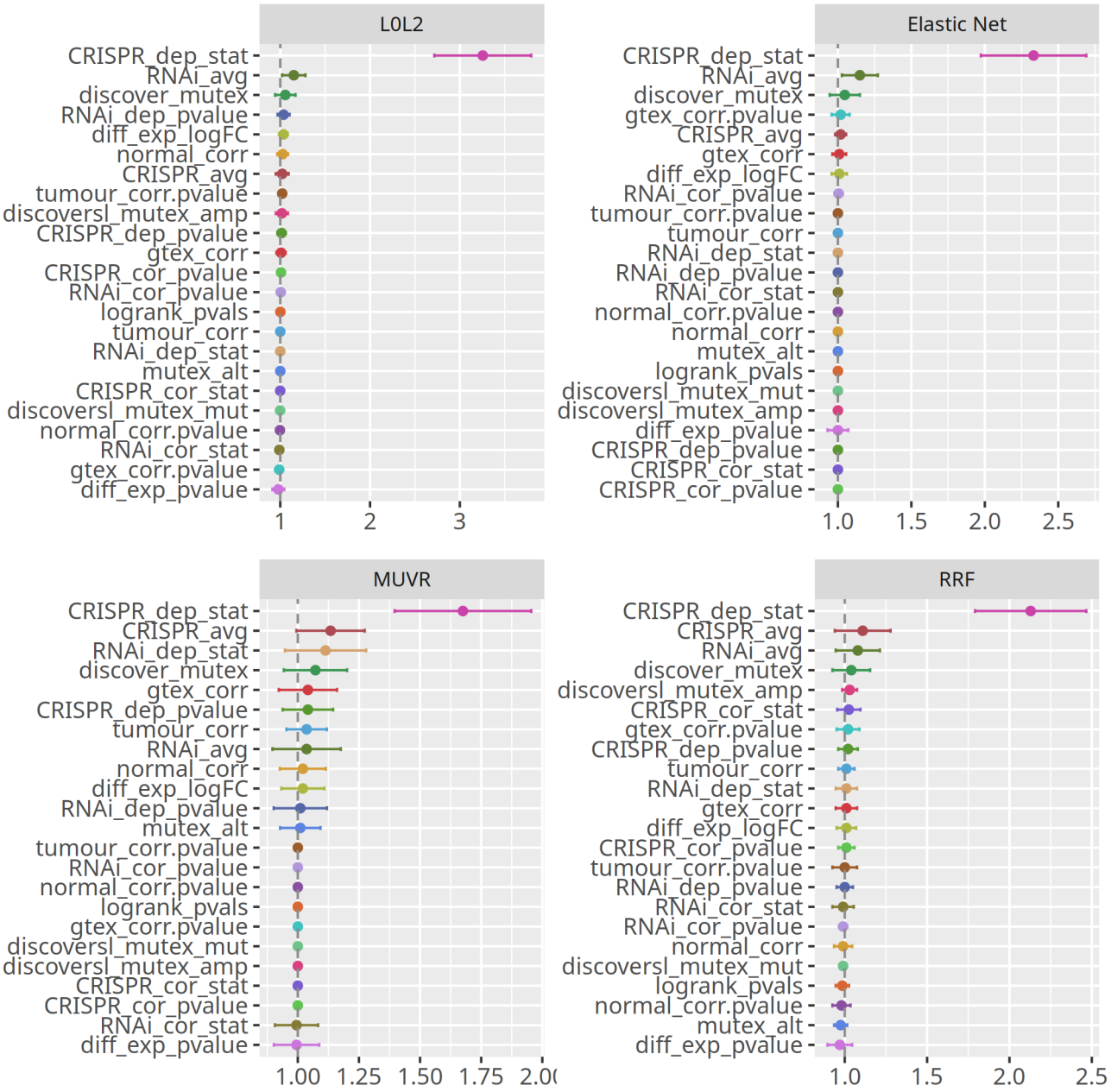


Figure S8: Median feature importance scores for the BRCA one-cancer models.These were scored using 100 repetitions of the model agnostic permutation feature importance algorithm. The bars represent the distribution of scores between the lower and upper 5% quantiles of importance values from the repetitions.

1.9 Median feature importance scores for the LUAD one-cancer models

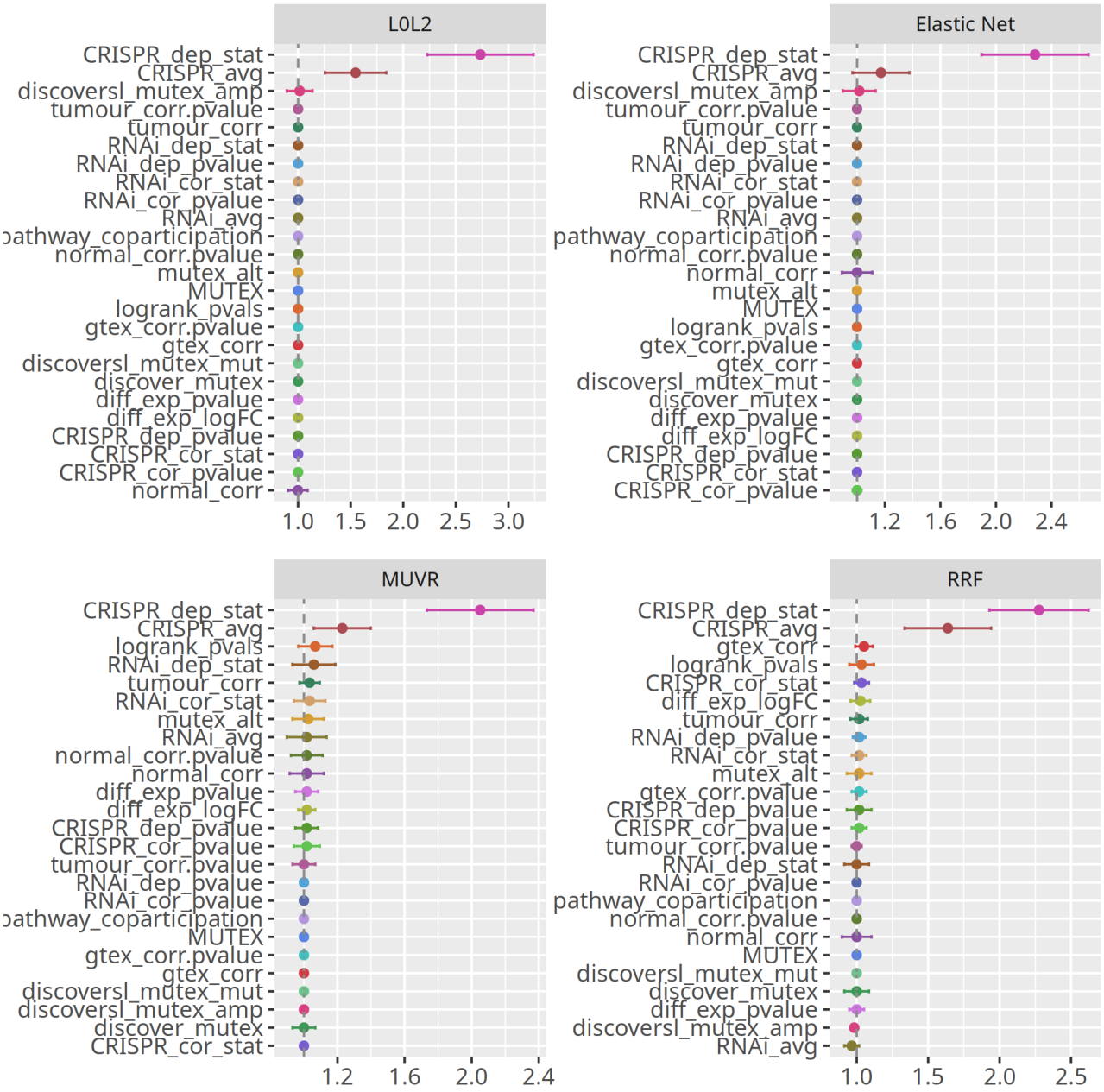


Figure S9: Median feature importance scores for the LUAD one-cancer models.These were scored using 100 repetitions of the model agnostic permutation feature importance algorithm. The bars represent the distribution of scores between the lower and upper 5% quantiles of importance values from the repetitions.

1.10 Median feature importance scores for the BRCA models trained without gene dependency-based features

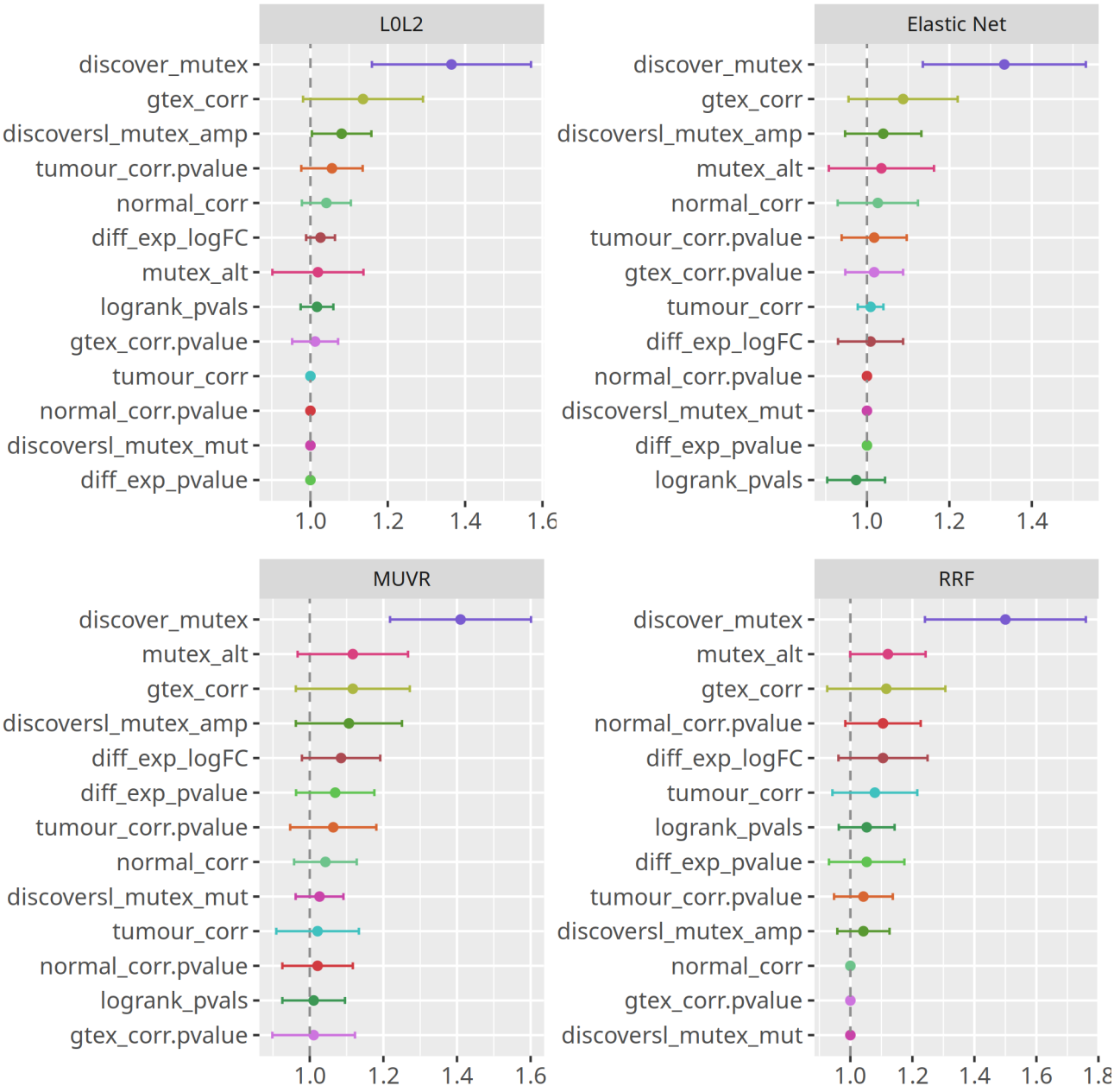


Figure S10: Median feature importance scores for the BRCA models trained without gene dependency-based features. These were scored using 100 repetitions of the model agnostic permutation feature importance algorithm. The bars represent the distribution of scores between the lower and upper 5% quantiles of importance values from the repetitions..

1.11 Median feature importance scores for the LUAD models trained without gene dependency-based features

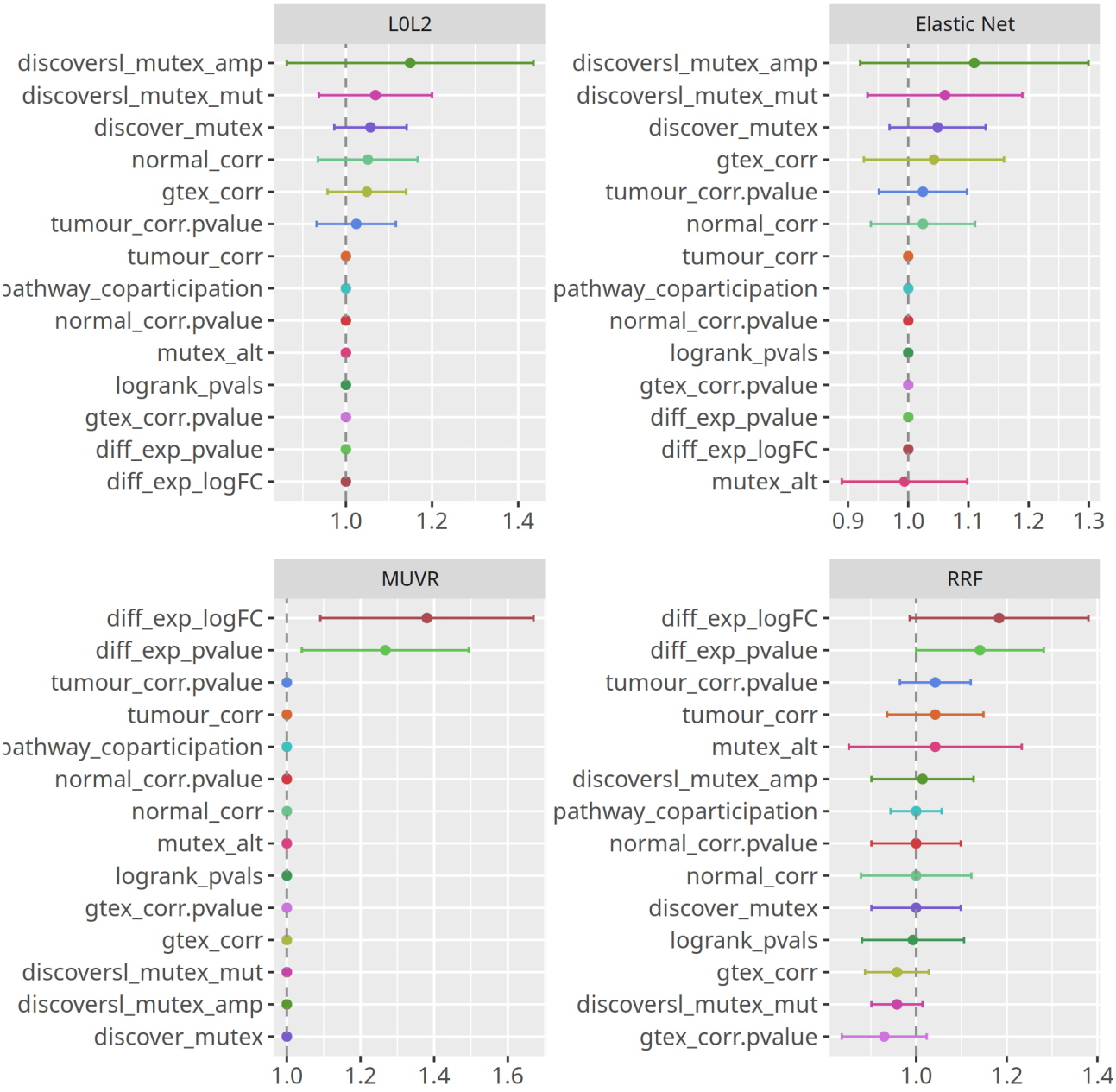
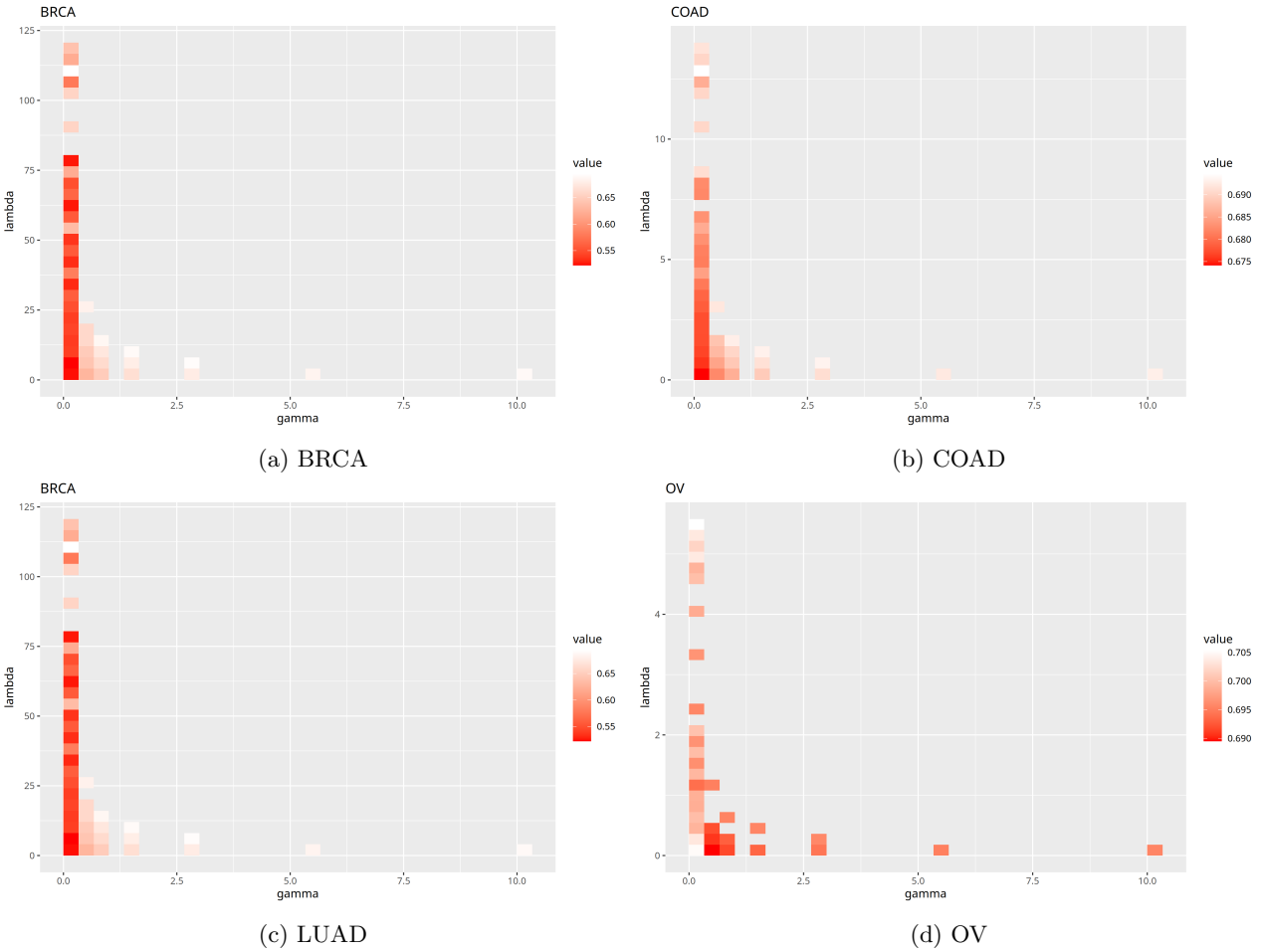


Figure S11: Median feature importance scores for the LUAD models trained without gene dependency-based features. These were scored using 100 repetitions of the model agnostic permutation feature importance algorithm. The bars represent the distribution of scores between the lower and upper 5% quantiles of importance values from the repetitions..

1.12 L0L2 regularisation per cancer type



(a) BRCA

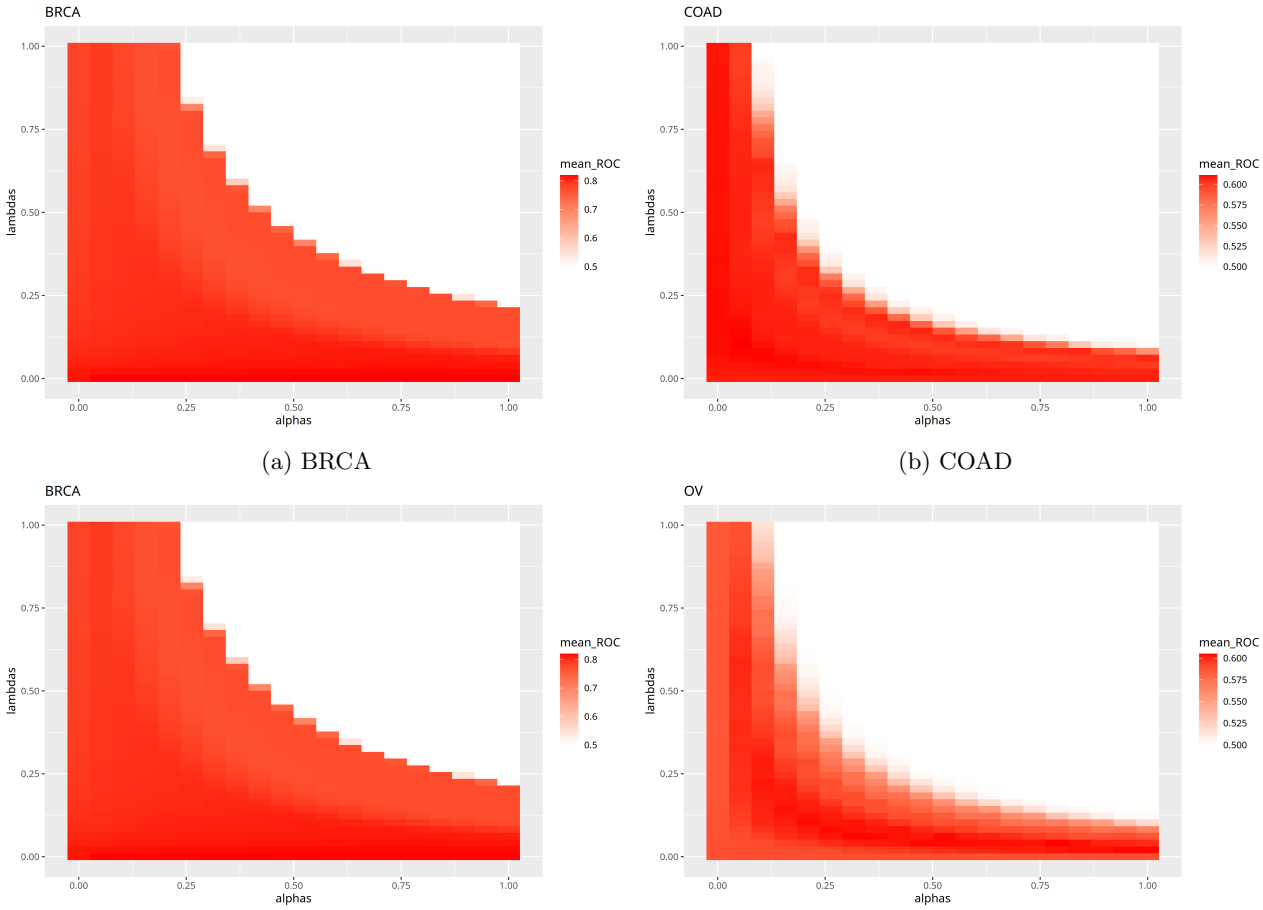
(b) COAD

(c) LUAD

(d) OV

Figure S12: L0L2 Cross Validation and Regularisation. Mean logistic loss values of the optimised local search cross validation results across each of the 10 folds across all 10 cross-validation runs for each cancer type. See Section 2.5. L0Learn implements *gamma* (x-axis) and *lambda* (y-axis) parameters control the regularisation. Deeper red values indicate lower mean logistic loss for that combination of *gamma* and *lambda*. Please refer to Section 2.4 for a theoretical description of the Elastic Net model. With respect to the Eq. 6, $\lambda_0 = lambda$, $\lambda_2 = gamma$. L0Learn uses a local search algorithm to find the optimal values for *lambda*, the use specifies a grid of *gamma* values.

1.13 Elastic Net regularisation per cancer type



(a) BRCA

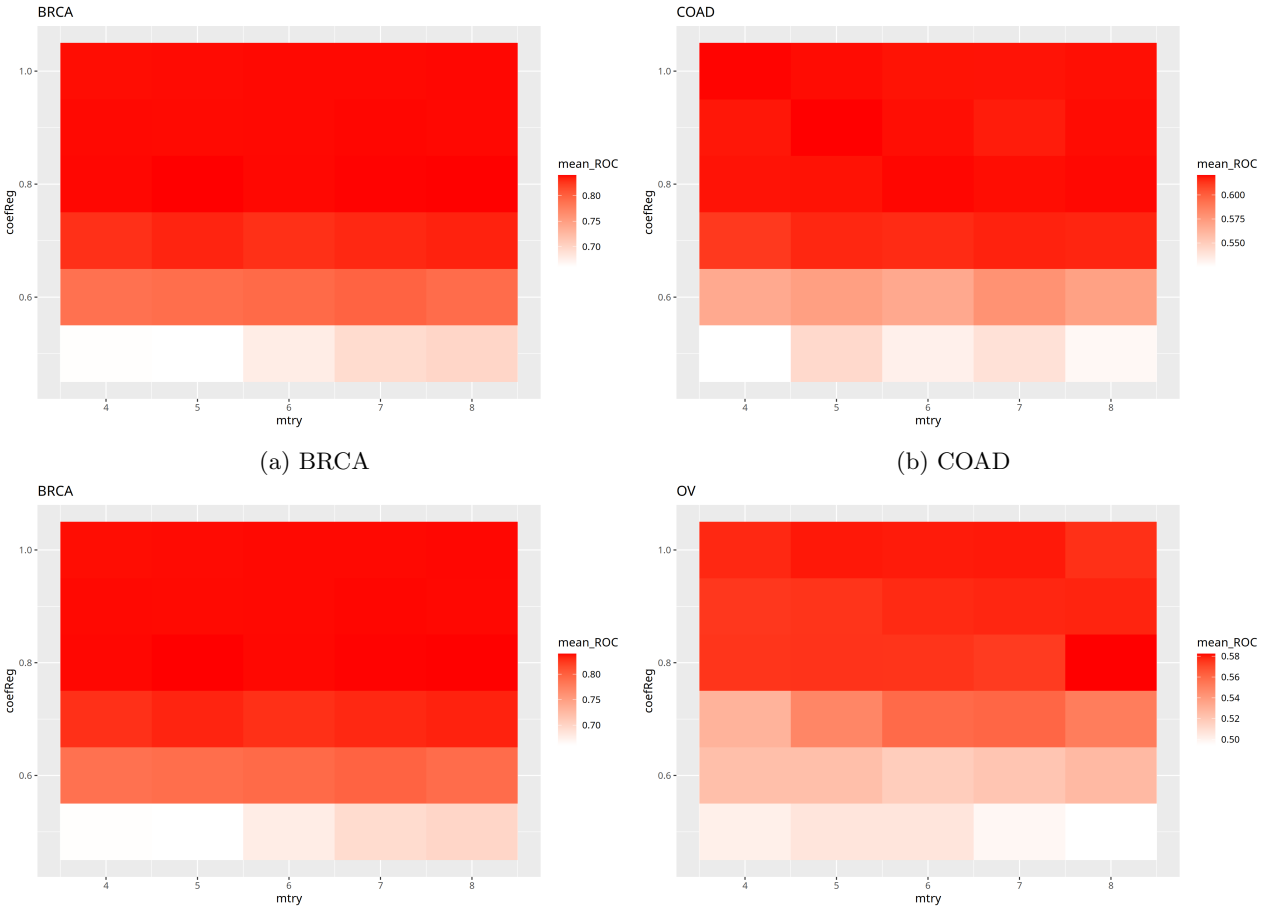
(b) COAD

(c) LUAD

(d) OV

Figure S13: Elastic Net Cross Validation and Regularisation. Mean AUROC grid search cross validation results across each of the 10 folds across all 10 cross-validation runs for each cancer type. See Section 2.5 in the main text. Glmnet implements *alpha* (x-axis) and *lambda* (y-axis) parameters control the regularisation. Deeper red values indicate higher mean AUROC for that combination of *alpha* and *lambda*. Please refer to Section 2.4 for a theoretical description of the Elastic Net model. With respect to the Eq. 7, $\lambda_1 = lambda \times alpha$, $\lambda_2 = lambda \times (1 - alpha)$. The *alpha* parameters controls the smooth switching from lasso regression (alpha=1) to ridge regression (alpha=0). The *lamdba* parameter controls the strength of the regularisation.

1.14 Regularised Random Forest regularisation per cancer type



(a) BRCA

(b) COAD

(c) LUAD

(d) OV

Figure S14: RRF Cross Validation and Regularisation. Mean AUROC grid search cross validation results across each of the 10 folds across all 10 cross-validation runs for each cancer type. See Section 2.5 in the main text. Caret implements mtry (x-axis) and $coefReg$ (y-axis) parameters control the regularisation. Deeper red values indicate higher mean AUROC for that combination of $coefReg$ and $mtry$. Please refer to Section 2.4 for a theoretical description of RRF.

2 Supplementary Tables

2.1 Cross-cancer experiment results

	BRCA	COAD	LUAD	OV
BRCA	$.82 \pm 0.01$	$.41 \pm 0.04$	$.69 \pm 0.04$	$.32 \pm 0.13$
COAD	$.5 \pm 0.01$	$.6 \pm 0.03$	$.52 \pm 0.01$	$.48 \pm 0.01$
LUAD	$.79 \pm 0.01$	$.35 \pm 0.07$	$.86 \pm 0.02$	$.28 \pm 0.12$
OV	$.43 \pm 0.01$	$.51 \pm 0.02$	$.46 \pm 0.03$	$.59 \pm 0.06$

(a) L0L2

	BRCA	COAD	LUAD	OV
BRCA	$.82 \pm 0.01$	$.43 \pm 0.02$	$.71 \pm 0.03$	$.38 \pm 0.11$
COAD	$.5 \pm 0.01$	$.6 \pm 0.02$	$.52 \pm 0.01$	$.47 \pm 0.02$
LUAD	$.8 \pm 0.01$	$.38 \pm 0.04$	$.86 \pm 0.02$	$.32 \pm 0.11$
OV	$.43 \pm 0.02$	$.53 \pm 0.01$	$.45 \pm 0.03$	$.6 \pm 0.07$

(b) Elastic Net

	BRCA	COAD	LUAD	OV
BRCA	$.86 \pm 0.01$	$.53 \pm 0.02$	$.53 \pm 0.02$	$.37 \pm 0.04$
COAD	$.5 \pm 0.01$	$.64 \pm 0.01$	$.51 \pm 0.01$	$.56 \pm 0.02$
LUAD	$.73 \pm 0.02$	$.54 \pm 0.02$	$.87 \pm 0.02$	$.39 \pm 0.05$
OV	$.43 \pm 0.02$	$.52 \pm 0.02$	$.51 \pm 0.02$	$.58 \pm 0.06$

(c) MUVR

	BRCA	COAD	LUAD	OV
BRCA	$.85 \pm 0.01$	$.53 \pm 0.02$	$.46 \pm 0.03$	$.38 \pm 0.05$
COAD	$.49 \pm 0.01$	$.63 \pm 0.02$	$.49 \pm 0.02$	$.56 \pm 0.02$
LUAD	$.72 \pm 0.02$	$.47 \pm 0.04$	$.87 \pm 0.02$	$.39 \pm 0.05$
OV	$.46 \pm 0.02$	$.54 \pm 0.02$	$.49 \pm 0.03$	$.58 \pm 0.08$

(d) RRF

Table S1: AUROC estimates for the single-cancer models trained against a single cancer type (left) and tested against another (top). Mean and standard deviation of 10 repetitions.

2.2 Leave-one-cancer-out experiment results

	AUROC
BRCA	0.69 ± 0.03
COAD	0.5 ± 0.02
LUAD	0.79 ± 0.01
OV	0.48 ± 0.01

(a) L0L2

	AUROC
BRCA	0.67 ± 0.02
COAD	0.5 ± 0.02
LUAD	0.78 ± 0.01
OV	0.45 ± 0.02

(b) Elastic Net

	AUROC
BRCA	0.52 ± 0.04
COAD	0.53 ± 0.01
LUAD	0.72 ± 0.02
OV	0.49 ± 0.02

(c) MUVR

	AUROC
BRCA	0.49 ± 0.04
COAD	0.53 ± 0.01
LUAD	0.72 ± 0.02
OV	0.51 ± 0.02

(d) RRF

Table S2: AUROC estimates for all models. Mean and standard deviation of 10 repetitions. In the left column is each cancer type that was held out for testing. The models were trained on a balanced dataset of all other cancers.

2.3 Gene holdout experiment results

Model	None	Single	Double
Elastic Net	0.84 ± 0.01	0.81 ± 0.02	0.71 ± 0.1
L0L2	0.84 ± 0.01	0.81 ± 0.02	0.74 ± 0.08
MUVR	0.86 ± 0.01	0.84 ± 0.02	0.72 ± 0.11
pca-gCMF	0.92 ± 0.01	0.87 ± 0.01	0.5 ± 0.04
Random Forest	0.86 ± 0.01	0.83 ± 0.02	0.72 ± 0.11

Table S3: AUROC estimates for each model trained under 3 conditions. Mean and standard deviation of 10 repetitions. Columns denote one of three scenarios: for "None", we only guaranteed that the training and test sets of gene pairs were disjoint; for "Single", only one of the genes out of every gene pair in the test set was present in the training set; and for "Double" neither gene of a gene pair in the test set appeared in the training set.

2.4 No gene dependency feature experiment results

Model	AUC
Elastic Net (No Dependencies)	0.67 ± 0.02
Elastic Net	0.83 ± 0.01
L0L2 (No Dependencies)	0.68 ± 0.02
L0L2	0.83 ± 0.01
Random Forest (No Dependencies)	0.76 ± 0.02
Random Forest	0.86 ± 0.01
MUVR (No Dependencies)	0.7 ± 0.05
MUVR	0.86 ± 0.02

(a) BRCA

Model	AUC
Elastic Net (No Dependencies)	0.64 ± 0.02
Elastic Net	0.84 ± 0.01
L0L2 (No Dependencies)	0.64 ± 0.03
L0L2	0.84 ± 0.01
Random Forest (No Dependencies)	0.76 ± 0.02
Random Forest	0.85 ± 0.02
MUVR (No Dependencies)	0.74 ± 0.02
MUVR	0.85 ± 0.02

(b) LUAD

Table S4: AUROC estimates for the one-cancer models that both do and do not include dependency score based features, trained and tested on BRCA and LUAD separately. Mean and standard deviation of 10 repetitions.