Learning Signature Exposures from Gene Expression at Single-Cell Resolution

Regular vs. Multitask Learning of Individual Regression Models

Ariel Potolski Eilat[1]

Supervisor(s): Joana Gonçalves[1], **Sara Costa**[1], **Ivan Stresec**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Ariel Potolski Eilat
Final project course: CSE3000 Research Project
Thesis committee: Joana Gonçalves, Sara Costa, Ivan Stresec, Catharine Oertel Genannt Bierbach

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

Abstract

Understanding the mutational processes active within cancer cells is essential to improve diagnosis and treatment strategies. This study investigates whether the activity levels of these processes, quantified as mutational signature exposures, can be predicted from single-cell gene expression data. Two regression-based learning paradigms are compared: regular independent modelling, where the different models of each mutational signature selects its own regularisation parameter and set of genes, and multitask modelling, where the different models agree on a set of genes to be used for the prediction of each signature, and the regularisation parameter is shared. We evaluate their predictive performance and interpretability using biologically informed metrics. Furthermore, we assess the models' robustness on unseen data by simulating real-world shifts through clustering-based data splits. Our results show that while both models achieve reasonable predictive accuracy, independently trained models offer greater flexibility and interpretability by identifying signature-specific genes and regularisation strengths. These findings suggest that gene expression carries meaningful information about a cell's mutational history and that signature-specific modelling may offer better biological insight into tumour heterogeneity.

1 Introduction

Cancer heterogeneity, the diversity of cancer cell populations within a tumour, is a major obstacle to effective treatment. This heterogeneity arises as cancer cells accumulate mutations, adapt, and evolve over time, resulting in distinct subpopulations with different molecular profiles and drug responses. Therefore, understanding the mechanisms that lead to this heterogeneous behaviour can lead to more effective treatment and diagnosis of tumours.

One of the methods to study these molecular changes is through mutational signature analysis, where patterns of somatic mutations (signatures) are linked to underlying mutagenic processes like DNA repair deficiencies or environmental exposures (e.g., smoking, UV light). For instance, C:G $>$ A:T mutations are predominant in lung cancer caused by smoking [1]. Another approach to study this is through gene expression, which refers to the level of activity of genes in a cell, quantified by the number of mRNA transcripts produced per gene. Gene expression profiles reflect the functional state of a cell, serving as a proxy for cellular behaviour, activity, and phenotype. As such, it provides a dynamic view of the cell's identity and can offer clues about the biological processes it is undergoing, including those related to mutational mechanisms.

Most mutational signature research has been performed on bulk sequencing data, averaging the mutation signals across thousands or millions of cells. Although this has revealed important biological processes, it fails to capture intra-tumour variation. Single-cell sequencing technology now enables the profiling of mutations and gene expression at the resolution of individual cells [2]. However, analysing mutational signatures from single-cell data remains methodologically challenging due to data sparsity, noise, and cost.

A study conducted in 2020 has indicated that the activity of certain mutational processes may be associated with changes in gene expression [3]. Moreover, a recent work explored the relationship between gene expression and mutational signatures in bulk samples using supervised machine learning [4]. In this study, the researchers successfully predicted the presence or absence of specific mutational signatures from gene expression profiles, framing the task as a classification problem. Their findings highlight that gene expression carries meaningful information about underlying mutational processes. More specifically, they found that gene expression profiles allow modelling of roughly 65% of known mutational signatures for each of the 33 cancer types analysed in the study.

However, existing studies have primarily relied on bulk sequencing data, which aggregates genetic signals across all cells within a tumour. This process ignores cell-to-cell variability, limiting the understanding of intra-tumour heterogeneity. Additionally, the paper mentioned in the previous paragraph frames the problem as a classification task, focusing on predicting the presence or absence of mutational signatures rather than quantifying their activity levels. While this has yielded useful insights, it fails to capture the relationship between gene expression and mutational signatures at the single-cell level.

Building on these limitations, this study explores whether mutational signature exposures in single-cell data can be predicted from gene expression profiles using regression models. By focusing on single cells and predicting continuous signature exposures, this work aims to offer a more nuanced and quantitative view of how gene activity reflects underlying mutational processes at cellular resolution. Such insights can help identify dominant mutational mechanisms driving tumour evolution, and potentially guide more targeted treatment strategies.

This paper addresses this critical gap by analysing whether mutational signature exposures of single-cell data are predictable from the cells' gene expression. To this end, we first compare two different modelling paradigms: a regular independent regression, where a separate model is trained for each signature with its own regularisation parameter and selected feature (gene) set; and multitask regression, where all signatures are modelled simultaneously using a shared regularisation parameter and a common set of predictive genes. Although each signature still learns its own weight vector in the multitask setup, the feature selection is coordinated across outputs. Furthermore, we assess how well these models generalise to unseen gene expression profiles by testing them on biologically distinct subsets of the data. This analysis offers insight into the robustness of the models under realistic shifts in data distribution, a key consideration for applications in diagnostics and personalised medicine, given the heterogeneous character of tumours.

The rest of this paper is structured as follows. In Section

2 we explain the methodology used to conduct the research. After that, Section 3 will explain the experimental setup used and present the results obtained from the experiments performed. Section 5 reflects on the ethical aspects of this study. Finally, a discussion of the results is presented in Section 4, followed by the conclusions, limitations and future work of this study in Section 6.

2 Methodology

2.1 Problem Definition

This study addresses the task of predicting mutational signature exposures from gene expression profiles in single-cell data. Mutational signatures are characteristic patterns of somatic mutations left behind by distinct mutagenic processes. Any single cancer cell can exhibit contributions from multiple such processes. The extent to which each mutational signature contributes to the observed mutations in a cell is quantified as its signature exposure - a non-negative, continuos value indicating the relative activity or influence of that signature [1]. These exposure values offer an overall view of the mutational history of each cell.

The data used in this study originates from single-cell RNA sequencing performed on 688 cells isolated from a single breast cancer tumour. Each cell was sequenced to obtain both its gene expression profile and its somatic mutation profile, allowing for integrative analysis at single-cell resolution. To derive mutational signature exposures, non-negative matrix factorisation (NMF) [5] was applied to the mutational profiles matrix M of each cell, which is a count matrix of the mutation types observed in a cell. This decomposition, $M \approx E \cdot S$, yields a matrix S of mutational signatures, where each row defines a signature as a probability-like distribution over mutation types, and an exposure matrix E, where each row corresponds to a cell and each column represents the exposure level of a specific signature.

In parallel, the gene expression of 687 of these cells was also obtained in the process, resulting in a gene expression matrix containing integer counts, where each row corresponds to a gene and each column to a cell. These counts reflect the number of times each of these genes' transcript was expressed in that specific cancer cell.

The objective is to learn a predictive mapping $f : \mathbb{R}^p \to \mathbb{R}^s$ such that $f(\boldsymbol{x_i}) \approx \boldsymbol{y_i}$, where p is the number of genes, s is the number of mutational signatures, $\boldsymbol{x_i}$ is the expression profile of cell i, and $\boldsymbol{y_i}$ is the corresponding vector of signature exposures. In this setup, the gene expression matrix serves as input to the model, while the signature exposure matrix serves as the prediction target.

Therefore, this setup is framed as a multivariate regression problem, where models are trained to predict multiple continuous outputs (signature exposures) for each input (gene expression profile).

To deepen our understanding of how well such models can capture this relationship, we address the following research aims:

First, we examine how the design of the model impacts performance. Specifically, we compare models trained independently per signature (i.e., learning a separate function for each column of Y) to models trained jointly on all signatures (i.e., learning a single multivariate model). This comparison assesses whether modelling signatures in isolation leads to improved predictive performance and interpretability compared to modelling the joint distribution.

Second, we assess the robustness of the models by evaluating their ability to generalise to biologically distinct subsets of the data. Instead of relying on a random split of cells into training, validation, and test sets, we cluster cells based on gene expression and perform a disjoint cluster-based split. This setup simulates a more realistic challenge in personalised medicine: predicting mutational processes in unseen or rare cell subtypes.

Together, these research questions aim to evaluate the predictability of mutational processes from gene expression data, the benefits of different modelling paradigms, and the robustness of the learned relationships.

It is important to note that the gene expression and signature exposure matrices used in this study were provided in processed form. These preprocessing steps were undertaken prior to the commencement of this work. As such, this work focuses solely on building predictive models using these preprocessed inputs.

All data processing, modelling, and analysis in this study were implemented in Python. Core preprocessing, machine learning, and data visualisation tasks were performed using standard scientific libraries, including NumPy, Pandas, SciPy, and scikit-learn. These libraries offer robust tools for handling sparse data, applying standard transformations, and training supervised learning models with cross-validation.

2.2 Data and Preprocessing

As described in the problem definition, this study uses two preprocessed matrices. The gene expression contains integer counts of 36601 genes and was derived from 687 breast cancer cells. The mutational signature exposure matrix was derived from 688 breast cancer cells and contains the exposure information of 96 different signatures, from which only 6 have non-zero exposure (SBS1, SBS5, SBS12, SBS26, SBS40c and SBS54). To prepare these matrices for regression modelling, several preprocessing steps were applied to address scale differences, sparsity, and distributional inconsistencies across features. These choices were made based on best practices in single-cell data analysis and to improve the performance and interpretability of downstream machine learning models [6].

First, it is important to address the discrepancy in the number of cells between the two matrices. To ensure consistency between input and target data, the unmatched cell was removed from the signature exposure matrix prior to any further processing. This alignment step reduced the dataset to 687 cells shared across both matrices.

Next, both matrices were split in three sets: a training set, a validation set, and a test set. This tripartite split is a common practice in supervised learning, as it enables proper model selection and evaluation: the training set is used to fit the models, the validation set is used for early model comparison and assess preprocessing variations, and the test set provides an unbiased estimate of the model's generalisation performance

on unseen data. This was done in two different ways, depending on the research question being addressed in the specific experiment.

In the first experiment, we address the first research question, that of comparing the two modelling paradigms (multitask vs. regular individual modelling). Since the goal here is to analyse and compare the performance of both models, both the gene expression and signature exposure matrices were randomly split into training (70%), validation (15%), and test (15%) sets. This ratio was selected to ensure a sufficient number of samples for training while preserving representative subsets for validation and unbiased evaluation.

The second research question investigates the robustness of predictive models when applied to biologically distinct subsets of cells, a scenario that closely reflects real-world applications such as personalised medicine, where a model may be deployed on previously unseen or rare tumour subtypes. To simulate this, cells were partitioned based on their transcriptional similarity using clustering instead of being randomly split in the second experiment. This approach ensures that the validation and test sets consist of biologically distinct cell groups that the model has not seen during training.

Dimensionality reduction was first applied to the raw gene expression matrix using Principal Component Analysis (PCA). Based on an explained variance plot, 40 principal components were retained, as this number captured over 95% of the cumulative variance and was also the point at which the silhouette scores stabilised across different clustering runs. This helped ensure that meaningful structure was preserved while reducing noise and computational complexity.

Next, KMeans clustering was applied to the PCA-reduced data. The optimal number of clusters was selected by evaluating silhouette scores across a range of values (see Figure 5). Five clusters were chosen, yielding a silhouette score of 0.3045, which indicated a reasonably well-separated clustering structure given the inherent noisiness of single-cell data. The number of clusters was kept small to minimise the overlap between the clusters. The clusters can be seen in Figure 6.

Once clusters were established, they were assigned to training, validation, and test sets in a disjoint manner, meaning that entire clusters of cells were held out to simulate previously unseen subtypes. While this type of split makes it more difficult to control the exact size of each subset, the resulting split was approximately 70% for training, and 15% each for validation and test sets. This ratio was chosen to balance the need for sufficient training data with the importance of evaluating generalisation on biologically distinct samples.

Following the split, the gene expression matrix was normalised using a method commonly applied in single-cell RNA sequencing data: counts per million (CPM) followed by a shifted logarithm transformation (log1p). This approach accounts for varying sequencing depth across cells and stabilises variance across genes, following the best practices in the field [7]. Without this normalisation, highly expressed genes or deeply sequenced cells could disproportionately influence the learning process, potentially leading to biased models. After normalisation, the data was standardised per gene (zero mean and unit variance) using the StandardScaler class from the scikit-learn library [8].

To avoid data leakage and ensure a fair evaluation, the normalisation and standardisation parameters were computed using only the training set. These transformations were then applied to the validation set using the statistics derived from the training data. Once the best model hyper-parameters were selected through cross-validation on the training set (as described in Subsection 2.3), the training and validation sets were merged, and the normalisation and standardisation were recomputed on this combined set. The final model was then retrained on the full `training+validation` data and evaluated on the test set using the updated transformations. This procedure ensures realistic generalisation evaluation, simulating a deployment scenario where future data is processed using only information available at training time, which prevents potential bias in the models.

In contrast, the signature exposure matrix was initially left raw, as these values already represent biologically meaningful proportions derived from non-negative matrix factorisation (NMF) and are typically interpreted on their original scale. Normalising exposures can lead to different biological interpretations compared to using absolute exposures. However, to assess whether differences in scale across signatures might bias model learning, additional experiments were conducted in which the same normalisation and standardisation pipeline used for the gene expression matrix was applied to the exposure matrix.

2.3 Modelling Approaches

To investigate the predictability of mutational signature exposures from gene expression data, this study employs two types of linear regression models: independent per-signature ridge regression and multitask regression using a variant of the lasso regression model from the scikit-learn library. These models were chosen for their interpretability, computational efficiency, and their differing assumptions about the relationships between output variables. This contrast allows us to investigate whether modelling each signature in isolation leads to improved predictive performance and interpretability.

Ridge Regression

Ridge regression is a linear model that includes L2 regularisation, which penalises large coefficients and helps prevent over-fitting in high-dimensional settings such as gene expression data. The model learns a weight vector $w \in \mathbb{R}^p$ for each target (signature) by minimising the following objective function:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^\top \mathbf{w}\right)^2 + \alpha\|\mathbf{w}\|_2^2 \tag{1}$$

Here, n denotes the number of training samples (cells), $\mathbf{x}_i \in \mathbb{R}^p$ is the gene expression vector of cell i, $y_i \in \mathbb{R}$ is the corresponding scalar exposure value for a given mutational signature, $\mathbf{w} \in \mathbb{R}^p$ is the weight vector to be learned, and $\alpha \in \mathbb{R}_{\geq 0}$ is the regularisation hyper-parameter controlling the strength of the L2 penalty. In this study, a separate Ridge regression model is trained for each signature, allowing for output-specific parametrisation and regularisation.

By training one model per signature, we can also directly assess which genes are most predictive for each mutational process separately, thus contributing to interpretability, which is essential when linking biological insights to model behaviour.

In this study, ridge regression was implemented using the RidgeCV class from the scikit-learn library [9]. This class extends normal ridge regression by performing internal cross-validation to automatically select the optimal parameter α from a predefined list of candidate values. Specifically, RidgeCV fits multiple ridge models with different values of α, and uses k-fold cross-validation to evaluate each model based on its mean squared error on the validation folds. The α value that minimises the average validation error is then selected, and a final ridge model is trained using this optimal α on the full training set. The k value used in the experiments presented in Section 3 is 10.

Since this study models each mutational signature independently, a separate RidgeCV instance was trained for each output variable (signature), allowing the regularisation strength to adapt to the unique signal and noise characteristics of each target. As a result, each signature gets its own regression model, with its own α value and weight vector. Biologically this reflects the idea that each signature may be influenced by different sets of genes.

MultiTaskLasso

MultiTaskLassoCV extends lasso regression to the multi-output setting, enabling the prediction of multiple targets from a shared set of features. It includes an L1/L2 mixed-norm regularisation to promote sparsity in the model [10], but does so in a way that enforces sparsity jointly across all outputs. The model's optimisation objective is:

$$\mathcal{L}(\mathbf{W}) = \frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \alpha \sum_{j=1}^{p} \sqrt{\sum_{k=1}^{s} W_{jk}^2} \quad (2)$$

Here, n is the number of samples (cells), p is the number of input features (genes), and s is the number of output tasks (mutational signatures). $\mathbf{X} \in \mathbb{R}^{n \times p}$ denotes the gene expression matrix, $\mathbf{Y} \in \mathbb{R}^{n \times s}$ the matrix of signature exposures, and $\mathbf{W} \in \mathbb{R}^{p \times s}$ the matrix of regression coefficients, where W_{jk} indicates the contribution of gene j to signature k. The regularisation parameter $\alpha \in \mathbb{R}_{\geq 0}$ controls the strength of the penalty. The second term in the objective, $\|\mathbf{W}\|_{2,1}$, promotes row-wise sparsity in $\mathbf{W}$, encouraging the model to select a common set of genes predictive across all signatures.

The MultiTaskLassoCV class from the scikit-learn library [11] performs k-fold cross-validation over a range of α values and selects the one that minimises the average mean squared error on the validation folds. Once selected, the final model is retrained on the entire training set using the best α. As done for the ridge model, a k value of 10 was used in the experiments conducted.

As a result of this formulation, although each signature has its own model, the alpha value is shared, and the set of genes that contribute to the prediction values is the same across all signatures.

2.4 Evaluation Metrics

To evaluate the performance of the models in predicting mutational signature exposures from gene expression data, we employ a set of complementary metrics that capture different aspects of predictive accuracy and structural alignment. Specifically, we use the coefficient of determination (R^2), and the mean squared error (MSE). In addition, we explore a mutation profile reconstruction metric to assess the biological relevance of predicted exposures.

Coefficient of Determination (R^2)

The R^2 score is a standard metric for regression tasks that measures the proportion of variance in the target variable that is explained by the model. For a single target y, it is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

We report both the overall R^2 score (averaged over all targets) and individual R^2 scores per signature. A score of 1 indicates perfect prediction, while scores near or below 0 indicate poor fit. This metric helps assess how well the model captures the variance in each signature's exposure values.

Mean Squared Error (MSE)

The MSE is used to quantify the average squared difference between predicted and true exposures. It penalizes larger errors more heavily than small ones, making it sensitive to outliers. While R^2 is scale-invariant, MSE provides a scale-sensitive view of prediction quality. We also report the MSE-to-variance ratio to assess how the prediction error compares to the natural variability in the data. A low ratio indicates that the model captures a large portion of the variation in the signature exposures.

3 Experimental Setup and Results

3.1 Predictive Performance Across Modelling Paradigms

To address the first research question — how regular independent per-signature modelling compares to multitask modelling in predicting mutational signature exposures from single-cell gene expression data — the focus was to explore the two different modelling strategies. The first selects a different set of genes for each signature, while the second encourages the models to select the same set of features across all signatures.

Both the data split and preprocessing pipeline were applied as described in Section 2.2. The preprocessing pipeline was only applied to the gene expression data. Both ridge and lasso models were trained as explained in Section 2.3 and evaluated with the metrics presented in Section 2.4. The prediction plots can be found in Appendix B.

Model Performance Comparison

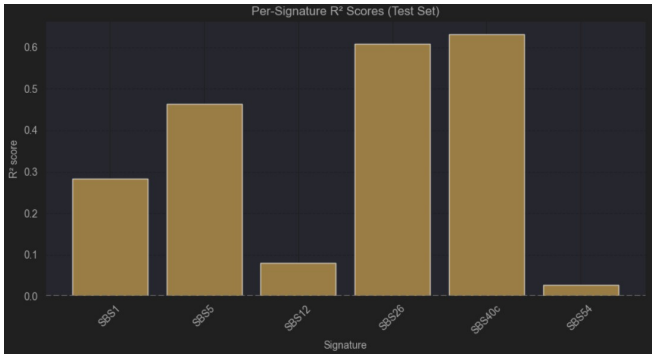
The overall performance of both models on the test set is summarised in Table 1. The independently trained RidgeCV models achieved a slightly higher R^2 score, as well as a lower mean squared error and MSE-to-variance ratio. These results

indicate that RidgeCV models captured a greater proportion of the variability in the data and produced more accurate exposure predictions.

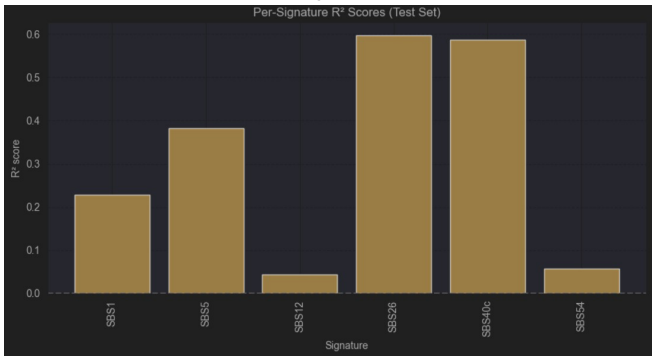
Table 1: Metrics calculated for both models on the test set for experiment 1.

Model	R^2	**MSE**	$\frac{MSE}{Variance}$
RidgeCV	0.35	1155.77	0.171
MultiTaskLassoCV	0.32	1234.42	0.182

To further explore the performance of each model, Figure 1 shows the R^2 scores achieved for each individual signature.



(a) Ridge model.



(b) Lasso model.

Figure 1: Individual R^2 scores per signature for experiment 1.

These plots suggest that signatures that are more present in the cells, like SBS26 and SBS40c, are predicted with more accuracy than rarer signatures, such as SBS12 and SBS54. Furthermore, the models appear to learn signatures with higher variance better as well, as the scores for SBS26 and SBS40c are higher than for SBS1 and SBS5.

From an interpretability perspective, the two approaches yielded distinct patterns of gene selection. The RidgeCV models identified different genes for each signature, enabling insights into signature-specific regulatory mechanisms. To identify the most influential genes, the top 10 were selected for each signature by ranking the absolute values of the learned regression coefficients (refer to Table 2. This approach revealed unique sets of genes for each mutational signature, reflecting the model's ability to uncover potentially distinct biological mechanisms underlying different mutational processes.

Table 2: Top 10 genes selected by RidgeCV for each mutational signature.

SBS1	**SBS5**	**SBS12**	**SBS26**	**SBS40c**	**SBS54**
LINC01228	KRTAP3-3	GAN	KRTAP3-3	KRTAP3-3	C8orf34
AC009831.3	GOLGA6L9	AC011461.1	AC012358.1	GOLGA6L9	AC005237.1
FAM166A	MAP1LC3B2	TRPC3	LINC01625	UGT1A1	C9orf147
NUP93	UGT1A1	AL512625.2	GOLGA6L9	MAP1LC3B2	LINC01987
AP003096.1	FGD4	AC092287.1	AC015727.1	MINAR2	HSD17B3
CUEDC1	CAVIN1	AP003096.1	KLHL2	SLC12A1	AP000462.3
EHMT2-AS1	AP000787.1	AL590822.2	OXR1	AL035427.1	DCTN1-AS1
NPFF	RASSF1-AS1	CDHR2	FGD4	FGD4	AC110609.1
AL512625.2	MINAR2	MTPAP	TIGD4	CAVIN1	AL731684.1
AL391001.1	AL392172.1	AC008393.1	WASF3	AP000787.1	AC087362.1

In contrast, MultiTaskLassoCV imposes a joint sparsity constraint across all outputs, resulting in a single set of predictive genes shared across signatures. Specifically, it selects genes with non-zero rows in the coefficient matrix, meaning these genes contribute to the prediction of at least one signature. In total, the MultiTaskLassoCV model selected 292 genes with non-zero coefficients. Among them, the top 10 most influential genes were determined by summing the absolute values of their coefficients across all signatures, thus highlighting genes with the strongest and most consistent contribution across the prediction tasks. Figure 2 shows the coefficients per signature of the top 10 genes selected by the MultiTaskLassoCV model.

Figure 2: Gene Coefficients per signature of the top 10 genes selected by the MultiTaskLassoCV model. Each row relates to a gene, and is labelled gene_name(gene_index).

Table 3 reports the Jaccard coefficient between the top 10 genes selected by each of the ridge models and the top 10 selected for the multitask lasso model. The Jaccard coefficient tells us how similar two sets are to each other, by calculating the ratio between their intersection and their union. Overlapping genes are also shown in the table. Notably, several genes were shared between the two models, such as AP003096.1, WASF3, and AC008393.1, suggesting that these genes may play a broad regulatory role across multiple mutational processes. On the other hand, many genes were unique to specific RidgeCV models, reinforcing the idea that individual signatures may be influenced by distinct gene sets that are not captured when a joint sparsity constraint is imposed.

Table 3: Jaccard coefficients between top genes selected by MultiTaskLassoCV and individual RidgeCV models. Overlapping genes are also listed.

Signature	Jaccard Coefficient	Overlapping Genes
SBS1	0.053	AP003096.1
SBS5	0.000	—
SBS12	0.111	AC008393.1, AP003096.1
SBS26	0.053	WASF3 (22994)
SBS40c	0.000	—
SBS54	0.000	—

These findings demonstrate that RidgeCV allows for more fine-grained gene attribution per signature, potentially offering better biological resolution when interpretability is a priority. Meanwhile, MultiTaskLassoCV identifies a concise subset of shared predictors, which may be advantageous for identifying global biological factors or developing compact predictive assays. Together, these gene-level comparisons highlight the trade-off between predictive specificity and interpretability inherent in regular versus multitask independent modelling strategies.

3.2 Generalisation to Biologically Distinct Subsets

Following the split, the same preprocessing pipeline described in Section 2.2 was applied, and the two modelling strategies from Section 2.3 were evaluated using the metrics defined in Section 2.4. By comparing model performance in this clustered-based split scenario, the experiment investigates whether predictions remain reliable under realistic shifts in gene expression distributions.

Table 4 shows the metrics that quantify predictive performance on unseen data. For the graphs, please refer to Appendix C.

Table 4: Metrics calculated for both models on the test set for experiment 2.

Model	R^2	MSE	$\frac{MSE}{Variance}$
RidgeCV	-0.20	1561.60	0.601
MultiTaskLassoCV	-0.23	1635.07	0.630

Both models exhibit a considerable drop in performance compared to the random split scenario (see Table 1). The R^2 scores became negative, indicating that the models performed worse than a naïve mean predictor for the test data.

It is important to highlight that clustering and data splitting can have a significant influence on the observed results. Since clusters are based on unsupervised patterns in gene expression, the biological distinctiveness of held-out clusters can vary, which introduces variability in the evaluation outcome. As such, while this split better simulates real-world generalisation, scores may vary notably depending on cluster configuration.

Overall, these findings reinforce the importance of evaluating models under biologically realistic conditions. Random splits may give overly optimistic estimates of generalisation, while cluster-based evaluation provides a more stringent and informative test of robustness.

3.3 Effect of Preprocessing Both Gene Expression and Signature Exposures

In this final experiment, the impact of applying normalisation and standardisation to both gene expression and signature exposure matrices was investigated. This setup tests whether eliminating scale discrepancies across signatures, by transforming the target data to a standard range, can improve regression performance or model stability.

The same random 70/10/20 train/validation/test split and preprocessing pipeline were used as in Experiment 1. The models were also trained in the same manner. The graphs can be found in Appendix **??**.

Table5 reports the global metrics for both the RidgeCV and MultiTaskLassoCV models.

Table 5: Metrics calculated for both models on the test set for experiment 3.

Model	R^2	MSE	$\frac{MSE}{Variance}$
RidgeCV	0.26	0.61	0.706
MultiTaskLassoCV	0.10	0.75	0.864

While both models experienced a drop in performance compared to their counterparts in Experiment 1, the ridge model retained better predictive power. In contrast, the MultiTaskLasso model struggled more under this configuration. This suggests that standardising the exposures, although useful for balancing target scales, may obscure biologically meaningful differences in signature prevalence.

Interpretability patterns also shifted. Figure 3 shows the top 10 genes selected by MultiTaskLassoCV based on the aggregated coefficient magnitude across all signatures, and their contributions to each of the signatures. The selected genes by each of the ridge models is shown in Table 6

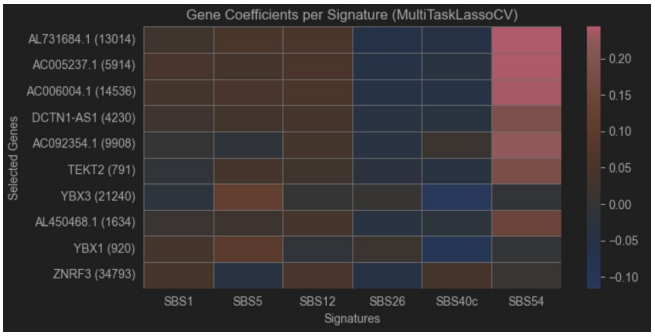


Figure 3: Gene Coefficients per signature of the top 10 genes selected by the MultiTaskLassoCV model. Each row relates to a gene, and is labelled ¡gene_name¿(¡gene_index¿).

Table 6: Top 10 genes selected by RidgeCV for each mutational signature (experiment 3).

SBS1	SBS5	SBS12	SBS26	SBS40c	SBS54
CUEDC1	KRTAP3-3	AC092287.1	AC092287.1	KRTAP3-3	AL157394.3
AL031005.2	GOLGA6L9	GAN	GAN	GOLGA6L9	C8orf34
SYCP3	MAP1LC3B2	MTPAP	MTPAP	MAP1LC3B2	AC087362.1
LINC01228	CAVIN1	AC015727.1	AC015727.1	CAVIN1	AL731684.1
RGL3	AP000787.1	AC008393.1	AC008393.1	FGD4	AL355338.1
AC092268.2	FGD4	WASF3	WASF3	AP000787.1	AC005237.1
HSPE1	UGT1A1	AL512625.2	NEK8	UGT1A1	AC092354.1
PFKFB1	RASSF1-AS1	NEK8	AC012358.1	RASSF1-AS1	FZD4-DT
AC016831.7	AL392172.1	FAM98C	LINC01625	AL392172.1	RHAG
TSIX	AC022007.1	LINC01625	FAM98C	AC022007.1	AC003101.1

Table 7 presents the Jaccard coefficients between the top genes selected by each of the ridge models and the top genes identified by MultiTaskLassoCV in Experiment 3. The overlapping genes are also shown in the table. Unlike in Experiment 1, where shared genes appeared across several signatures, here the overlap is almost exclusively associated with signature SBS54. Genes such as **AL731684.1**, **AC005237.1**, and **AC092354.1** were selected by both models, but not across multiple signatures. This may suggest that normalisation of both input and output altered the relative scale and influence of individual signatures, causing MultiTaskLassoCV to focus on a narrower subset of predictive features. Additionally, RidgeCV still selected distinct genes per signature, while MultiTaskLassoCV enforced joint sparsity, leading to more overlap in specific, but possibly dominant, signatures. This contrast highlights the impact of preprocessing choices on feature selection and model interpretability.

Table 7: Jaccard coefficients between top genes selected by MultiTaskLassoCV and individual RidgeCV models (Experiment 2). Overlapping genes are also listed.

Signature	Jaccard Coefficient	Overlapping Genes
SBS1	0.000	—
SBS5	0.000	—
SBS12	0.000	—
SBS26	0.000	—
SBS40c	0.000	—
SBS54	0.176	AC005237.1 (5914), AC092354.1 (9908), AL731684.1 (13014)

Overall, this experiment underscores a key trade-off: standardising targets may improve numerical stability and help some models converge, but it can also mask biologically important signal encoded in absolute exposure levels. For practical applications where interpretability and biological relevance are priorities, using raw or proportionally scaled exposure values may be preferable.

4 Discussion

4.1 Model Performance and Predictive Trade-offs

The comparison between independent RidgeCV models and the multitask MultiTaskLassoCV models revealed important differences in both predictive performance and interpretability. While RidgeCV slightly outperformed MultiTaskLassoCV in terms of R^2 and mean squared error (MSE) in both experimental settings, the latter offered a more compact and globally sparse solution. This trade-off highlights the difference between per-target flexibility and joint regularisation: RidgeCV adapts to the specific noise and signal characteristics of each signature independently, whereas MultiTaskLassoCV enforces sparsity across all outputs simultaneously, potentially at the cost of fine-grained accuracy.

Nonetheless, the multitask models still performed quite well compared to the regular ridge models. Several hypothesis can be made based on this. One is that the signatures may arise from related biological pathways, in which case a share gene set could still capture meaningful signals. Another possibility is that the gene expression data contains correlated features (genes). In this case, genes selected by the multitask models might be correlated to the ones selected by the ridge models. Alternatively, this could also be due to the high sparsity of single-cell data.

From an interpretability perspective, the two models produced distinct patterns of gene selection. The RidgeCV models identified different sets of top predictive genes for each signature, which may reflect signature-specific regulatory mechanisms. In contrast, MultiTaskLassoCV selected a shared set of 292 genes with non-zero coefficients, many of which contribute meaningfully across multiple signatures. This global sparsity constraint can aid biological interpretation by identifying a compact gene set potentially involved in multiple mutational processes.

Interestingly, several genes such as *WASF3*, *MAP1B*, and *AP003096.1* were selected by both modelling approaches, suggesting their relevance in explaining variation in signature exposures. These overlaps may point to common transcriptional responses to DNA damage or shared pathways activated across signatures, highlighting the potential for downstream biological investigation.

4.2 Robustness to Distribution Shift

The second experiment, based on stratified splits via clustering in PCA space, revealed a significant drop in model performance. Both models exhibited negative R^2 scores and substantially higher error metrics compared to the random split setting. This result underscores the challenge of generalising to unseen or rare cell subtypes, a common scenario in personalised medicine. The clustering-based split introduces a realistic distribution shift that tests the robustness of learned relationships. These findings suggest that training models on more diverse and representative cell populations, or incorporating techniques designed for domain generalisation, may improve predictive stability in real-world applications.

4.3 Limitations

Several limitations should be acknowledged. First, the analysis was restricted to a single tumour sample consisting of 688 cells, limiting the generalisability of the findings. Second, while gene selection results were presented and partially interpreted, no external biological validation was performed, such as enrichment analysis or cross-referencing with known cancer gene databases. Lastly, the number of PCA components and clusters used in the second experiment were chosen heuristically; these choices may influence generalisation performance and could be optimised more systematically in future work.

5 Responsible Research

5.1 Reproducibility

In this study, every effort was made to ensure that the data preprocessing, model training, and evaluation pipelines are transparent and reproducible. All data transformations — including normalisation, standardisation, and clustering — were implemented with fixed random seeds where applicable to ensure consistent results. Additionally, all models were trained using well-established libraries such as scikit-learn, with hyper-parameter selection performed via cross-validation using clearly defined procedures.

To further promote reproducibility, modular Python notebooks were developed with clear documentation, enabling independent reruns and parameter adjustments. Still, given the non-deterministic nature of some steps (e.g., clustering, data splits), slight variation in performance metrics may occur, particularly in experiments involving splits based on biological clusters.

5.2 Ethical Considerations in Data Usage

The data used in this study originates from single-cell RNA sequencing of a human breast cancer tumour. Despite the anonymisation of the data, it is essential to recognise that these cells were derived from a real patient. Ethical use of such data requires respecting the context in which it was collected and acknowledging the patient contribution that made this analysis possible.

Moreover, although the analysis is computational in nature, it ultimately seeks to extract insights that could inform biological understanding or, in the long term, clinical decision-making. Therefore, ethical responsibility extends to ensuring that models are interpreted cautiously, especially when derived from a limited sample without independent clinical validation.

5.3 Clinical Applicability and Caveats

Although the findings demonstrate that mutational signature exposures can be predicted from gene expression to a certain extent, several caveats must be acknowledged before applying these results in clinical contexts.

First, the data used in this study came from a single tumour, limiting generalisability. Models trained on this data may not perform similarly on other cancer types, subtypes, or patients. Second, although some interpretability is gained through gene selection analysis, these models should not be interpreted as proving causality. The identification of predictive genes does not necessarily imply that these genes are directly involved in the mutagenic processes they predict, and many more similar experiments would have to be performed before that could be said.

Lastly, predictive performance on unseen data — particularly under biologically distinct conditions — was significantly lower, as seen in the clustered data split experiment. This highlights the limitations of model robustness and underscores the need for further validation across diverse datasets before any clinical integration can be considered.

6 Conclusions and Future Work

6.1 Conclusions

This study investigated the predictability of mutational signature exposures from gene expression profiles at the single-cell level. By comparing two modelling paradigms, per-signature regression using RidgeCV and joint modelling using MultiTaskLassoCV, we addressed whether signatures are better modelled independently or collectively. The results suggest that both models are capable of learning predictive relationships, with RidgeCV achieving slightly higher R^2 and lower mean squared error on randomly split test sets. Importantly, RidgeCV also allowed for more interpretable results by selecting distinct gene sets per signature, potentially revealing signature-specific regulatory influences.

However, the MultiTaskLassoCV model, while slightly less accurate, identified a shared set of predictive genes, pointing to common underlying pathways. This model may be especially useful when seeking a sparse, biologically plausible set of predictors shared across mutational processes.

The second research question focused on the robustness of these models when applied to biologically distinct, unseen data. Using a clustered split based on PCA and k-means, both models showed significantly reduced performance, indicating limited generalisation capacity in the presence of distributional shifts. This finding underscores the challenges of deploying such models in clinical or diagnostic settings without broader training data and domain adaptation.

6.2 Future Work

Suggestions on how to extend this work include:

- **Reconstructing mutational catalogues from predicted exposures:** A valuable next step is to reconstruct the mutation profiles using predicted exposures and evaluate how well the models recover the original mutational landscape. This could provide a biologically grounded assessment of model utility.
- **Expanding to multi-patient datasets:** Generalisation could be improved by training and evaluating models across samples from multiple tumours or cancer types, allowing broader applicability and testing robustness across diverse contexts.
- **Run experiments with nonlinear models:** Running experiments with more powerful models such as SVR could yield different results, giving more insight into the relationship between gene expression and signature exposure at the single-cell level.
- **Gene enrichment analysis:** To better understand the biological functions associated with the top predictive genes selected by each model, gene enrichment analysis could be performed. This would reveal whether particular pathways, biological processes, or cellular components are overrepresented, potentially linking specific gene sets to mutagenic mechanisms or tumour phenotypes. Such analysis could further validate the relevance of the selected genes and provide insight into the molecular context of different signatures.

References

[1] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, *et al.*, "Deciphering signatures of mutational processes operative in human cancer," *Cell Reports*, vol. 3, no. 1, pp. 246–259, Jan. 2013. DOI: 10.1016/j.celrep.2012.12.008. [Online]. Available: https://doi.org/10.1016/j.celrep.2012.12.008.

[2] A. S. Nam, R. Chaligne, and D. A. Landau, "Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics," *Nature Reviews Genetics*, vol. 22, no. 1, pp. 3–18, 2021.

[3] J. Ping, O. Oyebamiji, H. Yu, *et al.*, "Mutex: A multifaced gateway for exploring integrative pan-cancer genomic data," English, *Briefings in Bioinformatics*, vol. 21, no. 4, pp. 1479–1486, Jul. 2020. DOI: 10.1093/bib/bbz084. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7373173/.

[4] L. Jiang, H. Yu, and Y. Guo, "Modeling the relationship between gene expression and mutational signature," *Quantitative Biology*, vol. 11, no. 1, pp. 31–43, Mar. 2023. DOI: 10.15302/J-QB-022-0309. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.15302/J-QB-022-0309.

[5] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999. DOI: 10.1038/44565.

[6] L. Heumos, A. Schaar, C. Lance, *et al.*, "Best practices for single-cell analysis across modalities," *Nature Reviews Genetics*, vol. 24, no. 8, pp. 550–572, 2023. DOI: 10.1038/s41576-023-00586-w. [Online]. Available: https://doi.org/10.1038/s41576-023-00586-w.

[7] A. Schaar, *Normalization*. [Online]. Available: https://www.sc-best-practices.org/preprocessing_visualization/normalization.html (visited on 05/02/2025).

[8] D. Cournapeau, *Standardscaler*, 2007. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[9] D. Cournapeau, *Ridgecv*, 2007. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeCV.html.

[10] D. Cournapeau, *Multi-task-lasso*, 2007. [Online]. Available: https://scikit-learn.org/stable/modules/linear_model.html#multi-task-lasso.

[11] D. Cournapeau, *Multitasklassocv*, 2007. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.MultiTaskLassoCV.html.

A Clustering graphs

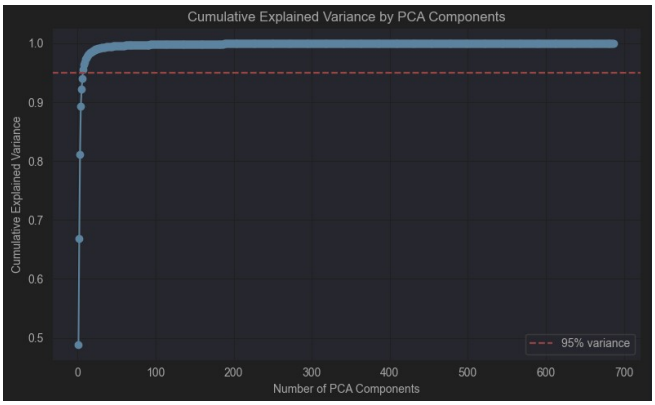


Figure 4: Explained variance by PCA components graph.

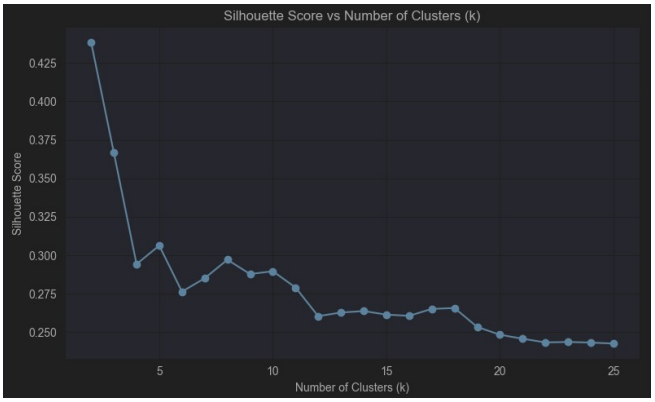


Figure 5: Number of clusters vs silhouette value graph.

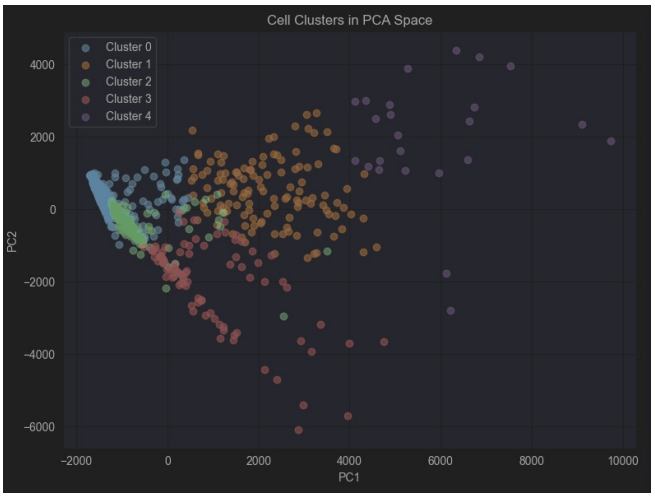


Figure 6: Clusters in the PC environment.

B Prediction plots for experiment one

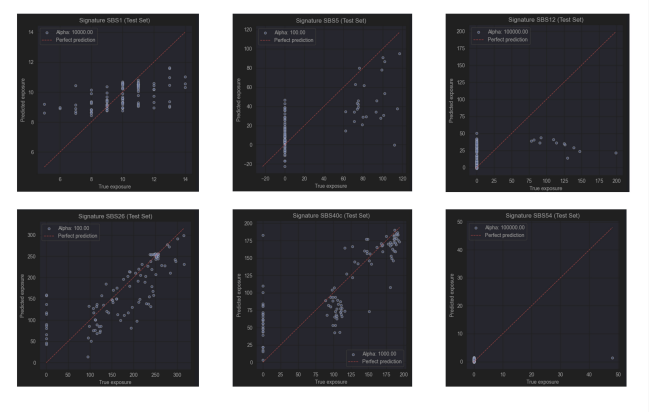


Figure 7: Predictions made by the RidgeCV model on the test set. Each point represents the prediction of one cell for that specific signature.

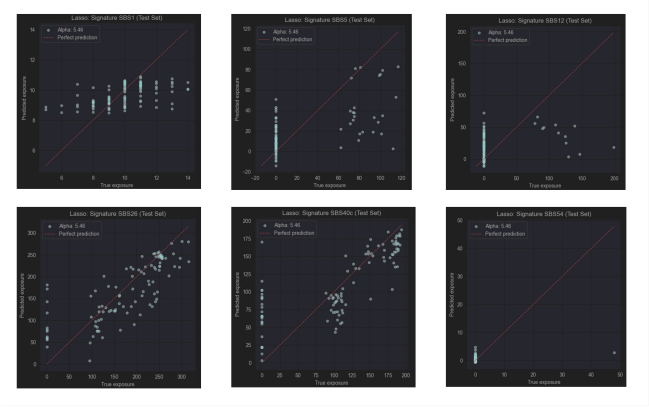


Figure 8: Predictions made by the MultiTaskLassoCV model on the test set. Each point represents the prediction of one cell for that specific signature.

C Graphs of the second experiment

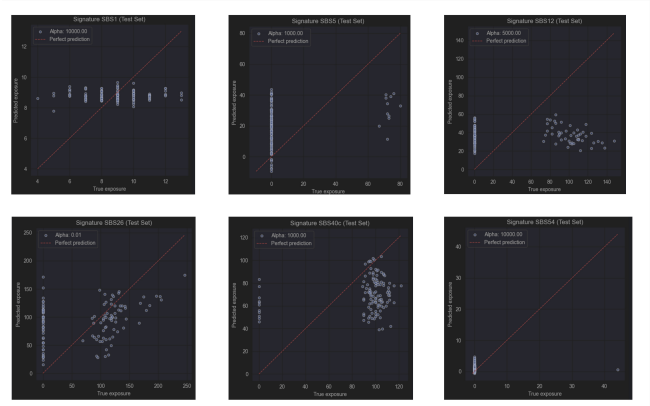


Figure 9: Predictions made by the RidgeCV model on the test set. Each point represents the prediction of one cell for that specific signature.

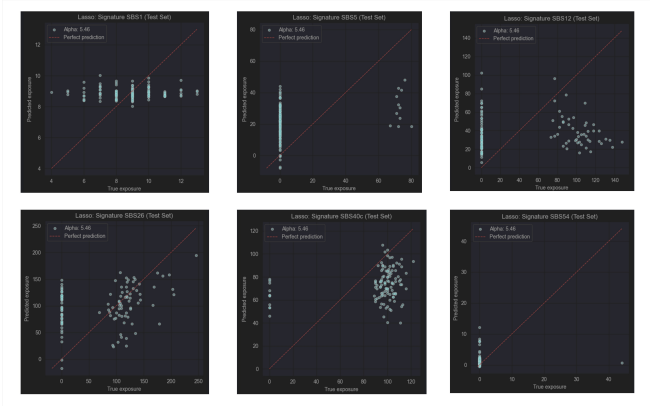
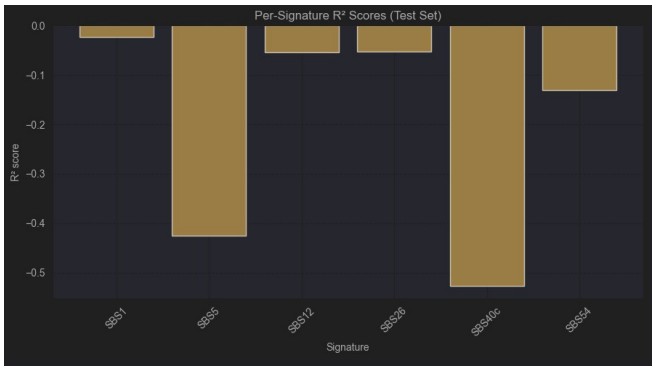
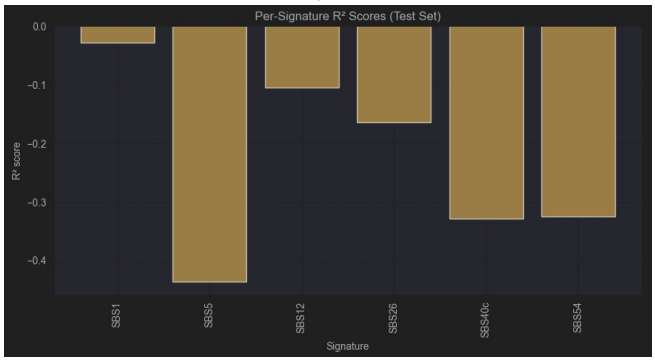


Figure 10: Predictions made by the MultiTaskLassoCV model on the test set. Each point represents the prediction of one cell for that specific signature.



(a) Ridge model.



(b) Lasso model.

Figure 11: Individual R^2 scores per signature for experiment 2.

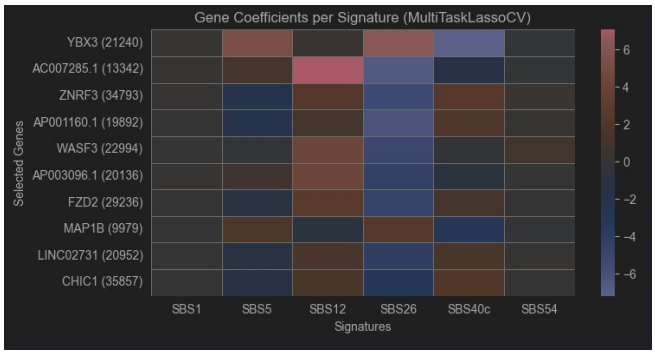


Figure 12: Gene Coefficients per signature of the top 10 genes selected by the MultiTaskLassoCV model. Each row relates to a gene, and is labelled ¡gene_name¿(¡gene_index¿).

Table 8: Top 10 genes selected by RidgeCV for each mutational signature (experiment 3).

SBS1	**SBS5**	**SBS12**	**SBS26**	**SBS40c**	**SBS54**
LINC01228	KRTAP3-3	AC007285.1	KRTAP3-3	KRTAP3-3	AC005237.1
CUEDC1	GOLGA6L9	TRPC3	GOLGA6L9	GOLGA6L9	AC087362.1
FAM166A	EFCAB5	AC011461.1	LINC01625	EFCAB5	SLC13A2
AP003096.1	UGT1A1	WASF3	AC012358.1	UGT1A1	AL731684.1
NXNL1	AC253576.2	AP003096.1	WASF3	MAP1LC3B2	CCDC196
SYCP3	MAP1LC3B2	CDHR2	FGD4	SLC12A1	AC073326.1
EHMT2-AS1	SLC12A1	GP2	AC007285.1	AC253576.2	ENPP3
NPFF	AC092268.2	AC004223.2	ZNF221	AL392172.1	CUZD1
FAM92B	AC093151.2	AC023886.1	AC022007.1	AC092268.2	TMEM95
MBD3	FGD4	AC090589.2	TEP1	AL035427.1	AP000915.1