Robustness of Fitted Mutational Signature Exposures in Single-Cell Data

Deciphering Cancer Heterogeneity with Machine Learning

Rebecca Nys[1]

Supervisors: 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: Rebecca Nys
Final project course: CSE3000 Research Project
Thesis committee: Joana Gonçalves, Sara Costa, Ivan Stresec, Catharine Oertel

Abstract

Tumor heterogeneity complicates mutational signature analysis at the single-cell level, where sparse catalogues and uneven mutation burdens can destabilise exposure estimates. This study quantifies the robustness of fitted mutational signatures in single-cell RNA-seq data from 688 breast-cancer cells. Known COSMIC v3.4 SBS96 signatures were assigned with SigProfilerAssignment and the input data was systematically perturbed by randomly deleting 5%, 10%, 20% and 40% of mutations, repeating each perturbation twenty times. Robustness was assessed with four complementary metrics: (i) persistence of each signature in the dataset, (ii) stability of the number of cells containing each signature, (iii) mean relative error of per-signature exposures, and (iv) per-cell cosine similarity between original and perturbed exposure vectors.

Six signatures (SBS1, 5, 12, 26, 40c and 54) were consistently recovered, even after 40% deletion, demonstrating that core biological signals may survive substantial data loss. Nevertheless, higher deletion levels triggered progressive overfitting: the number of additional signatures rose from three at 5% deletion to eighteen at 40%. Exposures seemed to shift between highly similar signature pairs (e.g., SBS12 and SBS26, SBS5 and SBS40c), and merging such pairs halved the mean relative error. Signature SBS54, detected in only eight cells and suspected to be artefactual, showed the poorest stability. Across cells, robustness scaled positively with the number of mutations per cell ($\rho \approx 0.38$ to 0.59) and negatively with entropy of the exposure vectors ($\rho \approx -0.27$ to -0.53), indicating that abundant or signature-dominated catalogues resist perturbation, whereas sparse or evenly distributed ones are more fragile. Together, our results indicate that while some signatures and cells can survive substantial data loss, signature exposures in sparse single-cell catalogues must be interpreted with caution.

1 Introduction

Not all cancer cells of a tumor are genetically identical, as different cells may carry different somatic mutations. Mutations are changes in the DNA sequence of a cell [1]. This genomic diversity within a single tumor is referred to as intra-tumor heterogeneity [2] and plays a central role in treatment resistance. Therapies that act on specific mutations may fail if those mutations are present only in a subset of tumor cells, allowing other subclones to survive. It is therefore important to understand heterogeneity in order to improve cancer treatments [3].

Cancer cells accumulate mutations caused by both exogenous sources, such as UV light and chemotherapy, and endogenous sources, including cellular processes like metabolism. These factors lead to DNA damage, which is normally corrected by DNA repair processes; however, if these repair processes are defective, the damage can persist. This, in turn, shapes the resulting pattern of mutations in the cancer genome. Mutational signatures are patterns of mutations that describe which mutational process generated a particular combination of mutations; e.g., tobacco smoking leaves a different imprint on the cancer genome than defective DNA mismatch repair does. A cancer's mutational catalogue is therefore a mixture of multiple signatures, and the number of mutations caused by a specific signature is called the signature exposure [4].

Mutational signatures can be extracted from genomic data using non-negative matrix factorization [5]. This de novo extraction technique can even discover previously unknown signatures [6]. Alternatively, a numerical optimization approach can assign a set of known signatures to a sample; in this case the signatures are being fitted to the data [7]. Such analyses have traditionally relied on bulk sequencing data, which aggregates mutations from an entire tumor sample so that it is impossible to determine which mutations occurred in which individual cells. Recently, however, mutations have been called at the single-cell level using single-cell RNA sequencing (scRNA-seq) data [8], promising deeper insights into tumor heterogeneity. It nevertheless remains uncertain whether the current mutational signature methodology works effectively on single-cell data [9]. The paper by Alexandrov et al. [4] already examined factors that influence the ability of extracting signatures; such as the number of genomes, the number of mutations per genome and the strength of exposure, but those experiments were limited to bulk data.

A particular challenge of scRNA-seq data is its sparsity, due to its low coverage and high dropout rate. This means only a small fraction of the genome is sequenced per cell and many mutations are not detected. After additional filtering to retain only high-quality calls [8], the resulting catalogue of mutations can be incomplete or biased, potentially destabilising signature fitting. Hence, the objective of this study is to assess the robustness of fitted mutational signature exposures in single-cell data under such loss of mutations. By exploring different levels of data loss, we aim to understand the sensitivity of the signature fitting process and identify which factors lead to greater or lesser stability of fitted signature exposures.

2 Methodology

2.1 Data

This study uses scRNA-Seq data from 688 breast cancer cells from a female donor aged 65 [10]. Variant calling was performed based on the approach described by Liu et al. [8], and the mutations were filtered using the best practices of the Genome Analysis Toolkit (GATK) pipeline [11]. This way, only high quality mutations are retained, ensuring that the analysis is done on reliable data. Both the variant calling and filtering were performed by my supervisors. This resulted in a dataset where each cell is represented as an individual Variant Call Format (VCF) file, containing the somatic mutations detected in that cell. Each VCF file specifies the chromosome number on which each mutation is located, the genomic position on the chromosome, a quality indicator showing whether

the mutation passed quality checks, the reference nucleotide at that genomic position, and the alternate nucleotide, which represents the observed mutation.

At this stage, minimal data preprocessing was required by me, consisting only of ensuring that the individual VCF files were correctly structured to meet the input requirements of the signature fitting tool that is described in the next section. The processed VCF files should list each mutation with the chromosome number, genomic position, cell identifier, reference nucleotide, and alternate nucleotide. An example of the resulting format is:

```
1 629896 Breast_Cancer_3p_LT_AATCACGAGAAATTGC-1 T A
1 631946 Breast_Cancer_3p_LT_AATCACGAGAAATTGC-1 A C
1 631983 Breast_Cancer_3p_LT_AATCACGAGAAATTGC-1 A C
```

2.2 Signature Fitting

Mutational signatures represent characteristic patterns of mutations caused by different mutational processes, such as exposure to UV light or failures in DNA repair mechanisms. The classification of mutations is based on the six types of base substitution-C>A, C>G, C>T, T>A, T>C, T>G-and also takes into account the nucleotides immediately 5' and 3' to the mutated base. This results in 96 possible mutation types [12].

One common approach to identify these signatures is de novo extraction, which can uncover new signatures and uses non-negative matrix factorization (NMF) [6]. Conceptually, NMF decomposes a matrix of mutation counts M, where each column represents a genome and each row represents a mutation type, into two smaller matrices: $M \approx P \times E$. P contains the mutational signatures, defined as a discrete probability density function over the mutation types, where the columns represent the signatures and the rows represent the mutation types. E contains the exposures, i.e., the number of mutations of a genome attributed to a specific signature, where each column represents a genome and each row represents a signature [4].

In contrast, this study uses signature fitting, a different approach in which the signatures are assumed to be known in advance. Specifically, we use SigProfilerAssignment to fit known reference signatures to the mutational profiles of the cells [7]. In this setting, the matrix P is fixed and represents a set of established mutational signatures, while the goal is to compute the matrix E, which estimates the exposure of each signature in each sample, defined as the number of mutations attributed to it. This results in a signature-by-sample exposure matrix representing how much each known signature contributes to the observed mutations in a sample. In our analysis, these exposures were afterwards normalized per sample, i.e., the exposures of each sample sum up to one, thus showing the proportion of mutations attributed to a signature.

The function in SigProfilerAssignment to assign mutational signatures is the `cosmic_fit` function. It accepts multiple input types. Since we used mutation calling files, the parameter `input_type` was set to `"vcf"` [13]. The reference signatures used in this study come from the Catalogue Of Somatic Mutations In Cancer (COSMIC), version 3.4, which is the most recent version at the time of writing. These signatures were extracted from large-scale cancer genome sequencing studies using SigProfiler [14]. We used the set of 96 single base substitution (SBS96) signatures [15] and the GRCh38 reference genome build, as our input mutation data consists of single base substitutions and is aligned to the GRCh38 reference genome.

2.3 Simulating Data Loss

To simulate the effects of data loss, i.e. a reduced number of mutations per cell, we performed controlled perturbation experiments. For each level of perturbation, specifically 5%, 10%, 20%, and 40%, the specified percentage of mutations was randomly deleted independently in each cell. These percentages were chosen as a compromise between covering a broad range of data loss and maintaining sufficient resolution at lower levels. The 5% and 10% thresholds allow us to observe more fine-grained effects of mild perturbations, while 20% and 40% simulate more substantial data loss. This range provides a practical way to simulate varying degrees of missingness in single-cell data.

The procedure was as follows:

1. **Baseline fit.** Signature fitting was first performed on the original, unperturbed dataset to serve as a baseline for comparison.
2. **Perturb-and-refit loop.** For each deletion fraction $d \in \{5, 10, 20, 40\}\%$:
   (a) Randomly delete $d\%$ of mutations independently in every cell.
   (b) Refit signature exposures to the perturbed input data.
   (c) Repeat steps (a)–(b) 20 times, each time with a different random seed to ensure stochastic variability. This yields 20 independent replicates for that deletion level.

2.4 Robustness Metrics

Signature Presence in the Dataset

We defined four metrics to evaluate the robustness of the fitted mutational signature exposures. The first metric assesses which signatures are detected in the dataset, i.e., across all cells, irrespective of its fitted exposure magnitude. For every perturbation run, the exposure matrix obtained from SigProfilerAssignment is converted to a binary indicator: an exposure value greater than zero is recoded as 1 (the signature is present in that cell) and 0 otherwise. A signature is considered present in the dataset if at least one cell carries a nonzero exposure. Repeating this procedure for the 20 independent runs at each deletion level yields, for each signature, the fraction of perturbation runs in which that signature is detected. To see which signatures are present in the unperturbed dataset, one can simply inspect the tumor mutational burden plot produced by SigProfilerAssignment. This plot shows exactly the set of signatures that are active in at least one cell. Additionally, this plot shows how many cells a signature is active in. That brings us to the next metric.

Consistency of Signature Presence across Cells

The second metric measures, for every mutational signature, how many single cells actually contain that signature and how

consistently that number is preserved when mutations are randomly deleted. Again, if a signature's exposure in a cell is greater than zero it is marked as active, otherwise inactive. For each signature, we compute the average number of cells in which the signature is active across the 20 runs of each deletion level, along with the standard deviation to capture run-to-run variability. This metric, like the first one, gives a high-level overview of the stability of the assigned signatures, but it definitely has limitations. The metric captures presence of signatures, but not magnitude of exposures. A signature might seem stable while its exposure value fluctuates widely. Conversely, a signature might appear unstable even if its exposure only changes a relatively small amount, e.g., from zero to just a tiny positive exposure. Although these metrics are a convenient starting point, its results should be interpreted alongside magnitude-sensitive measures to avoid over- or underestimating true robustness.

Per-Signature MRE Relative to Original Exposures

The third metric is such a magnitude-sensitive metric. It quantifies how much a signature's exposure deviates from the original exposure value as mutation loss increases. Per signature and per deletion level, we compute for each cell the mean relative error (MRE) between the original exposure and the 20 exposures obtained after the 20 perturbation runs. This MRE is then aggregated across all cells by taking the average and the standard deviation. The average will tell us, for this signature, how much the exposures deviate from the original on average across all cells. The standard deviation will tell us, for this signature, how much that deviation varies between cells. Specifically, for each signature and each cell, the MRE is defined as the average absolute difference between the original exposure and the refitted exposures across the 20 perturbation runs, divided by the original exposure:

$$\mathrm{MRE}_{s,c} = \frac{1}{20} \sum_{i=1}^{20} \frac{|x_i - y|}{y}$$

The error is thus expressed as a proportion of the original exposure, making it independent of the scale of the original exposure. This was preferred over the mean absolute error, which is scale-dependent. However, that does mean that this metric can only be used on cells in which the signature has an original exposure not equal to zero. Because it ignores cells where a signature is initially absent, a complementary measure is still needed to capture how frequently and to what extent new exposures emerge when they were originally zero. Nevertheless, the MRE is still useful as it provides a concise, scale-independent view of how the exposures of a core signature shift in magnitude; specifically in the set of cells in which that signature was originally detected.

Per-Cell Cosine Similarity between Exposure Vectors

The fourth metric shifts the perspective from signatures to cells, summarising stability per cell instead of per signature. For every cell, we treat its exposures as a single vector that captures how its mutations are distributed across signatures. At each deletion level, we take the cell's original exposure vector and compare it with each of the 20 perturbed vectors

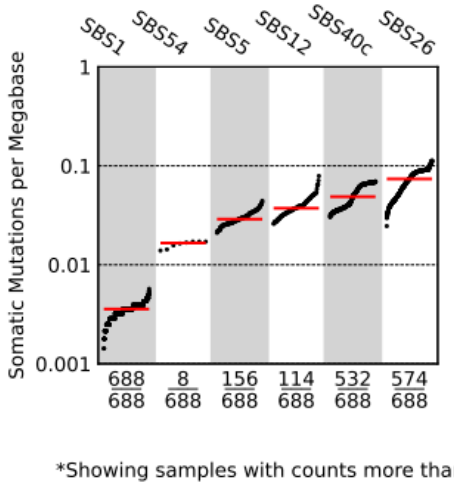

Figure 1: Tumor mutational burden (TMB) plot of the unperturbed dataset. Each column represents a mutational signature. In addition to showing which signatures are active in the dataset overall, it also illustrates how many cells each signature is active in; i.e., in how many cells a given signature has an exposure greater than zero. The y-axis is the somatic mutations per megabase.

using cosine similarity[1]. This gives 20 similarity scores per cell, which we then average to obtain one mean cosine similarity for that cell at that deletion level. Because cosine similarity ranges from 0 (completely different pattern) to 1 (identical pattern) and ignores scaling, a high mean value indicates that the cell's overall signature composition remains stable after mutation loss, whereas a low value shows that its exposure profile is easily reshuffled by perturbation.

3 Results and Discussion

3.1 Signature Presence in the Dataset

Six signatures are detected in the unperturbed dataset (Figure 1): SBS1 (deamination of 5-methylcytosine, clock-like), SBS5 (unknown aetiology, clock-like), SBS12 (unknown aetiology), SBS26 (defective DNA mismatch repair), SBS40c (unknown aetiology), and SBS54 (possible sequencing artefact) [15]. This indicates that the mutational activity in this tumor is likely driven by a limited set of biological processes, with the remaining COSMIC signatures playing no detectable role.

This baseline is now compared to the perturbed datasets to assess how signature presence changes under increasing levels of data loss. The results show that the 6 originally active signatures remain present in all perturbation runs at all deletion levels, including up to 40% mutation loss (Table 1). This demonstrates that SigProfilerAssignment is consistently able to detect the core signatures. This consistent detection suggests that strong biological signal can still be recovered even under substantial data loss.

[1]We also evaluated Jensen–Shannon divergence and L2 distance, but both metrics showed nearly identical trends to cosine similarity (Pearson r $>$0.9), so only cosine results are presented.

Deletion level:	5%	10%	20%	40%
SBS1	100%	100%	100%	100%
SBS5	100%	100%	100%	100%
SBS12	100%	100%	100%	100%
SBS26	100%	100%	100%	100%
SBS40c	100%	100%	100%	100%
SBS54	100%	100%	100%	100%
SBS87	35%	45%	100%	100%
SBS93	5%	10%	50%	95%
SBS37	15%	20%	50%	95%
SBS17a			30%	100%
SBS51			10%	40%
SBS21			5%	55%
SBS57			15%	90%
SBS19			5%	70%
SBS31				10%
SBS7d				50%
SBS23				15%
SBS33				15%
SBS32				15%
SBS88				20%
SBS7a				5%
SBS11				5%
SBS92				5%
SBS7b				5%

Table 1: Fraction of perturbation runs in which each signature is detected at the indicated deletion levels. Each row corresponds to a signature and the columns represent the different deletion levels. Each value shows the percentage of runs in which that signature is active in the dataset; i.e., it has a non-zero exposure in at least one cell. Signatures active in the original dataset are shaded green. Signatures that are never detected are excluded from the table for clarity.

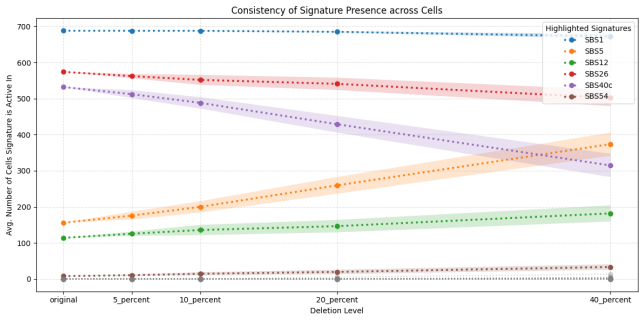

Figure 2: For each of the six originally active signatures, the average number of cells in which the signature is active across the 20 perturbation runs of each deletion level, with standard deviation across runs.

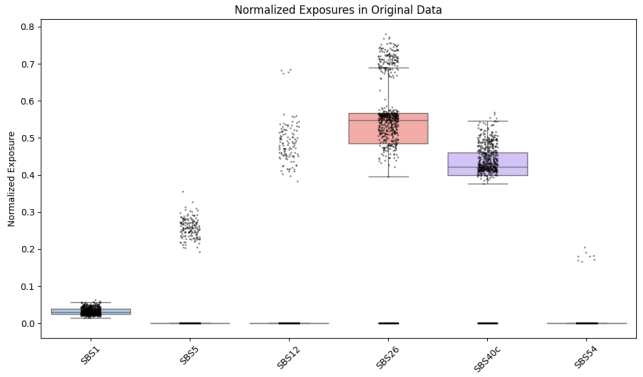

Figure 3: The distribution of normalized exposures in the original data, per signature.

However, as the level of mutation deletion increases, a growing number of additional signatures begin to appear in the fitted exposures. At 5% and 10% deletion, only 3 extra signatures emerge. At 20% deletion, this number rises to 8, and at 40%, a total of 18 non-original signatures are detected in the dataset. Underlying this pattern is the way SigProfilerAssignment minimises reconstruction error. In simple terms, the algorithm iteratively selects which signatures to include with fixed relative error thresholds. A signature is dropped if removing it raises the relative error by less than 0.01, and a signature is re-added if including it lowers the error by more than 0.05 [7]. In a sparse catalogue each remaining mutation accounts for a disproportionately large share of the 96-context profile, so deleting a signature that explains even a single mutation can breach the 1% drop limit, while adding a new signature to explain that mutation easily clears the 5 % add threshold. As mutation counts fall, the algorithm will become more eager to keep and add a signature to minimize the error, even if it contributes only a handful of mutations. From a statistical standpoint, this is similar to fitting a complex model with many parameters to a very small dataset, many different models could explain the data, and without a strict penalty for complexity, the optimizer chooses a complex model that almost perfectly explains the few data points. The consequence is overfitting: the additional signatures improve the fit on the given sample but are likely modeling sampling noise rather than real signal.

Additionally, several of the extra signatures have mutational profiles that are moderately similar to one of the six original signatures, e.g., SBS87 shares a cosine similarity of 0.75 with SBS1, while SBS37 is close to both SBS12 (0.82) and SBS26 (0.77). When mutations are removed, distinguishing contexts can disappear, allowing the model to swap a few remaining mutations onto another, similar signature while still improving the residual.

Future analyses could cross-check the biological plausibility of any newly fitted signature, e.g., SBS87 is known to be caused by thiopurine chemotherapy, so its presence in samples with no history of that treatment would likely indicate a false positive assignment.

3.2 Consistency of Signature Presence across Cells

In the original data, the six identified signatures are not necessarily active in all cells, e.g., SBS26 is active in 574 cells, while SBS54 is active in only 8 cells (Figure 1). We were interested to see whether these numbers remain similar to the original or if they change drastically when mutations are deleted. The obtained results show that this depends on the

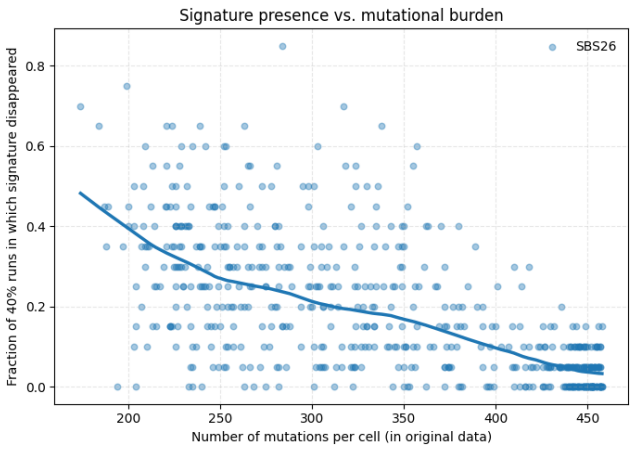

Figure 4: The fraction of 40% perturbation runs in which SBS26 disappeared plotted against mutational burden. Every point represents one cell that originally contained SBS26. The x-axis measures the total number of mutations that cell carried in the unperturbed data. The y-axis shows how often SBS26 vanished at the 40% deletion level (as a fraction of the 20 perturbation runs). The LOWESS (locally weighted scatterplot smoothing) curve is also shown.

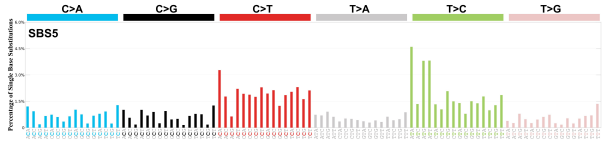

Figure 5: Mutational profile of SBS5 [16]

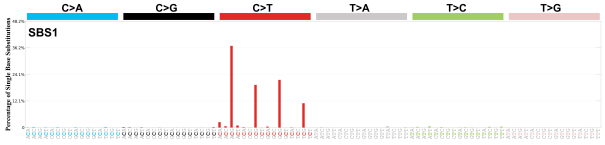

Figure 6: Mutational profile of SBS1 [17]

signature (Figure 2).

For example, SBS1 remains remarkably consistent across all deletion levels. The number of cells in which it is detected falls from 688 in the unperturbed dataset to an average of 672 after the 40% mutation dropout, which is only a 2% decrease. This may be explained by the fact that SBS1 has a relatively distinct mutational profile, especially compared to the other five original signatures, which all have a cosine similarity with SBS1 of less than 0.19. Perhaps this makes the signature less likely to be confused or replaced during fitting. One might expect signatures whose exposures are already near zero to be the first to drop below detection after downsampling; yet SBS1 actually has the lowest non-zero exposure levels among the six signatures (Figure 3), suggesting that profile distinctiveness may outweigh exposure magnitude in determining robustness.

In contrast, as data loss increases, SBS26 is detected in fewer cells, while SBS12 is detected in more. Interestingly, these two signatures share a high cosine similarity of 0.93, indicating that mutations that were originally assigned to SBS26 might be assigned to SBS12 instead because of limited context. Note that in the original fit every cell carries either SBS26 or SBS12, never both. We found that 93% of cells where SBS26 disappears in the 40% dropout level show positive exposure for SBS12 in every run where SBS26 disappears. A very similar trend can be seen for SBS40c and SBS5: the presence of SBS40c decreases while SBS5 increases; they share a cosine similarity of 0.91; every cell originally carries either one or the other; and 98% of cells where SBS40c disappears in the 40% dropout level show positive exposure for SBS5 in every run where SBS40c disappears. These observations further suggest that the model may confuse highly similar signatures under data loss. Future work could confirm this "signature swapping" by tracking the exact mutations that were assigned to SBS26 or SBS40c in the original fit and verifying whether those same mutations migrate to SBS12 or SBS5 after downsampling.

There could be several other factors that influence the consistency of signature presence across cells. A first possibility is the number of mutations per cell. For example, at the 40% deletion level, the exposure of SBS26 tends to drop to zero in cells with lower mutation counts more often than in cells with higher mutation counts (Figure 4). This suggests that low mutation counts make it harder to distinguish SBS26 from SBS12. A second possible factor is the flatness of a signature's mutational profile. Consider for instance SBS5, its profile is a relatively even distribution over all 96 mutation types (Figure 5) and is therefore a flatter signature than SBS1 for instance, whose profile has most of its signal in the C>T context (Figure 6). The hypothesis here is that flatter signatures are more sensitive to data loss because they lose the fine balance that lets the model choose one flat profile over another, while peaked signatures keep their diagnostic spike. This is consistent with recent benchmarking and methodological studies showing that flat signatures such as SBS5 and SBS40c are systematically harder to recover, especially when mutation counts are low or when other similar flat signatures are present, whereas sharply peaked signatures are fit more reliably [18],[19]. Future analyses could quantify signature flatness and its effects on the consistency of signature detection.

3.3 Per-Signature MRE Relative to Original Exposures

As one might expect, the mean relative error increases steadily as more mutations are deleted. Exposures deviate only modestly after a 5% loss but grow progressively larger at 10%, 20%, and reach their highest levels after 40% deletion (Figure 7). This shows that the higher the data loss, the more exposures deviate from their original values. Although the standard deviation also rises as mutations are removed, it increases more slowly than the mean relative error. Consequently, the relative spread (SD / mean) is highest at the 5% and 10% deletions, implying that, when only a small fraction of data is missing, a signature's exposure generally remains close to the original but drifts more in certain cells. By the time 20% and 40% of mutations are lost, the average error is large but the SD accounts for a smaller proportion of that mean, showing that deviations are now more uniformly high

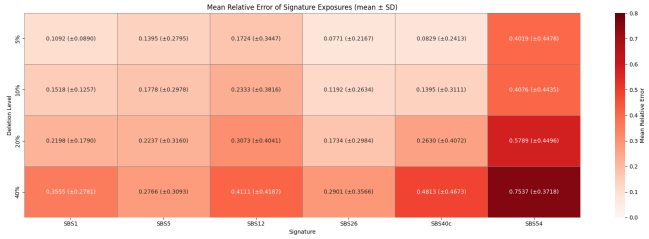

Figure 7: Heatmap of mean relative error. The rows are deletion levels and the columns are signatures. Per cell, the MRE is computed across 20 perturbation runs. The values in the heatmap show the average and standard deviation across all cells.

Figure 8: Heatmap of mean relative error. SBS5 is merged with SBS40c, and SBS12 is merged with SBS26.

across the cell population. In other words, lower data loss keeps exposures closer to their original values on average, but this varies more from cell to cell; higher data loss pushes exposures further away on average, and this is a more consistent shift across all cells.

Two more key points can be observed. First, the exposure of SBS54, detected in only eight cells and suspected to be a sequencing artefact, is noticeably less stable than that of the other signatures. Second, when highly similar signatures are merged and thus considered as one and the same signature (e.g., SBS5 + 40c and SBS12 + 26), their mean relative error drops by an average of 58% across all deletion levels (Figure 8), indicating that these signatures were often interchanged during fitting, just like we hypothesized previously.

3.4 Per-Cell Cosine Similarity between Exposure Vectors

Overall, exposure vectors of cells diverge progressively from the originals as data loss rises, yet the effect is uneven, some cells already show low similarity at 5% deletion, whereas others remain highly consistent even after 40% of their mutations are removed (Figure 9). Several factors can influence why some cells are more stable than others. One possible factor is the number of mutations a cell contains. When we plot mean cosine similarity against mutation count (Figure 10), we see that cells with more mutations tend to preserve their exposure vectors better. In fact, Spearman correlations between mutational burden and mean cosine similarity are positive and highly significant at every deletion level ($\rho \approx 0.38$–0.59 and $p \ll 10^{-24}$), indicating such a monotonic relationship. In other words, the absolute number of mutations matters. Losing 40% of 500 mutations still leaves plenty of signal to recover a similar exposure vector, whereas losing 40% of just a

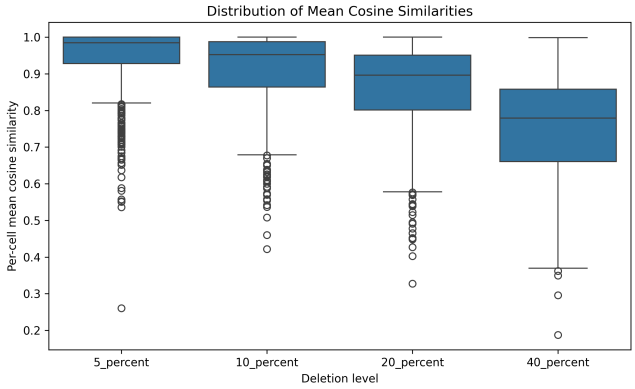

Figure 9: For each cell, the average cosine similarity between its original exposure vector and the 20 perturbed exposure vectors is computed. Per deletion level, the distribution of these mean cosine similarities is shown, with every data point representing a cell.

hundred mutations leaves very little data points, causing the exposures to drift.

Another possible factor is the Shannon entropy of a cell's original exposure vector, which measures how evenly its mutations are distributed across signatures. An entropy of 0 bits would mean every mutation is assigned to a single signature, whereas higher values reflect a flatter mix in which many signatures contribute smaller, similar fractions. Analysis shows that cells with higher-entropy exposure vectors tend to show lower mean cosine similarity after perturbation, and this effect strengthens as more data is removed. Spearman correlations support this observation, indicating a significant negative relationship, especially at higher deletion levels ($\rho \approx -0.27$ at 5% deletion and $\rho \approx -0.53$ at 40% deletion; $p \ll 10^{-12}$). If a single signature accounts for the bulk of mutations in a cell, that dominant process remains obvious even if some mutations are removed. In contrast, if a cell's mutations are divided among numerous signatures with small contributions, small perturbations can lead to disproportionate shifts: one signature might drop out or another might become relatively more prominent. As the deletion fraction rises, the vulnerability of high-entropy cells grows, explaining why the correlation becomes steadily more negative at 20% and 40% loss.

4 Responsible Research

4.1 Reproducibility

All elements required to replicate this study are specified in detail in the Methodology section (Section 2). The approach of variant calling is described by Liu et al. [8] and the GATK workflow for filtering is well-documented[2], but the exact pipeline is not available as this was done by my supervisors. However, the resulting per-cell VCF files, along with all scripts for preprocessing, signature fitting, data-loss perturbations and metric computations, are available in this project's

[2]GATK tutorial is available at gatk.broadinstitute.org

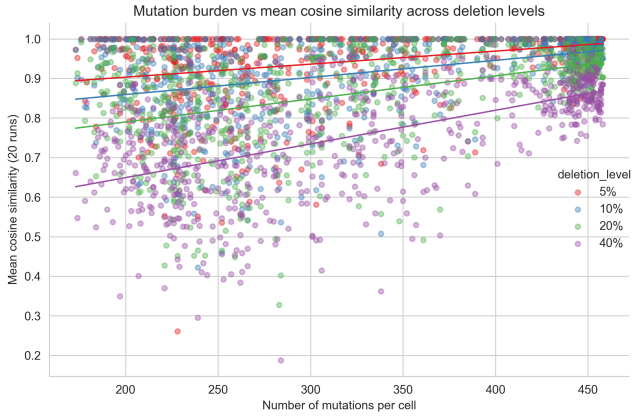

Figure 10: Relationship between number of mutations per cell and mean cosine similarity between exposure vectors. Each point represents one cell, plotting its number of mutations in the unperturbed dataset against the mean cosine similarity between its original and perturbed exposure vectors. Colours indicate the four deletion levels, and the corresponding ordinary-least-squares (OLS) linear regression lines highlight the correlation.

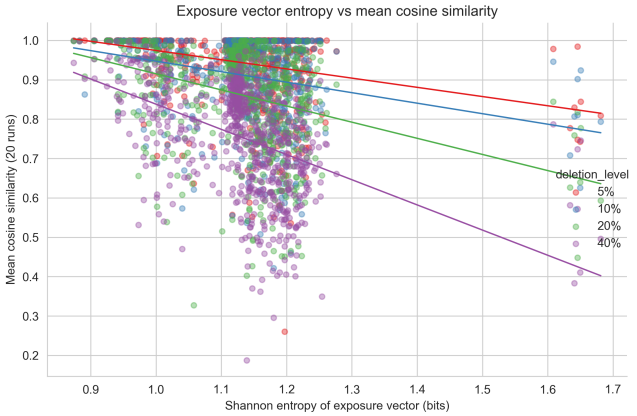

Figure 11: Relationship between entropy of the original exposure vector and mean cosine similarity between exposure vectors. Each point represents one cell, plotting its original exposure vector entropy against the mean cosine similarity between its original and perturbed exposure vectors. Colours indicate the four deletion levels, and the corresponding ordinary-least-squares (OLS) linear regression lines highlight the correlation.

GitLab repository[3] which is accessible to users authorized by TU Delft. Signature fitting uses the open-source tool SigProfilerAssignment[4] which internally provides the reference set of COSMIC v3.4 SBS96 signatures. The same signatures were also downloaded in numerical form to compute the pairwise cosine similarities between their mutational profiles[5]. Perturbation experiments were repeated twenty times per deletion level with random seeds uniformly drawn from the range 10,000 to 99,999. The complete seed log is included in the GitLab repository, ensuring that every stochastic decision can be replicated exactly.

4.2 Ethical Considerations

The dataset of raw scRNA-Seq reads is publicly available. Discovery Life Sciences [20] first procured the breast cancer cells from the 65-year-old female donor; those specimens were then supplied to the biotechnology company 10x Genomics, which performed the single-cell RNA-sequencing and released the resulting dataset[6]. In its public ethics statement, Discovery Life Sciences affirms that it "is committed to quality and integrity with CLIA-certified labs, stringent IRB and Ethics Committee compliance, and all of the applicable regulations, guidelines and best practices that meet or exceed the U.S. and international regulatory requirements" [21]. The company further emphasises strict adherence to personal data protection regulations, including the GDPR, and describes its commitment to respect and protect the privacy of individuals. These statements indicate that the biological material used in this study was collected through procedures that meet ethical and legal standards.

5 Conclusions and Future Work

This study investigates how progressively removing mutations affects the robustness of fitted mutational signature exposures and highlights several factors that possibly influence this robustness. After removing up to 40% of mutations SigProfilerAssignment still recovered the six signatures found in the original data, indicating that key biological signals can survive the sparsity typical of single-cell data. At the same time, the number of additional signatures assigned rises sharply with data loss, reflecting a growing risk of overfitting. Additionally, exposure seems to shift from one signature to another very similar signature, suggesting that distinguishing similar signatures is difficult with little data available. With only 5% or 10% loss, exposures generally remain close to their original values, although this depends on the cell. When 20% or 40% of mutations are removed, the exposure distributions of all cells consistently move further away from the original distribution. Two cell-level features can explain this behaviour. First, cells with more mutations were more stable because the fitting algorithm has more information to work with. Second, cells dominated by one or two signatures were more stable than cells whose mutations were spread evenly across many signatures.

Because the study is limited to a single tumor, employs one fitting tool and uses uniform random dropout, these findings should not be blindly generalised. A further limitation stems from the reference catalogue itself: COSMIC v3.4 SBS96 signatures were extracted de novo from bulk genomes across many cancer types, a context that differs from single-cell data. As such, the COSMIC library may contain signatures that do not accurately reflect the mutational processes active in individual cells, or may lack signatures that do. Re-

[3]Repository is available at https://gitlab.ewi.tudelft.nl/goncalveslab/bachelor-projects/bsc-rp-2425-rebecca-nys

[4]SigProfilerAssignment is available at https://github.com/AlexandrovLab/SigProfilerAssignment

[5]COSMIC signatures are available at cancer.sanger.ac.uk

[6]Raw scRNA-Seq reads are available at 10xgenomics.com

fining the reference set by extracting de novo signatures directly from single-cell catalogues and comparing them with COSMIC could reveal missing or mismatched profiles and ultimately provide a more accurate and robust framework for understanding mutational processes at the cellular level.

At least five lines of future work can follow this study: (i) repeat the perturbation experiments on tumors of different types and mutational burdens; (ii) track the consistency of signature exposures across perturbation runs; (iii) investigate how SigProfilerAssignment's reconstruction error changes with mutation loss; (iv) delete mutations from the input data in a biased way (e.g. certain chromosomes) or simulate noise and (v) carry out biological and clinical validation by checking that the single-cell signatures match those seen in bulk sequencing of the same tumor and by testing whether cells predicted to harbour, e.g., a DNA-repair defect actually behave accordingly. Addressing these points will strengthen the analysis of the robustness of fitted mutational signature exposures in single-cell data.

References

[1] National Human Genome Research Institute, "Mutation - genetics glossary." https://www.genome.gov/genetics-glossary/Mutation, 2025. Accessed: May 23, 2025.

[2] N. McGranahan and C. Swanton, "Clonal heterogeneity and tumor evolution: Past, present, and the future," *Cell*, vol. 168, pp. 613–628, Feb. 2017.

[3] M. Greaves, "Evolutionary determinants of cancer," *Cancer discovery*, vol. 5, no. 8, pp. 806–820, 2015.

[4] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton, "Deciphering signatures of mutational processes operative in human cancer," *Cell Reports*, vol. 3, pp. 246–259, Jan. 2013.

[5] J. G. Tate *et al.*, "Cosmic: the catalogue of somatic mutations in cancer," *Nucleic Acids Research*, vol. 47, pp. D941–D947, Jan. 2019.

[6] S. Islam, M. Díaz-Gay, Y. Wu, M. Barnes, R. Vangara, E. Bergstrom, Y. He, M. Vella, J. Wang, J. Teague, *et al.*, "Uncovering novel mutational signatures by de novo extraction with sigprofilerextractor," *Cell genomics*, vol. 2, no. 11, 2022.

[7] M. Díaz-Gay, R. Vangara, M. Barnes, X. Wang, S. Islam, I. Vermes, S. Duke, N. Narasimman, T. Yang, Z. Jiang, S. Moody, S. Senkin, P. Brennan, M. Stratton, and L. Alexandrov, "Assigning mutational signatures to individual samples and individual somatic mutations with sigprofilerassignment," *Bioinformatics*, vol. 39, p. btad756, 12 2023.

[8] X. Liu, J. I. Griffiths, I. Bishara, J. Liu, A. H. Bild, and J. T. Chang, "Phylogenetic inference from single-cell rna-seq data," *Scientific Reports*, vol. 13, p. 12854, Aug. 2023.

[9] J. Gonçalves, "Deciphering cancer heterogeneity with machine learning." https://projectforum.tudelft.nl/course_editions/118/generic_projects/5929, 2025. Accessed: April 22, 2025.

[10] 10x Genomics, "750 sorted cells from human invasive ductal carcinoma, single cell gene expression dataset by cell ranger v6.0.0." Mar. 2021.

[11] Broad Institute, "How to Filter variants either with VQSR or by hard filtering," 2023. Accessed: 2025-05-20.

[12] L. B. Alexandrov *et al.*, "Signatures of mutational processes in human cancer," *Nature*, vol. 500, no. 7463, pp. 415–421, 2013.

[13] M. Díaz-Gay, R. Vangara, M. Barnes, X. Wang, S. M. A. Islam, I. Vermes, N. B. Narasimman, T. Yang, Z. Jiang, S. Moody, S. Senkin, P. Brennan, M. R. Stratton, and L. B. Alexandrov, "Sigprofilerassignment wiki — osf." https://osf.io/mz79v/wiki/home/, 2023. Accessed: Jun. 17, 2025.

[14] K. J. H. N. Alexandrov, L.B. *et al.*, "The repertoire of mutational signatures in human cancer," *Nature*, vol. 578, p. 94–101, 2020.

[15] COSMIC, "Single base substitution (sbs) signatures." https://cancer.sanger.ac.uk/signatures/sbs/, 2023. Accessed: 2025-05-23.

[16] "Single base substitution signatures – sbs5," 2024. Accessed: Jun. 21, 2025.

[17] "Single base substitution signatures – sbs1," 2024. Accessed: Jun. 21, 2025.

[18] A. Medo, J. Smith, Q. Wang, *et al.*, "Comprehensive benchmarking of mutational-signature attribution tools reveals systematic challenges with flat signatures," *Nature Communications*, vol. 15, p. 1234, 2024.

[19] L. Jin, S. Gupta, E. Lee, *et al.*, "Musical improves mutational signature extraction and resolves ambiguities in flat signatures," *Nature Genetics*, vol. 56, pp. 789–801, 2024.

[20] Discovery Life Sciences, "Discovery life sciences—biospecimen procurement and analytical services," 2025. Accessed: 20-Jun-2025.

[21] Discovery Life Sciences, "Faq: Are your biospecimens ethically obtained?," 2025. Accessed: 20-Jun-2025.