THE EFFECT OF COSINE SIMILARITY ON SBS SIGNATURE EXTRACTION AND SBS SIGNATURE VALIDATION VIA THE LINEAR COMBINATION MODEL

Jonathan Markey

1805363

August 18 2023

Project Supervisor: Lewis Mitchell

In submitting this work I am indicating that I have read the University's Academic Integrity Policy. I declare that all material in this assessment is my own work except where there is clear acknowledgement and reference to the work of others.

I give permission for this work to be reproduced and submitted to other academic staff for educational purposes.

I give permission for this work to be reproduced and provided to future students as an exemplar report.

Abstract

This project aims to address two research questions pertaining to the cosine similarity of SBS signatures within human cancer genome samples. The first question (RQ1) explores the effect of cosine similarity on the results of the current framework, while the second question (RQ2) investigates the potential enhancement of results through a linear combination model. This project is important as the current NMF framework serves as a cornerstone for extracting signatures in human cancer genomes, directly impacting clinicians and medical researchers.

To answer these questions, a set of samples derived from SBS signatures of varying cosines similarity was analysed using the existing framework’s evaluation metrics and subsequently subjected to a linear combination model. Results revealed that the current framework was only able to reliably extract SBS signatures with a derived cosine similarity up to 0.763. A cosine similarity that exceeded this level resulted in incorrect SBS signatures being extracted. The linear combination model demonstrated the potential to increase the certainty of the results of the current framework in addition to providing a supplementary analysis of the intra-cluster distances of the extracted SBS signatures.

Despite limitations, this research advances the understanding of cosine similarity, SBS signatures, and the current framework's efficacy. Findings highlight framework limitations and refinement potential, encouraging future investigations to expand these results.

Introduction

Genetic mutations are pivotal in genetics and underlie disease mechanisms and genetic development [1]. Mutations, including substitutions, insertions, deletions, and repeat expansions, arise during cell division due to factors like UV light, age, and chemicals like those found in cigarettes and asbestos. Mutations can be inherited (germline) or occur during an organism's lifespan (somatic). Single-base substitutions (SBS) are a type of somatic mutation and are the focus of this research paper. In this type of mutation, six types of substitutions can occur; C>A, C>G, C>T, T>A, T>C and T>G. Within each type, there are 16 mutational contexts, totalling 96 unique context-dependent mutational types. Analysing DNA's mutational profile across genomes reveals patterns known as SBS signatures (Figure 1) and from these signatures, it is also possible to identify the factors that have caused that signature to arise [2][3][4].

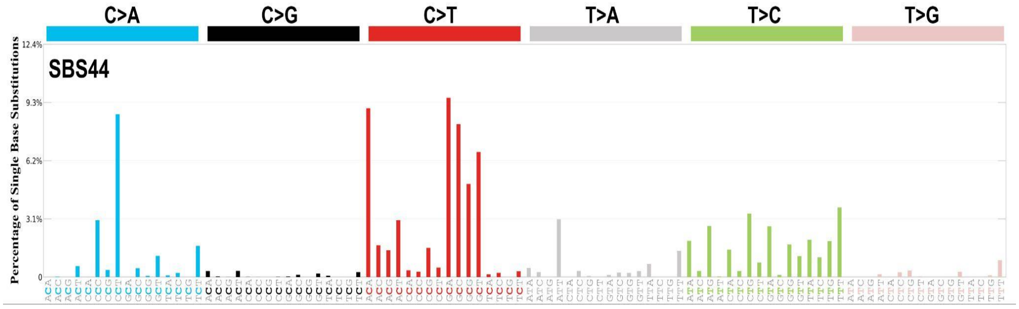


Figure 1: Mutational Profile of SBS 44 [5]

The foundation for deciphering SBS signatures present in cancer genomes was proposed by Alexandrov et al. (2013) [6]. The framework involves utilising processes such as dimensionality reduction, Monte Carlo bootstrap resampling, non-negative matrix factorization, K-means clustering, and several evaluation metrics to extract signatures from human genome samples. The purpose of this research project is to address the research questions:

- **RQ1**: How does the cosine similarity of SBS signatures in human cancer genome samples relate to the outcomes produced by the existing SBS signature extraction framework?
- **RQ2**: Can a linear combination model enhance the certainty of the results obtained from the current SBS signature extraction framework?

This research paper aims to analyse how samples derived from known SBS signatures with high cosine similarity affect the current framework due to its reliance on the silhouette width evaluation metric. The silhouette width is sensitive to cosine similarity as one of the two variables that are used to calculate silhouette width, inter-cluster distance, is equal to the cosine distance between the extracted signatures. Theoretically, if a human genome sample contains SBS signatures that are very similar to each other then the silhouette width would be lower, potentially causing the evaluation metric thresholds to not be met, which would cause the framework to fail.

The linear combination model is a supervised machine learning approach that uses minimisation to extract the coefficient values for known SBS signatures present in a genome sample. The linear combination model can potentially enhance the results of the current signature extraction framework by firstly increasing the certainty in the results extracted and secondly by serving as an evaluation mechanism for evaluating the variability in the signatures extracted.

This research is important as it is vital for clinicians and researchers to reliably and accurately identify SBS signatures present in cancers, allowing them to refine diagnostic tools, optimise treatment methods, and develop innovative therapies which drive advancements in cancer care and prevention.

Background

Academic Framework

The foundation for deciphering SBS signatures present in cancer genomes was proposed by Alexandrov, L.B. et al (2013) and is utilised in the program SigProfilerExtracted [6]. The framework utilises non-negative matrix factorisation (NMF) for de novo SBS signature extraction. Despite subsequent research into improving the model [7], in addition to explorations of alternative methods utilising linear regression [8], quadratic programming [9] and linear combination decomposition [10], the original NMF framework still remains the academic standard for deciphering SBS signatures. The framework has been used to identify many of the now-known SBS signatures including those in the current catalogue of known SBS signatures in the Catalogue Of Somatic Mutations In Cancer (COSMIC) database. This database primarily consists of SBS signatures extracted via the NMF framework [11]. The framework has also been used in multiple real-world scenarios, including in de Kanter, J.K. et al. (2021) to uncover potential SBS signatures arising from stem cell applications in regenerative medicine [12]. More recently the framework was used by Thatikonda, V. et al (2023) to uncover SBS signatures across 27 paediatric cancers [13].

The NMF framework is the process of decomposing a non-negative matrix into two non-negative matrices, such that their product approximates the original matrix (equation 1).

$$M \approx P \times E \tag{1}$$

$$M = \begin{pmatrix} m_1^1 & m_2^1 & \dots & m_G^1 \\ \vdots & \vdots & & \vdots \\ m_1^K & m_2^K & \dots & m_G^K \end{pmatrix},\ P = \begin{pmatrix} p_1^1 & p_2^1 & \dots & p_N^1 \\ \vdots & \vdots & & \vdots \\ p_1^K & p_2^K & \dots & p_N^K \end{pmatrix},\ E = \begin{pmatrix} e_1^1 & e_2^1 & \dots & e_G^1 \\ \vdots & \vdots & & \vdots \\ e_1^N & e_2^N & \dots & e_G^N \end{pmatrix}.$$

Figure 2: NMF Matrices M, P and E

M represents a $K \times G$ matrix of sample human cancer genomes (Figure 2). Each genome sample g in matrix M is defined as a vector $(m_g^1, \dots, m_g^K)^T$ where m_g^k is the number of mutations of type *k*. *K* equalling the 96 context-dependent mutational types. P represents a $K \times N$ matrix of extracted signatures. Each mutational signature n in P is defined as a vector $(p_n^1, \dots, p_n^K)^T$ where p_n^k represents the probability of a mutation type k occurring. *N* is equal to the total number of signatures that are extracted. Note that for each signature n, $\sum_{k=1}^{K} p_n^k = 1$ where $0 \le p_n^k \le 1$. E represents a $N \times G$ matrix of signature exposures (coefficients).

The NMF framework is a six-step process involving dimensionality reduction, Monte Carlo bootstrap resampling, NMF, iteration, K-means clustering and evaluation.

Dimensionality reduction is the first step in the process and is used to speed up the overall framework and reduce the potential impact of noise. This is achieved by reducing the 96 context-dependent mutational types in *M* by removing those which account for less than 1% of total mutations. Monte Carlo bootstrap resampling and iteration are used to increase the accuracy and reproducibility of the results by allowing the framework to analyse a much larger sample of generated synthetic data. This process functions by creating a derivation of matrix *M* by subjecting each genome in the matrix to resampling with replacement. This results in a new matrix *M* with genomes that differ slightly from the original genomes. Combined with iteration, a set of matrices S_M can be generated which overall should contain enough variability to be representative of the cancer genome being analysed. *M* is required to be bootstrap resampled 400 - 1000 times [6] to satisfy reproducibility requirements.

The set of matrices S_M can then be parsed through the NMF process to get two sets of matrices S_P and S_E. NMF achieves the decomposition by optimising the values in *P* and *E* to minimise the cost function, the Frobenius norm (equation 2). As NMF is not perfect, there is often an error term associated with the process which is equal to $M - P \times E$. The Frobenius norm measures the magnitude of this error by finding the square root of the sum of the squares of each element a_{gk} in the error matrix A.

$$||A||_F^2 = \sqrt{\sum_{g=1}^{G} \sum_{k=1}^{K} |a_{gk}|^2} \tag{2}$$

All the signatures in S_P can be clustered with K-means clustering. K-means clustering is a fairly simple machine-learning algorithm in which a data set is partitioned into K distinct clusters. K-means clustering functions by randomly assigning K cluster centroids to the signatures. Then, each signature is assigned to the nearest cluster based on cosine similarity. Cosine similarity measures similarity by calculating the cosine of the angle between the vectors. As all the signatures in the data are non-negative, the cosine similarity only ranges between 0 to 1, with similarity increasing closer to 1. The equation for cosine similarity is seen in equation (3) and is equal to the product of two vectors, divided by the product of the norms of the vectors.

$$Cos\,(A, B) = \frac{A \bullet B}{||A|| \bullet ||B||} \tag{3}$$

Once all signatures are assigned to the cluster centroids, new centroids are calculated to equal the mean of all the assigned signatures. This process is iteratively redone until there is no significant change in the centroids. The final cluster centroids are the extracted SBS signatures.

The evaluation metrics used to analyse the extracted signatures are silhouette width, Frobenius reconstruction error (FRE) and cosine similarity. Silhouette width is the primary evaluation metric used in the framework and is used to measure the overall reproducibility of the framework. The equation for silhouette width (equation 4) is equal to the difference between the intra-cluster distance a and the inter-cluster distance b, compared to the maximum of these distances. This is calculated on a per-cluster basis.

$$\text{Silhouette Width} = \frac{(b-a)}{max(a,b)} \tag{4}$$

The intra-cluster distance is equal to the cosine distance between a signature and all other signatures in the same cluster. The inter-cluster distance is equal to the cosine distance between a signature and the signatures in any other cluster. The distance is measured from -1 to 1 and measures how tightly clustered the signatures are around the centroid. Values closer to 1 indicate that a signature is distinct from the signatures in other clusters, while values closer to -1 indicate that a signature is more similar to the data points in other clusters than its own. For the framework to pass, the mean silhouette width is required to be greater than 0.8. In addition, the mean minimum silhouette width of any given cluster must be greater than 0.2. Values that are less than these will result in the extracted signatures being rejected [6][14].

FRE is equal to the Frobenius norm calculated during NMF. There is no definitive pass/fail metric associated with the FRE as the metric can be affected by factors such as noise in the original data. Instead, FRE is used to evaluate the accuracy with which the signatures were deciphered. The FRE percentage can also be measured and is equal to the Frobenius norm of the error matrix *A* compared to the Frobenius norm of the sample matrix *M.* For both metrics, values closer to zero indicate increased accuracy.

For cosine similarity, the extracted signatures are identified by calculating the similarity between the extracted signatures and the known SBS signatures. Signatures are identified when the cosine similarity is greater than 0.8, with values closer to 1 increasing the certainty [9][14].

Data

There are two data sets used for this project. The first is the known SBS signature data. This data is sourced from the COSMIC database [5][11]. The dataset contains 79 known SBS signatures, each of which is a 96-length vector, where each value corresponds to the mutational probability of a context-dependent mutational type. The second dataset is the human genome data. This dataset is similar to the known data with the exception that each element represents the number of mutations. For this project, five sets of synthetic human genome data have been generated to represent different combinations of signatures at different levels of cosine similarity. These sets include:

- Sample 1: SBS 6, SBS 12 and SBS 38 with a cosine similarity of 0.051.
- Sample 2: SBS 25, SBS 10c and SBS 6 with a cosine similarity of 0.279.
- Sample 3: SBS 30, SBS 44 and SBS 42 with a cosine similarity of 0.526.
- Sample 4: SBS 4, SBS 8 and SBS 29 with a cosine similarity of 0.763.
- Sample 5: SBS 56, SBS 10d and SBS 36 with a cosine similarity of 0.914.

For each of these combinations, the contribution of each signature is randomly distributed based on random weights between (5, 20). In addition, a 10% random noise is added to better simulate real-world data. One hundred genomes are generated per sample with each genome containing approximately 1100 mutations.

Research Justification

The main justification for this research can be seen when analysing the cosine similarity between SBS signatures and the effect that the metric has on the silhouette width evaluation metric.

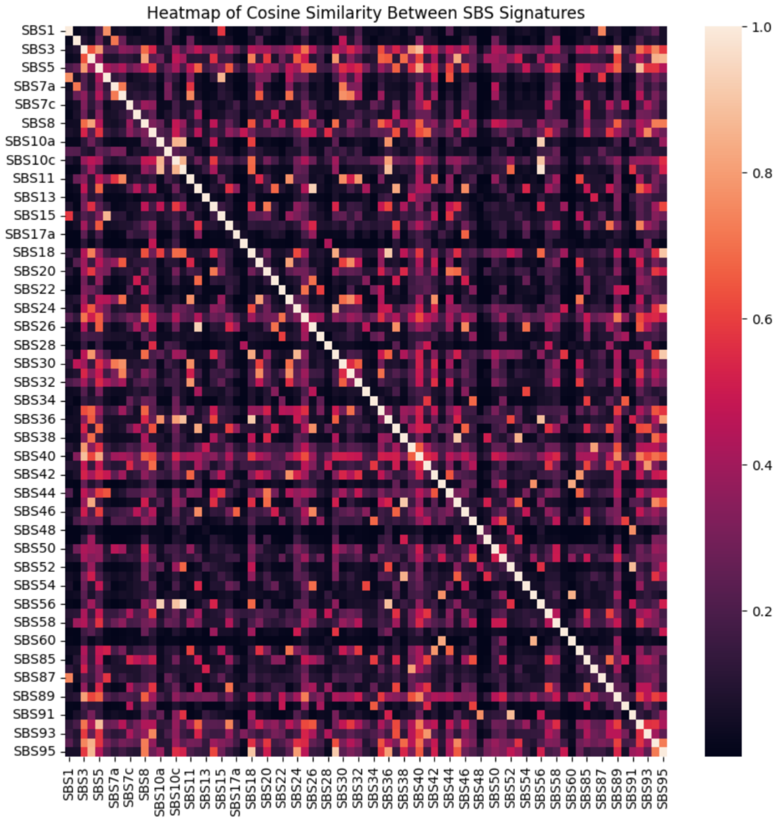


Figure 2: Heatmap of Cosine Similarity Between Known SBS Signatures

Analysing the cosine similarity heatmap for the known SBS signatures (Figure 2) shows that there is cosine similarity between all signatures. The mean similarity between signatures is 0.204, with SBS 40 having the highest mean of 0.410 and SBS 48 having the minimum mean of 0.058. The cosine similarity/distance is a potential problem as these metrics significantly influence NMF, particularly K-means clustering and silhouette width.

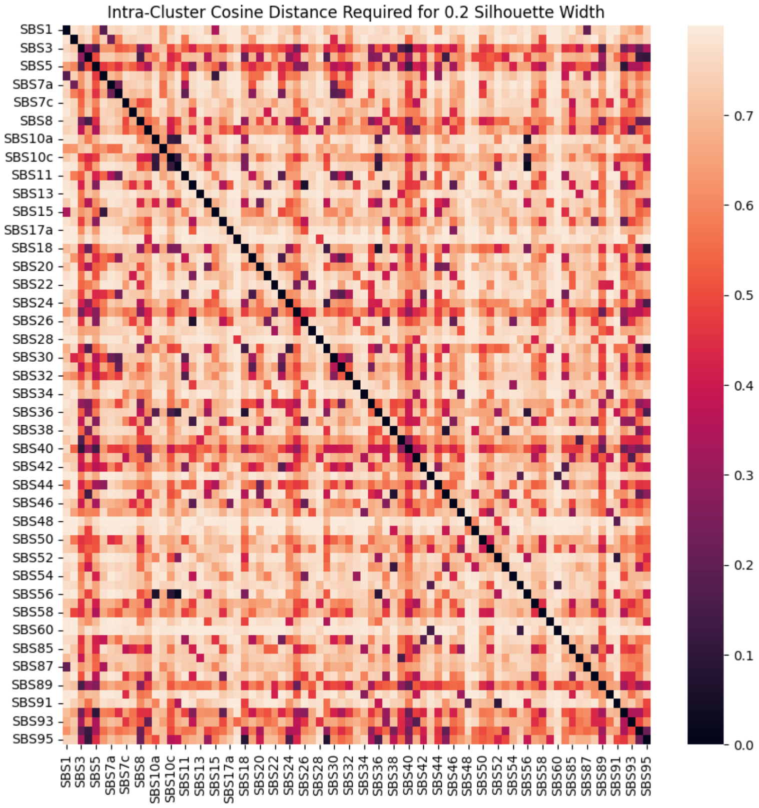


Figure 3: Intra-cluster Cosine Distance Required for Silhouette Width of 0.2

The intra-cluster distance required to equal a silhouette width of at least 0.2 (Figure 3) is notably high. Out of the 3081 unique combinations of signatures, there is only one combination of signatures, SBS 10d and 56, with a low tolerance of less than 0.05. 12 combinations have a moderate distance tolerance of less than 0.10.

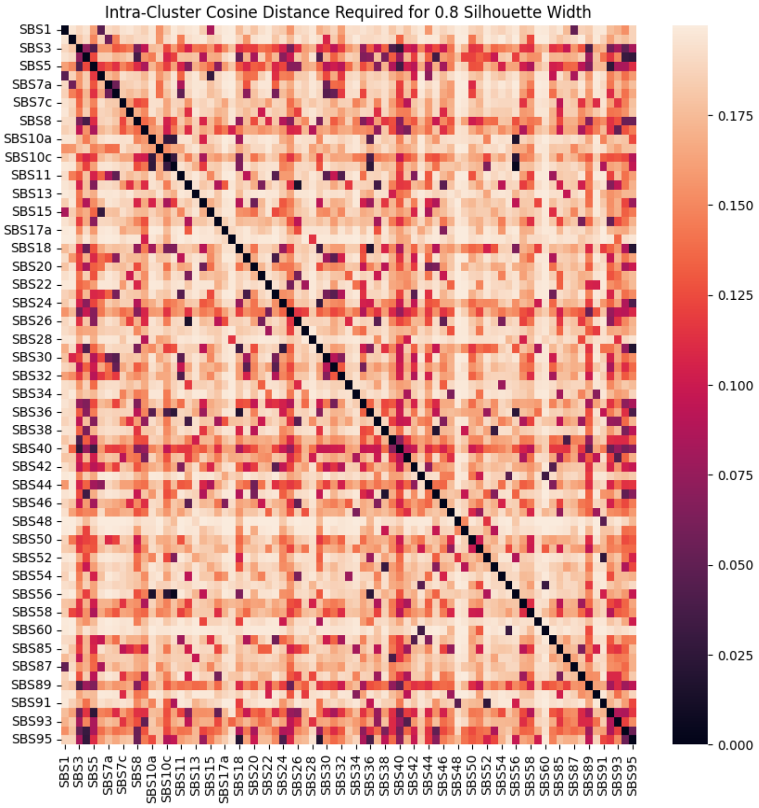


Figure 4: Intra-cluster Cosine Distance Required for Silhouette Width of 0.8

The intra-cluster cosine distance required to meet the average 0.8 silhouette width is significantly lower (Figure 4). However, 53 combinations of signatures have a low cosine distance tolerance of 0.05. Notable is SBS 4 with a low cosine distance tolerance with SBS 8, SBS 29, SBS 45, SBS 94 and SBS 95. The number of combinations with a moderate cosine distance tolerance of less than 0.10 has increased to 273.

These figures are calculated based on whether there is a perfect match between an extracted SBS signature and a known SBS signature. This may be unrealistic as factors such as noise, in addition to the processes used in the framework, may cause a wider variety in the signatures extracted. This poses an issue if clusters have 0.75 inter-cluster cosine distance and 0.15 intra-cluster distance, as width is 0.8. Yet, if clusters have 0.25 inter-cluster cosine distance and 0.15 intra-cluster distance, the width reduces to 0.40. Even with equal intra-cluster distance, silhouette width varies significantly.

Linear Combination Model

The proposed linear combination model (LCM) functions similarly to NMF as it aims to solve equation (1) by minimising the Frobenius norm. However, unlike in NMF, the LCM uses supervised machine learning techniques by only solving the matrix of coefficients *E*. In the LCM, *M* and *E* are individual vectors containing a single sample genome and a single set of coefficients while *P* contains the 79 known SBS signatures. To solve the minimisation problem, the model iteratively updates the coefficients in *E* until the Frobenius norm does not significantly decrease. The purpose of the LCM is twofold. The first is to increase the certainty of the cosine similarity evaluation metrics and the second is to analyse the variation of the signatures in the clustered groups.

As the SBS signature identification only requires a cosine similarity of 0.8, there can be uncertainty when an extracted signature has a high cosine similarity with multiple known signatures. The LCM could be used to reduce this uncertainty as the LCM will contribute the known SBS signature that most truly identifies with the analysed signature with the highest coefficient value. The theory behind this is that the model will analyse multiple combination levels of known SBS signatures and return the best combination for the analysed SBS signature. In this process, the model will maximise the contribution of the known SBS signature that best suits the data and minimise the contribution of those that don't.

For this same reason the LCM can be used as an additional metric to evaluate the variation within a cluster group. After applying an entire cluster group of extracted SBS signatures to the LCM, the intra-cluster distance should be reflected in the spread of the coefficient values. Clusters with a small intra-cluster distance should have a tight spread of coefficient values in one known SBS signature. Clusters with a larger intra-cluster distance should have a wider spread potentially across multiple signatures.

These factors indicate a potential issue with the framework. If the current framework cannot reliably extract SBS signatures with high cosine similarity, then its value to medical researchers and clinicians is limited. The ability to reliably extract a set of SBS signatures from a cancer genome is important, as determining the factors that cause certain cancers can allow researchers and clinicians to develop treatment and prevention measures.

Methods

Implantation of NMF Framework in Python

To answer the research questions, the current NMF framework outlined in Alexandrov, et al. (2013) [4] needed to be replicated. As the framework provided only a methodology for how to extract SBS signatures, it was replicated in Python. The implemented framework followed the outline explained in the background section of this report but with slight alterations for ease of use and flow. A GitHub repository of the code and the generated samples used in this research project is available here: https://github.com/Jmarkey11/RP_2023

Step 1: Dimensionality reduction involved removing the context-dependent mutational types that totalled less than 1% of the total mutations across all types.

Step 2: The Monte Carlo bootstrap resampling step was accomplished using the resample function from the sklearn.utils library. The main parameter required was that replacement equals true. Iteration was incorporated by repeating the resampling process 400 times. No seed value was set as this would only result in the same resampling being executed. This was not an issue for reproducibility as the resampling resulted in a Gaussian distribution.

Step 3: NMF was accomplished using the NMF function from the sklearn.decomposition library. The parameters required for the function included:

- N_components: The number of SBS signatures extracted. As the optimal level is unknown, multiple components needed to be tested and compared. For a full explanation refer to the following section.
- Max_iter: The maximum iterations the algorithm completed before stopping. This was set to 1,000,000.
- Init: How the matrices were initialised. ‘Nndsvda’ was used.
- Random_state: The seed value was set to ‘1234’.
- Tol: The stopping tolerance. Set to equal when the cost function does not deviate by 1e-5.
- Beta_loss: Set to ‘frobenius’ to calculate the Frobenius norm.

The matrices *P* and *E* were extracted using the nmf.components_ and nmf.transform() functions respectively. The FRE was extracted using the nmf.reconstruction_err_ function and the Frobenius reconstruction % was calculated by comparing the error to the Frobenius norm of the sample. Iteration was incorporated by repeating the process for all of the bootstrapped samples. As NMF is an unsupervised machine learning technique, the optimal level for n_components was unknown and needed to be derived. To achieve this, steps 3-5 were repeated at different values and the results were analysed and compared using the evaluation metrics. The optimal n_components was equal to the point where the Silhouette width was maximised and the FRE was minimised without violating the 0.8 and 0.2 silhouette width rules. Given that the samples tested in this research paper were derived from a known number of SBS signatures, the result of the analysis should match this number. In order to reduce the overall runtime of the project, this analysis only involved tuning between 2-6 n_components.

Step 4: As no library in Python clusters according to cosine similarity, a manual version was developed. In this process, cluster centroids were initialised as random signatures extracted from the NMF process using the np.random.choice() function from the Numpy library. The seed value was set prior, using the np.random.seed() function with the value ‘1234’. Each extracted signature’s cosine similarity was calculated and the signature was assigned to the cluster with the highest value. Once all signatures were assigned, the cluster centroid was updated to equal the mean of all assigned signatures. The stopping tolerance for the clustering was calculated to be when the absolute difference between the old and new centroids was less than 1e-10 for all centroids, or when a maximum of 100 iterations were completed. Once completed, the cluster labels and cluster centroids were returned.

Step 5: The final process was to calculate the intra-cluster distance, inter-cluster distance, silhouette width and cosine similarity metrics. The distance metrics were calculated using the cosine_distances() function from the sklearn.metrics library. Before calculating the cosine similarity, the extracted SBS signatures were converted back into their full dimensions. Following this, the cosine similarities between the extracted SBS signatures were calculated in addition to the cosine similarity to the known SBS signatures. For each of the evaluation metrics the mean, standard deviation, min, 25%, 50%, 75%, and max summary statistics were calculated.

Methodology Required to Answer Research Questions

The methodology required to answer research questions 1 and 2 involved analysing the FRE, silhouette width, intra and inter-cluster distances and cosine similarity evaluation metrics. To answer these research questions a comparative analysis was used for each sample. Included in this was an analysis of the mean, median, standard deviation, interquartile range, minimum and maximum for each evaluation metric. To measure the effect of cosine similarity on the results, an ordinary linear regression analysis was completed that plotted the cosine similarity against the evaluation metrics. It was assumed that there was a linear relationship between cosine similarity and the metrics. The statsmodels library was used to create each regression model and derive the metrics. The units of analysis for the linear regression analysis involved analysing the R^2 value, p-value, standard error, MSE and confidence intervals.

The methodology for the implantation of the LCM was centred around the utilisation of the minimize() function from the scipy.optimize library. The minimize() function required certain parameters in order to derive the set of SBS signature coefficients including:

- Fun: The objective function that was optimised. For this parameter, a callable function was created that returned the Frobenius norm using the np.linalg.norm() function from the Numpy library with the equation $M - np.dot(P, E)$.
- X0: The initial guess for the vector of coefficients in E. A vector was initialised with the values for each element in the vector equaling 1/96.
- Bounds: This parameter represents the boundaries for the coefficients. The boundaries needed to be equal to [0, 1] so that the minimisation ensured all SBS signatures were non-negative, and did not exceed 100% for any given signature.
- Constraints: A dictionary of additional constraints minimization. Two additional constraints were set for the type and fun constraints to

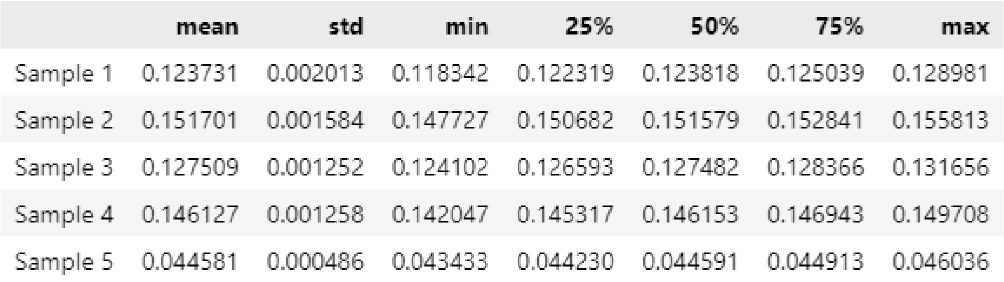
ensure that all values were treated equally and that the sum of all the coefficients was equal to 1.

To complete the minimisation, each cluster’s associated SBS signatures were parsed and a set of coefficients were derived for N cluster groups. The results of the LCM were analysed by examining the location, spread and outliers statistics.

Results

Frobenius Reconstruction Error

Note: As some samples had more mutational types removed during the dimensionality reduction step, the mean probability of each remaining mutational type was variable. Samples with a smaller number of mutational types tended to have larger FREs. As such, the FRE% would be used as the primary evaluation metric for comparative analysis. Analysis of the silhouette widths and FREs (Figures B1-B5, Appendix B) of each sample at a variable number of components indicated that the optimal number of components was equal to 3 across all samples.



	mean	std	min	25%	50%	75%	max
Sample 1	0.123731	0.002013	0.118342	0.122319	0.123818	0.125039	0.128981
Sample 2	0.151701	0.001584	0.147727	0.150682	0.151579	0.152841	0.155813
Sample 3	0.127509	0.001252	0.124102	0.126593	0.127482	0.128366	0.131656
Sample 4	0.146127	0.001258	0.142047	0.145317	0.146153	0.146943	0.149708
Sample 5	0.044581	0.000486	0.043433	0.044230	0.044591	0.044913	0.046036

Table 1: Frobenius Reconstruction Error Percentage Summary Statistics

Table 1 shows the results of the FRE analysis. Results indicated that the current framework was able to most accurately decipher the signatures in Sample 5, as indicated by the low mean and standard deviations. All samples had a consistent mean and standard deviation. For all samples, the mean and median were very similar, indicating that the FRE% had a central tendency. The interquartile range was tight for all signatures indicating that the deciphering accuracy was consistent. There was no obvious direction in the results and the FRE% appeared to be consistent across all samples except Sample 5.

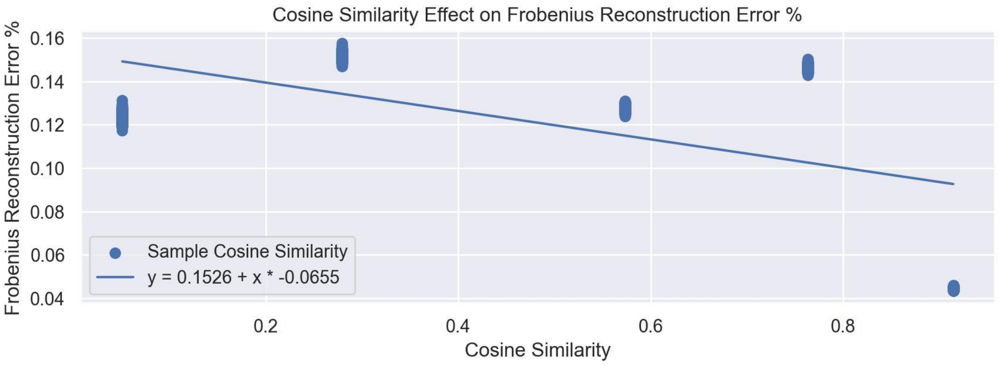
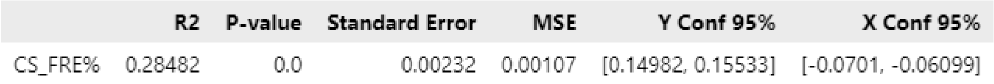


Figure 5: Frobenius Reconstruction Error % vs Cosine Similarity Linear Regression Model

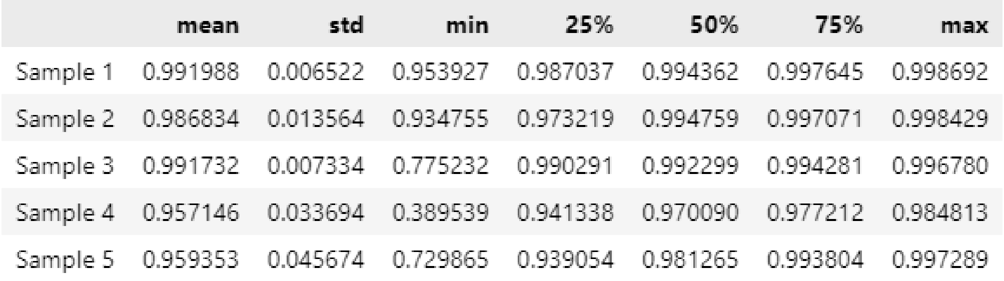


	R2	P-value	Standard Error	MSE	Y Conf 95%	X Conf 95%
CS_FRE%	0.28482	0.0	0.00232	0.00107	[0.14982, 0.15533]	[-0.0701, -0.06099]

Table 2: Linear Regression Analysis Results

The slope of the model (Figure 5) indicates a negative linear relationship between the variables. The R^2 value (Table 2) indicates that only a low proportion of the variability in FRE% can be explained by the cosine similarity. The p-value indicates that there was statistically significant evidence to suggest the relationship between the variables was not random. The standard error and MSE suggest that the model was able to predict values with relatively low error. The confidence intervals indicate that with 95% certainty, when cosine similarity equals 0, the FRE% will equal between 0.1498 and 0.1553, and when cosine similarity equals 1, the FRE% will decrease by -0.0701 to -0.0609.

Silhouette Width



	mean	std	min	25%	50%	75%	max
Sample 1	0.991988	0.006522	0.953927	0.987037	0.994362	0.997645	0.998692
Sample 2	0.986834	0.013564	0.934755	0.973219	0.994759	0.997071	0.998429
Sample 3	0.991732	0.007334	0.775232	0.990291	0.992299	0.994281	0.996780
Sample 4	0.957146	0.033694	0.389539	0.941338	0.970090	0.977212	0.984813
Sample 5	0.959353	0.045674	0.729865	0.939054	0.981265	0.993804	0.997289

Table 3: Silhouette Width Summary Statistics

The results of the silhouette width analysis (Table 3) indicated that the current framework had successfully extracted SBS signatures from all the samples. Samples 1 and 3 had the highest mean silhouette widths and lowest standard deviation values, indicating that the signatures extracted from these samples could be consistently deciphered from a similar sample if analysed. Samples 1, 2 and 3 all appeared to have a central tendency and a tight interquartile range. The moderate lower quartile range of Sample 3 indicated that there were potential outliers in the sample. Samples 4 and 5 had the lowest mean silhouette widths and the largest standard deviations indicating that there was more variation between the extracted signatures in these samples. The difference between the mean and median values also indicated that the samples' silhouette widths contained a slight negative skew. The high lower quartile range of Sample 4 indicated that there may be outliers in the sample. Analysis of these results indicated that there may be a slight negative relationship between silhouette width and cosine similarity.

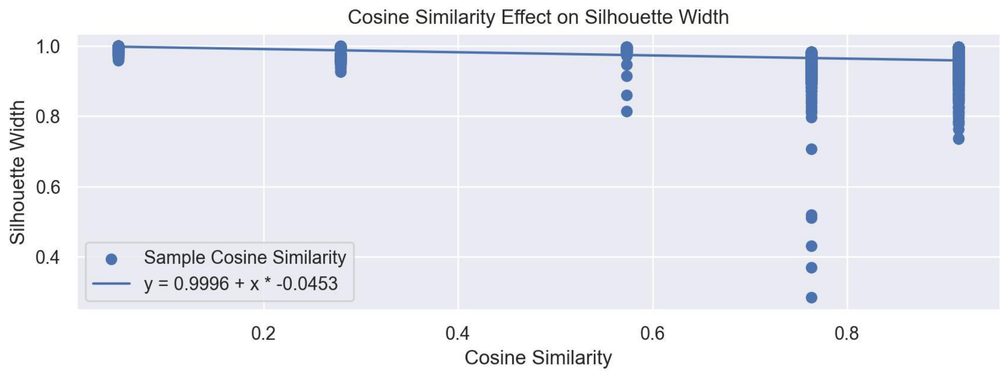
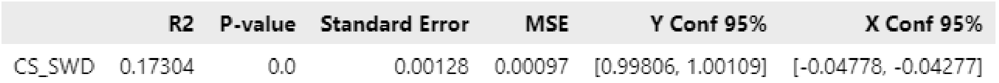


Figure 6: Silhouette Width vs Cosine Similarity Linear Regression

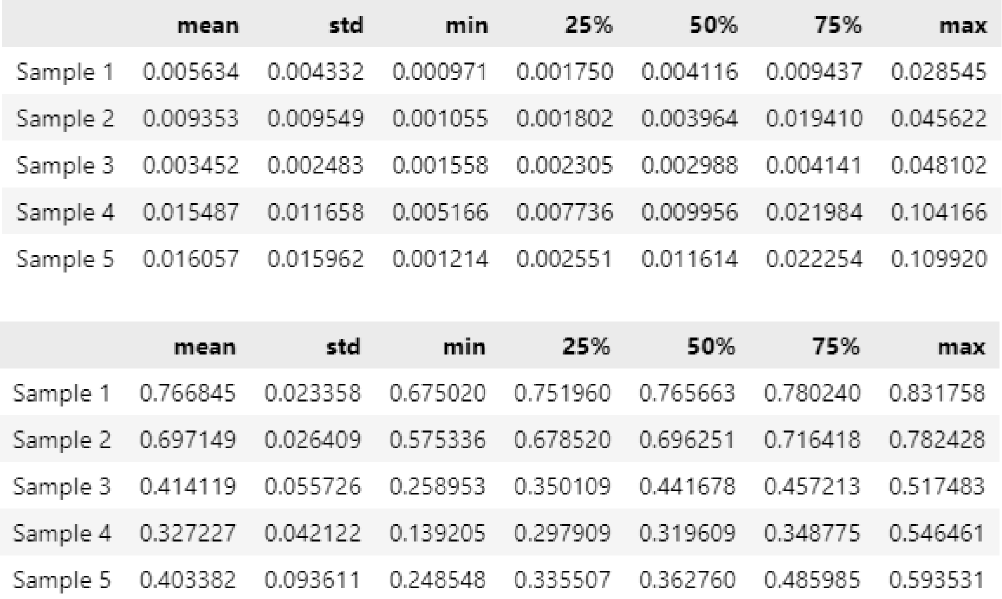


	R2	P-value	Standard Error	MSE	Y Conf 95%	X Conf 95%
CS_SWD	0.17304	0.0	0.00128	0.00097	[0.99806, 1.00109]	[-0.04778, -0.04277]

Table 4: Linear Regression Analysis Results

The linear regression model (Figure 6) indicated a small negative relationship between cosine similarity and silhouette width. The R^2 value (Table 4) indicated that a low proportion of the variance in the silhouette width could be explained by the cosine similarity. The p-value indicates that there was statistically significant evidence to suggest the relationship between the variables was not random. The standard error and MSE suggest that the model was able to predict values with relatively low error. The confidence intervals indicate that with 95% certainty, when cosine similarity equals 0, the silhouette width will equal between 0.9980 and 1.0010, and when cosine similarity equals 1, the silhouette width will decrease by -0.0477 to -0.04277.

Intra and Inter Cosine Distance



	mean	std	min	25%	50%	75%	max
Sample 1	0.005634	0.004332	0.000971	0.001750	0.004116	0.009437	0.028545
Sample 2	0.009353	0.009549	0.001055	0.001802	0.003964	0.019410	0.045622
Sample 3	0.003452	0.002483	0.001558	0.002305	0.002988	0.004141	0.048102
Sample 4	0.015487	0.011658	0.005166	0.007736	0.009956	0.021984	0.104166
Sample 5	0.016057	0.015962	0.001214	0.002551	0.011614	0.022254	0.109920

	mean	std	min	25%	50%	75%	max
Sample 1	0.766845	0.023358	0.675020	0.751960	0.765663	0.780240	0.831758
Sample 2	0.697149	0.026409	0.575336	0.678520	0.696251	0.716418	0.782428
Sample 3	0.414119	0.055726	0.258953	0.350109	0.441678	0.457213	0.517483
Sample 4	0.327227	0.042122	0.139205	0.297909	0.319609	0.348775	0.546461
Sample 5	0.403382	0.093611	0.248548	0.335507	0.362760	0.485985	0.593531

Table 5: Intra-Cluster Distance and Inter-Cluster Distance Summary Statistics

Analysis of the intra and inter-cluster cosine distances (Table 5) indicated that there was a potential relationship between cosine similarity and these metrics. For the intra-cluster cosine distance, the metrics indicated that the model had been able to extract a consistent set of SBS signatures across all samples, with all samples containing less than 2% variance on average. Samples 1 and 3 were the best-performing samples with the lowest mean, standard deviation and interquartile ranges. Samples 4 and 5 contained the largest amount of variance in their clusters with higher standard deviations and interquartile ranges. These samples appeared to be positively skewed with their means exceeding their medians and a substantial upper quartile range. With the exception of Sample 3, the intra-cluster cosine distance appeared to be increasing. The influence of cosine similarity was more prevalent in the intra-cluster cosine distance metrics. The mean inter-cluster distance decreased across all samples with the exception of Sample 5 where it increased. The variability in the intra-cluster distance also appeared to increase as the standard deviation and interquartile range increased across all samples. Finally, all samples appeared to have a mostly central tendency with similar means and medians.

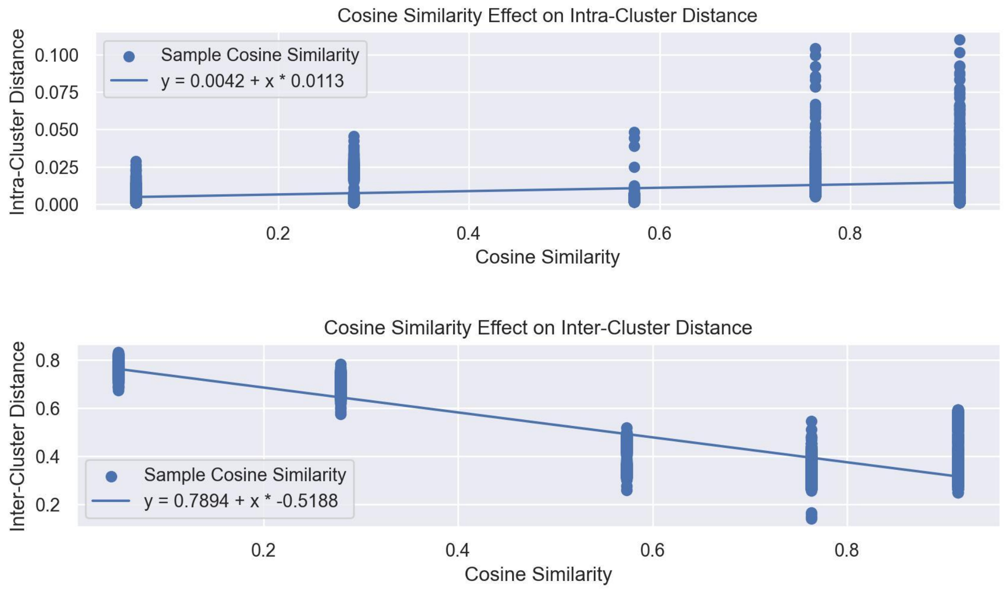
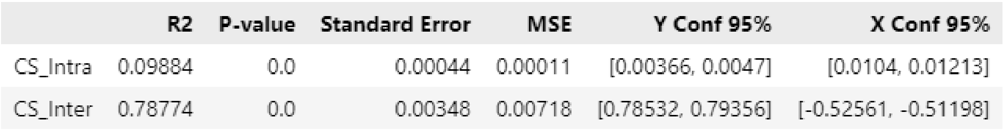


Figure 7: Intra-Cluster Distance and Inter-Cluster Distance vs Cosine Similarity Linear Regression

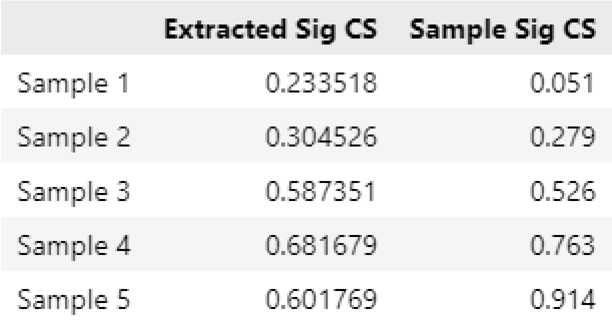


	R2	P-value	Standard Error	MSE	Y Conf 95%	X Conf 95%
CS_Intra	0.09884	0.0	0.00044	0.00011	[0.00366, 0.0047]	[0.0104, 0.01213]
CS_Inter	0.78774	0.0	0.00348	0.00718	[0.78532, 0.79356]	[-0.52561, -0.51198]

Table 6: Linear Regression Analysis Results

The linear regression models (Figure 7) indicated that there was a slight positive relationship between cosine similarity and intra-cluster distance and that there was a negative relationship between cosine similarity and inter-cluster distance. The R^2 value of the intra-cluster model indicated that cosine similarity explains an insignificant quantity of the variation, while the R^2 value of the inter-cluster model indicated that cosine similarity explained a significant quantity of the variation. The p-value indicates that there was statistically significant evidence to suggest the relationship between the variables was not random for both intra-cluster distance and inter-cluster distance. The standard error and MSE suggest that the model was able to predict values with relatively low error for both models. For the intra-cluster model, The confidence intervals indicate that with 95% certainty, when cosine similarity equals 0, the intra-cluster distance will equal between 0.0036 and 0.0047, and when cosine similarity equals 1, the intra-cluster distance will increase between 0.0104 to 0.0121. For the inter-cluster model, The confidence intervals indicate that with 95% certainty, when cosine similarity equals 0, the intra-cluster distance will equal between 0.7853 and 0.7935, and when cosine similarity equals 1, the intra-cluster distance will decrease between -0.5256 to -0.5119.

Cosine Similarity



	Extracted Sig CS	Sample Sig CS
Sample 1	0.233518	0.051
Sample 2	0.304526	0.279
Sample 3	0.587351	0.526
Sample 4	0.681679	0.763
Sample 5	0.601769	0.914

Table 7: Extracted Cosine Similarity vs Sample Cosine Similarity

Table 7 results indicate there is a significant difference in the cosine similarity of Sample 1 and Sample 5 versus their original mean cosine similarity. This change could explain why the inter-cluster cosine distance decreased between Samples 4 and 5 as seen in Table 5. This variance could potentially be explained by the dimensionality reduction. The dimensionality reduction process potentially changed the effect of the weighting of the remaining mutational types on the cosine similarity. This variance may also explain why the current NMF framework didn’t extract signatures that had a 100% cosine similarity with the original known SBS signatures.

The impact of dimensionality reduction could have continued after the reconversion of the signatures back into 96-length vectors. When reconverted, the previously removed mutational types were filled with 0 values. This could be an issue as insignificant values could have a substantial effect on cosine similarity while 0 values could significantly reduce cosine similarity. Further research is required to explore this relationship.

The known SBS signatures with the highest cosine similarity were (Appendix B, Figure B6):

- Sample 1: SBS 38, SBS 6 and SBS 12.
- Sample 2: SBS 10c, SBS 6 and SBS 25.
- Sample 3: SBS 42, SBS 30 and SBS 44.
- Sample 4: SBS 4, SBS 29 and SBS 8.
- Sample 5:SBS 36, SBS 56 and SBS 52.

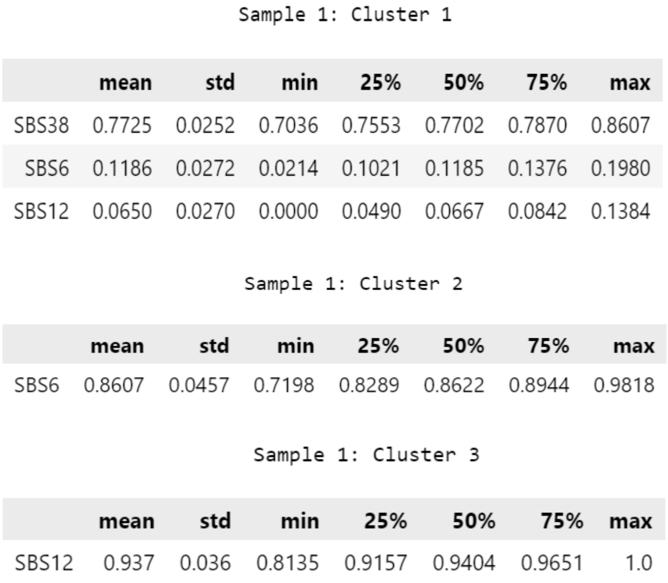
For Samples 1-4, all the contributing SBS signatures had the highest cosine similarity values. For Sample 5, SBS 52 had been incorrectly identified as one of the extracted SBS signatures. SBS 10d had been identified in other clusters but did not have a similarity greater than 0.8 for this cluster. As the cosine similarity of SBS 10d and SBS 56 was 0.9792, the framework may not have been able to distinguish between these results and extracted their contribution as a single signature.

Discussion of Evaluation Metrics Results

The overall results of the evaluation metrics indicated that there only appears to be a strong relationship between the cosine similarity and the inter-cluster distance, as expected. Results also indicated a weak relationship between cosine similarity and FRE%, silhouette width and intra-cluster distance. The cosine similarity analysis indicated that the NMF framework's processes cause the mutational profile of the extracted signatures to change, resulting in different cosine similarities between the extracted signatures. The cosine similarity analysis with the known signatures indicated that the current NMF framework was able to extract the correct SBS signatures from samples 1-4. The results of Sample 5 identified a potential limitation of the NMF framework. The framework may have a problem extracting the signatures present in samples containing signatures with very high cosine similarity. The dimensionality reduction may cause the NMF processes to be unable to separate the signatures with high cosine similarity and may combine them into a single signature. As the combined extracted signature would be very similar to the original signatures, the evaluation metrics would not identify the error. This also potentially indicates that the results from the evaluation metrics cannot be trusted for samples where the derived cosine similarity exceeds the similarity of Sample 4, 0.763. This is a significant problem as, in real-world scenarios, it is unknown what signatures comprise samples until they are parsed through the framework. A recommendation would be to question the results of the process if a signature is extracted that has a very high cosine similarity with any other known signature.

Linear Combination Model

The boxplots of the results of the LCM can be seen in Appendix B, Figures B7 - B21. Tables 8-12 indicate the SBS signatures with mean coefficient values greater than 0.05.



Sample 1: Cluster 1

	mean	std	min	25%	50%	75%	max
SBS38	0.7725	0.0252	0.7036	0.7553	0.7702	0.7870	0.8607
SBS6	0.1186	0.0272	0.0214	0.1021	0.1185	0.1376	0.1980
SBS12	0.0650	0.0270	0.0000	0.0490	0.0667	0.0842	0.1384

Sample 1: Cluster 2

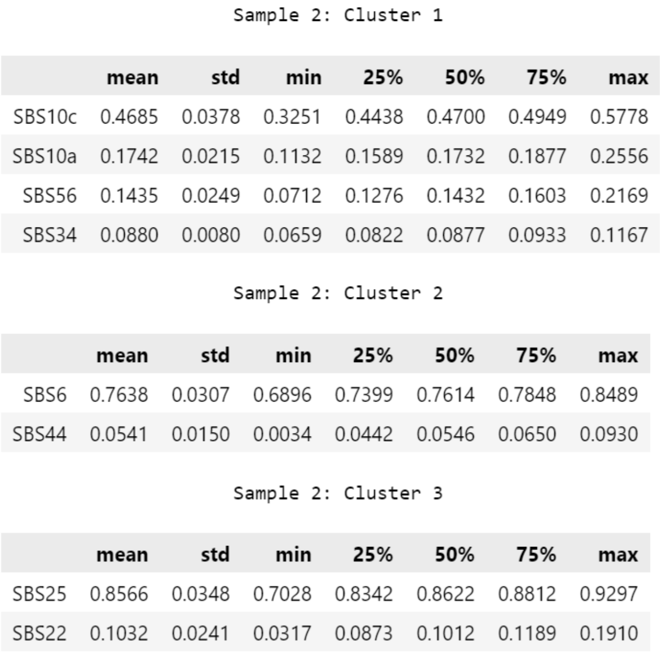
	mean	std	min	25%	50%	75%	max
SBS6	0.8607	0.0457	0.7198	0.8289	0.8622	0.8944	0.9818

Sample 1: Cluster 3

	mean	std	min	25%	50%	75%	max
SBS12	0.937	0.036	0.8135	0.9157	0.9404	0.9651	1.0

Table 8: Sample 1 LCM Results

The results for Sample 1 (Table 8) indicate that the LCM had been able to extract signatures used to derive the sample. The mean coefficient values were statistically significant enough to provide certainty that the identified signature was the actual SBS signature. The coefficient values had a narrow interquartile range for all the signatures. As the coefficients' interquartile ranges were narrow and the intra-cluster distance of the sample was high, the assumption that LCM could be used as a measurement of the intra-cluster distance was supported.



Sample 2: Cluster 1

	mean	std	min	25%	50%	75%	max
SBS10c	0.4685	0.0378	0.3251	0.4438	0.4700	0.4949	0.5778
SBS10a	0.1742	0.0215	0.1132	0.1589	0.1732	0.1877	0.2556
SBS56	0.1435	0.0249	0.0712	0.1276	0.1432	0.1603	0.2169
SBS34	0.0880	0.0080	0.0659	0.0822	0.0877	0.0933	0.1167

Sample 2: Cluster 2

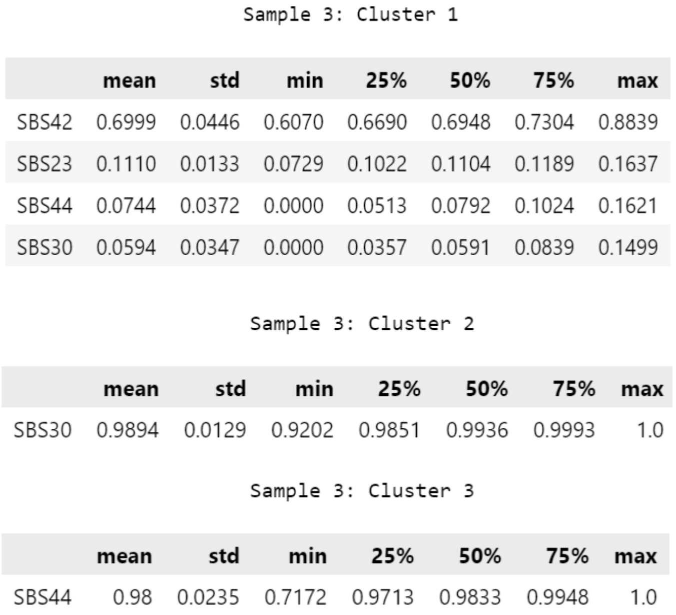
	mean	std	min	25%	50%	75%	max
SBS6	0.7638	0.0307	0.6896	0.7399	0.7614	0.7848	0.8489
SBS44	0.0541	0.0150	0.0034	0.0442	0.0546	0.0650	0.0930

Sample 2: Cluster 3

	mean	std	min	25%	50%	75%	max
SBS25	0.8566	0.0348	0.7028	0.8342	0.8622	0.8812	0.9297
SBS22	0.1032	0.0241	0.0317	0.0873	0.1012	0.1189	0.1910

Table 9: Sample 2 LCM Results

The results for Sample 2 (Table 9) indicate that the LCM had been able to extract signatures used to derive the sample. For clusters 2 and 3 the mean values were statistically significant enough to identify the cluster with certainty. Although lower, the coefficient value for SBS 10c was still the most significant when compared to the other signatures. Although the cluster was similar to the three signatures, it was closest to SBS 10c by a significant margin. The LCM's assumption was supported as both the interquartile ranges were narrow and the intra-cluster distance was small.



Sample 3: Cluster 1

	mean	std	min	25%	50%	75%	max
SBS42	0.6999	0.0446	0.6070	0.6690	0.6948	0.7304	0.8839
SBS23	0.1110	0.0133	0.0729	0.1022	0.1104	0.1189	0.1637
SBS44	0.0744	0.0372	0.0000	0.0513	0.0792	0.1024	0.1621
SBS30	0.0594	0.0347	0.0000	0.0357	0.0591	0.0839	0.1499

Sample 3: Cluster 2

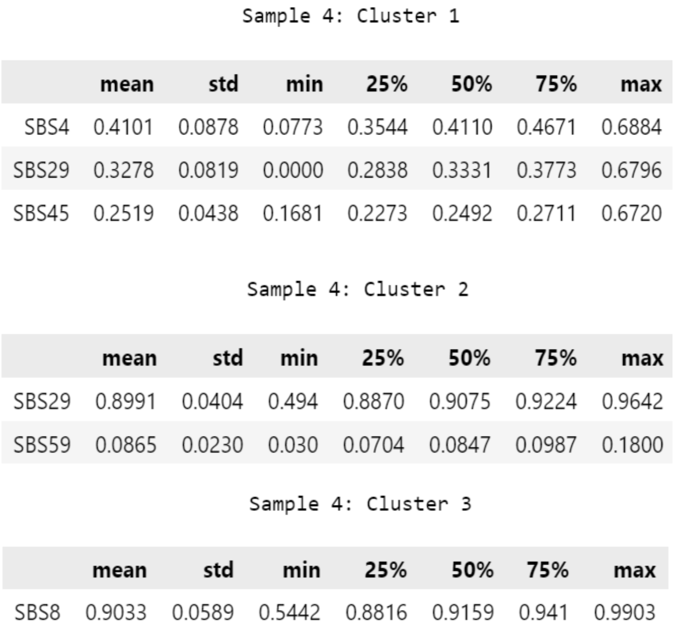
	mean	std	min	25%	50%	75%	max
SBS30	0.9894	0.0129	0.9202	0.9851	0.9936	0.9993	1.0

Sample 3: Cluster 3

	mean	std	min	25%	50%	75%	max
SBS44	0.98	0.0235	0.7172	0.9713	0.9833	0.9948	1.0

Table 10: Sample 3 LCM Results

The results for Sample 3 (Table 10) indicate that the LCM had been able to extract signatures used to derive the sample. The coefficient values are statistically significant enough to support the results of the cosine similarity analysis. The LCM's assumption was supported as both the interquartile ranges are narrow and the intra-cluster distance was small.



Sample 4: Cluster 1

	mean	std	min	25%	50%	75%	max
SBS4	0.4101	0.0878	0.0773	0.3544	0.4110	0.4671	0.6884
SBS29	0.3278	0.0819	0.0000	0.2838	0.3331	0.3773	0.6796
SBS45	0.2519	0.0438	0.1681	0.2273	0.2492	0.2711	0.6720

Sample 4: Cluster 2

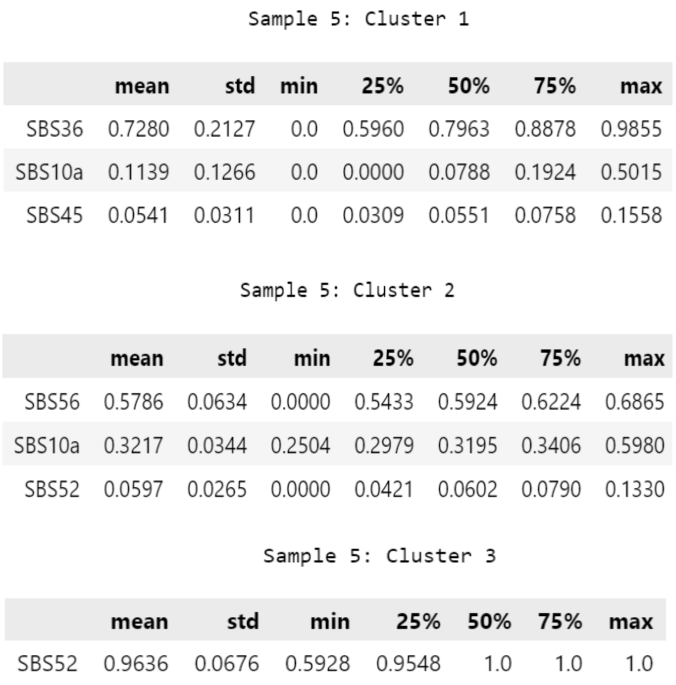
	mean	std	min	25%	50%	75%	max
SBS29	0.8991	0.0404	0.494	0.8870	0.9075	0.9224	0.9642
SBS59	0.0865	0.0230	0.030	0.0704	0.0847	0.0987	0.1800

Sample 4: Cluster 3

	mean	std	min	25%	50%	75%	max
SBS8	0.9033	0.0589	0.5442	0.8816	0.9159	0.941	0.9903

Table 11: Sample 4 LCM Results

The results for Sample 4 (Table 11) indicate that LCM had confidently extracted the signatures identified for clusters 2 and 3, but the certainty for cluster 1 was doubtful. The magnitude differences between the identified signatures were not large enough to provide certainty. The coefficient values range was also high indicating that there was variance in extracted signatures. This was supported by the upper quartile values of the intra-cluster distance for Sample 4 being high, ranging from 0.0219 to 0.1041. This supported the assumption of LCM, as the range in the coefficient values of the identified signatures in cluster 1 increased.



Sample 5: Cluster 1

	mean	std	min	25%	50%	75%	max
SBS36	0.7280	0.2127	0.0	0.5960	0.7963	0.8878	0.9855
SBS10a	0.1139	0.1266	0.0	0.0000	0.0788	0.1924	0.5015
SBS45	0.0541	0.0311	0.0	0.0309	0.0551	0.0758	0.1558

Sample 5: Cluster 2

	mean	std	min	25%	50%	75%	max
SBS56	0.5786	0.0634	0.0000	0.5433	0.5924	0.6224	0.6865
SBS10a	0.3217	0.0344	0.2504	0.2979	0.3195	0.3406	0.5980
SBS52	0.0597	0.0265	0.0000	0.0421	0.0602	0.0790	0.1330

Sample 5: Cluster 3

	mean	std	min	25%	50%	75%	max
SBS52	0.9636	0.0676	0.5928	0.9548	1.0	1.0	1.0

Table 12: Sample 5 LCM Results

The results for Sample 5 (Table 12) supported the results gathered during the cosine similarity analysis as LCM had identified a single SBS signature in each cluster with a high coefficient value. Although SBS 10d had not been identified, this may not have been an issue with the LCM, but rather with the NMF and dimensionality reduction steps as explained in the cosine similarity section. Although the results of Sample 5 had passed all the evaluation metrics of the current framework, there was still an underlying problem that was causing the extracted signatures to differ from the original signatures. The assumption of the LCM model was supported in this sample as the ranges for signatures identified in cluster 1 were high. This matched the intra-cluster distance metrics as, like Sample 4, the upper quartile of the metric was large, ranging between 0.0222 - 0.1099.

Discussion of LCM Results

Overall analysis of the LCM results indicated that the model was able to provide increased certainty for the results of the cosine similarity analysis and provide an indication of the intra-cluster distance. The LCM was able to provide increased certainty to the cosine similarity metrics by attributing a single signature with a high mean coefficient value compared to any other identified signatures. The model provided an indication of intra-cluster distance through the analysis of the spread of the coefficient values for each signature. The LCM was able to more consistently decipher the same set of coefficient values for clusters with a low intra-cluster distance but as the distance increased the range of coefficient values increased.

Conclusion

The purpose of this research project was to address two research questions:

- **RQ1**: How does the cosine similarity of SBS signatures in human cancer genome samples relate to the outcomes produced by the existing SBS signature extraction framework?
- **RQ2**: Can a linear combination model enhance the certainty of the results obtained from the current SBS signature extraction framework?

Answering these questions holds significant importance, given that the current NMF framework is the standard for extracting signatures in human cancer genomes, a methodology pivotal to clinicians and medical researchers [6][12][13]. Tackling these research questions directly contributes to enhancing the reliability and confidence in the outcomes, enabling these stakeholders to potentially develop more effective treatments and preventive strategies for diverse types of cancers.

To answer these questions a set of samples was generated from a list of SBS signatures of various cosine similarities. These samples were analysed using the current framework's evaluation metrics and then parsed through the LCM. These results revealed a strong relationship between cosine similarity and inter-cluster distance, and a weak relationship between cosine similarity and intra-cluster distance, silhouette width and FRE%. However, this analysis could only be trusted for samples that had a derived cosine similarity of 0.763 as it was also found that signatures were potentially inaccurate when extracted from samples containing signatures with a derived cosine similarity greater than 0.763. At a cosine similarity of 0.914, the current framework may not be able to decompose a sample into the original signatures. Instead, the framework may combine these signatures into a single extracted signature. The implication of this is that if the framework extracts a known SBS signature with a cosine similarity exceeding 0.763 with any other known signature, the results should be questioned.

The outcomes from the LCM indicate its potential to heighten the certainty of the results of cosine similarity analysis. A signature could be identified with high certainty when there was a high coefficient value in a single known signature compared to any other signature. Furthermore, analysis of the range statistics of coefficient values from extracted signatures unveiled a connection between the spread of coefficient values and intra-cluster distance. Wider coefficient value spreads aligned with larger intra-cluster distances, while tighter spreads corresponded to smaller intra-cluster distances. Although the silhouette width and intra-cluster distance analysis indicated that samples with a derived cosine similarity of 0.763 or less were unlikely to fail the current evaluation metrics, adding the LCM model to the current framework would be beneficial in improving the certainty and accuracy of results.

A primary limitation of this study was the number of samples tested. Despite efforts to optimise the runtime of the NMF framework, the complexity of algorithms and parameter tuning significantly extended processing times. The framework's execution for a single sample can take upwards of 30 minutes. Given that the samples were derived from 3 SBS signatures, and there exist potentially 79079 combinations of SBS signatures for testing, the projected runtime would be approximately 4.513 years. While the analysis of such an extensive range of combinations is unfeasible, future endeavours could consider a wider array of cosine similarities. The samples utilised in this study covered a cosine similarity interval of around 0.25. Expanding the sample range and narrowing the cosine similarity interval could validate and extend the conclusions drawn from this research.

In summary, this research advances our understanding of the relationship between cosine similarity, SBS signatures, and the efficacy of the current signature extraction framework. The implications of this work underscore a limitation of the current framework as well as the potential for refinement. While limitations of this research exist, they pave the way for future investigations that could amplify the significance and scope of these findings.

Word Count (Excluding Abstract): 6493

Appendices A

Acknowledgements

Lewis Mitchell

Paul Wang

Satyaveer Pattanaik

Shubham Bhatia

Links

Samples and code available here: https://github.com/Jmarkey11/RP_2023

Abbreviations:

DNA - Deoxyribonucleic Acid

FRE - Frobenius Reconstruction Error

COSMIC - Catalogue Of Somatic Mutations In Cancer

SBS - Single Base Substitution

NMF - Non-negative Matrix Factorization

LCM - Linear Combination Model

MSE - Mean Squared Error

Appendices B

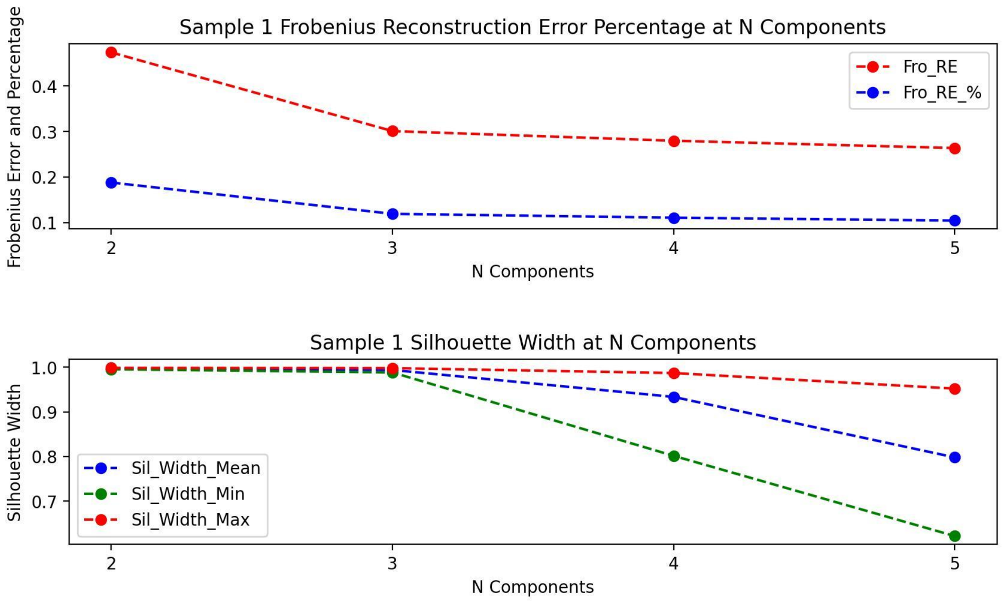


Figure B1: Sample 1 Silhouette Width and Frobenius Reconstruction Error

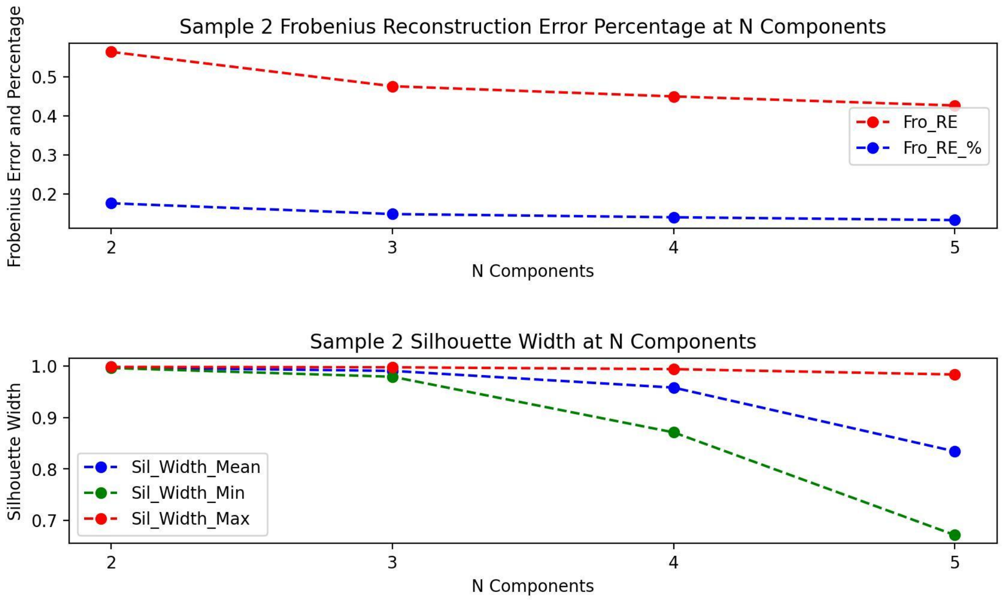


Figure B2: Sample 2 Silhouette Width and Frobenius Reconstruction Error

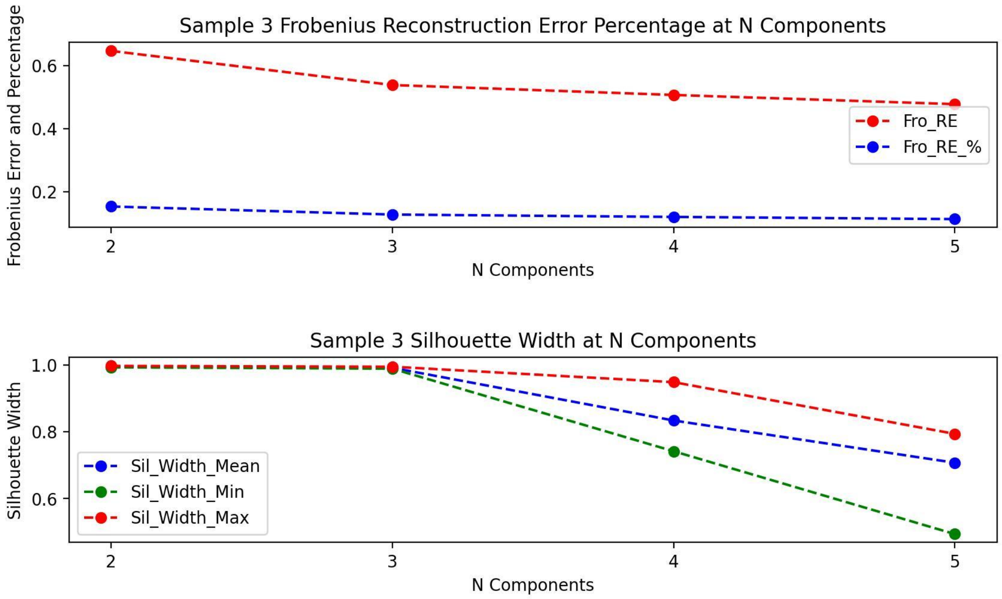


Figure B3: Sample 3 Silhouette Width and Frobenius Reconstruction Error

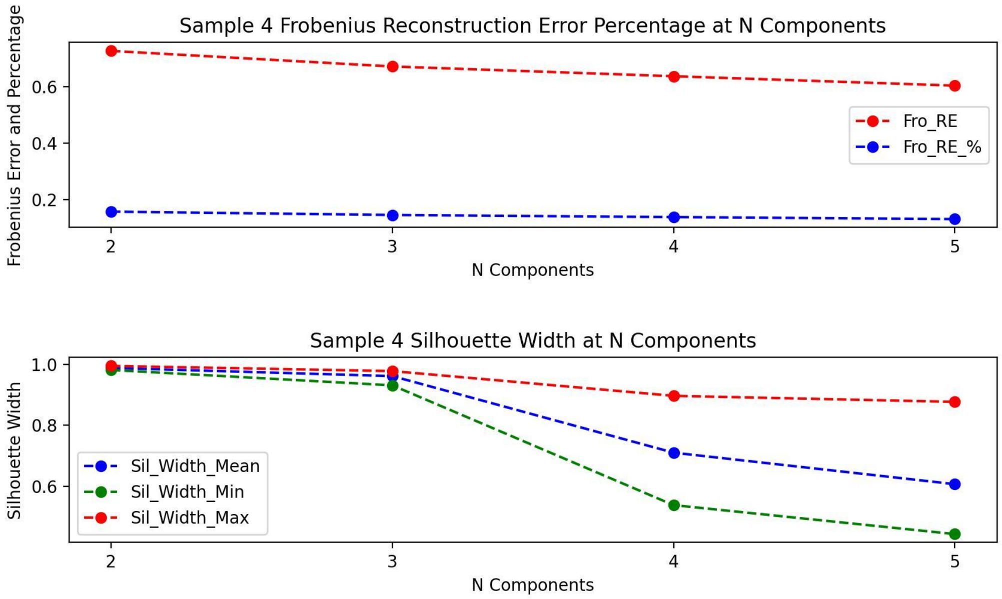
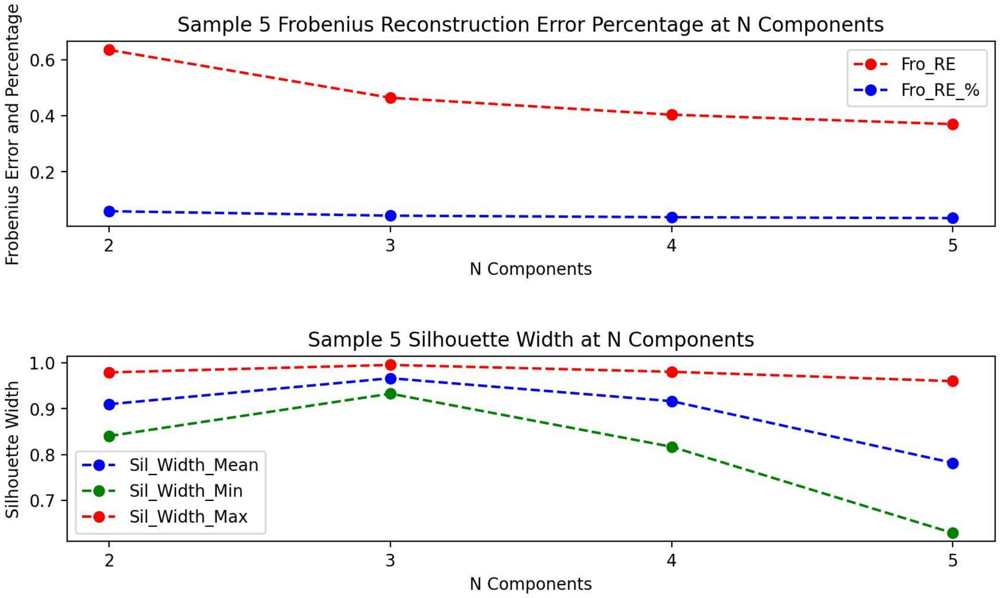


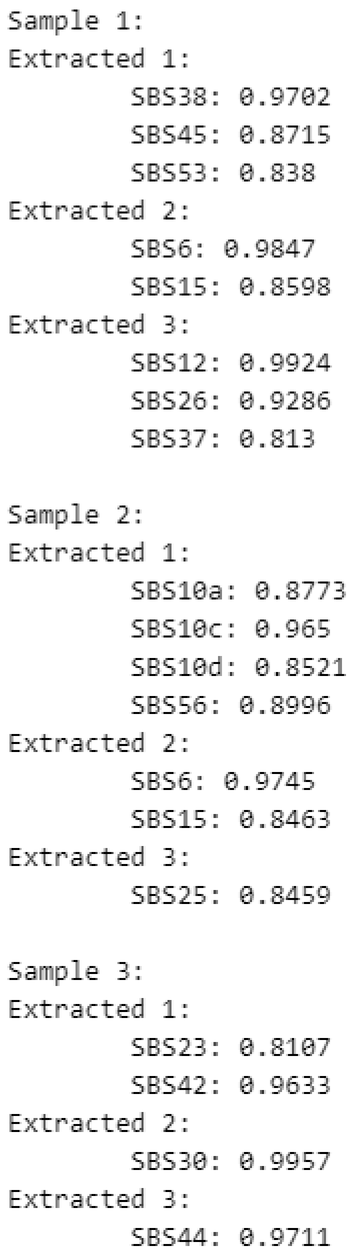
Figure B4: Sample 4 Silhouette Width and Frobenius Reconstruction Error



Sample 5 Frobenius Reconstruction Error Percentage at N Components

Sample 5 Silhouette Width at N Components

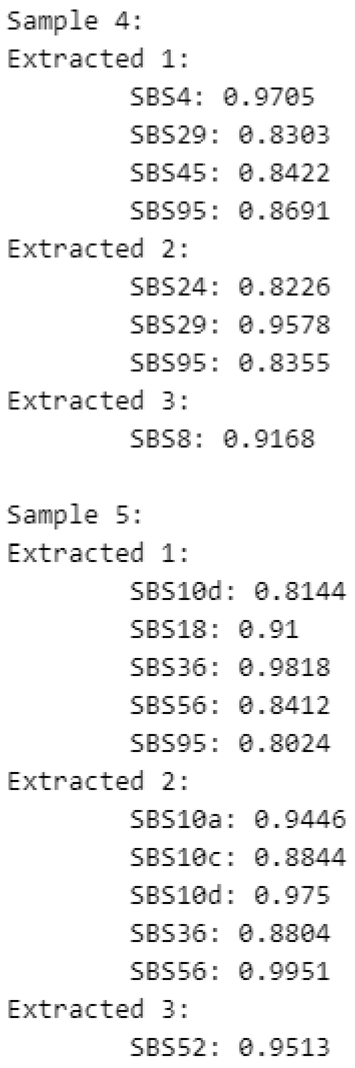
Figure B5: Sample 5 Silhouette Width and Frobenius Reconstruction Error



```
Sample 1:
Extracted 1:
        SBS38: 0.9702
        SBS45: 0.8715
        SBS53: 0.838
Extracted 2:
        SBS6: 0.9847
        SBS15: 0.8598
Extracted 3:
        SBS12: 0.9924
        SBS26: 0.9286
        SBS37: 0.813

Sample 2:
Extracted 1:
        SBS10a: 0.8773
        SBS10c: 0.965
        SBS10d: 0.8521
        SBS56: 0.8996
Extracted 2:
        SBS6: 0.9745
        SBS15: 0.8463
Extracted 3:
        SBS25: 0.8459

Sample 3:
Extracted 1:
        SBS23: 0.8107
        SBS42: 0.9633
Extracted 2:
        SBS30: 0.9957
Extracted 3:
        SBS44: 0.9711
```



```
Sample 4:
Extracted 1:
        SBS4: 0.9705
        SBS29: 0.8303
        SBS45: 0.8422
        SBS95: 0.8691
Extracted 2:
        SBS24: 0.8226
        SBS29: 0.9578
        SBS95: 0.8355
Extracted 3:
        SBS8: 0.9168

Sample 5:
Extracted 1:
        SBS10d: 0.8144
        SBS18: 0.91
        SBS36: 0.9818
        SBS56: 0.8412
        SBS95: 0.8024
Extracted 2:
        SBS10a: 0.9446
        SBS10c: 0.8844
        SBS10d: 0.975
        SBS36: 0.8804
        SBS56: 0.9951
Extracted 3:
        SBS52: 0.9513
```

Figure B6: Extracted Known SBS Signature Cosine Similarity

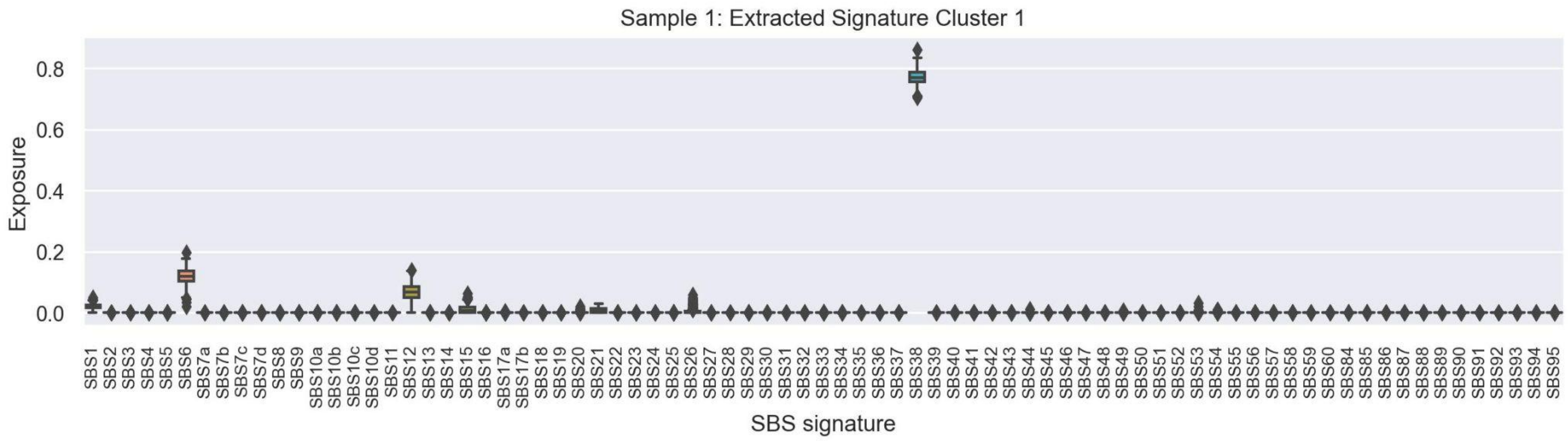


Figure B7: Sample 1 Extracted Cluster 1 Boxplot of LCM Coefficient Exposures

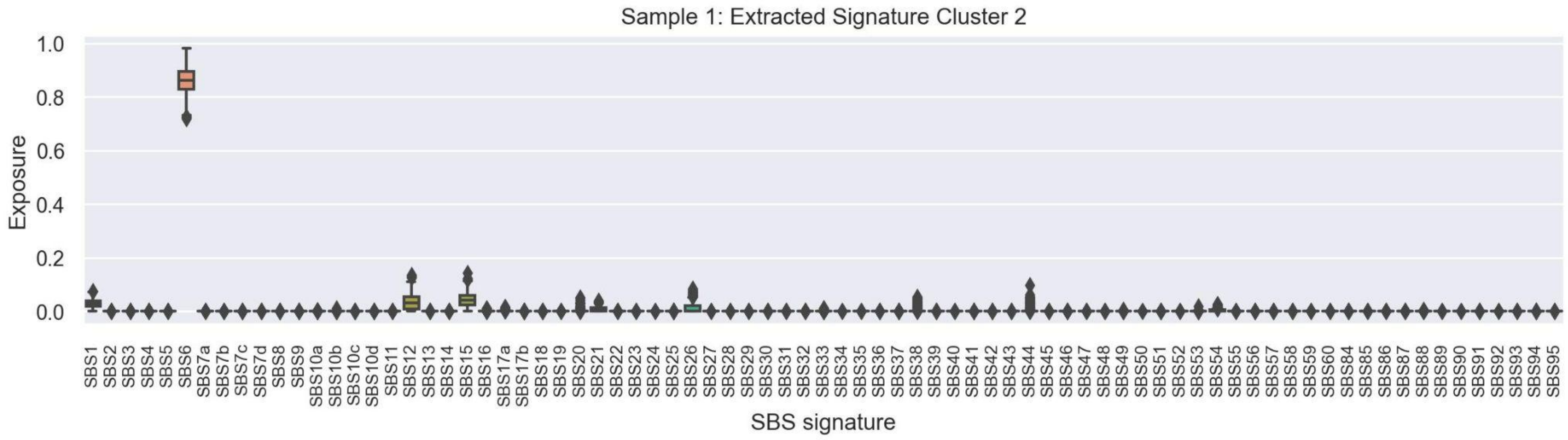


Figure B8: Sample 1 Extracted Cluster 2 Boxplot of LCM Coefficient Exposures

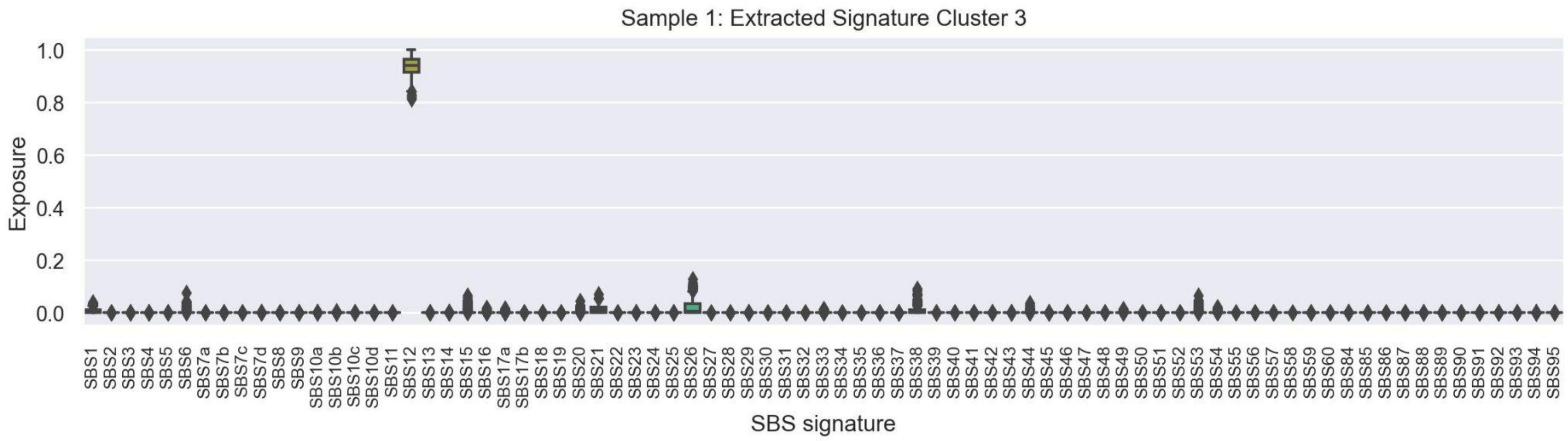


Figure B9: Sample 1 Extracted Cluster 3 Boxplot of LCM Coefficient Exposures

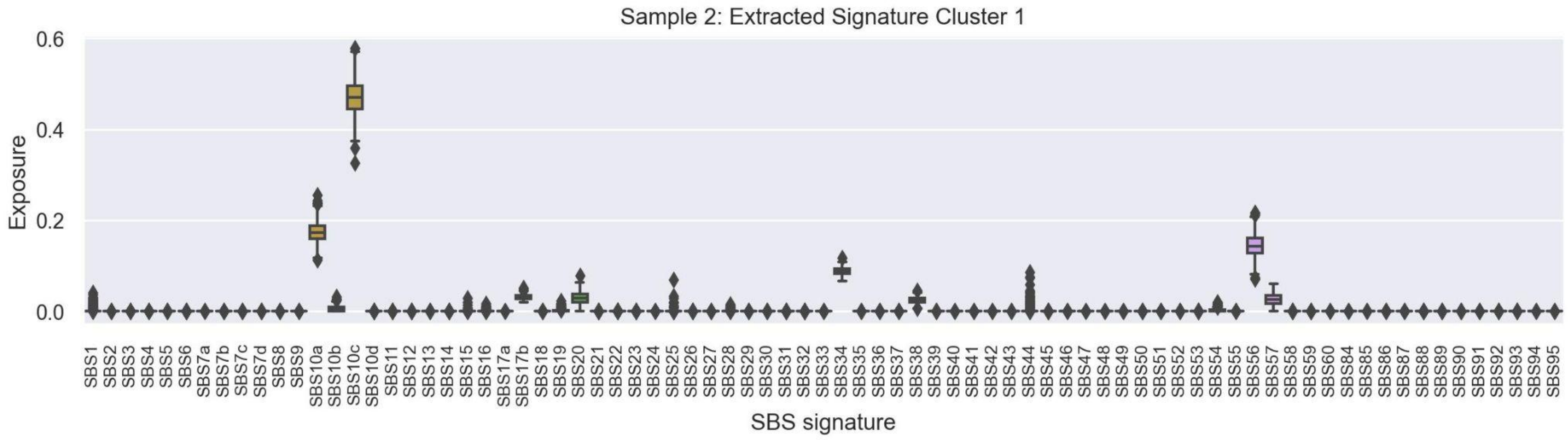


Figure B10: Sample 2 Extracted Cluster 1 Boxplot of LCM Coefficient Exposures

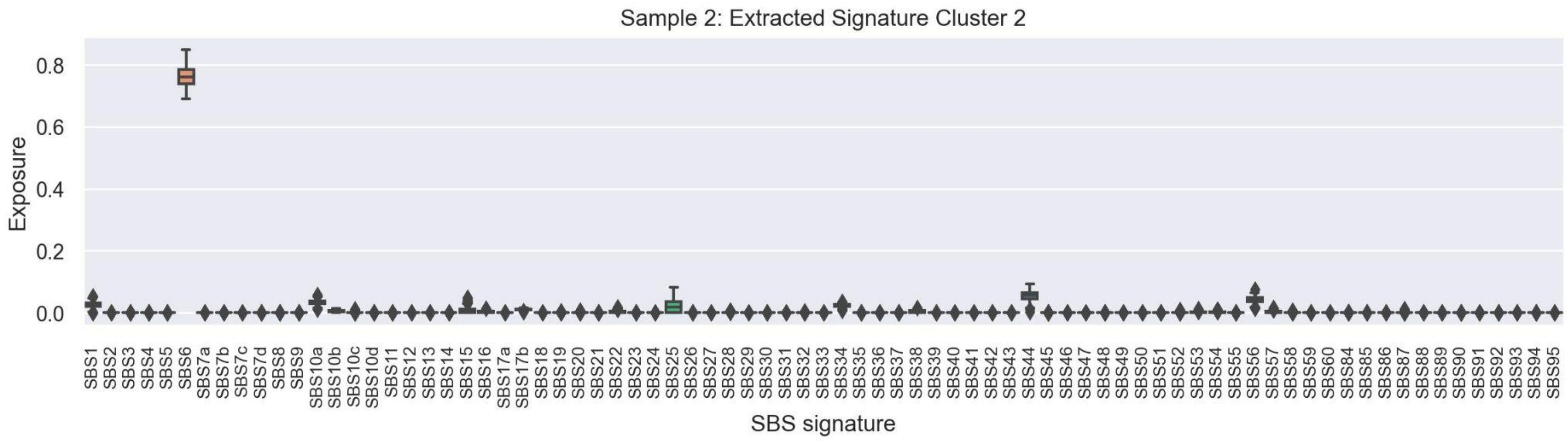


Figure B11: Sample 2 Extracted Cluster 2 Boxplot of LCM Coefficient Exposures

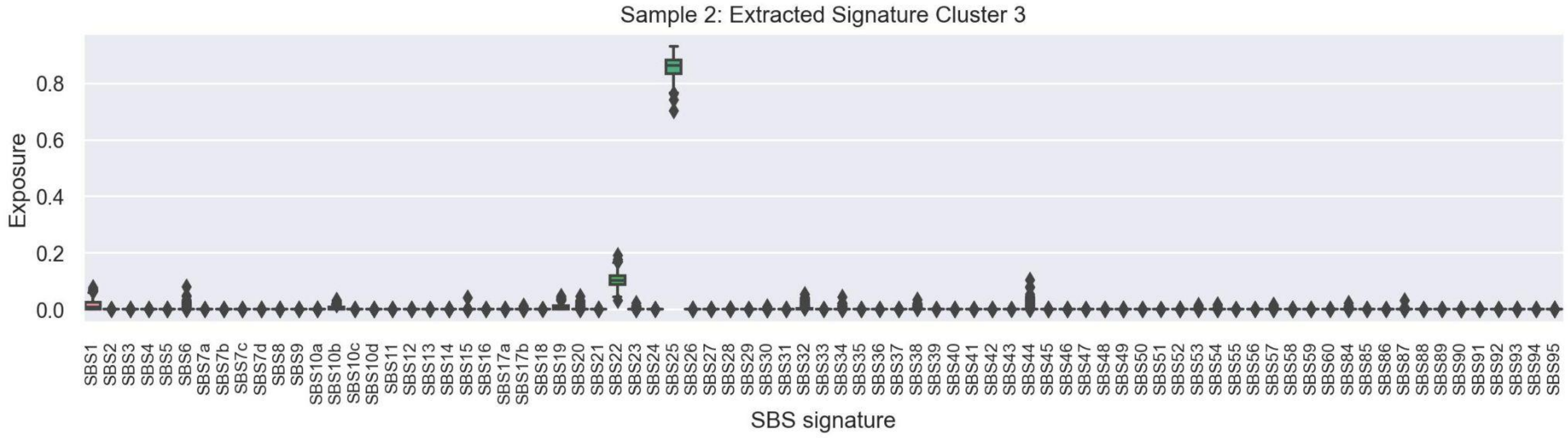


Figure B12: Sample 2 Extracted Cluster 3 Boxplot of LCM Coefficient Exposures

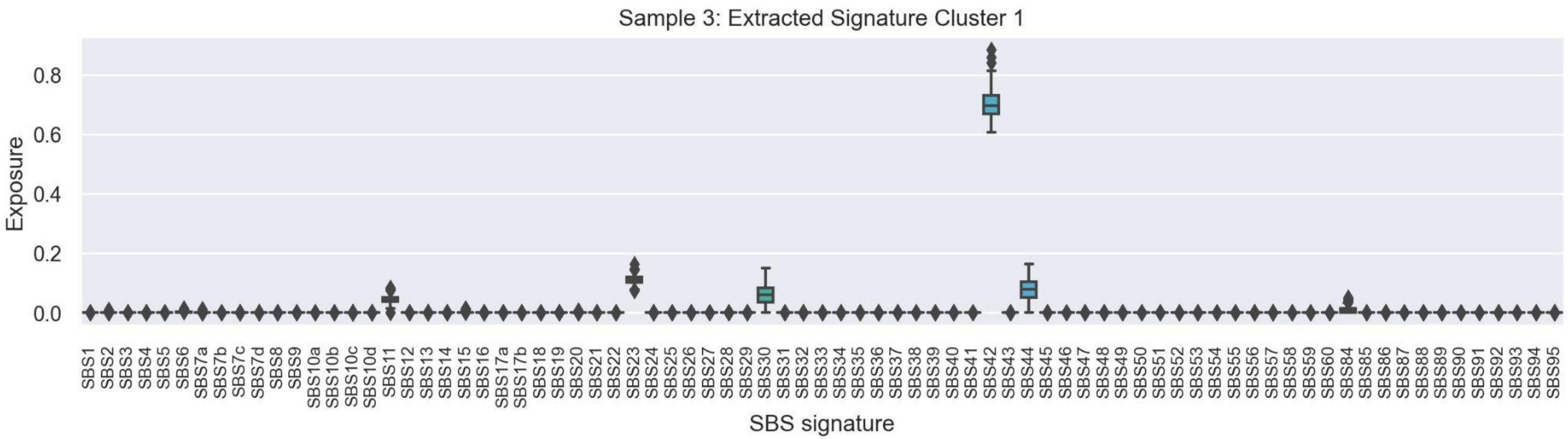


Figure B13: Sample 3 Extracted Cluster 1 Boxplot of LCM Coefficient Exposures

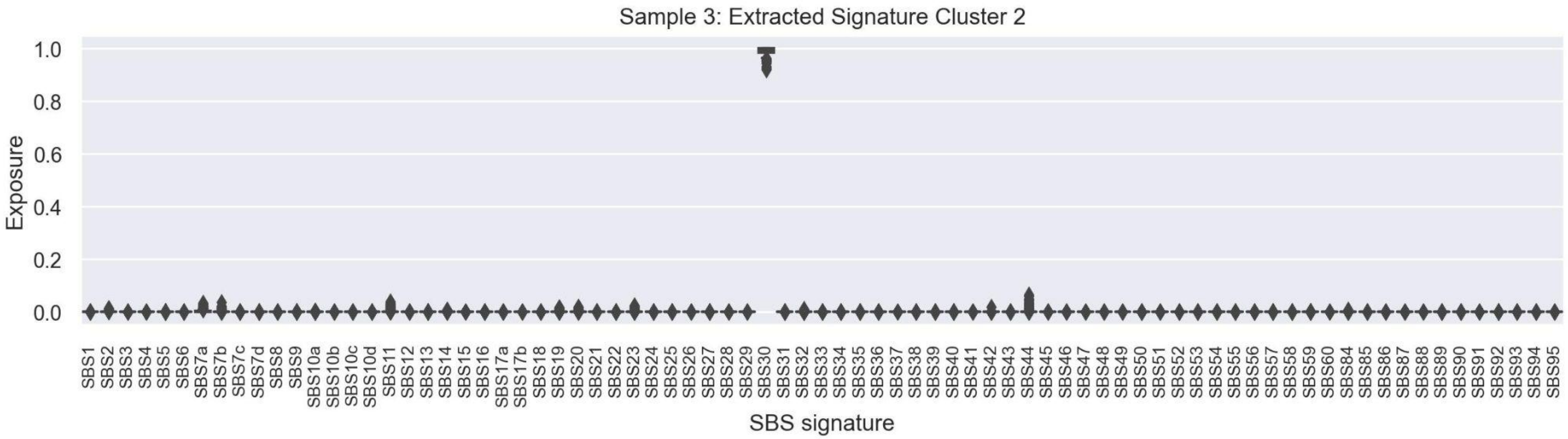


Figure B14: Sample 3 Extracted Cluster 2 Boxplot of LCM Coefficient Exposures

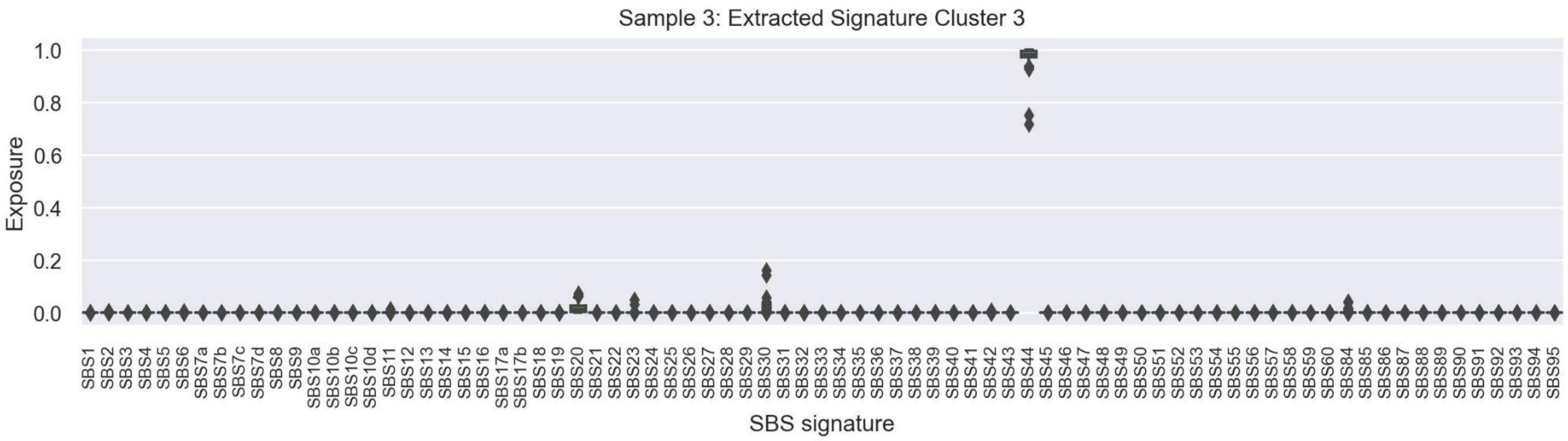


Figure B15: Sample 3 Extracted Cluster 3 Boxplot of LCM Coefficient Exposures

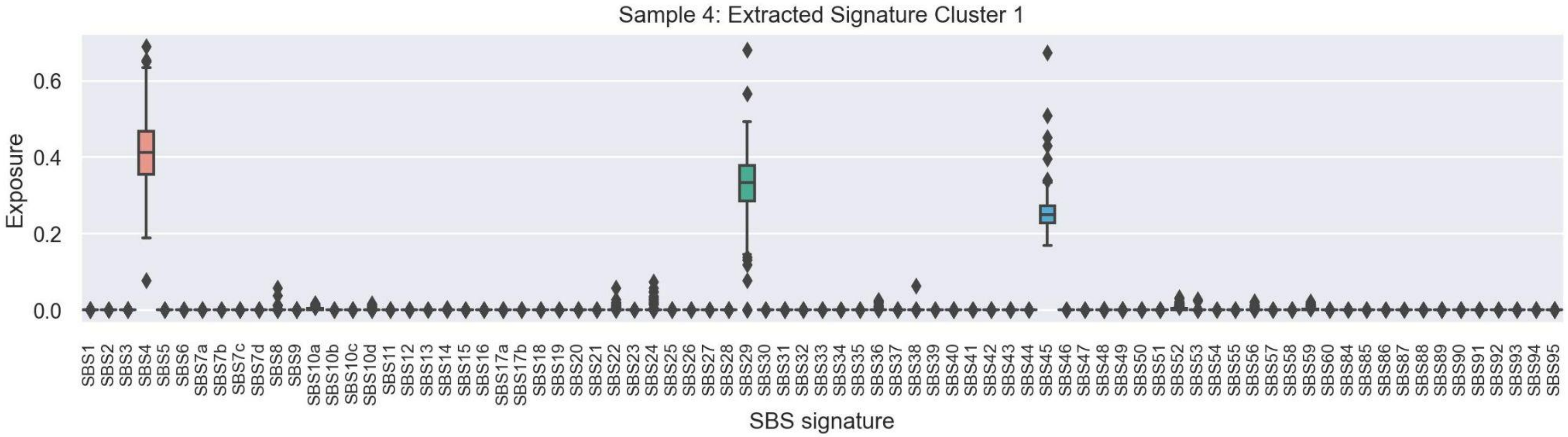


Figure B16: Sample 4 Extracted Cluster 1 Boxplot of LCM Coefficient Exposures

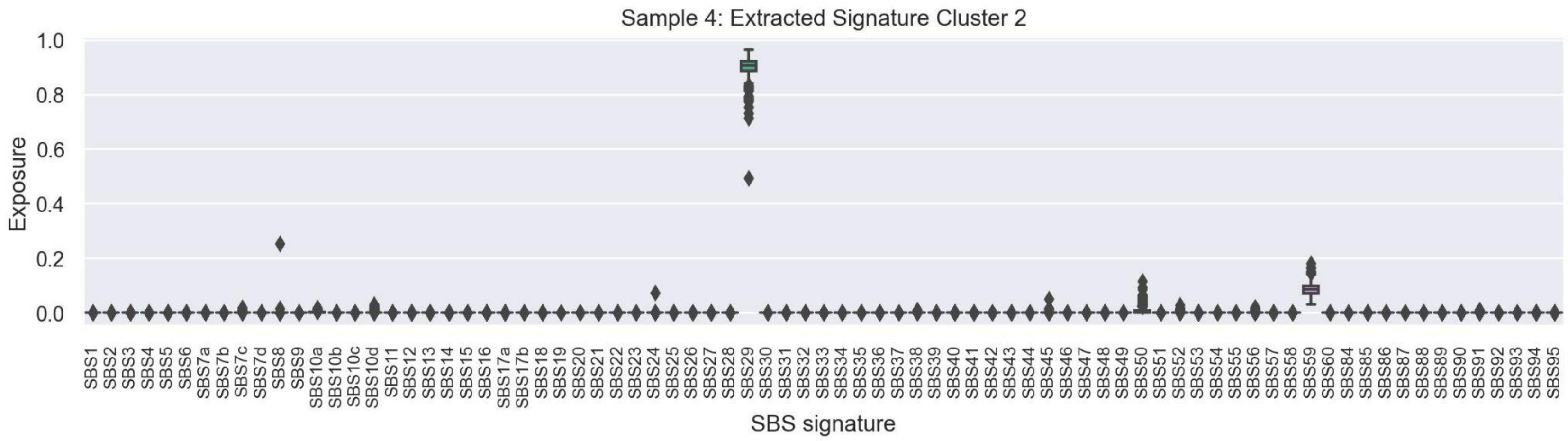

Sample 4: Extracted Signature Cluster 2

Figure B17: Sample 4 Extracted Cluster 2 Boxplot of LCM Coefficient Exposures

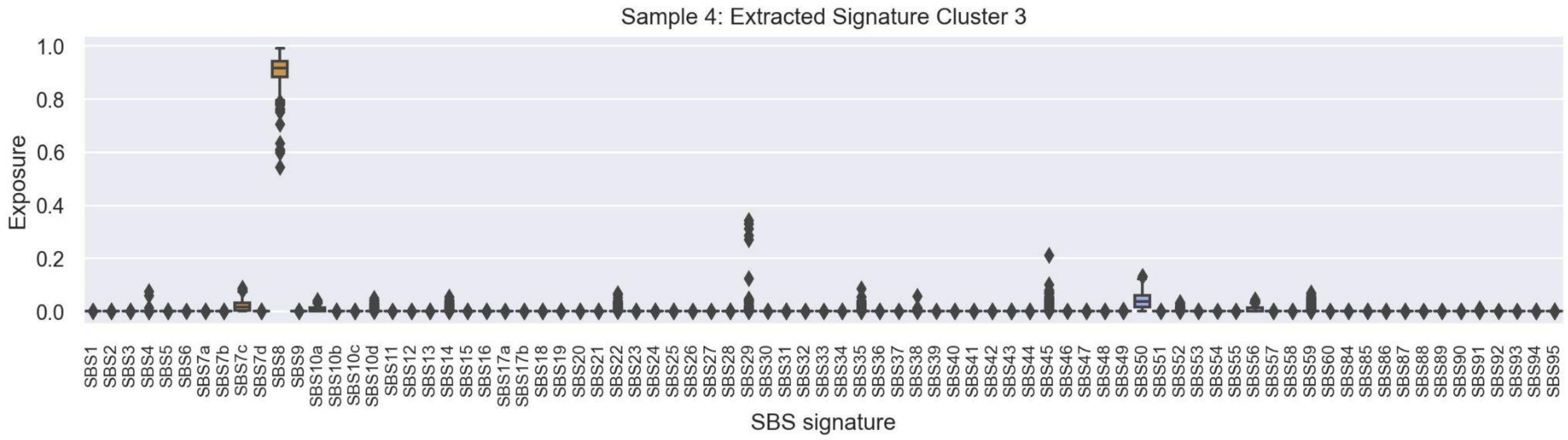

Sample 4: Extracted Signature Cluster 3

Figure B18: Sample 4 Extracted Cluster 3 Boxplot of LCM Coefficient Exposures

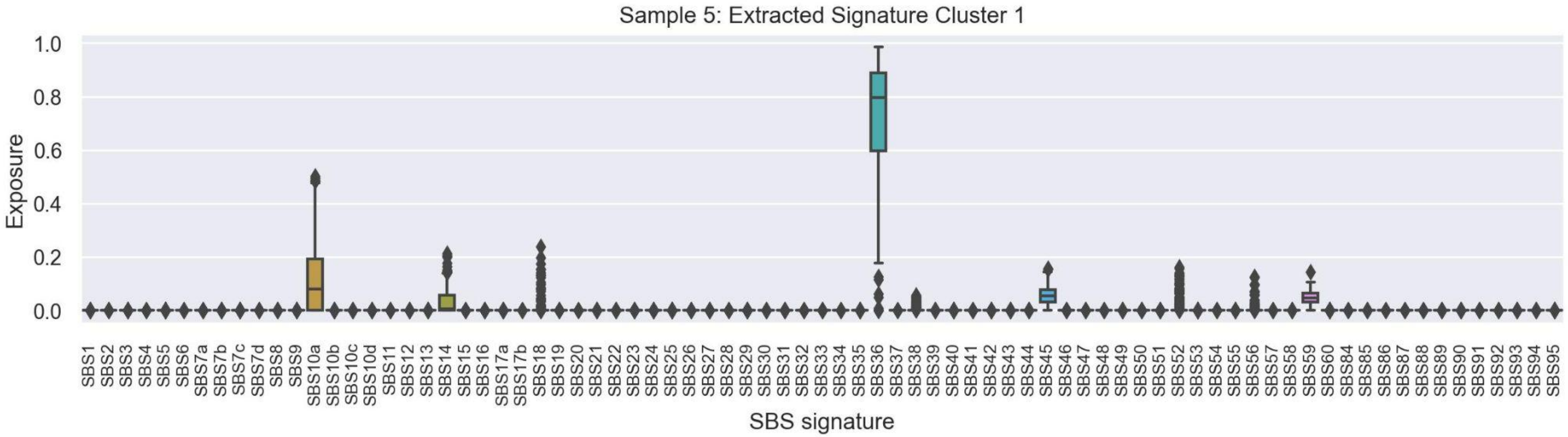


Figure B19: Sample 5 Extracted Cluster 1 Boxplot of LCM Coefficient Exposures

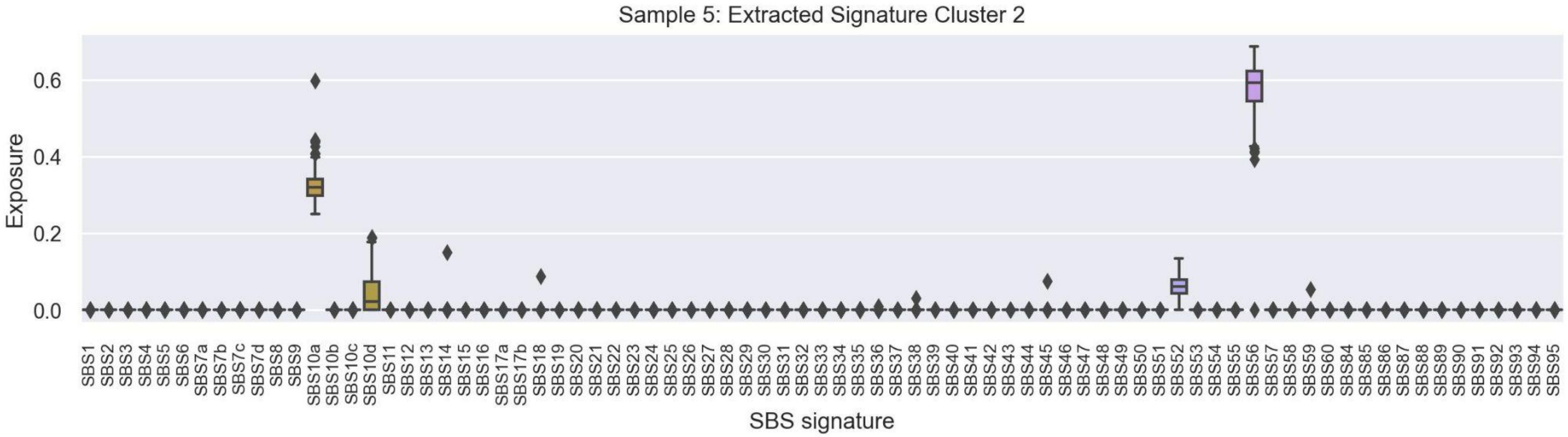


Figure B20: Sample 5 Extracted Cluster 2 Boxplot of LCM Coefficient Exposures

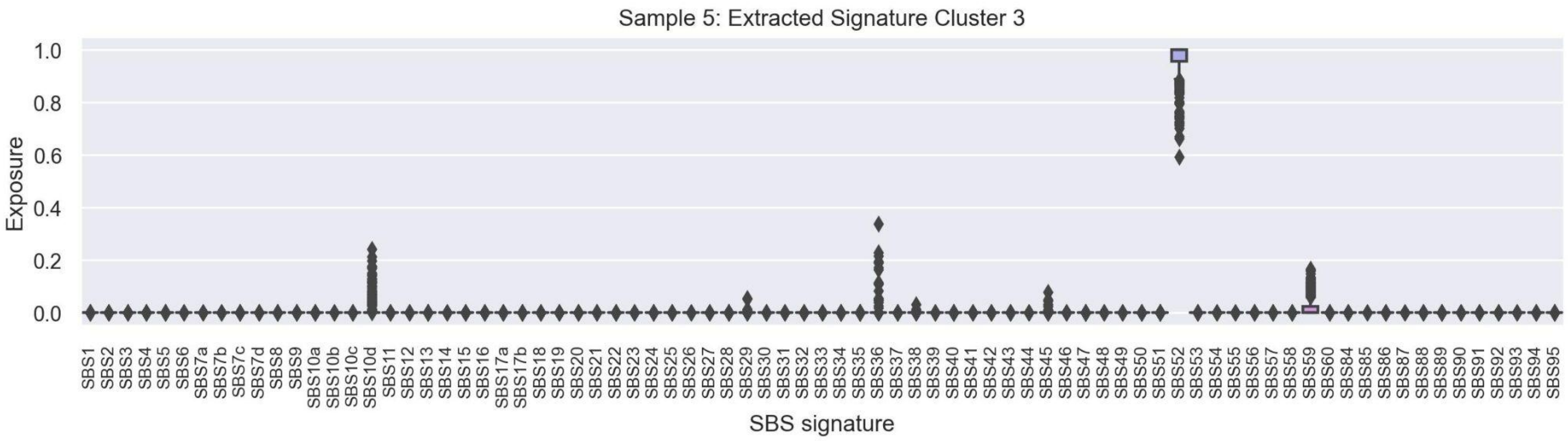


Figure B21: Sample 5 Extracted Cluster 3 Boxplot of LCM Coefficient Exposures

References

1. Helleday, T., Eshtad, S. and Nik-Zainal, S. (2014) ‘Mechanisms underlying mutational signatures in human cancers’, *Nature Reviews Genetics*, 15(9), pp. 585–598. doi:10.1038/nrg3729.
2. Alexandrov, L.B. et al. (2013) ‘Signatures of mutational processes in human cancer’, *Nature*, 500(7463), pp. 415–421. doi:10.1038/nature12477.
3. Fischer, A. *et al.* (2013) ‘EMU: Probabilistic inference of mutational processes and their localization in the cancer genome’, *Genome Biology*, 14. doi:10.1186/gb-2013-14-4-r39.
4. Gunnarsson, R. *et al.* (2022) ‘Single base substitution and insertion/deletion mutational signatures in adult core binding factor acute myeloid leukemia’, *Leukemia*, 36(6), pp. 1681–1684. doi:10.1038/s41375-022-01552-x.
5. *Mutational signatures (v3.3 - June 2022)* (2020) *COSMIC | SBS - Mutational Signatures*. Available at: https://cancer.sanger.ac.uk/signatures/downloads/ (Accessed: 15 August 2023).
6. Alexandrov, L.B. *et al.* (2013) ‘Deciphering signatures of mutational processes operative in human cancer’, *Cell Reports*, 3(1), pp. 246–259. doi:10.1016/j.celrep.2012.12.008.
7. Rosales, R.A. *et al.* (2017) ‘Signer: An empirical bayesian approach to mutational signature discovery’, *Bioinformatics*, 33(1), pp. 8–16. doi:10.1093/bioinformatics/btw572.
8. Rosenthal, R. *et al.* (2016) ‘DeconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution’, *Genome Biology*, 17(1). doi:10.1186/s13059-016-0893-4.
9. Huang, X., Wojtowicz, D. and Przytycka, T.M. (2017) ‘Detecting presence of mutational signatures in cancer with confidence’, *Bioinformatics*, 34(2), pp. 330–337. doi:10.1093/bioinformatics/btx604.
10. Omichessan, H., Severi, G. and Perduca, V. (2019) ‘Computational tools to detect signatures of mutational processes in DNA from tumours: A review and empirical comparison of performance’, *PLOS ONE*, 14(9). doi:10.1371/journal.pone.0221235.
11. Alexandrov, L.B. *et al.* (2020) ‘The repertoire of mutational signatures in human cancer’, *Nature*, 578, pp. 94–101. doi:10.1038/s41586-020-1943-3.
12. de Kanter, J.K. *et al.* (2021) ‘Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients’, *Cell Stem Cell*, 28(10). doi:10.1016/j.stem.2021.07.012.

13. Thatikonda, V. *et al.* (2023) ‘Comprehensive analysis of mutational signatures reveals distinct patterns and molecular processes across 27 pediatric cancers’, *Nature Cancer*, 4(2), pp. 276–289. doi:10.1038/s43018-022-00509-4.

14. Islam, S.M.A. *et al.* (2022) ‘Uncovering novel mutational signatures by de novo extraction with Sigprofilerextractor’, *Cell Genomics*, 2(11). doi:10.1016/j.xgen.2022.100179.