Supervised Non-negative Matrix Factorization

Integrated learning of mutational signatures and prediction of DNA repair pathway deficiencies

Sander Goossens

TU Delft

Supervised Non-negative Matrix Factorization

Integrated learning of mutational signatures and prediction of DNA repair pathway deficiencies

by

Sander Goossens

Student number:	4482514	
Masters programme:	Nanobiology & Computer Science, Bioinformatics specialization	
Faculty:	Electrical Engineering, Mathematics and Computer Science	
Project Duration:	September, 2021 - July, 2022	
Thesis committee:	Prof. dr. ir. Marcel Reinders	TU Delft
	Dr. Joana Gonçalves	TU Delft, supervisor
	Dr. Joris Pothof	Erasmus MC
	Dr. Megha Khosla	TU Delft
	MSc Yasin Tepeli	TU Delft, daily supervisor

TU Delft

Preface

During my study in nanobiology I became fascinated about the huge amounts and variety of biological data that is generated and the potential knowledge and applications that lie within it. I decided to pursue my interest and started a double masters degree in nanobiology and in computer science with the specialization bioinformatics. This thesis is the result of a 10 month research project in which I have been able to apply and integrate the knowledge and experience I gathered throughout both studies.

Around spring 2021, my thesis supervisor Joana Gonçalves introduced me to the idea of applying machine learning to predict DNA repair deficiencies based on the patterns of mutations in the DNA. After having conducted a literature survey on mutational signature decomposition techniques, I became enthusiastic about the idea of integrating mutational signature decomposition and the prediction of DNA repair deficiencies into a single machine learning model. During the journey thereafter leading up to this thesis I have learned a lot, from searching for an appropriate dataset and the challenges that come with analyzing biological data, to deriving, implementing and evaluating our novel algorithm: Supervised Non-negative Matrix Factorization. This introduction to scientific research has been an excellent learning experience for me, and will hopefully lead to an exciting future career in bioinformatics.

I would like to acknowledge the people without whom this work would not have been possible. First, I would like to thank Joana Gonçalves for her guidance, constructive feedback and engagement with this project. I enjoyed our discussions which always provided me with new insights, ideas and enthusiasm. I would also like to thank Yasin Tepeli for his support, helpful comments and always being approachable for questions throughout the project. Additionally, I would like to thank everyone else from the Goncalves lab, Attila Csala, Jurrian de Boer, Mathijs de Wolf, Aaron Wenteler, and in specific Colm Seale who guided me during the first half of this project and provided me with instructive discussions and advice. Furthermore, I would like to thank Marcel Reinders, Joris Pothof, and Megha Khosla for being part of my thesis defence committee. I look forward to discuss my work with them and learn from their criticism. I would like to express gratitude to my family and friends for their unconditional support during the entirety of my studies. Finally, I would like to thank my significant other, Sanne Verheul for all her love, patience, and encouragement.

Sander Goossens
Delft, June 2022

Integrated learning of mutational signatures and prediction of DNA repair pathway deficiencies

Sander Goossens [1], Yasin Tepeli [1], and Joana Gonçalves [1,]

[1] Pattern Recognition and Bioinformatics, Intelligent Systems Dept., EEMCS Faculty, Delft University of Technology, Netherlands

Abstract

Motivation: Many tumors show deficiencies in DNA damage repair. These deficiencies can play a role in the disease, but also expose vulnerabilities with therapeutic potential. Targeted treatments exploit specific repair deficiencies, for instance based on synthetic lethality. To decide which patients could benefit from such therapies requires the ability to determine the repair deficiency status of a tumor. It has been suggested that mutational signatures could be better predictors of DNA repair deficiency than loss of function in select genes. However, current models for prediction of repair deficiency rely on mutational signatures extracted using unsupervised learning techniques. As a result, the signatures are not optimized to discriminate between repair deficiency status or pathway. We argue that the supervised learning of mutational signatures guided by repair deficiency status could enable the identification of signatures that are predictive of repair deficiency, and capture underlying mechanisms of DNA repair.

Results: We propose S-NMF, a supervised non-negative matrix factorization method, which jointly optimizes two objectives: (1) learning of signatures shared across tumor samples using NMF, and (2) learning of signatures predictive of repair deficiency using logistic regression. We apply S-NMF to mutation profiles of human induced pluripotent cell lines carrying knockouts of genes involved in three DNA repair pathways: homologous recombination, base excision repair, and mismatch repair. We show that S-NMF achieves high prediction accuracy (0.971) and learns signatures that better distinguish the repair deficiency of a sample. Signatures extracted by S-NMF are similar to cancer-related signatures associated with the same repair deficiency. Additionally, S-NMF can capture signatures of deficiencies affecting distinct subpathways within a main repair pathway (e.g. OGG1 and UNG mechanisms in base excision repair).

Contact: a.c.h.goossens@student.tudelft.nl

1 Introduction

DNA damage can be caused by a variety of endogenous (e.g. DNA replication errors) and exogenous (e.g. environmental toxins or radiation) factors. These factors cause specific types of DNA damage such as base mismatches, single- and double-stranded breaks, or intra and interstrand crosslinks. There are multiple DNA repair pathways that recognize and repair specific types of DNA damage. For example, double-stranded breaks, base mismatches and single stranded breaks are respectively repaired by homologous recombination (HR), mismatch repair (MMR), and base excision repair (BER) (1, 2) (Fig. 1). However, repair mechanisms are not error free: they make mistakes which leave mutations in the DNA.

The accumulation of mutations can lead to genome instability, one of the enabling hallmarks of cancer(3). Mutations are also responsible for deficiencies in DNA repair mechanisms frequently occurring in tumor cells. For example, in breast cancer 1-5% of the tumors are attributed to inherited mutations in the *BRCA1* or *BRCA2* gene (*BRCA1/2*) (4, 5). The *BRCA1/2* genes are essential for repair of double-strand breaks (DSBs) mediated by homologous recombination (HR) (6). In tumors with an inherited *BRCA1/2* mutation, the second *BRCA1/2* allele is usually inactivated during tumorigenesis leading to HR repair deficiency (7, 8).

The relation between specific cancers and DNA repair deficiencies provides an opportunity for targeted treatment, for instance by exploiting known synthetic lethalities. Two genes are synthetically lethal when the inactivation of both simultaneously results in cell death, while inactivation of only one of the genes does not affect the viability of the cell (9). This means that, for example, patients with *BRCA1/2* deficient breast cancer can be treated with *PARP1* inhibitors, which cause more DSBs in the DNA (10). The accumulation of DSBs in the DNA results in an increased demand for proficient HR repair (11), which is absent in *BRCA1/2* deficient tumor cells. This accumulation of unrepaired DNA damage (12) is lethal, making *PARP* inhibitors an effective treatment for HR deficient tumors (13).

To apply targeted treatment interfering with repair pathway deficiency, it is necessary to identify whether the deficiency is present in the tumor. One of the ways this is done is by detecting mutation leading to loss of function in genes known as essential for the specific repair mechanism. However, this approach has limited sensitivity since not all genes involved in each repair pathway are known, and genes can be inactivated by mechanisms that are more difficult to identify, such as epigenetic modifications (14). To circumvent these limitations, methods have been developed that predict a repair deficiency based on the specific pattern of mutations it leaves on the genome of the tumor cells (15, 16). These patterns, also termed mutational signatures, are indicative of the different DNA damage and repair processes taking place during tumorigenesis.

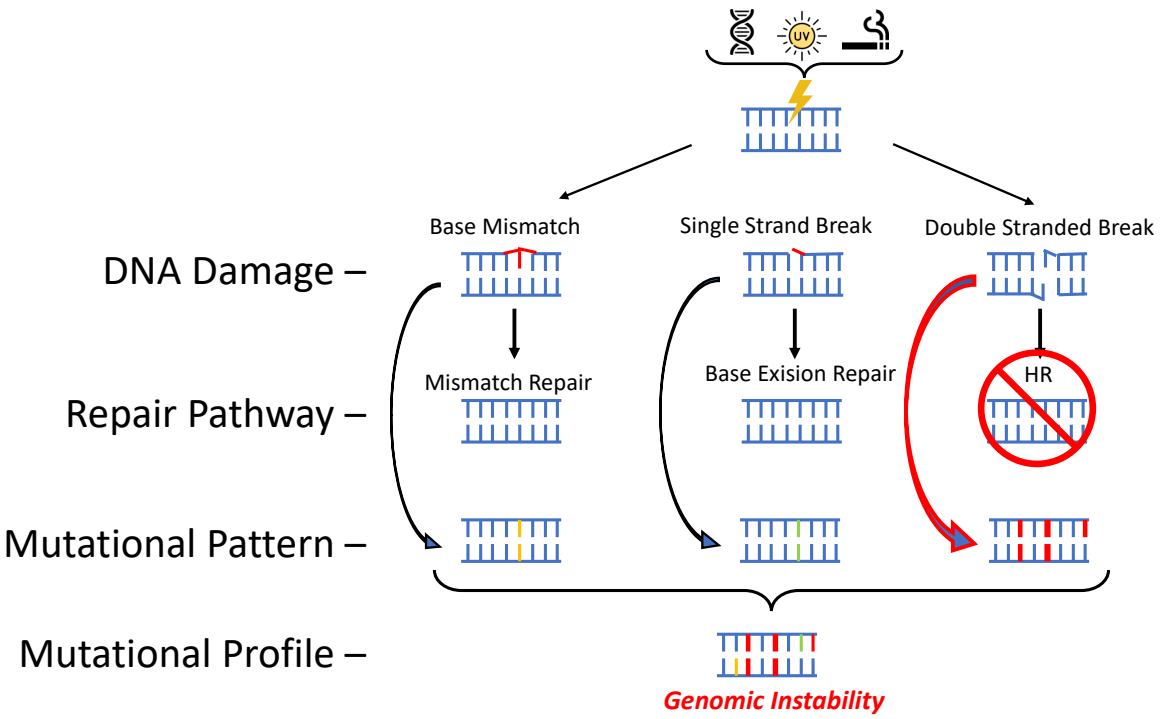


Fig. 1: **Schematic overview of DNA damage and DNA repair pathways (deficiency).** A variety of factors can cause specific types of DNA damage. In healthy cells DNA damage is recognized and repaired by specific repair pathways. This processes leave a mutational pattern in the genome specific to the type of DNA damage and repair pathway that have occurred. When a repair pathway is deficient, there is a elevated accumulation of mutations which could potentially result in genomic instability.

Mutational signature decomposition. Methods to identify mutational signatures and estimate their contributions (or exposures) to genome mutation profiles can be categorized into *de novo* and *refitting* approaches. *De novo* methods identify both signatures and exposures. This is typically done using non-negative matrix factorization, which decomposes an input matrix of mutation profiles from multiple genomes into two matrices: one with a set of mutational signatures, and the other with the exposures of those signatures for each genome (17). Extensions and alternatives include approaches that do bayesian inference of the number of mutational signatures (18) or take into account mutational opportunities (19; 20). *Refitting* methods estimate exposures of predefined signatures for new genomes (21; 22; 23). Commonly used are the signatures identified from cancer patient genomes from the 'Catalogue of Somatic Mutations in Cancer' (COSMIC)(24; 25). Several of the COSMIC signatures have been linked to specific aetiologies, including DNA repair deficiencies.

Non-integrated learning of signatures and prediction of repair deficiency. Recent methods have also been developed to predict HR or MMR pathway deficiency based on signature exposures estimated for new samples based on known COSMIC signatures (15; 16). These approaches estimate exposures for signatures previously identified using *de novo* signature decomposition methods (NMF), and then use supervised learning to build a model to predict repair deficiency for each sample based on its exposures.

Even though the signatures used by non-integrated exposure-based prediction models could be related to repair deficiency, they are not optimized to discriminate repair deficient from proficient tumors (or deficiency in different pathways). This is because current mutational signature decomposition methods (NMF) are unsupervised, meaning that they seek to best capture underlying mutation patterns without any prior knowledge of the biological processes they may be associated with. A supervised signature decomposition method would be able to exploit prior knowledge about the genomes (e.g. repair pathway deficiency) to potentially find more representative and discriminative signatures. We therefore reason that supervised mutational signature decomposition could identify mutational signatures that both (1) capture the underlying mutational processes, and (2) are predictive of DNA repair deficiency.

Supervised NMF for integrated learning of signatures and prediction tasks. Several supervised NMF methods have been proposed (26). The first supervised NMF approaches implemented Fisher discriminant constraints into NMF (Fisher NMF), which penalizes high scatter of samples within the same classes and reward high scatter between different classes (27; 28; 29). Other supervised NMF methods use the Frobenius loss (i.e. linear regression) to classify based on exposure (30), with an extension to semi-supervised NMF (31; 32). There are also supervised NMF approaches termed task-driven dictionary learning, that use cross-entropy loss (i.e. logistic regression) in combination with NMF applied to acoustic scene classification (33; 34).

Finally, Lyu et al. proposed supervised negative binomial NMF (SNBNMF) and applied it to learn mutational signatures predictive of cancer subtype. In SNBNMF, NMF is integrated with a support vector machine (SVM) (35). SNBNMF models mutation counts using a negative binomial distribution. However, this might allow the model to learn to classify samples based on mutation count instead of the actual mutation patterns. We reason that normalized counts should be used instead. As a result, SNBNMF is an unsuitable model for our problem, since normalized counts do not follow the negative binomial distribution.

Here we propose Supervised Non-negative Matrix Factorization (S-NMF), where we combine mutational signature decomposition using NMF, and sample classification using multinomial logistic regression in a single model. We do so by combining the reconstruction loss of NMF with the categorical cross-entropy loss of a logistic regression, resulting in

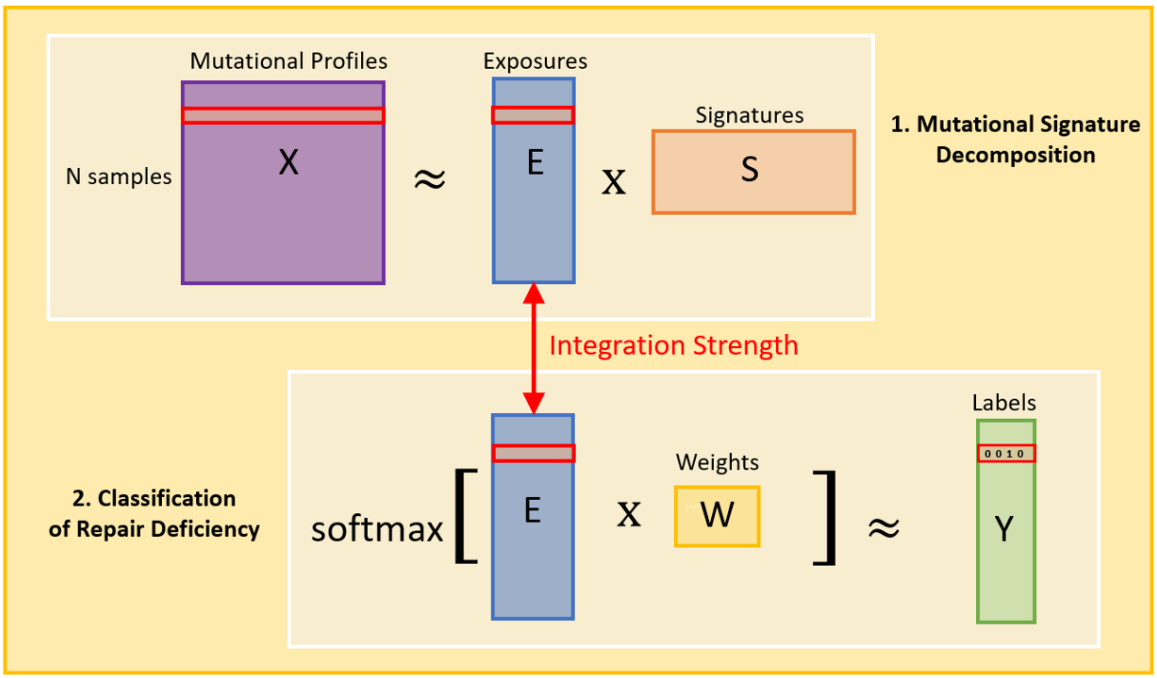


Fig. 2: **Schematic overview of S-NMF.** Integrating 1) Mutational signature decomposition, where the input mutational profiles are decomposed into exposures and signatures. 2) Prediction of DNA repair pathway deficiency, where the exposure are input to a logistic regression to make predictions about the repair pathway deficiency (labels, $\boldsymbol{Y}$)

an integrated loss function. We derive the corresponding update formulas to optimize the S-NMF model using gradient descent, and describe the procedure to train and evaluate an S-NMF model. We further investigate the effect of the hyperparameters on the S-NMF model, assess the effect of the integration on the prediction performance and the signatures, and interpret signatures obtained for prediction of DNA repair pathway deficiency.

2 Methods

Our aim is to learn a single model that is simultaneously able to (i) identify mutational signatures and their corresponding contributions to the mutational profile of a sample (also termed exposures), and (ii) make a prediction about the repair pathway status of the sample using the exposures to those signatures.

2.1 Learn mutational signatures of DNA repair deficiencies

To fulfill both goals, the model is learnt by jointly optimizing: the decomposition of the mutational profiles into signature and exposure latent spaces using non-negative matrix factorization, and the supervised learning of a linear combination of the exposures for the prediction task using multinomial logistic regression. We refer to this method as S-NMF, which stands for supervised non-negative matrix factorization.

2.1.1 Definitions

We define the input matrix of mutational profiles $\boldsymbol{X} \in [0,1]^{N \times T}$, where N is the number of input samples for which mutations have been profiled, and T is the number of mutation types used as input features. Each entry $x_{n,t}$ represents the relative frequency of mutation type t in the mutational profile of sample n. Each mutational profile, corresponding to a row of matrix $\boldsymbol{X}$, is then a probability distribution over all mutation types: $\sum_{t=1}^{T} x_{n,t} = 1, \forall n \in \{1, ..., N\}$. Additionally, we also define a matrix of corresponding DNA repair pathway deficiency status $\boldsymbol{Y} \in \{0,1\}^{N \times O}$, over the O different deficiencies or output classes that we would like the model to predict. The n^{th} row in $\boldsymbol{Y}$ represents the one-hot encoded repair deficiency status of sample n, where $\sum_{o=1}^{O} y_{n,o} = 1, \forall n \in \{1, ..., N\}$.

2.1.2 Problem formulation

Given a matrix of mutational profiles $\boldsymbol{X}$, its respective matrix of DNA repair pathway deficiency status $\boldsymbol{Y}$, and a number of signatures K ($K \leq \min(T, N)$), we aim to learn a model that simultaneously: (i) estimates mutational profiles in $\boldsymbol{X}$ as linear combinations of K underlying mutational signatures (matrix $\boldsymbol{S}$), with corresponding exposures (matrix $\boldsymbol{E}$), and (ii) finds a linear mapping between the exposures of each sample mutational profile in $\boldsymbol{E}$ and the corresponding output class or DNA repair pathway status in $\boldsymbol{Y}$. Once the model is built, it can be used to estimate exposures and make predictions about DNA repair deficiency status based on the mutational profiles of previously unseen samples.

2.1.3 Supervised Non-negative Matrix Factorization (S-NMF)

Mutational signature decomposition. To identify mutational signatures, matrix $\boldsymbol{X}$ is decomposed into two matrices, containing signatures and exposures. The mutational signature matrix $\boldsymbol{S} \in [0,1]^{K \times T}$ contains the relative frequencies of each of the T mutation types for each of the K estimated signatures or mutational patterns shared across samples.The contribution of each signature to each sample is represented in the exposure matrix $\boldsymbol{E} \in [0,1]^{N \times K}$. The mutational profiles in $\boldsymbol{X}$ can be approximated by the matrix product between exposures and signatures.

$$\boldsymbol{X} \approx \boldsymbol{E}\boldsymbol{S}$$

We use non-negative matrix factorization (NMF) to find matrices $\boldsymbol{S}$ and $\boldsymbol{E}$ (17; 36). Namely, we follow the multiplicative update algorithm to optimize the Frobenius reconstruction error of the signature decomposition ($\mathcal{L}_r$, eq. 1), under the constraint that both matrices are non-negative. **Reconstruction loss** (Frobenius reconstruction error):

$$\mathcal{L}_r = ||\boldsymbol{X} - \boldsymbol{E}\boldsymbol{S}||_F^2 \quad (1)$$

Repair deficiency prediction model. To learn a model for prediction of DNA repair deficiency status, we learn a linear mapping between mutational signature exposures $\boldsymbol{E}$ and repair deficiency status (class label) $\boldsymbol{Y}$ using multinomial logistic regression. This procedure estimates an additional matrix of weights $\boldsymbol{W} \in \mathbf{R}^{K \times O}$, with entry $w_{k,o}$ denoting the weight of a signature k with respect to the output class o in the prediction model. The prediction model is learned by minimizing the categorical cross-entropy loss (eq. 2), which quantifies the error between observed (y) and predicted ($\hat{y}$) repair deficiency status for any given sample. To mitigate overfitting to the training data, we include an L2 (ridge) regularization term in the classification loss (eq. 2), where hyperparameter λ_{L2} controls the regularization strength.

Classification Loss (categorical cross-entropy loss (L2-regularized):

$$\mathcal{L}_c = -\sum_{n=1}^{N}\sum_{o=1}^{O} y_{n,o}\log(\hat{y}_{n,o}) + \lambda_{L2}\sum_{w \in \boldsymbol{W}} w^2 \tag{2}$$

Each predicted label $\hat{y}_{n,o}$ (in matrix $\hat{\boldsymbol{Y}} \in [0,1]^{N \times O}$) is obtained by taking the softmax of the product between the n^{th} row of the exposure matrix $\boldsymbol{E}$ and the o^{th} column of the weight matrix $\boldsymbol{W}$ (eq. 3).

$$\text{with} \quad \hat{y}_{n,o} = softmax(\boldsymbol{E}_{n,*}\boldsymbol{W}_{*,o}) = \frac{e^{\boldsymbol{E}_{n,*}\boldsymbol{W}_{*,o}}}{\sum_{o=1}^{O} e^{\boldsymbol{E}_{n,*}\boldsymbol{W}_{*,o}}} \tag{3}$$

Symbol $*$ is a placeholder for all elements along a row or column of a matrix. Note that usually multinomial logistic regression is trained with fixed input data and only the weights are optimized. In contrast, S-NMF also optimizes the exposures (i.e. what would normally be fixed input).

Integrated (total) loss. To integrate both objectives, S-NMF optimizes a combined loss function (eq. 4), defined as the sum of the reconstruction (eq. 1, $\mathcal{L}_r$) and classification (eq. 2, $\mathcal{L}_c$) losses. The hyperparameter λ_c represents the integration strength between the two parts of the model. The larger the value of λ_c, the more influence the prediction objective will have on the exposures (and as a result on the signatures as well).Setting $\lambda_c = 0$ results in non-integrated signature decomposition and prediction.

S-NMF loss function (minimized by S-NMF):

$$\mathcal{L}_{tot} = \mathcal{L}_r + \lambda_c \mathcal{L}_c \tag{4}$$

Model optimization. To learn the S-NMF model, we minimize the total loss $\mathcal{L}_{tot}$ by iteratively applying gradient descent on $\mathcal{L}_{tot}$ with respect to $\boldsymbol{S}$, $\boldsymbol{E}$, and $\boldsymbol{W}$ using the updates in equations 5-7. Symbols η denote the respective learning rates, and ∇ the (partial-) derivatives of the total loss.

S-NMF update rules (for loss minimization by gradient descent):

$$\boldsymbol{S} \leftarrow \boldsymbol{S} - \eta_S \cdot \nabla_{\boldsymbol{S}} \mathcal{L}_{tot} \tag{5}$$

$$\boldsymbol{E} \leftarrow \boldsymbol{E} - \eta_E \cdot \nabla_{\boldsymbol{E}} \mathcal{L}_{tot} \tag{6}$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} - \eta_W \cdot \nabla_{\boldsymbol{W}} \mathcal{L}_{tot} \tag{7}$$

Derivatives. The derivatives of the total loss $\mathcal{L}_{tot}$ with respect to $\boldsymbol{S}$, $\boldsymbol{E}$ and $\boldsymbol{W}$ are given by equations 8-10 (full derivation in Appendix A.1). For $\boldsymbol{S}$, the derivative only has terms from the reconstruction loss, while for $\boldsymbol{E}$ the derivative contains terms from both the reconstruction and cross-entropy losses. For $\boldsymbol{W}$, the derivative has one term from the cross-entropy loss.

$$\nabla_{\boldsymbol{S}} \mathcal{L}_{tot} = -2\boldsymbol{E}^T\boldsymbol{X} + 2\boldsymbol{E}^T\boldsymbol{E}\boldsymbol{S} \tag{8}$$

$$\nabla_{\boldsymbol{E}} \mathcal{L}_{tot} = -2\boldsymbol{X}\boldsymbol{S}^T + 2\boldsymbol{E}\boldsymbol{S}\boldsymbol{S}^T + \lambda_c(\hat{\boldsymbol{Y}} - \boldsymbol{Y})\boldsymbol{W}^T \tag{9}$$

$$\nabla_{\boldsymbol{W}} \mathcal{L}_{tot} = \lambda_c(\boldsymbol{E}^T(\hat{\boldsymbol{Y}} - \boldsymbol{Y}) + 2\lambda_{L2}\boldsymbol{W}) \tag{10}$$

Learning rates. We use the same adaptive learning rates η_S and η_E for optimization of $\boldsymbol{S}$ and $\boldsymbol{E}$ as in (17).

We do not want the integration strength (λ_c) to effect the optimization of the regression weights $\boldsymbol{W}$, to prevent having a derivative of zero when setting $\lambda_c = 0$. Therefor, we divide our wanted constant learning rate μ_W by λ_c. This cancels out the integration strength term (λ_c) in the derivative $\nabla_{\boldsymbol{W}} \mathcal{L}_{tot}$ (eq.10) and leads to a constant learning rate μ_W in the final update formula.

$$\eta_S = \frac{\boldsymbol{S}}{2\boldsymbol{E}^T\boldsymbol{E}\boldsymbol{S}} \tag{11}$$

$$\eta_E = \frac{\boldsymbol{E}}{2\boldsymbol{E}\boldsymbol{S}\boldsymbol{S}^T} \tag{12}$$

$$\eta_W = \frac{\mu_W}{\lambda_c} \tag{13}$$

Multiplicative update formulas. The final multiplicative update formulas are obtained by substituting the derivatives of the loss and the learning rates in the gradient descent update formulas. The complete derivation of the update formulas is included in Appendix A.1 ($\odot$ and division are element-wise).

$$\boldsymbol{S} \leftarrow \boldsymbol{S} \odot \frac{\boldsymbol{E}^T\boldsymbol{X}}{\boldsymbol{E}^T\boldsymbol{E}\boldsymbol{S}} \tag{14}$$

$$\boldsymbol{E} \leftarrow \boldsymbol{E} \odot \frac{\boldsymbol{X}\boldsymbol{S}^T - \frac{\lambda_c}{2}(\hat{\boldsymbol{Y}} - \boldsymbol{Y})\boldsymbol{W}^T}{\boldsymbol{E}\boldsymbol{S}\boldsymbol{S}^T} \tag{15}$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} - \mu_W(\boldsymbol{E}^T(\hat{\boldsymbol{Y}} - \boldsymbol{Y}) + 2\lambda_{L2}\boldsymbol{W}) \tag{16}$$

2.1.4 Non-negativity constraint & stability

The S-NMF update formula for the exposure does not ensure non-negativity as a result of the subtraction of the predicted labels by the true labels ($\hat{\boldsymbol{Y}} - \boldsymbol{Y}$). To ensure the non-negativity constraint, at each exposure update new negative values are set to a very small value close to zero ($1 * 10^{-25}$). This small value is used instead of exactly 0 to prevent unstable solutions as result of getting a denominator equal to zero in the signature update (eq. 14). As a result of ensuring non-negativity of the exposures, the signatures are also ensured to be non-negative since they are updated by a matrix multiplication of non-negative matrices. However, the signatures are defined as a probability distribution over the T mutation types, implicating that each signature should sum to one ($\sum_{t=1}^{T} \boldsymbol{S}_{k,t} = 1, \forall k \in K$). To ensure this, after each update step the signatures are normalized to sum to one.

2.2 Experimental Setup

We implement S-NMF as an extension of the SigProfiler framework (17), which performs standard NMF along with convenient procedures to determine an optimal number of signatures K. The experimental procedure for S-NMF, comprises a training step (Algorithm 1), and a testing step (Algorithm 2). Input to the full experimental procedure are the mutational profiles and labels of the training and test samples ($\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{X_{test}}, \boldsymbol{Y_{test}}$), as well as the 3 hyperparameters (K, λ_c, and λ_{L2}).

2.2.1 S-NMF training

The training procedure consists of two steps. First, the S-NMF model is trained by iteratively updating $\boldsymbol{S}$, $\boldsymbol{E}$, and $\boldsymbol{W}$ according to equations 14-16, resulting in a set of K signatures with corresponding exposures and classification weights. To account for the random initialization of the $\boldsymbol{E}$ and $\boldsymbol{S}$ matrices, the first step of the training procedure is repeated 10 times (Algorithm 1, line 1-2). The 10 sets of K mutational signatures found in the different runs are then clustered. The applied partition clustering

approach (Algorithm 1, line 4) is a variation of K-means clustering, where each cluster gets assigned exactly one signature from each of the 10 runs. The the final set of signatures ($\boldsymbol{S_{final}}$) is obtained by averaging the signatures in each cluster (i.e. centroid of cluster). In the second step of the training procedure, the final signatures $\boldsymbol{S_{final}}$ are then fixed and fitted to obtain the corresponding exposure matrix $\boldsymbol{E}$ and final weight matrix $\boldsymbol{W_{final}}$ (Algorithm 1, line 5). This is done using non-integrated S-NMF (i.e. $\lambda_c = 0$), so that the exposures are not optimized with respect to the classification loss. The final signatures $\boldsymbol{S_{final}}$ and classification weights $\boldsymbol{W_{final}}$ are then stored and used in the testing step.

Algorithm 1: `Training with S-NMF`

Input : $\boldsymbol{X}, \boldsymbol{Y}, K, \lambda_c$, and λ_{L2}
Output: $\boldsymbol{S_{final}}$ and $\boldsymbol{W_{final}}$
1 **for** $i = 1$ *to* 10 **do**
2 $\quad \boldsymbol{S_i}, \boldsymbol{E}, \boldsymbol{W} \leftarrow$ S-NMF $(\boldsymbol{X}, \boldsymbol{Y}, K, \lambda_c, \lambda_{L2})$
3 **end**
4 $\boldsymbol{S_{final}} \leftarrow$ Partition-Clustering $(\boldsymbol{S_1}, \boldsymbol{...S_{10}})$
5 $\boldsymbol{E}, \boldsymbol{W_{final}} \leftarrow$ S-NMF $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S_{final}}, \lambda_{L2}, \lambda_c = 0)$

2.2.2 S-NMF testing

The test procedure comprises two steps. Firstly, the final signatures $\boldsymbol{S_{final}}$ are fitted to the unseen test samples $\boldsymbol{X_{test}}$ using non-negative least squares (NNLS) (37) to find the exposures for the test samples $\boldsymbol{E_{test}}$ (Algorithm 2, line 1). NNLS can be considered a special case of NMF, where for each sample the activity (i.e. exposure) is found with respect to a fixed dictionary (i.e. signatures), with the advantage that NNLS is computationally more efficient than NMF. The exposures of the test samples $\boldsymbol{E_{test}}$, together with the classification weights $\boldsymbol{W_{final}}$ learned in the train set, are then used to calculate the predicted probability of each class for each of the test samples according to equation 3 (Algorithm 2, line 2). This results in $\hat{\boldsymbol{Y}}$ containing a probability distribution for each sample over the possible repair pathway deficiencies (incl. control). The final predicted labels are determined according to the pathway deficiency with the highest probability $\left(\arg\max_{o \in \{1,...,O\}} \hat{y}_{n,o}, \forall n \in \{1, ..., N\}\right)$.

Algorithm 2: `Testing with S-NMF`

Input : $\boldsymbol{X_{test}}, \boldsymbol{Y_{test}}, \boldsymbol{S_{final}}, \boldsymbol{W_{final}}$
Output: $\hat{\boldsymbol{Y}}$
1 $\boldsymbol{E_{test}} \leftarrow$ NNLS $(\boldsymbol{X_{test}}, \boldsymbol{S_{final}})$
2 $\hat{\boldsymbol{Y}} \leftarrow$ Softmax $(\boldsymbol{W_{final}}, \boldsymbol{E_{test}})$

2.3 Evaluation

We evaluate the performance of S-NMF using three metrics: prediction accuracy on the test samples, stability of the signatures found over the 10 training runs, and reconstruction error on the test samples.

2.3.1 Accuracy of DNA repair deficiency prediction

We use accuracy measure classification performance, that is, how close the S-NMF model is to predicting the true DNA repair pathway deficiencies. The accuracy is the number of true positive and true negative test samples (# Correct predictions) divided by the total number of test samples (N_test) ($Accuracy = \frac{\text{\# Correct predictions}}{N_test}$).

2.3.2 Average stability of identified mutational signatures

Stability is used to evaluate the reproducibility of the signatures. The stability is calculated using the clusters obtained via partition clustering of the signatures found in the 10 training runs ($\boldsymbol{S_1}, ..., \boldsymbol{S_{10}}$). Specifically, stability is defined as the average silhouette width of the K clusters of signatures ($Cluster_1, ..., Cluster_K$), obtained as described in section 2.2.1.

$$\text{Stability} = \frac{1}{K} \sum_{k=1}^{K} \text{Silhouette}(Cluster_k)$$

With each cluster ($Cluster_k$) containing exactly 10 signatures, the silhouette width of a cluster is defined as:

$$\text{Silhouette}(Cluster_k) = \frac{1}{M} \sum_{i \in Cluster_k} \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Where $a(i)$ is the intra-cluster distance and $b(i)$ the mean nearest cluster-distance (38). The silhouette width ranges between -1 and 1, with 1 indicating that the model is consistently finding the same signature in each run, while a lower silhouette width (i.e stability) indicates a lower reproducibility of the signature. The cosine similarity is used as distance measure between mutational profiles/signatures. Given two mutational profiles/signatures (A and B) the cosine similarity is $d(A, B) = cossim(A, B) = \frac{A \cdot B}{||A|| ||B||}$. The cosine similarity is defined within the range of -1 to 1, with 1 indicating two completely similar profiles.

2.3.3 Reconstruction error of profiles given exposures & signatures

The reconstruction error is used to evaluate how well the mutational profiles of the test samples can be reconstructed from the exposures and (fixed) signatures. The Frobenius norm is used to calculate the reconstruction error, $||\boldsymbol{X} - \boldsymbol{ES}||_F^2$. A low reconstruction error indicates an accurate description of the original data by the reconstructed mutational profiles $\hat{\boldsymbol{X}} = \boldsymbol{ES}$. The downside of the reconstruction error is that choosing a larger number of signatures K tends to lead to a lower reconstruction error, as there are more latent dimensions (or degrees of freedom) to fit to the input data. Therefore, we consider the reconstruction subordinate to the average stability for measuring the performance of signature decomposition.

2.4 Benchmark models

We assess the integrated S-NMF model that jointly optimizes the mutational profile decomposition and DNA repair prediction goals (S-NMF with $\lambda_c > 0$) against two benchmark models: 1) direct logistic regression model, learnt directly from the mutation profiles; and 2) non-integrated exposure-based prediction (S-NMF with $\lambda_c = 0$).

Direct logistic regression: To evaluate prediction accuracy of our integrated S-NMF ($lambda_c > 0$), we compare it to a multinomial logistic regression model learnt directly using the mutation profiles $\boldsymbol{X}$ as input, with mutation types instead of signature exposures (used by S-NMF) as input features. The direct logistic regression therefore learns a coefficient matrix $\boldsymbol{W} \in \mathbf{R}^{T \times O}$, with entry $w_{t,o}$ denoting the weight of a mutation type t with respect to the output class o. Like in S-NMF, an (L2) regularization term was included to mitigate overfitting. This hyperparameter was determined using cross-validation (Methods 2.6).

Non-integrated exposure-based prediction To evaluate the effect of integrating the mutational signature decomposition and classification objectives, we compare the integrated S-NMF ($\lambda_c > 0$) to a non-integrated exposure-based prediction model learnt using S-NMF with the integration strength set to zero (i.e. $\lambda_c = 0$). This approach is equivalent to existing non-integrated approaches that first apply mutational signature decomposition to learn signatures and exposures, and then learn a classification model using the previously learnt exposures as features.

2.5 Data and Preprocessing

To evaluate the S-NMF algorithm, we use data generated by Zou et al.,(16), comprising mutation profiles for 42 human-induced pluripotent stem cell (hiPSC) lines with individual (biallelic) knockouts across 42 different genes involved in 12 DNA repair pathways. Knockout of gene *ATP2B4*, unrelated to DNA repair, was included as a control. Each sample was labeled according to the main repair pathway associated with the knocked out gene (or control). The knockouts (KOs) were induced using CRISPR-Cas9. For each gene KO cell line, multiple replicates were cultured (between 2 and 8), resulting in a total of 173 samples. Samples were cultured for 15 days to let mutations accumulate, after which they were processed and submitted for whole-genome sequencing (WGS). Sequenced genomes were aligned to the human reference GRCh37/hg19, and the CaVEMan algorithm [REF] was used to call somatic substitutions.

2.5.1 Mutation profile and mutation types

We focused on single base substitutions (SBS) as defined in (17). The SBS mutation types only consider the pyrimidines as reference bases (cytosine C or thymine T), since their complementary bases guanine G and adenine A represent the same SBS on the opposing strand. This results in 6 distinct SBSs (C>A, C>G, C>T, T>A, T>C, T>C), which are further specified by taking into account the two neighboring 5' and 3' bases as additional sequence context, resulting in 16 possible trinucleotide sequence contexts. Taken together, a mutation profile is characterized by 96 SBS (i.e. $T = 6 \times 16$) mutation types. To avoid that our model mainly exploits the total mutation count to classify the samples, we normalize the mutational profiles (i.e. divided the count of each mutation type by the total overall mutation types). The resulting mutational profiles represent a probability distribution over the 96 possible mutation types.

2.5.2 Selection of samples with distinctive mutation profiles

To allow for an informative evaluation of our model, we sought to select gene KOs from Zou et al. (16) that accumulated a sufficient number of mutations, and thus resulted in a reasonably distinctive mutational profile. We considered two criteria: mutation count, and cosine similarity between the mutation profile of the sample and the average mutational profile of the control samples. Distinctive mutational profiles were identified by plotting these two measures, and identified as those either having high mutation count and/or low cosine similarity with the control samples. Finally, we selected the gene KOs for which all replicates had a distinctive mutational profile (in addition to the control samples).

2.5.3 Reference COSMIC signatures

As preliminary validation, signatures found by S-NMF were compared to cancer-related signatures from the Catalogue of Somatic Mutations in Cancer (COSMIC), which were previously extracted using SigProfiler (17) from 2780 whole genomes by the Pan-Cancer Analysis of Whole Genome (PCAWG) Consortium (25, 24). The COSMIC database (v3.2) contains 78 signatures, some of which with proposed aetiology.

2.6 Data Augmentation and Hyperparameter Optimization

For the evaluation of the model we divided the data into disjoint train and test sets. The test set contained exactly one replicate of each gene KO and 2 replicates of the control samples. The remaining samples in the train set were then further divided into 3-folds for cross-validated hyperparameter optimization (section 2.6.2). The replicates per gene KO were evenly spread over the 3 folds (Supplementary fig. S2).

2.6.1 Data augmentation by bootstrapped oversampling

Since we only used the gene KOs that showed a distinctive mutation profile, the final number of available samples was limited. Therefore, we applied bootstrapped oversampling to generate additional artificial samples to improve the robustness of model learning. Each artificial sample has a high similarity to the gene KO of the original real sample but is slightly different because of the random bootstrapping. Bootstrapping was done per class (or repair deficiency) so that we could ensure a balanced dataset with an equal number of samples per class.

Bootstrapping procedure We predefined the number of samples per fold (N) and we have O classes. For each class we randomly drew ($\frac{N}{O}$ - number of real samples for that class) samples from the real sample of that class with replacement (i.e. oversampling). From the mutational profile of each of these samples (i.e. probability distribution over mutation types), we randomly drew the same number of mutations as the corresponding real sample had (i.e. bootstrapping). The counts of these new bootstrapped samples are then normalized so the mutational profiles sum up to 1. Finally, the bootstrapped and real samples are combined making a total of N samples with balanced class.

Choice of number of bootstrapped samples (train set) The downside of using more (bootstrapped) samples is that it makes the training of the S-NMF model slower. To balance this with the benefits of bootstrapped oversampling we settled for a total of 400 samples (real + bootstrapped) per fold. We applied the bootstrapping procedure to each of the 3 cross-validation folds. So during the cross-validation, each model is trained on two folds together having 800 samples and tested on the remaining validation fold with 400 samples.

Choice of number of bootstrapped samples (test set) To make obtain an informative and robust estimation of the prediction accuracy we also bootstrapped the test set according to the same procedure. Since the computational complexity of the test procedure is lower, the decrease in speed by adding bootstrapped sample is less of a problem. Therefore we decided to bootstrap up to 4000 samples for the test data. This will give the accuracy measurement a higher resolution and consistency.

2.6.2 Hyperparameter optimization via 3-fold cross-validation

The S-NMF model has three hyperparameters that need to be determined, namely the number of signatures K, integration strength λ_c, and L2-regularization strength λ_{L2}. We performed 3-fold cross-validation on the training data to find the optimal setting for these hyperparameters. We first determined the number of signatures by selecting the value of K that resulted in the highest prediction accuracy without a decrease in average stability. This should result in the number of signatures that best represent the patterns underlying the mutation profiles. Then, we selected the integration and regularization strengths. These two hyperparameters are dependent on each other, therefore the final values for both were chosen simultaneously. Since this is a multi-objective optimization (aiming for high accuracy and stability), it is not guaranteed to result in a single optimal setting. We therefore looked at the set of pareto optimal settings , containing all hyperparameter settings for which there did not exist another setting with both higher accuracy and higher stability. The optimal pareto setting represents a trade-off between accuracy and stability, where the best setting can be chosen based on giving priority to prediction accuracy or stability of mutational signature decomposition. Using the optimal hyperparameter settings from the cross-validation, we trained the final integrated S-NMF model on the entire train set.

3 Results and Discussion

To evaluate the usefulness of the proposed S-NMF method, we focused on three main aspects. First, we investigated the relationship between mutation profiles and their assigned DNA repair deficiency labels using exploratory data analysis. Second, we quantitatively assessed the performance of S-NMF to both identify latent mutational signatures shared

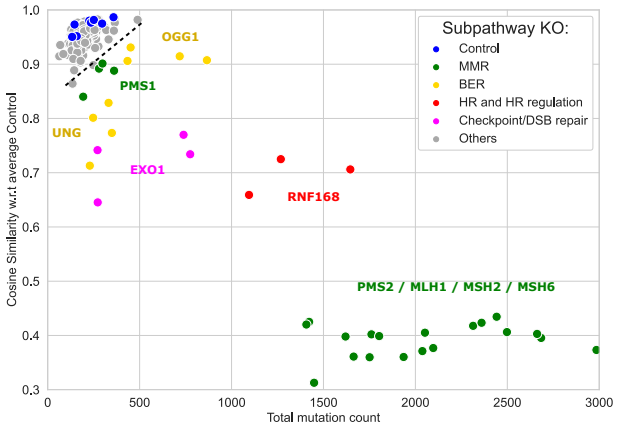


Fig. 3: **Distinctiveness of gene KO samples.** For each sample (dot), the plot shows the total mutation count (horizontal axis), and cosine similarity with average profile of control samples (vertical axis). Samples are colored according to repair pathway deficiency. The black dashed line indicates the threshold on the distinctiveness of mutation profiles, only gene KOs for which all samples were found to the right/below the dashed line were selected.

across mutation profiles, and predict DNA repair deficiency status for each mutation profile. Third, we analyzed the signatures of DNA repair deficiency identified by S-NMF in more detail to provide further biological interpretation and preliminary validation.

3.1 Mutation Profiles and DNA Repair Deficiencies

We identified which gene KOs in the data generated by Zou et al.(16) were sufficiently distinctive to be used to train the S-NMF model. We analyzed their mutational profiles and investigated how their associated DNA repair deficiency labels relate to existing literature on the functional role of the genes. In addition, we assessed the quality of the bootstrapped samples we generated to augment the original dataset.

3.1.1 Nine gene knockouts showed distinctive mutation profiles

We analyzed the distinctiveness of the mutational profiles obtained for the different samples of each gene KO (Section 2.5.2), to identify the KOs that would be informative to train the S-NMF model. We found that 9 gene KOs were sufficiently distinguishable from the control samples. The 9 gene KOs were annotated with 4 different repair pathway deficiencies, namely: mismatch repair (MMR) pathway (*MSH6, MSH2, MLH1, PMS2* and *PMS1*), base excision repair (BER) pathway (*UNG* and *OGG1*), homologous recombination (HR) repair pathway (*EXO1*), and double strand break repair (DSB) pathway (*RNF168*).

3.1.2 Mutation profiles and labels of distinctive gene KO samples

For each selected gene KO, we averaged the mutational profiles overall replicates and visualized the mutational profiles (Fig. 4). We analyzed these mutational profiles, related the gene KOs to existing literature, and decided on their final repair deficiency label.

HR deficiency: Two gene KOs, *RNF168* and *EXO1*, had a very similar mutational profile (cossim: 0.96), despite being labeled as different subpathways. *RNF168*, encoding E3 ubiquitin-protein ligase, is related to double-stranded break (DSB) repair. *EXO1*, encoding 5' to 3' exonuclease 1 protein, is related to multiple repair pathways, one of them being the repair of DSB by homologous recombination (HR). Based on the highly similar mutational profile, we assume *RNF168* and EXO1 gene KOs are affected by the same underlying mutational process. So since both genes are related to the HR pathway, we combined RNF168 and EXO1 into a single class labeled HR deficiency.

BER deficiency: We found two distinctive gene KOs in the BER pathway, for *OGG1* and *UNG*. However, *OGG1* has a relatively high average mutation count (617) but is still relatively similar to the control mutational profile (avg. cossim: 0.92). Contrarily, *UNG* has a lower average mutation count (289) but the mutational profile is less similar to the control (avg. cossim: 0.80). Even though *UNG* and *OGG1* relate to 2 different subpathways within BER, we still label both gene KO as BER to evaluate if the models can capture sub pathways in a repair deficiency since in other cases it might not be known that the gene KOs relate to two different sub pathways.

MMR deficiency: We found 5 distinctive MMR deficient gene KOs and which on average also had a very high total mutation count (1652). This can be explained by the fact that the MMR pathway repairs base mismatches commonly caused by DNA replication errors, one of the main endogenous DNA damaging factors. The replication errors are recognized by a MutS(α) or MutS(β) complex, respectively comprising MSH2/MSH6 and MSH2/MSH3 (39). Subsequently, MutS recruits the MutL complex, of which there are three variants. The most prominent variant in human cells is MutL(α), comprising MLH1 and PMS2 (40). In the other variants, MutL(β) and MutL(γ), MLH1 forms a complex with PMS1 and MLH3 respectively. Especially the *MLH1*, *MSH2*, and *MSH6* gene KOs had a very high average mutation count (1886, 2237, 2317) (Supplementary Fig. S1a) and similarity amongst each other (avg. cossim: 0.99). The less prominent role of MutL(β) could explain the relatively low average total mutation count of the *PMS1* gene KO (283).

3.1.3 bootstrapped-oversampling generates artificial samples with high similarity to corresponding real samples

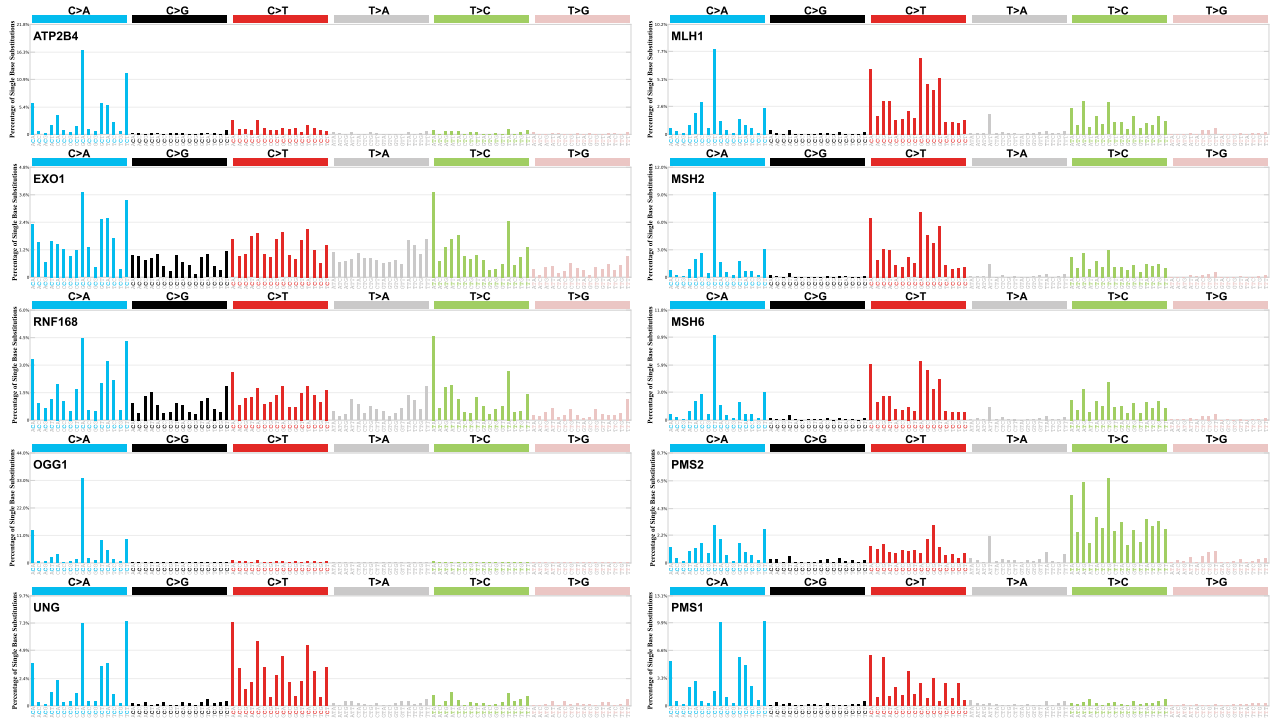
Since the size of the small dataset limited the ability to train robust models, we augmented it using bootstrapped oversampling (Section 2.6). We performed principal component analysis (PCA) to visually represent the relationship between real and bootstrapped training samples (Fig. 5, with samples colored by DNA repair deficiency label, where bootstrapped samples are lighter in color than real samples). Bootstrapped samples clustered together with the gene KO they were sampled from, as expected. We were also able to confirm that *RNF168* and *EXO1*, which we previously labeled as HR deficiency samples, indeed clustered together (Fig. 5, red). On the other hand, *OGG1* and *UNG* did not cluster together, forming two subpathways within BER (see Fig. 5, orange). We also saw that *PMS1* (Fig. 5, green cluster near control) was not as distinctive from the control samples as the other MMR gene KOs. Additionally, PMS2 did not cluster together with the other MMR gene KOs, yet it was distinctive from the control samples (Fig. 5, most right green cluster).

3.2 S-NMF Reconstruction and Prediction Performance

To better characterize the behavior and capabilities of the S-NMF model, we first analyzed the effect of the three hyperparameters (K, λ_c, and λ_{L2}) on the performance of the model. Additionally, we compared the performance of the final S-NMF trained using the best hyperparameters against two benchmark models. Lastly, we investigated factors that could be limiting the performance of S-NMF.

3.2.1 Optimal number of signatures guided by stability and accuracy

The number of signatures K had the largest effect on the performance metrics: prediction accuracy, average stability, and reconstruction error (Fig. 6, Supplementary Fig. S3). Therefore, we evaluated and selected the number of signatures, based on which we then analyzed the effect of the integration and regularization strengths. The search space for K was set



(a) (b)

Fig. 4: **Average mutational profiles** for the 9 distinctive gene KOs and control samples.

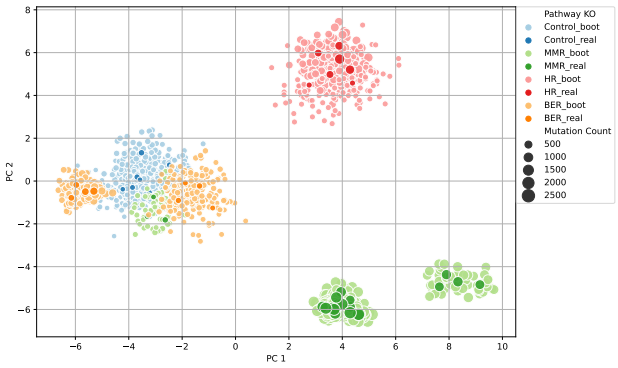


Fig. 5: **PCA plot of bootstrapped and real training samples**. Samples (dots) are projected according to the two main axes of variation, represented by PC1 and PC2. The color indicates repair deficiency. Darker dots represent the real samples, while lighter dots are bootstrapped samples. The dot size depicts the total mutation count. *RNF168* and *EXO1* form the HR cluster. *OGG1* (orange, left cluster) and *UNG* (orange, right cluster) are two subpathways within the BER pathway. *PMS1* (green, border control cluster) is not as distinct as other MMR samples, and *PMS2* (green, rightmost cluster) does not cluster together with other MMR gene KOs.

from 3 to 7 signatures. The cross-validation results showed that increasing K from 3 to 7 had a positive effect on the median prediction accuracy (0.78 vs 0.93), and on the median reconstruction error (5.74 vs 4.57). This was expected since more signatures (i.e. latent dimensions) allow for more degrees of freedom to fit the data. This effect became less pronounced above 5 signatures. Additionally, the median average stability improved when increasing K from 3 to 5 signatures (0.67 vs 0.99). However, using more than 5 signatures decreased the average stability (median 0.70 for $K = 7$). We therefore chose to use 5 signatures ($K = 5$), since this resulted in the most stable and reproducible signatures.

3.2.2 Larger integration strength can improve prediction performance

After fixing the number of signatures to 5, we further analyzed the effect of the integration and regularization strengths. These two hyperparameters had a much smaller effect on the performance compared to the number of signatures (Fig. 6). Nonetheless, increasing the integration strength improved the prediction accuracy from median accuracy of 0.906 with no integration ($\lambda_c = 0.0$) to 0.931 ($\lambda_c = 0.5$) (Fig. 6, second column, first row). However, further increasing the integration strength deteriorated the avg. stability of the signatures and reconstruction error. This is in line with the expectations since a larger integration strength shifts the importance from the signature decomposition more towards the classification performance during the model optimization. Taken together, there is an optimal integration strength between 0.1 to 0.5 where the trade-off between prediction accuracy and avg. stability is balanced.

3.2.3 Large regularization strengths lead to S-NMF underfitting

Using a small regularization strength slightly increased the median accuracy from 0.920 with no regularization to 0.938 with $\lambda_{L2} = 0.0001$. However, too strong regularization not only led to a decrease in median accuracy (0.899 for $\lambda_{L2} = 0.01$), it also worsened the stability and reconstruction error. This indicated that even though only the classifier weights are directly affected by the regularization strength, this effect is propagated downstream to the mutational signature decomposition. The intuition behind this is that a higher regularization strength makes the model simpler, thus limiting the extent to which the model can be optimized to accurately predict the training data. Since in S-NMF the exposures are trainable, using a high regularization in S-NMF may result in some exposures contributing very little to the model, where the remaining exposures will be forcibly optimized to still try to make as accurate predictions on the training data as possible. Changing the exposures subsequently leads to changes in the signatures, as these are jointly

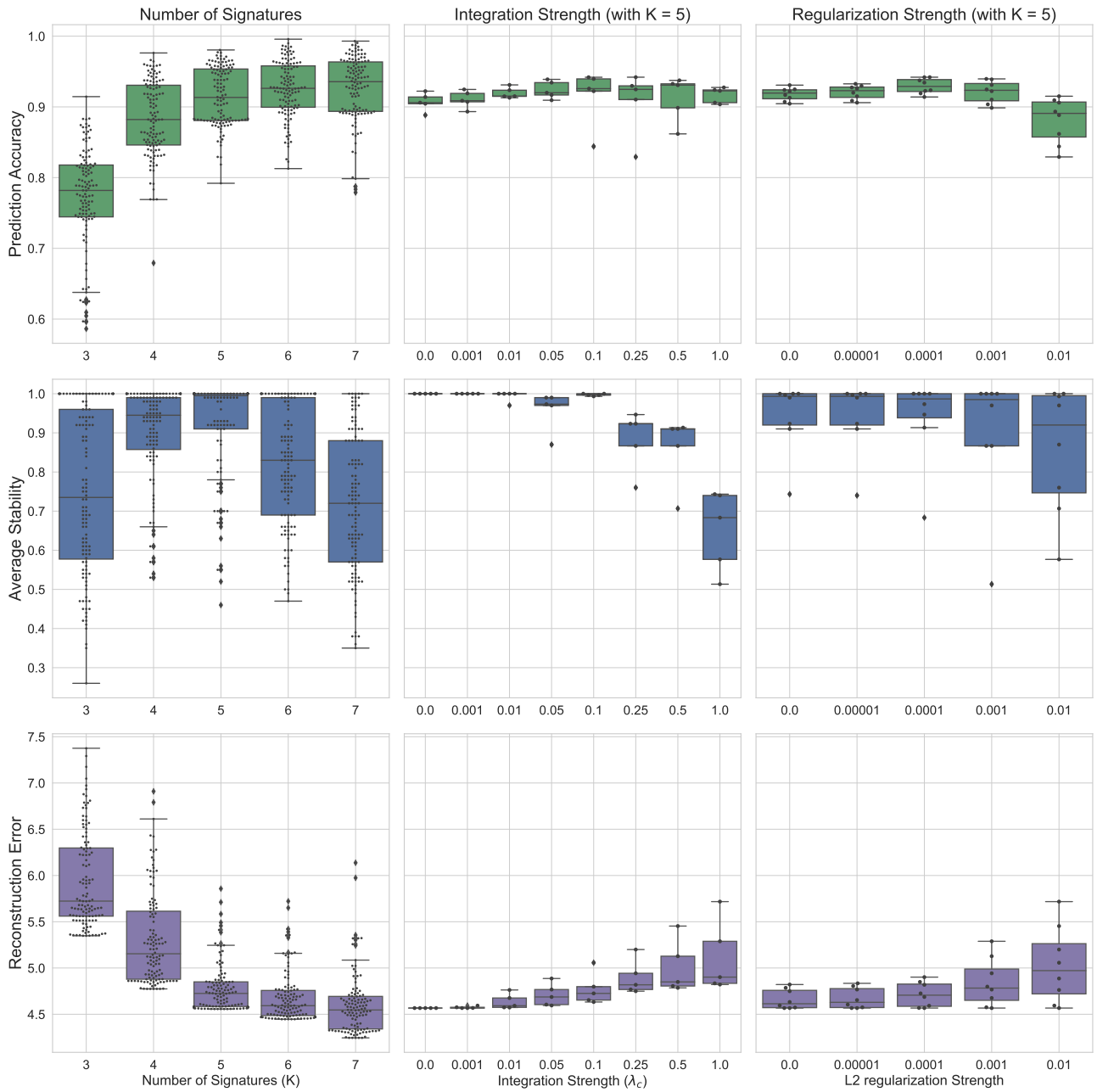


Fig. 6: **Boxplots summarizing hyperparameter optimization results of cross-validation.** Each dot represents the average of the three folds for one setting of hyperparameters. Rows show: (top row) prediction accuracy, (middle row) average stability of the signatures (over 10 training runs), (bottom row) reconstruction error of the final mutational signature decomposition. Columns show: (left column) effect of number of signatures K on performance, for all settings of integration and regularization strengths, (middle column) effect of integration strength on performances for all settings of regularization strength (with number of signatures $K = 5$), and (right column) effect of L2 regularization strength (5 settings) on performances, for all settings of integration strength (7 points).

optimized by the matrix decomposition part of the S-NMF algorithm, leading to worse reconstruction error and average stability (Fig. 6, right, bottom two rows). This effect especially occurred in combination with a high integration strength, which increases the influence of the classification on the exposures even more. Consequently, we need to balance the regularization strength in combination with the integration strength to improve the prediction accuracy while avoiding a drop in stability.

3.2.4 Integration and regularization balance accuracy and stability

To select the final hyperparameters for integration λ_c and regularization λ_{L2} after fixing $K = 5$, we analyzed the tradeoff between prediction accuracy and average stability. We looked at the pareto optimal solutions per number of signatures K (Fig. 7). The optimal settings together form a front per K, where the best performing setting is the closest to the theoretically perfect model with both accuracy and average stability equal to 1.0, Fig. 7 black ○). We then chose the combination of hyperparameters that achieved the highest average stability (1.0) and prediction accuracy (0.942) during cross-validation for $K = 5$. Specifically, this was

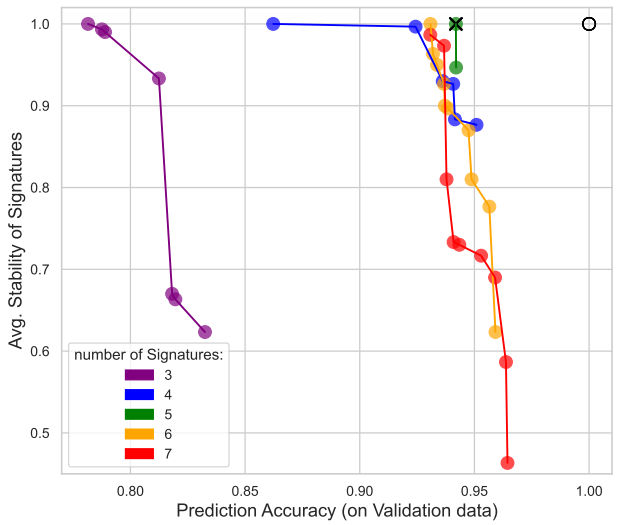


Fig. 7: Pareto optimal settings per number of signatures (K), taking into consideration the prediction accuracy (x-axis) and average stability of the signatures (y-axis). Each dot represents the performance using a certain hyperparameter setting. The settings for which there was a different setting with both superior accuracy and average stability are not shown (i.e. non-pareto settings). The lines connecting the settings for each number of signatures then represent the pareto optimal front which visualized the trade-off between accuracy and avg. stability. The black cross indicates the setting of the final model. The Unfilled black dot indicates the theoretical perfect setting.

integration strength $\lambda_c = 0.1$ and a regularization strength of $\lambda_{L2} = 0.0001$ (Fig. 7, black $\times$). Additionally, the pareto plot showed that setting the number of signatures K to 5 resulted in the pareto front closest to the theoretical optimum (Fig. 7, green), confirming that setting $K = 5$ resulted in the most stable signatures while still achieving a good accuracy.

3.2.5 Integrated S-NMF provides comparable prediction accuracy

Integrated S-NMF achieved high prediction accuracy (0.971), comparable to that of both non-integrated S-NMF (0.966) and direct logistic regression (0.968). This was accompanied by a decent mutational signature decomposition performance not very far from non-integrated S-NMF (average stability: 0.99 vs 1.0 and reconstruction error 4.57 vs 4.38) (Table 1). Since direct logistic regression does not perform signature decomposition, stability and reconstruction error could not be measured. This underlines the main drawback of direct logistic regression, which is not as easily relatable to the underlying biological processes as the S-NMF model.

3.2.6 Integrated S-NMF is sensitive to local optimal solutions

We analyzed the training curves, or the development of the loss function over the training epochs, for the 10 runs (Supplementary Fig. S4). The total loss was further broken down into its reconstruction loss and cross-entropy loss components. For the non-integrated S-NMF, all 10 runs converged to the same minimum for both the reconstruction and classification losses. However, the non-integrated S-NMF was only able to achieve a training prediction accuracy of 0.96 on average, whereas the integrated S-NMF immediately converged to a training accuracy of 1.0. Compared to the non-integrated S-NMF, the integrated S-NMF ($\lambda_c = 0.1$) converged to a solution with much lower cross-entropy loss, and slightly higher reconstruction loss (Supplementary Fig. S4, Middle). Furthermore, for the integrated S-NMF, not all runs converged to the same minimum. One run had both an increased reconstruction and cross-entropy loss. The remaining nine runs converged to two different sub-optima. Some runs resulted in a slightly lower reconstruction error with a slightly higher cross-entropy loss compared to other runs. However, when combining both losses, these nine runs showed a similar total loss. Similar effects could be seen for other settings with high accuracy but decreased average stability, for example when increasing the number of signatures. With $K = 7$ and the same settings as for the integrated S-NMF ($\lambda_c = 0.1$ and $\lambda_{L2} = 0.0001$), we saw that the losses of the runs converged to multiple, but very similar, (sub-)optima (Supplementary Fig. S4, Right). We hypothesize that these local optimal solutions for the different runs relate to the decrease in average stability, since the final signatures are averaged over the signatures from the 10 runs.

Model:	Prediction Accuracy	Avg. Stability	Reconstruction Error
Direct Logistic Regression	0.968	-	-
Non-integrated S-NMF	0.966	**1.0**	**4.38**
Integrated S-NMF	**0.971**	0.99	4.57

Table 1. Performance comparison of final models.

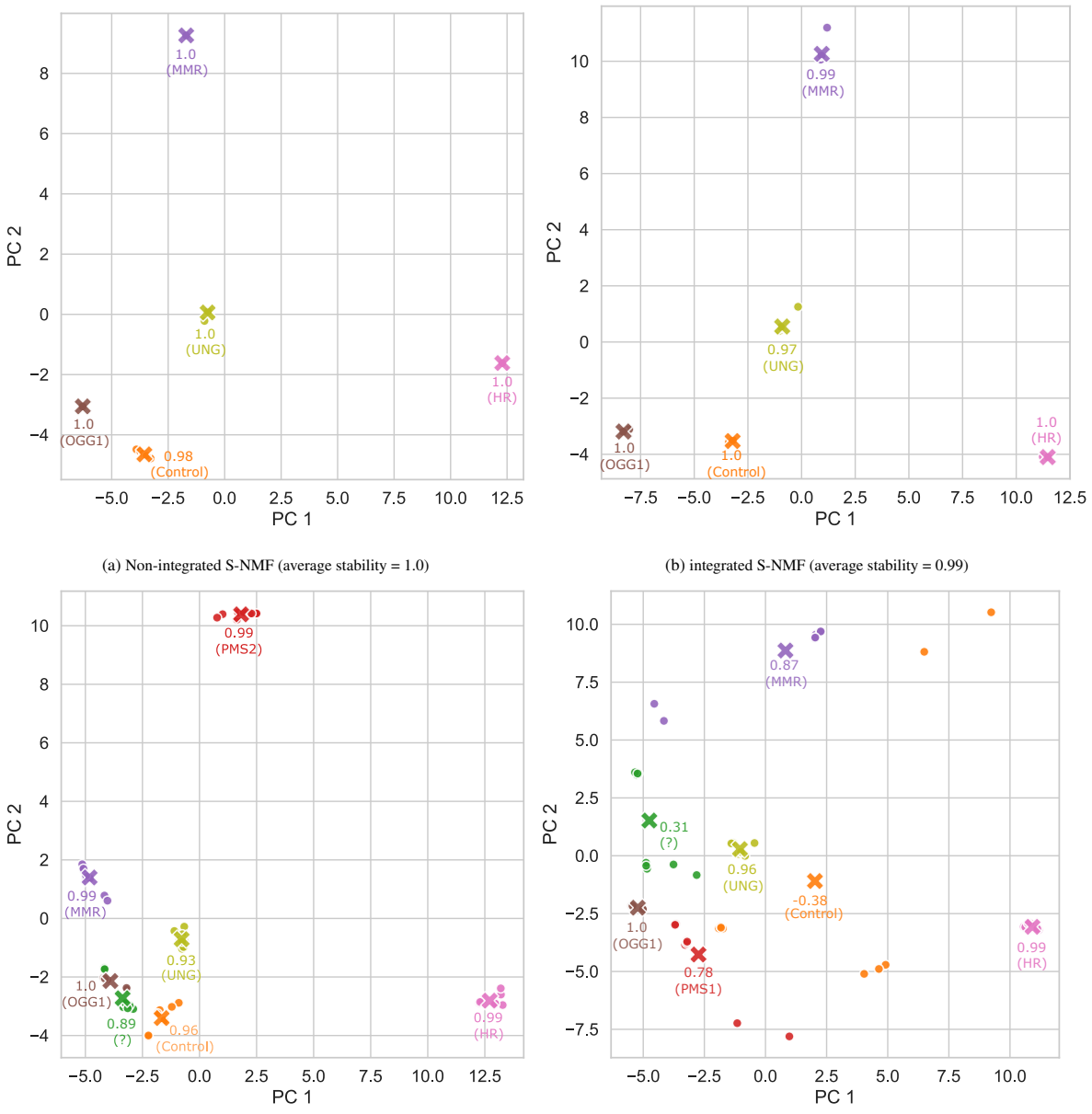
3.2.7 Decrease in stability could partially be attributed to clustering of runs.

We further investigated the local optimal runs seen in the learning curves (Section 3.2.6) and their relation to the drop in average stability by analyzing the signatures from the different runs. To allow for visualization, a PCA was performed on the signatures from the training runs and the final signatures (i.e. centroids). The signatures were colored according to the partition-clustering and annotated with the stability of the final signatures and the related (sub-) pathway (Fig. 8). It must be noted that only the first two principal components were visualized, which do not capture all variance between the signatures that are defined in the T dimensional 'mutation type space'.

Non-integrated S-NMF resulted in 5 closely grouped clusters with the average stability of 1.0 (Fig. 8a). The integrated S-NMF had one run with slightly different signatures for MMR and the *UNG* subpathway, resulting in average stability of 0.99 (Fig. 8b). This run corresponds to the run that had a significantly higher reconstruction loss as well as cross-entropy loss (Section 3.2.6, Supplementary fig. S4, middle column).

For the less stable setting of 7 signatures, the effect of integration became more pronounced. The non-integrated S-NMF with $K = 7$ resulted in reasonable stable signatures with average stability of 0.95. Five of the seven signatures were highly similar to the signatures previously found with $K = 5$. Of the additional 2 signatures, one was characterized by a G[C>A]A mutation and did not relate to any particular gene KO (Supplementary fig. S5b). The second additional signature captured the T>C mutations previously present in the MMR signatures. Correspondingly, this signature was mainly exposed in the *PMS2* gene KO since these are characterized by the T>C mutations.

The integrated S-NMF with $K = 7$ resulted in a much stronger decreased average stability (0.65) which was mainly caused by the control and two additional signatures (Fig. 8d, orange, red, green). Yet, under the influence of the classification component, integrated S-NMF found an additional signature that was mainly exposed in *PMS1* (Fig. 8d, red; and Supplementary fig. S5a), which resulted in a slightly higher overall accuracy (0.974) compared to the three previous settings with much better average stability. However, each run found different (subsets of) signatures, causing the decreased average stability. This mainly indicates that the underlying mutational patterns captured by the unstable signatures are not pronounced enough.



(a) Non-integrated S-NMF (average stability = 1.0)

(b) integrated S-NMF (average stability = 0.99)

(c) Non-integrated S-NMF with K = 7 (average stability = 0.95)

(d) integrated S-NMF with K = 7 (average stability = 0.65)

Fig. 8: **PCA plot of signatures of the 10 training runs and the final signatures.** Dots (•) represent the 10 training runs, crosses (×) the centroid of each cluster (i.e. the final signatures). The color indicate the clusters found by partition-clustering (coloring correspond to exposure in Fig. 10). Each centroid is annotated with the stability of the cluster and the repair pathway or gene KO in which the final signature had the highest exposure. Top row) $K = 5$. Bottom row) $K = 7$. Left column) non-integrated S-NF. Right column) integrated S-NMF.

However, we believe that using a different clustering technique could potentially mitigate the drop in average stability. The currently used partition-clustering algorithm imposed the constraint that each cluster must contain exactly one signature from each run. However, since not every signature is consistently found in each run, some final signatures were calculated over signatures that seemingly came from different clusters, resulting in unstable and probably less meaningful signatures (Fig. 8d, orange).

The challenge of dealing with less stable signatures could potentially be mitigated to some extent by using a different clustering approach. This challenge was already addressed in non-integrated mutational signature decomposition (41). They showed that hierarchical clustering, partition-clustering around medoids, or clustering with matching as an alternative to the original partition-clustering was better able to recover signatures from simulated mutational profiles. Additionally, they showed that filtering out runs with a higher loss based on a relative tolerance (RTOL) with regard to the run with the lowest loss, further improved the ability to recover more mutational signatures with higher stability. These alternative approaches might be particularly beneficial for S-NMF since integration seemed to negatively impact the average stability.

3.3 Interpretation of S-NMF Repair Deficiency Signatures

In this section, we will further interpret the results generated with the final S-NMF model, evaluate how it relates to existing literature. Additionally we will compare the resulting exposure found by the final S-NMF to

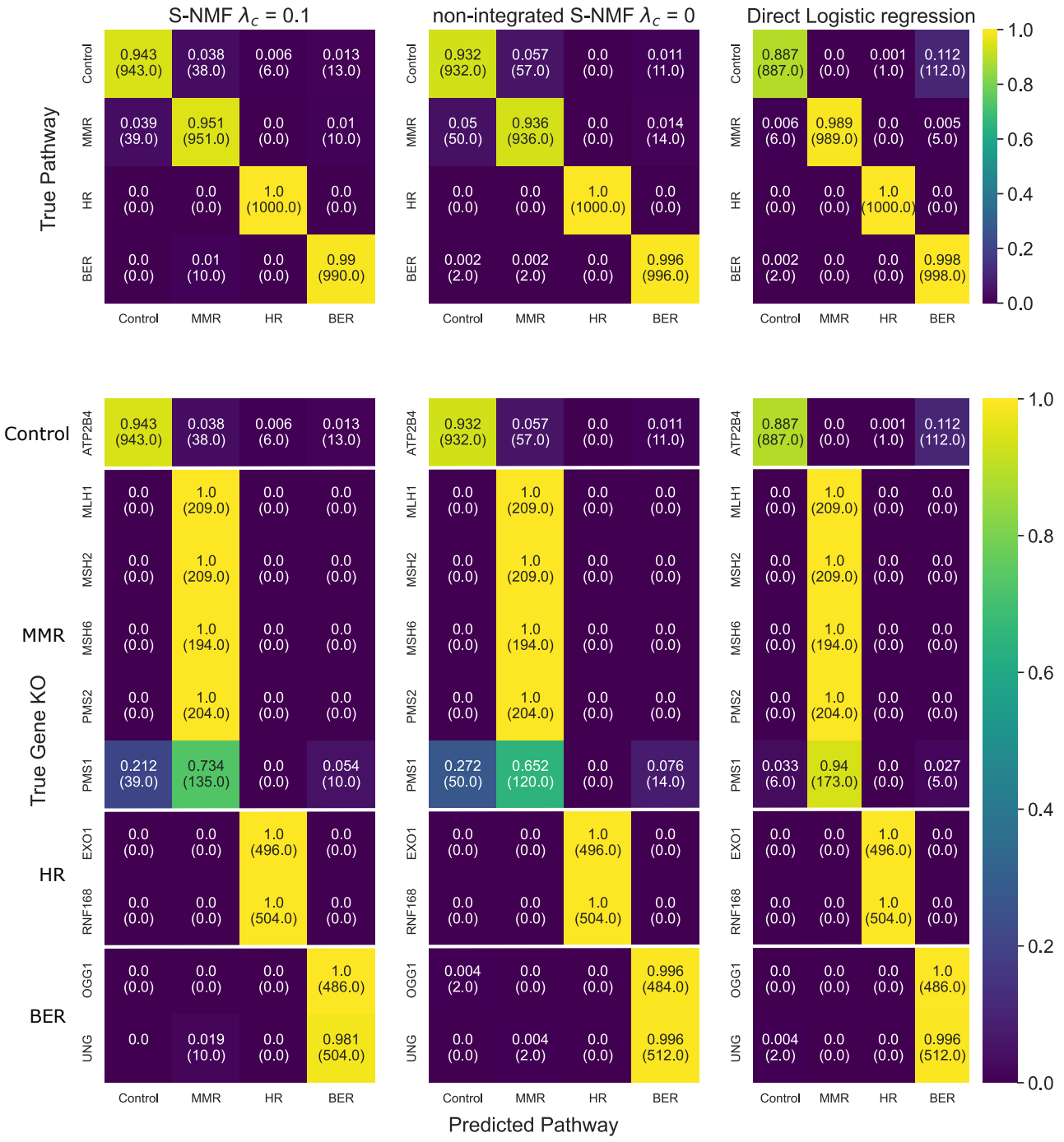
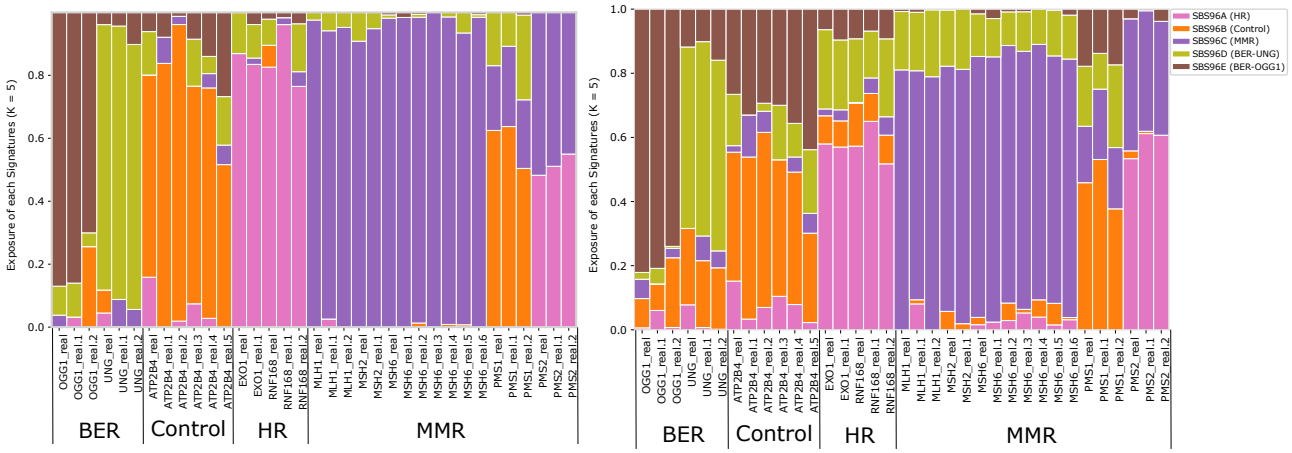


Fig. 9: **Confusion matrices with respect to pathway and gene KO.** Labeled with the percentage of sample that is predicted to the concerned pathway per true pathway/gene KO, the coloring is based on the same percentage. (In parentheses is the total number of sample with the concerned prediction) Overall prediction accuracy for integrated S-NMF is 0.971, non-integrated S-NMF is 0.966, and direct logistic regression 0.968.

the non-integrated S-NMF benchmark model to analyze the effect of integration on the mutational signature decomposition.

3.3.1 PMS1 gene KO is limiting classification performance

To evaluate the performance of the integrated S-NMF model on the individual gene KOs, we looked at per pathway and per gene KO confusion matrices of true vs. S-NMF predicted repair deficiencies (Fig. 9). Most incorrect predictions were made between the control and the MMR pathway. The confusion matrix per gene KO shows that this misclassification comes from the PMS1 gene KO samples, which are mainly misclassified as control (prediction accuracy 0.739). This is in line with what we saw in the PCA of all training samples, where PMS1 samples clustered close to the control samples, instead of the other MMR deficient gene KOs (fig.5, green samples close to control). Aside from PMS1, the other MMR and HR gene KO samples were predicted perfectly (1.0). For the BER pathway, prediction accuracy was also very high (0.99). Furthermore, all real samples were predicted correctly, meaning that misclassifications were made only on the bootstrapped samples.



(a) Exposure found by the integrated S-NMF ($\lambda_c = 0.1$).

(b) Exposure found by the non-integrated S-NMF ($\lambda_c = 0.0$).

Fig. 10: **Exposure of the signatures (K = 5) found by S-NMF**. Left) integrated S-NMF. Right) non integrated S-NMF. Only the real samples are shown to allow for practical visualization. Since we use normalized count as input data the sum of exposures sums to 1. Pink (SBS96A) relates to HR deficiency, Orange (SBS96B)to the control samples, Purple (SBS96C) to MMR deficiency, Green (SBS96D) to the BER-UNG (sub)pathway deficiency, and Brown (SBS96E) to the BER-OGG1 (sub)pathway deficiency.

We further compared the integrated S-NMF to the benchmark models. The non-integrated S-NMF (accuracy 0.966) made similar misclassifications between *PMS1* and control samples with even lower accuracy. Additionally, similar to the integrated S-NMF, all remaining gene KOs were predicted with very high accuracy. This limited the extent to which we could evaluate the effect of integration on the classification performance. The direct logistic regression (accuracy 0.968) achieved similar total accuracy to the non-integrated S-NMF. However, it mainly misclassified control samples as BER deficient, while the PMS1 samples were mostly predicted correctly (0.94).

3.3.2 S-NMF is able to identify DNA repair subpathways

We sought to interpret the signatures and exposures extracted by the integrated S-NMF model, focusing first on the exposures per gene KO (Fig. 10a). S-NMF was able to detect the two subpathways in BER we previously identified in the exploratory data analysis section (3.1.2). Even though *OGG1* and *UNG* were both predicted to be BER deficient with high accuracy, their largest exposures come from different signatures (Fig. 10, brown for *OGG1*; green for *UNG*). The ability to recognize these underlying subpathways is one of the main advantages of an NMF-based approach over the direct logistic regression, or any other supervised learning model that does not aim to represent the underlying patterns with valuable biological interpretation.

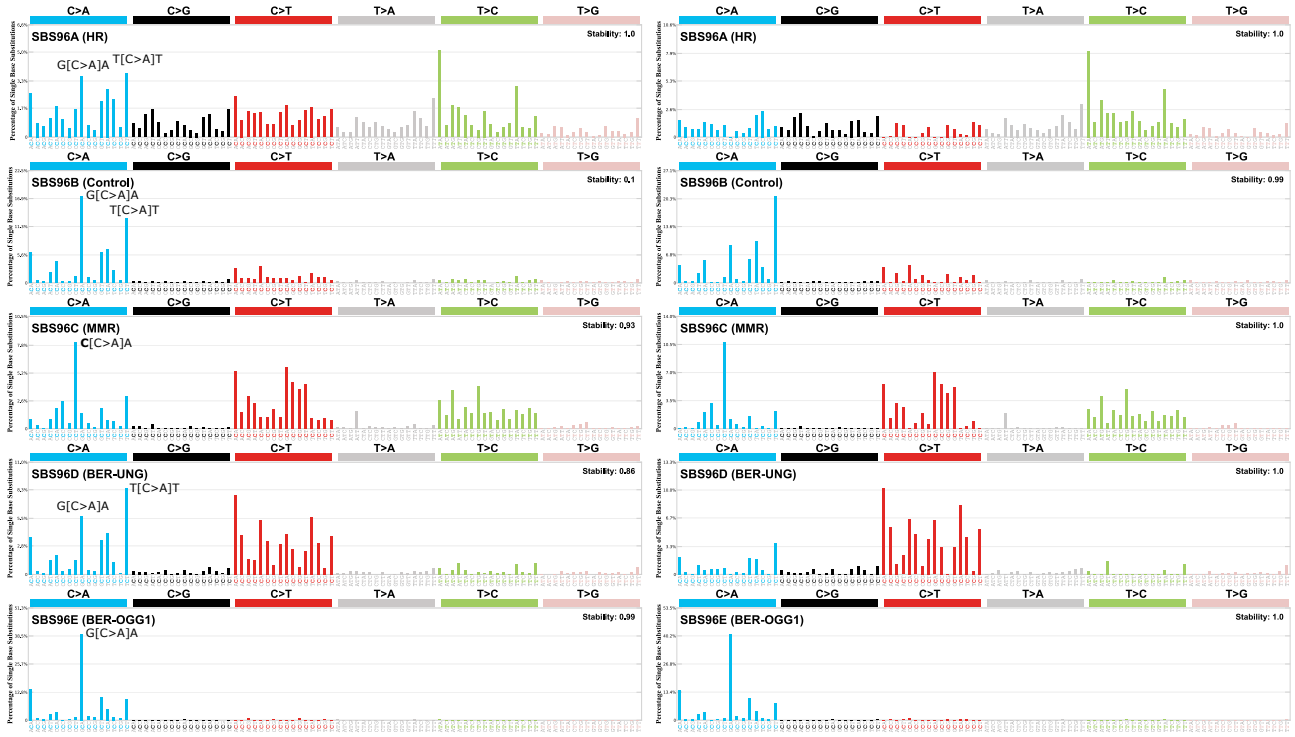
Genes *OGG1* and *UNG* both encode DNA glycosylase enzymes. However, *UNG* recognizes and excises uracil (i.e. deaminated bases). If *UNG* is deficient, unrepaired uracils lead to C>T mutations when the DNA is replicated (16). Whereas *OGG1* recognizes and removes 8-oxoG (i.e. oxidized bases) (42). If *OGG1* is deficient, unrepaired 8-oxoG results in G>T mutation (i.e. C>A mutations in mutational profile) (16). So even though they belong to the same family of enzymes and relate to the same DNA repair deficiency in the BER pathway, the exact type of DNA damage they recognize and repair is different.

For MMR deficiency, the *PMS2* and especially the *PMS1* gene KO are not captured as well by the MMR-related signature (Fig. 10, Purple). As a result, additional signatures are used to decompose their mutational profile. A higher number of signatures (e.g. $K = 7$) would be able to find *PMS1* and *PMS2* specific signatures. However, as the hyperparameter optimization showed, this led to a marked decrease in the average stability of the signatures.

3.3.3 S-NMF signatures are more representative of repair deficiency

To further analyze the effect of the integration in S-NMF, we compared the exposures found by the optimal integrated and non-integrated S-NMF (10b). Assuming that both models still capture similar underlying patterns, we paired each signature found by the non-integrated S-NMF to the integrated S-NMF signature with which it had the highest cosine similarity. This way we can make a direct (qualitative) comparison between the signatures and corresponding exposures. The main difference is that the integrated S-NMF mainly uses one signature as a representation of each sample, where on average the largest contribution signatures per sample had an exposure of 0.80 (Fig. 10a). In contrast, for the non-integrated this was 0.61, indicating a larger contribution of additional signatures to the samples (Fig. 10b). In this context, we could interpret the contribution of multiple signatures as an indication that the cells have multiple repair deficiencies, while we know that they have only one. For example, the exposure of the BER-UNG signature in MMR-deficient samples (Fig. 10b, green). The exposures of unrelated signatures are reduced in the integrated S-NMF compared the non-integrated approach. This suggests that the signatures learned by the integrated S-NMF better capture the underlying repair deficiency.

Besides the exposures, we also compared the signatures found by the integrated S-NMF (Fig. 11a) with the non-integrated S-NMF (Fig. 11b). The main difference is that in non-integrated S-NMF characteristic mutation types (i.e. few specific mutation types with very high probability) are more specific to a single signature. In contrast, in the integrated S-NMF, multiple signatures contained the same characteristic mutations types. This difference is especially clear in the characteristic C to A mutations. For non-integrated S-NMF, the control signature (SBS96B) is characterized by a T[C>A]T and the BER-OGG1 signature (SBS96C) by G[C>A]A. Correspondingly, the HR- (SBS96A) and BER-UNG (SBS96D) signatures have a low contribution of these two mutation types, since these mutations are captured by the (small) exposure of the control- and BER-OGG1 signatures (Fig. 10b, orange and green). In contrast, the integrated S-NMF decomposes each sample mainly into a single signature while the exposures of the other signatures are limited, as we saw in the previous section. Related to this, the signatures for HR and BER-UNG also need to contain these characteristic mutation types found in the control and BER-OGG1. This again could confirm that in the non-integrated S-NMF, (characteristic) mutations are attributed (wrongly) to signatures from a different repair deficiency because this slightly benefits the reconstruction



(a) Signatures found by the integrated S-NMF.

(b) Signatures found by the non-integrated S-NMF.

Fig. 11: **Signatures of the signatures (K = 5) found by S-NMF**. Left) integrated S-NMF. Right) non integrated S-NMF. The signatures of the non-integrated S-NMF were match based on the highest cosine similarity to the integrated S-NMF signatures. SBS96A (HR) has a cosine similarity of 0.834; SBS96B (Control) 0.893; SBS96C (MMR) 0.982; SBS96D (BER-UNG) 0.877; and SBS96E (BER-OGG1) 0.999.

error. In contrast, integrated S-NMF seems to prevent the exposure of unrelated signatures by exploiting the information contained in the repair deficiency labels.

3.3.4 S-NMF signatures akin to known cancer-related signatures

As preliminary validation, we investigated if the signatures identified by S-NMF had any resemblance to known cancer-related signatures, some of which have been previously associated with DNA repair deficiencies. For this, we calculated the cosine similarity between the signatures found by S-NMF and signatures from the COSMIC database (24) (Fig. 12). In this section, we compared the signatures found by S-NMF to the COSMIC signatures with a related aetiology (i.e. related to the same repair pathway deficiency) (Fig. 12, annotated with red square). Additionally, we evaluated the differences between the integrated S-NMF and the non-integrated S-NMF with regard to the similarities to the COSMIC signatures. In general, the results for the integrated S-NMF and non-integrated S-NMF are very similar. This shows that despite the classification influencing the signature decomposition, the resulting signatures still capture the same underlying mutational processes.

MMR deficiency. Seven COSMIC signatures have been suggested to relate to MMR deficiency (43). Of these signatures, six were experimentally validated using pms-2 and mlh-1 gene KOs in C.elegans (44). There is some overlap between several of the MMR-related COSMIC signatures. SBS6 and SBS15 are both characterized by C>T mutations. SBS14 is characterized by C>A mutations with a downstream T (i.e. N[C>A]T) and SBS20 by the more specific C[C>A]T mutation, with also an upstream C of the substitution. Both SBS21 and SBS26 are characterized by T>C mutations and have been more specifically related to the *PMS2* gene KO. (44). The last MMR-related COSMIC signature, SBS44, was identified and experimentally validated in colorectal cancer organoids with a *MLH1* gene KO (45). SBS44 seemingly comprises the characteristics of all previously described MMR-related signatures. The mutational signatures found by S-NMF to characterize MMR deficiency were most similar to SBS44 (cossim: 0.92 (optimal), 0.93 (non-integrated)) of all MMR COSMIC signatures. Combining our results with the findings of Drost et al.(45), might suggest that SBS44 is a more overarching signature to describe MMR deficiency.

HR deficiency. Of all COSMIC signatures, only SBS3 has been annotated as related to HR deficiency. In addition, although SBS8 has no proposed aetiology, Davies et al. suggest a relationship between SBS8 and HR deficiency (*BRCA1/BRCA2* deficient) (15). Subsequently, they use SBS8, in addition to SBS3, as a predictor of HR deficiency in their method HRDetect. For both integrated and non-integrated S-NMF the HR-related signatures show high similarity to SBS3 (cossim: 0.78 and 0.70) and SBS8 (cossim: 0.65 and 0.51). In both cases, the integrated S-NMF increased the similarity to the HR-related COSMIC signatures compared to the non-integrated S-NMF. This might suggest that the signatures found by the integrated S-NMF slightly better represent the underlying HR deficiency.

BER deficiency. As mentioned earlier, in both S-NMF models two different signatures were found to describe subpathways in BER. The signature related to *OGG1* deficiency and repair of oxidized bases is associated with two COSMIC signatures, SBS18 and SBS36. More specifically, SBS18 is related to DNA damage by reactive oxygen, while SBS36 is related to *MUTYH* deficiency. *MUTYH*, like *OGG1*, is a gene encoding a glycosylase protein that is related to the repair of oxidized bases. The BER-OGG1 signature found by the integrated S-NMF and non-integrated S-NMF had a high similarity to SBS18 (cossim: 0.76 and 0.74), and to an lesser extent to SBS36 (cossim: 0.52 and 0.50).

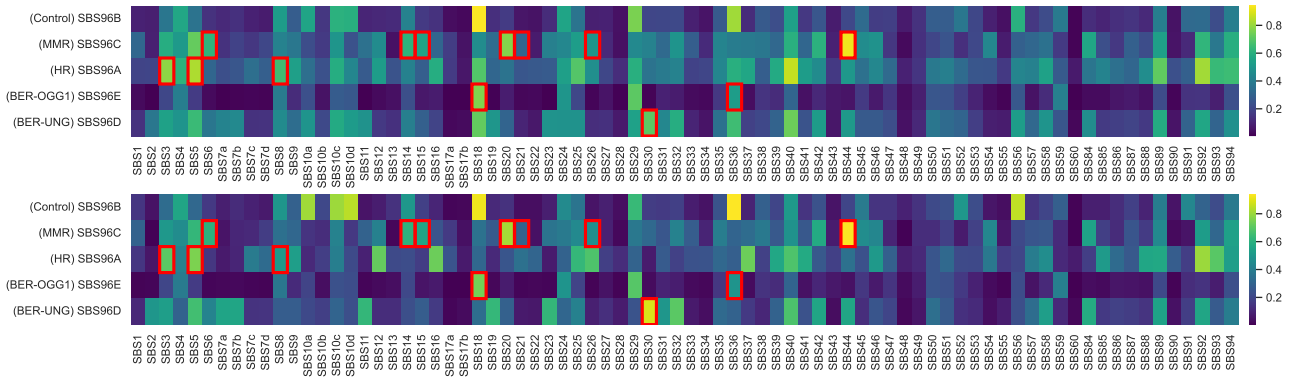


Fig. 12: **Heatmap showing the cosine similarity between the signatures found by S-NMF and the COSMIC signatures.** Top: integrated S-NMF. Bottom: non-integrated S-NMF. COSMIC signatures with aetiology related to the respective DNA repair pathways are annotated with red squares.

With regard to the signature related to *UNG* deficiency and repair of deaminated bases (uracil), we did not find a COSMIC signature with similar aetiology. However, COSMIC signature SBS30 is related to BER, specifically *NTHL1* deficiency (46), and is highly similar to the BER-UNG signatures found by S-NMF. Both *UNG* and *NTHL1* are BER glycosylases, however, *NTHL1* recognizes oxidized pyrimidines (i.e. T or C) (46), while *UNG* recognizes deaminated pyrimidines (U) (47). The BER-UNG signature found by the integrated S-NMF had a lower similarity than the signatures found by the non-integrated S-NMF (cossim: 0.70 and 0.89) . To conclude, even though both *UNG* and *NTHL1* are related to BER and result in a similar mutational profile, it might be the case that it is caused by a slightly different underlying mutational process.

4 Conclusion

We have implemented Supervised Non-negative Matrix Factorization (S-NMF), a novel approach that integrates mutational signature decomposition with a multinomial logistic regression classification. More specifically, we focused on the prediction of repair pathway deficiencies.

Our first aim was to learn signatures predictive of repair deficiency. By increasing the integration strength the importance of the cross-entropy loss (i.e. classification) becomes larger relative to the reconstruction loss (i.e. mutational signature decomposition). This trade-off was confirmed during hyperparameter optimization, where larger integration strength improved prediction accuracy at the cost of reconstruction error and stability of the signatures. However, the final integrated S-NMF model did improve prediction accuracy compared to benchmark models. Additionally, increasing the number of signatures above the most stable setting slightly improved the prediction accuracy but at the cost of a large decrease in stability. Taken together, the extent to which the prediction accuracy could be improved is limited by the decrease in the stability of the signatures. This is mainly attributed to the fact that the underlying mutational patterns that are being captured by (additional) signatures are not pronounced enough. However, we showed that the drop in stability could partially be accounted to the current partition-clustering approach of the training runs. We suggest that a different clustering technique could potentially mitigate the drop in stability, and as a consequence, allow for a higher accuracy given a particular stability. Additionally, we suggest that including a momentum term in the update formulas (e.g. Adam optimizer algorithm) makes the optimization by gradient descent less prone to local optima. Alternatively, methods could be explored to filter out sub-optimal runs based on their higher reconstruction and/or cross-entropy loss. These adjustments could mitigate the decrease in stability when increasing the integration strength or number of signatures and potentially allow for an improvement in prediction accuracy.

The second aim was to detect signatures that better represent the underlying mutational processes by exploiting the labels of the training samples. We showed that integrated S-NMF reduced the exposure of unrelated signatures and learned signatures with a more complete representation of the repair deficiencies. This effect is particularly strong if a signature is defined by a few characteristic mutation types. Furthermore, integrated S-NMF, like non-integrated S-NMF, captured signatures of repair deficiencies affecting distinct subpathways within the main repair pathway. Finally, the signatures found by S-NMF are similar to the cancer-related (COSMIC) signatures. However, there is no ground truth of the underlying mutational processes, which limits the assessment of the quality of the signatures.

The used data limited the performance evaluation of S-NMF due to three reasons. Firstly, the data from Zou et al. (16) had a limited number of distinctive samples. Therefore we had to rely upon bootstrapped oversampling to still perform a robust cross-validation and performance evaluation. Secondly, the prediction accuracy of all models was very high and the difference in accuracy was mainly attributed to samples from a single gene KO (*PMS1*). Thirdly, the cell lines were not exposed to any extrinsic DNA damage (e.g. using genotoxins), making intrinsic factors (e.g. DNA replication errors) the dominant cause of DNA damage. Together, this likely biased the mutational patterns of the repair deficiency and limited the number of repair pathways necessary to repair the damage. As a result, deficiencies in normally essential repair pathways probably went unnoticed. A solution could be to expose the gene KOs to a variety of DNA damaging factors that ideally mimic the *in vivo* DNA damage in humans.

The most realistic and unbiased approach to analyze the interplay between DNA damage and repair would be to use mutation data from cancer tumors. For example, data generated by The Cancer Genome Atlas (TCGA) or the International Cancer Genome Consortium (ICGC). Using cancer tumor data will likely result in a clinically more relevant model. However, cancer tumor data introduces two challenges. Firstly, the mutational processes occurring in human tumor cells are more convoluted which complicates the analysis of the effect of a repair pathway deficiency compared to the cell lines with only a single gene KO. On the other hand, tumor data might improve the evaluation of the integrated S-NMF since the benefit from exploiting the information contained in the annotations might be larger for the convoluted tumor mutation data.

The second challenge of tumor data is the limited sensitivity of current approaches to identify repair pathway deficiencies. Therefore, the labels used to train the S-NMF model will likely contain repair-deficient tumors

which are labeled as repair proficient. A semi-supervised version of S-NMF could be a potential solution. For example, by only considering tumors whose repair status was determined with high certainty in the categorical cross-entropy component of the loss function.

To conclude, we showed the potential benefit of integrating mutational signature decomposition with classification of samples using S-NMF. We expect this creates new opportunities for interpretable and in the long-term clinically relevant repair pathway prediction models.

References

[1]Hoeijmakers, J. H. J. DNA Damage, Aging, and Cancer. *The new england journal of medicine* **361**, 1475–1485 (2009).

[2]Ma, J., Setton, J., Lee, N. Y., Riaz, N. & Powell, S. N. The therapeutic significance of mutational signatures from DNA repair deficiency in cancer. *Nature Communications* 1–12 (2018). URL http://dx.doi.org/10.1038/s41467-018-05228-y.

[3]Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011).

[4]Anglian Breast Cancer Study Group. Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. *British Journal of Cancer* **83**, 1301–1308 (2000).

[5]Malone, K. E. *et al.* Prevalence and predictors of BRCA1 and BRCA2 mutations in a population-based study of breast cancer in White and Black American women ages 35 to 64 years. *Cancer Research* **66**, 8297–8308 (2006).

[6]Lord, C. J. & Ashworth, A. The DNA damage response and cancer therapy. *Nature* **481**, 287–294 (2012).

[7]Venkitaraman, A. R. Cancer suppression by the chromosome custodians, BRCA1 and BRCA2. *Science* **343**, 1470–1475 (2014).

[8]Knudson, A. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci USA* **68**, 820–823 (1971).

[9]Kaelin, W. G. The concept of synthetic lethality in the context of anticancer therapy. *Nature Reviews Cancer* **5**, 689–698 (2005).

[10]Prakash, R., Zhang, Y., Feng, W. & Jasin, M. Homologous Recombination and Human Health. *Perspectives in Biology* 1–29 (2015).

[11]Farmer, H. *et al.* Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434**, 917–921 (2005).

[12]Fong, P. C. *et al.* Inhibition of poly(adp-ribose) polymerase in tumors from brca mutation carriers. *New England journal of Medicine* 123–134 (2009).

[13]Bryant, H. E. *et al.* Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase.[erratum appears in Nature. 2007 May 17;447(7142):346]. *Nature* **434**, 913–917 (2005).

[14]Yang, X., Yan, L. & Davidson, N. E. DNA Methylation in Breast Cancer. *Endocrine-related Cancel* **8**, 115–127 (2001).

[15]Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nature Medicine* **23**, 517–525 (2017).

[16]Zou, X. *et al.* A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nature Cancer* **2**, 643–657 (2021).

[17]Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports* **3**, 246–259 (2013). URL http://dx.doi.org/10.1016/j.celrep.2012.12.008.

[18]Kasar, S. *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nature Communications* **6** (2015).

[19]Fischer, A., Illingworth, C. J. R., Campbell, P. J. & Mustonen, V. EMu : probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biology* **14** (2013).

[20]Rosales, R. A., Drummond, R. D., Valieris, R., Dias-neto, E. & Silva, I. T. Genome analysis signeR : an empirical Bayesian approach to mutational signature discovery. *Bioinformatics* **33**, 8–16 (2017).

[21]Rosenthal, R., Mcgranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. deconstructSigs : delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology* 1–11 (2016). URL http://dx.doi.org/10.1186/s13059-016-0893-4.

[22]Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016). URL http://dx.doi.org/10.1038/nature19768.

[23]Huang, X., Wojtowicz, D. & Przytycka, T. M. Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics* **34**, 330–337 (2018).

[24]Tate, J. G. *et al.* COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research* **47**, D941–D947 (2019).

[25]Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

[26]Chao, G., Luo, Y. & Ding, W. Recent Advances in Supervised Dimension Reduction: A Survey. *Machine Learning and Knowledge Extraction* **1**, 341–358 (2019).

[27]Wang, Y., Jia, Y., Hu, C. & Turk, M. Fisher non-negative matrix factorization for learning local features. *Proceedings of the Asian Conference on Computer Vision* 27–30 (2004). URL http://www.cs.ucsb.edu/~mturk/pubs/ACCVa2004.pdf.

[28]Yang, J., Zhang, Y. & Dong, X. A modified two-dimensional nonnegative matrix factorization algorithm for face recognition. *ICIC Express Letters, Part B: Applications* **6**, 7–12 (2015).

[29]Zafeiriou, S., Tefas, A., Buciu, I. & Pitas, I. Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. *IEEE Transactions on Neural Networks* **17**, 683–695 (2006).

[30]Jing, L., Zhang, C. & Ng, M. K. SNMFCA: Supervised NMF-based image classification and annotation. *IEEE Transactions on Image Processing* **21**, 4508–4521 (2012).

[31]Lee, H., Yoo, J. & Choi, S. Semi-Supervised Nonnegative Matrix Factorization. *IEEE Signal processing letters* **17**, 18–21 (2010).

[32]Haddock, J. *et al.* Semi-supervised NMF Models for Topic Modeling in Learning Tasks 1. *arXiv* 1–16 (2020).

[33]Serizel, R., Bisot, V., Essid, S. & Richard, G. Supervised group nonnegative matrix factorisation with similarity constraints and applications to speaker identification. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 36–40 (2017).

[34]Bisot, V., Serizel, R. & Essid, S. Feature Learning With Matrix Factorization Applied to Acoustic Scene Classification. *IEEE /ACM transactions on audio, speech, and language processing* **25**, 1216–1229 (2017).

[35]Lyu, X. *et al.* Mutational signature learning with supervised negative binomial non-negative matrix factorization. *Bioinformatics (Oxford, England)* **36**, i154–i160 (2020).

[36]Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. Nature. *Nature* **401**, 788–791 (2000).

[37]Lawson, C. & Hanson, R. Solving Least Squares Problems. *SIAM* **3**, 585–592 (1987).

[38]Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987).

[39]Drummond, J. T., Li, G. M., Longley, M. J. & Modrich, P. Isolation of an hMSH2-p160 heterodimer that restores DNA mismatch repair to tumor cells. *Science* **268**, 1909–1912 (1995).

[40]Prolla, T. A. *et al.* Tumour susceptibility and spontaneous mutation in mice deficient in Mlh1, Pms1 and Pms2 DNA mismatch repair. *Nature Genetics* **18**, 276–279 (1998).

[41]Degasperi, A. *et al.* A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies. *Nature Cancer* **1**, 249–263 (2020). URL http://dx.doi.org/10.1038/s43018-020-0027-5.

[42]Kim, Y.-J. & Wilson, D. M. Overview of Base Excision Repair Biochemistry. *Curr Mol Pharmacol.* **5**, 3–13 (2012).

[43]Alexandrov, L. B., Kim, J., Haradhvala, N. J. & Huang, M. N. The repertoire of mutational signatures in human cancer. *Nature* **578** (2020).

[44]Meier, B. *et al.* Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers. *Genome Research* **28**, 666–675 (2018).

[45]Drost, J. *et al.* Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science* **358**, 234–238 (2017).

[46]Grolleman, J. E. *et al.* Mutational Signature Analysis Reveals NTHL1 Deficiency to Cause a Multi-tumor Phenotype. *Cancer Cell* **35**, 256–266 (2019).

[47]Mol, C. D. *et al.* Crystal structure and mutational analysis of human uracil-DNA glycosylase: Structural basis for specificity and catalysis. *Cell* **80**, 869–878 (1995).

Supplementary Material

A Extended Methods: derivation of the update rules and derivatives for the S-NMF optimization algorithm

A.1 Matrix definitions

$\boldsymbol{X}_{N,T}$: Input matrix of (normalized) frequencies of T mutation types for each of the N samples (mutational profiles)

$\boldsymbol{Y}_{N,O}$: Output class label matrix, with O class labels for each of the N samples (one-hot encoded DNA repair deficiencies)

$\boldsymbol{S}_{K,T}$: Mutational signature matrix, denoting the frequencies of T mutation types for each of the K signatures

$\boldsymbol{E}_{N,K}$: Exposure matrix, containing the contribution of each of the K signatures to each of the N samples

$\boldsymbol{W}_{K,O}$: Coefficients of the logistic regression, weighing the contribution of each of the K exposures to the decision boundary of each of the O classes

with:

N : number of samples (and mutational profiles)

T : number of mutation types

K : number of signatures

O : number of output classes, corresponding to DNA repair pathway deficiencies and control

A.2 Loss function and update rules for optimization by gradient descent

The loss function $\mathcal{L}_{tot}$ optimized by S-NMF combines a reconstruction loss $\mathcal{L}_r$ with a classification loss $\mathcal{L}_c$ weighed by a factor λ_c as follows.

$$\mathcal{L}_{tot} = \mathcal{L}_r + \lambda_c \mathcal{L}_c$$

The reconstruction loss $\mathcal{L}_r$ is defined as the Frobenius reconstruction error of the decomposition of matrix $\boldsymbol{X}$ into matrices $\boldsymbol{E}$ and $\boldsymbol{S}$.

$$\mathcal{L}_r = ||\boldsymbol{X} - \boldsymbol{E}\boldsymbol{S}||_F^2$$

The classification loss $\mathcal{L}_c$ is the categorical cross-entropy loss:

$$\mathcal{L}_c = -\sum_{n=1}^{N}\sum_{o=1}^{O} y_{n,o} \log\left(\hat{y}_{n,o}\right),$$

where n and o are sample and output class indices, respectively, $y_{n,o}$ is the true class label (one-hot encoded), and $\hat{y}_{n,o}$ is the predicted soft class label calculated as follows.

$$\hat{y}_{n,o} = softmax\left(\boldsymbol{E_{*,o}}\boldsymbol{W_{n,*}}\right) = \frac{e^{\boldsymbol{E}_{n,*}\boldsymbol{W}_{*,o}}}{\sum_{o=1}^{O} e^{\boldsymbol{E}_{n,*}\boldsymbol{W}_{*,o}}}$$

The loss function $\mathcal{L}_{tot}$ is optimized using gradient descent, following the iterative updates below.

$$\boldsymbol{S} \leftarrow \boldsymbol{S} - \eta_S \cdot \nabla_{\boldsymbol{S}} \mathcal{L}_{tot}$$
$$\boldsymbol{E} \leftarrow \boldsymbol{E} - \eta_E \cdot \nabla_{\boldsymbol{E}} \mathcal{L}_{tot}$$
$$\boldsymbol{W} \leftarrow \boldsymbol{W} - \eta_W \cdot \nabla_{\boldsymbol{W}} \mathcal{L}_{tot}$$

In the next sections, we work out the derivatives for the update rules with respect to the the reconstruction and the classification losses.

A.3 Derivatives of the reconstruction loss:

Firstly, we calculate the derivative with respect to the reconstruction loss. This is identical to what existing methods that applied gradient descent to optimize NMF (36). For completeness, we will show the calculation.

Firstly, the frobenius reconstruction loss can be rewritten into four trace terms.

$$\begin{aligned} ||\boldsymbol{X} - \boldsymbol{ES}||_F^2 &= Tr((\boldsymbol{X}^T - \boldsymbol{S}^T\boldsymbol{E}^T)(\boldsymbol{X} - \boldsymbol{ES})) \\ &= Tr(\boldsymbol{X}^T\boldsymbol{X}) - Tr(\boldsymbol{X}^T\boldsymbol{ES}) - Tr(\boldsymbol{S}^T\boldsymbol{E}^T\boldsymbol{X}) + Tr(\boldsymbol{S}^T\boldsymbol{E}^T\boldsymbol{ES}) \\ since, & \\ ||\boldsymbol{X}||_F^2 &= Tr(\boldsymbol{X}^T\boldsymbol{X}) \\ Tr(A+B) &= Tr(A) + Tr(B) \end{aligned}$$

We take the derivative of each of the four trace terms in the loss function. Shown are the derivatives with respect to the exposure matrix $\boldsymbol{E}$, a similar procedure is followed to obtain derivatives with respect to $\boldsymbol{S}$ (applied mathematical rules between parentheses):

$$\begin{aligned} \nabla_S Tr(\boldsymbol{X}^T\boldsymbol{X}) &= 0 && - \\ \nabla_S Tr(\boldsymbol{X}^T\boldsymbol{ES}) &= \boldsymbol{E}^T\boldsymbol{X} && (\nabla_x Tr(\boldsymbol{AX}) = \boldsymbol{A}^T) \\ \nabla_S Tr(\boldsymbol{E}^T\boldsymbol{S}^T\boldsymbol{X}) &= \boldsymbol{E}^T\boldsymbol{X} && (\nabla_x Tr(\boldsymbol{X}^T\boldsymbol{A}) = \boldsymbol{A}) \\ \nabla_S Tr(\boldsymbol{S}^T\boldsymbol{E}^T\boldsymbol{ES}) &= ((\boldsymbol{E}^T\boldsymbol{E}) + (\boldsymbol{E}^T\boldsymbol{E})^T)\boldsymbol{S} = 2\boldsymbol{E}^T\boldsymbol{ES} && (\nabla_x Tr(\boldsymbol{X}^T\boldsymbol{AX}) = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{X}) \end{aligned}$$

The final derivatives $\nabla_{\boldsymbol{E}}\mathcal{L}_r$ and $\nabla_{\boldsymbol{S}}\mathcal{L}_r$ of the reconstruction loss $\mathcal{L}_r$ with respect to $\boldsymbol{E}$ and $\boldsymbol{S}$ are as follows.

$$\nabla_{\boldsymbol{S}}\mathcal{L}_r = -2\boldsymbol{E}^T\boldsymbol{X} + 2\boldsymbol{E}^T\boldsymbol{ES} \tag{1}$$

$$\nabla_{\boldsymbol{E}}\mathcal{L}_r = -2\boldsymbol{X}\boldsymbol{S}^T + 2\boldsymbol{ESS}^T \tag{2}$$

A.4 Derivatives of the classification loss:

Next, we calculate the gradients of classification component of the total loss. The main difference with a regular logistic regression is that in our case we not only apply gradient descent on the classification weights W but also the exposures E which could considered the input data in a regular logistic regression.

Here we focus on the categorical cross-entropy term and in section A.5 the gradient of the L2-regularization term will be calculated.

$$\mathcal{L}_c = -\sum_{n=1}^{N}\sum_{o=1}^{O} y_{n,o}\log(\hat{y}_{n,o}) + \lambda_{L2}\sum_{w\in\boldsymbol{W}} w^2 \tag{3}$$

For a more convenient calculation of the derivatives, we define $\boldsymbol{Z}$ as the product of the exposures and weights (eq. 5). $\boldsymbol{Z}$ can then be used as input to the softmax.

$$\text{with,} \quad \hat{y}_{n,o} = softmax(z_{n,o}) = \frac{e^{z_{n,o}}}{\sum_{o=1}^{O} e^{z_{n,o}}} \tag{4}$$

$$\text{with,} \quad \boldsymbol{Z} = \boldsymbol{EW} \tag{5}$$

Having defined $\boldsymbol{Z}$, we can define the derivative of the cross-entropy loss $d\mathcal{L}_c$ (eq. 6). Taking the derivative using matrices is more complex compared to scalars since the order of of the matrices is relevant. Taking this into account, the following rule can be applied:

$$d\mathcal{L}_c = \frac{\partial \mathcal{L}_c}{\partial \boldsymbol{Z}} : dZ \tag{6}$$

$$\text{with,} \quad \boldsymbol{A} : \boldsymbol{B} = \langle \boldsymbol{A}, \boldsymbol{B} \rangle_F$$

where : indicates the Frobenius inner product.

Next, both term in the derivative of the cross entropy loss $d\mathcal{L}_c$ (eq. 6 need to be calculated.

Firstly, the partial derivative of the cross-entropy loss w.r.t. $\boldsymbol{Z}$ ($\frac{\partial \mathcal{L}_c}{\partial \boldsymbol{Z}}$, eq. 7). This term is also derived for regular logistic regressions, therefore the partial derivative is know:

$$\frac{\partial \mathcal{L}_c}{\partial \boldsymbol{Z}} = (\hat{\boldsymbol{Y}} - \boldsymbol{Y}) \tag{7}$$

The second term in eq. 6, dZ, can be further defined in terms of the exposures and weights:

$$\boldsymbol{Z} = \boldsymbol{EW} \tag{8}$$

$$dZ = dE\boldsymbol{W} + \boldsymbol{E}dW \tag{9}$$

Taken together, $\frac{\partial \mathcal{L}_c}{\partial \boldsymbol{Z}}$ (eq. 7) and dZ (eq. 9) can be substituted in the derivative of the cross entropy loss (eq. 6).

$$\begin{aligned} d\mathcal{L}_c &= \frac{\partial \mathcal{L}_c}{\partial \boldsymbol{Z}} : dZ \\ &= (\hat{\boldsymbol{Y}} - \boldsymbol{Y}) : (dE\boldsymbol{W} + \boldsymbol{E}dW) \\ &= (\hat{\boldsymbol{Y}} - \boldsymbol{Y}) : dE\boldsymbol{W} + (\hat{\boldsymbol{Y}} - \boldsymbol{Y}) : \boldsymbol{E}dW \\ &= (\hat{\boldsymbol{Y}} - \boldsymbol{Y})\boldsymbol{W}^T : dE + \boldsymbol{E}^T(\hat{\boldsymbol{Y}} - \boldsymbol{Y}) : dW \end{aligned}$$

Finally, for the gradient w.r.t $\boldsymbol{E}$, $\boldsymbol{W}$ is constant (i.e. $dW = 0$) and for the w.r.t. $\boldsymbol{W}$, $\boldsymbol{E}$ is constant (i.e. $dE = 0$). This result in the final derivatives $\nabla_{\boldsymbol{E}}\mathcal{L}_c$ and $\nabla_{\boldsymbol{W}}\mathcal{L}_c$ of the cross-entropy loss w.r.t. $\boldsymbol{E}$ and $\boldsymbol{W}$:

$$\nabla_{\boldsymbol{W}}\mathcal{L}_c = \frac{\partial \mathcal{L}_c}{\partial \boldsymbol{W}} = \boldsymbol{E}^T(\hat{\boldsymbol{Y}} - \boldsymbol{Y}) \tag{10}$$

$$\nabla_{\boldsymbol{E}}\mathcal{L}_c = \frac{\partial \mathcal{L}_c}{\partial \boldsymbol{E}} = (\hat{\boldsymbol{Y}} - \boldsymbol{Y})\boldsymbol{W}^T \tag{11}$$

A.5 Derivative of L2-regularization term

Finally, the derivative of the L2 regularization term w.r.t. $\boldsymbol{W}$. Which is the same as standard derivations.

$$\frac{\partial(\lambda_{L2}\sum_{w\in\boldsymbol{W}} w^2)}{\partial \boldsymbol{W}} = 2\lambda_{L2}W \tag{12}$$

A.6 Final gradients

To conclude we combine the partial derivatives of individual terms in the loss function. Firslty, the signatures the loss term of the reconstruction loss (eq. 1). Secondly, for the exposure we have term from both the reconstrunction loss (eq. 2) and the cross-entropy loss (eq. 11). For the classifier weight we have a term from the cross-entropy loss (eq. 10) and from the L2-regularization (eq. 12) of the individual terms in the loss function to get the final gradients of the total loss $\mathcal{L}_{tot}$.

$$\nabla_{\boldsymbol{S}}\mathcal{L}_{tot} = -2\boldsymbol{E}^T\boldsymbol{X} + 2\boldsymbol{E}^T\boldsymbol{E}\boldsymbol{S} \tag{13}$$

$$\nabla_{\boldsymbol{E}}\mathcal{L}_{tot} = -2\boldsymbol{X}\boldsymbol{S}^T + 2\boldsymbol{E}\boldsymbol{S}\boldsymbol{S}^T + \lambda_c(\hat{\boldsymbol{Y}} - \boldsymbol{Y})\boldsymbol{W}^T \tag{14}$$

$$\nabla_{\boldsymbol{W}}\mathcal{L}_{tot} = \lambda_c(\boldsymbol{E}^T(\hat{\boldsymbol{Y}} - \boldsymbol{Y}) + 2\lambda_{L2}\boldsymbol{W}) \tag{15}$$

B Extended Results

(a) total number of single base substitutions (SBSs) per sample (b) total number of small insertions and deletions (Indels) per sample (c) total number of double base substitutions (DBSs) per sample

Fig. S1: **The mutations counts per sample categorized per gene KO**. Grey are the non-distinctive gene KOs. The other samples are the gene KO used for the evaluation of S-NMF colored by the related repair pathway.

Gene KO	DNA Repair pathway	Total # Samples	# Test	# Train	# Fold 1	# Fold 2	# Fold 3
ATP2B4	Control	8	2	6	2	2	2
MSH6	Mismatch Repair	8	1	7	3	2	2
MSH2		3	1	2	0	1	1
MLH1		4	1	3	1	1	1
PMS2		4	1	3	1	1	1
PMS1		4	1	3	1	1	1
EXO1	HR	3	1	2	1	0	1
RNF168		4	1	3	1	1	1
OGG1	Base Exision Repair	4	1	3	1	1	1
UNG		4	1	3	1	1	1
	Total:	46	11	35	12	11	12

Fig. S2: **Table showing the number of samples and how they are subdivided over the test set and training folds.** Shown are the Gene KOs (rows), their annotated DNA repair pathway deficiency and total number of samples/replicates of the with that gene KO. Next, how the total samples are divided over test and training data, and how the training data is further subdivided into 3 folds for cross-validation.

Fig. S3: **Heatmap showing hyperparameter optimization results.** First row) the average stability of the signatures. Second row) reconstruction error. Third row) prediction accuracy. Each cell in the square indicates a the score for that metric of a hyperparameter setting and is annotated and colored according to the the value for the particular metric. The main columns indicated the number of signatures K, ranging from 3 to 7. Within each square, the rows indicate the integration strength, and the columns the (L2) regularization strength.

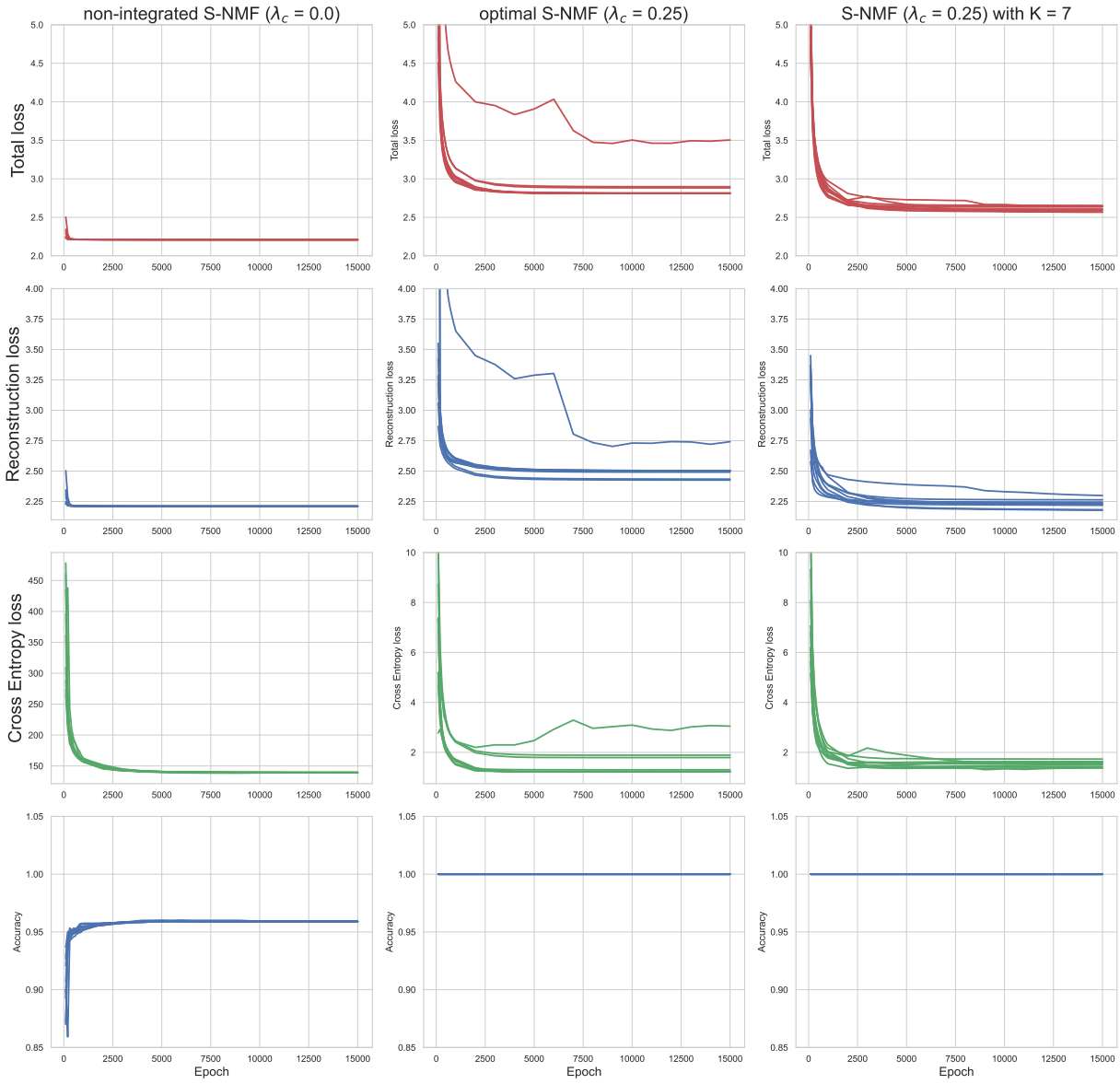
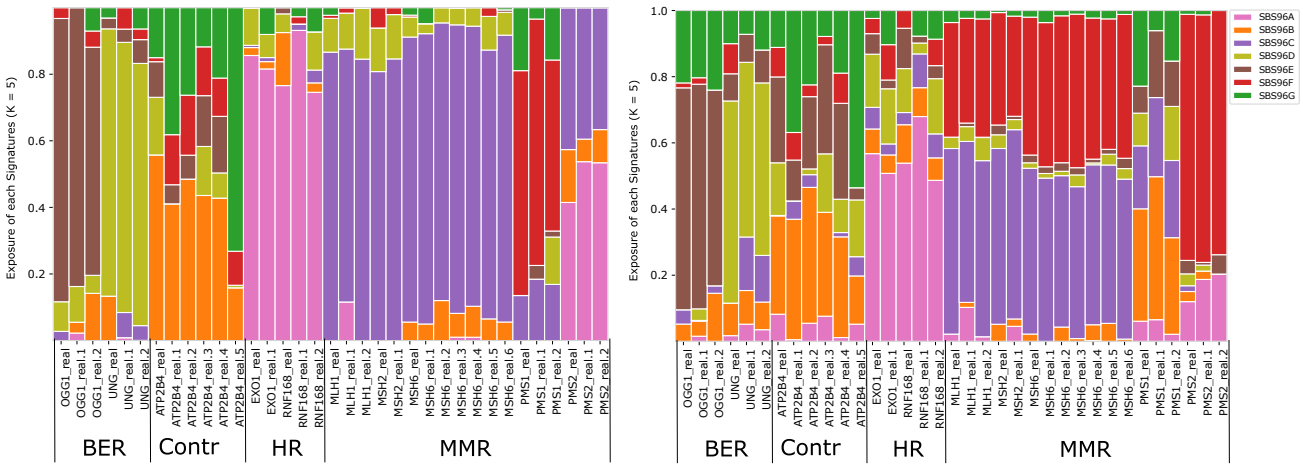


Fig. S4: **Training curves showing development of loss during the training epochs.** Left) non-integrated S-NMF. Middle) optimal S-NMF. Right) S-NMF with K=7. Top row) total loss. 2nd row) Reconstruction loss (frobenius reconstruction error). 3rd row) Cross-entropy loss. Bottom row) Prediction accuracy on training data



(a) Exposure found by the integrated S-NMF ($\lambda_c = 0.1$) with $K = 7$.

(b) Exposure found by the non-integrated S-NMF ($\lambda_c = 0.0$) with $K = 7$.

Fig. S5: **Exposure of the signatures (with $K = 7$) found by S-NMF**. Left) integrated S-NMF. Right) non integrated S-NMF. Only the real samples are shown to allow for practical visualization.