Mutational Signatures for Survival Prediction

Multi Task Auto Encoder for Survival Prediction using Mutational Signatures

Wouter Polet

Delft University of Technology

TU Delft

Mutational Signatures for Survival Prediction

Multi Task Auto Encoder for Survival Prediction using Mutational Signatures

by

Wouter Polet

Student number:	4730577	
Master's programme:	Computer Science, Bioinformatics specialization	
Faculty:	Electrical Engineering, Mathematics and Computer Science	
Project Duration:	February, 2023 - April, 2025	
Thesis committee:	Prof. dr. ir. Marcel Reinders	TU Delft
	Dr. Joana Gonçalves	TU Delft, supervisor
	Dr. Thomas Höllt	TU Delft
	MSc Yasin Tepeli	TU Delft, daily supervisor

Cover: Generated by Wouter Polet using Adobe Express

TU Delft

Preface

Throughout my master's I have had an interest in many of the departments at Computer Science. My interest for Bioinformatics stems partly from my interest in healthcare, which is definitely an influence of my family. This interest was strong enough to switch back to the Bioinformatics track after having chosen Algorithmics initially. This decision lead to some more courses I thoroughly enjoyed and ultimately to this project.

My thesis has been quite a journey which was, at times, quite challenging. I would like to thank Joana Gonçalves for being positive and helpful throughout the entire project, giving constructive feedback with every meeting. Furthermore, I want to thank Yasin Tepeli for always being available for a chat and for all your feedback and guidance throughout the thesis. I continued enjoying our meetings, which I always left with new ideas and motivation. Even though I feel that the thesis was the hardest part of my master's, it was also the most educational part thanks to you both. I should also not forget the Gonçalves lab, whose constructive feedback helped me improve a lot. Additionally, I would like to thank Marcel Reinders and Thomas Höllt for joining the thesis committee.

A big thank you to all my friends who supported me in various ways. I would not have gotten to VMB (or another university building) in the morning without some of you. Always having someone to ramble to about whatever aspect of the project I was busy with, helped me a lot during my thesis. Finally, I want to thank my family for always asking exactly once per visit how my thesis was progressing and being open to whatever the answer was.

Wouter Polet
Delft, April 2025

Multi Task Auto Encoder for Survival Prediction using Mutational Signatures

Wouter Polet*

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

Abstract

Motivation - Cancer remains one of the deadliest diseases worldwide and while advancements have been made in cancer treatment, cancer's heterogeneous nature makes it challenging to find a good treatment. Survival prediction for cancer patients can aid in choosing a treatment plan. Various machine learning methods have been employed to predict the survival of cancer patients, but they offer little insight into why a patient's survival is likely or not. Mutational signatures can offer an explanation on what a patient's cancer originates from, and can be linked to certain outside factors such as UV radiation. Even though mutational signatures have been employed in other problems, like predicting DNA repair pathway deficiencies, they have not been used in survival prediction. Integrating the survival problem with the extraction of mutational signatures could allow for extracting signatures that are particularly indicative of a patient's survival, providing a better prediction and more insight into why a patient's survival is predicted that way.

Results - We propose Multi-Task Auto-Encoder Cox (MTAE-Cox), which combines a non-negative auto-encoder for signature extraction with a Cox model for survival prediction and optimizes these in a multi-task manner. Our method jointly optimizes the auto-encoder's reconstruction error and the Cox loss, integrating the survival prediction problem into the signature extraction. MTAE-Cox is applied to four cancers of the TCGA dataset (GBM, HNSC, OV, SKCM) and its prediction performance is compared to Cox models using Gene Expression, Mutational Catalog, and exposures to COSMIC signatures. MTAE-Cox outperforms the generally applied gene expression (median C-index of 0.579 over 0.561 for gene expression) for GBM and outperforms Cox using non-integrated signatures derived by NMF for three of the four cancers. MTAE-Cox can extract biologically relevant signatures that are similar to COSMIC signatures that are known to be common in the specific type of cancer, for example SBS3 for ovarian cancer.

1. Introduction

Cancer is one of the deadliest diseases worldwide, being the leading cause of death in 2020, and continues to grow [1, 2]. While advancements in cancer treatment have been made in recent years with for example targeting genes in oncogene-driven cancers and immuno-oncology, challenges remain in the heterogeneous nature of cancer [3]. Because of the heterogeneity, cancer is divided into subtypes, which are used to determine the treatment path of patients [4]. However, even with the different treatments, not every cancer subtype (and therefore not every patient) can be treated effectively. Moreover, another consideration when choosing a cancer treatment is the side effects of the treatment [5]. For example, chemotherapy has more severe side effects than the aforementioned immuno-oncology [6]. Survival prediction of cancer patients can be used to aid the choice between treatments for the same subtype with different side effects, for example by estimating the effectiveness of a treatment [7] or by giving an indication of the lethality of a patient's cancer.

1.1. Survival prediction for cancer

Survival prediction currently involves training machine learning models to predict a patient's survival rate over time given the characteristics of the patient derived from genetic and clinical data. First, the survival prediction problem can be defined as a classification problem, which typically defines two classes for the patients: Low-risk and high-risk groups. Classifying patients into two or few classes simplifies the problem, but disregards crucial information such as the censoring of samples, the survival status, the survival time and typically requires defining low and high risk groups based on rules defined by the user (which can be dependent on the used data or cancer type). Second, methods that use regression models predict the survival time of patients. Although the regression definition takes survival time into account, it can still not incorporate censored data, as some patients may not have a survival time when they are still alive at the time of leaving the study. Finally, there are models that predict the hazard ratio (risk score) of patients over time. The predicted risk score combined with a baseline function leads to the survival function (hazard ratio over time) of a patient. Models solving this type of problem can incorporate censored samples into the training step. The standard model to predict the risk score is the linear Cox Proportional Hazards (Cox PH) model where the risk score is a linear combination of the covariates such as the expression of a selected set of genes [8, 9]. At present, non-linear versions of Cox PH have been proposed as well, where DeepSurv was the first non-linear method to improve over the linear Cox PH model [10].

Various machine learning methods have been employed using different types of data for survival prediction. Pathological images are used in combination with regularized machine learning methods to perform survival prediction by

distinguishing short-term and long-term survivors, for example for lung cancer patients [11]. More generic and readily available types of clinical data – such as the patient's age, the patient's geographic location, or the stage of a cancer – can be used for survival prediction as well [12, 13]. Next to clinical data, molecular or genetic data is used to predict the survival of cancer patients. In particular gene expression contains relevant information for the survival prediction problem and is widely used in the survival prediction problem [14]. While vastly leveraged, usage of gene expression may pose challenges due to its high dimensional nature. For many cancers, we have a limited number of patients ($\sim 100-1000$ for TCGA data), while gene expression typically involves more than 20.000 genes (in the TCGA dataset). Having much more features than samples may cause machine learning models to overfit, thus not fit properly and not provide an accurate survival prediction. To alleviate the overfitting, methods employ a feature selection step which is performed before training and using the machine learning model. The feature selection step can select features that are likely to be useful for the survival prediction, allowing the machine learning model to train on a smaller set of features that is closer to the number of samples in terms of size. One such a feature selection step is PKSFS, which selects features based on a priori knowledge and stability of the features to reduce the feature set [15]. Another example is proposed by Mosquera Orgueira et al. who iteratively prune the feature set based on the features' importance determined by the random forests models [13].

The survival of a patient is affected by multiple processes that drive their cancer. These processes include for example DNA repair deficiencies, external UV radiation, alcohol consumption, smoking cigarettes, etc [16, 17]. Each of these processes leaves behind mutations in the cells' DNA, where a nucleotide in the cell becomes different from the reference genome. Over time, the processes' mutations accumulate and a combination of mutations in certain places in the DNA can cause cells to become cancer cells. Alexandrov et al. have shown that a patient's mutations can be deciphered into mutational signatures and how much each of these signatures have attributed to the set of mutations, called exposures [18]. The deciphered signatures can be related to the processes causing mutations, giving an insight into what caused the cancer [19]. Various studies have related the mutational signatures of many cancer patients to the biological processes, whose results have been recorded in the Catalogue of Somatic Mutations of Cancer (COSMIC) database [20]. The mutational signatures, specifically the exposures to the signatures, have been used in the prediction of for example DNA repair deficiencies [21]. However, the mutational signatures have not been used for the survival prediction problem of cancer patients.

Deciphering mutational signatures from mutational catalogs (number of mutations per type of mutation) is typically done through nonnegative matrix factorization (NMF) [18]. NMF is an unsupervised method; therefore the derived signatures are not specific to a particular task, such as survival prediction, but are more general. Signatures that are specific to a task could increase the performance of the model for that task and provide more insight in the biologically relevant processes. Supervised versions of NMF to incorporate the prediction problem in the decomposition into signatures have been proposed [22, 23]. While these supervised NMF methods can extract signatures specific to a task, they cannot be integrated with some kinds of models such as deep-learning based methods and can therefore not be used for every task. To overcome this compatibility issue, we propose an auto-encoder model that uses a linear activation function and applies a non-negativity constraint to imitate NMF. The optimisation of the auto-encoder can be done through widely used deep learning optimisers such as SGD, Adam, or RMSProp, making it easier to integrate with other deep learning methods. The auto-encoder finds signatures in the weights of its decoder and the exposure in the encoding of the mutational catalog.

Using this linear auto-encoder we propose the Multi-Task Auto-Encoder Cox (MTAE-Cox) model to utilize and find mutational signatures for the survival prediction problem. MTAE-Cox uses the auto-encoder to extract signatures and exposures from the mutational catalog and uses these exposures in its Cox model to perform the survival prediction. We have implemented two versions: Auto-Encoder Cox model (AE-Cox), which derives signatures and performs the surival prediction disjointly, and the Multi-Task Auto-Encoder Cox model (MTAE-Cox) which derives signatures and performs the prediction jointly. Therefore, MTAE-Cox integrates the survival prediction problem into the signature extraction, yielding signatures specific to the survival prediction problem and improving the prediction performance. The performance of both AE-Cox and MTAE-Cox of the survival prediction problem is compared to various benchmark models. Finally, we find COSMIC signatures that relate to the found signatures to verify their origin.

2. Methods

In survival prediction, a survival function is defined as $h(t, \overrightarrow{x_i}) = \lambda_0(t)\eta(\overrightarrow{x_i})$ [10] where t is time, $\overrightarrow{x_i}$ is the feature vector of patient i, $\eta(\overrightarrow{x_i})$ is the risk function, $\lambda_0(t)$ is the baseline function, and $1 \leqslant i \leqslant N$ with N the number of patients. In this paper, we focus on finding the risk function, which provides a risk score based on the feature vector of the patient. This risk score multiplied by the baseline function provides the survival function of a patient. So given is a feature matrix $\boldsymbol{X} \in \mathbb{R}^{N \times P}$, where N is the number of samples (patients), P is the number of features, and $\overrightarrow{x_i}$ is row i of $\boldsymbol{X}$ corresponding to patient i. For the proposed models the feature matrix is the mutational catalog of single based substitutions with context (SBS96 mutations [18]), so $P = 96$.

In order to train a survival prediction model, a set of labeled samples is needed. Each sample has two labels; a time indicating the time between diagnosis and last interaction with the patient and a status indicating whether the patient is deceased (1) or not (0). A patient is censored when it is unknown what the result of their treatment was after the last follow-up date. Therefore, we know that a censored patient has survived until at least their survival time, but they may have survived for longer. The time and status give two matrices for labels; $\boldsymbol{T} \in \mathbb{R}^{N \times 1}$ for time and $\boldsymbol{S} \in \{0, 1\}^{N \times 1}$ for status.

2.1. AE-Cox and MTAE-Cox

Auto-Encoder Cox (AE-Cox) and Multi Task Auto-Encoder Cox (MTAE-Cox) introduce a novel approach to survival prediction of cancer patients using mutational signature extraction (Figure 1). AE-Cox first extracts novel mutational signatures for a patient group using a non-negative auto-encoder and after that predicts the survival function of patients using a Cox model. Thus, AE-Cox trains the auto-encoder and the cox model disjointly, where the Cox model uses the latent

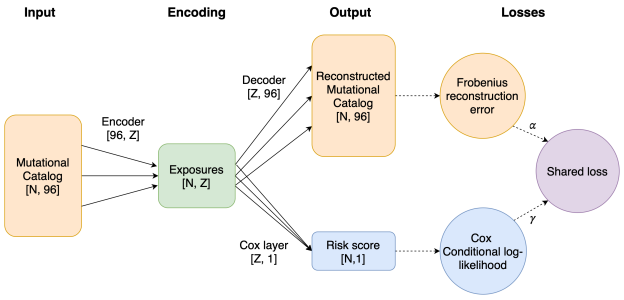


Fig. 1. Overview of the MTAE-Cox model's components. This includes the features (mutational catalog), the auto-encoder, its latent space (exposures) and the Cox model. The losses of the model are shown in the circles. The dimensions of the matrices for each component are shown in brackets.

embedding extracted from the auto-encoder as input. MTAE-Cox integrates the extraction of mutational signatures and the survival prediction by jointly training the auto-encoder and Cox model using a multi-task loss. Contrary to AE-Cox, MTAE-Cox derives signatures that can be influenced by the survival prediction problem.

2.1.1. Survival Prediction

AE-Cox and MTAE-Cox both include a Cox model, that estimates the risk function similarly to a Cox proportional hazards (CoxPH) model. The risk function is estimated by finding a set of unknown parameters such that $\eta(x_i) = e^{x_i\beta}$, where $\beta \in \mathbb{R}^{P\times 1}$ are the unknown parameters [8]. In other words, the risk function is a linear combination of the hazards (the latent embedding extracted from the auto-encoder) in the CoxPH model. Thus, in both AE-Cox and MTAE-Cox the Cox model is an output layer with a single output (the risk score) with a linear activation function. To estimate β, the conditional log-likelihood is optimised [8]

$$L_{\mathrm{COX}}(\beta) = \sum_{i=1}^{k} x_i\beta - \sum_{i=1}^{k} \log\left[\sum_{l\in R(t_i)} e^{x_l\beta}\right] \tag{1}$$

$$= \sum_{i=1}^{N} s_i \left(x_i\beta - \log\left[\sum_{l\in R(t_i)} e^{x_l\beta}\right]\right) \tag{2}$$

where s_i is the status of patient i in $\boldsymbol{S}$, $R(t_i)$ is the risk set of time point t_i, k is the number of uncensored patients. The sum from $i = 1$ to k is then the sum over the uncensored patients. The risk set is the set of patients that are still at risk at the given time point, including the censored patients. The Cox loss therefore uses both censored and uncensored patients to compute the log-likelihood of a sample, but only sums the log-likelihoods of the uncensored patients.

In AE-Cox the auto-encoder and the Cox model are trained disjointly. The auto-encoder and Cox model are instantiated separately, so the connection between the exposures and the Cox model in Figure 1 is not present directly. In training the non-negative auto-encoder is first fully trained, after which the Cox model is trained on the latent embedding (the exposures, $\boldsymbol{E} \in \mathbb{R}_{\geq 0}^{Z\times P}$) extracted from the trained auto-encoder. When a prediction is made using a trained AE-Cox model, the patient's mutational catalog's latent embedding is computed using the auto-encoder, which is then fed into the Cox model that finally computes the risk score.

2.1.2. Mutational signature extraction by non-negative auto-encoder

Typically mutational signatures are derived from the mutational catalog using non-negative matrix factorization (NMF) [18]. In AE-Cox and MTAE-Cox we use a non-negative auto-encoder instead, enabling us to integrate the mutational signature extraction with a larger variety of prediction models. The auto-encoder consists of the input layer, one hidden layer, and the output layer. Similar to NMF, the auto-encoder has a fixed number of signatures ($Z \in \mathbb{N}$ is the number of signatures). The input layer expects the mutational catalog of the patients, as defined in the problem definition $\boldsymbol{X}$. The number of nodes in the hidden layer is the number of signatures, which results in the exposure matrix $\boldsymbol{E} \in \mathbb{R}_{\geq 0}^{Z\times P}$. Finally the output layer has the same dimension as the input layer, so the output matrix $\hat{\boldsymbol{X}} \in \mathbb{R}^{N\times P}$. The encoder and decoder use linear activation with no bias, such that the auto-encoder simulates NMF as close as possible. Thus, the hidden layer and output layer are computed as

$$\boldsymbol{E} = \boldsymbol{X}\boldsymbol{S_{enc}} \tag{3}$$

$$\hat{\boldsymbol{X}} = \boldsymbol{E}\boldsymbol{S_{dec}} \tag{4}$$

where $\boldsymbol{E}$ is the matrix corresponding to the hidden layer, $\boldsymbol{S_{enc}} \in \mathbb{R}_{\geq 0}^{P\times Z}$ are the encoder weights, $\boldsymbol{S_{dec}} \in \mathbb{R}_{\geq 0}^{Z\times P}$ are the decoder weights, and $\hat{\boldsymbol{X}}$ is the reconstructed mutational catalog. The NMF originally proposed to extract mutational signatures uses Equation 4 to express the relation between the mutational catalog, exposures, and signatures [18]. Thus, we can extract the mutational signatures from the auto-encoder by taking the weights of the decoder, $\boldsymbol{S_{dec}}$. According to the definition of mutational signatures they behave like probability density functions, so they cannot have negative values and the sum of all values in a signature has to be 1 [18]. To align with the former, the non-negativity constraint is added to the encoder and decoder weights, which sets any negative value to zero after each training iteration. To adhere to the latter, we expect each row of the signature matrix $\boldsymbol{S_{dec}}$ to sum to 1. This is not enforced by a constraint during training, but instead $\boldsymbol{S_{dec}}$ is normalised when analysing the signatures. To optimise the weights of the auto-encoder the squared Frobenius reconstruction error is used (Equation 5).

$$F(\boldsymbol{X}, \hat{\boldsymbol{X}}) = ||\hat{\boldsymbol{X}} - \boldsymbol{X}||_F^2 \tag{5}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{P} |\hat{x}_{ij} - x_{ij}|^2 \tag{6}$$

As any auto-encoder the loss is minimized when the output $\hat{\boldsymbol{X}}$ is similar to the input $\boldsymbol{X}$.

2.1.3. Multi task model

MTAE-Cox integrates the training of the auto-encoder with the training of the Cox model, allowing the derived signatures to be influenced by the survival problem. To achieve this, the auto-encoder's latent embedding becomes the input for both the decoder and the Cox model as shown in Figure 1. Since the auto-encoder and Cox model are optimised at the same time, their losses have to be combined. The Cox loss (Equation 2) effectively only sums the losses of uncensored samples, while the auto-encoder loss (Equation 5) sums the error over all samples.

To accommodate this difference the losses are normalised as shown in Equation 7 and Equation 8.

$$\overline{L_{COX}}(\beta) = \frac{1}{k} L_{COX}(\beta) \tag{7}$$

$$\bar{F}(\boldsymbol{X}, \hat{\boldsymbol{X}}) = \frac{1}{N} F(\boldsymbol{X}, \hat{\boldsymbol{X}}) \tag{8}$$

Finally, the two losses are combined with two hyperparameters α and γ. The total loss is a weighted sum of the Cox loss and the reconstruction error, as shown in Equation 9.

$$L_{MTAE}(\boldsymbol{X}, \hat{\boldsymbol{X}}, \beta) = \alpha \bar{F}(\boldsymbol{X}, \hat{\boldsymbol{X}}) + \gamma \overline{L_{COX}}(\beta) \tag{9}$$

2.2. Experimental Setup

AE-Cox and MTAE-Cox give two types of outputs, which are both to be evaluated. First is the survival prediction of the patient, i.e. their risk score, determined by the Cox model. Second are the mutational signatures, generated by the auto-encoder part of the models. For the survival prediction we are interested in the performance of the prediction, or how accurate this performance is compared to the known labels. This requires the model to distinguish high risk patients with a low survival function from low risk patients well. The mutational signatures are analysed by their stability and compared to existing and well-known signatures.

2.2.1. TCGA data

The feature vector used by (MT)AE-Cox is derived from a single-omic, such as gene expression, mutational catalog, or exposures of signatures. For the experiments conducted in this paper data from The Cancer Genome Atlas (TCGA) is used [24]. More specifically, the data from all cancer types part of the Firehose Legacy dataset are obtained from the cBioPartal [25]. Since AE-Cox and MTAE-Cox need mutation data, a survival label, and a survival time, we consider only patients that contain those three types of data and patients that have at least 1 mutation. As a result of its loss function (Equation 2) the trainability of the Cox model is heavily dependent on the ratio of censored and uncensored patients. To select the cancer types to analyse, we therefore select the cancer types where more than 40% of the patients is uncensored. The selected cancer types are GBM (79%), HNSC (43%), LAML (65%), OV (67%), SKCM (52%). After filtering, we find a total of 285 GBM, 508 HNSC, 174 LAML, 314 OV, 360 SKCM patients. The auto-encoder part of the models depends on the number of mutations of the patients. The median number of mutations varies over the selected cancer types: 35 GBM, 74 HNSC, 6 LAML, 26 OV, 227 SKCM. Since LAML has a limited number of mutations per patient, it is excluded from further analysis.

Besides filtering the dataset on patients with useable data, one preprocessing step is performed. This preprocessing is constructing the mutational catalog for each of the patients, based on the TCGA mutation file. The mutational catalog is a count for each type of single base substitution, as defined by Alexandrov et al. [18]. For each type of substitution we consider the context, which are the two directly adjacent nucleotides. The 6 single base substitutions with each 16 possible contexts, yields 96 types of mutations that need to be counted. We combine the mutation and its location known from the TCGA data with the GRCh37.p13 reference genome from the Genome Reference Consortium [26] to count each of the mutation types in the mutational catalog.

2.2.2. Training

AE-Cox's auto-encoder and Cox model are trained disjointly. First, the auto-encoder is trained, whose features and labels are both the mutational catalog matrix $\boldsymbol{X}$. The weights are optimised using an Adam optimiser and updated using back-propagation [27, 28]. The training is run until the validation loss does not improve for 100 epochs, or until a maximum of 500 epochs. When the training is stopped, the weights for which the validation loss was lowest are selected. After the training of the auto-encoder is completed, the mutational catalog is passed through the encoder of the auto-encoder to obtain the embedding. Then the Cox model is trained using the embedding as input and the survival status and time as labels. This is also done with an Adam optimiser, until the validation loss does not improve for 100 epochs or until a maximum of 500 epochs.

MTAE-Cox is trained by optimising the joint loss of the auto-encoder and the Cox model. During training, the Cox model directly obtains the latent embedding of the mutational catalog from the auto-encoder, so the features of MTAE-Cox are the mutational catalog, $\boldsymbol{X}$, and the labels are the same mutational catalog for the auto-encoder and the survival time and status for the Cox model. Like AE-Cox, MTAE-Cox is optimised through back-propagation with the Adam optimiser. The training is performed until the validation loss does not improve after 300 epoch, with a maximum of 5,000 epochs allowing the training to continue for longer than AE-Cox. MTAE-Cox has a higher upper limit of epochs because the first part of training mostly focuses on improving the reconstruction error part of the loss, after which the Cox loss starts to decrease as well. Training more epochs allows for both of these optimisations and to find the best balance between the two losses. The training is performed in single batches per epoch for both AE-Cox and MTAE-Cox, allowing the Cox model to optimise better as it can use a larger number of uncensored patients per update iteration.

Mutational signatures

The mutational signatures are derived from the weights of the auto-encoders in AE-Cox and MTAE-Cox. After the training is completed, the weights from the decoder are extracted. There is no constraint on the weights to be probability density functions, but there is a non-negativity constraint. In order to create mutational signatures, we normalise each row of the weight matrix.

Determining the quality of the signatures is done by computing the stability of the signatures. Similar to Alexandrov et al.'s NMF method [18] the training is performed for a range of signatures (2 to 10 specifically), for 10 iterations per number of signatures providing 10 sets of signatures for each number of signatures. Per number of signatures, these signatures are clustered using a variant of k-means clustering in which each signature of an iteration is assigned to a different cluster. The stability is defined as the mean silhouette width of the clusters of mutational signatures. Additionally, the centroids of the clusters are used as the signatures derived from the models. Finally, the reconstruction error for these signatures is computed.

Using the stability over the number of signatures we can determine the optimal number of signatures for each model and dataset. Any number of signatures with a stability higher than 0.7 is considered and the one with the lowest reconstruction error is selected as optimal number of signatures. If there is no number of signatures above the threshold, the number of signatures with the highest stability is chosen.

Hyperparameters
Besides the number of signatures, the main hyperparameters to choose are the weights of the reconstruction error and of the cox loss for MTAE-Cox. These weights are determined by an exploratory analysis, where the training losses of each was analysed for a range of weights. We found that the weights did not greatly influence the model's performance, unless γ was so large that the reconstruction error would not decrease anymore. In that case, the auto-encoder is not training to find meaningful signatures anymore, which is not desirable. Hence, we chose to set both α and γ to 1.

2.2.3. Testing

After determining the optimal number of signatures, the trained models are tested and compared to established methods using 5-fold cross validation to allow for analysis in the stability of the methods. Finally, since the performance of the Cox model depends on the ratio of the (un-)censored samples, the folds are stratified on the survival status.

Evaluation metrics
The performance of the survival prediction is measured with the concordance-index (C-index). The C-index is computed by using the survival time of the samples. Using the predicted risk score the samples are sorted and their order is compared to the order when sorted using their survival time. The C-index is then computed by dividing the number of correctly ordered pairs over the total number of pairs as denoted in Equation 10 [29].

$$\text{C-index} = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\eta_j>\eta_i} \cdot S_j}{\sum_{i,j} 1_{T_j<T_i} \cdot S_j} \tag{10}$$

where S_j is the survival status (0 for censored, 1 for uncensored) of sample j and η_i is the predicted risk score of sample i.

The interpretation of the C-index is similar to that of the better-known AUC-index. When all pairs are predicted correctly, the C-index will be 1, while a random model will give a C-index of 0.5.

Comparison methods
We compare AE-Cox's and MTAE-Cox's performance with various existing methods. Both the performance of the Cox model and the signatures derived by the auto-encoders are evaluated.

The derived signatures are compared to Sigprofiler [30]. Sigprofiler uses a non-negative matrix factorization (NMF) to derive the signature matrix. This is used on the same dataset as AE-Cox and MTAE-Cox, with the same parameters for number of iterations and range of signatures. The optimal number of signatures is determined for Sigprofiler separately with the same stability threshold as is used for AE-Cox and MTAE-Cox of 0.7. To determine the similarity between signatures derived by NMF, AE-Cox, and MTAE-Cox the signatures of two methods are matched to each other, such that the highest average cosine similarity is obtained. Finding this allocation of signatures pairs is done by generating all possible combinations, calculating the average similarity for each combinations, and choosing the allocation resulting in the maximum average similarity.

The performance of the survival prediction is compared to a variety of models. The topology of the model does not change – (MT)AE-Cox is always compared to a Cox model – but the features of the Cox model vary. The following feature sets are used for the Cox models:

- Gene expression
- Mutational catalog
- Exposure to COSMIC signatures
- Exposure to SigProfiler's NMF signatures

Gene expression is used the most for survival prediction for cancer patients and is generally the best indicator for survival [14]. We use the z-score of the gene expression as input to the Cox model, as present in the TCGA dataset. The z-score computes the relative expression in a tumor compared to the gene expression of a control group. Thus it highlights the genes that are expressed in a non-standard way.

Mutational catalog is the counts of the mutations grouped as the SBS96 mutations. For this comparison method, the catalog is normalised and then directly used by the Cox model.

The COSMIC signatures are an established set of known mutational signatures, that have been found by a variety of studies [20]. To obtain the exposures of the samples in the dataset we use, we utilize SigProfiler's Assignment [31]. SigProfiler's Assignment allows the assigning of known signatures to a mutational catalog, yielding the exposure matrix. The version of the COSMIC signatures used is 3.4. We run the assignment on our entire dataset and obtain the exposure to COSMIC signature for each patient. Then the Cox model is trained and scored with these exposures as input.

The exposures of the Sigprofiler's NMF method are extracted from the signature derivation step. When NMF determines the mutational signatures, it simultaneously estimates the exposure matrix that the Cox model uses.

3. Results and Discussion

Throughout the results AE-Cox, MTAE-Cox, and the baseline methods where relevant, are applied to extract signatures and predict the survival from/of the patients across four cancer types (HNSC, GBM, OV, SKCM), as selected in subsubsection 2.2.1, unless stated otherwise. Furthermore, unless stated otherwise, the experiments are performed with 5-fold cross validation to be able to check for the stability of the methods' performance.

3.1. Hyperparameters and training

First of all, we evaluate how the balance between the auto-encoder's reconstruction error and the Cox loss impacts the MTAE-Cox's survival prediction performance (C-index). We analyse the influence of the ratio of the two losses' weights on both the prediction performance (C-index) and the training curves of the model, since the two weights are part of a weighted sum that forms the total loss. A 5-fold cross validation is run on ratios $\phi = \frac{\alpha}{\gamma}$ ranging from 10^{-3} to 10^{3}, where α is the weight for the Frobenius reconstruction error and γ is the weight for the Cox loss (Equation 9).

The impact of the loss ratio on the C-index performance is cancer dependent, as shown in Figure 2. MTAE-Cox has stable C-index performance for HNSC, ranging between a mean of 0.485 ($\phi = 10^{-2}$) and 0.522 ($\phi = 10^{2}$) and for SKCM, ranging between 0.558 ($\phi = 10^{-3}$) and 0.593 ($\phi = 10^{2}$). For GBM, the performance stays stable for ratios up to 1 (mean C-index of 0.569) and then the performance decreases as the ratio increases (mean C-index 0.495, $\phi = 10^{3}$). Performance on OV has the opposite trend, where the C-index is stable up to $\phi = 1$ as well, but the performance increases as the ratio increases after that (mean C-index 0.523 ($\phi = 1$) to 0.575($\phi = 10^{3}$)). The variance

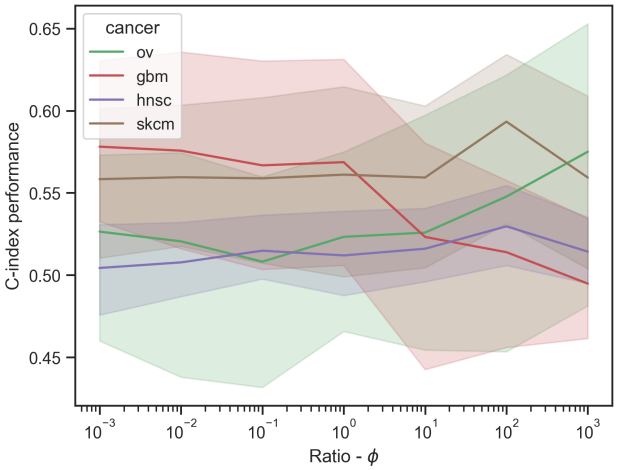


Fig. 2. Test C-index performance results of the weight ratios ϕ ranging from 10^{-3} to 10^{3}. Results are over a 5-fold cross validation, where the mean is plotted on the line.

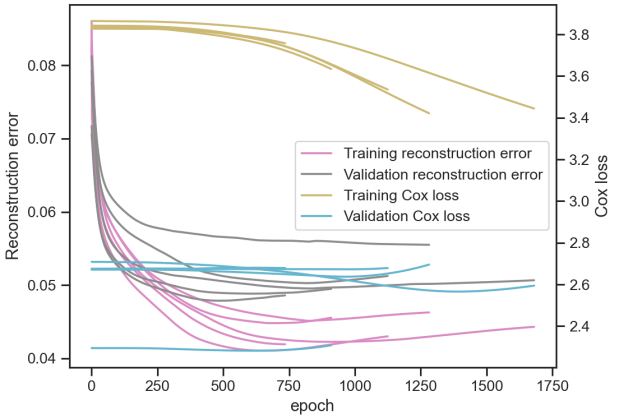


Fig. 4. Training plot of MTAE-Cox on ovarian cancer (OV) with weight ratio $\phi = 10$. Again a separate line per fold and both reconstruction error and Cox loss (both training and validation) are plotted.

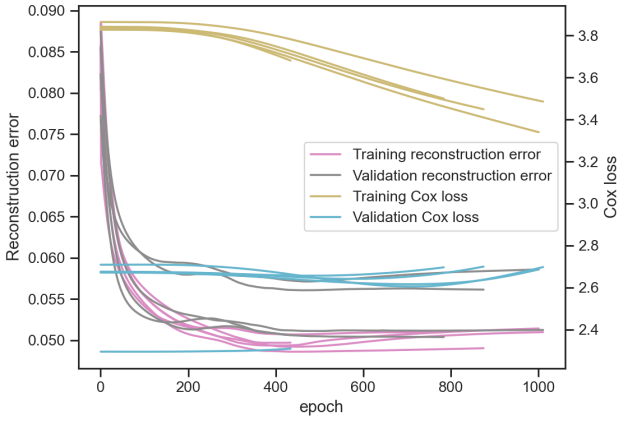


Fig. 3. Training plot of MTAE-Cox on ovarian cancer (OV), where a separate line per fold is plotted. Both reconstruction error and cox loss are plotted, with each the training and validation error/loss. The weight ratio ϕ is 1.

of the performance of the models over the folds is very high, making it impossible to reliably determine what ratio is better for the model's performance.

3.1.1. Analysis on reconstruction error and Cox loss weights

We also investigate how the MTAE-Cox model optimises the two losses during training to get more insight into the model's behaviour. The training curves follow similar trends for the four cancer types (one of which is shown in Figure 3). The reconstruction error of the training set decreases quickly in the first 200 epochs, after which it stabilizes. The reconstruction error over the validation set closely follows the behaviour of the error over the training set, albeit with a slightly higher value. The reconstruction error converges regardless of the weight ratio, however the higher ϕ the less epochs it needs to converge.

The Cox loss does not decrease rapidly at the beginning of training and is prone to overfitting after a certain number of epochs, where the training loss gradually decreases further but the validation loss does not. The validation Cox loss decreases before it starts increasing (and with that overfitting), so an optimum can be found during training. Contrary to the reconstruction error, the optimization of the Cox loss is dependent on the weight ratio (ϕ). When the ratio of the losses' weights' (ϕ) increases, the Cox loss over the training set starts decreasing after a larger number of epochs (Figure 4).

With a lower ratio, the training is stopped earlier where the reconstruction error contributed more to the total loss. The early stopping checks whether the total validation loss still decreases and stops the training if not. The validation reconstruction error keeps decreasing slightly, even after a high number of epochs. However, the Cox validation loss has an optimum, after which it starts increasing again. For higher weight ratios, the reconstruction error compensates for the increasing Cox loss, while for lower ratios the total validation loss increases because of the Cox loss increasing.

The optimal values of the validation losses generally do not change much with the weight ratio ϕ. However, the Cox loss requires more epochs to optimise, so to speed up the training process a lower ϕ can be chosen. Furthermore, for larger values of ϕ, the validation Cox loss does not reach its optimum anymore, so values of ϕ from at least 10^2 should be avoided.

3.2. Signatures of NMF, AE-Cox and MTAE-Cox

SigProfiler's NMF [18] and the auto-encoder in both AE-Cox and MTAE-Cox are used to derive the signatures from the mutational catalog. While they share the same purpose, the method of finding the signatures, and thus the signatures themselves, varies between them.

3.2.1. Selection of number of signatures

The stability of the mutational signatures, which gives an indication of how well the model can find similar signatures across repetitions, is used to determine the number of signatures to be used in NMF, AE-Cox, and MTAE-Cox. Specifically, we compare the stability of the signatures found by all these three models. The stability of the derived signatures is the average silhouette score of the clusters computed in the

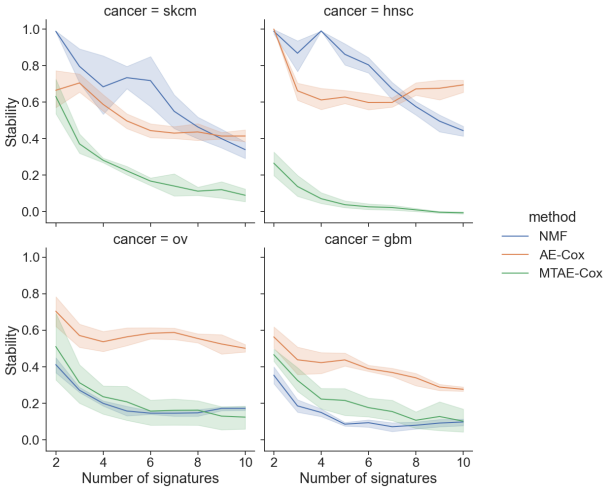


Fig. 5. The stability, as recorded when finding the optimal number of signatures, for each cancer type separately with values for the baseline NMF, AE-Cox, and MTAE-Cox. The number of signatures ranges from 2 to 10.

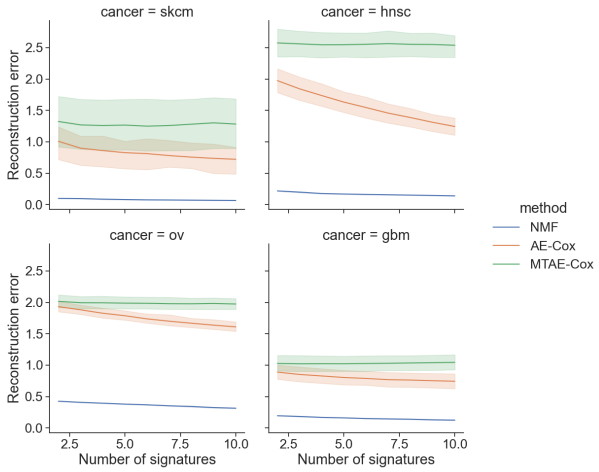


Fig. 6. The Frobenius reconstruction error, as recorded when finding the optimal number of signatures, for each cancer type separately with values for the baseline NMF, AE-Cox, and MTAE-Cox. AE-Cox's and MTAE-Cox's reconstruction error is multiplied by the number of samples per fold to compensate for the normalisation. The number of signatures ranges from 2 to 10.

clustering step as described in section 2.2.2. Across 4 cancers - SKCM, HNSC, OV, and GBM - we find two main behaviours with respect to the stability of the derived signatures (Figure 5) over five cross-validation folds.

For both SKCM and HNSC, NMF is more stable than AE-Cox for lower number of signatures, while MTAE-Cox yields the least stable signatures overall. NMF starts with a high average stability of 0.99 and steadily decreases to 0.34 for SKCM and ranges from a maximum stability of 0.99 to a minimum of 0.44 for HNSC. AE-Cox's stability ranges from 0.71 to 0.41 and 1.00 to 0.60 with cancers SKCM and HNSC respectively. Contrary to NMF, AE-Cox's stability stops decreasing towards 10 signatures for SKCM or even improves for HNSC. After a certain number of signatures, AE-Cox's stability is higher than NMF's stability (SKCM: at 9 signatures with mean stability of 0.41 (AE-Cox) and 0.40 (NMF); HNSC: at 8 signatures with mean stability of 0.67 (AE-Cox) and 0.58 (NMF)) Finally, MTAE-Cox's stability behaves similarly to NMF's stability as it decreases with higher numbers of signatures. However, it does so with a lower stability overall compared to the other two methods (SKCM: 0.63 to 0.09; HNSC 0.27 to -0.01).

On OV and GBM cancer, NMF shows a low stability which flattens out towards 10 signatures. Contrary to SKCM and HNSC, NMF's stability does not start at 1.00; instead, it ranges from 0.44 to 0.15 and from 0.35 to 0.07 with OV and GBM respectively. MTAE-Cox has a slightly higher stability than NMF; 0.51 to 0.13 with OV and 0.47 to 0.11 with GBM. AE-Cox has a higher stability than the other two methods, namely 0.70 to 0.50 for OV and 0.56 to 0.29 for GBM.

For OV and GBM cancers all methods are too unstable to use more than two signatures, following the procedure explained in section 2.2.2, as the stability is always below 0.7. Furthermore, MTAE-Cox is too unstable to use more than two signatures for all four cancers. For SKCM and HNSC cancers, 6 signatures can be chosen for NMF (stabilities of 0.72 SKCM and 0.80 HNSC) and for SKCM 3 signatures can be used with AE-Cox (stability of 0.71).

We investigated the impact of the number of signatures on the reconstruction error, which is expected to decrease when the number of signatures increases. When more signatures

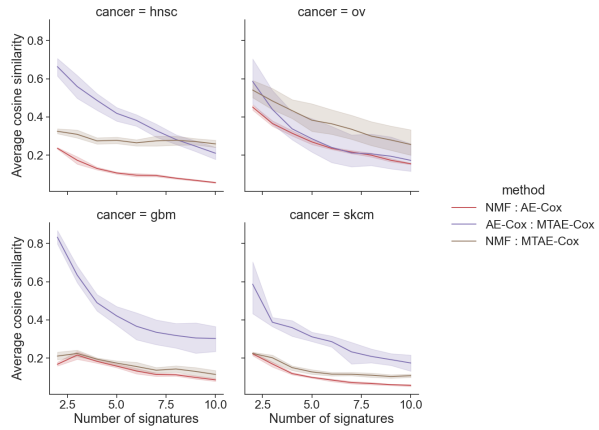
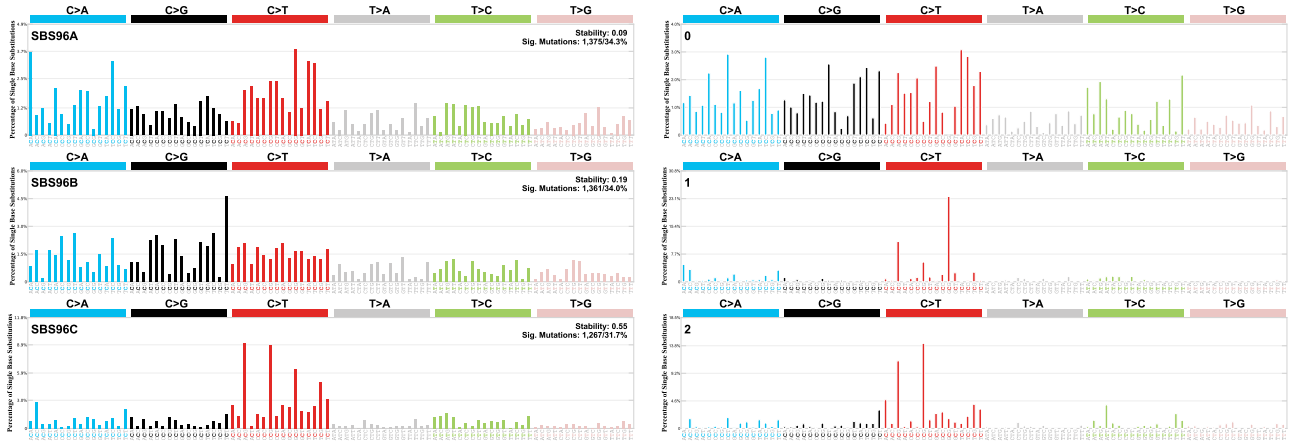


Fig. 7. Average cosine similarity of signatures between pairs of baseline method NMF, AE-Cox, and MTAE-Cox for different of signatures. In each fold, the signatures of two methods are matched with each other such that the average cosine similarity over the pairs of signatures is the highest. The plotted line is the mean of this average cosine similarity.

are used, the mutational catalogs can be decomposed in a more fine-grained way resulting into a smaller error. For all cancer types with both NMF and AE-Cox, the reconstruction error decreases with more signatures, as expected (Figure 6). However, NMF's reconstruction error starts and ends with lower values and differs less between the cancer types than AE-Cox's reconstruction error. For MTAE-Cox the reconstruction error does not decrease, but has very small changes instead which do not exclusively decrease. This can be explained by the optimization of the auto-encoder being determined not only by the reconstruction error itself, but also by the Cox loss. When a balance between both errors is found, the reconstruction error is no longer influenced by the optimization since decreasing it would increase the Cox loss, which would increase the total loss.



(a) Signatures found by SigProfiler's NMF

(b) Signatures found by AE-Cox

Fig. 8. Derived signatures for OV cancer on fold 3 by SigProfiler's NMF and AE-Cox during the determination of the number of signatures. The centroids of the clusters used to determine the stability are shown.

3.2.2. Similarity of signatures

To assess whether MTAE-Cox finds different signatures from AE-Cox and how closely AE-Cox's signatures relate to the baseline NMF signatures, we compare the cosine similarity between pairs of signatures derived across methods. For each pair of methods (NMF/AE-Cox, AE-Cox/MTAE-Cox, NMF/MTAE-Cox) the signatures of the two methods are matched per fold as explained in section 2.2.3, such that the average similarity is the highest (Figure 7).

For all cancer types, except for OV cancer, AE-Cox and MTAE-Cox have the most similar signatures, with a cosine similarity that ranges from 0.633 to 0.210 for HNSC, 0.833 to 0.303 for GBM, and 0.587 to 0.174 for SKCM. For OV cancer, the similarity between AE-Cox's and MTAE-Cox's signatures is closer to the cosine similarities between other comparisons (0.586 to 0.173). The similarity between NMF and MTAE-Cox is stable over the number of signatures for HNSC (averaged 0.325 to 0.259), although it is low. For other cancer types this similarity decreases with an increasing number of signatures (0.541 to 0.256 for OV, 0.211 to 0.114 for GBM, and 0.225 to 0.108 for SKCM). NMF's and AE-Cox's signatures' similarities decrease with a higher number of signatures for all cancer types (0.236 to 0.056 HNSC, 0.452 to 0.155 OV, 0.216 to 0.086 GBM, and 0.223 to 0.058 SKCM).

In summary, although the signatures derived by MTAE-Cox differ from the ones found by AE-Cox, they are more similar to each other than to the baseline NMF method for all cancers except ovarian. AE-Cox and MTAE-Cox having the most similar signatures can be explained by the fact that the same method is used to derive the signatures, namely the non-negative auto-encoder. The similarity between AE-Cox's and MTAE-Cox's signatures are comparable over all cancer types. Moreover, the signatures derived by AE-Cox and MTAE-Cox still differ, especially with a larger number of signatures, which can be caused by the impact of the survival prediction problem to the signatures of MTAE-Cox and by the larger instability that comes with more signatures.

The similarity between NMF and MTAE-Cox suffers from both the change in method (NMF to auto-encoder) and the integration of the Cox loss. However, MTAE-Cox's signatures are more similar to NMF's signatures than to AE-Cox's signatures, especially for HNSC. This could be due to the low stability of MTAE-Cox's signatures, which is highest for HNSC, causing the centroids of the clusters of signatures to be more noisy. Analyzing the derived signatures NMF and AE-Cox further indeed explains why the unstable signatures may be more similar to NMF's signatures. Most notably, AE-Cox isolates just mutations with frequency in new signatures from all other mutation types whereas NMF adds more low frequency mutations to all signatures (Figure 8). The more signatures there are, the more signatures show these isolated mutation types. The difference in behaviour could be a side effect of the non-negativity constraint set on the weights of the auto-encoder. The optimiser can optimise the weights to negative values, which are then set back to 0 after an epoch. This leads to weights of value 0 which corresponds to the mutation types that have 0% frequency in a signature.

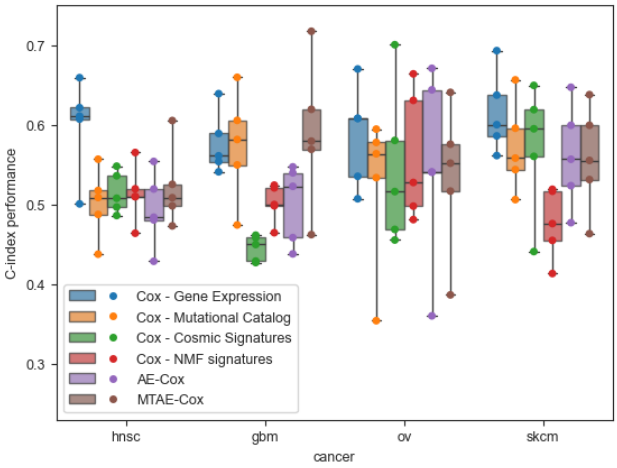


Fig. 9. C-index performance of the benchmark models, AE-Cox, and MTAE-Cox on the test set with 5-fold cross validation. The plotted dots are the C-index resulting from each fold. The models are run and plotted separately for each cancer.

3.3. Predictive performance

To assess the predictive performance of AE-Cox and MTAE-Cox, we compare their predictions' concordance index (C-index, see Figure 9) to that of a variety of benchmark methods across different feature types and extraction methods over the four selected cancer types.

3.3.1. Survival prediction performance comparison to benchmark models

The performance of the benchmark models, AE-Cox, and MTAE-Cox relative to each other varies between the cancer types and is thus dependent on the dataset.

For HNSC the Cox model on gene expression data (Cox-GE) achieves the best performance (median of 0.611). The Cox model with the mutational catalog (Cox-Catalog) performs similarly to the Cox model on exposures of COSMIC signatures (Cox-Cosmic) and the Cox model based on the exposures of NMF (Cox-NMF) (0.509, 0.508, 0.510 median respectively). AE-Cox drops in performance compared to the baseline methods (0.484), but MTAE-Cox again improves over AE-Cox to a similar performance (0.509) as the baseline models. The variance between runs is higher than the difference in performances, therefore we cannot conclude one method to perform better than another.

MTAE-Cox and Cox-Catalog outperform all other methods with GBM (median C-index of 0.579 and 0.581 respectively), including Cox-GE (0.561) which does not perform best only for GBM. Cox-Cosmic performs worst of the six methods (0.450), which is improved by Cox-NMF (0.500) and further improved by AE-Cox (0.521).

Cox-GE performs well with OV (median C-index of 0.608). Cox-Catalog performs worse (0.564) and Cox-Cosmic decreases the performance further (0.516). The signatures derived from the patients on which the survival prediction is run as well perform better than the exposures of COSMIC signatures (Cox-NMF 0.528, AE-Cox 0.541, MTAE-Cox 0.552).

Cox-GE performs best out of all the tested methods on SKCM (median C-index 0.600). Cox-Catalog performs worse (0.558), but Cox-Cosmic performs similar to Cox-GE (0.595). Cox-NMF performs worst of the methods (0.476) and AE-Cox performs similar to MTAE-Cox (0.557 and 0.555 respectively).

The Cox-GE method performs best in three out of the four cancer types (HNSC, OV, SKCM). First, this can be explained by gene expression data having more features than the other methods (20,530 genes compared to 96 mutation types, 79 COSMIC signatures, and less than 10 signatures for NMF, AE-Cox, and MTAE-Cox). Second, the gene expression data is obtained by measuring how much genes are transcribed into mRNA, which can then be turned into functioning proteins [32]. Gene mutations in a cell can cause the difference in gene expression, but the gene expression data capture the result of more complex processes than the plain gene mutations do. Using gene expression then allows for a better performing model; however, since it is harder to determine what caused a change in the gene expression, the model will be less interpretable. For well-known mutational signatures, on the other hand, processes behind the signatures are mostly known. This would allow for a better explanation for the prediction of the model.

For three out of four cancer types (HNSC, GBM, OV) MTAE-Cox performs better than AE-Cox and for another three cancer types (GBM, OV, SKCM) it outperforms Cox based on NMF. This is to be expected, as the signatures are tuned towards the survival prediction problem. For GBM MTAE-Cox sees the biggest increase in performance over Cox based on NMF and AE-Cox, while it has a small decrement in performance for HNSC compared to Cox on NMF and for SKCM compared to AE-Cox.

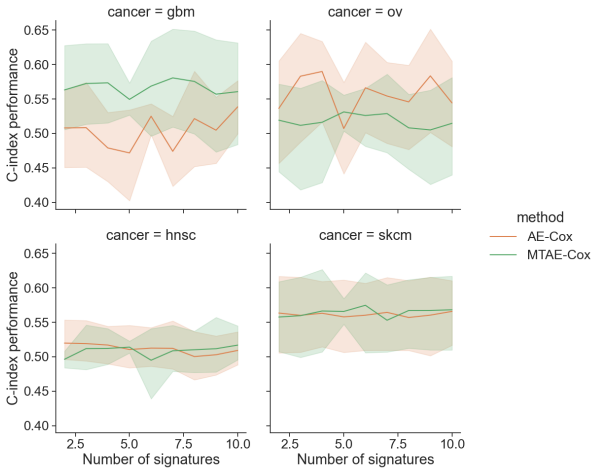


Fig. 10. C-index performance of the test set over the number of signatures in 5-fold cross validation. The range of number of signatures is 2 to 10. The performance for each cancer is plotted separately.

AE-Cox performs better than Cox with NMF signatures for GBM, OV, and SKCM. This indicates that the signatures found by the auto-encoder are better suited for the survival prediction problem than the signatures that are derived by the SigProfiler's NMF method.

3.3.2. Performance does not improve with unstable signatures

To analyse whether the performance of AE-Cox and MTAE-Cox would be better if we were to allow unstable signatures, the C-index over a range of number of signatures (from 2 to 10, Figure 10). For HNSC and SKCM, both AE-Cox and MTAE-Cox show a stable C-index, which is therefore independent of the number of signatures. Furthermore, the performances of AE-Cox and MTAE-Cox are very similar for these two cancer types. With GBM and OV, MTAE-Cox still has a stable performance over a varying number of signatures. AE-Cox has a performance that varies more, but there is no clear trend to be identified in its performance.

Since there is no significant improvement in the predictive performance of AE-Cox and MTAE-Cox when using unstable signatures, it is better to use a low number of signatures (i.e. 2) to obtain the most stable signatures that AE-Cox and MTAE-Cox can derive.

3.4. Comparison of AE-Cox and MTAE-Cox signatures to COSMIC signatures

In order to find biological explanations on the survival prediction of patients, we investigate the relation between the derived signatures to well-known COSMIC signatures, by analysing the cosine similarities between the derived signatures and the COSMIC signatures. The signatures AE-Cox and MTAE-Cox found could indicate what constitutes a higher or lower survival rate for patients, since the Cox model predicts the survival based on a patient's exposures to these signatures.

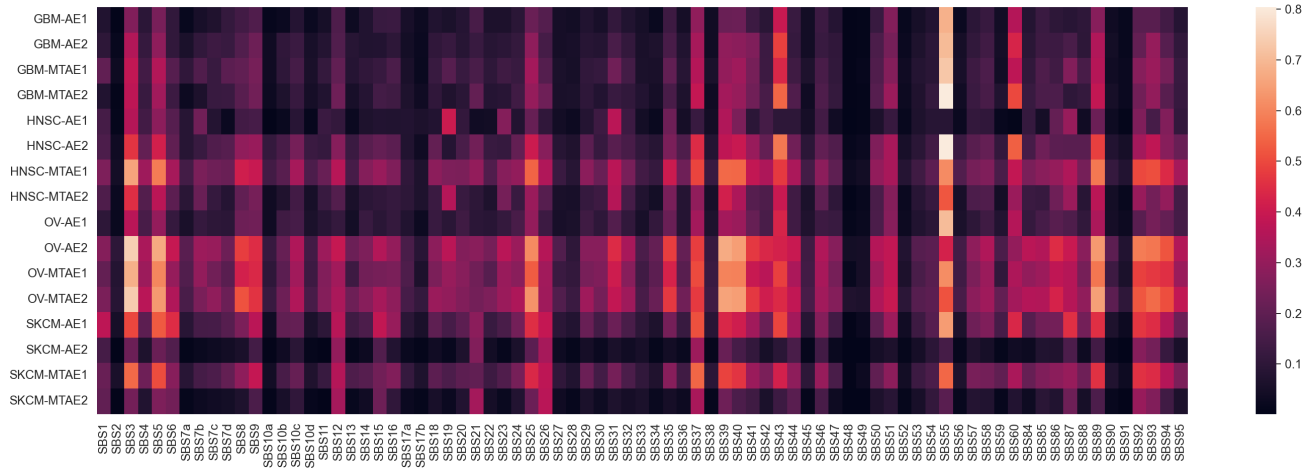


Fig. 11. Heatmap of the similarity between AE-Cox, MTAE-Cox signatures. Rows are the signatures from AE-Cox and MTAE-Cox, per cancer. The columns are the COSMIC signatures. The lighter the colour, the higher the cosine similarity between the two signatures.

Matching these signatures to COSMIC signatures can therefore uncover the biological processes linked to the cancer.

Furthermore, AE-Cox and MTAE-Cox find a small number of stable signatures, so we decompose these into COSMIC signatures to see whether they consist of multiple COSMIC signatures. The decomposition finds possible combinations of COSMIC signatures that can be combined into the found signatures.

3.4.1. Similarity based comparison between (MT)AE-Cox and COSMIC signatures

To compare the signatures found by AE-Cox and MTAE-Cox to the COSMIC signatures, the cosine similarity between these is shown in Figure 11. The signatures are derived by creating a single train-test split of 90%-10% and performing a stability analysis. The resulting centroids from the signature clusters are used in the comparison with COSMIC signatures. All methods and cancers reported a select number of signatures of 2, except for SKCM AE-Cox which selected 3 signatures. Since the predictive performance of AE-Cox and MTAE-Cox does not increase with less stable signatures, we have used 2 signatures for each method and cancer combination in the analysis of the signatures.

The signatures from OV cancer are the most similar to COSMIC signatures found (average similarity of cosmic signatures to the closest derived signature of 0.233 for AE-Cox and 0.301 for MTAE-Cox), followed by HNSC (0.161 and 0.222), SKCM (0.157 and 0.162) and lastly GBM (0.125 and 0.156). The MTAE signatures are more similar to COSMIC signatures than the AE signatures for all cancer types, suggesting that exposures to the COSMIC-like signatures can be a more valuable indicator for the survival prediction problem. OV has the largest increase in average similarity to COSMIC signatures from AE-Cox to MTAE-Cox with 0.068, which corresponds to a 29.1% increase. This is followed by HNSC (0.062, 38.4%), GBM (0.031, 24.9%), and finally SKCM (0.006, 3.63%).

We identify COSMIC signatures that are similar to AE-Cox's and MTAE-Cox's signatures, specifically the COSMIC signatures with a cosine similarity larger than 0.6. The most similar COSMIC signature to all found signatures across all cancer types is SBS55 (ranging from a cosine similarity 0.611 to 0.804 over all cancer types and both methods, where the highest similarity per method-cancer pair is chosen). SBS55 is not associated with a process (yet), but may just be a sequencing artifact in which case it would be expected to be present equally in all cancer types [33]. SBS3 has an average similarity of 0.45 and is mostly similar to the signatures extracted from the OV cancer, specifically to one of AE-Cox's signatures (cosine similarity of 0.744) and to both MTAE-Cox's signatures (0.679 and 0.736). SBS3 has been linked to somatic BRCA1 and BRCA2 mutations, in among others OV cancer [34, 35]. SBS5 has similar similarities to signatures from OV cancer as SBS3 (OV-AE2: 0.667, OV-MTAE1: 0.601, OV-MTAE2: 0.637). SBS5 may be related to tobacco smoking and is more commonly found in aged individuals [19]. SBS25, SBS39, SBS40, and SBS89 are all similar to one of the two signatures derived by either AE-Cox or MTAE-Cox (cosine similarities ranging from 0.608 to 0.670 for AE-Cox and 0.624 to 0.656 for MTAE-Cox). The processes behind these COSMIC signatures are generally unknown, but SBS25 could be caused by chemotherapy, a treatment often applied to cancer patients. With data from real patients, such as in the TCGA dataset, it would be possible that patients undergo chemotherapy and therefore have exposure to this signature.

Finding COSMIC signatures that are similar to AE-Cox's and MTAE-Cox's signatures, which are associated to biological processes that fit the cancer the signatures are derived from, indicate that AE-Cox and MTAE-Cox can derive signatures that are indicative of the processes relating to the cancer. Combined with MTAE-Cox's exposures, processes that are especially indicative of a patient's survival could be found.

3.4.2. Decomposition of MTAE-Cox's signatures into COSMIC signatures

The decomposition of signatures into COSMIC signatures finds whether the derived signatures are linear combinations of certain COSMIC signatures. We decompose MTAE-Cox's signatures into COSMIC signatures using NNLS from Sigprofiler which minimizes the L2 error to devise a list of COSMIC signatures for each of the signatures that are being decomposed [30].

The MTAE-Cox's signature from the OV cancer is decomposed the most successfully, judged by its highest cosine

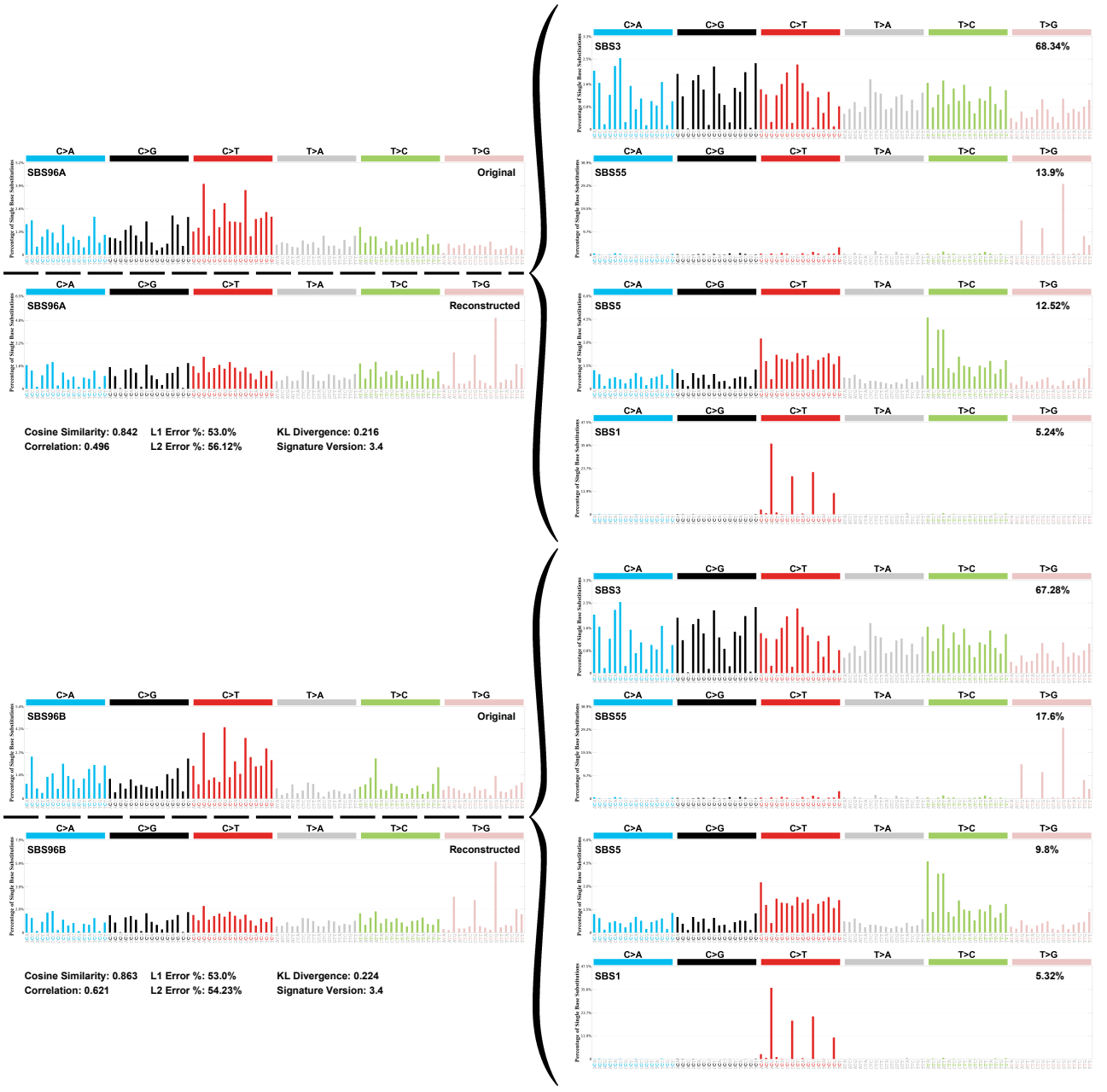


Fig. 12. Decomposition plot of MTAE-Cox's signatures from Ovarian cancer. The decomposition of two MTAE-Cox signatures are shown, denoted by SBS96A and SBS96B. On the left, the original signature and the reconstructed signature with the decomposition is shown. MTAE-Cox's signatures are decomposed in the signatures shown on the right, with the decomposition weights in the top right of each signature plot.

similarity between the original and reconstructed signatures. The other signatures' decomposition finds a less similar reconstructed signature to the original ones, indicating that these signatures could be new signatures, that incorporate more survival characteristics, instead. OV's signatures decomposition being the most successful is to be expected, since OV's signatures are also the most similar to the COSMIC signatures. Figure 12 shows the decomposition plots for the MTAE-Cox's signatures of OV cancer. The decomposition finds the same COSMIC signatures list for each signature (SBS3, SBS5, SBS55, and SBS1), but with different weights. SBS1 is the only signature we did not identify already from the cosine similarities. SBS1 has been linked to spontaneous deamination of 5-methylcytosine [34]. Furthermore, SBS1 may register the number of mitosises that a cell has experienced [36]. It has been found that the mitotic rate of cells can be linked to the survival time of patients [37]. SBS1 can therefore indeed be an indicative signature for the survival prediction. The additional COSMIC signature shows that, besides the similarity, the decomposition into COSMIC signatures can help to identify additional relevant biological processes related to the cancer.

4. Conclusion

We propose and have implemented two novel methods for both mutational signature extraction and survival prediction: Auto-Encoder Cox (AE-Cox) and Multi Task Auto-Encoder Cox (MTAE-Cox). AE-Cox combines a non-negative linear auto-encoder, that optimises its weights to signatures and computes the exposures as its latent space, with a Cox model to perform survival prediction disjointly. AE-Cox's signatures and exposures are therefore independent of the survival prediction problem. MTAE-Cox is a variation on AE-Cox that optimises the auto-encoder and the Cox model jointly, thereby integrating the survival prediction with the decomposition of the mutational catalog into mutational signatures and exposures.

The stability of signatures derived by AE-Cox, MTAE-Cox, and the baseline method NMF show no clear trend over the various cancer types. Integrating the survival prediction problem with the signature extraction using MTAE-Cox yields signatures that are better suited for survival prediction. MTAE-Cox outperforms AE-Cox in 3 of the 4 cancer types according to the C-index of its survival prediction. However, AE-Cox and MTAE-Cox do not improve survival prediction based on gene expression, which outperforms both in 3 out of 4 cancer types.

The signatures derived by AE-Cox and MTAE-Cox have found to be similar to certain COSMIC signatures. Overall, MTAE-Cox's signatures are more similar to COSMIC signatures than AE-Cox's signatures, indicating that COSMIC signatures can be indicative to a patient's survival. This effect is most noticeable in OV cancer, where the average cosine similarity over all COSMIC signatures increases by 36.1%. SBS55, which could be linked to sequencing artifacts, is the COSMIC signature that is most similar to all signatures derived by AE-Cox and MTAE-Cox (average cosine similarity of 0.55). Additionally, SBS3 (0.41) which is related to BRCA1 and BRCA2 mutations in OV cancer and SBS5 (0.41) which may be related to tobacco smoking, are found to be similar to the derived signatures. The decomposition of MTAE-Cox signatures indicate one more relevant COSMIC signature, SBS1, that could indicate the number of mitosises, which can be indicative of the survival of a patient.

AE-Cox and MTAE-Cox are limited by the number of features they require – survival time, survival status, mutation data – which decreases the number of samples that are usable from TCGA.

To improve the predictive performance of AE-Cox and MTAE-Cox a combination of multiple omics can be used for the survival prediction. The auto-encoder would not change, but the exposures computed by the auto-encoder would be combined with gene expression. This could combine a better prediction with the additional explainability that the mutational signatures offer.

To conclude, we have shown the possibility to derive mutational signatures using an auto-encoder enabling the integration of survival prediction with the mutational signature extraction. This does not yield a better prediction than currently available, but gives the opportunity to find new signatures that are relevant to the survival of patients.

References

1. WHO. Cancer. *World Health Organization*, 2022. https://www.who.int/news-room/fact-sheets/detail/cancer (accessed: 24 January 2025).
2. Kimberly D Miller, Leticia Nogueira, Angela B Mariotto, Julia H Rowland, K Robin Yabroff, Catherine M Alfano, Ahmedin Jemal, Joan L Kramer, and Rebecca L Siegel. Cancer treatment and survivorship statistics, 2019. *CA: a cancer journal for clinicians*, 69(5):363–385, 2019.
3. Jon Zugazagoitia, Cristiano Guedes, Santiago Ponce, Irene Ferrer, Sonia Molina-Pinelo, and Luis Paz-Ares. Current challenges in cancer treatment. *Clinical therapeutics*, 38(7):1551–1566, 2016.
4. Adrienne G Waks and Eric P Winer. Breast cancer treatment: a review. *Jama*, 321(3):288–300, 2019.
5. Mitch Golant, Tamara Altman, and Chloe Martin. Managing cancer side effects to improve quality of life: a cancer psychoeducation program. *Cancer nursing*, 26(1):37–44, 2003.
6. Volker Schirrmacher. From chemotherapy to biological therapy: A review of novel concepts to reduce the side effects of systemic cancer treatment. *International journal of oncology*, 54(2):407–419, 2019.
7. M Berkan Sesen, Timor Kadir, Rene-Banares Alcantara, John Fox, and Sir Michael Brady. Survival prediction and treatment recommendation with bayesian techniques in lung cancer. In *AMIA annual symposium proceedings*, volume 2012, page 838. American Medical Informatics Association, 2012.
8. David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.
9. Mohanad Mohammed, Innocent B Mboya, Henry Mwambi, Murtada K Elbashir, and Bernard Omolo. Predictors of colorectal cancer survival using cox regression and random survival forests models based on gene expression data. *PLoS One*, 16(12):e0261625, 2021.
10. Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18:1–12, 2018.
11. Kun-Hsing Yu, Ce Zhang, Gerald J Berry, Russ B Altman, Christopher Ré, Daniel L Rubin, and Michael Snyder. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature communications*, 7(1):12474, 2016.
12. Amir Sorayaie Azar, Samin Babaei Rikan, Amin Naemi, Jamshid Bagherzadeh Mohasefi, Habibollah Pirnejad, Matin Bagherzadeh Mohasefi, and Uffe Kock Wiil. Application of machine learning techniques for predicting survival in ovarian cancer. *BMC medical informatics and decision making*, 22(1):345, 2022.
13. Adrián Mosquera Orgueira, Marta Sonia González Pérez, José Ángel Díaz Arias, Beatriz Antelo Rodríguez, Natalia Alonso Vence, Ángeles Bendaña López, Aitor Abuín Blanco, Laura Bao Pérez, Andrés Peleteiro Raíndo, Miguel Cid López, et al. Survival prediction and treatment optimization of multiple myeloma patients using machine-learning models based on clinical and gene expression data. *Leukemia*, 35(10):2924–2935, 2021.
14. Wessel N Van Wieringen, David Kun, Regina Hampel, and Anne-Laure Boulesteix. Survival prediction using gene

expression data: a review and comparison. *Computational statistics & data analysis*, 53(5):1590–1603, 2009.
15. Fangzhou Yan and Yi Feng. A two-stage stacked-based heterogeneous ensemble learning for cancer survival prediction. *Complex & Intelligent Systems*, 8(6):4619–4639, 2022.
16. S Perwez Hussain, Lorne J Hofseth, and Curtis C Harris. Radical causes of cancer. *Nature Reviews Cancer*, 3(4):276–285, 2003.
17. Richard Doll and Richard Peto. The causes of cancer: quantitative estimates of avoidable risks of cancer in the united states today. *JNCI: Journal of the National Cancer Institute*, 66(6):1192–1308, 1981.
18. Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Peter J Campbell, and Michael R Stratton. Deciphering signatures of mutational processes operative in human cancer. *Cell reports*, 3(1):246–259, 2013.
19. Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolo Bolli, Ake Borg, Anne-Lise Børresen-Dale, et al. Signatures of mutational processes in human cancer. *nature*, 500(7463):415–421, 2013.
20. John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, et al. Cosmic: the catalogue of somatic mutations in cancer. *Nucleic acids research*, 47(D1):D941–D947, 2019.
21. Helen Davies, Dominik Glodzik, Sandro Morganella, Lucy R Yates, Johan Staaf, Xueqing Zou, Manasa Ramakrishna, Sancha Martin, Sandrine Boyault, Anieta M Sieuwerts, et al. Hrdetect is a predictor of brca1 and brca2 deficiency based on mutational signatures. *Nature medicine*, 23(4):517–525, 2017.
22. Sander Goossens, Yasin Tepeli, and Joana Gonçalves. Integrated learning of mutational signatures and prediction of dna repair deficiencies. Master's thesis, Delft University of Technology, 2022.
23. Tommaso Tofacchi, Sander Goossens, and Joana Gonçalves. Pseudo-labeling semi-supervised non-negative matrix factorization. Master's thesis, Delft University of Technology, 2024.
24. John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.
25. Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*, 2(5):401–404, 2012.
26. Deanna M Church, Valerie A Schneider, Tina Graves, Katherine Auger, Fiona Cunningham, Nathan Bouk, Hsiu-Chuan Chen, Richa Agarwala, William M McLaren, Graham RS Ritchie, et al. Modernizing reference genome assemblies. *PLoS biology*, 9(7):e1001091, 2011.
27. Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
28. David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.
29. Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.
30. SM Ashiqul Islam, Marcos Díaz-Gay, Yang Wu, Mark Barnes, Raviteja Vangara, Erik N Bergstrom, Yudou He, Mike Vella, Jingwei Wang, Jon W Teague, et al. Uncovering novel mutational signatures by de novo extraction with sigprofilerextractor. *Cell genomics*, 2(11), 2022.
31. Marcos Díaz-Gay, Raviteja Vangara, Mark Barnes, Xi Wang, SM Ashiqul Islam, Ian Vermes, Stephen Duke, Nithish Bharadhwaj Narasimman, Ting Yang, Zichen Jiang, et al. Assigning mutational signatures to individual samples and individual somatic mutations with sigprofilerassignment. *Bioinformatics*, 39(12):btad756, 2023.
32. Alvis Brazma and Jaak Vilo. Gene expression data analysis. *FEBS letters*, 480(1):17–24, 2000.
33. Ludmil B Alexandrov, Jaegil Kim, Nicholas J Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, Kyle R Covington, Dmitry A Gordenin, Erik N Bergstrom, et al. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, 2020.
34. Serena Nik-Zainal, Ludmil B Alexandrov, David C Wedge, Peter Van Loo, Christopher D Greenman, Keiran Raine, David Jones, Jonathan Hinton, John Marshall, Lucy A Stebbings, et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993, 2012.
35. Judit Zámborszky, Bernadett Szikriszt, Judit Zsuzsanna Gervai, Orsolya Pipek, Ádám Póti, Marcin Krzystanek, Dezső Ribli, János Márk Szalai-Gindl, István Csabai, Zoltán Szallasi, et al. Loss of brca1 or brca2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. *Oncogene*, 36(6):746–755, 2017.
36. Ludmil B Alexandrov, Philip H Jones, David C Wedge, Julian E Sale, Peter J Campbell, Serena Nik-Zainal, and Michael R Stratton. Clock-like mutational processes in human somatic cells. *Nature genetics*, 47(12):1402–1407, 2015.
37. John F Thompson, Seng-Jaw Soong, Charles M Balch, Jeffrey E Gershenwald, Shouluan Ding, Daniel G Coit, Keith T Flaherty, Phyllis A Gimotty, Timothy Johnson, Marcella M Johnson, et al. Prognostic significance of mitotic rate in localized primary cutaneous melanoma: an analysis of patients in the multi-institutional american joint committee on cancer melanoma staging database. *Journal of Clinical Oncology*, 29(16):2199–2205, 2011.