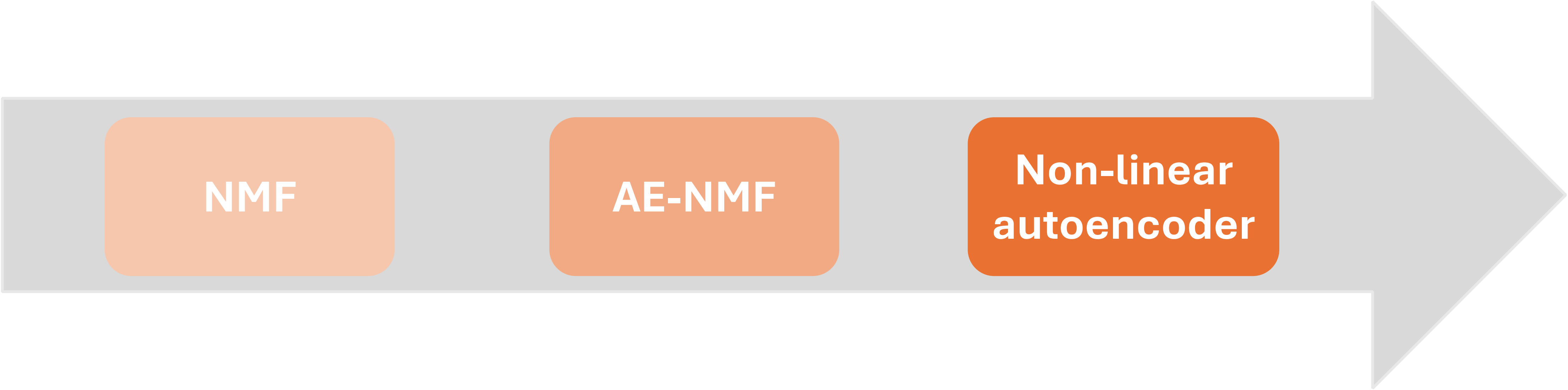
Autoencoders for mutational signature extraction

Cortinovis Nicola, Lucas Marta, Paladino Annalisa

UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

Introduction

Goal: Study mutational signature extraction using autoencoders

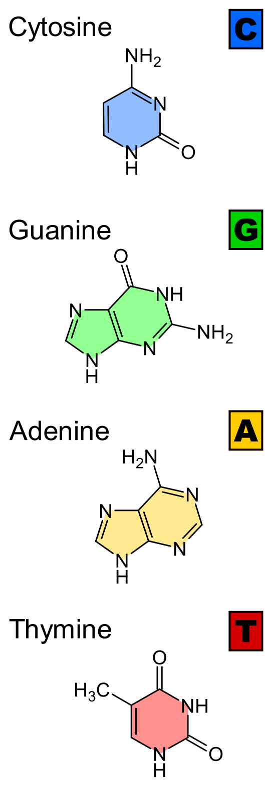


NMF

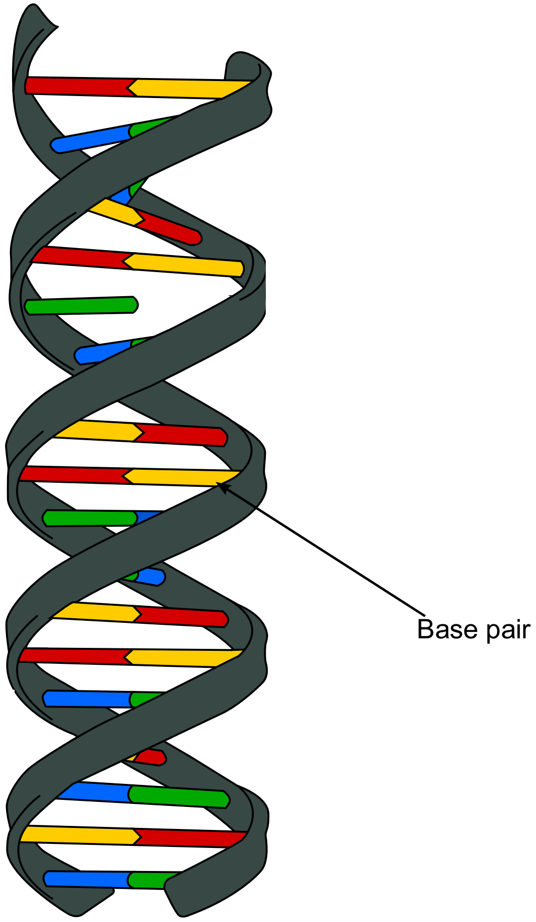
AE-NMF

Non-linear autoencoder

Mutational signatures



Nucleobases
of DNA



DNA
Deoxyribonucleic acid

Single Base Substitutions (SBS)

C > A   C > G   C > T   T > A   T > C   T > G

Context

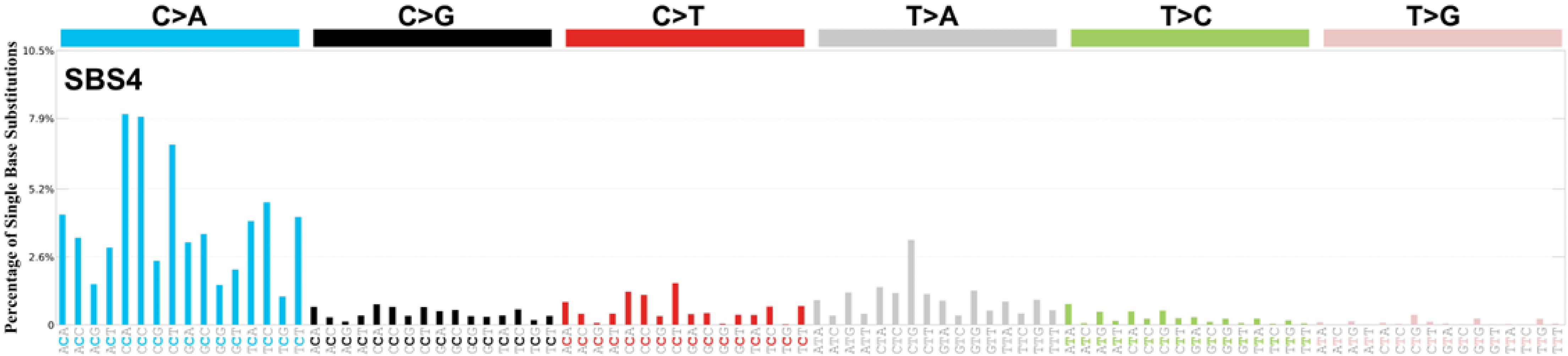
AC > AA
AC > AC
AC > AG
AC > AT
CC > AA
CC > AC
⋮

96 combinations

Types of mutational processes

Endogenous

Exogenous



SBS4 mutational signature linked to tobacco smoking, COSMIC (https://cancer.sanger.ac.uk/signatures/sbs/sbs4/)

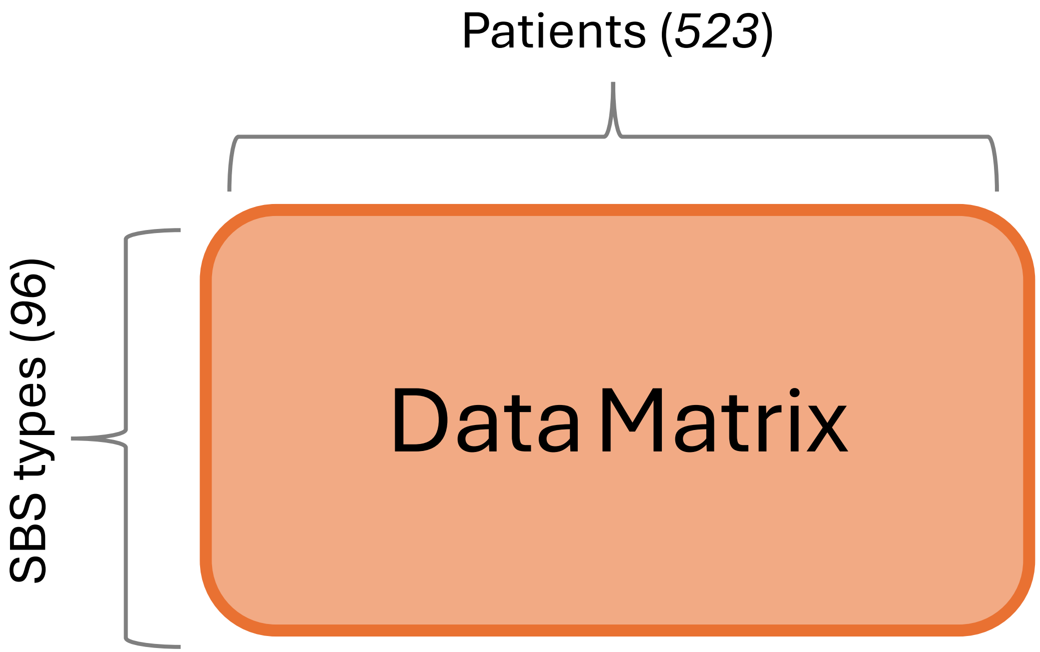
Data

Mutational catalogue of **ovarian cancer** from the **Genomics England** 100,000 Genomes Project (GEL)

523 whole genome sequences

Single Base Substitutions (SBS)

COSMIC v 3.4 SBS GRCh37 for signature *comparison*



Patients (*523*)

SBS types (96)

Data Matrix

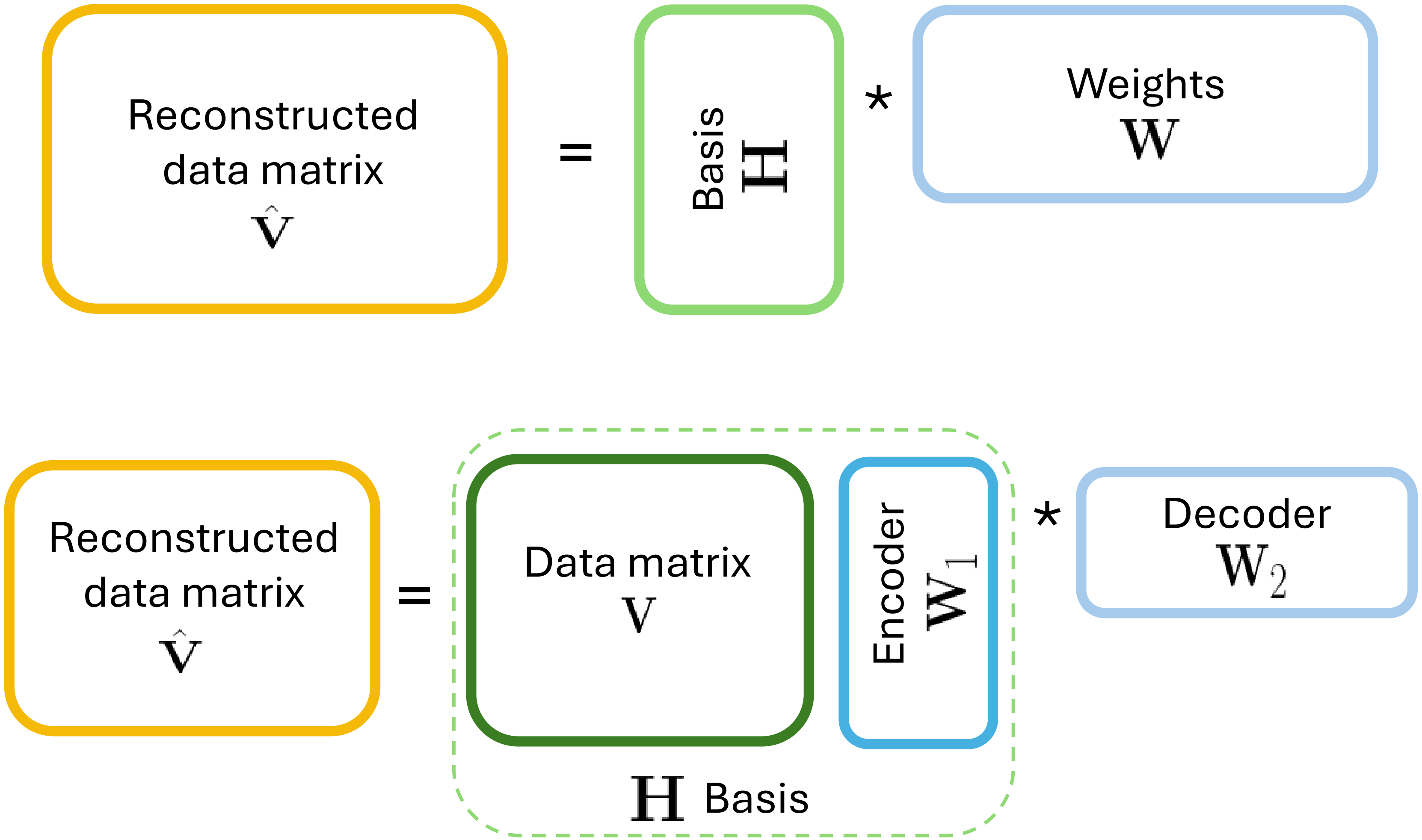
Non-Negative Matrix Factorization (NMF)

$$V \approx HW$$

NMF

$$V \approx VW_1W_2$$

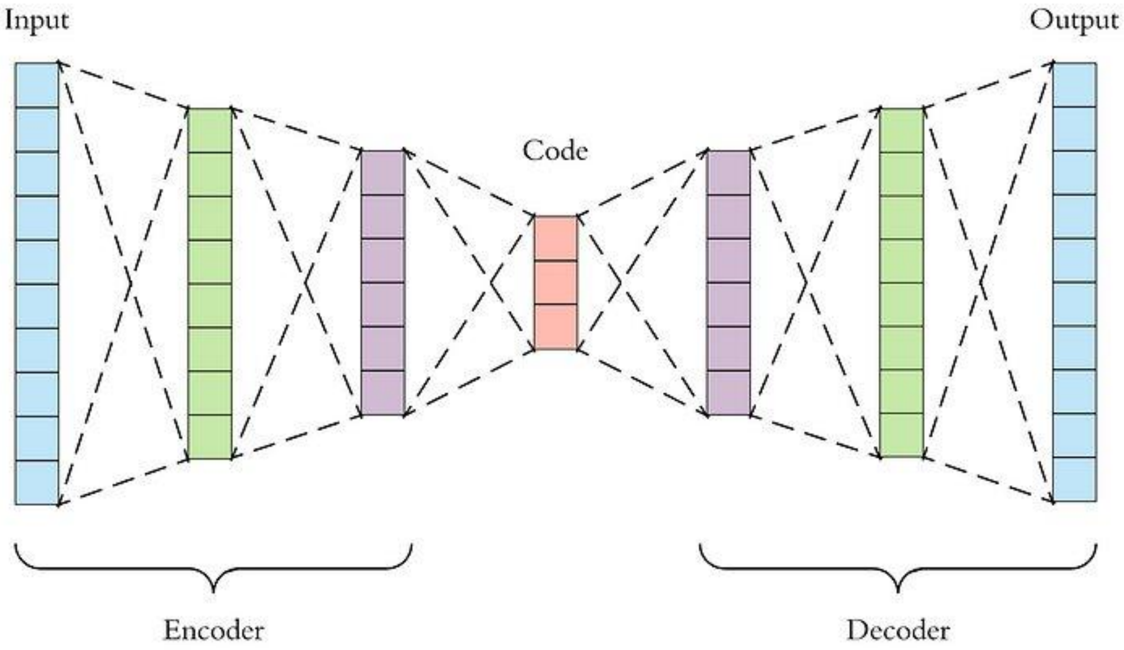
C-NMF/Autoencoder



Autoencoder

Autoencoder: neural network used for unsupervised learning to encode and decode data

- **Encoder:** compresses input into a lower-dimensional representation
- **Latent Space:** the compressed feature representation
- **Decoder:** reconstructs the input from the latent space



Non-negative autoencoder & Convex NMF *equivalence*

= 0

Shallow autoencoder: $$y_{pred} = \phi_{dec}(\phi_{enc}(\mathbf{V}\mathbf{W}_{enc}+\underline{\mathbf{b}_{enc}})\mathbf{W}_{dec}+\underline{\mathbf{b}_{dec}})$$

$$\phi_{enc}, \phi_{dec} : x \mapsto x$$

Weights constrained to be non-negative

$$\Downarrow$$

Convex NMF: $$y_{pred} = \mathbf{V}\mathbf{W}_1\mathbf{W}_2$$

Shallow-AE & PCA

PCA (Principal Component Analysis): method that finds the best lower-dimensional representation of data while preserving as much variance as possible

A shallow autoencoder with a linear activation function (identity) is mathematically **equivalent** to PCA:

- *Linear transformations*: autoencoder with an identity activation function only learns linear transformation, like PCA
- *Eigenvectors & Weights*: encoder learns a weight matrix that spans the same subspace as the PCA principal components
- *Recostruction Error*: the objective of training the autoencoder with MSE is to minimize the reconstruction error, same as PCA

PCA:

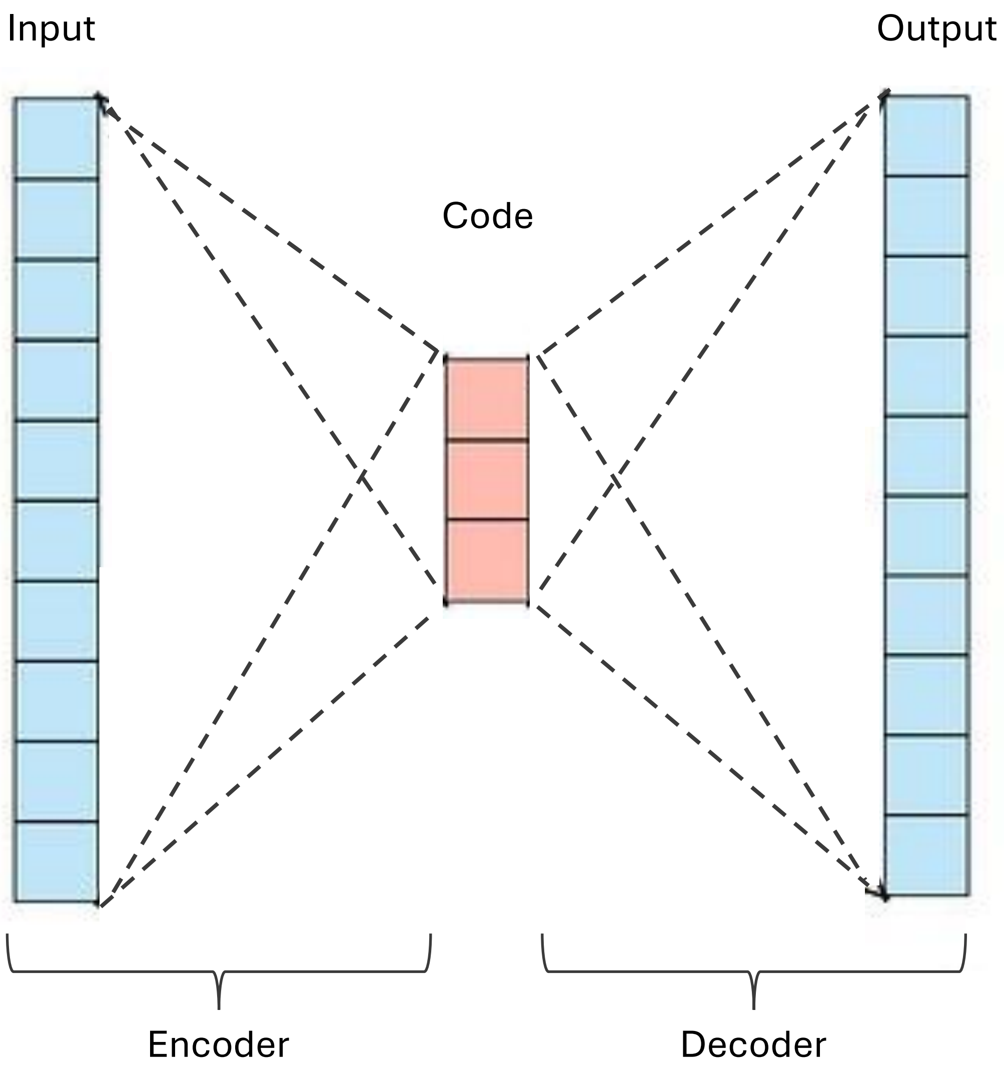
$$min_W ||X - XW^TW||^2$$

Shallow-AE:

$$min_W ||X - y_{pred}||^2 = ||X - XW_{enc}W_{dec}||^2$$

Autoencoder NMF (AE-NMF)

- **Input dimension:** 418
- **Latent dimension:** 4
- **Activation function:** Identity
- **Weight initialization:** $\mathcal{U}(0, 1)$



Training Procedure

- **Loss:** Frobenius Norm $\left\| X - \hat{X} \right\|_F$
- **Early stopping:** tolerance threshold (1e-10) on difference between subsequent loss values
- **ADAM** optimizer: learning rate fixed at 0.001
- **Weight clamping**: after gradient step and optimization, negative weights values are set to 0

Results AE-NMF vs NMF

- **Runs:** 30
- **Train-Test split :** 80% - 20%
- **Train error:** $\left\| X - \hat{X} \right\|_F$
- **Test procedure**: *freeze S,* randomly intialize *E,* minimize $\left\| X - \hat{X} \right\|_F$

Reconstruction Error	NMF	AE-NMF
Average train losses	1.48 * 10^4	1.86 * 10^4
Average test losses	1.05 * 10^5	1.64 * 10^4

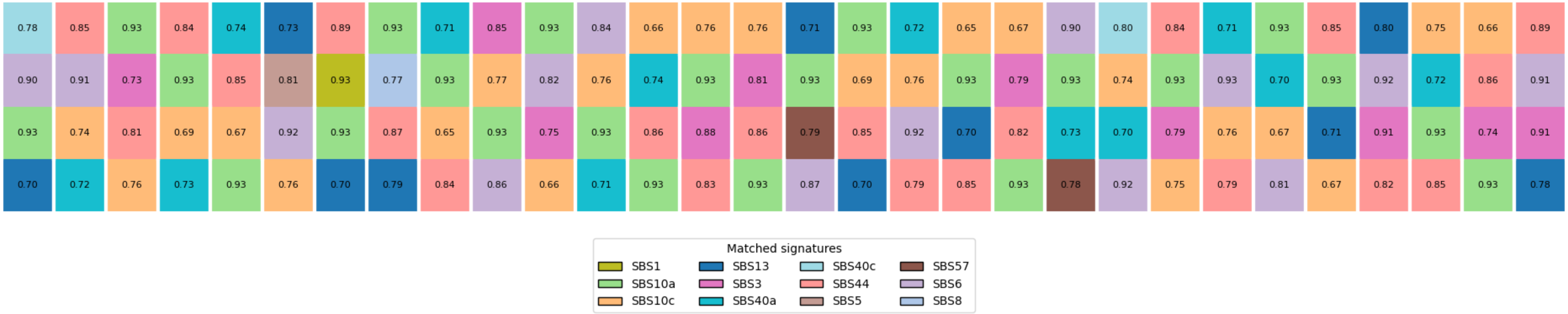
Cosine similarity with COSMIC

Reference for the extracted signature: **COSMIC v3.4 SBS GRCh37**

Cosine similarity matrix between *true* and *found* signatures $S_C(\tilde{\boldsymbol{h}}, \hat{\boldsymbol{h}}) = \frac{\tilde{\boldsymbol{h}} \cdot \hat{\boldsymbol{h}}}{||\tilde{\boldsymbol{h}}||||\hat{\boldsymbol{h}}||}$

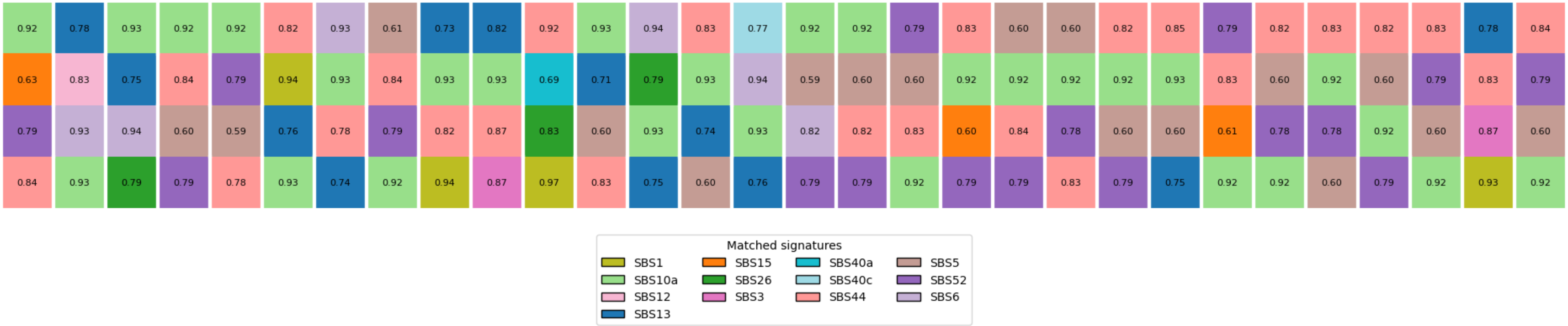
Signature matching is achieved through the **Hungarian algorithm** (linear assignment) ensuring the best available mapping

Cosine similarity comparison



Cosine similarity matrix NMF

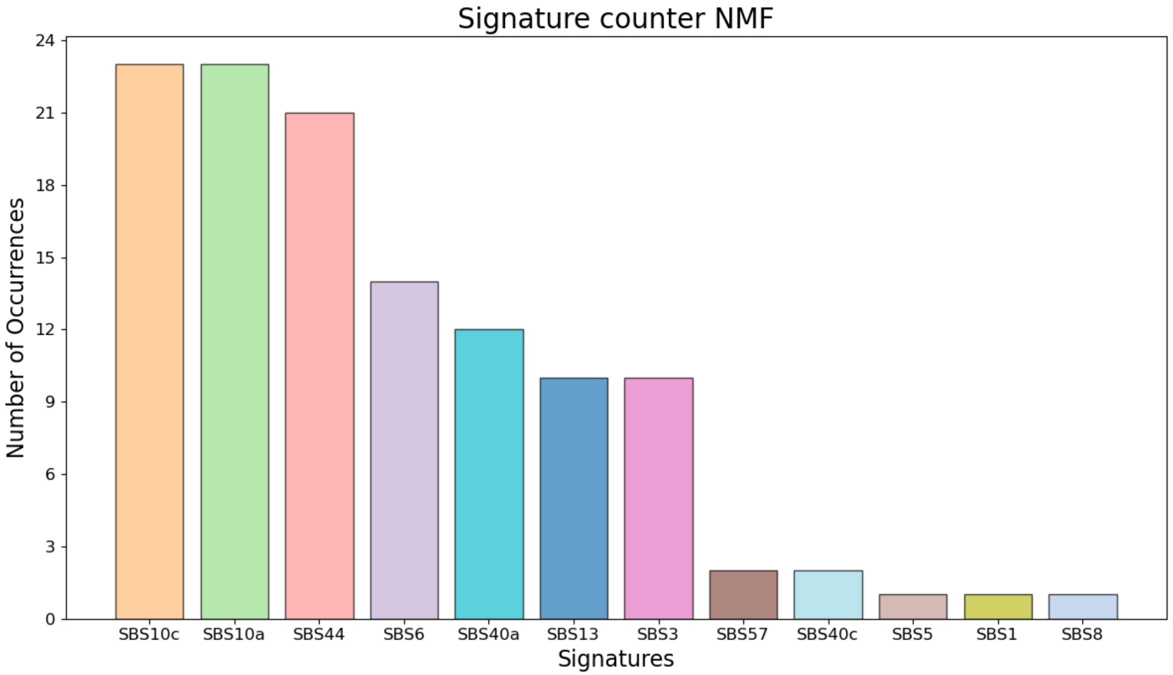
0.78	0.85	0.93	0.84	0.74	0.73	0.89	0.93	0.71	0.85	0.93	0.84	0.66	0.76	0.76	0.71	0.93	0.72	0.65	0.67	0.90	0.80	0.84	0.71	0.93	0.85	0.80	0.75	0.66	0.89
0.90	0.91	0.73	0.93	0.85	0.81	0.93	0.77	0.93	0.77	0.82	0.76	0.74	0.93	0.81	0.93	0.69	0.76	0.93	0.79	0.93	0.74	0.93	0.93	0.70	0.93	0.92	0.72	0.86	0.91
0.93	0.74	0.81	0.69	0.67	0.92	0.93	0.87	0.65	0.93	0.75	0.93	0.86	0.88	0.86	0.79	0.85	0.92	0.70	0.82	0.73	0.70	0.79	0.76	0.67	0.71	0.91	0.93	0.74	0.91
0.70	0.72	0.76	0.73	0.93	0.76	0.70	0.79	0.84	0.86	0.66	0.71	0.93	0.83	0.93	0.87	0.70	0.79	0.85	0.93	0.78	0.92	0.75	0.79	0.81	0.67	0.82	0.85	0.93	0.78

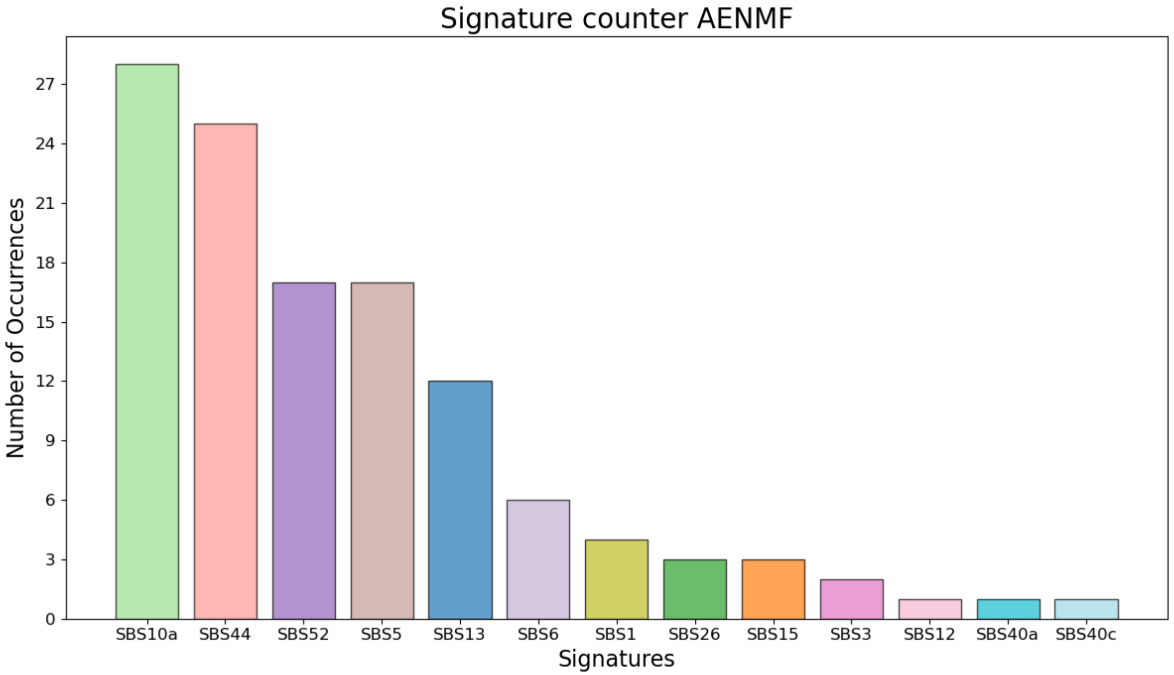


Cosine similarity matrix AENMF

0.92	0.78	0.93	0.92	0.92	0.82	0.93	0.61	0.73	0.82	0.92	0.93	0.94	0.83	0.77	0.92	0.92	0.79	0.83	0.60	0.60	0.82	0.85	0.79	0.82	0.83	0.82	0.83	0.78	0.84
0.63	0.83	0.75	0.84	0.79	0.94	0.93	0.84	0.93	0.93	0.69	0.71	0.79	0.93	0.94	0.59	0.60	0.60	0.92	0.92	0.92	0.92	0.93	0.83	0.60	0.92	0.60	0.79	0.83	0.79
0.79	0.93	0.94	0.60	0.59	0.76	0.78	0.79	0.82	0.87	0.83	0.60	0.93	0.74	0.93	0.82	0.82	0.83	0.60	0.84	0.78	0.60	0.60	0.61	0.78	0.78	0.92	0.60	0.87	0.60
0.84	0.93	0.79	0.79	0.78	0.93	0.74	0.92	0.94	0.87	0.97	0.83	0.75	0.60	0.76	0.79	0.79	0.92	0.79	0.79	0.83	0.79	0.75	0.92	0.92	0.60	0.79	0.92	0.93	0.92

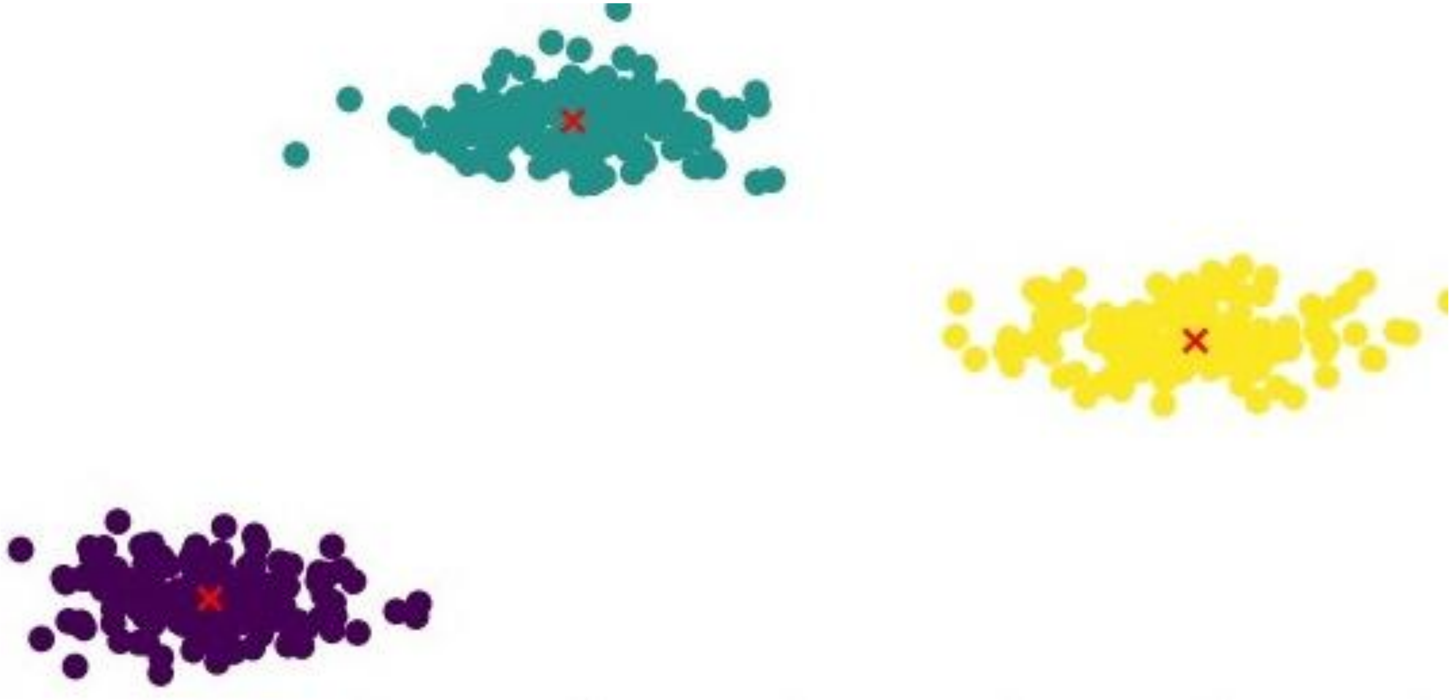
Signature counts comparison



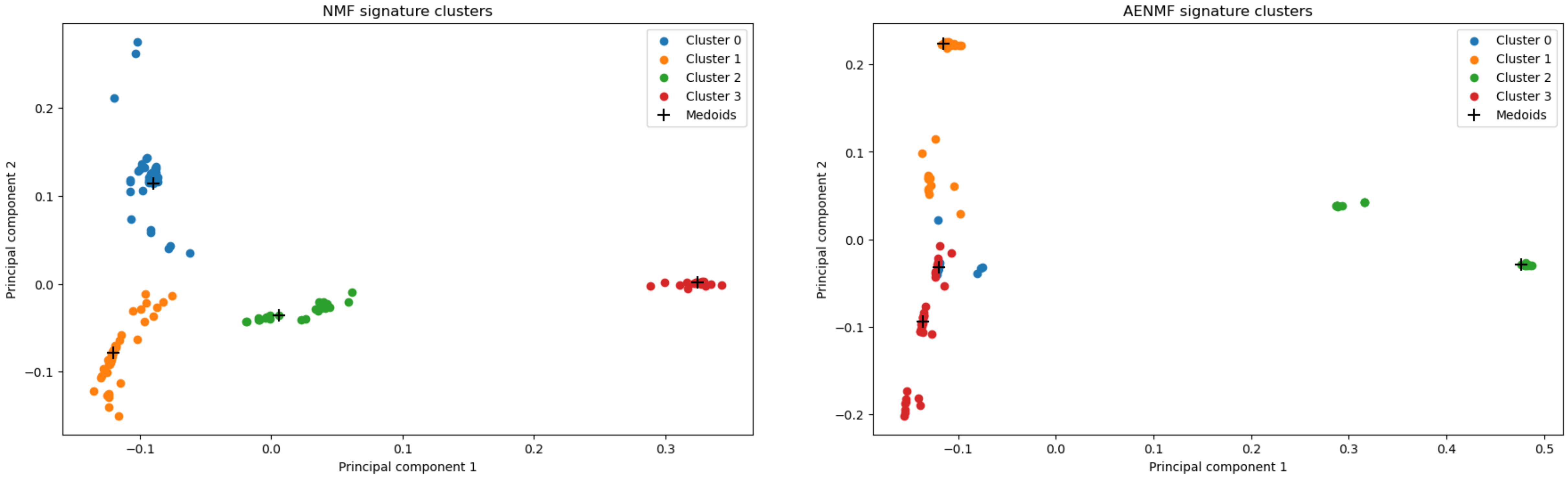


Consensus signatures

Consensus signatures are found via a **K-medoids** algorithm based around the **cosine similarity** measure between the extracted signatures

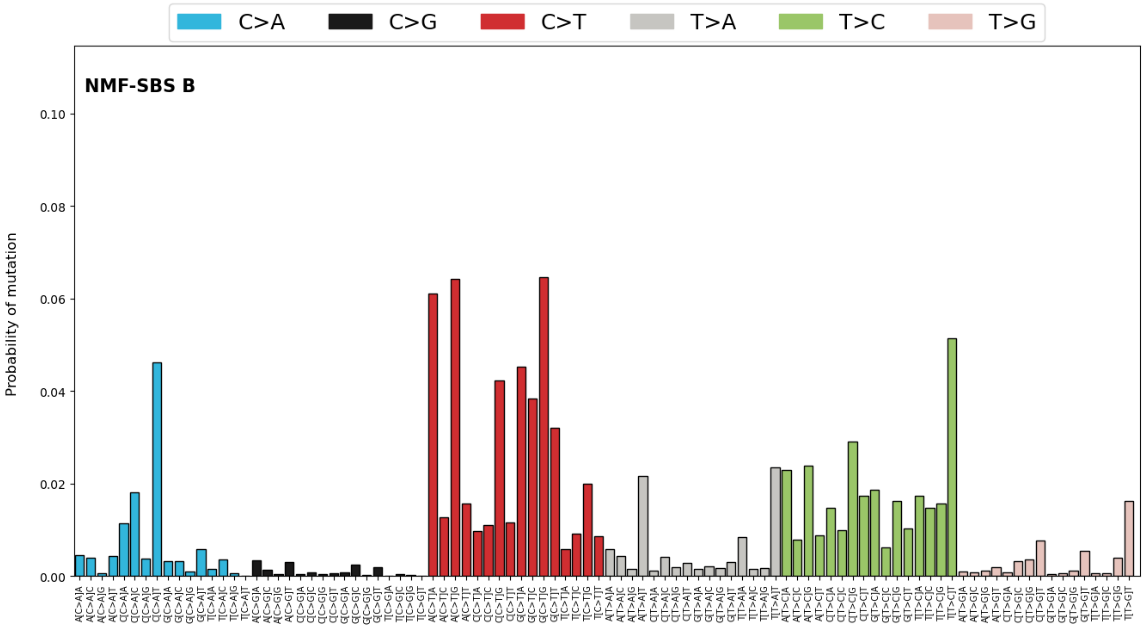


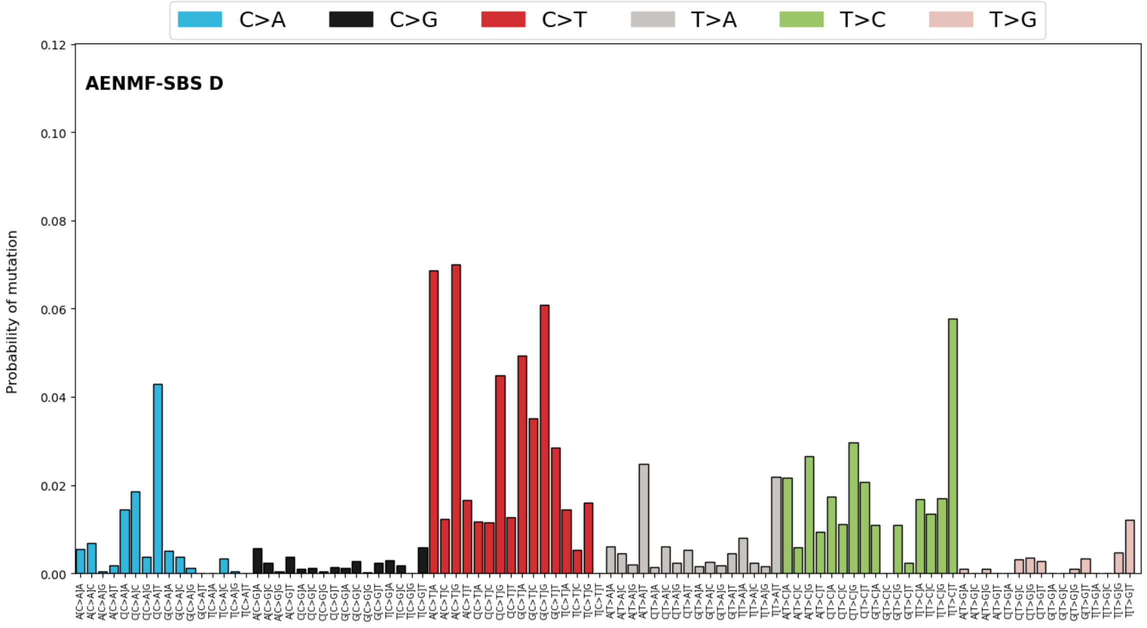
Consensus signatures comparison

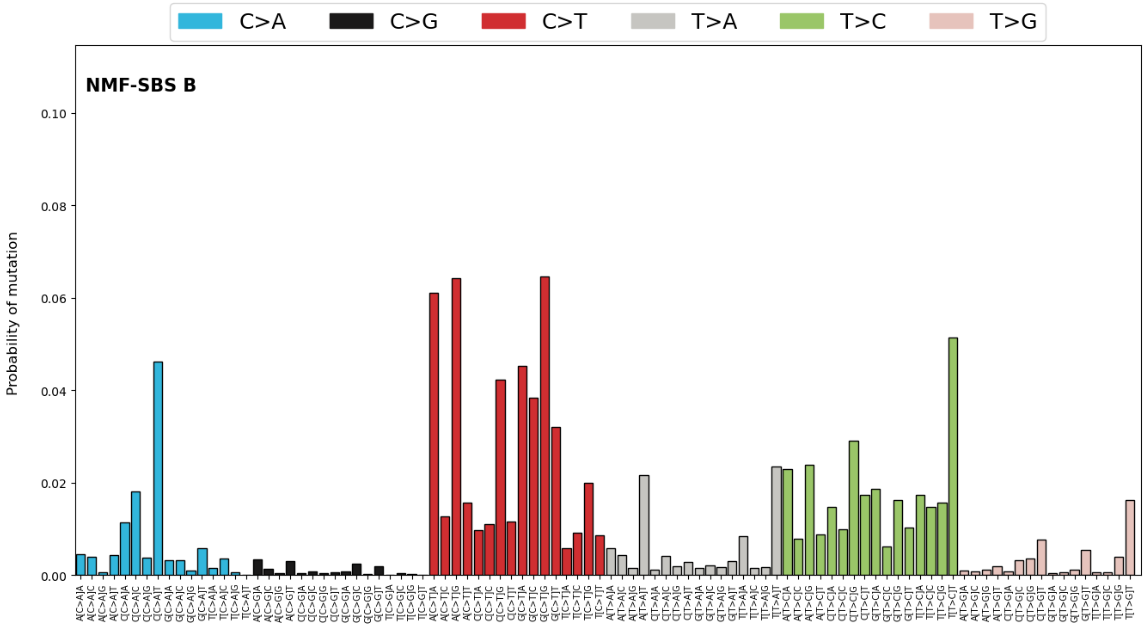


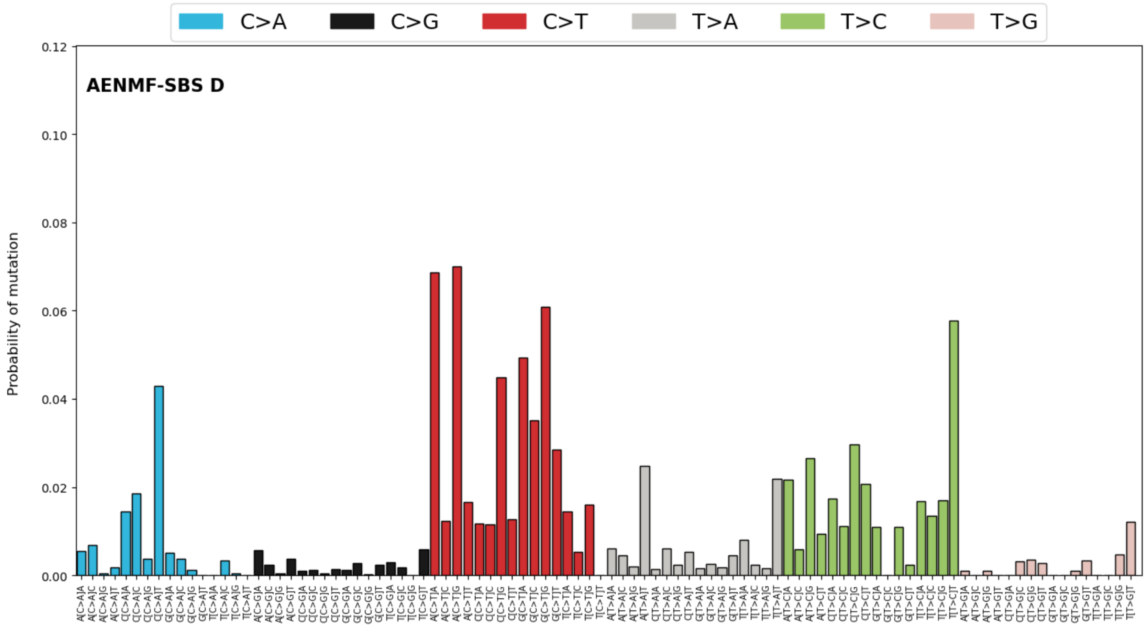
NMF Extracted	True	Cosine Similarity
SBS-A	SBS40a	0.74
SBS-B	**SBS44**	**0.85**
SBS-C	SBS10c	0.69
SBS-D	**SBS10a**	**0.93**

AENMF Extracted	True	Cosine Similarity
SBS-A	SBS5	0.60
SBS-B	SBS52	0.78
SBS-C	**SBS10a**	**0.92**
SBS-D	**SBS44**	**0.83**

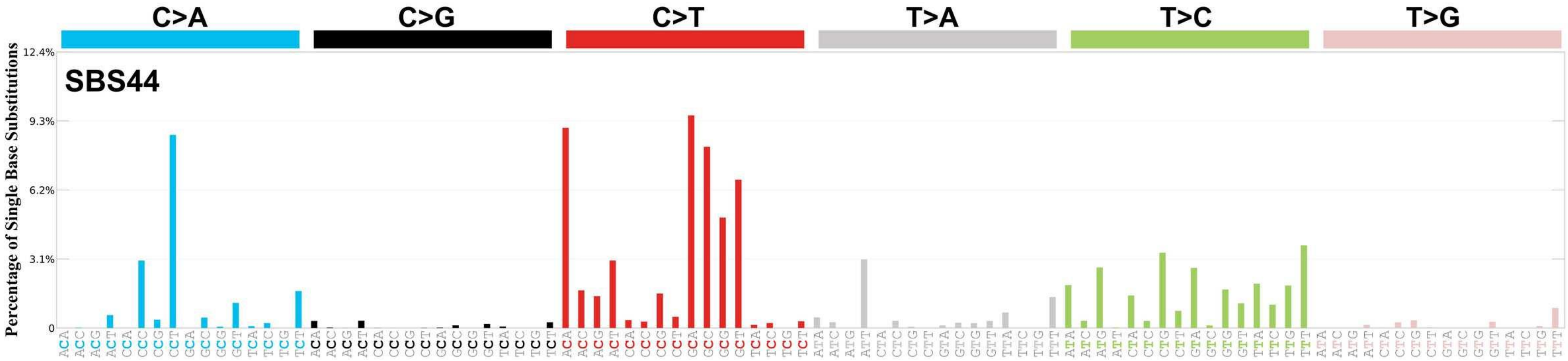








Comparison with COSMIC signature



Non-linear autoencoder

Why?

- A **non-linear autoencoder** can effectively model complex genomic interactions that could be missed by simpler, linear methods.

- A **Poisson-based** loss function naturally accommodates the count-based nature of mutational data.

- **Sparsity constraints** may improve the interpretability of extracted signatures, reducing overlap and facilitating biological interpretation.

Input data adjustment

$m = 96$

$n = 523$

Data matrix

$\mathbf{V}$

$$V \in \mathbb{R}^{m \times n}$$

$$V \approx HW$$

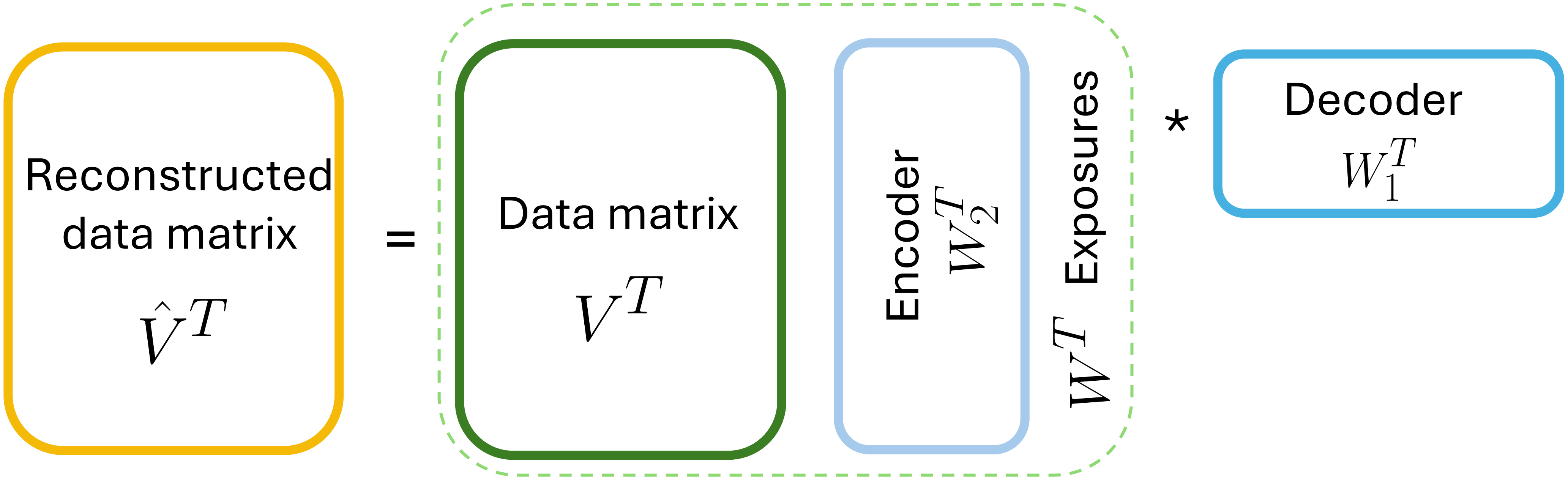
Transposed data matrix

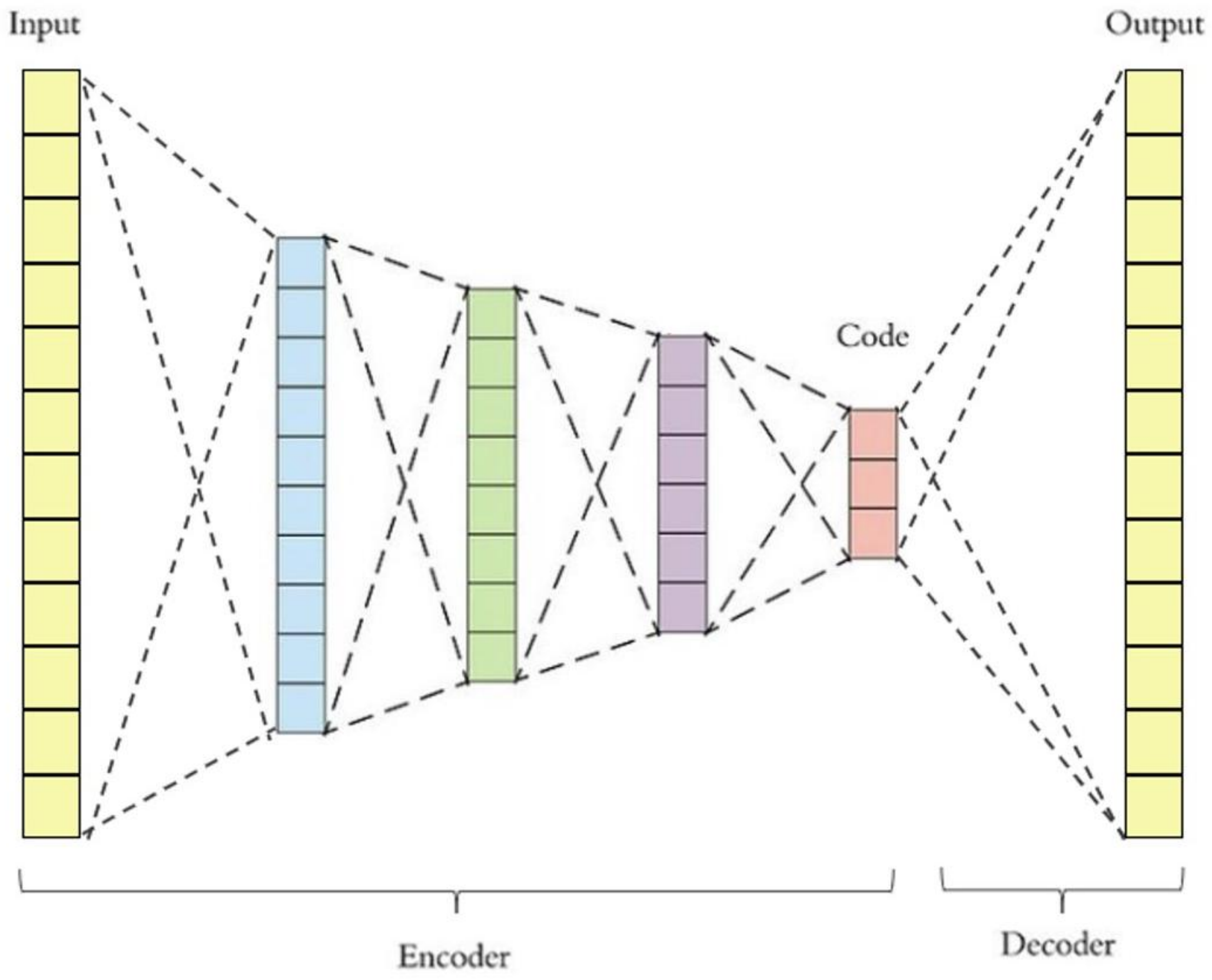
V^T

$$V^T \in \mathbb{R}^{n \times m}$$

$$V^T \approx W^T H^T$$

Signature and exposure inversion





Autoencoder Architecture

Encoder

Input dimension: 96

Three hidden layers: 128, 64, 32

Latent dimension: between 3 and 9

Decoder (shallow)

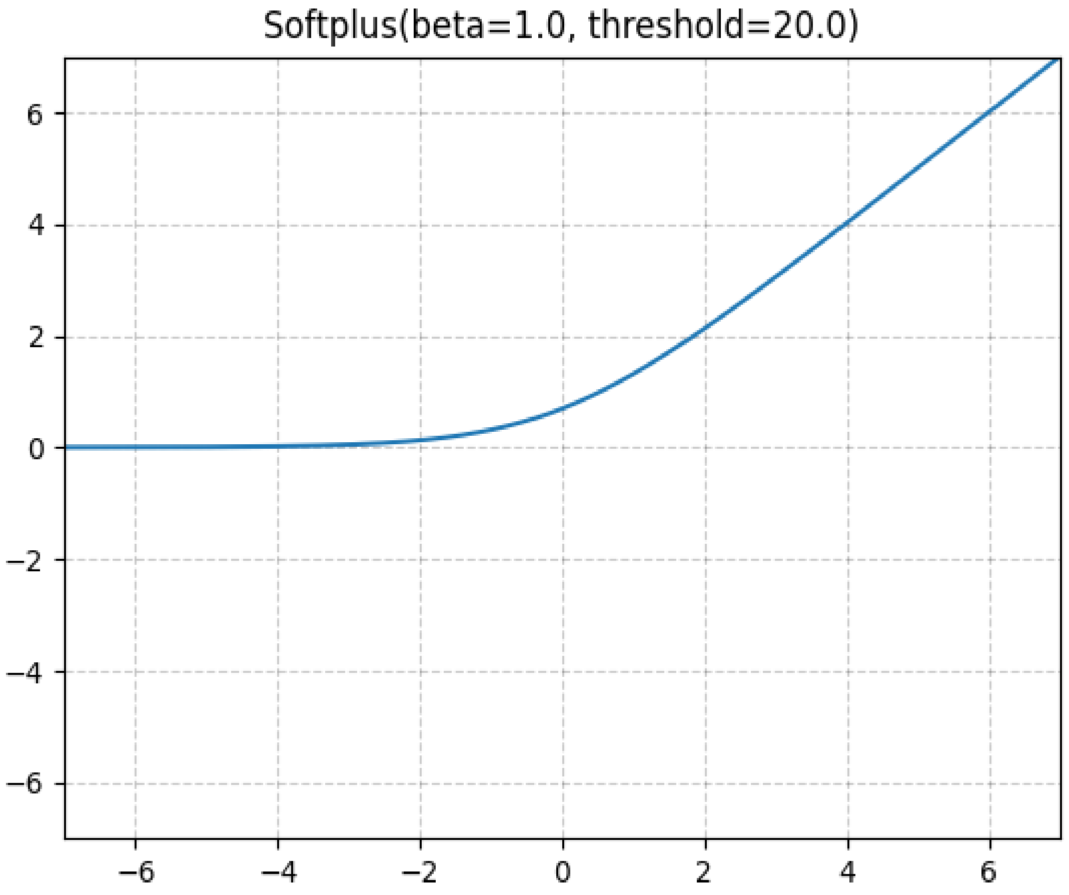
One linear layer

Output dimension: 96

Activation function

The **Softplus** activation function was used in all encoder layers *and* in the *latent layer*.

For the decoder an **identity** activation function was used



$$\frac{1}{\beta}\log\left(1+\exp(x\cdot\beta)\right)$$

Weight Initialization

The *xavier uniform* method was used to initialize all the layers weights

Clamping on the *decoder* weights to avoid negative values

$$w_{ij} \sim U\left(-\sqrt{\frac{6}{n_{\text{in}} + n_{\text{out}}}}, \sqrt{\frac{6}{n_{\text{in}} + n_{\text{out}}}}\right)$$

Loss optimization

Stochastic Gradient Descent with a batch size of 64 and **ADAM** optimizer: learning rate 0.001.

The β parameter is initially fixed at 0.001

$$L(x;\hat{x}) = \overbrace{-x\log(\hat{x})+\hat{x}}^{\text{Non-negative Poisson Likelihood}} \underbrace{+\beta\log\Big(\det(WW^{\top}+I)\Big)}_{\text{Minimum Volume Regularizer}} \quad \overbrace{\text{subject to} \quad W \geq 0}^{\text{Signature must be non negative}}$$

Multinomial Bootstrapping

Each i-th patient of the **augmented dataset** is drawn from the *i-th* Multinomial distribution $S_i \sim M_i(N_i, p_i)$ parametrized by:

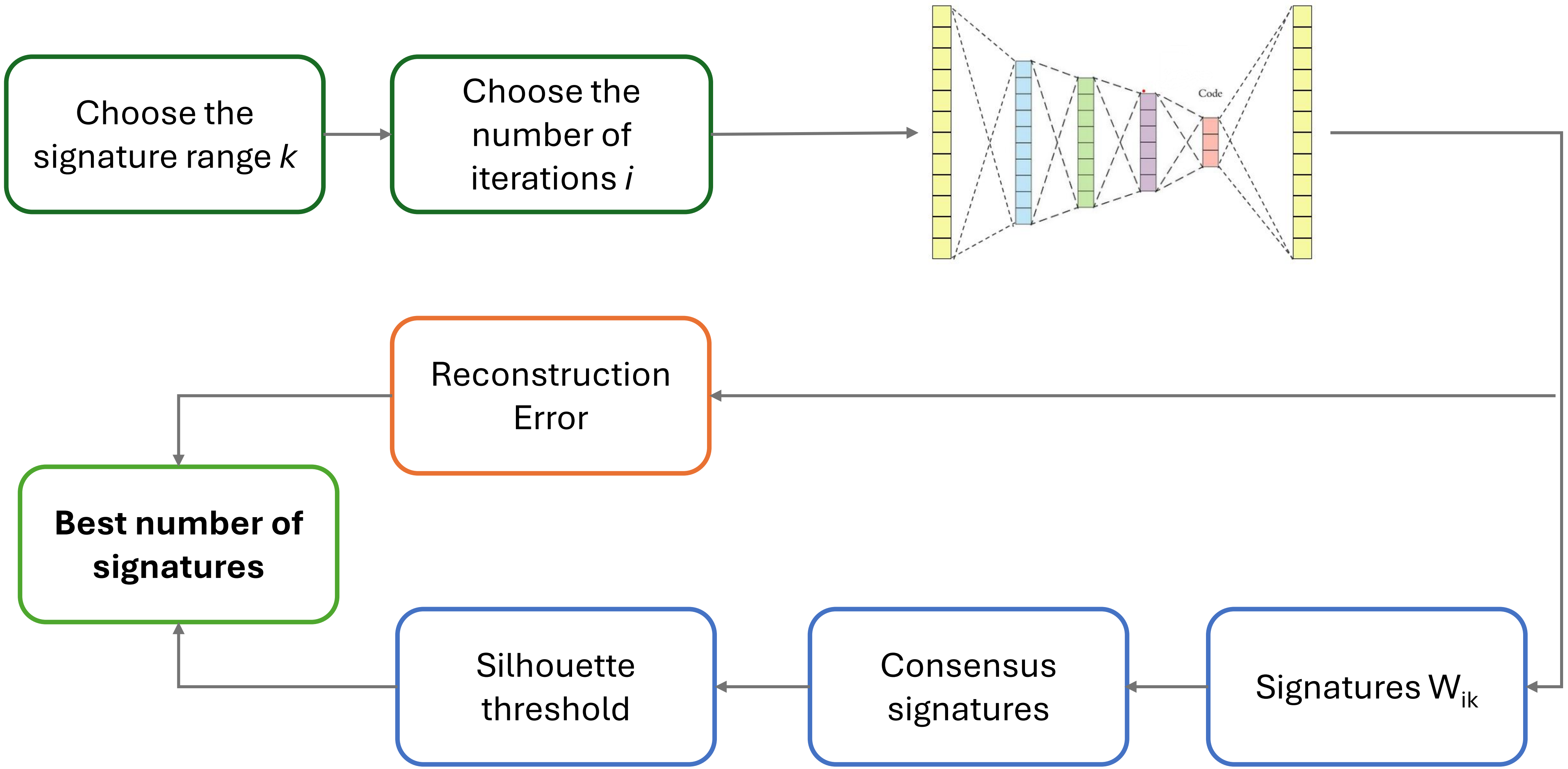
N_i Total **number** of mutations for patient *i* in the original dataset

p_i **Probablity** of each mutation for patient *i* in the original dataset

We generated **50 new datasets** and stacked them to create one larger augmented dataset to train the autoencoder with.

The original dataset is used as a validation set to check for overfitting via earlystopping

Choosing the correct number of signatures



Choose the signature range *k*

Choose the number of iterations *i*

Reconstruction Error

Best number of signatures

Signatures W_{ik}

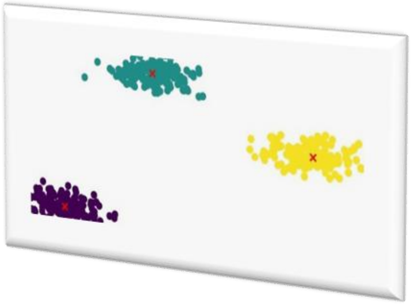
Consensus signatures

Silhouette threshold

Choosing the correct number of signatures

Consensus signatures

For each group of *n* mutational signatures run the *K-medoids* algorithm to find the consensus signatures



Silhouette threshold

Discard the consensus signatures which minimum and average silhouette score don’t satisfy the threshold (0.2 - 0.5)

$$th_{min} \geq \rho_{min}$$
$$th_{avg} \geq \rho_{avg}$$

Reconstruction Error

Computed using the *Frobenius* norm on the difference between the original data and the reconstructed one. Used for early stopping via a patience counter

$$\left\| X - \hat{X} \right\|_F$$

Bayesian Optimization

We want to efficiently explore the hyperparameter space towards the global optimum.

Goal: find the set of input parameters x that **maximize** a function $f(x)$

$$\underset{x \in A}{\operatorname{argmax}} f(x)$$

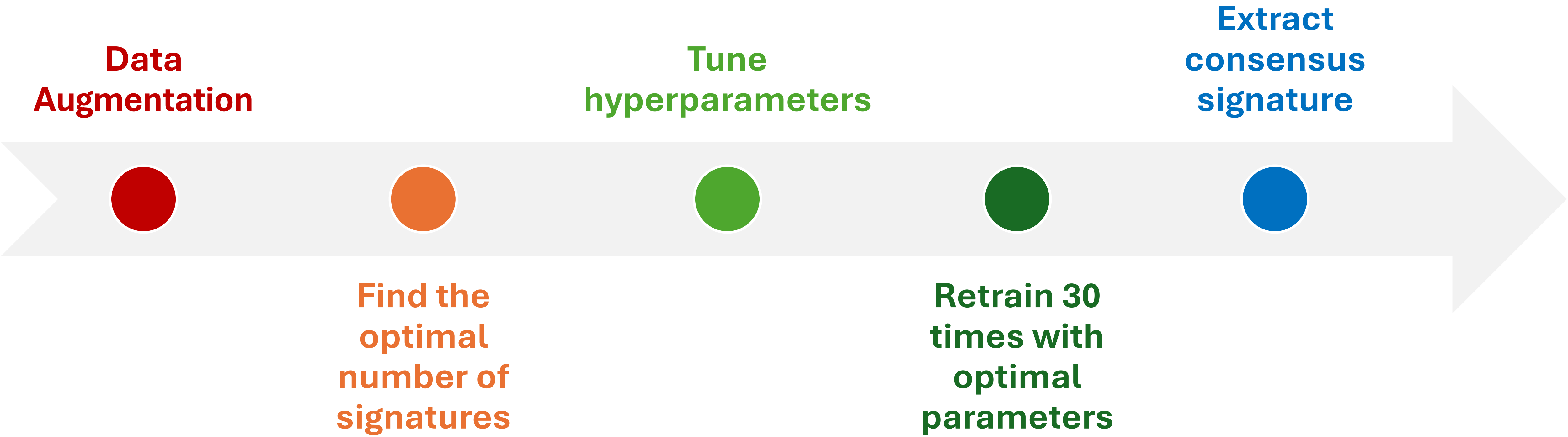
Components:

- **Surrogate model**: a statistical model to **approximate** the objective function $f(x)$ → *Gaussian Processes*
- **Acquisition function**: function that **guides** where to sample next → *Expected Improvement* (expected value of how much better the function value $f(x)$ at a given point x is compared to the best known function value achieved at step n)

Hyperparameter Tuning

Hyperparameters	Range
ADAM learning rate lr	[1e-4, 1e-2]
Regularization strength β	[1e-4, 1e-2]

Pipeline



Data Augmentation

Find the optimal number of signatures

Tune hyperparameters

Retrain 30 times with optimal parameters

Extract consensus signature

Results

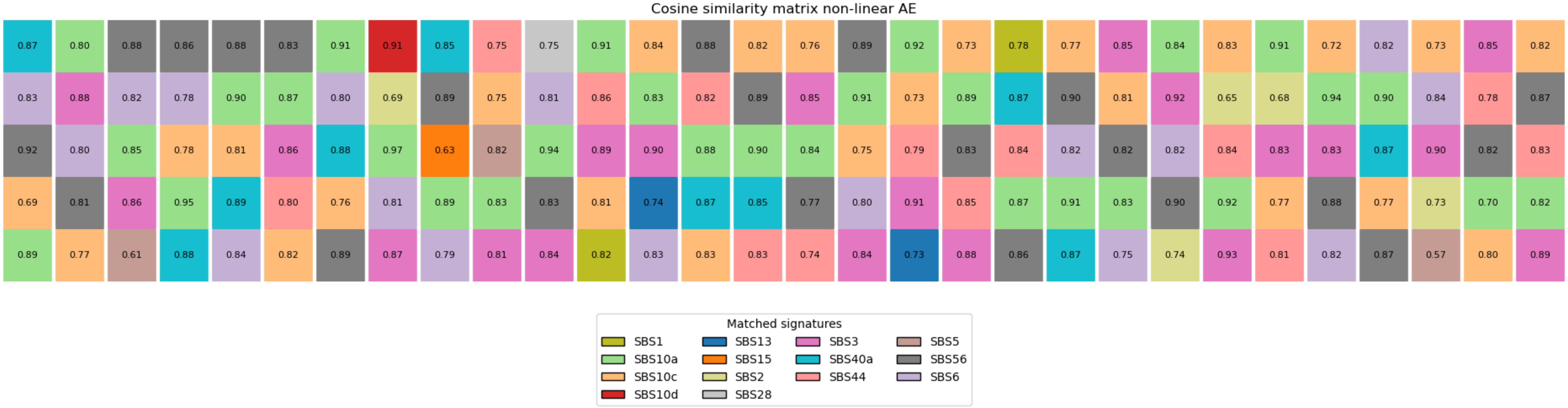
$k = 5$

lr = 0.008

B = 0.005

Extracted	True	Cosine Similarity
SBS-A	SBS10a	0.85
SBS-B	SBS40a	0.87
SBS-C	SBS10c	0.75
SBS-D	SBS56	0.89
SBS-E	SBS6	0.82

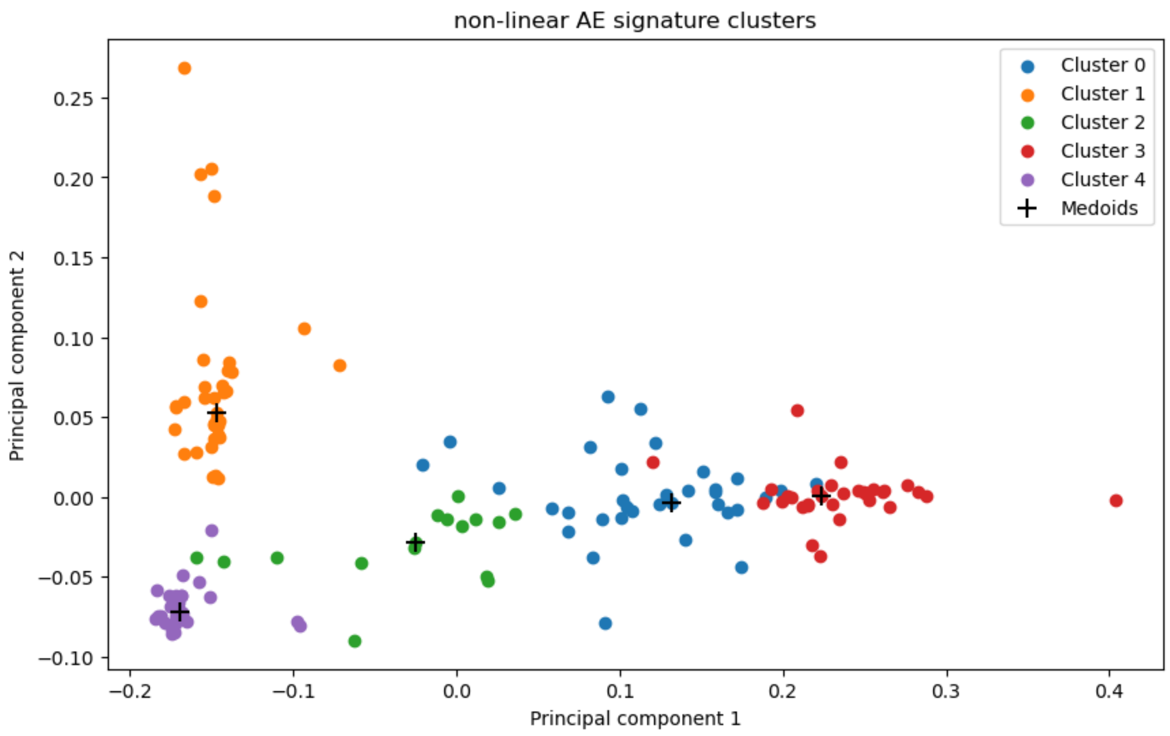
Cosine similarity comparison

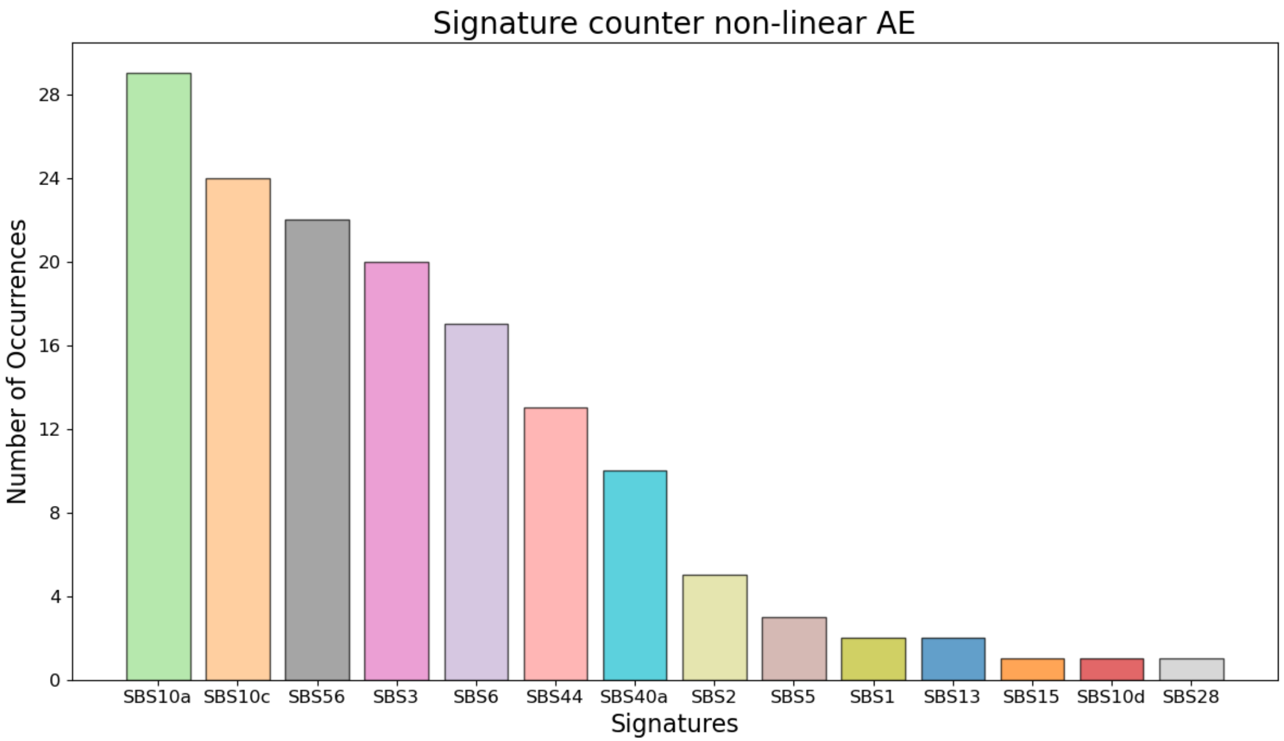


Cosine similarity matrix non-linear AE

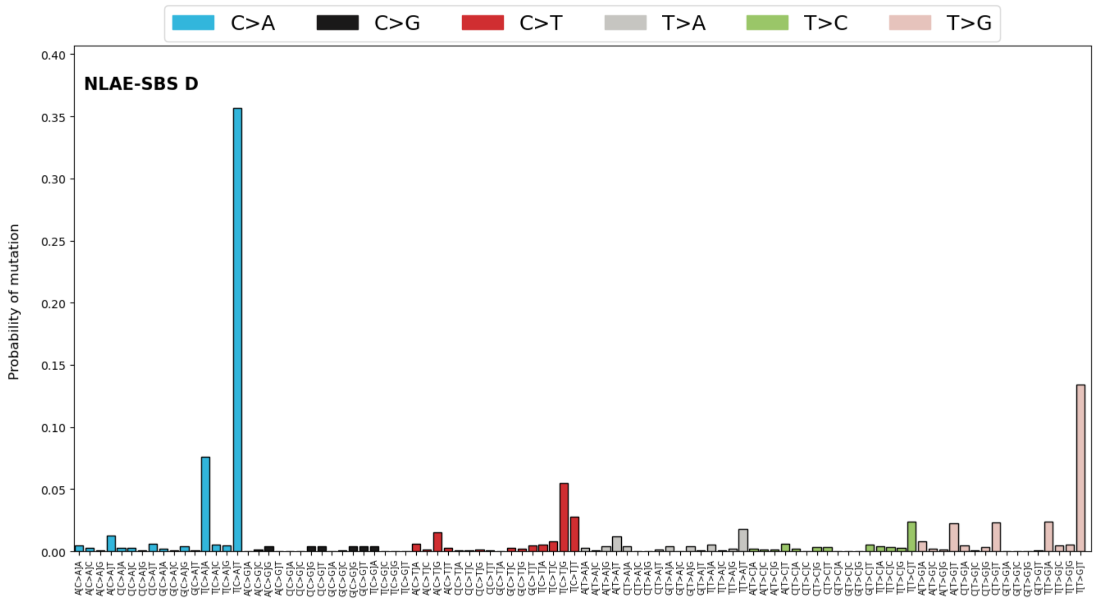
0.87	0.80	0.88	0.86	0.88	0.83	0.91	0.91	0.85	0.75	0.75	0.91	0.84	0.88	0.82	0.76	0.89	0.92	0.73	0.78	0.77	0.85	0.84	0.83	0.91	0.72	0.82	0.73	0.85	0.82
0.83	0.88	0.82	0.78	0.90	0.87	0.80	0.69	0.89	0.75	0.81	0.86	0.83	0.82	0.89	0.85	0.91	0.73	0.89	0.87	0.90	0.81	0.92	0.65	0.68	0.94	0.90	0.84	0.78	0.87
0.92	0.80	0.85	0.78	0.81	0.86	0.88	0.97	0.63	0.82	0.94	0.89	0.90	0.88	0.90	0.84	0.75	0.79	0.83	0.84	0.82	0.82	0.82	0.84	0.83	0.83	0.87	0.90	0.82	0.83
0.69	0.81	0.86	0.95	0.89	0.80	0.76	0.81	0.89	0.83	0.83	0.81	0.74	0.87	0.85	0.77	0.80	0.91	0.85	0.87	0.91	0.83	0.90	0.92	0.77	0.88	0.77	0.73	0.70	0.82
0.89	0.77	0.61	0.88	0.84	0.82	0.89	0.87	0.79	0.81	0.84	0.82	0.83	0.83	0.83	0.74	0.84	0.73	0.88	0.86	0.87	0.75	0.74	0.93	0.81	0.82	0.87	0.57	0.80	0.89

Signatures comparison



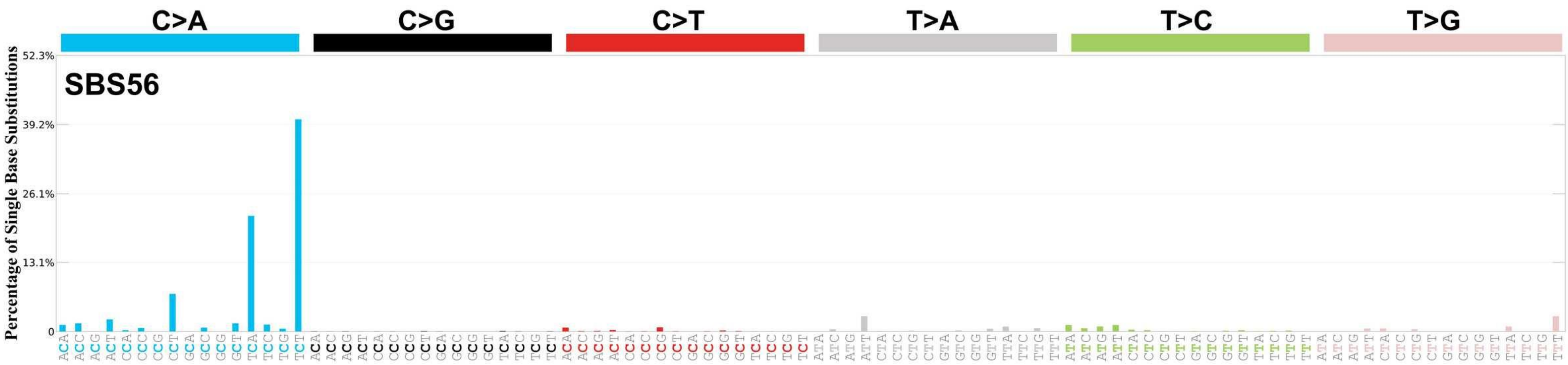


Comparison with COSMIC signature



Final reconstruction error

4.3 * 10^4



Aetiology

SBS3, while not in the consensus, was extracted frequently and is directly linked to the presence of ovarian cancer.

SBS44 is found in both NMF and AE-NMF, not by NLAE but it shares the same aetiology as **SBS6.**

Signature	Times found	Aetiology
SBS10a	29/30	Polymerase epsilon (POLE) exonuclease domain mutations
SBS10c	24/30	POLE exonuclease domain mutations
SBS56	22/30	Possible sequencing artifacts
SBS3	20/30	Defective homologous recombination-based DNA (BRCA1 and BRCA2)
SBS6	17/30	Defective DNA mismatch repair
SBS40a	10/30	Unknown

Conclusions

Overall, NLAE demonstrated consistency in its findings, repeatedly extracting the same signatures across multiple runs. It shares many signatures with NMF and may have captured an acquisition error.

The reconstruction error increased slightly, but it remains within the same order of magnitude as other methods.

The experiment was generally successful, but further validation on diverse datasets is needed to strengthen our conclusions. Additionally, exploring other techniques to mitigate overfitting could provide valuable insights.

Finally, investigating overdispersion in the reconstruction error and eventually adjusting the Poisson term in the loss to be a Negative Binomial one could be of interest

THANK YOU !

References:

- On the Relation Between Autoencoders and Non-negative Matrix Factorization, and Their Application for Mutational Signature Extraction ; Egendal et al. (2024)

- MUSE-XAE: MUtational Signature Extraction with eXplainable AutoEncoder enhances tumour types classification; Pancotti et al. (2024)

- Decoding whole-genome mutational signatures in 37 human pan-cancers by denoising sparse autoencoder neural network; Pei et al. (2020)