Supplementary Material: FATE - Feature-Agnostic Transformer-based Encoder for learning generalized embedding spaces in flow cytometry data

Lisa Weijler
TU Wien
lweijler@cvl.tuwien.ac.at

Florian Kowarsch
TU Wien
florian.kowarsch@gmail.com

Michael Reiter
TU Wien
rei@cvl.tuwien.ac.at

Pedro Hermosilla
TU Wien
phermosilla@cvl.tuwien.ac.at

Margarita Maurer-Granofszky
St. Anna CCRI
margarita.maurer@ccri.at

Michael Dworzak
St. Anna CCRI
michael.dworzak@stanna.at

1. Datasets

Tab. 1 gives a detailed overview of the number of patients and their samples collected per dataset, as well as city and years of acquisition. All samples from the datasets used for pre-training (CONTROL, DIA, ALL-MRD) are from different patients than those in the target dataset, AML-MRD, for a fair assessment of generalization between patients.

Fig. 1 shows four different FCM samples of different patients coming from the AML-MRD dataset. The plots are 2D projections of the sample on different features including forward- and side-scatter (see Sec. 2). Each point corresponds to the feature measurement vector of one cell. We can see the variation in cancer cells (red) regarding positioning and proportion as well as the variation in density of healthy cell populations (blue).

2. Features

All four datasets combined have 35 features in total including forward- (FSC) and side-scatter (SSC) of the laser light. FSC and SSC measure physical properties of the cell, like size and granularity. The rest of the features correspond to fluorescent-labeled surface markers, of which the concentration is measured. Different cell types bind to different surface markers and hence, those markers enable to separate and analyse cell-types. A detailed listing of what features in how many samples are present for each dataset is given in Tab. 2. ”CD” stands for cluster of differentiation. ”A”, ”H” an ”W” stands for area, hight and width, respectivley in the FSC and SSC. We just speak of FSC and SSC and do not necessarily differentiate between area, hight and width, since those are highly correlated. TIME indicates when the feature vector was acquired throughout the acquisition process of one sample and hence is not a strong discriminative feature. In total we can see that the AML-MRD and ALL-MRD datasets have the least in common with only 2 discriminative features excluding FSC and SSC throughout all samples, CD45 and CD34.

3. Patient Cross-Validation

The number of samples used for training, validation and testing are given in Tab. 3. Each patient determines one split. After precision p, recall r and F_1-score have been determined for each sample in each test split, metrics are averaged over all samples to get the final results that are reported.

4. FATE-Decoder

Our Decoder proposed for the masked autoencoder (MAE) experiments can be divided into a set-decoder and a feature-decoder part as described in the main text. Fig. 2 shows a detailed illustration of the two parts of the decoder.

5. Induced Self-Attention Block

We use the ISAB layers as proposed in [1]. Fig. 3 shows an illustration of the detailed architecture, where n denotes the number of input set elements and m the number of learnt queries i.e. induced points.

6. General Embedding of the FATE-MAE

Fig. 4 and Fig. 5 visualize the embedding of two different samples generated with the pretrained FATE-MAE with the CONTROL, DIA and ALL-MRD dataset with a masking ratio of 0.25%. We visualize healthy (blue) versus cancerous (red) cells (row 1 and 4) as well as clustering of cell populations (row 2 and 3). The figures show that the learnt embedding forms meaningful clusters of cells with respect to clusters of the original feature space.

Table 1. Description of the FCM datasets with respect to the city of the laboratory and the time-span the samples where acquired, the number of samples per dataset as well as the number of patients the samples were taken from.

Dataset	**City**	**Years**	**Samples**	**Patients**
AML-MRD	Vienna	2021-2022	71	12
CONTROL	Essen	2016	52	17
	Padua	2016-2021	88	22
	Vienna	2016-2021	168	57
DIA	Essen	2016	24	12
	Padua	2016	6	3
	Vienna	2016-2022	80	30
ALL-MRD	Vienna	2009-2014	200	200
	Berlin	2015	72	72
	Buenos Aires	2016-2017	66	66

References

[1] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019. 1, 7

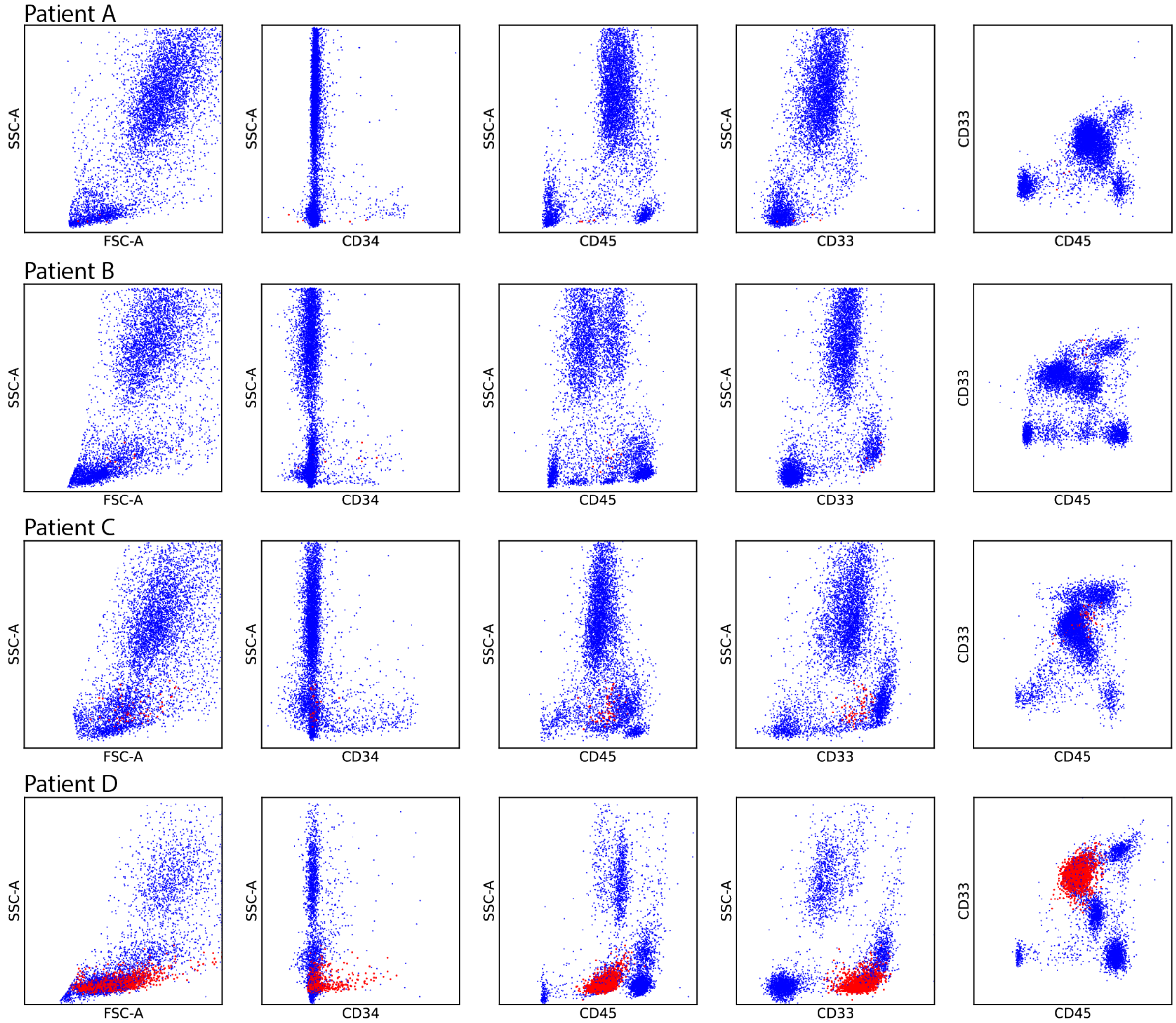

Figure 1. This figure shows 2D projections of samples from the AML-MRD dataset on different combination of features. Each dot corresponds to one event i.e. the feature measurement vector of one cell. Cancer cells are marked in red, healthy in blue. Each row corresponds to a different patient.

Table 2. List of features (markers) present in the datasets and their number of occurrences per dataset.

Feature	**AML-MRD**	**CONTROL**	**DIA**	**ALL-MRD**
FSC-A	71	308	110	338
FSC-H	66	131	58	0
FSC-W	71	253	99	338
SSC-A	71	308	110	338
SSC-W	67	185	52	1
CD38	32	164	52	338
CD371	32	163	52	0
CD34	71	308	110	338
CD117	71	308	110	0
CD33	71	308	110	0
CD71	5	6	28	0
CD123	33	165	53	0
CD45RA	32	164	52	0
HLA-DR	71	308	26	0
CD45	71	308	110	338
TIME	71	308	110	338
CD99	34	180	63	0
CD10	2	0	12	338
CD133	3	0	8	0
CD15	39	161	57	0
CD11A	18	11	13	0
CD7CD56	5	1	3	0
CD14	39	161	57	0
CD11B	39	161	57	0
CD13	14	132	35	0
NG2	3	2	4	0
CD56	12	0	0	0
CD7	15	15	11	0
CD312	5	0	1	0
CD16	2	0	14	0
CD48	1	0	2	0
CD58	0	0	0	138
CD19	0	3	3	338
CD20	0	0	0	338
SY41	0	0	0	338

Table 3. The sample of the AML-MRD dataset are divided into training-, evaluation-, and test-set. Each patient's samples exclusively form the test set, as indicated by the rightmost column specifying the sample count for each patient. The rest of the samples from different patients are allocated to the training and evaluation sets with a proportion of approximately 0.8 and 0.2, respectively.

Patient	**Train**	**Eval**	**Test**
A	56	14	1
B	51	15	5
C	47	14	10
D	54	13	4
E	53	13	5
F	49	11	11
G	42	11	18
H	56	13	2
I	54	15	2
J	51	11	9
K	56	13	2
L	54	15	2

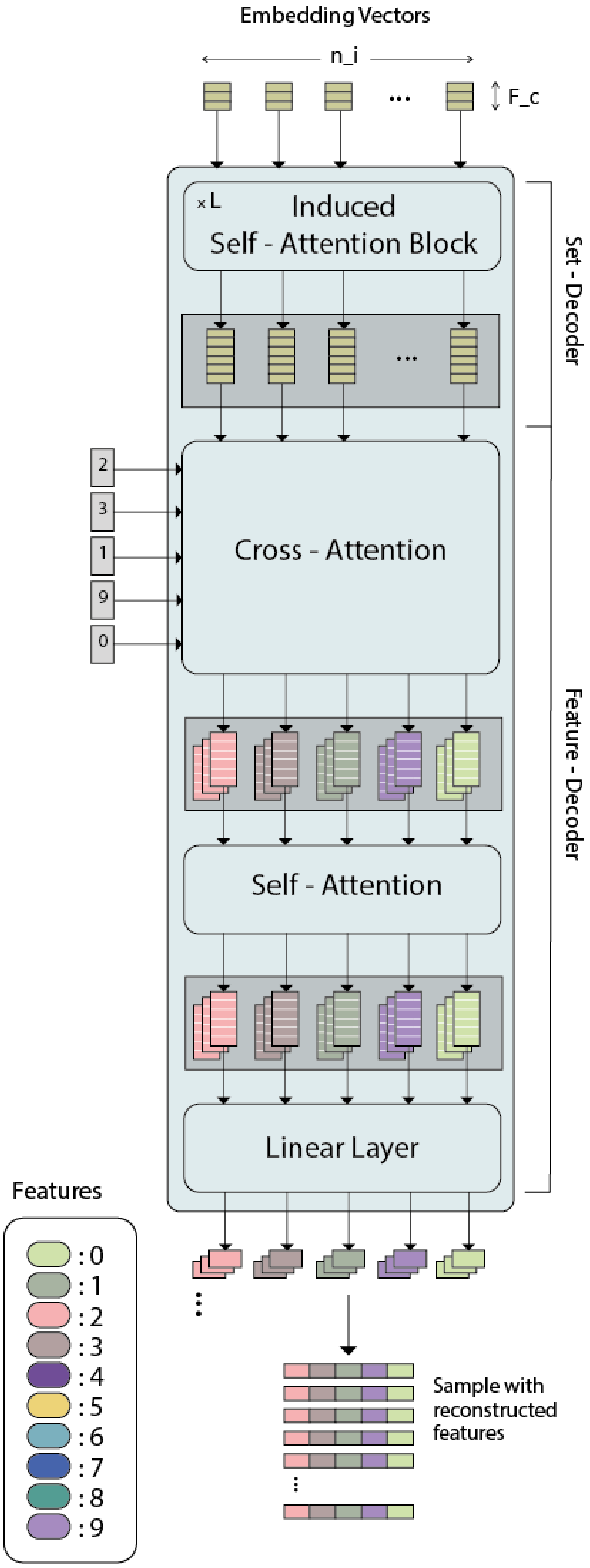

Figure 2. This figure illustrates the architecture of the decoder proposed for the MAE experiments. Input are the embedding vectors, which are put into context to each other with ISAB layers. It follows cross attention with the feature encodings of the features that should be reconstructed. Finally, the learnt vectors each corresponding to one feature value are attended to each other and mapped to its final reconstructed scalar feature value.

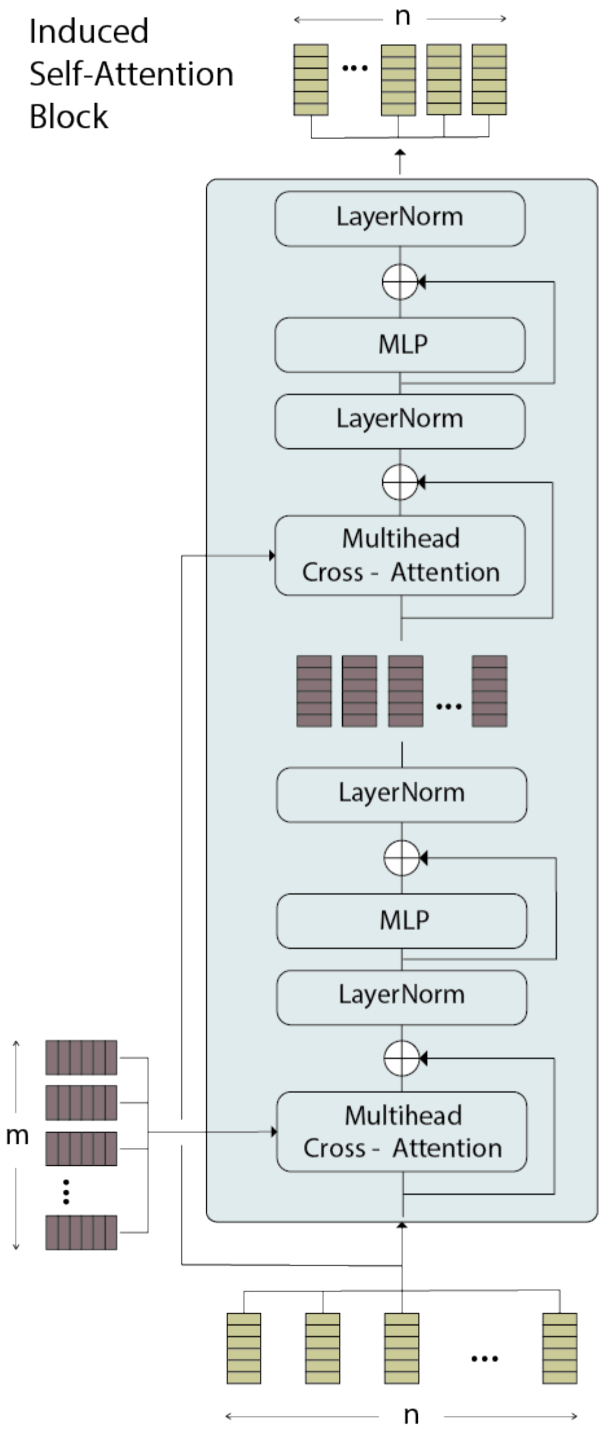

Figure 3. Illustration of the ISAB layer with an input set of n elements and m induced points [1]. Plus sign indicates additive skip connections.

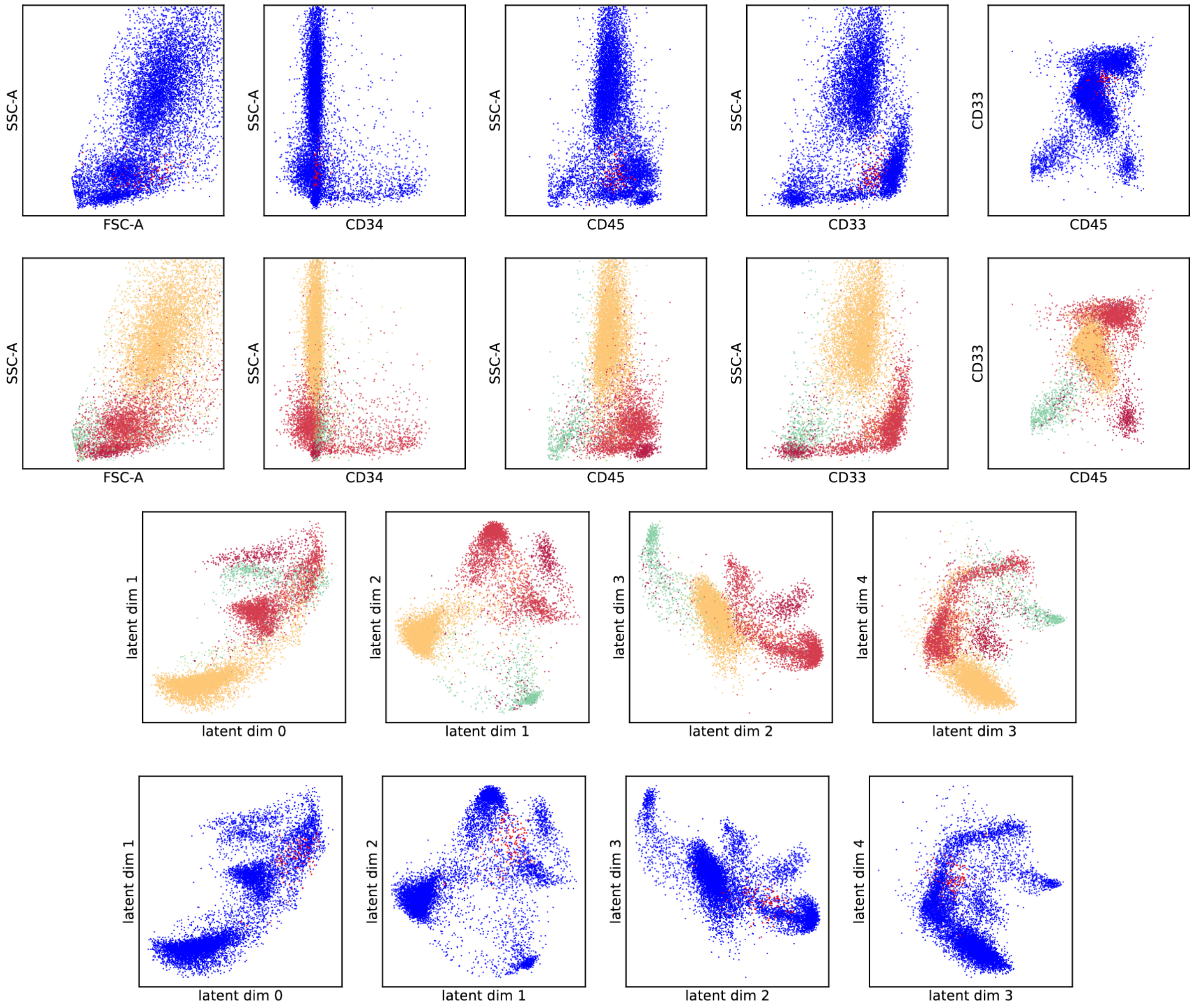

Figure 4. Visualization of a sample of Patient C from the AML-MRD dataset in its original feature space (row 1-2) and in the learnt general embedding space (row 3-4). Blue denotes healthy cells, red cancerous cells. The colours in row 2 and 3 indicate clusters of cell populations.

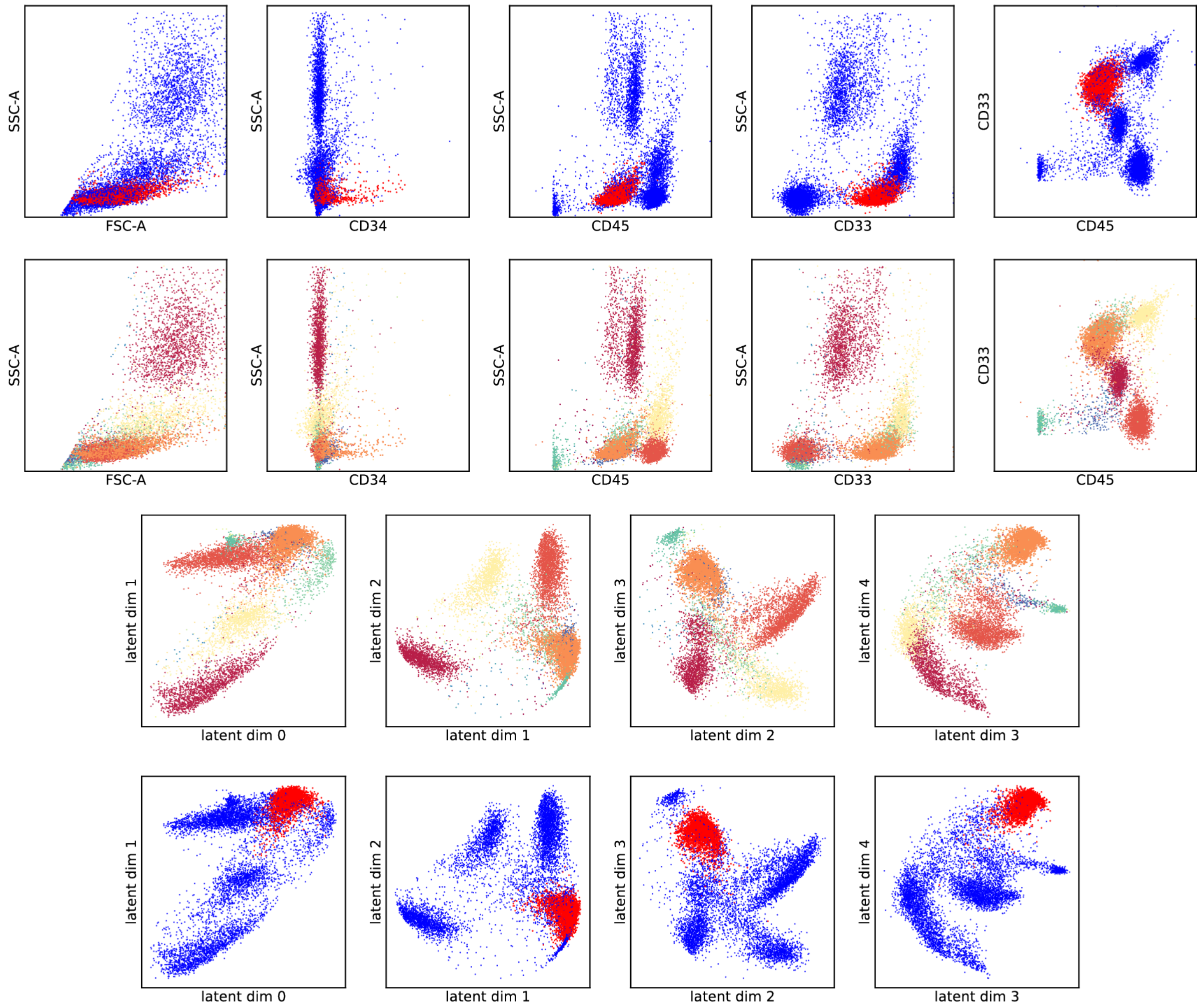

Figure 5. Visualization of a sample of Patient D from the AML-MRD dataset in its original feature space (row 1 -2) and in the learnt general embedding space (row 3-4). Blue denotes healthy cells, red cancerous cells. The colours in row 2 and 3 indicate clusters of cell populations.