



Finanziato  
dall'Unione europea  
NextGenerationEU



Ministero  
dell'Università  
e della Ricerca



Italiadomani

PIANO NAZIONALE  
DI RIPRESA E RESILIENZA



# Exploiting the latent space of deep AutoEncoders for the identification of signal pulses in noisy time-series

**Gioacchino Alex Anastasi<sup>1,2</sup>**, **Paules Zakhary<sup>1,2</sup>**, **Noemi Pino<sup>3</sup>**,  
**Sebastiana Maria Puglia<sup>1,2,4</sup>**, **Marzio De Napoli<sup>1,2</sup>**, **Alessia Tricomi<sup>1,2,4</sup>**,  
**Sebastiano Albergo<sup>1,2,4</sup>**

*1. University of Catania, Department of Physics and Astronomy*

*2. Istituto Nazionale di Fisica Nucleare (INFN), Catania section*

*3. Laboratori Nazionali del Sud*

*4. Il Centro Siciliano di Fisica Nucleare e Struttura della Materia (CSFNSM), Catania*

XIV International Conference on  
New Frontiers in Physics (ICNFP 2025)  
- Special session on Machine Learning -



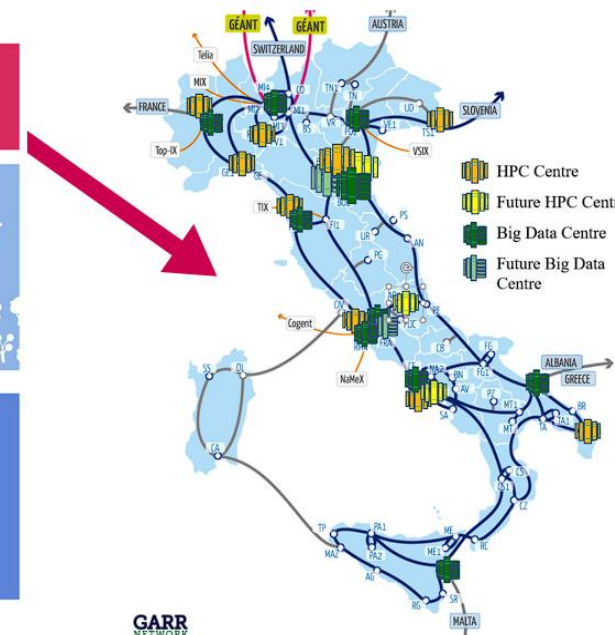
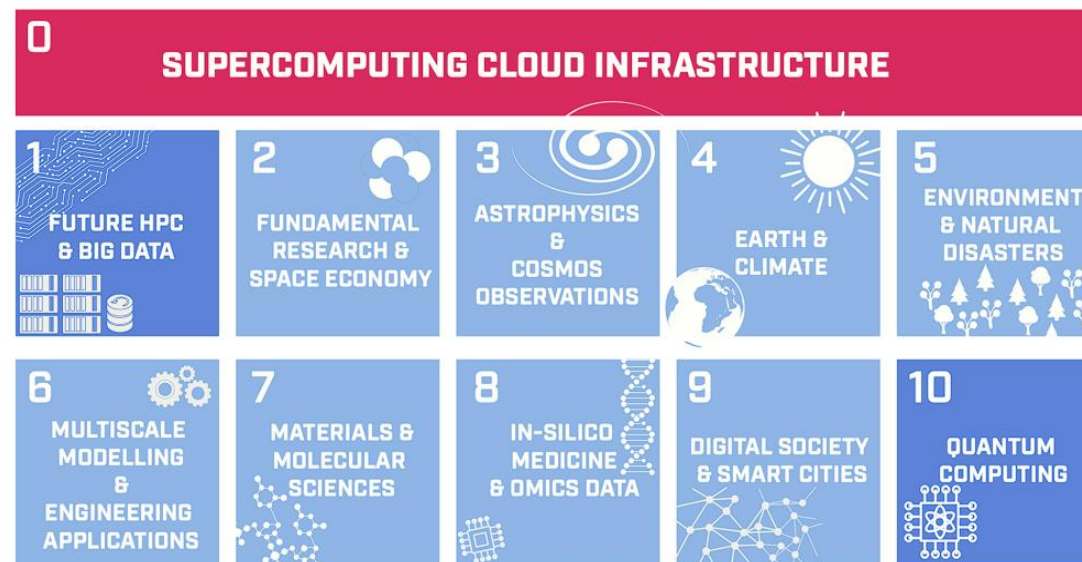
# The National Centre for HPC, Big Data and Quantum Computing

Managed by the [ICSC Foundation](#), one of the five National Centres founded under the Italian PNRR.

**Goal** : a long-term, national, distributed infrastructure for cutting-edge research and innovation in high-performance and high-throughput computing.

**A total investment of almost 320 million Euros**  
Sept. 2022 – Aug. 2025

Over 50 founding members, to foster synergy between scientific and industrial sectors.



One cross spoke, Spoke 0 ("Supercomputing Cloud Infrastructure"), and 10 thematic spokes.



## The use-case DAIDREAM within Spoke 2 – WP3

***DA**ta-driven **ID**entification of **R**are **E**vents in **A**stroparticle physics through **M**achine learning techniques*

Employ **self-supervised or weakly supervised deep-learning** to fully exploit experimental data in at least 2 distinct (yet contiguous) experimental settings:

- searches for WIMPs with mass up to ~10 TeV in dual-phase Liquid Argon TPCs;
- search for rare or anomalous air-shower footprints in ground-based observatories.



Project developed on the [INFN-Cloud infrastructure](#).

The **Spoke-2** (*Fundamental Research & Space Economy*) is coordinated by **INFN** and organized in 6 Work Packages.

WP1: algorithms and tools for theor. physics

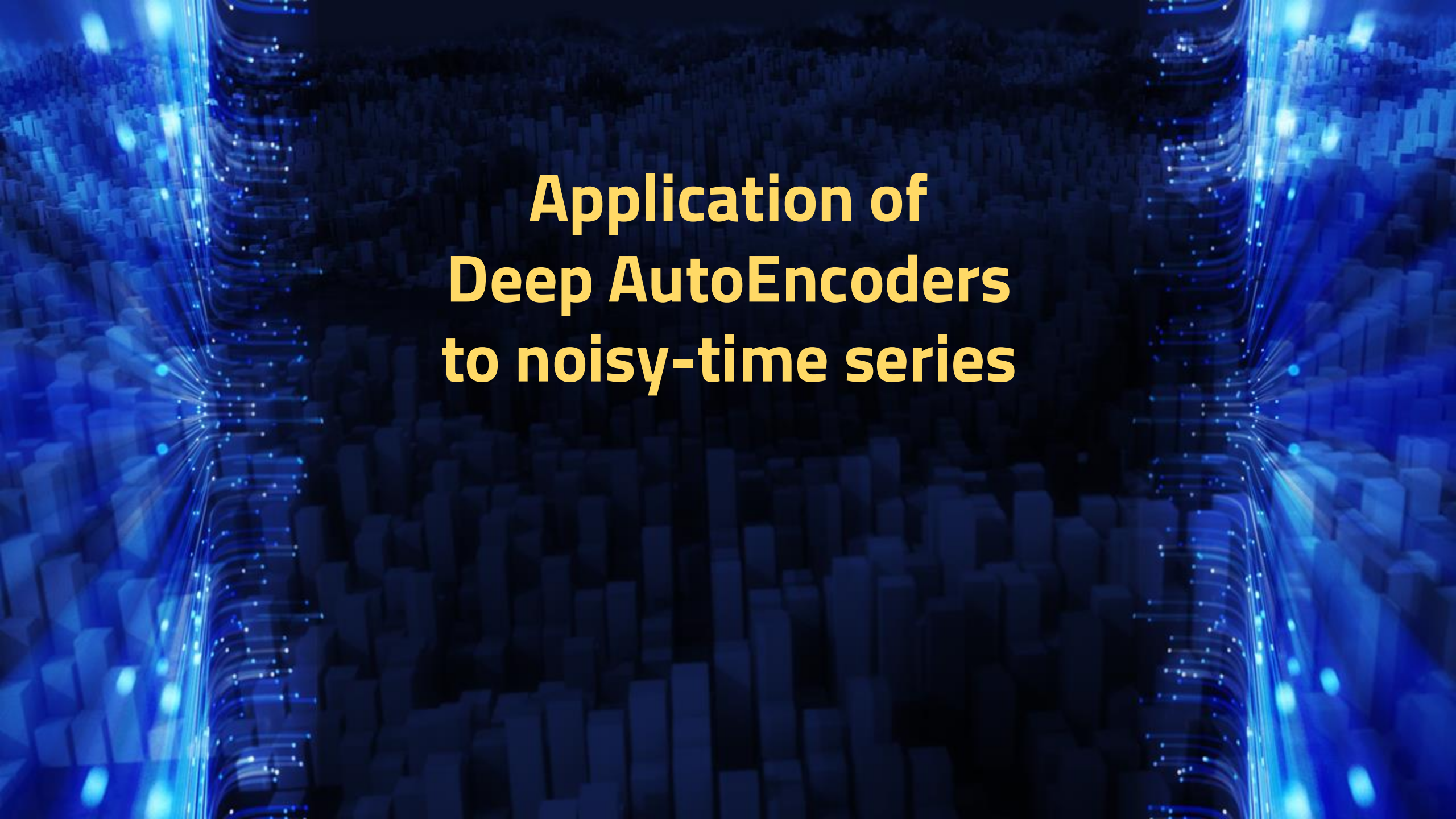
WP2: applications for exp. high-energy physics

WP3: applications for exp. astroparticle and G.W.

WP4: boost computational performances & porting to GPU, FPGA, etc.

WP5: support for data management & distributed data-lake infrastructure

WP6: Cross Domain initiatives & Space Economy



**Application of  
Deep AutoEncoders  
to noisy-time series**



## Reference work

**General goal:** extract features from raw waveforms studying the representation in an optimized latent space, possibly avoiding the need for a dedicated pulse finder.

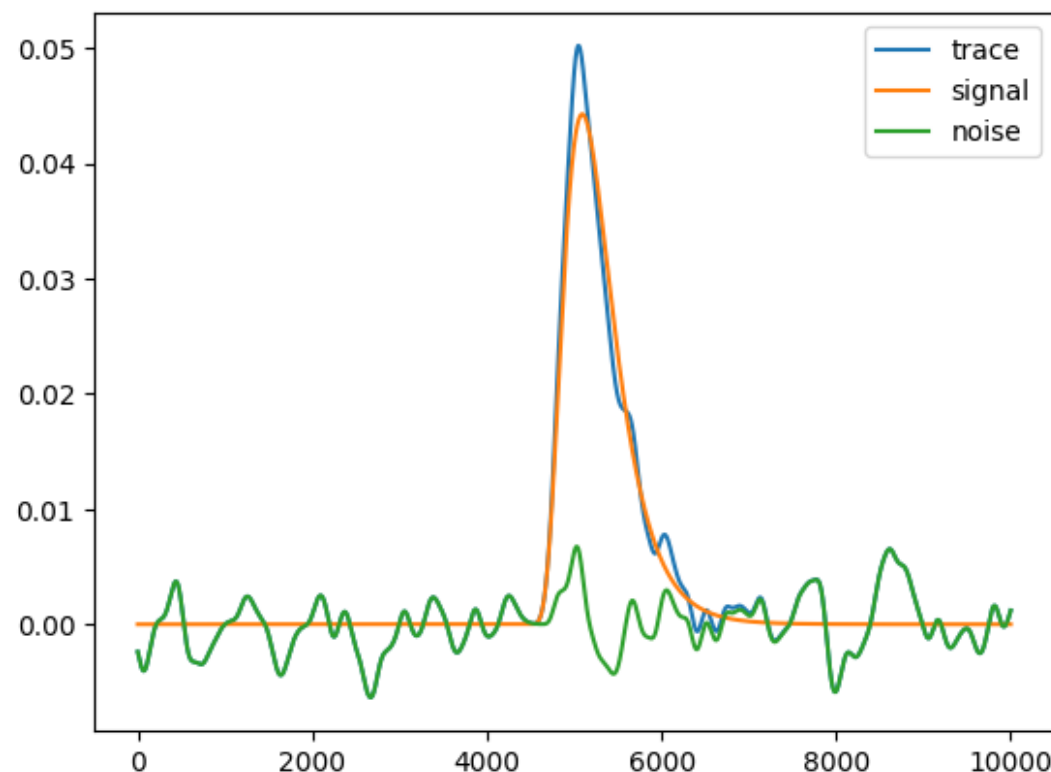
- Presentation of the method: [G.A. Anastasi, CRIS-MAC 2024](#)
- Application to the data from the *Recoil Directionality (ReD)* apparatus, in the context of the Darkside-20k experiment: [N. Pino, IDM 2024](#)

**In this talk we present a study of the performances of this method on a dataset of synthetic waveforms with controlled noise level and signal pulses.**

## Synthetic dataset

**Synthetic waveform:** single-peaked log-normal shaped signal on top of non-gaussian noise.

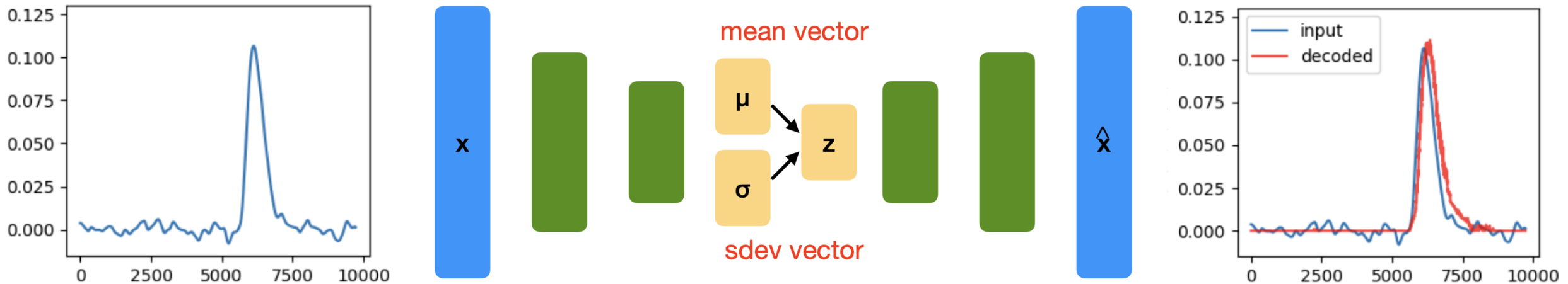
- Generated **signals** have amplitude (i.e. peak value) distributed uniformly in  $\log_{10}$  in  $[0.001, 1]$ , and fixed shape and location parameters.
- The noise is composed by the sum of many (log-normal) pulses randomly positioned, each with fixed amplitude, shape and location parameters.



□ Dataset of 15000 w.f. : (6000 + 1500) for training & validation, 7500 for testing.

## Variational Convolutional AutoEncoders

**Self-supervised neural network** architecture where data are compressed into a low dimensionality *latent space*, then reconstructed minimizing differences between original and output.



- The encoder outputs a vector of means  $\mu$  and a vector of standard deviations  $\sigma$ , **both of dimension  $k$** .
- The **sampling layer  $z$**  produces a vector of features by sampling  $k$  values, each from an independent Normal distribution  $N(\mu, \sigma)$ .
- The decoder uses the sampled  $z$  values to reconstruct the inputs.



## Architecture

3 Conv1D + avg. pooling layers, followed by 1 flattened Dense layer.

Sampling layer of dimension 2 to have a (working) latent space as simple as possible.

Optimizer: *ADAM* - initial Learning Rate 0.001

Reconstruction Loss: sum of square differences btw input and output.

Probabilistic Loss: standard *K-L divergence*.

Activation function: *ReLU*

Implemented in **Keras** with **Tensorflow** backend.

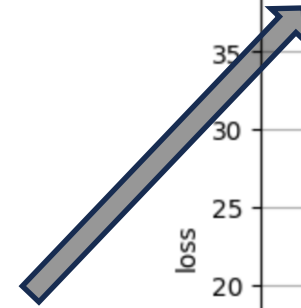
Layer (type)	Output Shape	Param #
encoder (Functional)	[(None, 2), (None, 2), (None, 2)]	6496
input (InputLayer)	[(None, 9728, 1)]	0
conv1 (Conv1D)	(None, 2432, 4)	132
average_pooling1d (AveragePooling1D)	(None, 1216, 4)	0
conv2 (Conv1D)	(None, 304, 8)	1032
average_pooling1d_1 (AveragePooling1D)	(None, 152, 8)	0
conv3 (Conv1D)	(None, 38, 16)	4112
average_pooling1d_2 (AveragePooling1D)	(None, 19, 16)	0
flatten (Flatten)	(None, 304)	0
z_mean (Dense)	(None, 2)	610
z_log_var (Dense)	(None, 2)	610
sampling (Sampling)	(None, 2)	0
decoder (Functional)	(None, 9728, 1)	6173
input_1 (InputLayer)	[(None, 2)]	0
dense_decoded (Dense)	(None, 304)	912
reshape (Reshape)	(None, 19, 16)	0
up_sampling1d (UpSampling1D)	(None, 38, 16)	0
deconv3 (Conv1DTranspose)	(None, 152, 8)	4104
up_sampling1d_1 (UpSampling1D)	(None, 304, 8)	0
deconv2 (Conv1DTranspose)	(None, 1216, 4)	1028
up_sampling1d_2 (UpSampling1D)	(None, 2432, 4)	0
deconv1 (Conv1DTranspose)	(None, 9728, 1)	129

Total params: 12675 (49.51 KB)

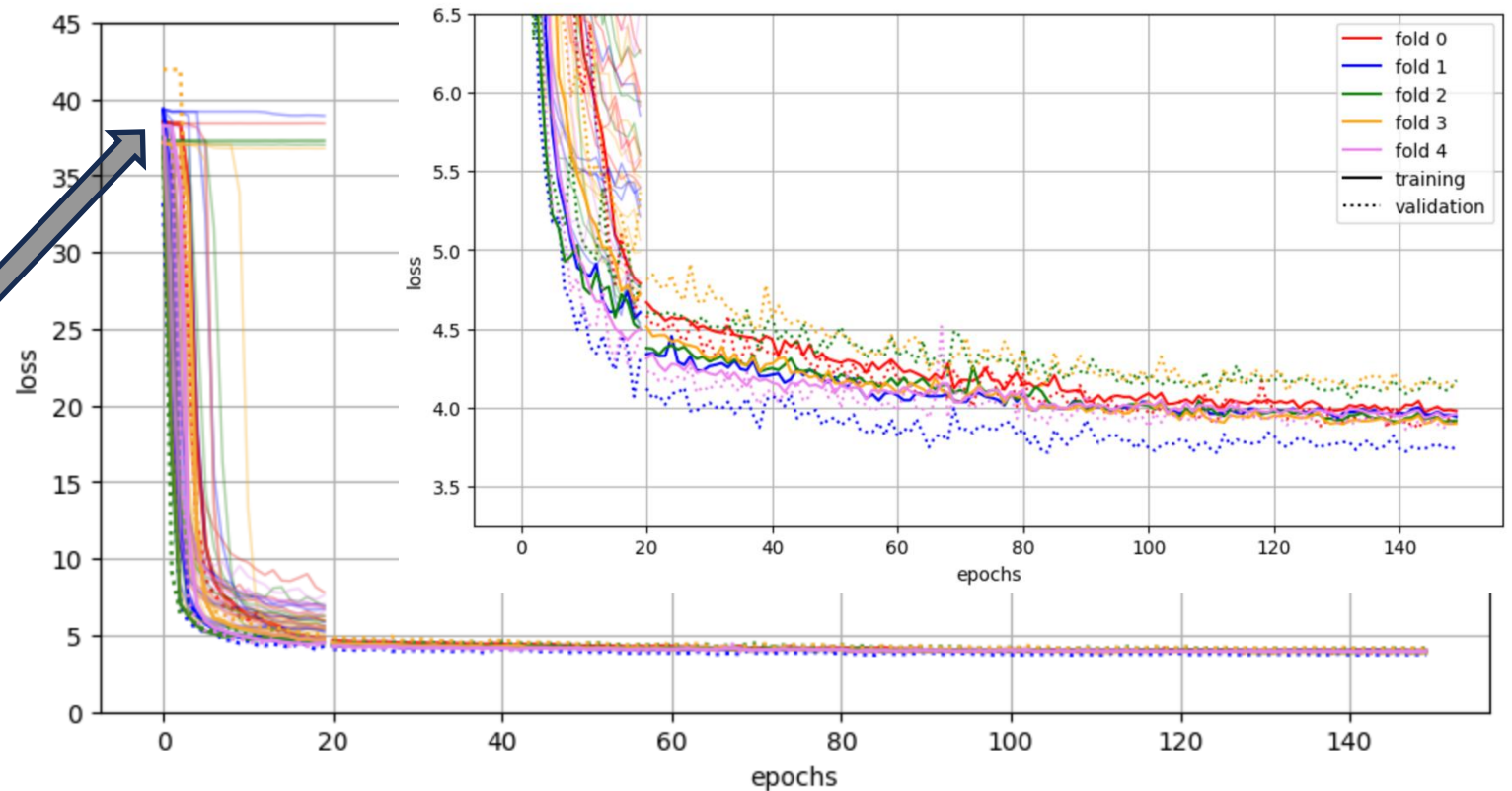
## Training

**5-fold cross-validation** training scheme, where for each fold:

- 10 models with different random initialization are trained for 20 epochs, selecting the one with **lowest validation loss**;
- the best model from the previous step is trained for 130 epochs.

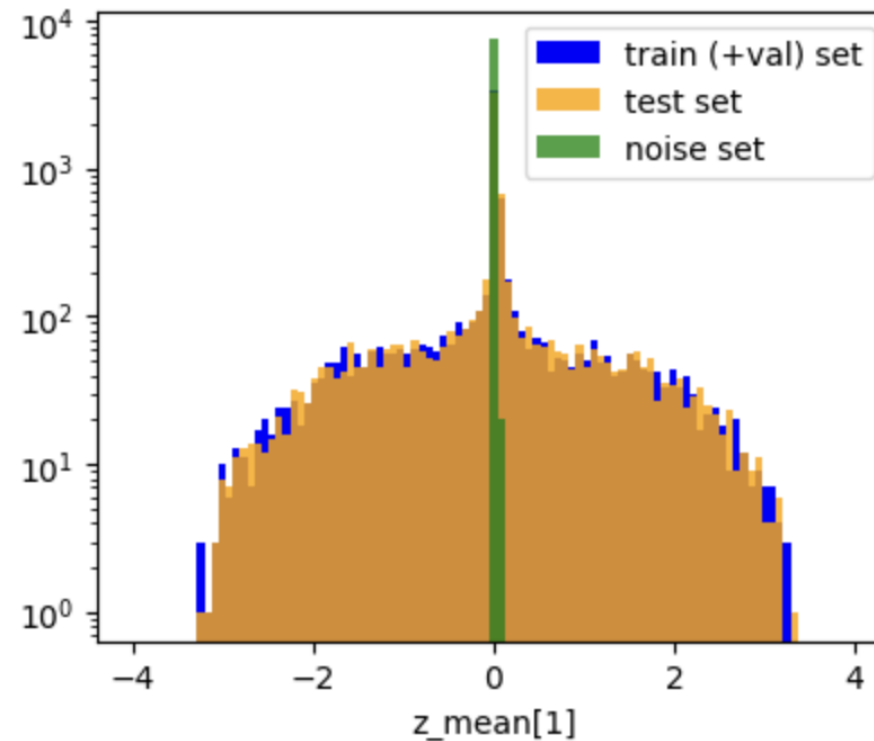
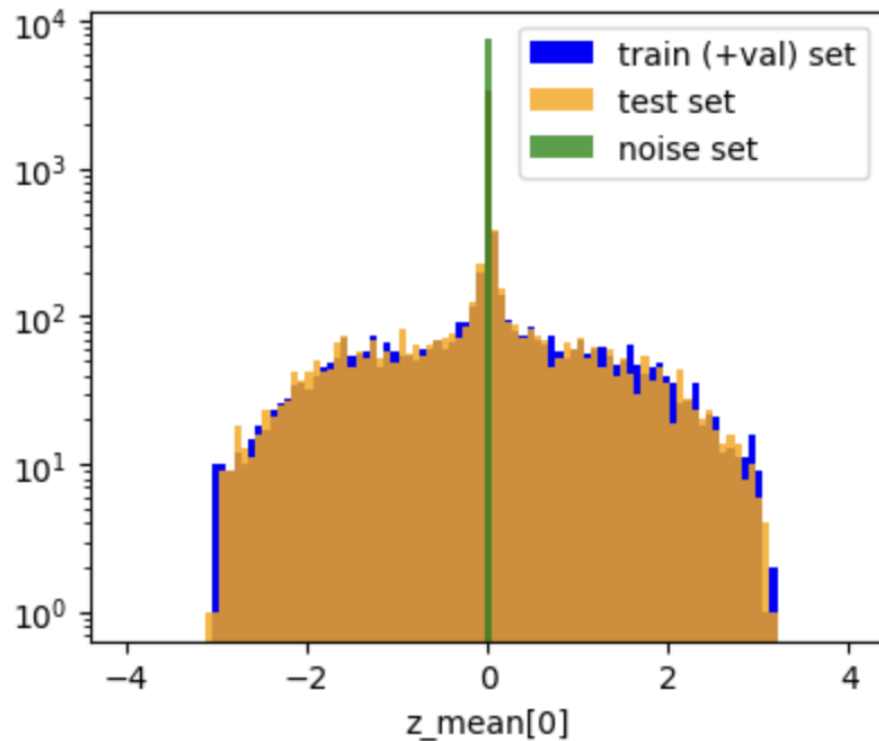


In ~20% of the cases the model never starts to learn.  
Work is in progress to improve this issue.



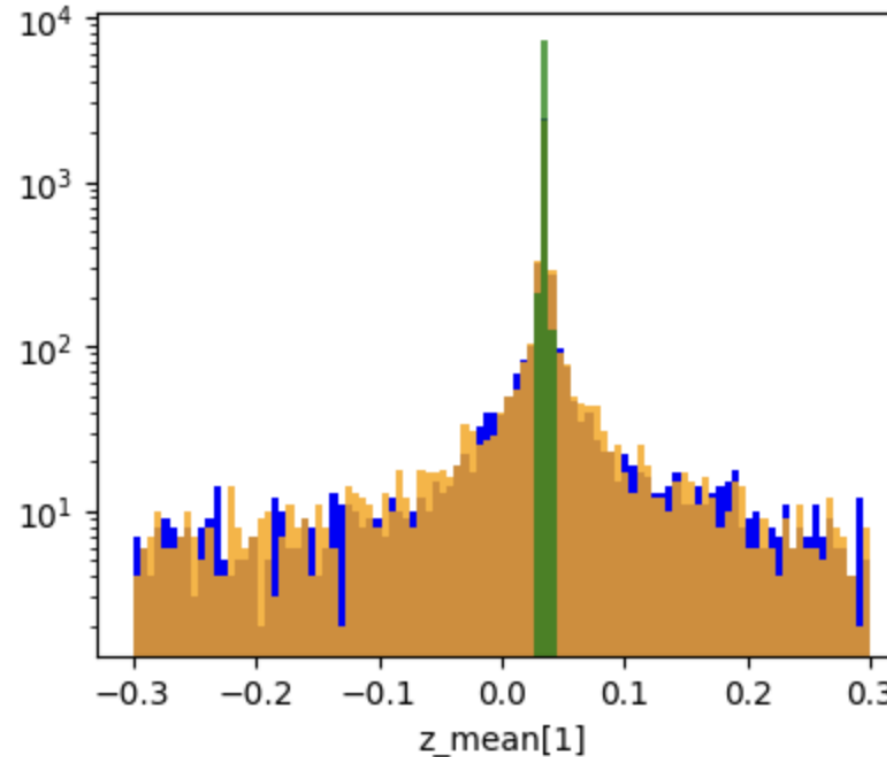
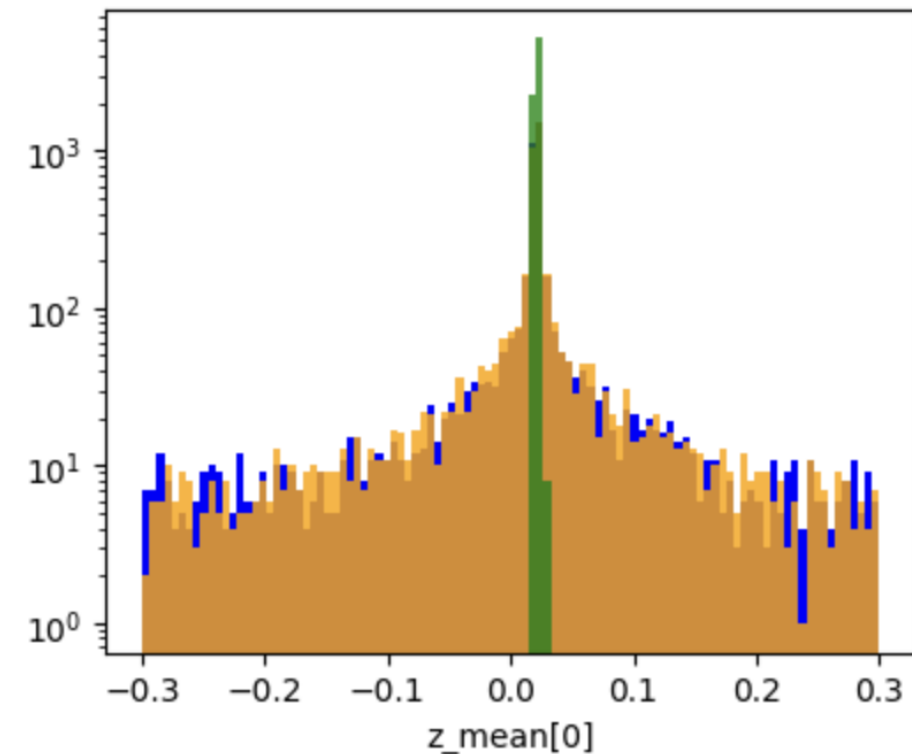
## Latent space

**Characterizing result:** waveforms with **negligible signals (*noise-only*)** are encoded into a limited region of the latent space where the  $z\_means$  simultaneously assume specific values.



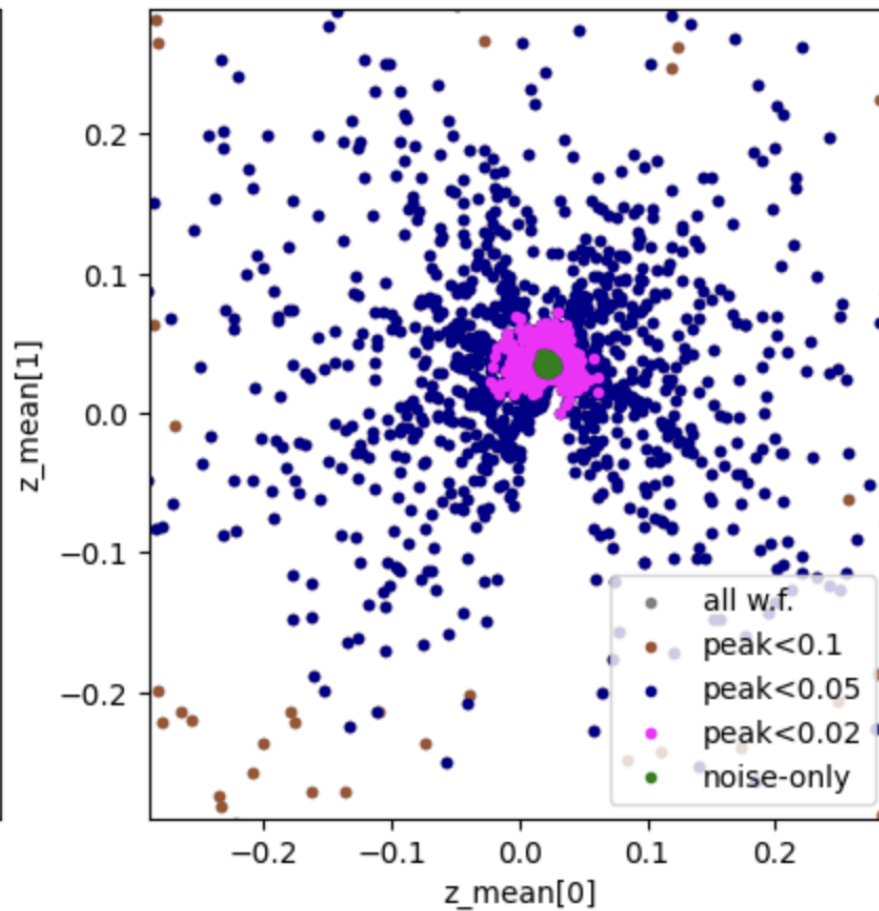
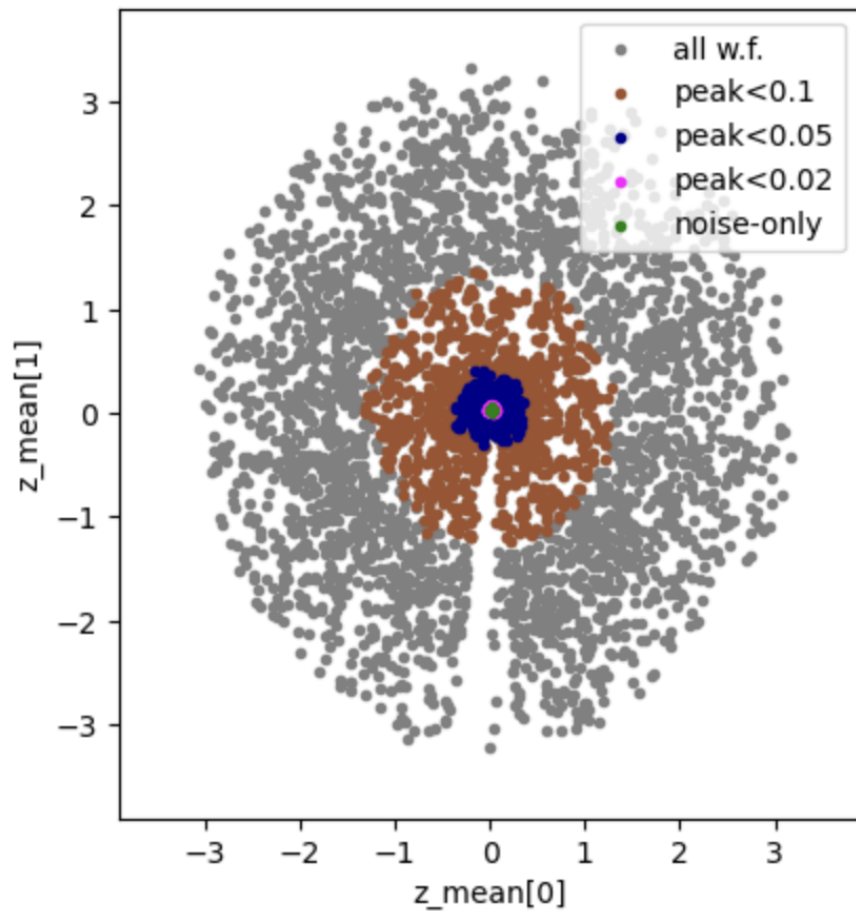
## Latent space

**Characterizing result:** waveforms with **negligible signals (*noise-only*)** are encoded into a limited region of the latent space where the  $z\_means$  simultaneously assume specific values.



**Notice that such *characteristic values* are not zeros !**

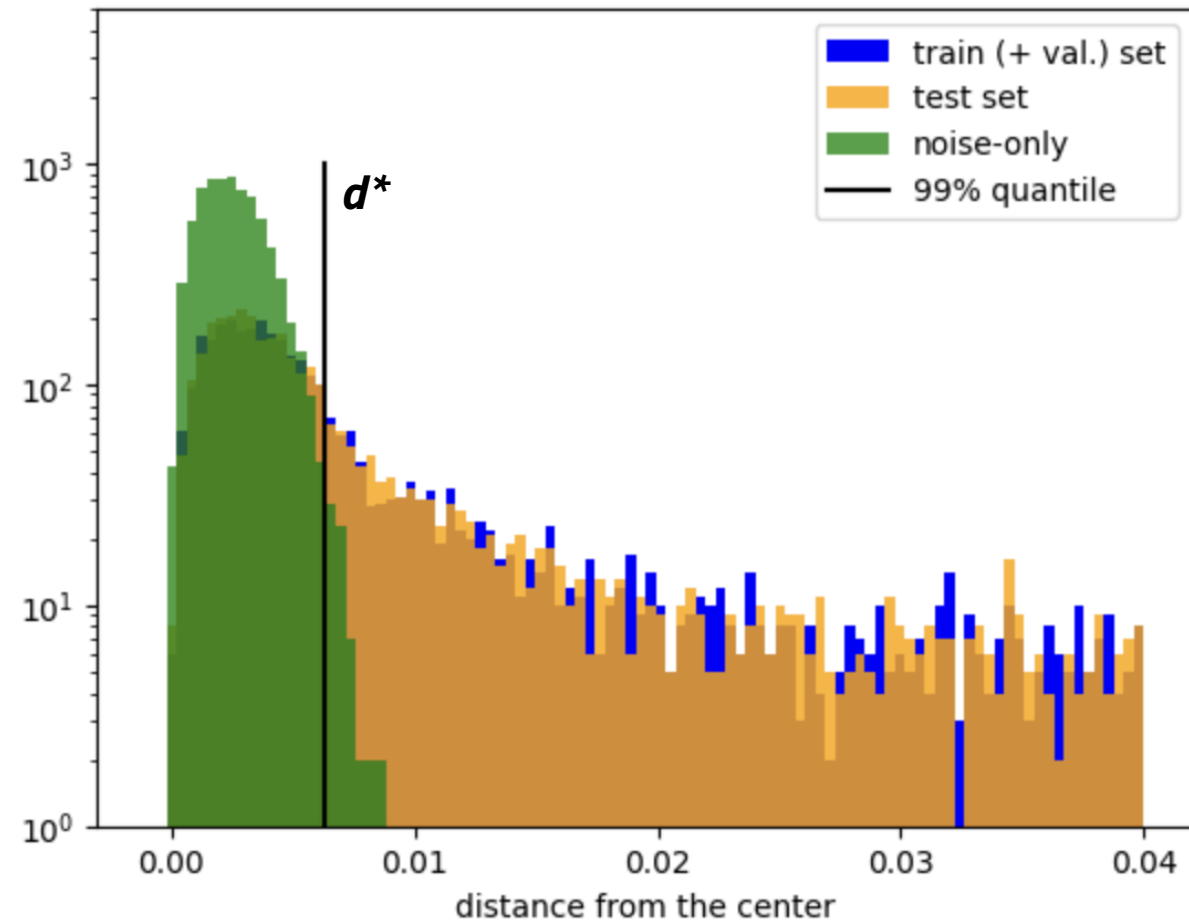
## Waveforms representation in the latent space



Notably, the smaller the signal (considering the peak in the true signal pulse) the nearest the corresponding w.f. is encoded to region (i.e. to the *characteristic values*) of noise-only cases.

## Definition of a region of “noise-only” waveforms

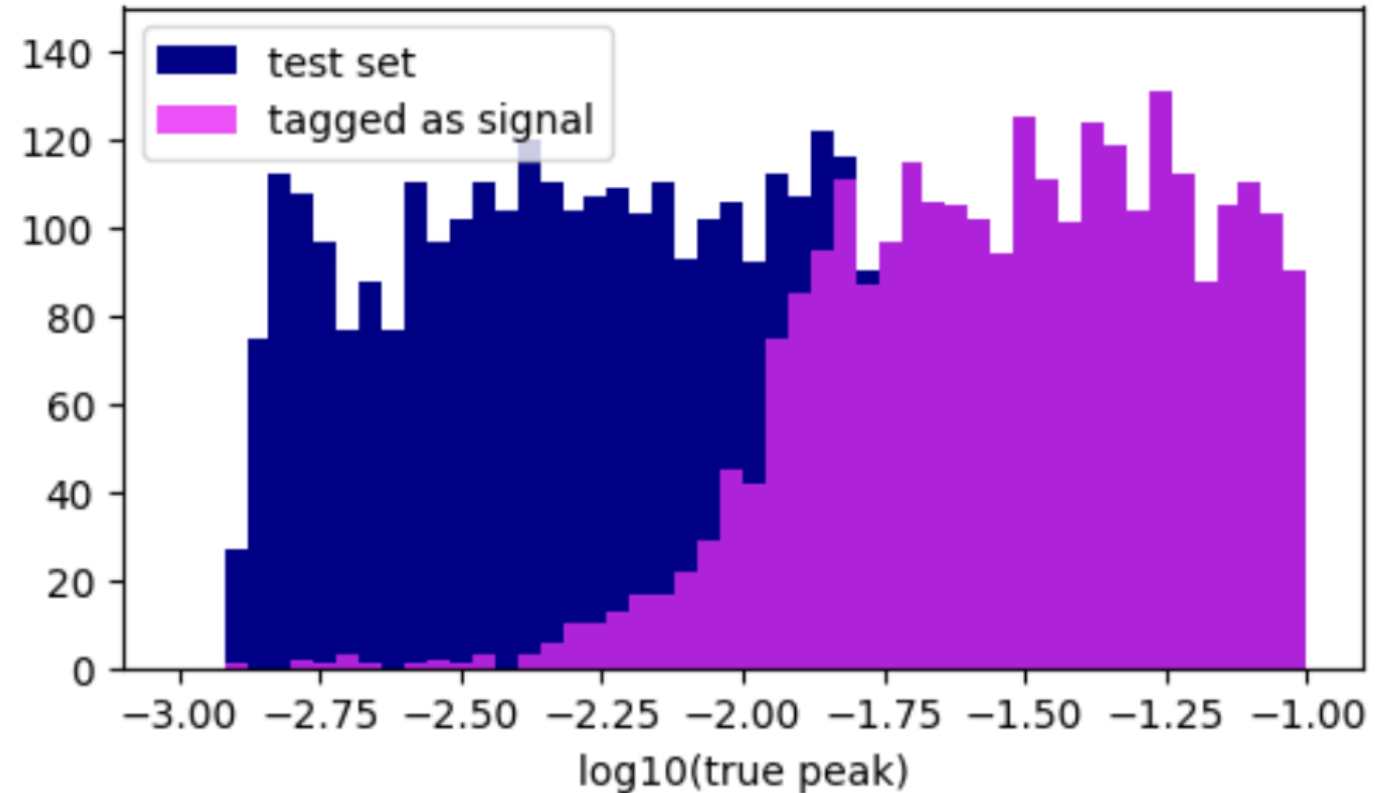
- ❖ Find the **center** (i.e. the couple of  $z\_mean[0]$  &  $z\_mean[1]$ ) of the latent space region where **noise-only** waveforms are encoded.
- ❖ Define a **distance from such a center** simply as
$$d = \sqrt{(z_{mean}[0] - z_{mean}^{center}[0])^2 + (z_{mean}[1] - z_{mean}^{center}[1])^2}$$
- ❖ Find the one-sided 99% quantile  $d^*$  in the distribution of distances for the noise-only waveforms (i.e. 99% of these w.f. are encoded within a region of the latent space where  $d < d^*$ ).



## Selection method to identify waveforms with signal

1. Process the waveform with the trained VCAE and extract the 2  $z_{\text{mean}}$  values;
2. Calculate the distance  $d$  from the centre of the noise-only region (as defined in the previous slide);
3. **If  $d > d^*$ , the w.f. is tagged as signal;**  
if not, the w.f. is tagged as noise.

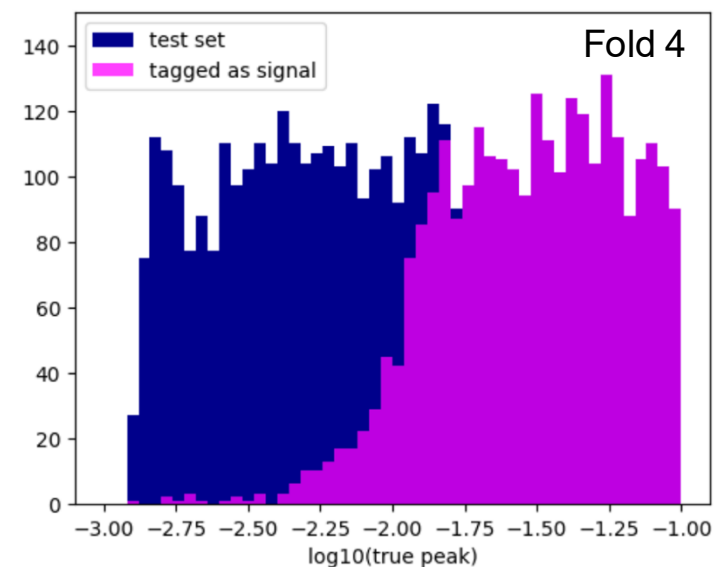
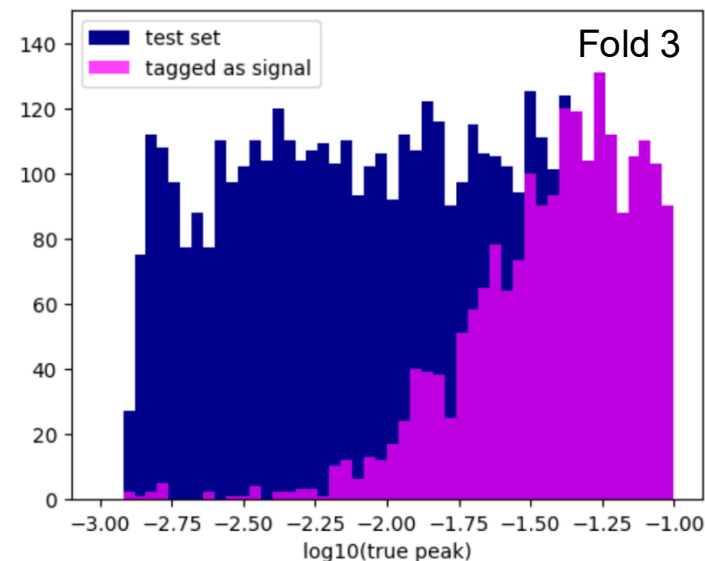
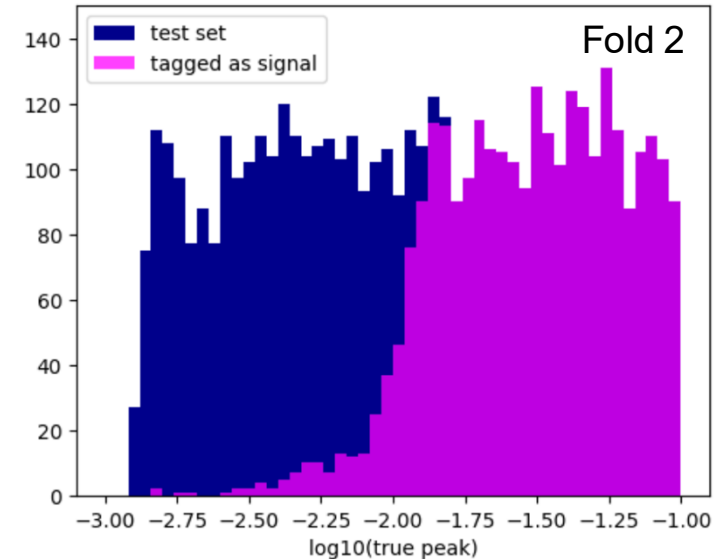
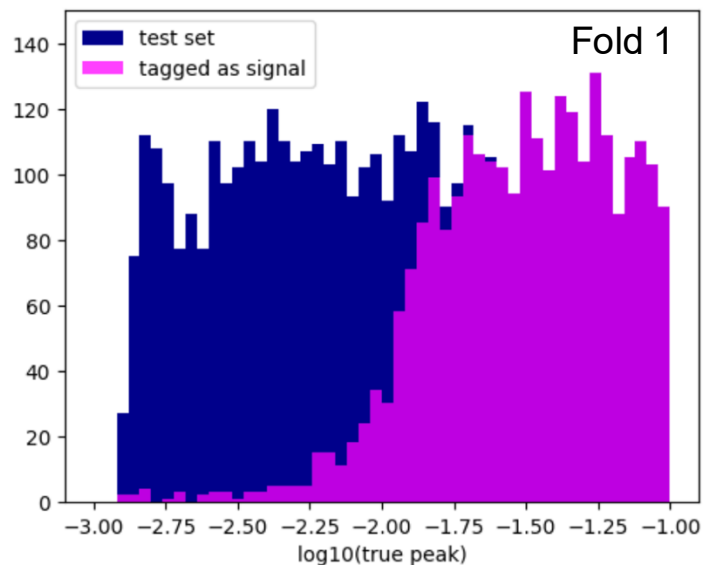
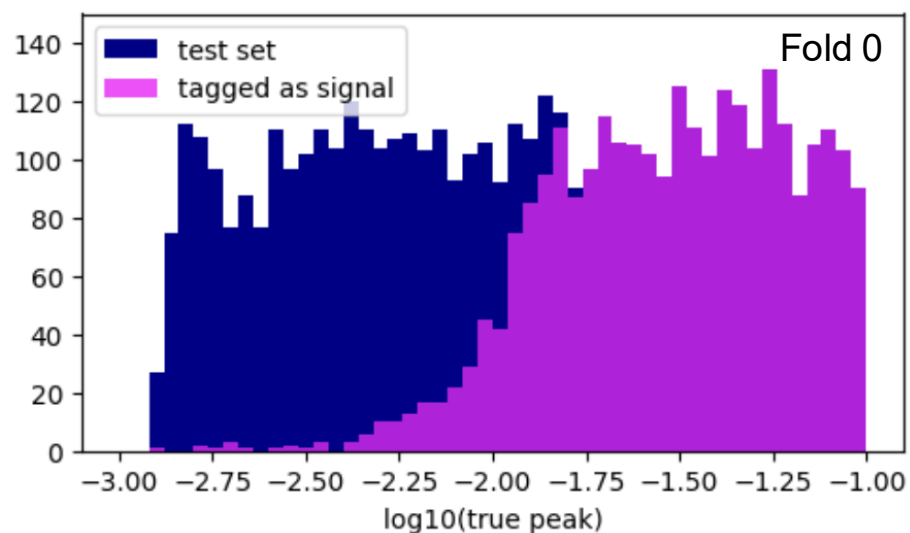
- **False positive rate = 1% by definition**
- **False negative rate:**  
function of signal magnitude (*for instance, peak in the true signal pulse*) and partially on the training itself (*see next slide*).



## Final result for different folds

The described procedure is repeated across the 5 trained models.

Differences in the results reflect the diverse organization of the latent space learned in each fold (*see also backup*).





## Future prospects

- ❑ Systematic study of the current VCAE model architecture under **different noise levels**, to assess its generalization capabilities and the (expected) differences in the performances.
- ❑ Investigation of VCAE models with **higher-dimensional latent spaces**, possibly resulting in a (more complicated but) more expressive structure of the latent representation.
- ❑ **Application of the improved VCAE model to the data of the *ReD* experiment**, and possibly in other contexts where the signal is difficult to separate from background noise, such as in low-energy direct dark matter detection experiments.



## Conclusions

- The reported observations highlight the potential of this (relatively simple) VCAE model for the identification of signal pulses in noisy time-series.
- Leveraging deeper or more expressive models may lead to further improvements in both accuracy and scalability.
- The **computational resources and infrastructure provided by ICSC** will allow extensive experimentation and further refinement of these models.

# Thanks for the attention!

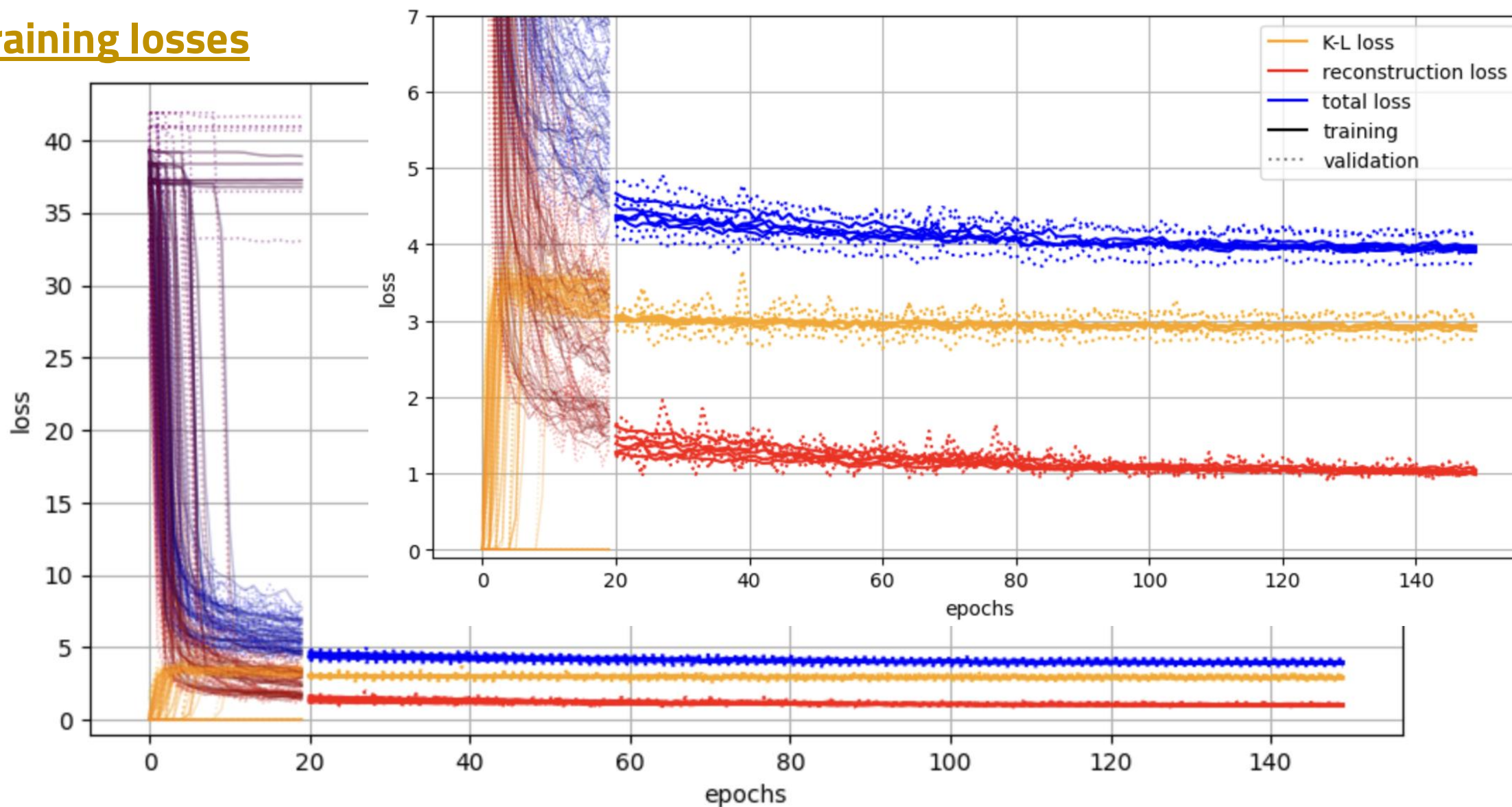
This work is supported by ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, funded by European Union – NextGenerationEU

The background features a dark blue, futuristic digital landscape. It is composed of a dense grid of small, light blue rectangular blocks that create a 3D effect. Two prominent, glowing blue light trails run vertically down the center, curving slightly towards the edges. These trails are made of many thin, parallel lines of light, with small, bright blue dots scattered along them, suggesting data flow or network connections. The overall atmosphere is high-tech and digital.

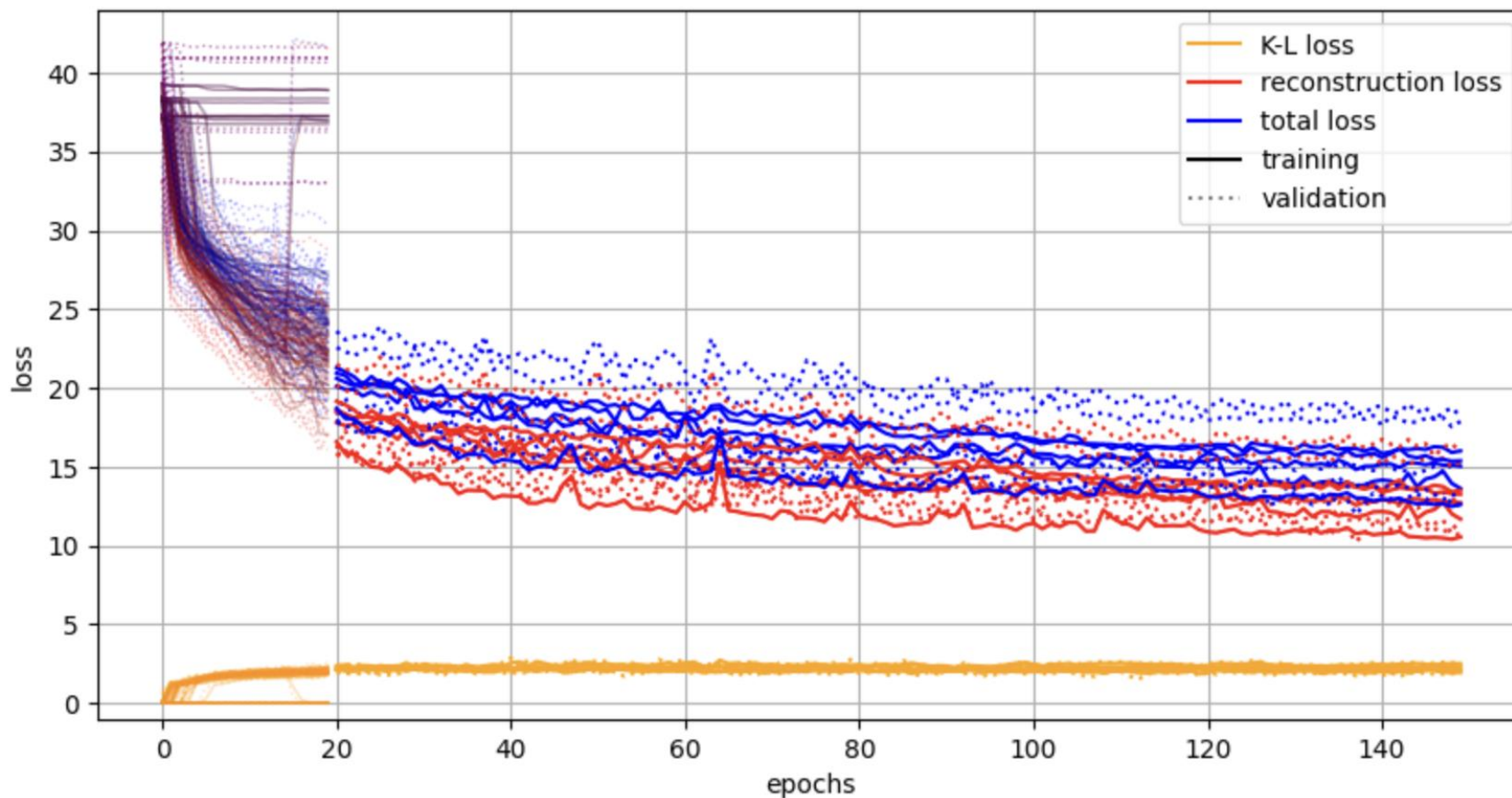
**Backup**



## Training losses



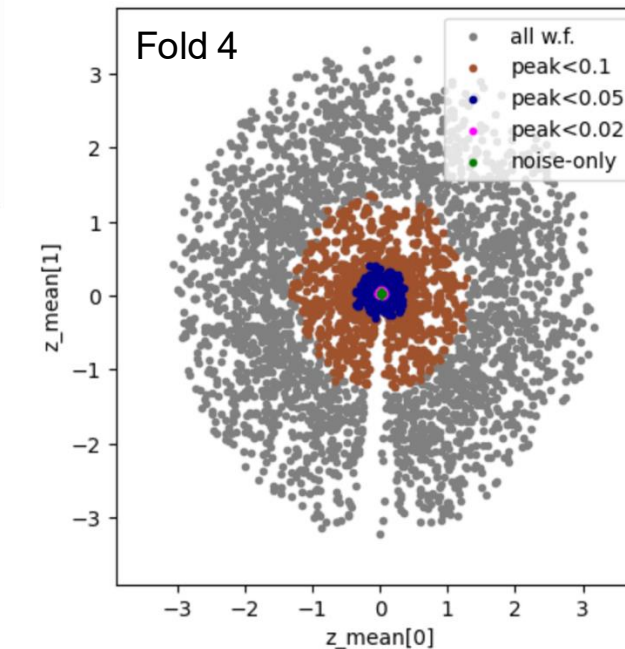
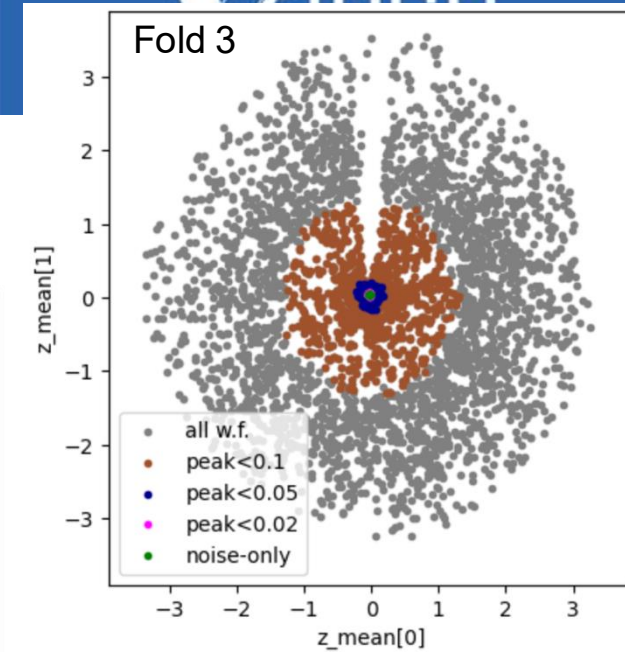
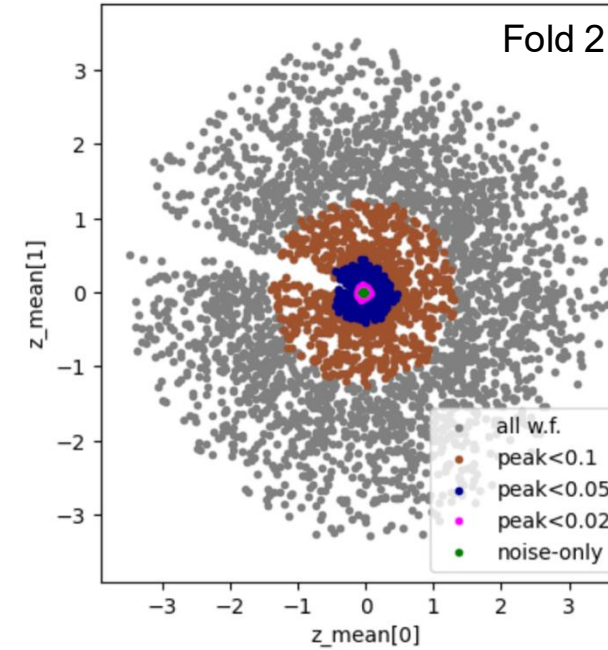
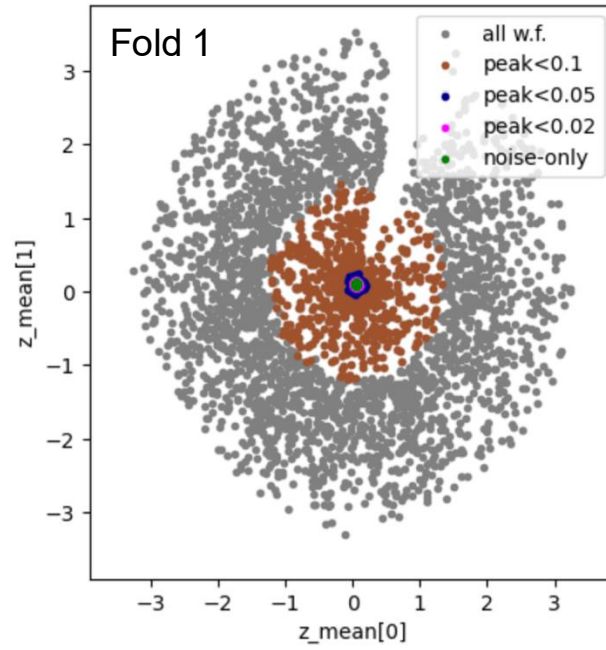
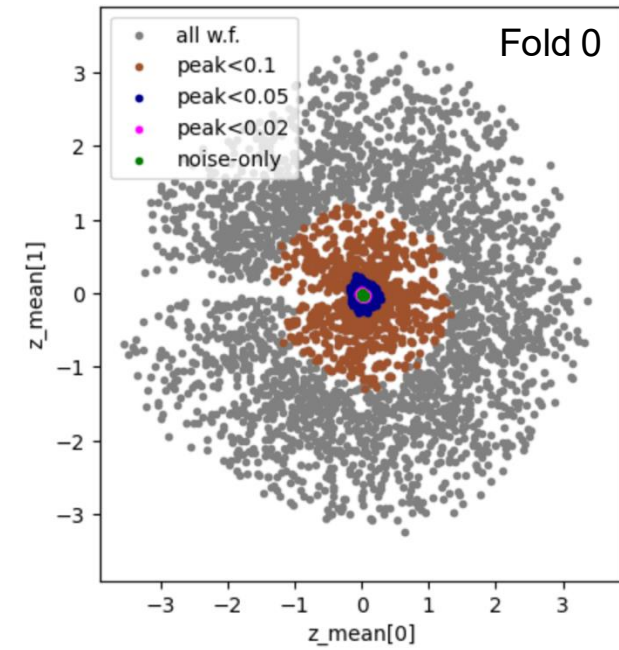
## VCAE with latent space dimension 1



**Much worse reconstruction loss!**

With only one sampling parameter, the model cannot efficiently re-build the waveforms.

## Latent space structure across folds

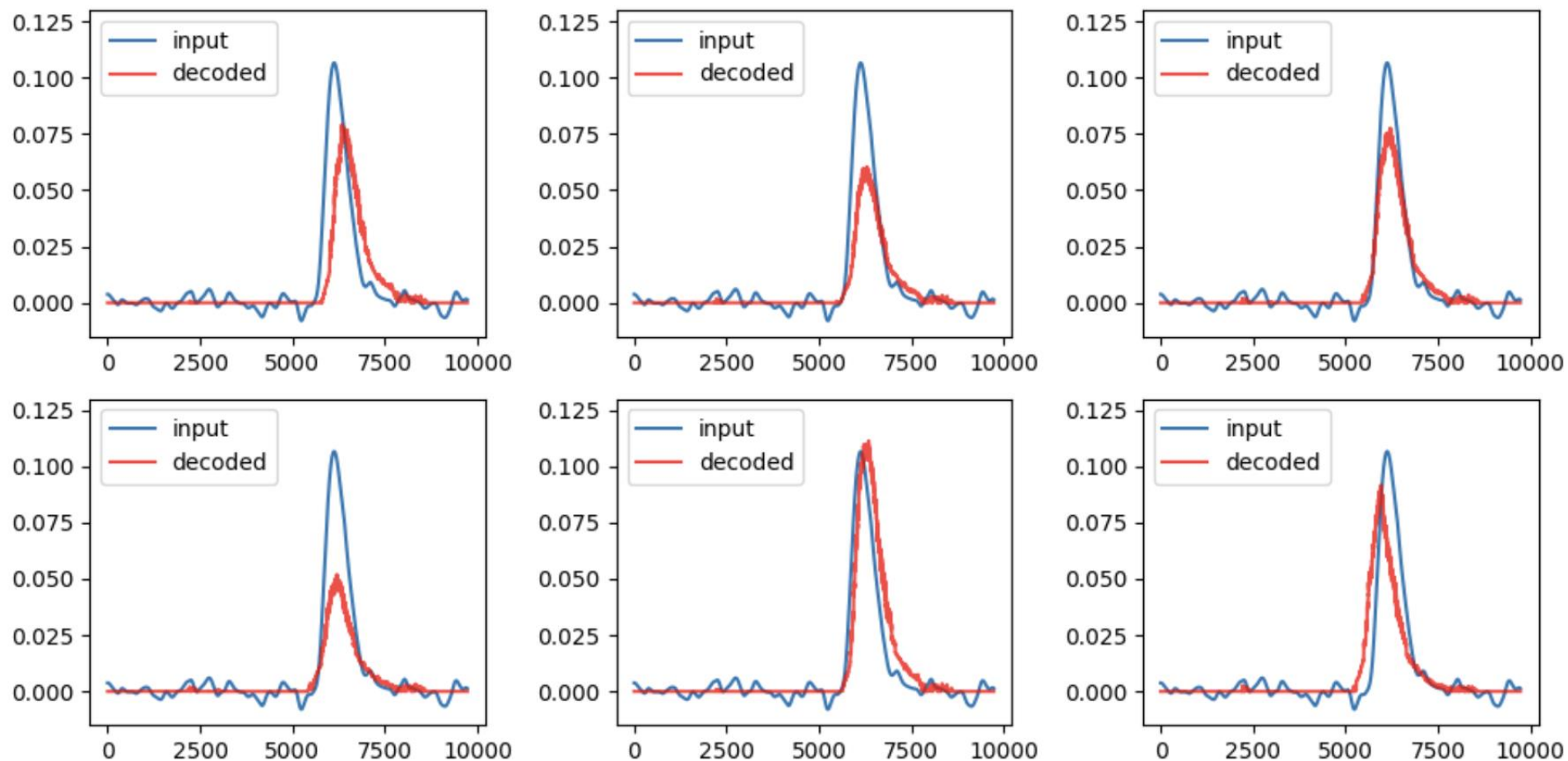


Empirical finding ([L. Moschella et al., ICLR 2023](#)): while global orientation varies, angles between encodings are preserved

Proposed solution: represent each sample by its similarity to a fixed set of anchors, yielding more stable and transferable embeddings.



## Reconstructed waveforms



Given the sampling process happening at the bottleneck of a Variational AE, the decoded w.f. are slightly different each time a new sample is drawn (i.e. the decoder acts as a *generative network*).