Deep Learning Applications for collider physics
Lecture 5

Maurizio Pierini

European Research Council

Plan for this week

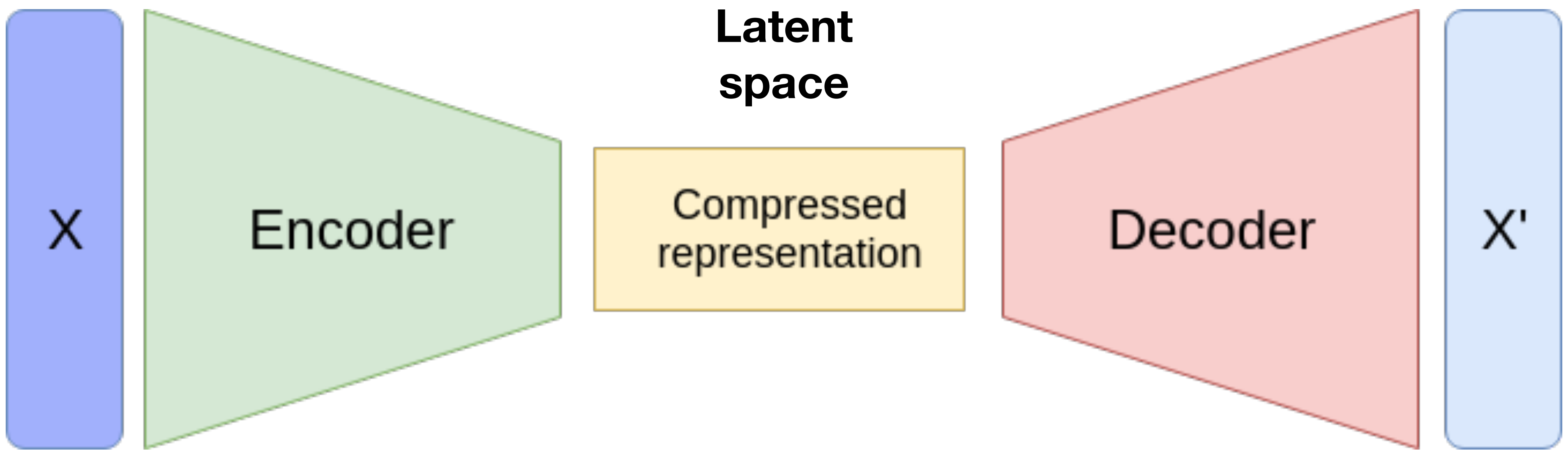
	Day1	Day2	Day3	Day4	Day5
Lecture	Introduction	ConvNN	RNNs	Graphs	Unsupervised Learning
Tutorial	Fully Connected Classifier	ConvNN Classifier	RNNs Classifier	Graphs Classifier	Anomaly Detection

Autoencoders

- *Autoencoders are networks with a typical "bottleneck" structure, with a symmetric structure around it*
    - *They go from $\mathbb{R}^n \rightarrow \mathbb{R}^n$*
    - *They are used to learn the identity function as $f^{-1}(f(x))$*

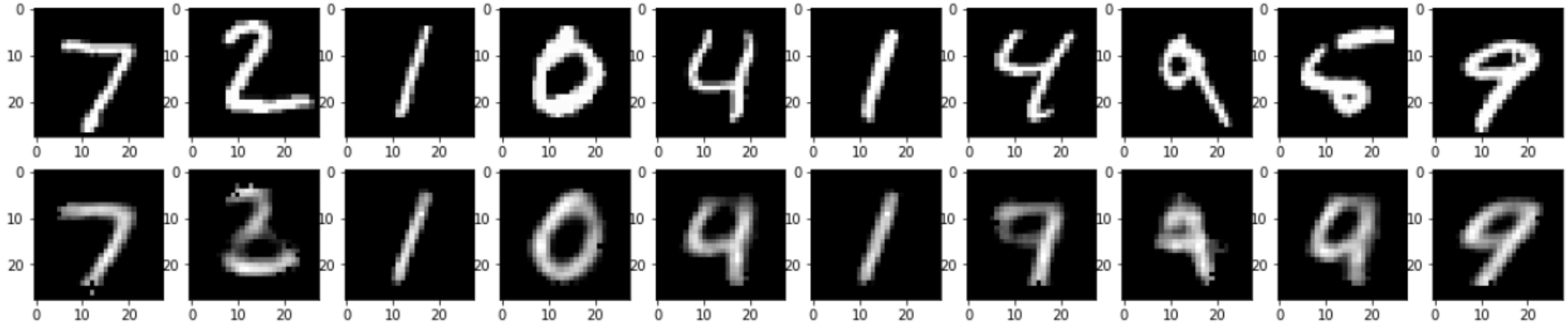
where $f: \mathbb{R}^n \rightarrow \mathbb{R}^k$ and $f^{-1}: \mathbb{R}^k \rightarrow \mathbb{R}^n$

- *Autoencoders are essential tools for unsupervised studies*



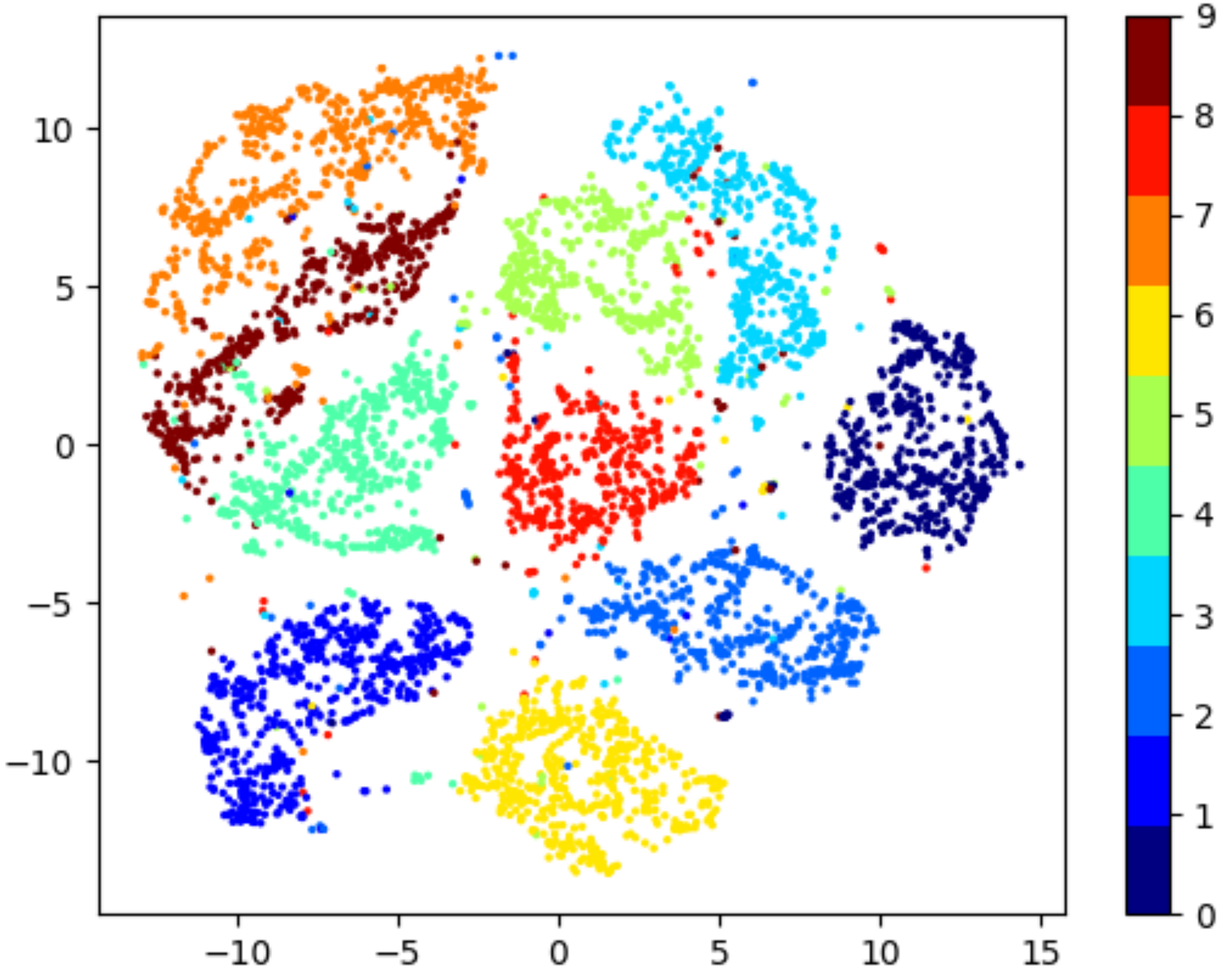
Dimensional Reduction

- Autoencoders can be seen as compression algorithms
  - The n inputs are reduced to k quantities by the encoder
  - Through the decoder, the input can be reconstructed from the k quantities
- As a compression algorithm, an auto encoder allows to save (n-k)/n of the space normally occupied by the input dataset



Clustering

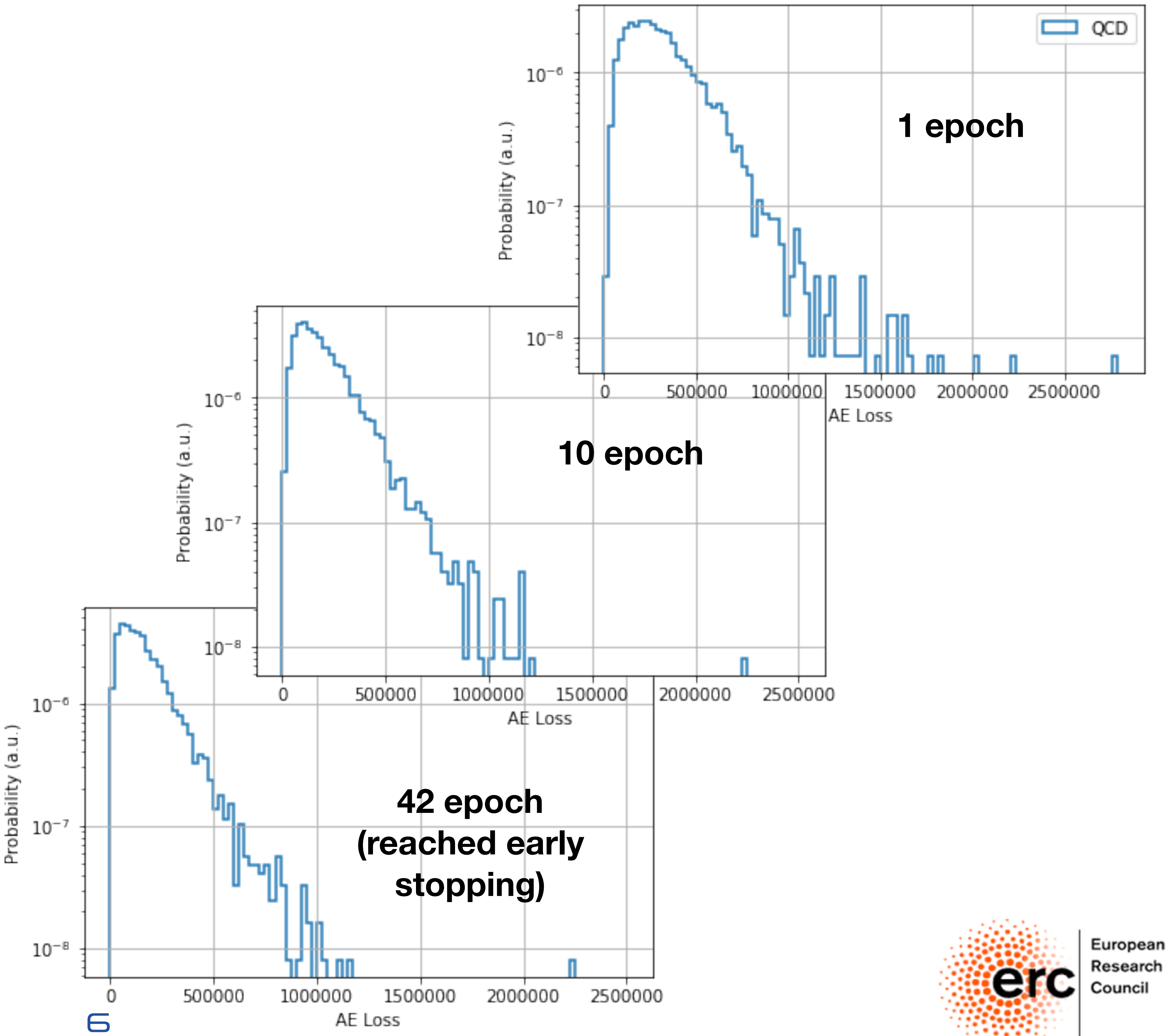
- The auto encoder can be used as a clustering algorithm
  - Alike inputs tend to populate the same region of the latent space
  - Different inputs tend to be far away



European Research Council

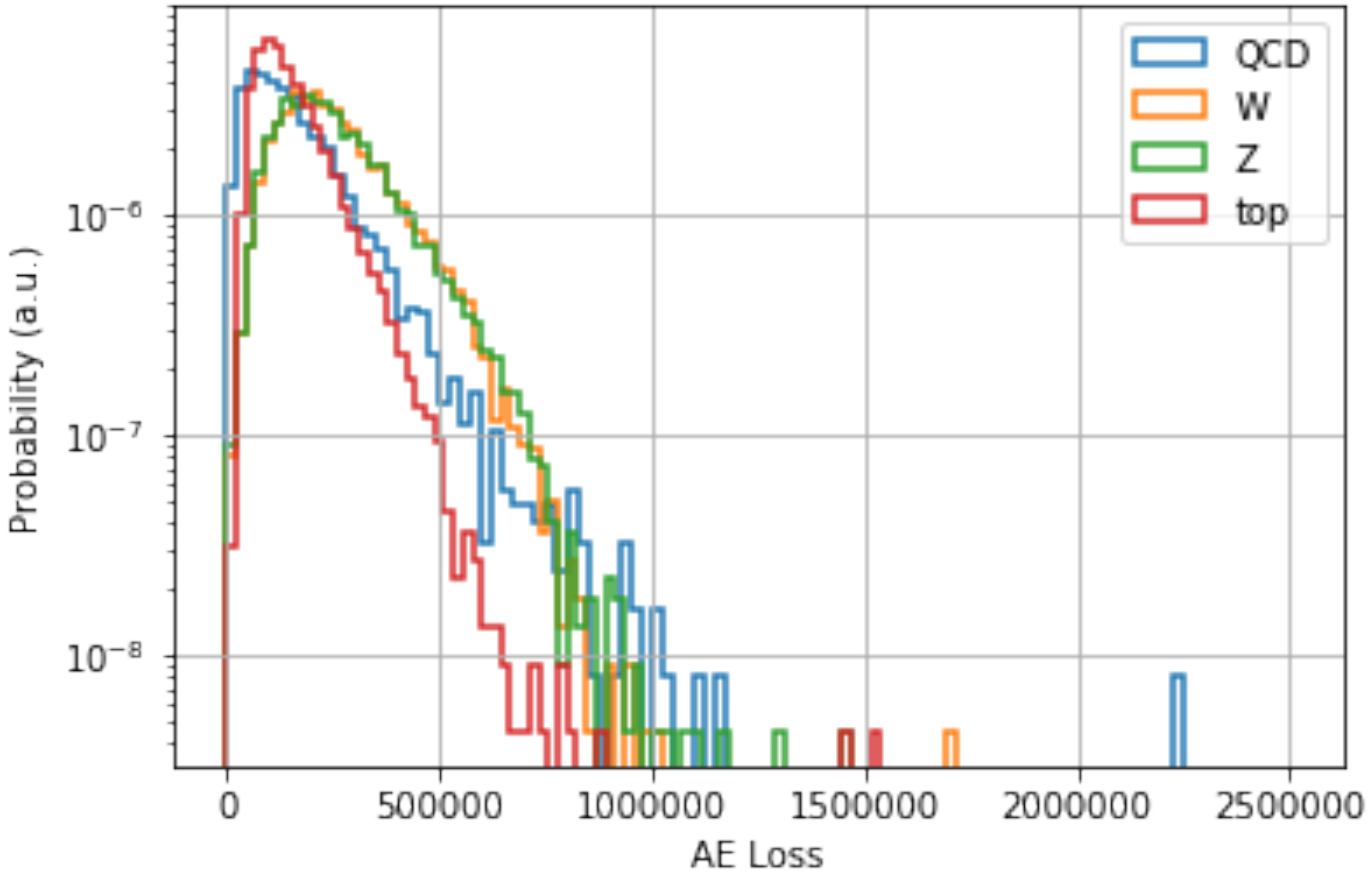
Training an Autoencoder

- AEs are training minimizing the distance between the inputs and the corresponding outputs
- The loss function represents some distance metric between the two
    - e.g., MSE loss
- A minimal distance guarantees that the latent representation + decoder is enough to reconstruct the input information



Anomaly detection

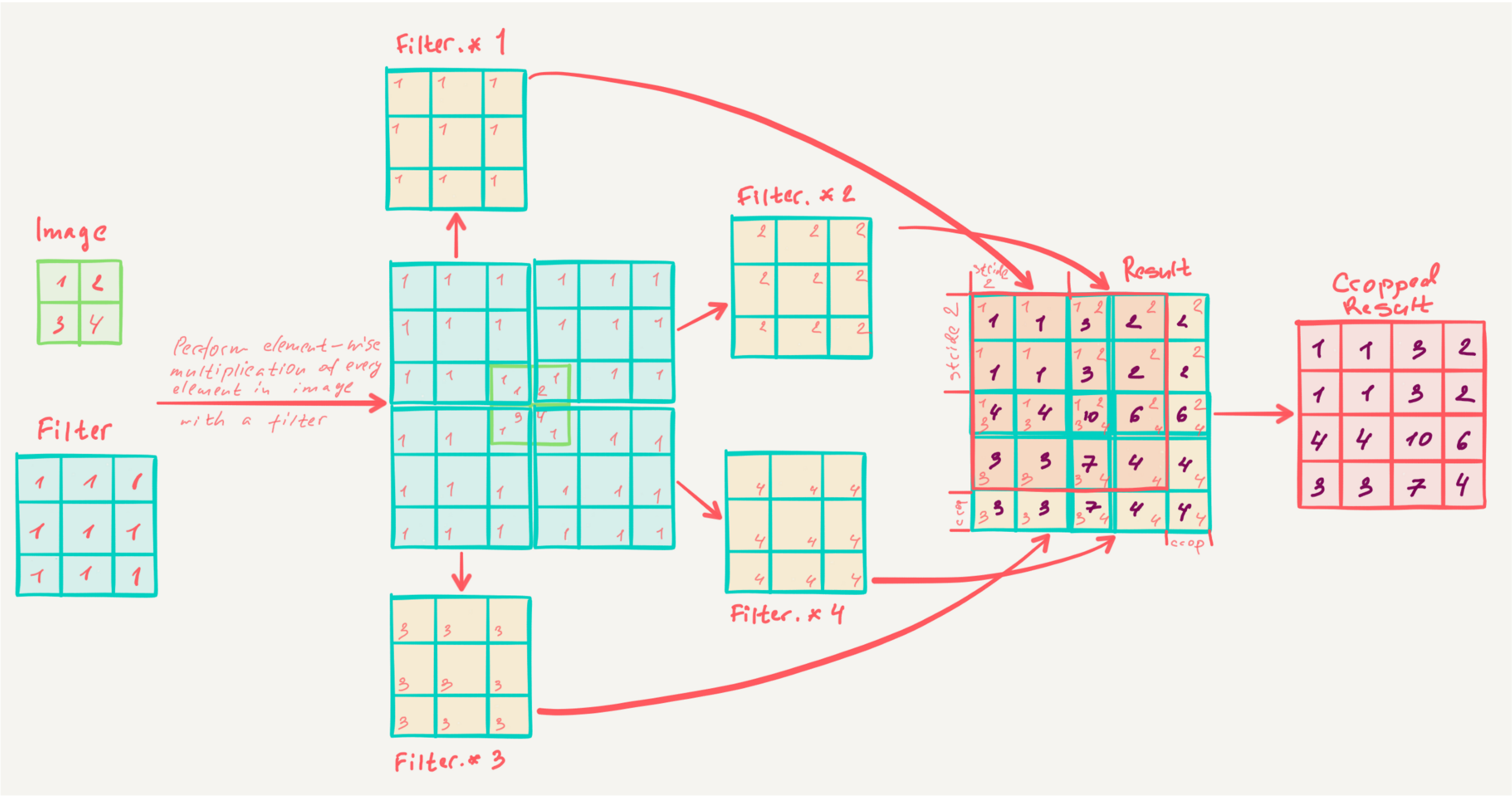
- Once trained, an autoencoder can reproduce new inputs of the same kind of the training dataset
  - The distance between the input and the output will be small
- If presented an event of some new kind (anomaly), the encoding-decoding will tend to fail
  - In this circumstance, the loss (=distance between input and output) will be bigger



European Research Council

Convolutional Autoencoders

- Conv Autoencoders take images as input
- Through Conv and MaxPooling, they reduce it to some latent-space 1D array
- This 1D array is expanded using the inverse of the encoder functions
  - ConvTranspose (aka “Deconvolution”)
  - Upsampling



“Bed of Nails”

1	2
3	4

Input: 2 x 2

1	0	2	0
0	0	0	0
3	0	4	0
0	0	0	0

Output: 4 x 4

Nearest Neighbor

1	2
3	4

Input: 2 x 2

1	1	2	2
1	1	2	2
3	3	4	4
3	3	4	4

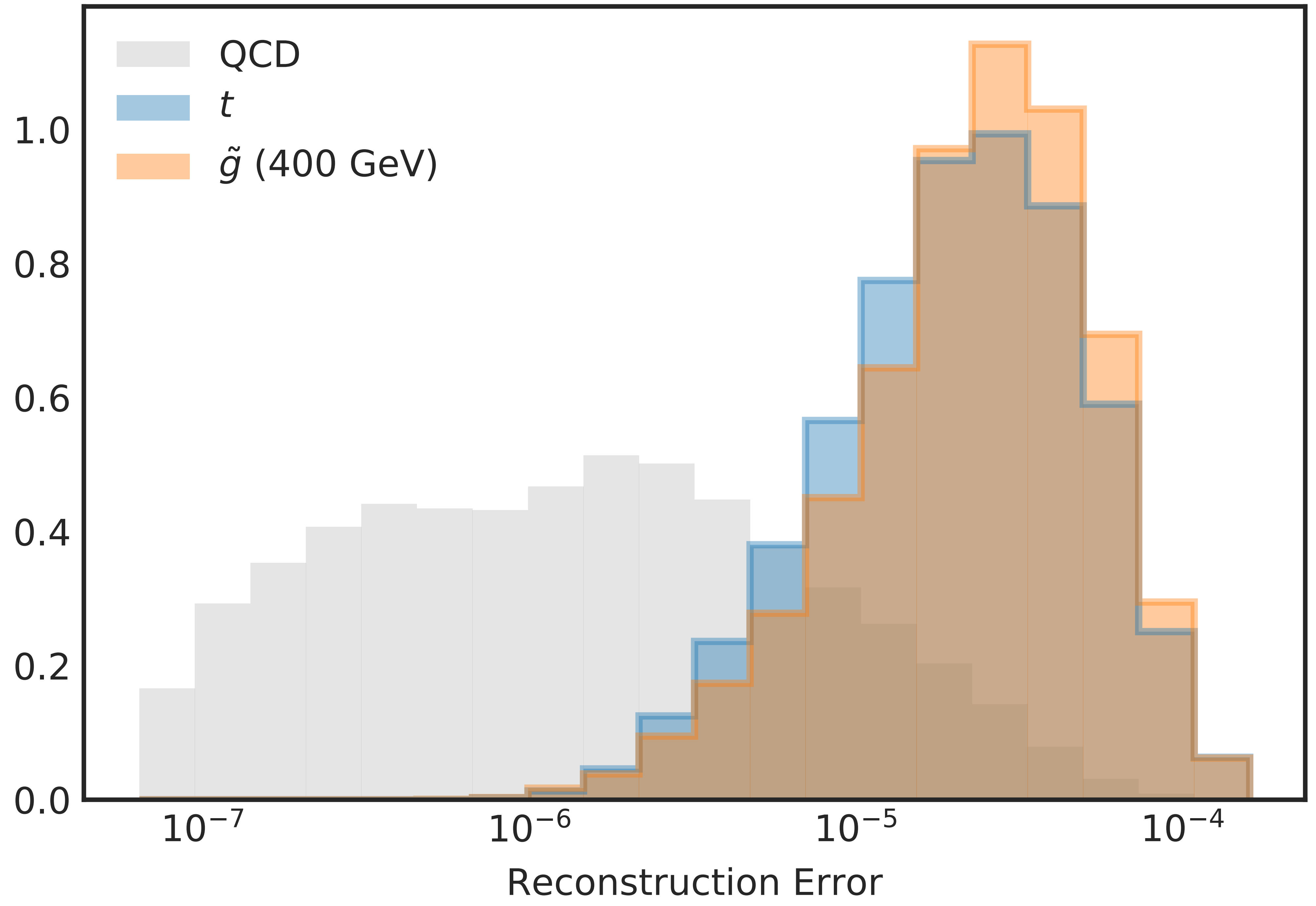
Output: 4 x 4

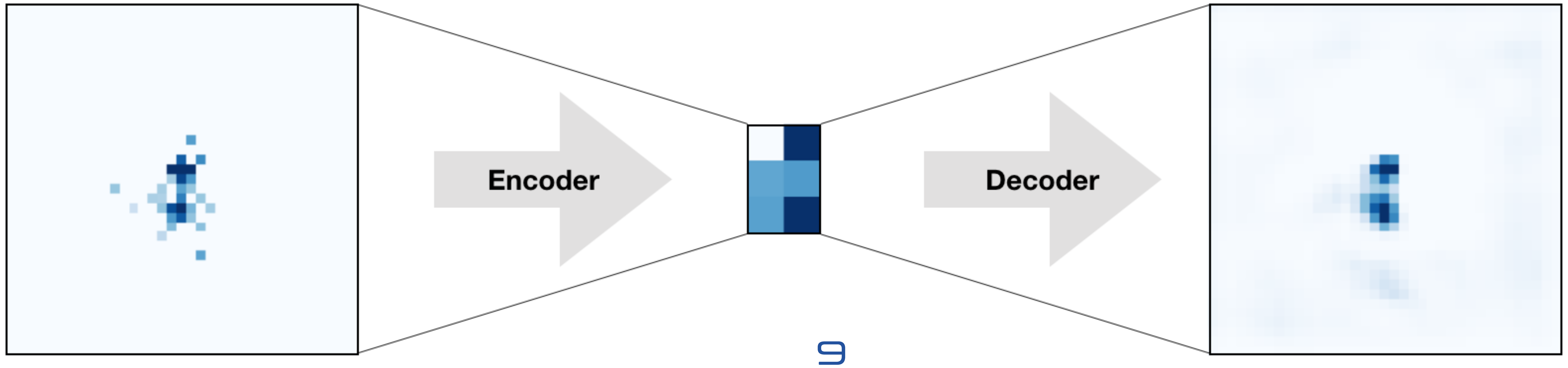
Example: Jet autoencoders

[Farina et al., arXiv:1808.08992](https://arxiv.org/abs/1808.08992)

[Heimel et al., arXiv:1808.08979](https://arxiv.org/abs/1808.08979)

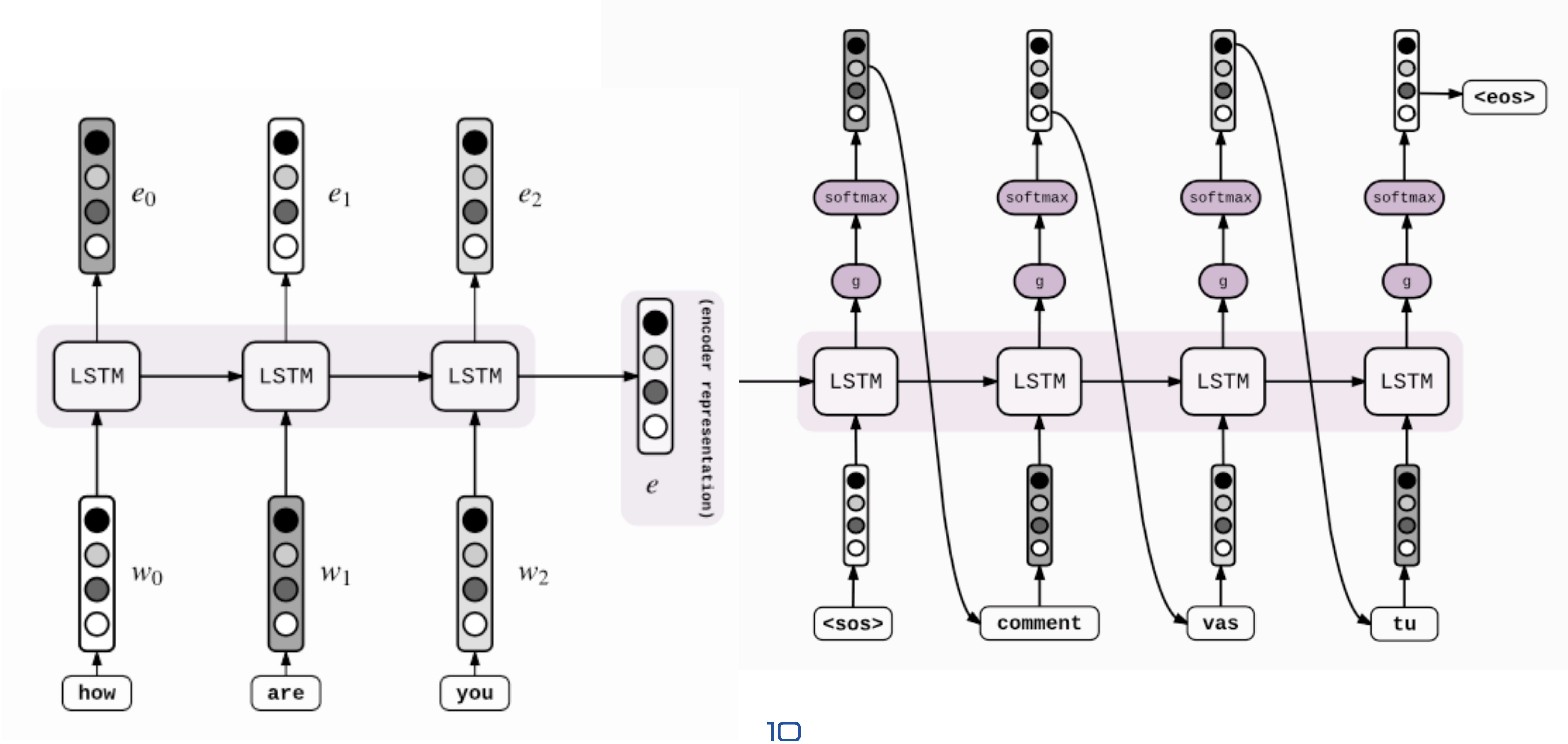
- *Idea applied to tagging jets, in order to define a QCD-jet veto*
- *Applied in a BSM search (e.g., dijet resonance) could highlight new physics signal*
- *Based on image and physics-inspired representations of jets*

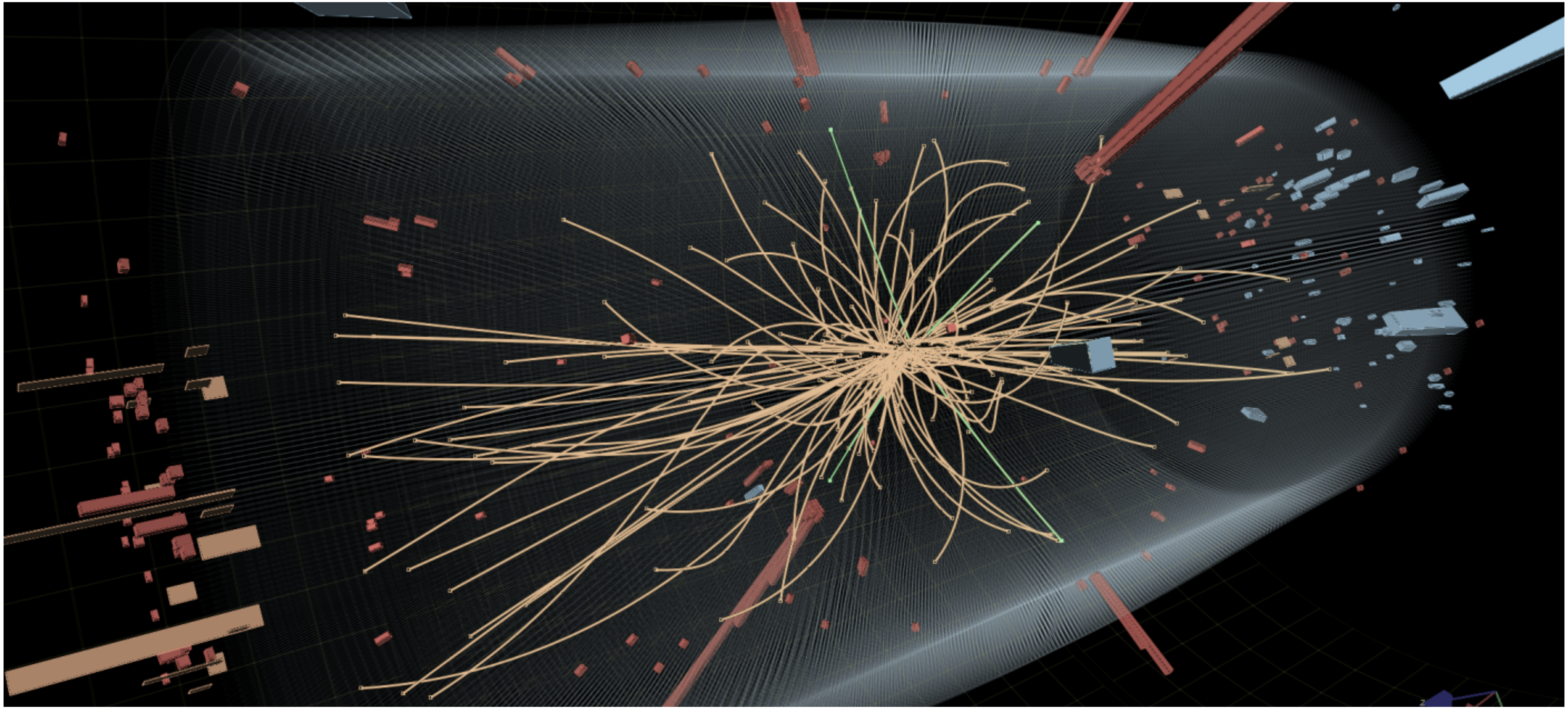




Recurrent Autoencoders

- When given as input a sequence, the AE needs a recurrent layer to process it
  - The encoder is similar to the classifier we already saw
  - What about the decoder? This is where the serial output of the RNN comes in

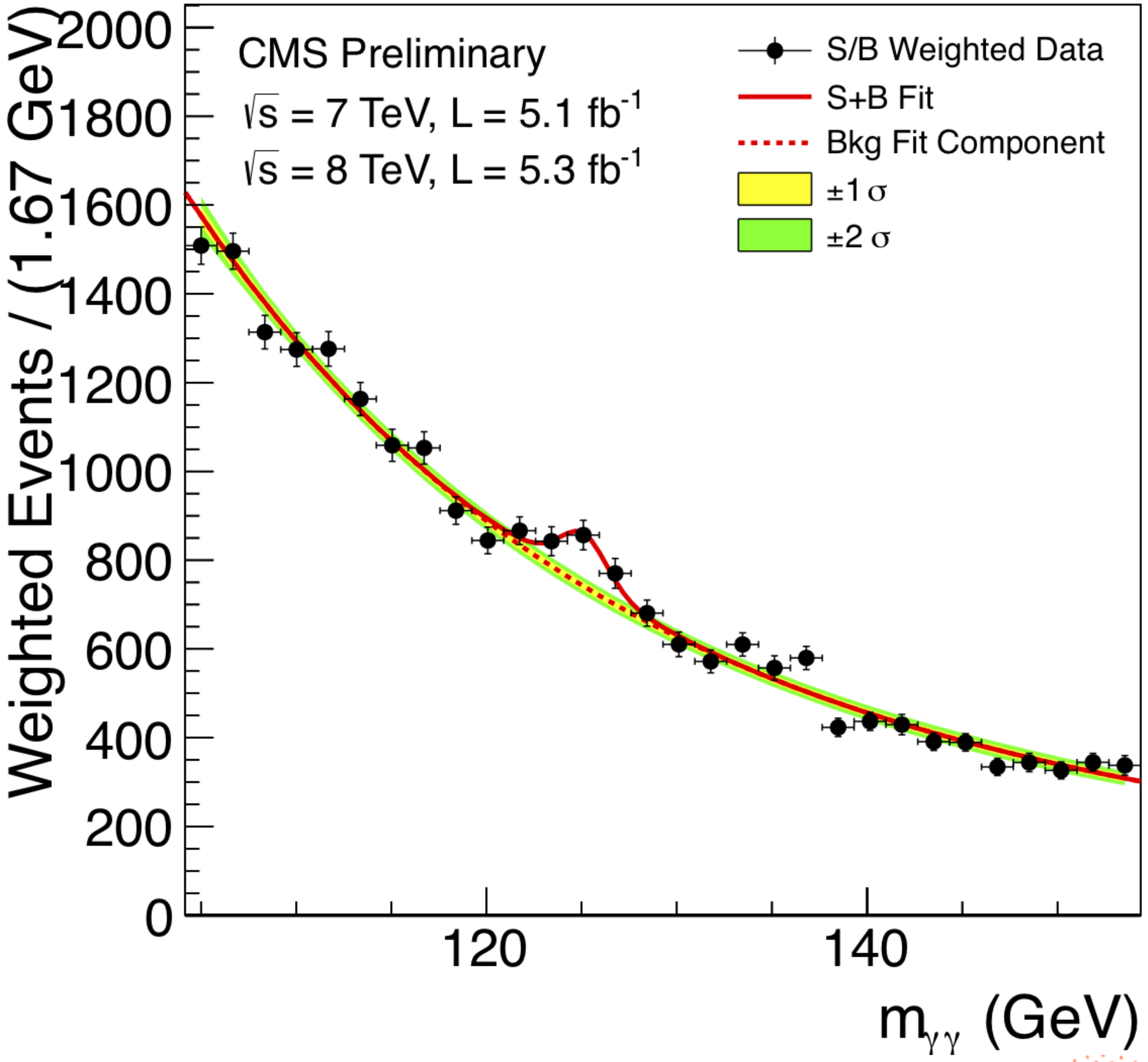




Autoencoders for New Physics searches

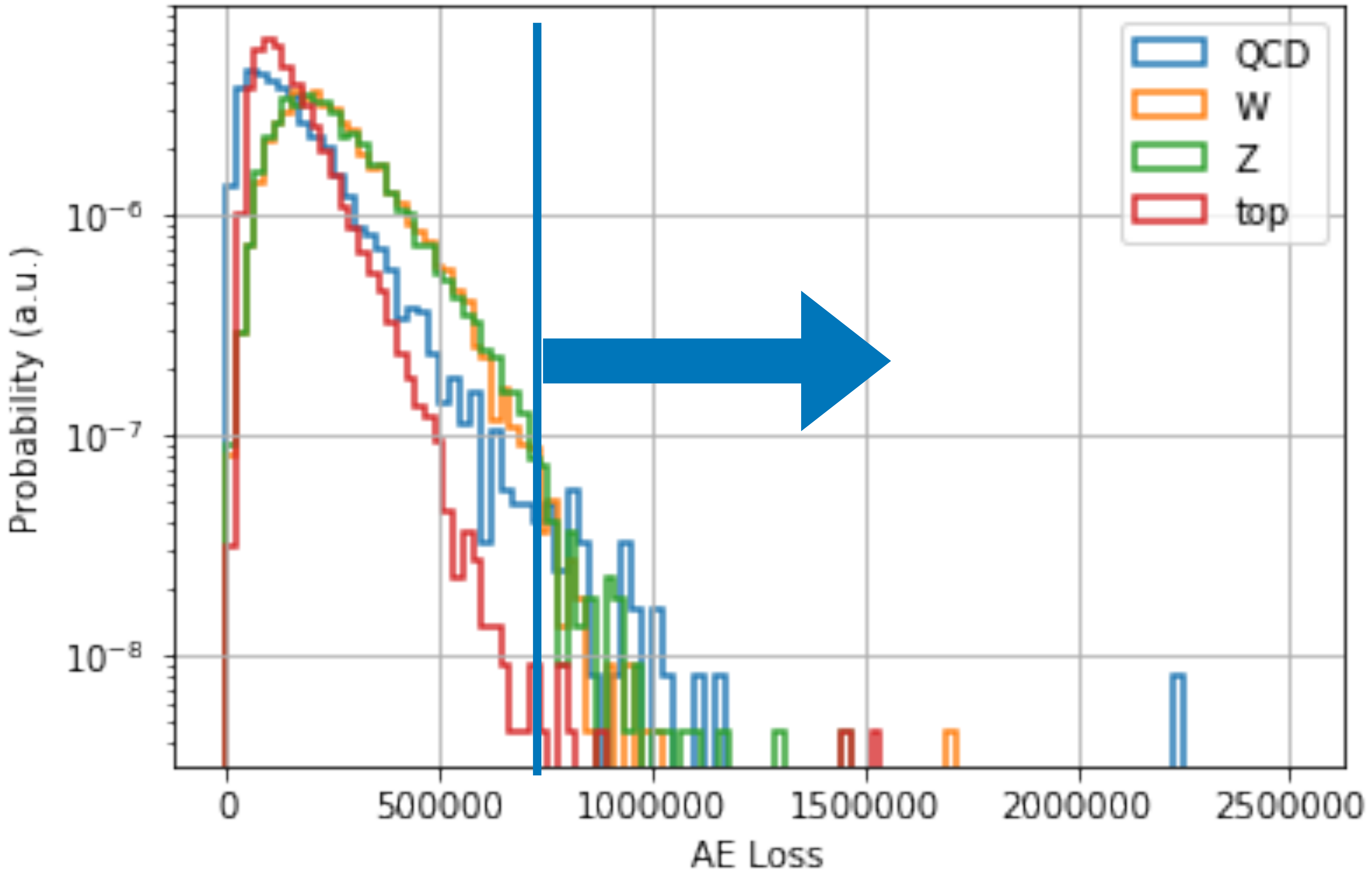
Supervised search for new physics

- *Searches for new physics are typically supervised*
    - *One knows what to look for*
    - *MC simulation provides labelled datasets to model the signal and the background*
    - *The analysis is performed as hypothesis testing*
- *The bias (what to look for) enters very early in the game (often already at trigger level). What if we are looking in the wrong place?*



Unsupervised search for new physics

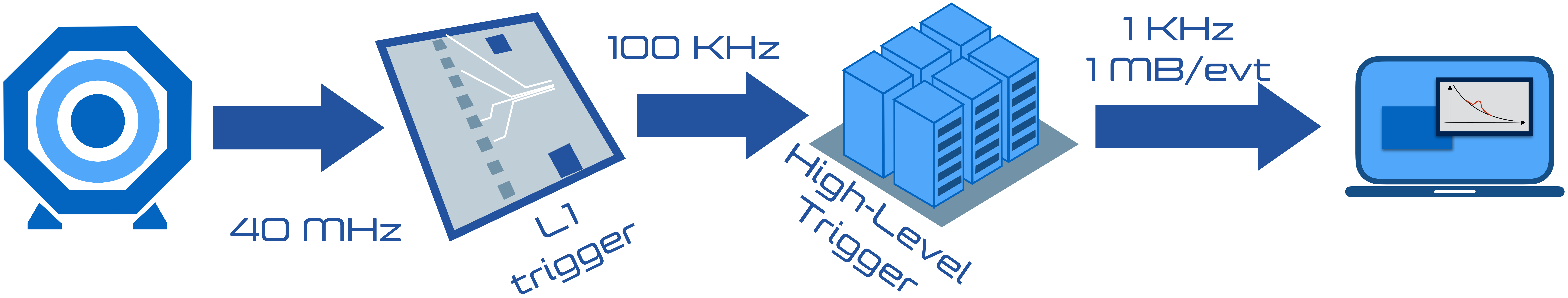
- *One can use Autoencoders to relax the assumption on the nature of new physics*
  - *Train on standard events*
  - *Run autoencoder on new events*
  - *Consider as anomalous all events with loss > threshold*



Running in the trigger

- *One needs the unsupervised algorithm to run before data are discarded*

- *This would allow to possibly notice recurrent patterns across events -> suggest explanations (new models) -> runs a classic supervised search (+ dedicated trigger) on the data to come*



Our use case: ℓ+X @HLT

- *Consider a stream of data coming from L1*
  - *Passed L1 because of 1 lepton (e,m) with pT>23 GeV*
  - *At HLT, very loose isolation applied*
  - *Sample mainly consists of W, Z, tt & QCD (for simplicity, we ignore the rest)*

Standard Model processes					
Process	Acceptance	Trigger efficiency	Cross section [nb]	Events fraction	Event /month
W	55.6%	68%	58	59.2%	110M
QCD	0.08%	9.6%	$1.6 \cdot 10^5$	33.8%	63M
Z	16%	77%	20	6.7%	12M
$t\bar{t}$	37%	49%	0.7	0.3%	0.6M

- *We consider 21 features, typically highlighting the difference between these SM processes (no specific BSM signal in mind)*

- The isolated-lepton transverse momentum p_T^ℓ.
- The three isolation quantities (CHPFISO, NEUPFISO, GAMMAPFISO) for the isolated lepton, computed with respect to charged particles, neutral hadrons and photons, respectively.
- The lepton charge.
- A boolean flag (ISELE) set to 1 when the trigger lepton is an electron, 0 otherwise.
- S_T, i.e. the scalar sum of the p_T of all the jets, leptons, and photons in the event with $p_T > 30$ GeV and $|\eta| < 2.6$. Jets are clustered from the reconstructed PF candidates, using the FASTJET [23] implementation of the anti-k_T jet algorithm [24], with jet-size parameter R=0.4.
- The number of jets entering the S_T sum (N_J).
- The invariant mass of the set of jets entering the S_T sum (M_J).
- The number of these jets being identified as originating from a b quark (N_b).
- The missing transverse momentum, decomposed into its parallel ($p^{\text{miss}}_{T,\parallel}$) and orthogonal ($p^{\text{miss}}_{T,\perp}$) components with respect to the isolated lepton direction. The missing transverse momentum is defined as the negative sum of the PF-candidate p_T vectors:

$$\vec{p}_T^{\text{miss}} = -\sum_q \vec{p}_T^{\,q} \,. \tag{2}$$

- The transverse mass, M_T, of the isolated lepton ℓ and the E_T^{miss} system, defined as:

$$M_T = \sqrt{2 p_T^\ell E_T^{\text{miss}} (1 - \cos\Delta\phi)} \,, \tag{3}$$

with $\Delta\phi$ the azimuth separation between the lepton and $\vec{p}_T^{\text{miss}}$ vector, and E_T^{miss} the absolute value of $\vec{p}_T^{\text{miss}}$.

- The number of selected muons (N_μ).
- The invariant mass of this set of muons (M_μ).
- The total transverse momentum of these muons ($p^\mu_{T,TOT}$).
- The number of selected electrons (N_e).
- The invariant mass of this set of electrons (M_e).
- The total transverse momentum of these electrons ($p^e_{T,TOT}$).
- The number of reconstructed charged hadrons.
- The number of reconstructed neutral hadrons.

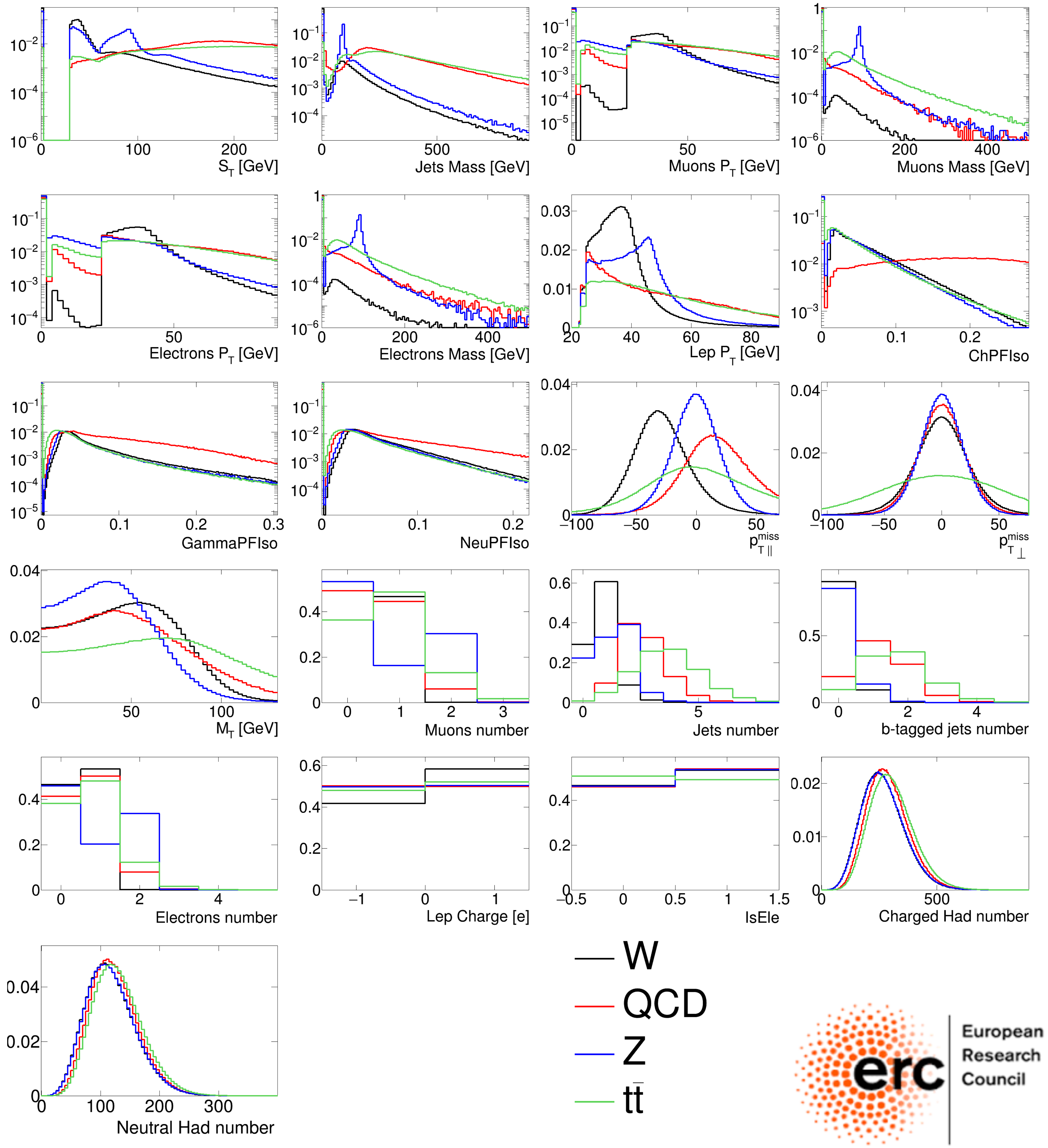
Europea
Researc
Council

Our use case: ℓ+X @HLT

- *Consider a stream of data coming from L1*
  - *Passed L1 because of 1 lepton (e,m) with pT>23 GeV*
  - *At HLT, very loose isolation applied*
  - *Sample mainly consists of W, Z, tt & QCD (for simplicity, we ignore the rest)*

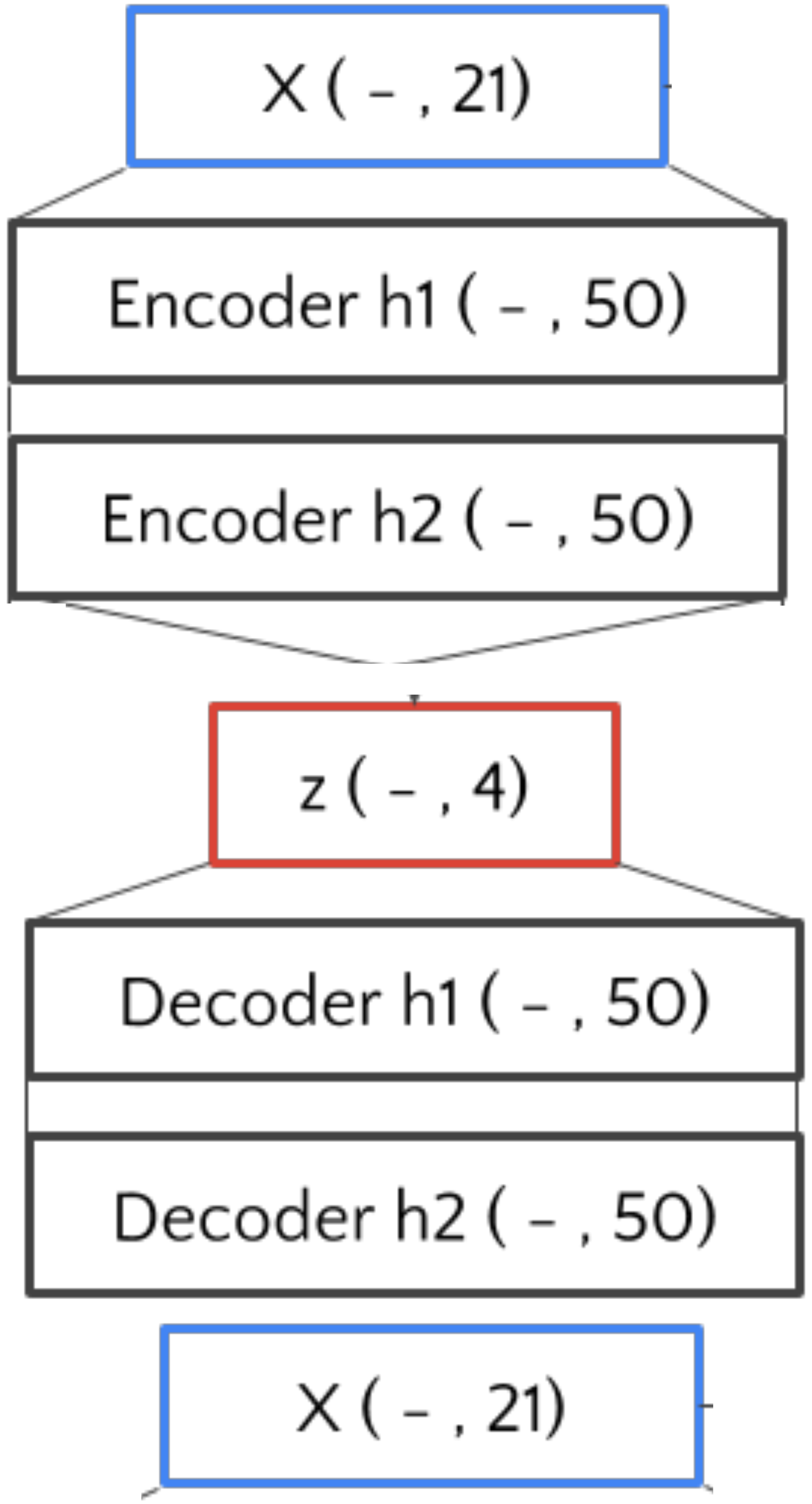
Standard Model processes					
Process	Acceptance	Trigger efficiency	Cross section [nb]	Events fraction	Event /month
W	55.6%	68%	58	59.2%	110M
QCD	0.08%	9.6%	$1.6 \cdot 10^5$	33.8%	63M
Z	16%	77%	20	6.7%	12M
$t\bar{t}$	37%	49%	0.7	0.3%	0.6M

  - *We consider 21 features, typically highlighting the difference between these SM processes (no specific BSM signal in mind)*



Standard Model AE

- *We train a VAE on a cocktail of SM events (weighted by xsec)*
- ***ENCODER:*** *21 inputs, 2 hidden layers → 4Dim latent space*
- ***DECODER:*** *from a random sample in the 4D space → 2 hidden layers → 21 outputs*

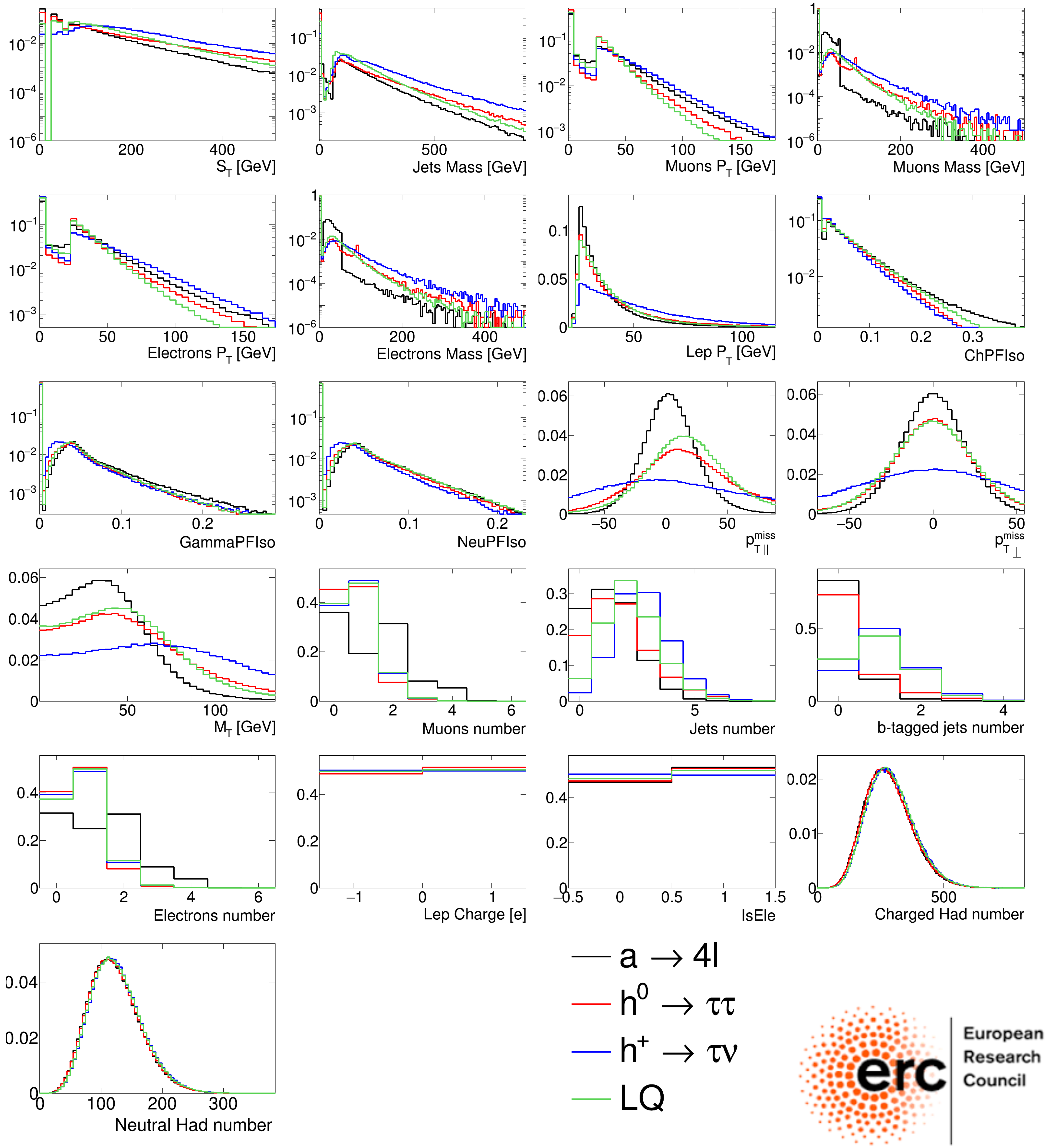


European Research Council

Some BSM benchmark

- We consider four BSM benchmark models, to give some sense of VAEs potential
  - leptoquark with mass 80 GeV, LQ→bτ
  - A scalar boson with mass 50 GeV, a→Z*Z*→4ℓ
  - A scalar scalar boson with mass 60 GeV, h→ττ
  - A charged scalar boson with mass 60 GeV, h±→τν

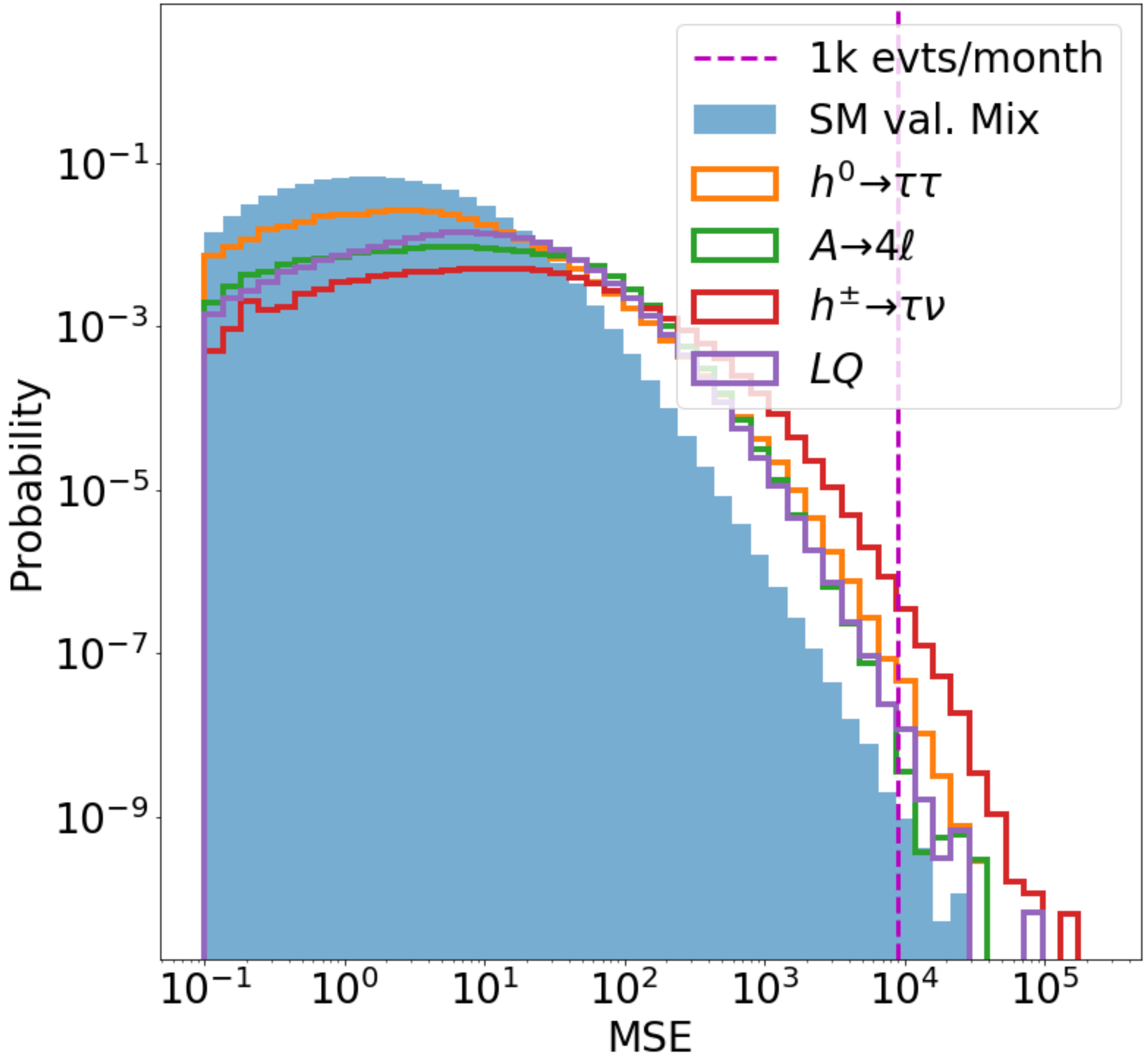
BSM benchmark processes				
Process	Acceptance	Trigger efficiency	Total efficiency	Cross-section 100 events/month
$h^0 \to \tau\tau$	9%	70%	6%	335 fb
$h^0 \to \tau\nu$	18%	69%	12%	163 fb
$LQ \to b\tau$	19%	62%	12%	166 fb
$a \to 4\ell$	5%	98%	5%	436 fb



European Research Council

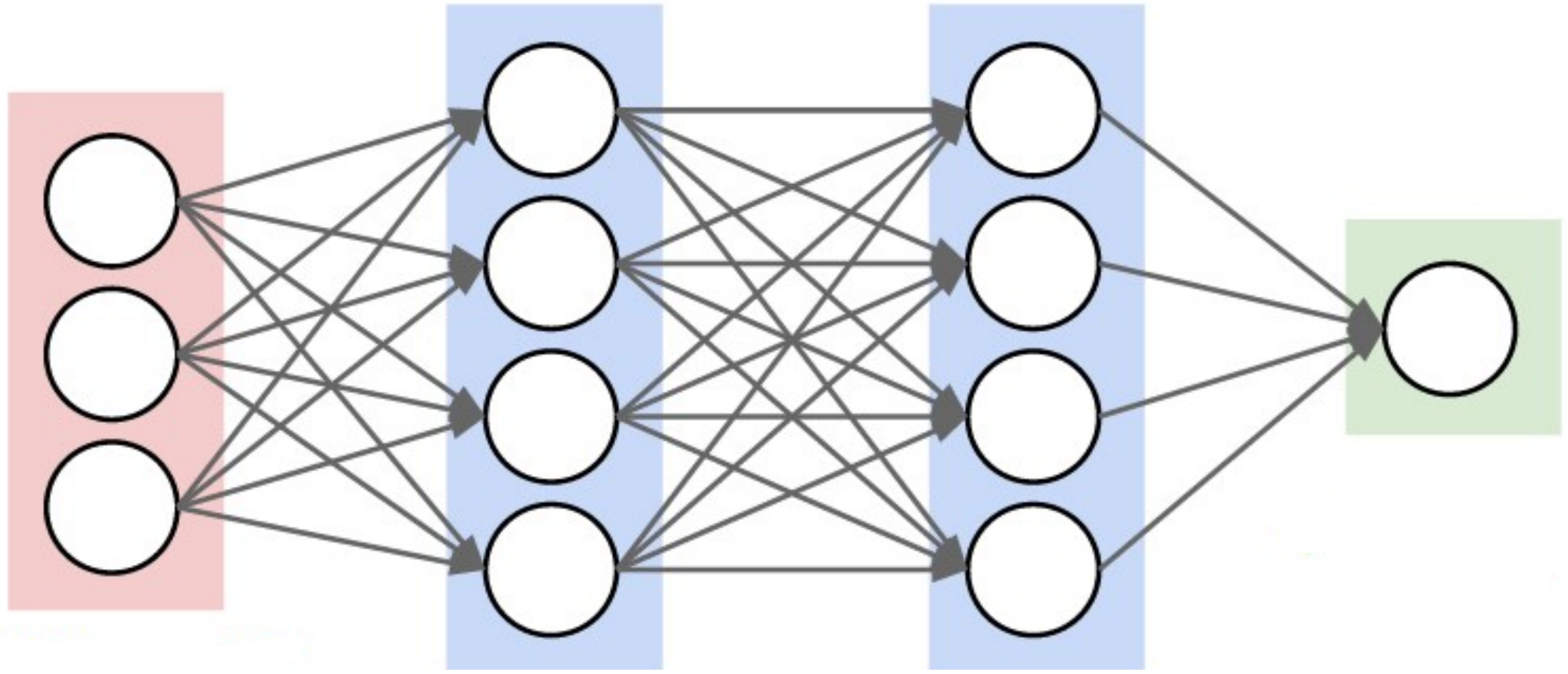
Defining anomaly

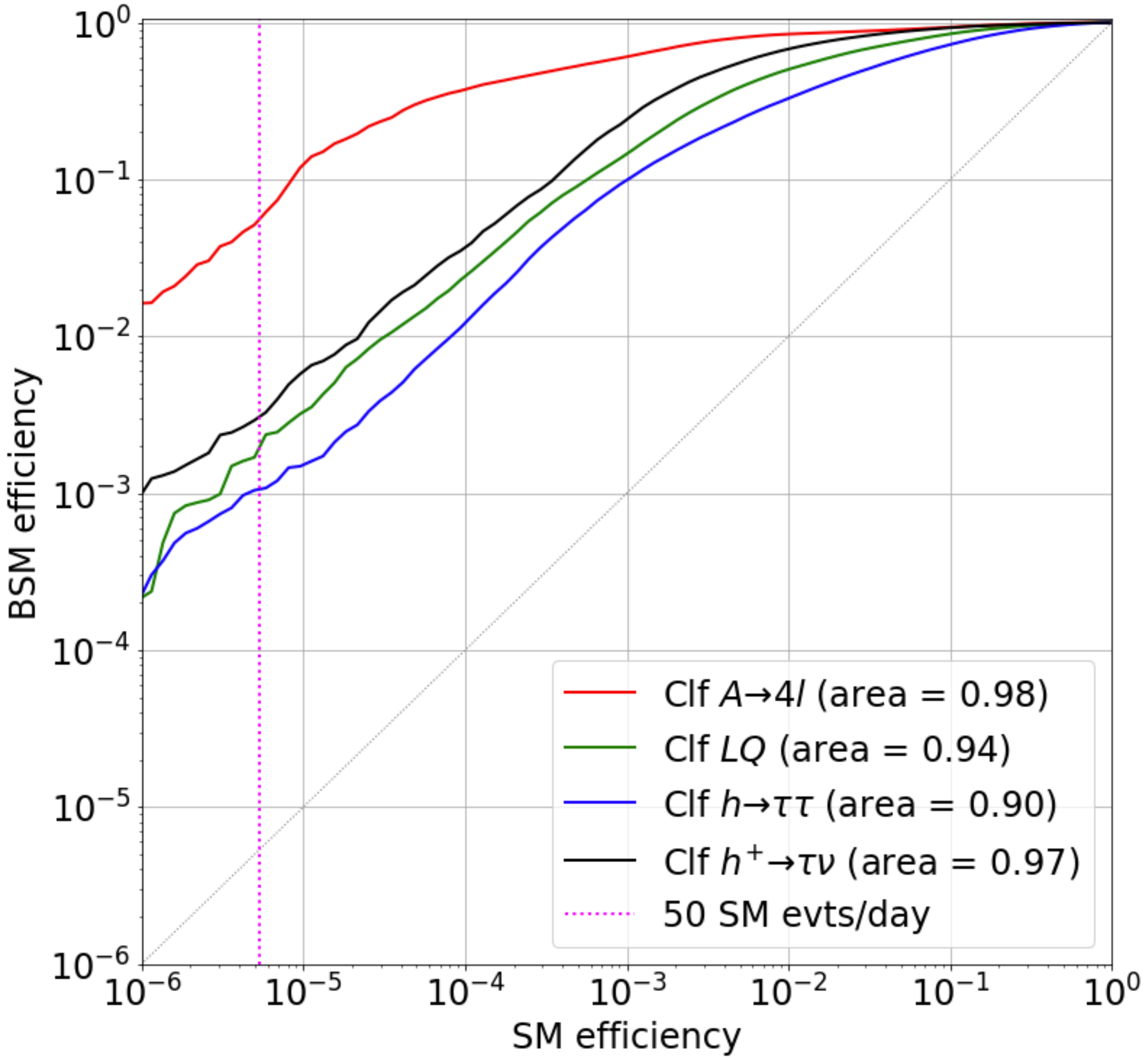
- Anomaly defined as a p-value threshold on a given test statistics
  - Loss function an obvious choice
  - Some part of a loss could be more sensitive than others
  - We tested different options and found the total loss to behave better



Benchmark comparison

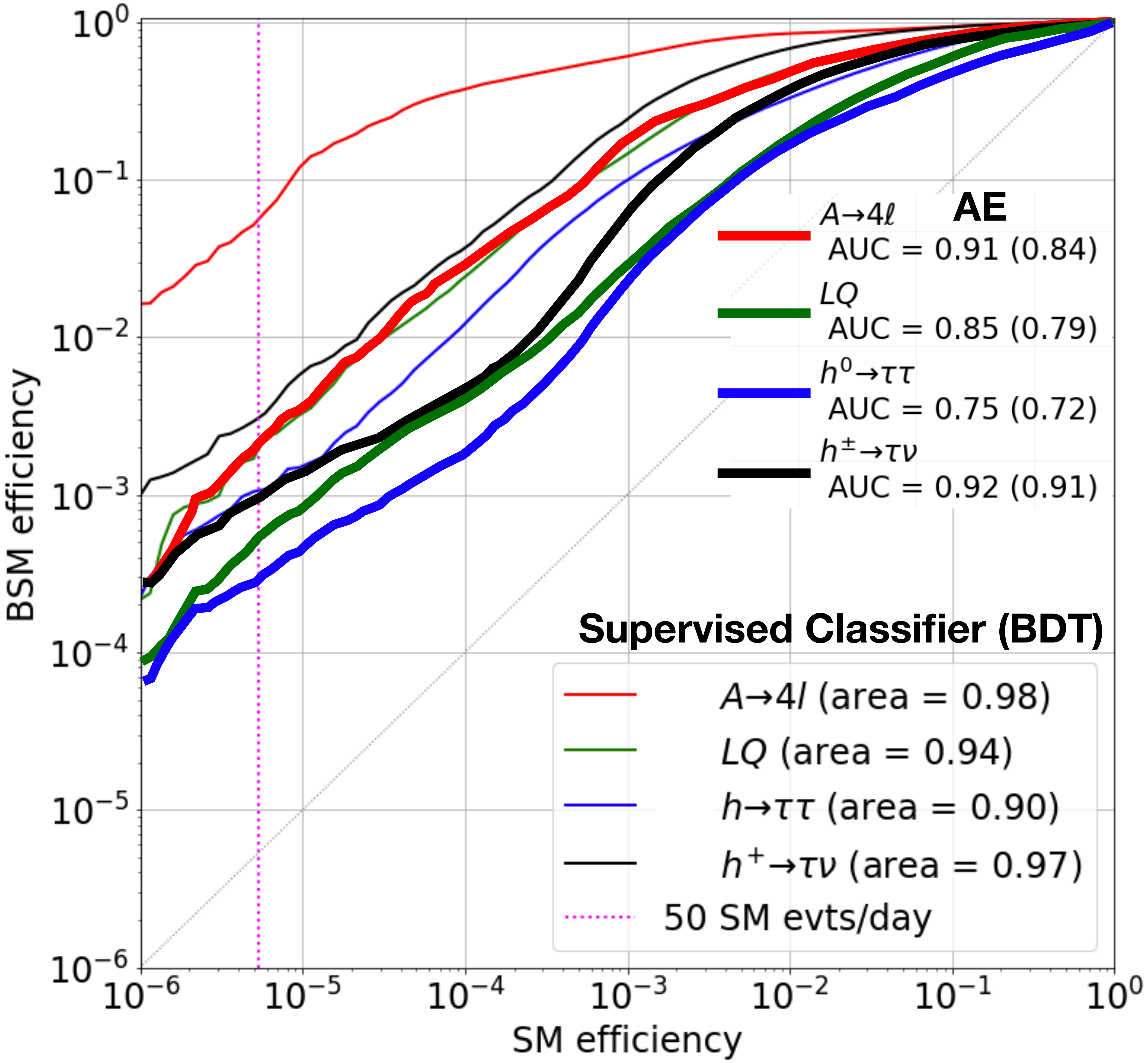
- *VAE's performances benchmarked against supervised classifiers*
- *For each BSM model*
  - *take same inputs as VAE*
  - *train a fully-supervised classifier to separate signal from background*
  - *use supervised performances as a reference to aim to with the unsupervised approach*
  - *Done for our 4 BSM models using dense neural networks*

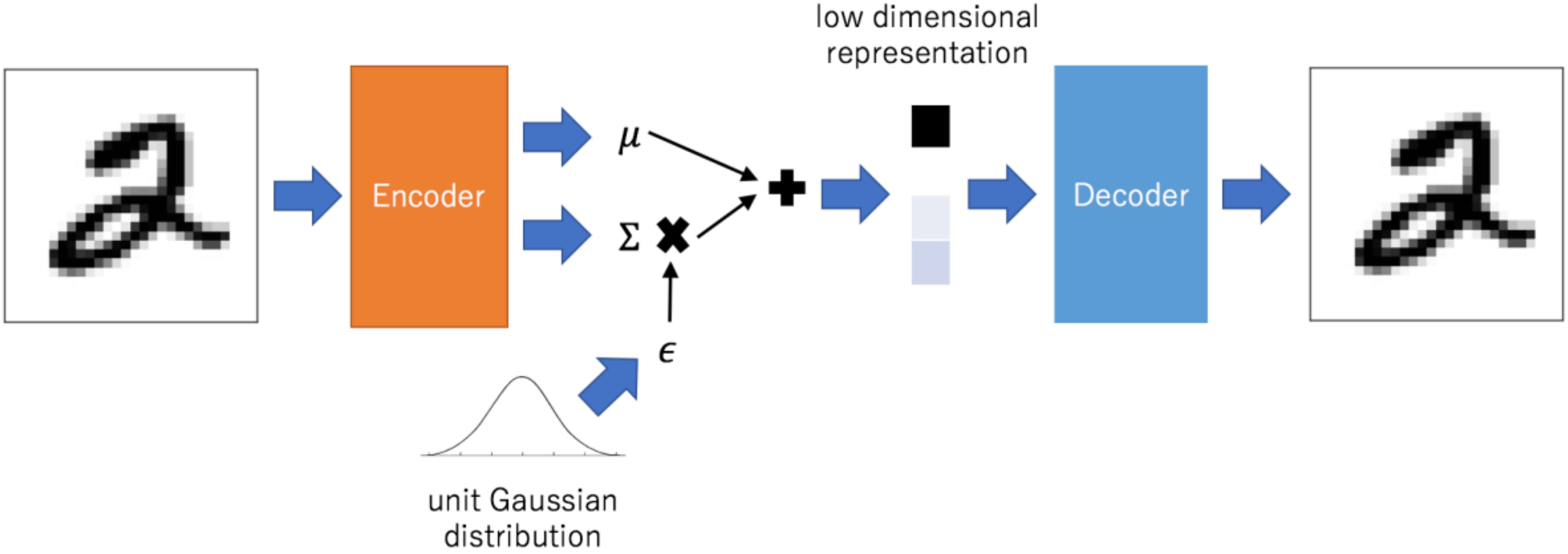




Performances

- Evaluate general discrimination power by ROC curve and area under curve (AUC)
  - clearly worse than supervised
  - but not so far
- Fixing SM acceptance rate at 50 events/day
  - competitive results considering unsupervised nature of the algorithm

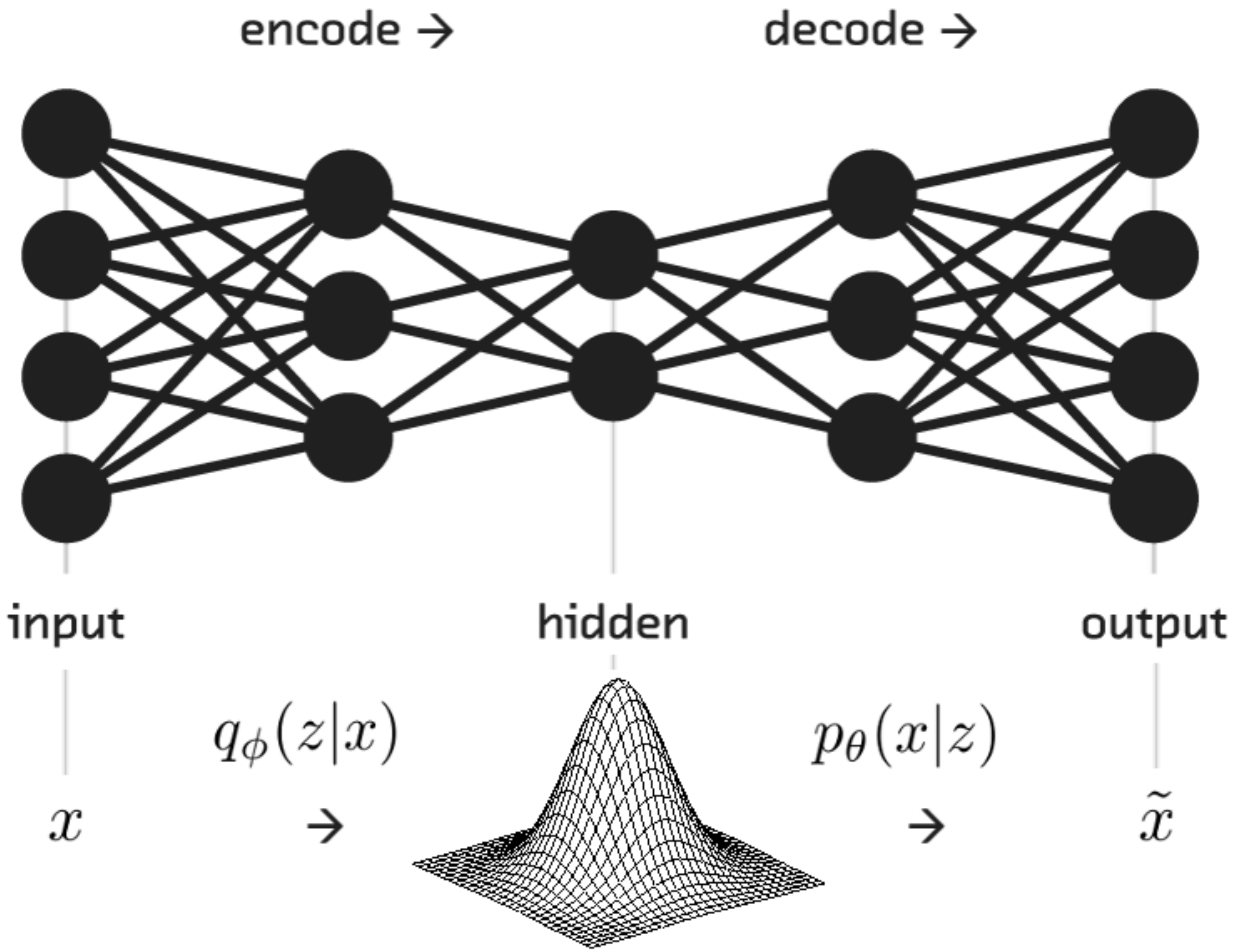




Variational Autoencoders

Variational Autoencoders

- *We investigated variational autoencoders*
- *Unlike traditional AEs, VAEs try to associate a multi-Dim pdf to a given image*
  - *can be used to generate new examples*
  - *comes with a probabilistic description of the input*
  - *tends to work better than traditional AEs*



The Loss Function

- *Loss function described as the sum of two terms (scaled by a tuned λ parameter that makes the two contribution numerically similar)*
  - *Reconstruction loss (e.g. MSE(output-input))*
  - *KL loss: distance between Gaussian pdfs (assumption on prior here)*
  - *Why Gaussian? KL loss can be written analytically*

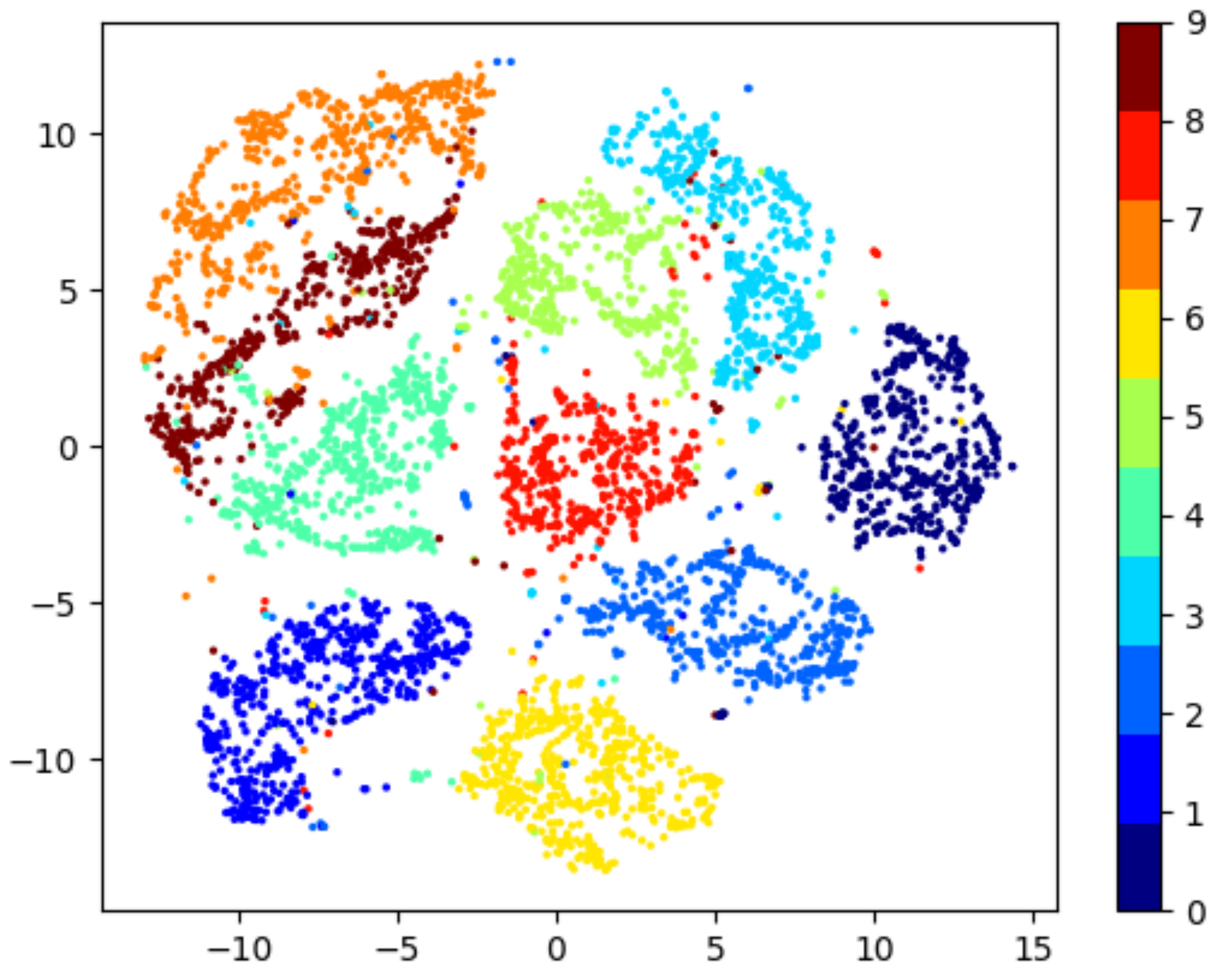
$$\mathrm{Loss}_{\mathrm{Tot}} = \mathrm{Loss}_{\mathrm{reco}} + \beta D_{\mathrm{KL}}$$

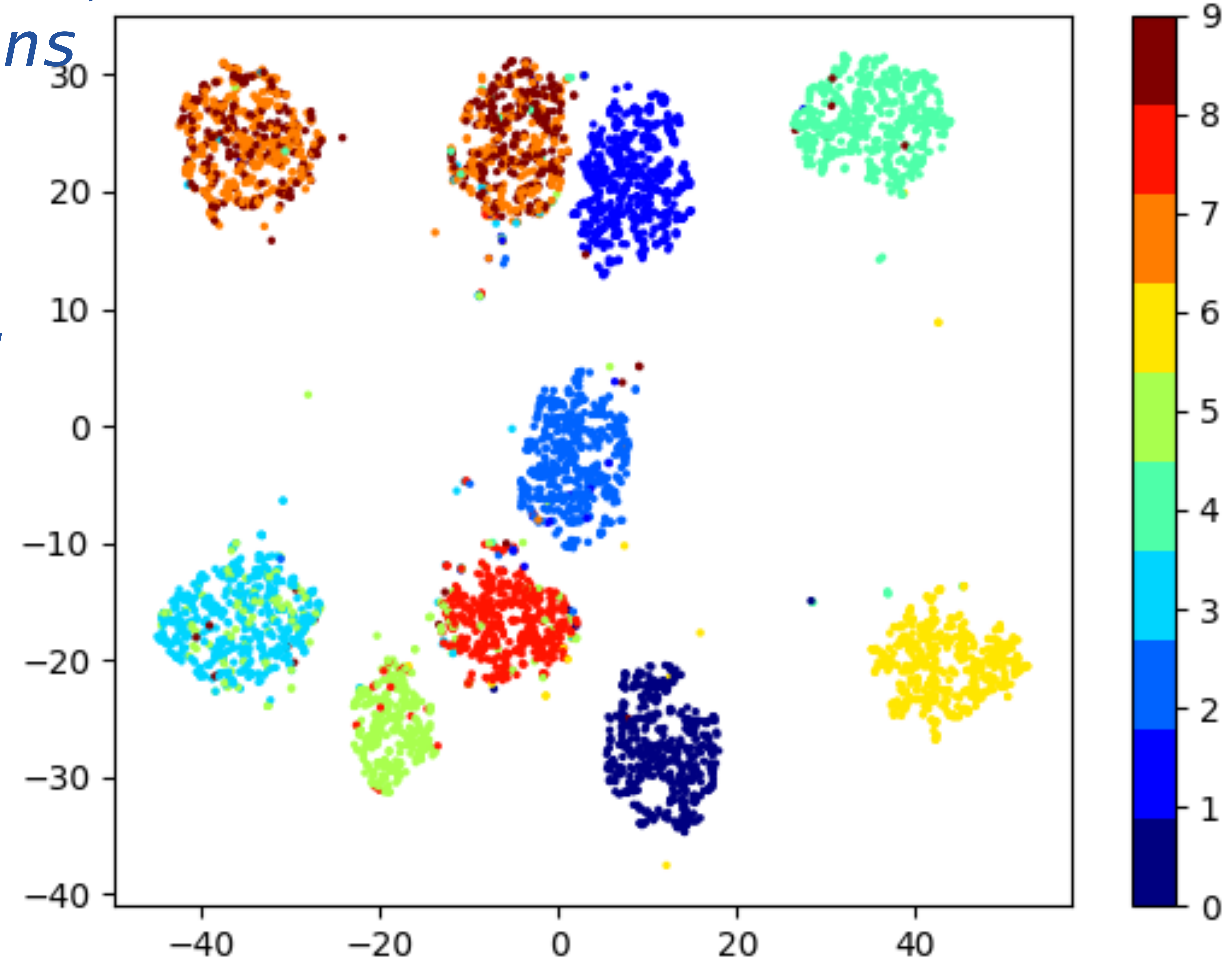
$$D_{\mathrm{KL}} = \frac{1}{k}\sum_i D_{\mathrm{KL}}\left(N(\mu_z^i, \sigma_z^i) \;||\; N(\mu_P, \sigma_P)\right)$$

$$= \frac{1}{2k}\sum_{i,j}\left(\sigma_P^j \sigma_z^{i,j}\right)^2 + \left(\frac{\mu_P^j - \mu_z^{i,j}}{\sigma_P^j}\right)^2 + \ln\frac{\sigma_P^j}{\sigma_z^{i,j}} - 1$$

Clustering with VAE

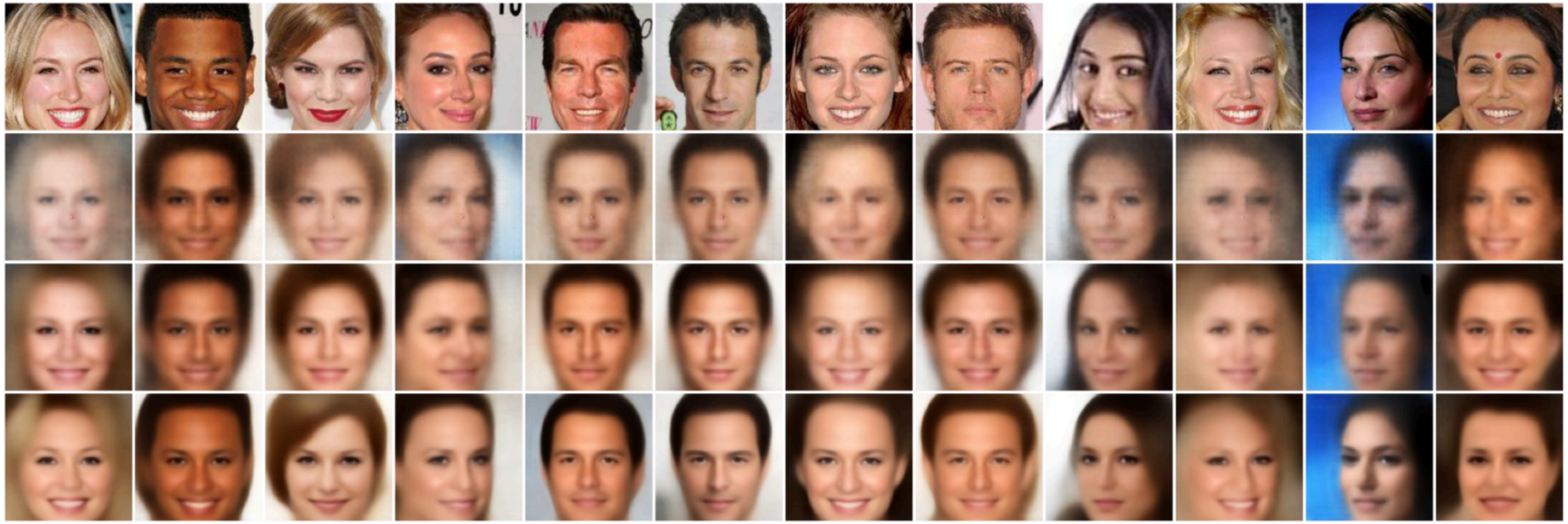
- In the clustering example, the different populations are forced on sums of Gaussian distributions
- This gives more regular shape for the clusters



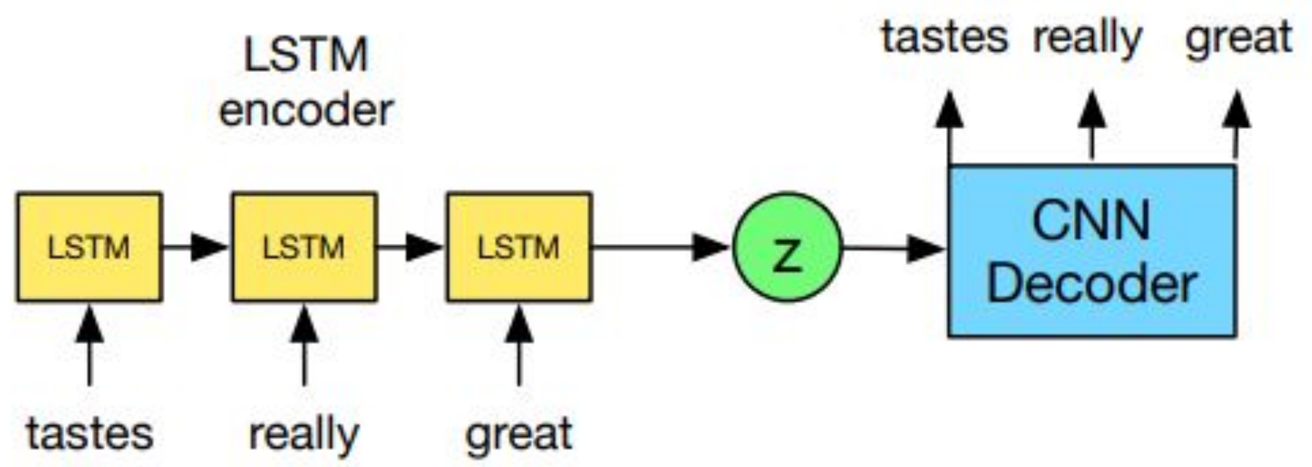


A Generative model

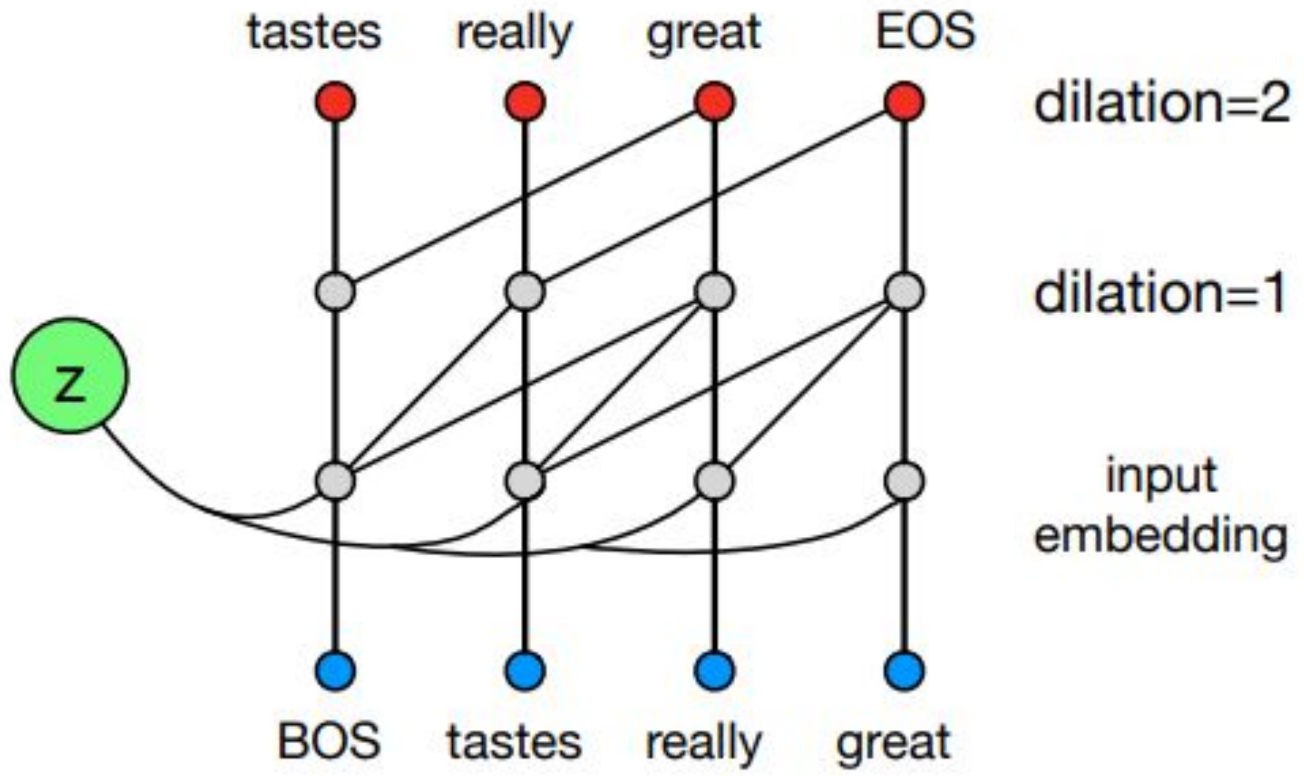
- *Now that we have a probabilistic description of the latent space, we can sample points from it*
- *These points, propagated through the decoder, will provide new examples*
- *We have defined a generative model*



More effective with sequential data

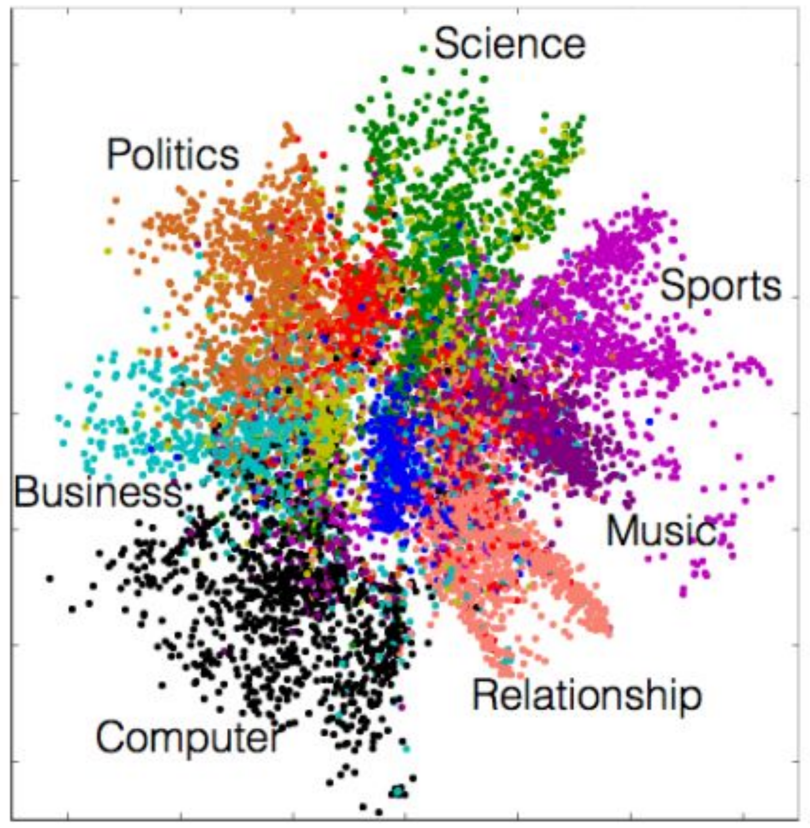


(a) VAE training graph using a dilated CNN decoder.

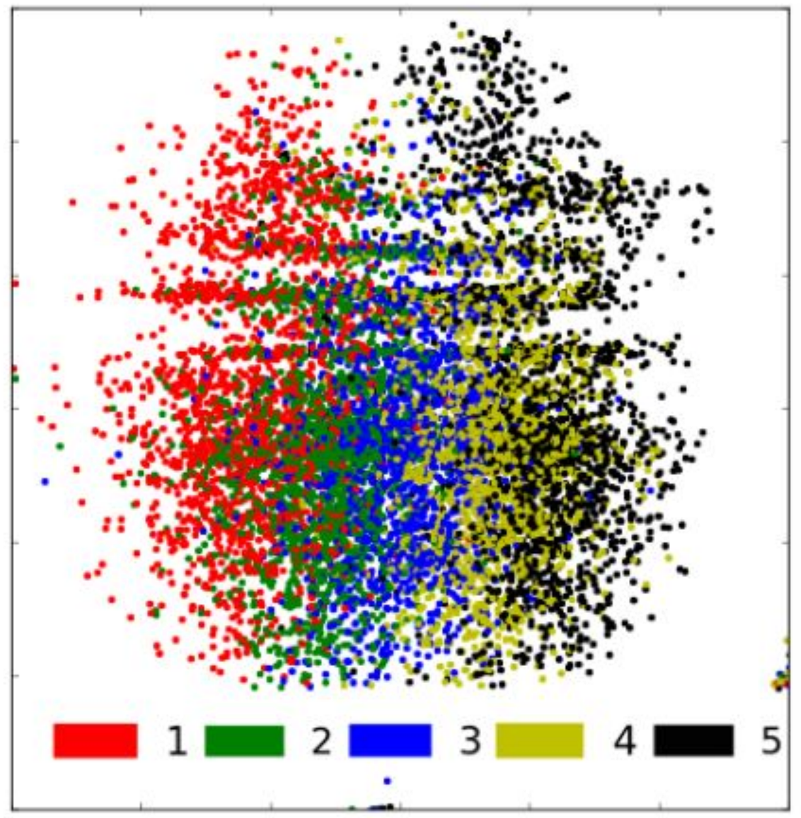


(b) Digram of dilated CNN decoder.

1 star	the food was good but the service was horrible . took forever to get our food . we had to ask twice for our check after we got our food . will not return .
2 star	the food was good , but the service was terrible . took forever to get someone to take our drink order . had to ask 3 times to get the check . food was ok , nothing to write about .
3 star	came here for the first time last night . food was good . service was a little slow . food was just ok .
4 star	food was good , service was a little slow , but the food was pretty good . i had the grilled chicken sandwich and it was really good . will definitely be back !
5 star	food was very good , service was fast and friendly . food was very good as well . will be back !



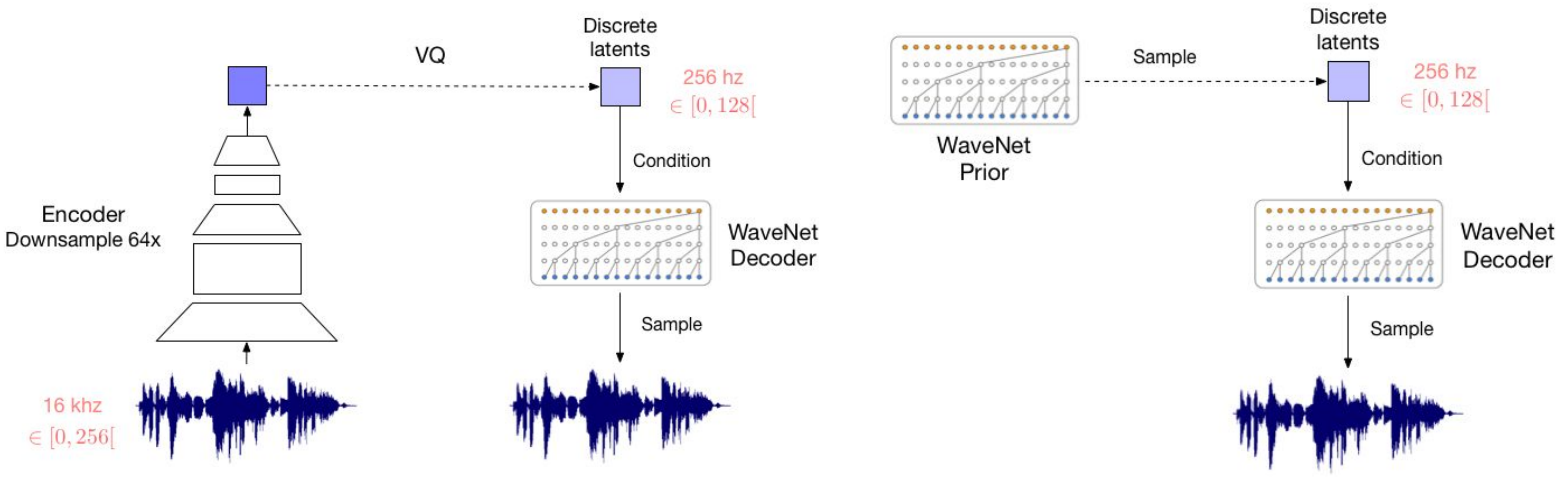
(a) Yahoo



(b) Yelp

Yang, Z., Hu, Z., Salakhutdinov, R., & Berg-Kirkpatrick, T. (2017). Improved variational autoencoders for text modeling using dilated convolutions. *ICML 2017*

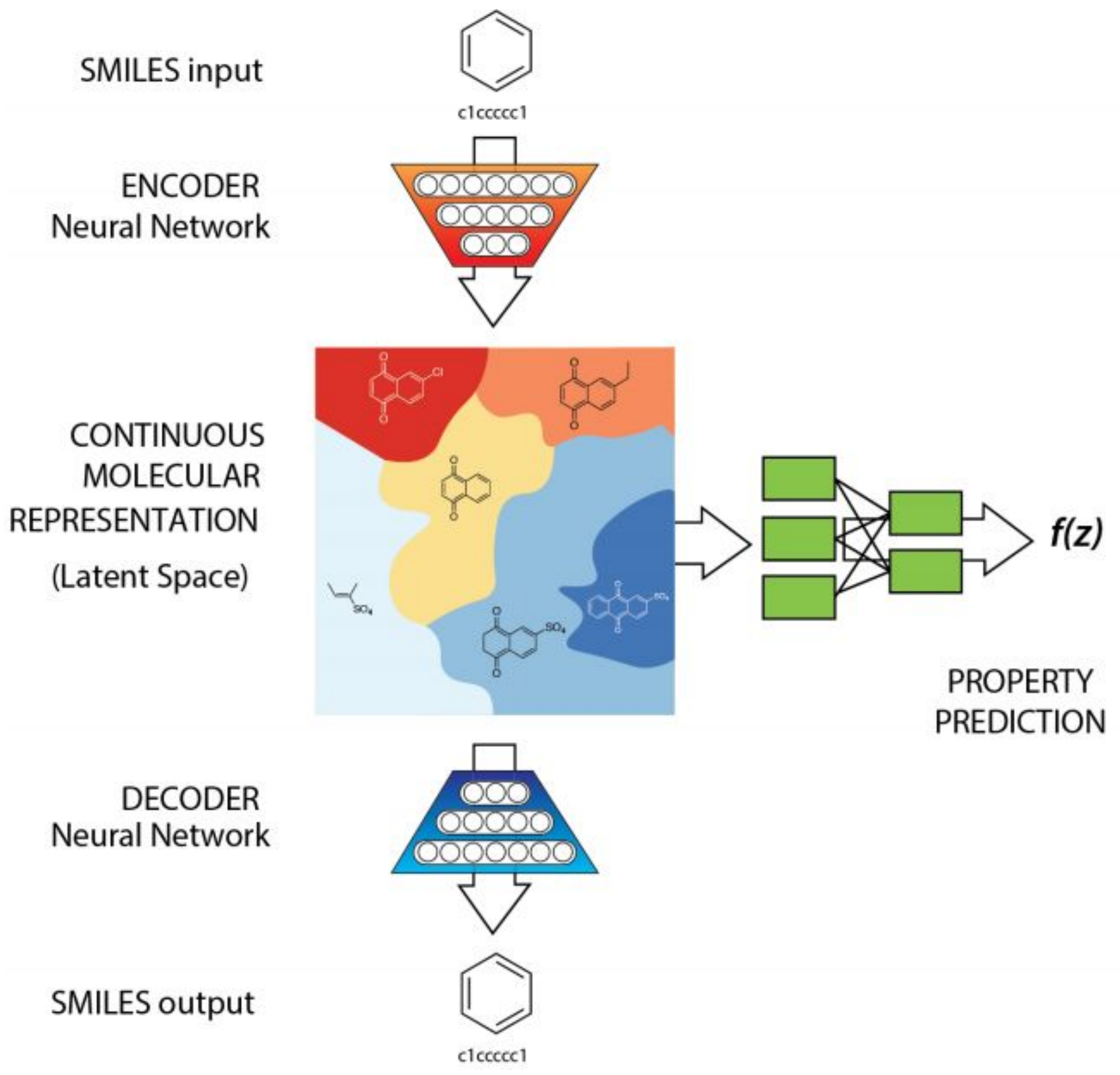
More effective with sequential data

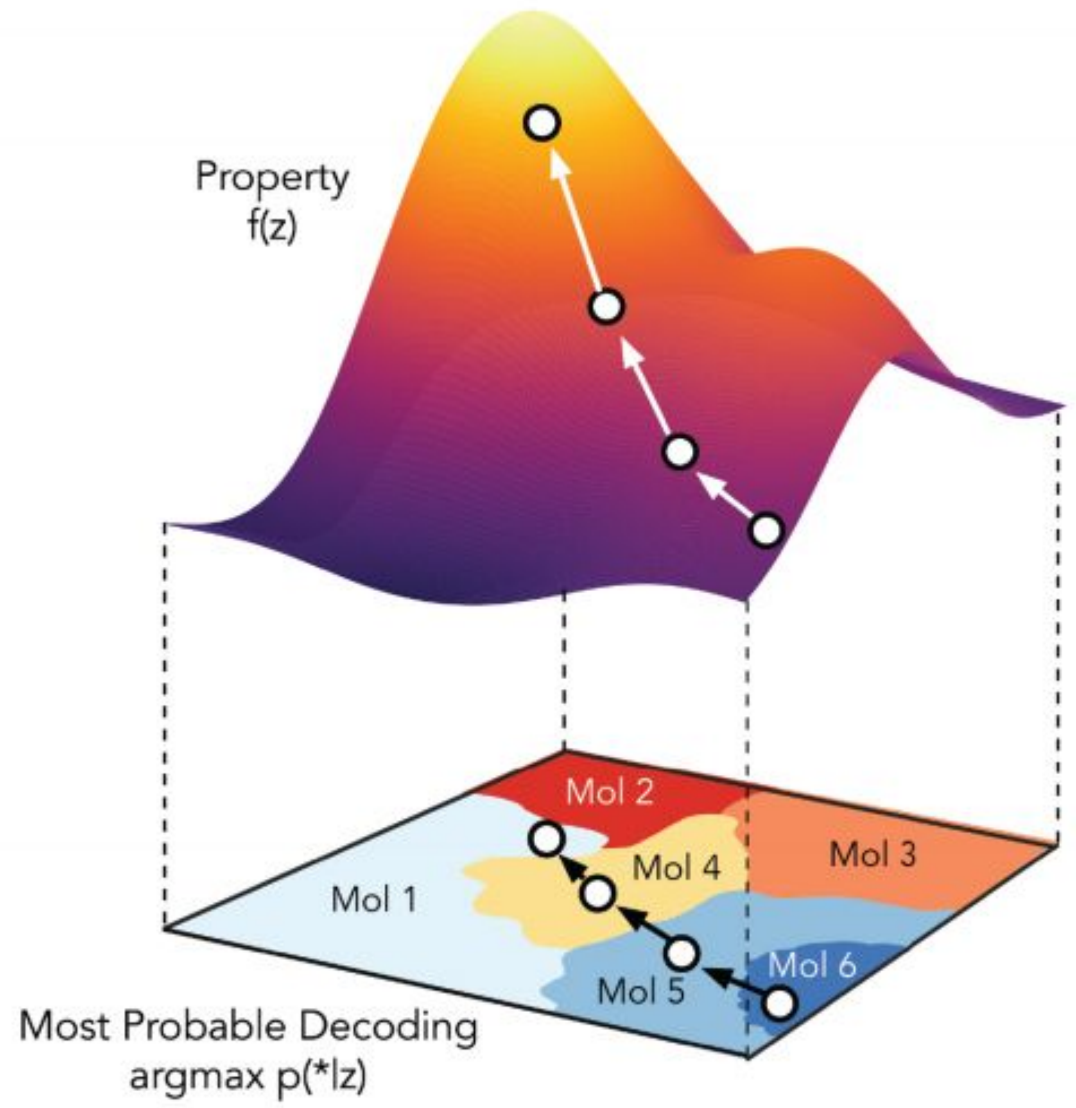


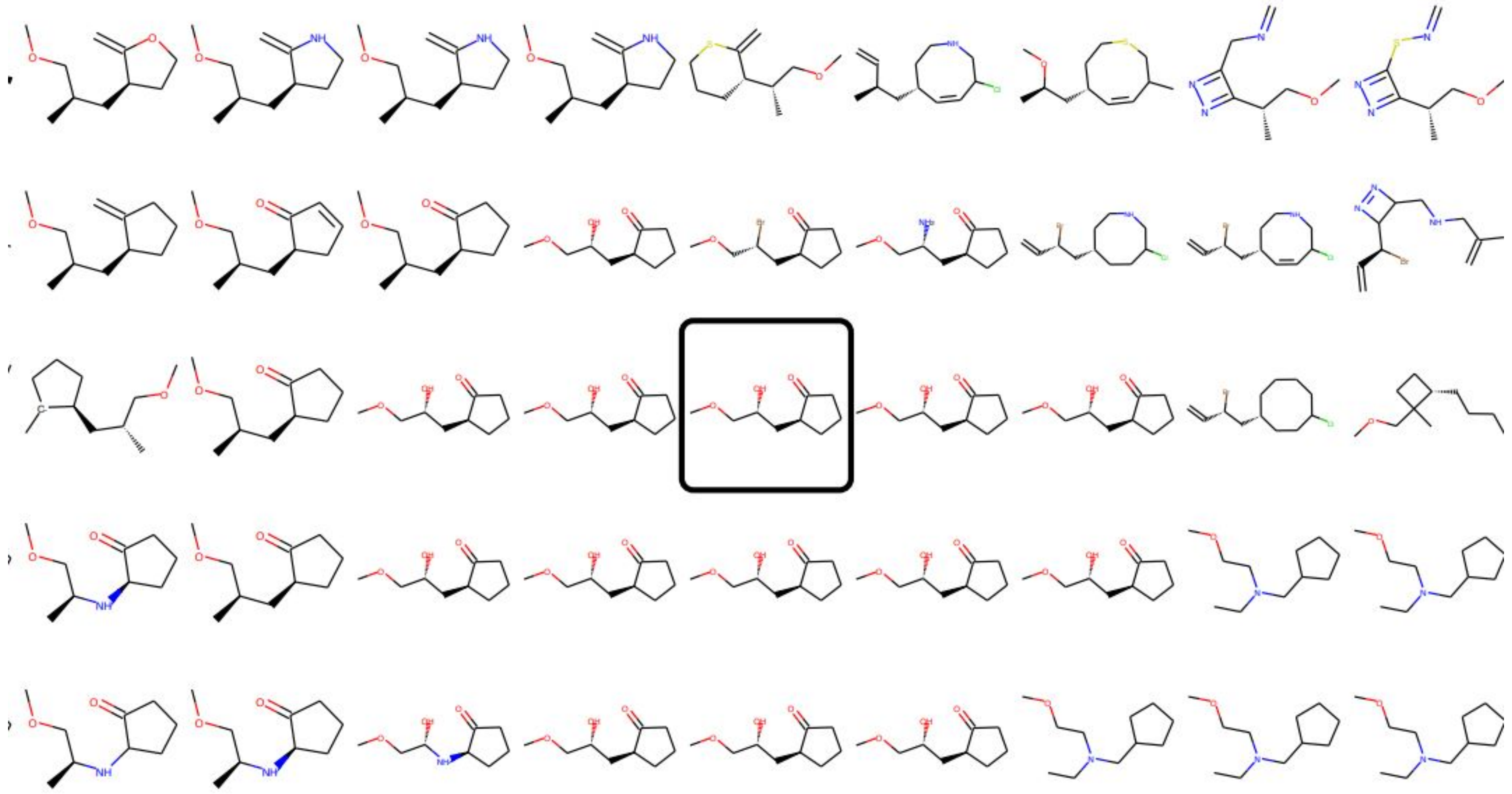
reconstruction **generation**

van den Oord, A., & Vinyals, O. (2017). Neural discrete representation learning. *NIPS 2017*.

More effective with sequential data

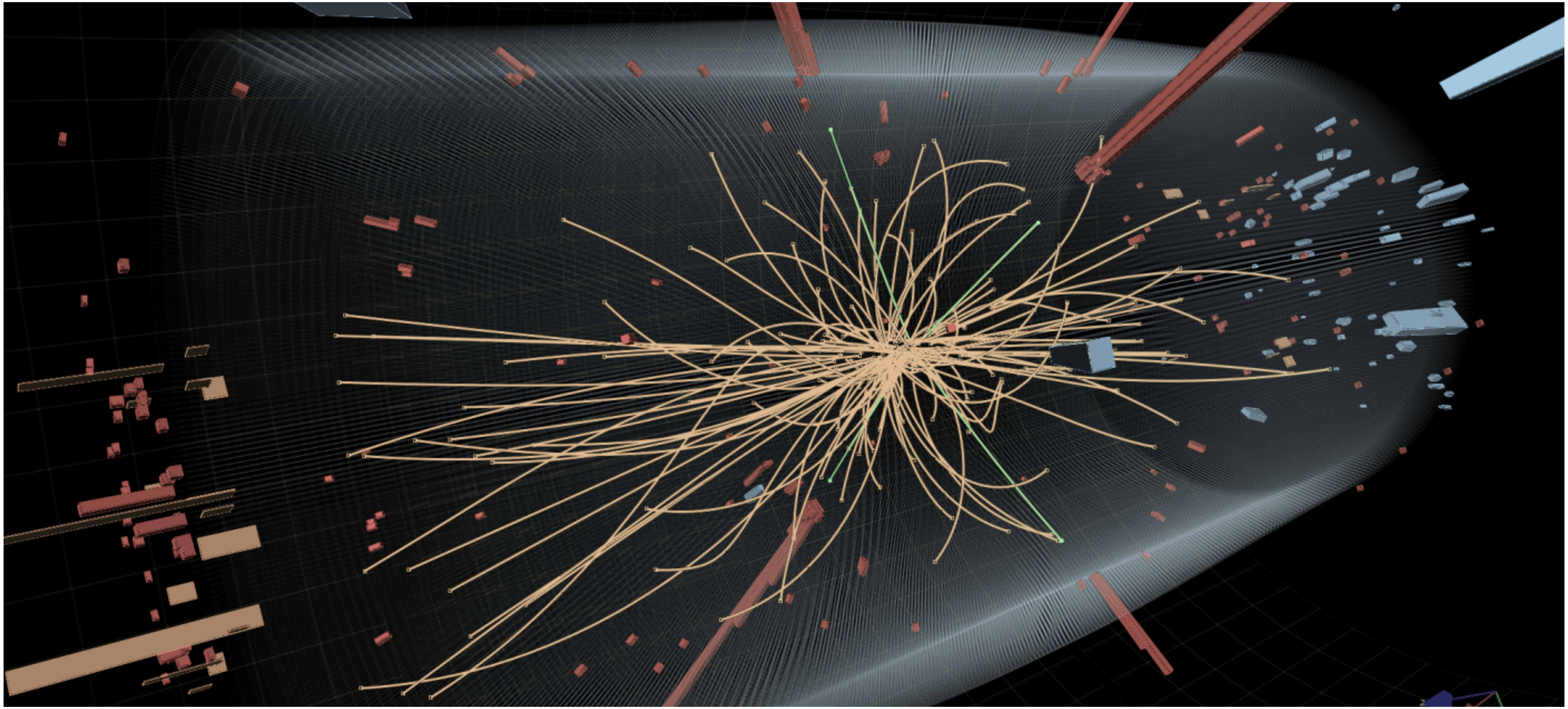






Gómez-Bombarelli, R., et al. (2018). Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules ACS Cent.

Kusner, M. J., Paige, B., & Hernández-Lobato, J. M. (2017). Grammar variational autoencoder. *arXiv preprint arXiv:1703.01925*.

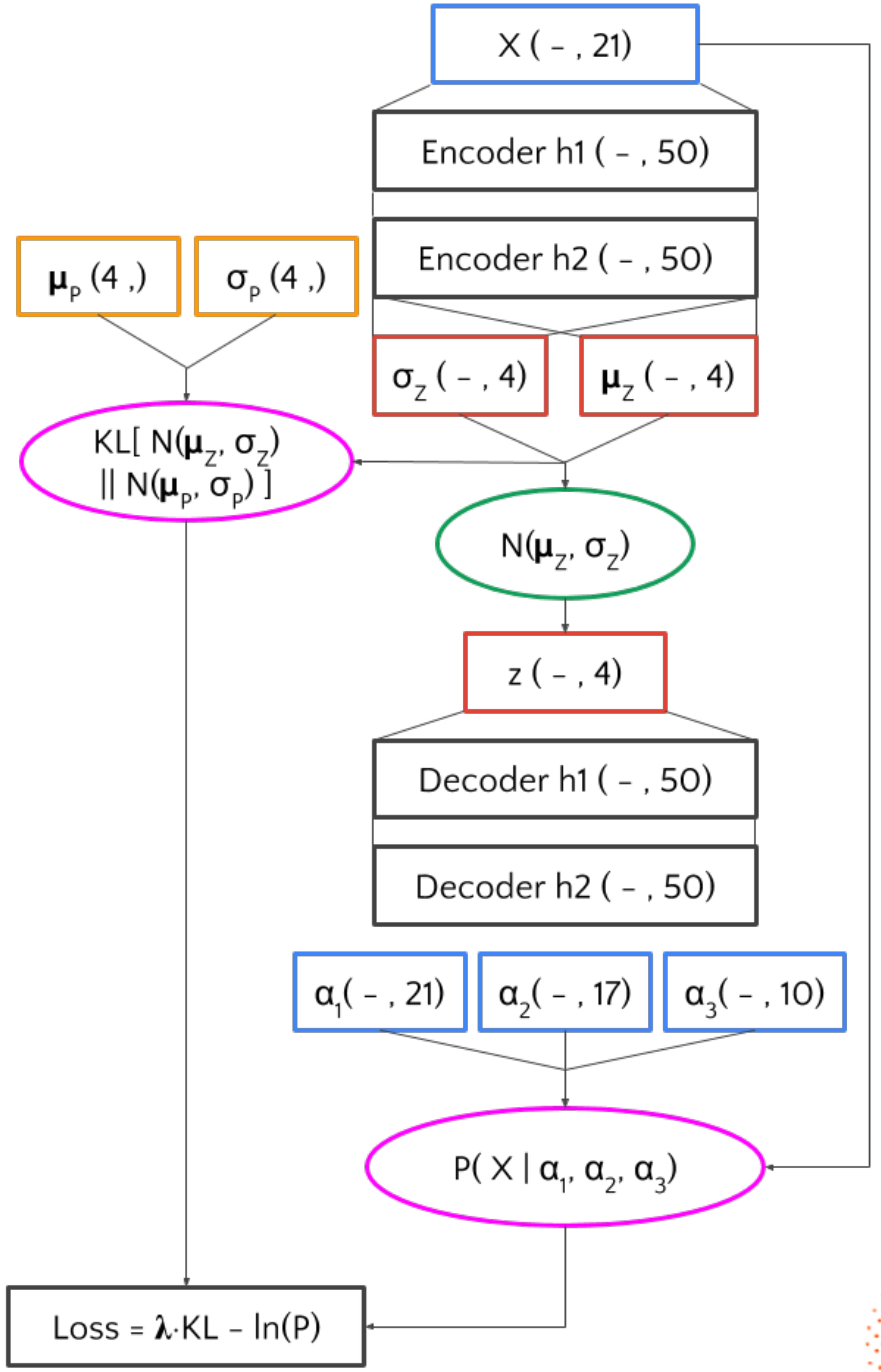


Variational Autoencoders for particle physics

European Research Council

Back to our example

- *We train a VAE on a cocktail of SM events (weighted by xsec)*
- ***ENCODER:*** *21 inputs, 2 hidden layers → 4Dim latent space*
  - *hidden nodes = μ and σ of the Gaussian pdfs describing the hidden variables*
- ***DECODER:*** *from a random sample in the 4D space → 2 hidden layers → parameters describing the shape of the 21Dim input space*



European Research Council

The Loss Function

- *Loss function described as the sum of two terms (scaled by a tuned λ parameter that makes the two contribution numerically similar)*
  - *Reconstruction loss: likelihood of the input 21Dim point, given the shape parameters reconstructed from it*
  - *KL loss: distance between the pdf in the latent space and an nDim Gaussian*

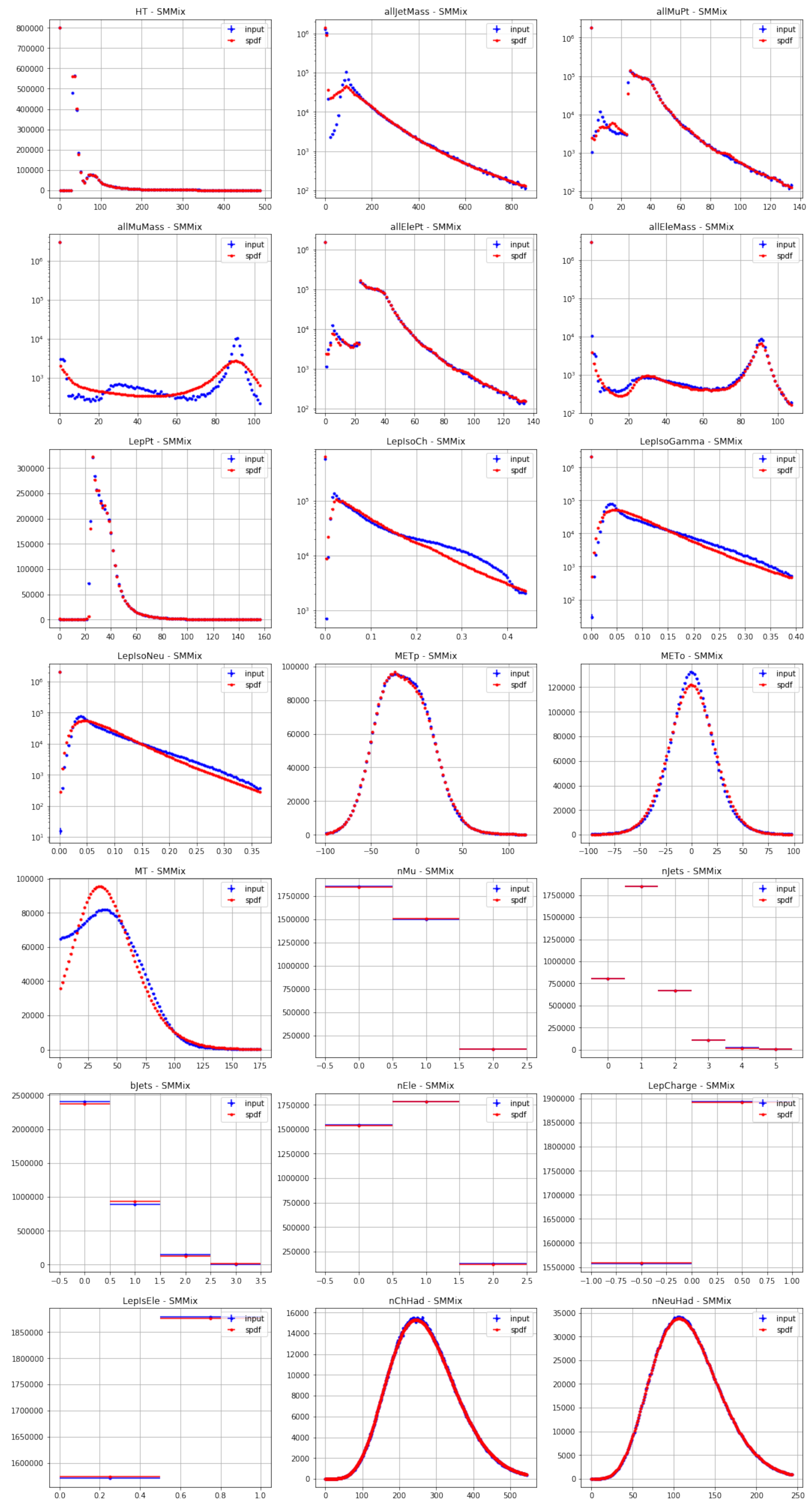
$$\mathrm{Loss}_{\mathrm{Tot}} = \mathrm{Loss}_{\mathrm{reco}} + \beta D_{\mathrm{KL}}$$

$$\begin{aligned}\mathrm{Loss}_{\mathrm{reco}} &= -\frac{1}{k}\sum_i \ln\left(P(x \mid \alpha_1, \alpha_2, \alpha_3)\right) \\ &= -\frac{1}{k}\sum_{i,j} \ln\left(f_j(x_{i,j} \mid \alpha_1^{i,j}, \alpha_2^{i,j}, \alpha_3^{i,j})\right)\end{aligned}$$

$$\begin{aligned}D_{\mathrm{KL}} &= \frac{1}{k}\sum_i D_{\mathrm{KL}}\left(N(\mu_z^i, \sigma_z^i) \;||\; N(\mu_P, \sigma_P)\right) \\ &= \frac{1}{2k}\sum_{i,j}\left(\sigma_P^j \sigma_z^{i,j}\right)^2 + \left(\frac{\mu_P^j - \mu_z^{i,j}}{\sigma_P^j}\right)^2 + \ln\frac{\sigma_P^j}{\sigma_z^{i,j}} - 1\end{aligned}$$

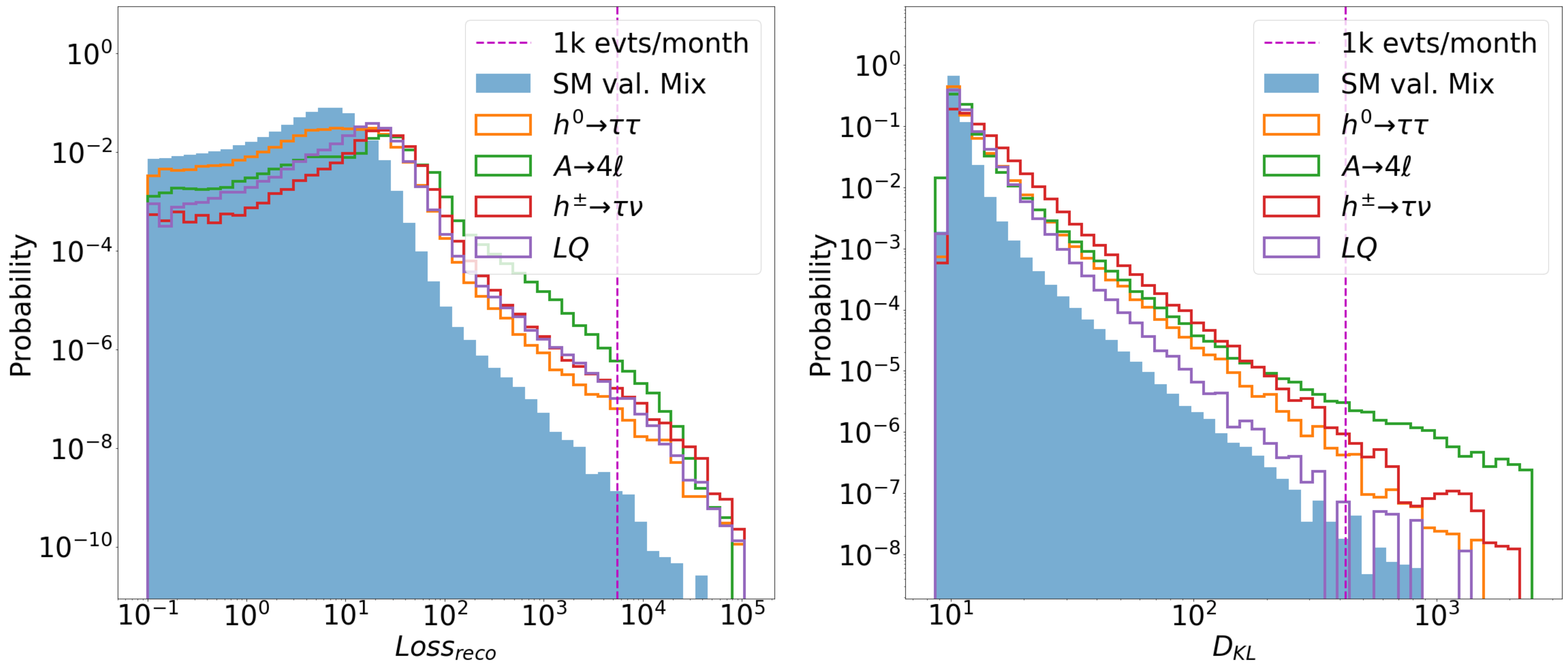
Standard Model encoding

- First post-training check consists in verifying encoding-decoding capability, comparing input data to those generated sampling from decoder
- Reasonable agreement observed, with small discrepancy here and there
- NOTICE THAT: this would be a suboptimal event generator, but we want to use it for anomaly detection
  - no guarantee that the best autoencoder is the best anomaly detector (no anomaly detection rate in the loss function)
  - pros & cons of an unsupervised/ semisupervised approach



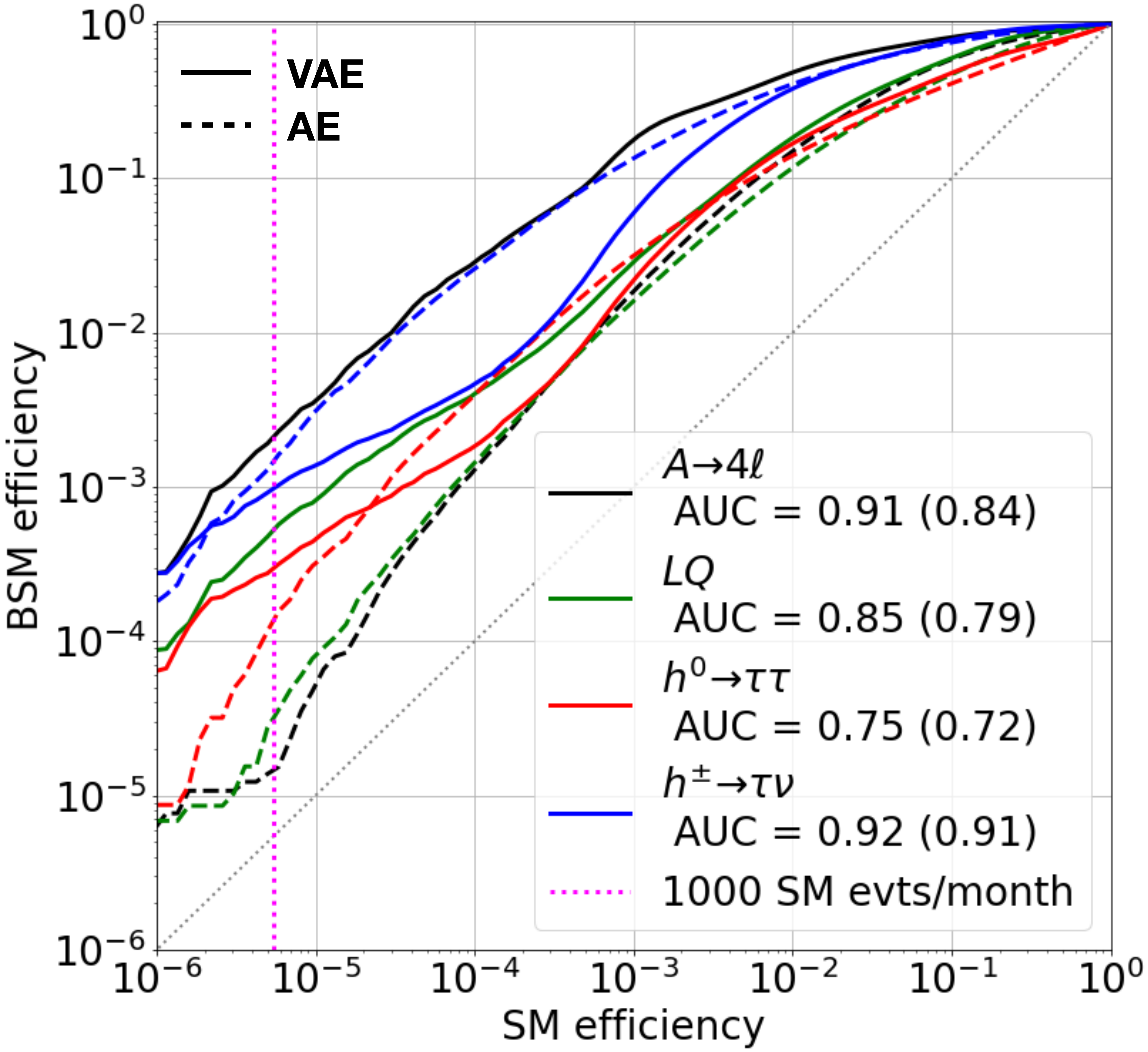
Defining anomaly

- Anomaly defined as a p-value threshold on a given test statistics
  - Loss function an obvious choice
  - Some part of a loss could be more sensitive than others
  - We tested different options and found the total loss to behave better



Performances

- Evaluate general discrimination power by ROC curve and area under curve (AUC)
  - clearly worse than supervised
  - but not so far
- Fixing SM acceptance rate at 50 events/day
  - competitive results considering unsupervised nature of the algorithm



Performances

- *Small efficiency but still much larger than for SM processes*

- *Allows to probe 10-100 pb cross sections for reasonable amount of collected signal events*

Process	Efficiency for ~30 evt/day	xsec for 100 evt/ month [pb]	xsec for S/B~1/3 [pb]
$a \to 4\ell$	$2.8 \cdot 10^{-3}$	7.1	27
$LQ \to \tau b$	$6.5 \cdot 10^{-4}$	31	120
$h \to \tau\tau$	$3.6 \cdot 10^{-4}$	56	220
$h^{\pm} \to \tau\nu$	$1.2 \cdot 10^{-3}$	17	67

1/2 way to model independence

- Procedure designed to be model independent
  - Training done only on SM
  - Algorithm that defines anomaly tuned only on number of selected SM events (false positive rate)
- Still, residual model dependence present
  - Based on physics-motivated observables
  - List not tailored on specific models and general enough to offer good performances in principle
  - But one cannot prove that performances on specific BSM models will generalise
- Can we go beyond this limitation and define something really BSM agnostic?

Summary

- Autoencoders are NNs for unsupervised problems
  - Clustering
  - Dimensional reduction
  - Anomaly detection
- When adding variational functionality
  - Can be used as generators
  - Can improve robustness (e.g., anomaly detection performance)
- Could be relevant to reduce model dependence in searches for new physics at the LHC

European Research Council