arXiv:1811.10276

ACAT 2019, Saas Fee (Ch), March 14th 2019

Variational Autoencoders for New Physics Mining at the LHC

Olmo Cerri [a], T. Q. Nguyen [a], M. Pierini [b], M. Spiropulu [a], and J. R. Vlimant [a]

[a] California Institute of Technology, [b] CERN

Model-independent tagger for unexpected events

Save events that does not come from SM processes, despite their nature or particular features

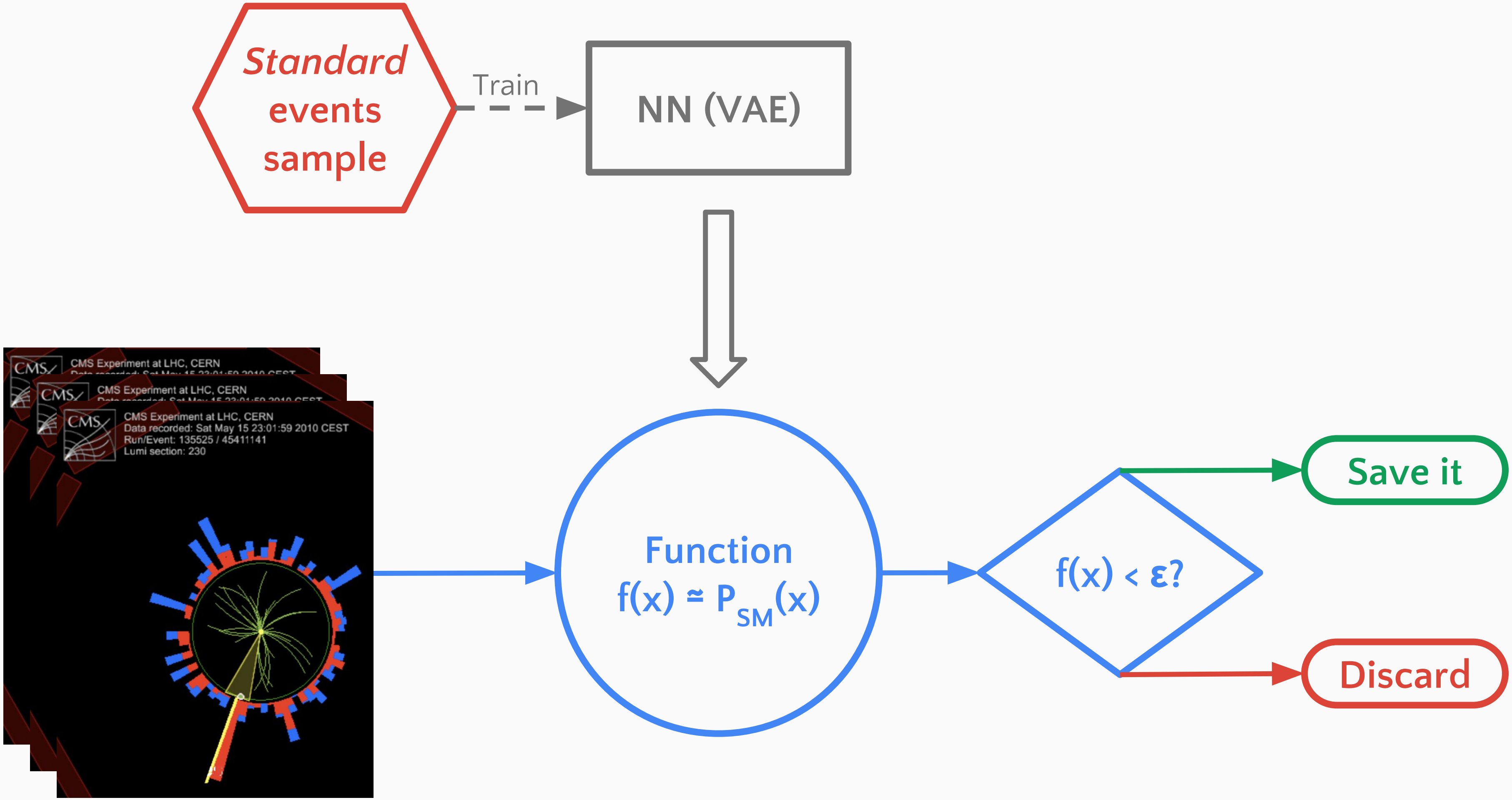
1. Set the stage
2. Results overview
3. How it works
4. Performances

Physics anomaly detection

- Data mining concept
  - Often: PCA, AE
- Based on Variational Auto-Encoders [1]

1. Define what is "standard" through a set of example events
   - The Standard Model
2. Fit a function which gives the p-value of belonging to the standard set
   - No assumption on the anomaly
     - Completely agnostic on BSM
3. Use this function to tag new events
   - Anomaly: low probability of belonging to the standard set
   - SM tails or BSM

[1]: https://arxiv.org/abs/1312.6114


Standard events sample
Train
NN (VAE)
CMS Experiment at LHC, CERN
Data recorded: Sat May 15 23:01:59 2010 CEST
Run/Event: 135525 / 45411141
Lumi section: 230
Function
f(x) ≃ P_SM(x)
f(x) < ε?
Save it
Discard

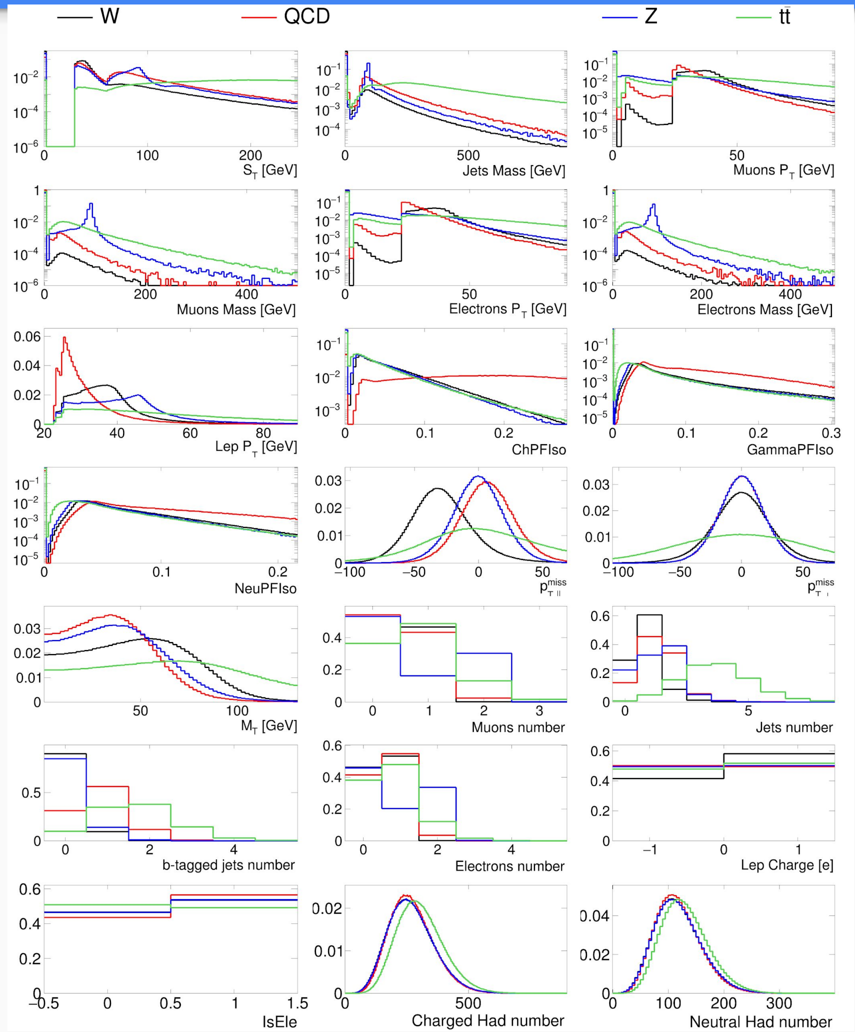
Workflow

A use case: ℓ+X

- Stream of data with at least one interesting lepton (e or μ)
  - p_T > 23 GeV & ISO < 0.45
- SM contribution:

Process	Event fraction in the stream	Events/month
W	59%	110M
QCD	34%	63M
Z	6.7%	12M
tt	0.3%	0.6M

- Events represented by 21 high level features (HLF)
  - Broad general choice, not BSM tailored

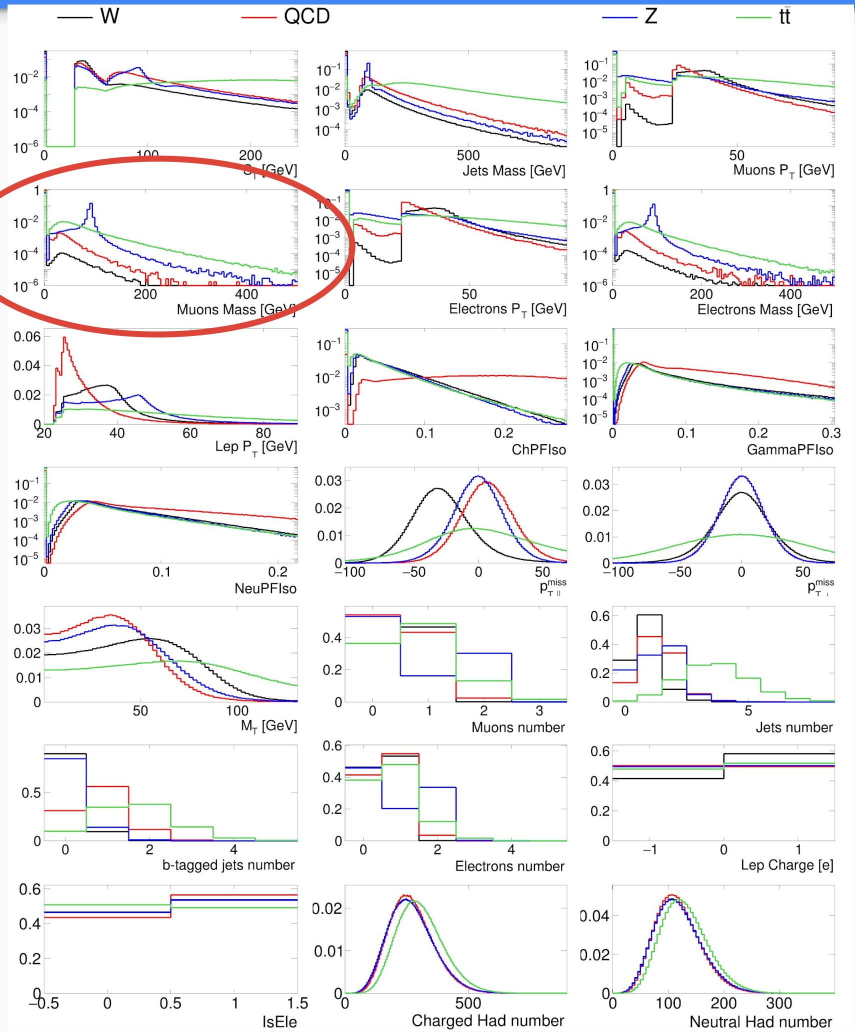


A use case: ℓ+X

- Stream of data with at least one interesting lepton (e or μ)
  - p_T > 23 GeV & ISO < 0.45
- SM contribution:

Process	Event fraction in the stream	Events/month
W	59%	110M
QCD	34%	63M
Z	6.7%	12M
tt	0.3%	0.6M

- Events represented by 21 high level features (HLF)
  - Broad general choice, not BSM tailored



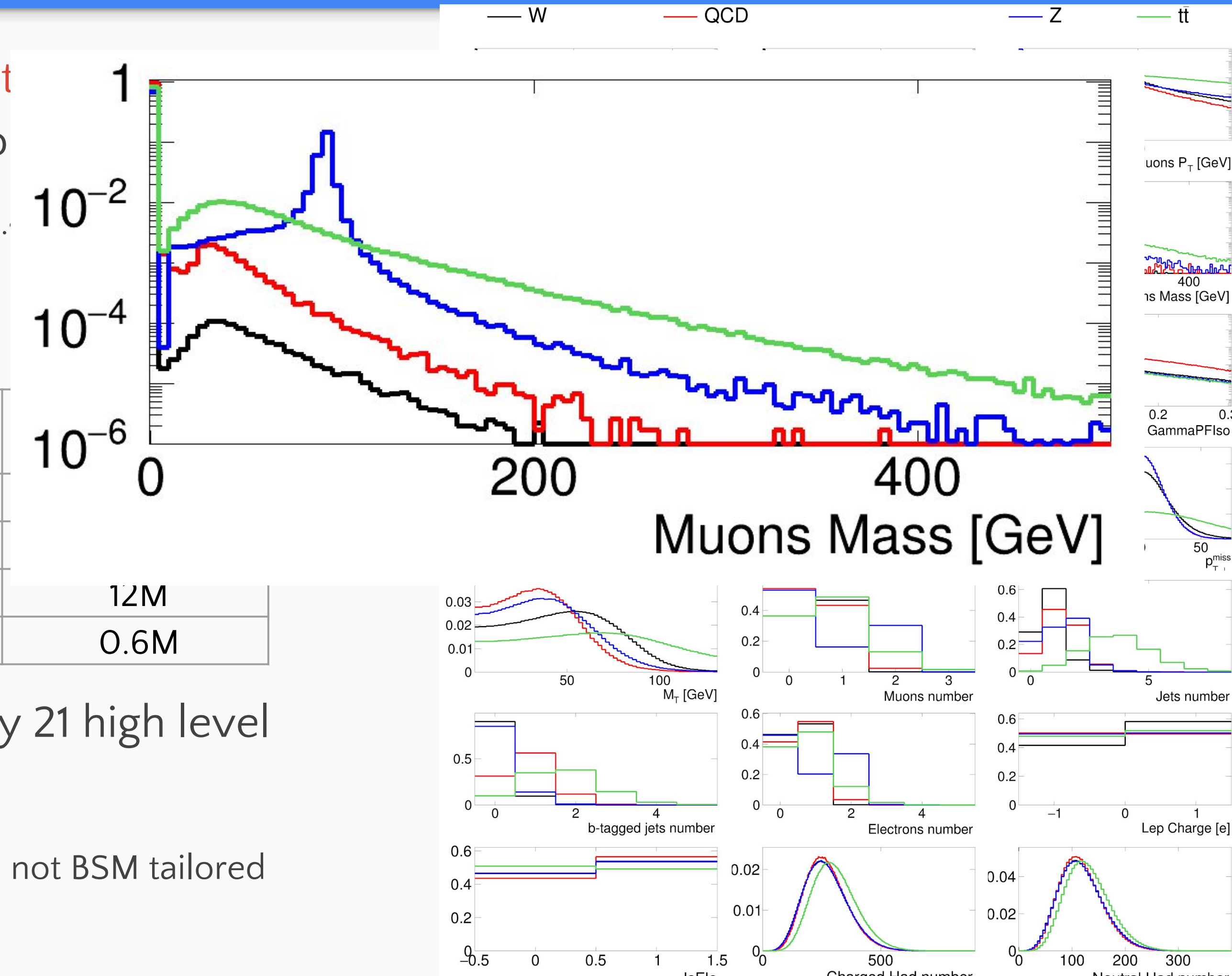
A use case: ℓ+X

- Stream of data with at
  interesting lepton (e o
    - p_T > 23 GeV & ISO < 0.

- SM contribution:

Process	Event fraction in the stream	
W	59%	
QCD	34%	
Z	6.7%	12M
tt	0.3%	0.6M

- Events represented by 21 high level features (HLF)
    - Broad general choice, not BSM tailored



How to deploy it

- VAE trained only on SM
- VAE does not see the BSM (if any) until it's evaluated on new events

1. Train one (or more) VAE(s):
   a. Train on MC (pure SM)
   b. Training on data (robust against signal injection)

2. Put the VAE(s) online in the trigger
   a. Evaluate each event
   b. Acceptance threshold such that O(10) SM events/day are triggered

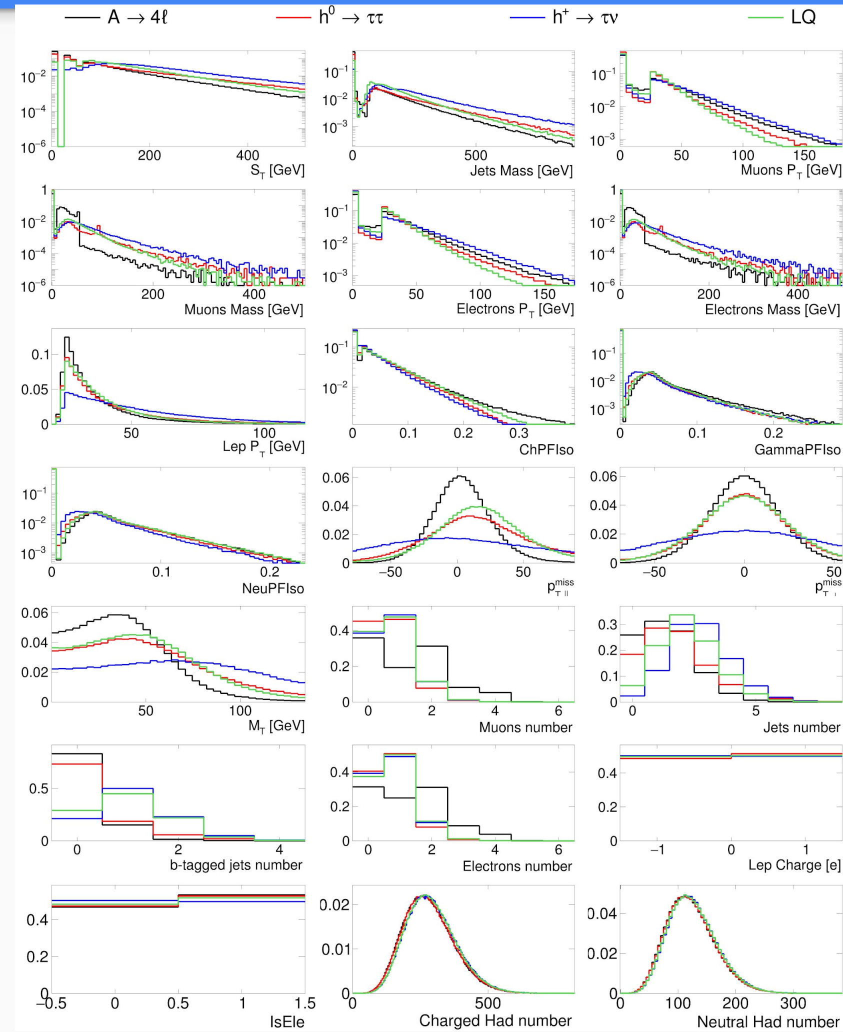
3. Collect events in a dedicated dataset
   a. Visual inspection
   b. Develop targeted analysis

BSM benchmark models

Light BSM which are usually very hard to trigger with standard strategies

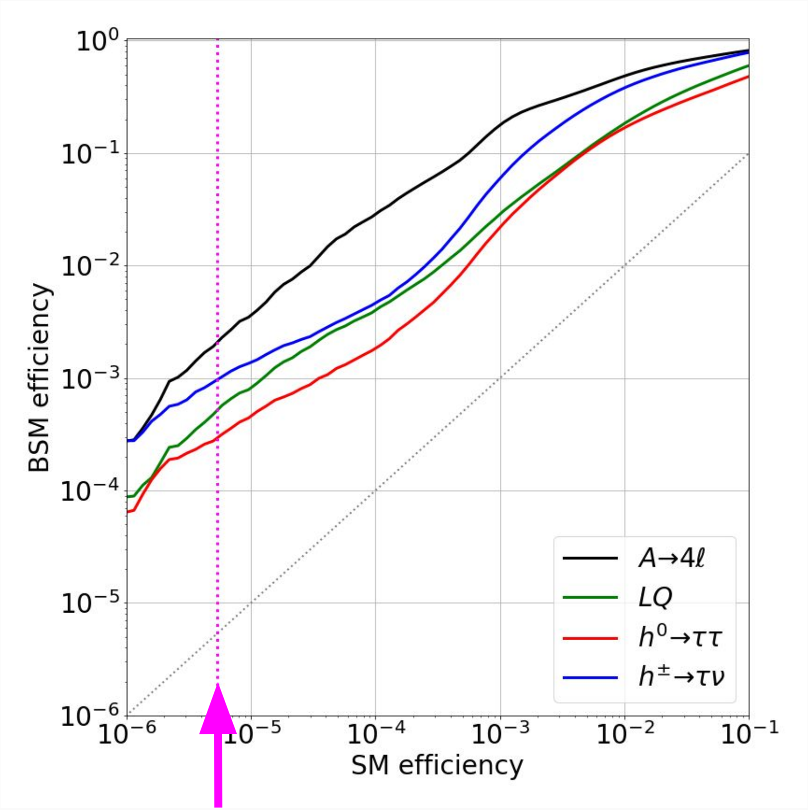
- A ⟶ 4ℓ: neutral scalar, M = 50 GeV
- LQ⟶ bτ: leptoquark, M = 80 GeV
- h^0 ⟶ ττ: neutral scalar, M = 60 GeV
- $h^{\pm}$ ⟶τν: charged scalar, M = 60 GeV

BENCHMARKING ONLY, NOT USED FOR TRAINING



Given the model independent nature, there is no unique way to define benchmarks.

Results

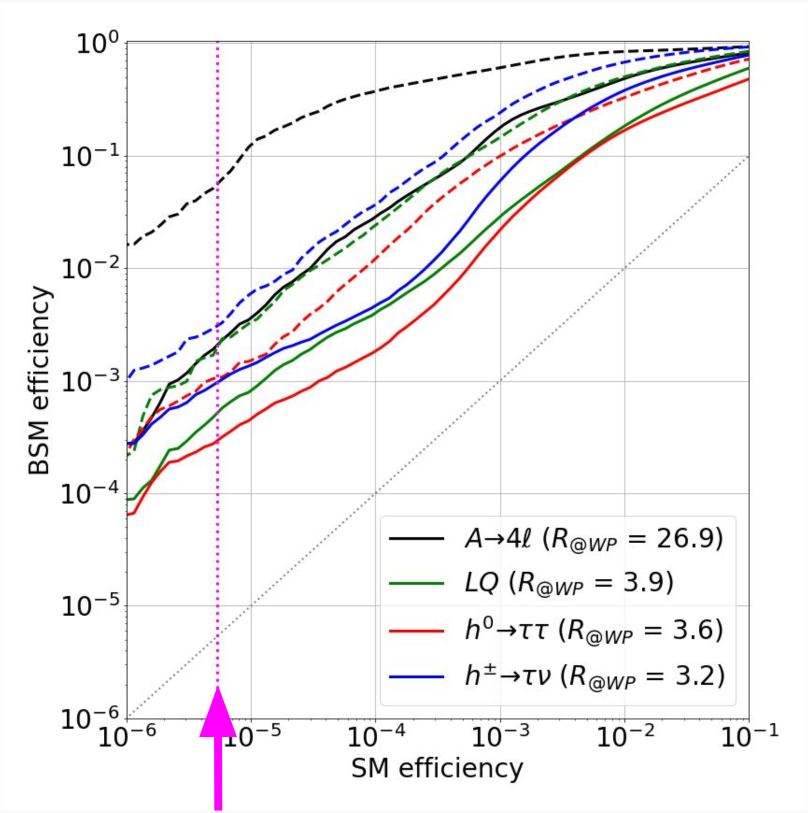


$\varepsilon_{SM} = 5.4 \cdot 10^{-6} \Leftrightarrow$ 30 evts/day

- VAE
  - A single one, trained only on SM
  - Applied to all the BSM

Results

--- Model dep. ——— VAE



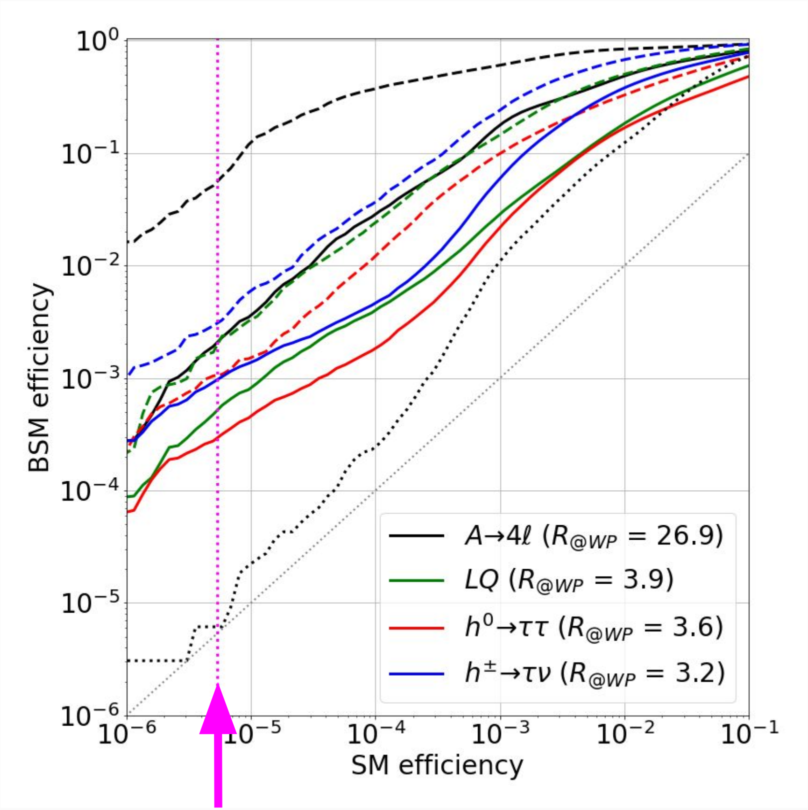
$\varepsilon_{SM} = 5.4 \cdot 10^{-6} \Leftrightarrow$ 30 evts/day

- VAE
  - A single one, trained only on SM
  - Applied to all the BSM
- Model dependent clf
  - 4 in total, each one trained on a specific BSM vs SM
  - Set target performances

Results

--- **Model dep.** ——— **VAE**

··· Model dep. on a different model



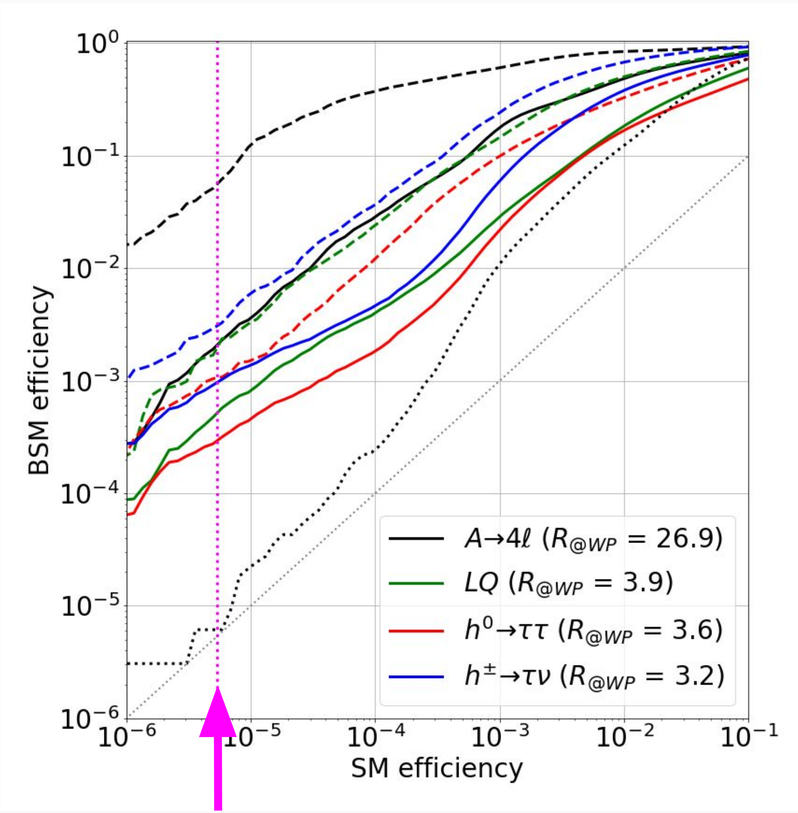
$\varepsilon_{SM} = 5.4 \cdot 10^{-6} \Leftrightarrow$ 30 evts/day

- VAE
  - A single one, trained only on SM
  - Applied to all the BSM
- Model dependent clf
  - 4 in total, each one trained on a specific BSM vs SM
  - Set target performances
- Model dep. clf applied to a different BSM model

Results

--- **Model dep.** ——— **VAE**

··· Model dep. on a different model



$\varepsilon_{SM} = 5.4 \cdot 10^{-6} \Leftrightarrow$ 30 evts/day

Standard Model processes			
Process	VAE selection	Sample composition	Event/month
W	$3.6 \pm 0.7 \cdot 10^{-6}$	32%	379 ± 74
QCD	$6.0 \pm 2.3 \cdot 10^{-6}$	29%	357 ± 143
Z	$21 \pm 3.5 \cdot 10^{-6}$	21%	256 ± 43
$t\bar{t}$	$400 \pm 9 \cdot 10^{-6}$	18%	212 ± 5
Tot			1204 ± 167

BSM benchmark processes			
Process	VAE selection efficiency	Cross-section 100 events/month [pb]	Cross-section S/B = 1/3 [pb]
$A \to 4\ell$	$2.8 \cdot 10^{-3}$	7.1	27
$LQ \to b\tau$	$6.7 \cdot 10^{-4}$	30	110
$h^0 \to \tau\tau$	$3.6 \cdot 10^{-4}$	55	210
$h^\pm \to \tau\nu$	$1.2 \cdot 10^{-3}$	17	65

Efficiency drop $\lesssim$ 10 w.t.r. to model-dependent classifier (i.e. optimal limit)

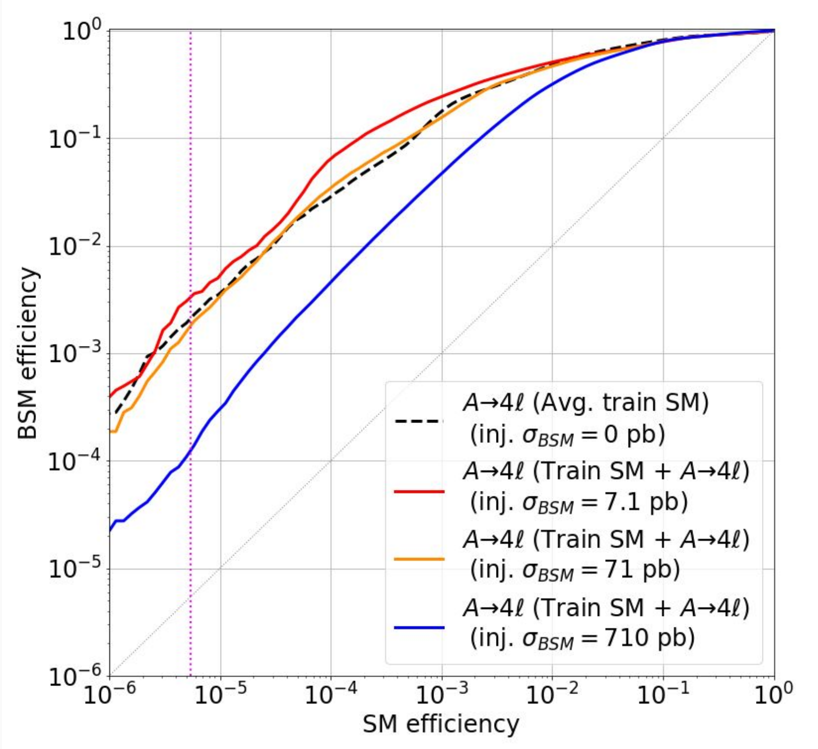
Train on data

If BSM is rare enough, having it in the training sample will not spoil performances.

- Train on a dataset with signal injected:

Injected evts	Training set fraction	VAE selected evts/month	Anomaly fraction
700	$2 \cdot 10^{-4}$	134	12%
7k	$2 \cdot 10^{-3}$	957	48%
70k	$2 \cdot 10^{-2}$	6	0.6%

- SM size: 3.5M evts ≃ 100 pb^{-1} ≃ few hours



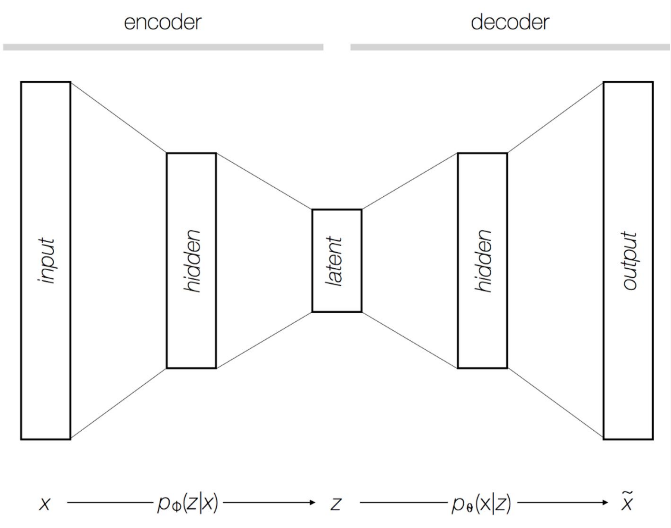
No performance drop up to 10^{-3} signal contamination in training set (**huge, S/B = 1**):

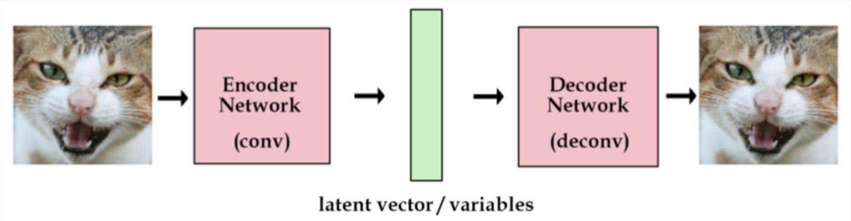
⇒ Can be trained on data without impacting BSM efficiency

Let's open the box

Auto-encoders in one slide

- Data coding algorithms which learn to describe a given dataset in a latent space
- Unsupervised algorithm, used for data compression, generation, clustering, etc.
- Anomaly: any event whose output is “far” from the input



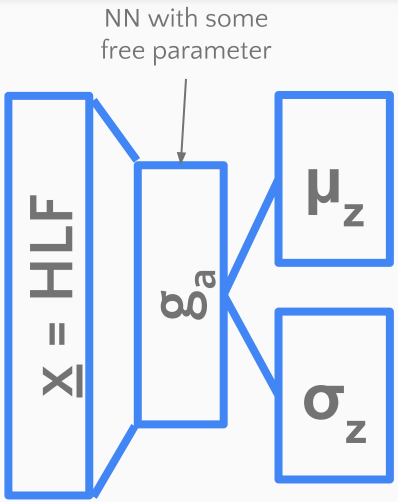


The Variational Auto-Encoder

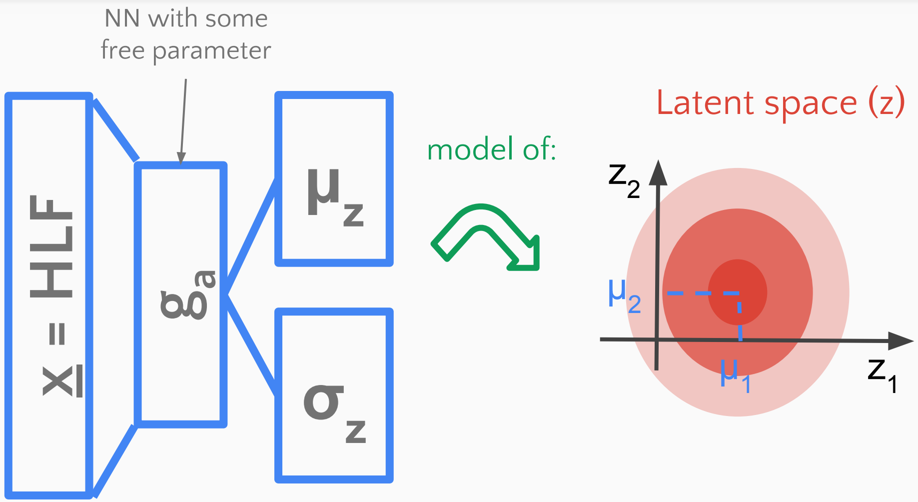
The Variational Auto-Encoder

$\underline{x}$ = HLF

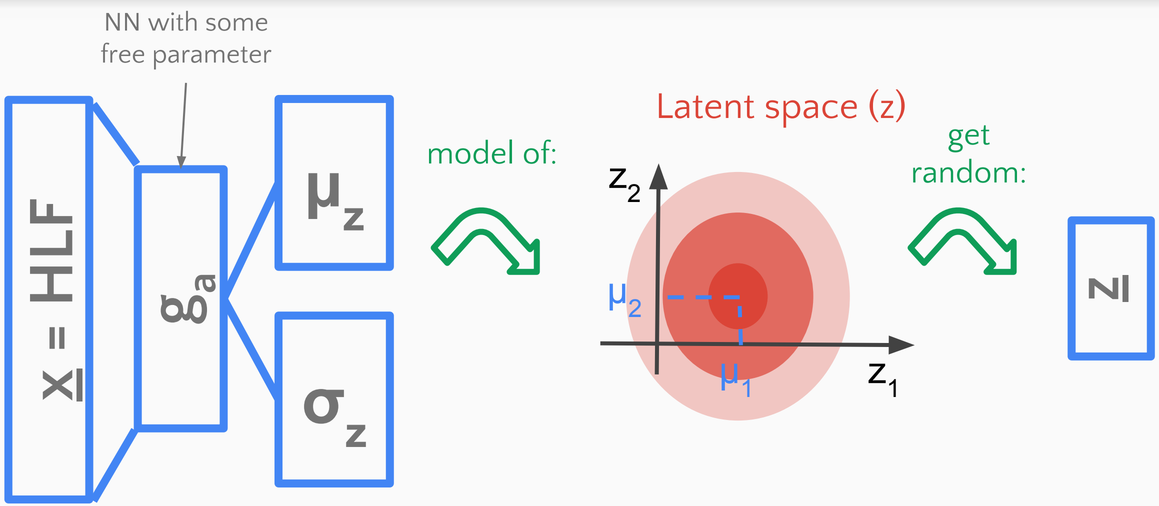
The Variational Auto-Encoder



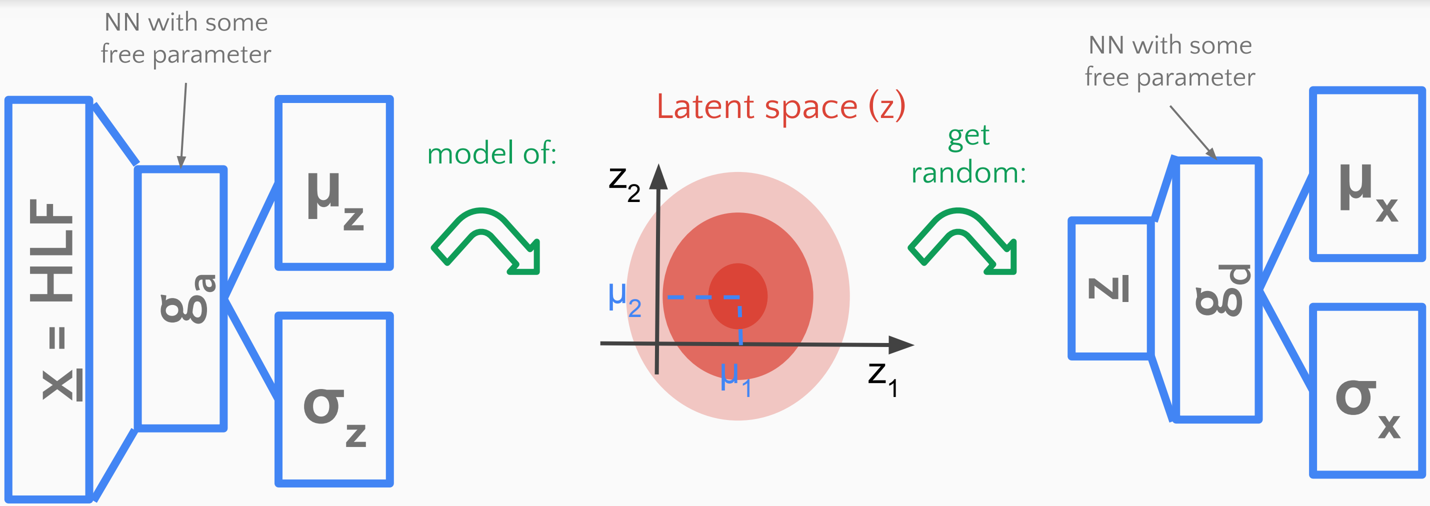
The Variational Auto-Encoder



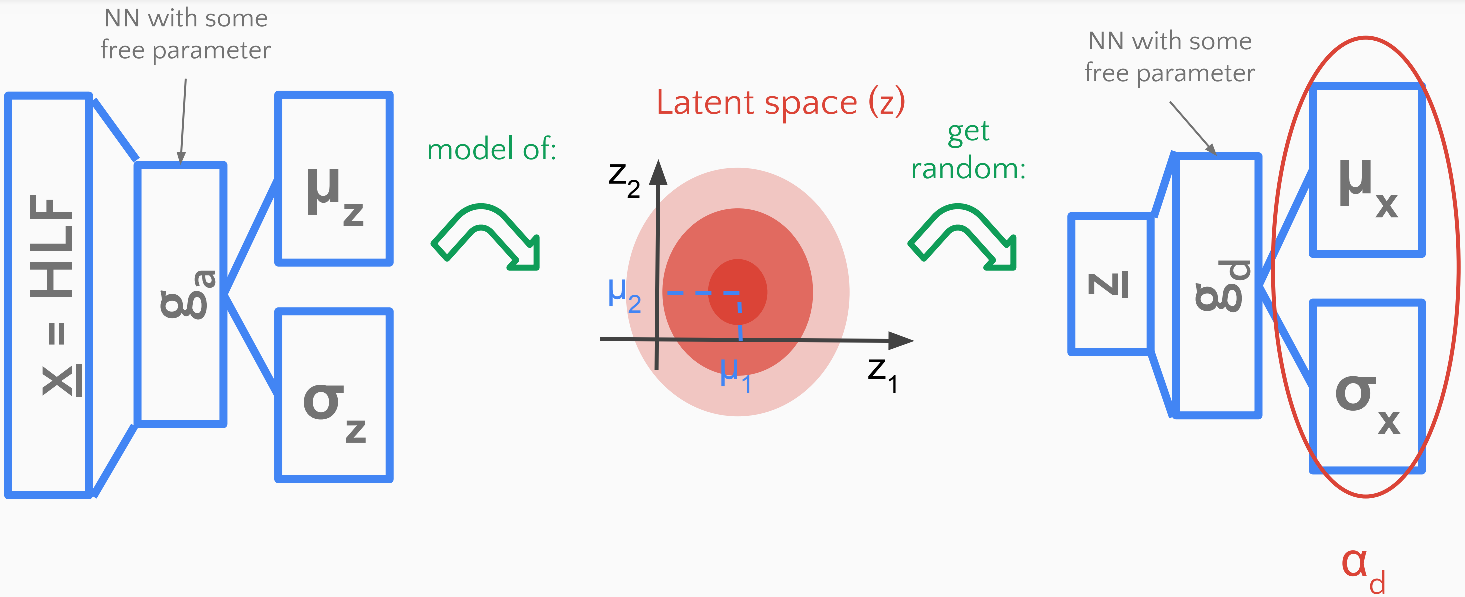
The Variational Auto-Encoder



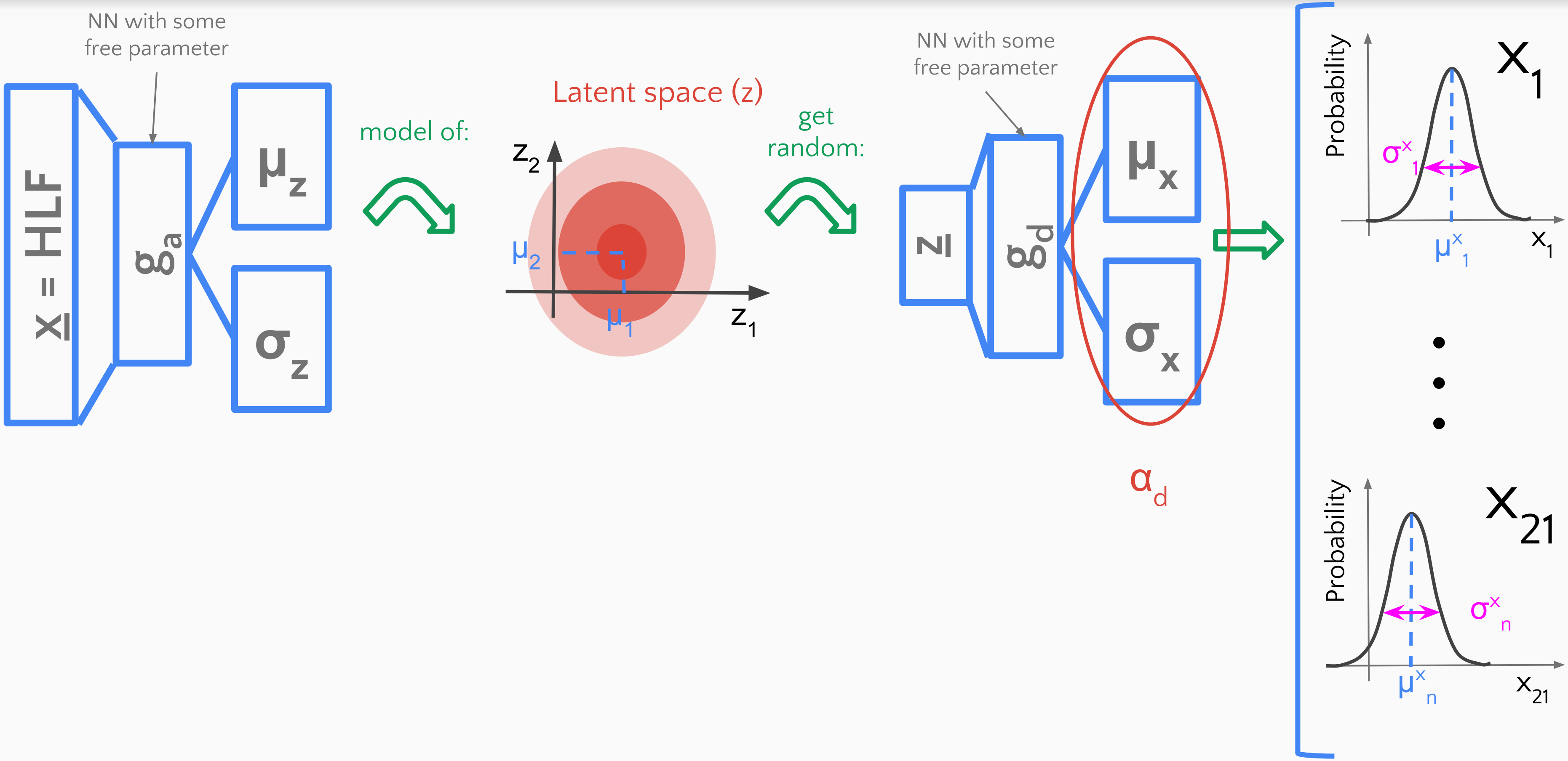
The Variational Auto-Encoder



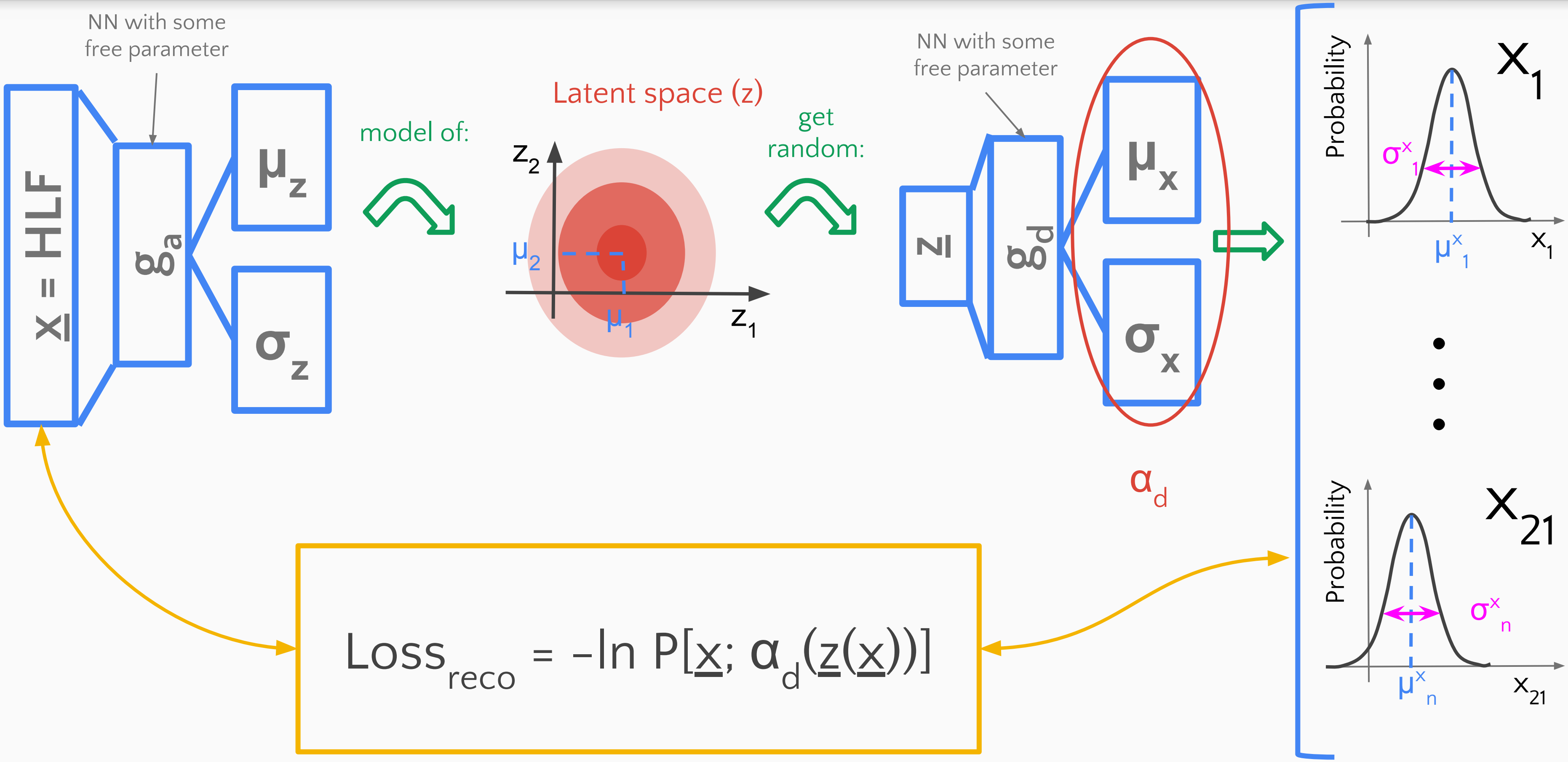
The Variational Auto-Encoder



The Variational Auto-Encoder

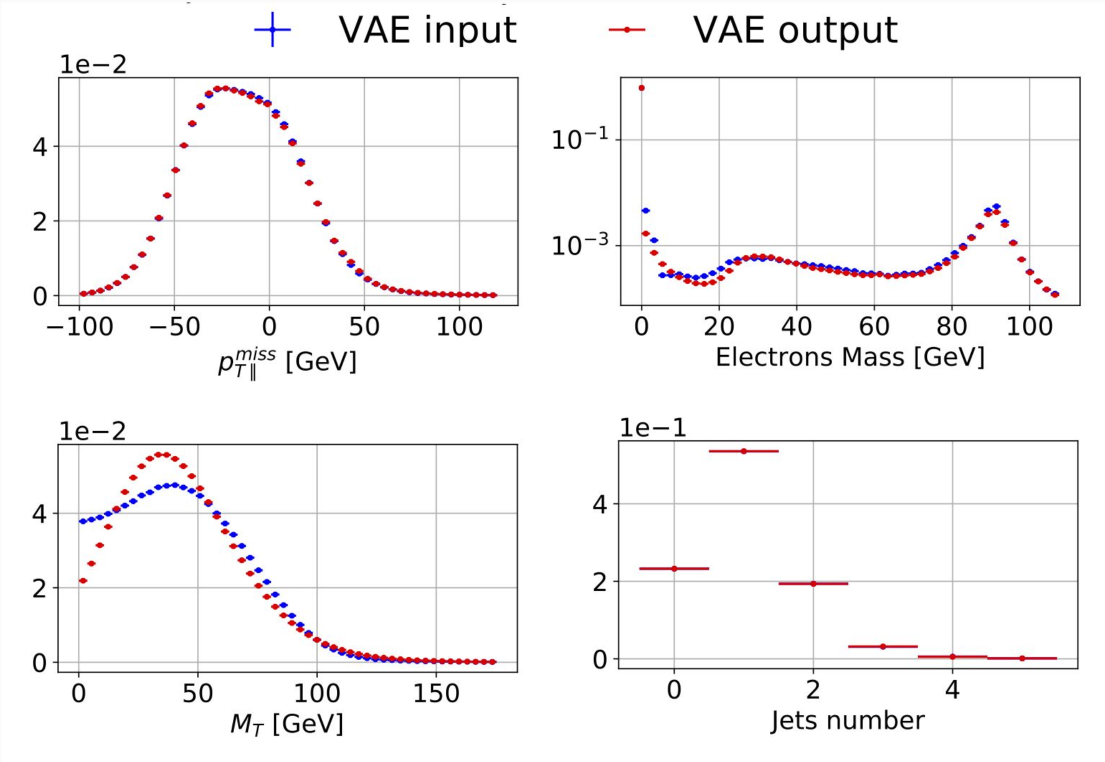


The Variational Auto-Encoder



Convergence check: SM auto-encoding

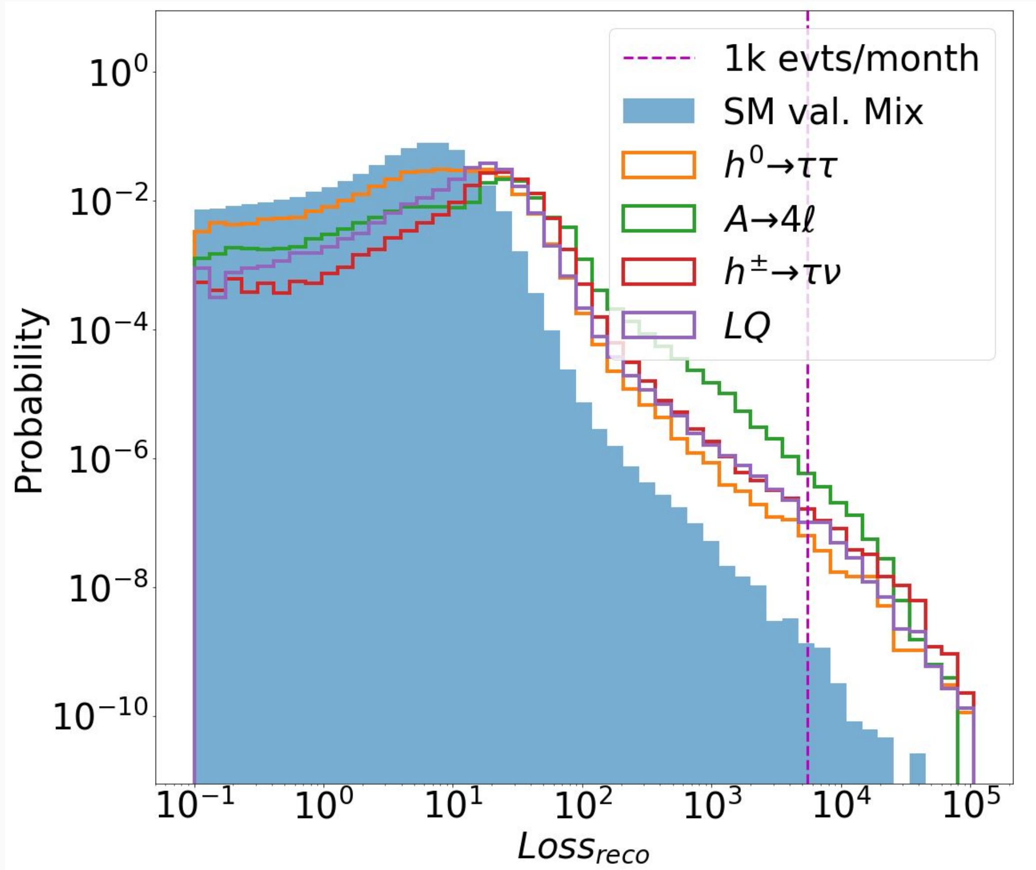
- Verifying encoding-decoding on validation set
  - Distributions of input vs generated from decoder
- Good agreement, with small discrepancy here and there
- Best autoencoder is not necessarily the best anomaly detector



Defining anomaly

- Anomaly defined by a p-value threshold on a given test statistics

- VAE loss function is the natural choice for the test statistics



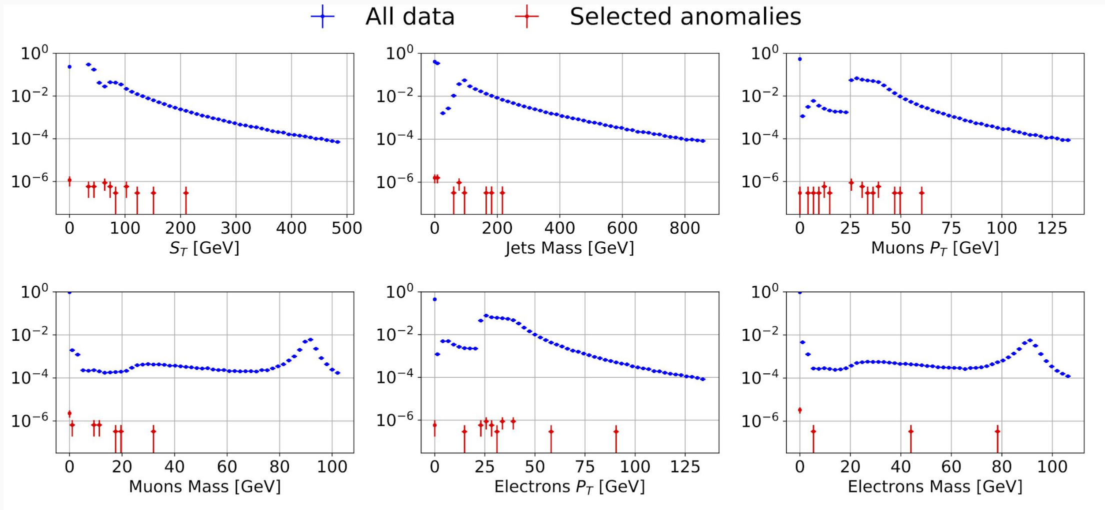
$\text{Loss}_{\text{reco}}$ used as test statistics.

Not a tail-cut algorithm

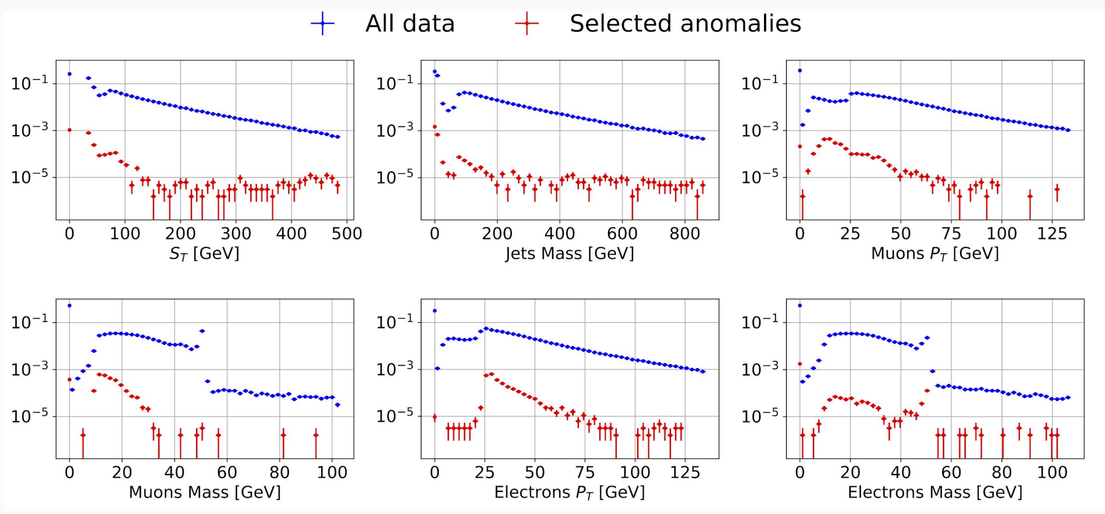
- Selected events stand on the core of 1D distributions

- Expand the possibility w.t.r. to classical anomaly detection triggers

SM Mix

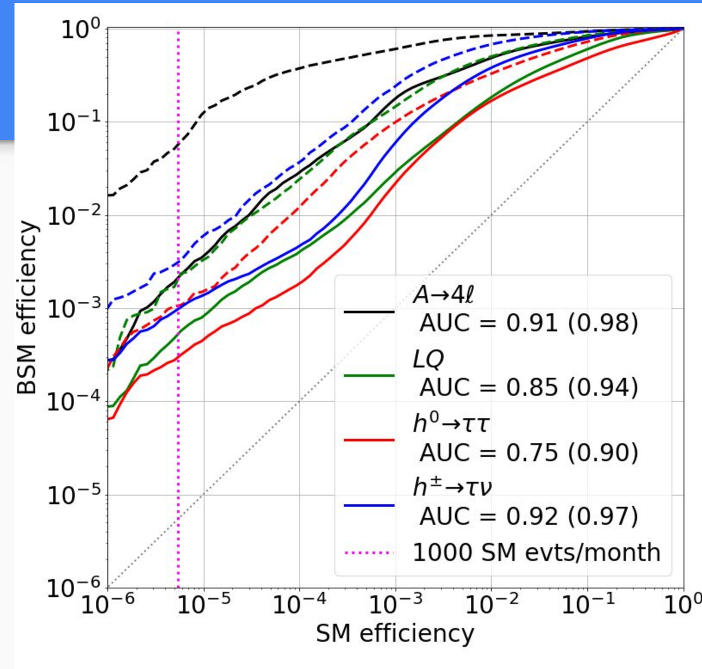


$A \rightarrow 4\ell$



Conclusions

- VAE as model-independent BSM trigger
  - Train just on SM, no need to specify a BSM model
  - Can be trained on data
- Select 30 events/day and create a dataset of anomalous events
  - Further study within and outside the collaborations
- Allows (benchmark models) to probe 10-100 pb cross section
  - Alternative strategy, parallel to canonical approaches
- Might open new physics directions



Standard Model processes			
Process	VAE selection	Sample composition	Event/month
W	$3.6 \pm 0.7 \cdot 10^{-6}$	32%	379 ± 74
QCD	$6.0 \pm 2.3 \cdot 10^{-6}$	29%	357 ± 143
Z	$21 \pm 3.5 \cdot 10^{-6}$	21%	256 ± 43
$t\bar{t}$	$400 \pm 9 \cdot 10^{-6}$	18%	212 ± 5
Tot			1204 ± 167

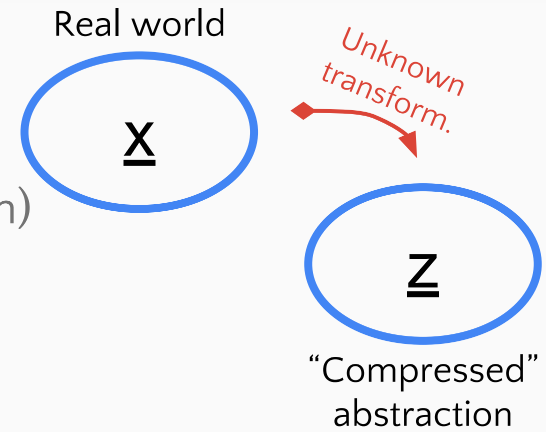
BSM benchmark processes			
Process	VAE selection efficiency	Cross-section 100 events/month [pb]	Cross-section S/B = 1/3 [pb]
$A \to 4\ell$	$2.8 \cdot 10^{-3}$	7.1	27
$LQ \to b\tau$	$6.7 \cdot 10^{-4}$	30	110
$h^0 \to \tau\tau$	$3.6 \cdot 10^{-4}$	55	210
$h^\pm \to \tau\nu$	$1.2 \cdot 10^{-3}$	17	65

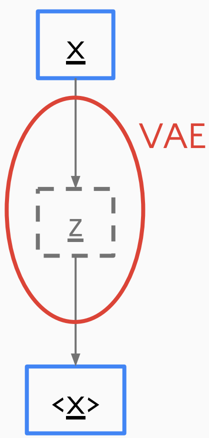
BACKUP

The Variational Auto-Encoder (1/2)

Working hypothesis:

- Each event has a set of features: $\underline{x} \in \mathbb{R}^n$
- Relevant information can be summarized in: $\underline{z} \in \mathbb{R}^m$ (n>m)
  - Lost information for is somehow stored in the encoding/decoding function





Goal:

- Creating a function that, ON THE STD DATASET, allow to consistently compress and decompress the event information
  - the VAE should underperform on a different dataset because the lost information is different from the one of the training
- Consistency can be directly checked by comparing input and output

Training loss function technicalities

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

Reconstruction likelihood :

- "True" loss (NLL)
- Force the autoencoded distribution to describe the x
- The goodness of the VAE depends on the ability of f_j to describe p(x | z)

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

Regularization term:

- Force the z distribution to a Normal
- To avoid strange latent variable

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

The Variational Auto-Encoder

Encoder:

- For each value of $\underline{x}$, tell what is the pdf of $\underline{z}$
- Practically:
  - A functional form $f_e[\underline{z};\, \alpha_e(\underline{x})]$ is fixed

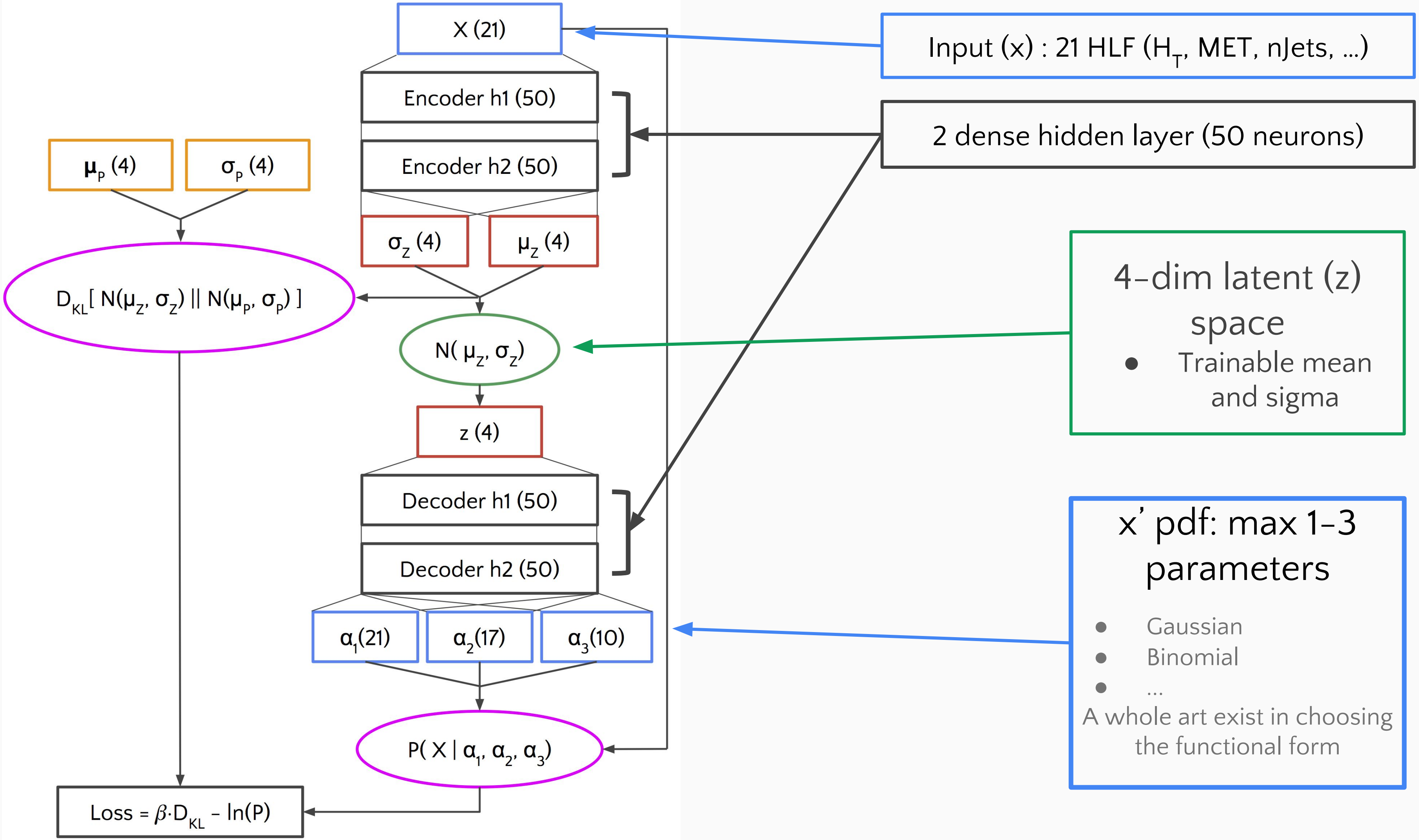
The encoder function $g_e\colon \underline{x} \longrightarrow \alpha_e$ gives the value of the $\underline{z}$ distribution parameters

!!: $\underline{x}$ and $\underline{z}$ are swapped w.r.t. to Encoder

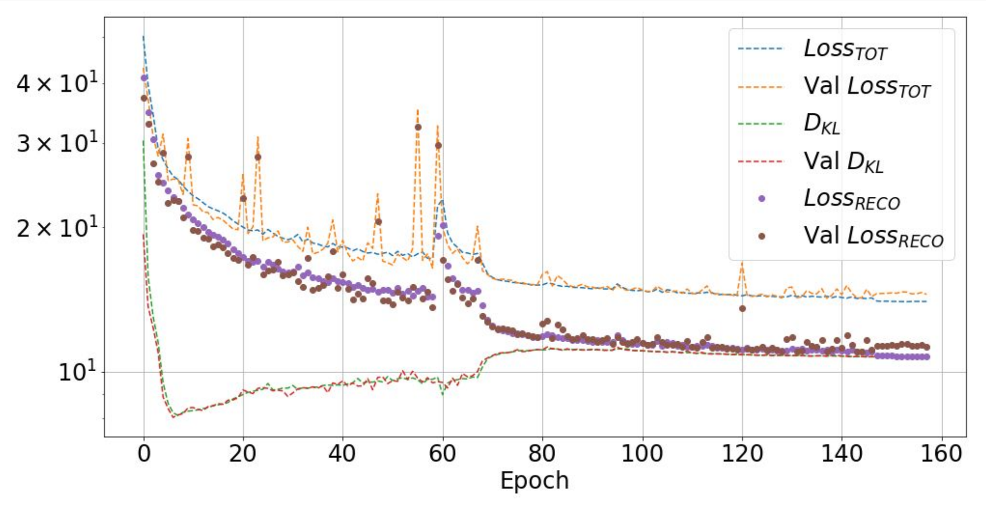
Decoder:

- For each value of $\underline{z}$, tell what is the pdf of $\underline{x}$
- Practically:
  - A functional form $f_d[\underline{x};\, \alpha_d(\underline{z})]$ is fixed

The encoder function $g_d\colon \underline{z} \longrightarrow \alpha_d$ gives the value of the $\underline{x}$ distribution parameters

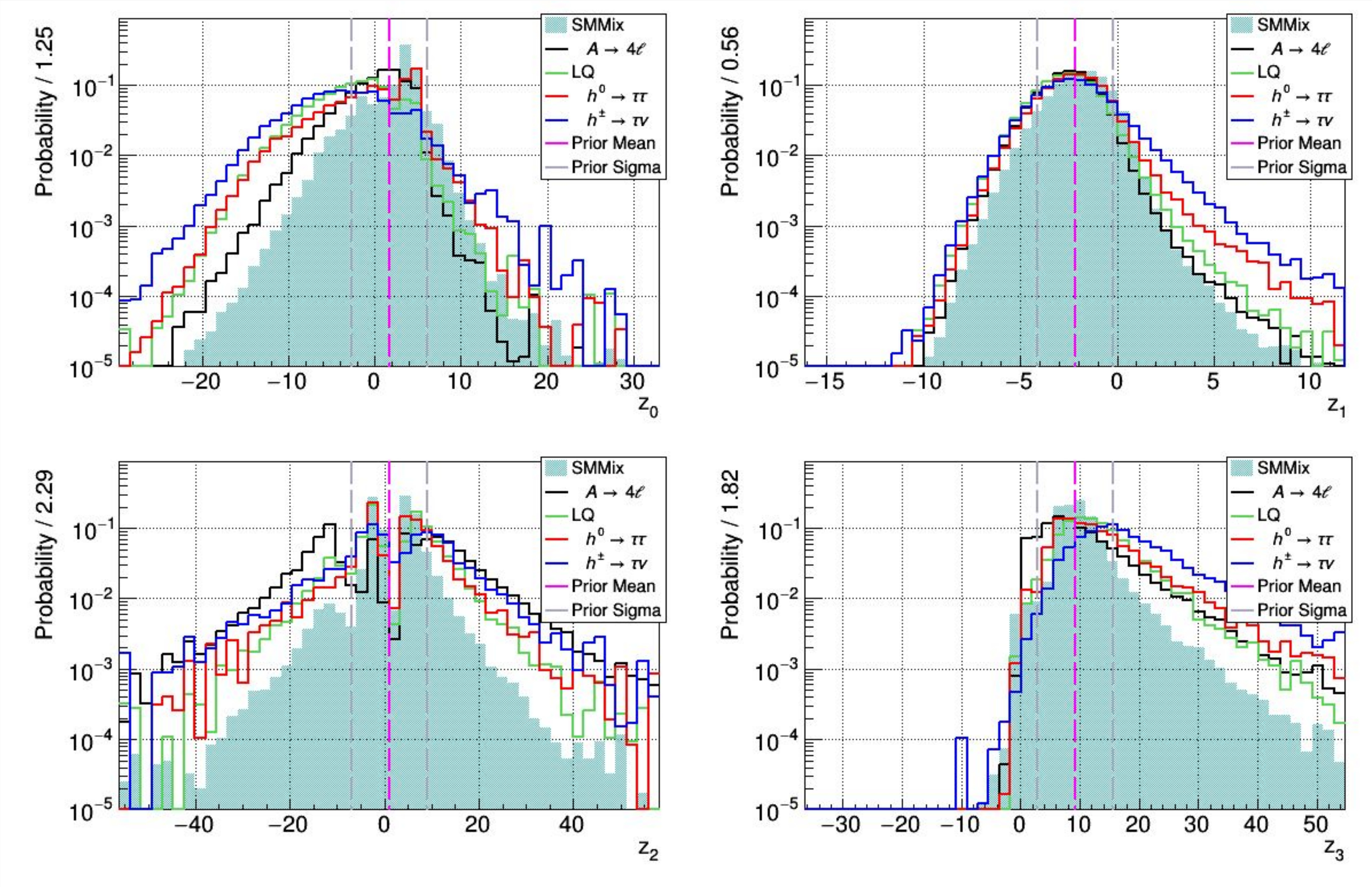

X (21)
Encoder h1 (50)
Encoder h2 (50)
μp (4)
σp (4)
σZ (4)
μZ (4)
DKL[N(μZ, σZ) || N(μP, σP)]
N(μZ, σZ)
z (4)
Decoder h1 (50)
Decoder h2 (50)
α1(21)
α2(17)
α3(10)
P(X | α1, α2, α3)
Loss = β·DKL - ln(P)
Input (x) : 21 HLF (HT, MET, nJets, ...)
2 dense hidden layer (50 neurons)
4-dim latent (z) space
Trainable mean and sigma
x' pdf: max 1-3 parameters
Gaussian
Binomial
...
A whole art exist in choosing the functional form

Training: not a easy beast



- Optimizer
  - Adam
  - Callbacks
- Samples
  - 3.5 M event for training
  - 3.5 M for validation
  - # evt/# par >> 10
- The training
  - Not long, about 1h
  - Spike not unusual
  - Delicate equilibrium of training parameters

Latent space distribution



Ops. conditions

Simulation details:

- Pythia 8
- Delphes
  - CMS phase II default card
- Training on 3.5 M of SM
  - Equivalent of 100 pb^{-1}

Machine working conditions:

- 8 months of data taking per year
- $L_{TOT} = 40\ fb^{-1}$
- $\langle L_{inst} \rangle = 2.8 \cdot 10^{33}\ cm^{-2}s^{-1}$
- $\langle PU \rangle = 20$
- $E_{CM} = 13$ TeV

The 21 considered features

- The absolute value of 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` [24] implementation of the anti-k_T jet algorithm [25], 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 ($p_{T,\parallel}^{\text{miss}}$) and orthogonal ($p_{T,\perp}^{\text{miss}}$) 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:

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

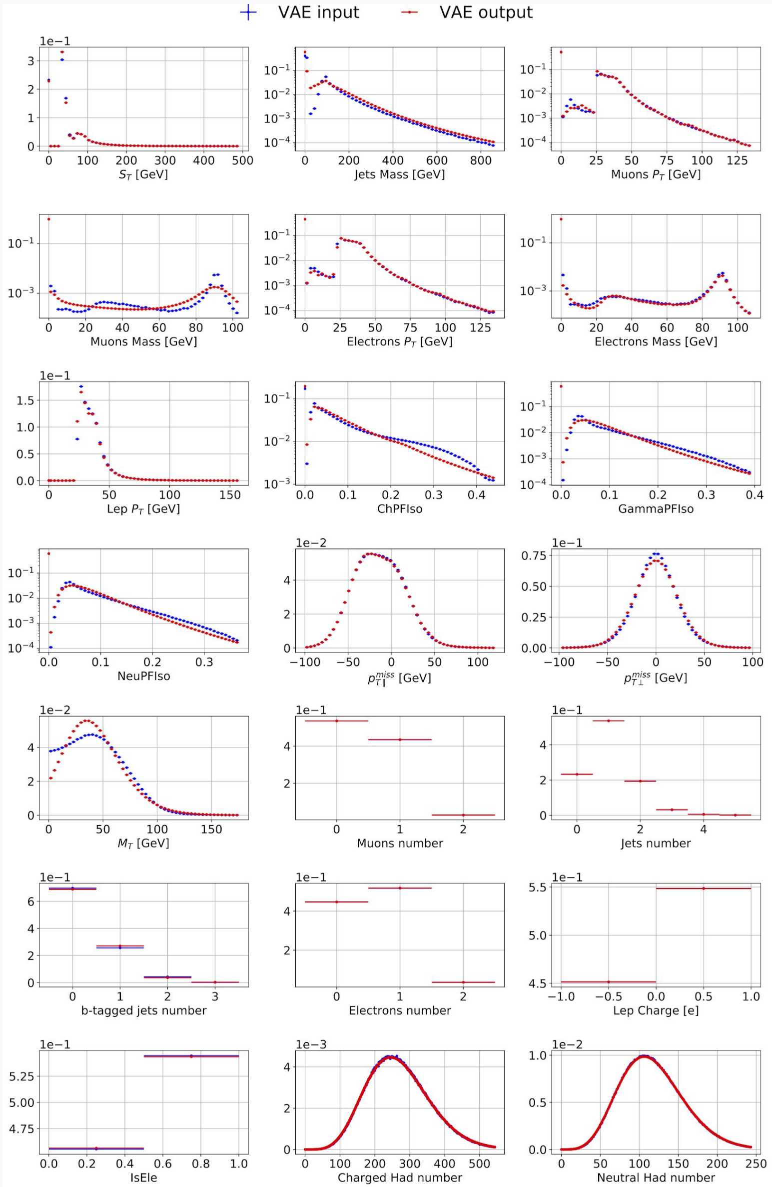
- The transverse mass, M_T, of the isolated lepton ℓ and the E_T^{miss} system, defined as:

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

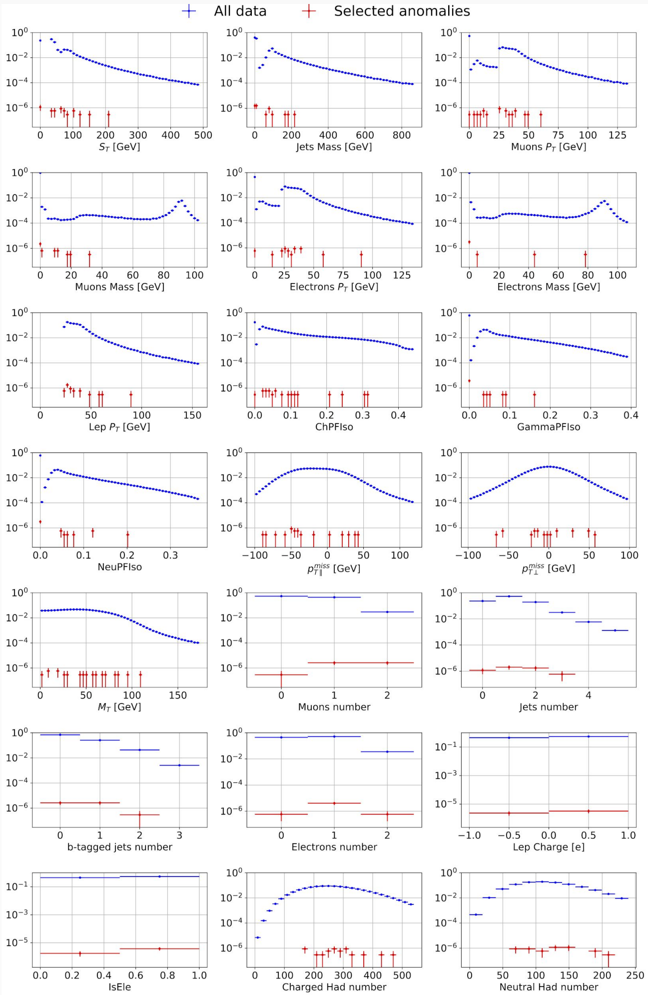
with $\Delta\phi$ the azimuth separation between the $\vec{p}_T^{\,\ell}$ and $\vec{p}_T^{\,\text{miss}}$ vectors, and E_T^{miss} the absolute value of $\vec{p}_T^{\,\text{miss}}$.

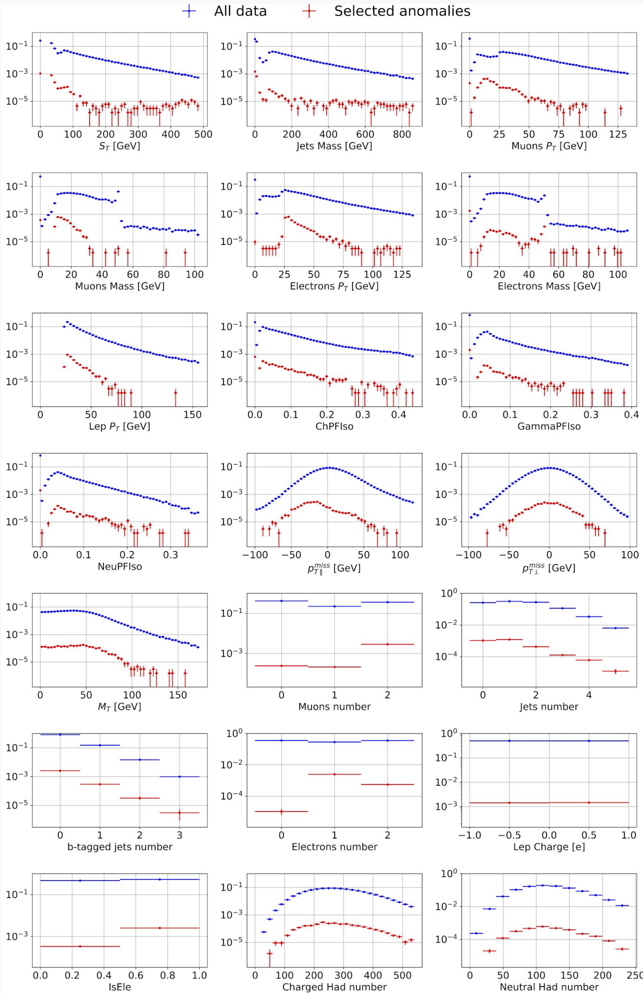
- The number of selected muons (N_μ).
- The invariant mass of this set of muons (M_μ).
- The absolute value of the total transverse momentum of these muons ($p_{T,TOT}^\mu$).
- The number of selected electrons (N_e).
- The invariant mass of this set of electrons (M_e).
- The absolute value of the total transverse momentum of these electrons ($p_{T,TOT}^e$).
- The number of reconstructed charged hadrons.
- The number of reconstructed neutral hadrons.

VAE auto-encoding cross-check

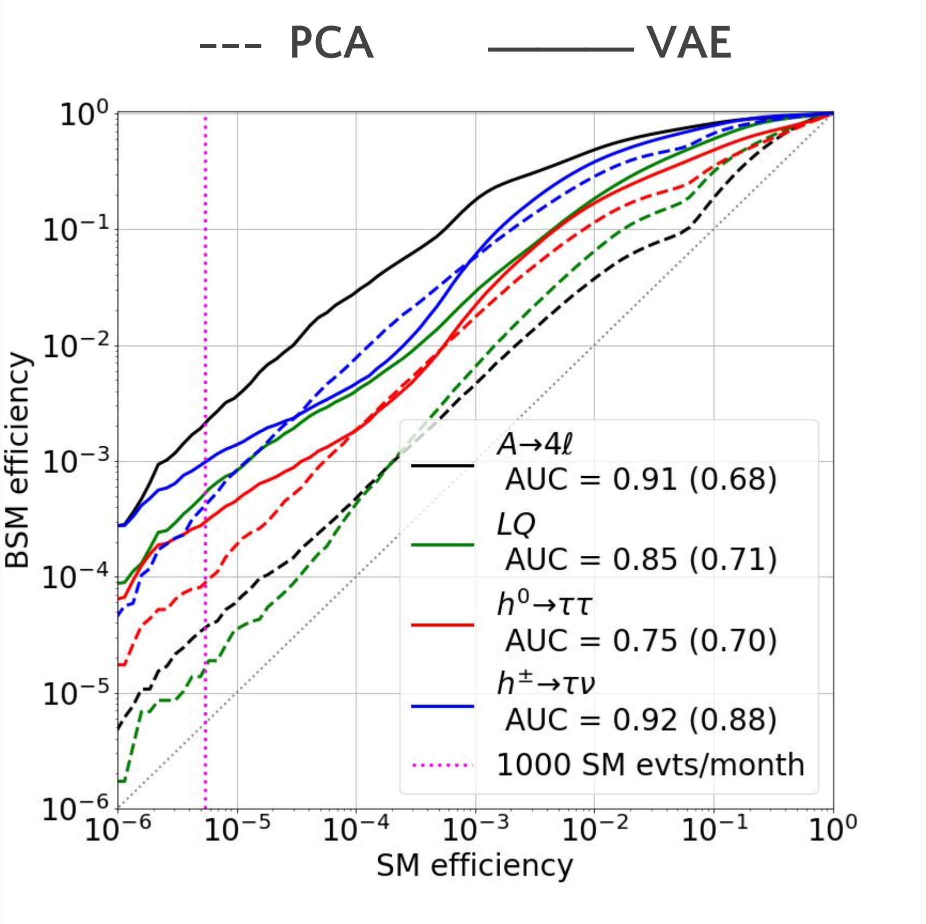


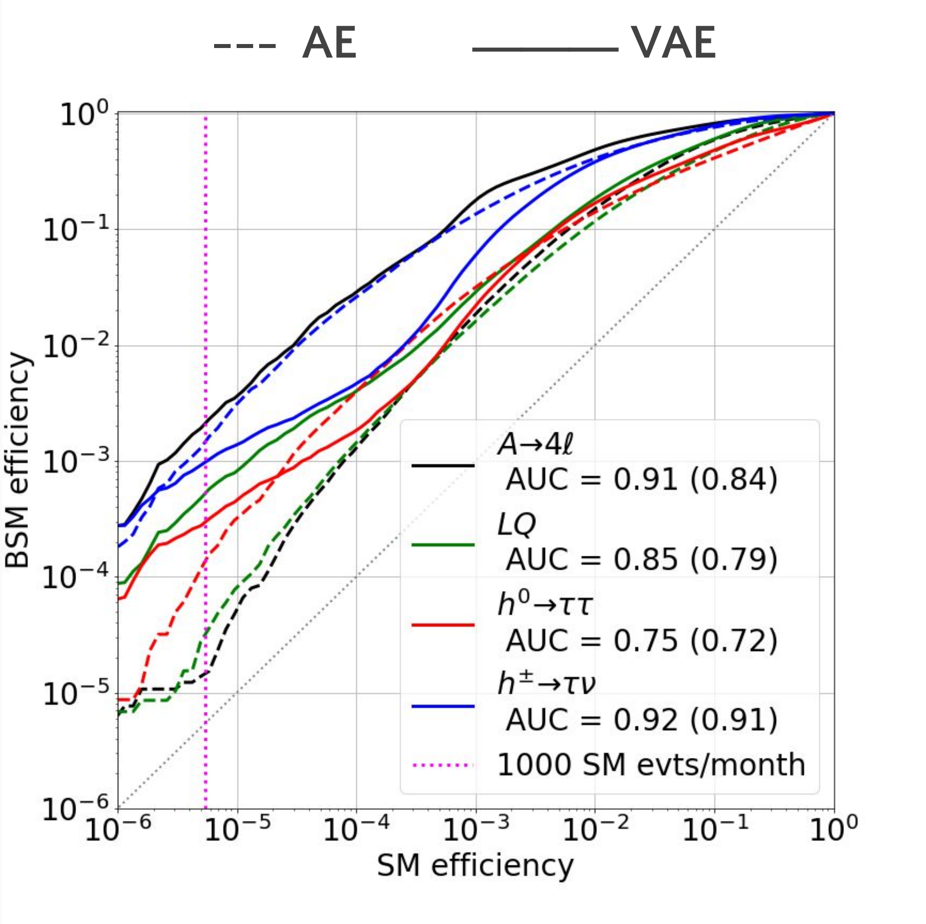
Not a tail-cut algorithm





Other algorithms comparison





Scenario w/o the VAE trigger

Reasonable cuts for single muon full trigger path (i.e. what we can really save on disk):

- p_T > 27 GeV
- ISO < 0.25

Efficiency

	SM	$A \rightarrow 4\ell$	$h \rightarrow \tau\tau$	$h \rightarrow \tau\nu$	LQ
VAE	5e-6	3e-3	4e-4	1e-3	7e-4
Single muon trigger	0.6	0.5	0.6	0.7	0.6

VAE trigger improves S/N ratio of 2-3 order of magnitude

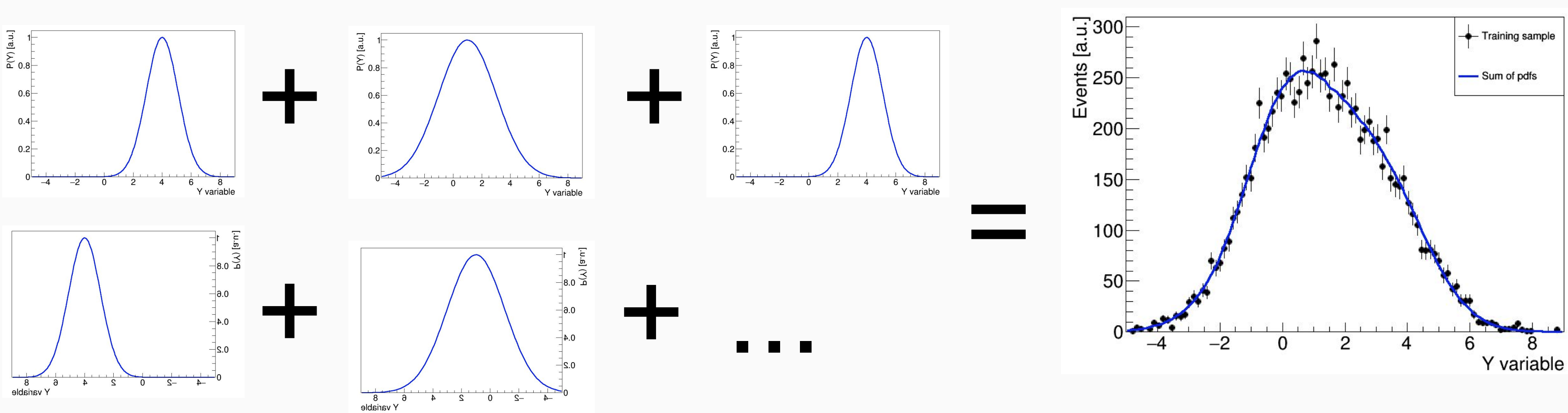
The great advantage of VAE is not only the ability to select BSM events but also to produce a high purity sample

Checking the convergence: sum of pdfs

High input dimension ⇒ Global convergence check

Predicted pdf for the single event

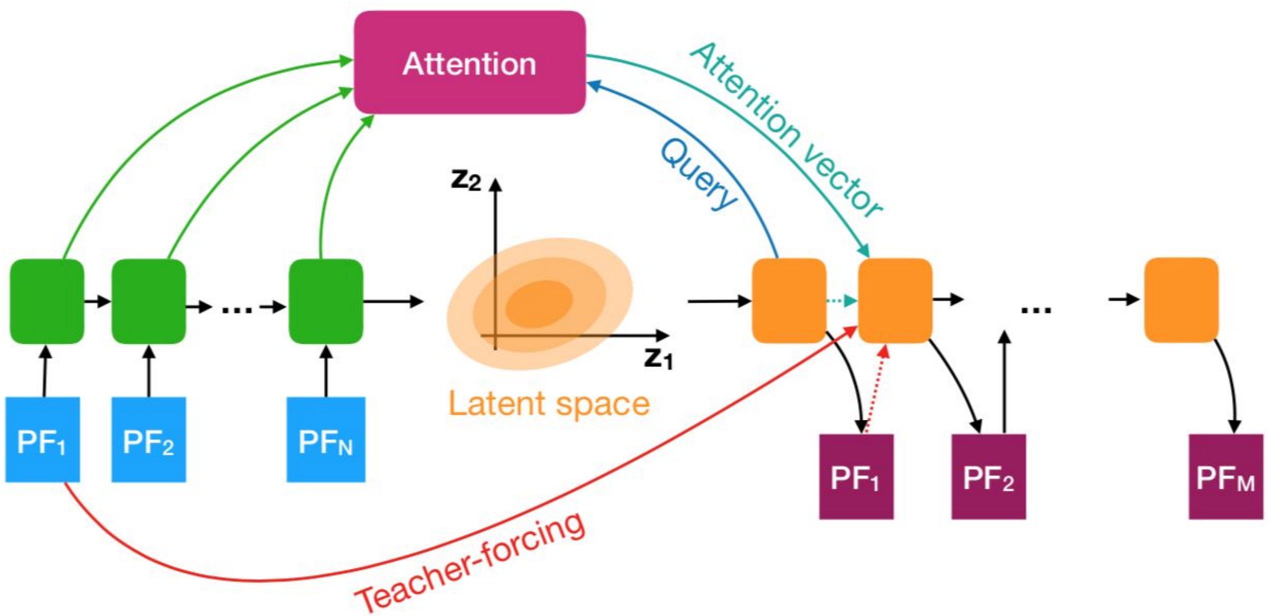
Obtain the distribution of the input as sum of all the predicted pdf



Particle Based VAE

Attentional Particle-VAE

- Attention: a function of both list of input particles and the current hidden state of the decoder's RNN cell.



Particle based VAE performance

Performance (2/2)

- Roughly 10 times worse than the VAE trained on HLFs.
- Optimization in progress, could be improved much further (more data + optimized loss functions).

SM p-value cutoff: 1.0E-5

Sample	Efficiency	Rate [Hz]	evts/month
ttbar	2.3E-3 +/- 1.5E-4	5.7E-3	4.8E+3 +/- 3.2E+2
QCD	1.0E-5 +/- 1.0E-5	2.5E-3	2.1E+3 +/- 2.1E+3
Wlnu	0.0E+1 +/- 0.0E+1	0.0E+1	0.0E+1 +/- 0.0E+1

Expected evts/month: 6883 +/- 5228

Sample	Efficiency	xsec (10 evts/month) [fb]	xsec (S/B = 0.3) [fb]
Ato4l	3.3e-4 +/- 8.6e-5	7.2E+3	1.5E+6
leptoquark	5.8e-4 +/- 7.6e-5	4.1E+3	8.5E+5
HiggsToTauTau	1.1e-3 +/- 1.5e-4	2.2E+3	4.5E+5
ChHiggsToTauNu	1.4e-3 +/- 1.7e-4	1.7E+3	3.4E+5

PRELIMINARY