(Variational) Auto-Encoders for New Physics Mining at the LHC

Olmo Cerri

Model-independent tagger for unexpected events

Flag non-SM events, despite their nature or particular features

1. Auto-Encoders & Anomaly Detection
2. Online trigger
3. Offline analysis

Physics anomaly detection

- Data mining concept
- 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 probability of belonging to the standard set
   - No assumption on the anomaly
3. Use this function to tag new events
   - Anomaly: low probability of belonging to the standard set
   - SM rare region or BSM

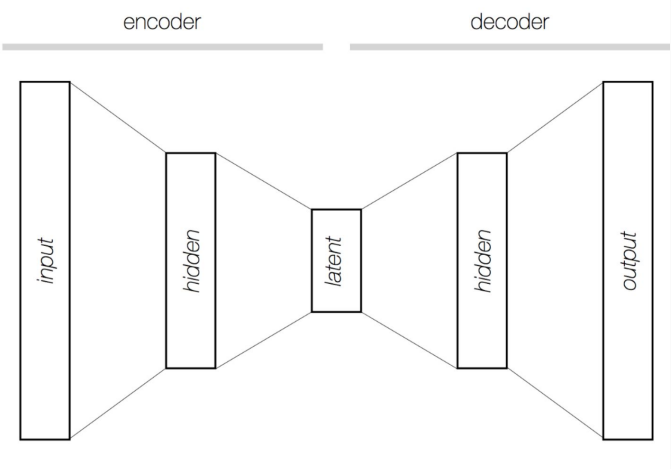
[1]: https://arxiv.org/abs/1312.6114

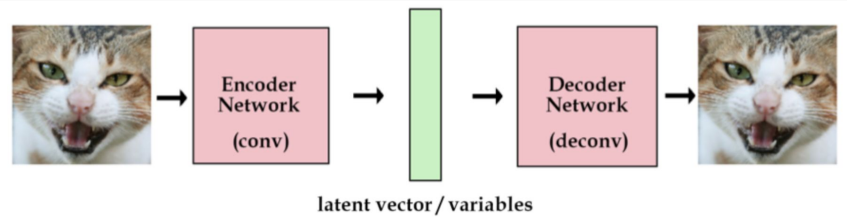
Auto-encoders in one slide

- **Map an input onto itself** passing through a latent representation

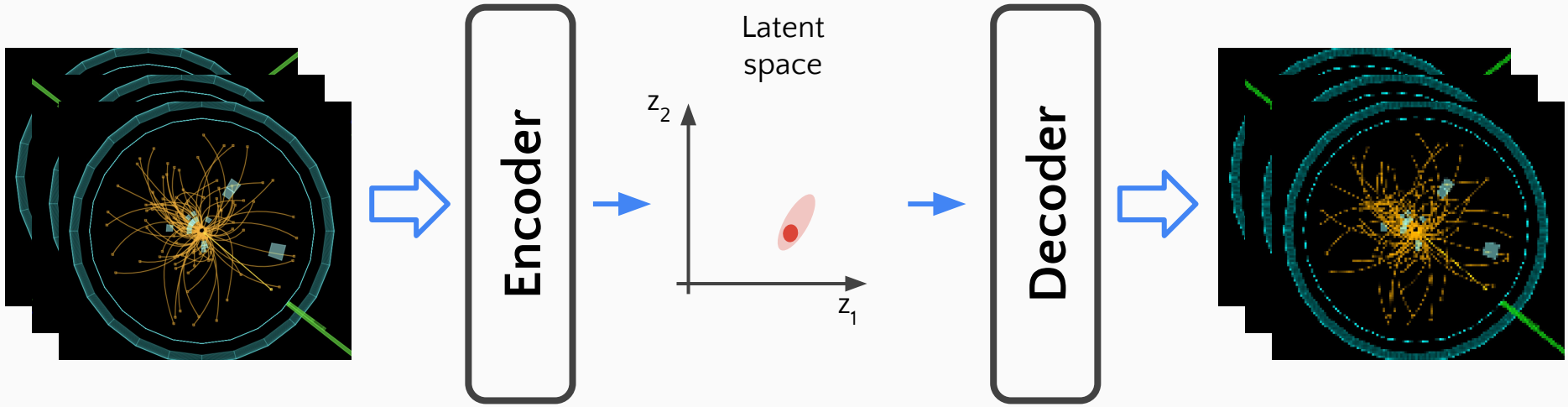
- **Unsupervised algorithm**, used for data compression, generation, clustering, etc.

- **Anomaly**: any event whose output is "far" from the input

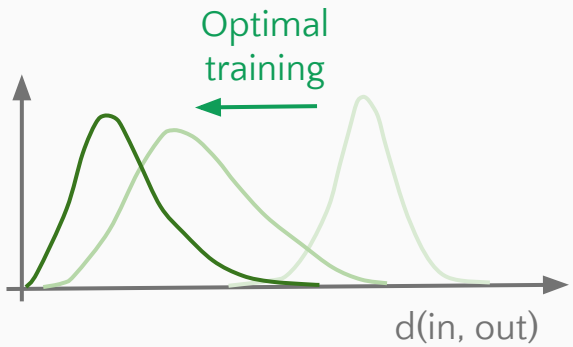




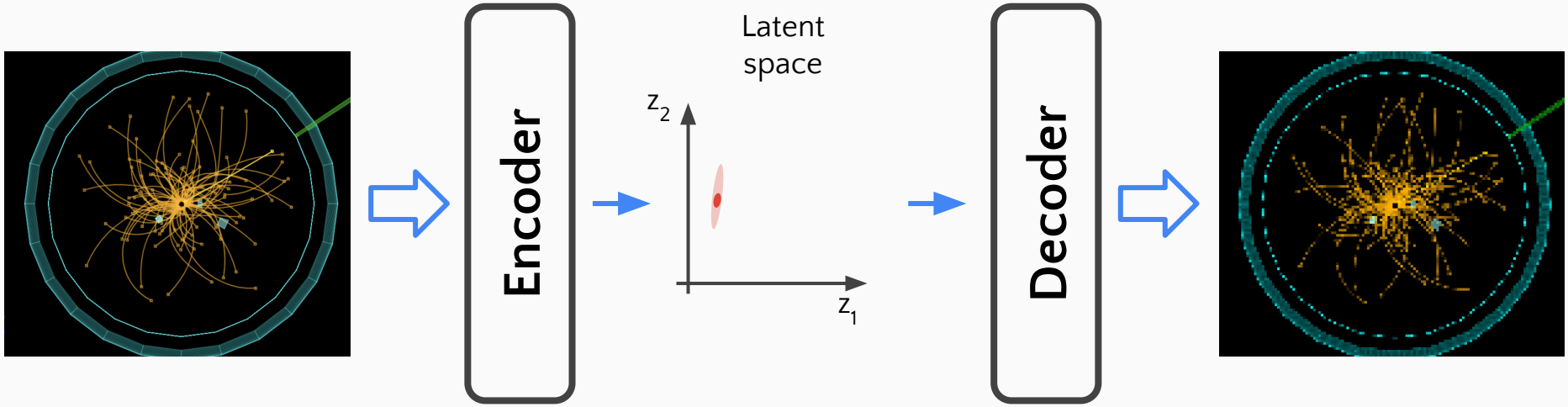
AE anomaly detection: Training



Training:
Fit the VAE params to minimize the input-output distance

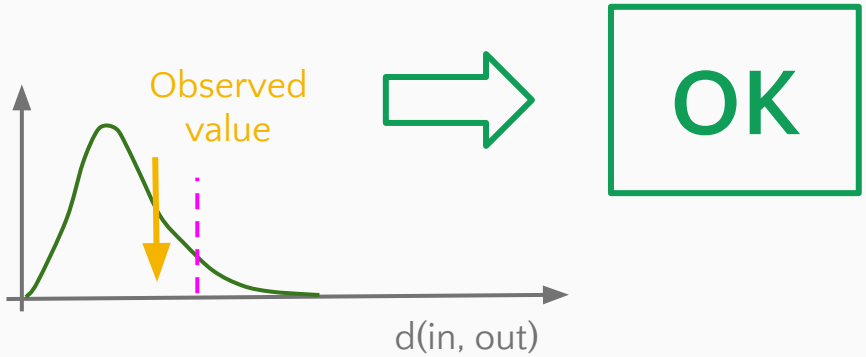


AE anomaly detection: Inference

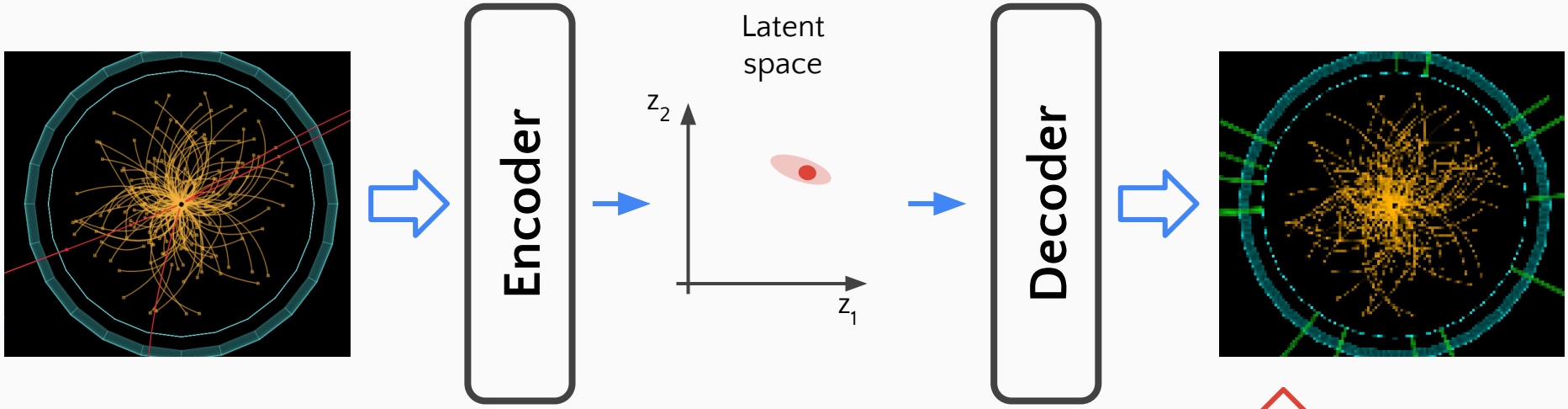


Evaluate:

One-side hypothesis test on the input-output distance

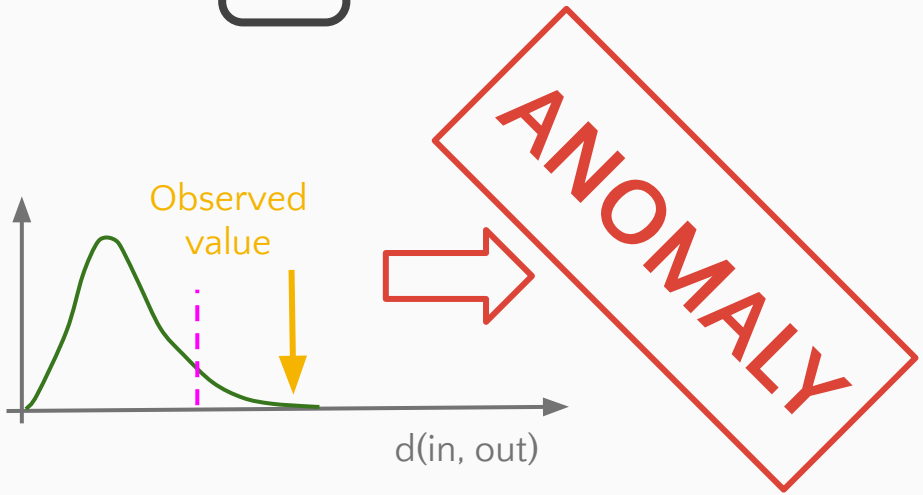


AE anomaly detection: Inference



Evaluate:

One-side hypothesis test on the input-output distance



Online anomaly detection: a model independent trigger

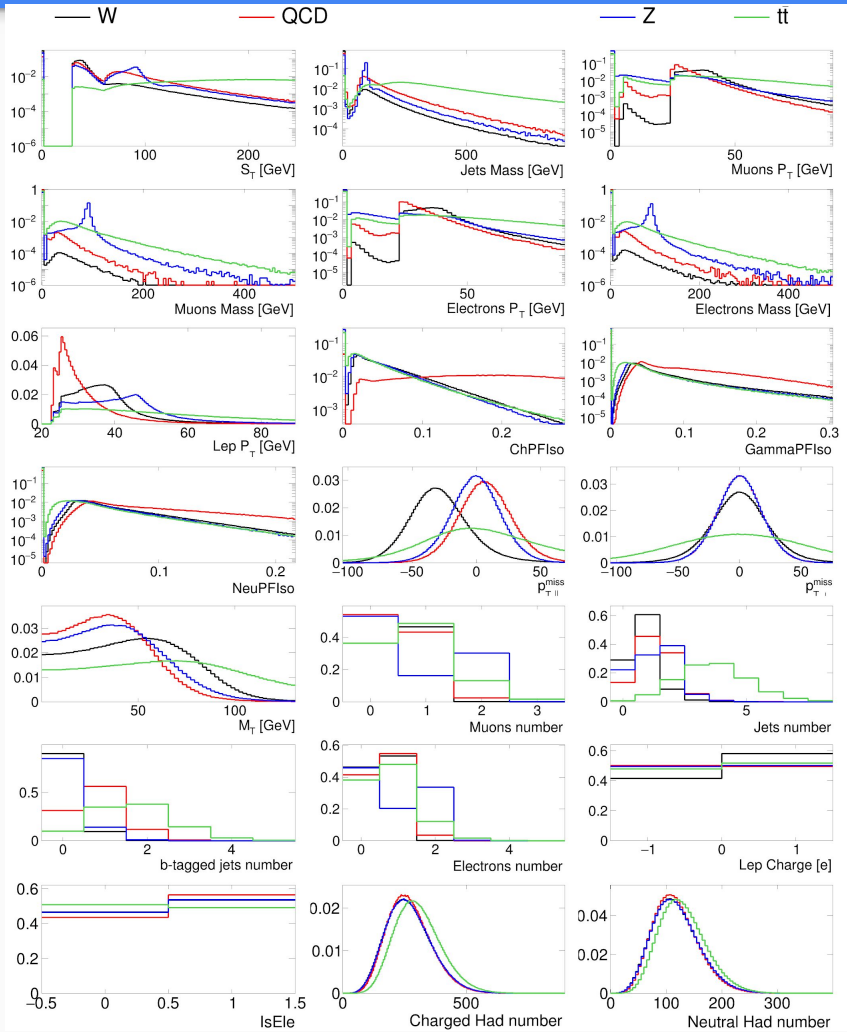
O. Cerri [a], T. Q. Nguyen[a], M. Pierini[b], M. Spiropulu[a], J. R. Vlimant[a]

[a] CalTech, [b] CERN

J. High Energy Phys. (2019) 2019: 36

Anomalies in lepton stream

- Stream of data with at least one interesting lepton (e or μ)
  - p_T > 23 GeV & ISO < 0.45
  - Too high rate to be tolerated at LHC
- SM contribution
  - W (59%), QCD (34%), Z (6%), tt (1%)
- Events represented by 21 high level features (HLF)
  - Broad general choice, not BSM tailored



Procedure

1. **Train** one (or more) VAE(s):
   a. Train on MC or **on data** (robust against signal injection – backup or ask)

2. Run the **VAE(s) in the High Level Trigger**
   a. Evaluate each event
   b. Acceptance threshold tuned to O(10) SM events/day

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

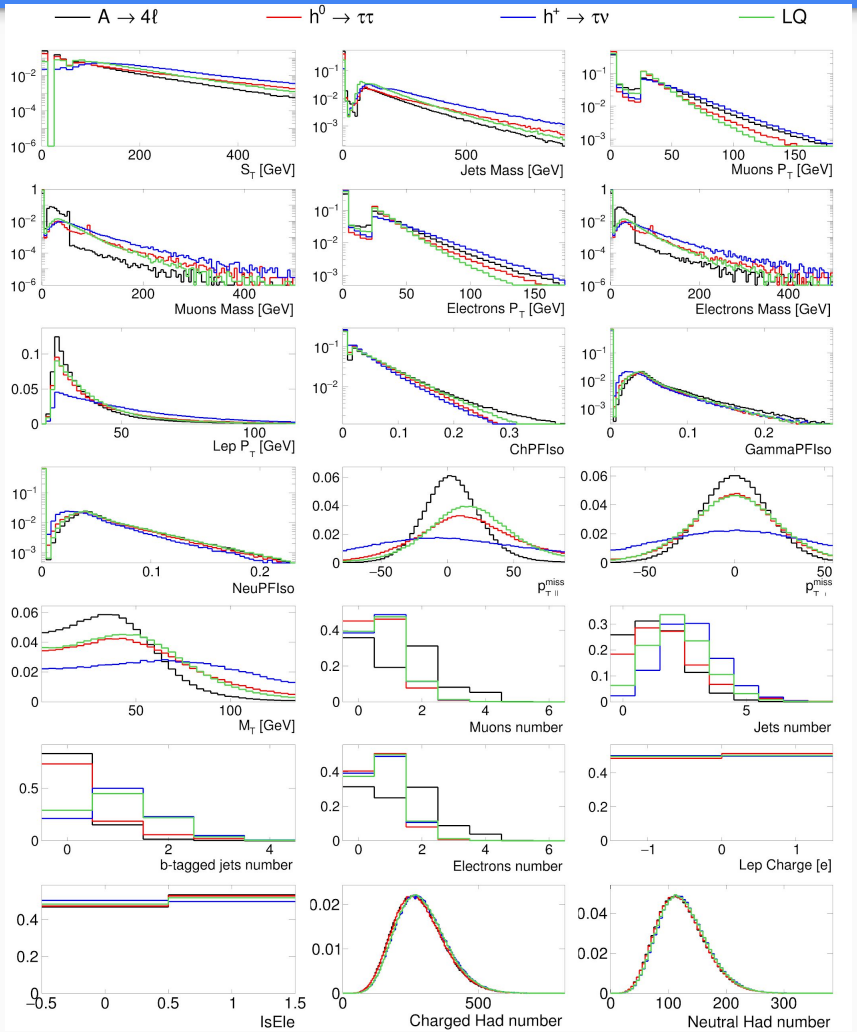
Remind: VAE does not see the BSM (if any) until it's evaluated on new events

BSM benchmark models

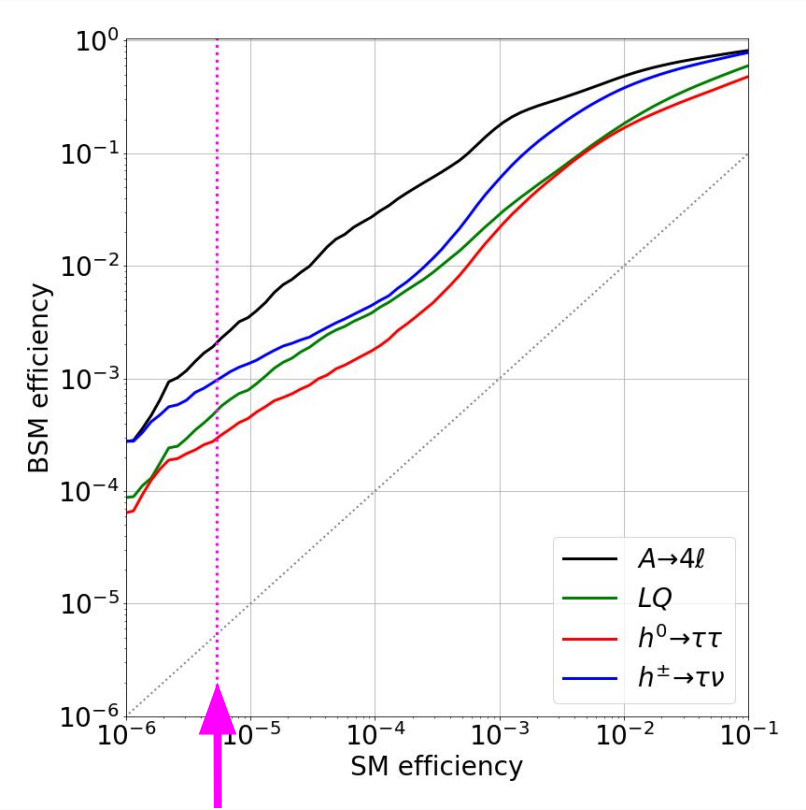
Light BSM which are usually very hard to trigger with standard strategies

- $A \rightarrow 4\ell$: neutral scalar, M = 50 GeV
- $LQ \rightarrow b\tau$: leptoquark, M = 80 GeV
- $h^0 \rightarrow \tau\tau$: neutral scalar, M = 60 GeV
- $h^{\pm} \rightarrow \tau\nu$: charged scalar, M = 60 GeV

BENCHMARKING ONLY, NOT USED FOR TRAINING



Results

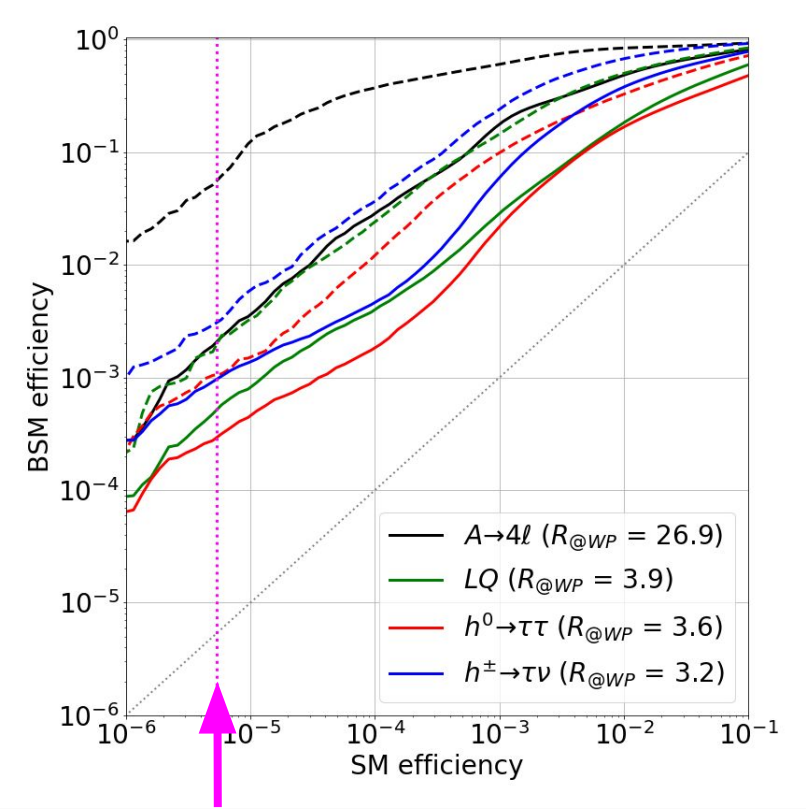


$\varepsilon_{SM} = 5.4 \cdot 10^{-6} \Leftrightarrow$ 30 evts/day

- (Unsupervised) VAE
  - Single one, trained only on SM
  - No info on BSM → less powerful
  - Applies to any BSM → more general

Results

--- Model dep. ——— VAE



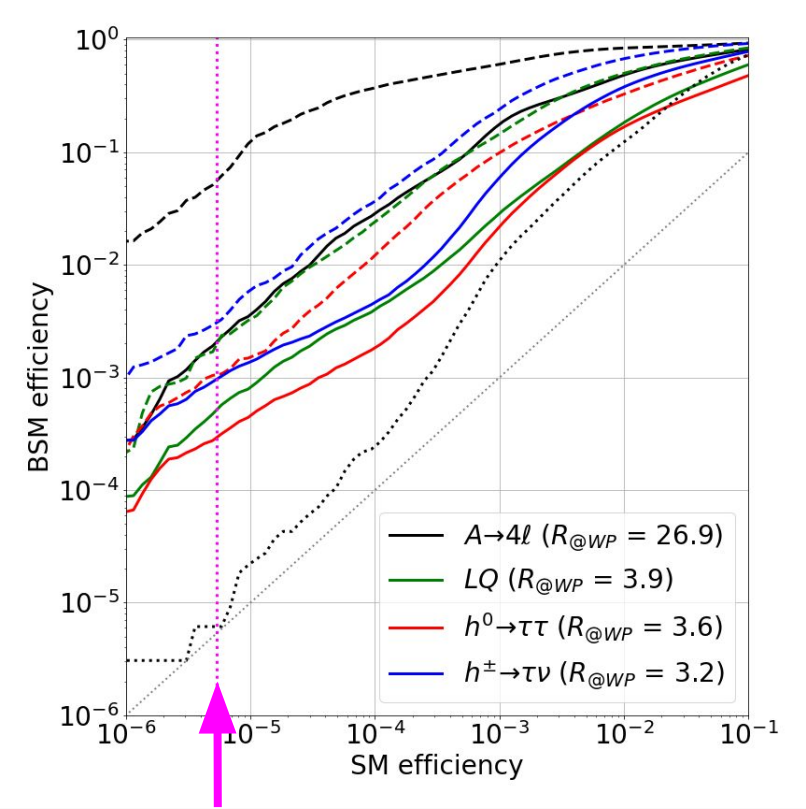
$\varepsilon_{SM} = 5.4 \cdot 10^{-6} \Leftrightarrow$ 30 evts/day

- (Unsupervised) VAE
  - Single one, trained only on SM
  - No info on BSM → less powerful
  - Applies to any BSM → more general
- (Supervised) Model-dependent classifiers
  - 4 algorithms, each trained on a specific BSM scenario
  - Set optimal performances

Results

--- Model dep. ——— VAE

··· Model dep. on a different model



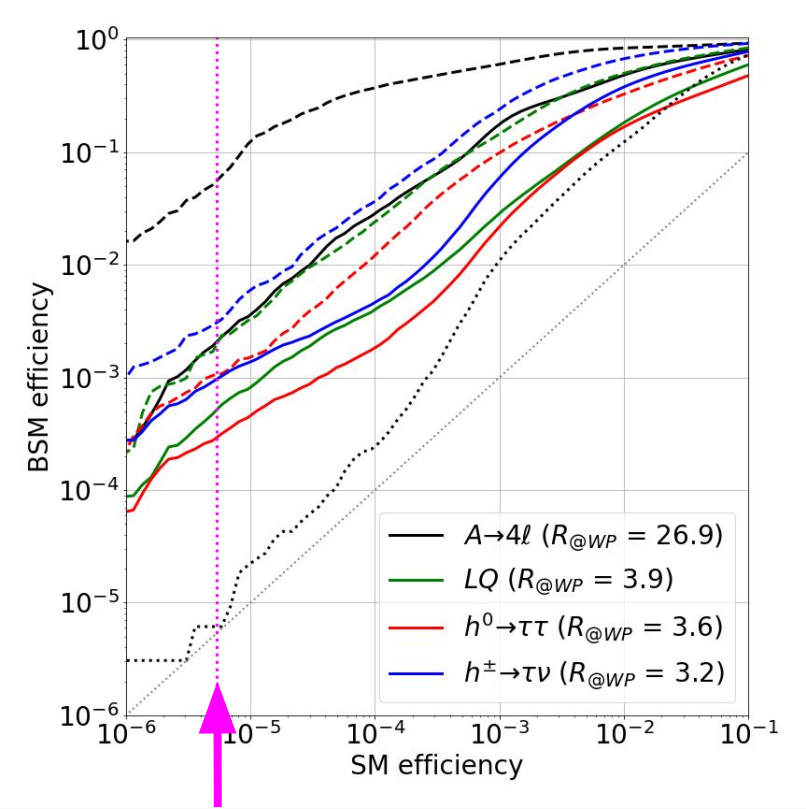
$\varepsilon_{SM} = 5.4 \cdot 10^{-6} \Leftrightarrow$ 30 evts/day

- (Unsupervised) VAE
  - Single one, trained only on SM
  - No info on BSM → less powerful
  - Applies to any BSM → more general
- (Supervised) Model-dependent classifiers
  - 4 algorithms, each trained on a specific BSM scenario
  - 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)

PRELIMINARY

Offline anomaly detection: a model independent analysis

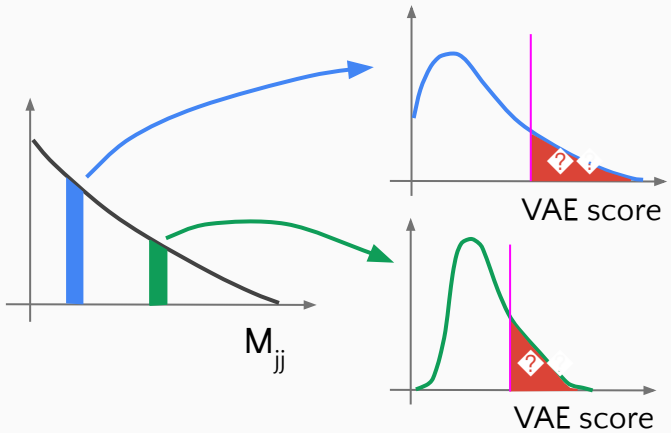
O. Cerri [a], J. M. G. Duarte [b], J. Ngadiuba [c], T. Q. Nguyen [a], M. Pierini [c], M. Spiropulu [a], J. R. Vlimant [a], K. A. Wozniak [c,d]

[a] CalTech, [b] Fermilab, [c] CERN, [d] U. of Vienna

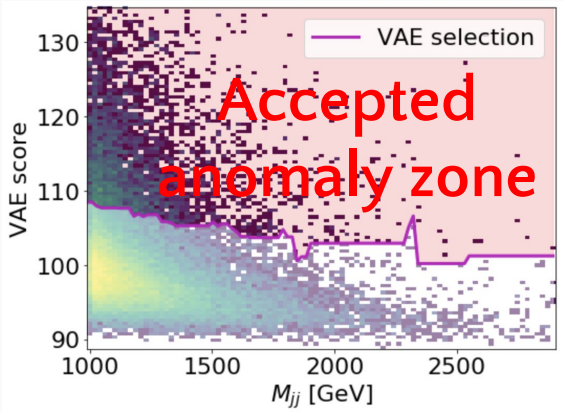
arXiv:19XX.XXXXX

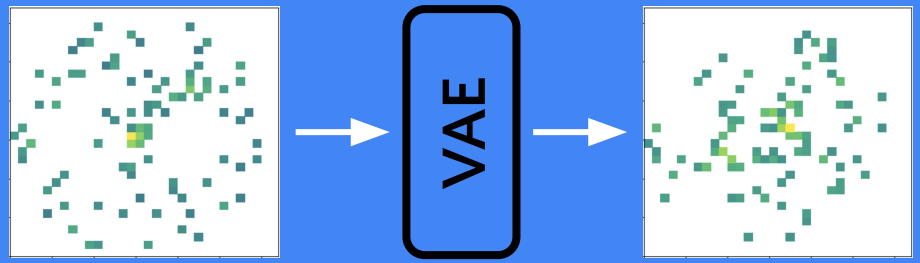
Extend an existing supervised search

Di-Jet bump hunt



Constant efficiency





1. Train a VAE on SM jets from a control region [1,2]

2. Choose a VAE selection with constant efficiency for the control region as a function of a variable of interest (Mjj) [3]

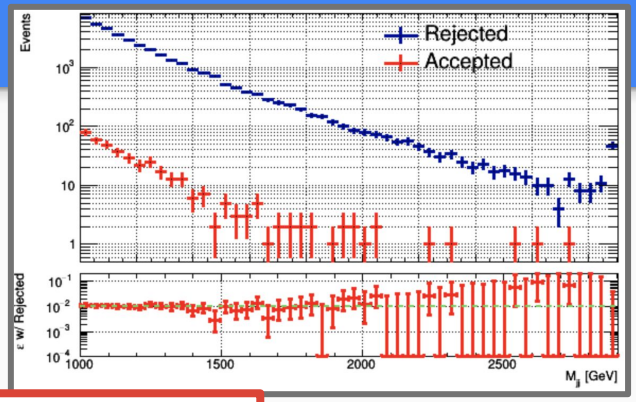
[1] Heimel et al. https://arxiv.org/abs/1808.08979
[2] Franco et al. http://arxiv.org/abs/arXiv:1808.0829
[3] Dolen et al. https://arxiv.org/abs/1603.00027

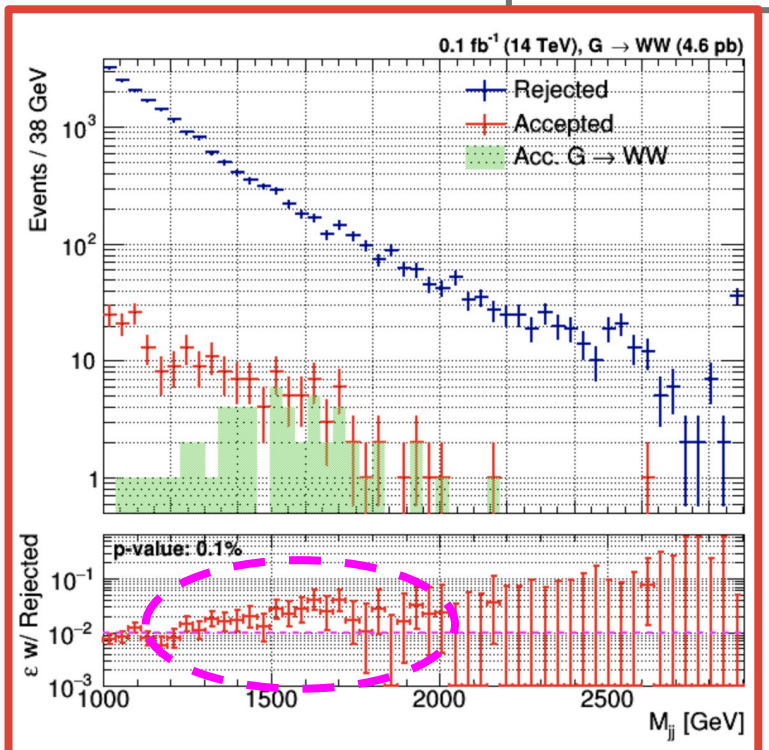
New search strategy

- Selection force the rejected and accepted events to have the same mass spectrum
  - In control region by construction
  - In signal region, if no BSM is present

- BSM-agnostic search:
  - Hypothesis test comparing rejected and accepted shapes
  - Improve state-of-art [1]

SM side band (training region)





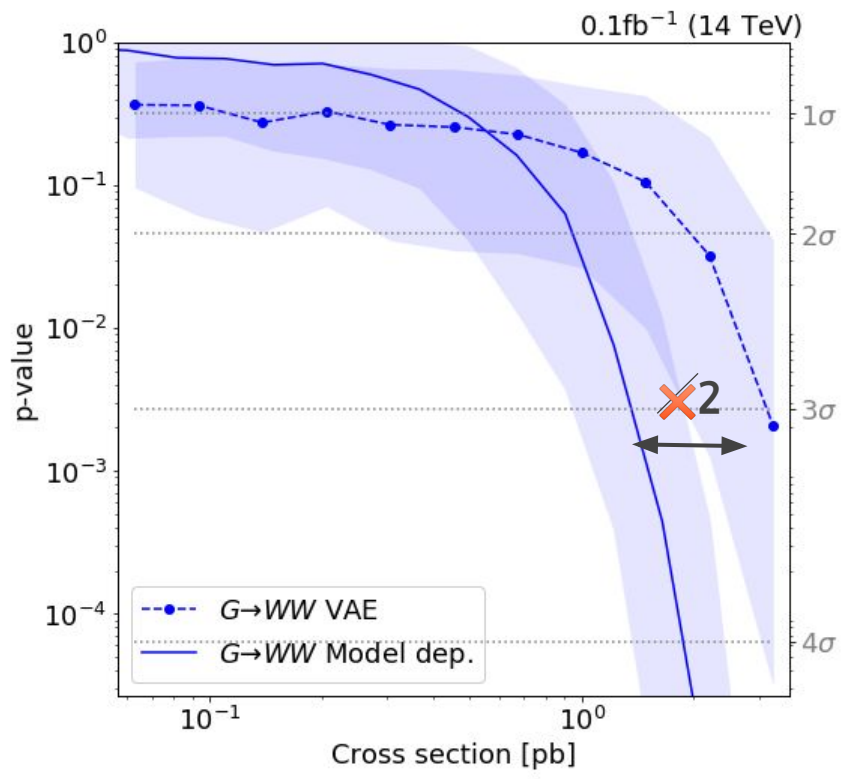
Signal region

[1] Collins et al. https://arxiv.org/abs/1902.02634

Results

- Di-jet events
  - Injected 1.5 TeV graviton
  - Control region $|\Delta\eta_{jj}|>1.4$
- Image-based AE
- Defined threshold to retain 1% of SM events as anomalous
  - Model independent: comparing standard to anomaly M_{jj} spectrum
- Model Dependent bump hunt

Discovery significance



Only a factor 2 drop wtr to optimal performances

Conclusions

- **VAE as model-independent BSM trigger**
  - Can be **trained on data**
  - Select otherwise lost anomalous events for further studies
  - Allows benchmark models to be **probed down to 10-100 pb cross section**

- **VAE based model-independent analysis**
  - New method proposed to turn each LHC search with jets into a model-independent test
  - Provides **data driven** estimation of the background shape
  - Define anomalies and provide a test to look for non-SM processes

- **Alternative strategies** might open new directions

BACKUP

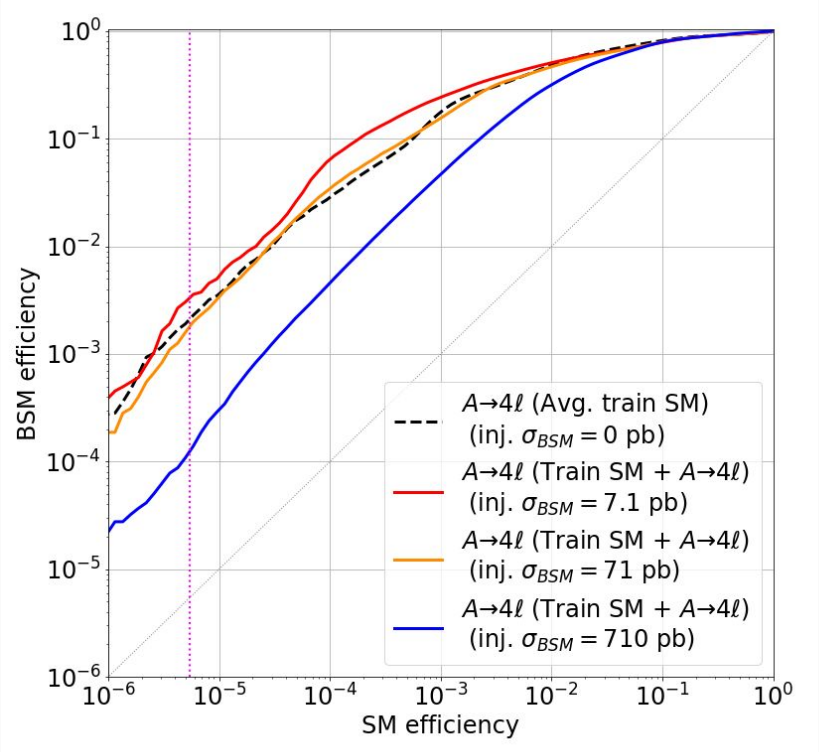
Train on data

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

- Train on a dataset with signal injected:

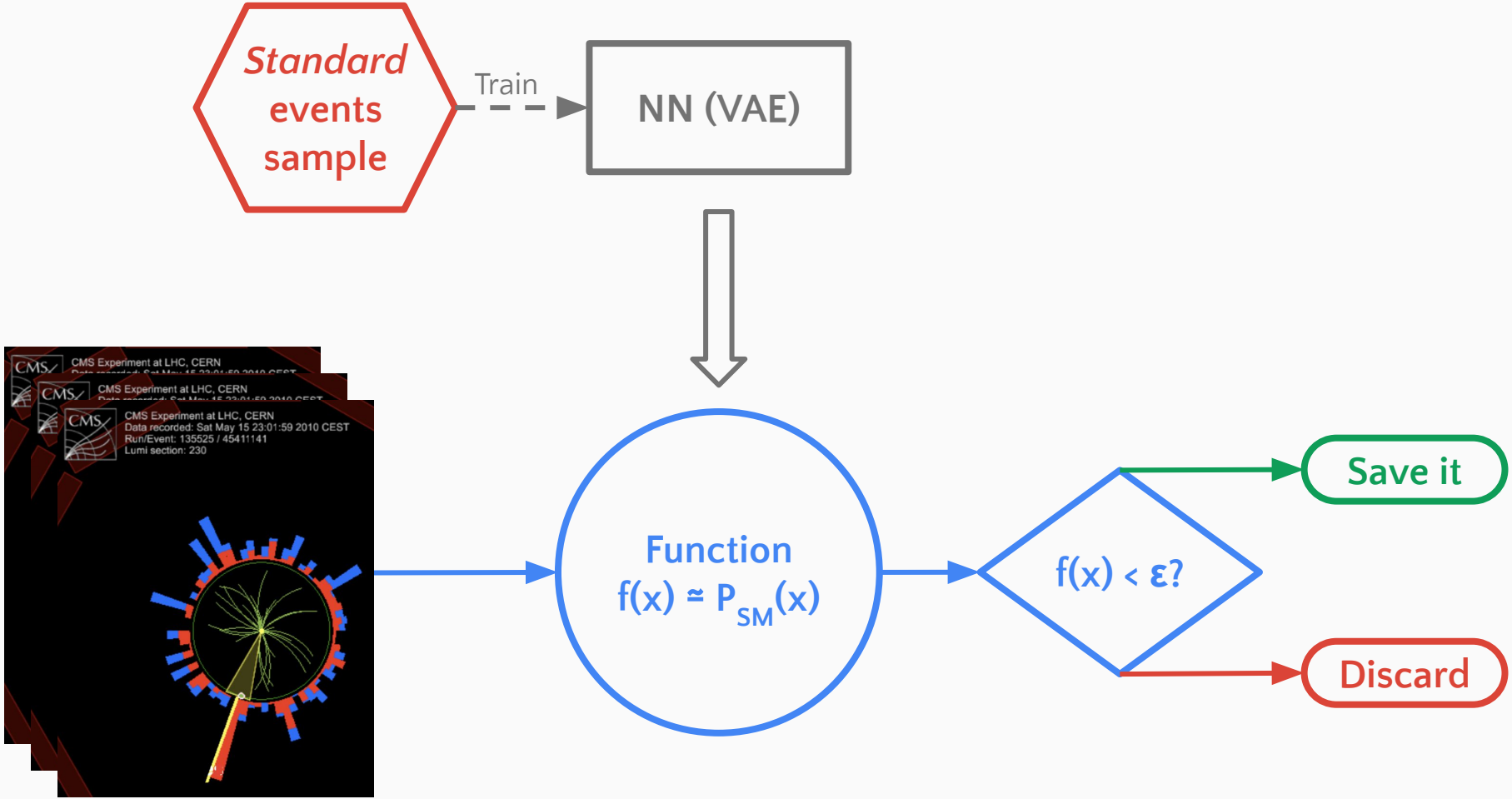
Injected evts	Training set fraction	VAE selected evts/month	Anomaly fraction
700	$2 \cdot 10^{-4}$	134	12%
7k	$2 \cdot 10^{-3}$	957	48%
70k	$2 \cdot 10^{-2}$	6	0.6%

- 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

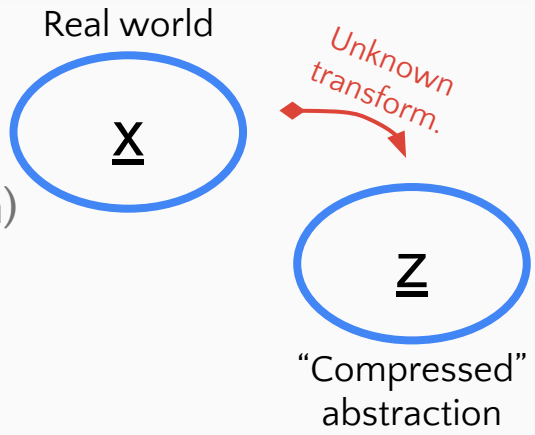

Standard events sample
Train
NN (VAE)
CMS Experiment at LHC, CERN
Data recorded: Sat May 15 23:01:59 2010 CEST
Run/Event: 135525 / 45411141
Lumi section: 230
Function
f(x) ≃ P_SM(x)
f(x) < ε?
Save it
Discard

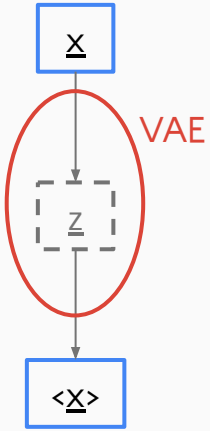
Workflow

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)

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

Regularization term:

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

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

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

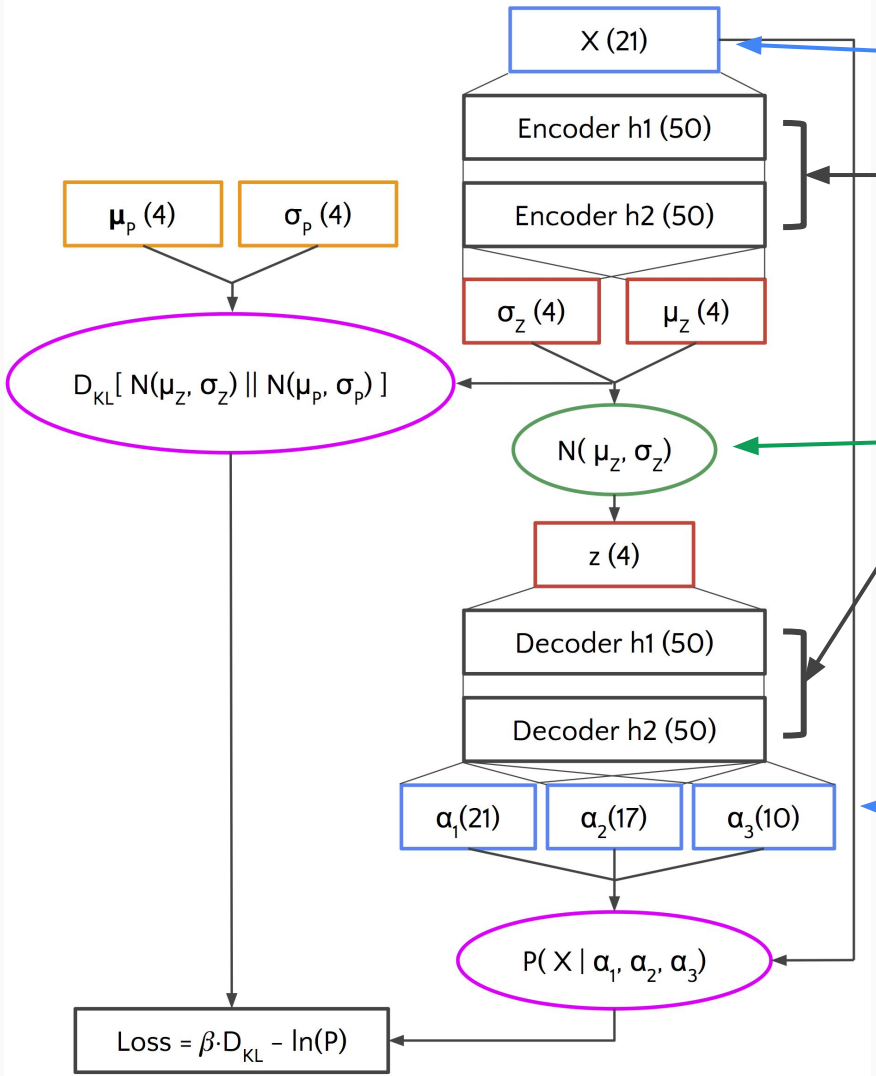
The encoder function $g_e: \underline{x} \longrightarrow \alpha_e$ gives the value of the $\underline{z}$ distribution parameters

!!: $\underline{x}$ and $\underline{z}$ are swapped w.t.r. 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 $g_d: \underline{z} \longrightarrow \alpha_d$ gives the value of the $\underline{x}$ distribution parameters



Input (x) : 21 HLF (H_T, MET, nJets, ...)

2 dense hidden layer (50 neurons)

4-dim latent (z) space

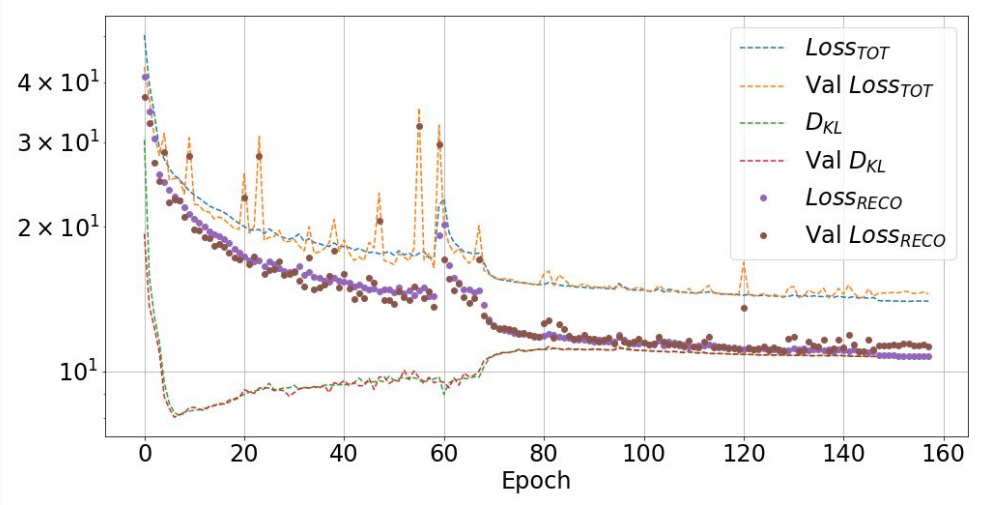
- Trainable mean and sigma

x' pdf: max 1-3 parameters

- Gaussian
- Binomial
- ...

A whole art exist in choosing the functional form

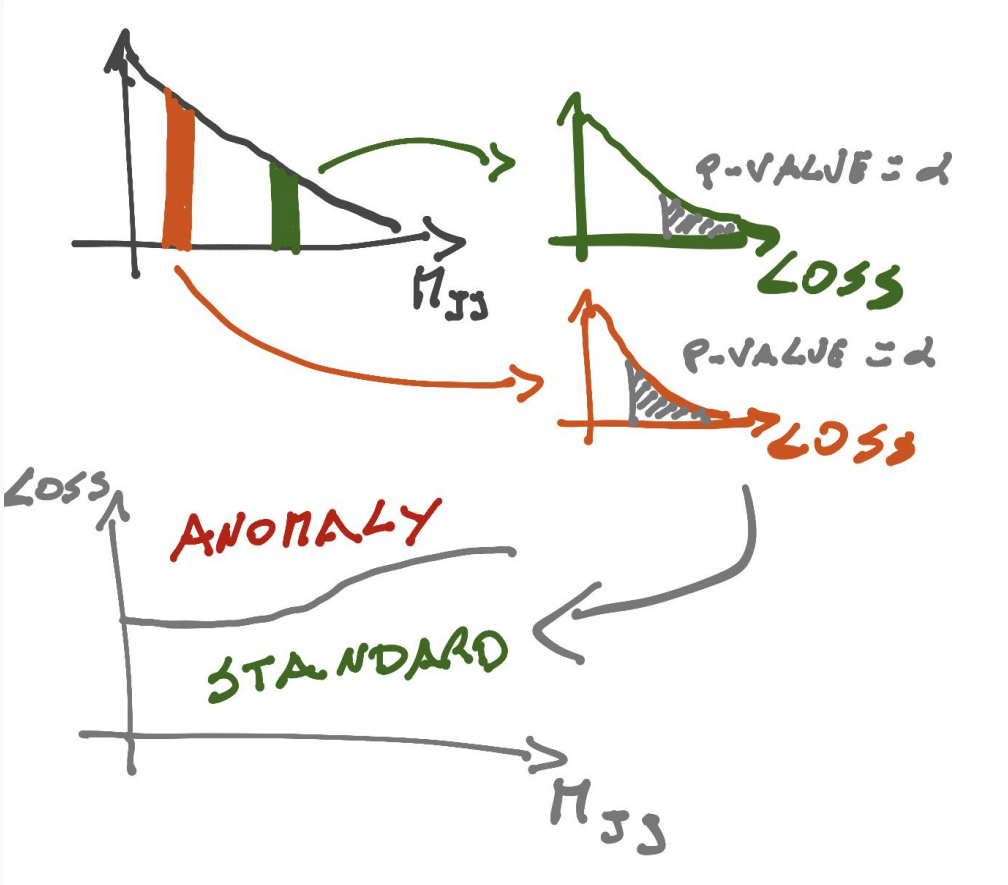
Training: not a easy beast



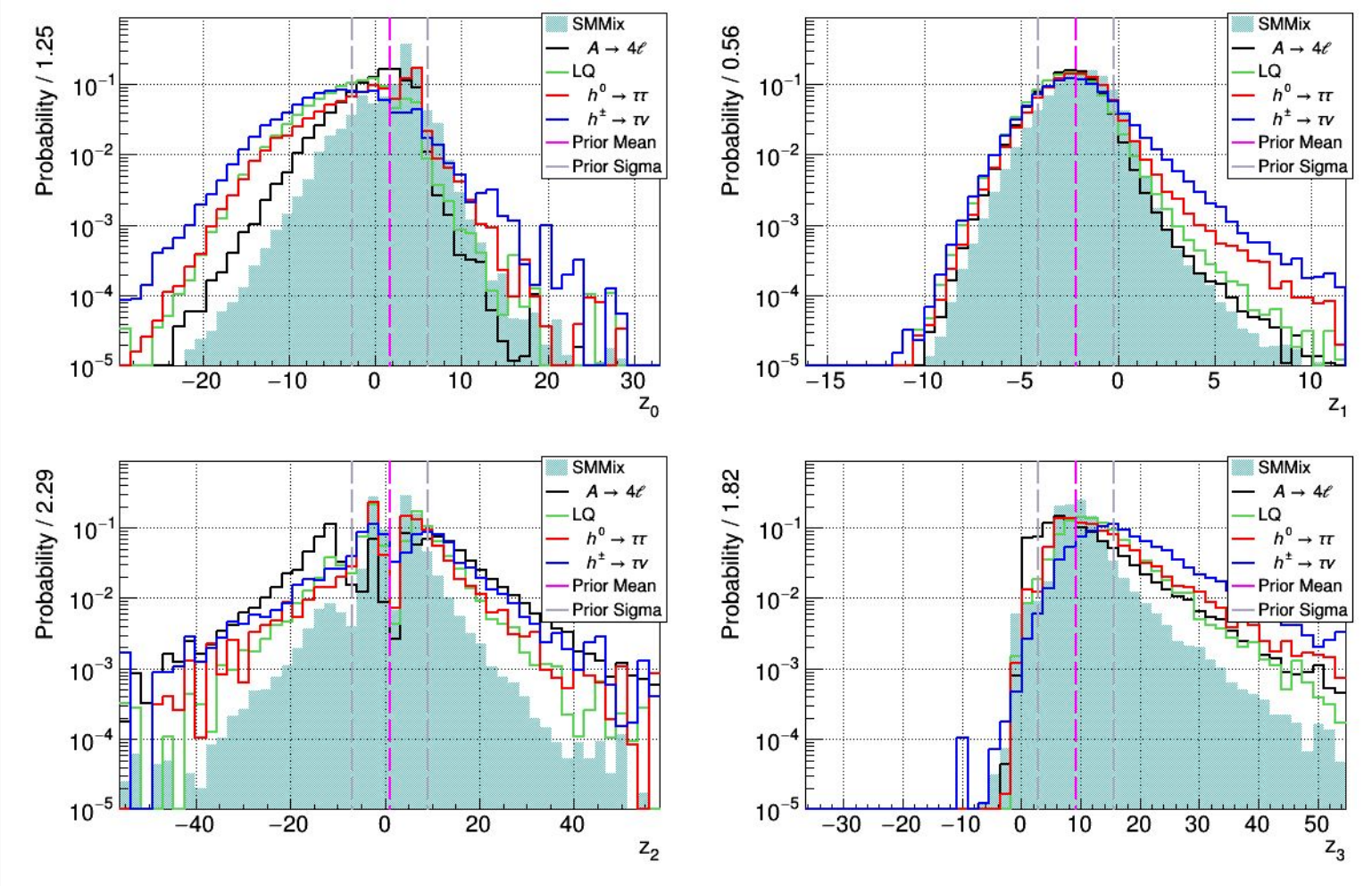
- Optimizer
  - Adam
  - Callbacks
- Samples
  - 3.5 M event for training
  - 3.5 M for validation
  - # evt/# par >> 10
- The training
  - Not long, about 1h
  - Spike not unusual
  - Delicate equilibrium of training parameters

VAE for new search strategies

1. Start from an existing supervised search with jets (e.g., dijet bumphunt)
2. Train a VAE to "learn" QCD jets on a control region (e.g. $|\Delta\eta_{jj}|>1.4$)
   a. Heimel et al. https://arxiv.org/abs/1808.08979
   b. Franco et al. http://arxiv.org/abs/arXiv:1808.0829
3. Modulate the cut on the VAE loss as a function of a variable of interest (Mjj) using DDT
   a. Selected events -> anomaly-enriched sample
   b. Rejected events -> background control region



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.

Output pdf

Clipped Log-normal + δ function: used to describe S_T, M_J, p_T^μ, M_μ, p_T^e, M_e, p_T^ℓ, ChPFIso, NeuPFIso and GammaPFIso:

$$P(x \mid \alpha_1, \alpha_2, \alpha_3) = \begin{cases} \alpha_3 \delta(x) + \frac{1-\alpha_3}{x\alpha_2\sqrt{2\pi}} \exp\left(\frac{(\ln x - \alpha_1)^2}{2\alpha_2^2}\right) \; for\; x \geq 10^{-4} \\ 0 \; for\; x < 10^{-4} \end{cases} . \tag{9}$$

Gaussian: used for $p_{T,\parallel}^{\text{miss}}$ and $p_{T,\perp}^{\text{miss}}$:

$$P(x \mid \alpha_1, \alpha_2) = \frac{1}{\alpha_2\sqrt{2\pi}} \exp\left(\frac{(x-\alpha_1)^2}{2\alpha_2^2}\right) . \tag{10}$$

Truncated Gaussian: a Gaussian function truncated for negative values and normalized to unit area for $X > 0$. Used to model M_T:

$$P(x \mid \alpha_1, \alpha_2) = \Theta(x) \cdot \frac{1 + 0.5 \cdot (1 + \text{erf}\frac{-\alpha_1}{\alpha_2\sqrt{2}})}{\alpha_2\sqrt{2\pi}} \exp\left(\frac{(x-\alpha_1)^2}{2\alpha_2^2}\right) . \tag{11}$$

Discrete truncated Gaussian: like the truncated Gaussian, but normalized to be evaluated on integers (i.e. $\sum_{n=0}^{\infty} P(n) = 1$). This function is used to describe N_μ, N_e, N_b and N_J. It is written as:

$$P(n \mid \alpha_1, \alpha_2) = \Theta(x) \left[\text{erf}\left(\frac{n + 0.5 - \alpha_1}{\alpha_2\sqrt{2}}\right) - \text{erf}\left(\frac{n - 0.5 - \alpha_1}{\alpha_2\sqrt{2}}\right)\right] \mathcal{N} , \tag{12}$$

where the normalization factor $\mathcal{N}$ is set to:

$$\mathcal{N} = 1 + \frac{1}{2}\left(1 + \text{erf}\left(\frac{-0.5 - \alpha_1}{\alpha_2\sqrt{2}}\right)\right) . \tag{13}$$

Binomial: used for `IsEle` and lepton charge:

$$P(n \mid p) = \delta_{n,m} p + \delta_{n,l}(1-p) \tag{14}$$

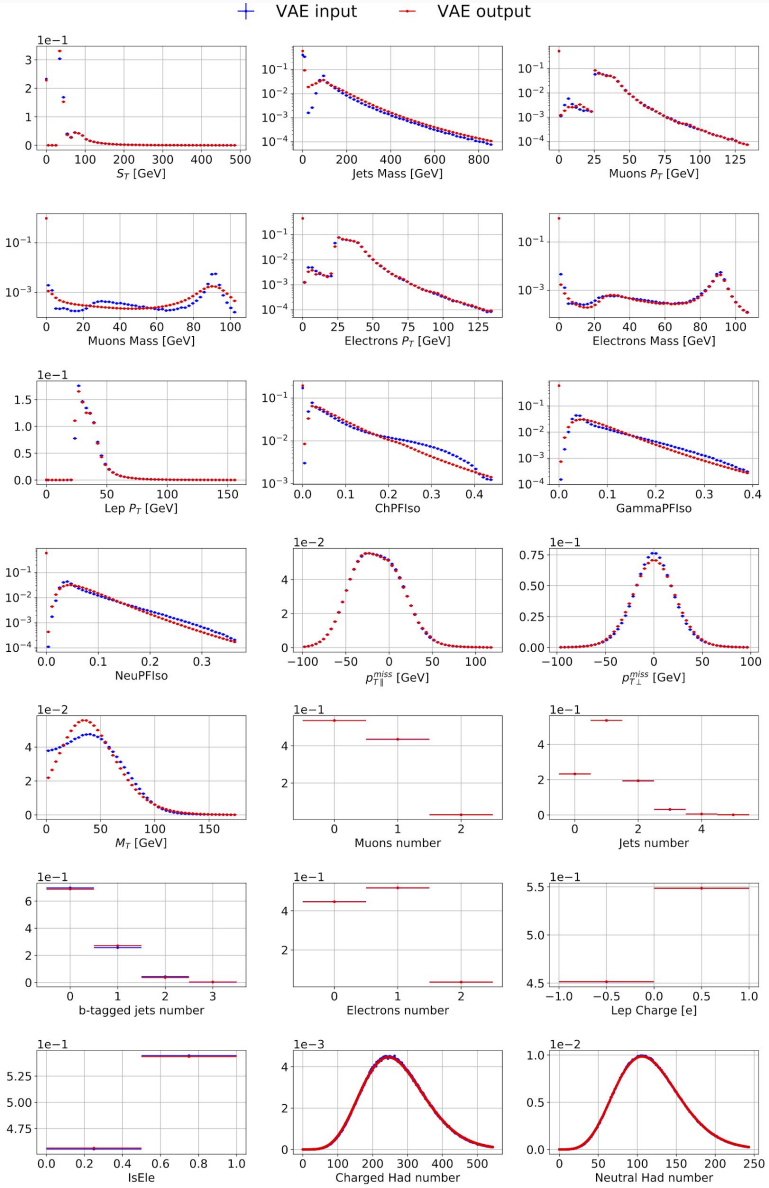
where m and l are the two possible values of the variable (0 or 1 for `IsEle` and -1 or 1 for lepton charge) and $p = C_{\tanh}(\alpha_1)$

Poisson: used for charged-particle and neutral-hadron multiplicities:

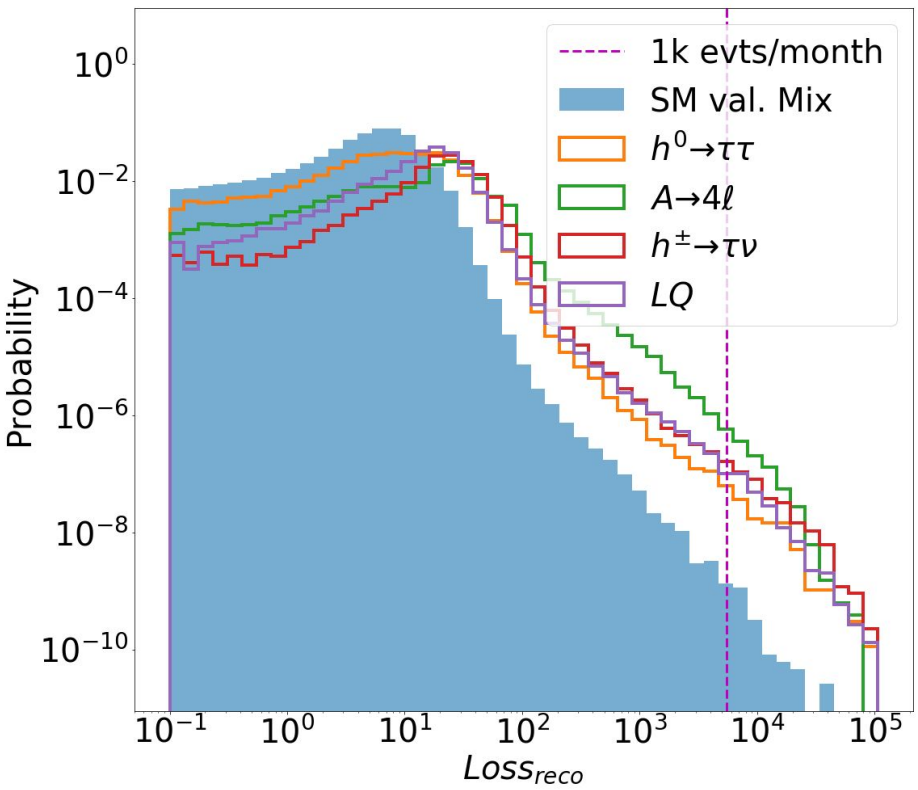
$$P(n \mid \mu) = \frac{\mu^n e^{-\mu}}{\Gamma(n+1)} \tag{15}$$

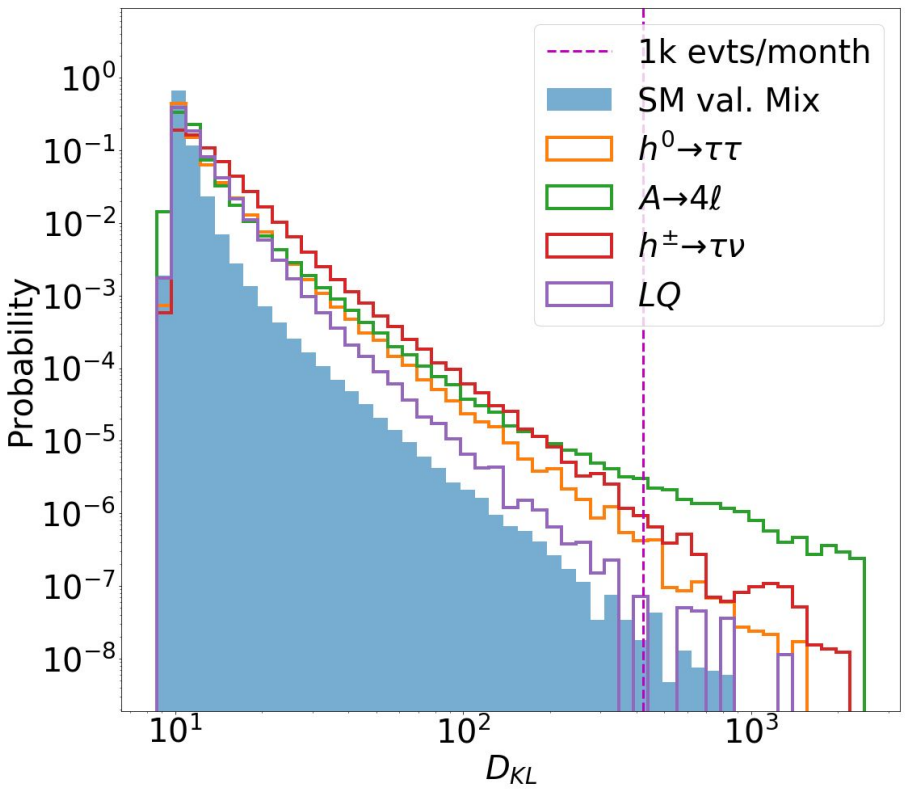
where $\mu = \text{p-ISRLu}(\alpha_1)$.

VAE auto-encoding cross-check

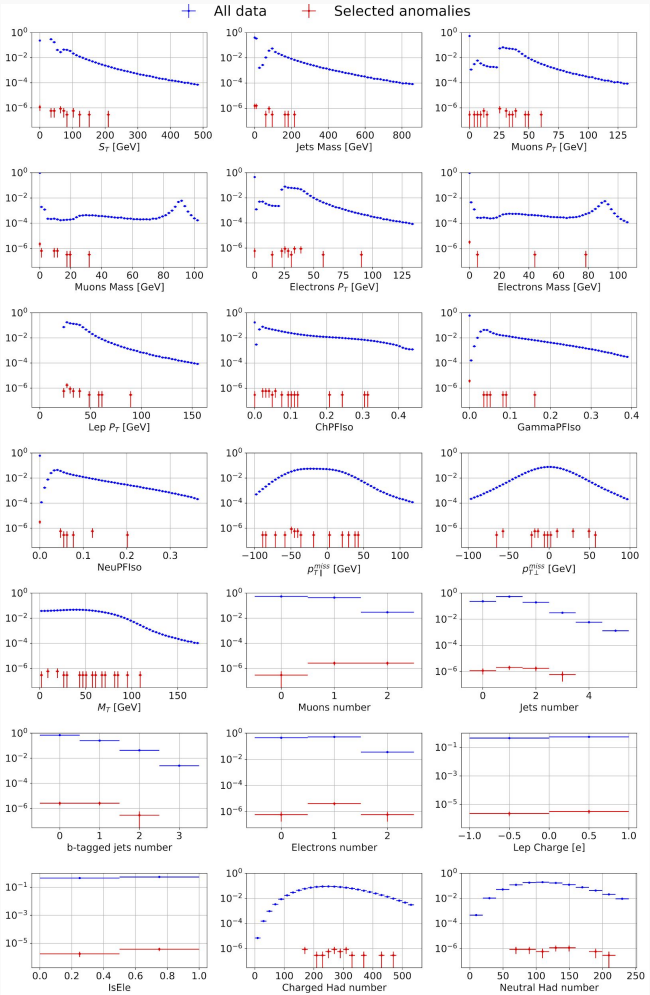


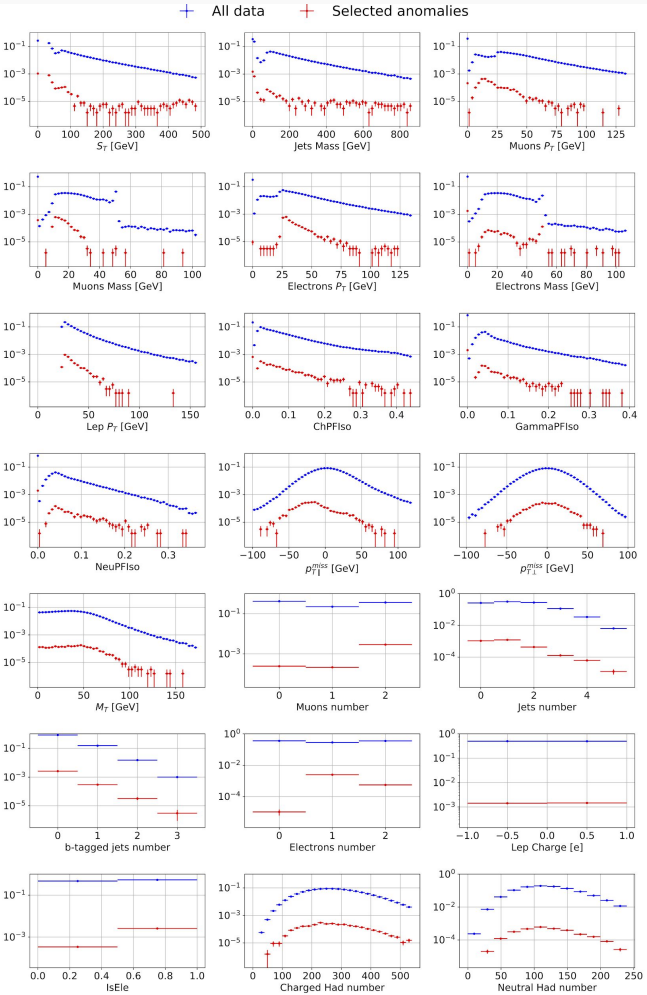
Loss distributions



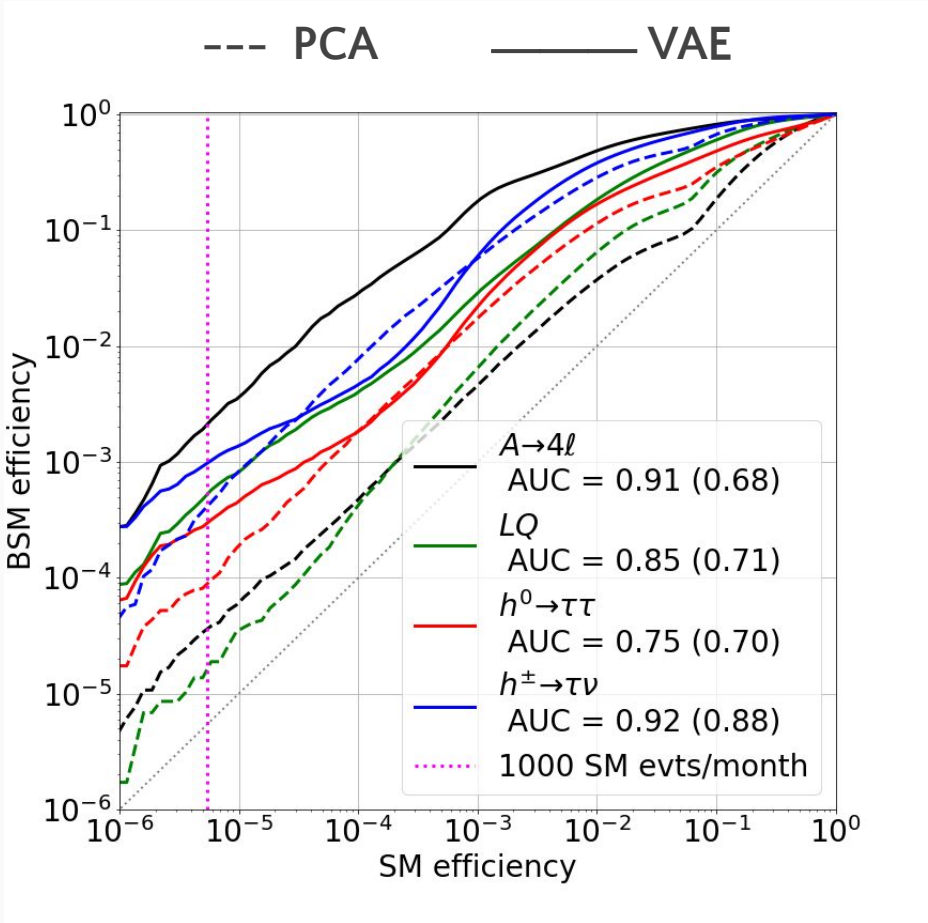


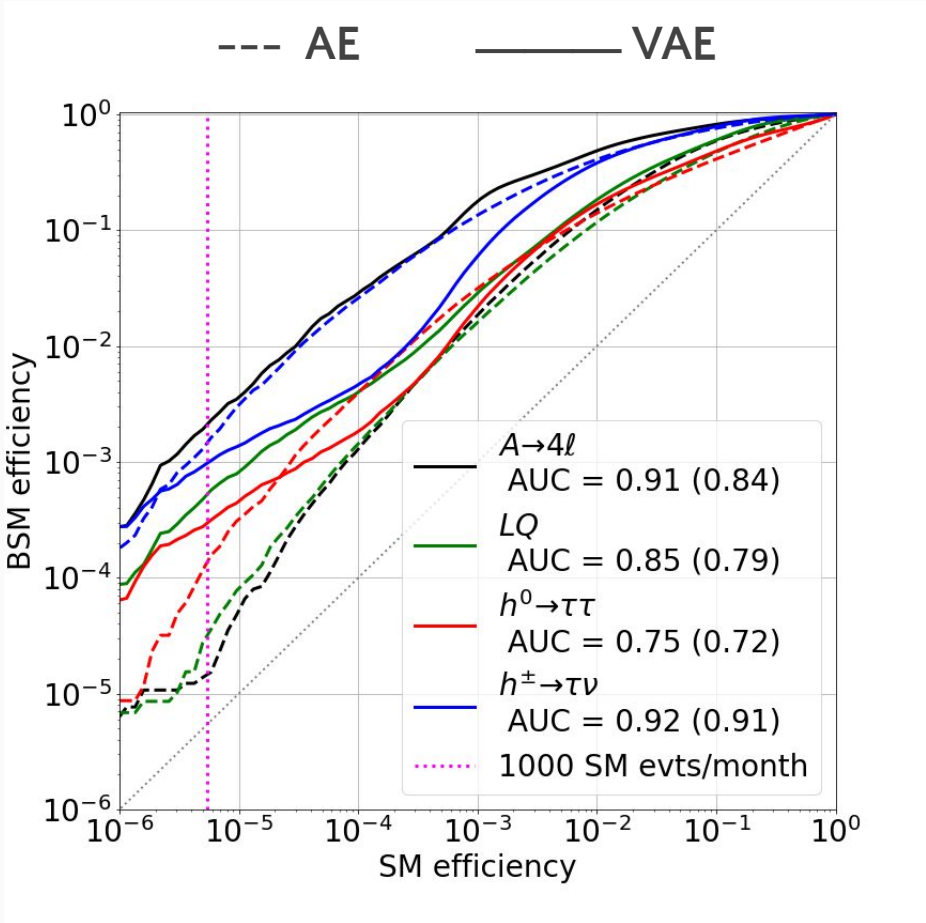
Not a tail-cut algorithm





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	$A \to 4\ell$	$h \to \tau\tau$	$h \to \tau\nu$	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

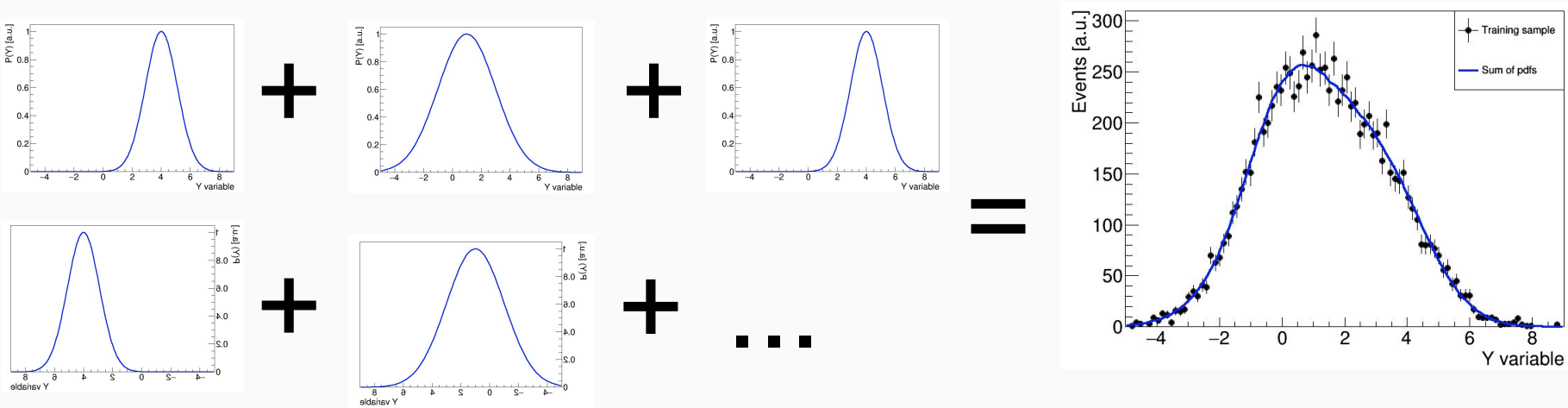
The great advantage of VAE is not only the ability to select BSM events but also to produce a high purity sample

Checking the convergence: sum of pdfs

High input dimension ⇒ Global convergence check

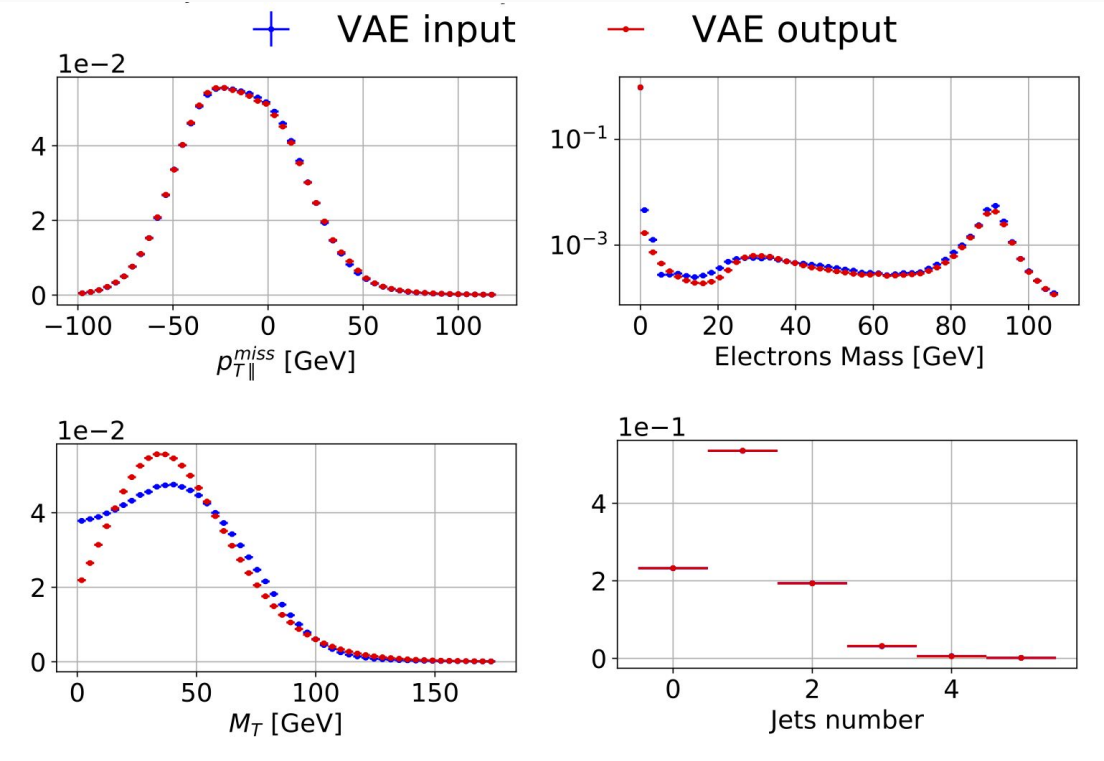
Predicted pdf for the single event

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



Convergence check: SM auto-encoding

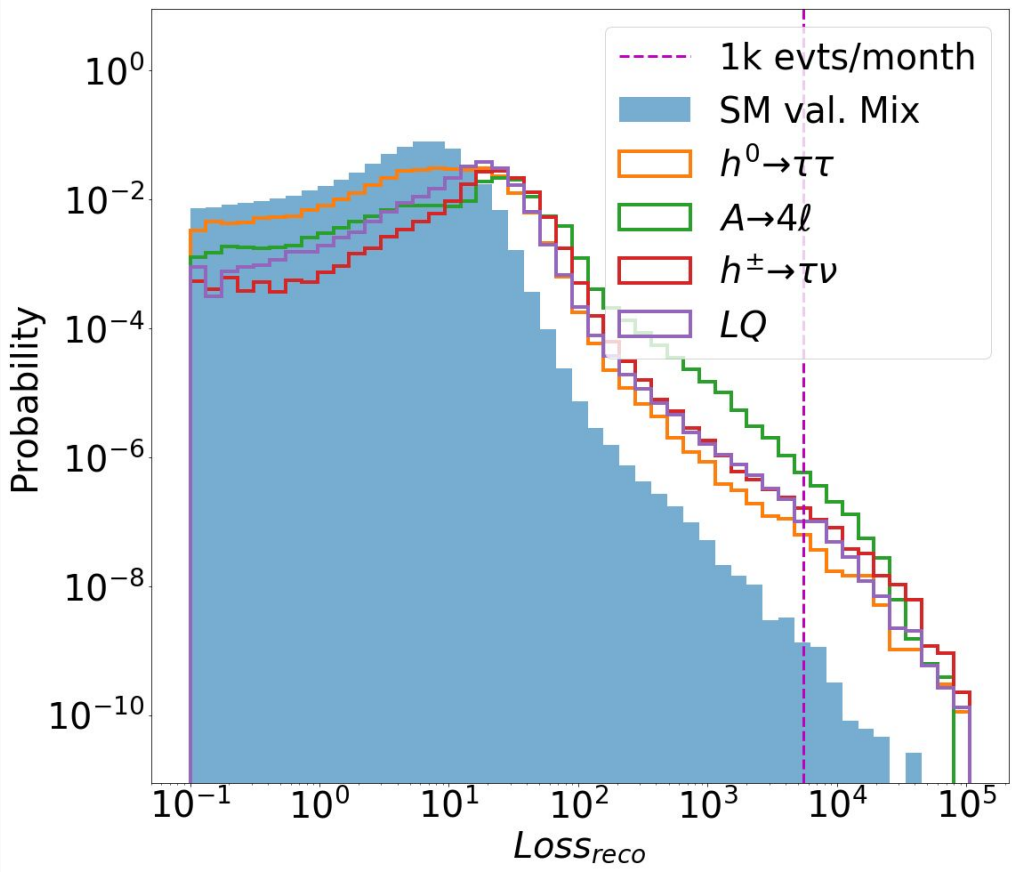
- Verifying encoding-decoding on validation set
  - Distributions of input vs generated from decoder
- Good agreement, with small discrepancy here and there
- Best autoencoder is not necessarily the best anomaly detector



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



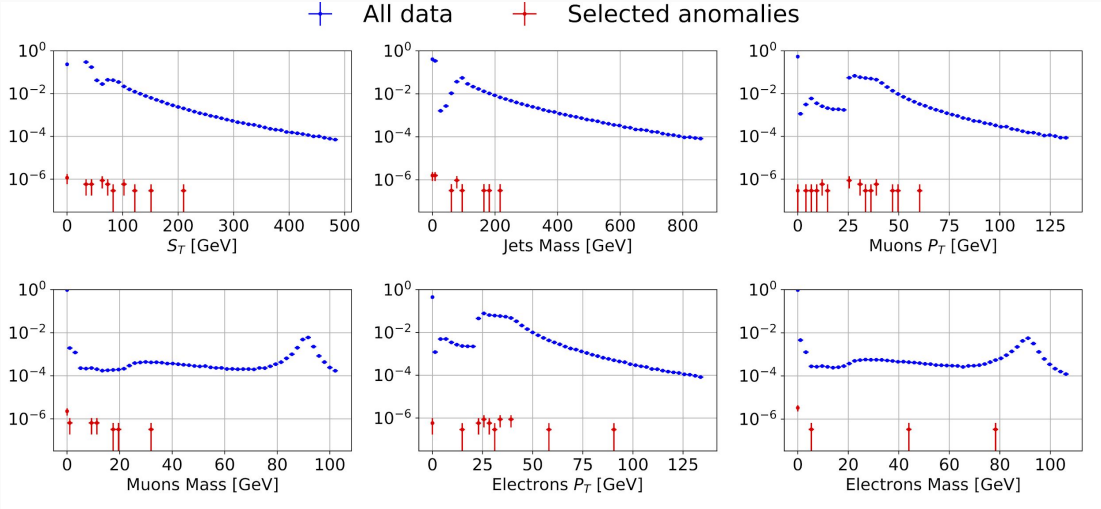
$Loss_{reco}$ used as test statistics.

Not a tail-cut algorithm

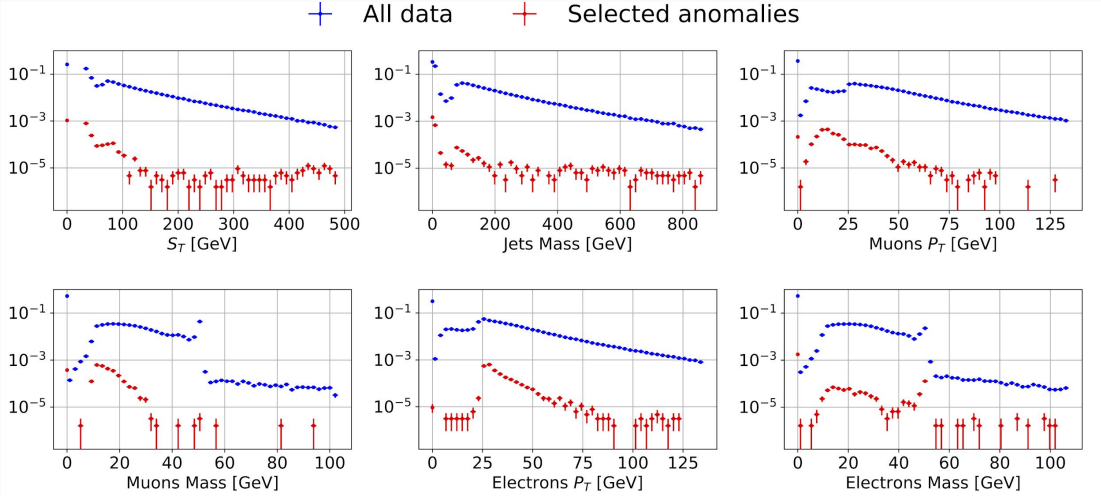
- Selected events stand on the core of 1D distributions

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

SM Mix

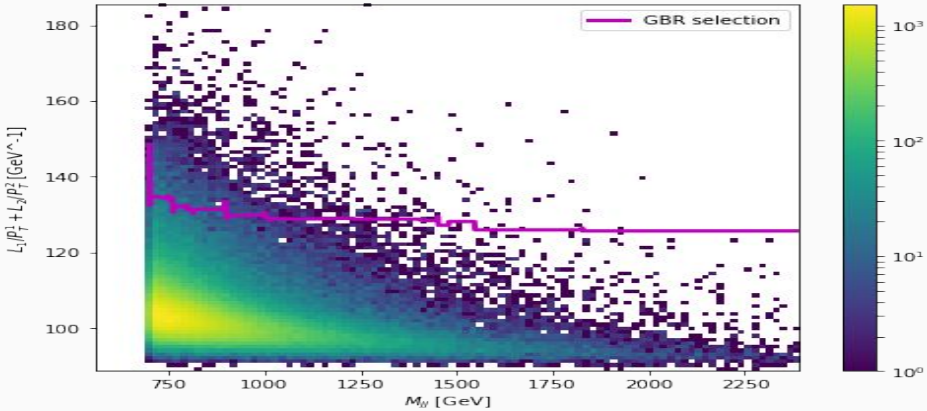


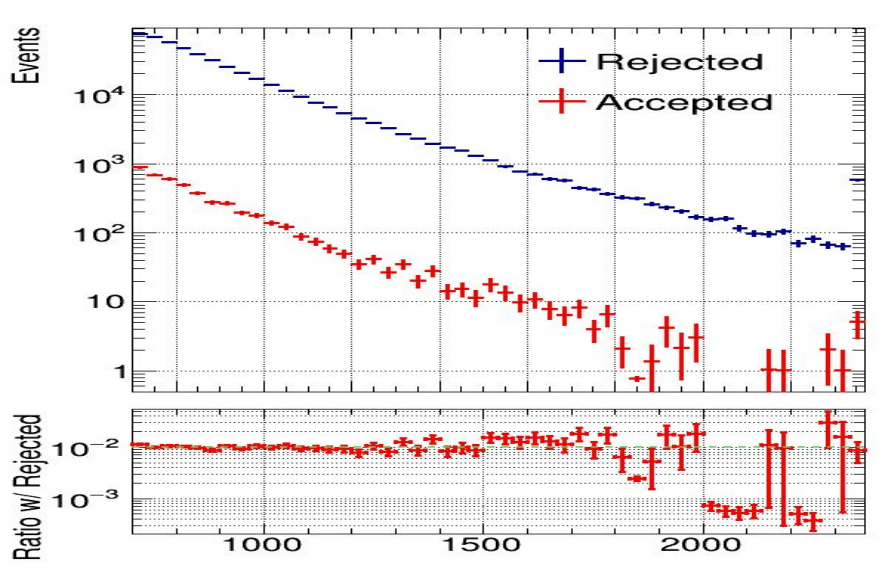
$A \to 4\ell$



Results

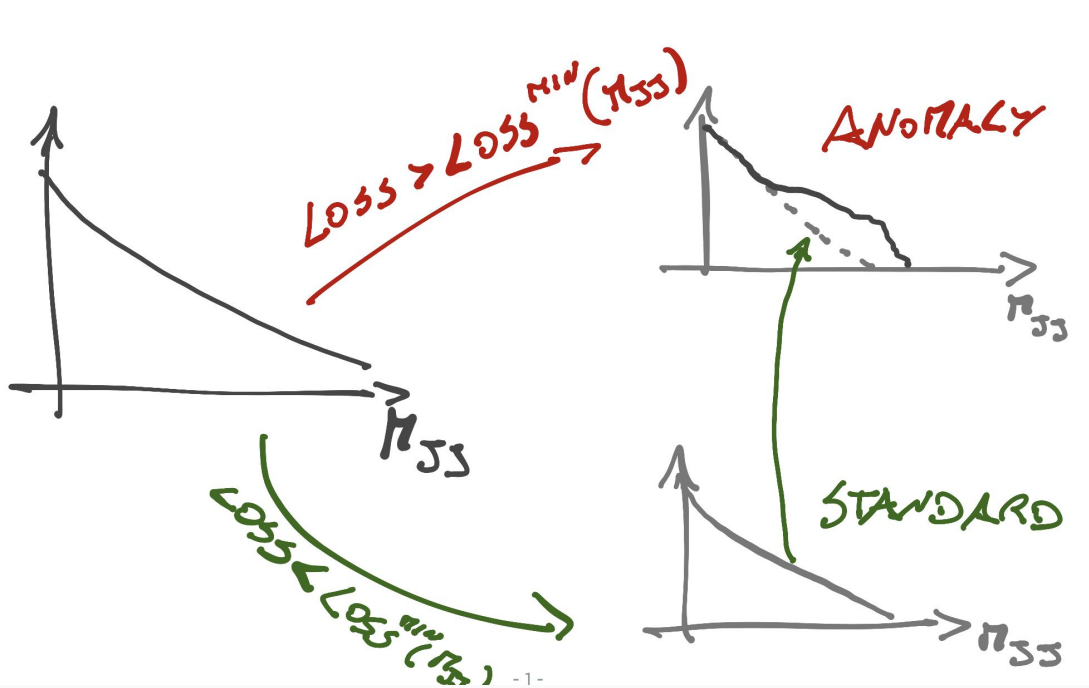
- Applied to a sample of Dijet events
- Used Image-based Autoencoders
- Defined threshold to retain 1% of the events as anomalous
  - Model Dependent: Bump Hunt on inclusive sample
  - Model Independent: GoF comparing standard to anomaly Mjj distributions
  - (Not Shown) VAE-enhanced Model Dependent: Bump Hunt in the anomaly Mjj distribution (using standard to predict background)





New search strategies

- BSM-agnostic search: goodness-of-fit test comparing the two shapes
- BSM-specific search: classic search with simultaneous fit (e.g., bump hunt)
- Similar to [Collins et al (COwLa)]
  - but uses data only once
  - generalizes to any analysis with jets
  - no assumption on signal shape

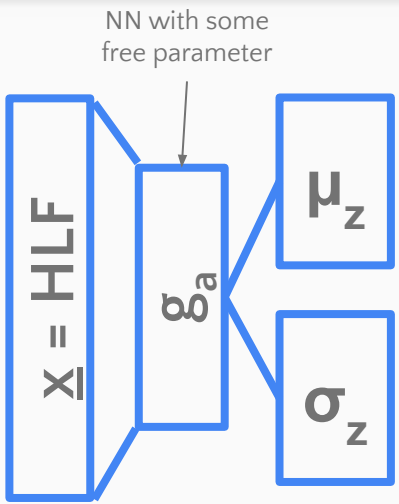


The Variational Auto-Encoder

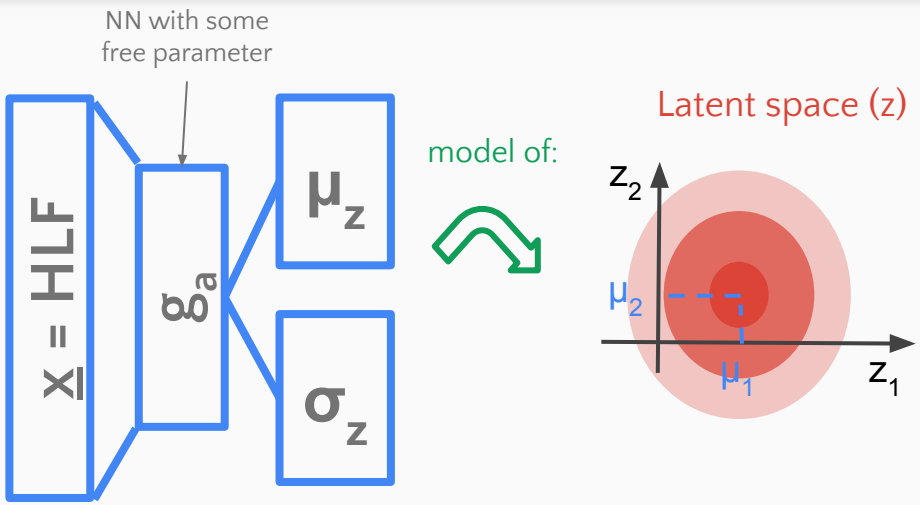
The Variational Auto-Encoder

$\underline{x}$ = HLF

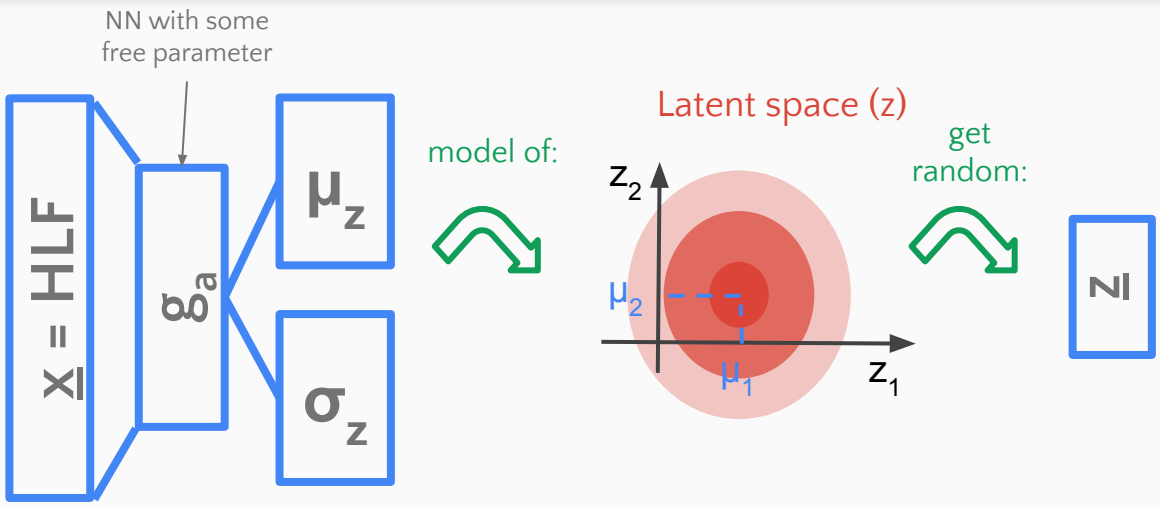
The Variational Auto-Encoder



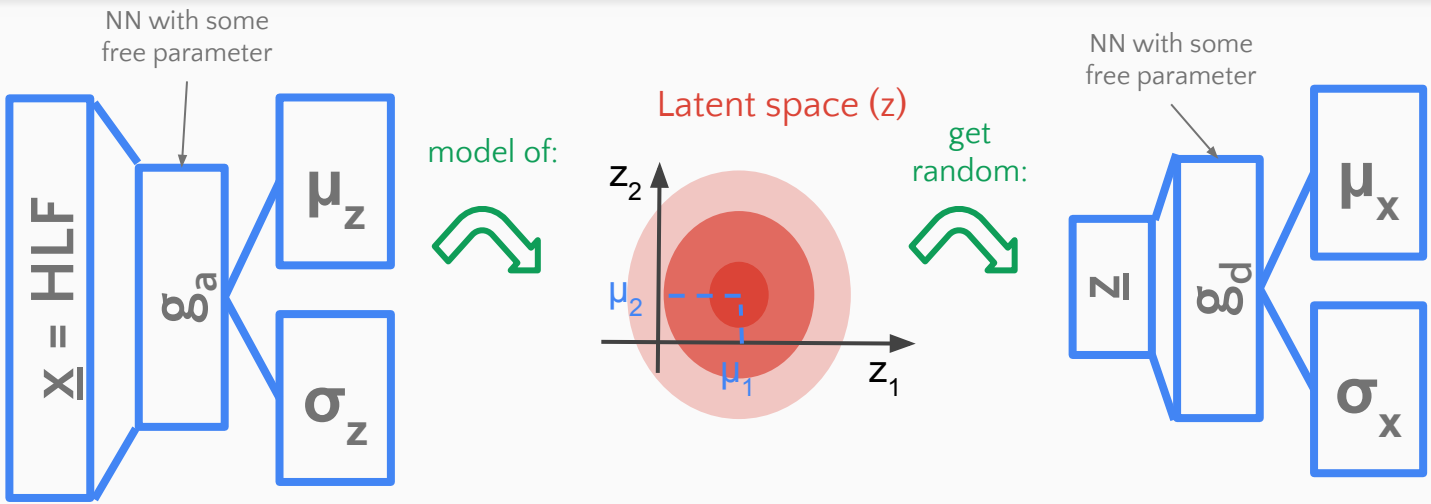
The Variational Auto-Encoder



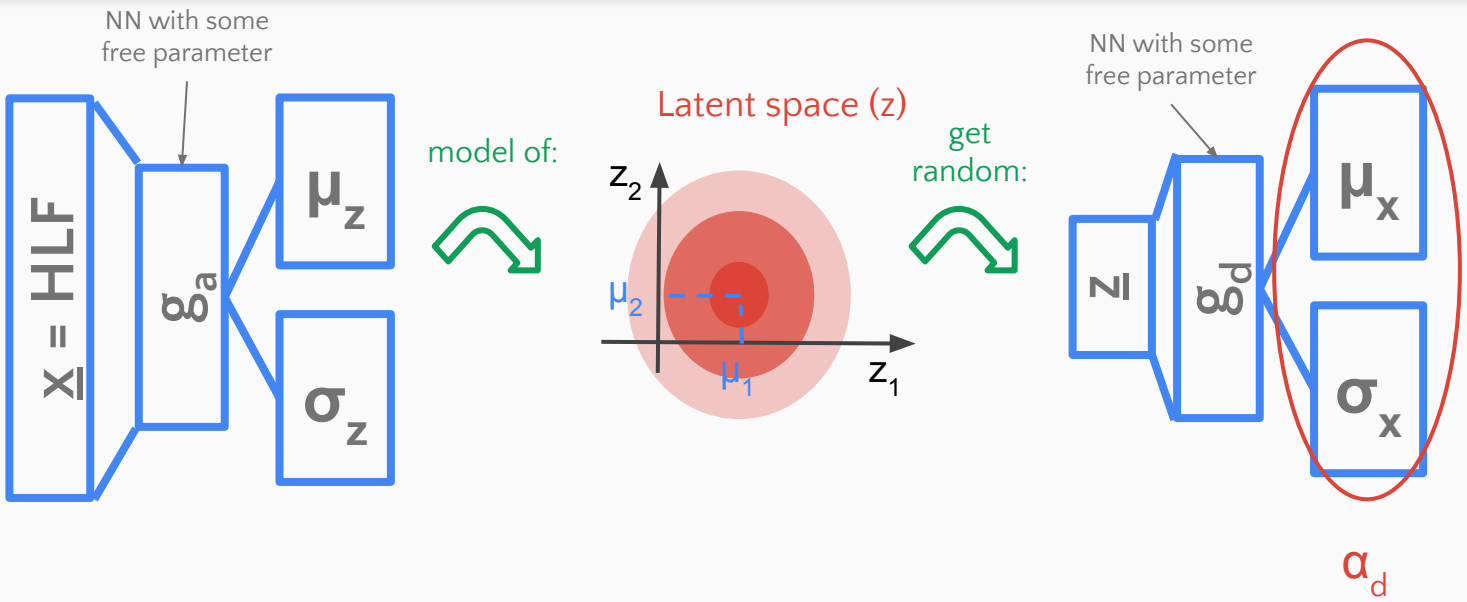
The Variational Auto-Encoder



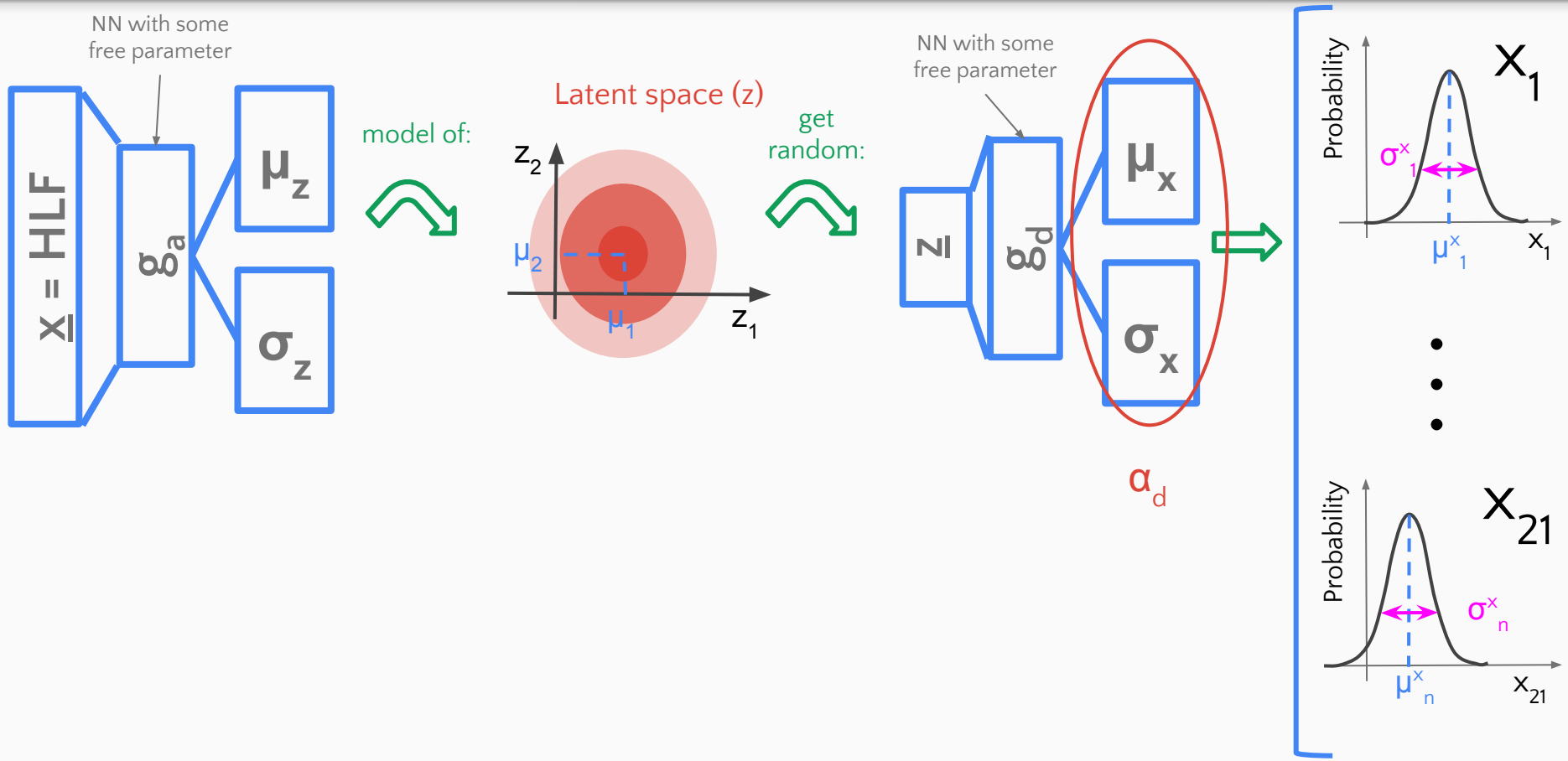
The Variational Auto-Encoder



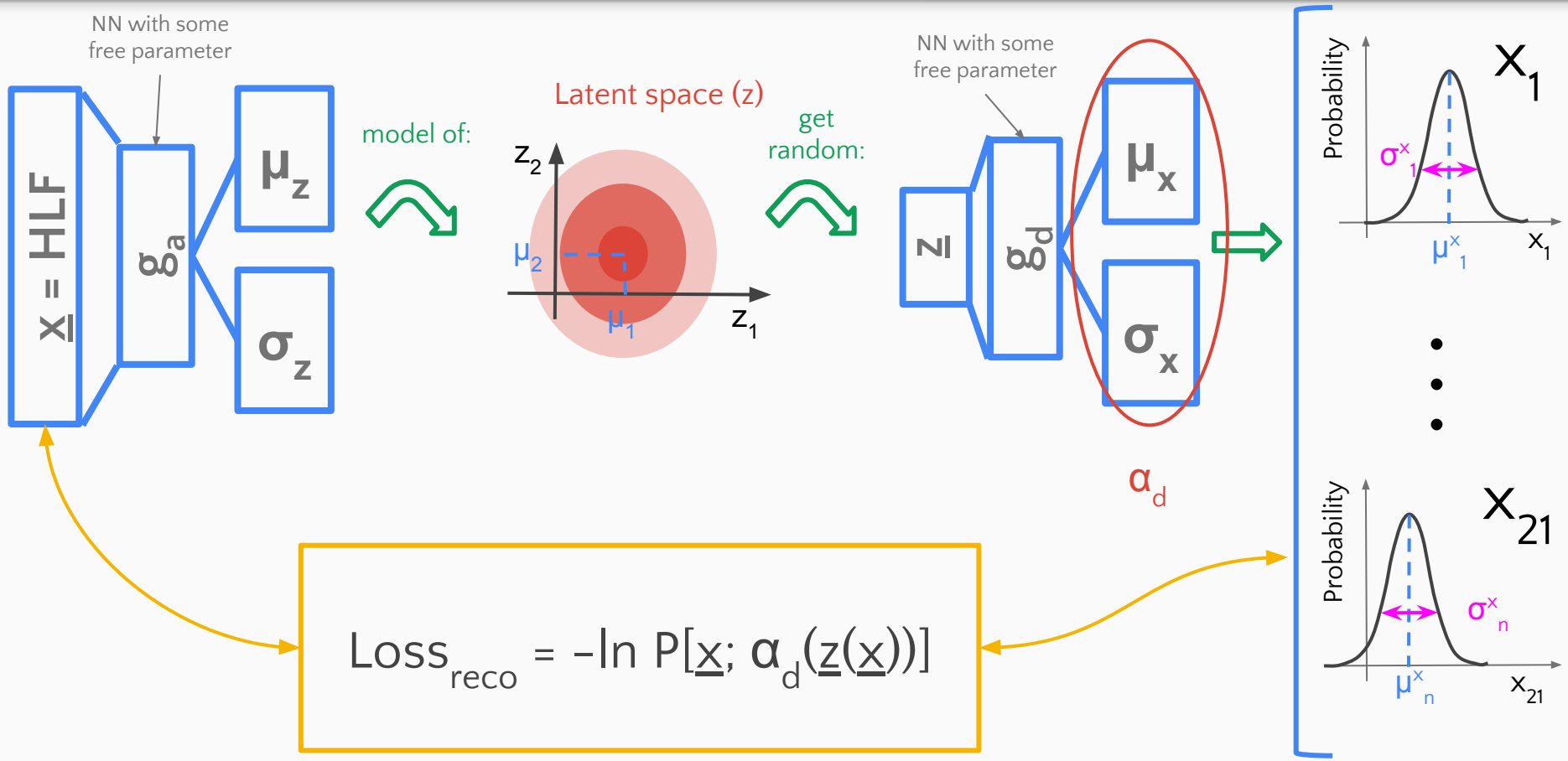
The Variational Auto-Encoder



The Variational Auto-Encoder



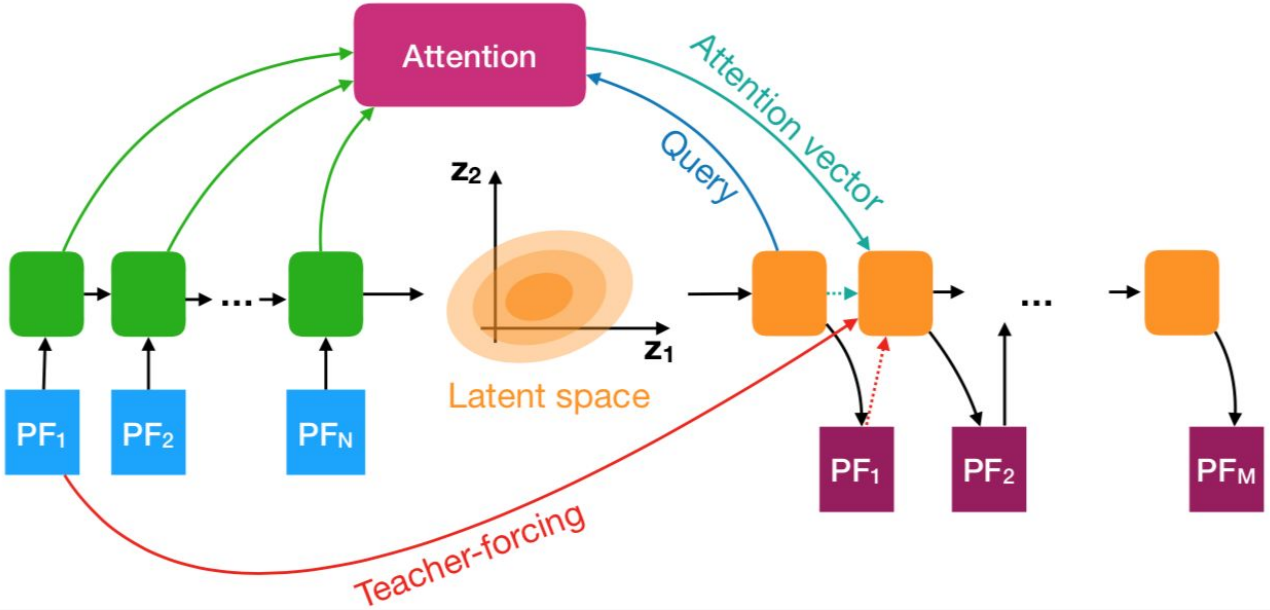
The Variational Auto-Encoder



$$\text{Loss}_{\text{reco}} = -\ln P[\underline{x}; \alpha_d(\underline{z}(\underline{x}))]$$

Attentional Particle-VAE

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



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