Particle-based Fast Simulation of Jets at the LHC with Variational Autoencoders

10th International Conference on New Frontiers in Physics

Mini-Workshop on Machine Learning for Particle Physics

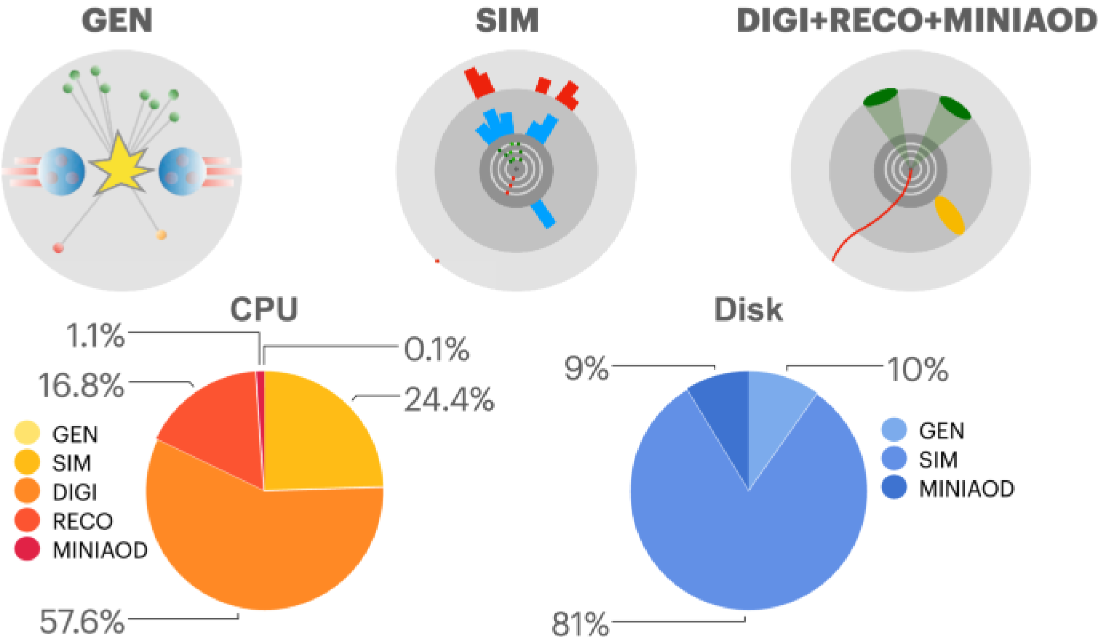
25/08/2021

Mary Touranakou [1, 2], Maurizio Pierini [1], Dimitrios Gunopulos [2], Breno Orzari [3], Thiago Tomei [3], Raghav Kansal [4], Javier Duarte [4], Jean-Roch Vlimant [5]

[1]CERN, [2]NKUA, [3]SPRACE-UNESP, [4]UCSD, [5]Caltech

Deep Learning for Particle Physics

- High Energy Physics (HEP) data analysis heavily relies on the production and the storage of large datasets of simulated events
- HL-LHC upgrade: number of collisions will increase, need for more accurate, synthetic data for analysis

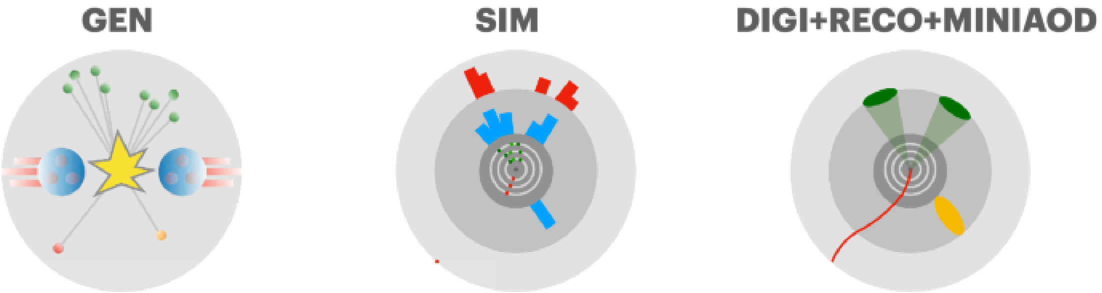


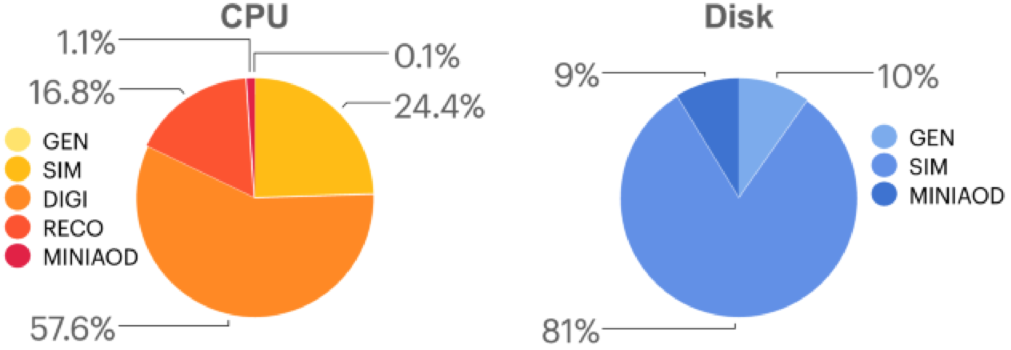
Deep Learning for Particle Physics

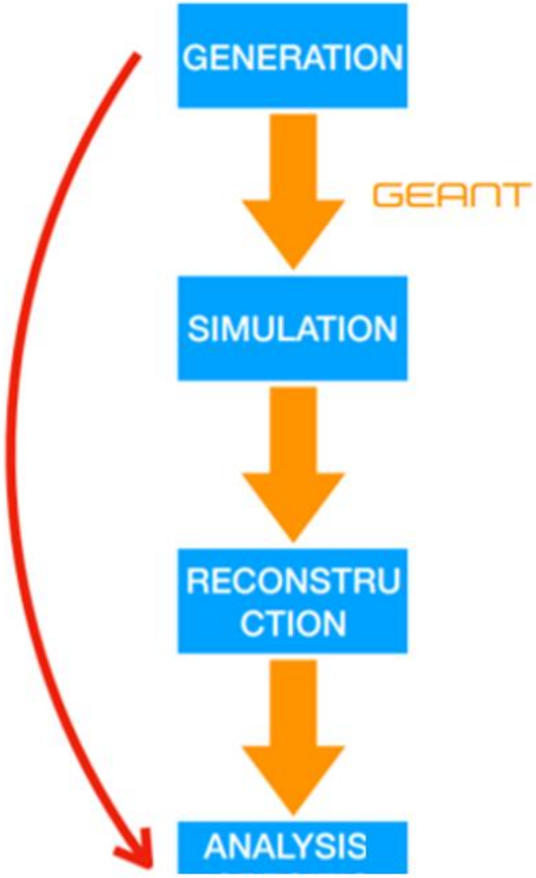
Current HEP workflows are very accurate, but:

- Costly in time & storage resources
- Cannot scale with fixed budget

→Deep neural networks proposed as a tool for speedup

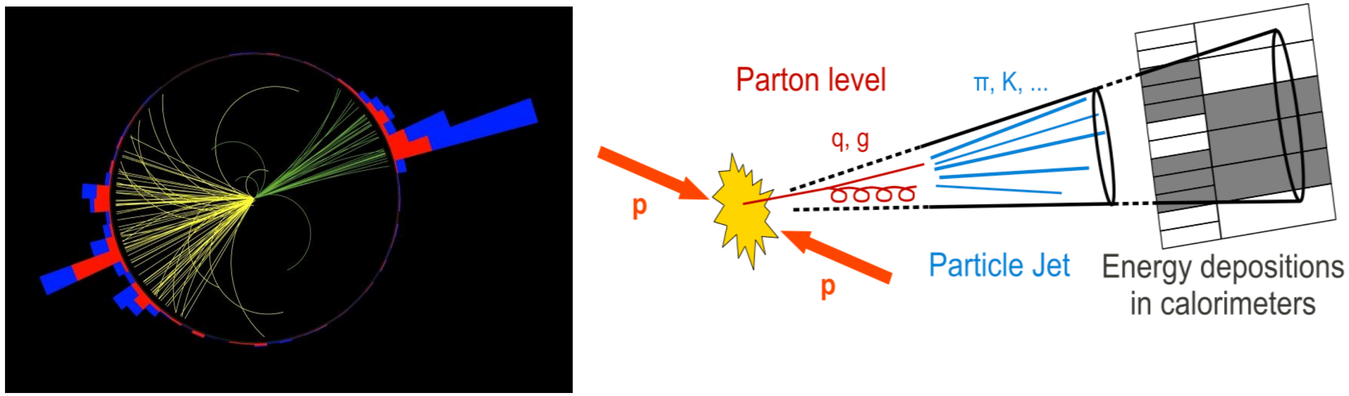






Data Representation of Jets

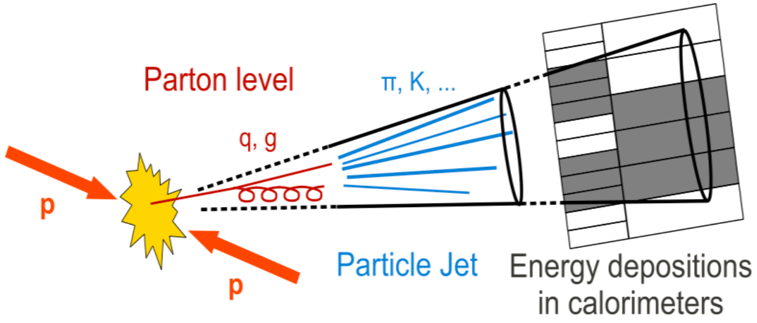
- The representation of data is of crucial importance for machine learning applications. It needs to fit the nature of the data as well as its underlying properties.
- HEP high-level objects like jets are of particular interest: sparse, high granular, irregularly distributed in space data



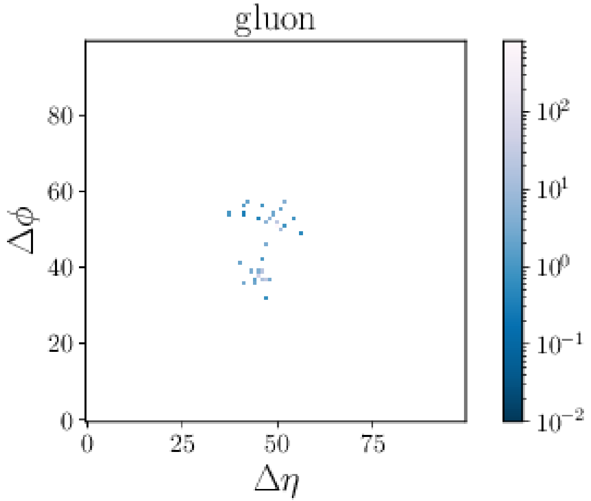
Jets as collimated sprays of particles following a pp-collision

Data Representation of Jets

- Jets can be seen as sparse sets of constituents that are intrinsically unordered. Although, an ordering might be given to the data, it is important to preserve its permutation invariance
- We represent each jet as a list of particles where each particle is characterized by its cartesian momenta p_{x}, p_{y}, p_{z}



Jet at CMS

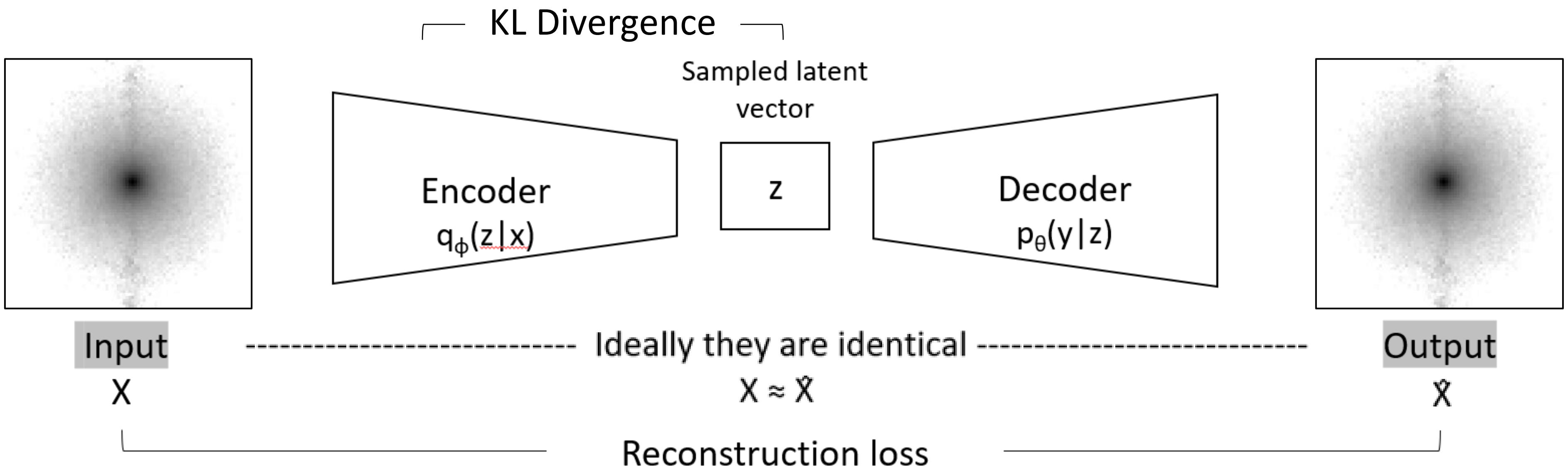


Jet represented as an image

p_{x_1}	p_{x_2}	p_{x_3}	...	p_{x_N}
p_{y_1}	p_{y_2}	p_{y_3}	...	p_{y_N}
p_{z_1}	p_{z_2}	p_{z_3}	...	p_{z_N}

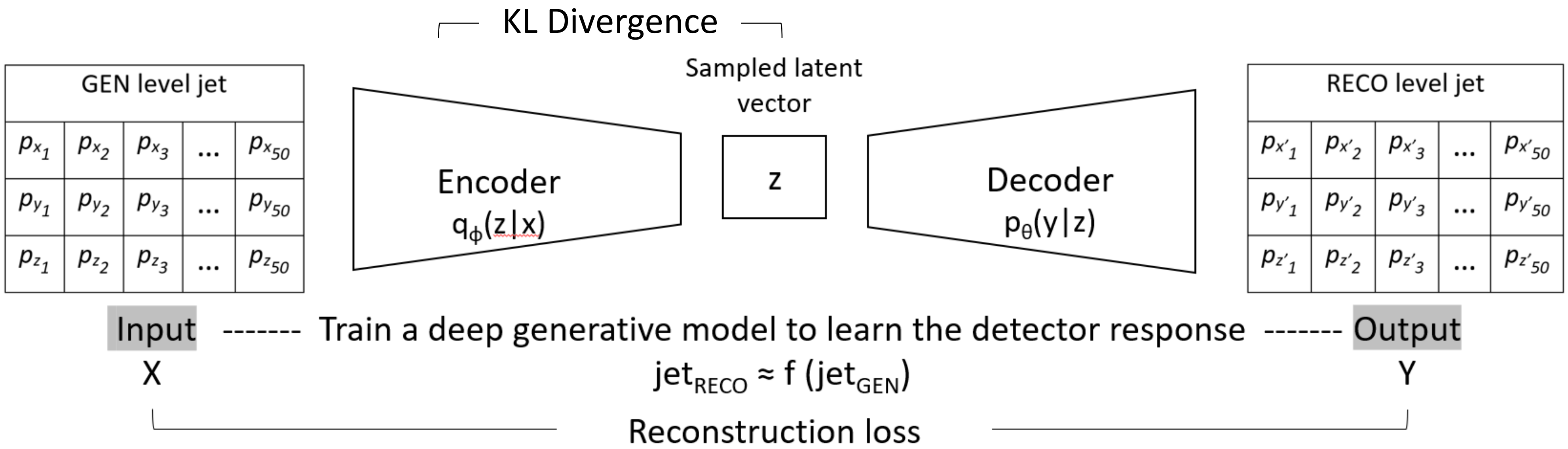
Jet represented as a list of N particles with features p_{x}, p_{y}, p_{z}

Variational Autoencoders



- Generative models capable of generating new samples by modeling the underlying distribution of the original data and its most important attributes.
- Variational Autoencoder (VAE) traditionally designed to reconstruct its input through a compressed lower representation of it.

Particle-based Fast Simulation of Jets with VAEs



KL Divergence

GEN level jet				
p_{x_1}	p_{x_2}	p_{x_3}	...	$p_{x_{50}}$
p_{y_1}	p_{y_2}	p_{y_3}	...	$p_{y_{50}}$
p_{z_1}	p_{z_2}	p_{z_3}	...	$p_{z_{50}}$

Encoder $q_\phi(z|x)$

Sampled latent vector

z

Decoder $p_\theta(y|z)$

RECO level jet				
$p_{x'_1}$	$p_{x'_2}$	$p_{x'_3}$	...	$p_{x'_{50}}$
$p_{y'_1}$	$p_{y'_2}$	$p_{y'_3}$	...	$p_{y'_{50}}$
$p_{z'_1}$	$p_{z'_2}$	$p_{z'_3}$	...	$p_{z'_{50}}$

Input X ------- Train a deep generative model to learn the detector response ------- Output Y

$jet_{RECO} \approx f\ (jet_{GEN})$

Reconstruction loss

- Proposed idea: VAE to reconstruct a detector-level view (RECO) of jet constituents starting from a generator-level (GEN) view of them through modeling in an unsupervised manner the noise function of the detector effects.
- Customization of a permutation-invariant reconstruction loss to increase accuracy and impose physics constraints to the model.

Loss function

Total loss defined as

$$L_{VAE} = L_r + \beta D_{KL}$$

with β = 0.5. The reconstruction loss is defined as a combination of two terms

$$L_r = L_{CD} + L_J$$

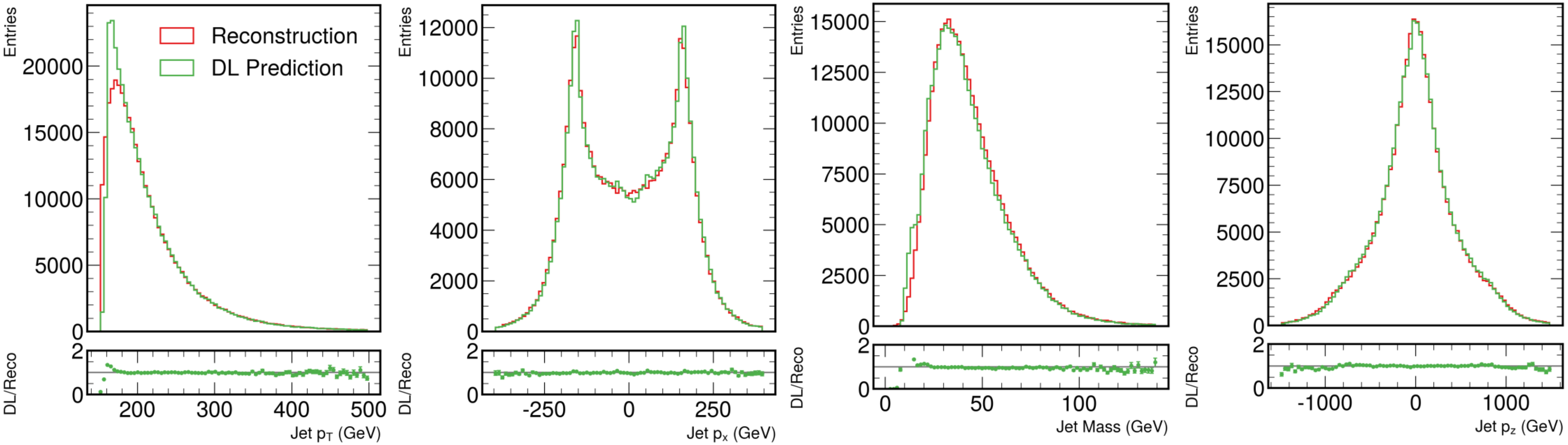
where L_{CD} is the permutation-invariant Chamfer distance of the particles' features representing the input and output jets as sets of particles

$$L_{CD} = \sum_{i \in X} \min_{j \in Y} (p_i - p_j)^2 + \sum_{j \in Y} \min_{i \in X} (p_i - p_j)^2$$

and L_J mean squared error terms on the jet transverse momentum and jet mass that impose physics constraints to the model

$$L_J = MSE(p_T^{jet}, \hat{p_T}^{jet}) + MSE(m^{jet}, \hat{m}^{jet})$$

Results



- Evaluating learning capability of the model through jet features' distributions matching based on Earth mover's distance (EMD) between the output (DL Prediction) and the target (Reconstruction). Most jet features matched with fair fidelity, room for further improvement.
- Early stopping employed as a learning regularization mechanism based on EMD

Conclusion

- Presented the use case of employing deep generative models for particle-based fast simulation of jets at the LHC. A Variational Autoencoder with a permutation-invariant reconstruction loss was considered and jets was represented as sets of particles.
- Results pointed to a promising direction; the potential of such models to bypass the detector simulation, and the event reconstruction steps of a traditional event processing, potentially speeding up significantly the events generation workflow.

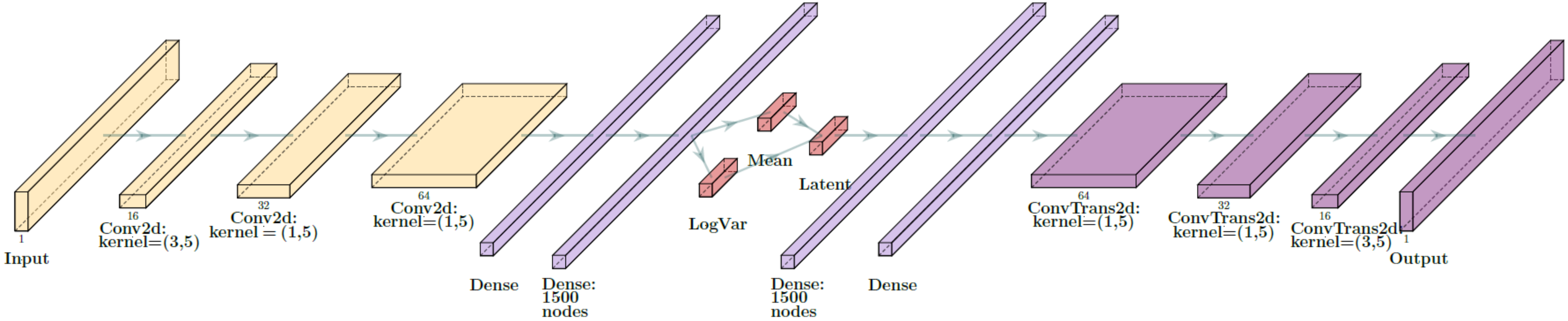
Next steps

- Leveraging the capabilities of VAEs with graph neural networks
- Working with varied length inputs that would better fit the nature of the data

Thank you for listening

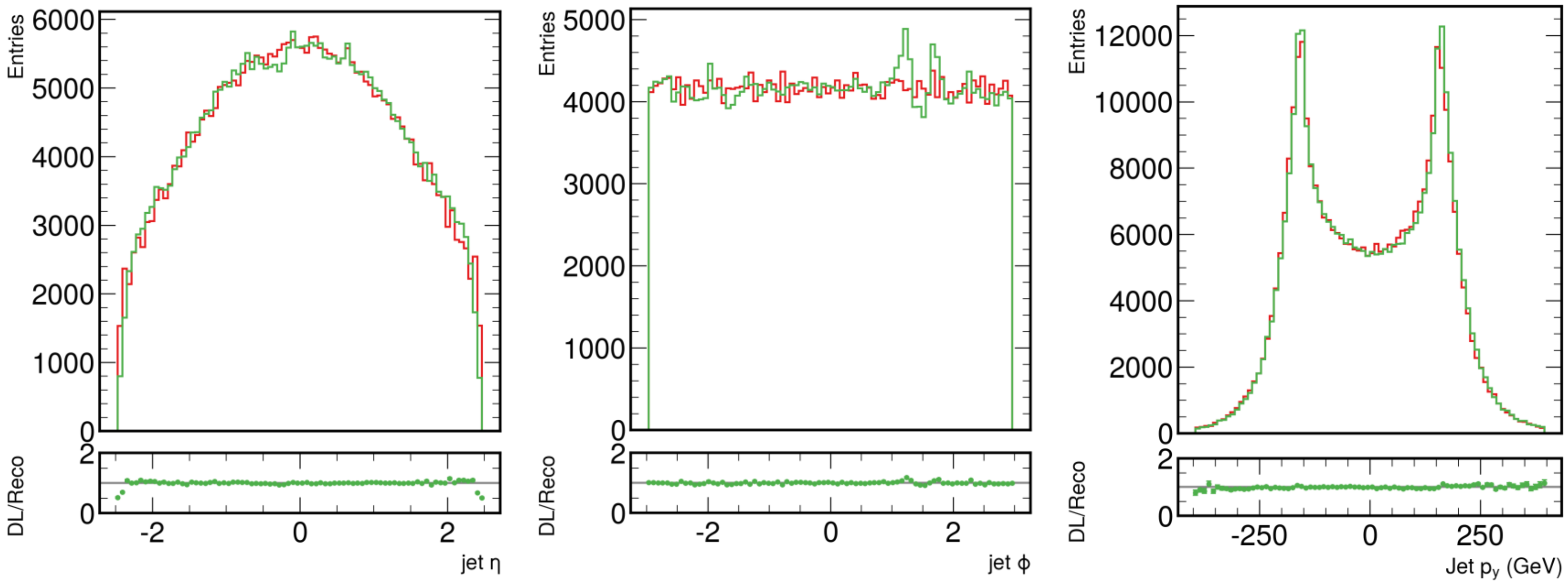
Backup

VAE architecture



- ReLU is being used as the activation function on all layers except for the last layer where linear activation is used.
- Adam optimization with a learning rate = 0.0001, latent space dimension = 20, and loss function tuned with β = 0.5, coeff$_{\text{particles}}$ = 0.015, coeff$_{\text{mass}}$ = 1.0, coeff$_{p_T}$ = 0.1.

More results


Entries
DL/Reco
jet η
Entries
DL/Reco
jet ϕ
Entries
DL/Reco
Jet p_y (GeV)