Particle-based Fast Jet Simulation with Variational Autoencoders

ML4Jets

07/07/21

Mary Touranakou, Maurizio Pierini, Dimitrios Gunopulos, Breno Orzari, Thiago Tomei, Raghav Kansal, Javier Duarte, Jean-Roch Vlimant

DEPARTMENT OF INFORMATICS & TELECOMMUNICATIONS

HELLENIC REPUBLIC
National and Kapodistrian
University of Athens
EST. 1837

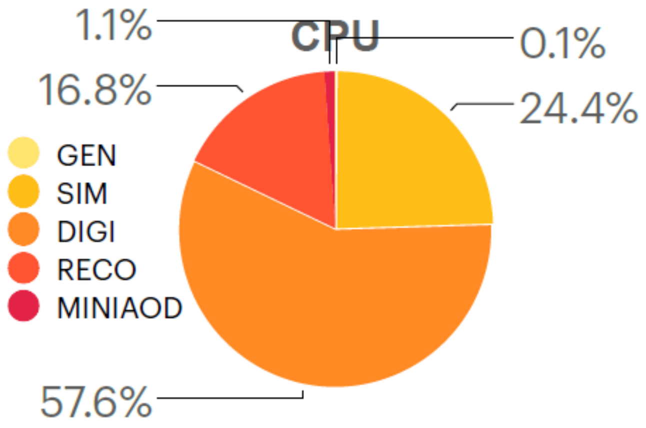
mPP

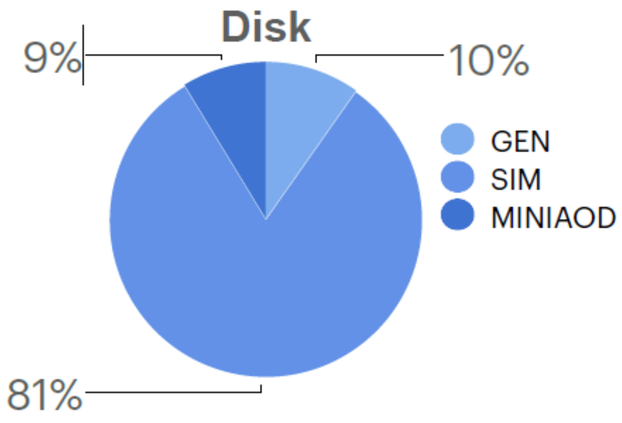
European Research Council
Established by the European Commission

CERN

Motivation

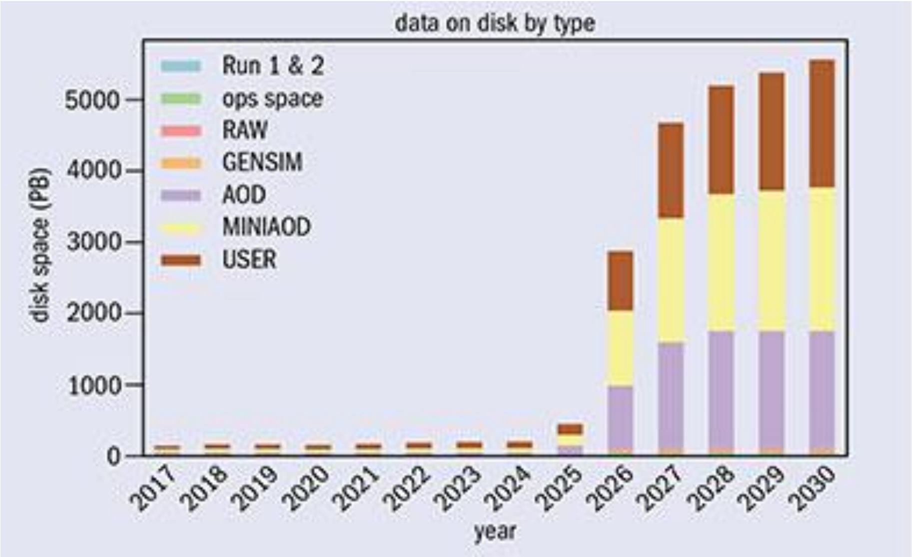
- HEP data analysis heavily relies on the production and the storage of large datasets of simulated events
- Current HEP solutions are very accurate, but also very slow
  - ➢ Costly in time resources
  - ➢ Costly in storage (as data needs to be saved till researchers need it)
  - ➢ Cannot be used as on-demand services

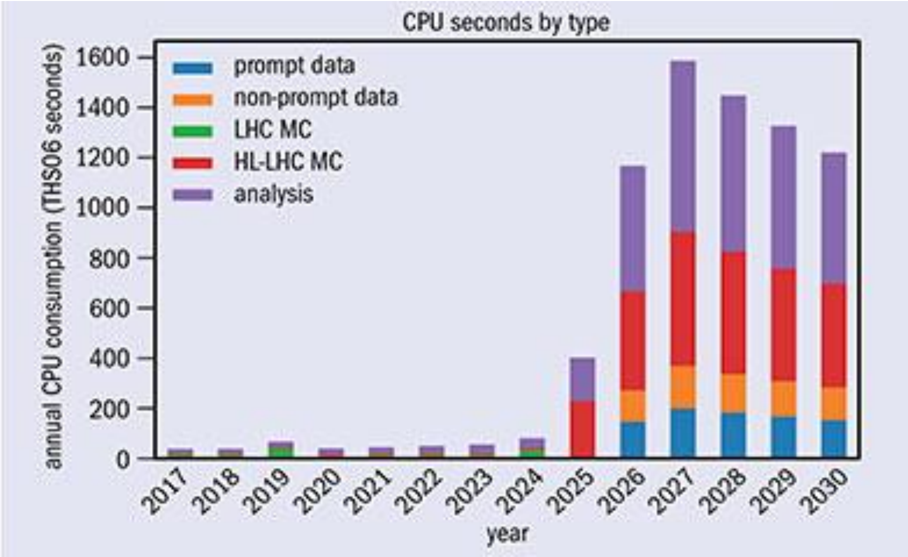




Motivation

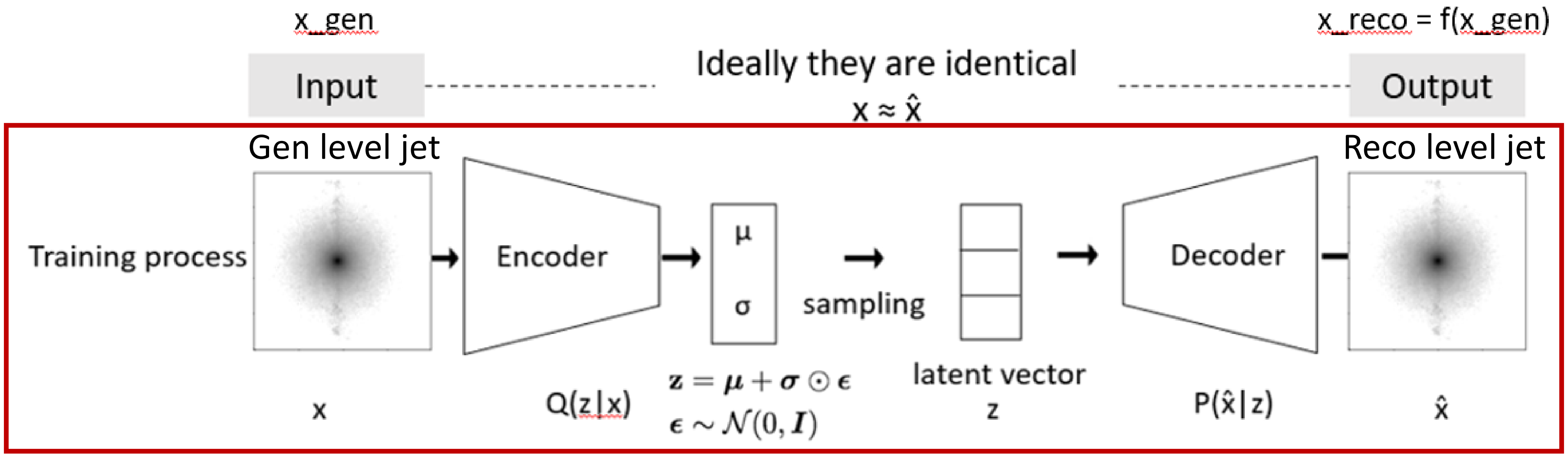
- Foreseen upgrade for the High-Luminosity phase of the LHC will increase the number of collisions and the need of more accurate, synthetic data for analysis
- Current solutions do not scale with fixed budget → Deep Generative models proposed as a tool for speeding up HEP workflows





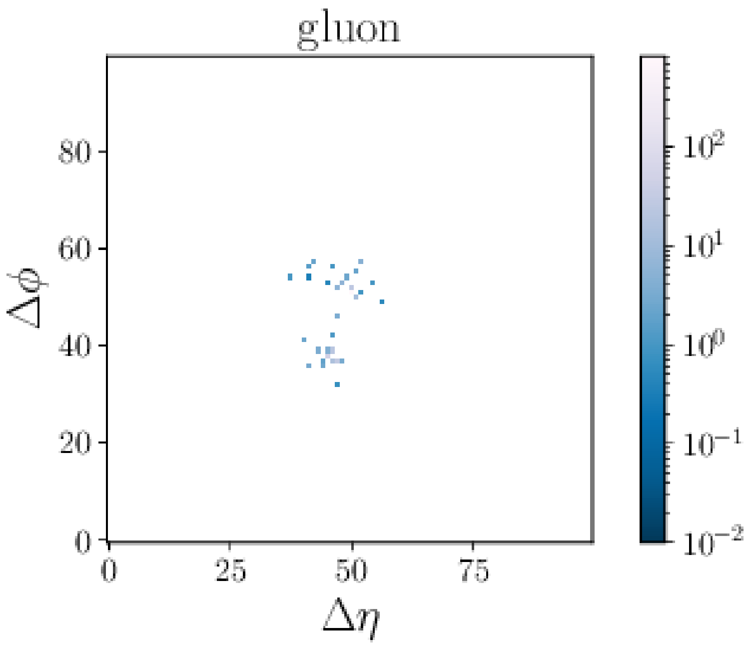
Particle-based Fast Jet Simulation with VAEs

Starting from a simulation of the jet before detector effects, we train a VAE to return the corresponding list of constituents after detection.



Run end-to-end VAE (from GEN to RECO) to learn a parametric description of the detector response

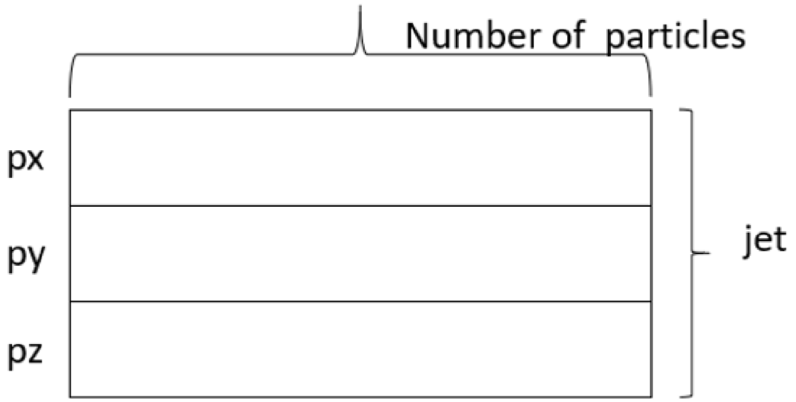
Data Representation of Jets



Graphical representation of a gluon jet (arXiv:1908.05318)

- Jets can be characterized as sparse sets of particles that are intrinsically unordered
- Jets can be represented as a list of constituents characterized by their momenta:
  - in cartesian coordinates p_x, p_y, p_z
  - in hadronic coordinates p_T, η, φ
- Although, sometimes, an ordering might be given to the data (e.g. ordering particles by decreasing p_T), it is also important to preserve the permutation invariance in it (depending on application-specific requirements)

Data Representation of Jets



Structural representation of a jet

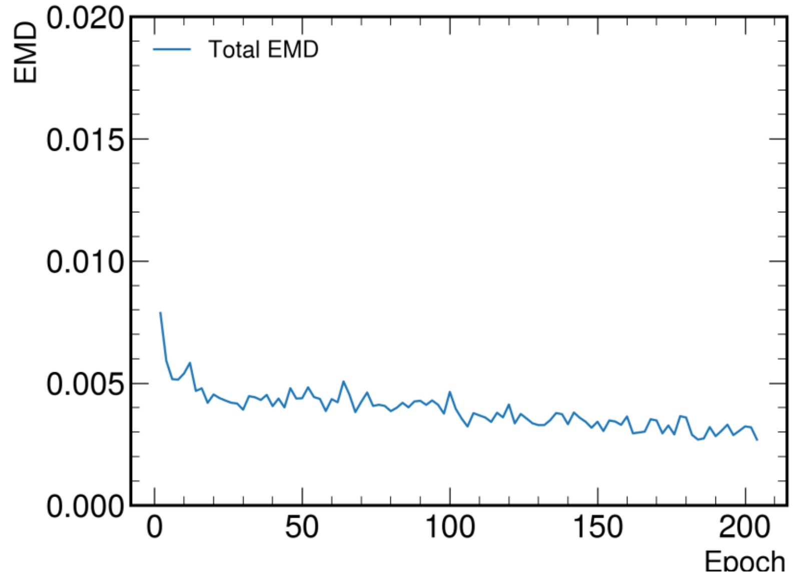
- We represent jets as a list of N constituents, characterized by their momenta p_x, p_y, p_z. If less than N particles are present in the jet, zero-padding is applied.
- We apply feature-dependent standardization such that each feature (p_x, p_y, p_z) has zero mean and unit variance.
- Our dataset consists of W boson jets with up to 50 particles (originally 100 particles with a cut on the first 50 particles).

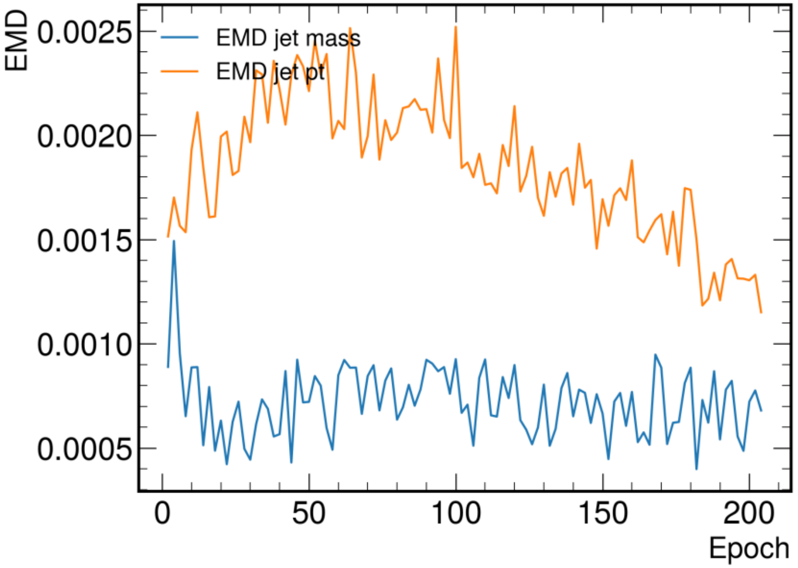
Reconstruction Loss Function

- Use of a permutation-invariant, pairwise Nearest Neighbour Distance (NND) known as the Chamfer distance (arXiv:1906.02795) for the reconstruction loss
- Customization of the reconstruction loss by adding two extra terms, the jet mass and the jet p_T to enforce the model to learn jet kinematics
- The jet mass and the jet p_T (target and reconstructed) are computed from the sum of the momenta of the particles on the jet

$$L_{reco} = \sum_i min[d_{eucl}(X_i, \hat{X})]^2 + \sum_i min[d_{eucl}(X, \hat{X}_i)]^2$$
$$+ \sum_j [d_{eucl}(p_T^{jet}, \hat{p}_T^{jet})]^2 + \sum_j [d_{eucl}(m^{jet}, \hat{m}^{jet})]^2$$

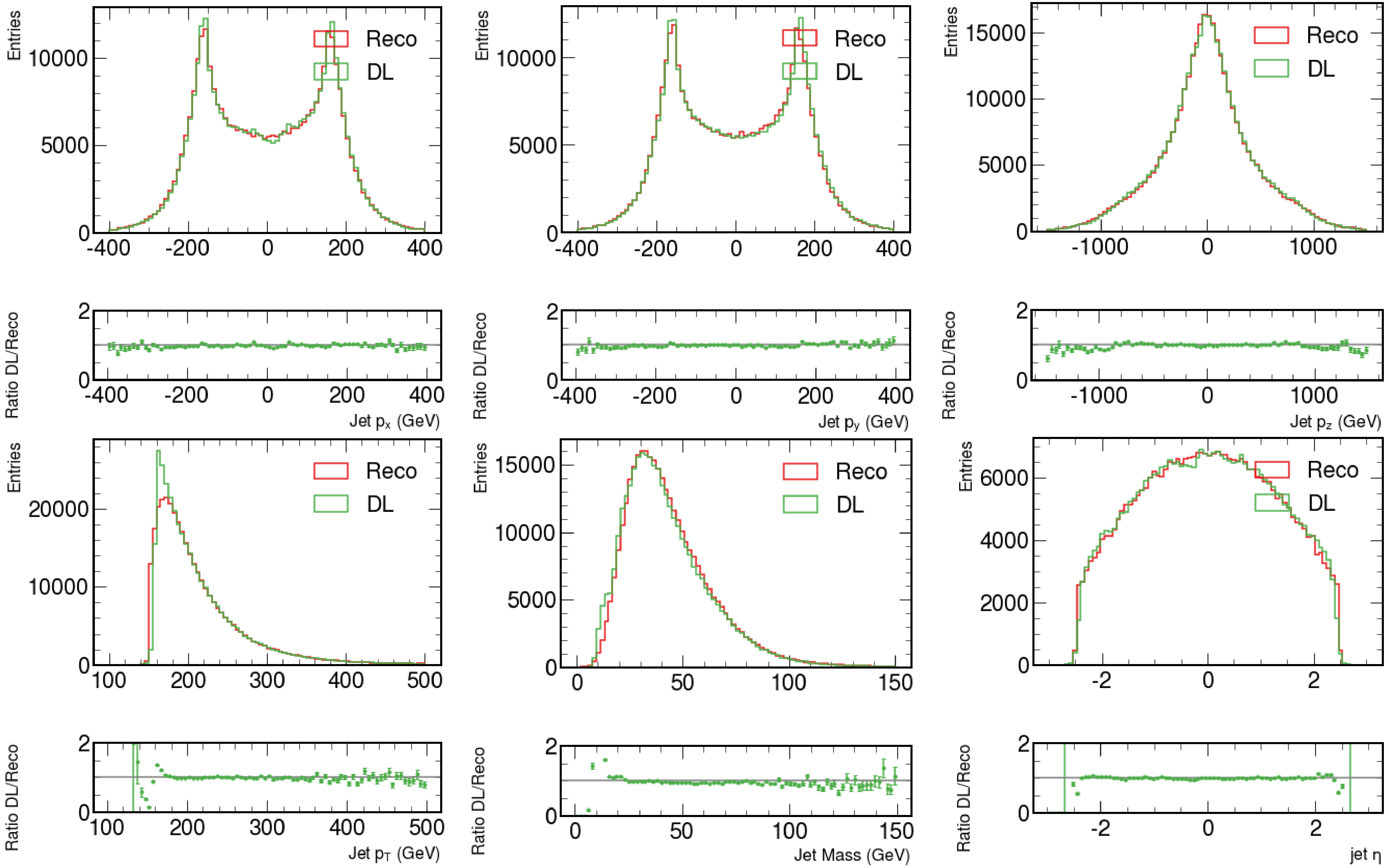
Model Detection & Evaluation





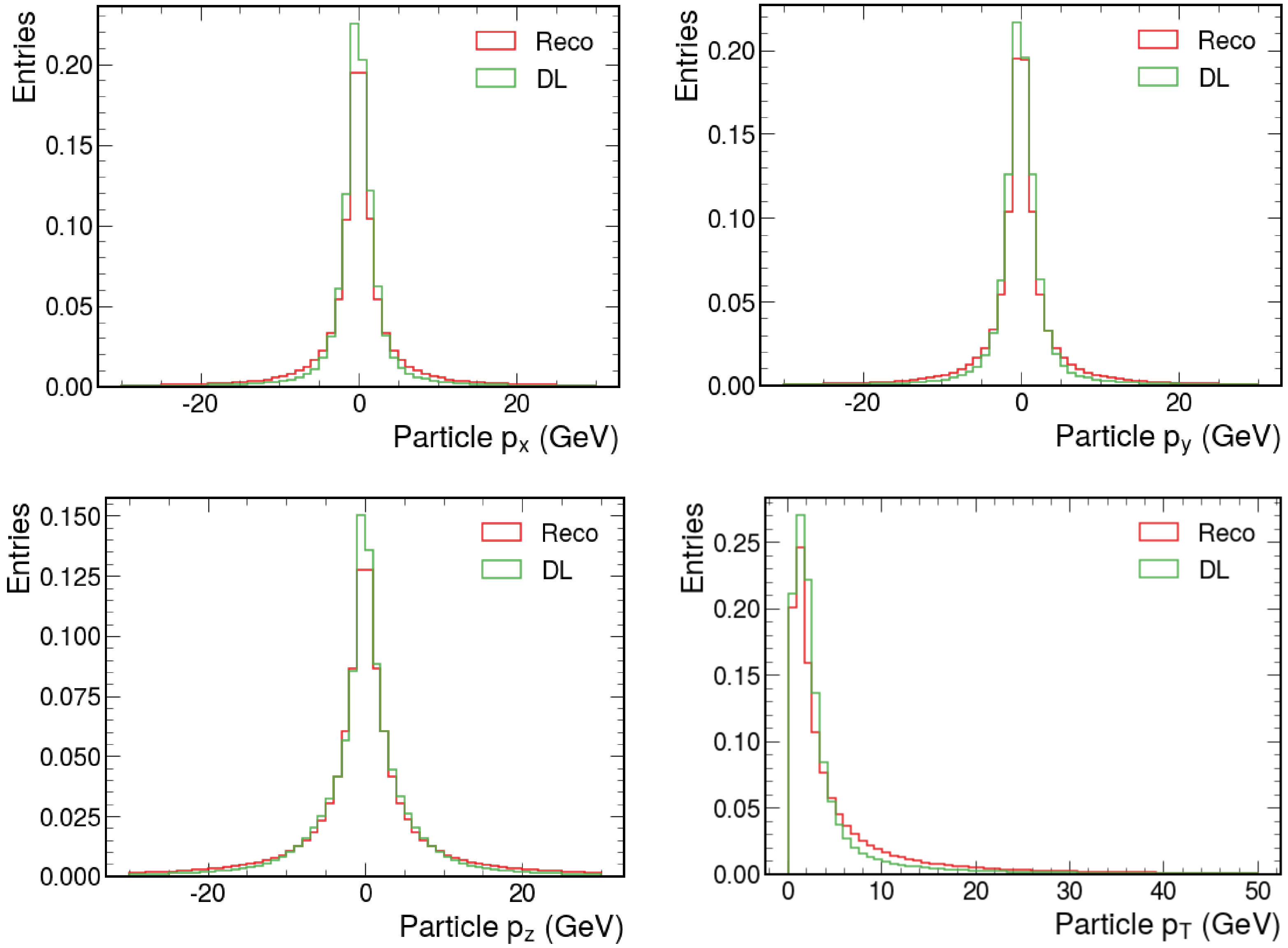
- Earth mover's distance (EMD) between the output (DL) and the target (Reco) jet features histograms used as a quantitative metric to select the best model for analysis
- Define as best the model that satisfies joint conditions for EMD metrics:
  - minimizes EMD sum of all jet features
  - keeps the EMD sum of jet mass and jet p_T below an empirically-set threshold
- EMD as a figure of merit to find a compromise between the multi-objective reco loss

Results: Simulation of Reco level Jets (jet level features)

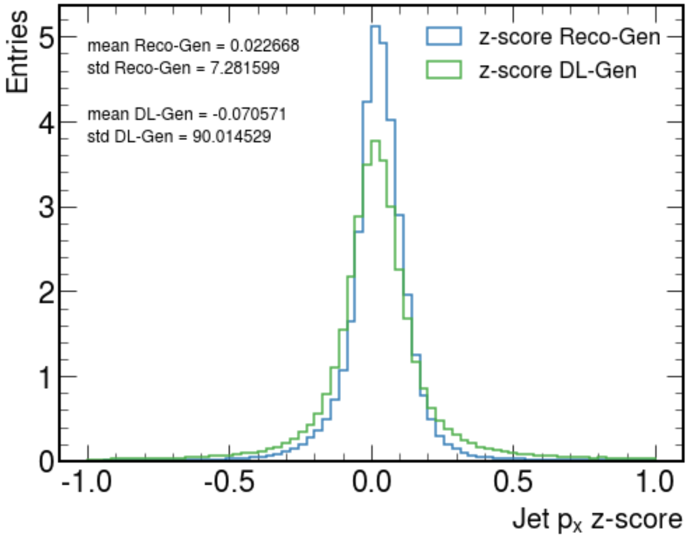


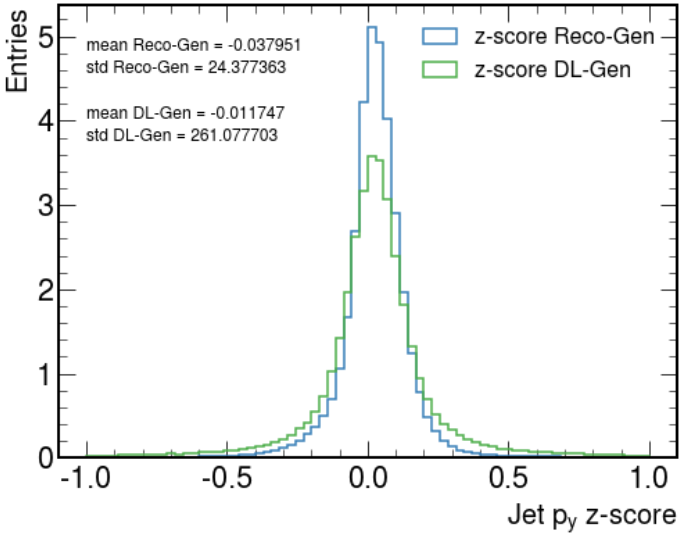
Jet features are matched with high fidelity, except for the jet pT that could be further improved.

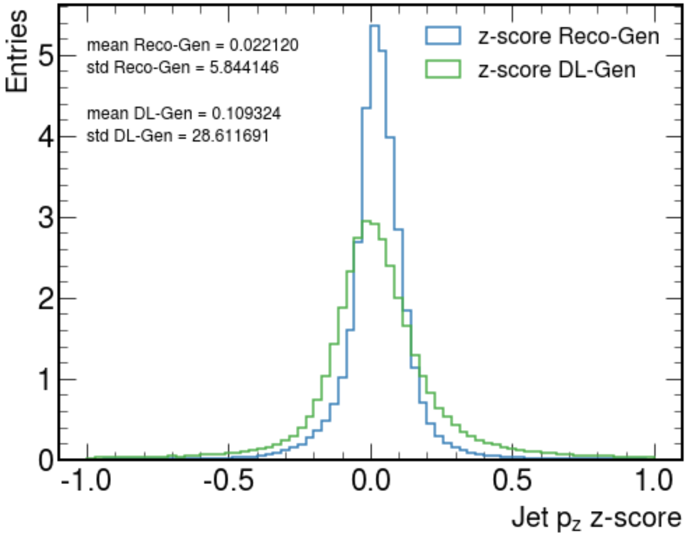
Results: Simulation of Reco level Jets (particle level features)



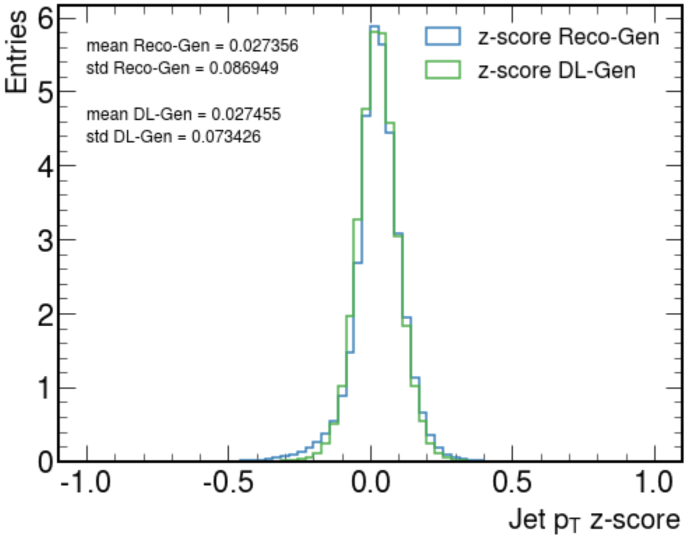
Particle features are matched with lower accuracy, but most of the not well-matched particles are low-energy objects carrying little information of the jet global features → Acceptance if jet features are OK.

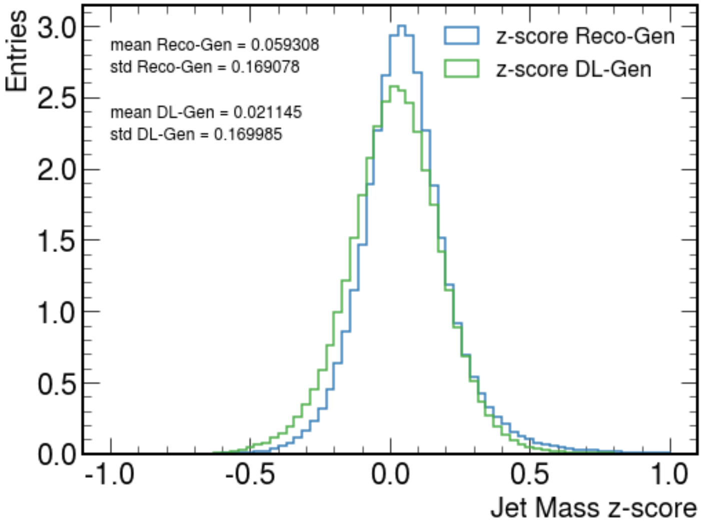






z-score: the difference between the target/output and the input, divided by the input.





For our domain-specific application, we focus on modeling the jet features with high fidelity (within 10% of accepted uncertainty) and some fair fidelity on a jet-per-jet basis. → Most jet features are modeled within the desired uncertainty keeping some degree of connection between the output and the target.

Work Status & Challenges

- CNN VAE
  - ➢Working baseline model with a permutation-invariant loss
  - ➢Main challenge: fixed-length tensors and zero-padding
  - ➢Possible solutions: learnable masking feature as described in the Deep Set Prediction Networks paper to deal with zero-padding, or Graph NNs.
- Graph VAE
  - ➢Moving into a graph implementation of the VAE
  - ➢Permutation-invariant architecture + permutation-invariant loss → Replace CNN VAE with a Graph VAE + Chamfer loss
  - ➢ Varied length tensors instead of fixed length → Avoid zero-padding

Summary

- We presented a Convolutional Variational Autoencoder with a permutation-invariant reconstruction loss function for particle-based fast jet simulation.
- We discussed the customization of the loss to improve accuracy and the use of EMD for best model detection and analysis.
- The CNN architecture and the Chamfer loss seem to serve as a working baseline for the problem in study though possibly limited.
- We are looking into moving into a permutation-invariant architecture to improve the current baseline.

Backup

VAE Loss Function

$$L^{VAE} = L_{reco} + \beta D_{KL}$$

where

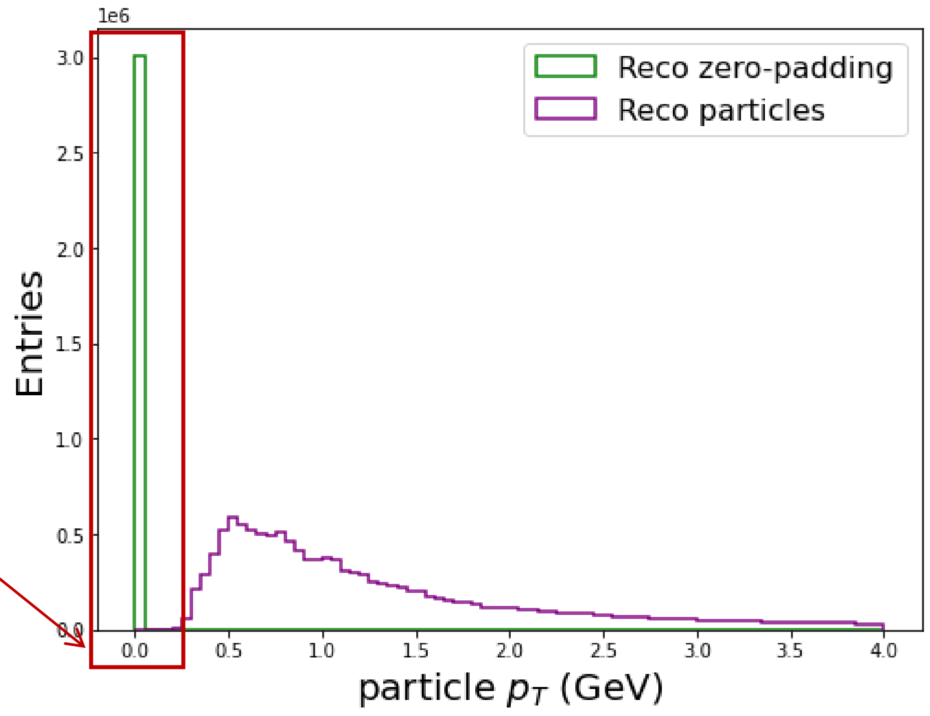
$$D_{KL}(\underbrace{q_\phi(z|x)}_{\sim N(\mu,\, \sigma)} \,||\, \underbrace{p_\theta(z)}_{\sim N(0,\, 1)})$$

The loss function of a VAE consists of two terms:

1. The reconstruction loss (e.g. traditionally a generic loss function such as the MSE) – the error between the original samples (inputs/targets) and the produced outputs (reconstructed inputs).
2. The Kullback-Leibler (KL) divergence used as a loss function between the encoder's distribution $q_\phi(z|x)$ and the $p_\theta(z)$ that optimizes the probability distribution parameters (μ and σ) to closely resemble those of the target distribution.

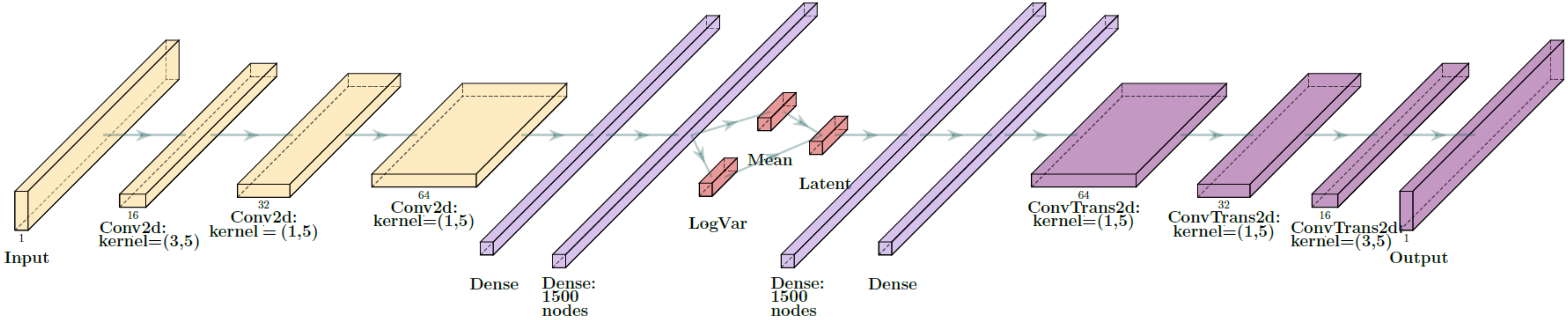
CNNs & Permutation Invariance

- <u>Order invariance</u>
  - ➢ Our workflow should not depend on any ordering principle (even if an arbitrary ordering is imposed to the input data) → Chamfer loss seems like a natural choice for the VAE due to its permutation invariance.
- <u>Zero-padding</u>
  - ➢ CNNs work with fixed length tensors which introduced zero-padding*. → Complication: trying to learn these zeros, the network finds a compromise between the two populations.
  - ➢ Possible Solutions: binary flag to differentiate the two populations, or alternative architecture to encode varied-length data.



*Two populations of particles that are isolated.

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 $\beta = 0.5$, $coeff_{particles} = 0.015$, $coeff_{mass} = 1.0$, $coeff_{pT} = 0.1$.
- VAE model trained in Pytorch with Early Stopping based on min EMD sum between the output (DL) and the target (Reco) jet features histograms with a patience of 50 epochs.