University of CINCINNATI
ARTS AND SCIENCES

Modeling Hadronization using Machine Learning

Phenomenology 2022 Symposium

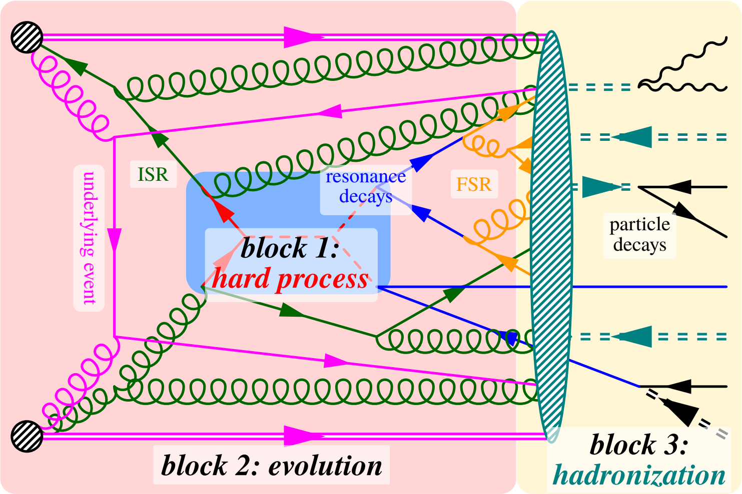
Ahmed Youssef

Based on 2203.04983 with P. Ilten, T. Menzo and J. Zupan

May 9, 2022

University of Cincinnati Department of Physics

Motivation



ISR
underlying event
resonance decays
FSR
particle decays
block 1: hard process
block 2: evolution
block 3: hadronization

Motivation

- The hard process and the parton shower are perturbative in their nature

  $\Rightarrow$ Theoretical under control

- Hadronization is inherently non-perturbative

  $\Rightarrow$ Forced to use phenomenological models

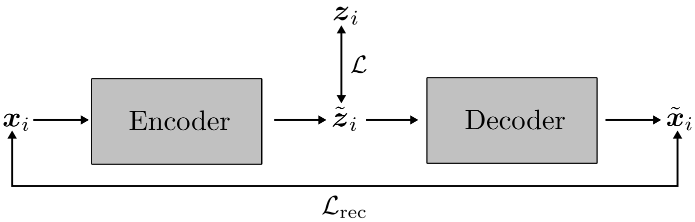
- First step: Create a Machine Learning (ML) Architecture that is able to reproduce the simplified Lund String Model

- Goal: **Train on real experimental data** and replace the Hadronization model in PYTHIA

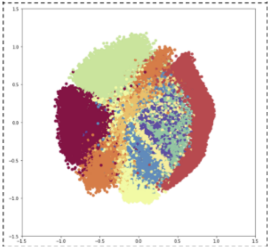
MLHAD

VAE

- VAE is a commonly used generative model:
  - $\longrightarrow$ Not flexible with the latent representation
  - $\longrightarrow$ kl-divergence limits latent distribution to a simple analytical form (e.g. Gaussian)

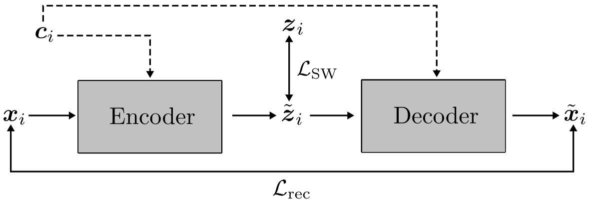


(a) Vanilla VAE
(arXiv: 1312.6114)

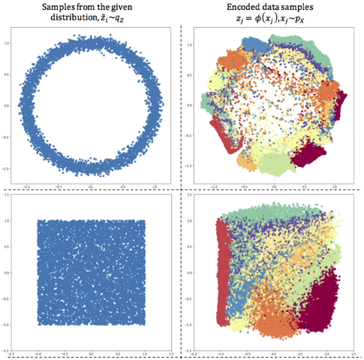


(b) VAE latent space

cSWAE



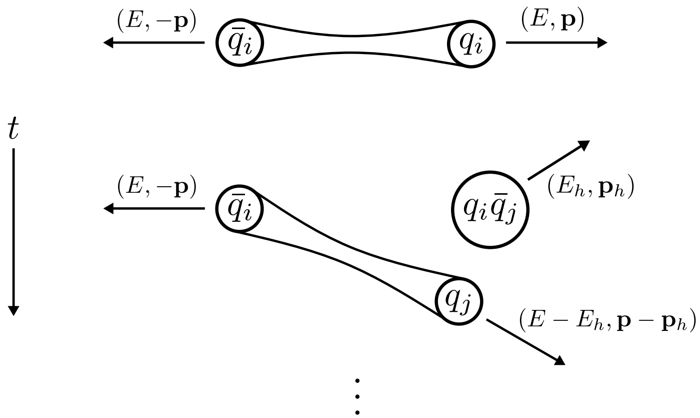
(a) cSWAE architecture



(b) SWAE latent space
(arXiv: 1804.01947)

Total loss: $\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{SWD}$

Lund-String model



Lund-String model

- Selection of the flavor and hadron kinematics are independent until the final stages
- Final stages of hadronization: String energy is close to the nonperturbative scale $\longrightarrow$ the flavor and kinematic selection become interwined

  $\Longrightarrow$ CM string energy cutoff $E_{\text{cut}} = 5$ GeV
- Coordinate system is oriented such the z-axis and the initial string direction are aligned

$$p_x = p_T \cos\varphi \qquad p_y = p_T \sin\varphi \qquad p_T = \sqrt{p_x^2 + p_y^2}$$

- String breaking and hadron emission are axial symmetric in PYTHIA

  $\Longrightarrow$ Reduces to a 2 variable problem: p_z and p_T
- Input data $\mathbf{x_i} \in \left[p_z^{(i)}, p_T^{(i)}\right]$; conditioned on the initial string energy E_i
- Limit the training on light quark flavors and only pions as final state hadrons

MLHAD overview

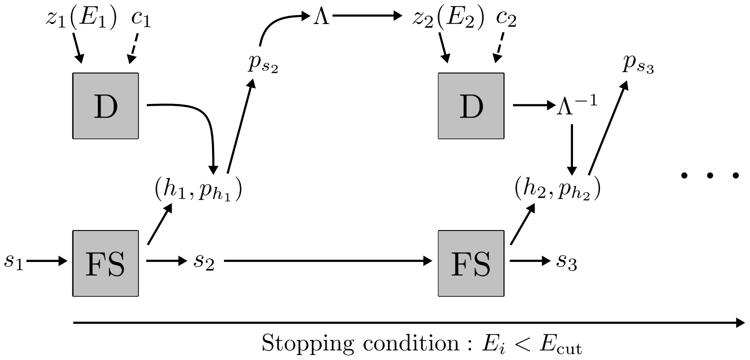


Illustration of the conditional generation

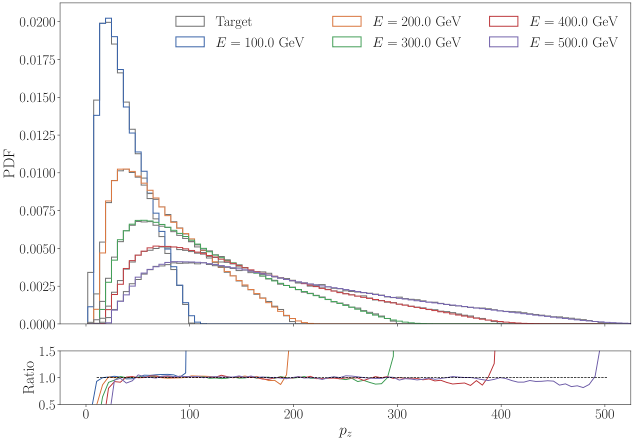
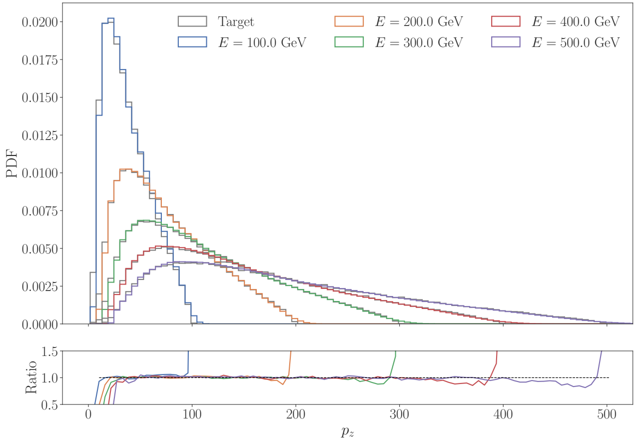
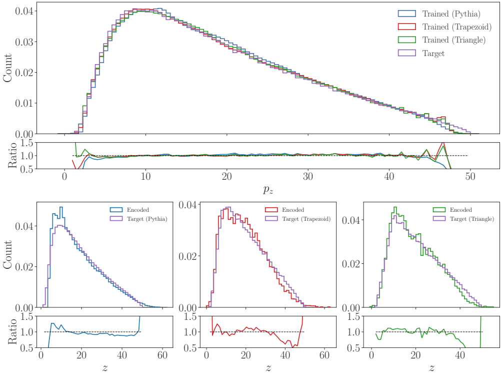


Illustration of the conditional generation

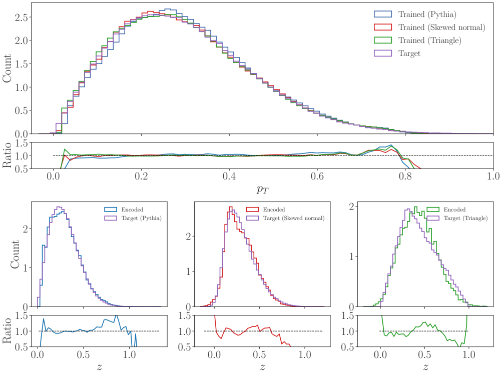


$\Longrightarrow$ **Fixed initial string energy:** $E_i = 50$ GeV

Results

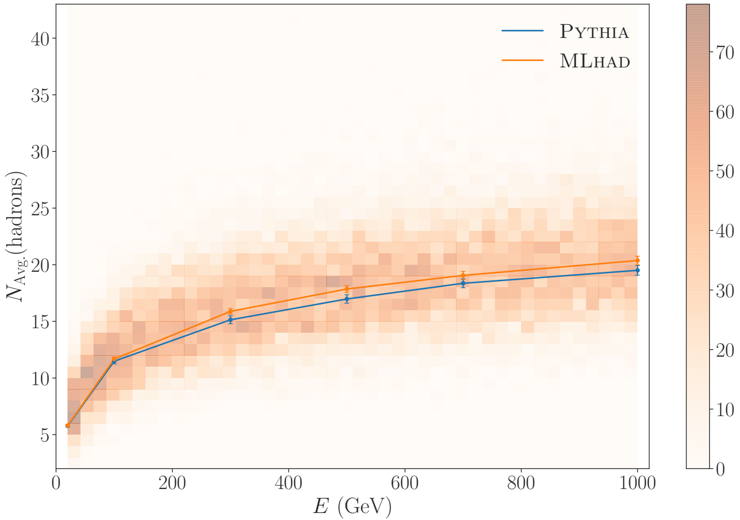


(a) MLHAD generated p_z and latent distribution



(b) MLHAD generated p_T and latent distribution

Results



Pythia
MLhad
$N_{\text{Avg.}}$ (hadrons)
E (GeV)
0
200
400
600
800
1000
5
10
15
20
25
30
35
40
0
10
20
30
40
50
60
70

Conclusion and Outlook

- MLHAD based on the cSWAE architecture was succesfully trained on a simplified PYTHIA Hadronization model
  - $\Rightarrow$ limited to light quark flavor endings of the string and only pions as final state hadrons
  - $\Rightarrow$ MLHAD is extendable to handle all possible string flavors and kinematics
- Flexibility in the choice of the latent space distribution (does not need to have an analytical form)
- Public code available: `https://gitlab.com/uchep/mlhad`

Work in progress [1]

- Performing training on physically accessible observables to train MLHAD on **real experimental data**
- Extending MLHAD e.g. additional conditional labels allow:
  - $\rightarrow$ Additional conditional labels allow to generate hadron flavors with kinematic dependence
  - $\rightarrow$ Replacing PYTHIA's FS by a ML architecture
  - $\rightarrow$ Interpolating to different string energies
- Exploring different architectures (e.g. based on normalizing flows and RNNs)

[1] In collaberation with P. Ilten, T. Menzo, J. Zupan, M. Szewc and S. Mrenna

Back up

University of CINCINNATI

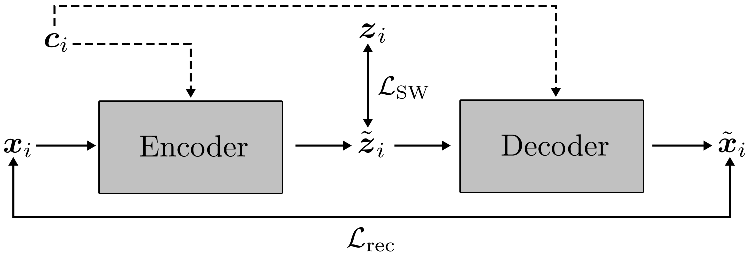
cSWAE Training-process

- conditioned on initial string energy $E_i \rightarrow c_i = (\bar{c}_i, 1 - \bar{c}_i)$:

$$E_i = E_{min}\bar{c}_i + E_{max}(1 - \bar{c}_i) \quad \Rightarrow \quad \bar{c}_i = \frac{E_{max} - E_i}{E_{max} - E_{min}}$$

- Encoder ϕ:
  - Input data $\mathbf{x_i}$ is a $N_e = 100$ dimensional vector, where $\mathbf{x_i} \in [p_{z,k}^{(i)}, p_{T,k}^{(i)}]$
  - p_z and p_T are uncorrelated and treated seperately
  - Takes as input x_i and c_i; returns the latent space vector $\bar{z}_i = \phi(x_i, c_i)$
- Decoder ψ:
  - Takes as input $\bar{z}_i$ and returns $\bar{x}_i = \psi(\phi(x_i, c_i))$
- Limit the training on light quark flavors and only pions as final state hadrons

cSWAE architecture



$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{N_{\mathrm{tr}}} \sum_{i=1}^{N_{\mathrm{tr}}} \left[\frac{1}{Q} d_2^2(\boldsymbol{x}_i, \psi(\phi(\boldsymbol{x}_i, \boldsymbol{c}_i))) + d_1(\boldsymbol{x}_i, \psi(\phi(\boldsymbol{x}_i, \boldsymbol{c}_i))) \right],$$

$$\mathcal{L}_{\mathrm{SW}} = \frac{\lambda}{L N_{\mathrm{tr}}} \sum_{\ell=1}^{L} \sum_{i=1}^{N_{\mathrm{tr}}} d_{\mathrm{SW}}(\boldsymbol{\theta}_\ell \cdot \boldsymbol{z}_{[i]_\ell}, \boldsymbol{\theta}_\ell \cdot \phi(\boldsymbol{x}_{[i]_\ell}, \boldsymbol{c}_i)),$$

cSWAE architecture