BERKELEY LAB
Bringing Science Solutions to the World

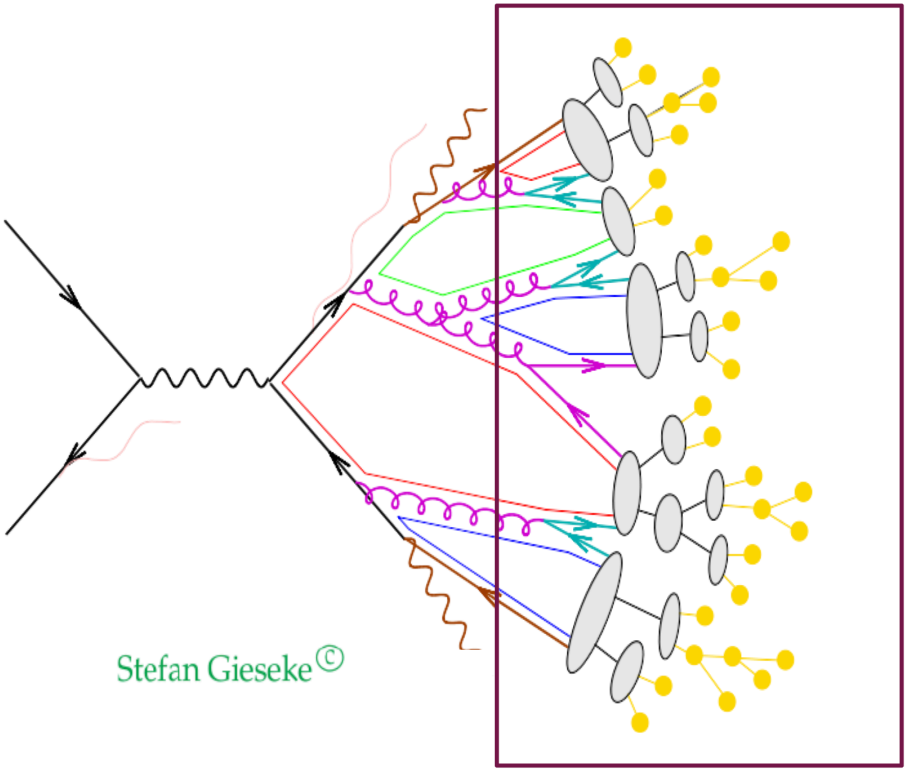
U.S. DEPARTMENT OF ENERGY
Office of Science

# A Deep Generative Model for Hadronization

Jay Chan[1], Xiangyang Ju[1], Adam Kania[2], Ben Nachman[1], Vishnu Sangli[1], Andrzej Siodmok[2]

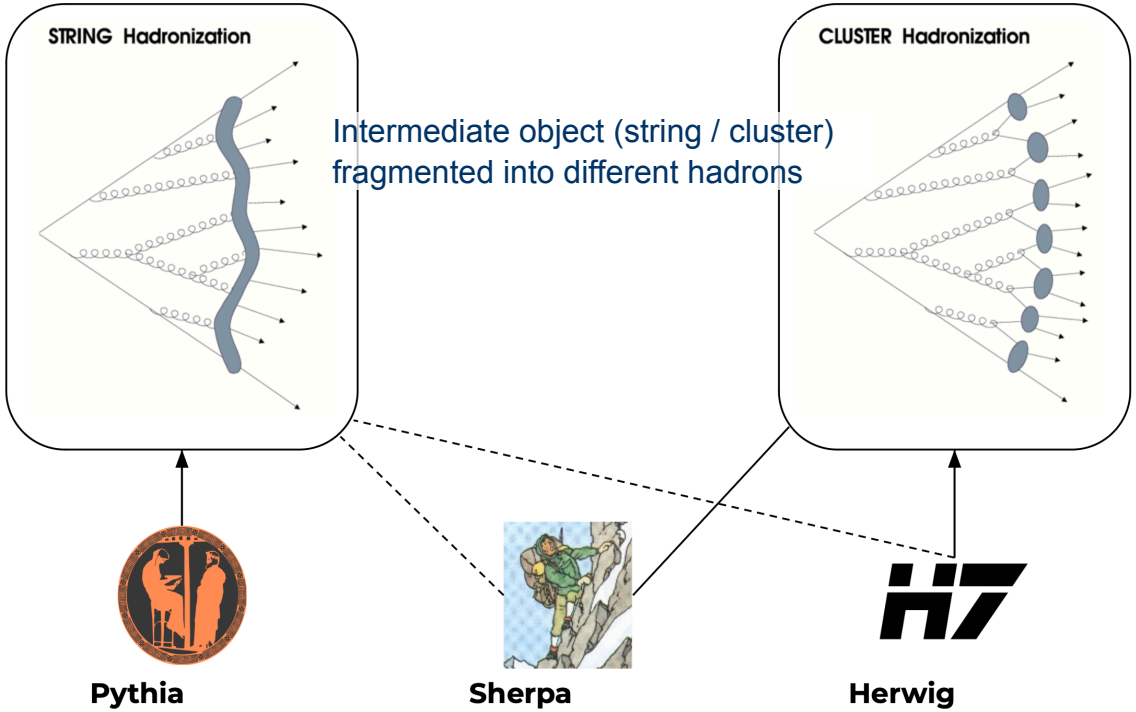
**[1]Lawrence Berkeley National Lab, [2]Jagiellonian University**

# Hadronization is one of the least understood problems

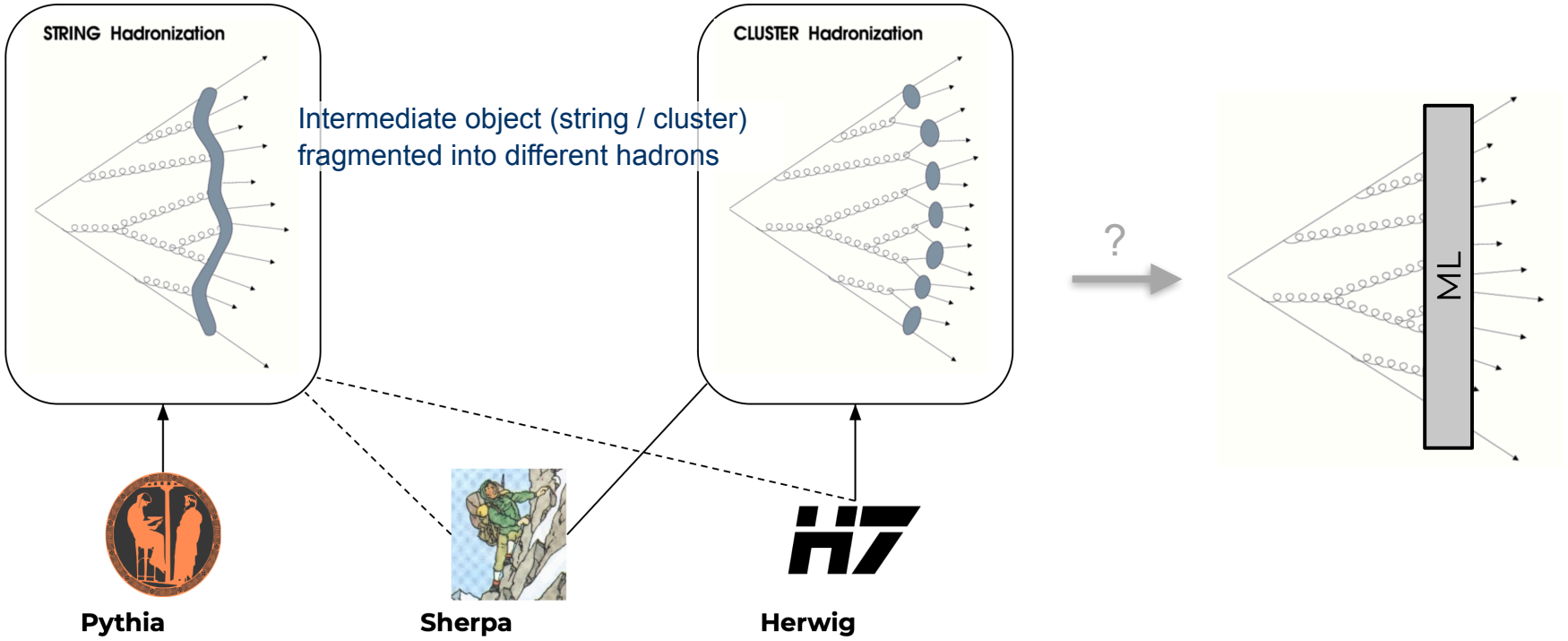



- Formation of hadrons out of quarks / gluons
- In MC simulation, this is done after parton shower
- The QCD of hadronization not yet fully understood

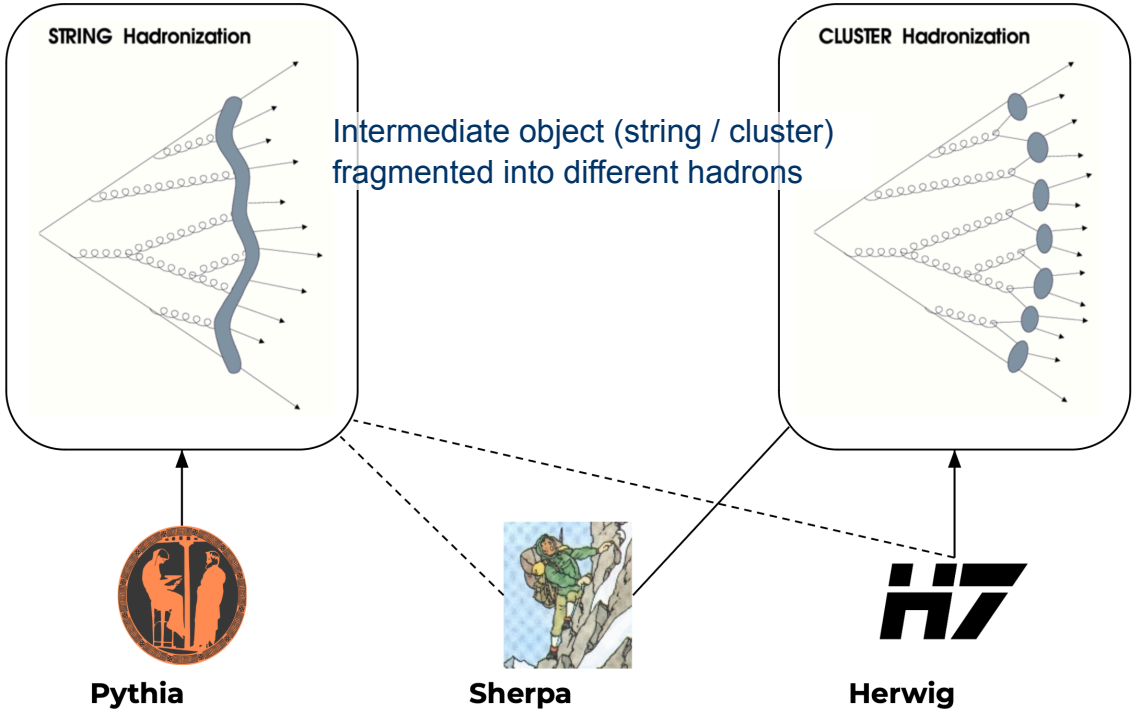
# Most common hadronization models are based on physically-inspired parametrization

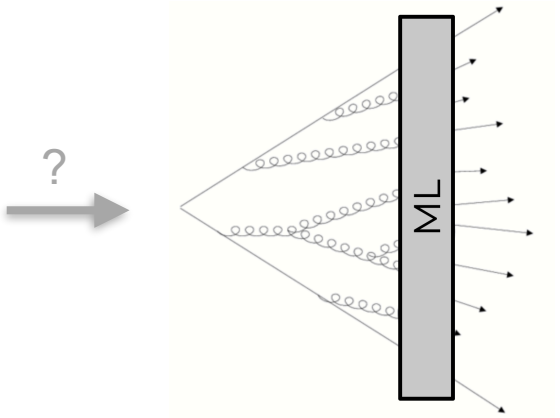


# Most common hadronization models are based on physically-inspired parametrization



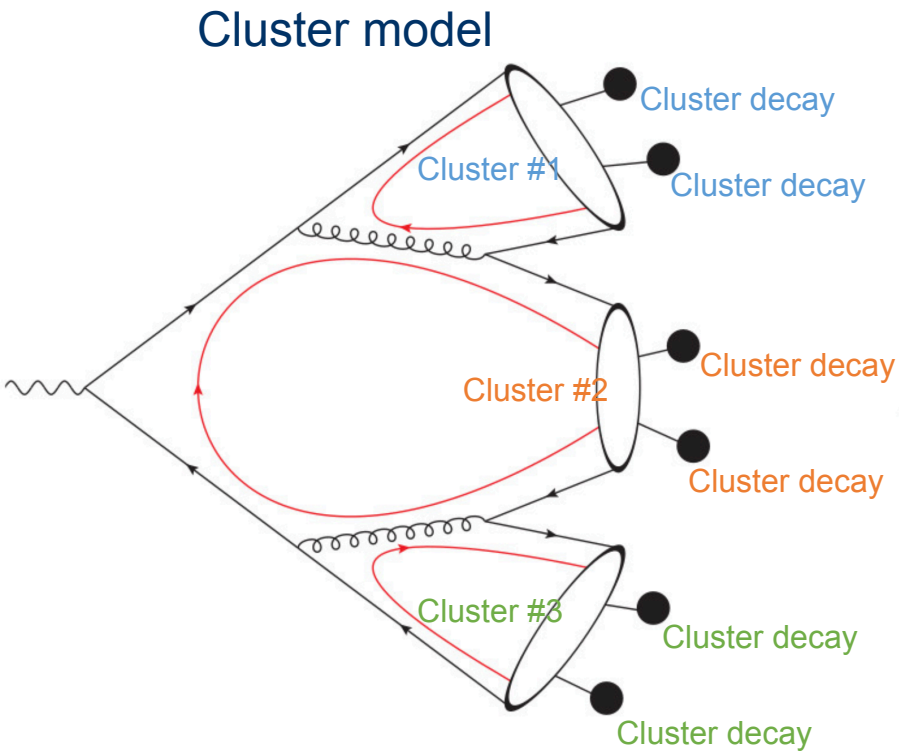
# Most common hadronization models are based on physically-inspired parametrization

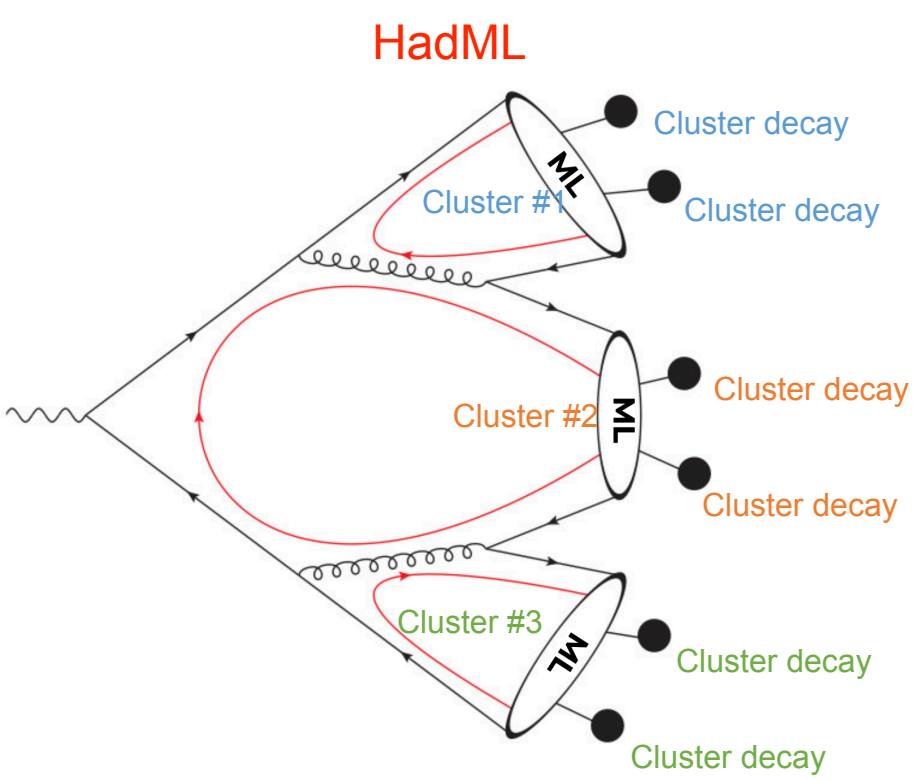


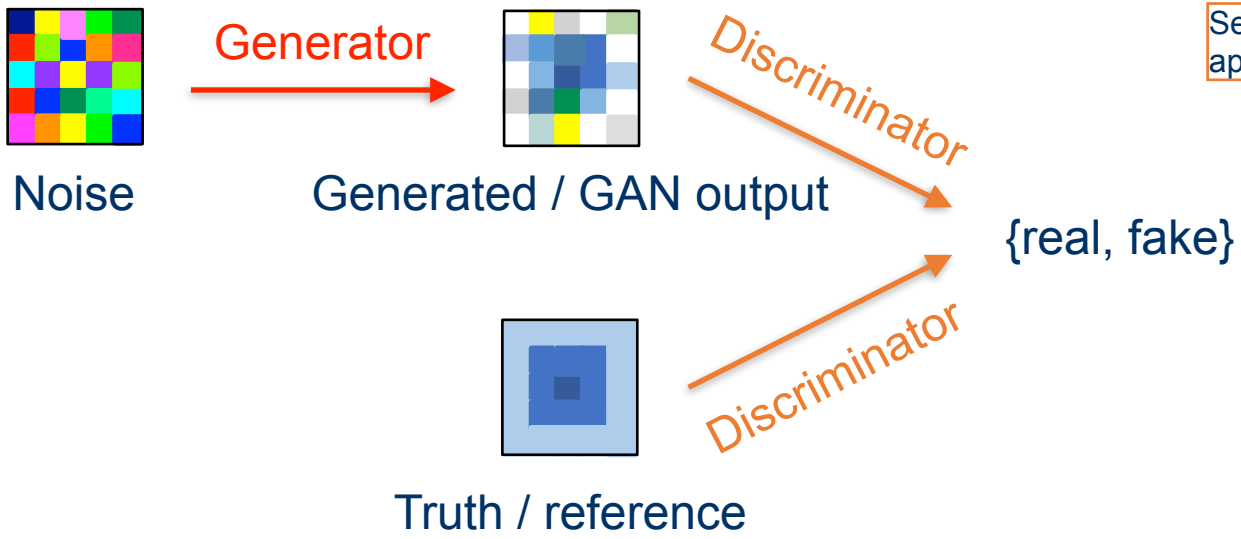

- More flexible
- Can work on unbinned data

# HadML: Generate cluster decays with neural networks



# Our tool of choice: Generative Adversarial Network (GAN)

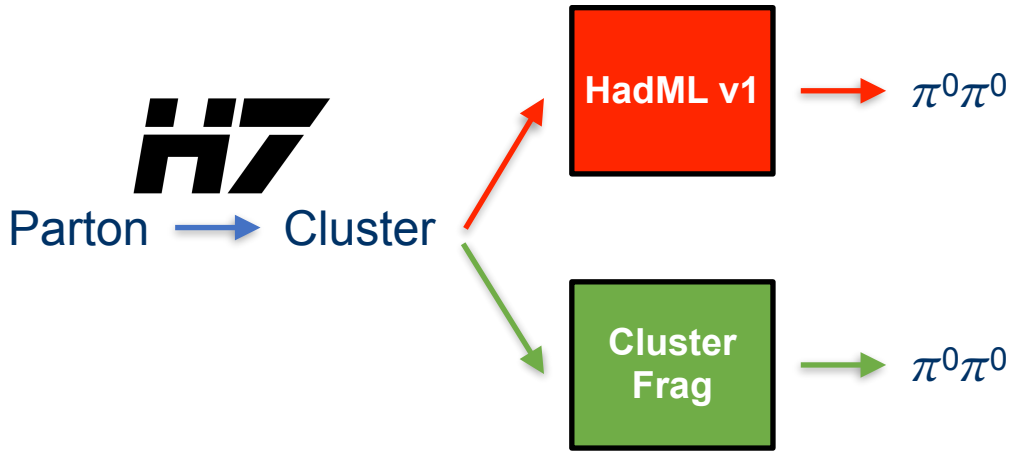



See Tony's presentation on the approach with normalizing flow

- A two-network game where one (generator) maps noise to structures and one (discriminator) classifies images as fake or real
- Allows to take observed data as reference (fit to data!)

# HadML v1: proof of concept

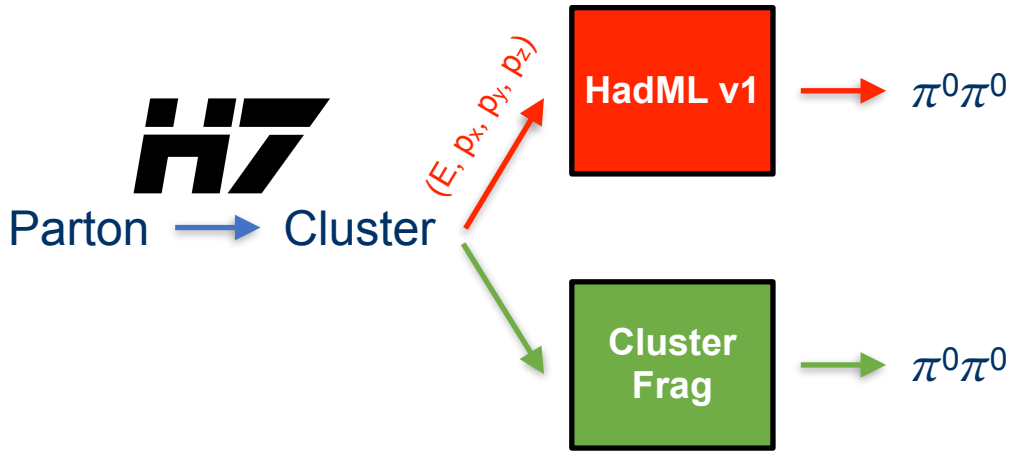
arXiv:2203.12660




- Take e+e- collision at 91.2 GeV simulated by Herwig7 as training sample
- Focus on two-body decays of clusters to pions (the majority)

# HadML v1: proof of concept

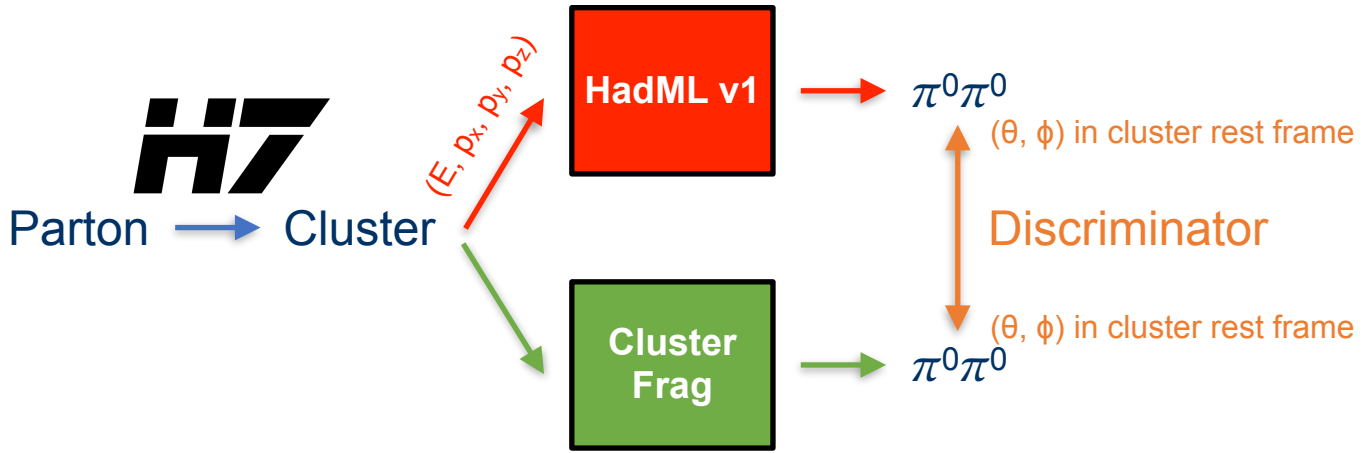
arXiv:2203.12660




- Take e+e- collision at 91.2 GeV simulated by Herwig7 as training sample
- Focus on two-body decays of clusters to pions (the majority)
- GAN model conditioned on cluster kinematics (4-momentum)

# HadML v1: proof of concept

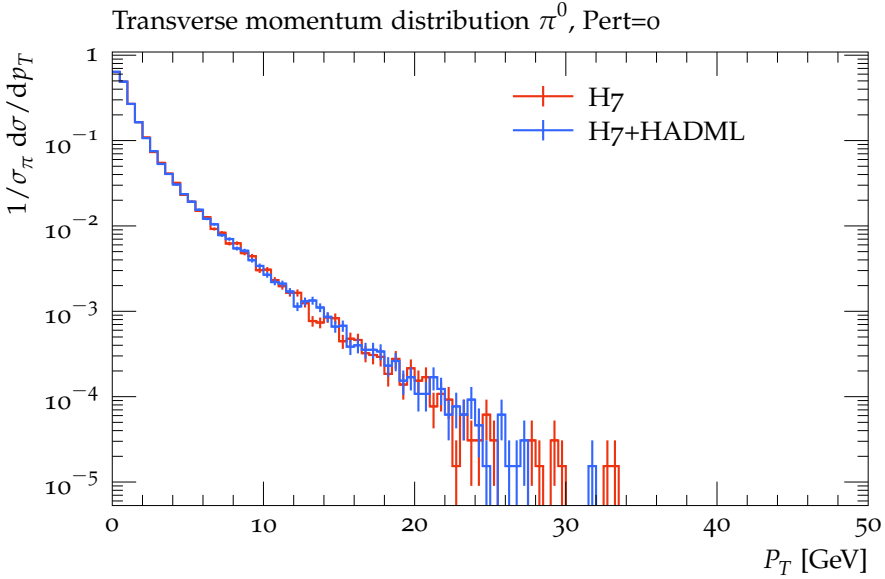
arXiv:2203.12660

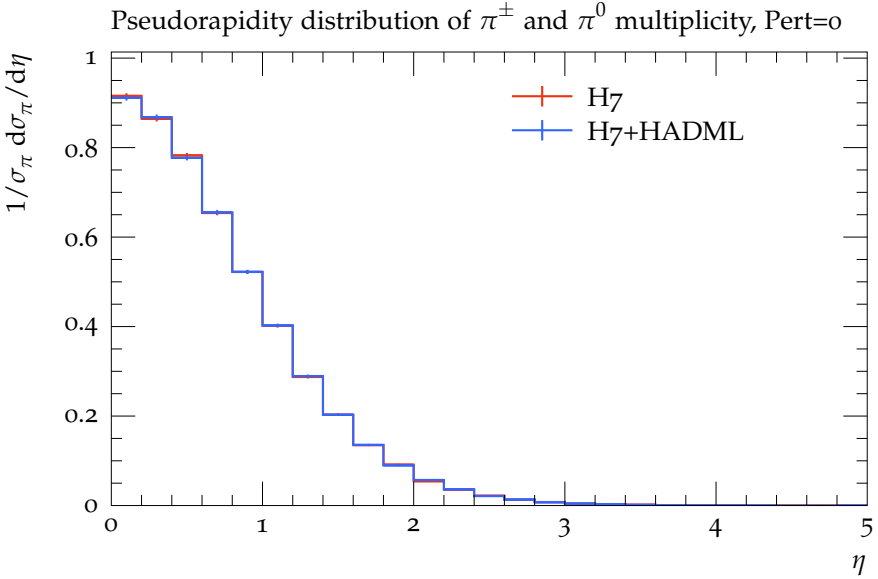



- Take e+e- collision at 91.2 GeV simulated by Herwig7 as training sample
- Focus on two-body decays of clusters to pions (the majority)
- GAN model conditioned on cluster kinematics (4-momentum)
- Discriminator sees the hadron kinematics per cluster (a pair of pions)

# Distribution of hadron kinematics is well learned

arXiv:2203.12660



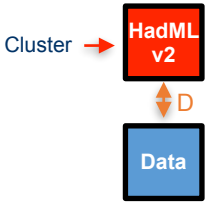




- Pion kinematics distributions generated by HadML v1 (Herwig clusters + HadML cluster decays) compared with pure Herwig7

# Problems with HadML v1

- The GAN model is fit to *Herwig7 simulation*
  - Discriminator takes each cluster (pion pair) as inputs but in real data all clusters are mixed together (we don't know which pions are paired together)
- All clusters have two-body decays
- All clusters decay to only pions

# Problems with HadML v1

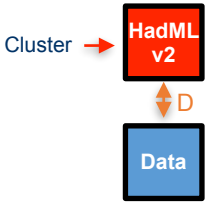



arXiv:2305.17169

**HadML v2** modifies the protocol to fit the GAN model in a realistic setting

- The GAN model is fit to *Herwig7 simulation*
  - Discriminator takes each cluster (pion pair) as inputs but in real data all clusters are mixed together (we don't know which pions are paired together)
- All clusters have two-body decays
- All clusters decay to only pions

# Problems with HadML v1




arXiv:2305.17169

**HadML v2** modifies the protocol to fit the GAN model in a realistic setting

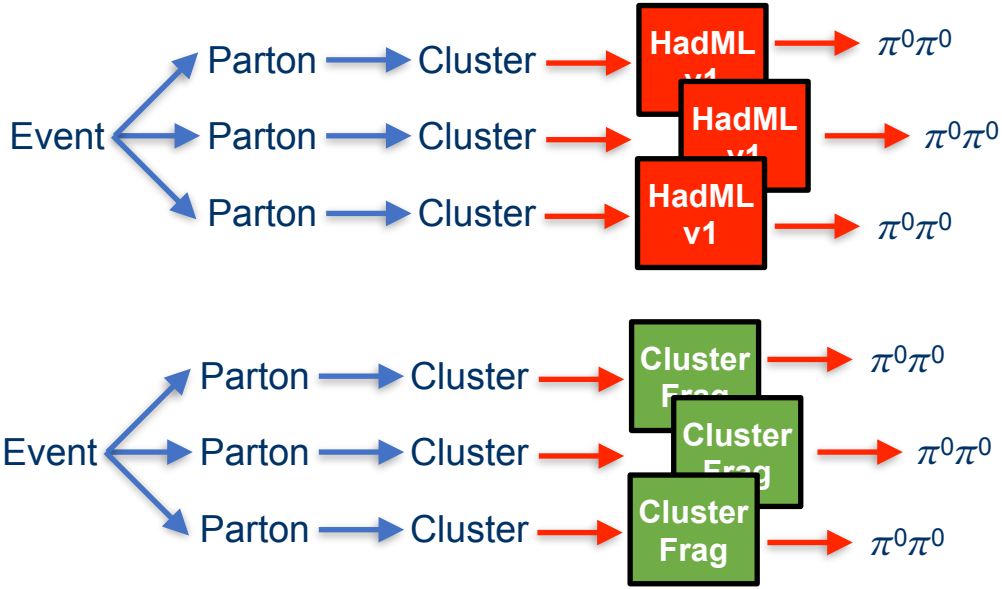
- The GAN model is fit to *Herwig7 simulation*
  - Discriminator takes each cluster (pion pair) as inputs but in real data all clusters are mixed together (we don't know which pions are paired together)
- All clusters have two-body decays
- All clusters decay to only pions


**HadML v3** adds hadron type into the generation

arXiv:2312.08453

# HadML v2: fitting GAN to data

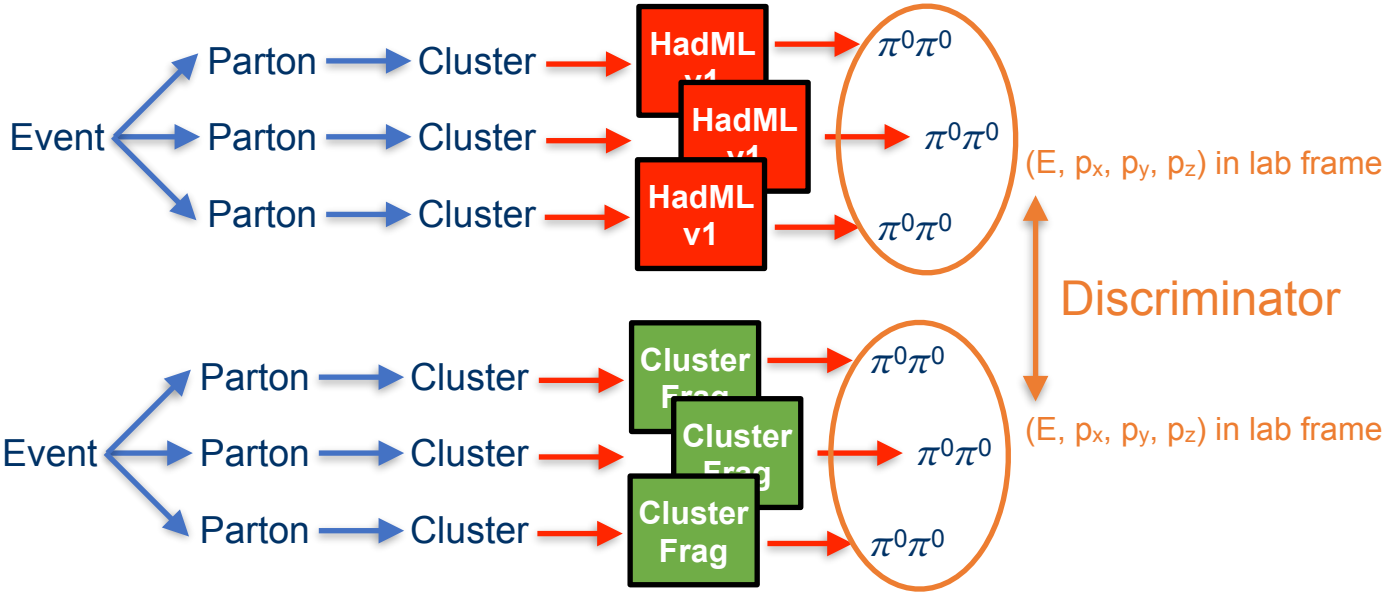
arXiv:2305.17169




- "Data" as in H7 simulation *without* cluster information

# HadML v2: fitting GAN to data

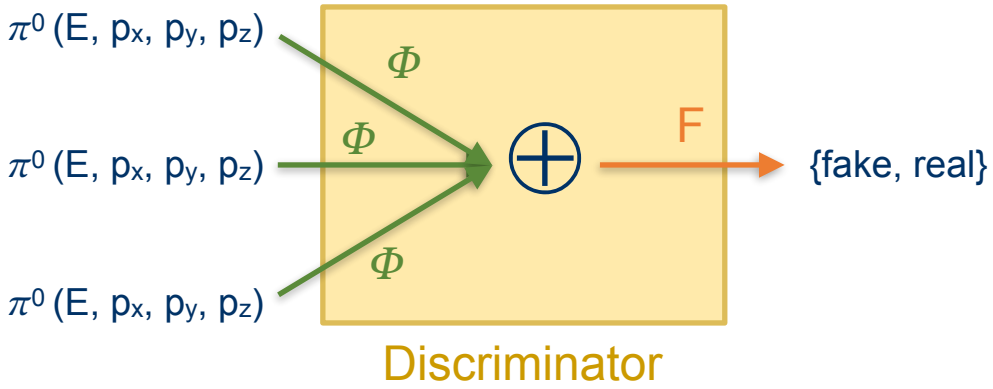
arXiv:2305.17169




- "Data" as in H7 simulation *without* cluster information
- Discriminator sees all hadrons in each event

# A deep set-based discriminator

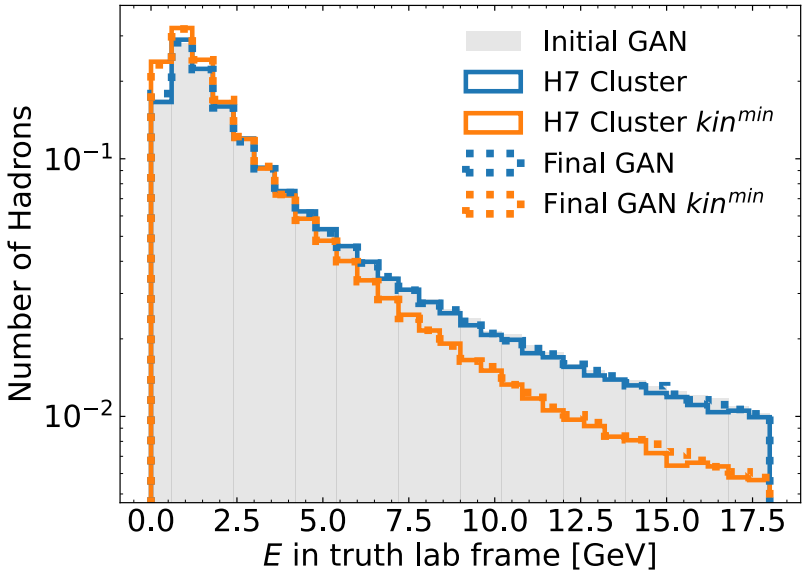
arXiv:2305.17169

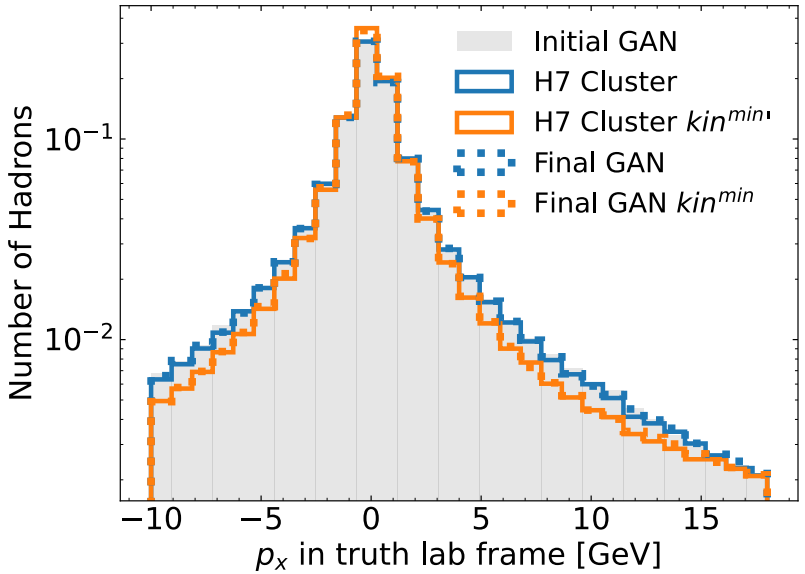



- Discriminator is a deep set model
- Invariant under permutation of hadrons

# Performance tested on independent datasets

arXiv:2305.17169

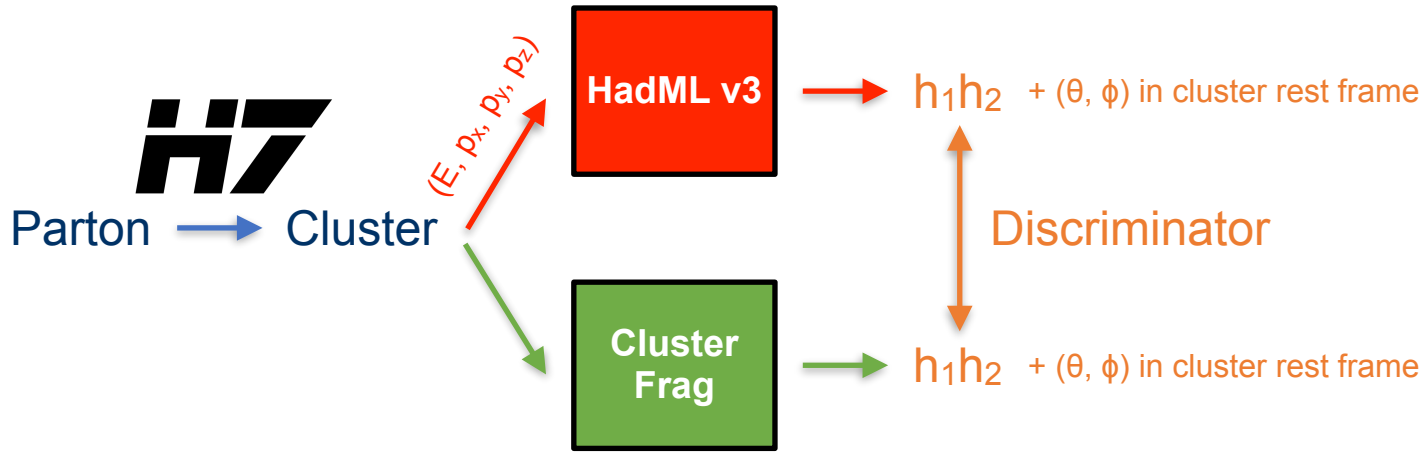



- Fit to two datasets with different cluster fragmentation settings (default and variation)
- The GAN models are able to adapt to different data distributions

# HadML v3: generate hadron types as well

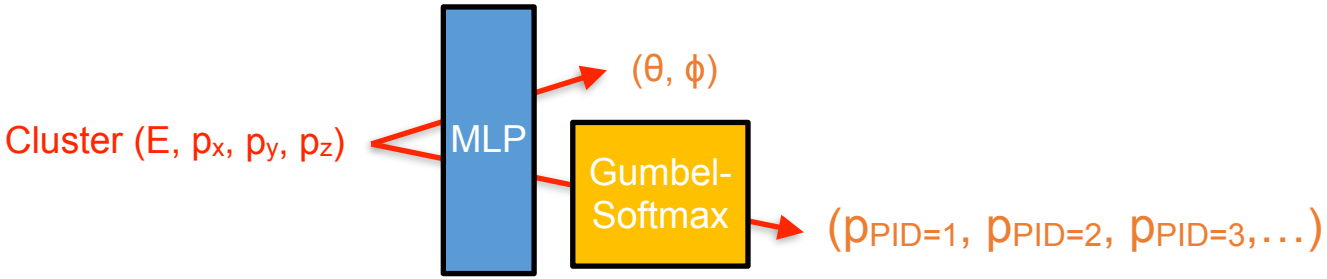
arXiv:2312.08453




- Same as HadML v1 but can predict hadron types other than pion
- Discriminator sees hadron kinematics as well as hadron types (h1, h2)

# Gumbel-Softmax for hadron type prediction

arXiv:2312.08453



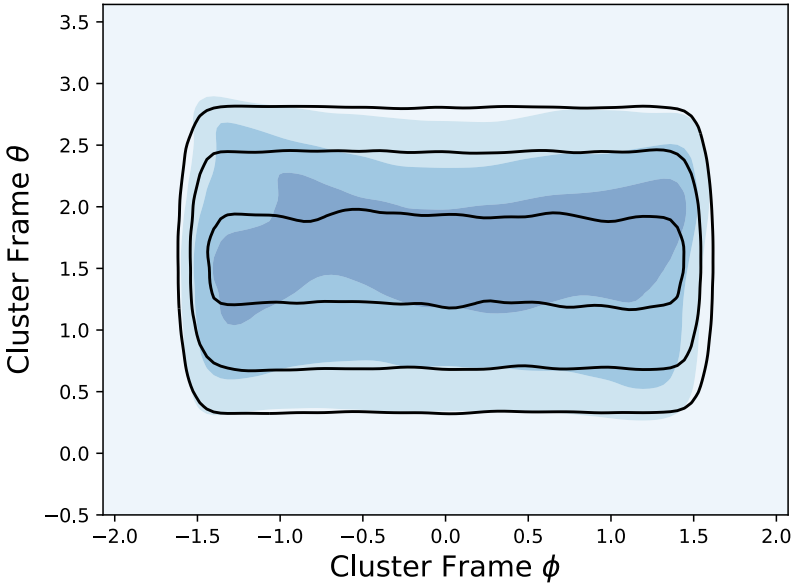

- Gumbel-Softmax activation used to approximate the hadron type distribution (discrete)

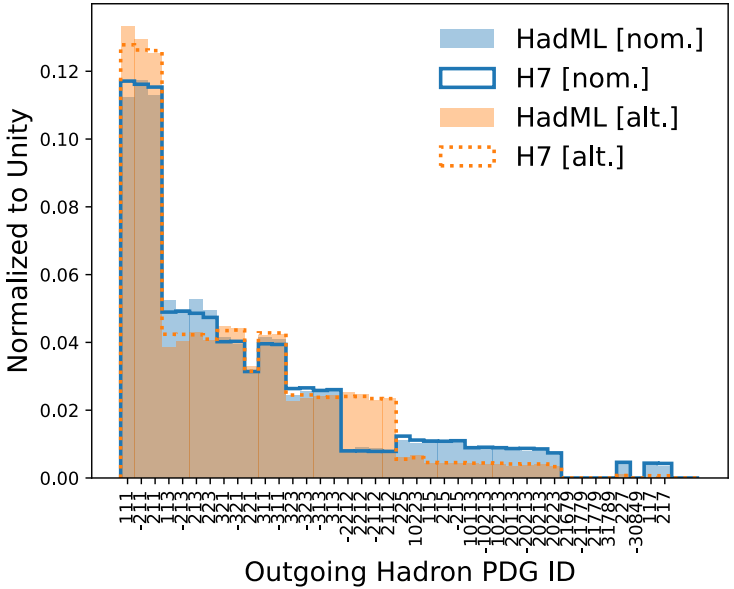
$$y_i = \frac{\exp\left(\left(\log \pi_i + g_i\right)/\tau\right)}{\sum_i \exp\left(\left(\log \pi_i + g_i\right)/\tau\right)}$$

- $g_i \sim \text{Gumbel}(0, 1)$
- $\tau$ decreases from 1 → 0.1 during training (0.1 for inference)

# Distribution of hadron types is well learned

arXiv:2312.08453

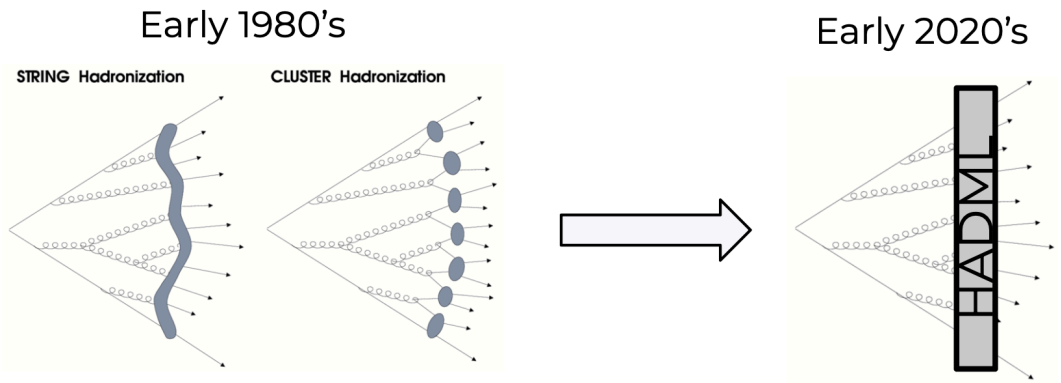



- Stress test with two different hadron type distributions (both perform well)

# What is next for HADML?

- HadML v3 can be combined with HadML v2 to fit the hadron type distributions to data
- Go beyond two-body decays (variable number of hadrons?)
- Increase model flexibility to accommodate strings model and beyond
- Hyperparameter optimization and explore alternative generative models
- A multi-year program ahead: ***stay tuned!***



# Backups