MACHINE LEARNING HADRONIZATION

Jure Zupan
U. of Cincinnati

MLHAD

based on 2203.04983, 2307.nnnnn, ...; in collaboration with Phil Ilten, Tony Menzo, Steve Mrenna, Manuel Szewc, Michael K. Wilkinson, Ahmed Youssef

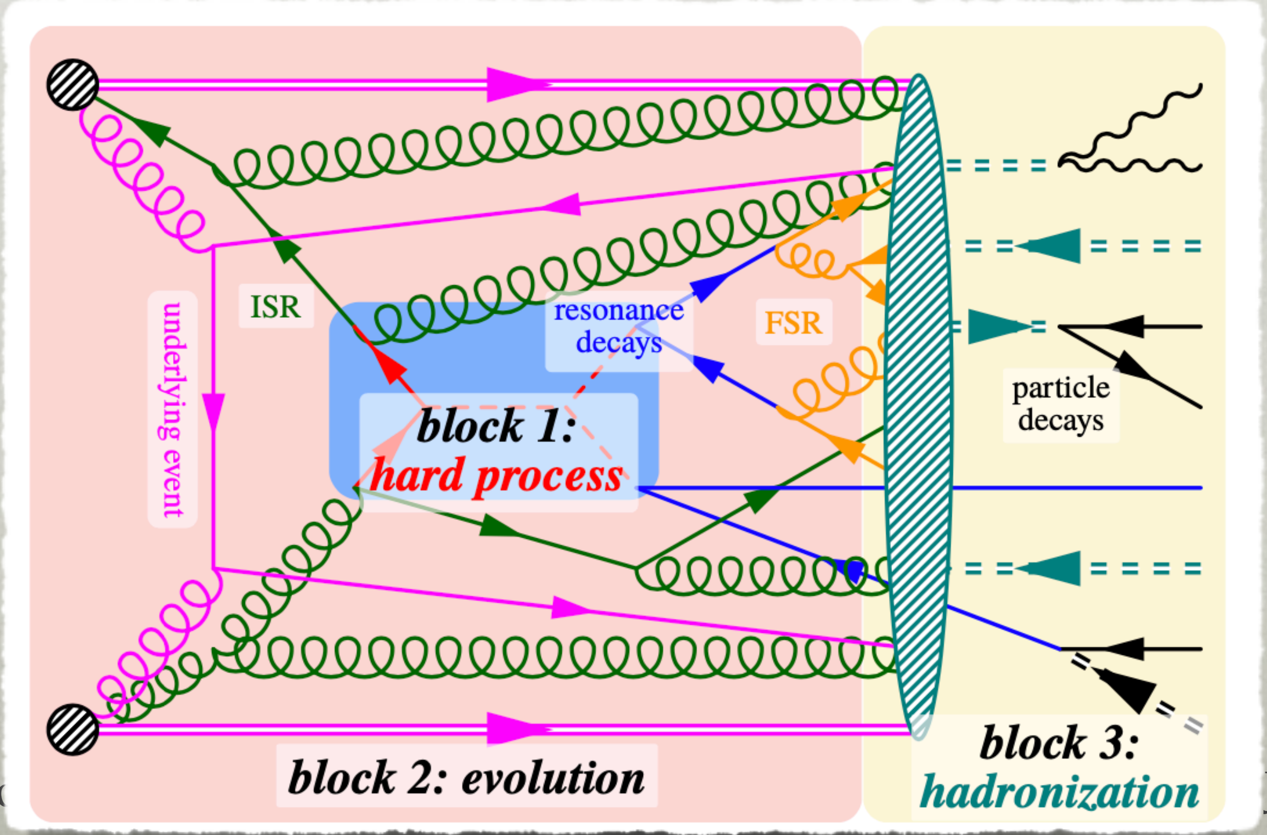
SM@LHC, Fermilab, July 10 2023

MONTE CARLO HEP EVENT

- block structure of HEP Monte Carlo
  - hard process
  - shower

  } under good perturbative control and systematically improvable

  - hadronization } modeling of nonperturbative physics
  - (detector simulation)



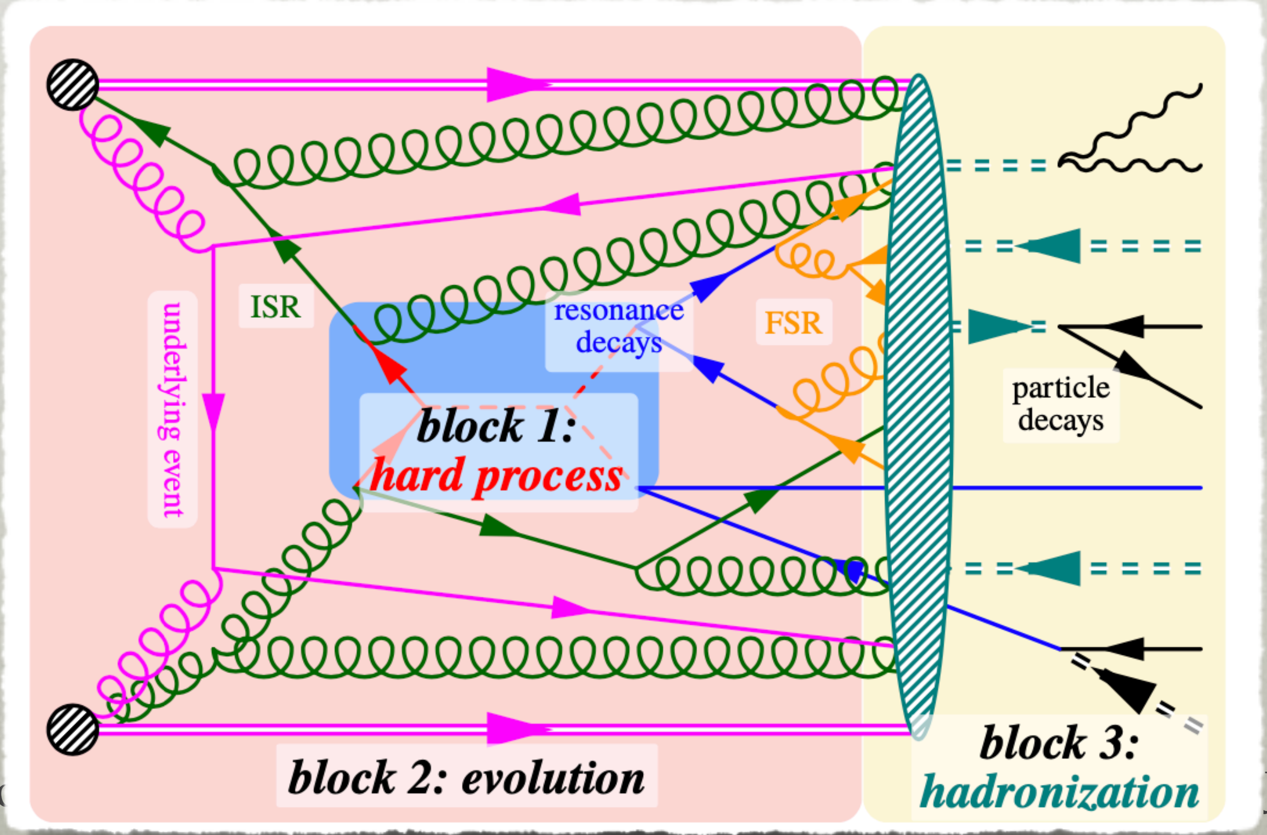
MONTE CARLO HEP EVENT

- block structure of HEP Monte Carlo
  - hard process
  - shower
  - hadronization
  - (detector simulation)

} under good perturbative control and systematically improvable

} modeling of no physics

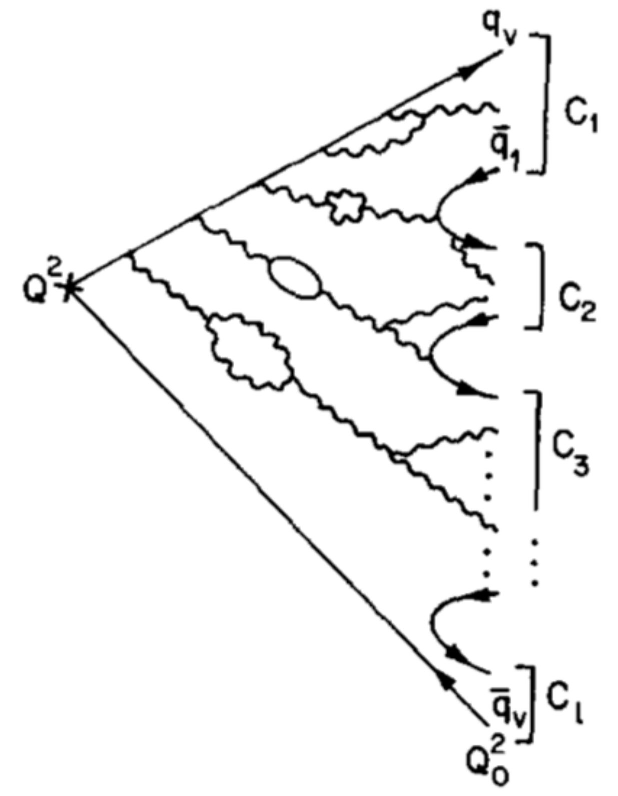
use Machine Learning



HADRONIZATION

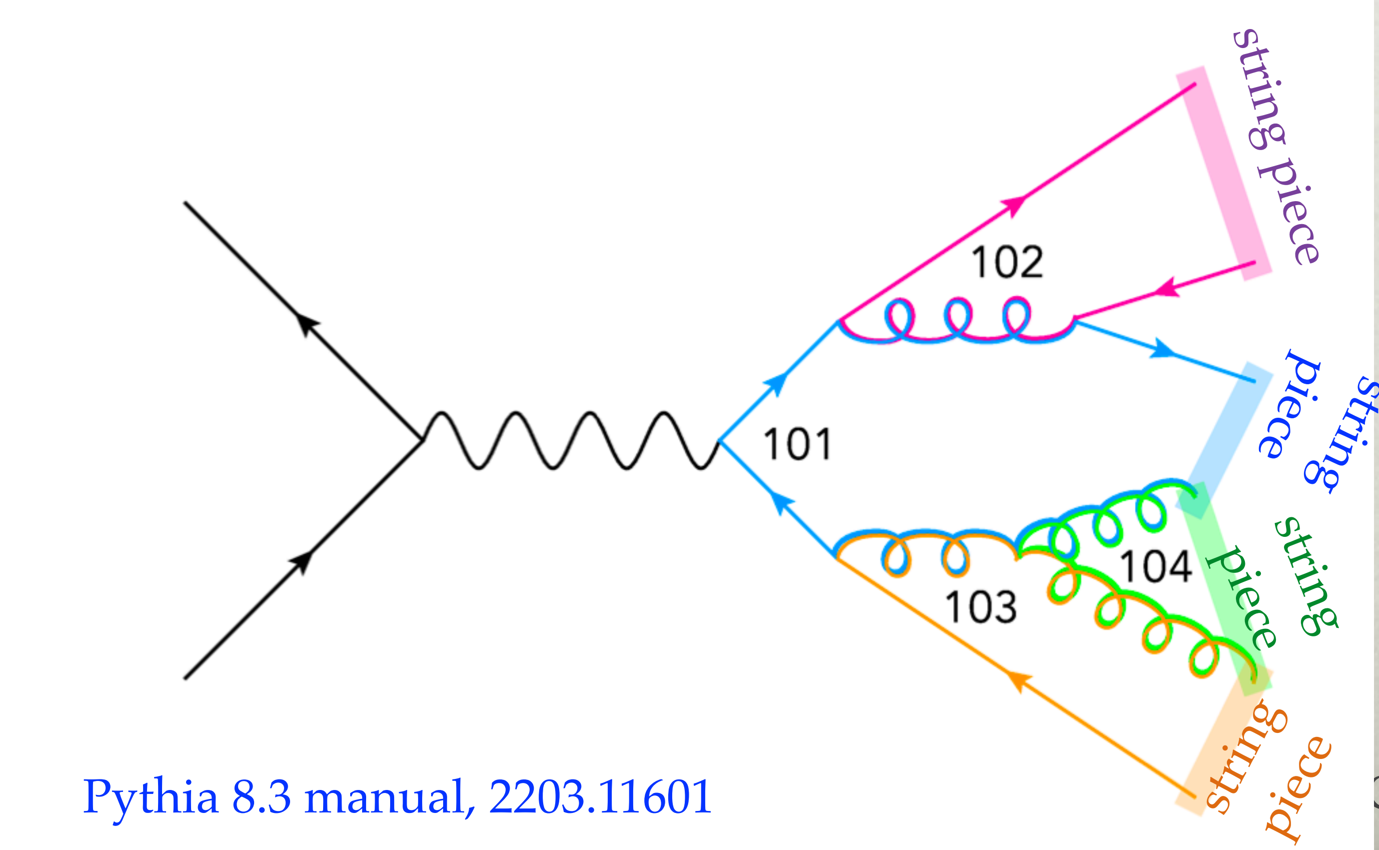
- two main models for hadronization
  - Lund string model (Pythia)
  - cluster hadronization model (Herwig)
- both have as a starting point "colour preconfinement" stage of QCD shower
  - stop shower at some scale Q_0
  - in large $N_c \to \infty$ limit planar graphs
  - groups final $q, \bar{q}, g$ in QCD singlet clusters

Amati, Veneziano, PLB83, 1979



LUND STRING MODEL

- strings connect $q\bar{q}$ systems
- gluons kinks in strings
  - split gluons to a collinear $q\bar{q}$ pair ⇒string pieces
- string pieces break into hadrons (model dep.)
  - controlled by Lund string fragmentation function
- Pythia Lund string model: many parameters, $\mathcal{O}(200)$
  - many of these related to color reconnection



Pythia 8.3 manual, 2203.11601

WHEN SHOULD WE CARE ABOUT HADRONIZATION?

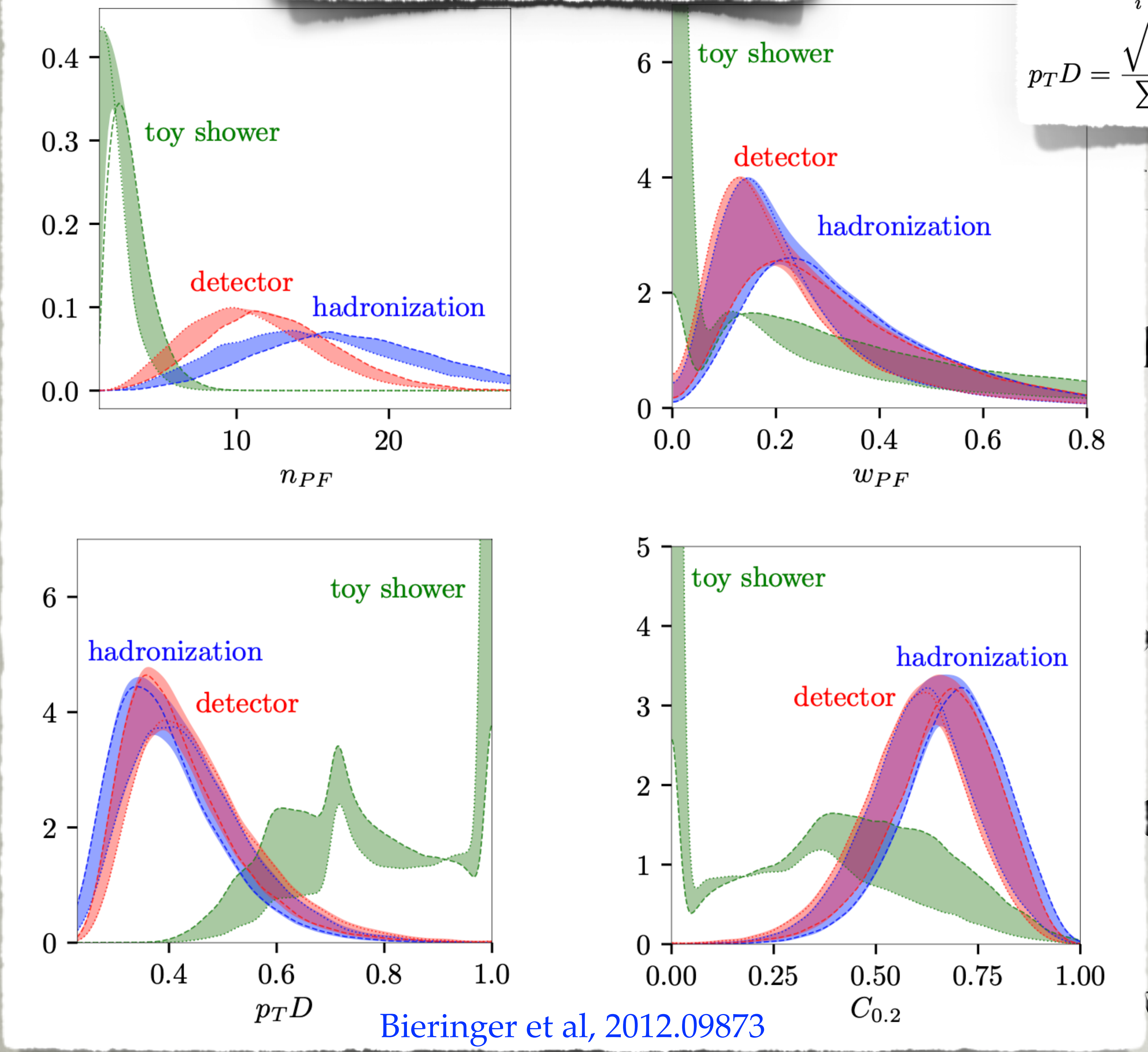
- if observables/measurements inclusive enough no need for modeling hadronization
- not the situation in the real world
  - experimental cuts, detectors not perfect, resonances decay in different ways
  - modeling well hadronization step essential for precision studies
- some measurements more sensitive than others
  - e.g., number of charged particles, correlations between exclusive states, etc.

$e^+e^- \to q\bar{q} \qquad \text{with} \quad q = u, d, s$

$$n_{\rm PF} = \sum_i 1 \qquad w_{\rm PF} = \frac{\sum_i p_{T,i} \Delta R_{i,\rm jet}}{\sum_i p_{T,i}}$$

$$p_T D = \frac{\sqrt{\sum_i p_{T,i}^2}}{\sum_i p_{T,i}} \qquad C_{0.2} = \frac{\sum_{ij} E_{T,i} E_{T,j} (\Delta R_{ij})^{0.2}}{\sum_i E_{T,i}^2} \, .$$

ZATION?



Bieringer et al, 2012.09873

lusive enough no

t perfect, resonances

tep essential for

ve than others

- e.g., number of charged particles, correlations between exclusive states, etc.

ML for hadronization

MLhad: Ilten, Menzo,Youssef, JZ, 2203.04983, https://gitlab.com/uchep/mlhad

see also HadML: (Chan, Ghosh,) Ju, (Kania), Nachman, (Sangli,) Siodmok, 2203.12660, 2305.17169

- MLhad: the long term goal
  - use ML to "parametrize our ignorance" about hadronization, use data
- more immediate
  - reproduce simplified version of Pythia Lund string model

ML for hadronization

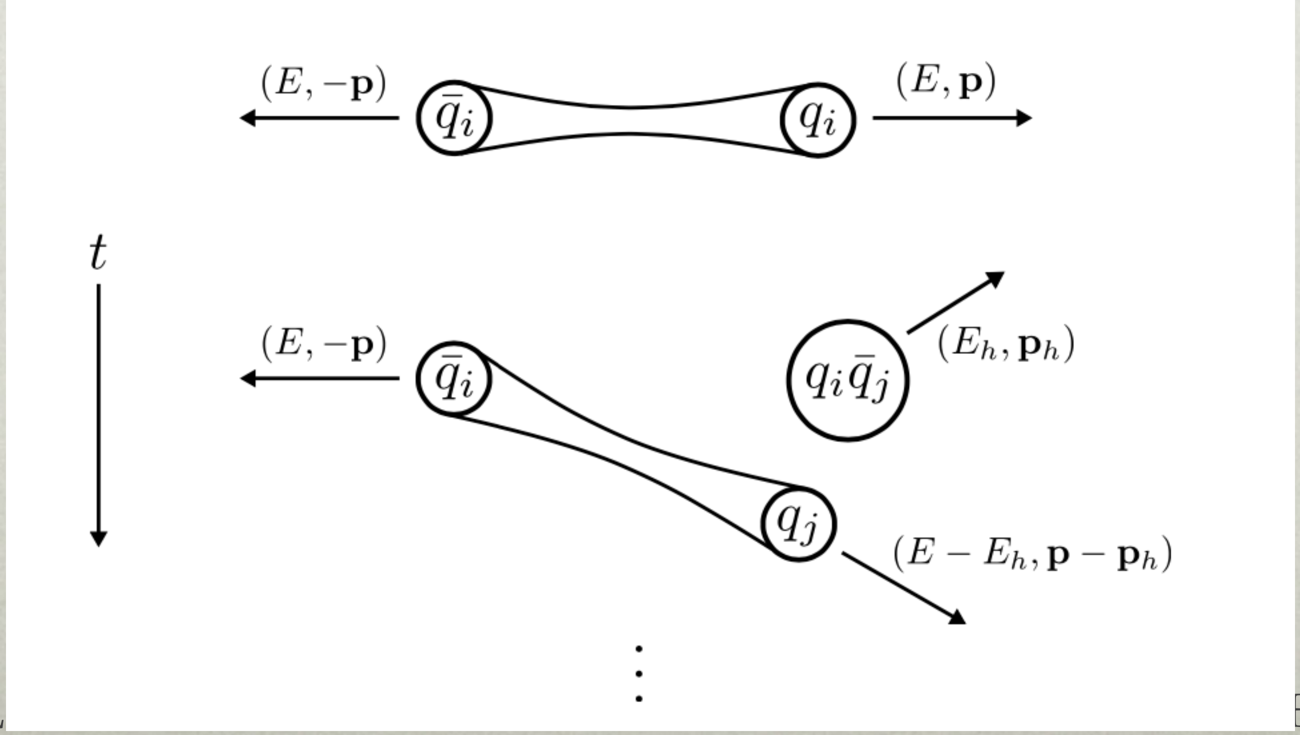
- a series of progressive steps to be done before practically useful in Pythia/MC simulations
  - ML architecture that mimicks a simplified Lund string hadronization model
    - train ML on truth level Pythia output (not obs. in exp)
  - develop a framework to propagate errors
  - improved ML architecture with full hadron flavor selector
  - train on mock data (i.e. just observable information)
  - train on real data (i.e. just already measured information)
  - replace/supplement Pythia string model

we are here

Ilten, Menzo,Youssef, JZ, 2203.04983

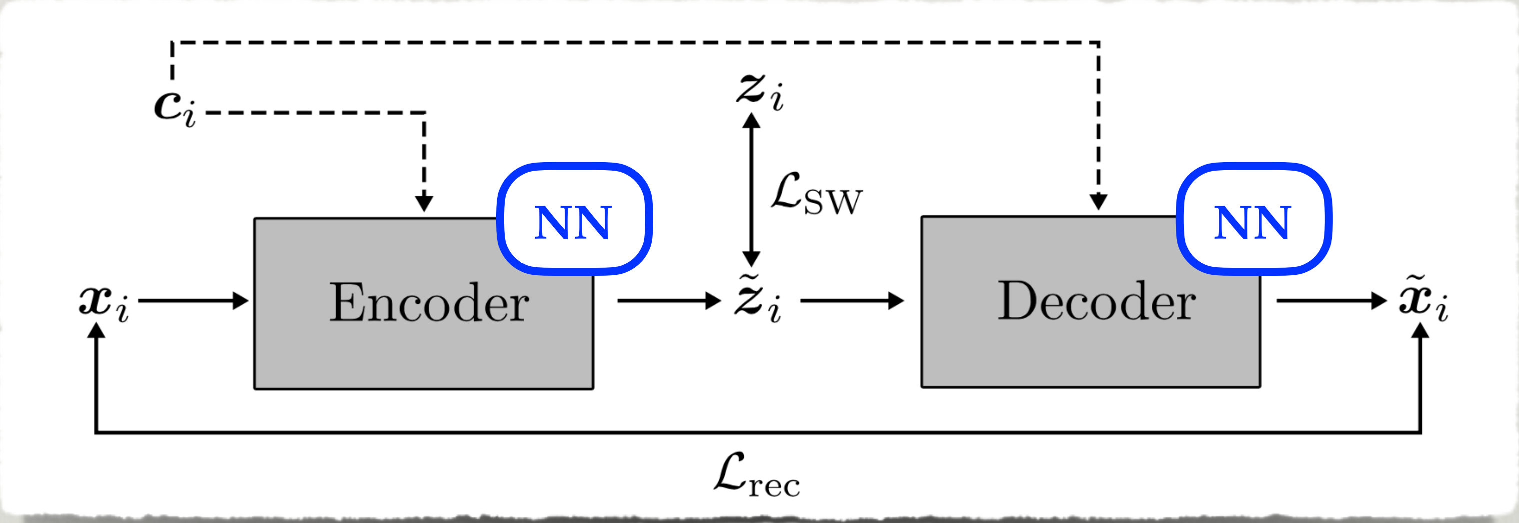
SIMPLIFIED STRING HADRONIZATION MODEL

- assume that color reconnection done correctly by Pythia
- want to reproduce first hadron emission from a string piece with $q, \bar{q}$ ends
  - the whole hadronization chain is then reproduced by iterating
  - the string is labeled by $q, \bar{q}$ flavor and its energy in cms, $2E$
- simplified flavor selector: only emission of pions
- have an IR cut-off of 5 GeV, at which hadronization chain terminates



cSWAE

- use conditional Sliced-Wasserstein Autoencoder
  - SW gives flexibility in the use of latent space distributions


$\boldsymbol{c}_i$
$\boldsymbol{z}_i$
$\mathcal{L}_{\mathrm{SW}}$
NN
NN
$\boldsymbol{x}_i$
Encoder
$\tilde{\boldsymbol{z}}_i$
Decoder
$\tilde{\boldsymbol{x}}_i$
$\mathcal{L}_{\mathrm{rec}}$

- string energy E_i is encoded in a label $\bar{c}_i$

$$\bar{c}_i = \frac{E_{\mathrm{max}} - E_i}{E_{\mathrm{max}} - E_{\mathrm{min}}},$$

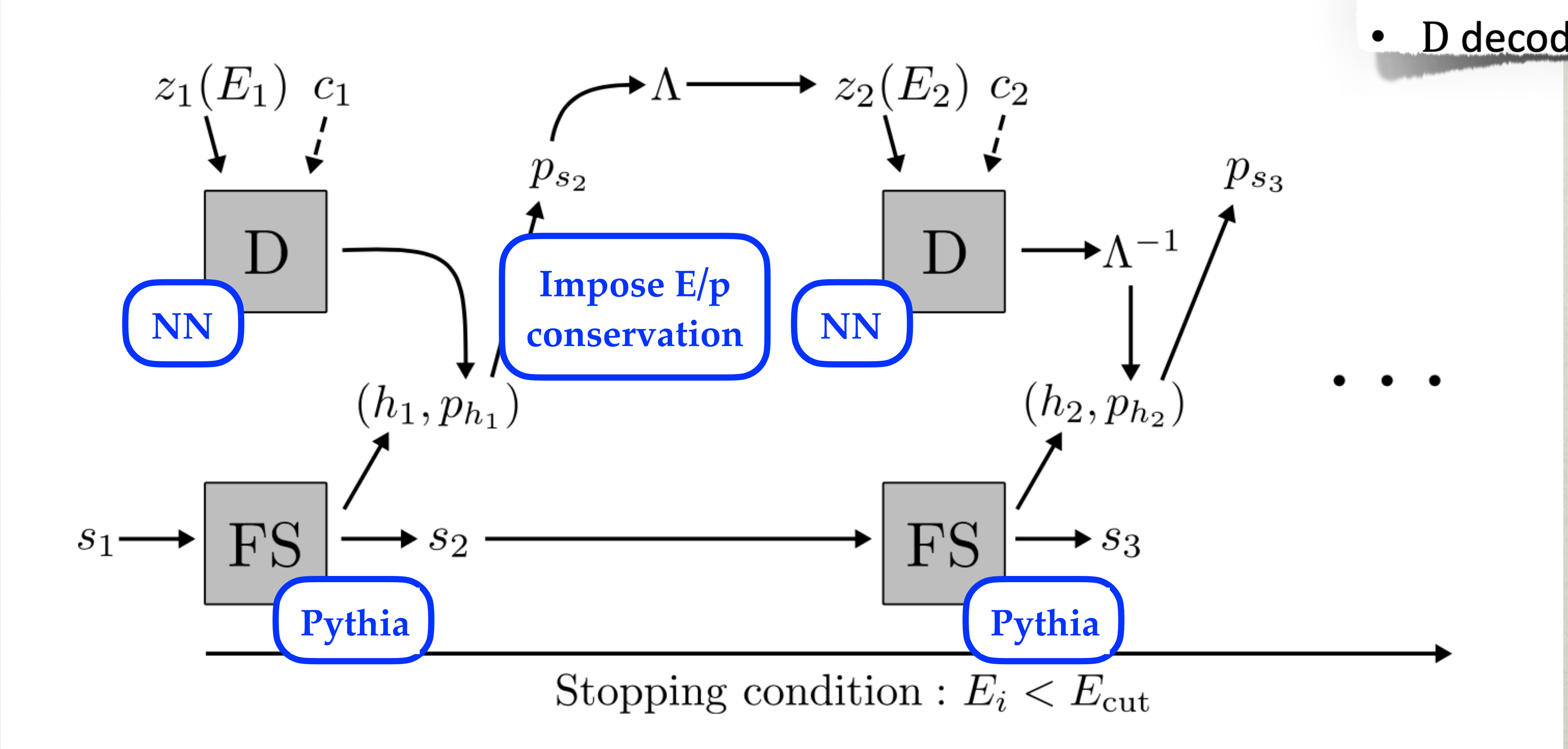
- training data: $\mathbf{x}_i$ sorted vector of 100 first emission
  - either p_z or p_T values
- loss function

$$\mathcal{L}(\psi, \phi) = \mathcal{L}_{\mathrm{rec}} + \mathcal{L}_{\mathrm{SW}},$$

MLhad as a generator

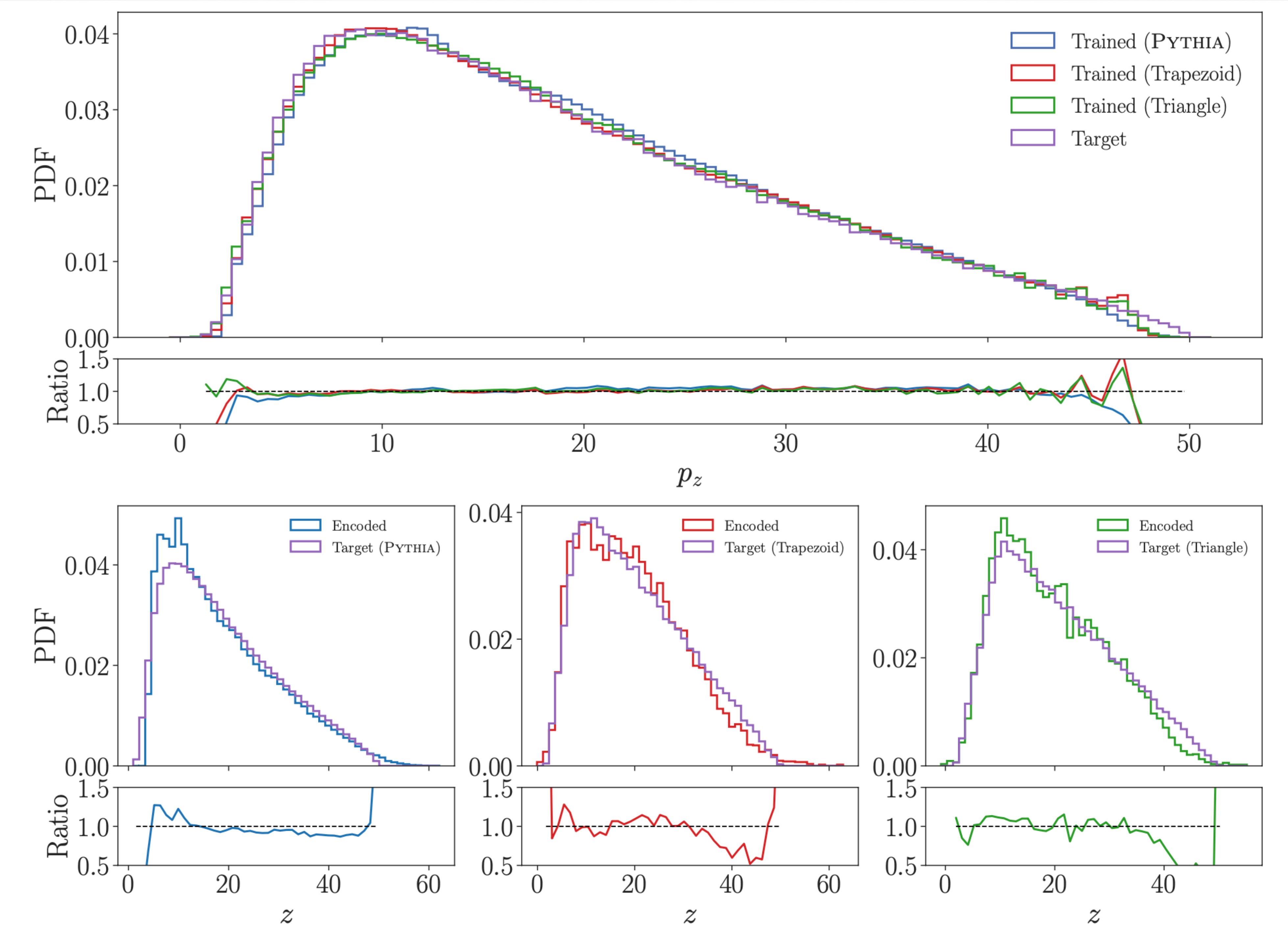
- MLhad as a generator of the hadronization chains

- h_i hadron
- s_i string fragment
- p_j 4-momentum
- Λ Lorentz transform
- FS flavor-selector
- D decoder



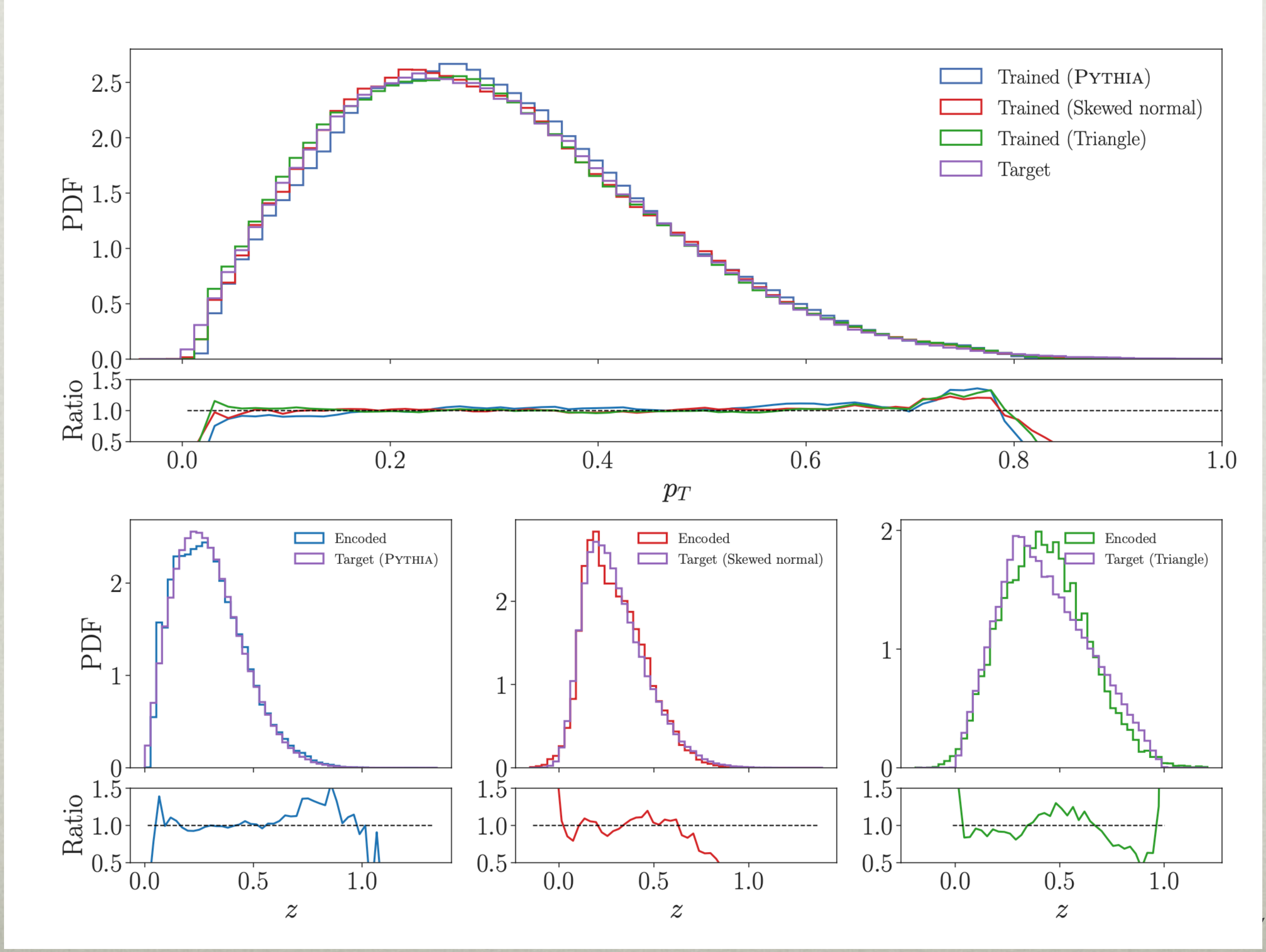
RESULTS - FIRST EMISSION

- MLhad generated p_z distribs.



RESULTS - FIRST EMISSION

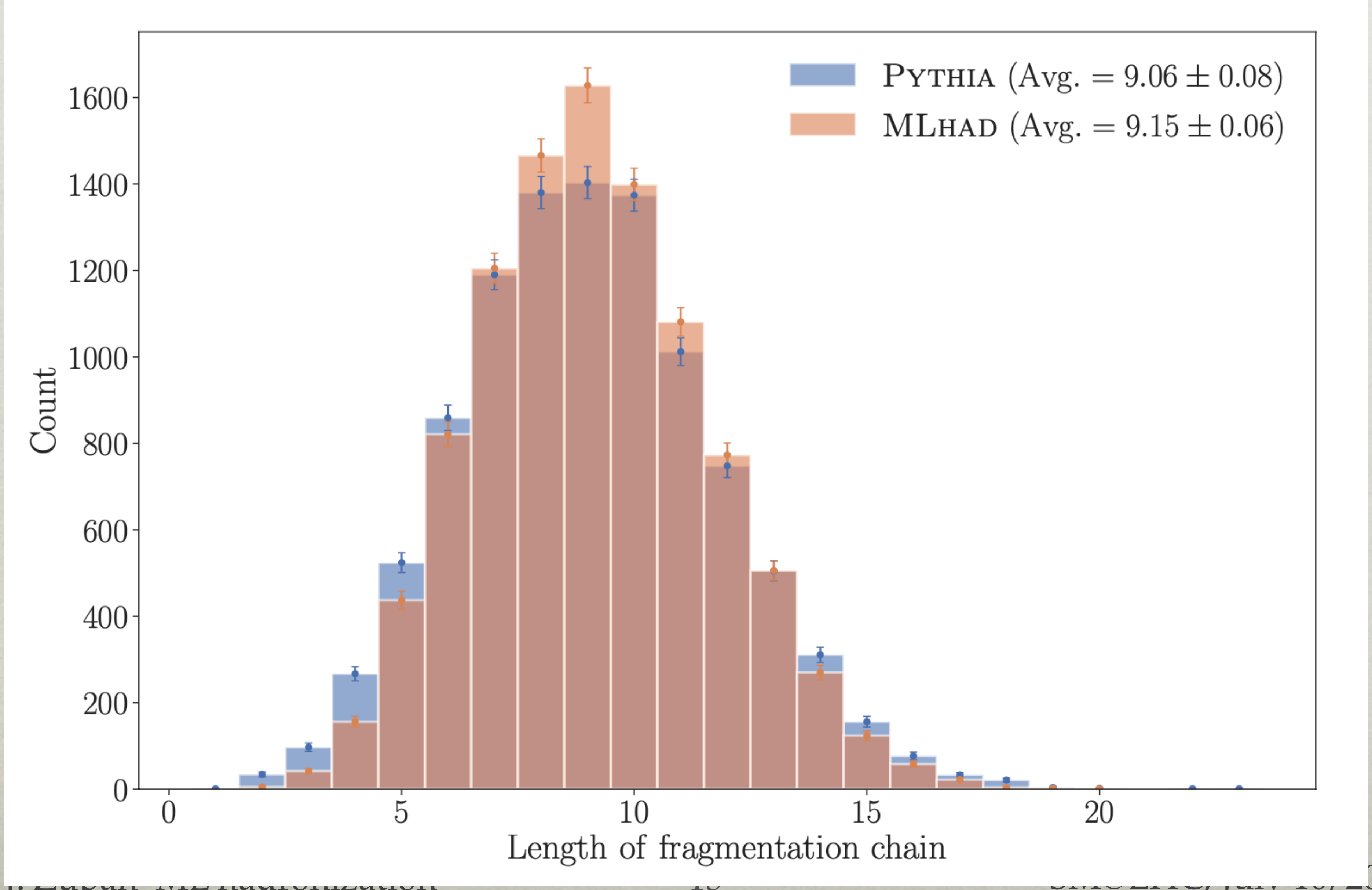
- MLhad generated p_T distribs.



2023

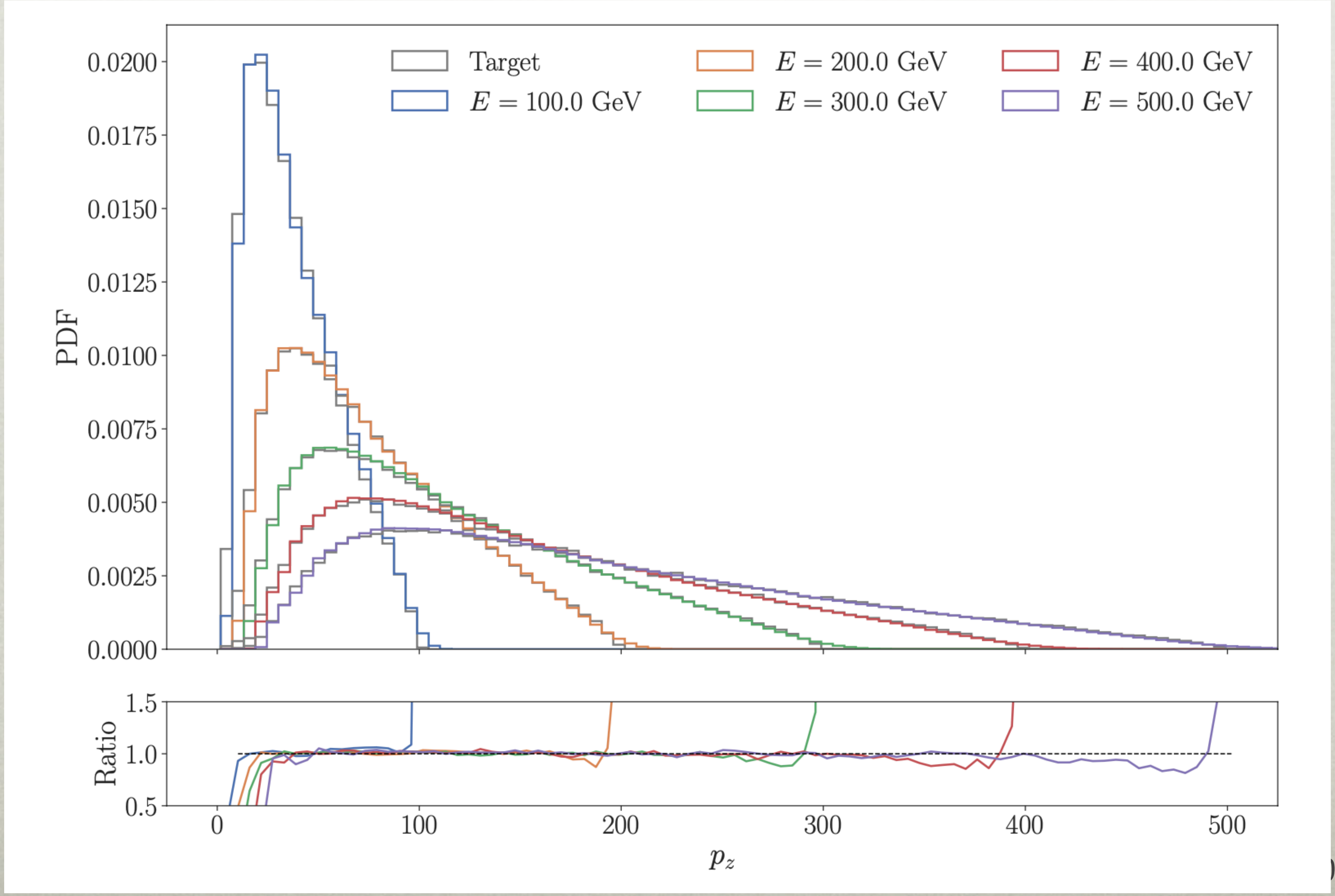
GENERATING HADRONIZATION CHAINS

- number of hadrons produced in hadronization of 50 GeV string



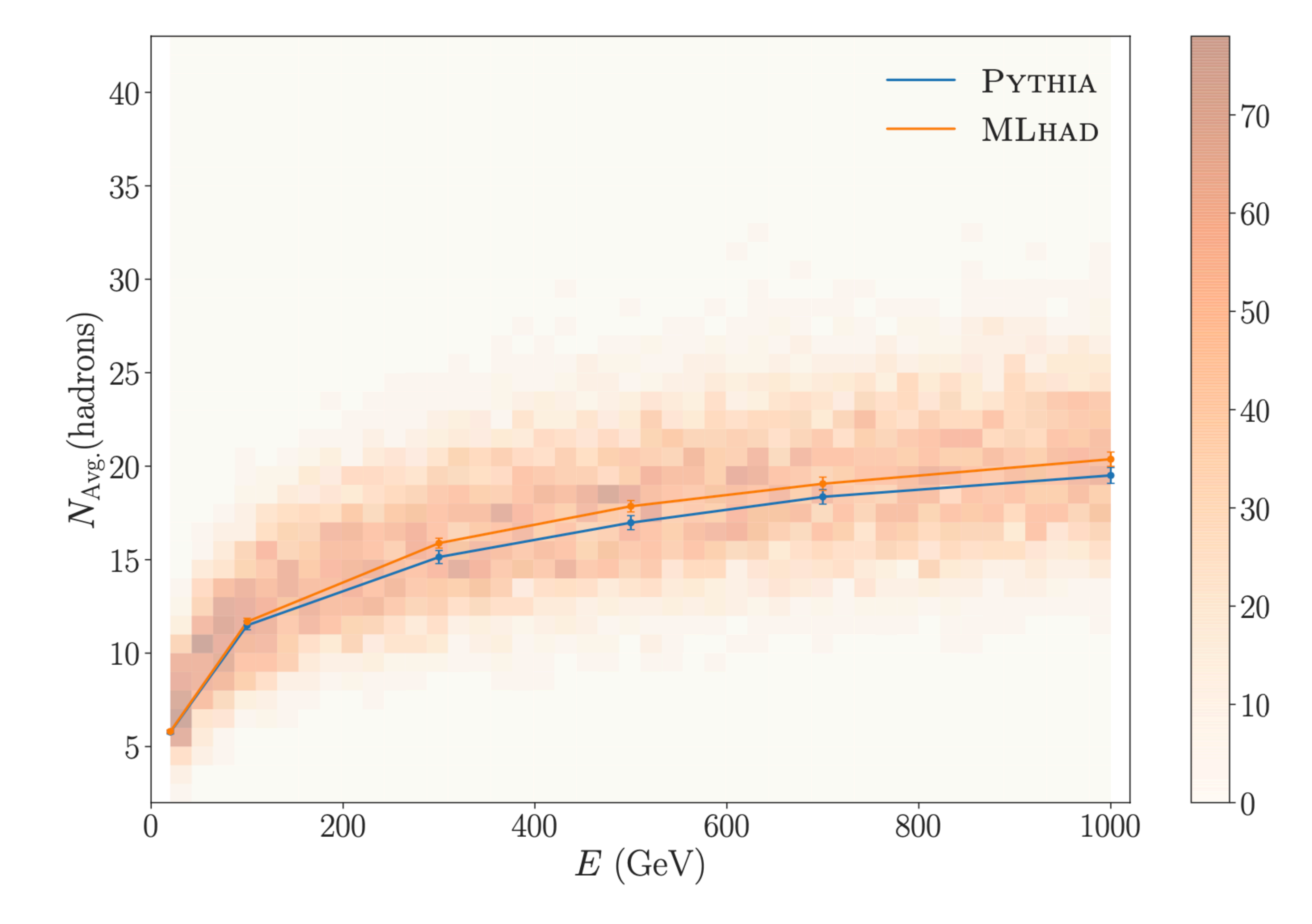
E dependent distributions

- train on first hadron emissions at $E = \{5, 30, 700, 1000\}$ GeV
- generate at a different set of string energies



GENERATING HADRONIZATION CHAINS

- the distributions match over a range of string energies



RECAP

- MLhad architecture captures well (simplified) Pythia Lund string model
- proof of principle - need to see how this ports to training on data

NEXT STEPS

- to train on data
  - want fast evaluation of parameter dependency
  - use reweighting method
  - first implementation in Pythia for Lund string model (to be released soon in Pythia)

Ilten et al, 2307.nnnnn

- propagation of errors
  - alternative ML architecture with Bayesian normalizing flows

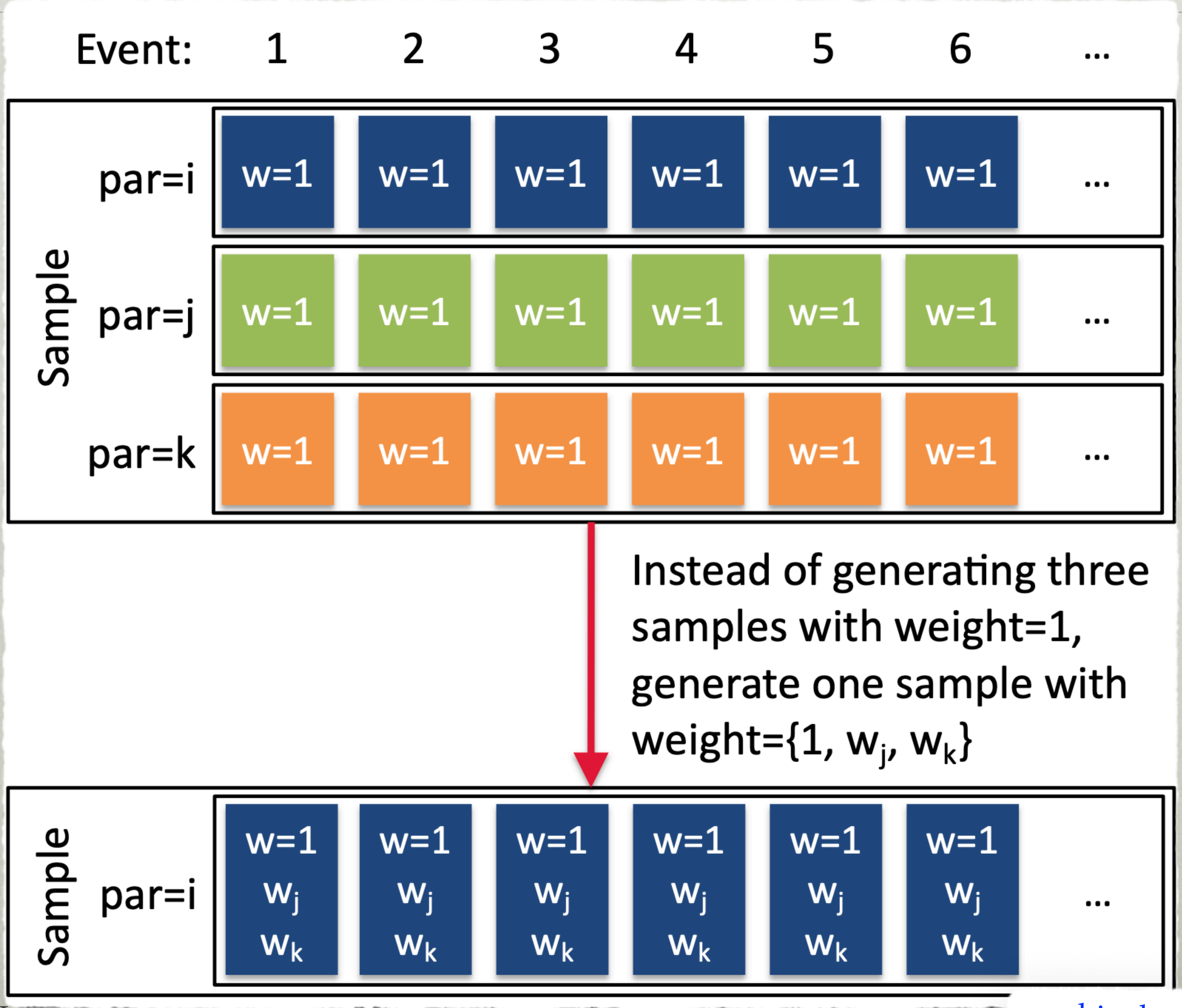
Ilten et al, 2308.nnnnn, see backup slides

REWEIGHTING HADRONIZED PYTHIA EVENTS

Ilten et al, 2307.nnnnn

- event generation is time-consuming
  - want to reweight events without regenerating
- in Pythia the Lund string fragmentation function sampled via standard veto algorithm
  - if rejected instances are kept ⇒
  - a modified veto algorithm ⇒ new event weights for diff. hadronization params.

REWEIGHTING HADRONIZED PYTHIA EVENTS



Event:	1	2	3	4	5	6	…
Sample par=i	w=1	w=1	w=1	w=1	w=1	w=1	…
Sample par=j	w=1	w=1	w=1	w=1	w=1	w=1	…
Sample par=k	w=1	w=1	w=1	w=1	w=1	w=1	…

Instead of generating three samples with weight=1, generate one sample with weight={1, w_j, w_k}

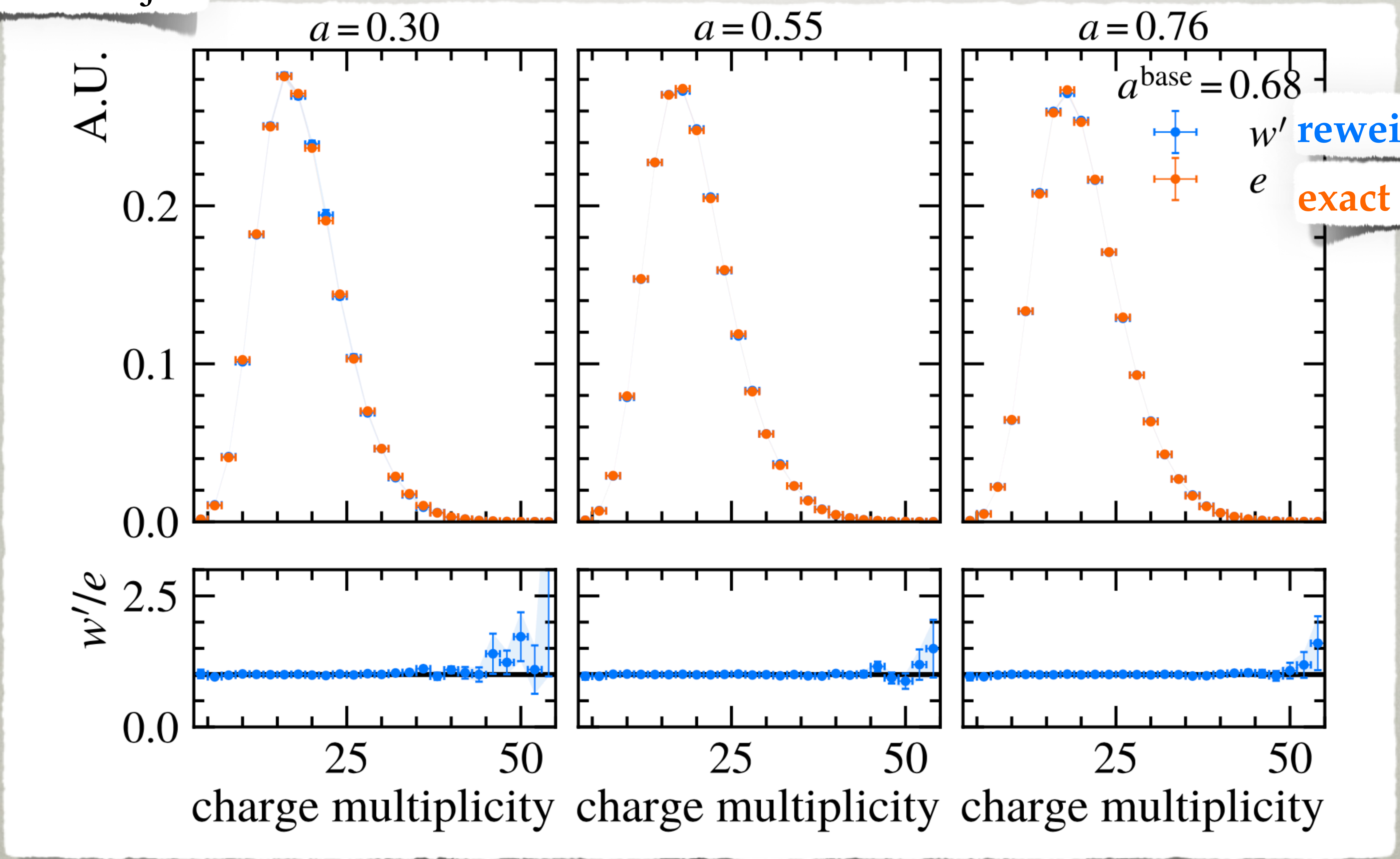
Sample par=i	w=1 w_j w_k	w=1 w_j w_k	w=1 w_j w_k	w=1 w_j w_k	w=1 w_j w_k	w=1 w_j w_k	…

graphics by M. K. Wilkinson

REWEIGHTING HADRONIZED PYTHIA EVENTS

$e^+e^- \to Z \to$ jets

Ilten et al, 2307.nnnnn

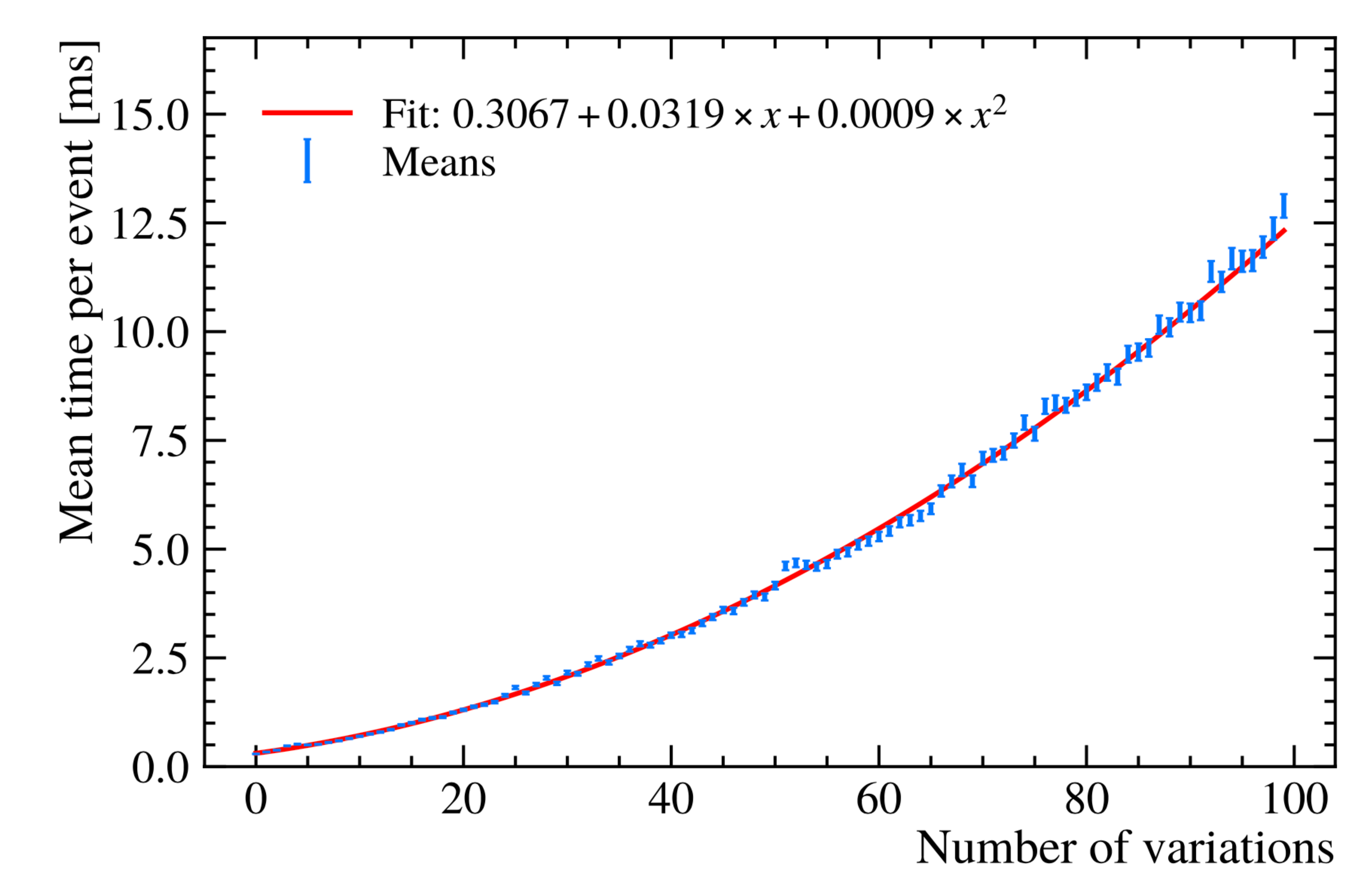


reweighted

exact calc.

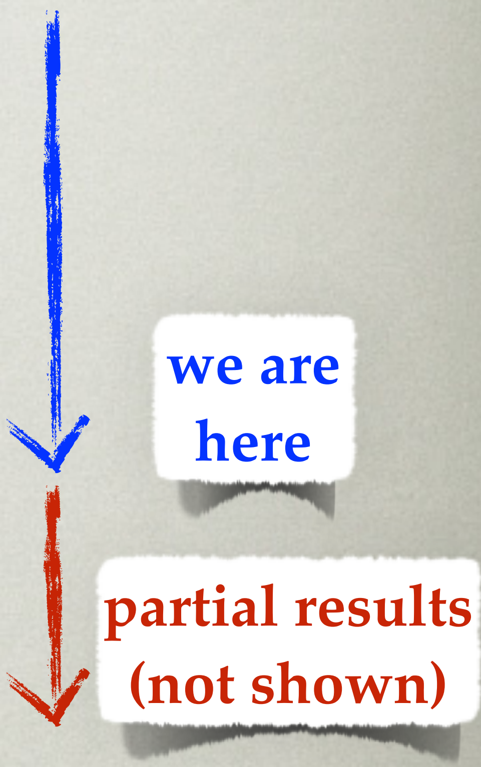
REWEIGHTING HADRONIZED PYTHIA EVENTS

Ilten et al, 2307.nnnnn



NEXT STEPS

- a series of progressive steps to be done before practically useful in Pythia simulations
    - ML architecture that mimicks a simplified Lund string hadronization model
        - train ML on truth level Pythia output (not obs. in exp)
    - develop a framework to propagate errors
    - improved ML architecture with full hadron flavor selector
    - train on mock data (i.e. just observable information)
    - train on real data (i.e. just already measured information)
    - replace/supplement Pythia string model



we are here

partial results (not shown)

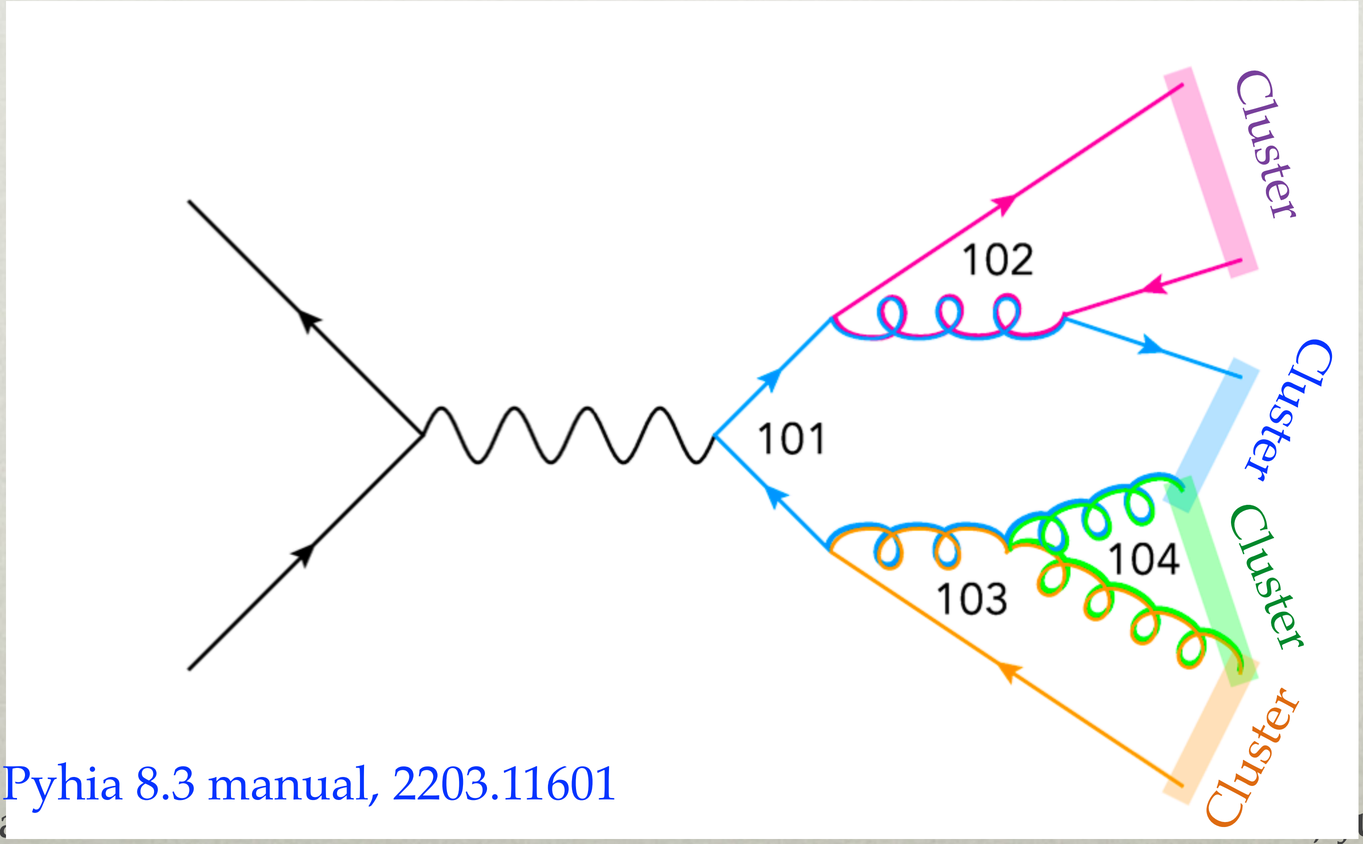
CONCLUSIONS

- MLhad: first steps in creating ML based description of hadronization
    - cSWAE reproduces simplified first hadron emission model
    - efficient parameter variation of Pythia hadronized events through reweighting
- long term: achieve a full fledge ML based description of hadronization

BACKUP SLIDES

CLUSTER MODEL

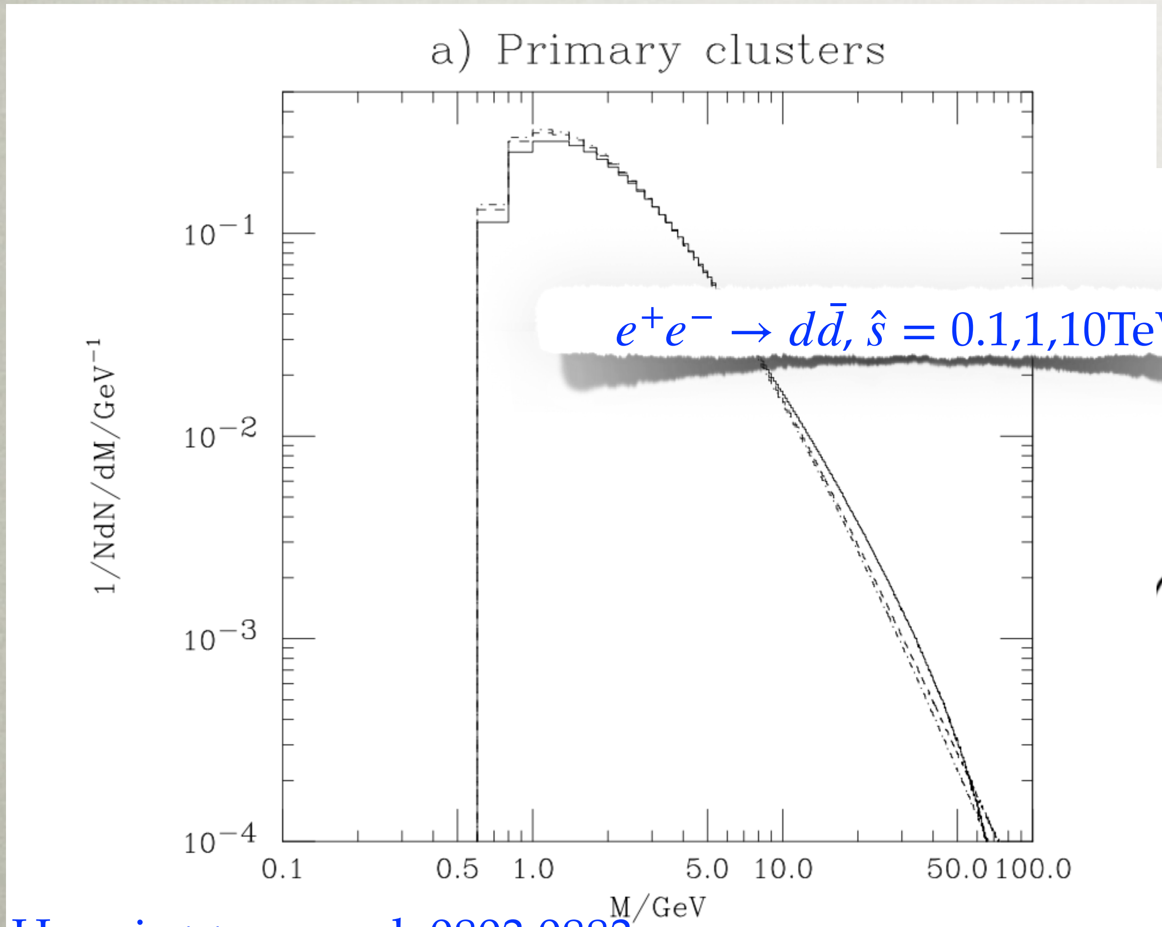
- assign mass to gluons, decay them to $q\bar{q}$ pairs
  - these are color singlets: *primary clusters*
  - primary clusters have universal mass distrib
- heavier clusters are decayed to lighter ones (model dep. step)
- relatively small set of params, $\mathcal{O}(30)$



Pyhia 8.3 manual, 2203.11601

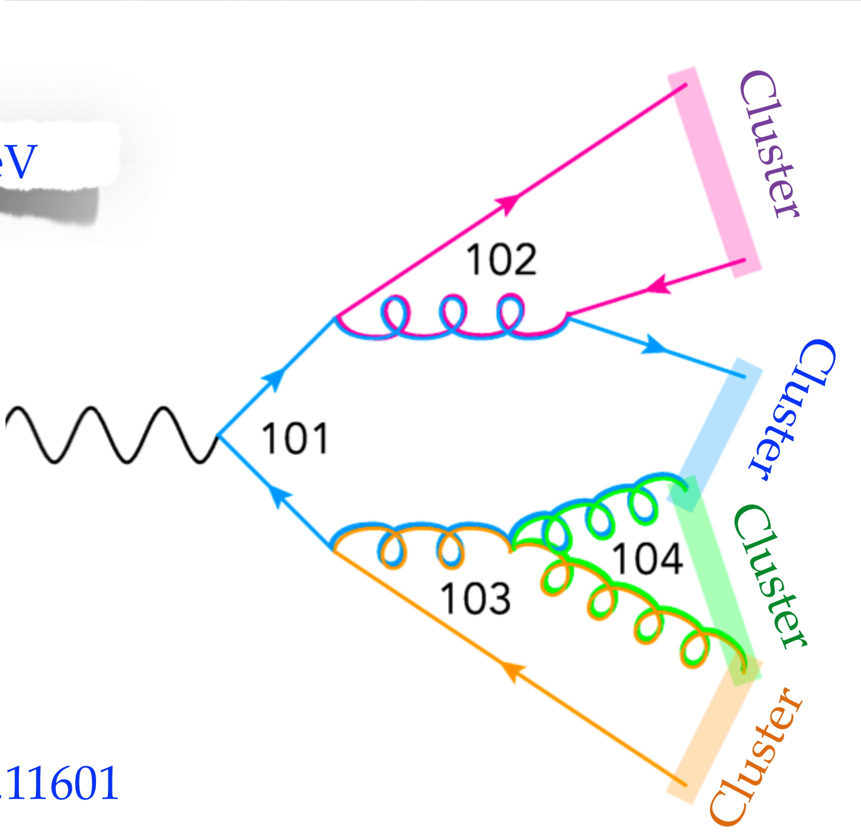
CLUSTER MODEL

- assign mass to gluons, decay them to $q\bar{q}$ pairs
  - these are color singlets: *primary clusters*
  - primary clusters have universal mass distrib
- heavier clusters are decayed to lighter ones (model dep. step)
- relatively small set of params, $\mathcal{O}(30)$



$e^+e^- \to d\bar{d}, \hat{s} = 0.1,1,10$TeV

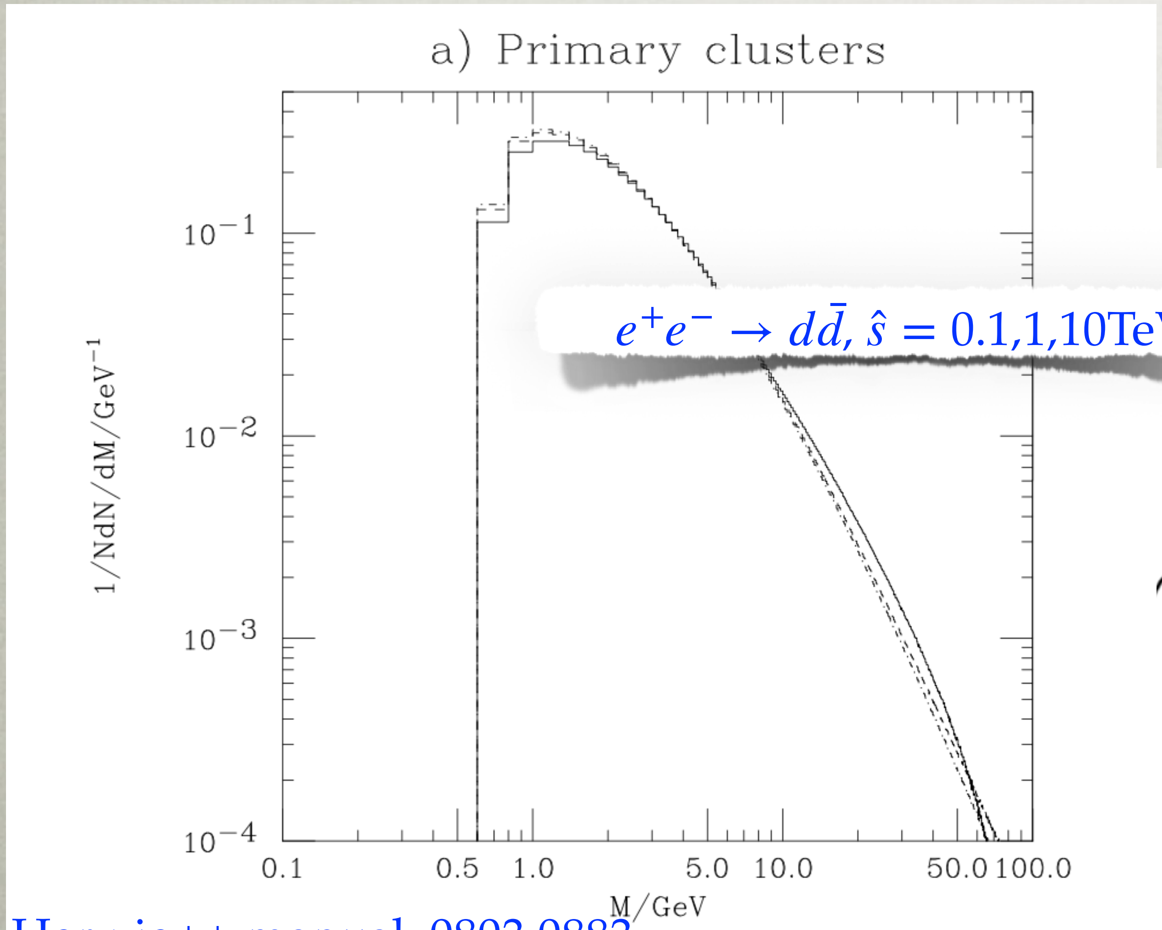
Herwig++ manual, 0803.0883



Pythia 8.3 manual, 2203.11601

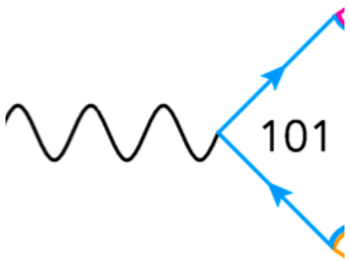
CLUSTER MODEL

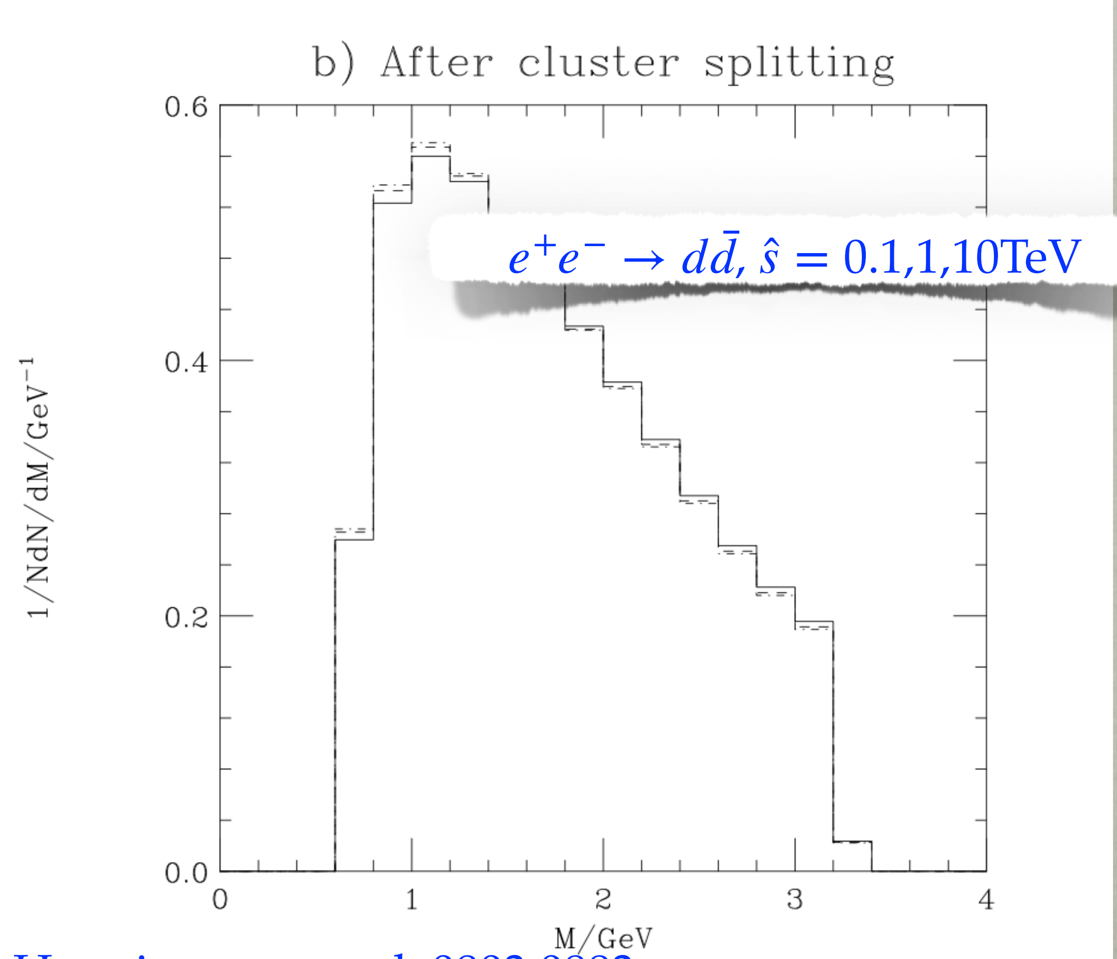
- assign mass to gluons, decay them to $q\bar{q}$ pairs
  - these are color singlets: *primary clusters*
  - primary clusters have universal mass distrib
- heavier clusters are decayed to lighter ones (model dep. step)
- relatively small set of params, $\mathcal{O}(30)$



$e^+e^- \to d\bar{d}$, $\hat{s} = 0.1,1,10$TeV

Herwig++ manual, 0803.0883



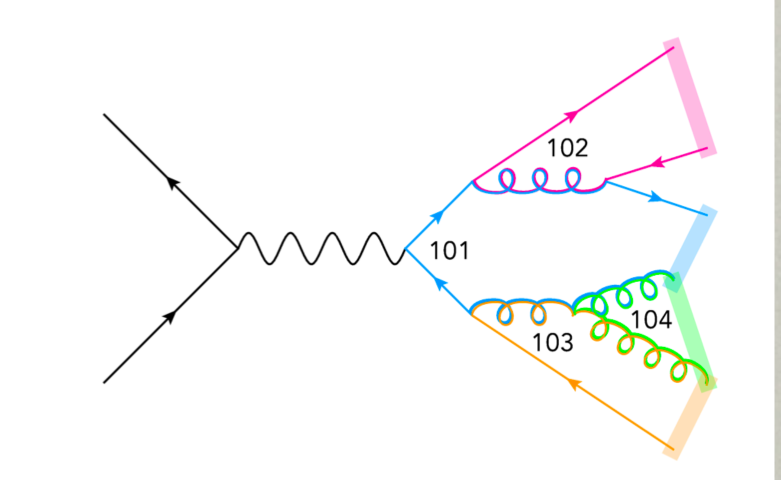


$e^+e^- \to d\bar{d}$, $\hat{s} = 0.1,1,10$TeV

Herwig++ manual, 0803.0883

ythia 8.3 manual, 2203.11601

COLOR RECONNECTION



- all perturbative predictions in leading color approximation ($N_c \rightarrow \infty$ with $\alpha_s N_c$ fixed)
  - direct mapping of color flow to strings
- color reconnection: inclusion of $1/N_c$ suppressed terms (model dep.)
  - reassing colors, not change in parton momenta

    Pyhia 8.3 manual, 2203.11601
  - several examples where important

    Fritzch, 1977; Ali et al, 1979
    - first historic mention: for charmonium production in B decays
    - for multiple parton interactions (Pythia MPI model) Sjöstrand, Zijl, 1987
    - $e^+e^- \rightarrow W^+W^- \rightarrow 4j$ at LEP 2 excludes no CR hypothesis 1302.3415
    - top quark mass determination from hadronic tops
  - several color reconnection models in Pythia

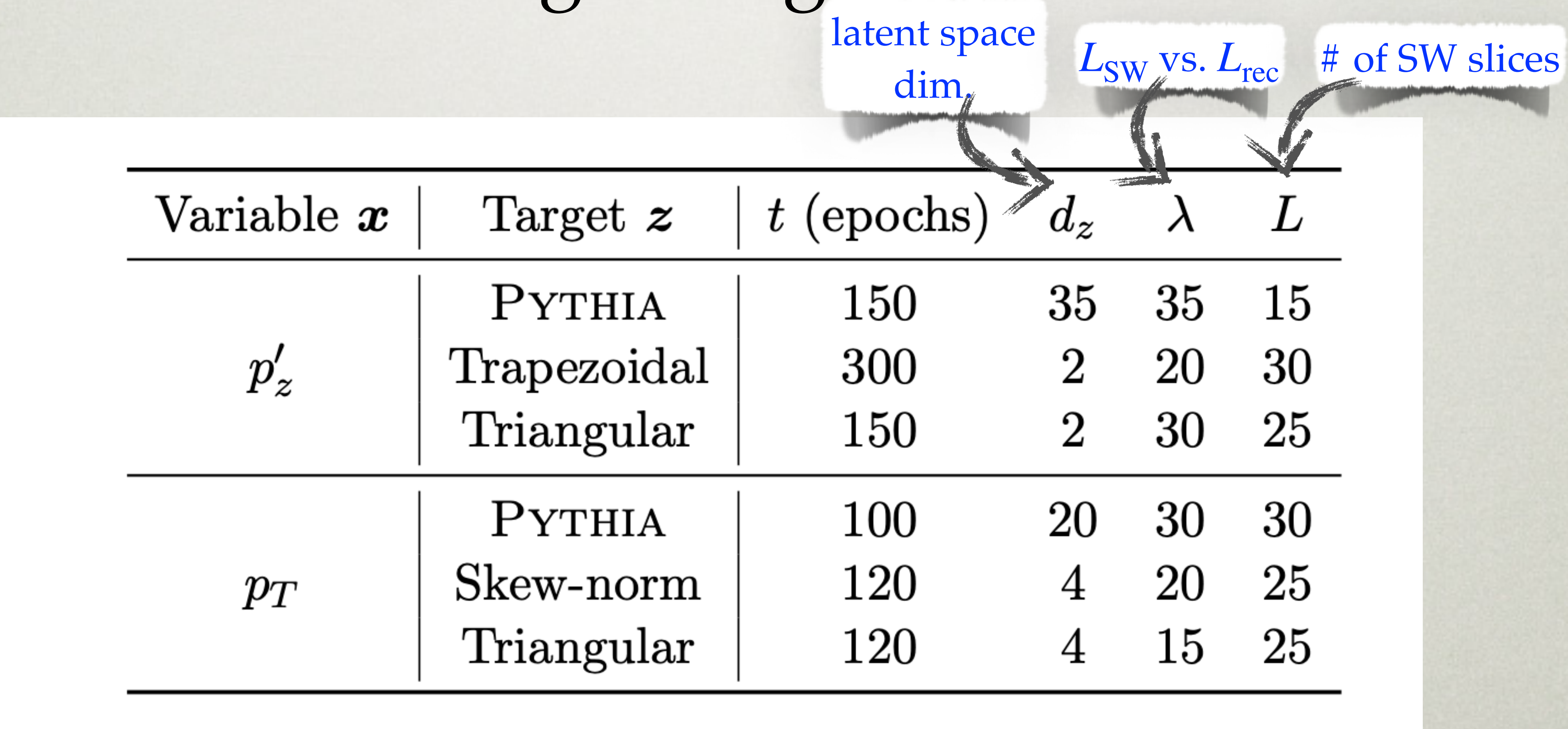
CHALLENGES FOR HADRONIZATION MODELS

Fischer, Sjostrand, 1610.09818

- in general out of the box hadronizations models work within *20-50%*
- some challenges for Pythia
  - change of flavor composition with event multiplicity
    - high multiplicity events have higher strangenesss content
    - no mechanism in Pythia to mimic it
  - average $\langle p_T \rangle$ larger for heavier particles, trend ok in Pythia, but numerically not large enough
  - charge particle p_T spectrum not correctly modelled at low p_T
    - partially can be fixed by tunes, but then a problem at interm. p_T
  - there is a peak in Λ/K p_T spectrum at $p_T \sim 2.5$ GeV, not reproduced by Pythia
  - the observation of the ridge in *pp* requires collective effects
- at least some of them addressed in Pythia 8.3 by introducing more involved models of string interactions, thermodynamical string fragmentation model, etc.
- Herwig has a different set of challenges, e.g., predicting heavy baryon distributions

RESULTS - FIRST EMISSIONS

- three different latent space distributions used
- cSWAE training configurations



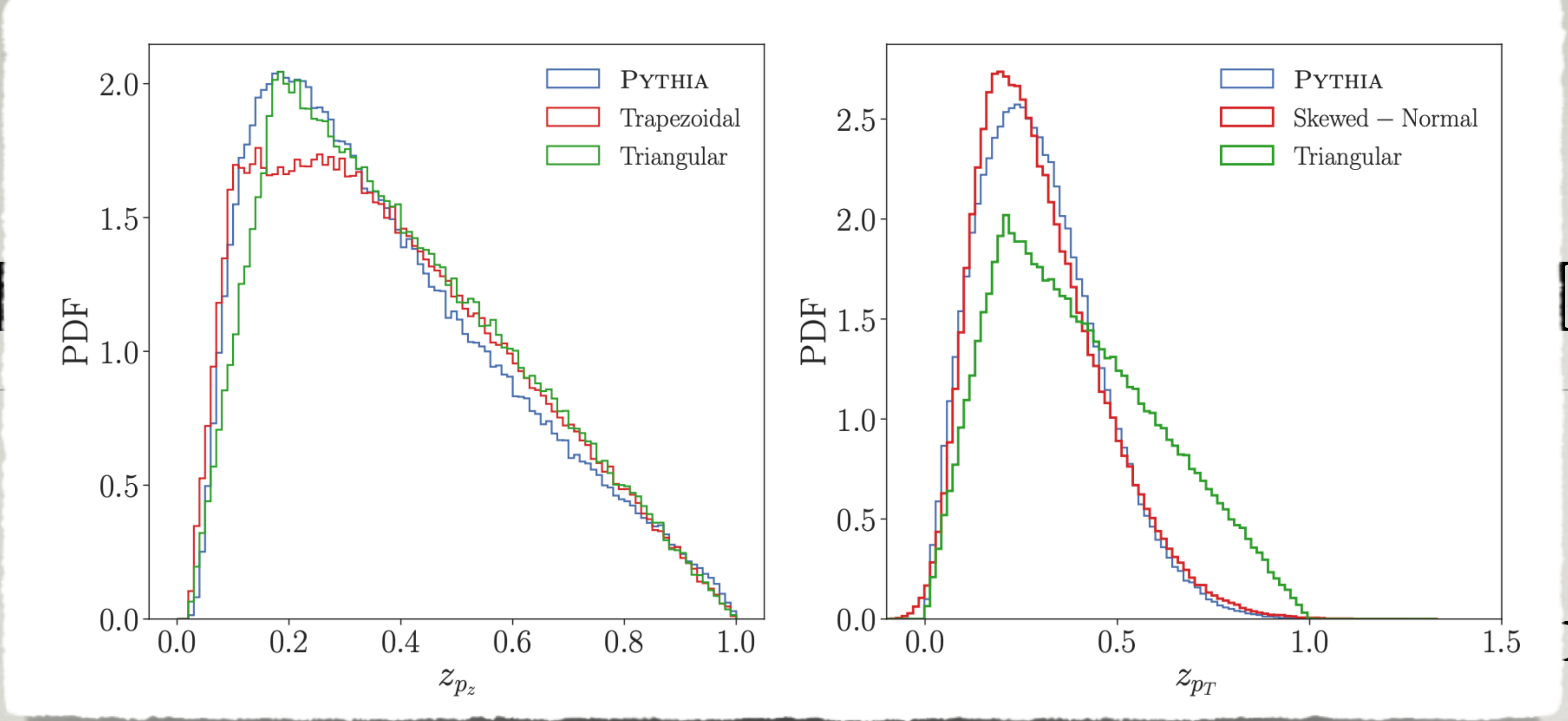
latent space dim.

L_{SW} vs. L_{rec}

# of SW slices

Variable $\boldsymbol{x}$	Target $\boldsymbol{z}$	t (epochs)	d_z	λ	L
p'_z	Pythia	150	35	35	15
	Trapezoidal	300	2	20	30
	Triangular	150	2	30	25
p_T	Pythia	100	20	30	30
	Skew-norm	120	4	20	25
	Triangular	120	4	15	25

RI IONS

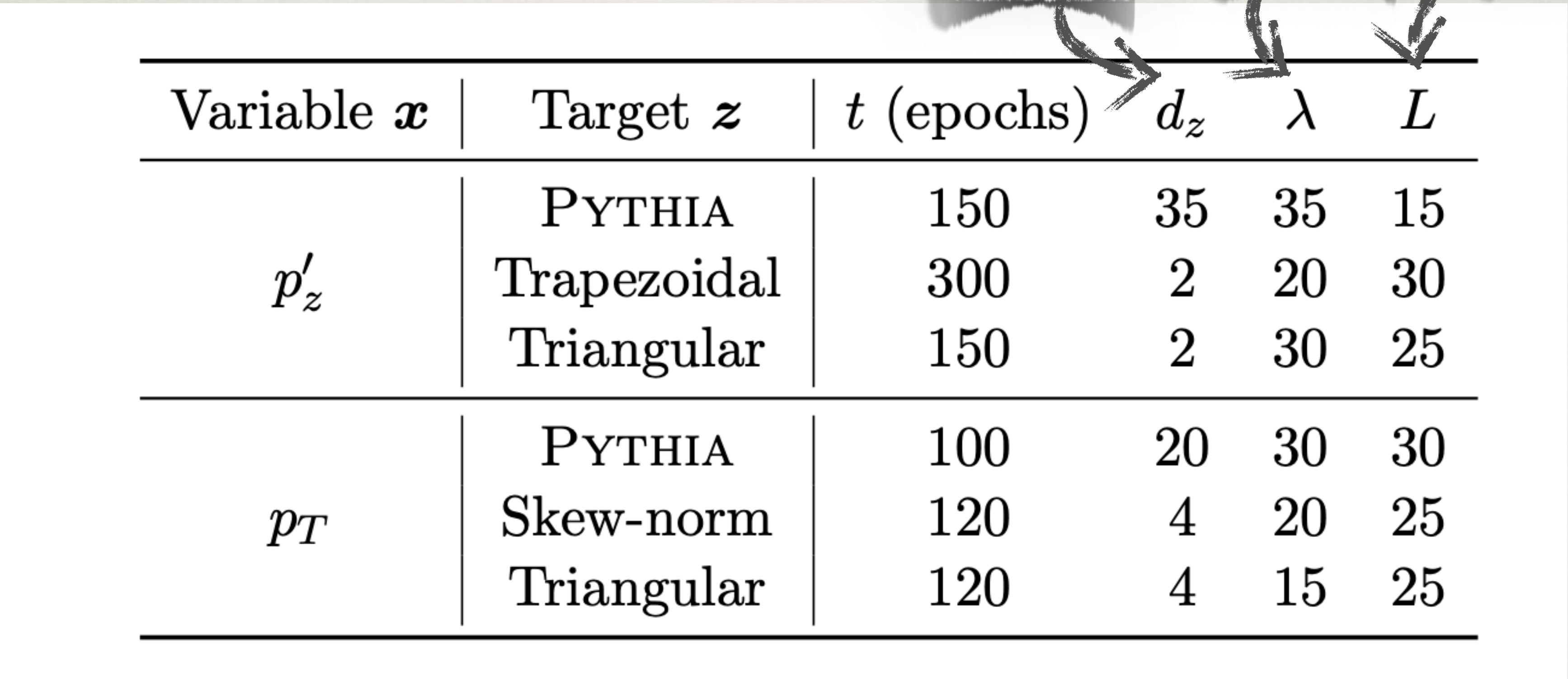


- ns used
- cSWAE training configurations

latent space dim.

L_{SW} vs. L_{rec}

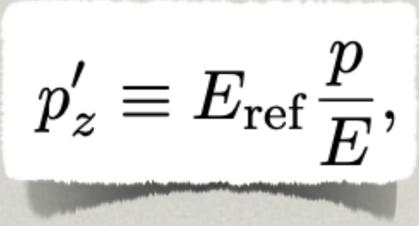
# of SW slices



Variable $\boldsymbol{x}$	Target $\boldsymbol{z}$	t (epochs)	d_z	λ	L
p'_z	Pythia	150	35	35	15
	Trapezoidal	300	2	20	30
	Triangular	150	2	30	25
p_T	Pythia	100	20	30	30
	Skew-norm	120	4	20	25
	Triangular	120	4	15	25

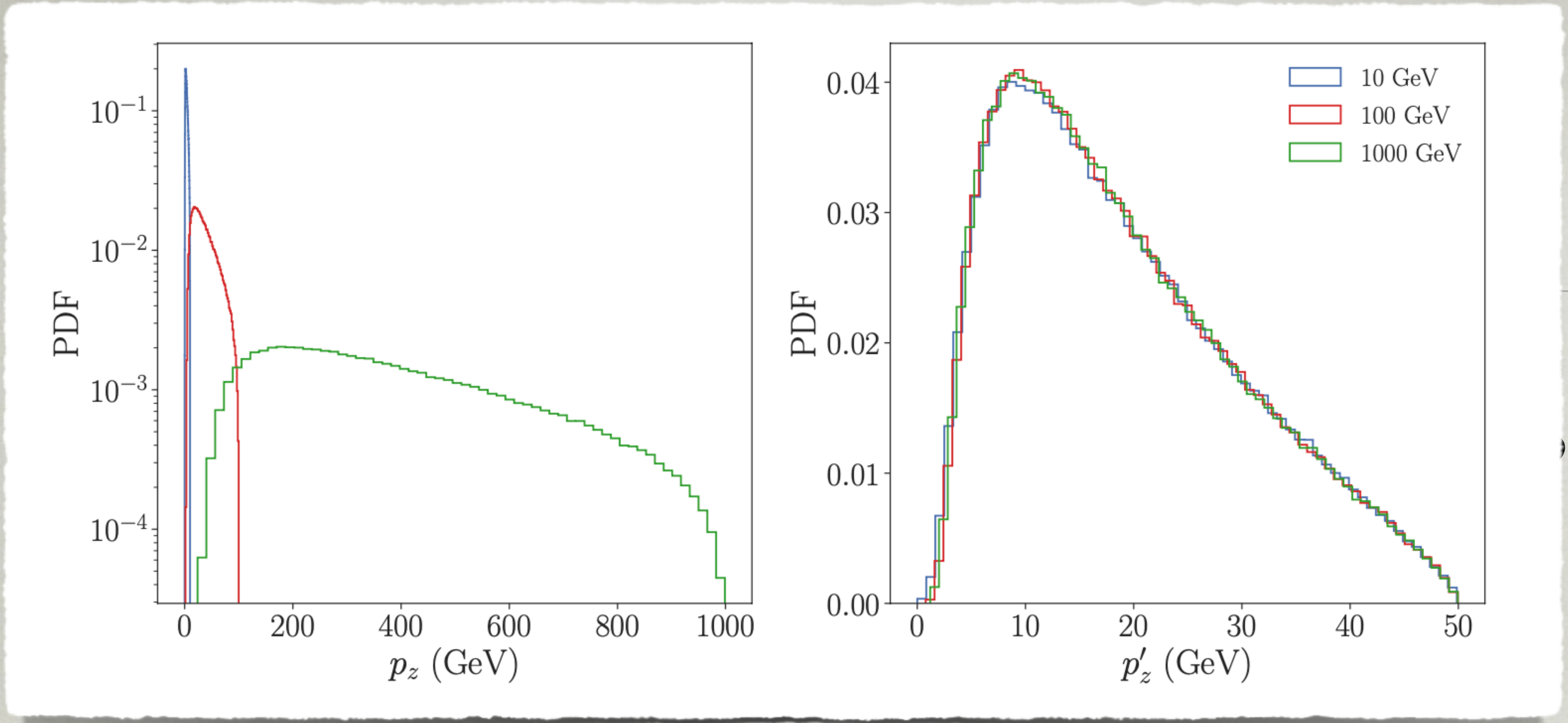
MLhad

- right now trained directly on Pythia first emission output
  - hadron mom. described by p_z, p_T
- the IR cut-off has two effects
  - p_z and p_T distributions are uncorellated
  - makes the problem scale invariant in p_Z
    - enough to train at one string mass, $2E_{\text{ref}}$
    - for other energies can rescale



$$p'_z \equiv E_{\text{ref}} \frac{p}{E},$$

- this is relaxed in the end, E dependence can be recovered



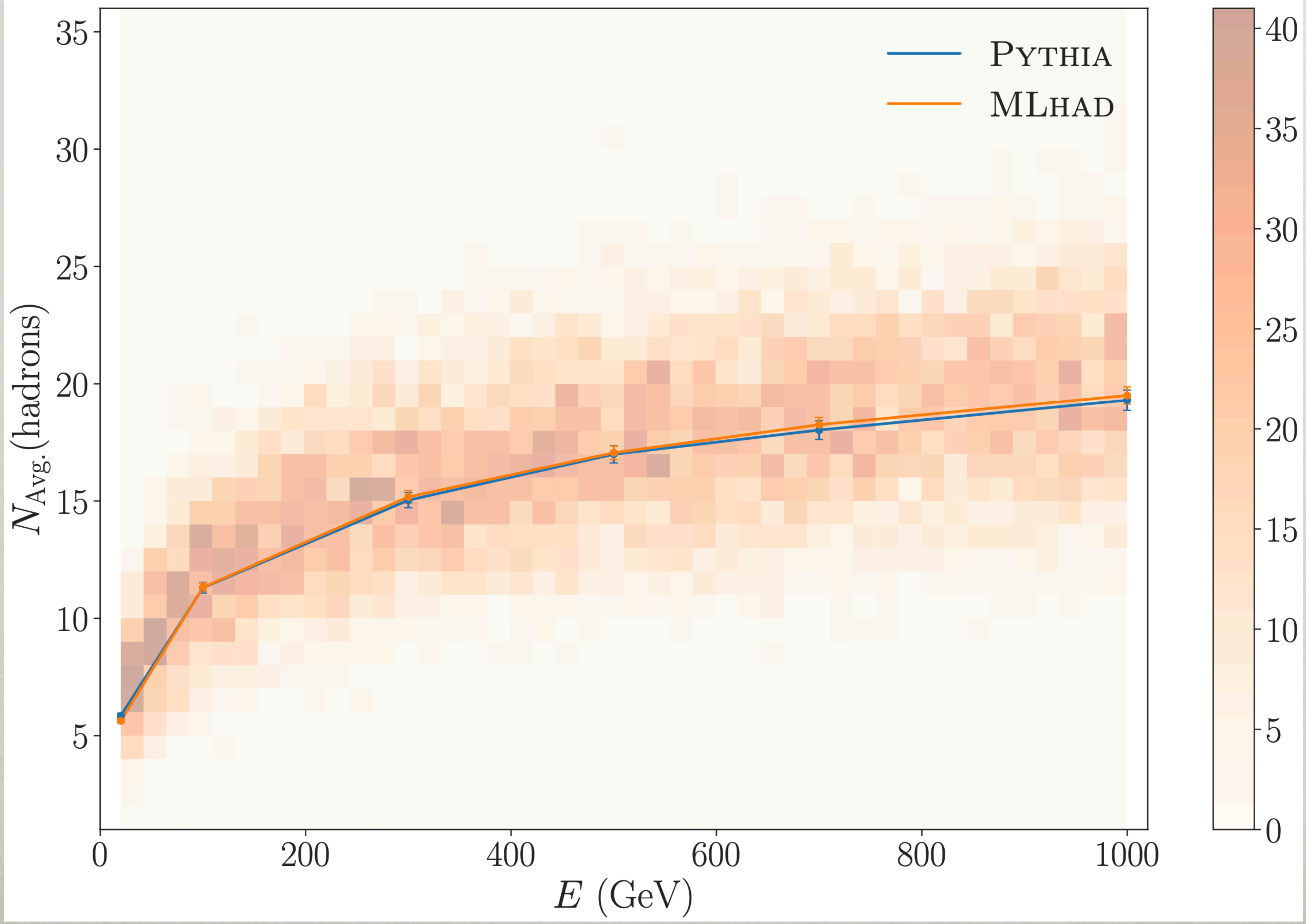
ut

- p_z and p_T distributions are uncorellated
- makes the problem scale invariant in p_Z
  - enough to train at one string mass, $2E_{\text{ref}}$
  - for other energies can rescale

$$p'_z \equiv E_{\text{ref}} \frac{p}{E},$$

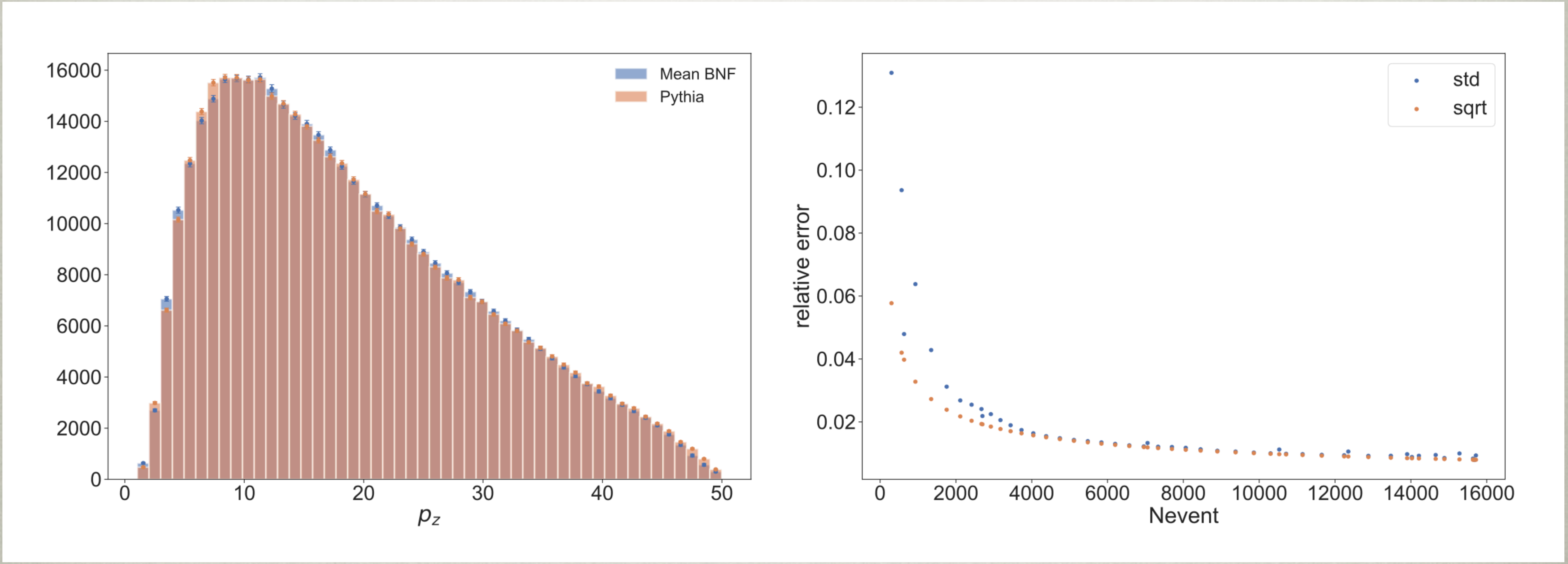
- this is relaxed in the end, E dependence can be recovered

MLhad with normalizing flows



MLhad with normalizing flows

- Bayesian NF captures well the uncertainties


Mean BNF
Pythia
p_z
std
sqrt
relative error
Nevent