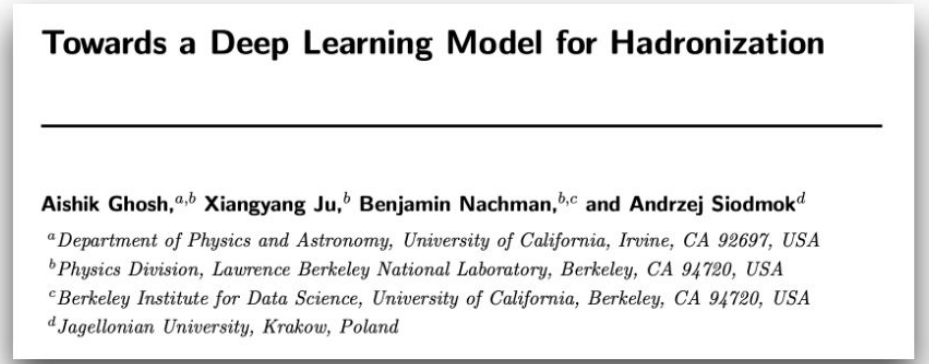
Construction and Fitting of a Deep Generative Hadronization Model

Andrzej Siódmok

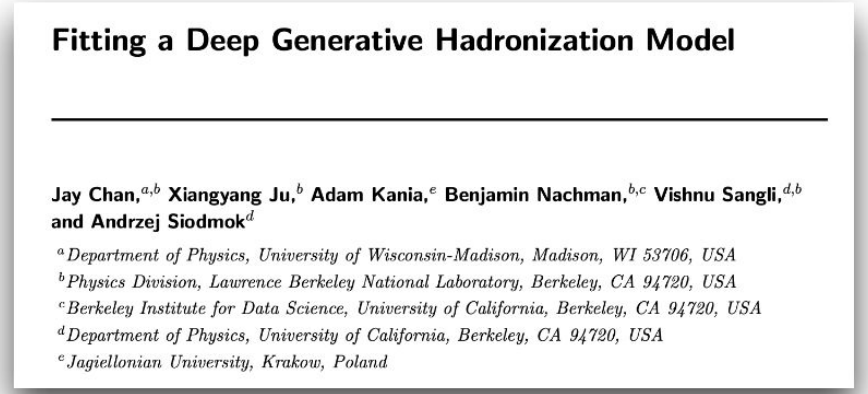


Towards a Deep Learning Model for Hadronization

Aishik Ghosh,[a,b] Xiangyang Ju,[b] Benjamin Nachman,[b,c] and Andrzej Siodmok[d]

[a] *Department of Physics and Astronomy, University of California, Irvine, CA 92697, USA*
[b] *Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[c] *Berkeley Institute for Data Science, University of California, Berkeley, CA 94720, USA*
[d] *Jagellonian University, Krakow, Poland*

2203.12660



Fitting a Deep Generative Hadronization Model

Jay Chan,[a,b] Xiangyang Ju,[b] Adam Kania,[e] Benjamin Nachman,[b,c] Vishnu Sangli,[d,b] and Andrzej Siodmok[d]

[a] *Department of Physics, University of Wisconsin-Madison, Madison, WI 53706, USA*
[b] *Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[c] *Berkeley Institute for Data Science, University of California, Berkeley, CA 94720, USA*
[d] *Department of Physics, University of California, Berkeley, CA 94720, USA*
[e] *Jagiellonian University, Krakow, Poland*

2305.17169

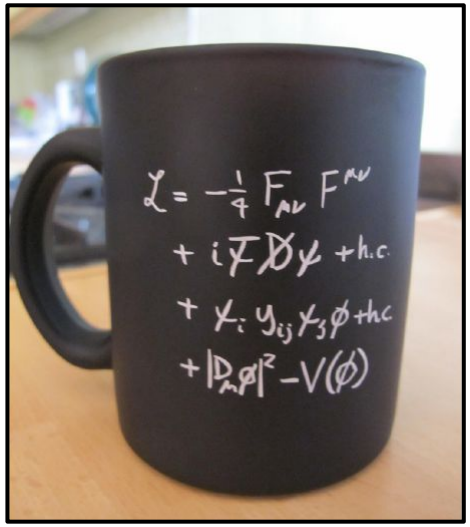
JAGIELLONIAN UNIVERSITY IN KRAKÓW
UCZELNIA BADAWCZA INICJATYWA DOSKONAŁOŚCI
NARODOWE CENTRUM NAUKI
NCN: 2019/34/E/ST2/00457
MCnet

Motivation - Monte Carlo Event Generators (MCEG)

There is a **huge gap** between a one-line formula of a fundamental theory, like the Lagrangian of the SM, and the experimental reality that it implies

Theory

Standard Model Lagrangian



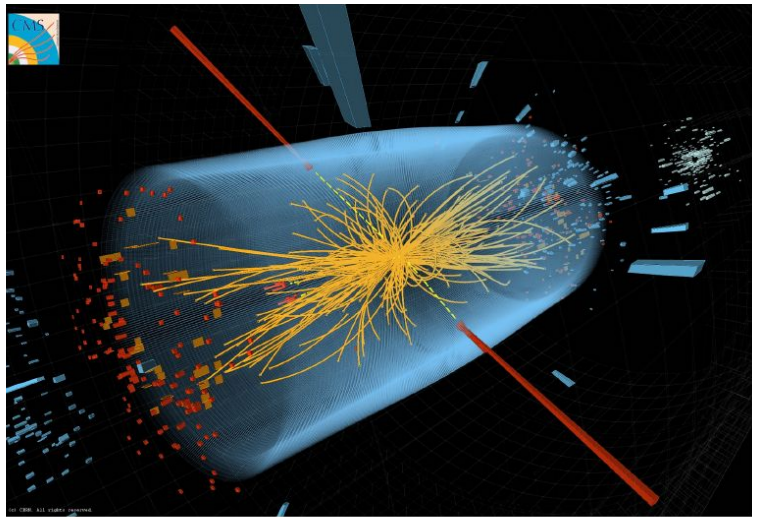
Data makes you smarter

It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong.

Richard P. Feynman

Experiment

LHC event

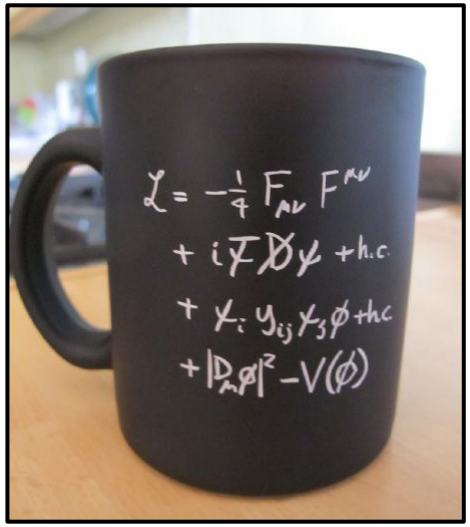


Motivation - Monte Carlo Event Generators (MCEG)

There is a **huge gap** between a one-line formula of a fundamental theory, like the Lagrangian of the SM, and the experimental reality that it implies

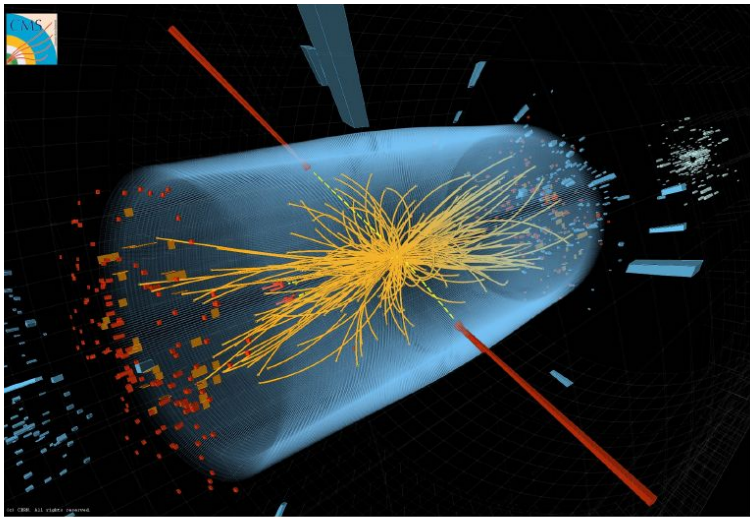
Theory

Standard Model Lagrangian



Experiment

LHC event



- MC event generators are designed to bridge that **gap**
- “Virtual collider” ⇒ Direct comparison with data

Almost all **HEP measurements and discoveries** in the modern era have **relied on MCEG**, most notably the discovery of the Higgs boson.

Published papers by ATLAS, CMS, LHCb: **2252**
Citing at least 1 of 3 existing MCEG: **1888** (**84%**)

Motivation - Monte Carlo Event Generators (MCEG)

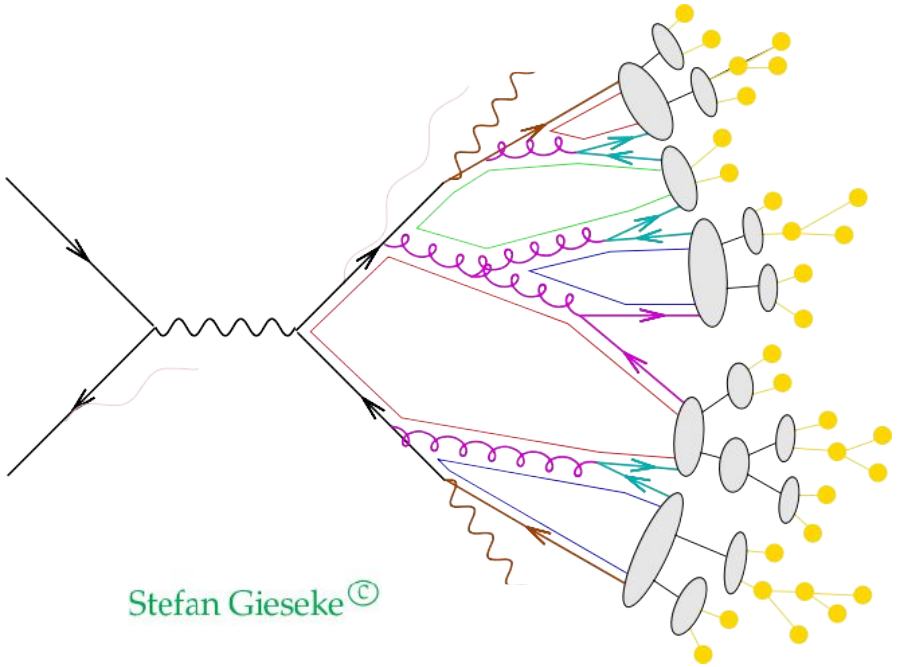
QCD correctly describes strong interactions in each energy range but its complex mathematical structure makes it very difficult to obtain precise predictions (Millennium Prize Problem $1,000,000)

High energy

- **perturbative QCD**
- **in theory we know what to do**
- **in practice very difficult**

Low energy

- **non-perturbative QCD**
- **we don't know what to do**
- **phenomenological models (with many free parameters)**

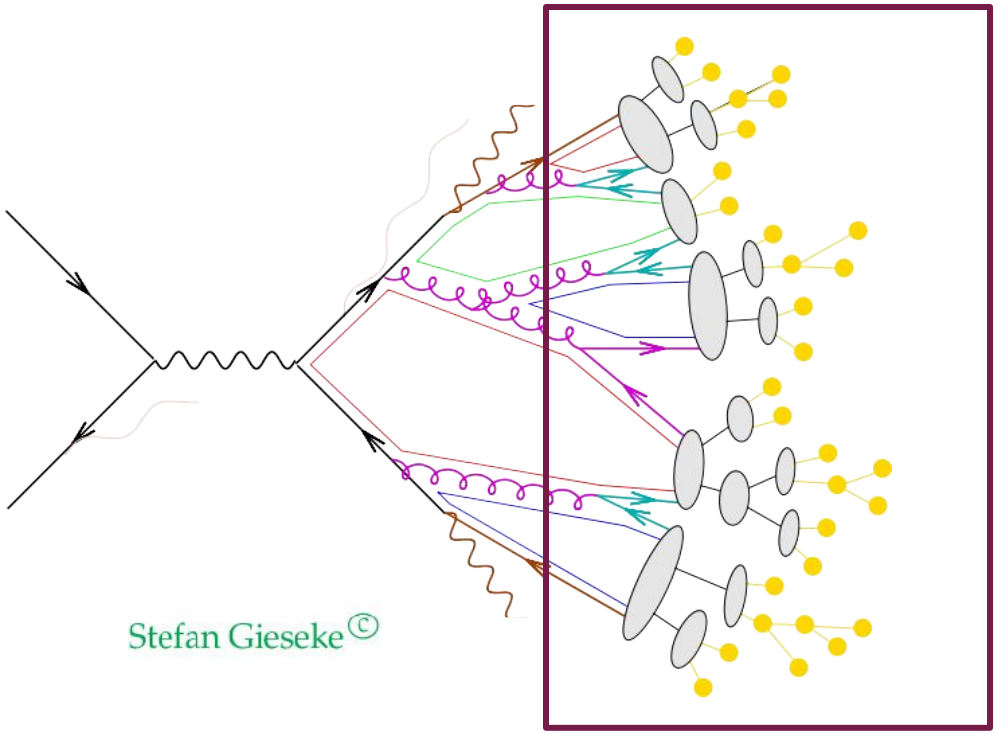


Why hadronization?

QCD correctly describes strong interactions in each energy range but its complex mathematical structure makes it very difficult to obtain precise predictions (Millennium Prize Problem $1,000,000)

High energy
- **perturbative QCD**
- **in theory we know what to do**
- **in practice very difficult**

Low energy
- **non-perturbative QCD**
- **we don't know what to do**
- **phenomenological models (with many free parameters)**



Hadronization:
one of the least understood elements of MCEG

Motivation - Hadronization

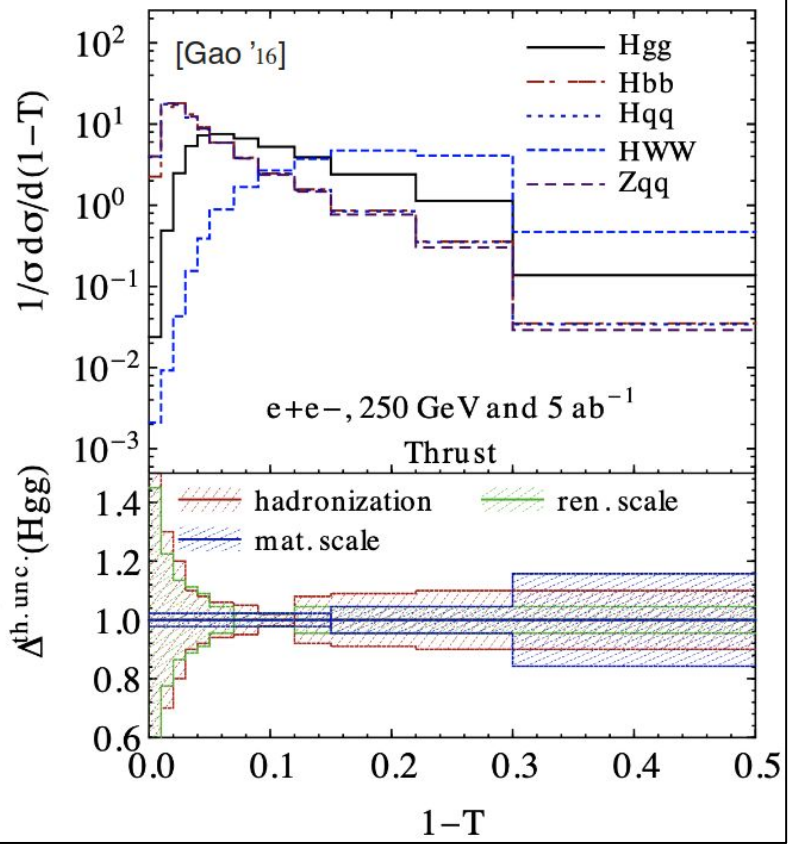
Hadronization:

➔ Increased control of perturbative corrections ⇒ more often measurements are limited by non-perturbative components, such as hadronization.

- W mass measurement using a new method [Freytsis at al. JHEP 1902 (2019) 003]
- Extraction of the strong coupling in [M. Johnson, D. Maître, Phys.Rev. D97 (2018) no.5]
- Top mass [S. Argyropoulos, T. Sjöstrand, JHEP 1411 (2014) 043]
- ...

Pier Moni's talk
FCC Physics Workshop 2023

- However, hadronisation remains the main bottleneck
  - e.g. thrust in Higgs decays (MC variation in plot)
- Increase in energy insufficient for suppression ($Q \sim m_H$)
- Runs at lower energies are essential for a robust tuning of NP models in MCs
- Also crucial for training of ML algorithms for jet tagging, instrumental in extraction of Higgs couplings



Motivation - Hadronization

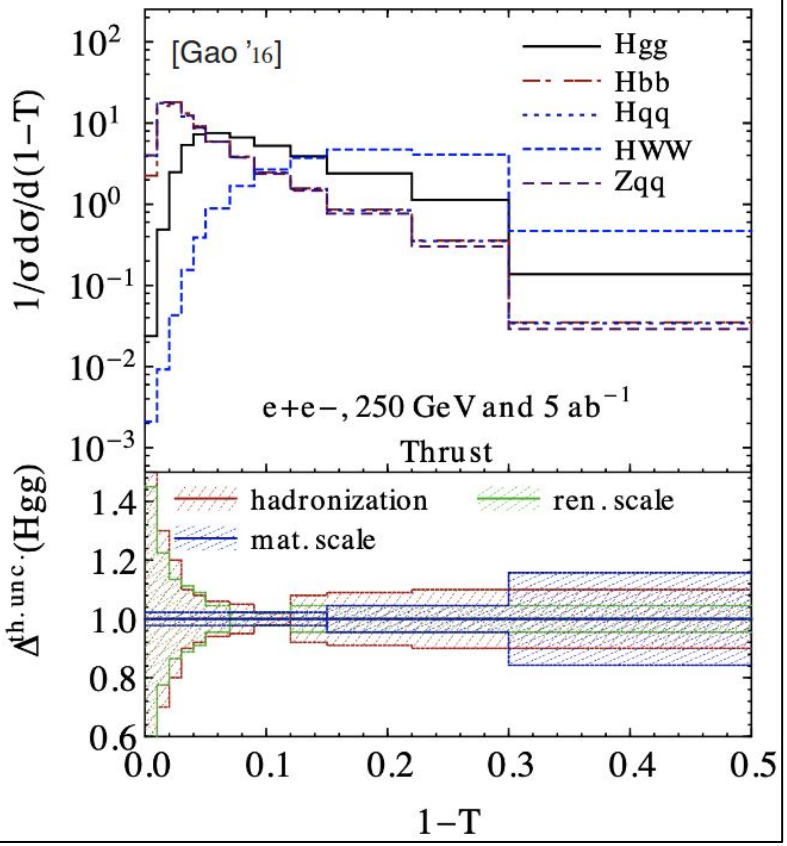
Hadronization:

➔ Increased control of perturbative corrections ⇒ more often LHC measurements are limited by non-perturbative components, such as hadronization.

- W mass measurement using a new method [Freytsis at al. JHEP 1902 (2019) 003]
- Extraction of the strong coupling in [M. Johnson, D. Maître, Phys.Rev. D97 (2018) no.5]
- Top mass [S. Argyropoulos, T. Sjöstrand, JHEP 1411 (2014) 043]
- ...

Pier Moni's talk
FCC Physics Workshop 2023

- However, hadronisation remains the main bottleneck
  - e.g. thrust in Higgs decays (MC variation in plot)
- Increase in energy insufficient for suppression ($Q \sim m_H$)

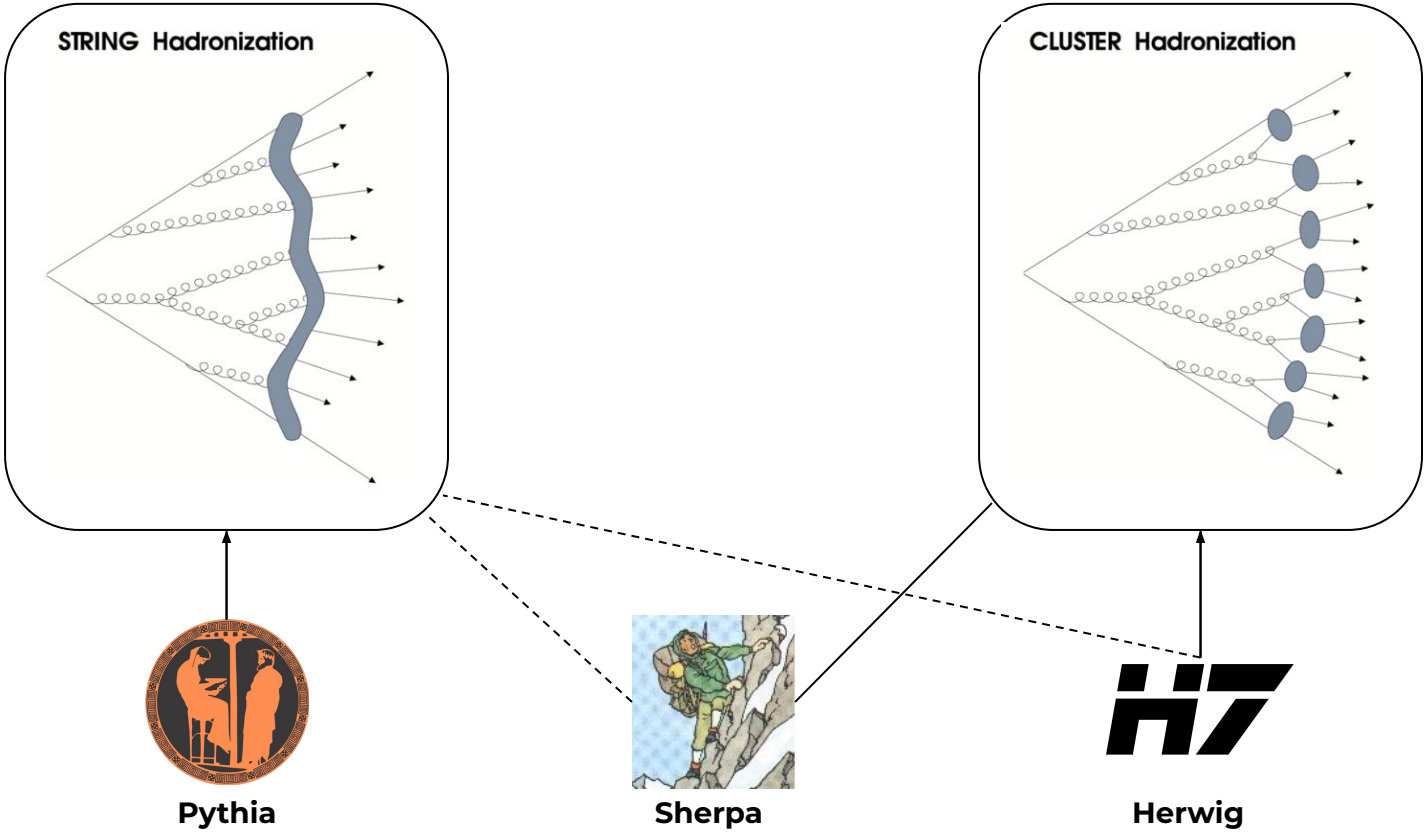


Motivation - Hadronization

Hadronization:

➔ Increased control of perturbative corrections ⇒ more often LHC measurements are limited by non-perturbative components, such as hadronization.

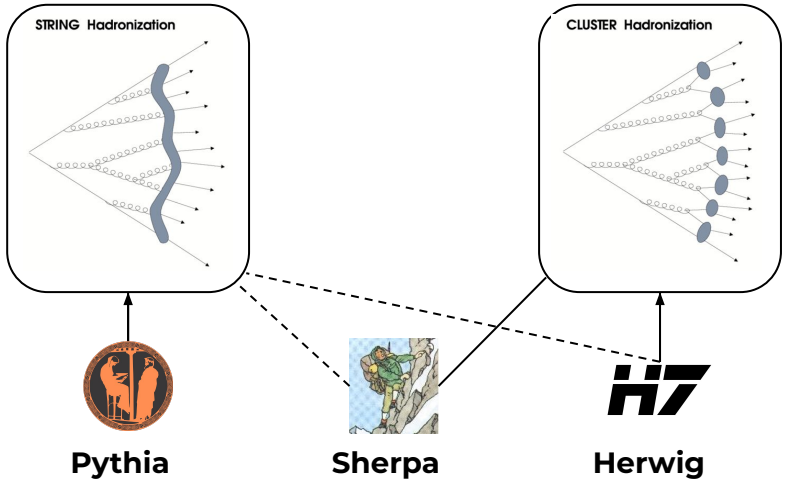
- W mass measurement using a new method [Freytsis at al. JHEP 1902 (2019) 003]
- Extraction of the strong coupling in [M. Johnson, D. Maître, Phys.Rev. D97 (2018) no.5]
- Top mass [S. Argyropoulos, T. Sjöstrand, JHEP 1411 (2014) 043]
- ...



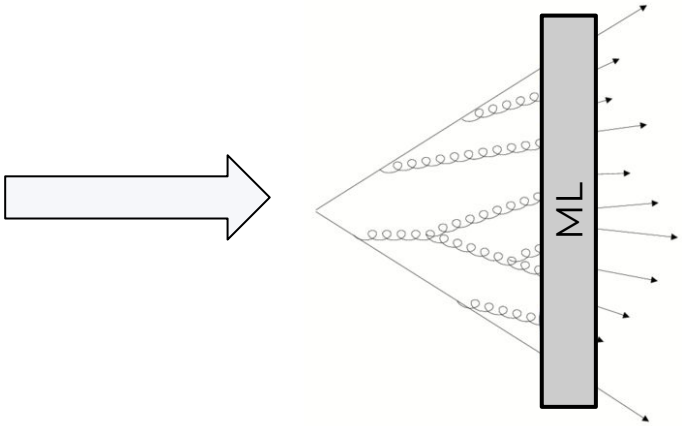
Hadronization models

Hadronization:

Early 1980's
(since then little development)



Early 2020's
(lot of progress in ML)



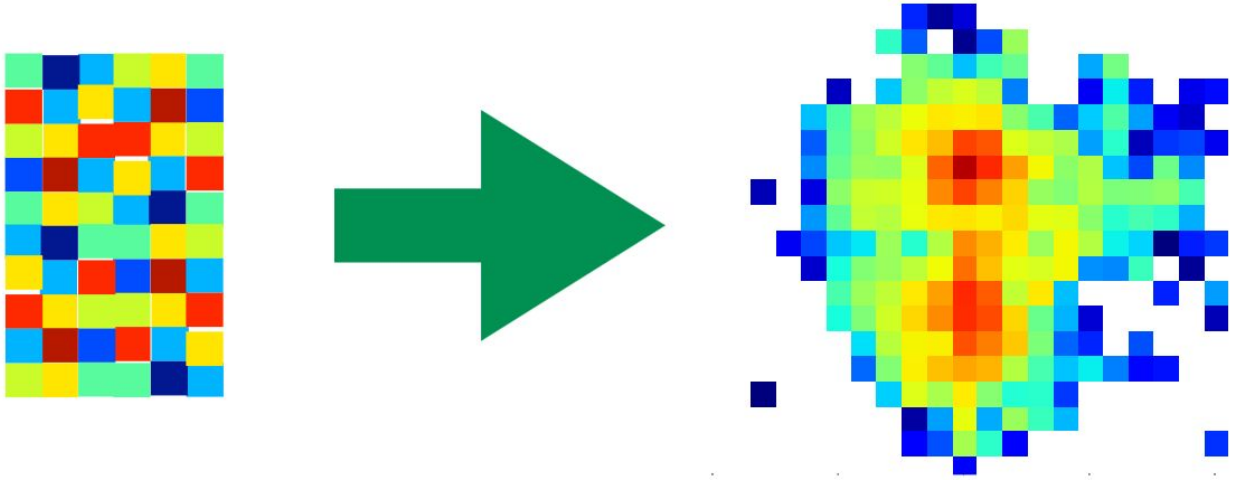
Idea of using Machine Learning (ML) for hadronization.

Hadronization is a fitting problem
Can ML be more flexible and extract more information from data?

See also PDFs and the pioneer NNPDF

What is a deep generative model?

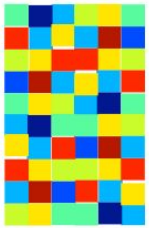
A **generator** is nothing other than a function that maps random numbers to structure.



Deep generative models: the map is a deep neural network.

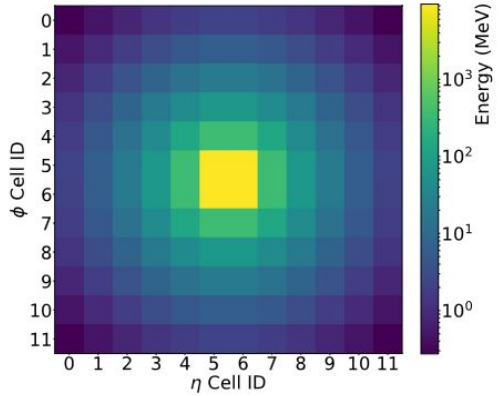
Our tool of choice: GANs

Generative Adversarial Networks (GANs):
*A two-network game where one **maps noise to structure** and one **classifies images as fake or real**.*



noise

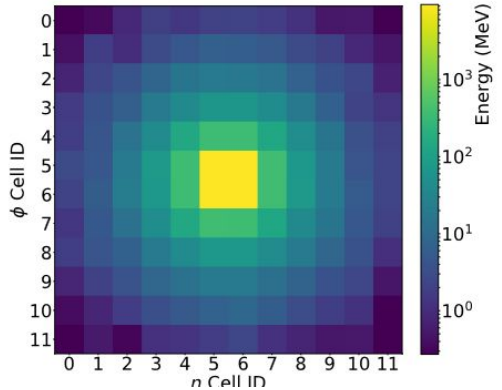
G



D

{real,fake}

When **D** is maximally confused, **G** will be a good generator



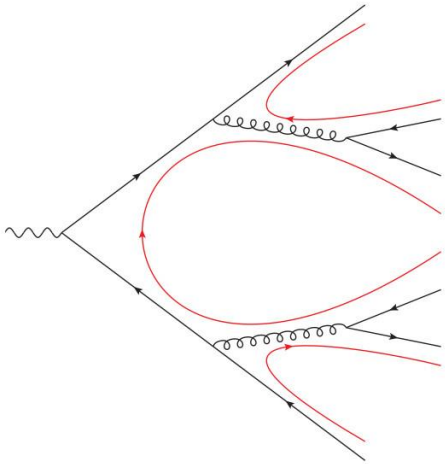
D

Physics-based simulator or data

Cluster hadronization model

The philosophy of the model: use information from perturbative QCD as an input for hadronization.

QCD **pre-confinement** discovered by Amati & Veneziano:

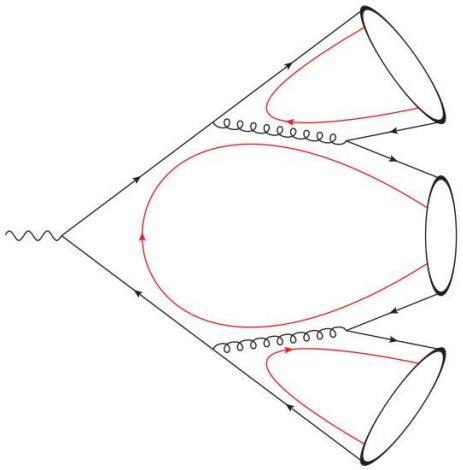


- QCD provide pre-confinement of colour

Cluster hadronization model

The philosophy of the model: use information from perturbative QCD as an input for hadronization.

QCD **pre-confinement** discovered by Amati & Veneziano:

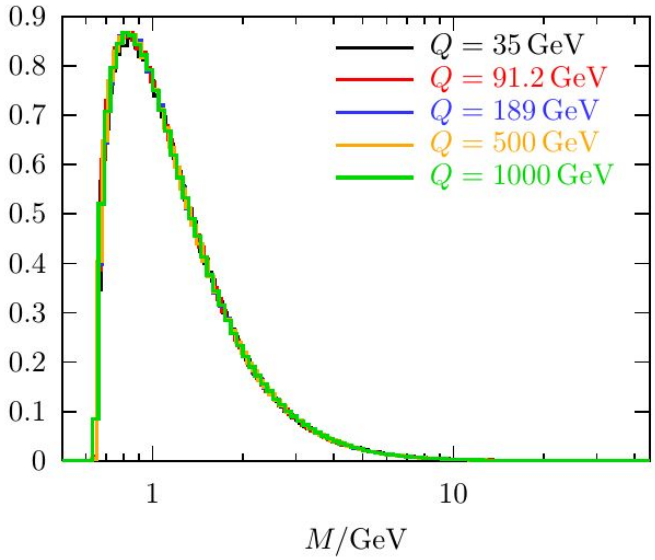


- QCD provide pre-confinement of colour
- Colour-singlet pair end up close in phase space and form highly excited hadronic states, the clusters

Cluster hadronization model

The philosophy of the model: use information from perturbative QCD as an input for hadronization.

QCD **pre-confinement** discovered by Amati & Veneziano:

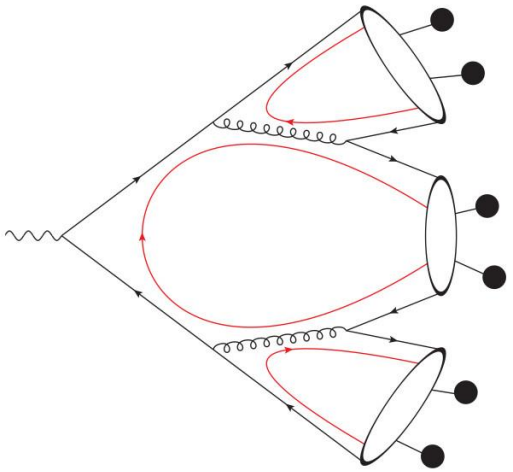


- QCD provide pre-confinement of colour
- Colour-singlet pair end up close in phase space and form highly excited hadronic states, the clusters
- Pre-confinement states that the spectra of clusters are independent of the hard process and energy of the collision

Cluster hadronization model

The philosophy of the model: use information from perturbative QCD as an input for hadronization.

QCD **pre-confinement** discovered by Amati & Veneziano:

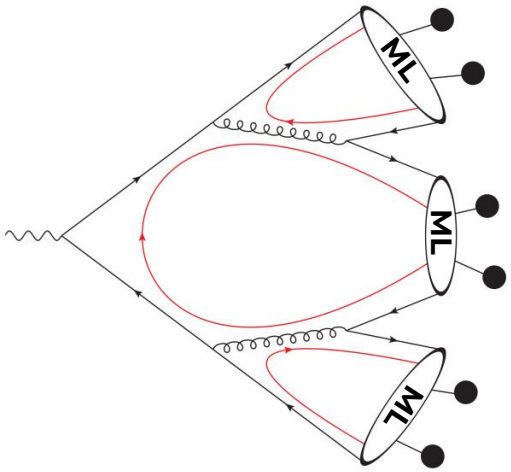


- QCD provide pre-confinement of colour
- Colour-singlet pair end up close in phase space and form highly excited hadronic states, the clusters
- Pre-confinement states that the spectra of clusters are independent of the hard process and energy of the collision
- Peaked at low mass (1-10 GeV) typically decay into 2 hadrons

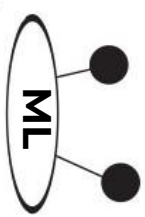
Cluster hadronization model

The philosophy of the model: use information from perturbative QCD as an input for hadronization.

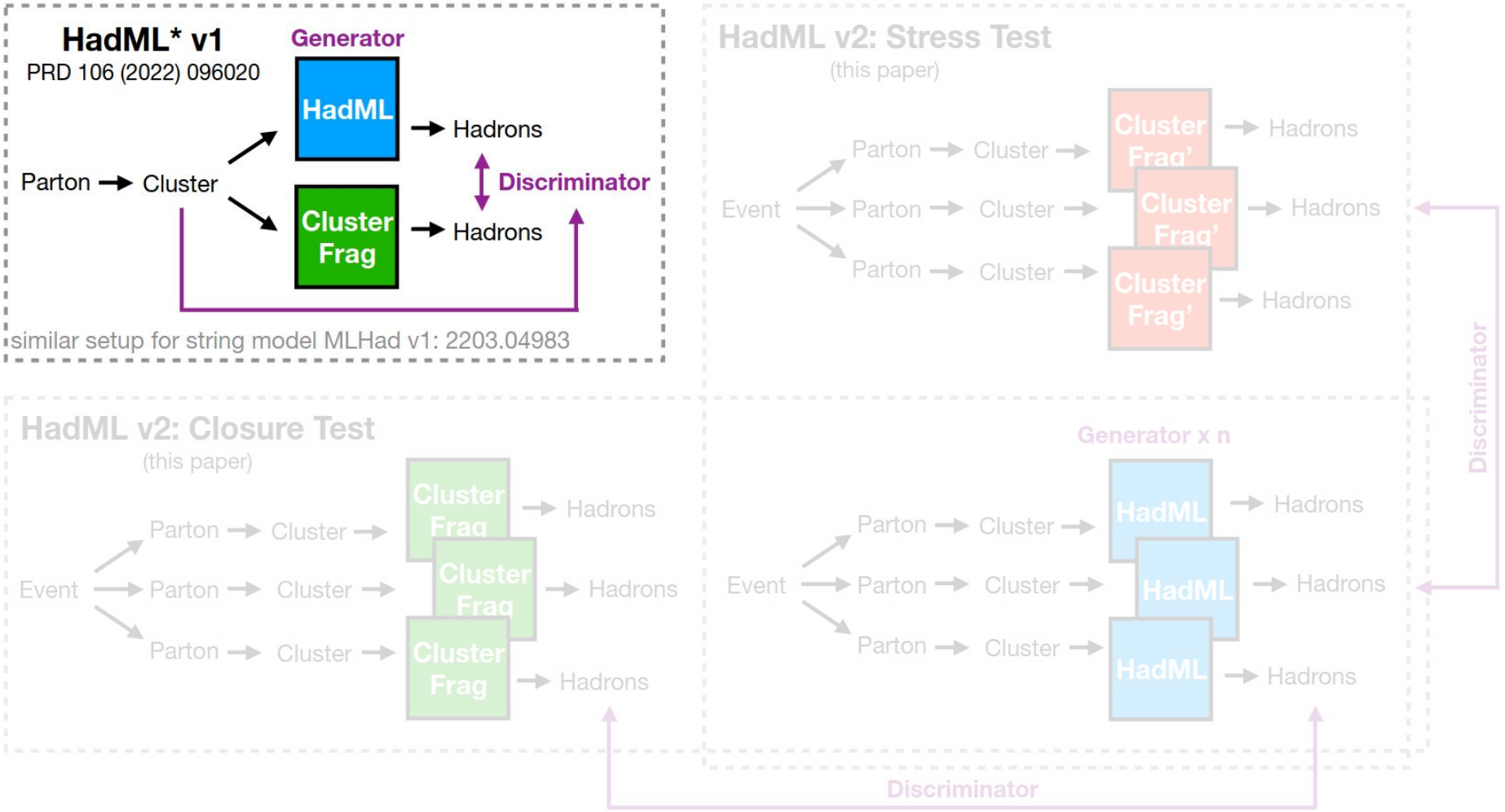
QCD **pre-confinement** discovered by Amati & Veneziano:



- QCD provide pre-confinement of colour
- Colour-singlet pair end up close in phase space and form highly excited hadronic states, the clusters
- Pre-confinement states that the spectra of clusters are independent of the hard process and energy of the collision
- Peaked at low mass (1-10 GeV) typically decay into 2 hadrons
- **ML hadronization**
  1st step: generate kinematics of a cluster decay:



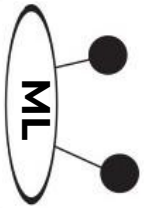
Road map for today



Towards a Deep Learning Model for Hadronization

ML hadronization

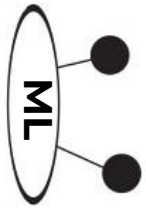
1st step: generate kinematics of a cluster decay to 2 hadrons



Towards a Deep Learning Model for Hadronization

ML hadronization

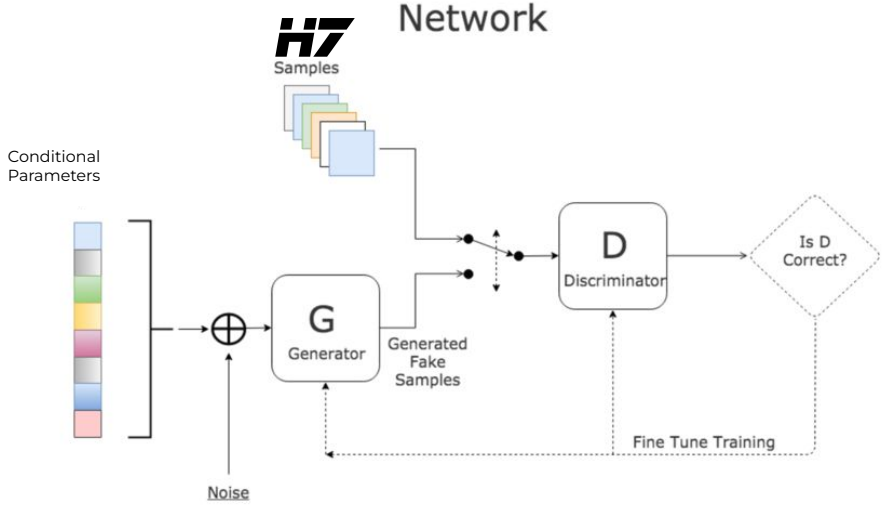
1st step: generate kinematics of a cluster decay to 2 hadrons



How?

We have a conditional GAN, with cluster 4-vector input and two hadron 4-vector outputs.

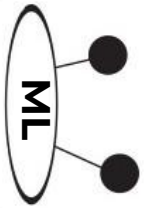
Generative Adversarial Network



Towards a Deep Learning Model for Hadronization

ML hadronization

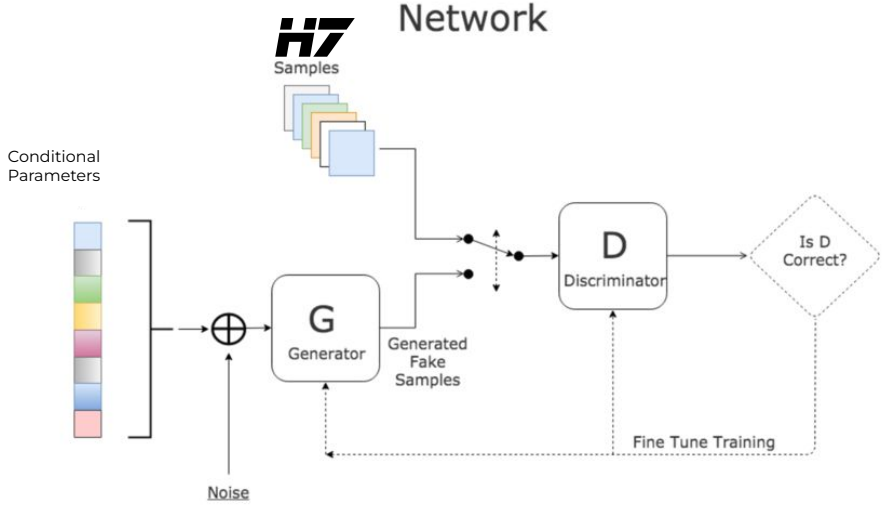
1st step: generate kinematics of a cluster decay to 2 hadrons



How?

We have a conditional GAN, with cluster 4-vector input and two hadron 4-vector outputs.

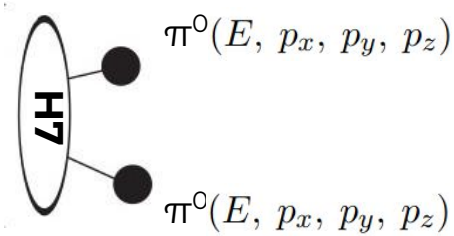
Generative Adversarial Network



Training data:

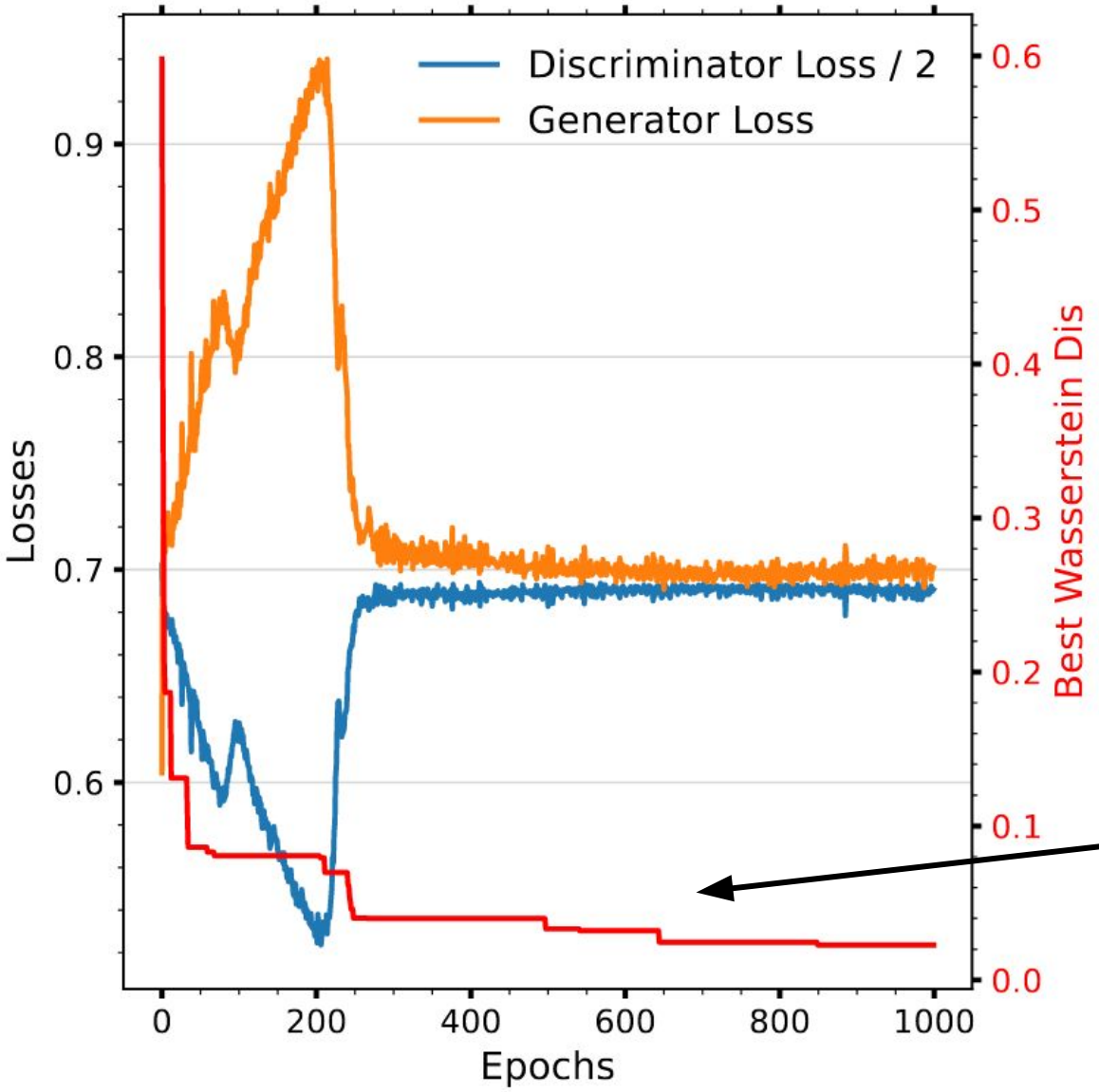
e^+e^- collisions at $\sqrt{s} = 91.2$ GeV

Cluster $(E,\ p_x,\ p_y,\ p_z)$



Simplification: considering only pions and generating two angles in the cluster rest frame.

Training HADML v1

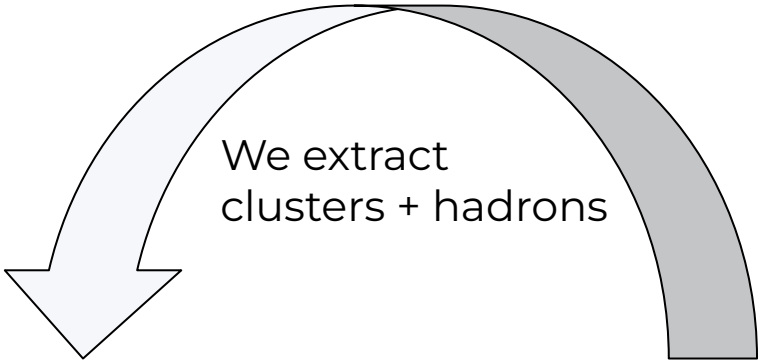


We have a conditional GAN, with cluster 4-vector input and two hadron 4-vector outputs.

<u>Simplification:</u>
considering only pions and generating two angles in the cluster rest frame.

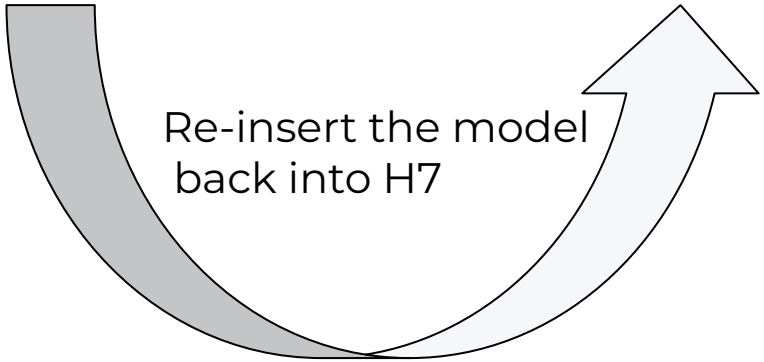
This is a typical learning curve for GAN training

Integration into Herwig



Training

Event generation

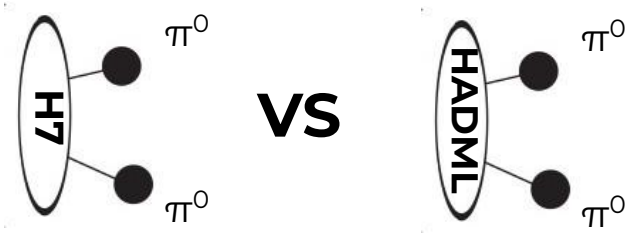


This then allows us to run a full event generator and produce plots

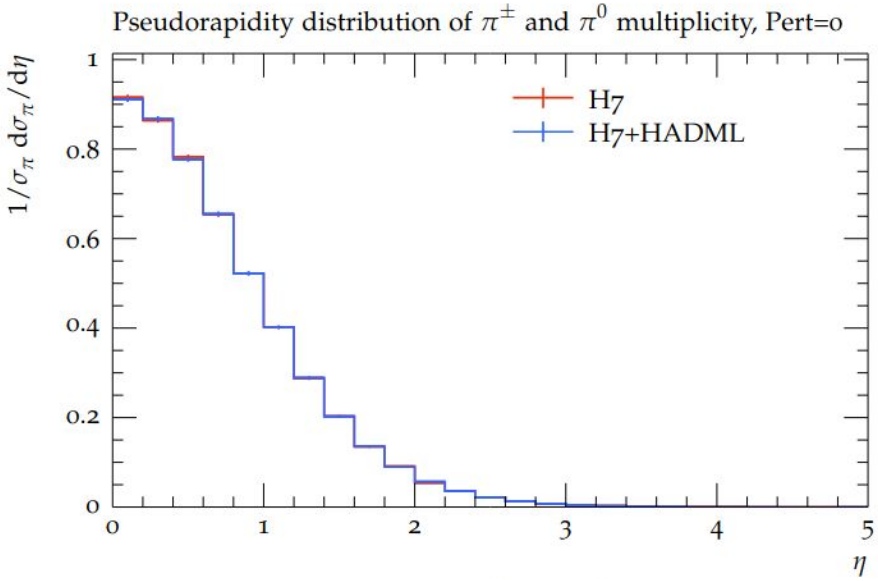
Performance: Pions

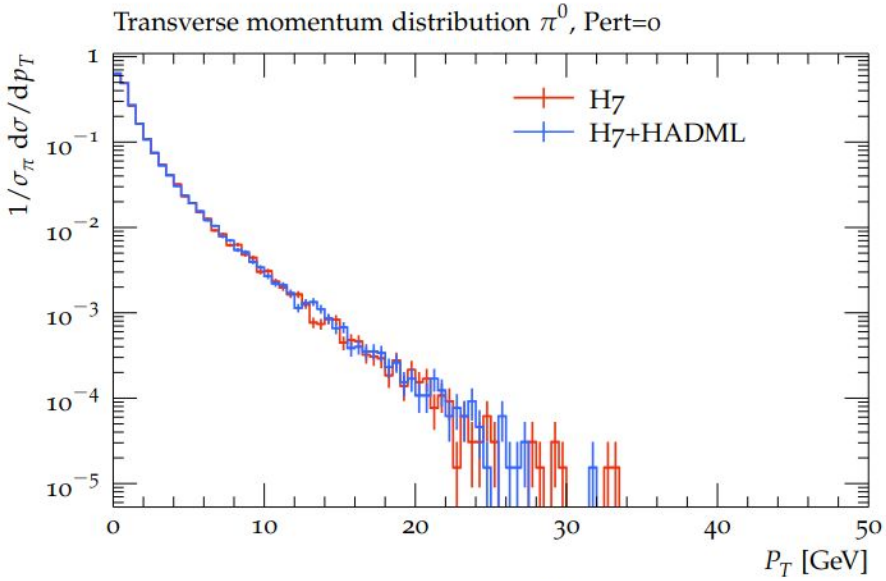
Low-level Validation
(similar to training data)

e^+e^- collisions at
$\sqrt{s} = 91.2$ GeV



π^0 kinematic variables

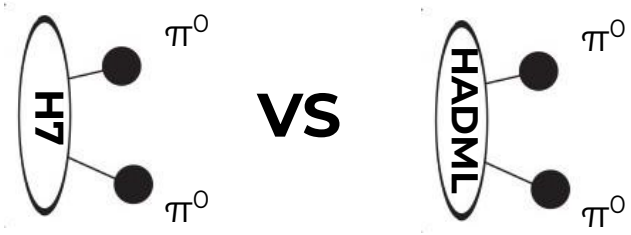




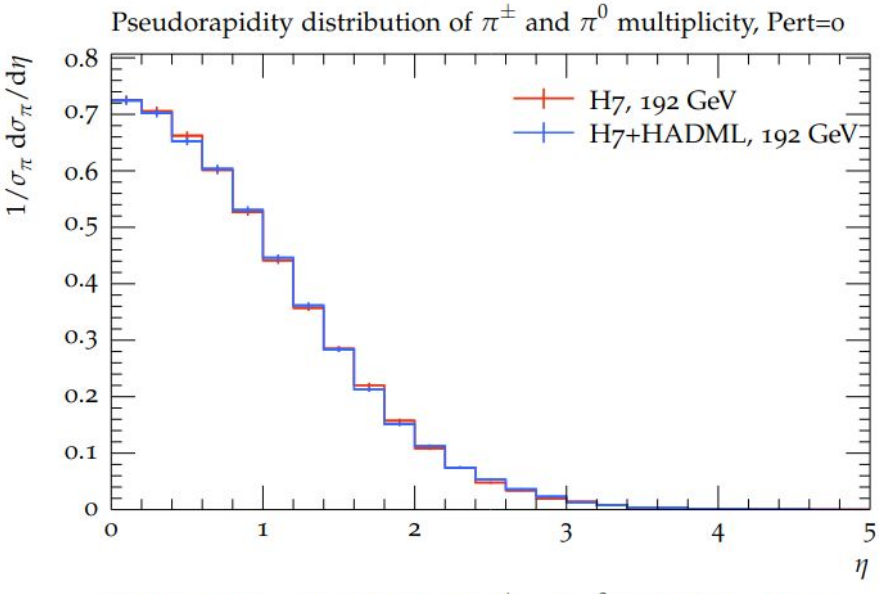
Performance: Energy of the collisions

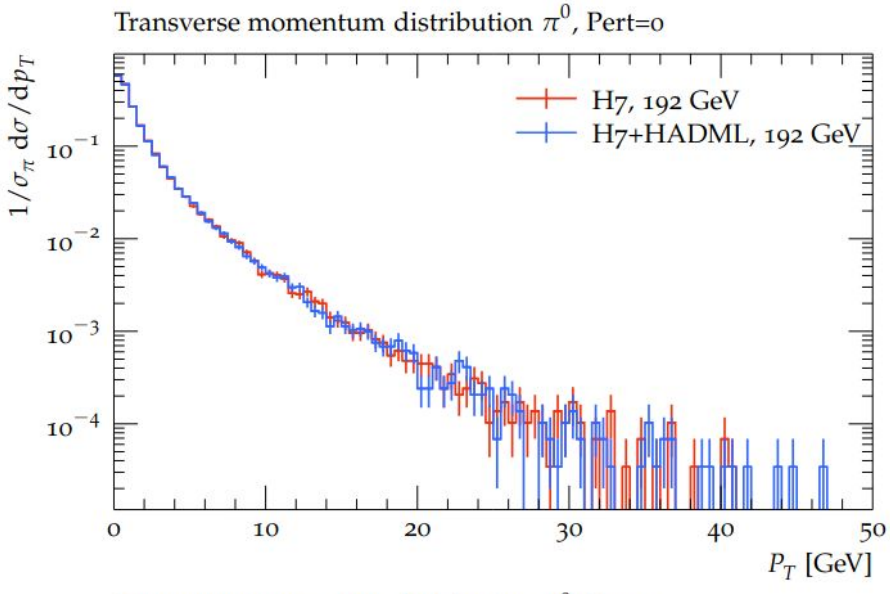
Low-level Validation
(**beyond** training data **different energy**)

e^+e^- collisions at
$\sqrt{s} = 192$ GeV



π^0 kinematic variables

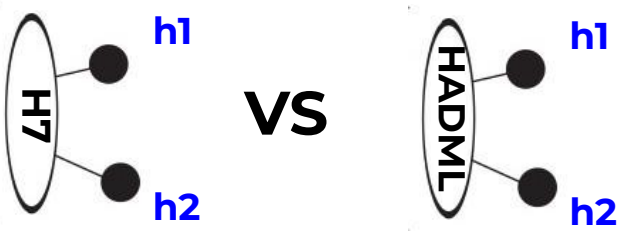




Performance: All Hadrons

Low-level Validation
(**beyond** training data **different hadrons**)

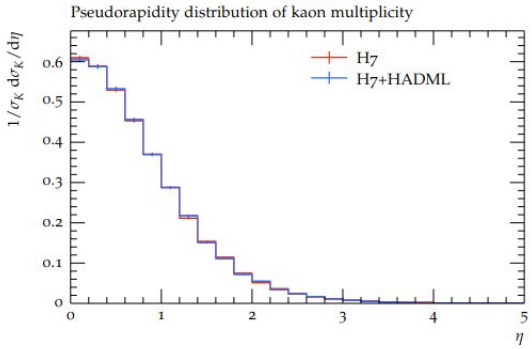
e^+e^- collisions at
$\sqrt{s} = 91.2$ GeV

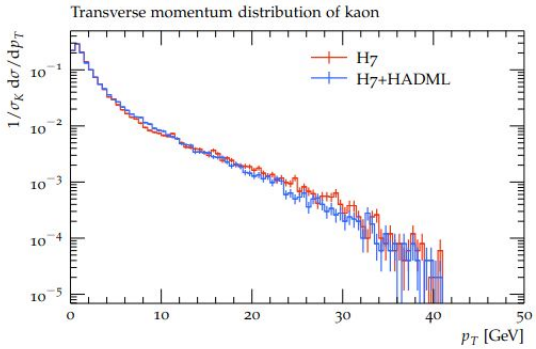


h kinematic variables

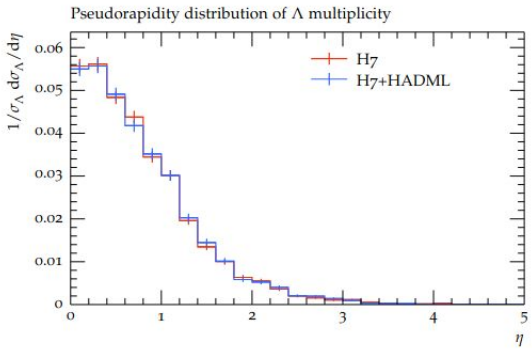
As a crude "full" model, we simply take the PIDs from Herwig and the kinematics from the GAN.

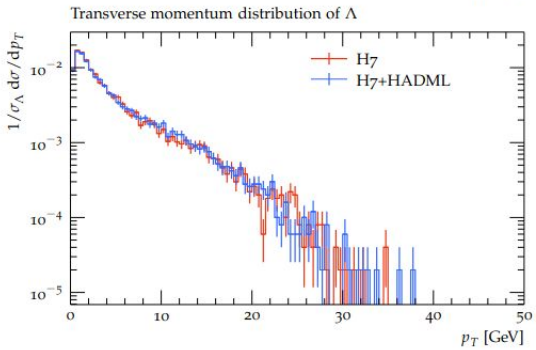
Kaons





Lambda

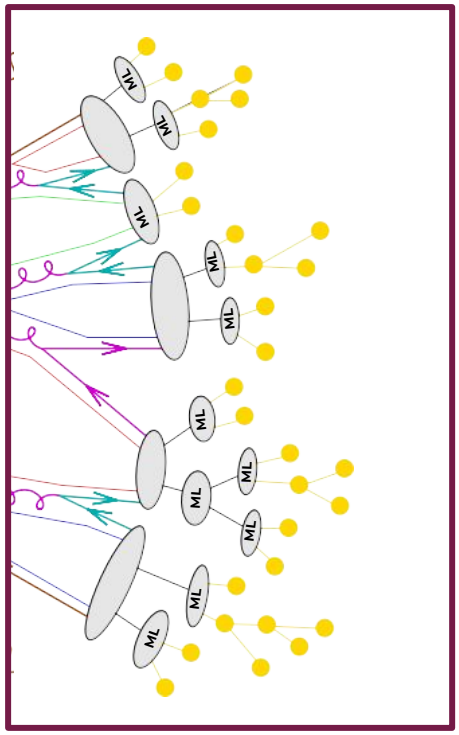


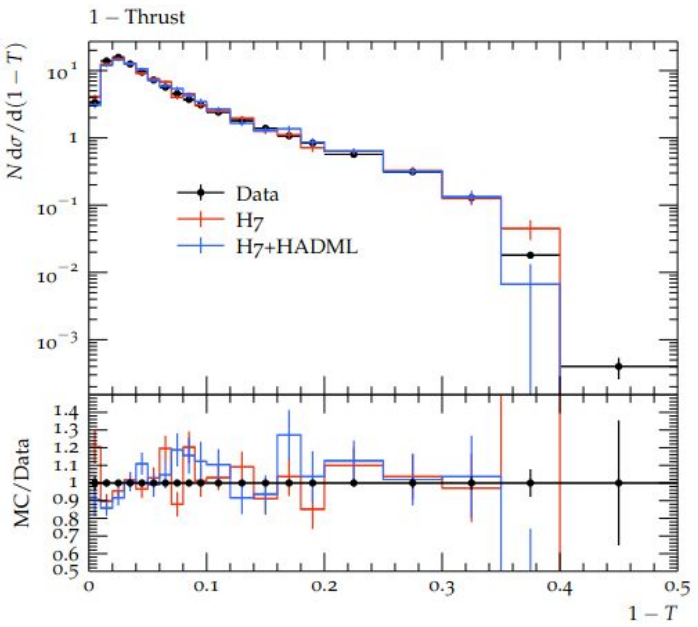


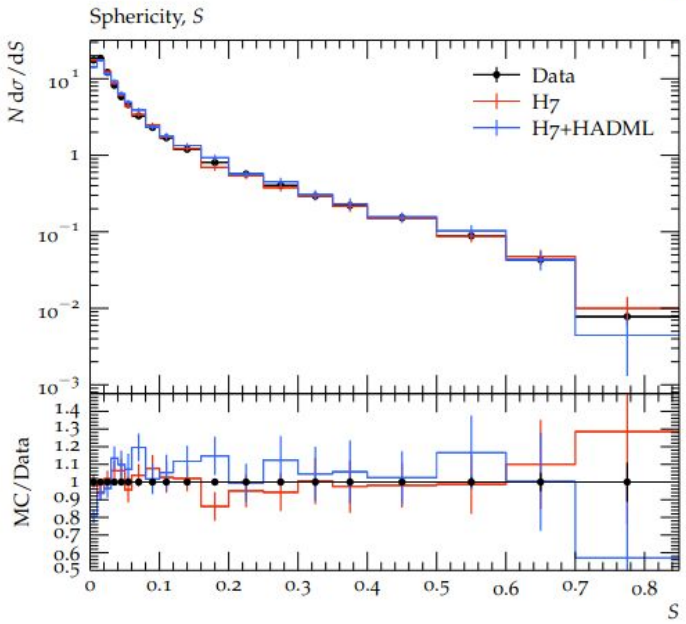
Performance: Data!

With a “full” model, we can compare directly to data!

LEP DELPHI **Data**

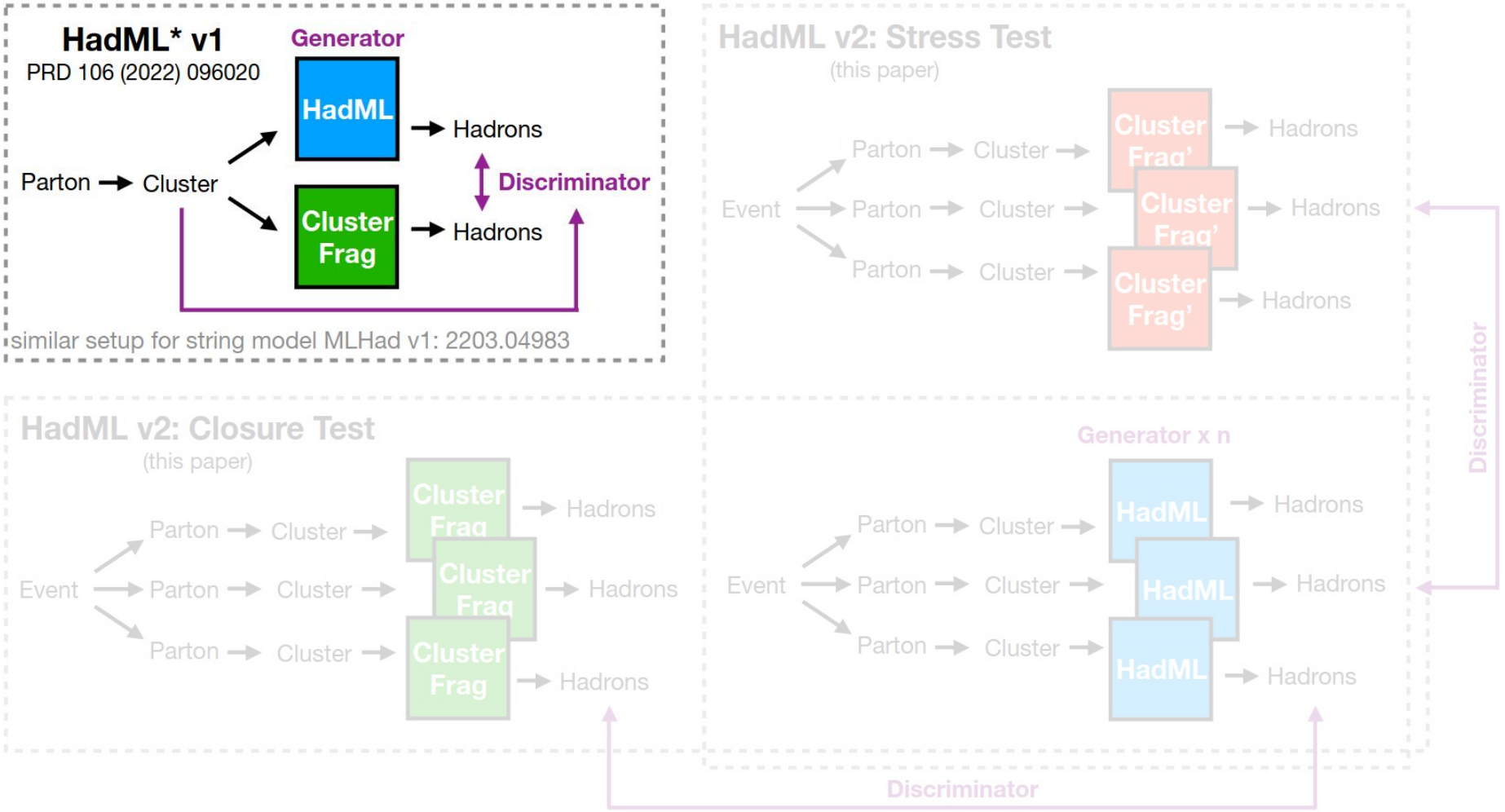




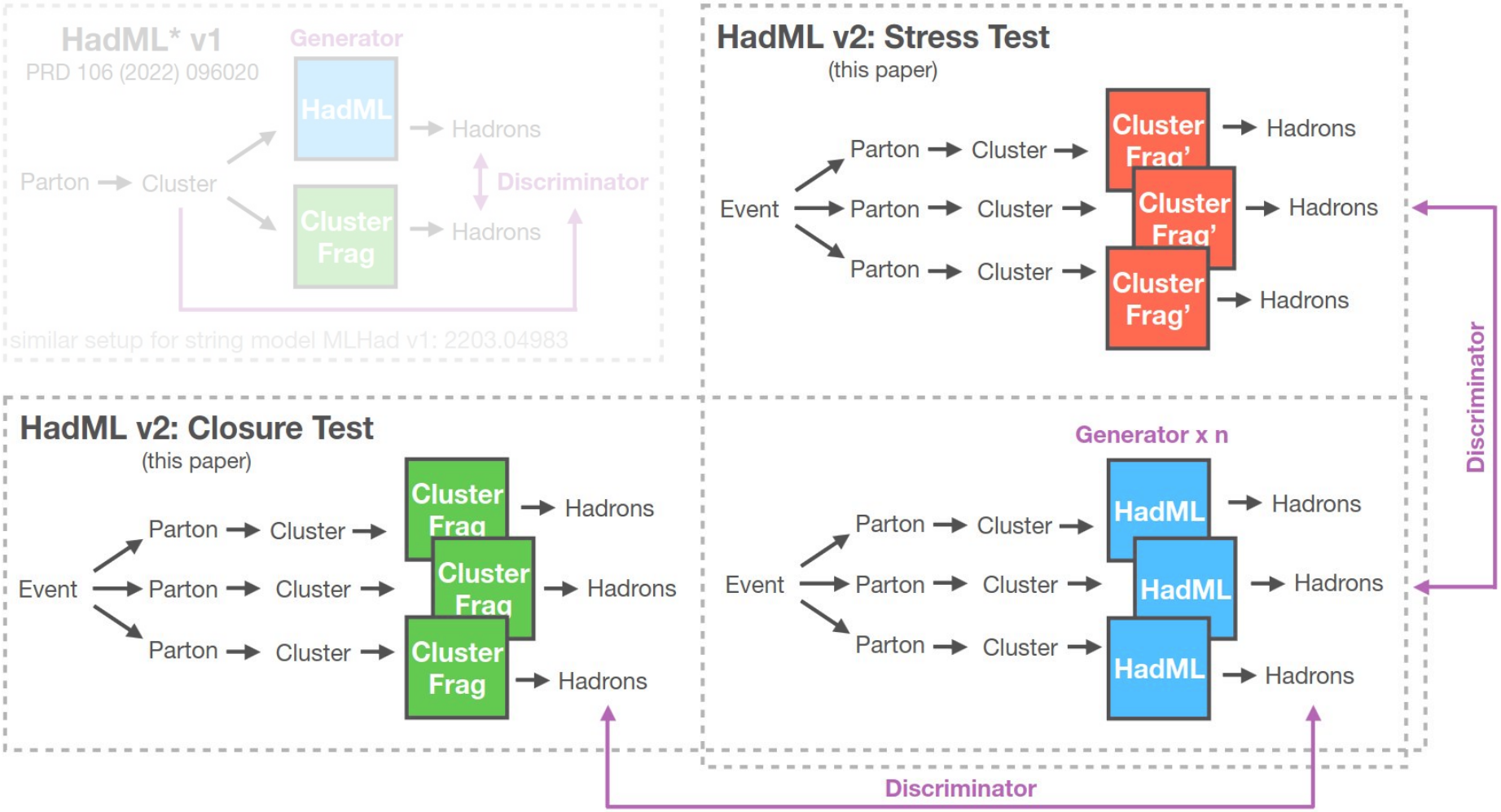


N.B. we have trained on H7, so we don’t expect to be any better than it at modeling the data.

Road map for today

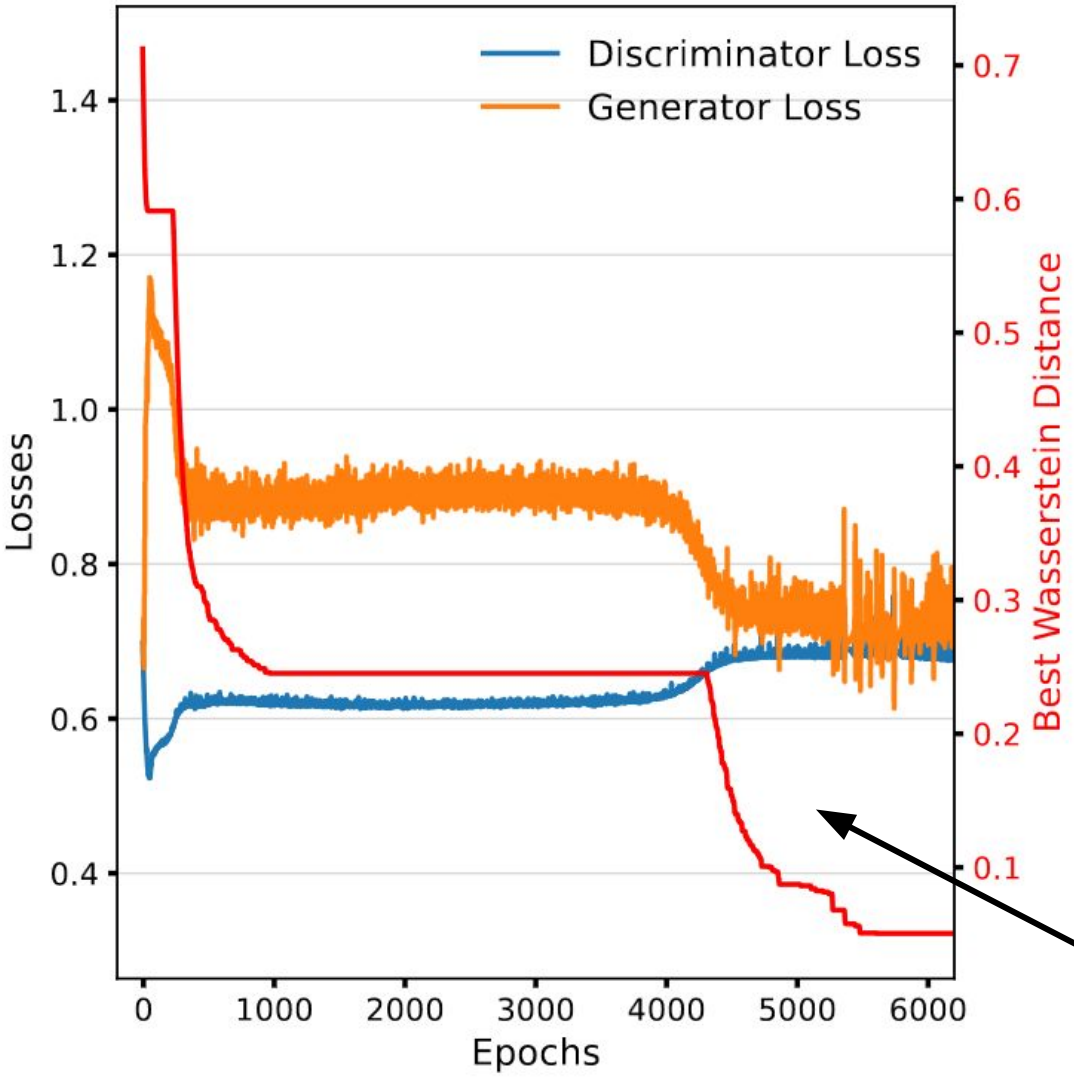


Road map for today



Protocol for fitting a deep generative hadronization model in a realistic data setting, where we only have access to a set of hadrons in data.

Training HADML v2



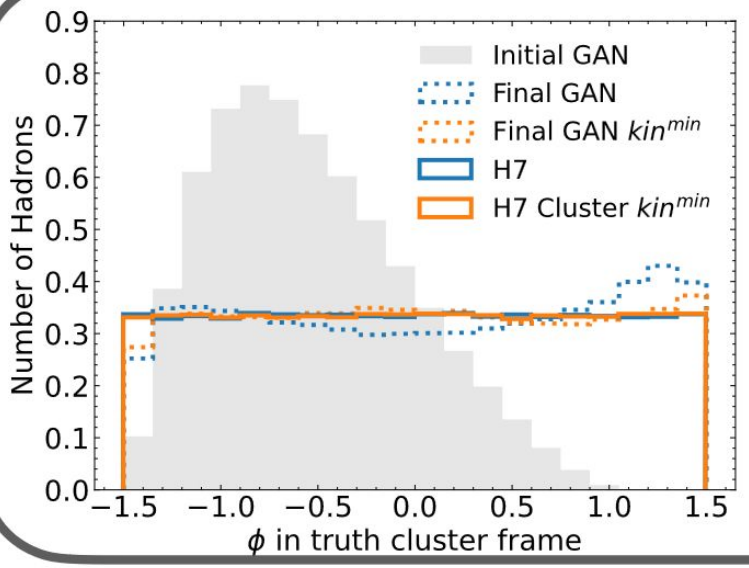
Now, the generator is local (per cluster), but the discriminator is global (whole event).

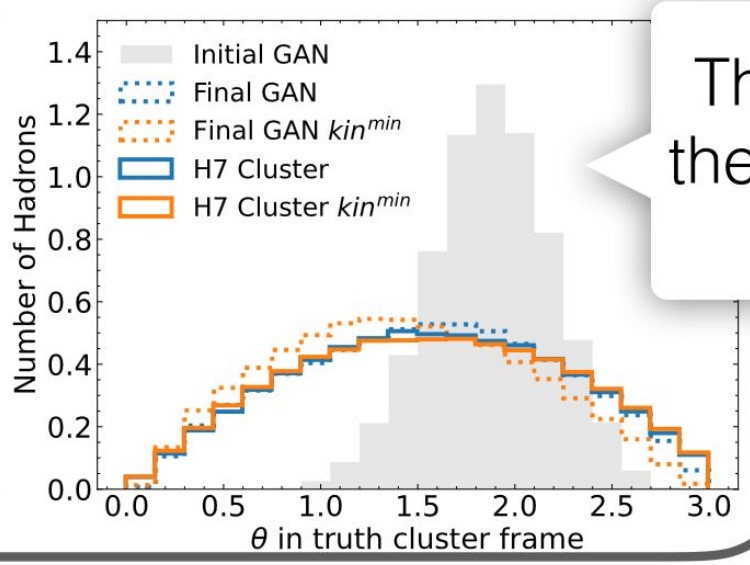
Discriminator is a permutation-invariant architecture called Deep Sets.

Simplification only Pions

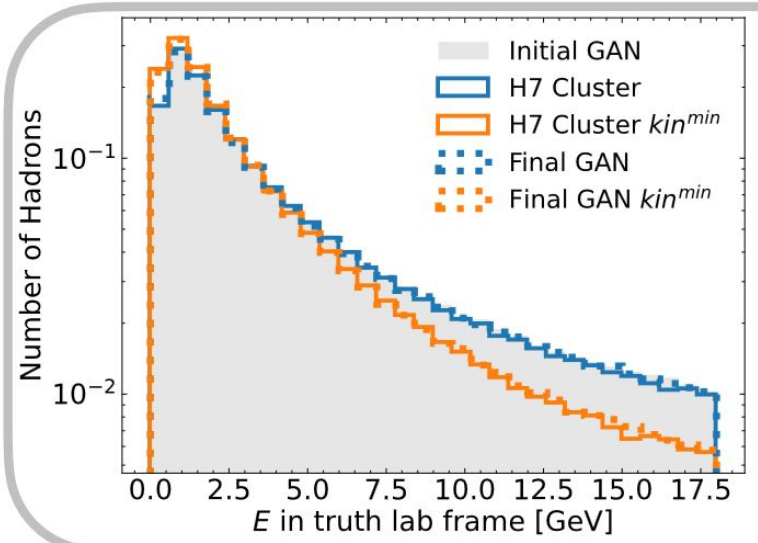
Still works !

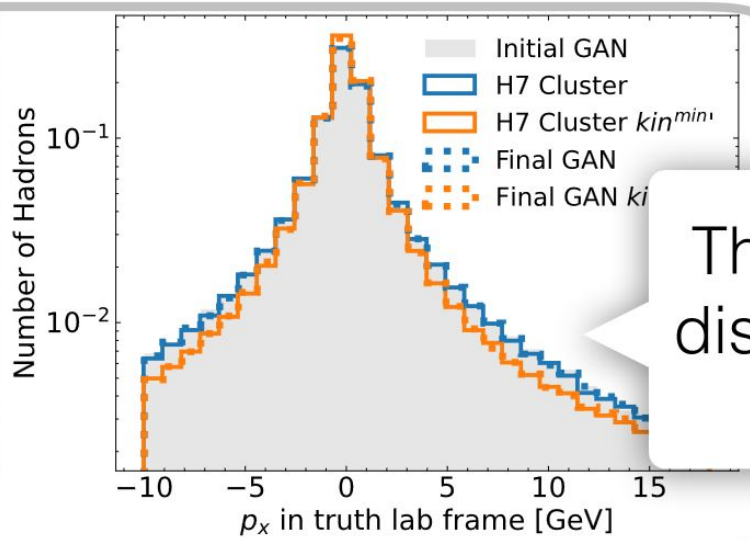
Performance





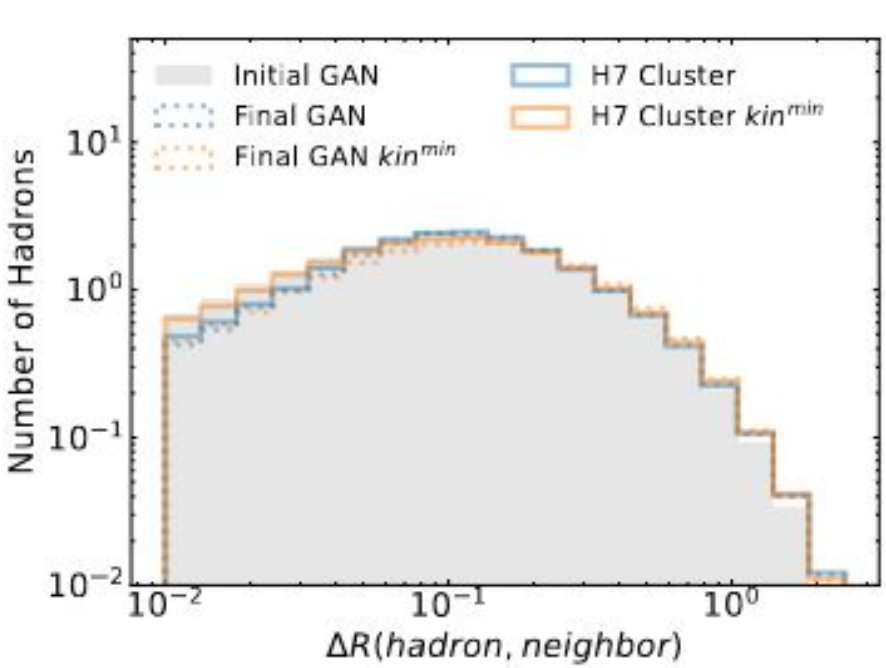
This is what the generator "sees"

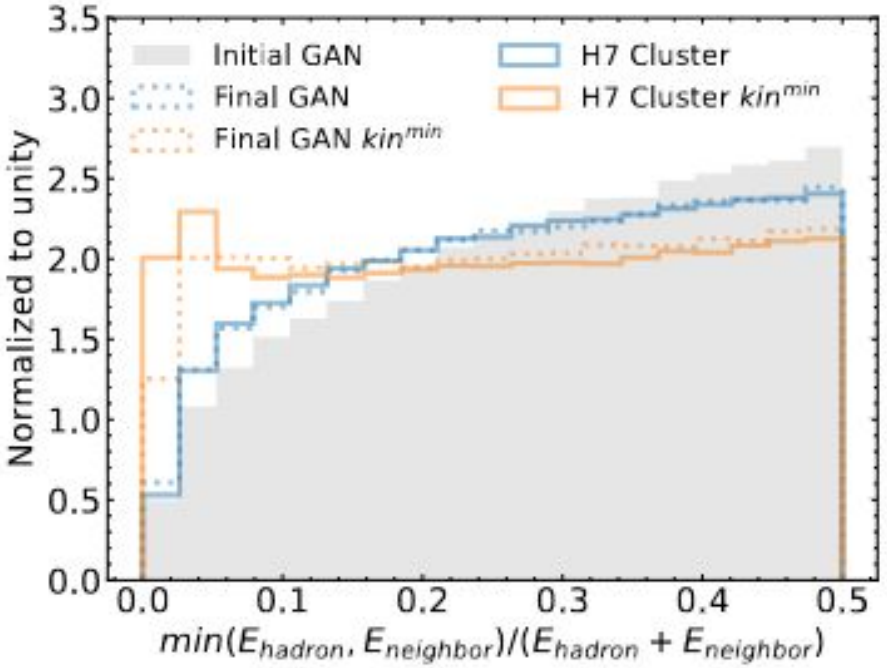




This is what discriminator "sees"

Performance: going beyond inputs and outputs

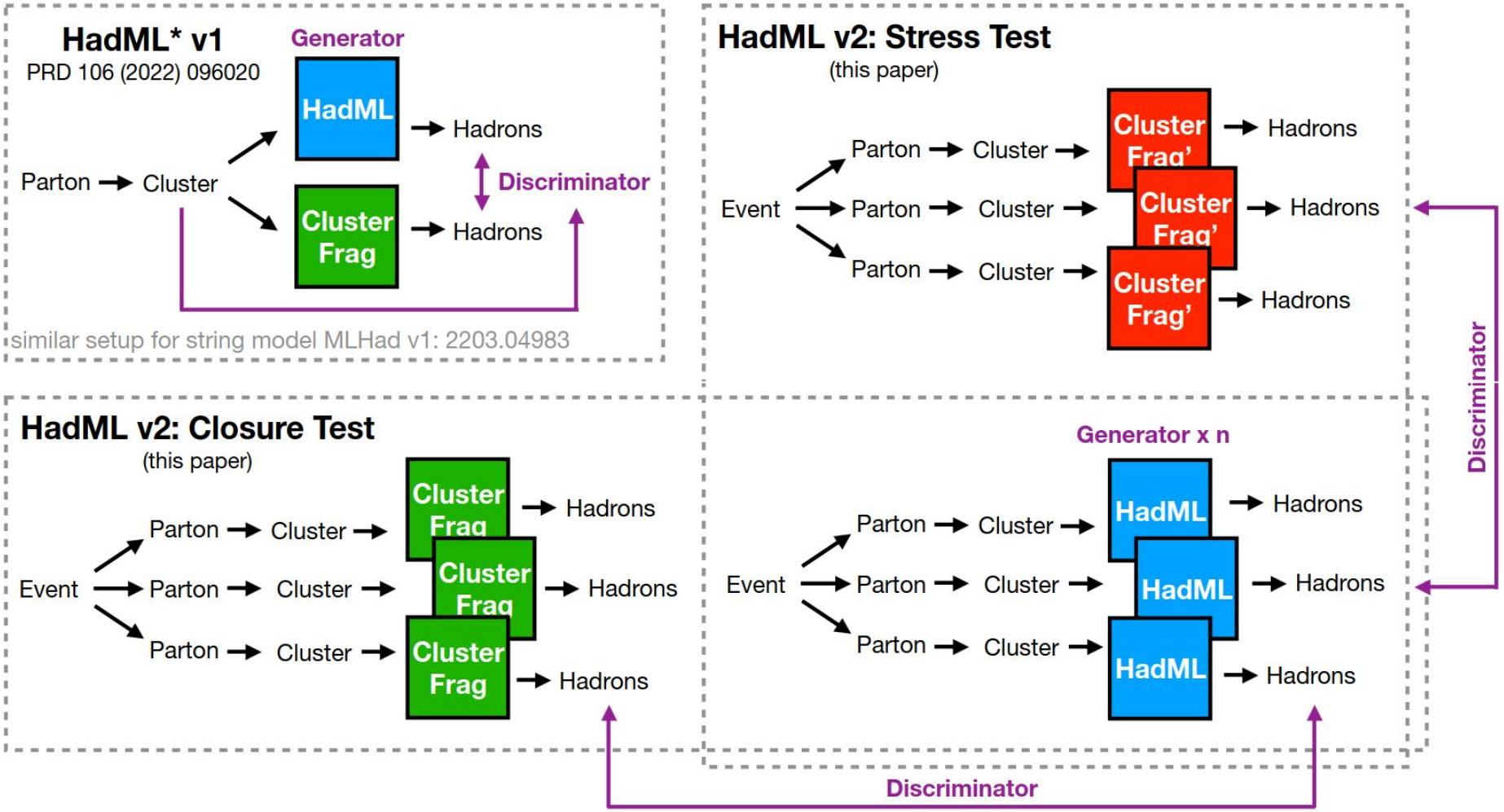




MINIMAL $\Delta R^2 = \Delta\phi^2 + \Delta\eta^2$

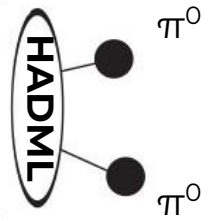
A key advantage of this fitting protocol over other methods is that it can accommodate unbinned and high-dimensional inputs.

Summary


HadML* v1
PRD 106 (2022) 096020
Generator
HadML
Hadrons
Parton
Cluster
Discriminator
Cluster Frag
Hadrons
similar setup for string model MLHad v1: 2203.04983
HadML v2: Stress Test
(this paper)
Cluster Frag'
Hadrons
Event
Parton
Cluster
Cluster Frag'
Hadrons
Parton
Cluster
Cluster Frag'
Hadrons
Discriminator
HadML v2: Closure Test
(this paper)
Cluster Frag
Hadrons
Parton
Cluster
Event
Parton
Cluster
Cluster Frag
Hadrons
Parton
Cluster
Cluster Frag
Hadrons
Generator x n
HadML
Hadrons
Parton
Cluster
Event
Parton
Cluster
HadML
Hadrons
Parton
Cluster
HadML
Hadrons
Discriminator

Outlook

For HADML, we have made significant progress,
but there are still multiple steps to build and tune a full-fledged hadronization model.

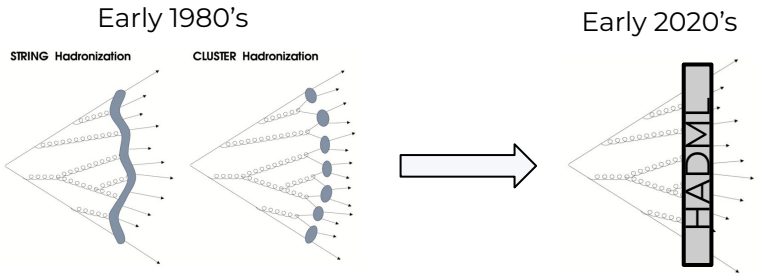


What is next?

- Number of technical and methodological step needed:

  ➔ Directly accommodate multiple hadron species with their relative probabilities
  ➔ Hyperparameter optimization, including the investigation of alternative generative models
  ➔ More flexible model with a capacity to mimic the cluster or string models as well as go beyond either model.

There is still a multi-year program ahead of us, but it will be worth it!



So Stay tuned!

Advertisement

A postdoc in ML/HEP position

JAGIELLONIAN UNIVERSITY
IN KRAKÓW

If you are interested please contact me:
andrzej.siodmok@cern.ch