HadML

A Novel Deep Generative Approach to Hadronisation

Andrzej Siódmok

HADML

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*

Phys.Rev.D 106 (2022)

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*

JHEP 09 (2023) 084

Integrating Particle Flavor into Deep Learning Models for Hadronization

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

[a] *Scientific Data Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, 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*

Phys.Rev.D (2025)

JAGIELLONIAN UNIVERSITY IN KRAKÓW

H7

CERN

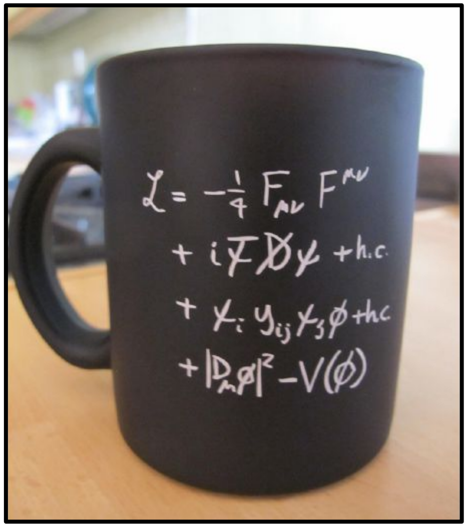
Motivation - Monte Carlo Event Generators (MCEG)

Standard Model

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



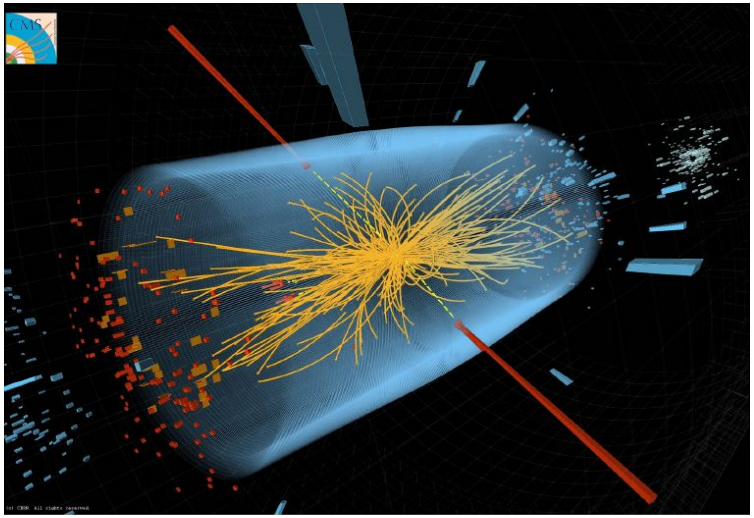
Data makes you smarter

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



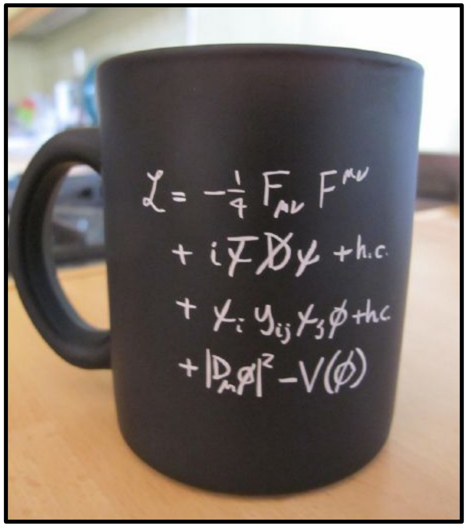
Motivation - Monte Carlo Event Generators (MCEG)

Standard Model

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

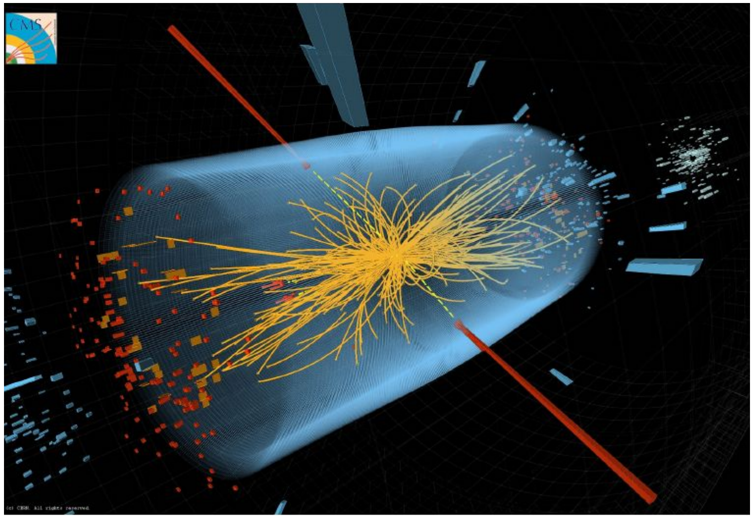
Theory

Standard Model Lagrangian



Experiment

LHC event



- MC event generators are designed to bridge the 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.

(Herwig and Sherpa)@**UJ**, Pythia

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

Motivation - Monte Carlo Event Generators (MCEG)

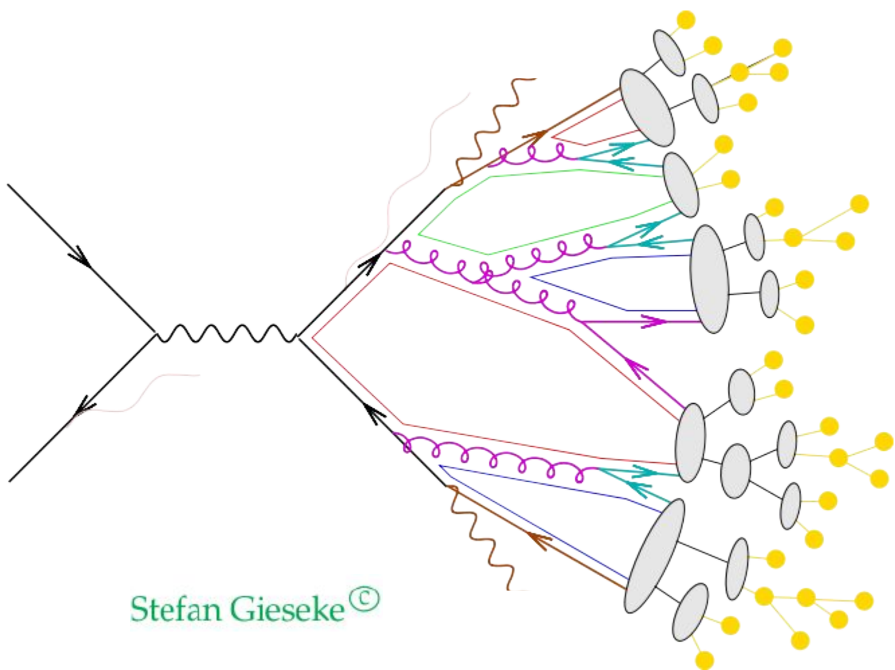
QCD correctly describes strong interactions in each energy range but its complex mathematical structure makes it very difficult to obtain precise predictions (Millennium Prize Problem $1,000,000)

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)**

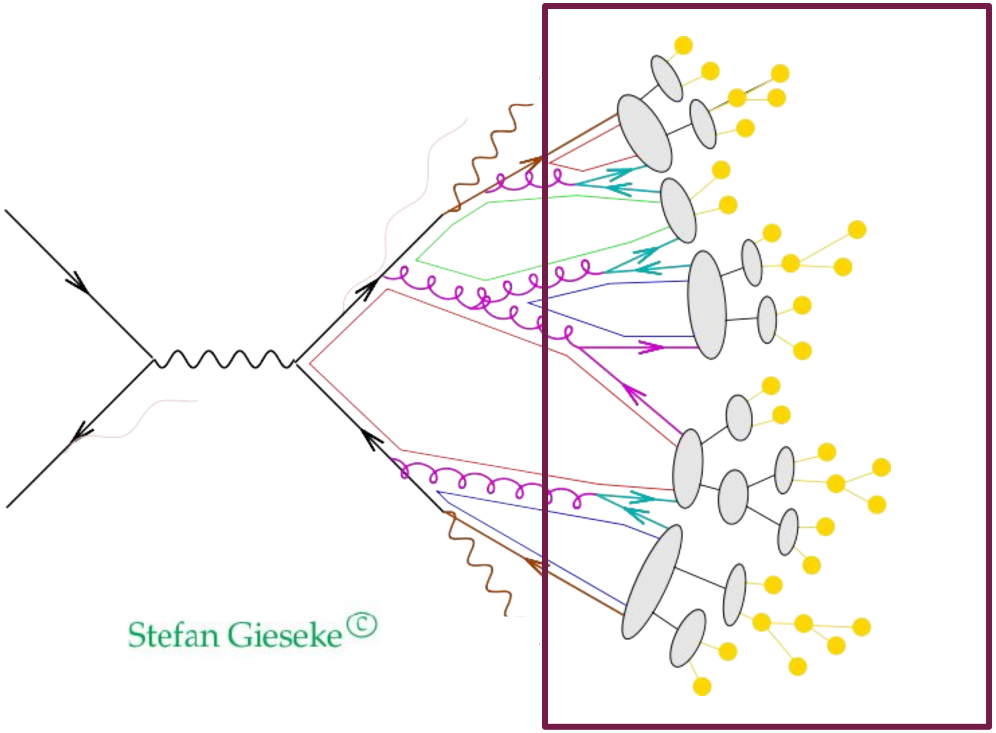


Why hadronization?

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

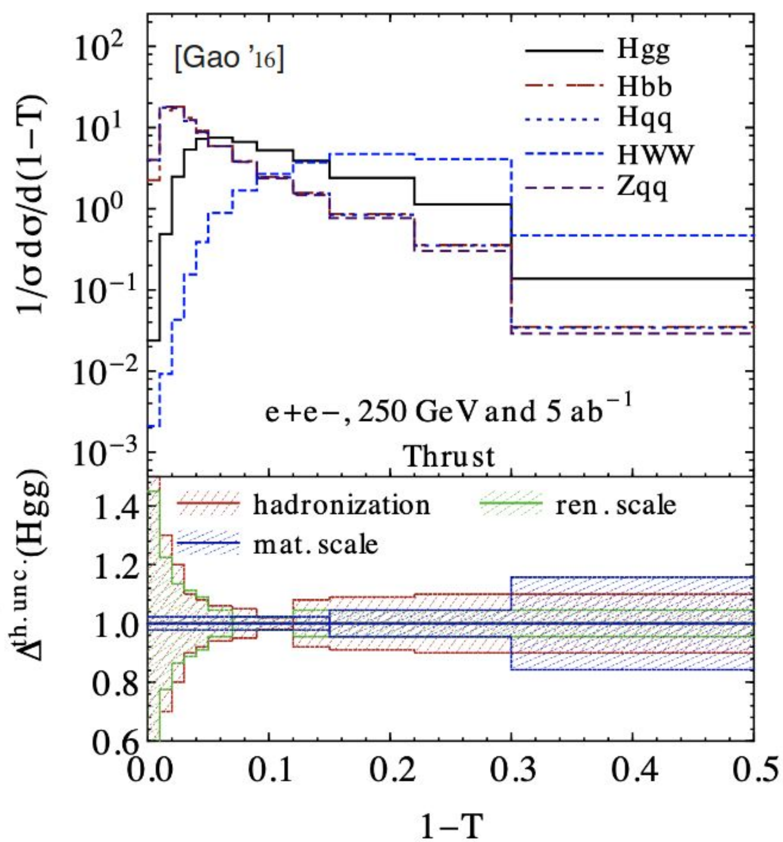
Hadronization:

➔ 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 ($Q \sim m_H$)
- 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



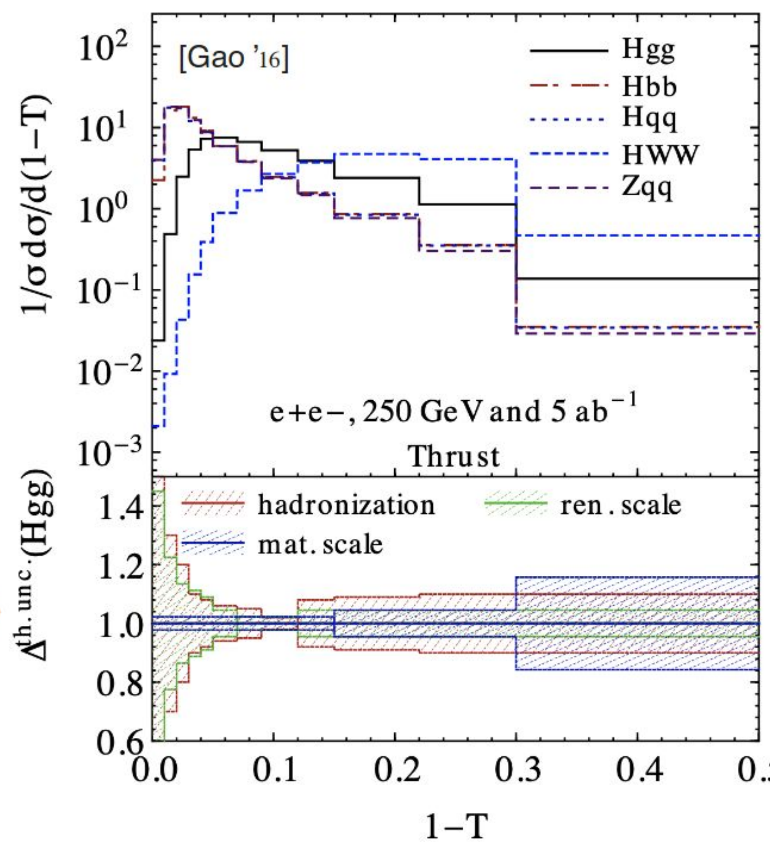
Hadronization:

➔ 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

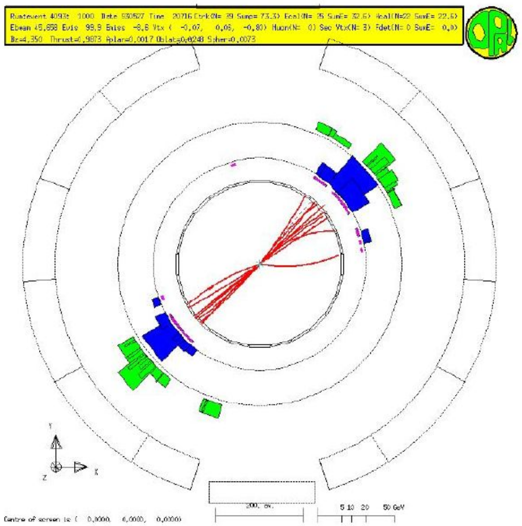
- However, hadronisation remains the main bottleneck
  - e.g. thrust in Higgs decays (MC variation in plot)
- Increase in energy insufficient for suppression ($Q \sim m_H$)



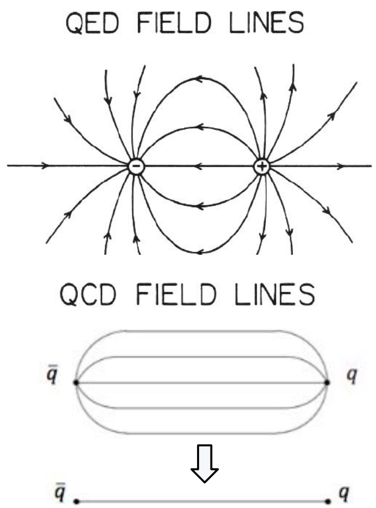
String Model in nutshell [Andersson, Gustafson, Ingelman, Sjostrand, Phys.Rept.97(1983)31]

Originally invented without perturbative physics of parton showers in mind.

We start with 2-jet events in
$e^+ e^- \to$ hadrons.



Self coupling of gluons
“attractive field line”

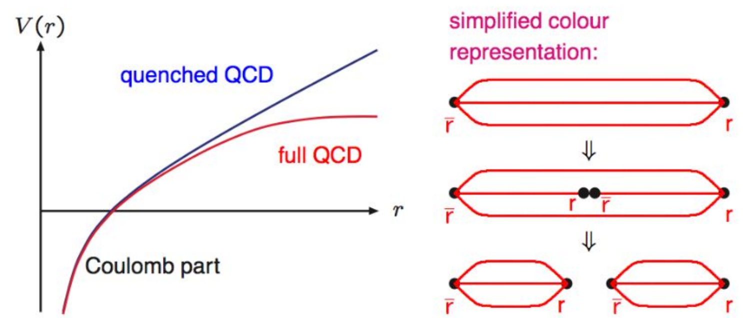


1+1 - dim. object: **string**

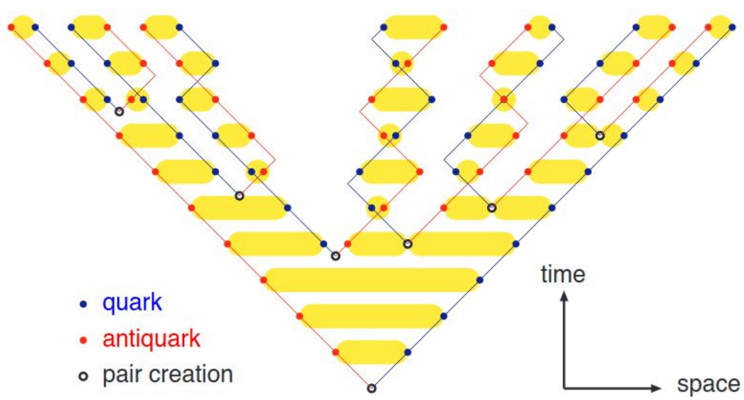
Linear static potential:

$$F(r) \approx \text{const} = \kappa \approx 1\ \text{GeV/fm} \quad \Longleftrightarrow \quad V(r) \approx \kappa r$$

Picture supported by lattice QCD



Lund string model: like rubber band that is pulled apart and breaks into pieces



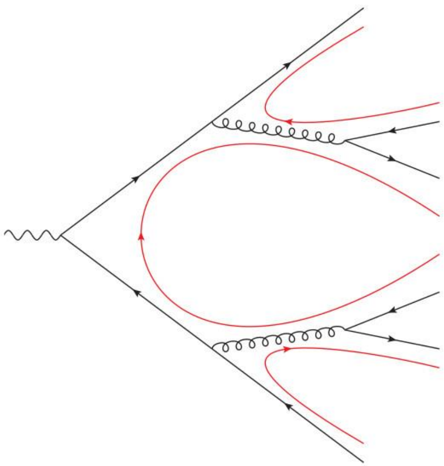
Plots from T. Sjostrand

Cluster model in nutshell

[Webber NPB238(1984)492]

What if we have PS (more perturbative input before hadronization).

The philosophy of the model: use information from perturbative QCD as an input for hadronization. QCD **pre-confinement** discovered by Amati & Veneziano [*Phys.Lett.B* 83 (1979) 87-92]:



- QCD provide pre-confinement of colour

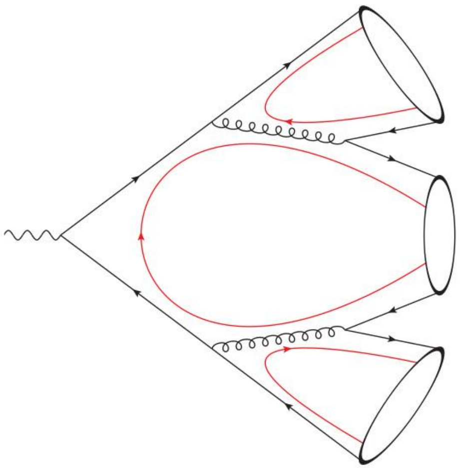
. . .

Cluster model in nutshell

[Webber NPB238(1984)492]

What if we have PS (more perturbative input before hadronization).

The philosophy of the model: use information from perturbative QCD as an input for hadronization. QCD **pre-confinement** discovered by Amati & Veneziano [*Phys.Lett.B* 83 (1979) 87-92]:



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

. . .

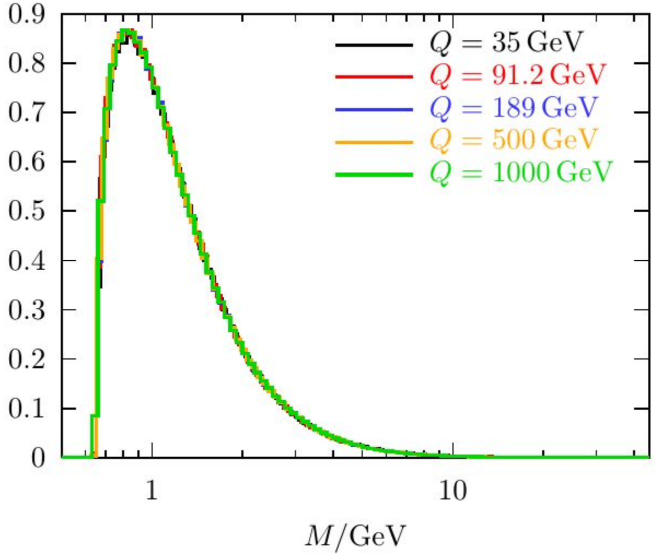
Cluster model in nutshell

[Webber NPB238(1984)492]

What if we have PS (more perturbative input before hadronization).

The philosophy of the model: use information from perturbative QCD as an input for hadronization.
QCD **pre-confinement** discovered by Amati & Veneziano [*Phys.Lett.B* 83 (1979) 87-92]:

[S. Gieseke, A. Ribon, MH Seymour,
P Stephens, B Webber JHEP 0402 (2004) 005]



- 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

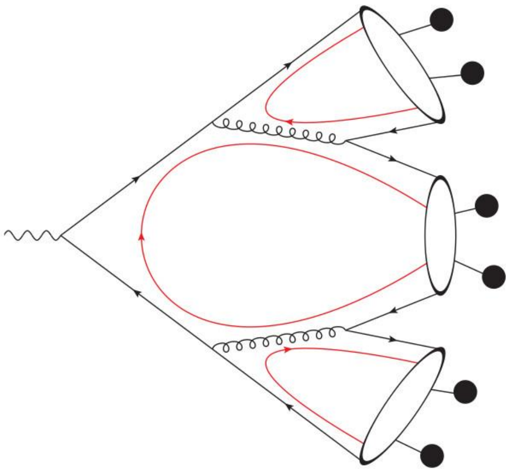
. . .

Cluster model in nutshell

[Webber NPB238(1984)492]

What if we have PS (more perturbative input before hadronization).

The philosophy of the model: use information from perturbative QCD as an input for hadronization. QCD **pre-confinement** discovered by Amati & Veneziano [*Phys.Lett.B* 83 (1979) 87-92]:



- 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

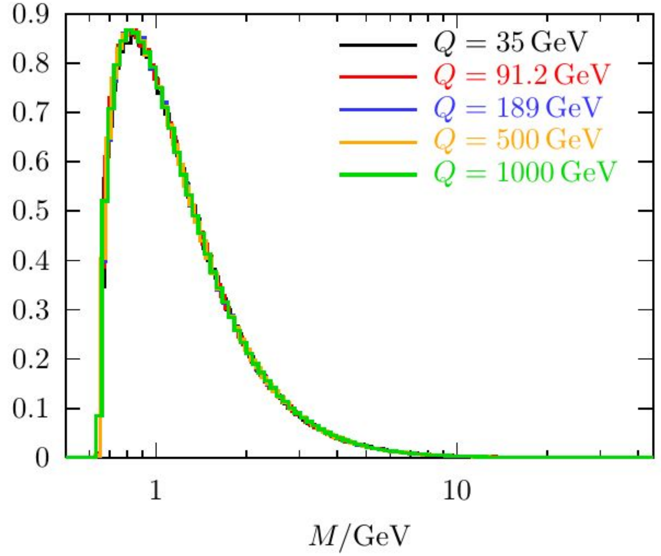
Cluster model in nutshell

[Webber NPB238(1984)492]

What if we have PS (more perturbative input before hadronization).

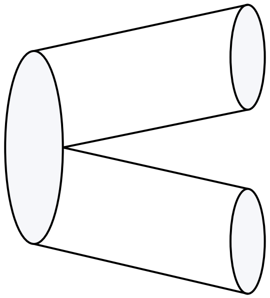
The philosophy of the model: use information from perturbative QCD as an input for hadronization.
QCD **pre-confinement** discovered by Amati & Veneziano [*Phys.Lett.B* 83 (1979) 87-92]:

[S. Gieseke, A. Ribon, MH Seymour, P Stephens, B Webber JHEP 0402 (2004) 005]

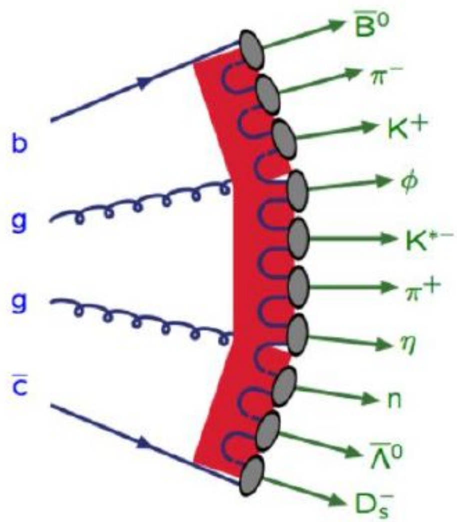


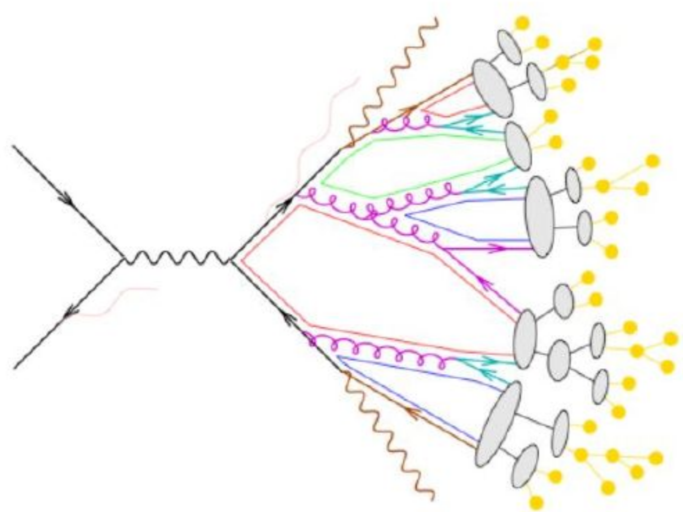
- QCD provide pre-confinement of colour
- Colour-singlet pair end up close in phase space and form highly excited hadronic states, the clusters
- Pre-confinement states that the spectra of clusters are independent of the hard process and energy of the collision
- Peaked at low mass (1-10 GeV) typically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

- Small fraction of clusters too heavy for isotropic two-body decay, heavy cluster decay first into lighter cluster C → CC, or radiate a hadron C → HC, it is rather string-like.
- ~ 15% of primary clusters get split but ~ 50% of hadrons come from them!



String vs Cluster model





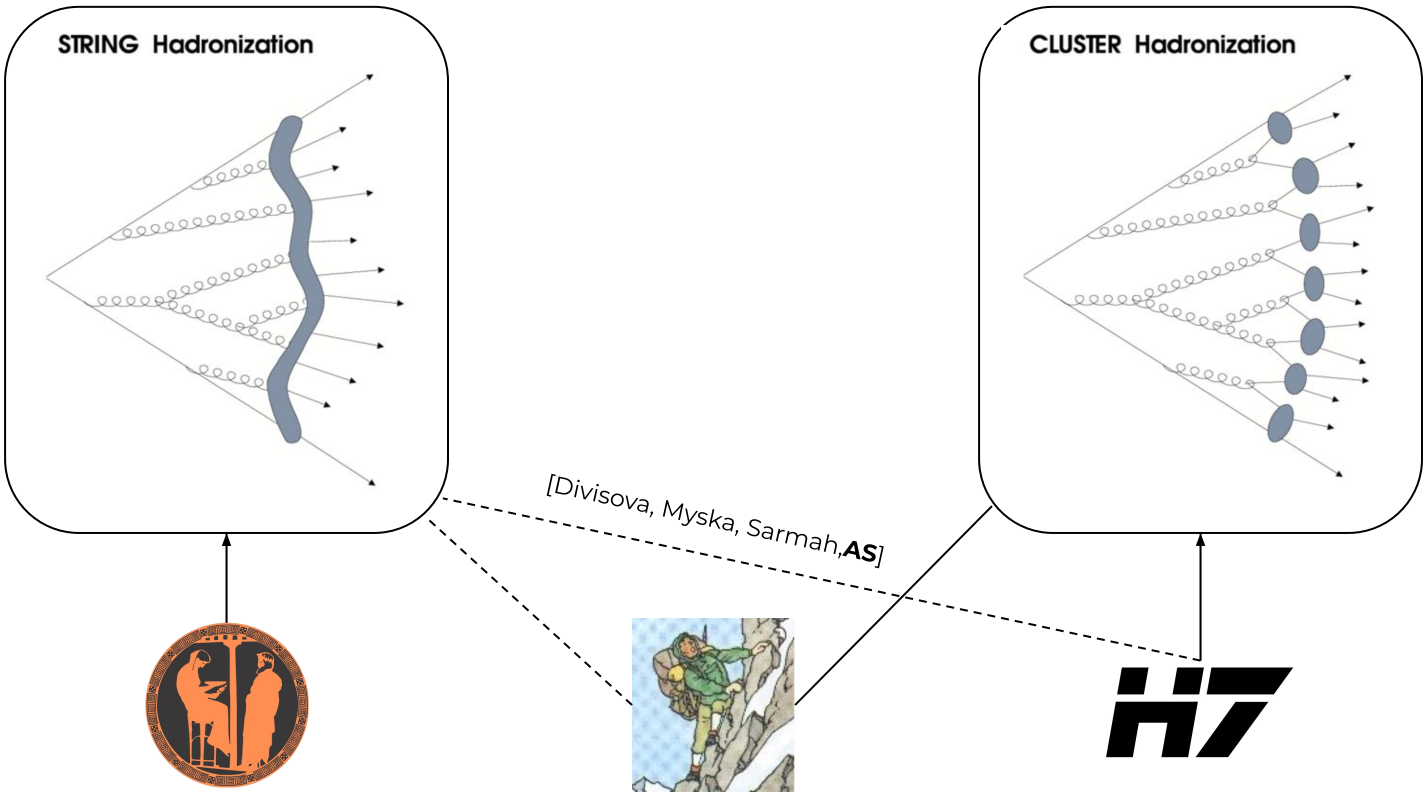
program	PYTHIA	Herwig
model	string	cluster
energy–momentum picture	powerful	simple
	predictive	unpredictive
parameters	few	many
flavour composition	messy	simple
	unpredictive	in-between
parameters	many	few

Taken from T. Sjostrand

Hadronization:

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

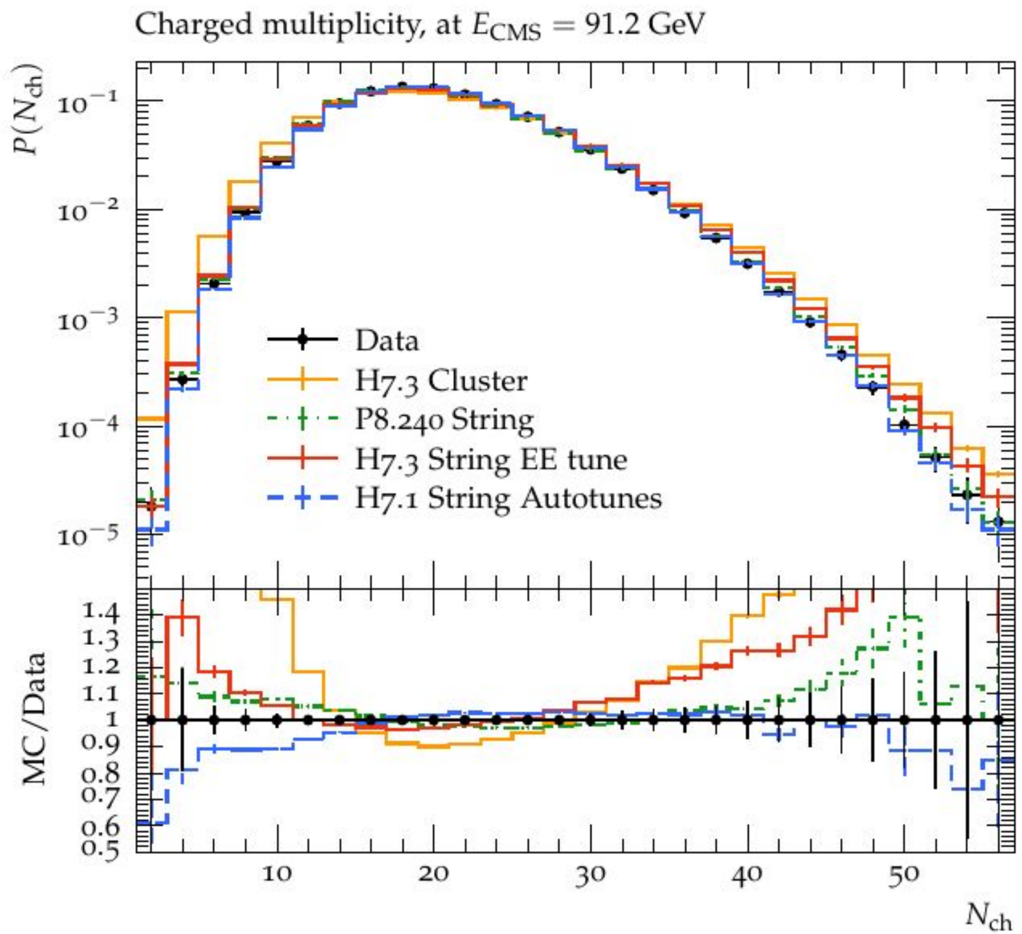
- W mass measurement using a new method [Freytsis at al. JHEP 1902 (2019) 003]
- Extraction of the strong coupling in [M. Johnson, D. Maître, Phys.Rev. D97 (2018) no.5]
- Top mass [S. Argyropoulos, T. Sjöstrand, JHEP 1411 (2014) 043]
- ...



Pythia **Sherpa** **Herwig**

Hadronisation effects with Angular Ordered Parton Shower

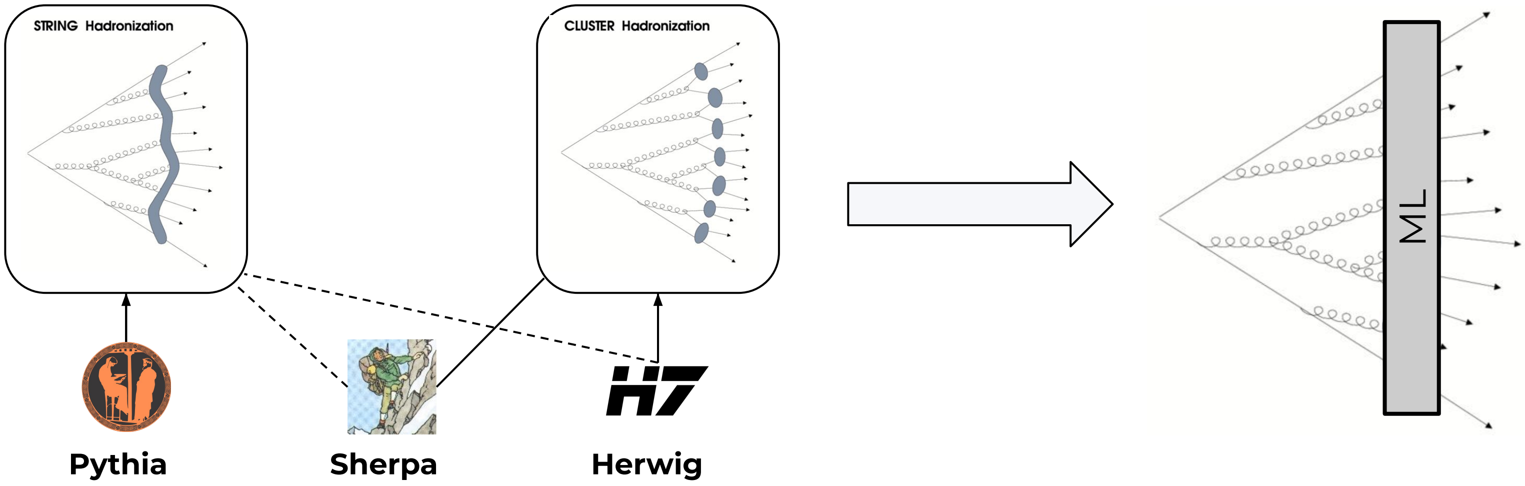
Michaela Divisova[a,1], Miroslav Myska[b,1], Pratixan Sarmah[c,2], Andrzej Siódmok[d,2]



Hadronization:

Early 1980's
(limited progress)

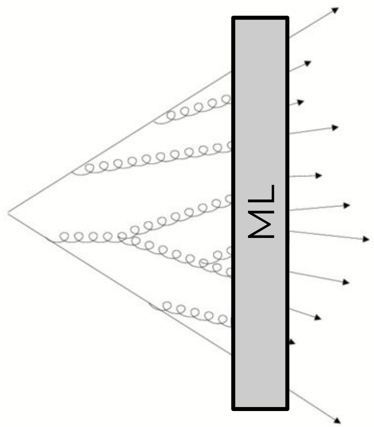
Early 2020's
(lot of progress in ML)



Idea of using Machine Learning (ML) for hadronization.

Motivation for Machine learning hadronization

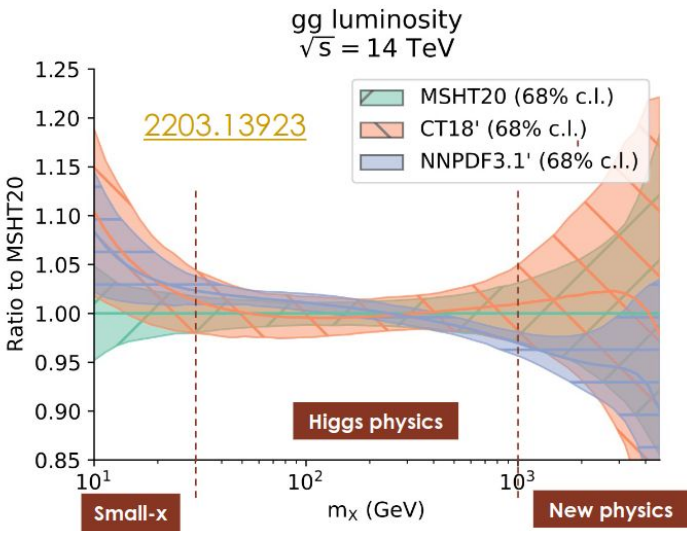
Idea of using Machine Learning (ML) for hadronization.



- Existing hadronization models are highly parameterized functions.
- Hadronization is a fitting problem
  - Can ML hadronization be more flexible to fit the data?
  - Can ML hadronization extract more information from the data? [can accommodate unbinned and high-dimensional inputs]

NNPDF

NNPDF used successfully ML to nonperturbative Parton Density Functions (PDF).
Hadronization is closely related to fragmentation functions (FF) which were considered the counterpart of PDFs.



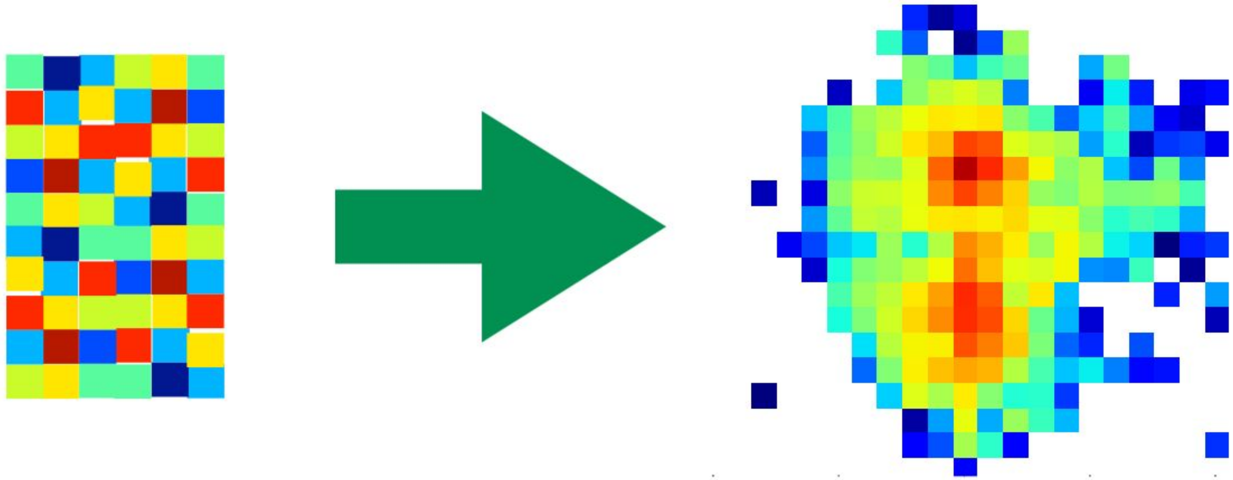
First steps for ML hadronization:

- HADML - [A. Ghosh, Xi. Ju, B. Nachman **AS**, *Phys.Rev.D* 106 (2022) 9]
- MLhad - [P. Ilten, T. Menzo, A. Youssef and J. Zupan, SciPost Phys. 14, 027 (2023)]

	MLhad	HADML
Deep generative model:	Variational Autoencoder	Generative Adversarial Networks
Trained on:	String model	Cluster model
Recent progress:	*“Reweighting Monte Carlo Predictions and Automated Fragmentation Variations in Pythia 8”* [Bierlich, Ilten, Menzo, Mrenna, Szewc, Wilkinson, Youssef, Zupan, 2308.13459] ...	*“Fitting a Deep Generative Hadronization Model”* [J. Chan, X. Ju, A. Kania, B. Nachman, V. Sangli and **AS**, JHEP 09 (2023) 084] ...

MLhadML (2411.02194)

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

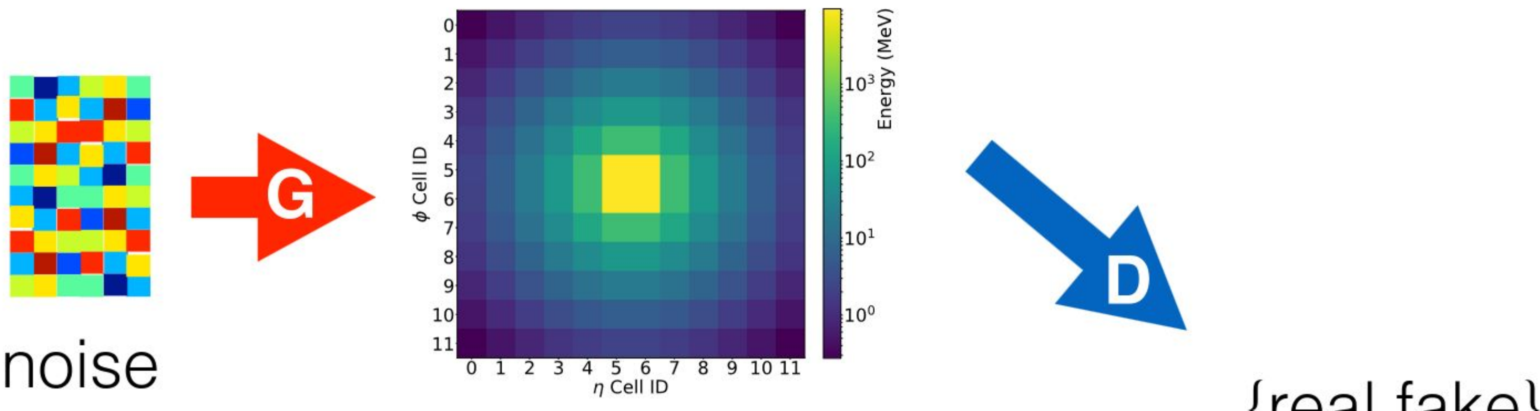


Deep generative models: the map is a deep neural network.

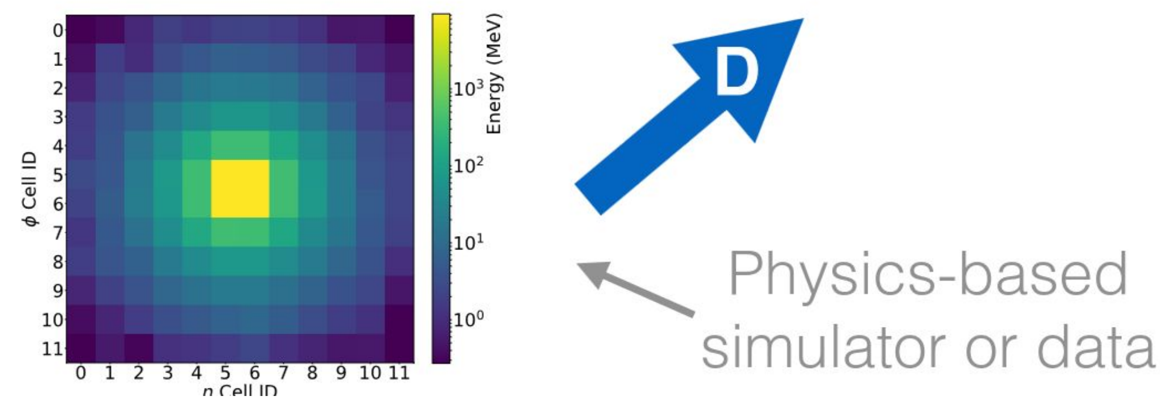
Our tool of choice: GANs

[Goodfellow et al. “Generative adversarial nets”. arxiv:1406.2661]

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

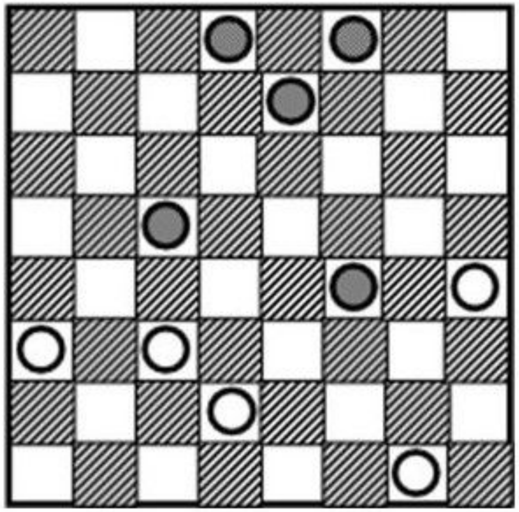


When **D** is maximally confused, **G** will be a good generator



Adversarial Networks

Arthur Lee Samuel (1959) wrote a program that learnt to play checkers well enough to beat him.



- He popularized the term **"machine learning"** in 1959.
- The program chose its move based on a **minimax** strategy, meaning it made the move assuming that the opponent was trying to optimize the value of the same function from its point of view.
- He also had it play thousands of **games against itself** as another way of learning.

Adversarial Networks

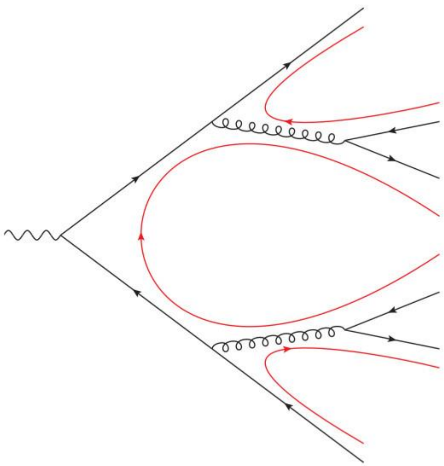
Demis Hassabis

CBE FRS FREng FRSA

By playing **games against itself**, AlphaGo Zero surpassed the strength of AlphaGo Lee in three days by winning 100 games to 0.

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

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



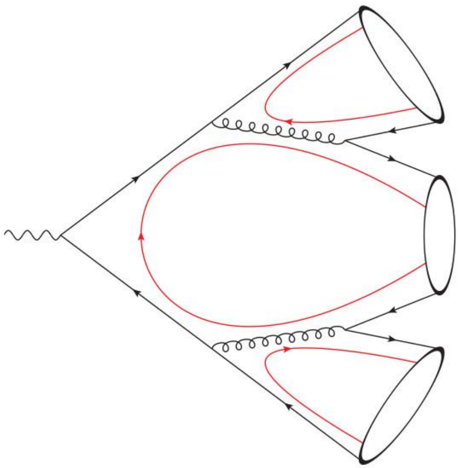
- QCD provide pre-confinement of colour

. . .

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:

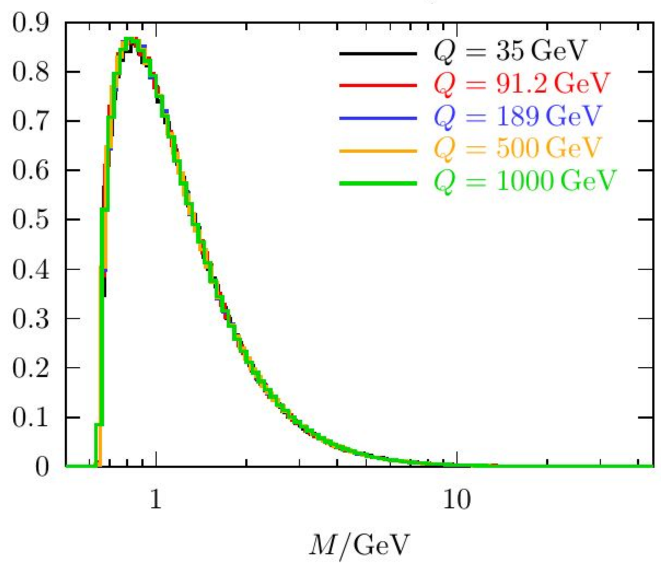


- QCD provide pre-confinement of colour
- Colour-singlet pair end up close in phase space and form highly excited hadronic states, the clusters

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:



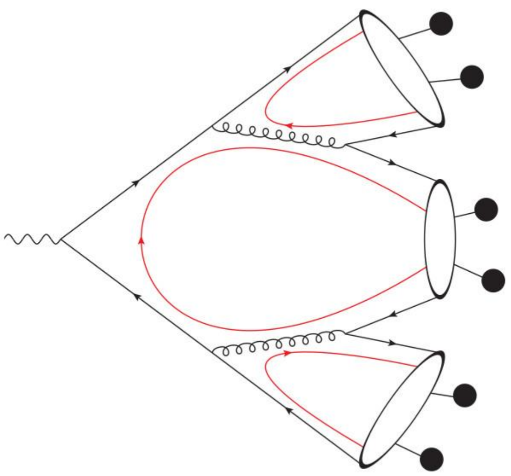
[S. Gieseke, A. Ribon, MH Seymour,
P Stephens, B Webber JHEP 0402 (2004) 005]

- 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:

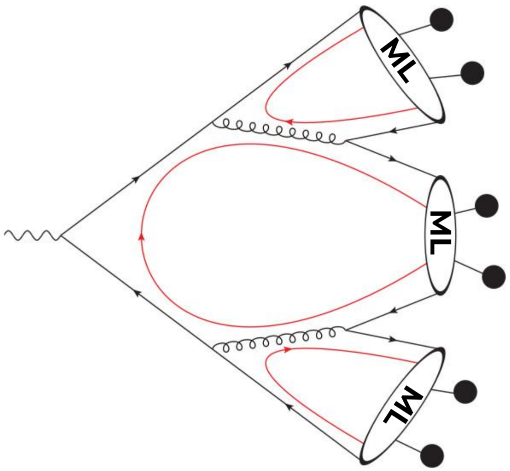


- QCD provide pre-confinement of colour
- Colour-singlet pair end up close in phase space and form highly excited hadronic states, the clusters
- Pre-confinement states that the spectra of clusters are independent of the hard process and energy of the collision
- Peaked at low mass (1-10 GeV) typically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

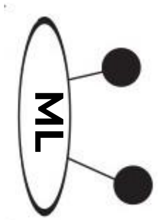
Cluster hadronization model

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

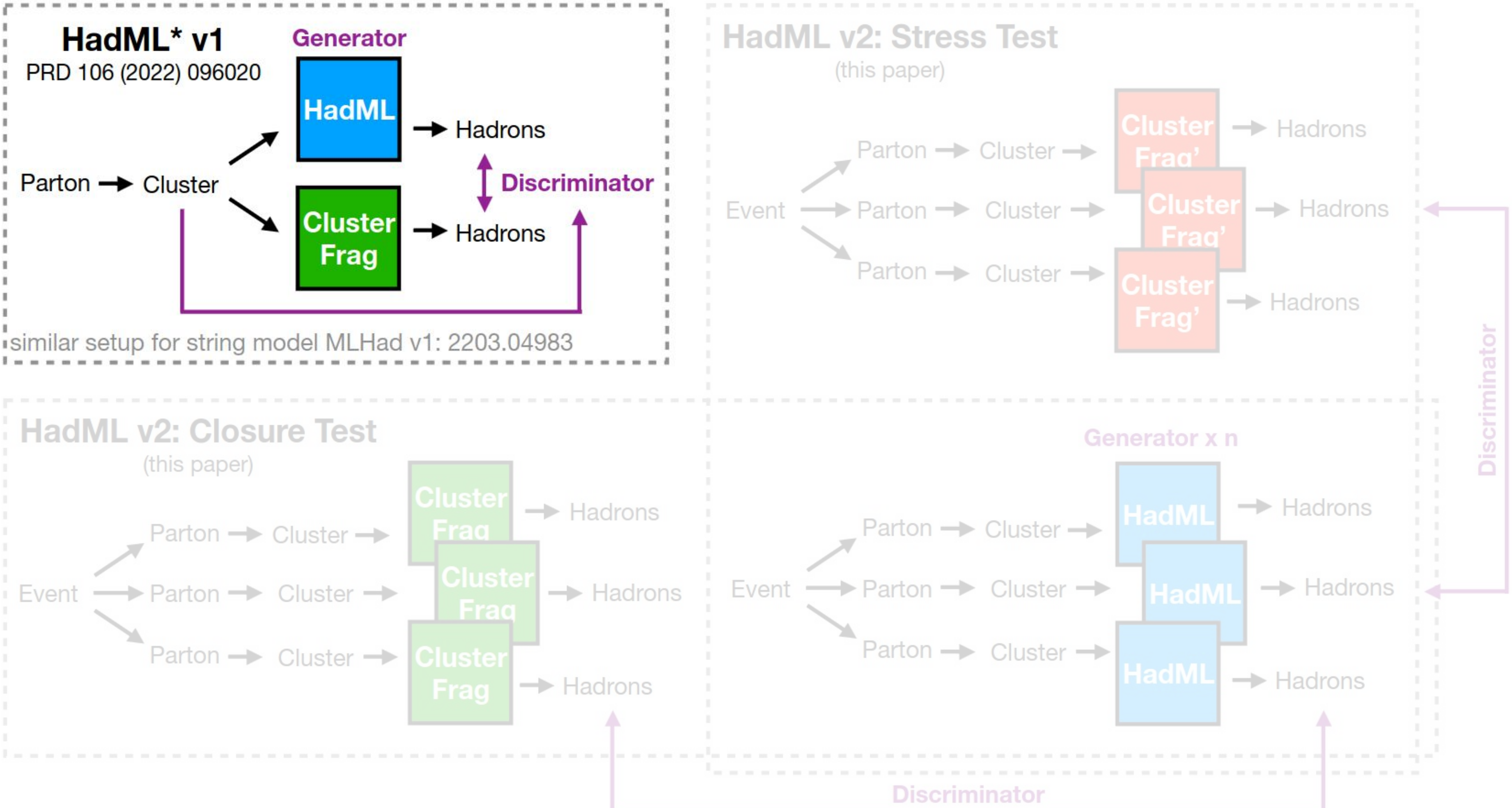
QCD **pre-confinement** discovered by Amati & Veneziano:



- 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:

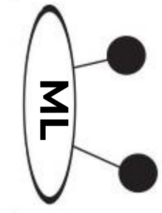


Road map for today



ML hadronization

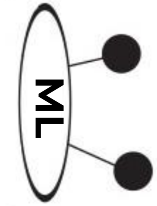
1st step: generate kinematics of a cluster decay to 2 hadrons



Towards a Deep Learning Model for Hadronization

ML hadronization

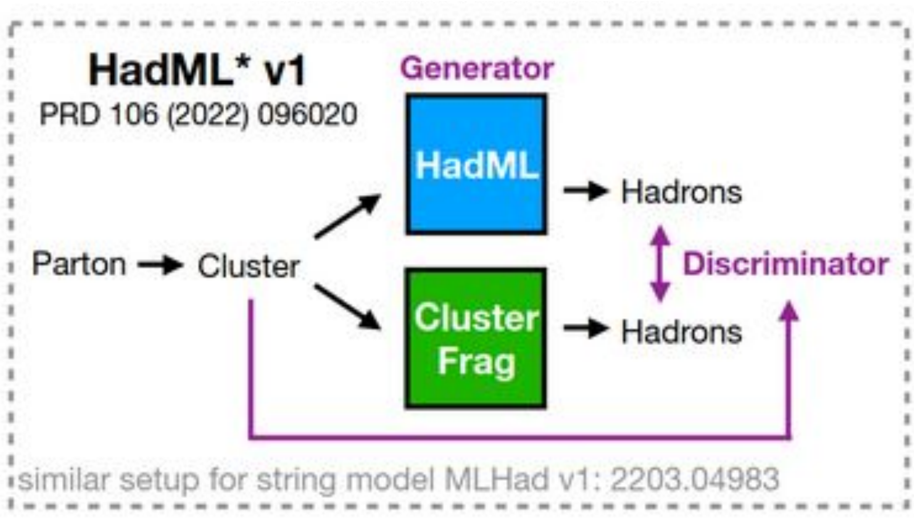
1st step: generate kinematics of a cluster decay to 2 hadrons



How?

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

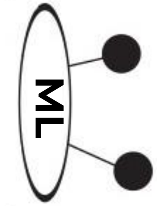
Generative Adversarial Net



Towards a Deep Learning Model for Hadronization

ML hadronization

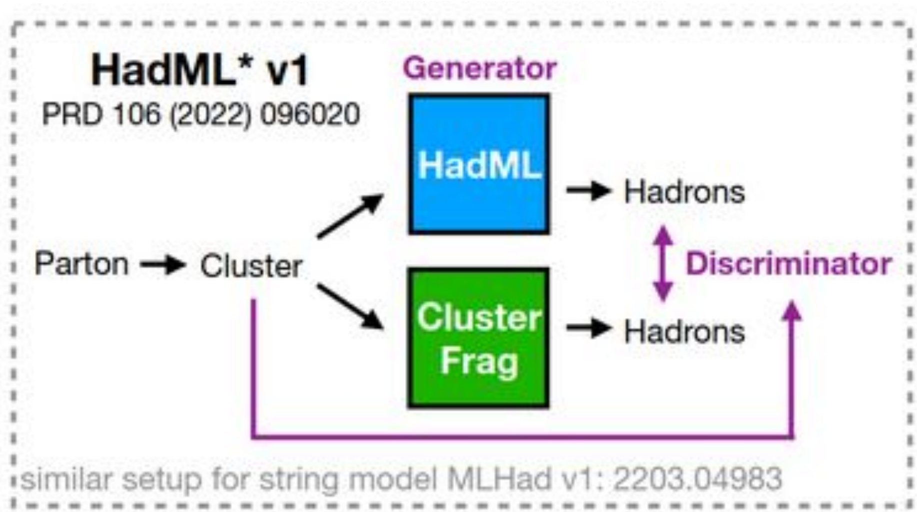
1st step: generate kinematics of a cluster decay to 2 hadrons



How?

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

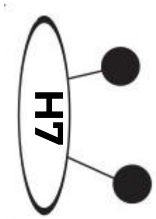
Generative Adversarial Net



Training data:

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

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



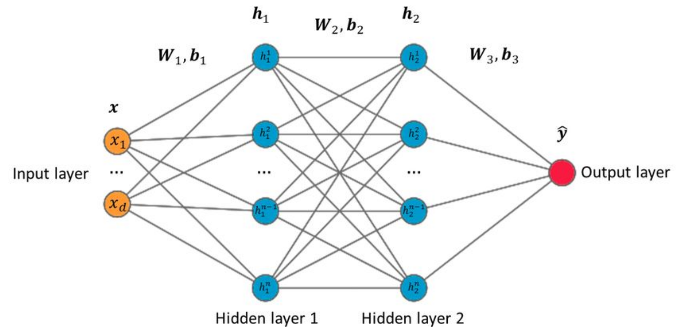
$\pi^0(E,\ p_x,\ p_y,\ p_z)$

$\pi^0(E,\ p_x,\ p_y,\ p_z)$

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

Architecture: conditional GAN

Generator and the Discriminator are composed of two-layer perceptron
(each a fully connected, hidden size 256, a batch normalization layer, LeakyReLU activation function)



Generator

Input

Cluster $(E,\ p_x,\ p_y,\ p_z)$ and 10 noise features sampled from a Gaussian distribution

Output (in the cluster frame)

$\left.\begin{array}{lll} \phi & - & \text{polar angle} \\ \theta & - & \text{azimuthal angle} \end{array}\right\}$ we reconstruct the four vectors of the two outgoing hadrons

Discriminator

Input

ϕ and θ labeled as signal (generated by Herwig) or background (generated by Generator)

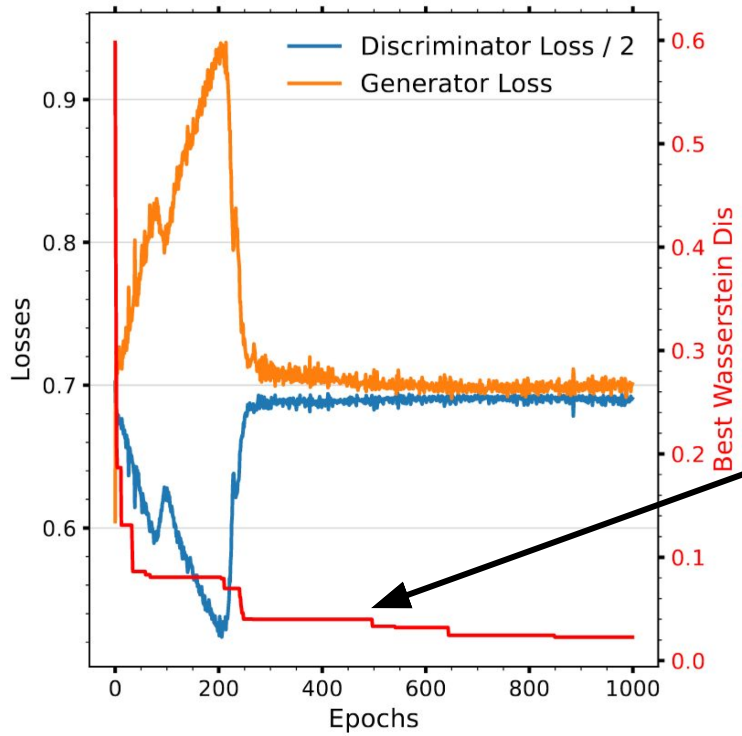
Output

Score that is higher for events from Herwig and lower for events from the Generator

Training

- **Data normalization:**
  cluster's four vector and angular variables are scaled to be between -1 and 1 (tanh activation function as the last layer of the Generator)

- **Discriminator** and the **Generator** are trained separately and alternately by two independent Adam optimizers with a learning rate of 10^{-4}, for 1000 epochs



This is a typical learning curve for GAN training

- **The best model** for events with partons of Pert = 0, is found at the epoch 849 with a total Wasserstein distance of 0.0228.

Integration into Herwig

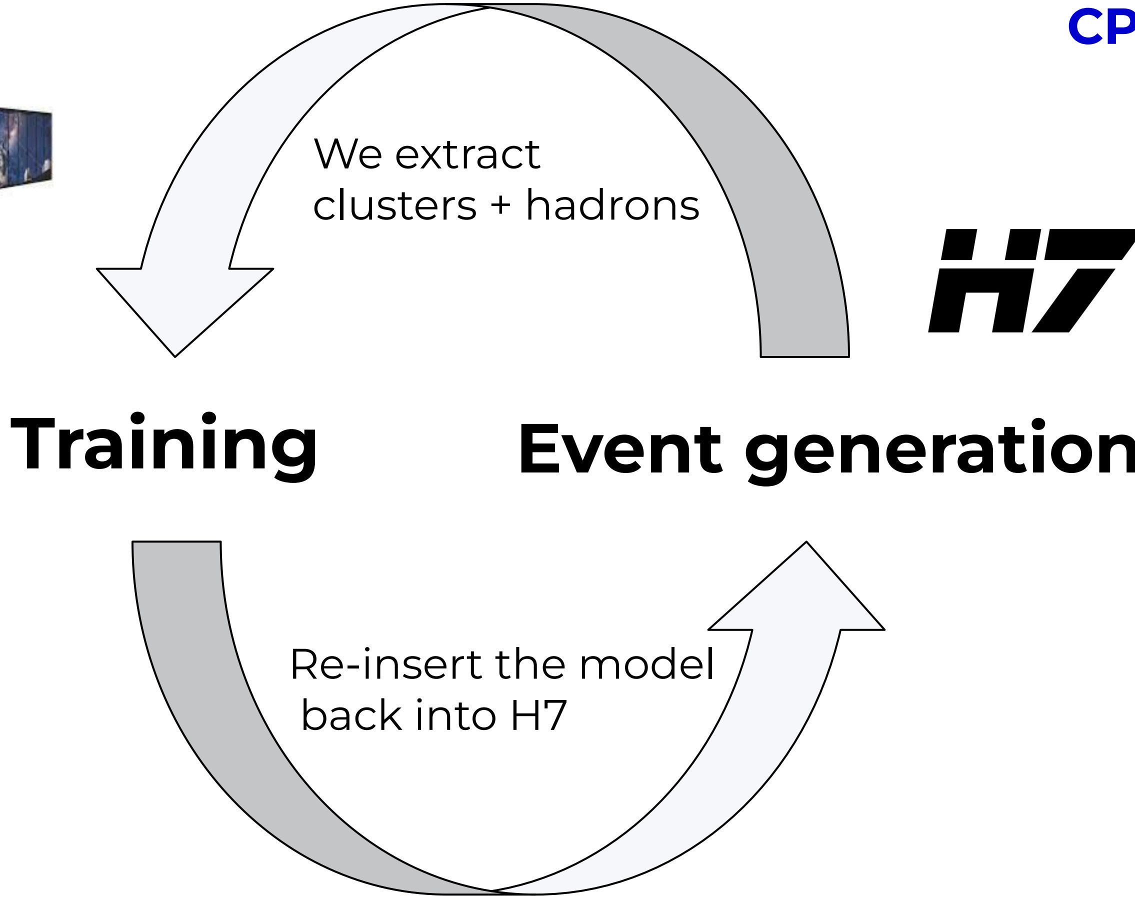
GPU

TensorFlow

python

CPU

H7

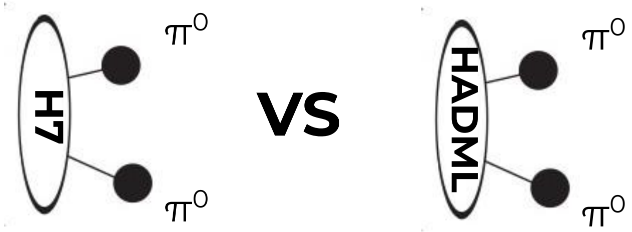


ONNX RUNTIME

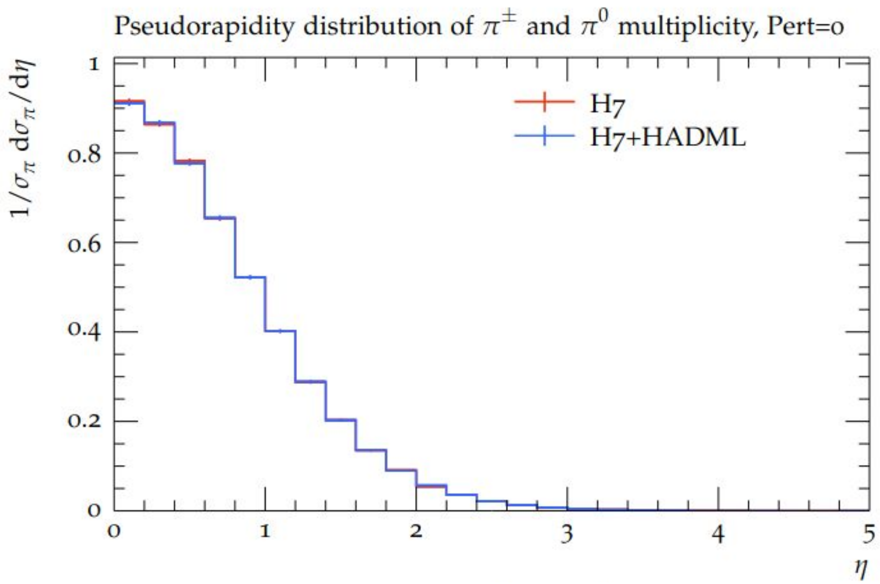
This then allows us to run a full event generator and produce plots

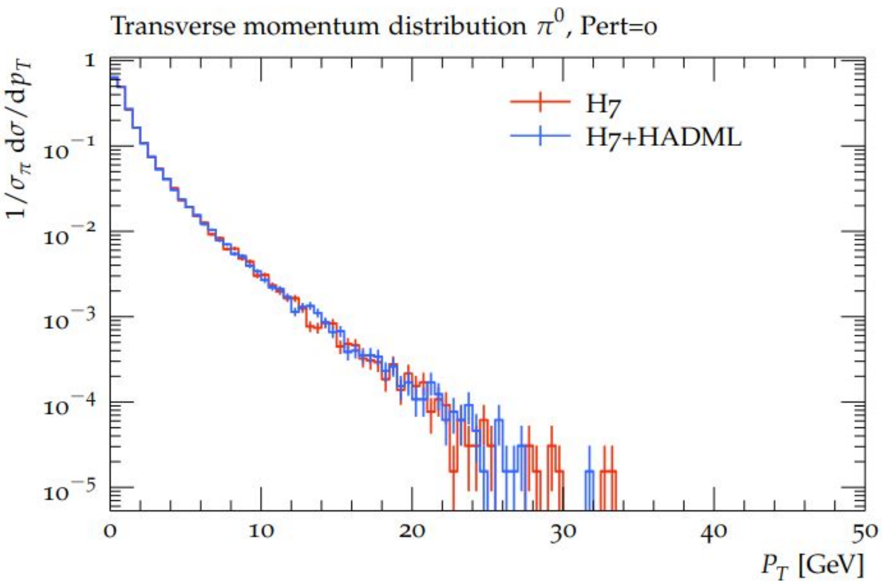
Low-level Validation
(similar to training data)

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



π^0 kinematic variables

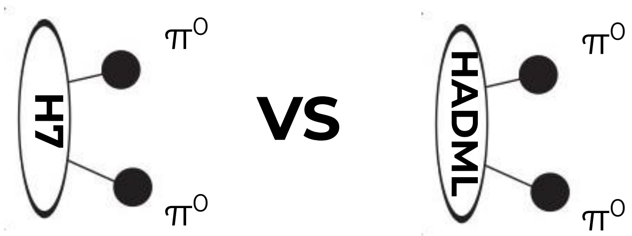




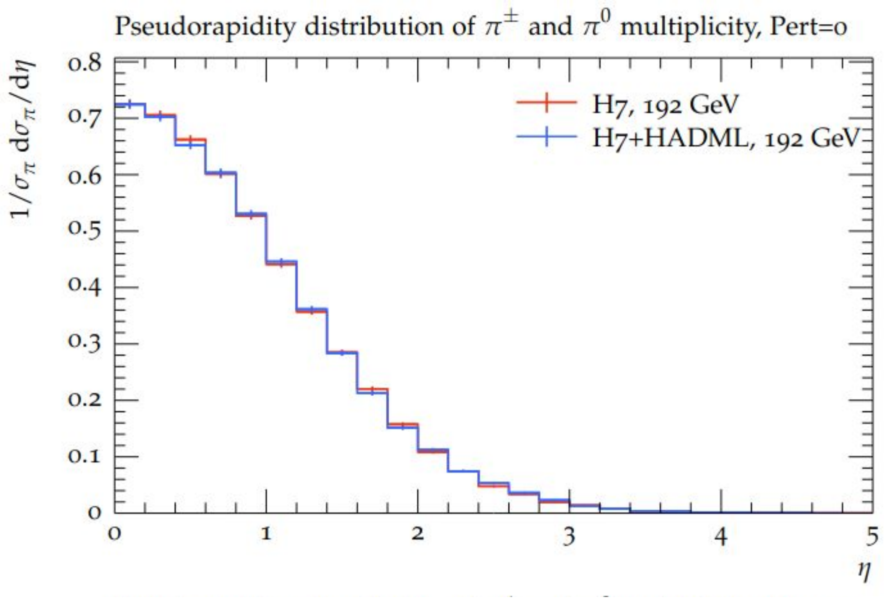
Performance: Energy of the collisions

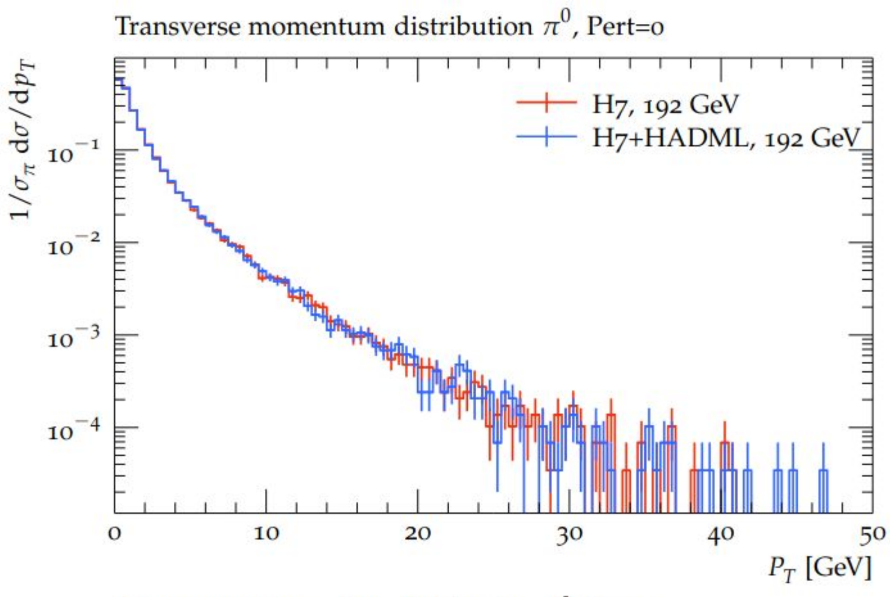
Low-level Validation
(**beyond** training data **different energy**)

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



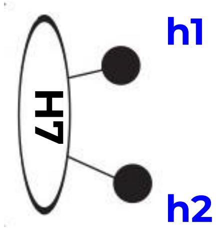
π^0 kinematic variables



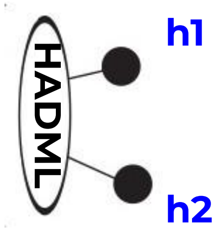


Low-level Validation
(**beyond** training data **different hadrons**)

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



VS

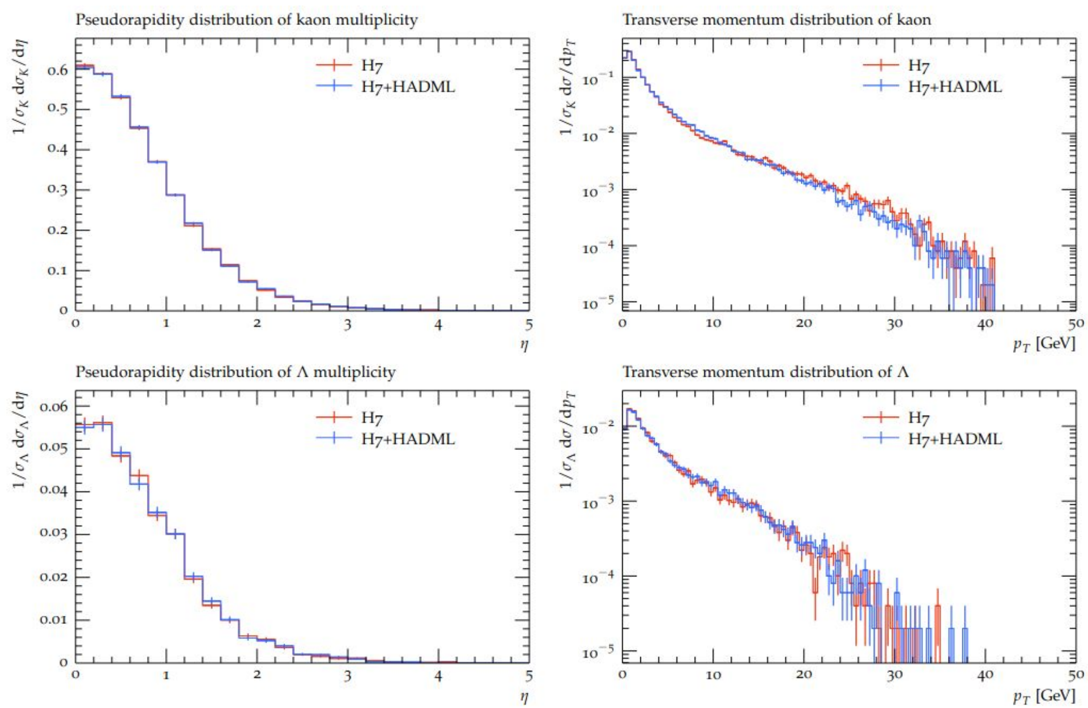


h kinematic variables

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

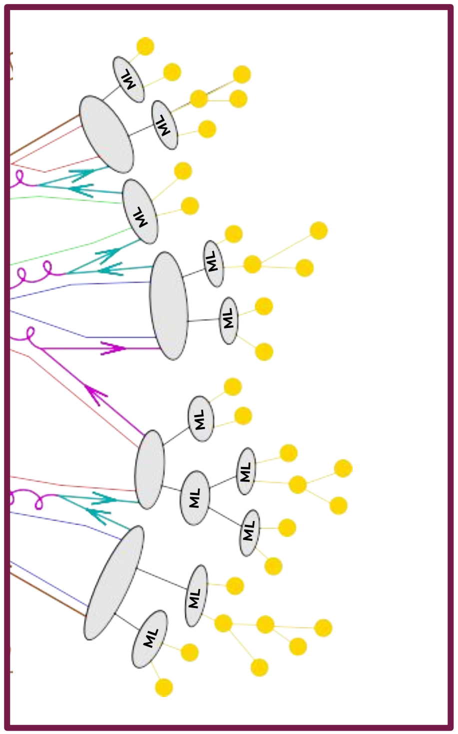
Kaons

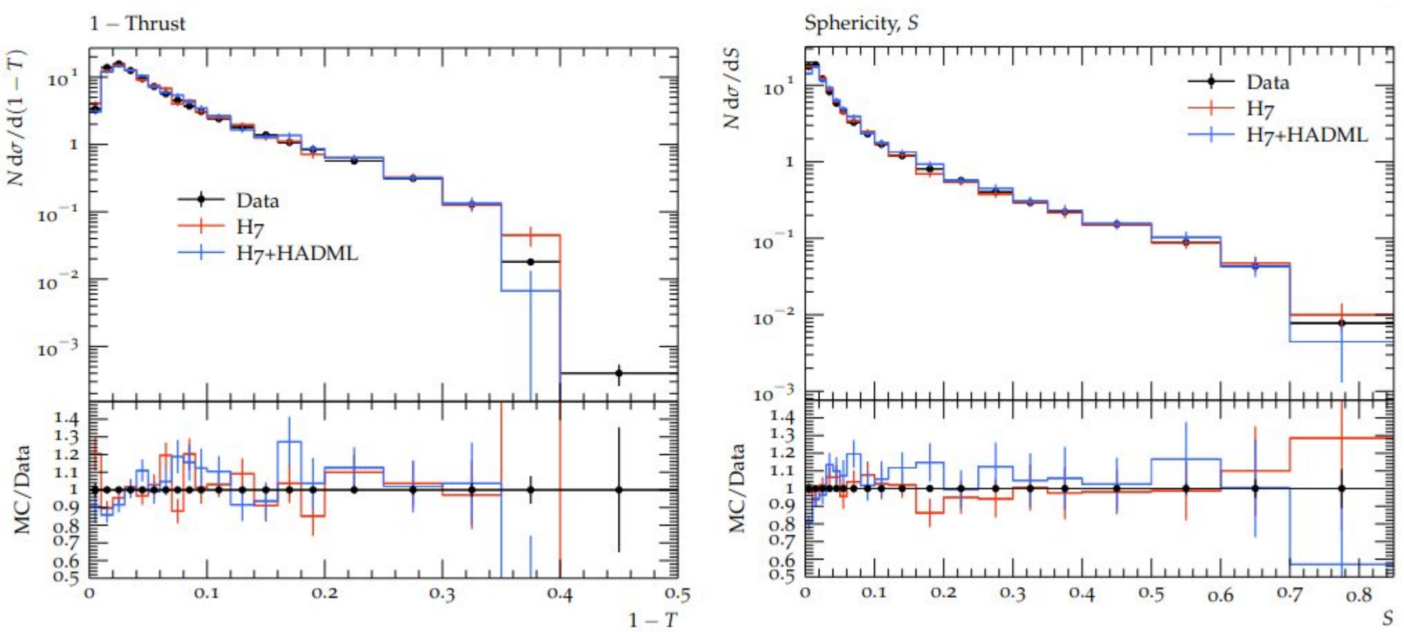
Lambda



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

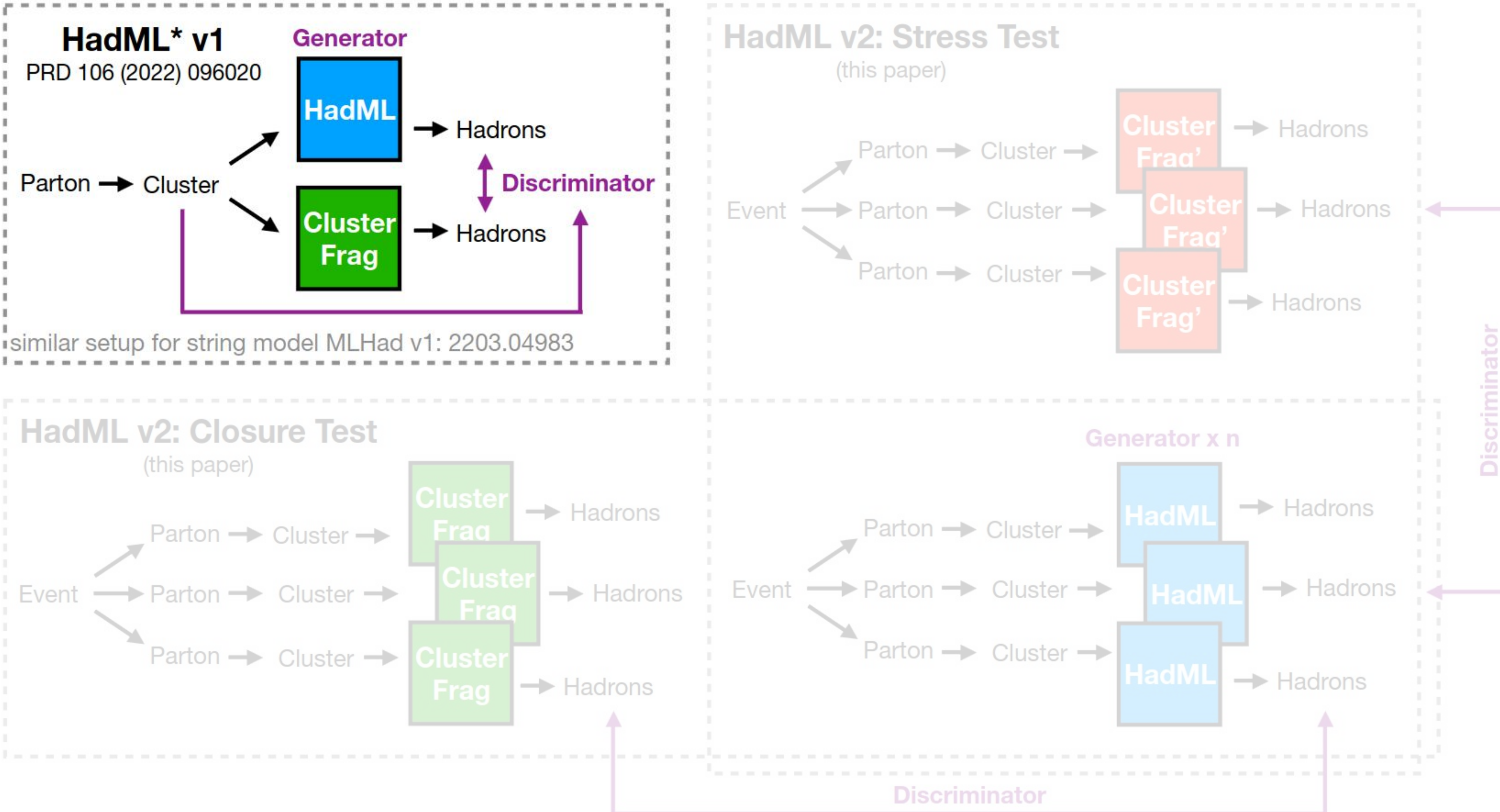
LEP DELPHI **Data**



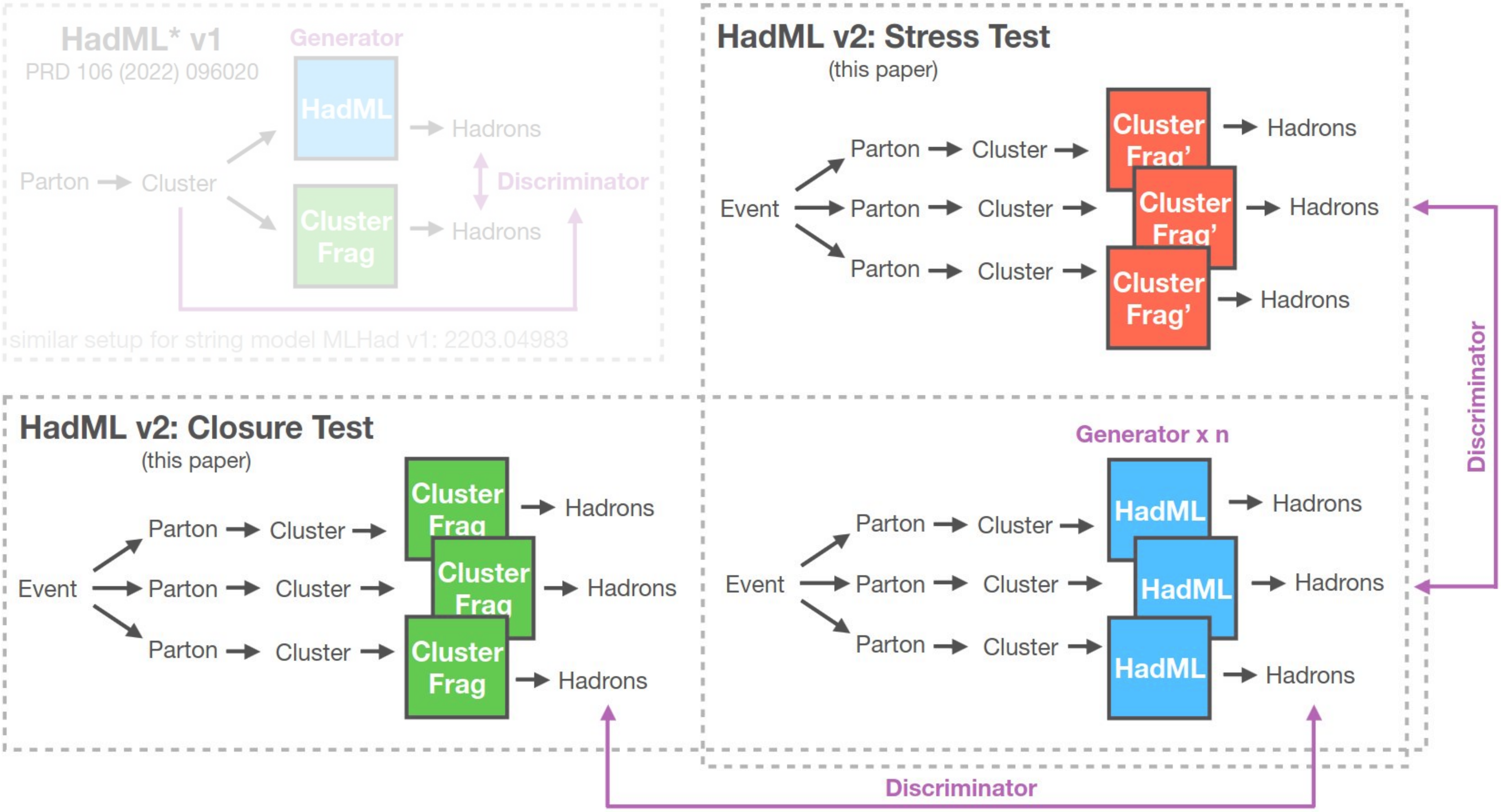


N.B. we have trained on H7, so we don’t expect to be any better than it at modeling the data.

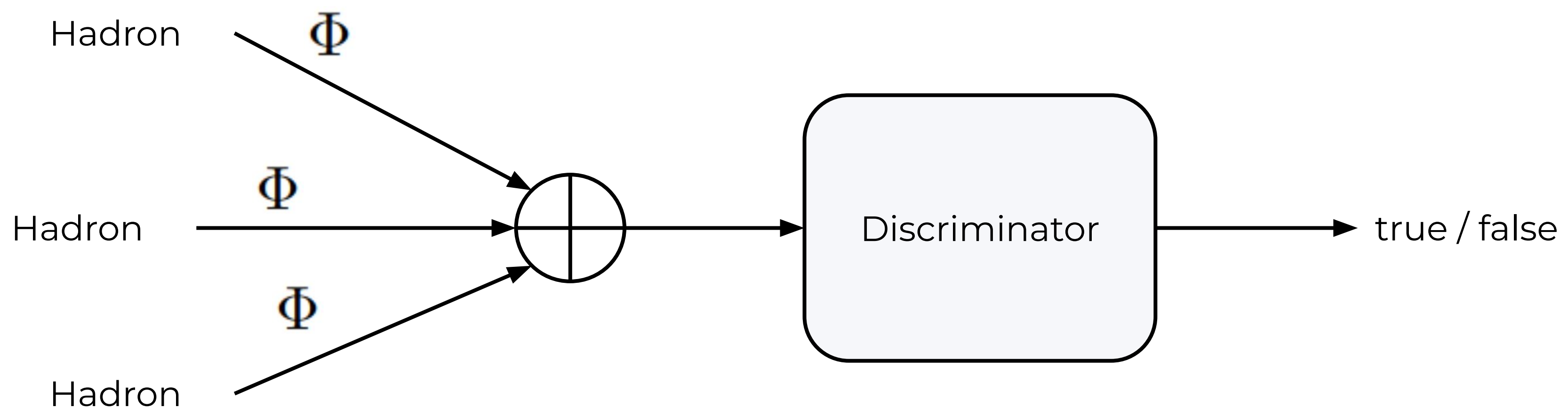
Road map for today



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.



The discriminator function is modified, we parameterize is as a Deep Sets model

$$D_E(x) = F\left(\frac{1}{n}\sum_{i=1}^{n} \Phi\left(h_i, \omega_{D_\Phi}\right), \omega_F\right)$$

invariant under permutations of hadrons

HadML v1

The loss function:

$$L = - \sum_{\lambda \sim \textsc{Herwig},\, z \sim p(z)} \left(\log\left(D\left(\tau\left(\lambda\right)\right)\right) + \log\left(1 - D\left(G\left(z, \lambda\right)\right)\right)\right)$$

HadML v2

The discriminator function is modified, we parameterize is as a Deep Sets model

$$D_E\left(x\right) = F\left(\frac{1}{n}\sum_{i=1}^{n}\Phi\left(h_i, \omega_{D_\Phi}\right), \omega_F\right)$$

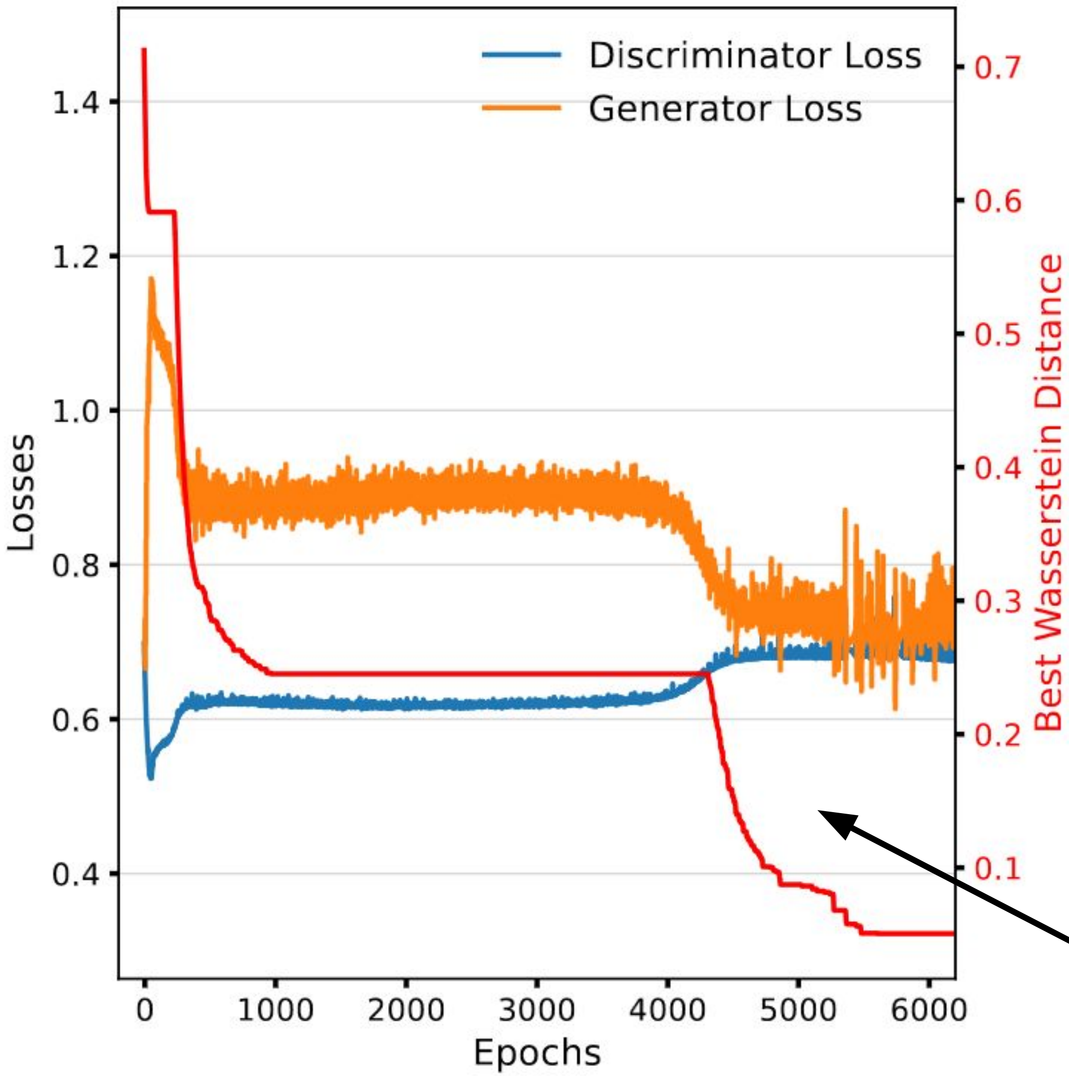
invariant under permutations of hadrons

Φ embeds a set of hadrons into a fixed-length latent space and F acts on the average

$$L = - \sum_{x \sim \text{data}} \log\left(D_E\left(x\right)\right) - \sum_{\{G\} \sim \textsc{Herwig},\, z \sim p(z)} \log\left(1 - D_E\left(\{G\left(z, \lambda\right)\}\right)\right)$$

The approach could also be used to fit (without binning) data to a parametric physics model (for example cluster) as well. However, this would require making the cluster model differentiable.

Training HADML v2

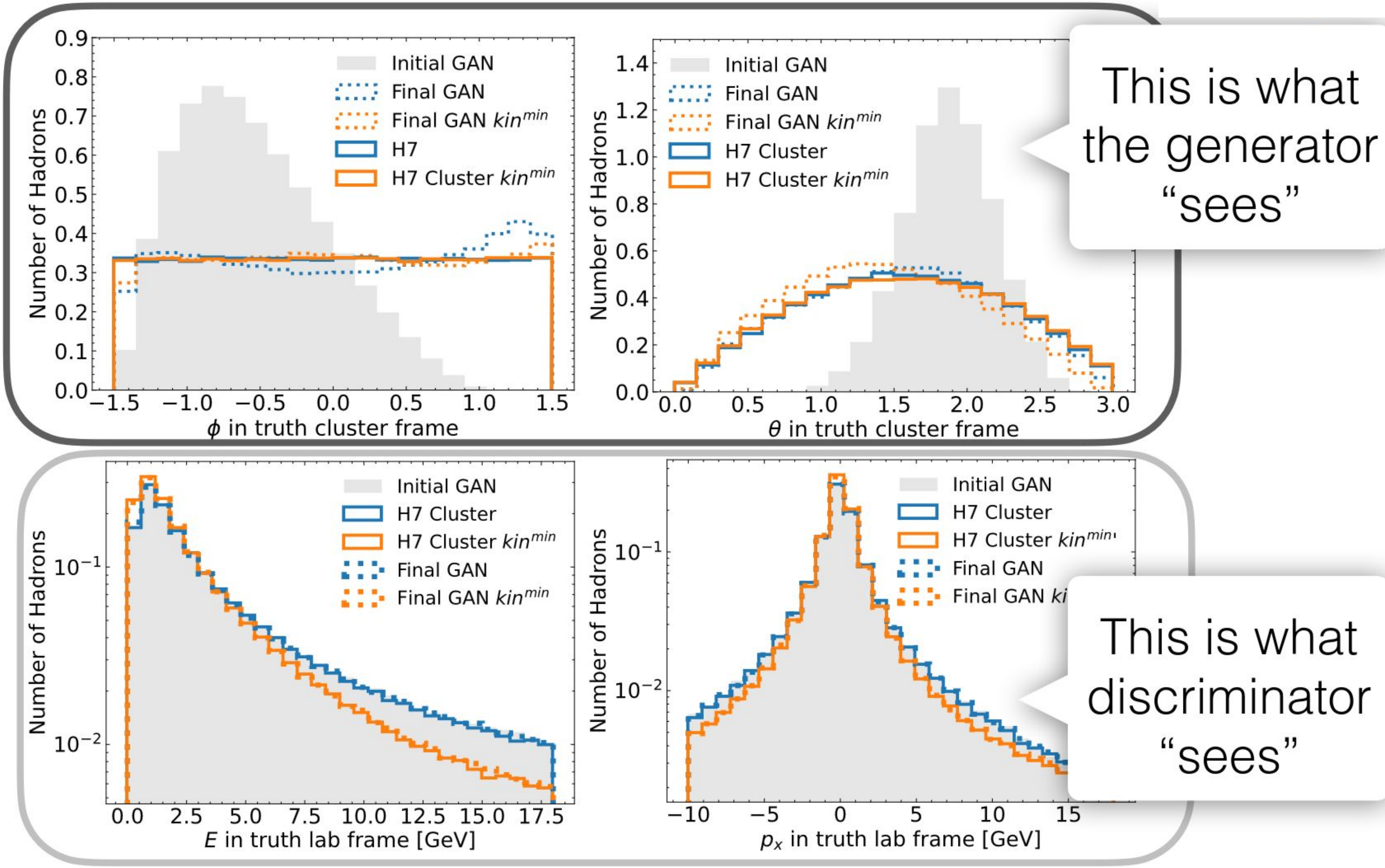


Now, the generator is local (per cluster), but the discriminator is global (whole event).

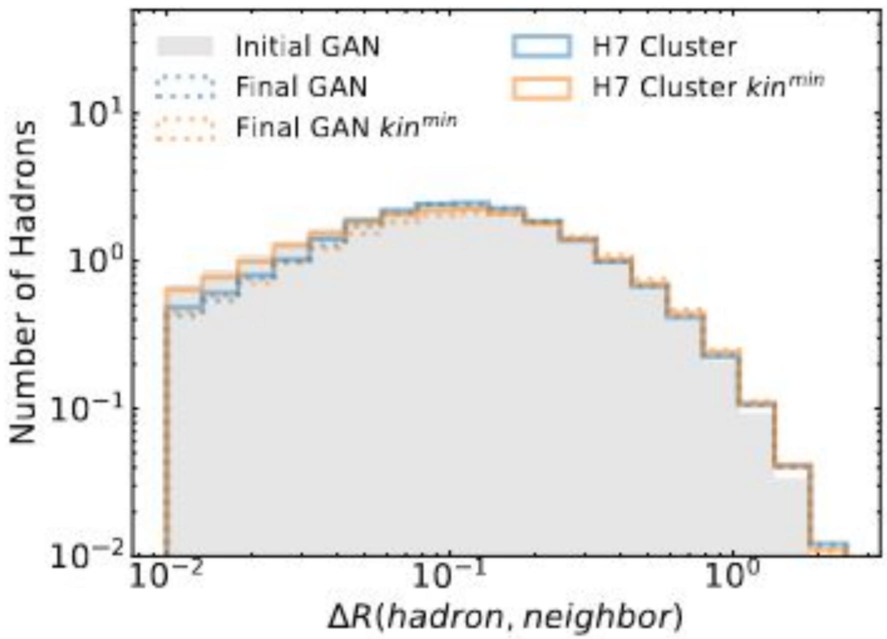
Discriminator is a permutation-invariant architecture called Deep Sets.

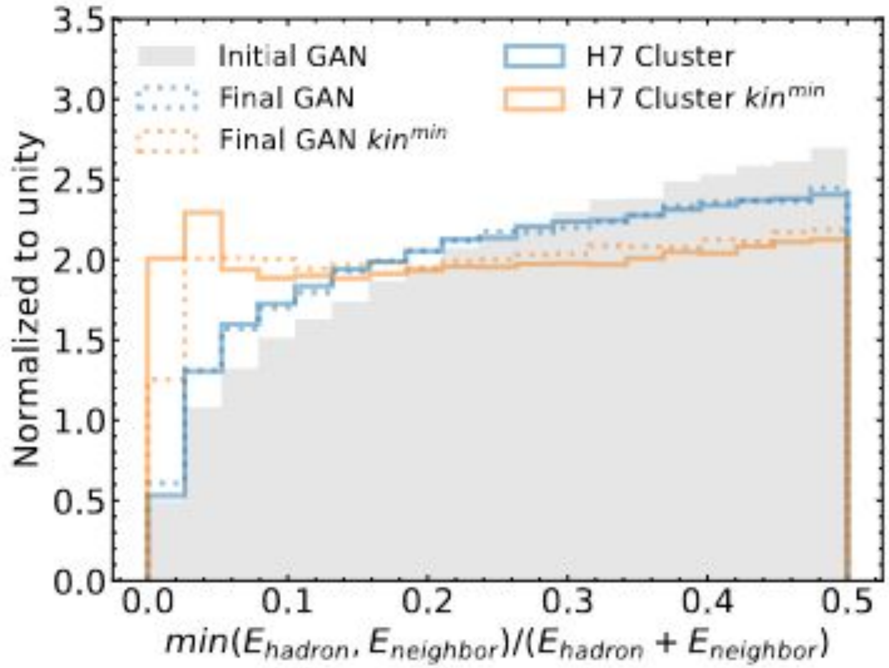
Simplification only Pions

Still works !


Initial GAN
Final GAN
Final GAN kin^{min}
H7
H7 Cluster kin^{min}
Number of Hadrons
ϕ in truth cluster frame
H7 Cluster
θ in truth cluster frame
This is what the generator “sees”
E in truth lab frame [GeV]
p_x in truth lab frame [GeV]
This is what discriminator “sees”

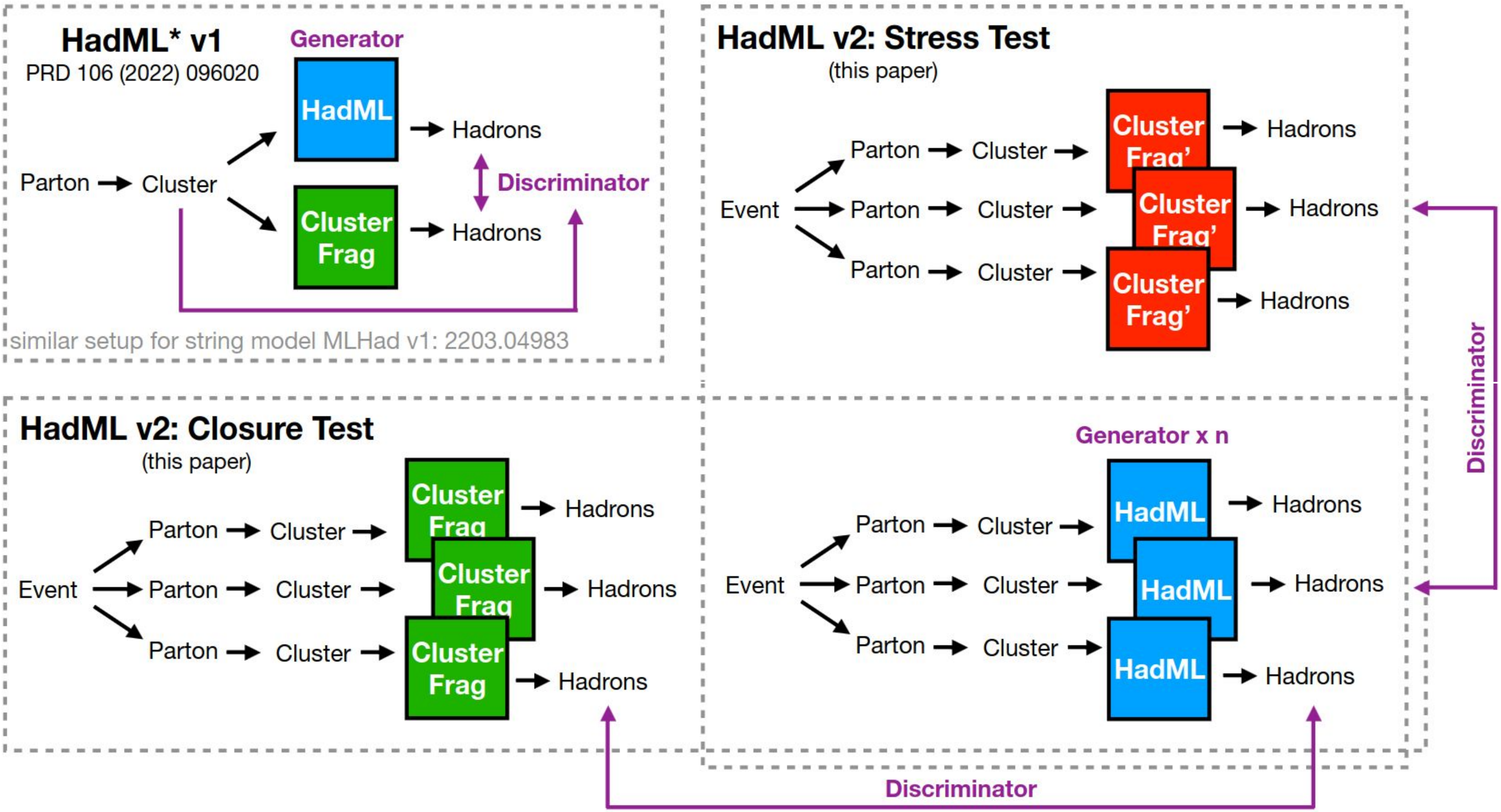
Performance: going beyond inputs and outputs





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

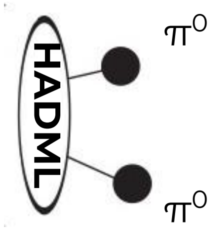
Summary



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

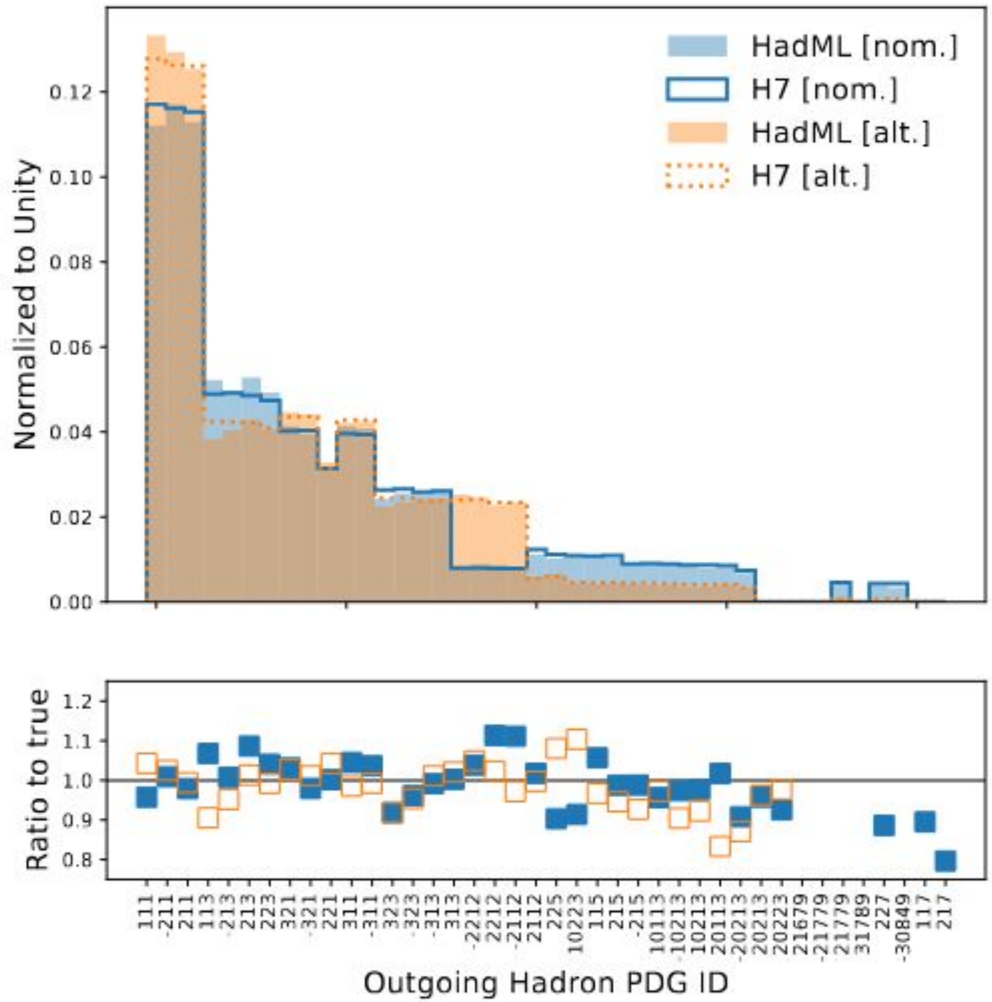
The approach could also be used to tune (without binning) data to a parametric physics model (for example cluster) as well. However, this would require making the cluster model differentiable.

- 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



[Chan, Ju, Kania, Nachman, **AS**, PRD (2025)]

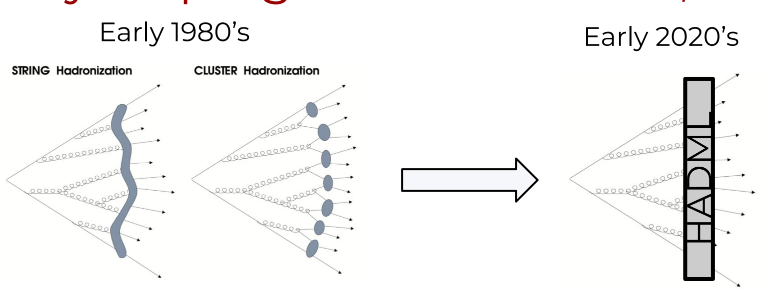
Outlook

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

What is next?

- Number of technical and methodological step needed:
  - ➔ Include heavy clusters (so far done by Herwig)
  - ➔ Hyperparameter optimization, including the investigation of alternative generative models
  - ➔ More flexible model with a capacity to mimic the cluster or string models and beyond.
  - ➔ MLhadML joining idea of MLhad reweighting and HadML fitting deep generative models [Heller, Ilten, Menzo, Mrenna, Nachman, AS, Szewc, Youssef, arxiv 2411.02194]
  - ➔ Tune to the LEP data

There is still a multi-year program ahead of us, so please stay tuned!



Backup

String Model in nutshell

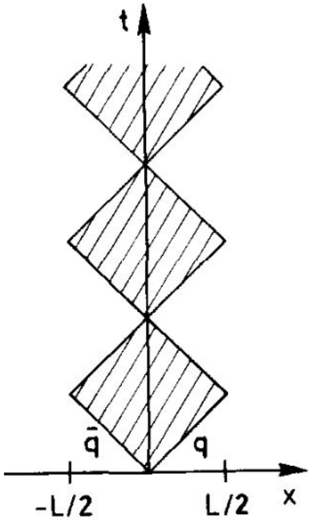
[Andersson, Gustafson, Ingelman, Sjostrand, Phys.Rept.97(1983)31]

String motion

From linear static potential $V(r) \approx \kappa r$ and linearity between space–time and energy–momentum:

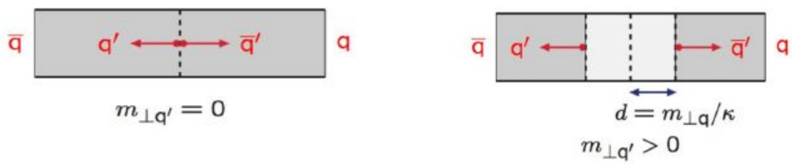
$$\left|\frac{\mathrm{d}E}{\mathrm{d}z}\right| = \left|\frac{\mathrm{d}p_z}{\mathrm{d}z}\right| = \left|\frac{\mathrm{d}E}{\mathrm{d}t}\right| = \left|\frac{\mathrm{d}p_z}{\mathrm{d}t}\right| = \kappa$$

We get a “YoYo“ state which we interpret as a meson.



String breakdowns

The quarks obtain a mass and a transverse momentum in the breakup through a tunneling mechanism



with a probability:

$$\mathcal{P} \propto \exp\left(-\frac{\pi m_{\perp q}^2}{\kappa}\right) = \exp\left(-\frac{\pi p_{\perp q}^2}{\kappa}\right) \exp\left(-\frac{\pi m_q^2}{\kappa}\right)$$

- Suppression of heavy quarks: uu : dd : ss : cc ≈ 1 : 1 : 0.3 : 10^{-11}
- Common Gaussian pT spectrum, <pT>~ 0.4 GeV
- Diquark (qq - q̄q̄ breakups) ~ antiquark ⇒ simple model for baryon production.

Iterative process (left-right symmetry) leads to distribution of momentum fraction taken by each hadron as:

$$f(z) \propto \frac{(1-z)^a}{z} \exp\left(-\frac{bm^2}{z}\right)$$

Summary:

String model has very good energy-momentum picture however it is unpredictive in understanding of hadron mass effects ⇒ many parameters, 10-30 depending on how you count.

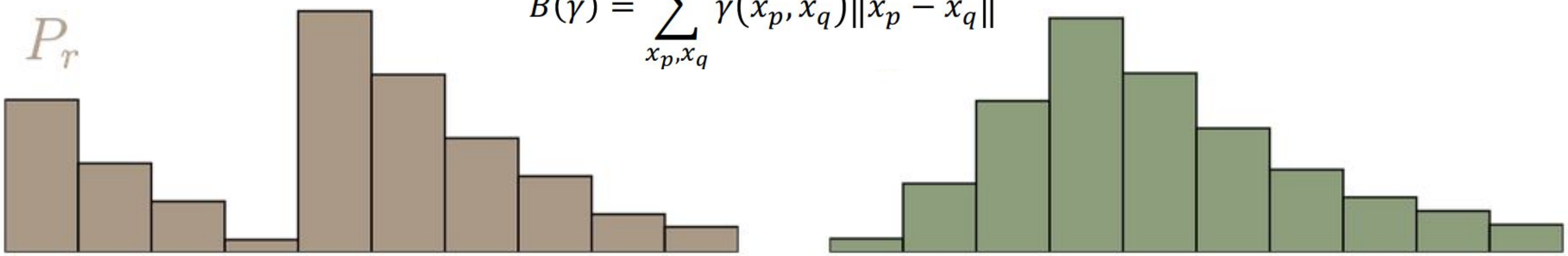
The Wasserstein distance

- For discrete probability distributions, the Wasserstein distance is called the earth mover's distance (EMD):
- EMD is the minimal total amount of work it takes to transform one heap into the other.

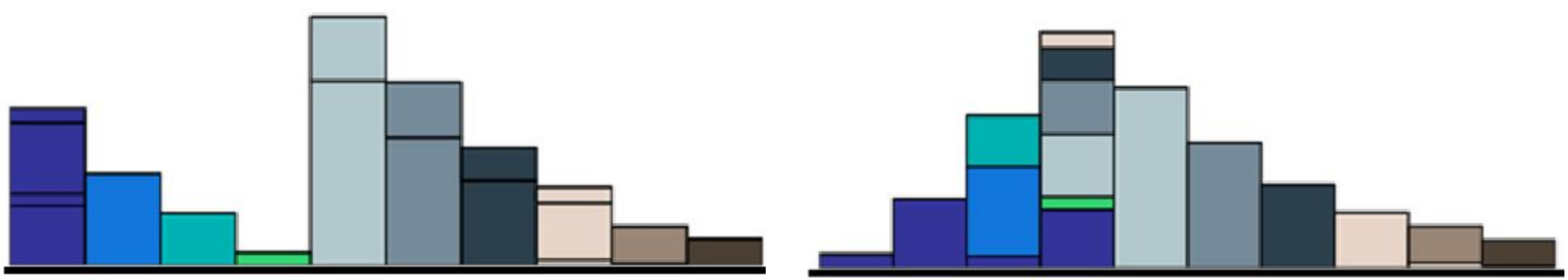
$$W(P, Q) = \min_{\gamma \in \Pi} B(\gamma)$$

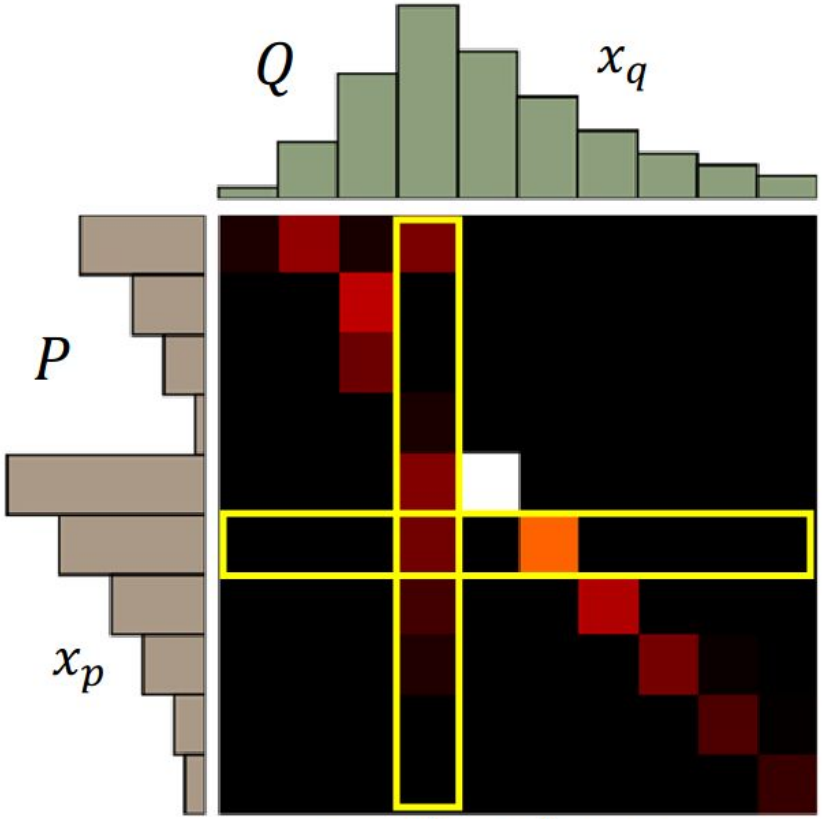
- Work is defined as the amount of earth in a chunk times the distance it was moved.

$$B(\gamma) = \sum_{x_p, x_q} \gamma(x_p, x_q) \|x_p - x_q\|$$



Best "moving plans" of this example





moving plan γ

All possible plan Π

A "moving plan" is a matrix

The value of the element is the amount of earth from one position to another.

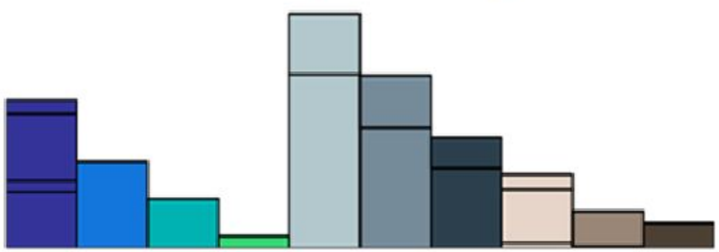
Average distance of a plan γ:

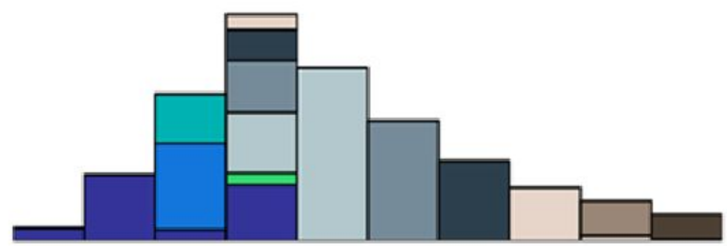
$$B(\gamma) = \sum_{x_p, x_q} \gamma(x_p, x_q) \| x_p - x_q \|$$

Earth Mover's Distance:

$$W(P, Q) = \min_{\gamma \in \Pi} B(\gamma)$$

The best plan





Minimax Loss

In the paper that introduced GANs, the generator tries to minimize the following function while the discriminator tries to maximize it:

$$E_x[log(D(x))] + E_z[log(1 - D(G(z)))]$$

In this function:

- `D(x)` is the discriminator's estimate of the probability that real data instance x is real.
- E_x is the expected value over all real data instances.
- `G(z)` is the generator's output when given noise z.
- `D(G(z))` is the discriminator's estimate of the probability that a fake instance is real.
- E_z is the expected value over all random inputs to the generator (in effect, the expected value over all generated fake instances G(z)).
- The formula derives from the cross-entropy between the real and generated distributions.

The generator can't directly affect the `log(D(x))` term in the function, so, for the generator, minimizing the loss is equivalent to minimizing `log(1 - D(G(z)))`.

2.2 Machine Learning Implementation

The generator (G) and discriminator (D) functions are both parametrized as neural networks. Each of them is a fully connected network with four hidden layers, each with a width of 1,000 neurons. All intermediate layers in these networks use a LeakyReLU [24] activation function.

The non-discrete conditional inputs of G are normalized to the range of $(-1,\ 1)$, whereas the noise prior p is a Gaussian distribution with a mean of 0 and width of 1. The noise dimension N_z is set to 64. The last layer of G is divided into the variables $\kappa \in \mathbb{R}^2$, which correspond to hadron kinematics, and the variables $\pi \in \mathbb{R}^{2N_t}$, which correspond to hadron types. Here, N_t is the number of hadron types considered. For simplicity, we consider only the 40 most common hadron types (i.e. $N_t = 40$)[‡]. The hadron kinematics θ_{h_1} and ϕ_{h_1} are extracted from κ with a tanh activation function, as in the previous work [15]. The hadron type, on the other hand, is a categorical variable. In order to avoid zero gradients when using *argmax* in training, we use the Gumbel-Softmax [25] distribution to approximate the distribution of hadron types:

$$y_i = \frac{\exp\left(\left(\log \pi_i + g_i\right)/\tau\right)}{\sum_i \exp\left(\left(\log \pi_i + g_i\right)/\tau\right)}, \tag{2.4}$$

where g_i are independent and identically distributed samples drawn from Gumbel(0, 1). τ is a temperature parameter and as it approaches 0 the Gumbel-Softmax distribution becomes identical to the categorical distribution. We anneal τ by linearly decreasing it from 1.0 to 0.1 during training. The hadron type distributions y from the Gumbel-Softmax distribution are then taken as the inputs for D, in addition to the hadron kinematics θ_{h_1} and ϕ_{h_1}. During inference, the generated hadron types are obtained from the Gumbel-Softmax distribution with the *argmax* operation. The last layer of D uses a sigmoid activation function.

All neural networks are implemented and trained using PyTorch [26]. The generator and discriminator are optimized alternately with Adam [27] with a learning rate of 3×10^{-4} for both networks. The training uses a batch size of 40,000 and is performed for 25 epochs. The hyperparameters are optimized with Weights and Biases [28].