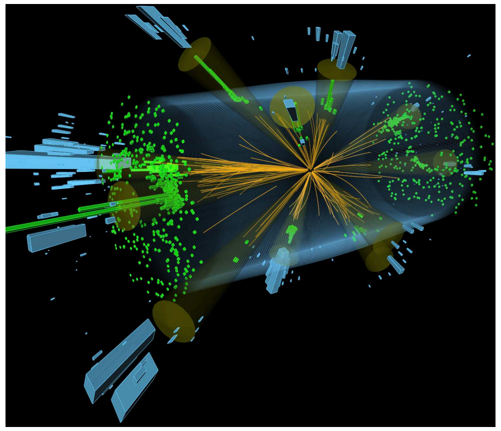
The Metric Space of Collider Events

Patrick T. Komiske III
Massachusetts Institute of Technology
Center for Theoretical Physics

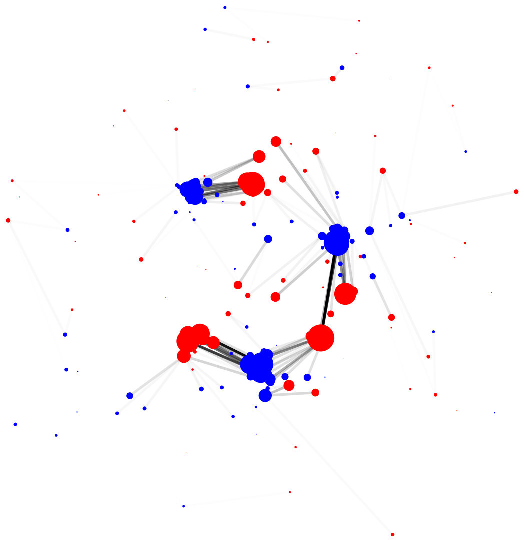
Deep Learning in the Natural Sciences
Hamburg, Germany
March 1, 2019

Collaborators: Eric Metodiev and Jesse Thaler

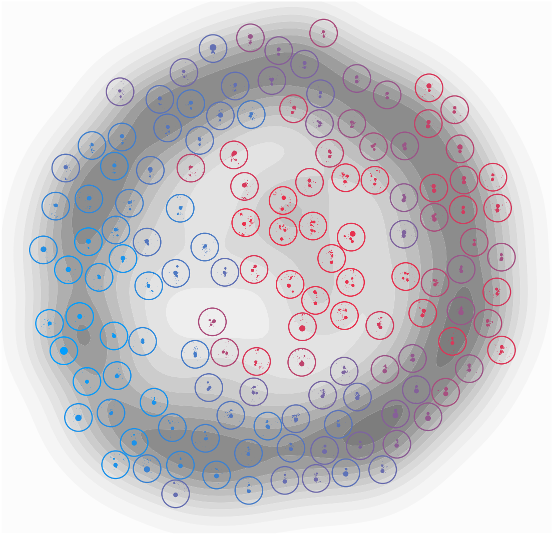
1902.02346



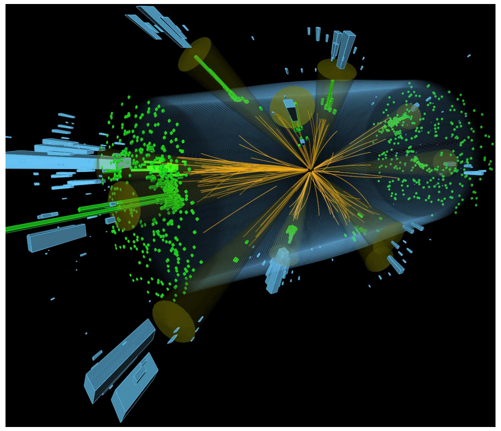
The Space of Collider Events



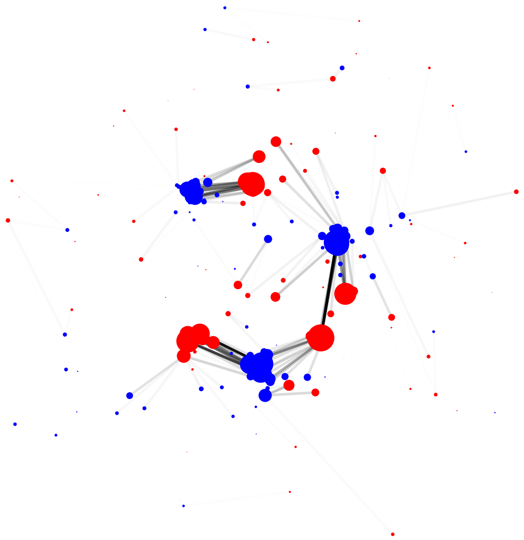
The Energy Mover's Distance



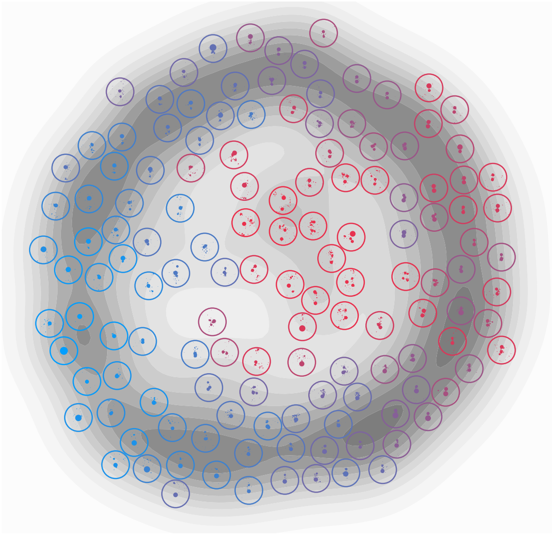
Particle Physics Applications



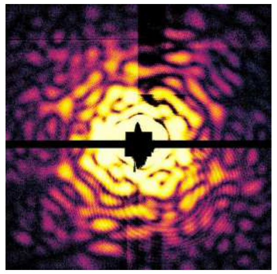
The Space of Collider Events



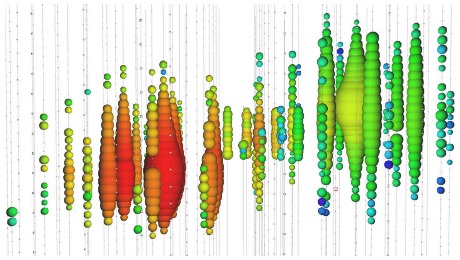
The Energy Mover’s Distance



Particle Physics Applications

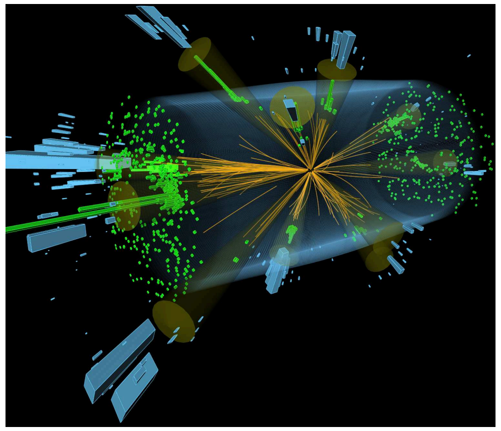


?

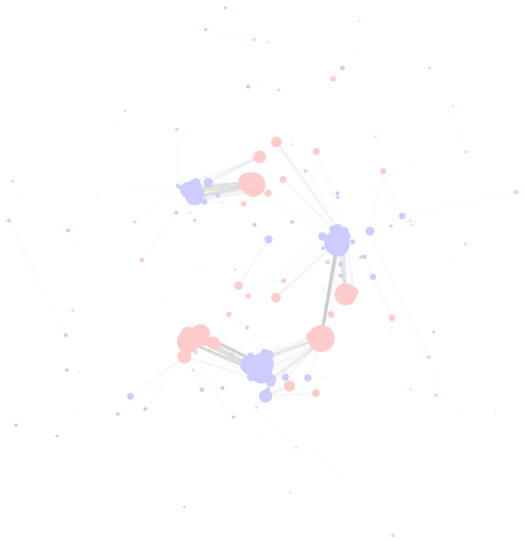


?

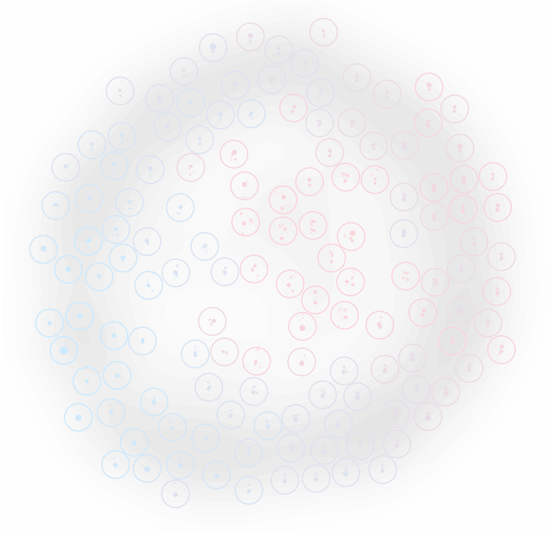
[Gorkhover, et al], [IceCube]



The Space of Collider Events



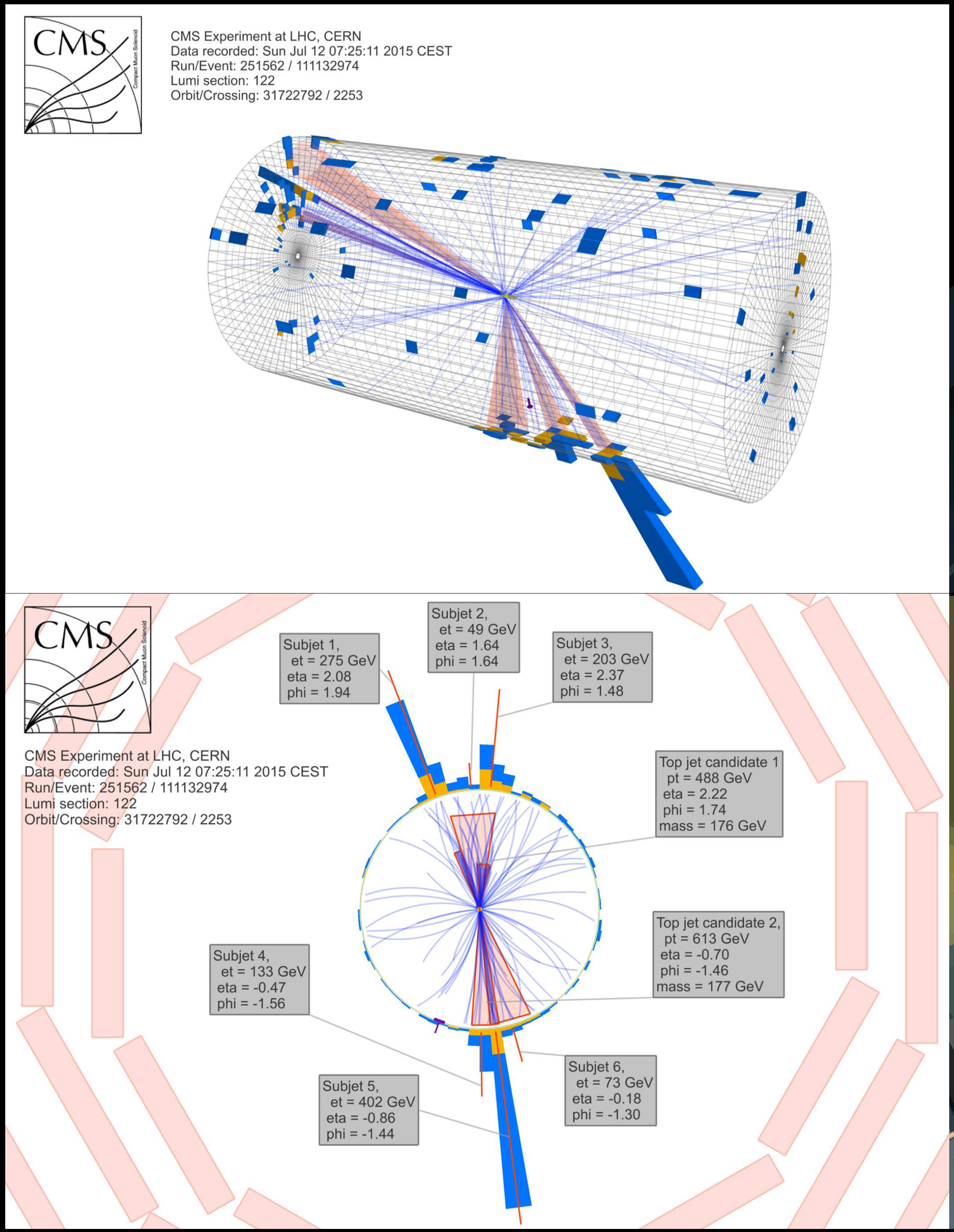
The Energy Mover's Distance



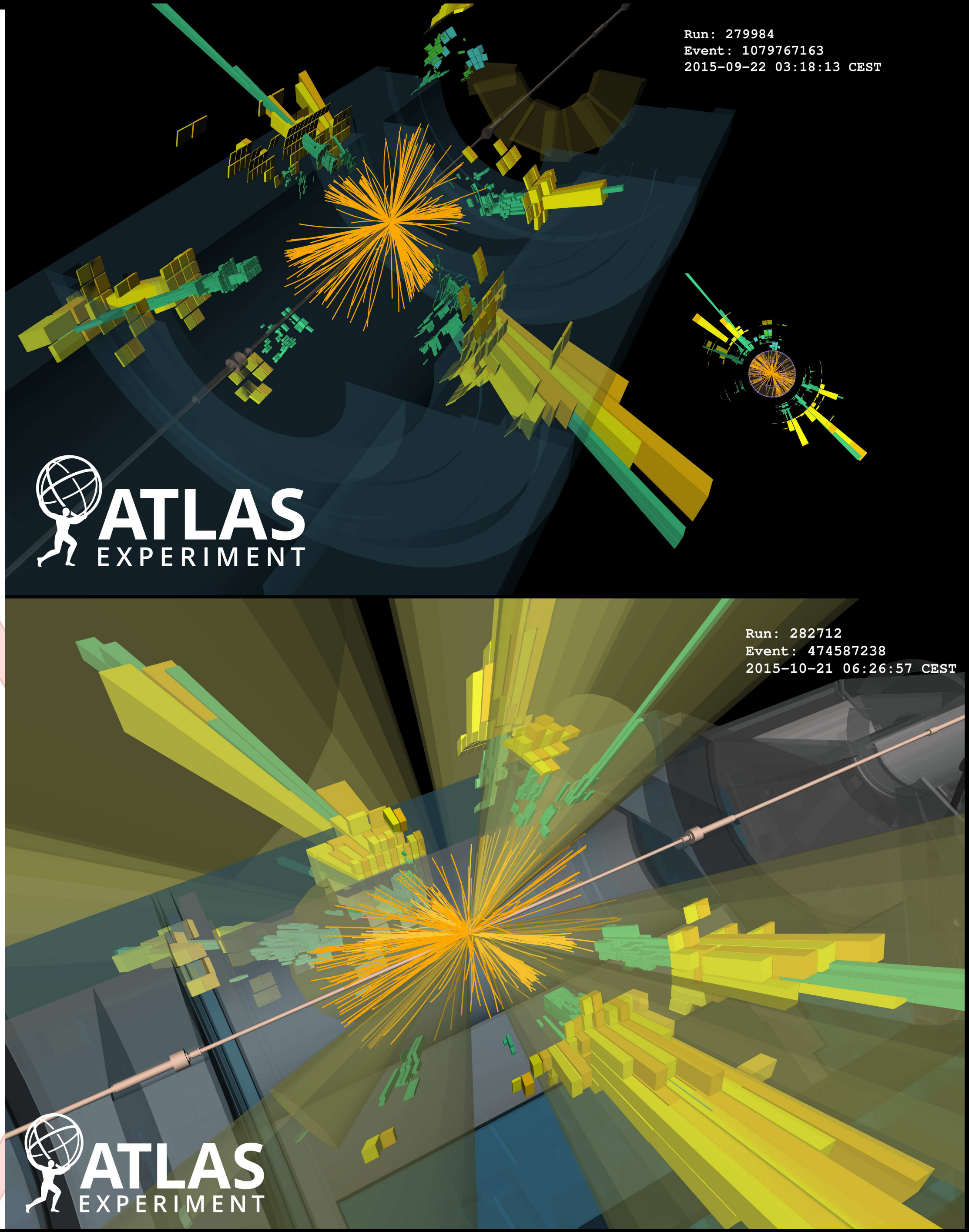
Particle Physics Applications

Fascinating Event Topologies at the LHC

New physics searches involve complicated final states including jets (collimated sprays of hadrons)



CMS hadronic $t\bar{t}$ event



ATLAS high jet multiplicity events

Jet Formation in Theory

Hard collision

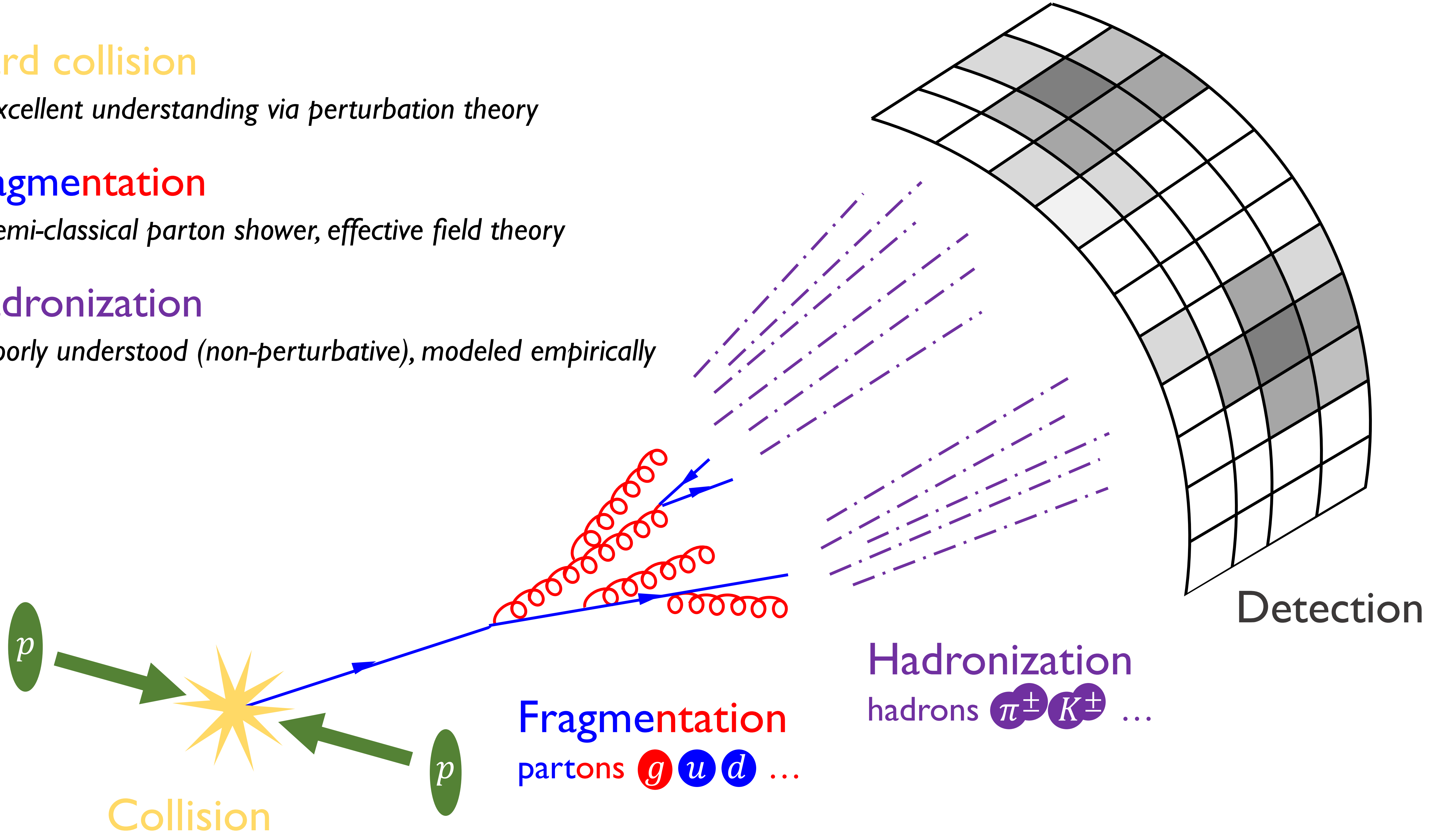
Excellent understanding via perturbation theory

Fragmentation

Semi-classical parton shower, effective field theory

Hadronization

Poorly understood (non-perturbative), modeled empirically

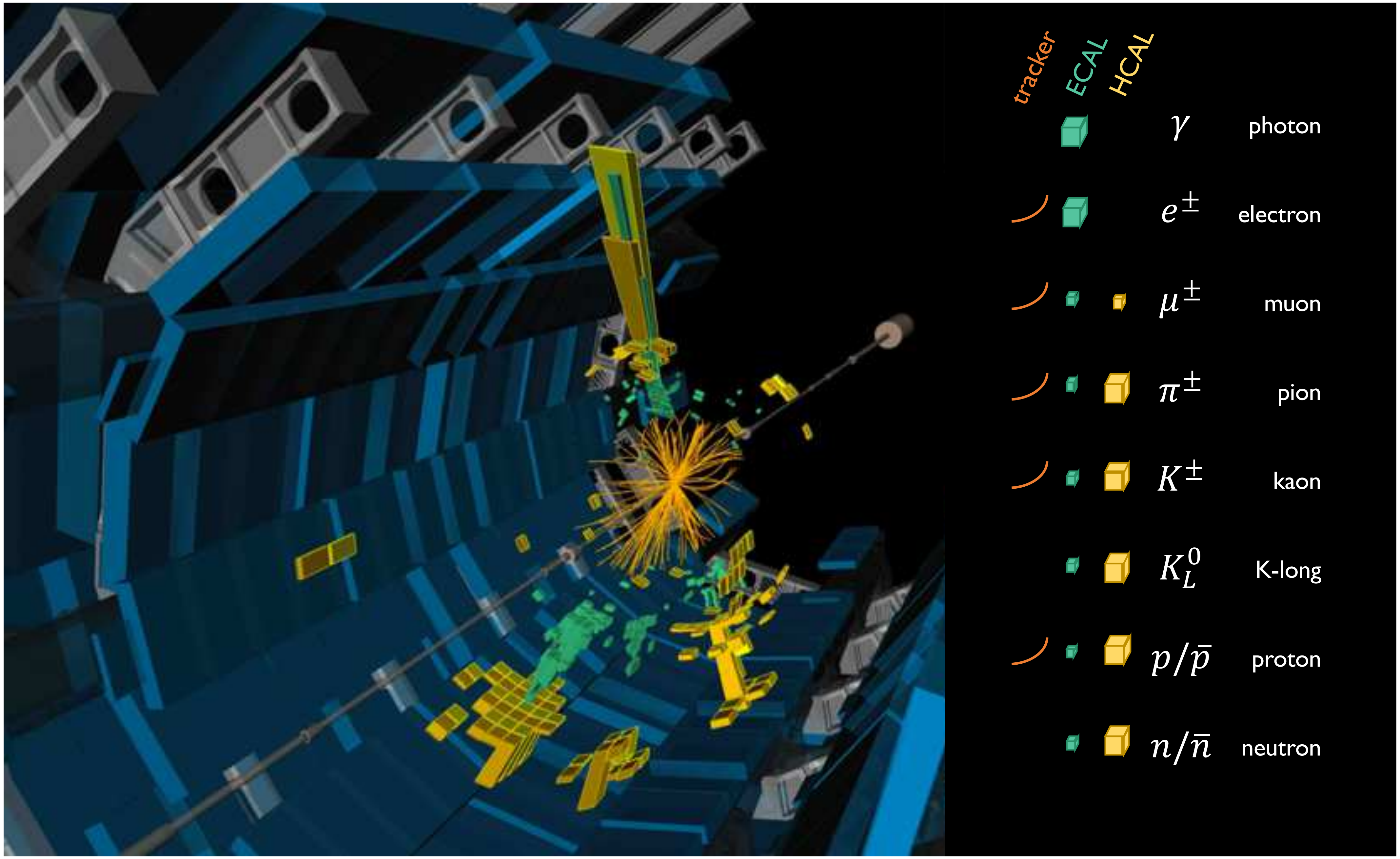


Cartoon of jet formation as a multi-scale process

Diagram by Eric Metodiev

Jet Detection in Experiment

Reconstruct event by synthesizing information from many detector systems



What information is both theoretically and experimentally robust?

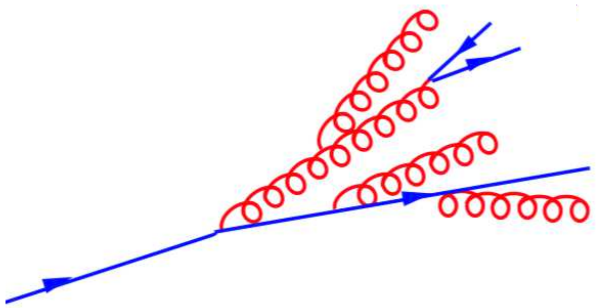
Diagram by Eric Metodiev

Energy Flow

Events, Theoretically

$$|\mathcal{E}\rangle = |(p_1^\mu, \vec{q}_1); (p_2^\mu, \vec{q}_2); \ldots\rangle$$

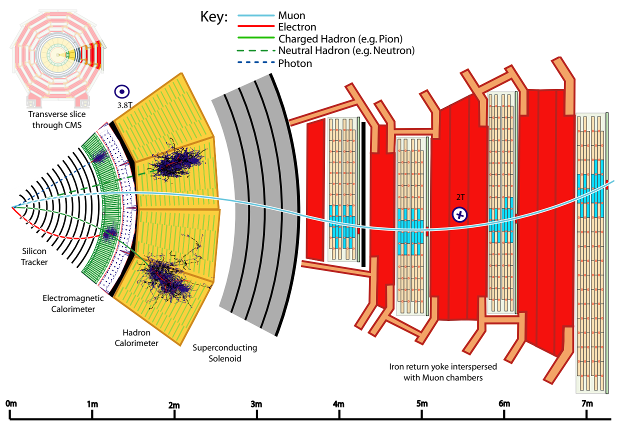
quantum state?



parton branching history?

Events, Experimentally

O(10 million) electrical signals?



set PF candidates?

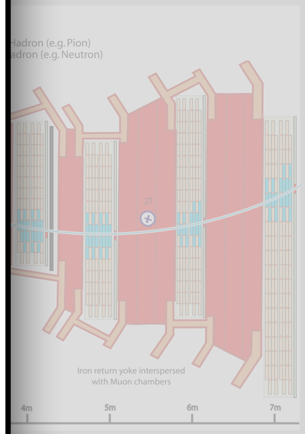
The energy flow (distribution of energy) is robust to fragmentation, hadronization, detector effects

$$\sum_{i=1}^{M} \underset{\substack{\uparrow\\ \text{Energy}}}{E_i}\,\delta(\underset{\substack{\uparrow\\ \text{Direction}}}{\hat{p}_i})$$

Energy Flow ⟷ Infrared and Collinear Safe Information

Energy Flow

Events, Theoretically | Events, Experimentally



didates?

When are collider events similar?

💡 Use a distance metric! 💡

Symmetric, non-negative, pairwise function $d(x,y)$

$$\underbrace{d(x,x) = 0}_{\text{Identity of Indiscernibles}}\,, \qquad \underbrace{d(x,y) \leq d(x,z) + d(z,y)}_{\text{Triangle Inequality}}, \qquad \forall\, x,y,z$$

e.g. the Euclidean metric

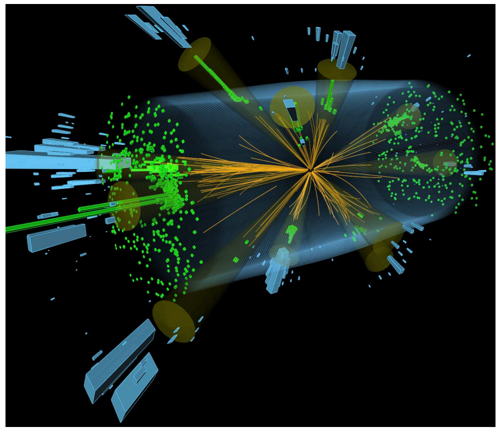
Th … *n,*

$\sum_{i=1}^{M} E_i$

Energy

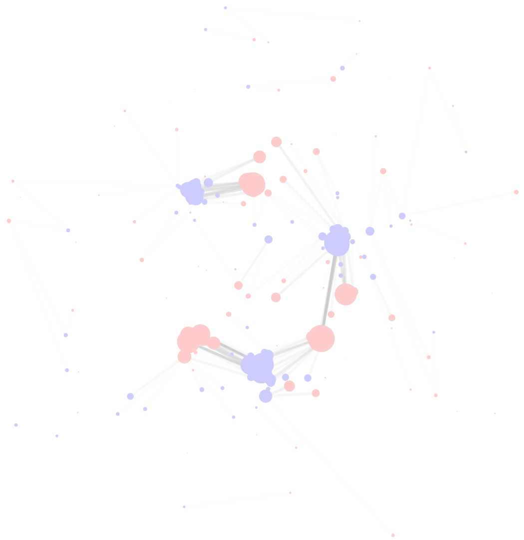
Direction

Energy Flow $\longleftrightarrow$ Infrared and Collinear Safe Information

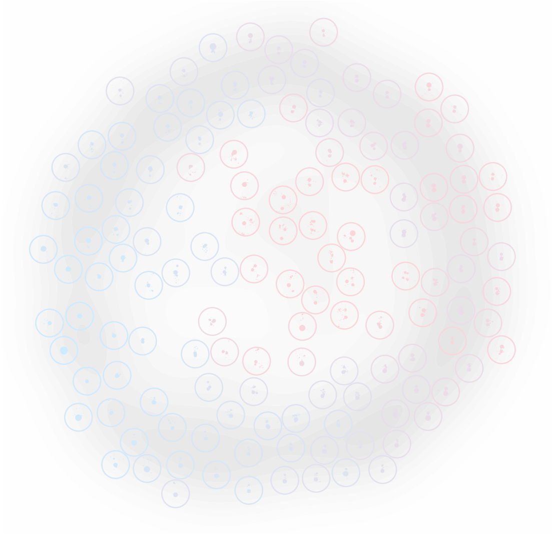


The Space of Collider Events

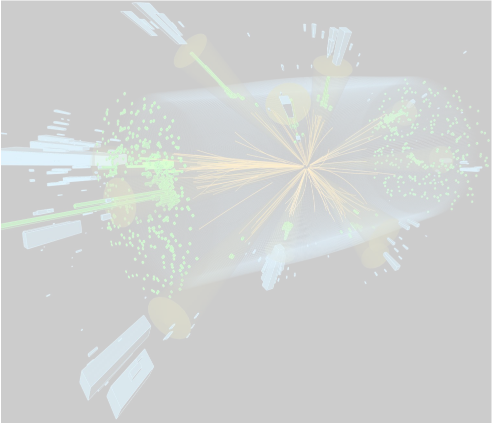
Space of events ≈ IRC-safe energy flows



The Energy Mover’s Distance

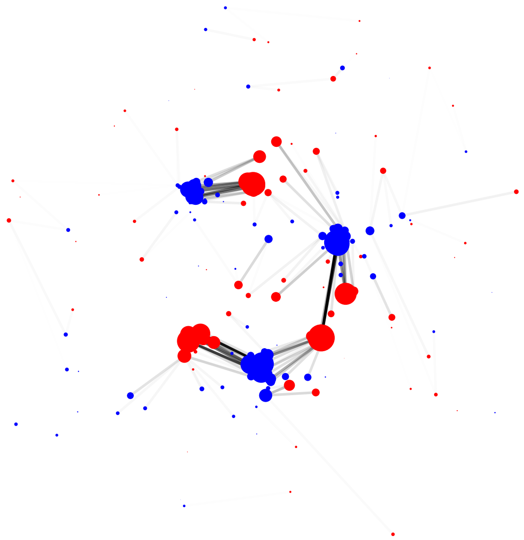


Particle Physics Applications

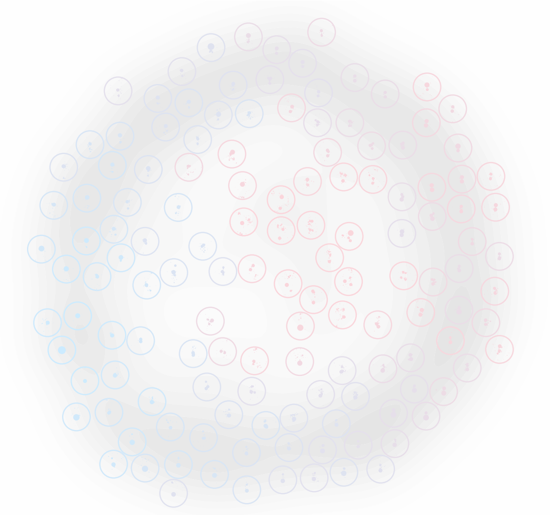


The Space of Collider Events

Space of events ≈ IRC-safe energy flows



The Energy Mover's Distance

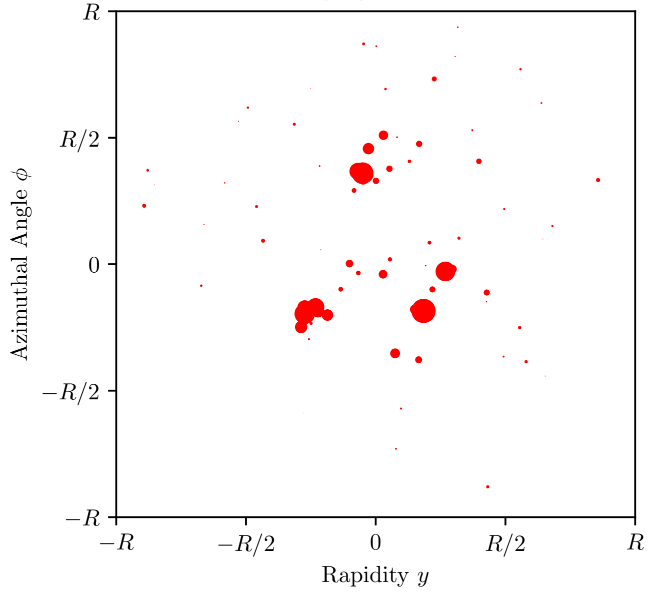


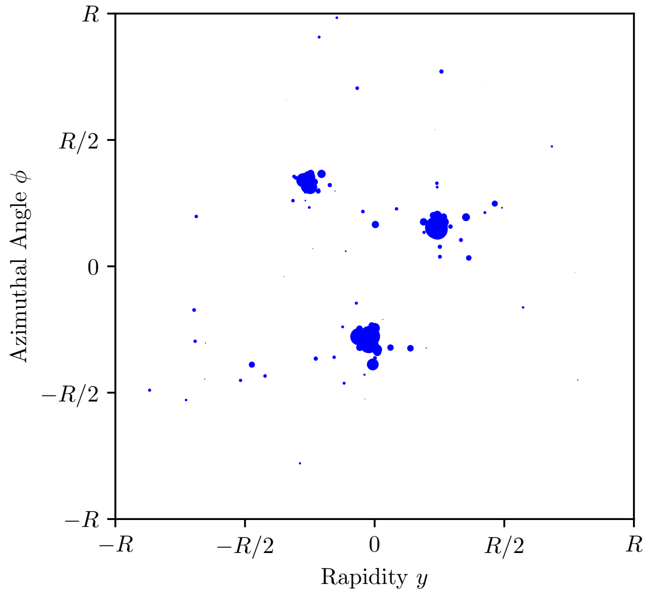
Particle Physics Applications

Optimal Transport – Earth Mover’s Distance

Earth Mover’s Distance is a metric on distributions*

The "work" (stuff x distance) required to most efficiently transport supply to demand


Top Jet 1


Top Jet 2

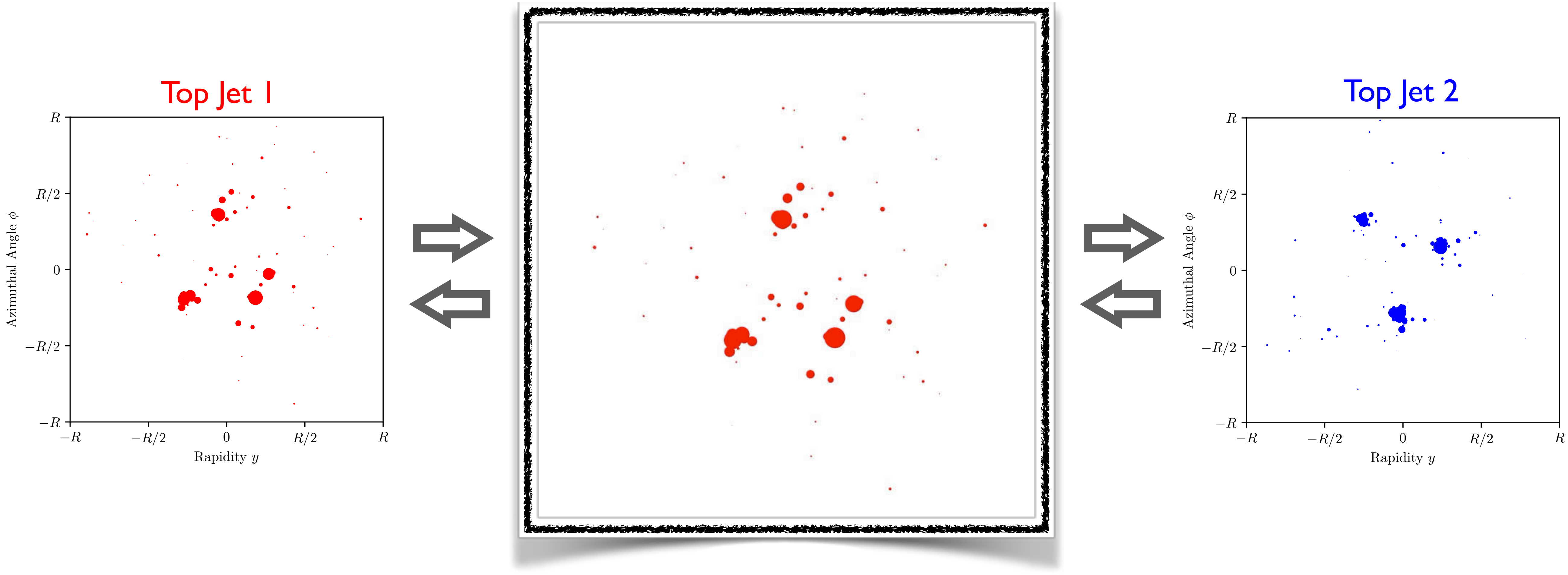
[Peleg, Werman, Rom; Pele, Werman]

*Also known as the 1-Wasserstein distance

Optimal Transport – Earth Mover’s Distance

Earth Mover’s Distance is a metric on distributions*

The "work" (stuff x distance) required to most efficiently transport supply to demand



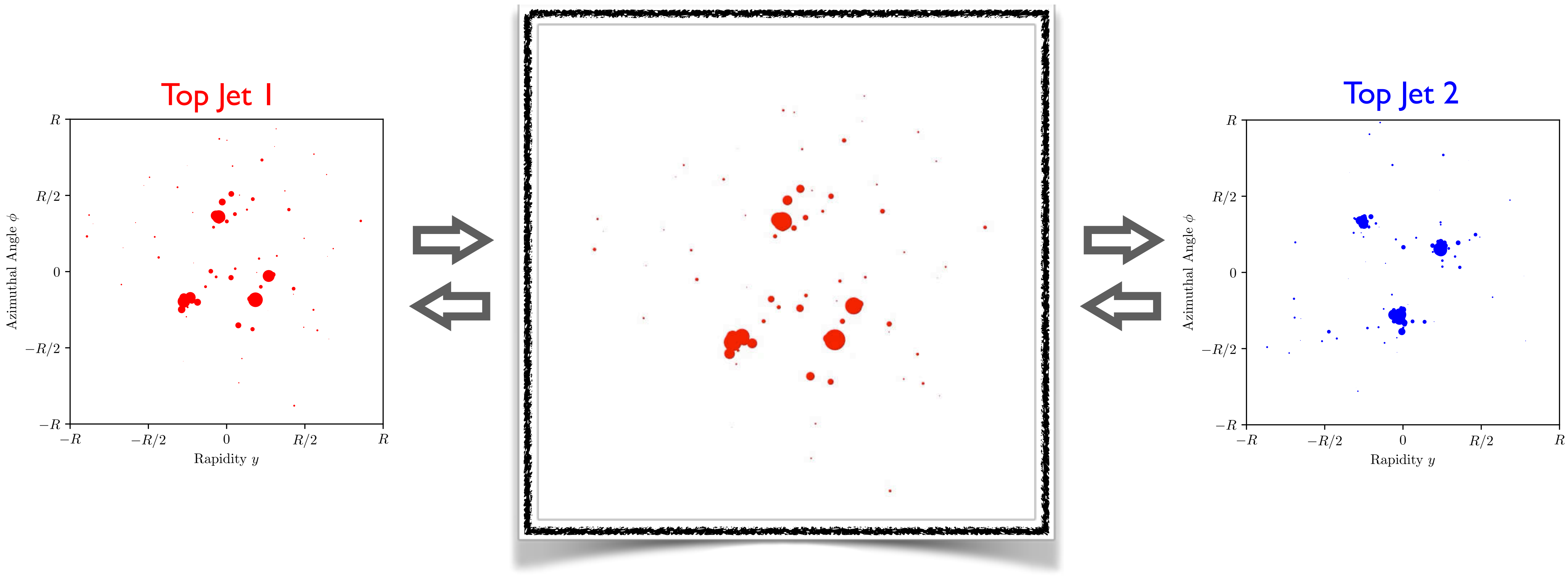
[Peleg, Werman, Rom; Pele, Werman]

*Also known as the 1-Wasserstein distance

Optimal Transport – Earth Mover's Distance

Earth Mover's Distance is a metric on distributions*

The "work" (stuff x distance) required to most efficiently transport supply to demand



Collider event metric: treat events as distributions of energy and find optimal transport

[Peleg, Werman, Rom; Pele, Werman]

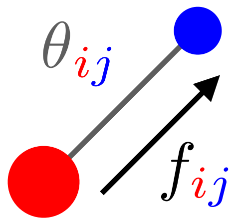
*Also known as the 1-Wasserstein distance

The Energy Mover's Distance (EMD)

[PTK, Metodiev, Thaler, 1902.02346]

$$\text{EMD}(\mathcal{E},\mathcal{E}') = \min_{\{f_{ij}\geq 0\}} \sum_i \sum_j f_{ij} \frac{\theta_{ij}}{R} + \left| \sum_i E_i - \sum_j E'_j \right|$$

$$\sum_j f_{ij} \leq E_i, \quad \sum_i f_{ij} \leq E'_j, \quad \sum_{ij} f_{ij} = \min\left(\sum_i E_i, \sum_j E'_j\right)$$



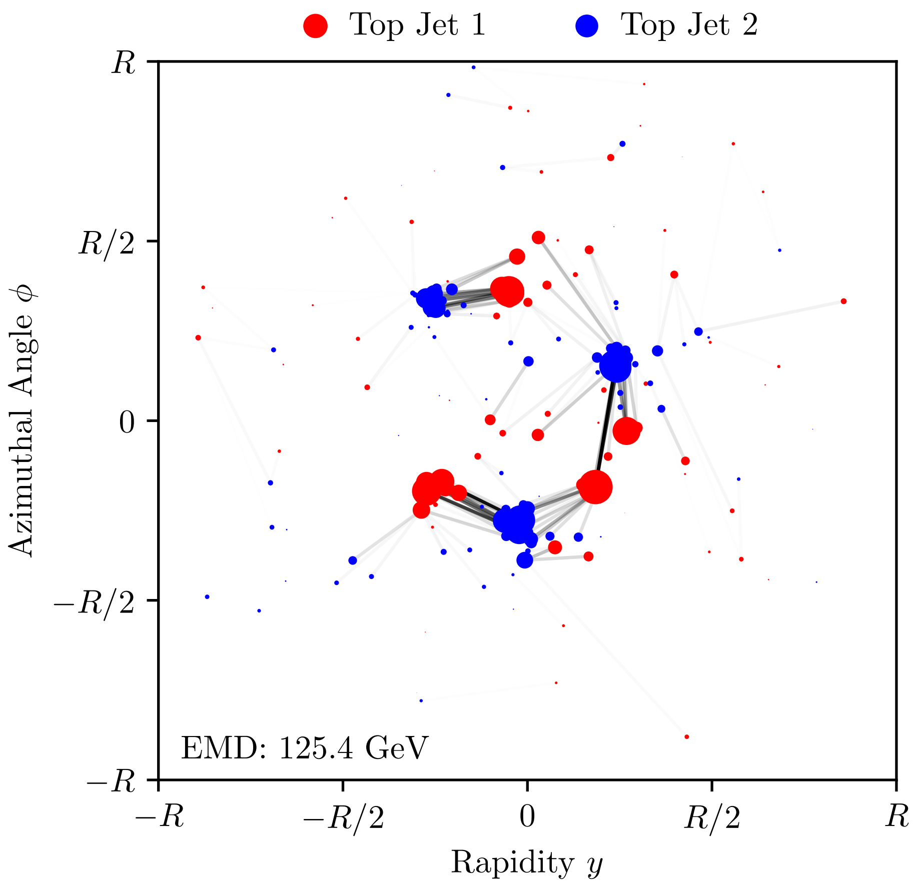
EMD has dimensions of energy

Satisfies triangle inequality as long as $R \geq d_{\max}/2$

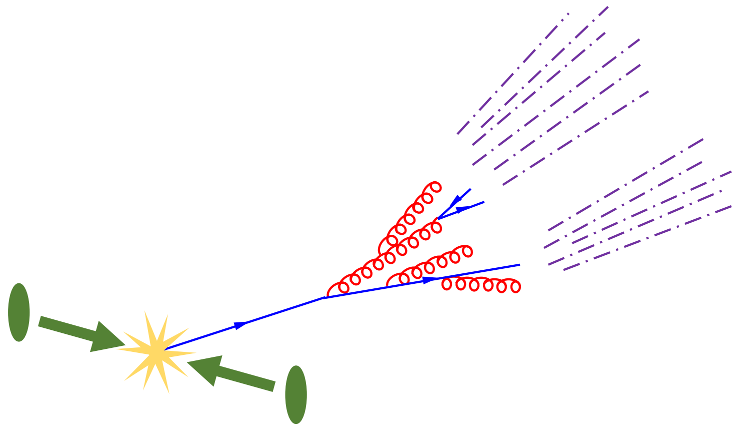
Solvable via network simplex algorithm (polynomial time)

~1 ms for two 100 particle jets on a typical CPU

Alternative to pixel based metric for images



Visualizing Jet Formation – QCD Jets

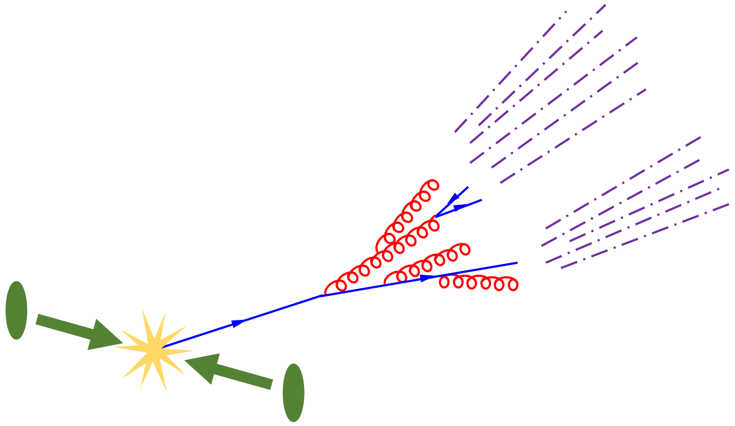


Compare initiating particle to partons from fragmentation to final state hadrons

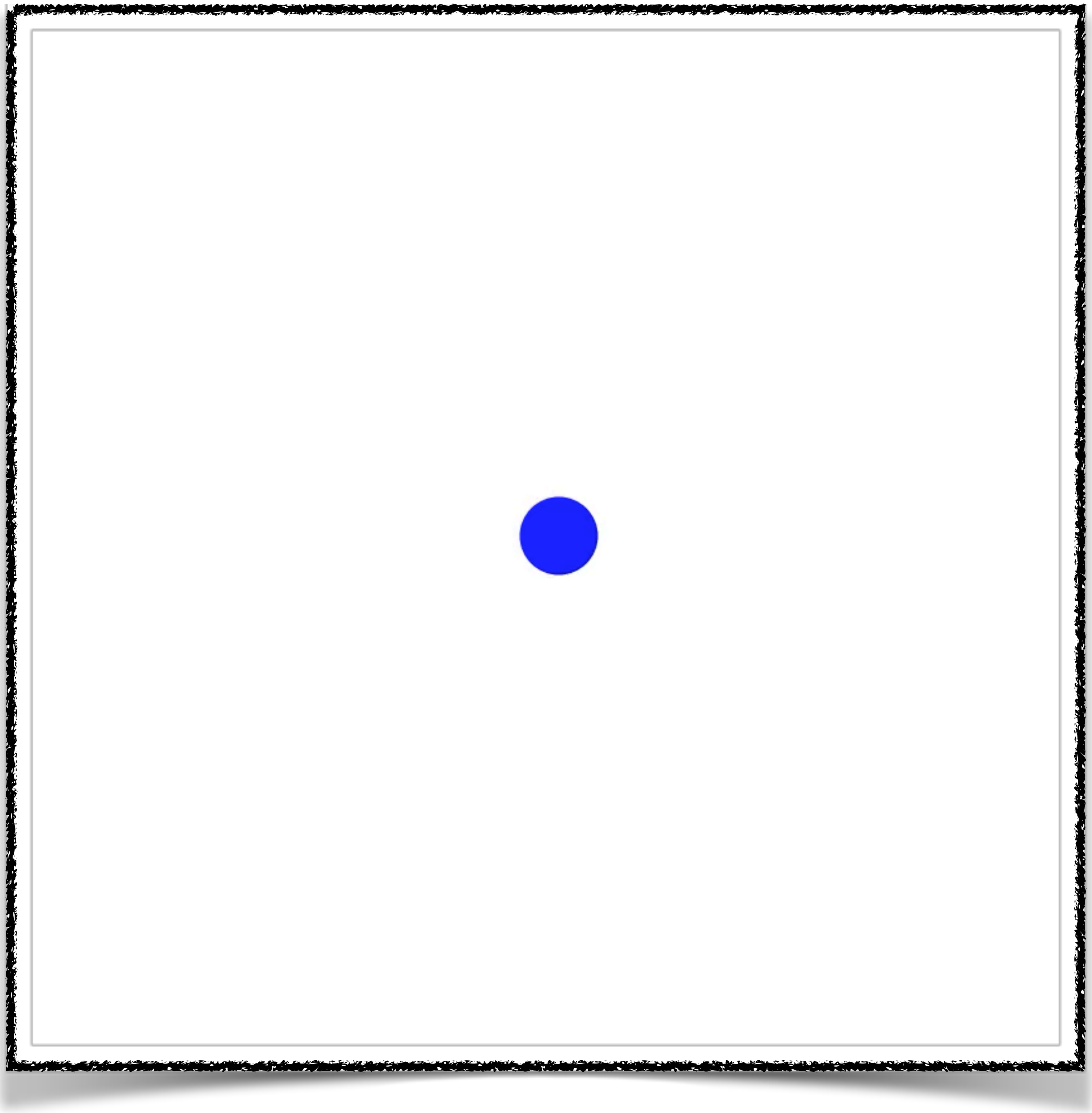
fragmentation

hadronization

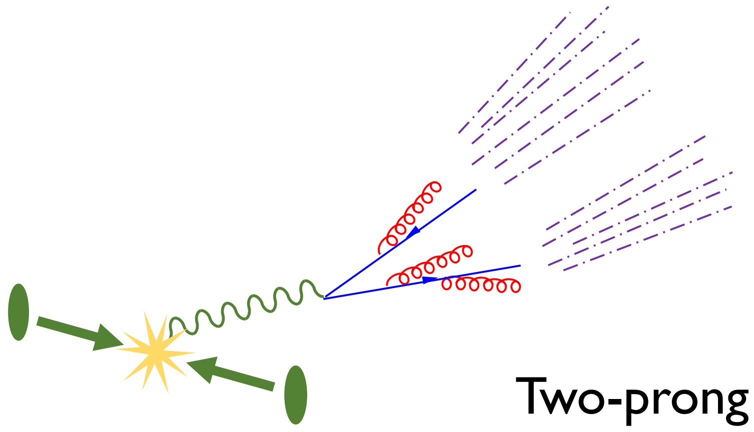
Visualizing Jet Formation – QCD Jets



Compare initiating particle to partons
from fragmentation to final state hadrons

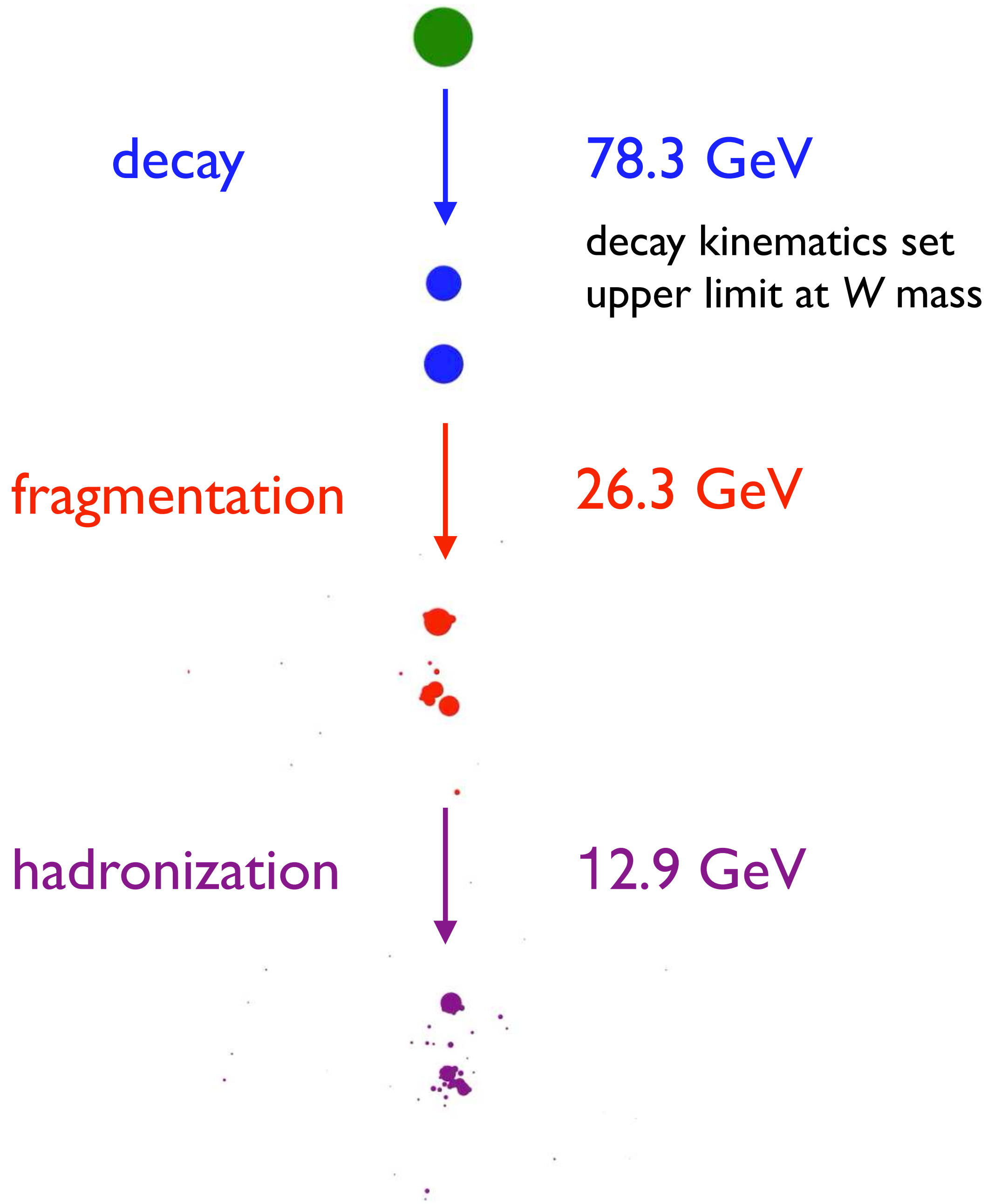


Visualizing Jet Formation – *W* Jets



Two-prong decay

Compare initiating particle to decay products to partons from fragmentation to final state hadrons



decay

78.3 GeV

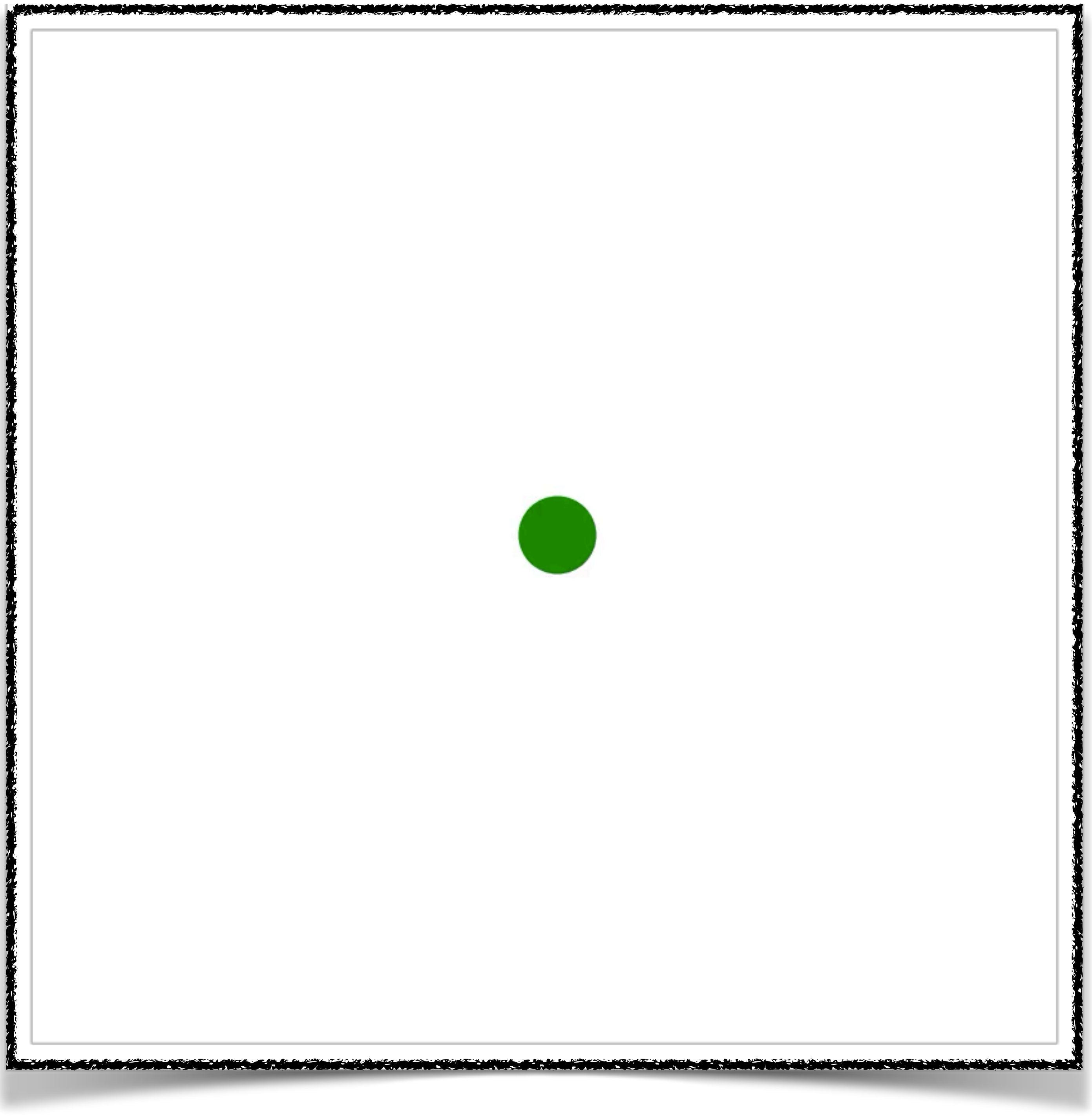
decay kinematics set upper limit at *W* mass

fragmentation

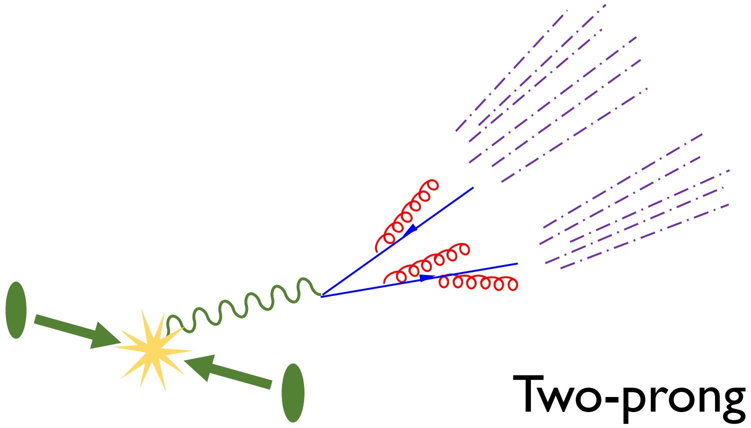
26.3 GeV

hadronization

12.9 GeV

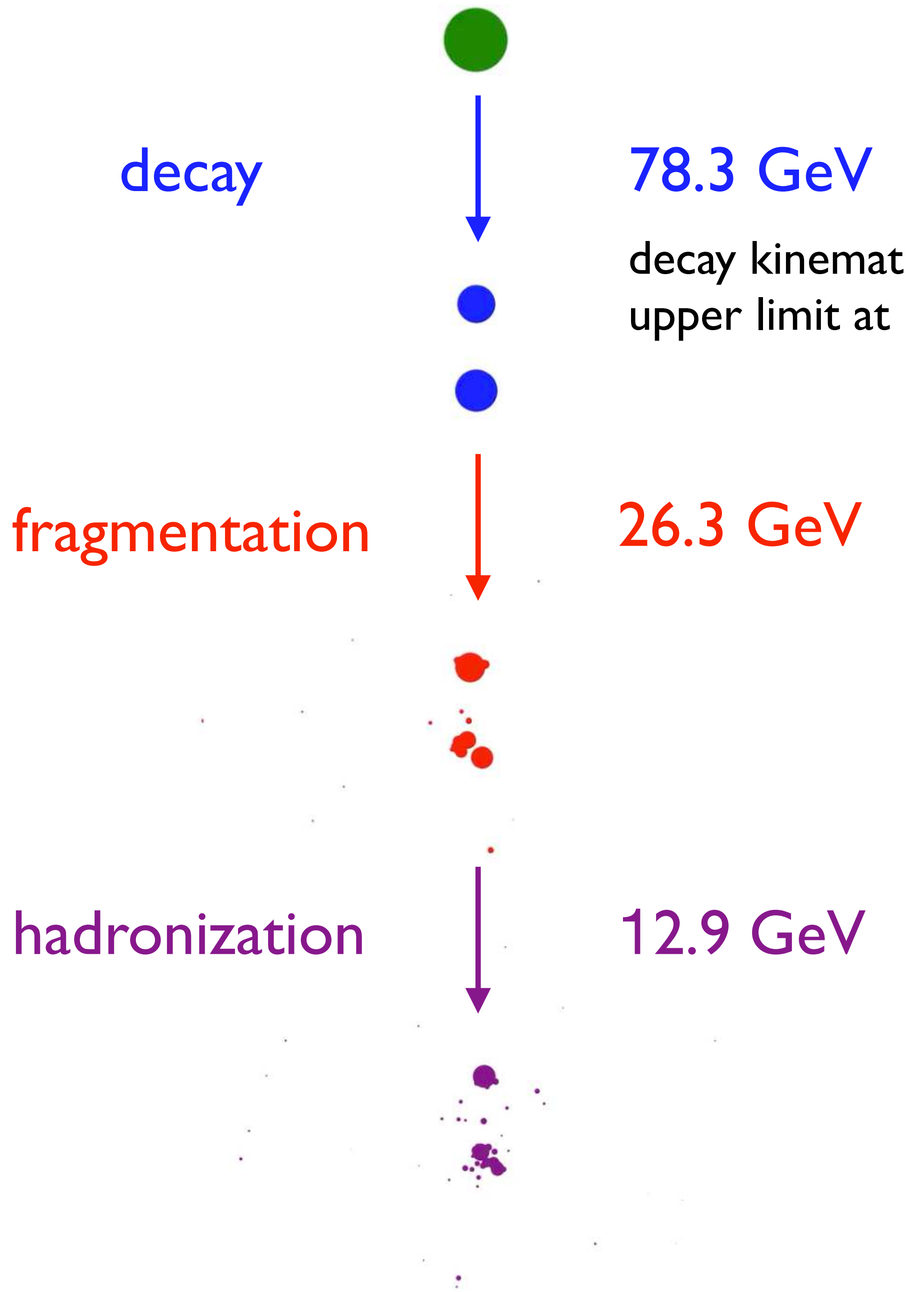


Visualizing Jet Formation – *W* Jets

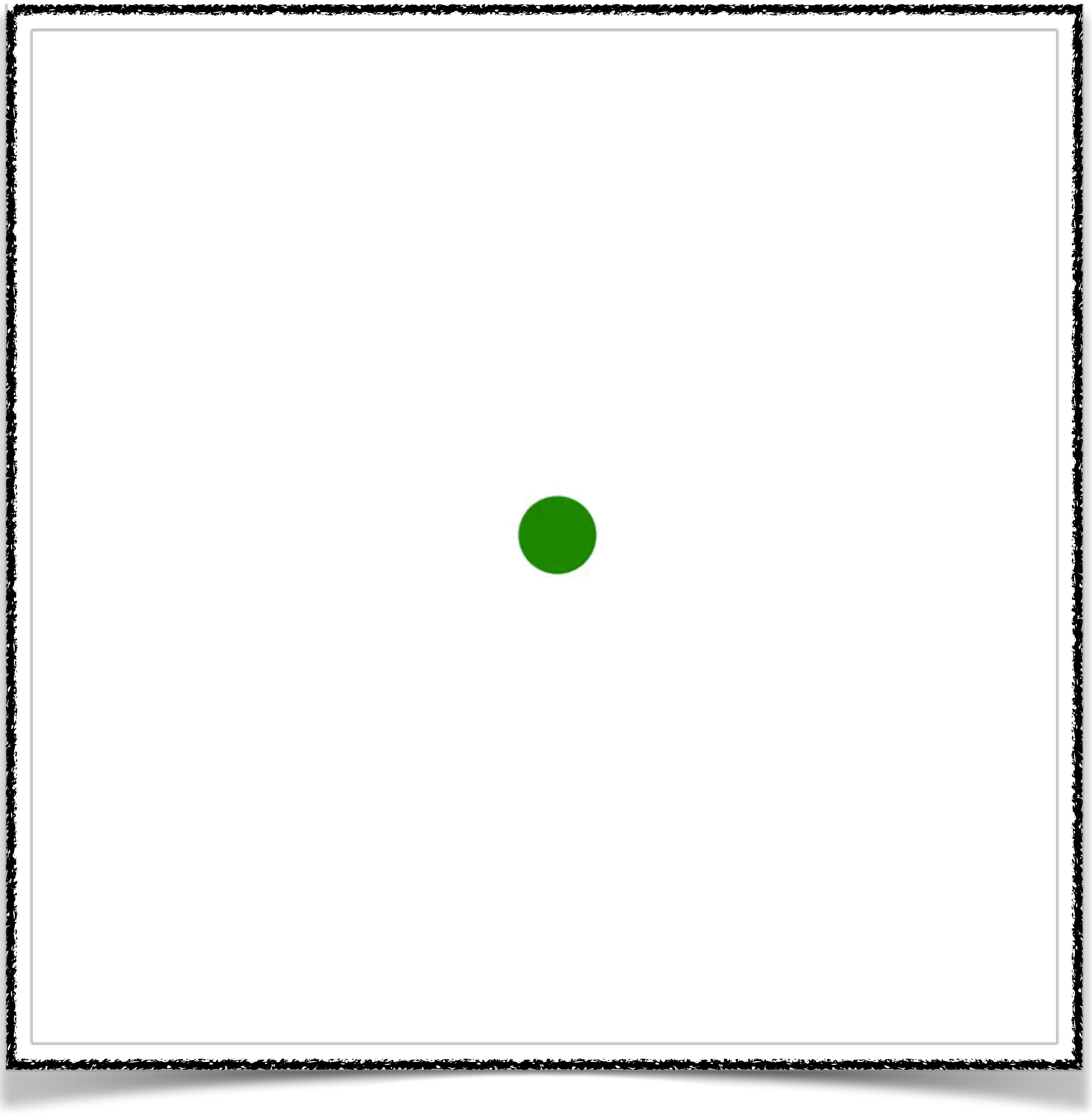


Two-prong decay

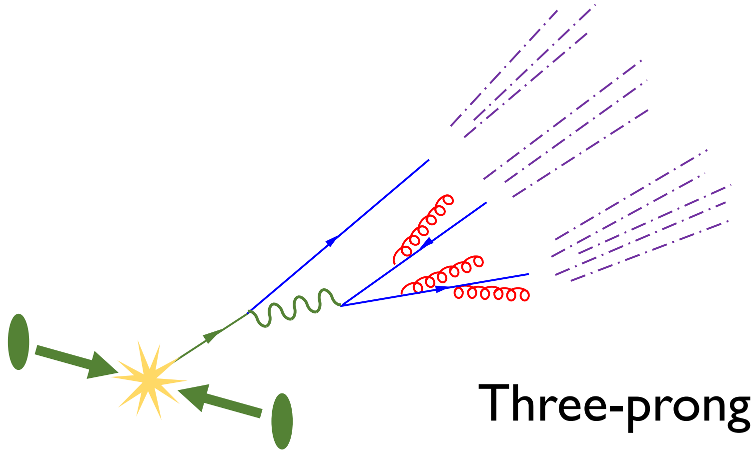
Compare initiating particle to decay products to partons from fragmentation to final state hadrons



decay kinematics set upper limit at *W* mass

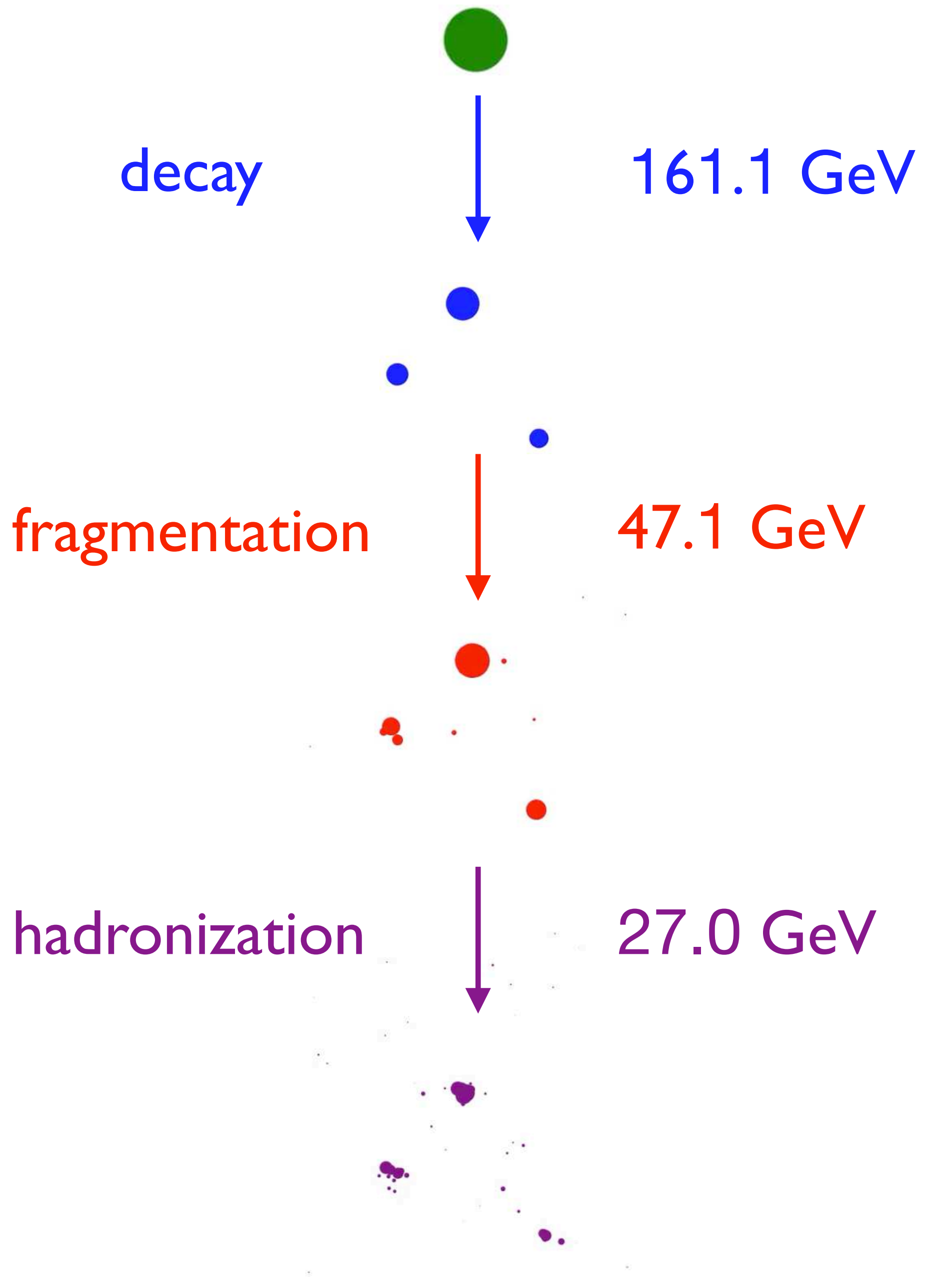


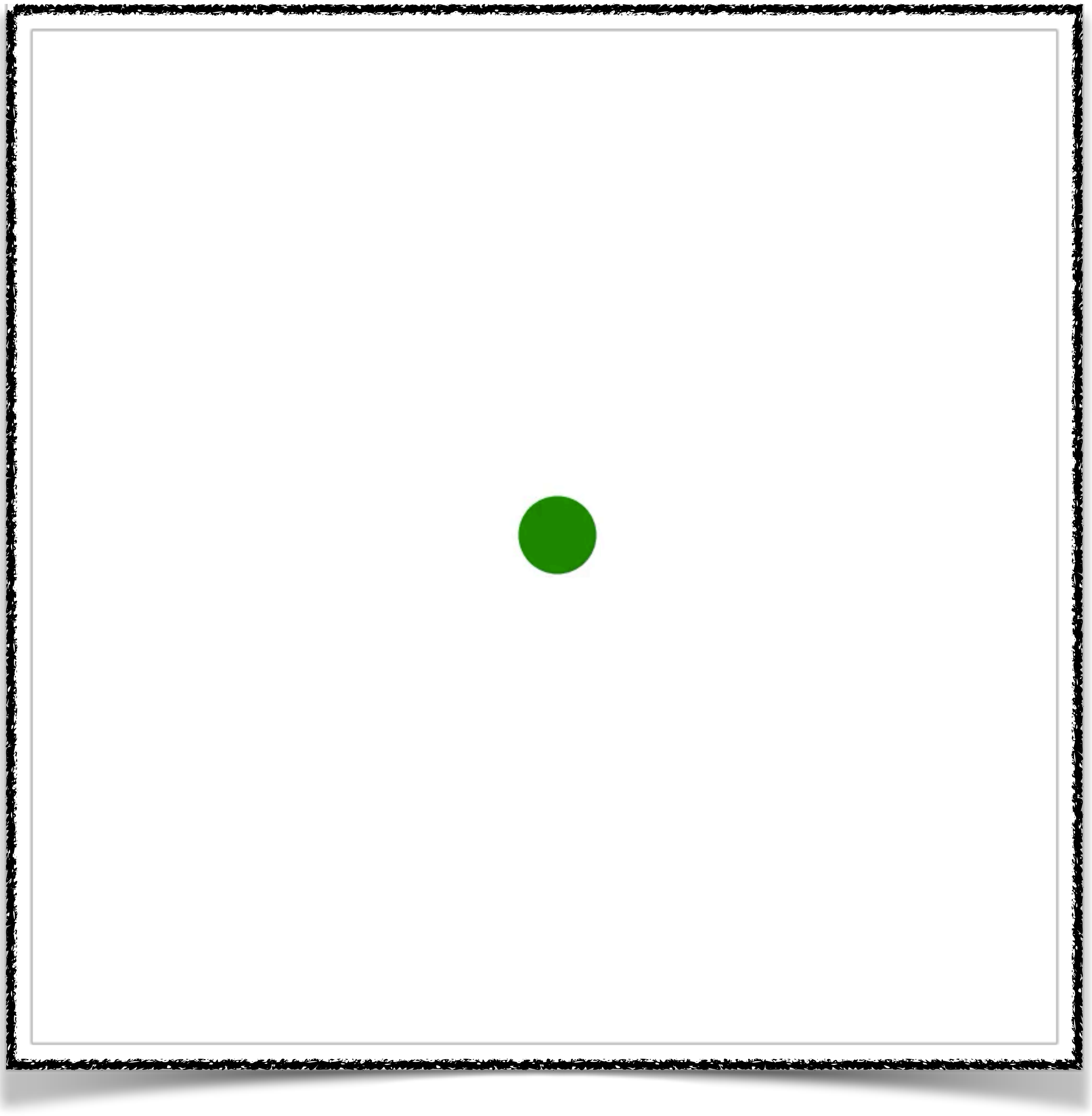
Visualizing Jet Formation – Top Jets



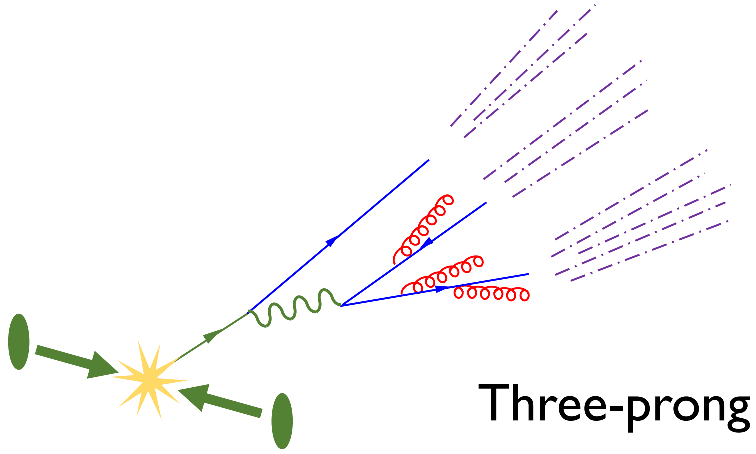
Three-prong decay

Compare initiating particle to decay products to partons from fragmentation to final state hadrons



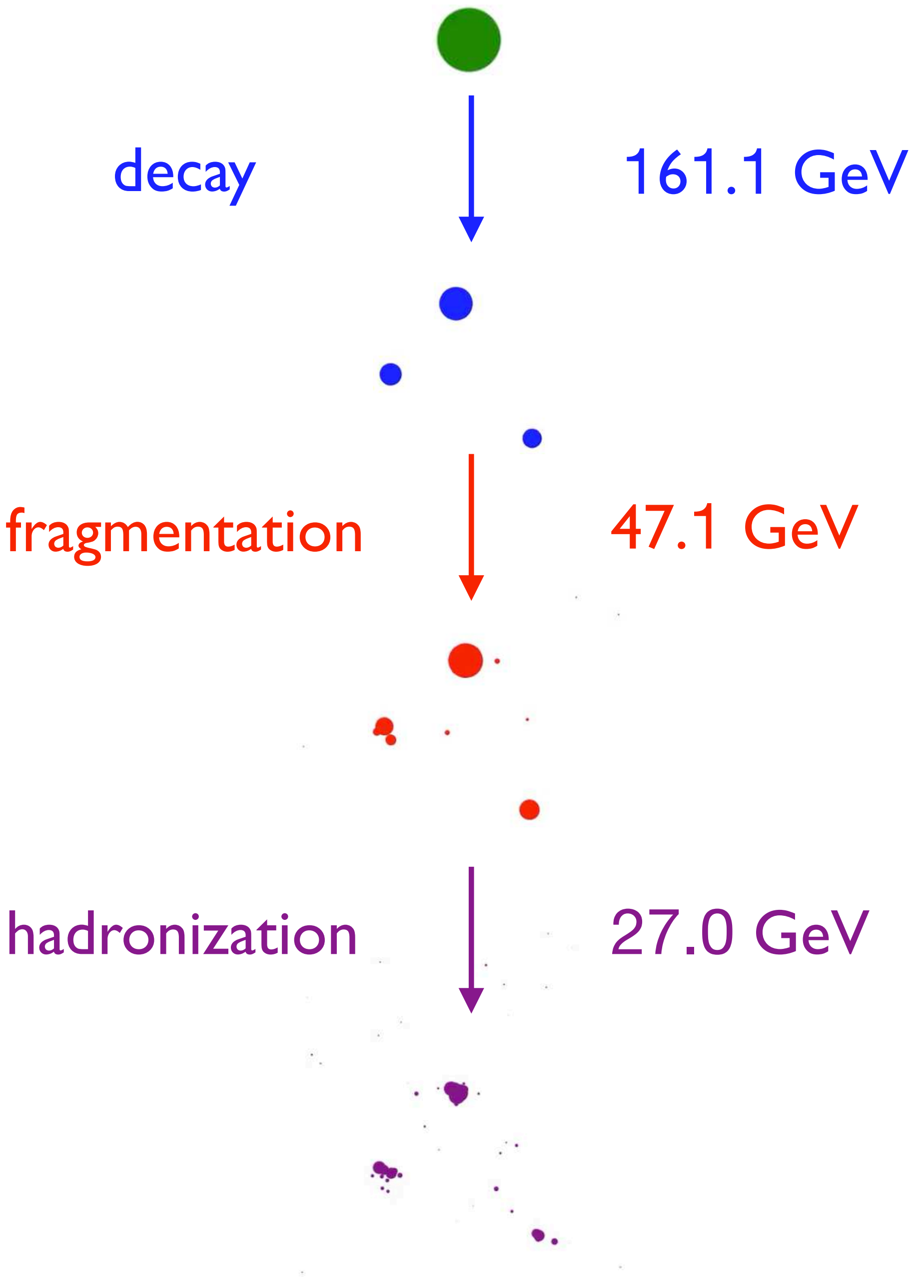


Visualizing Jet Formation – Top Jets



Three-prong decay

Compare initiating particle to decay products to partons from fragmentation to final state hadrons



decay

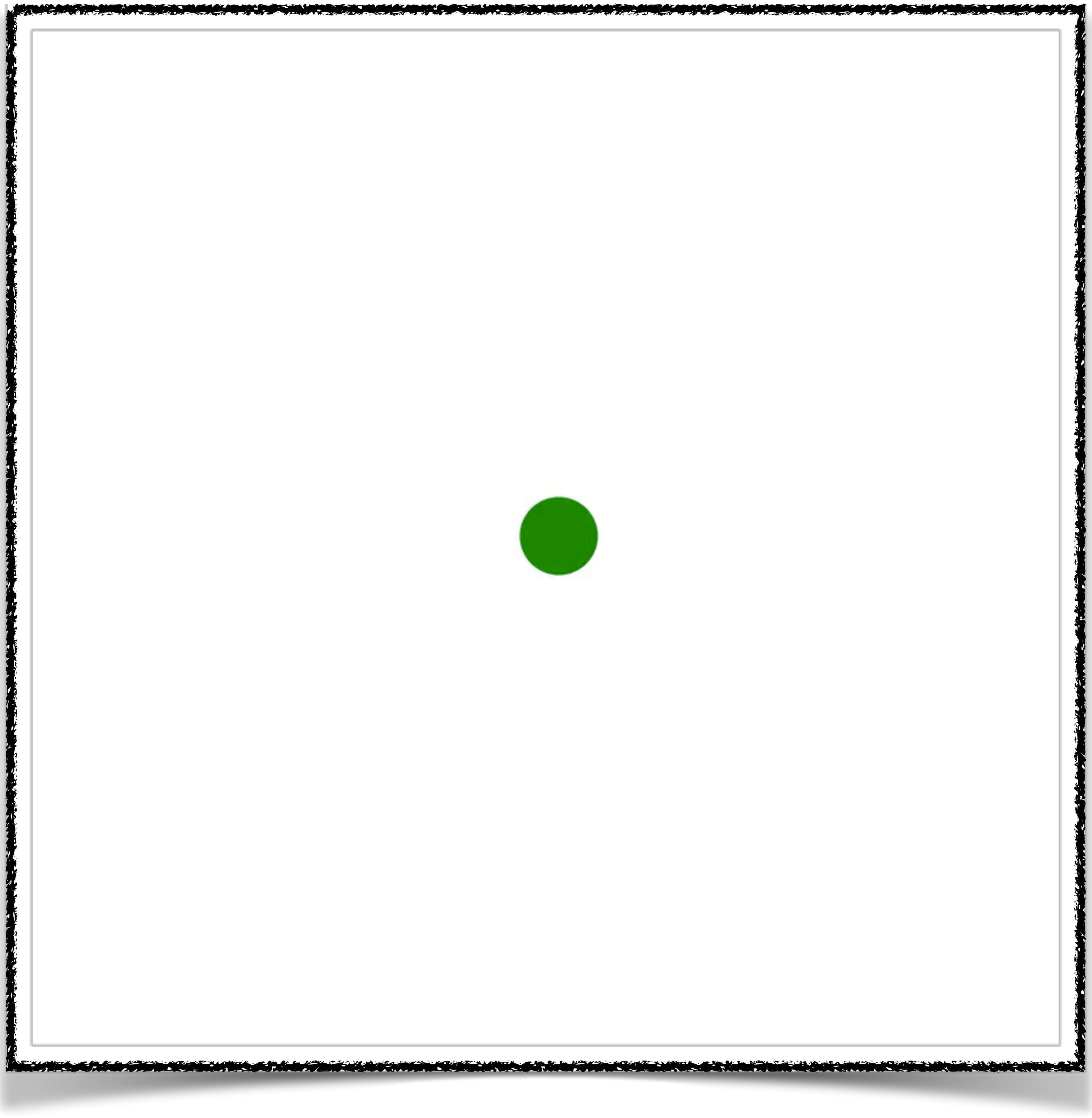
161.1 GeV

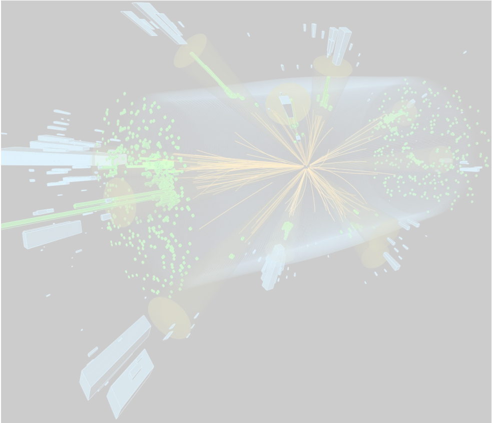
fragmentation

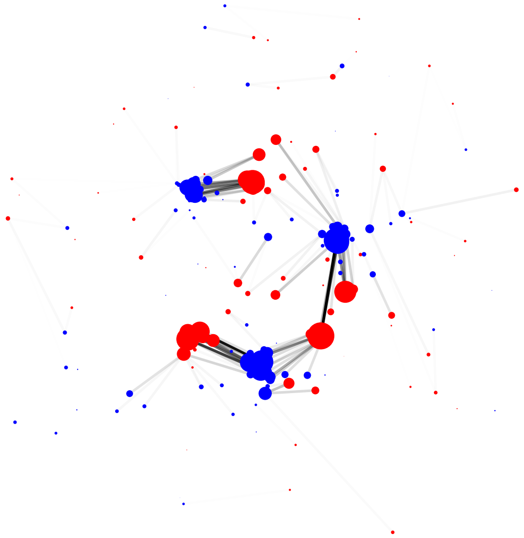
47.1 GeV

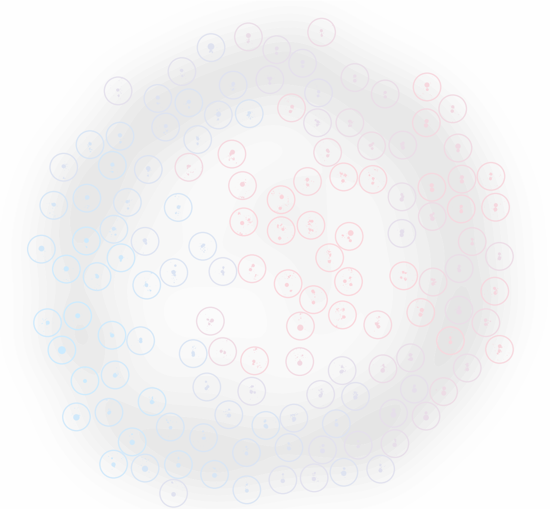
hadronization

27.0 GeV









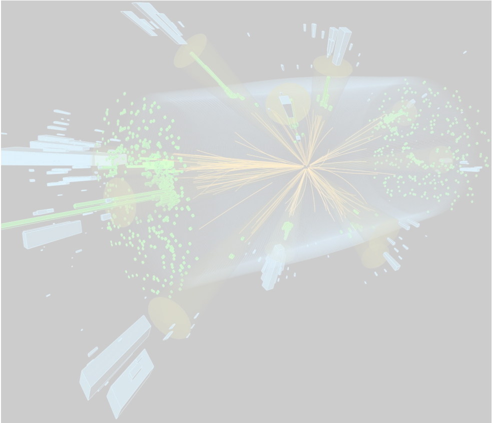
The Space of Collider Events

Space of events ≈ IRC-safe energy flows

The Energy Mover's Distance

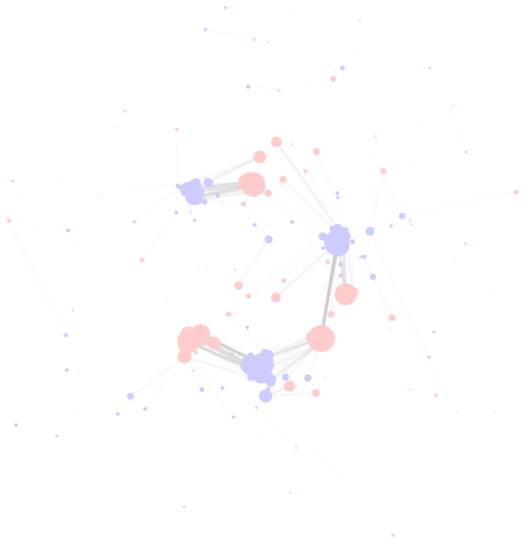
Quantifies the difference in radiation pattern between events

Particle Physics Applications



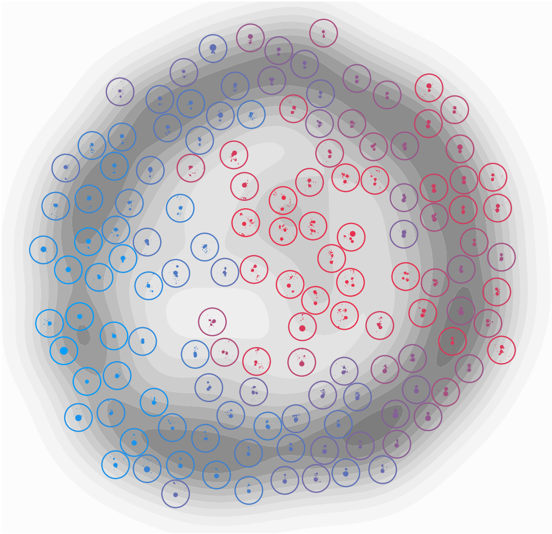
The Space of Collider Events

Space of events ≈ IRC-safe energy flows



The Energy Mover's Distance

Quantifies the difference in radiation pattern between events

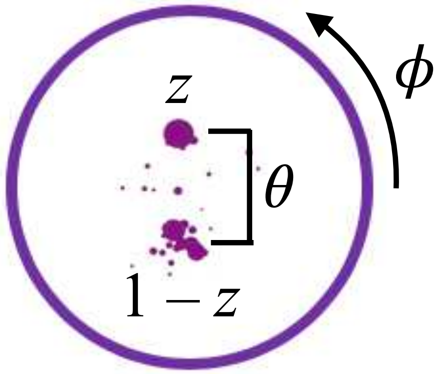


Particle Physics Applications

Visualizing the Metric Space of *W* Jets

Metric spaces have intrinsic structure (e.g. triangulation of points in R^3 from pairwise distances)

Cartoon *W* jets are two-pronged and have three degrees of freedom

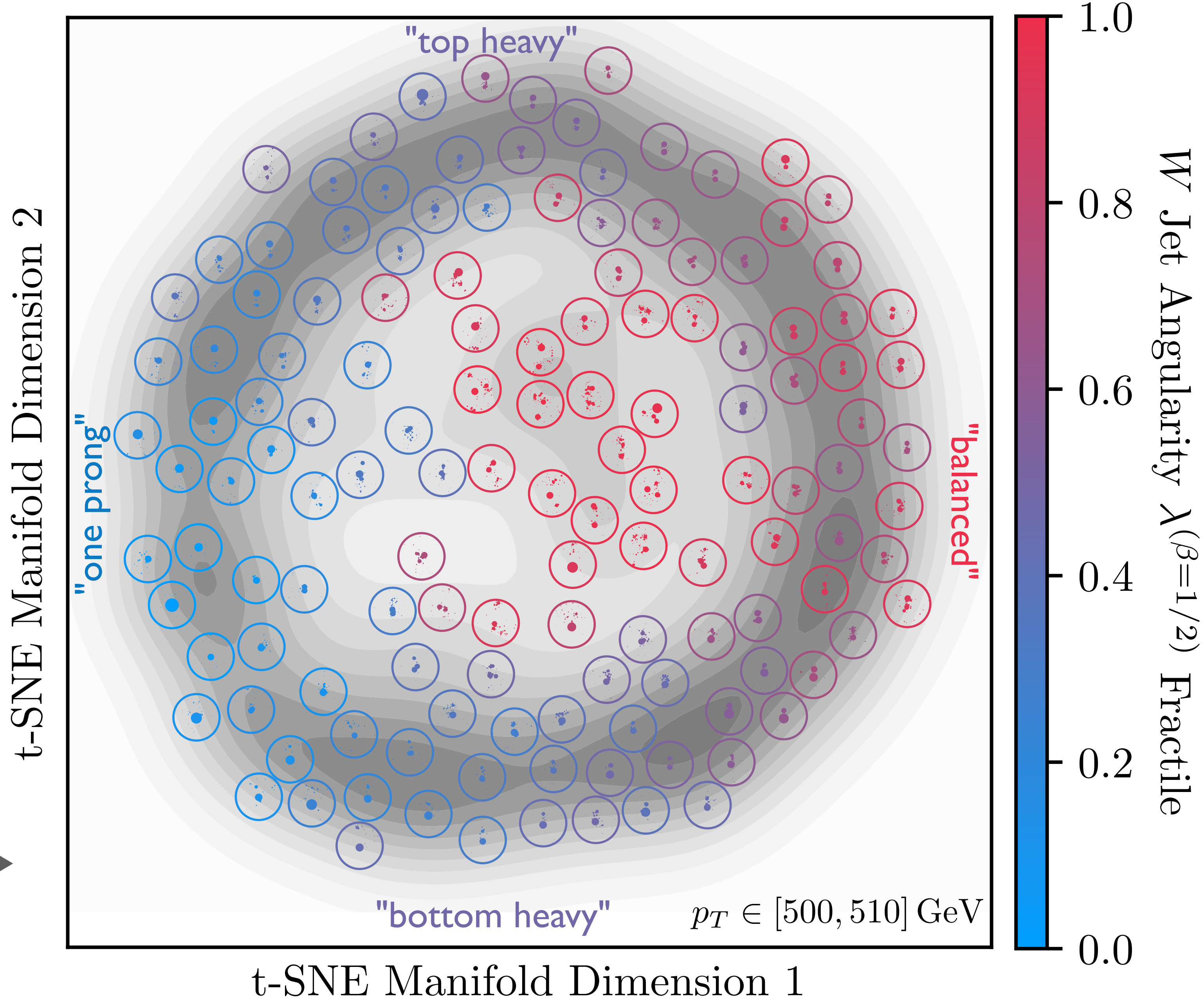


Mass constraint & rotational preprocessing remove two DOF

One dimensional manifold appears as a ring with weird events at the center

Gray contours show event density, example jets sprinkled throughout

t-SNE finds 2d embedding that attempts to respect distances according to a given metric

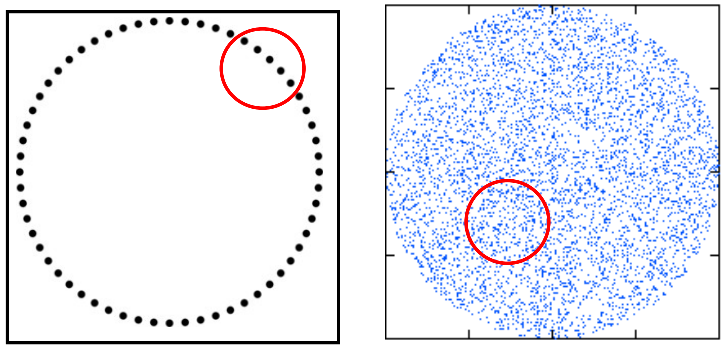


[L. van der Maaten, G. Hinton]

Manifold Dimensions of Event Space

What is the dimension of the manifold of QCD, *W*, or top jets?

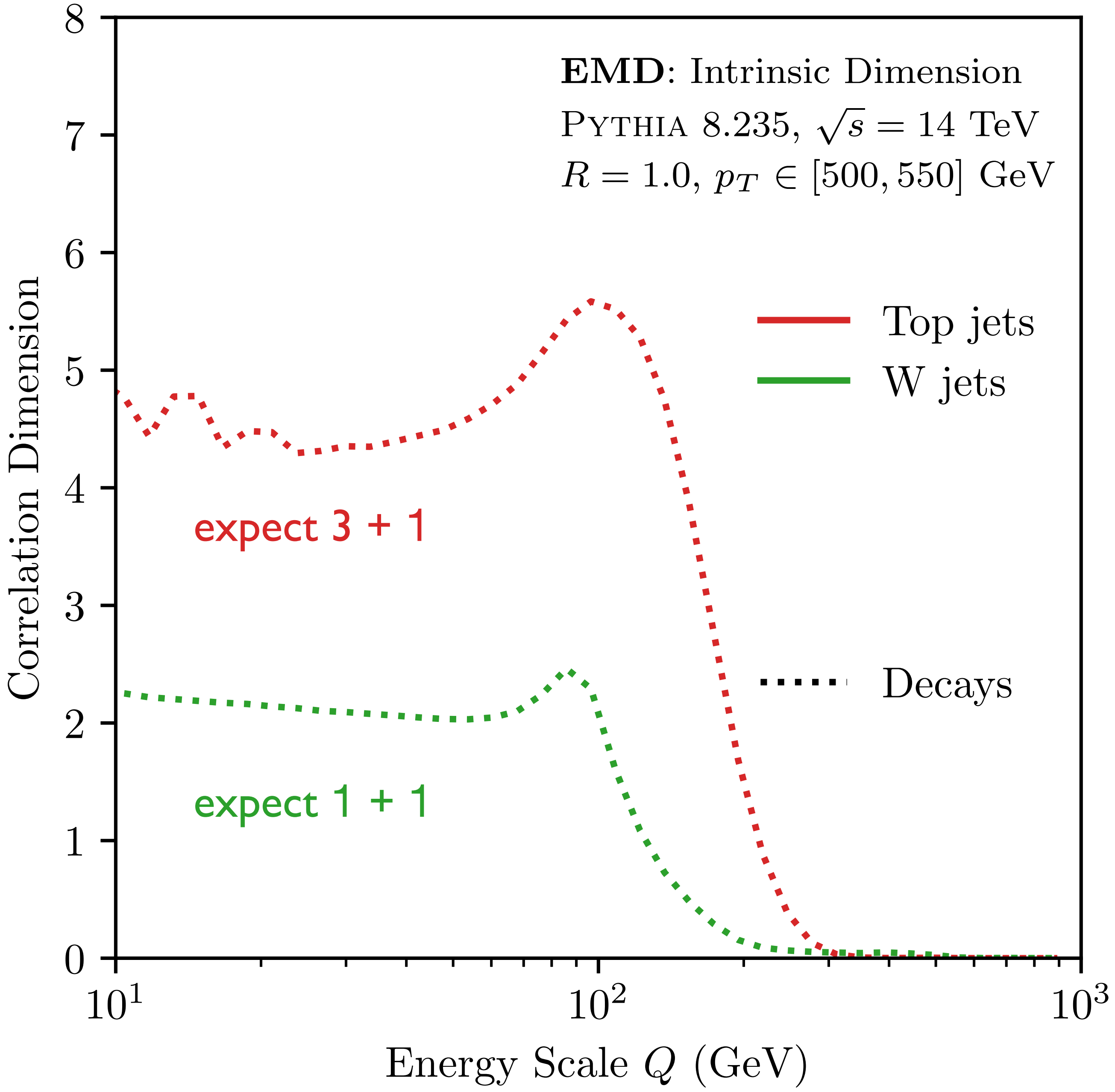
<u>Correlation dimension</u>: how does the # of elements within a ball of size *Q* change?



$$N_{\text{neigh.}}(Q) \propto Q^{\text{dim}} \quad \Longrightarrow \quad \text{dim}(Q) = \frac{d}{d\ln Q} N_{\text{neigh.}}(Q)$$

Correlation dimension lessons:

Complexity hierarchy: QCD < W < Top
Decays are "constant" dim. at low *Q*
Fragmentation increases dim. at smaller scales
Hadronization important around 20-30 GeV

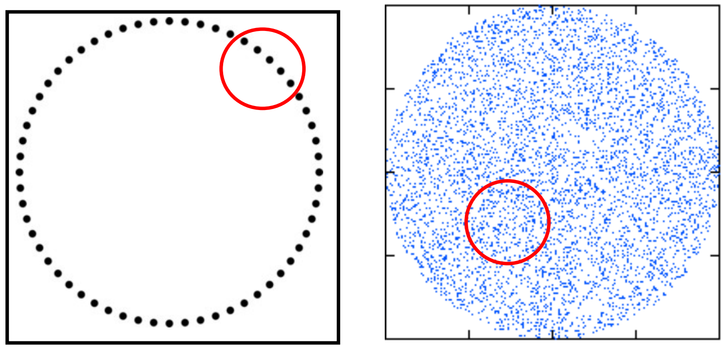


[Grassberger, Procaccia]

Manifold Dimensions of Event Space

What is the dimension of the manifold of QCD, *W*, or top jets?

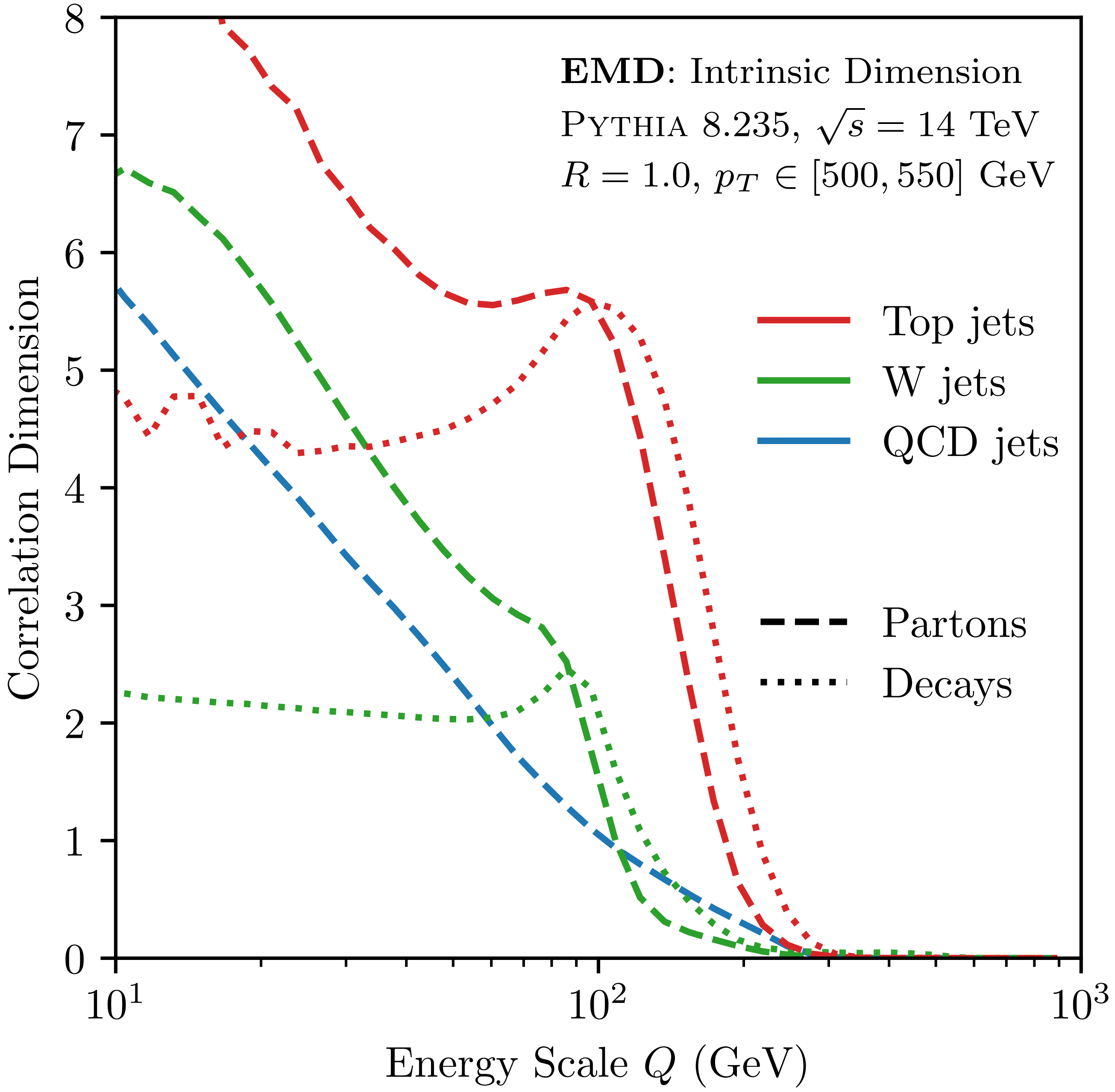
<u>Correlation dimension</u>: how does the # of elements within a ball of size *Q* change?



$$N_{\text{neigh.}}(Q) \propto Q^{\text{dim}} \implies \text{dim}(Q) = \frac{d}{d\ln Q} N_{\text{neigh.}}(Q)$$

Correlation dimension lessons:

- Complexity hierarchy: QCD < W < Top
- Decays are "constant" dim. at low *Q*
- Fragmentation increases dim. at smaller scales
- Hadronization important around 20-30 GeV

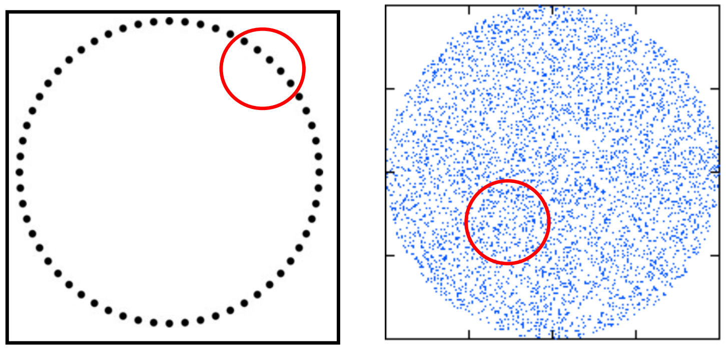


[Grassberger, Procaccia]

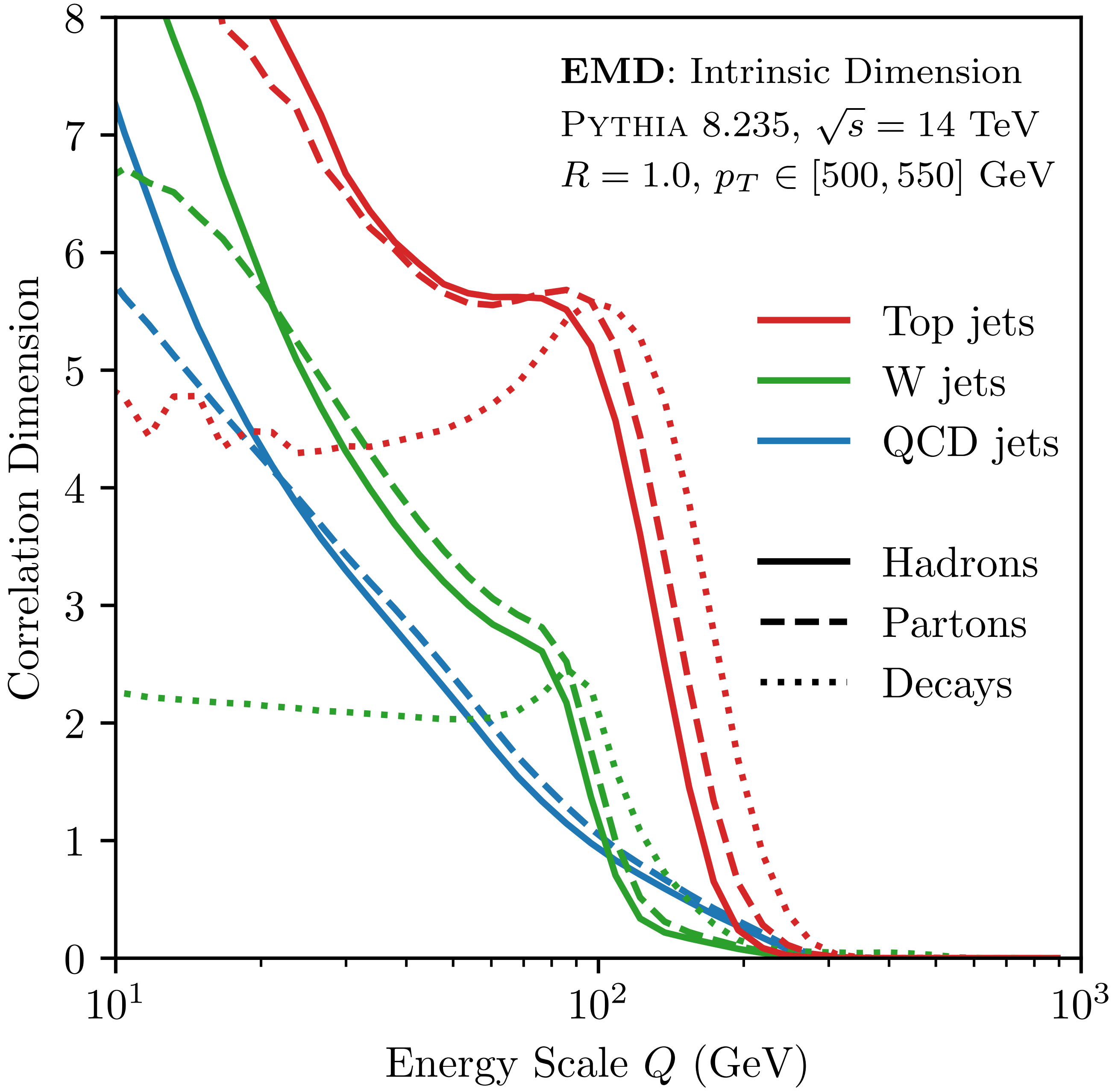
Manifold Dimensions of Event Space

What is the dimension of the manifold of QCD, *W*, or top jets?

<u>Correlation dimension</u>: how does the # of elements within a ball of size *Q* change?



$$N_{\text{neigh.}}(Q) \propto Q^{\text{dim}} \implies \text{dim}(Q) = \frac{d}{d\ln Q} N_{\text{neigh.}}(Q)$$

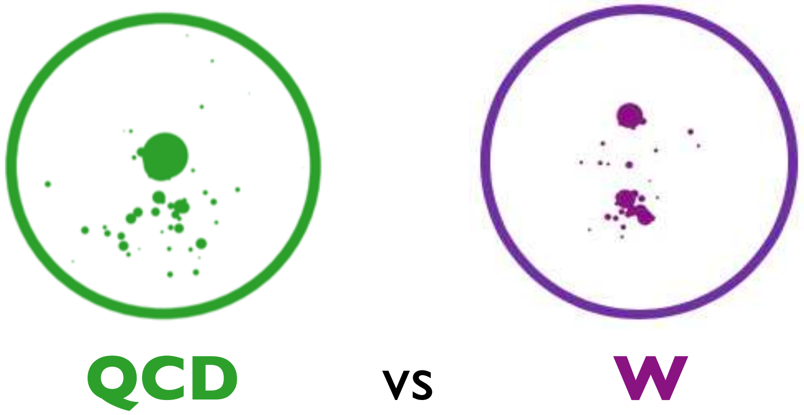


Correlation dimension lessons:

- Complexity hierarchy: QCD < W < Top
- Decays are "constant" dim. at low *Q*
- Fragmentation increases dim. at smaller scales
- Hadronization important around 20-30 GeV

[Grassberger, Procaccia]

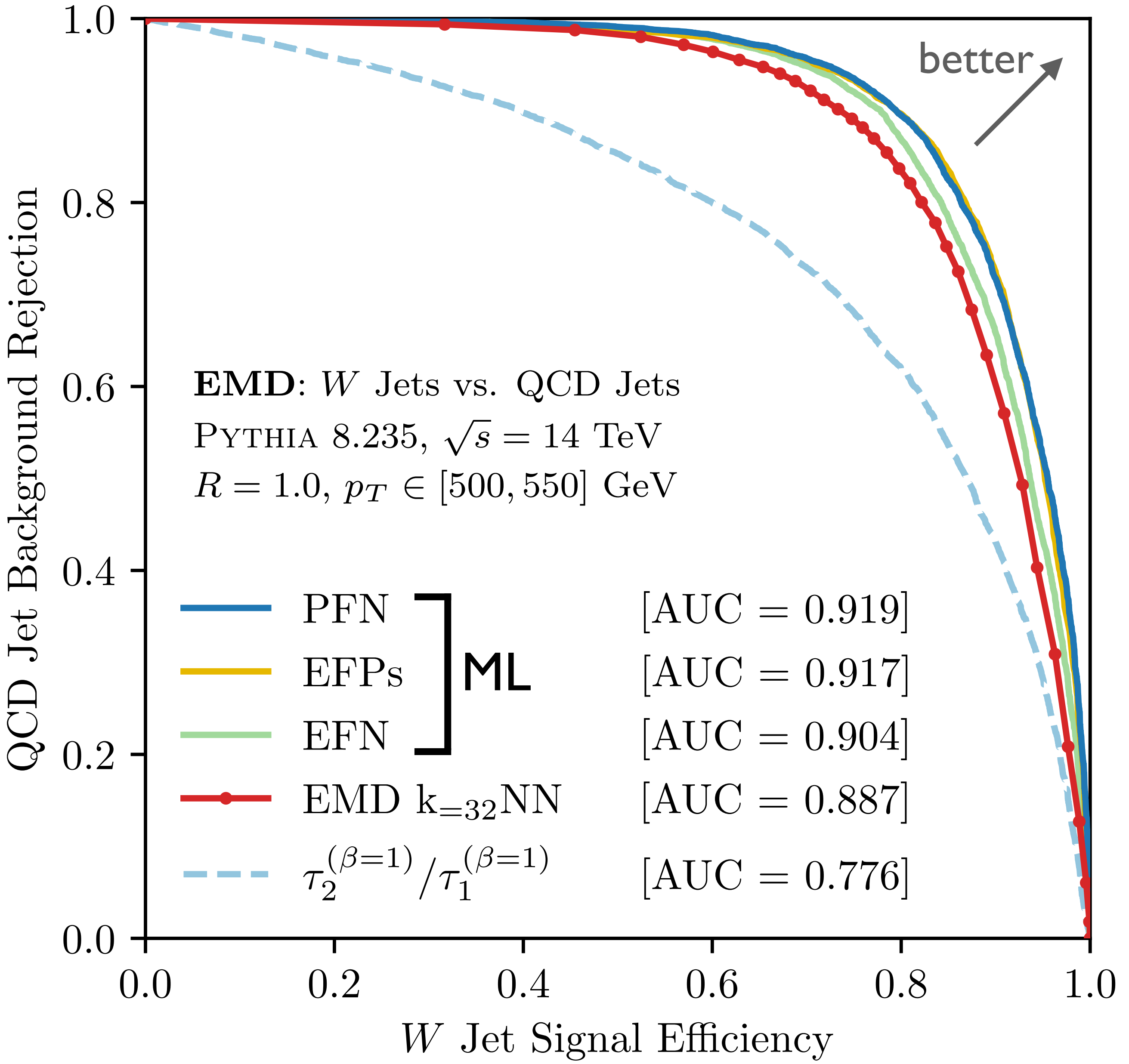
Nearest Neighbor Density Estimation for Jet Classification



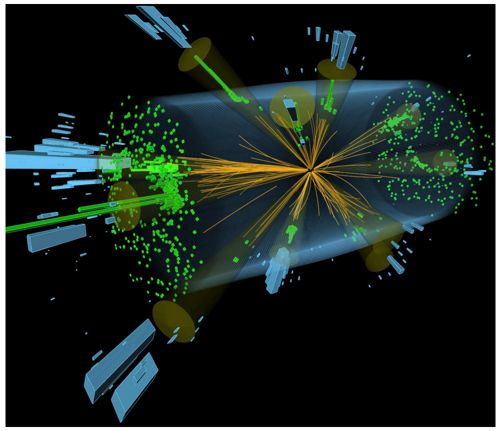
Given a reference sample of two kinds of jets, classify test jets based on k-nearest neighbors

Optimal IRC-safe classifier with enough data

kNN performance approaches that of ML

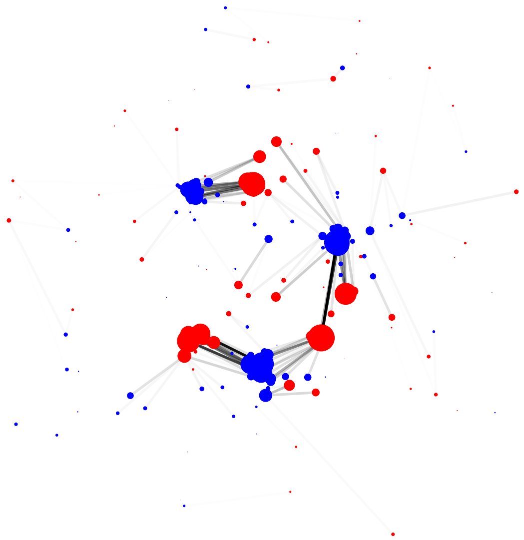


[PTK, Metodiev, Thaler, 1810.05165]



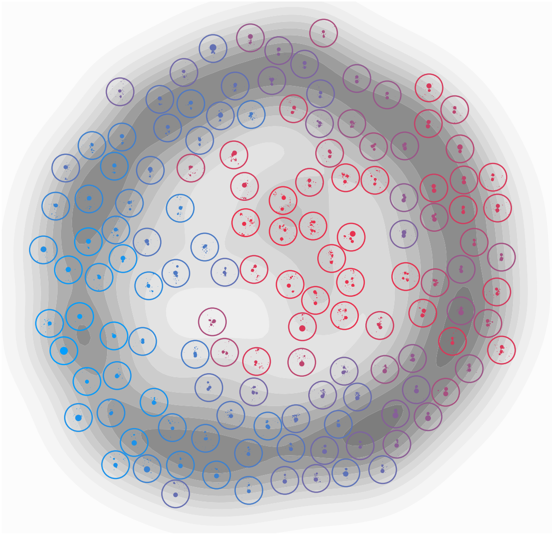
The Space of Collider Events

Space of events ≈ IRC-safe energy flows



The Energy Mover’s Distance

Quantifies the difference in radiation pattern between events



Particle Physics Applications

Visualizing and quantifying event manifolds, kNN classification

Further Directions

EMD quantifies energy flow – use it to quantify observables*?

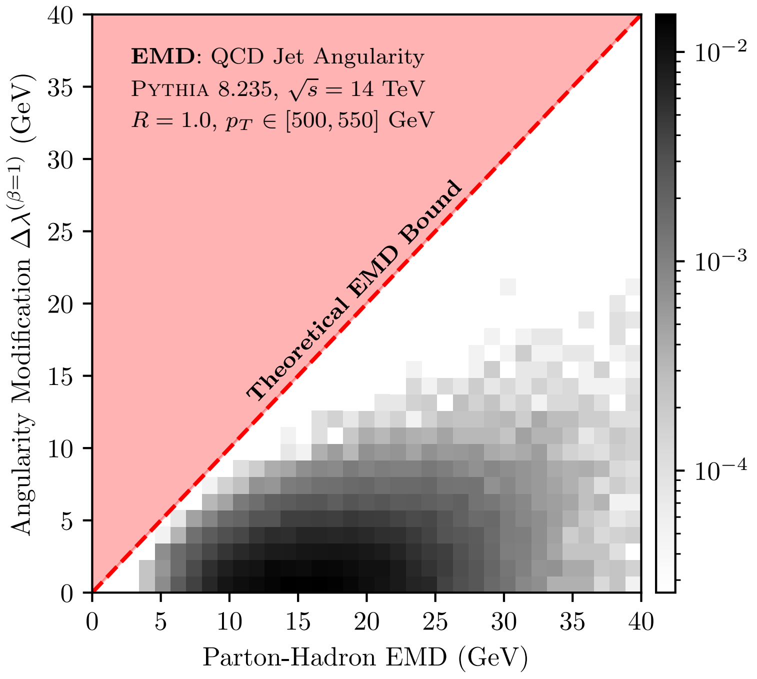
(# of LHC events) >> 1 – distill most representative events?

(# of LHC events)^2 ~ ∞, speed up using triangle inequality?

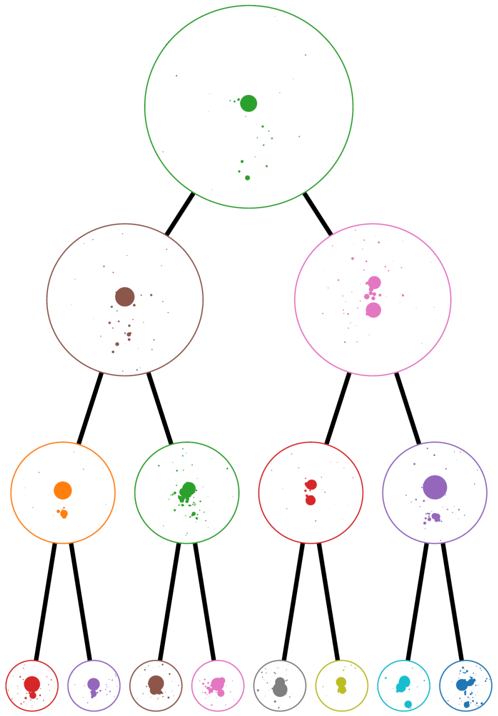
Interesting physics in correlation dimension – can we calculate it?

EMD is IRC safe – include unsafe information e.g. flavor?

EMD quantifies differences – use as ML loss function?



*More in backup



Metric tree

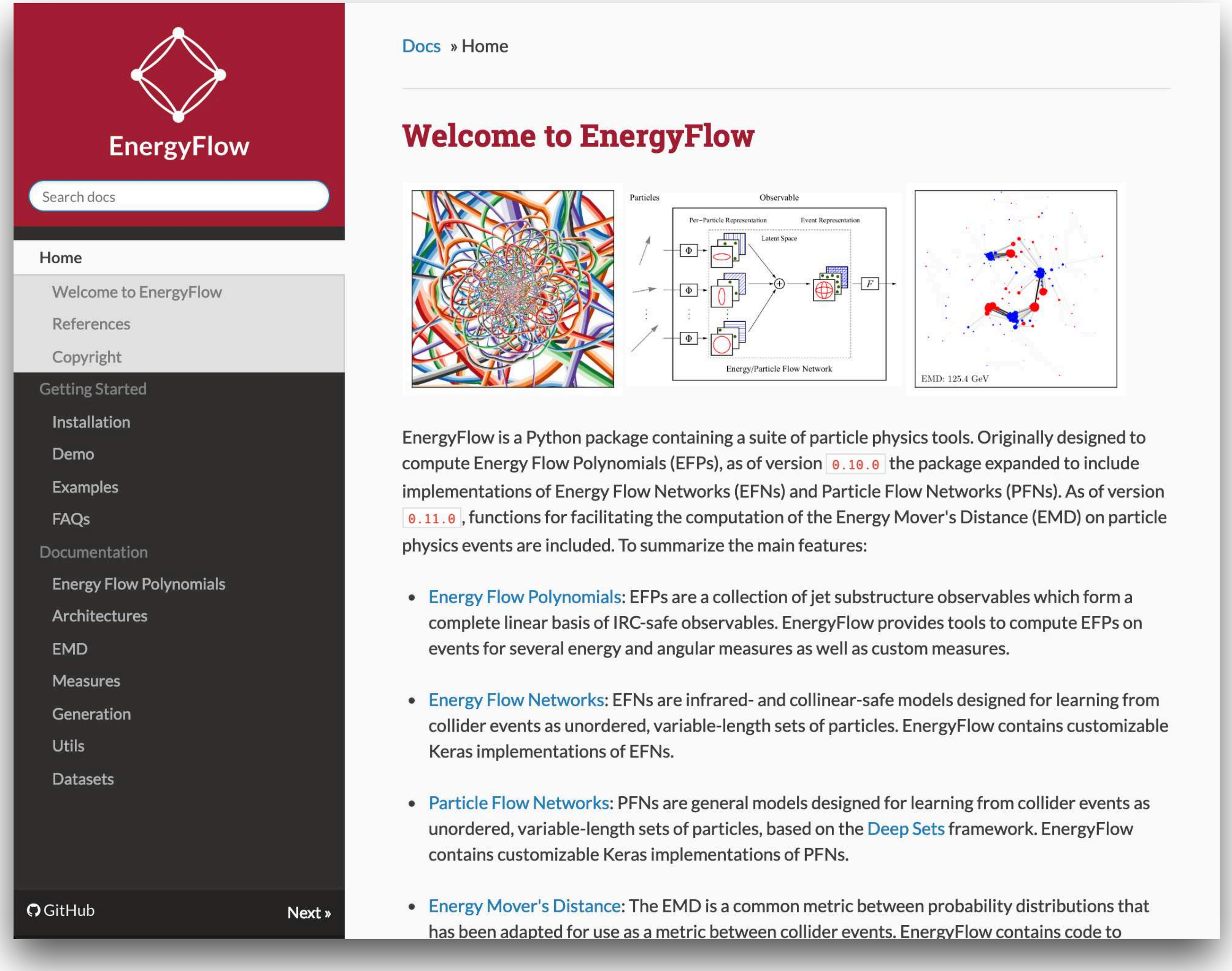
Backup Slides

EnergyFlow Python Package

Convenient functions for calculating EMD using the Python Optimal Transport library

Keras implementations of EFNs, PFNs, DNNs, CNNs, efficient EFP computation

Several detailed examples demonstrating common use cases and visualization procedures



EnergyFlow

Search docs

Home

- Welcome to EnergyFlow
- References
- Copyright

Getting Started

- Installation
- Demo
- Examples
- FAQs

Documentation

- Energy Flow Polynomials
- Architectures
- EMD
- Measures
- Generation
- Utils
- Datasets

GitHub

Next »

Docs » Home

Welcome to EnergyFlow

EnergyFlow is a Python package containing a suite of particle physics tools. Originally designed to compute Energy Flow Polynomials (EFPs), as of version `0.10.0` the package expanded to include implementations of Energy Flow Networks (EFNs) and Particle Flow Networks (PFNs). As of version `0.11.0`, functions for facilitating the computation of the Energy Mover's Distance (EMD) on particle physics events are included. To summarize the main features:

- Energy Flow Polynomials: EFPs are a collection of jet substructure observables which form a complete linear basis of IRC-safe observables. EnergyFlow provides tools to compute EFPs on events for several energy and angular measures as well as custom measures.
- Energy Flow Networks: EFNs are infrared- and collinear-safe models designed for learning from collider events as unordered, variable-length sets of particles. EnergyFlow contains customizable Keras implementations of EFNs.
- Particle Flow Networks: PFNs are general models designed for learning from collider events as unordered, variable-length sets of particles, based on the Deep Sets framework. EnergyFlow contains customizable Keras implementations of PFNs.
- Energy Mover's Distance: The EMD is a common metric between probability distributions that has been adapted for use as a metric between collider events. EnergyFlow contains code to

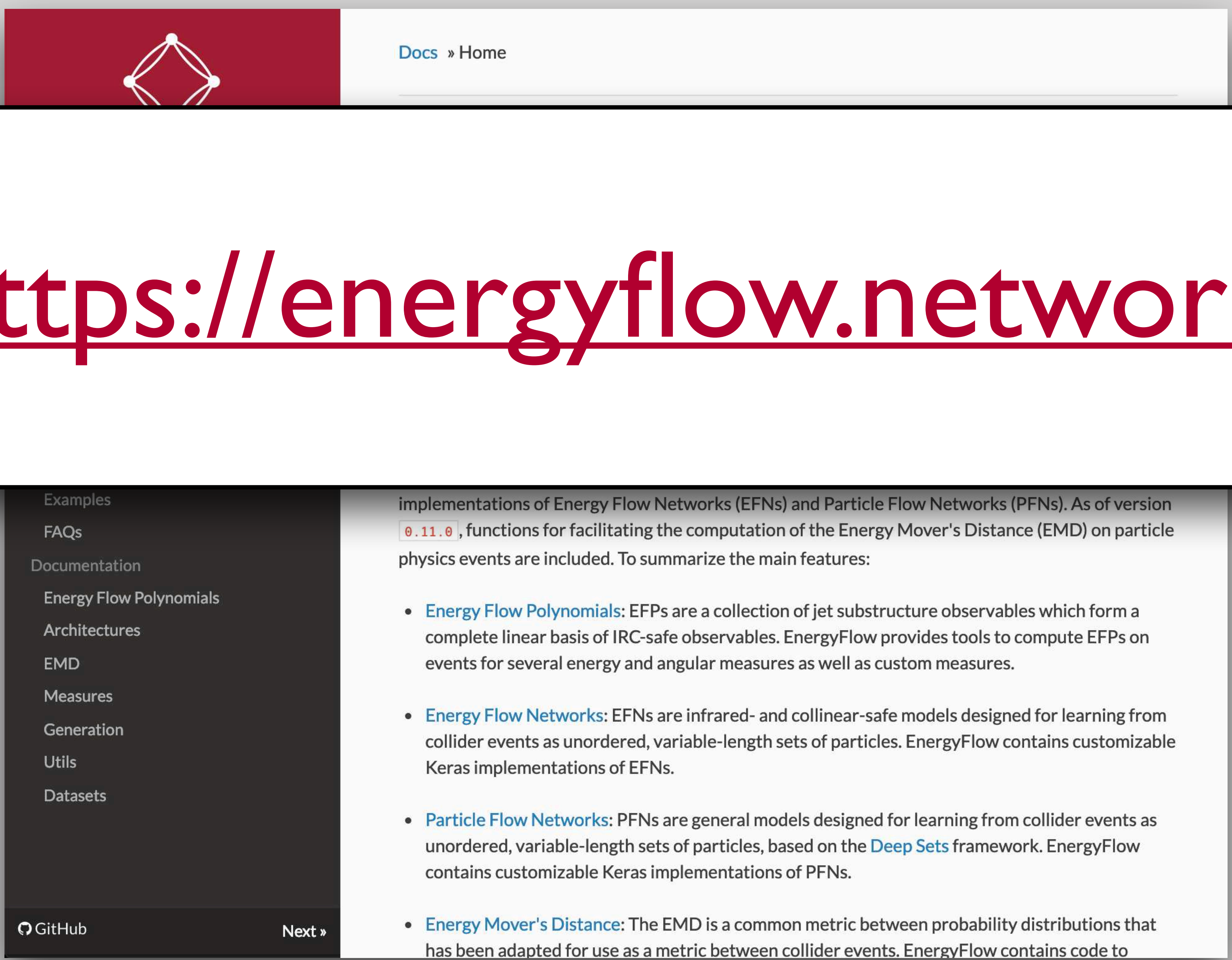
EnergyFlow Python Package

Convenient functions for calculating EMD using the Python Optimal Transport library

Keras implementations of EFNs, PFNs, DNNs, CNNs, efficient EFP computation

Several detailed examples demonstrating common use cases and visualization procedures

https://energyflow.network



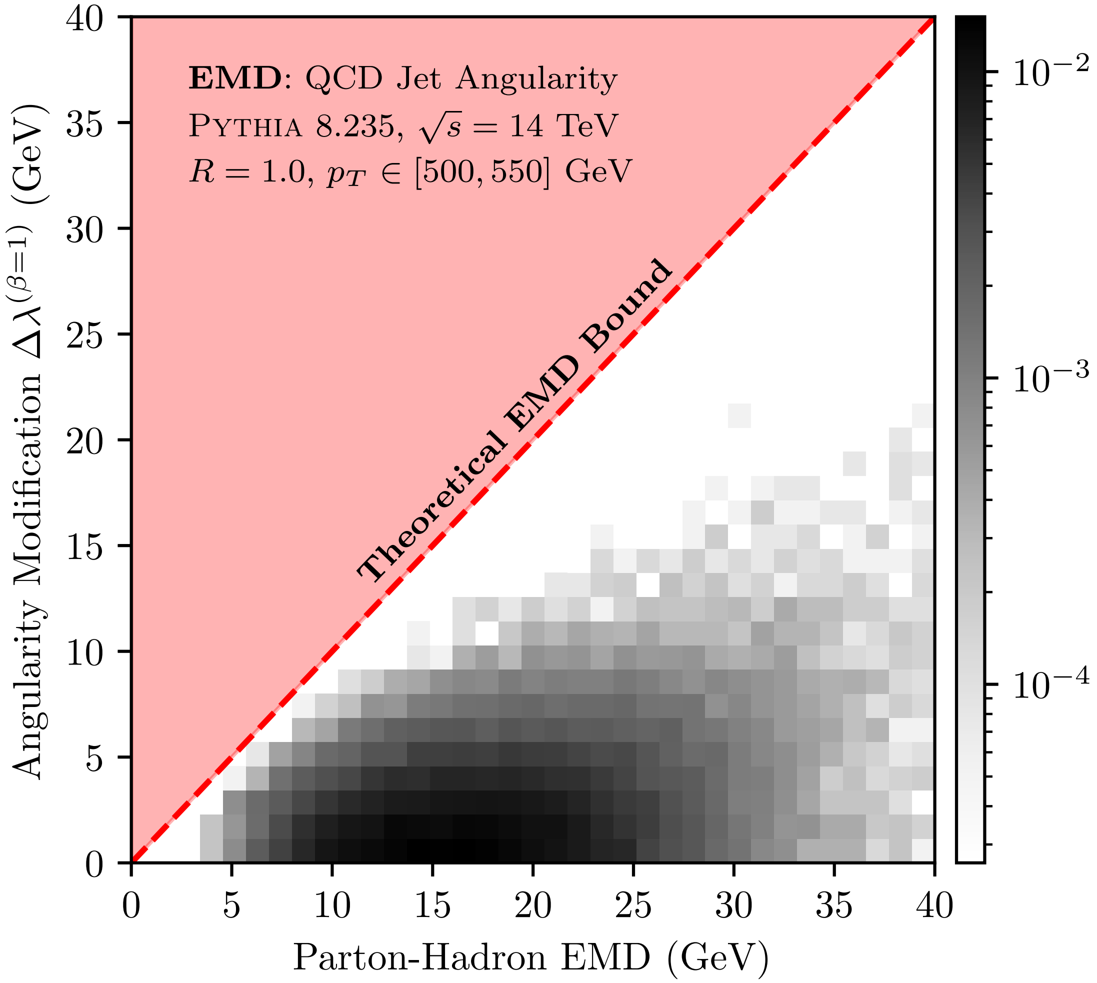
implementations of Energy Flow Networks (EFNs) and Particle Flow Networks (PFNs). As of version `0.11.0`, functions for facilitating the computation of the Energy Mover's Distance (EMD) on particle physics events are included. To summarize the main features:

- Energy Flow Polynomials: EFPs are a collection of jet substructure observables which form a complete linear basis of IRC-safe observables. EnergyFlow provides tools to compute EFPs on events for several energy and angular measures as well as custom measures.
- Energy Flow Networks: EFNs are infrared- and collinear-safe models designed for learning from collider events as unordered, variable-length sets of particles. EnergyFlow contains customizable Keras implementations of EFNs.
- Particle Flow Networks: PFNs are general models designed for learning from collider events as unordered, variable-length sets of particles, based on the Deep Sets framework. EnergyFlow contains customizable Keras implementations of PFNs.
- Energy Mover's Distance: The EMD is a common metric between probability distributions that has been adapted for use as a metric between collider events. EnergyFlow contains code to

Quantifying Event Modifications – e.g. Hadronization

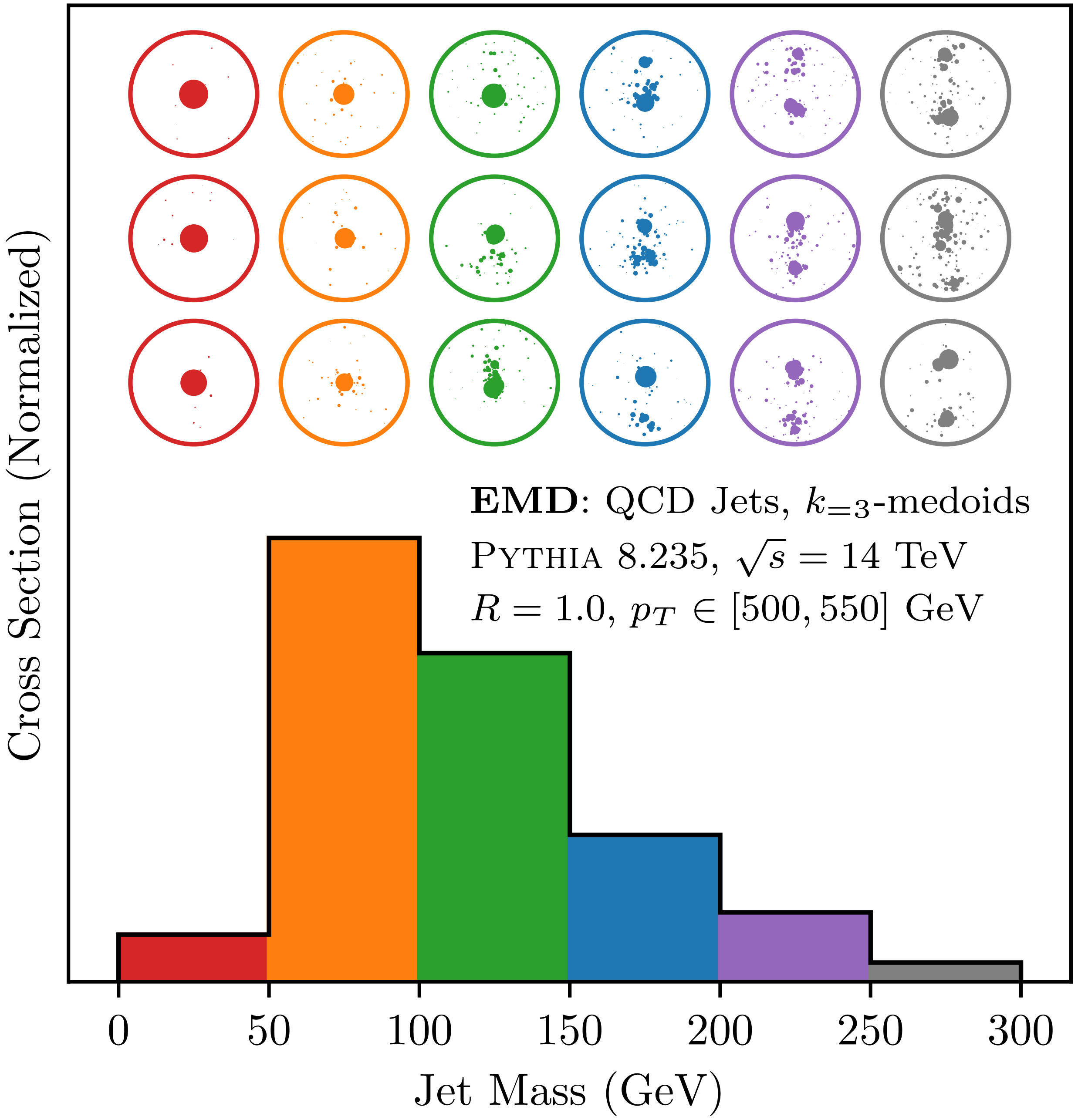
Hadronization affects all hadronic final states and yet is poorly understood

$$\mathrm{EMD}(\mathcal{E},\mathcal{E}') \geq \frac{1}{RL}\left|\sum_i E_i \Phi(\hat{p}_i) - \sum_j E'_j \Phi(\hat{p}'_j)\right| = \frac{1}{RL}\,|\mathcal{O}(\mathcal{E}) - \mathcal{O}(\mathcal{E}')|$$



Finding Representative Events

K-medoids finds representative events, for instance in different histogram bins



Infrared and Collinear (IRC) Safety

QCD has soft and collinear divergences associated with gluon radiation

$$dP_{i\to ig} \simeq \frac{2\alpha_s}{\pi} C_a \frac{d\theta}{\theta} \frac{dz}{z}$$

$$C_q = C_F = 4/3$$
$$C_g = C_A = 3$$

KLN Theorem: IRC safety of an observable is sufficient to guarantee that soft/collinear divergences cancel at each order in perturbation theory

Infrared (IR) safety – observable is unchanged under addition of a soft particle

$$S(\{p_1^\mu, \dots, p_M^\mu\}) = S(\{p_1^\mu, \dots, (1-\lambda)p_M^\mu, \lambda p_M^\mu\}), \quad \forall\, \lambda \in [0,1]$$

Collinear (C) safety – observable is unchanged under a collinear splitting of a particle

$$S(\{p_1^\mu, \dots, p_M^\mu\}) = \lim_{\epsilon\to 0} S(\{p_1^\mu, \dots, p_M^\mu, \epsilon p_{M+1}^\mu\}), \quad \forall\, p_{M+1}^\mu$$

IRC safety is a key theoretical *and experimental* property of observables