Metric Methods with Open Collider Data

Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, Jesse Thaler

Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Introduction

Machine learning and particle physics are on a collision course, producing exciting new ideas.

Exploring **public collider data** inspired new fundamental questions, with answers coming from an unlikely place: **optimal transport**.

Q: What's the "distance" between two collisions?

A: The "work" to rearrange one into another!

Equipping collider data with a metric unlocks new unsupervised and visualization techniques.

Remarkably, this connection sheds new light on fundamental concepts in quantum field theory. Optimal transport provides a geometric foundation for 60 years of collider techniques!

CMS Open Data

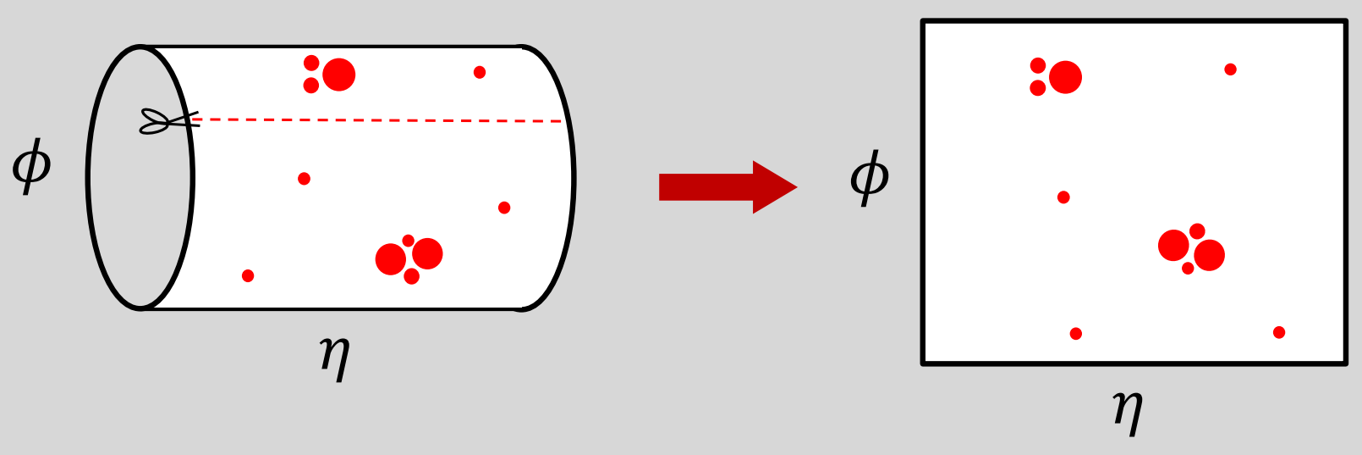
The CMS Experiment at the Large Hadron Collider (LHC) has begun publicly releasing research-grade open collider data.

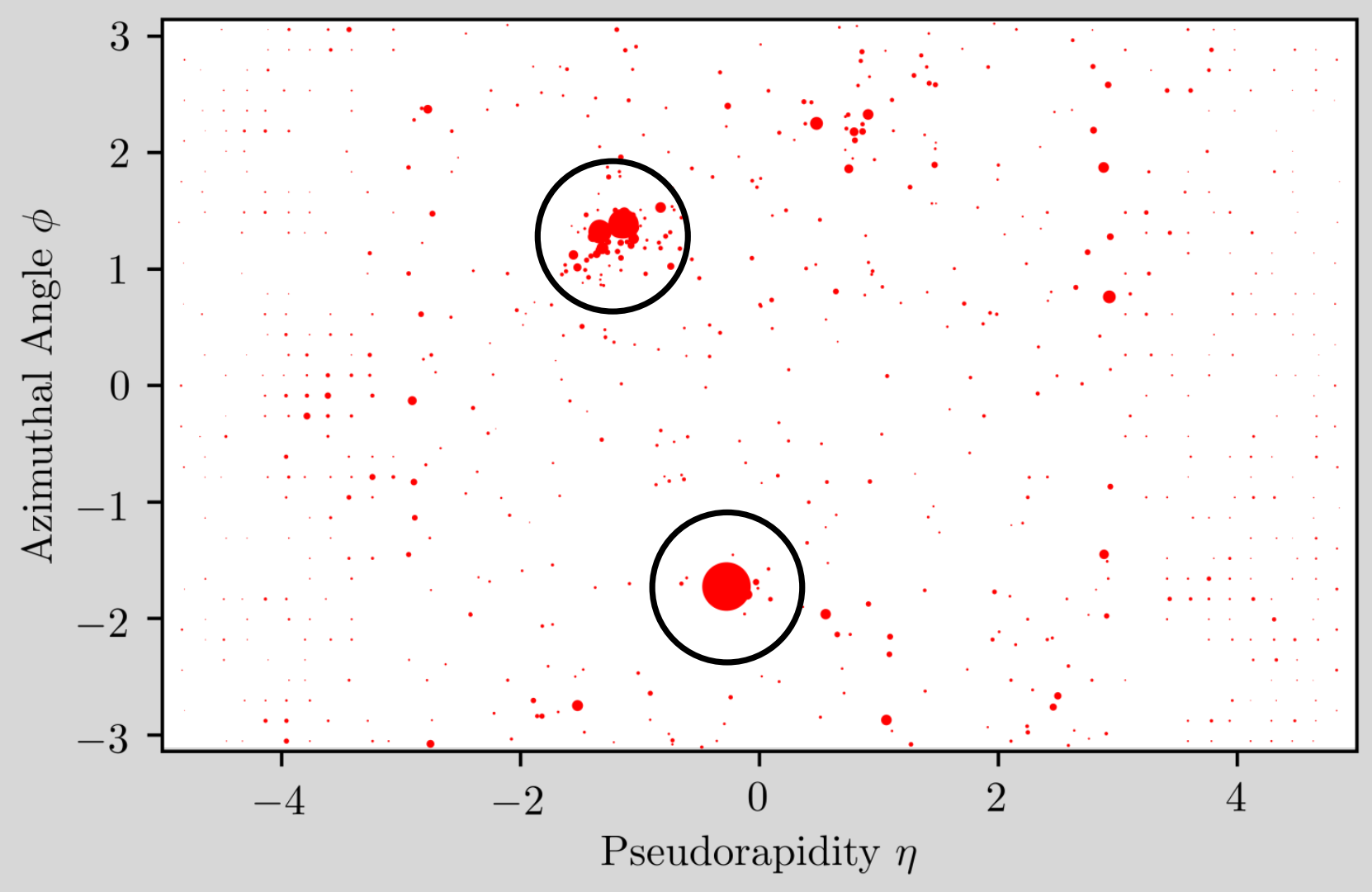
opendata CERN

opendata.cern.ch

Getting started with CMS Open Data is easy!

1. **Download** an "Analysis Object Data" file.
2. **Read** in the file with the uproot package.
3. **Start** looking at events!

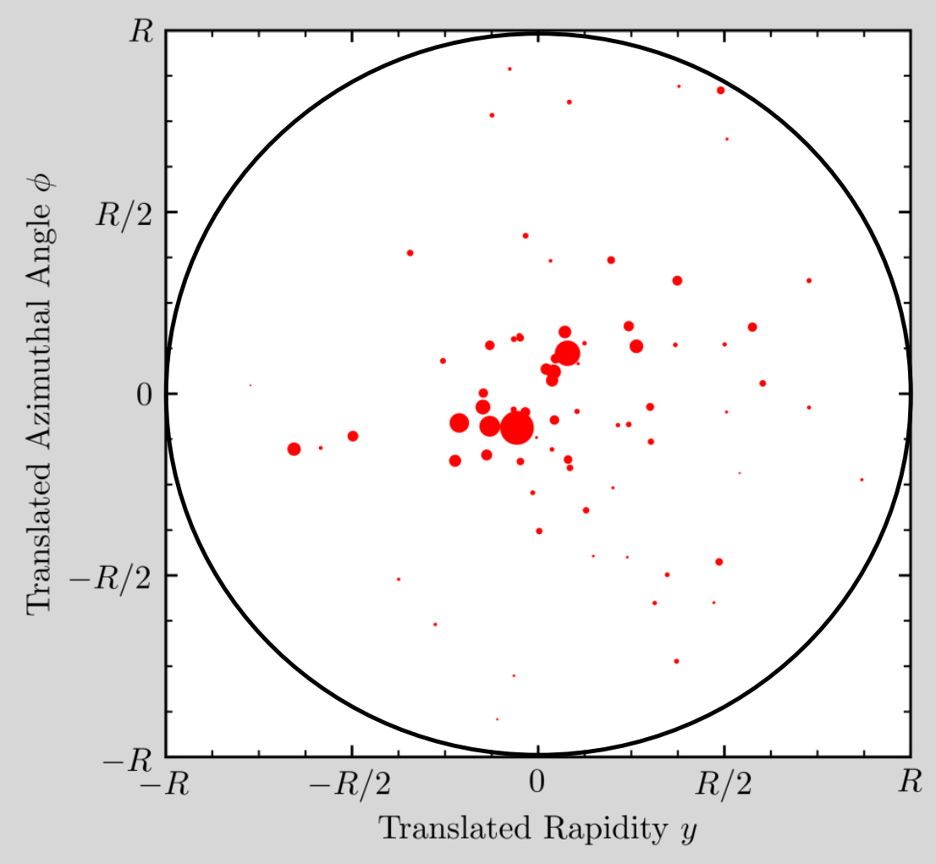




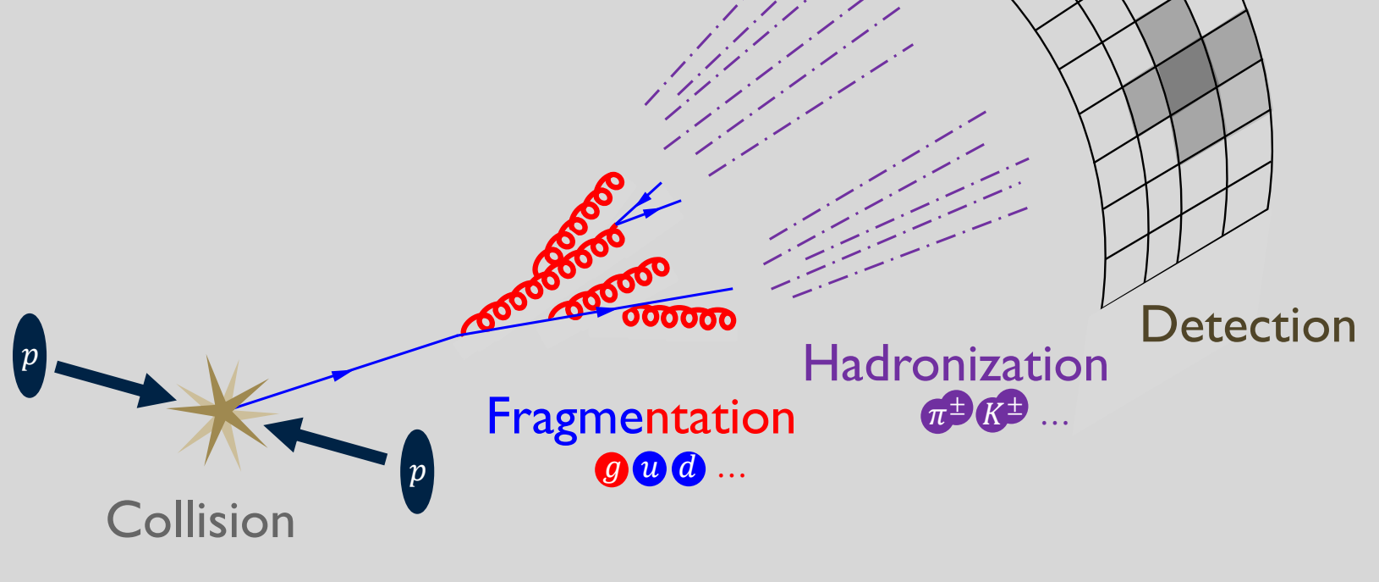
A real collision event recorded by the CMS detector.

Jets and their Substructure

Jets are collimated sprays of particles that originate from high energy quarks and gluons.

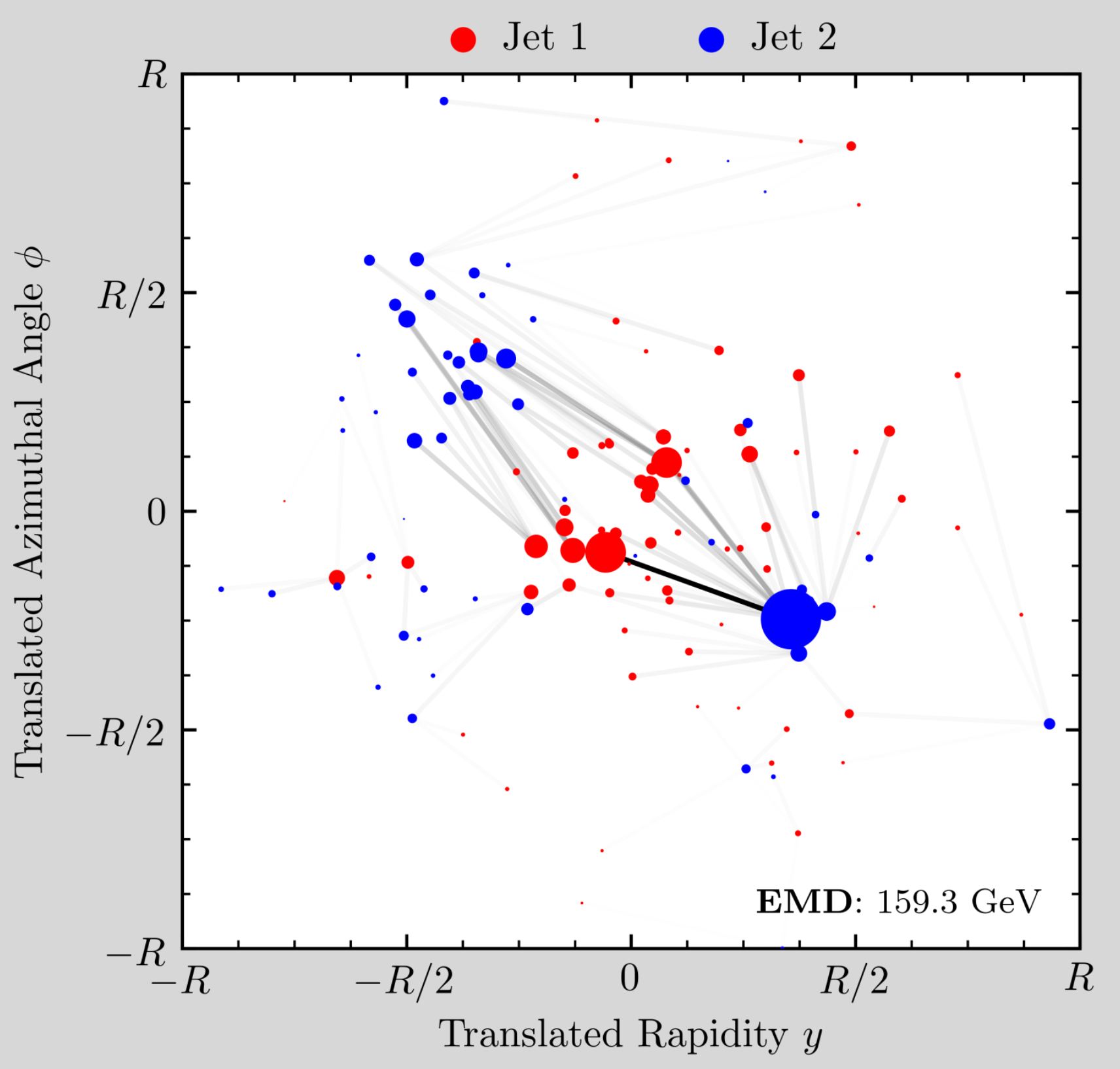


Jets and their substructure are crucial for understanding the strong force and searching for new physics at the LHC.



A Metric for Collider Data

When are two particle collisions similar? Or when are two jets similar?



Natural question with no satisfying answer in physics.

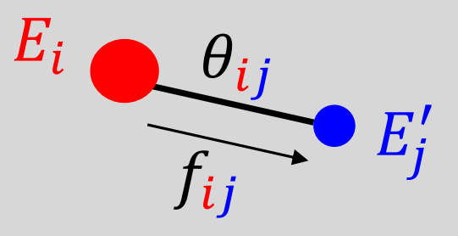
Image-based pixel comparisons are unstable under tiny perturbations of the particles.

Observable (i.e. feature) comparisons can have zero distance for very different events or jets.

The Earth (or Energy) Mover's Distance (EMD) provides a natural answer. Solving for the EMD is an optimal transport problem.

The "work" to rearrange one event into another!

$$\text{EMD}(\mathcal{E},\mathcal{E}') = \min_{\{f\}} \sum_{i=1}^{M}\sum_{j=1}^{M'} f_{ij}\frac{\theta_{ij}}{R} + \left|\sum_{i=1}^{M} E_i - \sum_{j=1}^{M'} E'_j\right|$$



Collider Physics and Optimal Transport

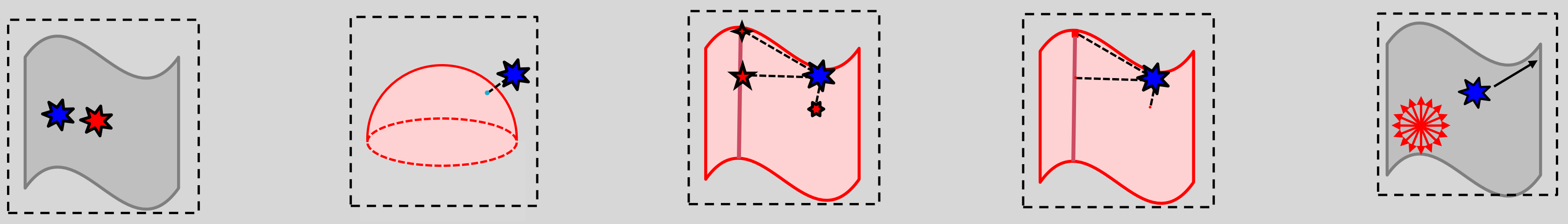
Six decades of collider techniques can be naturally cast as geometry in the "space of events" with EMD.

Smooth functions of energy distribution are finite in QFT

Event shapes as distances to the 2-particle manifold

Jets are N-particle event approximations

Subtract pileup as a uniform distribution

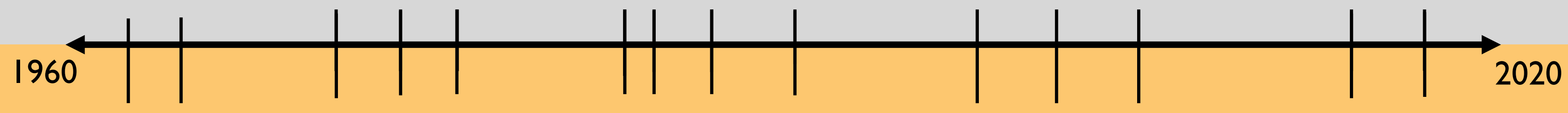


$$\text{EMD}(\mathcal{E},\mathcal{E}') < \delta \rightarrow |\mathcal{O}(\mathcal{E}) - \mathcal{O}(\mathcal{E}')| < \epsilon$$

$$t(\mathcal{E}) = \min_{|\mathcal{E}'|=2} \text{EMD}(\mathcal{E},\mathcal{E}')$$

$$\mathcal{J}(\mathcal{E}) = \operatorname{argmin}_{|\mathcal{E}'|=N} \text{EMD}(\mathcal{E},\mathcal{E}')$$

$$\mathcal{E} - \mathcal{U}$$



1960s
Taming Infinites

1970s-1980s
The Shape of Events

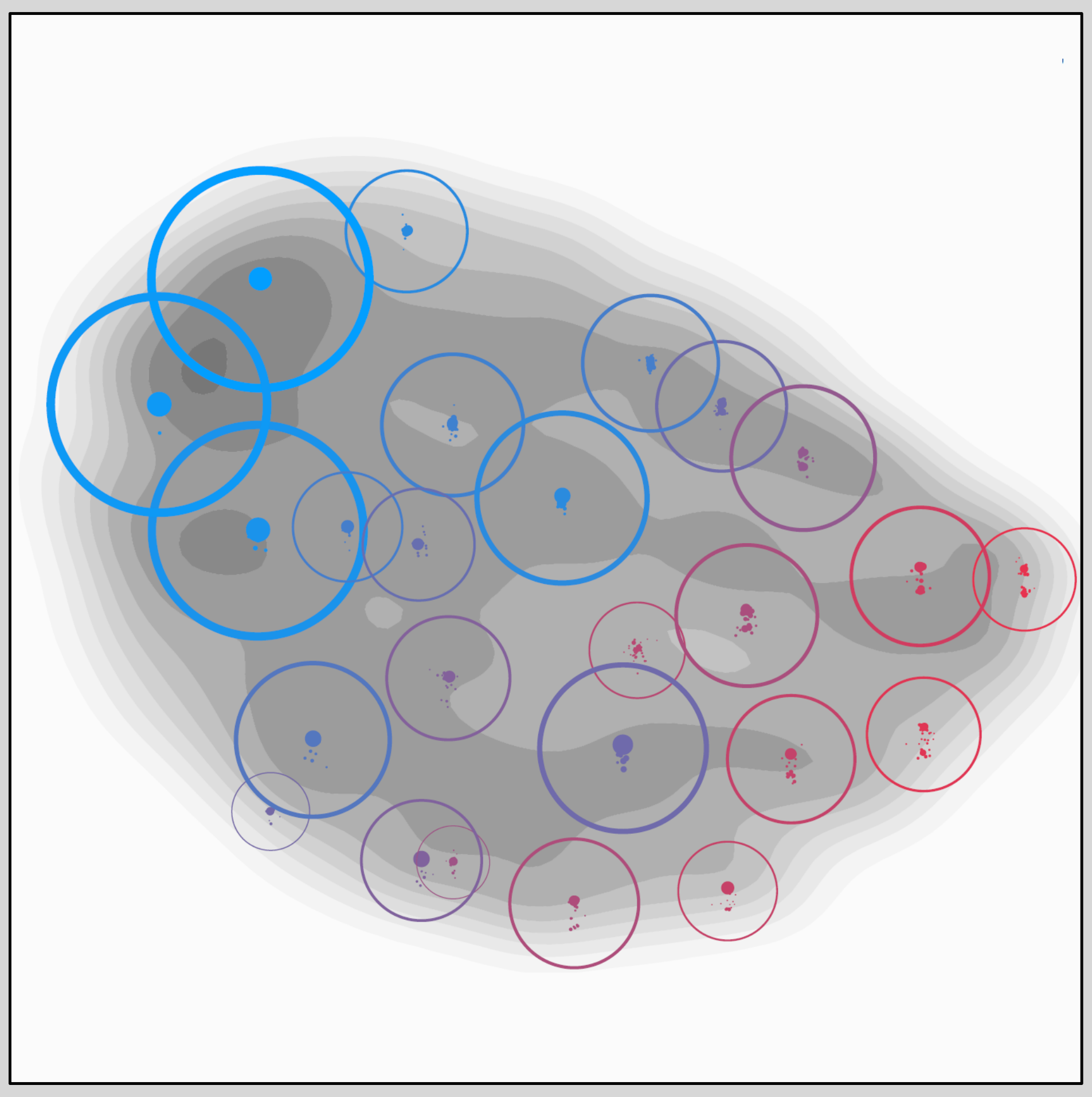
1990-2000s
Jet Clustering

2010s
Jet Substructure

2010s
Pileup Subtraction

Exploring the Space of Jets

The "space of jets" can be visualized by embedding the jet dataset with t-SNE.



25-medoid jets shown, sized by importance.

A peak of one-prong jets with a tail of two-pronged jets naturally emerges.

A natural consequence of the QCD splitting function. The rate for a quark to emit a gluon of energy E at angle θ:

$$P(E,\theta)\, dE\, d\theta = \frac{8\alpha_s}{3\pi}\frac{dE}{E}\frac{d\theta}{\theta}$$

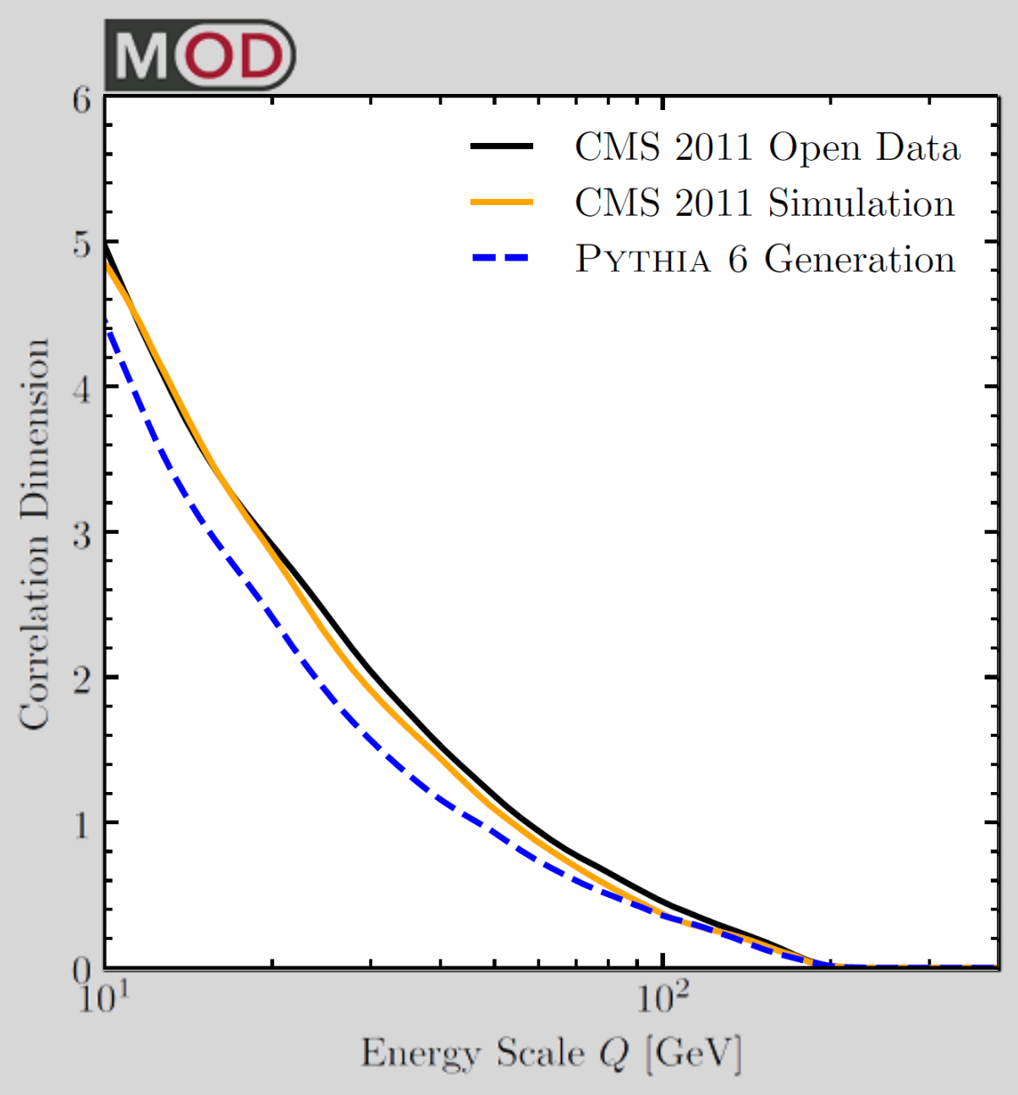
With infrared and collinear divergences.

Jets with balanced prongs are above. Jets with asymmetric prongs are below.

The Fractal Dimension of Jets

The correlation (fractal) dimension of the dataset is defined with pairwise distances:

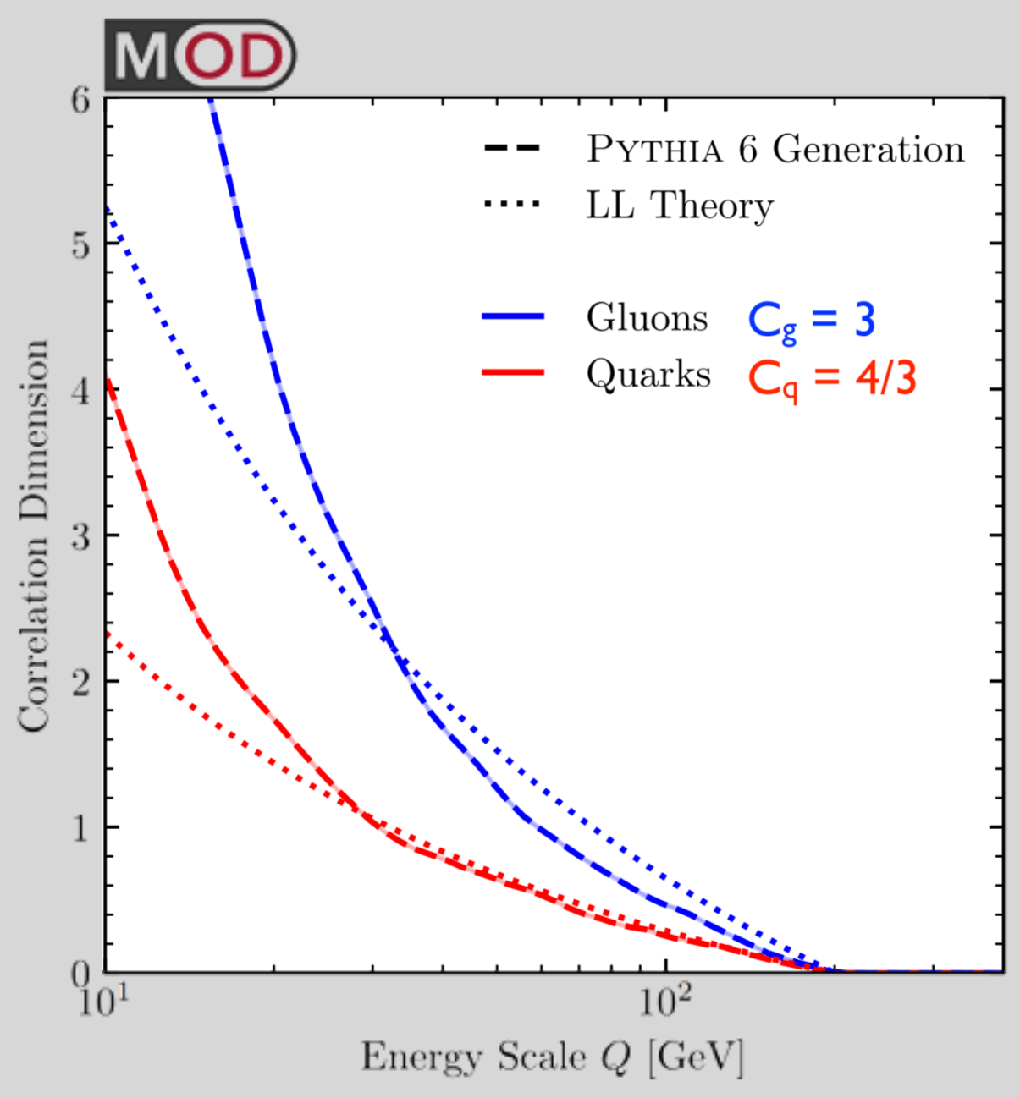
$$\dim(Q) = Q\frac{\partial}{\partial Q}\ln \sum_{i=1}^{N}\sum_{j=1}^{N} \Theta[\text{EMD}(\mathcal{E}_i,\mathcal{E}_j) < Q]$$



Jets become more complex at lower energies.

Jets are "more than fractal" since the correlation dimension doesn't level off.

We can begin to theoretically calculate it!



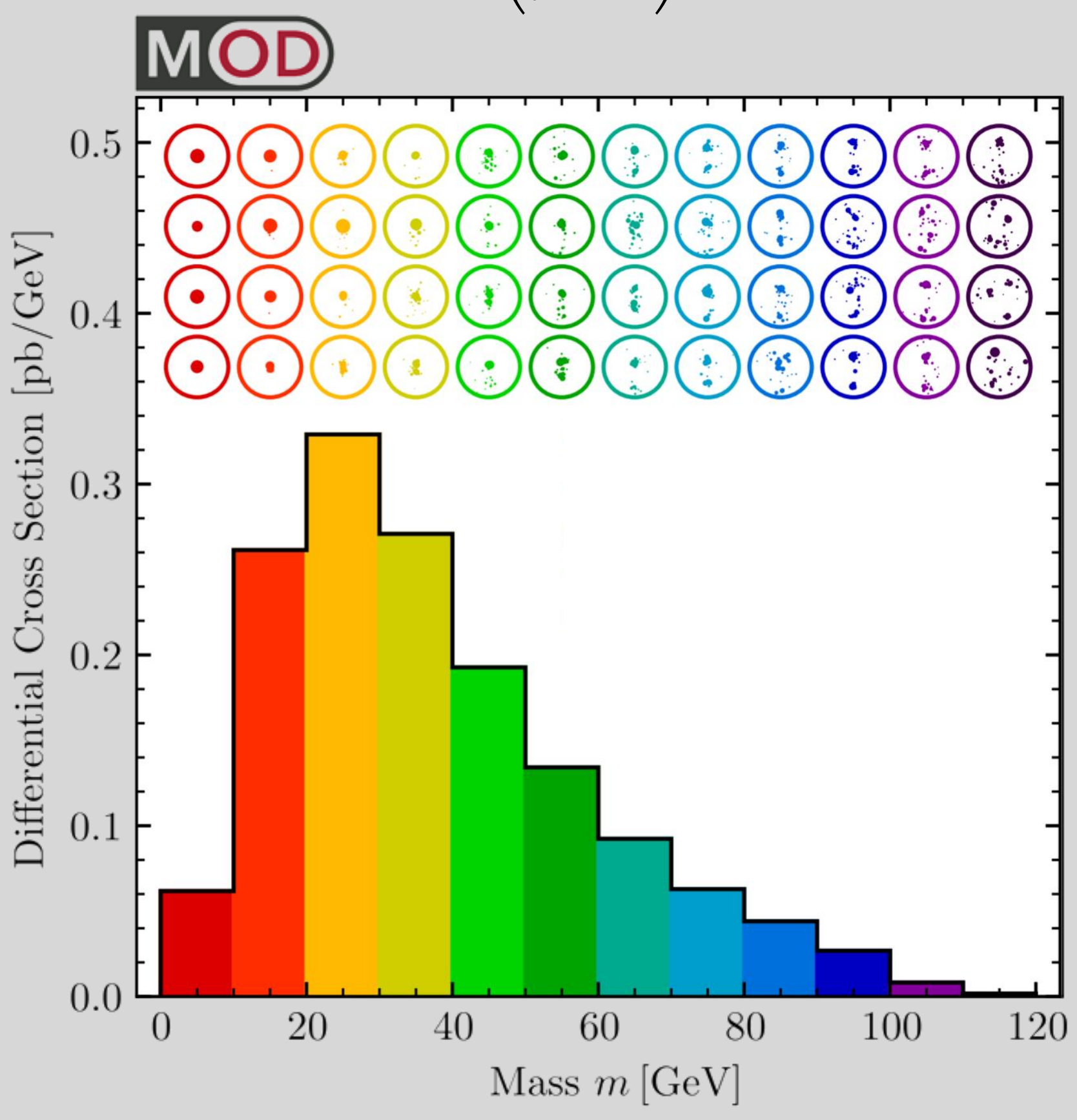
Visualizing Substructure

The substructure of jets is traditionally probed by computing histograms of "observables".

The most representative jets in each bin, determined via the metric, illustrate the physics that governs the observable.

The Jet Mass m probes how "wide" the jet is.

$$m^2 = \left(\sum_{i=1}^{M} p_i^\mu\right)^2$$



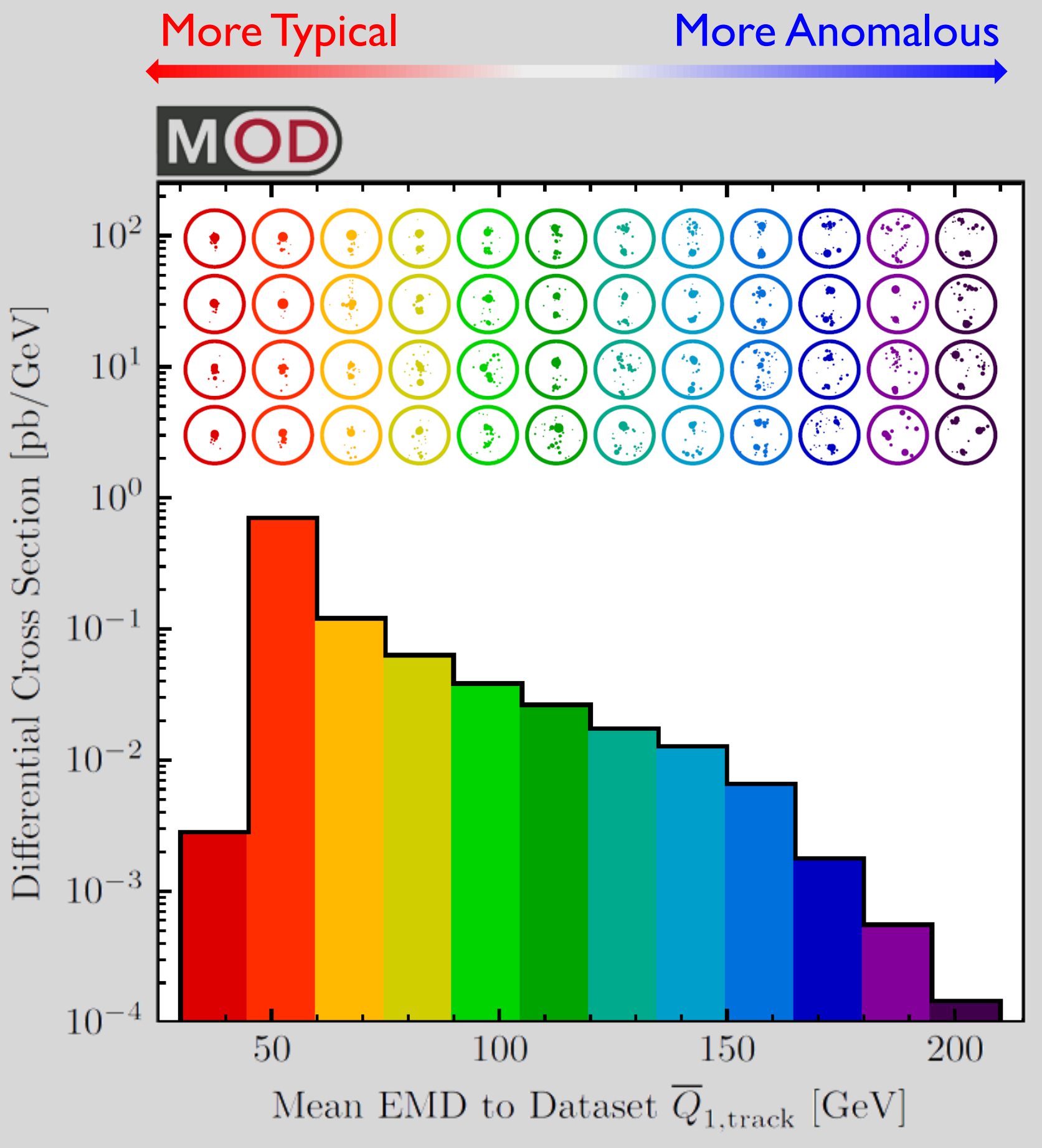
Towards Anomaly Detection

The lack of new physics at the LHC has stimulated interest in model-independent anomaly detection.

Using the metric, we can identify the "most typical" and "least typical" jets based on their average distance to the dataset.

$$\bar{Q}(\mathcal{E}) = \frac{1}{N}\sum_{i=1}^{N} \text{EMD}\,(\mathcal{E},\mathcal{E}_i)$$

A step towards anomaly detection at the LHC.



Selected References

[1] CERN Open Data Portal. opendata.cern.ch

[2] Patrick T. Komiske, Eric M. Metodiev, Jesse Thaler. Metric Space of Collider Events. PRL **123** 041801, 2019.

[3] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, Jesse Thaler. Exploring the Space of Jets with CMS Open Data. arXiv:1908.08542

[4] Patrick T. Komiske, Eric M. Metodiev, Jesse Thaler. The Hidden Geometry of Particle Collisions. *To appear.*

Contact Information

Eric M. Metodiev
email: eric.metodiev@gmail.com
web: ericmetodiev.com