

The Hidden Geometry of Particle Collisions

Patrick T. Komiske III

Massachusetts Institute of Technology
Center for Theoretical Physics

Based on work with Eric Metodiev and Jesse Thaler
[1902.02346](#) (PRL)
[2004.04159](#) (JHEP)

BSM Pandemic Double Feature

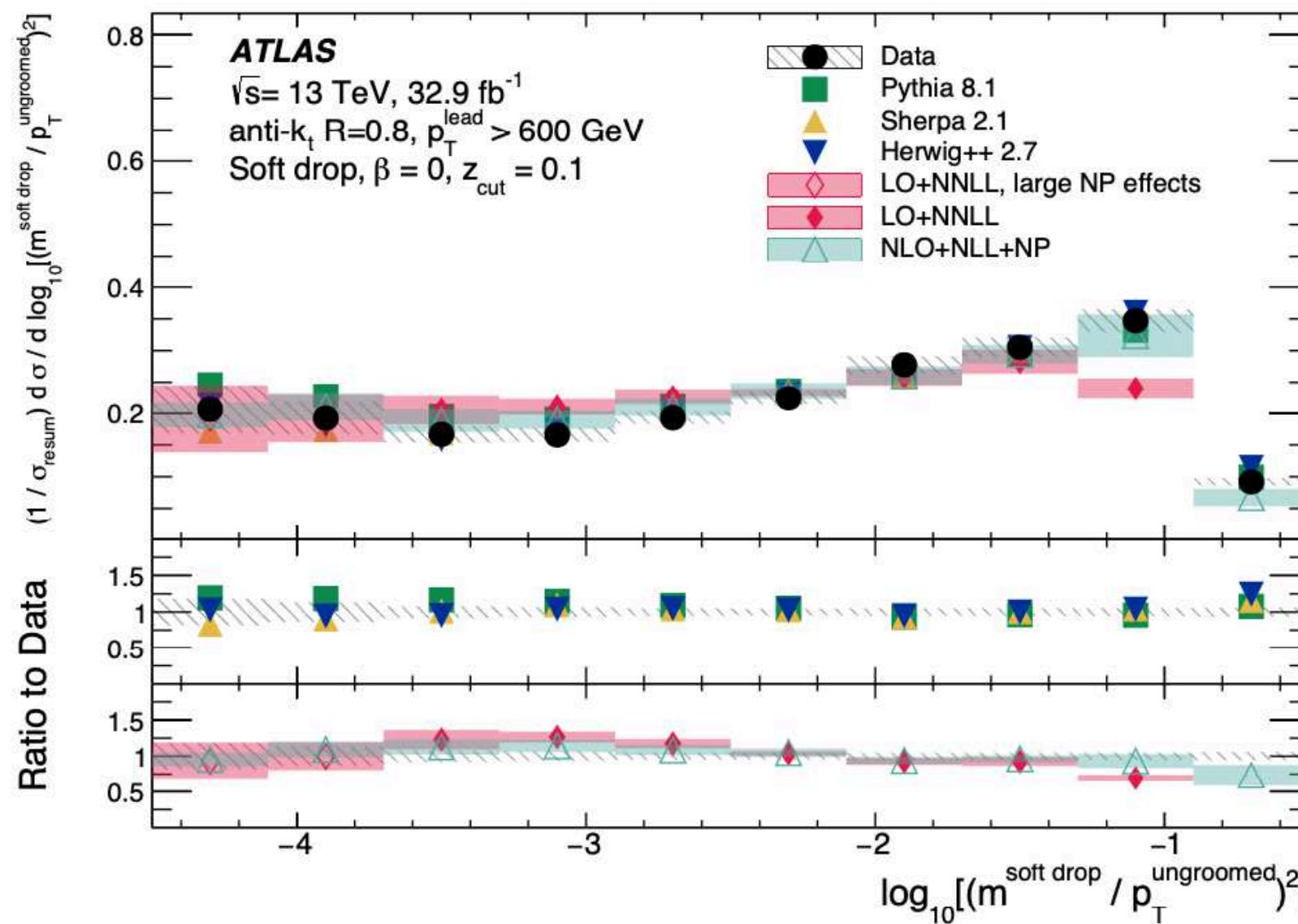
December 1, 2020

Developing New Analysis Frameworks

Particle theory *makes predictions and invents new models*

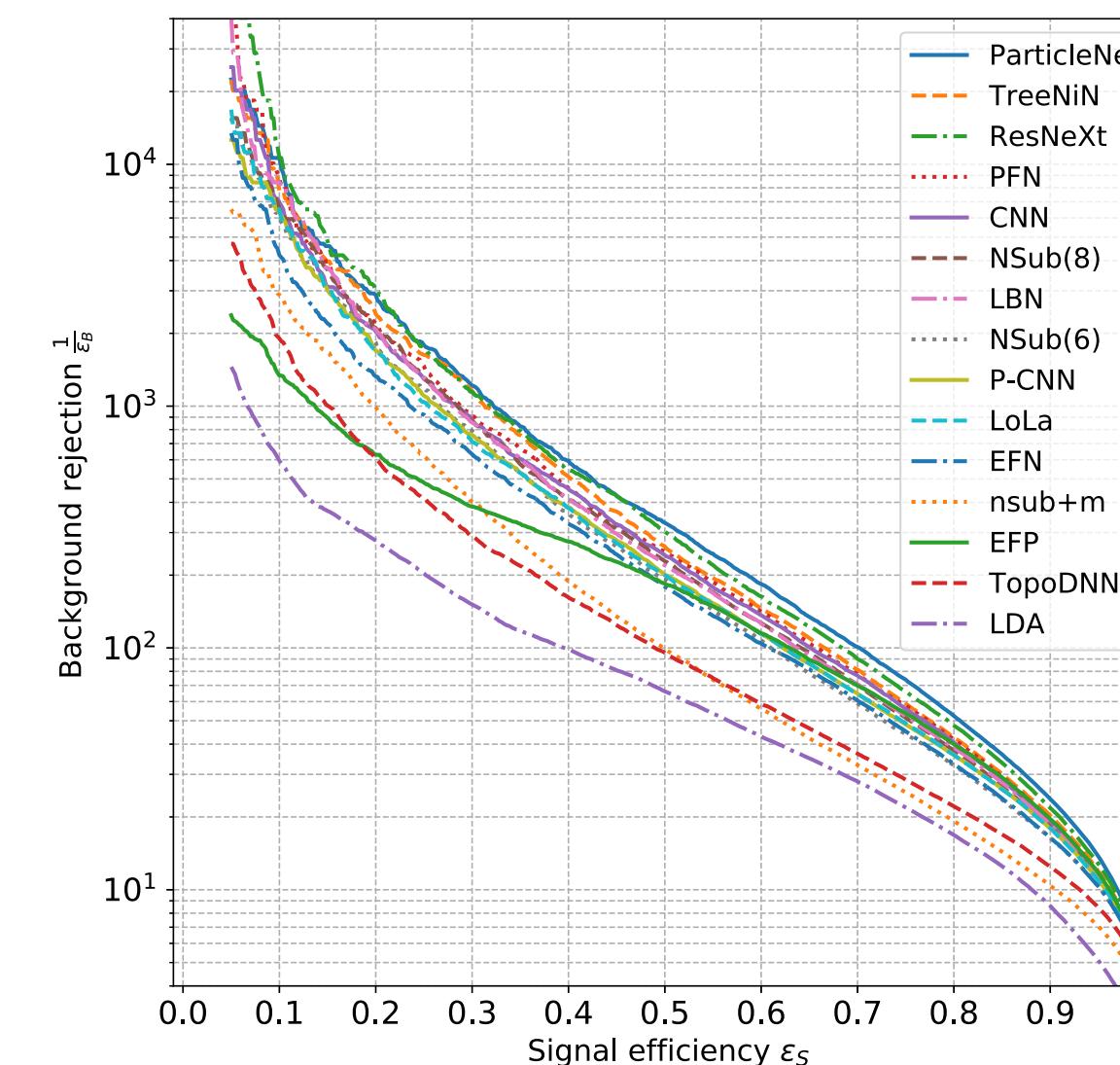
Particle experiment *conducts measurements and searches for new physics*

Soft Drop jet mass measurement



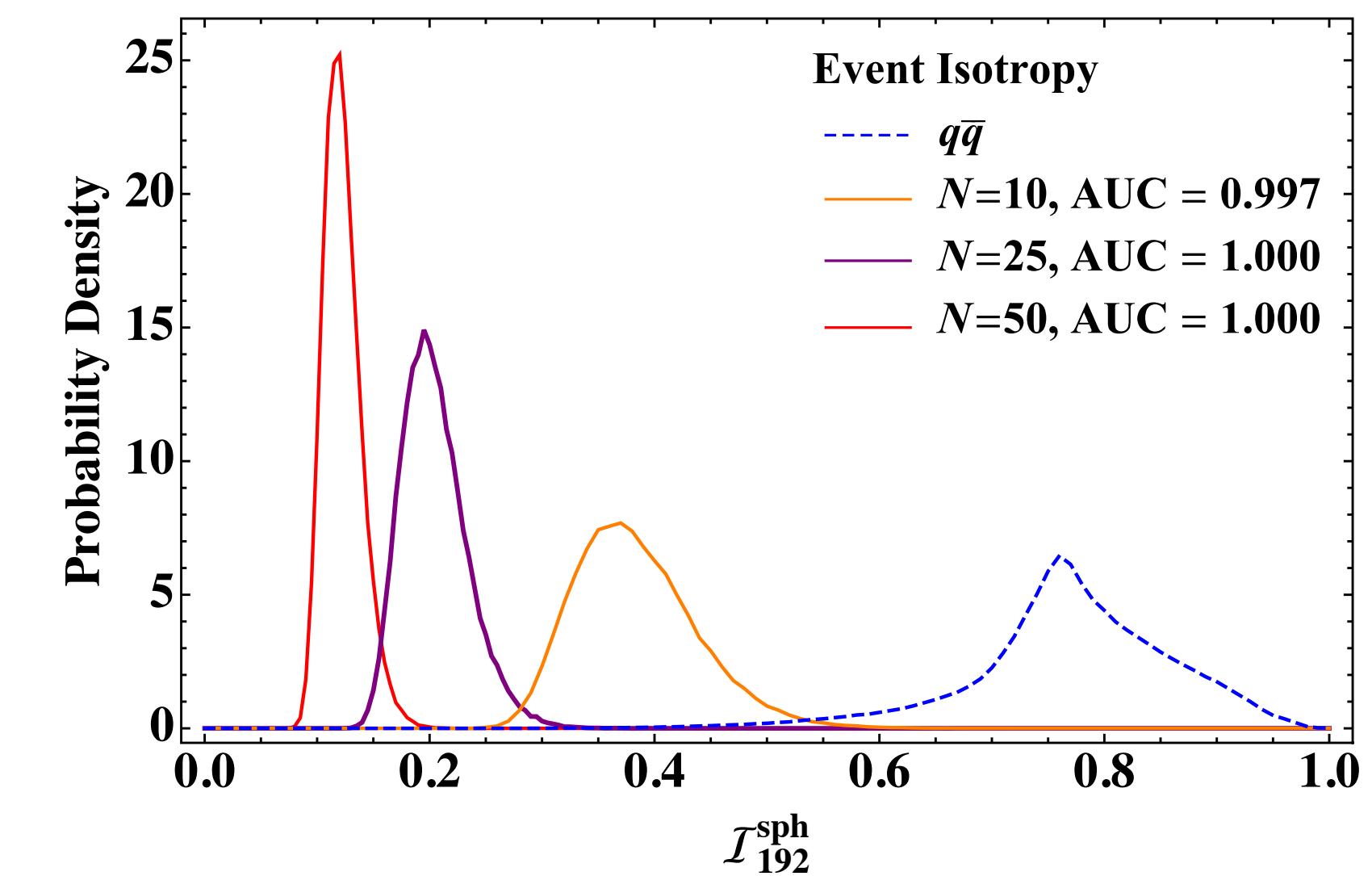
[Larkoski, Marzani, Soyez, Thaler, [JHEP 2014](#);
ATLAS, [PRL 2018](#)]

Comparison of ML top taggers



[Kasieczka, Plehn, et al., [1902.09914](#);
using PTK, Metodiev, Thaler, [1810.05165](#), and others]

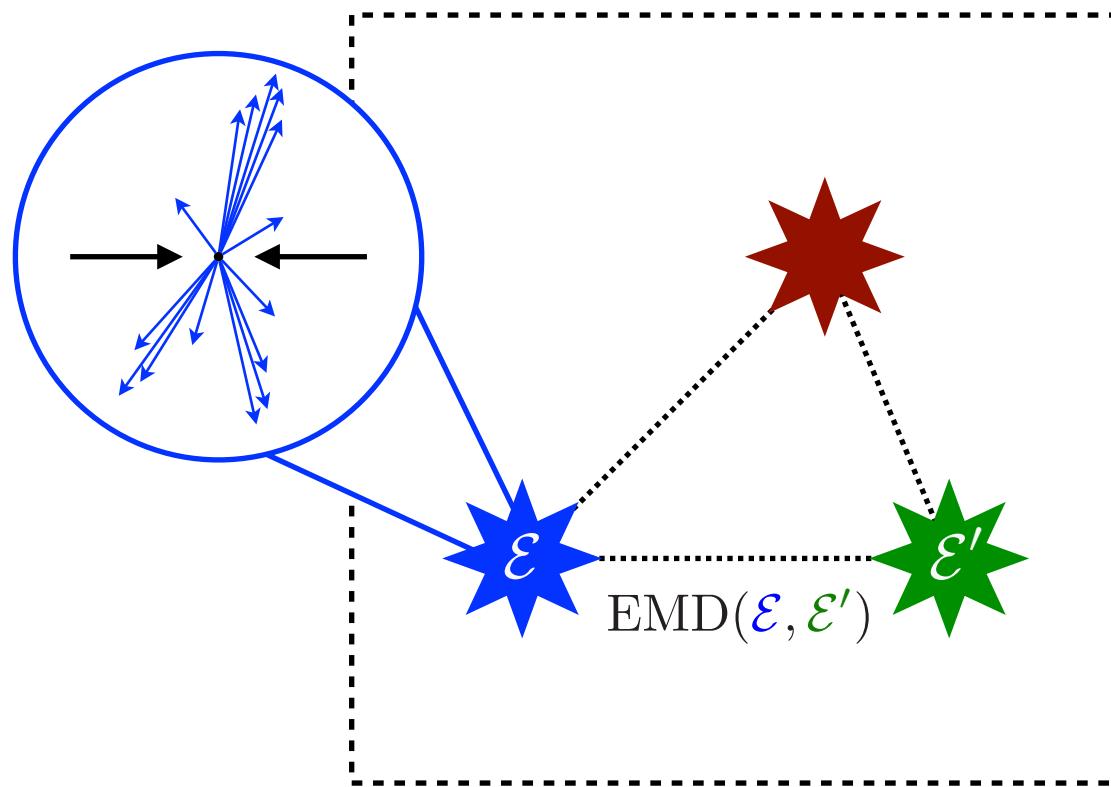
New observables sensitive to new signals



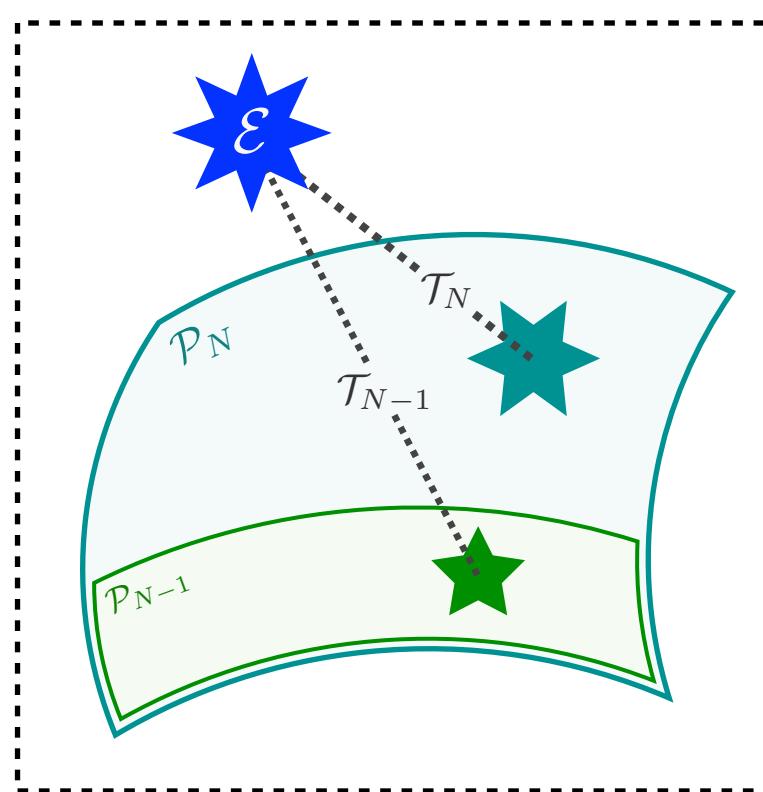
[Cesarotti, Thaler, [JHEP 2020](#);
utilizing PTK, Metodiev, Thaler, [PRL 2019](#)]

Twenty more years of the LHC + any future collider(s)

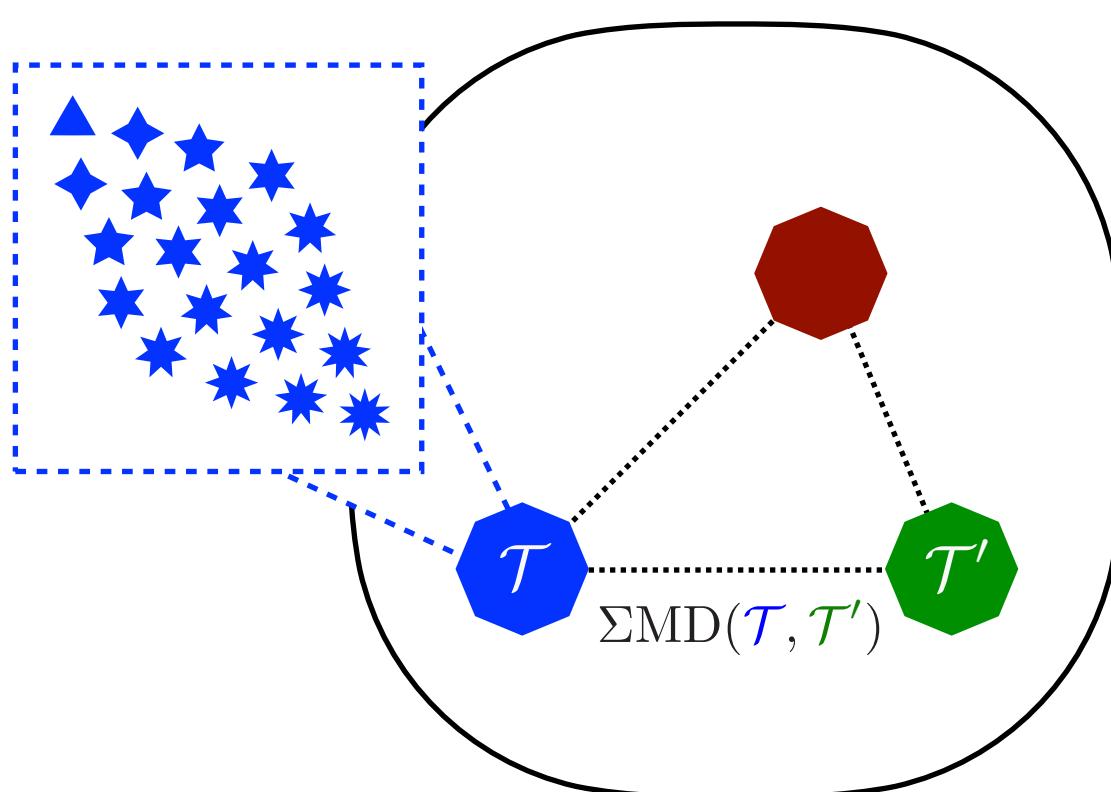
Maximizing physics potential will require insights from data science and ML



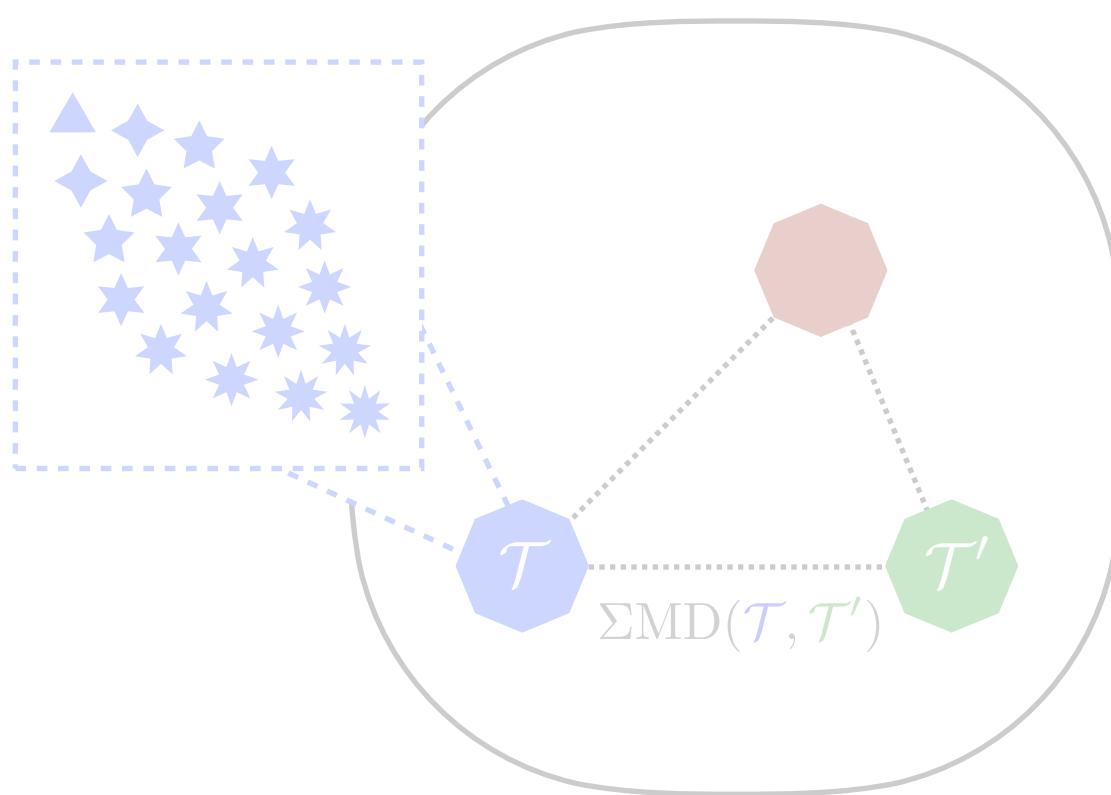
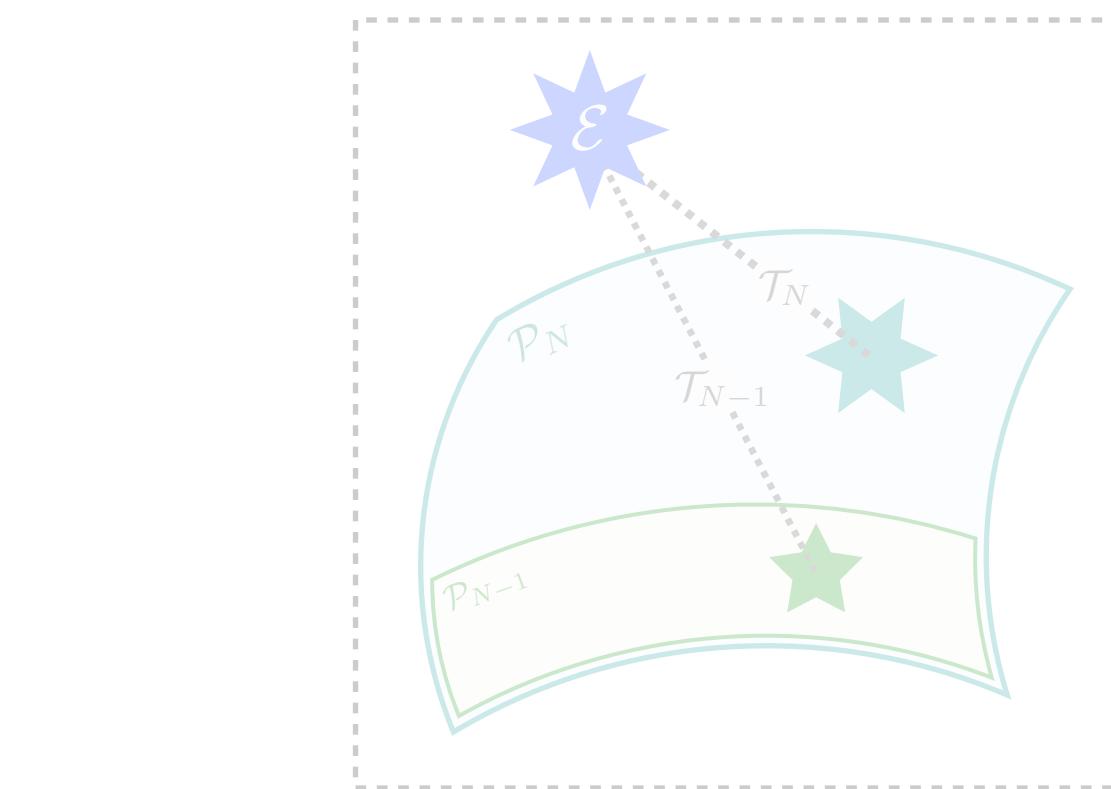
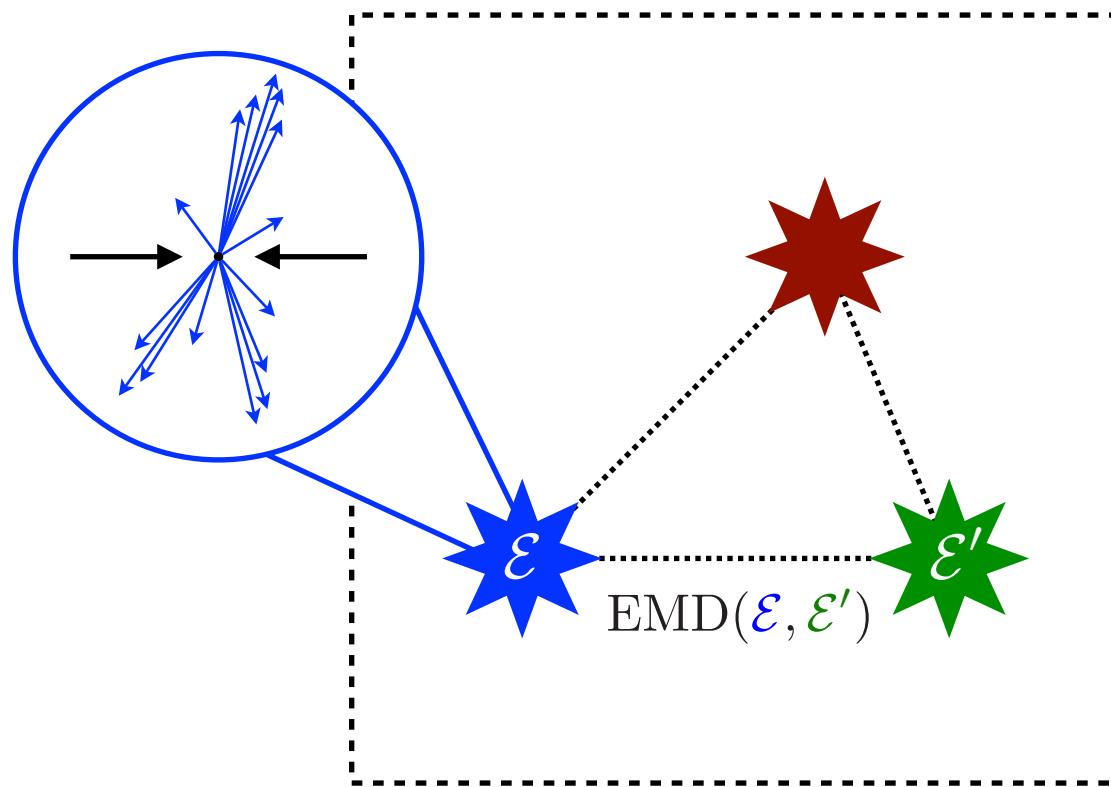
The (Metric) Space of Events



Revealing Hidden Geometry



[Theory Space]



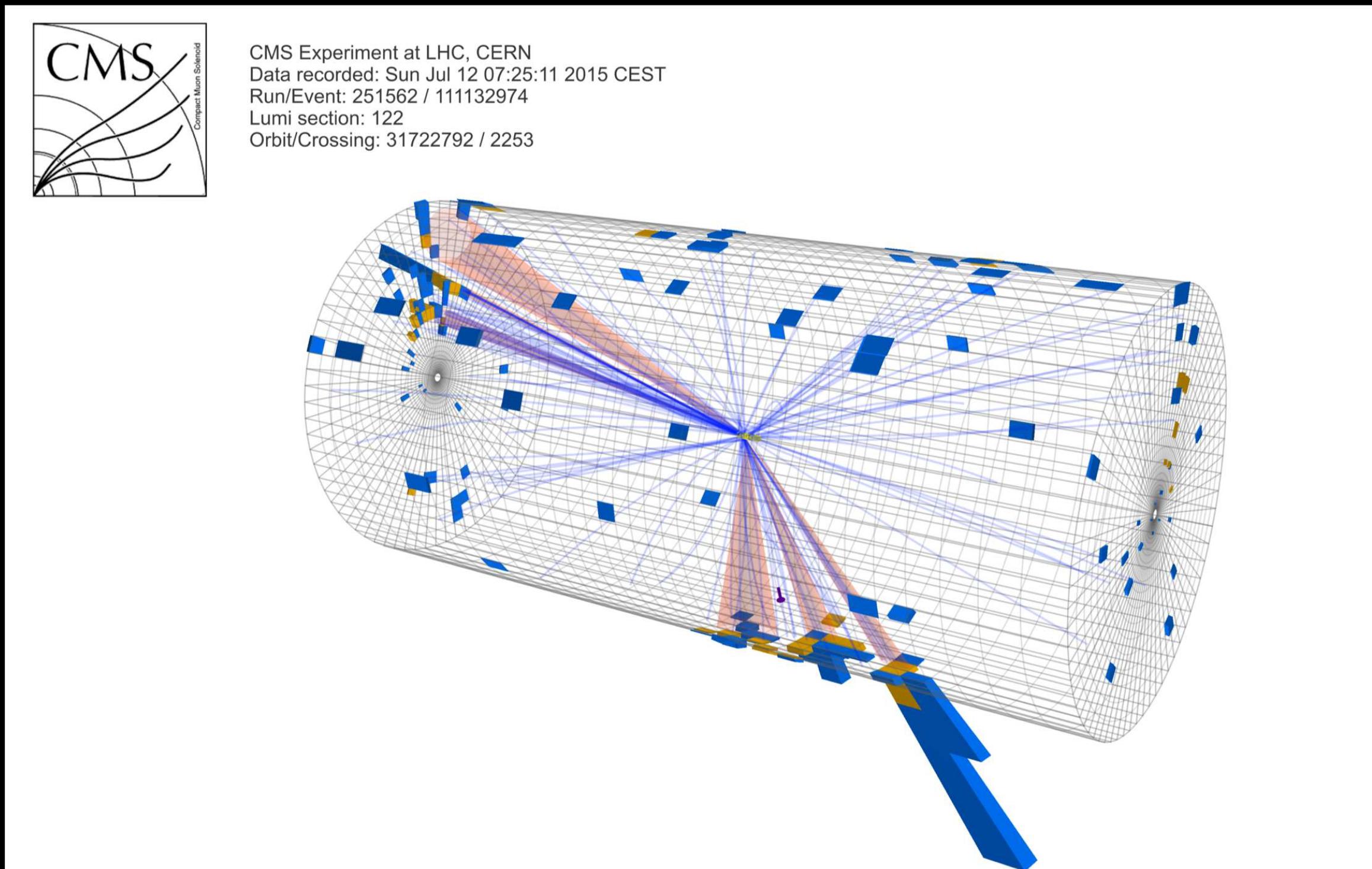
The (Metric) Space of Events

Revealing Hidden Geometry

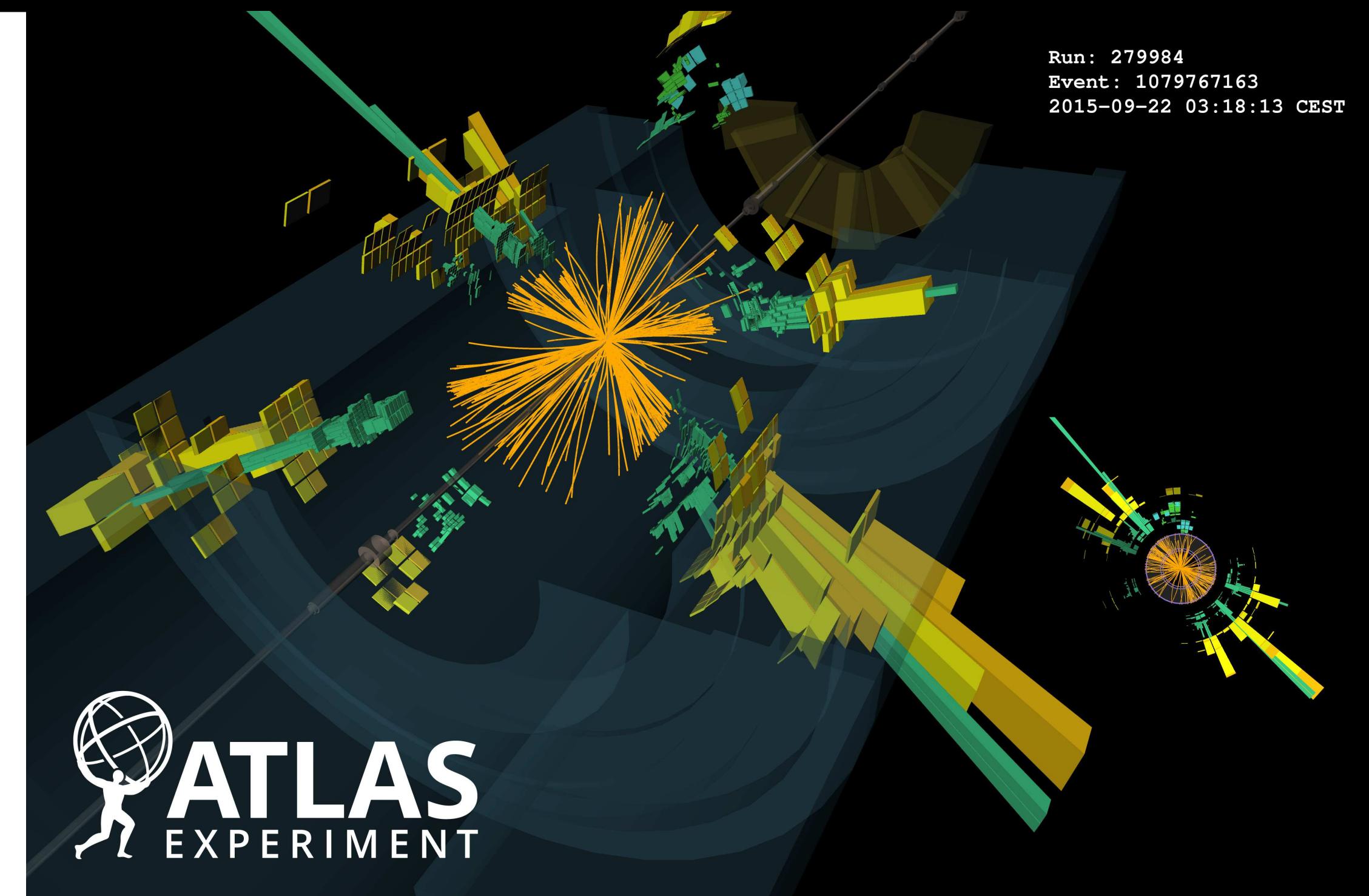
[Theory Space]

Explicit Geometry – Individual Events at the LHC

High-energy collisions produce final state particles with
energy, direction, charge, flavor, and other quantum numbers



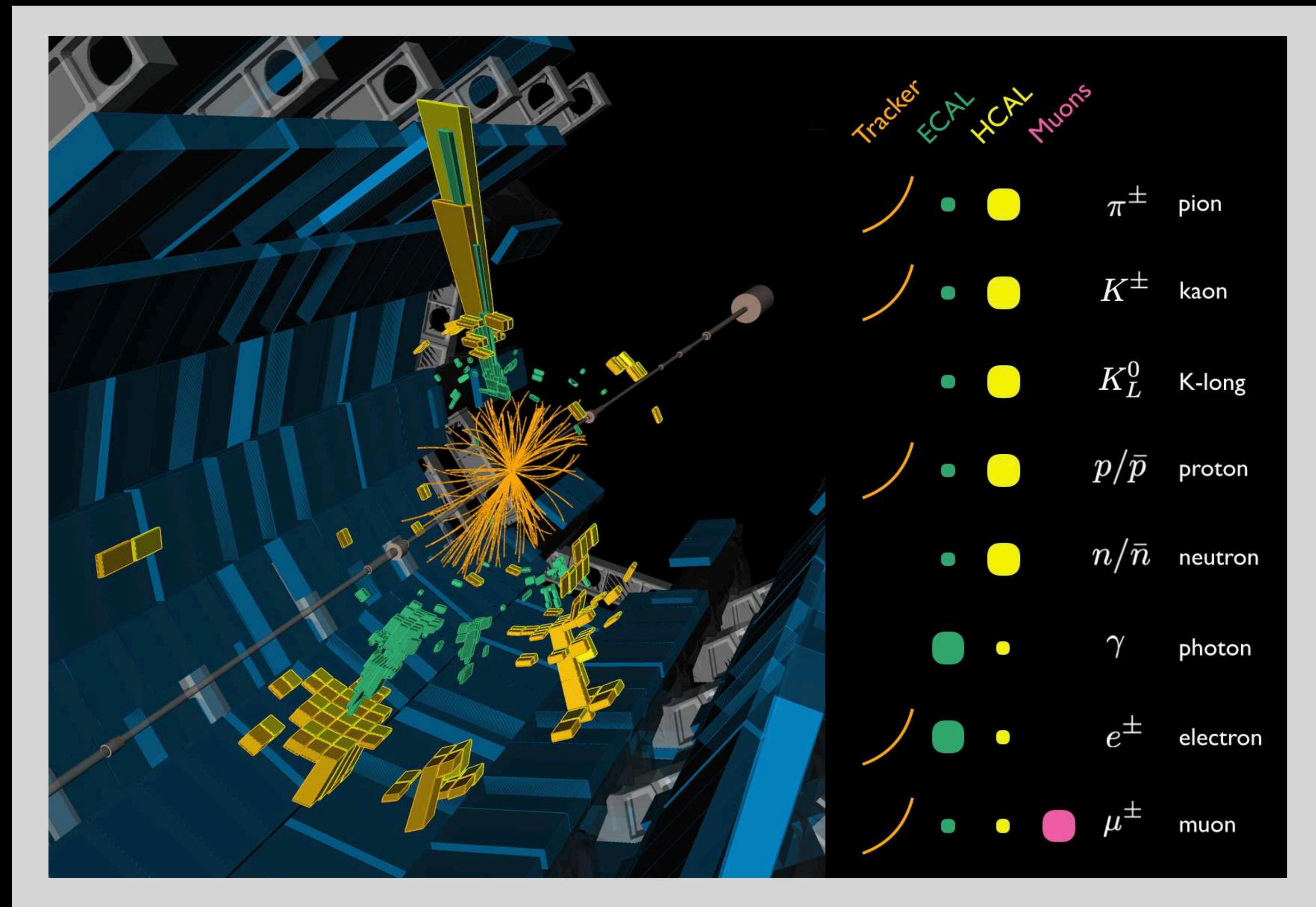
CMS hadronic $t\bar{t}$ event



ATLAS high jet-multiplicity event

Explicit Geometry – Individual Events at the LHC

High-energy collisions produce final state particles with
energy, direction, charge, flavor, and other quantum numbers

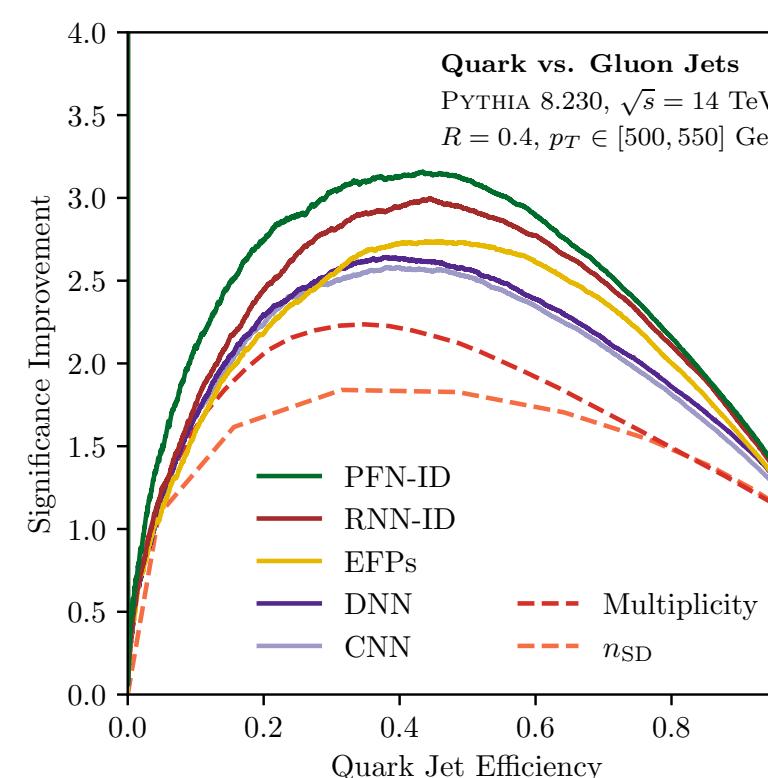
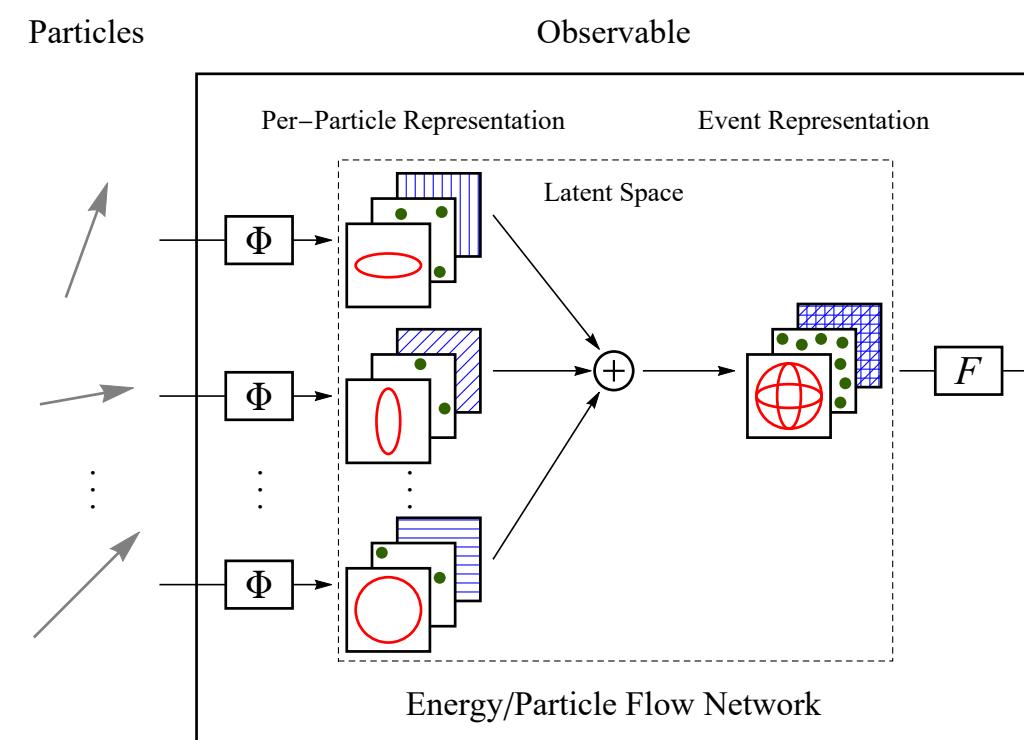


Machine-Learning-Inspired Methods for Particle Physics

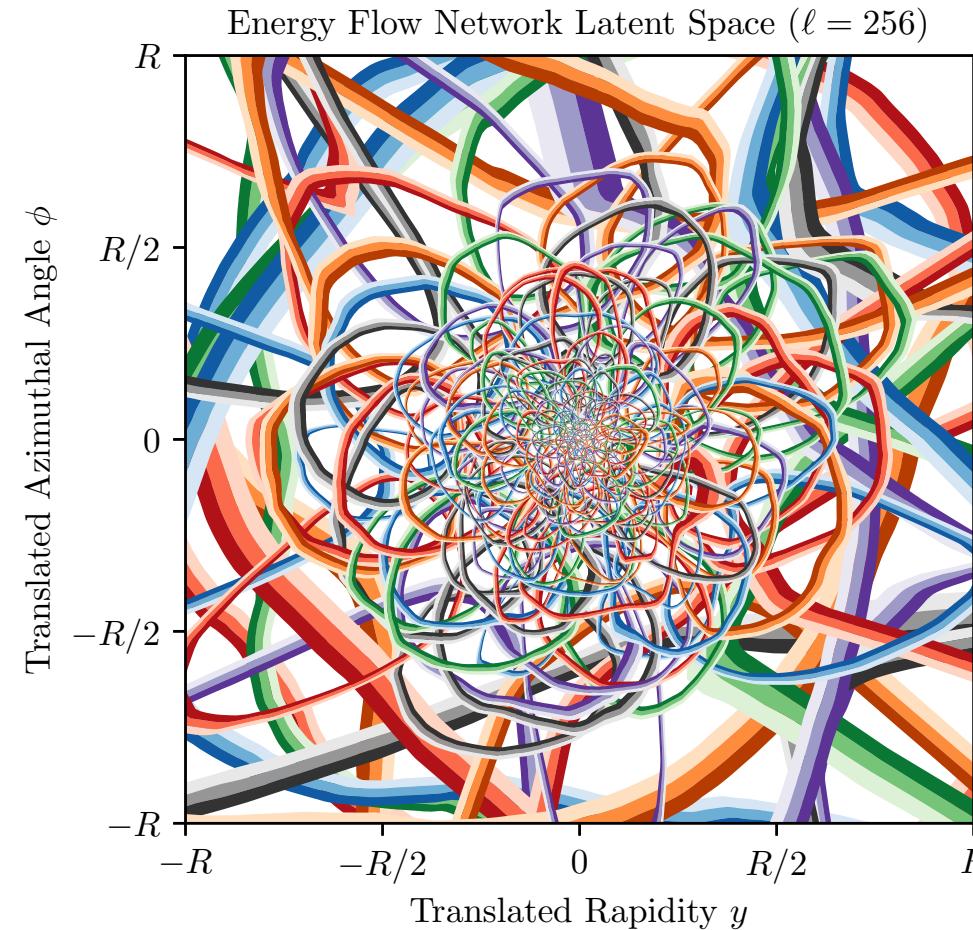
Energy/Particle Flow Networks (EFNs/PFNs)

[PTK, Metodiev, Thaler, [JHEP 2019](#)]

Permutation symmetric neural network architecture
for events with variable numbers of particles



Can be used to build
powerful taggers



Latent space visualization reveals
what the network has learned

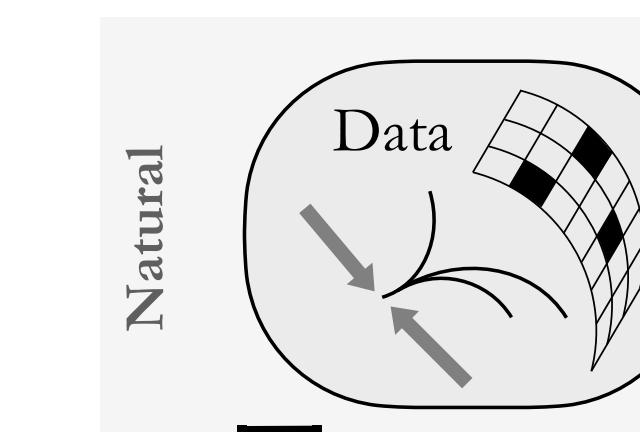
Dynamic pixel sizing related to
collinear singularity of QCD!

OmniFold

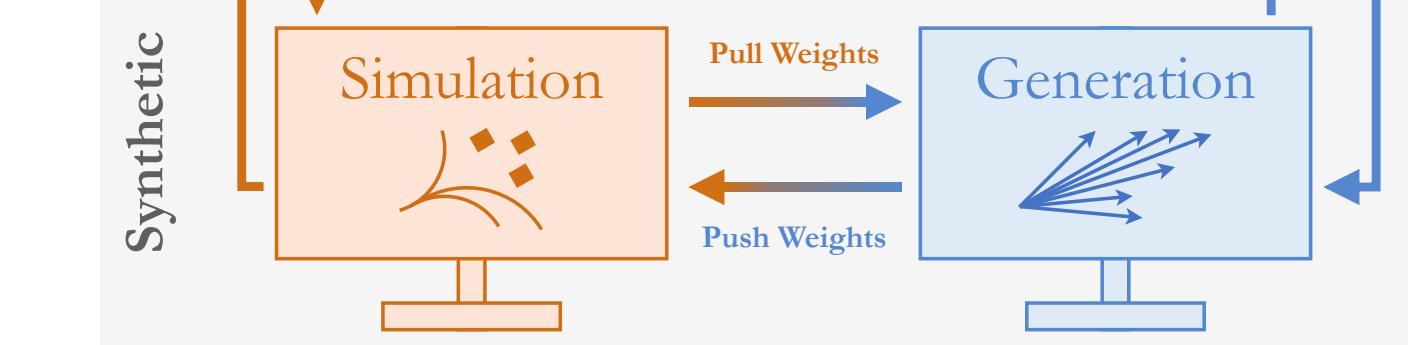
[Andreassen, PTK, Metodiev, Nachman, Thaler, [PRL 2020](#)]

Unbinned, full-phase space unfolding of all observables simultaneously

Detector-level

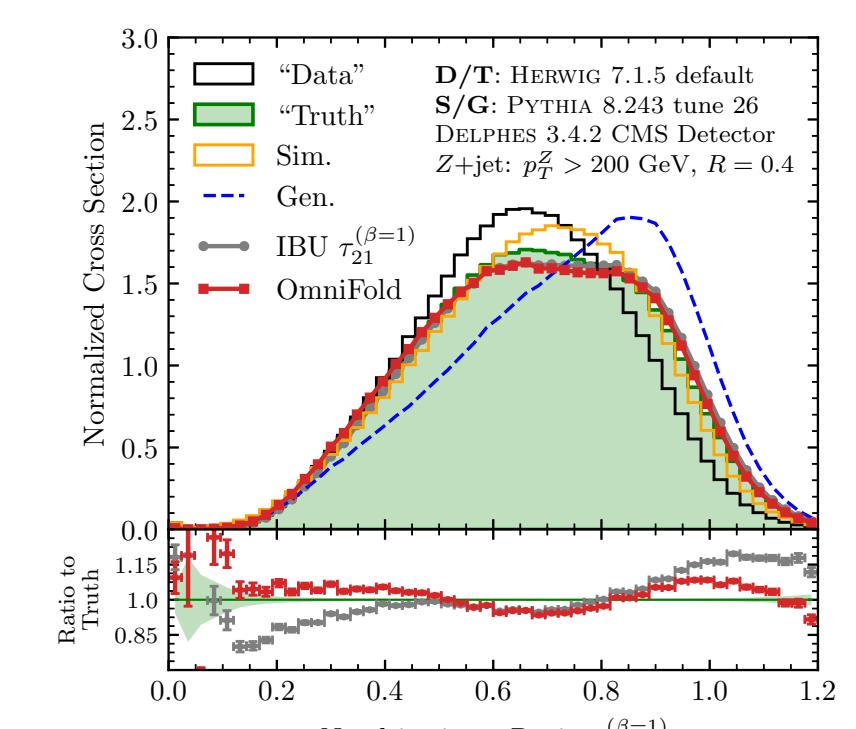
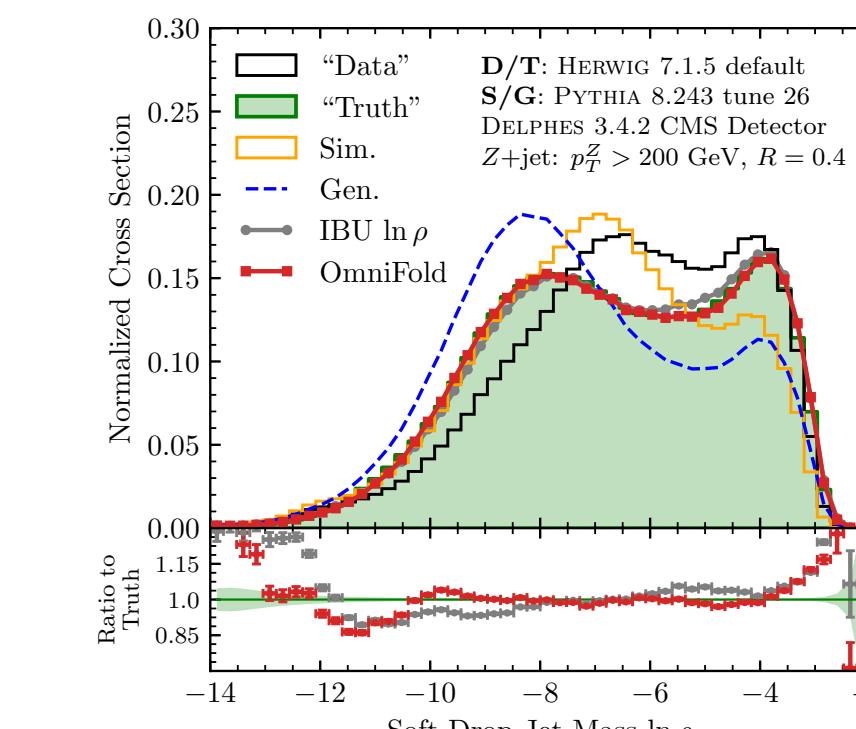


Synthetic



Particle-level

Single application of OmniFold succeeds where multiple IBUs are required!

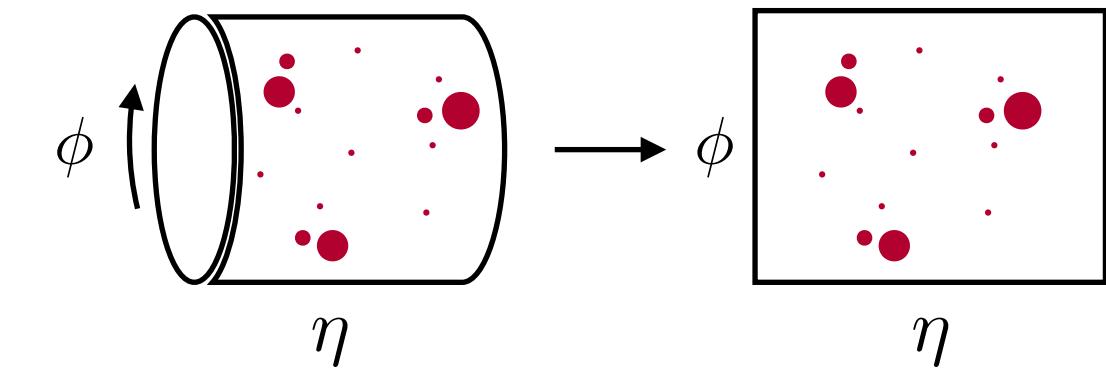


IRC safe

Sudakov safe

Back to Explicit Geometry – Events as Distributions of Energy

[PTK, Metodiev, Thaler, JHEP 2019; PTK, Metodiev, Thaler, JHEP 2020]

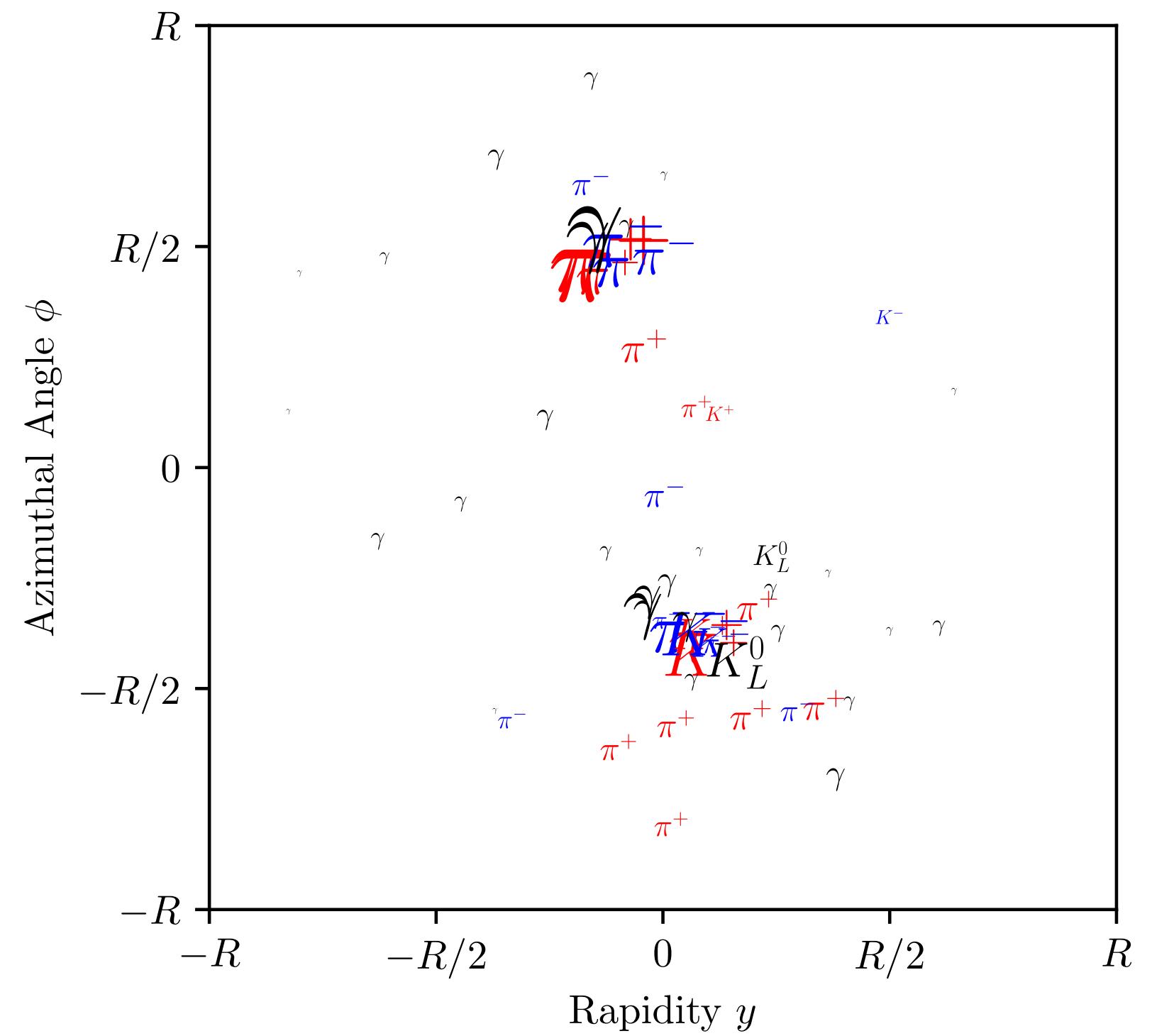


Energy flow distribution fully captures IRC-safe information

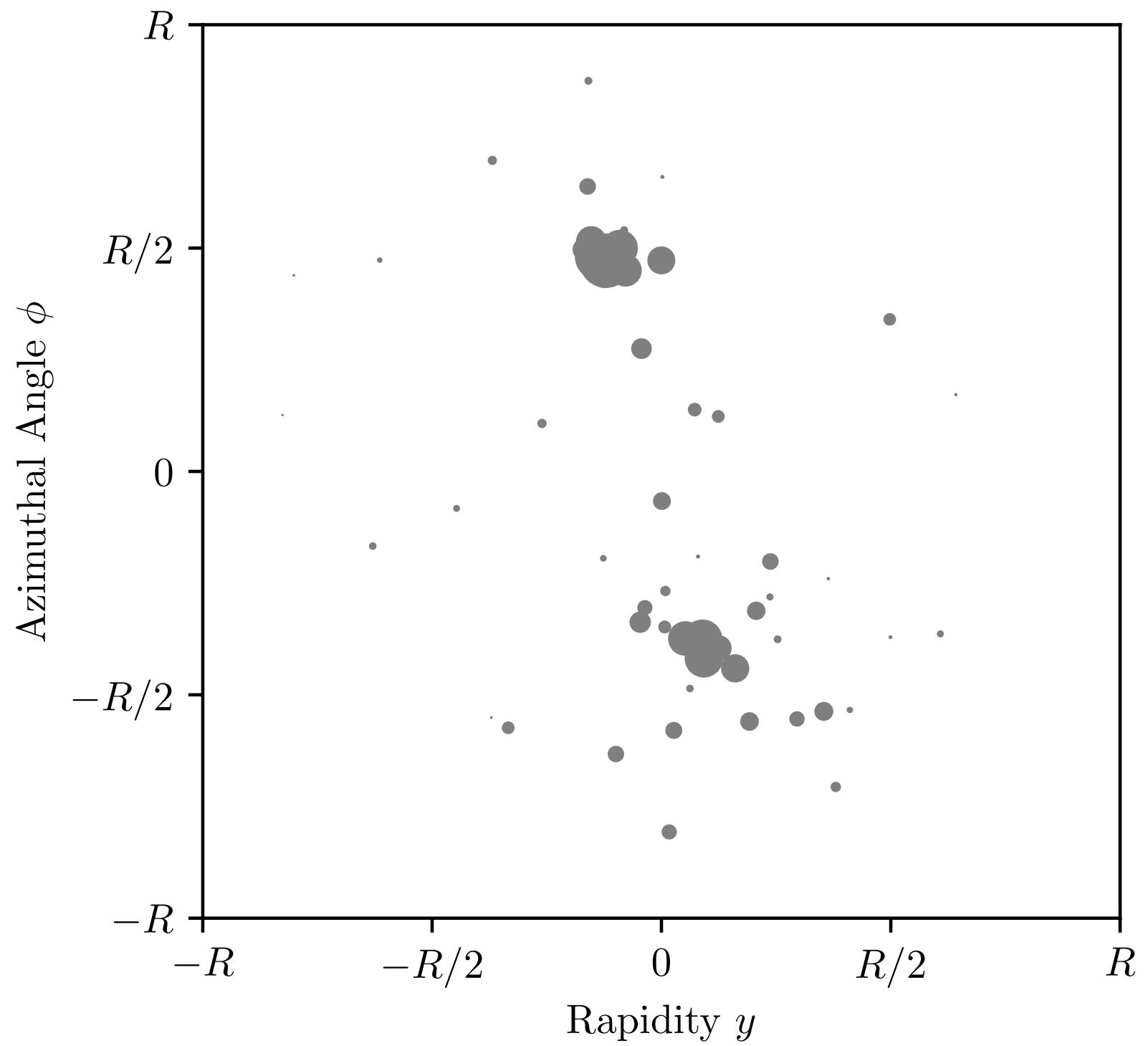
$$\mathcal{E}(\hat{n}) = \sum_{i=1}^M E_i \delta(\hat{n} - \hat{n}_i)$$

↑
 Energy Flow
 Distribution ↑
 Energy
 (p_T) Direction
 (y, φ)

Full event is a set of particles having momentum and charge/flavor

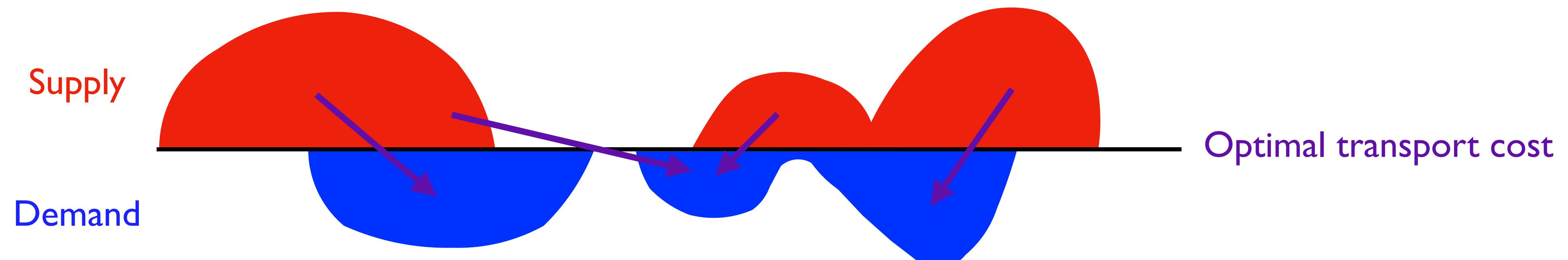
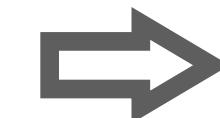


The **energy** flow is unpixelized and ignores charge/flavor information



Very Basic Question – When are Two Distributions Similar?

Optimal transport minimizes the “**work**” (**stuff** x distance) required to transport **supply** to **demand**



[Monge, 1781; Vaserštejn, 1969; Peleg, Werman, Rom, [IEEE 1989](#); Rubner, Tomasi, Guibas, [ICCV 1998](#), [ICCV 2000](#); Pele, Werman, [ECCV 2008](#); Pele, Taskar, [GSI 2013](#)]

The Energy Mover's Distance (EMD)

[PTK, Metodiev, Thaler, PRL 2019]

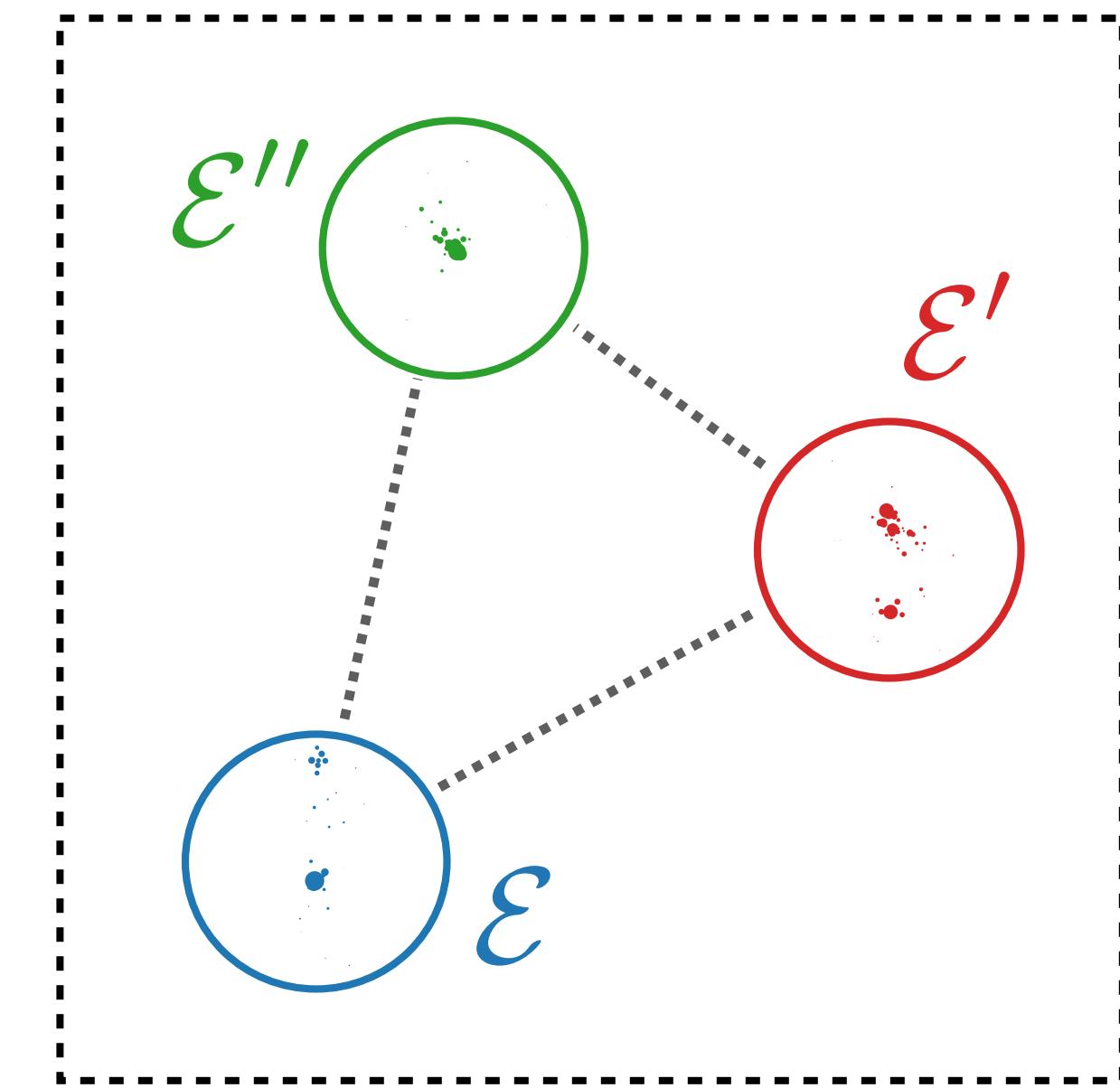
EMD between energy flows defines a metric on the space of events

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

Cost of optimal transport Cost of energy creation

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

Capacity constraints to ensure proper transport



R : controls cost of transporting energy vs. destroying/creating it

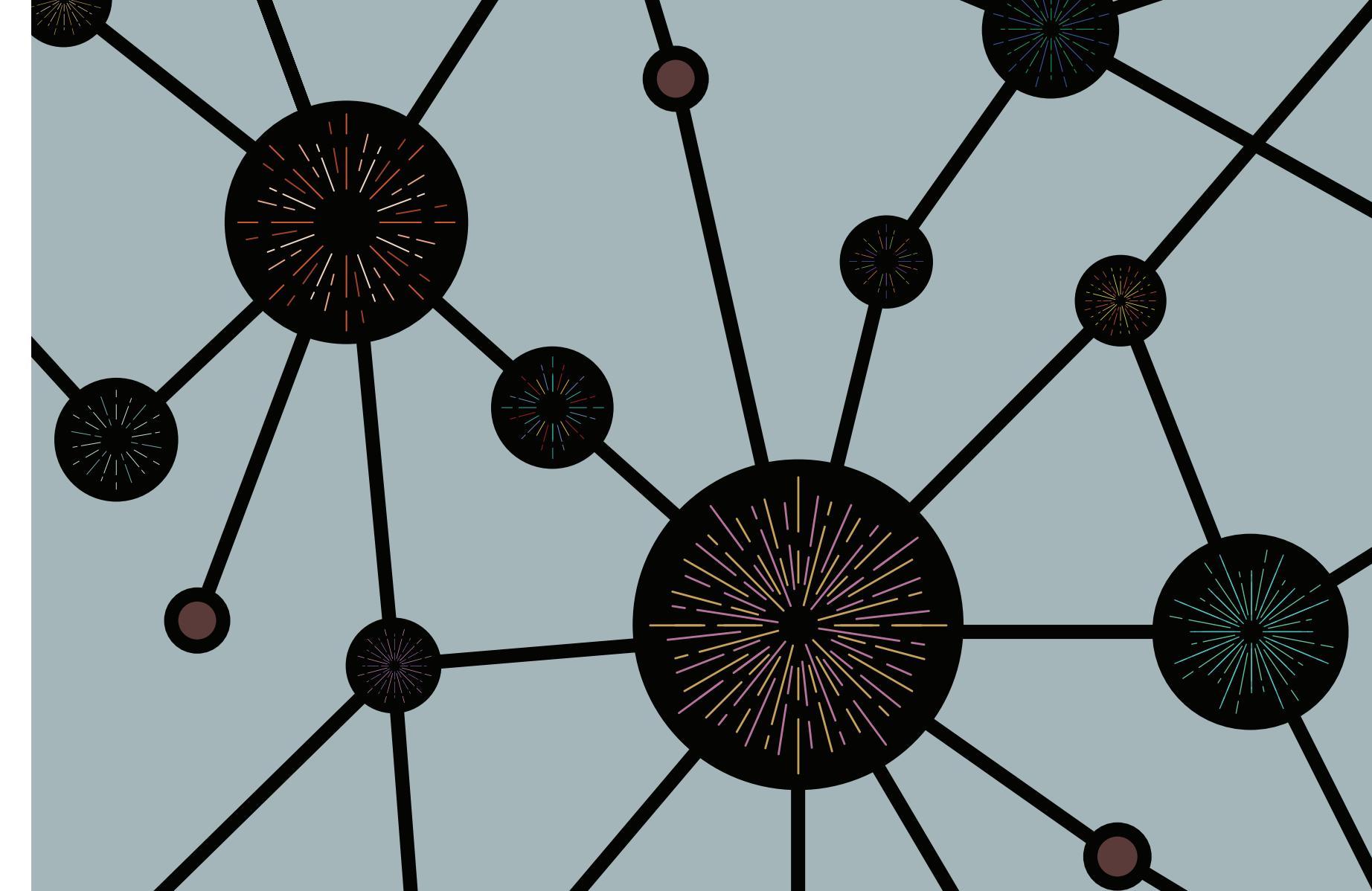
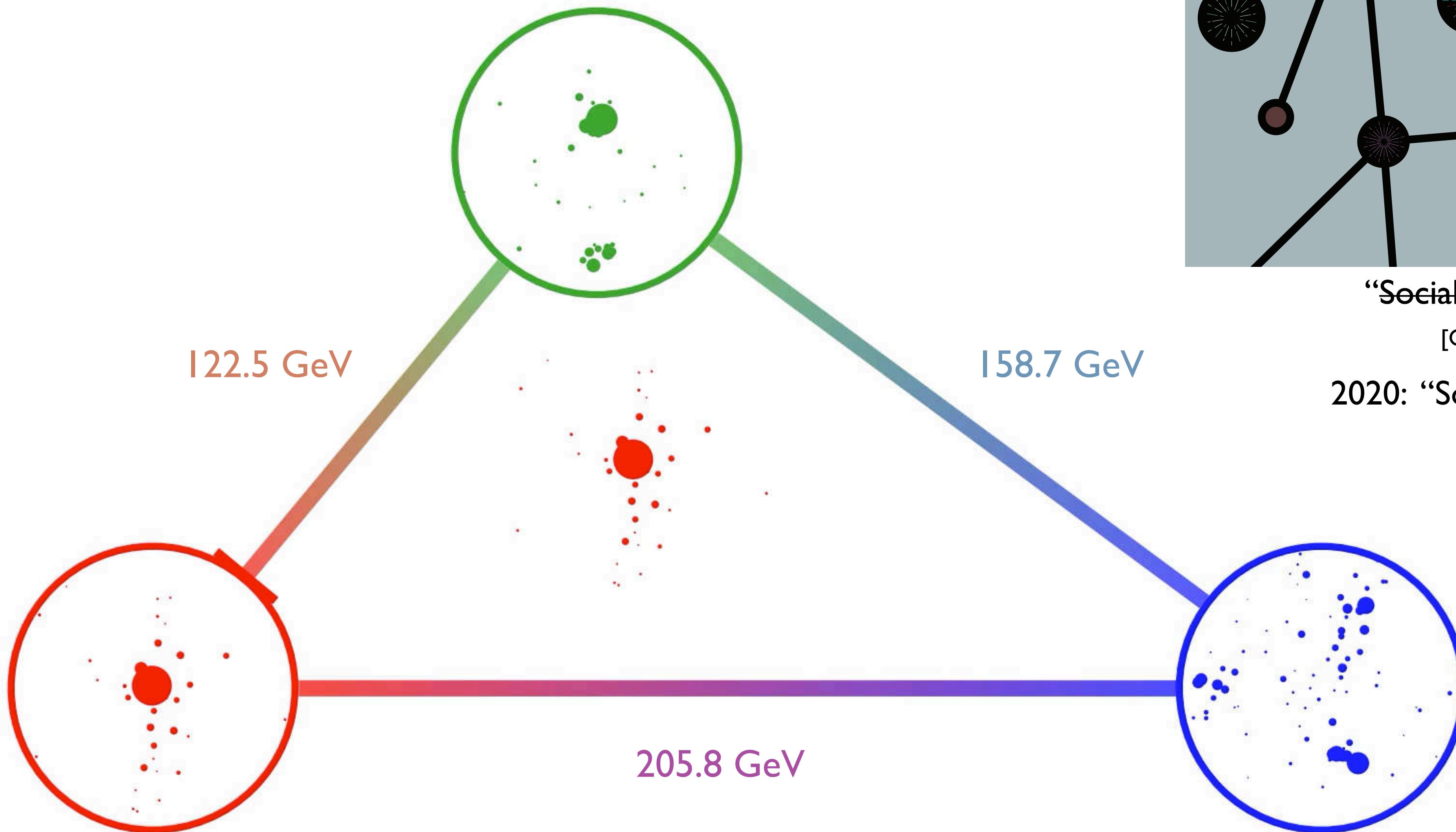
β : angular weighting exponent

Triangle inequality satisfied for $R \geq d_{\max}/2$

$$0 \leq \text{EMD}(\mathcal{E}, \mathcal{E}') \leq \text{EMD}(\mathcal{E}, \mathcal{E}'') + \text{EMD}(\mathcal{E}'', \mathcal{E}')$$

i.e. $R \geq$ jet radius for conical jets

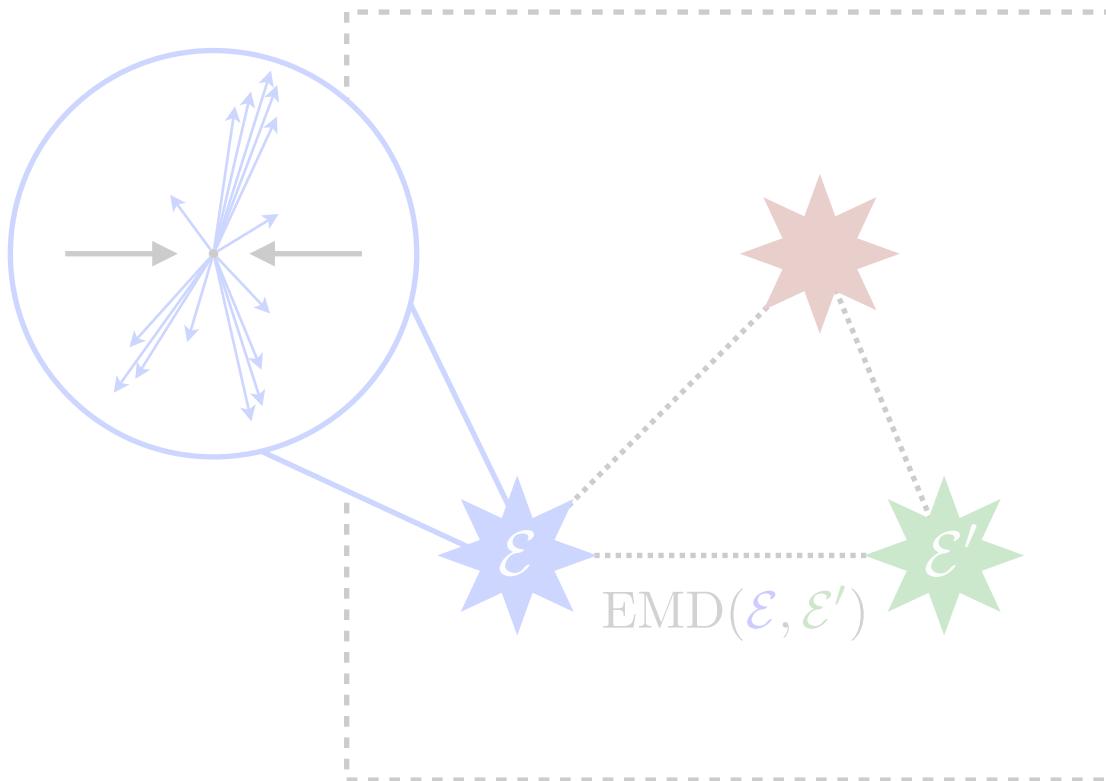
Geodesics in the Space of Events



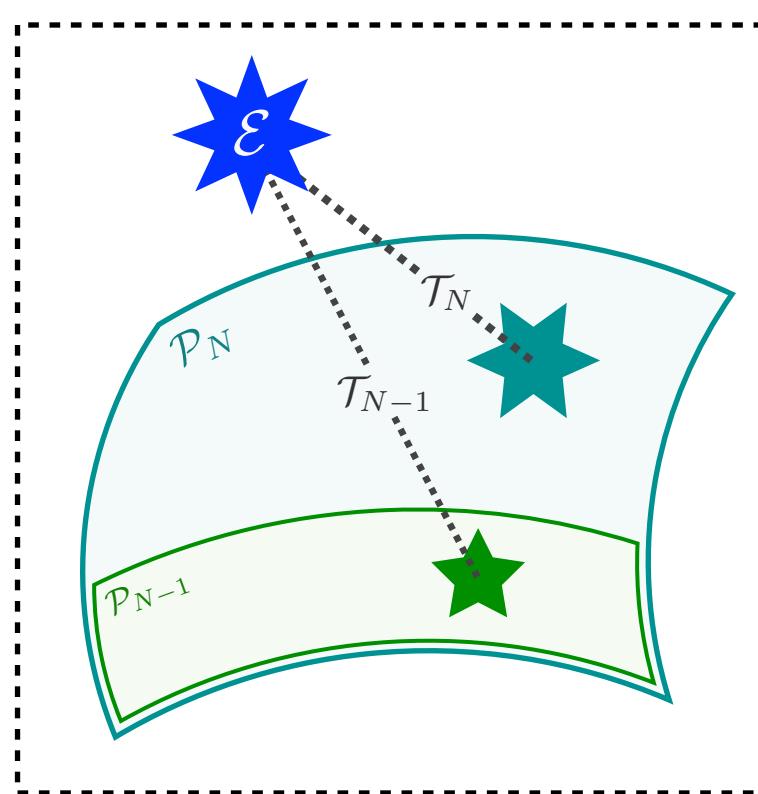
"Social networking of jets"

[Chu, MIT News 2019]

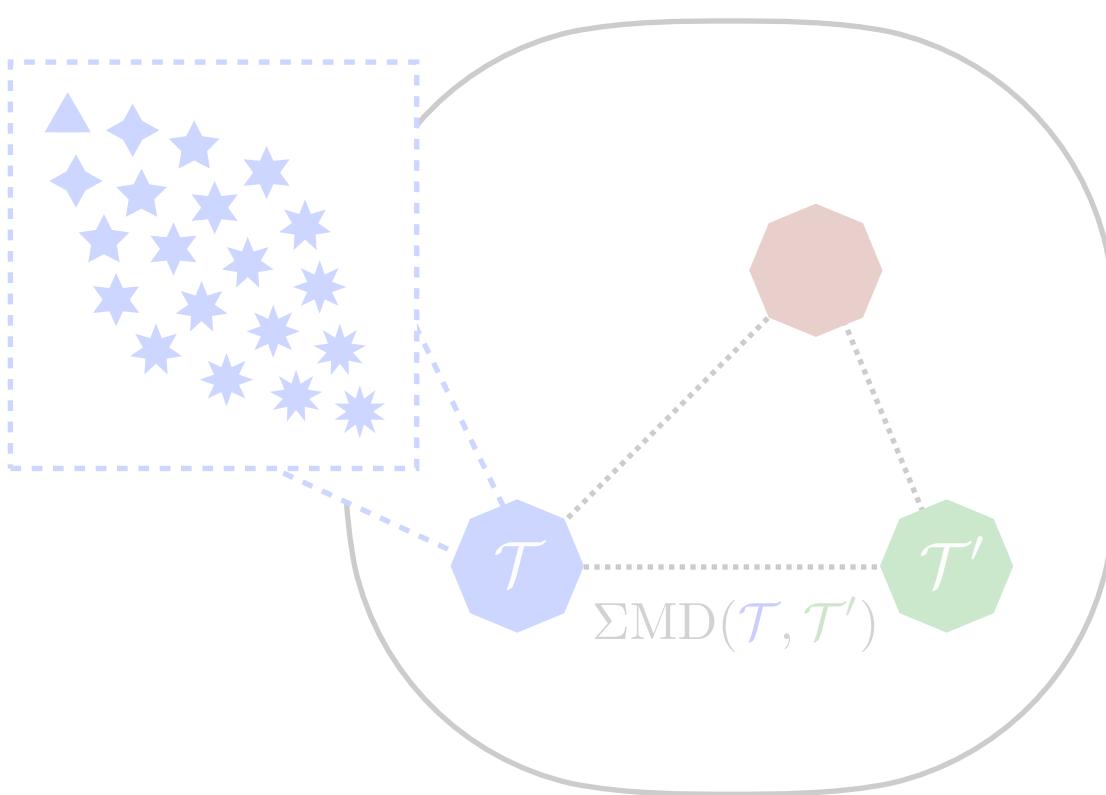
2020: "Social distancing of jets"



The (Metric) Space of Events



Revealing Hidden Geometry

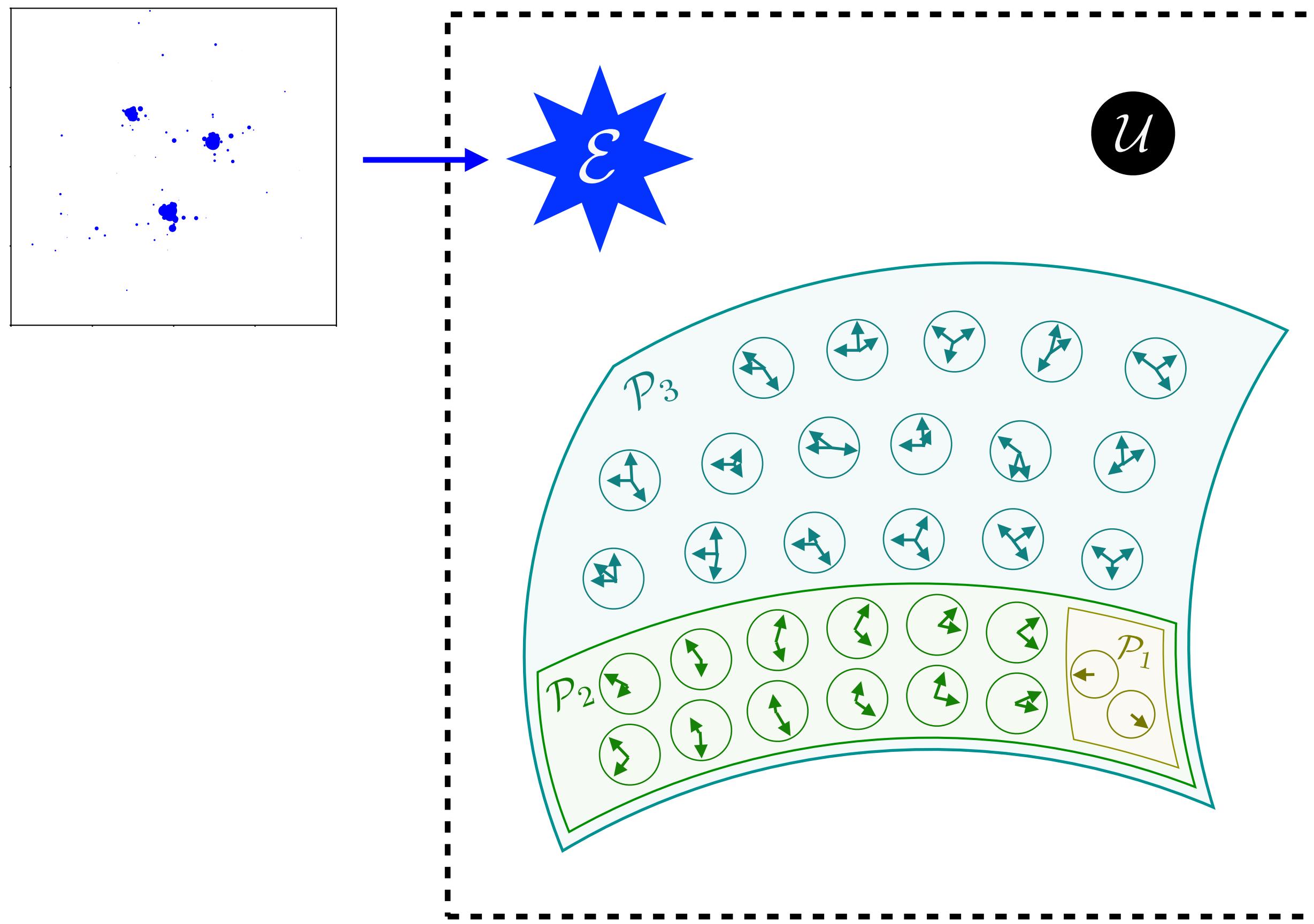


[Theory Space]

N-particle Manifolds in the Space of Events

[PTK, Metodiev, Thaler, JHEP 2020]

$$\mathcal{P}_N = \text{set of all N-particle configurations} = \left\{ \sum_{i=1}^N E_i \delta(\hat{n} - \hat{n}_i) \mid E_i \geq 0 \right\}$$



\mathcal{P}_1 : manifold of events with one particle

\mathcal{P}_2 : manifold of events with two particles

\mathcal{P}_3 : manifold of events with three particles

⋮

$\mathcal{P}_N \supset \mathcal{P}_{N-1} \supset \dots \supset \mathcal{P}_3 \supset \mathcal{P}_2 \supset \mathcal{P}_1$

by soft and collinear limits

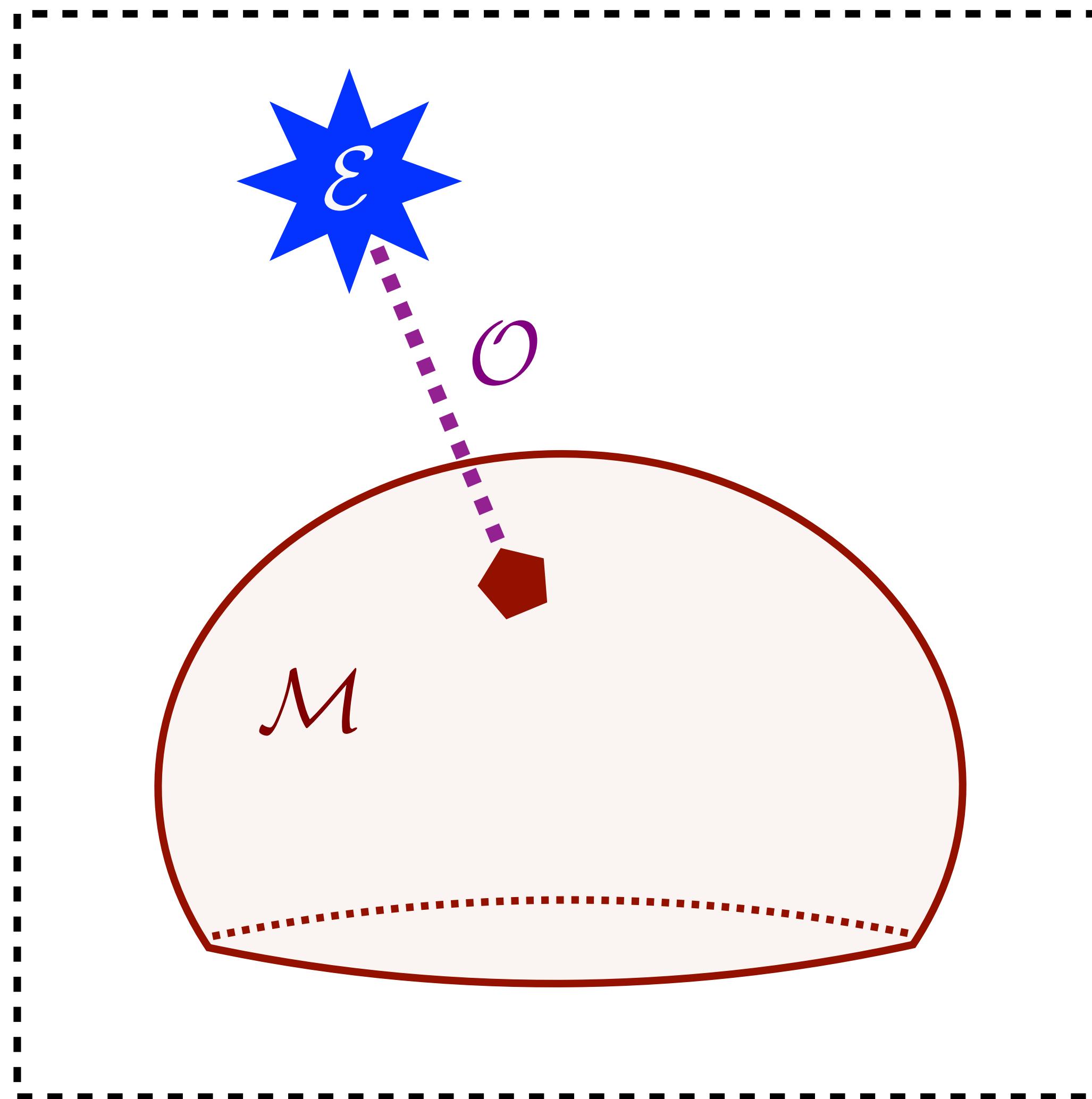
$$\longleftrightarrow = \xrightarrow{\epsilon \rightarrow 0} = \xrightarrow{1-\lambda} \xrightarrow{\lambda}$$



Uniform event, not contained in any \mathcal{P}_N

Defining Observables via Event Space Geometry

[PTK, Metodiev, Thaler, JHEP 2020]



Many common *observables* are distance of closest approach from event to a specific *manifold*

$$\mathcal{O}(\mathcal{E}) = \min_{\mathcal{E}' \in \mathcal{M}} \text{EMD}_{\beta,R}(\mathcal{E}, \mathcal{E}')$$

EMD variant for equal-energy events

$$\text{EMD}_\beta(\mathcal{E}, \mathcal{E}') = \lim_{R \rightarrow \infty} R^\beta \text{EMD}_{\beta,R}(\mathcal{E}, \mathcal{E}') = \min_{\{f_{ij} \geq 0\}} \sum_{i=1}^M \sum_{j=1}^{M'} f_{ij} \theta_{ij}^\beta$$

Enforces equal energy (else infinity)

on equal-energy events

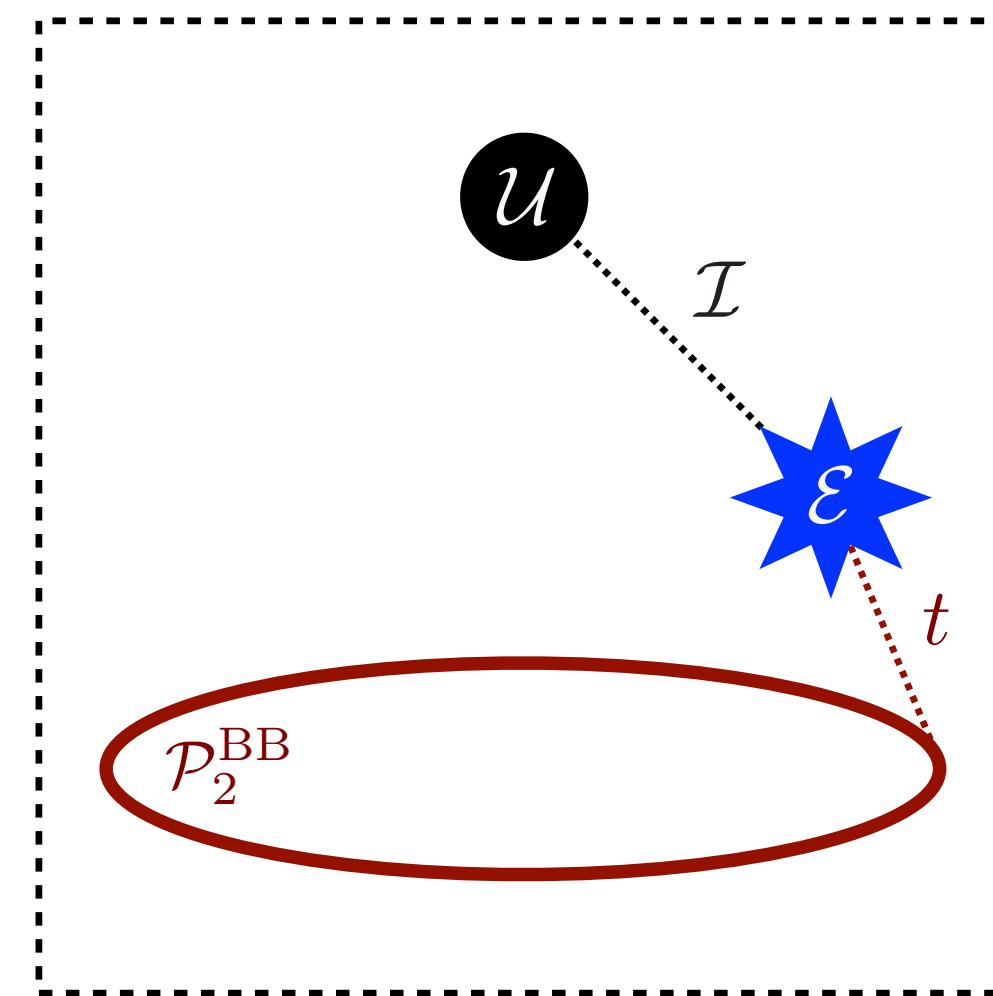
Defining Observables via Event Space Geometry

$$\mathcal{O}(\mathcal{E}) = \min_{\mathcal{E}' \in \mathcal{M}} \text{EMD}_{\beta,R}(\mathcal{E}, \mathcal{E}')$$

[PTK, Metodiev, Thaler, [JHEP 2020](#)]

Thrust, spherocity, isotropy*

*Distance of closest approach
to a specific manifold*



$$t(\mathcal{E}) = \min_{\mathcal{E}' \in \mathcal{P}_2^{\text{BB}}} \text{EMD}_2(\mathcal{E}, \mathcal{E}')$$

$$\sqrt{s(\mathcal{E})} = \min_{\mathcal{E}' \in \mathcal{P}_2^{\text{BB}}} \text{EMD}_1(\mathcal{E}, \mathcal{E}')$$

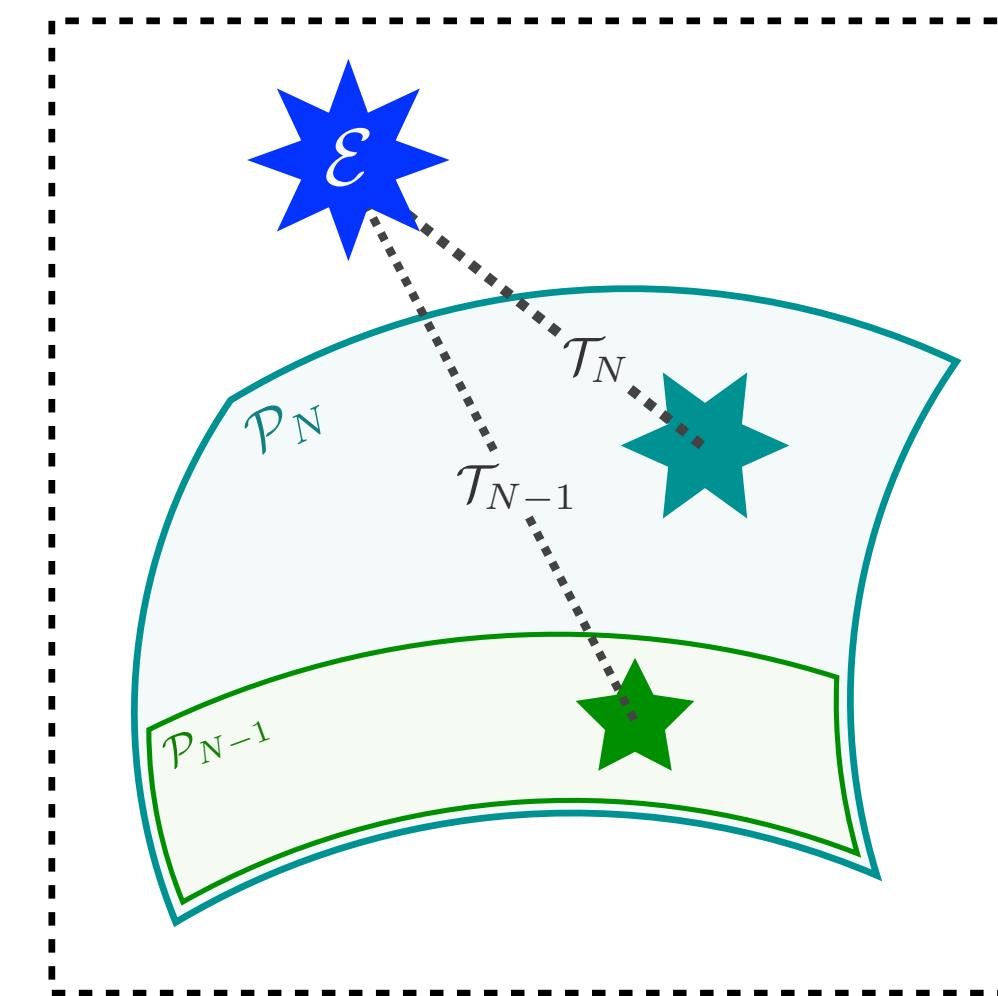
$$\mathcal{I}^{(\beta)}(\mathcal{E}) = \min_{\mathcal{E}' \in M_U} \text{EMD}_\beta(\mathcal{E}, \mathcal{E}')$$

[Farhi, [PRL 1977](#); Georgi, Machacek, [PRL 1977](#)]

*New! [Cesarotti, Thaler, [2004.06125](#)]

N-jettiness

*Minimum distance from event
to N-particle manifold*



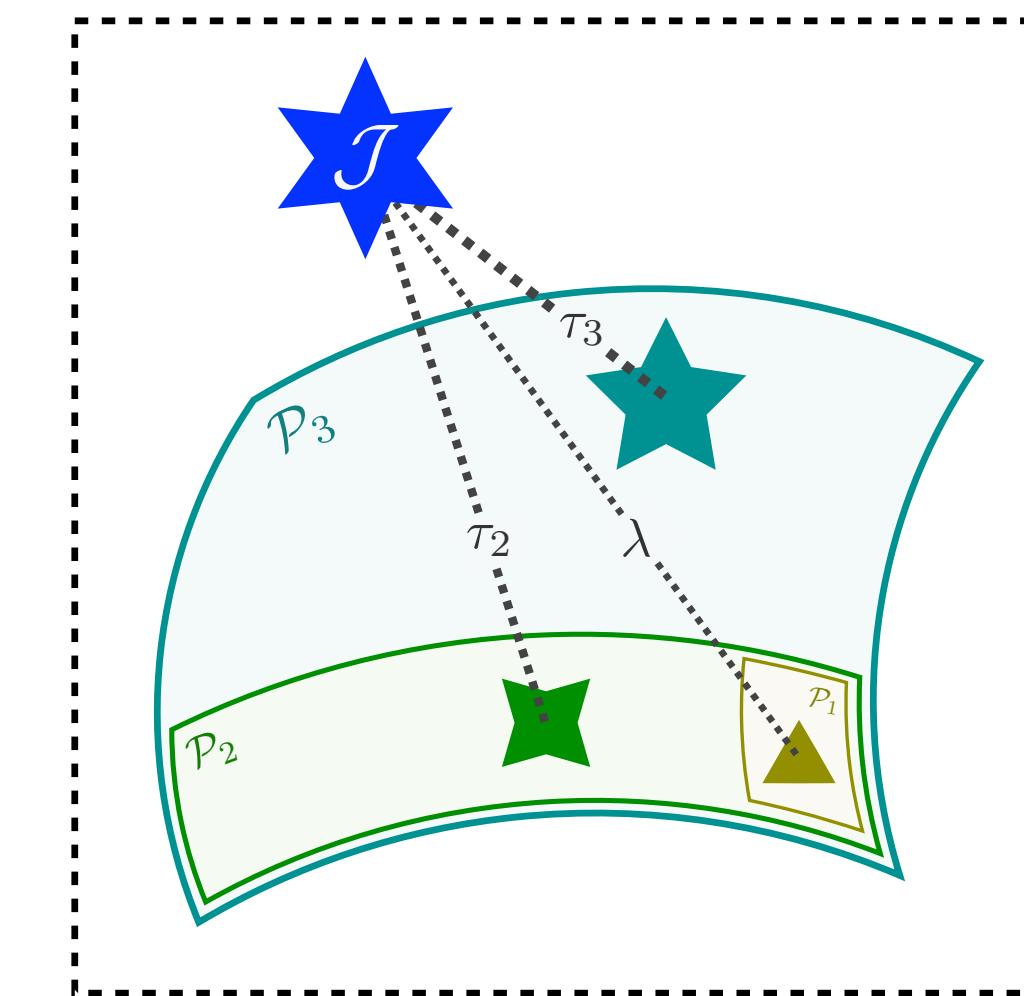
$$\mathcal{T}_N^{(\beta)}(\mathcal{E}) = \min_{\mathcal{E}' \in \mathcal{P}_N} \text{EMD}_\beta(\mathcal{E}, \mathcal{E}')$$

$$\mathcal{T}_N^{(\beta,R)}(\mathcal{E}) = \min_{\mathcal{E}' \in \mathcal{P}_N} \text{EMD}_{\beta,R}(\mathcal{E}, \mathcal{E}')$$

[Brandt, Dahmen, [Z. Phys 1979](#);
Stewart, Tackmann, Waalewijn, [PRL 2010](#)]

N-subjettiness, angularities

*Smallest distance from jet to
N-particle manifold*



$$\lambda_\beta(\mathcal{J}) = \min_{\mathcal{J}' \in \mathcal{P}_1} \text{EMD}_\beta(\mathcal{J}, \mathcal{J}')$$

$$\tau_N^{(\beta)}(\mathcal{J}) = \min_{\mathcal{J}' \in \mathcal{P}_N} \text{EMD}_\beta(\mathcal{J}, \mathcal{J}')$$

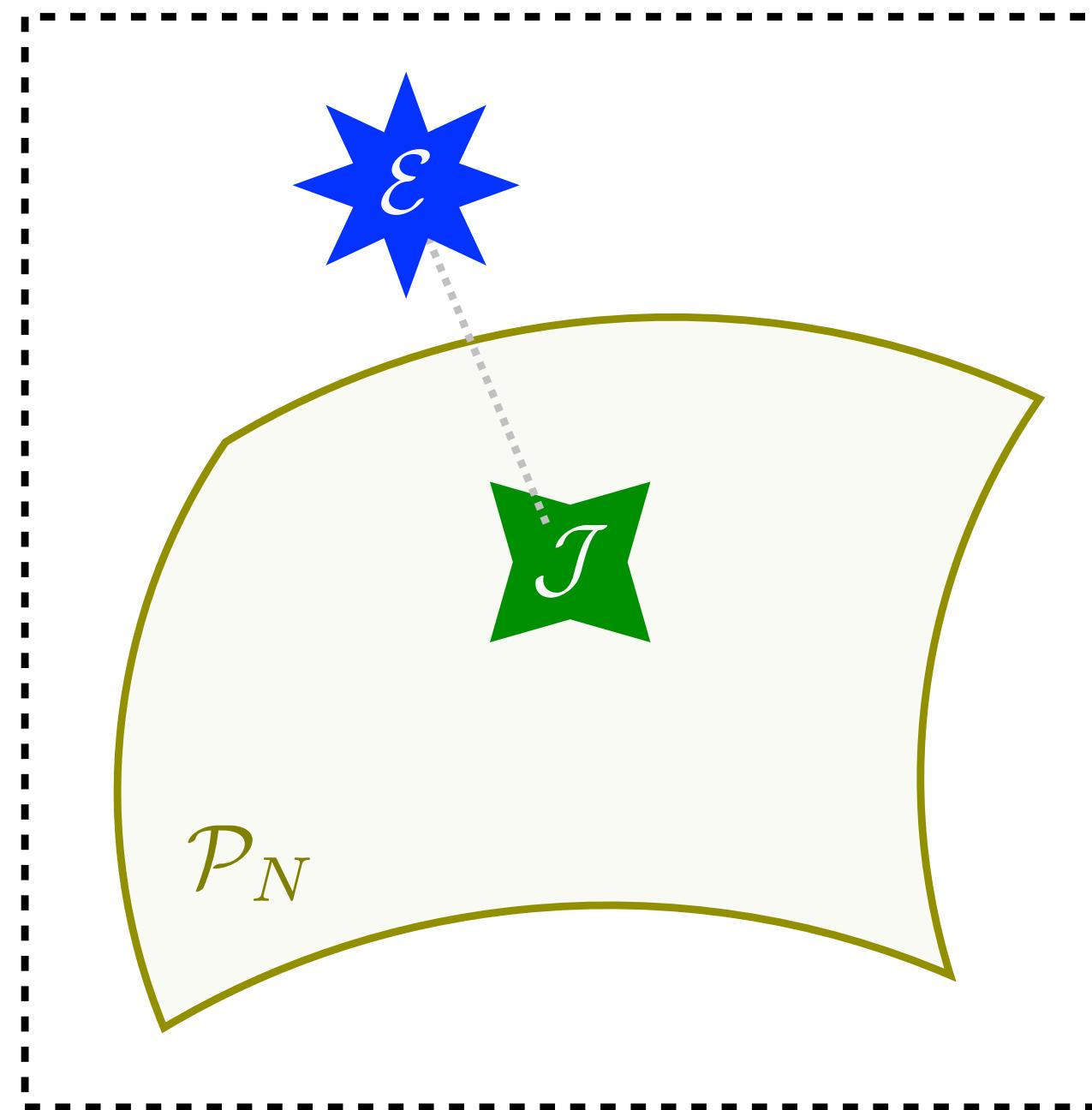
[Ellis, Vermilion, Walsh, Hornig, Lee, [JHEP 2010](#);
Thaler, Van Tilburg, [JHEP 2011](#), [JHEP 2012](#)]

Jets in the Space of Events – The Closest N -particle Description of an M -particle Event

[PTK, Metodiev, Thaler, [JHEP 2020](#)]

Exclusive cone finding

XCone finds N jets by
minimizing N -jettiness

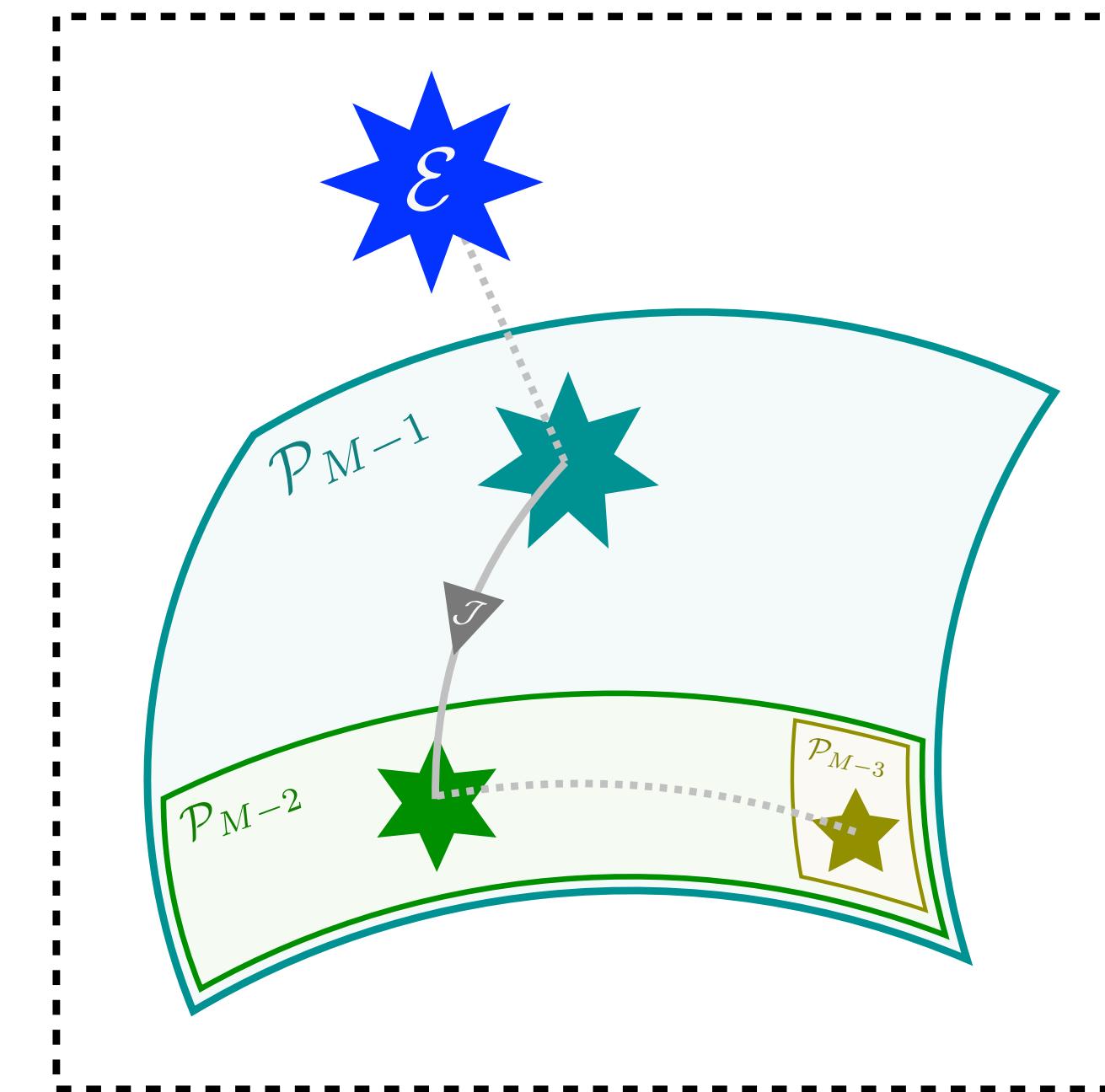


$$\mathcal{J}_{N,\beta,R}^{\text{XCone}}(\mathcal{E}) = \arg \min_{\mathcal{J} \in \mathcal{P}_N} \text{EMD}_{\beta,R}(\mathcal{E}, \mathcal{J})$$

[Stewart, Tackmann, Thaler, Vermilion, Wilkason, [JHEP 2015](#);
Thaler, Wilkason, [JHEP 2015](#)]

Sequential recombination

Iteratively merges particles or
identifies a jet



event with one fewer particle after one step

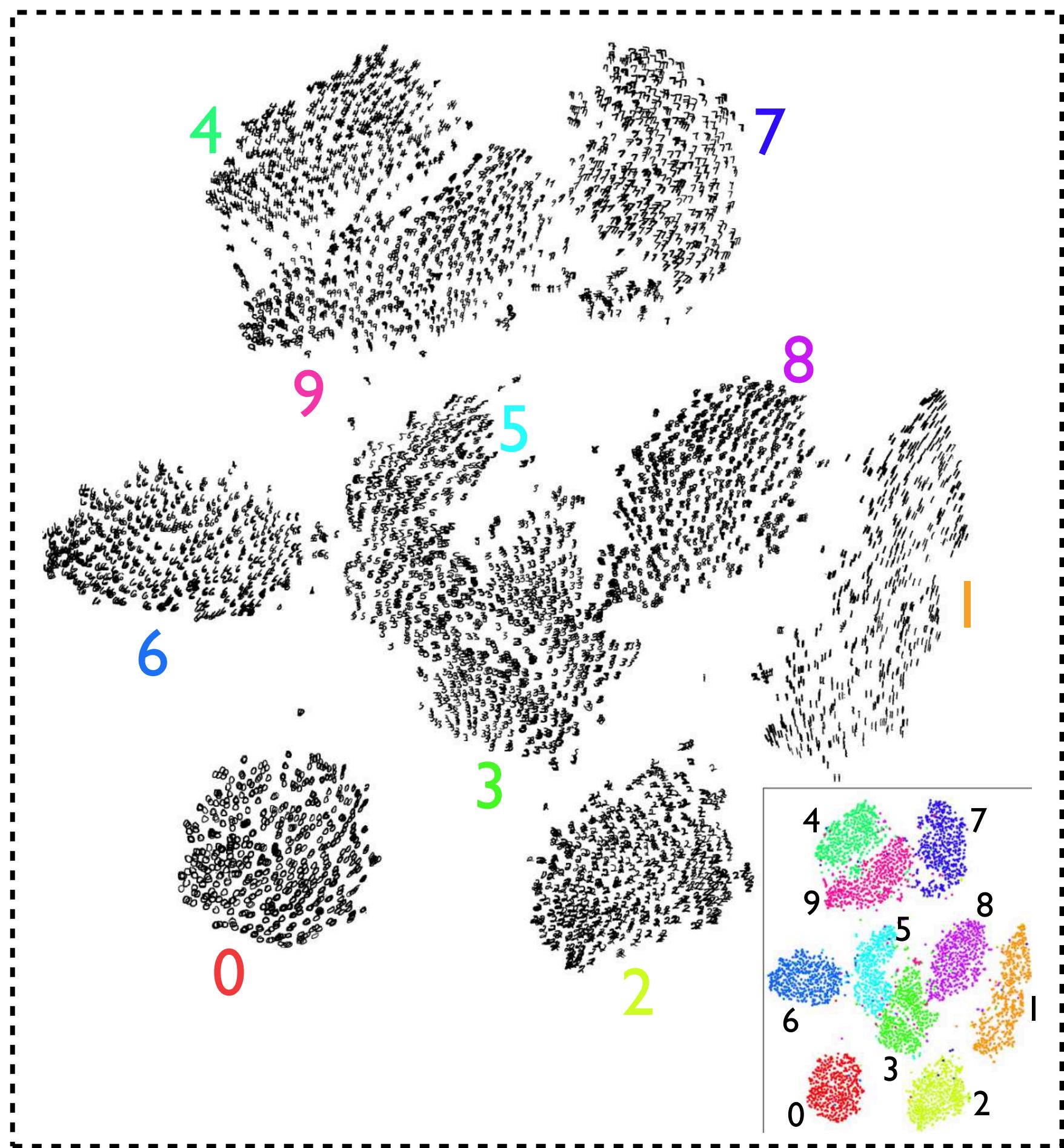
$$\mathcal{E}_{M-1}^{(\beta,R)}(\mathcal{E}_M) = \arg \min_{\mathcal{E}'_{M-1} \in \mathcal{P}_{M-1}} \text{EMD}_{\beta,R}(\mathcal{E}_M, \mathcal{E}'_{M-1})$$

[Catani, Dokshitzer, Seymour, Webber, [Nucl. Phys. B 1993](#);
Ellis, Soper, [PRD 1993](#);
Dokshitzer, Leder, Moretti, Webber, [JHEP 1997](#);
Cacciari, Salam, Soyez, [JHEP 2008](#)]

Visualizing Geometry in the Space of Events

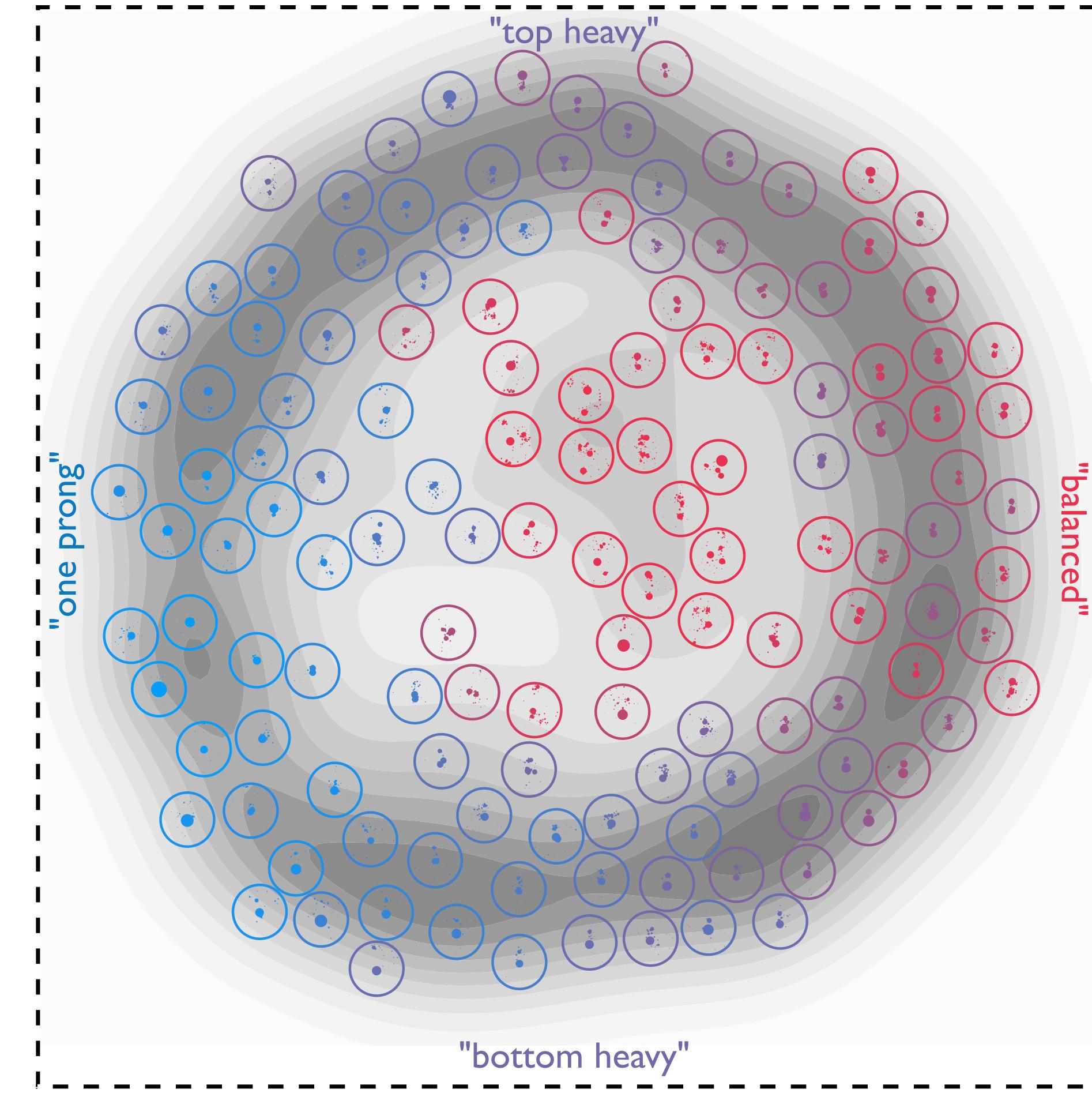
[PTK, Metodiev, Thaler, PRL 2019]

t-Distributed Stochastic Neighbor Embedding (t-SNE)
MNIST handwritten digits

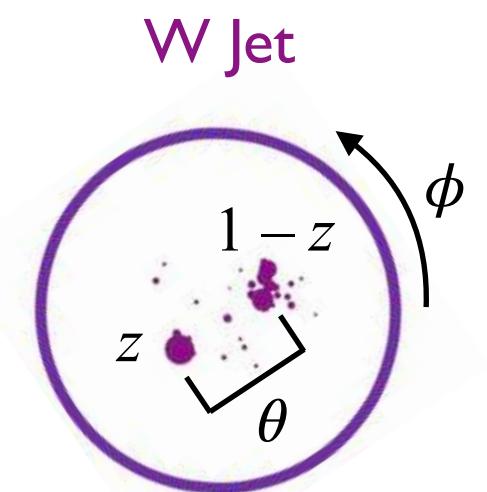


[L. van der Maaten, G. Hinton, JMLR 2008]

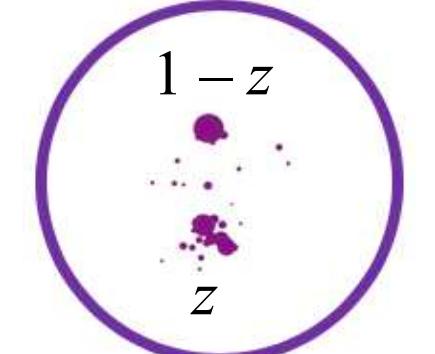
Geometric space of W jets



Gray contours represent the density of jets
Each circle is a particular W jet

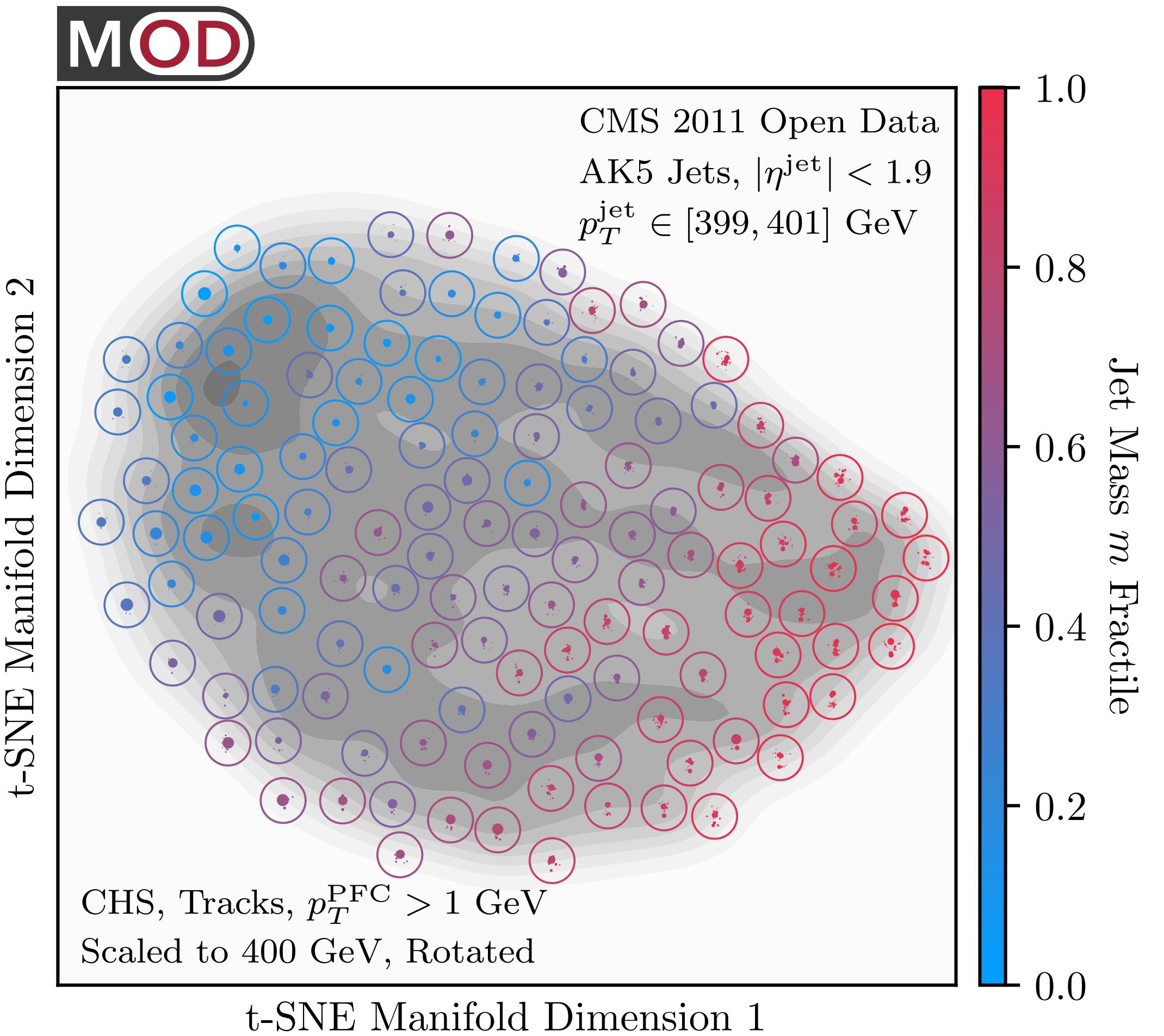


Constraints: W Mass and
 $\phi = 0$ preprocessing

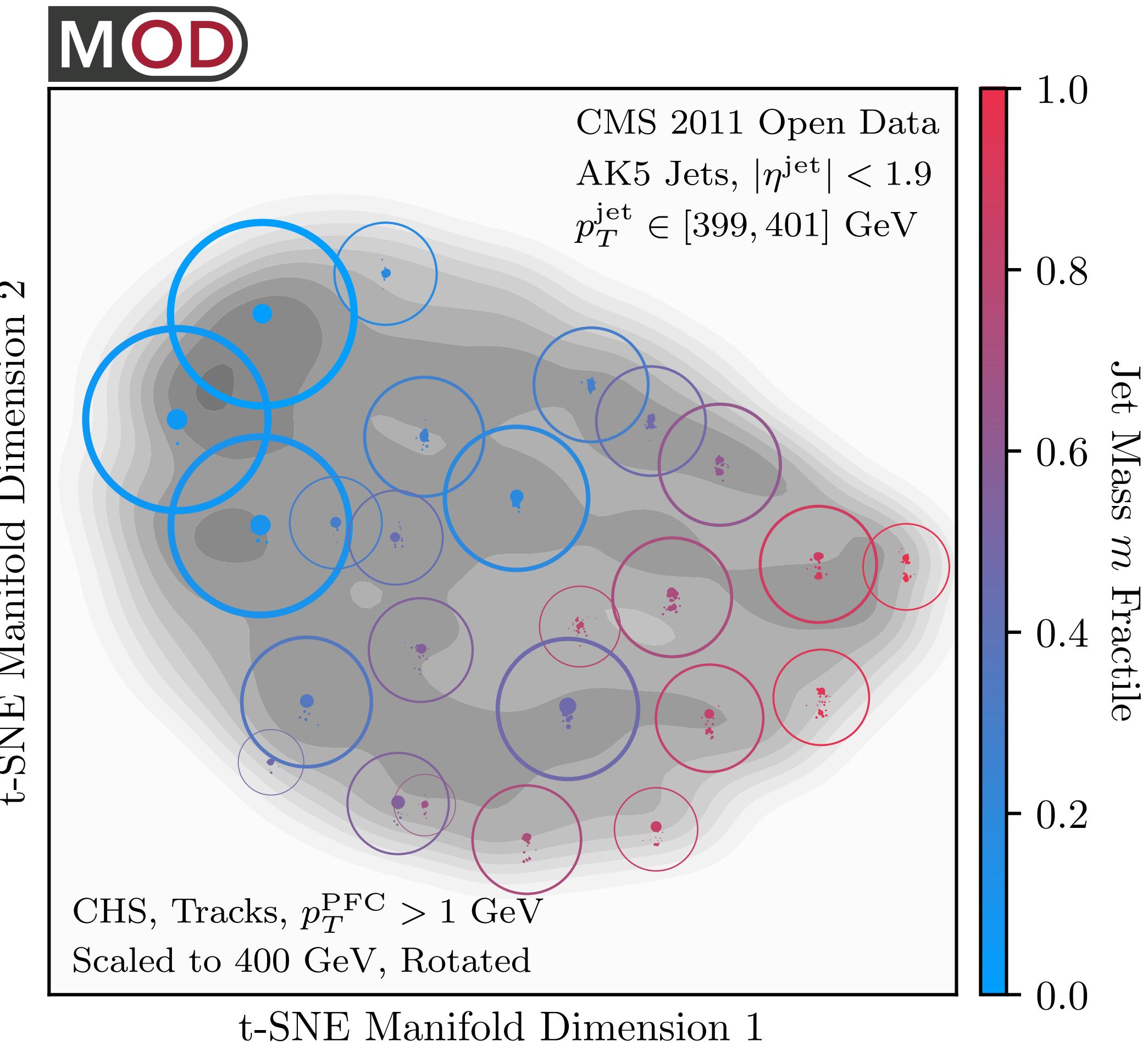


Visualizing Geometry in CMS Open Data

[PTK, Mastandrea, Metodiev, Naik, Thaler, PRD 2019; code and datasets at energyflow.network]



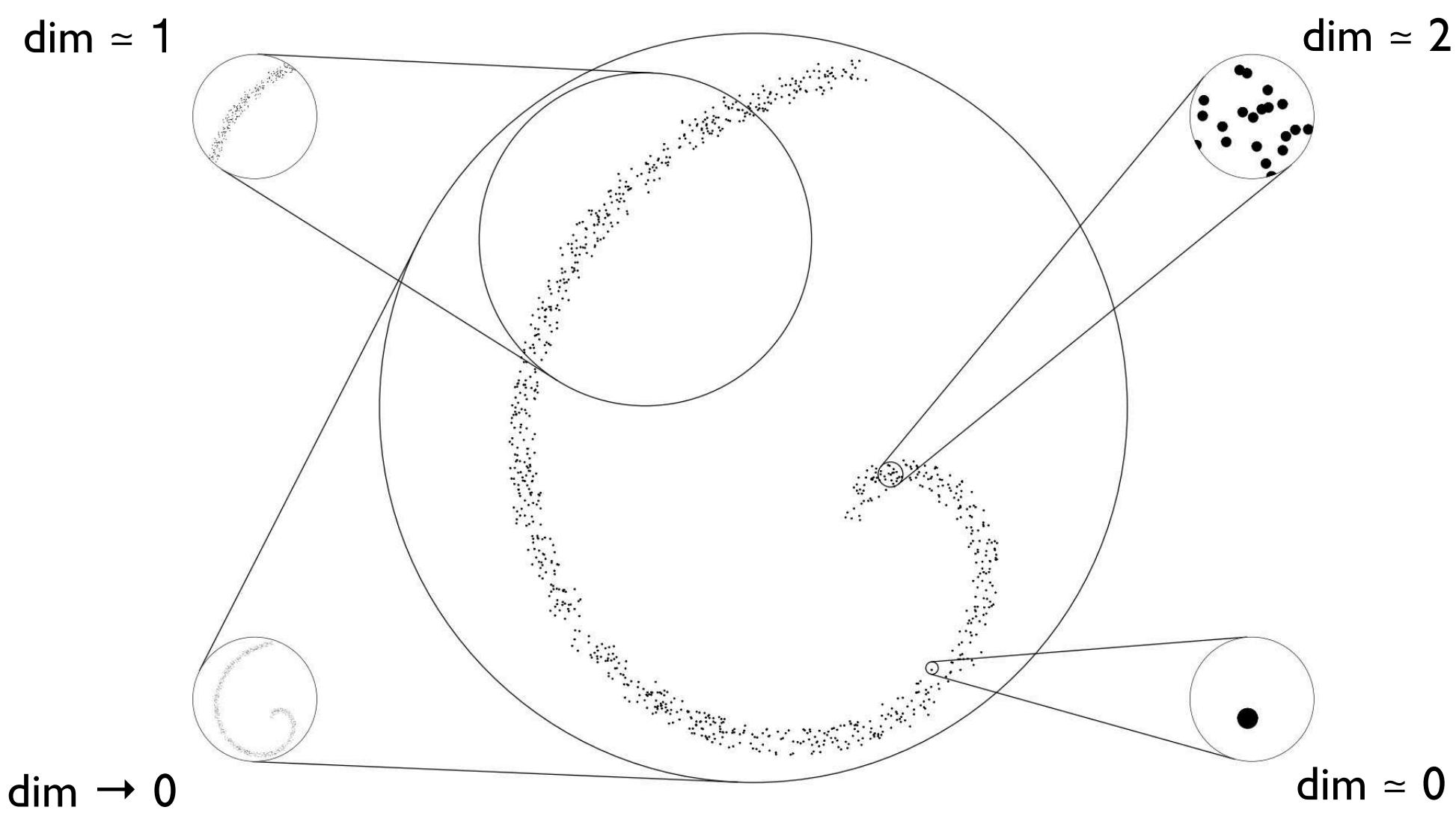
Example jets sprinkled throughout



25 most representative jets ("medoids")
Size is proportional to number of jets associated to that medoid

Quantifying Event-Space Manifolds

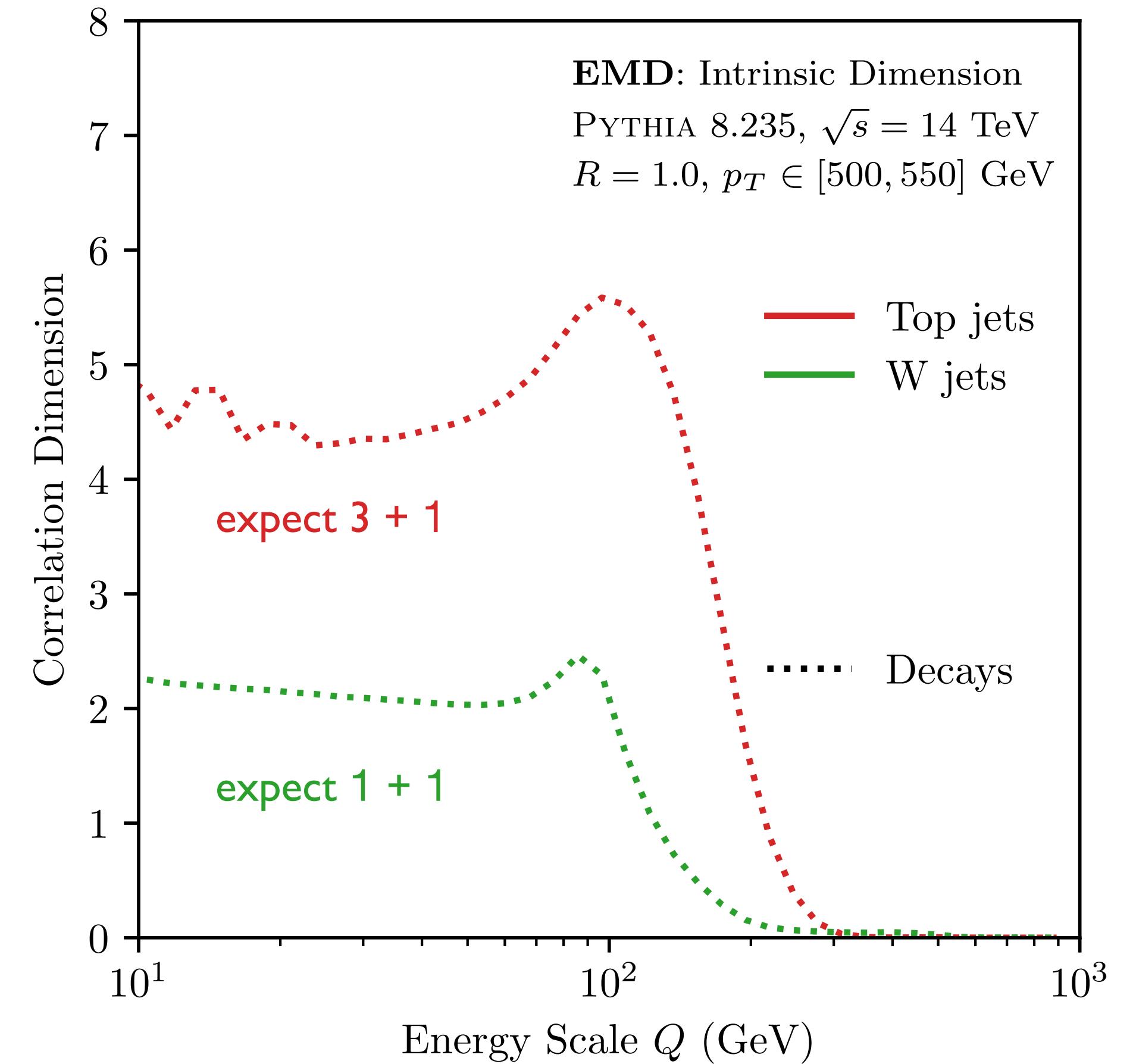
Correlation dimension: how does the # of elements within a ball of size Q change?



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

Correlation dimension lessons:
Decays are "constant" dim. at low Q

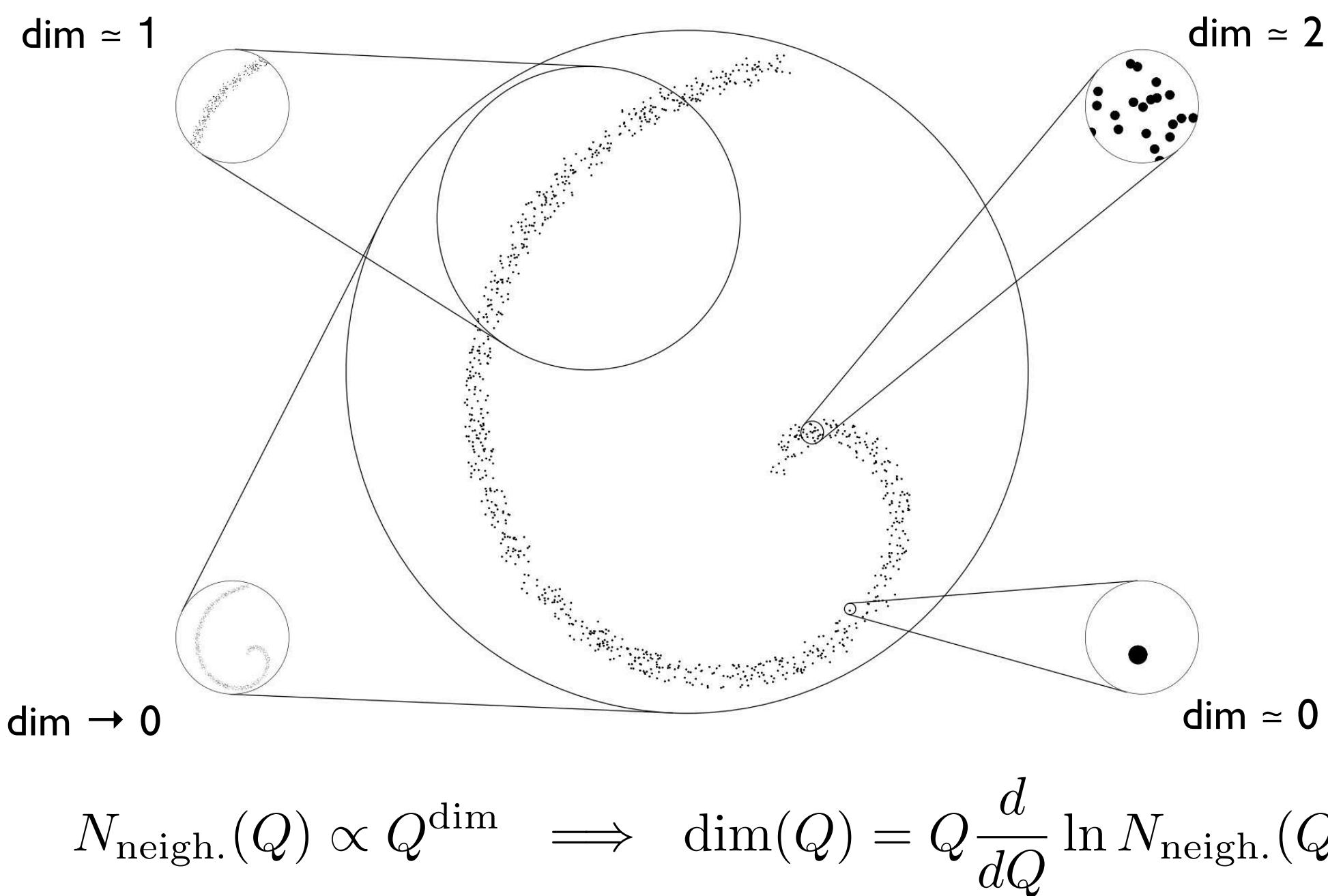
$$\dim(Q) = Q \frac{\partial}{\partial Q} \ln \sum_i \sum_j \Theta(\text{EMD}(\mathcal{E}_i, \mathcal{E}'_j) < Q)$$



[Grassberger, Procaccia, [PRL 1983](#); PTK, Metodiev, Thaler, [PRL 2019](#)]

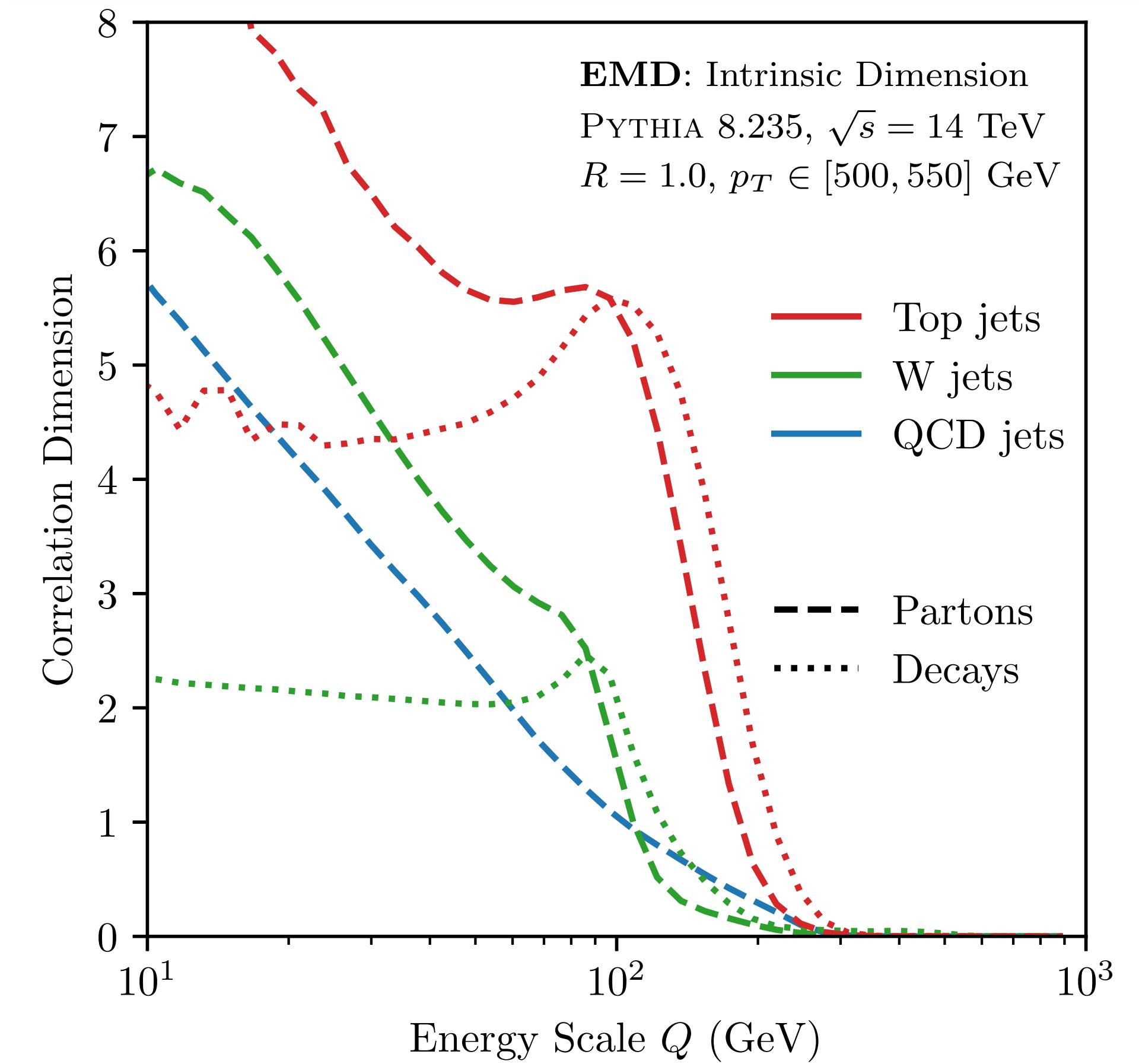
Quantifying Event-Space Manifolds

Correlation dimension: how does the # of elements within a ball of size Q change?



Correlation dimension lessons:
 Decays are "constant" dim. at low Q
 Complexity hierarchy: QCD < W < Top
 Fragmentation increases dim. at smaller scales

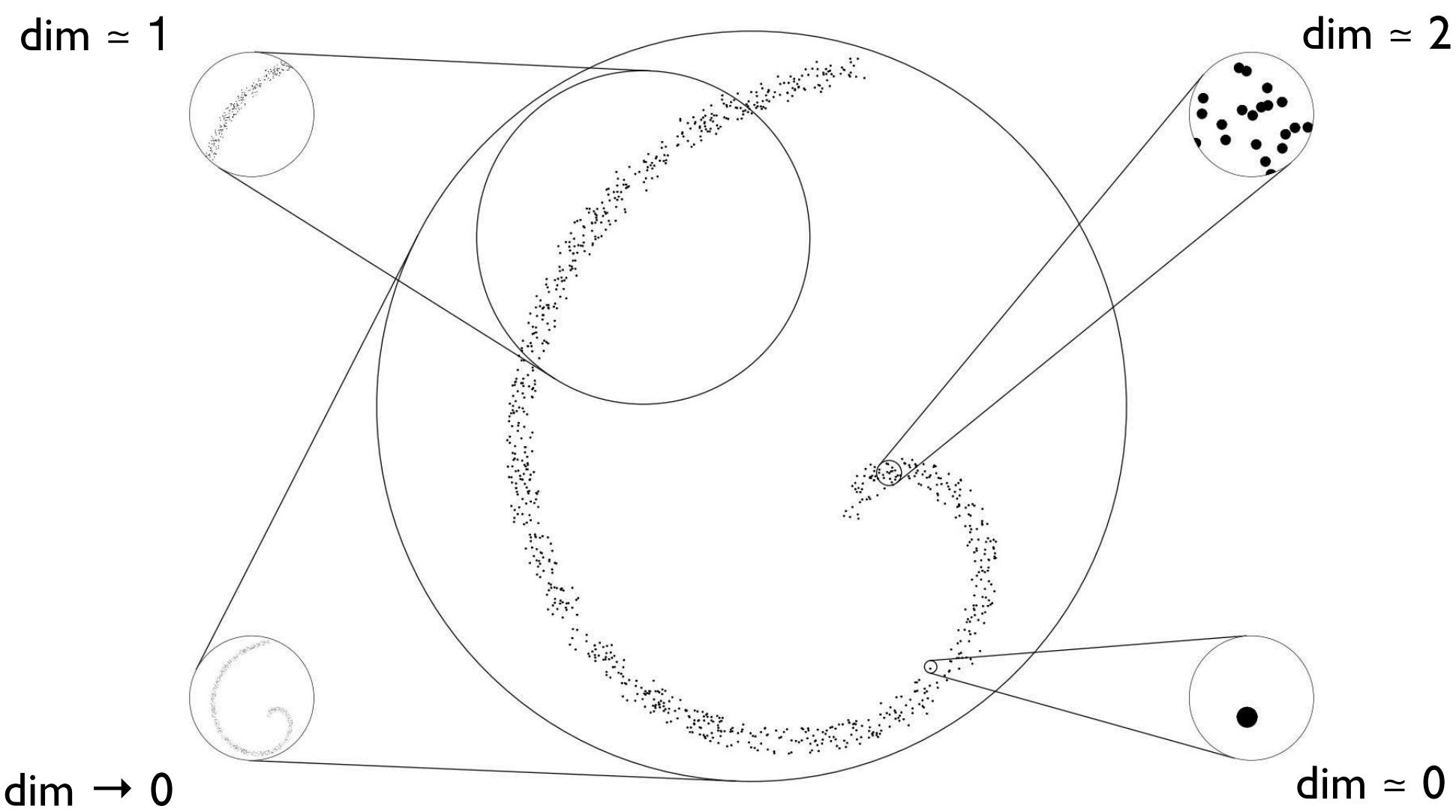
$$\text{dim}(Q) = Q \frac{\partial}{\partial Q} \ln \sum_i \sum_j \Theta(\text{EMD}(\mathcal{E}_i, \mathcal{E}'_j) < Q)$$



[Grassberger, Procaccia, [PRL 1983](#); PTK, Metodiev, Thaler, [PRL 2019](#)]

Quantifying Event-Space Manifolds

Correlation dimension: how does the # of elements within a ball of size Q change?



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

Correlation dimension lessons:

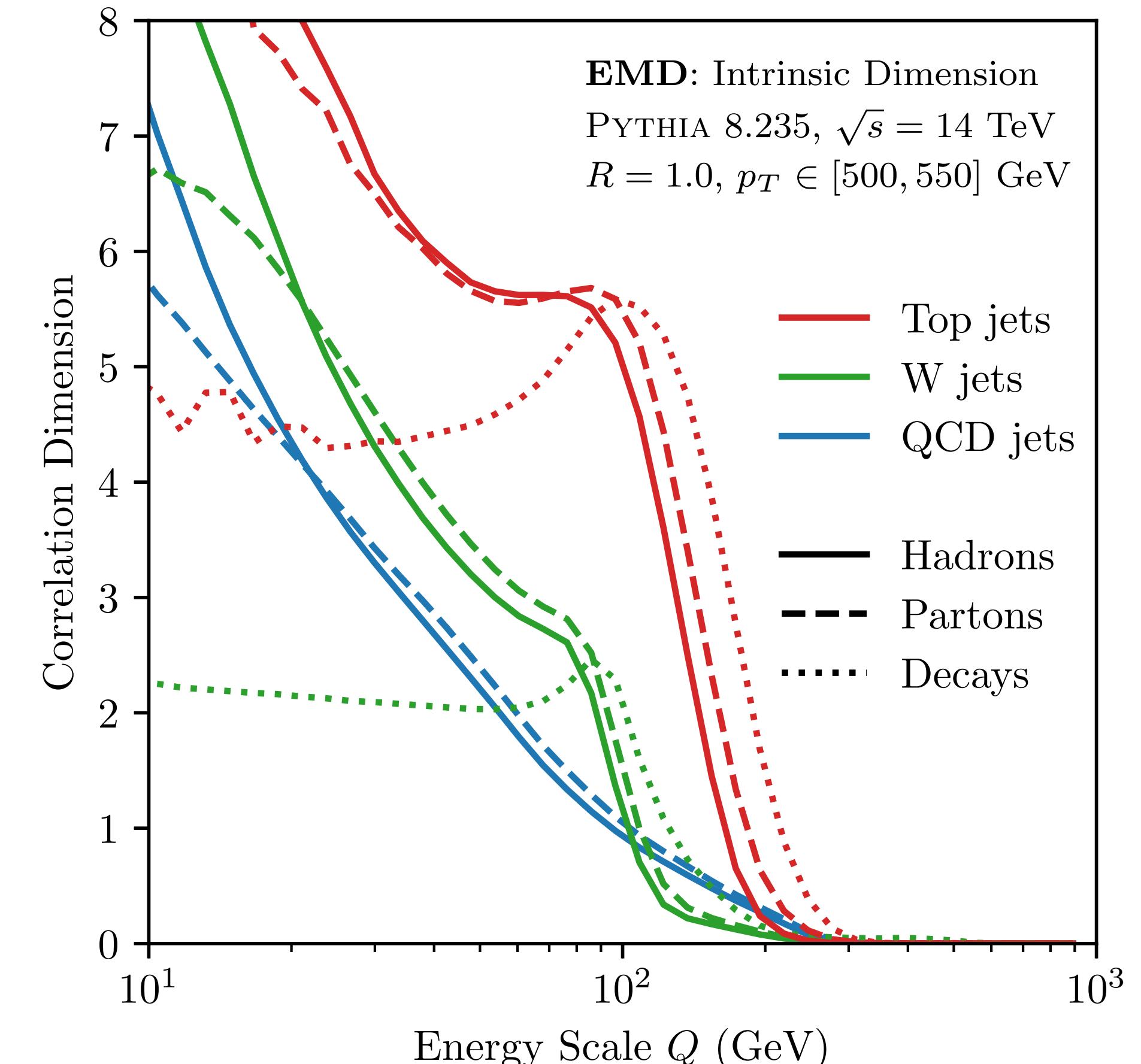
Decays are "constant" dim. at low Q

Complexity hierarchy: QCD < W < Top

Fragmentation increases dim. at smaller scales

Hadronization important around 20-30 GeV

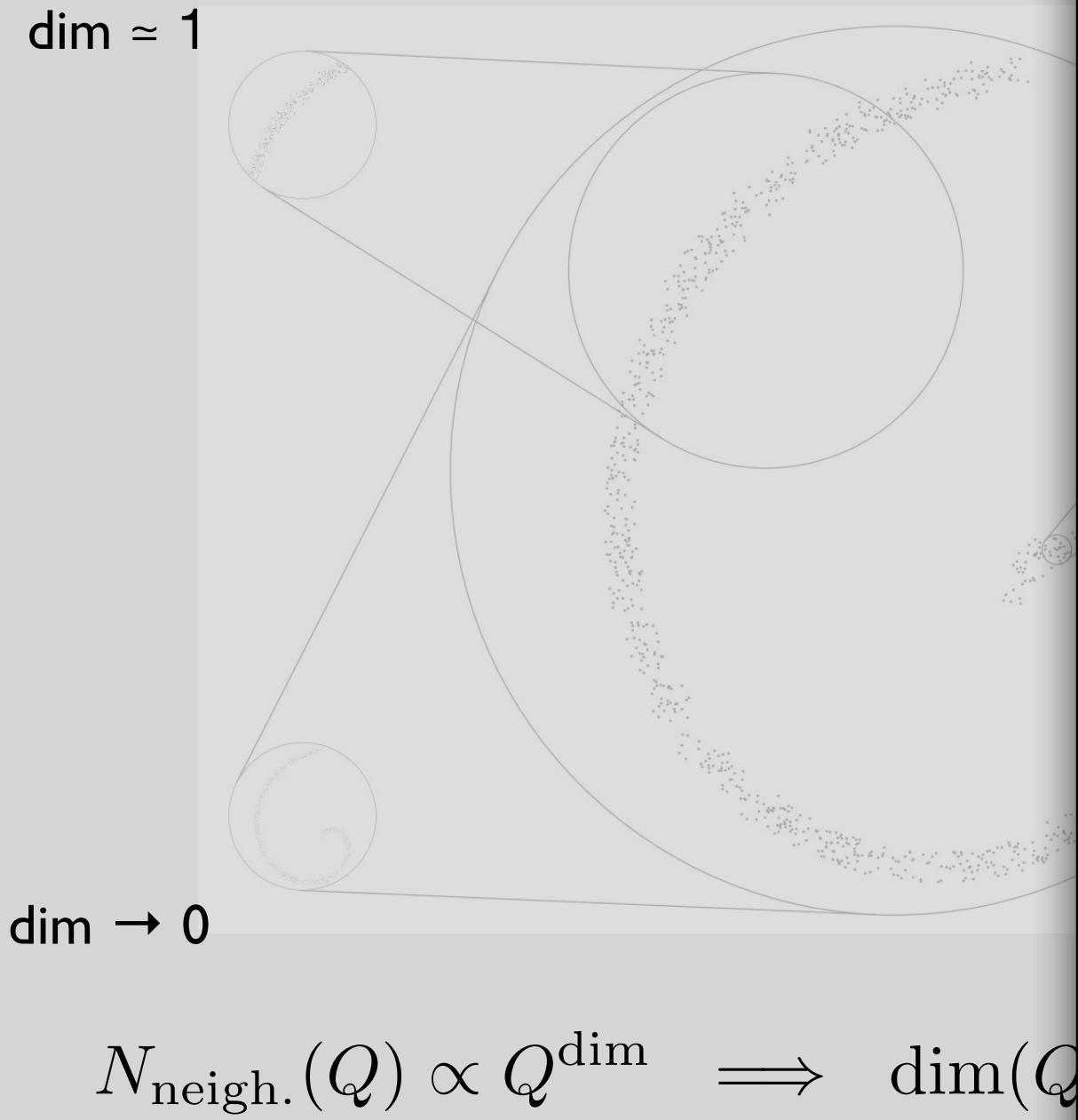
$$\dim(Q) = Q \frac{\partial}{\partial Q} \ln \sum_i \sum_j \Theta(\text{EMD}(\mathcal{E}_i, \mathcal{E}'_j) < Q)$$



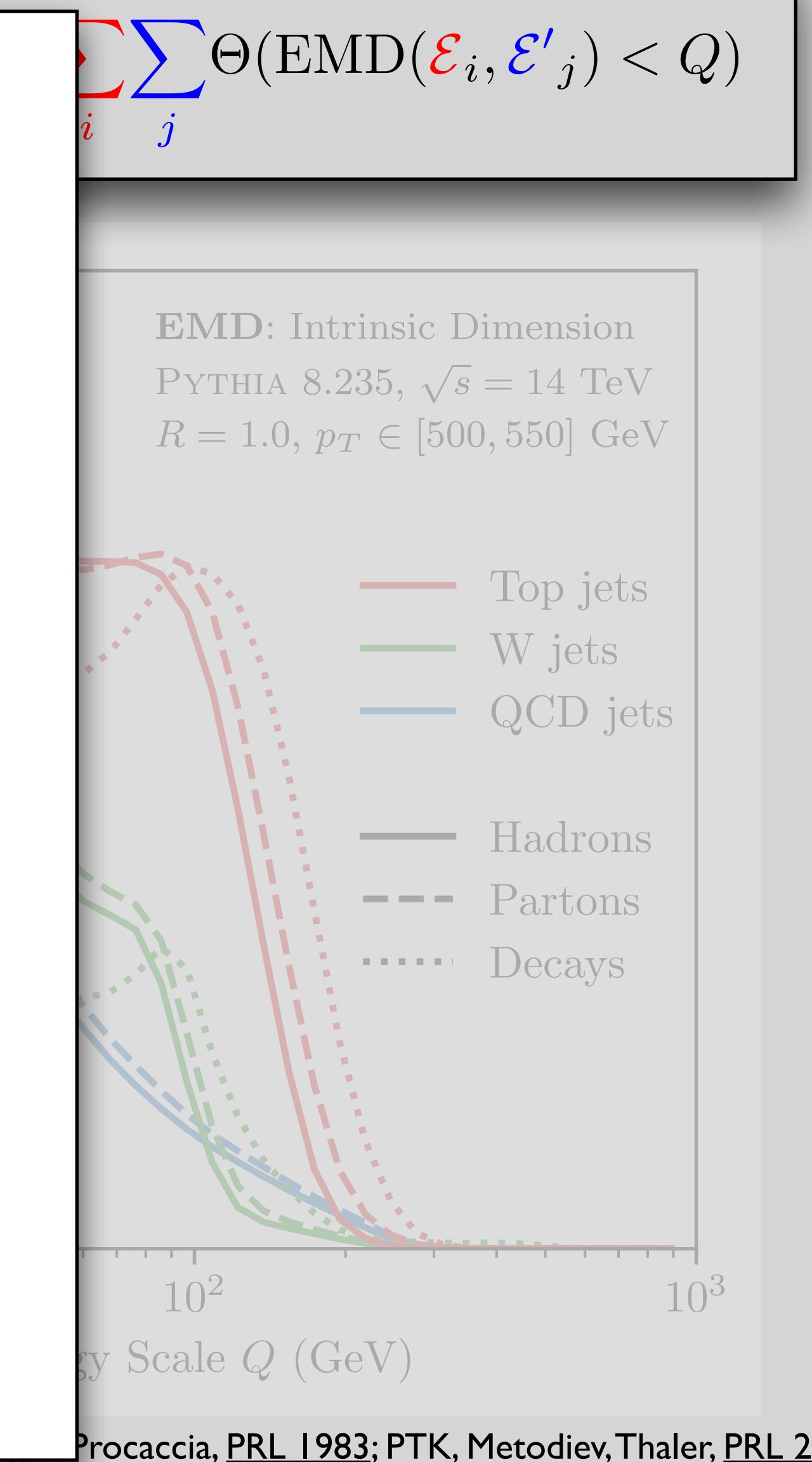
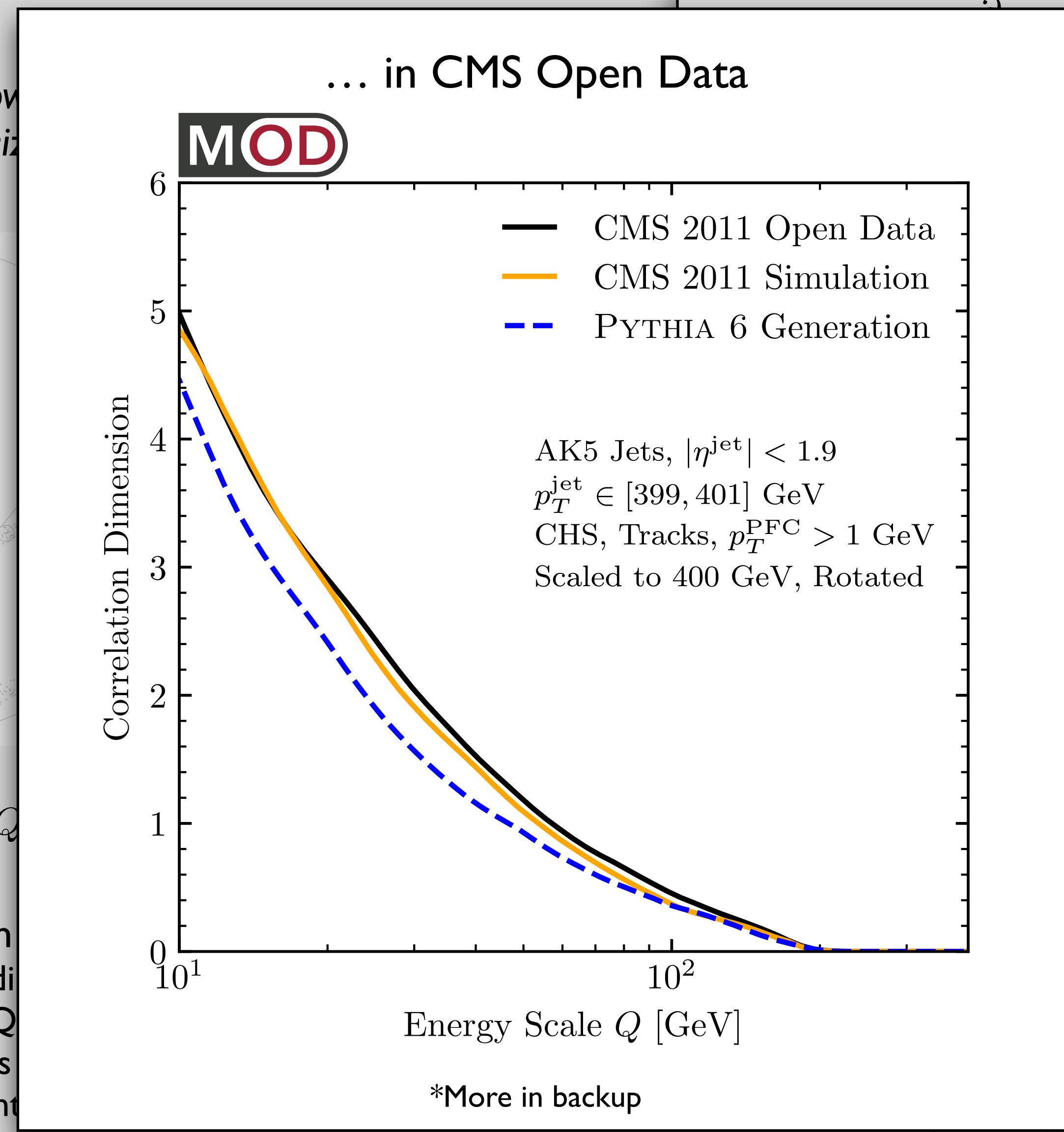
[Grassberger, Procaccia, [PRL 1983](#); PTK, Metodiev, Thaler, [PRL 2019](#)]

Quantifying Event-Space Manifolds

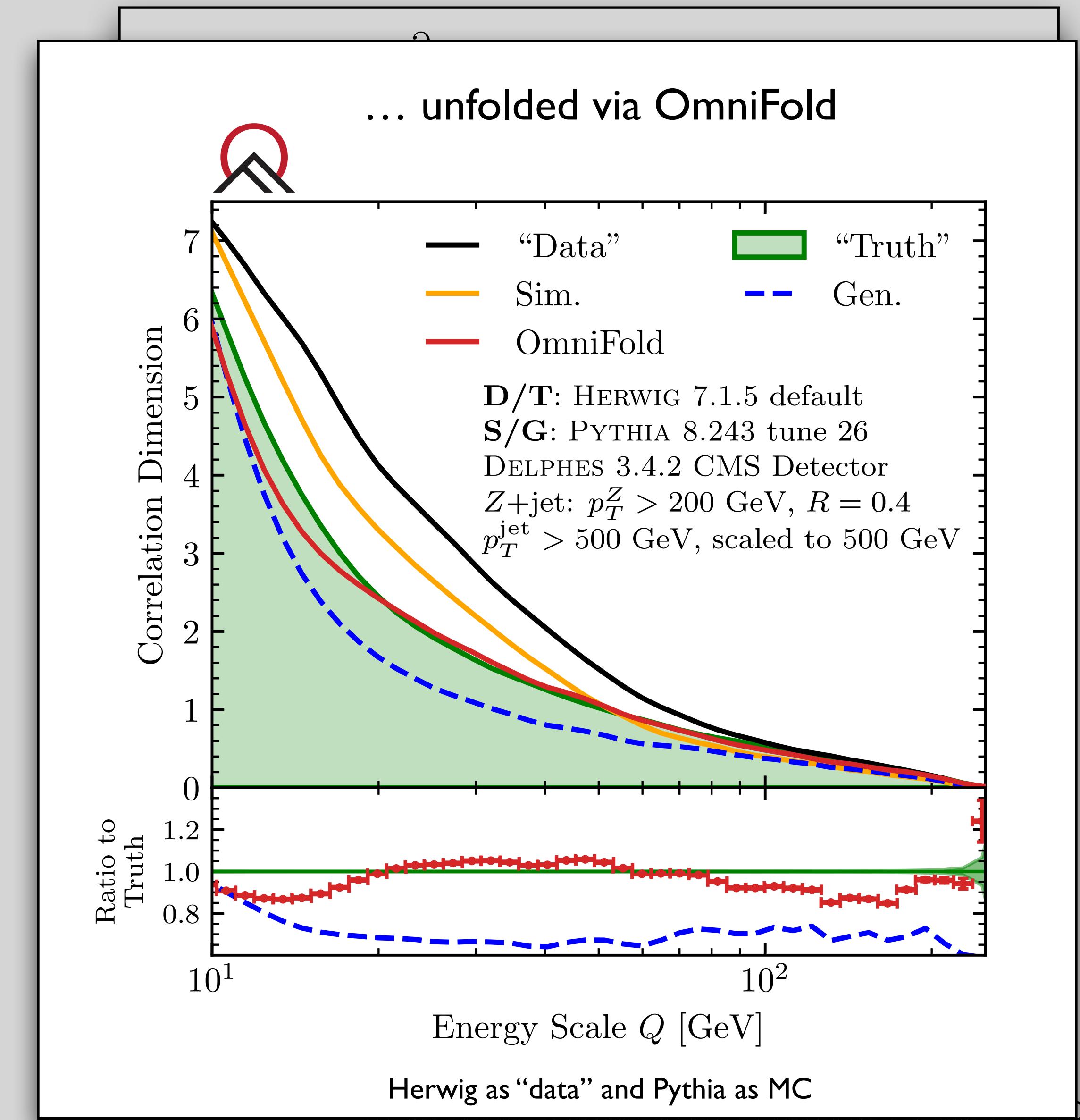
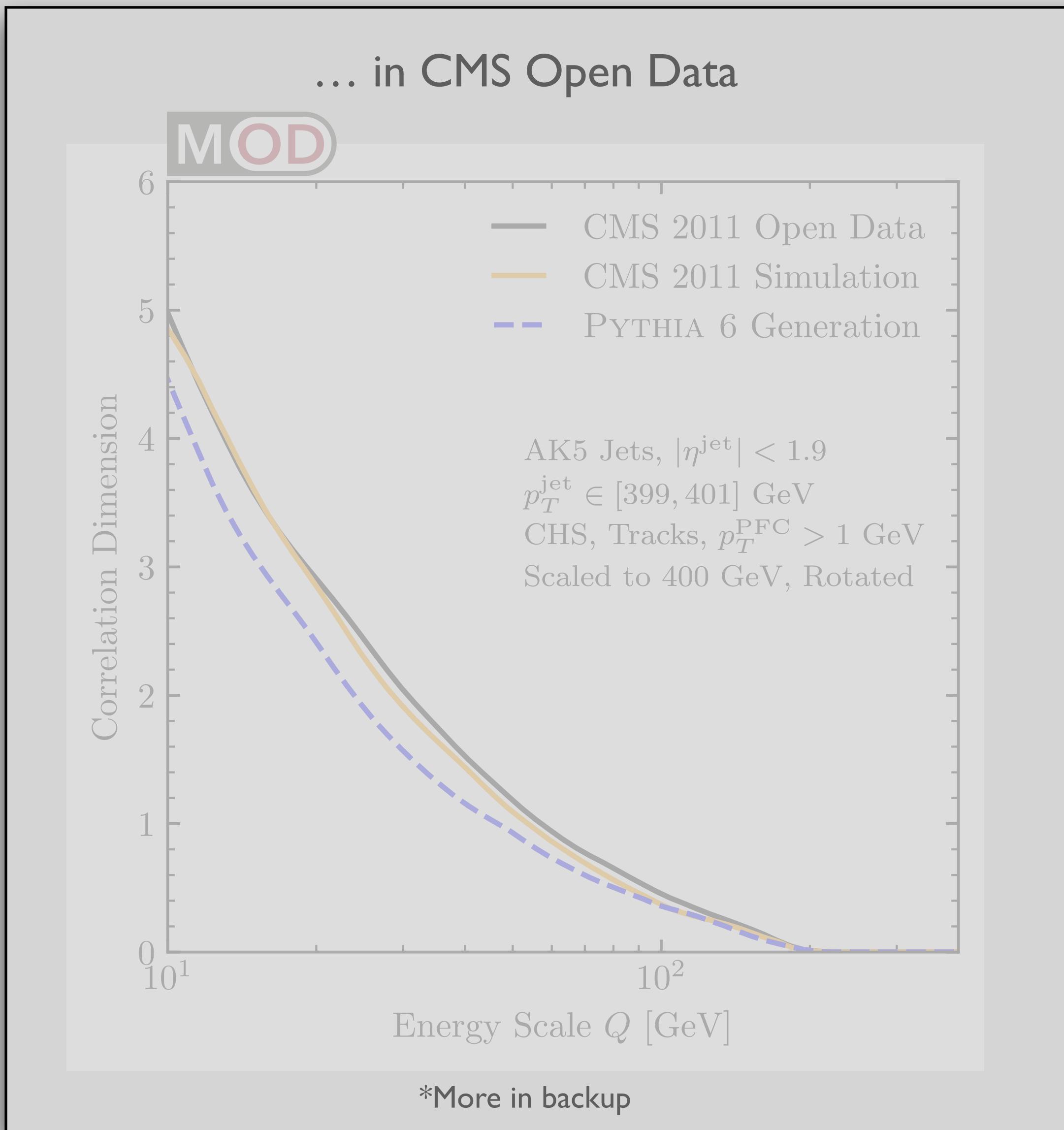
Correlation dimension: how many elements within a ball of size Q

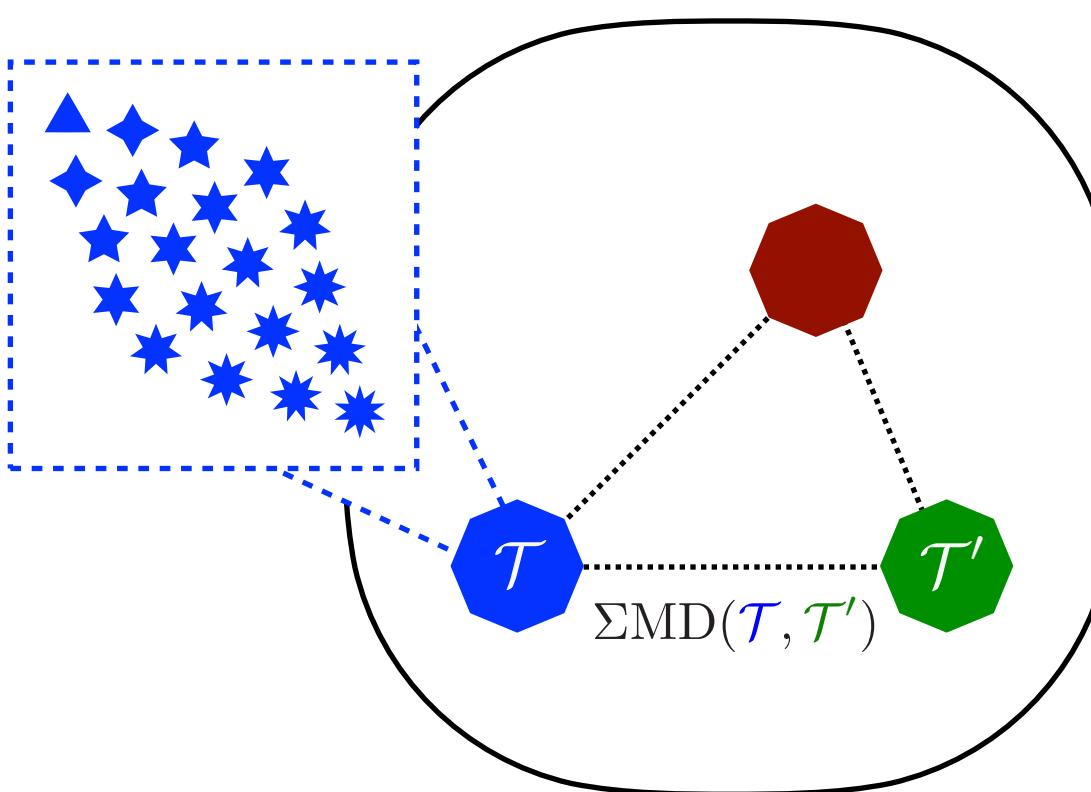
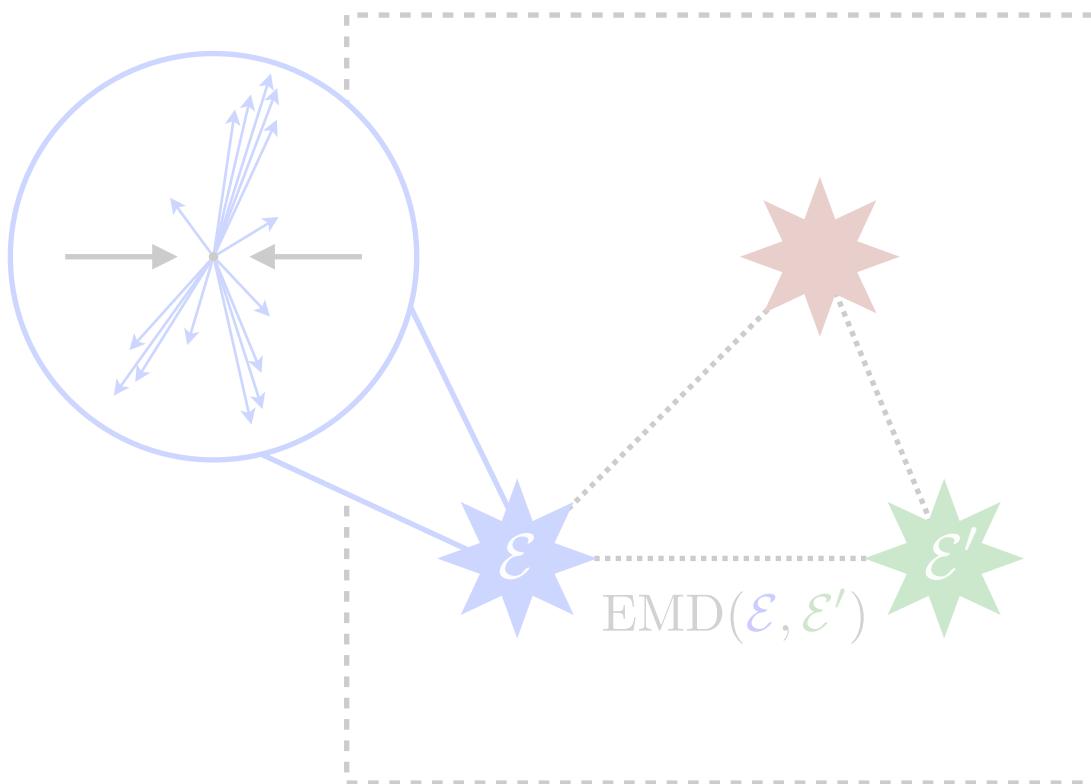


Correlation dimension
Decays are "constant" dim
Complexity hierarchy: Q
Fragmentation increases
Hadronization important



Quantifying Event-Space Manifolds





The (Metric) Space of Events

Revealing Hidden Geometry

[Theory Space]

Templated Metric Construction

Inputs

- “Points” that live in the ground metric space
- “Ground metric” that measures point distances
- “Weights” associated to each point

Output

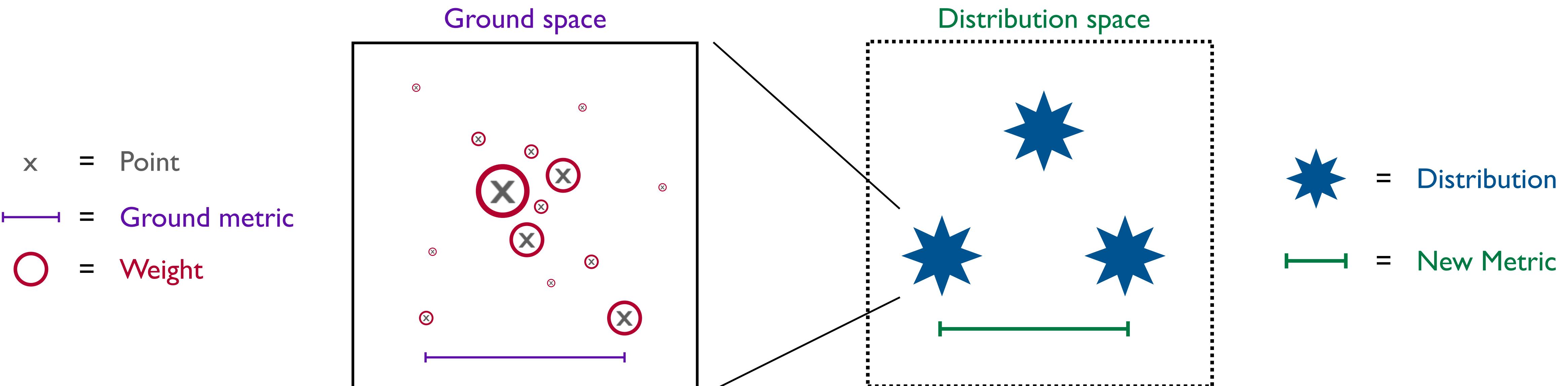
- A new metric for collections of weighted points
- A metric space where these distributions live

p-Wasserstein metric from optimal transport theory

$$W_p(\mu, \nu) = \left(\inf_{J \in \mathcal{J}(\mu, \nu)} \int_{M \times M} d(x, y)^p dJ(x, y) \right)^{1/p}$$

(M, d) , metric space

$\mathcal{J}(\mu, \nu)$, space of joint distributions with marginals μ, ν



Templated Metric Construction – Energy Mover’s Distance

[PTK, Metodiev, Thaler, PRL 2019]

Inputs

- “Points” that live in the ground metric space
- “Ground metric” that measures point distances
- “Weights” associated to each point

p-Wasserstein metric from optimal transport theory

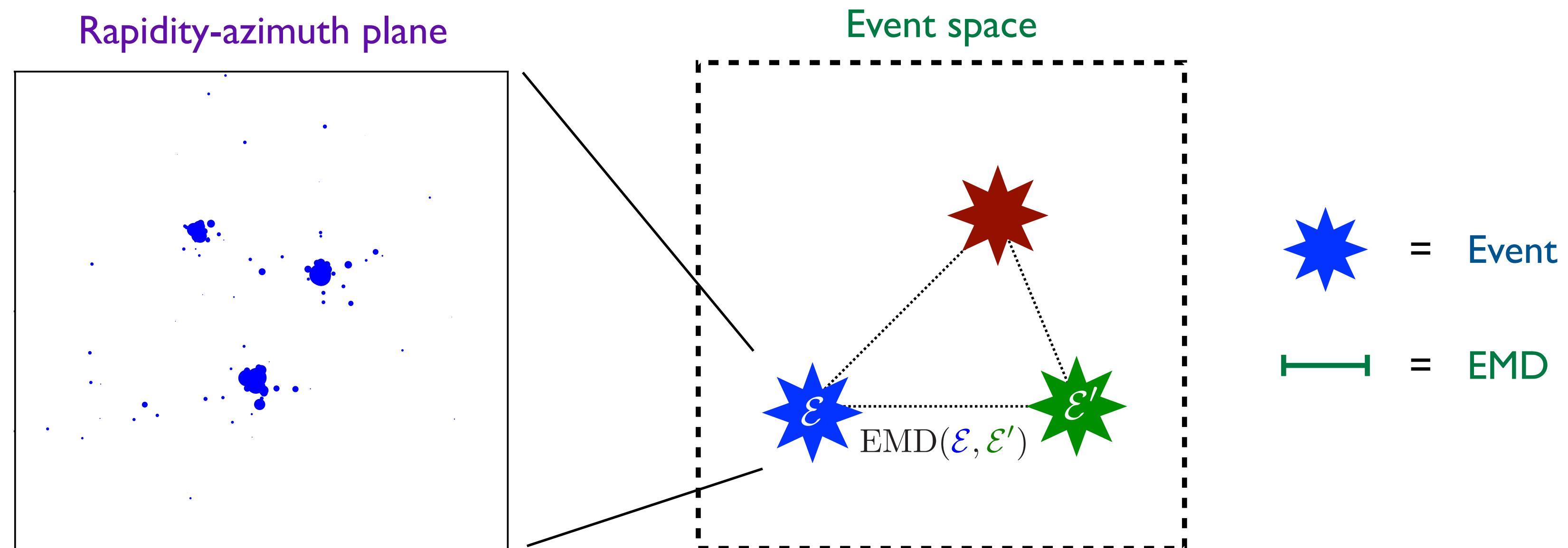
$$W_p(\mu, \nu) = \left(\inf_{J \in \mathcal{J}(\mu, \nu)} \int_{M \times M} d(x, y)^p dJ(x, y) \right)^{1/p}$$

(M, d) , metric space

$\mathcal{J}(\mu, \nu)$, space of joint distributions
with marginals μ, ν

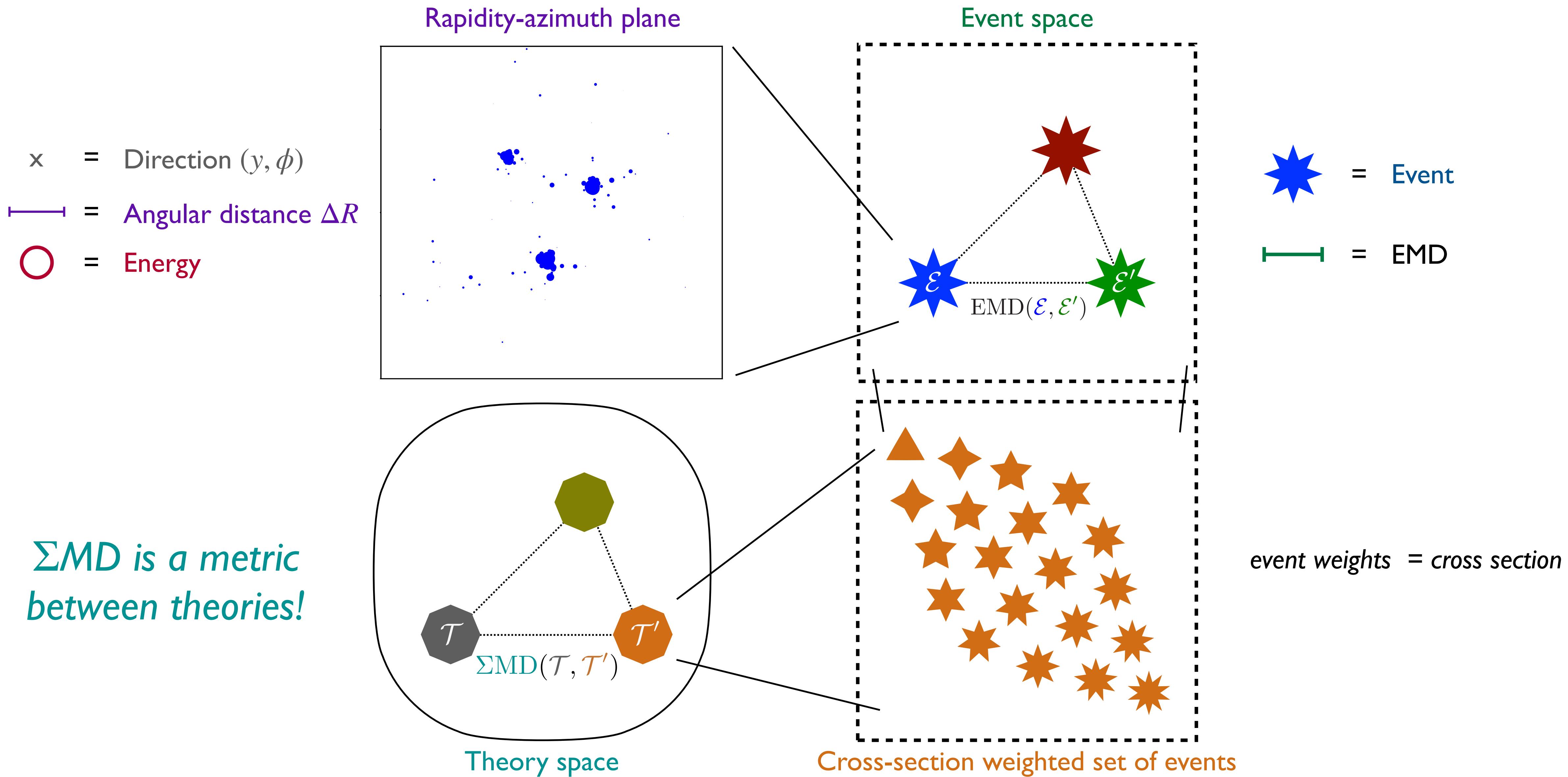
Output

- A new metric for collections of weighted points
- A metric space where these distributions live



Bootstrapping to the Cross-Section Mover's Distance (Σ MD)

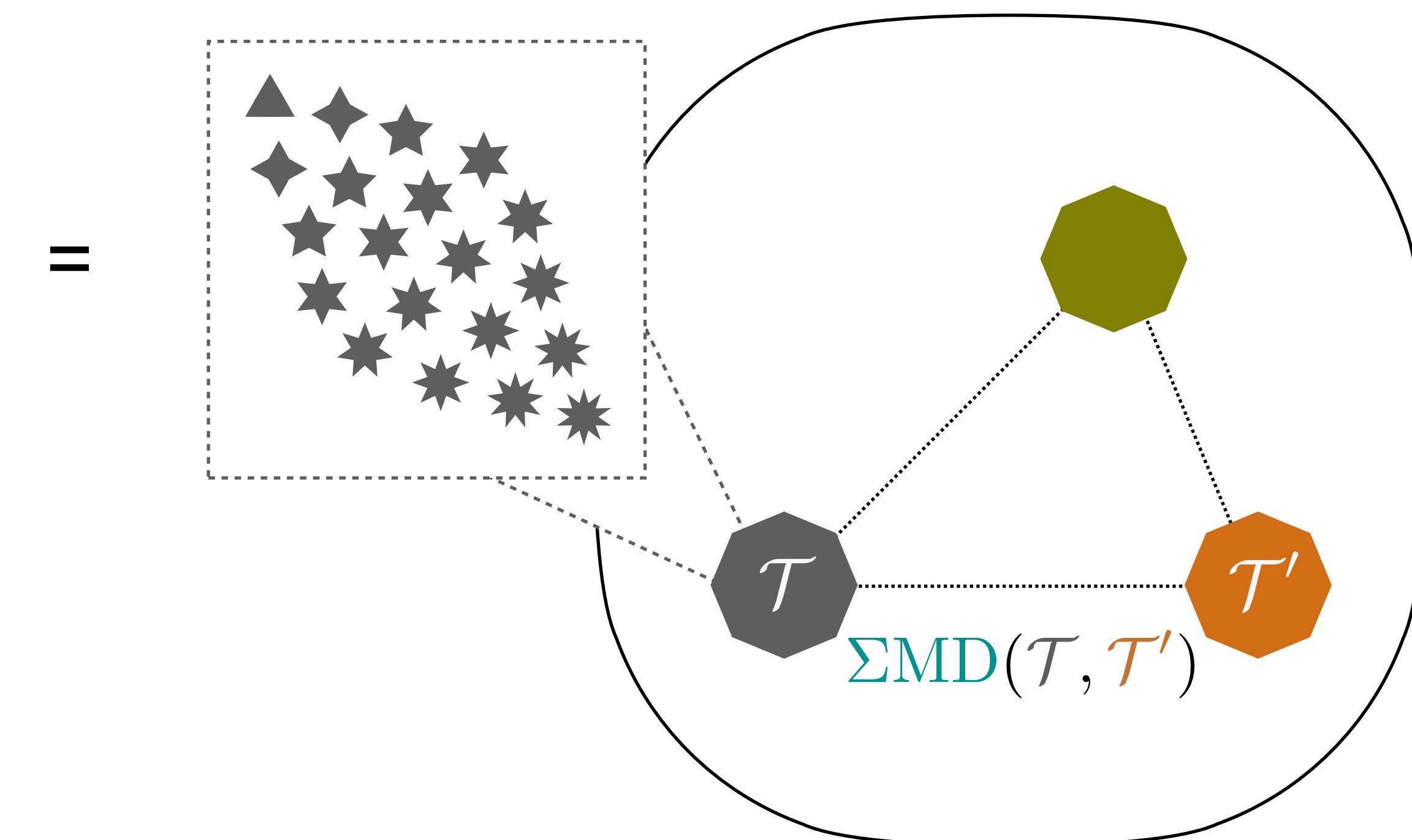
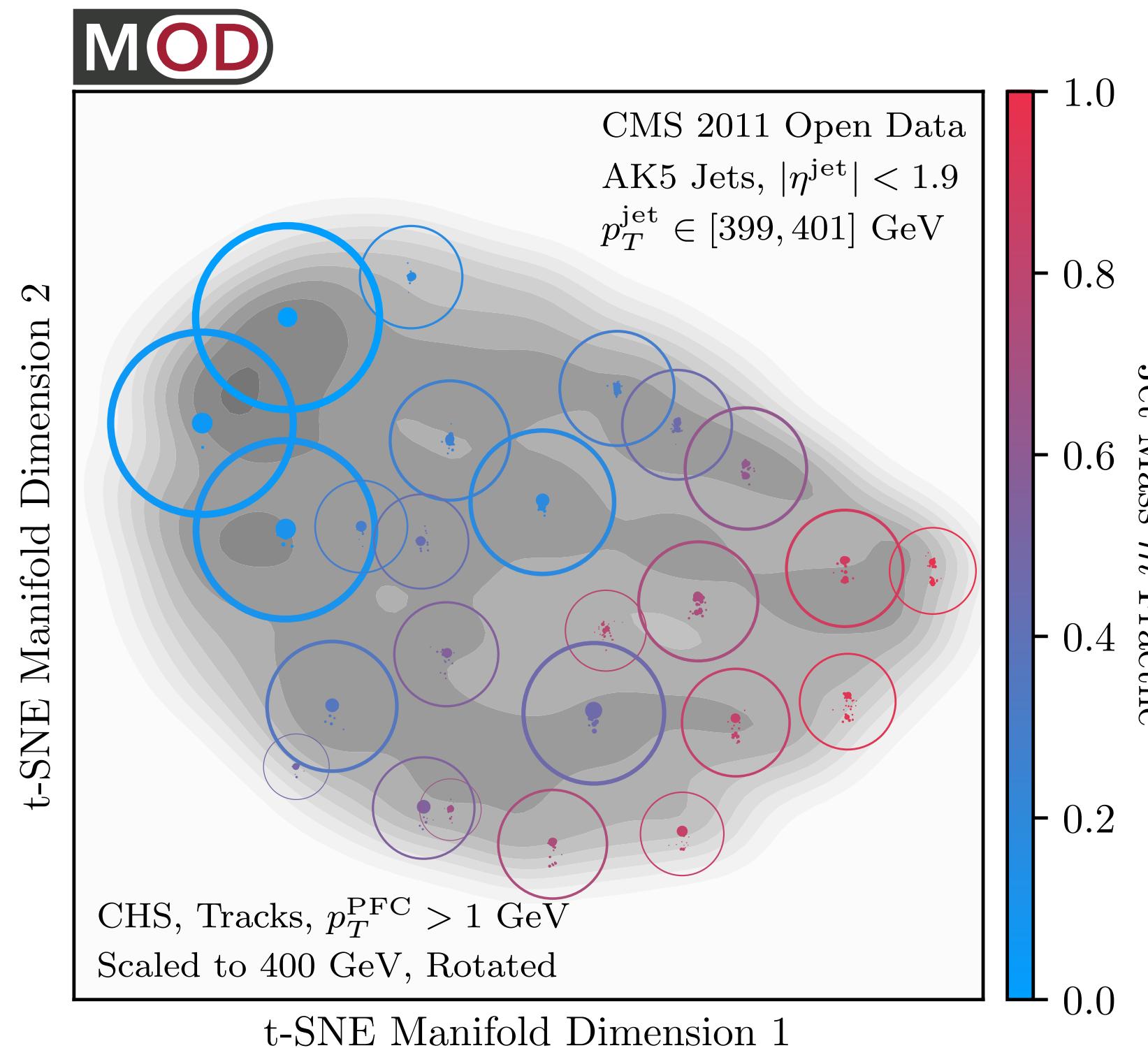
[PTK, Metodiev, Thaler, 2004.04159]



The Space of Theories

[PTK, Metodiev, Thaler, 2004.04159]

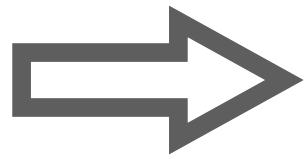
ΣMD provides a rigorous construction of theory space



*Theories are distinguished by their energy flows only

Applications of Σ MD and the Space of Theories

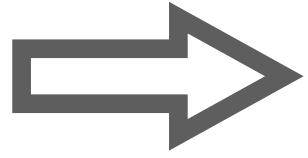
N -(sub)jettiness



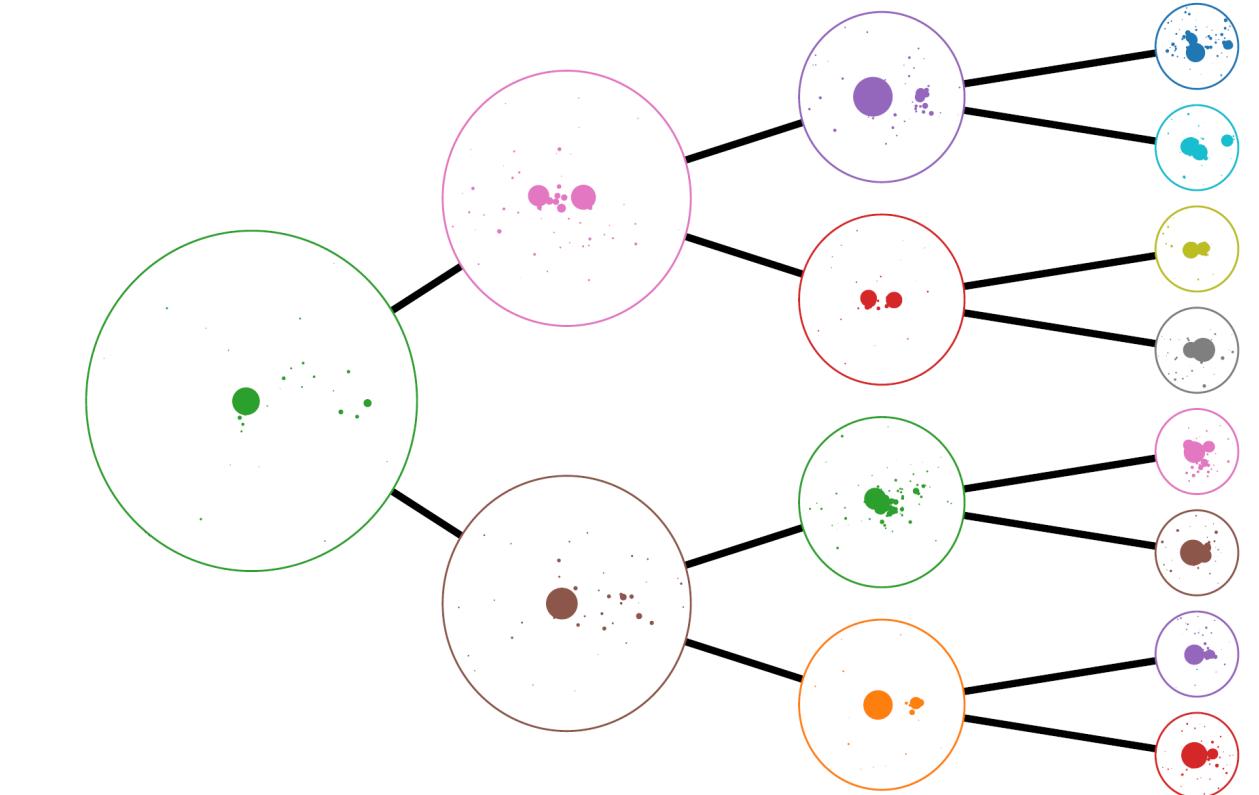
k -eventiness defined

$$\begin{aligned}\mathcal{V}_k^{(\gamma)}(\{\sigma_i, \mathcal{E}_i\}) &= \min_{\mathcal{K}_1, \dots, \mathcal{K}_k} \sum_{i=1}^N \sigma_i \min \{\text{EMD}(\mathcal{E}_i, \mathcal{K}_1), \dots, \text{EMD}(\mathcal{E}_i, \mathcal{K}_k)\}^\gamma \\ \mathcal{V}_k^{(\gamma)}(\mathcal{T}) &= \min_{|\mathcal{T}'|=k} \Sigma\text{MD}_\gamma(\mathcal{T}, \mathcal{T}')\end{aligned}$$

Jet clustering

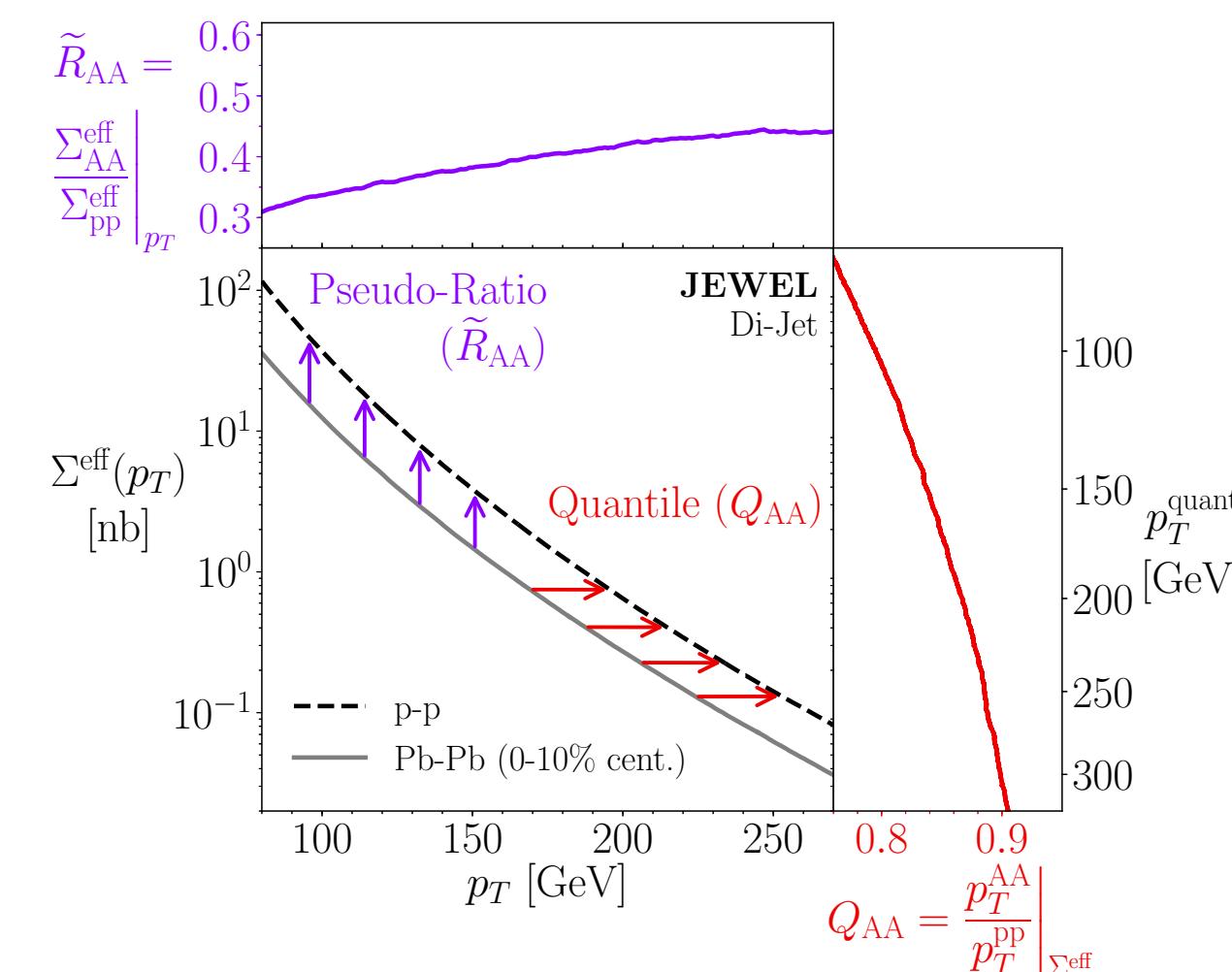
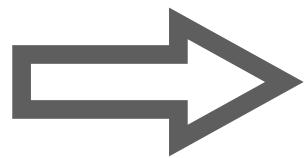


Event clustering enabled
– Exclusive cone finding
– Sequential recombination



Jet quenching in HI collisions

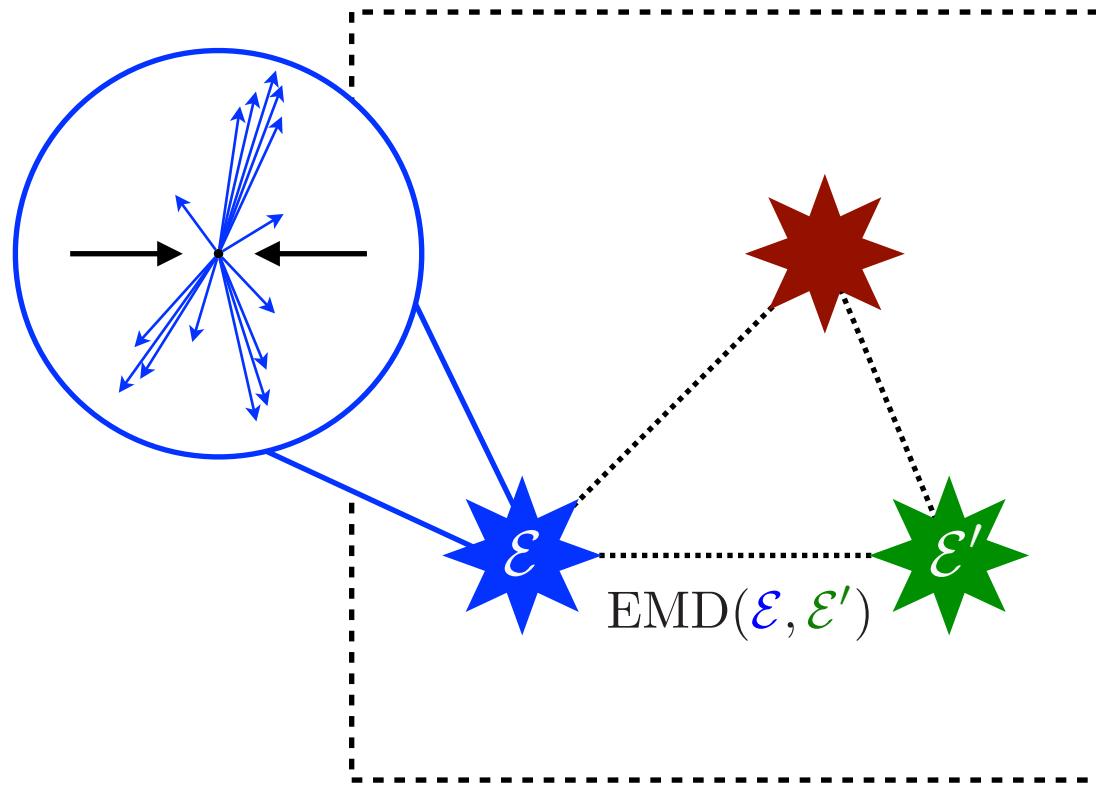
[Brewer, Milhano, Thaler, PRL 2019]



Quantile matching:
 $\Sigma_{\text{pp}}^{\text{eff}}(p_T^{\text{quant}}) \equiv \Sigma_{\text{AA}}^{\text{eff}}(p_T^{\text{AA}})$

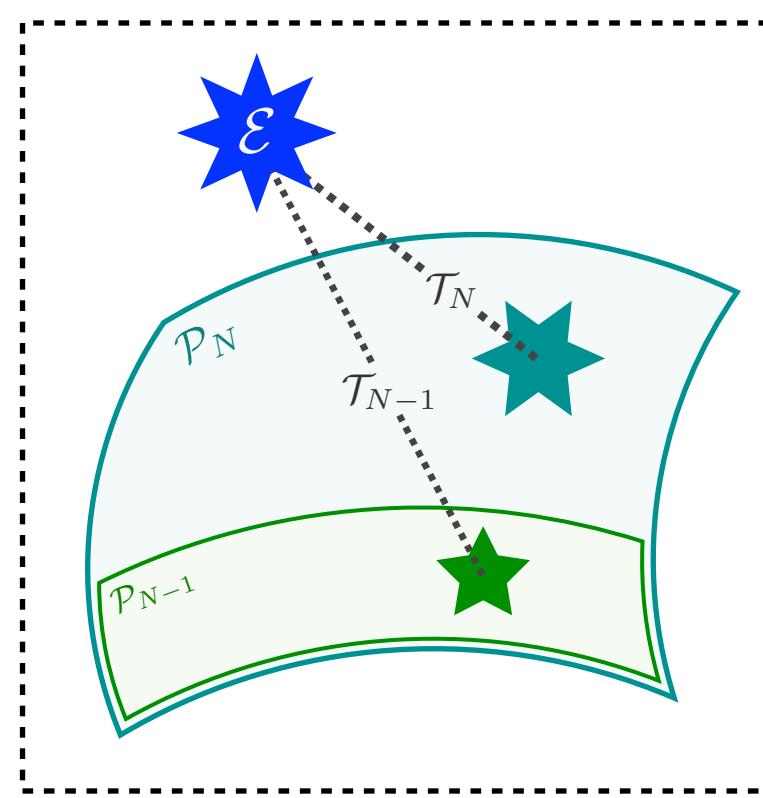
...is exactly a theory moving problem!
 $p_T^{\text{quant}} = \text{TM}(\mathcal{T}_{\text{AA}}, \mathcal{T}_{\text{pp}})[p_T^{\text{AA}}]$

↑
optimal p_T -only theory movement



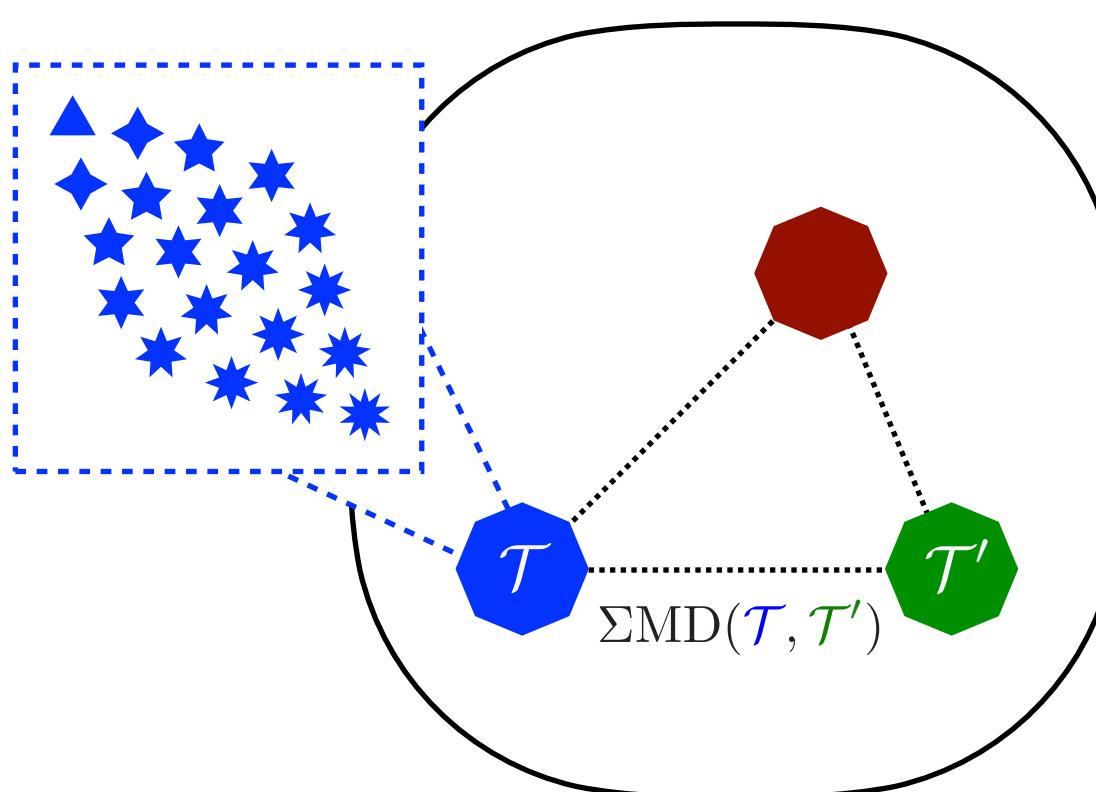
The (Metric) Space of Events

- Energy flow is theoretically and experimentally robust
- EMD metrizes the space of energy flows (events)
- Manifolds in the space of events can be visualized and quantified



Revealing Hidden Geometry

- Event space exhibits a rich geometry that can be probed using the EMD
- Decades worth of collider techniques are naturally described in this geometry
- Many new techniques are suggested, and new light is shed on old ones



[Theory Space]

- Is rigorously constructed using the cross-section mover's distance Σ MD
- Σ MD uses the EMD as ground metric and cross sections as weights
- Theories can be explored with tools developed for events (e.g. k -eventiness)

EnergyFlow Python Package

pip3 install energyflow wasserstein

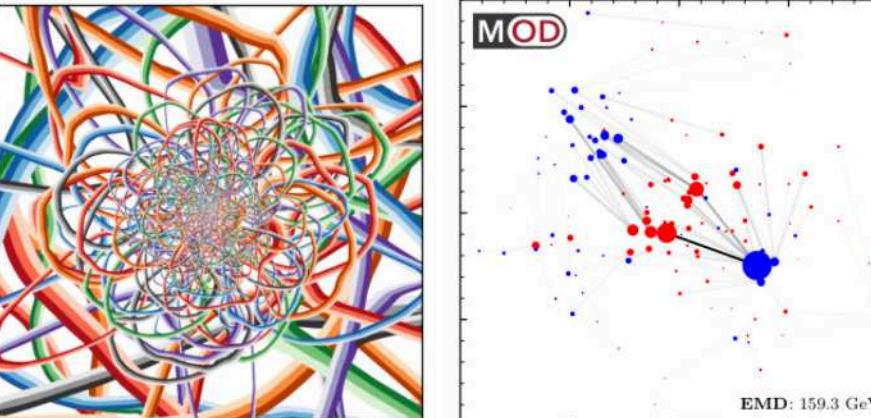
Parallelized EMD calculations via the Wasserstein library
EFN/PFN implementations in TensorFlow/Keras
Detailed examples, demos, and documentation

Interfaces with CMS 2011A Jet Primary Dataset hosted on Zenodo

 EnergyFlow

Docs » Home

Welcome to EnergyFlow



EnergyFlow is a Python package containing a suite of particle physics tools:

- **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. Available from version `0.10.0` onward.
- **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. Available from version `0.10.0` onward.
- **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 facilitate the computation of the EMD between events based on an underlying implementation provided by the Python Optimal Transport (POT) library. Available from version `0.11.0` onward.
- **Energy Flow Moments:** EFM are moments built out of particle energies and momenta that can be evaluated in linear time in the number of particles. They provide a highly efficient means of implementing $\beta = 2$ EFPs and are also very useful for reasoning about linear redundancies that appear between EFPs. Available from version `1.0.0` onward.

The EnergyFlow package also provides easy access to particle physics datasets and useful supplementary features:

- **CMS Open Data in MOD HDF5 Format:** Reprocessed datasets from the CMS Open Data,

GitHub Next »

hub.gke.mybinder.org/user/pkomiske-energyflow-0ls4z3ee/notebooks/demos/EMD%20Demo.ipynb

jupyter EMD Demo (autosaved)

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

EMD Demo

[EnergyFlow website](#)

In this tutorial, we demonstrate how to compute EMD values for particle physics events. The core of the computation is done using the [Python Optimal Transport](#) library with EnergyFlow providing a convenient interface to particle physics events. Batching functionality is also provided using the builtin multiprocessing library to distribute computations to worker processes.

Energy Mover's Distance

The Energy Mover's Distance was introduced in [1902.02346](#) as a metric between particle physics events. Closely related to the Earth Mover's Distance, the EMD solves an optimal transport problem between two distributions of energy (or transverse momentum), and the associated distance is the "work" required to transport supply to demand according to the resulting flow. Mathematically, we have

$$\text{EMD}(\mathcal{E}, \mathcal{E}') = \min_{\sum_j f_{ij} \leq E_i, \sum_i f_{ij} \leq E'_j, \sum_j f_{ij} = \min\left(\sum_i E_i, \sum_j E'_j\right)} \sum_{i,j} f_{ij} \frac{\theta_{ij}}{R} + \sum_i E_i - \sum_j E'_j,$$

Imports

```
In [1]: import numpy as np
%matplotlib inline
import matplotlib.pyplot as plt
import energyflow as ef
```

Plot Style

```
In [2]: plt.rcParams['figure.figsize'] = (4,4)
plt.rcParams['figure.dpi'] = 120
plt.rcParams['font.family'] = 'serif'
```

Load EnergyFlow Quark/Gluon Jet Samples

```
In [3]: # load quark and gluon jets
x, y = ef.eg_jets.load(2000, pad=False)
num = 750

# the jet radius for these jets
R = 0.4

# process jets
G, Qs = [], []
for arr,events in [(G, X[y==0]), (Qs, X[y==1])]:
    for i,x in enumerate(events):
        if i > num:
            break

        # ignore padded particles and removed particle id information
        x = x[x[:,0] > 0,:]

        # center jet according to pt-centroid
        yphi_avg = np.average(x[:,1:3], weights=x[:,0], axis=0)
        x[:,1:3] -= yphi_avg

        # mask out any particles farther than R=0.4 away from center (rare)
        x = x[np.linalg.norm(x[:,1:3], axis=1) <= R]

        # add to list
        G.append(x)
        Qs.append(x)
```

zenodo

August 8, 2019

CMS 2011A Open Data | Jet Primary Dataset | pT > 375 GeV | MOD HDF5 Format

Komiske, Patrick, Mastandrea, Radha, Metodiev, Eric, Naik, Preksha, Thaler, Jesse

A dataset of 1,785,625 jets from the Jet Primary Dataset of the CMS 2011A Open Data reprocessed into the MOD HDF5 format. Jets are selected from the hardest two anti- $R=0.5$ jets in events passing the Jet300 High Level Trigger and are required to have $p_T^{jet} > 375$ GeV, where p_T^{jet} includes a jet energy correction factor. Particle Flow Candidates (PFCs) for each jet are provided and include information about the PFC kinematics, PDG ID, and vertex. Additionally, jets have metadata describing their kinematics and provenance in the original CMS AOD files.

For additional details about the dataset, please see the accompanying paper, Exploring the Space of Jets with CMS Open Data. There, jets were further restricted to have $|y^{jet}| < 1.9$ to ensure tracking coverage and have 'medium' quality to reject fake jets.

The supported method for downloading, reading, and using this dataset is through the EnergyFlow Python package, which has additional documentation about how to read and use this and related datasets. Should any problems be encountered, please submit an issue on GitHub.

There are corresponding datasets of simulated jets organized by hard parton \hat{p}_T also available on Zenodo:

SIM/GEN QCD Jets 170-300 GeV	274	432
SIM/GEN QCD Jets 300-470 GeV	views	downloads
SIM/GEN QCD Jets 470-600 GeV	See more details...	
SIM/GEN QCD Jets 600-800 GeV		
SIM/GEN QCD Jets 800-1000 GeV		
SIM/GEN QCD Jets 1000-1400 GeV		
SIM/GEN QCD Jets 1400-1800 GeV		
SIM/GEN QCD Jets 1800-oo GeV		

Files (2.0 GB)

Name	Size
CMS_Jet300_pT375-infGeV_0_compressed.h5	111.2 MB
CMS_Jet300_pT375-infGeV_10_compressed.h5	110.8 MB
CMS_Jet300_pT375-infGeV_11_compressed.h5	111.3 MB
CMS_Jet300_pT375-infGeV_12_compressed.h5	111.7 MB
CMS_Jet300_pT375-infGeV_13_compressed.h5	111.3 MB
CMS_Jet300_pT375-infGeV_14_compressed.h5	111.2 MB
CMS_Jet300_pT375-infGeV_15_compressed.h5	111.0 MB

OpenAIRE

Publication date: August 8, 2019

DOI: DOI: 10.5281/zenodo.3340205

Keyword(s): cms open data hpc jet substructure hep physics

Related identifiers: Supplement to arXiv:1908.08542

License (for files): Creative Commons Attribution 4.0 International

Versions

Version v0	Aug 8, 2019
10.5281/zenodo.3340205	

Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.3340204. This DOI represents all versions, and will always resolve to the latest one. [Read more](#).

Share

GitHub Next »

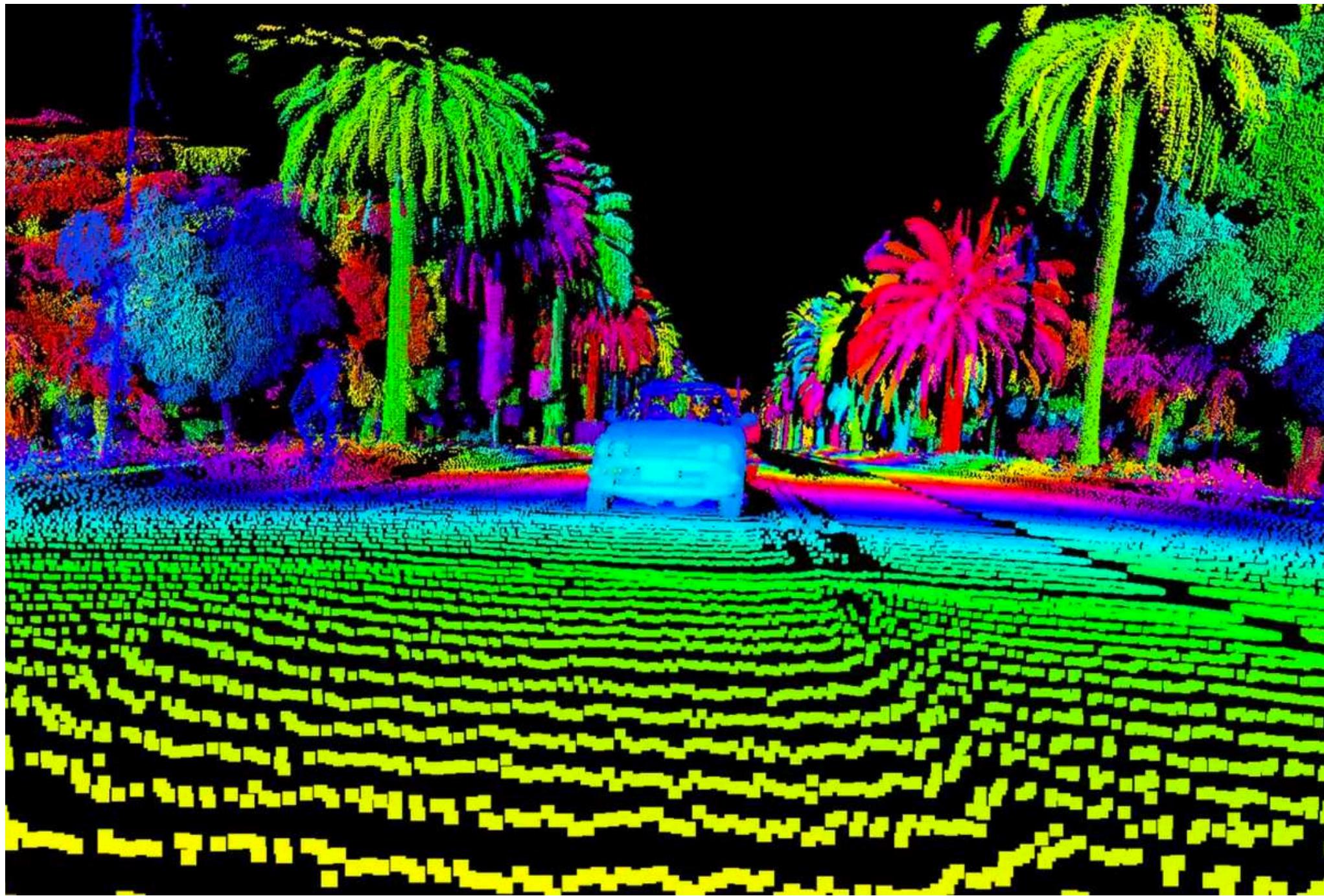
Additional Slides

Neural Network Architectures for Particle Physics

Maximally appropriate ML architectures respect symmetries of the underlying data

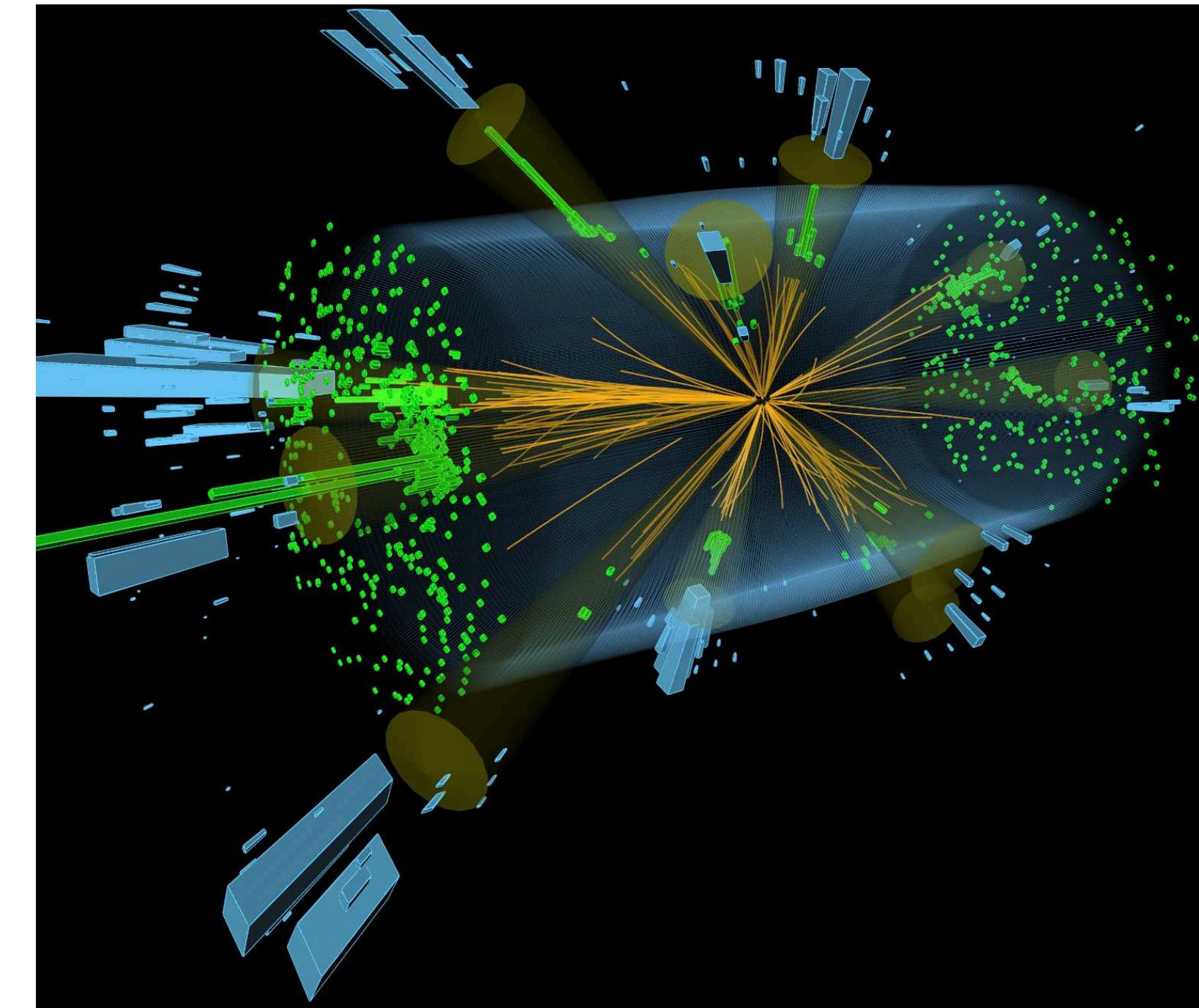
Particle physics events are naturally point clouds (alternatively, images e.g. calorimeters)

Point cloud: "A set of data points in space" –Wikipedia



LIDAR data from self-driving car sensor

An **unordered, variable length** collection of particles

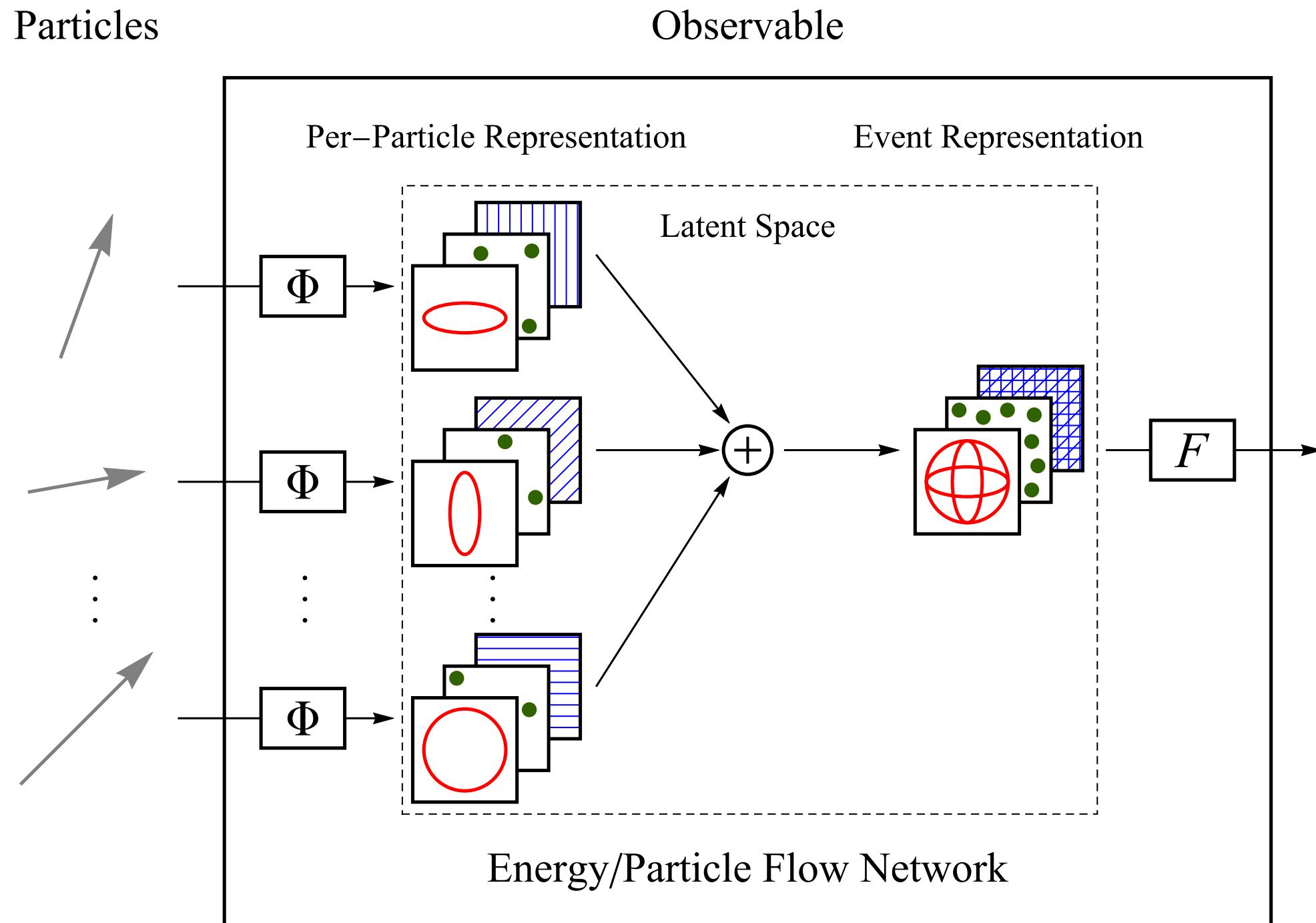


Due to quantum-mechanical indistinguishability

Due to probabilistic nature of event formation

Energy/Particle Flow Networks (EFNs/PFNs)

[Zaheer, Kottur, Ravanbakhsh, Póczos, Salakhutdinov, Smola, [I703.06114](#);
 PTK, Metodiev, Thaler, [I810.05165](#);
[EnergyFlow Python Package](#)]



Particle Flow Network (PFN)

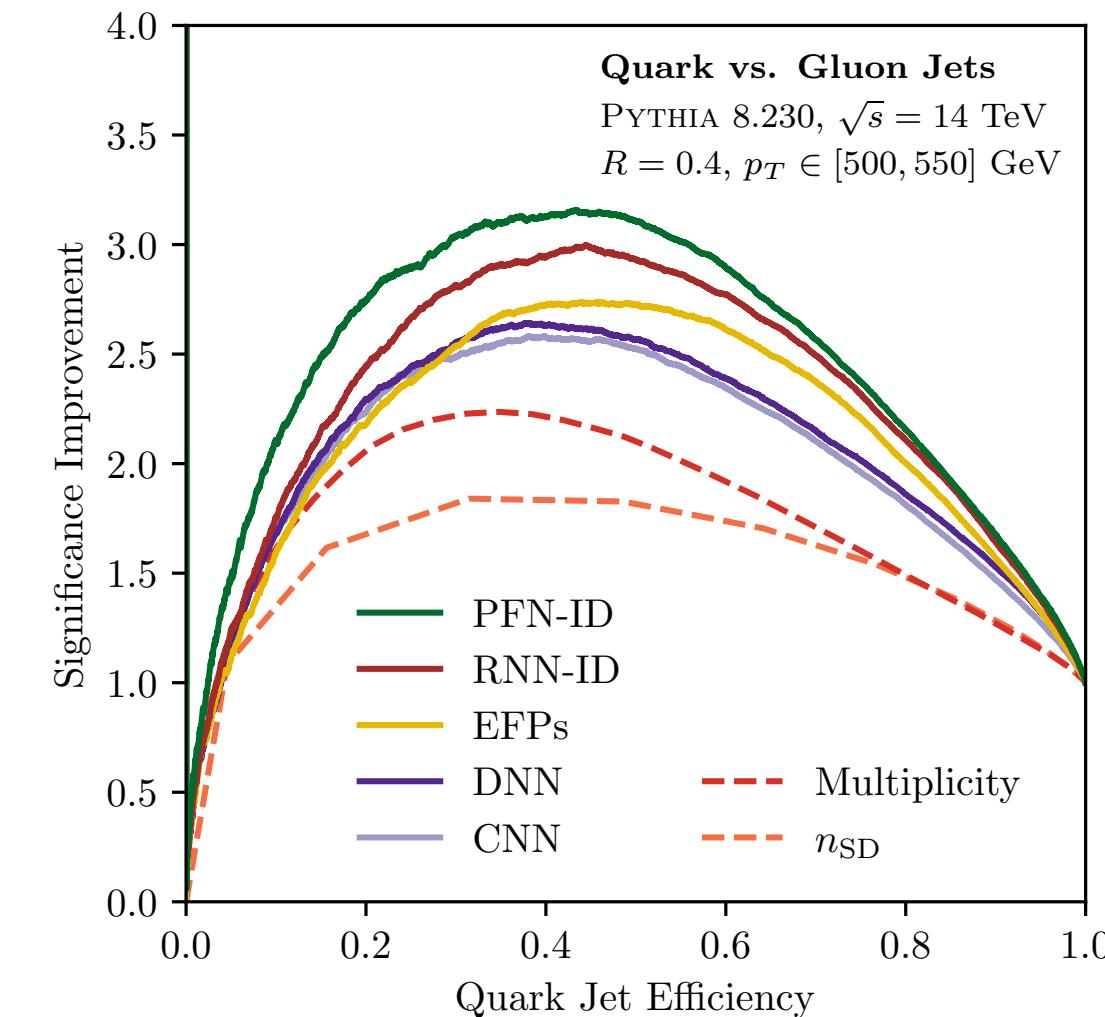
$$\text{PFN}(\{p_1^\mu, \dots, p_M^\mu\}) = F \left(\sum_{i=1}^M \Phi(p_i^\mu) \right)$$

Fully general latent space

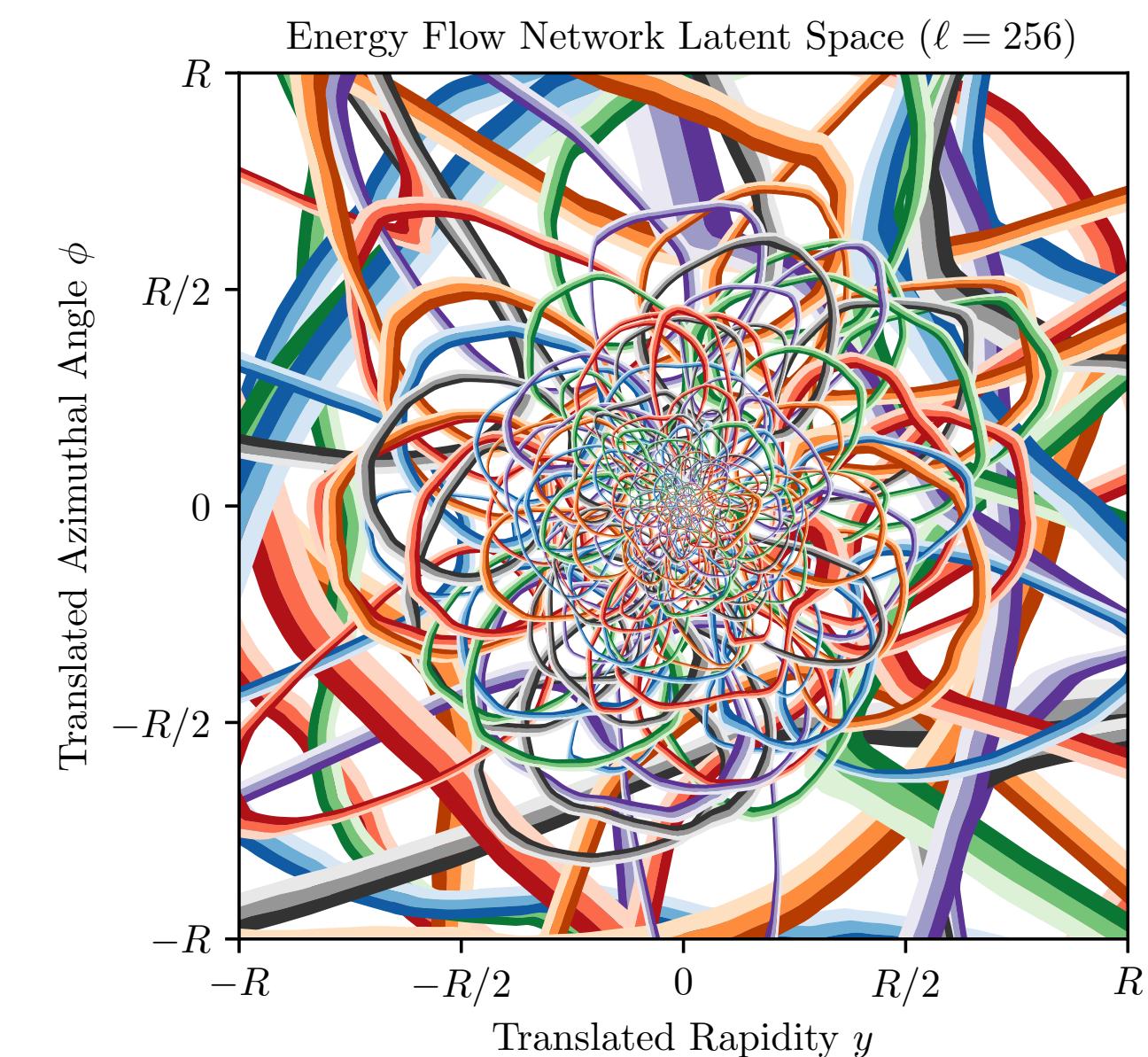
Energy Flow Network (EFN)

$$\text{EFN}(\{p_1^\mu, \dots, p_M^\mu\}) = F \left(\sum_{i=1}^M \textcolor{brown}{z}_i \Phi(\hat{p}_i) \right)$$

IRC-safe latent space



Improved performance (and training)
compared to RNN and CNN



Latent space visualization reveals
what the network has learned

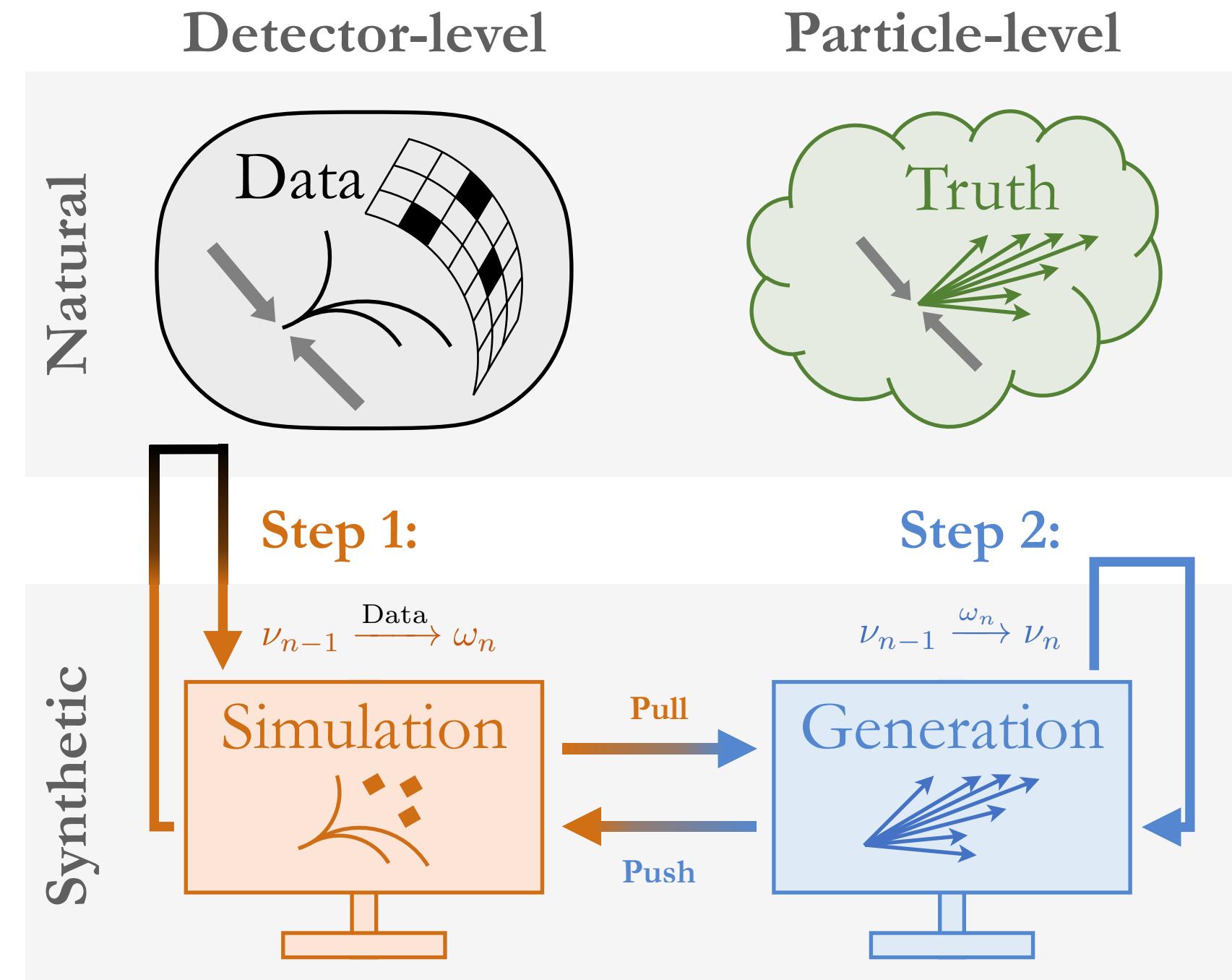
Dynamic pixel sizing related to
collinear singularity of QCD!

OmniFold – Unbinned, Full Phase-Space Unfolding



[Andreassen, PTK, Metodiev, Nachman, Thaler, [1911.09107](#); PTK talk at ML4Jets 2020]

OmniFold weights particle-level **Gen** to be consistent with **Data** once passed through the detector



Step 1 – Reweights **Sim_{n-1}** to data, pulls weights back to particle-level **Gen_{n-1}**

Step 2 – Reweights **Gen_{n-1}** to (step 1)-weighted gen_{n-1}, pushes weights to detector-level **Sim_n**

OmniFold – i.e. continuous IBU

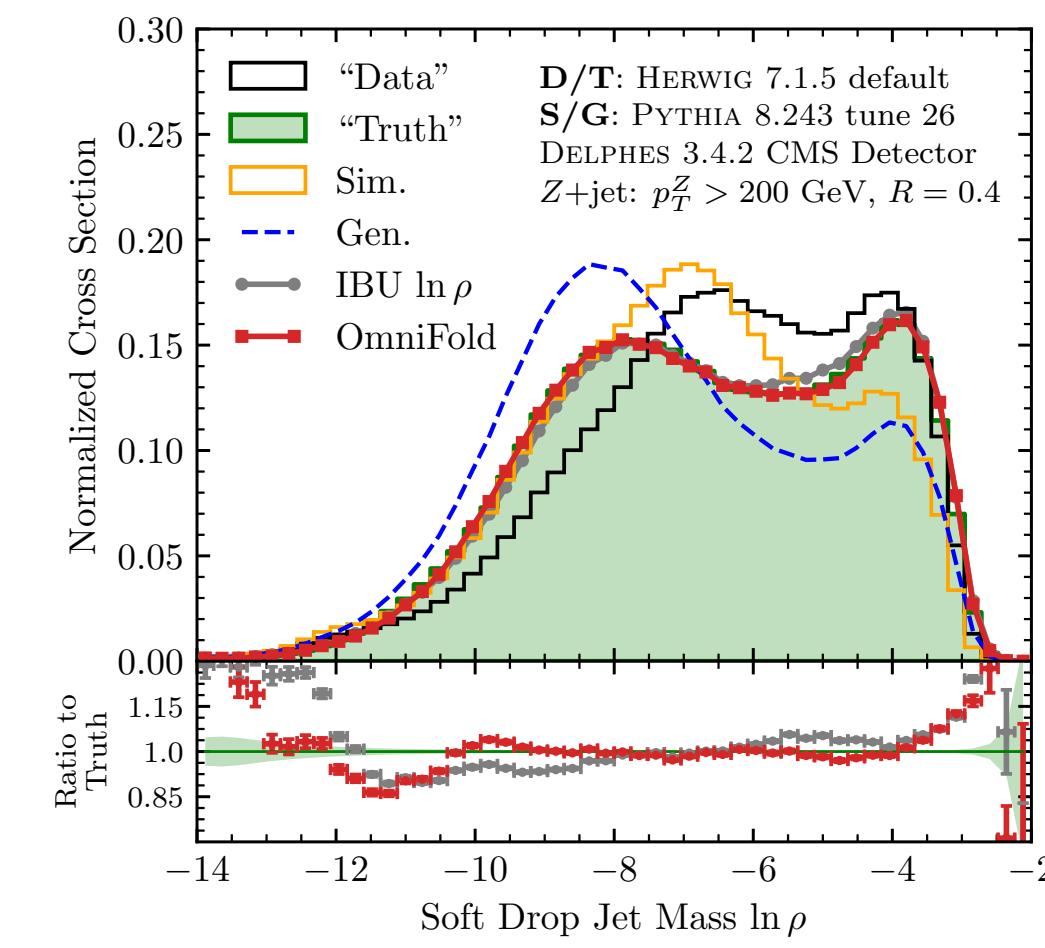
$$\text{Step 1} - \omega_n(m) = \nu_{n-1}^{\text{push}} \times L[(1, \text{Data}), (\nu_{n-1}^{\text{push}}, \text{Sim})](m)$$

$$\text{Step 2} - \nu_n(t) = \nu_{n-1}(t) \times L[(\omega_n^{\text{pull}}, \text{Gen}), (\nu_{n-1}, \text{Gen})](t)$$

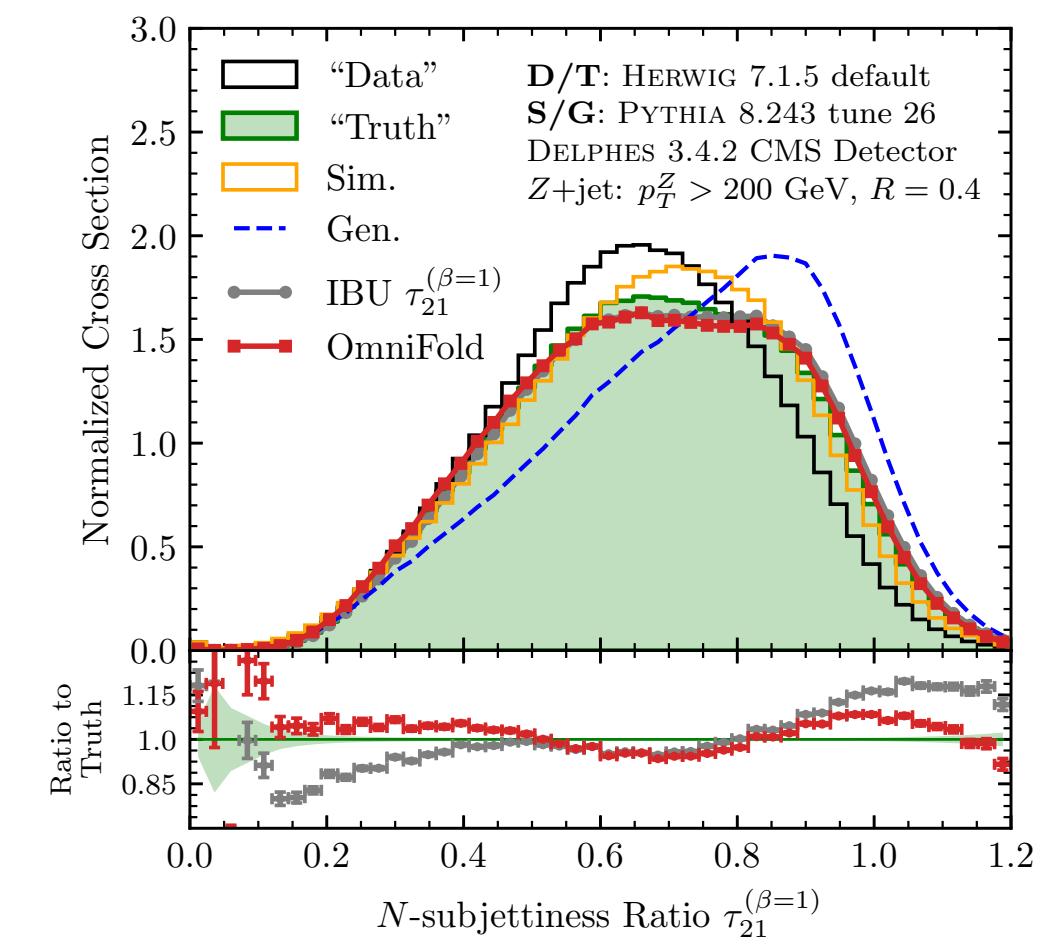
Unfold any* observable $p_{\text{Gen}(t)}$ using universal weights $\nu_n(t)$

$$p_{\text{unfolded}}^{(n)}(t) = \nu_n(t) \times p_{\text{Gen}}(t)$$

*Observables should be chosen responsibly



IRC safe



Sudakov safe

Explicit Geometry – Individual Events in Theory

Hard collision

Good understanding via perturbation theory

Fragmentation

Semi-classical parton shower, effective field theory

Hadronization

Poorly understood (non-perturbative), modeled empirically

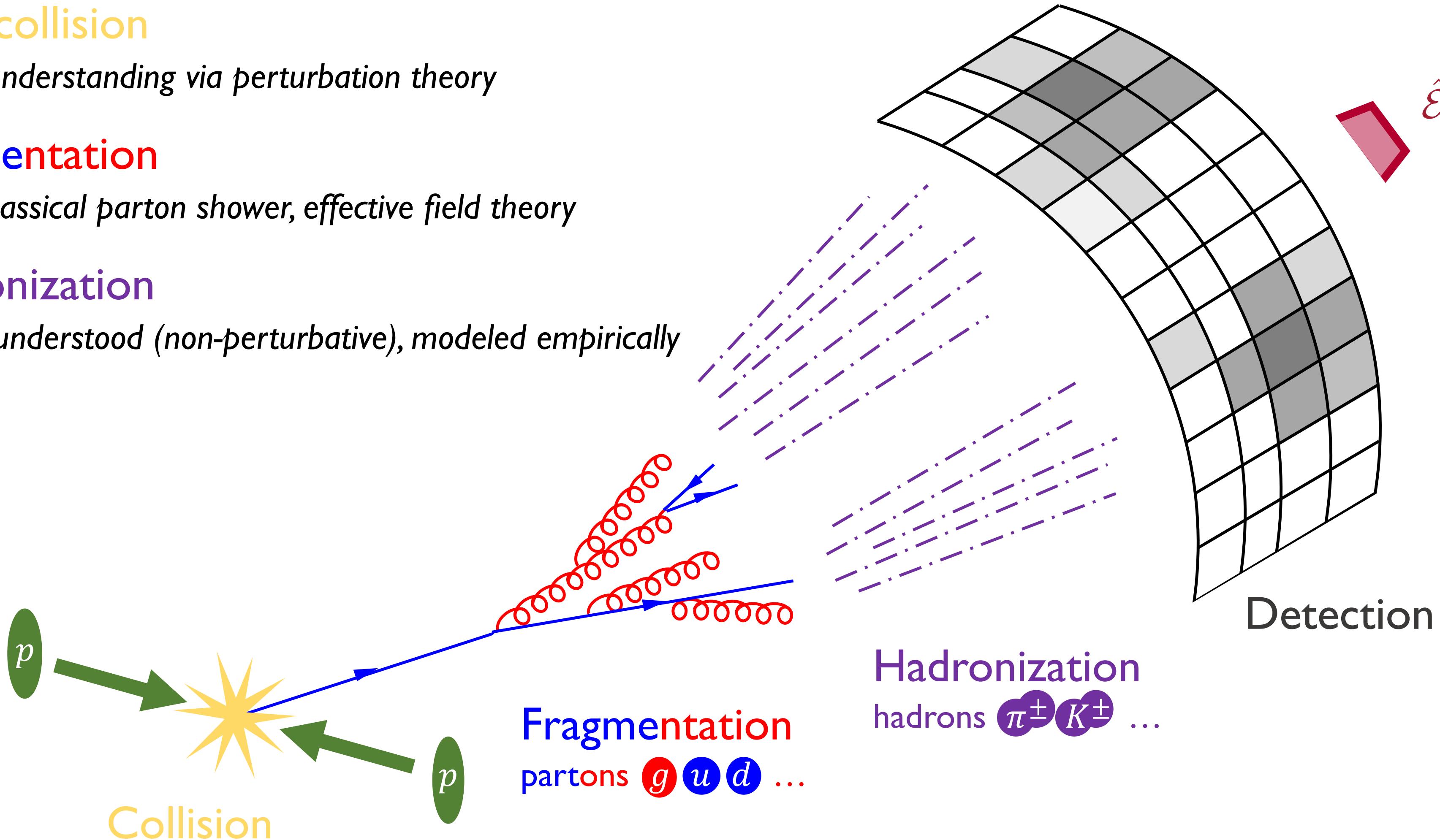


Diagram by Eric Metodiev

[Sveshnikov, Tkachov, [PLB 1996](#); Hofman, Maldacena, [JHEP 2008](#); Mateu, Stewart, Thaler, [PRD 2013](#); Belitsky, Hohenegger, Korchemsky, Sokatchev, Zhiboedov, [PRL 2014](#); Chen, Moult, Zhang, Zhu, [2004.11381](#); Dixon, PTK, Moult, Thaler, Zhu, [to appear soon, see more here](#)]

$$\hat{\mathcal{E}} \simeq \lim_{t \rightarrow \infty} \hat{n}_i T^{0i}(t, vt\hat{n})$$

Stress-energy flow

Robust to non-perturbative and detector effects

Well-defined for massless gauge theories

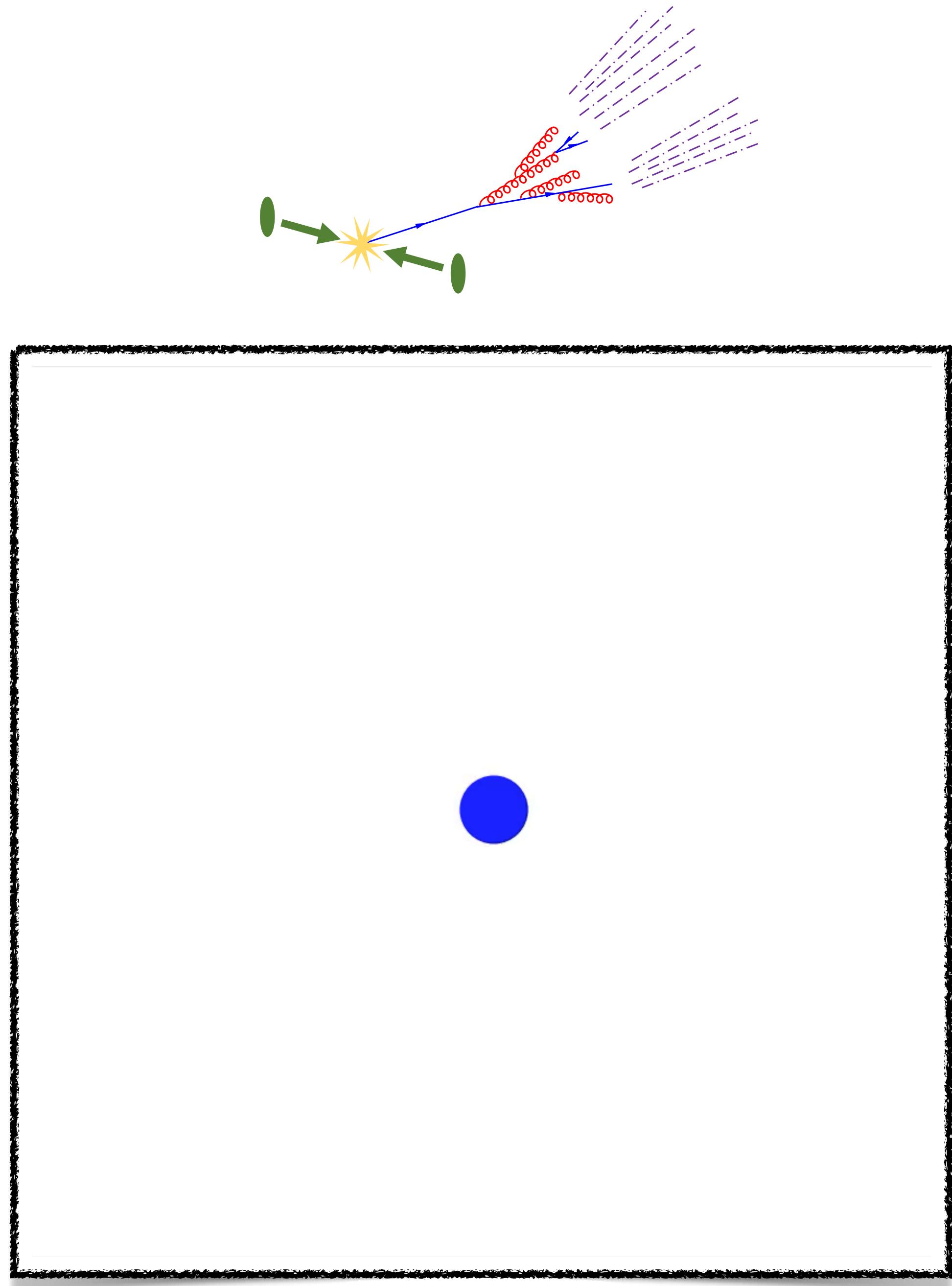
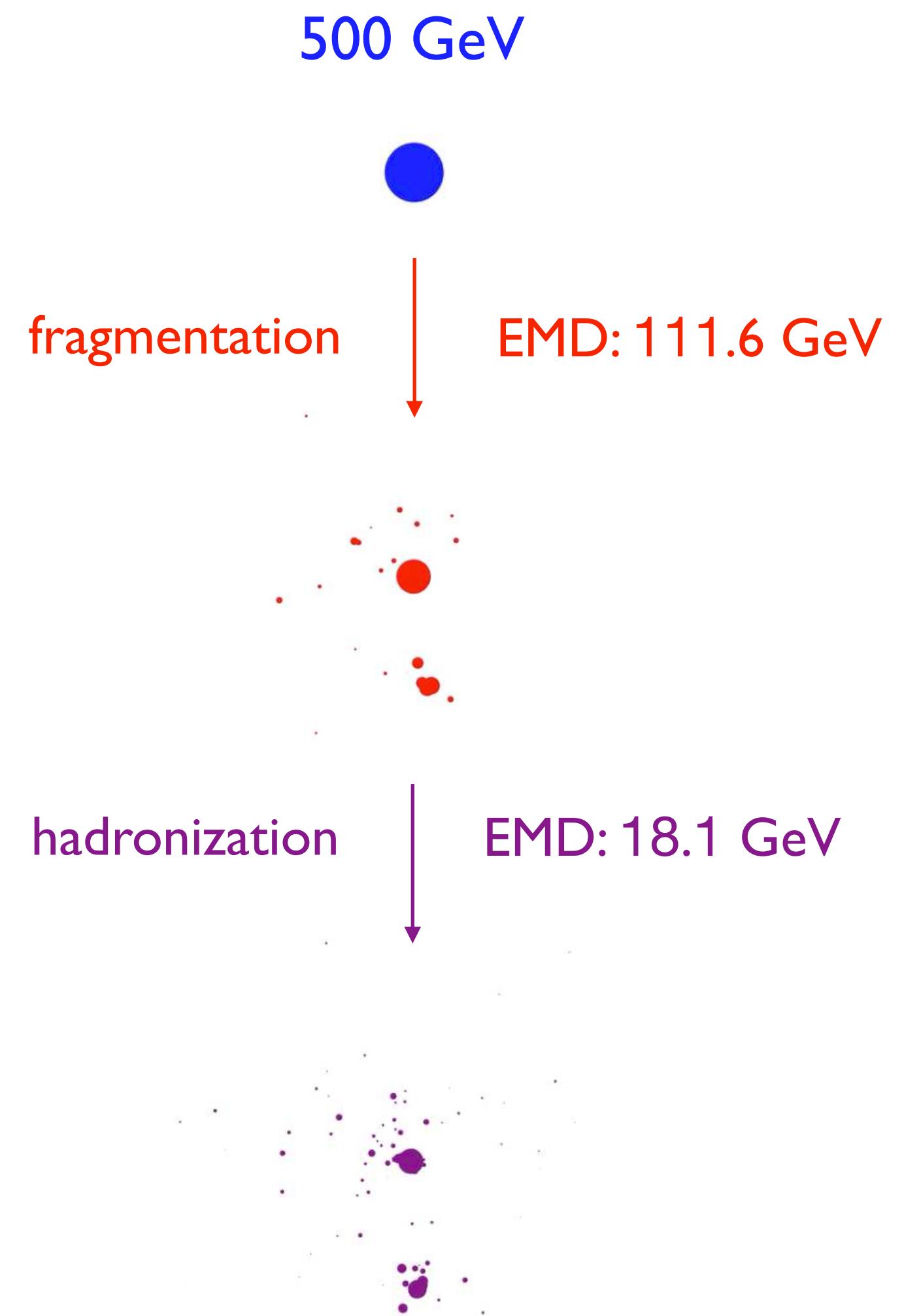
Correlation functions calculated in $N=4$ SYM and QCD

Table of Observables Defined via Event Space Geometry

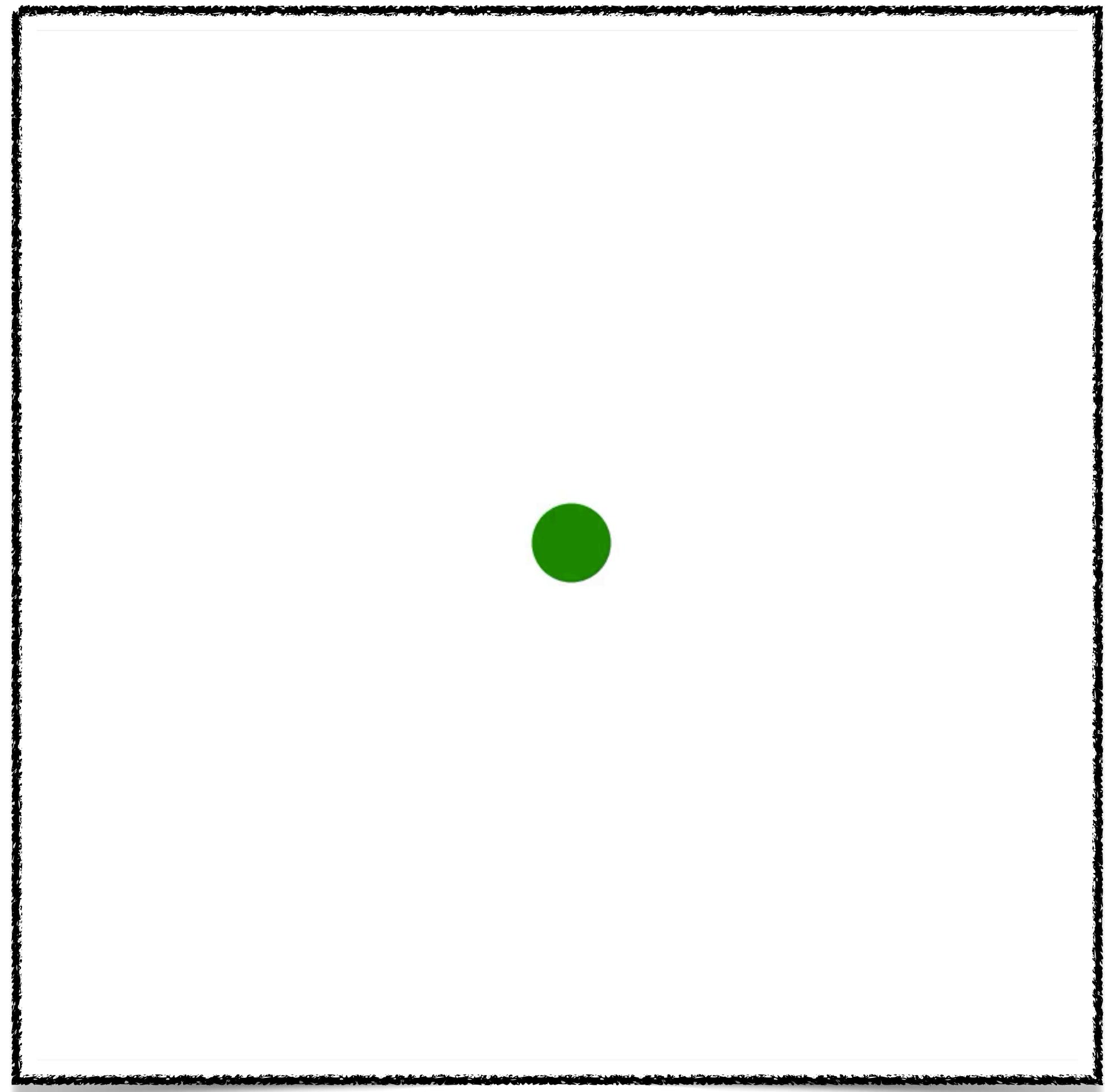
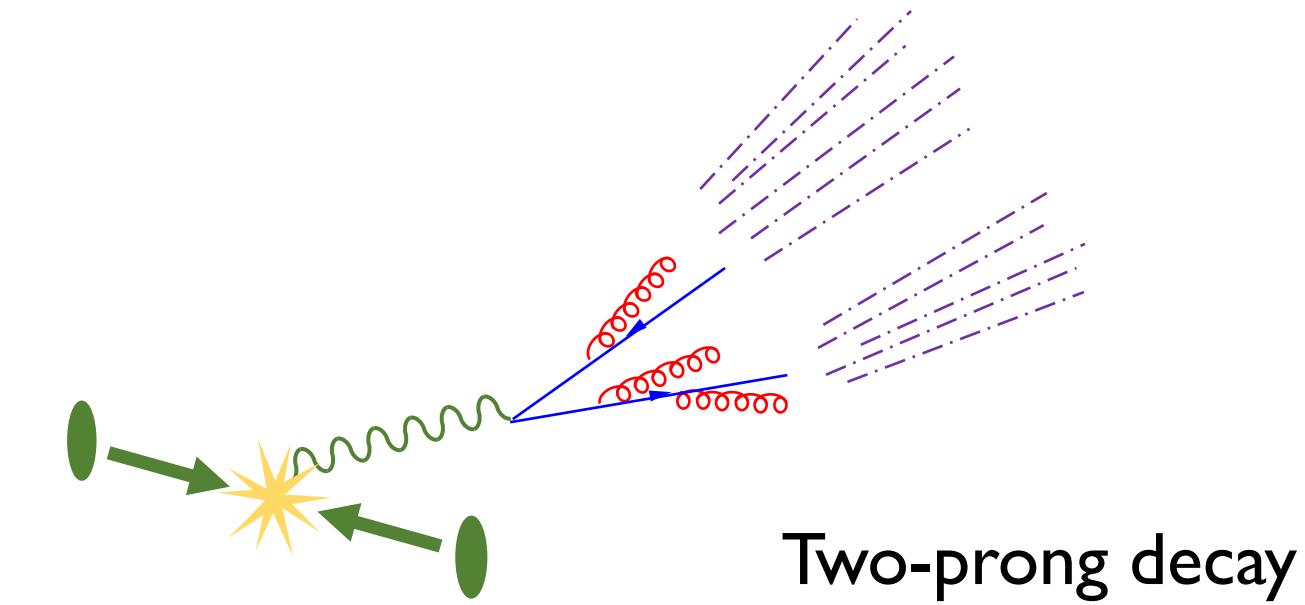
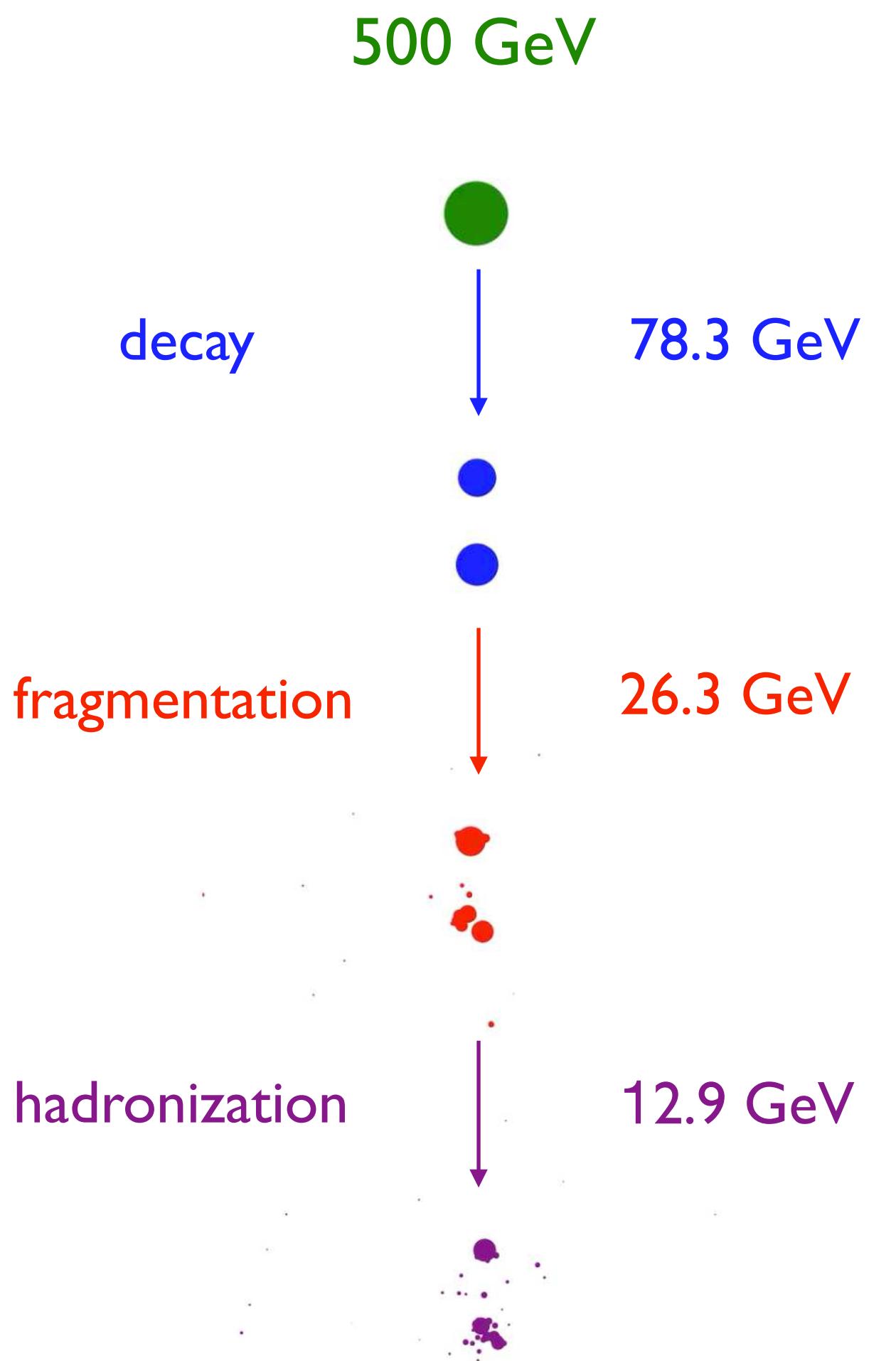
$$\mathcal{O}(\mathcal{E}) = \min_{\mathcal{E}' \in \mathcal{M}} \text{EMD}_{\beta, R}(\mathcal{E}, \mathcal{E}')$$

Name	$\mathcal{O}(\mathcal{E}) = \min_{\mathcal{E}' \in \mathcal{M}} \text{EMD}_{\beta}(\mathcal{E}, \mathcal{E}')$	β	Manifold \mathcal{M}
Thrust	$t(\mathcal{E})$	2	$\mathcal{P}_2^{\text{BB}}$: 2-particle events, back to back
Spherocity	$\sqrt{s(\mathcal{E})}$	1	$\mathcal{P}_2^{\text{BB}}$: 2-particle events, back to back
Broadening	$b(\mathcal{E})$	1	\mathcal{P}_2 : 2-particle events
N -jettiness	$\mathcal{T}_N^{(\beta)}(\mathcal{E})$	β	\mathcal{P}_N : N -particle events
Isotropy	$\mathcal{I}^{(\beta)}(\mathcal{E})$	β	$\mathcal{M}_{\mathcal{U}}$: Uniform events
Jet Angularities	$\lambda_{\beta}(\mathcal{J})$	β	\mathcal{P}_1 : 1-particle jets
N -subjettiness	$\tau_N^{(\beta)}(\mathcal{J})$	β	\mathcal{P}_N : N -particle jets

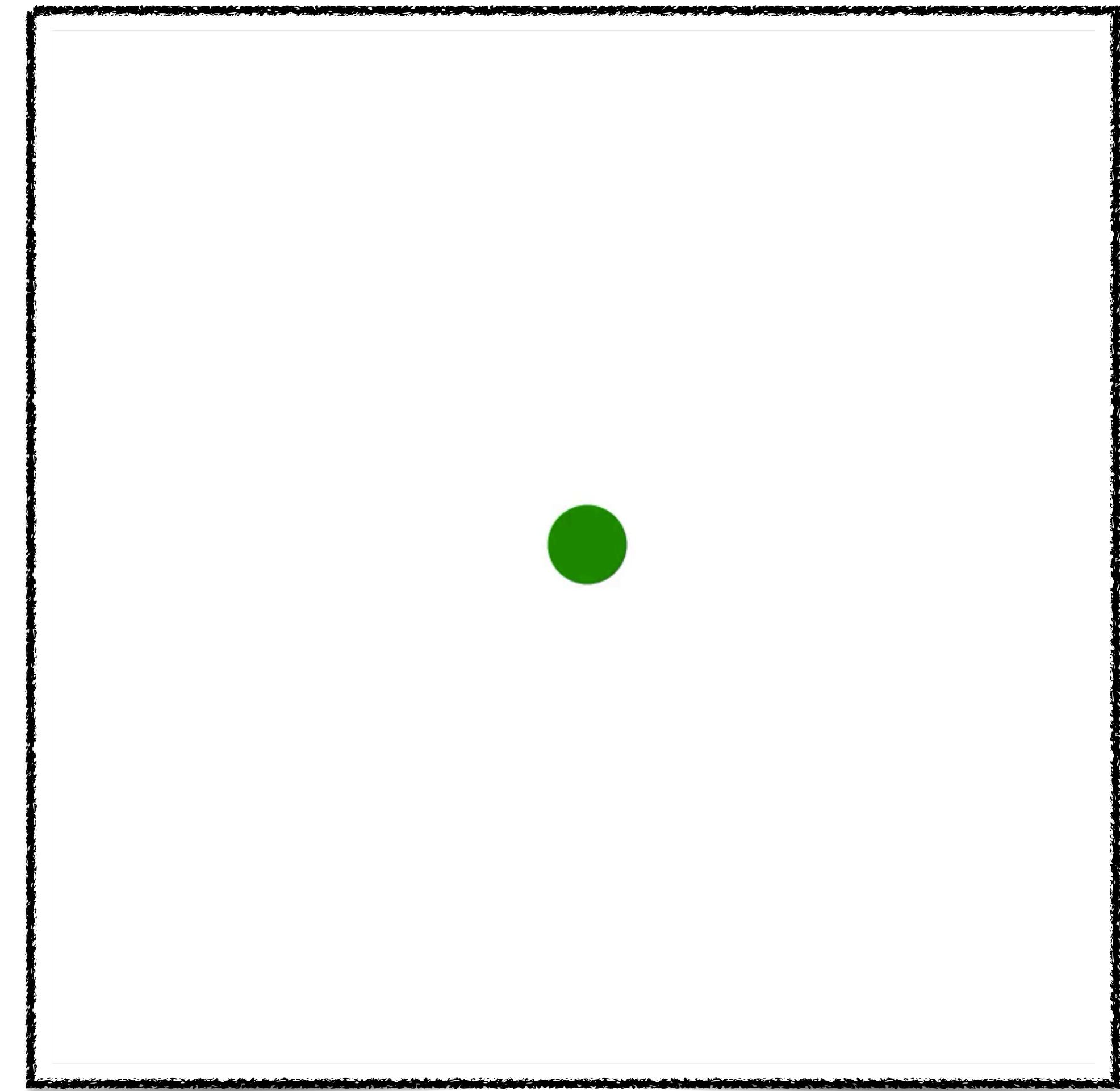
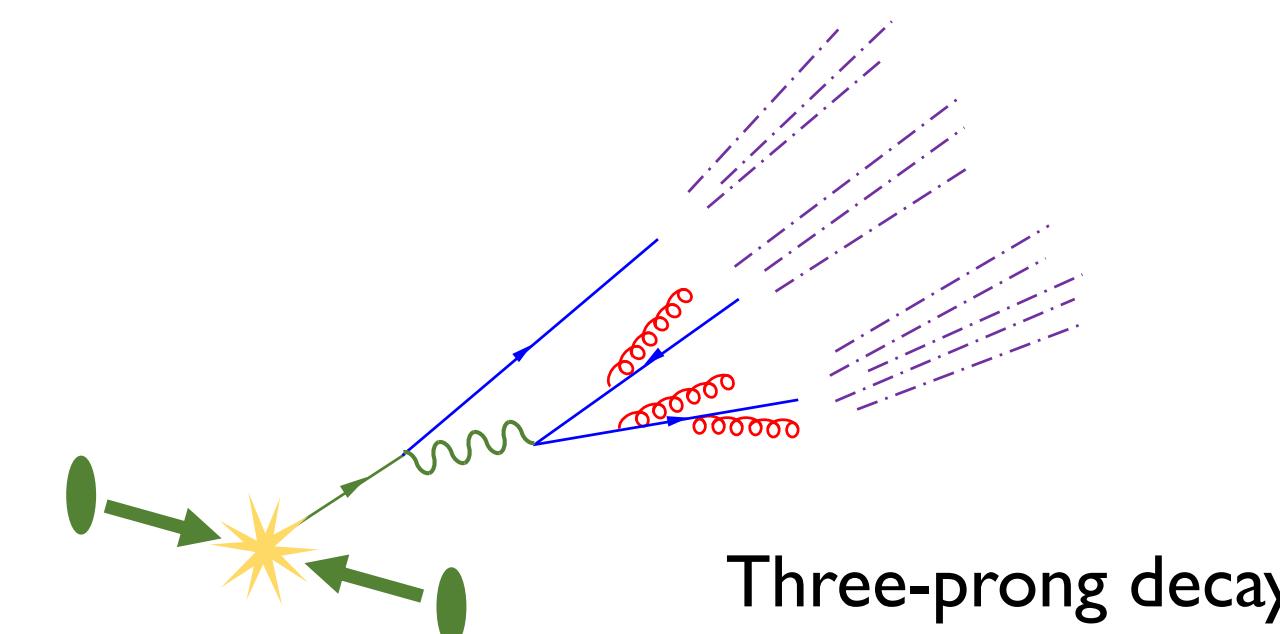
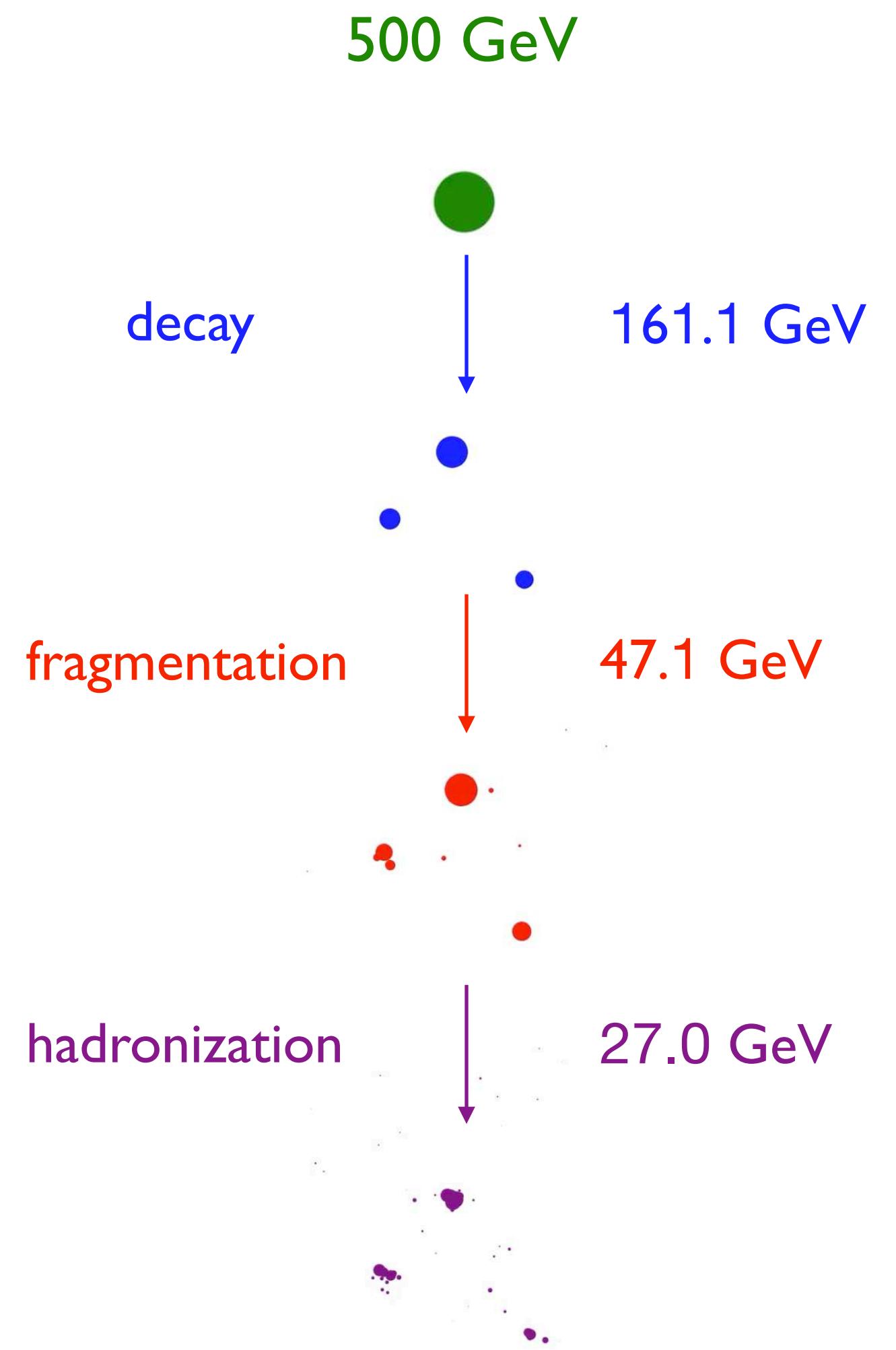
Visualizing Jet Formation – QCD Jets



Visualizing Jet Formation – W Jets



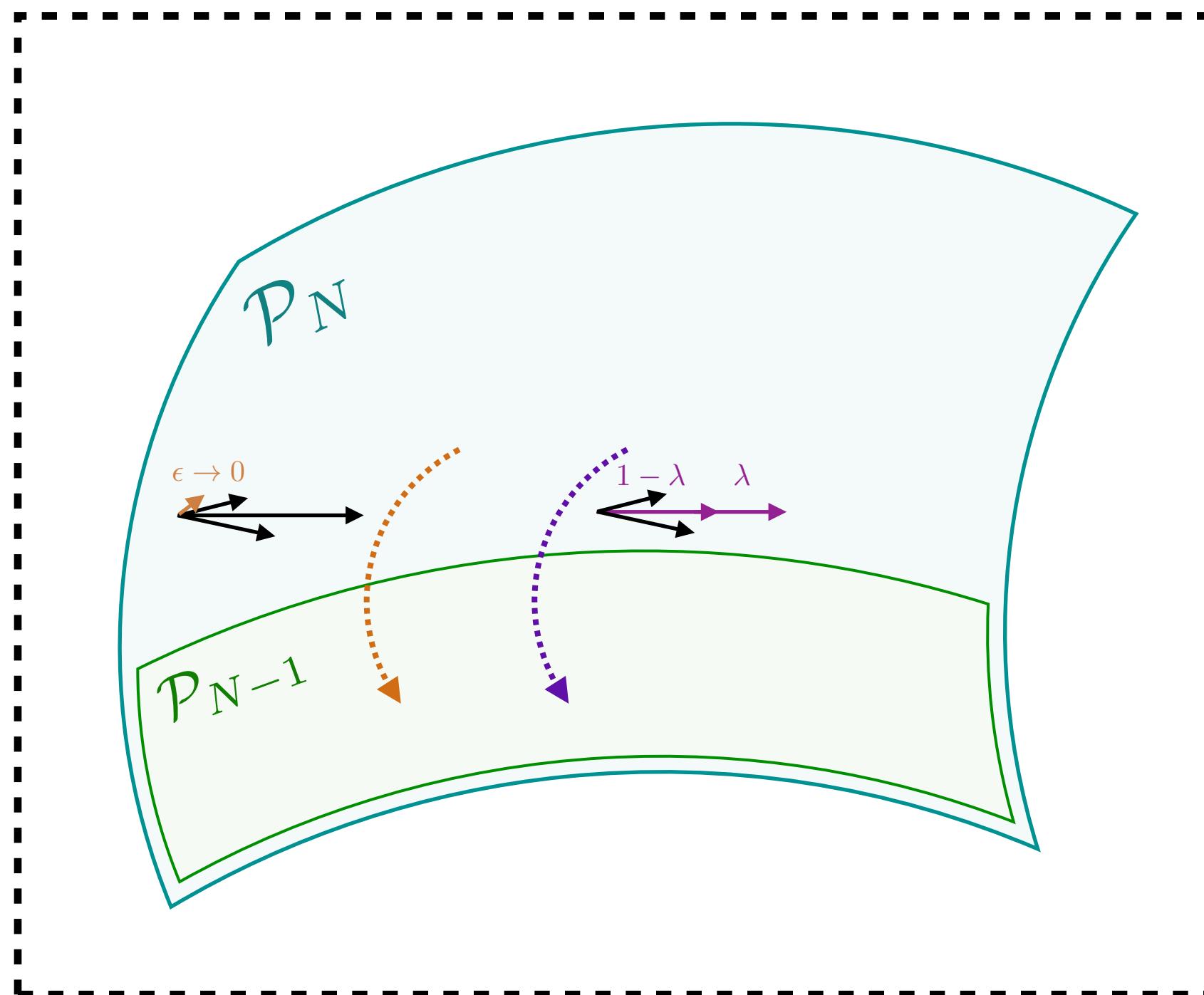
Visualizing Jet Formation – Top Jets



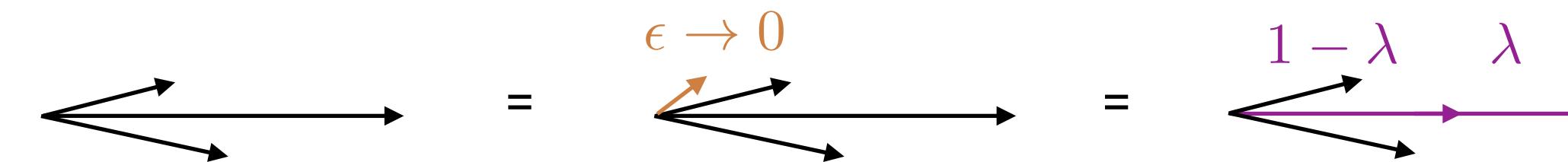
N-particle Manifolds in the Space of Events – Infrared Divergences

[PTK, Metodiev, Thaler, JHEP 2020]

$$\mathcal{P}_N = \text{set of all } N\text{-particle configurations} = \left\{ \sum_{i=1}^N E_i \delta(\hat{n} - \hat{n}_i) \mid E_i \geq 0 \right\}$$



Energy flow is unchanged by exact soft/collinear emissions



Functions of **energy** flow automatically satisfy exact **IRC** invariance!

Real and virtual divergences appear naturally together

$$\mathcal{P}_N \supset \mathcal{P}_{N-1} \supset \cdots \supset \mathcal{P}_3 \supset \mathcal{P}_2 \supset \mathcal{P}_1$$

by **soft** and **collinear** limits

Defining IRC Safety Precisely

[Sterman, Weinberg, [PRL 1997](#); Sterman, [PRD 1978](#); Banfi, Salam, Zanderighi, [JHEP 2005](#)]

Infrared and collinear safety is a proxy for perturbative calculability of an observable

Exact **IRC** invariance

$$\mathcal{O}(p_1^\mu, \dots, p_M^\mu) = \mathcal{O}(0p_0^\mu, p_1^\mu, \dots, p_M^\mu)$$

$$\mathcal{O}(p_1^\mu, \dots, p_M^\mu) = \mathcal{O}(\lambda p_1^\mu, (1 - \lambda)p_1^\mu, \dots, p_M^\mu)$$

Guarantees observable is well-defined on **energy** flows

Allows for pathological observables, e.g. pseudo-multiplicity

Smooth **IRC** invariance

$$\mathcal{O}(p_1^\mu, \dots, p_M^\mu) = \lim_{\epsilon \rightarrow 0} \mathcal{O}(\epsilon p_0^\mu, p_1^\mu, \dots, p_M^\mu)$$

$$\mathcal{O}(p_1^\mu, \dots, p_M^\mu) = \lim_{p_0^\mu \rightarrow p_1^\mu} \mathcal{O}(\lambda p_0^\mu, (1 - \lambda)p_1^\mu, \dots, p_M^\mu)$$

Eliminates common observables with hard boundaries

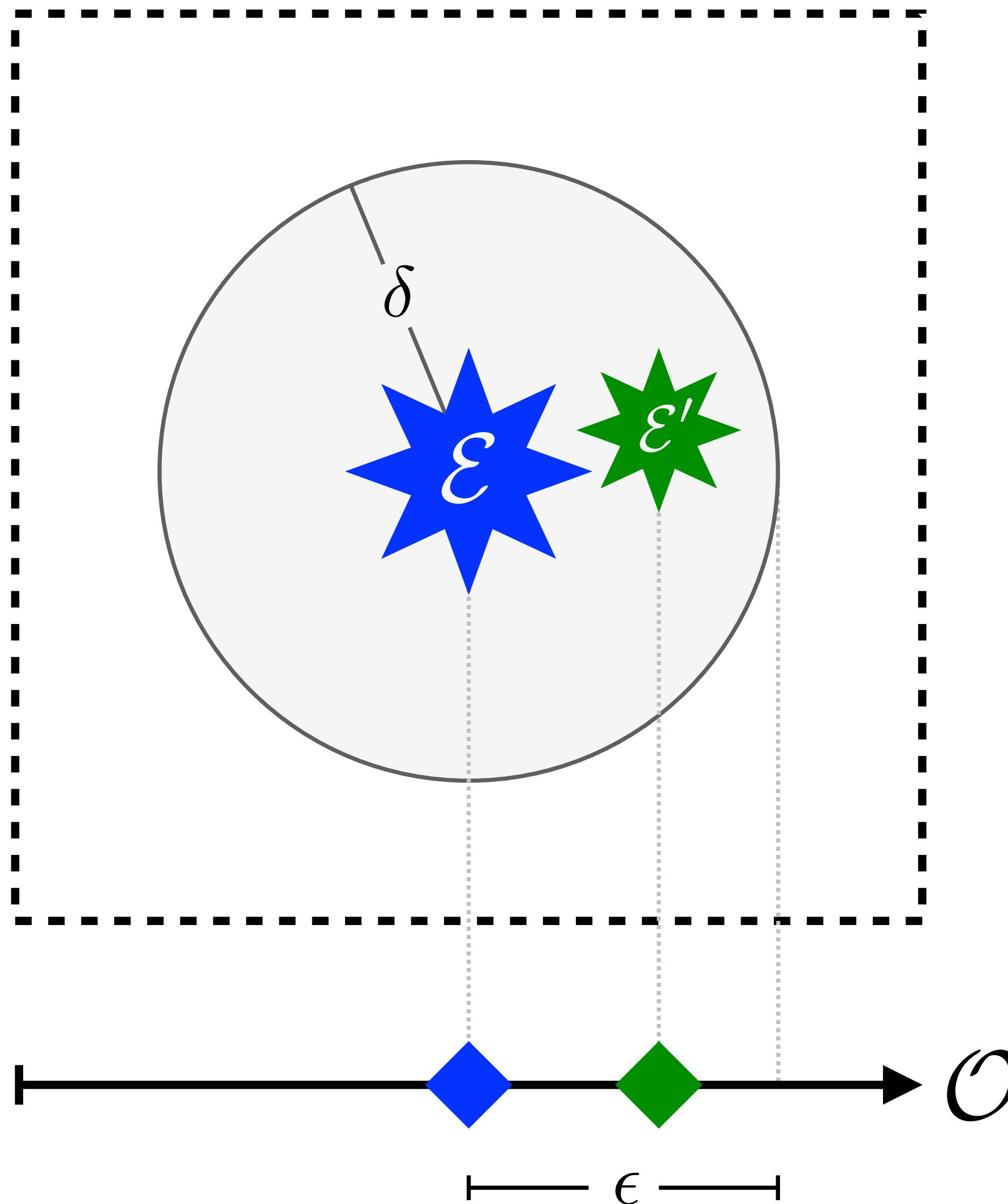
All Observables	Comments
Multiplicity ($\sum_i 1$)	IR unsafe and C unsafe
Momentum Dispersion [65] ($\sum_i E_i^2$)	IR safe but C unsafe
Sphericity Tensor [66] ($\sum_i p_i^\mu p_i^\nu$)	IR safe but C unsafe
Number of Non-Zero Calorimeter Deposits	C safe but IR unsafe

Defined on Energy Flows		
Pseudo-Multiplicity ($\min\{N \mid \mathcal{T}_N = 0\}$)		Robust to exact IR or C emissions

Infrared & Collinear Safe		
Jet Energy ($\sum_i E_i$)		Disc. at jet boundary
Heavy Jet Mass [67]		Disc. at hemisphere boundary
Soft-Dropped Jet Mass [38, 68]		Disc. at grooming threshold
Calorimeter Activity [69] (N_{95})		Disc. at cell boundary

More EMD Geometry – Continuity in the Space of Events

[PTK, Metodiev, Thaler, 2004.04.159]



Classic $\epsilon - \delta$ definition of continuity in a metric space

An observable \mathcal{O} is **EMD continuous** at an event \mathcal{E} if, for any $\epsilon > 0$, there exists a $\delta > 0$ such that for all events \mathcal{E}' :

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

Towards a geometric definition of **IRC Safety**

IRC Safety = EMD Continuity*

*on all but a negligible set[‡] of events

[‡]a negligible set is one that contains no positive-radius EMD-ball

⋮

Perturbation Theory in the Space of Events

[PTK, Metodiev, Thaler, 2004.04.159]

Sudakov safety

[Larkoski, Thaler, JHEP 2014; Larkoski, Marzani, Thaler, PRD 2015]

Some observables have discontinuities on P_N for some N

A resummed **IRC-safe companion** can mitigate the divergences

$$p(\mathcal{O}_{\text{Sudakov}}) = \int d\mathcal{O}_{\text{Comp.}} p(\mathcal{O}_{\text{Sudakov}} | \mathcal{O}_{\text{Comp.}}) p(\mathcal{O}_{\text{Comp.}})$$

Event geometry suggests N -(sub)jettiness as universal companion

Fixed-order calculability

[Sterman, PRD 1979; Banfi, Salam, Zanderighi, JHEP 2005]

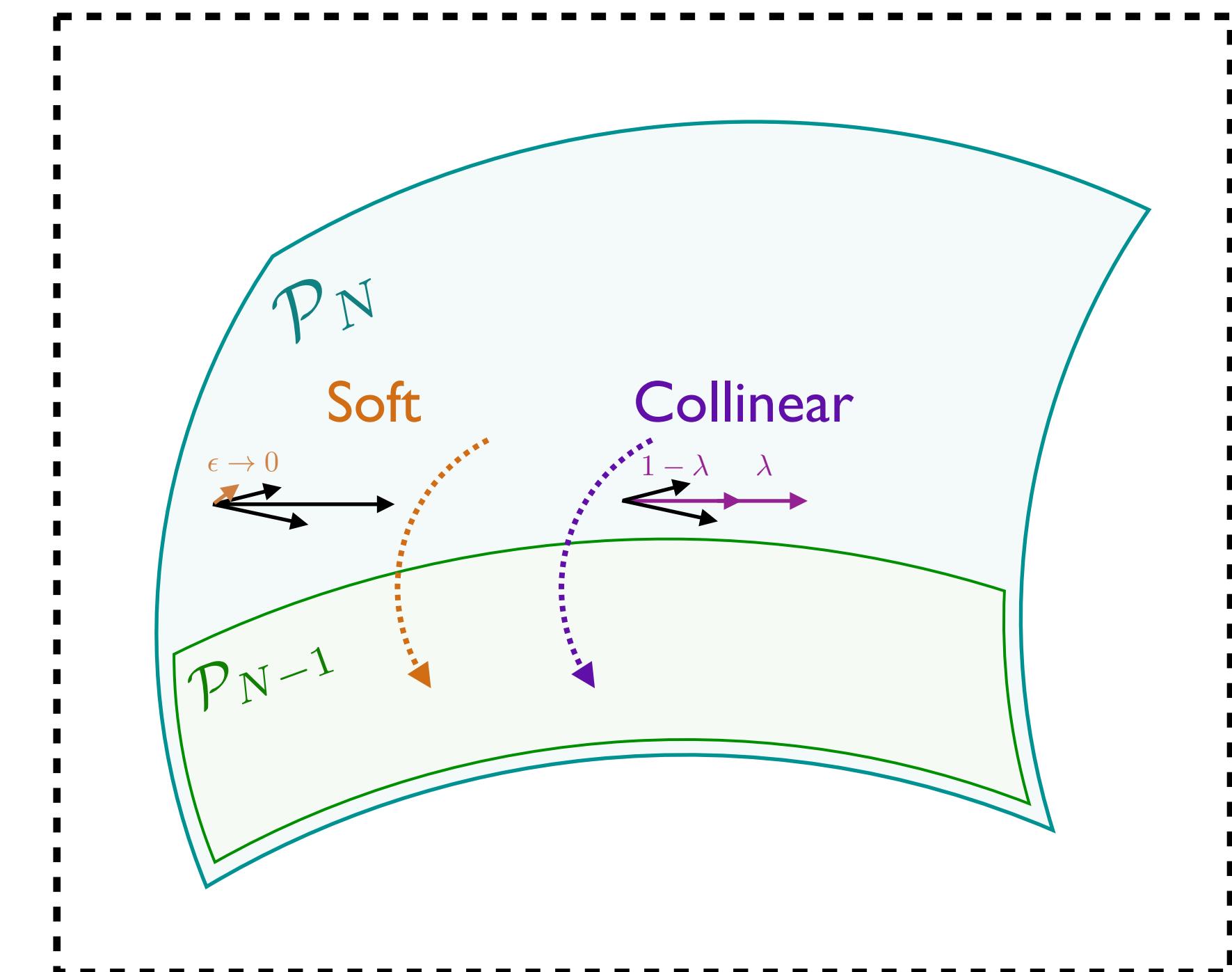
Is a statement of integrability on each P_N

EMD continuity must be upgraded to EMD-Hölder continuity on each P_N

$$\lim_{\mathcal{E} \rightarrow \mathcal{E}'} \frac{\mathcal{O}(\mathcal{E}) - \mathcal{O}(\mathcal{E}')}{\text{EMD}(\mathcal{E}, \mathcal{E}')^c} = 0, \quad c > 0$$

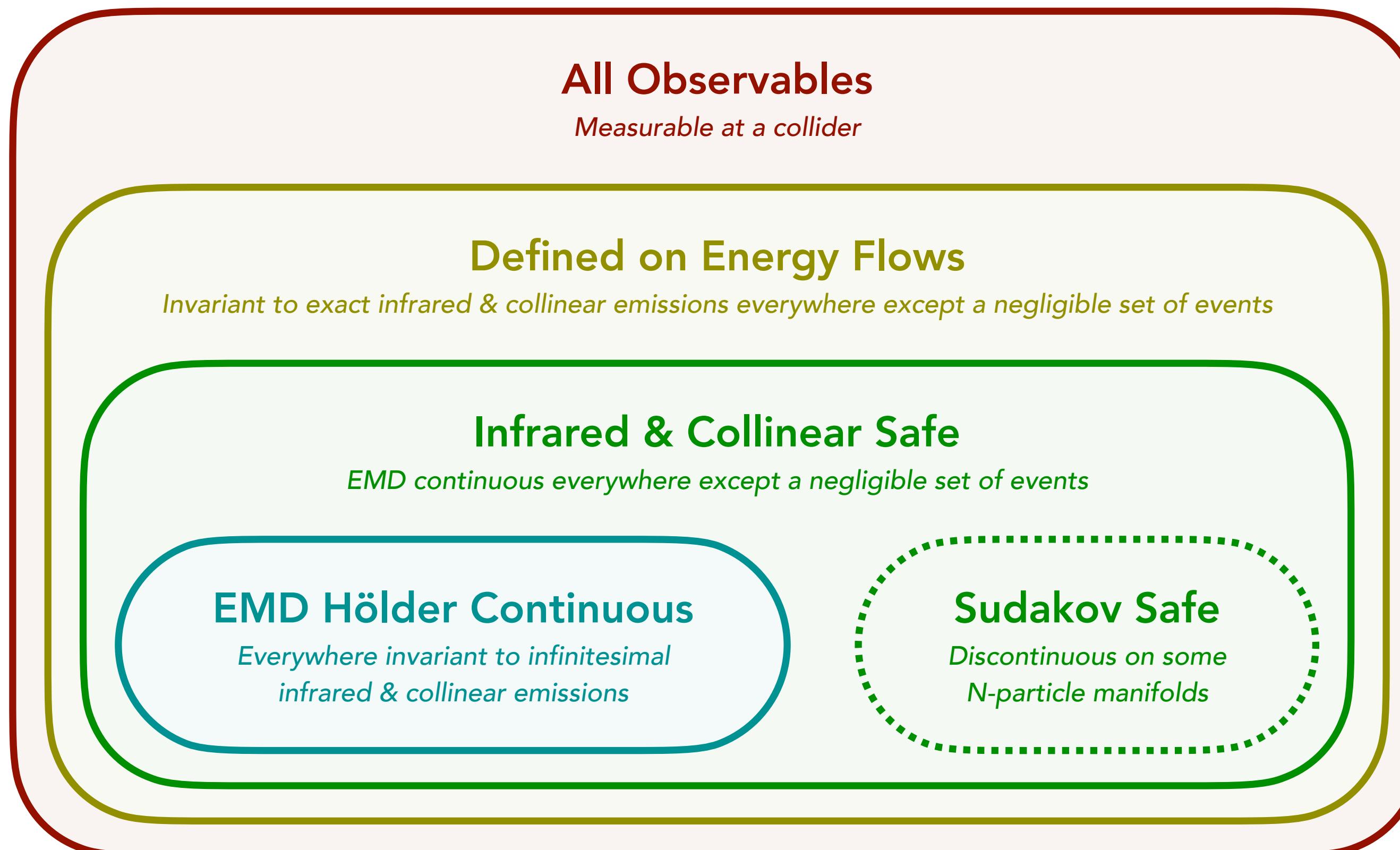
Example: $V(\mathcal{E}) = \mathcal{T}_2(\mathcal{E}) \left(1 + \frac{1}{\ln E(\mathcal{E})/\mathcal{T}_3(\mathcal{E})} \right)$ is EMD continuous but not EMD Hölder continuous (it is Sudakov safe)

Infrared singularities of massless gauge theories appear on each P_N



Hierarchy of IRC Safety Definitions

[PTK, Metodiev, Thaler, 2004.04.159]

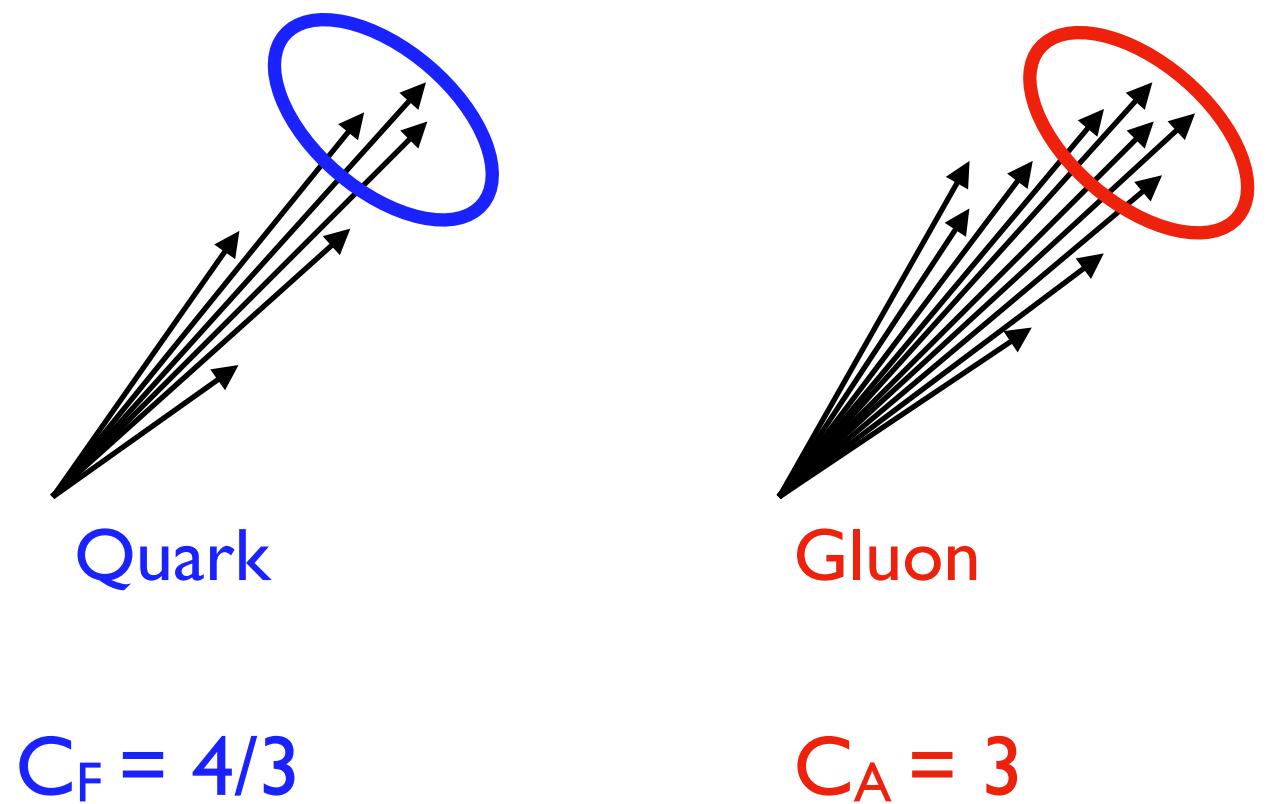


All Observables	Comments
Multiplicity ($\sum_i 1$)	IR unsafe and C unsafe
Momentum Dispersion [65] ($\sum_i E_i^2$)	IR safe but C unsafe
Sphericity Tensor [66] ($\sum_i p_i^\mu p_i^\nu$)	IR safe but C unsafe
Number of Non-Zero Calorimeter Deposits	C safe but IR unsafe
Defined on Energy Flows	
Pseudo-Multiplicity ($\min\{N \mid \mathcal{T}_N = 0\}$)	Robust to exact IR or C emissions
Infrared & Collinear Safe	
Jet Energy ($\sum_i E_i$)	Disc. at jet boundary
Heavy Jet Mass [67]	Disc. at hemisphere boundary
Soft-Dropped Jet Mass [38, 68]	Disc. at grooming threshold
Calorimeter Activity [69] (N_{95})	Disc. at cell boundary
Sudakov Safe	
Groomed Momentum Fraction [39] (z_g)	Disc. on 1-particle manifold
Jet Angularity Ratios [37]	Disc. on 1-particle manifold
N -subjettiness Ratios [47, 48] (τ_{N+1}/τ_N)	Disc. on N -particle manifold
V parameter [36] (Eq. (2.11))	Hölder disc. on 3-particle manifold
EMD Hölder Continuous Everywhere	
Thrust [40, 41]	
Sphericity [42]	
Angularities [70]	
N -jettiness [44] (\mathcal{T}_N)	
C parameter [71–74]	Resummation beneficial at $C = \frac{3}{4}$
Linear Sphericity [72] ($\sum_i E_i n_i^\mu n_i^\nu$)	
Energy Correlators [36, 75–77]	
Energy Flow Polynomials [15, 17]	

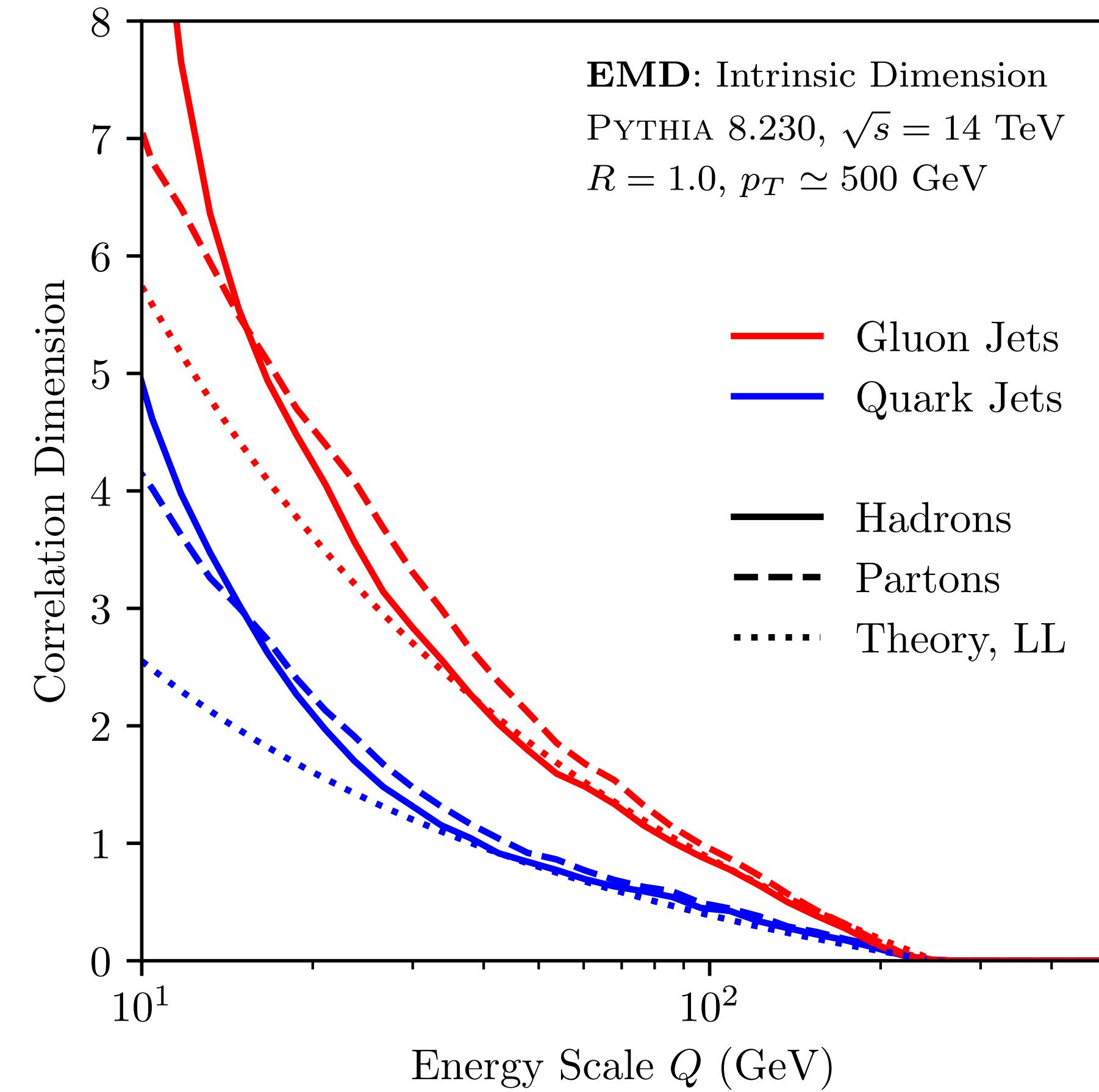
Quark and Gluon Correlation Dimensions

Leading log (single emission) calculation:

$$\dim_i(Q) \simeq -\frac{8\alpha_s}{\pi} C_i \ln \frac{Q}{p_T/2}$$



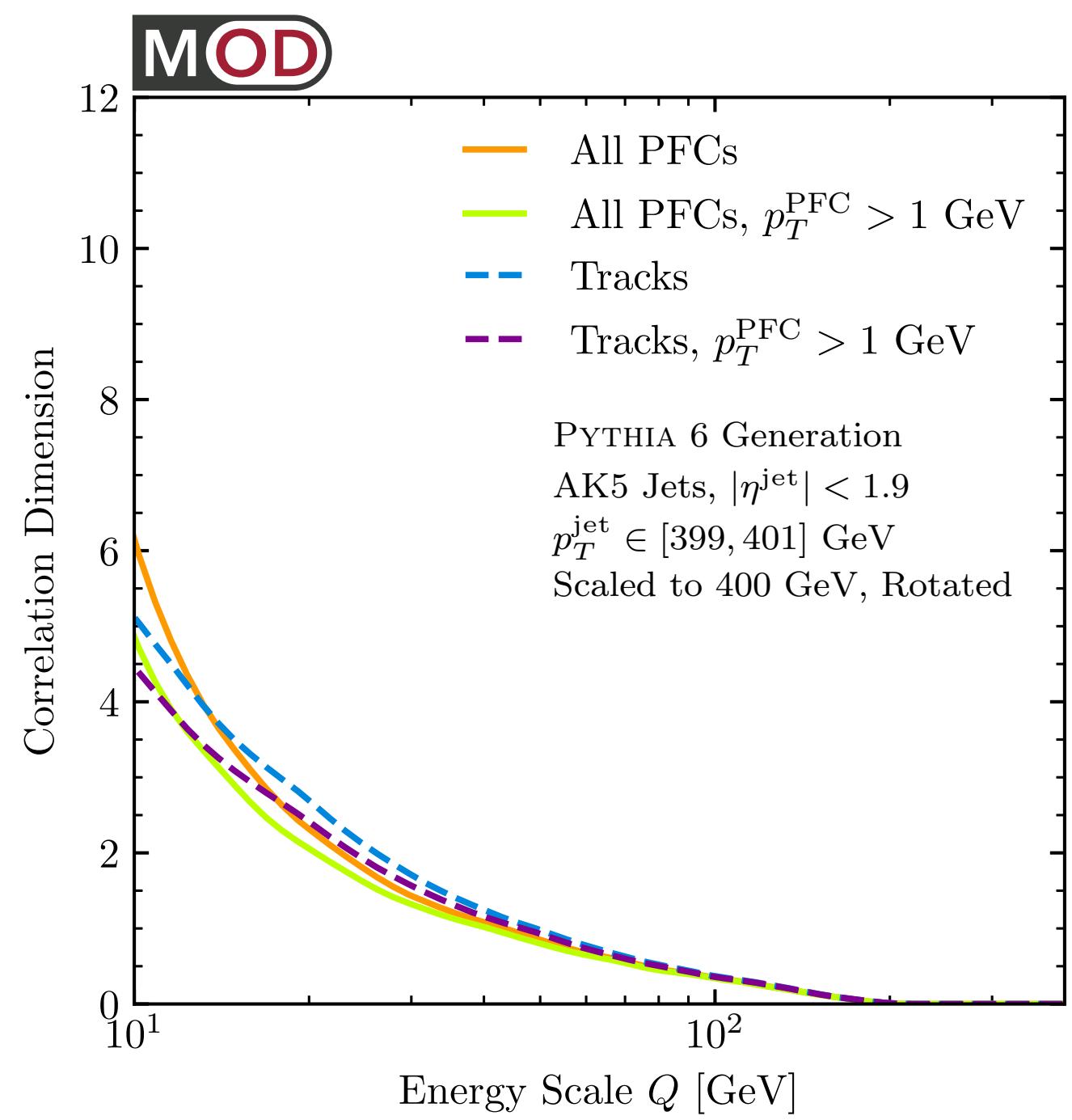
$$\dim(Q) = Q \frac{\partial}{\partial Q} \ln \sum_i \sum_j \Theta(\text{EMD}(\mathcal{E}_i, \mathcal{E}'_j) < Q)$$



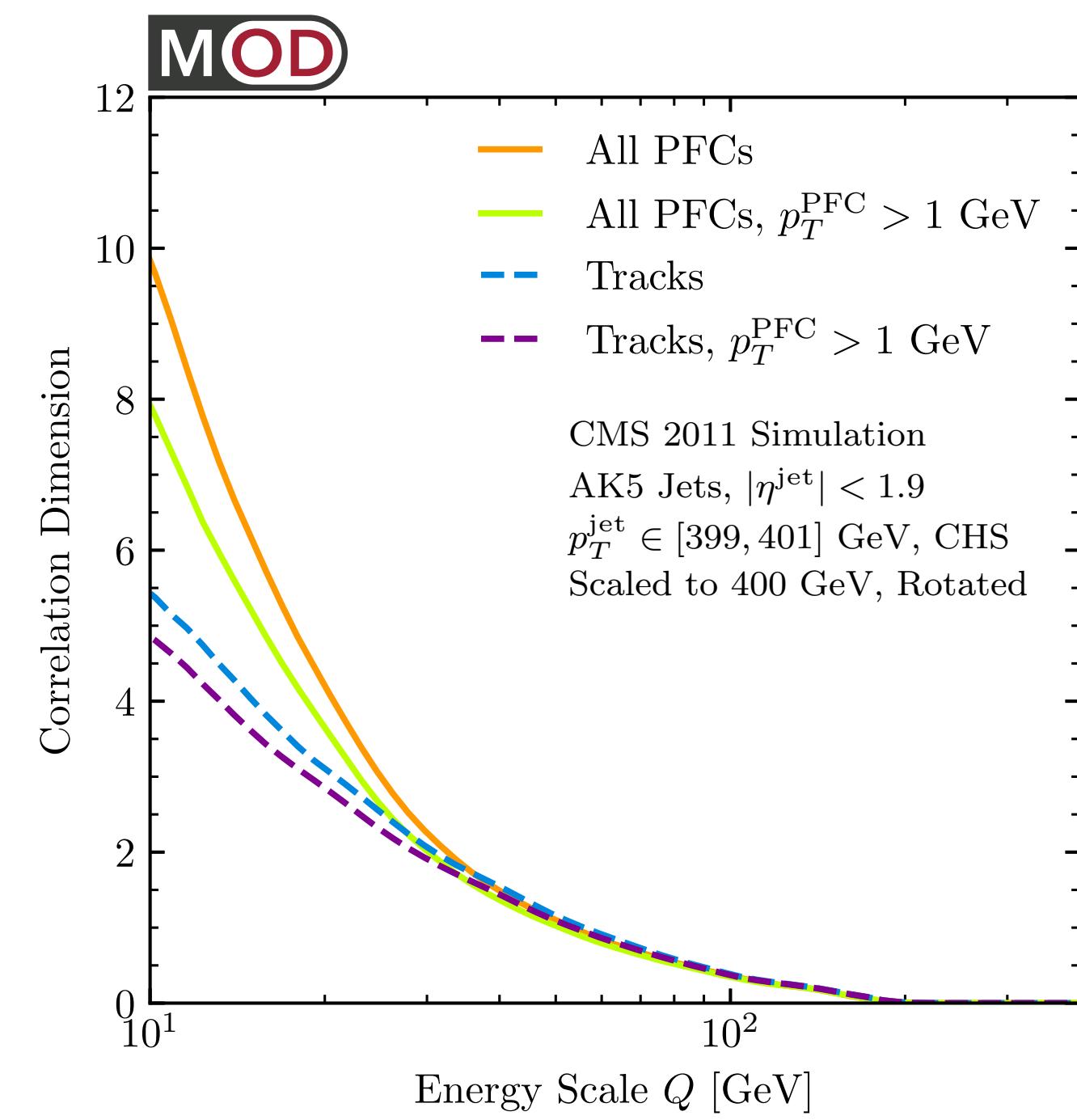
[PTK, Metodiev, Thaler, to appear soon]

Correlation Dimension at Particle and Detector Levels

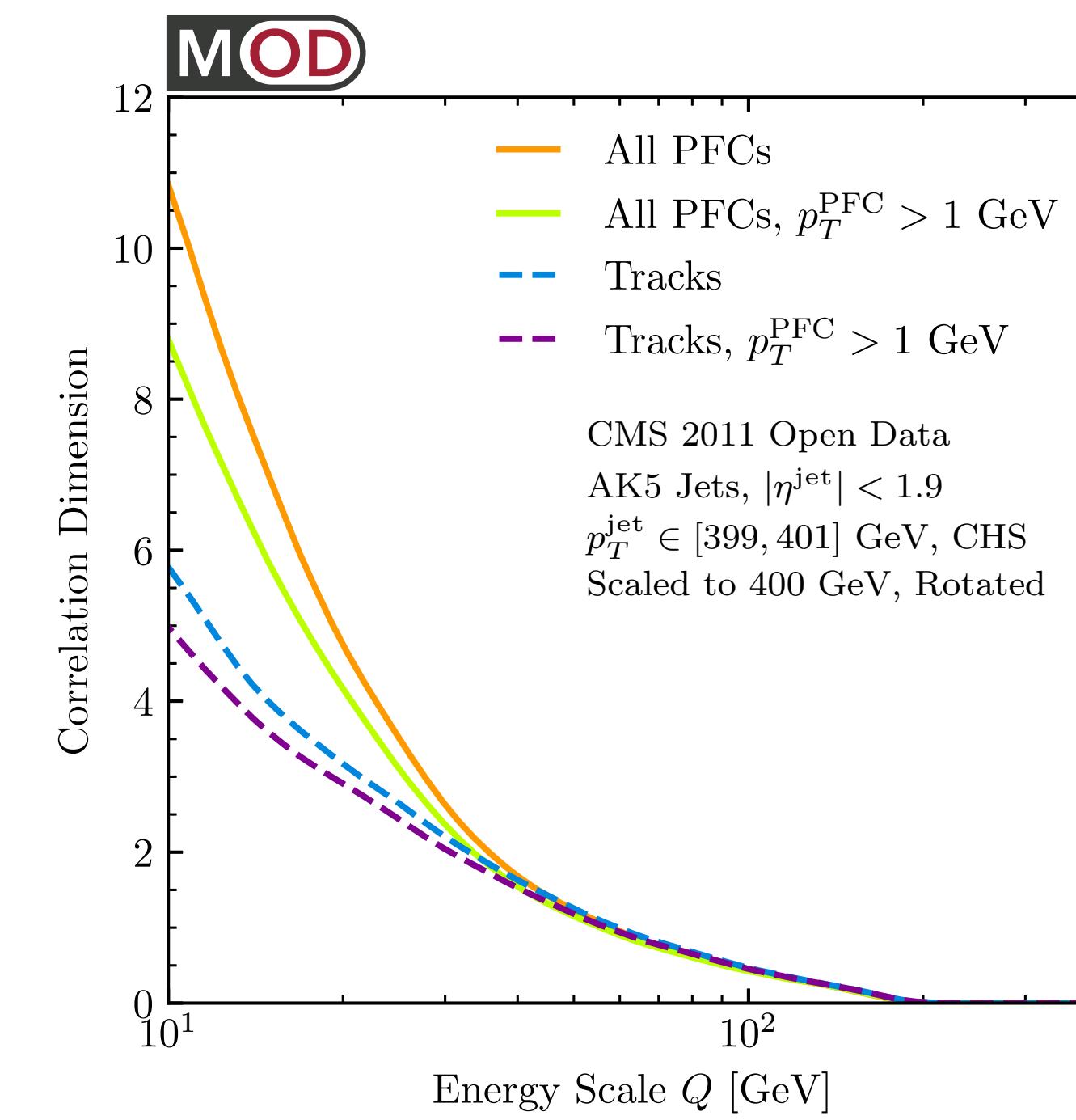
Particle-level (PYTHIA)



Detector-level (PYTHIA + GEANT 4)

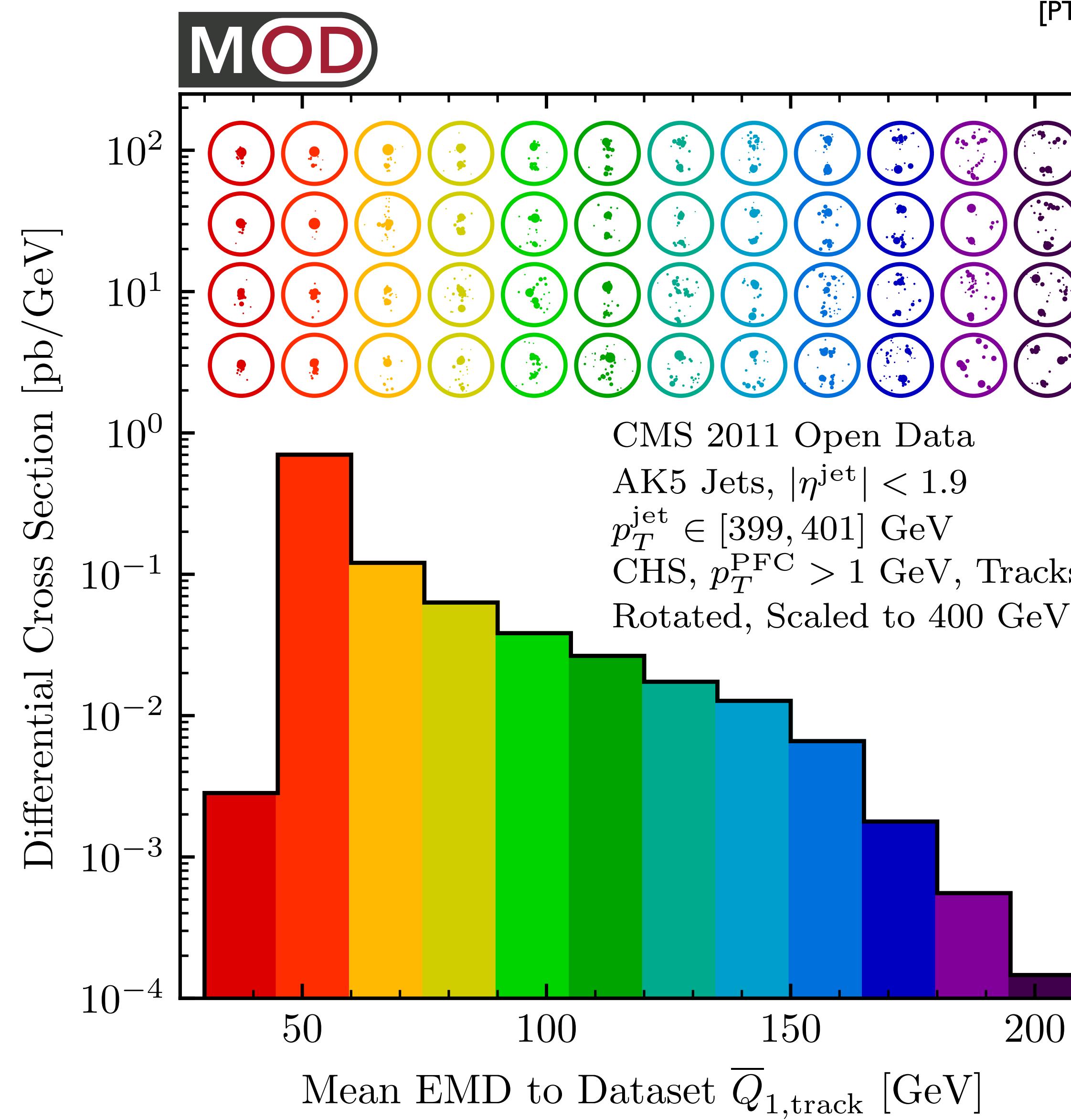


CMS Open Data



Visualizing Geometry in CMS Open Data

[PTK, Mastandrea, Metodiev, Naik, Thaler, PRD 2019; code and datasets at energyflow.network]



EMD for anomaly detection

4 medoids in each bin of anomaliness \bar{Q}_1

n^{th} moment of EMD distribution for a dataset

$$\bar{Q}_n(\mathcal{I}) = \sqrt[n]{\frac{1}{N} \sum_{k=1}^N (\text{EMD}(\mathcal{I}, \mathcal{J}_k))^n}$$

How far does this go?

$$\mathcal{V}_k = \frac{1}{N} \sum_{i=1}^N \min \{ \text{EMD}(\mathcal{J}_i, \mathcal{K}_1), \dots, \text{EMD}(\mathcal{J}_i, \mathcal{K}_k) \}$$

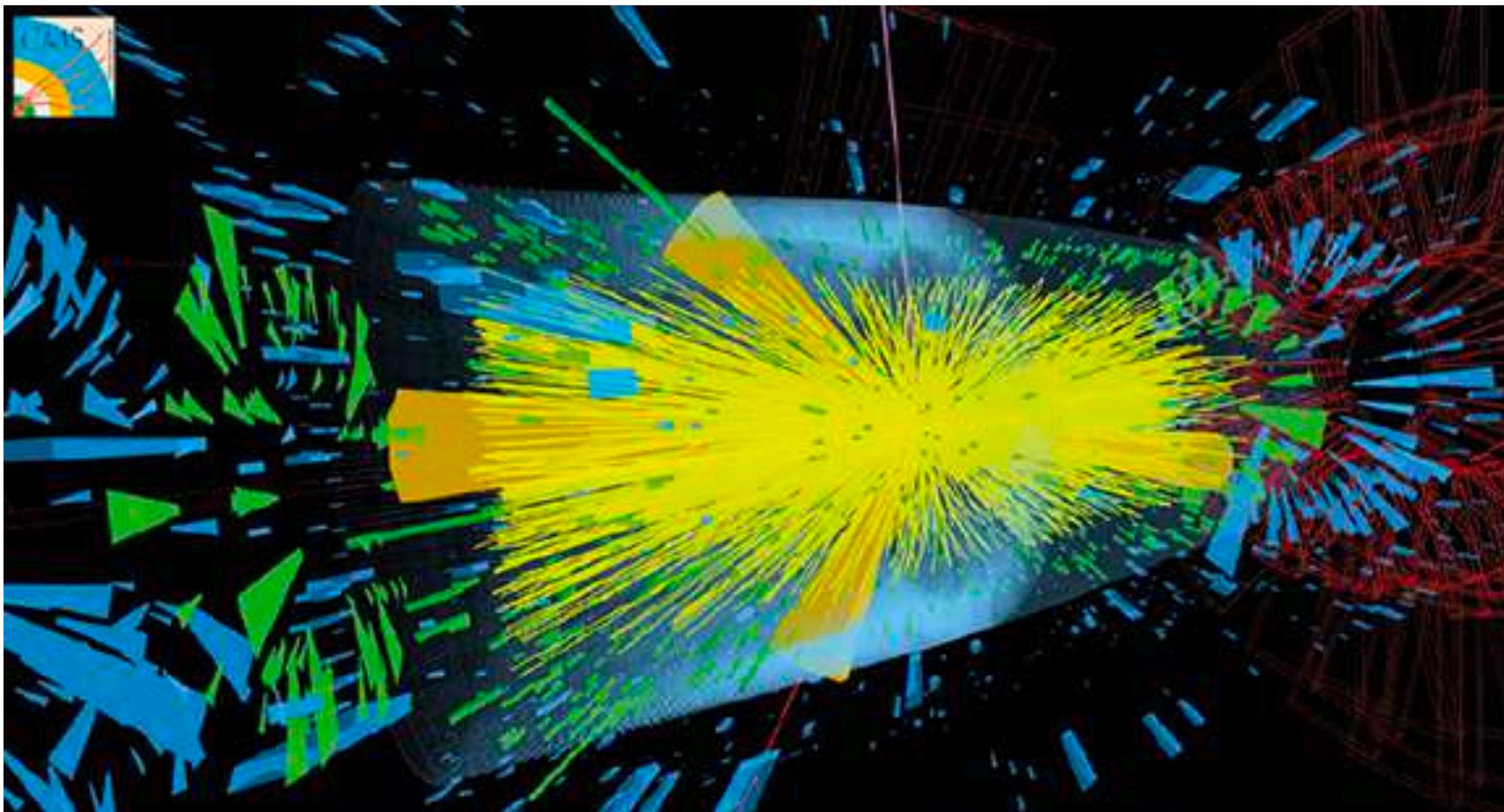
k -eventiness?

jet from dataset

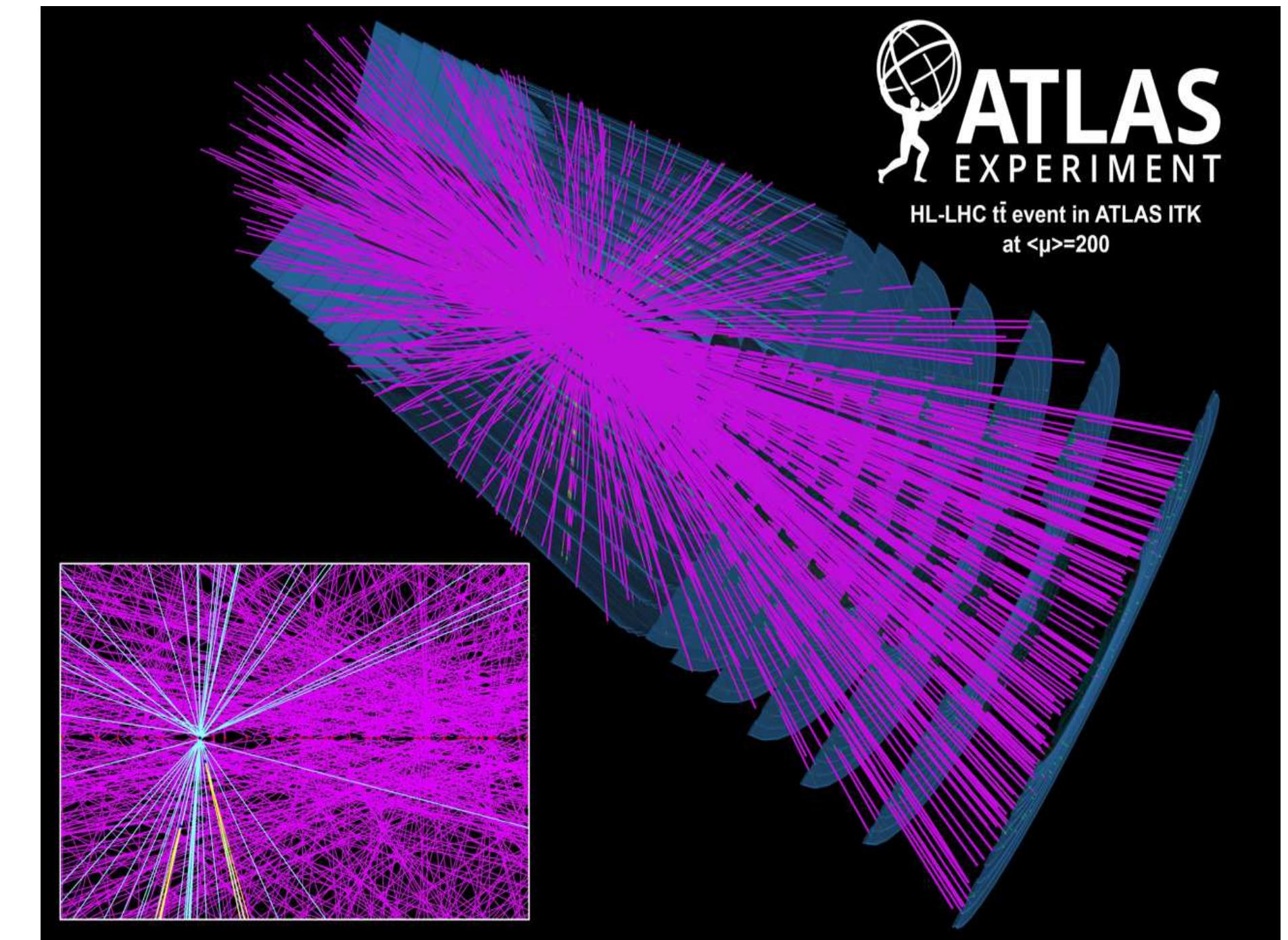
medoids

Pileup at the (HL-)LHC

Pileup is uniform (on average) radiation from additional proton-proton collisions



VBF Higgs + 200 pileup vertices

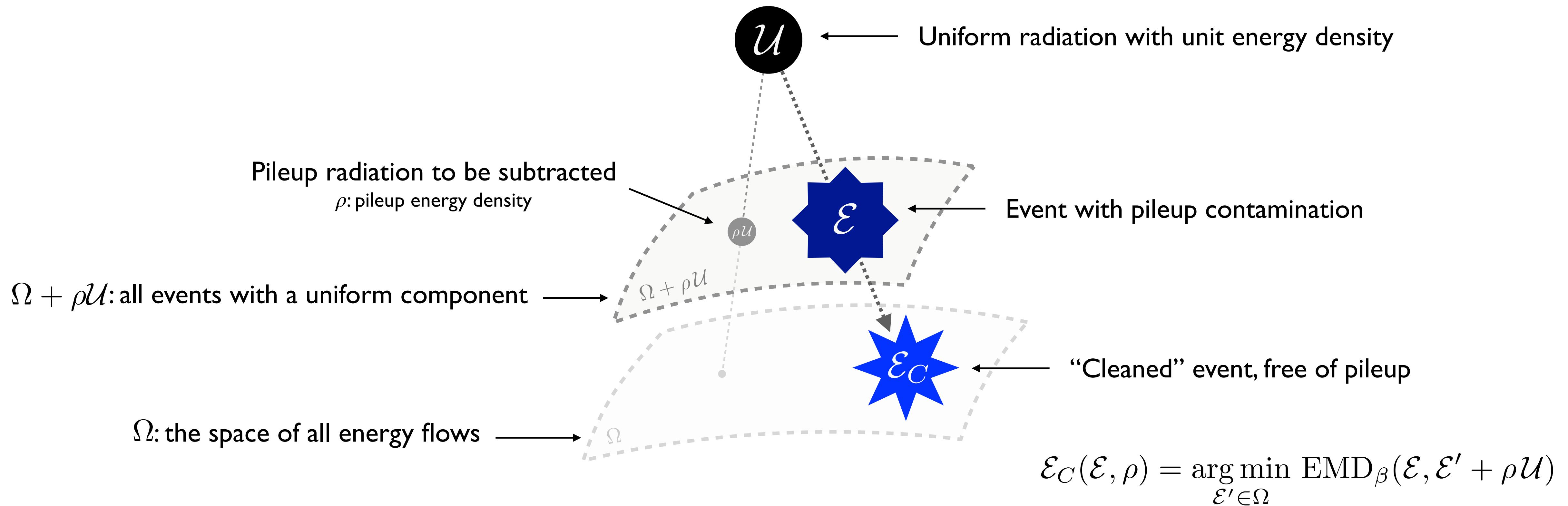


$t\bar{t}$ + 200 pileup vertices

Pileup Mitigation in Event Space

Pileup: uniform (on average) radiation from additional proton-proton collisions

Pileup mitigation: “moving away” from the uniform event

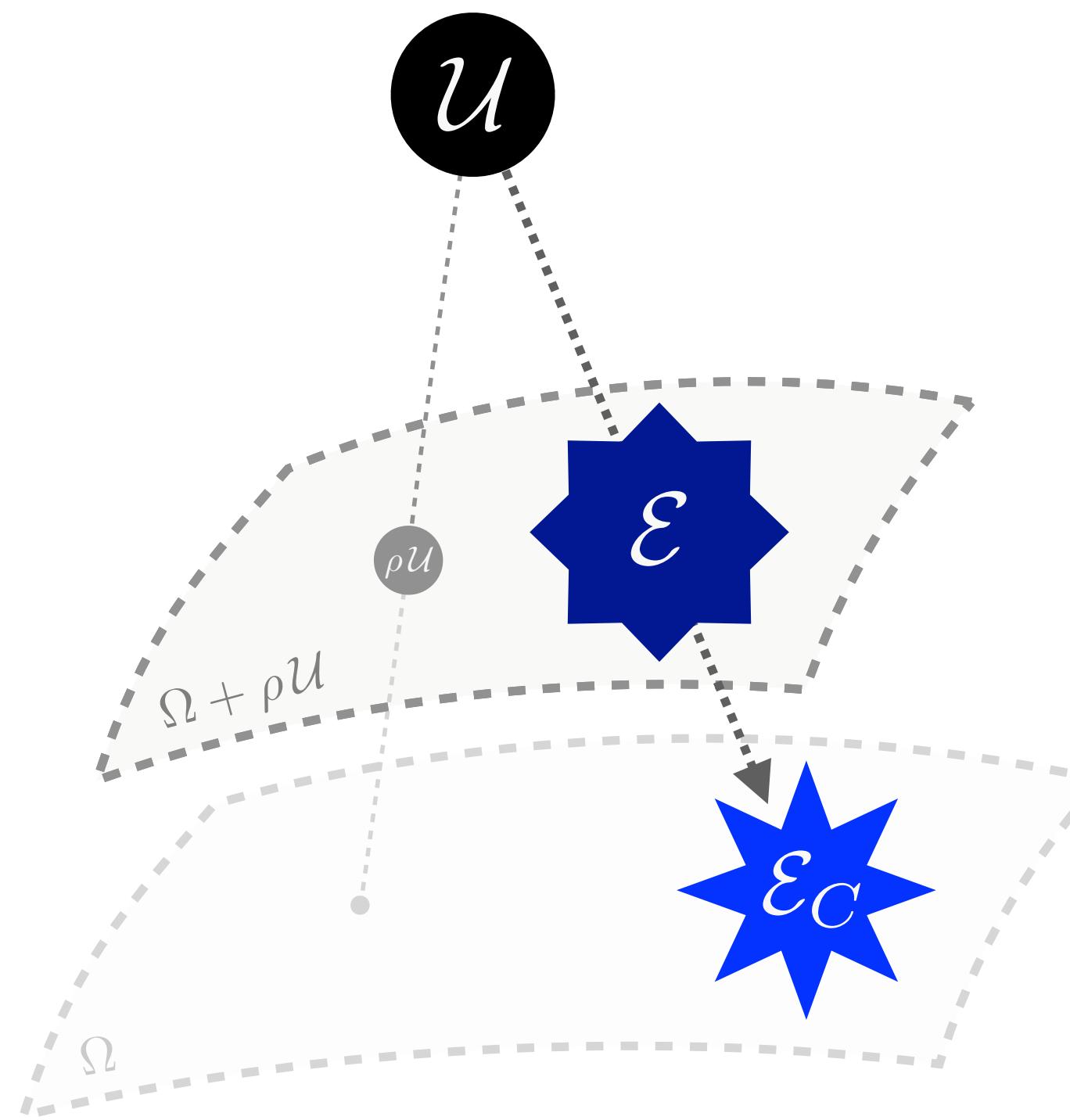


Pileup Mitigation in Event Space

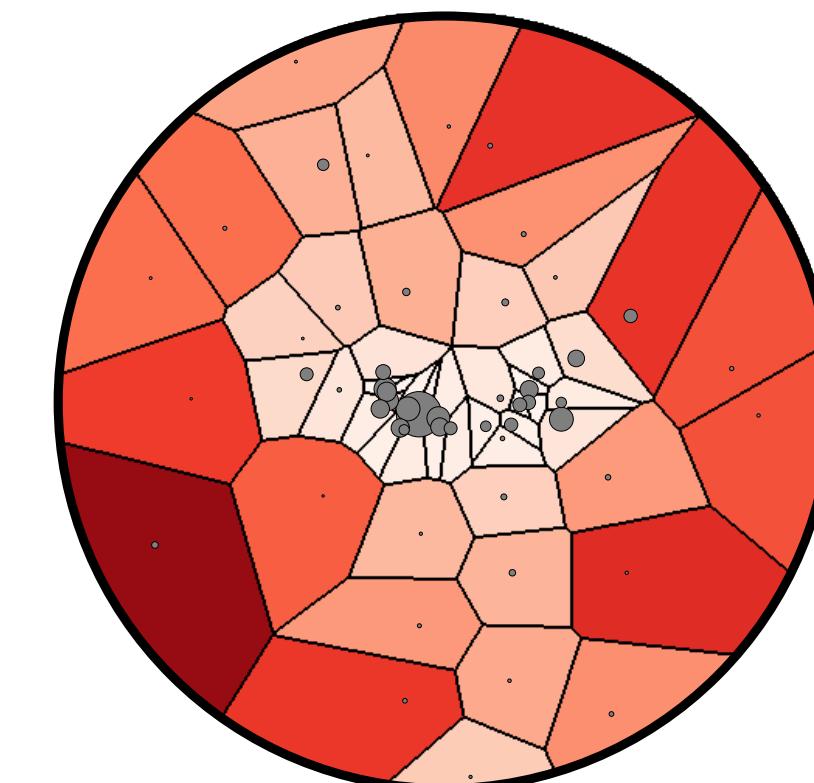
Area subtraction

Particle energies corrected proportional to area of associated region

$$\mathcal{E}_C(\mathcal{E}, \rho) = \arg \min_{\mathcal{E}' \in \Omega} \text{EMD}_{\beta}(\mathcal{E}, \mathcal{E}' + \rho \mathcal{U})$$



Voronoi

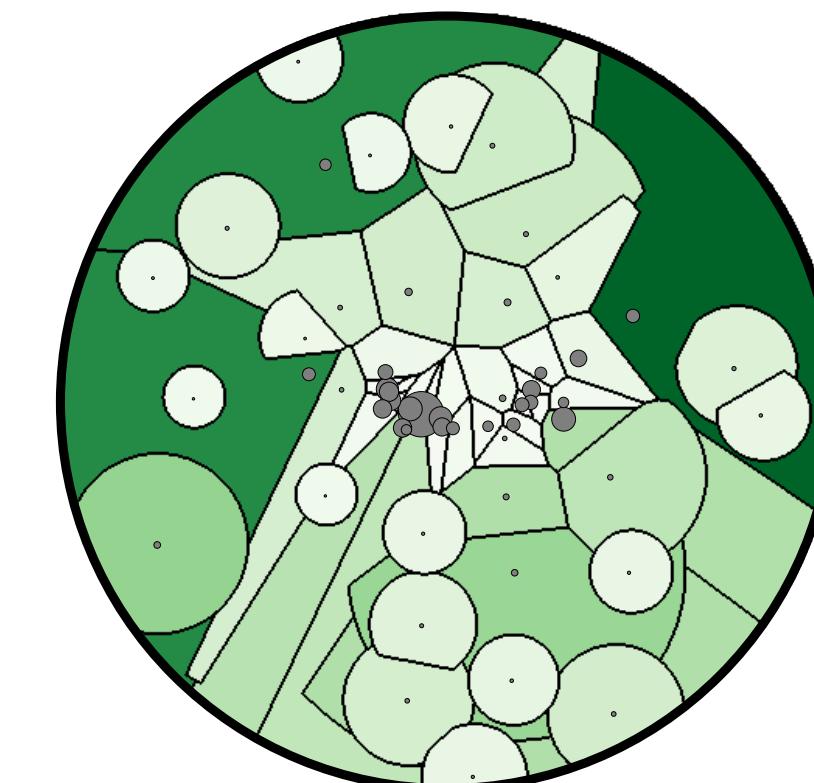


[Cacciari, Salam, Soyez, JHEP 2008]

Voronoi regions IRC unsafe

Sensitive to small modifications

Constituent subtraction



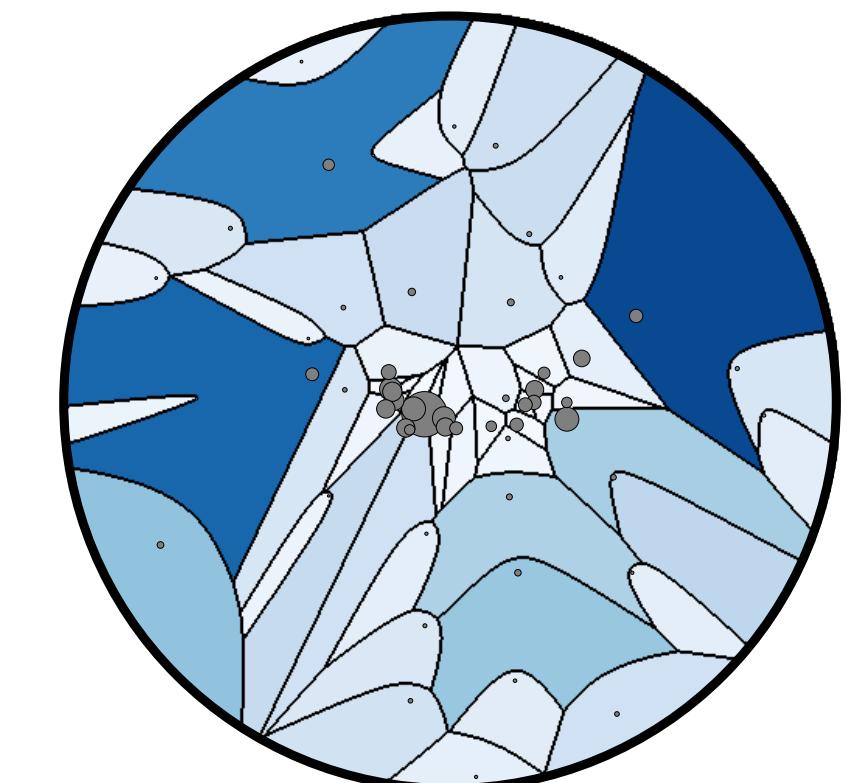
[Berta, Spousta, Miller, Leitner, JHEP 2008]

Lays down grid of “ghost” particles

Ghosts associate to nearest particle

Vanished particles don't attract ghosts

Apollonius



[PTK, Metodiev, Thaler, 2004.04159]

Ghosts are optimally assigned to particles by minimizing EMD

Apollonius regions have an understood continuum limit

Beyond Observables via Weighted Cross Sections

Standard observable (e.g. EFPs)

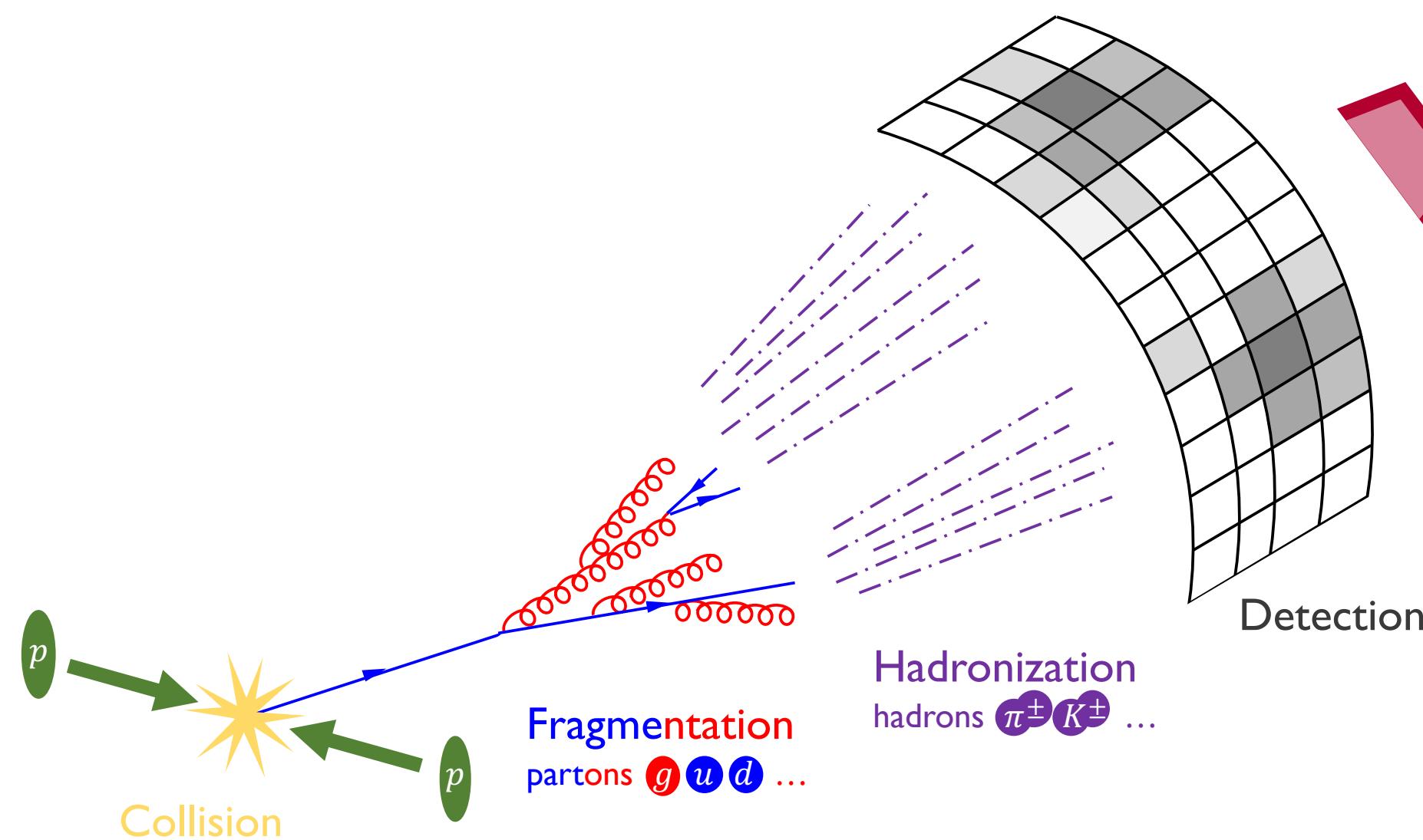
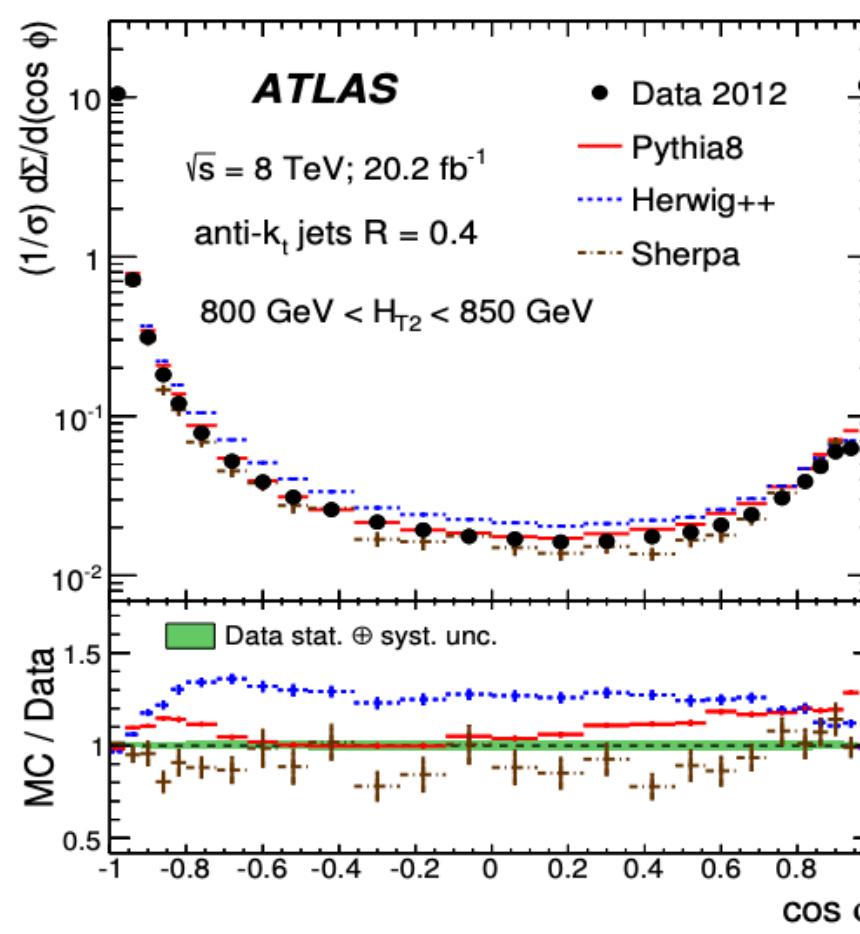
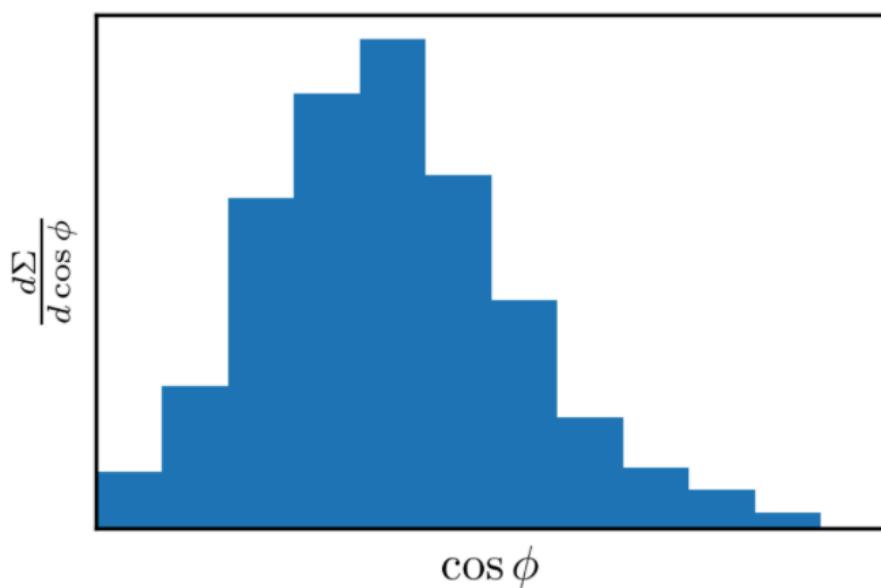
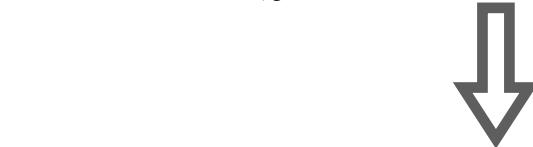
*Calculate a single number for each jet/event
and study distribution of values*

Weighted cross section

*Calculate a distributional quantity per event
and study the mean distribution*

e.g. energy-energy correlator (EEC)

$$\frac{d\Sigma}{d\cos\phi} = \sum_{i,j} \int d\sigma \frac{E_i E_j}{Q^2} \delta(\cos\theta_{ij} - \cos\phi)$$

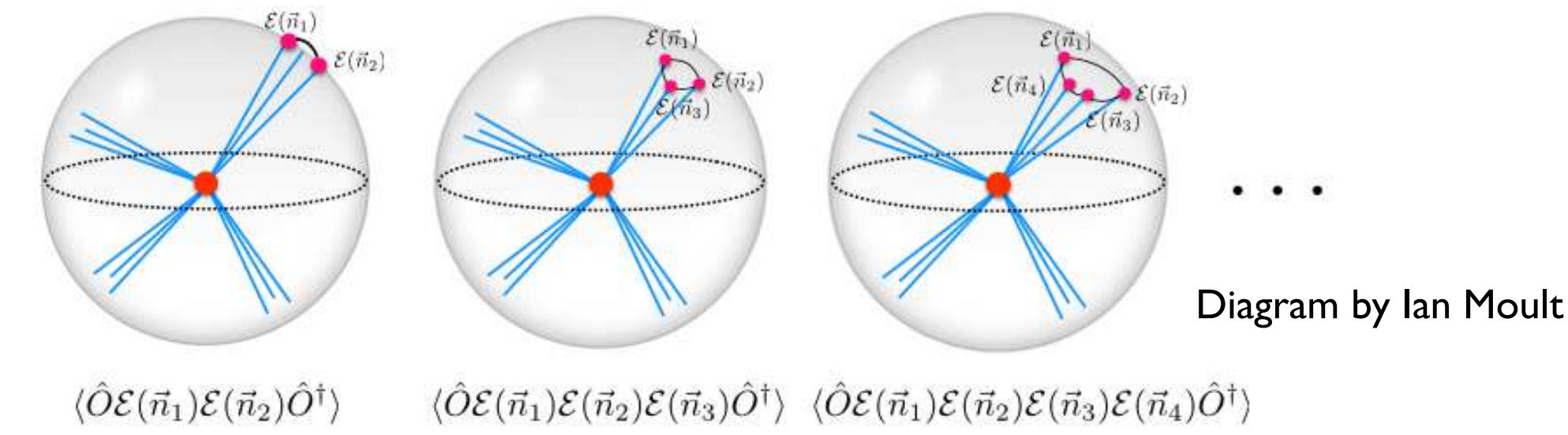


$$\mathcal{E}(\hat{n}) = \int_0^\infty dt \lim_{r \rightarrow \infty} r^2 n^i T_{0i}(t, r\hat{n})$$

Stress-energy flow

$$\frac{1}{\sigma_{\text{tot}}} \frac{d\sigma}{d\hat{n}_1 \cdots d\hat{n}_N} = \frac{\langle \mathcal{O}\mathcal{E}(\hat{n}_1) \cdots \mathcal{E}(\hat{n}_N)\mathcal{O}^\dagger \rangle}{\langle \mathcal{O}\mathcal{O}^\dagger \rangle}$$

Correlations of energy flow operators can be directly studied!



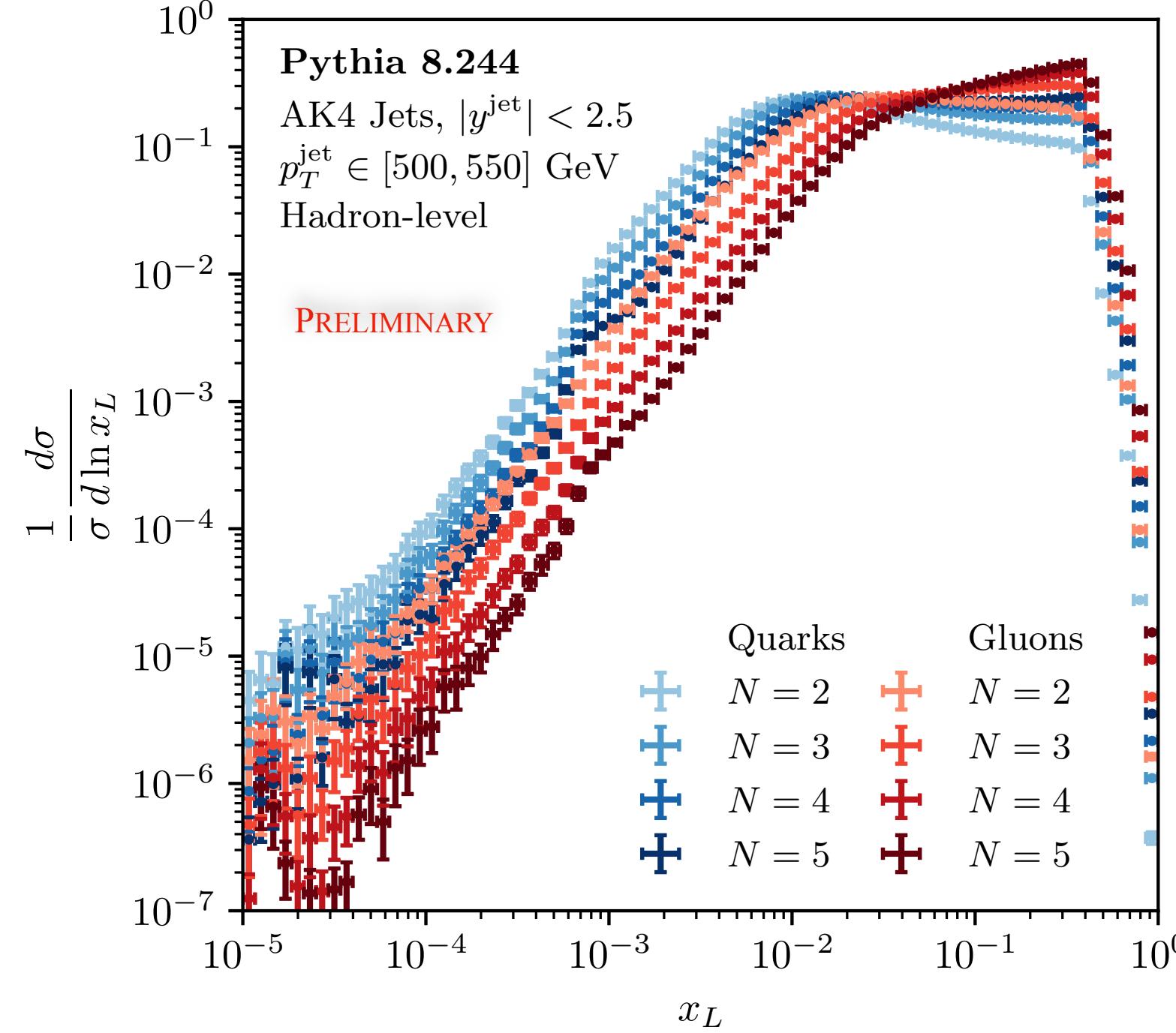
Energy-Energy Correlators – Projection to Longest Side

[PTK, Moult, Thaler, Zhu, to appear soon]

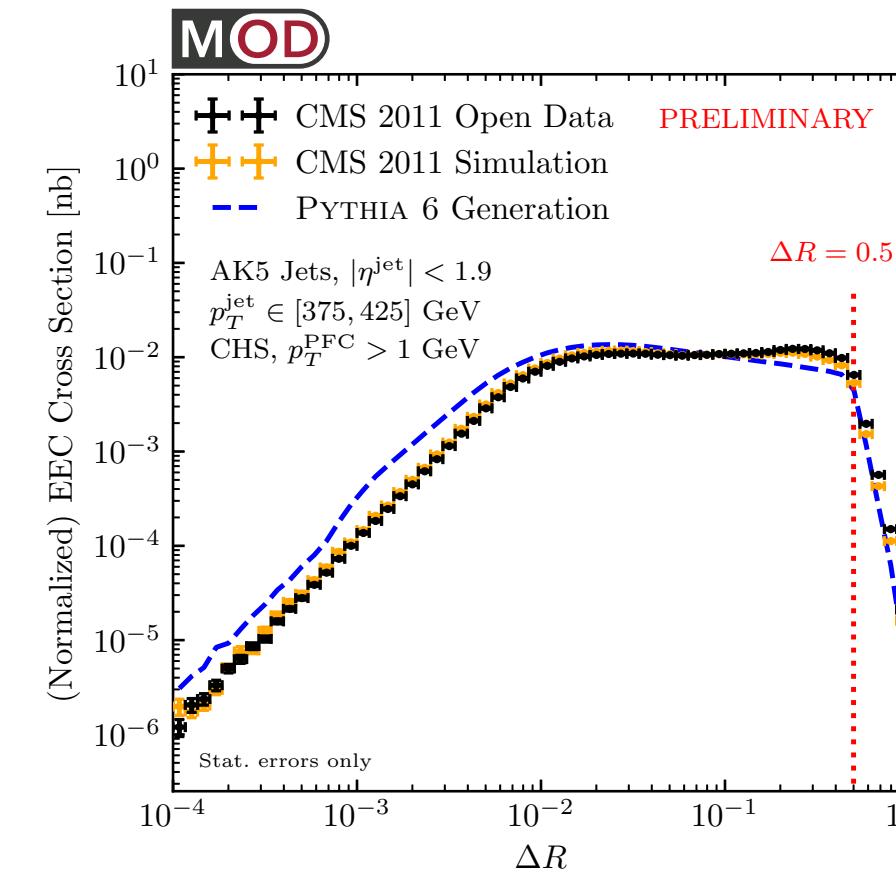
Integrate out shape dependence but keep overall size dependence

$$\frac{d\Sigma[N]}{dx_L} = \sum_n \sum_{1 \leq i_1 \leq \dots \leq i_N \leq n} \int d\sigma_n \frac{E_{i_1} \cdots E_{i_N}}{Q^N} \delta(x_L - \max_{1 \leq j < k \leq N} \{\theta_{i_j i_k}\})$$

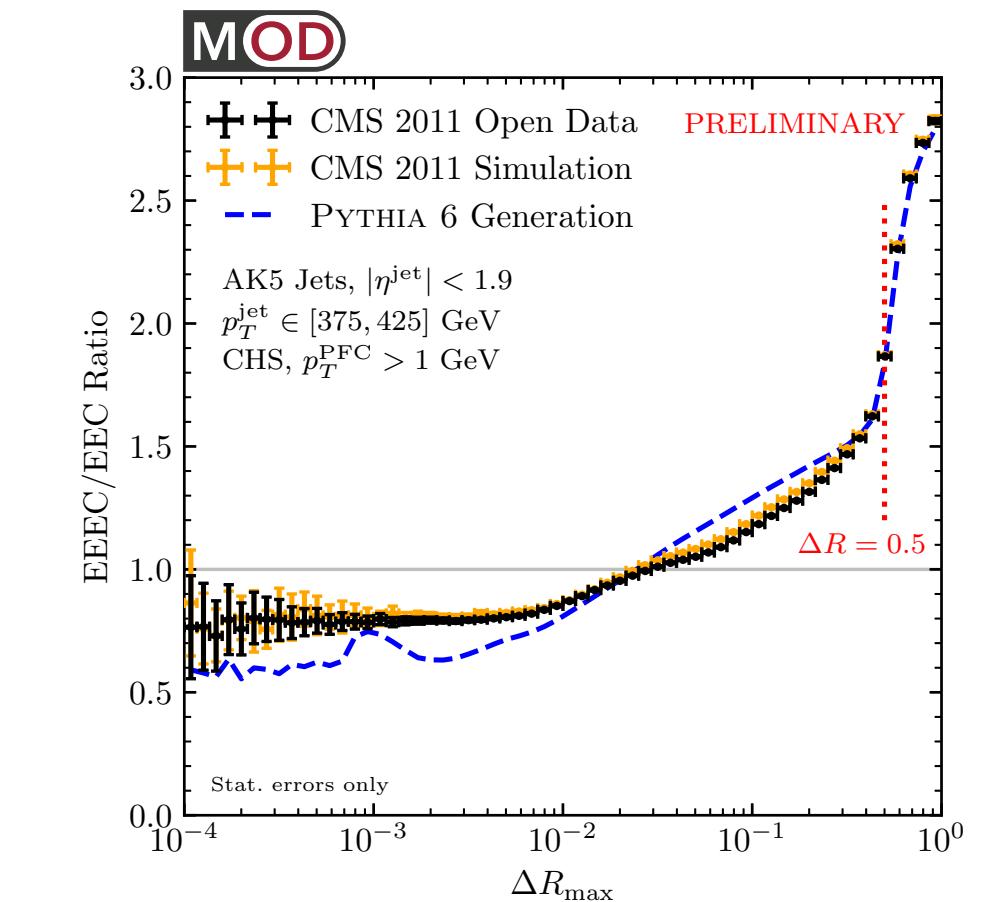
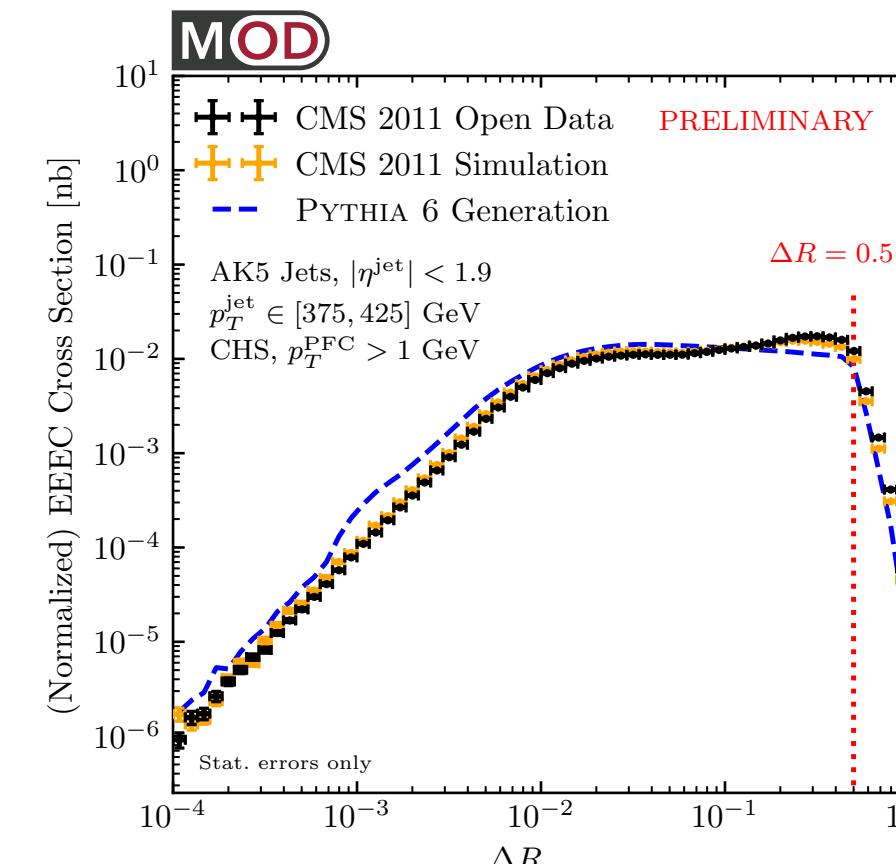
EEEC/EEC Ratio



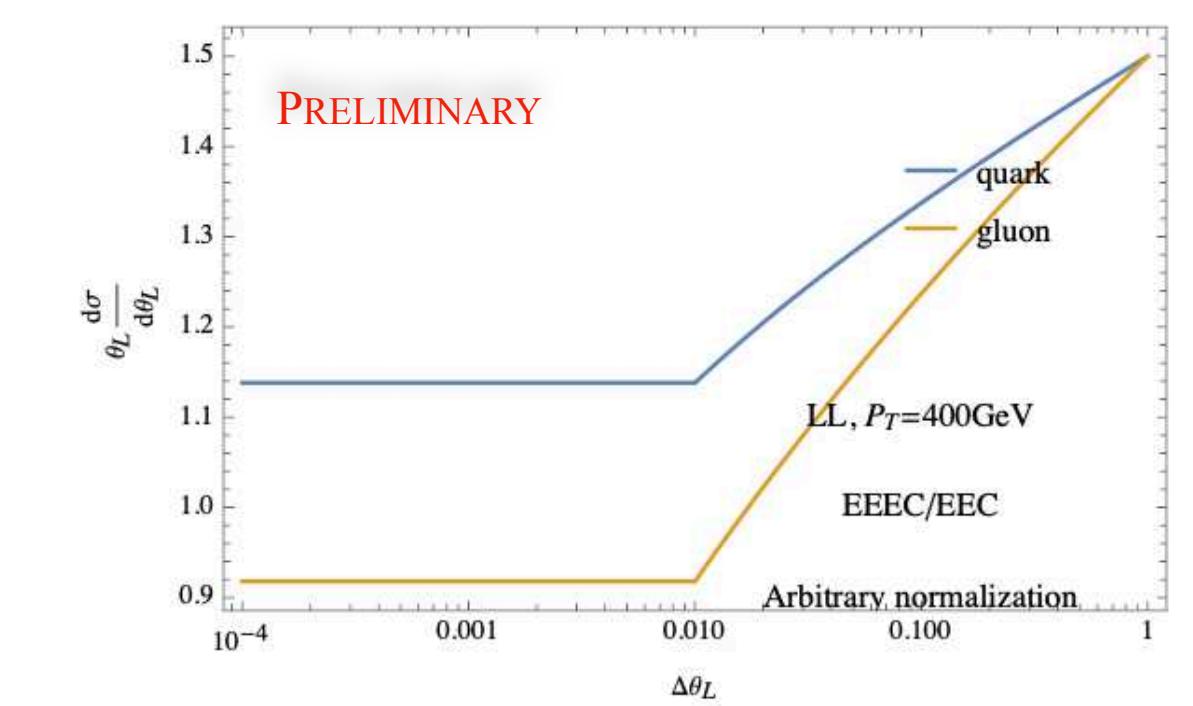
$N = 2$



$N = 3$

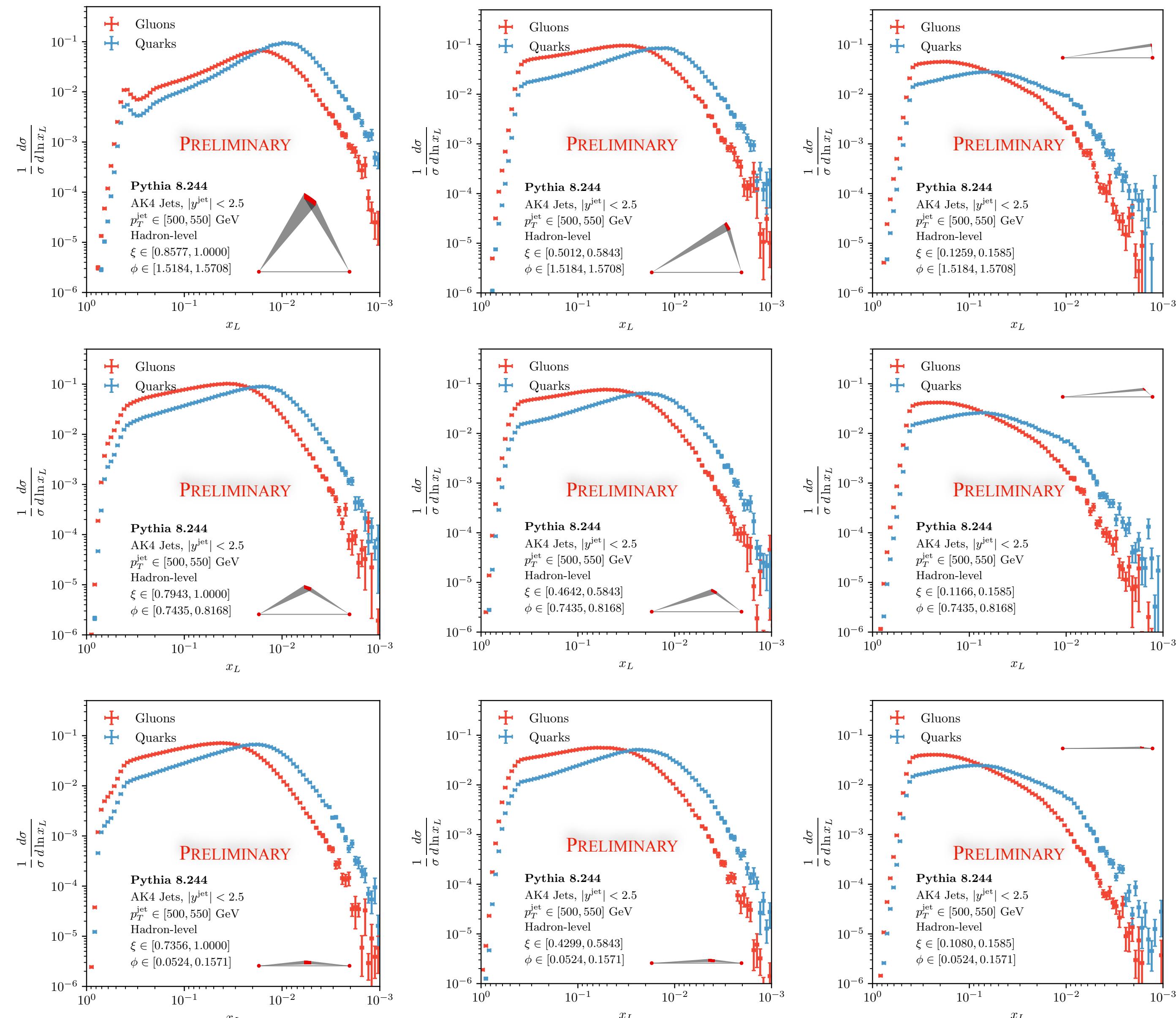
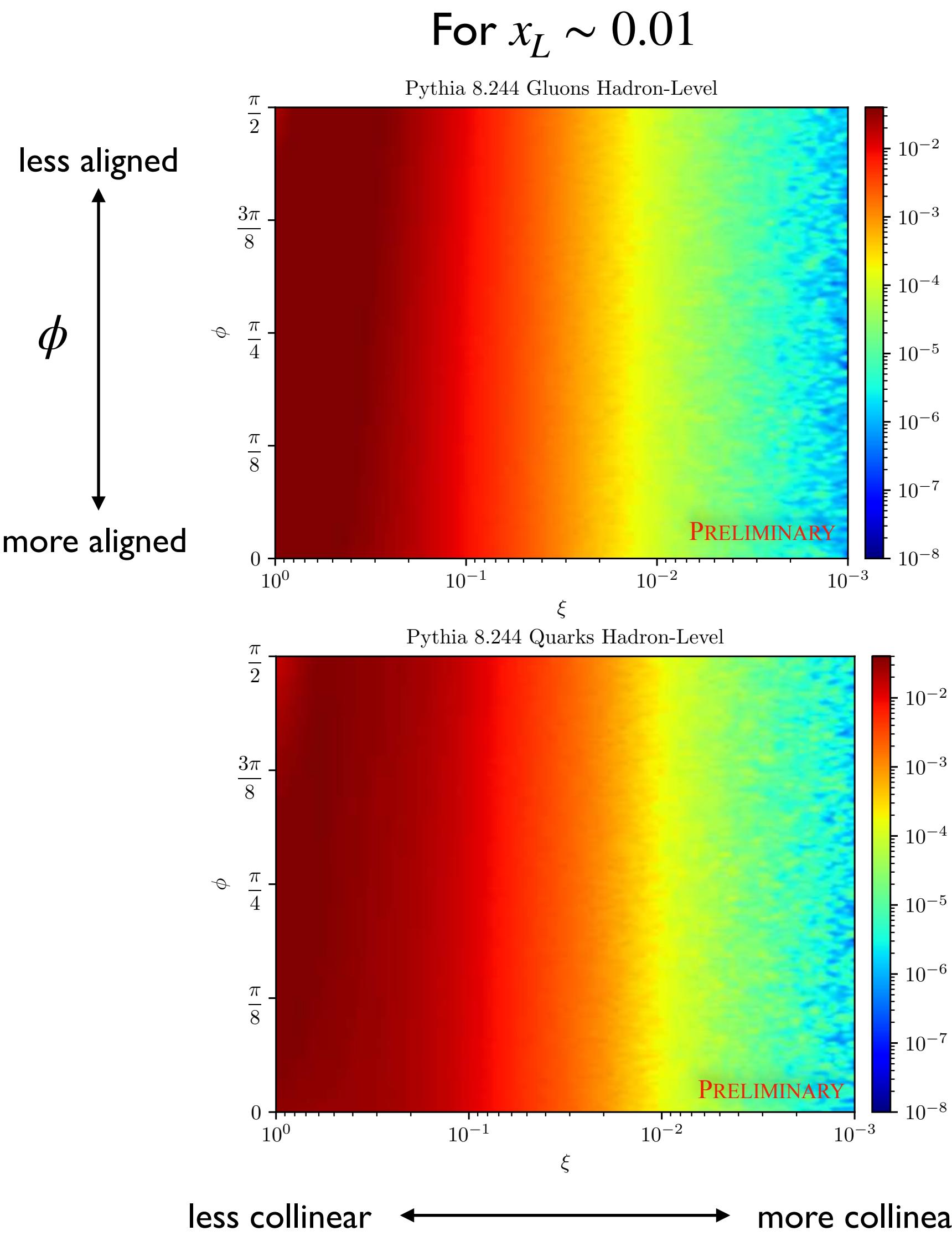


LL prediction of ratio



EEEC – Full Shape Dependence

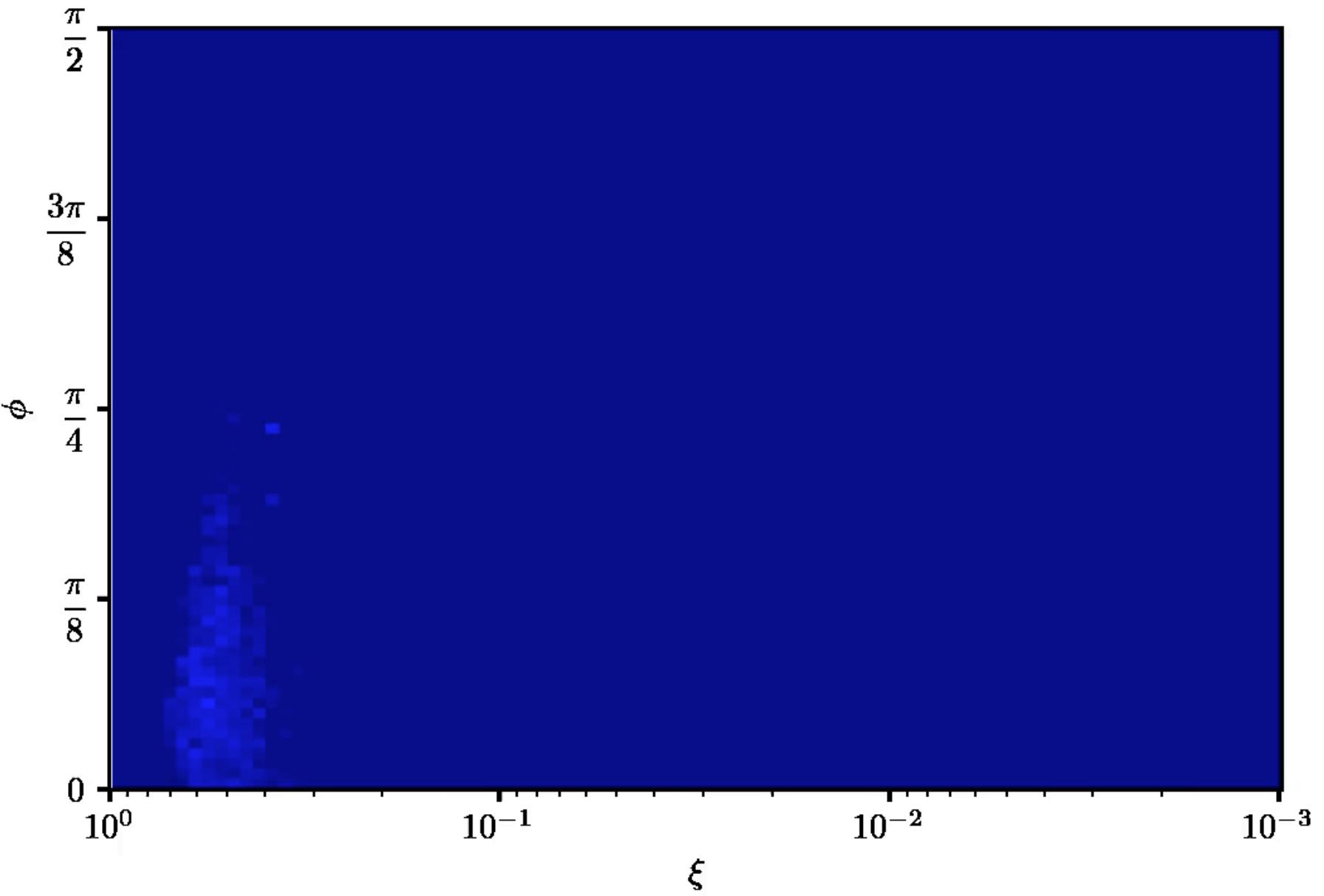
[PTK, Moult, Thaler, Zhu, to appear soon]



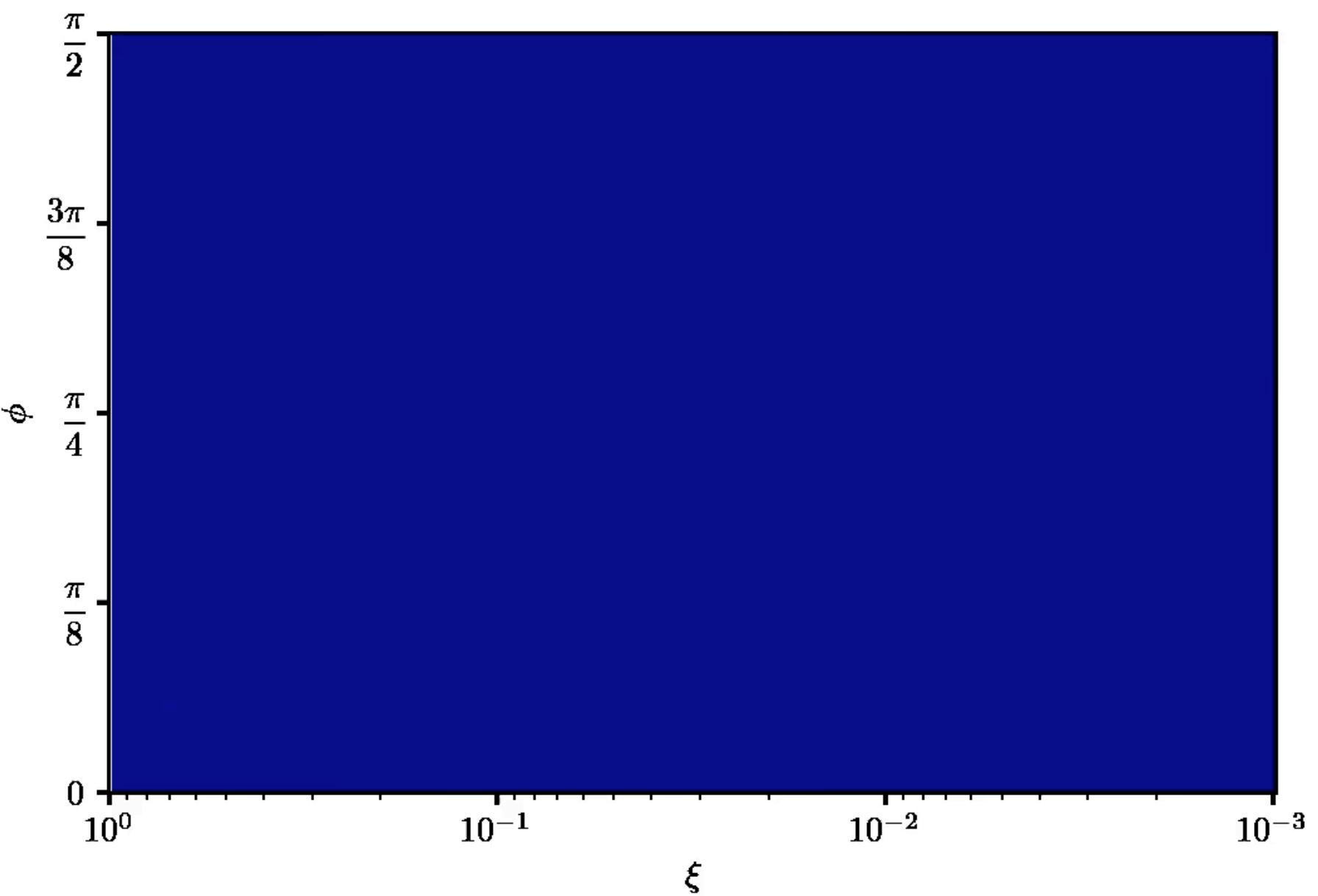
Visualizing the 3D EEEC

[PTK, Moult, Thaler, Zhu, to appear soon]

Pythia **Gluon** Jets
 $p_T \in [500,550]$ GeV



Pythia **Quark** Jets
 $p_T \in [500,550]$ GeV



Time in the videos corresponds to
 $\ln x_L$ going from 0 to $-\infty$

Color corresponds to log of EEC
(red is large, blue is small)

Uniformly persistent **red** is roughly the
perturbatively accessible region