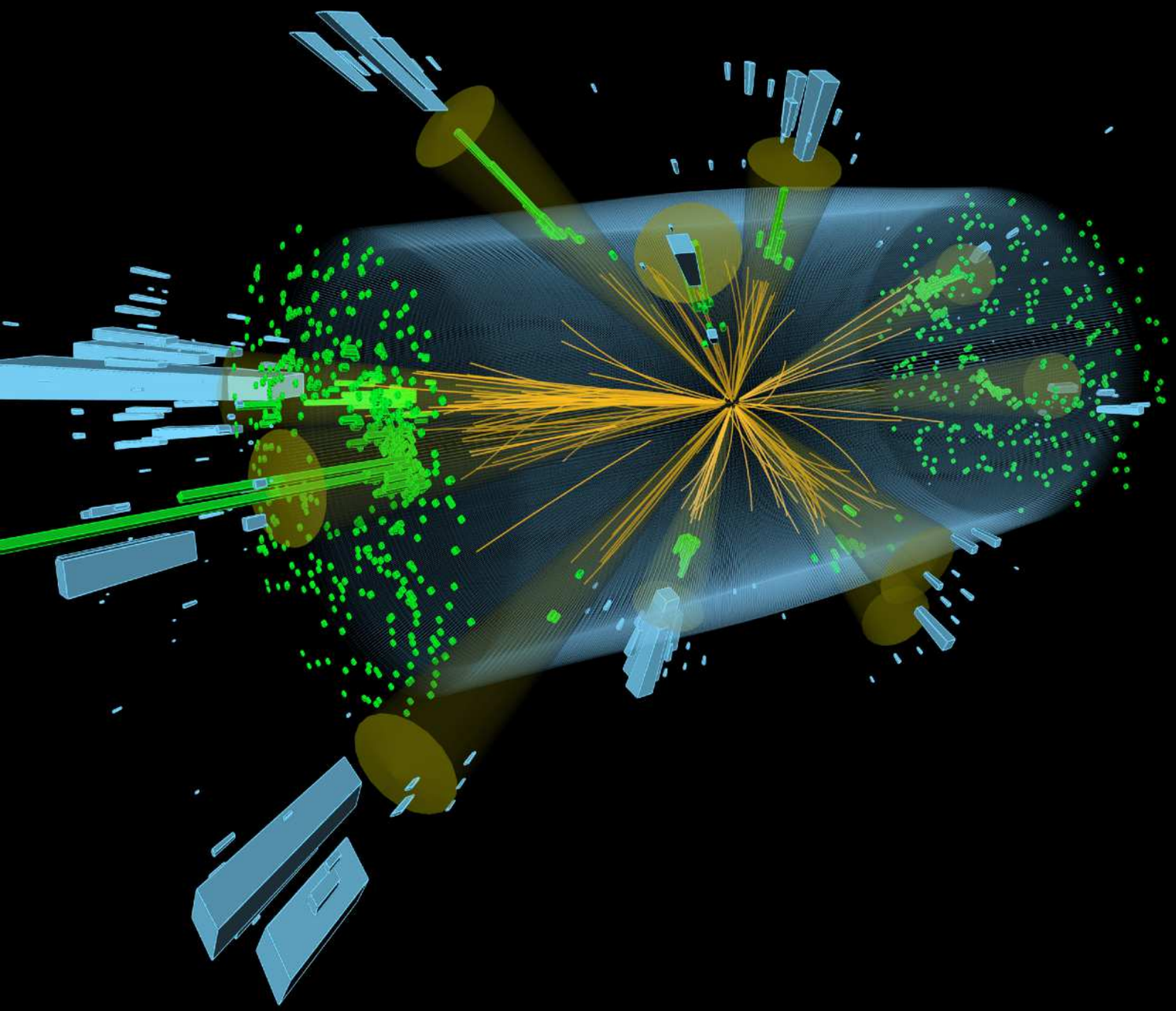
The Space of Collider Events

Jesse Thaler

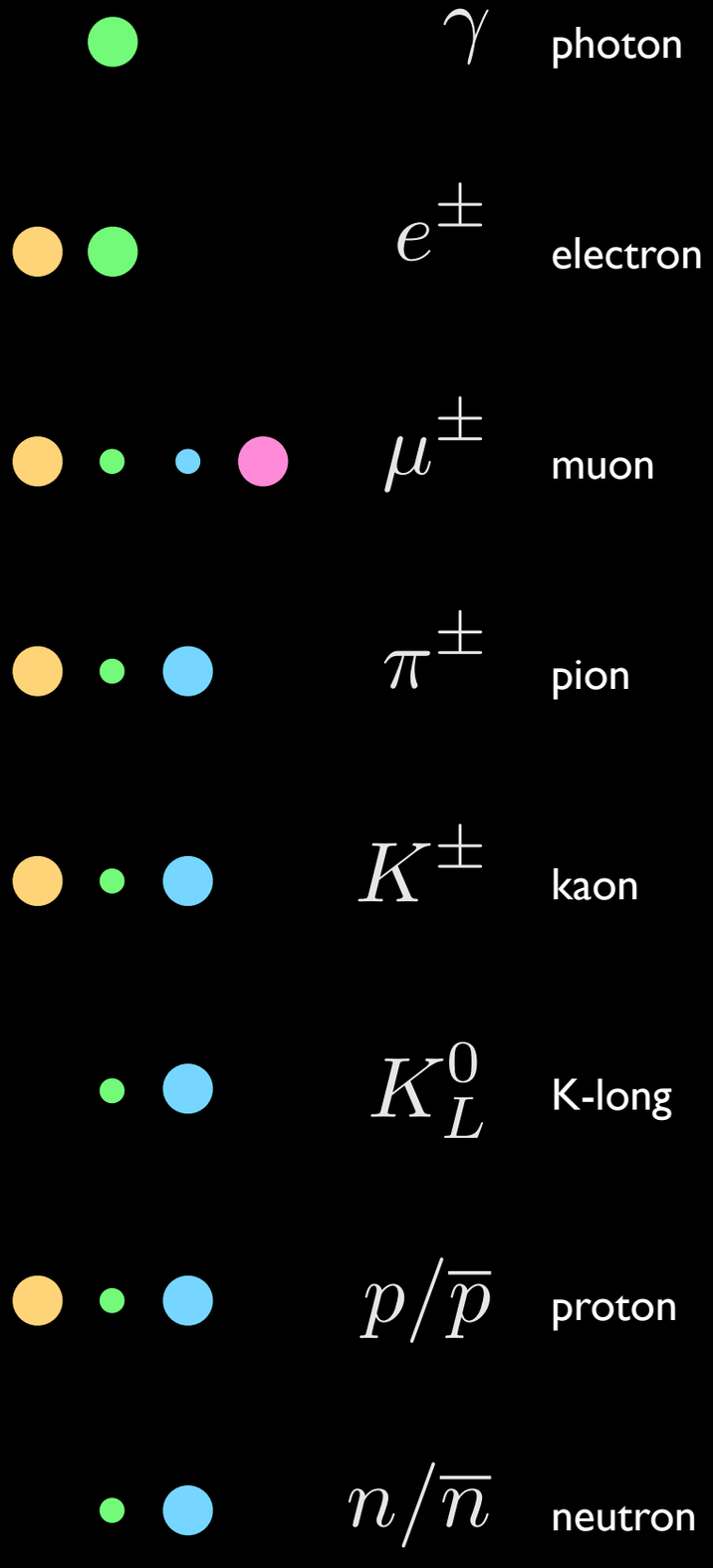
with Patrick Komiske & Eric Metodiev, 1902.02346

Theory Seminar, Jožef Stefan Institute, Ljubljana — June 28, 2019

One Collider Event



T E H M

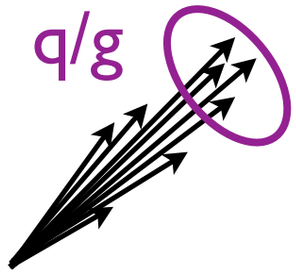


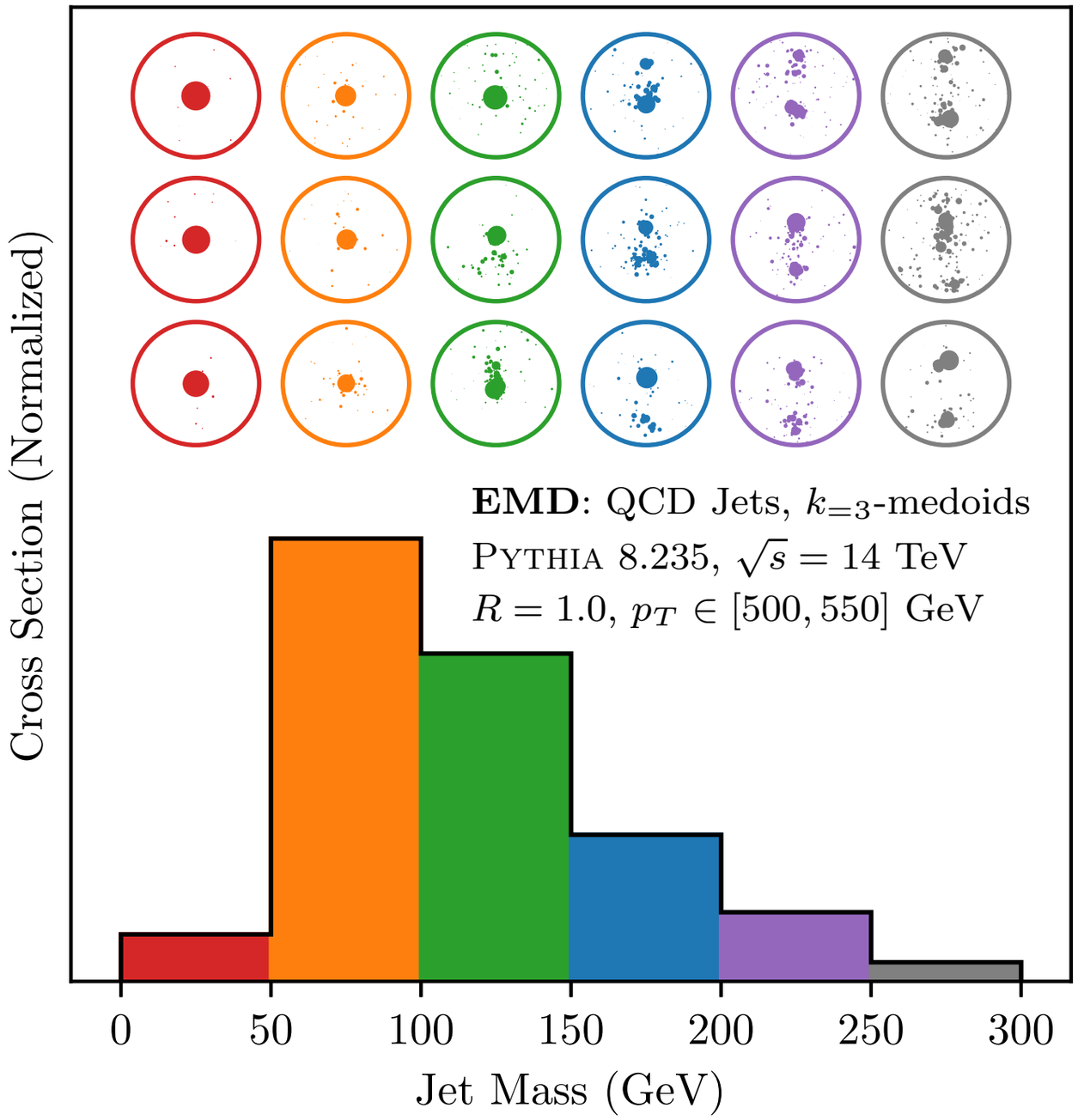
One Collider Event (for an Event)

[BOOST 2019: July 22-26, 2019, MIT]

N Events ⇒ Histogram

Here, N Jets

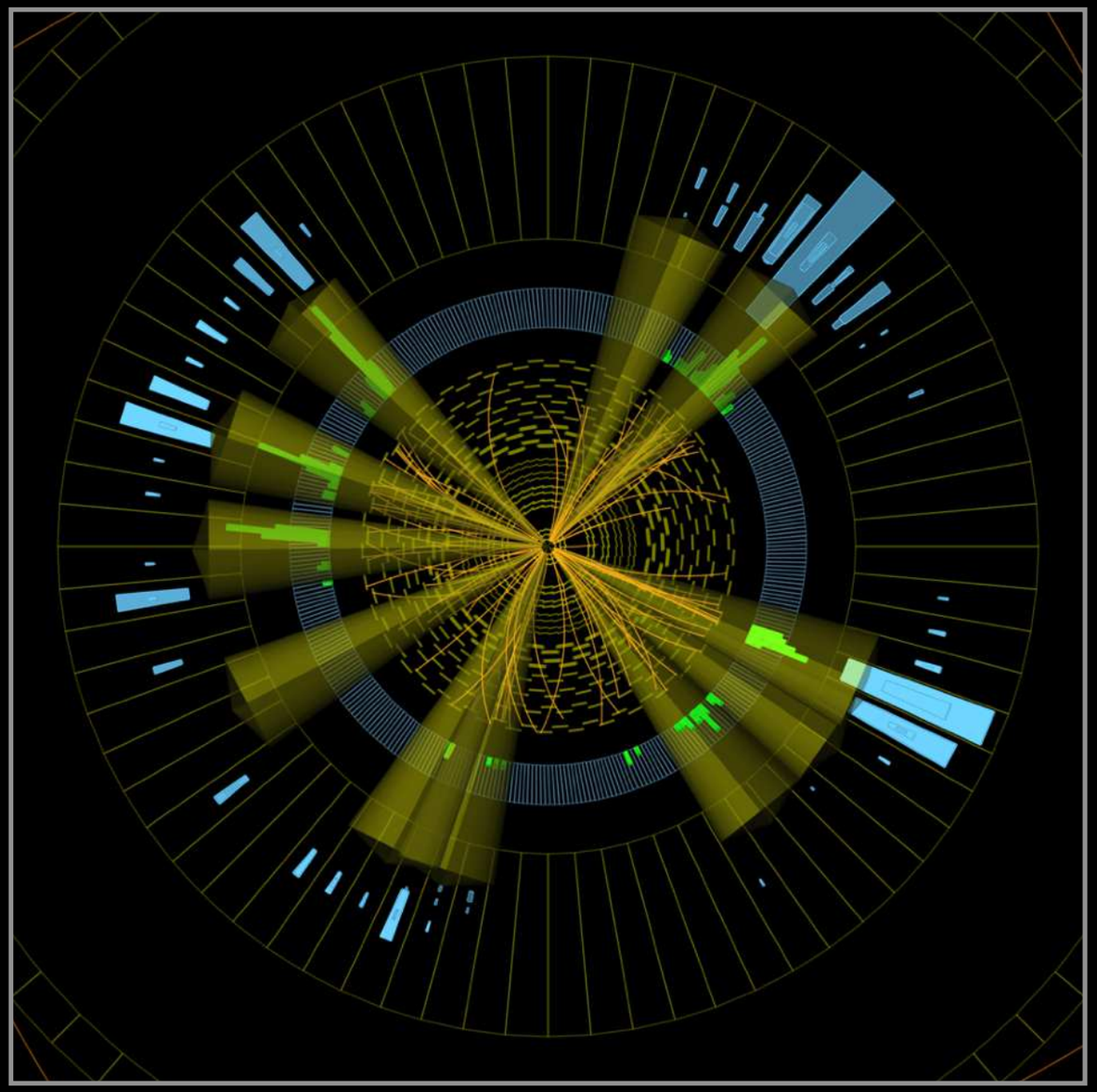


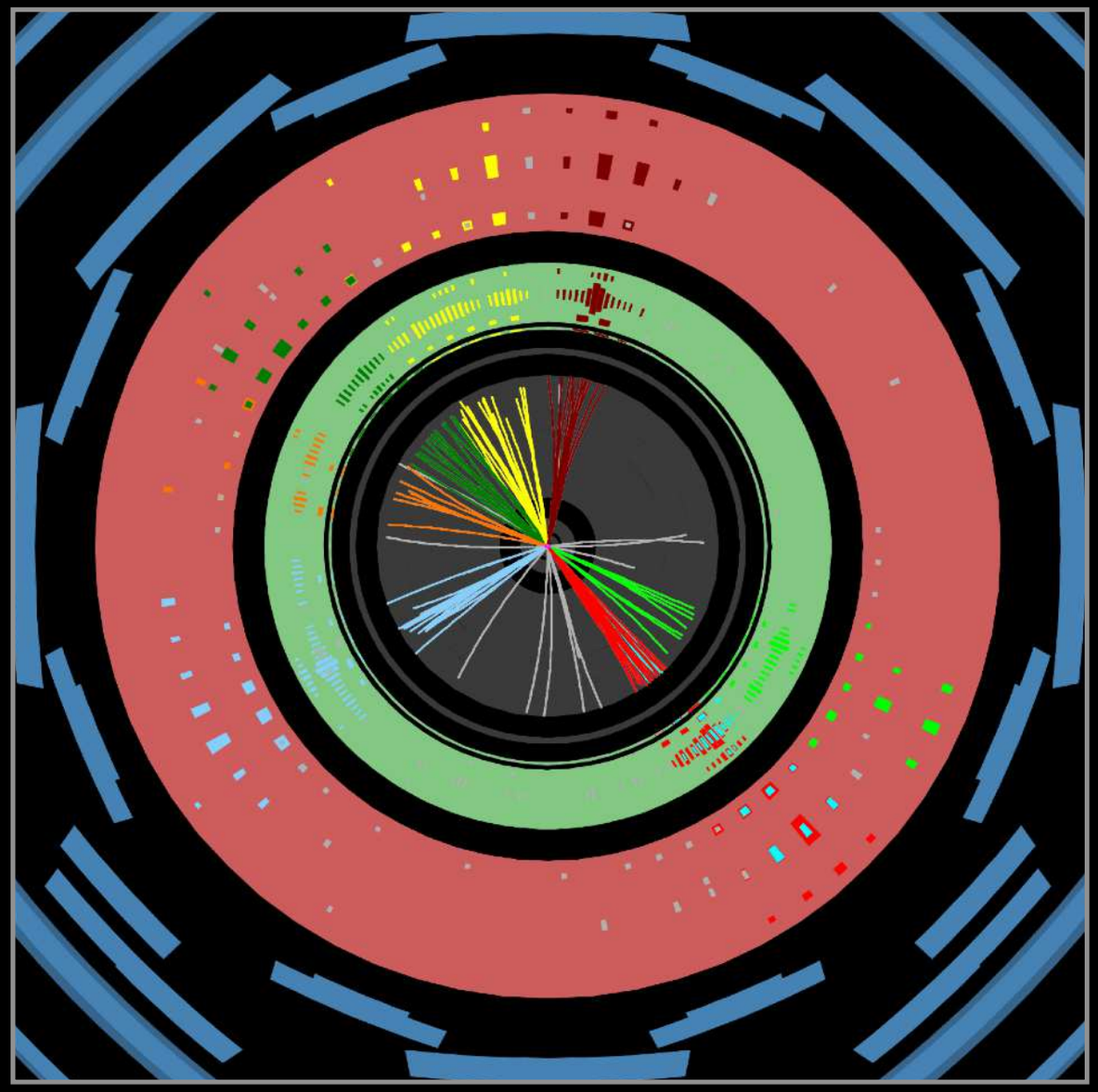


Three most representative jets in each bin (more later)

[Komiske, Metodiev, JDT, 1902.02346]

Two Collider Events

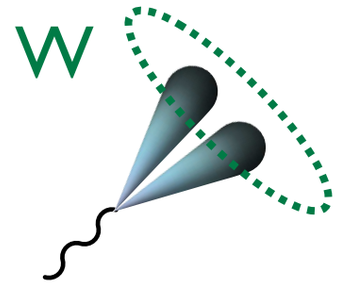


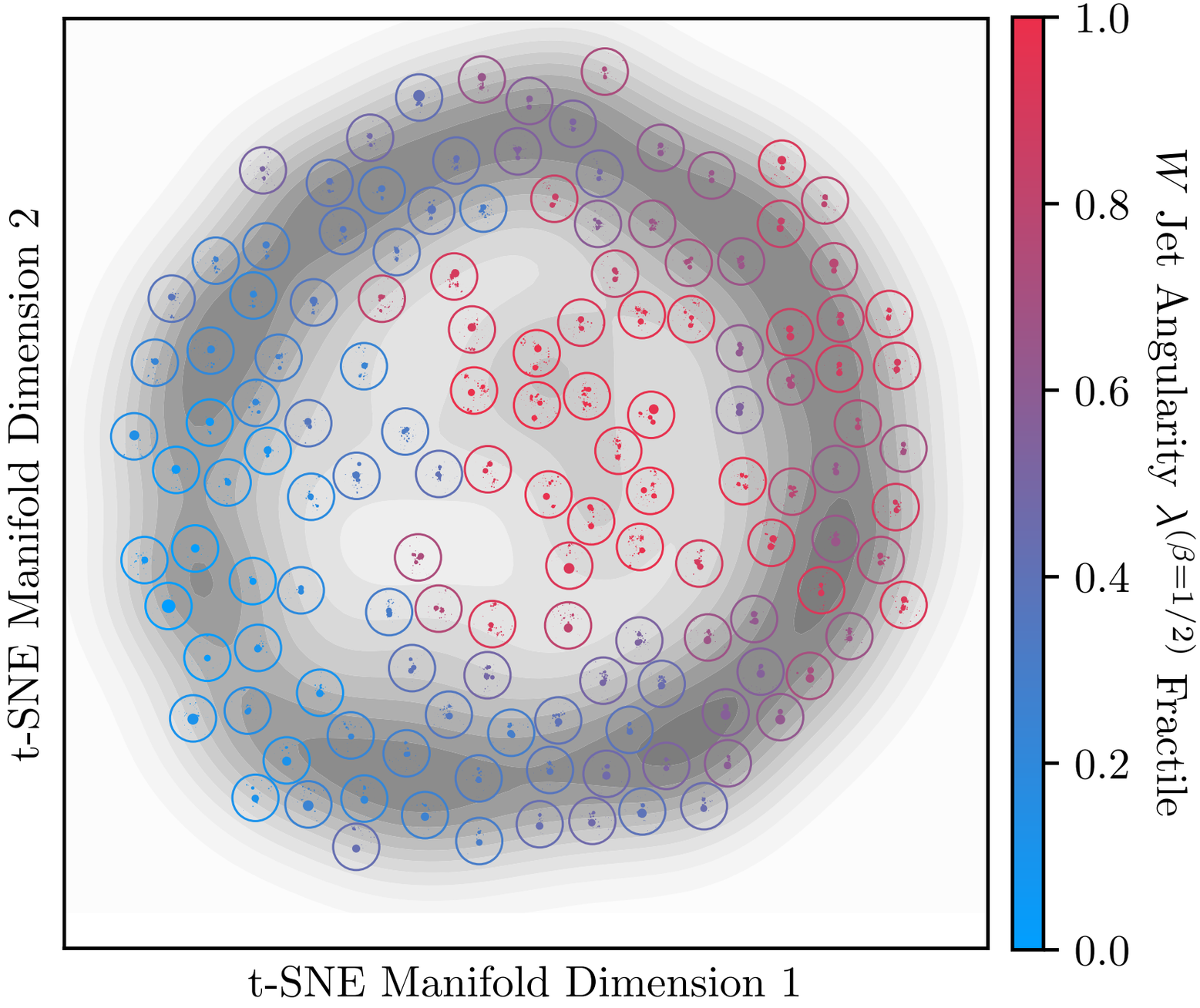


What is the "distance" between these events? (8.5 km?)

N² Distances ⇒ Metric Space

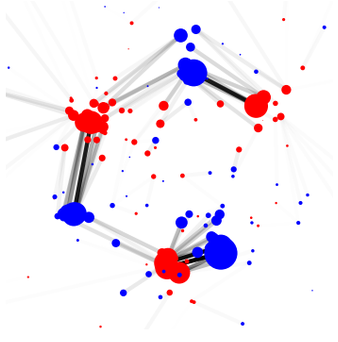
E.g. Boosted W Jets



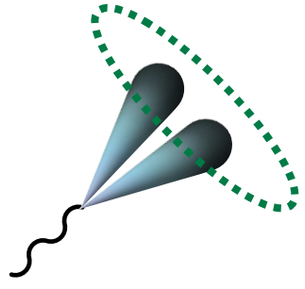


[Komiske, Metodiev, JDT, 1902.02346]

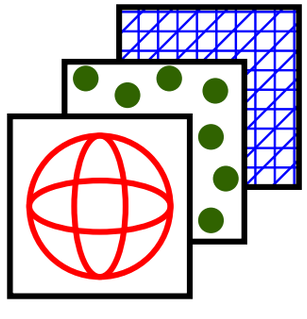
Outline



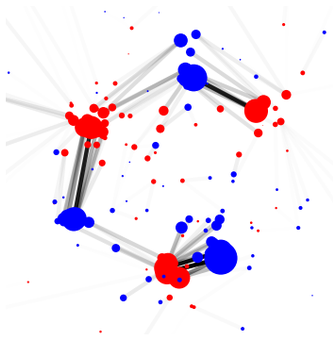
The Energy Mover's Distance



Initial Physics Studies



(Broader Comments)

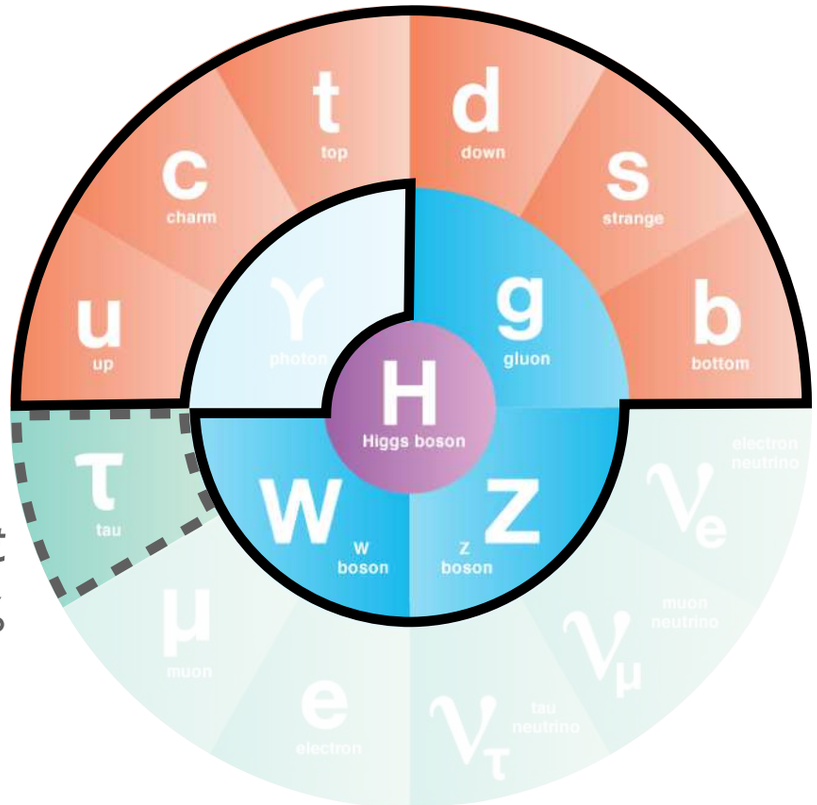


The Energy Mover's Distance

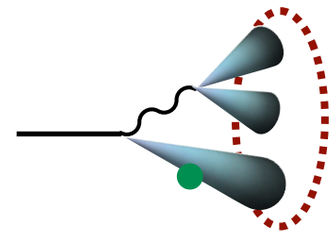
Initial Physics Studies

(Broader Comments)

Jets at the LHC

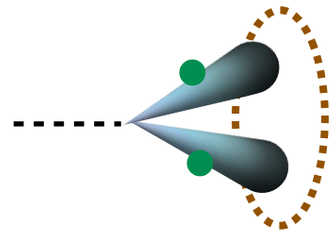


micro-jet
≈ 65%



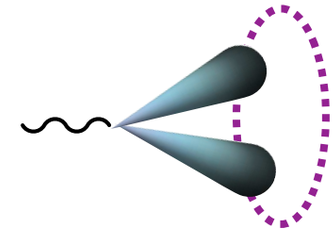
t
≈ 70%

173 GeV++



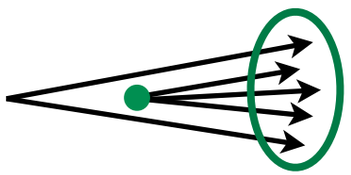
H
≈ 60%

125 GeV



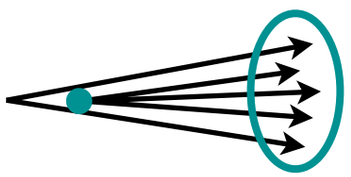
W/Z
≈ 70%

80/91 GeV



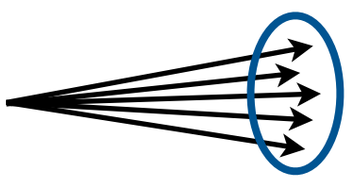
b

4.2 GeV++



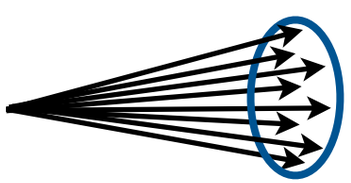
c

1.3 GeV++



u,d,s

100 MeV++



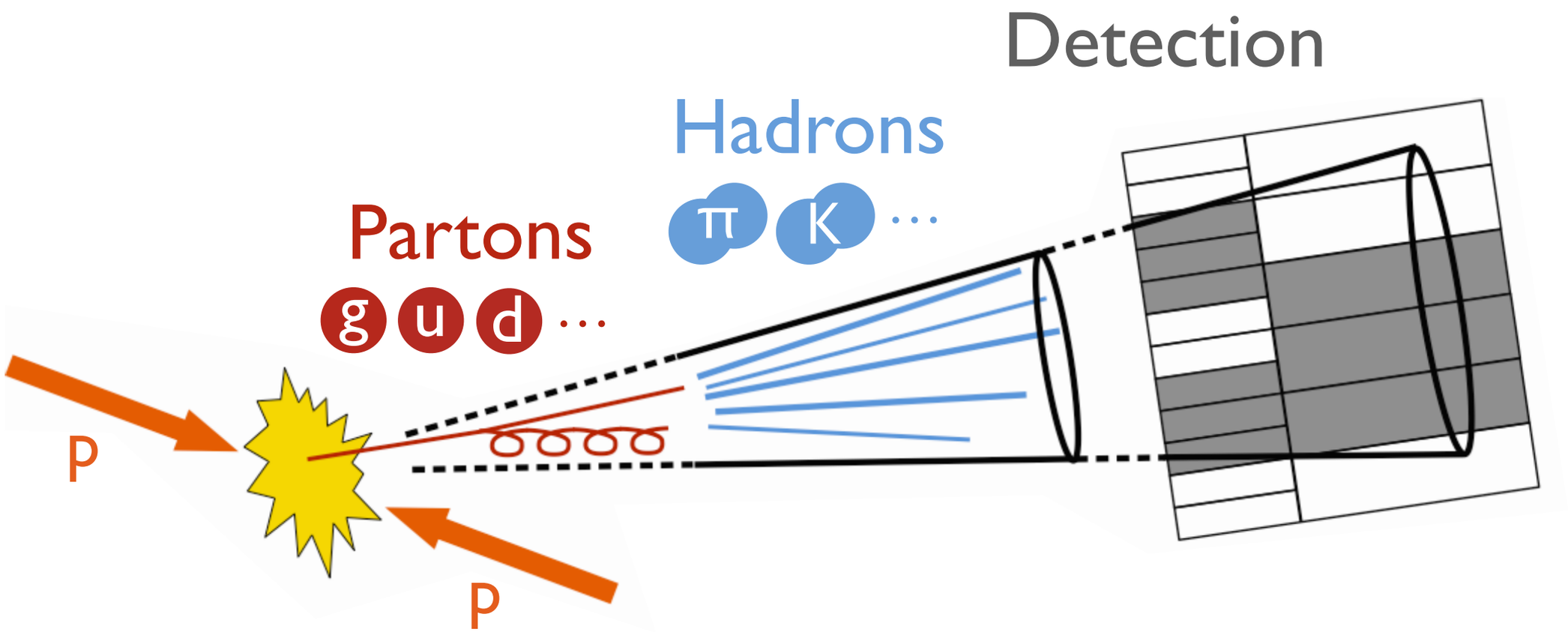
g

0++

++ = Mass from QCD Radiation

[see review in Larkoski, Moult, Nachman, 1709.04464]

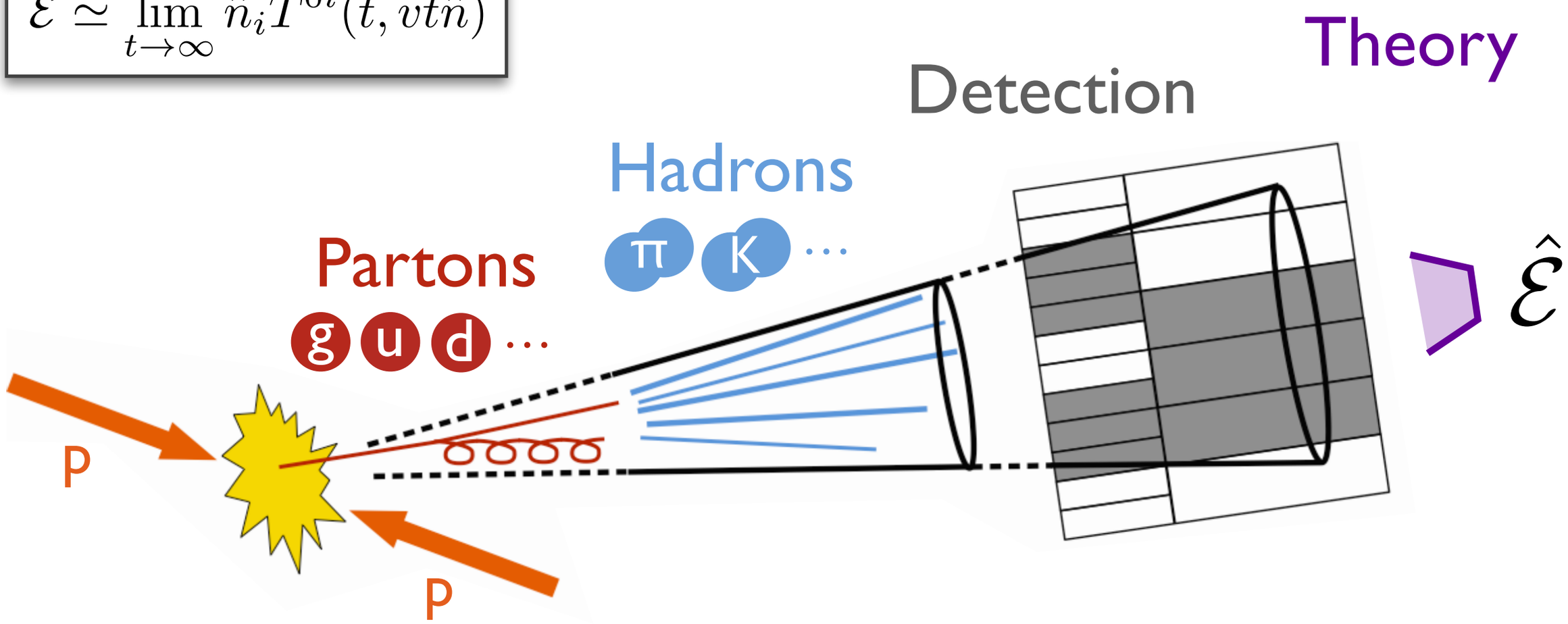
Focus on Energy Flow



[Sveshnikov, Tkachov, hep-ph/9512370; Hofman, Maldacena, 0803.1467; Mateu, Stewart, JDT, 1209.3781; Komiske, Metodiev, JDT, 1712.07124, 1810.05165]

Focus on Energy Flow

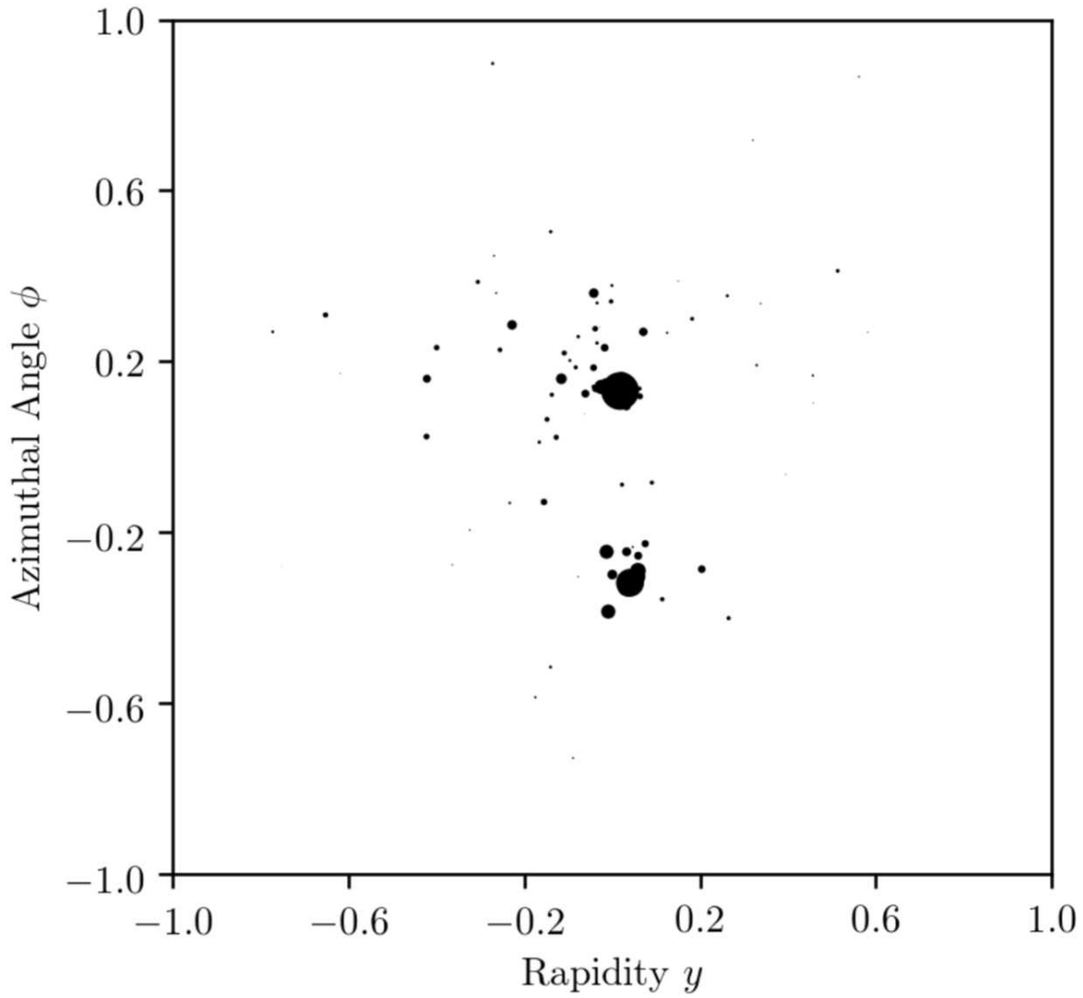
$$\hat{\mathcal{E}} \simeq \lim_{t\to\infty} \hat{n}_i T^{0i}(t, vt\hat{n})$$



Stress-energy flow: *Measure of event/jet structure robust to non-perturbative and detector effects (i.e. IRC safe)*

[Sveshnikov, Tkachov, hep-ph/9512370; Hofman, Maldacena, 0803.1467; Mateu, Stewart, JDT, 1209.3781; Komiske, Metodiev, JDT, 1712.07124, 1810.05165]

Focus on Energy Flow



Represent jet as:

$$\rho(\hat{p}) = \sum_{i \in \text{jet}} E_i \, \delta(\hat{p} - \hat{p}_i)$$

Energy (p_T)

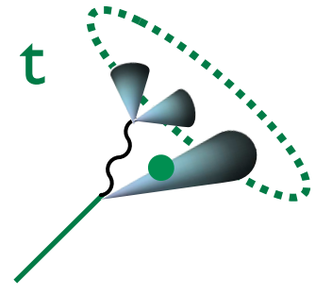
Direction (y,φ)

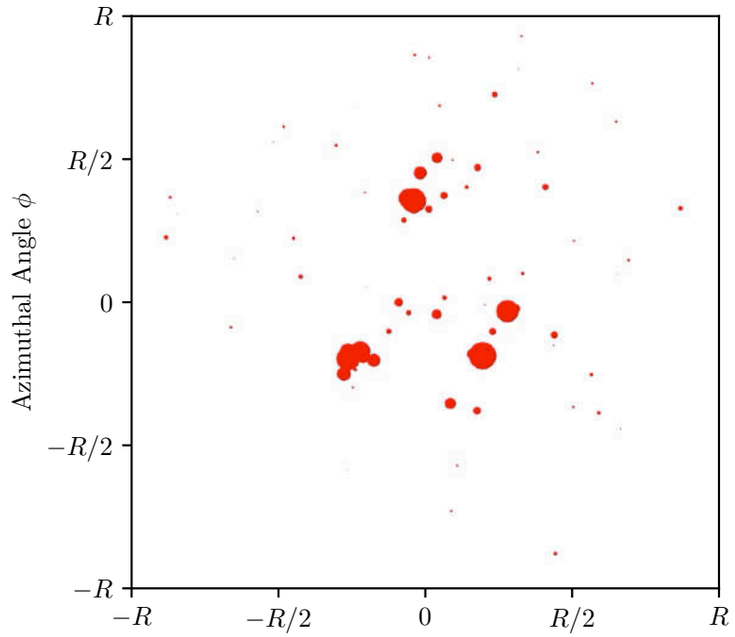
Safe to infrared & collinear splittings
No flavor/charge information
No pixelation needed

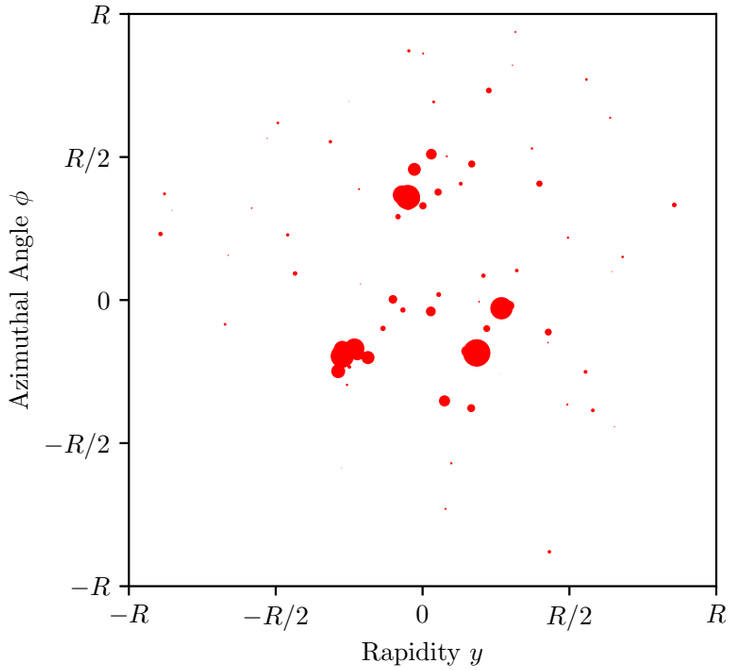
Stress-energy flow: *Measure of event/jet structure robust to non-perturbative and detector effects (i.e. IRC safe)*

[Sveshnikov, Tkachov, hep-ph/9512370; Hofman, Maldacena, 0803.1467; Mateu, Stewart, JDT, 1209.3781; Komiske, Metodiev, JDT, 1712.07124, 1810.05165]

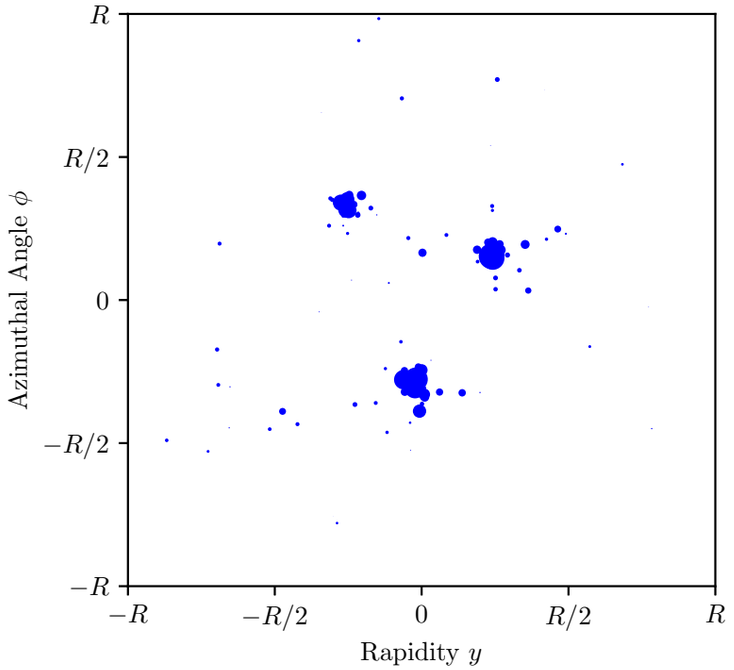
Similarity of Two Energy Flows?







125.4 GeV



[h/t @pkomiske & @EricMetodiev for animations]

The Earth Mover's Distance

Optimal Transport:

[Peleg, Werman, Rom, IEEE 1989;
Rubner, Tomasi, Guibas, ICCV 1998, ICJV 2000;
Pele, Werman, ECCV 2008; Pele Taskar, GSI 2013]

Minimum "work" (stuff x distance) to make one distribution …

The Earth Mover's Distance

Optimal Transport:

[Peleg, Werman, Rom, IEEE 1989;
Rubner, Tomasi, Guibas, ICCV 1998, IJCV 2000;
Pele, Werman, ECCV 2008; Pele Taskar, GSI 2013]

Minimum "work" (stuff x distance) to make one distribution look like another distribution

The Earth Mover's Distance

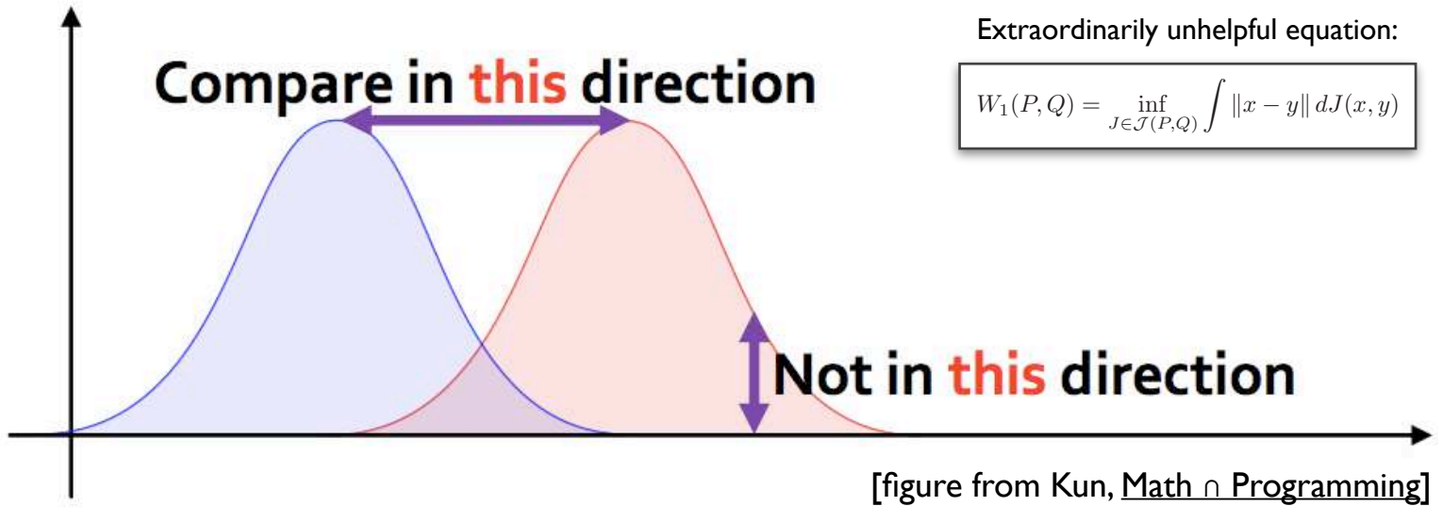
Optimal Transport:

[Peleg, Werman, Rom, IEEE 1989;
Rubner, Tomasi, Guibas, ICCV 1998, ICJV 2000;
Pele, Werman, ECCV 2008; Pele Taskar, GSI 2013]

Minimum "work" (stuff x distance) to make one distribution look like another distribution

Equivalent to 1-Wasserstein:

Metric on normalized distributions
(symmetric, non-negative,
triangle inequality, zero iff equal)



[figure from Kun, Math ∩ Programming]

Extraordinarily unhelpful equation:

$$W_1(P,Q) = \inf_{J \in \mathcal{J}(P,Q)} \int \|x - y\| \, dJ(x,y)$$

Very popular in ML applications:

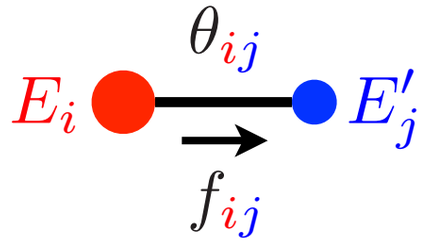
e.g. Wasserstein Generative Adversarial Networks

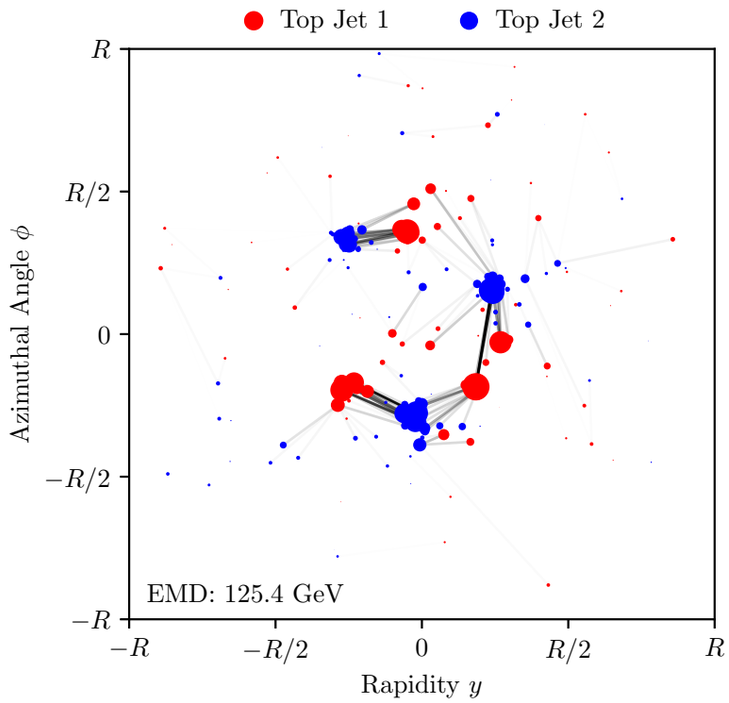
[Arjovsky, Chintala, Bottou, 1701.07875;
see also Erdmann, Geiger, Glombitza, Schmidt, 1802.03325; Erdmann, Glombitza, Quast, 1807.01954]

e.g. Wasserstein(-Wasserstein) Autoencoders

[Tolstikhin, Bousquet, Gelly, Schoelkopf, 1711.01558; Zhang, Gao, Jiao, Liu, Wang, Yang, 1902.09323]

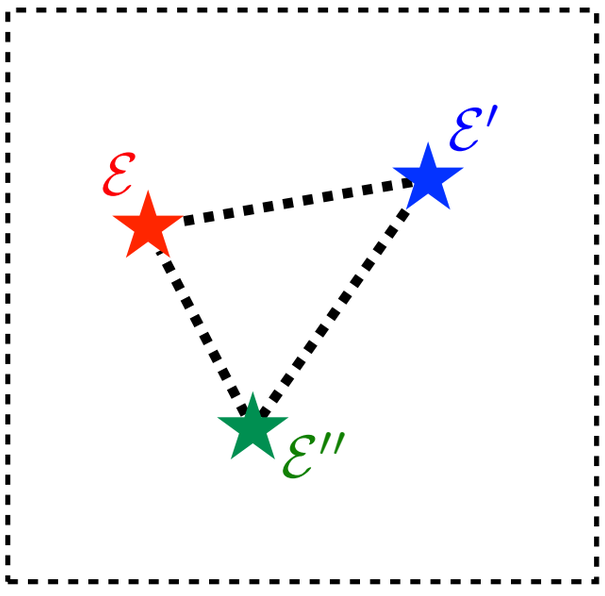
The Energy Mover's Distance





Optimal transport between energy flows…

$$\underbrace{\mathrm{EMD}(\mathcal{E},\mathcal{E}')}_{\text{in GeV}} = \min_{\{f\}} \underbrace{\sum_i \sum_j f_{ij} \frac{\theta_{ij}}{R}}_{\text{Cost to move energy}} + \underbrace{\left| \sum_i E_i - \sum_j E'_j \right|}_{\text{Cost to create energy}}$$

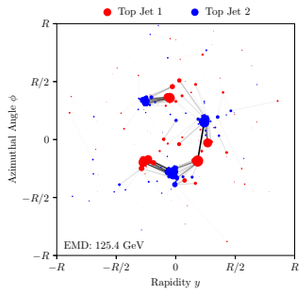


…defines a metric on the space of events

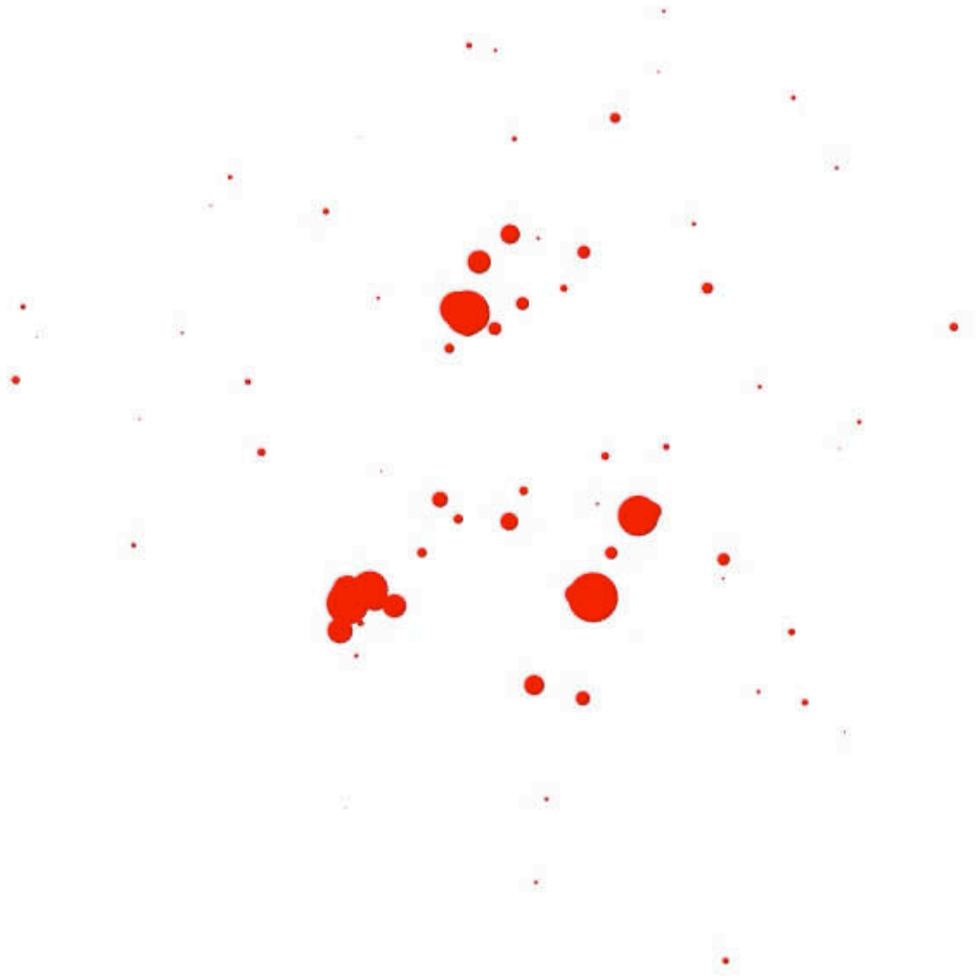
$$0 \leq \mathrm{EMD}(\mathcal{E},\mathcal{E}') \leq \mathrm{EMD}(\mathcal{E},\mathcal{E}'') + \mathrm{EMD}(\mathcal{E}',\mathcal{E}'')$$

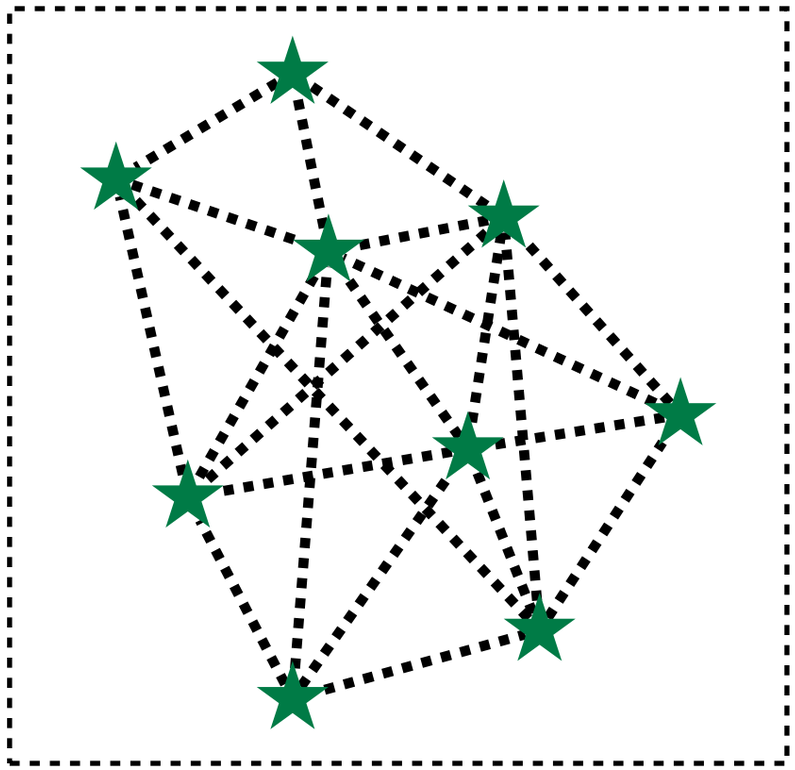
(assuming $R \geq \theta_{max}/2$, i.e. $R \geq$ jet radius for conical jets)

[Komiske, Metodiev, JDT, 1902.02346;
see also Pele, Werman, ECCV 2008; Pele Taskar, GSI 2013]



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

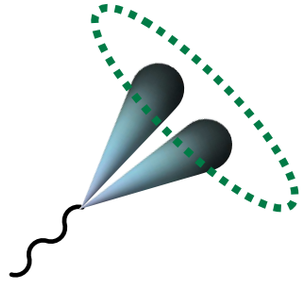




We defined a (metric) space for collider events

Now what?

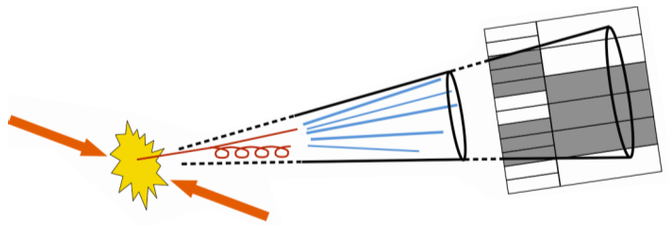
The Energy Mover's Distance



Initial Physics Studies

(Broader Comments)

Visualizing Jet Evolution



500 GeV

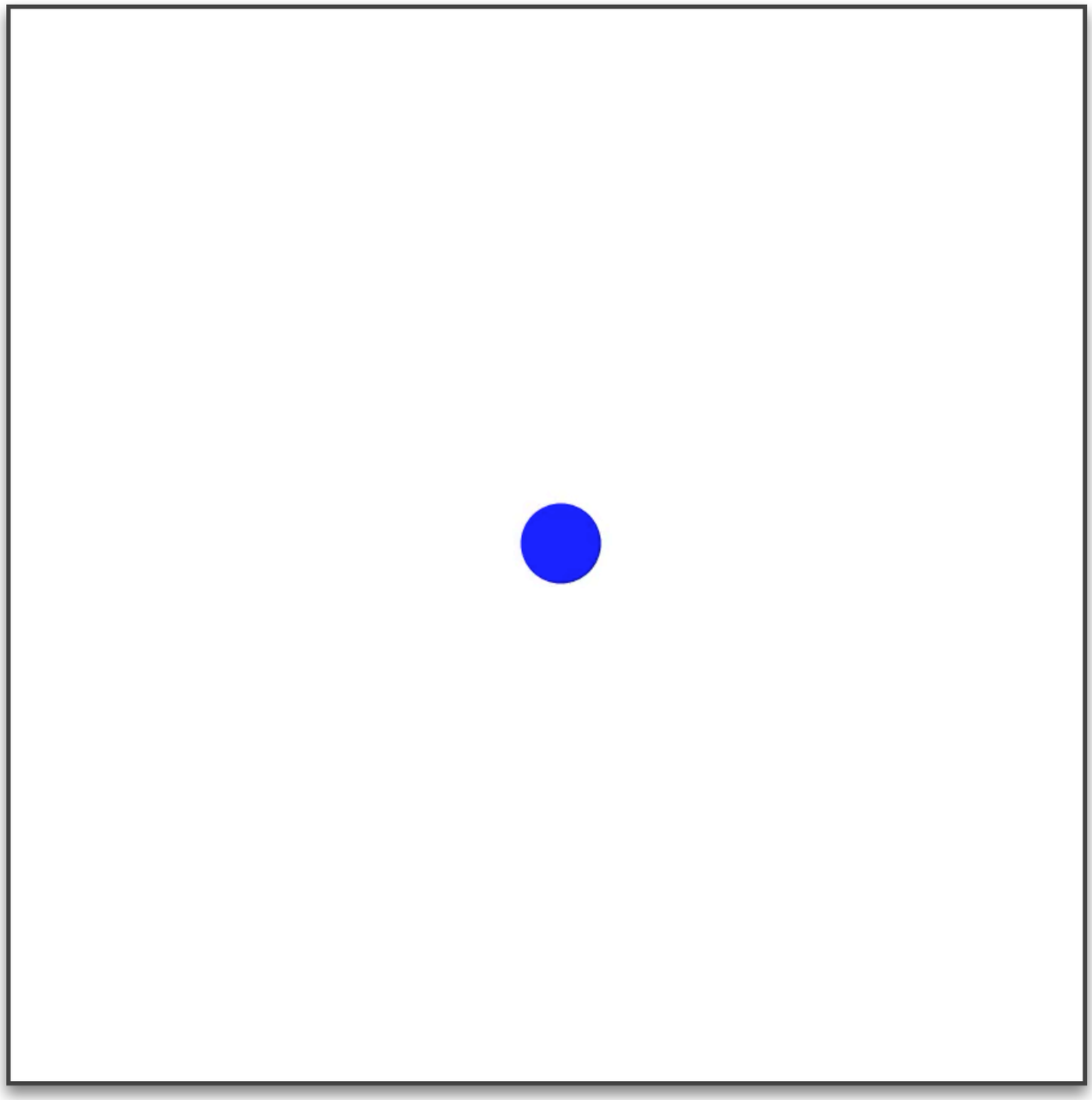
Hard Quark

Showering
EMD: 111.6 GeV

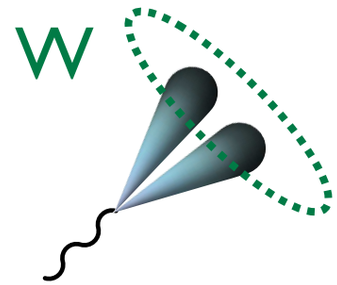
Partons

Hadronization
EMD: 18.1 GeV

Hadrons



Visualizing W Boson Evolution



500 GeV

W Boson

Decay
EMD: 78.3 GeV

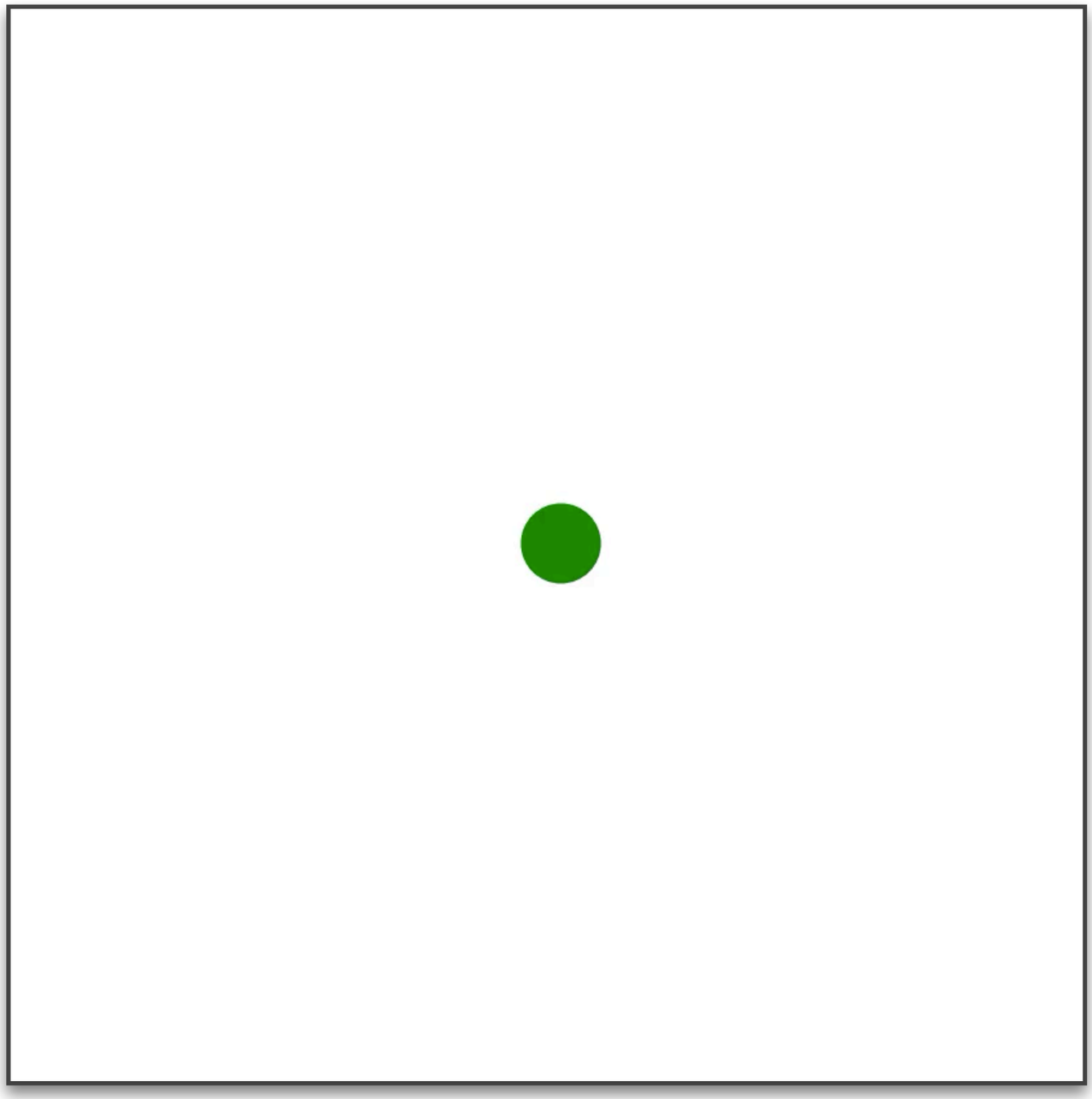
Two Quarks

Showering
EMD: 26.3 GeV

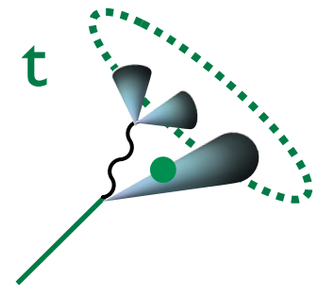
Partons

Hadronization
EMD: 12.9 GeV

Hadrons



Visualizing Top Quark Evolution



500 GeV

Top Quark

Decay
EMD: 161.1 GeV

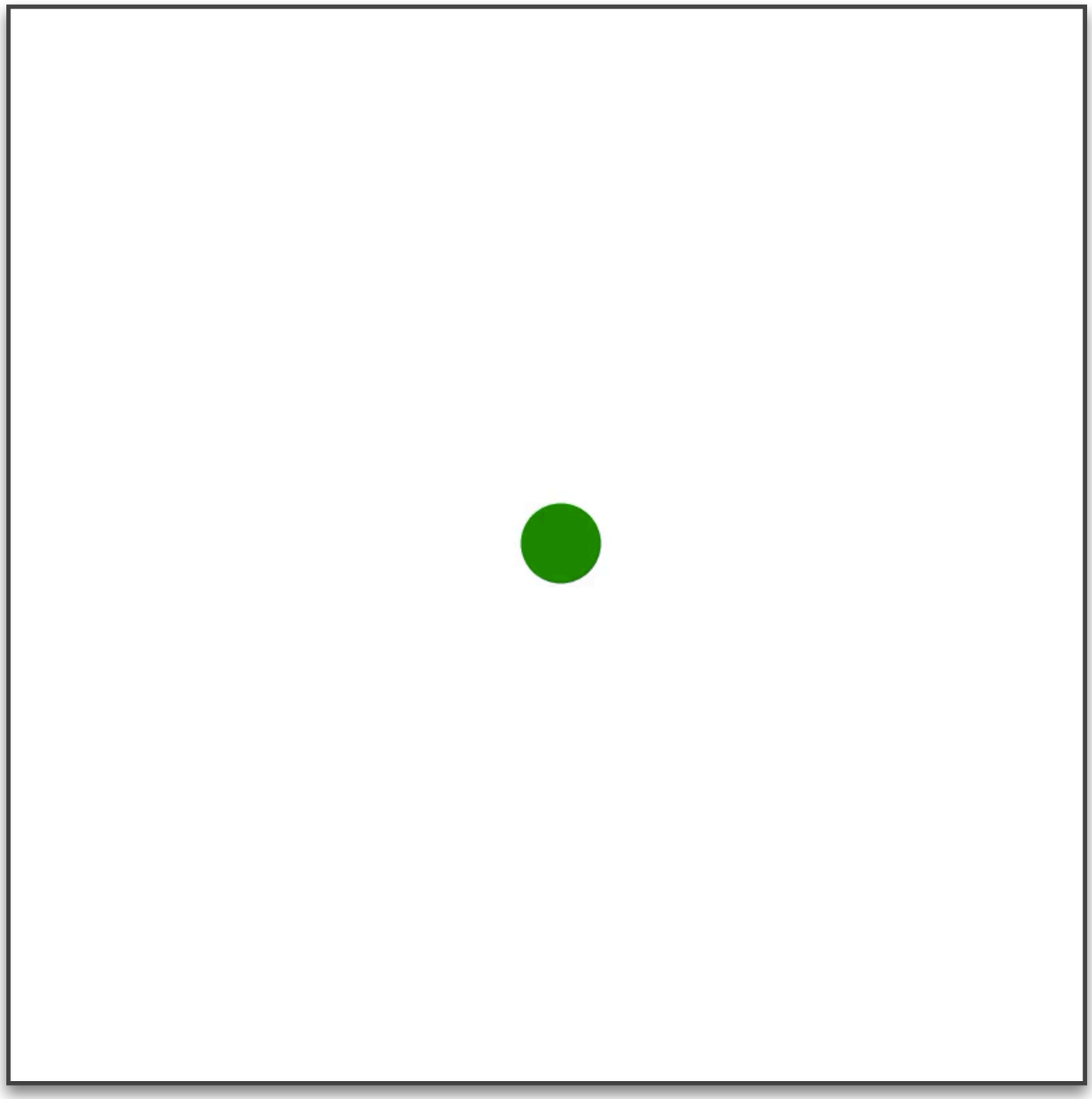
Three Quarks

Showering
EMD: 47.1 GeV

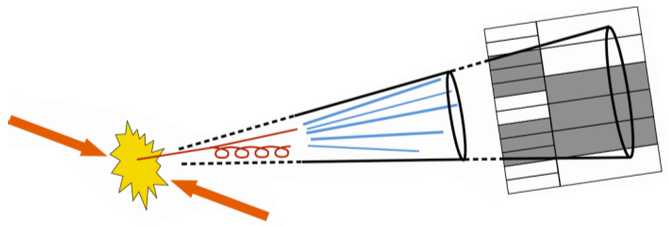
Partons

Hadronization
EMD: 27.0 GeV

Hadrons



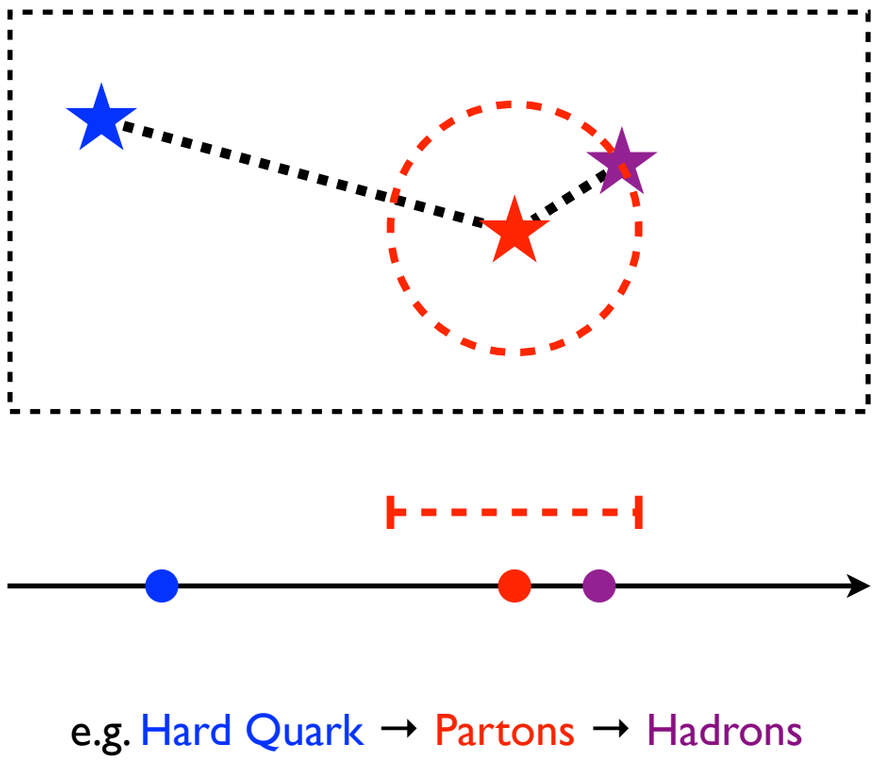
Quantifying Jet Evolution



Events close in EMD are close in any* IRC-safe observable

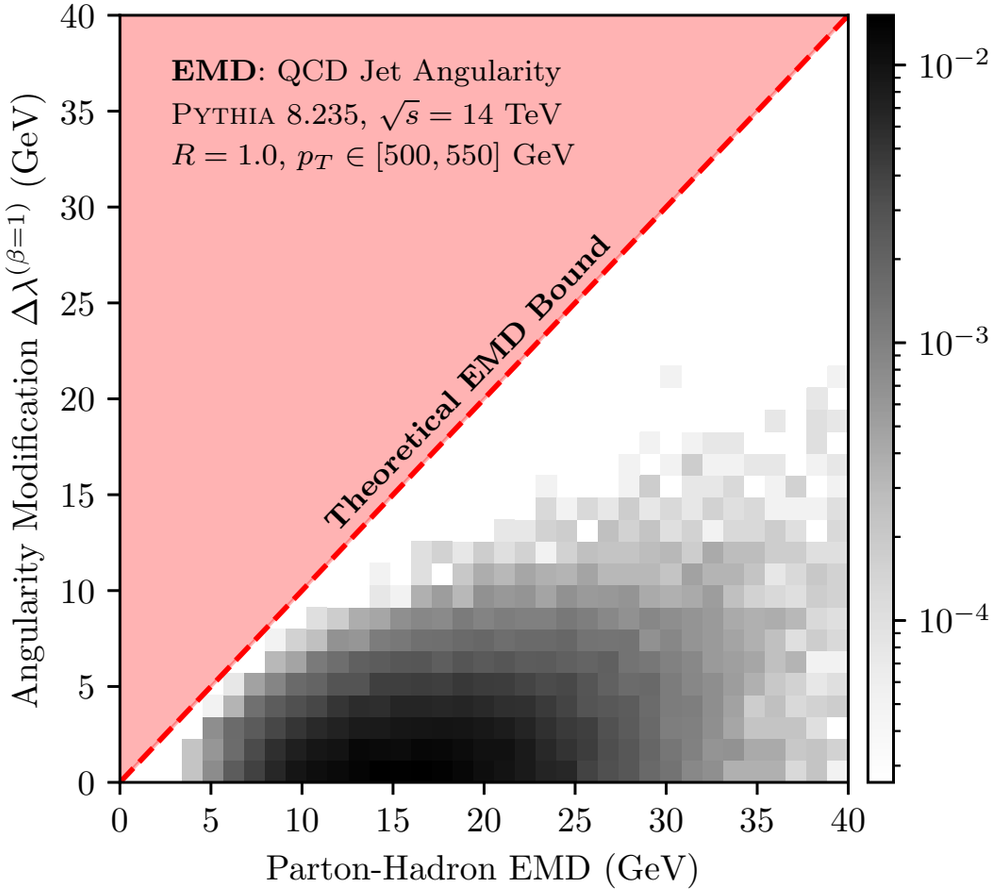
Using Kantorovich-Rubinstein duality

$$\left|\lambda^{(\beta)}(\mathcal{E}) - \lambda^{(\beta)}(\mathcal{E}')\right| \leq \beta\,\mathrm{EMD}(\mathcal{E},\mathcal{E}')$$



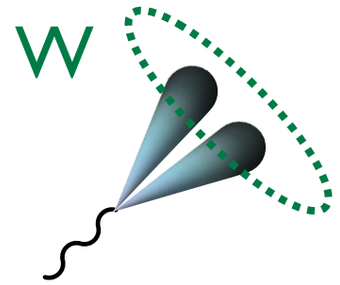
e.g. Hard Quark → Partons → Hadrons

e.g. Angularities (β≥1) $\lambda^{(\beta)}(\mathcal{E}) = \sum_i E_i \theta_i^\beta$

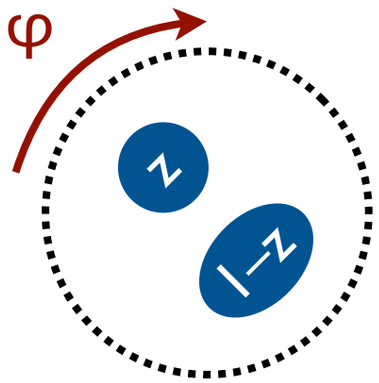


[Komiske, Metodiev, JDT, 1902.02346; using Berger, Kucs, Sterman, hep-ph/0303051; Ellis, Vermilion, Walsh, Hornig, Lee, 1001.0014; Larkoski, JDT, Waalewijn, 1408.3122]

The Space of Boosted W Bosons



Identify low-D manifold in high-D space?

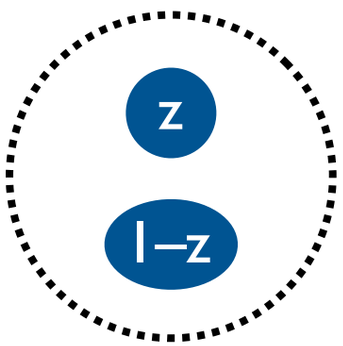


2-body kinematics

↓

2D

(fixed mass)



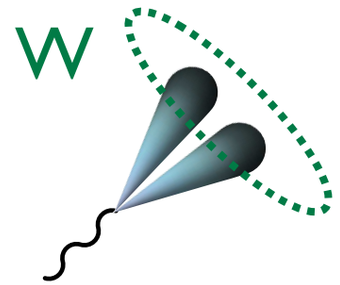
Rotate to φ = 0

↓

1D

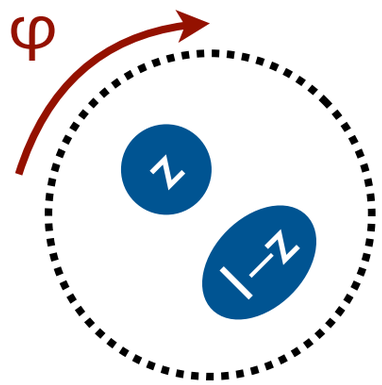
(up to fragmentation)

The Space of Boosted W Bosons

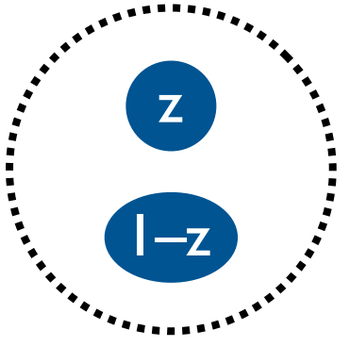


Identify low-D manifold in high-D space?

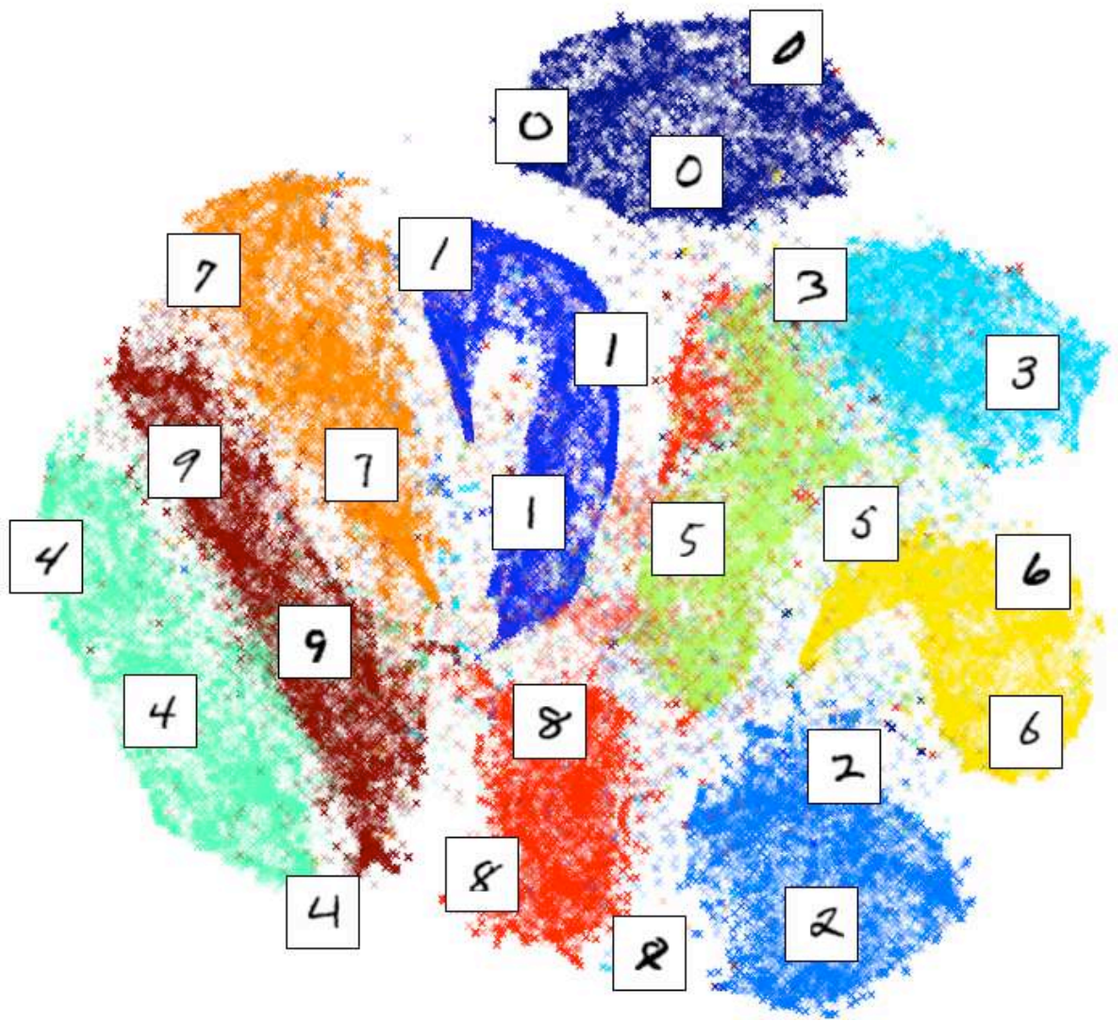
e.g. t-SNE
(t-distributed Stochastic Neighbor Embedding)



2-body kinematics
↓
2D
(fixed mass)

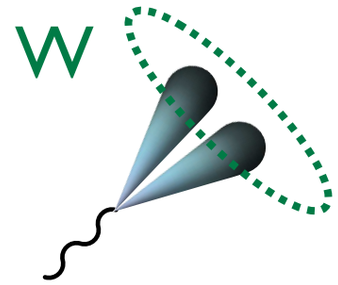


Rotate to φ = 0
↓
1D
(up to fragmentation)

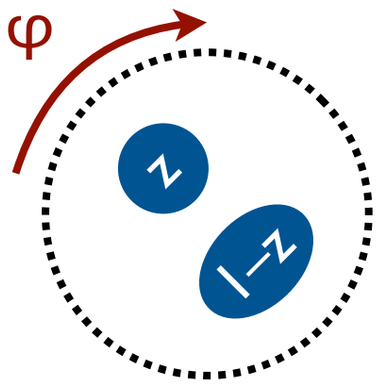


[van der Maaten, Hinton, JMLR 2008; figure from BigSnarf]

The Space of Boosted W Bosons



Identify low-D manifold in high-D space?

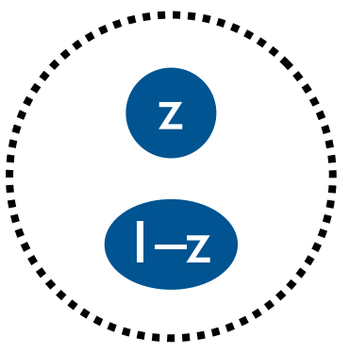


2-body kinematics

↓

2D

(fixed mass)



Rotate to φ = 0

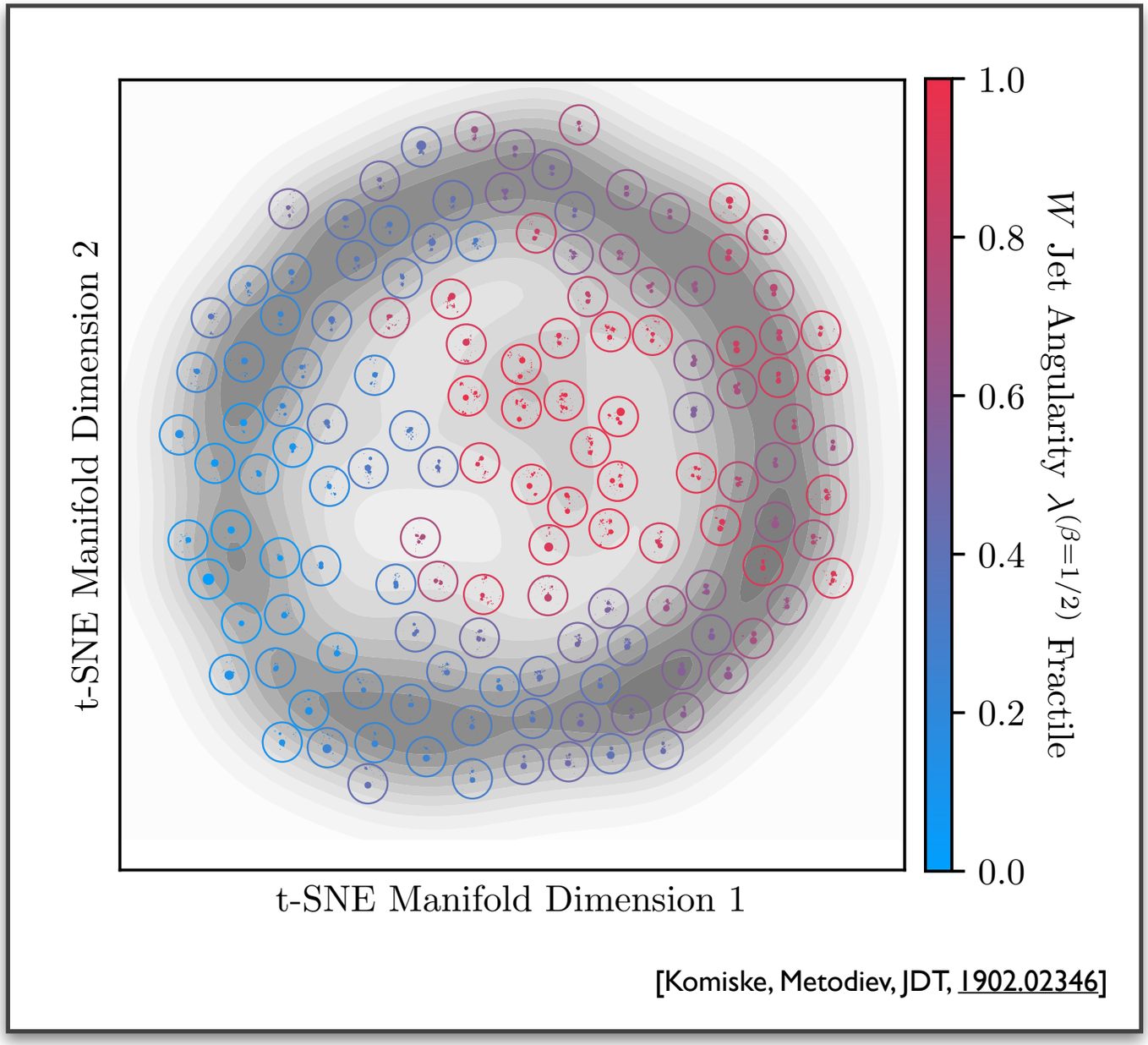
↓

1D

(up to fragmentation)

e.g. t-SNE

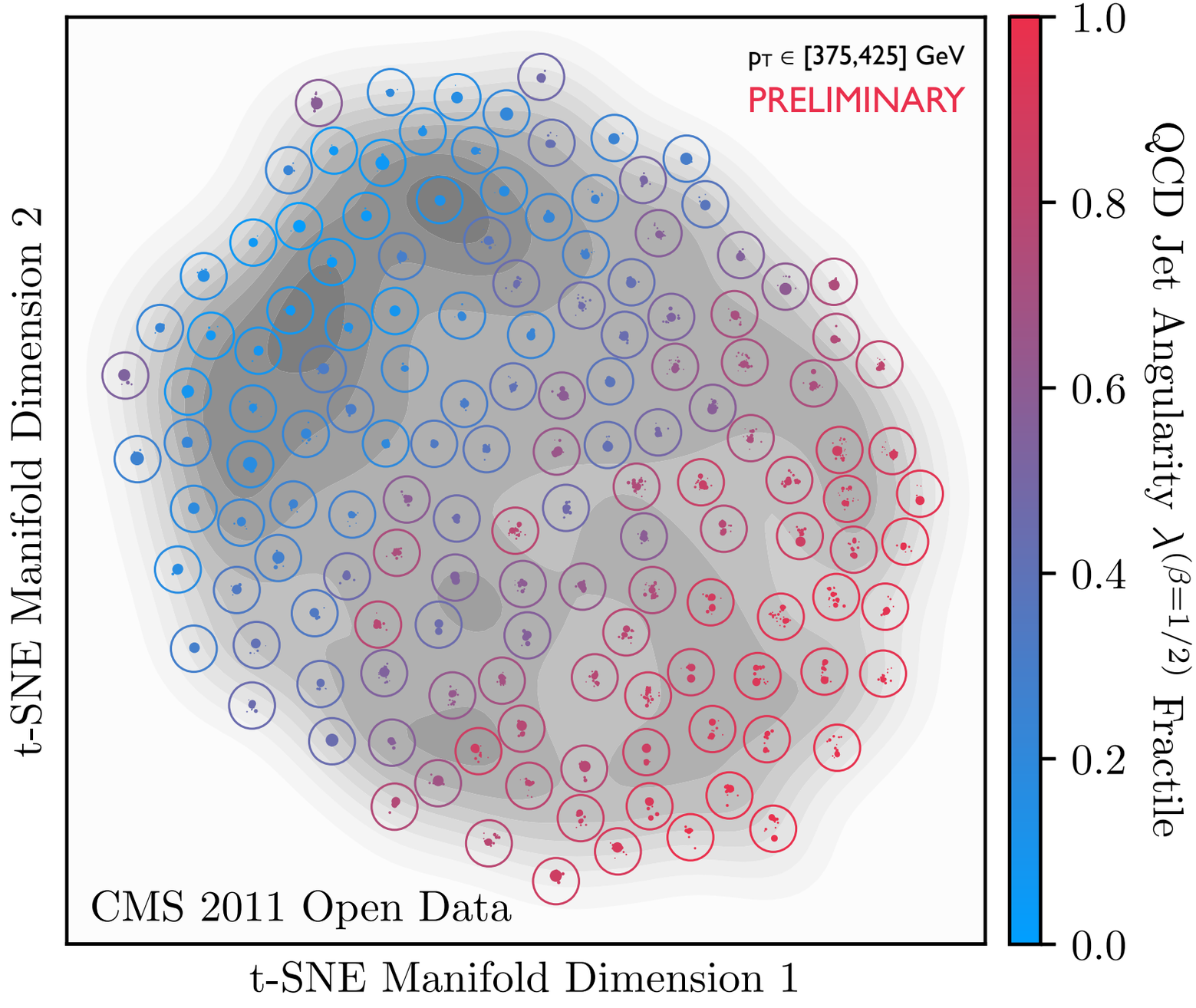
(t-distributed Stochastic Neighbor Embedding)

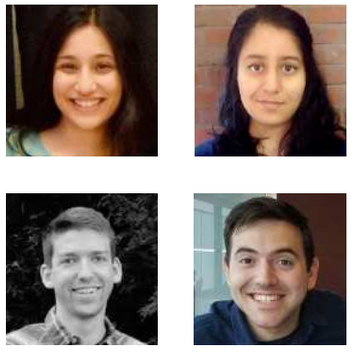


[Komiske, Metodiev, JDT, 1902.02346]

[van der Maaten, Hinton, JMLR 2008; figure from BigSnarf]

CMS Open Data: The Space of Jets



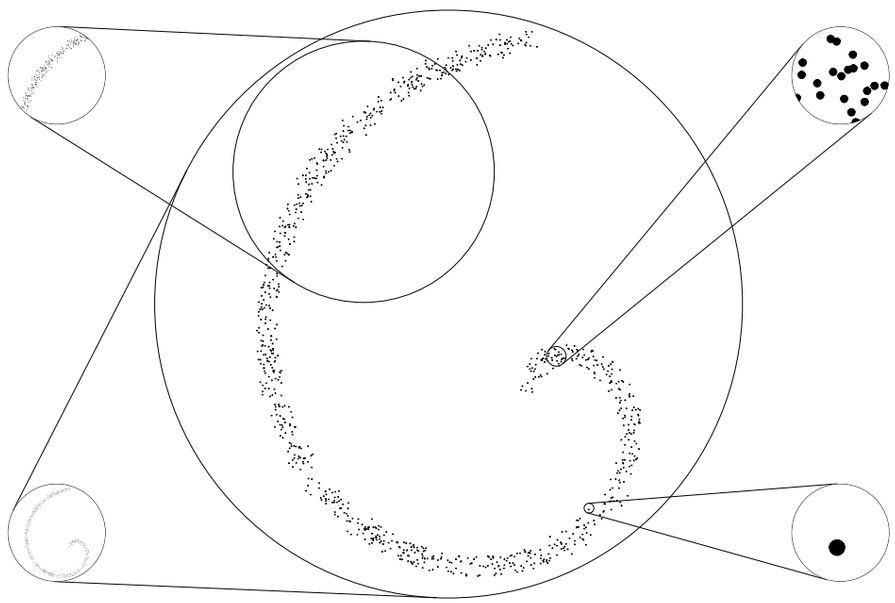


[Mastandrea, Naik, Komiske, Metodiev, JDT, in preparation]

Quantifying Dimensionality

Correlation Dimension: $\dim(Q) = Q\frac{\partial}{\partial Q}\ln\sum_i\sum_j\Theta\big(\mathrm{EMD}(\mathcal{E}_i,\mathcal{E}_j)<Q\big)$

dim≃1



dim≃2

(eventually 0)

dim≃2

$$N_{\text{neighbors}}(r) \sim r^{\dim}$$

$$\Downarrow$$

$$\dim(r) \sim r\frac{\partial}{\partial r}\ln N_{\text{neighbors}}(r)$$

dim≃0

[Grassberger, Procaccia, PRL 1983; Kégl, NIPS 2002]

Decay-Level Dimension

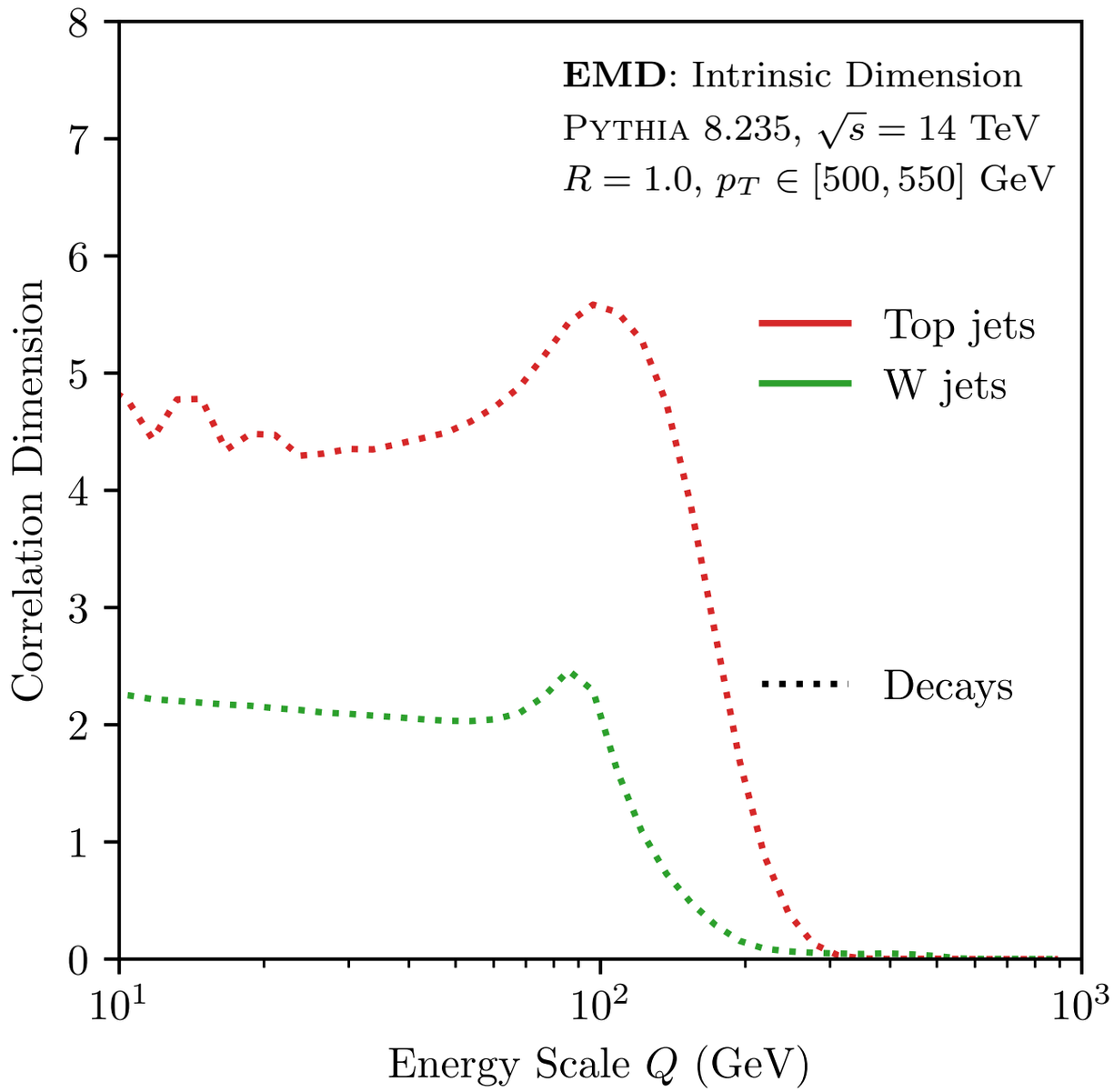
$$\dim(Q) = Q\frac{\partial}{\partial Q}\ln\sum_i\sum_j\Theta\big(\mathrm{EMD}(\mathcal{E}_i,\mathcal{E}_j) < Q\big)$$

3-body phase space

with W mass constraint

2-body phase space

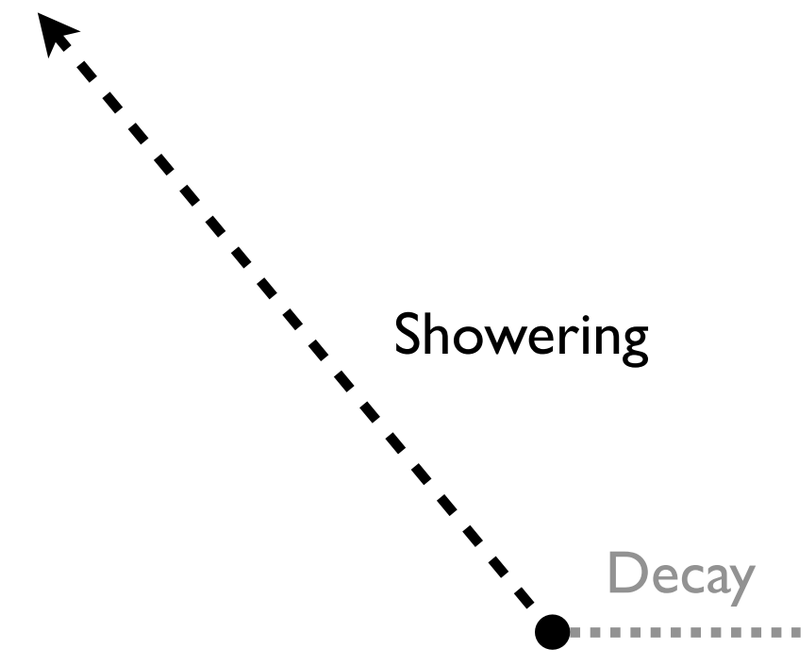
Increasing complexity: multi-body phase space

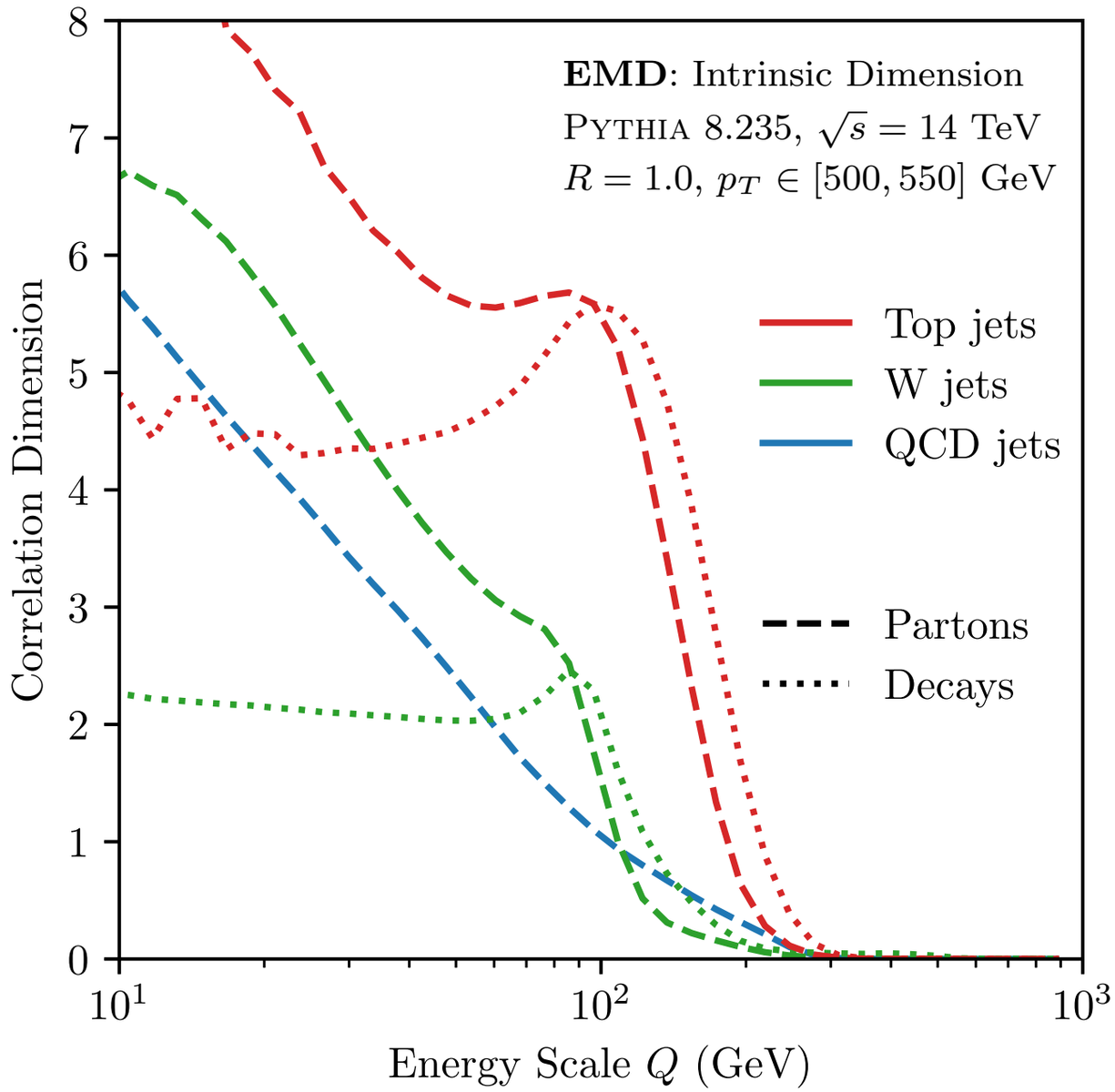


[Komiske, Metodiev, JDT, 1902.02346]

Parton-Level Dimension

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





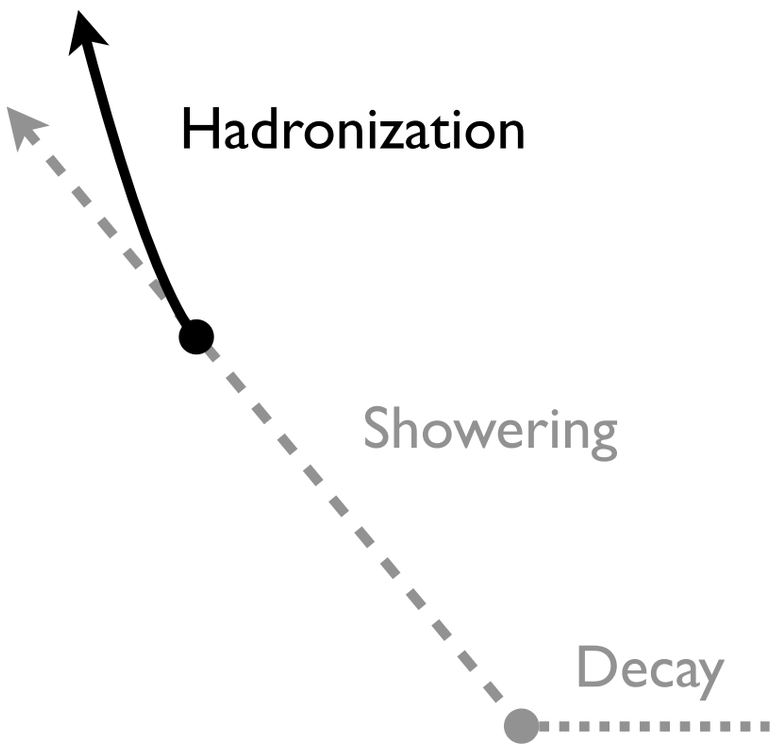
Increasing complexity: multi-body phase space

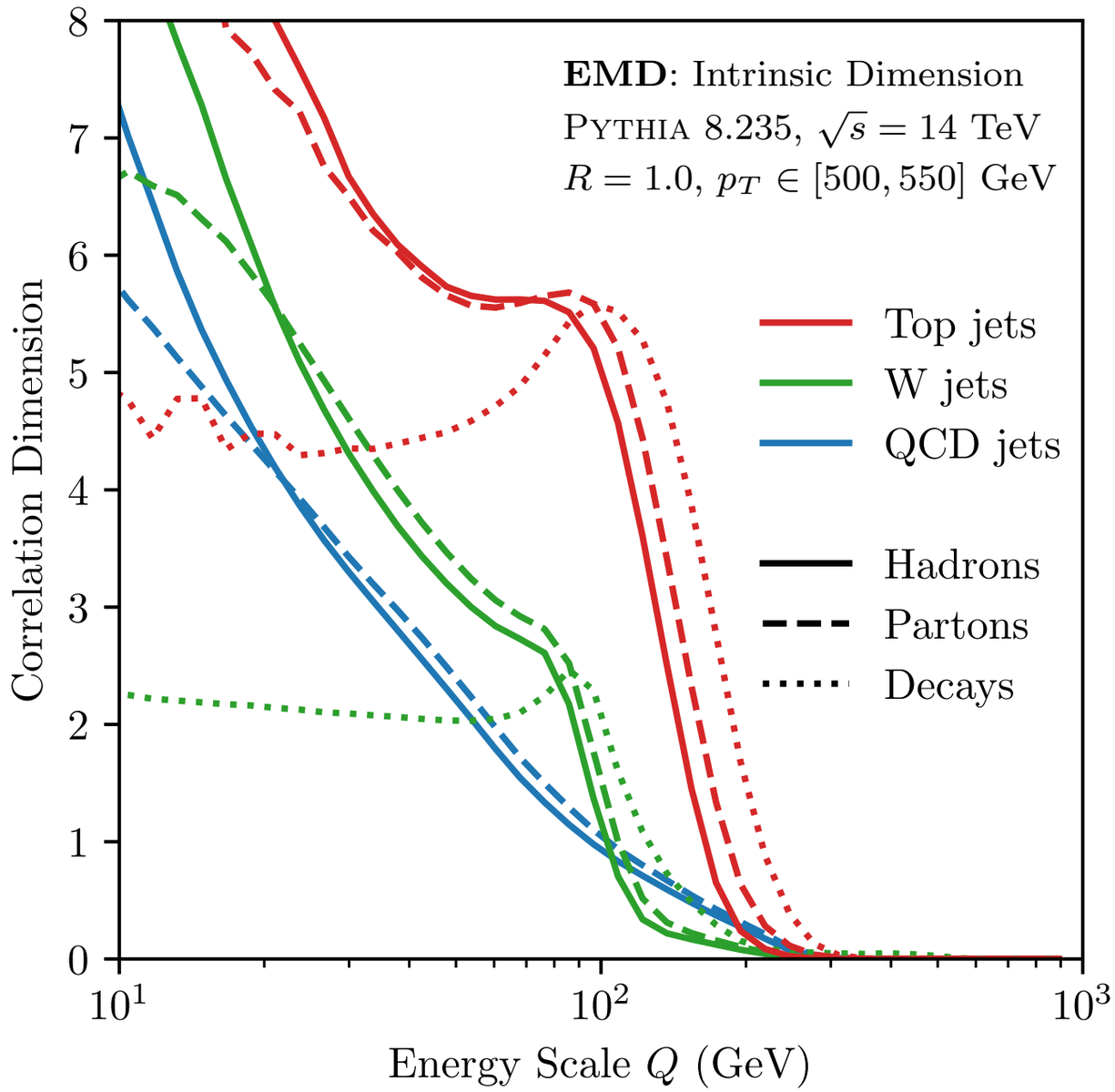
perturbative emissions

[Komiske, Metodiev, JDT, 1902.02346]

Hadron-Level Dimension

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





Increasing complexity: multi-body phase space

perturbative emissions

non-perturbative dynamics

[Komiske, Metodiev, JDT, 1902.02346]

Preliminary Calculation

Leading Log:

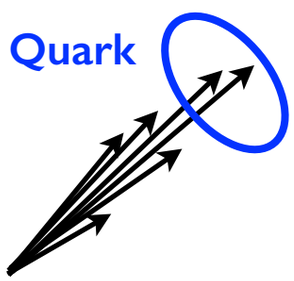
(single log, since dim has derivative)

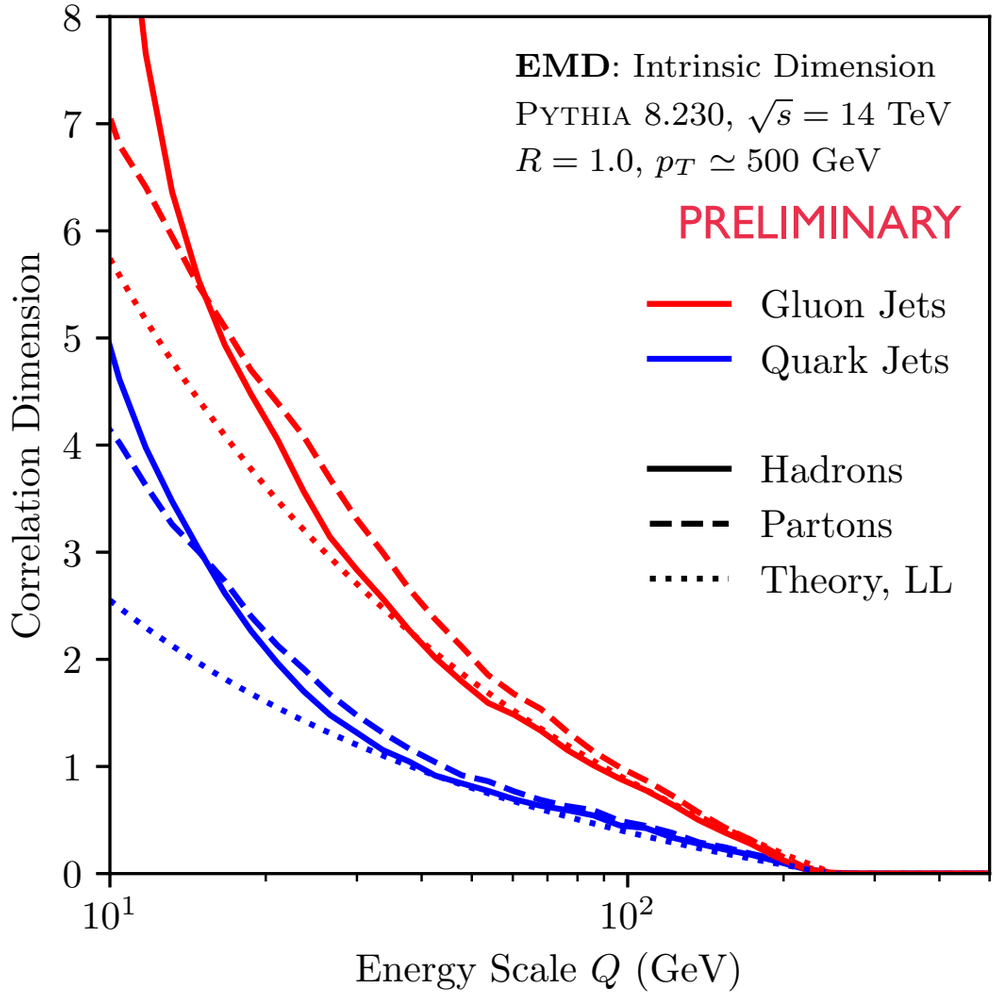
$$\dim_i(Q) \simeq -\frac{8\alpha_s}{\pi} C_i \ln \frac{Q}{p_T}$$

Color Factor

$C_A = 3$

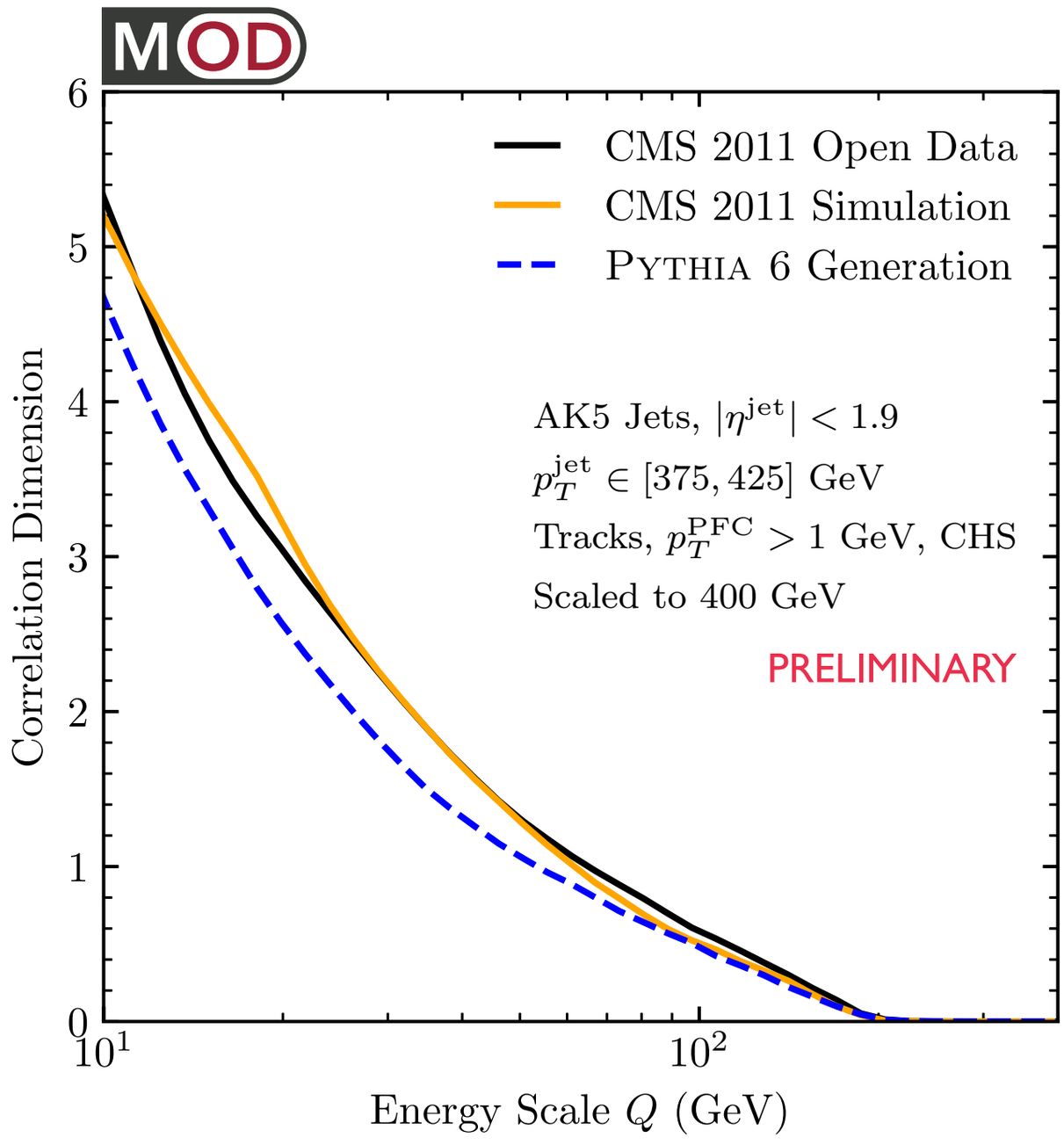
$C_F = 4/3$





CMS Open Data: Dimension of Jets

Fascinating impact of detector effects

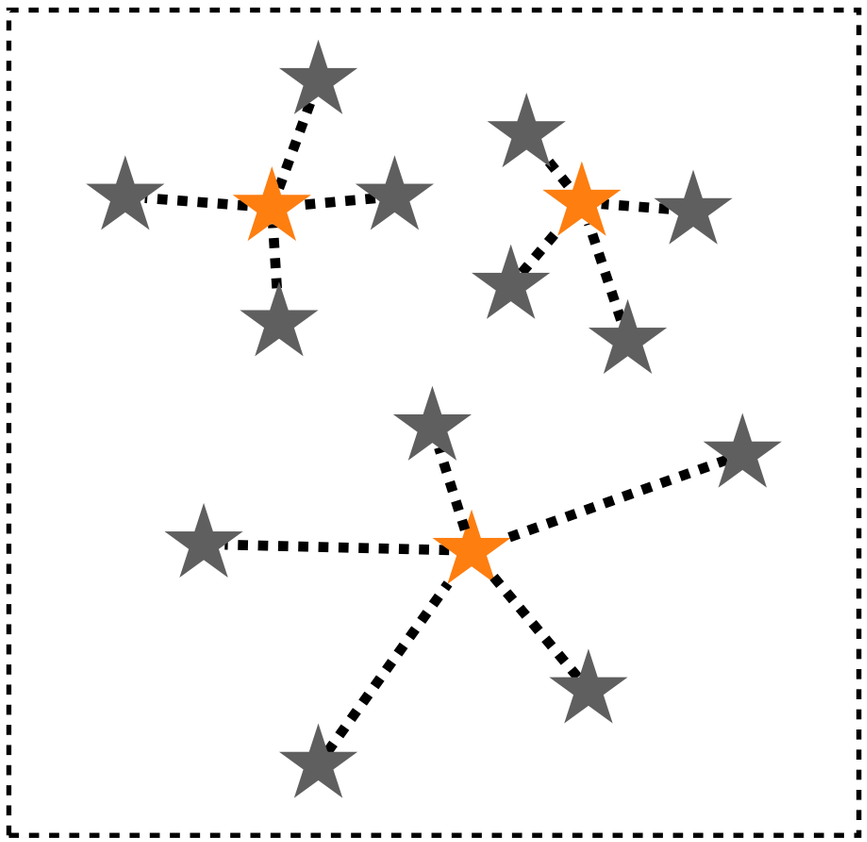


[Mastandrea, Naik, Komiske, Metodiev, JDT, in preparation]

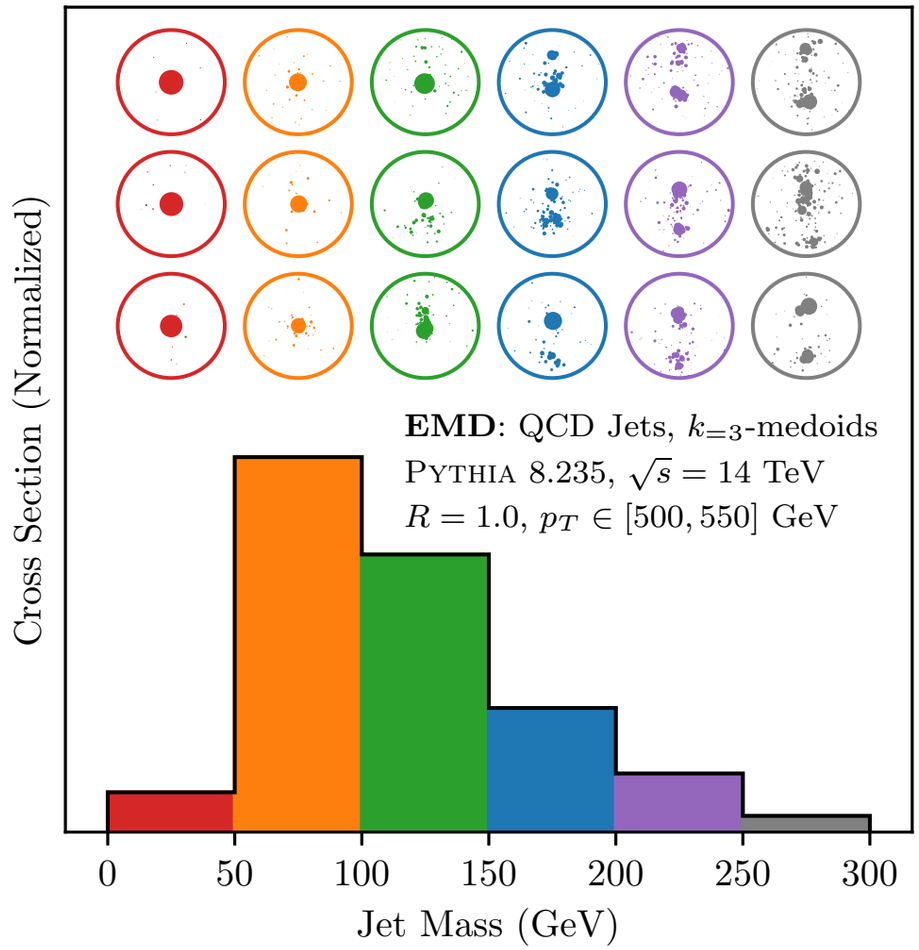
Identifying Representative Jets

medoid: jet that minimizes total EMD to other jets

k-medoids: k clusters to minimize total medoid distance

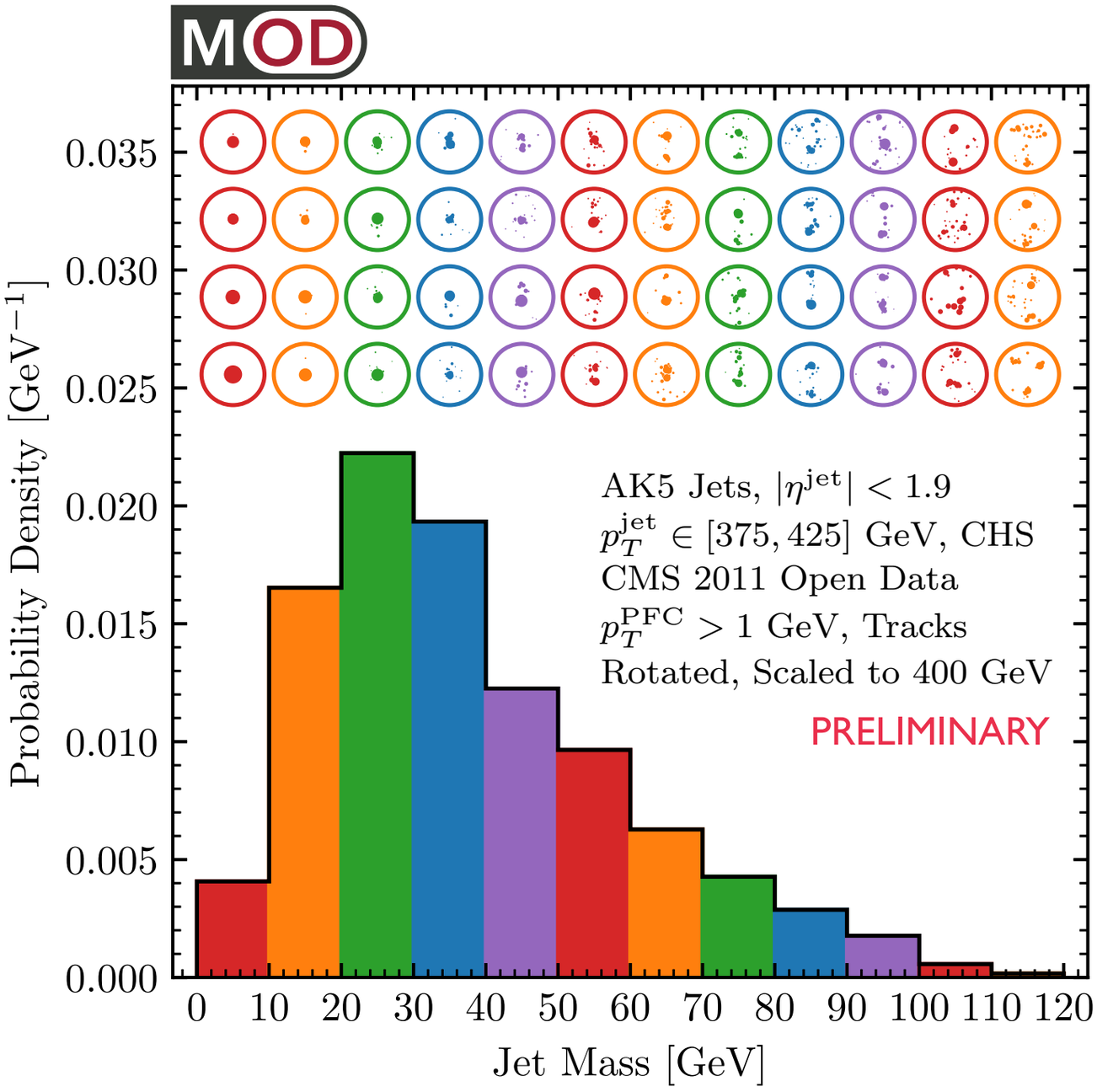


3-medoid



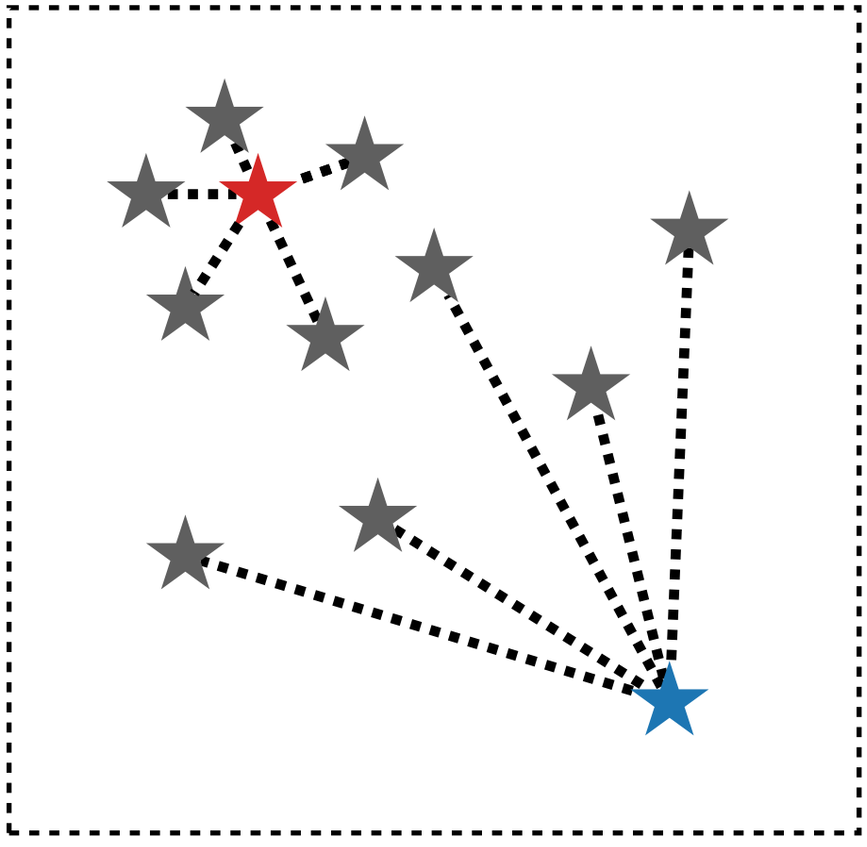
[Komiske, Metodiev, JDT, 1902.02346]

CMS Open Data: Jet Mass

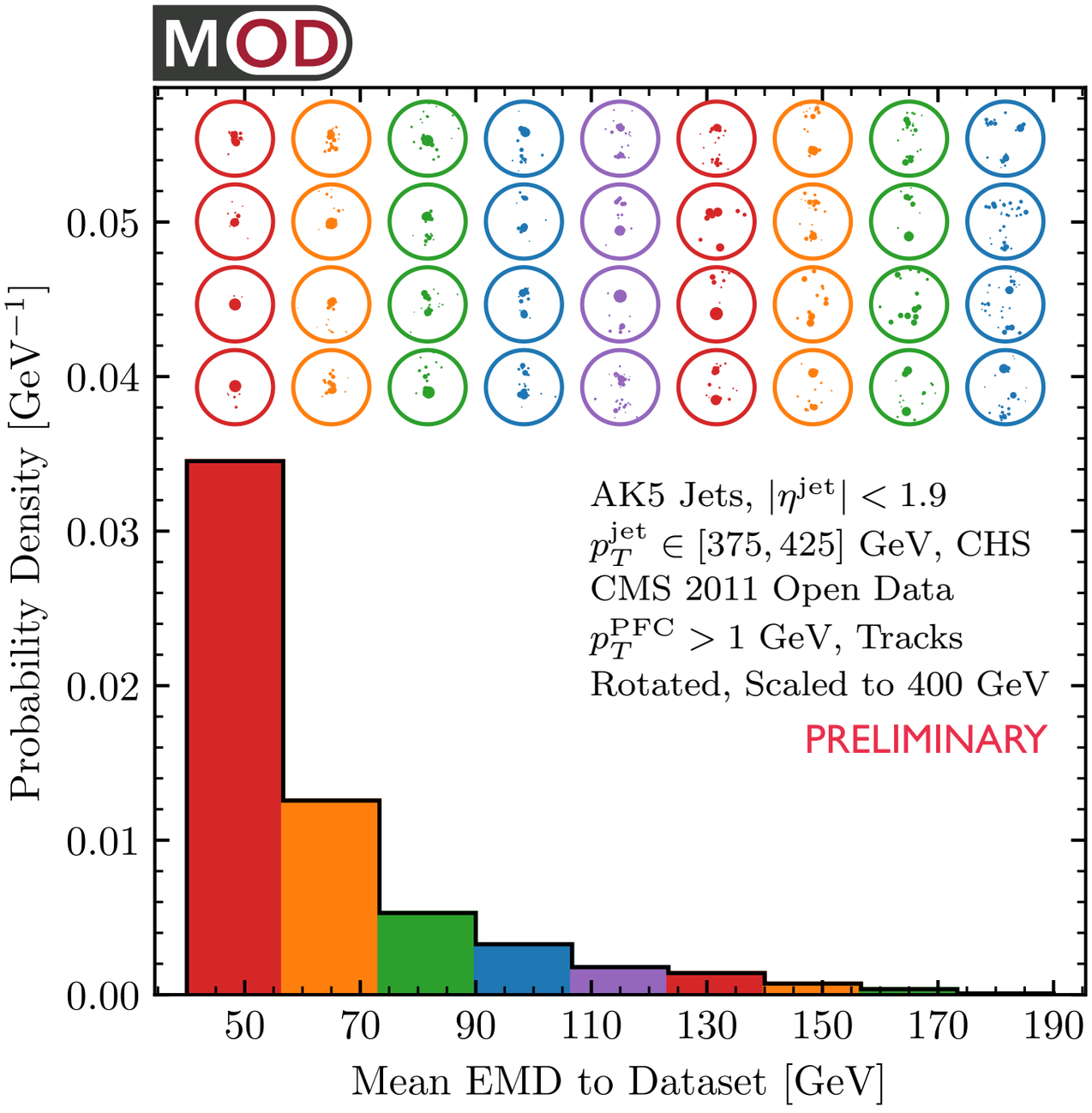


[Mastandrea, Naik, Komiske, Metodiev, JDT, in preparation]

CMS Open Data: Mean EMD



A first step towards
EMD-based anomaly detection

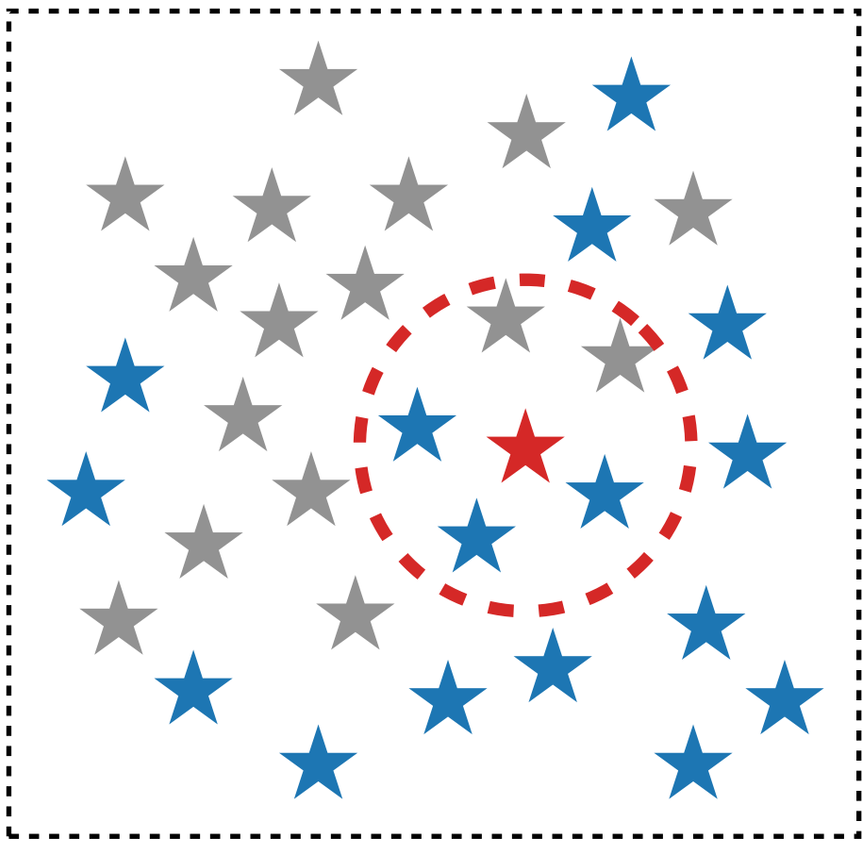


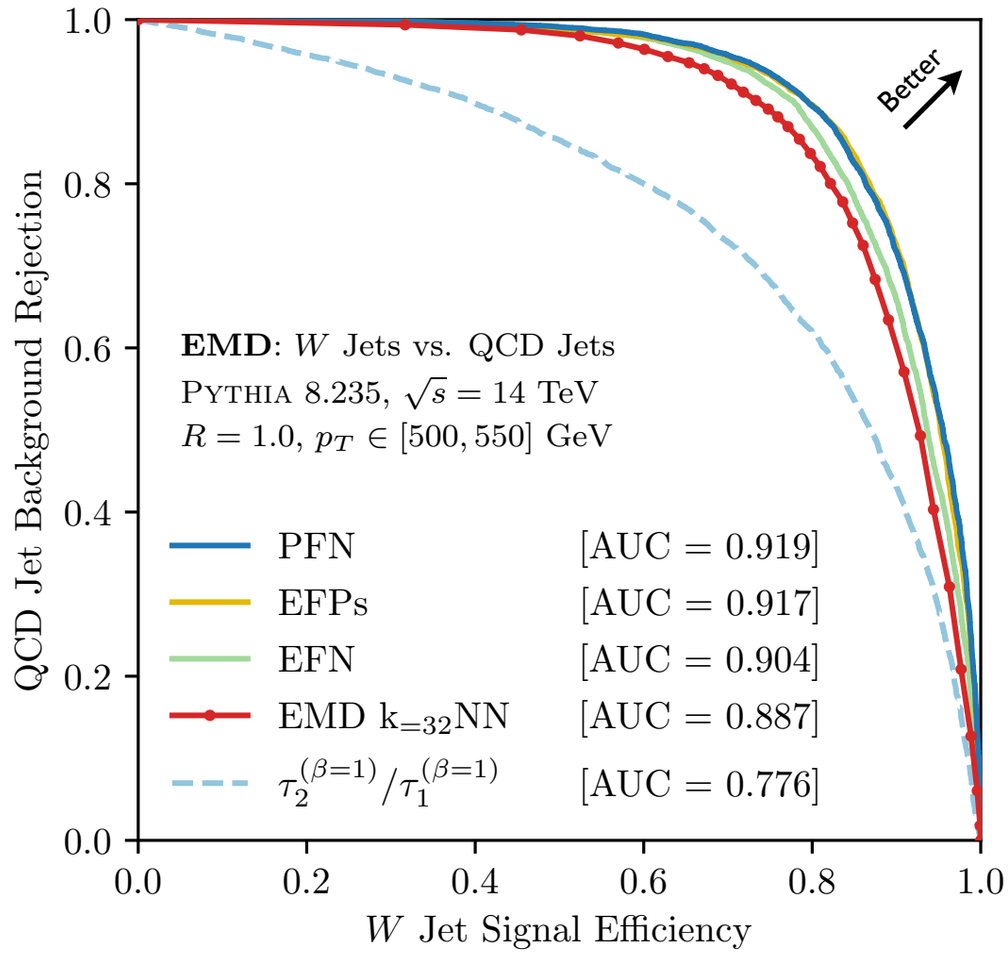
[Mastandrea, Naik, Komiske, Metodiev, JDT, in preparation]

Jet Classification

Estimate jet label by k nearest neighbors in training data

Approaches performance of modern machine learning





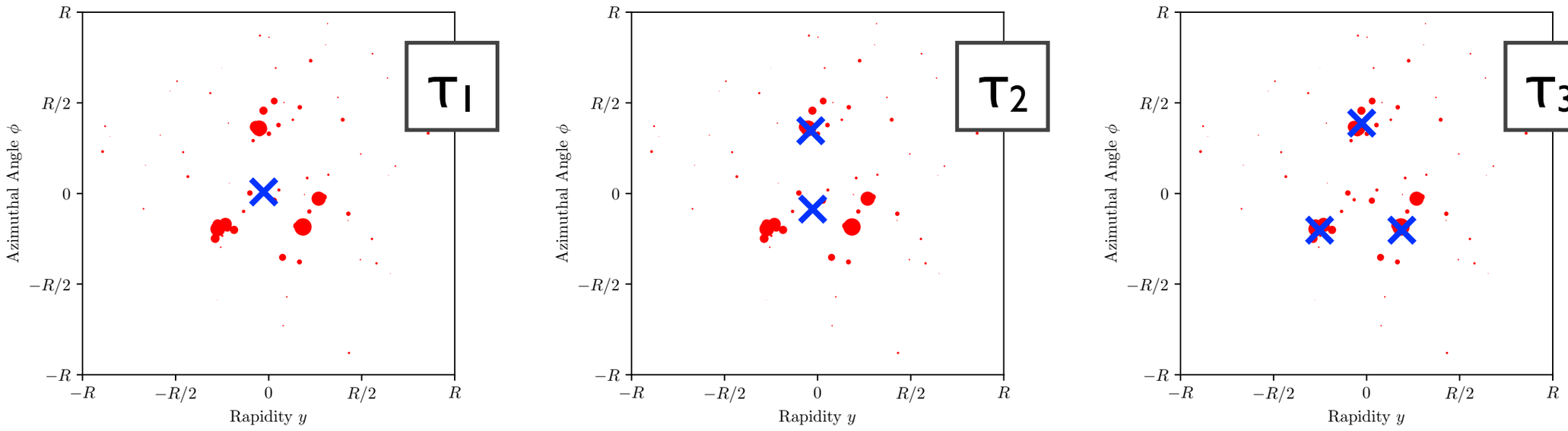
[Komiske, Metodiev, JDT, 1902.02346;
comparison to JDT, Van Tilburg, 1011.2268, 1108.2701; Komiske, Metodiev, JDT, 1712.07124, 1810.05165]

Insight into N-subjettiness

$$\tau_N^{(\beta)}(\mathcal{E}) = \min_{N \text{ axes}} \sum_i E_i \min \left\{ \theta_{1,i}^{\beta}, \theta_{2,i}^{\beta}, \ldots, \theta_{N,i}^{\beta} \right\}$$

IRC safe

kind of arbitrary



$$\tau_N(\mathcal{E}) = \min_{|\mathcal{E}'|=N} \text{EMD}(\mathcal{E}, \mathcal{E}')$$

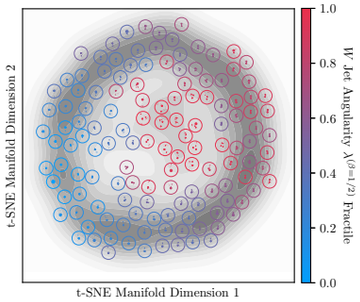
for β = 1

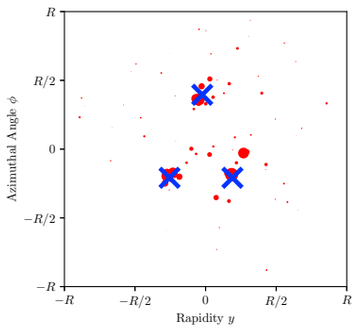
Related to p-Wasserstein metric for p = β > 1

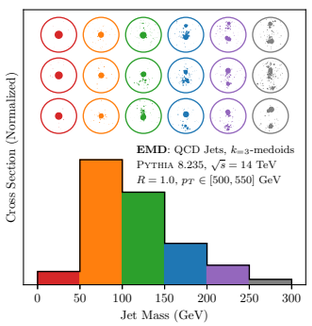
very satisfying

[JDT, Van Tilburg, 1011.2268, 1108.2701; based on Brandt, Dahmen, ZPC 1979; Stewart, Tackmann, Waalewijn, 1004.2489]

Future Directions







Quantify (and mitigate?) pileup/detector effects

Non-parametric density estimates (& unfolding?)

Automated data compression (& triggering?)

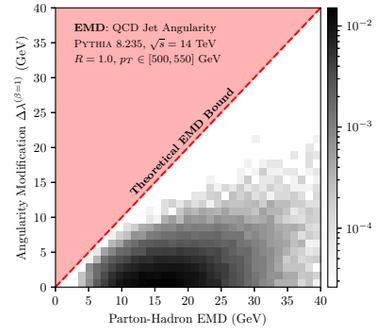
Define new observables through EMD geometry

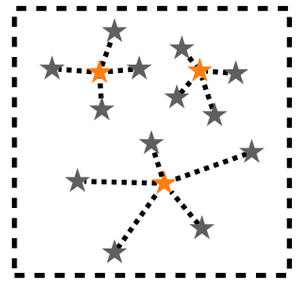
Precision QCD calculations of pairwise metrics

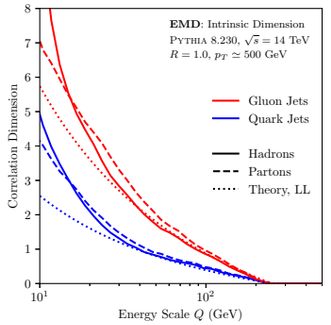
Novel analyses with and beyond histograms

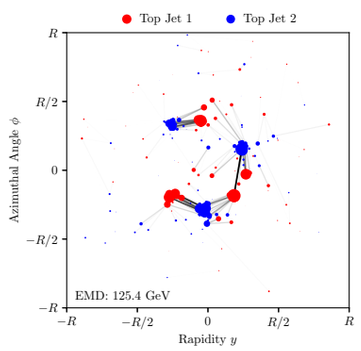
Event Mover's Distance between ensembles

…



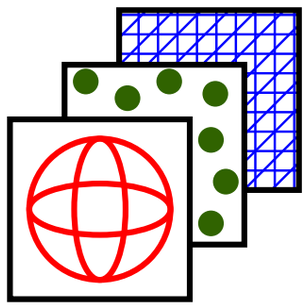






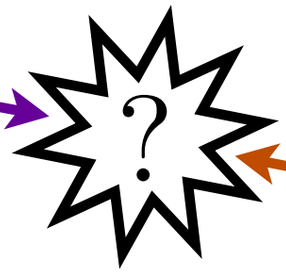
The Energy Mover's Distance

Initial Physics Studies



(Broader Comments)

High-Energy
Particle
Physics



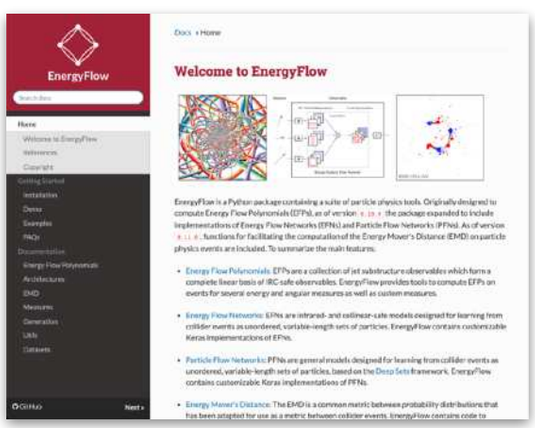
Mathematics,
Statistics,
Computer Science

Patrick Komiske

Eric Metodiev

High-Energy Particle Physics

Mathematics, Statistics, Computer Science

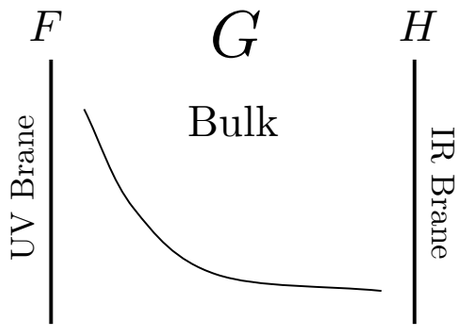


Energy Flow Package

https://energyflow.network/

Evolution of a "Model Builder"

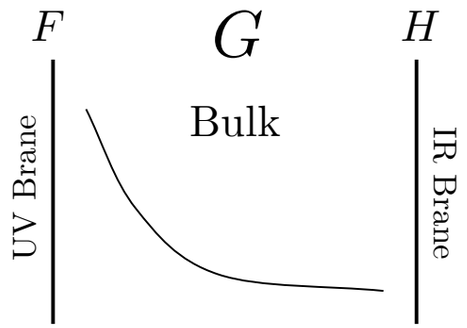
$\mathcal{L}$

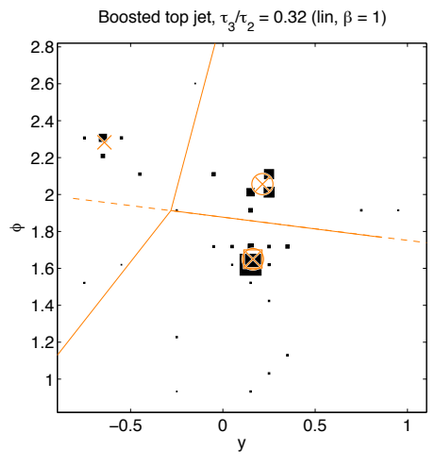


Build Lagrangians & Paradigms

[images from JDT, hep-ph/0502175; JDT, Van Tilburg, 1108.2701; Komiske, Metodiev, JDT, 1810.05165]

Evolution of a "Model Builder"



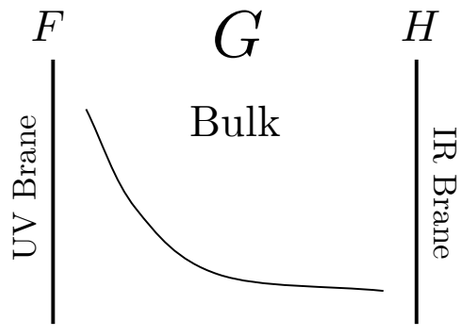


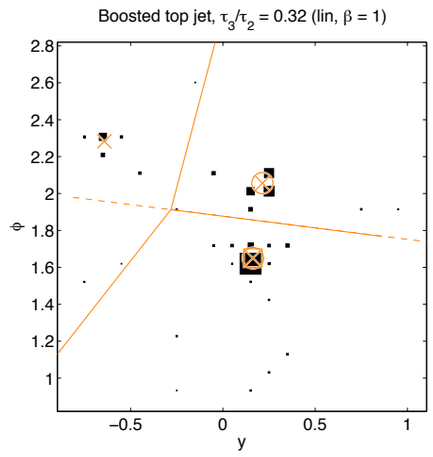
Build Lagrangians & Paradigms

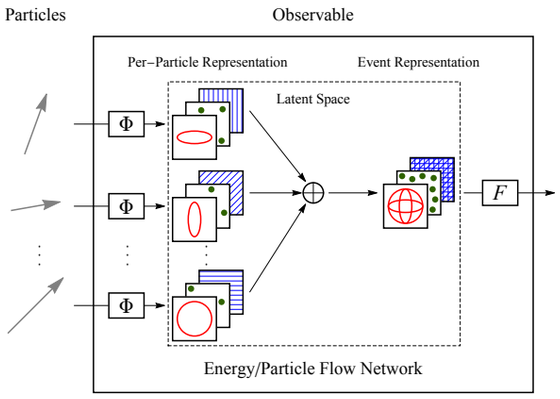
Build Observables & Algorithms

[images from JDT, hep-ph/0502175; JDT, Van Tilburg, 1108.2701; Komiske, Metodiev, JDT, 1810.05165]

Evolution of a "Model Builder"







Build Lagrangians & Paradigms

Build Observables & Algorithms

Build Models & Loss Functions

[images from JDT, hep-ph/0502175; JDT, Van Tilburg, 1108.2701; Komiske, Metodiev, JDT, 1810.05165]

Evolution of a "Model Builder"

Given current status of the LHC, which strategy:

Makes maximal verifiable use of collider data?

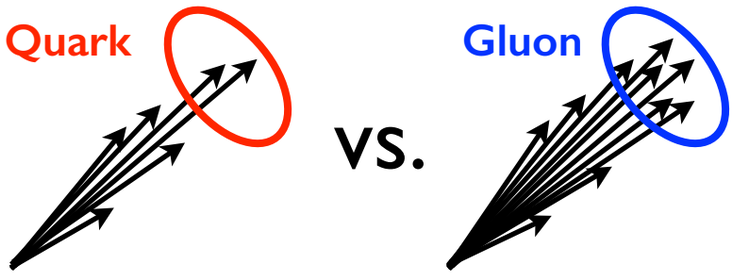
Can scale up to the challenges of HL-LHC?

Has the greatest discovery potential?

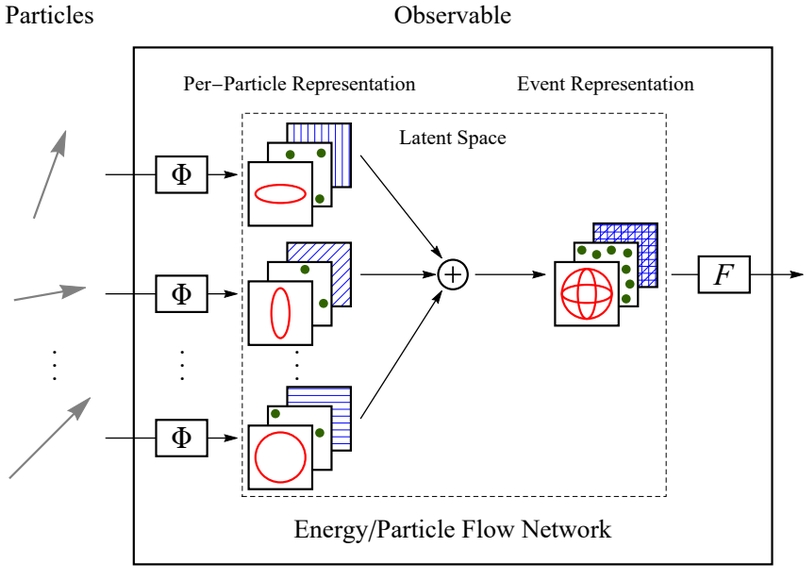
Offers new insights into fundamental physics?

[images from JDT, hep-ph/0502175; JDT, Van Tilburg, 1108.2701; Komiske, Metodiev, JDT, 1810.05165]

Opportunities for Visualization

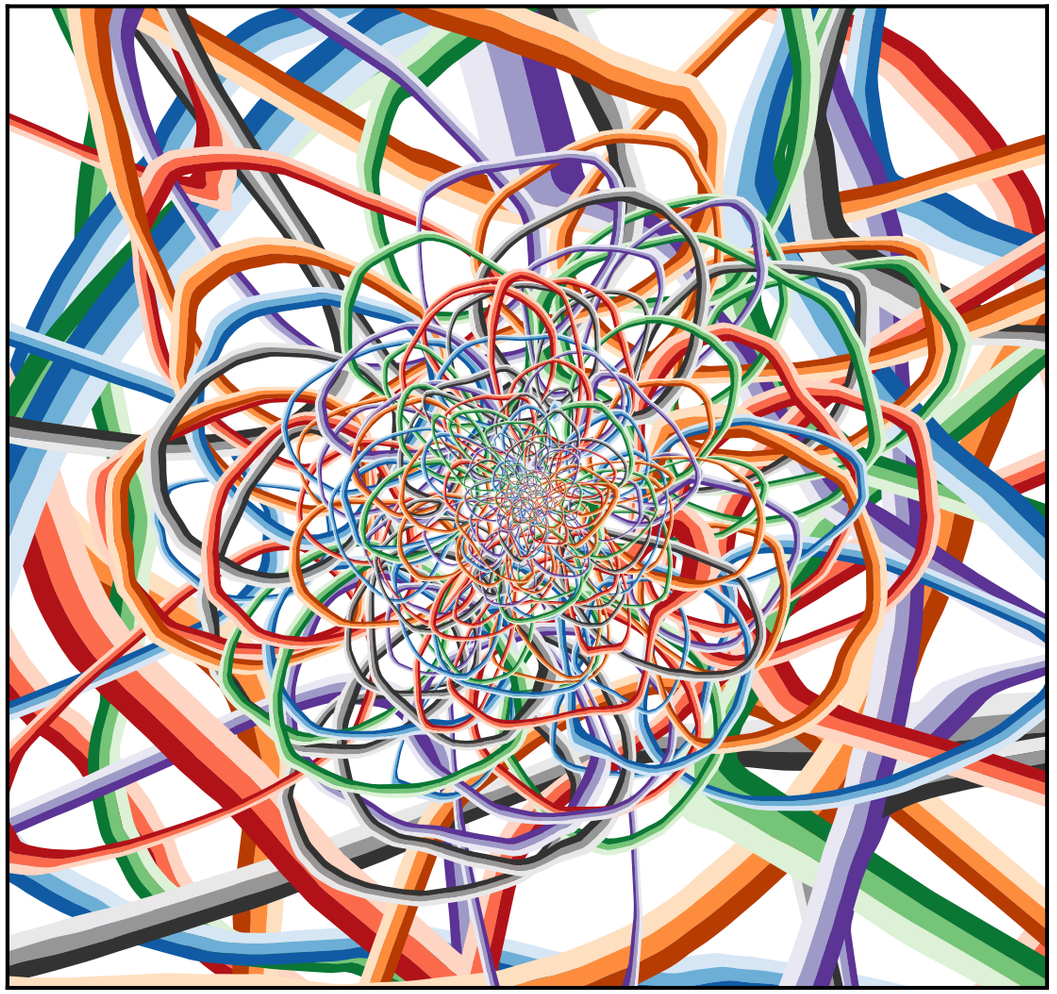


+



⇒

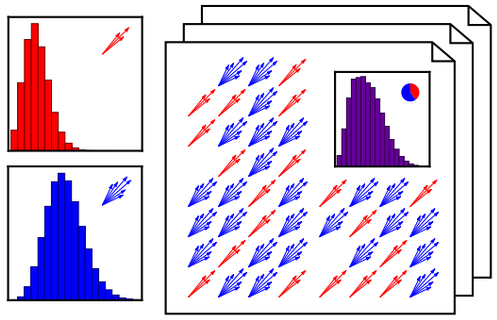
"Seeing" the collinear singularity of QCD



Augmenting our exceptional human/scientific ability to recognize patterns

[Komiske, Metodiev, JDT, 1810.05165]

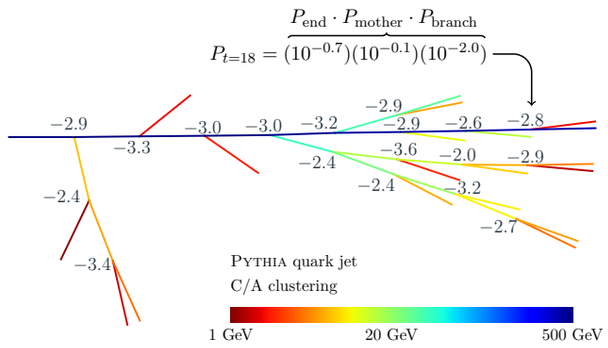
The Rise of Unsupervised Learning



Jet Topics

Blind Source Separation

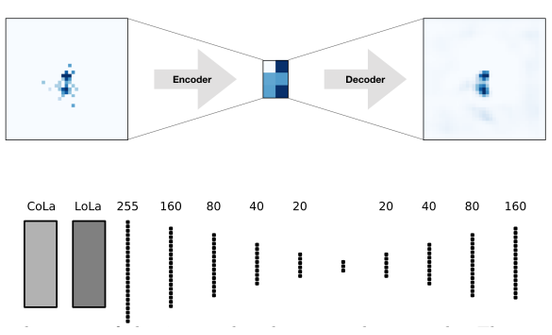
[Metodiev, JDT, 1802.00008; Komiske, Metodiev, JDT, 1809.01140;
see also Metodiev, Nachman, JDT, 1708.02949; Dillon, Faroughy, Kamenik, 1904.04200]



JUNIPR

Probability Modeling

[Andreassen, Feige, Frye, Schwartz, 1804.09720;
see also Monk, 1807.03685]



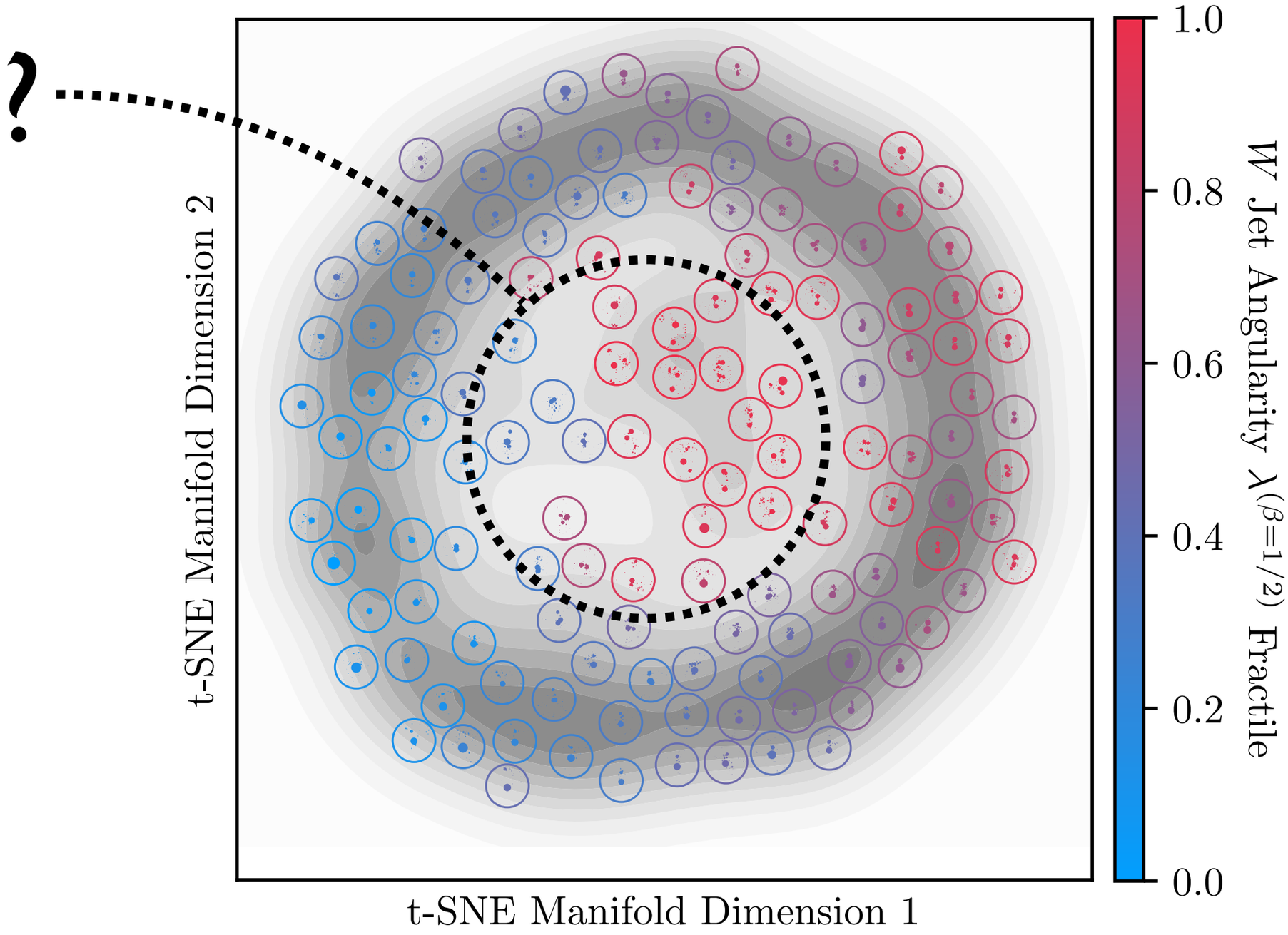
Autoencoders

Anomaly Detection

[Hajer, Li, Liu, Wang, 1807.10261; Heimel, Kasieczka, Plehn, Thompson, 1808.08979;
Farina, Nakai, Shih, 1808.08992; Cerri, Nguyen, Pierini, Spiropulu, Vlimant, 1811.10276;
see also Collins, Howe, Nachman, 1805.02664, 1902.02634; De Simone, Jacques, 1807.06038]

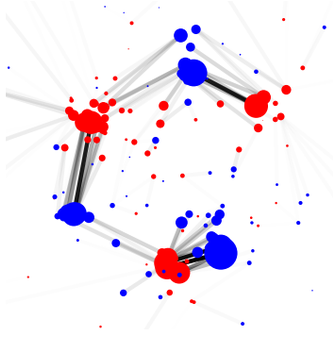
Common theme: Analyze event ensembles, not individual events

Visualization meets Anomaly Detection



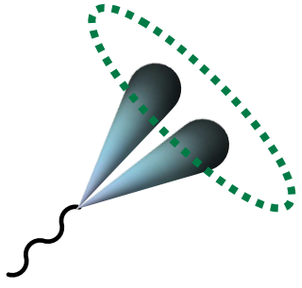
[Komiske, Metodiev, JDT, 1902.02346]

Summary



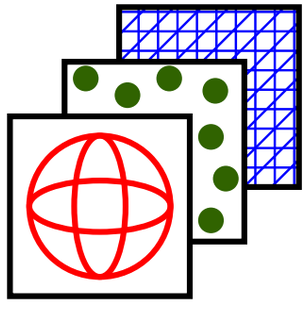
The Energy Mover's Distance

A metric on the space of IRC-safe energy flows



Initial Physics Studies

A geometric language for collider data exploration



(Broader Comments)

Machine learning as a pathway to LHC discoveries