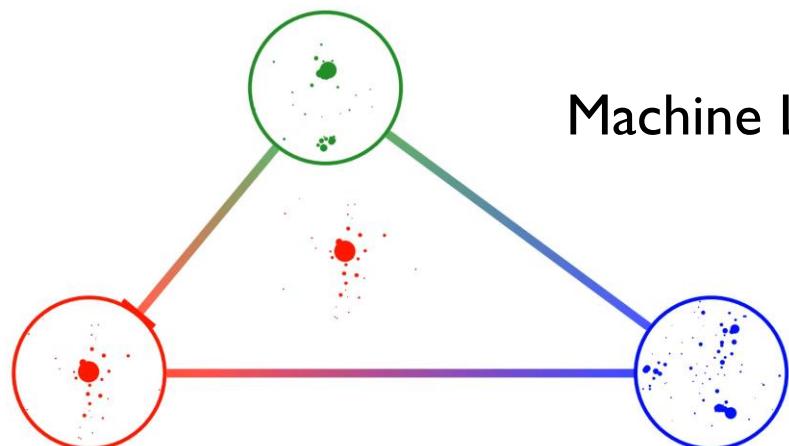


# Metric Methods with Open Collider Data



Machine Learning and the Physical Sciences, NeurIPS 2019

Eric M. Metodiev

Center for Theoretical Physics  
Massachusetts Institute of Technology



Patrick  
Komiske



Radha  
Mastandrea

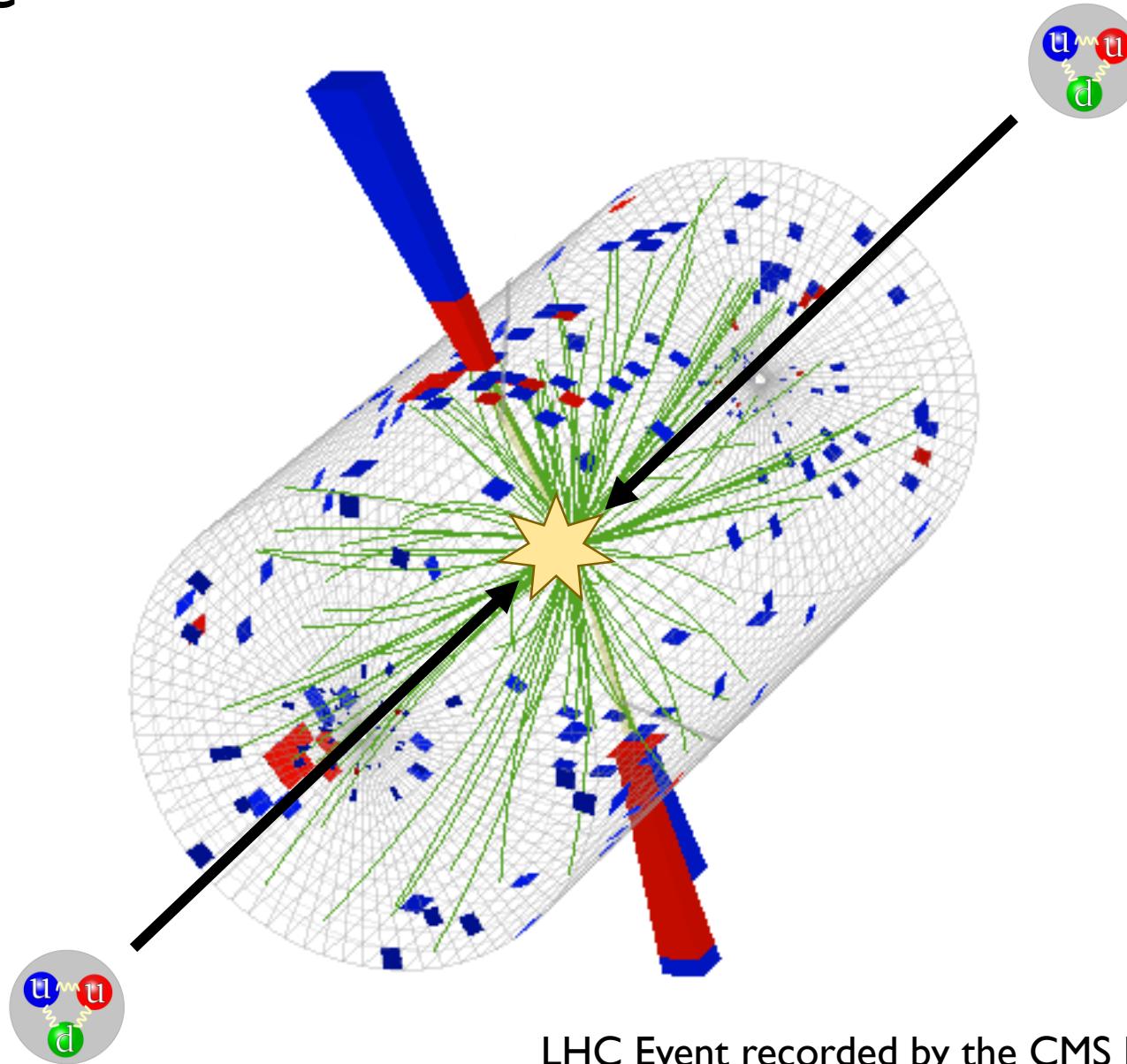


Preksha  
Naik



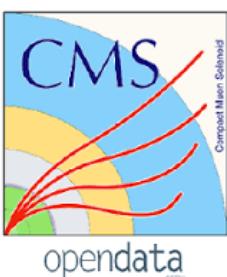
Jesse  
Thaler

# Collision Course

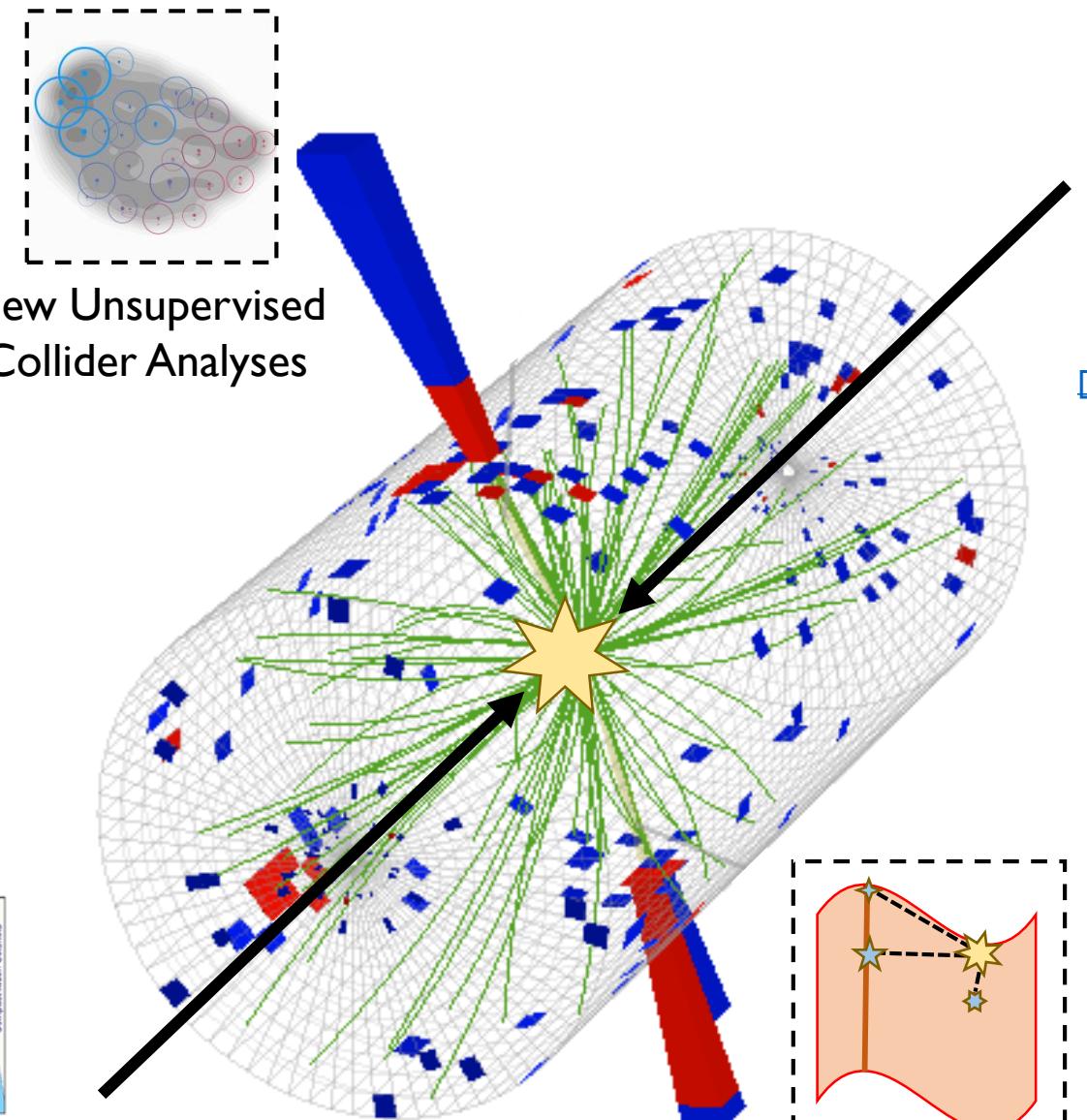


LHC Event recorded by the CMS Experiment at CERN

# Collision Course



Public Collider Data  
[\[opendata.cern.ch\]](https://opendata.cern.ch)



New Insights into  
Quantum Field Theory

[h/t Jesse Thaler]

# opendata.cern.ch

Explore more than **two petabytes**  
of open data from particle physics!

jet primary dataset

search examples: [collision datasets](#), [keywords:education](#), [energy:7TeV](#)

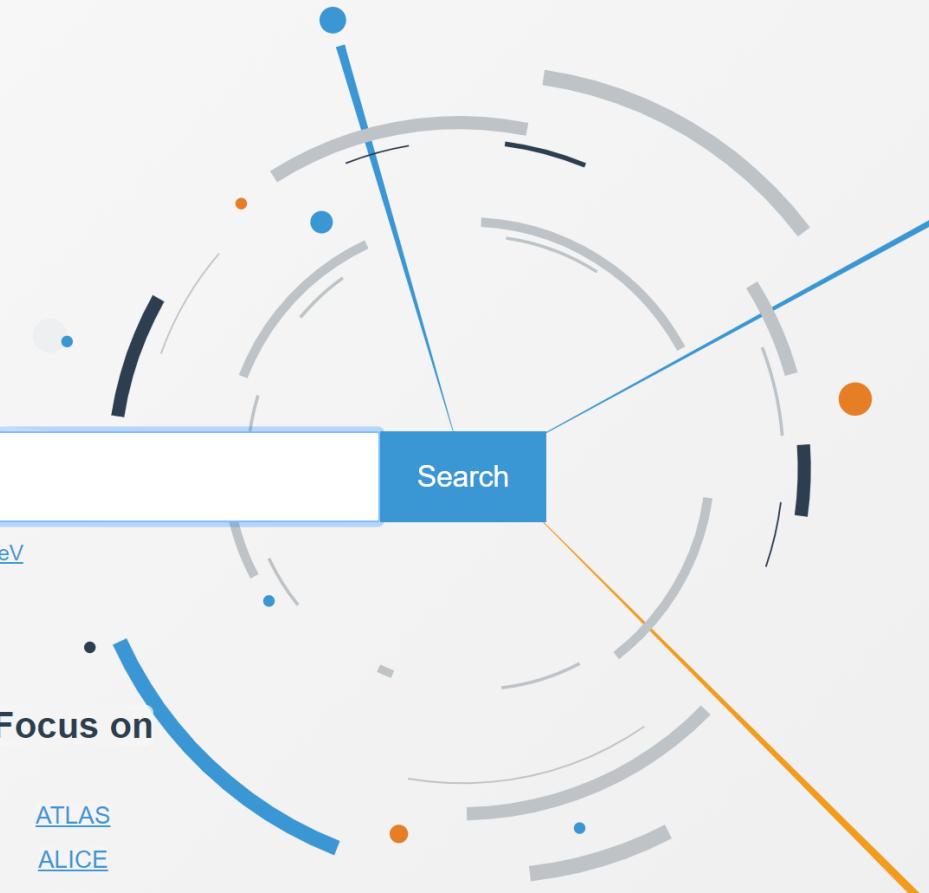
## Explore

[datasets](#)  
[software](#)  
[environments](#)  
[documentation](#)

## Focus on

[ATLAS](#)  
[ALICE](#)  
[CMS](#)  
[LHCb](#)  
[OPERA](#)  
[Data Science](#)

▼ Get started ▼



# CMS Open Data

Download a CMS “AOD” file: [2011A Jet Primary Dataset](#)

04913DA0-8B3F-E311-924F-0025901AD38A.root

966.3 MB



Fifteen lines of code later...

```
import uproot

# Load in the specified file with uproot
file = uproot.open('~/Downloads/04913DA0-8B3F-E311-924F-0025901AD38A.root')
events = file[b'Events;1']

# read particle transverse momenta (pts), pseudorapidity (eta), and azimuth (phi)
PFCkey = b'recoPFCandidates_particleFlow__RECO.obj'
pts = events[PFCkey + b'.pt_'].array()
etas = events[PFCkey + b'.eta_'].array()
phis = events[PFCkey + b'.phi_'].array()
```

```
import numpy as np
import matplotlib.pyplot as plt

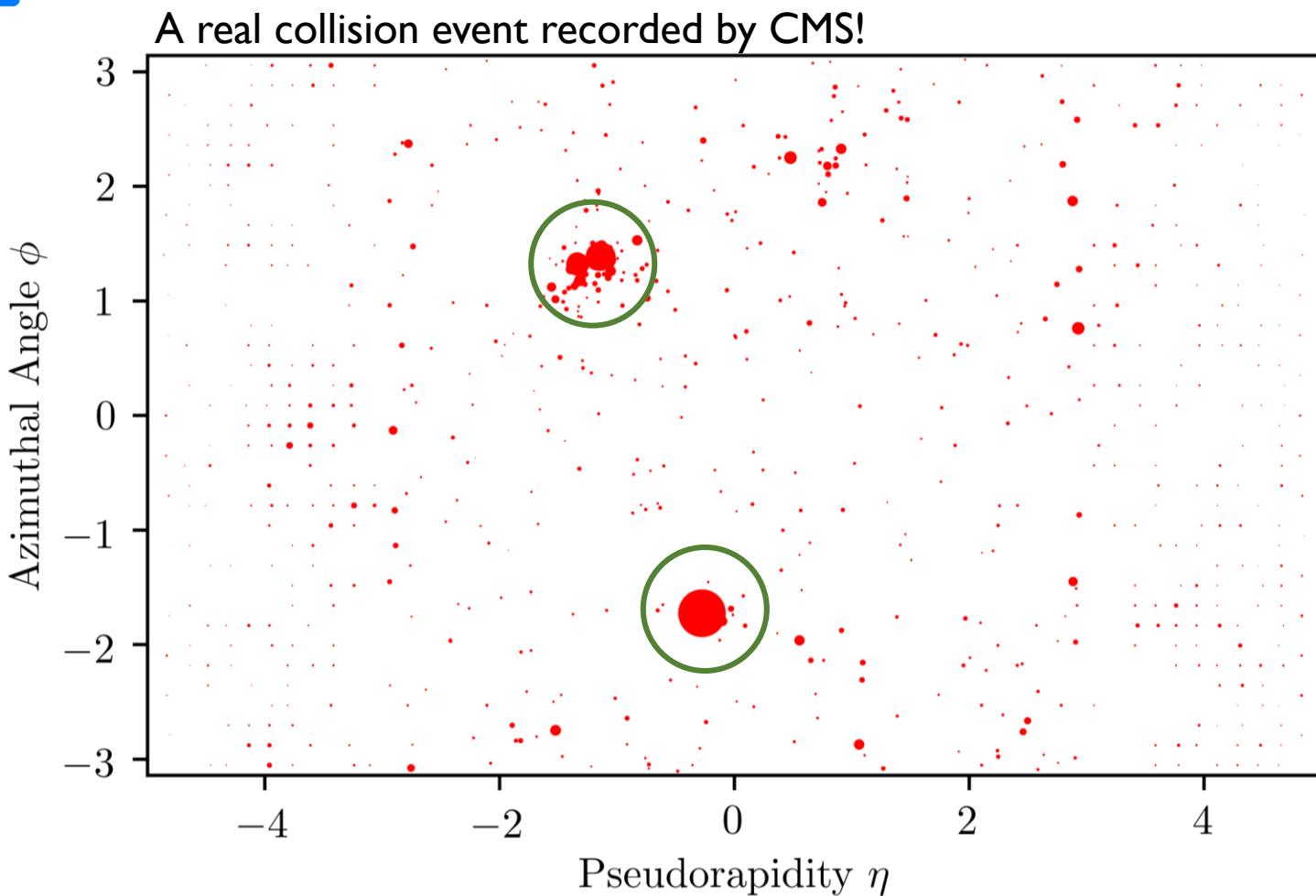
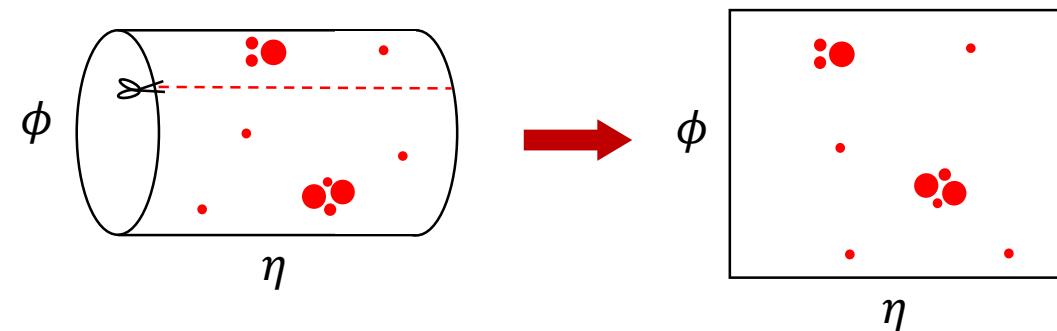
# choose an event
ind = 6457

# plot the collision event of interest
plt.scatter(etas[ind], phis[ind], s=pts[ind], lw=0, color='red')

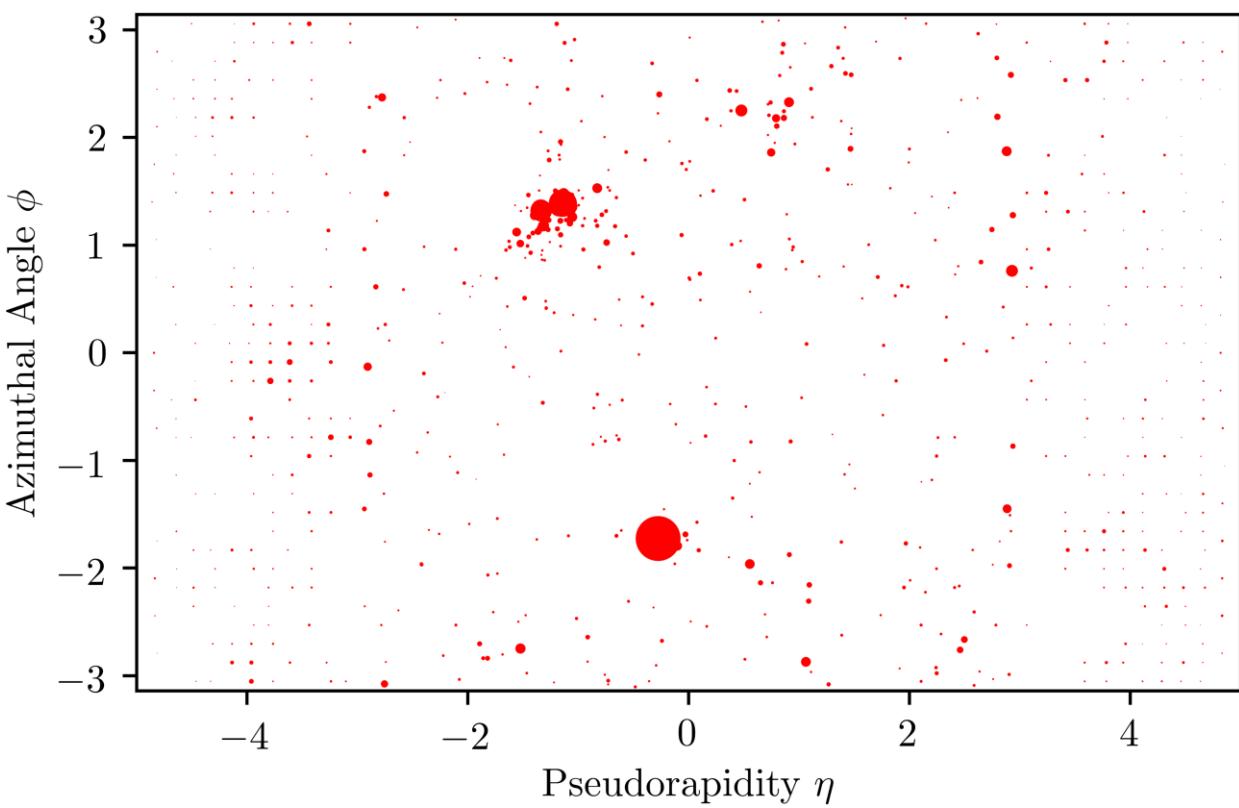
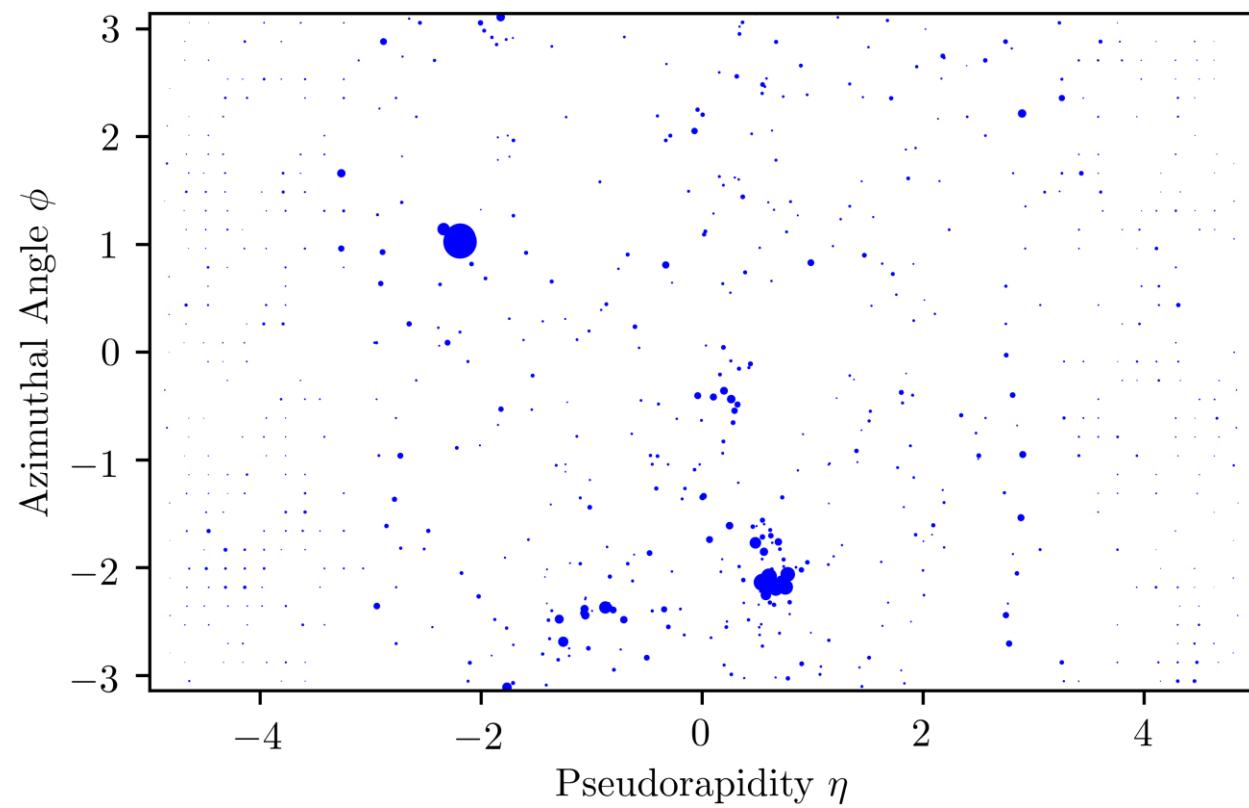
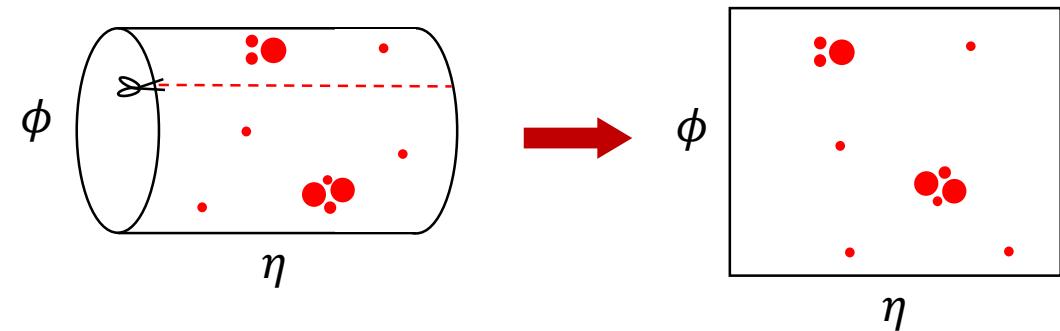
# plot settings
plt.xlim(-5, 5)
plt.ylim(-np.pi, np.pi)
plt.xlabel('Pseudorapidity $\eta$')
plt.ylabel('Azimuthal Angle $\phi$')

plt.show()
```

Thanks to the [uproot](#) package!

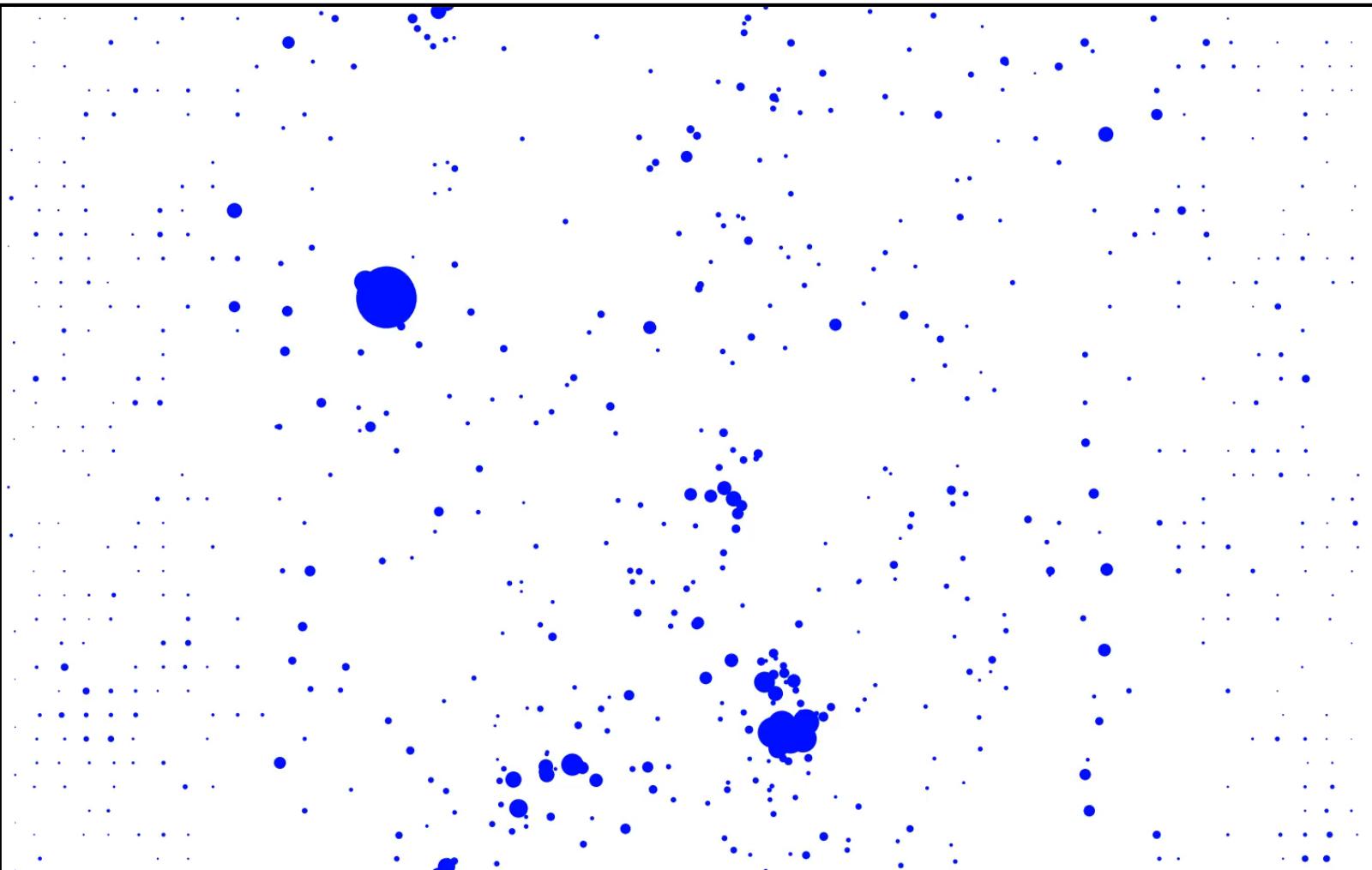
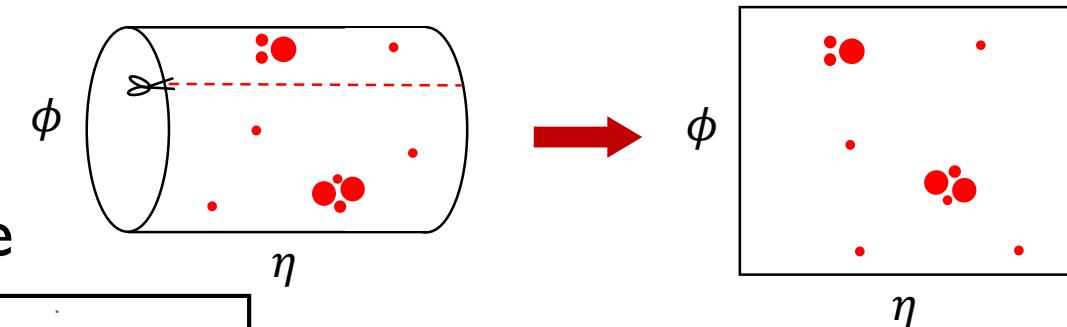


# When are two collisions similar?



# When are two collisions similar?

The Earth Mover's (or Wasserstein) Distance



The “work” required to rearrange one collision event into another!

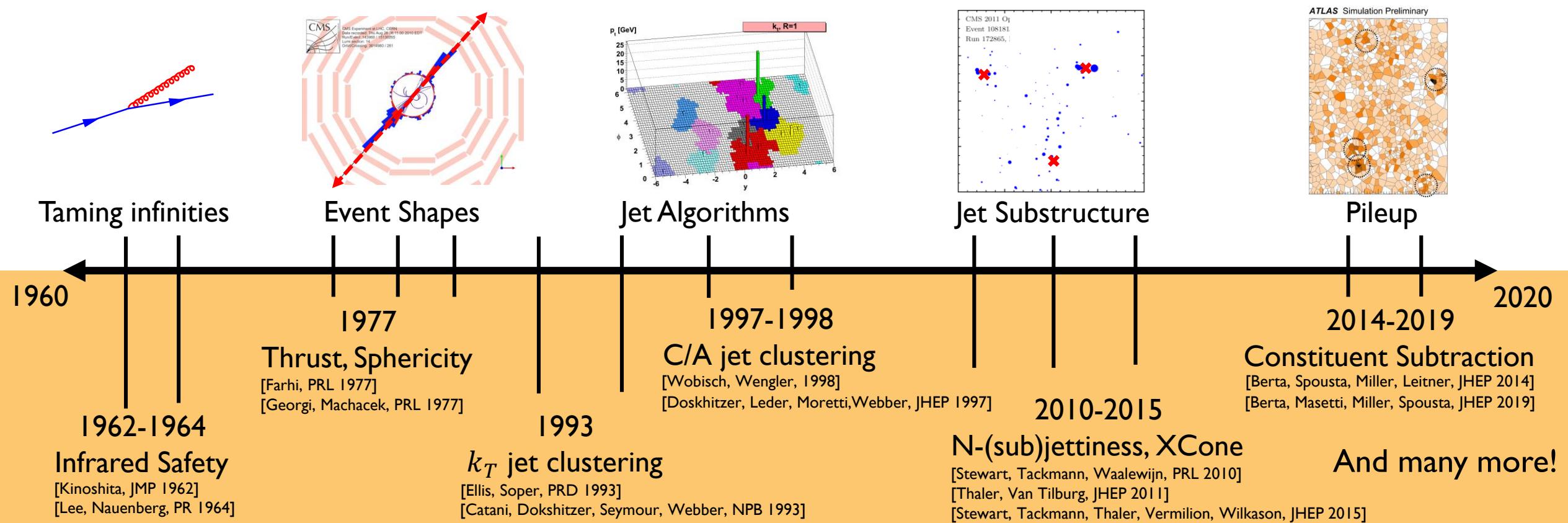
Plus a cost to create or destroy energy.

Optimal Transport Problem

Here using [python optimal transport](#)

[Komiske, EMM, Thaler, PRL 2019]

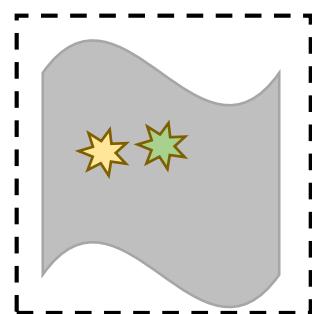
# Six Decades of Collider Techniques



# Six Decades of Collider Techniques as Optimal Transport!

[Komiske, EMM, Thaler, to appear]

Smooth function of energy distribution are finite in QFT



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

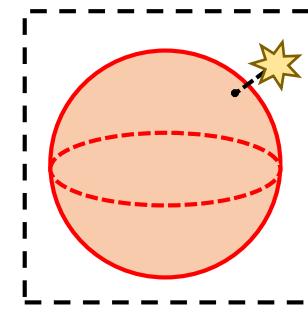
Taming infinities

1960

1962-1964

Infrared Safety  
[Kinoshita, JMP 1962]  
[Lee, Nauenberg, PR 1964]

Event shapes as distances to the 2-particle manifold



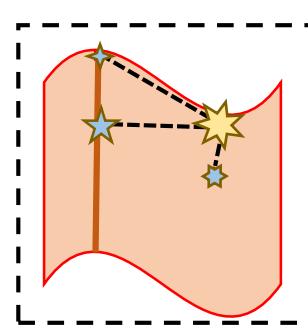
$$t(\mathcal{E}) = \min_{|\mathcal{E}'|=2} \text{EMD}(\mathcal{E}, \mathcal{E}')$$

Event Shapes

1977

Thrust, Sphericity  
[Farhi, PRL 1977]  
[Georgi, Machacek, PRL 1977]

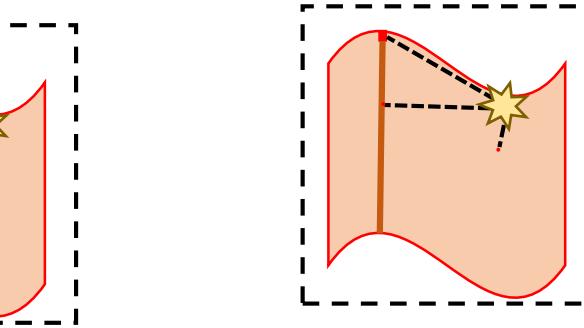
Jets are N-particle event approximations



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

Jet Algorithms

$k_T$  jet clustering  
[Ellis, Soper, PRD 1993]  
[Catani, Dokshitzer, Seymour, Webber, NPB 1993]

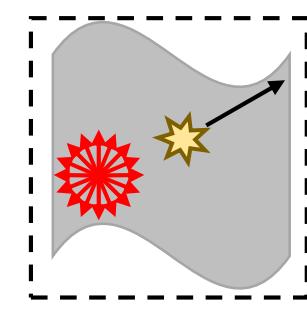


Jet Substructure

2010-2015

N-(sub)jettiness, XCone  
[Stewart, Tackmann, Waalewijn, PRL 2010]  
[Thaler, Van Tilburg, JHEP 2011]  
[Stewart, Tackmann, Thaler, Vermilion, Wilkason, JHEP 2015]

Subtract a pileup as a uniform distribution



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

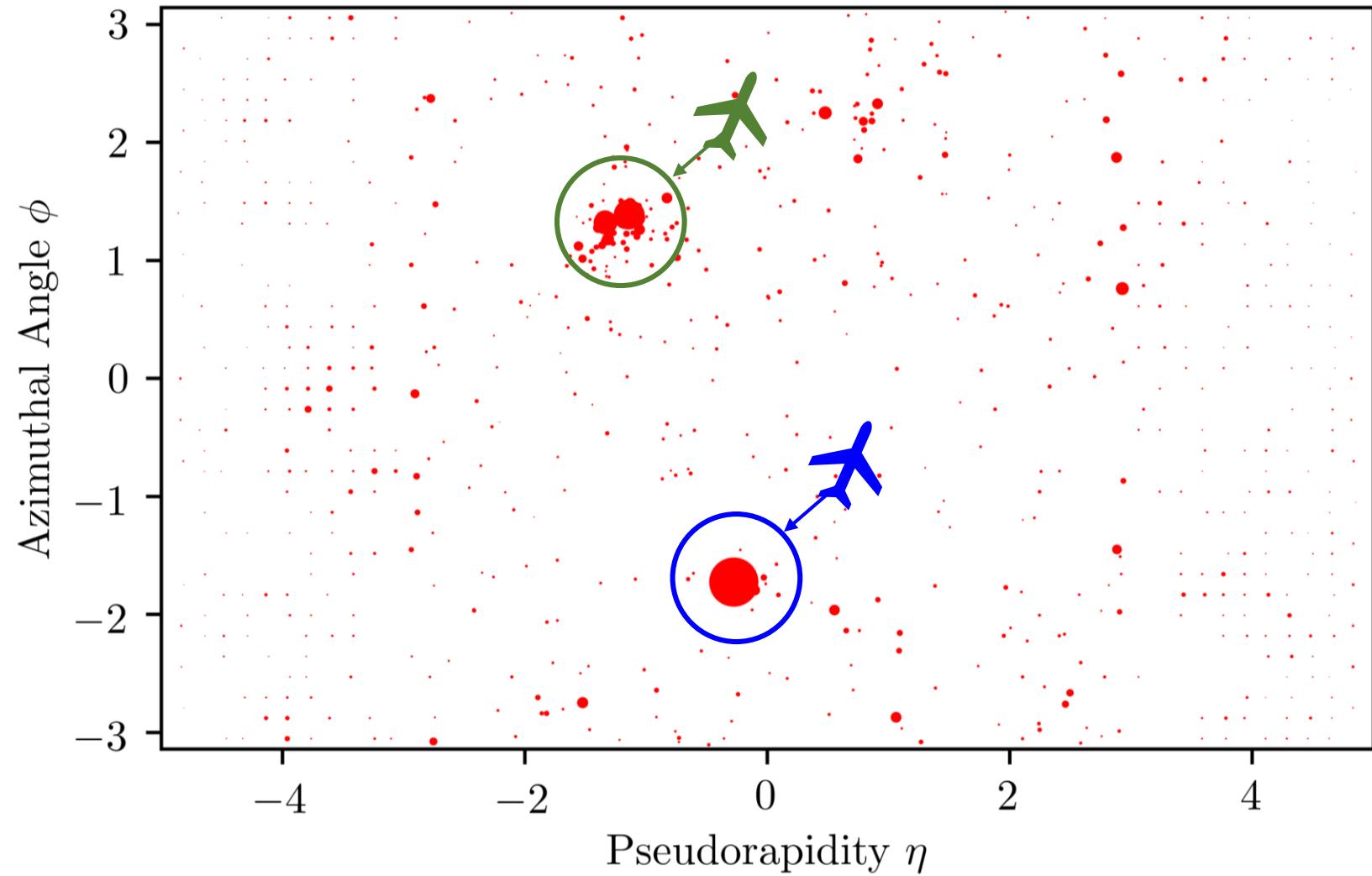
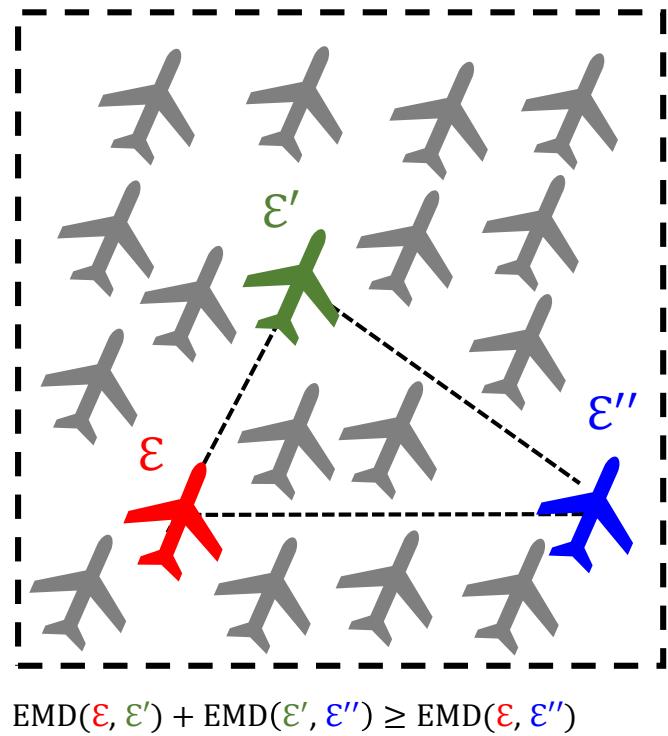
Pileup

2014-2019

Constituent Subtraction  
[Berta, Spousta, Miller, Leitner, JHEP 2014]  
[Berta, Masetti, Miller, Spousta, JHEP 2019]

And many more!

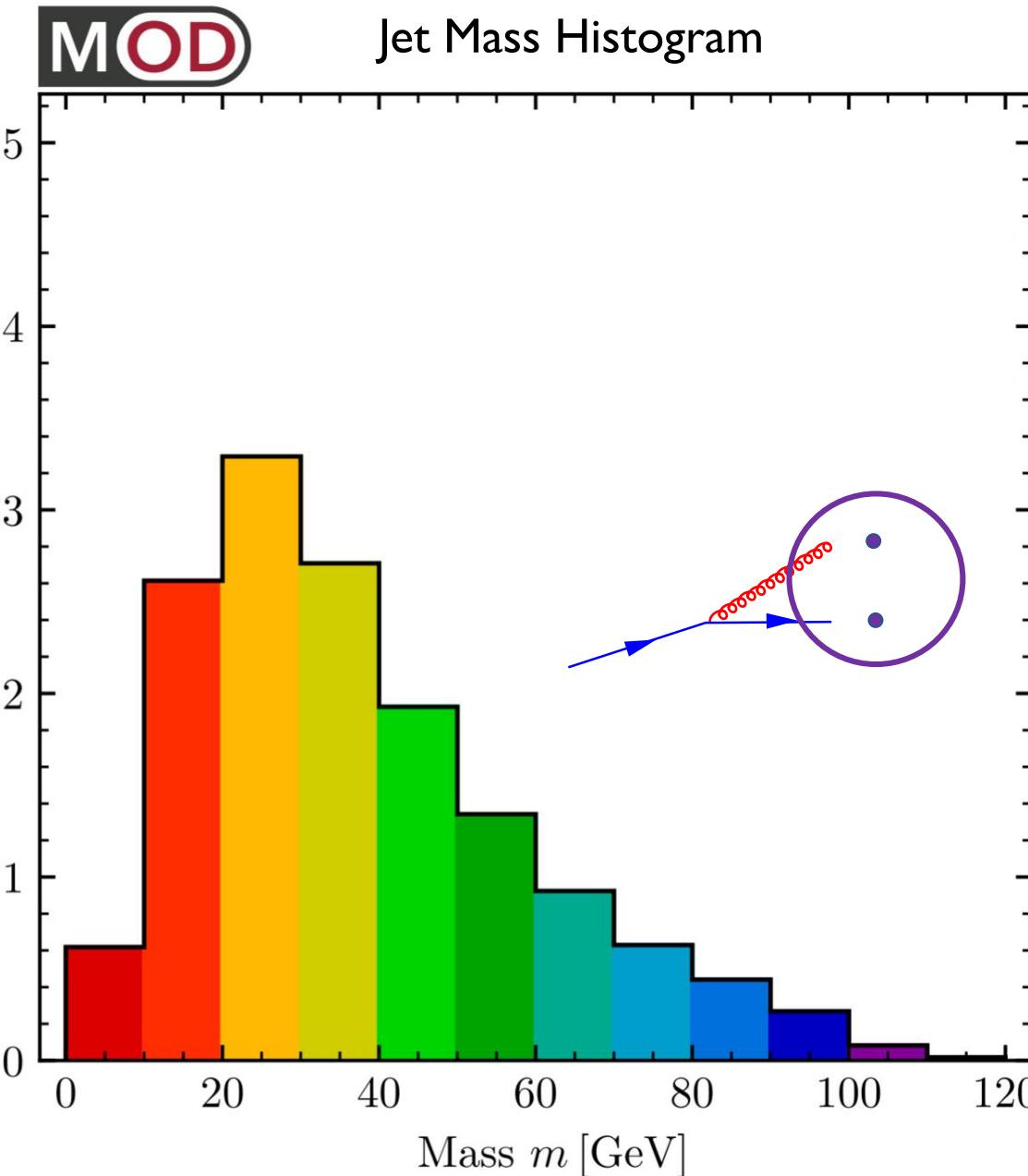
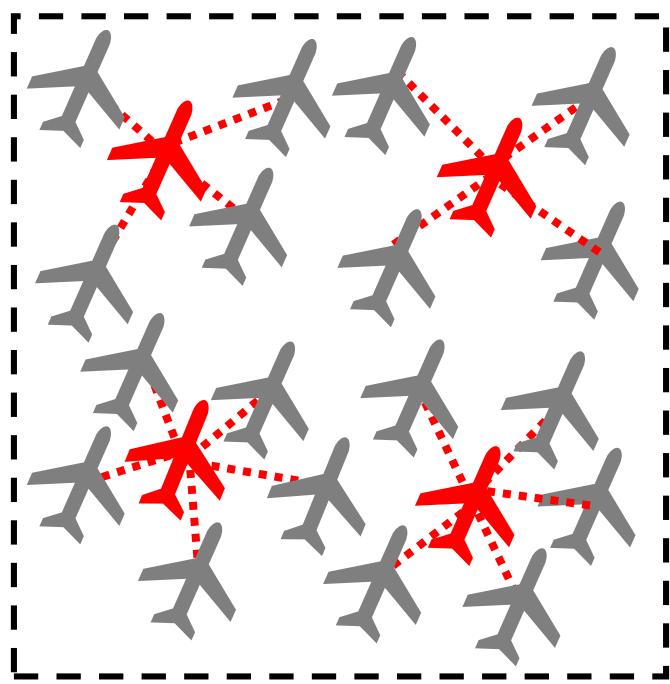
# Exploring the Space of Jets



# Most Representative Jets

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

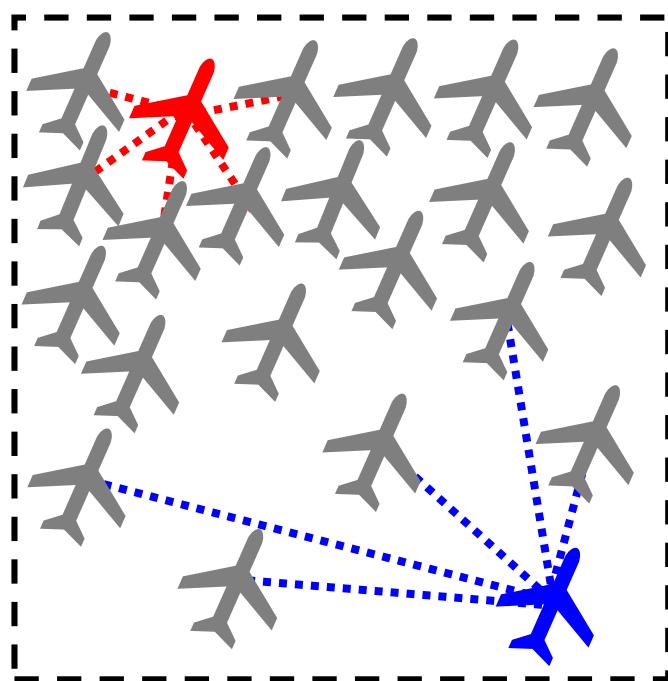
Measures how “wide” the jet is.



# Towards Anomaly Detection

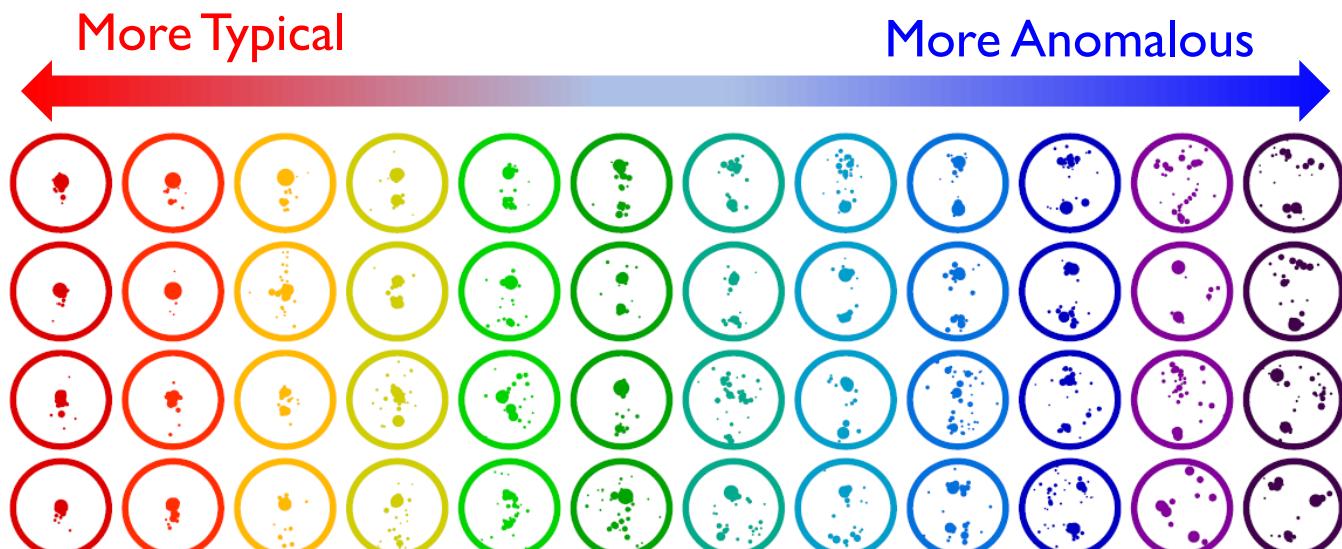
Mean EMD to Dataset:

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



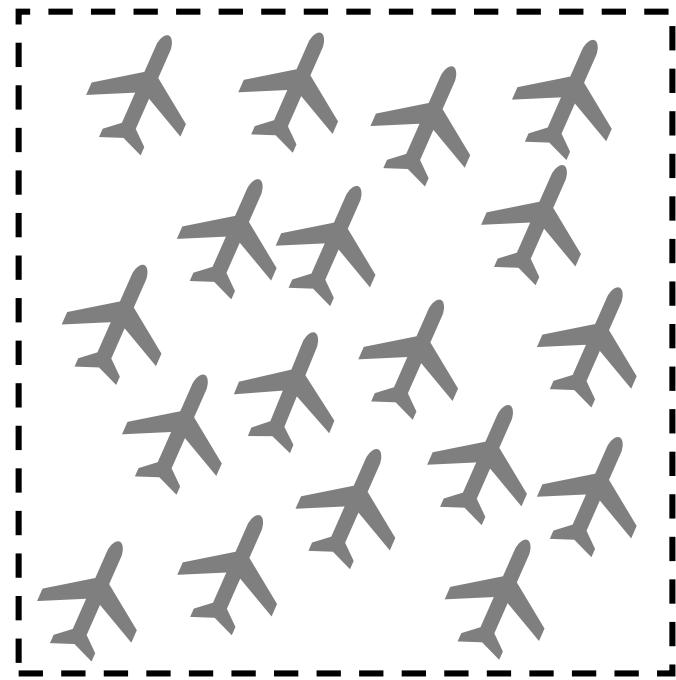
Complements recent developments in anomaly detection for collider physics.

- [\[Collins, Howe, Nachman, 1805.02664\]](#)
- [\[Heimel, Kasieczka, Plehn, Thompson, 1808.08979\]](#)
- [\[Farina, Nakai, Shih, 1808.08992\]](#)
- [\[Cerri, Nguyen, Pierini, Spiropulu, Vlimant, 1811.10276\]](#)

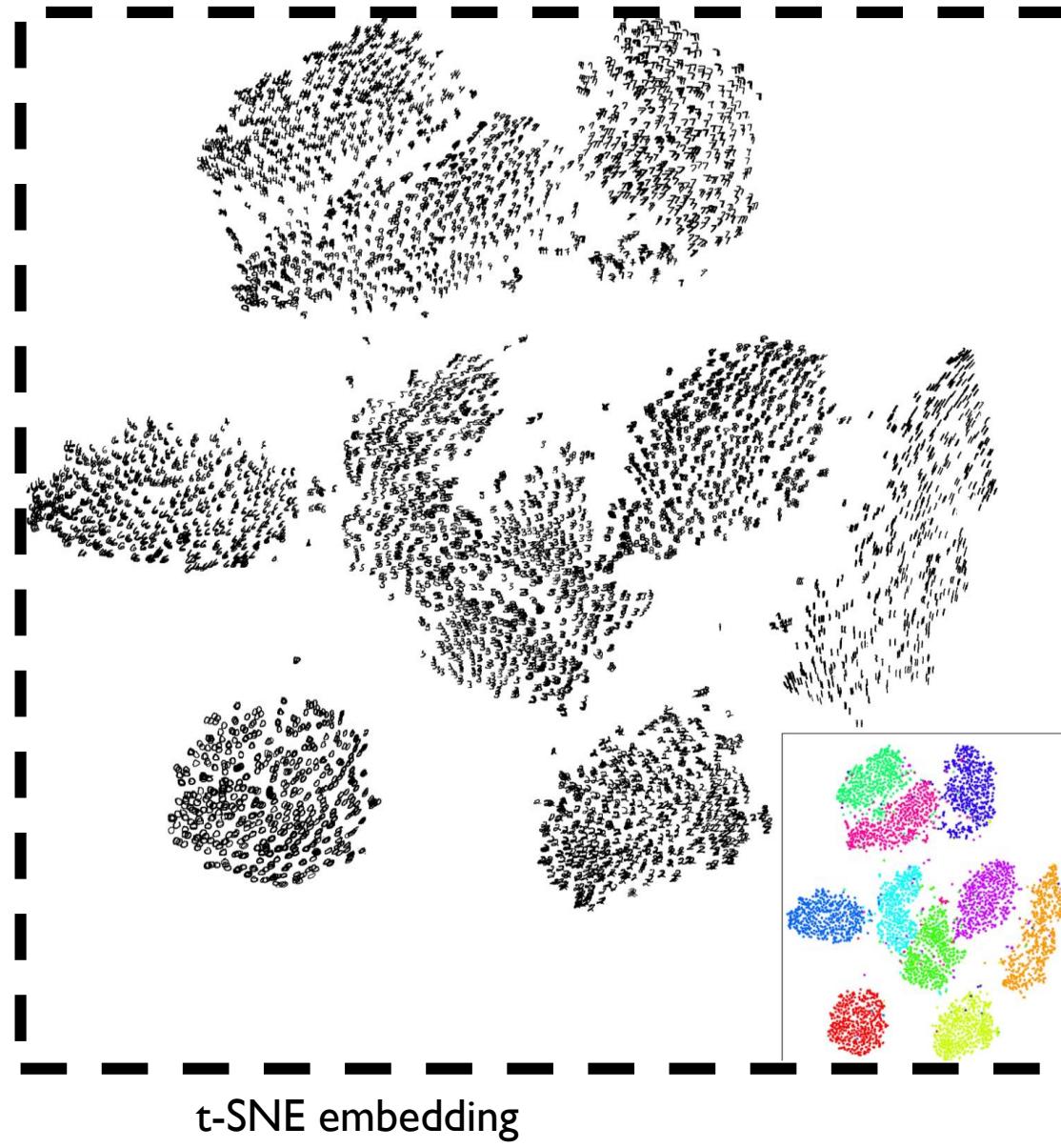


# Visualizing the Manifold

What does the space of jets look like?

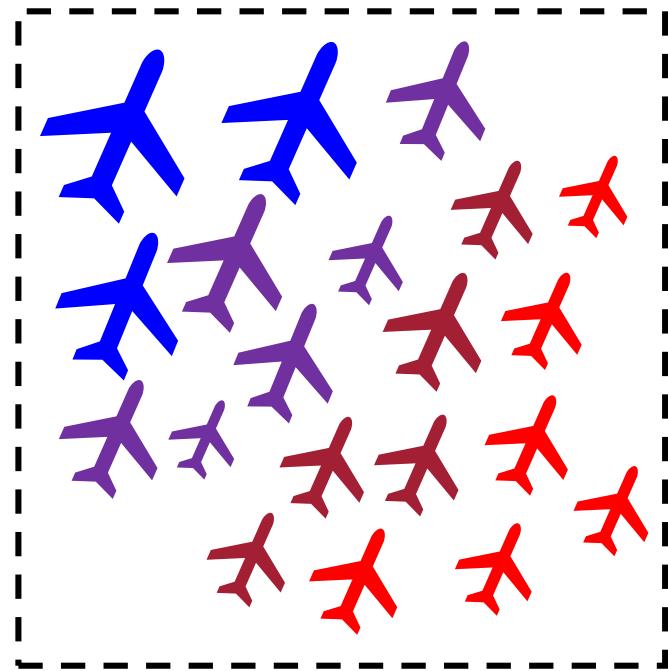


[van der Maaten, Hinton, JMLR 2008]



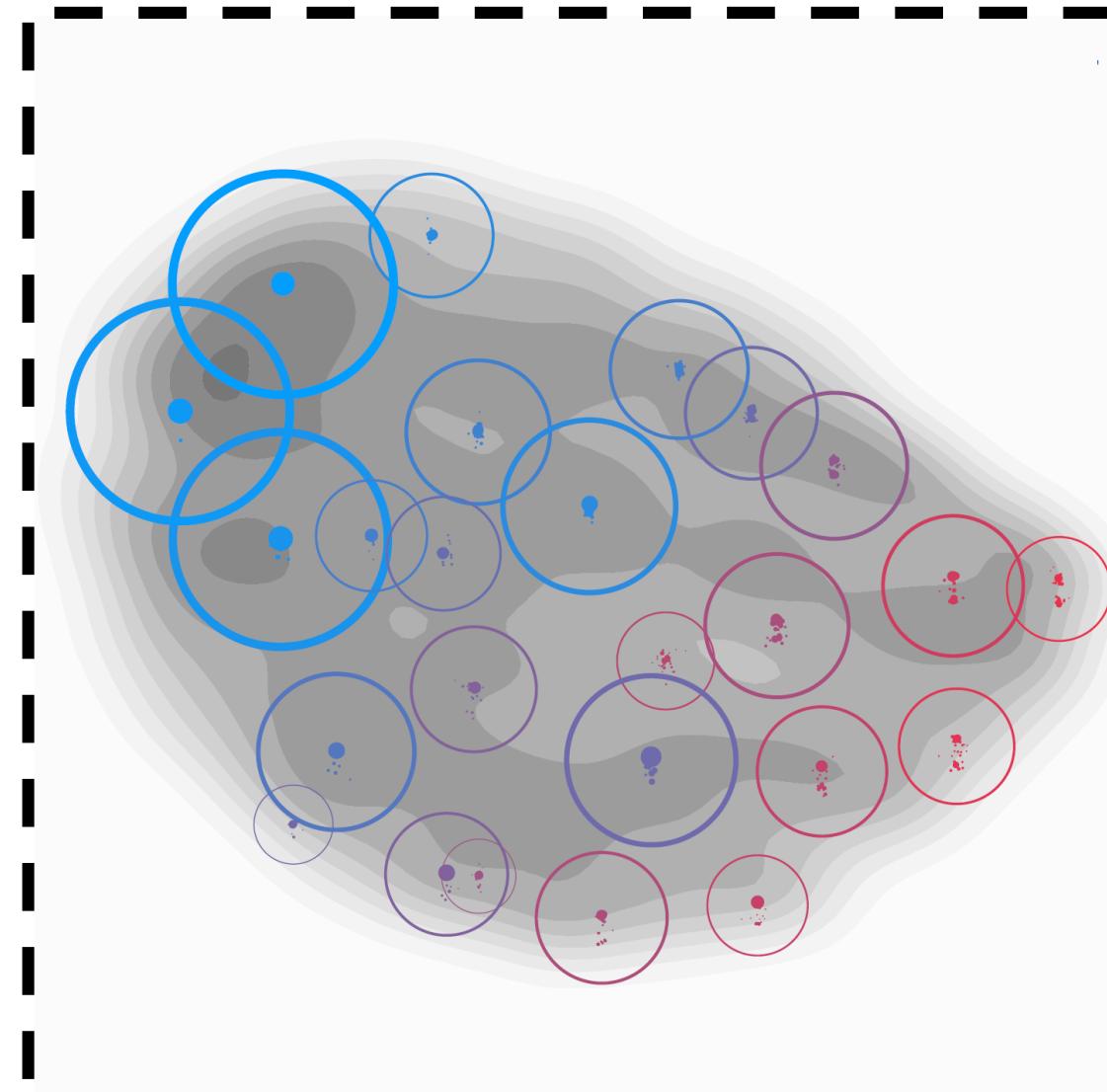
# Visualizing the Manifold

What does the space of jets look like?



[van der Maaten, Hinton, JMLR 2008]

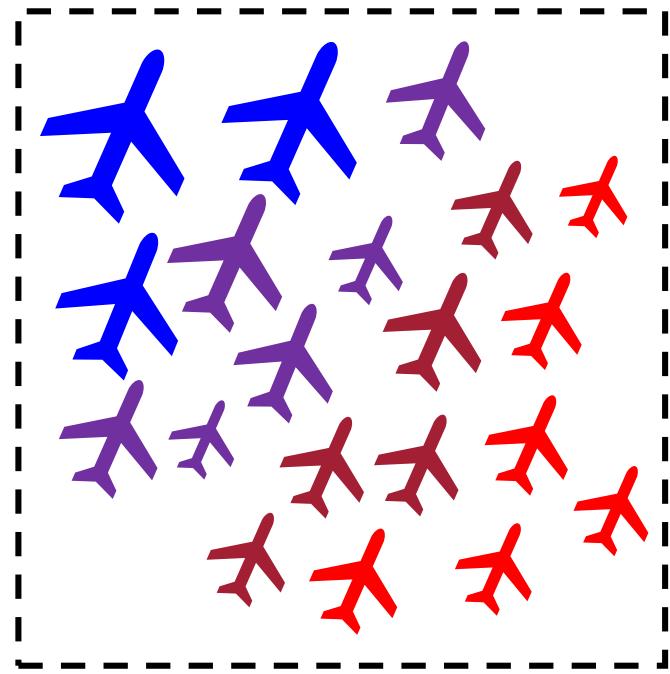
[\[Komiske, Mastandrea, EMM, Naik, Thaler, 1908.08542\]](#)



t-SNE embedding: 25-medoid jets shown

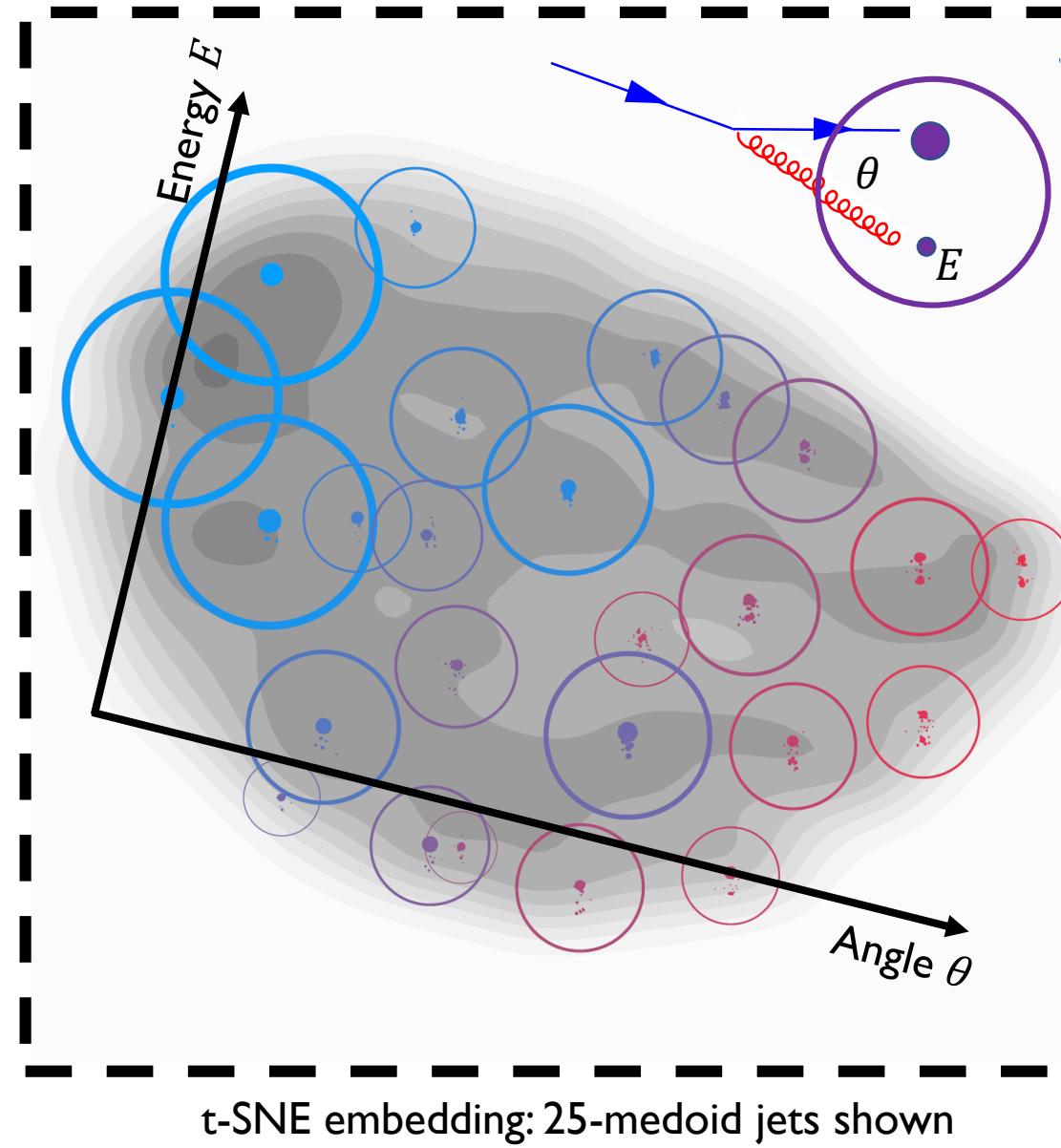
# Visualizing the Manifold

What does the space of jets look like?



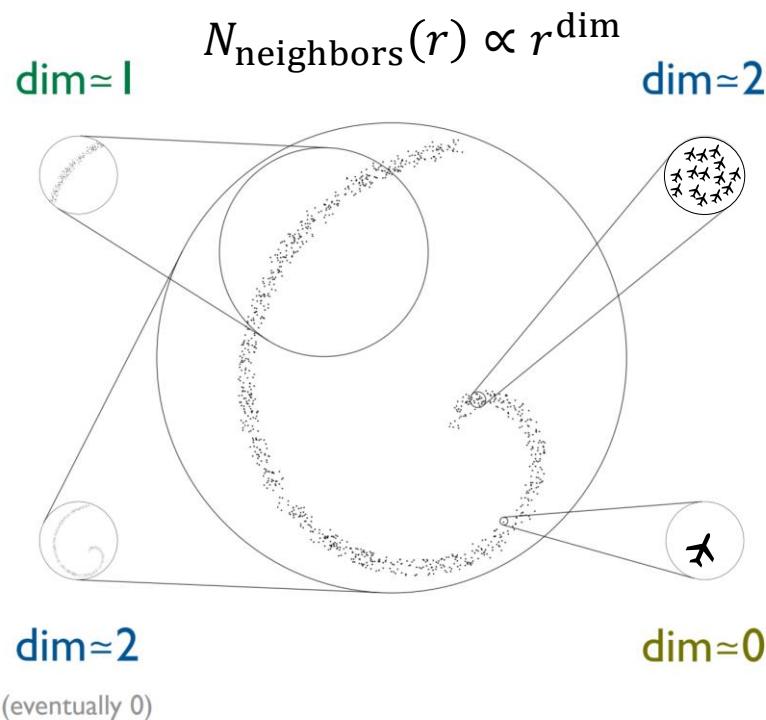
[van der Maaten, Hinton, JMLR 2008]

[Komiske, Mastandrea, EMM, Naik, Thaler, 1908.08542]



# Correlation Dimension

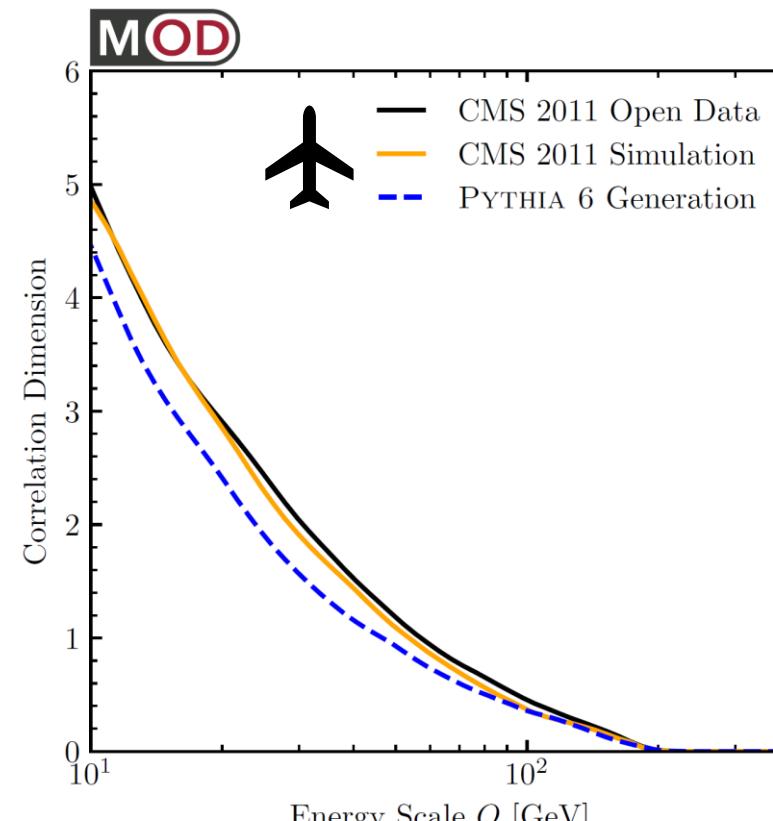
## Conceptual Idea



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

[Grassberger, Procaccia, PRL 1983] [Kegl, NeurIPS 2002]

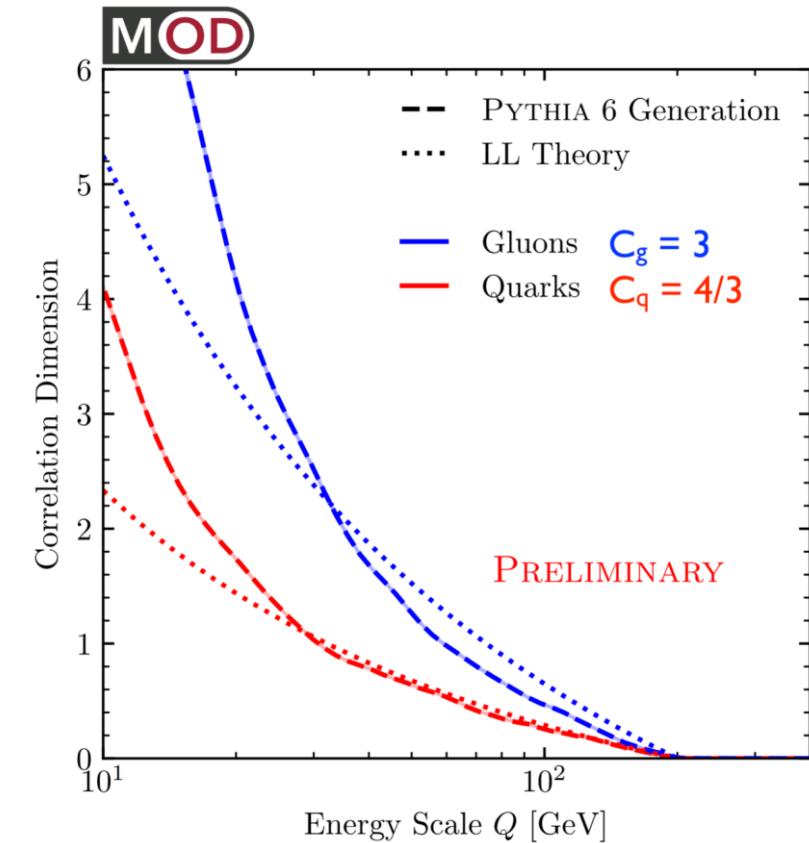
## Experimental Data



Dimension blows up at low energies.

[Komiske, Mastandrea, EMM, Naik, Thaler, 1908.08542]

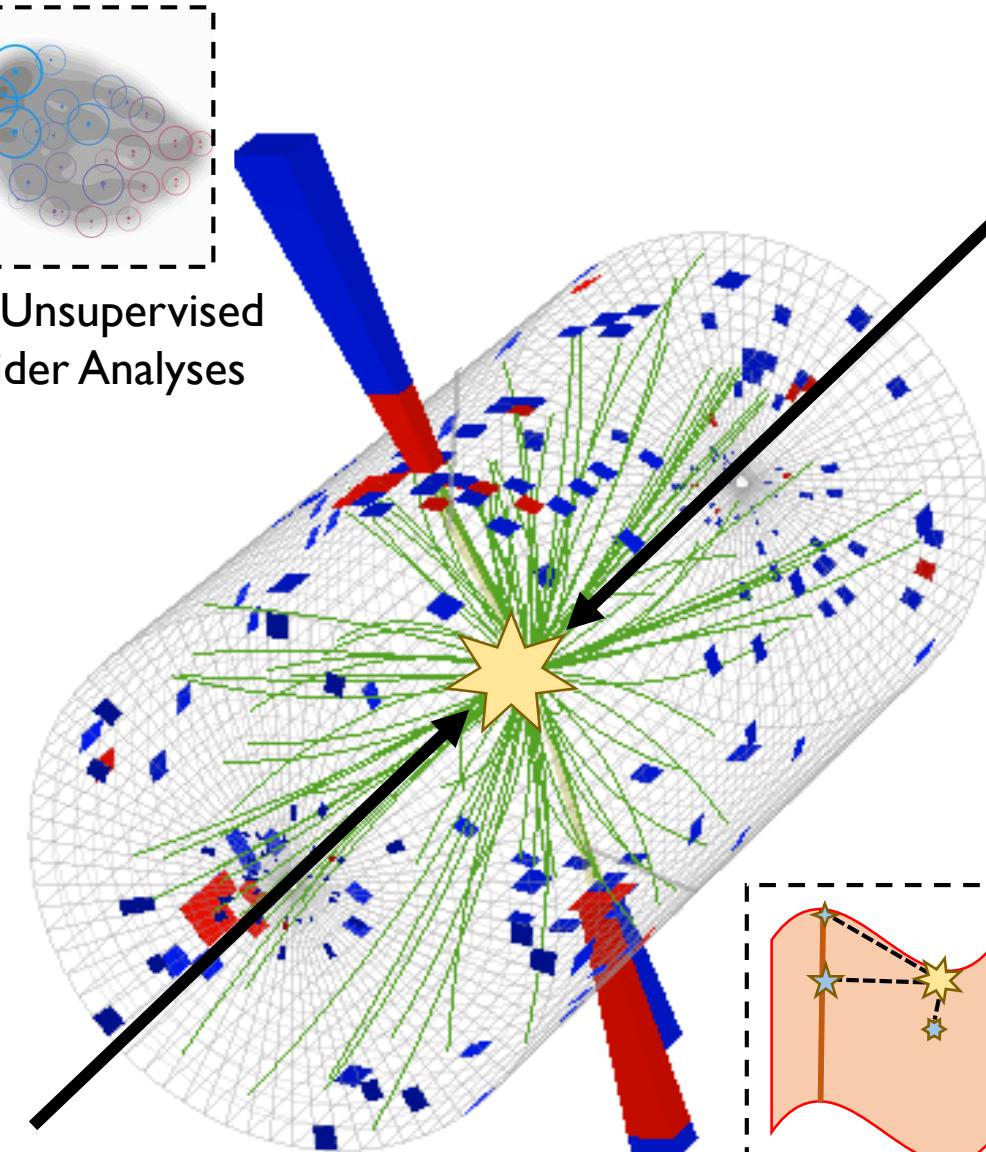
## Theoretical Calculation



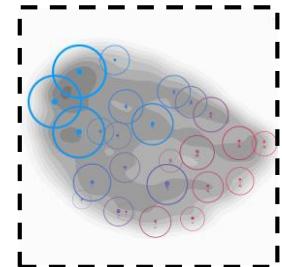
# Thank You!



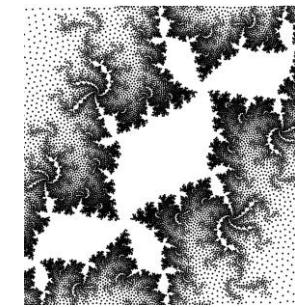
Public Collider Data  
[\[opendata.cern.ch\]](https://opendata.cern.ch)



New Unsupervised  
Collider Analyses



New Insights into  
Quantum Field Theory



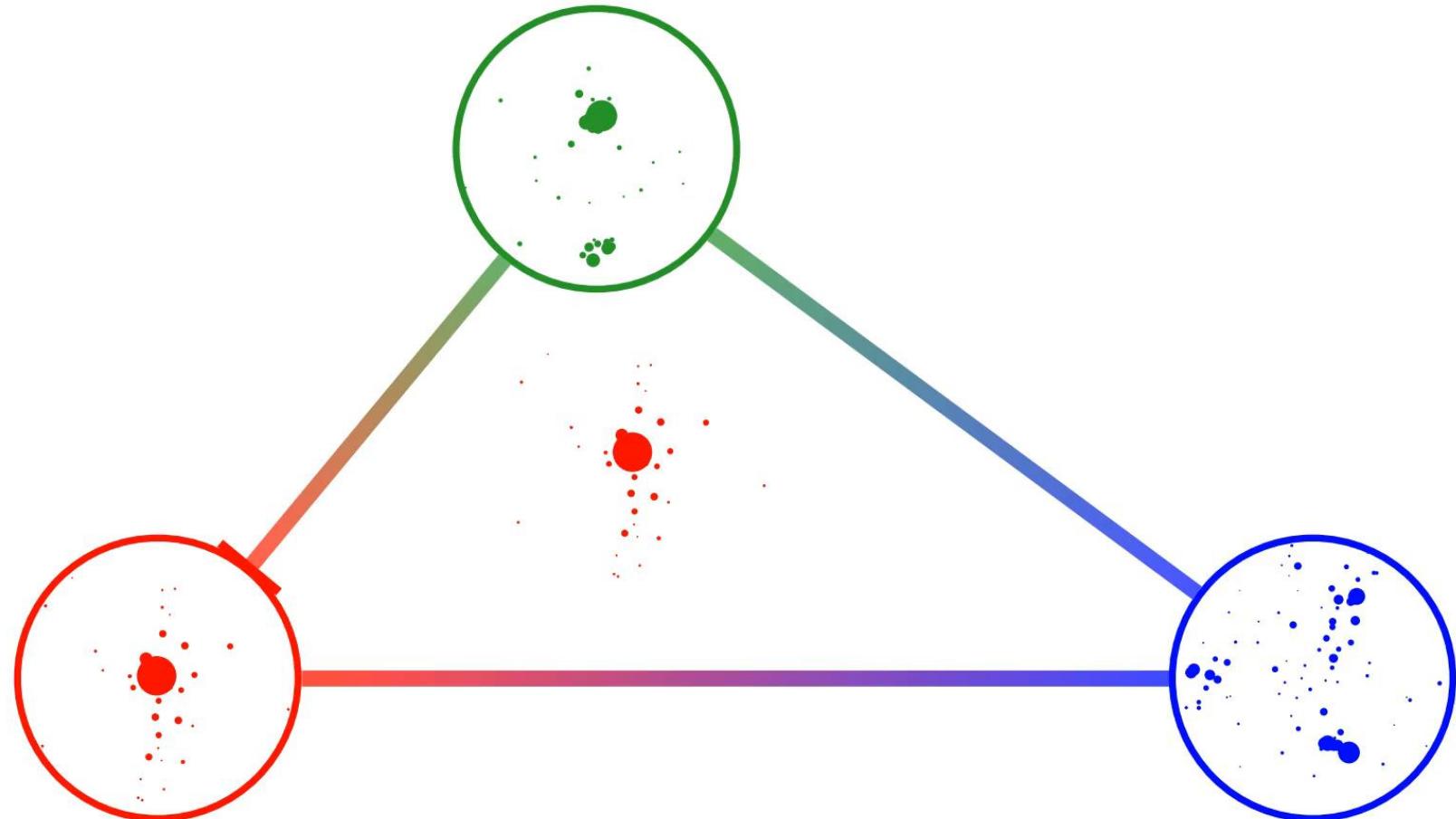
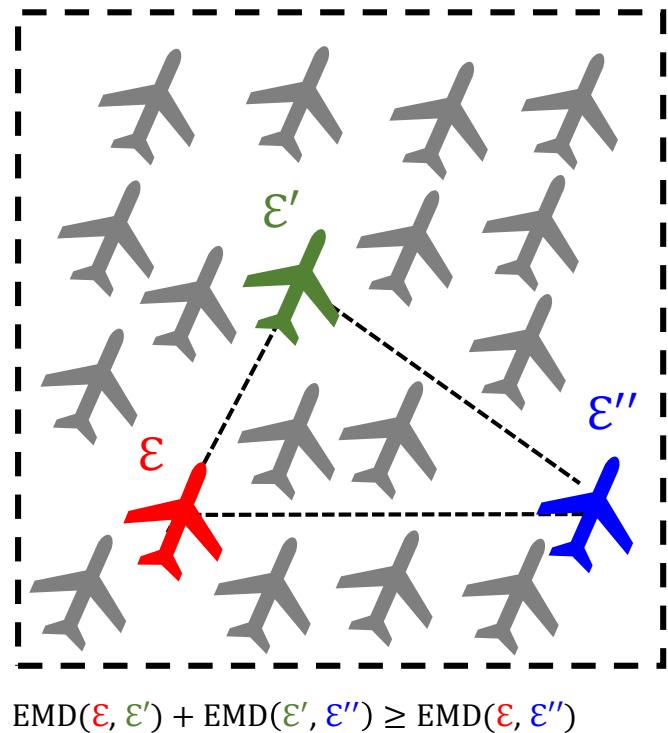
Optimal Transport  
[\[OTML Workshop, NeurIPS 2019\]](#)



[Publicly released](#)  
[jet dataset](#)

# Extra Slides

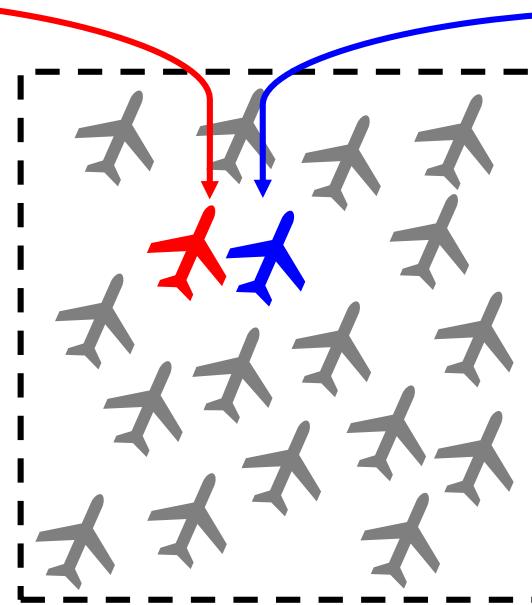
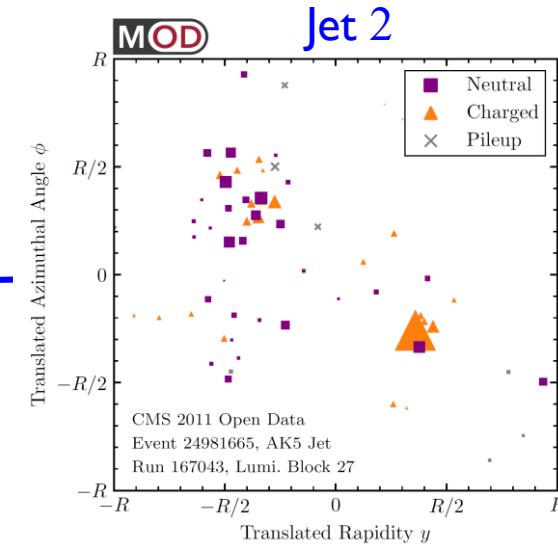
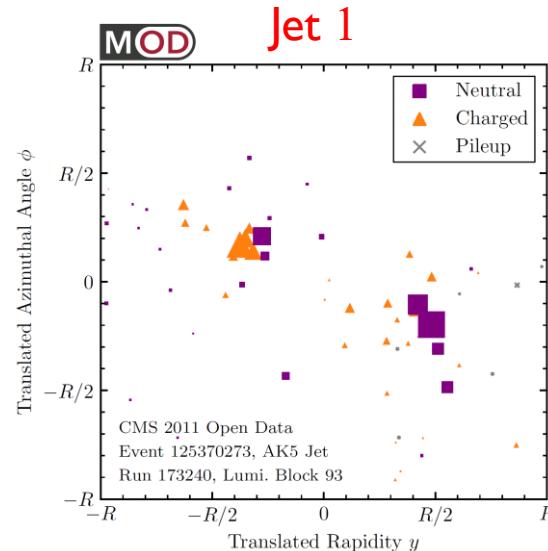
# Exploring the Space of Jets



# When are two events similar?

These two jets “look” similar, but have different numbers of particles, flavors, and locations.

How do we quantify this?

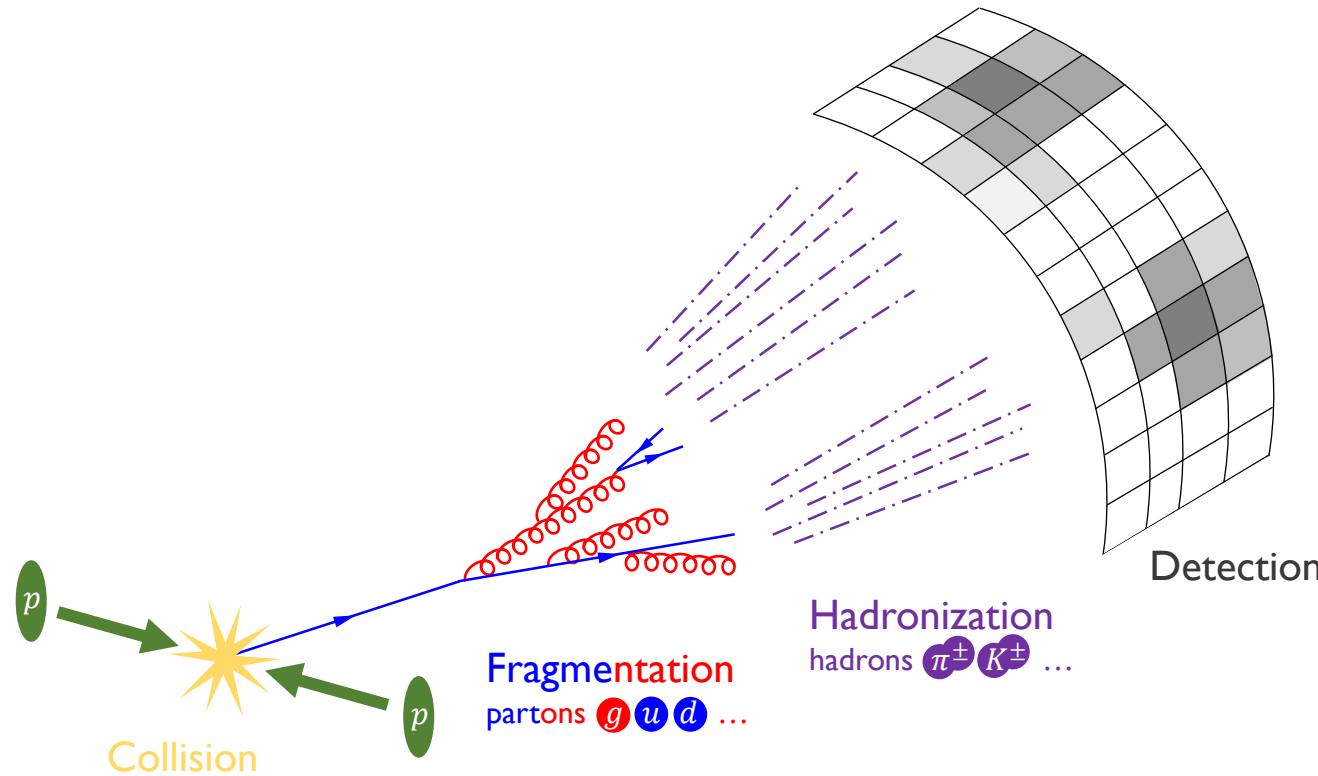


“Space of Jets”

400 GeV  $R = 0.5$  anti-  $k_T$  Jets from CMS Open Data

# When are two events similar?

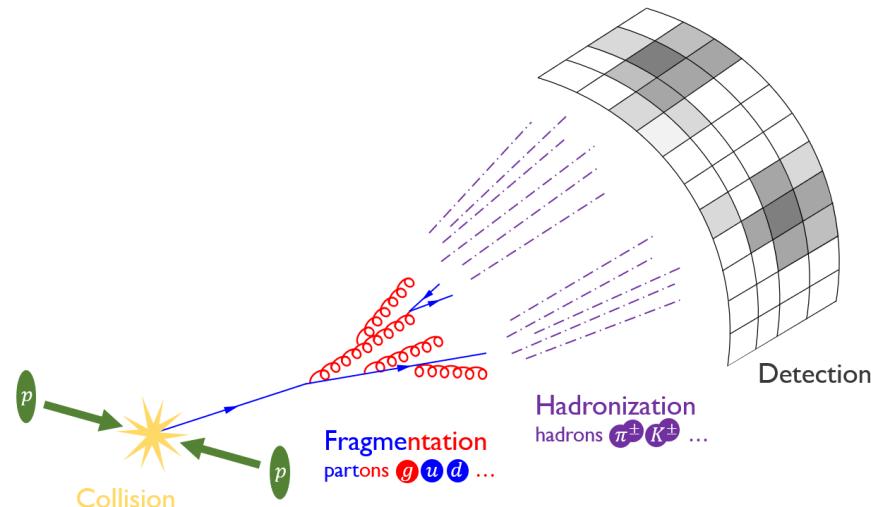
*How a jet gets its shape*



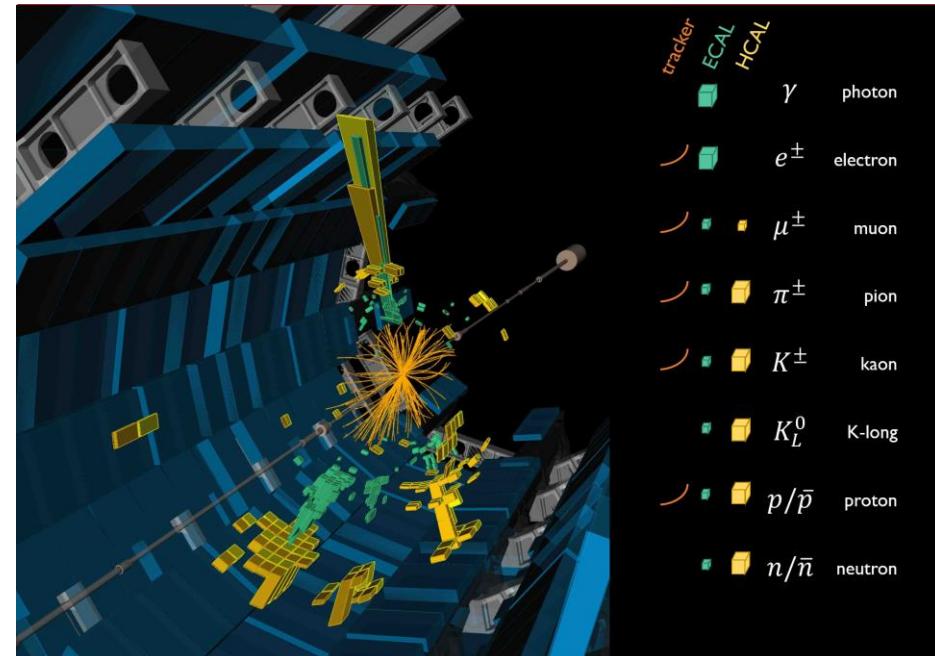
# When are two events similar?

An event is...

Theoretically: very complicated



Experimentally: very complicated



However:

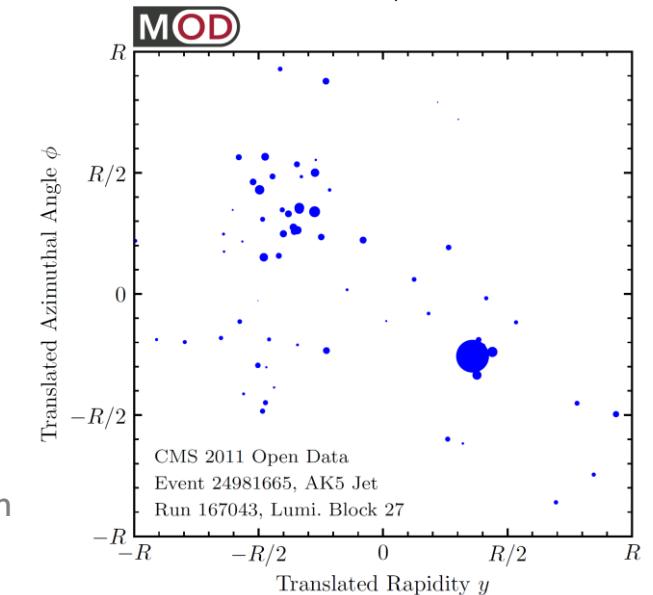
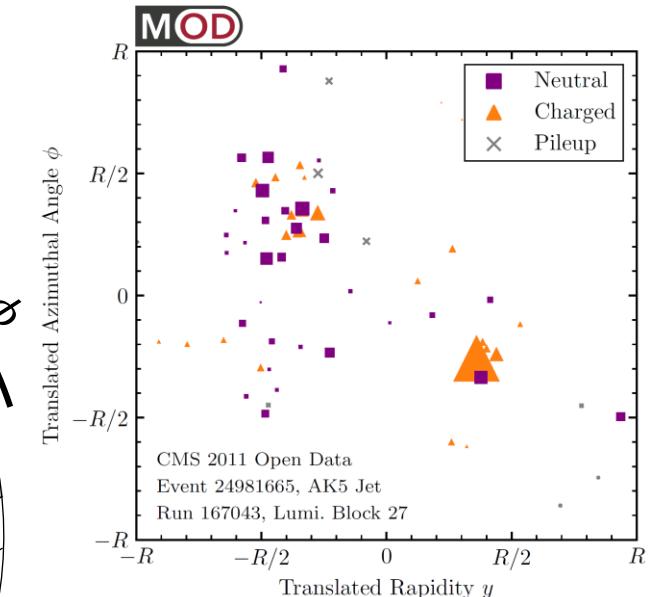
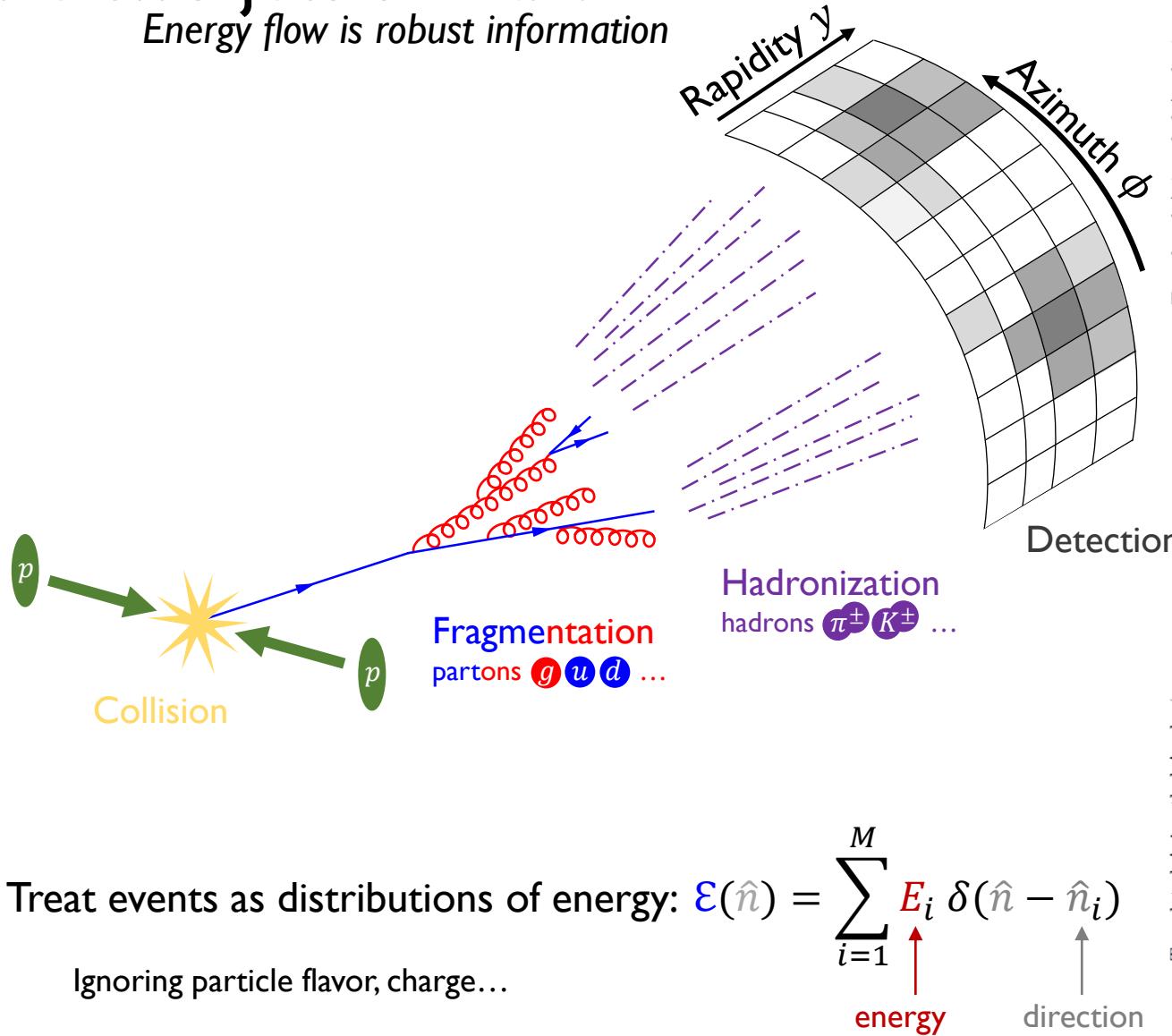
The energy flow (distribution of energy) is the information that is robust to:  
fragmentation, hadronization, detector effects, ...

[\[N.A. Sveshnikov, F.V. Tkachov, 9512370\]](#)  
[\[F.V. Tkachov, 9601308\]](#)  
[\[P.S. Cherzor, N.A. Sveshnikov, 9710349\]](#)

Energy flow  $\Leftrightarrow$  Infrared and Collinear (IRC) Safe information

# When are two jets similar?

*Energy flow is robust information*



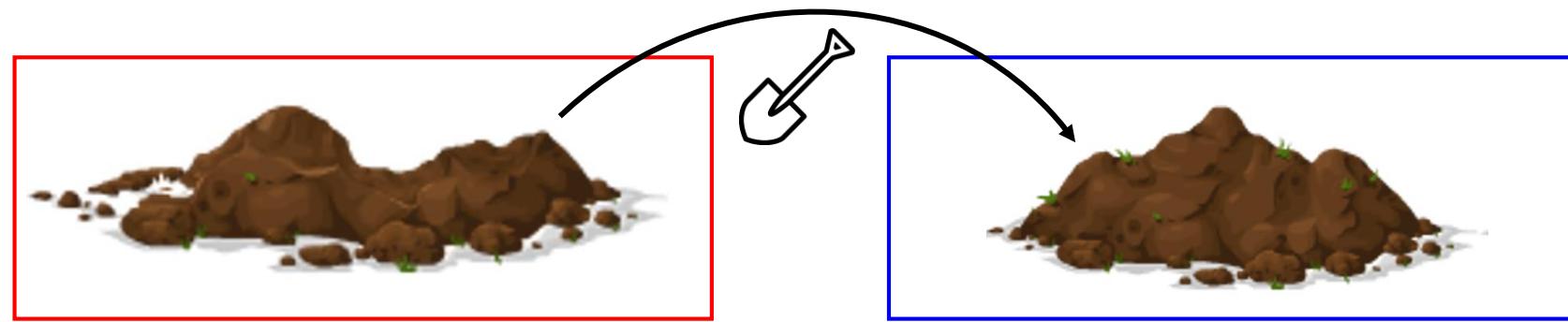
# The Energy Mover's Distance

Review: *The Earth Mover's Distance*

**Earth Mover's Distance:** the minimum “work” (stuff  $\times$  distance) to rearrange one pile of dirt into another

[Peleg, Werman, Rom]

[Rubner, Tomasi, Guibas]



Metric on the space of (normalized) distributions: *symmetric, non-negative, triangle inequality*

Distributions are close in EMD  $\Leftrightarrow$  their expectation values are close.

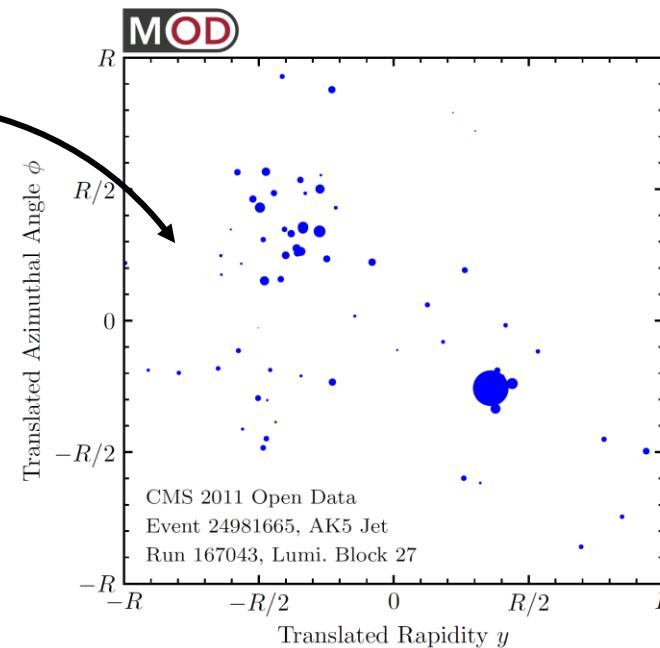
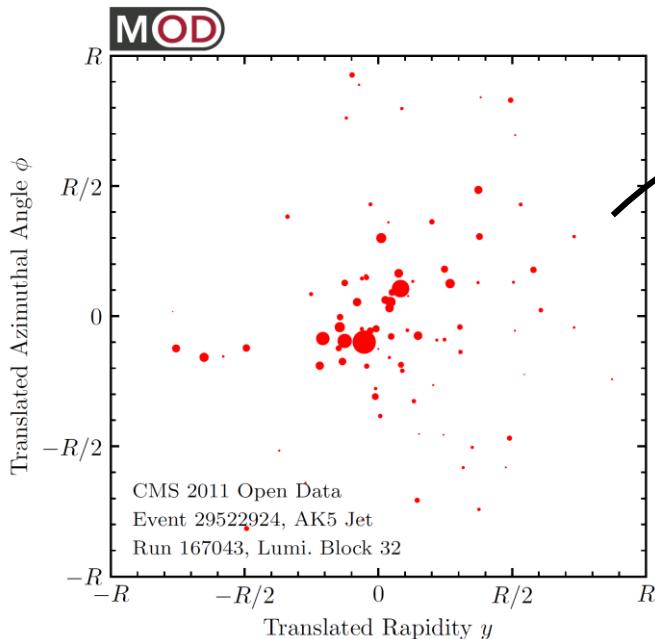
Also known as the 1-Wasserstein metric.

# The Energy Mover's Distance

From Earth to Energy

**Energy Mover's Distance:** the minimum “work” (**energy**  $\times$  angle) to rearrange one jet (pile of energy) into another

[Komiske, EMM, Thaler, 1902.02346]



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

Difference in radiation pattern      Difference in total energy

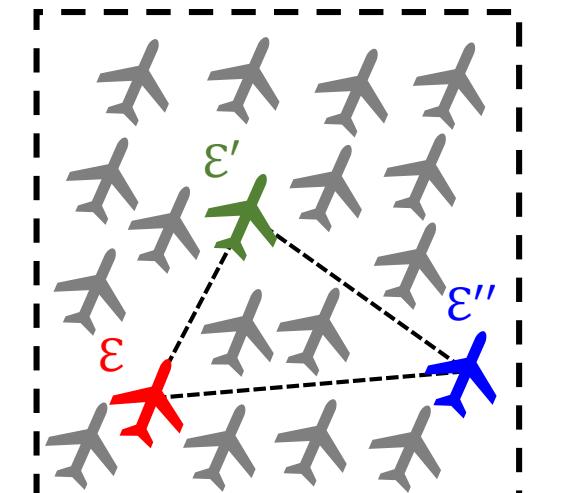
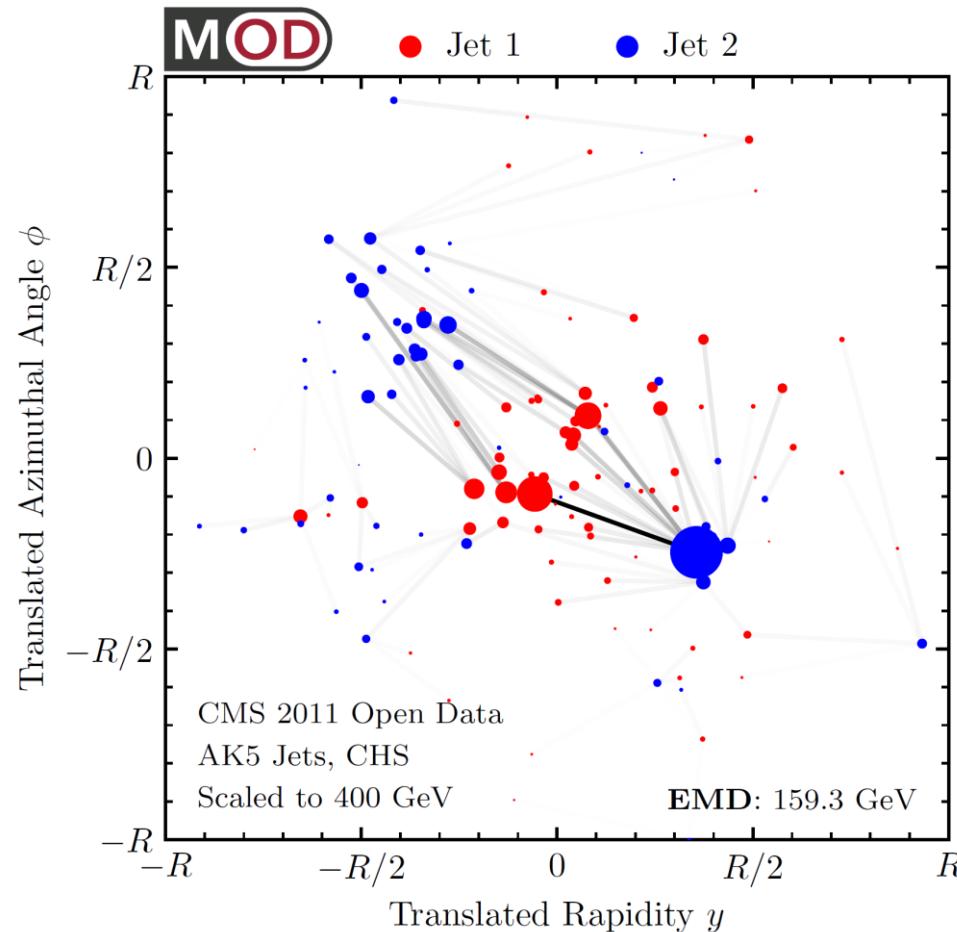
$E_i$   $\theta_{ij}$   $f_{ij}$   $E'_j$

# The Energy Mover's Distance

From Earth to Energy

**Energy Mover's Distance:** the minimum “work” (**energy**  $\times$  angle) to rearrange one event (pile of energy) into another

[Komiske, EMM, Thaler, 1902.02346]



$$\text{EMD}(\varepsilon, \varepsilon') + \text{EMD}(\varepsilon', \varepsilon'') \geq \text{EMD}(\varepsilon, \varepsilon'')$$

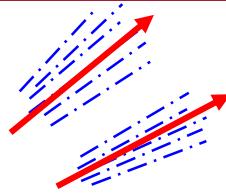
EMD has dimensions of energy

True metric as long as  $R \geq \frac{1}{2}\theta_{\max}$   
 $R \geq$  the jet radius, for conical jets

Solvable via Optimal Transport problem.

~1ms to compute EMD for two jets with 100 particles.

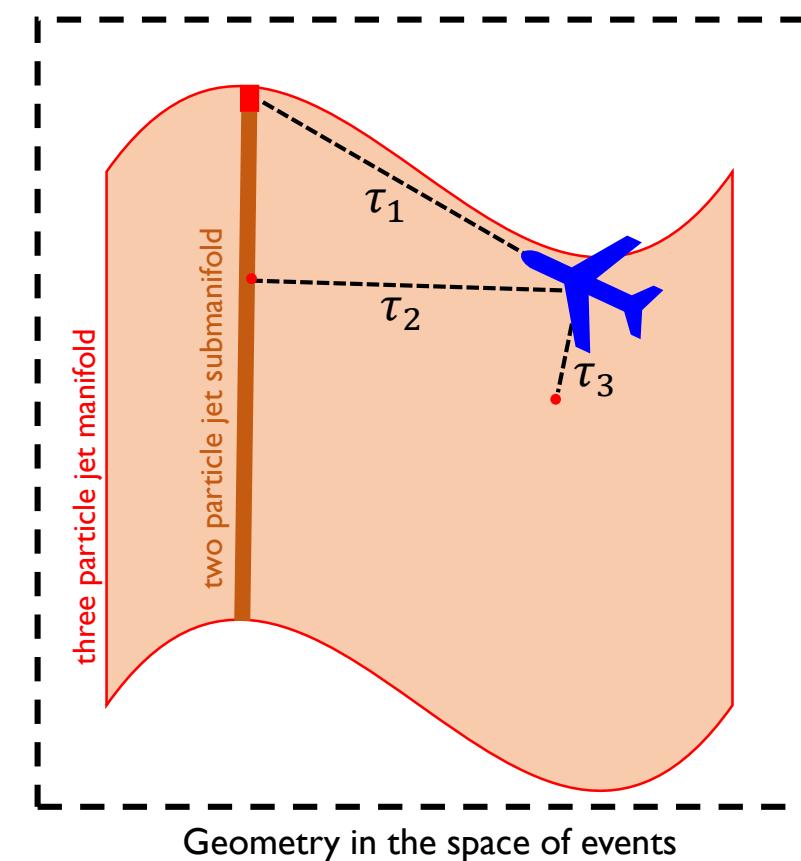
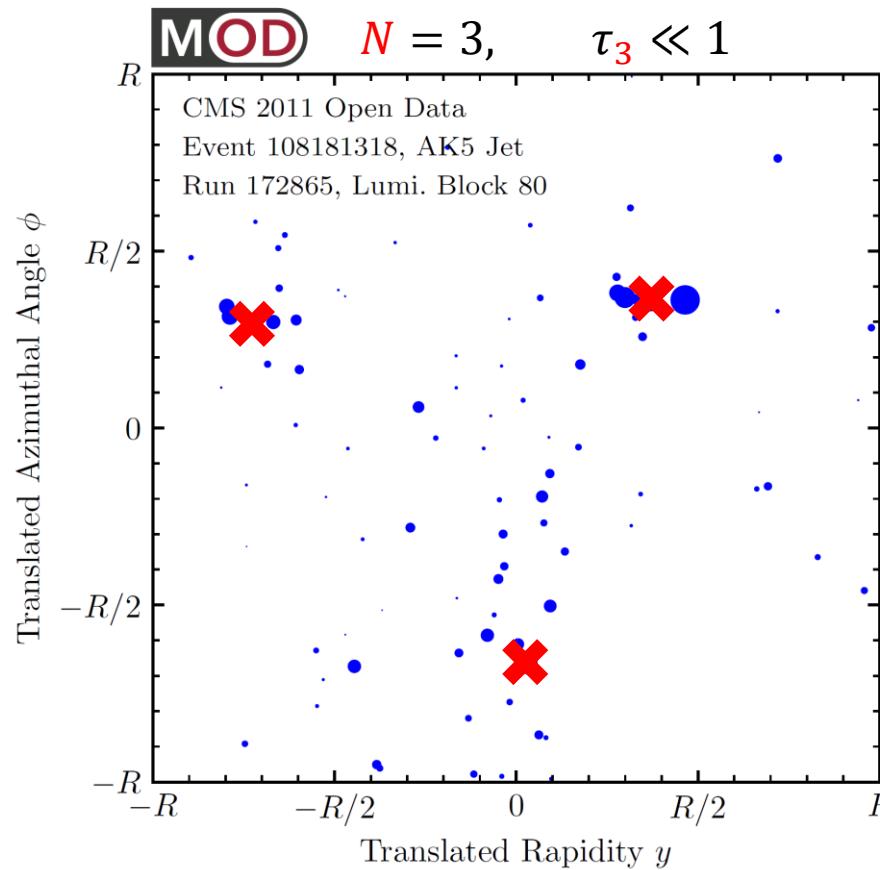
# A Geometric Language for Observables



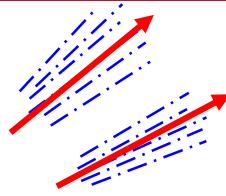
**$N$ -(sub)jettiness** is the EMD between the event and the closest  $N$ -particle event.

$$\tau_N(\mathcal{E}) = \min_{N \text{ axes}} \sum_{i=1}^M E_i \min\{\theta_{1,i}^\beta, \theta_{2,i}^\beta, \dots, \theta_{N,i}^\beta\} \longrightarrow \tau_N(\mathcal{E}) = \min_{|\mathcal{E}'|=N} \text{EMD}(\mathcal{E}, \mathcal{E}').$$

$\beta$ -Wasserstein distance



# A Geometric Language for Observables



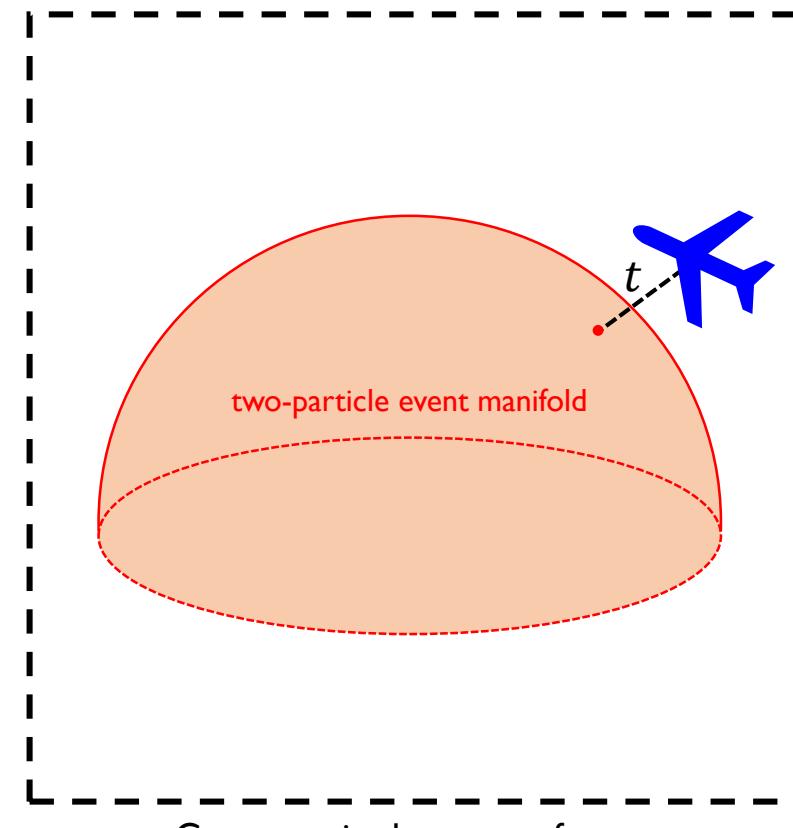
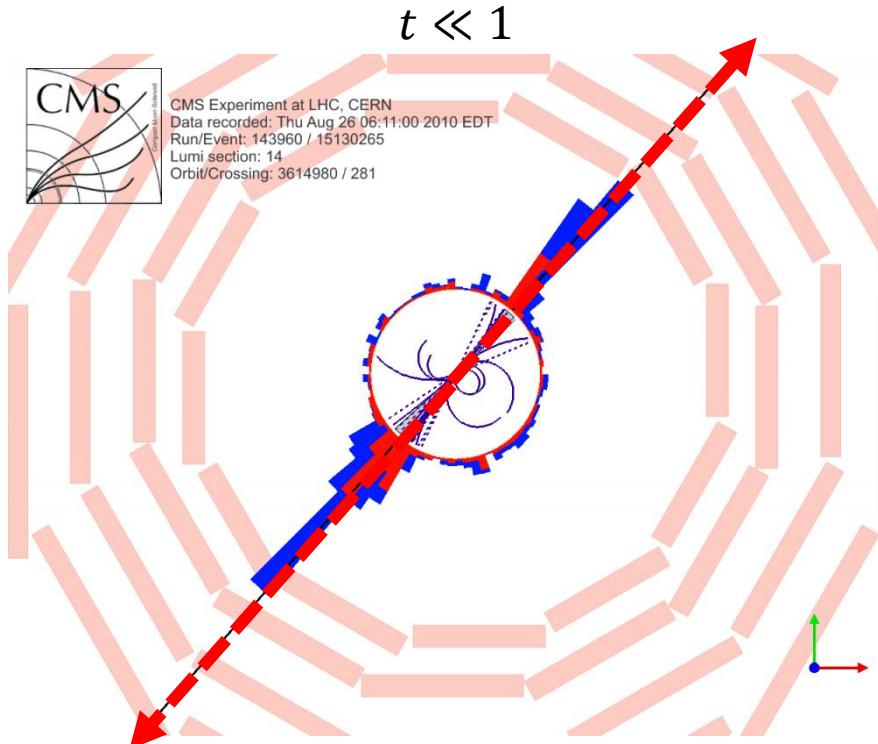
**Thrust** is the EMD between the **event** and the closest **two-particle** event.

$$t(\mathcal{E}) = E - \max_{\hat{n}} \sum_i |\vec{p}_i \cdot \hat{n}|$$

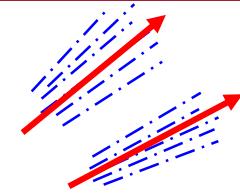


$$t(\mathcal{E}) = \min_{|\mathcal{E}'|=2} \text{EMD}(\mathcal{E}, \mathcal{E}')$$

$$\text{with } \theta_{ij} = \hat{n}_i \cdot \hat{n}_j, \quad \hat{n} = \vec{p}/E$$



# A Geometric Language for Observables

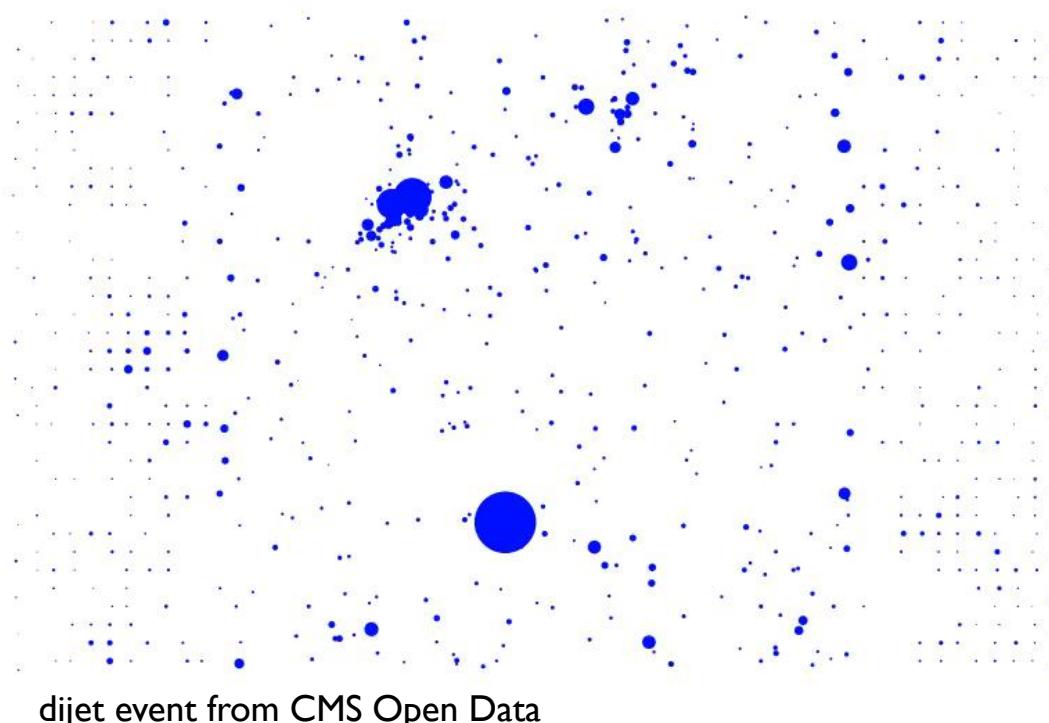


**Isotropy** is a new observable to probe how “uniform” an event is.

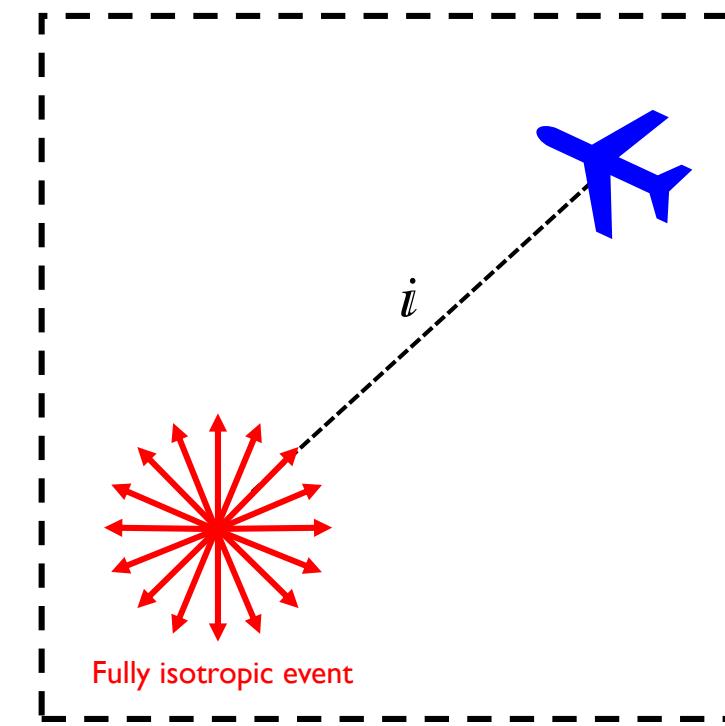
It is sensitive to very different new physics signals than existing event shapes.

e.g. uniform radiation from micro black holes [\[Cari Cesarotti and Jesse Thaler, coming soon!\]](#)

$$i(\mathcal{E}) = \text{EMD}(\mathcal{E}, \mathcal{E}_{\text{iso}}) \text{ where } \mathcal{E}_{\text{iso}} \text{ is a fully isotropic event}$$

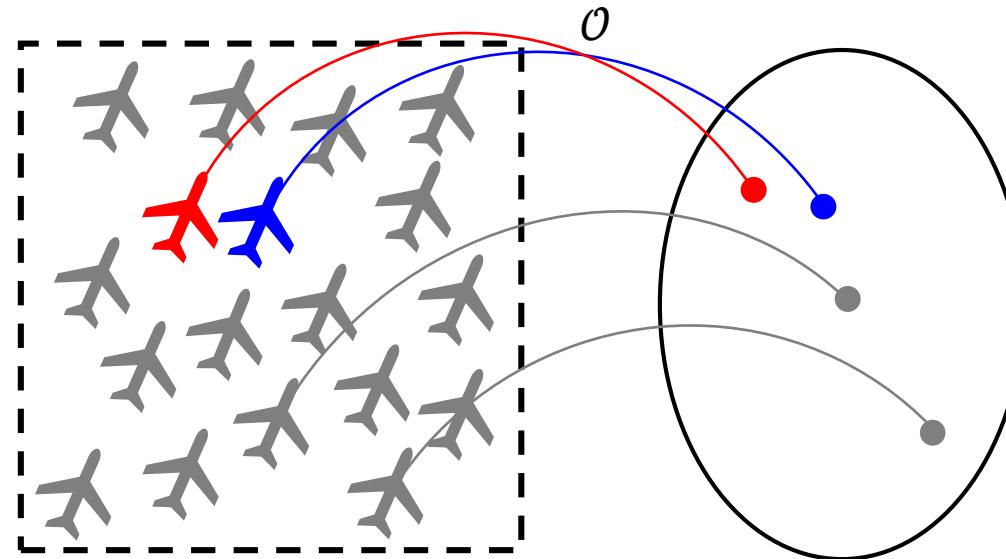


dijet event from CMS Open Data



# A Geometric Language for Observables

Events close in EMD are close in any infrared and collinear safe observable!



Additive IRC-safe observables:  $\mathcal{O}(\mathcal{E}) = \sum_{i=1}^M \textcolor{red}{E}_i \Phi(\hat{n}_i)$

Energy Mover's  
Distance

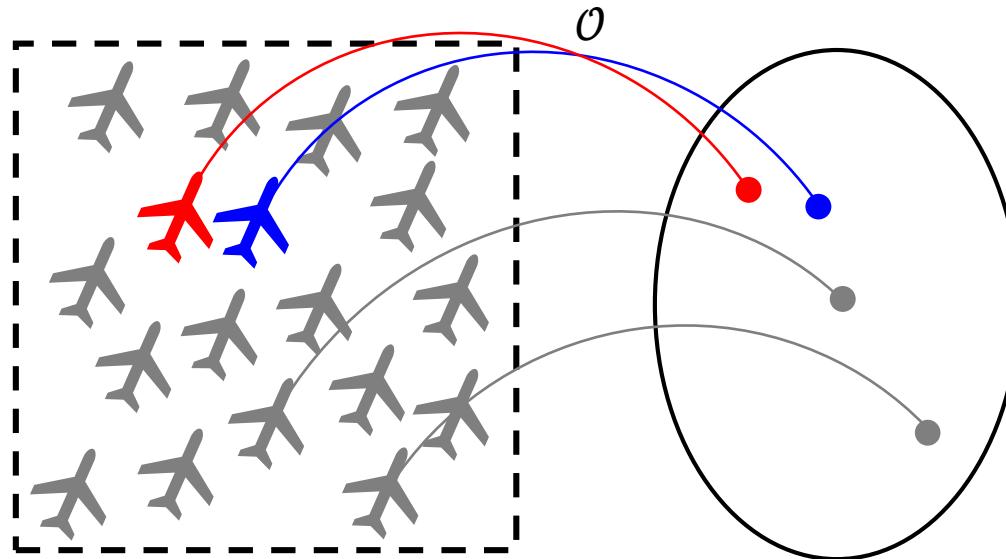
$$\text{EMD}(\mathcal{E}, \mathcal{E}') \geq \frac{1}{RL} |\mathcal{O}(\mathcal{E}) - \mathcal{O}(\mathcal{E}')|$$

“Lipschitz constant” of  $\Phi$   
i.e. bound on its derivative

Difference in  
observable values

# A Geometric Language for Observables

Events close in EMD are close in any infrared and collinear safe observable!



Jet angularities with  $\beta \geq 1$ :

[\[C. Berger, T. Kucs, and G. Sterman, 0303051\]](#)

[\[A. Larkoski, J. Thaler, and W. Waalewijn, 1408.3122\]](#)

$$\lambda^{(\beta)} = \sum_{i=1}^M E_i \theta_i^\beta$$

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

# Exploring the Space of Jets: Visualizing the Manifold

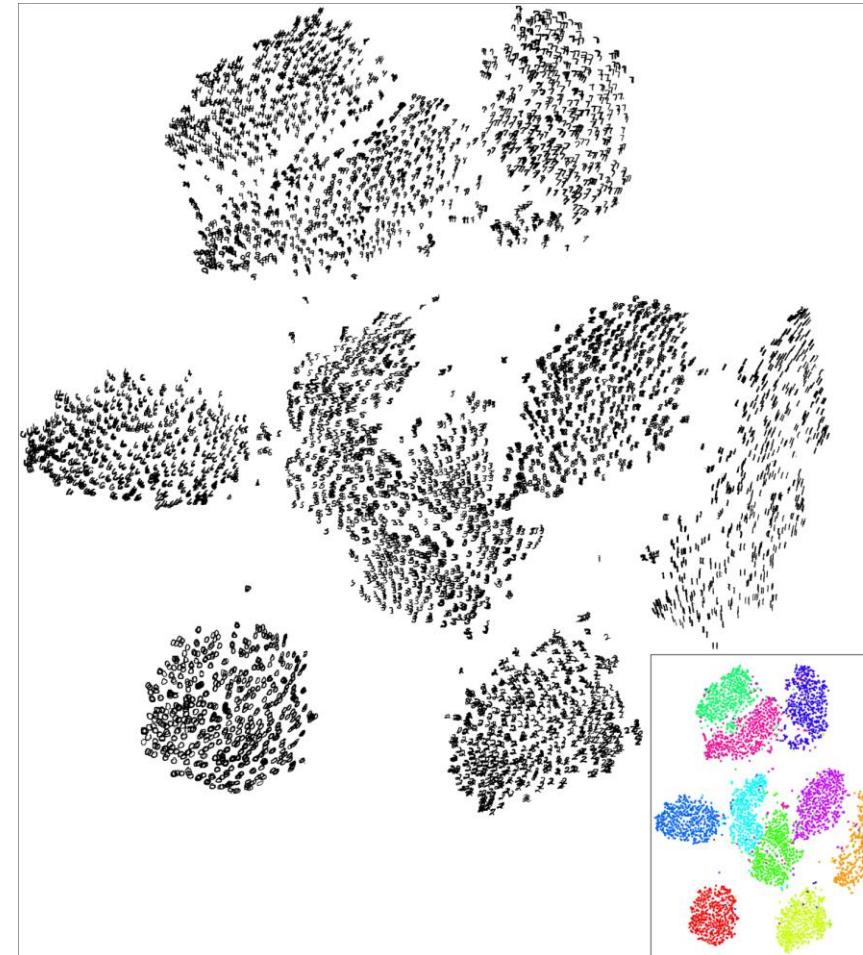
Visualize the space of events with t-Distributed Stochastic Neighbor Embedding (t-SNE).

[\[L. van der Maaten, G. Hinton\]](#)

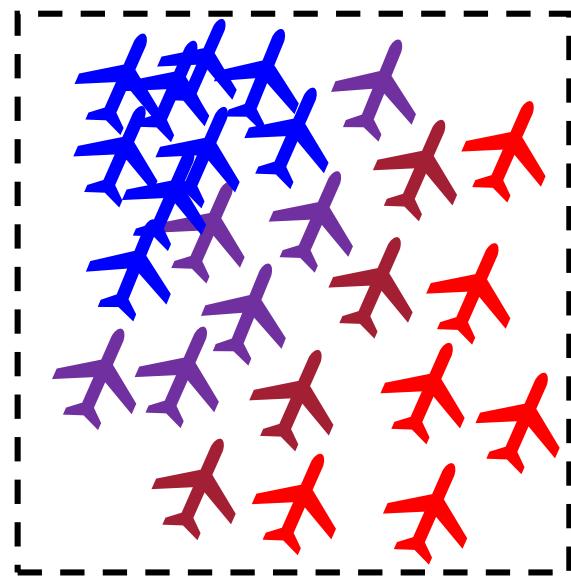
Finds an embedding into a low-dimensional manifold that respects distances.



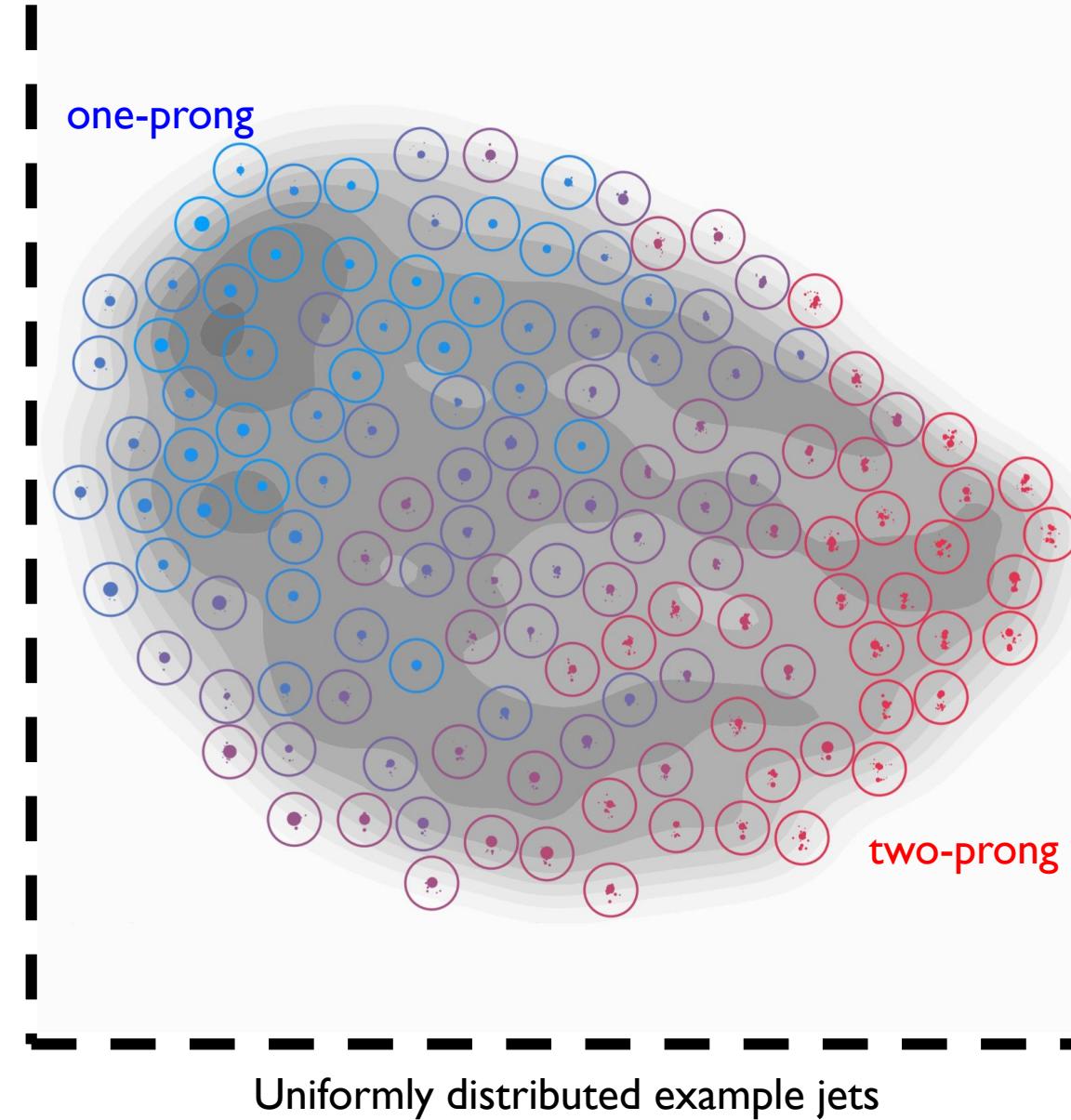
What does the space  
of jets look like?



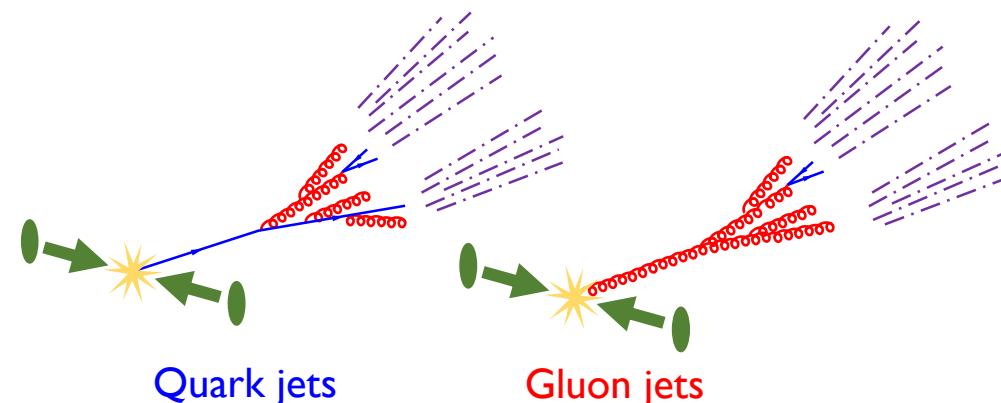
# Exploring the Space of Jets: Visualizing the Manifold



What does the space  
of jets look like?



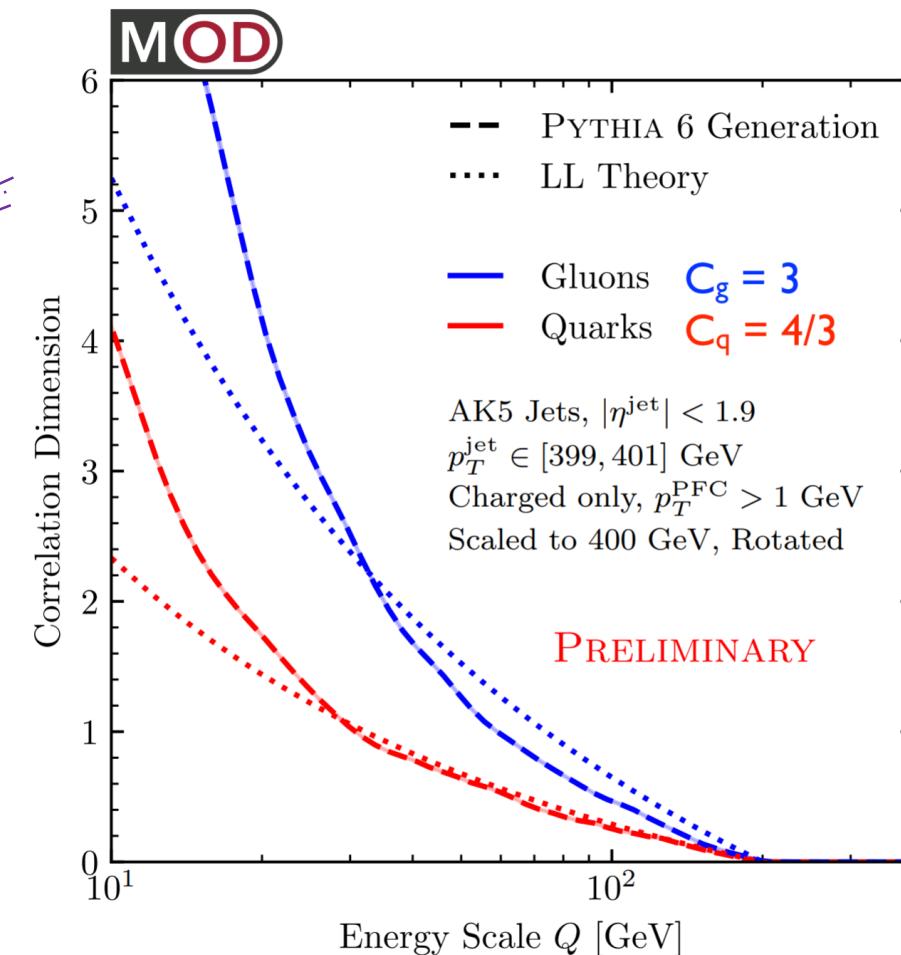
# Exploring the Space of Jets: Correlation Dimension



$$\text{At LL: } \dim_{q/g}(Q) = -\frac{8\alpha_s C_{q/g}}{\pi} \ln \frac{Q}{p_T/2}$$

+ 1-loop running of  $\alpha_s$

$$C_q = C_F = \frac{4}{3}$$
$$C_g = C_A = 3$$



See extra slides for sketch of calculation.