Data ex Machina

Machine Learning with Public Collider Data

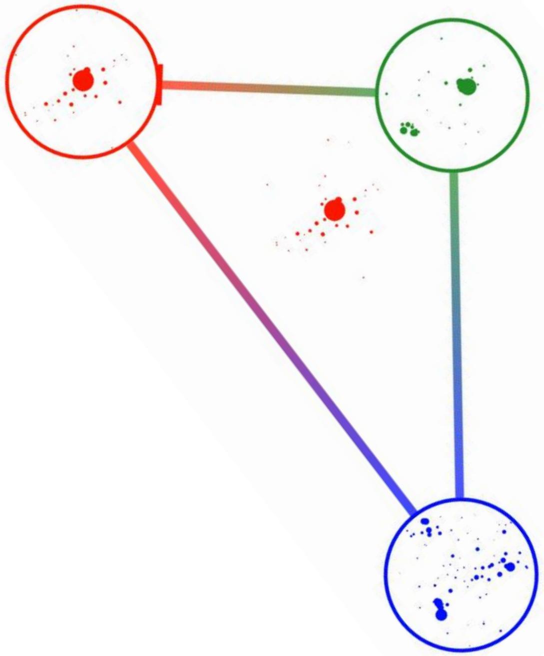
AI & Physics, Applied Machine Learning Days 2020

Eric M. Metodiev

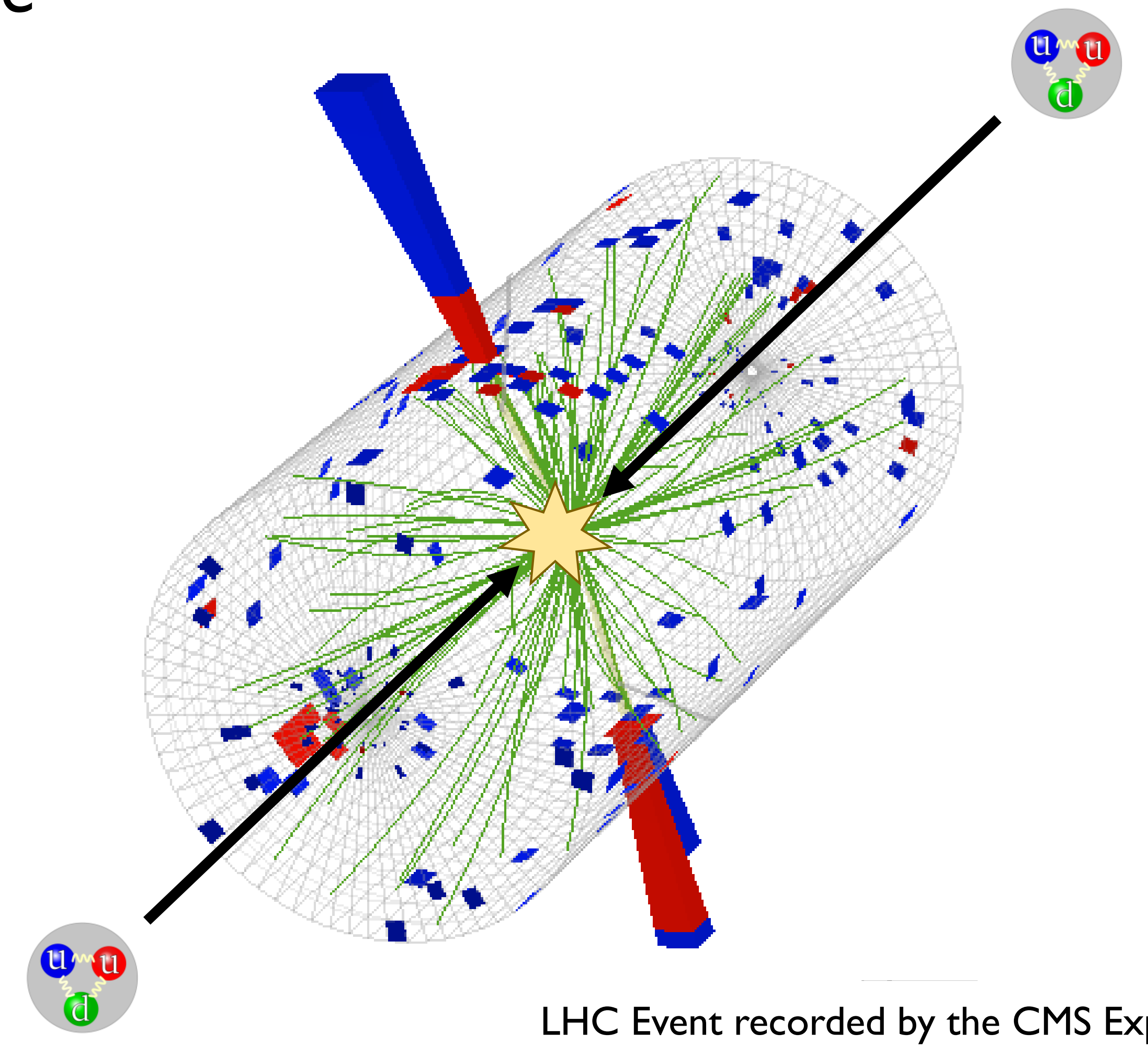
Center for Theoretical Physics

Massachusetts Institute of Technology

Patrick Komiske | Radha Mastandrea | Preksha Naik | Jesse Thaler

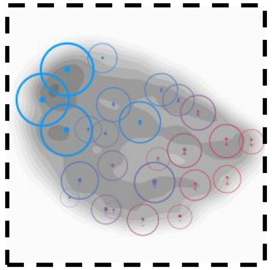


Collision Course

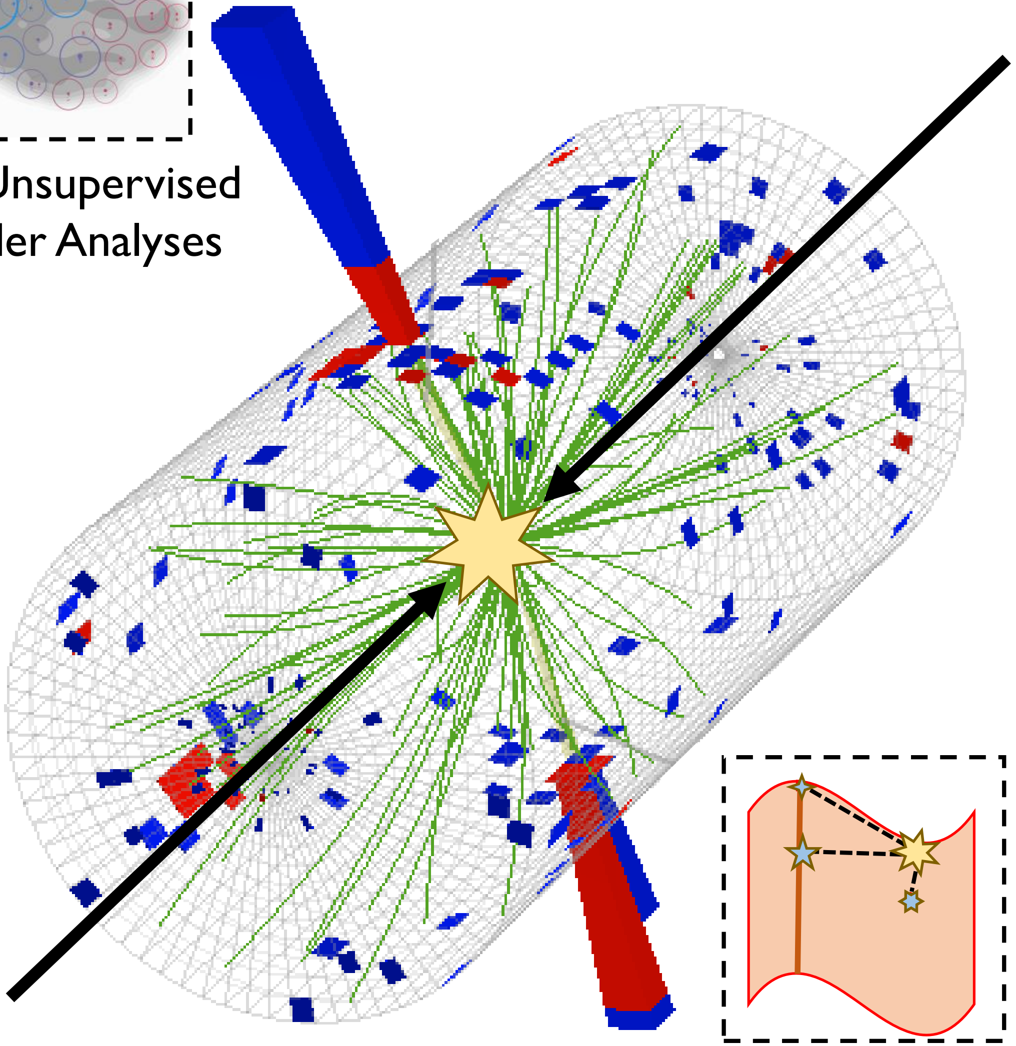


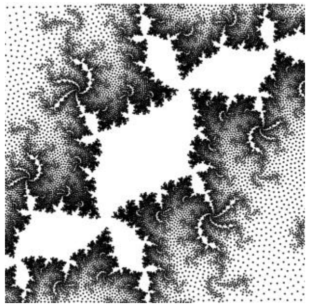
LHC Event recorded by the CMS Experiment at CERN

Collision Course



New Unsupervised Collider Analyses





Optimal Transport
[OTML Workshop, NeurIPS 2019]

Public Collider Data
[opendata.cern.ch]

New Insights into Quantum Field Theory

[h/t Jesse Thaler]

opendata CERN

About

opendata.cern.ch

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

jet primary dataset

Search

search examples: collision datasets, keywords:education, energy:7TeV

Explore

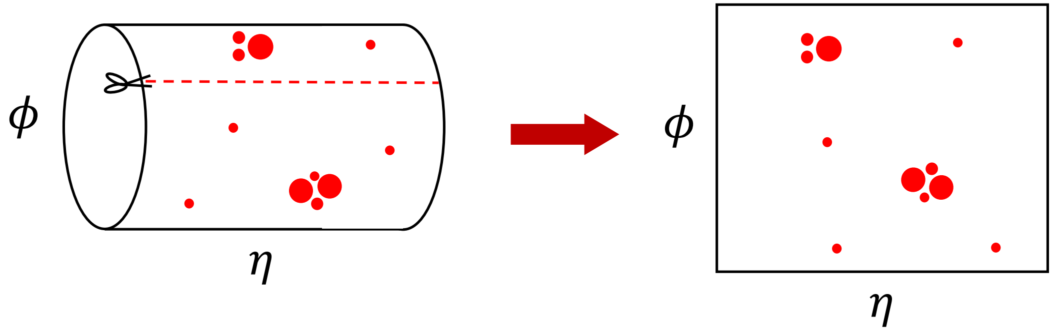
datasets
software
environments
documentation

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

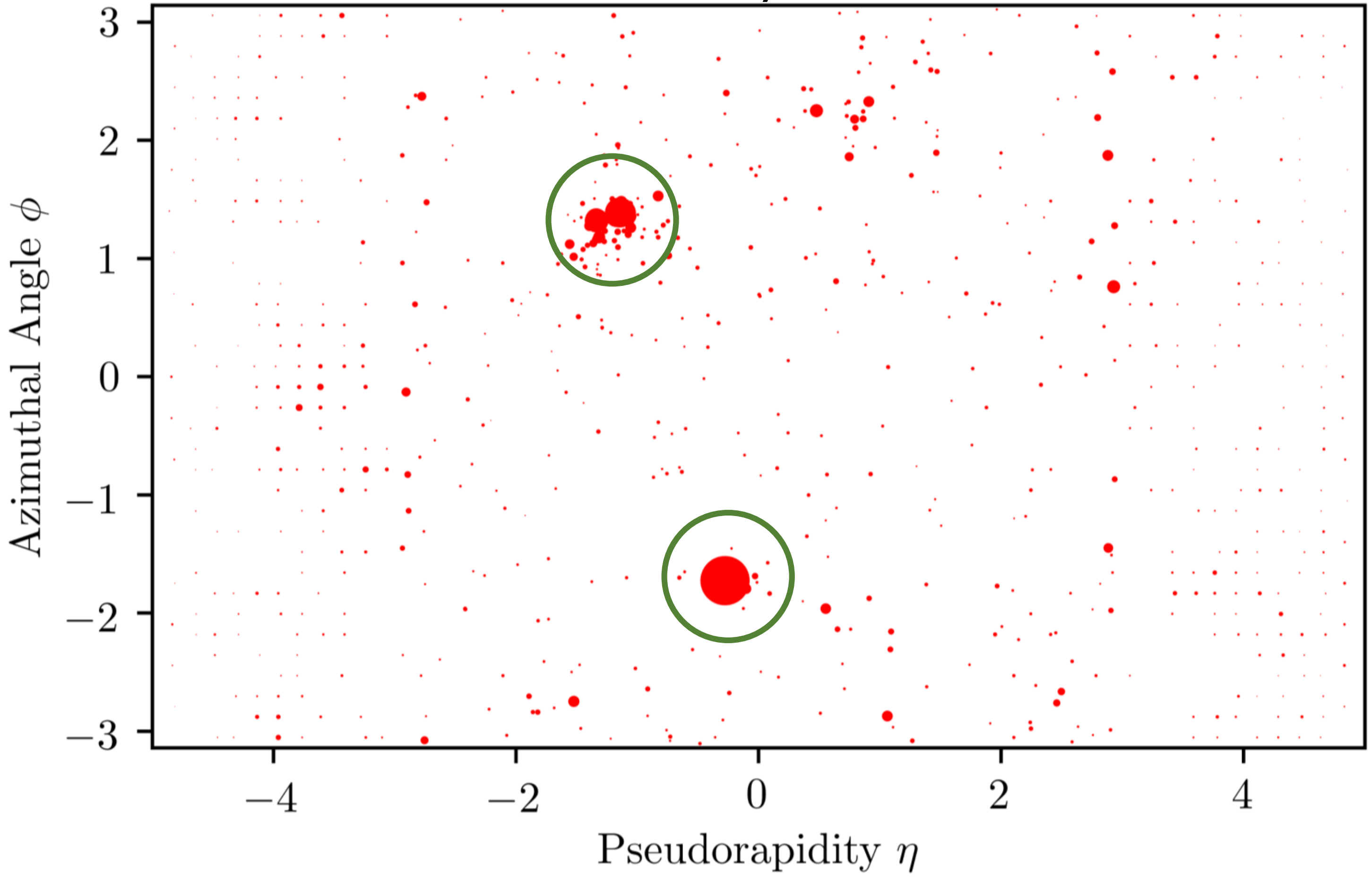
Fifteen lines of code later…

```
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!

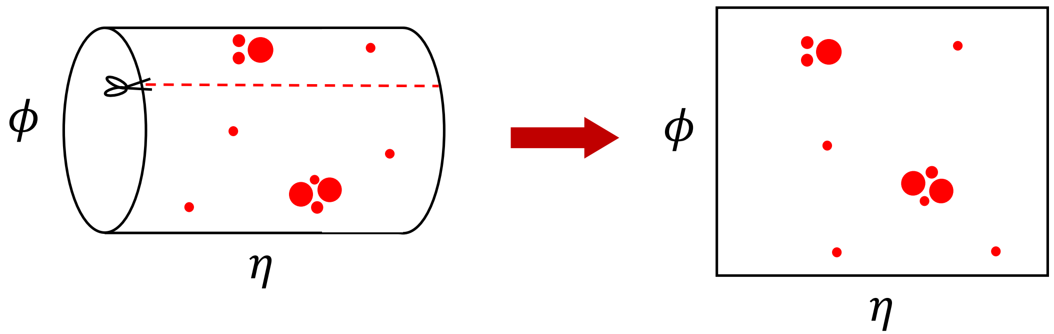
A real collision event recorded by CMS!

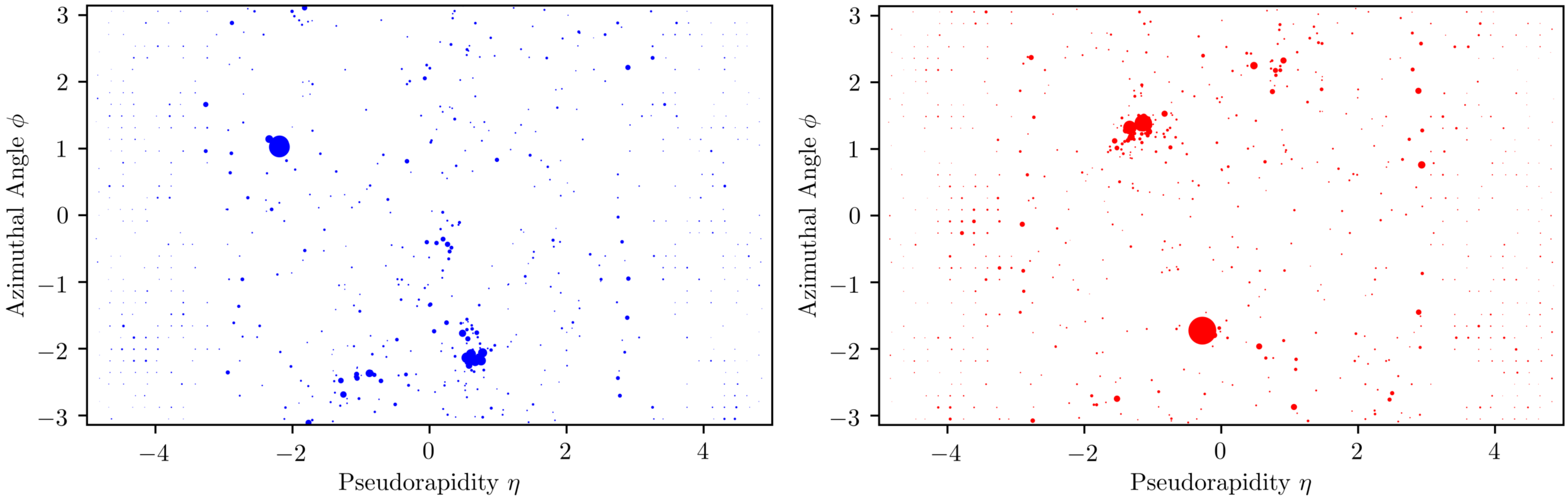


When are two collisions similar?

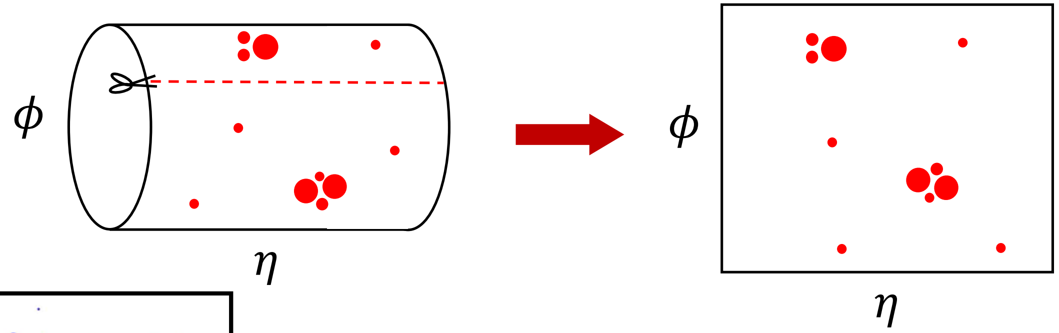
Many unsupervised methods rely on a **distance matrix.**

Need a physically-sensible **metric** between events!

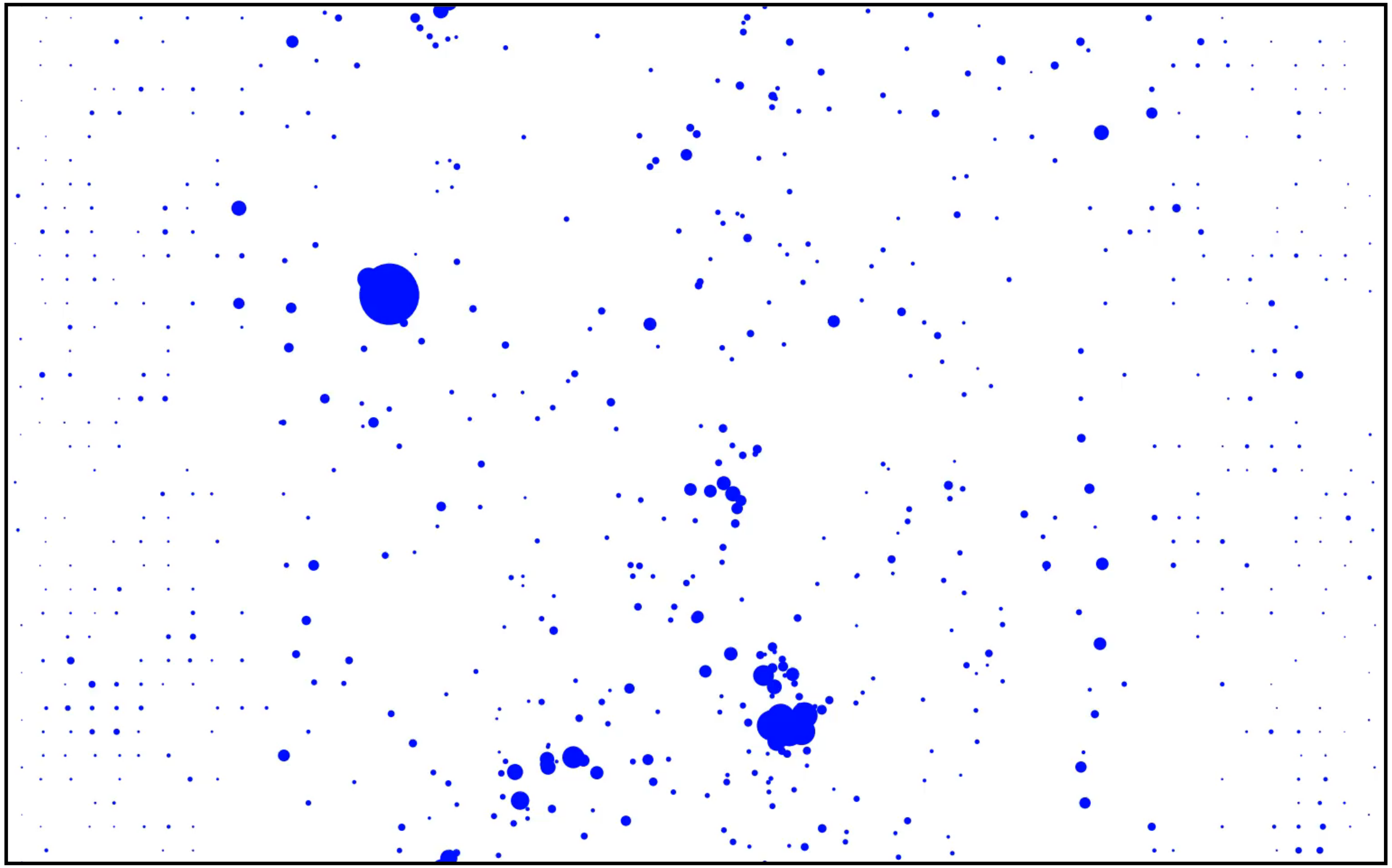




When are two collisions similar?



The Earth Mover's (or Wasserstein) Distance



The "work" required to rearrange one collision event into another!

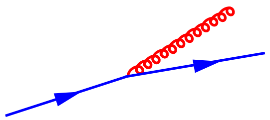
Plus a cost to create or destroy energy.

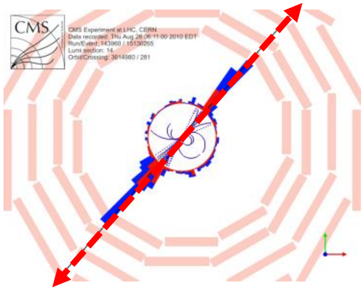
Optimal Transport Problem

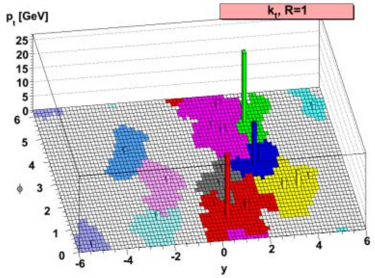
Here using [python optimal transport]

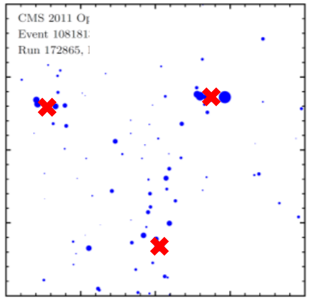
[Komiske, **EMM**, Thaler, PRL 2019]

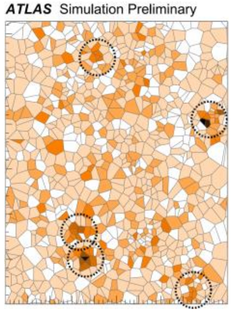
Six Decades of Collider Techniques











Taming infinities | Event Shapes | Jet Algorithms | Jet Substructure | Pileup

1960 — 2020

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

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

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

1997-1998
C/A jet clustering
[Wobisch, Wengler, 1998]
[Doskhitzer, Leder, Moretti, Webber, JHEP 1997]

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

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

And many more!

Six Decades of Collider Techniques as Optimal Transport!

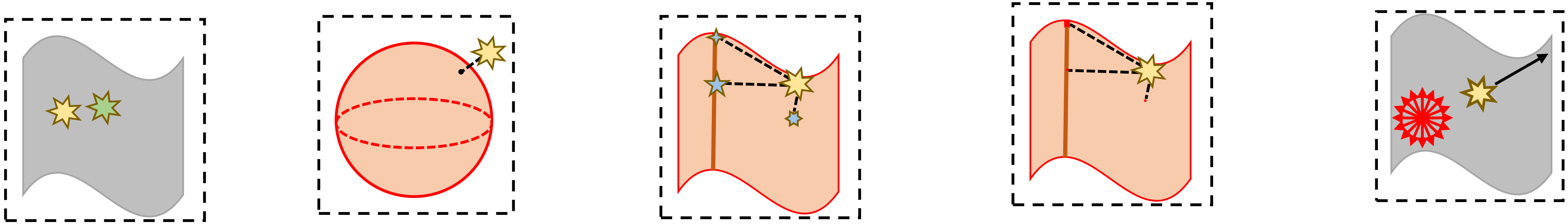
[Komiske, **EMM**, Thaler, *to appear*]

Smooth function of energy distribution are finite in QFT

Event shapes as distances to the 2-particle manifold

Jets are N-particle event approximations

Subtract a pileup as a uniform distribution



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

$t(\mathcal{E}) = \min\limits_{|\mathcal{E}'|=2} \mathrm{EMD}(\mathcal{E}, \mathcal{E}')$

$\mathcal{J}(\mathcal{E}) = \operatorname*{argmin}\limits_{|\mathcal{E}'|=N} \mathrm{EMD}(\mathcal{E}, \mathcal{E}')$

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

Taming infinities | Event Shapes | Jet Algorithms | Jet Substructure | Pileup

1960 — 2020

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

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

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

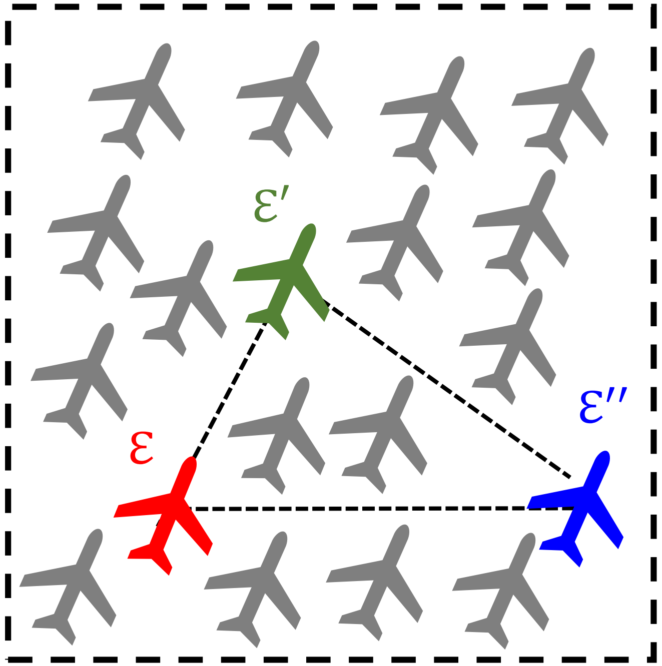
1997-1998
C/A jet clustering
[Wobisch, Wengler, 1998]
[Doskhitzer, Leder, Moretti, Webber, JHEP 1997]

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

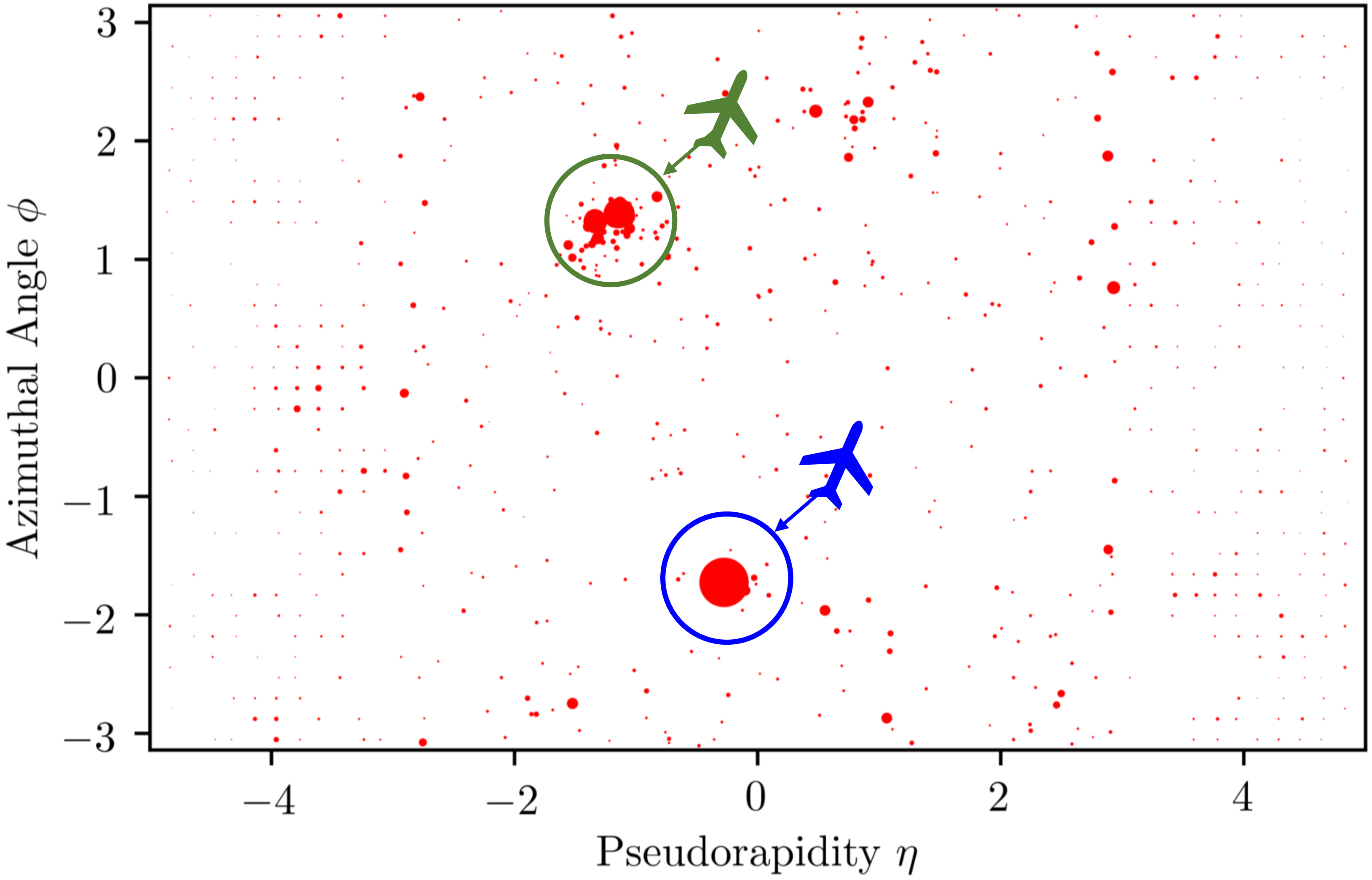
2014-2019
Constituent Subtraction
[Berta, Spousta, Miller, Leitner, JHEP 2014]
[Berta, Masetti, Miller, Spousta, JHEP 2019]

And many more!

Exploring the Space of Jets



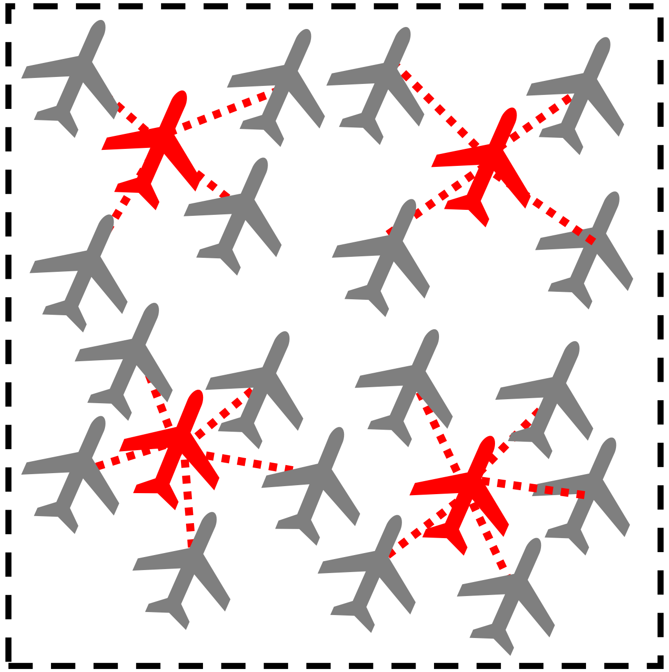
$\mathrm{EMD}(\mathcal{E}, \mathcal{E}') + \mathrm{EMD}(\mathcal{E}', \mathcal{E}'') \geq \mathrm{EMD}(\mathcal{E}, \mathcal{E}'')$



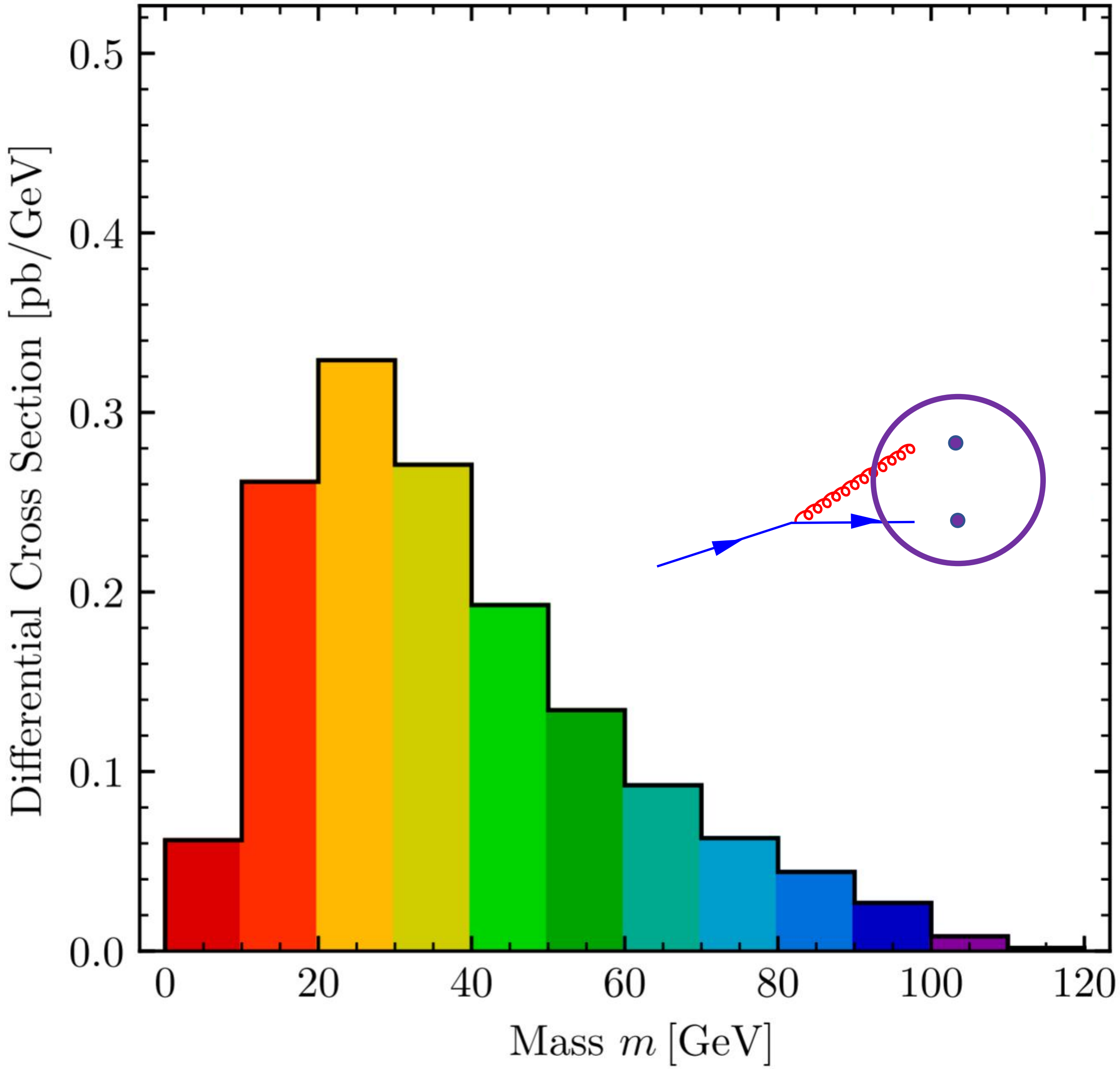
Most Representative Jets

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

Measures how "wide" the jet is.



Jet Mass Histogram

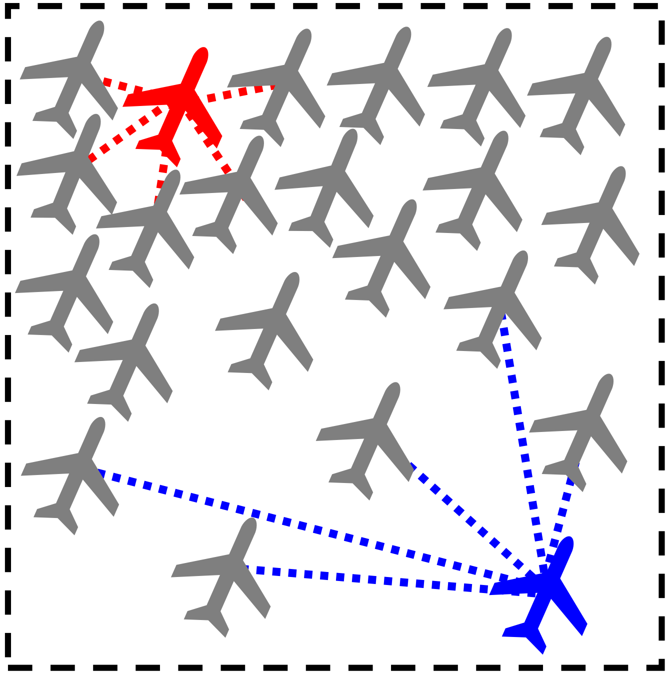


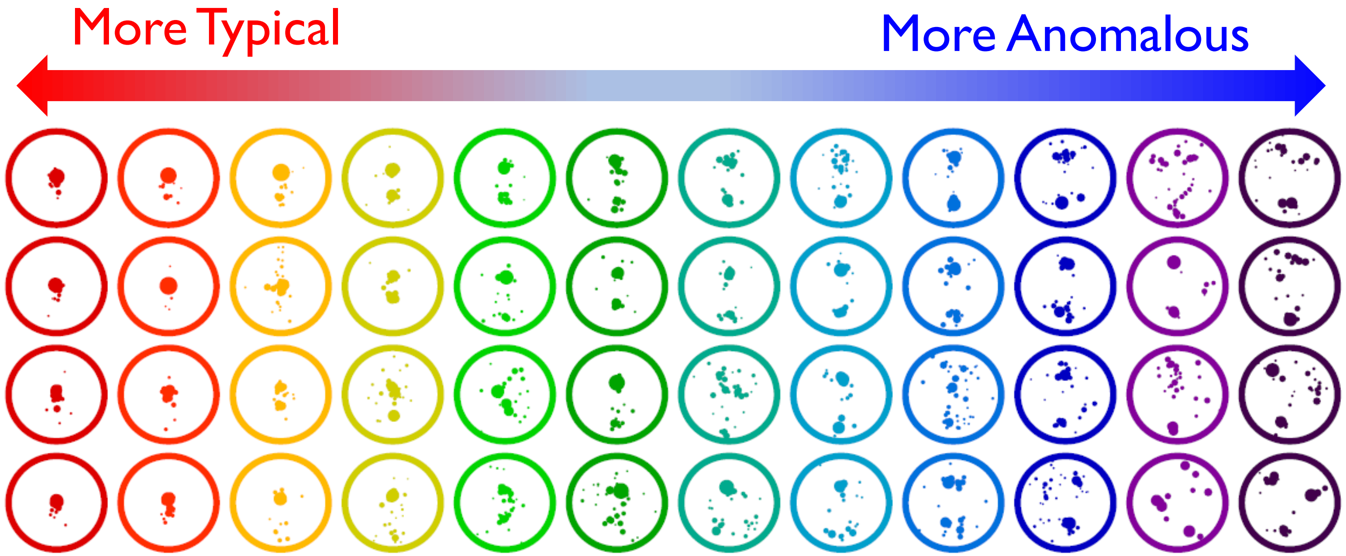
[Komiske, Mastandrea, **EMM**, Naik, Thaler, 1908.08542]

Towards Anomaly Detection

Mean EMD to Dataset:

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



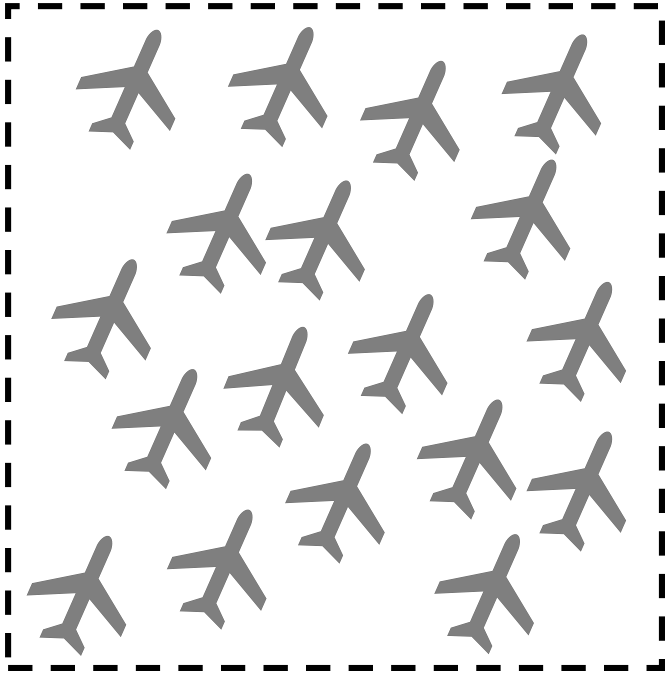


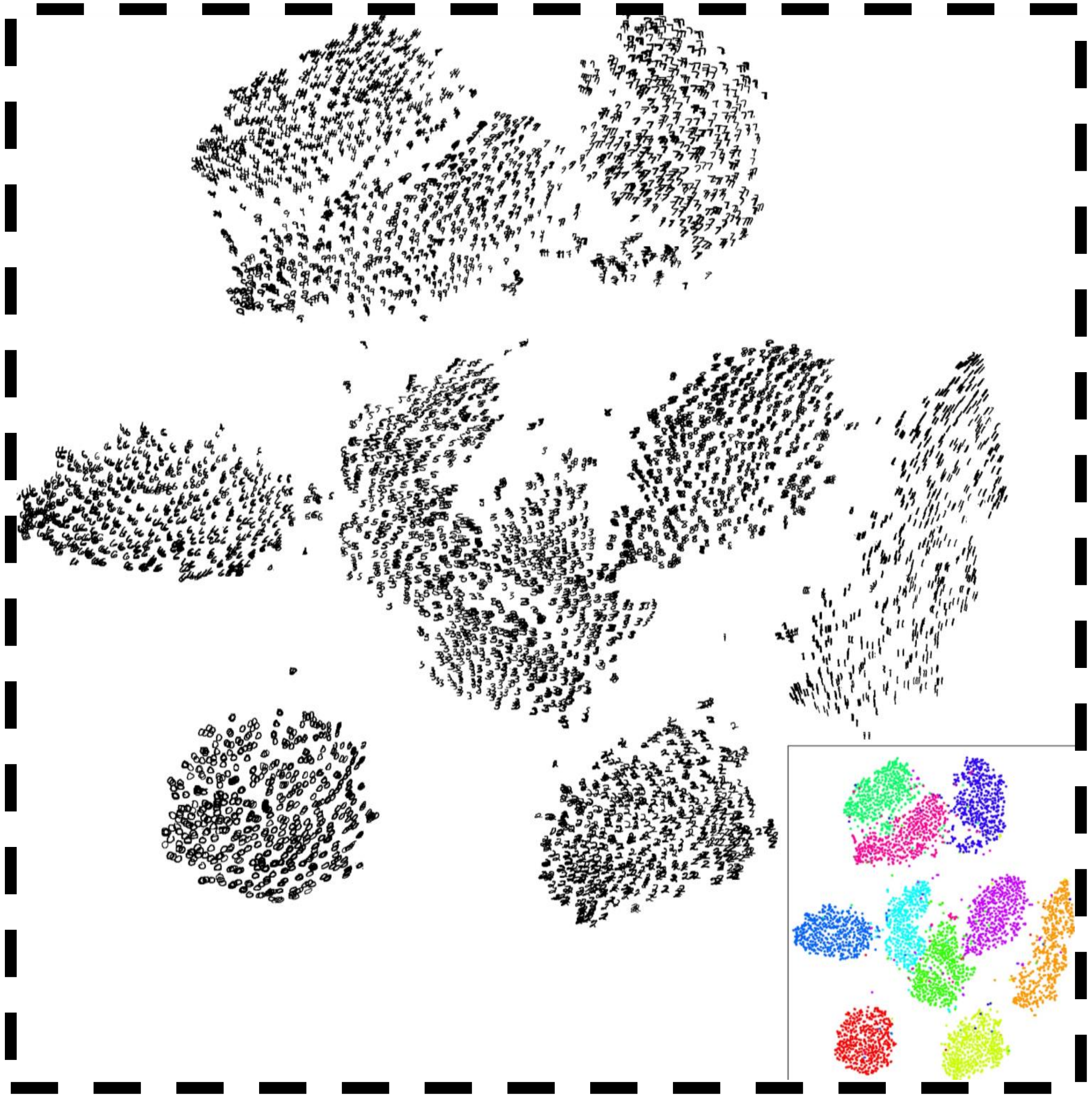
Complements recent developments in anomaly detection for collider physics.

[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?



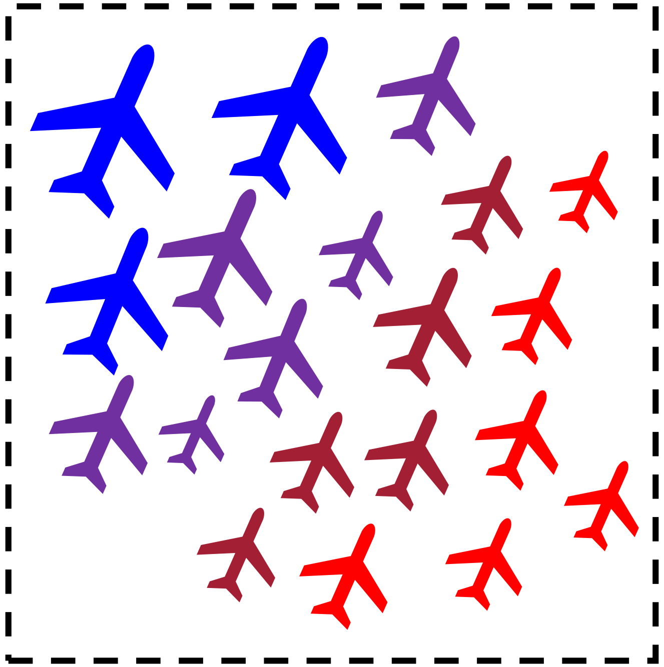


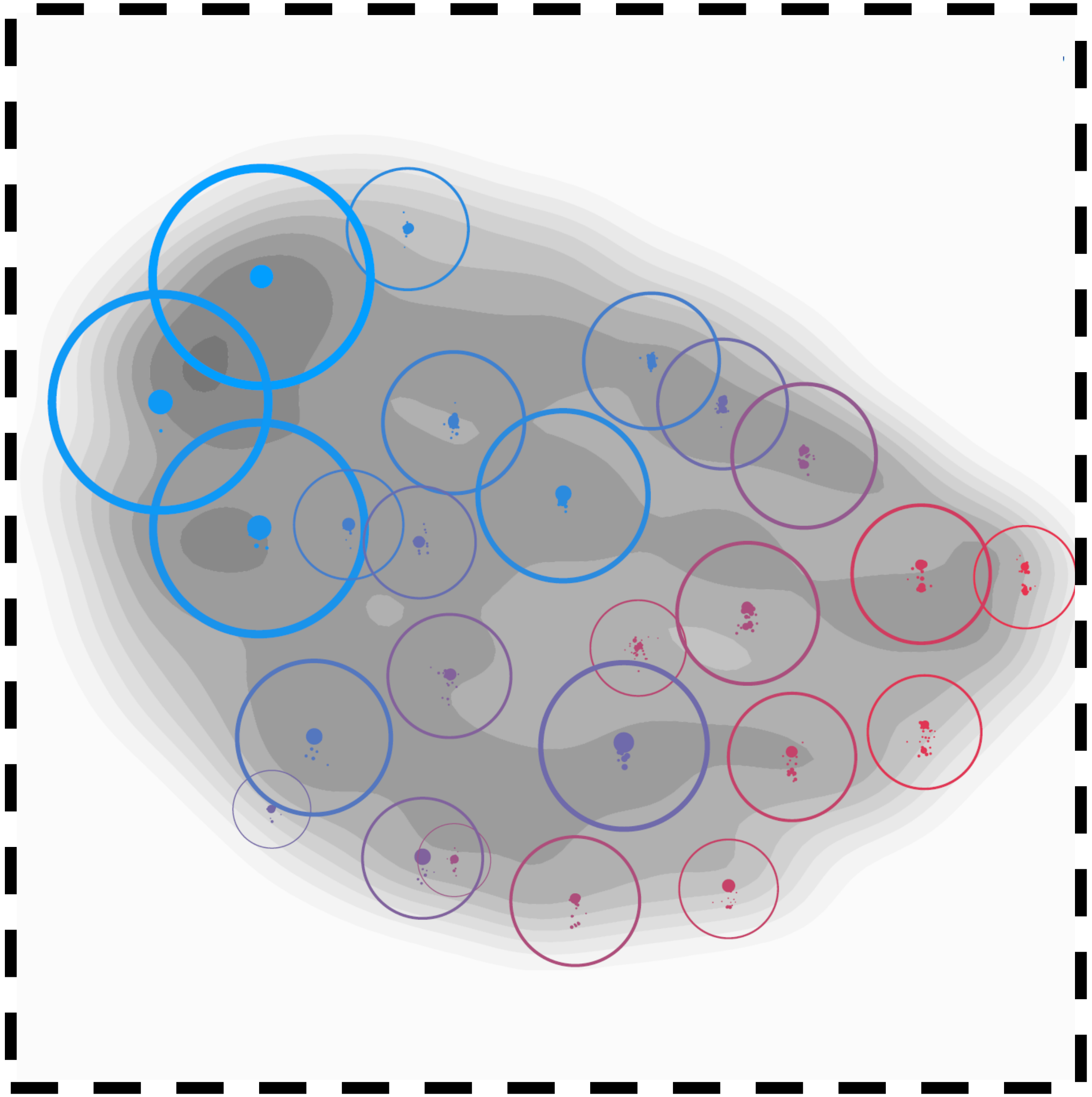
t-SNE embedding

[van der Maaten, Hinton, JMLR 2008]

Visualizing the Manifold

What does the space of jets look like?



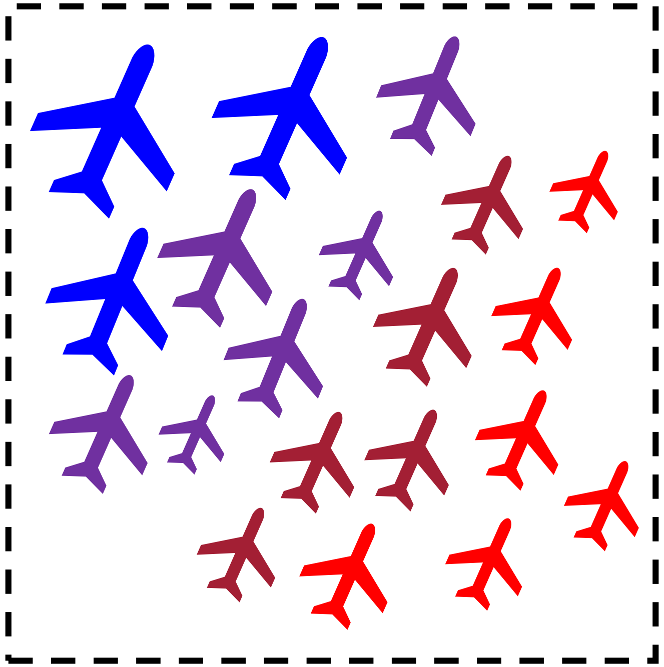


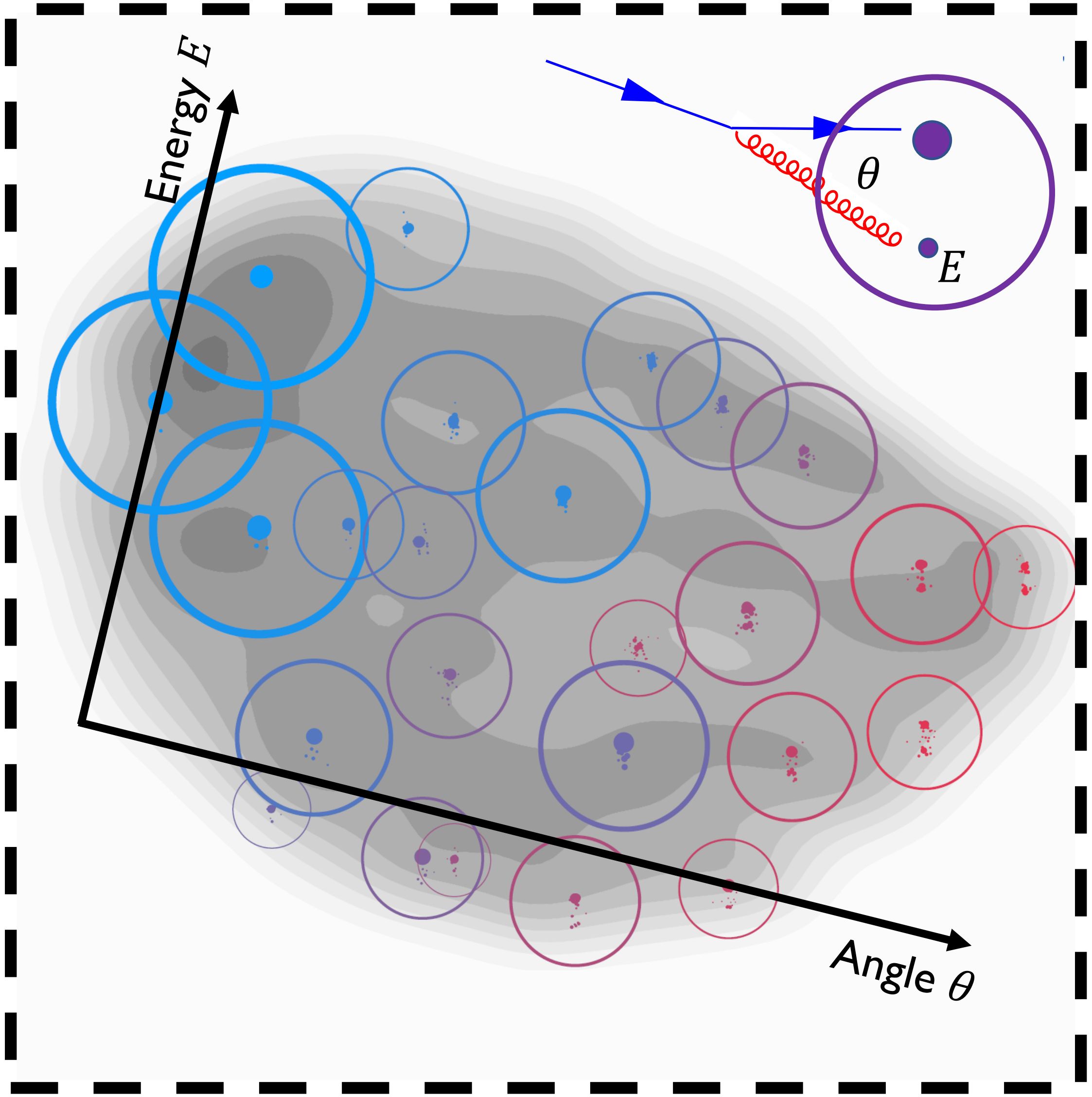
t-SNE embedding: 25-medoid jets shown

[van der Maaten, Hinton, JMLR 2008]
[Komiske, Mastandrea, **EMM**, Naik, Thaler, 1908.08542]

Visualizing the Manifold

What does the space of jets look like?





t-SNE embedding: 25-medoid jets shown

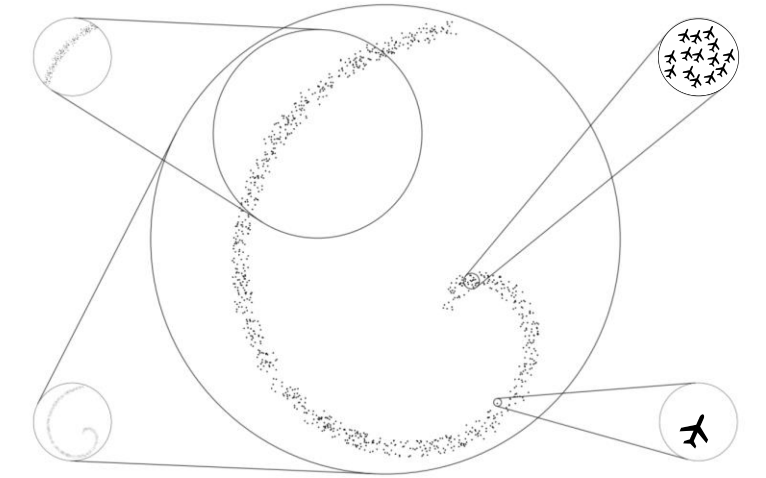
[van der Maaten, Hinton, JMLR 2008]

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

Correlation Dimension

Conceptual Idea

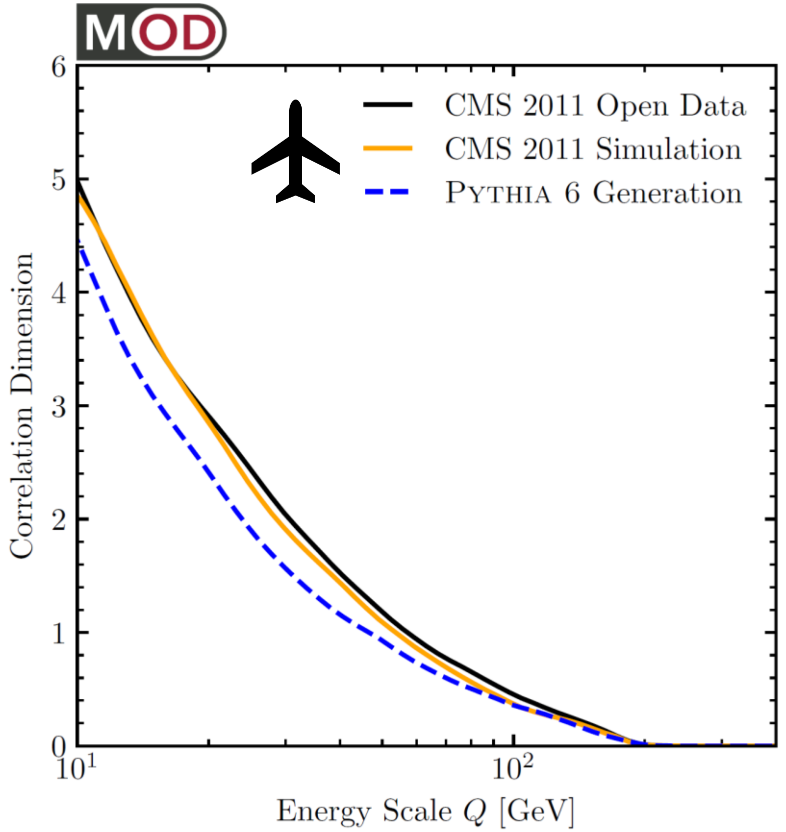
$N_{\text{neighbors}}(r) \propto r^{\dim}$



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

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

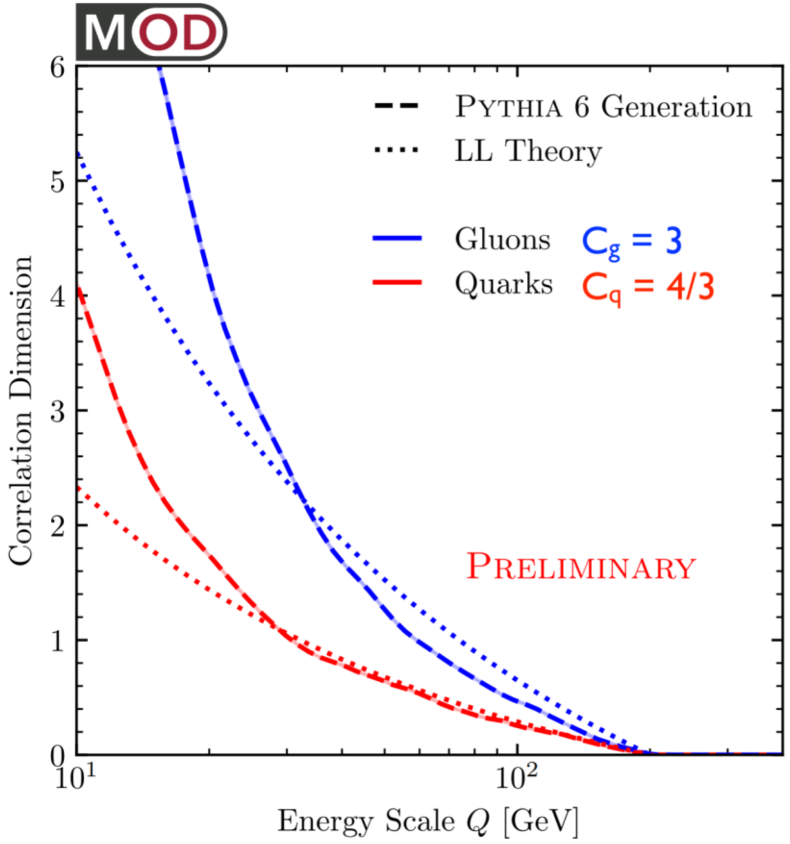
Experimental Data



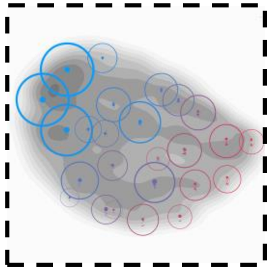
Dimension blows up at low energies.

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

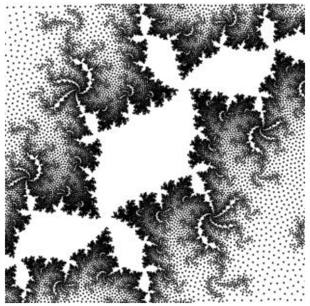
Theoretical Calculation



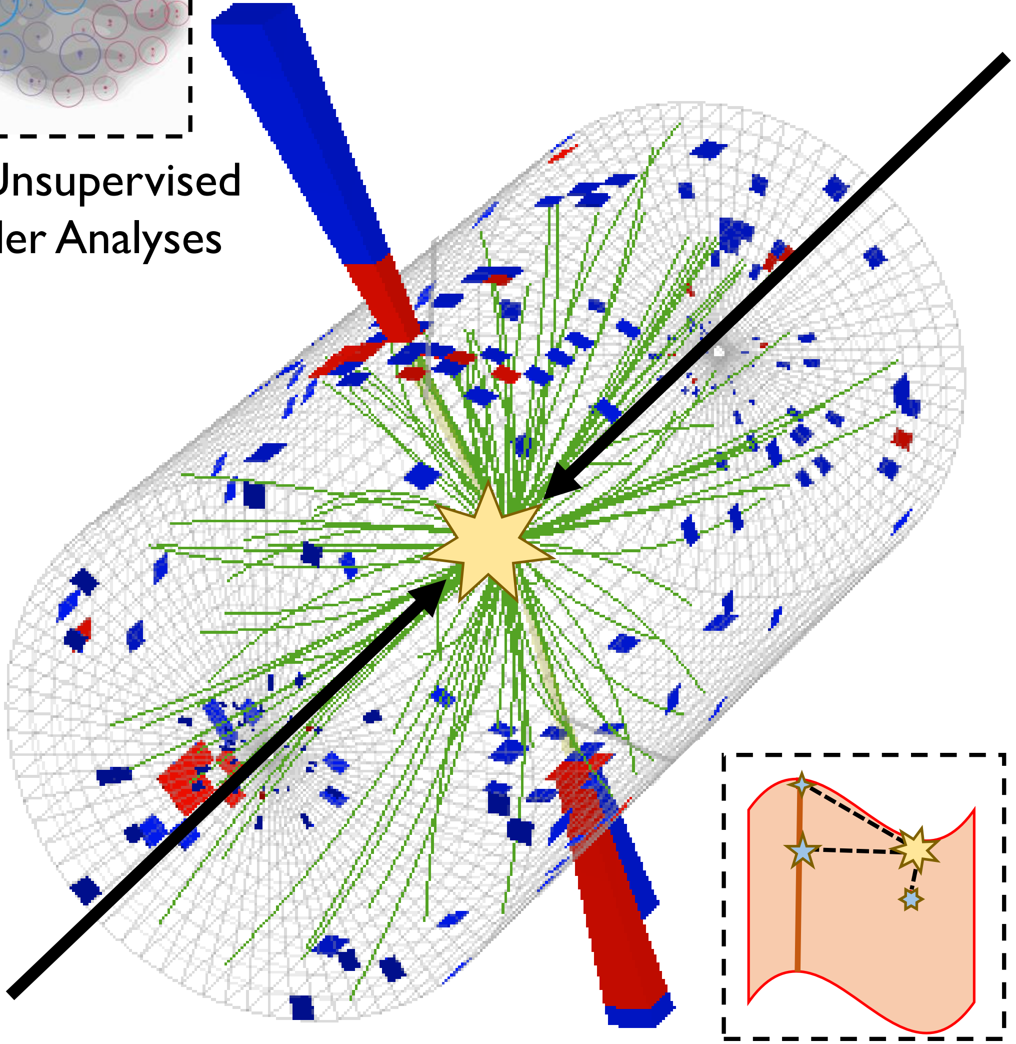
Thank You!



New Unsupervised Collider Analyses



Optimal Transport
[OTML Workshop, NeurIPS 2019]

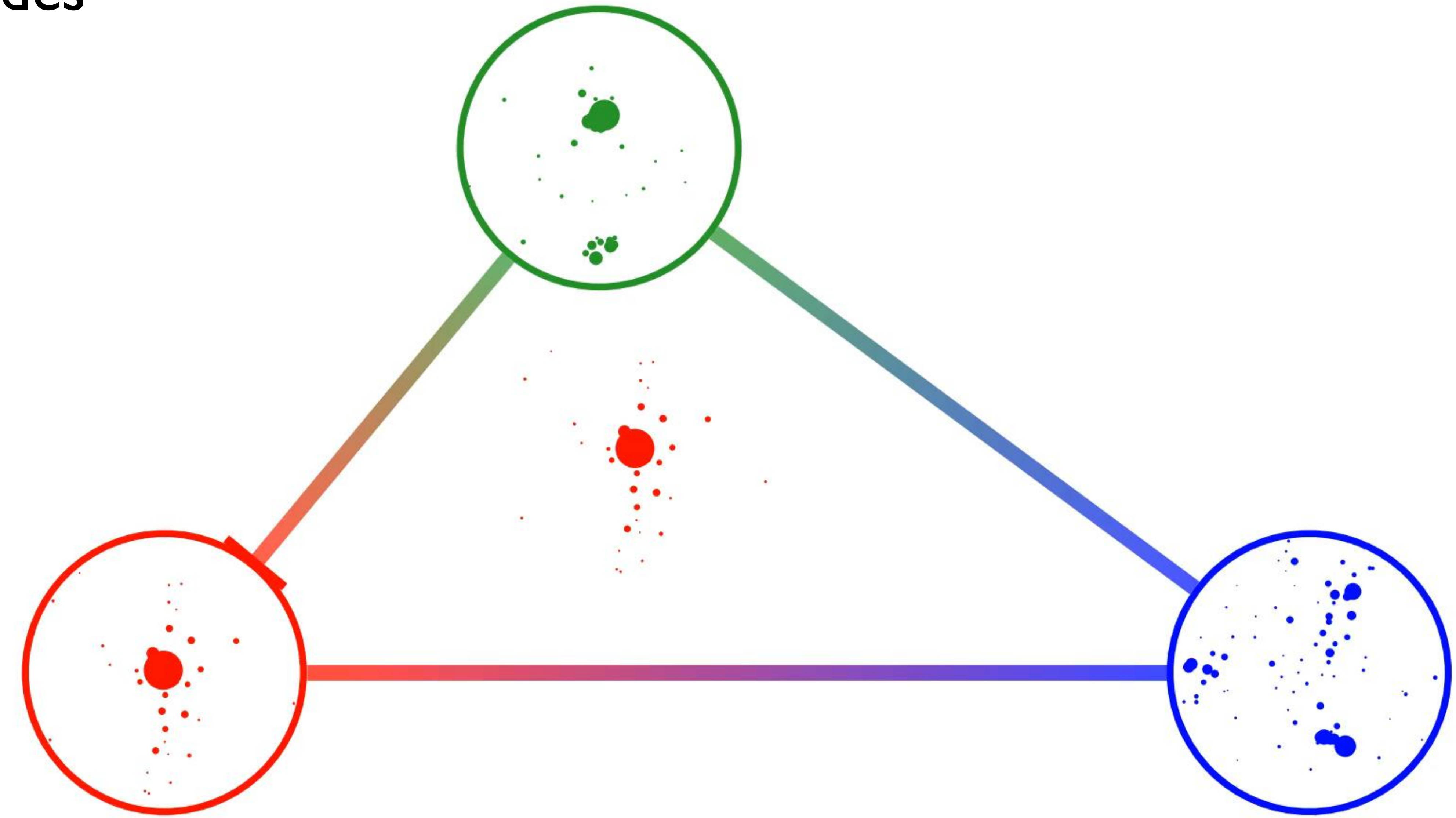


Public Collider Data
[opendata.cern.ch]

New Insights into Quantum Field Theory

Publicly released jet dataset

Extra Slides



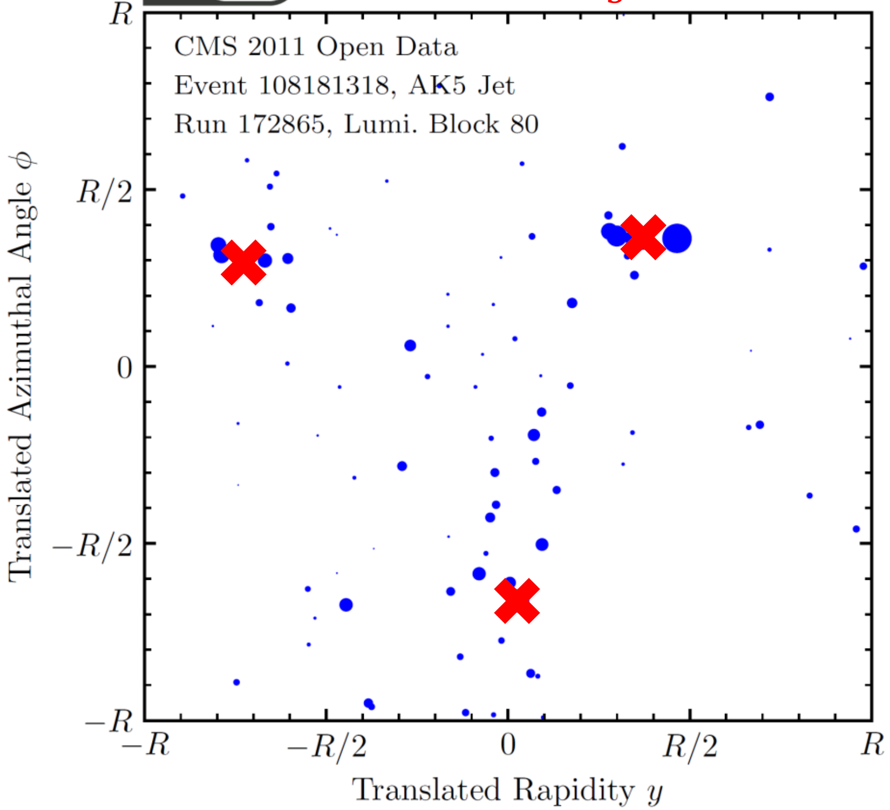
A Geometric Language for Observables

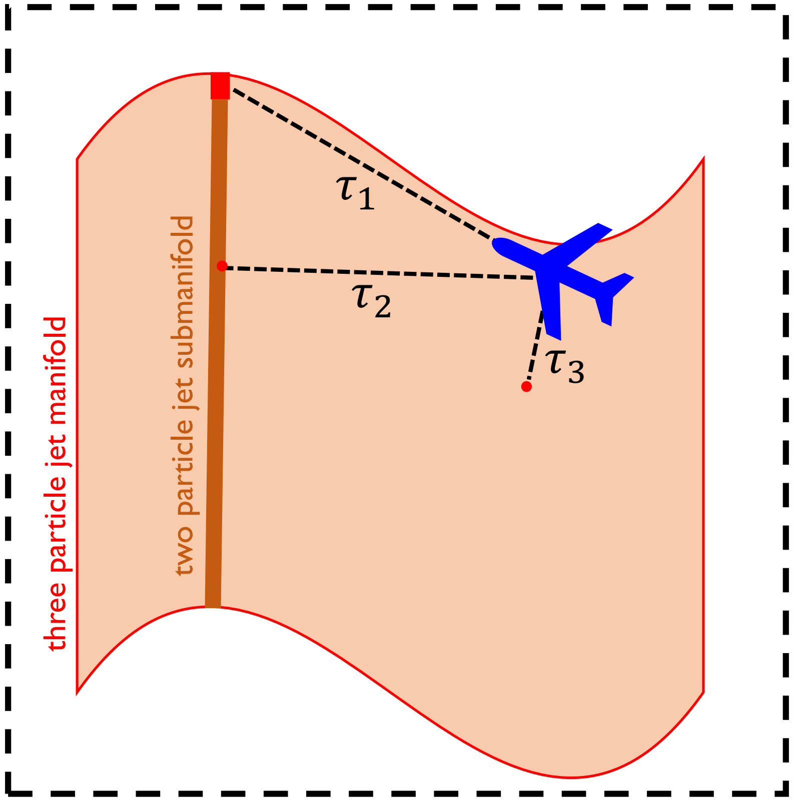
***N*-(sub)jettiness** is the EMD between the event and the closest N-particle event.

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

β-Wasserstein distance

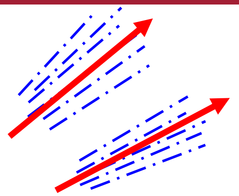
$N = 3, \quad \tau_3 \ll 1$





Geometry in the space of events

A Geometric Language for Observables

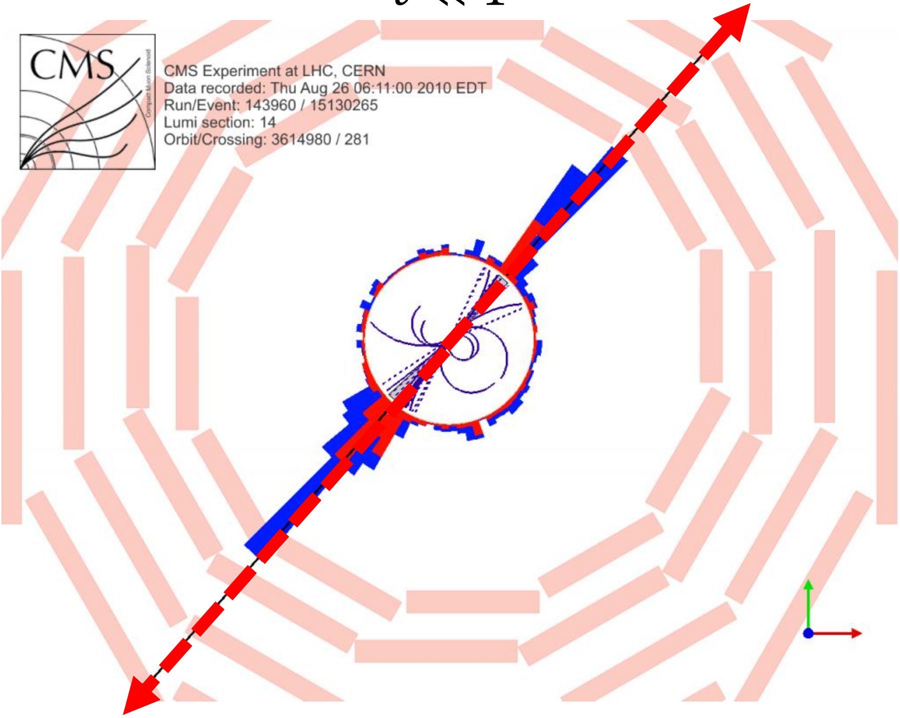


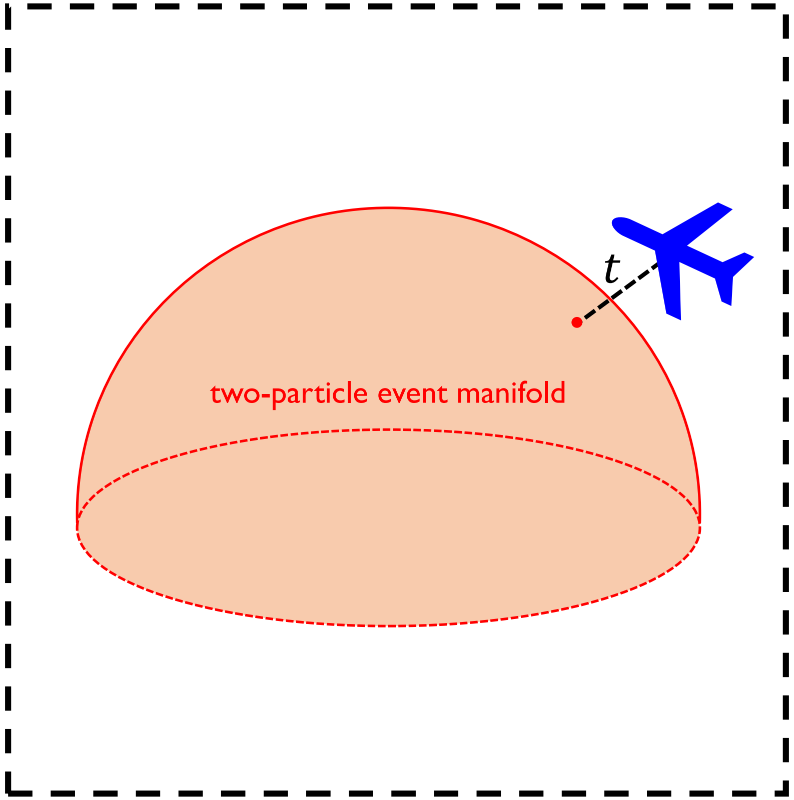
Thrust is the EMD between the event and the closest two-particle event.

$$t(\mathcal{E}) = E - \max_{\hat{n}} \sum_i |\vec{p}_i \cdot \hat{n}| \longrightarrow t(\mathcal{E}) = \min_{|\mathcal{E}'|=2} \mathrm{EMD}(\mathcal{E}, \mathcal{E}')$$

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

$t \ll 1$





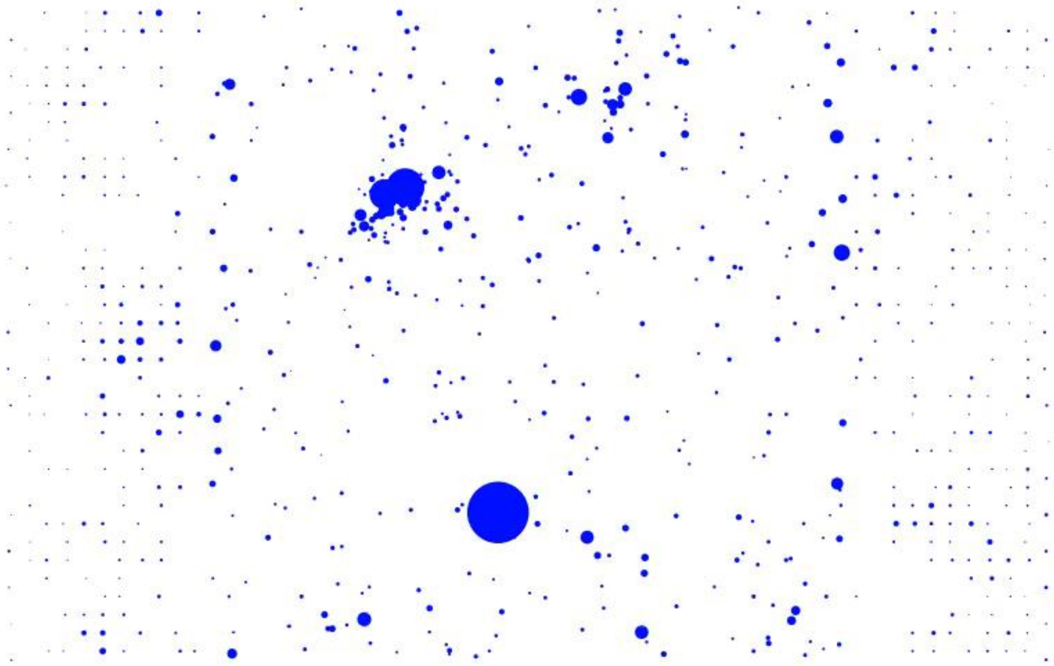
Geometry in the space of events

A Geometric Language for Observables

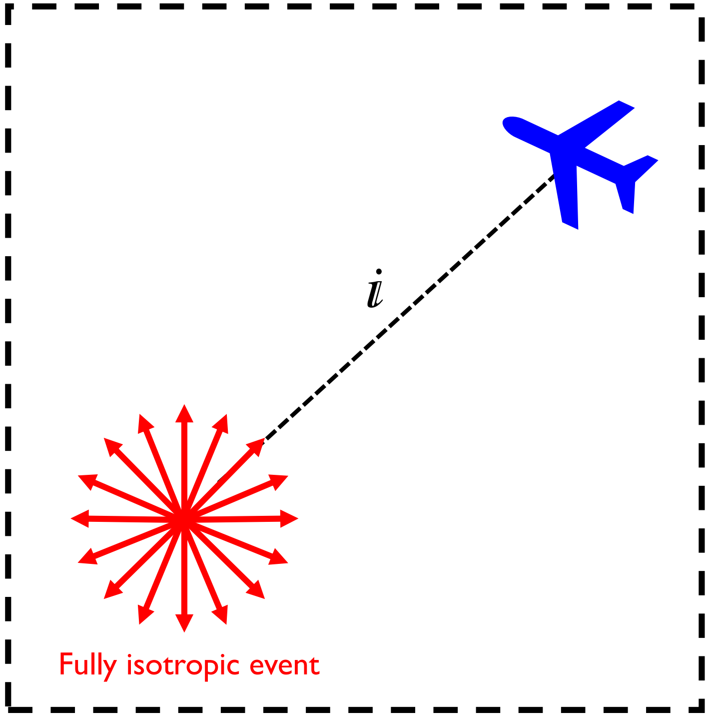
Isotropy is a new observable to probe how "uniform" an event is.

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!]

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

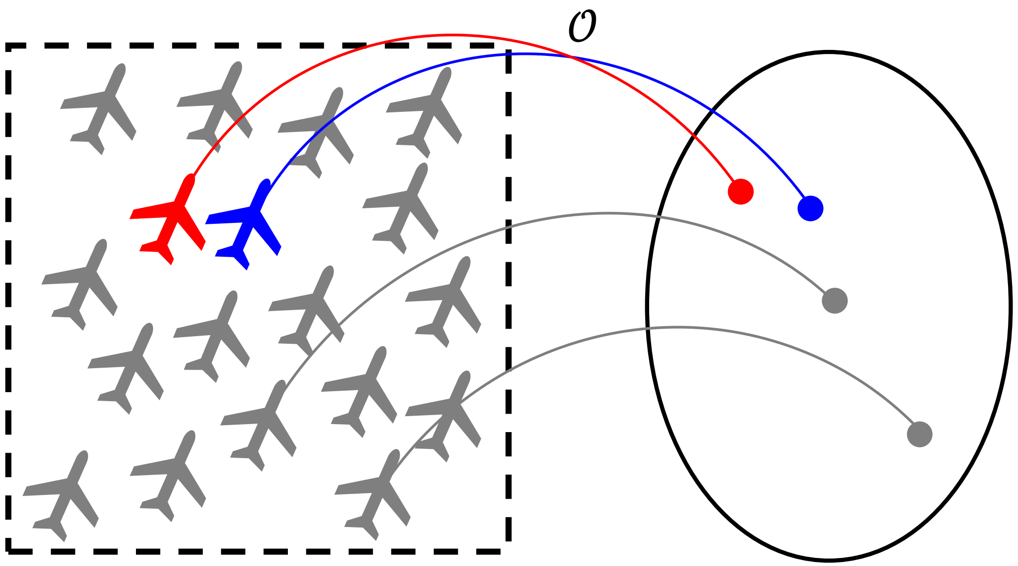


dijet event from CMS Open Data



A Geometric Language for Observables

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



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

Energy Mover's Distance

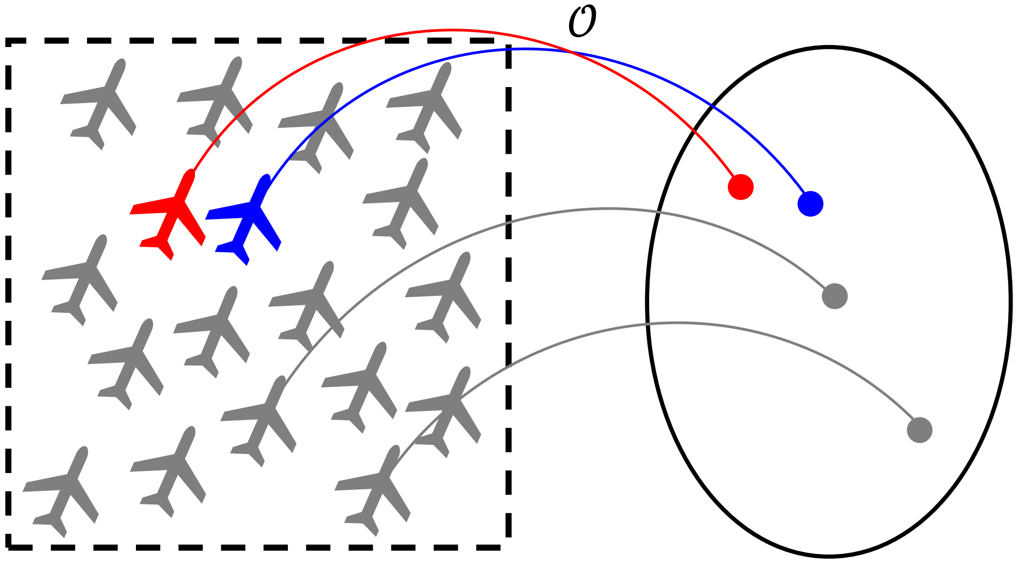
$$\boxed{\mathrm{EMD}(\mathcal{E}, \mathcal{E}') \geq \frac{1}{RL}\, |\mathcal{O}(\mathcal{E}) - \mathcal{O}(\mathcal{E}')|}$$

Difference in observable values

"Lipschitz constant" of Φ
i.e. bound on its derivative

A Geometric Language for Observables

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



Jet angularities with $\beta \geq 1$:

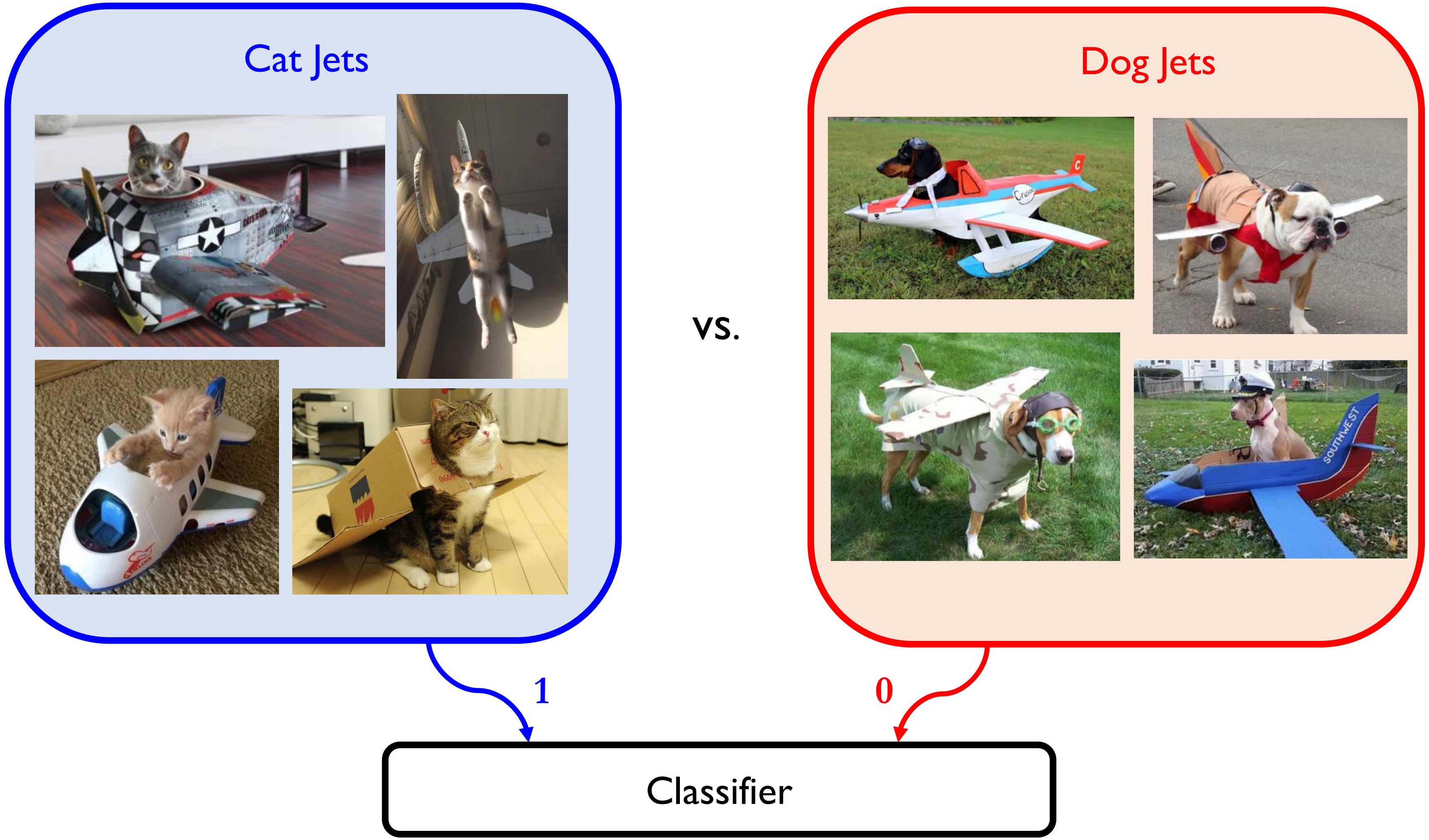
[C. Berger, T. Kucs, and G. Sterman, 0303051]

[A. Larkoski, J. Thaler, and W. Waalewijn, 1408.3122]

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

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

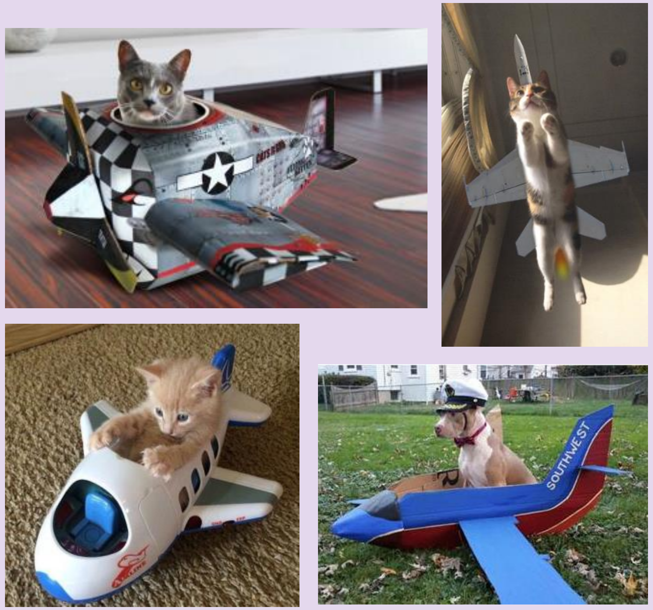
Training on pure samples: Cat jets vs. Dog jets



Training on mixed samples: Cat jets vs. Dog jets

Classification Without Labels (CWoLa)

Cat-enriched Jets



vs.

Dog-enriched Jets

1

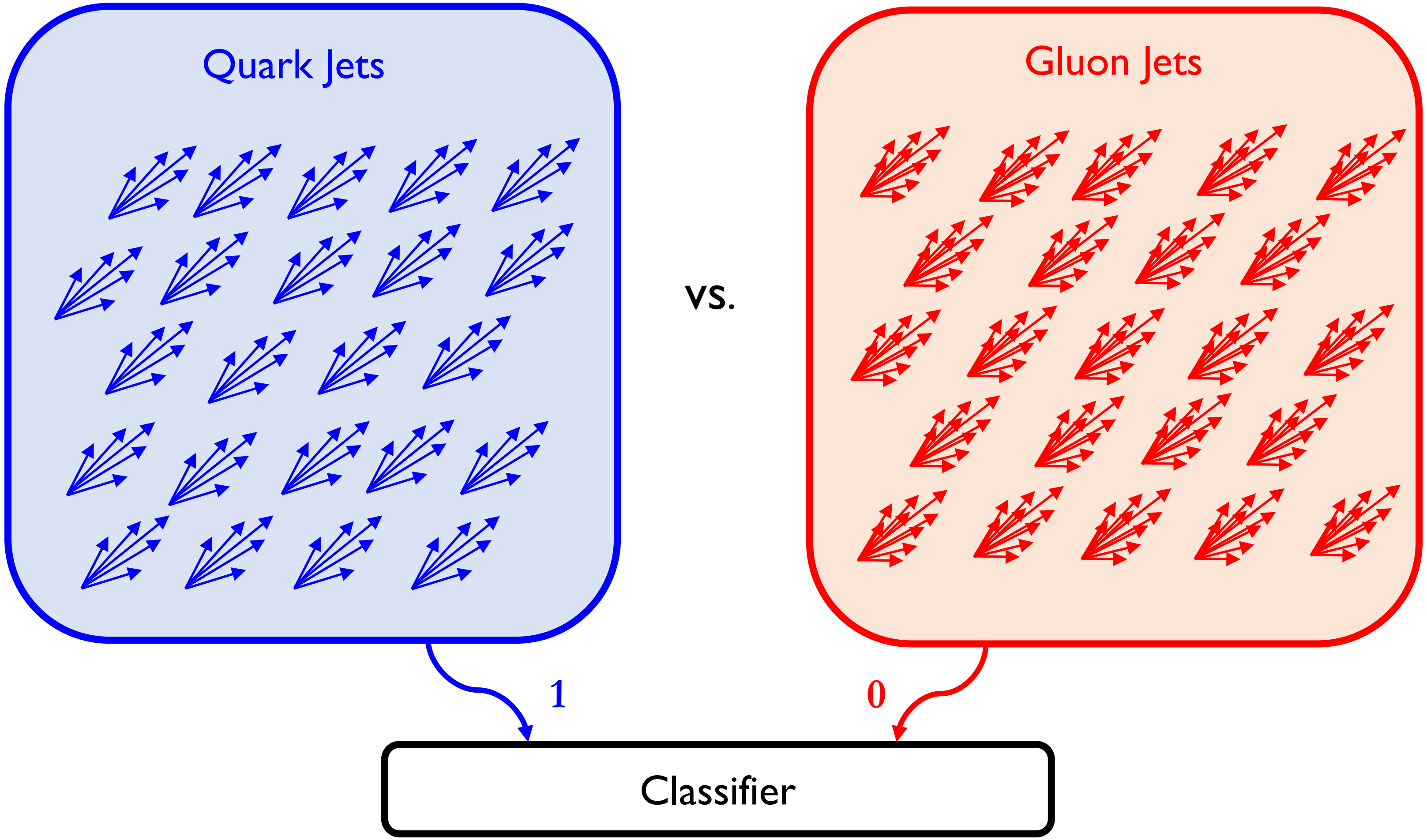
0

Classifier

Used by CMS for ttbb! [CMS 1909.05306]

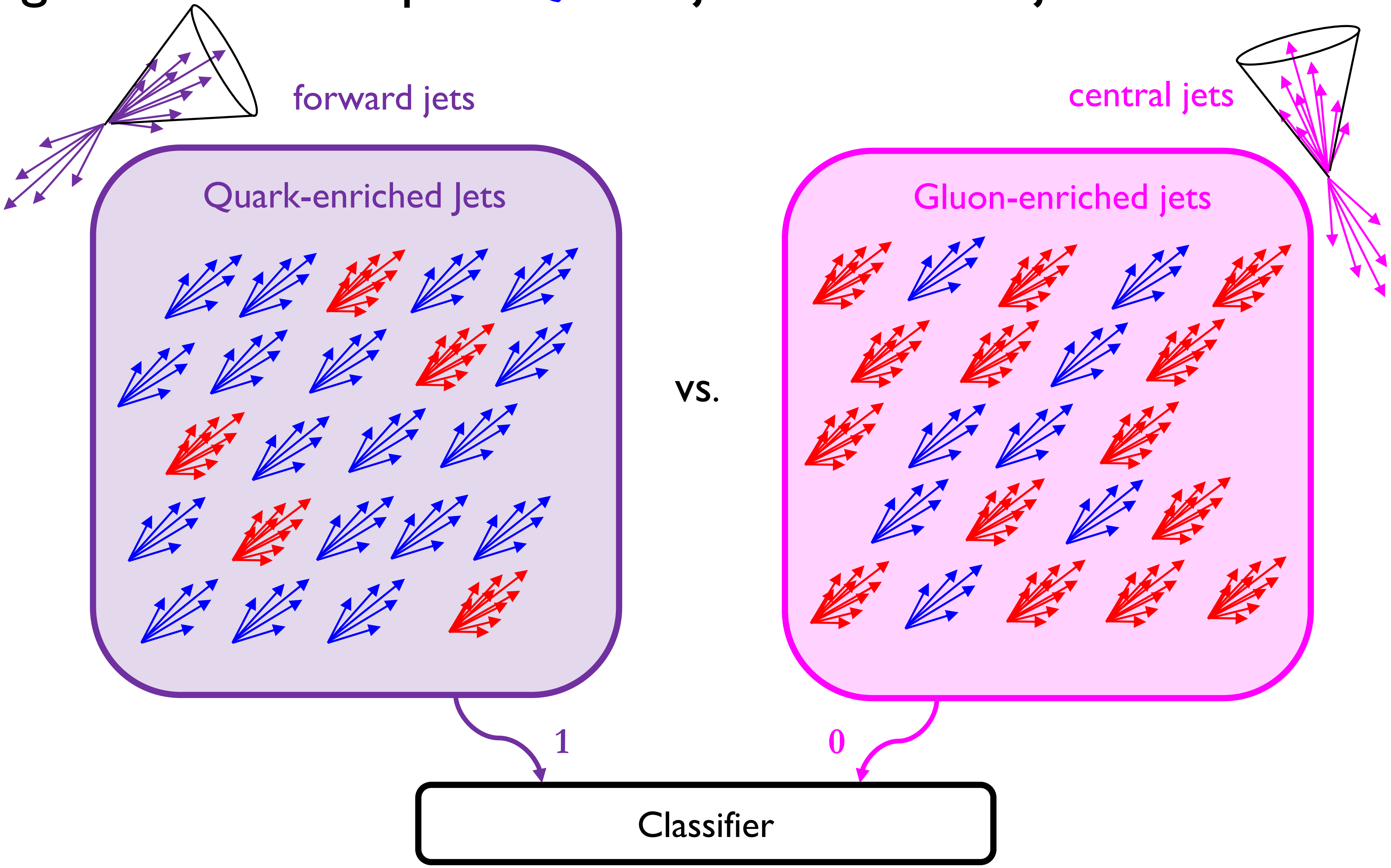
This defines an equivalent classifier to the pure case!

Training on pure samples: Quark jets vs. Gluon jets



Training on mixed samples: Quark jets vs. Gluon jets

Classification Without Labels (CWoLa)

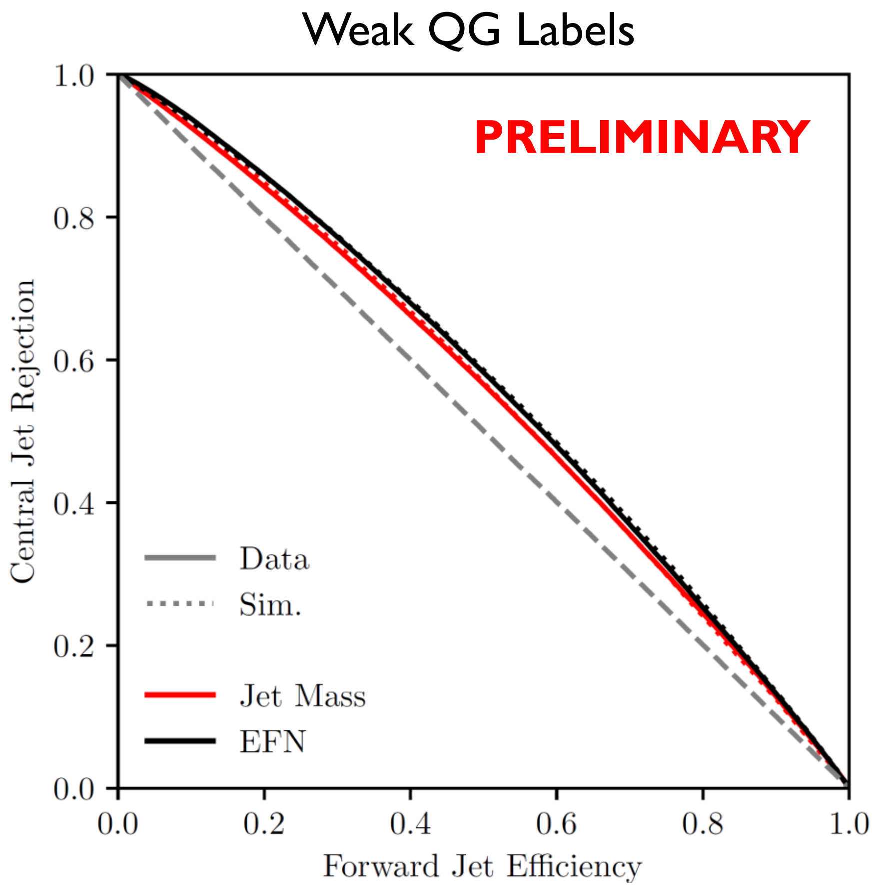


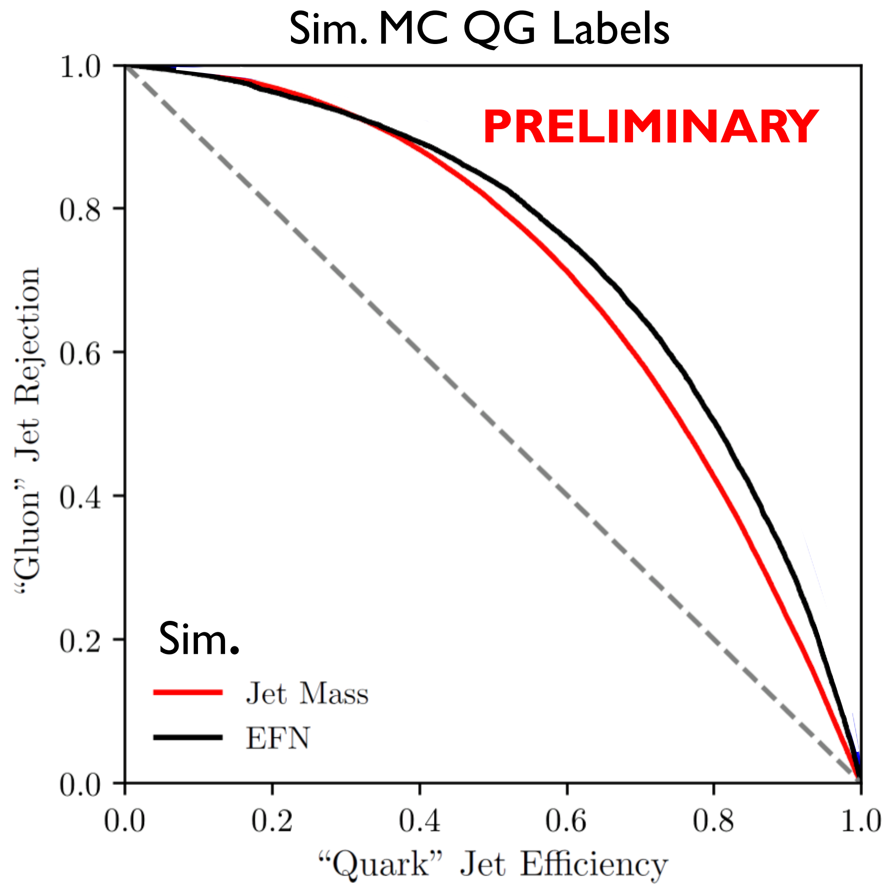
[**EMM**, B. Nachman, J. Thaler, 1708.02949]
[P.T. Komiske, **EMM**, B. Nachman, M.D. Schwartz, 1801.10158]
[L. Dery, B. Nachman, F. Rubbo, A. Schwartzman, 1702.00414] [T. Cohen, M. Freytsis, B. Ostdiek, 1706.09451]

Training on Data!

Central Jets $(|\eta^{\mathrm{jet}}| < 0.7)$: ~45% quark jets

Forward Jets $(|\eta^{\mathrm{jet}}| > 0.7)$: ~65% quark jets





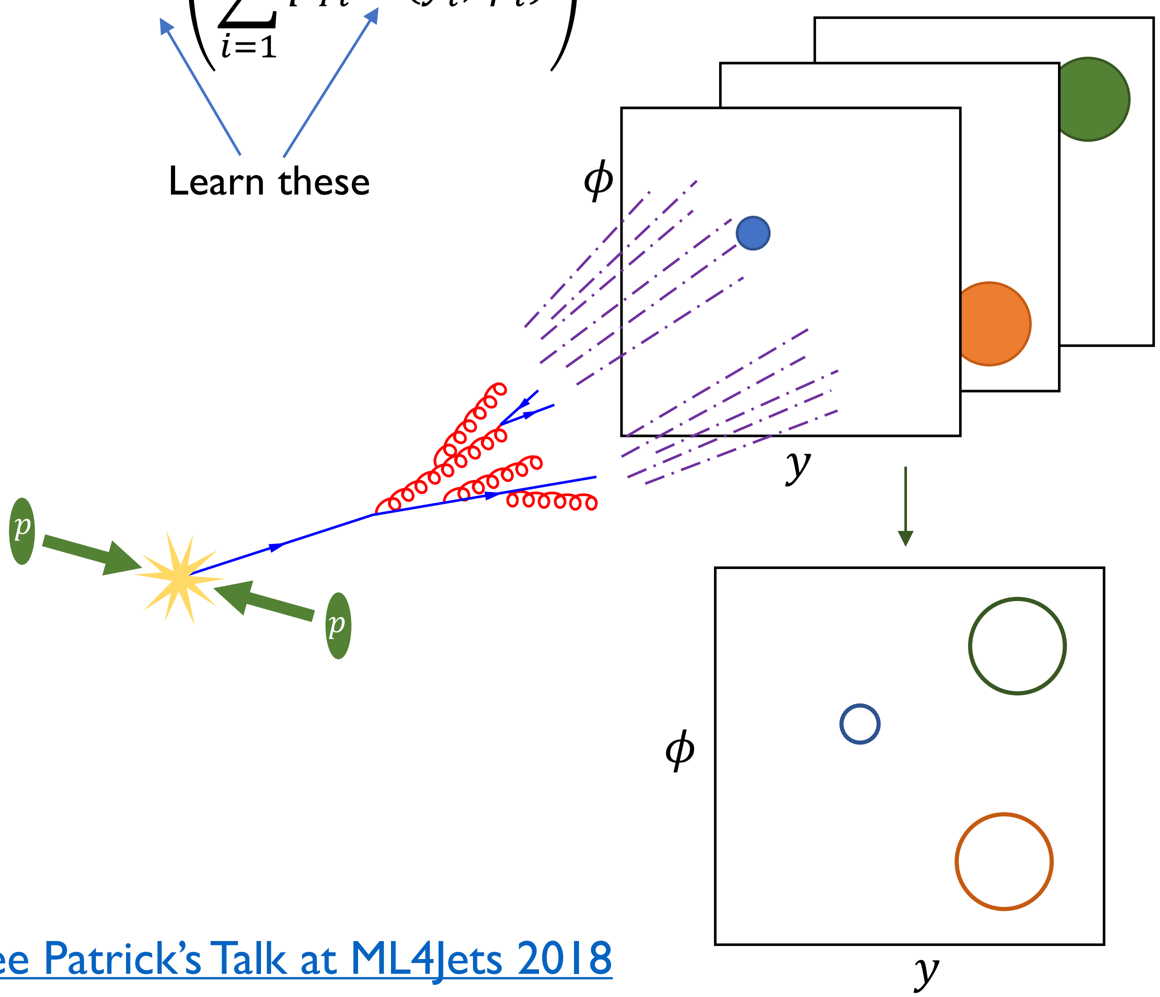
To reduce sample dependence, we train an EFN on tracks with $p_T^{\mathrm{PFC}} > 1$ GeV and remove pileup.

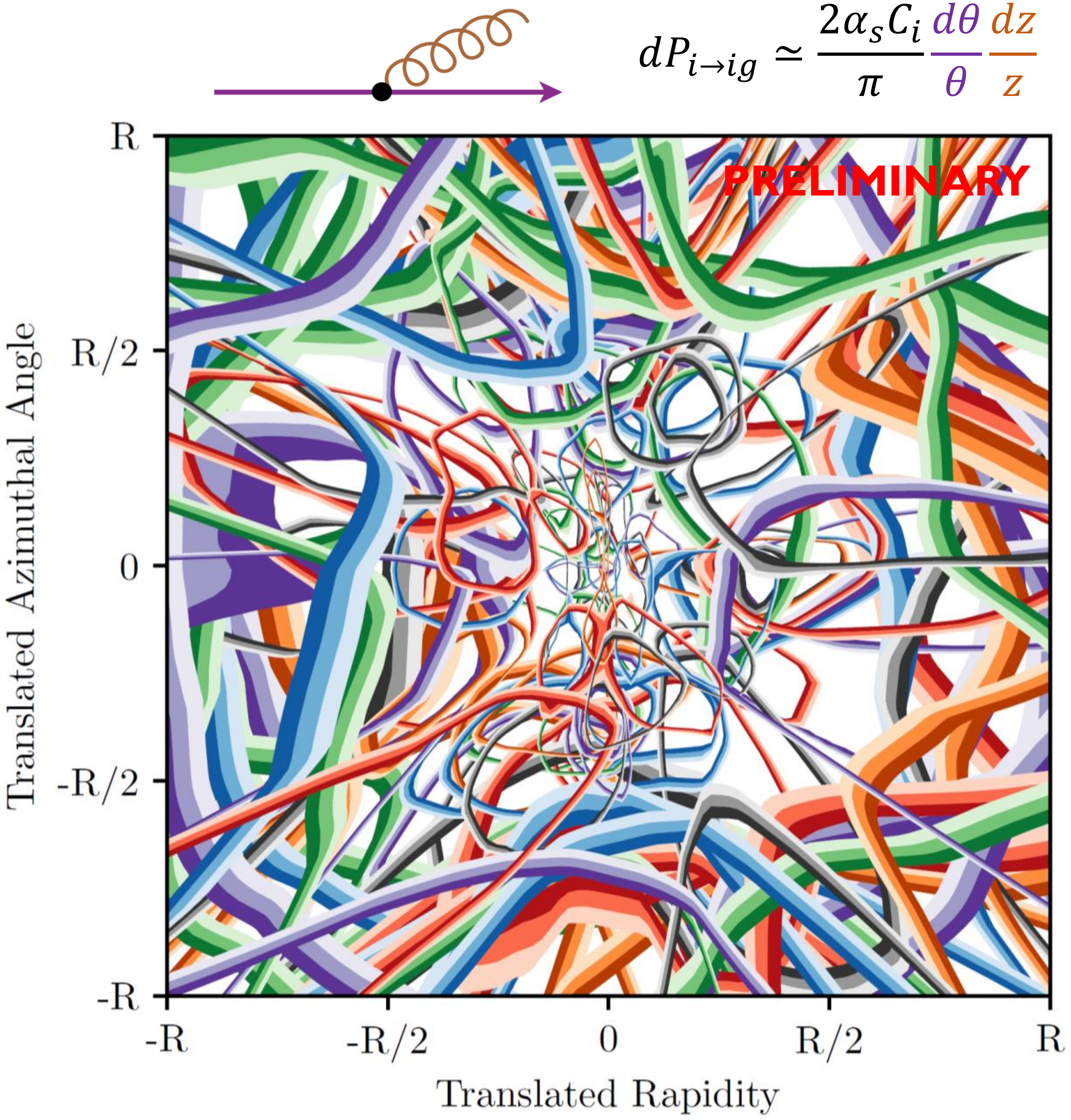
Or high-dimensional unfolding? [See Patrick's Talk]

What is the model learning?

$$\text{EFN} = F\left(\sum_{i=1}^{M} p_{Ti}\, \mathbf{\Phi}(y_i, \phi_i)\right)$$

Learn these





Visualizing 256 filters for EFN (weakly) trained on data

See Patrick's Talk at ML4Jets 2018

Exploring the Space of Jets: Correlation Dimension

Sketch of leading log (one emission) calculation:

$$\dim_{q/g}(Q) = Q\frac{\partial}{\partial Q}\ln\sum_{i=1}^{N}\sum_{j=1}^{N}\Theta[\mathrm{EMD}(\mathcal{E}_i,\mathcal{E}_j) < Q]$$

$$= Q\frac{\partial}{\partial Q}\ln \Pr\,[\mathrm{EMD} < Q]$$

$$= Q\frac{\partial}{\partial Q}\ln \Pr\,[\lambda^{(\beta=1)} < Q; C_{q/g} \to 2\,C_{q/g}]$$

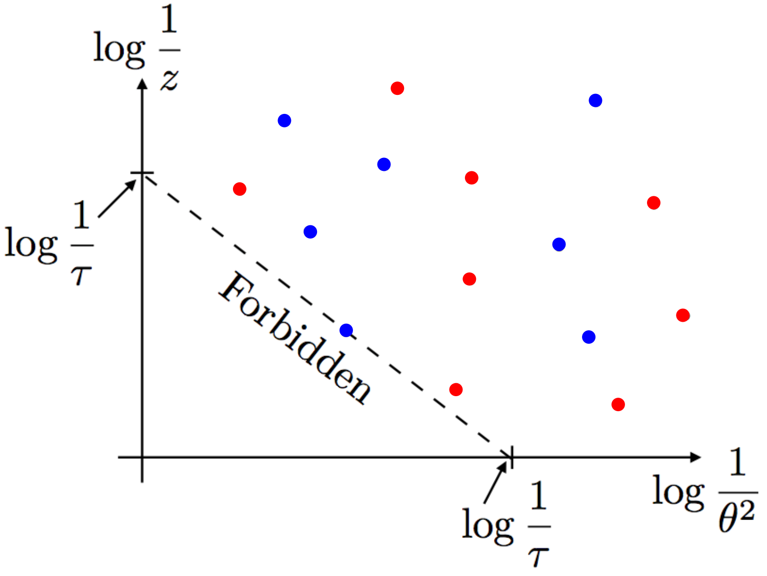
$$= Q\frac{\partial}{\partial Q}\ln \exp\left(-\frac{4\alpha_s C_{q/g}}{\pi}\ln^2\frac{Q}{p_T/2}\right)$$

$$= -\frac{8\alpha_s C_{q/g}}{\pi}\ln\frac{Q}{p_T/2}$$

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

$$C_g = C_A = 3$$

+ 1-loop running of α_s



[A. Larkoski, 1709.06195]