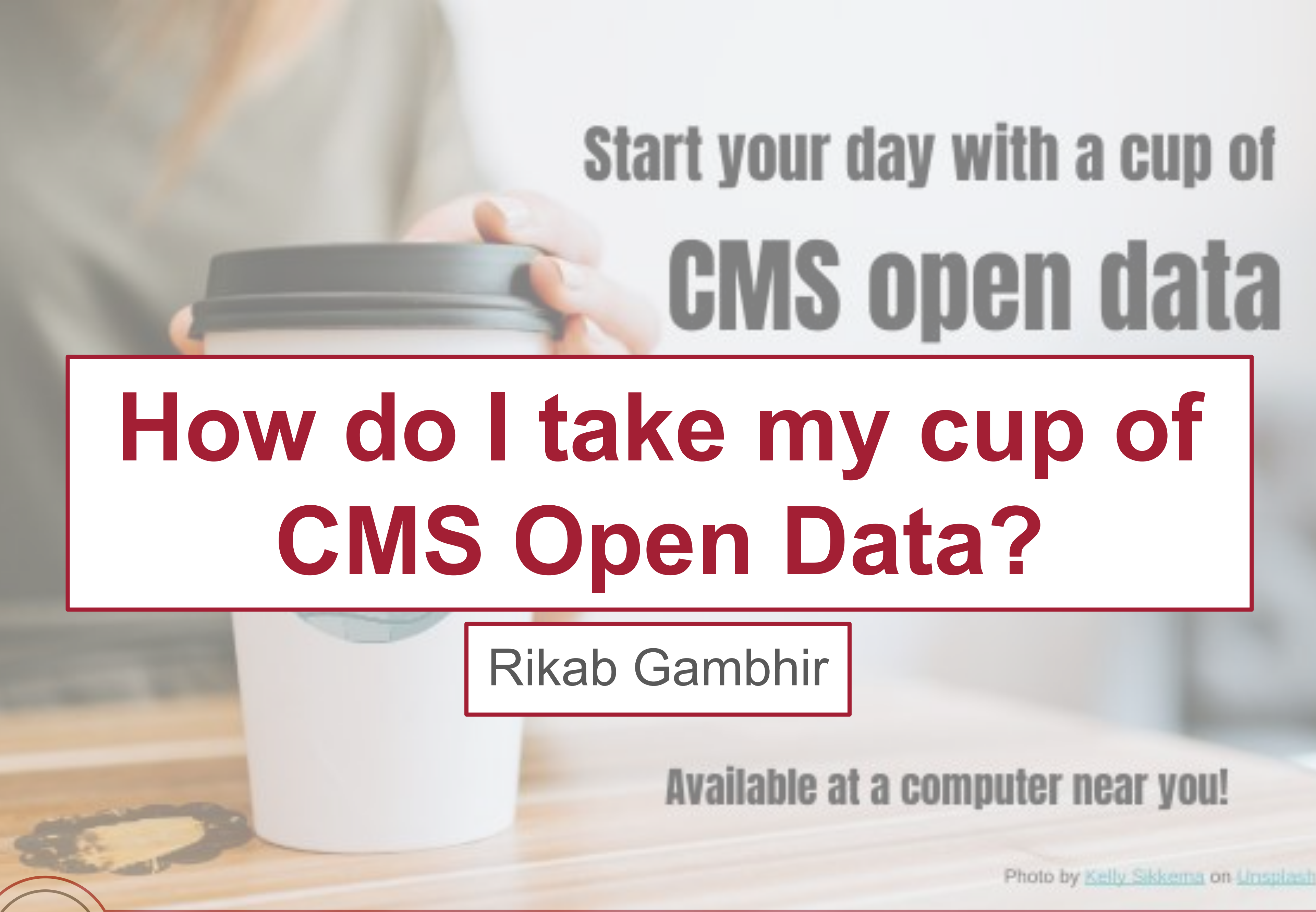

Start your day with a cup of
CMS open data
Available at a computer near you!

Photo by Kelly Sikkema on Unsplash

How do I take my cup of CMS Open Data?

Rikab Gambhir

I like my **CMS Open Data** like I like my coffee …

Start your day with a cup of
CMS open data
Available at a computer near you!

Photo by Kelly Sikkema on Unsplash

I like my **CMS Open Data** like I like my coffee …

- Very easily accessible anywhere I am
- Takes only a few seconds to minutes to set up
- Highly preprocessed and prepackaged
- Don’t have to understand all the details of how it was made
- Helps me make plots
- Can order online
- Made by somebody else
- Contains flavor information

Admittedly, the last few are a stretch

Start your day with a cup of
CMS open data
DD
DUNKIN'
DONUTS®
Available at a computer near you!

Photo by Kelly Sikkema on Unsplash

I like my CMS ... like I like my coffee

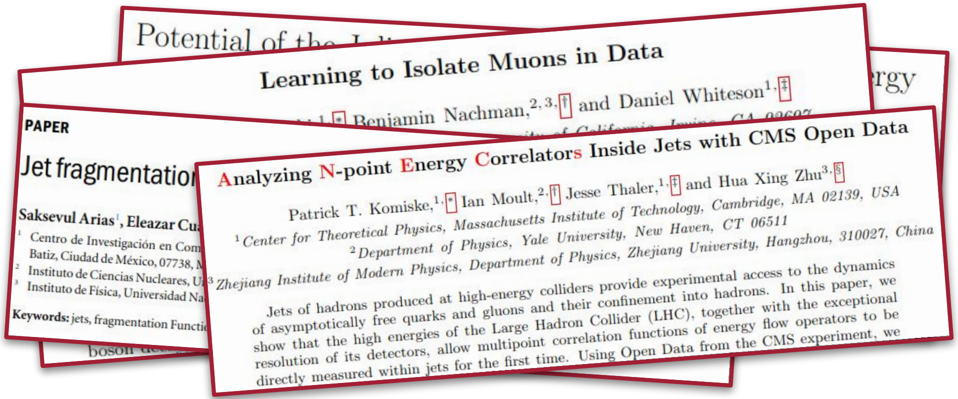
Start your day with a cup of

Don't have to understand all

Photo by Kelly Sikkema on Unsplash

This Talk

CMS Open Data, who uses it, and how it's being used



Potential of the I

Learning to Isolate Muons in Data

Benjamin Nachman,[2,3,†] and Daniel Whiteson[1,‡]

rgy

PAPER

Jet fragmentation

Saksevul Arias[1], Eleazar Cu

[1] Centro de Investigación en Com Batiz, Ciudad de México, 07738, M

[2] Instituto de Ciencias Nucleares, U

[3] Instituto de Física, Universidad Na

Keywords: jets, fragmentation Functio

boson

Analyzing N-point Energy Correlators Inside Jets with CMS Open Data

Patrick T. Komiske,[1,*] Ian Moult,[2,†] Jesse Thaler,[1,‡] and Hua Xing Zhu[3,§]

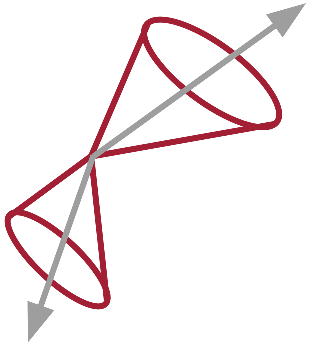
[1] Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2] Department of Physics, Yale University, New Haven, CT 06511

[3] Zhejiang Institute of Modern Physics, Department of Physics, Zhejiang University, Hangzhou, 310027, China

Jets of hadrons produced at high-energy colliders provide experimental access to the dynamics of asymptotically free quarks and gluons and their confinement into hadrons. In this paper, we show that the high energies of the Large Hadron Collider (LHC), together with the exceptional resolution of its detectors, allow multipoint correlation functions of energy flow operators to be directly measured within jets for the first time. Using Open Data from the CMS experiment, we

My own experiences and anecdotes with **CMS Open Data**

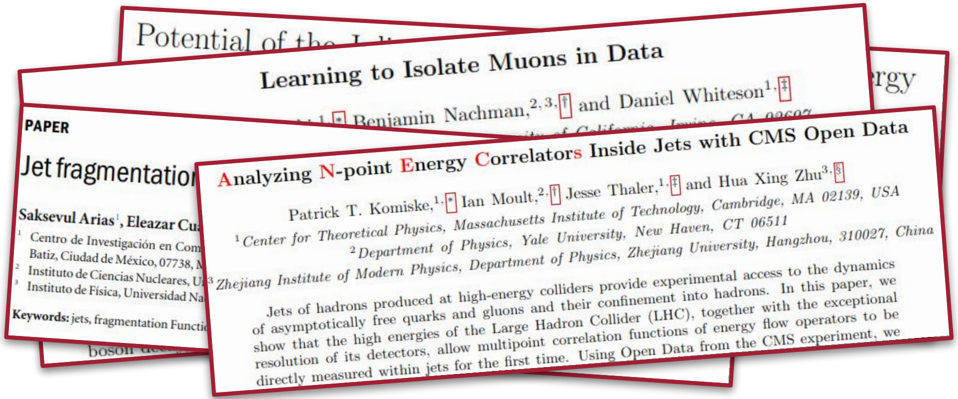


CMS

Compact Muon Solenoid

This Talk

CMS Open Data, who uses it, and how it's being used



Potential of the I

rgy

Learning to Isolate Muons in Data

1,[*] Benjamin Nachman,[2,3,†] and Daniel Whiteson[1,‡]

PAPER

Jet fragmentation

Saksevul Arias[1], Eleazar Cu

[1] Centro de Investigación en Com Batiz, Ciudad de México, 07738, M

[2] Instituto de Ciencias Nucleares, U

[3] Instituto de Física, Universidad Na

Keywords: jets, fragmentation Functio

boson

Analyzing N-point Energy Correlators Inside Jets with CMS Open Data

Patrick T. Komiske,[1,*] Ian Moult,[2,†] Jesse Thaler,[1,‡] and Hua Xing Zhu[3,§]

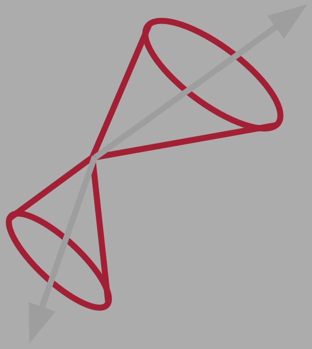
[1] Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2] Department of Physics, Yale University, New Haven, CT 06511

[3] Zhejiang Institute of Modern Physics, Department of Physics, Zhejiang University, Hangzhou, 310027, China

Jets of hadrons produced at high-energy colliders provide experimental access to the dynamics of asymptotically free quarks and gluons and their confinement into hadrons. In this paper, we show that the high energies of the Large Hadron Collider (LHC), together with the exceptional resolution of its detectors, allow multipoint correlation functions of energy flow operators to be directly measured within jets for the first time. Using Open Data from the CMS experiment, we

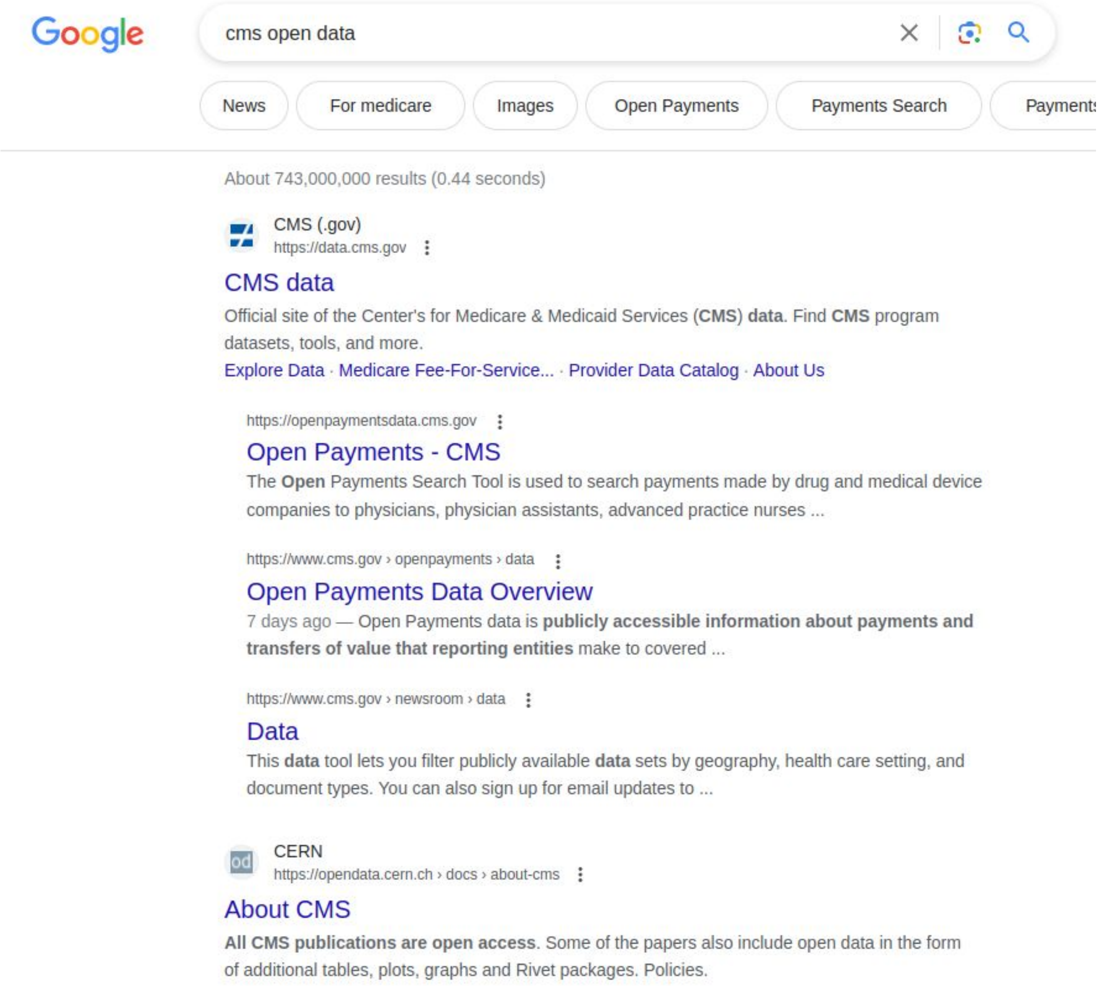
My own experiences and anecdotes with **CMS Open Data**



CMS

Compact Muon Solenoid

CMS Open Data



About 743,000,000 results (0.44 seconds)

CMS (.gov)
https://data.cms.gov

CMS data

Official site of the Center's for Medicare & Medicaid Services (**CMS**) **data**. Find **CMS** program datasets, tools, and more.

Explore Data · Medicare Fee-For-Service... · Provider Data Catalog · About Us

https://openpaymentsdata.cms.gov

Open Payments - CMS

The **Open** Payments Search Tool is used to search payments made by drug and medical device companies to physicians, physician assistants, advanced practice nurses ...

https://www.cms.gov › openpayments › data

Open Payments Data Overview

7 days ago — Open Payments data is **publicly accessible information about payments and transfers of value that reporting entities** make to covered ...

https://www.cms.gov › newsroom › data

Data

This **data** tool lets you filter publicly available **data** sets by geography, health care setting, and document types. You can also sign up for email updates to ...

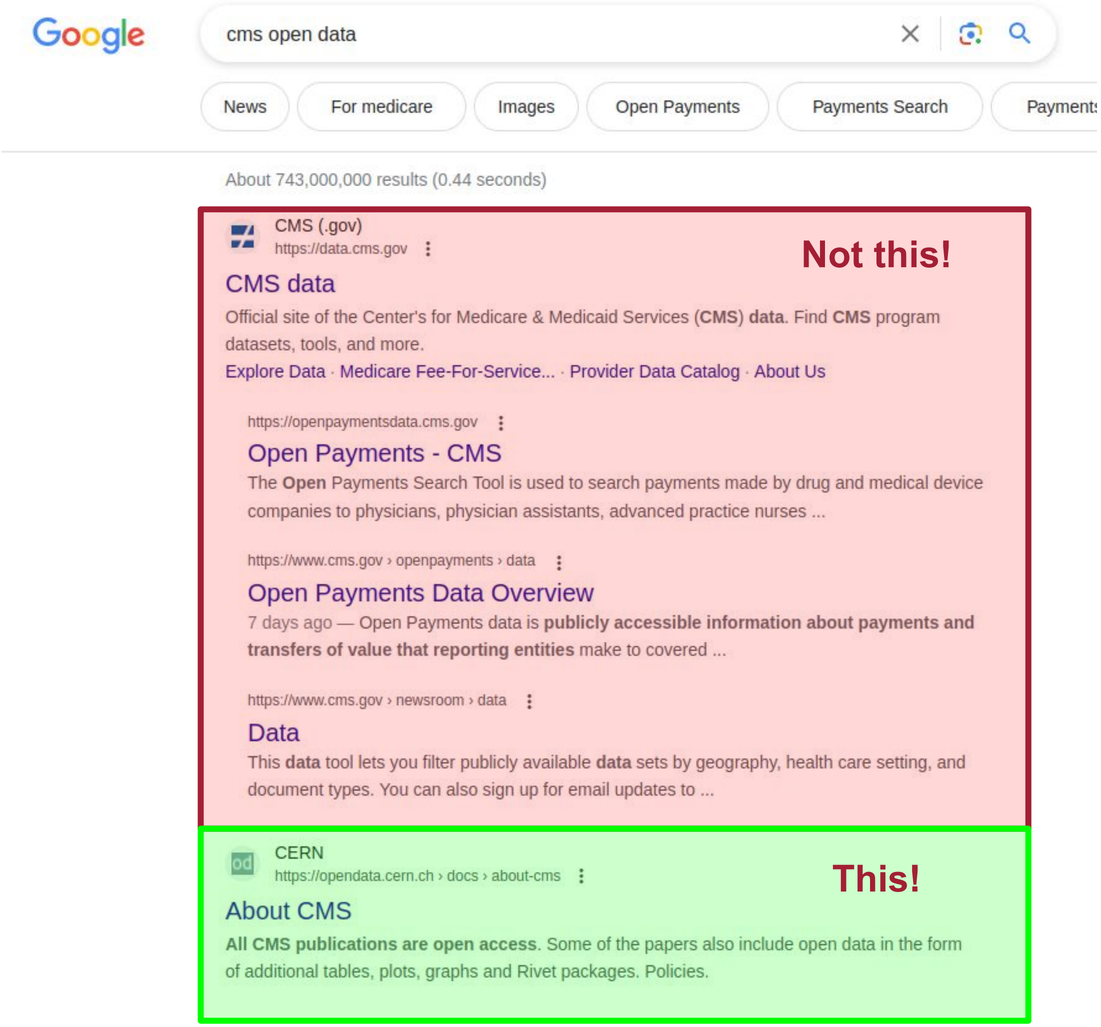
CERN
https://opendata.cern.ch › docs › about-cms

About CMS

All CMS publications are open access. Some of the papers also include open data in the form of additional tables, plots, graphs and Rivet packages. Policies.

According to Google…

CMS Open Data



cms open data

About 743,000,000 results (0.44 seconds)

CMS (.gov)
https://data.cms.gov

Not this!

CMS data

Official site of the Center's for Medicare & Medicaid Services (**CMS**) **data**. Find **CMS** program datasets, tools, and more.

Explore Data · Medicare Fee-For-Service... · Provider Data Catalog · About Us

https://openpaymentsdata.cms.gov

Open Payments - CMS

The **Open** Payments Search Tool is used to search payments made by drug and medical device companies to physicians, physician assistants, advanced practice nurses ...

https://www.cms.gov › openpayments › data

Open Payments Data Overview

7 days ago — Open Payments data is **publicly accessible information about payments and transfers of value that reporting entities** make to covered ...

https://www.cms.gov › newsroom › data

Data

This **data** tool lets you filter publicly available **data** sets by geography, health care setting, and document types. You can also sign up for email updates to ...

CERN
https://opendata.cern.ch › docs › about-cms

This!

About CMS

All CMS publications are open access. Some of the papers also include open data in the form of additional tables, plots, graphs and Rivet packages. Policies.

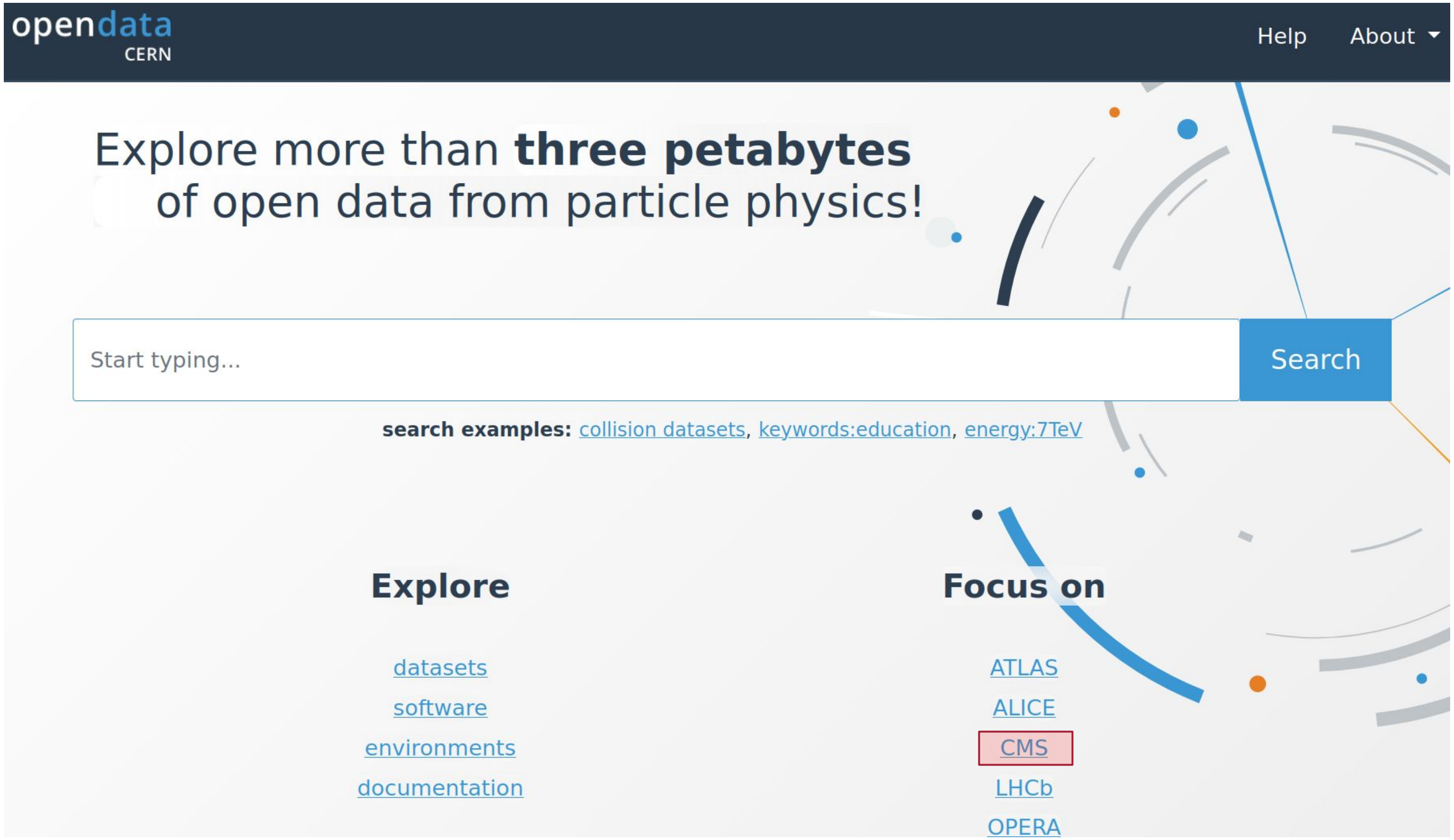
According to Google…

✗

✓

CMS Open Data

http://opendata.cern.ch/



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

Start typing...

Search

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

Explore

datasets

software

environments

documentation

Focus on

ATLAS

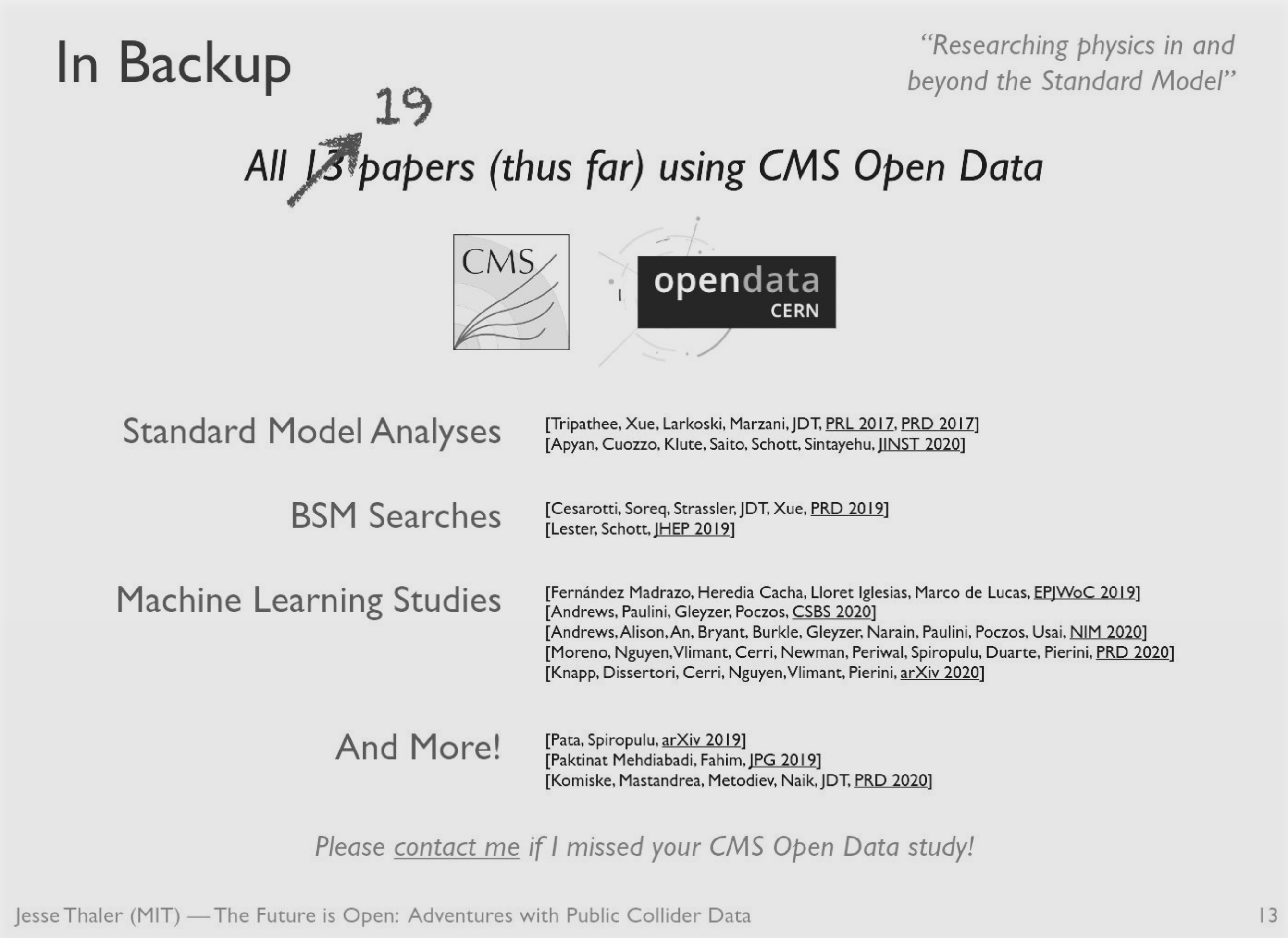
ALICE

CMS

LHCb

OPERA

In 2020…



In Backup

"Researching physics in and beyond the Standard Model"

All ~~13~~ 19 papers (thus far) using CMS Open Data

Standard Model Analyses
[Tripathee, Xue, Larkoski, Marzani, JDT, PRL 2017, PRD 2017]
[Apyan, Cuozzo, Klute, Saito, Schott, Sintayehu, JINST 2020]

BSM Searches
[Cesarotti, Soreq, Strassler, JDT, Xue, PRD 2019]
[Lester, Schott, JHEP 2019]

Machine Learning Studies
[Fernández Madrazo, Heredia Cacha, Lloret Iglesias, Marco de Lucas, EPJWoC 2019]
[Andrews, Paulini, Gleyzer, Poczos, CSBS 2020]
[Andrews, Alison, An, Bryant, Burkle, Gleyzer, Narain, Paulini, Poczos, Usai, NIM 2020]
[Moreno, Nguyen, Vlimant, Cerri, Newman, Periwal, Spiropulu, Duarte, Pierini, PRD 2020]
[Knapp, Dissertori, Cerri, Nguyen, Vlimant, Pierini, arXiv 2020]

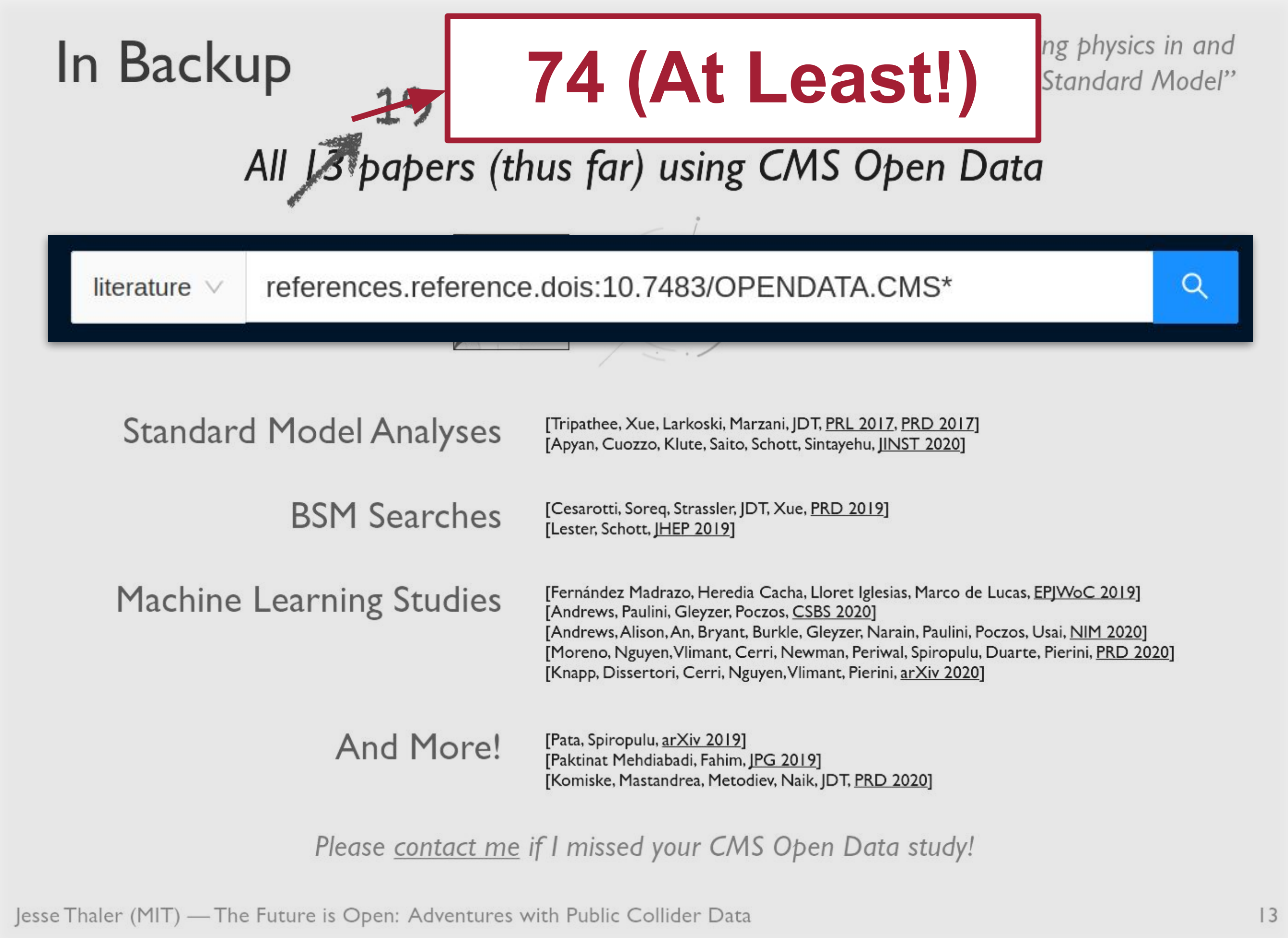
And More!
[Pata, Spiropulu, arXiv 2019]
[Paktinat Mehdiabadi, Fahim, JPG 2019]
[Komiske, Mastandrea, Metodiev, Naik, JDT, PRD 2020]

Please contact me if I missed your CMS Open Data study!

Jesse Thaler (MIT) — The Future is Open: Adventures with Public Collider Data 13

[Thaler, Adventures with Public Collider Data (2020)]

In 2023…



In Backup

74 (At Least!)

…ng physics in and …Standard Model”

19

All ~~13~~ papers (thus far) using CMS Open Data

references.reference.dois:10.7483/OPENDATA.CMS*

Standard Model Analyses

[Tripathee, Xue, Larkoski, Marzani, JDT, PRL 2017, PRD 2017]
[Apyan, Cuozzo, Klute, Saito, Schott, Sintayehu, JINST 2020]

BSM Searches

[Cesarotti, Soreq, Strassler, JDT, Xue, PRD 2019]
[Lester, Schott, JHEP 2019]

Machine Learning Studies

[Fernández Madrazo, Heredia Cacha, Lloret Iglesias, Marco de Lucas, EPJWoC 2019]
[Andrews, Paulini, Gleyzer, Poczos, CSBS 2020]
[Andrews, Alison, An, Bryant, Burkle, Gleyzer, Narain, Paulini, Poczos, Usai, NIM 2020]
[Moreno, Nguyen, Vlimant, Cerri, Newman, Periwal, Spiropulu, Duarte, Pierini, PRD 2020]
[Knapp, Dissertori, Cerri, Nguyen, Vlimant, Pierini, arXiv 2020]

And More!

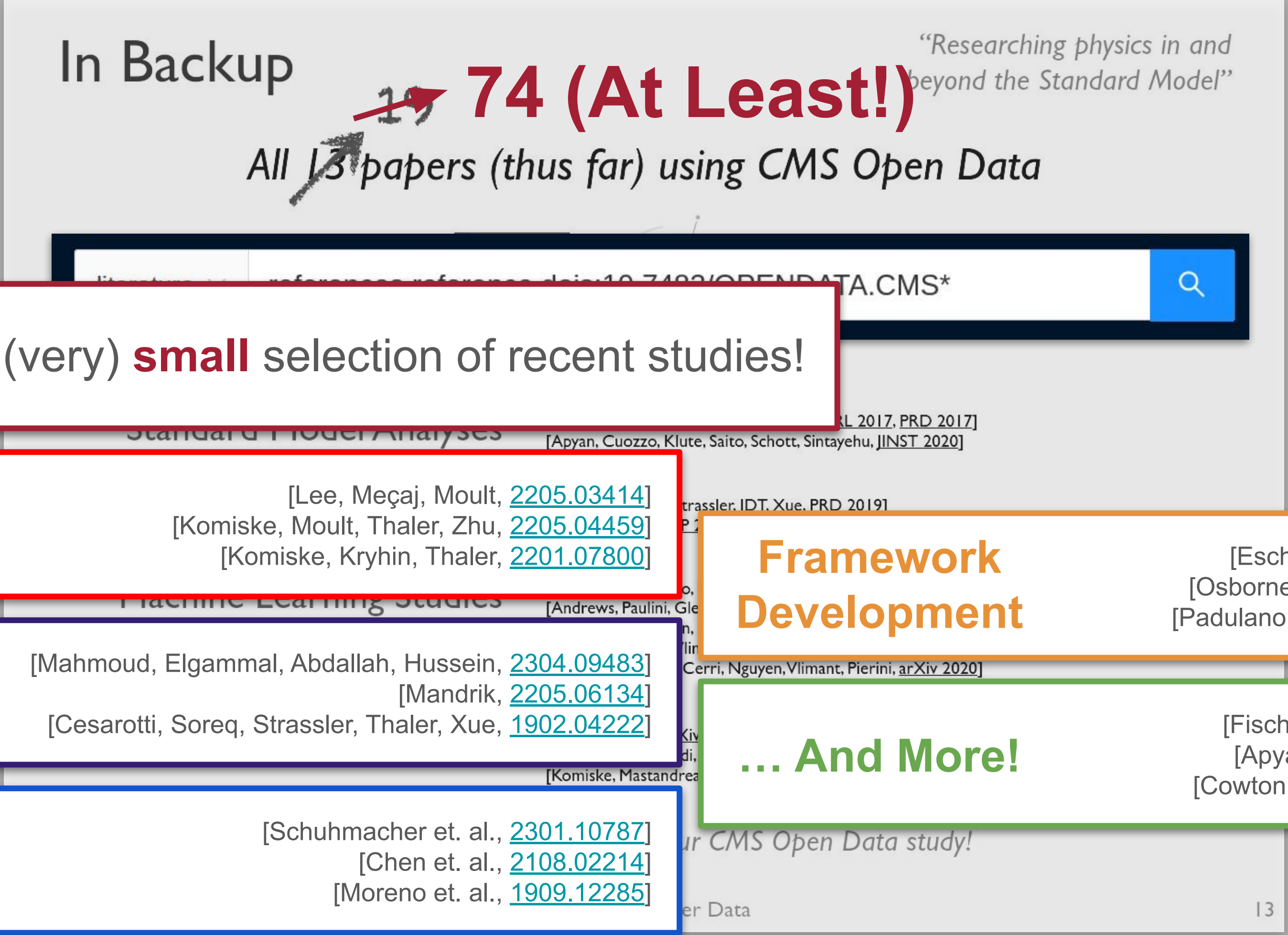
[Pata, Spiropulu, arXiv 2019]
[Paktinat Mehdiabadi, Fahim, JPG 2019]
[Komiske, Mastandrea, Metodiev, Naik, JDT, PRD 2020]

Please contact me if I missed your CMS Open Data study!

Jesse Thaler (MIT) — The Future is Open: Adventures with Public Collider Data 13

[Thaler, Adventures with Public Collider Data (2020)]

In 2023…



In Backup

“Researching physics in and beyond the Standard Model”

~~19~~ → 74 (At Least!)

All ~~13~~ papers (thus far) using CMS Open Data

literature
references:reference.doi:10.7483/OPENDATA.CMS*
Standard Model Analyses
RL 2017, PRD 2017]
[Apyan, Cuozzo, Klute, Saito, Schott, Sintayehu, JINST 2020]
trassler, JDT, Xue, PRD 2019]
Machine Learning Studies
[Andrews, Paulini, Gle
Cerri, Nguyen, Vlimant, Pierini, arXiv 2020]
[Komiske, Mastandrea
ur CMS Open Data study!
er Data
13

Just a (very) **small** selection of recent studies!

QCD

[Lee, Meçaj, Moult, 2205.03414]
[Komiske, Moult, Thaler, Zhu, 2205.04459]
[Komiske, Kryhin, Thaler, 2201.07800]

BSM

[Mahmoud, Elgammal, Abdallah, Hussein, 2304.09483]
[Mandrik, 2205.06134]
[Cesarotti, Soreq, Strassler, Thaler, Xue, 1902.04222]

ML/AI

[Schuhmacher et. al., 2301.10787]
[Chen et. al., 2108.02214]
[Moreno et. al., 1909.12285]

Framework Development

[Eschle et. al., 2306.03675]
[Osborne, Pivarski, 2302.0986]
[Padulano et. al., CDS-2856552]

… And More!

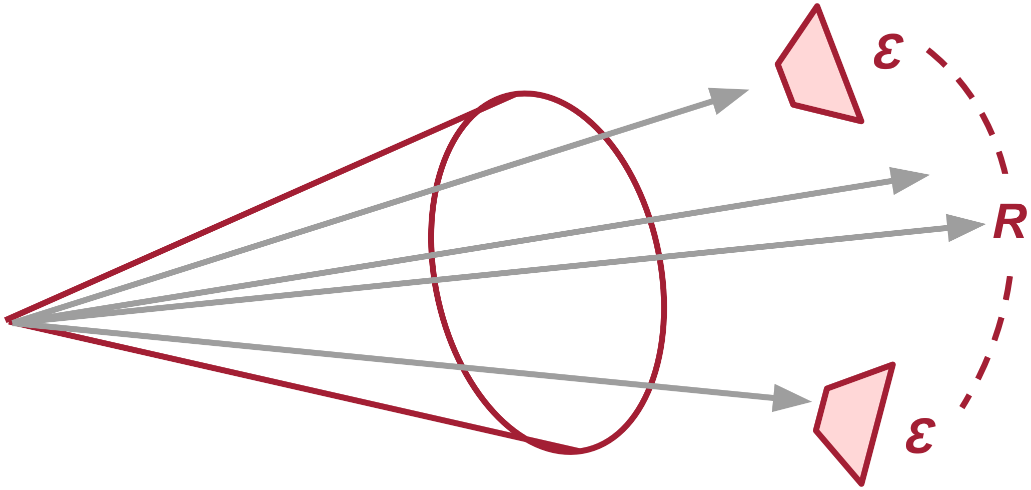
[Fischer et. al., 2109.06065]
[Apyan et. al., 1907.08197]
[Cowton et. al., CDS-2134548]

[Thaler, Adventures with Public Collider Data (2020)]

Some Fun Recent Highlights

*With a strong bias towards research done my members of my group/home institution

Example 1: Energy-Energy Correlators



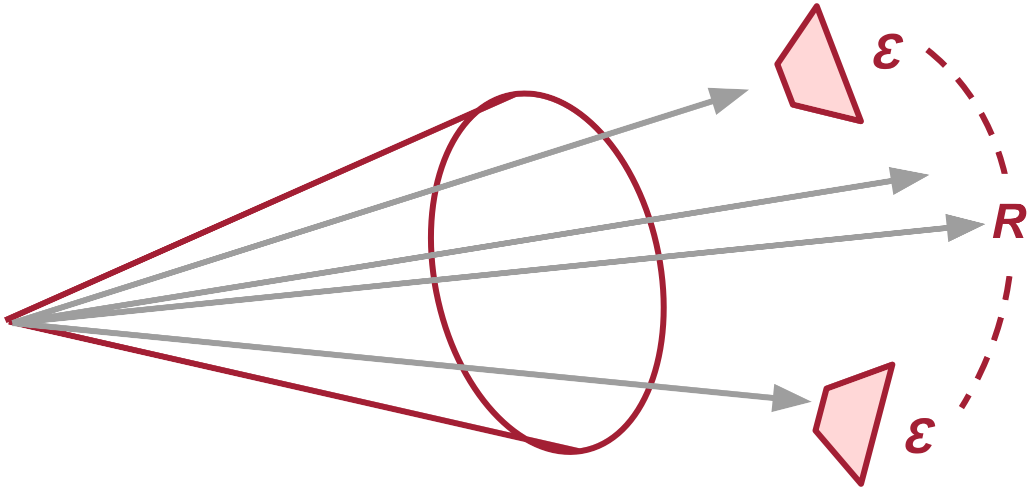
$$\mathrm{ENC}(R_L) = \left(\prod_{k=1}^{N}\int d\Omega_{\vec{n}_k}\right)\delta(R_L - \Delta\hat{R}_L) \cdot \frac{1}{(E_{\text{jet}})^N}\langle\mathcal{E}(\vec{n}_1)\mathcal{E}(\vec{n}_2)\ldots\mathcal{E}(\vec{n}_N)\rangle$$

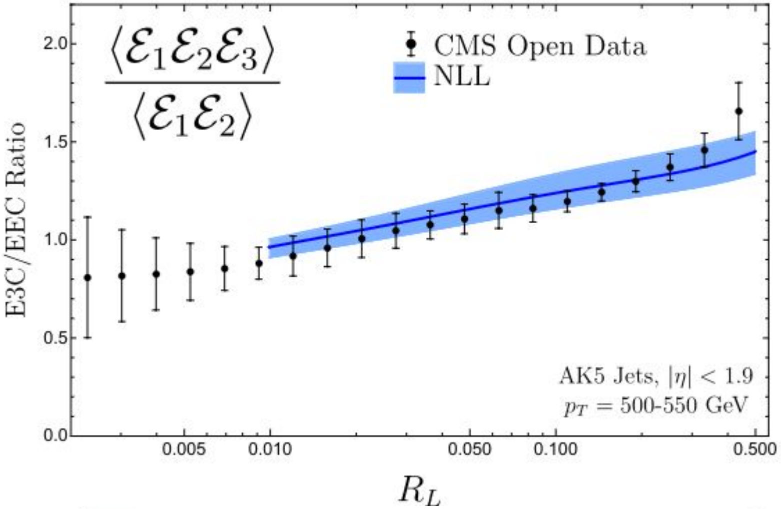
Energy-Energy Correlators (*EEC*s (and *E^N^C*s)) let us explore different aspects of QCD, including scaling behavior, collinear structure, phase transitions, and more

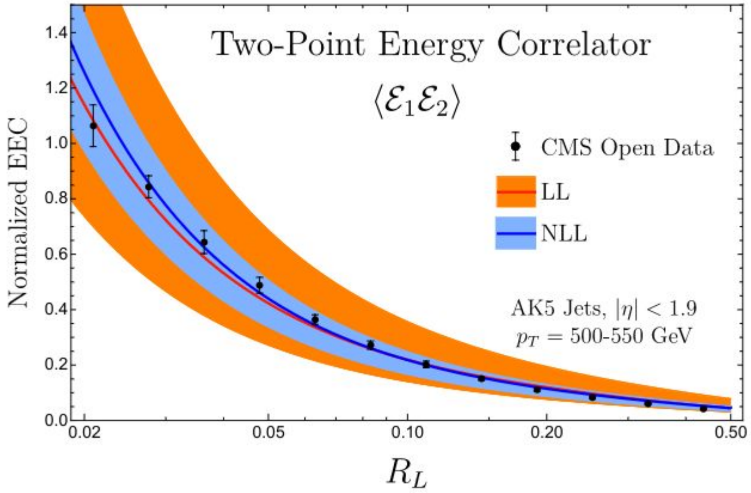
Explored in CMS Open Data!

[Lee, Meçaj, Moult, 2205.03414]

Example 1: Energy-Energy Correlators

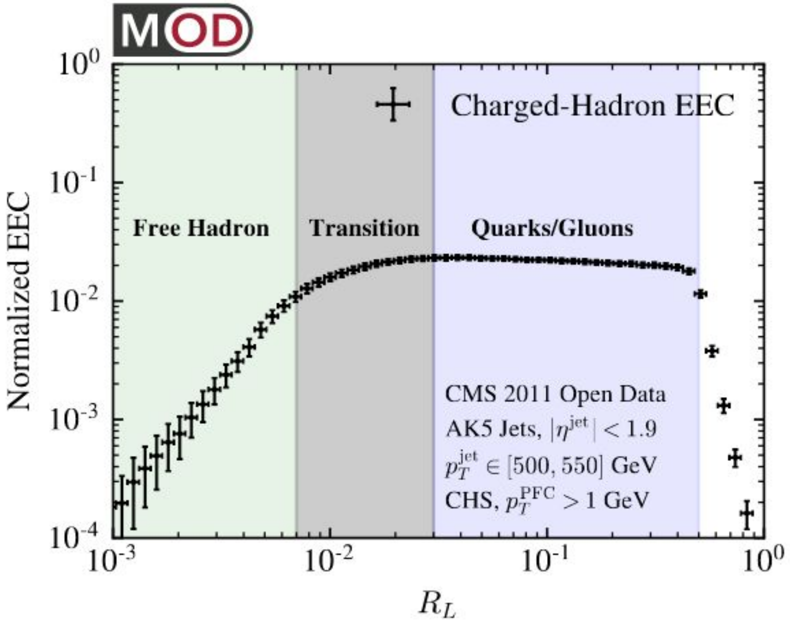


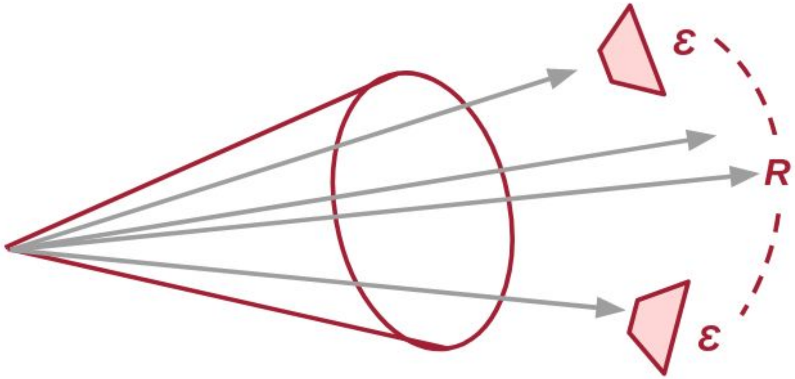




Possible for phenomenologists to compare calculations to data *directly*!

Example 1: Energy-Energy Correlators



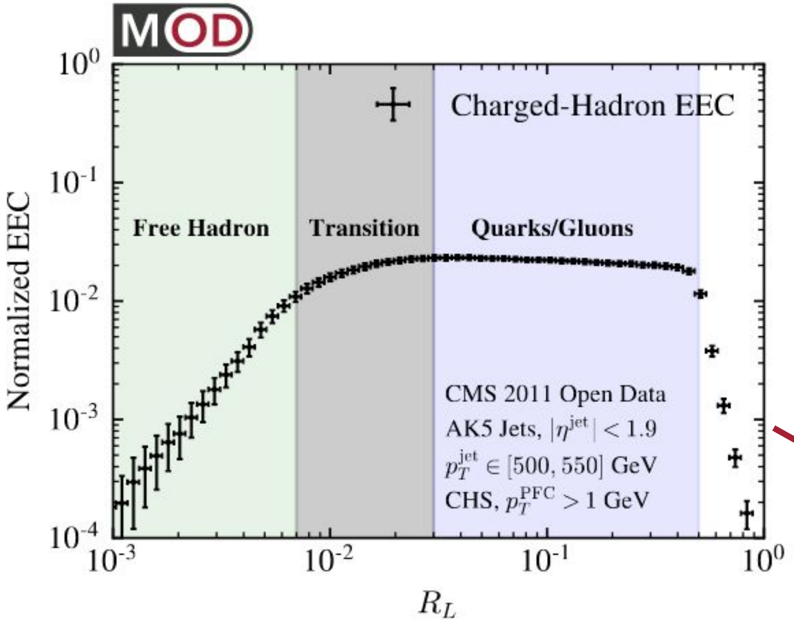


Different length scales probe different regimes of QCD!

Can see it all within Open Data!

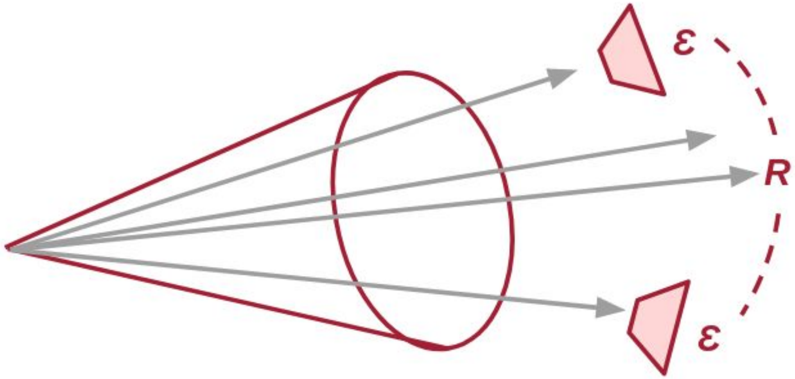
[Komiske, Moult, Thaler, Zhu, 2205.04459]

Example 1: Energy-Energy Correlators



Different length scales probe different regimes of QCD!

Can see it all within Open Data!



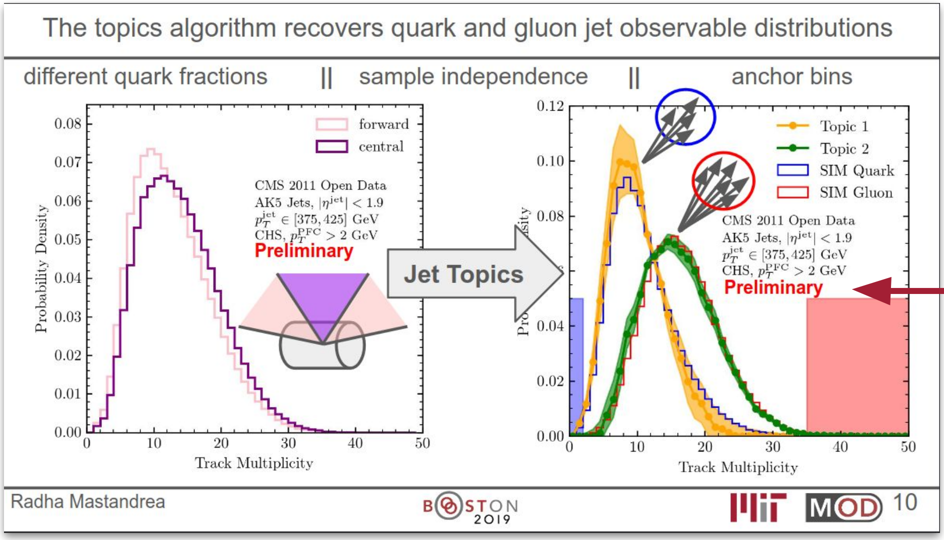
A similar measurement is now being done by CMS for Run II Data (see APS April talk).

Open Data analyses can inspire measurements!

[Metodiev, Thaler, 1802.00008]
[Komiske, Kryhin, Thaler, 2201.07800]

Example 2: Jet Topics

Slight difference in detector response for forward vs. central **quark** vs. **gluon** jets



The topics algorithm recovers quark and gluon jet observable distributions

different quark fractions || sample independence || anchor bins

Radha Mastandrea

BOSTON 2019

10

[Mastandrea, Analyzing CMS Open Collider Data through Topic Modeling (2019)]

A jet *x* is never purely a **quark** jet or a **gluon** jet, but rather a mixture:

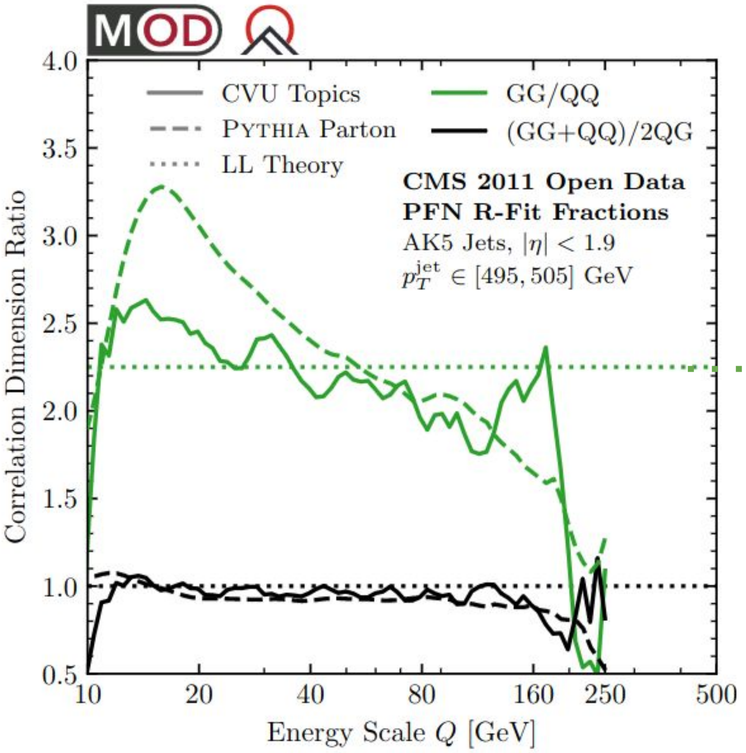
$$p_{\text{mixed}}(\vec{x}) = f_q\, p_{\text{quark}}(\vec{x}) + (1-f_q)\, p_{\text{gluon}}(\vec{x})$$

Can be used to *operationally* define quark/gluon categories, slightly different from Pythia labels, using **Topic Modeling**!

[Metodiev, Thaler, 1802.00008]
[Komiske, Kryhin, Thaler, 2201.07800]

Example 2: Jet Topics

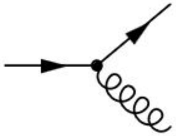
Can turn these quark/gluon distributions into measurements of **fundamental constants** of QCD in CMS Open Data!



$$dP_{i\to jk} = \frac{2\alpha_s}{\pi} C_i \frac{d\theta}{\theta}\frac{dE}{E}$$

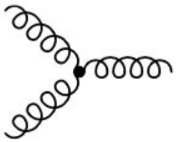
$$\frac{9}{4} = \frac{C_A}{C_F}$$ ✓

Quarks



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

Gluons

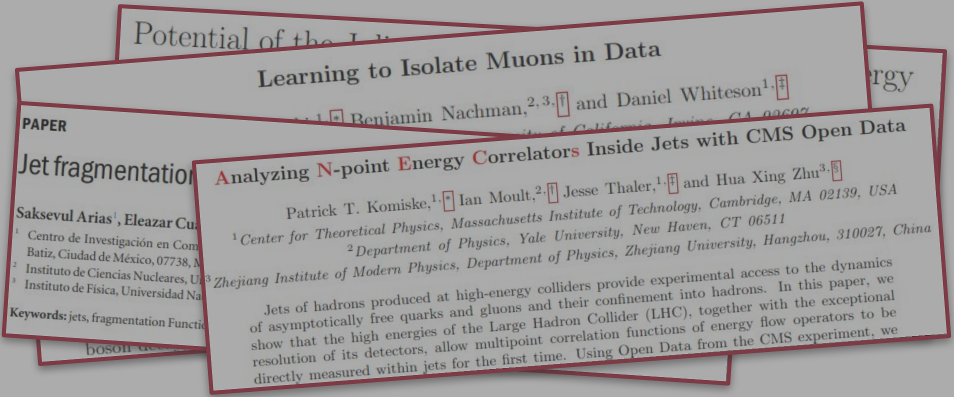


$$C_A = 3$$

Correlation dimensions are defined using Wasserstein geometry, ask me about it later!

This Talk

CMS Open Data, who uses it, and how it's being used



Potential of the I

ergy

Learning to Isolate Muons in Data

Benjamin Nachman,[2,3,†] and Daniel Whiteson[1,‡]

PAPER

Jet fragmentation

Saksevul Arias[1], Eleazar Cu

[1] Centro de Investigación en Com Batiz, Ciudad de México, 07738, M

[2] Instituto de Ciencias Nucleares, U

[3] Instituto de Física, Universidad Na

Keywords: jets, fragmentation Functi

boson

Analyzing N-point Energy Correlators Inside Jets with CMS Open Data

Patrick T. Komiske,[1,*] Ian Moult,[2,†] Jesse Thaler,[1,‡] and Hua Xing Zhu[3,§]

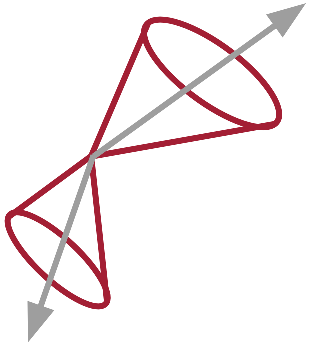
[1] Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2] Department of Physics, Yale University, New Haven, CT 06511

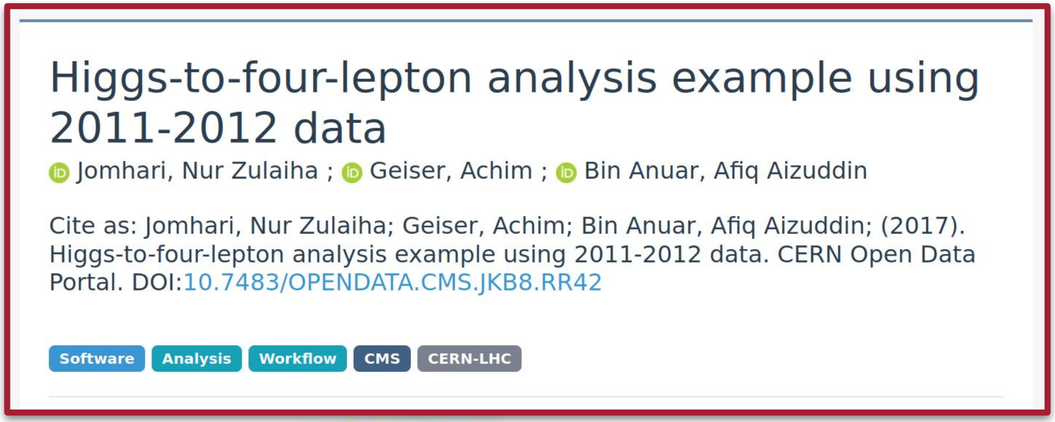
[3] Zhejiang Institute of Modern Physics, Department of Physics, Zhejiang University, Hangzhou, 310027, China

Jets of hadrons produced at high-energy colliders provide experimental access to the dynamics of asymptotically free quarks and gluons and their confinement into hadrons. In this paper, we show that the high energies of the Large Hadron Collider (LHC), together with the exceptional resolution of its detectors, allow multipoint correlation functions of energy flow operators to be directly measured within jets for the first time. Using Open Data from the CMS experiment, we

My own experiences and anecdotes with **CMS Open Data**



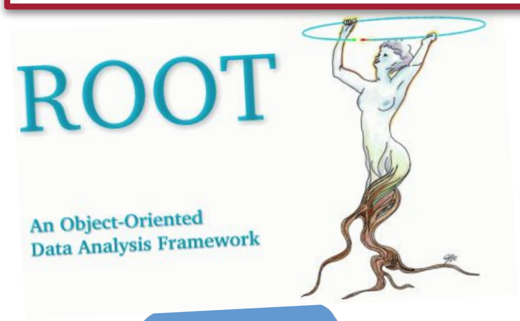
Open Data as a **teaching tool**

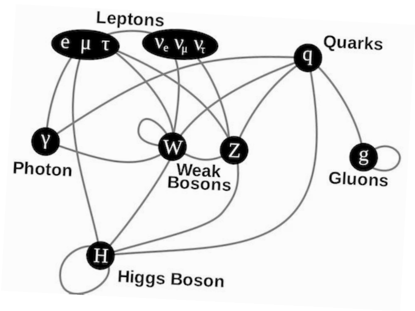


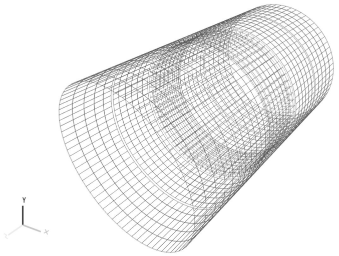
Higgs-to-four-lepton analysis example using 2011-2012 data

Jomhari, Nur Zulaiha ; Geiser, Achim ; Bin Anuar, Afiq Aizuddin

Cite as: Jomhari, Nur Zulaiha; Geiser, Achim; Bin Anuar, Afiq Aizuddin; (2017). Higgs-to-four-lepton analysis example using 2011-2012 data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.JKB8.RR42

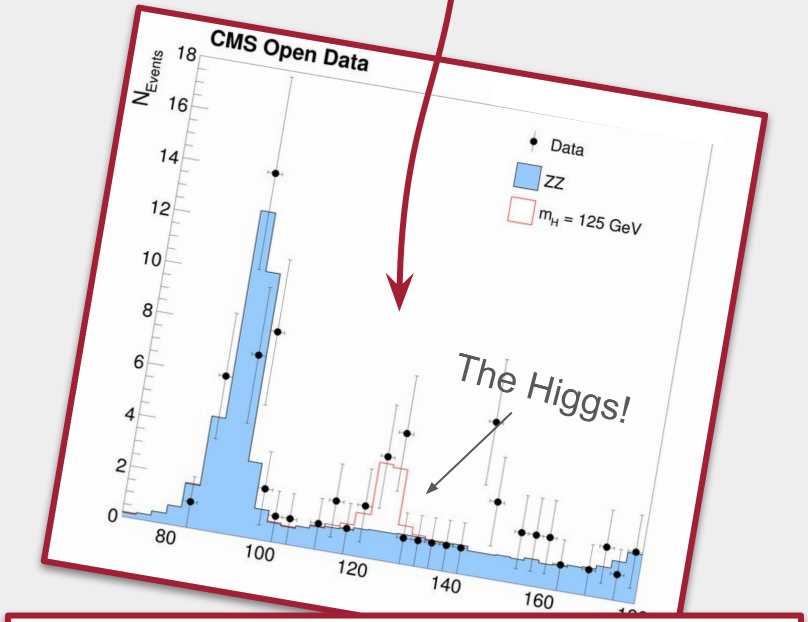






Me as an undergrad in 2018 joining the Rutgers CMS B2G Group

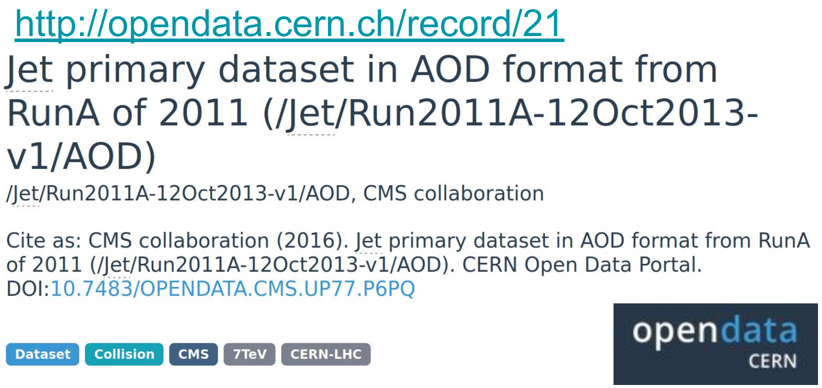
this is actually me



The Higgs!

One of my ever first plots!

My favorite dataset: CMS2011AJets



http://opendata.cern.ch/record/21

Jet primary dataset in AOD format from RunA of 2011 (/Jet/Run2011A-12Oct2013-v1/AOD)

/Jet/Run2011A-12Oct2013-v1/AOD, CMS collaboration

Cite as: CMS collaboration (2016). Jet primary dataset in AOD format from RunA of 2011 (/Jet/Run2011A-12Oct2013-v1/AOD). CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.UP77.P6PQ

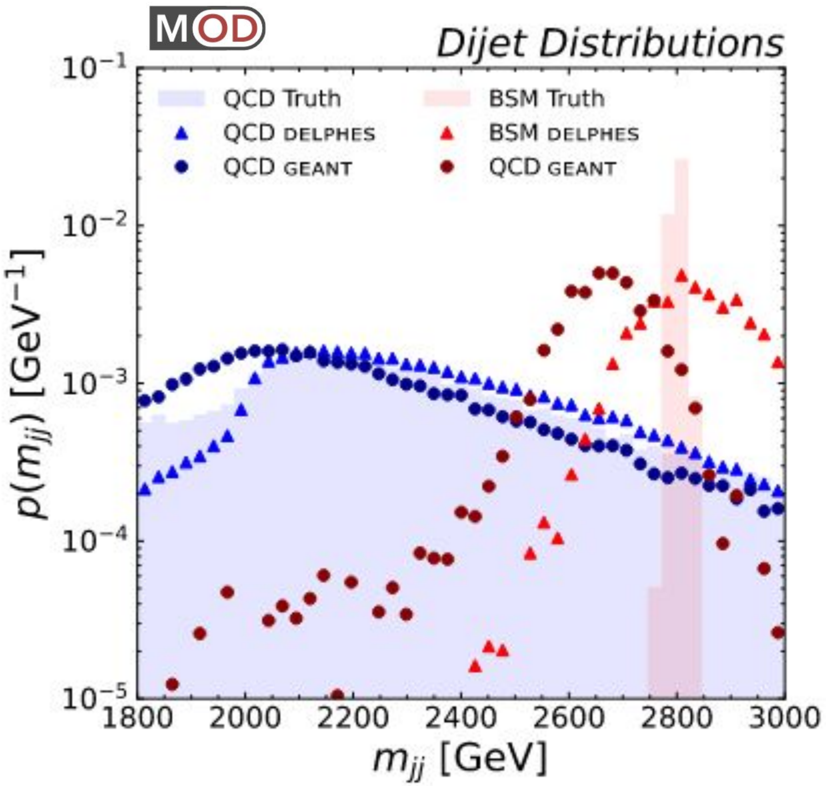
Jet Data collected in 2011 Run A

Applied `HLT Jet300` single-jet trigger

AK5 Jets with $p_T > 375$ GeV

AOD files located at Record 21, with associated MC (in both SIM/GEN varieties) at Records 1364 - 1369

Perfect for QCD & Jet studies!



Pictured: Dijet mass of QCD samples from CMS Open Sim at truth and detector level, [**RG**, Nachman, Thaler, 2205.05084]

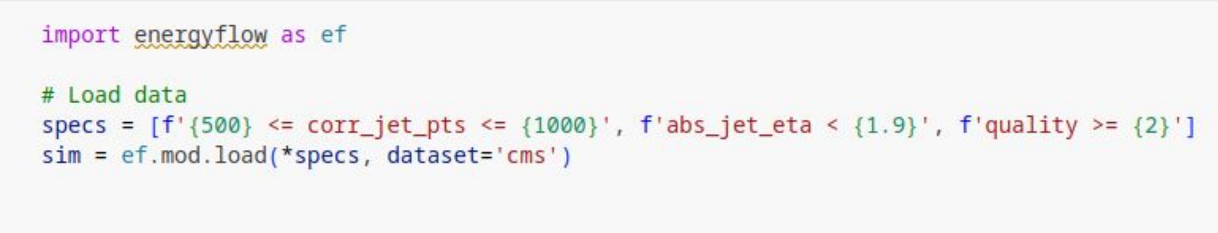
https://energyflow.network/docs/datasets/
[Komiske, Mastandrea, Metodiev, Naik, Thaler, 1908.08542]
[Tripathee, Xue, Larkoski, Naik, Marzani, 1704.05842]

My favorite way to *access* open data: The **MIT Open Data (MOD)** Format

Processed AOD files into manageable "MOD HDF5" text files hosted at https://zenodo.org/record/3340205

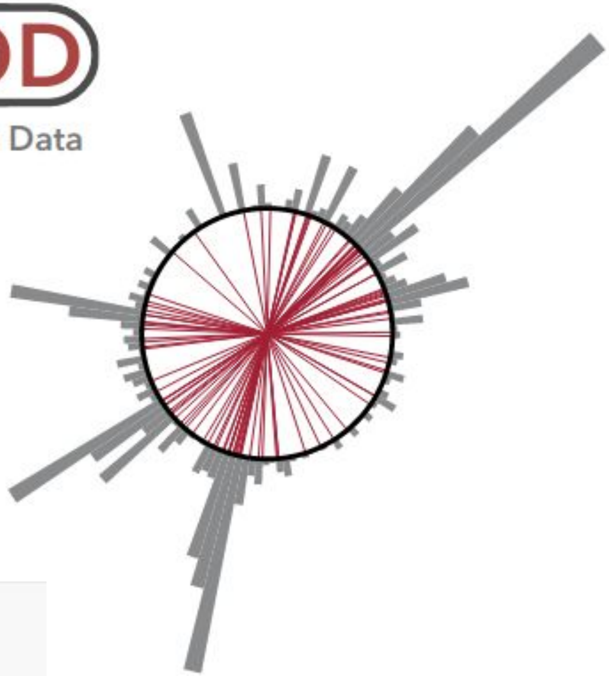
Very **easy to access** – no CMSSW, no virtual machines, no ROOT, no complicated AODs …

Can easily download *anywhere* on *any machine* with *energyflow*:



```
import energyflow as ef

# Load data
specs = [f'{500} <= corr_jet_pts <= {1000}', f'abs_jet_eta < {1.9}', f'quality >= {2}']
sim = ef.mod.load(*specs, dataset='cms')
```

MIT Open Data



Try *pip install energyflow*

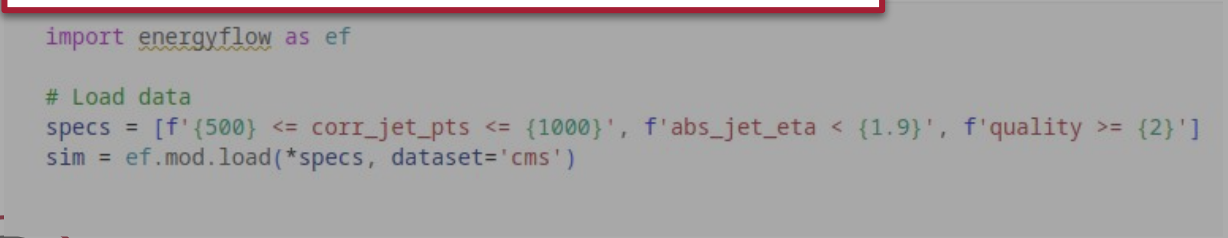
https://energyflow.network/docs/datasets/
[Komiske, Mastandrea, Metodiev, Naik, Thaler, 1908.08542]
[Tripathee, Xue, Larkoski, Naik, Marzani, 1704.05842]

My favorite way to *access* open data: The MIT Open Data (MOD) Format

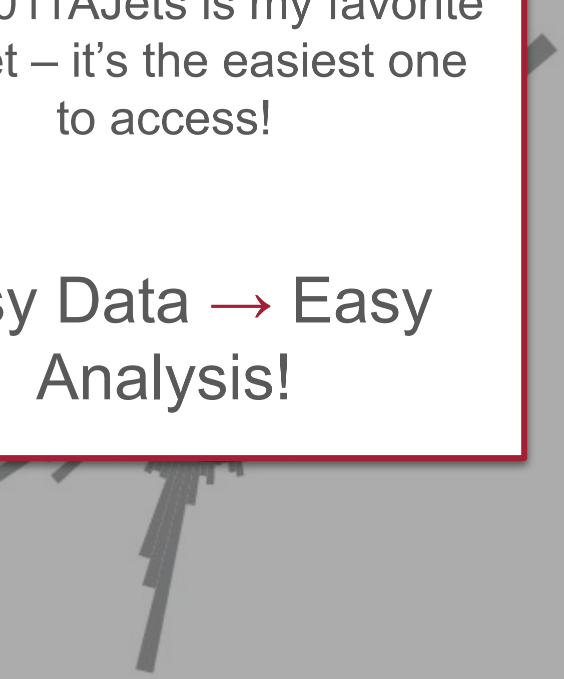
Processed AOD files into manageable "MOD HDF5" text files hosted at https://zenodo.org/record/3340205

Very **easy to access** – no CMSSW, no virtual machines, no ROOT, no complicated AODs …

Can easily download *anywhere* on *any machine* with *energyflow*:



```
import energyflow as ef

# Load data
specs = [f'{500} <= corr_jet_pts <= {1000}', f'abs_jet_eta < {1.9}', f'quality >= {2}']
sim = ef.mod.load(*specs, dataset='cms')
```



This is the reason why CMS2011AJets is my favorite dataset – it's the easiest one to access!

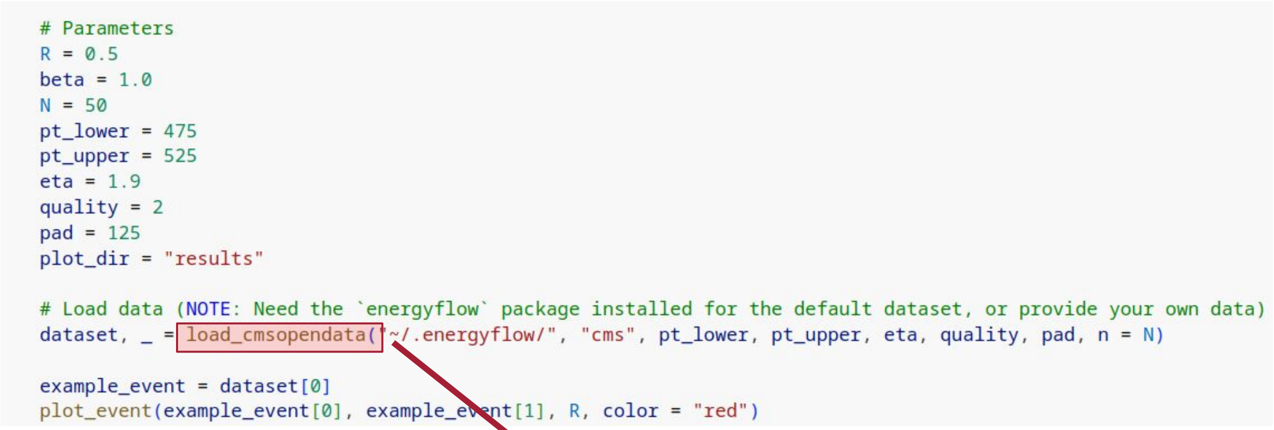
Easy Data → Easy Analysis!

Try *pip install energyflow*

This code follows the example analysis I made at
https://github.com/rikab/SHAPER/blob/main/examples/example.ipynb
[Ba, Dogra, **RG**, Tasissa, Thaler, 2302.12266]
Try *pip install pyshaper*

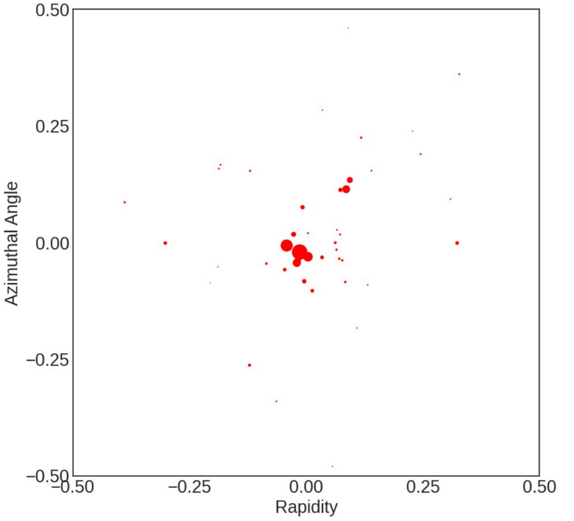
My typical **workflow**:

Step 1: Download CMS Open Data!



```
# Parameters
R = 0.5
beta = 1.0
N = 50
pt_lower = 475
pt_upper = 525
eta = 1.9
quality = 2
pad = 125
plot_dir = "results"

# Load data (NOTE: Need the `energyflow` package installed for the default dataset, or provide your own data)
dataset, _ = load_cmsopendata("~/.energyflow/", "cms", pt_lower, pt_upper, eta, quality, pad, n = N)

example_event = dataset[0]
plot_event(example_event[0], example_event[1], R, color = "red")
```

Pictured: An AK5 Jet measured during Run A in 2011



Downloads the **CMS2011AJets** dataset using MOD, does minor preprocessing, and converts to *np* arrays

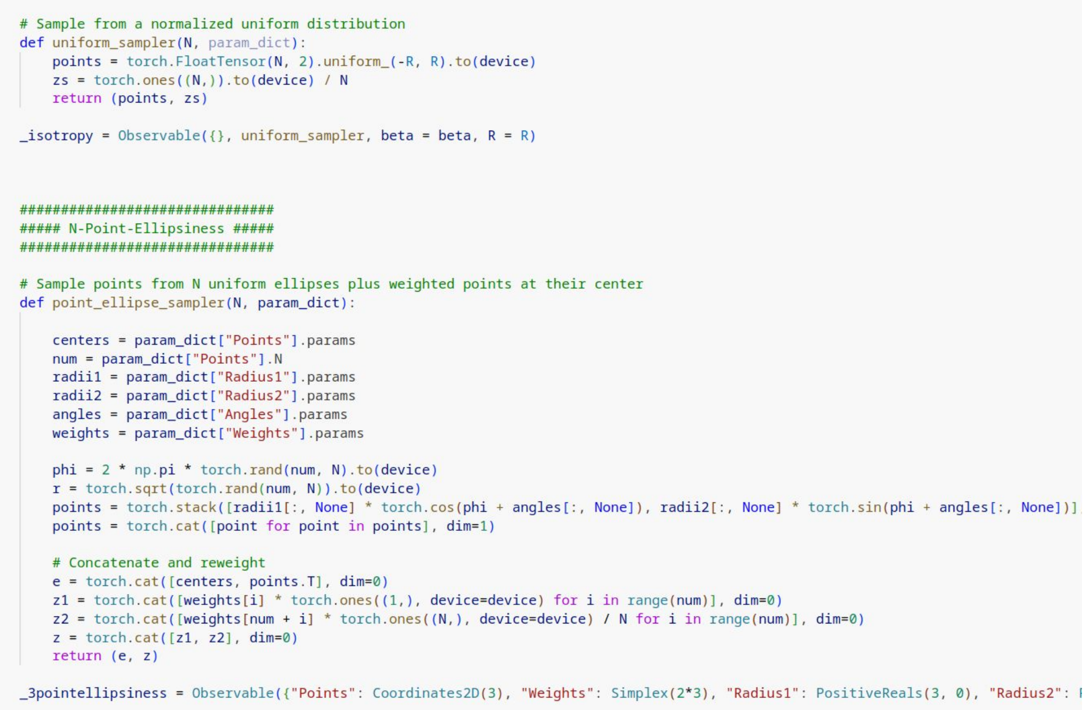
On a fresh machine, takes only 5 minutes to download a 100,000 jet sample

Ease of download makes open data great as an example data set (especially for **tutorials**)! I don't have to worry about Pythia, Geant, etc …

This code follows the example analysis I made at
https://github.com/rikab/SHAPER/blob/main/examples/example.ipynb
[Ba, Dogra, **RG**, Tasissa, Thaler, 2302.12266]
Try *pip install pyshaper*

My typical **workflow**:

Step 2: Set up calculations, e.g.



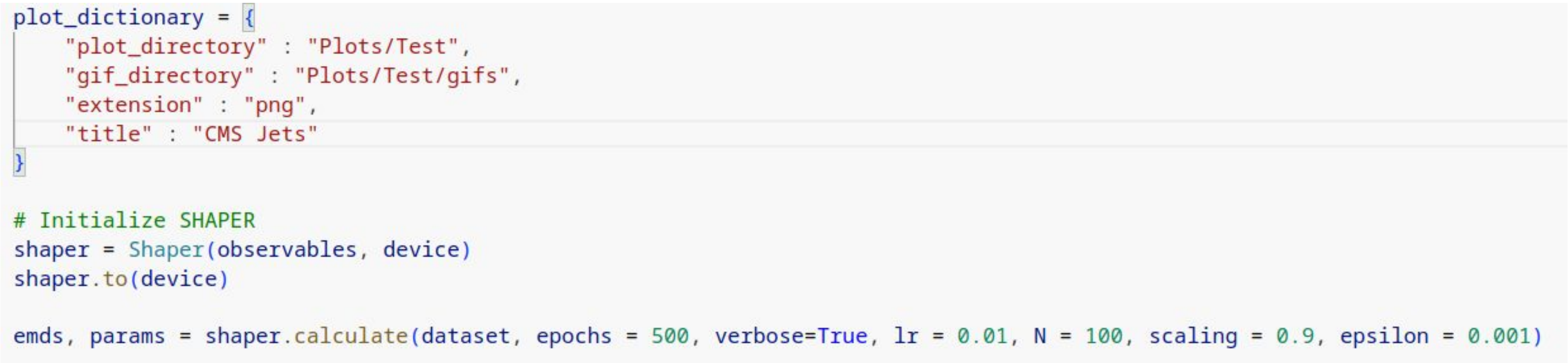
```
# Sample from a normalized uniform distribution
def uniform_sampler(N, param_dict):
    points = torch.FloatTensor(N, 2).uniform_(-R, R).to(device)
    zs = torch.ones((N,)).to(device) / N
    return (points, zs)

_isotropy = Observable({}, uniform_sampler, beta = beta, R = R)



###############################
##### N-Point-Ellipsiness #####
###############################

# Sample points from N uniform ellipses plus weighted points at their center
def point_ellipse_sampler(N, param_dict):

    centers = param_dict["Points"].params
    num = param_dict["Points"].N
    radii1 = param_dict["Radius1"].params
    radii2 = param_dict["Radius2"].params
    angles = param_dict["Angles"].params
    weights = param_dict["Weights"].params

    phi = 2 * np.pi * torch.rand(num, N).to(device)
    r = torch.sqrt(torch.rand(num, N)).to(device)
    points = torch.stack([radii1[:, None] * torch.cos(phi + angles[:, None]), radii2[:, None] * torch.sin(phi + angles[:, None])]
    points = torch.cat([point for point in points], dim=1)

    # Concatenate and reweight
    e = torch.cat([centers, points.T], dim=0)
    z1 = torch.cat([weights[i] * torch.ones((1,), device=device) for i in range(num)], dim=0)
    z2 = torch.cat([weights[num + i] * torch.ones((N,), device=device) / N for i in range(num)], dim=0)
    z = torch.cat([z1, z2], dim=0)
    return (e, z)

_3pointellipsiness = Observable({"Points": Coordinates2D(3), "Weights": Simplex(2*3), "Radius1": PositiveReals(3, 0), "Radius2":
```

For me, this usually involves defining QCD observables or building ML tools to act on the data – This is where all the physics happens!

This code follows the example analysis I made at
https://github.com/rikab/SHAPER/blob/main/examples/example.ipynb
[Ba, Dogra, **RG**, Tasissa, Thaler, 2302.12266]
Try *pip install pyshaper*

My typical **workflow**:

Step 3: Run all calculations on the data!



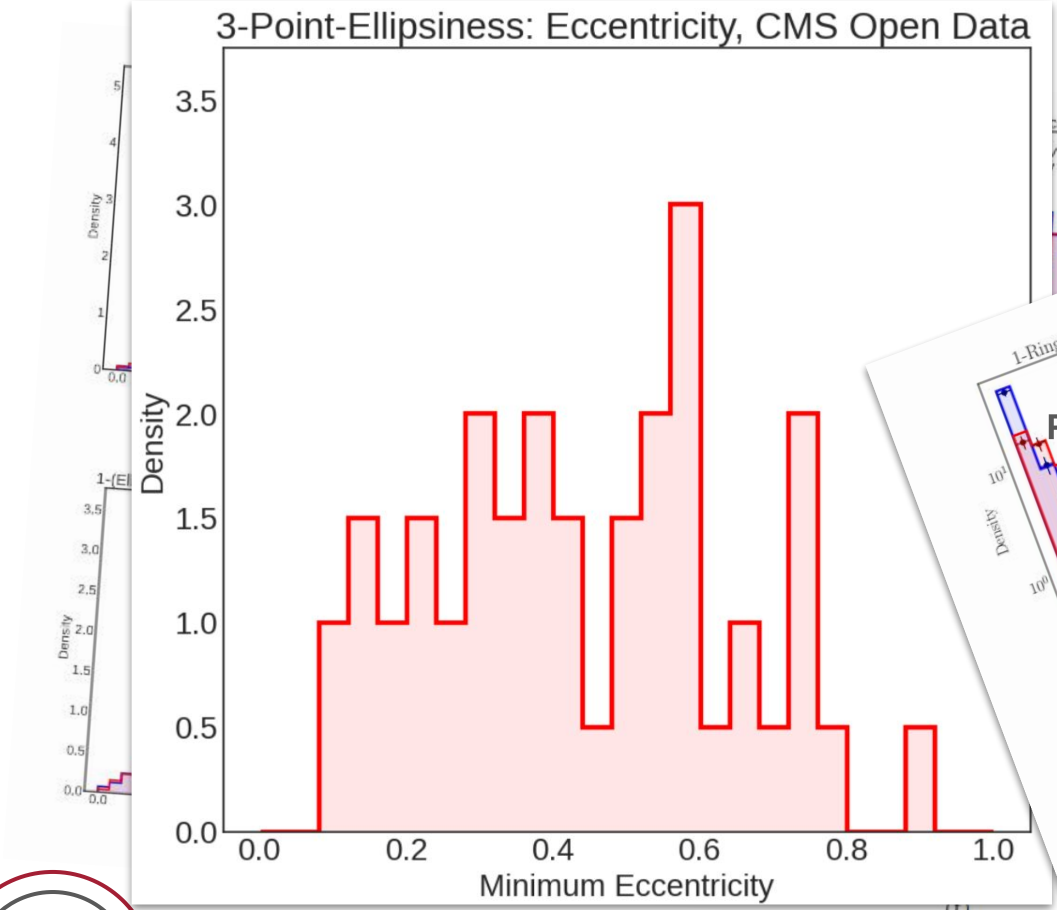
```
plot_dictionary = {
    "plot_directory" : "Plots/Test",
    "gif_directory" : "Plots/Test/gifs",
    "extension" : "png",
    "title" : "CMS Jets"
}

# Initialize SHAPER
shaper = Shaper(observables, device)
shaper.to(device)

emds, params = shaper.calculate(dataset, epochs = 500, verbose=True, lr = 0.01, N = 100, scaling = 0.9, epsilon = 0.001)
```

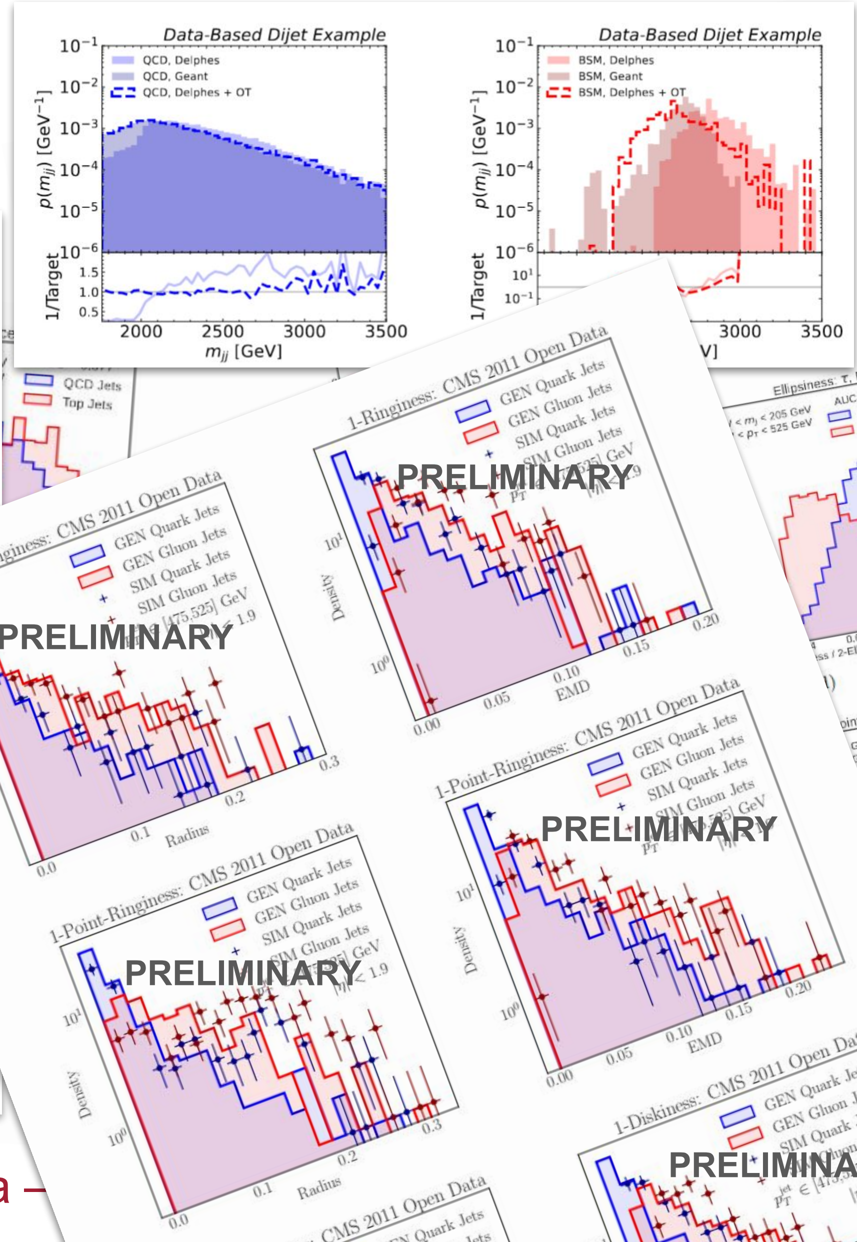
(Often done on a big cluster rather than a Jupyter notebook …)

This code follows the example analysis I made at
https://github.com/rikab/SHAPER/blob/main/examples/example.ipynb
[Ba, Dogra, **RG**, Tasissa, Thaler, 2302.12266]
[**RG**, Thaler, Wu, WIP]

My typical **workflow**:

Step 4: Plots Plots Plots Plots Plots!





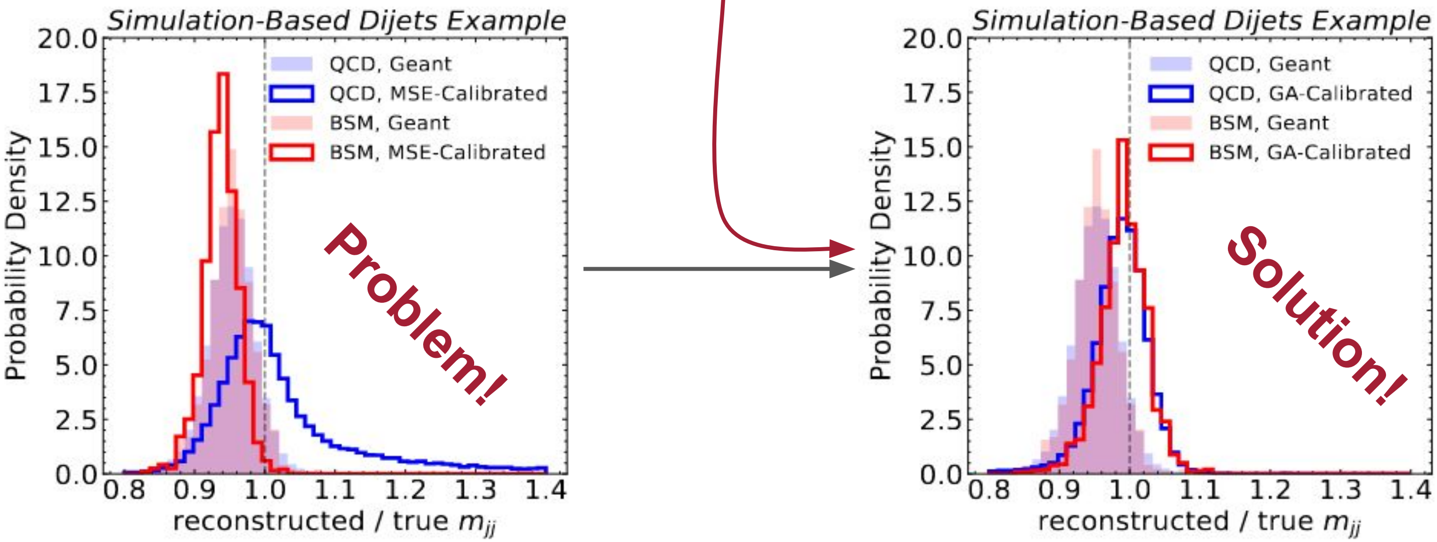
[RG, Nachman, Thaler, 2205.05084]

Try `pip install GaussianAnsatz`

CMS Open Sim for Calibration

$$
\begin{aligned}
T(x,z) = A(x)& \\
&+ (z - B(x)) \cdot D(x) \\
&+ \frac{1}{2}(z - B(x))^T \cdot C(x,z) \cdot (z - B(x))
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{L}_{\mathrm{DVR}}[T] = -\Big(&\mathbb{E}_{P_{XZ}}[T] - \log\left(\mathbb{E}_{P_X \otimes P_Z}\left[e^T\right]\right)\Big) \\
&+ \lambda_D \mathbb{E}_{P_{XZ}} |D(X)|
\end{aligned}
$$



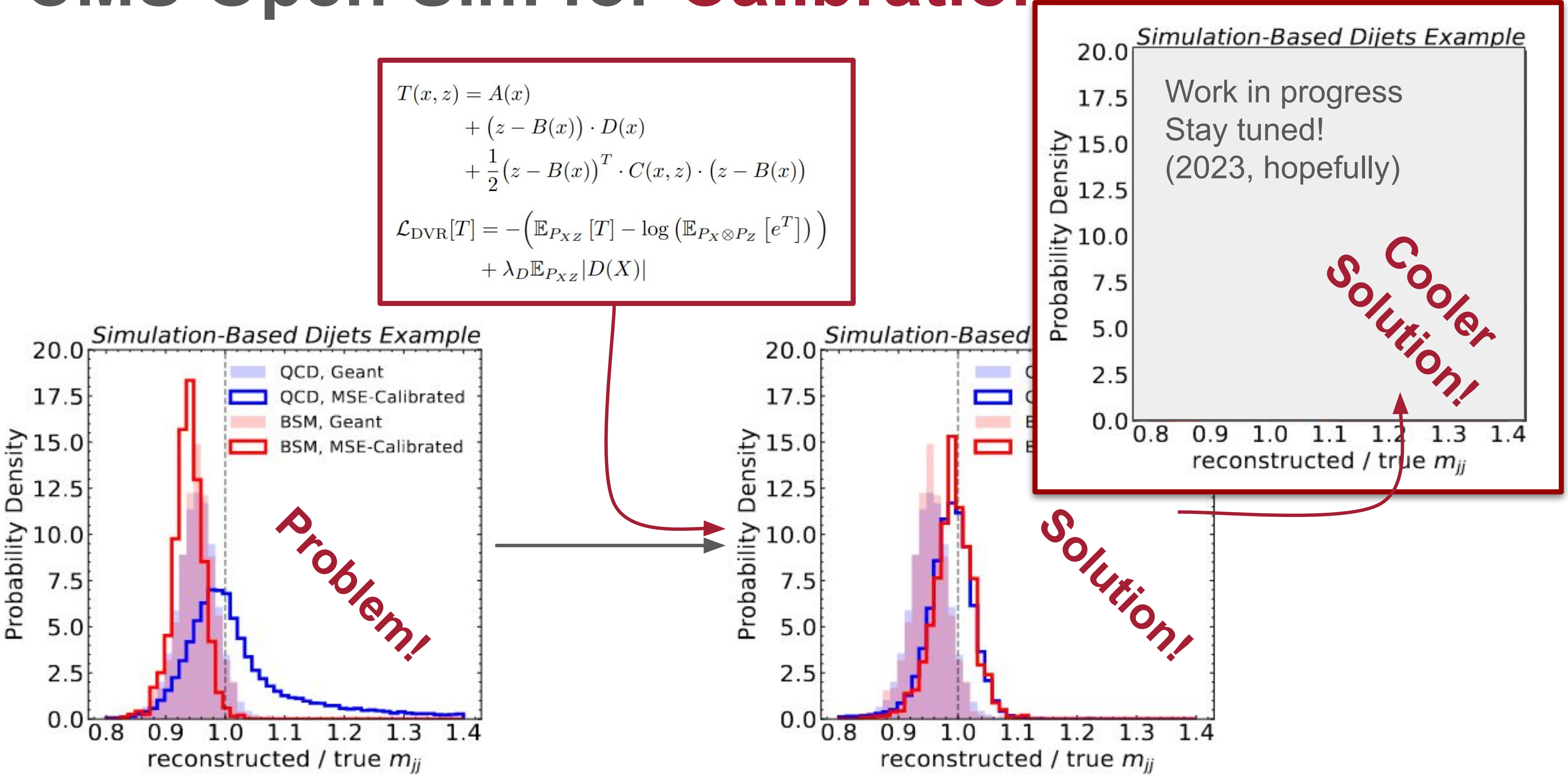
Problem!

Solution!

… Using Open Data as an easy, realistic example dataset for **ML studies** and **calibration**!

[**RG**, Nachman, Thaler, 2205.05084
[**RG**, Nachman, Thaler, 2205.05084]
[**RG**, Nachman, Thaler, WIP]
Try `pip install GaussianAnsatz`

CMS Open Sim for **Calibration**



$$T(x,z) = A(x) + (z - B(x)) \cdot D(x) + \frac{1}{2}(z - B(x))^T \cdot C(x,z) \cdot (z - B(x))$$

$$\mathcal{L}_{\mathrm{DVR}}[T] = -\left(\mathbb{E}_{P_{XZ}}[T] - \log\left(\mathbb{E}_{P_X \otimes P_Z}\left[e^T\right]\right)\right) + \lambda_D \mathbb{E}_{P_{XZ}}|D(X)|$$

Problem!

Solution!

Work in progress
Stay tuned!
(2023, hopefully)

Cooler Solution!

… Using Open Data as an easy, realistic example dataset for **ML studies** and **calibration**!

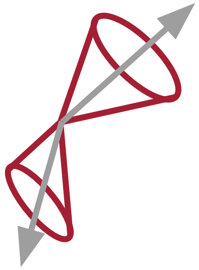
[RG, Nachman, Thaler, 2205.03413]

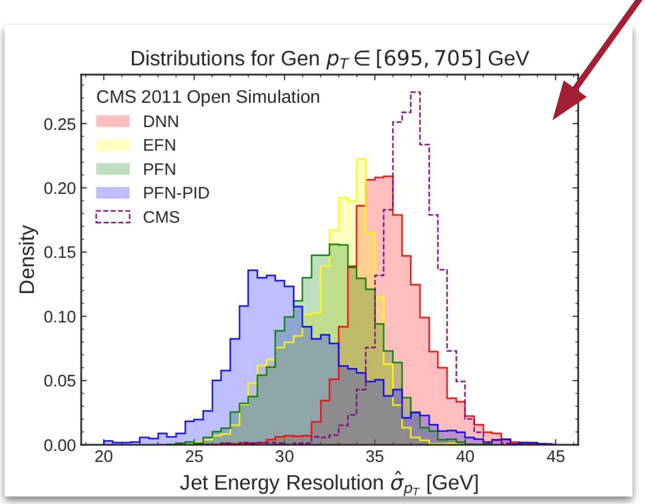
Try `pip install GaussianAnsatz`

CMS Open Sim for **Uncertainty Estimation**

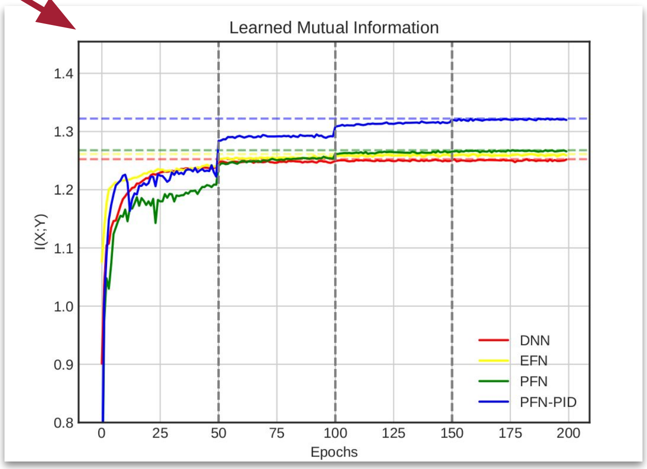
MOD
MIT Open Data

Gaussian Ansatz




Distributions for Gen $p_T \in [695, 705]$ GeV
CMS 2011 Open Simulation
DNN
EFN
PFN
PFN-PID
CMS
Density
0.00
0.05
0.10
0.15
0.20
0.25
20
25
30
35
40
45
Jet Energy Resolution $\hat{\sigma}_{p_T}$ [GeV]

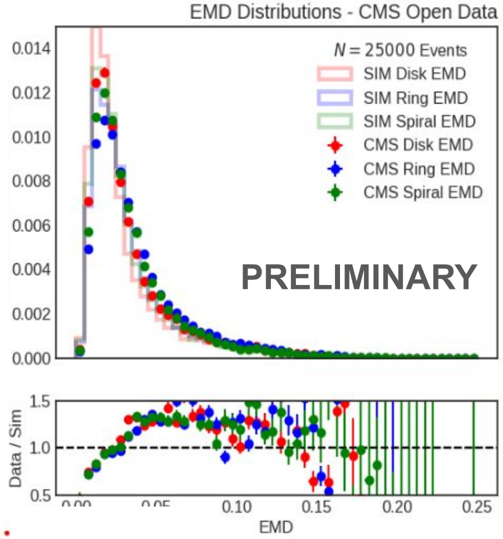
Model	Avg Resolution [GeV]	Mutual Information *I(X;Z)*
DNN	35.7 ± 2.1	1.23
EFN	32.6 ± 2.3	1.26
PFN	32.5 ± 2.5	1.27
PFN-PID	**30.8 ± 3.6**	**1.32**
CMS Open Data	36.9 ± 1.7	—

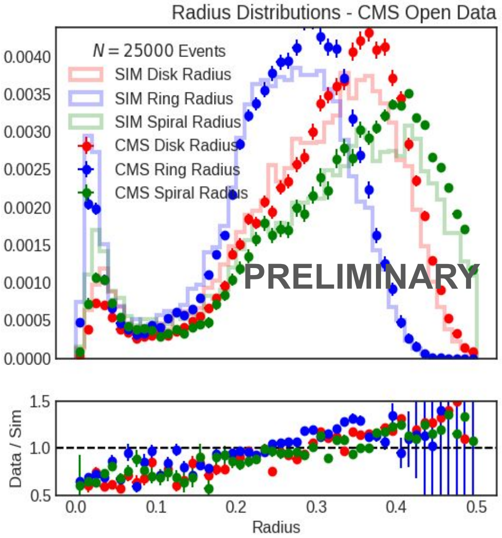


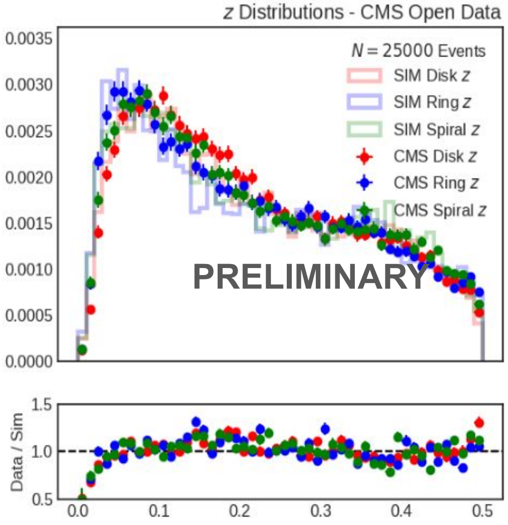
… Using Open Data to understand **detector efforts** and quantify **uncertainties and correlations** with **ML**!

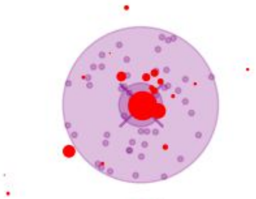
[Ba, Dogra, **RG**, Tasissa, Thaler, 2302.12266]

Hearing the **Shapes of Jets**

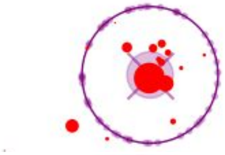




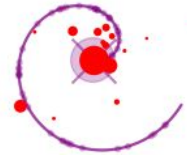




Disk +
δ-function



Ring + δ-function

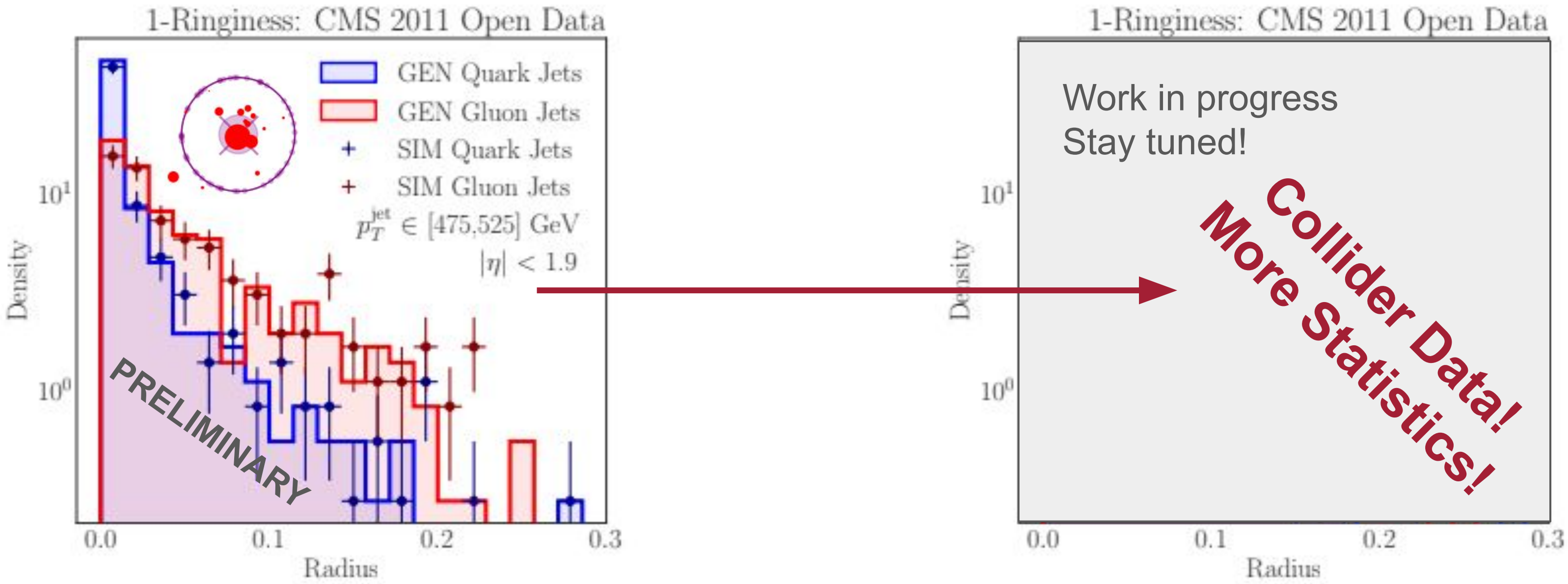


Spiral + δ-function

Probing **collinear** (δ-function) and **soft** (shape) structure!
Are some aspects of substructure not well modeled in Pythia? Check against data!

[RG, Thaler, Wu, WIP]

How **wide** are QCD jets?



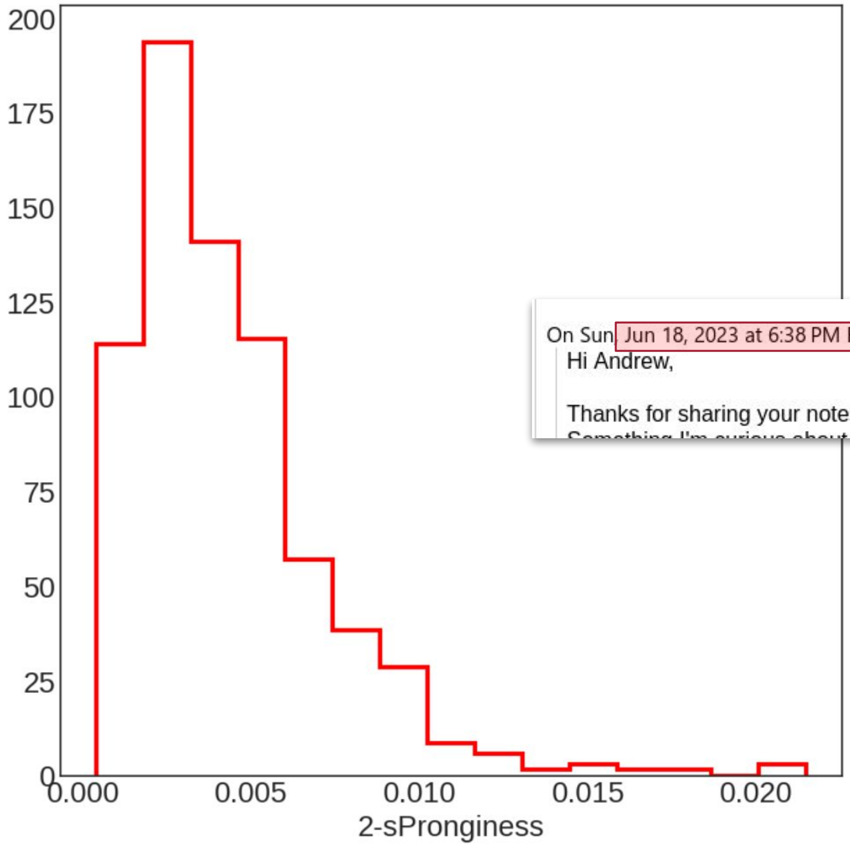
Work in progress
Stay tuned!

Collider Data!
More Statistics!

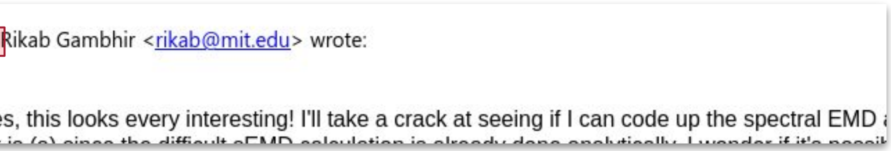
Determining the radius distribution of q/g jets in data with an **MIT Summer Research Program undergrad** (Xinyue Wu)! From zero to this in a few weeks!

[RG, Larkoski, Thaler, WIP]

Prototyping new metrics



Easy to use CMS2011AJets as a realistic dataset to prototype new things without having to generate my own data!

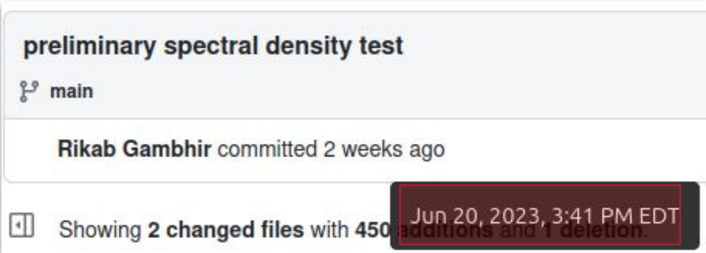


On Sun, Jun 18, 2023 at 6:38 PM Rikab Gambhir <rikab@mit.edu> wrote:

Hi Andrew,

Thanks for sharing your notes, this looks every interesting! I'll take a crack at seeing if I can code up the spectral EMD a

Something I'm curious about is (a) since the difficult sEMD calculation is already done analytically, I wonder if it's possi



preliminary spectral density test

main

Rikab Gambhir committed 2 weeks ago

Showing 2 changed files with 450 additions and 1 deletion.

Jun 20, 2023, 3:41 PM EDT

How do I take my cup of CMS Open Data?

- Very easily accessible anywhere I am
- Takes only a few seconds to minutes to set up
- Highly preprocessed and prepackaged
- Don’t have to understand all the details of how it was made
- Helps me make plots
- Can order online
- Made by somebody else
- Contains flavor information

Admittedly, the last few are a stretch

Start your day with a cup of
CMS open data
DD
DUNKIN' DONUTS®
Available at a computer near you!

Photo by Kelly Sikkema on Unsplash

How do I take my

Conclusion

Open

Data?

Start your day with a cup of

But it’s good to have some variety in coffee!

How can we enable more datasets to be made easily accessible and useable?

Photo by Kelly Sikkema on Unsplash

Admittedly, the last few are a stretch