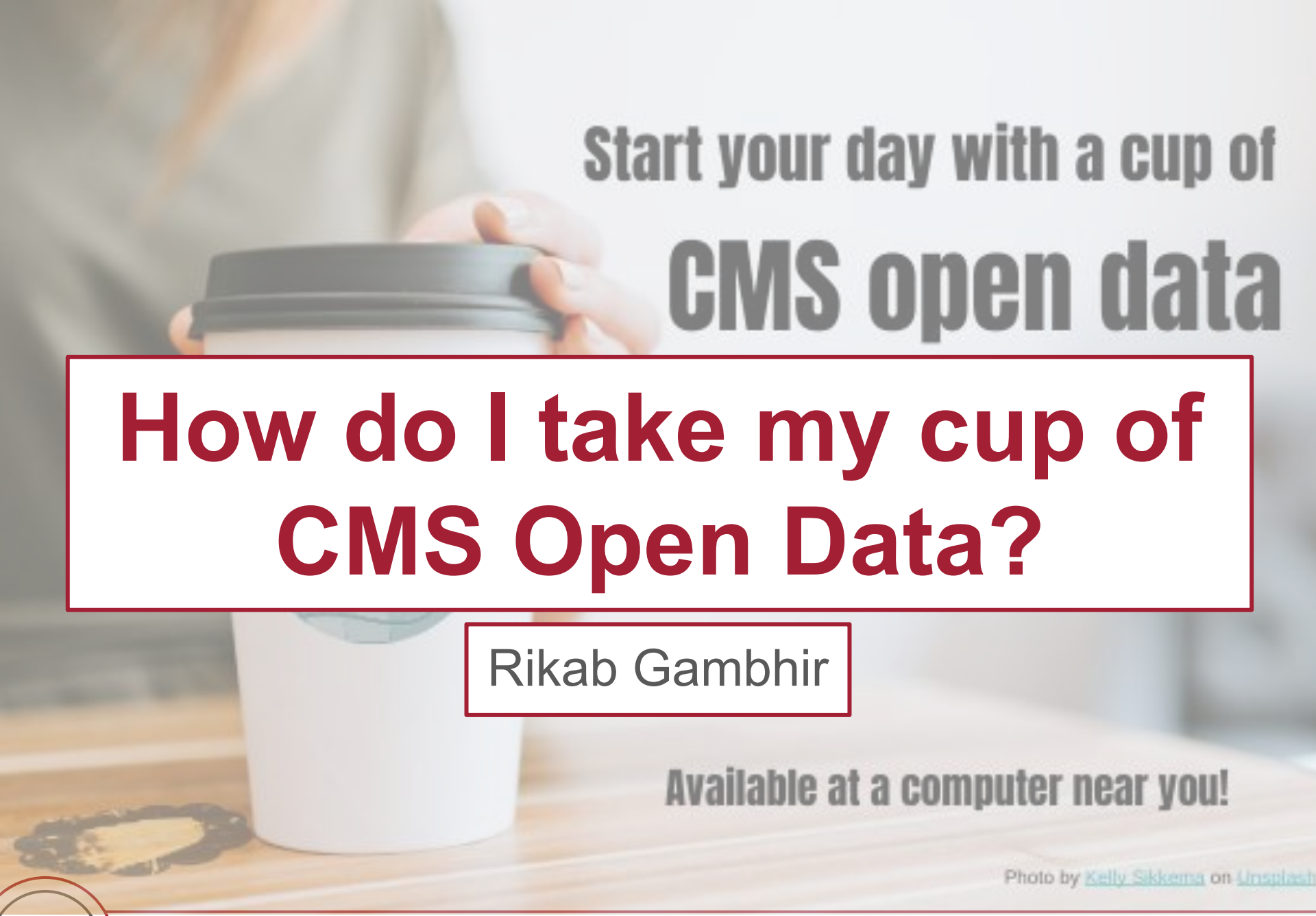


How do I take my cup of CMS Open Data?

Rikab Gambhir

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

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

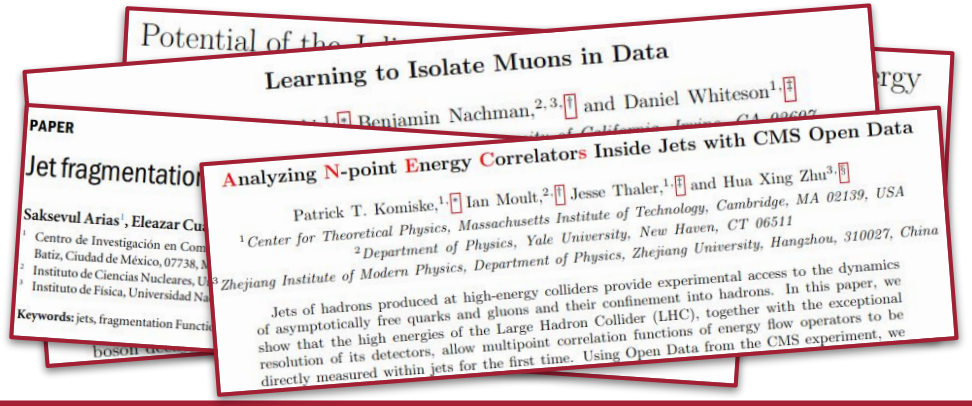
Admittedly, the last few are a stretch

I like my CMS
like I
like my coffee

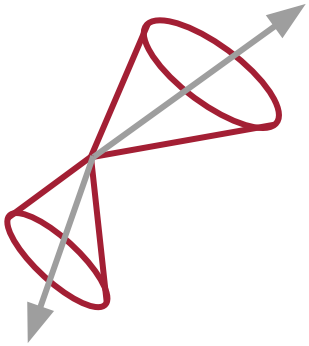
Start your day with a cup of

This Talk

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



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



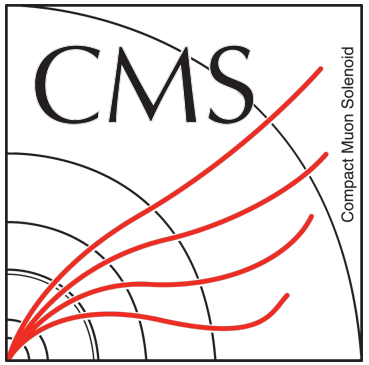
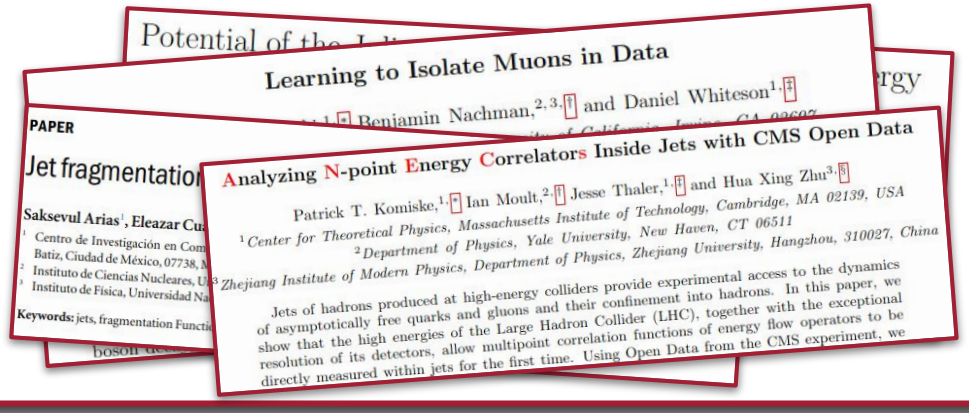


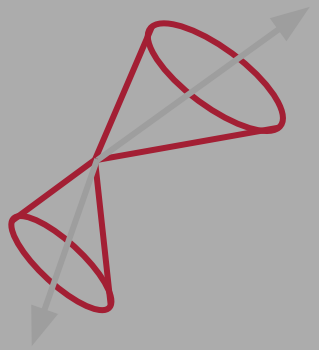
Photo by Kelly Sikkema on Unsplash

This Talk

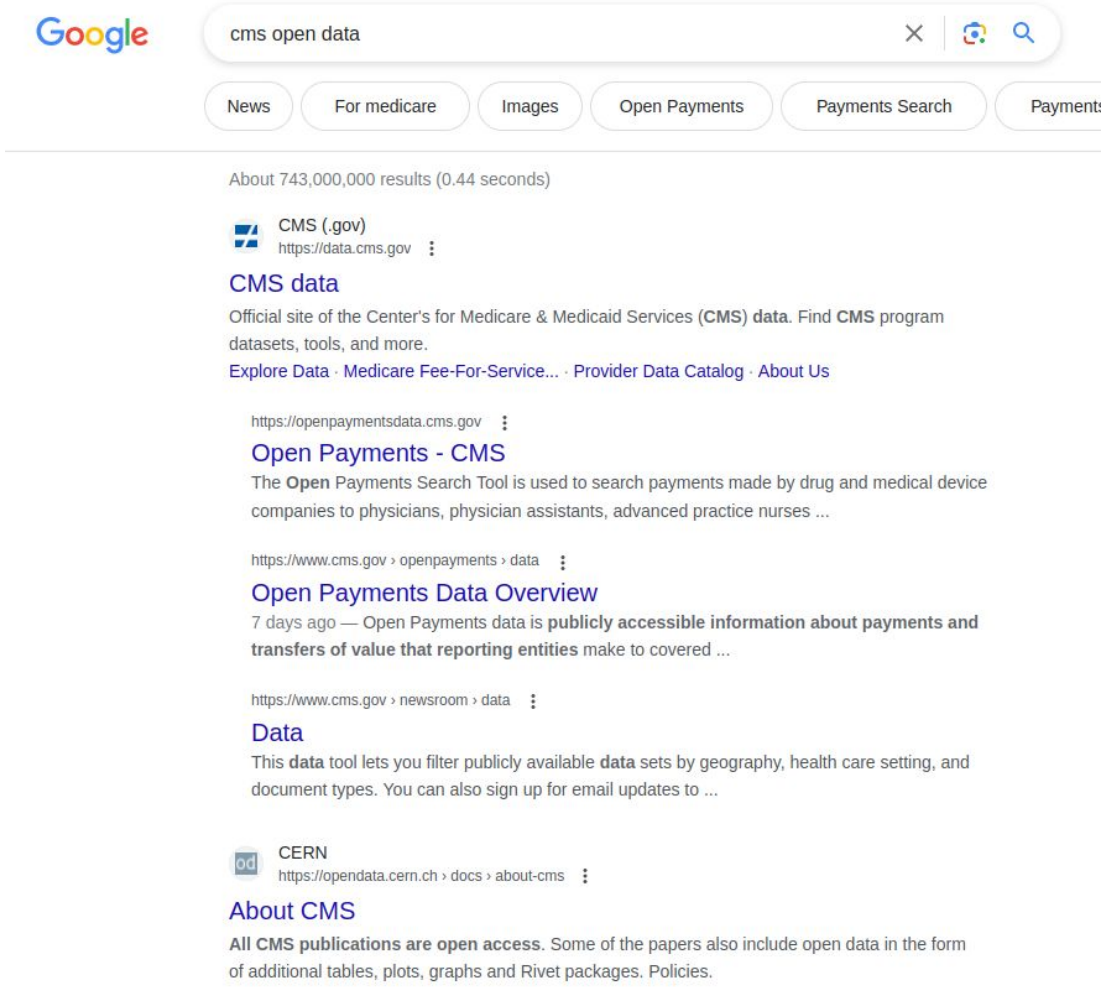
CMS Open Data, who uses it, and how it's being used



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

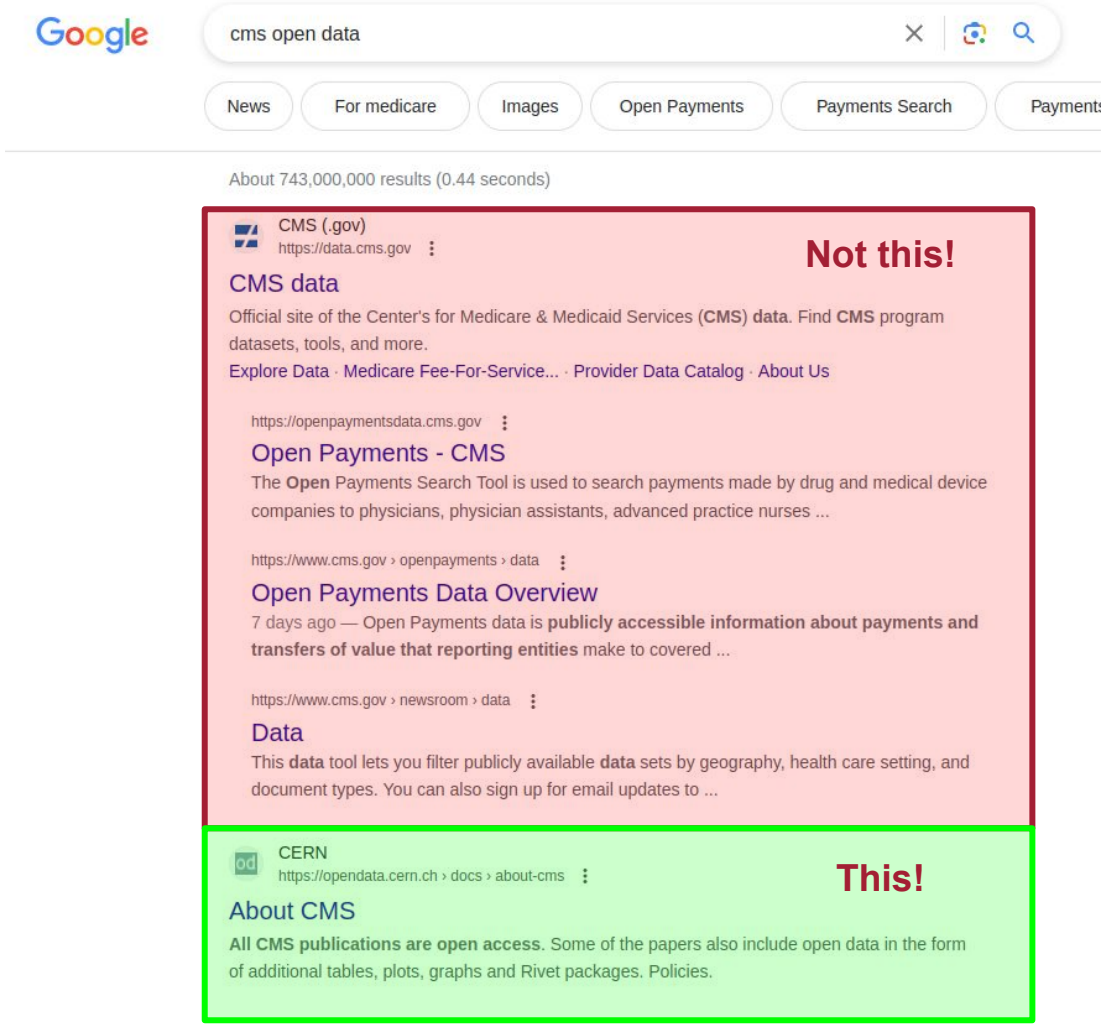


CMS Open Data

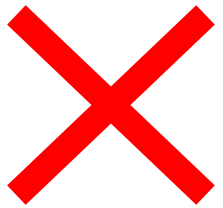


According to Google…

CMS Open Data

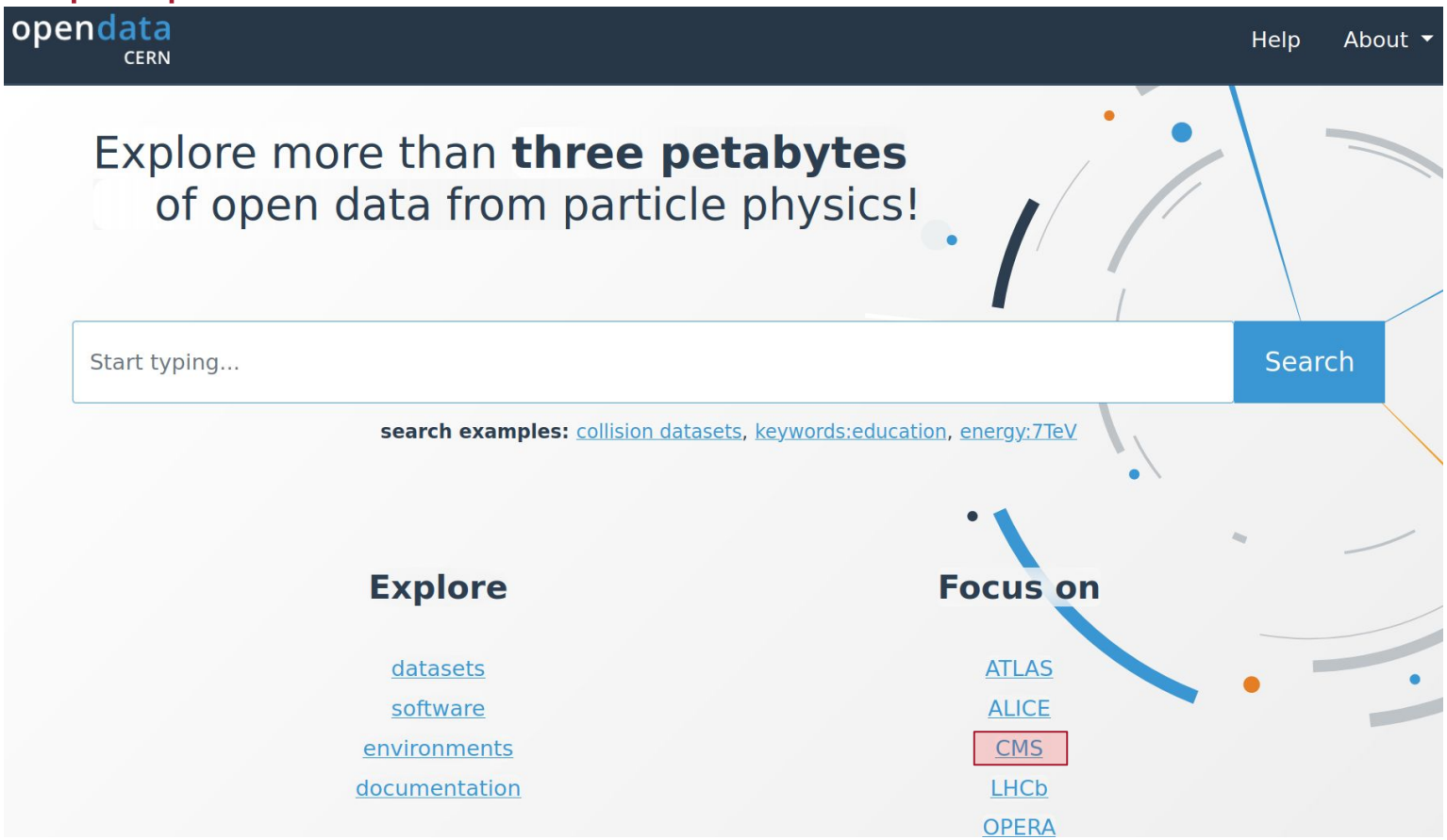


According to Google…



CMS Open Data

http://opendata.cern.ch/



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!)

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

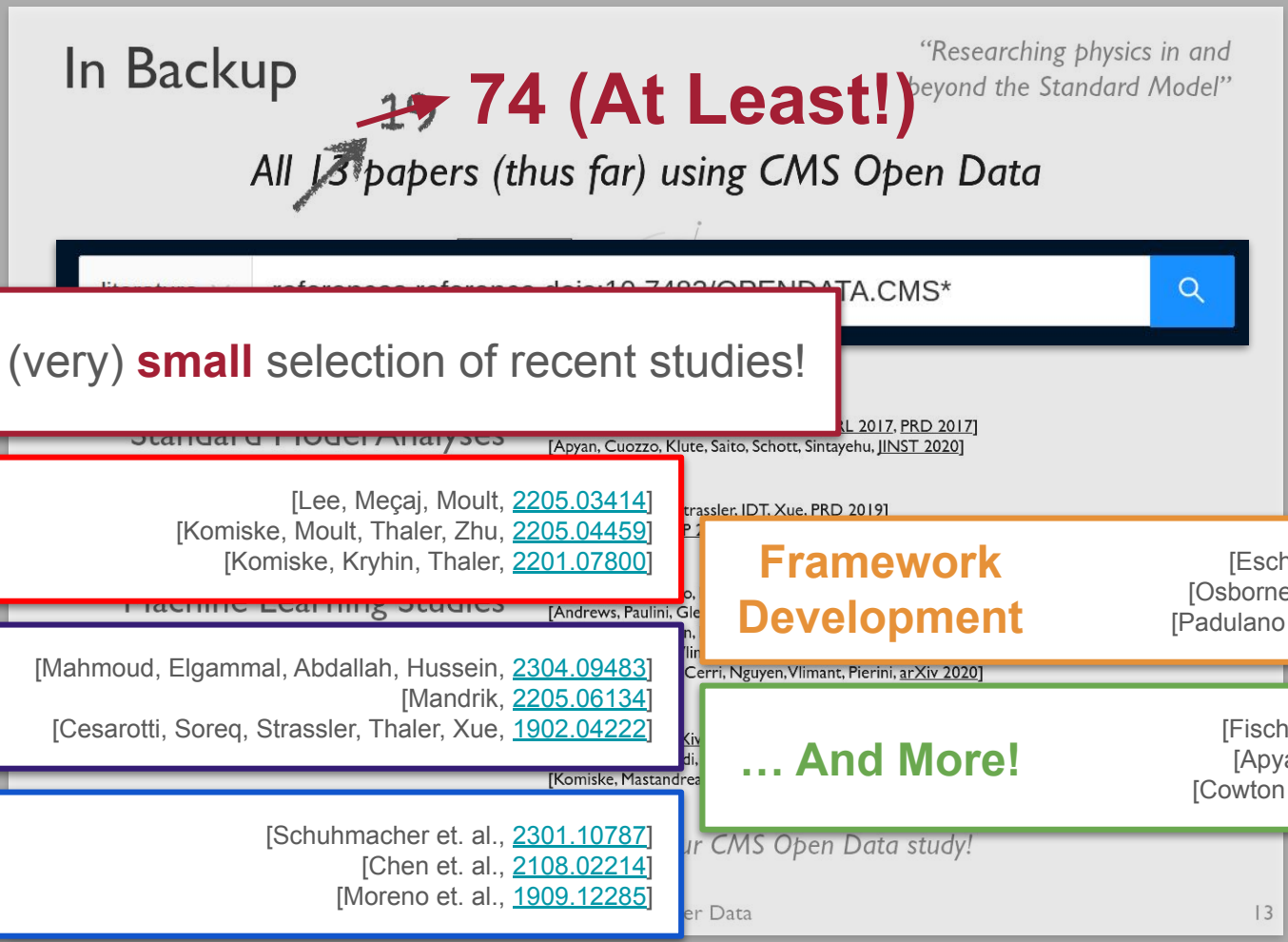
Standard Model Analyses	[Tripathee, Xue, Larkoski, Marzani, JDT, PRL 2017, PRD 2017] [Apyan, Cuozzo, Klute, Saito, Schott, Sintayehu, JINST 2020]
BSM Searches	[Cesarotti, Soreq, Strassler, JDT, Xue, PRD 2019] [Lester, Schott, JHEP 2019]
Machine Learning Studies	[Fernández Madrazo, Heredia Cacha, Lloret 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…

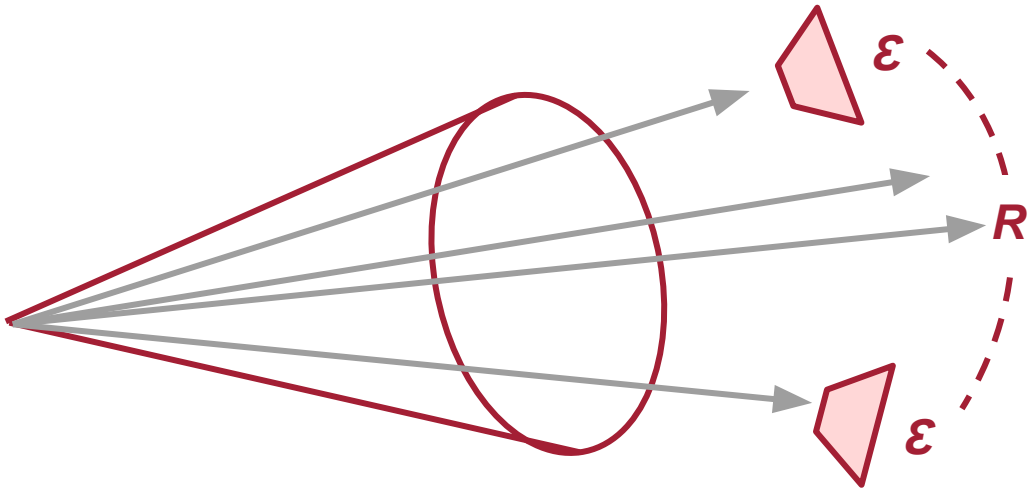


[Thaler, Adventures with Public Collider Data (2020)]

Some Fun Recent Highlights

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

Example 1: Energy-Energy Correlators



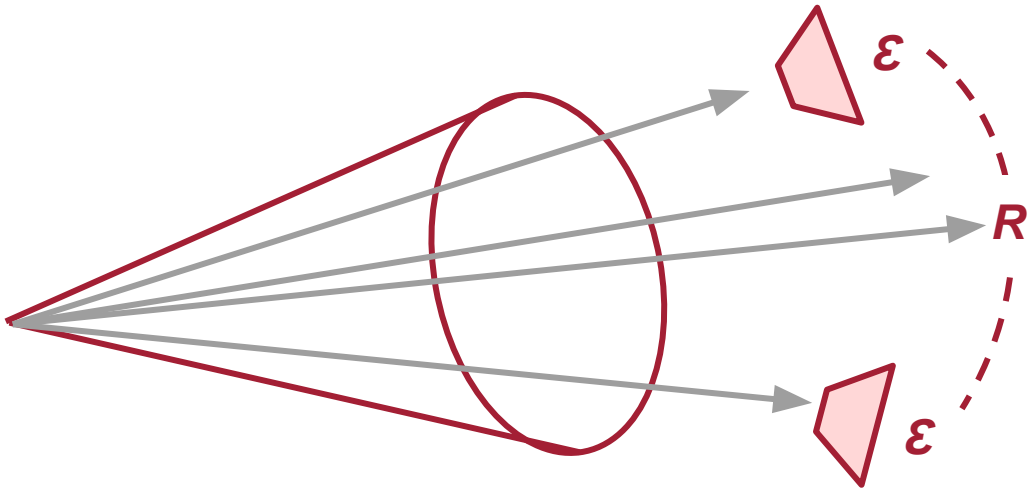
$$\mathrm{ENC}(R_L) = \left(\prod_{k=1}^{N}\int d\Omega_{\vec{n}_k}\right)\delta(R_L - \Delta\hat{R}_L) \cdot \frac{1}{(E_{\mathrm{jet}})^N}\langle\mathcal{E}(\vec{n}_1)\mathcal{E}(\vec{n}_2)\ldots\mathcal{E}(\vec{n}_N)\rangle$$

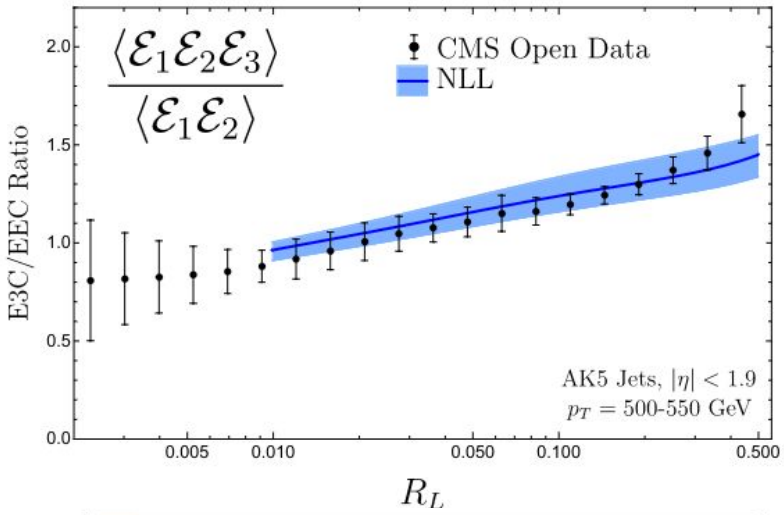
Energy-Energy Correlators (*EEC*s (and $E^N C$s)) let us explore different aspects of QCD, including scaling behavior, collinear structure, phase transitions, and more

Explored in CMS Open Data!

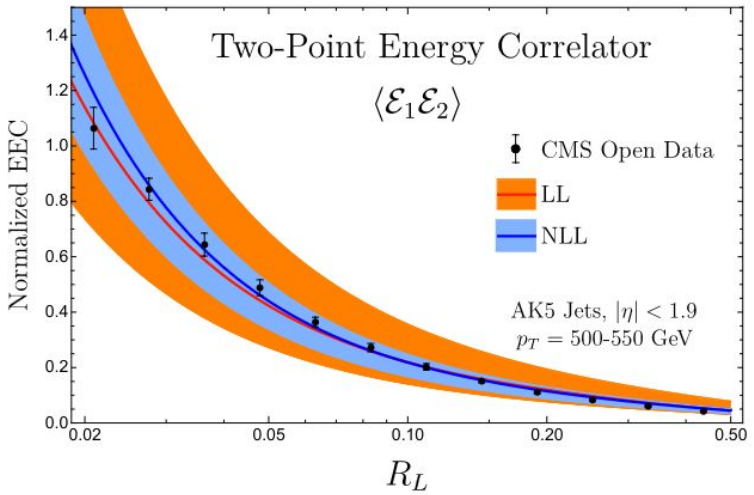
[Lee, Meçaj, Moult, 2205.03414]

Example 1: Energy-Energy Correlators



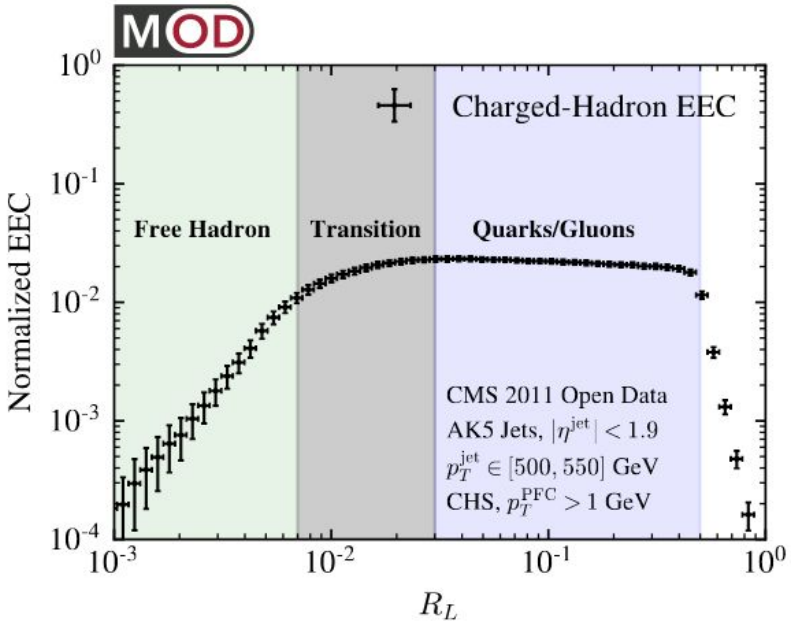


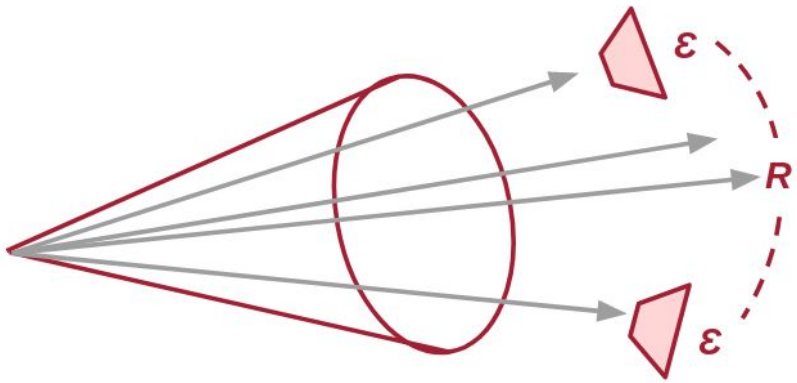
Possible for phenomenologists to compare calculations to data *directly*!



[Komiske, Moult, Thaler, Zhu, 2205.04459]

Example 1: Energy-Energy Correlators



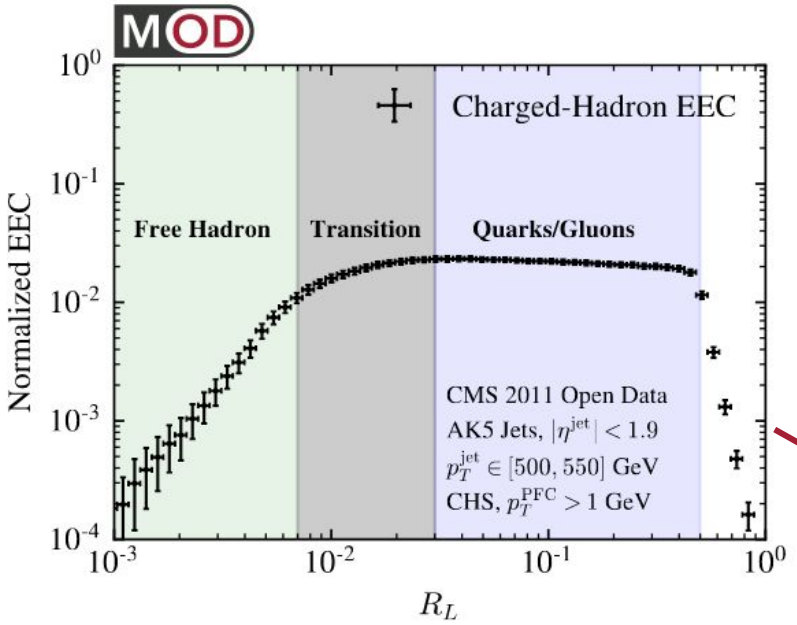


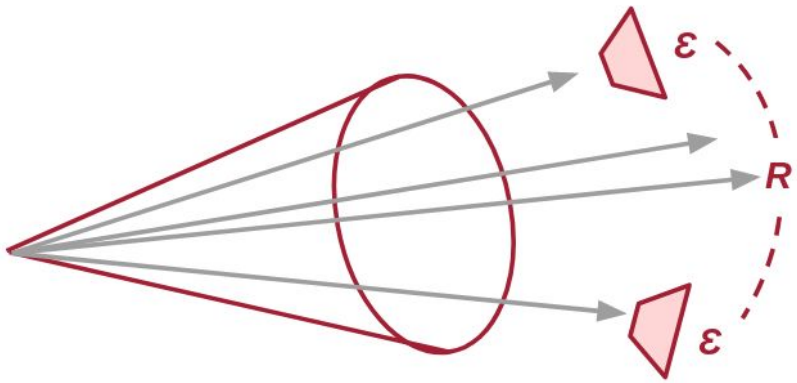
Different length scales probe different regimes of QCD!

Can see it all within Open Data!

[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!

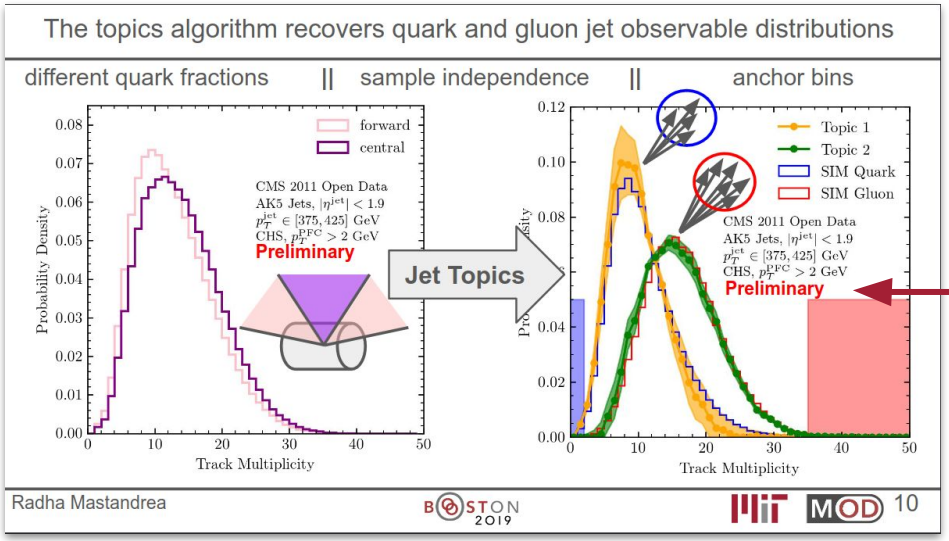
A similar measurement is now being done by CMS for Run II Data (see APS April talk).

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



[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:

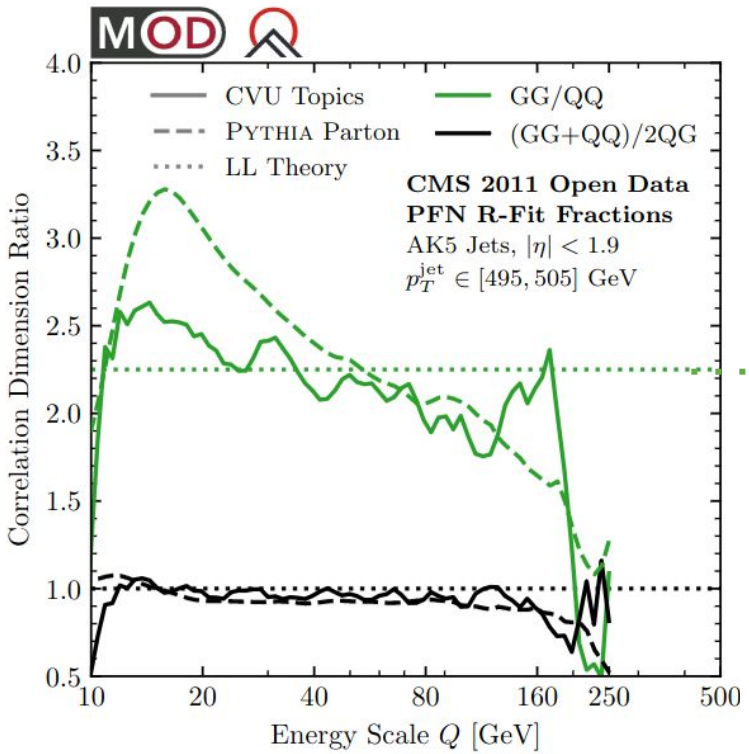
$$p_{\text{mixed}}(\vec{x}) = f_q\, p_{\text{quark}}(\vec{x}) + (1 - f_q)\, p_{\text{gluon}}(\vec{x})$$

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

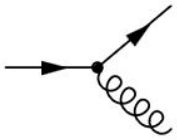
Can turn these quark/gluon distributions into measurements of **fundamental constants** of QCD in CMS Open Data!



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

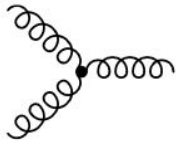
$$\frac{9}{4} = \frac{C_A}{C_F}$$

Quarks



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

Gluons

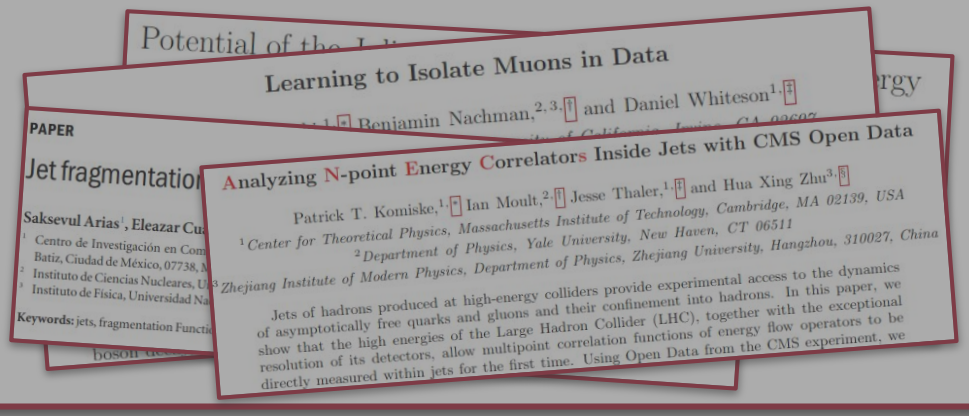


$$C_A = 3$$

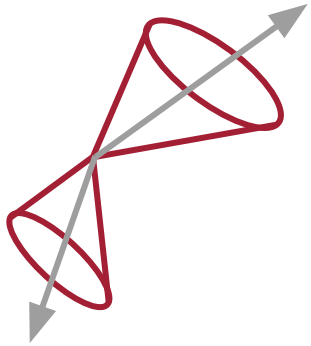
Correlation dimensions are defined using Wasserstein geometry, ask me about it later!

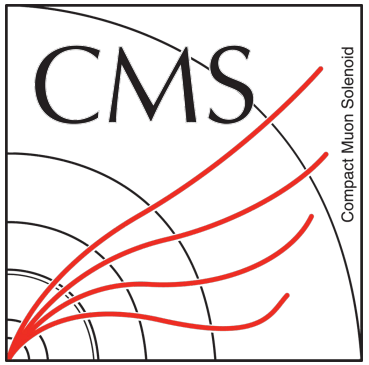
This Talk

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



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





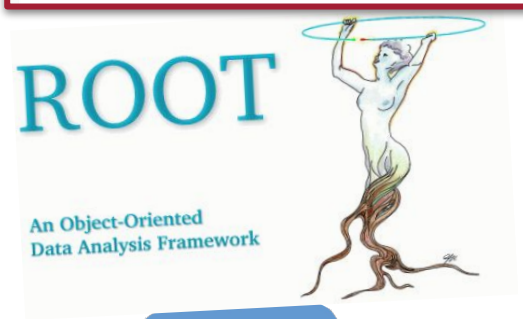
Open Data as a **teaching tool**

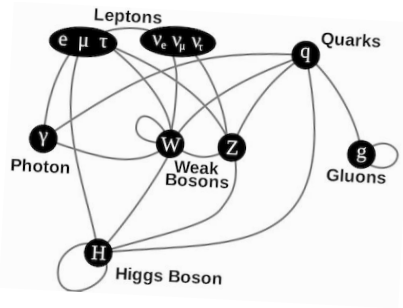
Higgs-to-four-lepton analysis example using 2011-2012 data

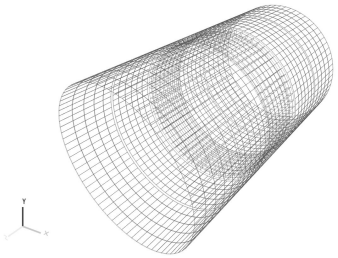
Jomhari, Nur Zulaiha ; Geiser, Achim ; Bin Anuar, Afiq Aizuddin

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

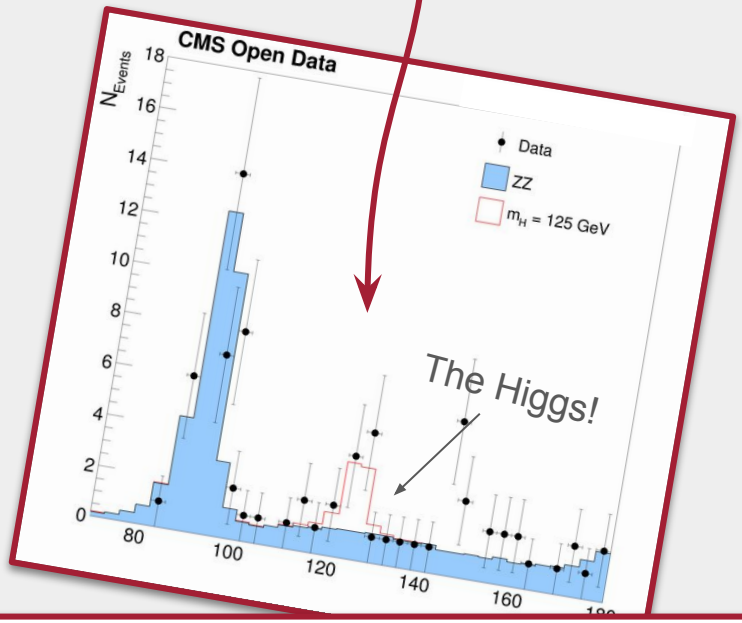
Software | Analysis | Workflow | CMS | CERN-LHC







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



One of my ever first plots!

My favorite dataset: CMS2011AJets

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!

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

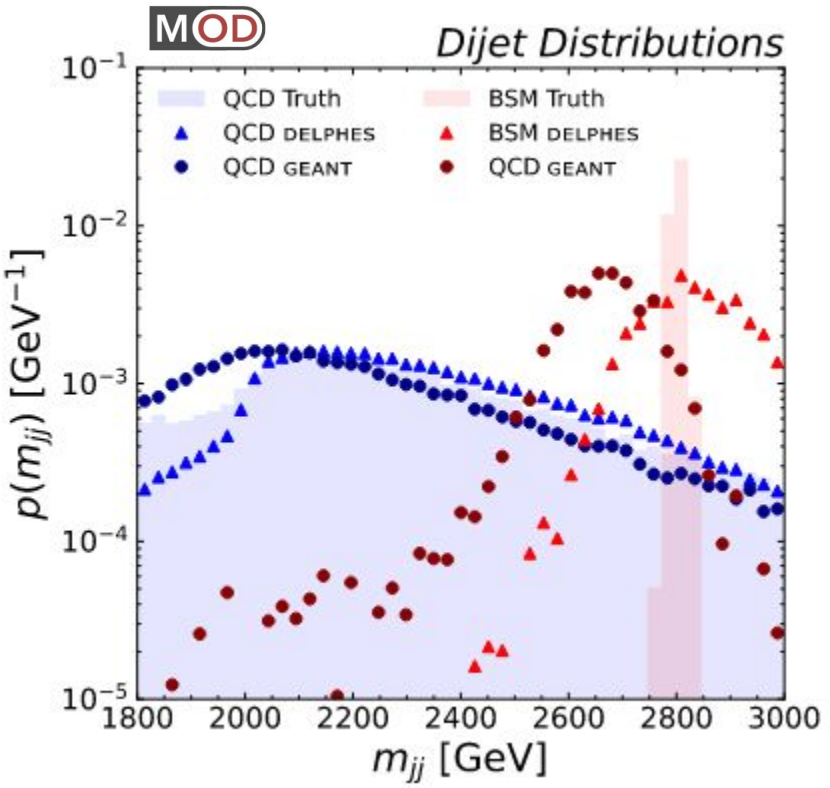
Jet primary dataset in AOD format from RunA of 2011 (/Jet/Run2011A-12Oct2013-v1/AOD)

/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

Dataset | Collision | CMS | 7TeV | CERN-LHC

opendata CERN



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

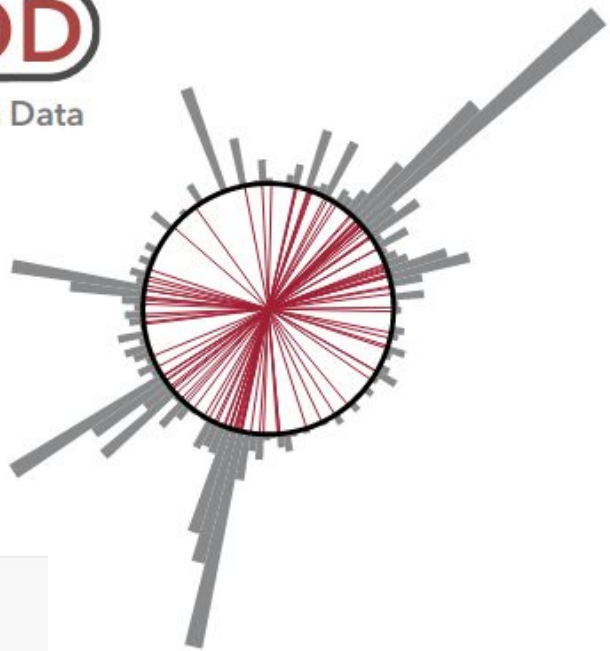
opendata CERN

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`:

MOD
MIT Open Data



```
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')
```

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

opendata CERN

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`:

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

Easy Data → Easy Analysis!

```
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')
```

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]

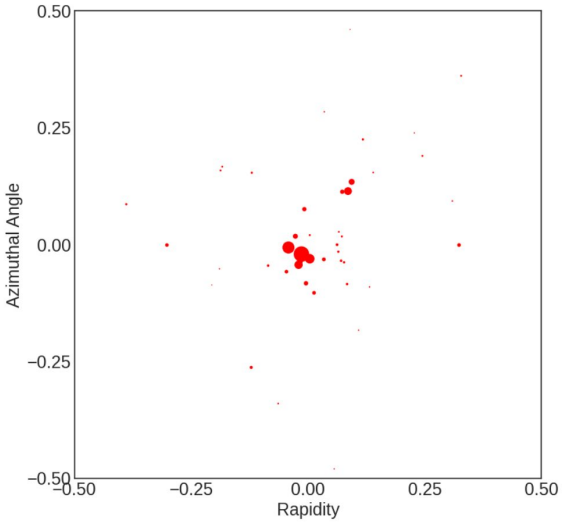
Try *pip install pyshaper*

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.

```python
# 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": P
```

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!

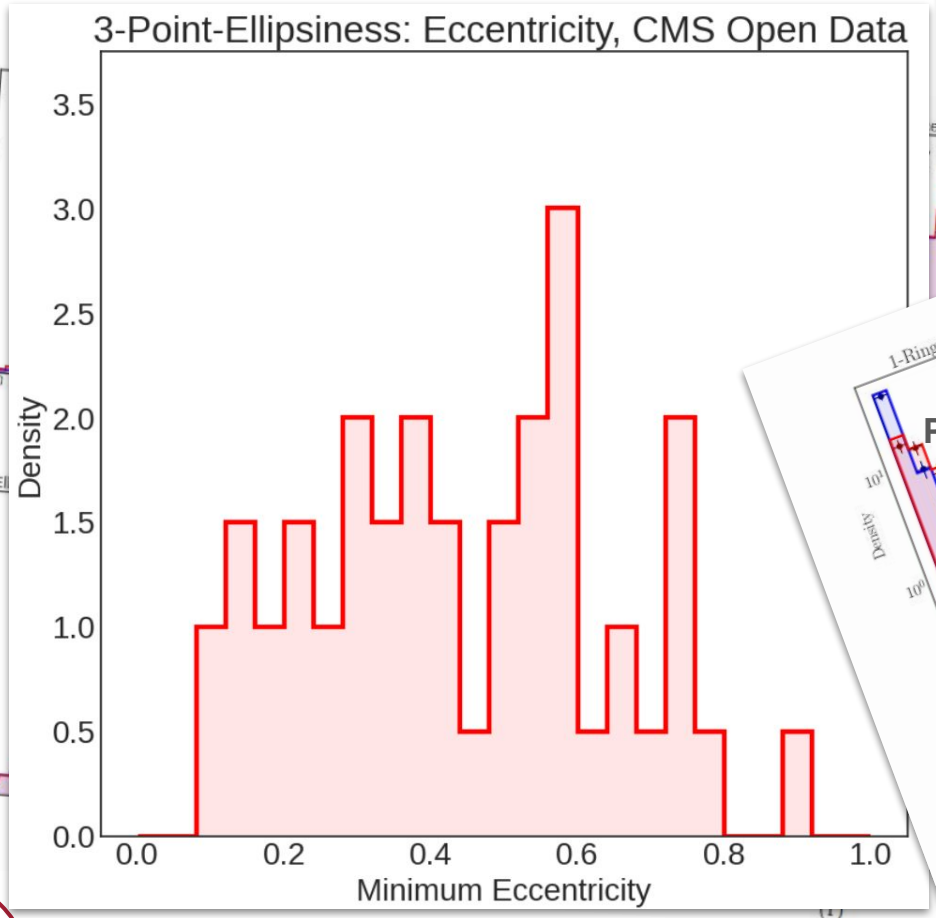
```
plot_dictionary = {
    "plot_directory" : "Plots/Test",
    "gif_directory" : "Plots/Test/gifs",
    "extension" : "png",
    "title" : "CMS Jets"
}

# 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)
```

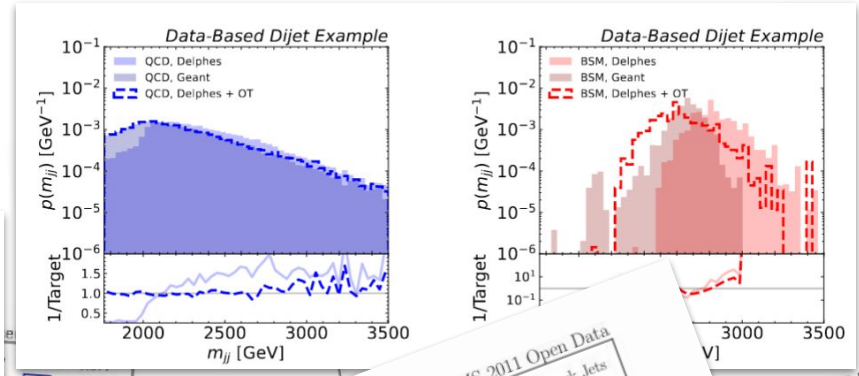
(Often done on a big cluster rather than a Jupyter notebook …)

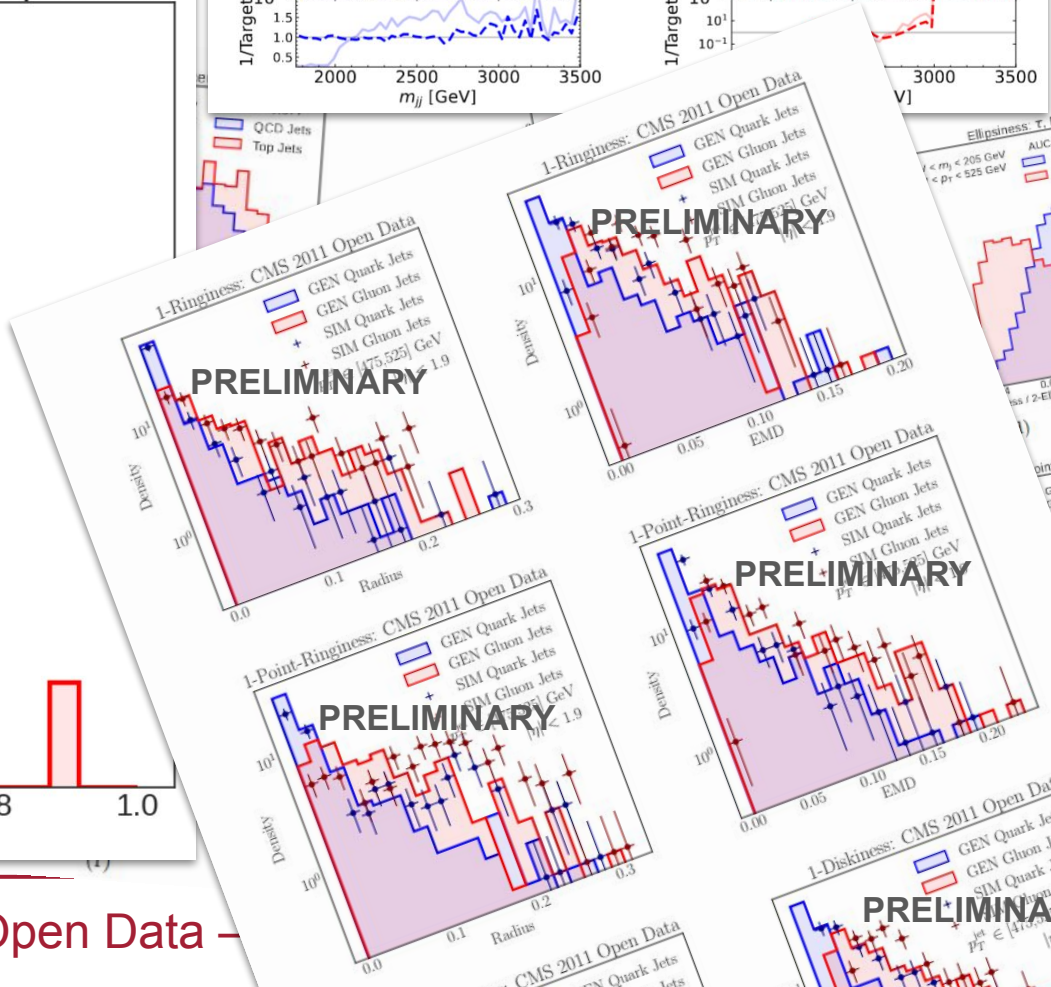
This code follows the example analysis I made at
https://github.com/rikab/SHAPER/blob/main/examples/example.ipynb
[Ba, Dogra, **RG**, Tasissa, Thaler, 2302.12266]
[**RG**, Thaler, Wu, WIP]

My typical **workflow**:

Step 4: Plots Plots Plots Plots Plots!







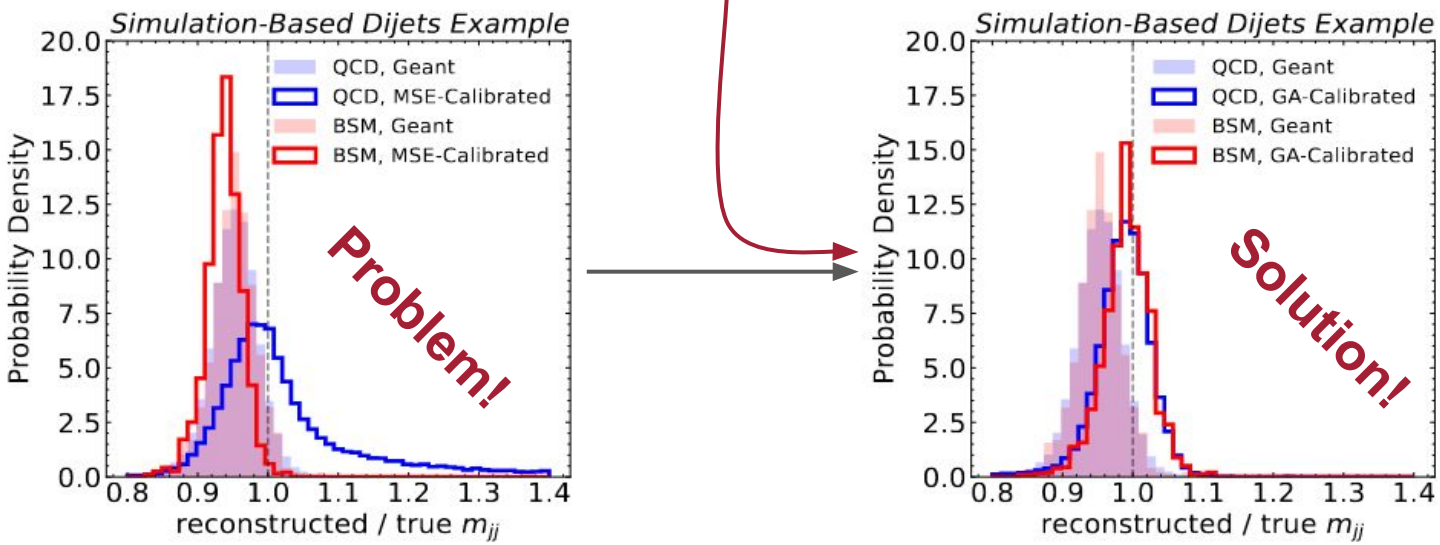
[**RG**, Nachman, Thaler, 2205.05084]

Try *pip install GaussianAnsatz*

CMS Open Sim for **Calibration**

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

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



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

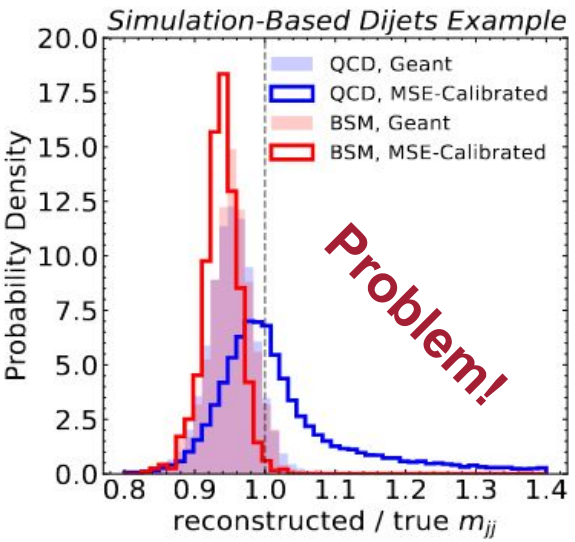


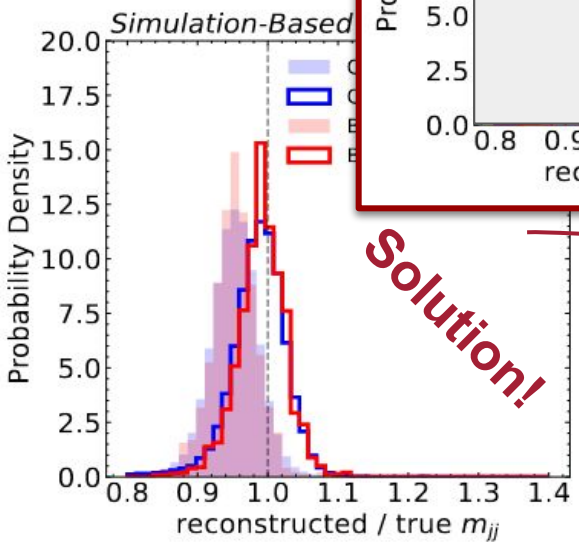
[**RG**, Nachman, Thaler, 2205.05084]
[**RG**, Nachman, Thaler, 2205.05084]
[**RG**, Nachman, Thaler, WIP]
Try `pip install GaussianAnsatz`

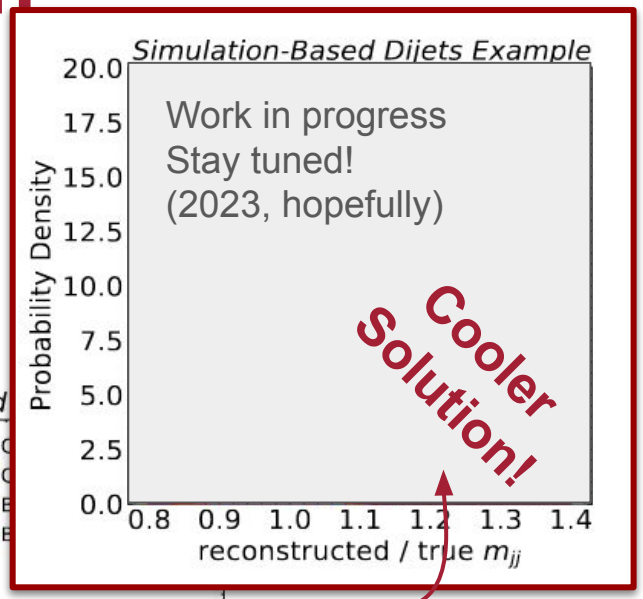
CMS Open Sim for **Calibration**

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

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





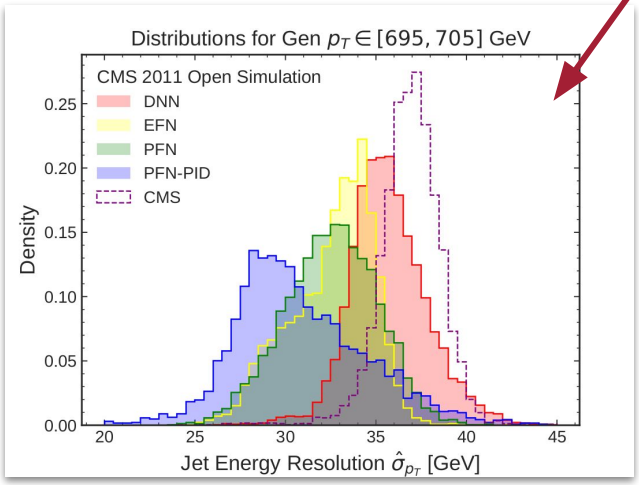


… Using Open Data as an easy, realistic example dataset for **ML studies** and **calibration**!

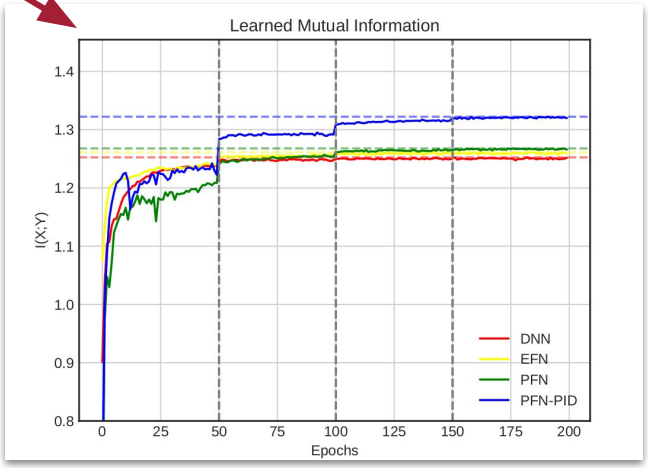
[RG, Nachman, Thaler, 2205.03413]

Try *pip install GaussianAnsatz*

CMS Open Sim for Uncertainty Estimation



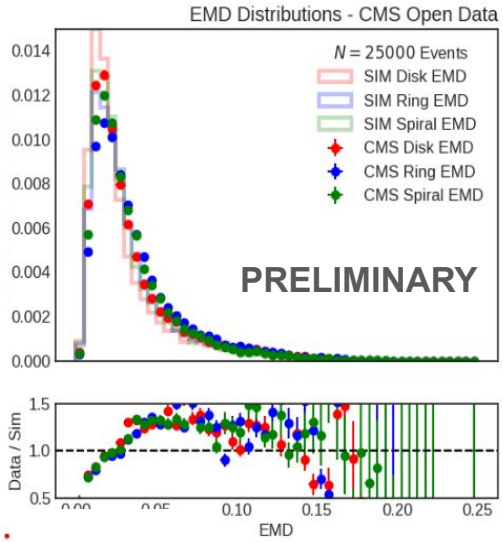
Model	Avg Resolution [GeV]	Mutual Information *I(X;Z)*
DNN	35.7 ± 2.1	1.23
EFN	32.6 ± 2.3	1.26
PFN	32.5 ± 2.5	1.27
PFN-PID	**30.8 ± 3.6**	**1.32**
CMS Open Data	36.9 ± 1.7	—

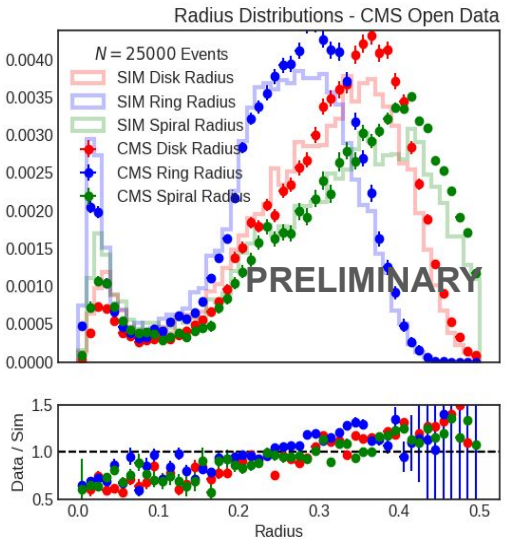


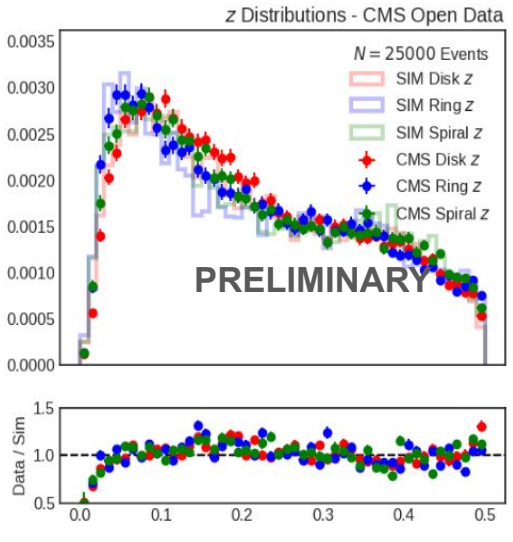
… Using Open Data to understand **detector efforts** and quantify **uncertainties and correlations** with **ML**!

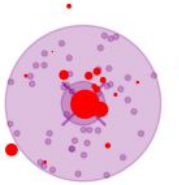
[Ba, Dogra, **RG**, Tasissa, Thaler, 2302.12266]

Hearing the Shapes of Jets

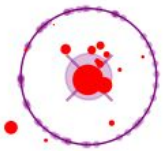




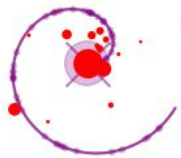




Disk + δ-function



Ring + δ-function

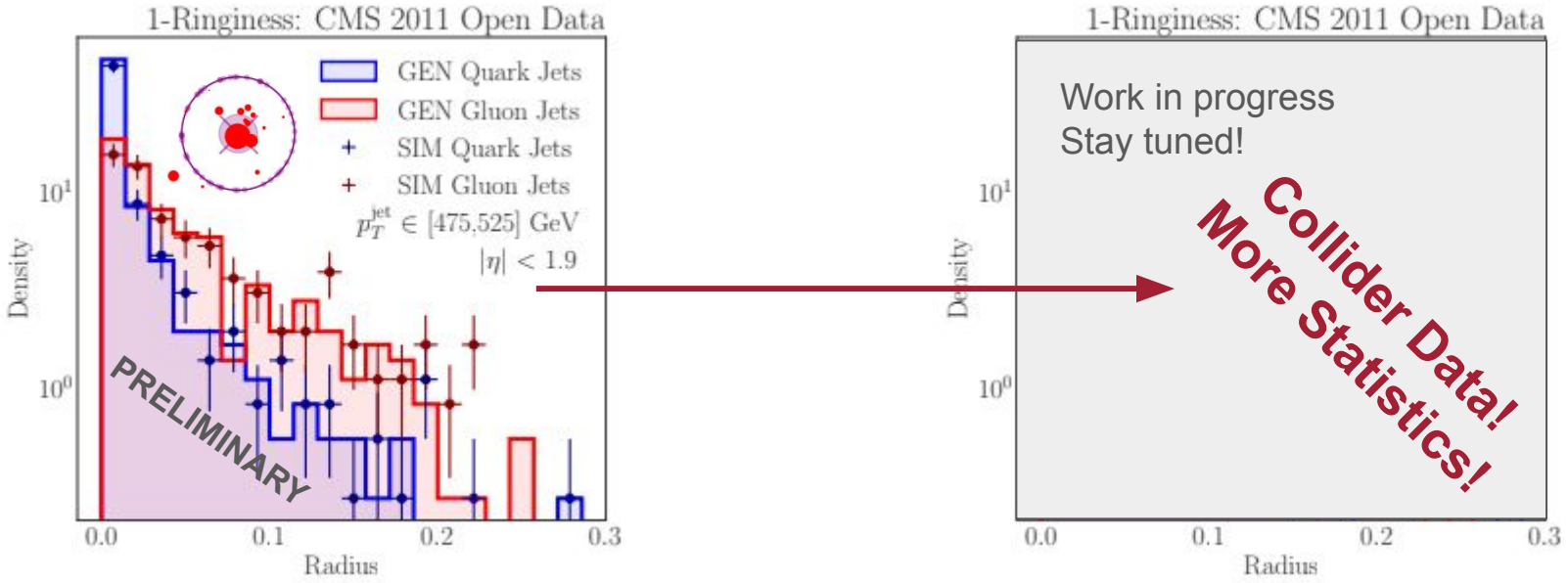


Spiral + δ-function

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]

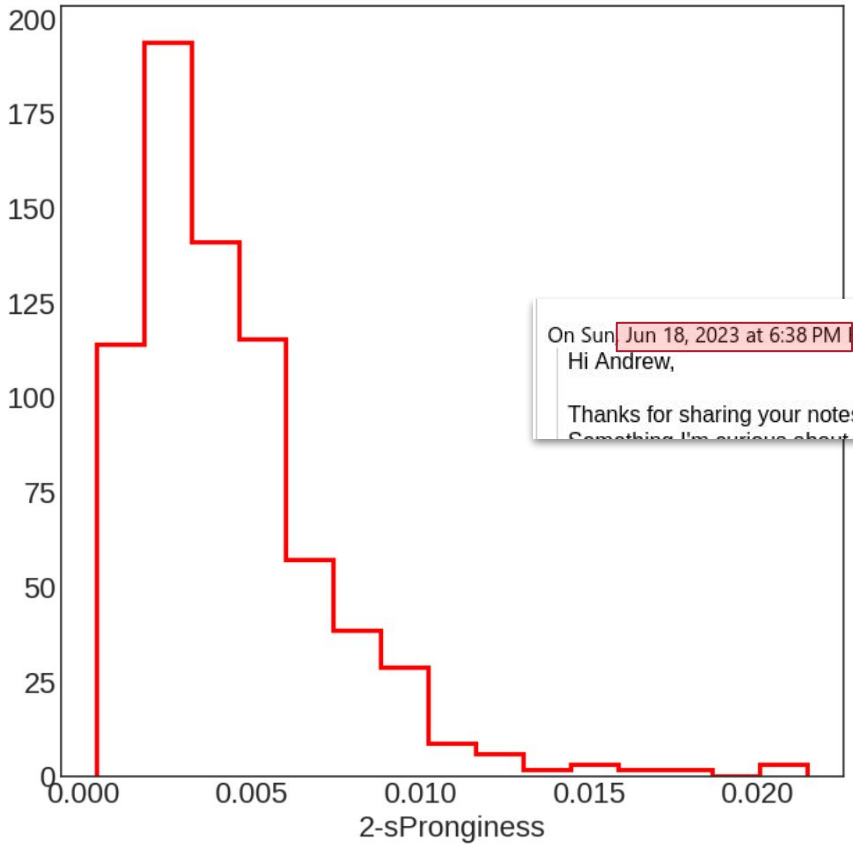
How wide are QCD jets?



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

preliminary spectral density test

main

Rikab Gambhir committed 2 weeks ago

Jun 20, 2023, 3:41 PM EDT

Showing **2 changed files** with **450** additions and **1** deletion.

How do I take my Open Data?

Conclusion

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

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

Admittedly, the last few are a stretch