The Future is Open

Jet Substructure with CMS Public Data

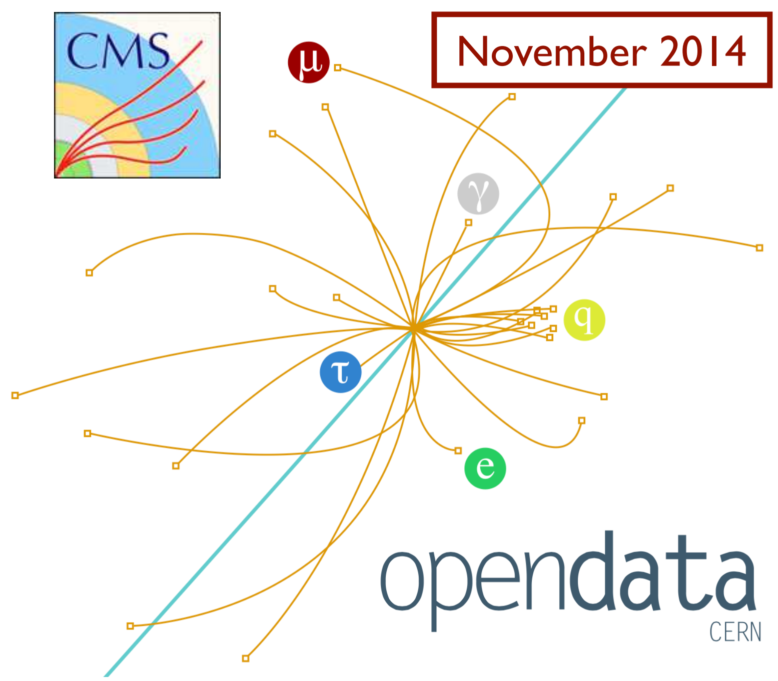
Jesse Thaler

MIT

LHC and the Standard Model: Physics and Tools, CERN TH — June 14, 2017

Last time in 4-3.006…

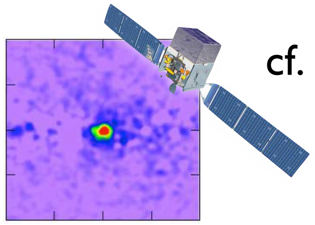
The Future is Open



CMS 2010:

Unique data set
with very low pileup

Accelerating science
through public data

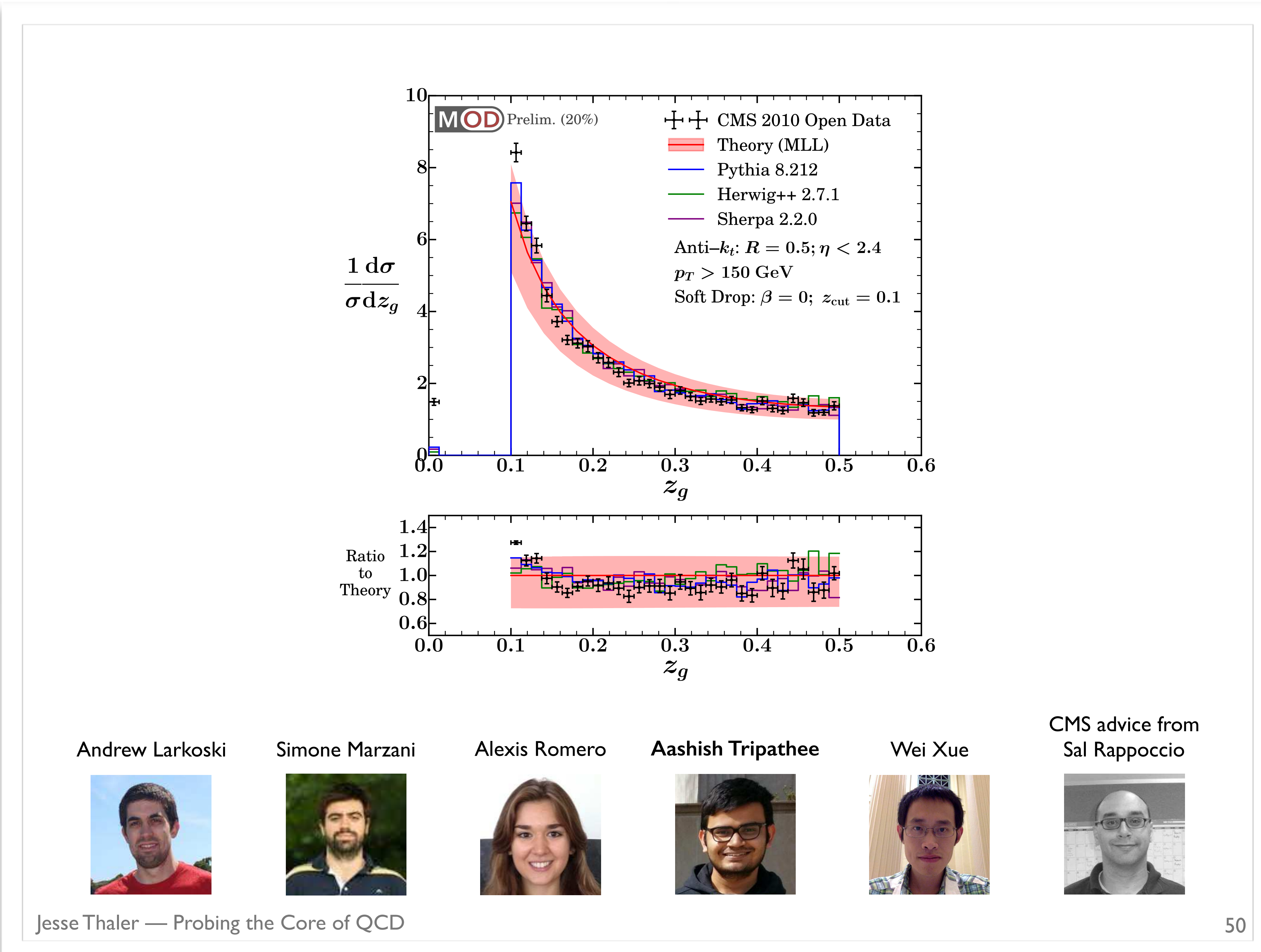


cf. Fermi

Jesse Thaler — Probing the Core of QCD 48

[March 2016]

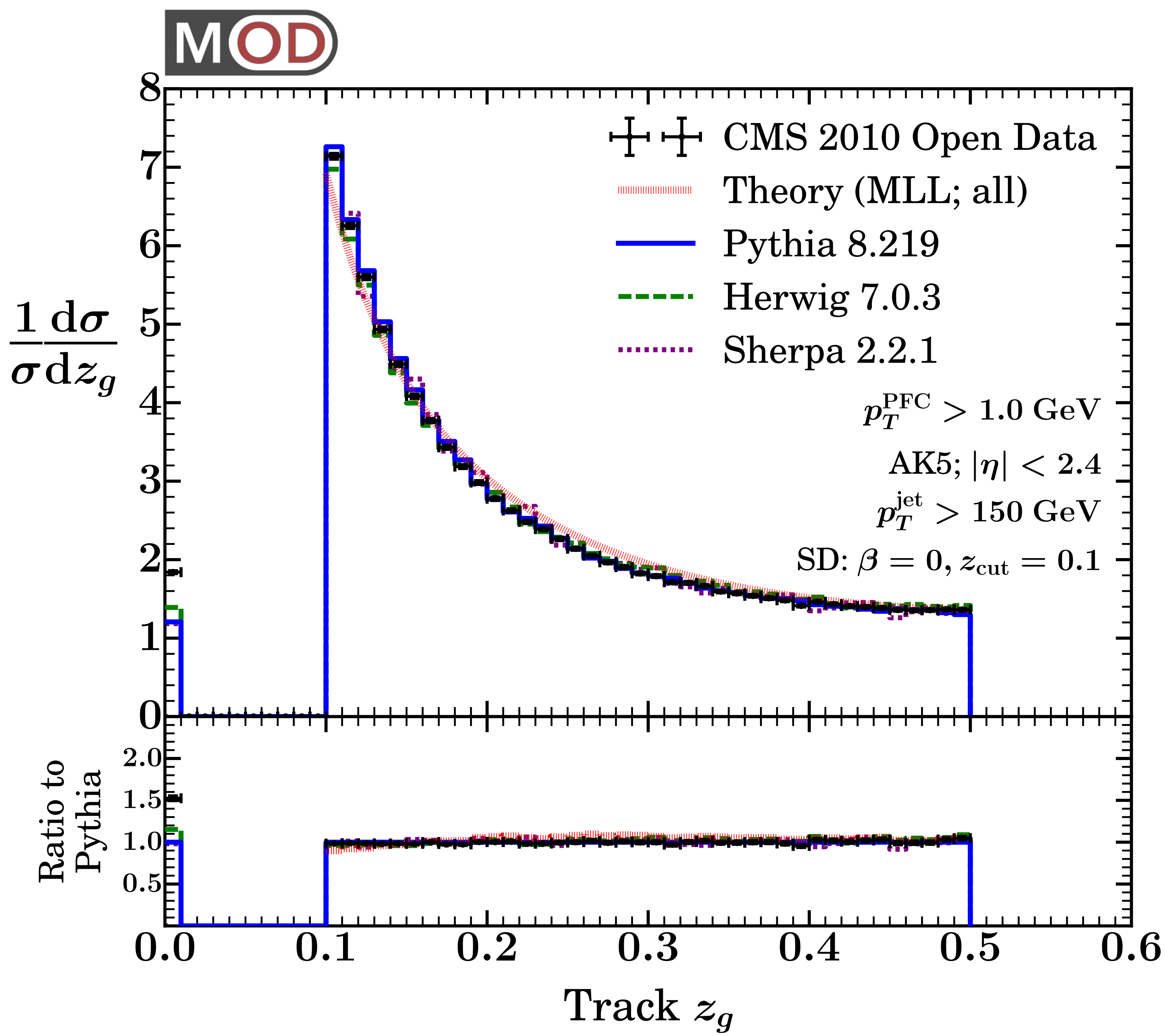
Last time in 4-3.006…



[March 2016]

First Analysis using CMS Open Data

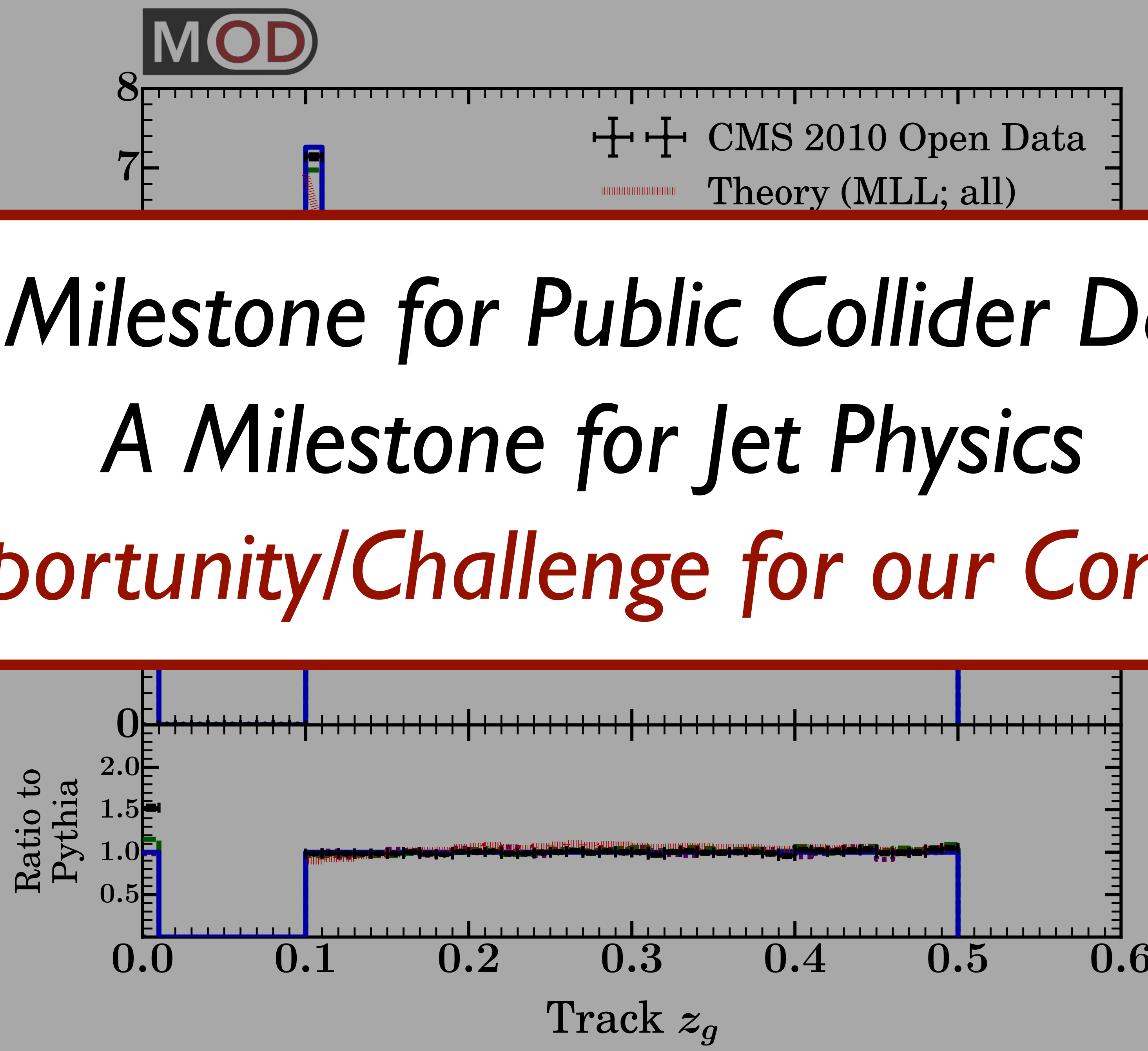
First Measurement* of Groomed Momentum Fraction*



[Larkoski, Marzani, JDT, Tripathee, Xue, 1704.05066, 1704.05842]

First Analysis using CMS Open Data

First Measurement* of Groomed Momentum Fraction*



A Milestone for Public Collider Data

A Milestone for Jet Physics

An Opportunity/Challenge for our Community

[Larkoski, Marzani, JDT, Tripathee, Xue, 1704.05066, 1704.05842]

Jet Substructure Studies with CMS Open Data

Aashish Tripathee,[1,*] Wei Xue,[1,†] Andrew Larkoski,[2,‡] Simone Marzani,[3,§] and Jesse Thaler[1,¶]

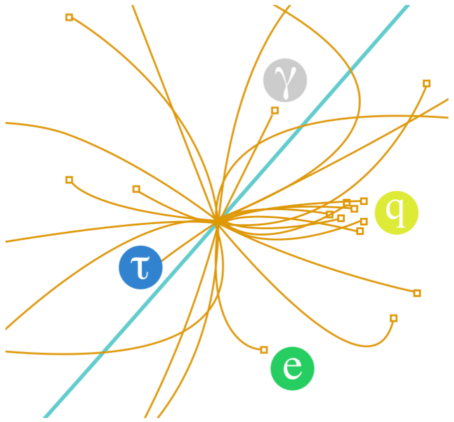
[1] *Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[2] *Physics Department, Reed College, Portland, OR 97202, USA*
[3] *University at Buffalo, The State University of New York, Buffalo, NY 14260-1500, USA*

VI. CONCLUSION

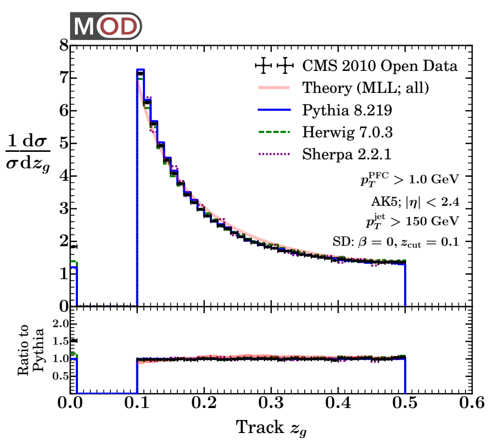
As the LHC explores the frontiers of scientific knowledge, its primary legacy will be the measurements and discoveries made by the LHC detector collaborations. But there is another potential legacy from the LHC that could be just as important: granting future generations of physicists access to unique high-quality data sets from proton-proton collisions at 7, 8, 13, and 14 TeV.

[Tripathee, Xue, Larkoski, Marzani, JDT, 1704.05842]

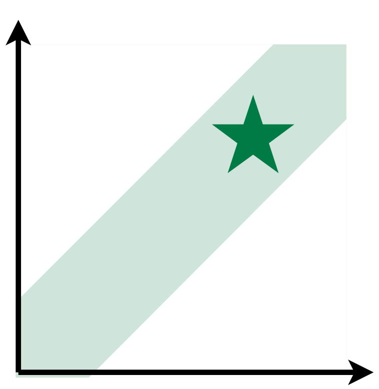
Outline



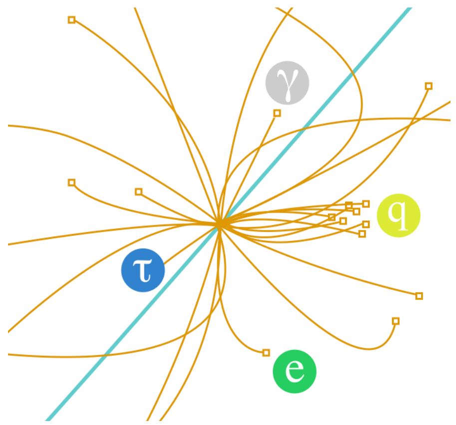
Introducing the CMS Open Data



Jet Substructure and QCD Splittings



(The Future of Public Collider Data)

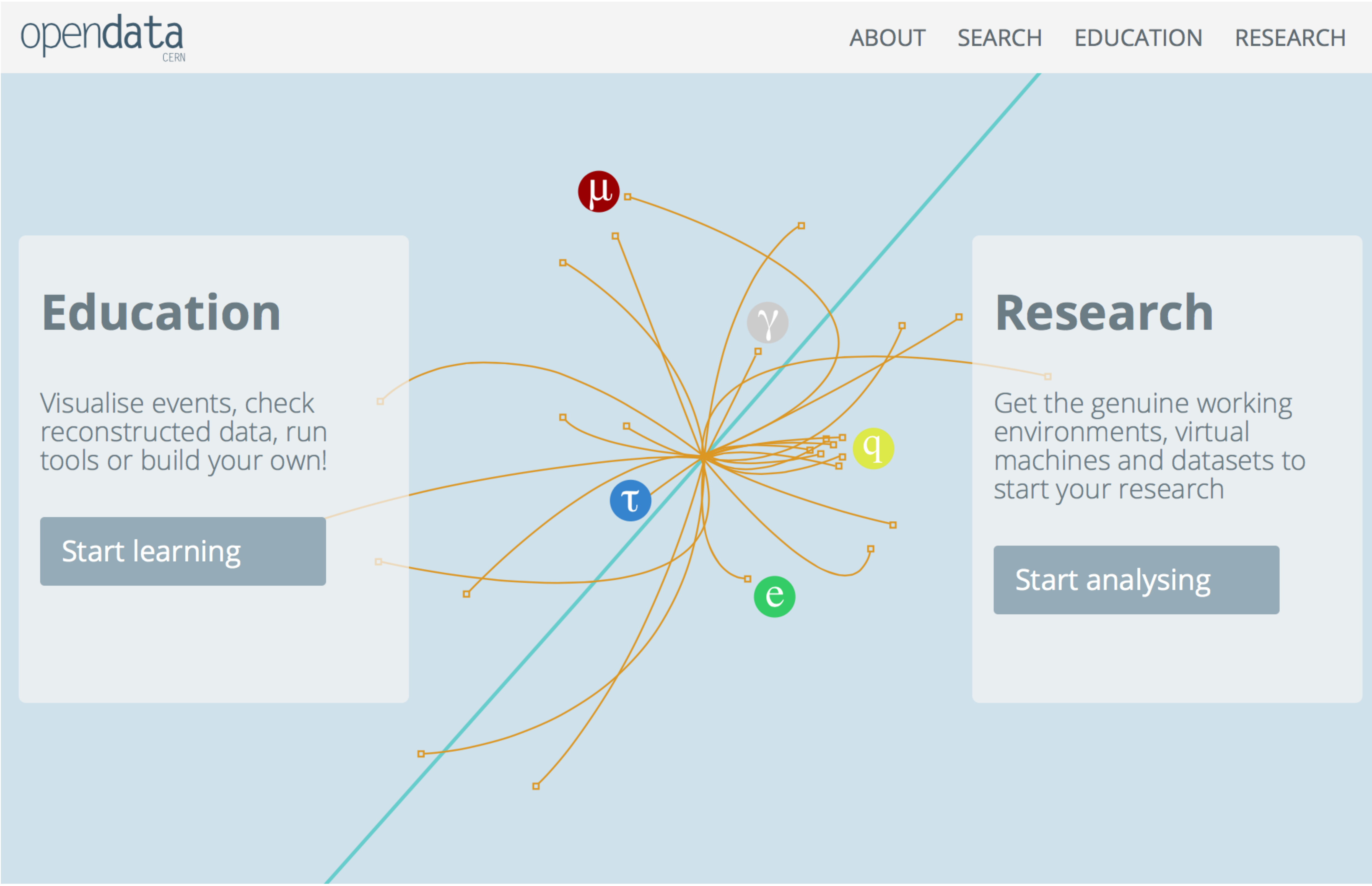


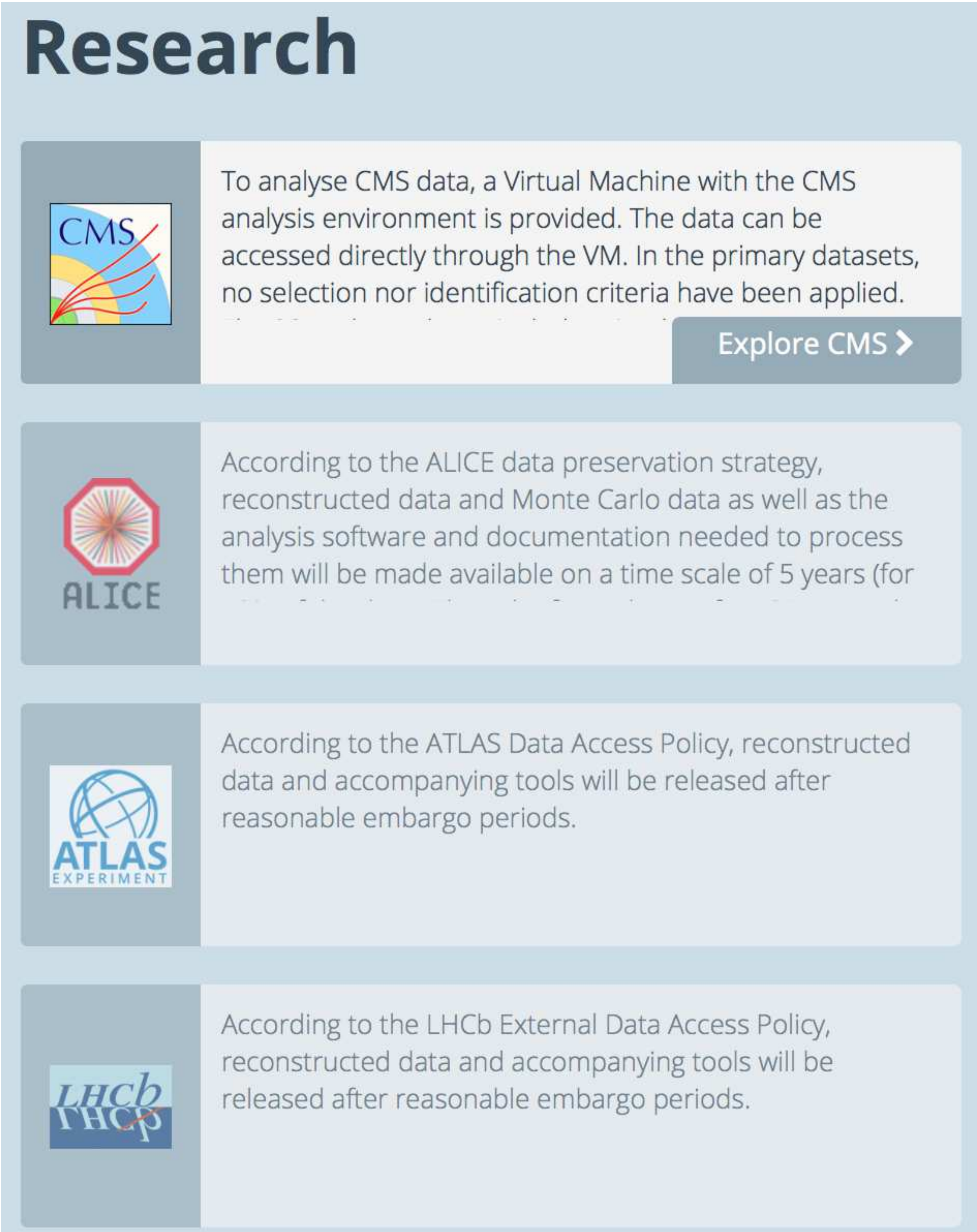
Introducing the CMS Open Data

Jet Substructure and QCD Splittings

(The Future of Public Collider Data)

http://opendata.cern.ch/





November 2014:

Run 2010B
7 TeV, 32 pb^{-1}

>20 TB, no MC

(Today's Talk)

April 2016:

Run 2011A
7 TeV, 2.5 fb^{-1}

>100 TB, with MC

Translating to "MIT Open Data"

Jet Primary Dataset

CernVM + CMSSW 4.2.8

AOD Format (CMS Root)

RAW → RECO → "Analysis Object Data"

Access via XRootD

2.0 TB

20,022,826 events
1664 files

MODAnalyzer + FastJet 3.1.3

MOD Format (ASCII + gzip)

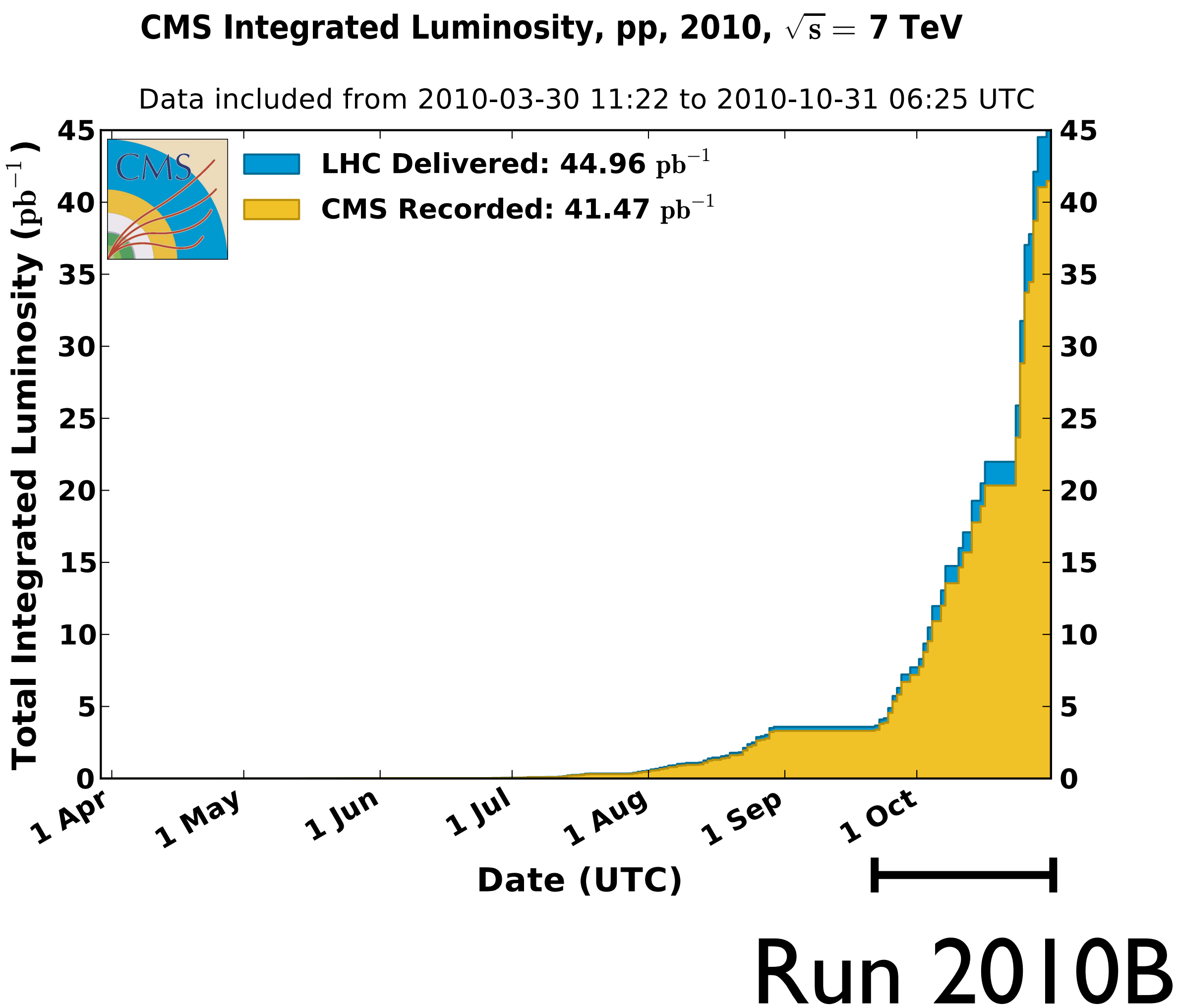
Cross-check with flat Root n-tuples

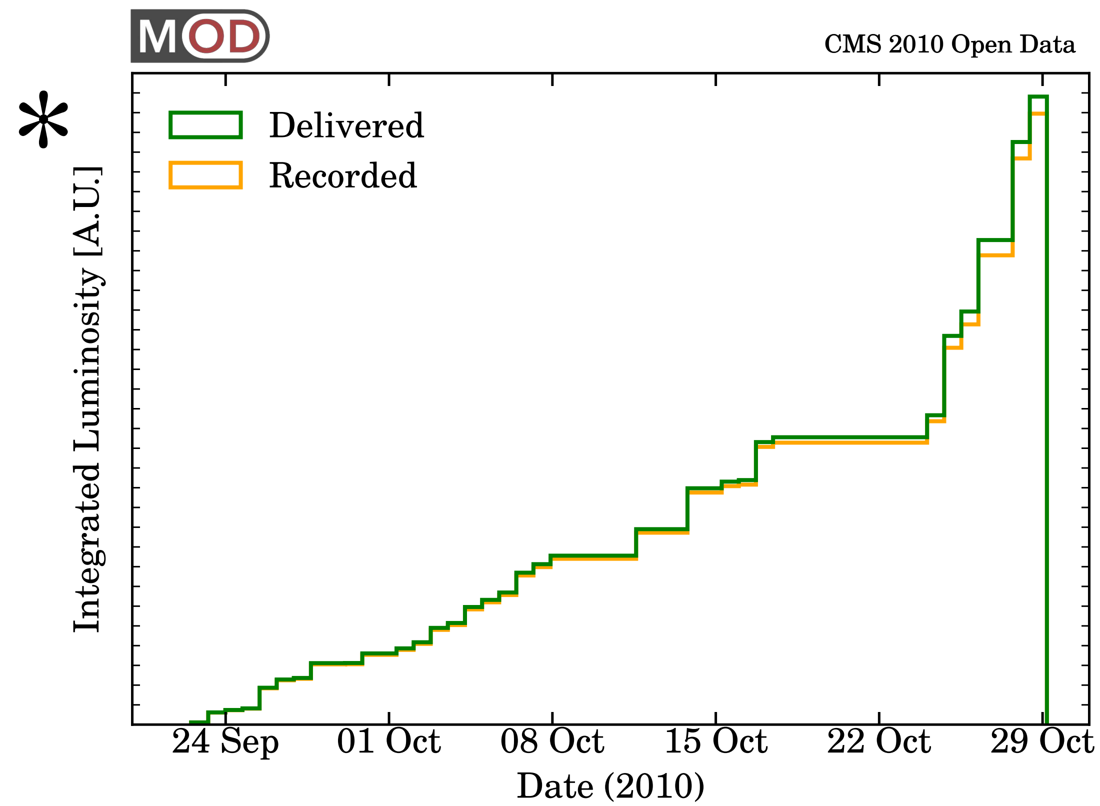
Access via External Hard Drive

200 GB

20 GB after
baseline selection

Integrated Luminosity





* AOD luminosity information is qualitative, not official

Demonstrates value of stress-testing archival strategies while collaboration is active

MOD Format

```
BeginEvent  Version  5  CMS_2010  Jet_Primary_Dataset
 # Cond     RunNum  EventNum  LumiBlock  validLumi  intgDelLumi  intgRecLumi  AvgInstLumi     NPV   timestamp    msOffset
   Cond     147926  188160899       201          1     21496.19     21208.58        92.03       4  1287023343    516890
 #  Trig                    Name  Prescale_1  Prescale_2     Fired?
    Trig  HLT_DiJetAve100U_v1              1           1          0
    Trig      HLT_DiJetAve15U            500          10          0
    Trig      HLT_DiJetAve30U              1         500          0
    Trig      HLT_DiJetAve50U              1          65          0
    Trig   HLT_DiJetAve70U_v2              1          25          0
    Trig        HLT_Jet100U_v2             1           1          1
 #  AK5        px        py        pz    energy        jec      area  no_of_const  chrg_multip  neu_had_frac  neu_em_frac  chrg_had_frac  chrg_em_frac
    AK5      9.31     -3.42     27.29     29.21       1.03      0.82            8            4          0.35         0.18           0.46          0.00
    AK5      6.77      2.40     13.35     15.30       0.99      0.72            6            4          0.55         0.11           0.33          0.00
    AK5      7.08      0.93    -61.18     61.62       0.93      0.82            2            0          1.00         0.00           0.00          0.00
 #  PFC        px        py        pz    energy      pdgId
    PFC     -0.95     -0.05      0.65      1.16       -211
    PFC     -0.75     -0.24     -1.06      1.33       -211
    PFC      1.27     -1.27    -11.10     11.25        130
    PFC     -0.00     -0.59      0.50      0.79        211
    PFC     -0.41      0.54      0.59      0.91       -211
    PFC      1.55      0.57      5.99      6.22        211
    PFC      0.12     -0.52      1.36      1.47       -211
    PFC      0.76      0.36     -1.59      1.81        211
    PFC      0.43      0.78      2.04      2.23        211
    PFC      1.90     -0.09      5.88      6.19        130
    PFC      0.71      1.71      0.94      2.08        211
EndEvent
```

See backup slides for more technical details

MOD Format

```
BeginEvent  Version  5  CMS_2010  Jet_Primary_Dataset
 # Cond     RunNum   EventNum  LumiBlock  validLumi  intgDelLumi  intgRecLumi  AvgInstLumi   NPV   timestamp    msOffset
   Cond     147926  188160899        201          1     21496.19     21208.58        92.03     4  1287023343      516890
 #  Trig                 Name  Prescale_1  Prescale_2     Fired?
    Trig HLT_DiJetAve100U_v1           1           1          0
    Trig     HLT_DiJetAve15U         500          10          0
    Trig     HLT_DiJetAve30U           1         500          0
    Trig     HLT_DiJetAve50U           1          65          0
    Trig  HLT_DiJetAve70U_v2           1          25          0
    Trig        HLT_Jet100U_v2         1           1          1
 #  AK5       px       py        pz    energy      jec   area  no_of_const  chrg_multip  neu_had_frac  neu_em_frac  chrg_had_frac  chrg_em_frac
    AK5     9.31    -3.42     27.29     29.21     1.03   0.82            8            4          0.35         0.18           0.46          0.00
    AK5     6.77     2.40     13.35     15.30     0.99   0.72            6            4          0.55         0.11           0.33          0.00
    AK5     7.08     0.93    -61.18     61.62     0.93   0.82            2            0          1.00         0.00           0.00          0.00
 #  PFC       px       py        pz    energy    pdgId
    PFC    -0.95    -0.05      0.65      1.16     -211
    PFC    -0.75    -0.24     -1.06      1.33     -211
    PFC     1.27    -1.27    -11.10     11.25      130
    PFC    -0.00    -0.59      0.50      0.79      211
    PFC    -0.41     0.54      0.59      0.91     -211
    PFC     1.55     0.57      5.99      6.22      211
    PFC     0.12    -0.52      1.36      1.47     -211
    PFC     0.76     0.36     -1.59      1.81      211
    PFC     0.43     0.78      2.04      2.23      211
    PFC     1.90    -0.09      5.88      6.19      130
    PFC     0.71     1.71      0.94      2.08      211
EndEvent
```

Triggers

Jets: anti-k_t R = 0.5

Particle Flow Candidates

See backup slides for more technical details

MOD Format

```
BeginEvent  Version  5  CMS_2010  Jet_Primary_Dataset
 # Cond    RunNum   EventNum  LumiBlock  validLumi  intgDelLumi  intgRecLumi  AvgInstLumi   NPV   timestamp   msOffset
   Cond    147926  188160899        201          1     21496.19     21208.58        92.03     4  1287023343     516890
 #  Trig                Name  Prescale_1  Prescale_2     Fired?
    Trig  HLT_DiJetAve100U_v1          1           1          0
    Trig     HLT_DiJetAve15U         500          10          0
    Trig     HLT_DiJetAve30U           1         500          0
    Trig     HLT_DiJetAve50U           1          65          0
    Trig   HLT_DiJetAve70U_v2          1          25          0
    Trig        HLT_Jet100U_v2          1           1          1
 #  AK5       px        py        pz    energy     jec    area  no_of_const  chrg_multip  neu_had_frac  neu_em_frac  chrg_had_frac  chrg_em_frac
    AK5     9.31     -3.42     27.29     29.21    1.03    0.82            8            4          0.35         0.18           0.46          0.00
    AK5     6.77      2.40     13.35     15.30    0.99    0.72            6            4          0.55         0.11           0.33          0.00
    AK5     7.08      0.93    -61.18     61.62    0.93    0.82            2            0          1.00         0.00           0.00          0.00
 #  PFC       px        py        pz    energy   pdgId
    PFC    -0.95     -0.05      0.65      1.16    -211
    PFC    -0.75     -0.24     -1.06      1.33    -211
    PFC     1.27     -1.27    -11.10     11.25     130
    PFC    -0.00     -0.59      0.50      0.79     211
    PFC    -0.41      0.54      0.59      0.91    -211
    PFC     1.55      0.57      5.99      6.22     211
    PFC     0.12     -0.52      1.36      1.47    -211
    PFC     0.76      0.36     -1.59      1.81     211
    PFC     0.43      0.78      2.04      2.23     211
    PFC     1.90     -0.09      5.88      6.19     130
    PFC     0.71      1.71      0.94      2.08     211
EndEvent
```

Jet Quality Criteria

Jet Energy Corrections

See backup slides for more technical details

MOD Format

Low Pileup

Jet Area

```
BeginEvent  Version  5  CMS_2010  Jet_Primary_Dataset
 # Cond     RunNum  EventNum  LumiBlock  validLumi  intgDelLumi  intgRecLumi  AvgInstLumi      NPV   timestamp    msOffset
   Cond     147926  188160899       201          1     21496.19     21208.58        92.03        4  1287023343      516890
 #  Trig                 Name  Prescale_1  Prescale_2    Fired?
    Trig HLT_DiJetAve100U_v1           1           1         0
    Trig     HLT_DiJetAve15U         500          10         0
    Trig     HLT_DiJetAve30U           1         500         0
    Trig     HLT_DiJetAve50U           1          65         0
    Trig   HLT_DiJetAve70U_v2          1          25         0
    Trig       HLT_Jet100U_v2          1           1         1
 #  AK5        px        py        pz    energy       jec     area  no_of_const  chrg_multip  neu_had_frac  neu_em_frac  chrg_had_frac  chrg_em_frac
    AK5      9.31     -3.42     27.29     29.21      1.03     0.82            8            4          0.35         0.18           0.46          0.00
    AK5      6.77      2.40     13.35     15.30      0.99     0.72            6            4          0.55         0.11           0.33          0.00
    AK5      7.08      0.93    -61.18     61.62      0.93     0.82            2            0          1.00         0.00           0.00          0.00
 #  PFC        px        py        pz    energy     pdgId
    PFC     -0.95     -0.05      0.65      1.16      -211
    PFC     -0.75     -0.24     -1.06      1.33      -211
    PFC      1.27     -1.27    -11.10     11.25       130
    PFC     -0.00     -0.59      0.50      0.79       211
    PFC     -0.41      0.54      0.59      0.91      -211
    PFC      1.55      0.57      5.99      6.22       211
    PFC      0.12     -0.52      1.36      1.47      -211
    PFC      0.76      0.36     -1.59      1.81       211
    PFC      0.43      0.78      2.04      2.23       211
    PFC      1.90     -0.09      5.88      6.19       130
    PFC      0.71      1.71      0.94      2.08       211
EndEvent
```

See backup slides for more technical details

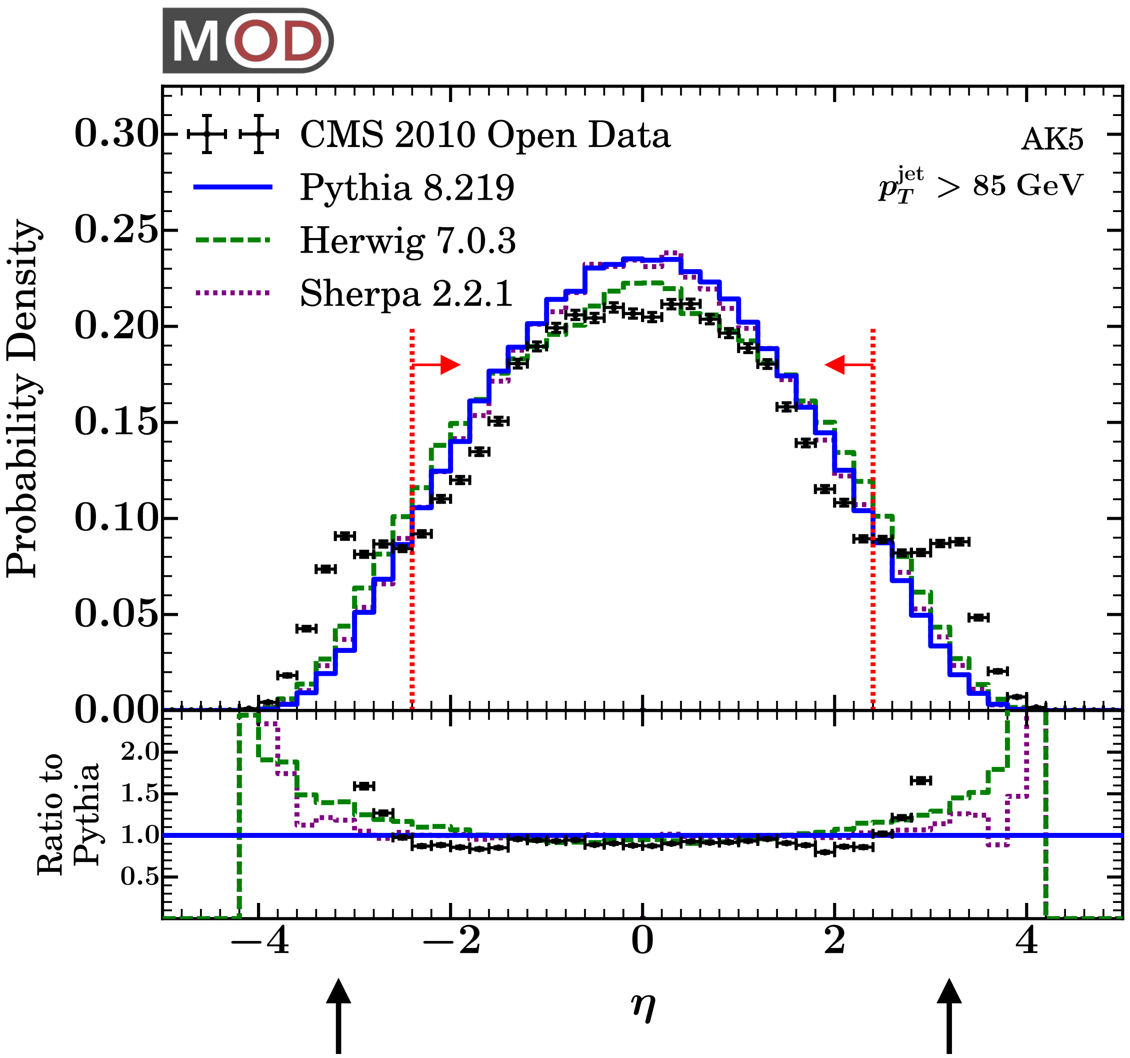
Warning: the following plots **cannot** be interpreted like standard experimental results

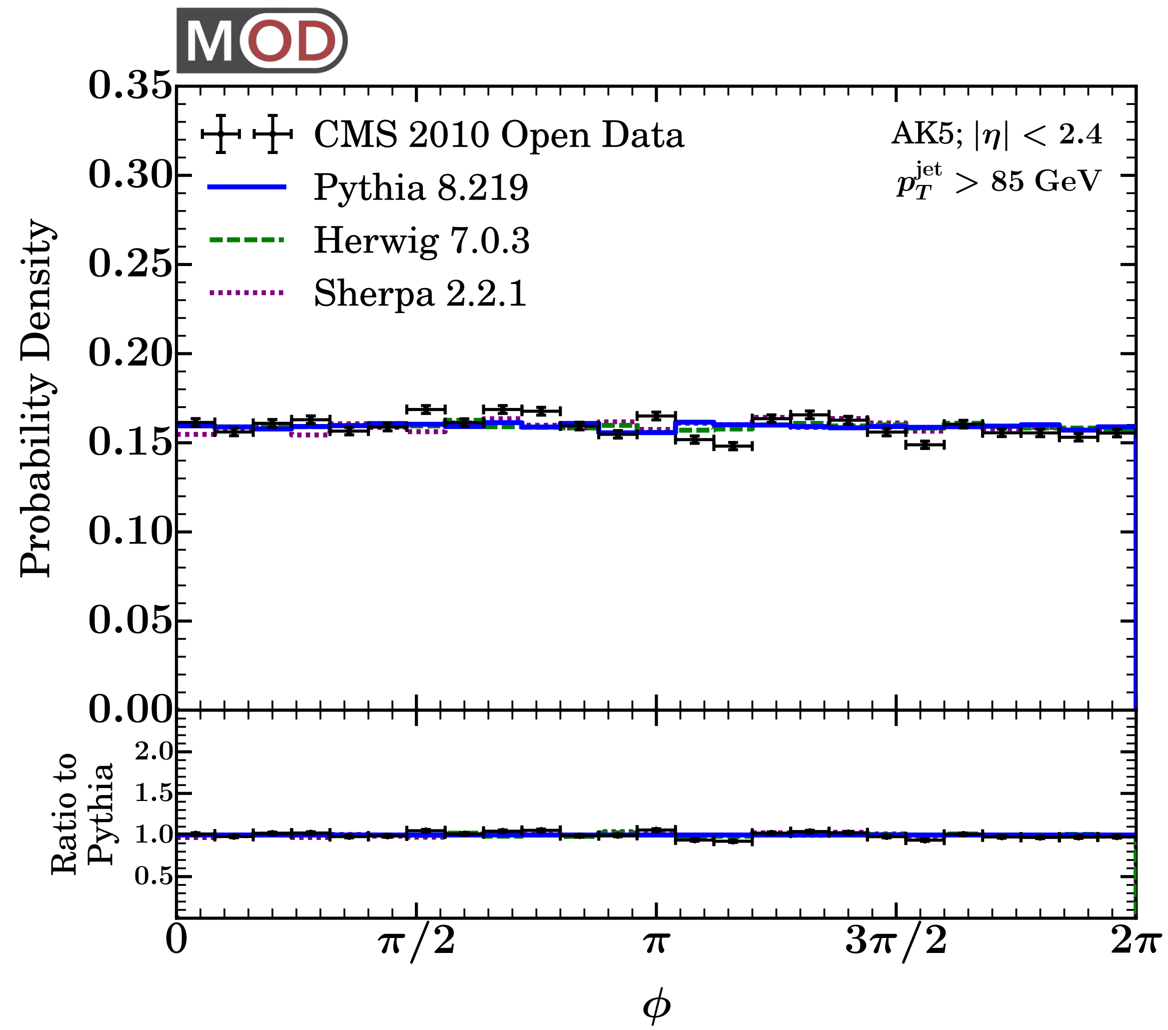
Run 2010B data does not include information for calibration/unfolding beyond JEC factors (Run 2011A does include MC)

We are **forced** to commit a cardinal sin:

Detector-object data (with statistical errors only)
overlaid on
Truth-hadron parton shower generators (no simulation)

Hardest Jet Kinematics

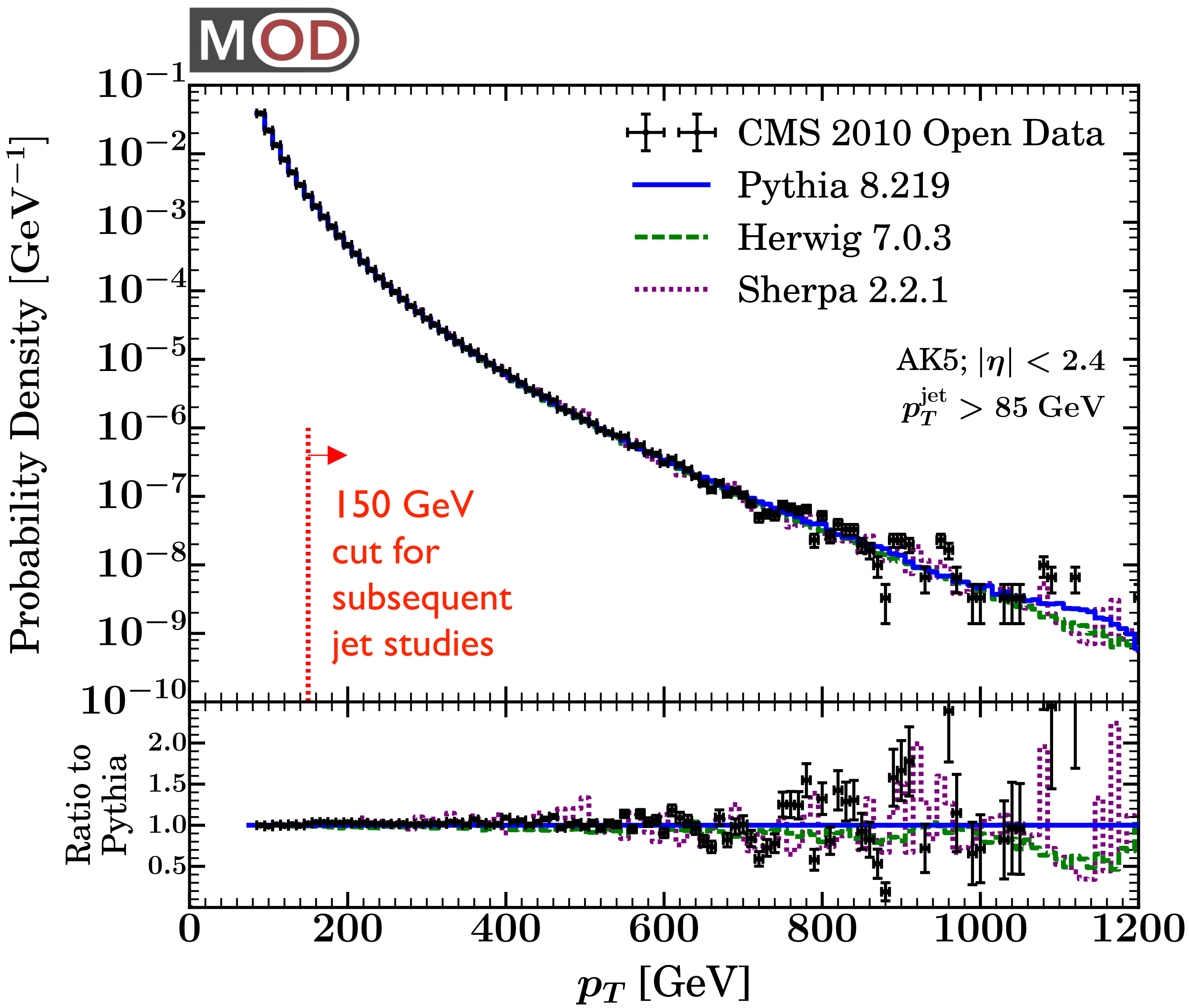




Bad-quality jets beyond tracking acceptance

Comparison to 3 parton shower generators
with default tuning parameters

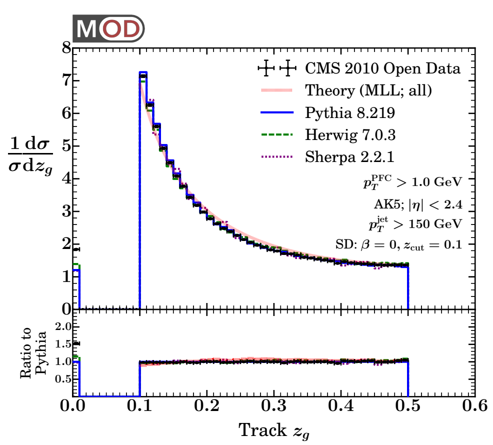
Hardest Jet p_T Spectrum



Largest (high-quality) jet p_T encountered: 1277 GeV

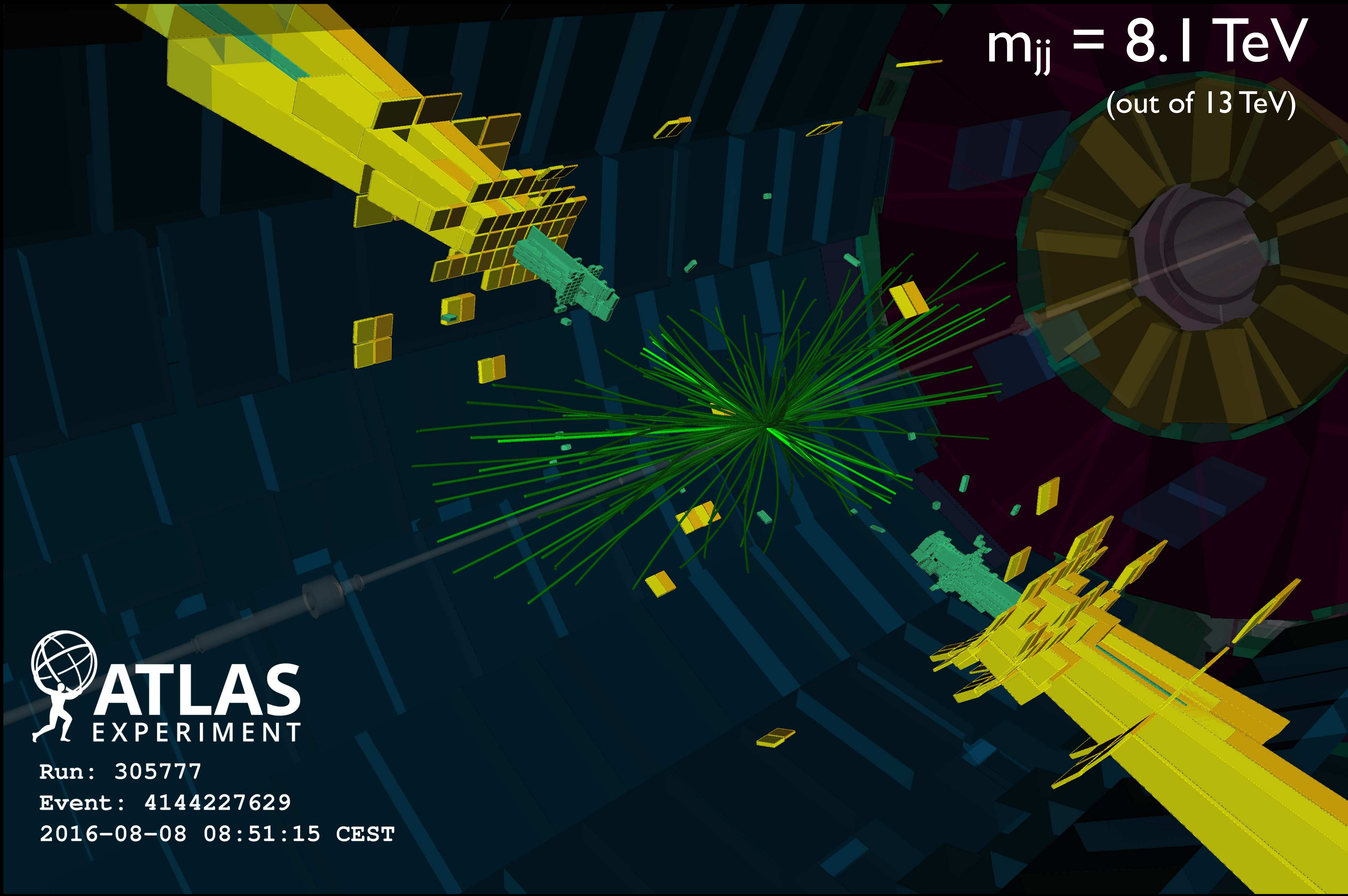
5 trigger merging took ≈2 years of debugging

Introducing the CMS Open Data

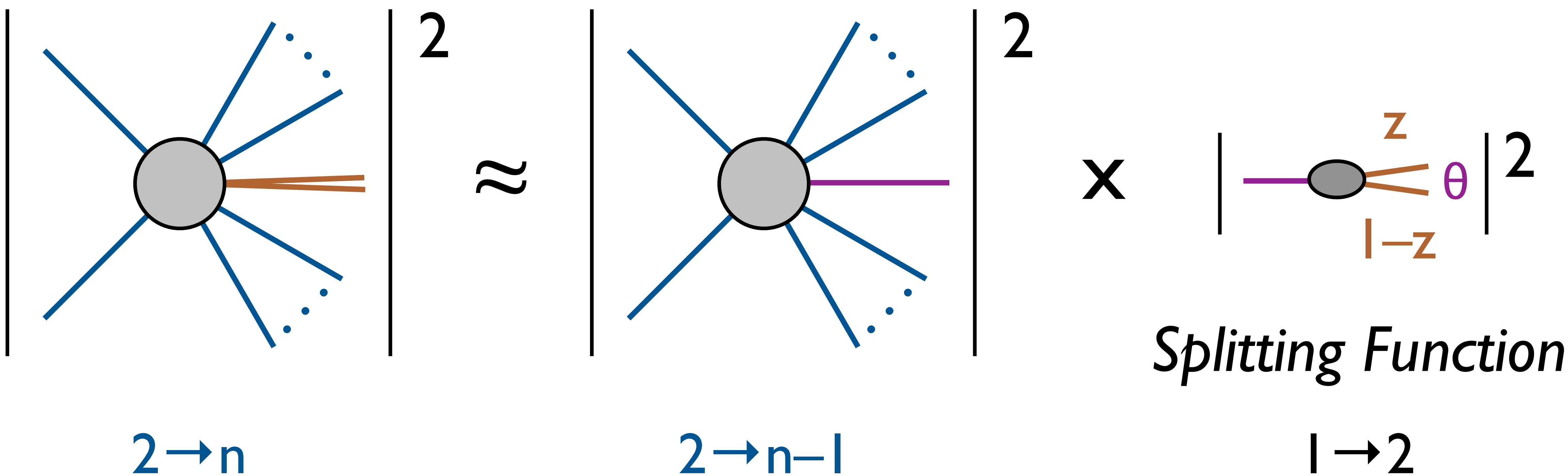


Jet Substructure and QCD Splittings

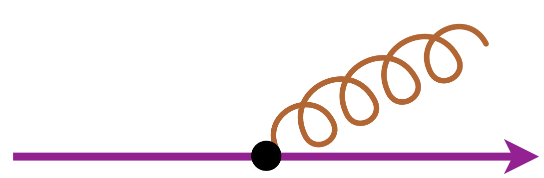
(The Future of Public Collider Data)


m_{jj} = 8.1 TeV
(out of 13 TeV)
ATLAS
EXPERIMENT
Run: 305777
Event: 4144227629
2016-08-08 08:51:15 CEST

Textbook QCD: Universal Collinear Limit



For this talk:



$C_q = 4/3$
$C_g = 3$

$$\mathrm{d}P_{i\to ig} \simeq \frac{2\alpha_s}{\pi} C_i \frac{\mathrm{d}\theta}{\theta} \frac{\mathrm{d}z}{z}$$

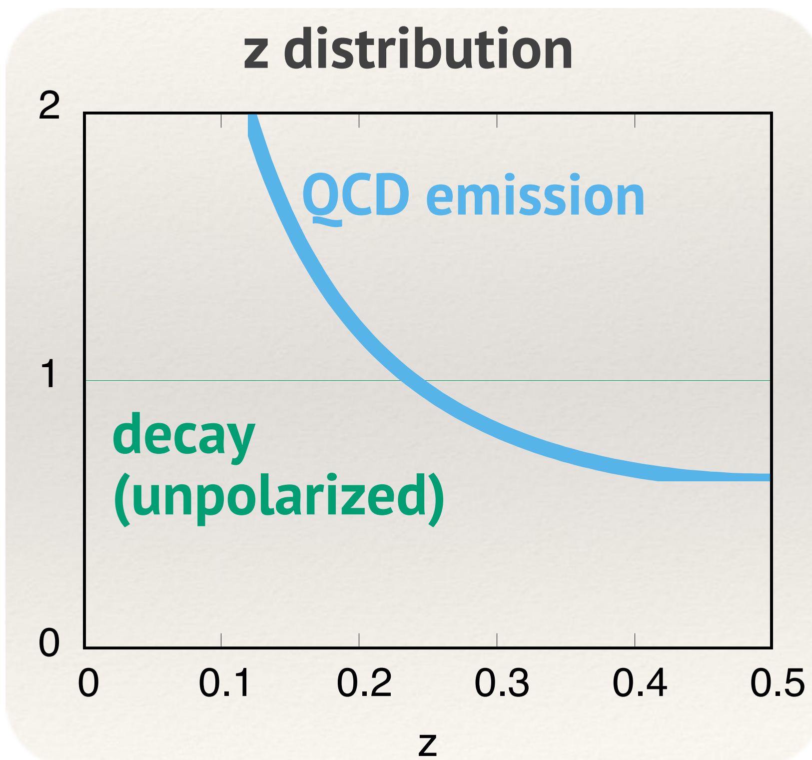
Collinear singularity

Soft singularity

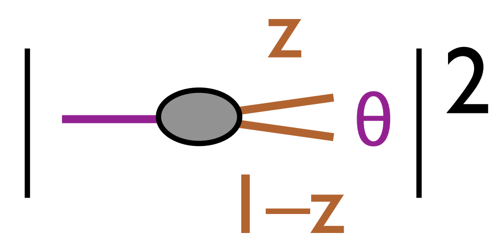
QCD Splitting Functions

Basis for DGLAP evolution of PDFs, parton shower generators, fixed-order subtractions, k_t jet clustering…

Jet Substructure Discrimination



[From Gavin's FCC talk, March 2015]



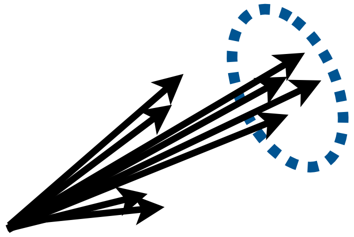
Splitting Function

1→2

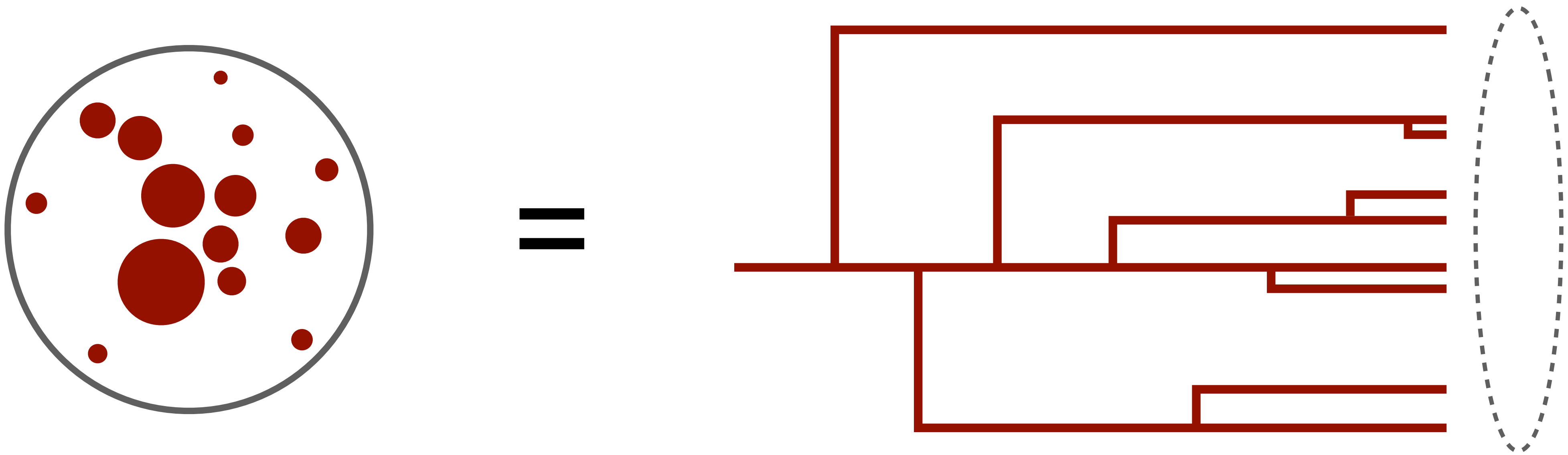
$$\frac{2\alpha_s}{\pi} C_i \frac{\mathrm{d}\theta}{\theta} \frac{\mathrm{d}z}{z}$$

Collinear singularity — Soft singularity

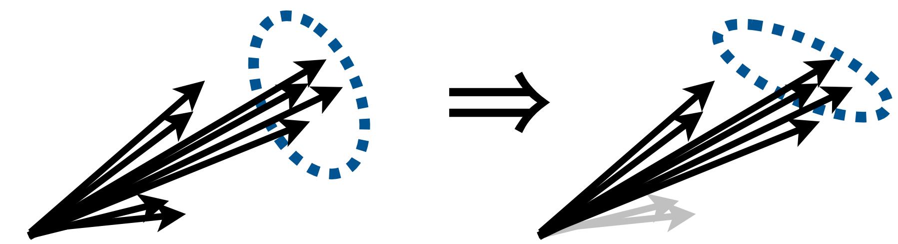
Soft Drop Declustering



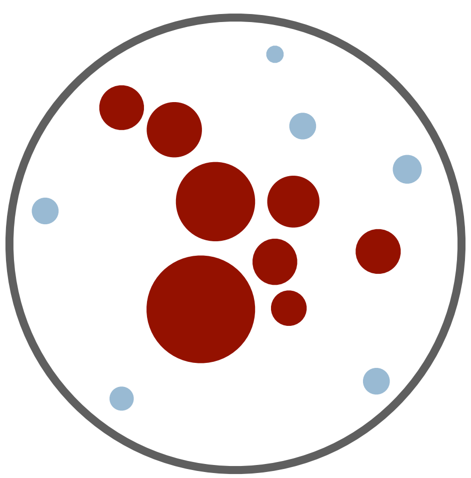
Original Jet　　　　　　　　Clustering Tree



Soft Drop Declustering

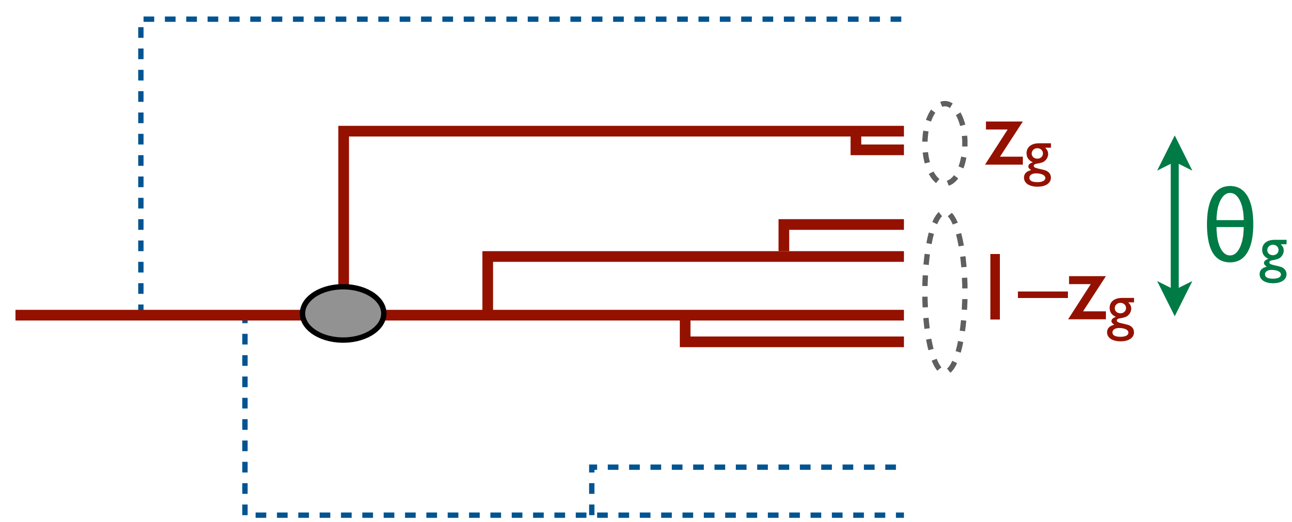


Groomed Jet



=

Groomed Clustering Tree

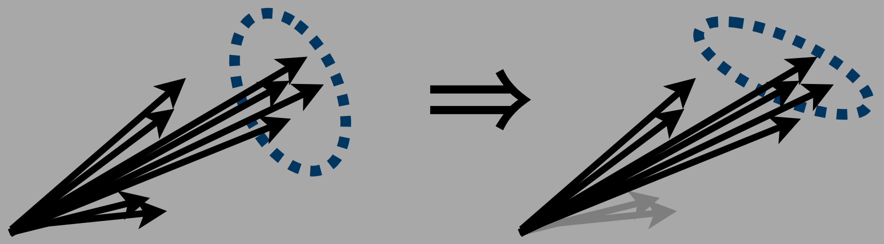


$$z_g > z_{cut}\,\theta_g^{\beta}$$

$\beta = 0$:
mMDT

[Larkoski, Marzani, Soyez, JDT, 1402.2657; see also Butterworth, Davison, Rubin, Salam, 0802.2470; Dasgupta, Fregoso, Marzani, Salam, 1307.0007]

Soft Drop Declustering



From my previous talk at CERN TH:

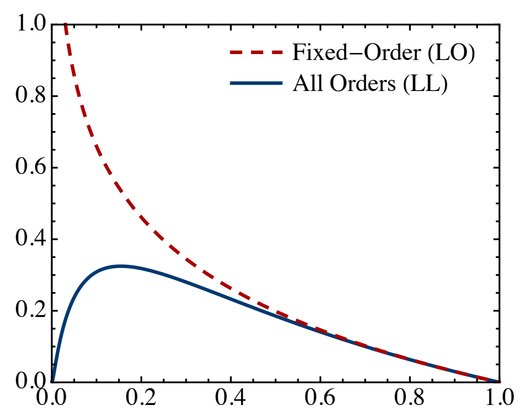
"Sudakov Safe"

Calculable order-by-order in α_s

(!)

$$p(z_g) = \int \mathrm{d}\theta_g \, p(\theta_g) \, p(z_g | \theta_g) \simeq \frac{1}{z_g}$$

Form factor suppresses singularities at all orders in α_s

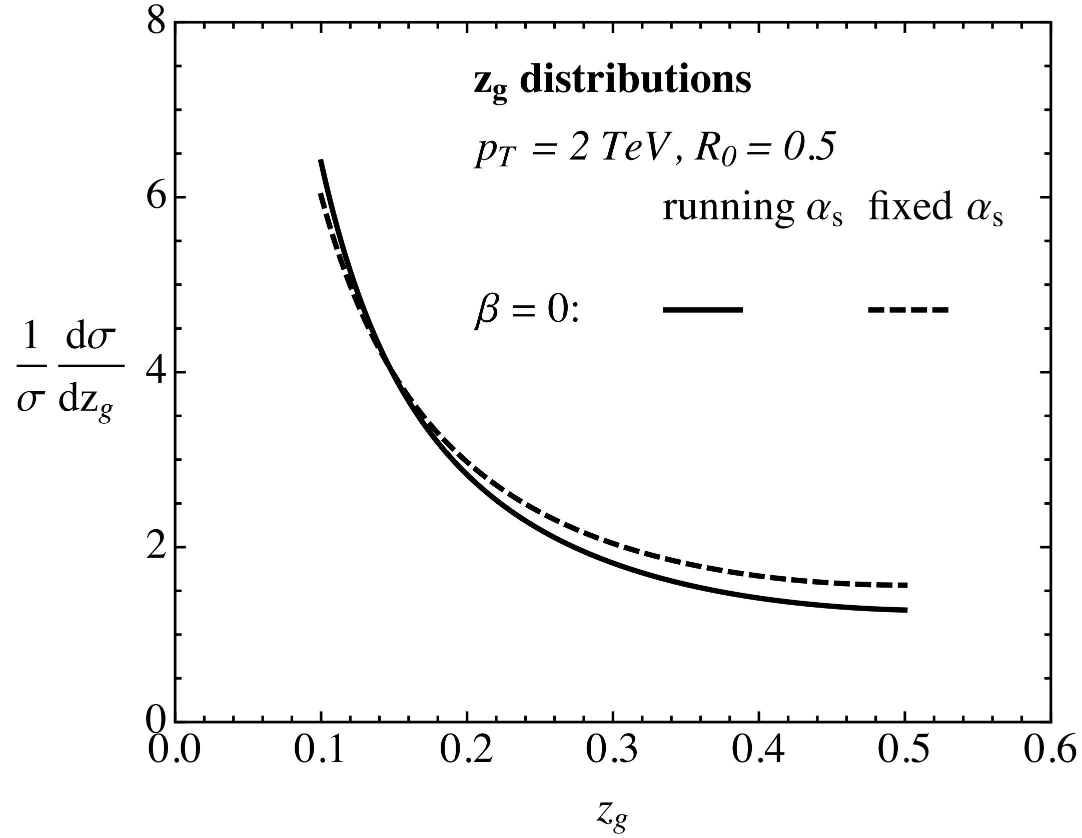


in mMDT
β=0 limit

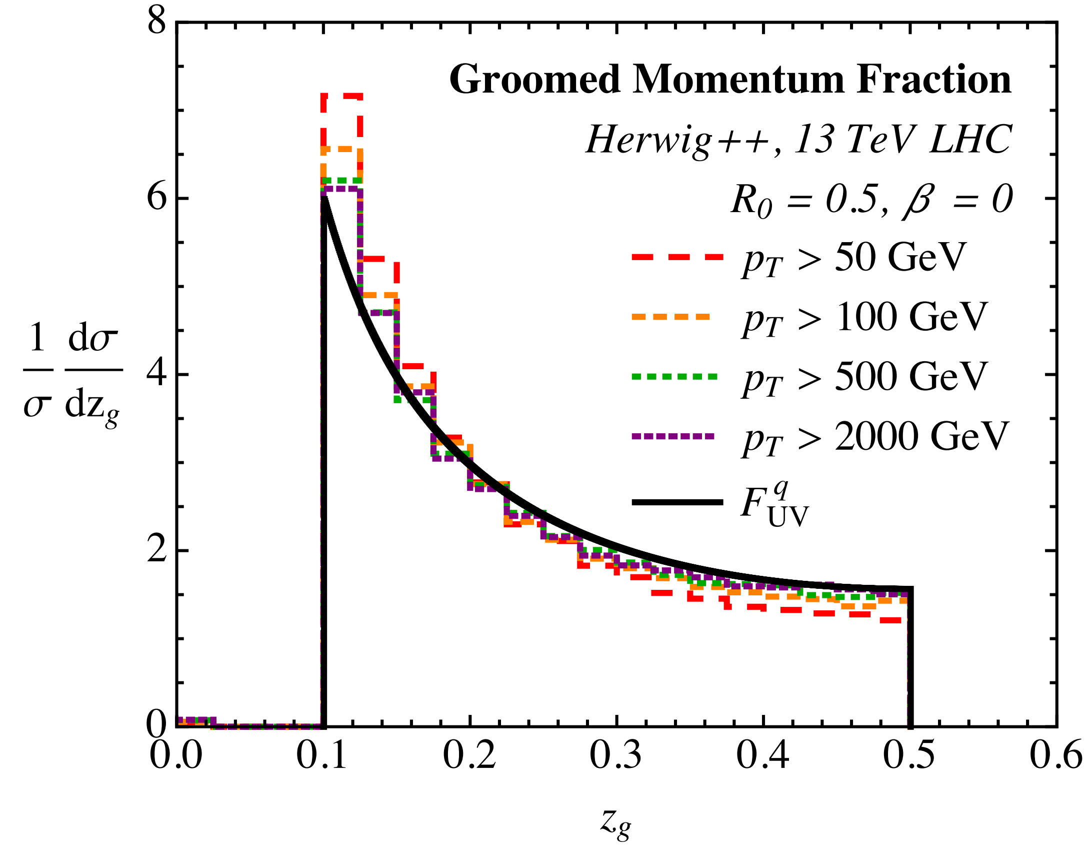
[Larkoski, JDT, 1307.1699; Larkoski, Marzani, JDT, 1502.01719]

[Larkoski, Marzani, Soyez, JDT, 1402.2657; see also Butterworth, Davison, Rubin, Salam, 0802.2470; Dasgupta, Fregoso, Marzani, Salam, 1307.0007]

"Unsafe but Calculable"



Verified with Parton Shower

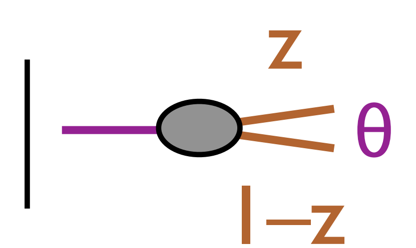


Core Feature of QCD: $\simeq \frac{1}{z_g}$

≈ independent of α_s (!)
≈ independent of jet energy/radius
≈ same for quarks/gluons

$$dP_{i\to ig} \simeq \frac{2\alpha_s}{\pi} C_i \frac{d\theta}{\theta} \frac{dz}{z}$$

cf. $\left| \right|^2$



[Larkoski, Marzani, JDT, 1502.01719; using Larkoski, JDT, 1307.1699]

Perfect application of CMS Open Data

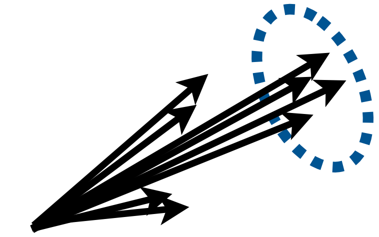
2010 data ⇒ 2014 release ⇒ 2015 idea ⇒ 2017 analysis

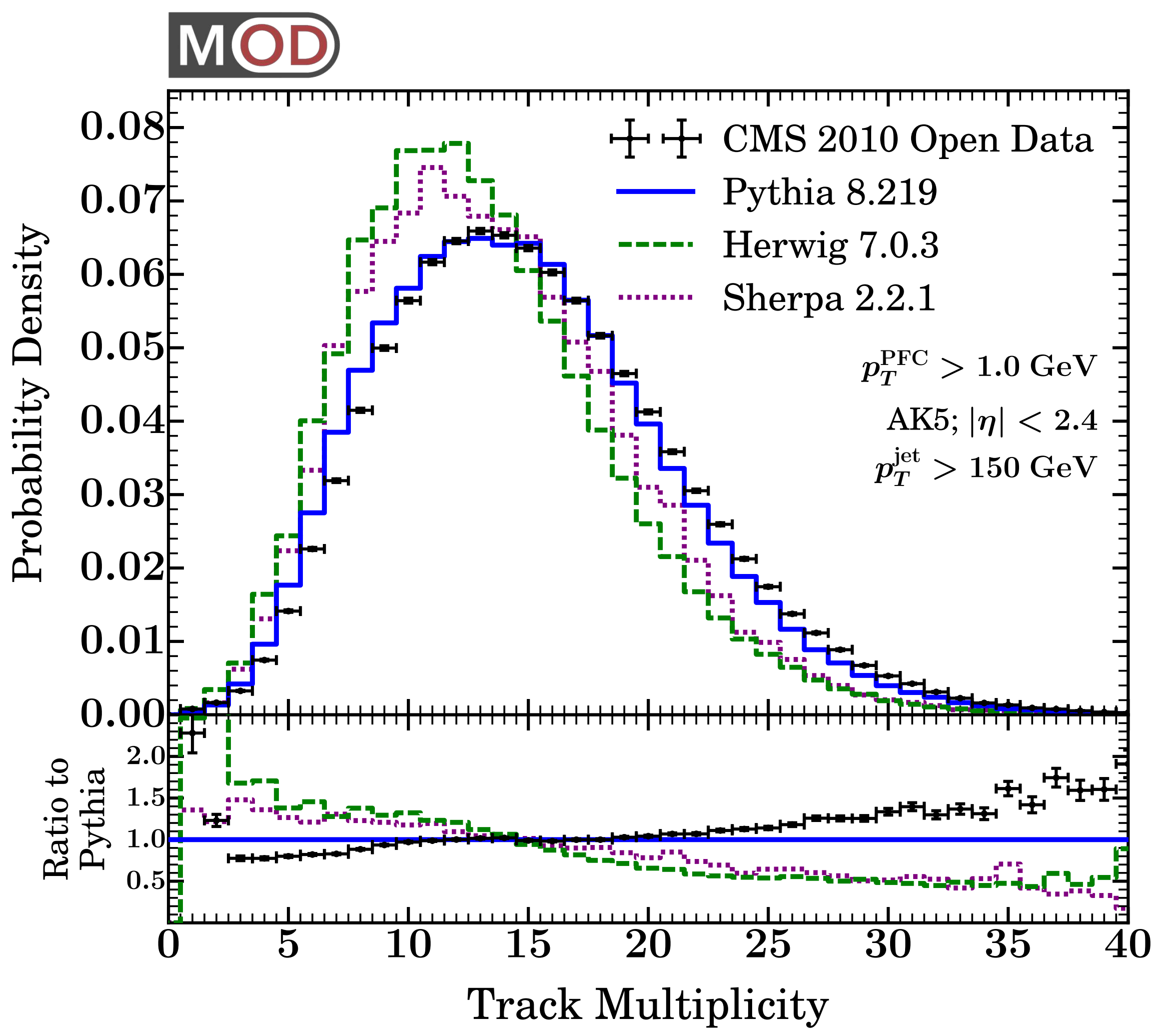
Benefits from low trigger thresholds and low pileup

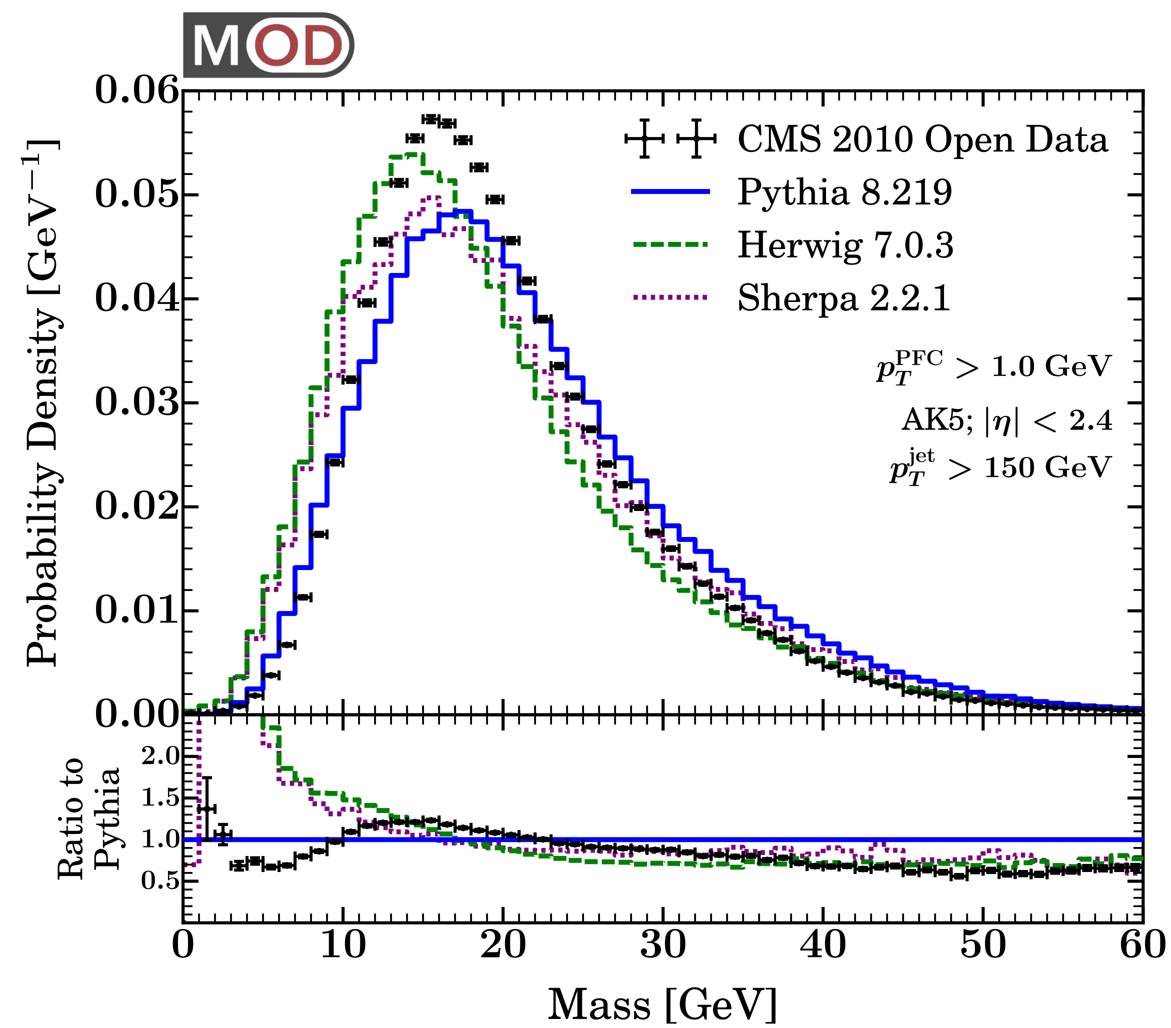
See backup slides for description of particle flow objects and many other jet substructure distributions

Basic Substructure

No grooming applied





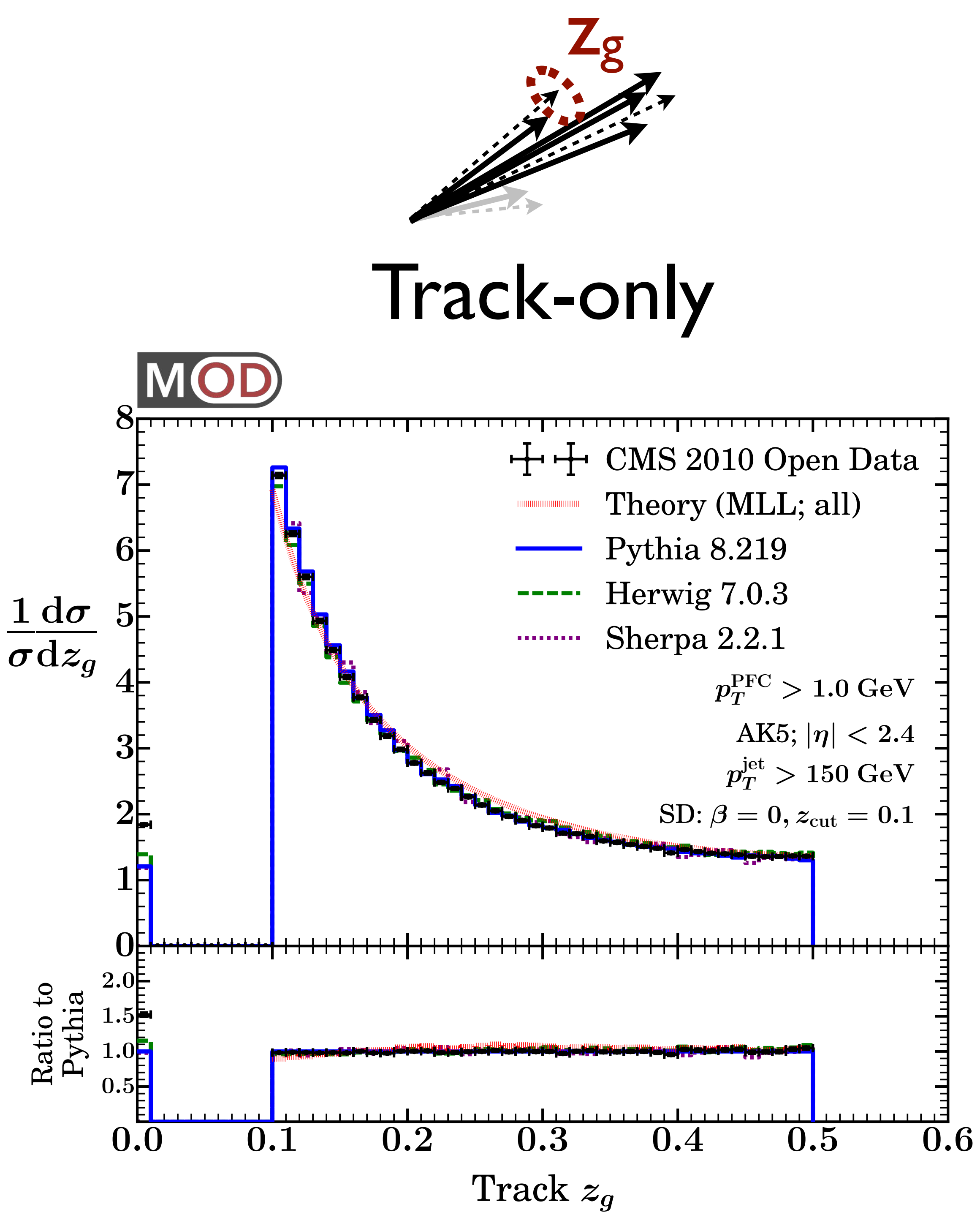


Careful! Can't assess data/MC (dis)agreement without unfolding

Still interesting to investigate MC/MC differences

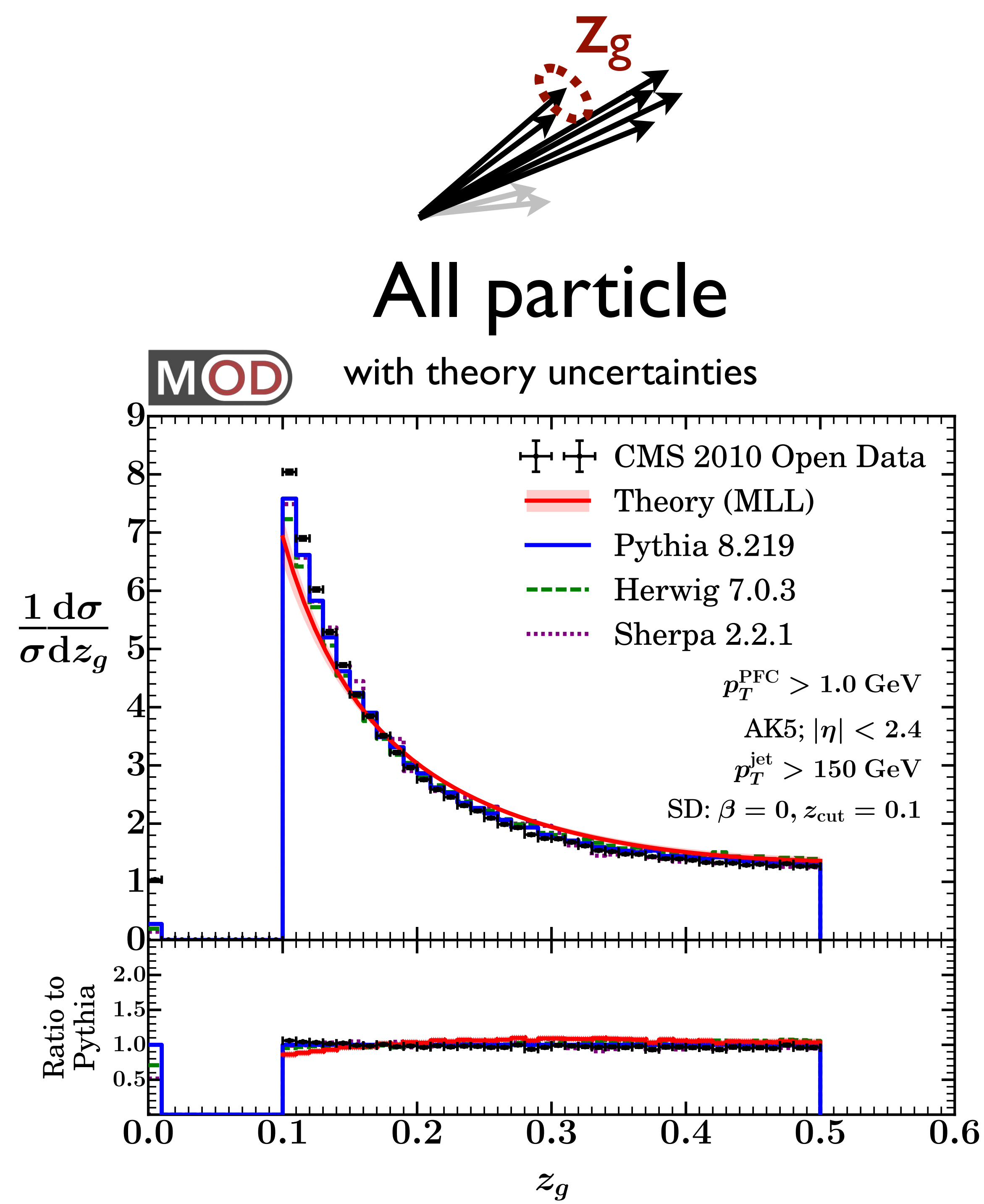
Exposing the QCD Splitting Function

Track-only



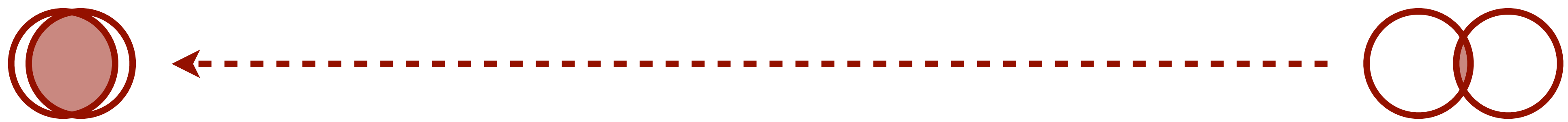
All particle

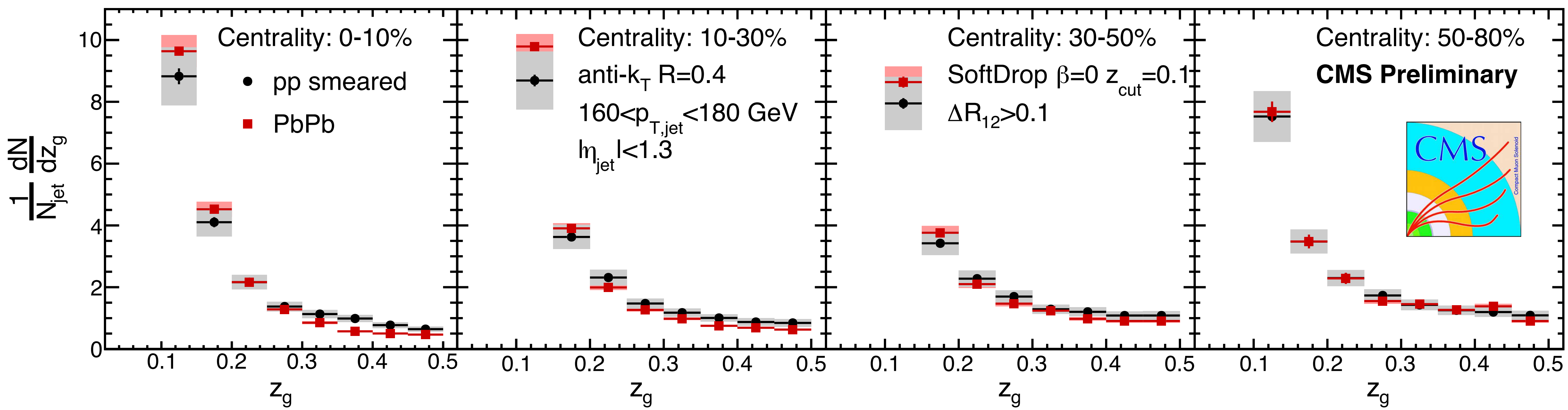
with theory uncertainties

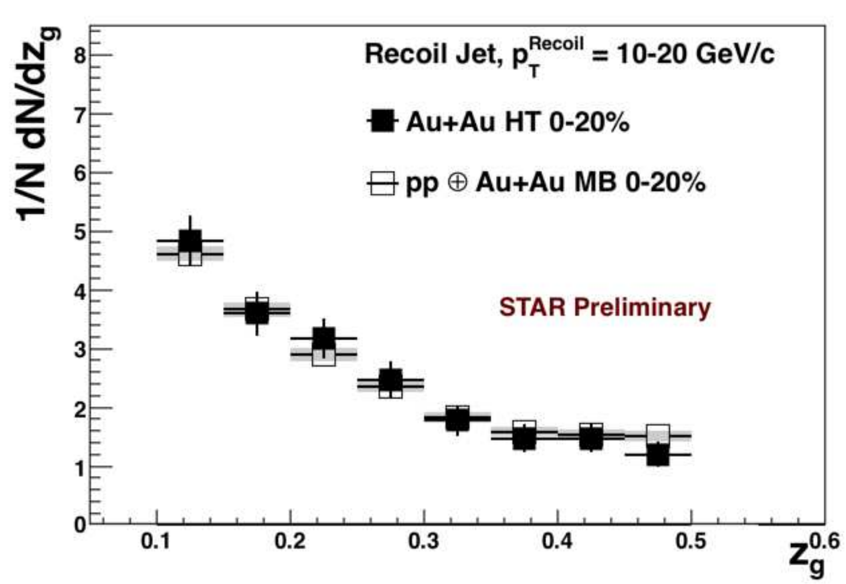


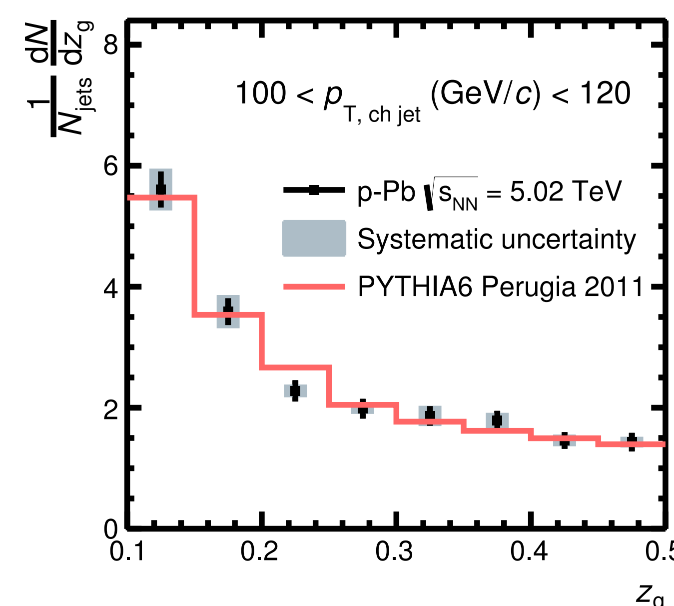
[Larkoski, Marzani, JDT, Tripathee, Xue, 1704.05066]

Preliminary Results from Heavy Ions





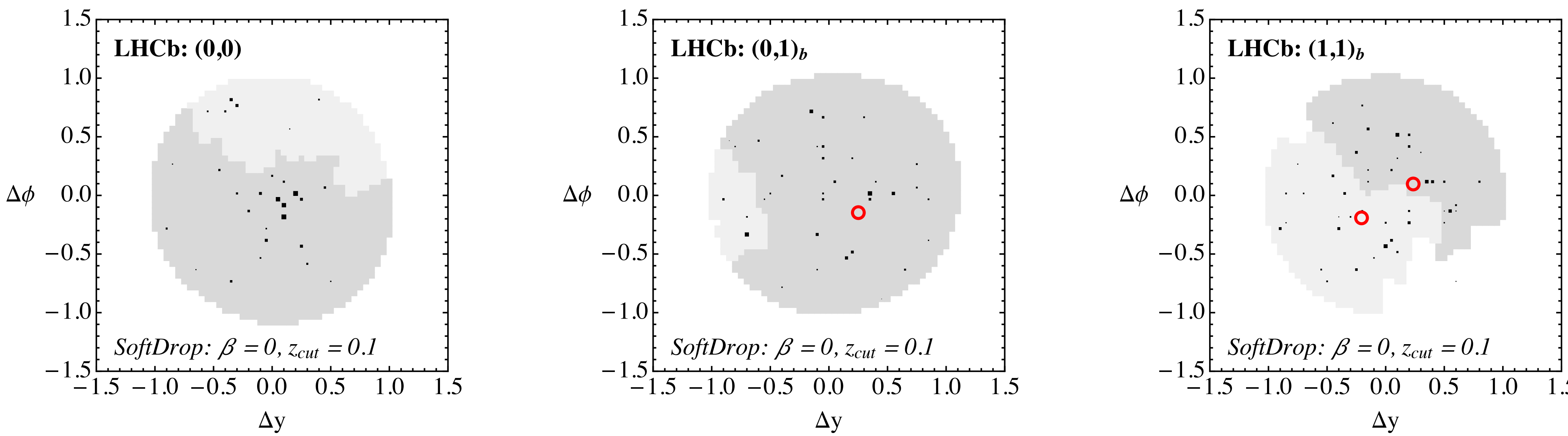


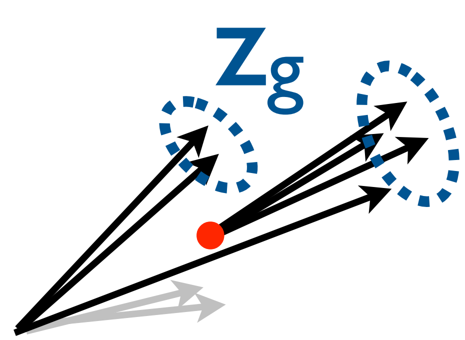


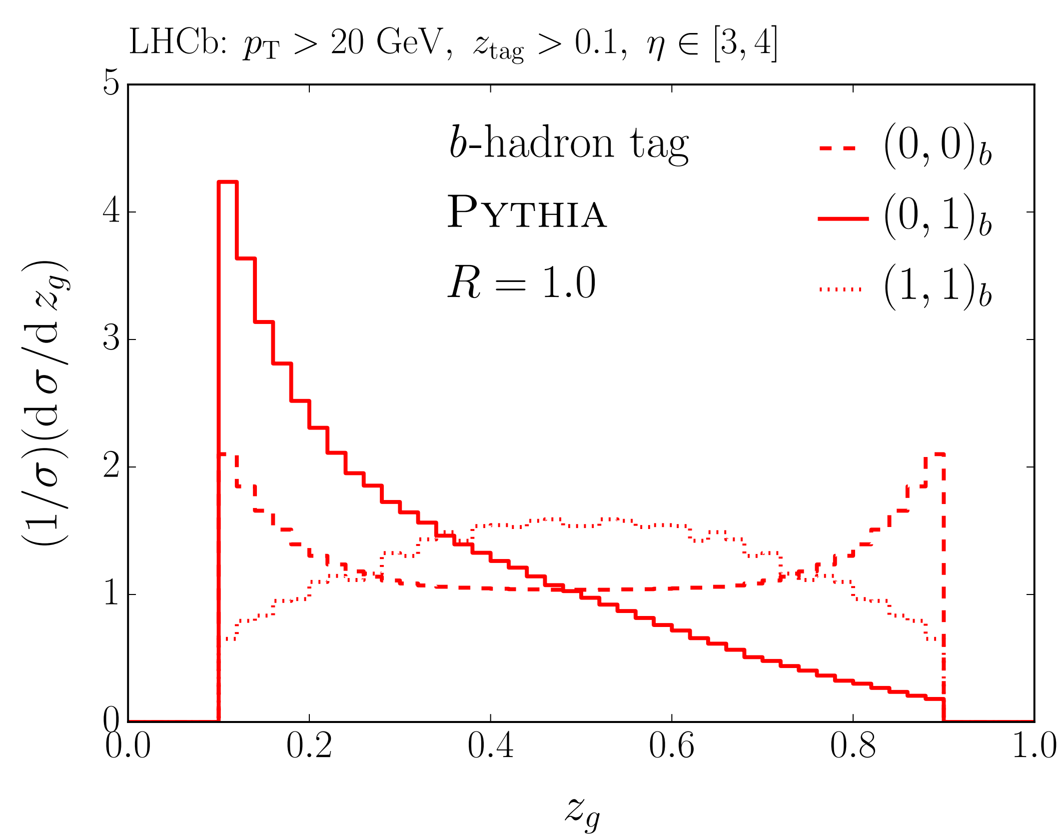
[CMS-PAS-HIN-16-006, STAR preliminary, ALICE preliminary]

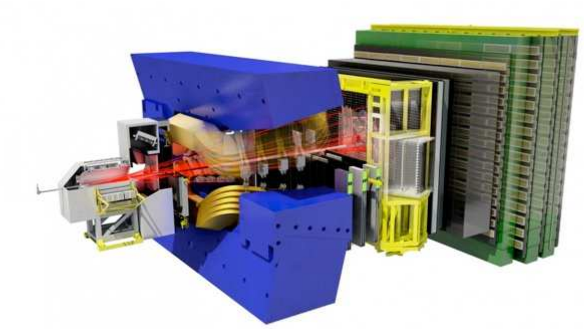
Possibilities for Heavy Flavor

$g \to gg$ $\quad$ $b \to bg$ $\quad$ $g \to b\bar{b}$





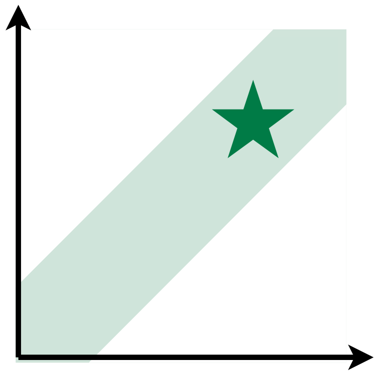




[Ilten, Rodd, JDT, Williams, 1702.02947]

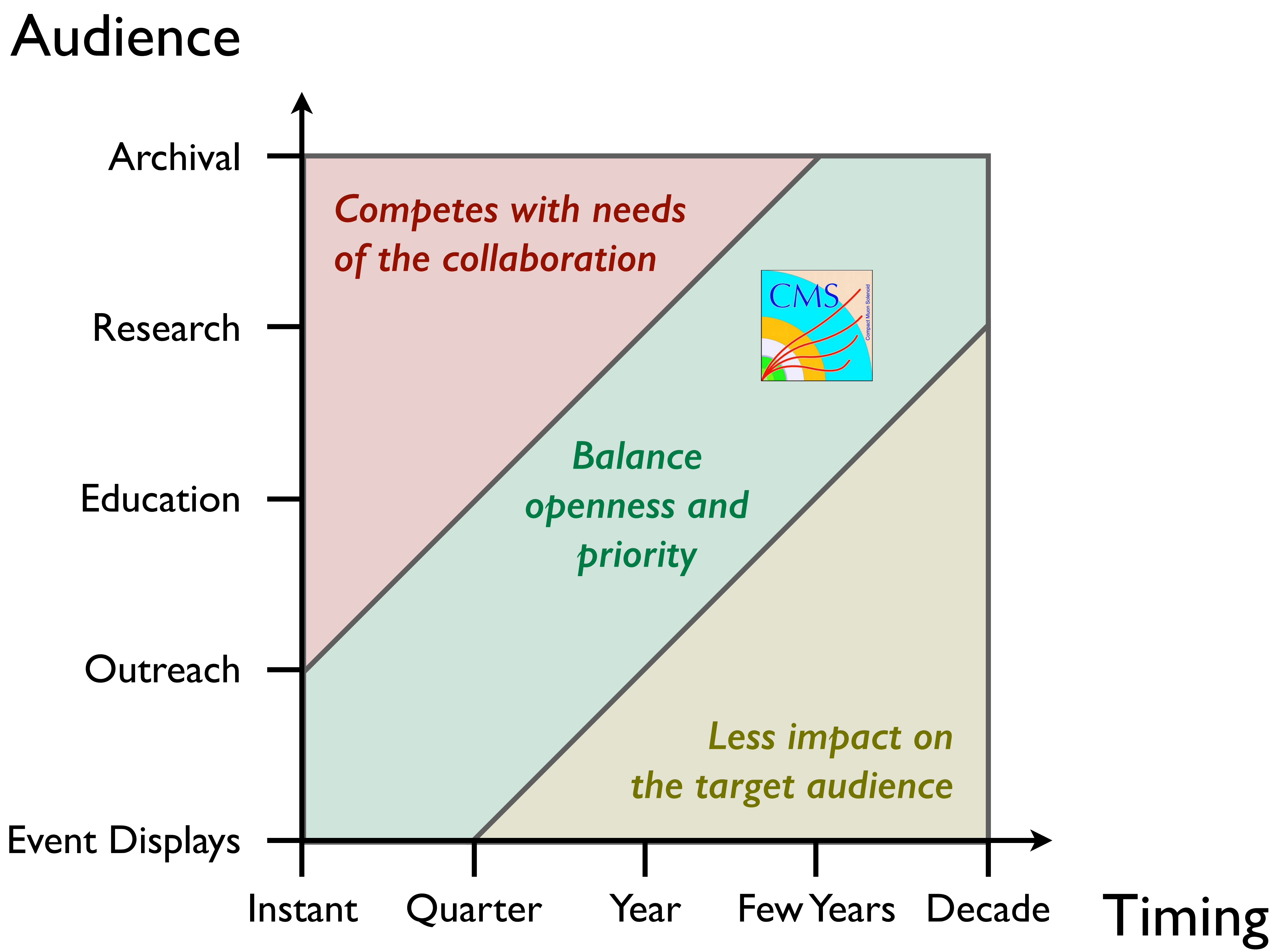
Introducing the CMS Open Data

Jet Substructure and QCD Splittings



(The Future of Public Collider Data)

Different Options for "Public Data"



Viability of CMS Open Data (and expansion to other experiments) depends on interest/enthusiasm of particle physics community

Data preservation (and outside analyses) require significant resources: people, time, ideas, and money

Important to address (valid) concerns about public data for collider physics

Confronting the Steep Learning Curve

"I support open data on principle, but it seems to require an excessive amount of effort to use the CMS Open Data"

CMS Primary Datasets

CMS primary datasets are AOD (Analysis Object Data) files, which contain the information that is needed for analysis

Years: 2010, 2011

Total records:

33

CMS Simulated Datasets

This collection contains CMS Simulated Datasets.

Years: 2010, 2011

Total records:

381

vs.

CMS Learning Resources

This collection includes learning resources that use CMS public data

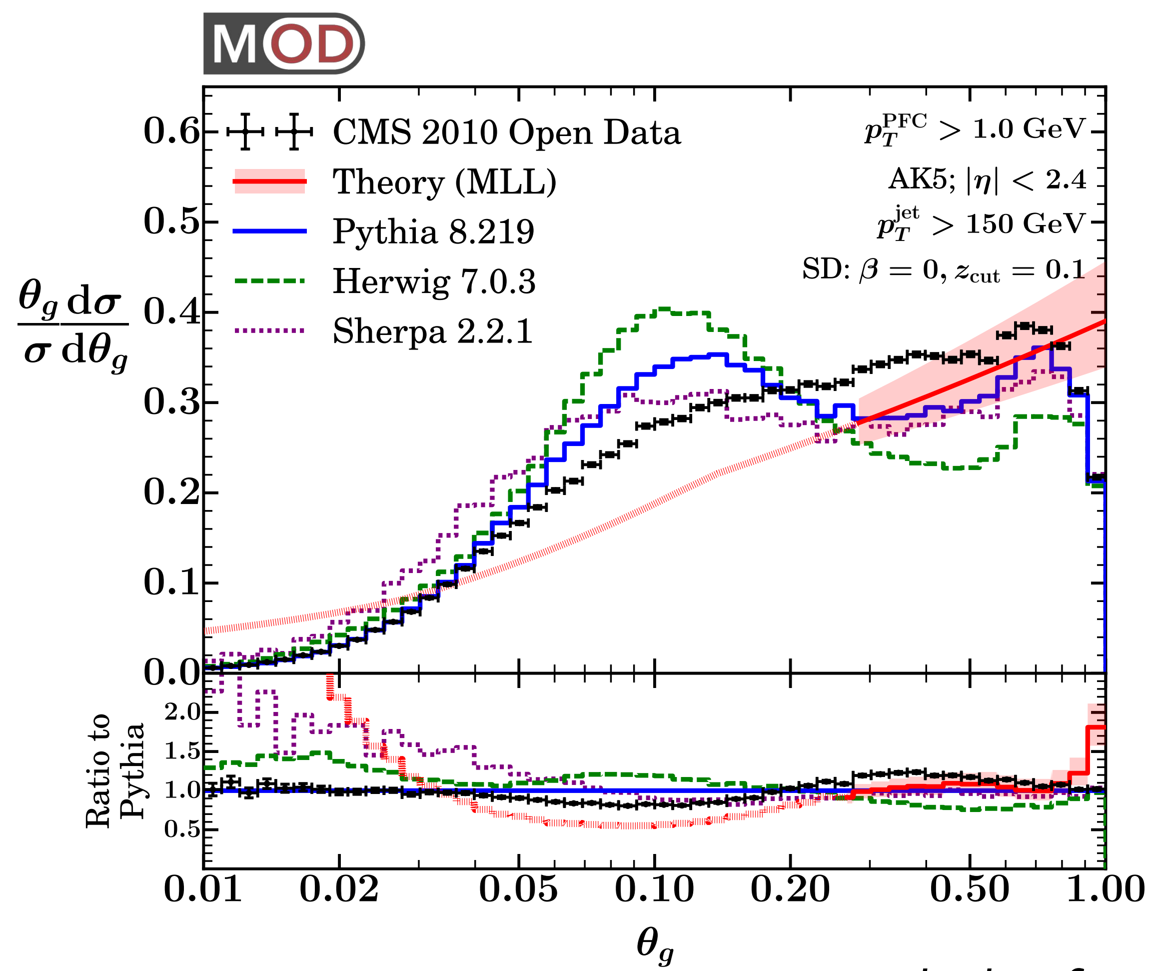
Total records:

7

With a suitable investment, open data could be as straightforward to parse and interpret as detector-simulated Monte Carlo (cf. MOD data format)

Balance between Sophistication and Exploration

"There is no way you can do an external analysis with the same degree of sophistication as within the collaboration"



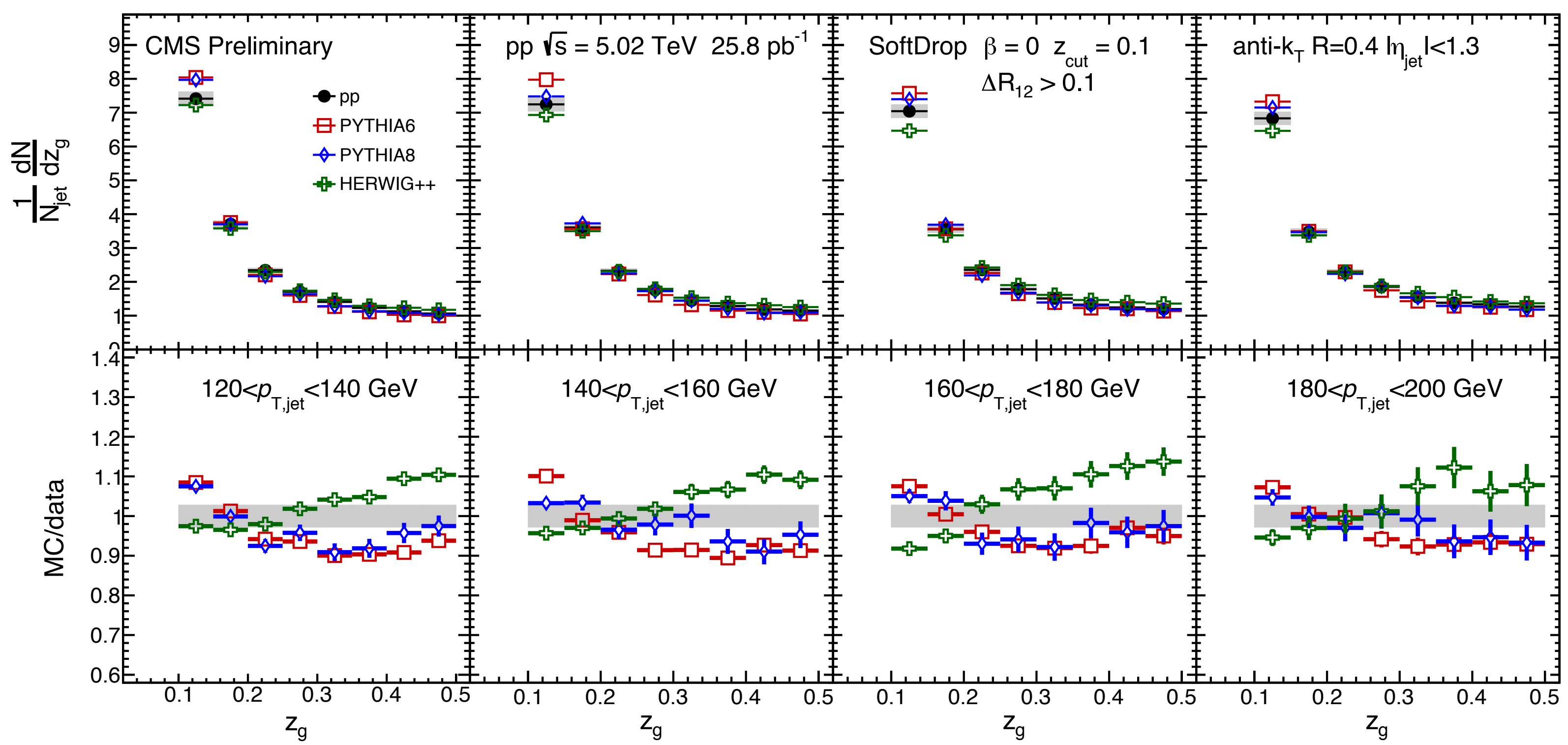
see backup for joint (z_g, θ_g)

Agreed, but with unexpected theoretical and experimental issues at play, valuable to explore data/calculations before precision studies

[Tripathee, Xue, Larkoski, Marzani, JDT, 1704.05842]

Synergy between Internal and External Efforts

"This work competes with ongoing collaboration analyses and does not meet the standards of experimental particle physics"



I will be heartbroken if our work impedes experimental progress on z_g, and I will be thrilled if our work inspires more rigorous investigations into the QCD splitting function

[CMS-PAS-HIN-16-006]

Value of Open-Ended Investigations

"If you really wanted to do this jet substructure measurement, you should have joined CMS as a short term associate"

Getting started with CMS 2010 data

"I have installed the CERN Virtual Machine: now what?"

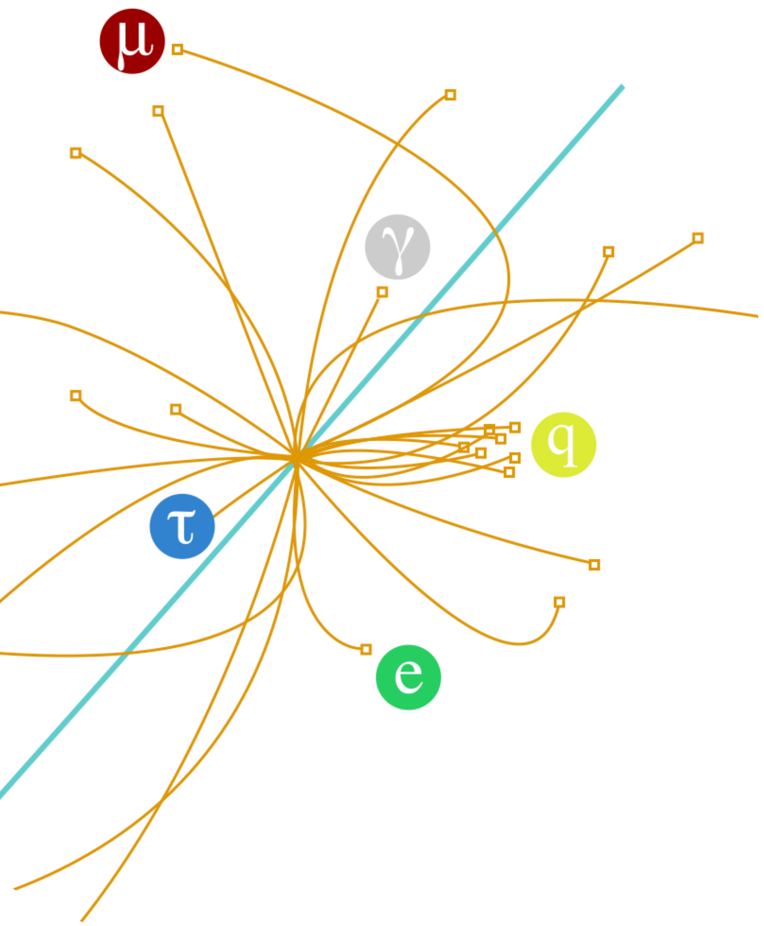
To analyse CMS data collected in 2010, you need **version 4.2.8** of CMSSW, supported only on **Scientific Linux 5**. If you are unfamiliar with Linux, take a look at this short introduction to Linux or try this interactive command-line bootcamp. Once you have installed the CMS-specific CERN Virtual Machine, execute the following command in the terminal if you haven't done so before; it ensures that you have this version of CMSSW running:

```
$ cmsrel CMSSW_4_2_8
```

Agreed, but what I really wanted to do is figure out the answer to this question (curiosity-driven research)

My View

The CMS Open Data is a fantastic resource, with many exciting applications



Educating future scientists

Stress-testing archival data strategies

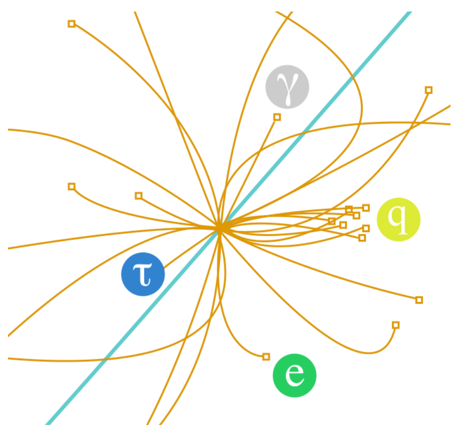
Enabling exploratory/proof-of-principle studies

Facilitating dialogue between theory and experiment

Researching physics in and beyond the standard model

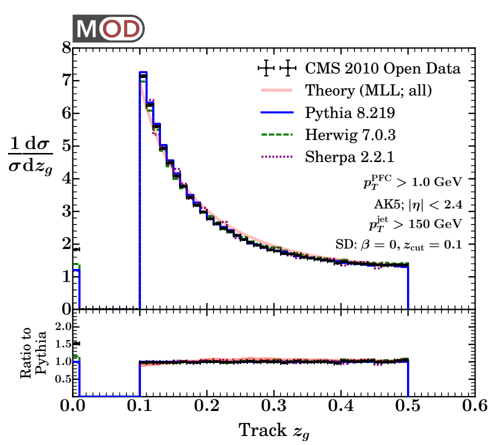
These are only possible with sustained investments in public data initiatives

Summary



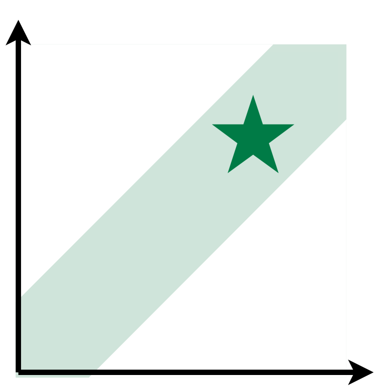
Introducing the CMS Open Data

Unique collider data set, ideal for exploratory studies



Jet Substructure and QCD Splittings

Exposing the universal singularity structure of gauge theories



(The Future of Public Collider Data)

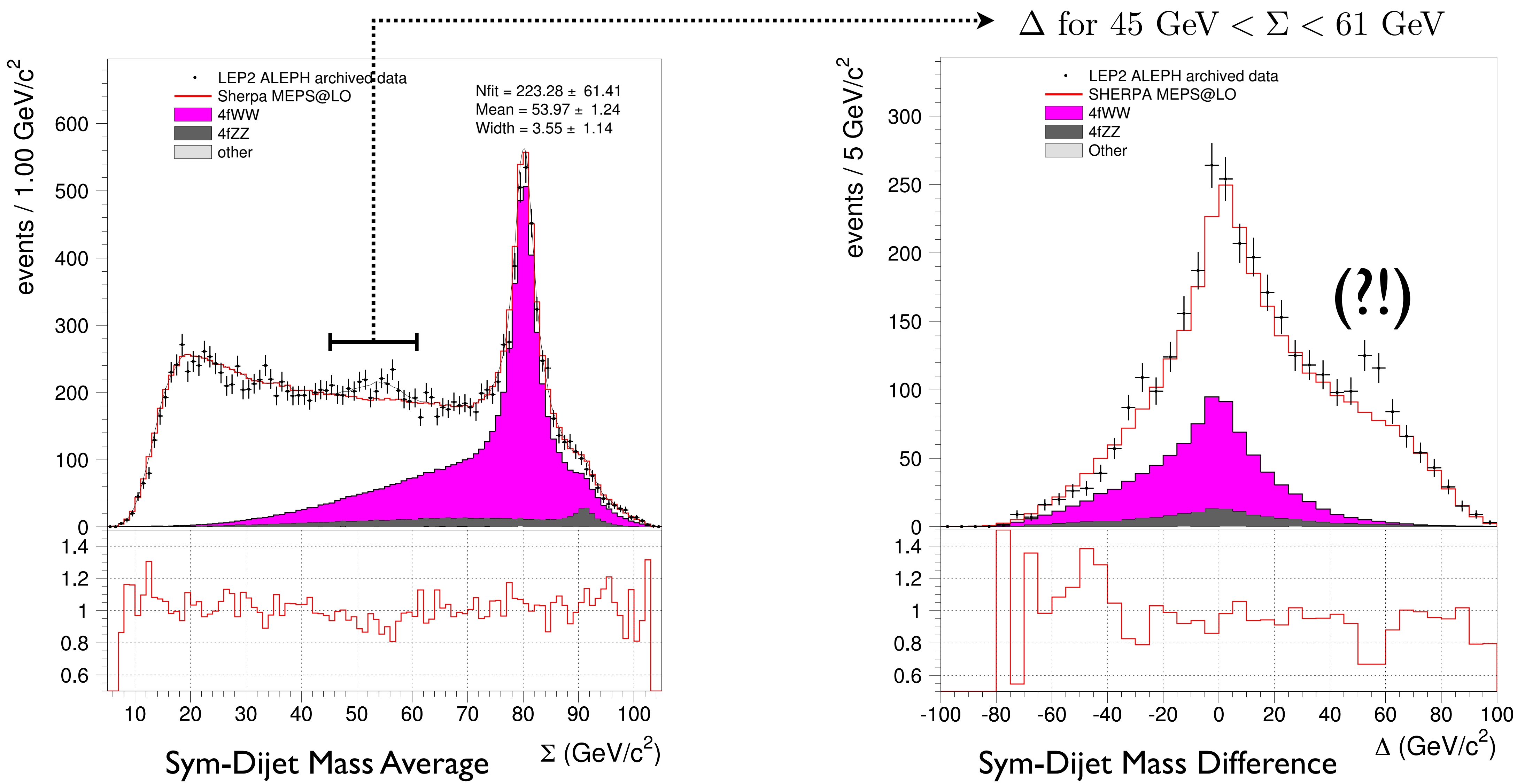
Sustained investment from outreach to research to archives

ACKNOWLEDGMENTS

We applaud CERN for the historic launch of the Open Data Portal, and we congratulate the CMS collaboration for the fantastic performance of their detector and the high quality of the resulting public data set. We thank Alexis Romero for collaboration in the early stages of this work. We are indebted to Salvatore Rappoccio and Kati Lassila-Perini for helping us navigate the CMS software framework. We benefitted from code and encouragement from Tim Andeen, Matt Bellis, Andy Buckley, Kyle Cranmer, Sarah Demers, Guenther Dissertori, Javier Duarte, Peter Fisher, Achim Geiser, Giacomo Govi, Phil Harris, Beate Heinemann, Harri Hirvonsalo, Markus Klute, Greg Landsberg, Yen-Jie Lee, Elliot Lipeles, Peter Loch, Marcello Maggi, David Miller, Ben Nachman, Christoph Paus, Alexx Perloff, Andreas Pfeiffer, Maurizio Pierini, Ana Rodriguez, Gunther Roland, Ariel Schwartzman, Liz Sexton-Kennedy, Maria Spiropulu, Nhan Tran, Ana Trisovic, Chris Tully, Marta Verweij, Mikko Voutilainen, and Mike Williams.

Backup Slides

A Quad-Jet Puzzle in Archival ALEPH Data



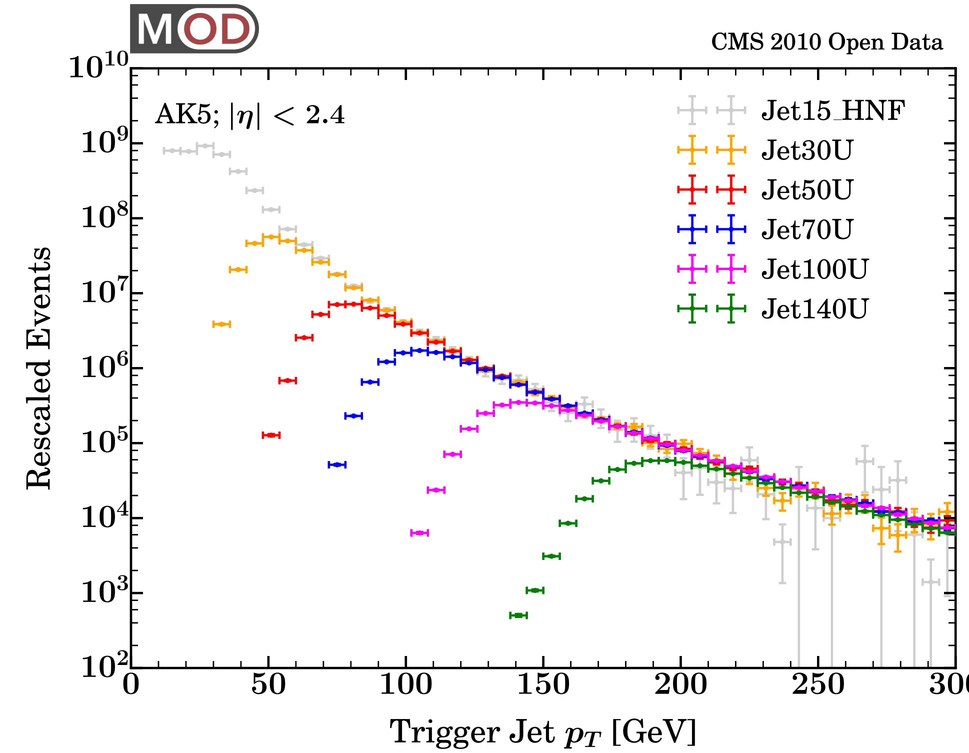
Science thrives on openness, reproducibility, and intense scrutiny

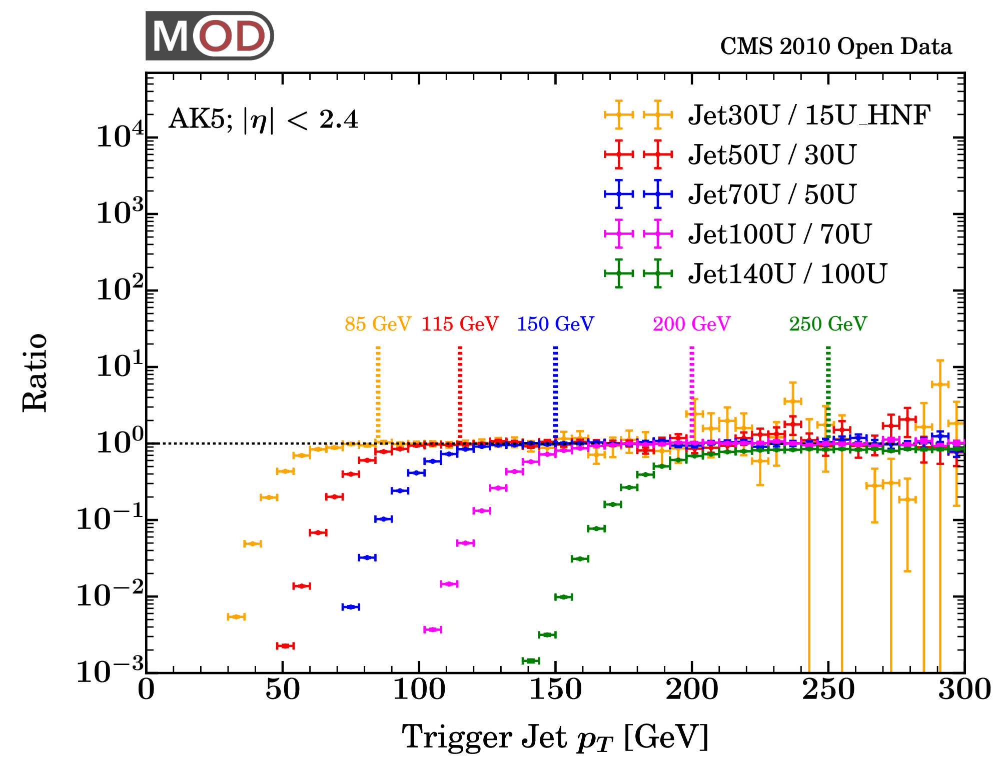
What role should legacy data sets play in collider physics?

[Kile, von Wimmersperg-Toeller, 1706.02242, 1706.02255, 1706.02269]

Trigger Selection and Efficiency

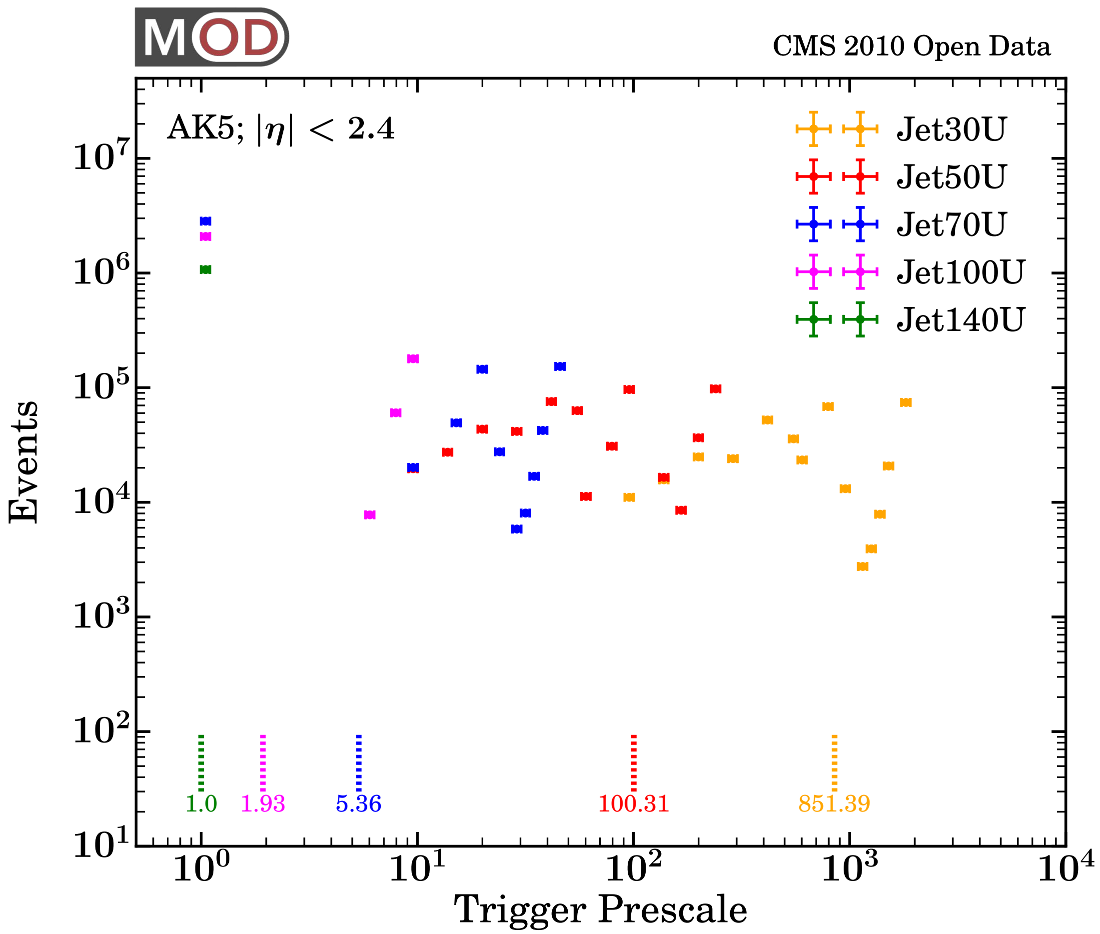
	Trigger	Present?	Fired?
Single-jet	HLT_Jet15U	16,341,190	1,342,155
	* HLT_Jet15U_HNF	16,341,190	1,341,930
	* HLT_Jet30U	16,341,190	604,287
	* HLT_Jet50U	16,341,190	870,649
	* HLT_Jet70U	16,341,190	5,257,339
	* HLT_Jet100U	16,341,190	3,689,951
	* HLT_Jet140U	5,989,945	1,898,874
	HLT_Jet180U	2,595,038	553,331
Di-jet	HLT_DiJetAve15U	16,341,191	1,067,561
	HLT_DiJetAve30U	16,341,191	648,000
	HLT_DiJetAve50U	16,341,191	859,292
	HLT_DiJetAve70U	16,341,191	2,310,033
	HLT_DiJetAve100U	5,989,945	1,252,661
	HLT_DiJetAve140U	2,595,038	452,222
Quad-jet	HLT_QuadJet20U	10,351,245	677,451
	HLT_QuadJet25U	10,351,244	219,256
H_T	HLT_HT100U	10,351,245	7,369,985
	HLT_HT120U	10,351,245	4,090,218
	HLT_HT140U	10,351,245	2,430,208
	HLT_EcalOnly_SumEt160	10,351,246	208,718





Event Selection, Prescale Factors, Pileup

Hardest Jet p_T	Trigger Name	Events	⟨Prescale⟩
[85, 115] GeV	HLT_Jet30U	33,375	851.514
[115, 150] GeV	HLT_Jet50U	66,412	100.320
[150, 200] GeV	HLT_Jet70U	365,821	5.362
[200, 250] GeV	HLT_Jet100U	216,131	1.934
> 250 GeV	HLT_Jet100U	34,736	1.000
	HLT_Jet140U	177,891	1.000



	Jet Primary Dataset		Hardest Jet Selection	
N_{PV}	Events	Fraction	Events	Fraction
1	4,716,494	0.289	190,277	0.248
2	4,814,495	0.295	246,387	0.321
3	3,630,413	0.222	180,021	0.234
4	1,933,832	0.118	93,587	0.122
5	819,835	0.050	38,598	0.050
6	294,612	0.018	13,805	0.018
7	93,714	0.006	4,318	0.006
8	27,550	0.002	1,242	0.002
9	7,481	0.000	330	0.000
10	2,041	0.000	91	0.000
11	540	0.000	21	0.000
12	125	0.000	6	0.000
13	41	0.000	3	0.000
14	9	0.000	1	0.000
≥ 15	5	0.000	0	0.000

Workflow

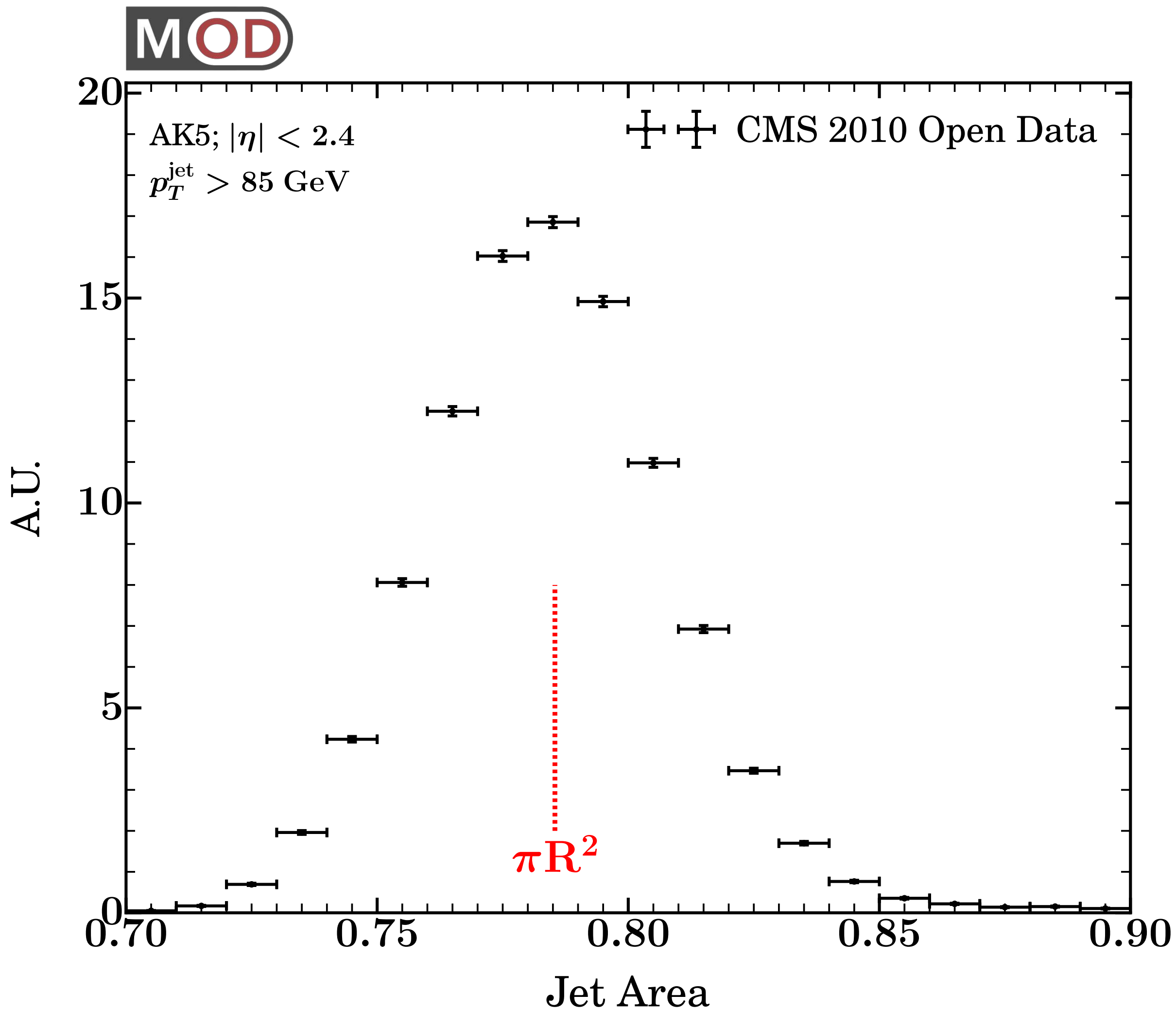
Instrumentation

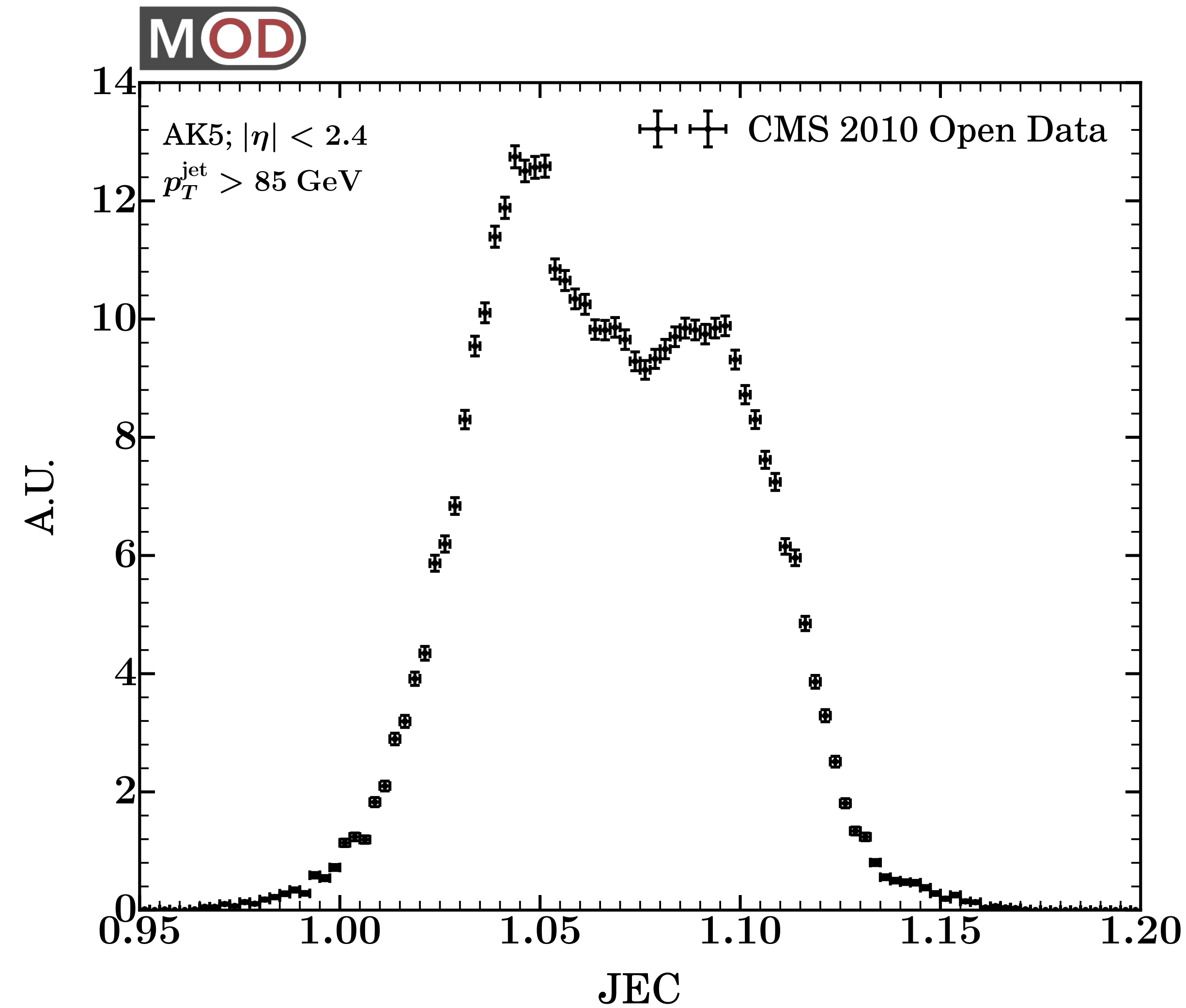
	Events	Fraction	
Jet Primary Dataset	20,022,826	1.000	
Validated Run	16,341,187	0.816	Provided by CMS
Assigned Trigger Fired (Table II)	894,366	0.045	Derived by us, consistent with CMS
Loose Jet Quality (Table V)	843,129	0.042	Provided by CMS
AK5 Match	843,128	0.042	Numerical rounding issue
$\lvert\eta\rvert < 2.4$	768,687	0.038	Central jets
Passes Soft Drop ($z_g > z_{\text{cut}}$)	760,055	0.038	Jet grooming (more later)

Physics

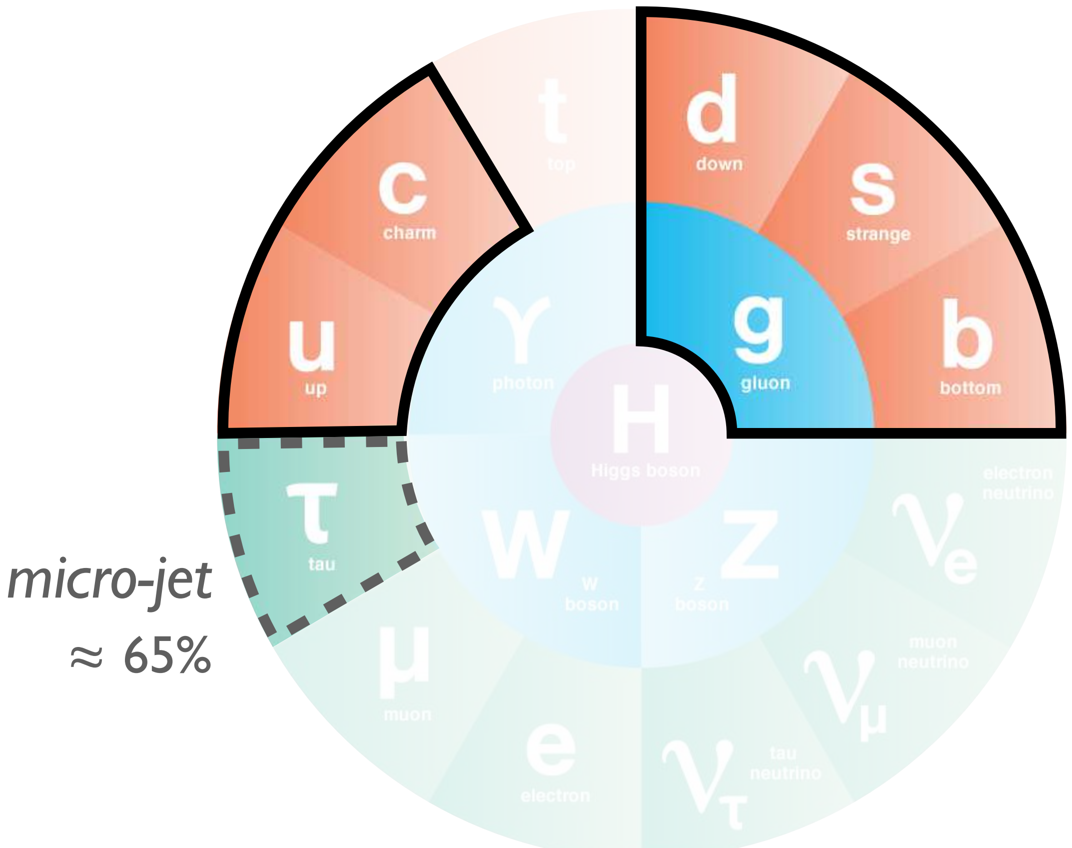
Factor of 20 reduction in events by using $\approx$100% efficient triggers on high-quality jets

Jet Corrections



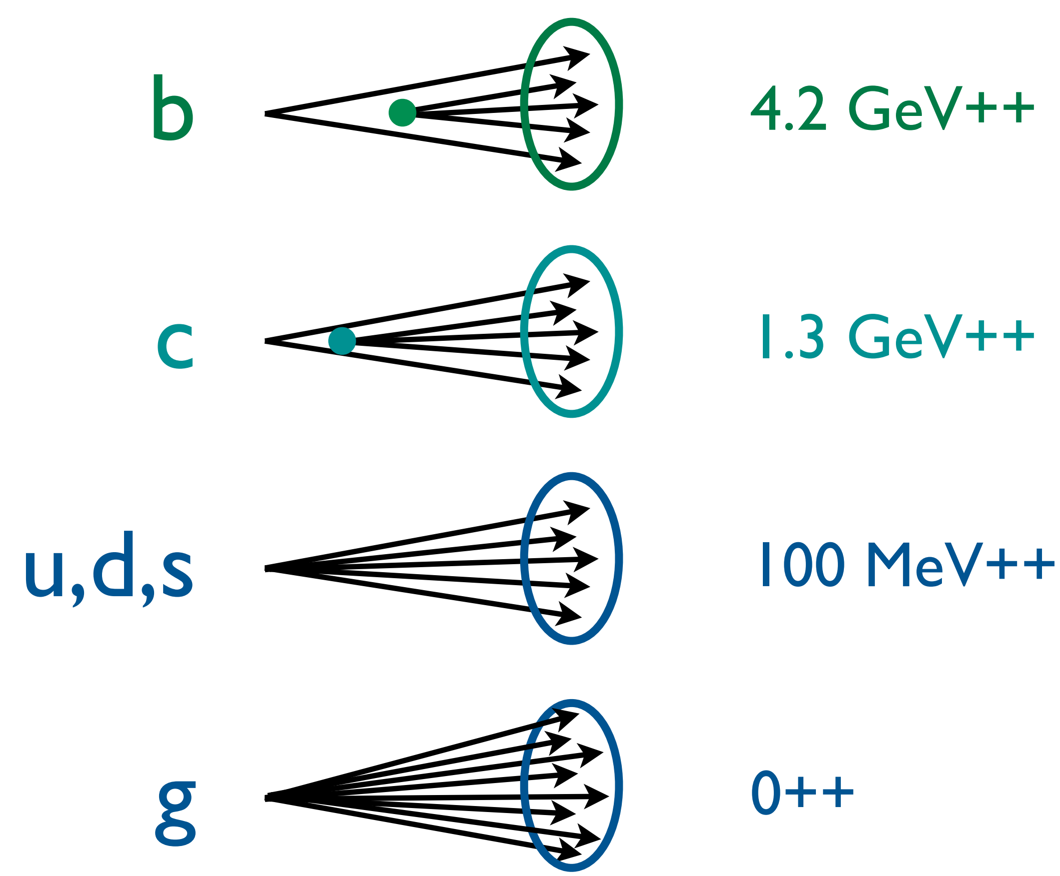


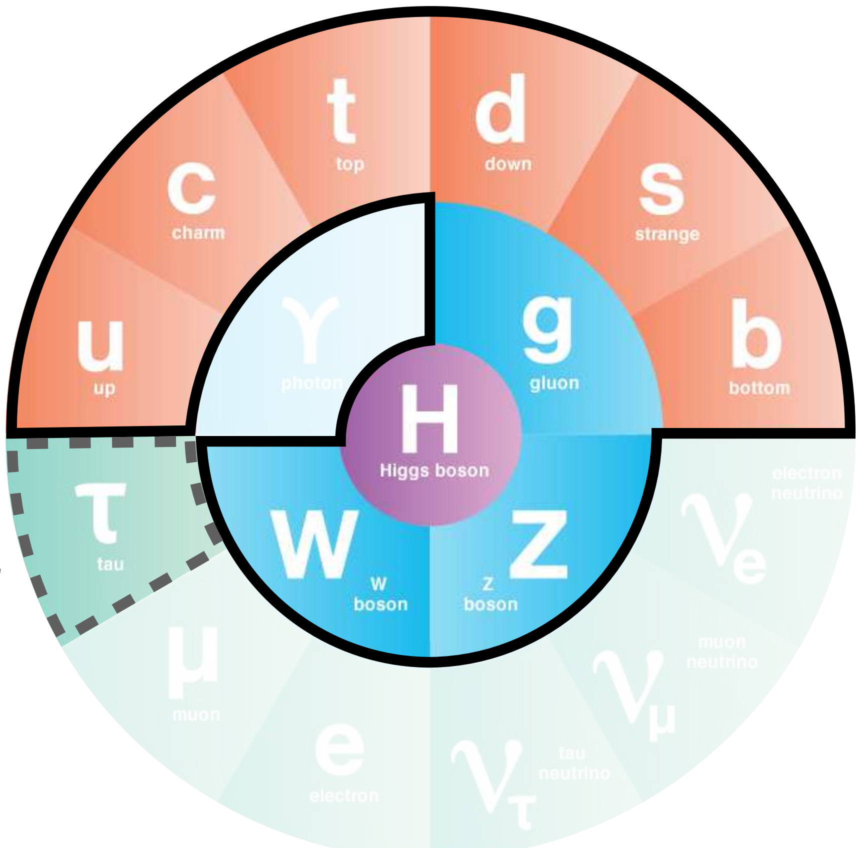
Jet Area Subtraction ------→ Jet Energy Corrections



Jets from the Standard Model

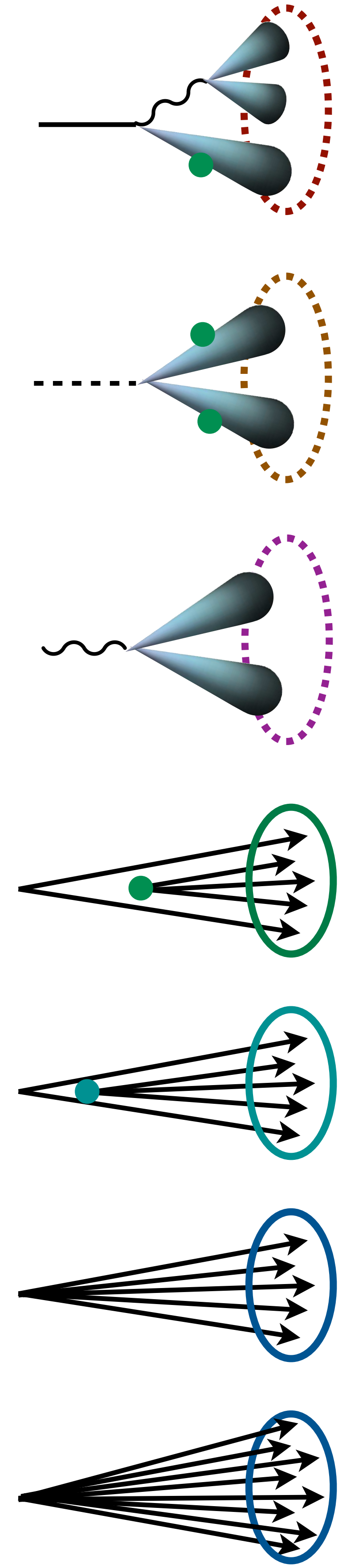
++ = plus gluonic radiation





micro-jet
≈ 65%

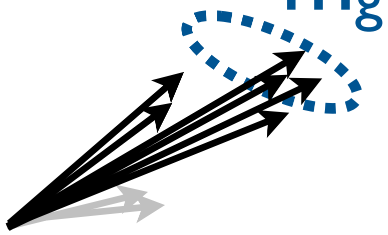
Jets from the Standard Model



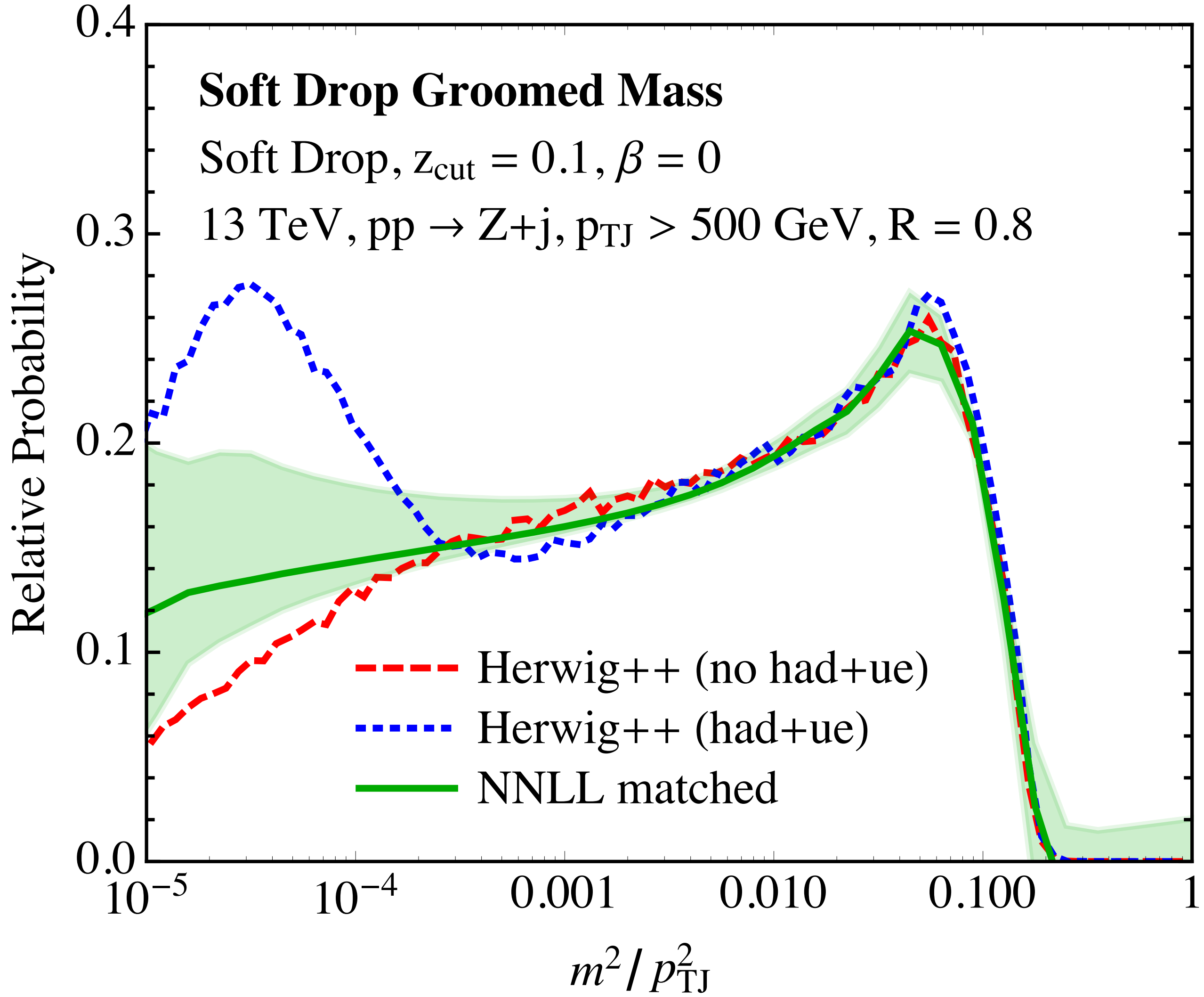
t ≈ 70%		173 GeV++
H ≈ 60%		125 GeV
W/Z ≈ 70%		80/91 GeV
b		4.2 GeV++
c		1.3 GeV++
u,d,s		100 MeV++
g		0++

++ = plus gluonic radiation

Soft Drop Jet Mass



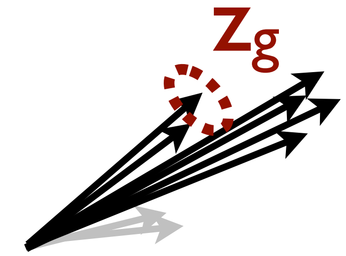
First NNLL + O(α_s^2) result for substructure in pp (!)



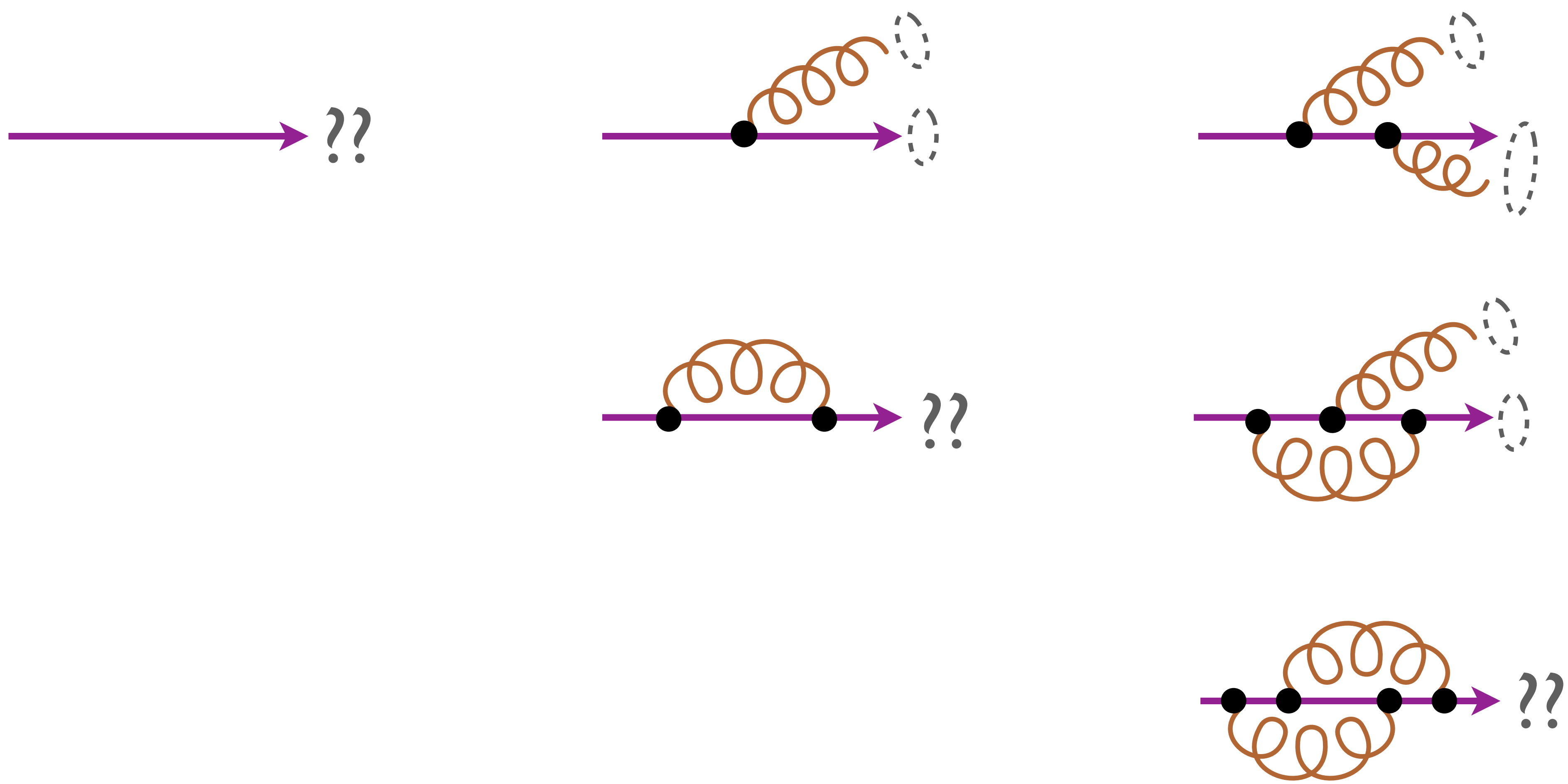
Grooming *simplifies* structure of calculation, *reduces* NP effects

[Frye, Larkoski, Schwartz, Yan, 1603.06375, 1603.09338; see also Marzani, Schunk, Soyez, 1704.02210]

Soft Drop Momentum Fraction



$$\frac{\mathrm{d}\sigma}{\mathrm{d}z_g} = \Big(\textit{undefined} \Big) + \alpha_s \Big(\textit{infinity} \Big) + \alpha_s^2 \Big(\textit{infinity}^2 \Big) + \ldots$$



z_g

Collinear Unsafe*

Can't make prediction from perturbative QCD (?)

*unless you simultaneously restrict jet mass

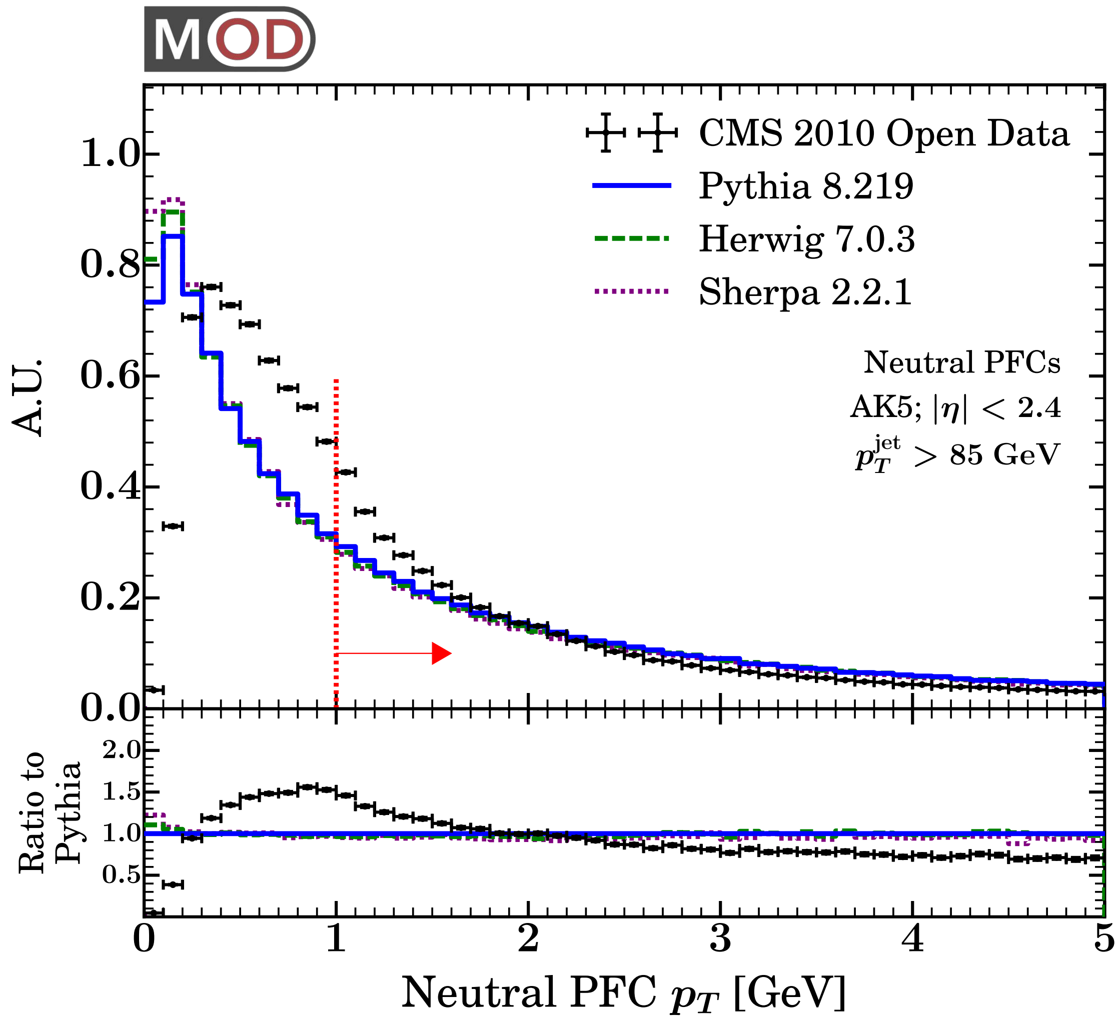
Particle Flow Reconstruction

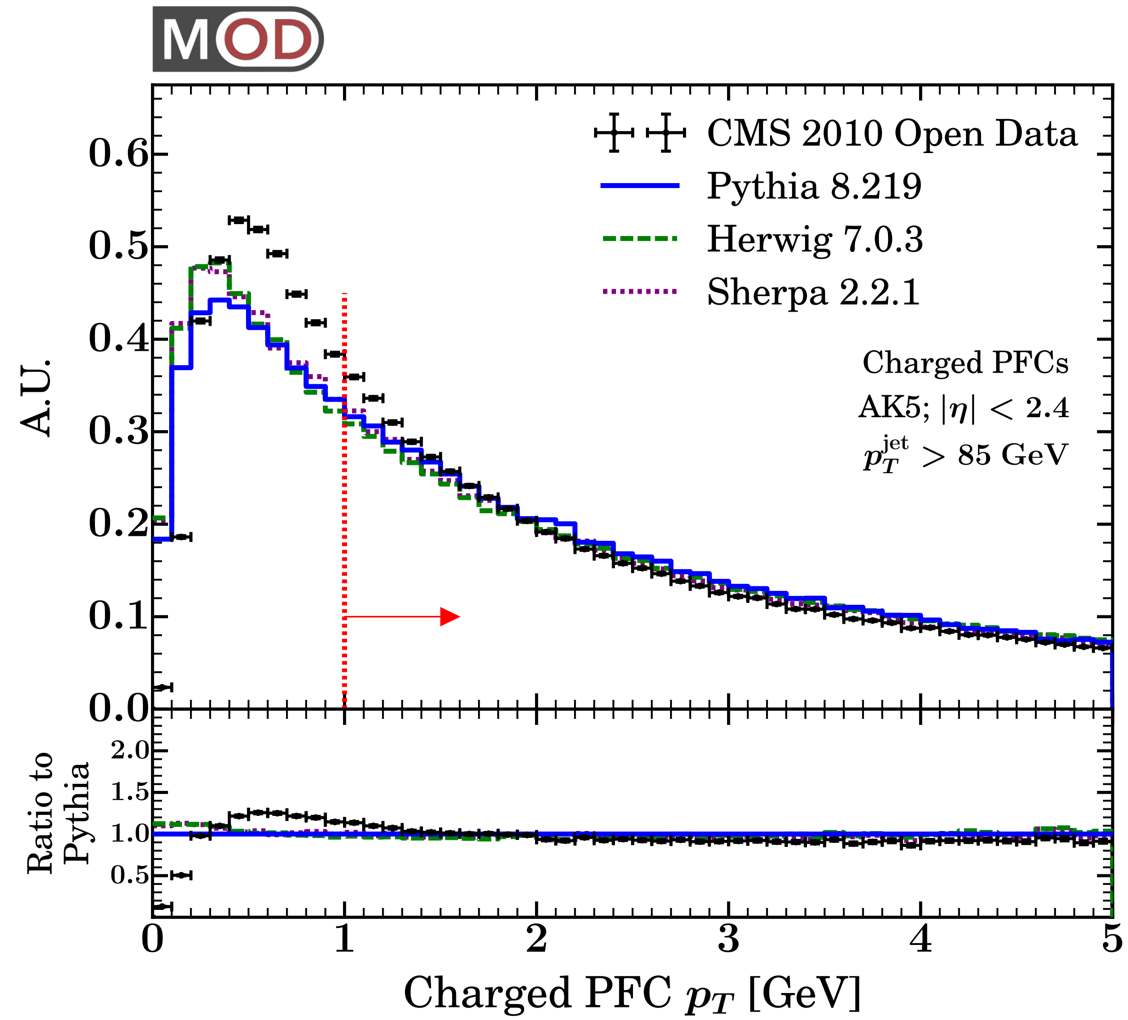
Workhorse of every CMS substructure analysis

Code	Candidate	Total Count	$p_T > 1$ GeV
11	electron (e^-)	32,917	32,900
−11	positron (e^+)	32,984	32,968
13	muon (μ^-)	12,941	12,653
−13	antimuon (μ^+)	13,437	13,110
211	positive hadron (π^+)	6,908,914	5,183,048
−211	negative hadron (π^-)	6,729,328	5,027,146
22	photon (γ)	9,436,530	4,805,173
130	neutral hadron (K_L^0)	2,214,385	1,658,892

Without detector simulation, difficult to assess performance of neutral hadrons (esp. $\pi_0 \to \gamma\gamma$)

Particle Flow Fiducialization

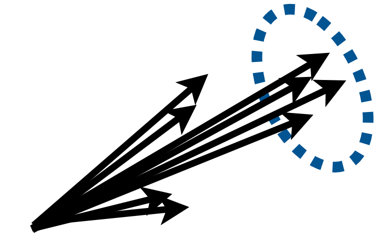


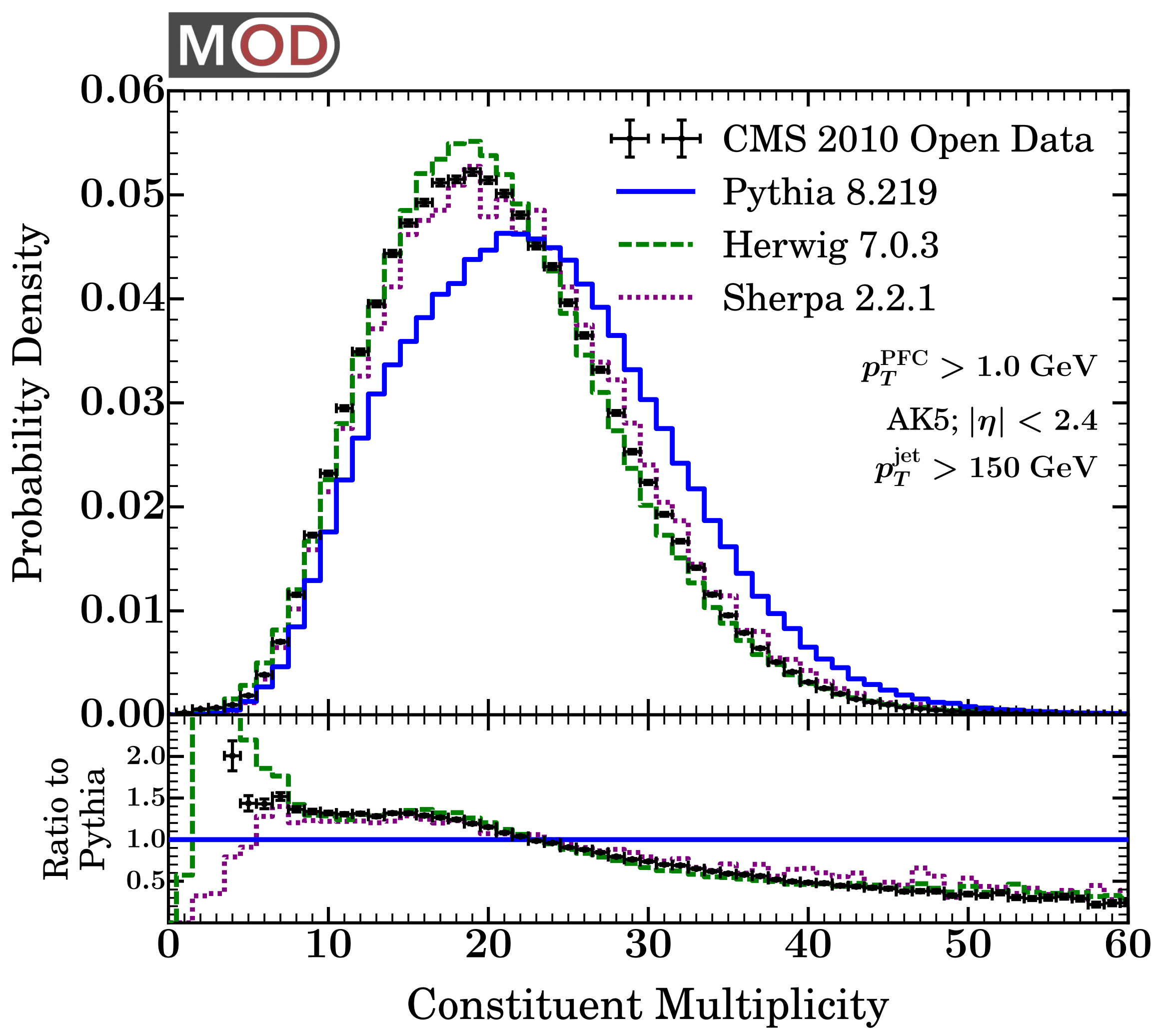


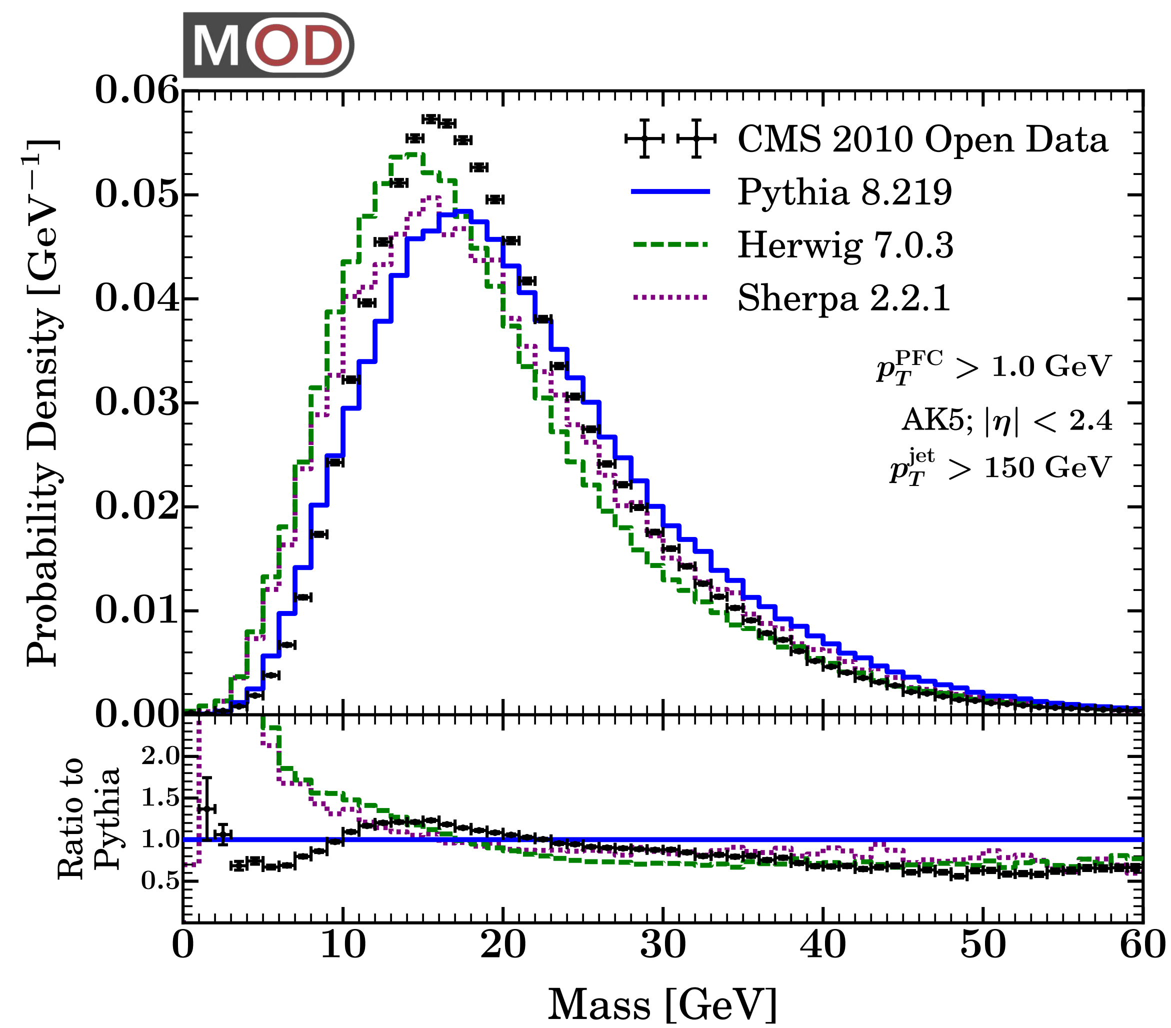
Motivation to focus on charged PFCs with p_T > 1 GeV

Basic Substructure

No grooming applied





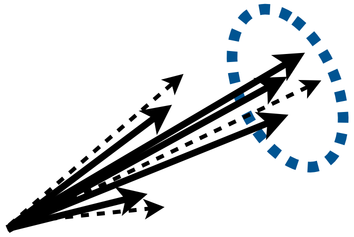


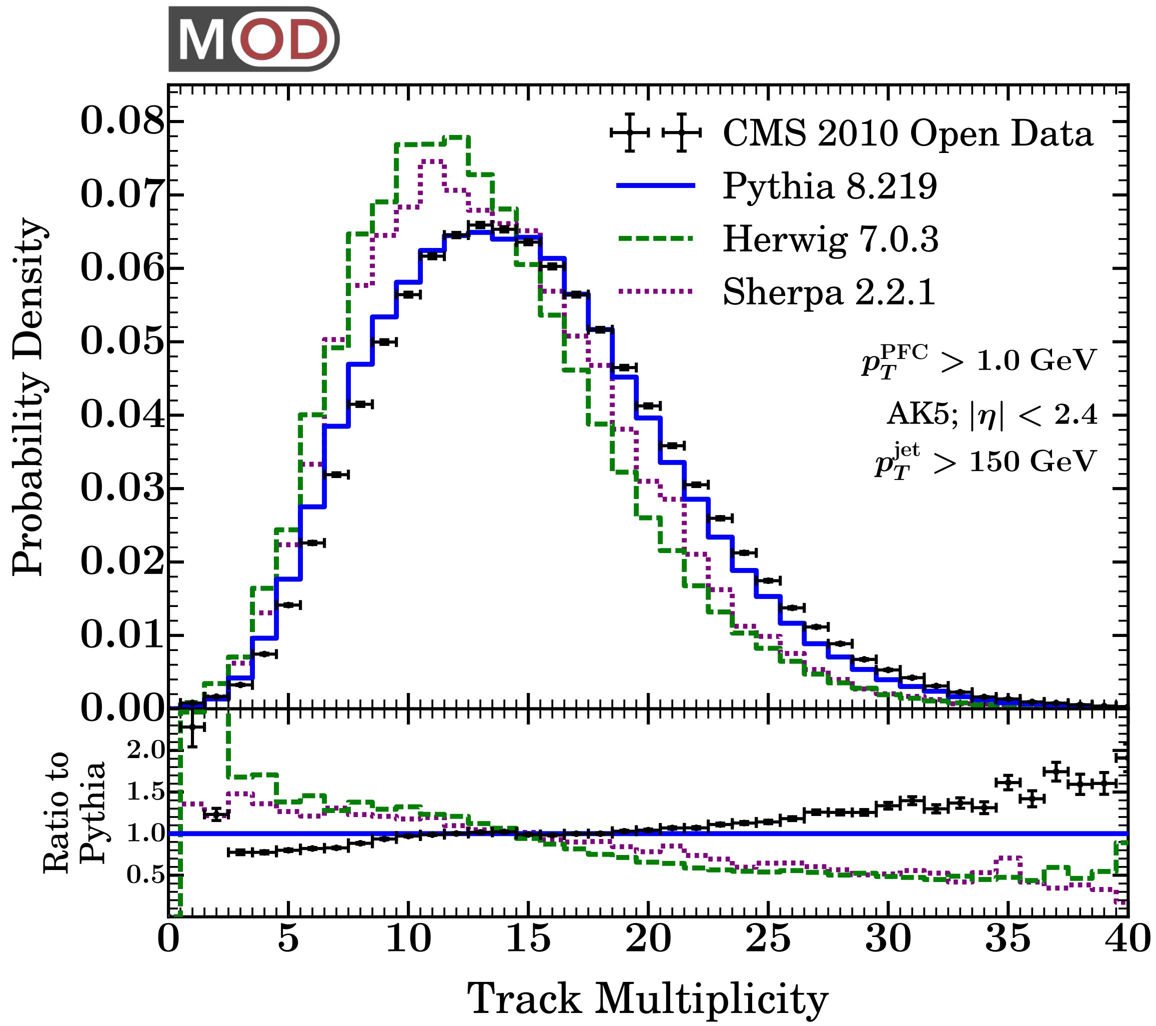
Careful! Can't assess data/MC (dis)agreement without unfolding

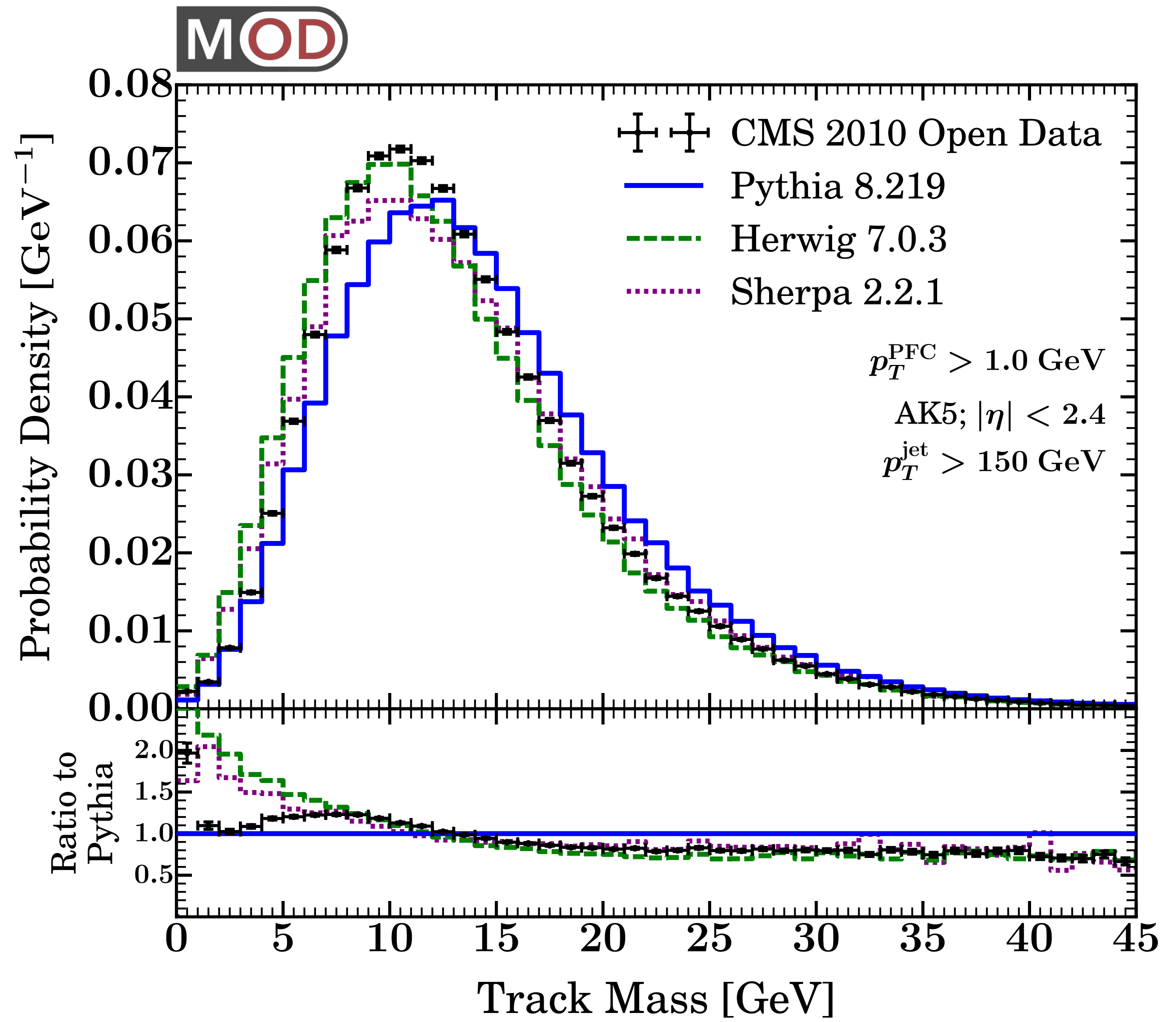
Still interesting to investigate MC/MC differences

Track-Based Substructure

No grooming applied



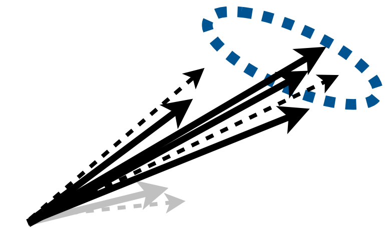


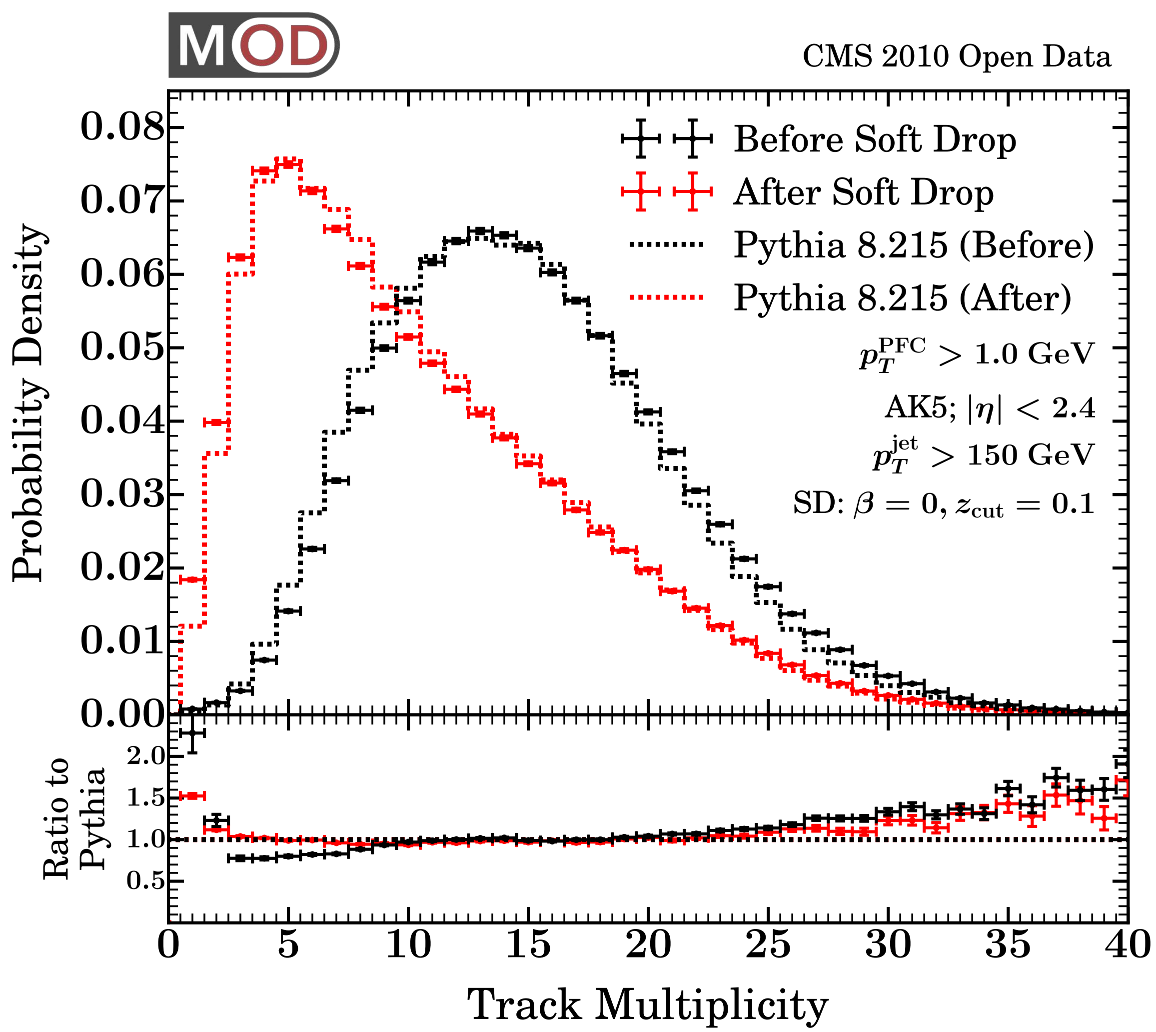


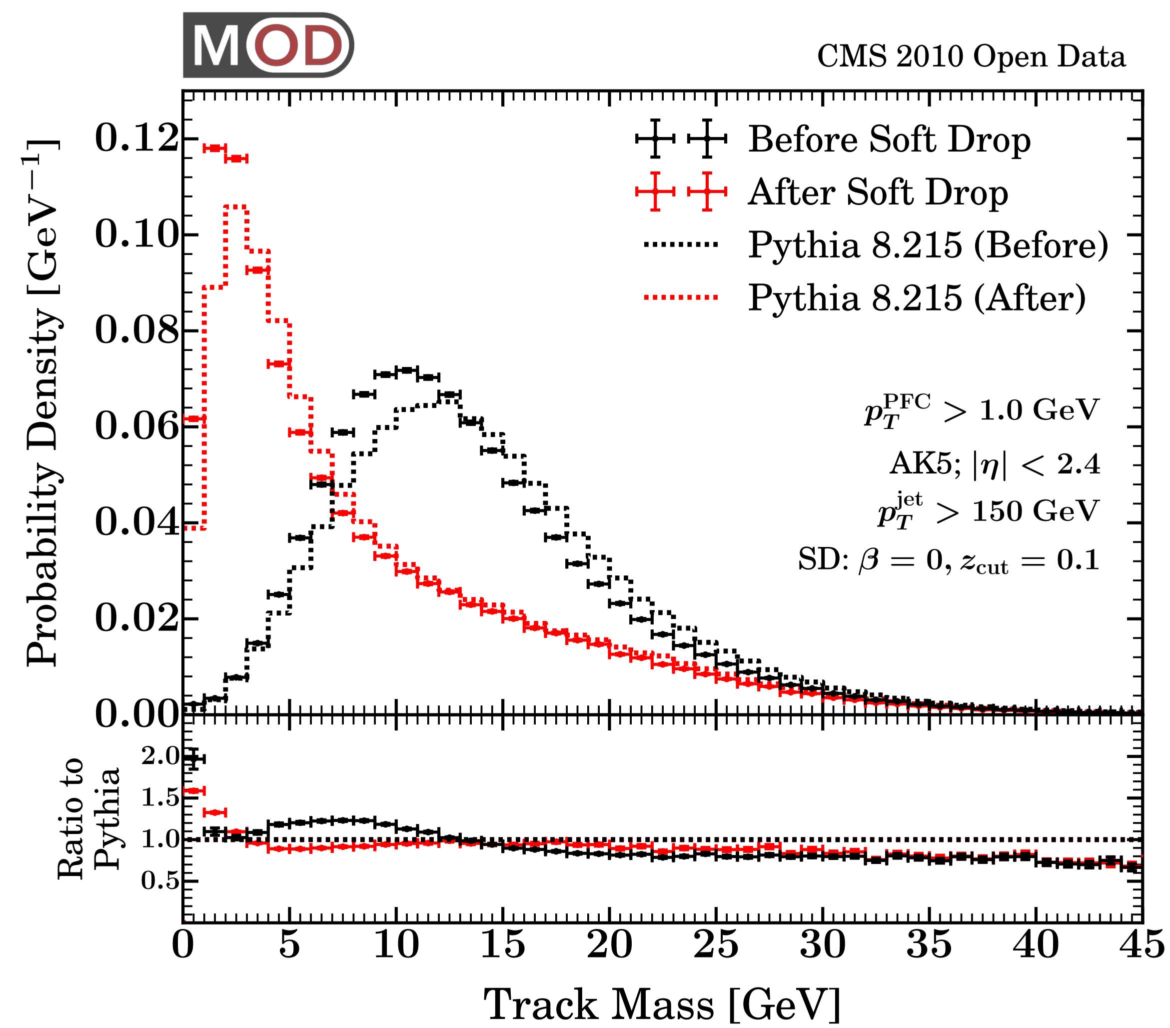
Restricting to charged particles typically improves data/MC agreement (but not always)

Track-Based Substructure

With and without soft drop grooming

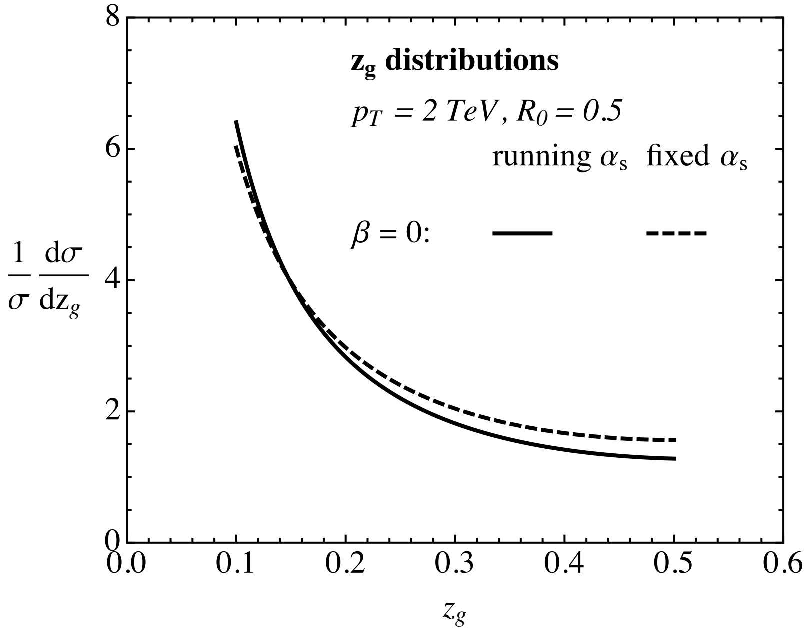




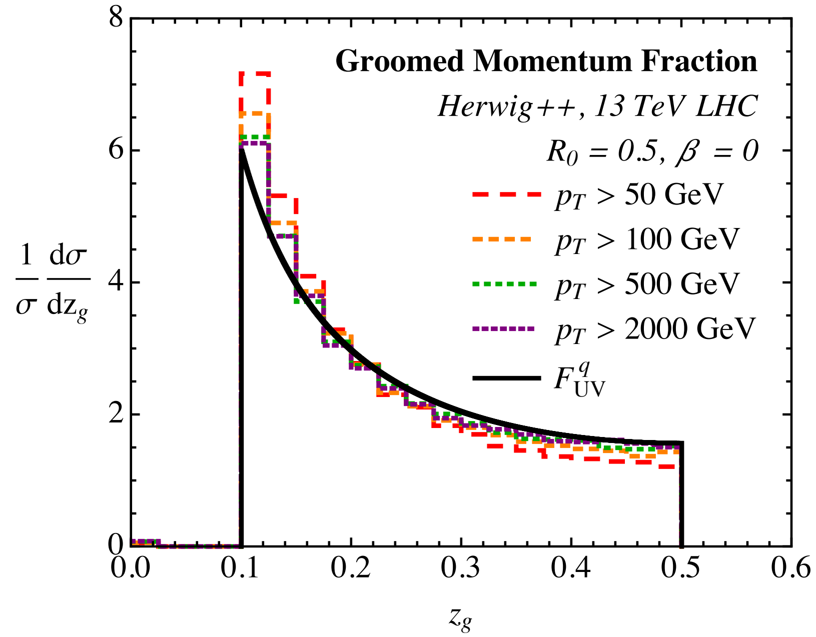


Jet grooming does not typically affect data/MC agreement

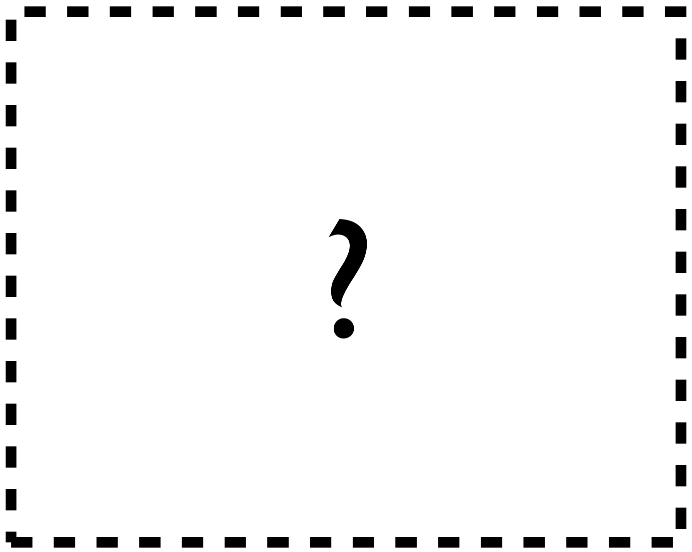
First-principles QCD



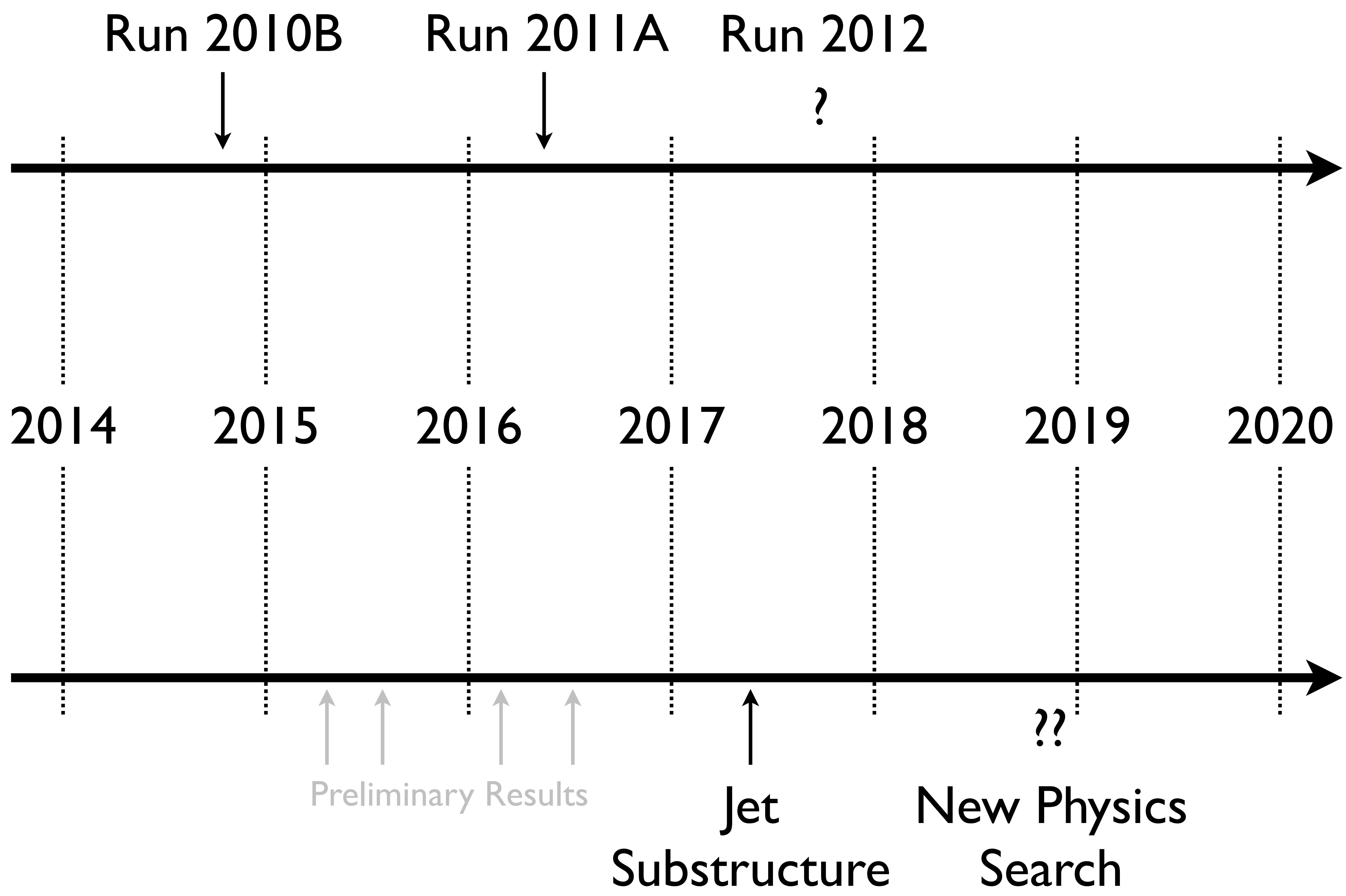
Parton Shower Study



Collider Data



The Open Data Pipeline



CMS Experiment CERN @CMSexperiment · Apr 19

Here's the first-ever physics analysis published using CMS #opendata! arxiv.org/abs/1704.05066 More: opendata.cern.ch/research/CMS #cernopendata

20 21

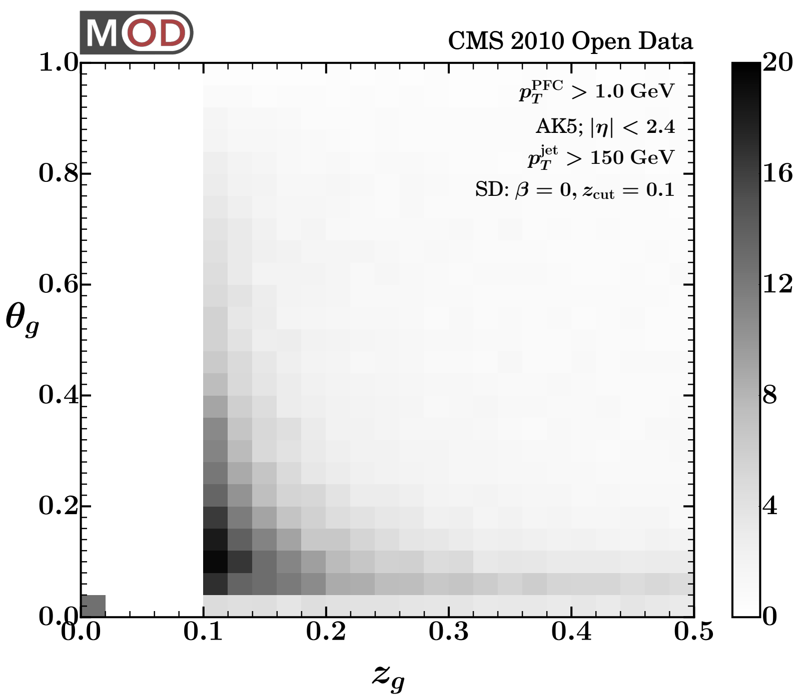
Steven Lowette @StevenLowette · Apr 19

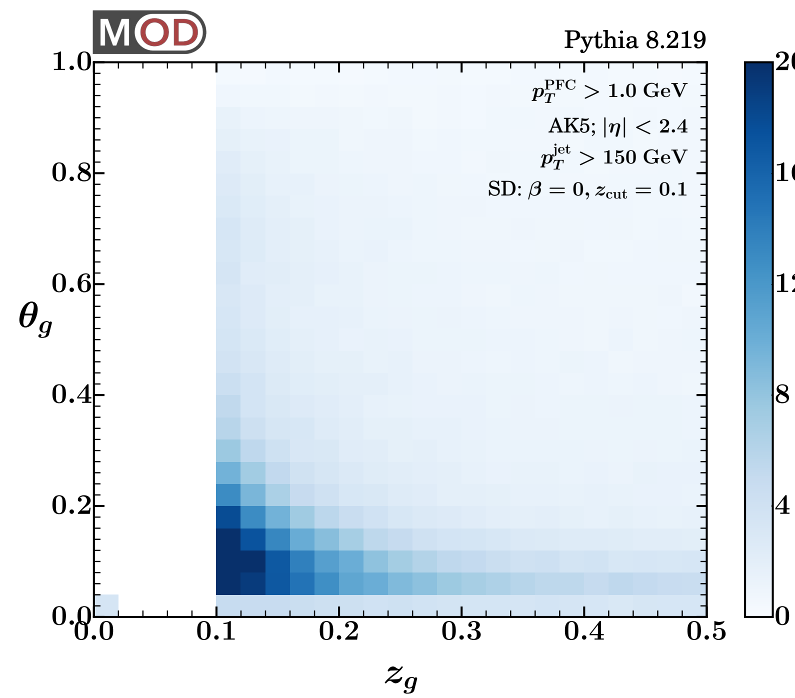
Forget the R(K*) ambulance chasing, this is the interesting paper of the day, using **CMS open data**: arxiv.org/abs/1704.05066

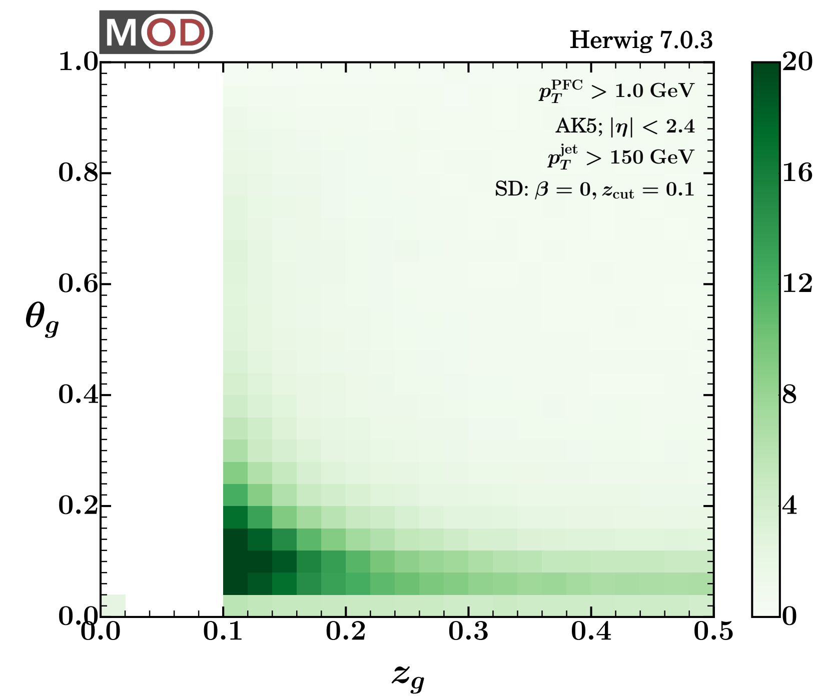
2 4

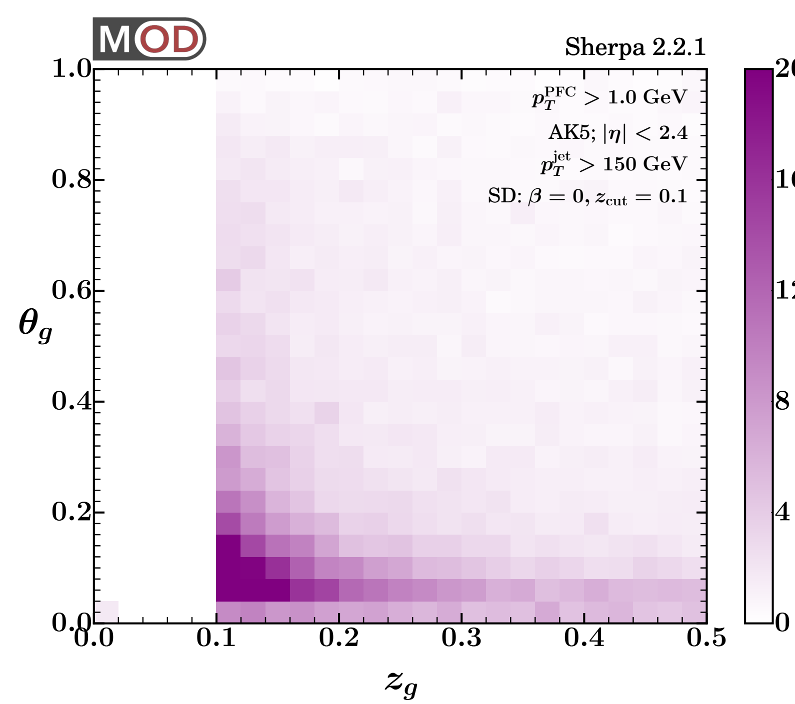
Visualizing the Singularity Structure of QCD

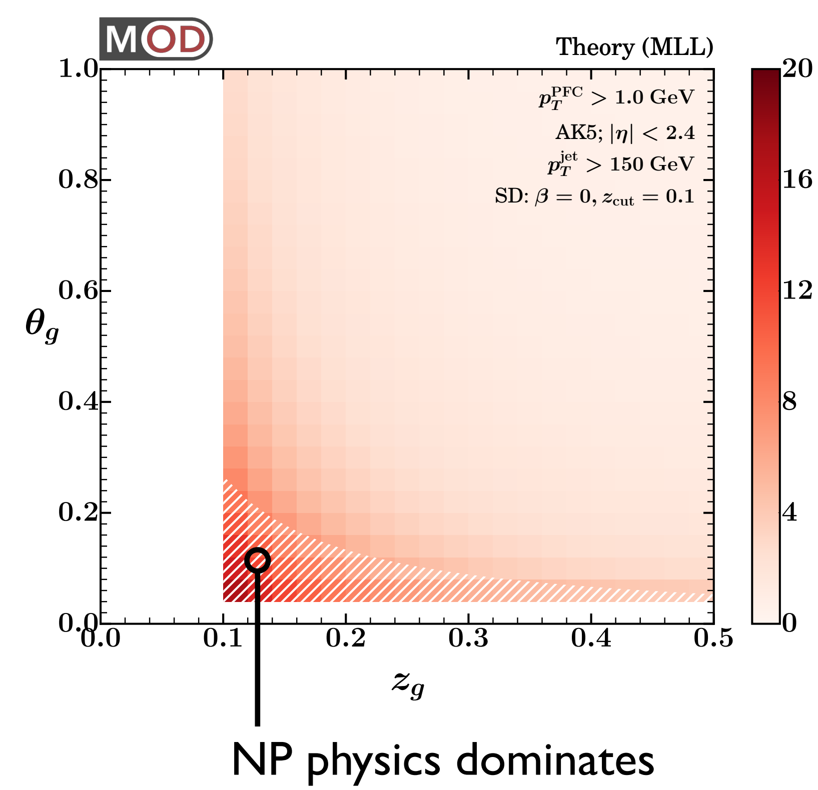
Linear scale







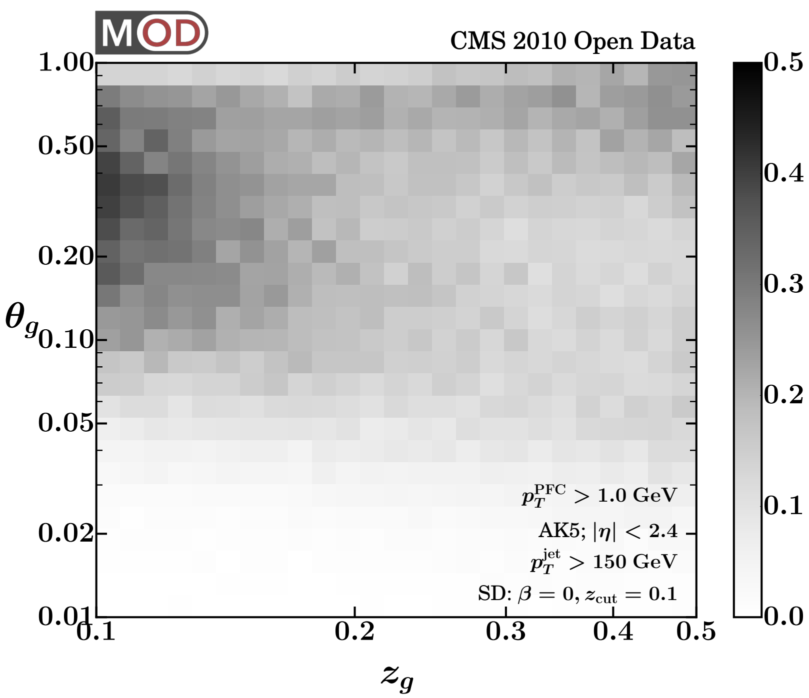


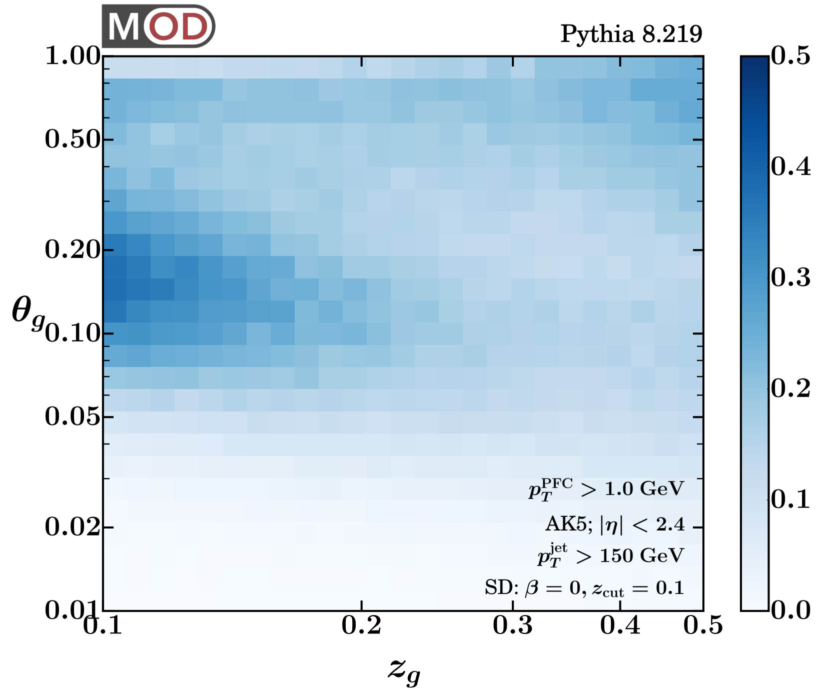


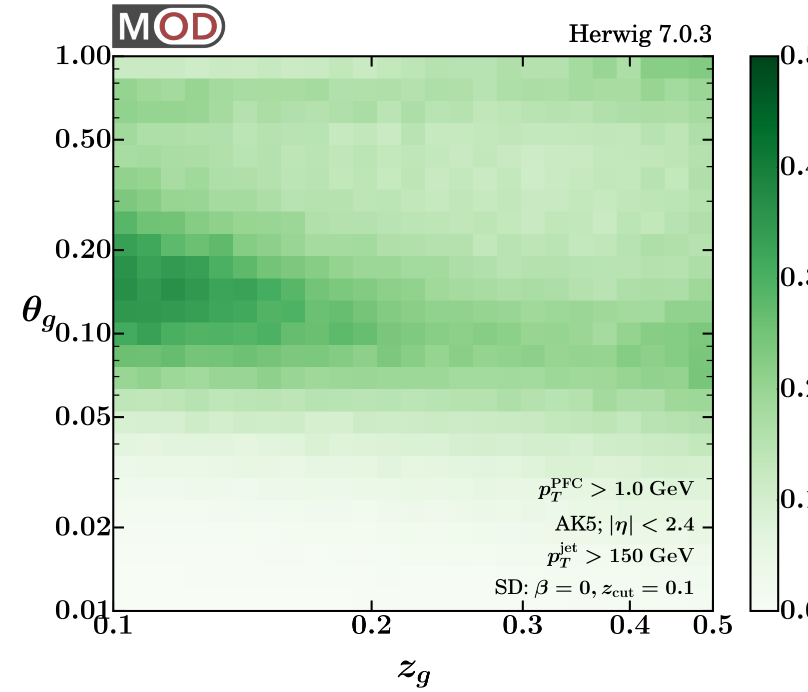
$$\mathrm{d}P_{i\to ig} \simeq \frac{2\alpha_s}{\pi} C_i \frac{\mathrm{d}\theta}{\theta} \frac{\mathrm{d}z}{z}$$

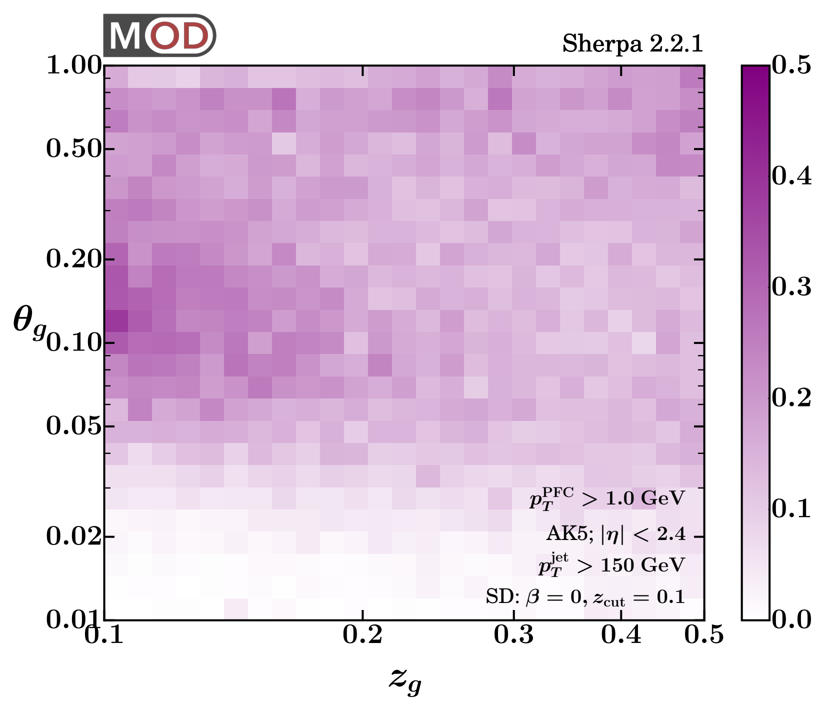
Visualizing the Singularity Structure of QCD

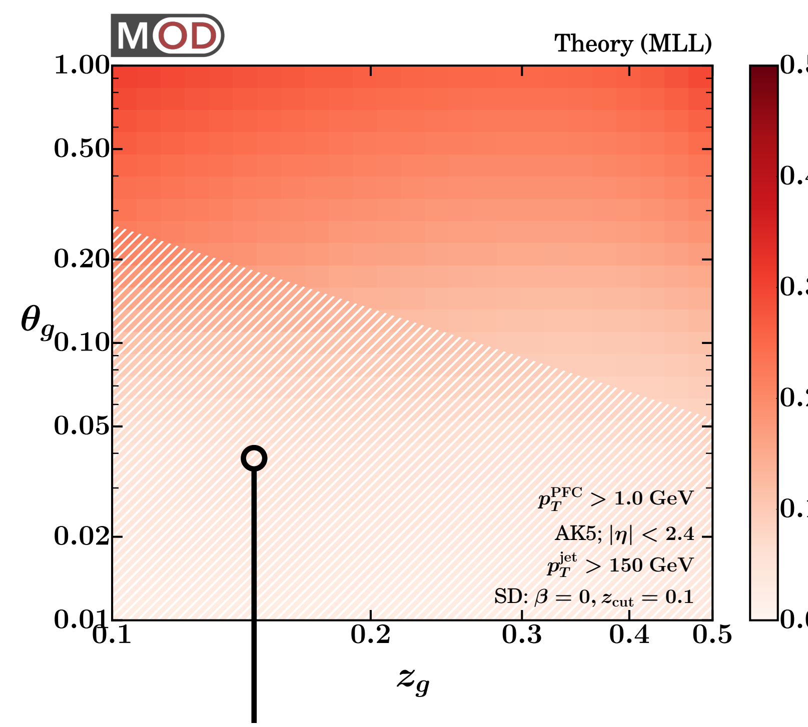
Logarithmic scale









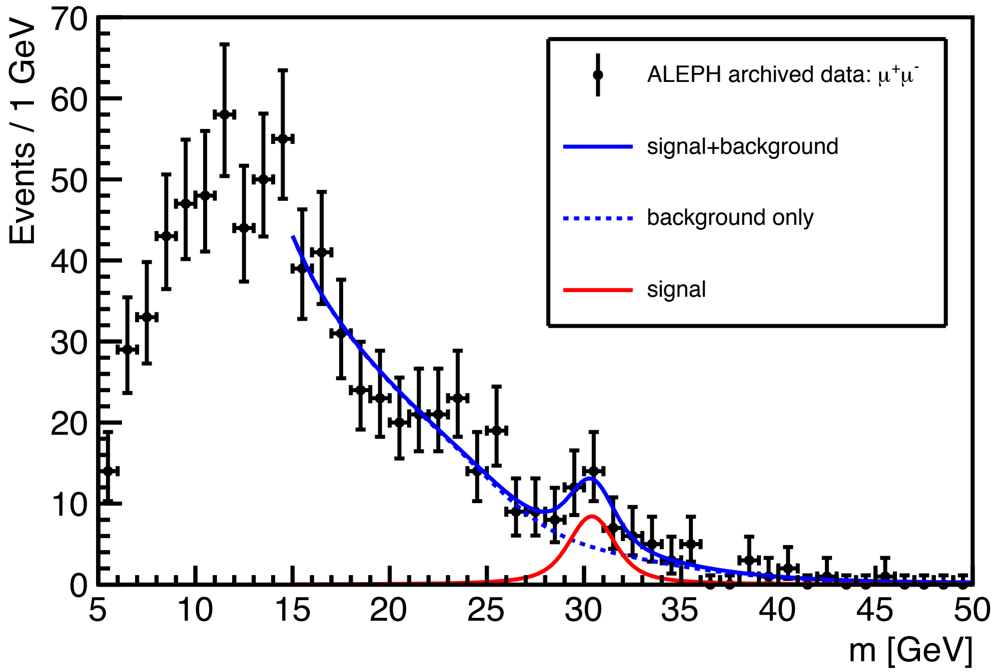


NP physics dominates

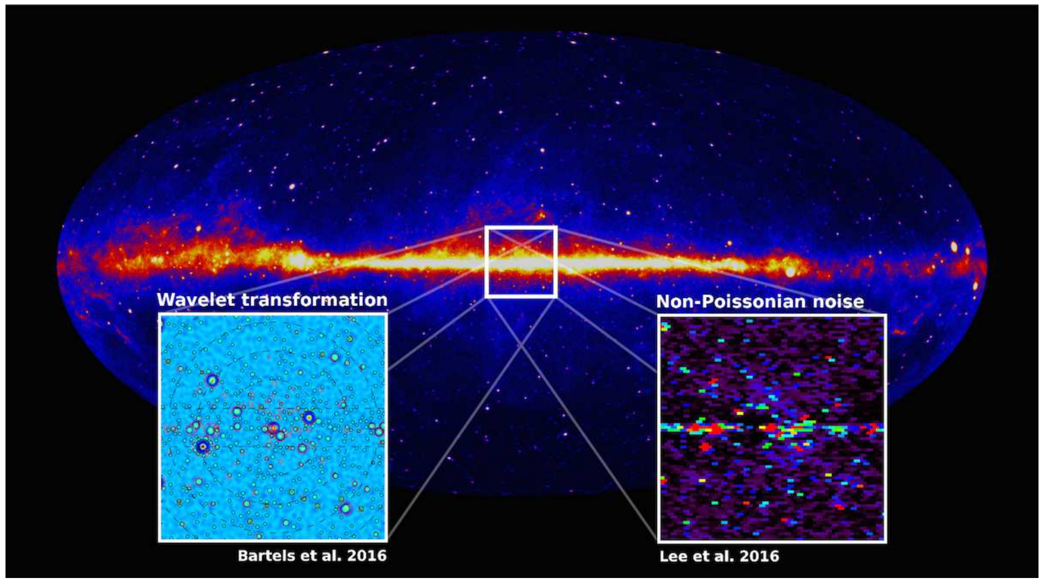
$$\mathrm{d}P_{i\to ig} \simeq \frac{2\alpha_s}{\pi} C_i \frac{\mathrm{d}\theta}{\theta}\frac{\mathrm{d}z}{z}$$

Openness as a Vehicle for Scrutiny

"Thank goodness the CMS Open Data is so hard to use, otherwise there would be countless rogue analyses"



vs.

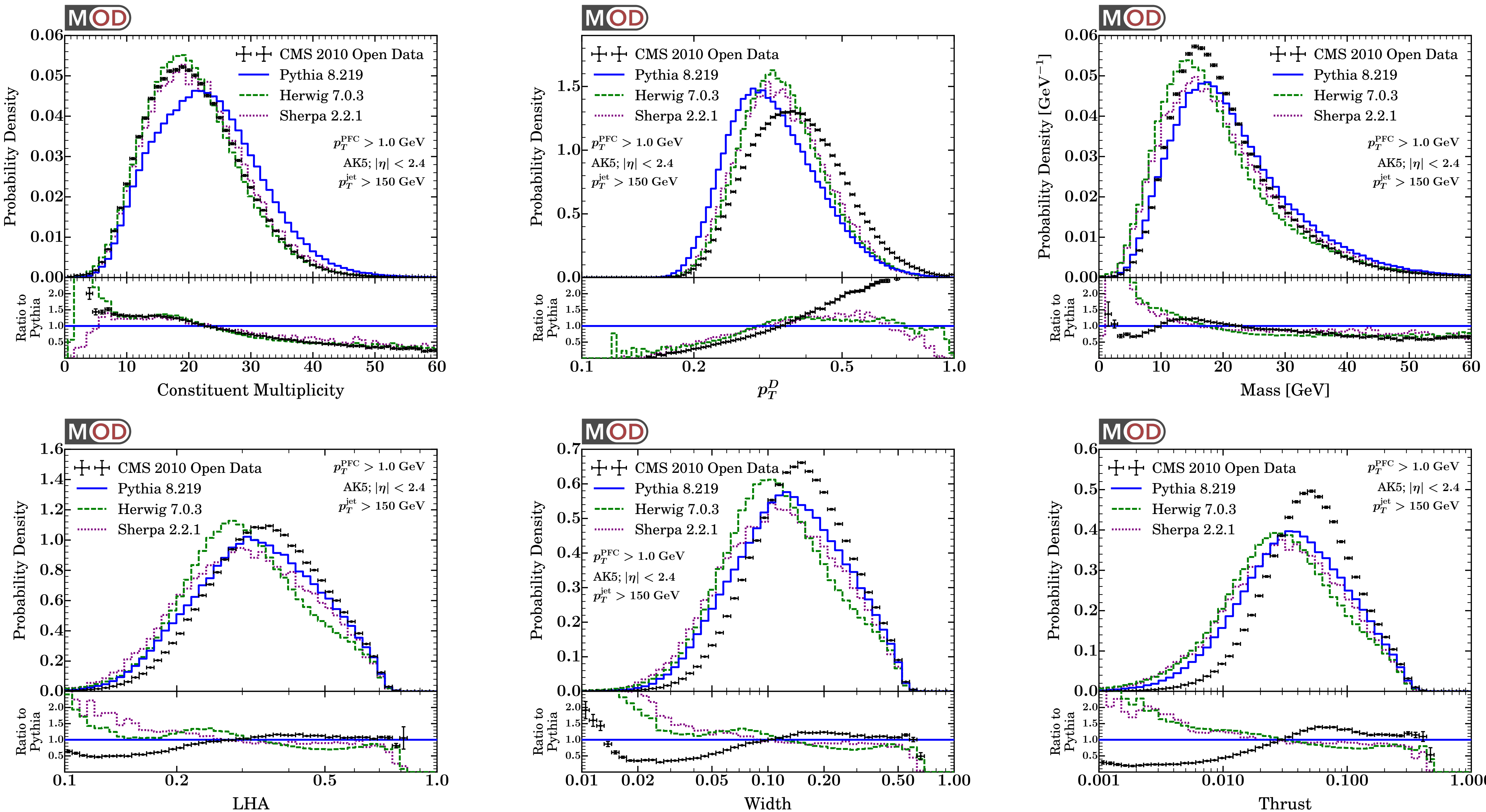


The easier the data is to use,
the more likely it will be used correctly
and the results cross-checked by other groups

[Heister, 1610.06536; Bartels, Krishnamurthy, Weniger, 1506.05104; Lee, Lisanti, Safdi, Slatyer, Xue, 1506.05124]

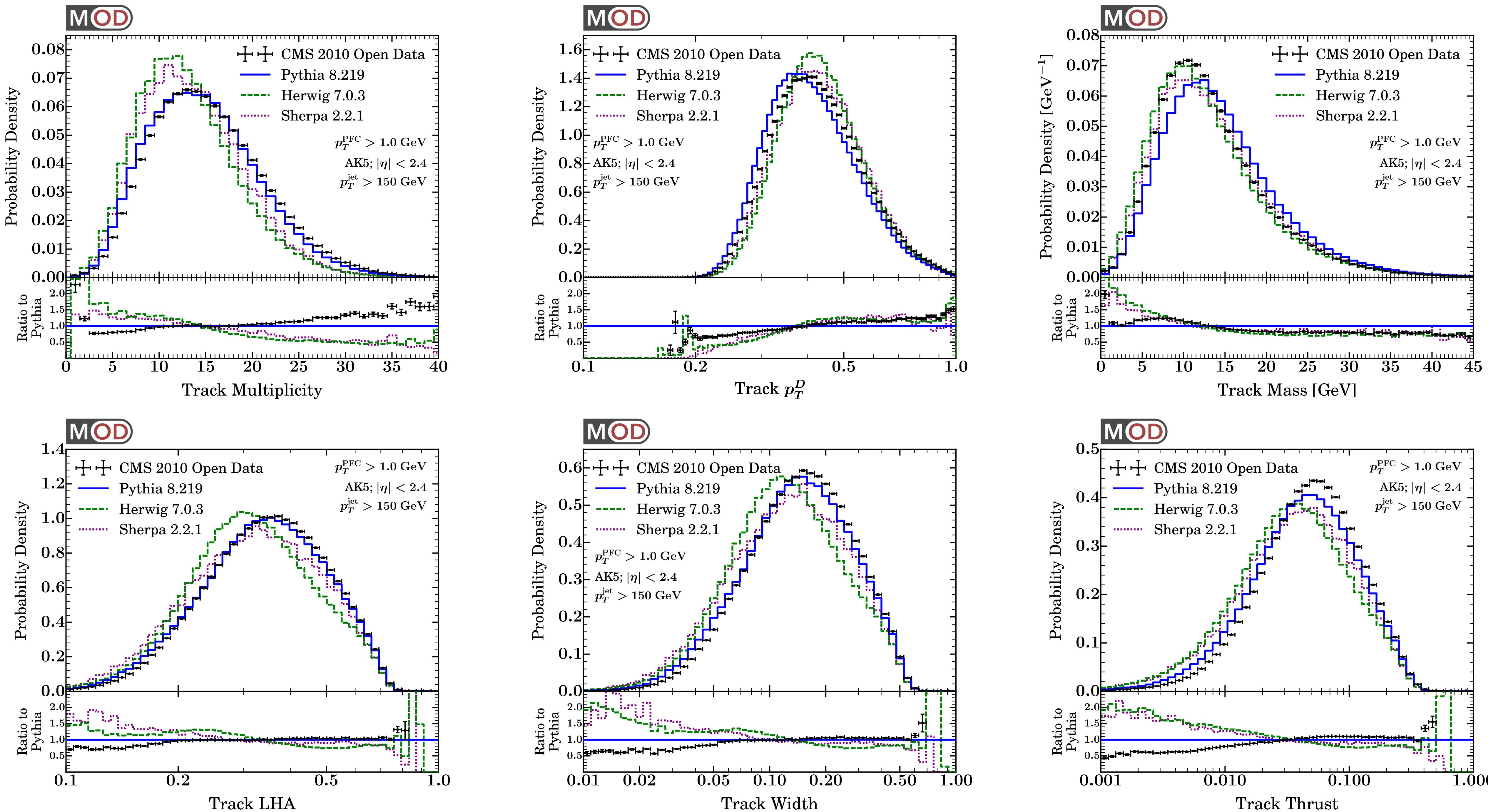
All-Particle Observables

No grooming applied



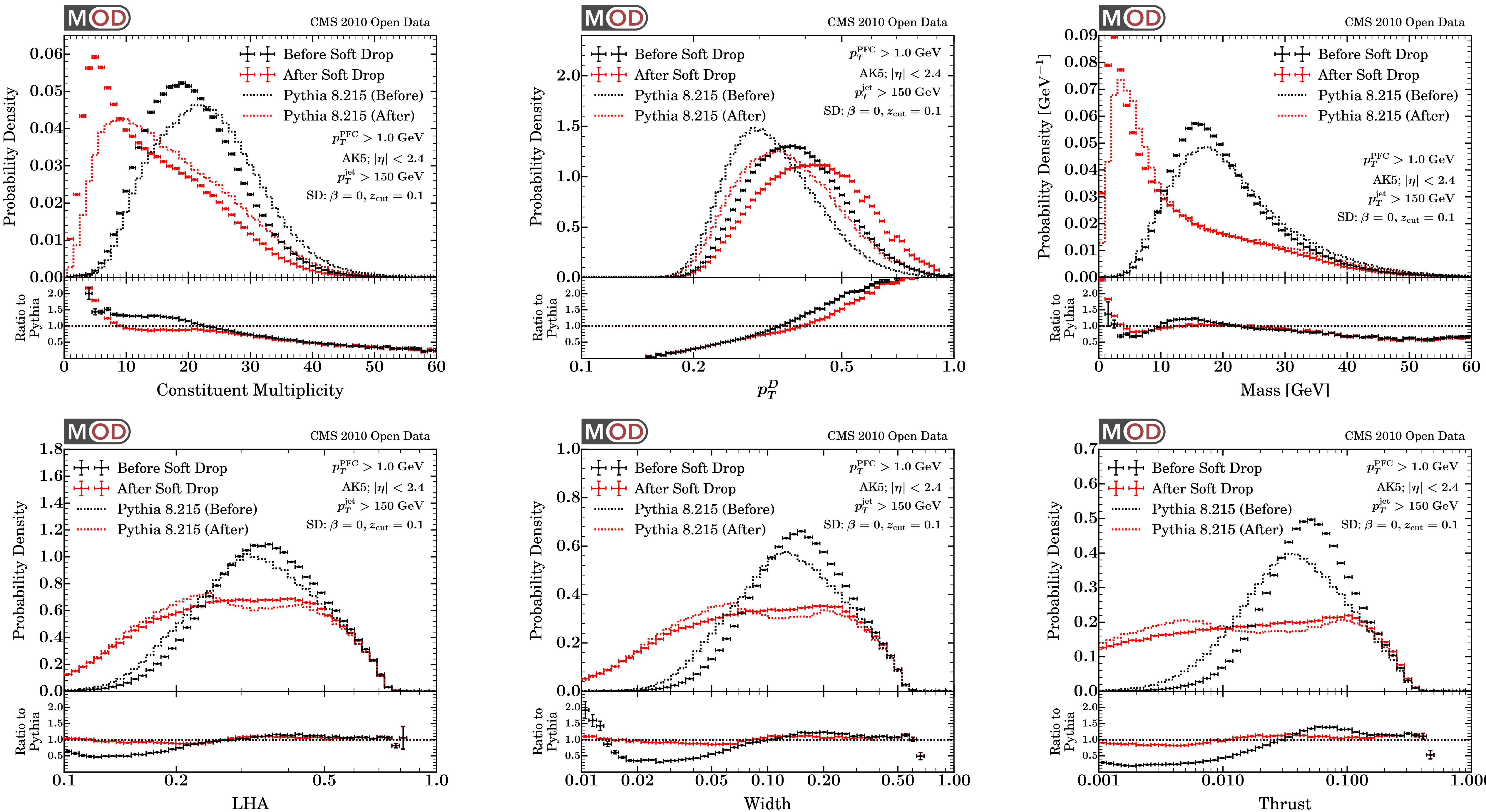
Track-Based Observables

No grooming applied



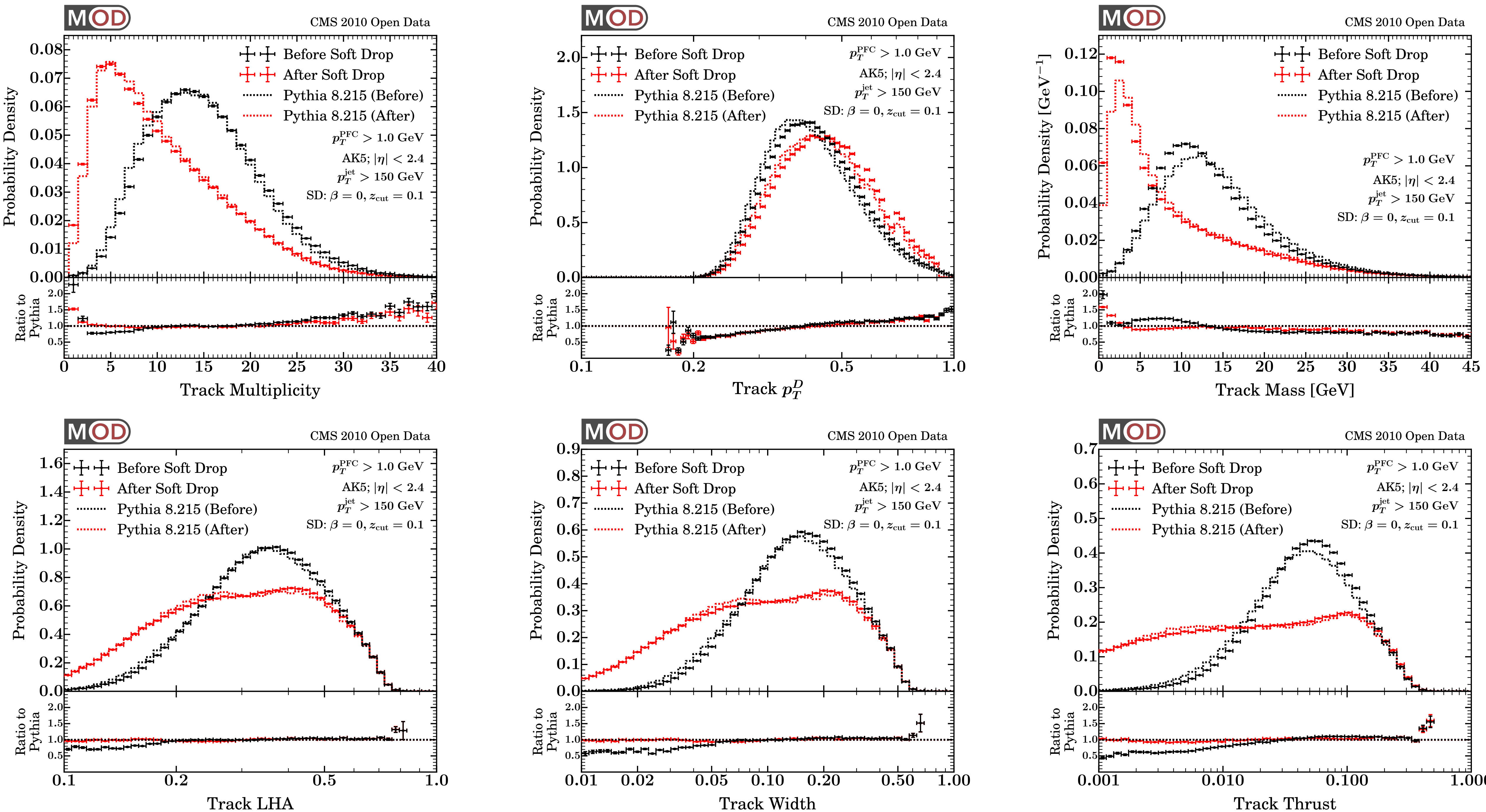
All-Particle Observables

Impact of Grooming



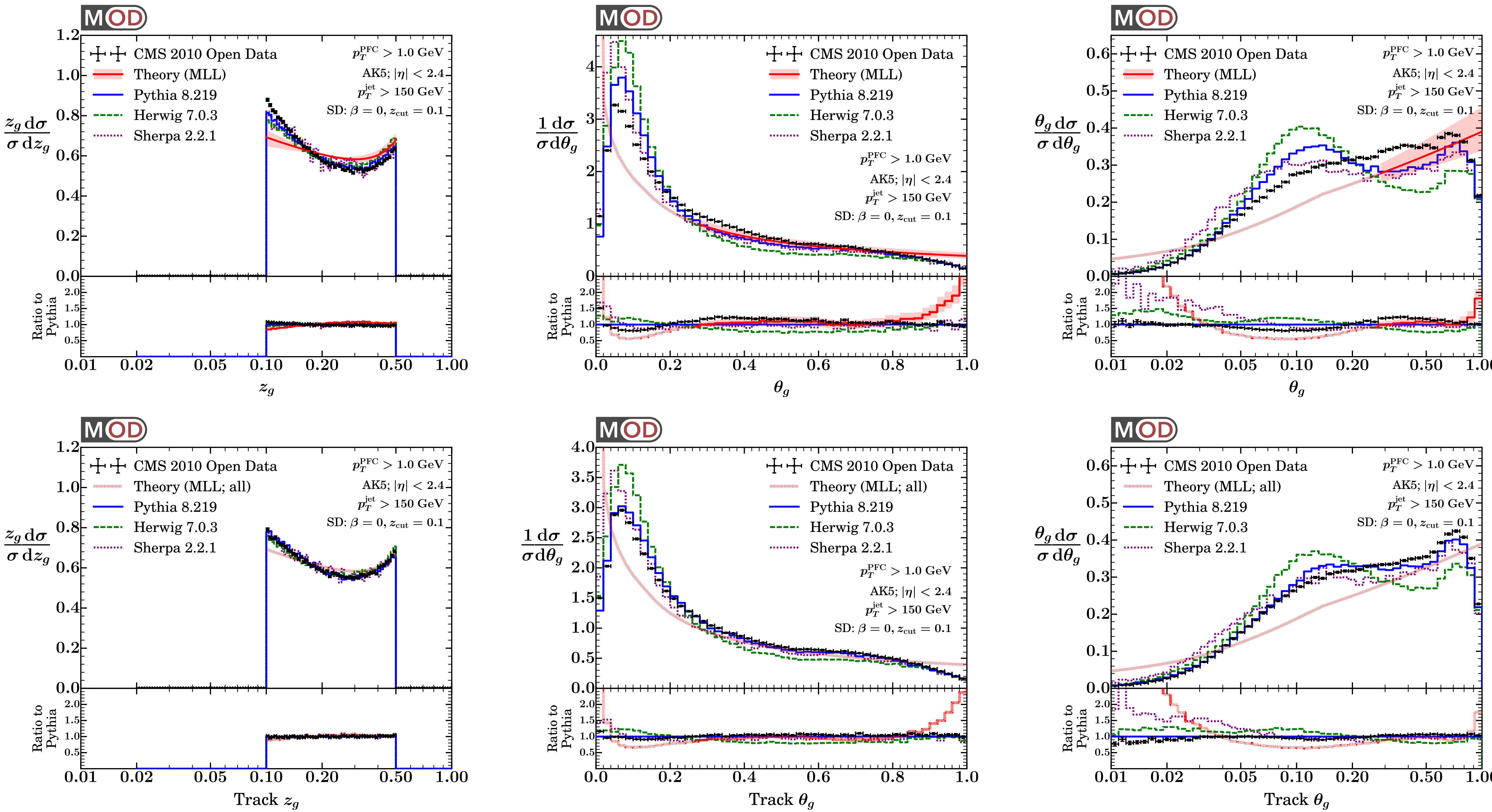
Track-Based Angularities

Impact of grooming



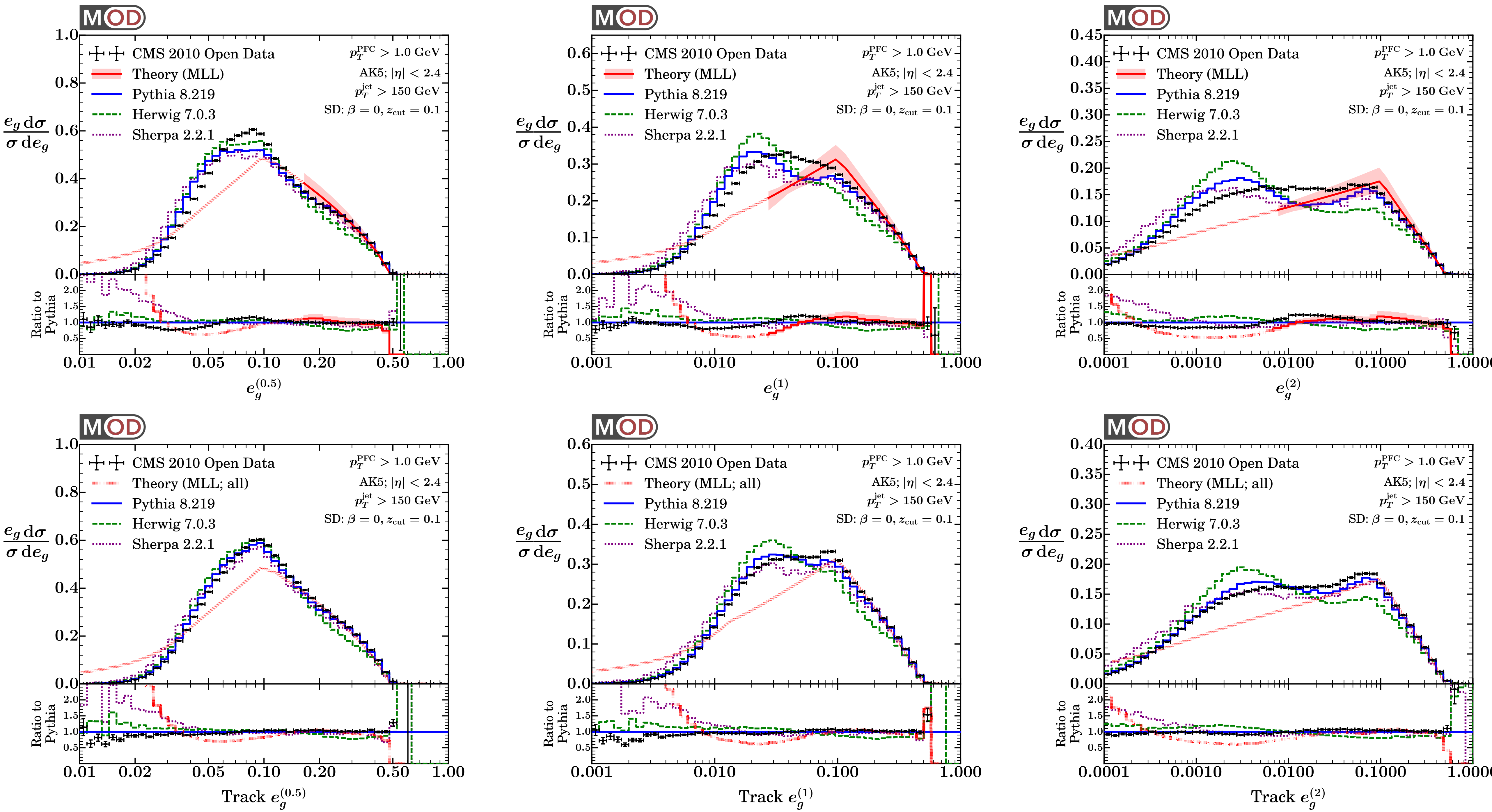
Soft Drop Momentum Balance and Angle

With comparison to MLL calculation



Soft Drop Observables

With comparison to MLL calculation



Miscellaneous Plots

Did I mention each of these has 7 different p_T ranges?