

The Future is Open

Jet Substructure with CMS Public Data

Jesse Thaler

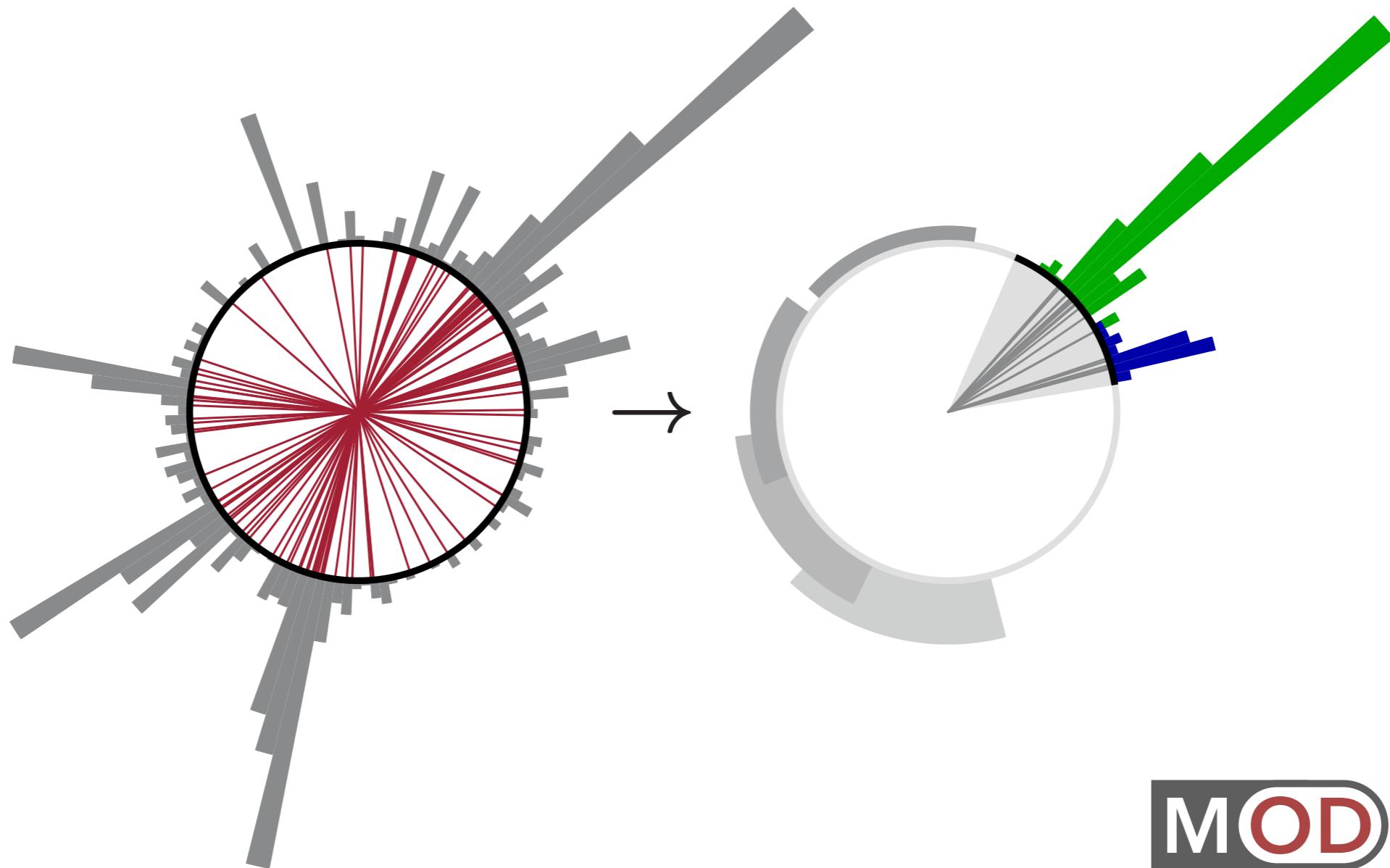


Yale ATLAS Informal Seminar — December 1, 2017



Exposing the QCD Splitting Function with CMS Open Data

Andrew Larkoski,^{1,*} Simone Marzani,^{2,†} Jesse Thaler,^{3,‡} Aashish Tripathee,^{3,§} and Wei Xue^{3,||}





Exposing the QCD Splitting Function with CMS Open Data

Andrew Larkoski,^{1,*} Simone Marzani,^{2,†} Jesse Thaler,^{3,‡} Aashish Tripathee,^{3,§} and Wei Xue^{3,||}

A Milestone for Public Collider Data
A Milestone for Jet Physics
An Opportunity/Challenge for our Community





Jet substructure studies with CMS open data

Aashish Tripathee,^{1,*} Wei Xue,^{1,†} Andrew Larkoski,^{2,‡} Simone Marzani,^{3,§} and Jesse Thaler^{1,||}

V. ADVICE TO THE COMMUNITY

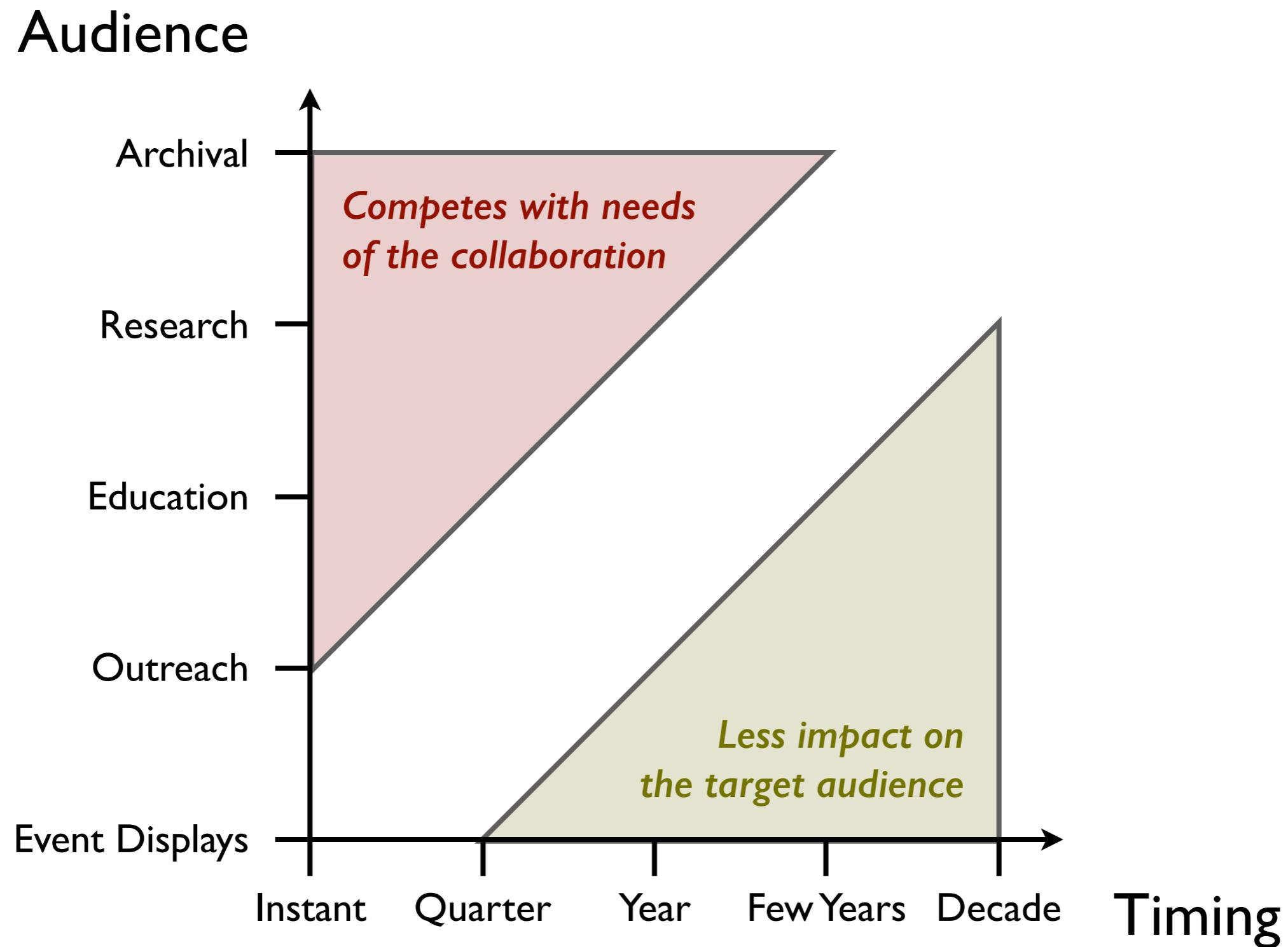
A. Challenges

B. Recommendations

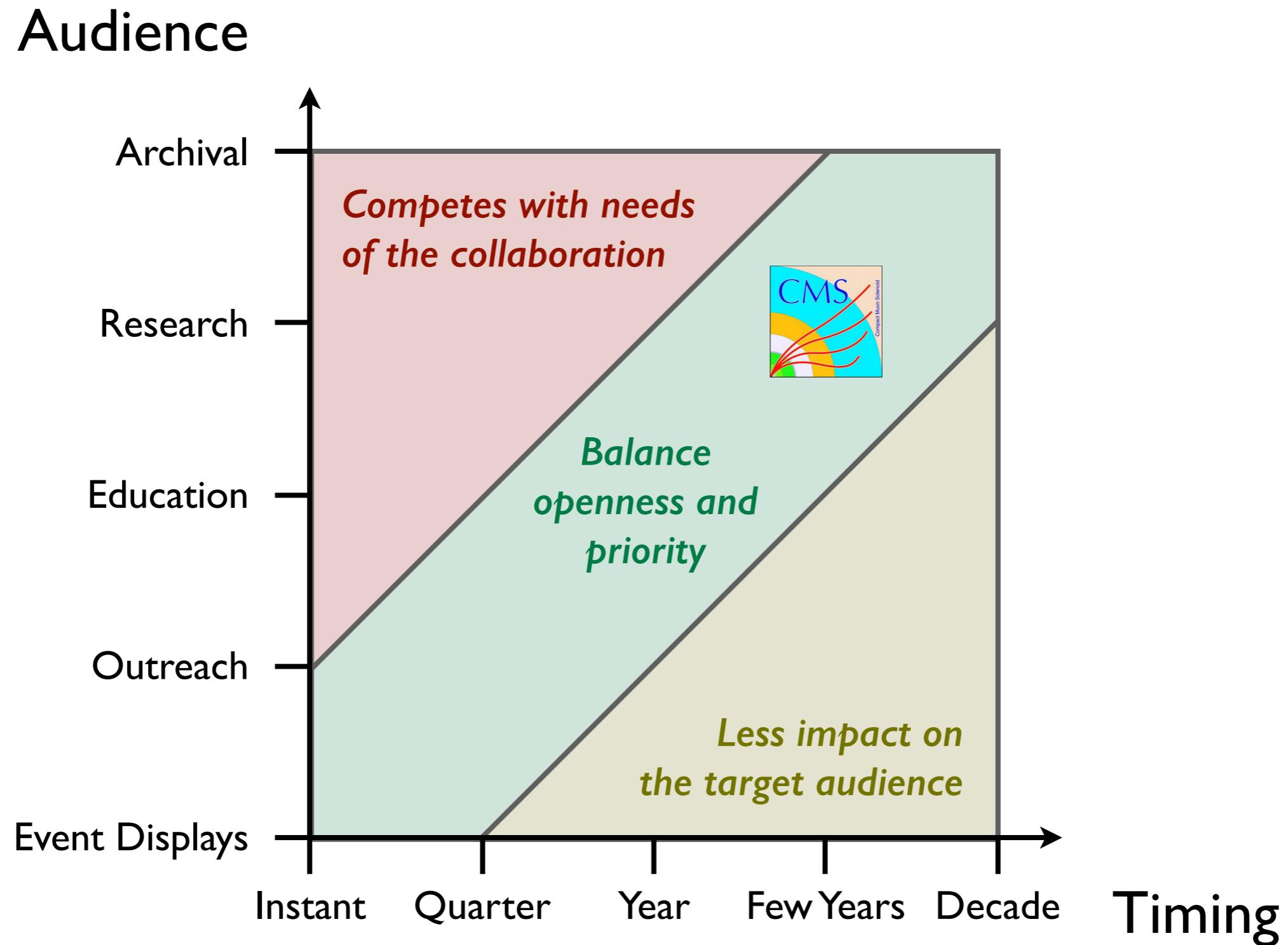
VI. CONCLUSION

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.

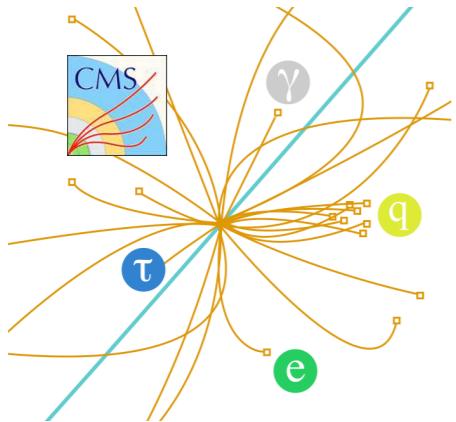
Different Options for “Public Data”



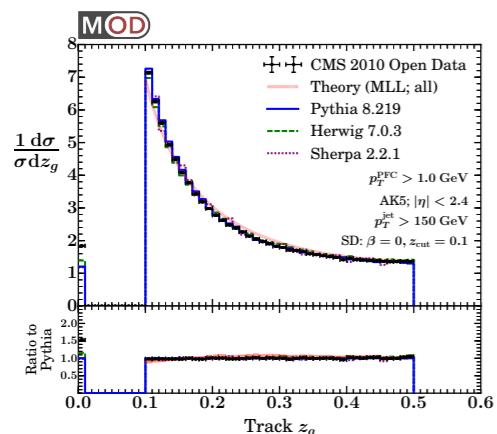
Different Options for “Public Data”



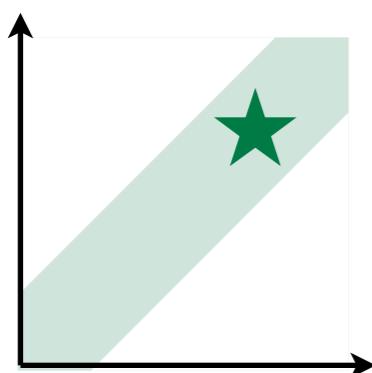
Outline



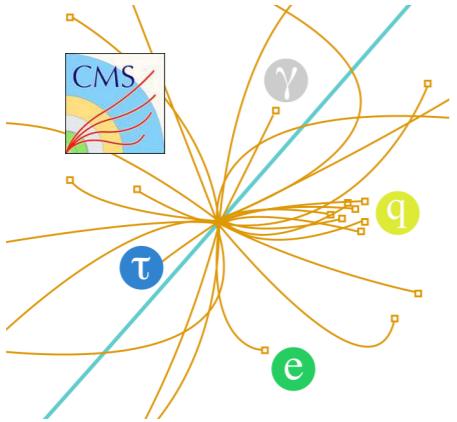
Using the CMS Open Data



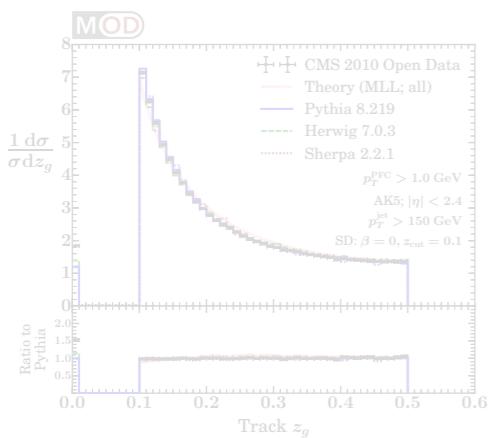
Jet Substructure and QCD Splittings



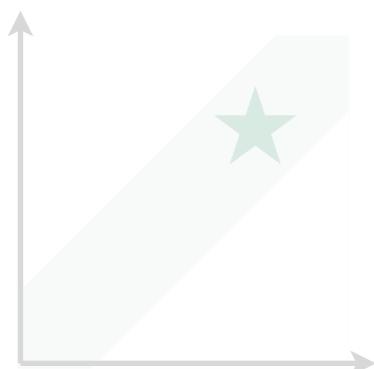
The Future of Public Collider Data



Using the CMS Open Data



Jet Substructure and QCD Splittings



The Future of Public Collider Data



opendata
CERN

opendata.cern.ch/research/CMS

November 2014:

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

$>20 \text{ TB}$, no MC
(Today: QCD)

April 2016:

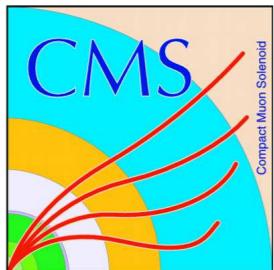
Run 2011A
 $7 \text{ TeV}, 2.5 \text{ fb}^{-1}$

$>100 \text{ TB}$, with MC
(In the Pipeline: BSM)

Translating to “MIT Open Data”

Jet Primary
Dataset

CernVM + CMSSW 4.2.8



AOD Format (CMS Root)

RAW → RECO → “Analysis Object Data”

2.0 TB

20,022,826 events

1664 files

Access via XRootD

MODAnalyzer + FastJet 3.1.3



MOD Format (ASCII + gzip)

Cross-check with flat Root n-tuples

200 GB

20 GB after
baseline selection

Access via External Hard Drive

```

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

```

Luminosity*

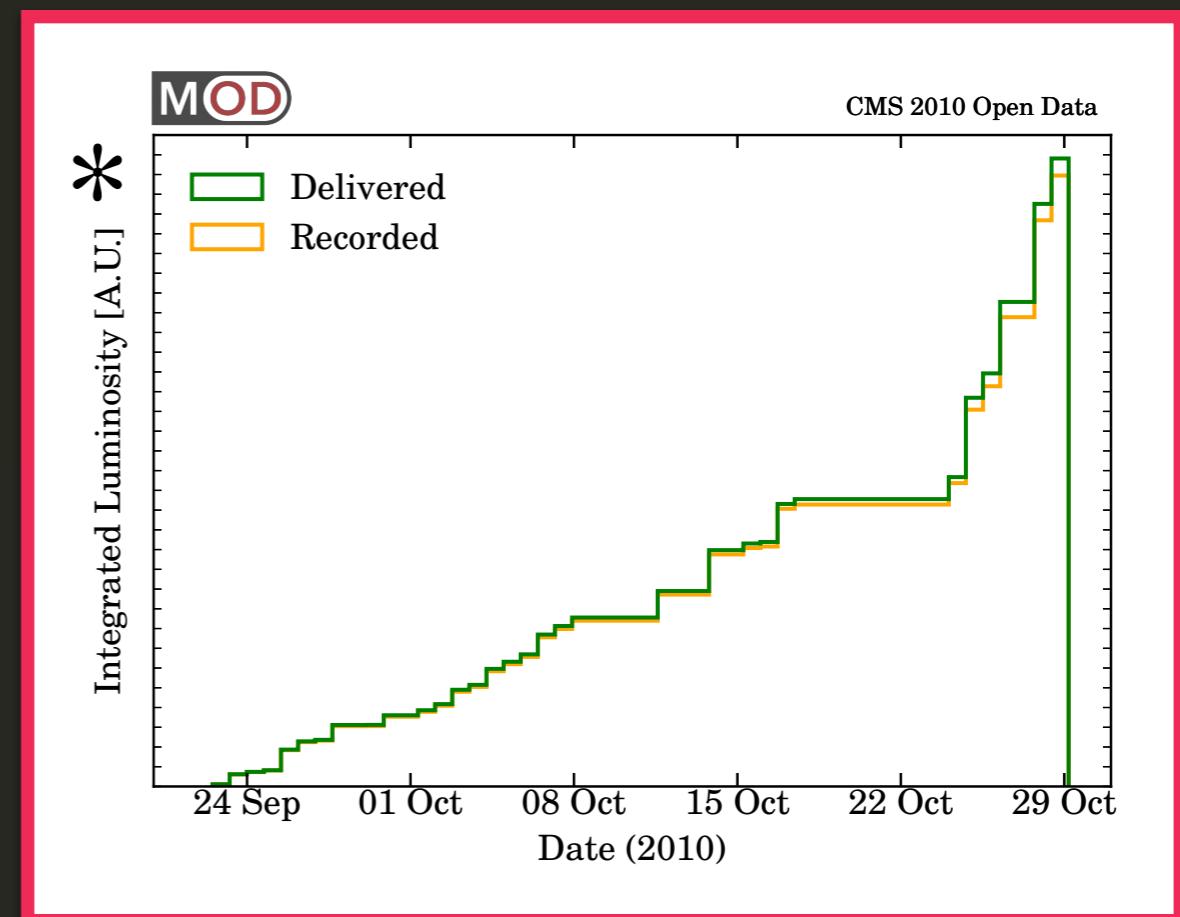


```

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

```

* Not official luminosity table!
Important to stress-test
while collaboration is active



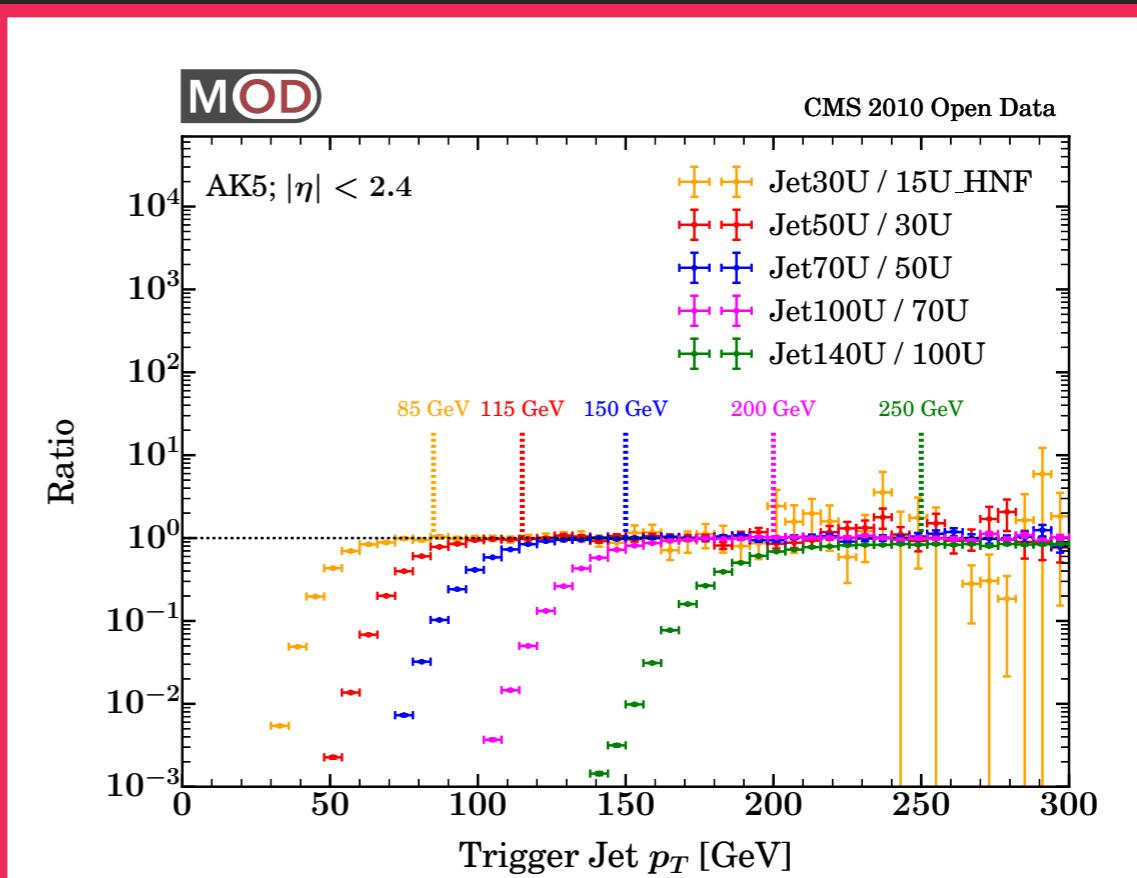
```

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

```

“No magic”
*Have to reproduce most aspects
of official analysis pipeline*

Triggers



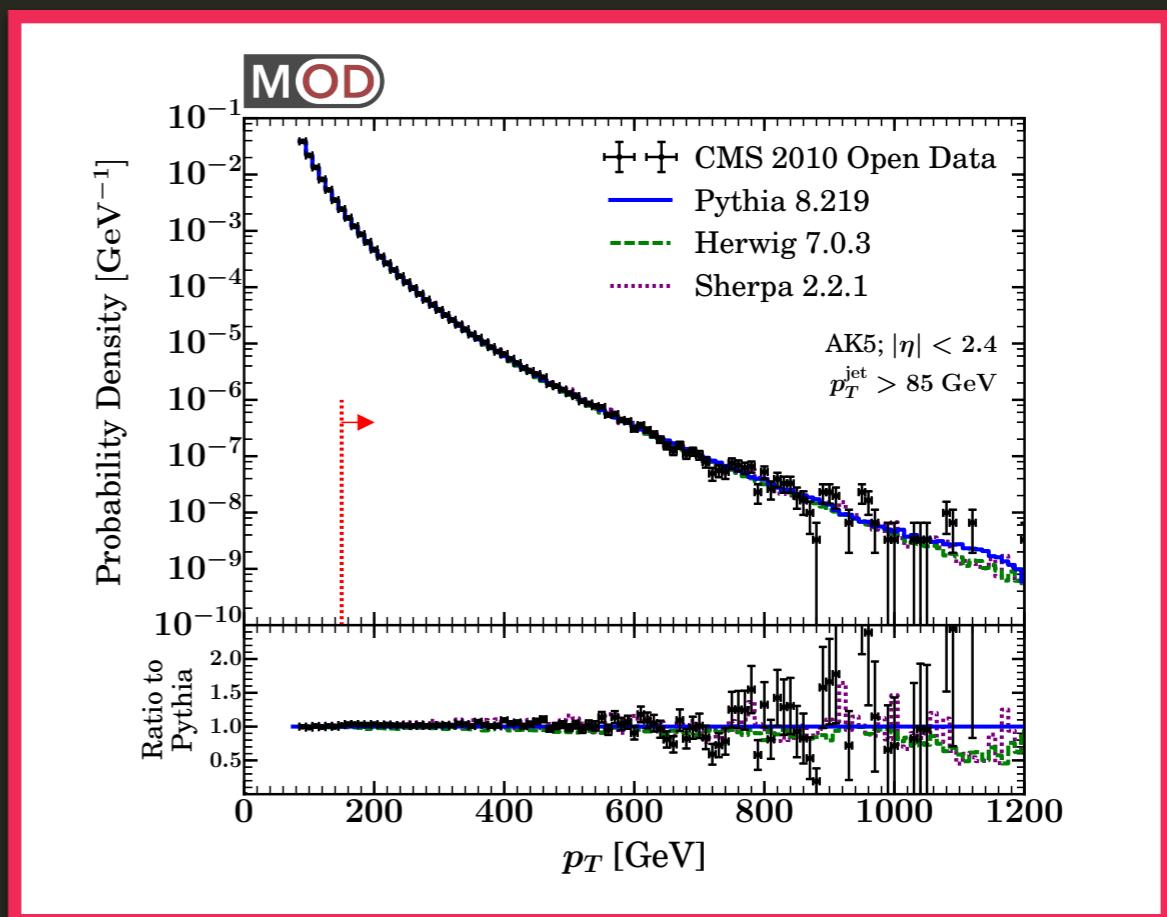
```

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

```

Jets: anti- k_t R = 0.5

Remarkable!
Good data/MC agreement
“out of the box”...

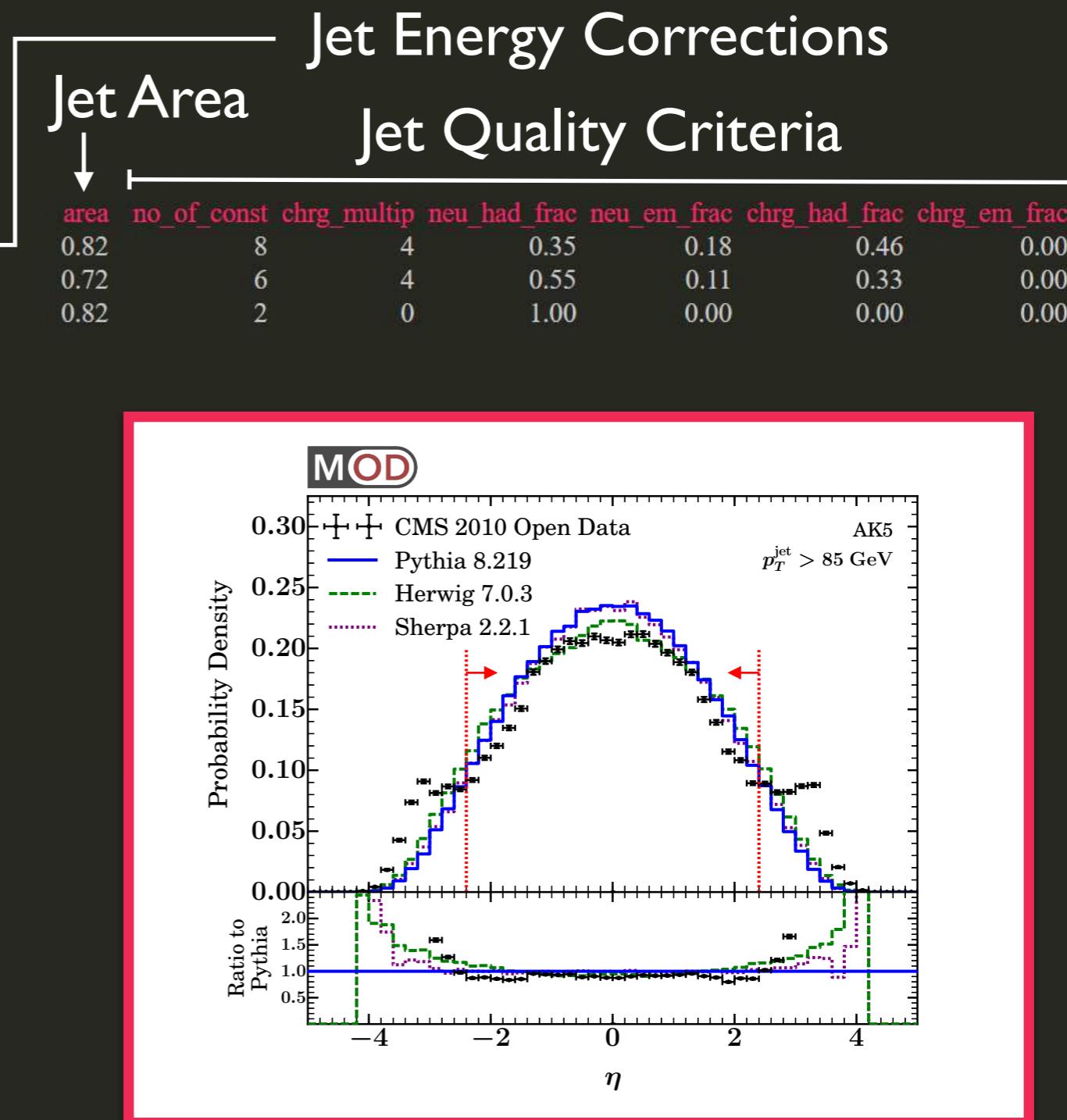


```

1 BeginEvent Version 5 CMS_2010 Jet_Primary_Dataset
2 # Cond RunNum EventNum LumiBlock validLumi intgDelLumi intgRecLumi AvgInstLumi      NPV   timestamp   msOffset
3   Cond    147926 188160899       201        1    21496.19    21208.58      92.03        4 1287023343      516890
4 # Trig          Name Prescale_1 Prescale_2 Fired?
5 Trig HLT_DiJetAve100U_v1           1           1       0
6 Trig HLT_DiJetAve15U             500          10       0
7 Trig HLT_DiJetAve30U             1          500       0
8 Trig HLT_DiJetAve50U             1           65       0
9 Trig HLT_DiJetAve70U_v2           1           25       0
10 Trig HLT_Jet100U_v2              1           1       1
11 # AK5          px     py     pz   energy   jec
12 AK5   9.31   -3.42   27.29   29.21   1.03
13 AK5   6.77    2.40   13.35   15.30   0.99
14 AK5   7.08    0.93  -61.18   61.62   0.93
15 # PFC          px     py     pz   energy   pdgId
16 PFC   -0.95   -0.05   0.65    1.16  -211
17 PFC   -0.75   -0.24   -1.06   1.33  -211
18 PFC   1.27   -1.27  -11.10   11.25  130
19 PFC   -0.00   -0.59   0.50    0.79  211
20 PFC   -0.41   0.54    0.59    0.91  -211
21 PFC   1.55   0.57    5.99    6.22  211
22 PFC   0.12   -0.52   1.36    1.47  -211
23 PFC   0.76   0.36   -1.59   1.81  211
24 PFC   0.43   0.78    2.04   2.23  211
25 PFC   1.90   -0.09   5.88    6.19  130
26 PFC   0.71   1.71    0.94   2.08  211
27 EndEvent

```

*...oh wait, no magic.
Important to consolidate
information on best practices*

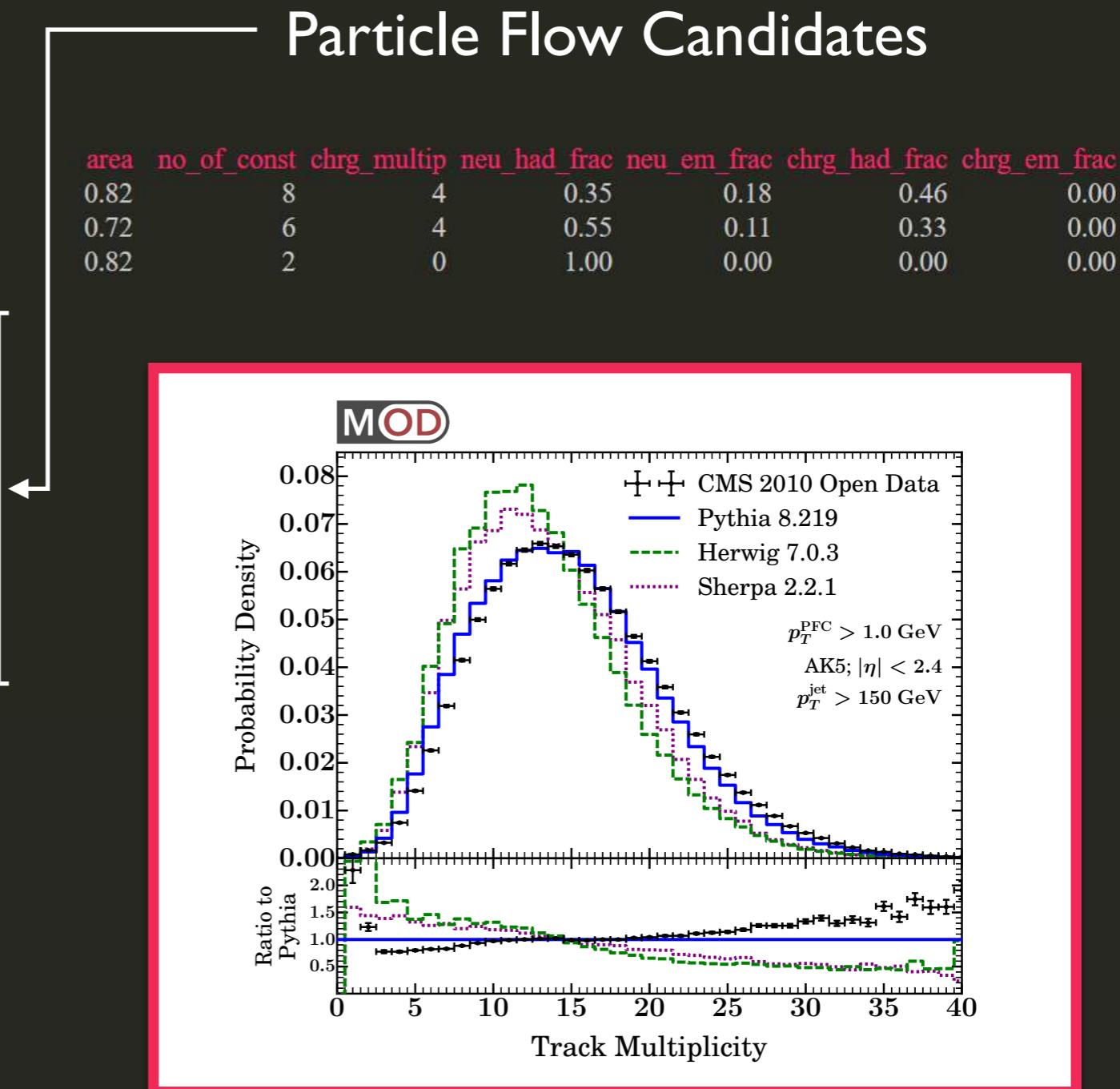


```

1 BeginEvent Version 5 CMS_2010 Jet_Primary_Dataset
2 # Cond RunNum EventNum LumiBlock validLumi intgDelLumi intgRecLumi AvgInstLumi      NPV   timestamp   msOffset
3   Cond    147926 188160899       201        1    21496.19    21208.58      92.03      4 1287023343      516890
4 # Trig          Name Prescale_1 Prescale_2 Fired?
5 Trig HLT_DiJetAve100U_v1           1           1       0
6 Trig HLT_DiJetAve15U             500          10       0
7 Trig HLT_DiJetAve30U             1           500       0
8 Trig HLT_DiJetAve50U             1            65       0
9 Trig HLT_DiJetAve70U_v2           1            25       0
10 Trig HLT_Jet100U_v2              1           1       1
11 # AK5          px     py     pz     energy     jec
12 AK5    9.31   -3.42   27.29   29.21    1.03
13 AK5    6.77    2.40   13.35   15.30    0.99
14 AK5    7.08    0.93  -61.18   61.62    0.93
15 # PFC          px     py     pz     energy     pdgId
16 PFC    -0.95   -0.05   0.65    1.16   -211
17 PFC    -0.75   -0.24   -1.06   1.33   -211
18 PFC    1.27   -1.27  -11.10   11.25   130
19 PFC    -0.00   -0.59   0.50    0.79    211
20 PFC    -0.41   0.54    0.59    0.91   -211
21 PFC    1.55    0.57   5.99    6.22    211
22 PFC    0.12   -0.52   1.36    1.47   -211
23 PFC    0.76    0.36  -1.59    1.81    211
24 PFC    0.43    0.78   2.04    2.23    211
25 PFC    1.90   -0.09   5.88    6.19   130
26 PFC    0.71    1.71   0.94    2.08    211
27 EndEvent

```

Opportunity: Novel analyses
Challenge: Calibration



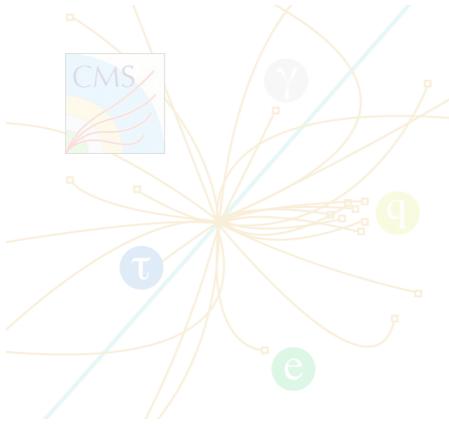
For this study:

Detector-object data (with statistical errors only)
overlaid on

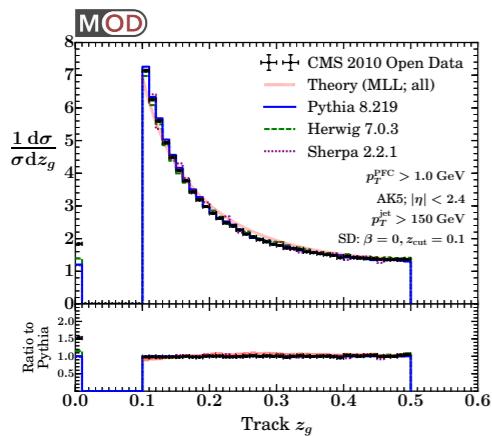
Truth-hadron parton shower generators (no simulation)

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

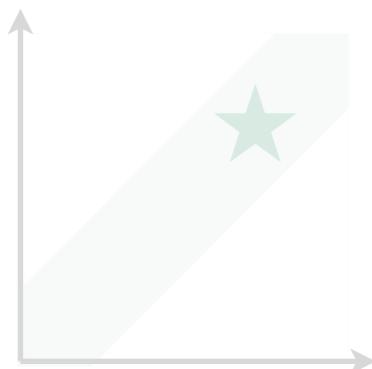
*Run 2011A does include MC,
allowing an estimate of systematic uncertainties*



Using the CMS Open Data



Jet Substructure and QCD Splittings



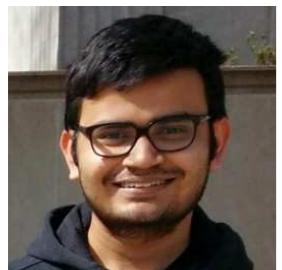
The Future of Public Collider Data



opendata
CERN

Kati Lassila-Perini,
Achim Geiser, ...

MOD



Aashish Tripathee



Wei Xue



Andrew Larkoski



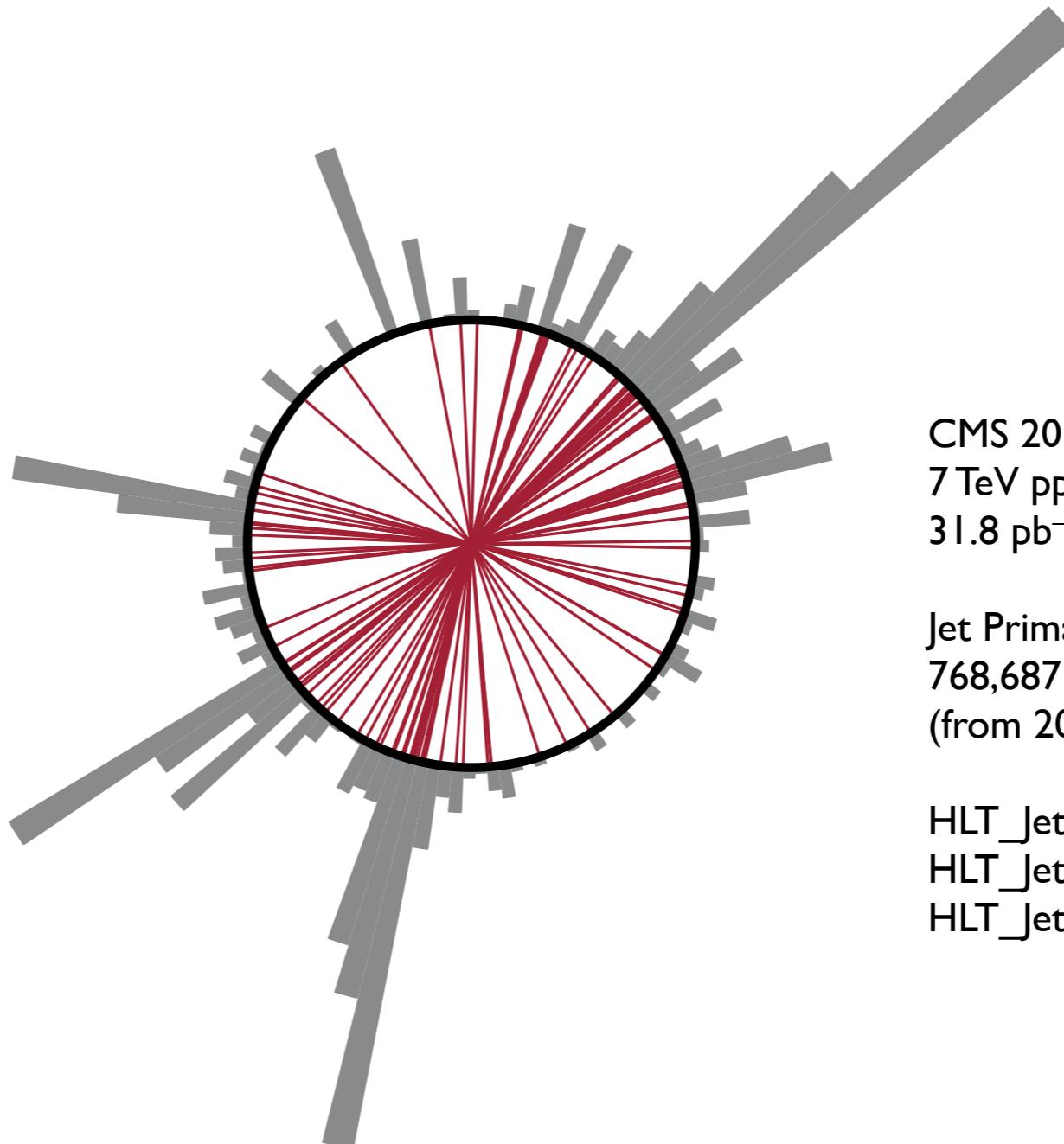
Simone Marzani



Summer intern:
Alexis Romero



CMS advice:
Sal Rappoccio



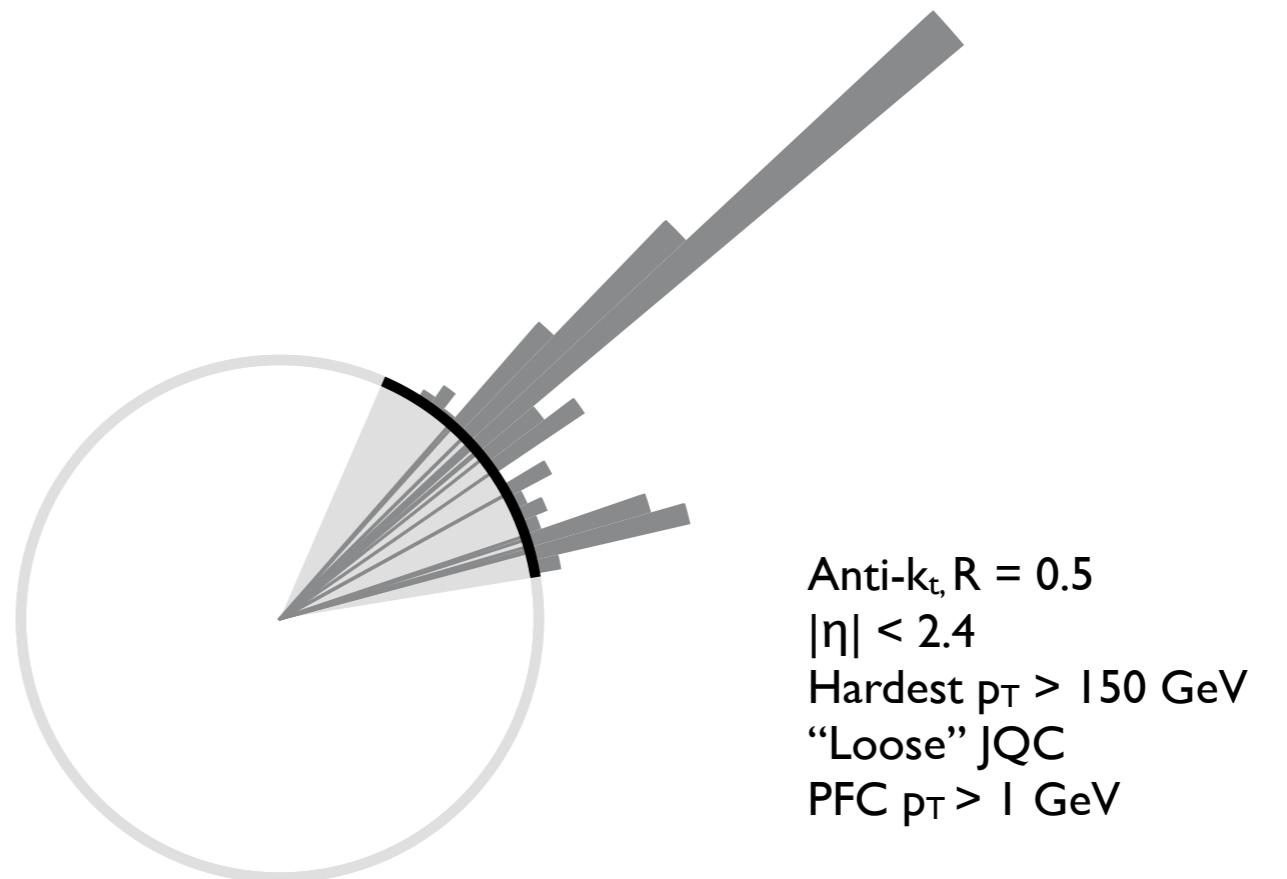
CMS 2010 Run B
7 TeV pp
 31.8 pb^{-1}

Jet Primary Dataset
768,687 events
(from 20 million)

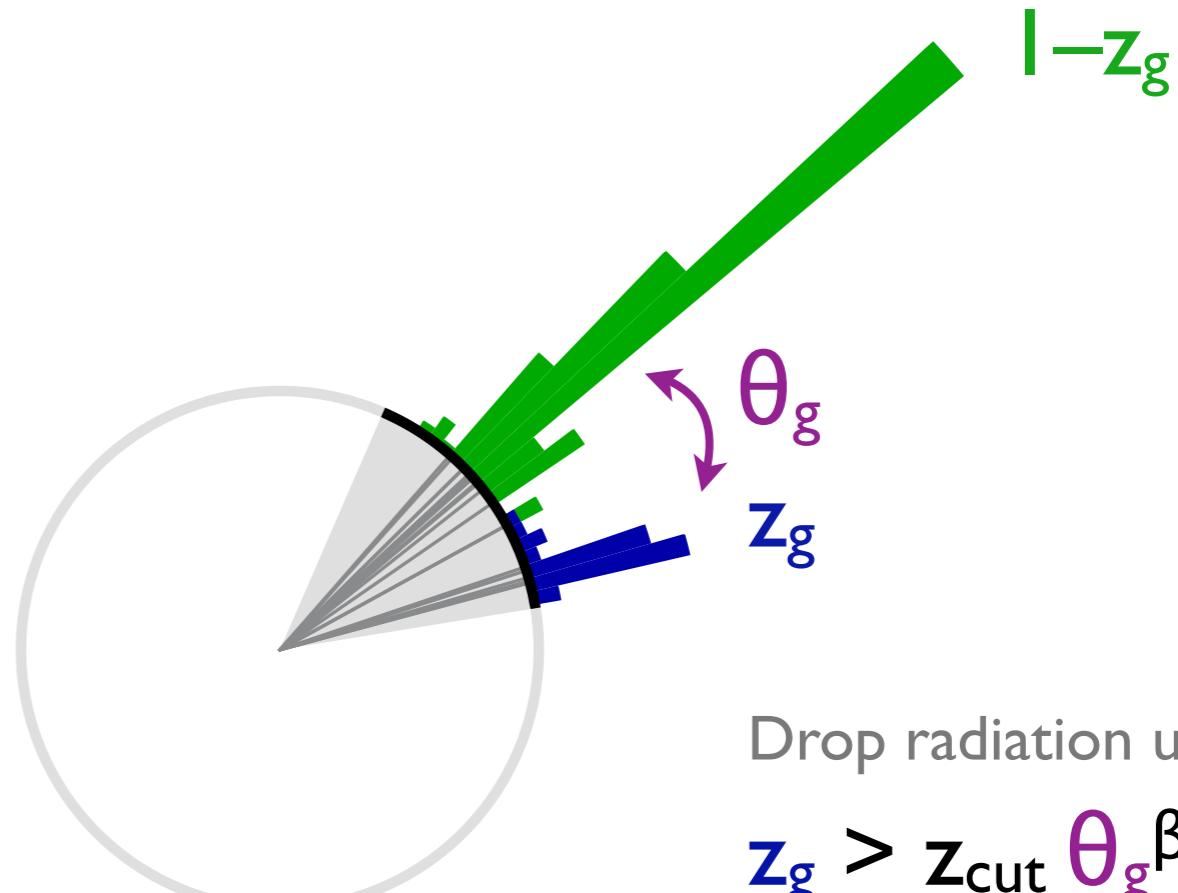
HLT_Jet70U
HLT_Jet100U
HLT_Jet140U



Anti- k_t , $R = 0.5$
 $|\eta| < 2.4$
 $p_T > 20 \text{ GeV}$
“Loose” JQC

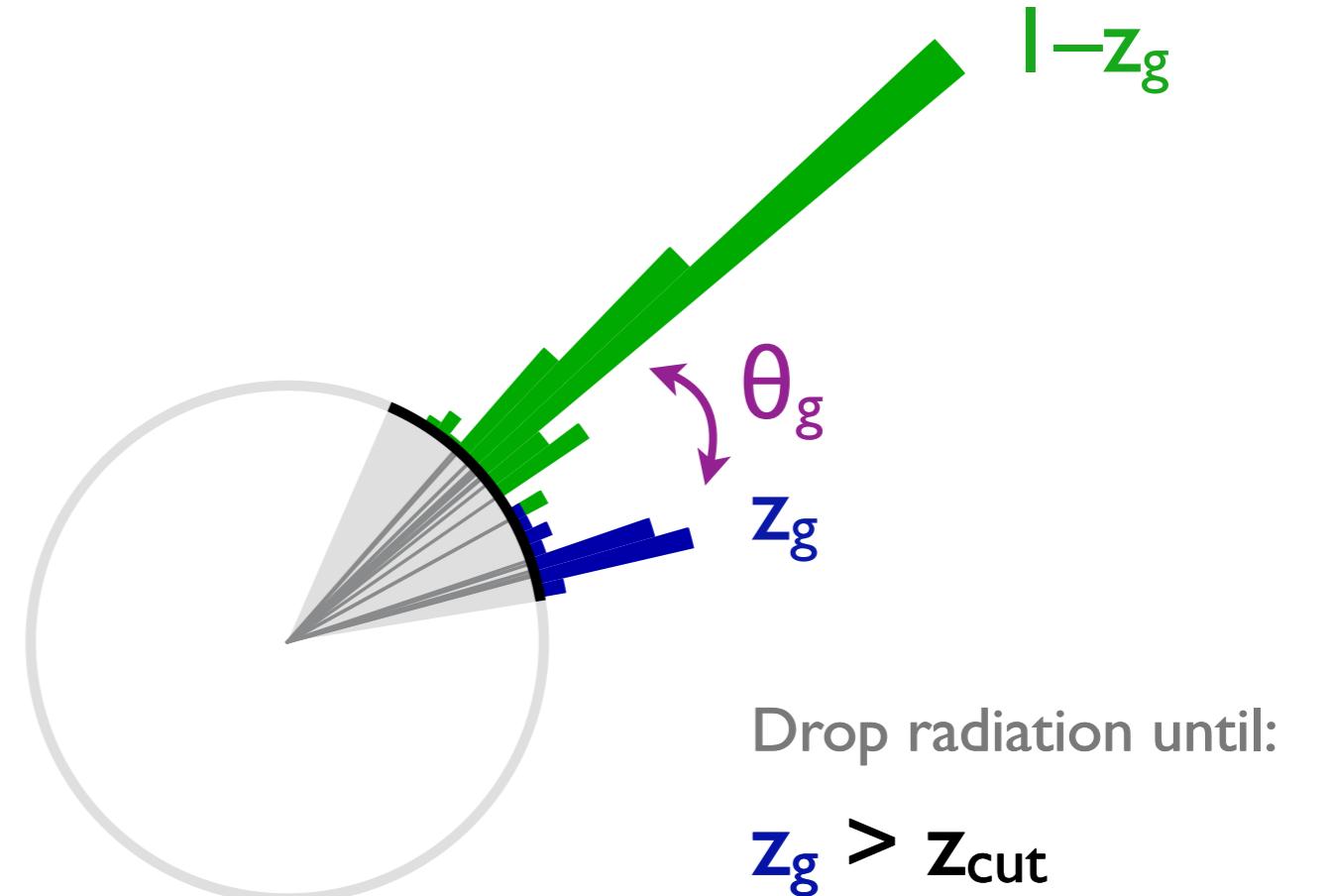
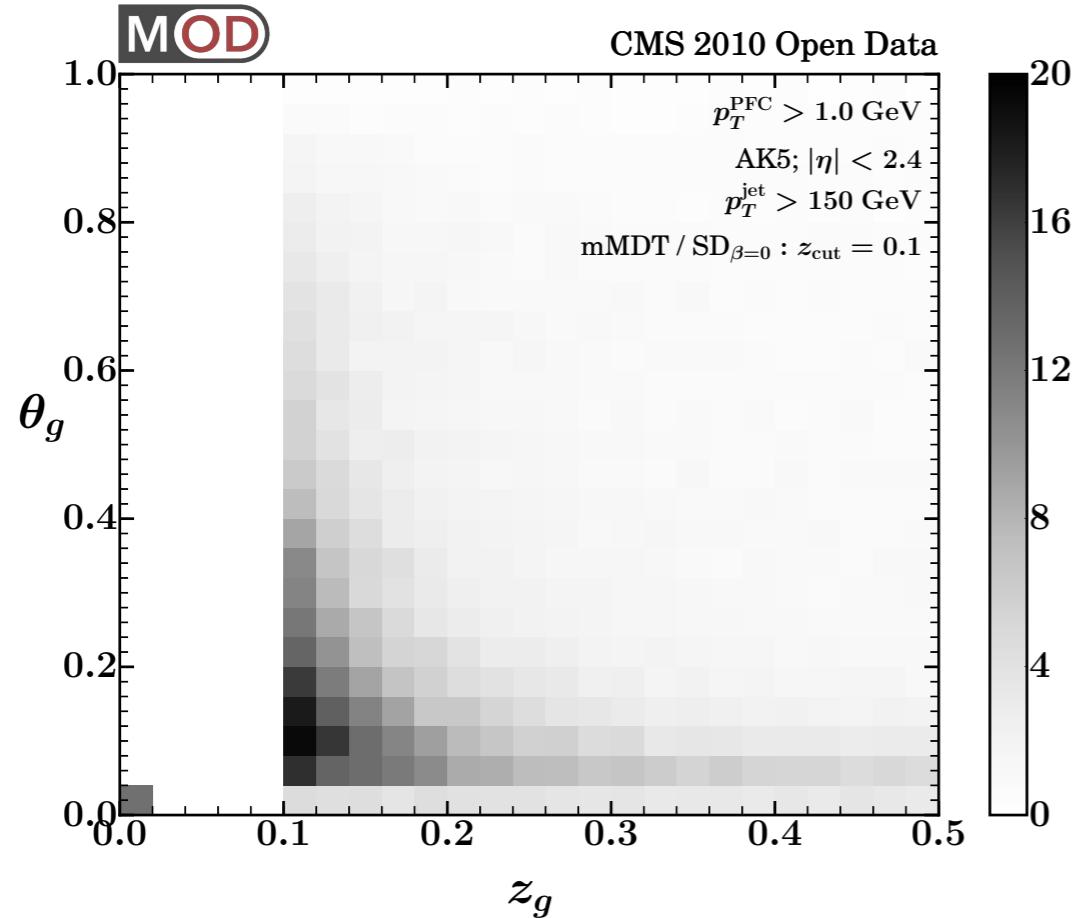


mMDT/Soft Drop



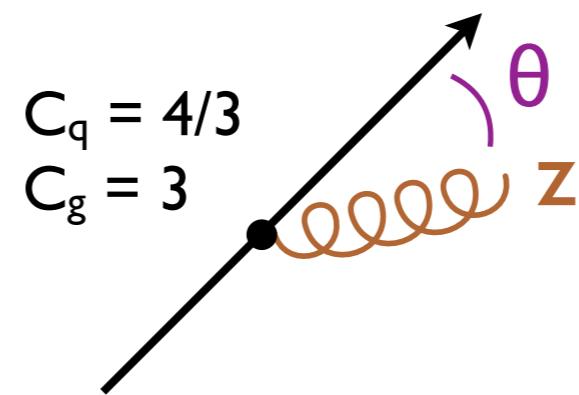
[Larkoski, Marzani, Soyez, JDT, 1402.2657; Dasgupta, Fregoso, Marzani, Salam, 1307.0007;
see also Butterworth, Davison, Rubin, Salam, 0802.2470]

Grooming Highlights Soft/Collinear Behavior



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

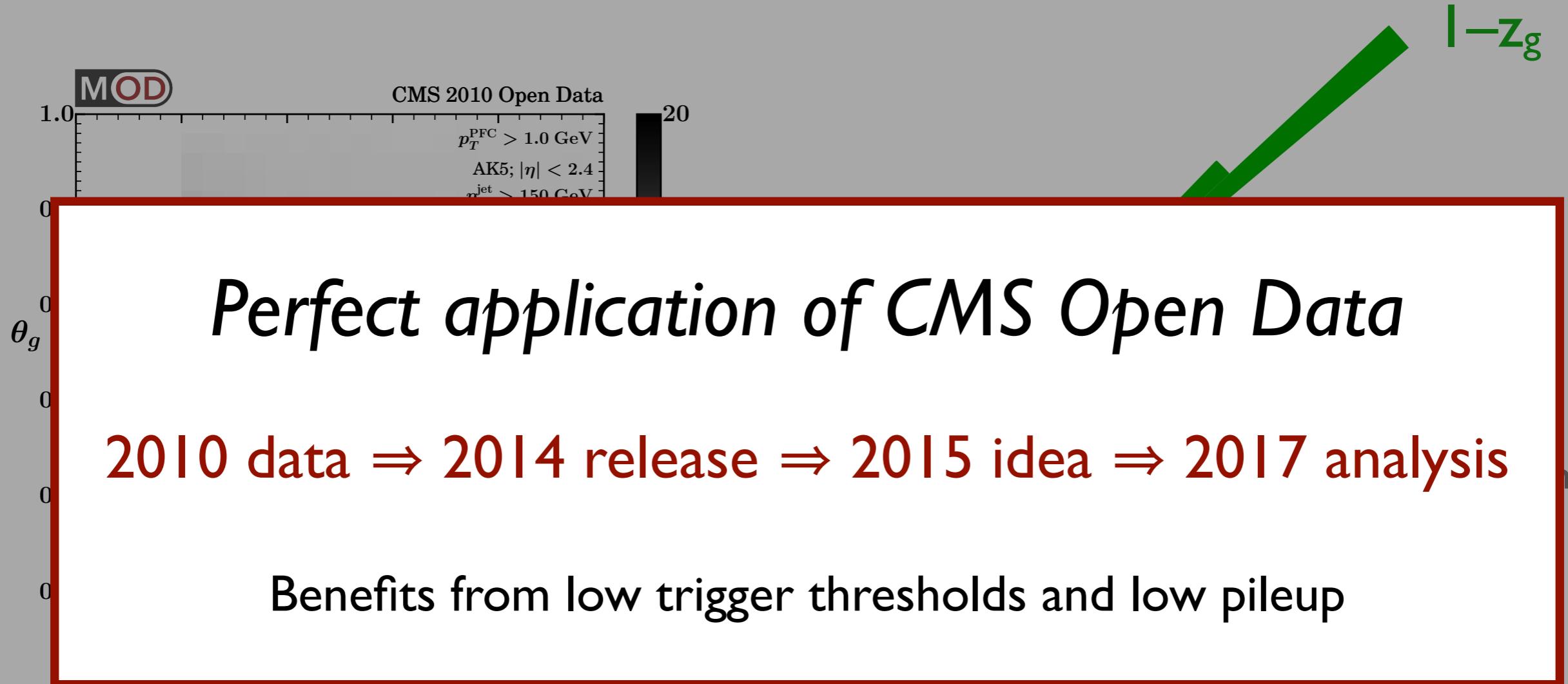
— Collinear — Soft



z ≈ z_g (!)

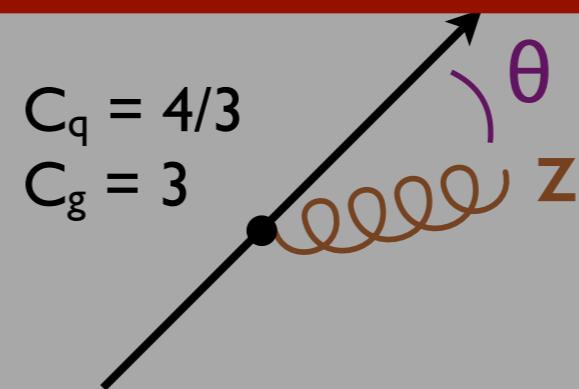
[Larkoski, Marzani, JDT, 1502.01719;
 see also Larkoski, JDT, 1307.1699]

Grooming Highlights Soft/Collinear Behavior



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

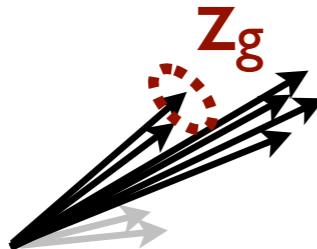
Collinear Soft



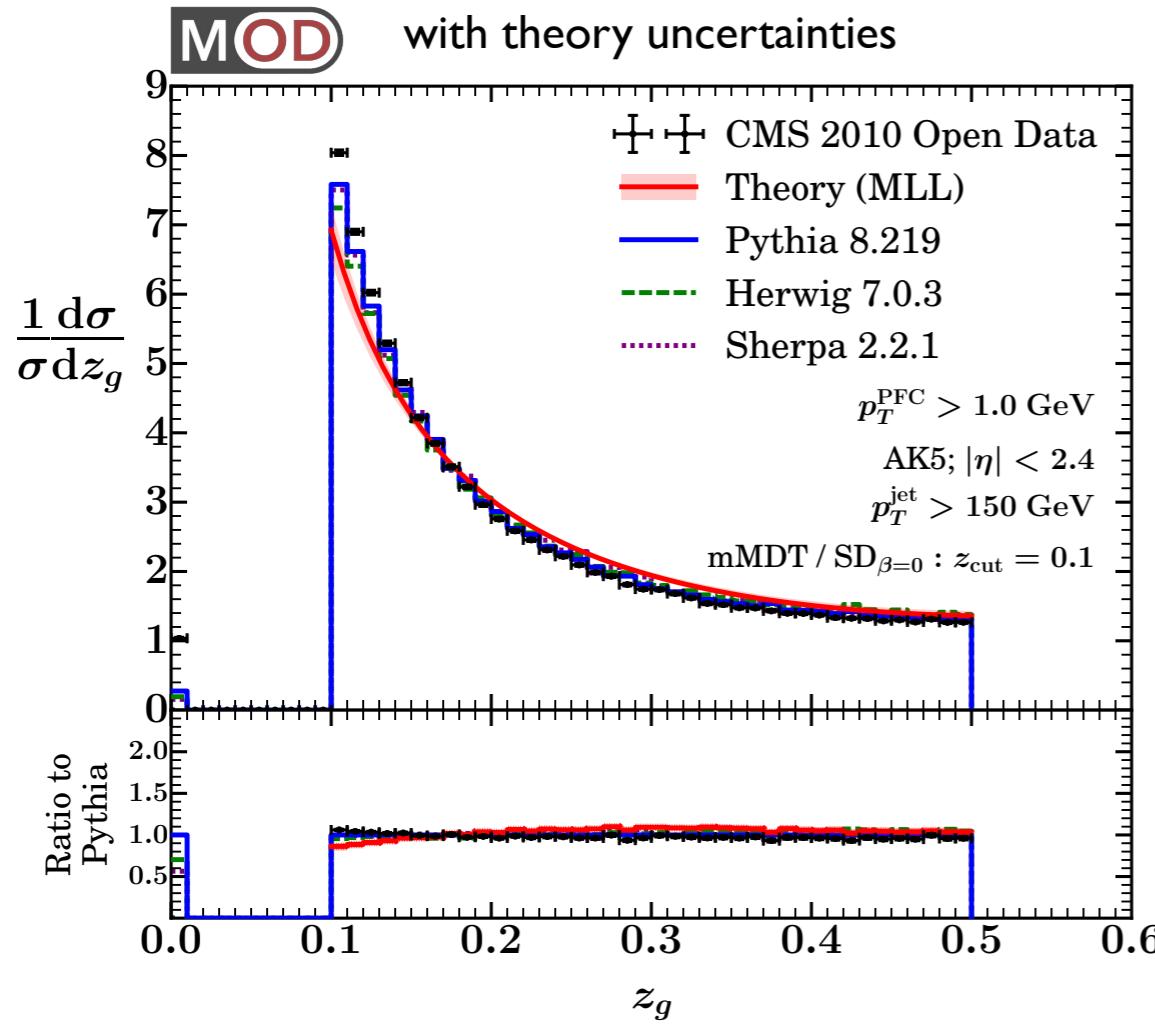
$$z \approx z_g \quad (!)$$

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

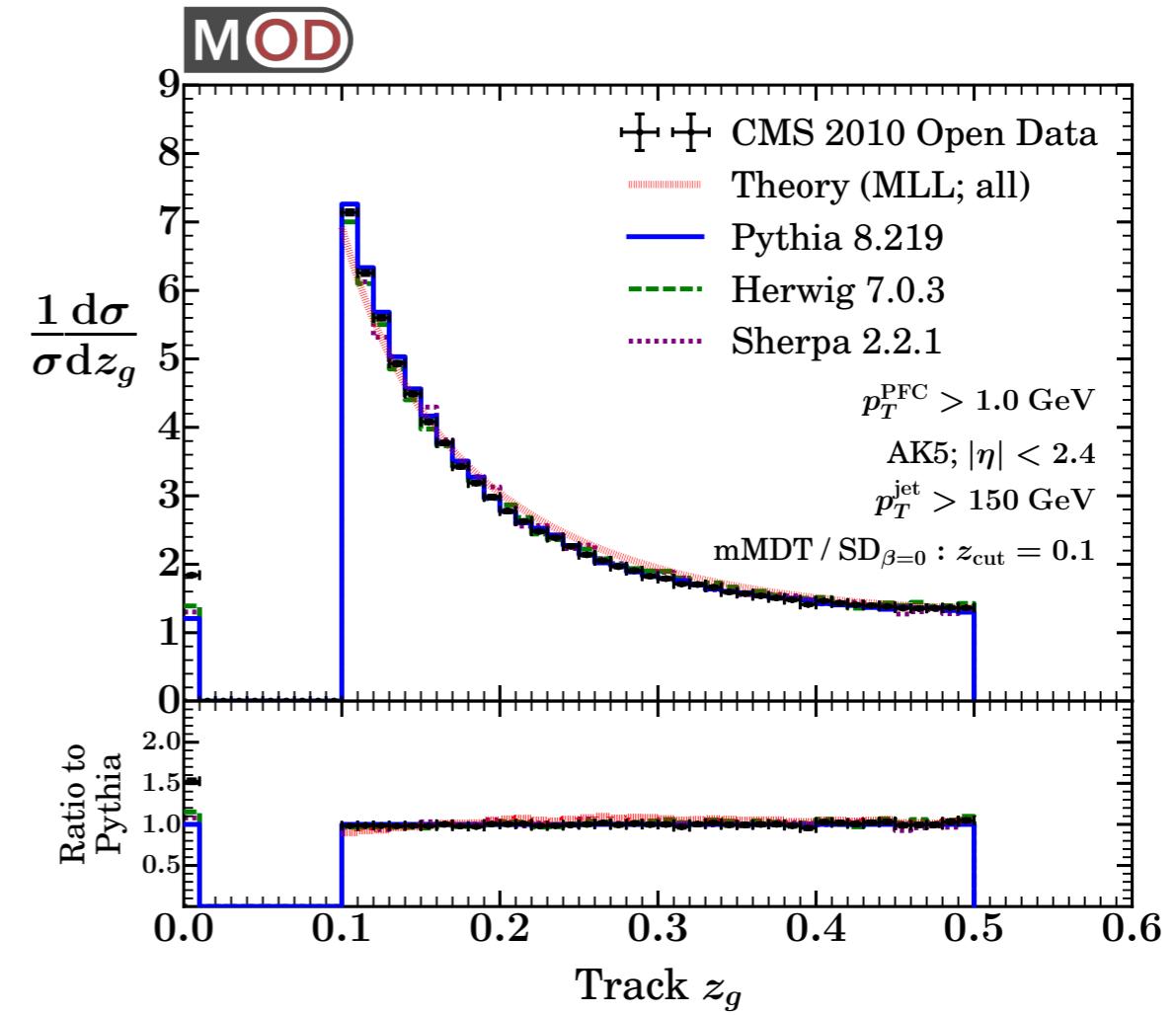
Exposing the QCD Splitting Function



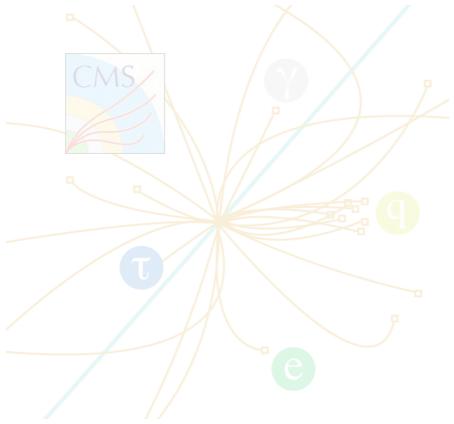
All particle



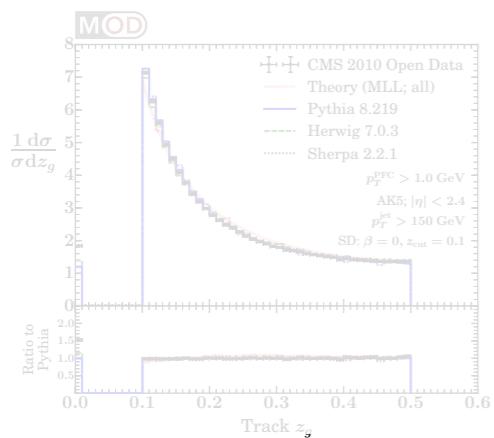
Track-only



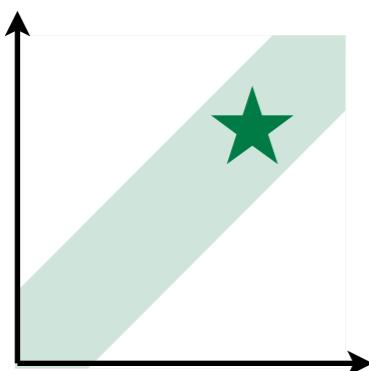
[Larkoski, Marzani, JDT, Tripathee, Xue, 1704.05066]



Using the CMS Open Data

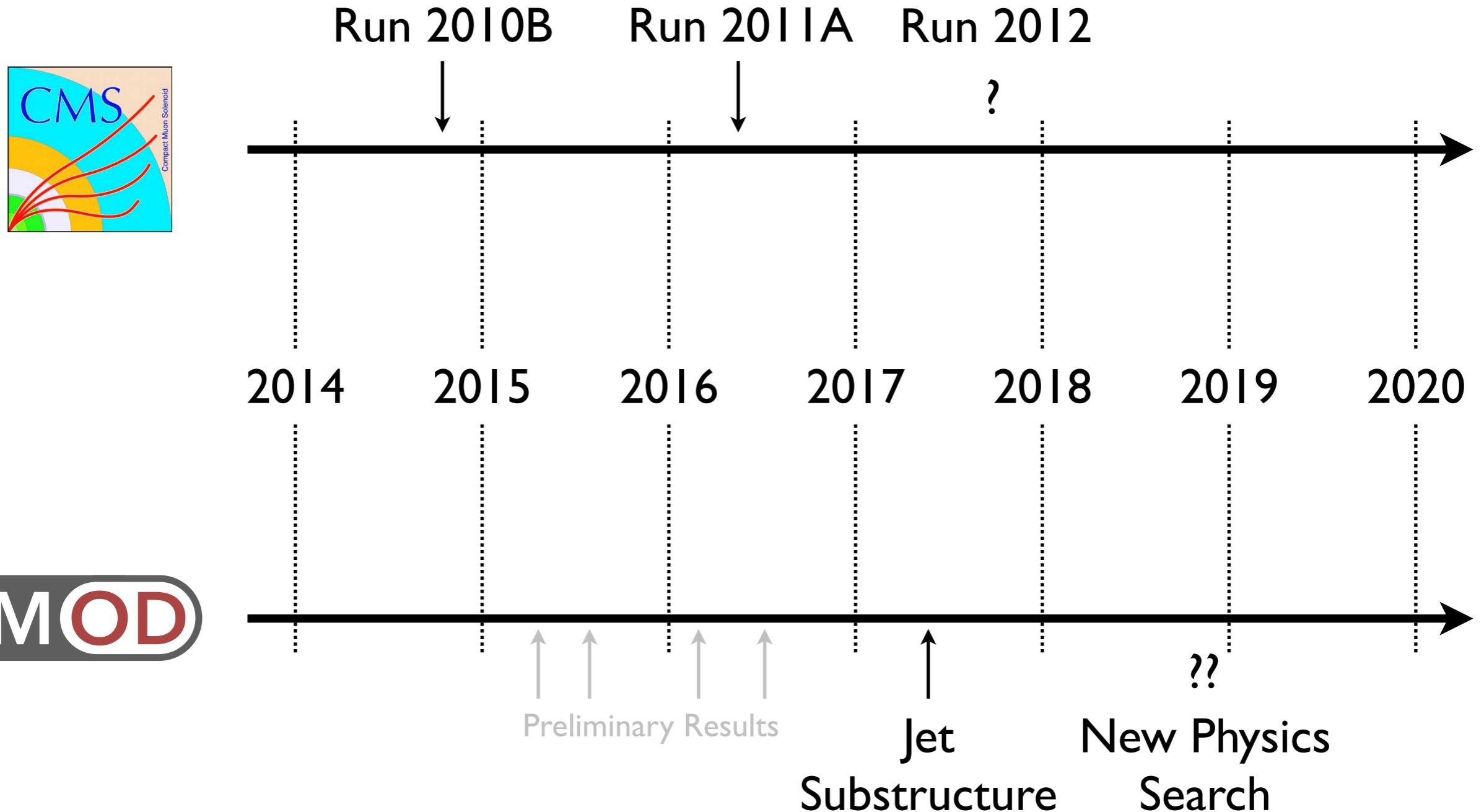


Jet Substructure and QCD Splittings



The Future of Public Collider Data

The Open Data Pipeline



Data preservation (and outside analyses)
require significant resources:

People, time, ideas, and money

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

*Let me address some concerns about public data raised by our work
(not exact quotes, not exhaustive, not really representative)*

Confronting the Steep Learning Curve

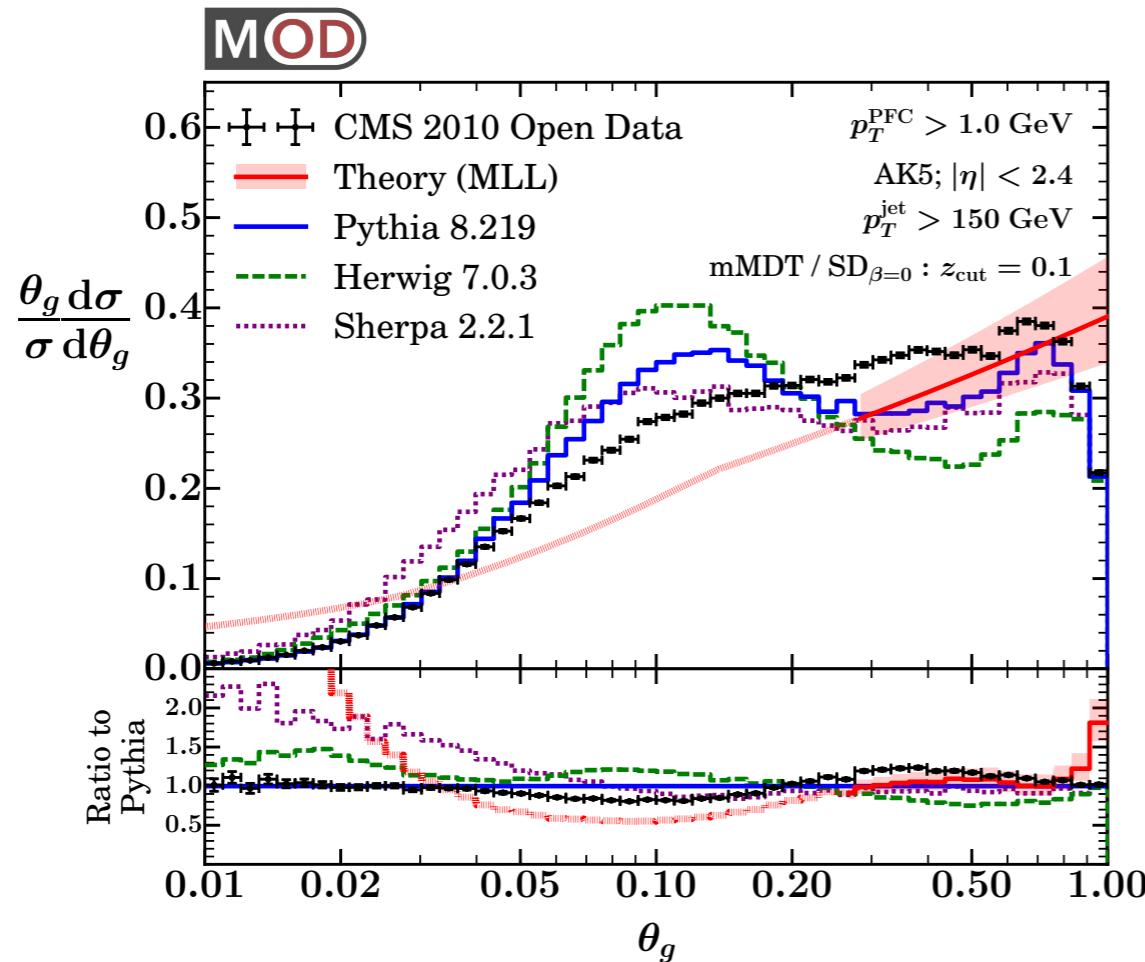
“I support open data on principle, but it seems to require an excessive amount of effort to use the CMS Open Data”

CMS Primary Datasets	CMS Simulated Datasets	CMS Learning Resources
CMS primary datasets are AOD (Analysis Object Data) files, which contain the information that is needed for analysis	This collection contains CMS Simulated Datasets.	This collection includes learning resources that use CMS public data
Years: 2010, 2011	Years: 2010, 2011	VS.
Total records: 33	Total records: 381	Total records: 7

No magic, but with suitable investment, open data could be as straightforward to parse and interpret as detector-simulated Monte Carlo (e.g. MOD 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”



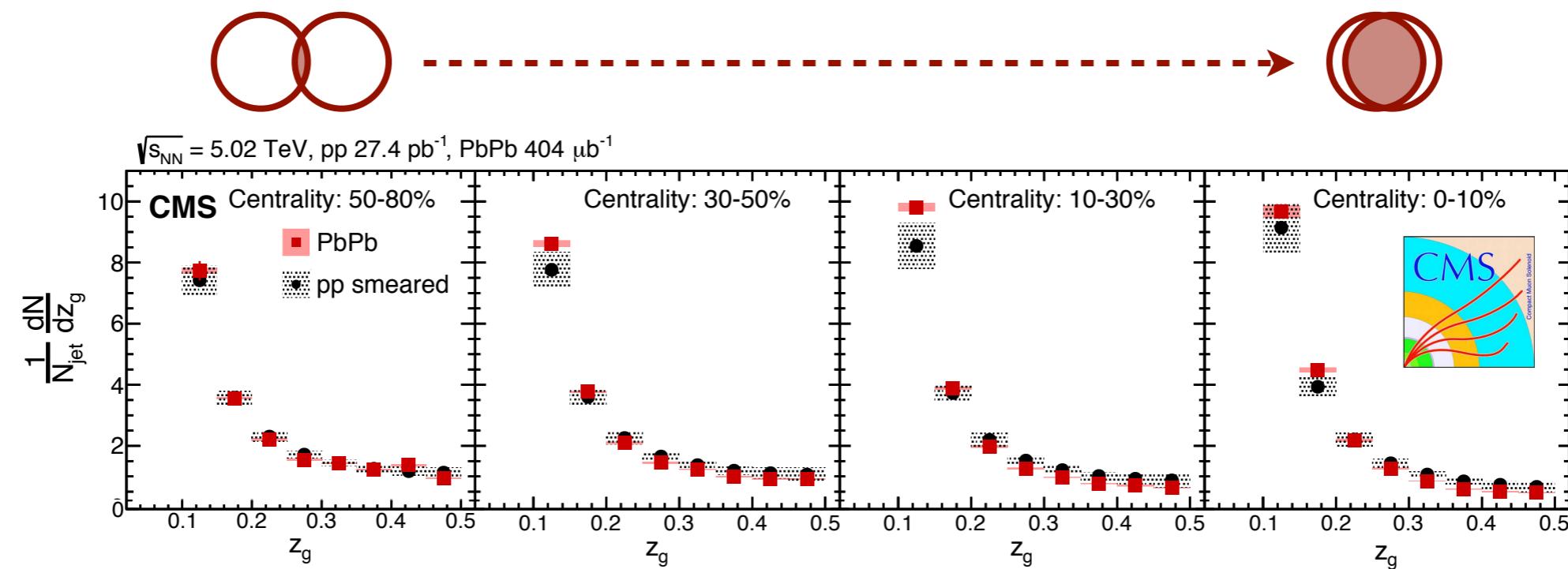
Agreed (mostly)

But with unexpected
theoretical/experimental
issues at play, value in
exploratory studies

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

Synergy between Internal and External Efforts

“This work competes with ongoing collaboration analyses without the scrutiny of internal review; careers are at stake”



I will be heartbroken if open data adversely impacts experimental progress, and I hope our work inspires more vigorous investigations into jets and QCD

[CMS, 1708.09429]

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?" ←

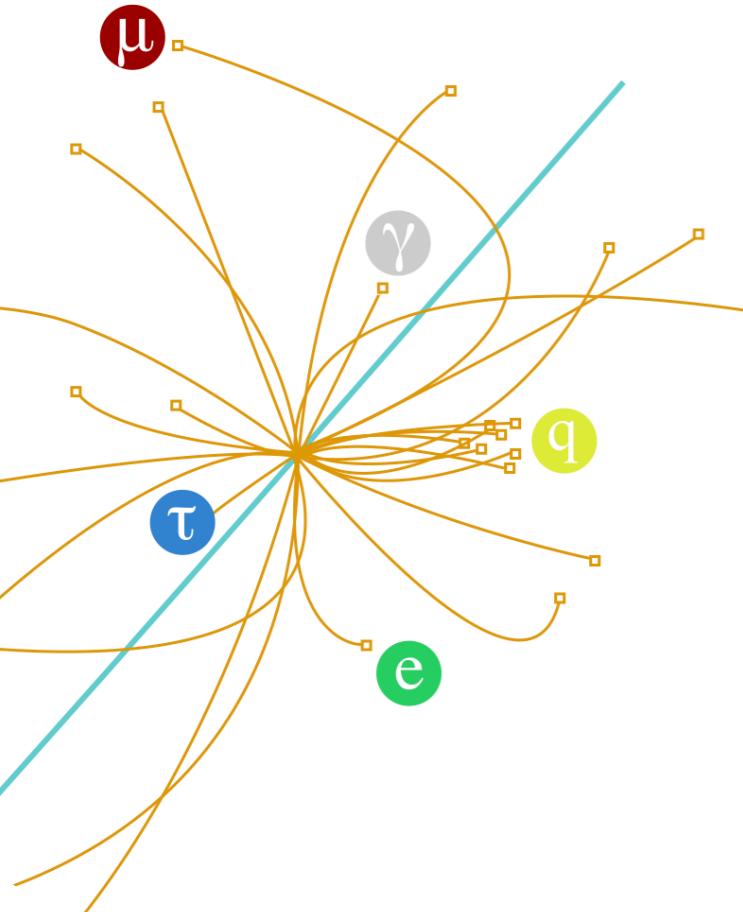
To analyse CMS data collected in 2010, you need **version 4.2.8** of CMSSW, supported only on **Scientific Linux 5**. If you are unfamiliar with Linux, take a look at [this short introduction to Linux](#) or try this interactive [command-line bootcamp](#). Once you have installed the CMS-specific [CERN 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

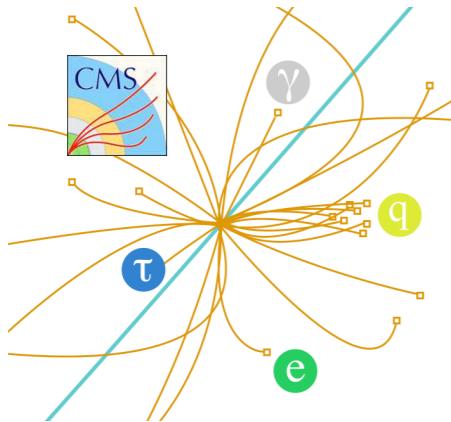
Enabling exploratory/proof-of-principle studies

Facilitating dialogue between theory and experiment

Researching physics in and beyond the standard model

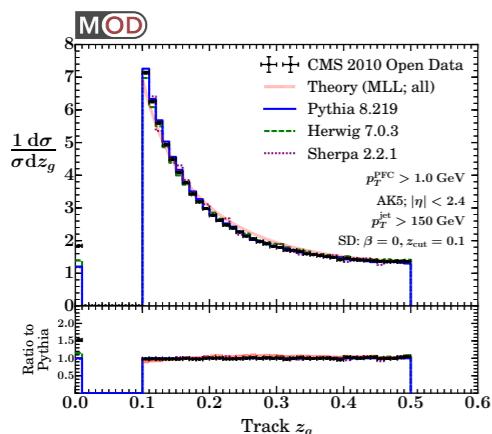
*These are only possible with sustained
investment in public data initiatives*

Summary



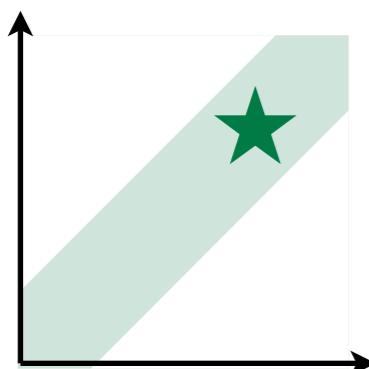
Using the CMS Open Data

Unique collider data set, ideal for exploratory studies



Jet Substructure and QCD Splittings

Exposing the universal singularity structure of gauge theories

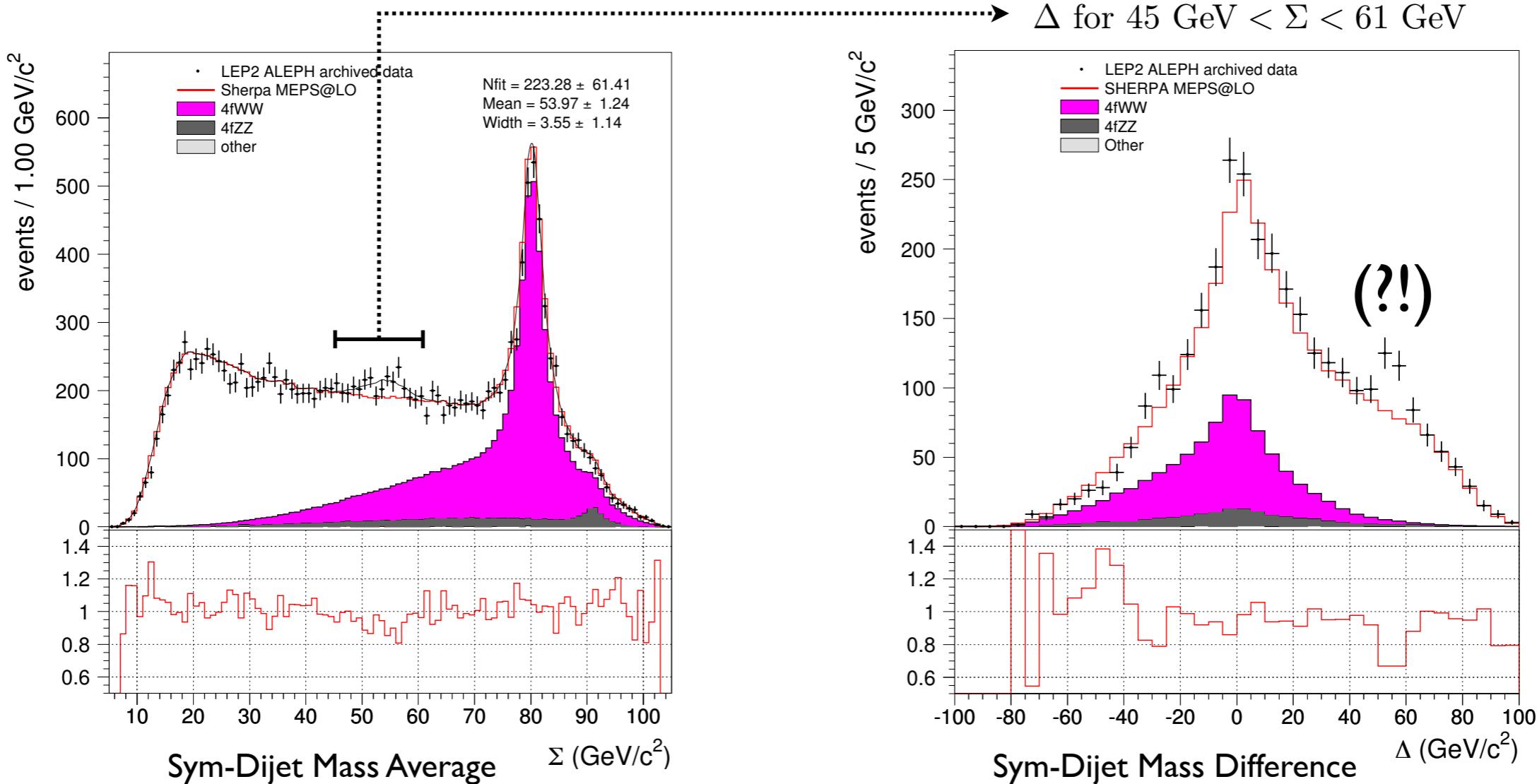


The Future of Public Collider Data

Sustained investment from outreach to research to archives

Backup Slides

A Quad-Jet Puzzle in Archival ALEPH Data

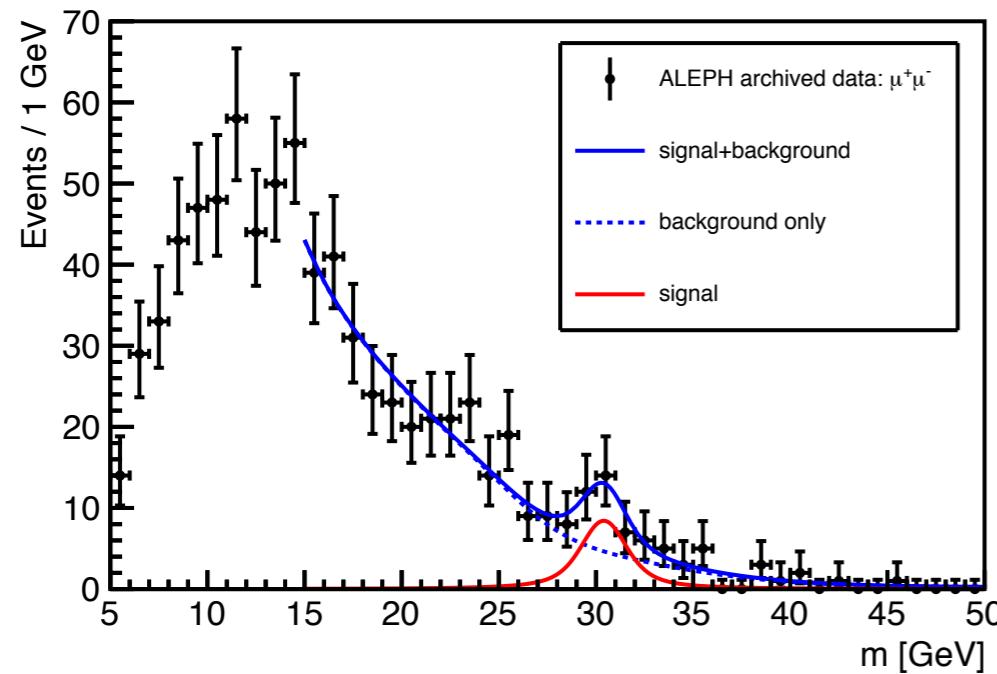


*Science thrives on openness, reproducibility, and intense scrutiny
What role should legacy data sets play in collider physics?*

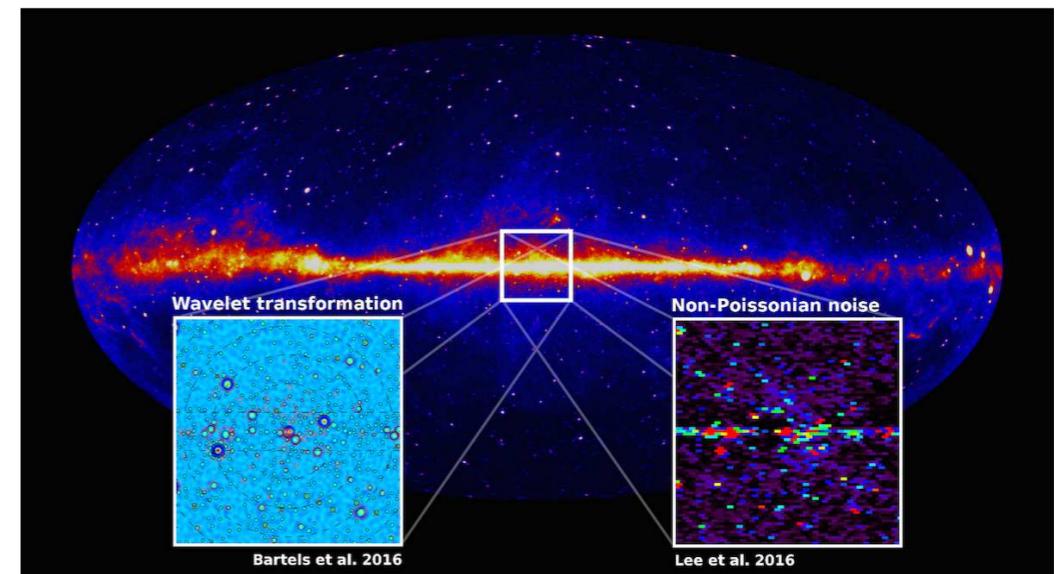
[Kile, von Wimmersperg-Toeller, 1706.02242, 1706.02255, 1706.02269]

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.

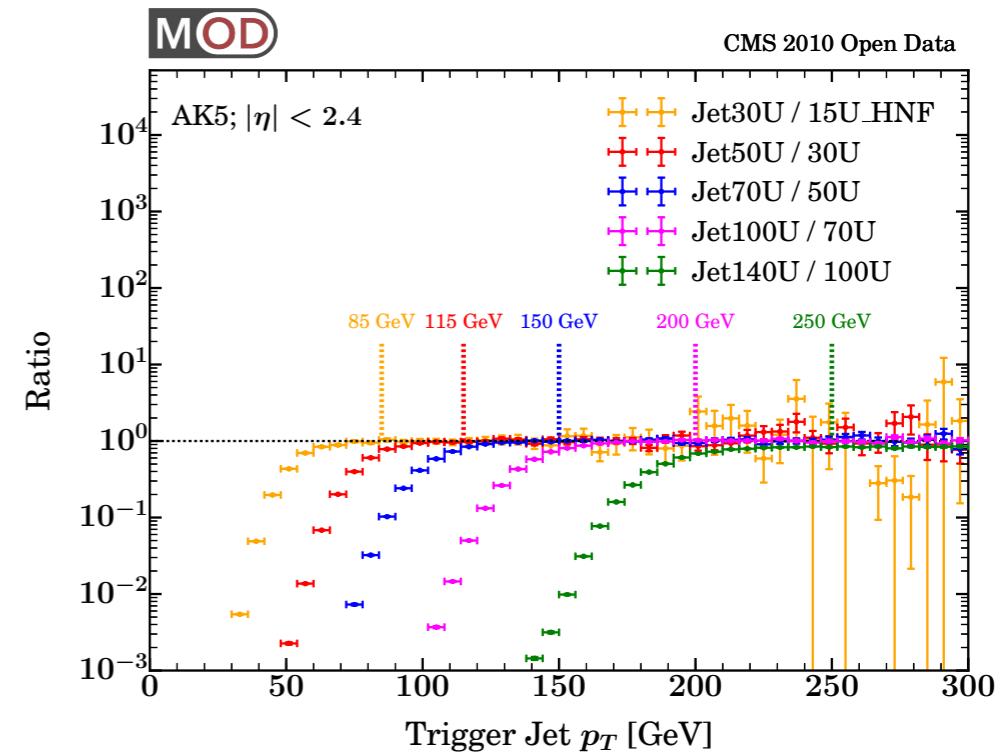
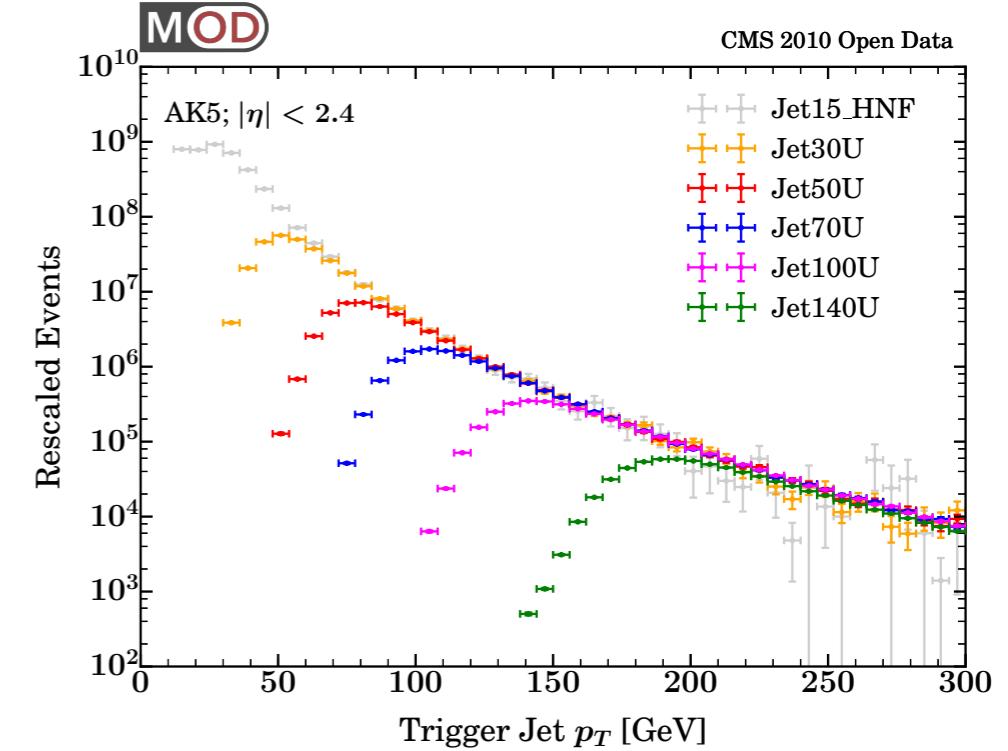


Possibly the opposite: the easier the data is to use, the more likely it will be used correctly and the results cross-checked by other groups

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

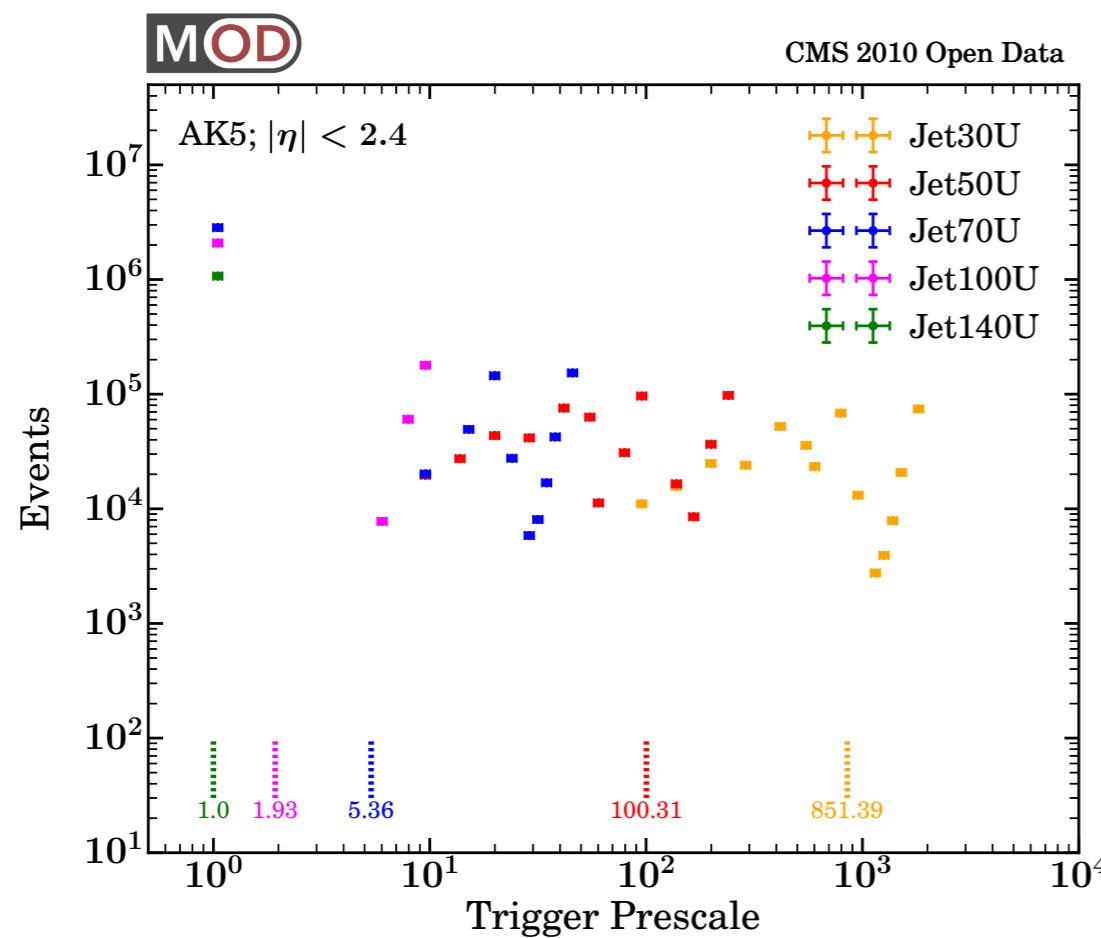
Trigger Selection and Efficiency

	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
Di-jet	HLT_DiJetAve15U	2,595,038	553,331
	HLT_DiJetAve30U	16,341,191	1,067,561
	HLT_DiJetAve50U	16,341,191	648,000
	HLT_DiJetAve70U	16,341,191	2,310,033
	HLT_DiJetAve100U	5,989,945	1,252,661
	HLT_DiJetAve140U	2,595,038	452,222
	HLT_QuadJet20U	10,351,245	677,451
H_T	HLT_QuadJet25U	10,351,244	219,256
	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



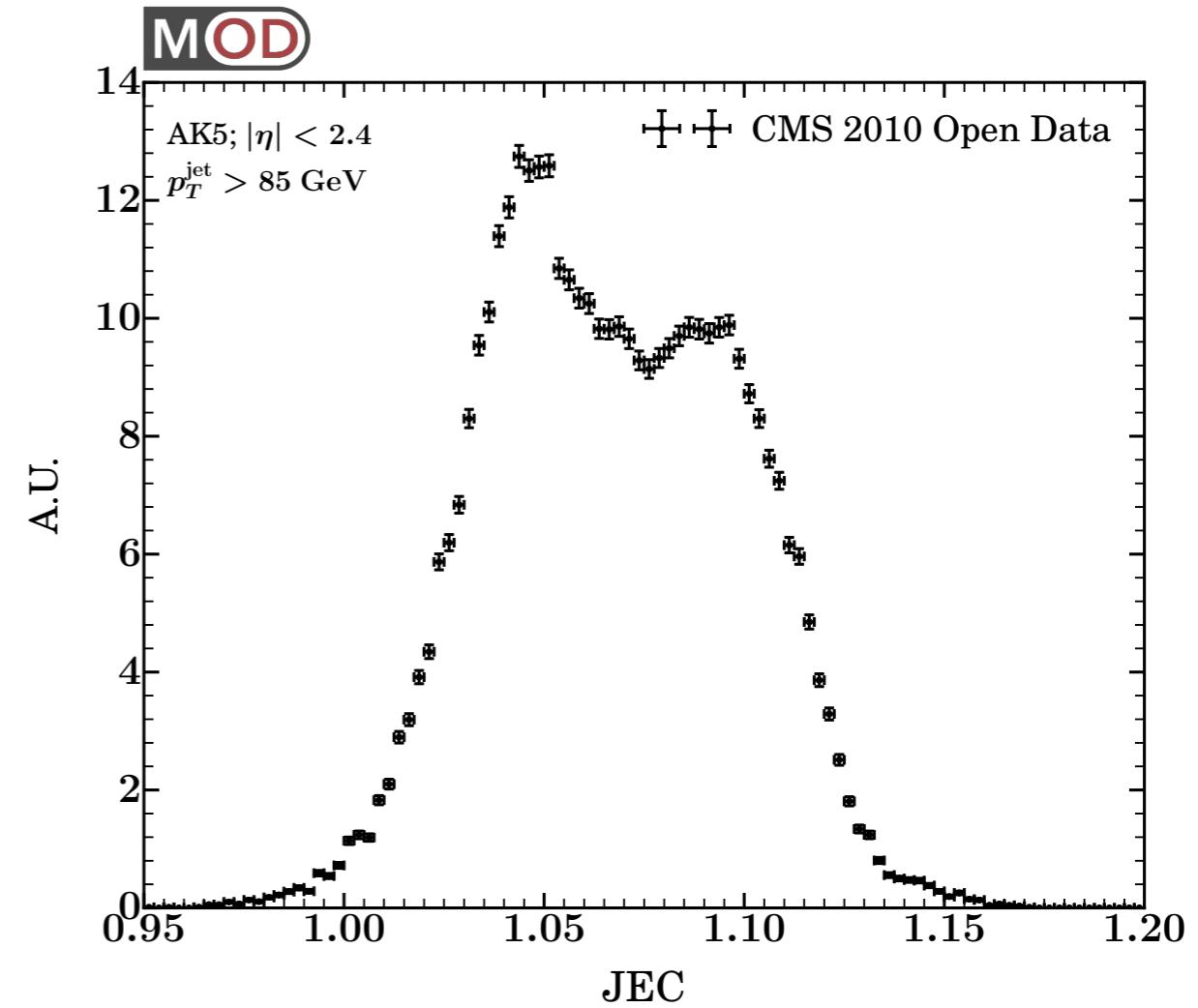
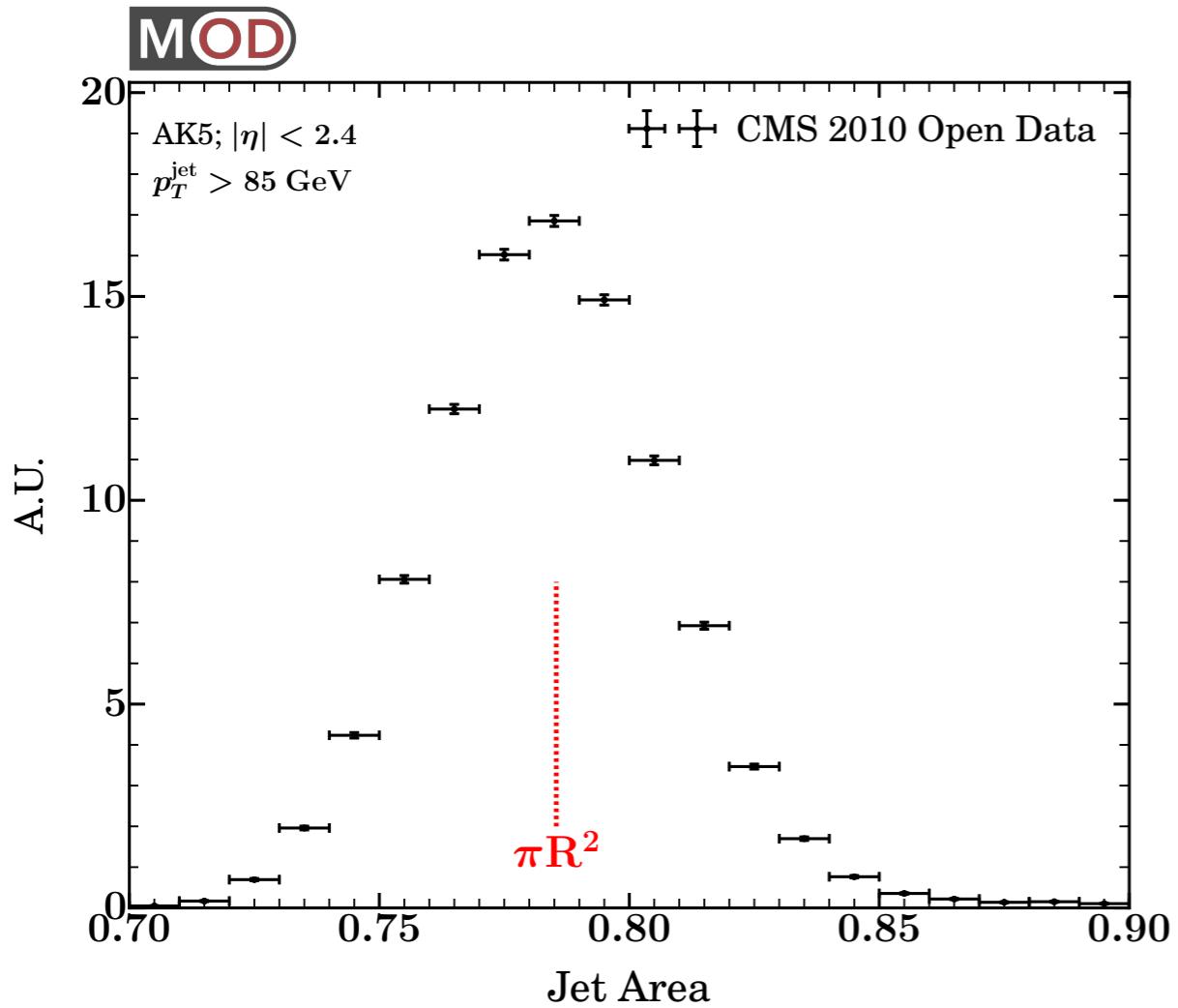
Event Selection, Prescale Factors, Pileup

Hardest Jet p_T	Trigger Name	Events	$\langle \text{Prescale} \rangle$
[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



N_{PV}	Jet Primary Dataset		Hardest Jet Selection	
	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

Jet Corrections



Jet Area Subtraction



Jet Energy Corrections

Workflow

Instrumentation

Physics

	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
$ \eta < 2.4$	768,687	0.038	Central jets
Passes Soft Drop ($z_g > z_{\text{cut}}$)	760,055	0.038	Jet grooming (more later)

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

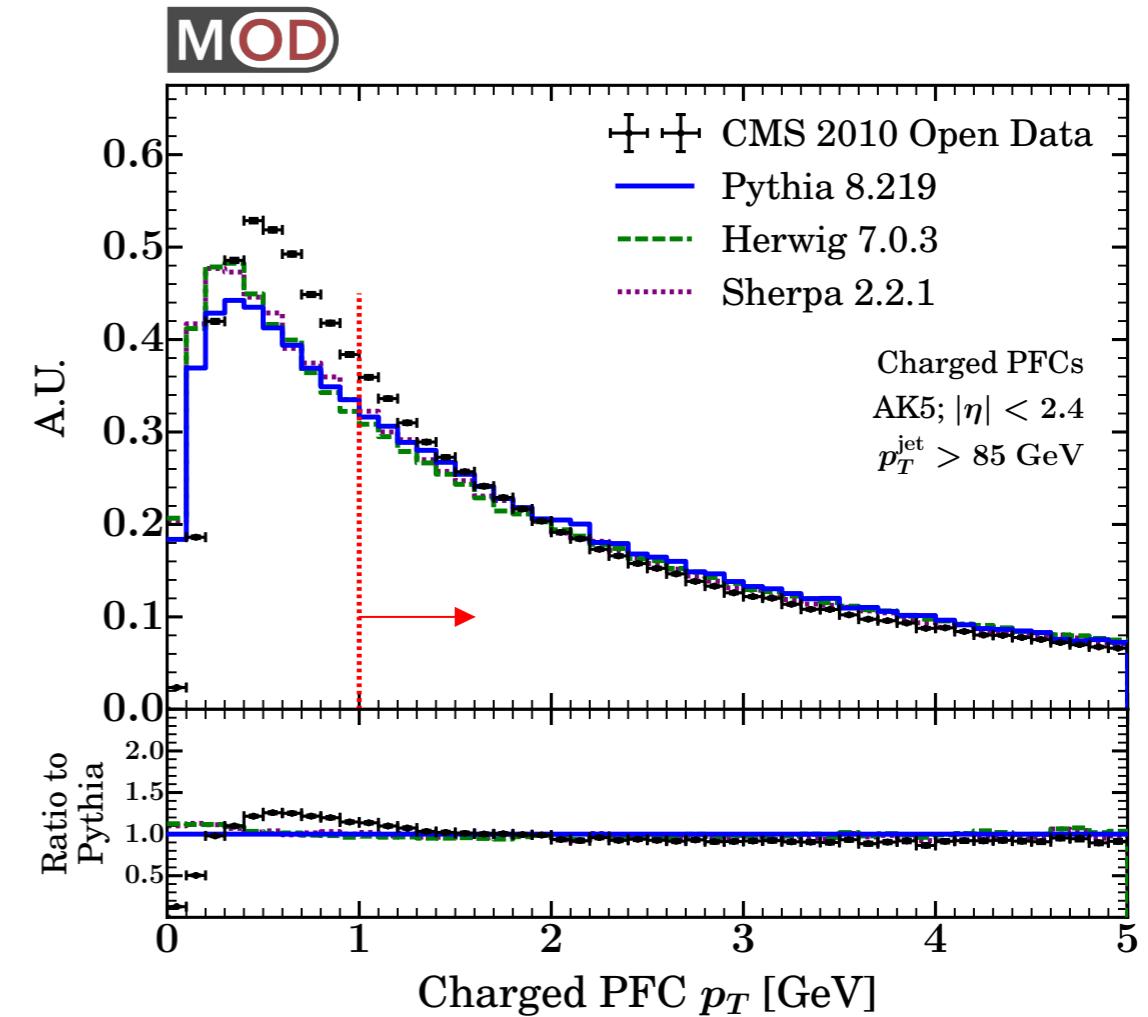
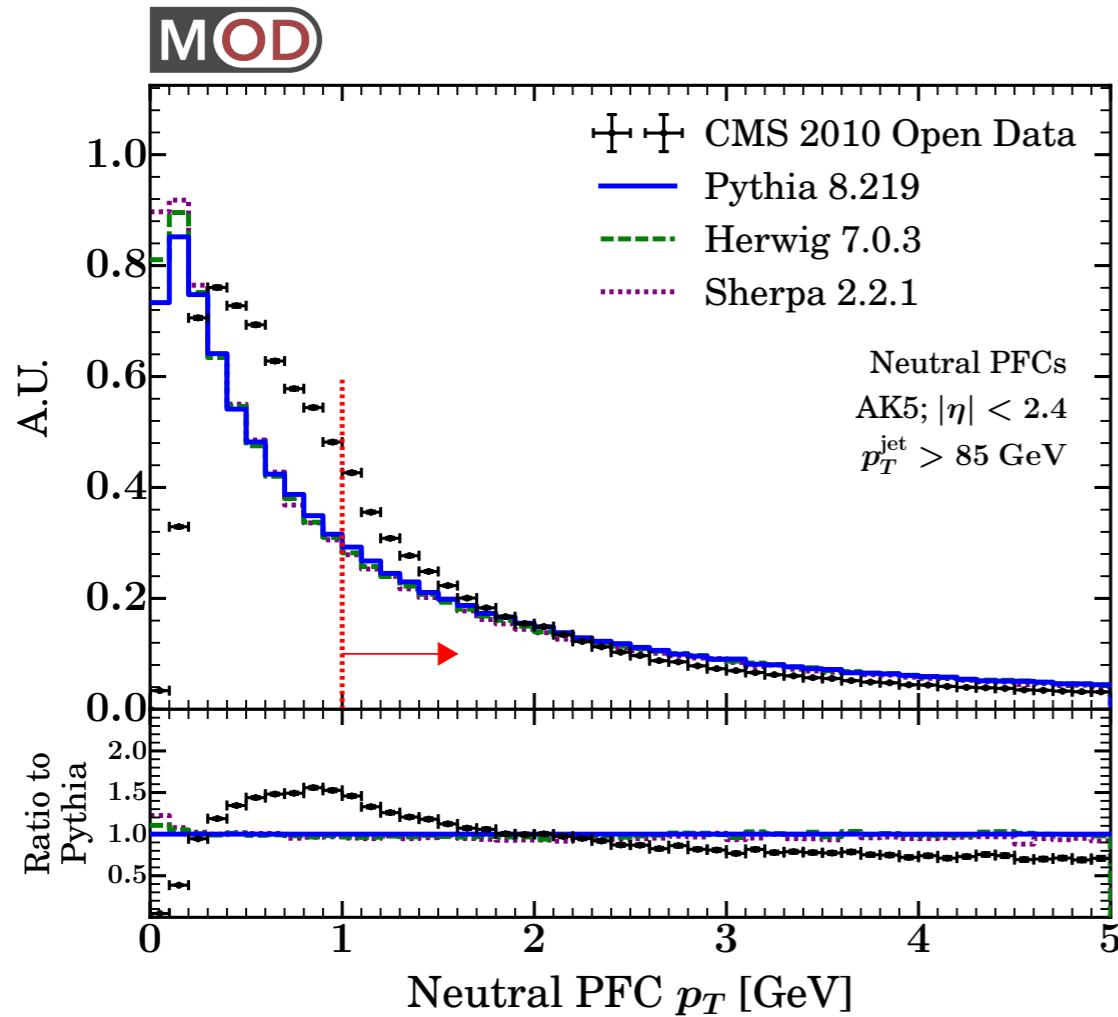
Particle Flow Reconstruction

Workhorse of every CMS substructure analysis

Code	Candidate	Total Count	$p_T > 1 \text{ GeV}$
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. $\pi_0 \rightarrow \gamma\gamma$)

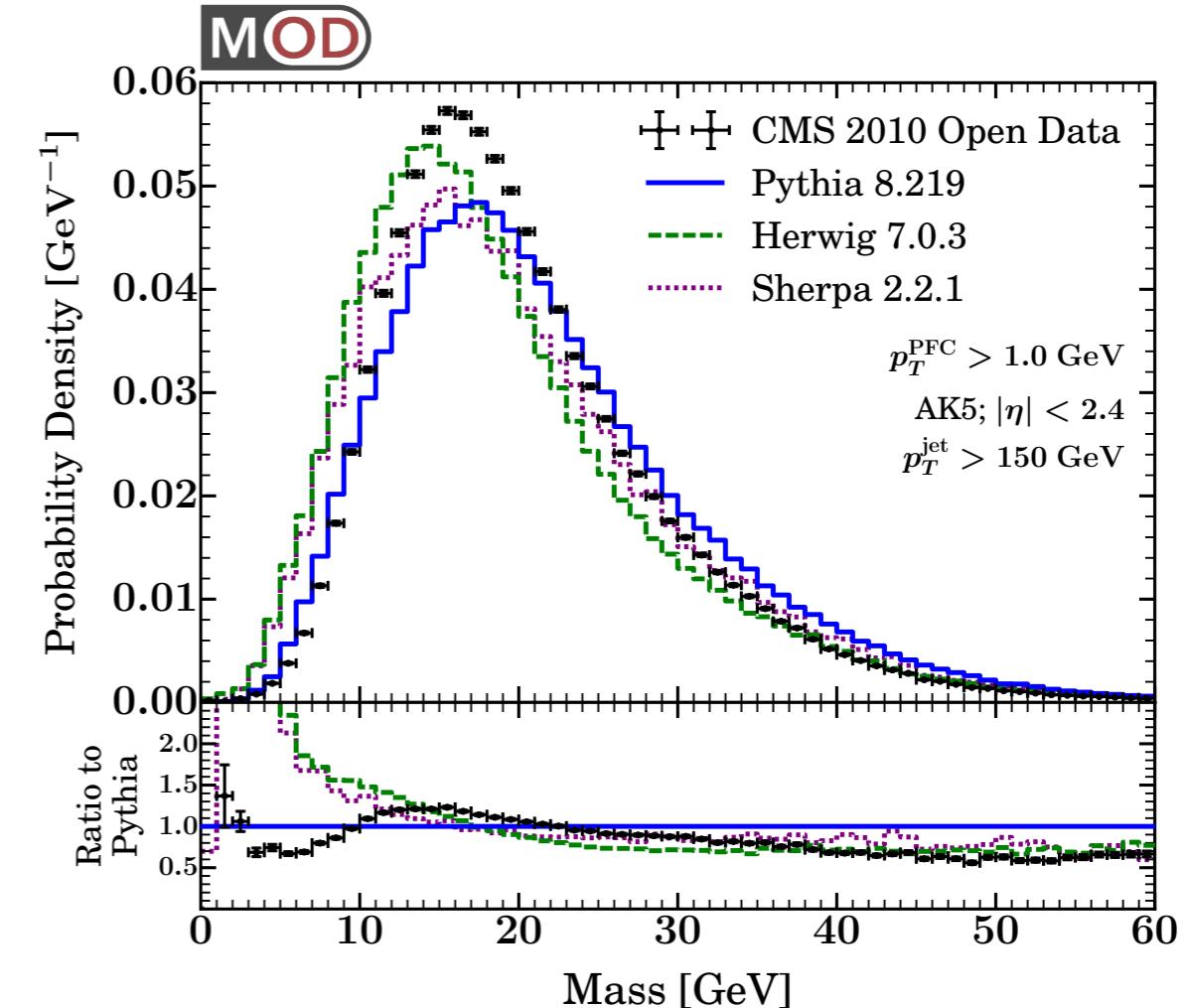
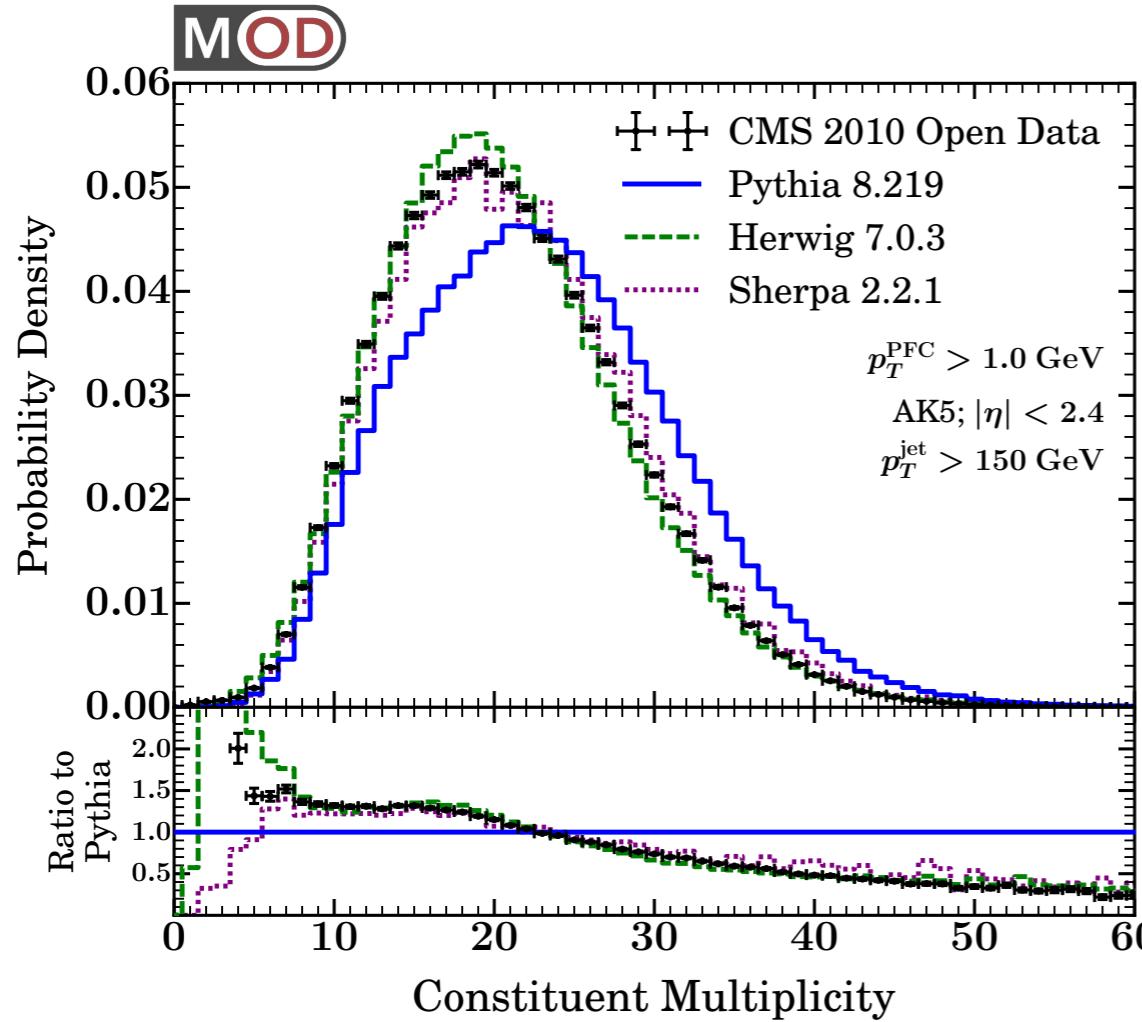
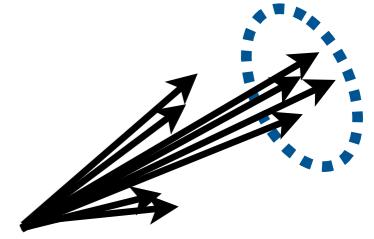
Particle Flow Fiducialization



Motivation to focus on charged PFCs with $p_T > 1 \text{ GeV}$

Basic Substructure

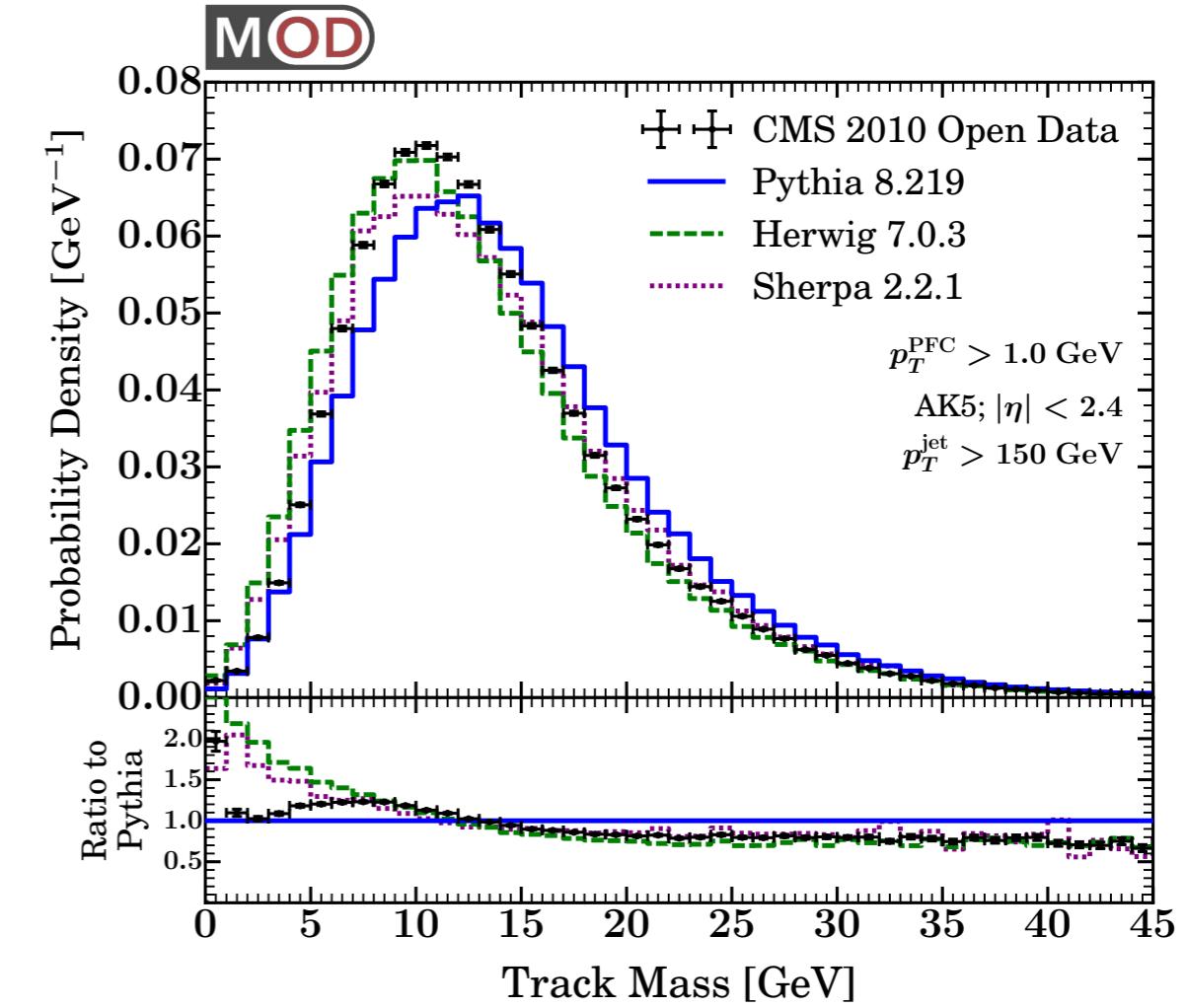
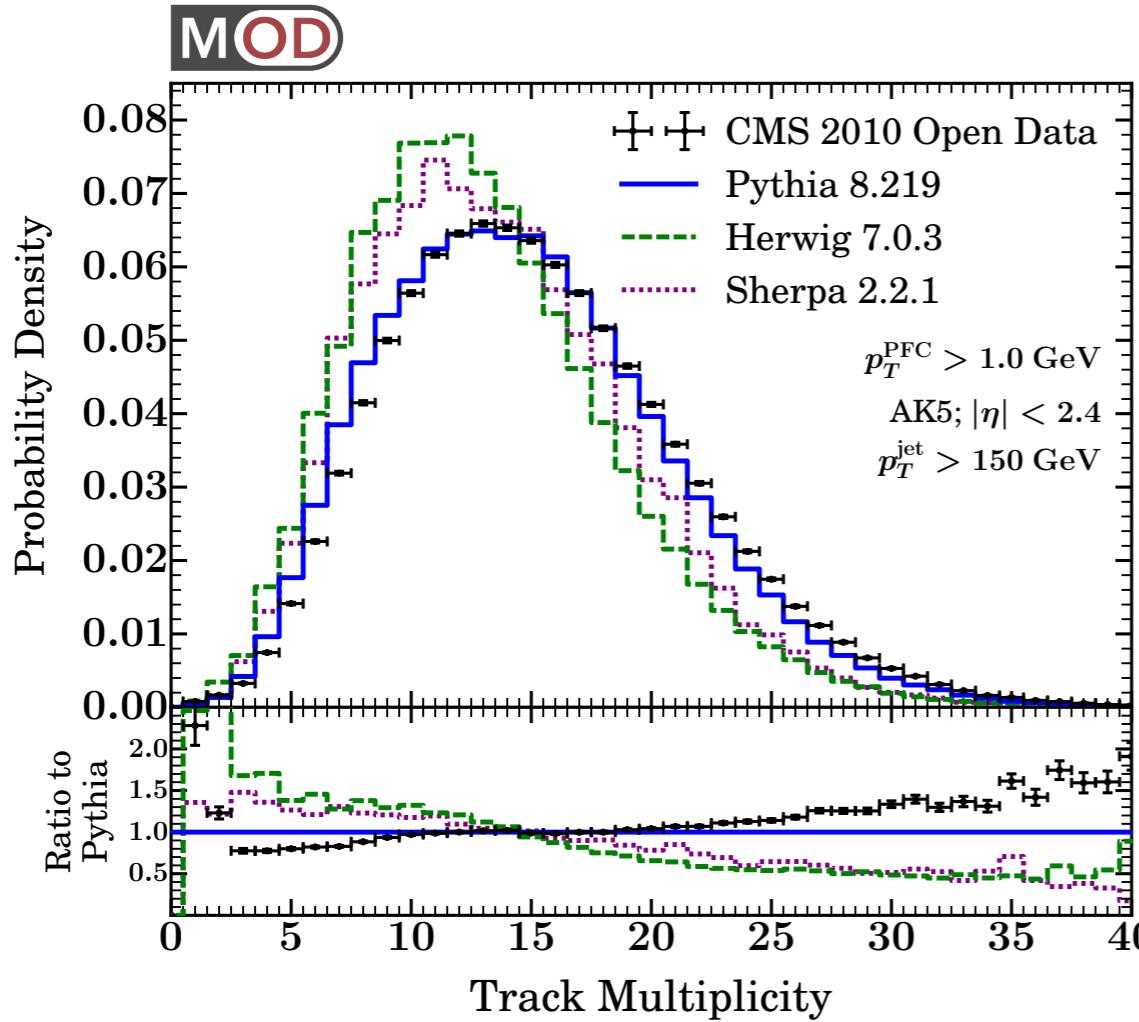
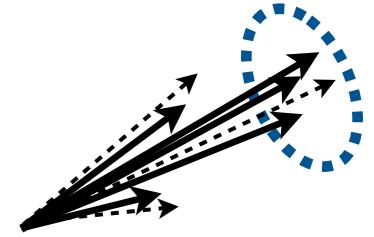
No grooming applied



Careful! Can't assess data/MC (dis)agreement without unfolding
Still interesting to investigate MC/MC differences

Track-Based Substructure

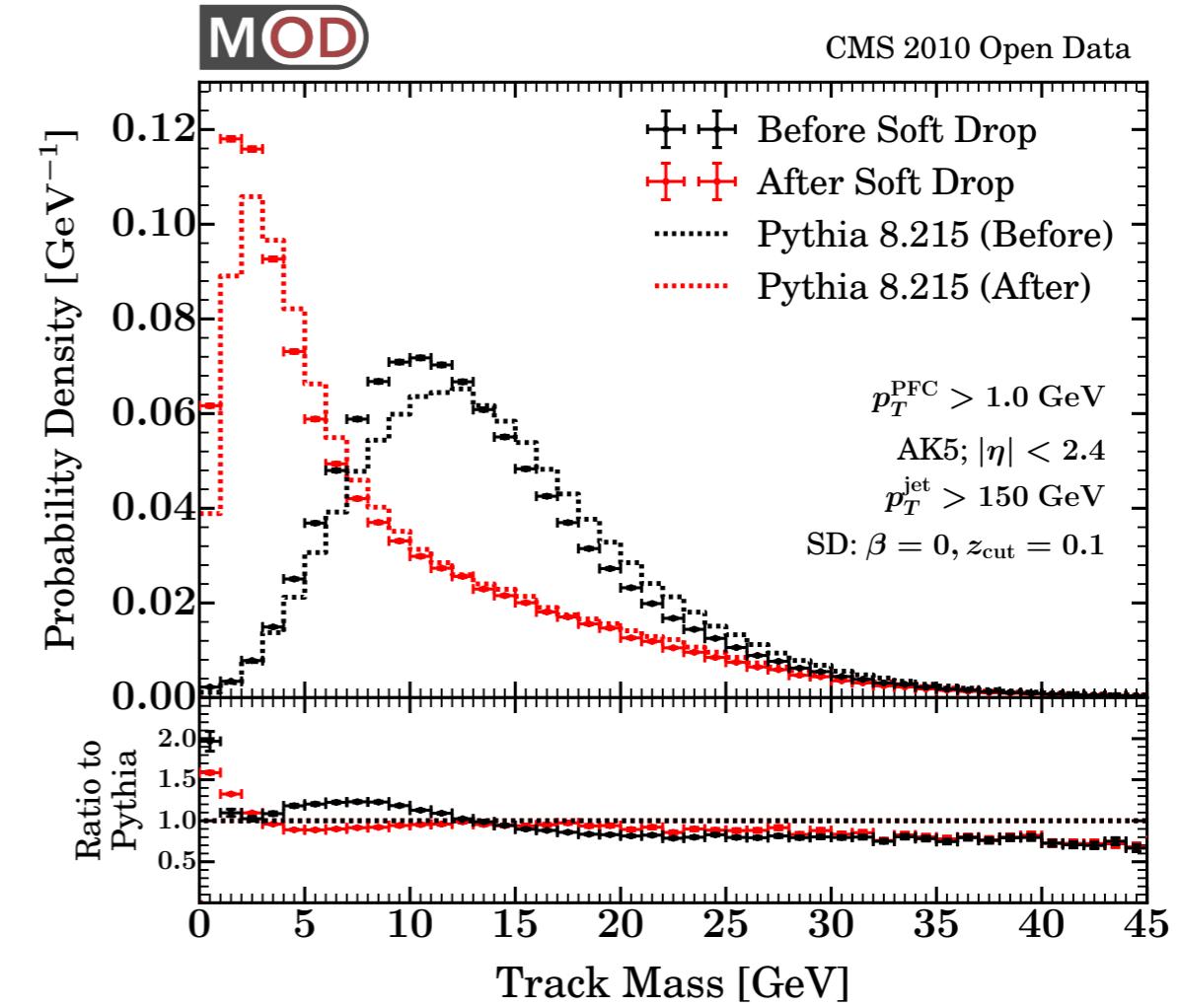
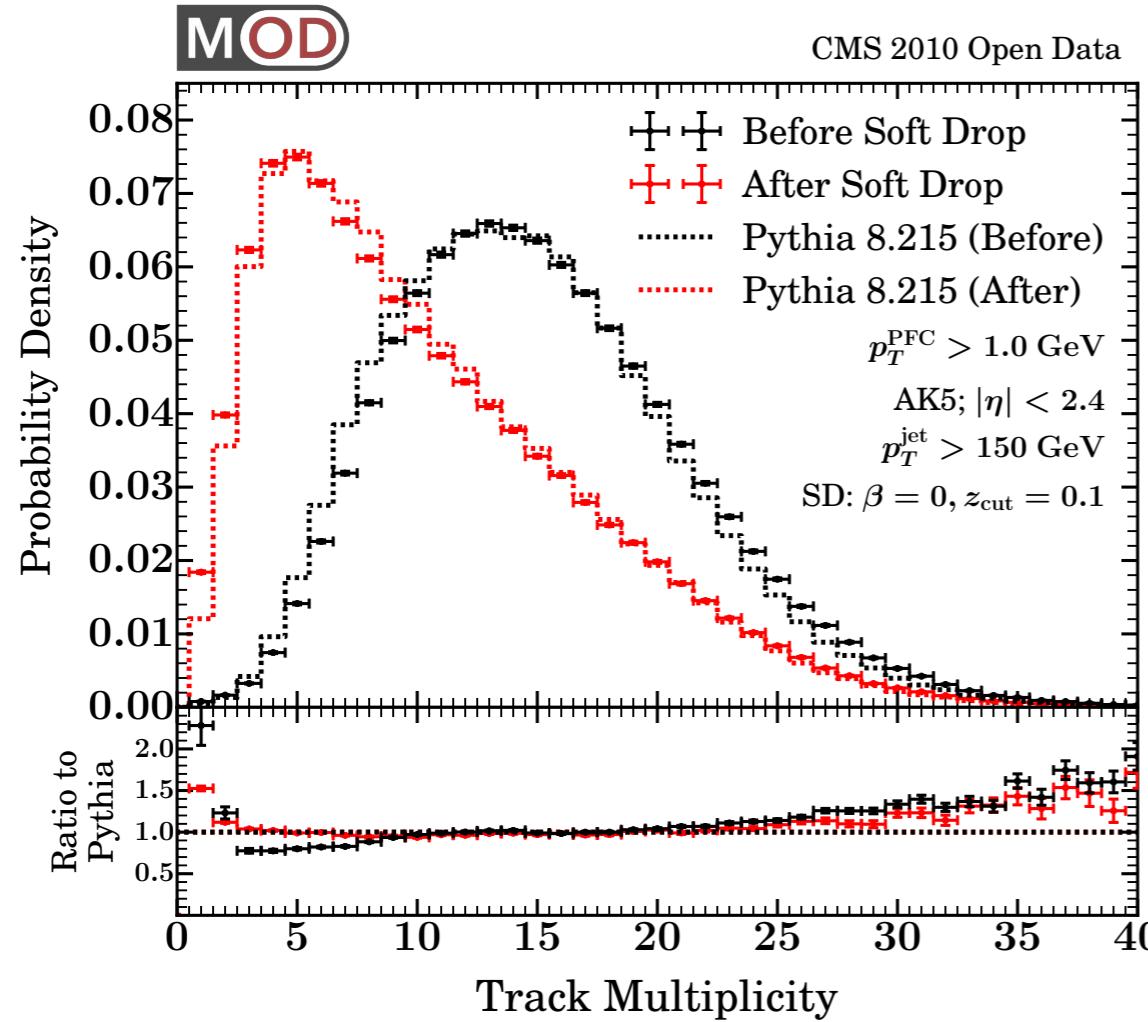
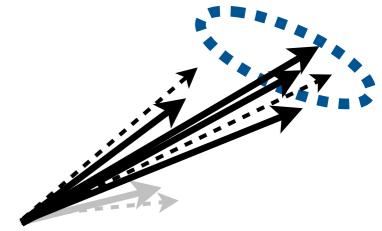
No grooming applied



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

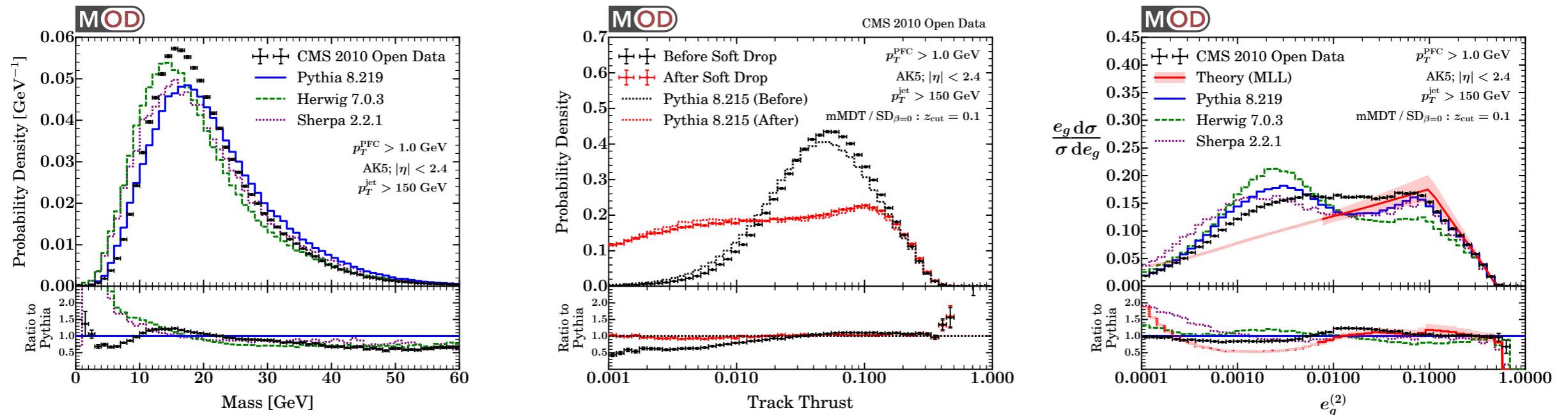
Track-Based Substructure

With and without soft drop grooming



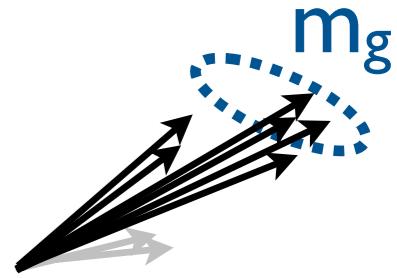
Jet grooming does not typically affect data/MC agreement

Explorations in Jet Substructure

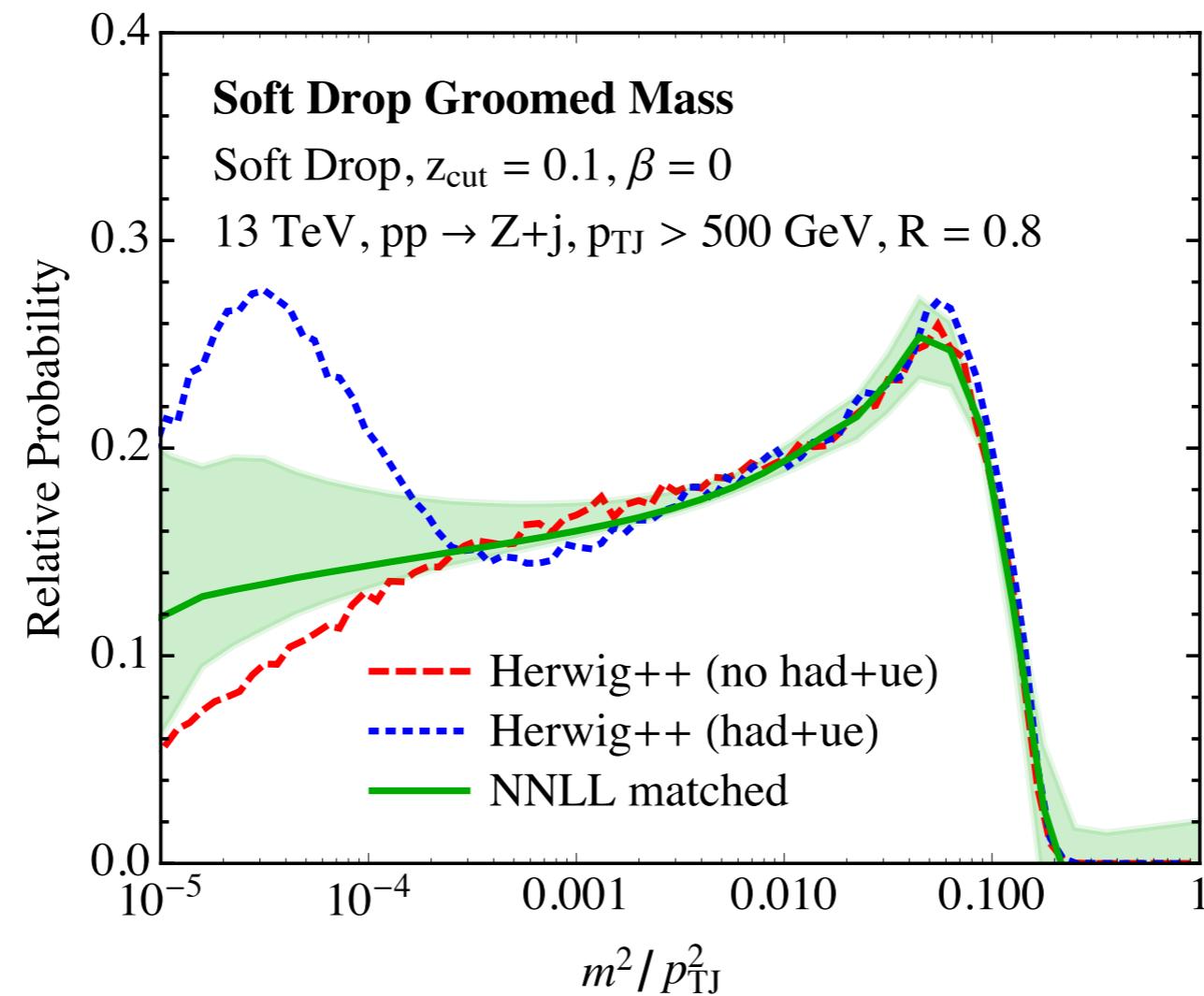


*Three different probes of jet mass
with different experimental and theoretical challenges*

Soft Drop Jet Mass



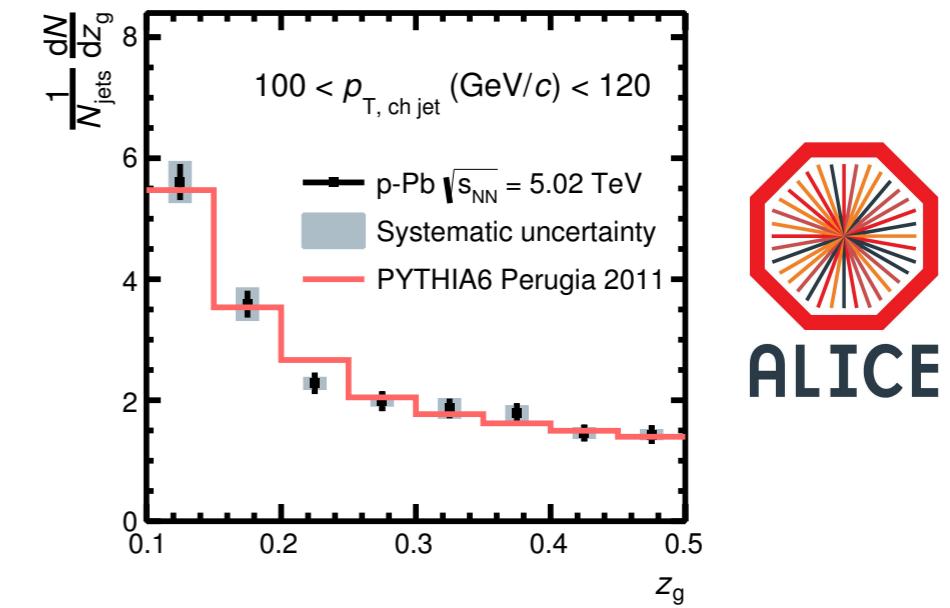
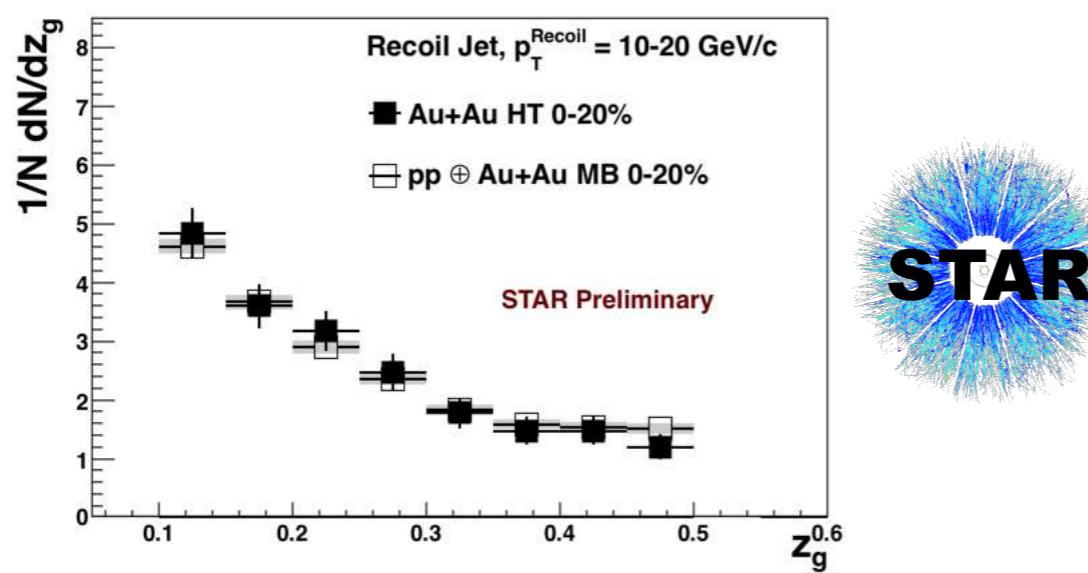
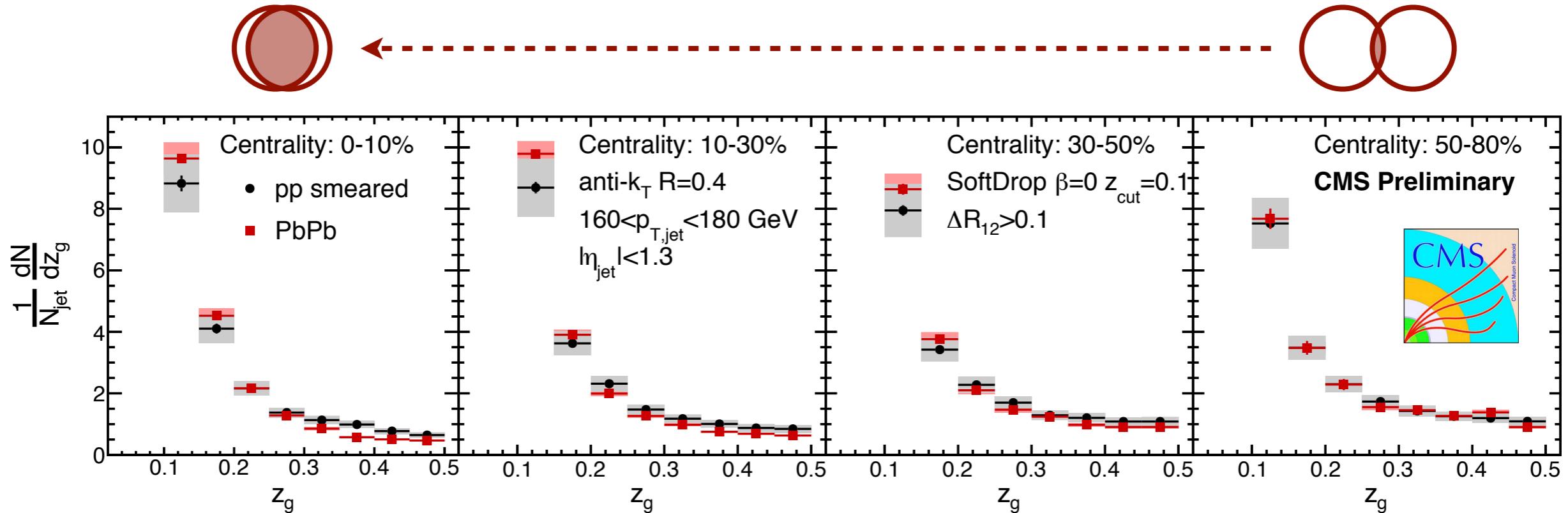
First NNLL + $\mathcal{O}(\alpha_s^2)$ result for substructure in pp (!)



Grooming *simplifies* structure of calculation, reduces NP effects

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

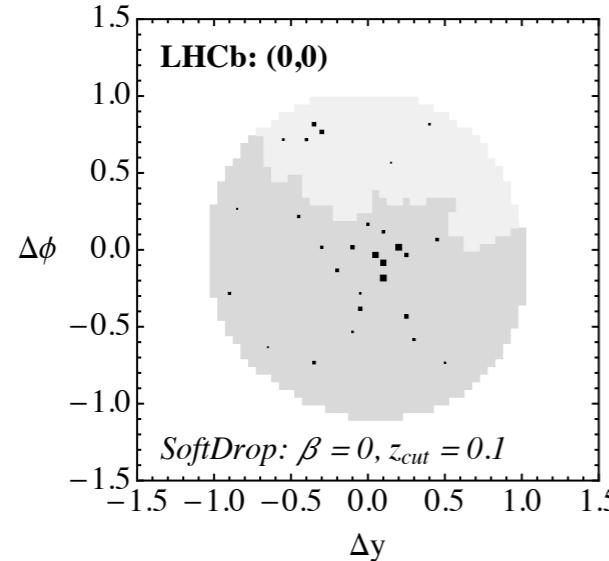
Preliminary Results from Heavy Ions



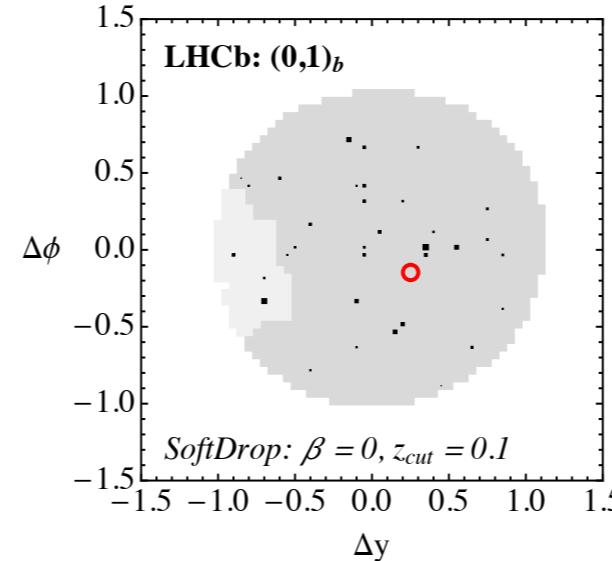
[CMS-PAS-HIN-16-006, STAR preliminary, ALICE preliminary]

Possibilities for Heavy Flavor

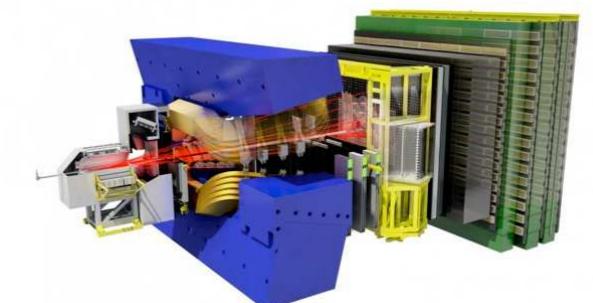
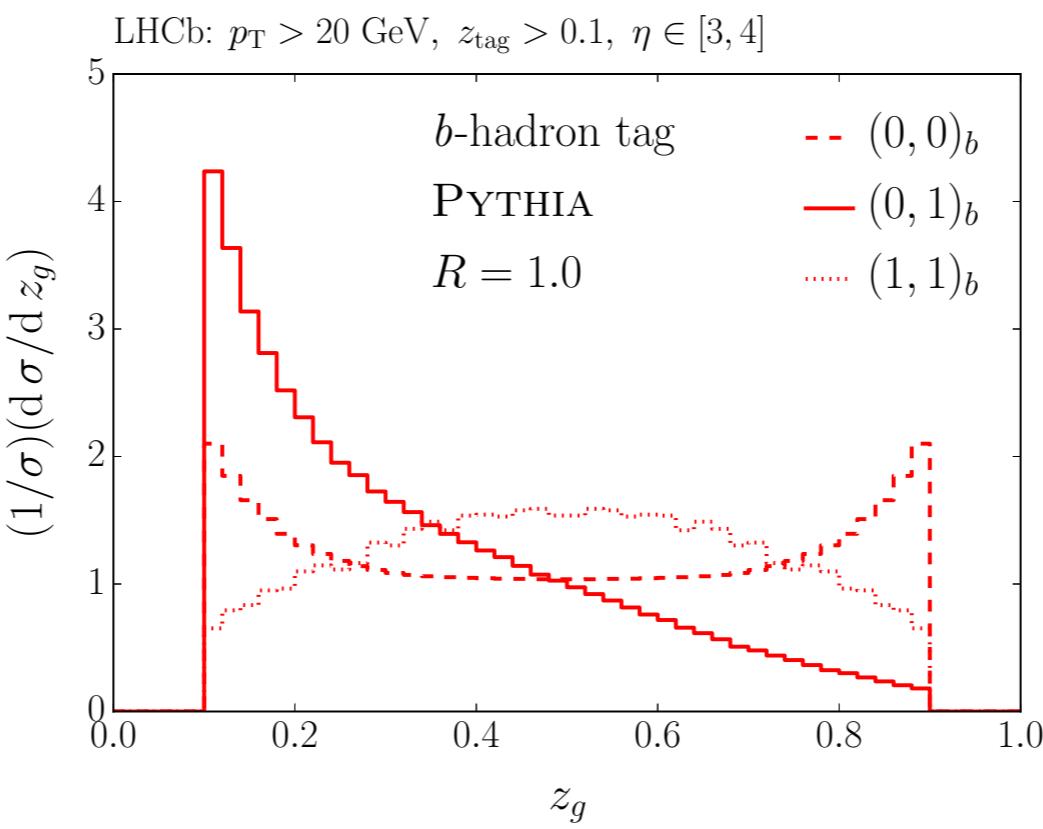
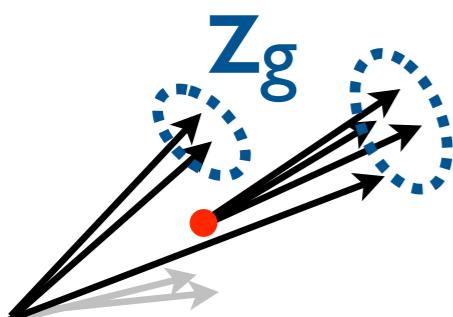
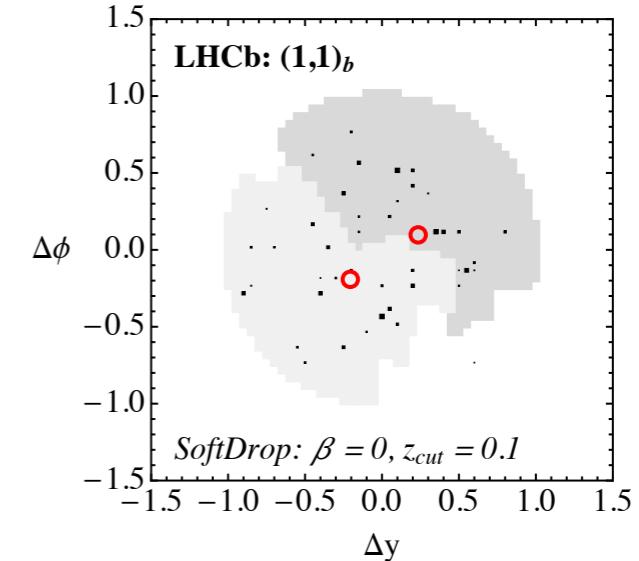
$g \rightarrow gg$



$b \rightarrow bg$

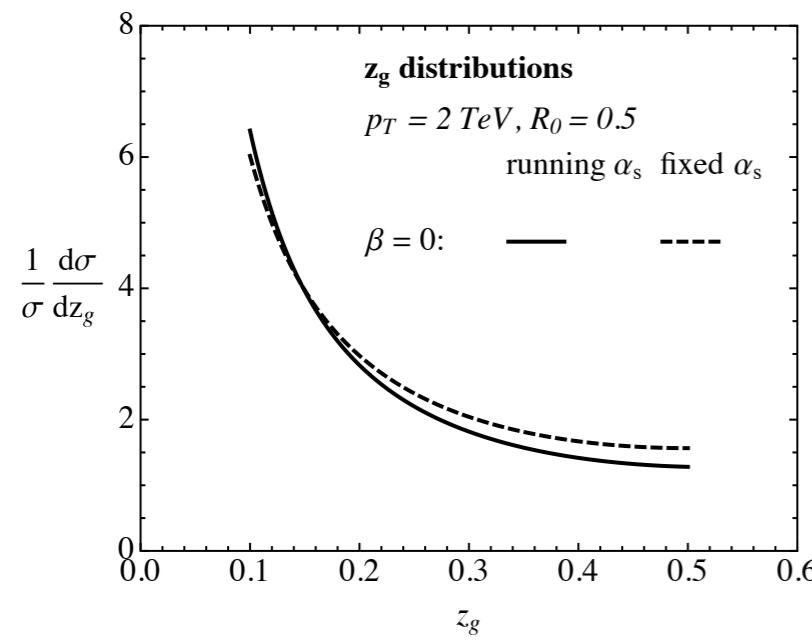


$g \rightarrow b\bar{b}$

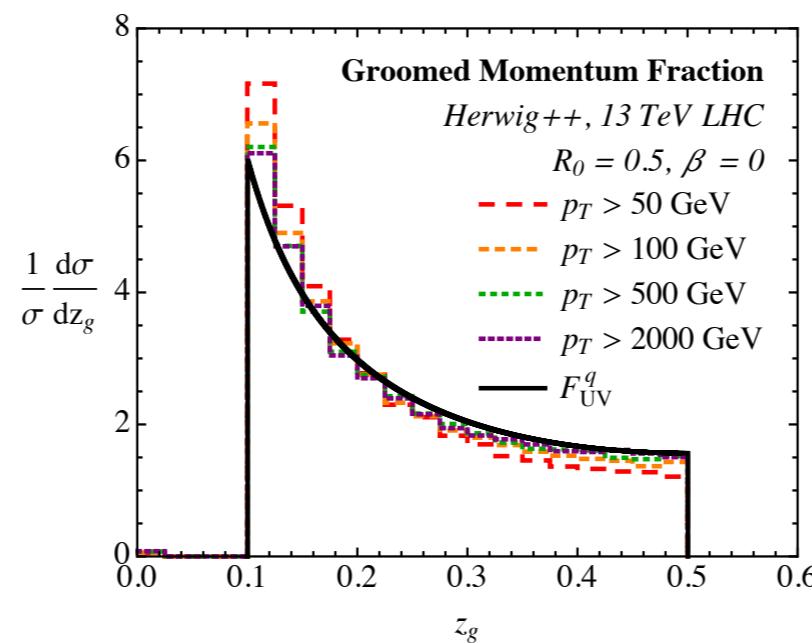


[Ilten, Rodd, JDT, Williams, I702.02947]

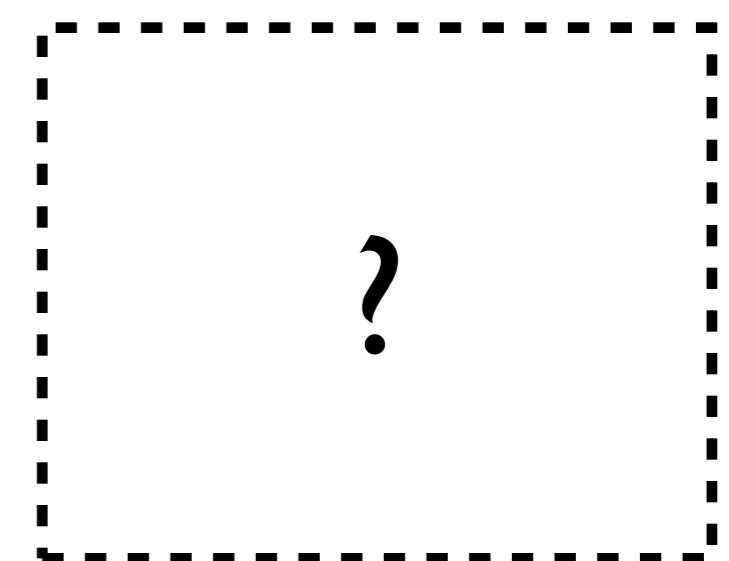
First-principles QCD



Parton Shower Study



Collider Data





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



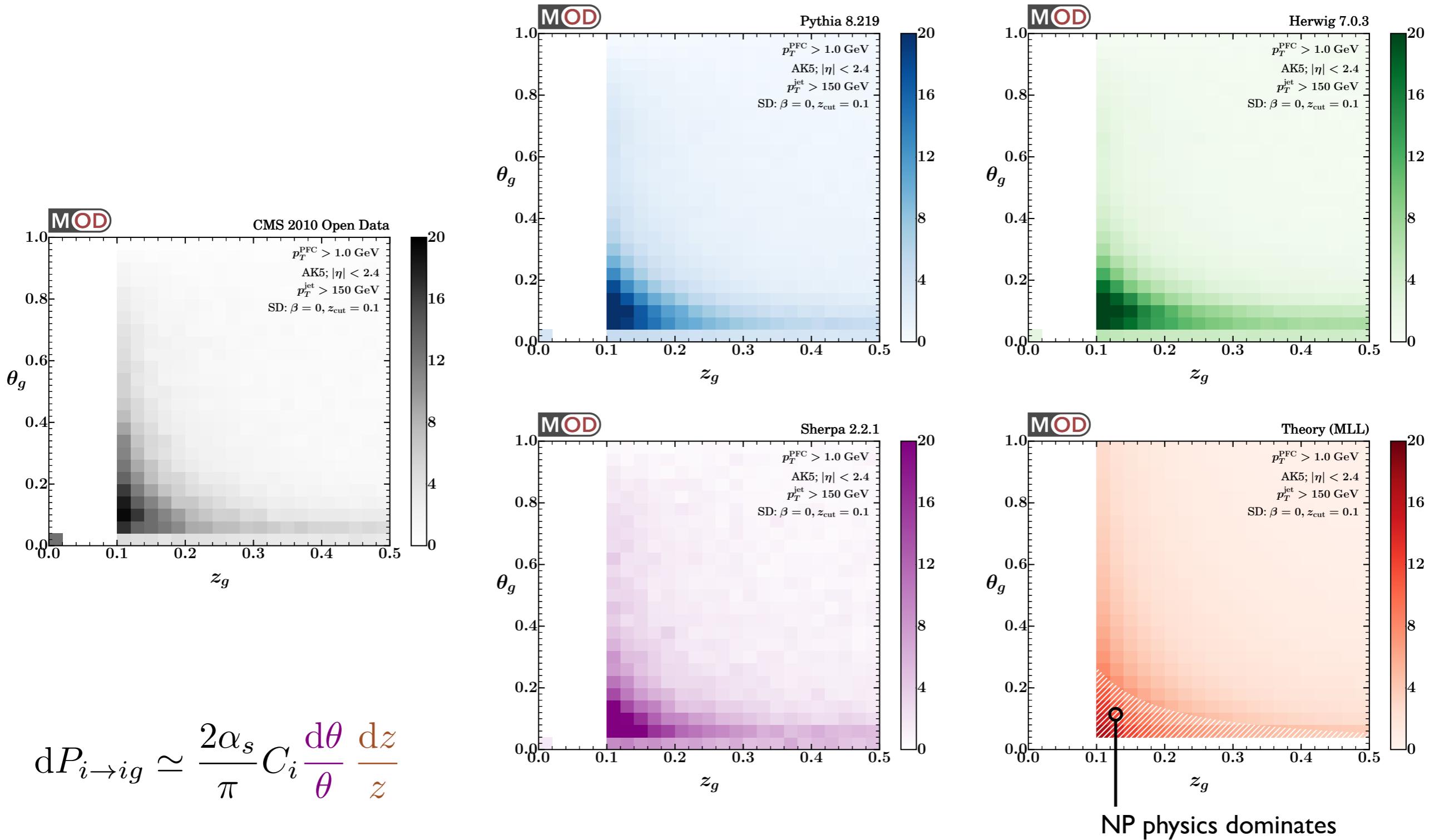
2



4

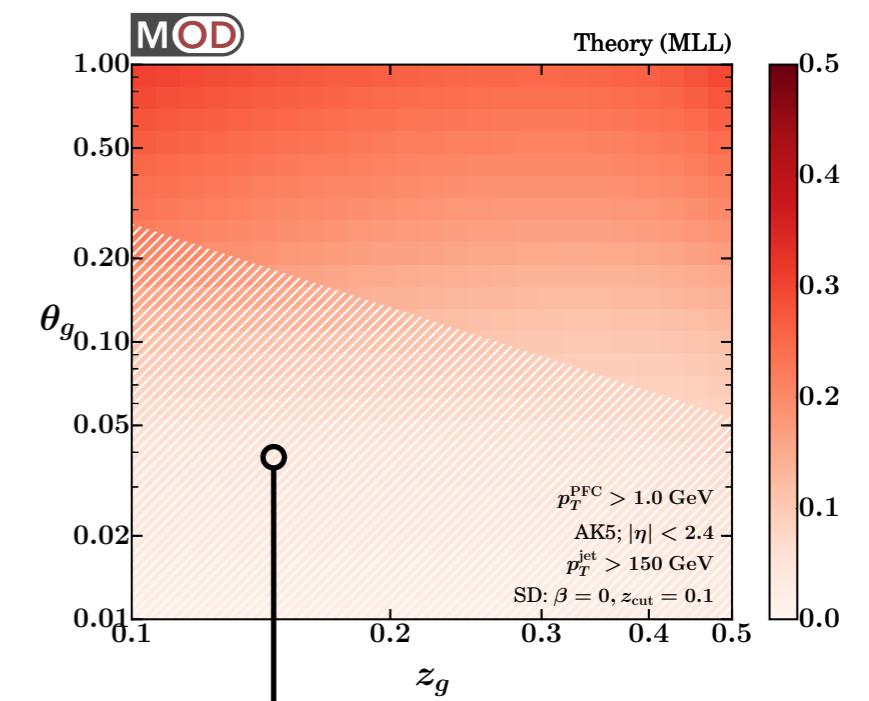
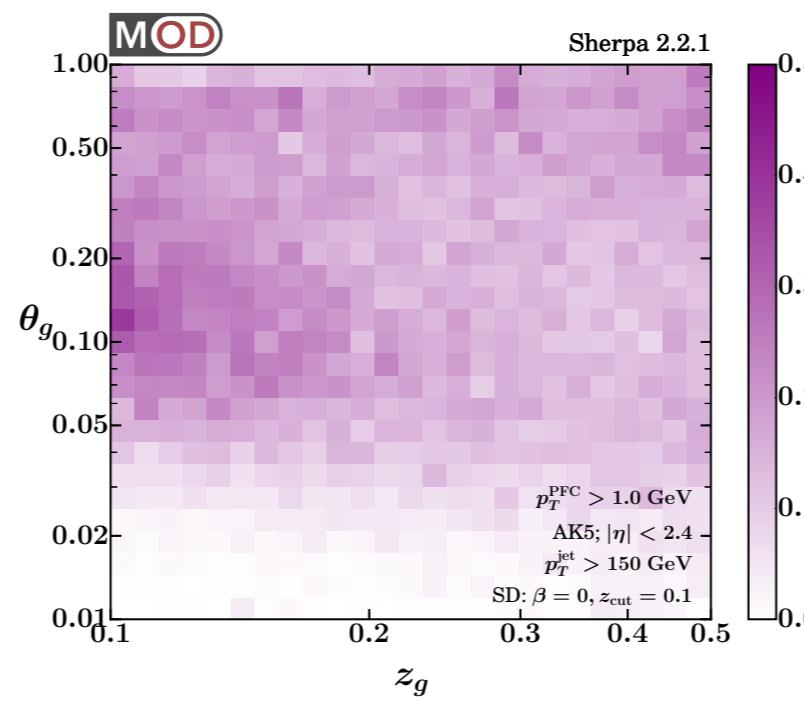
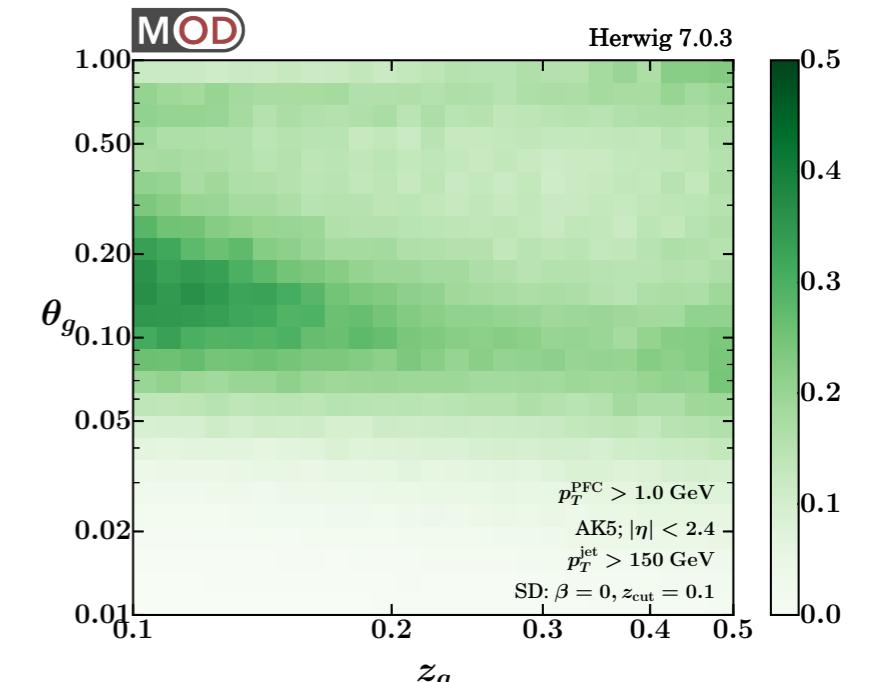
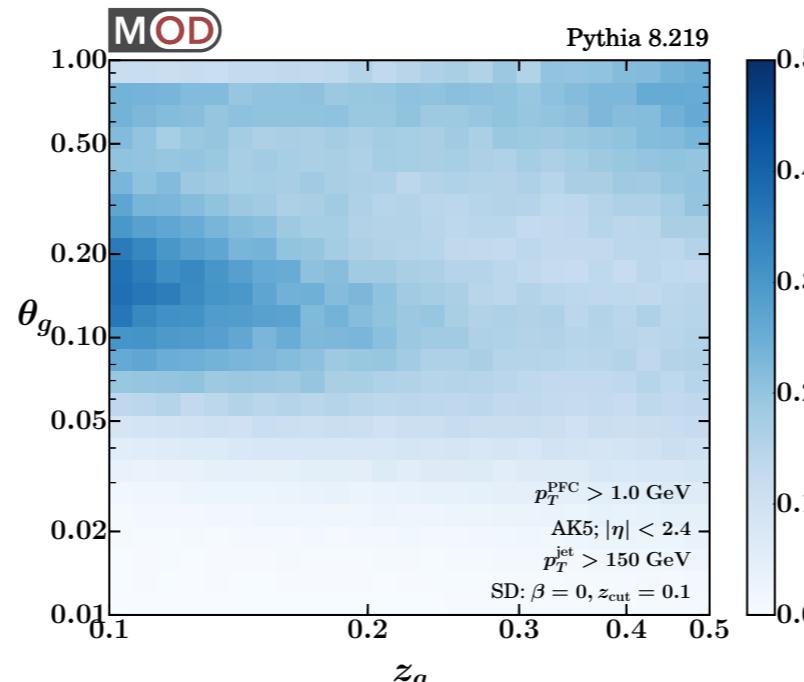
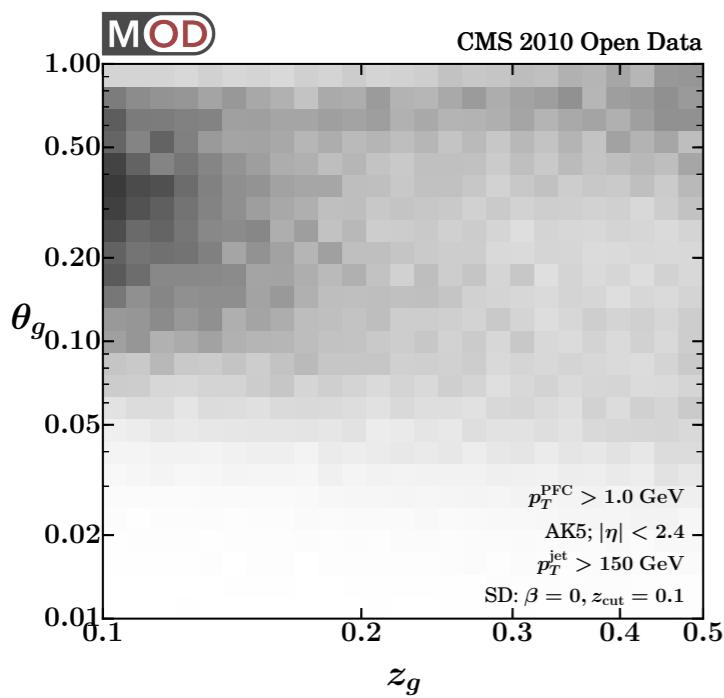
Visualizing the Singularity Structure of QCD

Linear scale



Visualizing the Singularity Structure of QCD

Logarithmic scale



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