Unbinned differential cross section measurements: *towards a common format*

Benjamin Nachman

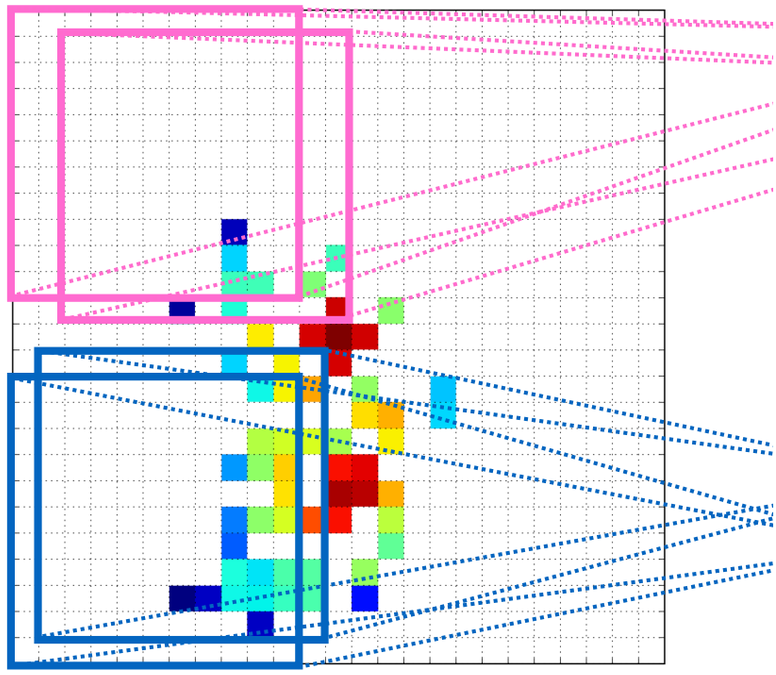
Lawrence Berkeley National Laboratory

bpnachman.com

bpnachman@lbl.gov

@bpnachman

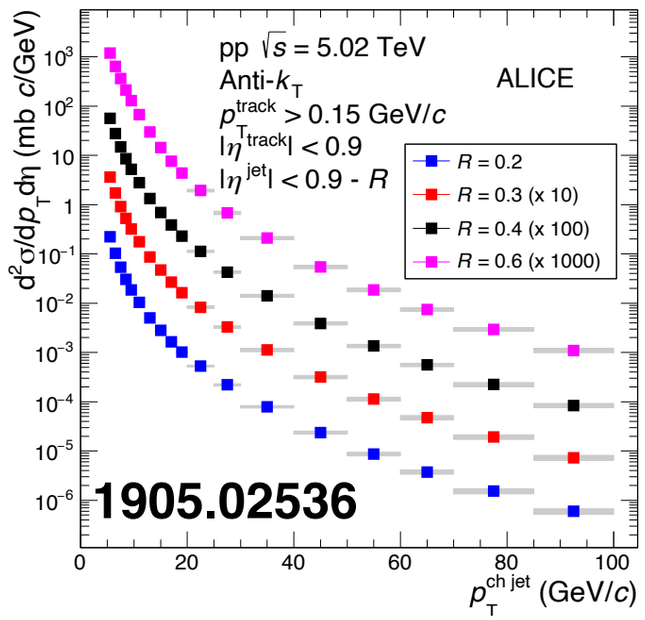
bnachman

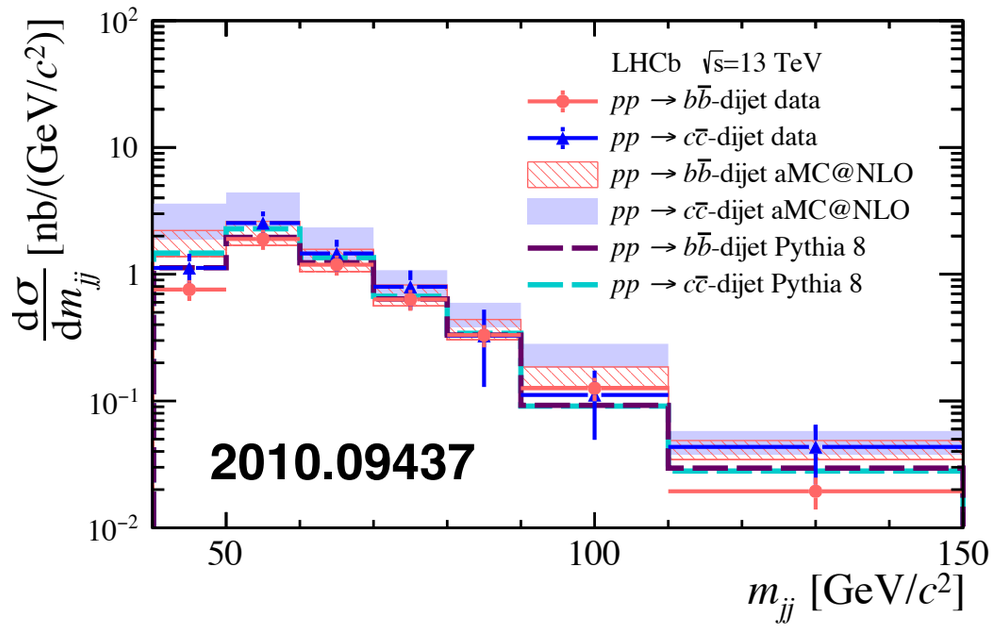


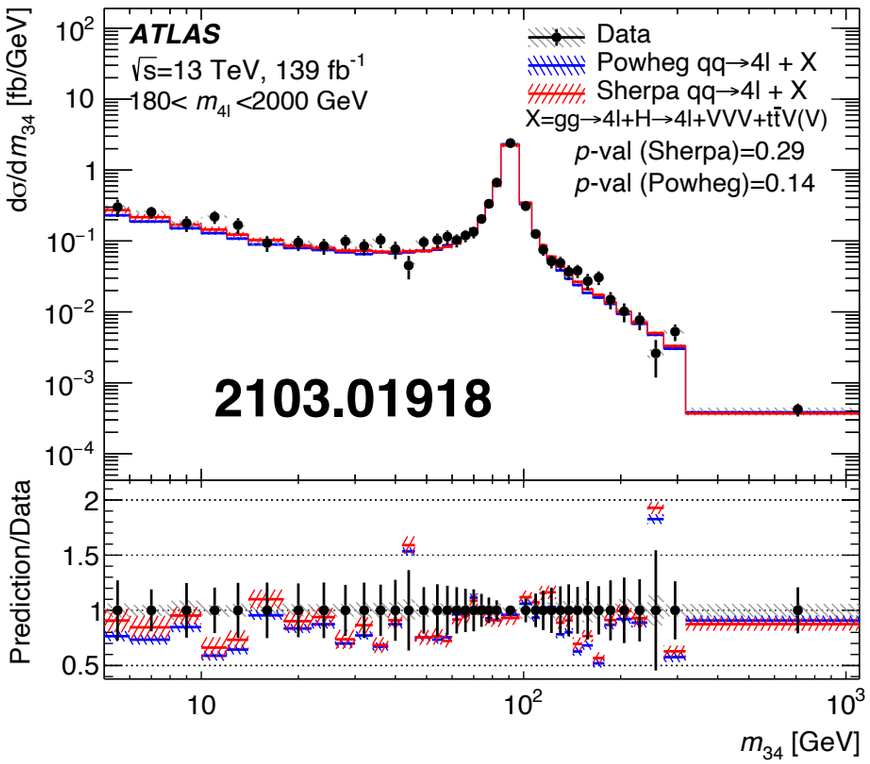
Virtual Houches

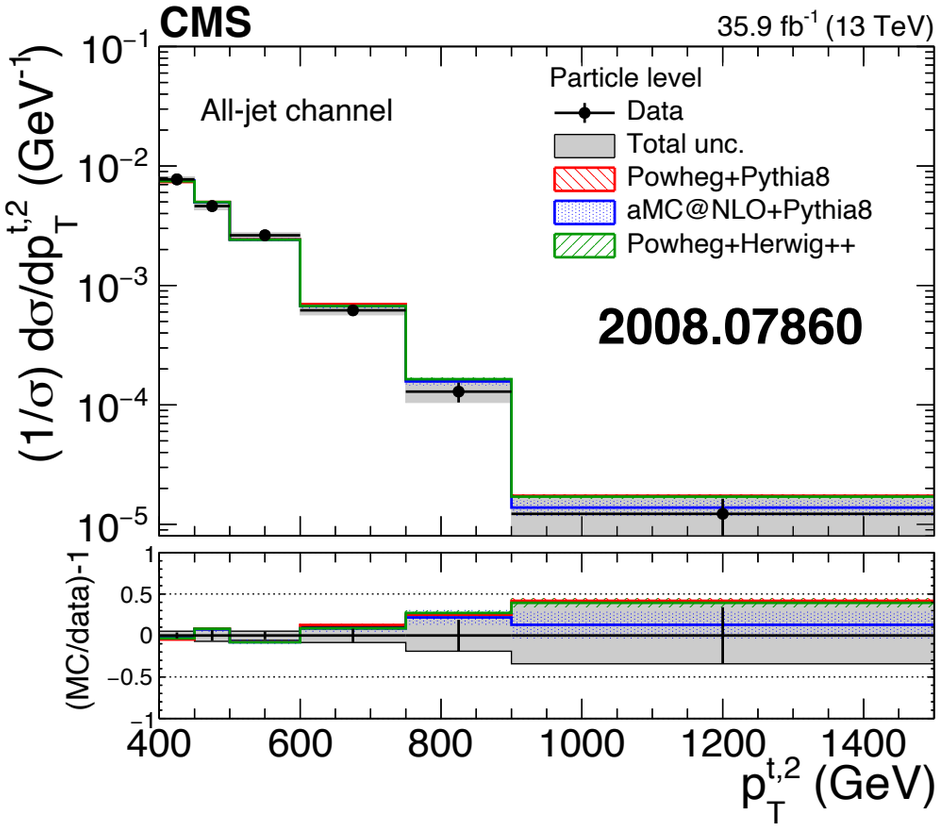
June 17, 2021

Differential Cross Section Measurements

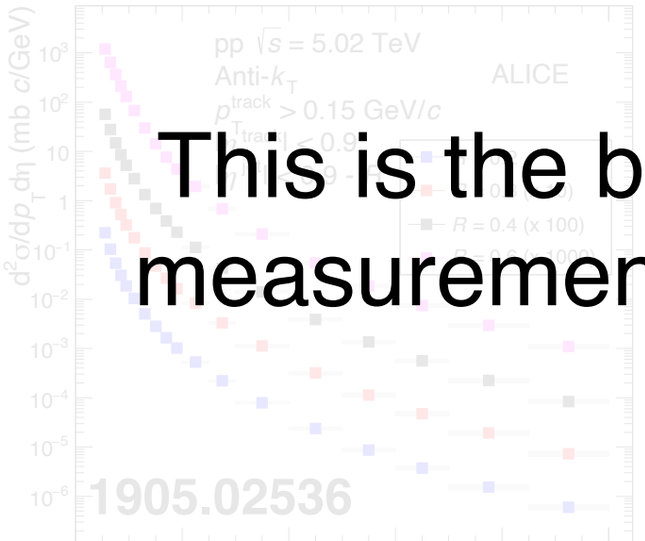


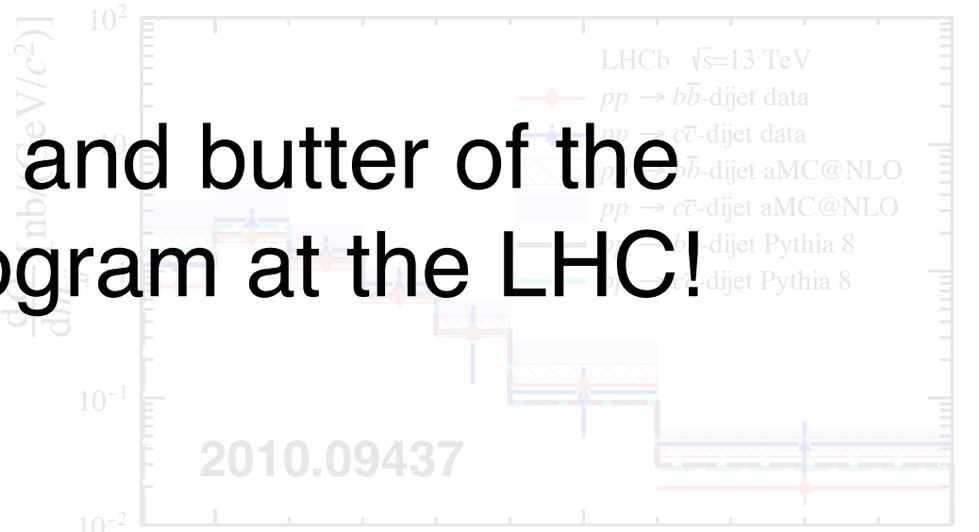






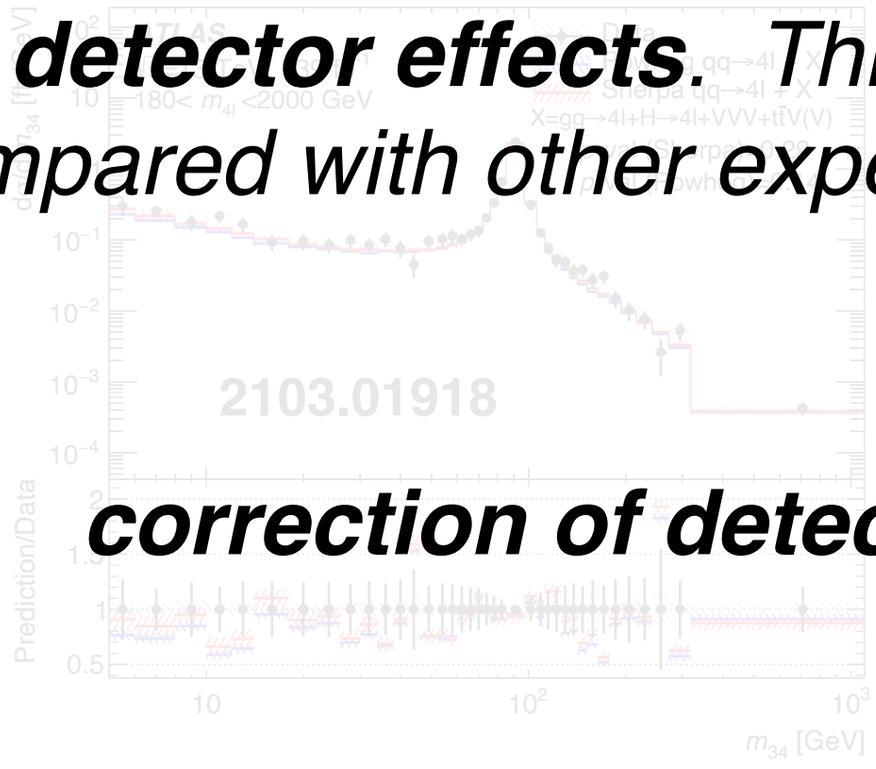
Differential Cross Section Measurements

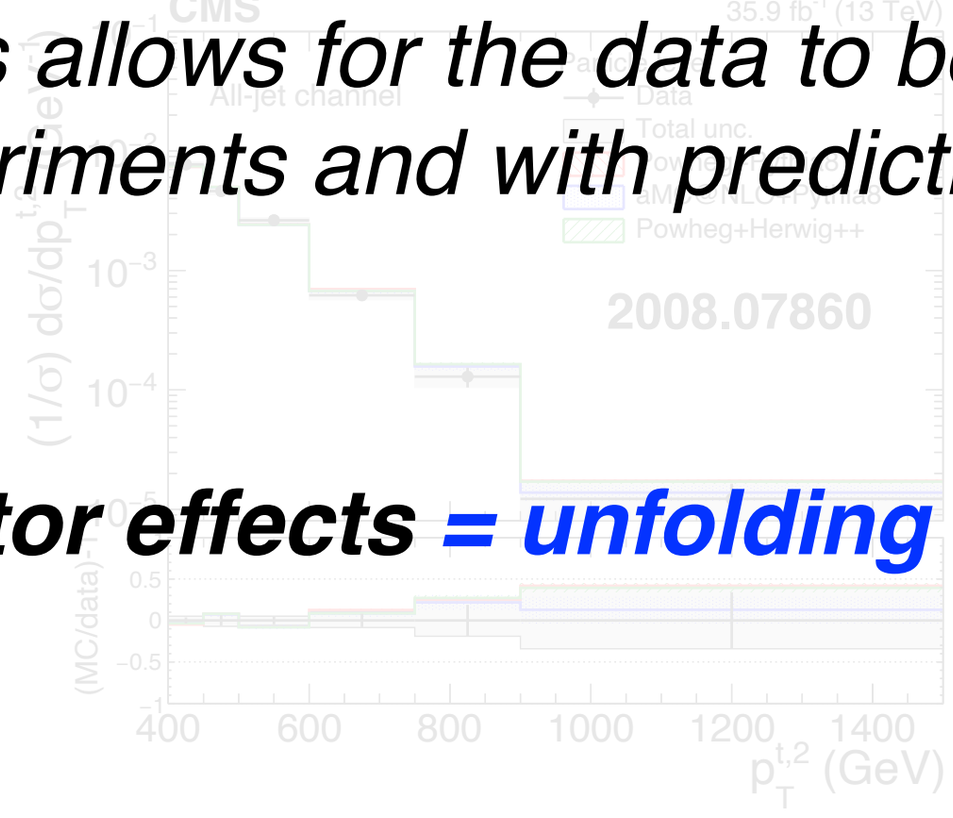




This is the bread and butter of the measurement program at the LHC!

*The key component of this is the **correction of detector effects**. This allows for the data to be compared with other experiments and with predictions.*

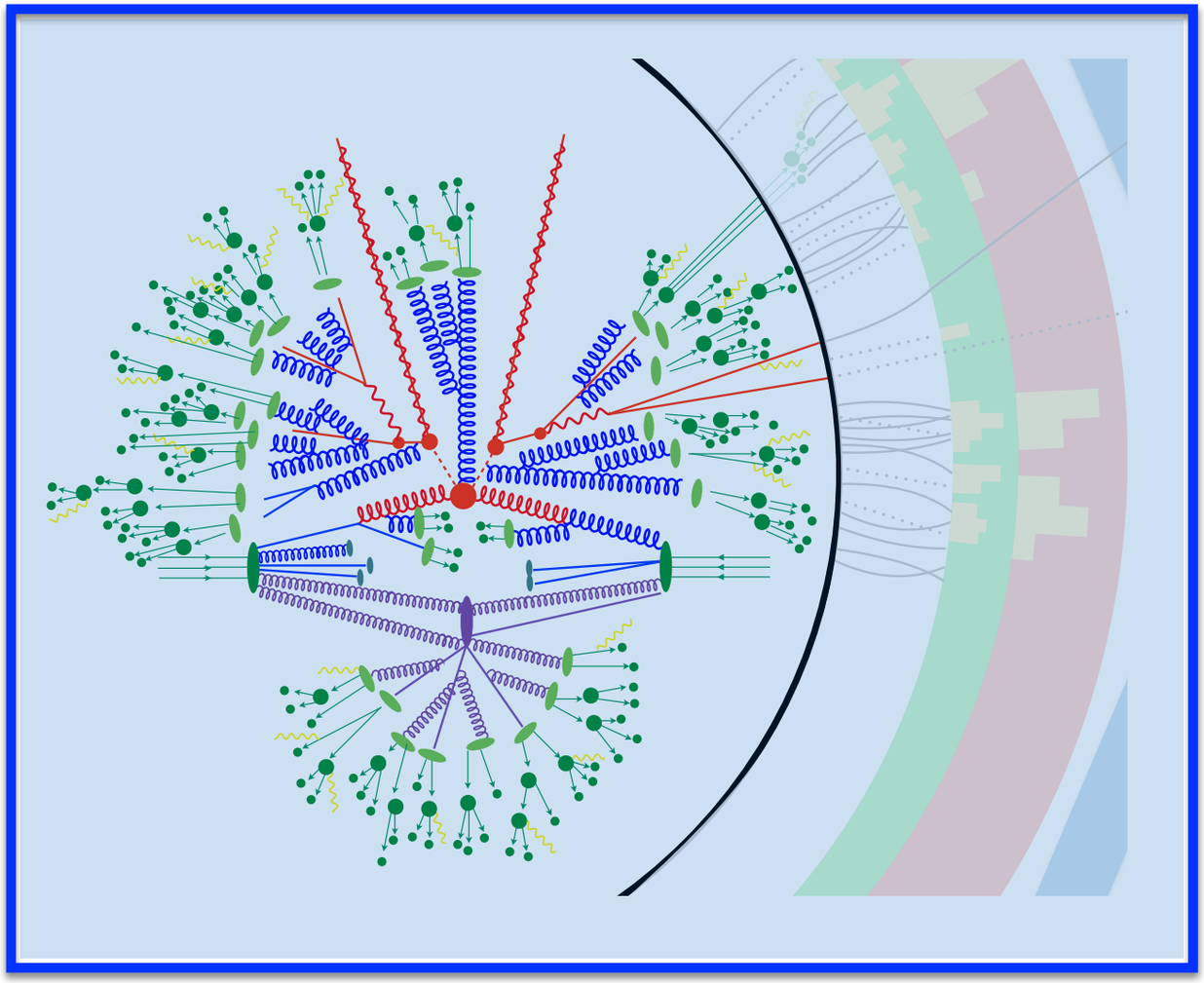




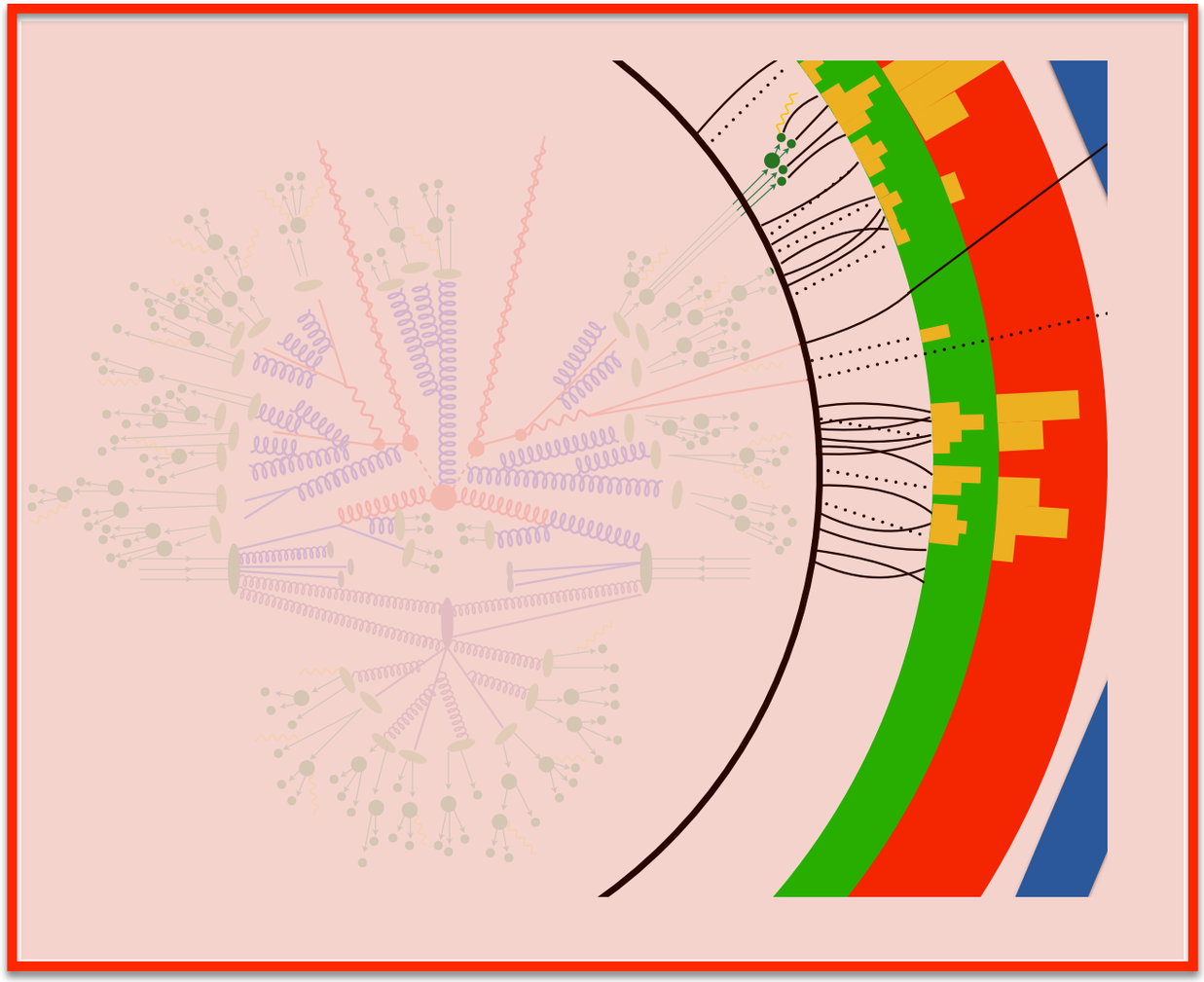
correction of detector effects* = *unfolding

The Unfolding Challenge

Want this

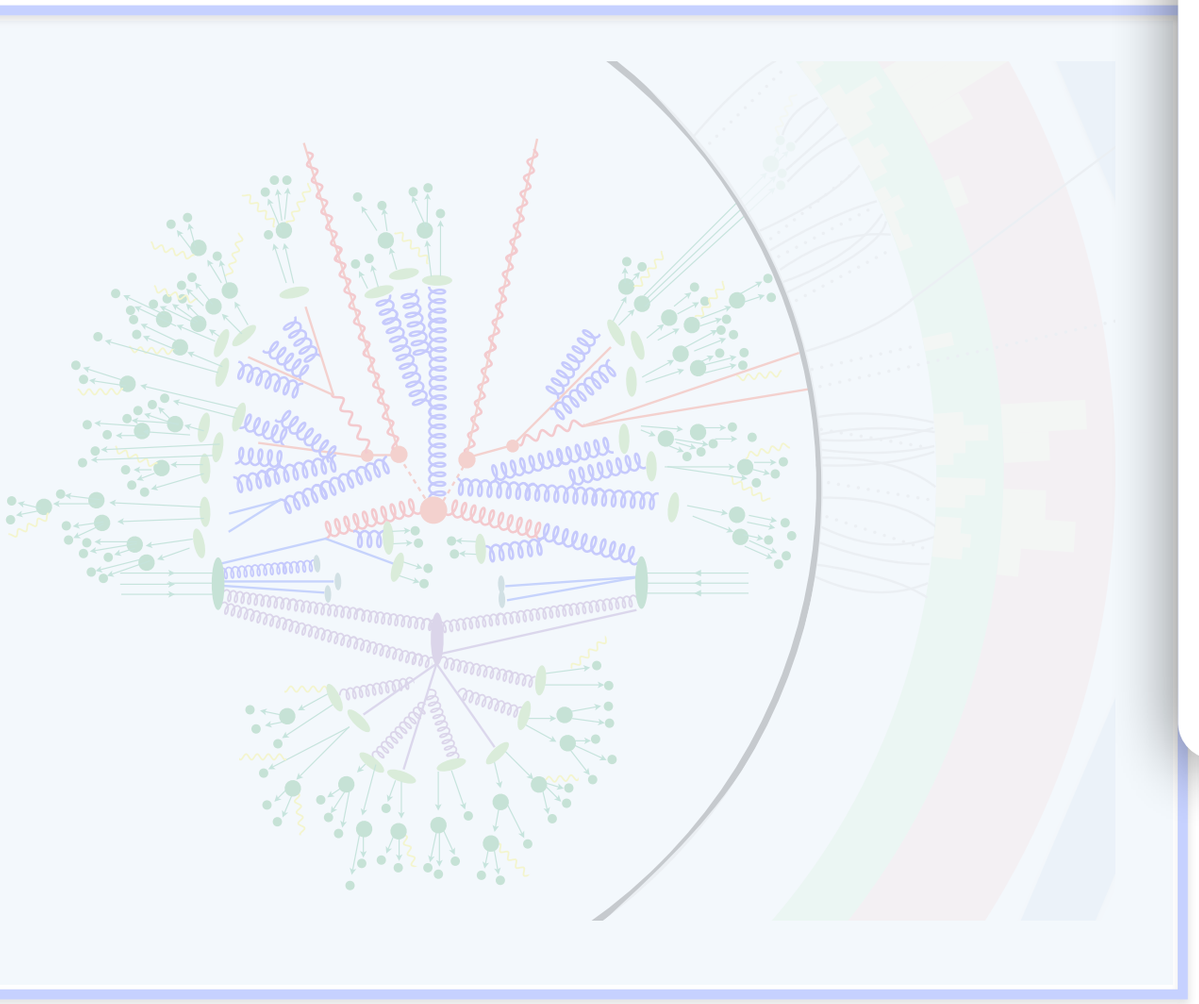


Measure this

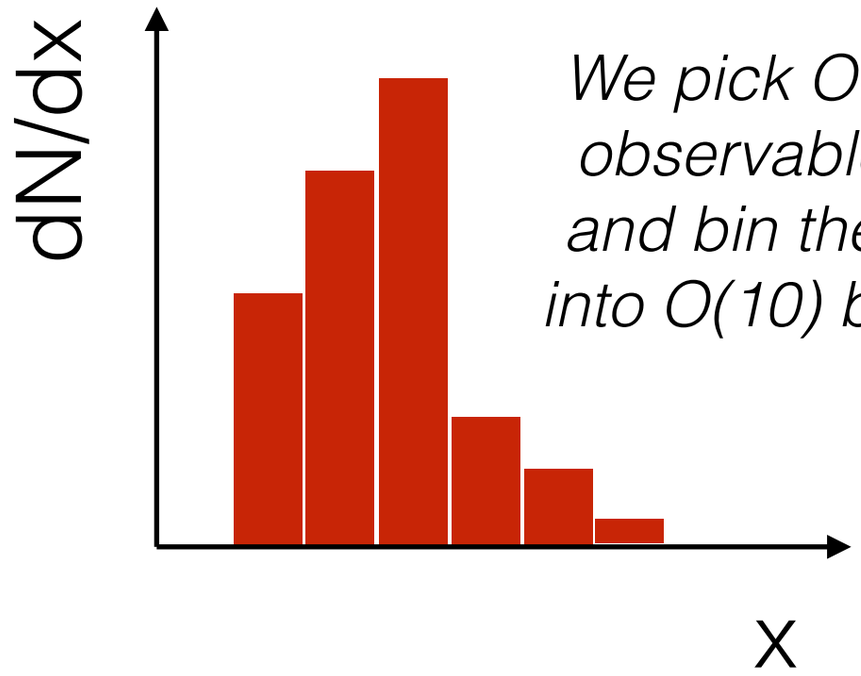


The Unfolding Challenge

Want this



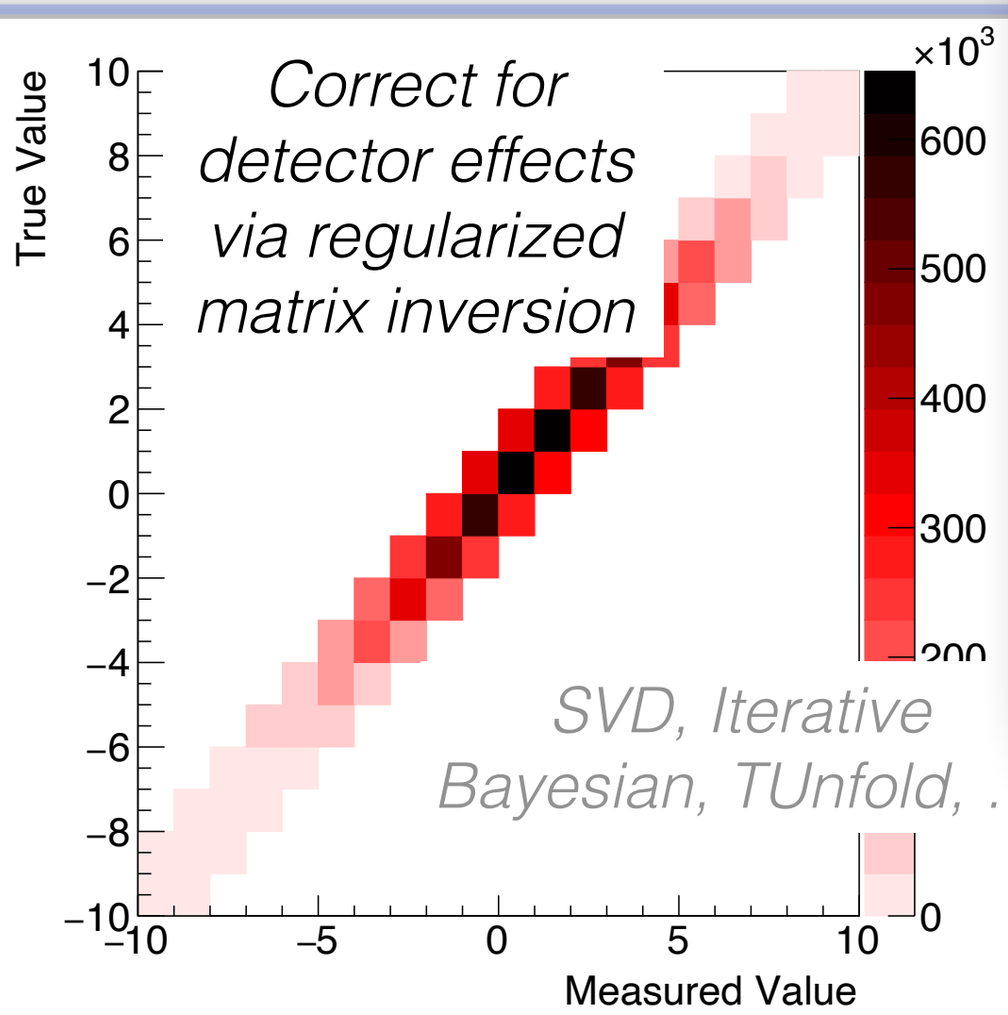
Usual solution:



We pick O(1) observables and bin them into O(10) bins

The Unfolding Challenge

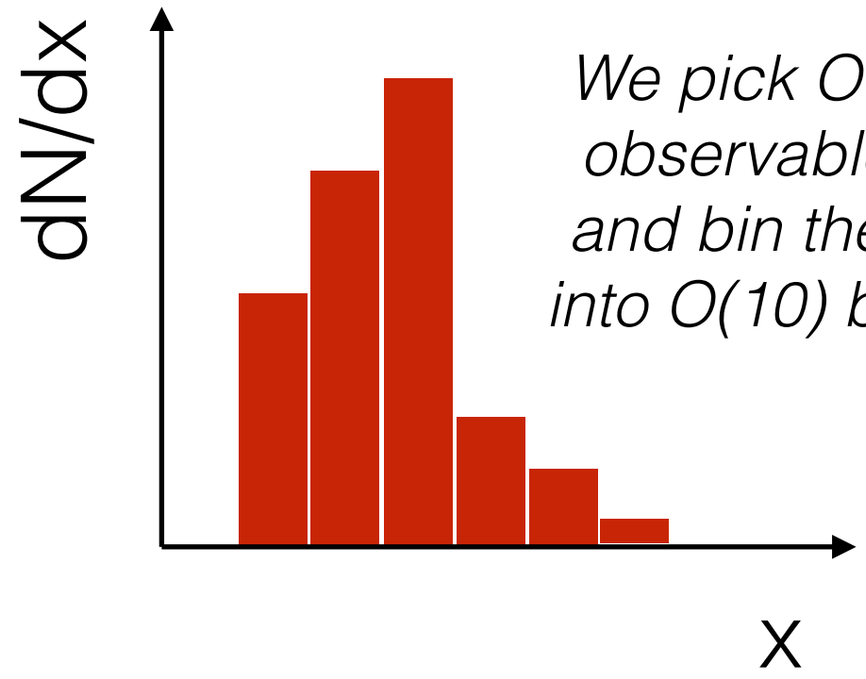
Want this



Correct for detector effects via regularized matrix inversion

SVD, Iterative Bayesian, TUnfold, …

Usual solution:



We pick O(1) observables and bin them into O(10) bins

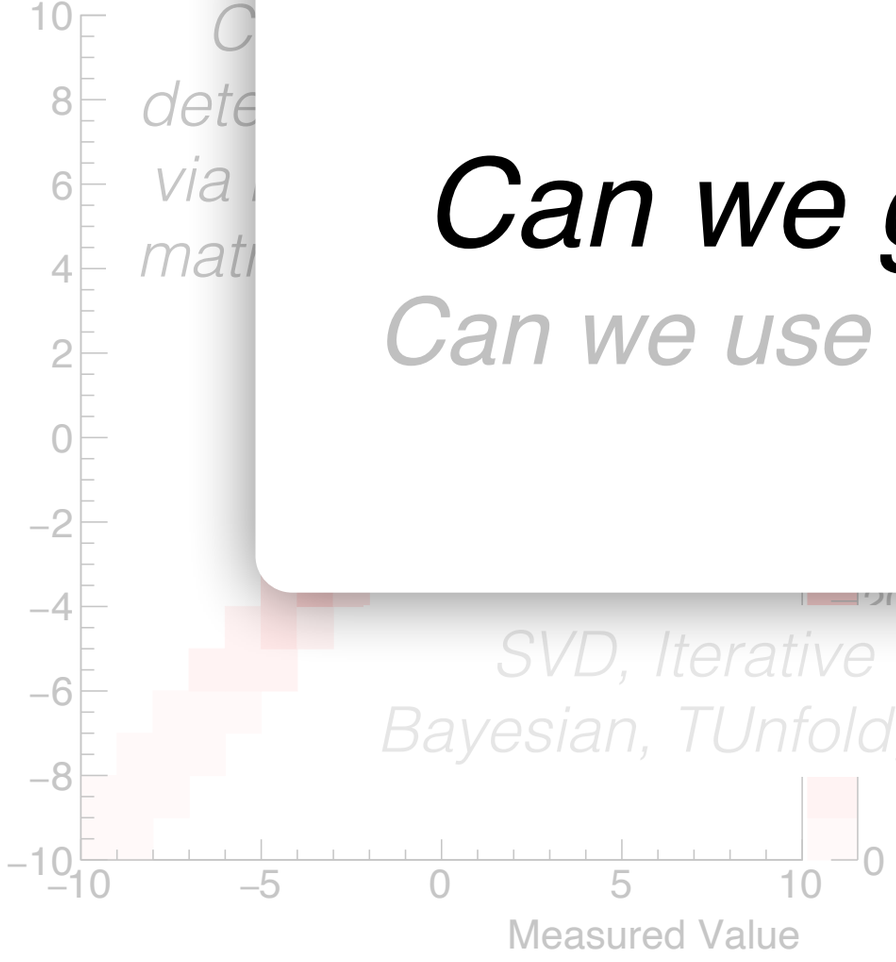
The Unfolding Challenge

Want this

Usual solution:

Can we go unbinned?

Can we use many dimensions?



SVD, Iterative
Bayesian, TUnfold, …

ick O(1)
ervables
bin them
O(10) bins

X

Unbinned Unfolding

There were some early proposals for unbinned unfolding* but as far as I am aware, they were not used for any measurements.

However, recent innovations in machine learning and resulted in new methods for unbinned unfolding, which are being used for data analysis+ (!)

The goal of this discussion is to propose a common way for publishing unbinned results to maximize their science potential

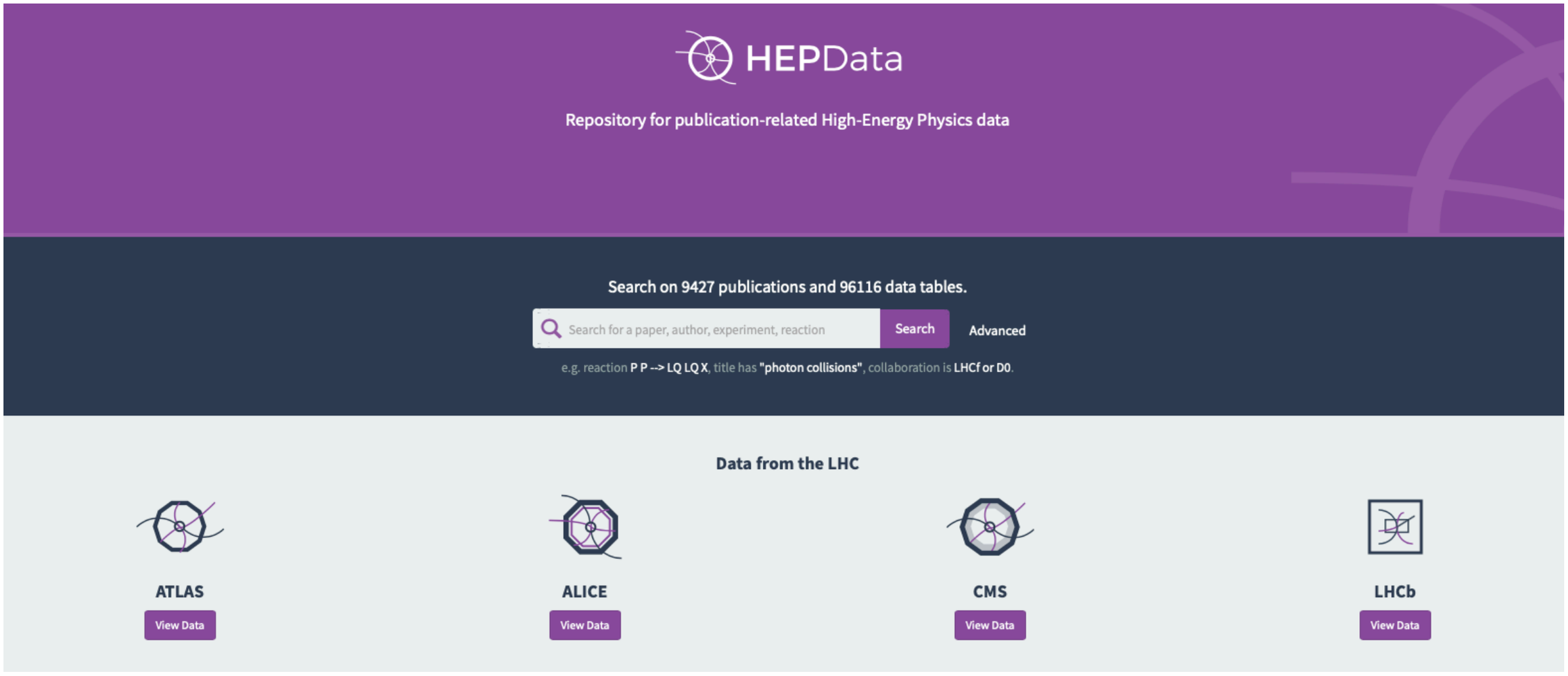
We need input from both experimentalist and theorists (!)

*see L. Lindemann and G. Zech, NIM A 354 (1995) 516 & related

+see https://www-h1.desy.de/h1/www/publications/htmlsplit/H1prelim-21-031.long.html

Publishing Binned Results

How do we publish binned results?



Publishing Binned Results

Hide Publication Information

Properties of $g \to b\bar{b}$ at small opening angles in pp collisions with the ATLAS detector at $\sqrt{s} = 13$ TeV

The ATLAS collaboration

Aaboud, Morad , Aad, Georges , Abbott, Brad , Abbott, Dale Charles , Abdinov, Ovsat , Abeloos, Baptiste , Abhayasinghe, Deshan Kavishka , Abidi, Syed Haider , Abouzeid, Ossama , Abraham, Nicola

Phys.Rev.D 99 (2019) 052004, 2019.

https://doi.org/10.17182/hepdata.85697

Journal | INSPIRE | Resources

Rivet Analysis

Abstract (data abstract)
CERN-LHC.

The fragmentation of high energy gluons at small opening angles is largely unconstrained by present measurements. Gluon splitting to b-quark pairs is a unique probe into the properties of gluon fragmentation because identified b-tagged jets provide a proxy for the quark daughters of the initial gluon. In this study, key differential distributions related to the g->bb process are measured using 33/fb of sqrt(s)=13 TeV pp collision data recorded by the ATLAS experiment at the LHC in 2016. Jets constructed from charged-particle tracks, clustered with the jet anti-kt algorithm with radius parameter R = 0.2, are used to probe angular scales below the R = 0.4 jet radius. The observables are unfolded to particle level in order to facilitate direct

Download All

View Analyses

Filter 4 data tables

Table 1
Data from Figure 6A
10.17182/hepdata.85697.v1/t1
Normalisaed differential cross section, $(1/\sigma_{\mathrm{fid}})d\sigma_{\mathrm{fid}}/d\Delta R(b,b)$, as a function of $\Delta R(b,b)$ - the angle in η and ϕ between...

Table 2
Data from Figure 6B
10.17182/hepdata.85697.v1/t2
Normalisaed differential cross section, $(1/\sigma_{\mathrm{fid}})d\sigma_{\mathrm{fid}}/d\Delta\theta_{\mathrm{gpp,gbb}}/\pi$, the angle between production (gpp) and decay (gbb) planes ($\Delta\theta_{\mathrm{gpp,gbb}}$).

Table 3
Data from Figure 6C
10.17182/hepdata.85697.v1/t3
Normalisaed differential cross section, $(1/\sigma_{\mathrm{fid}})d\sigma_{\mathrm{fid}}/dz(p_{\mathrm{T}})$, as a function of $z(p_{\mathrm{T}}) = p_{\mathrm{T},2}/(p_{\mathrm{T},1} + p_{\mathrm{T},2})$.

Table 4
Data from Figure 6D
10.17182/hepdata.85697.v1/t4
Normalized differential cross section, $(1/\sigma_{\mathrm{fid}})d\sigma_{\mathrm{fid}}/d\log(m_{bb}/p_{\mathrm{T}})$, as a function of $\log(m_{bb}/p_{\mathrm{T}})$ for m_{bb} the invariant mass of the two b-jets.

Table 1

10.17182/hepdata.85697.v1/t1

https://www.hepdata.net/rec | JSON

Data from Figure 6A

Normalisaed differential cross section, $(1/\sigma_{\mathrm{fid}})d\sigma_{\mathrm{fid}}/d\Delta R(b,b)$, as a function of $\Delta R(b,b)$ - the angle in η and ϕ between the two b-tagged jets.

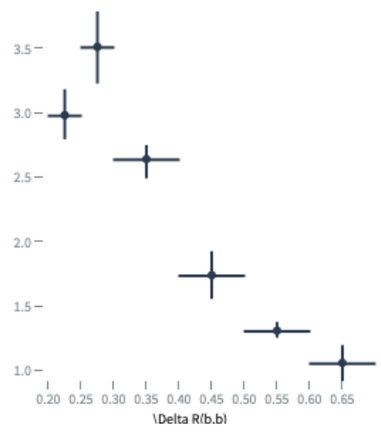
cmenergies: 13000.0

observables: DSIG/DDR

reactions: P P --> g + X | g --> b + b

RE	P P --> g < b b > X
SQRT(S)	13000.0 GEV
$\Delta R(b,b)$	$(1/\sigma_{\mathrm{fid}})d\sigma_{\mathrm{fid}}/d\Delta R(b,b)$
0.2 - 0.25	2.98 ±0.0016 stat +0.076 -0.019 sys,Calorimeterjetenergy +0.041 -0.042 sys,Flavortagging + 4 more errors Show all
0.25 - 0.3	3.51 ±0.0019 stat +0.062 -0.1 sys,Calorimeterjetenergy +0.034 -0.034 sys,Flavortagging + 4 more errors Show all
0.3 - 0.4	2.64 ±0.0017 stat +0.02 -0.089 sys,Calorimeterjetenergy +0.021 -0.018 sys,Flavortagging + 4 more errors Show all
0.4 - 0.5	1.74 ±0.0018 stat +0.039 -0.024 sys,Calorimeterjetenergy +0.0092 -0.0084 sys,Flavortagging + 4 more errors Show all
0.5 - 0.6	1.31 ±0.0013 stat +0.036 -0.0067 sys,Calorimeterjetenergy +0.013 -0.013 sys,Flavortagging + 4 more errors Show all
0.6 - 0.7	1.06 ±0.0013 stat +0.018 -0.026 sys,Calorimeterjetenergy +0.014 -0.017 sys,Flavortagging + 4 more errors Show all

Visualize



Sum errors ☑ Log Scale (X) ☐ Log Scale (Y) ☐

Publishing Binned Results

Hide Publication Information

Properties of $g \to b\bar{b}$ at small opening angles in pp collisions with the ATLAS detector at $\sqrt{s} = 13$ TeV

The ATLAS collaboration

Aaboud, Morad , Aad, Georges , Abbott, Brad , Abbott, Dale Charles , Abdinov, Ovsat , Abeloos, Baptiste , Abhayasinghe, Deshan Kavishka , Abidi, Syed Haider , Abouzeid, Ossama , Abraham, Nicola

Phys.Rev.D 99 (2019) 052004, 2019.

https://doi.org/10.17182/hepdata.85697

Journal | INSPIRE | Resources

Rivet Analysis

Abstract (data abstract)
CERN-LHC.

The fragmentation of high energy gluons at small opening angles is largely unconstrained by present measurements. Gluon splitting to b-quark pairs is a unique probe into the properties of gluon fragmentation because identified b-tagged jets provide a proxy for the quark daughters of the initial gluon. In this study, key differential distributions related to the g->bb process are measured using 33/fb of sqrt(s)=13 TeV pp collision data recorded by the ATLAS experiment at the LHC in 2016. Jets constructed from charged-particle tracks, clustered with the jet anti-kt algorithm with radius parameter R = 0.2, are used to probe angular scales below the R = 0.4 jet radius. The observables are unfolded to particle level in order to facilitate direct

Download All

View Analyses

Filter 4 data tables

Table 1
Data from Figure 6A
10.17182/hepdata.85697.v1/t1
Normalisaed differential cross section, $(1/\sigma_{\rm fid})d\sigma_{\rm fid}/d\Delta R(b,b)$, as a function of $\Delta R(b,b)$ - the angle in η and ϕ between...

Table 2
Data from Figure 6B
10.17182/hepdata.85697.v1/t2
Normalisaed differential cross section, $(1/\sigma_{\rm fid})d\sigma_{\rm fid}/d\Delta\theta_{\rm gpp,gbb}/\pi$, the angle between production (gpp) and decay (gbb) planes ($\Delta\theta_{\rm gpp,gbb}$).

Table 3
Data from Figure 6C
10.17182/hepdata.85697.v1/t3
Normalisaed differential cross section, $(1/\sigma_{\rm fid})d\sigma_{\rm fid}/dz(p_{\rm T})$, as a function of $z(p_{\rm T}) = p_{\rm T,2}/(p_{\rm T,1} + p_{\rm T,2})$.

Table 4
Data from Figure 6D
10.17182/hepdata.85697.v1/t4
Normalized differential cross section, $(1/\sigma_{\rm fid})d\sigma_{\rm fid}/d\log(m_{b\bar{b}}/p_{\rm T})$, as a function of $\log(m_{b\bar{b}}/p_{\rm T})$ for $m_{b\bar{b}}$ the invariant mass of the two b-jets.

Table 1 10.17182/hepdata.85697.v1/t1
Data from Figure 6A

https://www.hepdata.net/rec | JSON

Normalisaed differential cross section, $(1/\sigma_{\rm fid})d\sigma_{\rm fid}/d\Delta R(b,b)$, as a function of $\Delta R(b,b)$ - the angle in η and ϕ between the two b-tagged jets.

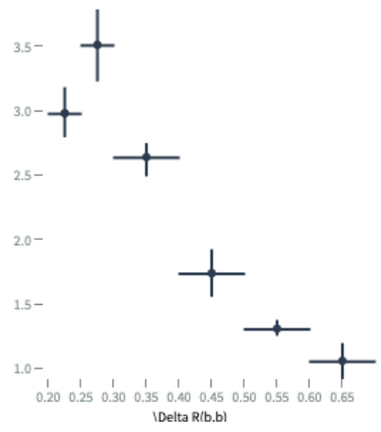
cmenergies: 13000.0

observables: DSIG/DDR

reactions: P P --> g + X | g --> b + b

RE	P P --> g < b b > X
SQRT(S)	13000.0 GEV
$\Delta R(b,b)$	$(1/\sigma_{\rm fid})d\sigma_{\rm fid}/d\Delta R(b,b)$
0.2 - 0.25	2.98 ±0.0016 stat +0.076 -0.019 sys,Calorimeterjetenergy +0.041 -0.042 sys,Flavortagging + 4 more errors Show all
0.25 - 0.3	3.51 ±0.0019 stat +0.062 -0.1 sys,Calorimeterjetenergy +0.034 -0.034 sys,Flavortagging + 4 more errors Show all
0.3 - 0.4	2.64 ±0.0017 stat +0.02 -0.089 sys,Calorimeterjetenergy +0.021 -0.018 sys,Flavortagging + 4 more errors Show all
0.4 - 0.5	1.74 ±0.0018 stat +0.039 -0.024 sys,Calorimeterjetenergy +0.0092 -0.0084 sys,Flavortagging + 4 more errors Show all
0.5 - 0.6	1.31 ±0.0013 stat +0.036 -0.0067 sys,Calorimeterjetenergy +0.013 -0.013 sys,Flavortagging + 4 more errors Show all
0.6 - 0.7	1.06 ±0.0013 stat +0.018 -0.026 sys,Calorimeterjetenergy +0.014 -0.017 sys,Flavortagging + 4 more errors Show all

Visualize



Sum errors ☑ Log Scale (X) ☐ Log Scale (Y) ☐

YAML with resource files

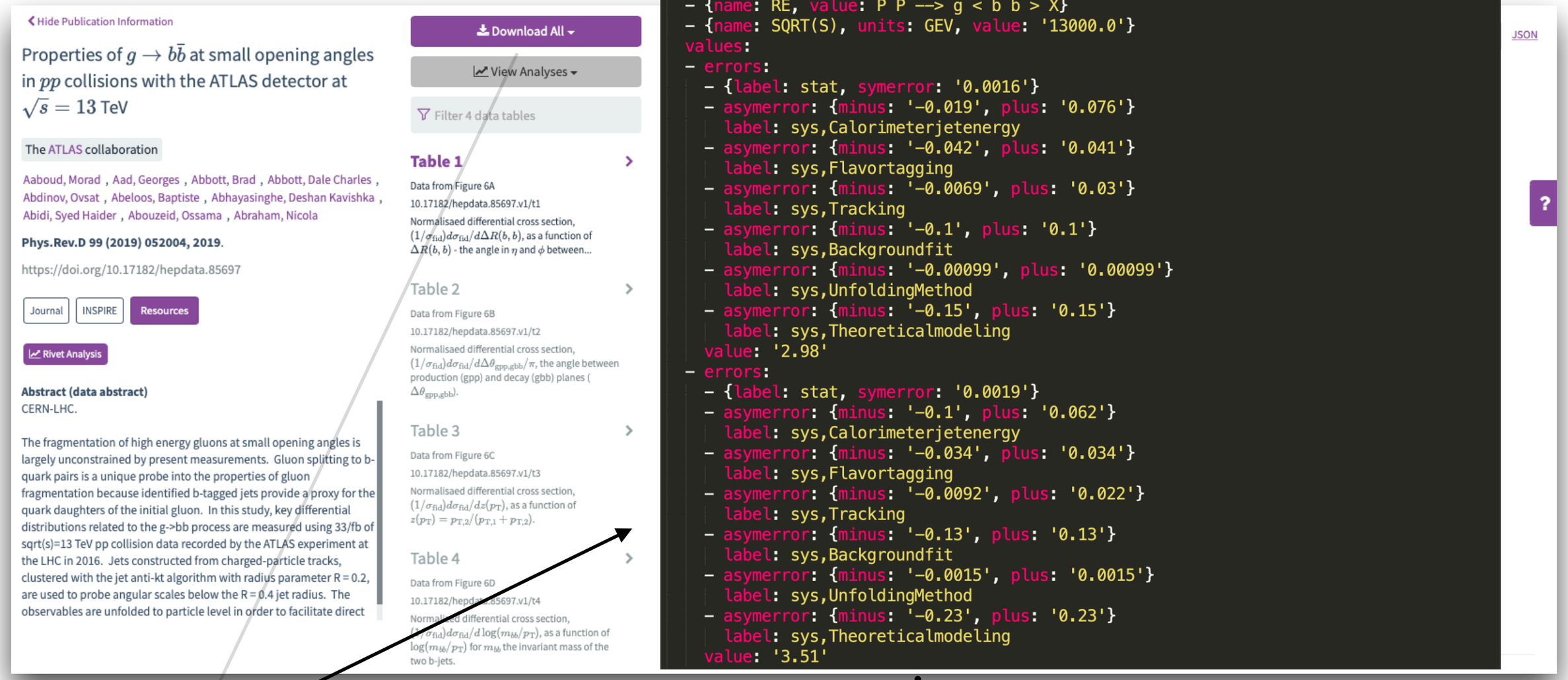
YAML

YODA

ROOT

CSV

Publishing Binned Results




```yaml
dependent_variables:
- header: {name: '$(1/\sigma_\text{fid})d\sigma_\text{fid}/d\Delta R(b,b)$'}
  qualifiers:
  - {name: RE, value: P P --> g < b b > X}
  - {name: SQRT(S), units: GEV, value: '13000.0'}
  values:
  - errors:
    - {label: stat, symerror: '0.0016'}
    - asymerror: {minus: '-0.019', plus: '0.076'}
      label: sys,Calorimeterjetenergy
    - asymerror: {minus: '-0.042', plus: '0.041'}
      label: sys,Flavortagging
    - asymerror: {minus: '-0.0069', plus: '0.03'}
      label: sys,Tracking
    - asymerror: {minus: '-0.1', plus: '0.1'}
      label: sys,Backgroundfit
    - asymerror: {minus: '-0.00099', plus: '0.00099'}
      label: sys,UnfoldingMethod
    - asymerror: {minus: '-0.15', plus: '0.15'}
      label: sys,Theoreticalmodeling
    value: '2.98'
  - errors:
    - {label: stat, symerror: '0.0019'}
    - asymerror: {minus: '-0.1', plus: '0.062'}
      label: sys,Calorimeterjetenergy
    - asymerror: {minus: '-0.034', plus: '0.034'}
      label: sys,Flavortagging
    - asymerror: {minus: '-0.0092', plus: '0.022'}
      label: sys,Tracking
    - asymerror: {minus: '-0.13', plus: '0.13'}
      label: sys,Backgroundfit
    - asymerror: {minus: '-0.0015', plus: '0.0015'}
      label: sys,UnfoldingMethod
    - asymerror: {minus: '-0.23', plus: '0.23'}
      label: sys,Theoreticalmodeling
    value: '3.51'
```

```yaml
independent_variables:
- header: {name: '$\Delta R(b,b)$'}
  values:
  - {high: 0.25, low: 0.2}
  - {high: 0.3, low: 0.25}
  - {high: 0.4, low: 0.3}
  - {high: 0.5, low: 0.4}
  - {high: 0.6, low: 0.5}
  - {high: 0.7, low: 0.6}
```

YAML with resource files

YAML

YODA

ROOT

CSV

YAML files with metadata, bin contents, and uncertainties

How to represent unbinned data?

If the data can be fit with a function, you could publish the function (e.g. if it is a NN, you could publish the architecture and weights).

Another natural representation that doesn't require a function fit is to publish data sampled from the unfolded result.

My proposal is based on this idea.

Proposal

As in HEPData, I propose there is a “submission” YAML file with the same measurement metadata.

Each submeasurement* also has some metadata & points to a data file. In HEPData, the data file is itself a YAML file.

The files will have data with the “shape” [(M+1) x N(k+1)]

…where N is the number of sampled events

and M is the number of systematic uncertainties

and k is the number of dimensions per event

*this could be a single observable, or many observables

Proposal

The files will have data with the “shape” [(M+1) x N(k+1)]

Each event has k floats* and 1 event weight

There are N events

This is repeated for each of the M systematic uncertainties

For representations that don’t have weights, the weights will be set to 1. For representations that only use weights, there will be M copies of the original array.

I have not thought deeply about file formats (npy, root, hdf5) and would be happy to hear opinions.

*For variable-length measurements, perhaps should use variable-length arrays like awkward for storage

Proposal - uncertainties

The submission YAML should give metadata about which uncertainties are included.

For statistical uncertainties, there should be Q replicas and the uncertainty in a given bin is computed by taking the standard deviation over replicas.

For systematic uncertainties, the difference between the nominal and varied bin content is the uncertainty.

There should be warnings in metadata and/or inflated uncertainties in regions of phase space that should not be studied with the data.

Proposal - where to store?

Zenodo is a very natural location. Maybe the submission YAML can also be hosted on HEPData and linked to Zenodo for each searching?

Proposal - example

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:
```
x = np.random.normal(0,1,10000)
```

In [3]:
```
w = np.abs(x)**0.2
w_syst_up = np.abs(x)**0.3
w_syst_dn = np.abs(x)**0.1
```

In [4]:
```
plt.hist(x,bins=np.linspace(-3,3,10),alpha=0.5)
plt.hist(x,bins=np.linspace(-3,3,10),weights=w,histtype="step",color="black")
n_syst_up,b=np.histogram(x,bins=np.linspace(-3,3,10),weights=w_syst_up)
n_syst_dn,_=np.histogram(x,bins=np.linspace(-3,3,10),weights=w_syst_dn)
for i in range(len(b)-1):
    plt.fill_between([b[i],b[i+1]],n_syst_dn[i],n_syst_up[i],color="black",alpha=0.3)
plt.xlabel("X")
```

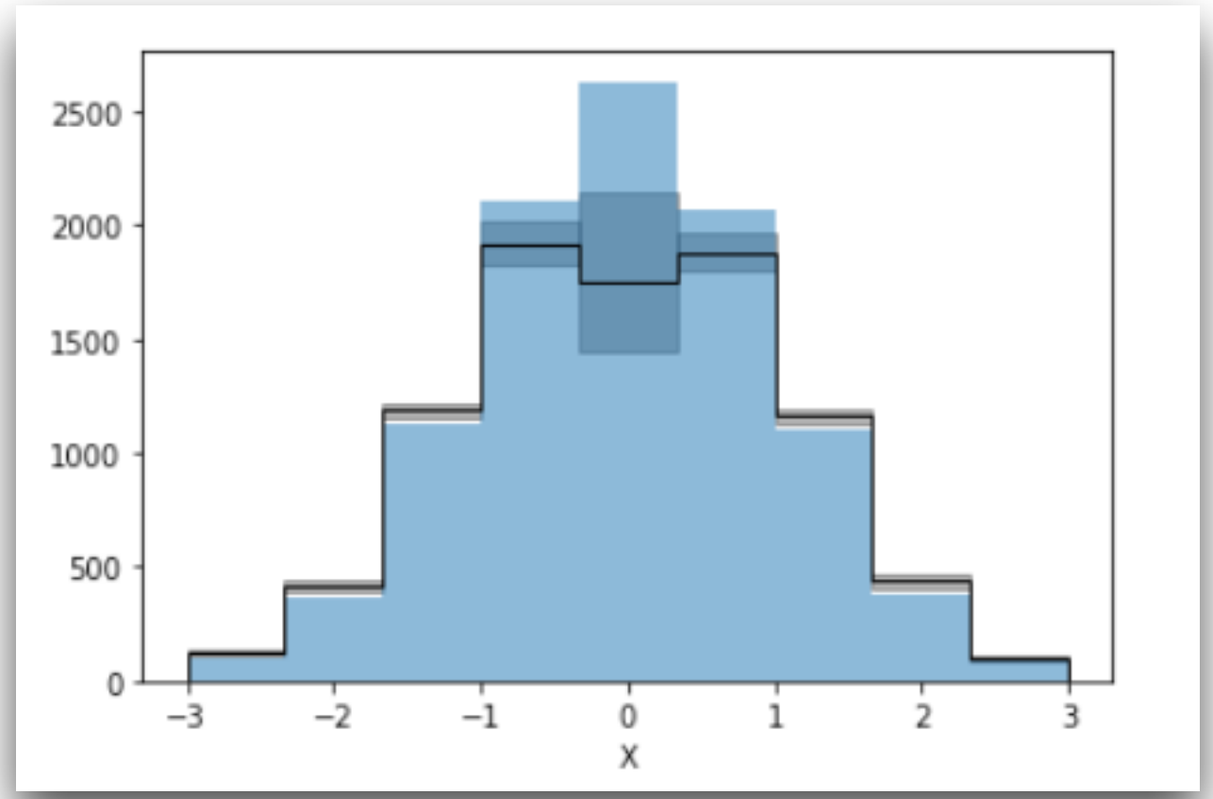
In [5]:
```
d = {"nominal":x,
     "nominalw":w,
     "syst_up":x,
     "syst_upw":w_syst_up,
     "syst_dn":x,
     "syst_dnw":w_syst_dn}
```

In [6]:
```
df = pd.DataFrame(data=d)
```

In [7]:
```
df
```

Out[7]:

	nominal	nominalw	syst_up	syst_upw	syst_dn	syst_dnw
0	0.731914	0.939490	0.731914	0.910622	0.731914	0.969273
1	0.146232	0.680783	0.146232	0.561711	0.146232	0.825096
2	-0.629654	0.911634	-0.629654	0.870423	-0.629654	0.954795
3	0.581001	0.897089	0.581001	0.849675	0.581001	0.947148
4	-0.321038	0.796730	-0.321038	0.711159	-0.321038	0.892597
...	...	...	...	...	...	...



Discussion!