Systematic Uncertainties from Synthetic Datasets

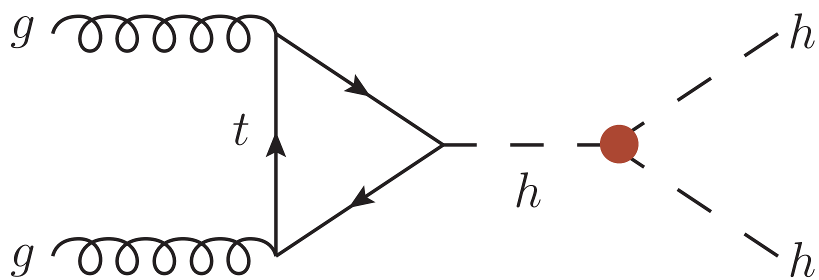
A case study with $HH \rightarrow 4b$

John Alison

Carnegie Mellon University

Motivation

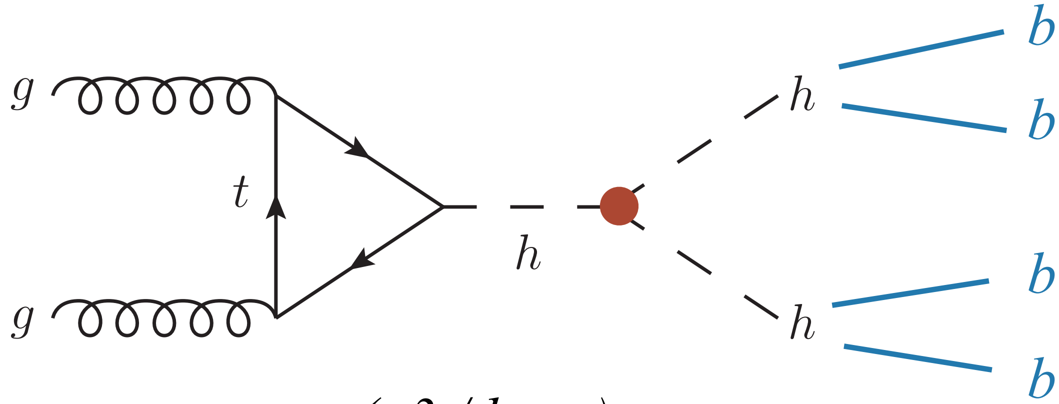
Measuring the Higgs self-coupling λ major goal of the HL-LHC
Di-Higgs production most direct and most sensitivity way measure λ



(~3 / hour)

Motivation

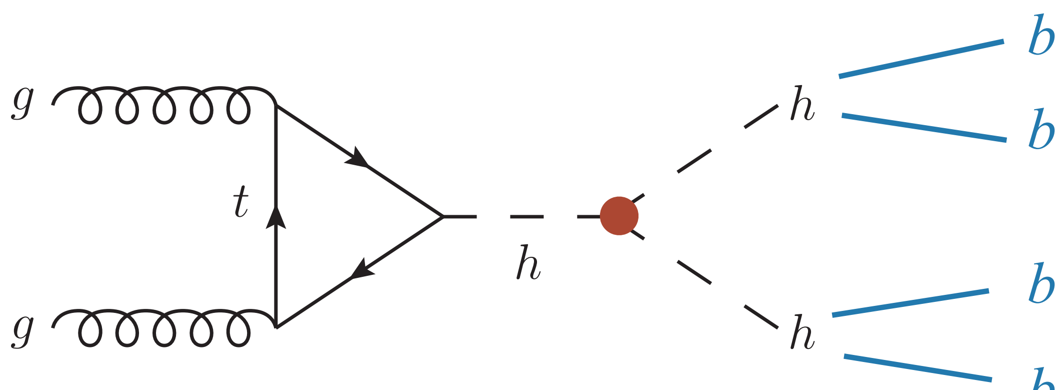
Measuring the Higgs self-coupling λ major goal of the HL-LHC
Di-Higgs production most direct and most sensitivity way measure λ



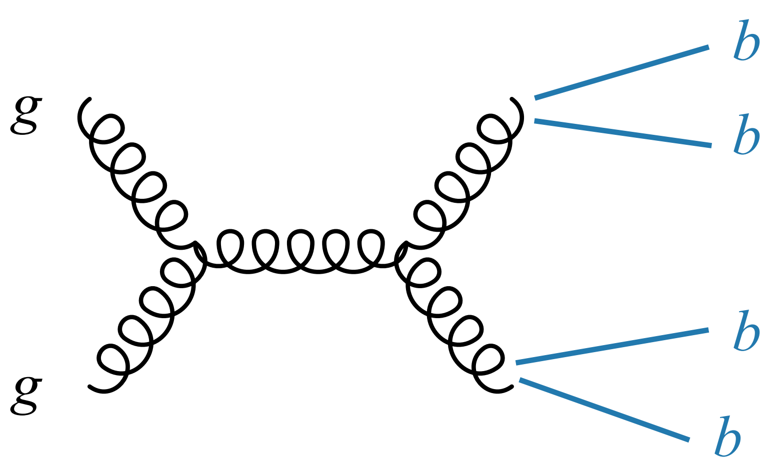
(~3 / hour)

Motivation

Measuring the Higgs self-coupling λ major goal of the HL-LHC
Di-Higgs production most direct and most sensitivity way measure λ



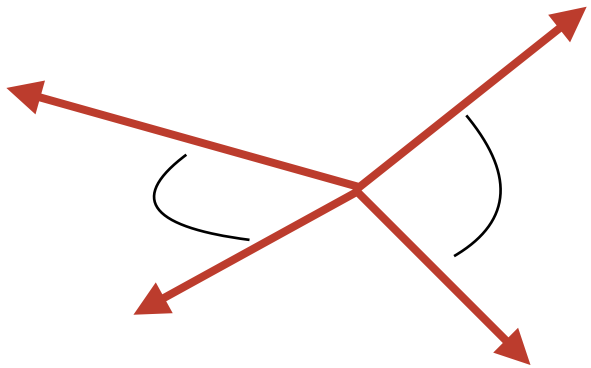
(~3 / hour)



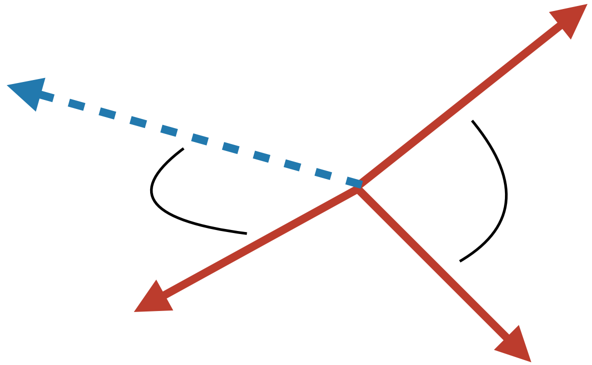
(~30,000 / hour)

Data Driven Background: *ABCD*

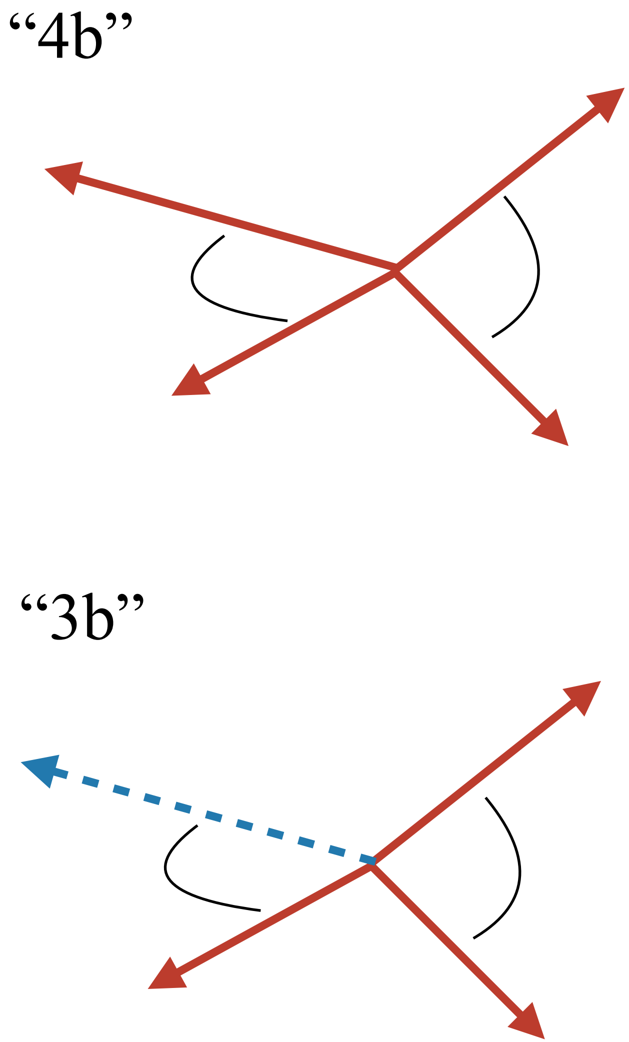
"4b"

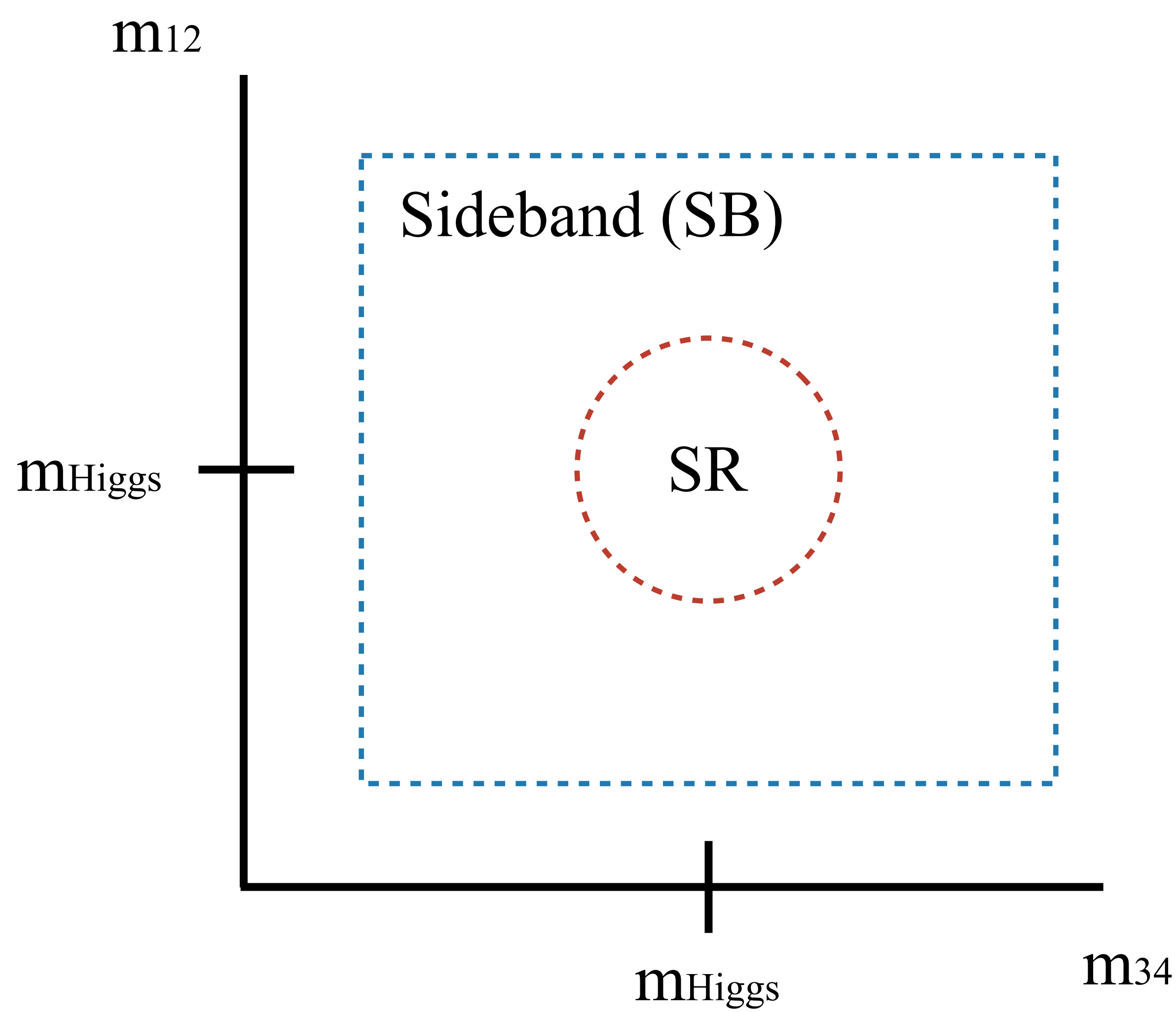


"3b"

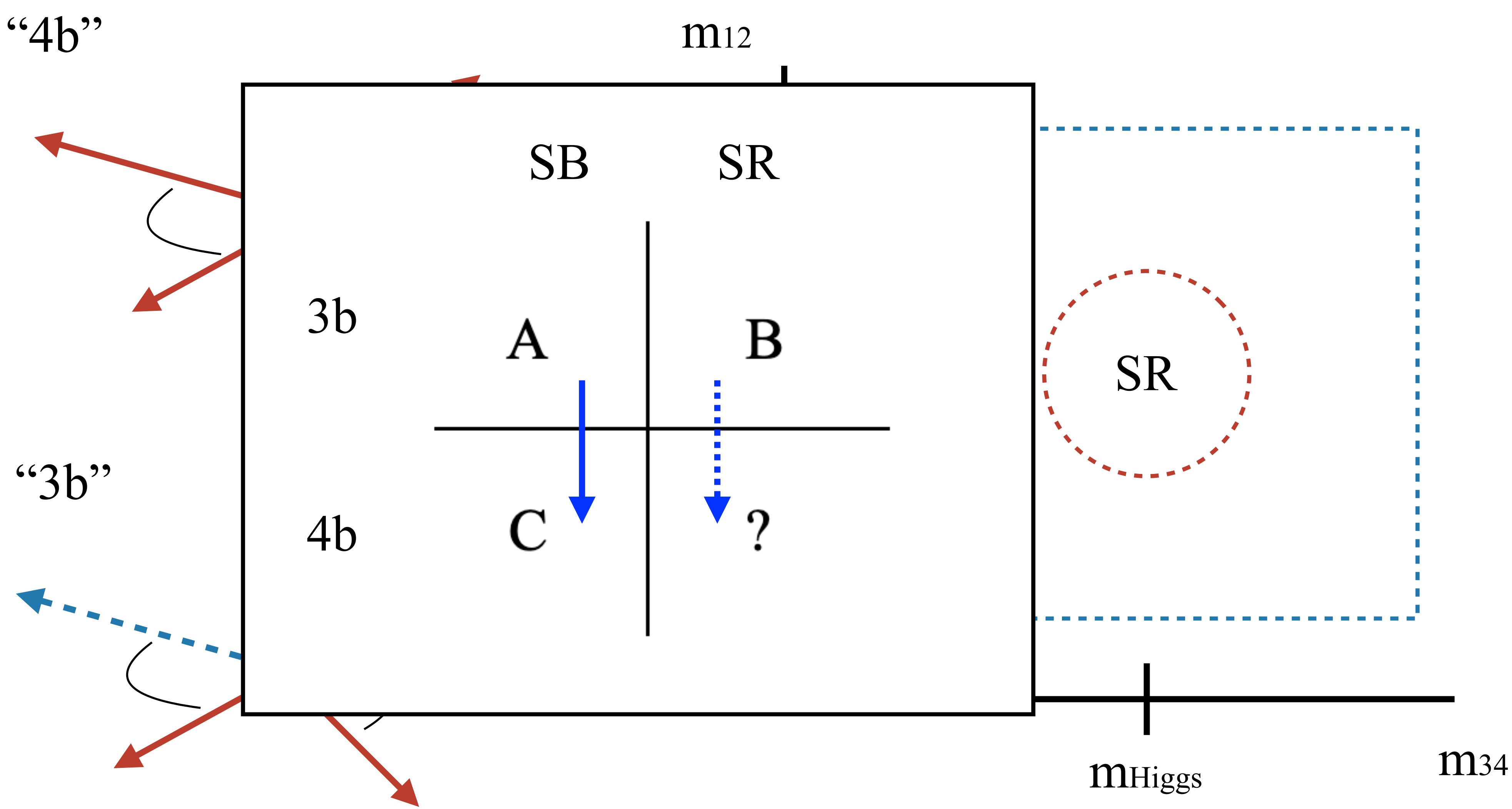


Data Driven Background: *ABCD*

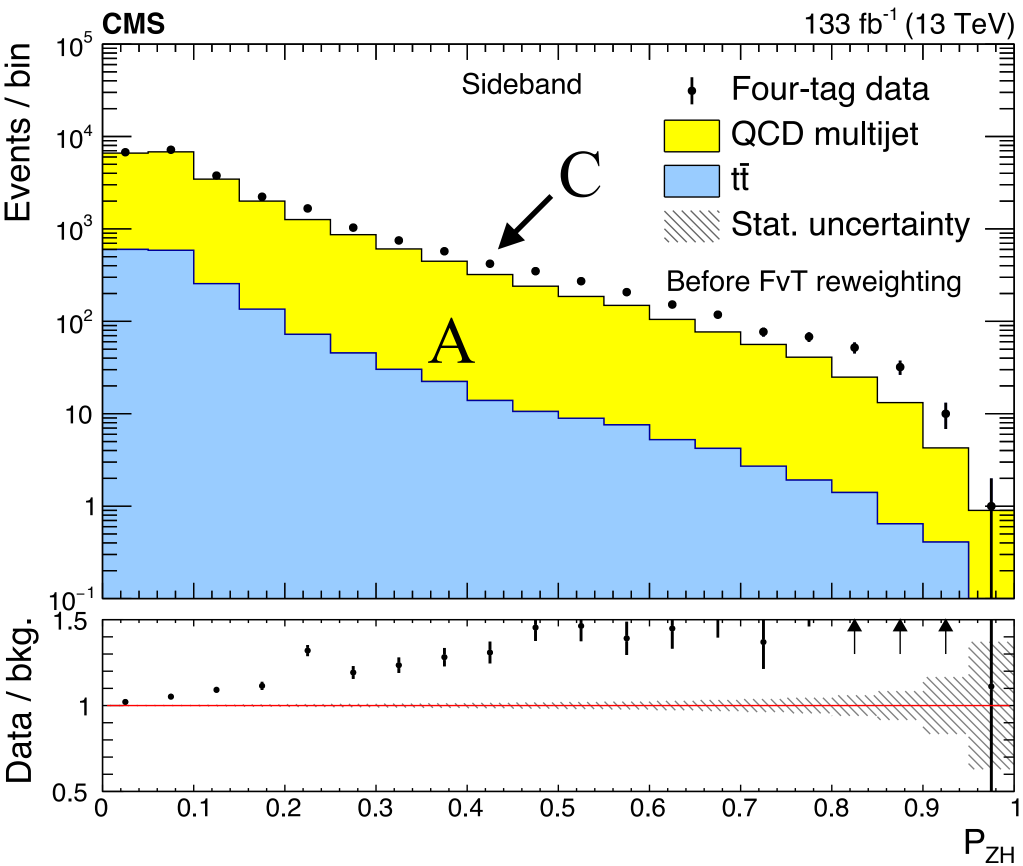


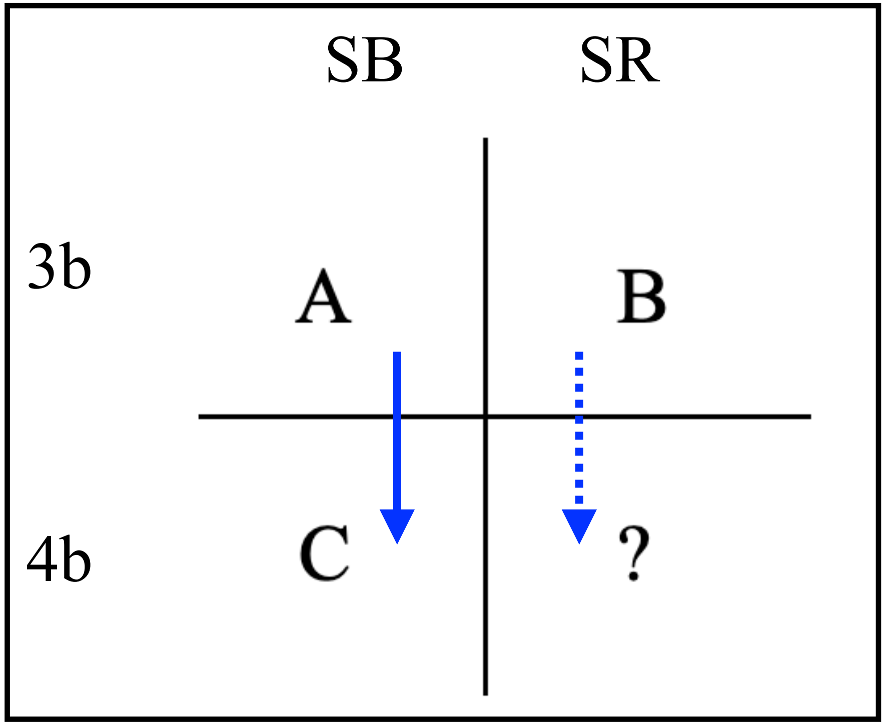


Data Driven Background: *ABCD*

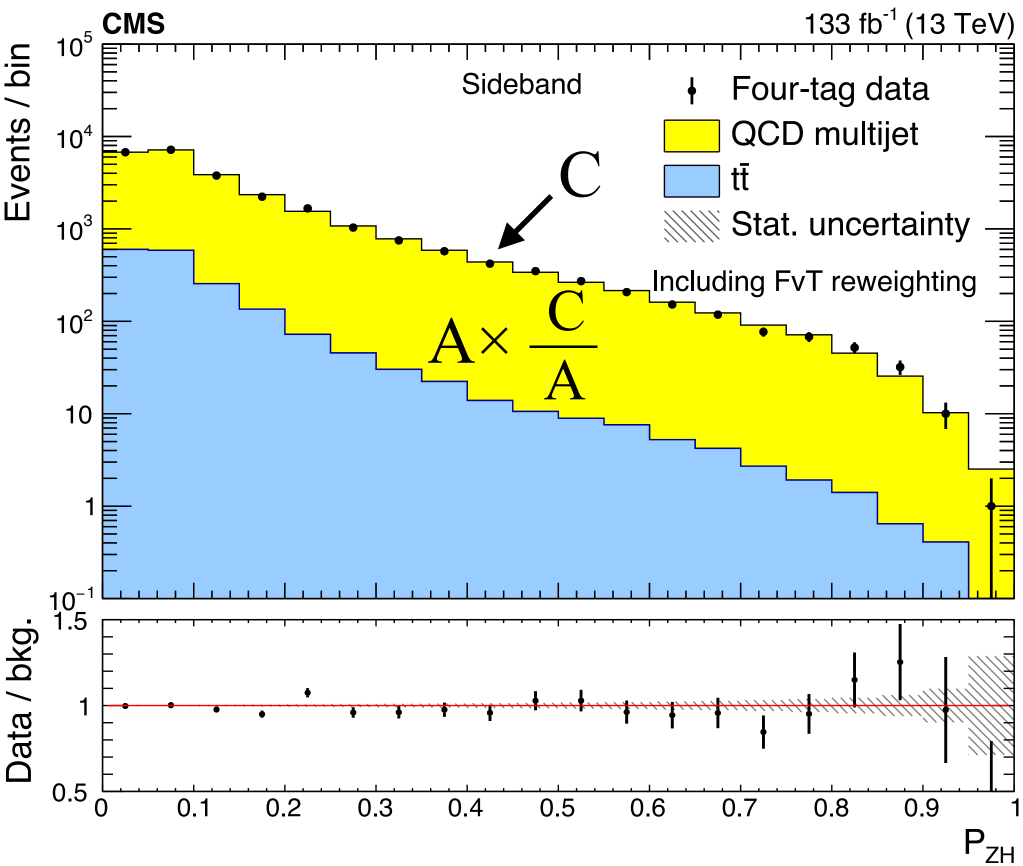


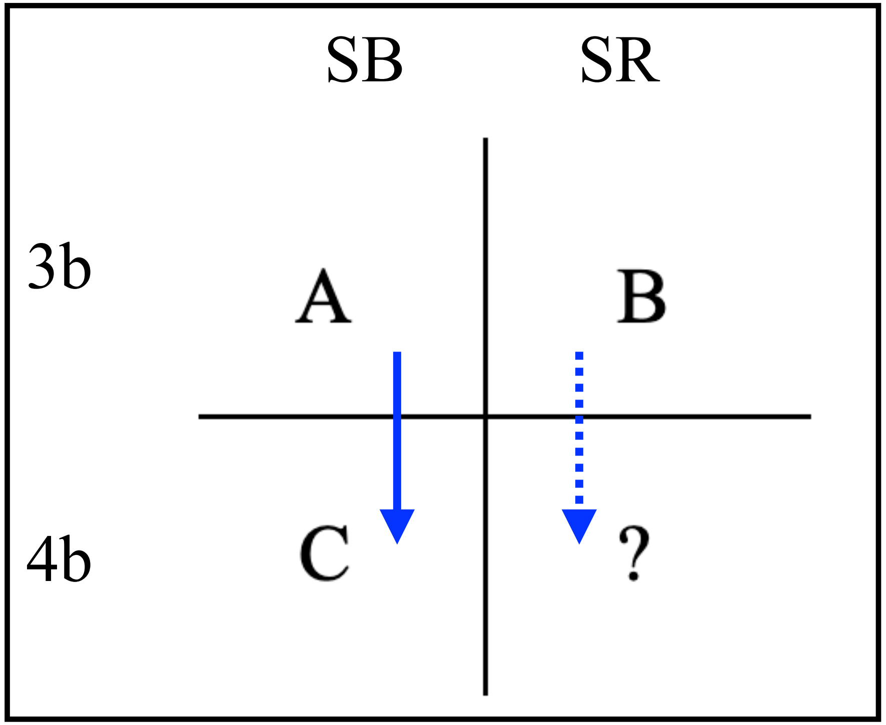
ABCD Method Works



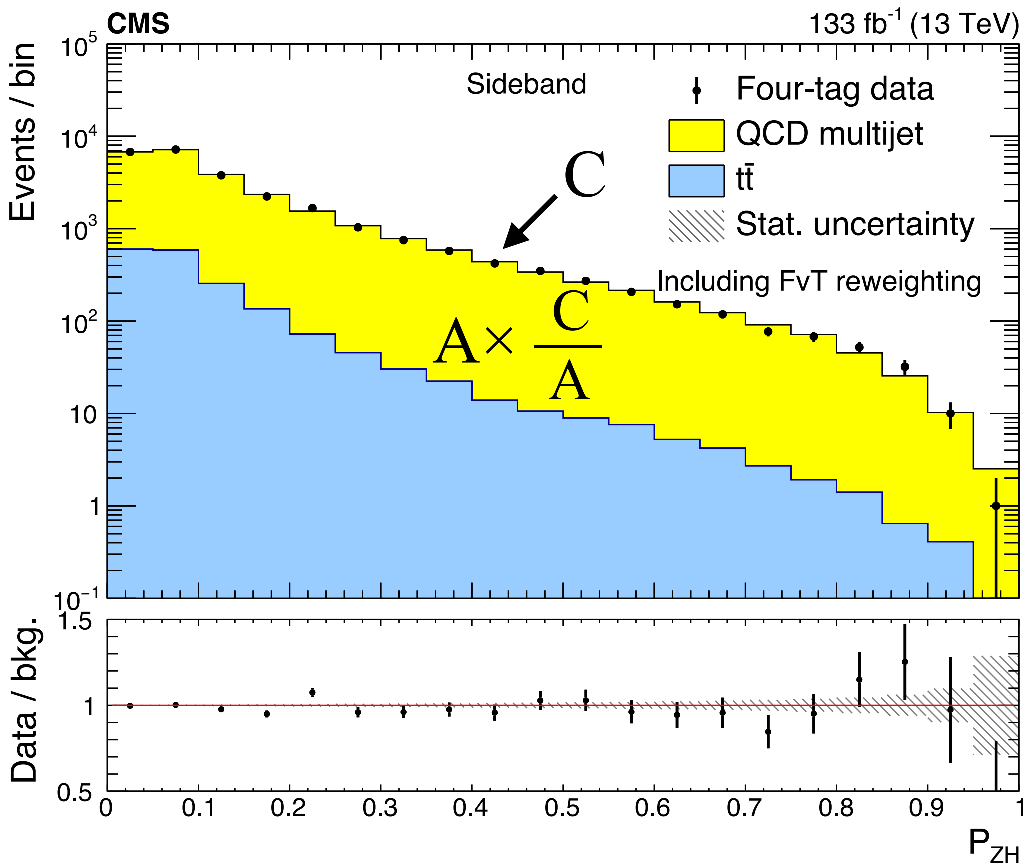


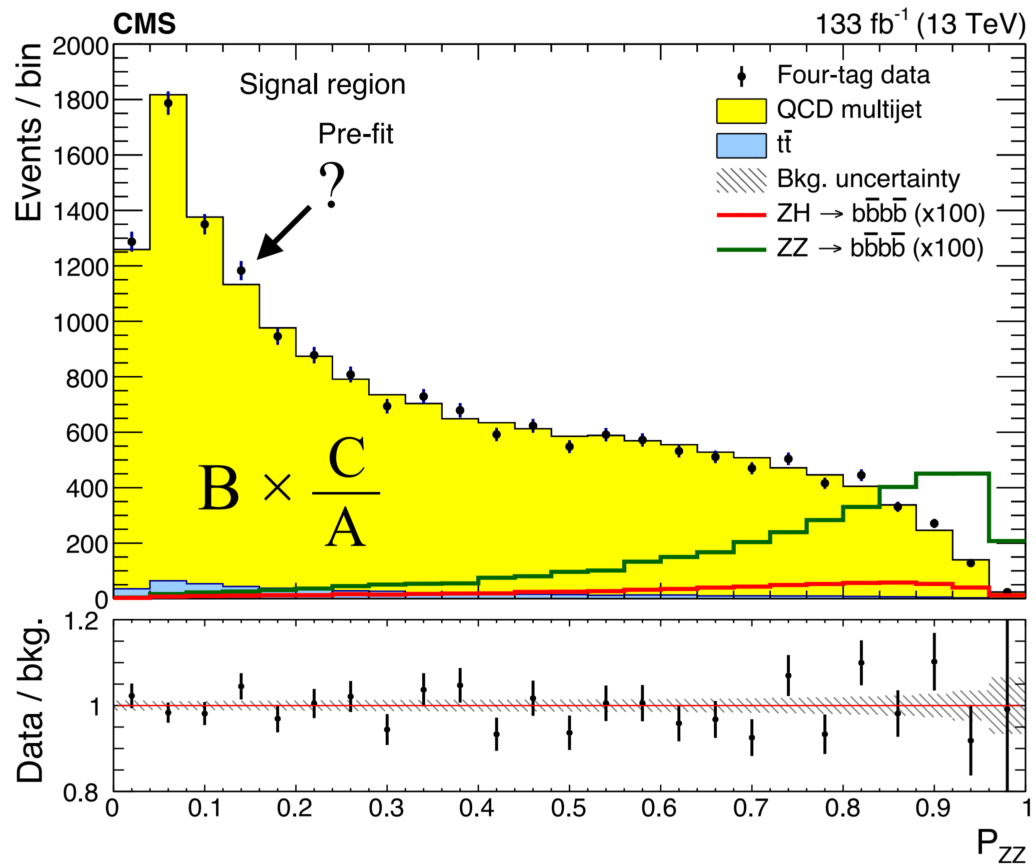
ABCD Method Works



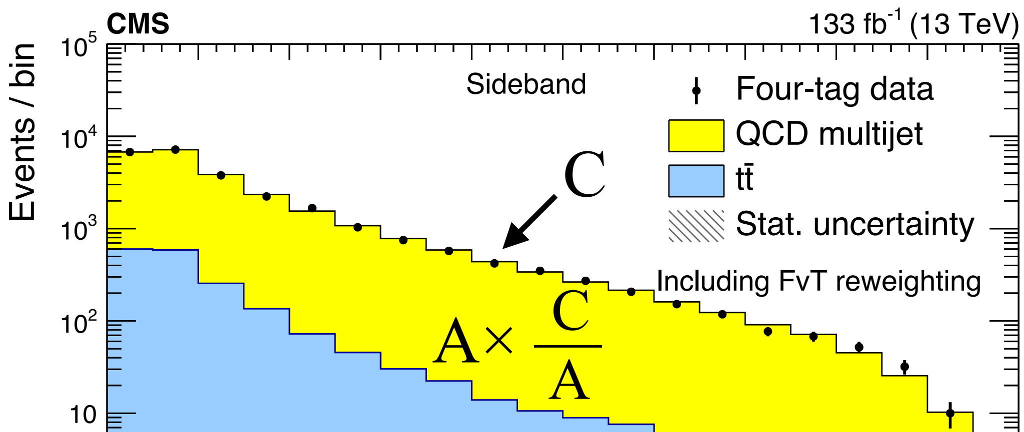


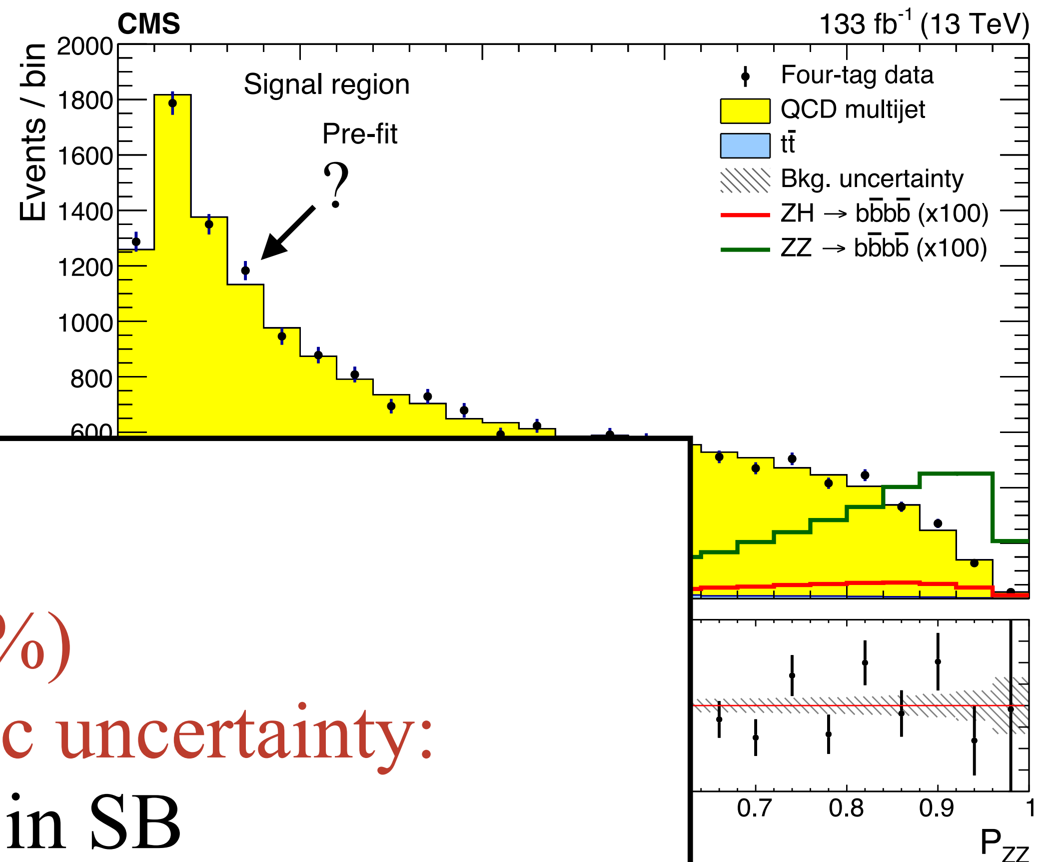
ABCD Method Works





ABCD Method Works





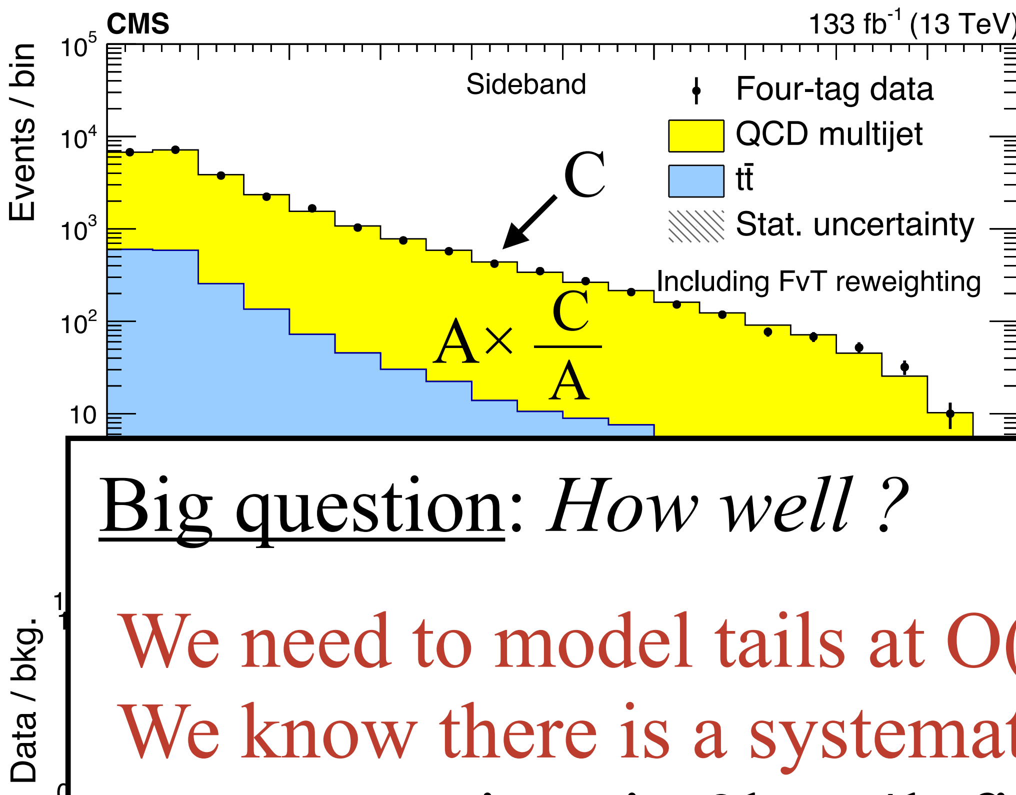
Big question: *How well ?*

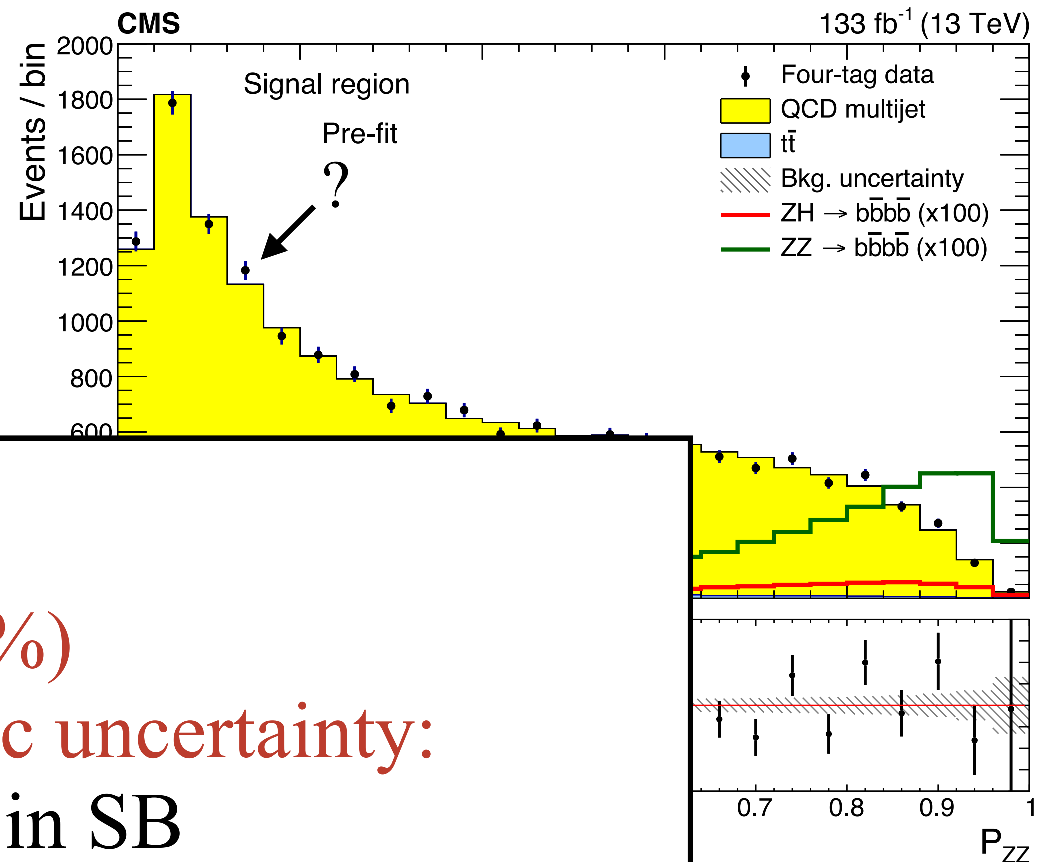
We need to model tails at O(%)

We know there is a systematic uncertainty:

- Uncertainty in 3b→4b fit in SB
- Extrapolation (domain shift) from CR→SR
- …

ABCD Method Works





<u>Big question</u>: *How well ?*

We need to model tails at O(%)

We know there is a systematic uncertainty:

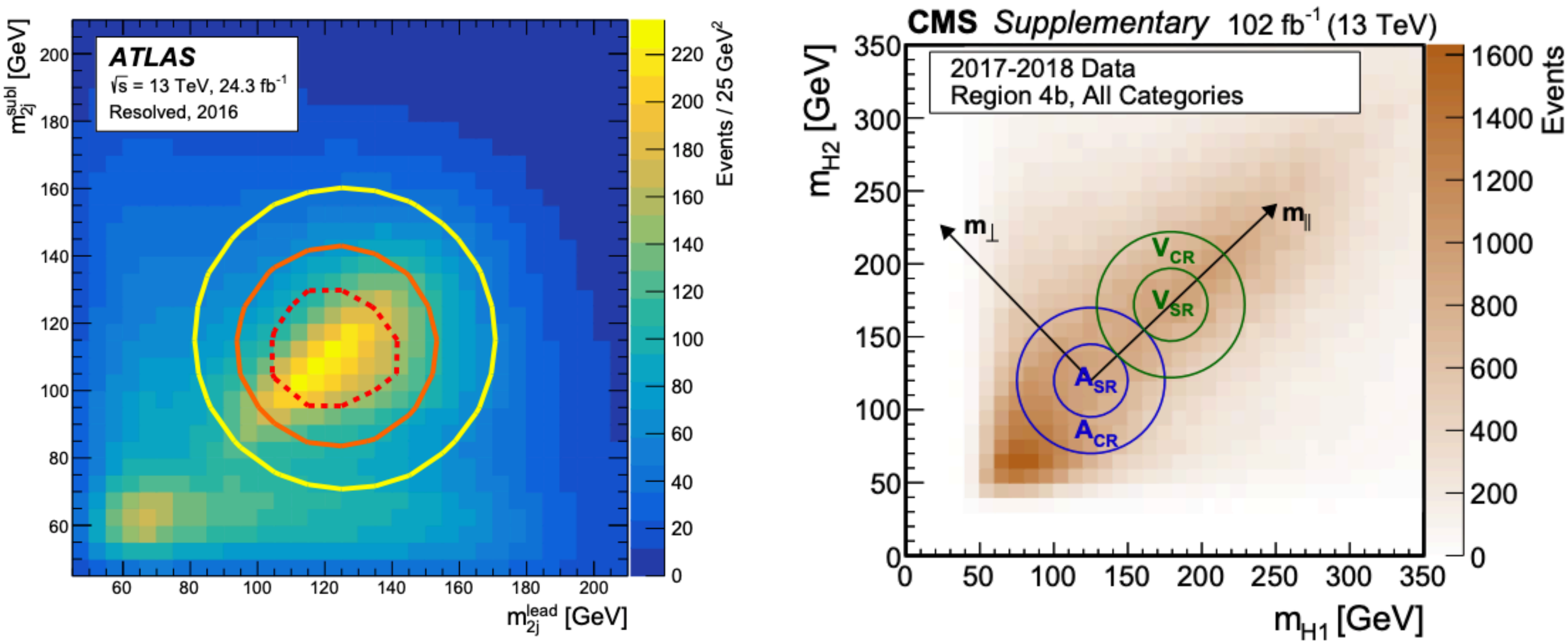
- Uncertainty in 3b→4b fit in SB
- Extrapolation (domain shift) from CR→SR
- ...

Rest of this talk is an idea for validating this uncertainty

Standard Solution: Validation Region

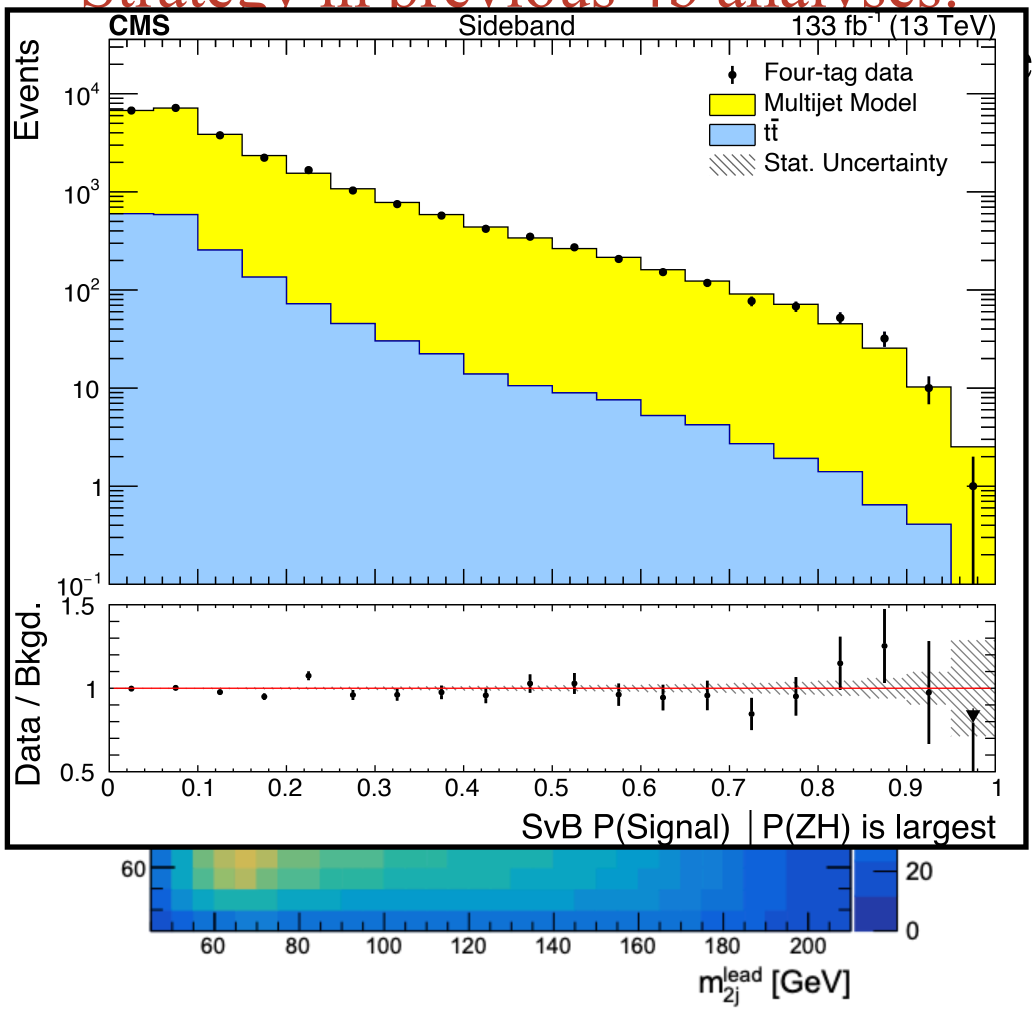
Strategy in previous 4b analyses:

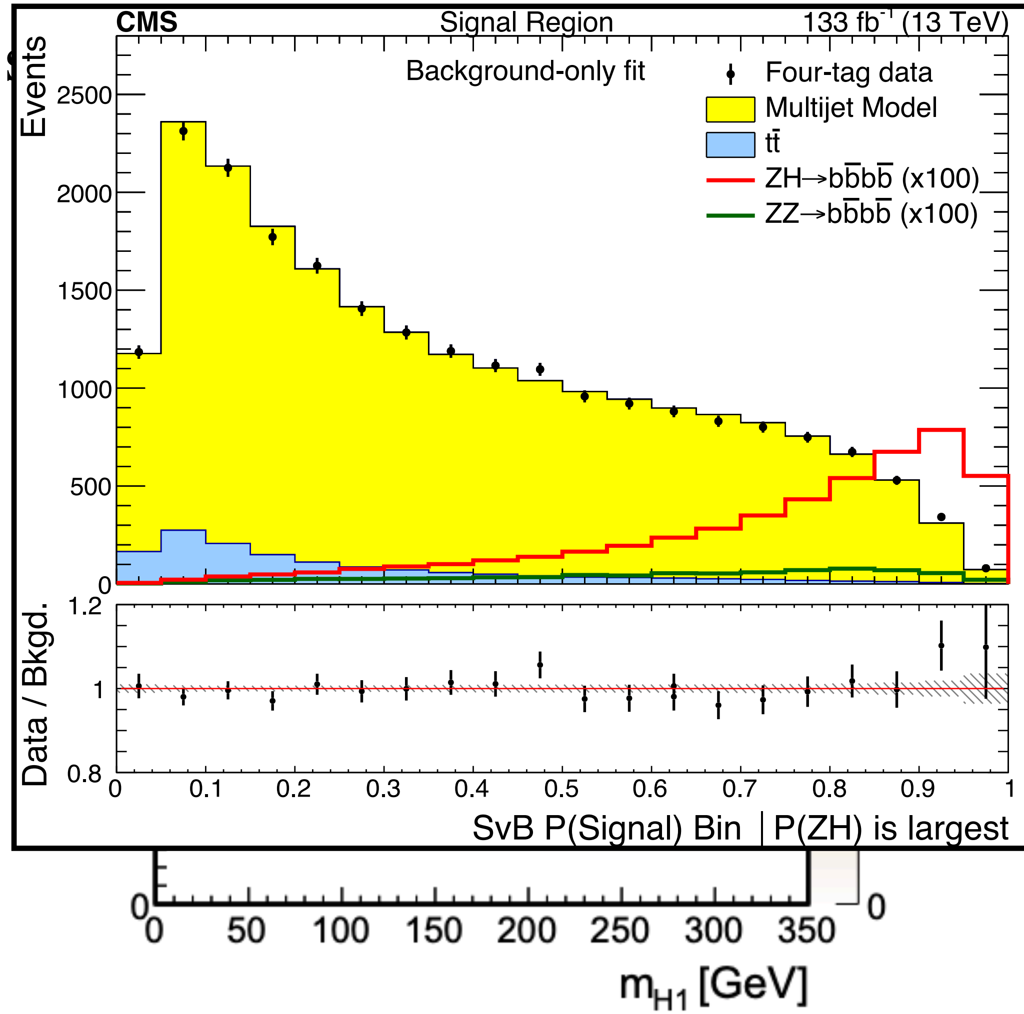
Validated prediction in alternative signal-free region



Standard Solution: Validation Region

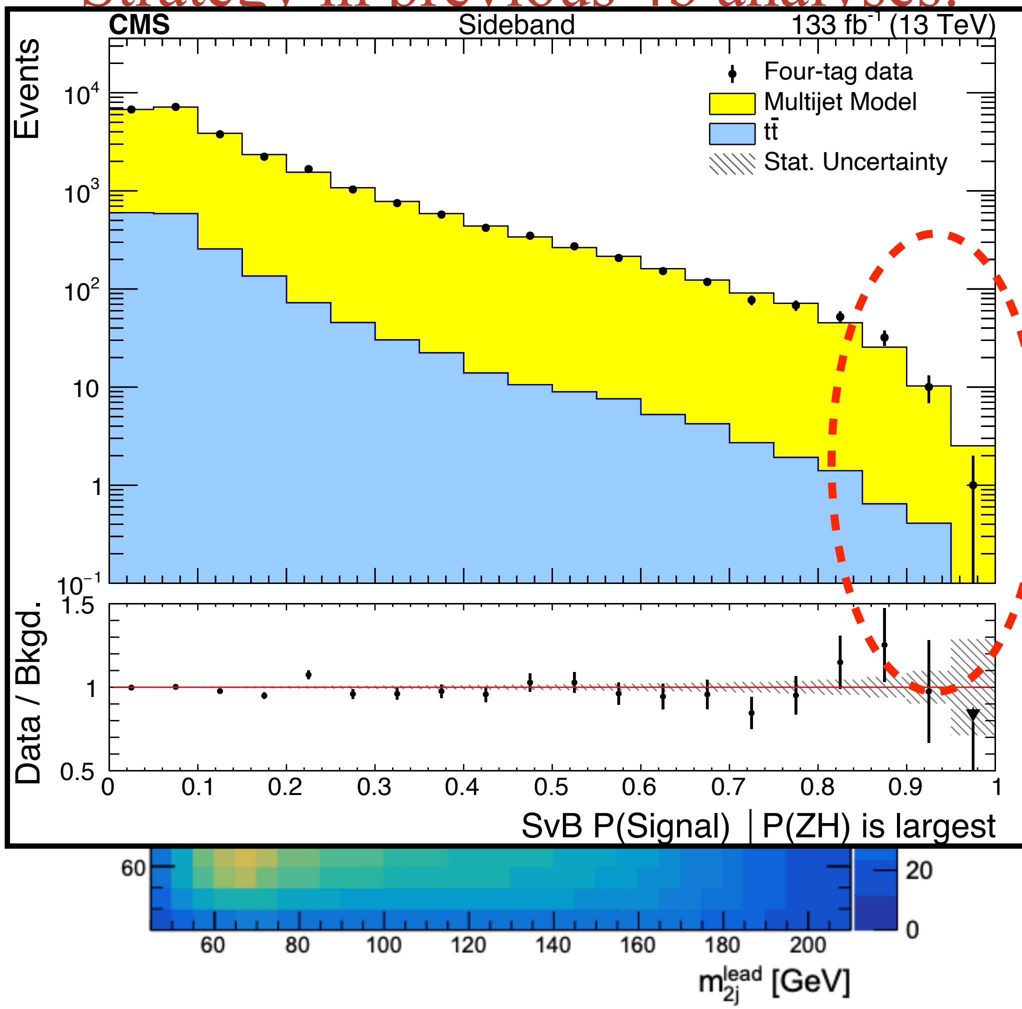
Strategy in previous 4b analyses:

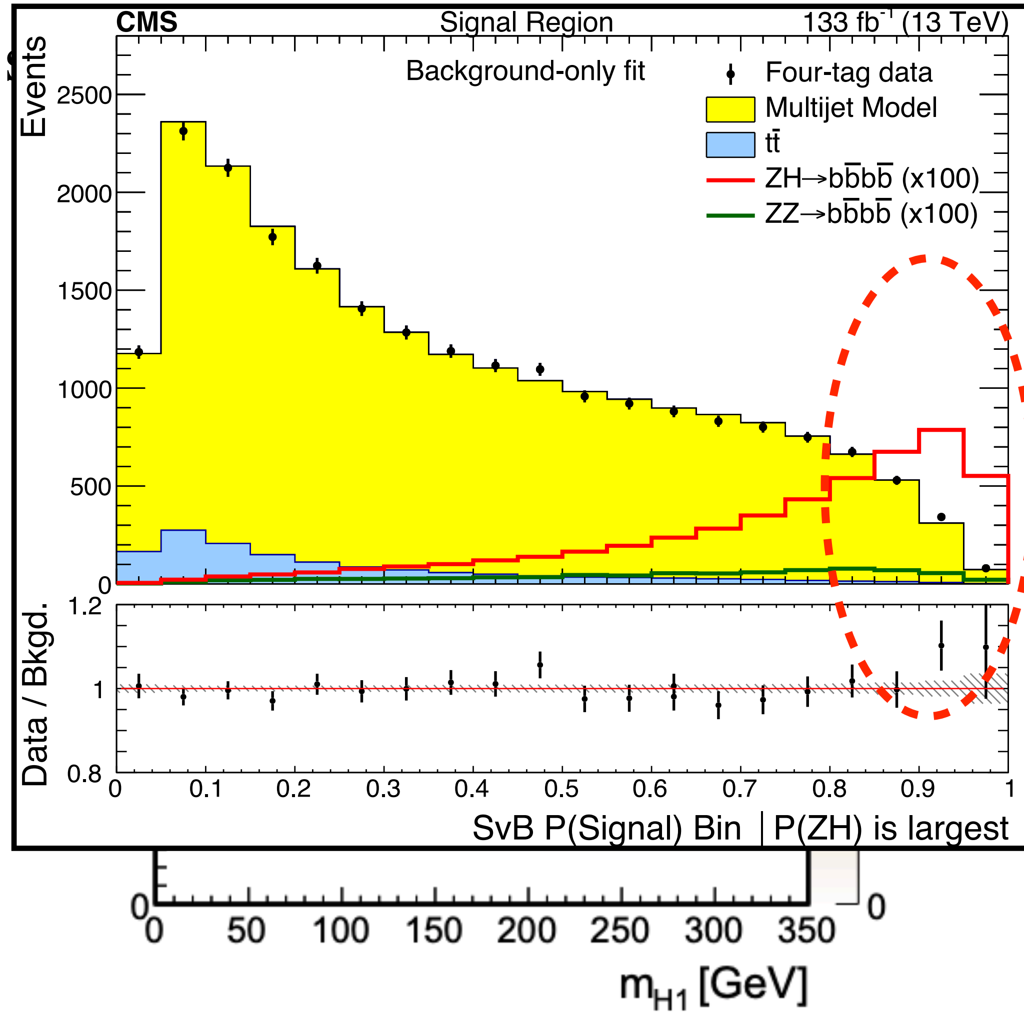




Standard Solution: Validation Region

Strategy in previous 4b analyses:





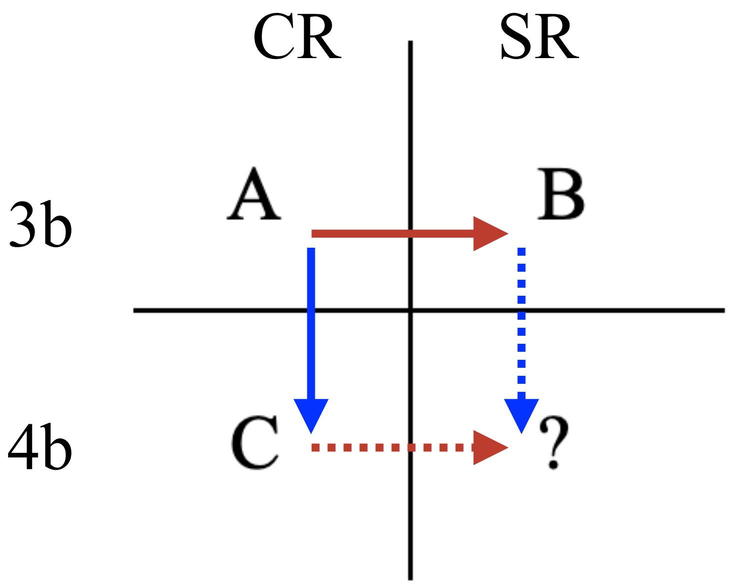
Aside: Solution with Optimal Transport

BACKGROUND MODELING FOR DOUBLE HIGGS BOSON PRODUCTION: DENSITY RATIOS AND OPTIMAL TRANSPORT

BY TUDOR MANOLE[a], PATRICK BRYANT[d], JOHN ALISON[e], MIKAEL KUUSELA[b], AND LARRY WASSERMAN[c]

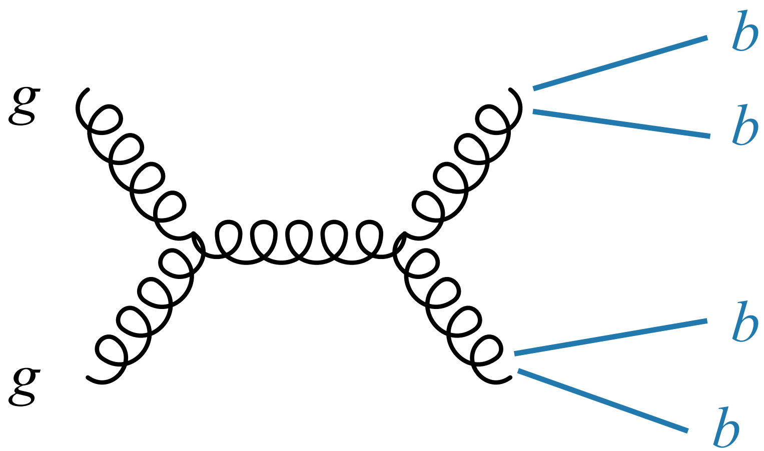
Department of Statistics and Data Science and NSF AI Planning Institute for Data-Driven Discovery in Physics, Carnegie Mellon University
[a]*tmanole@andrew.cmu.edu*; [b]*mkuusela@andrew.cmu.edu*; [c]*larry@stat.cmu.edu*

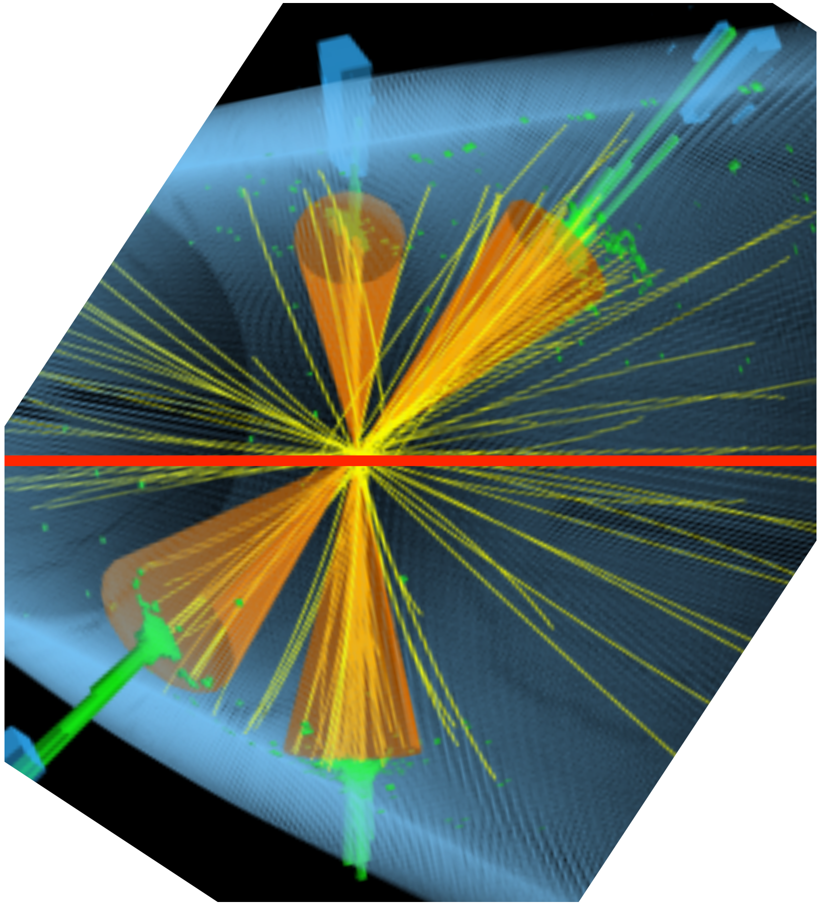
Department of Physics and NSF AI Planning Institute for Data-Driven Discovery in Physics, Carnegie Mellon University
[d]*pbryant2@andrew.cmu.edu*; [e]*johnalison@cmu.edu*



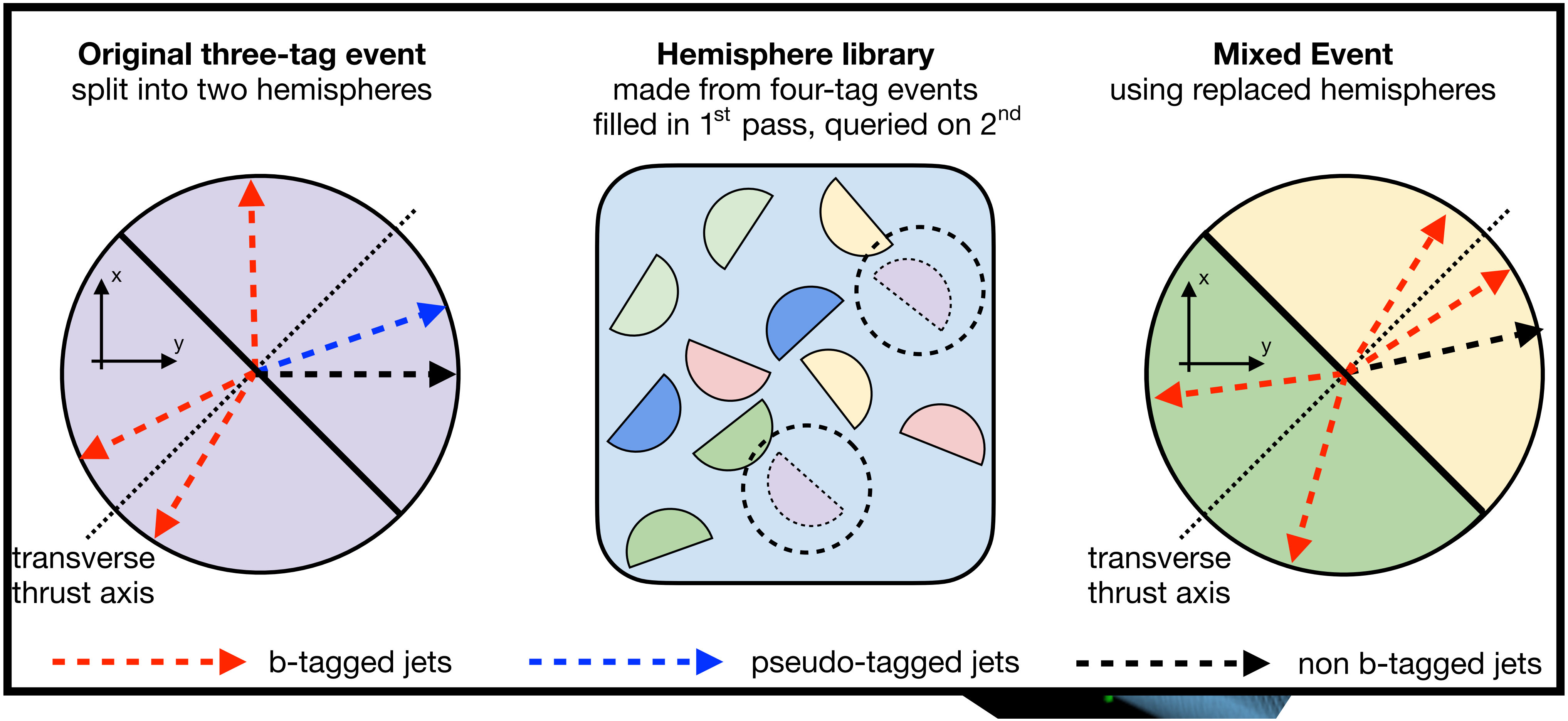
arXiv:2208.02807

Synthetic Datasets: Event Mixing

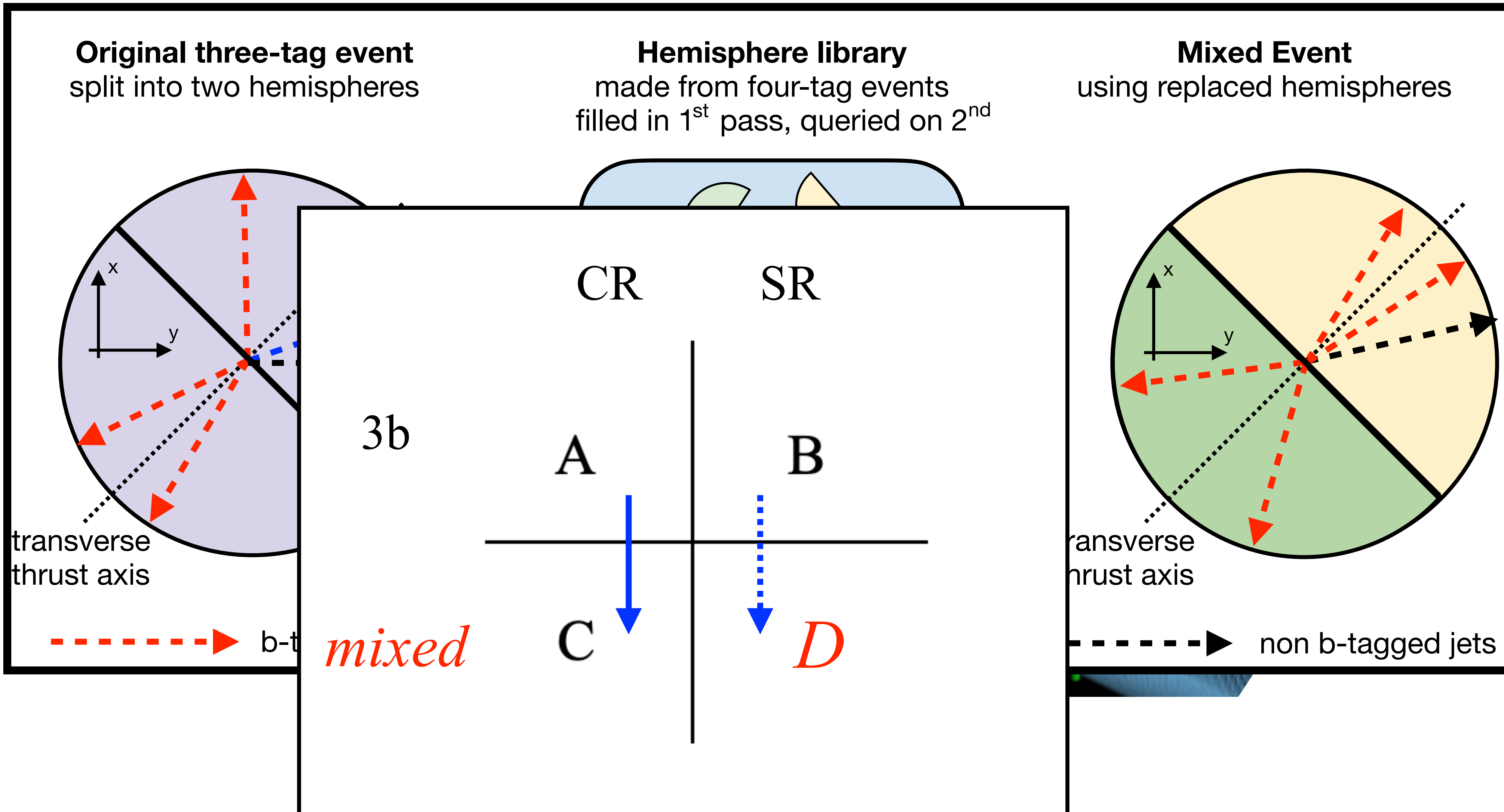


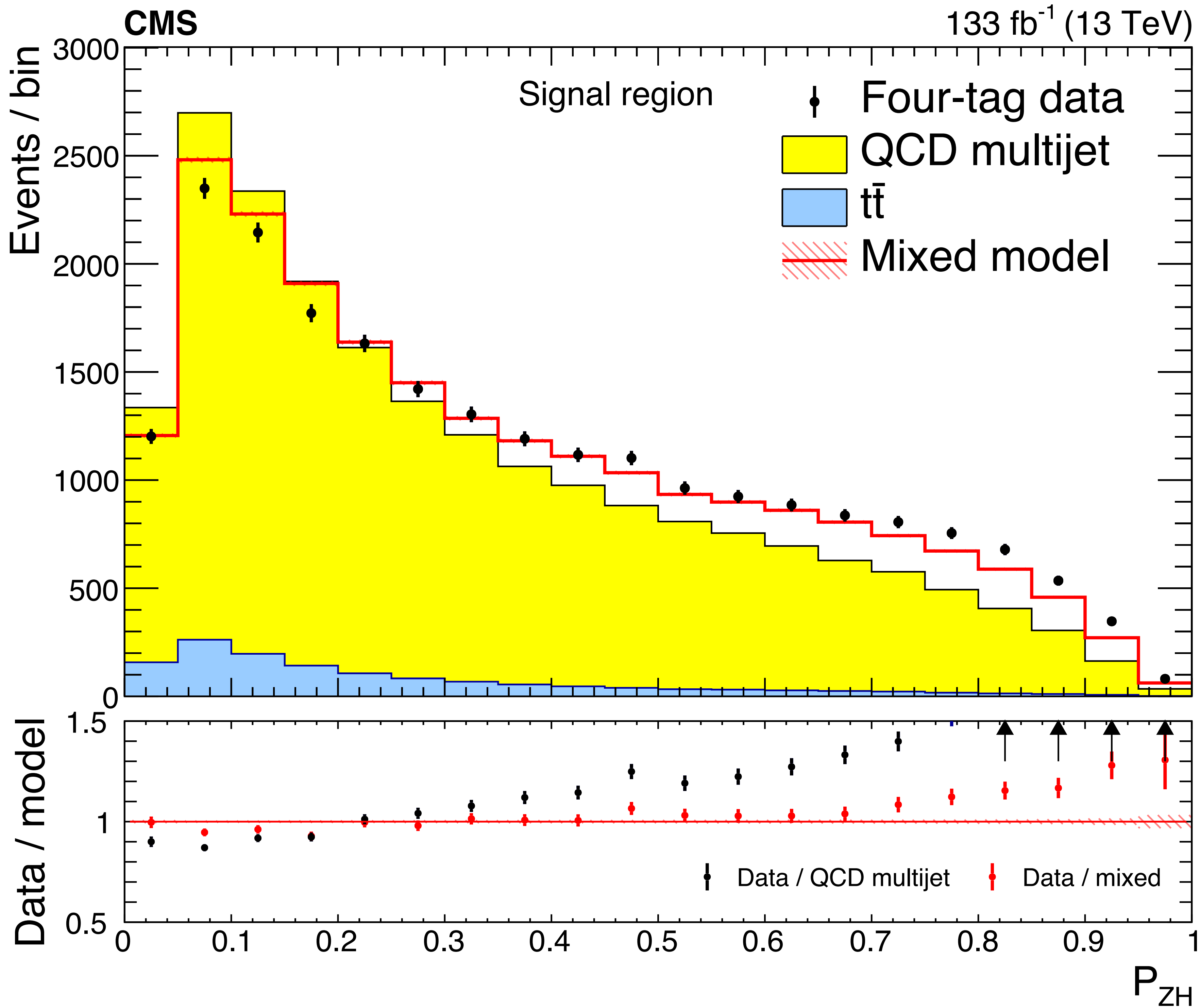


Synthetic Datasets: Event Mixing

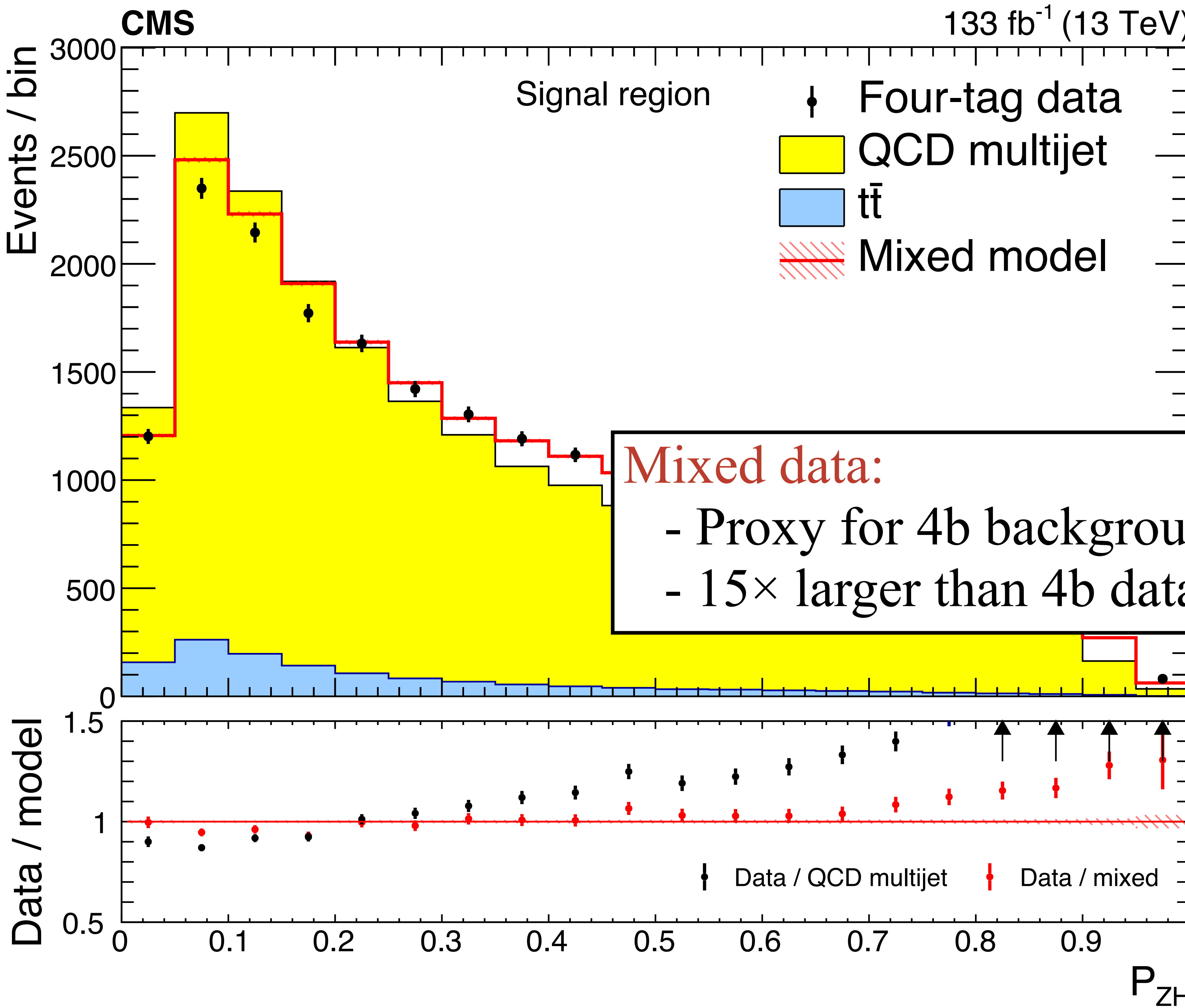


Synthetic Datasets: Event Mixing





CMS
133 fb^{-1} (13 TeV)
Signal region
Four-tag data
QCD multijet
$t\bar{t}$
Mixed model
Events / bin
Data / model
Data / QCD multijet
Data / mixed
P_{ZH}



Mixed data:

- Proxy for 4b background
- 15× larger than 4b dataset

Systematics with Mixed Data

Consider three sources of potential systematic uncertainty

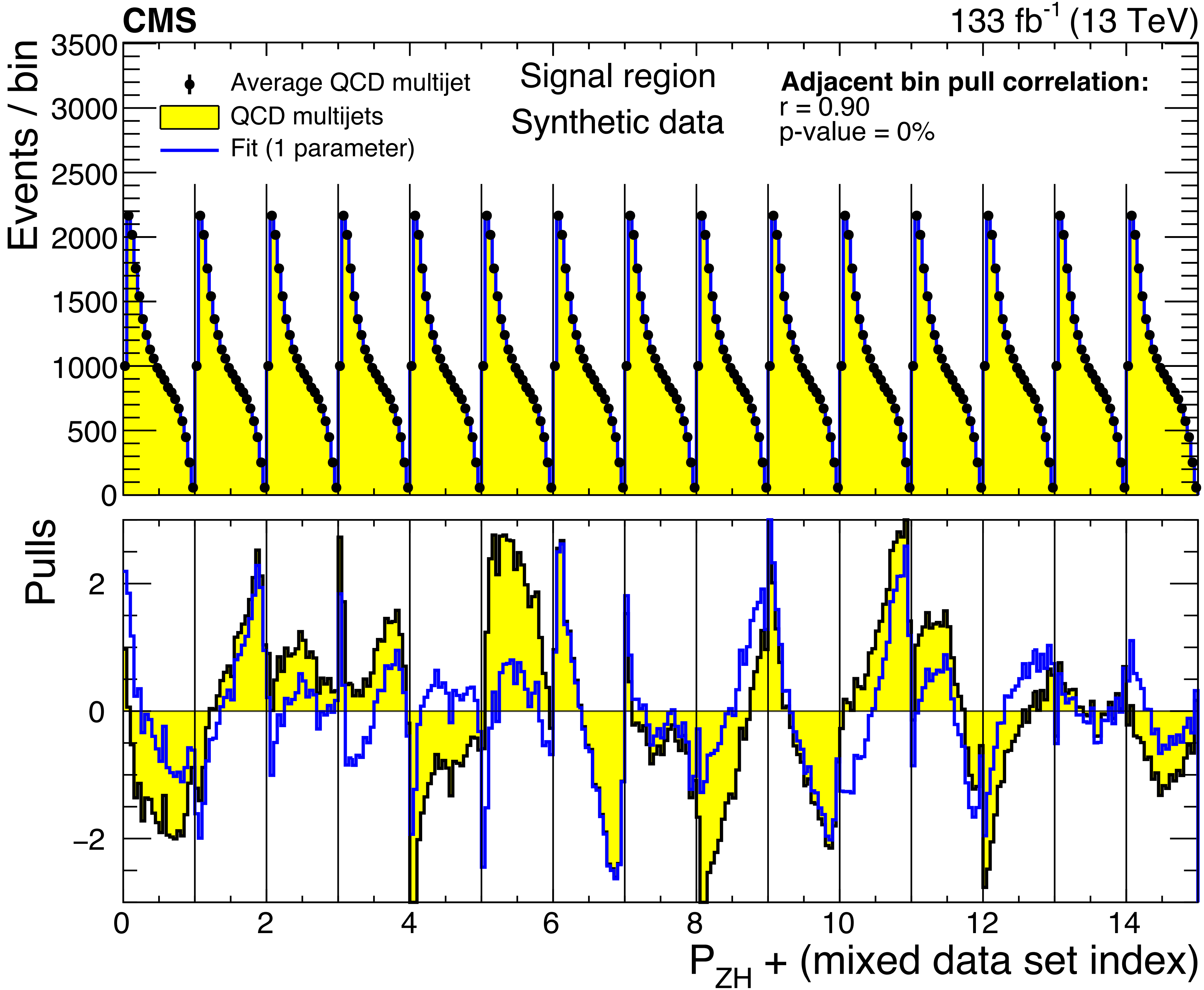
Variance: Arises from multi-variate classifier fit finite dataset CR

Bias: Assumption that density ratio measured in CR is same as SR

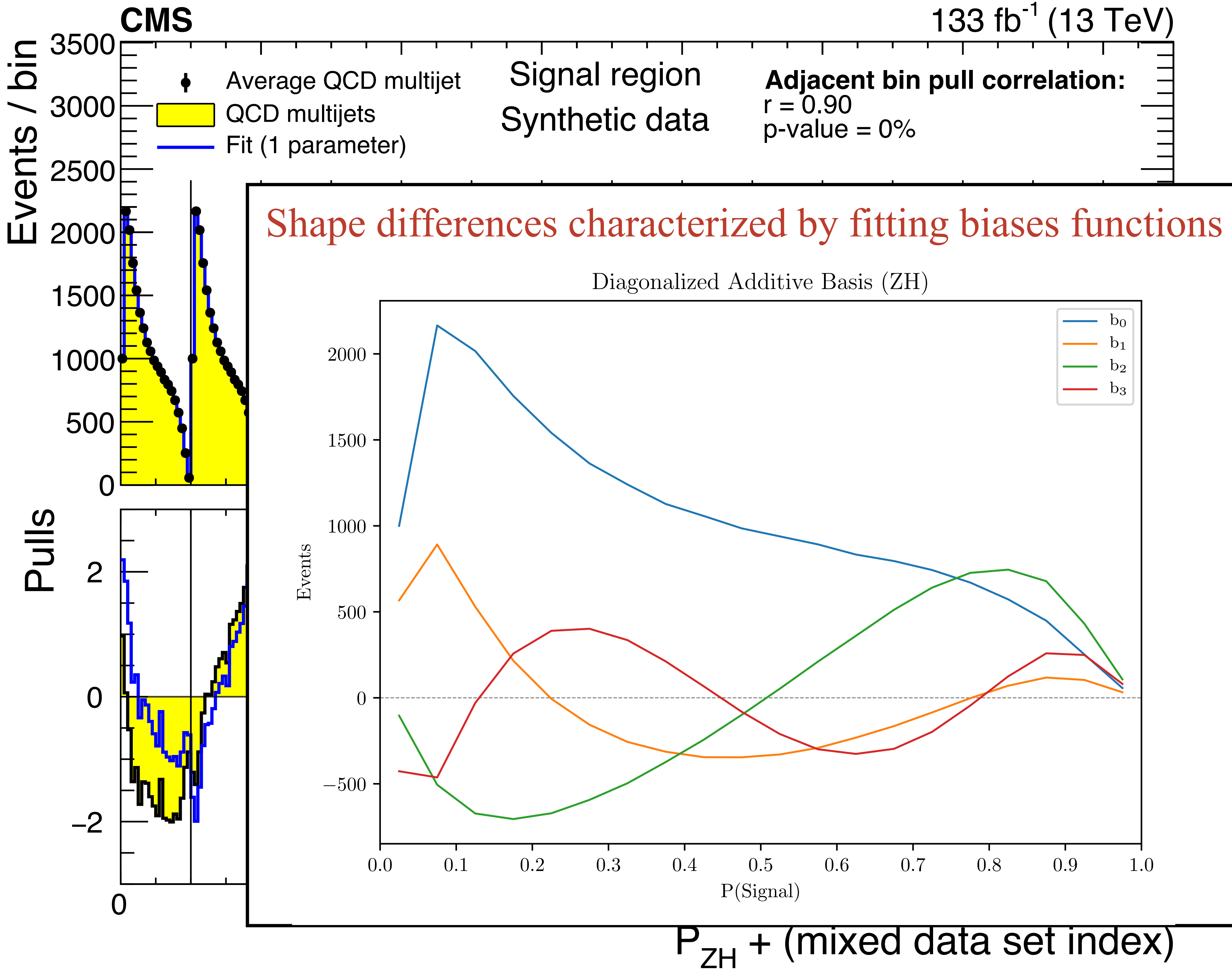
Spurious Signal Can misspecification of the background model look like a signal under the null hypothesis ?

(See backup)

Assumptions rigorously defined for stats audience: arXiv:2208.02807

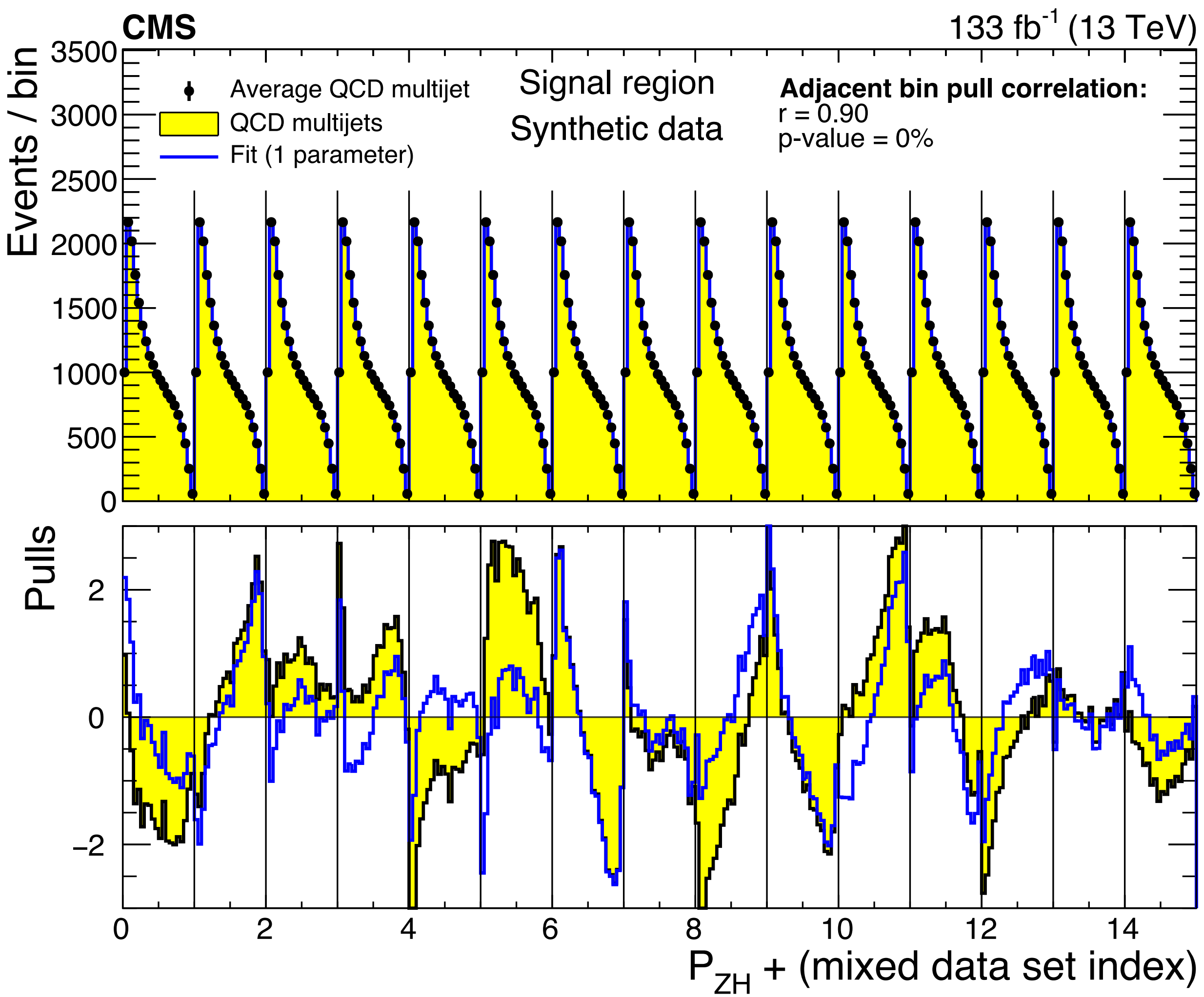


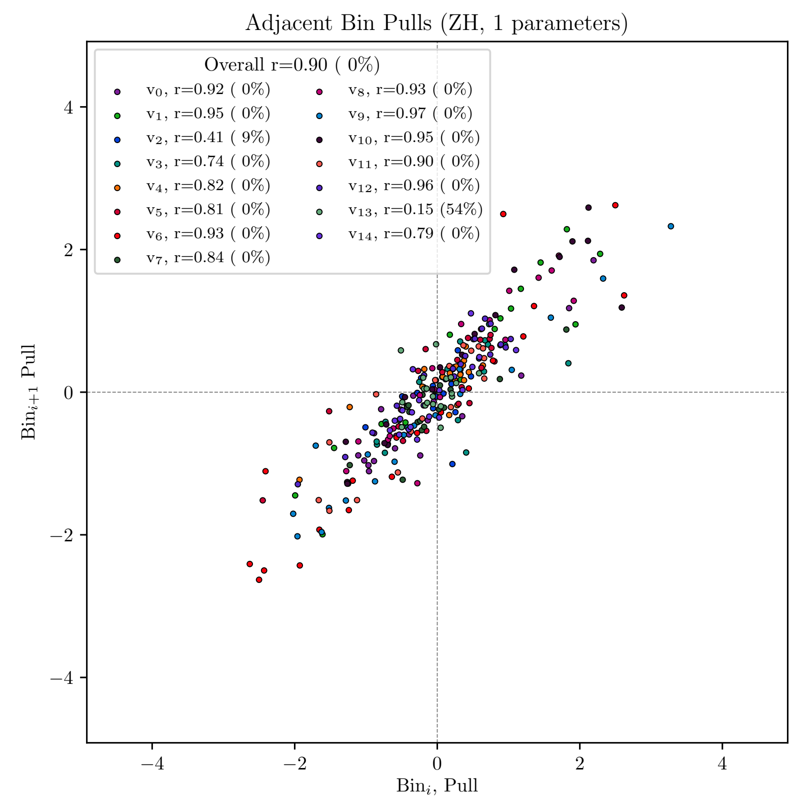
CMS
133 fb-1 (13 TeV)
Events / bin
Average QCD multijet
QCD multijets
Fit (1 parameter)
Signal region
Synthetic data
Adjacent bin pull correlation:
r = 0.90
p-value = 0%
Pulls
P_ZH + (mixed data set index)


CMS
133 fb-1 (13 TeV)
Events / bin
Average QCD multijet
QCD multijets
Fit (1 parameter)
Signal region
Synthetic data
Adjacent bin pull correlation:
r = 0.90
p-value = 0%
Pulls
Shape differences characterized by fitting biases functions
Diagonalized Additive Basis (ZH)
Events
b0
b1
b2
b3
P(Signal)
P_ZH + (mixed data set index)

Quantifying Variance

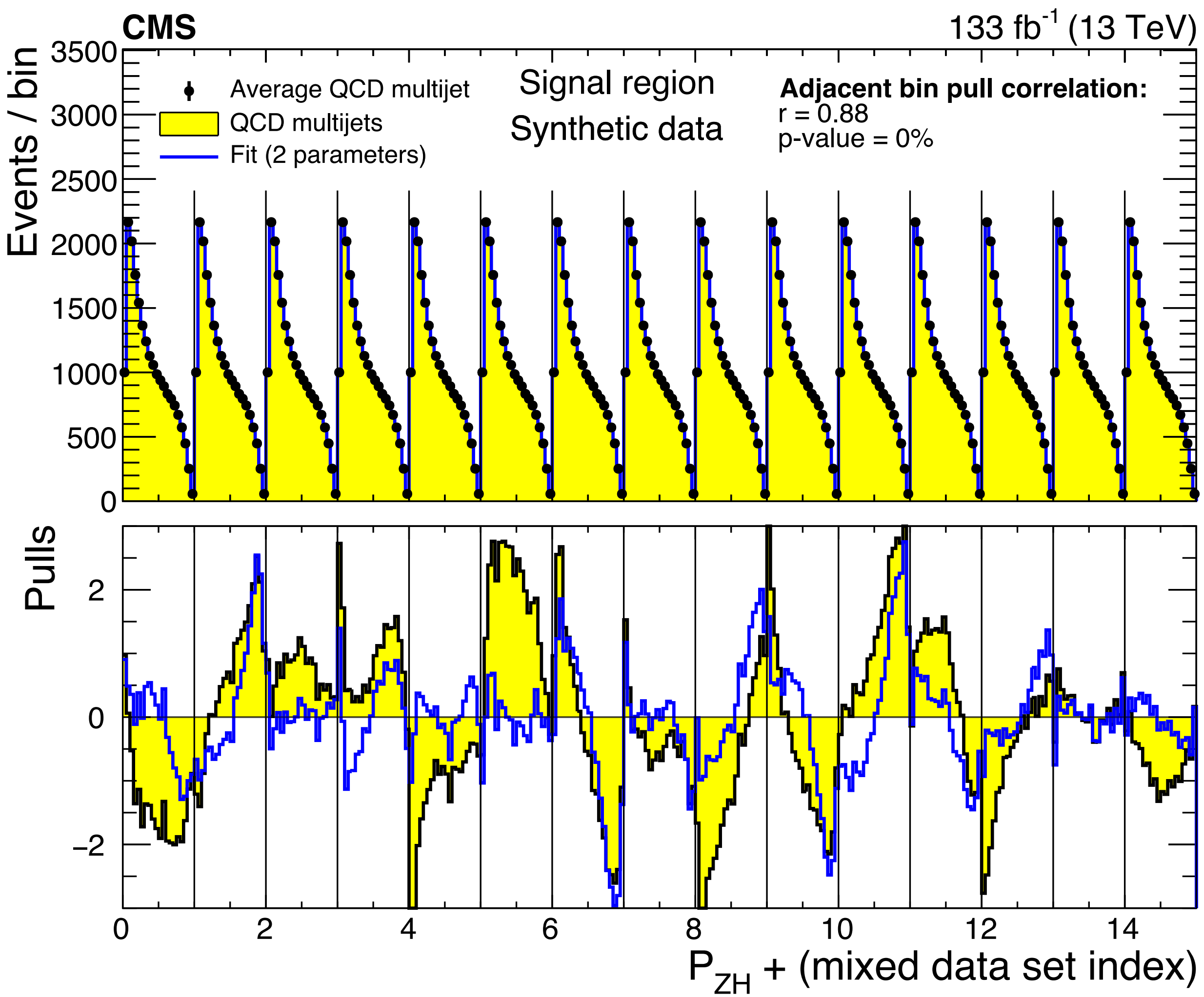
Compare each of the background predictions to the average

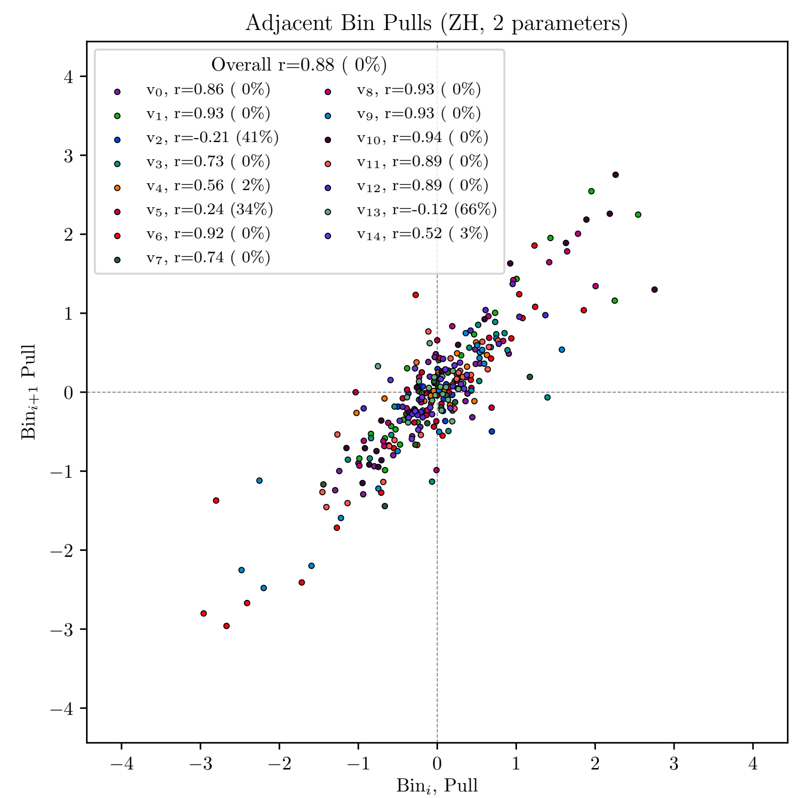




Quantifying Variance

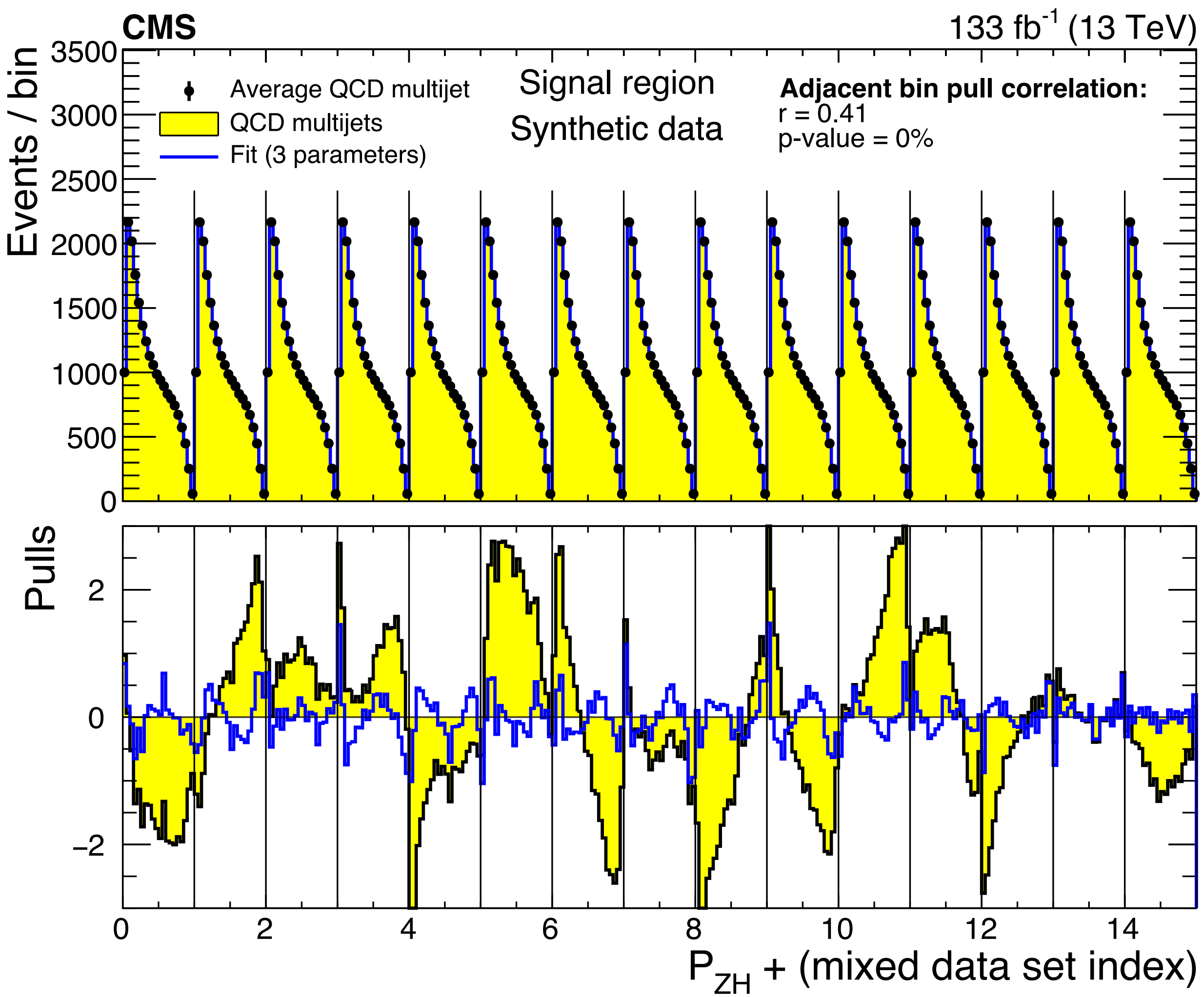
Compare each of the background predictions to the average

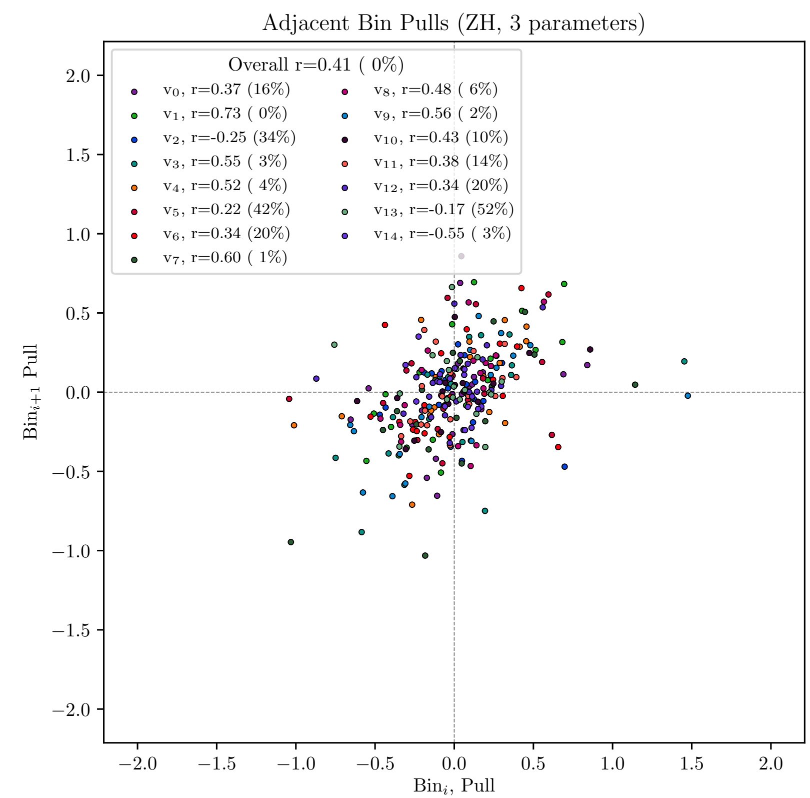




Quantifying Variance

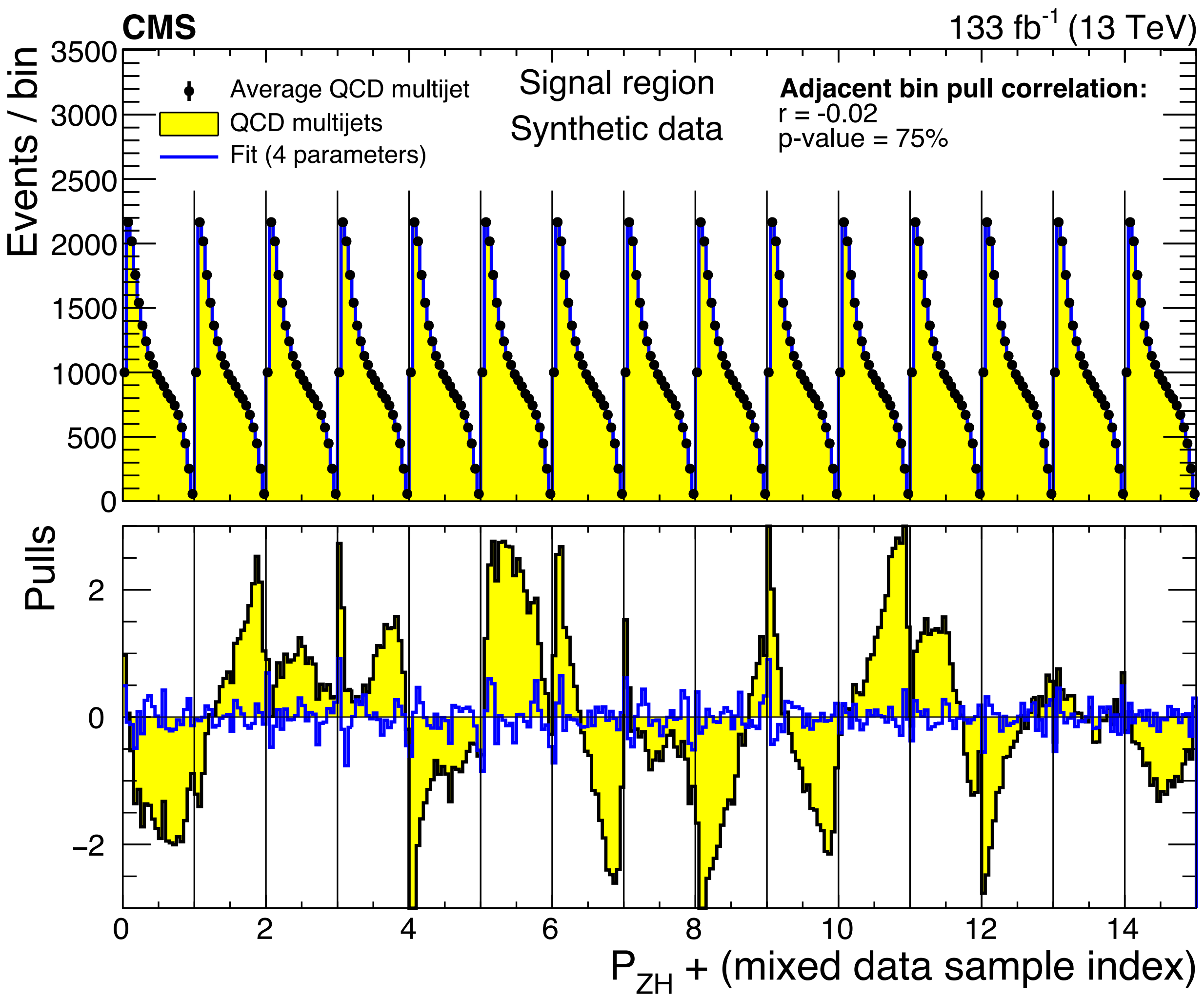
Compare each of the background predictions to the average

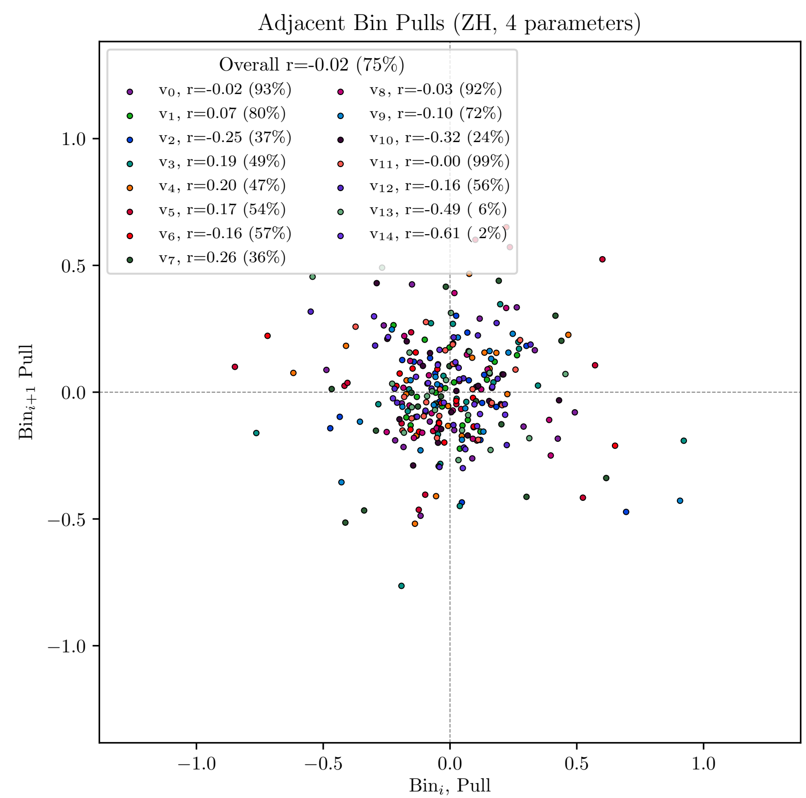




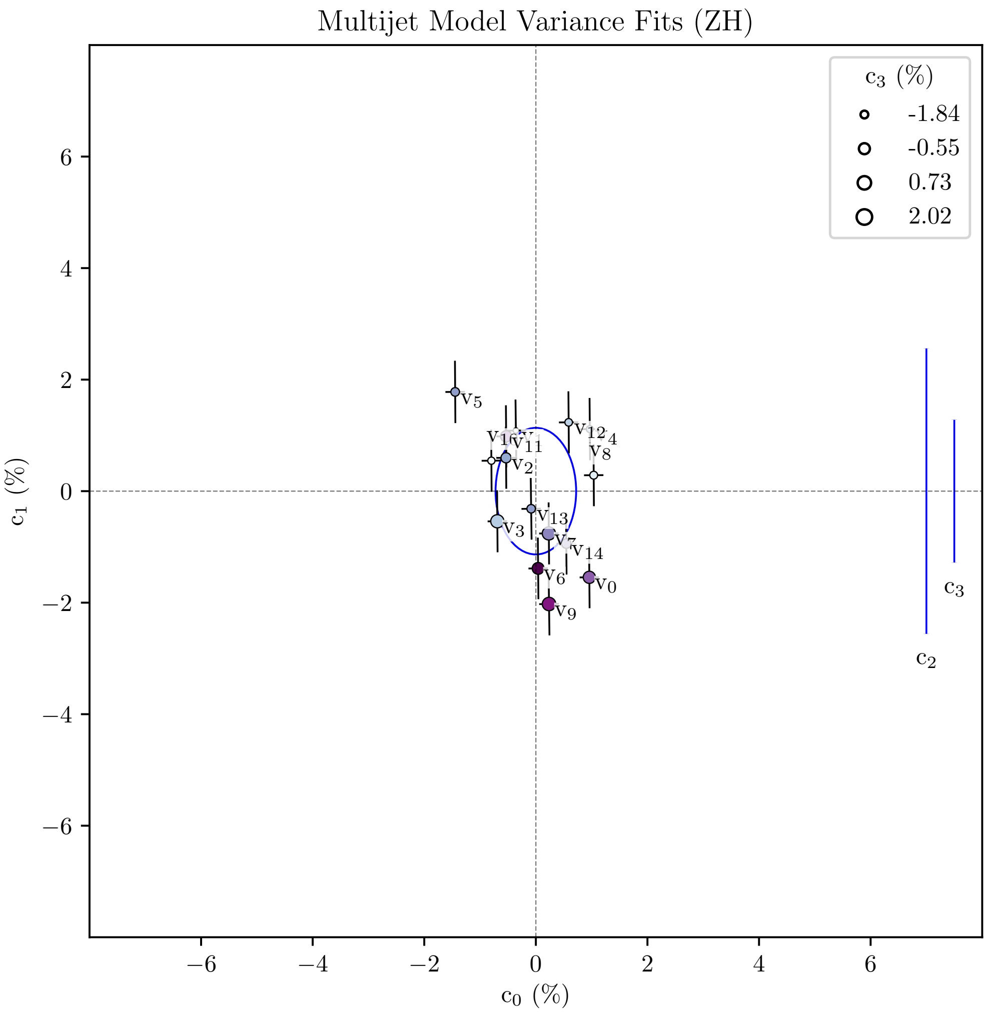
Quantifying Variance

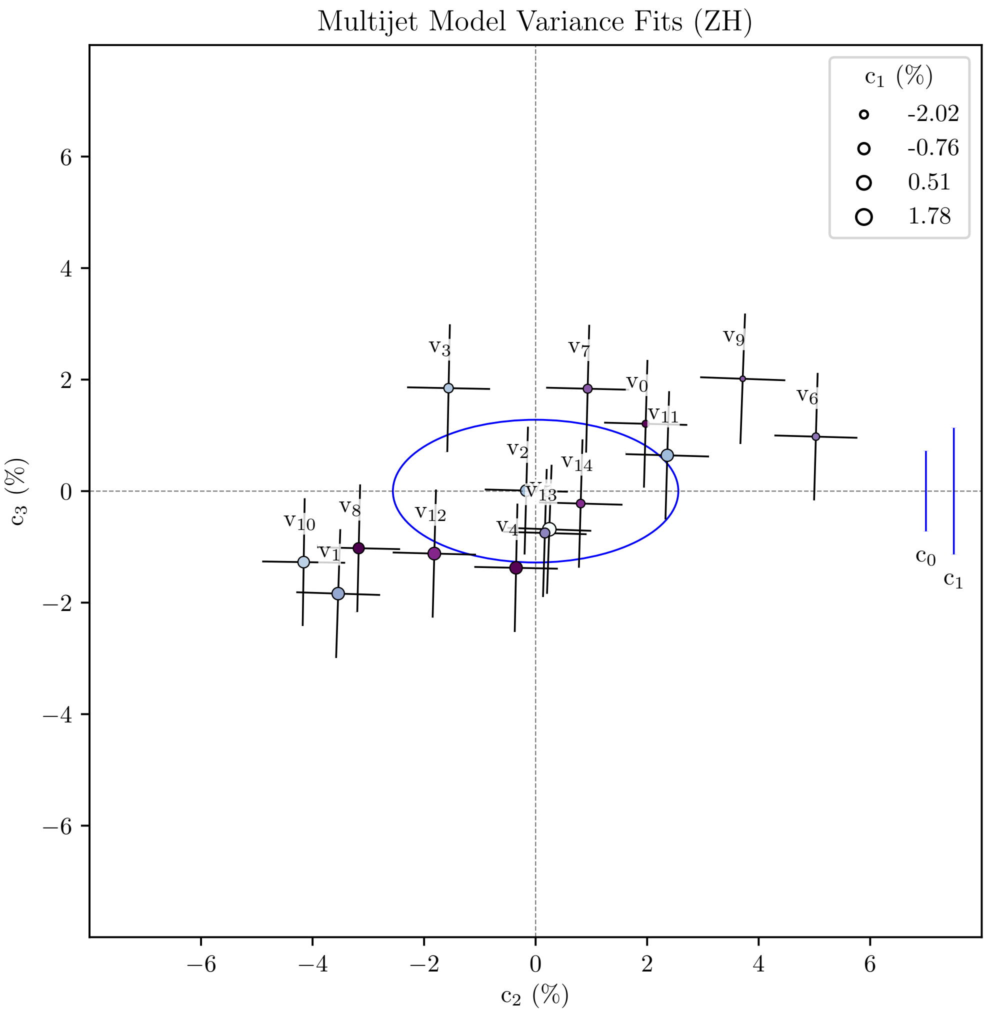
Compare each of the background predictions to the average





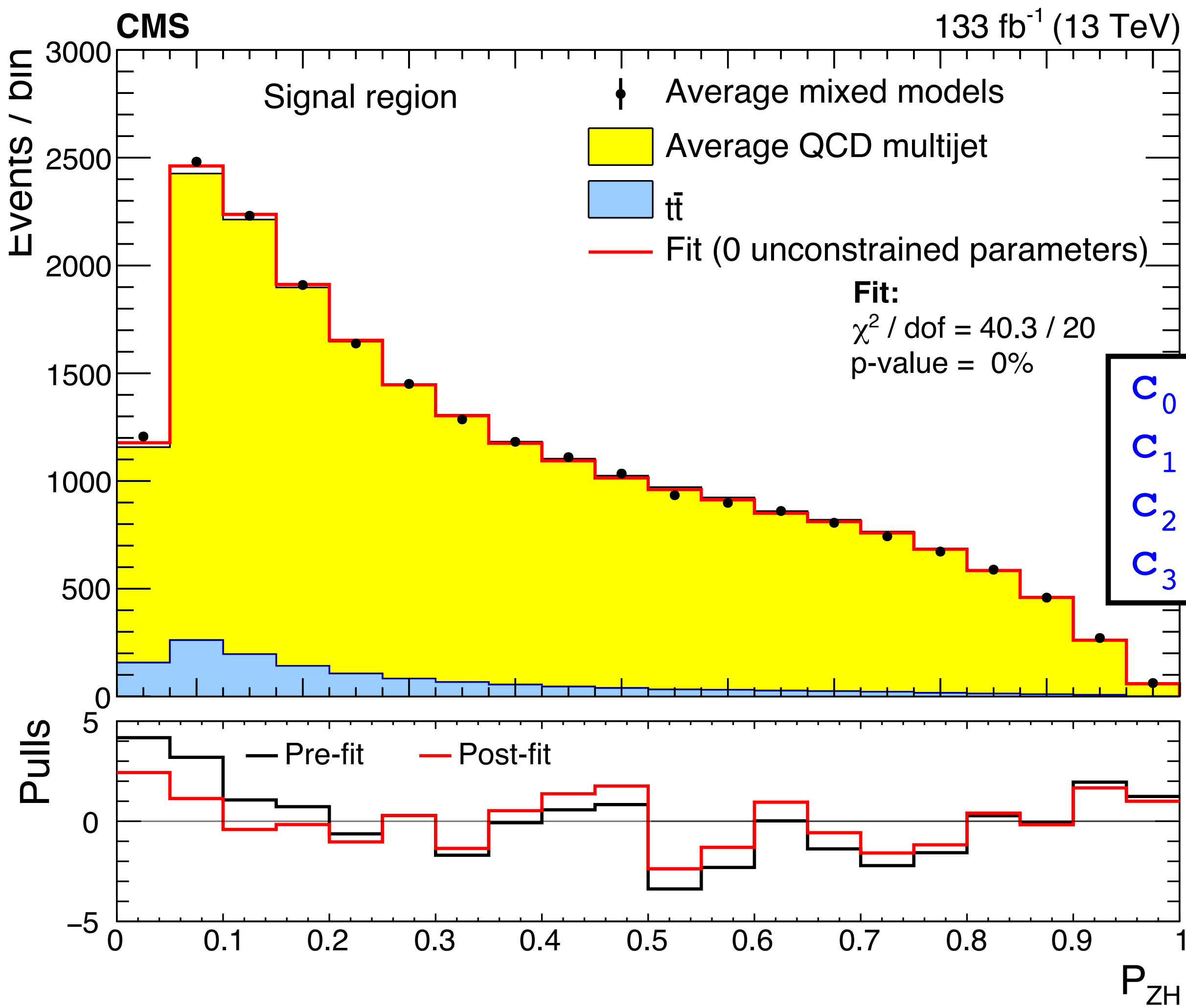
Variance Uncertainties





Extrapolation Uncertainty

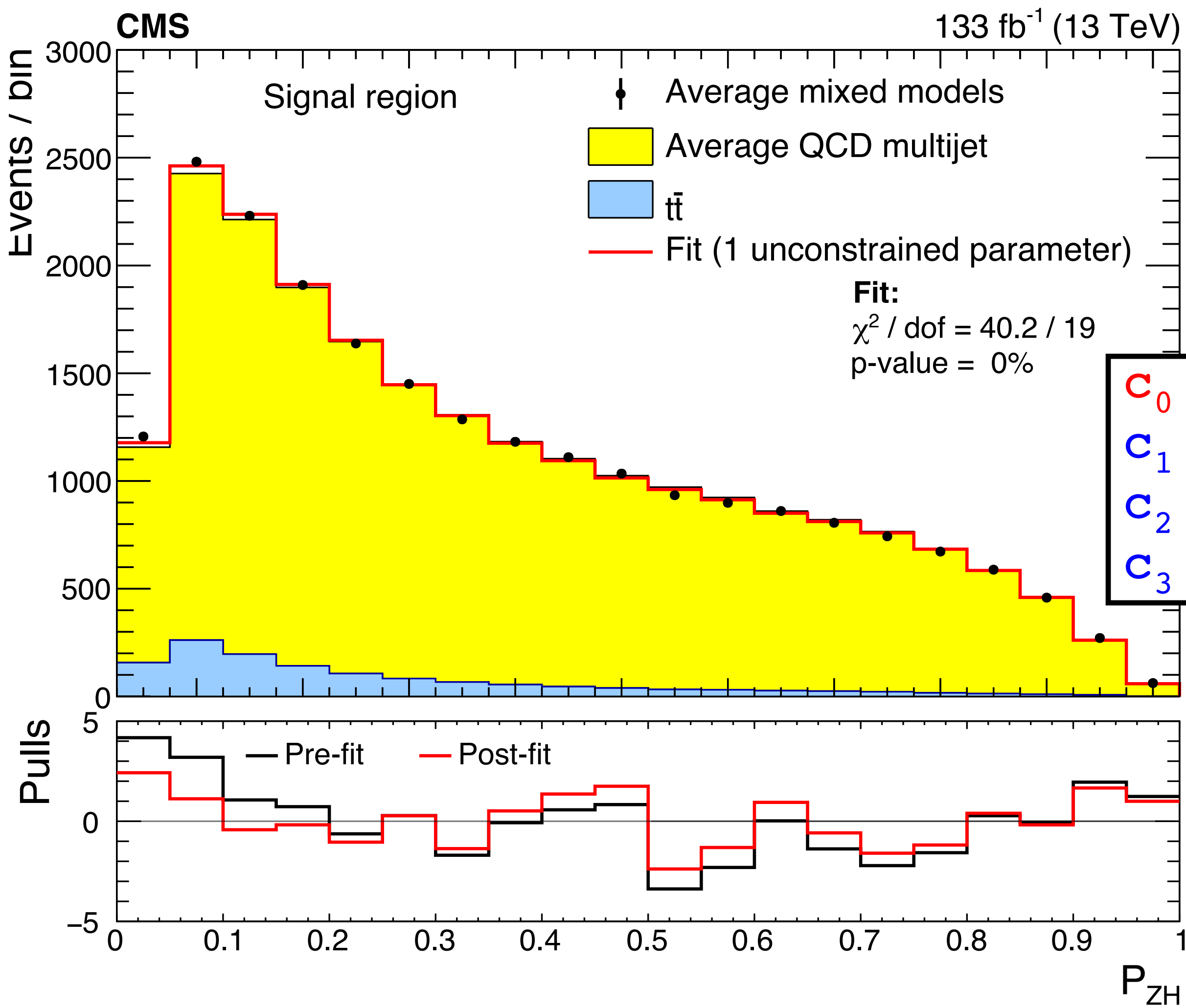
Compare average background predictions to observed yield in (mixed-data) signal region



c_0 = 0.1% : 0.4σ
c_1 = 3.2% : 4.8σ
c_2 =−1.3% : 1.4σ
c_3 = 0.3% : 0.3σ

Extrapolation Uncertainty

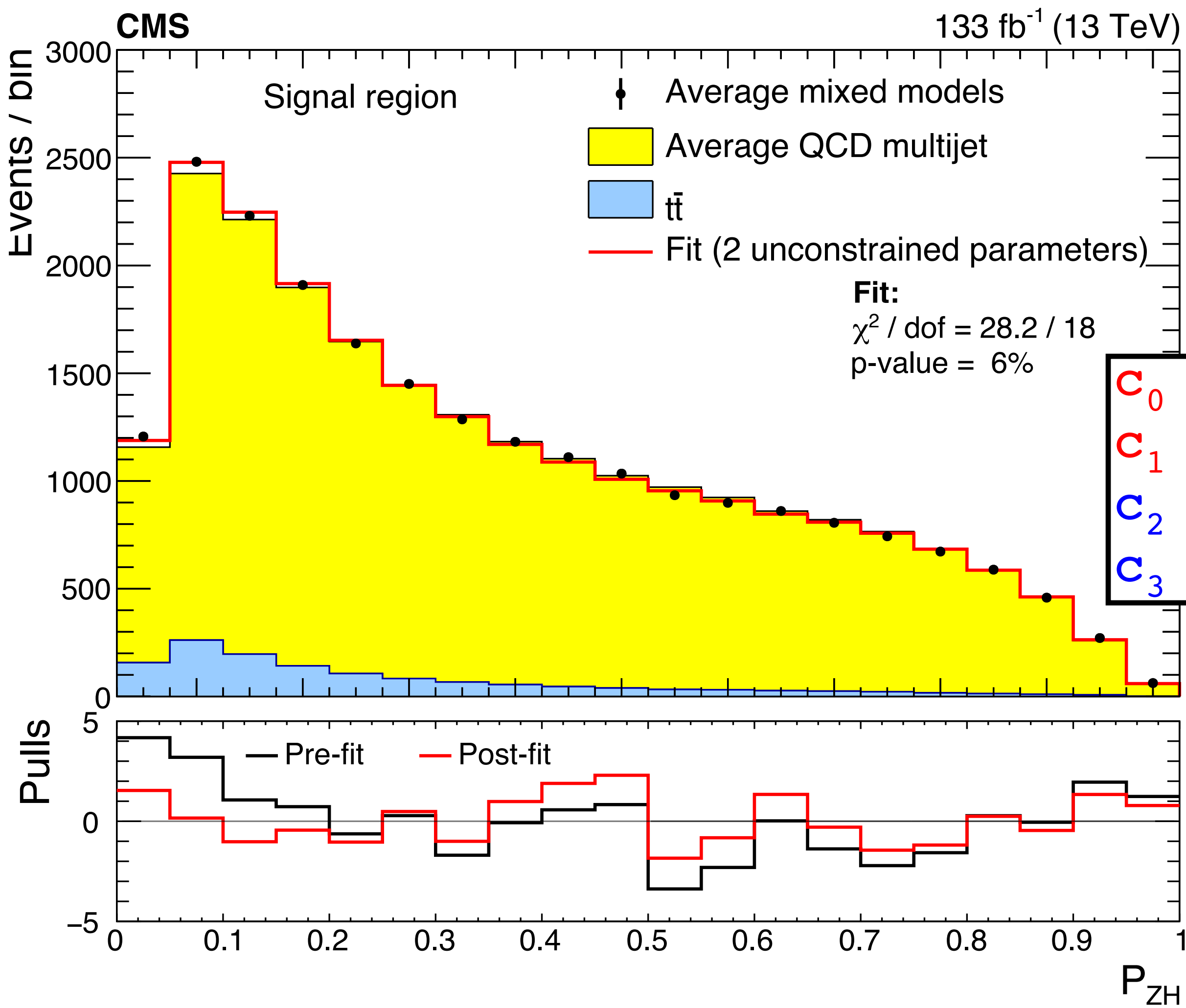
Compare average background predictions to observed yield in (mixed-data) signal region



c_0 = 0.1% : 0.4σ
c_1 = 3.2% : 4.8σ
c_2 =-1.3% : 1.4σ
c_3 = 0.3% : 0.3σ

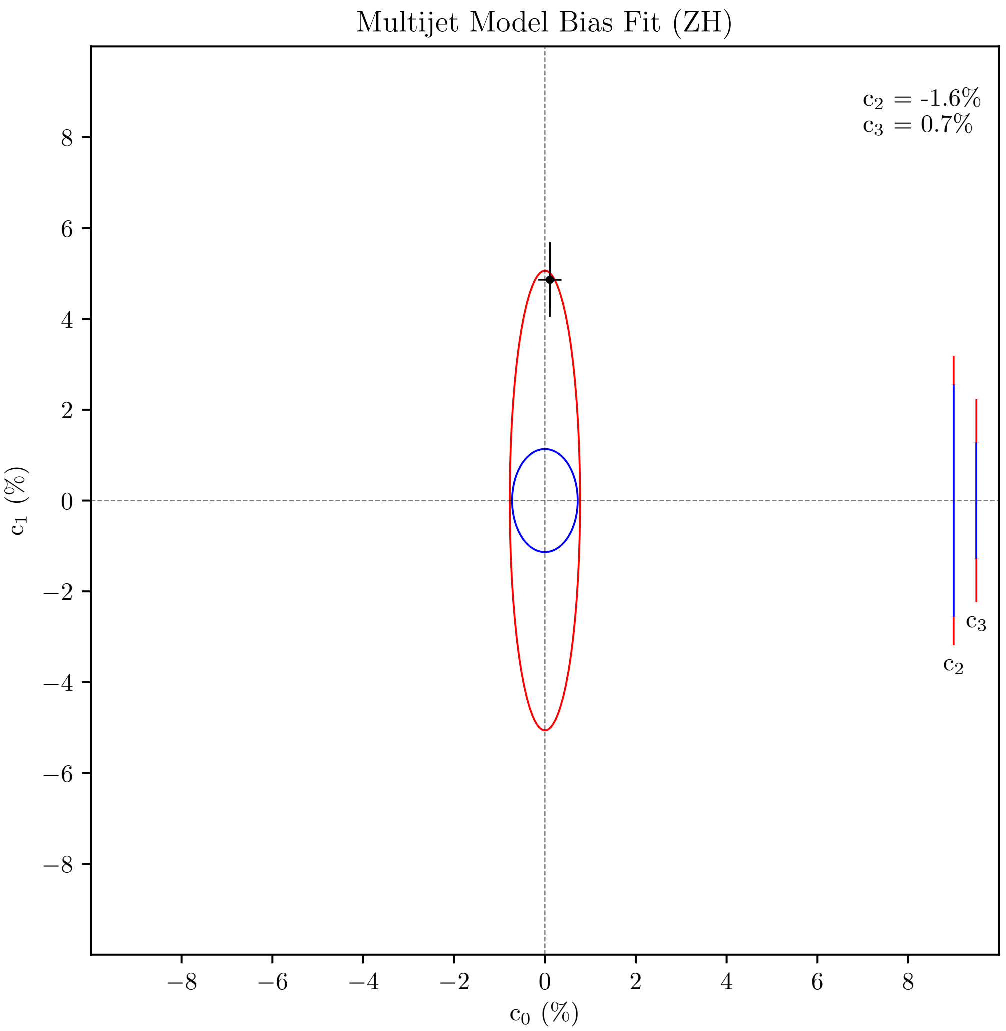
Extrapolation Uncertainty

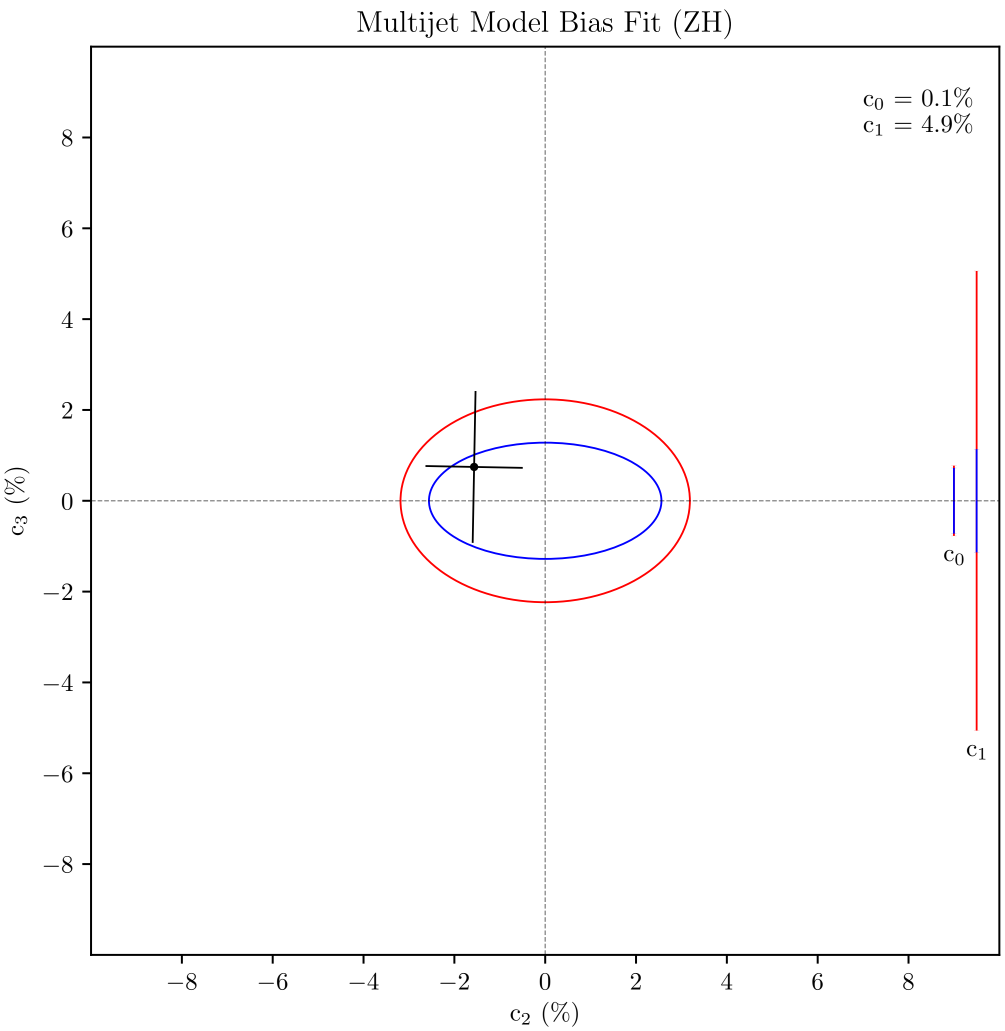
Compare average background predictions to observed yield in (mixed-data) signal region



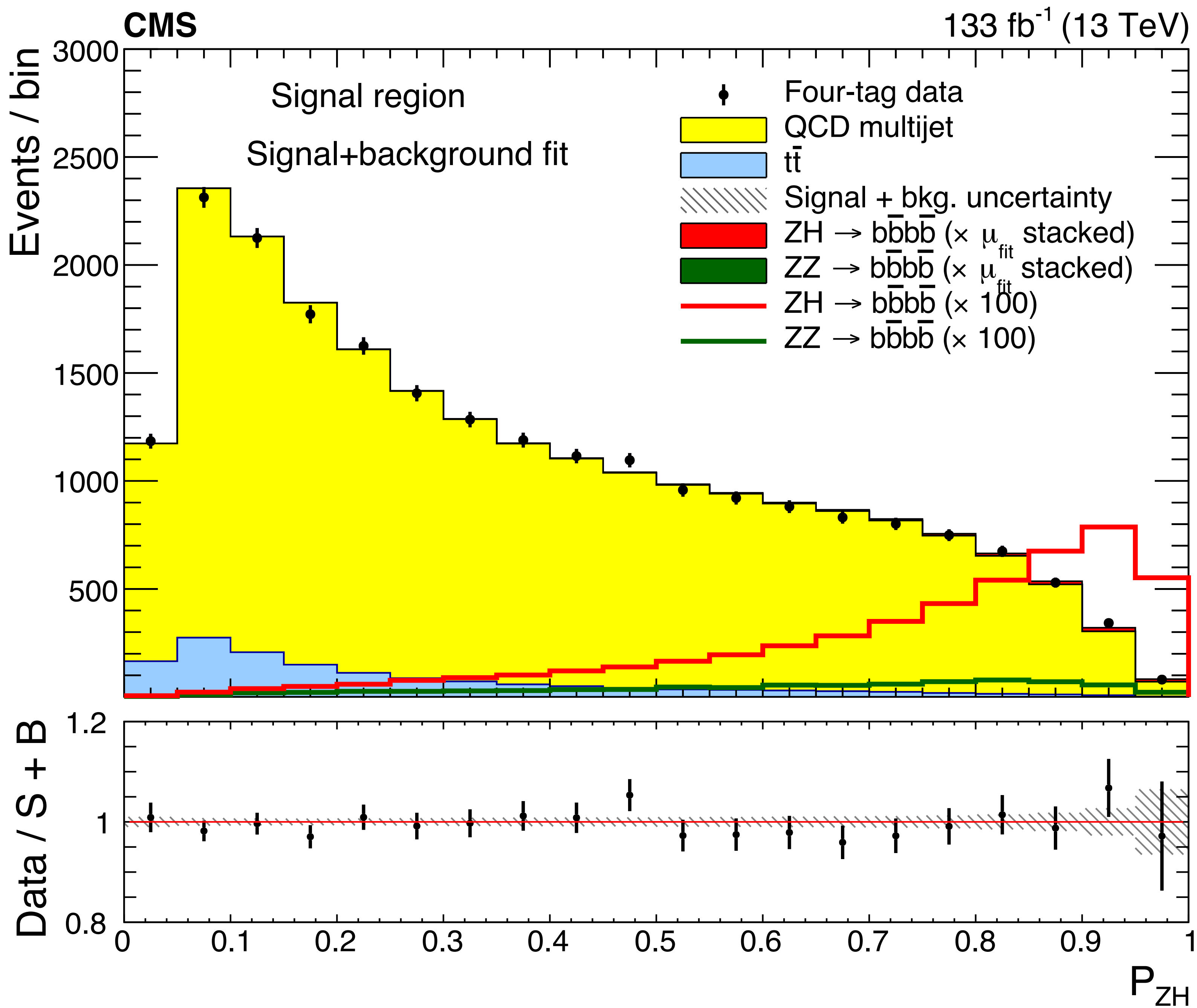
$c_0 = 0.1\% : 0.4\sigma$
$c_1 = 4.9\% : 5.9\sigma$
$c_2 = -1.3\% : 1.4\sigma$
$c_3 = 0.3\% : 0.3\sigma$

Bias Uncertainties





Analysis of 4b Signal Region



Conclusions

Data-driven background ubiquitous in particle physics

Require assumptions w/large hard-to-quantify systematic uncertainties

Synthetic datasets can provide more principled assessment of systematics
  Believe synthetic datasets will be increasing important in future

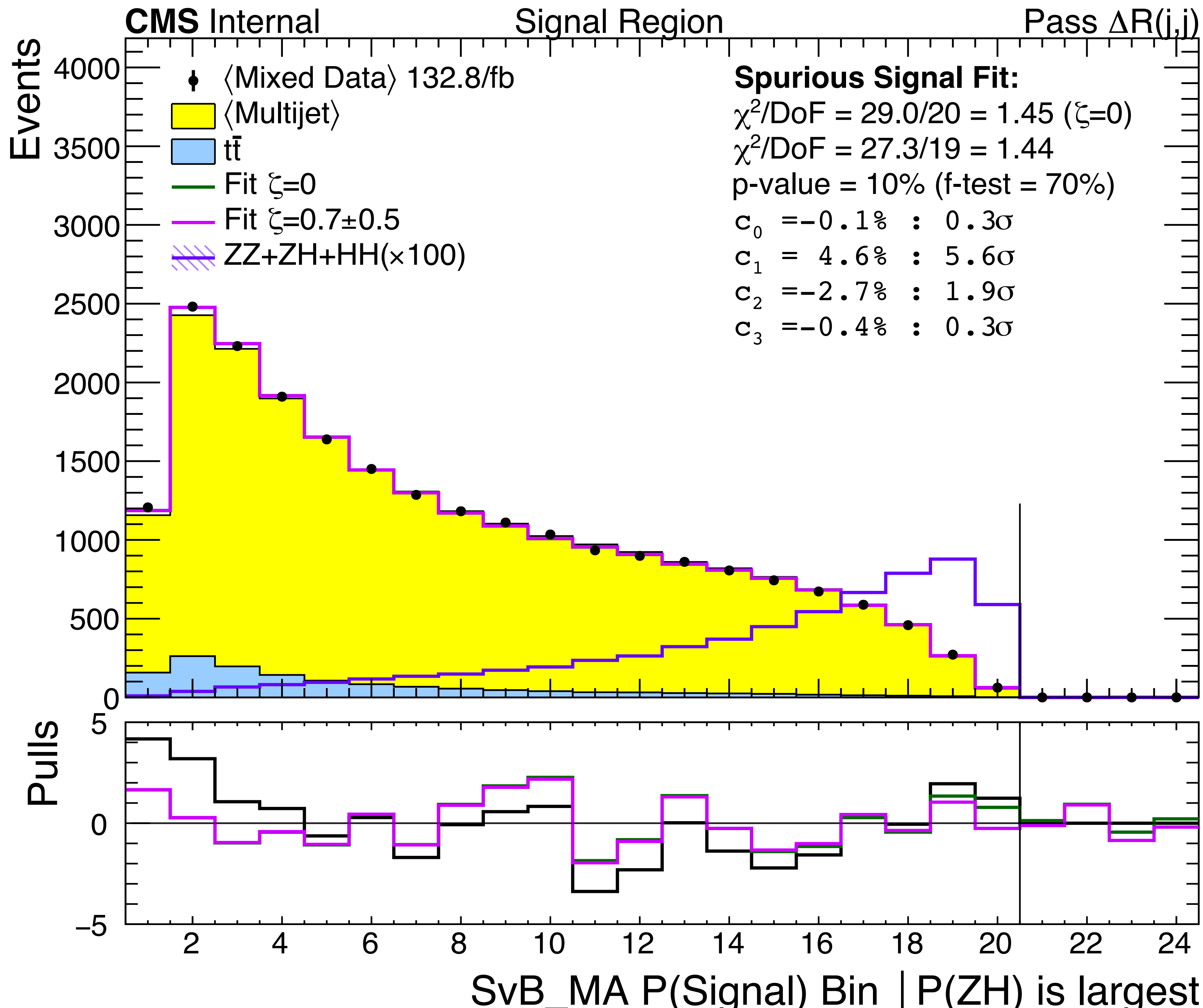
Case study in search for HH→4b more details: arXiv:2403.20241

Believe concept can be generalized beyond HH and high-energy physics

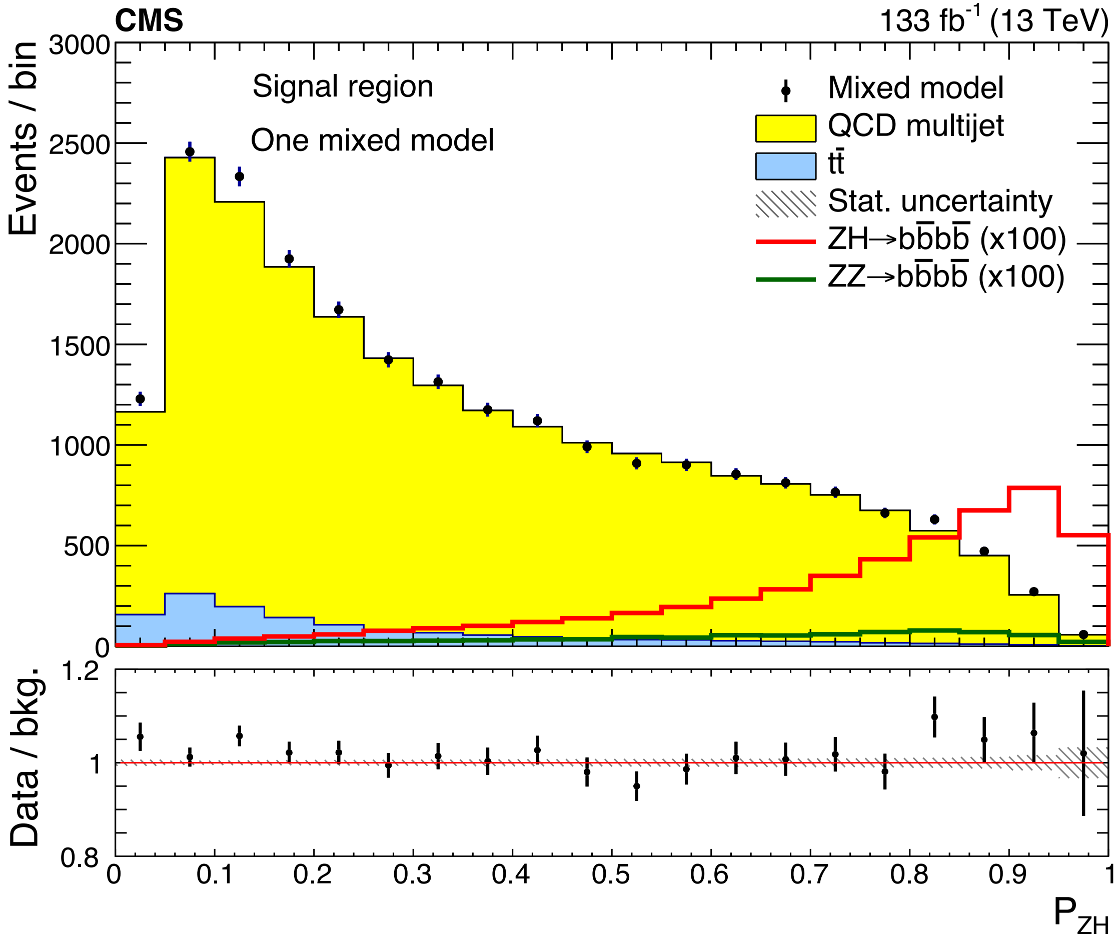
Future directions:
- Reduce variance by k-folding
- Correct bias, take smaller uncertainty
- Larger higher fidelity synthetic datasets

Backup

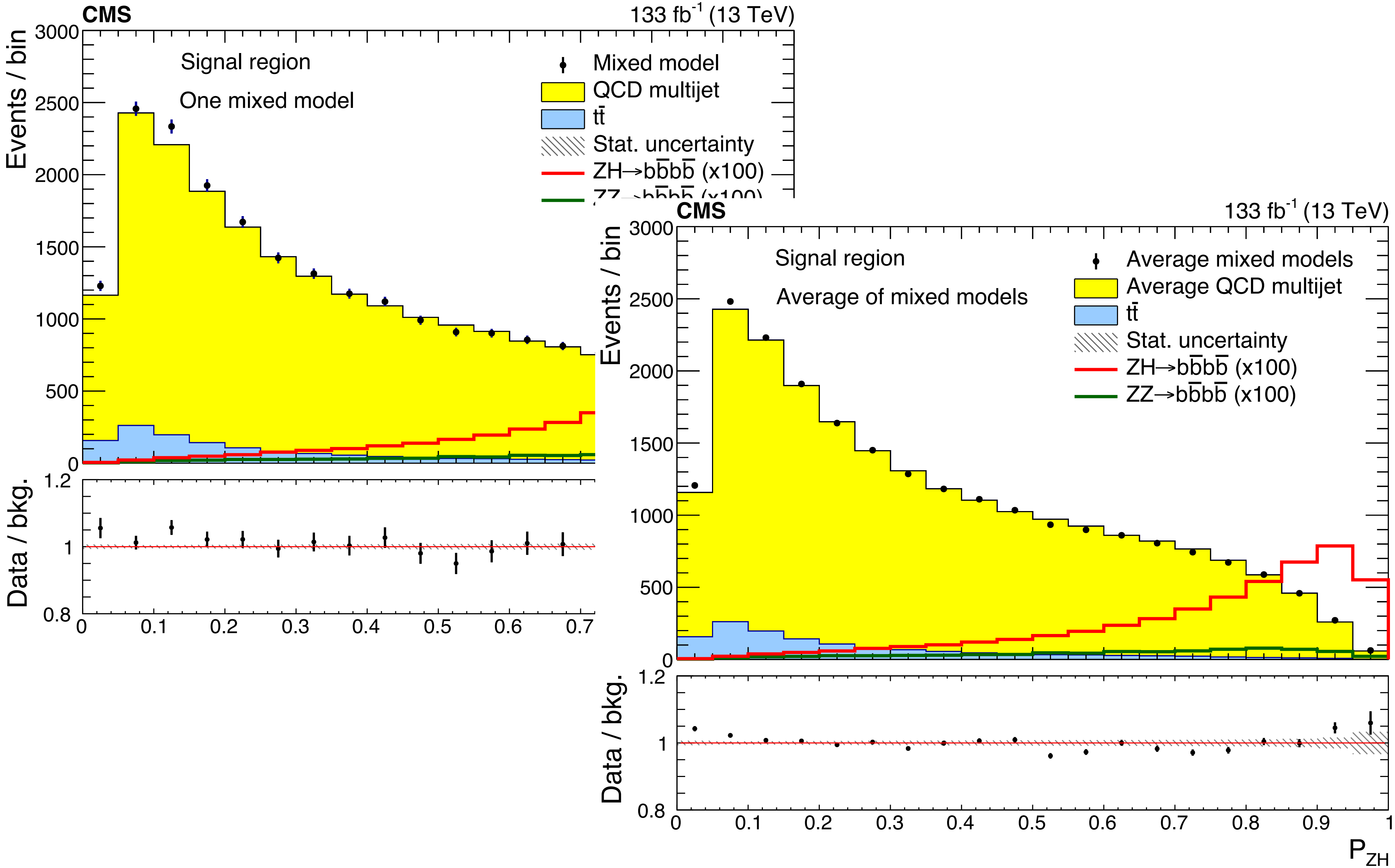
Spurious Signal



Mixed data:


CMS
133 fb^{-1} (13 TeV)
Events / bin
Signal region
One mixed model
Mixed model
QCD multijet
$t\bar{t}$
Stat. uncertainty
ZH→$b\bar{b}b\bar{b}$ (x100)
ZZ→$b\bar{b}b\bar{b}$ (x100)
Data / bkg.
P_{ZH}

Mixed data:



CMS
133 fb^{-1} (13 TeV)
Signal region
One mixed model
Mixed model
QCD multijet
$t\bar{t}$
Stat. uncertainty
ZH→$b\bar{b}b\bar{b}$ (x100)
Events / bin
Data / bkg.
CMS
133 fb^{-1} (13 TeV)
Signal region
Average of mixed models
Average mixed models
Average QCD multijet
$t\bar{t}$
Stat. uncertainty
ZH→$b\bar{b}b\bar{b}$ (x100)
ZZ→$b\bar{b}b\bar{b}$ (x100)
Events / bin
Data / bkg.
P_{ZH}