University of Bari
PhD in Physics
XXXVIII cycle

XXXIV International School "Francesco Romano" on Nuclear, Subnuclear and Astroparticle Physics

Donato Troiano

23 September 2023

Glossary

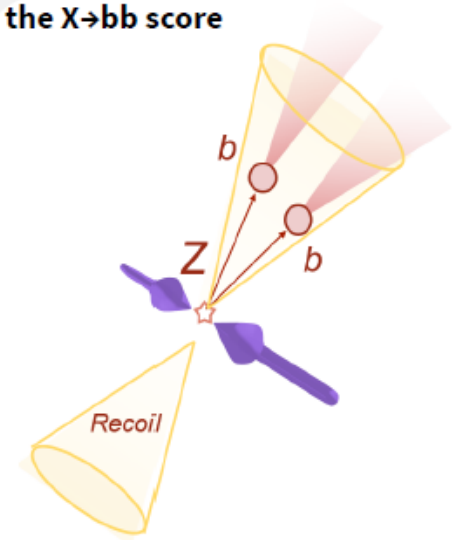
- **AK8 (AK4) jets**: Jets clustered with the anti-*k*T algorithm using a distance parameter of 0.8 (0.4).

- **Soft-drop mass (m_{SD}):** The groomed jet mass obtained from the "soft drop" algorithm with $\beta = 0$ and $z_{cut} = 0.1$.

- **ParticleNet-MD (PN-MD):** A mass-decorrelated particle identification algorithm designed for identifying hadronic decays of highly Lorentz-boosted particles (e.g., X→bb, X→cc, X→qq).
  - PN-MD_BBvsQCD =p(X→bb) / [p(X→bb) + p(QCD)]

AK8 heavy-flavour X→bb tagger activity

Z→bb jet is the handle to check the X→bb score



- Di-jet topology
- Z mass constraint to suppress QCD events
- compare data/MC on high X→bb/cc score region: excess over QCD events should be from Z(bb/cc)+jets events (or W+jets)

- Goal: isolating the Z+jet contribution in Data from the overwhelming QCD backgrounds, comparing with the MC modelling of Z+jets
  - PN-MD not commissioned since CMS Run 2

Trigger selection: PFHT1050, PFJet500, AK8PFJet500, AK8PFJet400_TrimMass30, AK8PFJet420_TrimMass30, AK8PFHT800_TrimMass50

Event selection:

- Leading-p_T AK8 jet: $p_T > 450$ GeV $\wedge$ $|\eta| < 2.4$
- Sub-leading-p_T AK8 jet: $p_T > 200$ GeV $\wedge$ $|\eta| < 2.4$
- $N_e = N_\mu = 0$
- No b-tagged AK4 jet: $p_T > 30$ GeV $\wedge$ $|\eta| < 2.4$ $\wedge$ ΔR(AK4 jet, leading AK8 jet) > 0.8
  - DeepCSV P(b) + P(bb) > 0.4184 [1]

[1] https://github.com/BariGEMJetTau/Hcc/blob/main/HccAna/src/HccAna.cc#L937

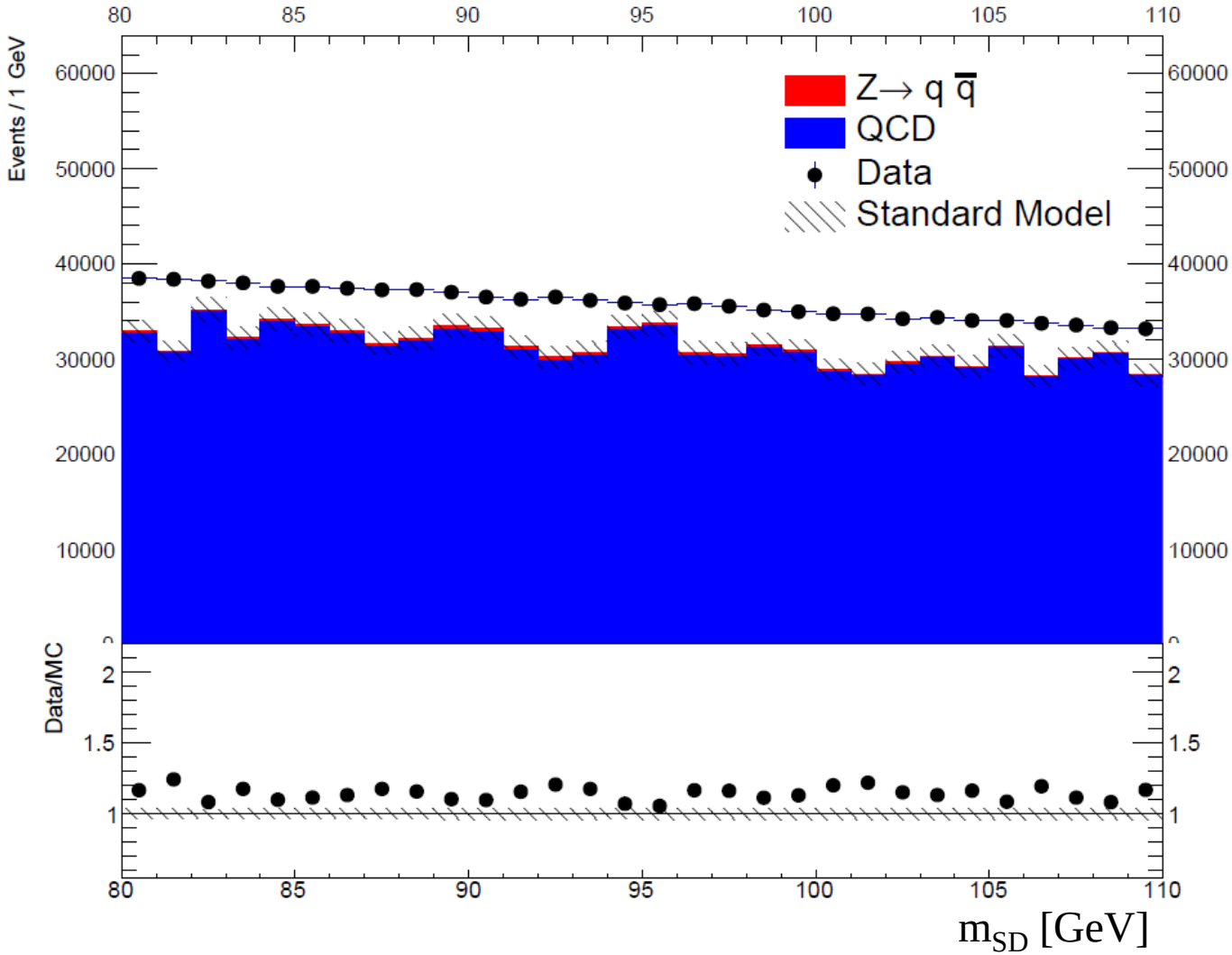
Comparison Data-MC: m_{SD} leading jet

SAMPLES

- MC Zqq pre–Era E (preEE) produced privately [1]
  - $2x10^6$ events
- MC QCD preEE (Run3Summer22MiniAODv3)
  - $19x10^6$ events
- Era C ReReco (5507 pb^{-1})
- Era D ReReco (3417 pb^{-1})

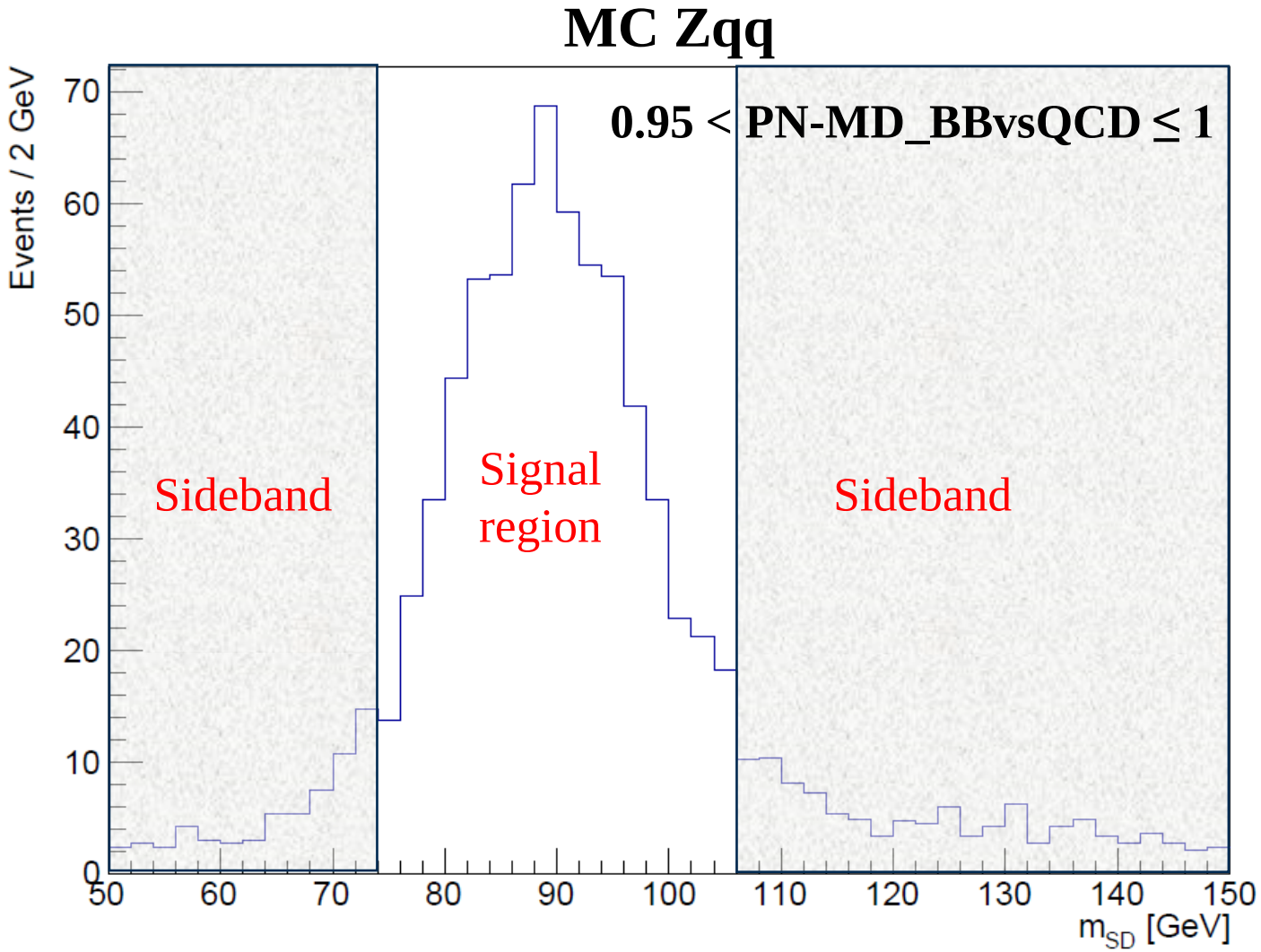
- MC QCD samples are unsuitable to describe Data:
  - Bad Data-MC agreement;
- QCD must be estimated with Data-driven techniques.

[1] DAS



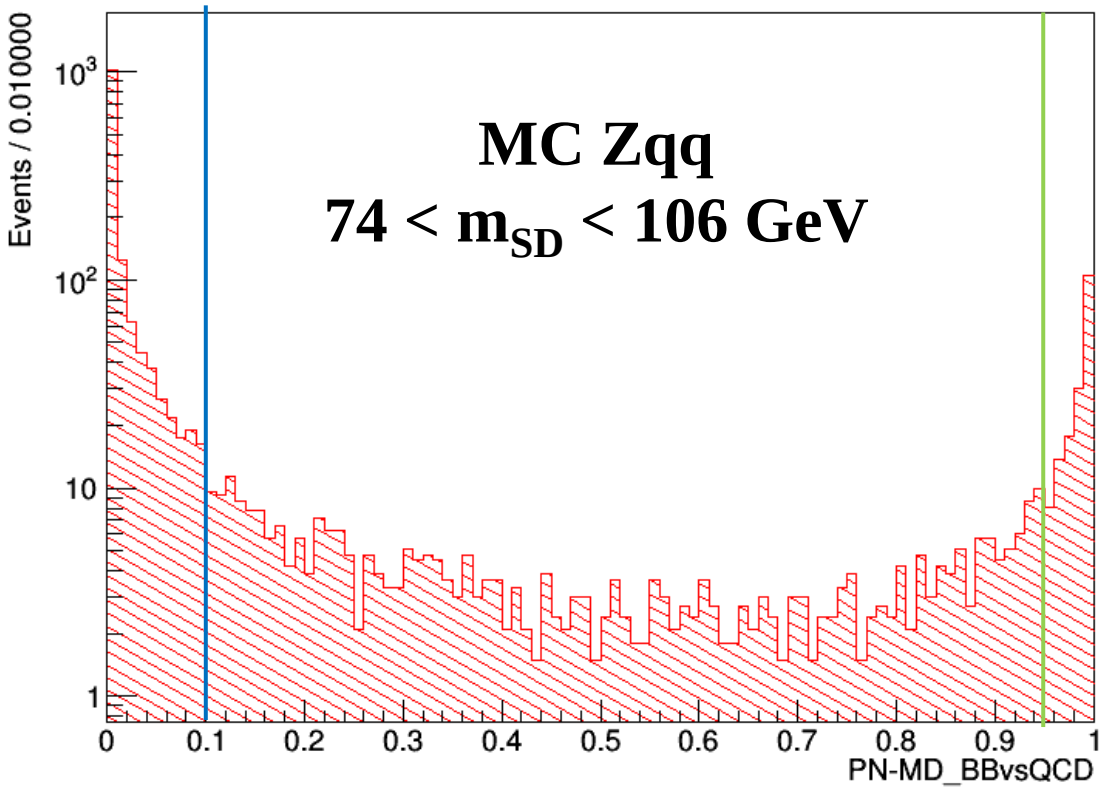
QCD Data-driven estimation phase space

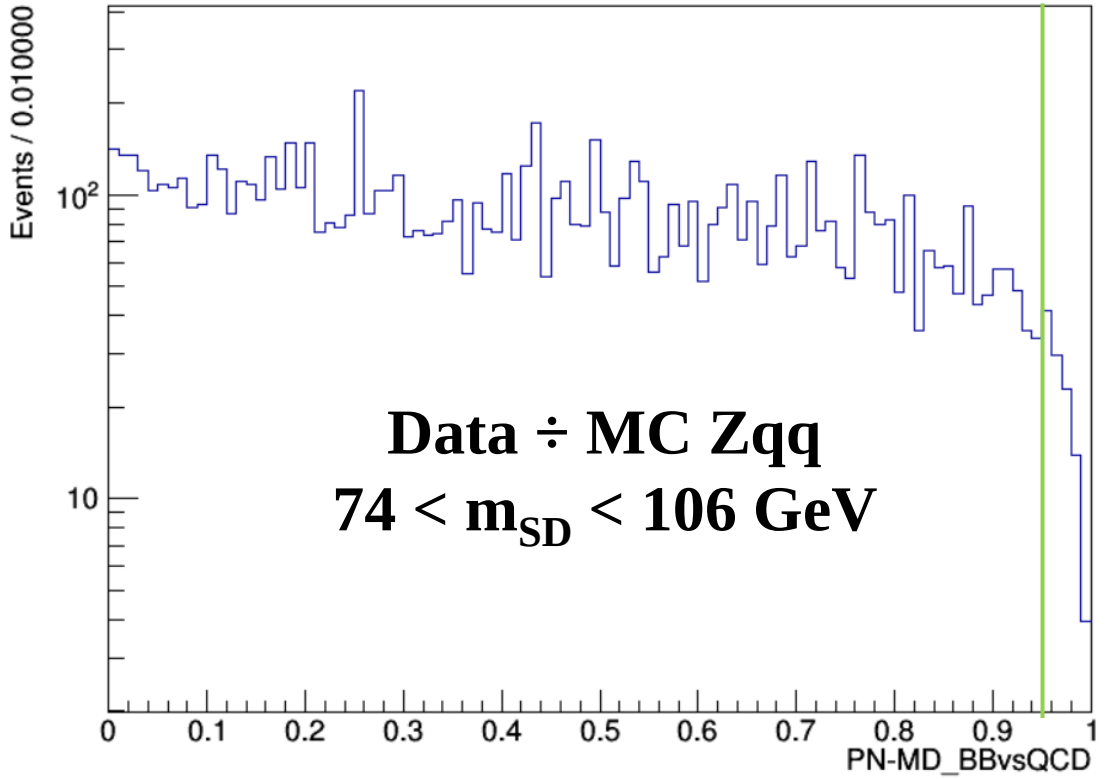
- m_{SD} ranges from 50 to 150 GeV.
- m_{SD} bin width chosen: 2 GeV.
- Signal region from 74 to 106 GeV.
  - MC Z peak almost fully contained within the signal region.



PN-MD_BBvsQCD score regions

- PN-MD_BBvsQCD score regions: 0-0.1 || 0.1-0.26 || 0.26-0.65 || 0.65-0.95 || 0.95-1;
  - Up to 74% located in the region PN-MD_BBvsQCD < 0.1
  - The ratio Data-MC Zqq distribution has a drop at PN-MD_BBvsQCD = 0.95.
  - From 0.1 to 0.95, score regions defined in order to have an equal number of events in each MC Zqq peak region.





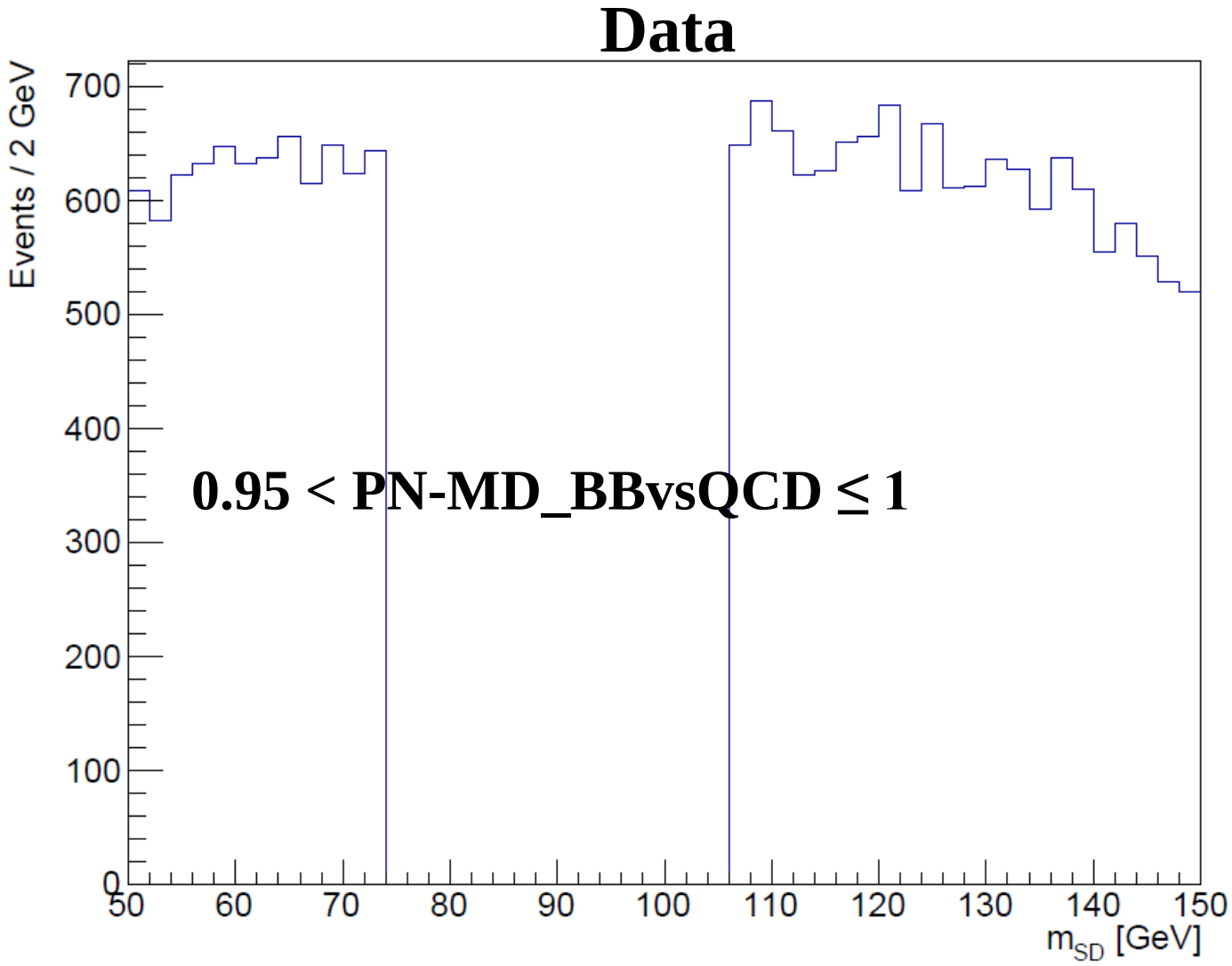
QCD Data-driven estimation technique

1. Plotting, in each PN-MD_BBvsQCD region, the full Data m_{SD} distribution covering the signal region.
   - Data in the sideband regions are mainly QCD events.

2. Fitting the m_{SD} distribution with several functions.

3. Extrapolating in the signal region the QCD m_{SD} distribution from each of the fitting functions.

4. Choosing the "best" one.



Fitting functions

- Several fitting functions adopted:
  1. Polynomials
  2. Chebyshev polynomials
  3. CMS empirical fit function (on the right) [1]

- Number of parameters obtained with the Fisher test (CL of 5%).

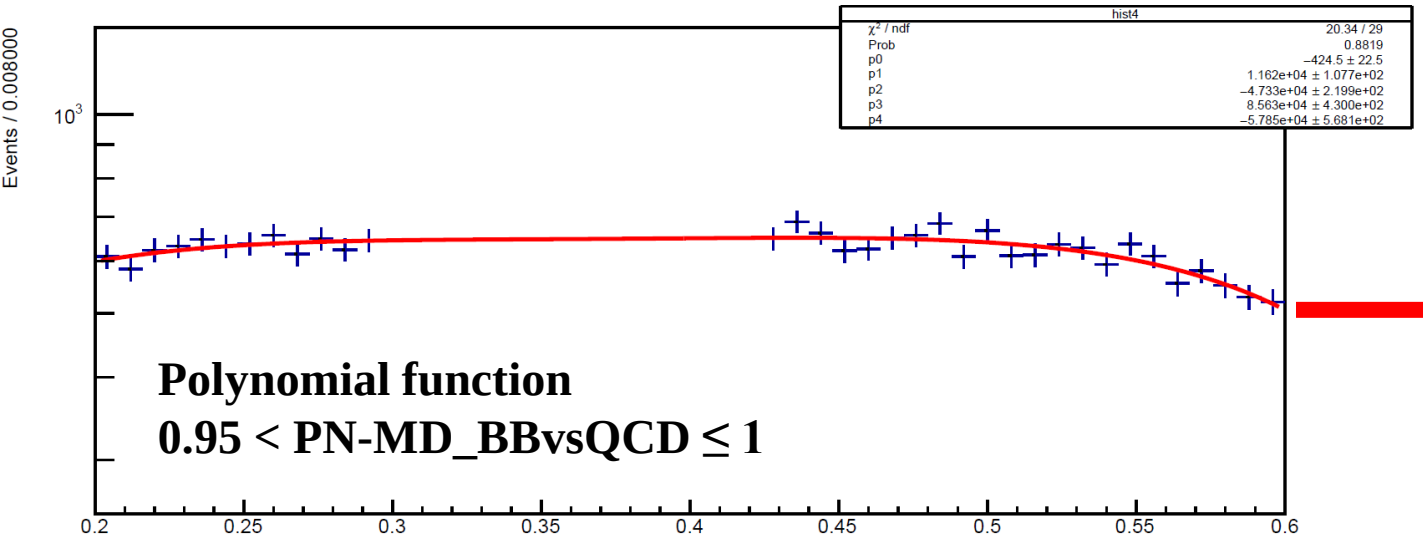
- Fit variable $x = m_{SD}/250$ GeV.

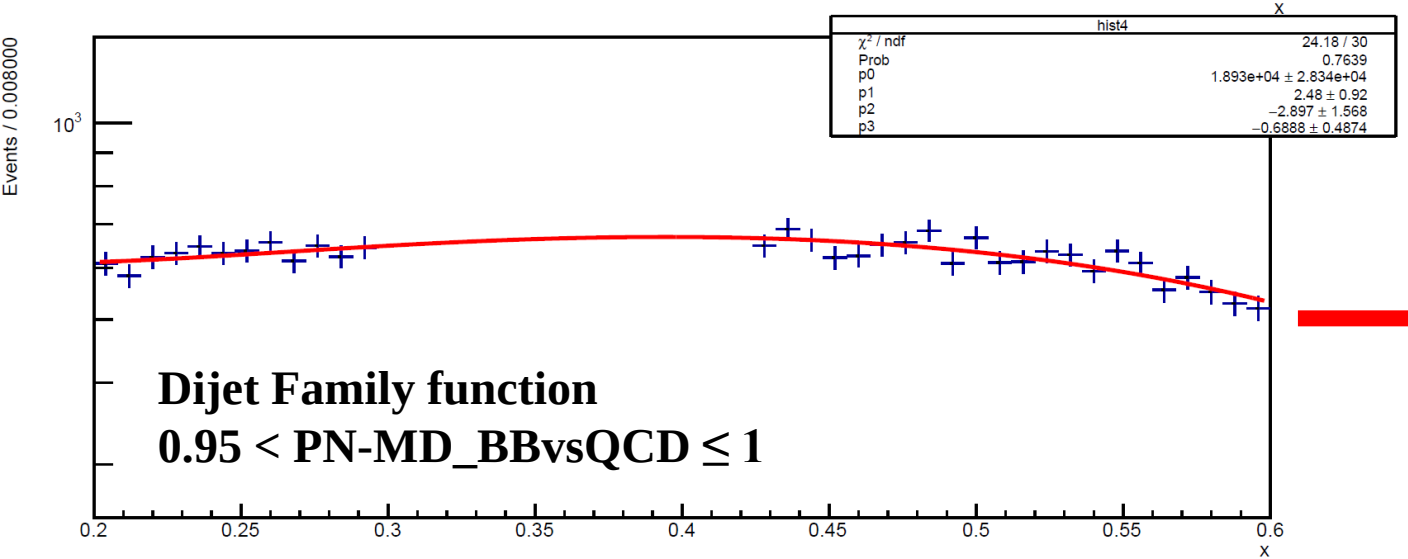
- "Best" function: m_{SD} QCD distribution with the smallest propagated error.

Number of Parameters	Definition
Dijet Family	
3	$p^0(1-x)^{p_1}x^{-p_2}$
4	$p^0(1-x)^{p_1}x^{-(p_2+p_3\log(x))}$
5	$p^0(1-x)^{p_1}x^{-(p_2+p_3\log(x)+p_4\log^2(x))}$
6	$p^0(1-x)^{p_1}x^{-(p_2+p_3\log(x)+p_4\log^2(x))+p_5\log^3(x)}$
Modified Dijet Family	
3	$p^0((1-x)^{1/3})^{p_1}x^{-p_2}$
4	$p^0((1-x)^{1/3})^{p_1}x^{-(p_2+p_3\log(x))}$
5	$p^0((1-x)^{1/3})^{p_1}x^{-(p_2+p_3\log(x)+p_4\log^2(x))}$
6	$p^0((1-x)^{1/3})^{p_1}x^{-(p_2+p_3\log(x)+p_4\log^2(x))+p_5\log^3(x)}$
Polynomial Power Family	
3	$p^0(1+p_1x)^{-p_2}$
4	$p^0(1+p_1x+p_2x^2)^{-p_3}$
5	$p^0(1+p_1x+p_2x^2+p_3x^3)^{-p_4}$
Polynomial Extension Family	
5	$p^0(1-x)^{p_1}(1+p_4x)x^{-(p_2+p_3\log(x))}$
6	$p^0(1-x)^{p_1}(1+p_4x+p_5x^2)x^{-(p_2+p_3\log(x))}$
UA2/ATLAS Family	
4	$p^0\exp-(p_2x+p_3x^2)x^{-p_1}$
5	$p^0\exp-(p_2x+p_3x^2+p_4x^3)x^{-p_1}$
6	$p^0\exp-(p_2x+p_3x^2+p_4x^3+p_5x^4)x^{-p_1}$

[1] https://indico.cern.ch/event/1275872/

Comparison of the fitting functions



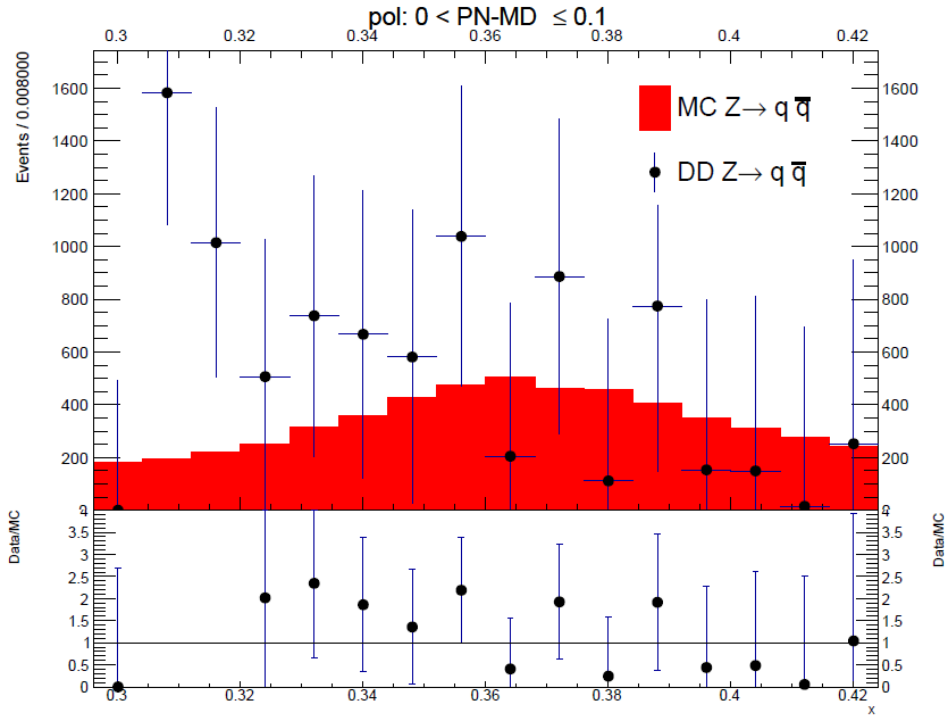


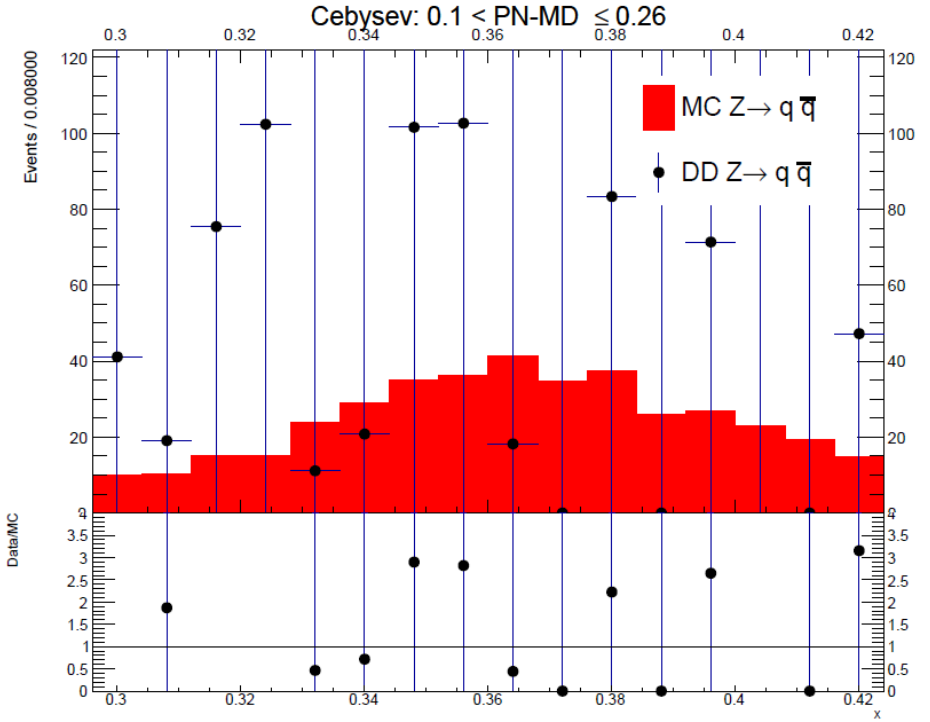
The different function line shapes seem to be quite like each other.

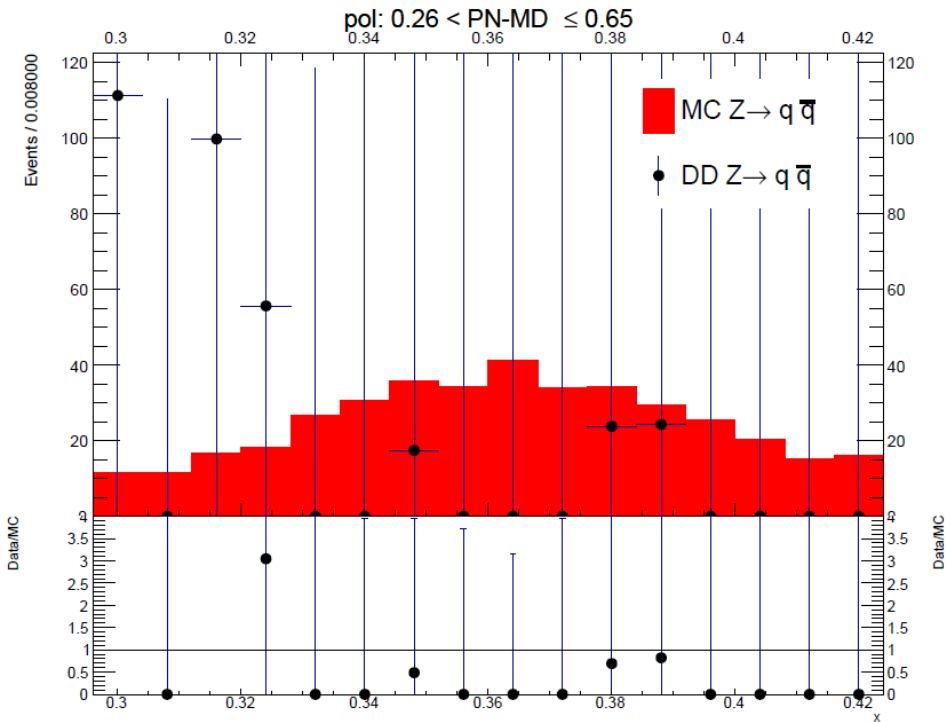
QCD = 10362 ± 236

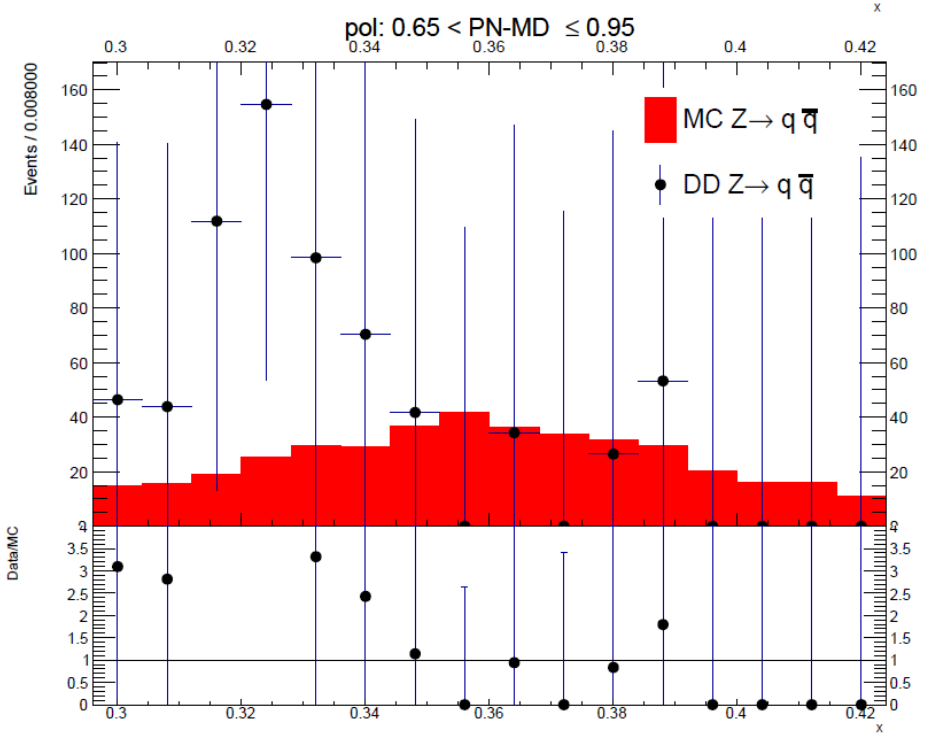
Chosen the QCD distributions with the lower error

QCD = 10605 ± 6105



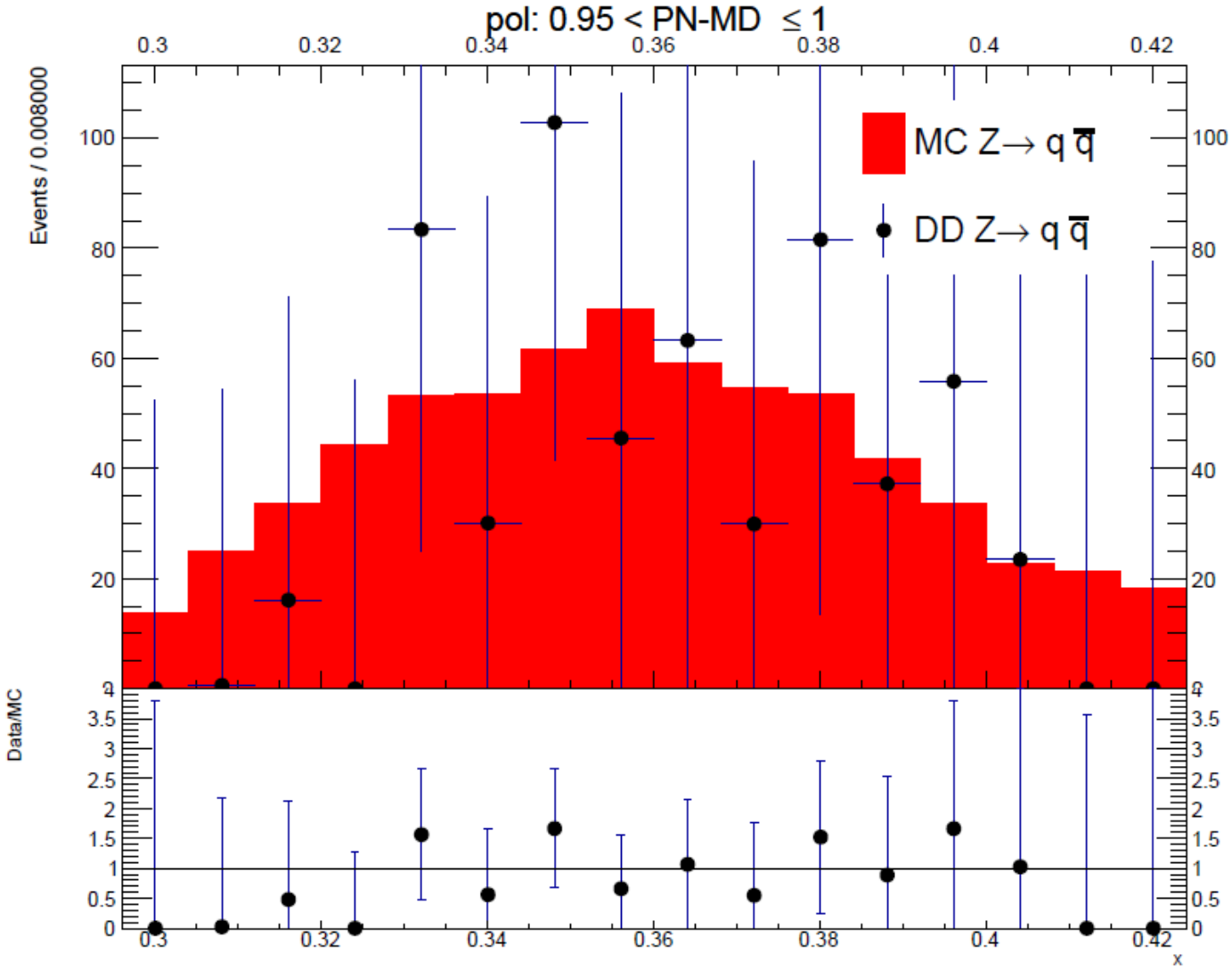






- The functions that suit most well in the different score regions are the polynomials, except in the 0.1-0.26 region where it is the Chebyshev polynomial.
- Data-driven (DD) Zqq estimated as difference between Data and QCD.
  - Bin filled with 0 if there are more QCD events.
- No Data-MC agreement.

PN-MD_BBvsQCD highest score region



- Fair Data-MC agreement.
- Fit function is not as accurate as it was expected since there are many more Data events (mostly QCD) than MC Zqq events.
  - m_{SD} bin Data ~ 600 events
  - m_{SD} bin MC Zqq ~ 60 events

Conclusions

- I have compared the Z+jet Data-driven estimation with the MC modelling for the 2022 preEE.
- Fair Data-MC agreement only at high scores.
- The agreement is not completely fair since there are too much Data events respect to the MC Zqq ones.
- A possible solution to increase the Data-MC agreement is tightening the event selection to further reduce the Data-MC Zqq ratio.

Thanks for your attention