European Research Council
Established by the European Commission

MARIA UBIALI
UNIVERSITY OF CAMBRIDGE

EVALUATING THE FAITHFULNESS OF PDF UNCERTAINTIES IN THE PRESENCE OF INCONSISTENT DATA

Based on arXiv:2503.17447 by A. Barontini, M. Costantini, G. De Crescenzo, S. Forte, MU

OUTLINE

- Closure testing a PDF fit
- Modelling experimental inconsistencies
- How the NNPDF fit responds to inconsistencies
- A new criterion for detecting experimental inconsistencies
- Conclusions and outlook

Based on arXiv:2503.17447 by A. Barontini, M. Costantini, G. De Crescenzo, S. Forte, MU

CLOSURE TESTS

- Assume a given underlying law of Nature: e.g. NNLO QCD predictions for partonic cross section and a given PDF set (for example a random NNPDF replica)

- Generate data with central values given by the "true" law of Nature, and distributed according to experimental covariance matrix.

- Run a fit with a given methodology on this dataset

- Do statistics on "runs of the universe": generate y_0 N_{fit} times with different random noise drawn from experimental distribution

L1 data, each corresponding to a "run of the universe"

$$y_0 = f + \eta$$

$$f = G(w^0)$$

Experimental noise

True value of observables, unknown in real life but identified with G in clos. test

$$\hat{\sigma}_{\rm NNLO} \otimes (f \otimes f)_{\text{“true” PDF}}$$

L2 pseudo-data (to propagate uncertainties in Monte Carlo fits)

$$\mu^{(k)} = y_0 + \epsilon^{(k)}$$

$$k = 1, ..., N_{\rm rep}$$

STATISTICAL ESTIMATORS

$$u_{*,k} = \underset{u_k}{\text{argmin}} \left[\chi^{2(k)}_{\text{val}} | \underset{u_k}{\text{argmin}}\, \chi^{2(k)}_{\text{tr}} \right], \quad k = 1, ..., N_{\text{rep}}$$

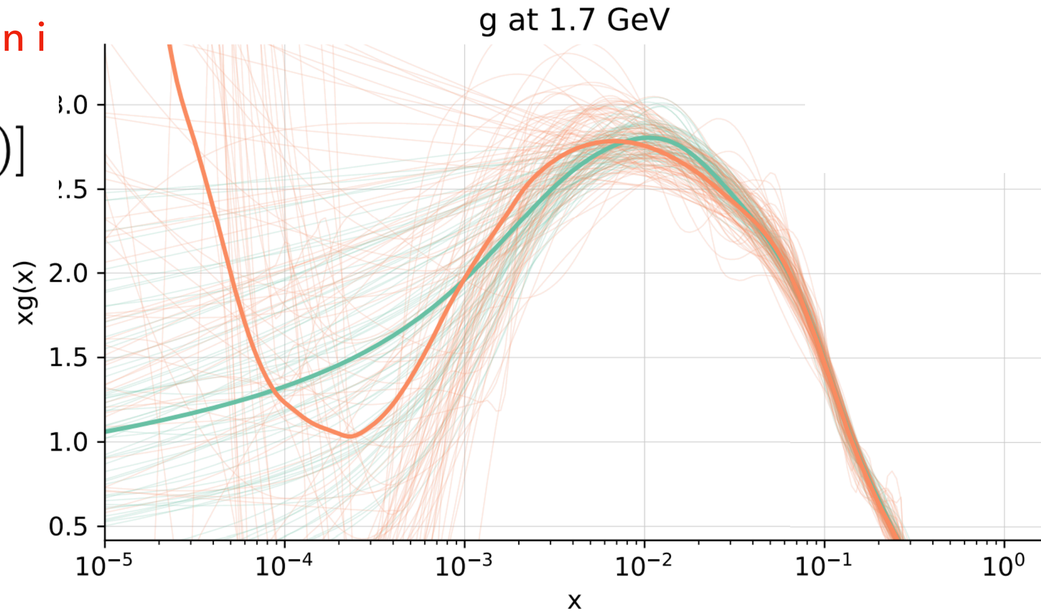
Best fit for each data replica in Monte Carlo approach

$$\left(\bar{\mathcal{G}}\right)_i = \mathbb{E}_\epsilon \left[\mathcal{G}(u_{*,k})_i\right]$$ → Central value of prediction i

$$\left(\Delta_{\text{PDF}}\right)_i = \mathbb{E}_\epsilon \left[\mathcal{G}(u_{*,k})_i - \mathbb{E}_\epsilon \mathcal{G}(u_{*,k})_i\right]$$ → PDF uncertainty of prediction i

$$\left(C_{\text{PDF}}\right)_{ij} = \frac{N_{\text{reps}}}{N_{\text{reps}} - 1} \mathbb{E}_\epsilon \left[\left(\mathcal{G}(u_{*,k})_i - \mathbb{E}_\epsilon \mathcal{G}(u_{*,k})_i\right)\left(\mathcal{G}(u_{*,k})_j - \mathbb{E}_\epsilon \mathcal{G}(u_{*,k})_j\right)\right]$$

PDF covariance matrix of predictions i and j
(PDF induced correlation among observables)



- Key estimate is the normalised **bias**:
  measure mean square deviation of predictions from the "truth" in units of predicted standard deviation

$$B^{(l)}(C_{\text{PDF}}) = \frac{1}{N_{\text{data}}} \sum_{i,j=1}^{N_{\text{data}}} \left(\mathbb{E}_\epsilon \mathcal{G}(u^{(l)}_{*,k})_i - f_i\right) (\overline{C}_{\text{PDF}})^{-1}_{ij} \left(\mathbb{E}_\epsilon \mathcal{G}(u^{(l)}_{*,k})_j - f_j\right)$$

$$R_b = \sqrt{\mathbb{E}_\eta B^{(l)}(\overline{C}_{\text{PDF}})}$$

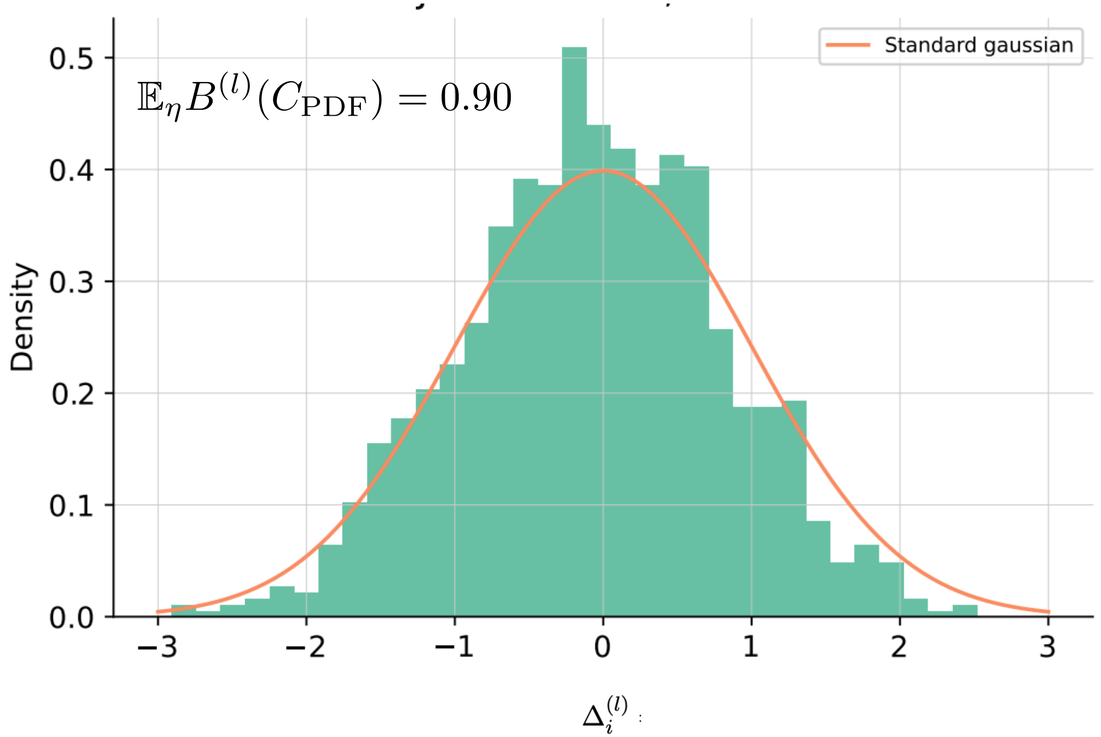
Bias averaged over N_{fit} L1 data, i.e. runs of the universe

$R_b \sim 1$ faithfully estimated PDF uncertainties
$R_b < 1$ overestimated PDF uncertainties
$R_b > 1$ underestimated PDF uncertainties

CONSISTENT CLOSURE TEST

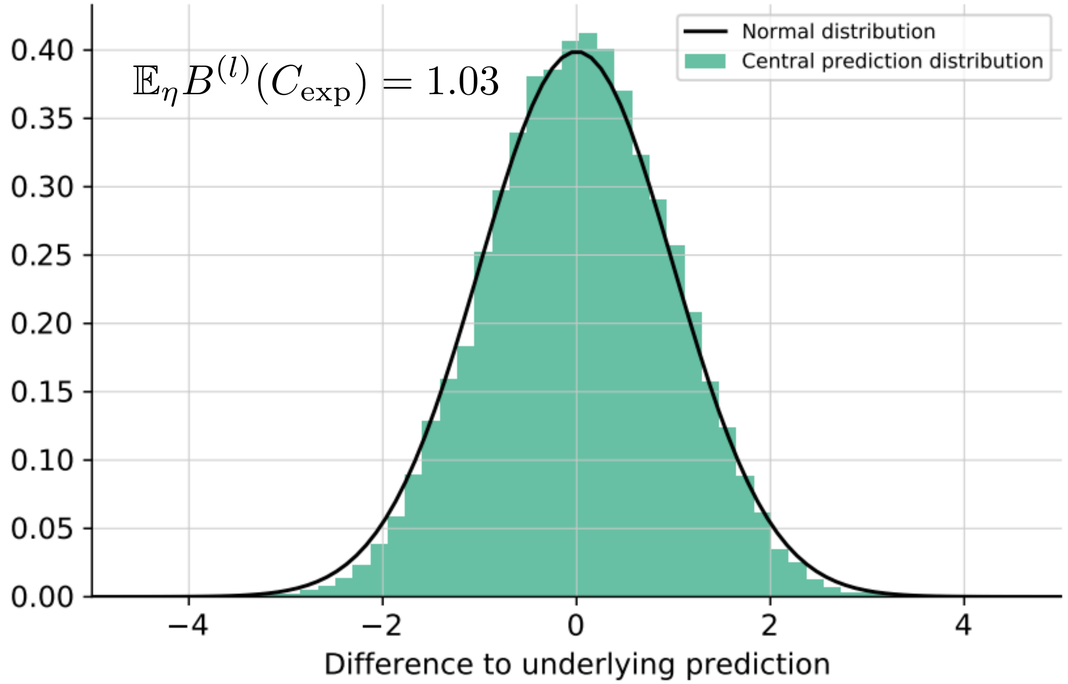
$$\Delta_i^{(l)} = \sum_{j=1}^{N_{\rm data}} \left(\mathbb{E}_\epsilon \mathcal{G}(u_{*,k}^{(l)})_j - f_j \right) v_j^{(i)}$$

NNPDF4.0 global fit (new better estimator)



Barontini et al, 2503.17447

NNPDF4.0 global fit (old estimator)



NNPDF4.0 paper, 2109.02653

$$B^{(l)}(C_{\rm PDF}) = \frac{1}{N_{\rm data}} \sum_{i,j=1}^{N_{\rm data}} \left(\mathbb{E}_\epsilon \mathcal{G}(u_{*,k}^{(l)})_i - f_i \right) (\overline{C}_{\rm PDF})_{ij}^{-1} \left(\mathbb{E}_\epsilon \mathcal{G}(u_{*,k}^{(l)})_j - f_j \right)$$

B ~ 1 faithfully estimated PDF uncertainties
B < 1 overestimated PDF uncertainties
B > 1 underestimated PDF uncertainties

MODELLING EXPERIMENTAL INCONSISTENCIES

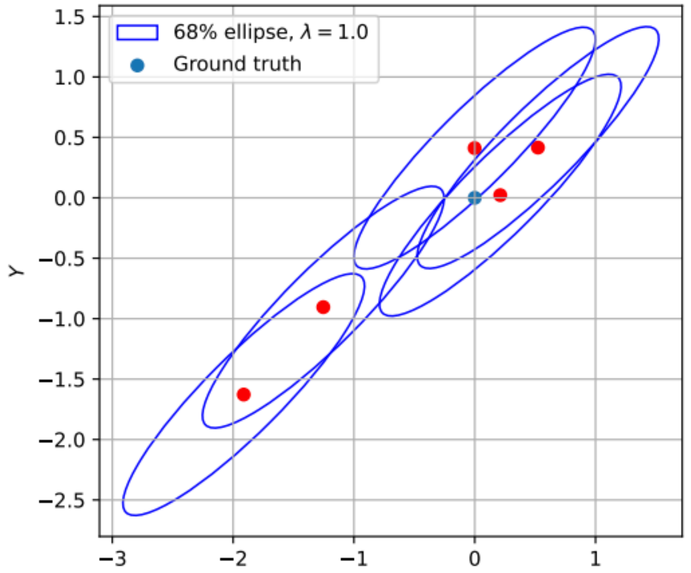
- How to model an inconsistency due to some underestimated experimental systematics?

- Generate L1 data with “true” experimental covariance matrix

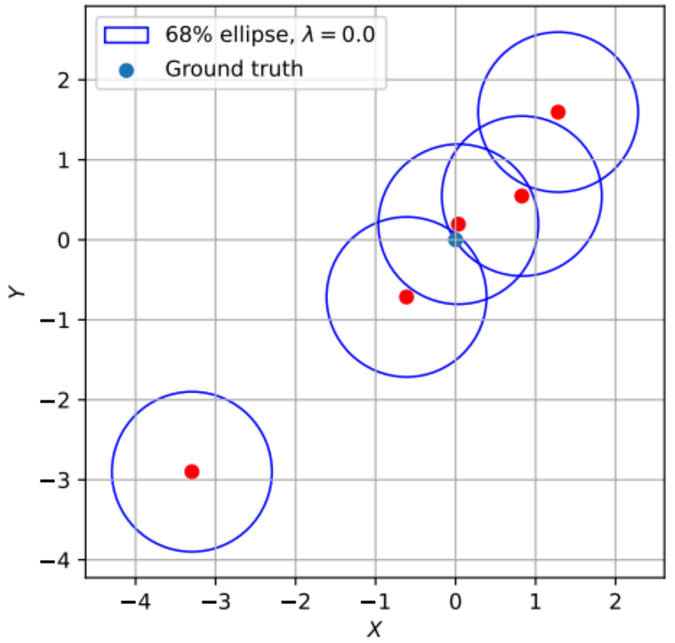
$$(C_{\text{exp}})_{ij} = \delta_{ij}\sigma_i^{(\text{uncorr})}\sigma_j^{(\text{uncorr})} + \sum_{k=1}^{N_{\text{corr}}} \sigma_{i,k}^{(\text{corr})}\sigma_{j,k}^{(\text{corr})}$$

- Fit the data using the rescaled experimental covariance matrix (both in pseudo data generation and in the loss function)

$$(C_{\text{exp}}^{\lambda})_{ij} = \delta_{ij}\sigma_i^{(\text{uncorr})}\sigma_j^{(\text{uncorr})} + \sum_{k=1}^{N_{\text{corr}}} \lambda_{i,k}\sigma_{i,k}^{(\text{corr})}\lambda_{j,k}\sigma_{j,k}^{(\text{corr})}$$



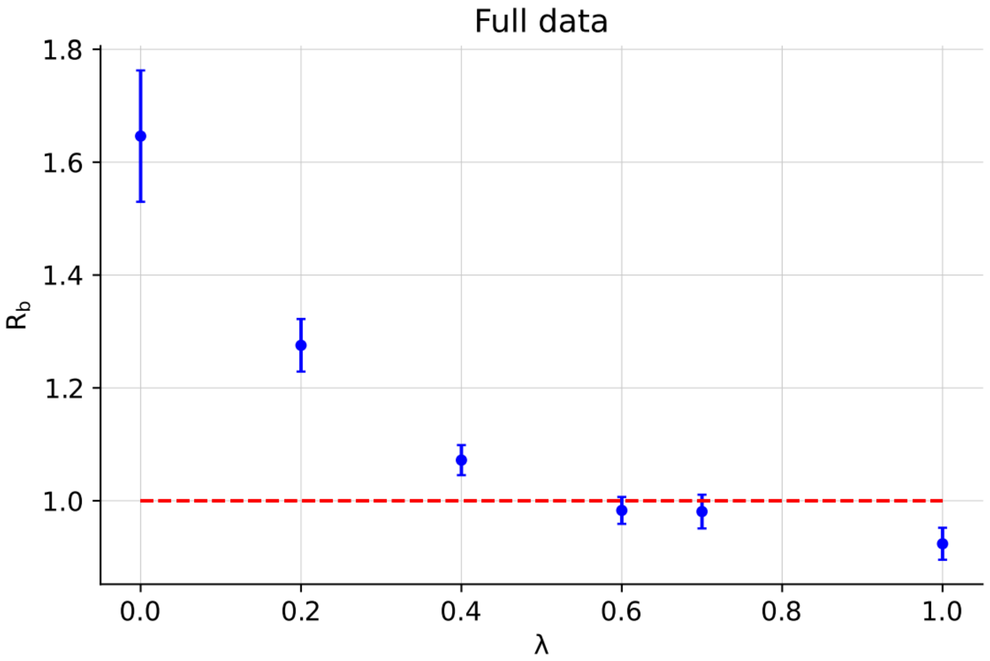
Consistent
$\lambda = 1$

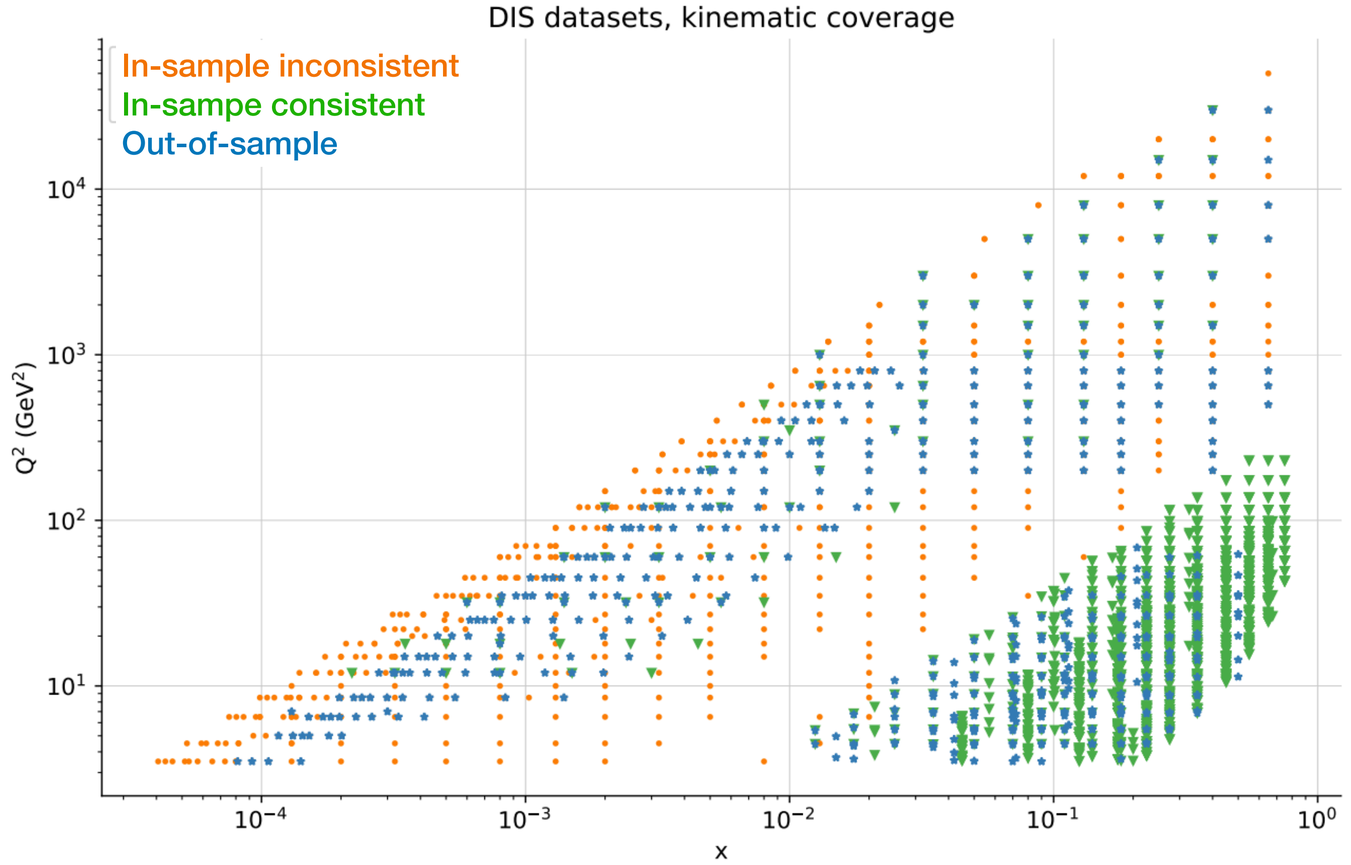


Extreme inconsistency
$\lambda = 0$

BULK INCONSISTENCY

- DIS only fit, in-sample HERA NC data are inconsistent.
- **860 out of 2576** inconsistent datapoints, with all systematic uncertainties underestimated by the same factor λ.

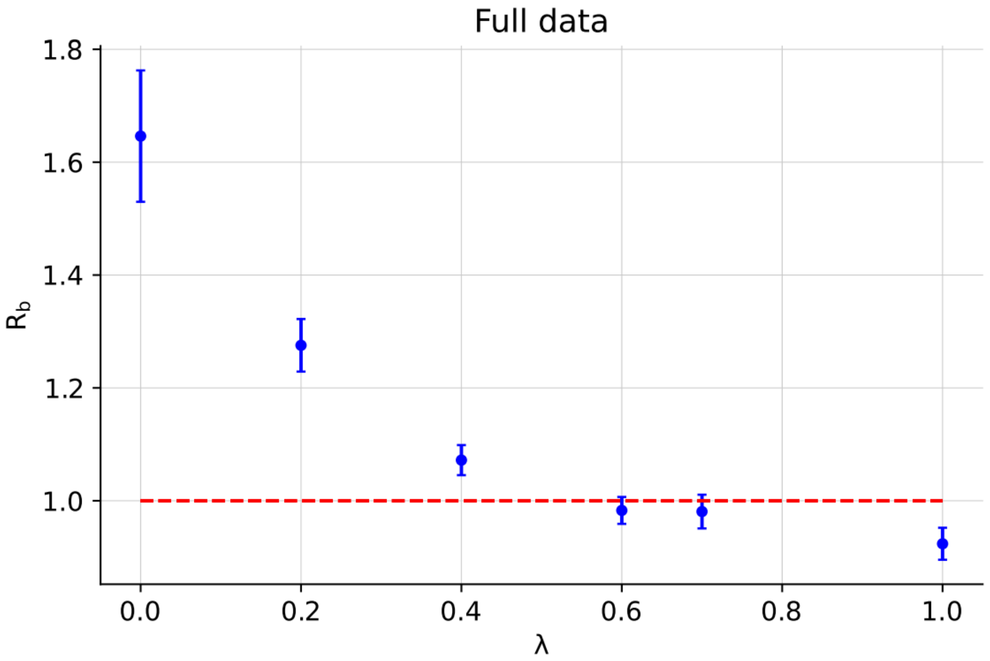


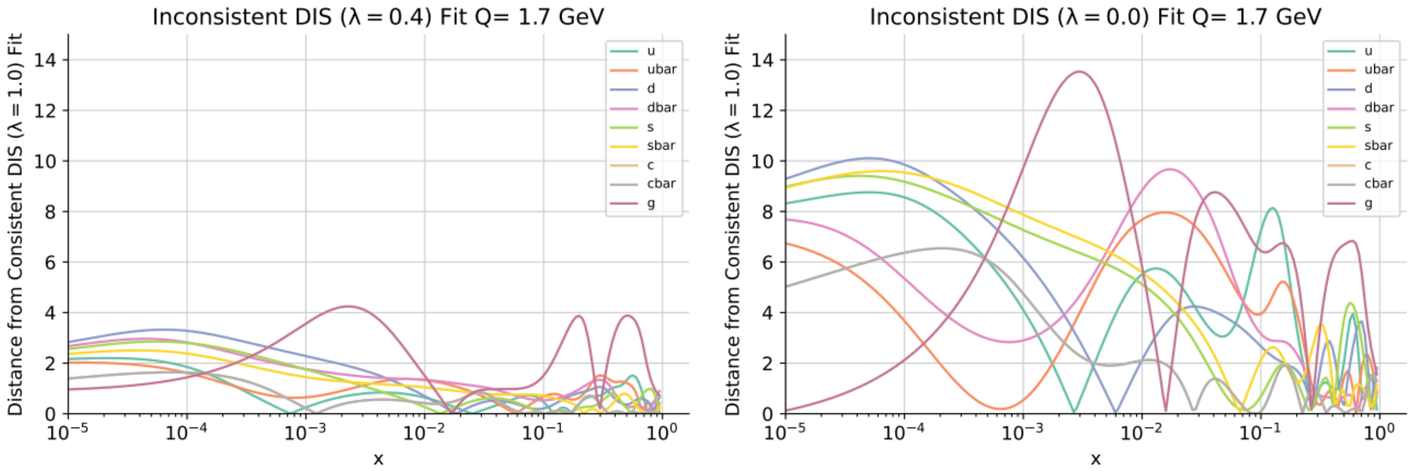


- Highly non linear behaviour: for $\lambda \geq 0.4$ despite inconsistency, PDF uncertainties remain faithful, but then sharply rises.
- In-sample and out-of-sample datasets behave in a similar way => NN model effective at generalising.

BULK INCONSISTENCY

- DIS only fit, in-sample HERA NC data are inconsistent.
- **860 out of 2576** inconsistent datapoints, with all systematic uncertainties underestimated by the same factor λ.





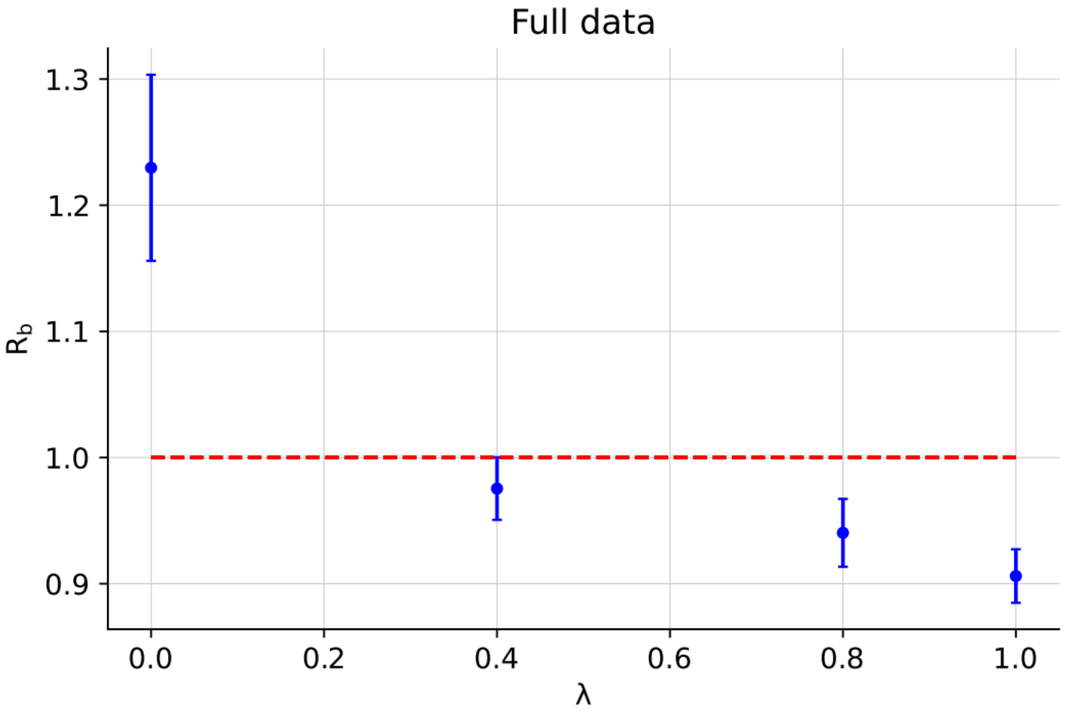
$\lambda \geq 0.4$: Model corrects for underestimated uncertainty in the inconsistent data, PDF uncertainties do not decrease despite the reduced data uncertainty.

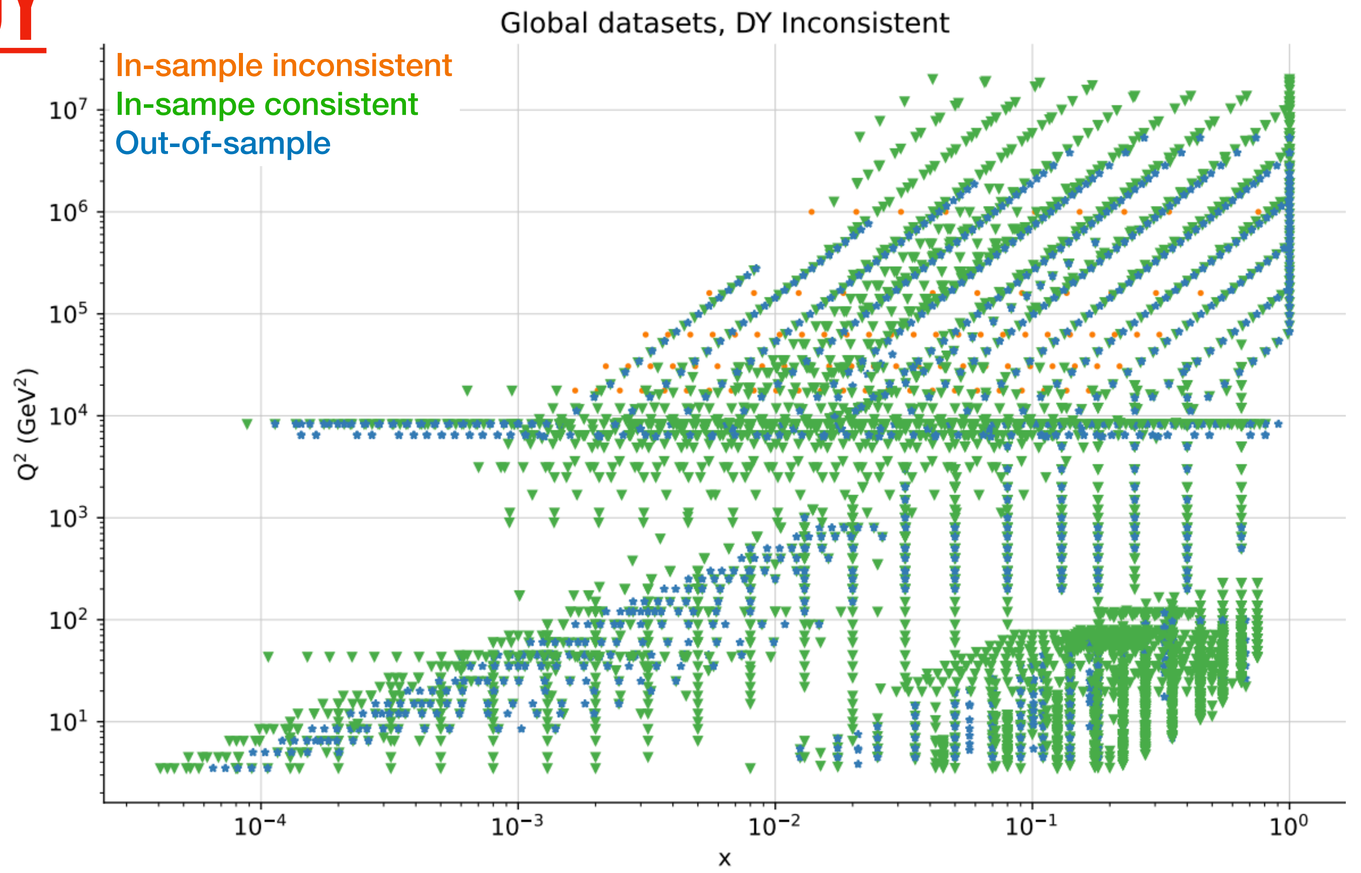
$\lambda < 0.4$: PDF uncertainty shrinks => underestimated PDF uncertainties, largest shifts in the gluon and quark singlet combinations.

- Highly non linear behaviour: for $\lambda \geq 0.4$ despite the inconsistency, PDF uncertainties remain faithful, but then sharply rises.
- In-sample and out-of-sample datasets behave in a similar way => NN model effective at generalising.

SINGLE DATASET INCONSISTENCY: DY

- Global fit, in-sample with one of the in-sample datasets inconsistent: ATLAS double differential high-mass DY cross section.
- **48 (607) out of 3772** inconsistent datapoints (considering correlations)

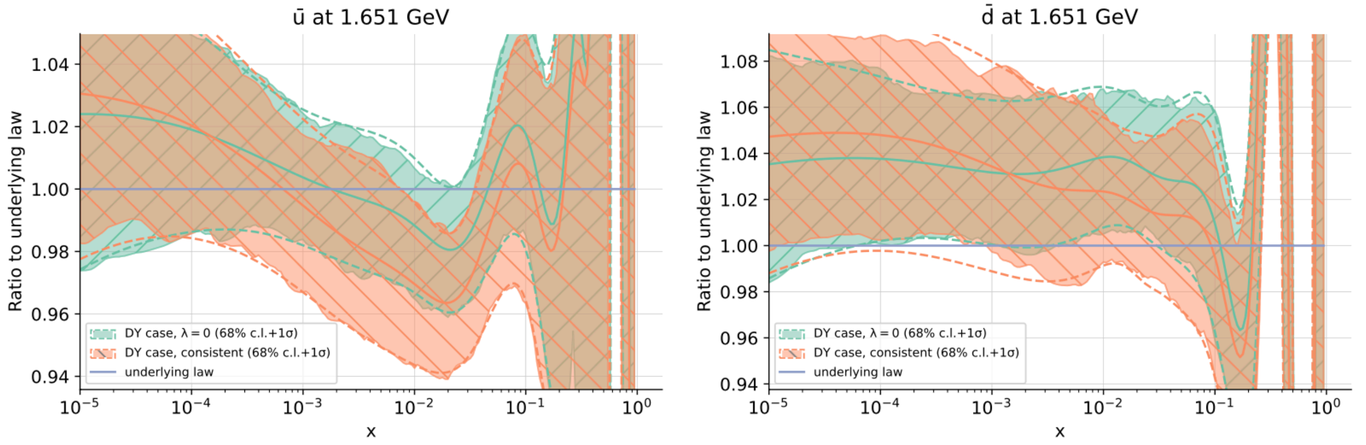




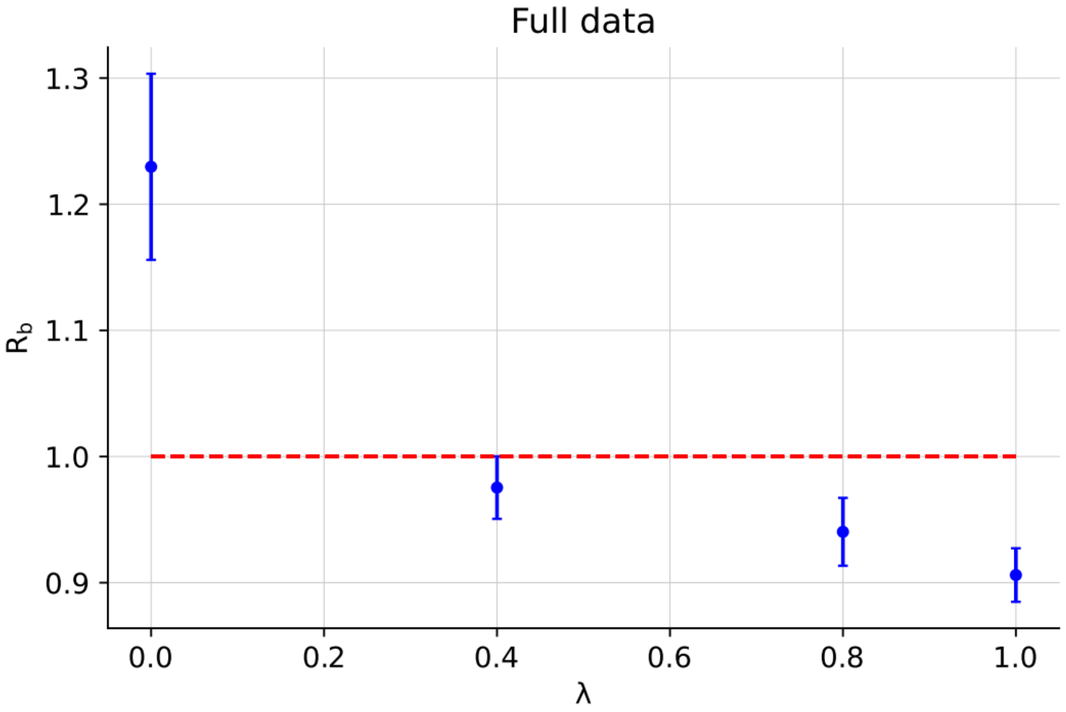
- Down to λ ≈ 0.2 model corrects for inconsistency
- λ = 0: leap, model inaccurate and uncertainties somewhat inaccurate.
- λ = 0: PDF uncertainties unchanged, but especially antiup and antidown PDFs undergo significant shift

SINGLE DATASET INCONSISTENCY: DY

- Global fit, in-sample with one of the in-sample datasets inconsistent: ATLAS double differential high-mass DY cross section.
- **48 (607) out of 3772** inconsistent datapoints (considering correlations)



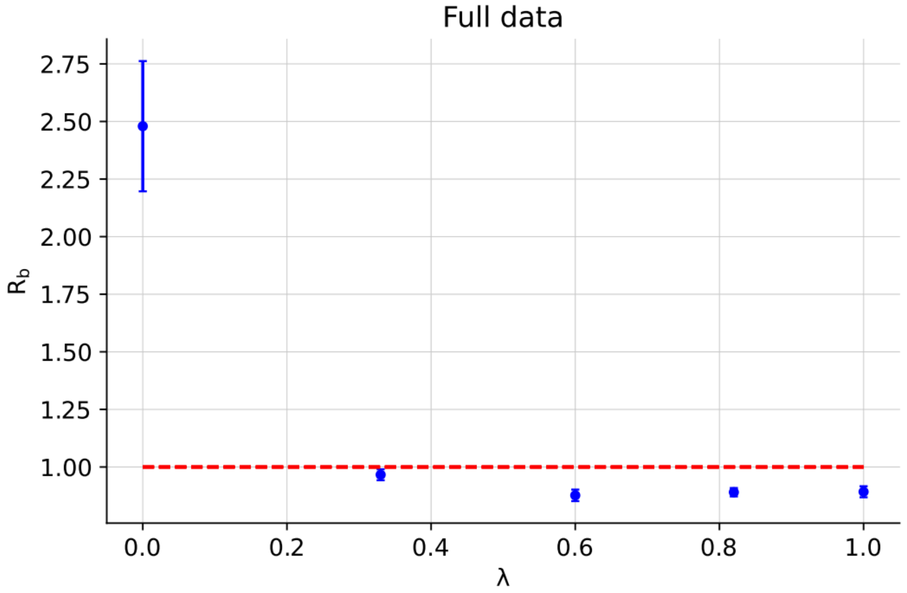
λ = 0: PDF uncertainties unchanged, but especially antiup and antidown PDFs shift.

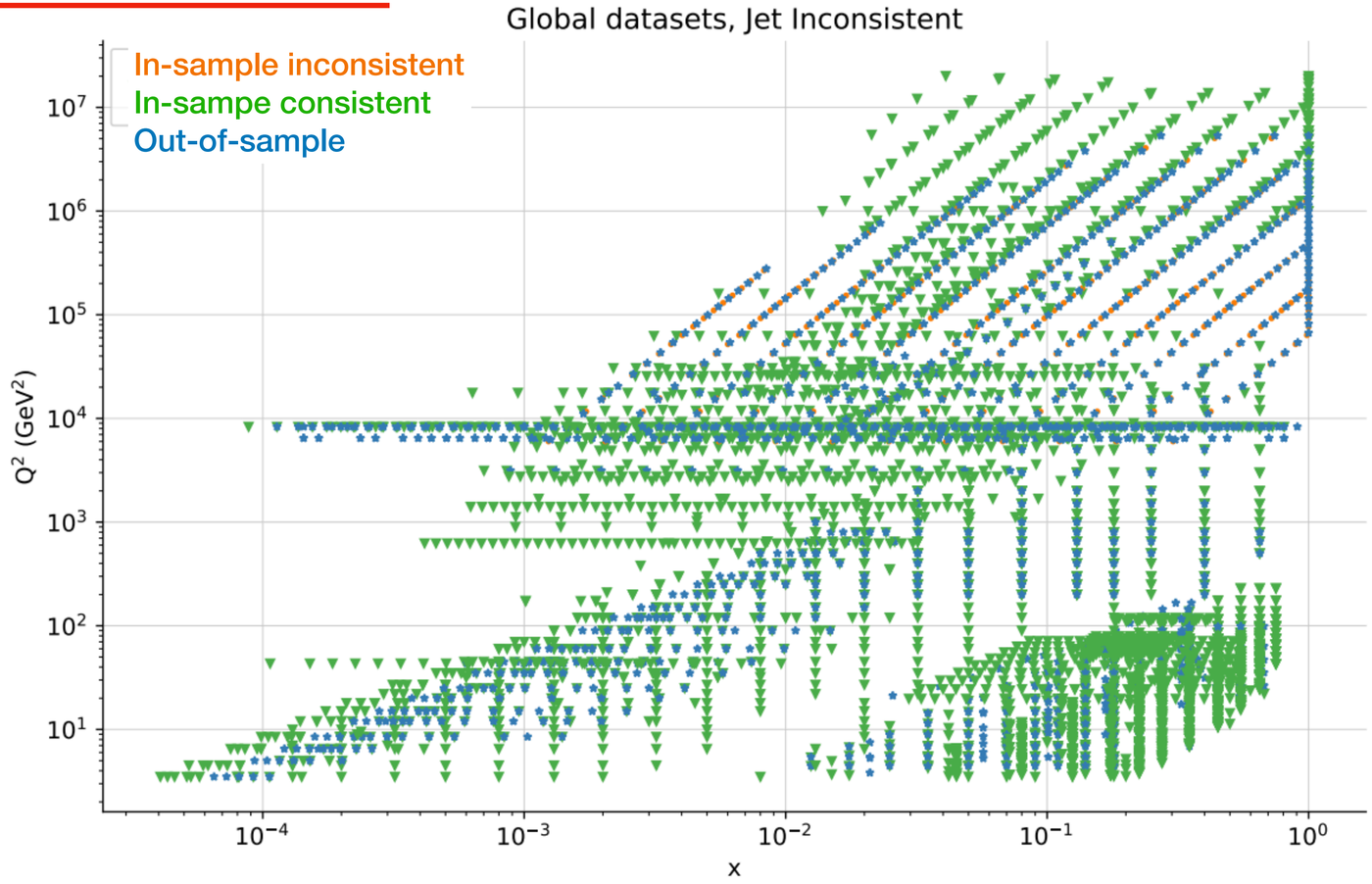


- Down to λ ≈ 0.2 model corrects for inconsistency
- λ = 0: leap, model inaccurate and uncertainties somewhat inaccurate.

HIGH-IMPACT DATASET INCONSISTENCY: JETS

- Global fit, in-sample with one of the in-sample datasets inconsistent: ATLAS single inclusive jet data. CMS inclusive jet data out of sample.
- **171 (607) out of 3793** inconsistent datapoints (considering correlations)

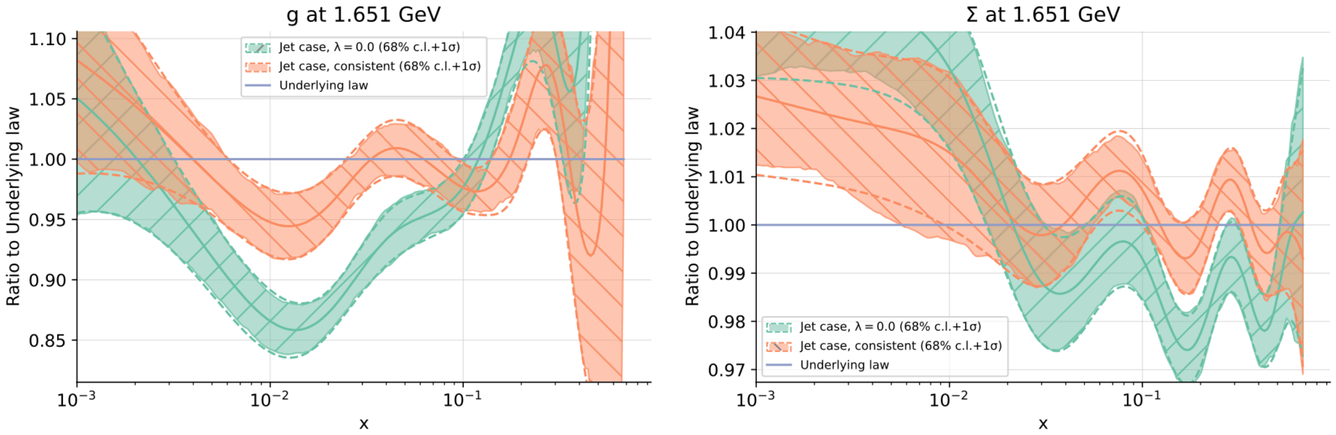




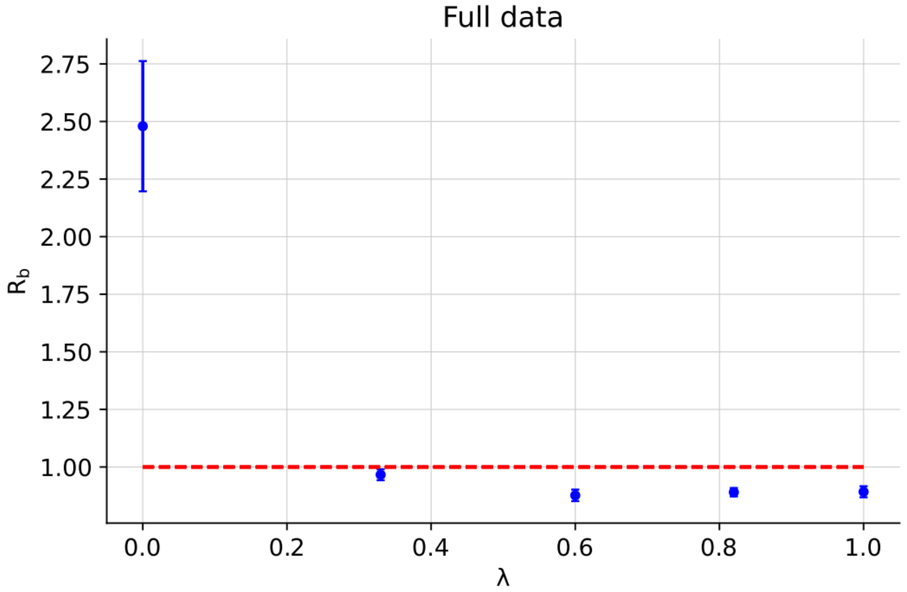
- Down to $\lambda \approx 0.3$ model corrects for inconsistency
- $\lambda = 0$: phase transition, model fails and uncertainties are completely wrong.

HIGH-IMPACT DATASET INCONSISTENCY: JETS

- Global fit, in-sample with one of the in-sample datasets inconsistent: ATLAS single inclusive jet data. CMS inclusive jet data out of sample.
- **171 (607) out of 3793** inconsistent datapoints (considering correlations)

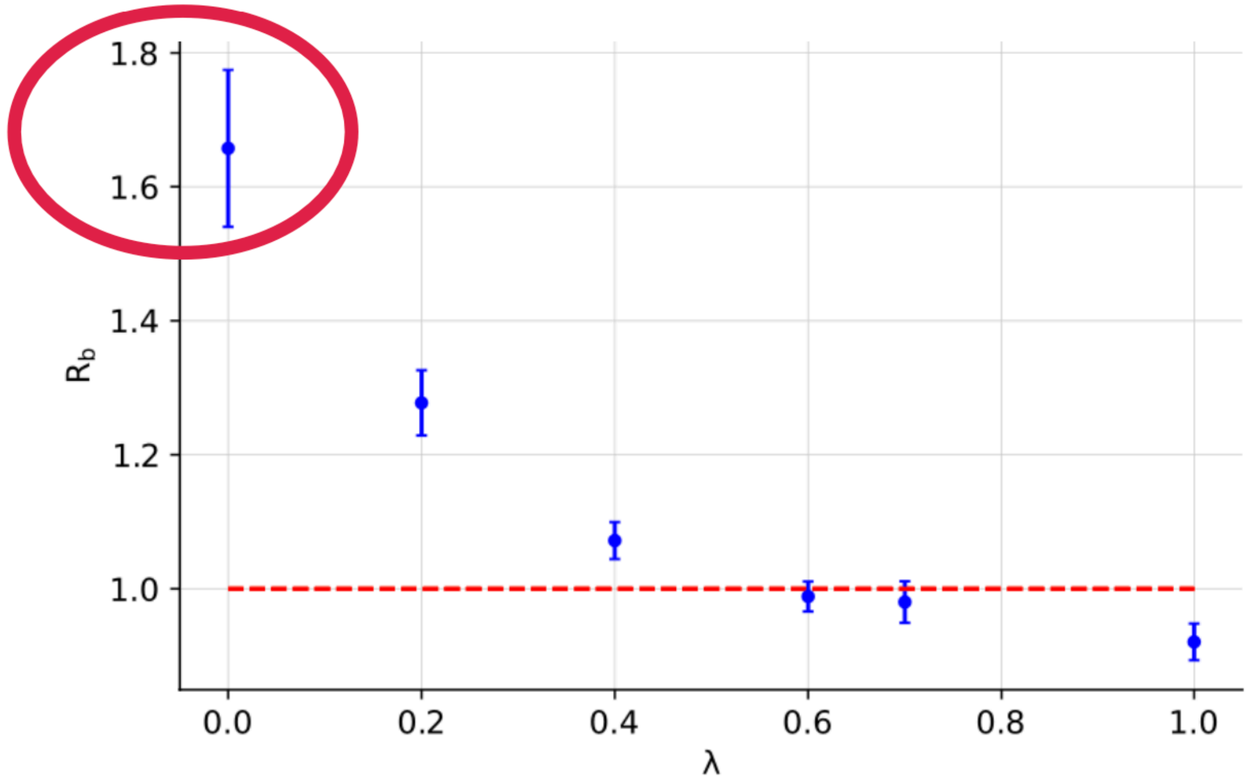


λ = 0: gluon shows large deviation from "true" value, but no significant increase in its uncertainty compared to consistent case. Milder effect for singlet, coupled to gluon in DGLAP evolution.



- Down to λ ≈ 0.3 model corrects for inconsistency
- λ = 0: phase transition, model fails, uncertainties are completely wrong.

A CRITERION TO SPOT EXPERIMENTAL INCONSISTENCIES



- So far learn that NN model cures experimental inconsistencies until they are not too big
- If strong inconsistencies model fails, as either PDF uncertainties shrink or PDFs shift far from underlying law.
- In real life no access to the “truth”, normalised bias cannot be computed and only a single run of the Universe.
- How to spot inconsistencies in a actual PDF fit?

A CRITERION TO SPOT EXPERIMENTAL INCONSISTENCIES

Large R_b → Large $\chi 2$ of the inconsistent dataset and of consistent datasets correlated to it by the PDFs

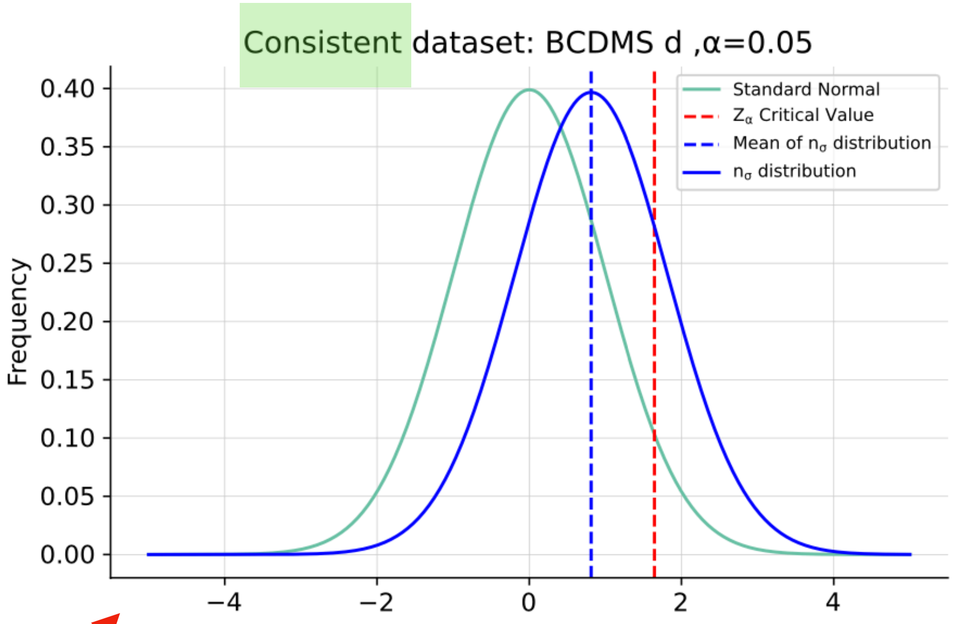
$$n_\sigma^{(i)} = \frac{\chi_i^2 - N_{\text{data}}^{(i)}}{\sqrt{2/N_{\text{data}}^{(i)}}}$$

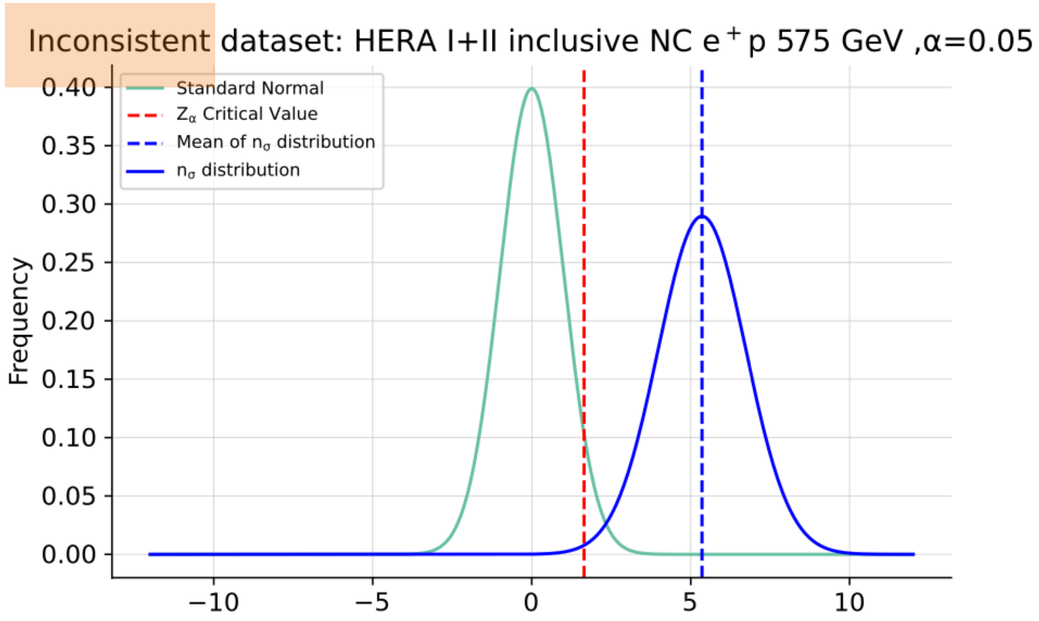
Normalised deviation of $\chi 2$ from its mean

$$\mu_i = \left\langle n_\sigma^{(i)} \right\rangle_{N_{\text{fit}}}$$

$$\mu_i > Z$$

Z: Critical threshold





Distribution of n_σ values across N_{fits} in DIS case with maximal inconsistency ($\lambda = 0$)

A CRITERION TO SPOT EXPERIMENTAL INCONSISTENCIES

Large R_b → Large $\chi 2$ of the inconsistent dataset and of consistent datasets correlated to it by the PDFs

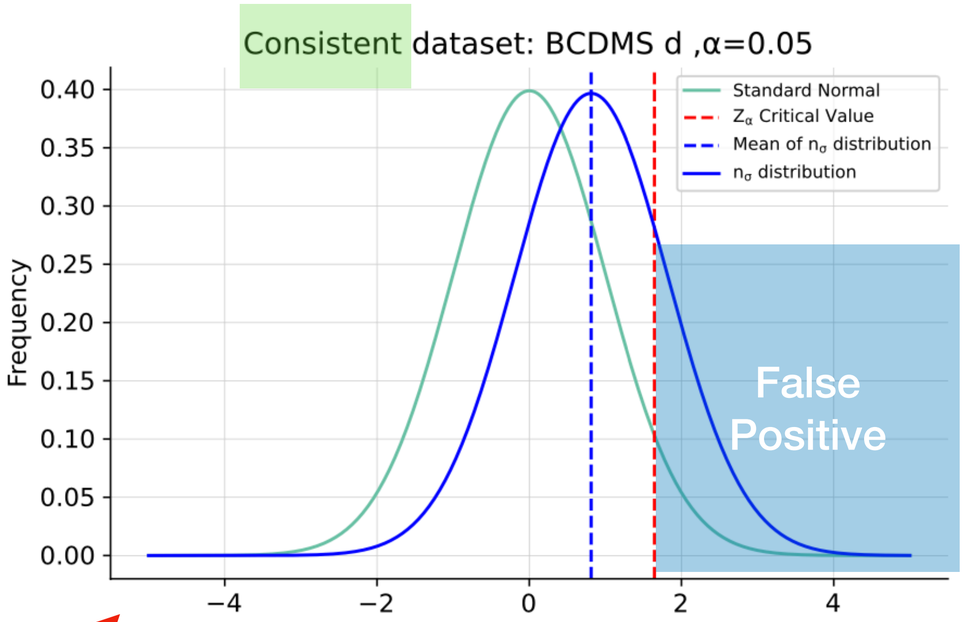
$$n_\sigma^{(i)} = \frac{\chi_i^2 - N_{\text{data}}^{(i)}}{\sqrt{2/N_{\text{data}}^{(i)}}}$$

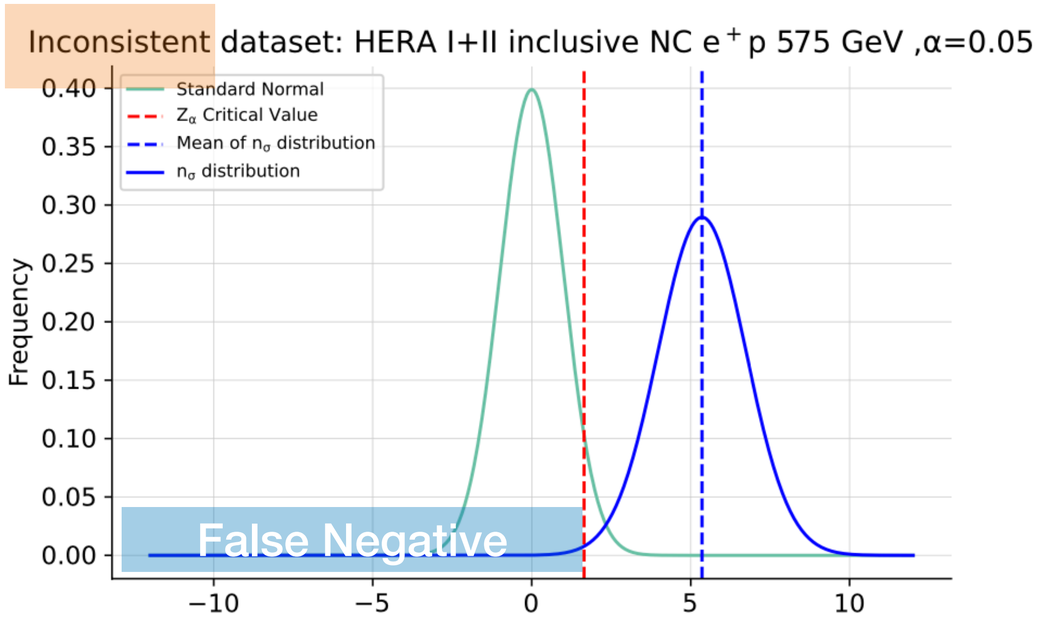
Normalised deviation of $\chi 2$ from its mean

$$\mu_i = \left\langle n_\sigma^{(i)} \right\rangle_{N_{\text{fit}}}$$

$$\mu_i > Z$$

Z: Critical threshold





Distribution of n_σ values across N_{fits} in DIS case with maximal inconsistency ($\lambda = 0$)

A CRITERION TO SPOT EXPERIMENTAL INCONSISTENCIES

A single criterion S_1 given by n_σ threshold Z not enough to minimise false positive and false negative

$$n_\sigma^{(i)} = \frac{\chi_i^2 - N_{\rm data}^{(i)}}{\sqrt{2/N_{\rm data}^{(i)}}}$$

$$\chi^{2,(i)}_{\rm weighted} = \frac{1}{N_{\rm data} - N_{\rm data}^{(j)}} \sum_{j=1, j\neq i}^{N_{\rm exp}} N_{\rm data}^{(j)} \chi_j^2 + w^{(i)} \chi_i^2$$

$$w^{(i)} = N_{\rm data}/N_{\rm data}^{(i)}$$

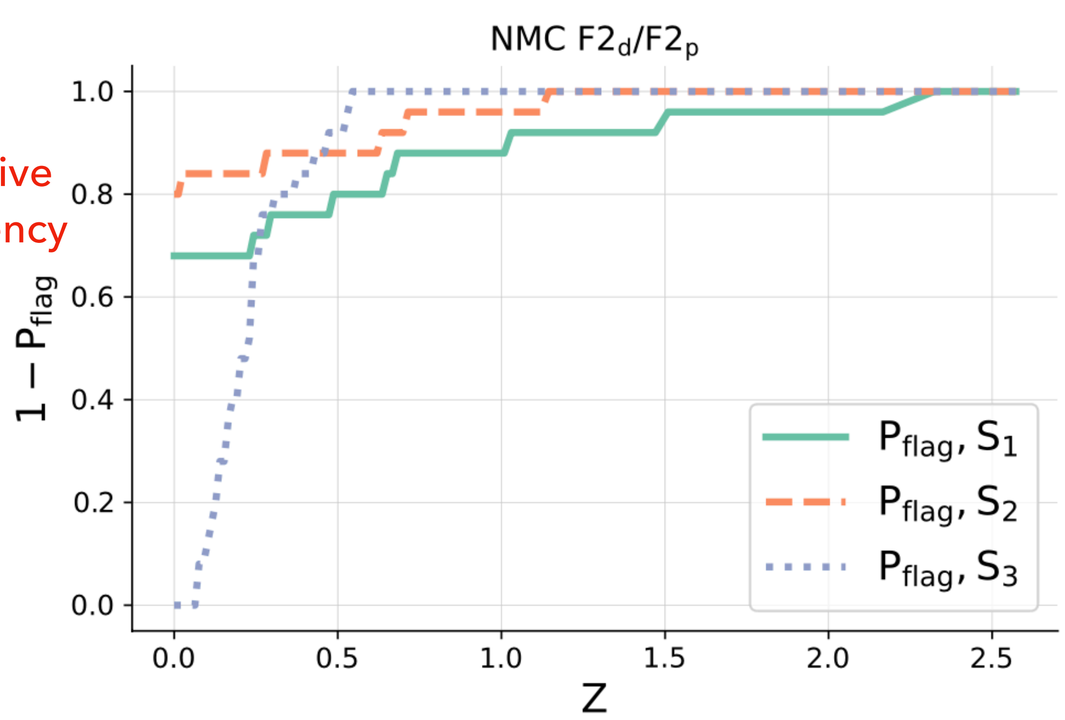
Extra criterion: what happens if a large weight is given to inconsistent/consistent dataset?
Expect that the inconsistent dataset, if given extra weight in the fit, will either fail to improve or will improve but spoil $\chi 2$ of other consistent datasets, while the consistent dataset will not.

S_1 $\mu_i > Z$ **S_2** $n_\sigma^{\rm weighted,(i)} > Z$ **S_3** $n_\sigma^{\rm weighted,(j)} - n_\sigma^{(j)} > Z \qquad \forall j \neq i$

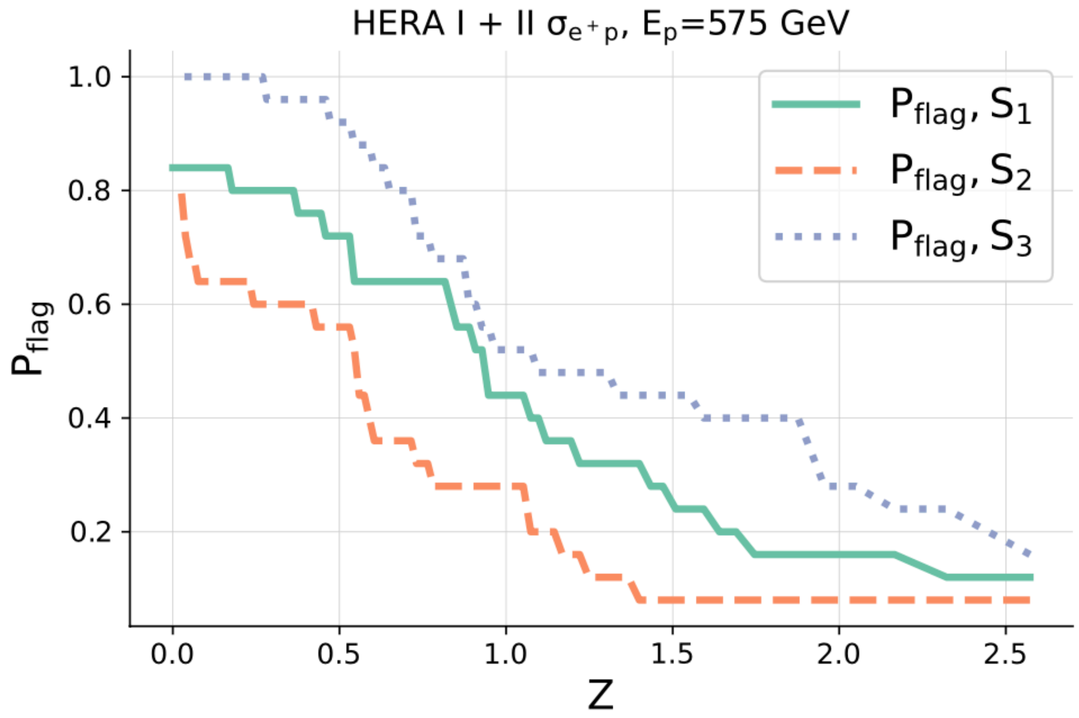
A CRITERION TO SPOT EXPERIMENTAL INCONSISTENCIES

Test efficiency of criteria in the DIS case for extreme inconsistency ($\lambda = 0$) on consistent and inconsistent datasets

True negative efficiency



True positive efficiency



S_1 $\mu_i > Z$

S_2 $n_\sigma^{\text{weighted},(i)} > Z$

S_3 $n_\sigma^{\text{weighted},(j)} - n_\sigma^{(j)} > Z \qquad \forall j \neq i$

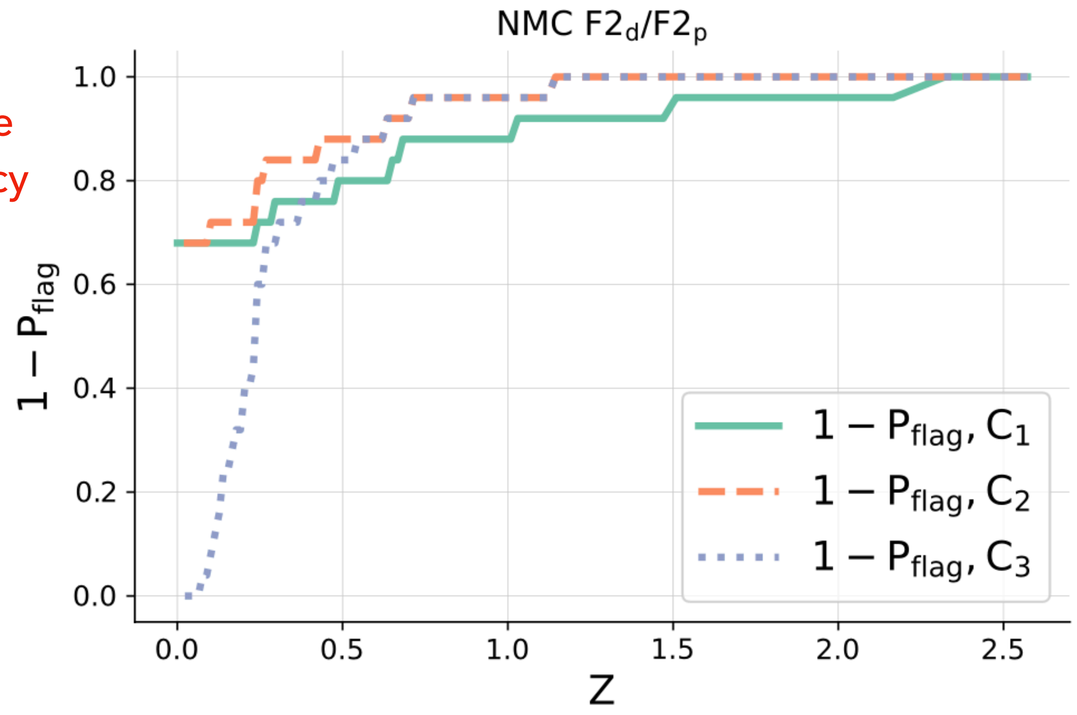
OPTIMAL INCONSISTENCY DETECTION

C_1: condition S_1 satisfied

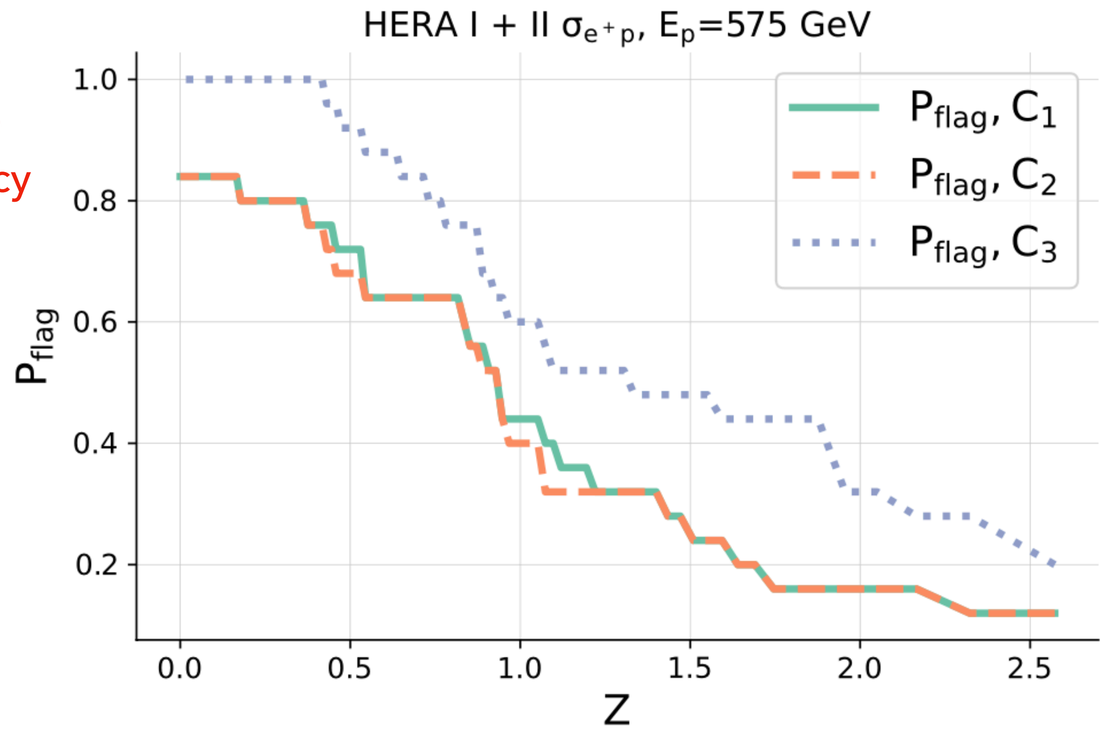
C_2: condition S_1 satisfied, and in a weighted fit either S_2 or S_3 are satisfied (NNPDF4.0 criterion)

C_3: in weighted fit either S_2 or S_3 are satisfied

True negative efficiency



True positive efficiency



S_1 $\mu_i > Z$

S_2 $n_\sigma^{\text{weighted},(i)} > Z$

S_3 $n_\sigma^{\text{weighted},(j)} - n_\sigma^{(j)} > Z \qquad \forall j \neq i$

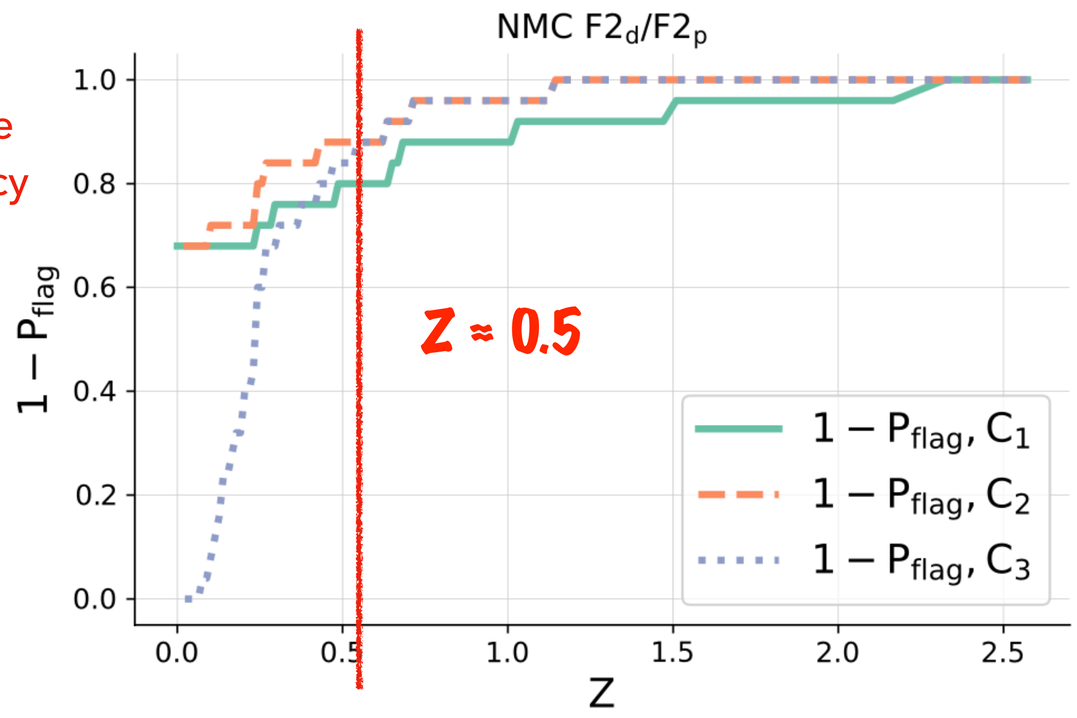
OPTIMAL INCONSISTENCY DETECTION

C_1: condition S_1 satisfied

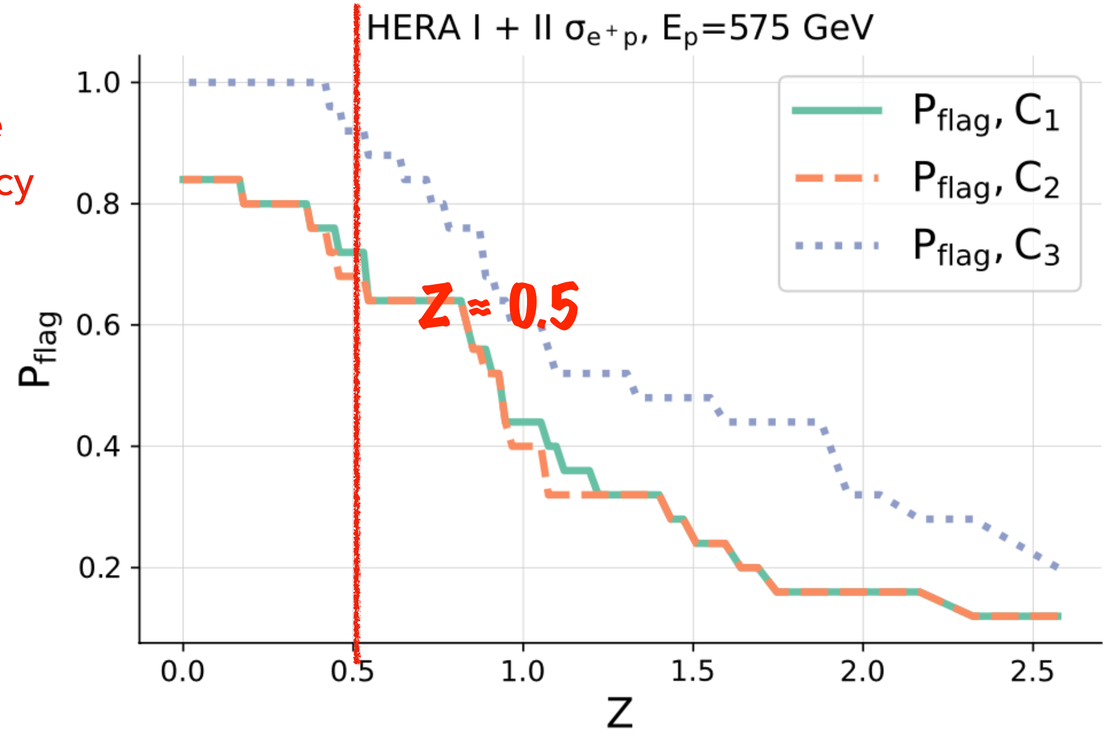
C_2: condition S_1 satisfied, and in a weighted fit either S_2 or S_3 are satisfied (NNPDF4.0 criterion)

C_3: in weighted fit either S_2 or S_3 are satisfied

True negative efficiency



True positive efficiency



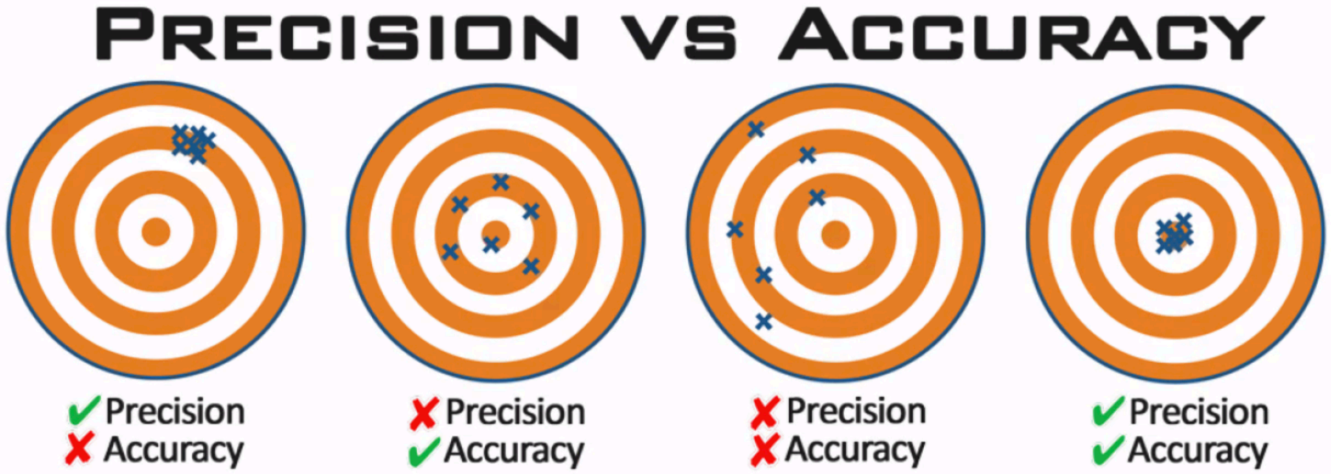
In this case for Z ≈ 0.5, 90% probability of NOT flagging a consistent dataset as inconsistent and 95% probability of flagging an inconsistent dataset as inconsistent.

CONCLUSIONS

- Tensions and incompatibilities between experimental measurements may significantly affect the reliability of PDFs and their uncertainty
- By analysing a built-in inconsistency injected in the data in a controlled way could measure effect of such inconsistency in a PDF fit in conditions close to real-life scenarios
- NNPDF methodology manages to correct for moderate to medium inconsistencies & generalises well to out-of-sample datasets
- For strong inconsistency, effect visible and distorts PDFs in regions in which experimental measurements are systematics-dominated
- Developed a new procedure to detect cases in which uncorrected inconsistencies are present
- Criterion will be applied to NNPDF4.1 for dataset selection
- Claim: in NNPDF approach explicit tolerance not needed if dataset selection criterion is effective

EXTRA MATERIAL

THE PRECISION VERSUS ACCURACY CHALLENGE



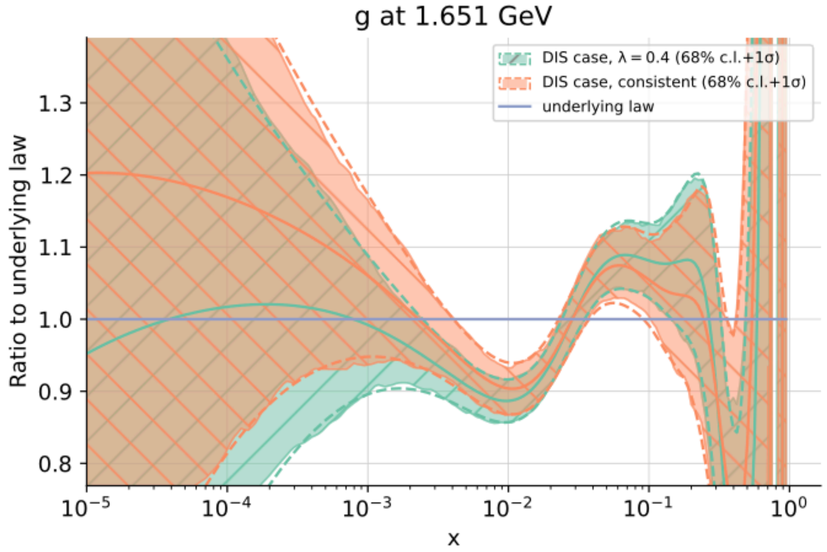
PDF4LHC21, 2203.05506

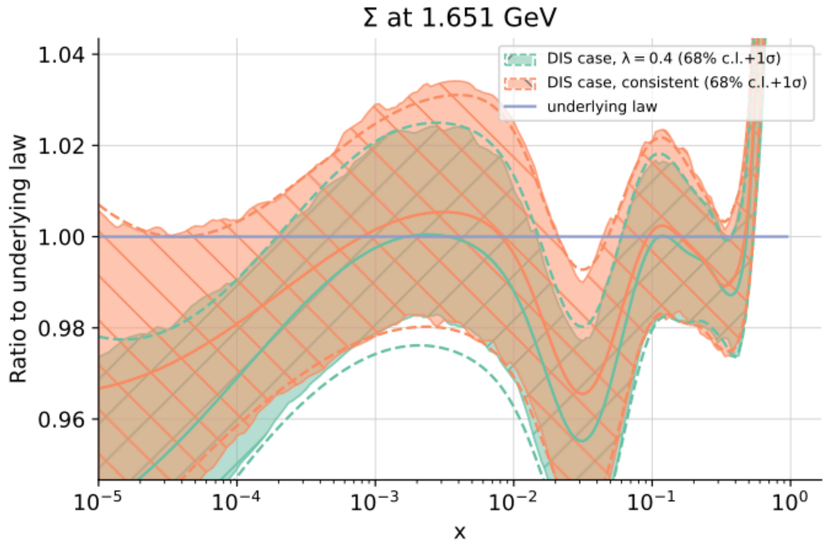
Dataset	$N_{\rm pt}$	$\chi^2/N_{\rm pt}$		
		CT18	MSHT20	NNPDF3.1
BCDMS F_2^p	$329/163^{\dagger\dagger}/325^{\dagger}$	1.06	1.00	1.21
BCDMS F_2^d	$246/151^{\dagger\dagger}/244^{\dagger}$	1.06	0.88	1.10
NMC F_2^d/F_2^p	$118/117^{\dagger}$	0.93	0.93	0.90
NuTeV dimuon $\nu + \bar{\nu}$	38+33	0.79	0.83	1.22
HERAI+II	1120	1.23	1.20	1.22
E866 $\sigma_{pd}/(2\sigma_{pp})$	15	1.24	0.80	0.43
LHCb 7 TeV & 8TeV W,Z	29+30	1.15	1.17	1.44
LHCb 8 TeV $Z \to ee$	17	1.35	1.43	1.57
ATLAS 7 TeV W,Z (2016)	34	1.96	1.79	2.33
D0 Z rapidity	28	0.56	0.58	0.62
CMS 7 TeV electron $A_{\rm ch}$	11	1.47	1.52	0.76
ATLAS 7 TeV W,Z(2011)	30	1.03	0.93	1.01
CMS 8TeV incl. jet	$185/174^{\dagger\dagger}$	1.03	1.39	1.30
Total $N_{\rm pt}$	—	2263	1991	2256
Total $\chi^2/N_{\rm pt}$	—	1.14	1.15	1.20

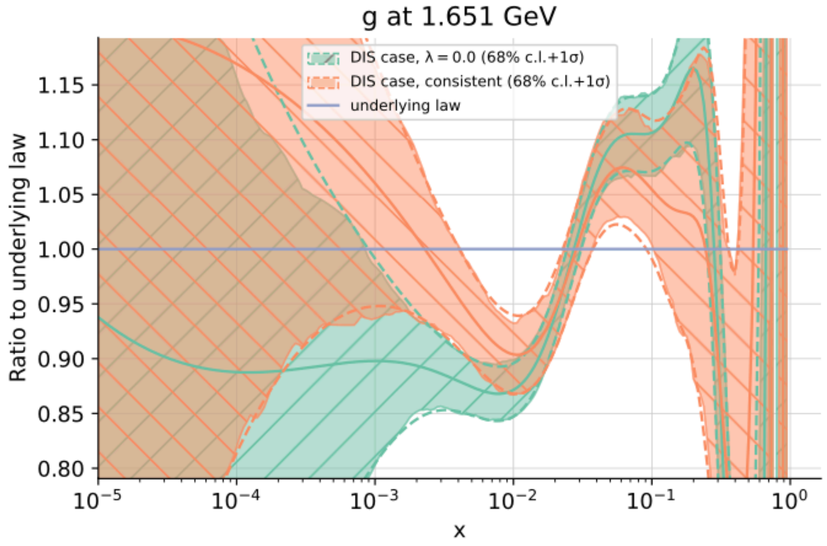
Challenges

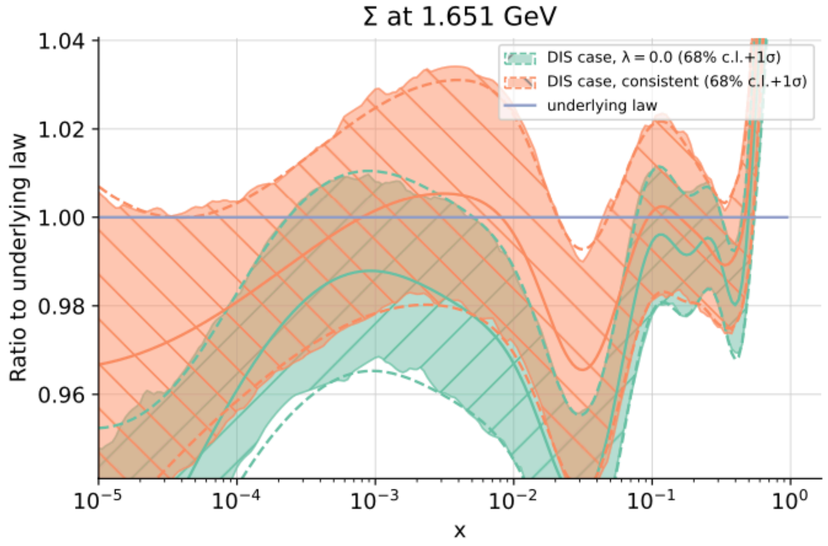
- Inconsistency or tension in data of experimental origin (underestimate of systematics…)
- Deficiencies in fitting methodology (data-driven parametrisation change, optimisation issues, overfitting…)
- Inaccuracy in theoretical framework
  - Missing higher order uncertainties (QCD, EW)
  - Other corrections (nuclear, higher-twist, non-perturbative effects…)
- Fitting away possible BSM signals

DIS: BULK INCONSISTENT









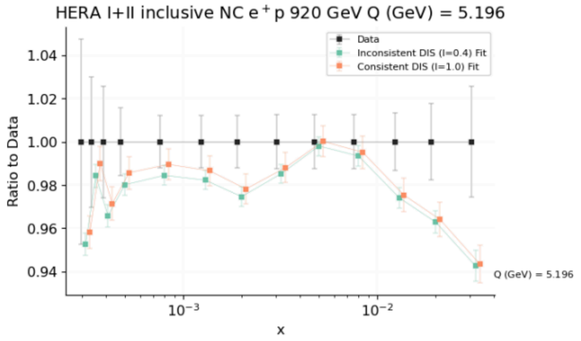
EFFECT ON OBSERVABLES

- $\lambda \approx 0.4/0.3$ model corrects for inconsistency, predictions do not move
- $\lambda = 0$: model fails
- DIS case (bulk inconsistency) and JETS case (high-impact inconsistency): predictions off with unchanged uncertainties.
- DY case (single dataset inconsistency): uncertainties shrink driven by smaller high mass ATLAS dataset

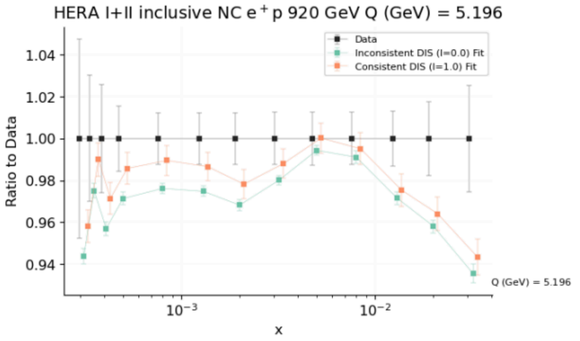
S. Forte's slides
PDF4LHC Dec 2024

DIS: BULK INCONSISTENCY

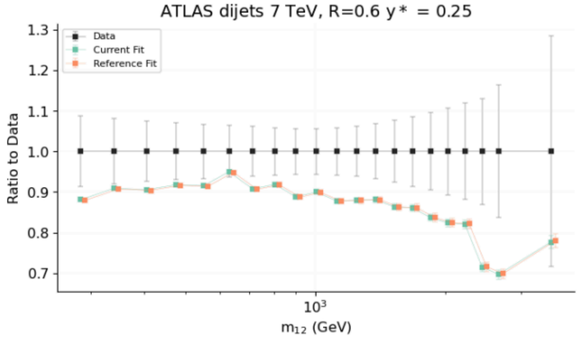
$\lambda = 0.4$: MODEL CORRECTS

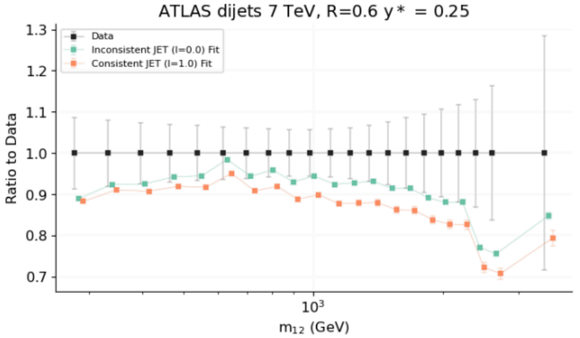


$\lambda = 0$: MODEL FAILS



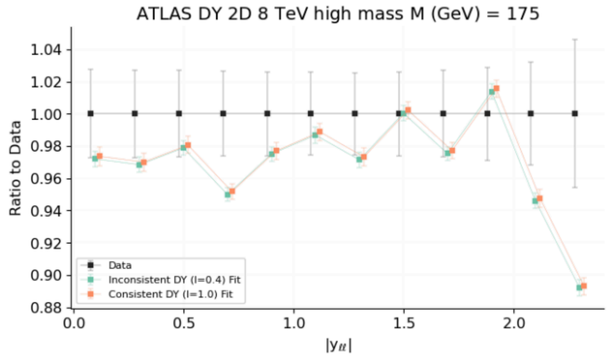
JETS: HIGH-IMPACT INCONSISTENCY





DY: SINGLE DATASET INCONSISTENCY

$\lambda = 0.4$: MODEL CORRECTS



$\lambda = 0$: MODEL FAILS