NNPDF

CLOSURE TESTING

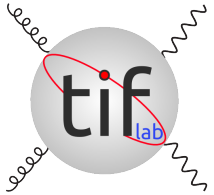
THE IMPACT OF DATA INCONSISTENCY

STEFANO FORTE

UNIVERSITÀ DI MILANO & INFN

UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI FISICA



PDF4LHC MEETING

CERN, DECEMBER 3, 2024

CLOSURE TESTS

- ASSUME TRUTH UNDERLYING PDF $\Rightarrow$ E.G. SOME RANDOM PDF REPLICA
- GENERATE DATA DISTRIBUTED ACCORDING TO EXPERIMENTAL COVARIANCE MATRIX
- RUN WHOLE UNMODIFIED METHODOLOGY ON THESE DATA
- DO STATISTICS ON "RUNS OF THE UNIVERSE"

STATISTICAL INDICATORS

PDFS AND THEIR UNCERTAINTY (REMINDER)

i-TH DATA PREDICTION (REPLICA AVERAGE): $f_i^{\rm pred} = \frac{1}{N_{\rm rep}} \sum_n f_i^{(\rm n)}$

PDF UNCERTAINTY (REPLICA COVARIANCE): $C_{ij}^{\rm PDF} = \frac{1}{N_{\rm rep}-1} \sum_n \left(f_i^{(\rm n)} - f_i^{\rm pred}\right)\left(f_j^{(\rm n)} - f_j^{\rm pred}\right)$

BIAS-VARIANCE RATIO

BIAS: $b^2 = \frac{1}{N_{\rm dat}} \sum_{ij} \left(f_i^{\rm pred} - f_i^{\rm true}\right) C_{ij}^{-1} \left(f_j^{\rm pred} - f_j^{\rm true}\right)$; C_{ij} exp. covmat

VARIANCE: $\sigma^2 = \frac{1}{N_{\rm dat}} \frac{1}{N_{\rm rep}-1} \sum_{ij} \left(f_i^{(\rm n)} - f_i^{\rm pred}\right) C_{ij}^{-1} \left(f_j^{(\rm n)} - f_j^{\rm pred}\right)$

$R_{bv} = b/v$ DEVIATION FROM TRUTHS IN UNITS OF UNCERTAINTY

IN DATA EIGENVECTOR BASIS

N-σ QUANTILE

ξ_n: FRACTION OF BIAS DISTN. WITHIN $n\sigma$ OF TRUTH

NNPDF4.0 CLOSURE TEST RESULTS:

NUMBERS

BIAS/VARIANCE RATIO AND ONE-σ QUANTILE

DATA-SPACE, DATA COVARIANCE MATRIX, OUT-OF-SAMPLE

Dataset	$\sqrt{\mathrm{b/v}}$	$\xi_{1\sigma}^{(\mathrm{data})}$	$\mathrm{erf}(R_{bv}/\sqrt{2})$
DY	0.99 ± 0.08	0.69 ± 0.02	0.69 ± 0.04
Top-pair	0.75 ± 0.06	0.75 ± 0.03	0.82 ± 0.03
Jets	1.14 ± 0.05	0.63 ± 0.03	0.62 ± 0.02
Dijets	0.99 ± 0.07	0.70 ± 0.03	0.69 ± 0.04
Direct photon	0.71 ± 0.06	0.81 ± 0.03	0.84 ± 0.03
Single top	0.87 ± 0.07	0.69 ± 0.04	0.75 ± 0.04
Total	1.03 ± 0.05	0.68 ± 0.02	0.67 ± 0.03

PDF-SPACE & COV MATRIX

flavour	$\xi_{1\sigma}^{(\mathrm{pdf})}$
Σ	0.82 ± 0.04
g	0.70 ± 0.05
V	0.65 ± 0.05
V_3	0.63 ± 0.05
V_8	0.72 ± 0.04
T_3	0.71 ± 0.05
T_8	0.71 ± 0.05
Total	0.71 ± 0.02

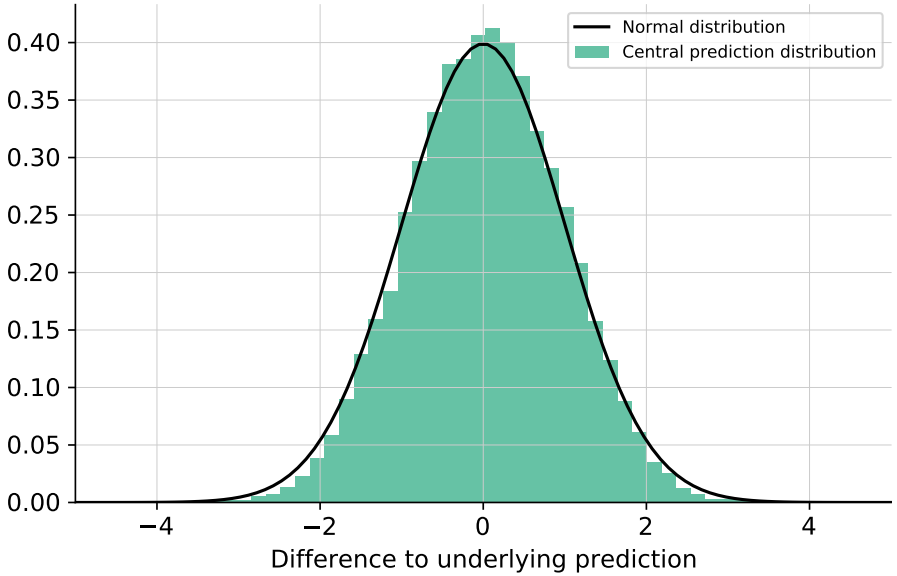
- 25 "UNIVERSE RUNS", 45 REPLICAS EACH
- IN-SAMPLE DATA: PRE 2015
- OUT OF SAMPLE DATA: 2015-2020, MOSTLY LHC
- PDFS HIGHLY CORRELATED $\Rightarrow$ SAMPLED AT 4 POINTS EACH

NNPDF4.0 CLOSURE TEST RESULTS:

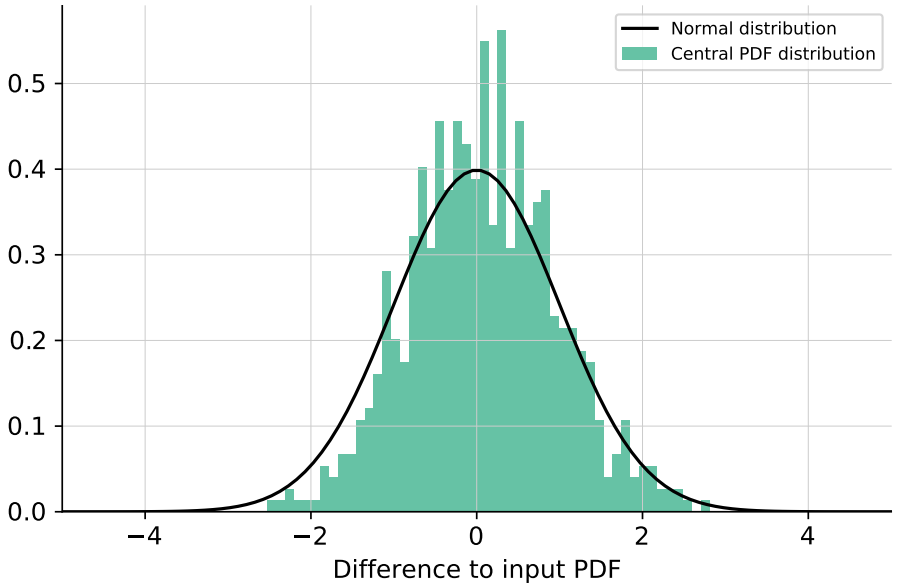
PICTURES

DISTRIBUTION OF DEVIATIONS FROM TRUTH

DATA SPACE (OUT OF SAMPLE)



PDF SPACE



- PDF-SPACE MORE NOISY THAN DATA SPACE

NEW STATISTICAL INDICATORS

A. Barontini, M. Costantini, G. De Crescenzo, S.F., M. Ubiali, in preparation

PDFS AND THEIR UNCERTAINTY (REMINDER)

i-TH DATA PREDICTION (REPLICA AVERAGE): $f_i^{\rm pred} = \frac{1}{N_{\rm rep}} \sum_n f_i^{(n)}$

PDF UNCERTAINTY (REPLICA COVARIANCE): $C_{ij}^{\rm PDF} = \frac{1}{N_{\rm rep}-1} \sum_n \left(f_i^{(n)} - f_i^{\rm pred}\right)\left(f_j^{(n)} - f_j^{\rm pred}\right)$

NORMALIZED BIAS

BIAS: $b^2 = \frac{1}{N_{\rm dat}} \sum_{ij} \left(f_i^{\rm pred} - f_i^{\rm true}\right) C_{ij}^{-1} \left(f_j^{\rm pred} - f_j^{\rm true}\right)$; C_{ij} exp. covmat

NORMALIZED BIAS: $R_b^2 = \frac{1}{N_{\rm dat}} \sum_{ij} \left(f_i^{\rm pred} - f_i^{\rm true}\right) {C_{ij}^{\rm PDF}}^{-1} \left(f_j^{\rm pred} - f_j^{\rm true}\right)$;

DEVIATION FROM TRUTHS IN UNITS OF UNCERTAINTY IN PDF EIGENVECTOR BASIS

- CORRELATES EXPERIMENTALLY UNCORRELATED DATA SENSITIVE TO SAME PDF
- WEIGHS EXPERIMENTS BASED ON NUMBER OF DATA POINT

N-σ QUANTILE

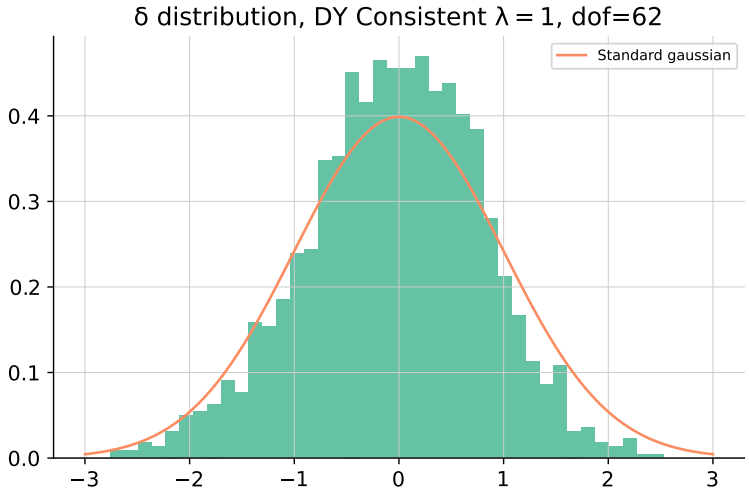
ξ_n: FRACTION OF NORMALIZED BIAS DISTN WITHIN $n\sigma$ OF TRUTH

NEW CONSISTENT CLOSURE TEST

- 25 "UNIVERSE RUNS", 100 REPLICAS EACH
- FULL NNPDF4.0 DATASET, VARIOUS IN-SAMPLE OUT OF SAMPLE PARTITIONS BASED ON REPRESENTATIVE FOLDS
- PDF COVARIANCE MATRIX HIGHLY CORRELATED $\Rightarrow$ PCA REDUCTION

NORMALIZED BIAS DISTRIBUTION

FULL DATASET



$$R_b = 0.88 \pm 0.02; \qquad \xi_{1\sigma} = 0.75 \pm 0.02$$

NNPDF4.0 UNCERTAINTIES SLIGHTLY OVERESTIMATED

THE INCONSISTENCY

THE MODEL

- GENERATE DATA WITH STAT+SYST UNCERTAINTIES $C = C^{\rm stat} + C^{\rm syst}$;

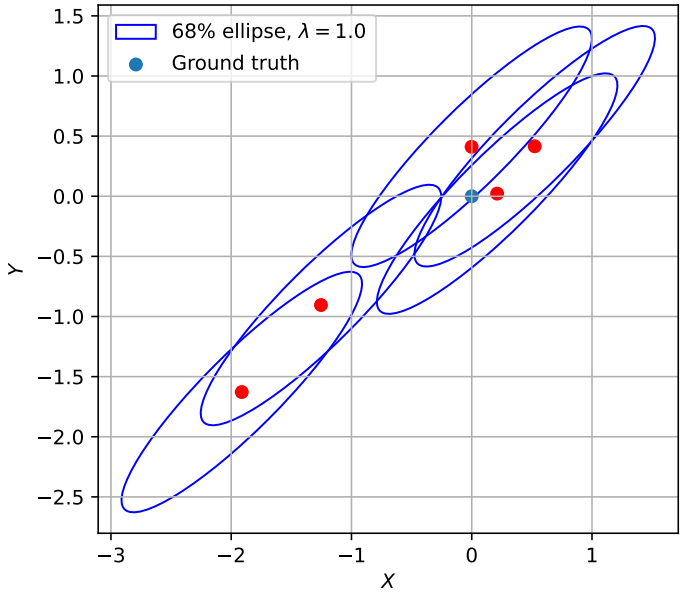
  $C^{\rm syst}_{ij} = \sum_k \Delta^k_i \Delta^k_j$; $\Delta^k_i \Rightarrow$ k-TH SYSTEMATICS ON i-TH DATAPOINT

- ASSUME SYSTEMATICS UNDERESTIMATED $\Rightarrow$ PERFORM FIT WITH $\Delta^k_i \to \lambda \Delta^k_i$

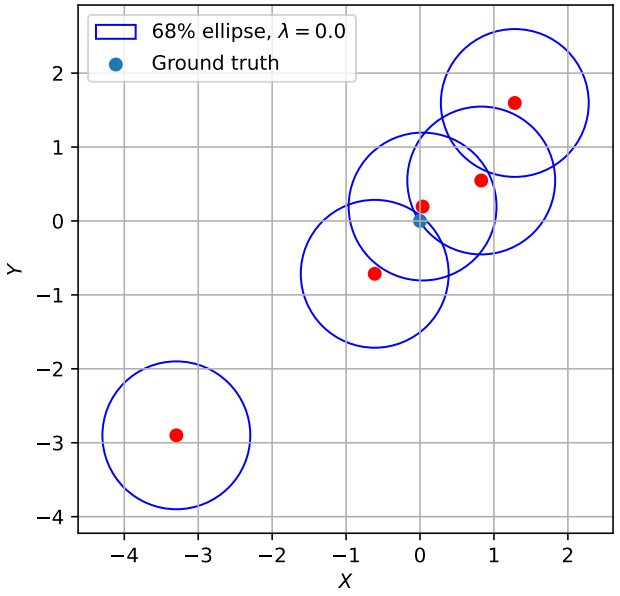
  $\lambda = 1 \Rightarrow$ CONSISTENT; $\lambda = 0 \Rightarrow$ EXTREME INCONSISTENCY

PREDICTED UNCERTAINTY ON GENERATED DATA

CONSISTENT



EXTREME INCONSISTENCY



THE INCONSISTENCY

SCENARIOS

- "DIS": DIS ONLY FIT, IN-SAMPLE HERA DATA INCONSISTENT

  BULK INCONSISTENCY

- "DY": GLOBAL FIT, HIGH-MASS ATLAS 2D DY INCONSISTENT

  SINGLE DATASET INCONSISTENCY

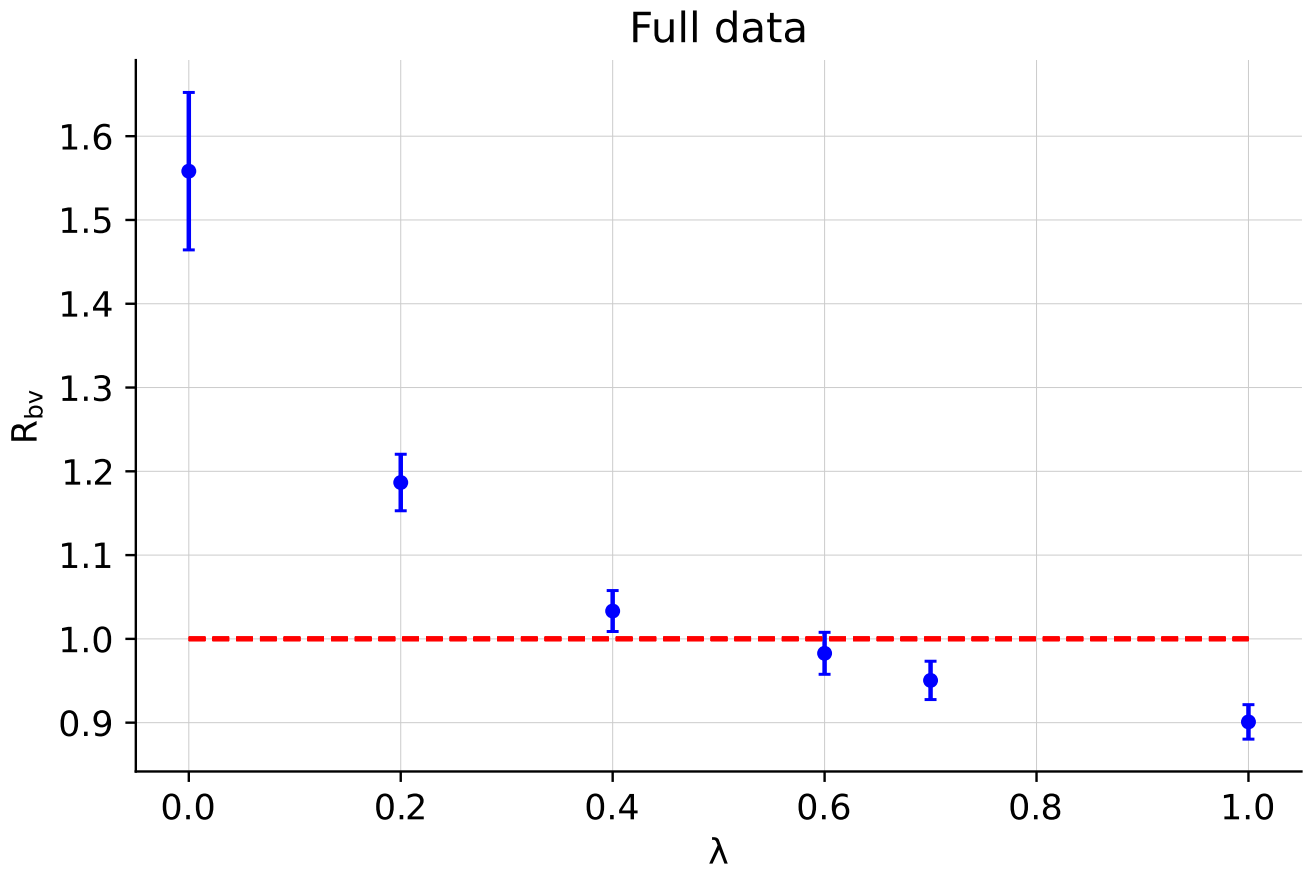
- "JETS": GLOBAL FIT, ALL IN-SAMPLE SINGLE-INCLUSIVE JETS INCONSISTENT

  INCONSISTENCY OF DATA WITH HIGH-IMPACT ON A PDF

THE NORMALIZED BIAS

DIS: BULK INCONSISTENCY

BIAS VS INCONSISTENCY (FULL DATASET)

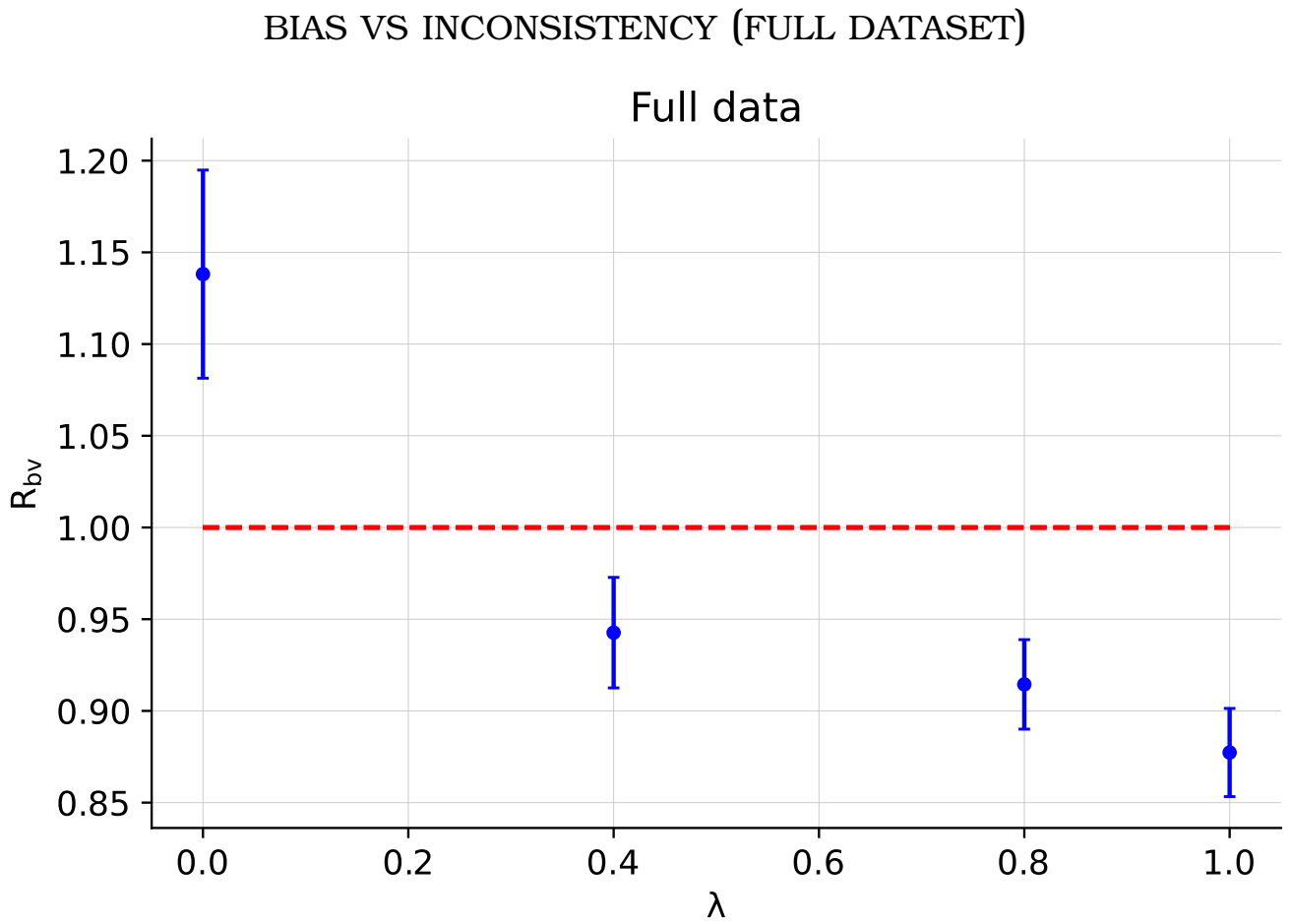


- $\lambda \gtrsim 0.3 \Rightarrow$ MODEL PARTLY CORRECTS INCONSISTENCY

- $\lambda = 0$ NONLINEAR GROWTH: MODEL POOR (UNCERTAINTIES UNDERESTIMATED)

THE NORMALIZED BIAS

DY: SINGLE DATASET INCONSISTENCY

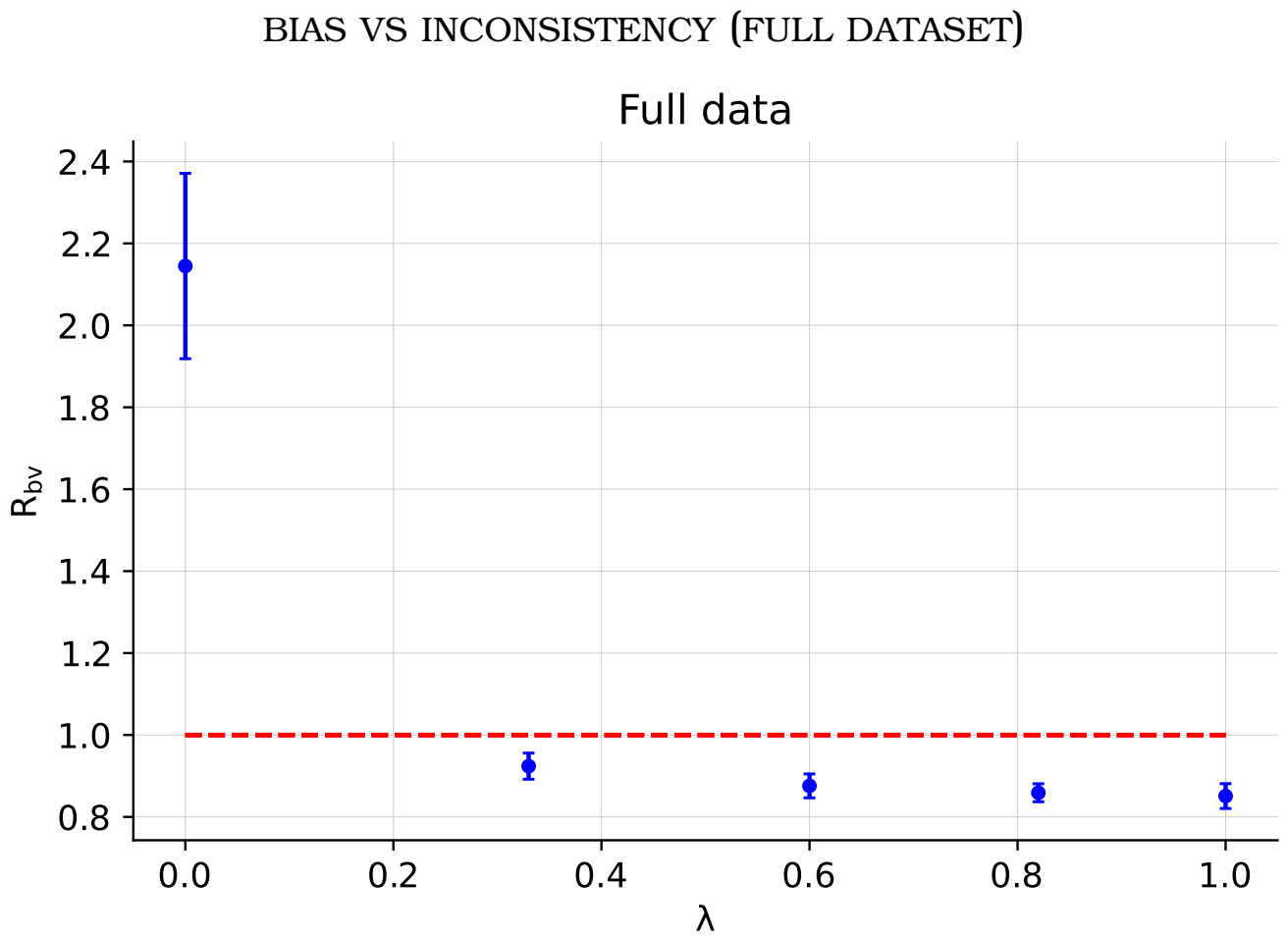


- $\lambda \gtrsim 0.2 \Rightarrow$ MODEL CORRECTS INCONSISTENCY

- $\lambda = 0$ LEAP: MODEL INACCURATE (UNCERTAINTIES SOMEWHAT UNDERESTIMATED)

THE NORMALIZED BIAS

JETS: HIGH-IMPACT INCONSISTENCY



- $\lambda \gtrsim 0.3 \Rightarrow$ MODEL FULLY CORRECTS INCONSISTENCY

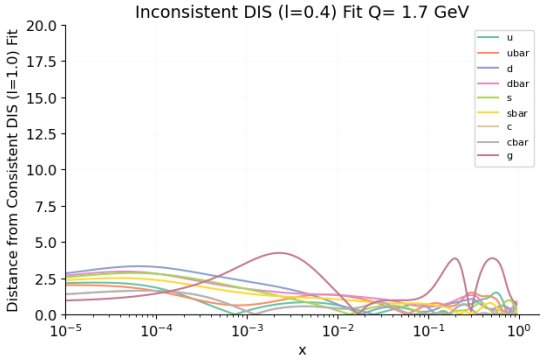
- $\lambda = 0$ PHASE TRANSITION: MODEL FAILS (UNCERTAINTIES COMPLETELY WRONG)

THE PDFs

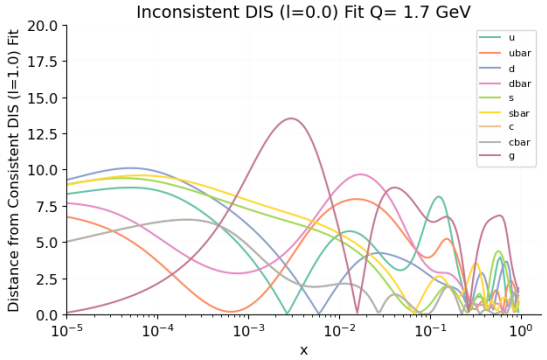
DISTANCES: $\frac{f^{\text{consistent}} - f^{\text{inconsistent}}}{\sigma^{\text{PDF}}/\sqrt{N_{\text{rep}}}}$; $d \sim 1 \Rightarrow$ STATISTICAL EQUIVALENCE

DIS: BULK INCONSISTENCY

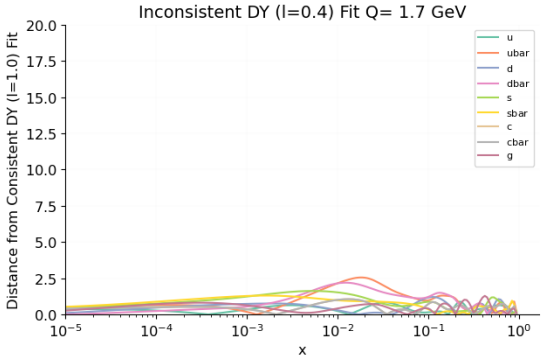
$\lambda = 0.4$: MODEL CORRECTS

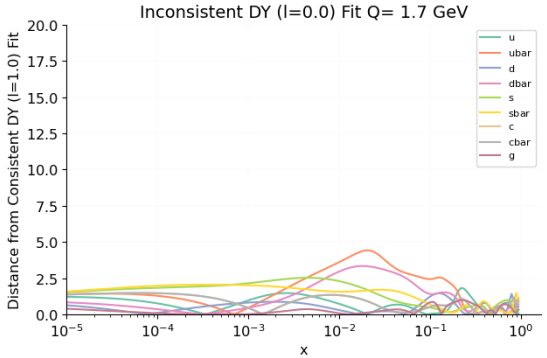


$\lambda = 0$: MODEL FAILS

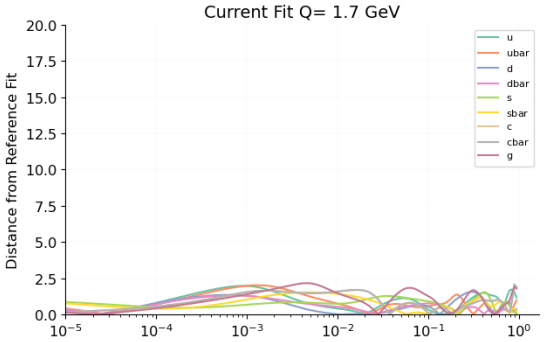


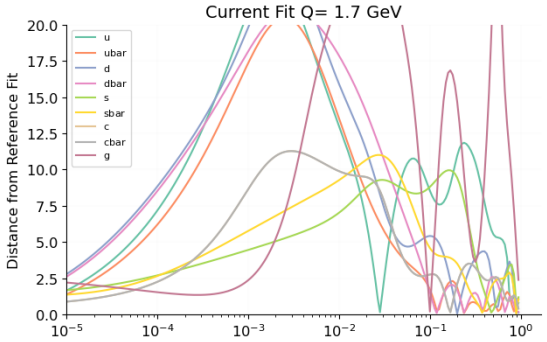
DY: SINGLE DATASET INCONSISTENCY





JETS: HIGH-IMPACT INCONSISTENCY



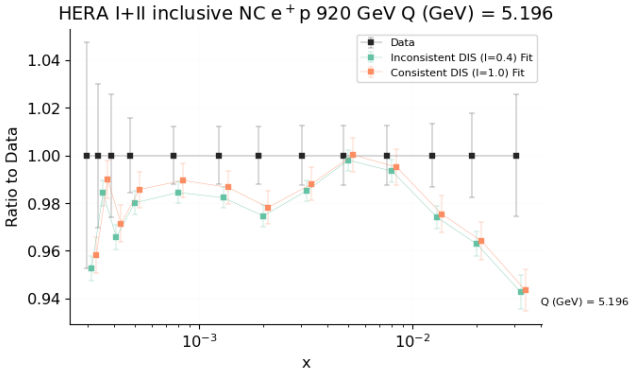


DATA vs. PREDICTION I

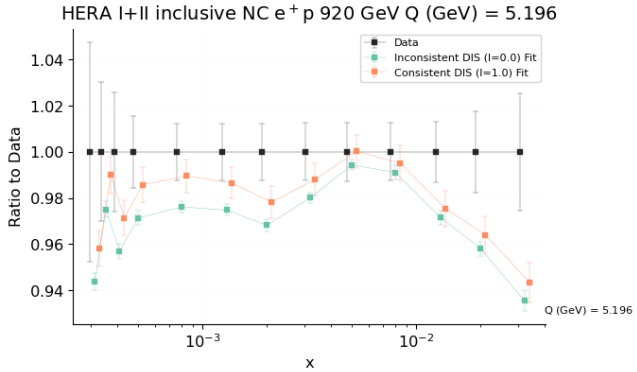
INCONSISTENT VS. CONSISTENT PDF PREDICTIONS FOR INCONSISTENT DATASET

DIS: BULK INCONSISTENCY

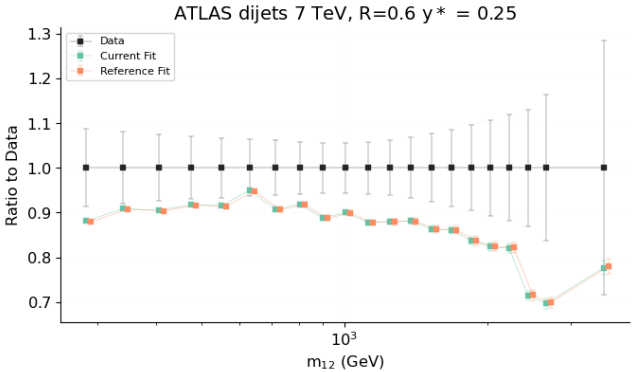
$\lambda = 0.4$: MODEL CORRECTS

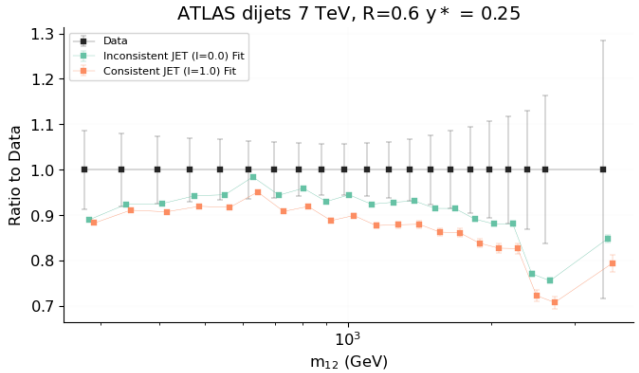


$\lambda = 0$: MODEL FAILS



JETS: HIGH-IMPACT INCONSISTENCY





- $\lambda = 0.4$: MODEL CORRECTS, PREDICTION DOES NOT MOVE

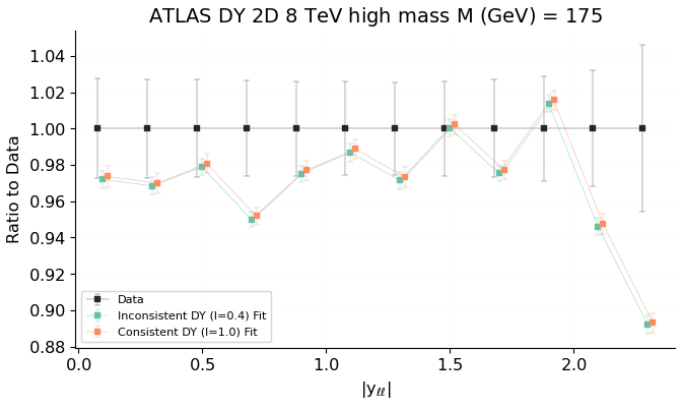
- $\lambda = 0$: MODEL FAILS, PREDICTION OFF WITH UNCHANGED UNCERTAINTY

DATA vs. PREDICTION II

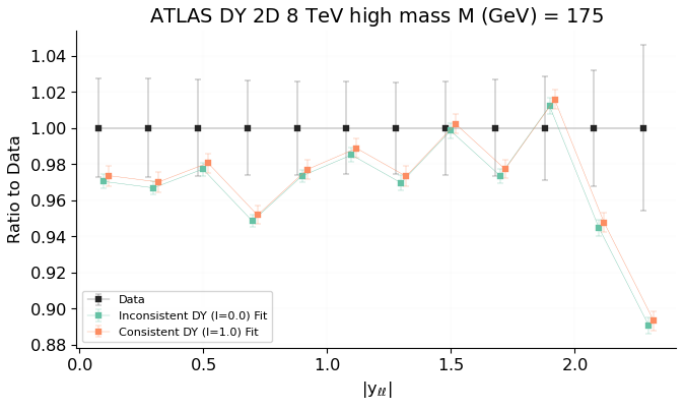
INCONSISTENT VS. CONSISTENT PDF PREDICTIONS FOR INCONSISTENT DATASET

DY: SINGLE DATASET INCONSISTENCY

$\lambda = 0.4$: MODEL CORRECTS



$\lambda = 0$: MODEL FAILS



- PREDICTION DOES NOT MOVE
- $\lambda = 0.4$: MODEL CORRECTS, UNCERTAINTY UNCHANGED

  DESPITE SMALLER DATA UNCERTAINTY
- $\lambda = 0$: MODEL FAILS, UNCERTAINTY SHRINKS DRIVEN BY SMALLER DATA UNCERTAINTY

FLAGGING INCONSISTENCIES

VALIDATING THE NNPDF4.0 DATASET SELECTION METHOD

- LARGE $R_b \Rightarrow$ LARGE χ^2_p OF INCONSISTENT DATASET
- χ^2_p OF CONSISTENT DATASETS SENSITIVE TO SAME PDFS ALSO LARGE

EXAMPLE:

INCONSISTENT ATLAS JET DATA WITH IN-SAMPLE CMS JET DATA:

- $n_\sigma = \frac{\chi^2_p - 1}{\sqrt{N_{\rm dat}/2}}$: ATLAS TRUE POSITIVE: $n_\sigma = 4.8$; CMS FALSE POSITIVE: $n_\sigma = 2.7$

REPEAT FIT WITH LARGE WEIGHT TO PUTATIVE INCONSISTENT DATASETS

- WEIGHTED FIT:
  - ATLAS INCONSISTENT TRUE POSITIVE: $n_\sigma = 2.8$ REMAINS INCONSISTENT
  - CMS CONSISTENT FALSE POSITIVE: $n_\sigma = 0.7$ BECOMES CONSISTENT

CONCLUSIONS

- NNPDF4.0 PDF UNCERTAINTIES SOMEWHAT OVERESTIMATED ($\sim 10\%$)
- ML MODEL CORRECTS FOR INCONSISTENCY EXCEPT IN EXTREME CASES
- EXTREME INCONSISTENCY SIGNALED BY LARGE χ^2,

  DETECTABLE BY WEIGHTED FIT