NNPDF

DETERMINATION AND VALIDATION

OF

MODELING & THEORY UNCERTAINTIES

STEFANO FORTE

UNIVERSITÀ DI MILANO & INFN

UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI FISICA

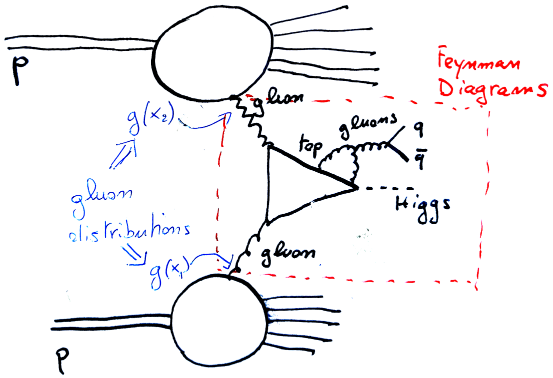
INFN Istituto Nazionale di Fisica Nucleare

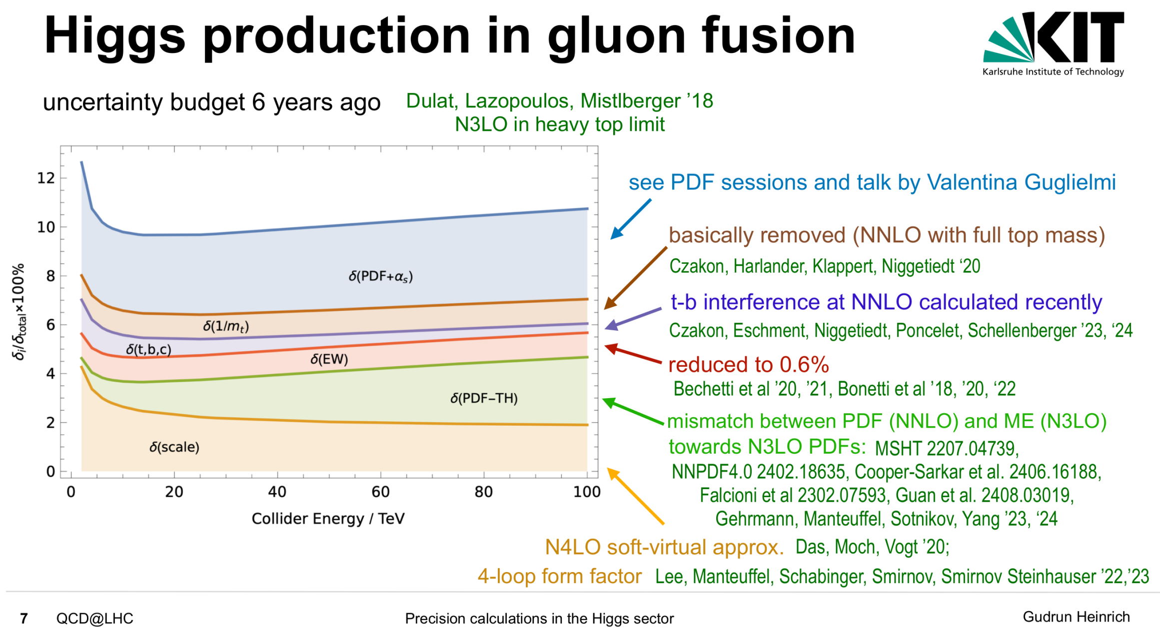
UQ4ML COMETA WORKSHOP

CEA SACLAY, SEPTEMBER 16, 2025

PDFs+α_s UNCERTAINTY

QCD FACTORIZATION





G. Heinrich, QCDLHC 2024

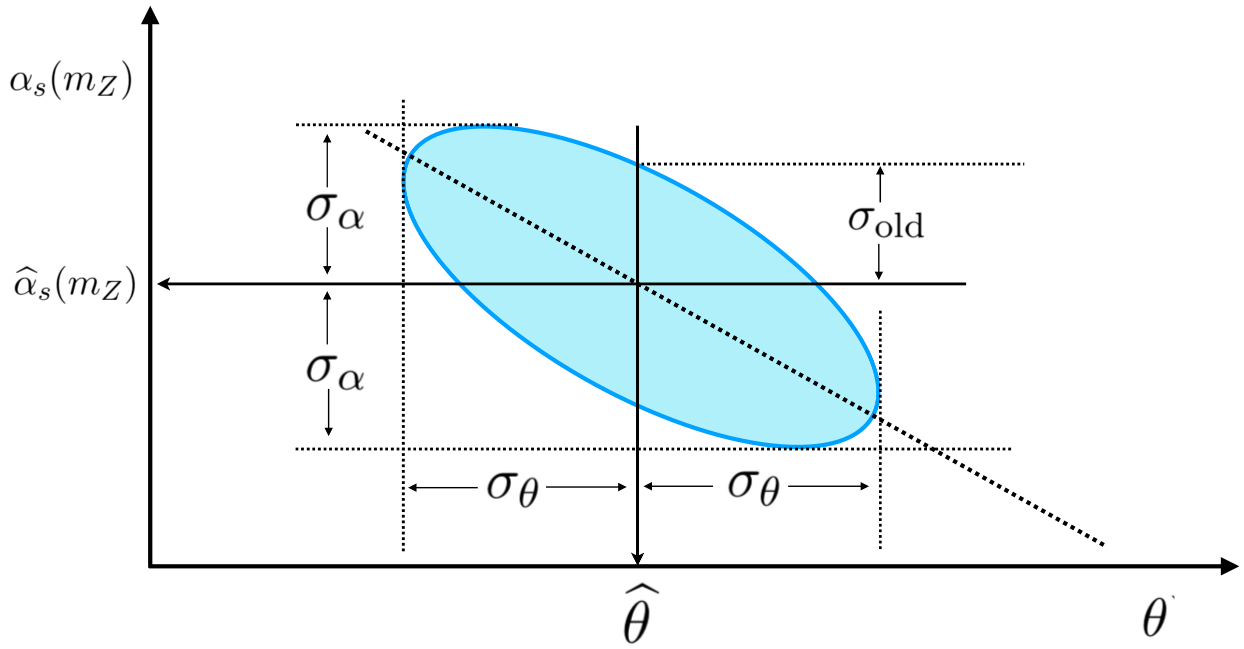
- FACTORIZED "PROBABILITY" OF A QUARK OR GLUON (PARTONS) TO PARTICIPATE IN HARD INTERACTION
- DOMINANT UNCERTAINTY: UNKNOWN TRUE VALUE
  - THEORY PREDICTION
  - PDF MODEL

TEST PROBLEM: DETERMINATION OF α_s

THE PROBLEM

- DETERMINING THEORY & MODEL UNCERTAINTIES
  - COVARIANCE MATRIX VS. NUISANCE PARAMETERS
  - BAYESIAN VS. FREQUENTIST
- VALIDATION & BIAS DETECTION
  - CLOSURE TESTING
  - "FUTURE" TESTS

α_s DETERMINATION



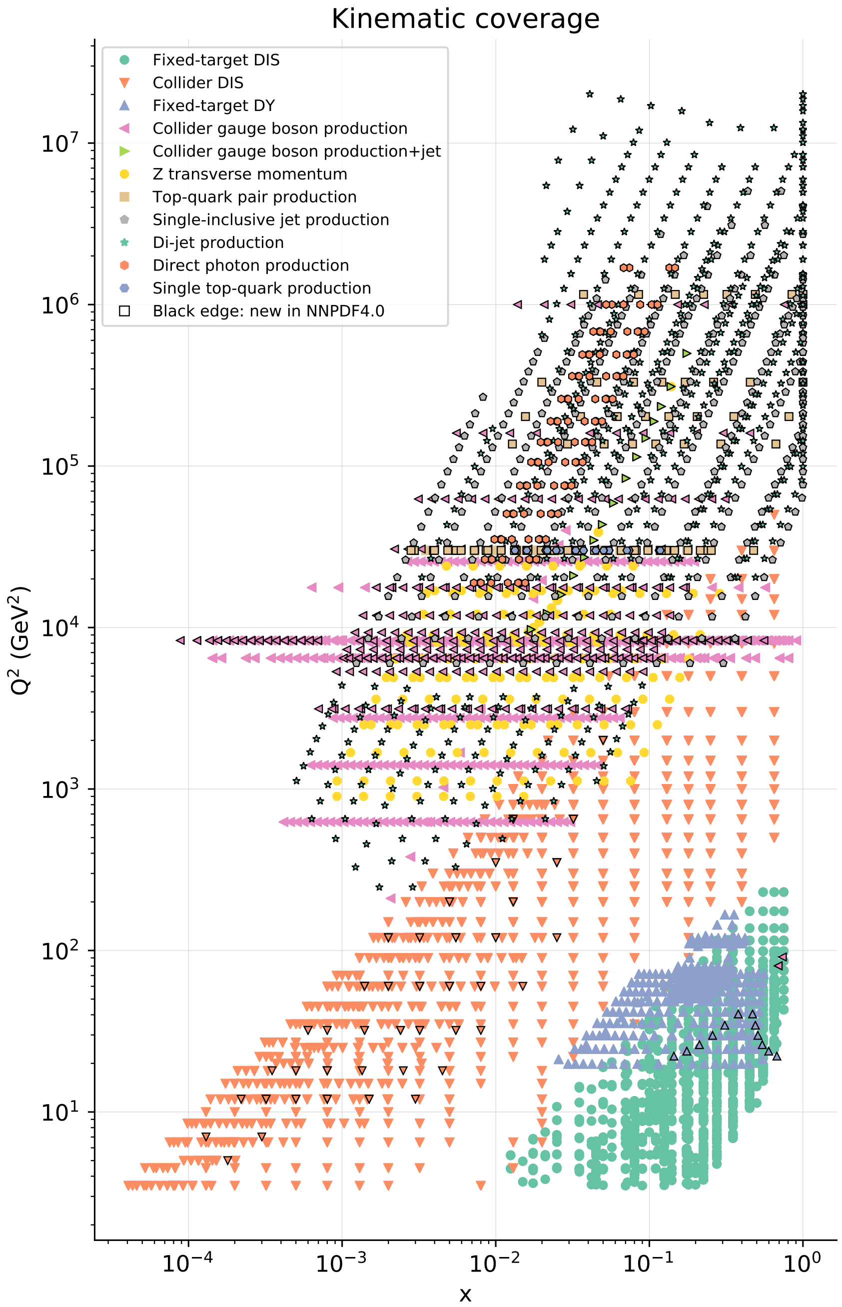
$\theta \Leftrightarrow$ PDF

- MINIMUM DETERMINED ALONG THE "BEST PDF" LINE $\Rightarrow \sigma_{\text{old}}$ UNDERESTIMATE α_s UNCERTAINTY
- NEED SIMULTANEOUS MINIMZATION IN (PDF, α_s) SPACE

REGRESSION FROM DATA

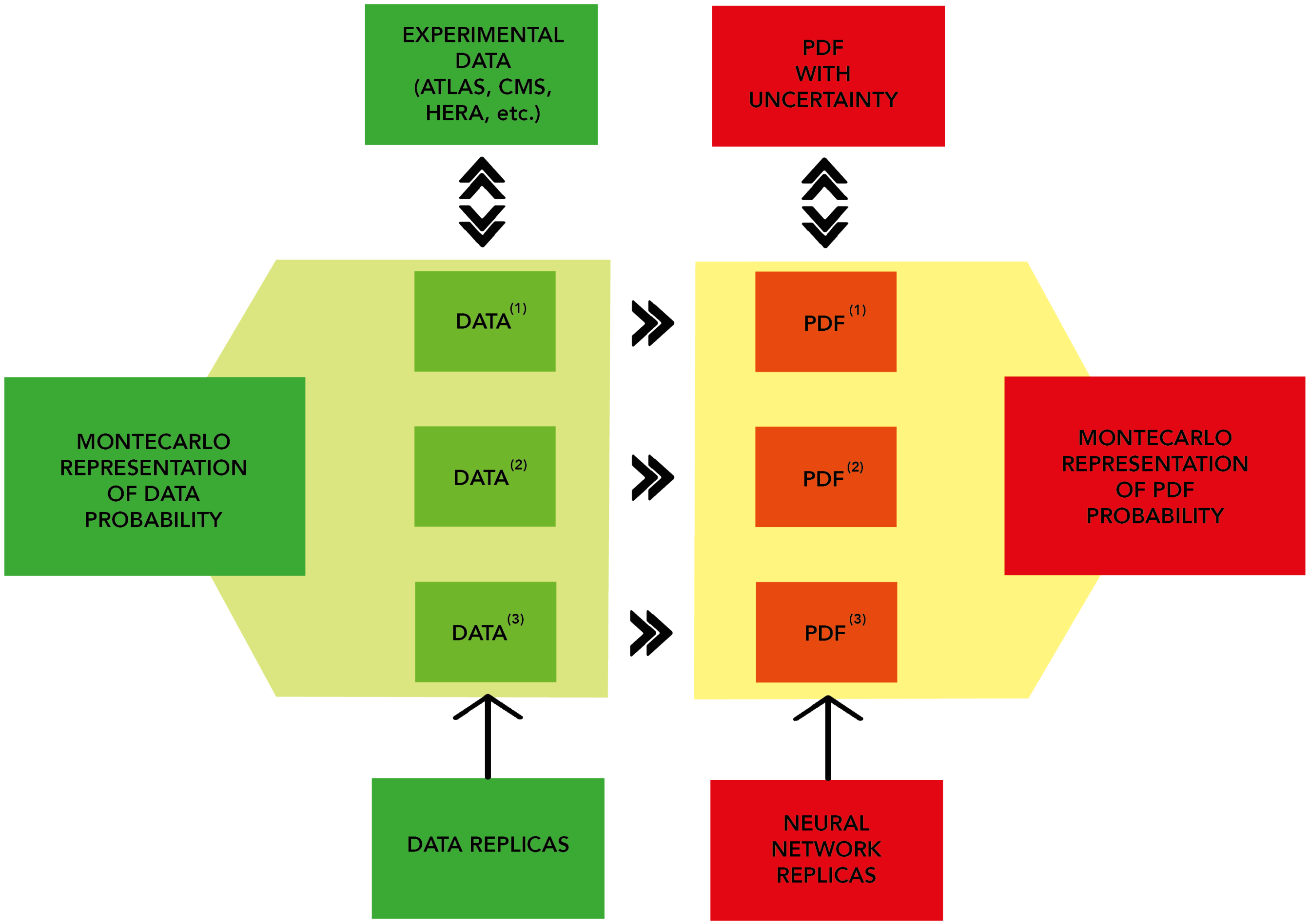
- LHC CROSS SECTION:
  - $\sigma = \sum_{ij} \hat{\sigma}_{ij} \otimes f_i^{(1)} \otimes f_j^{(2)}$
  - $\hat{\sigma}_{ij}(\alpha_s(Q^2))$ PARTONIC CROSS SECTION, INCOMING PARTONS i, j
  - $f_i^{(j)} = \sum_{i'} \Gamma_{ii'}(\alpha_s(Q^2)) f_{i'}^{(j)}(x)$ PDF FOR PARTON OF SPECIES i IN j-TH INCOMING PROTON
  - $\Gamma_{ii'}(\alpha_s(Q^2))$ EVOLUTION FACTOR
  - $\otimes$ CONVOLUTION OVER x
- PARTONIC CROSS SECTION & EVOLUTION FACTOR PERTURBATIVE SERIES IN α_s
- PDFs $f_i(x)$ REGRESSED GIVEN DATA & THEORY
  - ABOUT 4600 DATAPOINTS
  - LEPTOPRODUCTION & HADROPRODUCTION, COLLIDER & FIXED-TARGET

THE NNPDF4.0 DATASET



PROBABILITY REGRESSION

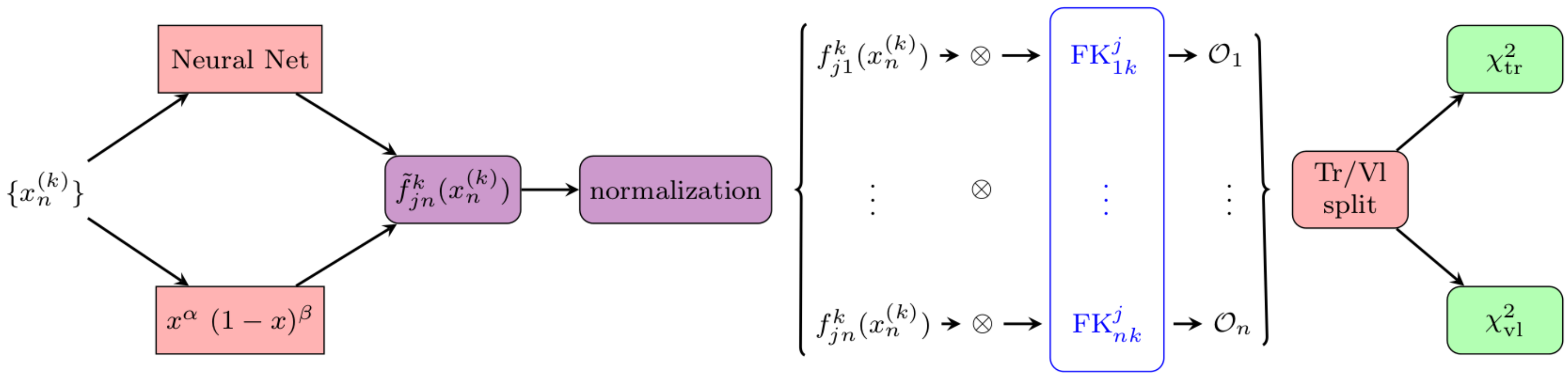
REPLICA SAMPLE OF FUNCTIONS $\Leftrightarrow$ PROBABILITY DENSITY IN FUNCTION SPACE
KNOWLEDGE OF LIKELIHHOD SHAPE (FUNCTIONAL FORM) NOT NECESSARY

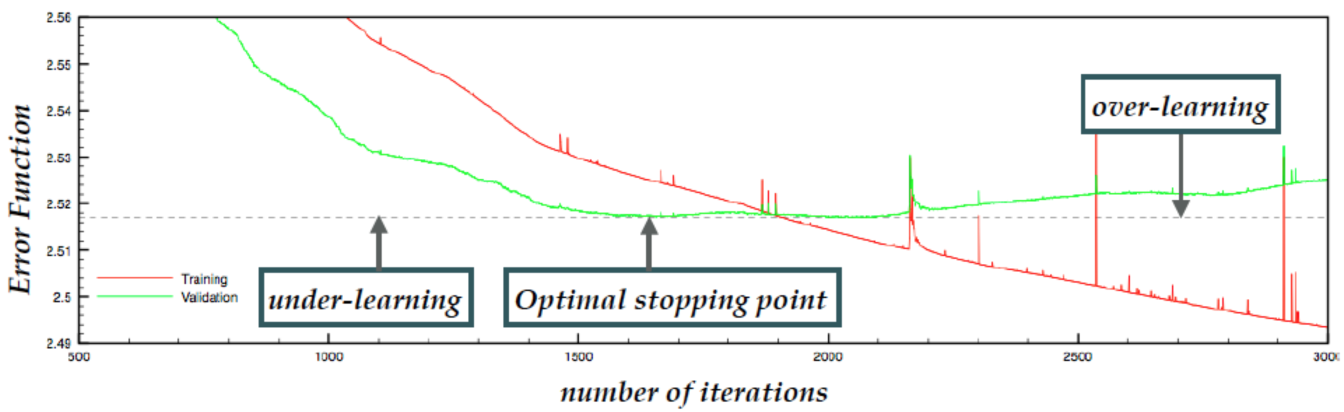


FINAL PDF SET: $f_i^{(a)}(x, \mu)$;
i =up, antiup, down, antidown, strange, antistrange, charm, gluon; $j = 1, 2, \ldots N_{\rm rep}$

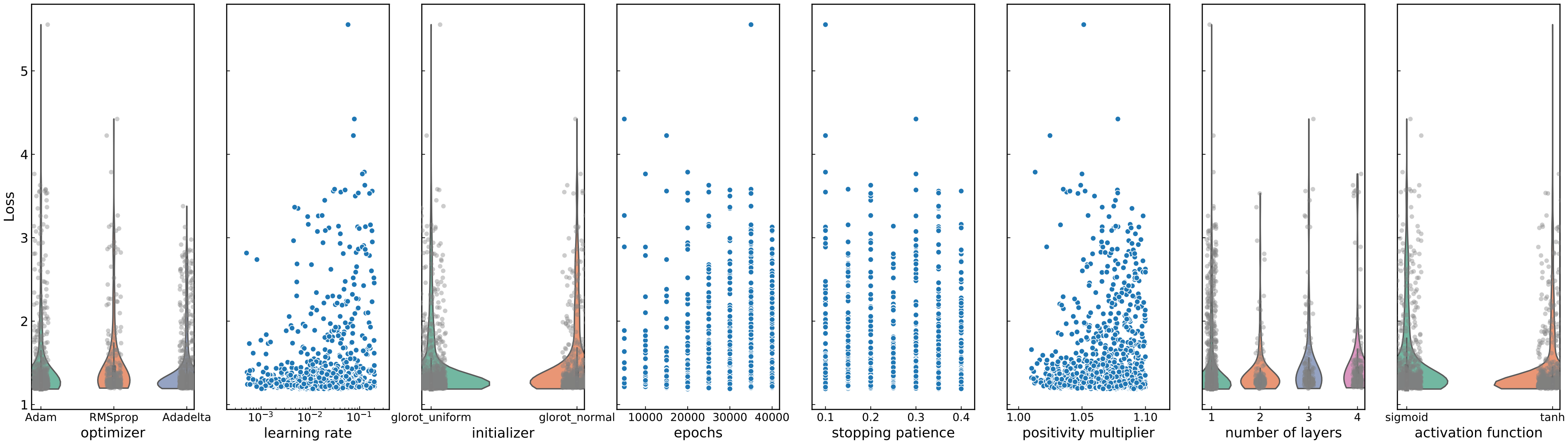
CROSS-VALIDATED LEARNING

- PDFs modeled by a feed-forward neural network
- Neural net parameters determined by χ^2 minimization through gradient descent
- Random training-validation split, χ^2 to training data replicas minimized
- Lowest validation χ^2 $\Rightarrow$ optimal fit





METHODOLOGY HYPEROPTIMIZATION

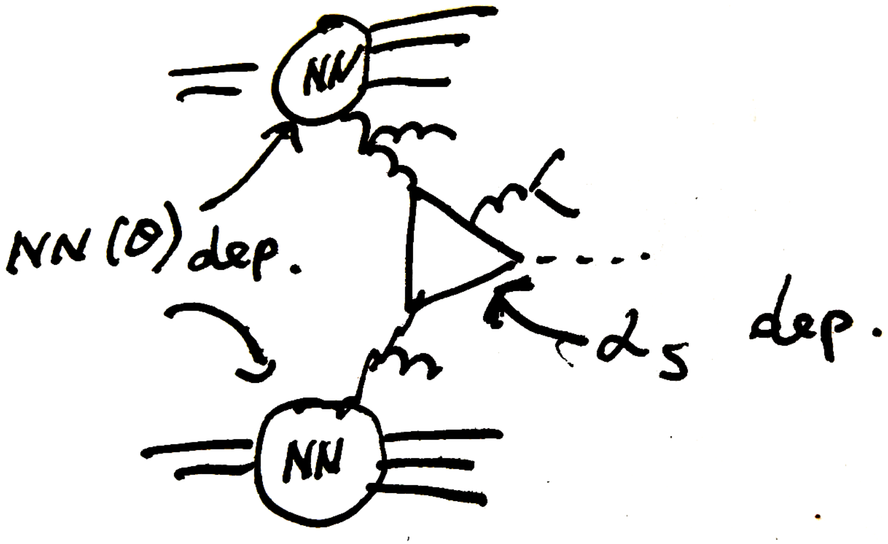


HYPEROPT PARAMETERS

NEURAL NETWORK	FIT OPTIONS
NUMBER OF LAYERS (*)	OPTIMIZER (*)
SIZE OF EACH LAYER	INITIAL LEARNING RATE (*)
DROPOUT	MAXIMUM NUMBER OF EPOCHS (*)
ACTIVATION FUNCTIONS (*)	STOPPING PATIENCE (*)
INITIALIZATION FUNCTIONS (*)	POSITIVITY MULTIPLIER (*)

- SCAN PARAMETER SPACE
- OPTIMIZE FIGURE OF MERIT: K-FOLDING LOSS

SIMULTANEOUS PDF-α_s DETERMINATION
THE PROBLEM



$$\sigma = \sum_{ij} \hat{\sigma}_{ij}(\alpha_s) \otimes f_i^{(1)}(\theta) \otimes f_j^{(2)}(\theta)$$

- PDF PARAMETERS θ ENTER ML MODEL OF PDFs $f_i^{(a)}(\theta)$
- α_s ENTERS THEORY PREDICTION $\hat{\sigma}_{ij}(\alpha_s)$

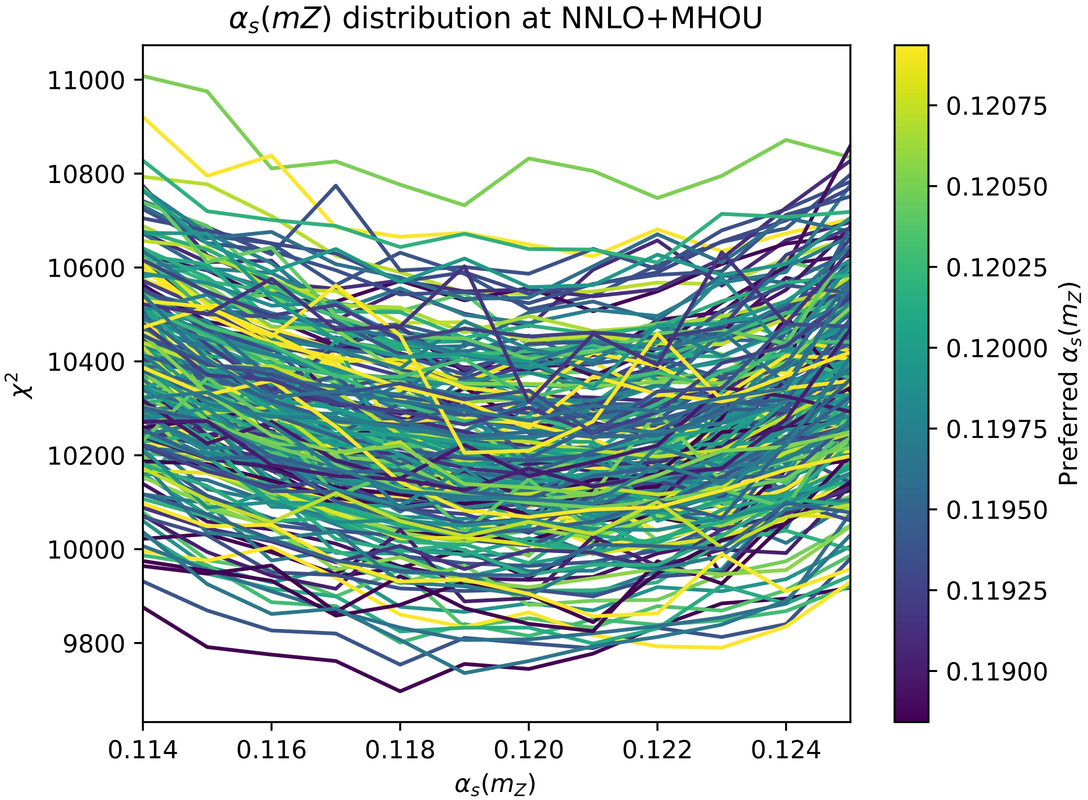
SOLUTIONS

- ✗ THEORY INTERPOLATION ⇒ SIMUNET (S. Iranipour and M. Ubiali, 2022)
- ✔ FREQUENTIST MC RESAMPLING ⇒ CORRELATED REPLICA METHOD (NNPDF, 2018)
- ✔ BAYESIAN POSTERIOR PARM. ESTIMATION ⇒ THEORY COVARIANCE MATRIX METHOD (R. D Ball and R. Pearson, 2021)

THE CORRELATED REPLICA METHOD
NNPDF3.1 (2018)

- DETERMINE BEST-FIT PDF REPLICA TO EACH DATA REPLICA FOR SEVERAL (DISCRETE) α_s VALUES
- EACH DATA-REPLICA $\Rightarrow$ χ^2 PROFILE $\Rightarrow$ α_s VALUE
- MC SAMPLE OF α_s REPLICAS

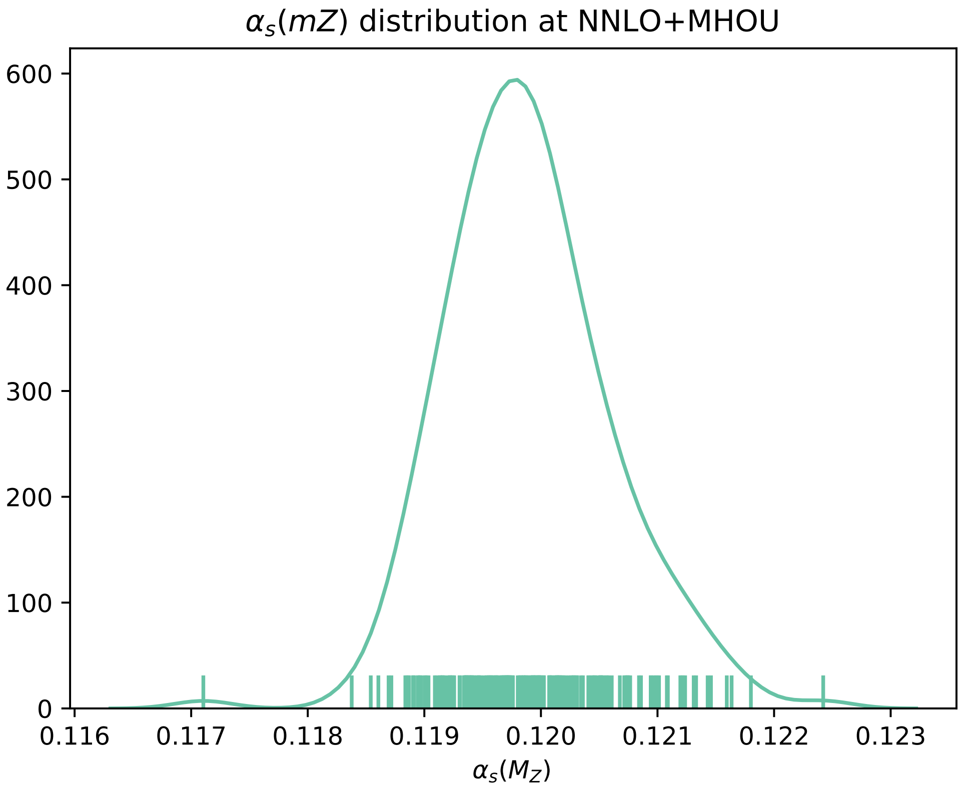
ENSEMBLE OF α_s VALUES



THE CORRELATED REPLICA METHOD
NNPDF3.1 (2018)

- DETERMINE BEST-FIT PDF REPLICA TO EACH DATA REPLICA FOR SEVERAL (DISCRETE) α_s VALUES
- EACH DATA-REPLICA $\Rightarrow$ χ^2 PROFILE $\Rightarrow$ α_s VALUE
- MC SAMPLE OF α_s REPLICAS

ENSEMBLE OF α_s VALUES



THE THEORY COVARIANCE METHOD

CORRELATED UNCERTAINTY COVMAT S $\Leftrightarrow$ NUISANCE PARAMETER λ ON PREDICTION T

$$S_{ij} = \beta_i \beta_j \qquad T \to T + \lambda\beta$$

- PROBABILITY OF PREDICTION T GIVEN DATA & λ
  $P(T|D,\lambda) \propto \exp\left[-\frac{1}{2}(T+\lambda\beta-D)^T C^{-1}(T+\lambda\beta-D)\right]$
- UNCERTAINTY ON NUISANCE PARM $P(\lambda) \propto \exp\left(-\frac{\lambda^2}{2}\right)$
- PROBABILITY OF PREDICTION T GIVEN DATA
  $P(T|D) \propto \exp\left[-\frac{1}{2}(T-D)^T (C+S)^{-1}(T-D)\right]$
- POSTERIOR DISTRIBUTION OF λ!: $P(\lambda|T,D) \propto \exp\left(-\frac{1}{2}Z^{-1}\left(\lambda-\bar{\lambda}(T,D)\right)^2\right)$,
  WITH $\bar{\lambda}(T,D) = \beta^T (C+S)^{-1}(D-T), \quad Z = 1-\beta^T (C+S)^{-1}\beta$

IDEA

TREAT $\alpha_s - \alpha_s^0$ AS NUISANCE PARAMETER WITH PRIOR $P(\alpha_s)$
CENTERED ABOUT PRIOR α_s^0

VALIDATION:
CLOSURE TESTS

- ASSUME "TRUE" UNDERLYING PDF & α_s VALUE
- GENERATE DATA DISTRIBUTED ACCORDING TO EXPERIMENTAL COVARIANCE MATRIX
- RUN WHOLE METOHDOLOGY ON THESE DATA
- DO STATISTICS ON "RUNS OF THE UNIVERSE":
  TURNING BAYESIAN INTO FREQUENTIST
  - BIAS/VARIANCE: MEAN SQUARE DEVIATION WR TO TRUTH VS UNCERTAINTY
    $$R_{\rm bv} = \sqrt{\frac{1}{N_r}\sum_{j=1}^{N_r}\left(\frac{\alpha_s^{(j)} - \alpha^{\bar{\rm true}}{}_s}{\sigma_\alpha^{(j)}}\right)^2}$$
  - ONE-SIGMA QUANTILE $\xi_{1\sigma}$: IS TRUTH WITHIN ONE SIGMA 68% OF TIMES?

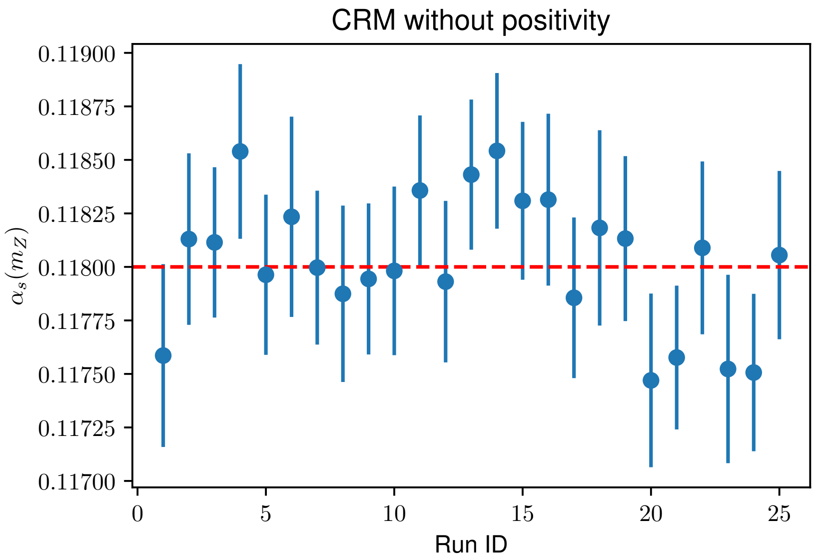
CLOSURE TEST RUNS

- CRM (FREQ.) 250 MC REPLICAS $\times$ 12 VALUES $0.114 \leq \alpha_s \leq 0.123$;
  25 RUNS OF THE UNIVERSE

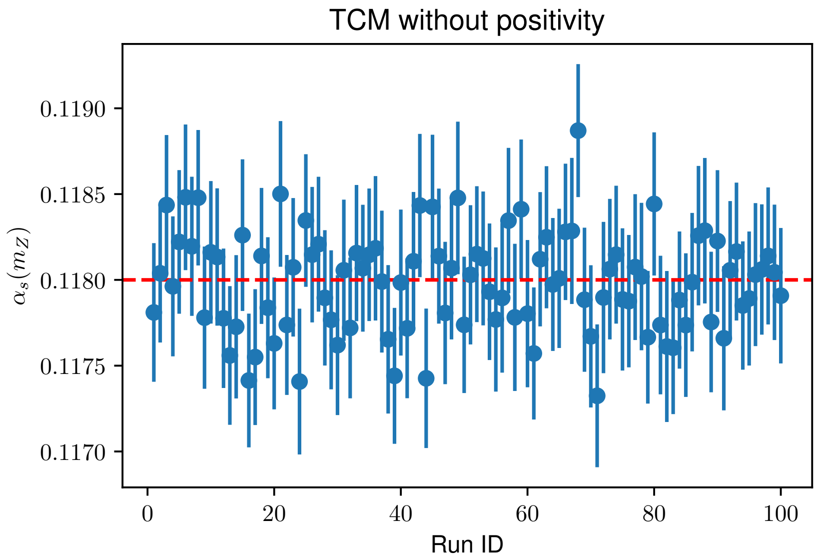
- TCM (BAYS.) 550 MC REPLICAS; 100 RUNS OF THE UNIVERSE

N_r RUNS $\Rightarrow$ DETERMINE $\langle \alpha_s \rangle$, UNCERTAINTY $\frac{\sigma_\alpha}{\sqrt{N_r}}$; PULL $P = \frac{\langle \alpha_s \rangle - \alpha_s^{\mathtt{true}}}{\sigma_\alpha / \sqrt{N_r}}$

CRM RUNS



TCM RUNS

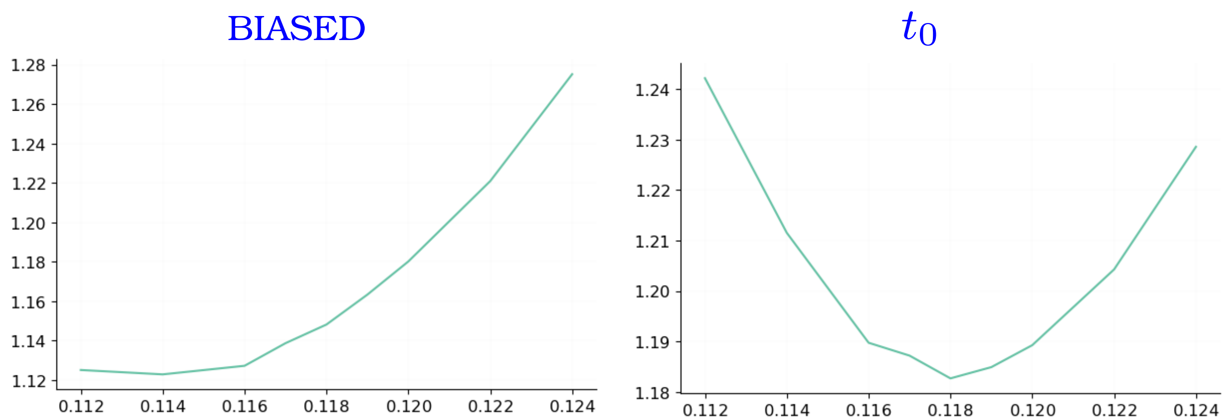


STABILITY TESTS

- CRM
  - PARABOLA $\Rightarrow$ HIGHER ORDER POLYNOMIALS
  - INDEPENDENT VARIABLE: $\ln \alpha_s$ VS. α_s
  - MULTI-BATCH MINIMIZATION
- TCM
  - CHOICE OF PRIOR (WIDTH AND CENTERING)
- EXP. VS t_0 COVMAT IN DATA GENERATION

DETECTING BIAS
MULTIPLICATIVE UNCERTAINTIES
A KNOWN PROBLEM

- MULTIPLICATIVE UNCERTAINTIES (E.G. LUMI): $\mathrm{cov}_{ij} = \sigma_i^{\mathcal{L}} \sigma_j^{\mathcal{L}} D_i D_j$; D_i DATA
- CANNOT BE USED IN LIKELIHOOD $\Rightarrow$ BIAS (d'Agostini, 1994)
- SOLUTION (t_0): $\mathrm{cov}_{ij} = \sigma_i^{\mathcal{L}} \sigma_j^{\mathcal{L}} t_i^{(0)} t_j^{(0)}$; $t_i^{(0)}$ THEORY PRED. FROM PREVIOUS



A NEW PROBLEM

- THEORY PREDICTION DEPENDS ON α_s
- VARYING $t^{(0)}(\alpha_s)$ OR FIXED $t^{(0)}$?

INPUT $\alpha_s = 0.118$; CRM 25 RUNS

VARYING $t^{(0)}(\alpha_s)$

$\langle \alpha_s \rangle = 0.119450$; $\langle \sigma_\alpha \rangle / \sqrt{N_r} = 0.000077$

$P = 19$ $R_{\rm bv} = 3.8 \pm 0.16$

FIXED $t^{(0)}$

$\langle \alpha_s \rangle = 0.118152$; $\langle \sigma_\alpha \rangle / \sqrt{N_r} = 0.000070$

$P = 2.2$ $R_{\rm bv} = 0.97 \pm 0.11$

DETECTING BIAS

INPUT $\alpha_s = 0.118$; TCM 100 RUNS

POSITIVITY

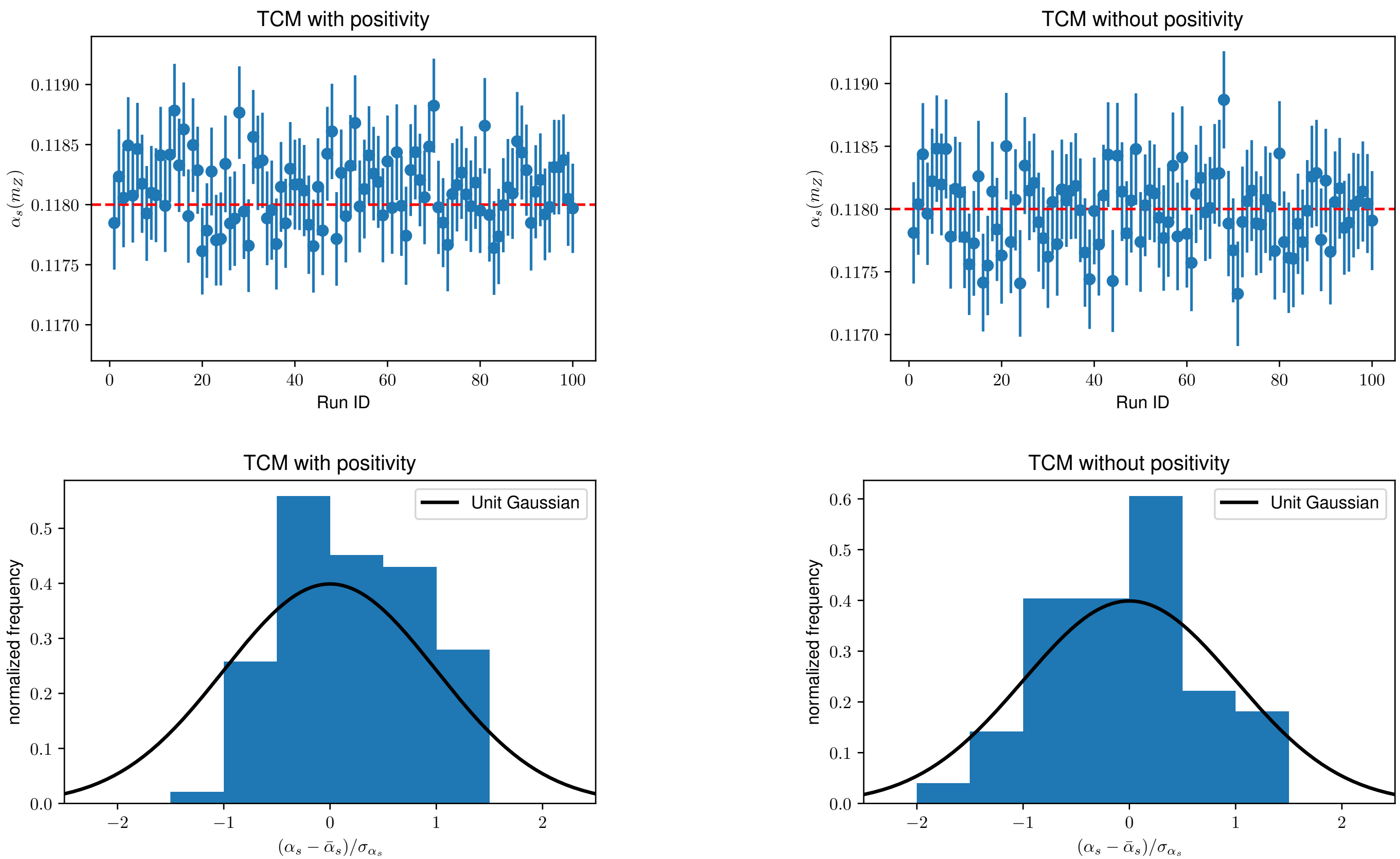
$\langle \alpha_s \rangle = 0.118132$; $\langle \sigma_\alpha \rangle / \sqrt{N_r} = 0.000039$

$P = 3.4$ $R_{\rm bv} = 0.80 \pm 0.06$

NO POSITIVITY

$\langle \alpha_s \rangle = 0.117984$; $\langle \sigma_\alpha \rangle / \sqrt{N_r} = 0.000041$

$P = 0.39$ $R_{\rm bv} = 0.71 \pm 0.05$



- POSITIVITY OF OBSERVABLES: PERFECT UNCERTAINTIES; BIASED CENTRAL VALUE
- DETECTED WITH MULTI-RUNS
- POSITIVITY $\Leftrightarrow$ NONGAUSSIAN BEHAVIOR

FINAL RESULTS

INPUT $\alpha_s = 0.118$

TCM 100 RUNS

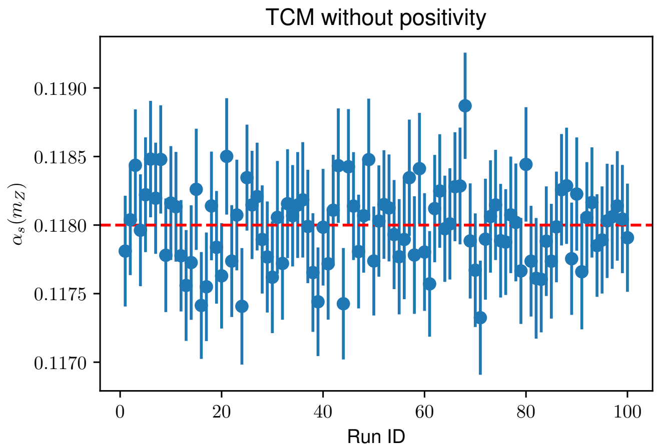
$\langle \alpha_s \rangle = 0.117984; \langle \sigma_\alpha \rangle / \sqrt{N_r} = 0.000041$

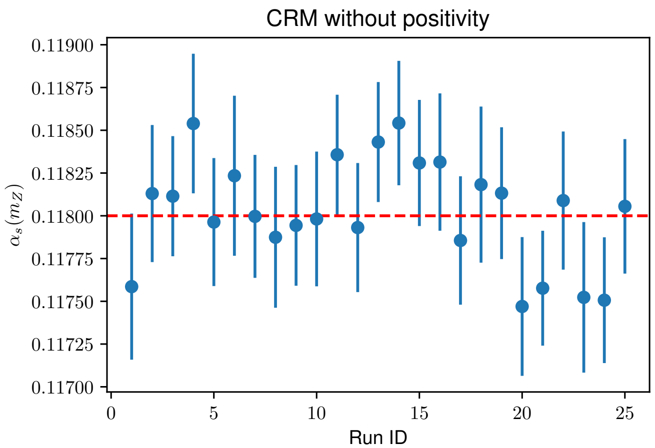
$P = 0.39 \; R_{\rm bv} = 0.71 \pm 0.05$

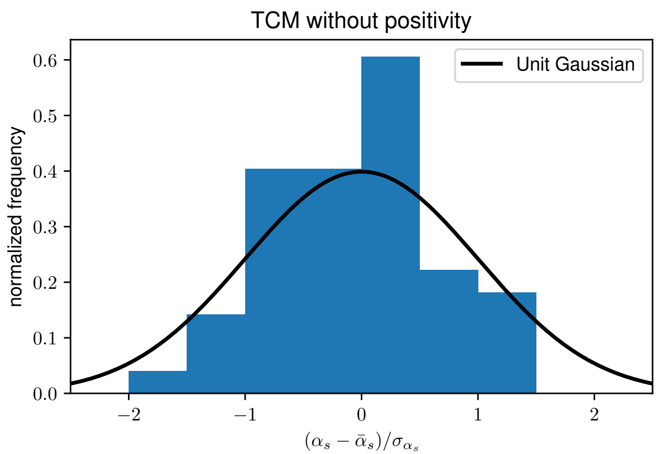
CRM 25 RUNS

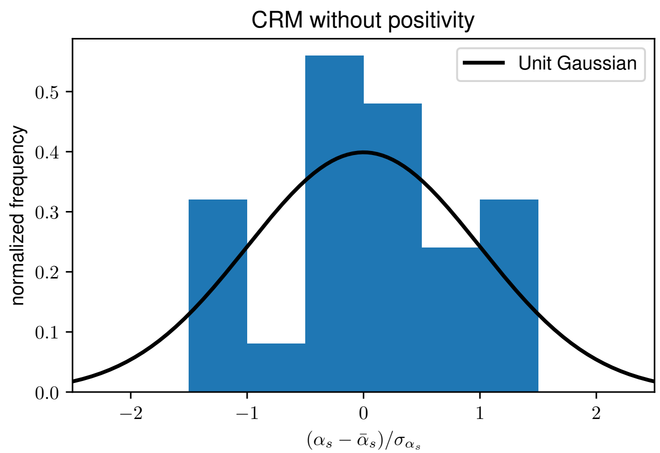
$\langle \alpha_s \rangle = 0.118029; \langle \sigma_\alpha \rangle / \sqrt{N_r} = 0.000077$

$P = 0.38 \; R_{\rm bv} = 0.80 \pm 0.09$









- BAYESIAN (TCM) AND FREQUENTIST (CRM) IN PERFECT AGREEMENT
- VALIDATION OF THEORY UNCERTAINTIES

CONCLUSIONS

- UNCERTAINTIES DETERMINED USING
  - BAYESIAN THEORY COVARIANCE
  - FREQUENTIST REPLICA SAMPLING
- VALIDATED THROUGH CLOSURE TEST
- BIAS DETECTED & REMOVED
- AGREEMENT BETWEEN TWO METHODS