WHAT DOES PROTON STRUCTURE TEACH US ON MACHINE LEARNING AND VICE-VERSA

STEFANO FORTE
UNIVERSITÀ DI MILANO & INFN

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI FISICA

Istituto Nazionale di Fisica Nucleare

THEORY SEMINAR

TUM, DECEMBER 14, 2023

SUMMARY

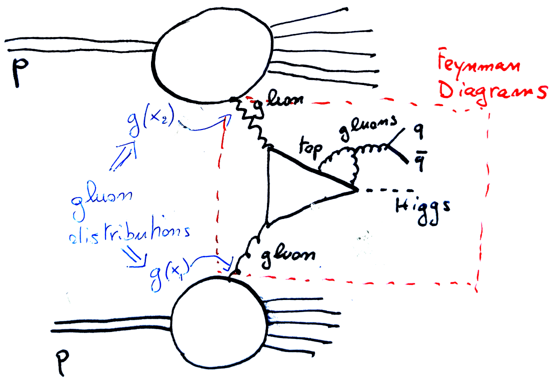
- THE PROBLEM
  - PDFS AND THEIR UNCERTAINTIES
  - WHY WE NEED MACHINE LEARNING
- WHAT PDFs TEACH US ON MACHINE LEARNING
  - PDF DETERMINATION AS MACHINE LEARNING
  - TESTING UNCERTAINTIES
  - TOWARDS XAI
- WHAT MACHINE LEARNING TEACHES US ON PDFs
  - THE NATURE OF PDF UNCERTAINTIES
  - CORRELATING INFORMATION
  - SERENDIPITOUS DISCOVERY

PROLOGUE

WHAT'S THE PROBLEM WITH PDFs?

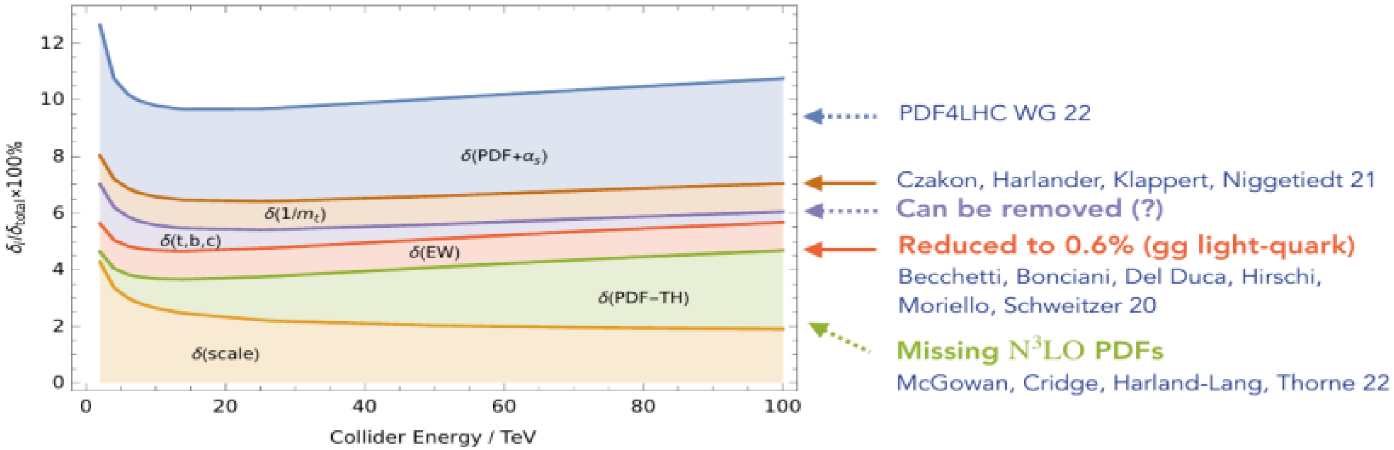
PDFs AND THEIR UNCERTAINTIES

QCD FACTORIZATION



UNCERTAINTIES:

HIGGS IN GLUON FUSION



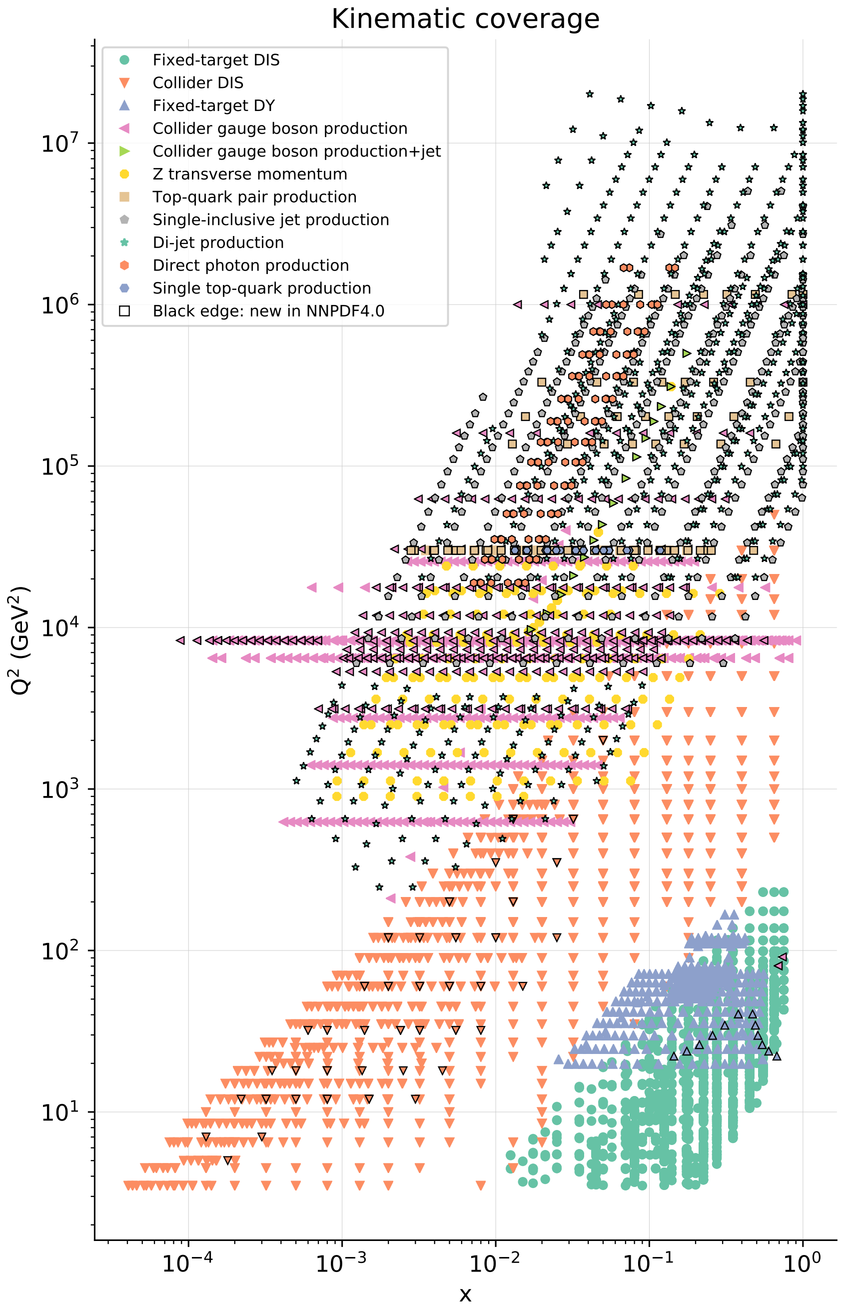
(R. Röntsch, Les Houches 2023)

- FACTORIZED "PROBABILITY" OF A QUARK OR GLUON (PARTONS) TO PARTICIPATE IN HARD INTERACTION
- REQUIRED FOR THE COMPUTATION OF ANY PROCESS AT THE LHC
- DOMINANT SOURCE OF UNCERTAINTY

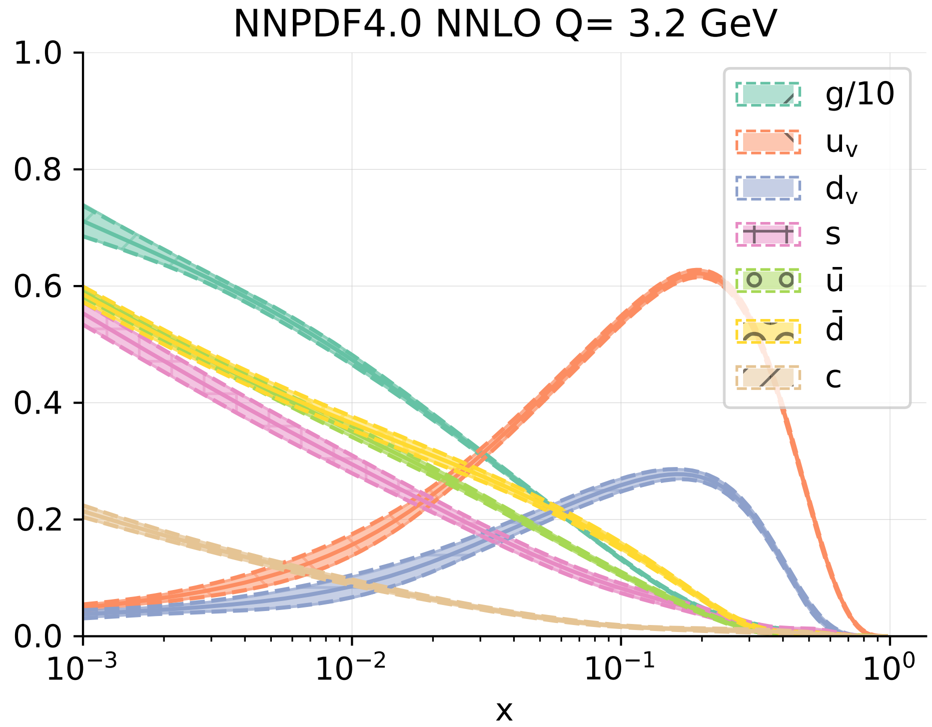
PDFs AND DATA

- LHC CROSS SECTION:
  - $\sigma = \sum_{ij} \hat{\sigma}_{ij} \otimes f_i^{(1)} \otimes f_j^{(2)}$
  - $\hat{\sigma}_{ij}$ PARTONIC CROSS SECTION, INCOMING PARTONS i, j
  - $f_i^{(j)}(x, Q^2)$ PDF FOR PARTON OF SPECIES i IN j-TH INCOMING PROTON
  - $\otimes$ CONVOLUTION OVER x
  - PDF DEPENDS ON Q^2 AND x, OTHER KINEMATIC VARIABLES IN $\hat{\sigma}$
- PARTONIC CROSS SECTION COMPUTED PERTURBATIVELY
- PDFs DETERMINED COMPARING σ TO DATA
  - ABOUT 4600 DATAPOINTS
  - LEPTOPRODUCTION & HADROPRODUCTION, COLLIDER & FIXED-TARGET

THE NNPDF4.0 DATASET



PDFs: THE STATE OF THE ART (NNPDF4.0, 2021)



- A SET OF PROBABILITY DISTRIBUTIONS OF PROBABILITY DISTRIBUTIONS
- FULL (INFINITE DIMENSIONAL) COVARIANCE MATRIX
- MUST BE DETERMINED FROM FINITE SET OF DISCRETE DATA

WHY WE NEED MACHINE LEARNING I

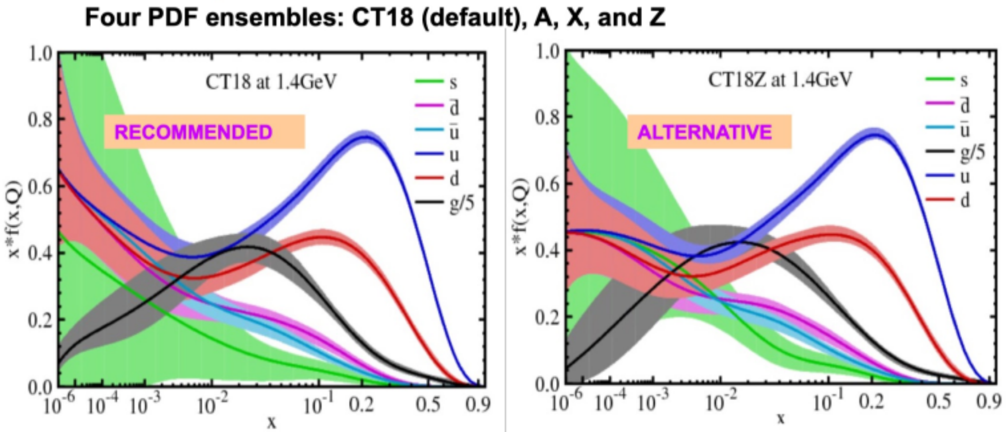
ALTERNATIVE: A MODEL-DEPENDENT APPROACH

PARAMETRIZATIONS

- CTEQ5 2002: $xg(x, Q_0^2) = A_0 x^{A_1}(1-x)^{A_2}(1 + A_3 x^{A_4})$
- MRST-HERALHC 2005: $xg(x, Q_0^2) = A_g x^{\delta_g}(1-x)^{\eta_g}(1 + \epsilon_g x^{0.5} + \gamma_g x) + A_{g'} x^{\delta_{g'}}(1-x)^{\eta_{g'}}$
- CT18: $g(x, Q = Q_0) = x^{a_1 - 1}(1-x)^{a2}\left[a_3(1-y)^3 + a_4 3y(1-y)^2 + a_5 3y^2(1-y) + y^3\right]$;
  $y = \sqrt{x}$; $a_5 = (3 + 2a_1)/3$.

MORE DATA $\Rightarrow$ BIGGER PARAMETRIZATION (?)

PROLIFERATION OF PDF SETS



- The CT18 family of PDFs includes LHC data available up to 2018, i.e. mostly 7 and 8 TeV data
- CT18 is the primary PDF; CT18A includes the ATLAS 7 TeV W/Z data (excluded from CT18 due to very poor fit); CT18X includes scale to simulate effects of low x resummation for DIS; CT18Z includes both effects
- CT18As (new) allows a more flexible parametrization for strange
- CT18As_Lat (new) adds lattice constraint

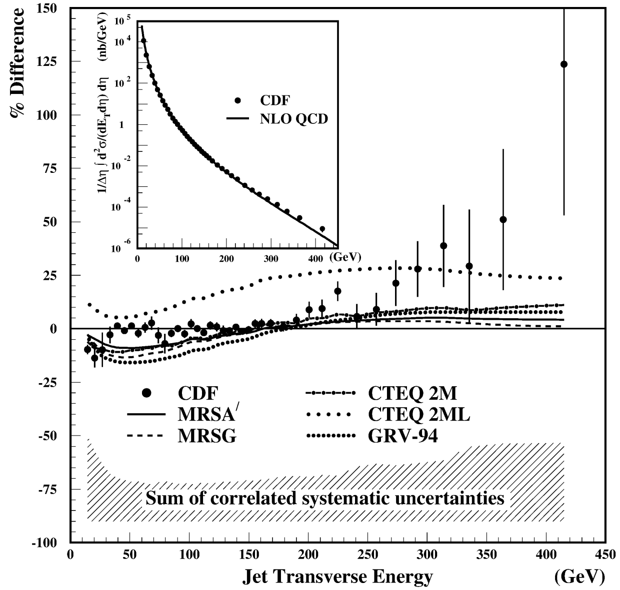
(J. Huston, PDF4LHC 11/2023)

MORE DATA $\Rightarrow$ BIGGER UNCERTAINTIES (!)

WHY WE NEED MACHINE LEARNING II

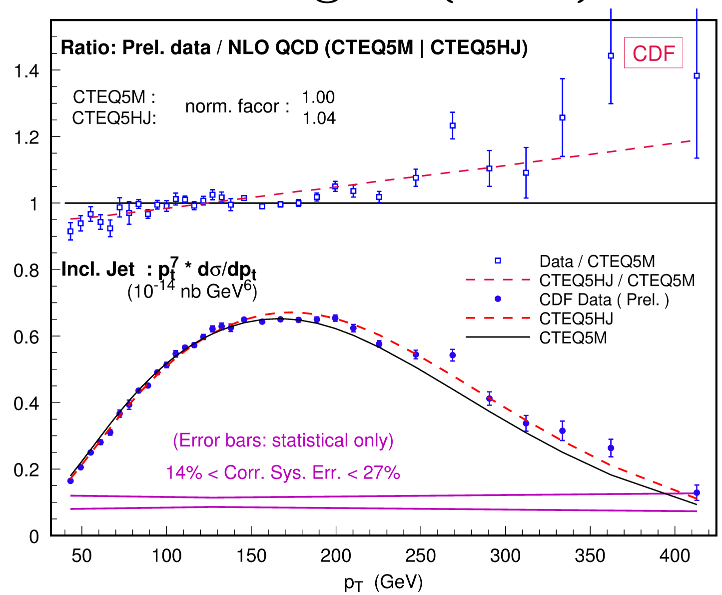
DISCOVERY PHYSICS 1995

- DISCREPANCY BETWEEN QCD CALCULATION AND CDF JET DATA (1995)
- ~~EVIDENCE FOR QUARK COMPOSITENESS~~
- NO INFO ON PARTON UNCERTAINTY $\Rightarrow$ RESULT STRONGLY DEPENDS ON GLUON AT $x \gtrsim 0.1$

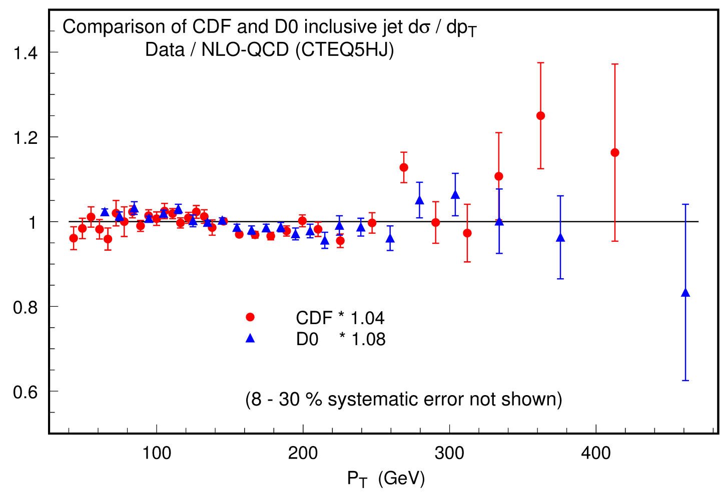


DISCREPANCY REMOVED IF JET DATA INCLUDED IN THE FIT

NEW CTEQ FIT (1996)



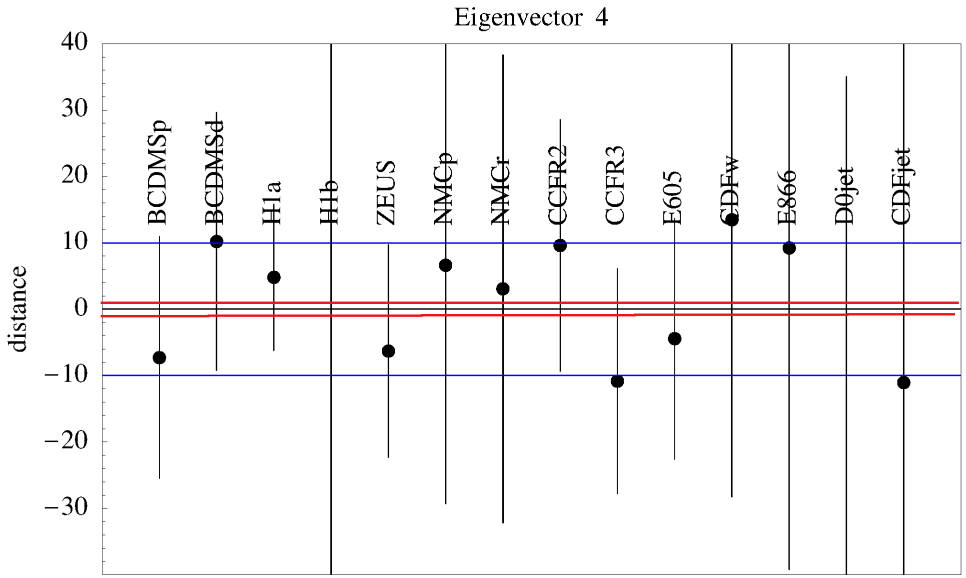
FINAL CTEQ FIT (1998)



WHY WE NEED MACHINE LEARNING III
“TOLERANCE”

FIRST PDFs WITH UNCERTAINTIES (2002)

one sigma & ten sigma intervals for typical covariance matrix eigenvalue vs best value and uncertainty from individual experiments



MSHT PDFS (2020)

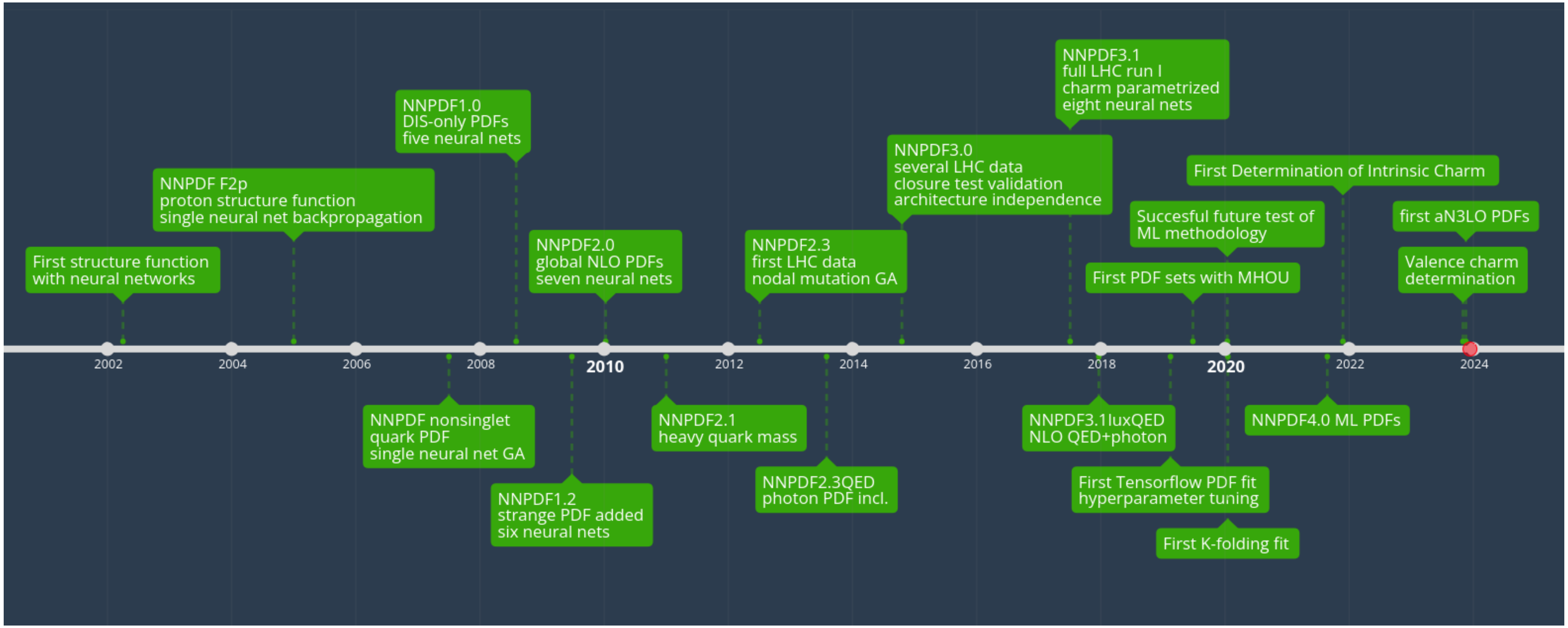
e− vector	+ t	+ T	Most constraining data set	− t
1	3.71	3.75	ATLAS 7 TeV high prec. W,Z	4.76
2	3.12	3.33	NuTeV $\nu N \to \mu\mu X$	2.85
3	2.48	2.58	NuTeV $\nu N \to \mu\mu X$	4.07
4	3.61	3.60	CMS 8 TeV W	2.93
5	2.64	3.00	ATLAS 7 TeV high prec. W,Z	2.72
6	5.22	5.46	ATLAS 8 TeV double dif Z	5.01
7	4.07	4.37	NMC/... F_L	2.90
8	3.90	3.50	LHCb 2015 W,Z	3.90
9	5.48	5.59	LHCb 2015 W,Z	3.73
10	3.55	3.58	BCDMS μp F_2	4.87
11	3.06	2.91	DØ W asym.	4.83
12	1.42	1.71	DØ W asym.	3.40
13	3.87	4.10	CMS asym. $p_T > 25, 30$ GeV	4.38
14	1.36	1.50	E866/NuSea pd/pp DY	3.67
15	5.53	5.89	E866/NuSea pd/pp DY	3.17
16	1.89	0.52	E866/NuSea pd/pp DY	5.64
17	2.51	2.54	E866/NuSea pd/pp DY	2.69
18	1.80	1.88	DØ W asym.	2.47
19	2.47	2.18	CMS 8 TeV W	1.37
20	1.82	2.22	DØ W asym.	4.69
21	4.41	5.36	ATLAS 8 TeV Z p_T	4.68
22	3.49	3.23	DØ W asym.	3.04
23	1.84	2.43	ATLAS 8TeV sing dif $t\bar{t}$ dilep	4.96
24	0.99	1.23	E866/NuSea pd/pp DY	4.61
25	2.01	1.35	DØ W asym.	2.77
26	2.25	2.51	NuTeV νN xF_3	2.06
27	2.83	3.65	ATLAS 8 TeV $t\bar{t}$, dilepton	2.64
28	1.74	1.92	DØ W asym.	2.65
29	2.57	2.85	CMS 7 TeV $W + c$	1.79
30	4.76	3.92	CCFR $\nu N \to \mu\mu X$	2.25
31	2.79	4.81	ATLAS 7TeV high prec W,Z	2.07
32	2.57	4.27	CCFR $\nu N \to \mu\mu X$	2.58

- PDF UNCERTAINTIES RESCALED BY “TOLERANCE” $T \sim 4 \div 10$
- DETERMINED FROM SPREAD OF BEST-FIT FROM DIFFERENT DATA

ACT I
PDF lessons on ML

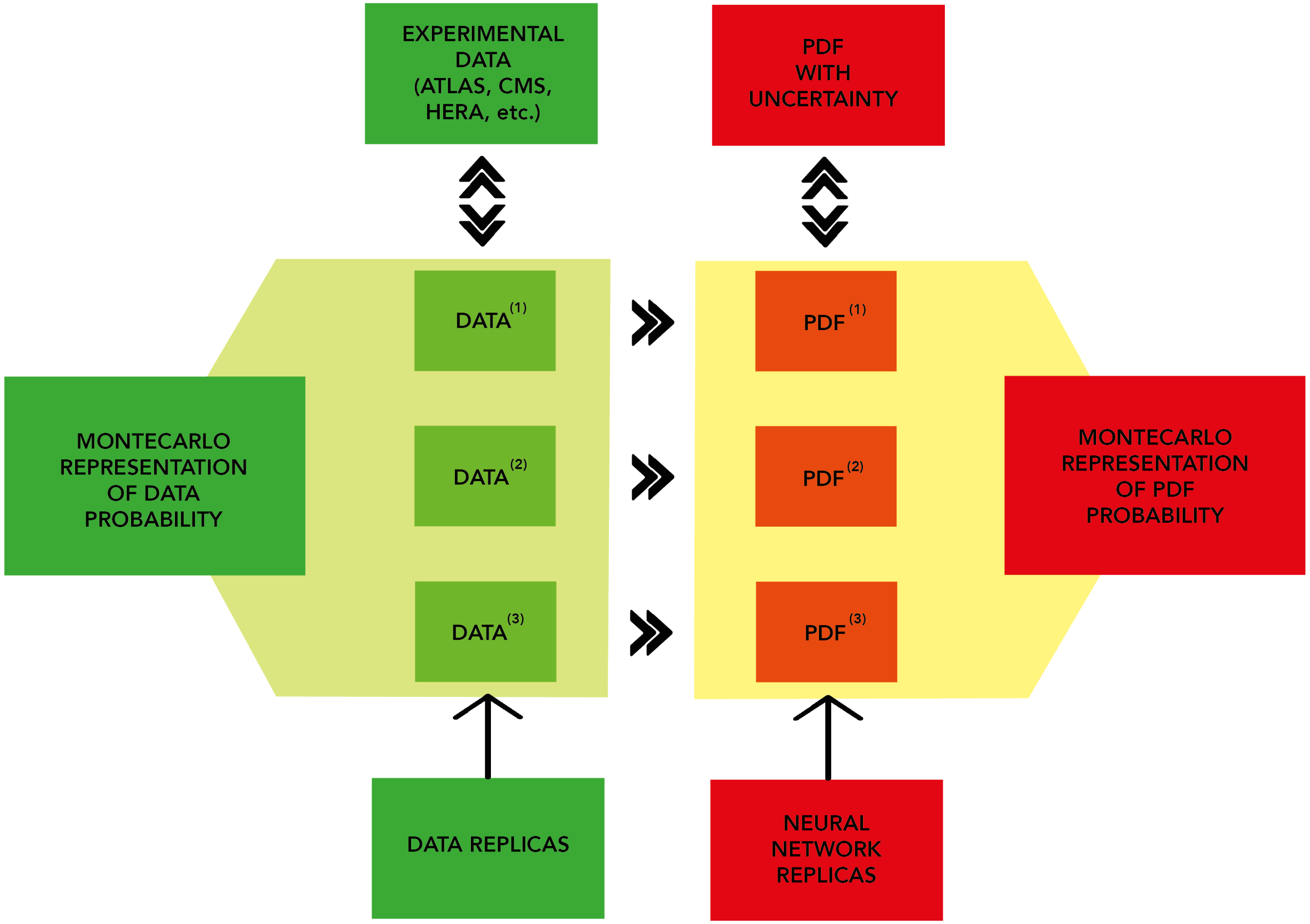
PROTON STRUCTURE AS A ML PROBLEM

NNPDF



PROBABILITY REGRESSION

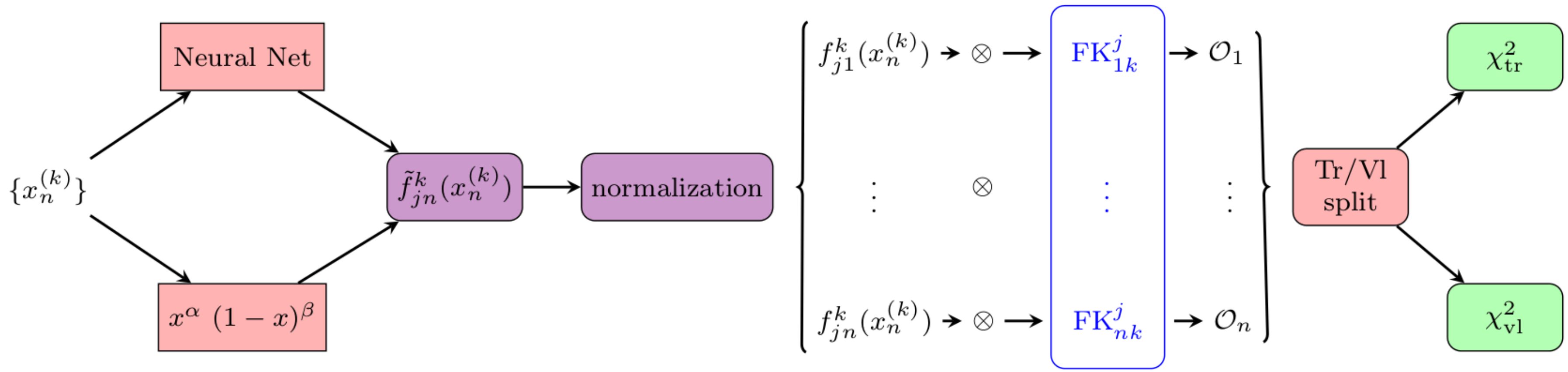
REPLICA SAMPLE OF FUNCTIONS $\Leftrightarrow$ PROBABILITY DENSITY IN FUNCTION SPACE
KNOWLEDGE OF LIKELIHHOD SHAPE (FUNCTIONAL FORM) NOT NECESSARY

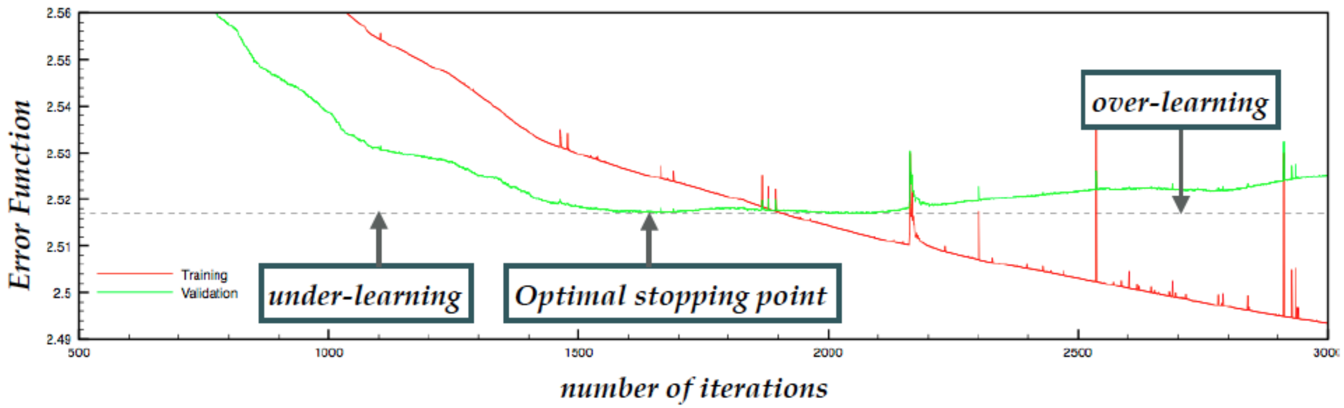


FINAL PDF SET: $f_i^{(a)}(x,\mu)$;
i =up, antiup, down, antidown, strange, antistrange, charm, gluon; $j = 1, 2, \ldots N_{\rm rep}$

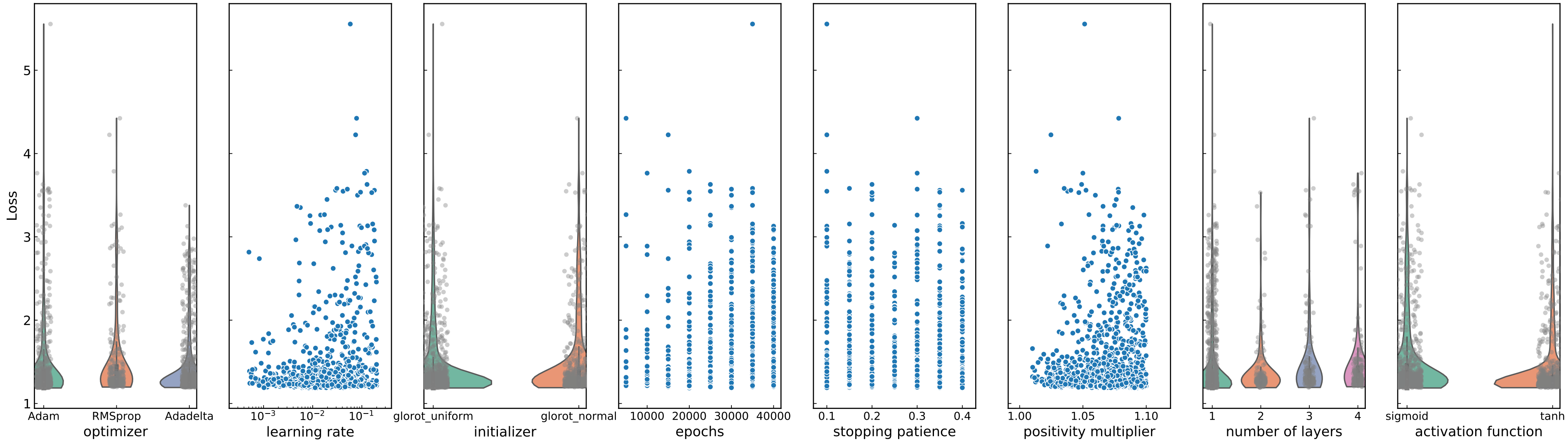
CROSS-VALIDATED LEARNING

- NEURAL NET PARAMETERS DETERMINED BY χ^2 MINIMIZATION THROUGH GRADIENT DESCENT
- RANDOM TRAINING-VALIDATION SPLIT, χ^2 TO TRAINING DATA REPLICAS MINIMIZED
- TRAINING STOPS IF VALIDATION χ^2 GROWS FOR A WHILE (PATIENCE)
- LOWEST VALIDATION χ^2 $\Rightarrow$ OPTIMAL FIT





METHODOLOGY HYPEROPTIMIZATION

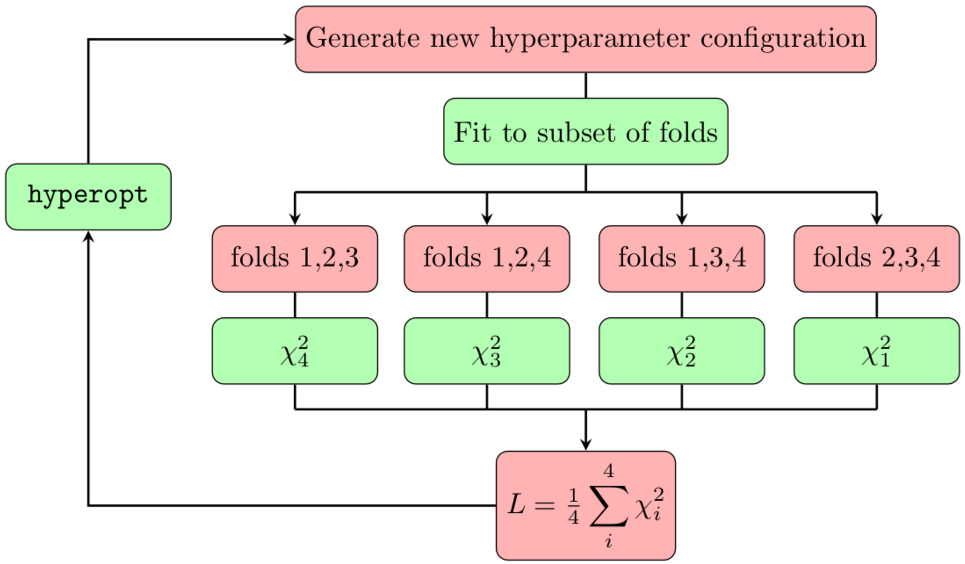


HYPEROPT PARAMETERS

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

K-FOLD OPTIMIZATION



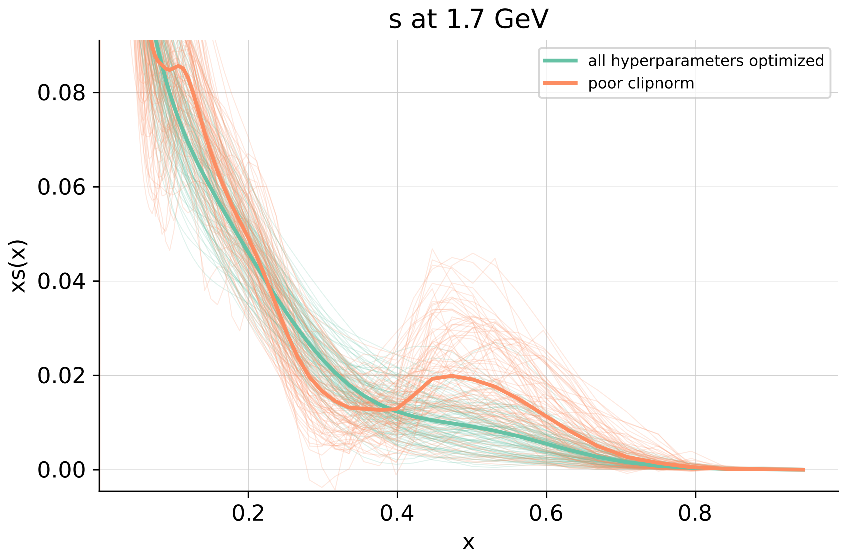
Fold 1		
CHORUS σ^{ν}_{CC}	HERA I+II inc NC e^+p 920 GeV	BCDMS p
LHCb Z 940 pb	ATLAS W, Z 7 TeV 2010	CMS Z p_T 8 TeV (p_T^{ll}, y_{ll})
DY E605 $\sigma^p_{\rm DY}$	CMS Drell-Yan 2D 7 TeV 2011	CMS 3D dijets 8 TeV
ATLAS single-$\bar{t}$ y (normalised)	ATLAS single top R_t 7 TeV	CMS $t\bar{t}$ rapidity $y_{t\bar{t}}$
CMS single top R_t 8 TeV		

Fold 2		
HERA I+II inc CC e^-p	HERA I+II inc NC e^+p 460 GeV	HERA comb. $\sigma^{\rm red}_{b\bar{b}}$
NMC p	NuTeV $\sigma^{\bar{\nu}}_c$	LHCb $Z \to ee$ 2 fb
CMS W asymmetry 840 pb	ATLAS Z p_T 8 TeV (p_T^{ll}, M_{ll})	D0 $W \to \mu\nu$ asymmetry
DY E886 $\sigma^p_{\rm DY}$	ATLAS direct photon 13 TeV	ATLAS dijets 7 TeV, R=0.6
ATLAS single antitop y (normalised)	CMS $\sigma^{\rm tot}_{tt}$	CMS single top $\sigma_t + \sigma_{\bar{t}}$ 7 TeV

Fold 3		
HERA I+II inc CC e^+p	HERA I+II inc NC e^+p 575 GeV	NMC d/p
NuTeV σ^{ν}_c	LHCb $W, Z \to \mu$ 7 TeV	LHCb $Z \to ee$
ATLAS W, Z 7 TeV 2011 Central selection	ATLAS W^++jet 8 TeV	ATLAS HM DY 7 TeV
CMS W asymmetry 4.7 fb	DYE 866 $\sigma^d_{\rm DY}/\sigma^p_{\rm DY}$	CDF Z rapidity (new)
ATLAS $\sigma^{\rm tot}_{tt}$	ATLAS single top y_t (normalised)	CMS $\sigma^{\rm tot}_{tt}$ 5 TeV
CMS $t\bar{t}$ double diff. $(m_{t\bar{t}}, y_t)$		

Fold 4		
CHORUS $\sigma^{\bar{\nu}}_{CC}$	HERA I+II inc NC e^+p 820 GeV	LHCb $W, Z \to \mu$ 8 TeV
LHCb $Z \to \mu\mu$	ATLAS W, Z 7 TeV 2011 Fwd	ATLAS W^-+jet 8 TeV
ATLAS low-mass DY 2011	ATLAS Z p_T 8 TeV (p_T^{ll}, y_{ll})	CMS W rapidity 8 TeV
D0 Z rapidity	CMS dijets 7 TeV	ATLAS single top y_t (normalised)
ATLAS single top R_t 13 TeV	CMS single top R_t 13 TeV	

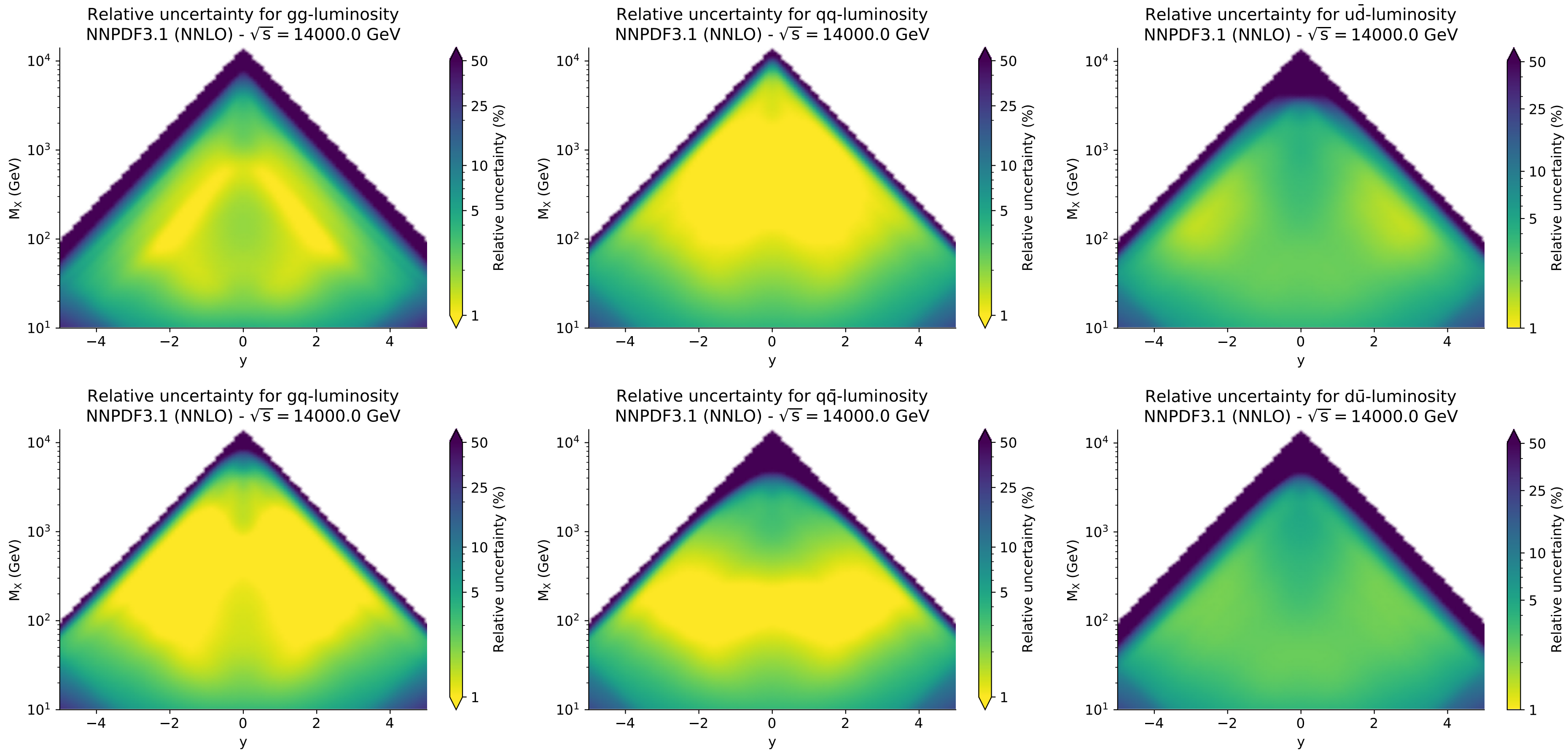
K-FOLDING VS NO K-FOLDING



- EACH FOLD REPRODUCES FEATURES OF FULL DATASET
- LOSS: AVERAGE χ^2 OF NON-FITTED FOLDS
- OVERFITTING REMOVED $\Rightarrow$ CORRECT GENERALIZATION

UNCERTAINTIES 2016

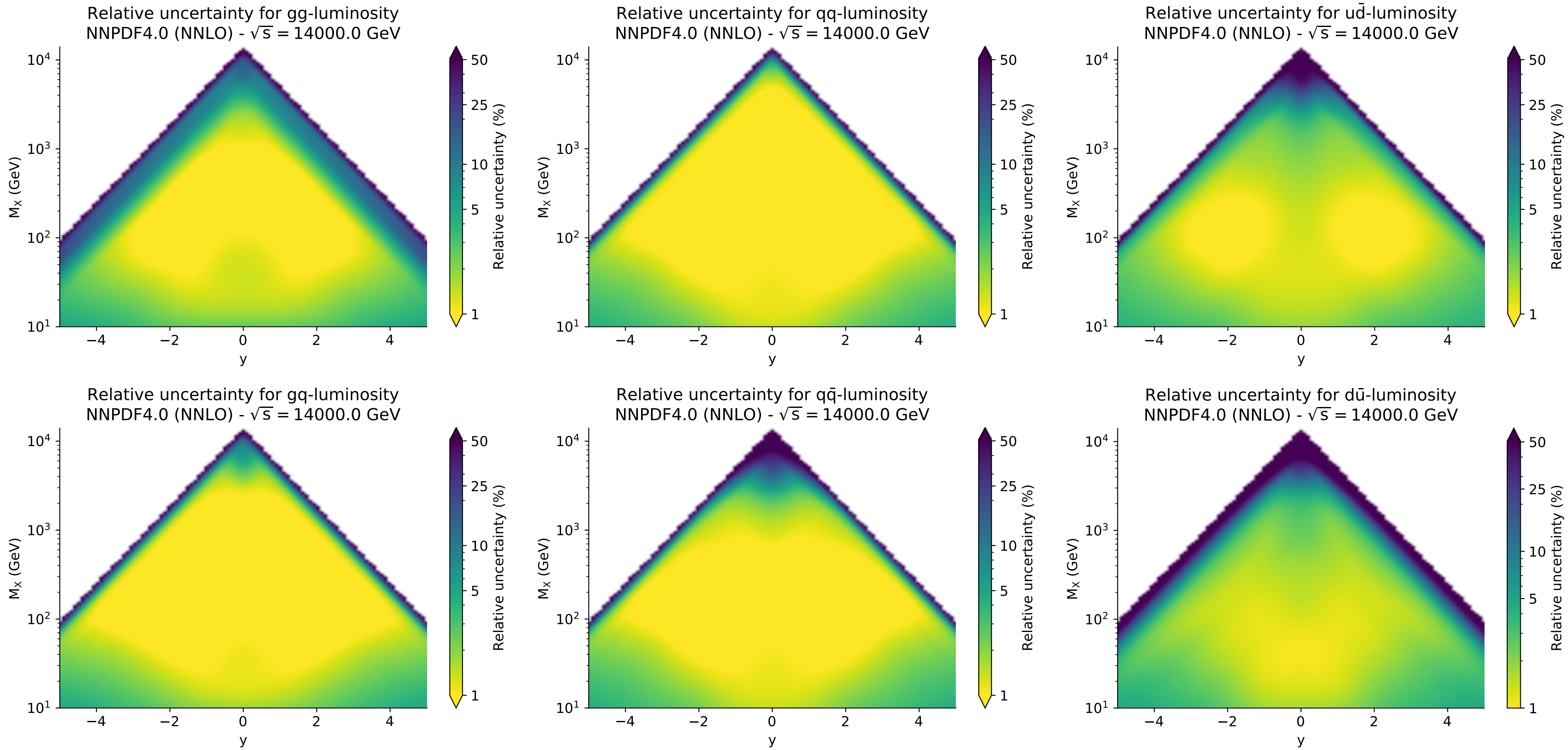
GLUON SINGLET FLAVORS



- TYPICAL UNCERTAINTIES IN DATA REGION: SINGLET $\sim 3\%$, NONSINGLET $\sim 5\%$
- DATA REGION: $10^2 \lesssim M_X \lesssim 10^3$ TeV, $-2 \lesssim y \lesssim 2$

UNCERTAINTIES 2022

GLUON SINGLET FLAVORS



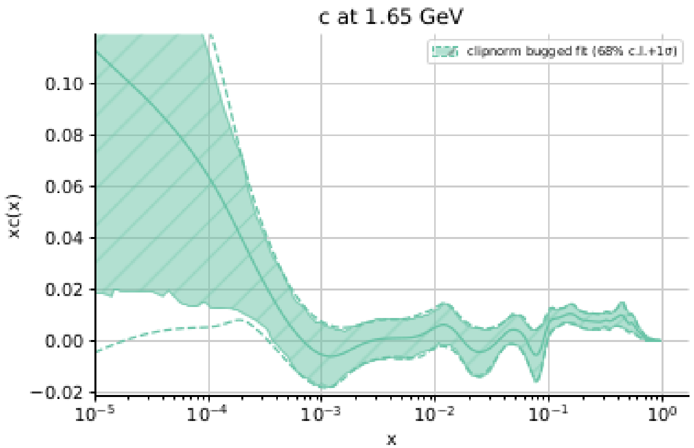
- TYPICAL UNCERTAINTIES IN DATA REGION: SINGLET $\sim 1\%$, NONSINGLET $\sim 2-3\%$
- DATA REGION: $10 \lesssim M_X \lesssim 3 \cdot 10^3$ TeV, $-4 \lesssim y \lesssim 4$

VALIDATING UNCERTAINTIES I: OVERFITTING METRIC

- RECOMPUTE VALIDATION $\chi^2_{\mathrm{val}'}$ FOR ALL DATA REPLICAS
  - KEEPING SAME TRAINING-VALIDATION SPLIT
  - BUT DIFFERENT FLUCTUATED VALIDATION DATA
- COMPUTE AVERAGE OVER REPLICAS $\langle\chi^2_{\mathrm{val}'}\rangle$ &
  DETERMINE DIFFERENCE TO STANDARD VALIDATION χ^2_{val}
  OVERFITNESS: $\mathcal{R}_O = \chi^2_{\mathrm{val}} - \langle\chi^2_{\mathrm{val}'}\rangle$
- NEGATIVE OVERFITNESS $\mathcal{R}_O \Rightarrow$ OVERFIT

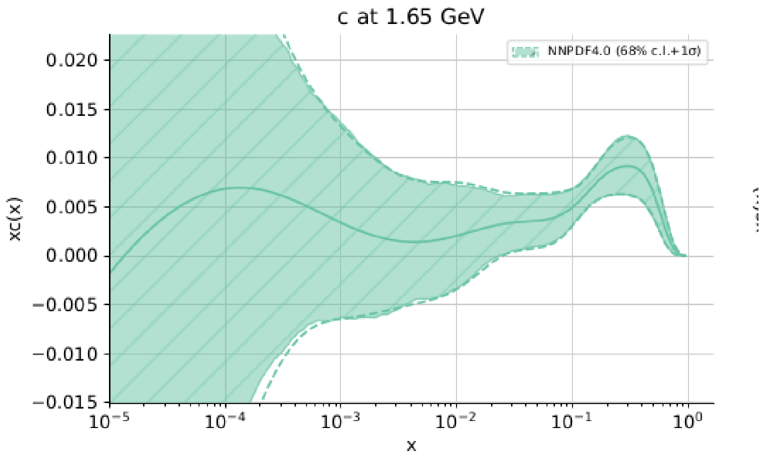
CHARM PDF

OVERFIT (NO CLIPNORM)



$\mathcal{R}_O = -0.024 \pm 0.012$

PROPER FIT (NNPDF4.0)



$\mathcal{R}_O = -0.001 \pm 0.013$

VALIDATING UNCERTAINTIES II:
CLOSURE TESTS

FAITHFUL UNCERTAINTIES IN DATA REGION?

- ASSUME "TRUE" UNDERLYING PDF ⇒ E.G. SOME RANDOM PDF REPLICA
- GENERATE DATA DISTRIBUTED ACCORDING TO EXPERIMENTAL COVARIANCE MATRIX
- RUN WHOLE METHDOLOGY ON THESE DATA
- DO STATISTICS ON "RUNS OF THE UNIVERSE", POSSIBLE THANKS TO EFFICIENT METHDOLOGY: COMPARE TO TRUE VALUES OF OBSERVABLES (NOT FITTED)
  - BIAS/VARIANCE: MEAN SQUARE DEVIATION WR TO TRUTH VS UNCERTAINTY
  - $\xi^{(\mathrm{pdf})}_{1\sigma}$: IS TRUTH WITHIN ONE SIGMA 68% OF TIMES?
  - $\mathrm{erf}(R_{bv}/\sqrt{2})$: IS THE ¿UE ONE-SIGMA GAUSSIAN QUANTILE 68%?

CLOSURE TEST RESULTS:

NUMBERS

BIAS/VARIANCE RATIO AND ONE-σ QUANTILE

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

Dataset	$\sqrt{\mathrm{b/v}}$	$\xi^{(\mathrm{data})}_{1\sigma}$	$\mathrm{erf}(R_{bv}/\sqrt{2})$
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	$\xi^{(\mathrm{pdf})}_{1\sigma}$
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

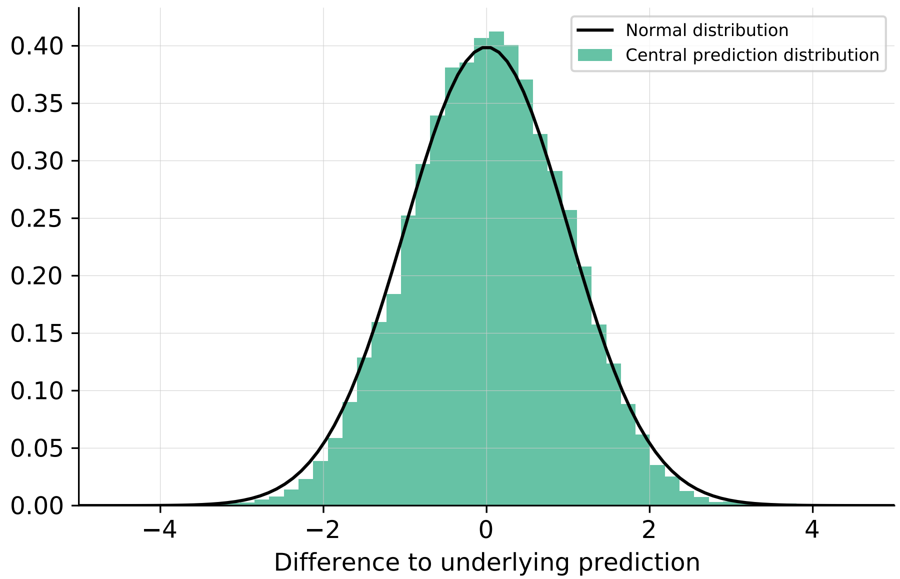
- 25 "UNIVERSE RUNS", 45 REPLICAS EACH
- IN-SAMPLE DATA: PRE 2015
- OUT OF SAMPLE DATA: 2015-2020, MOSTLY LHC
- PDFS HIGHLY CORRELATED $\Rightarrow$ SAMPLED AT 4 POINTS EACH

CLOSURE TEST RESULTS:

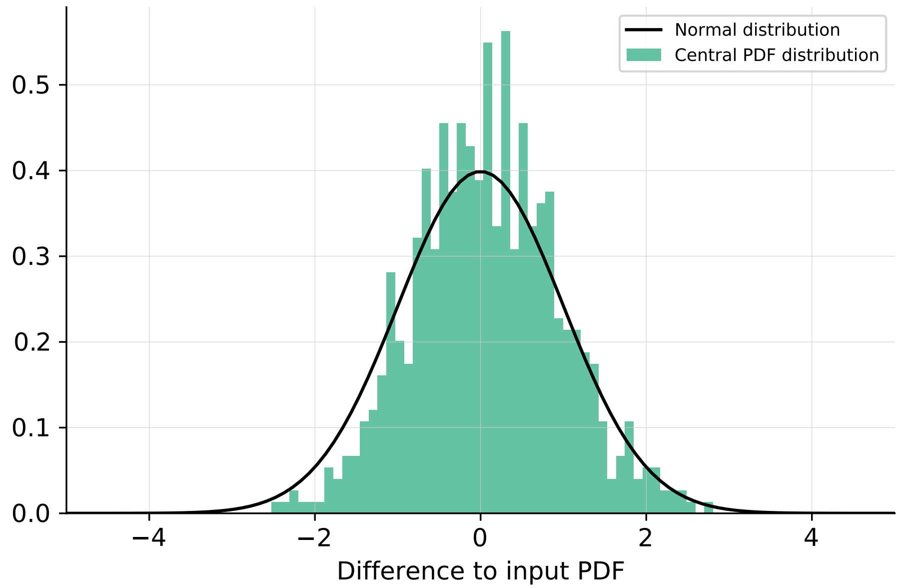
PICTURES

DISTRIBUTION OF DEVIATIONS FROM TRUTH

DATA SPACE (OUT OF SAMPLE)



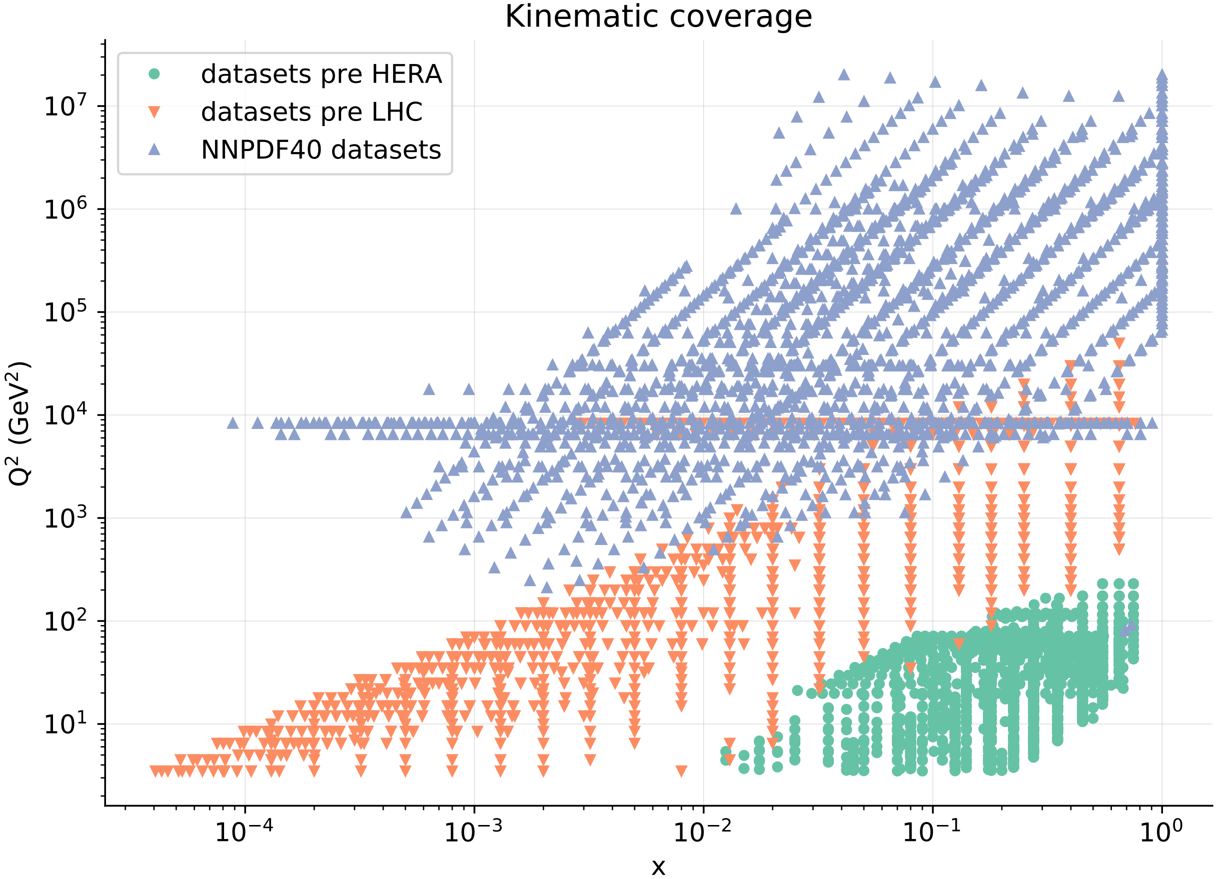
PDF SPACE



- PDF-SPACE MORE NOISY THAN DATA SPACE

VALIDATING UNCERTAINTIES III: FUTURE TESTS

HOW DO WE TEST UNCERTAINTIES IN EXTRAPOLATION?



- DEFINE "PRE-HERA", " PRE-LHC" AND "CURRENT" DATASETS
  EACH LATER DATASET IS EXTRAPOLATION OF PREVIOUS
- DETERMINE PDFS & COMPARE TO "FUTURE" DATA
- COMPUTE χ^2 TO FUTURE DATA:
  - WITHOUT PDF UNCERTAINTIES $\Rightarrow$ IF $\gg 1$, MISSING INFORMATION
  - WITH PDF UNCERTAINTY $\Rightarrow$ IF ~ 1, TEST PASSED
    MISSING INFO REPRODUCED BY UNCERTAINTY

ASSESSING EXTRAPOLATION UNCERTAINTIES

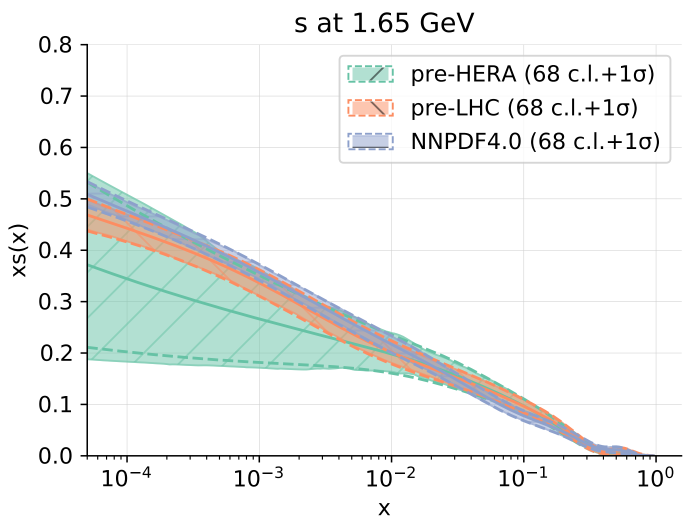
FUTURE TEST

RESULTS (NNPDF4.0)

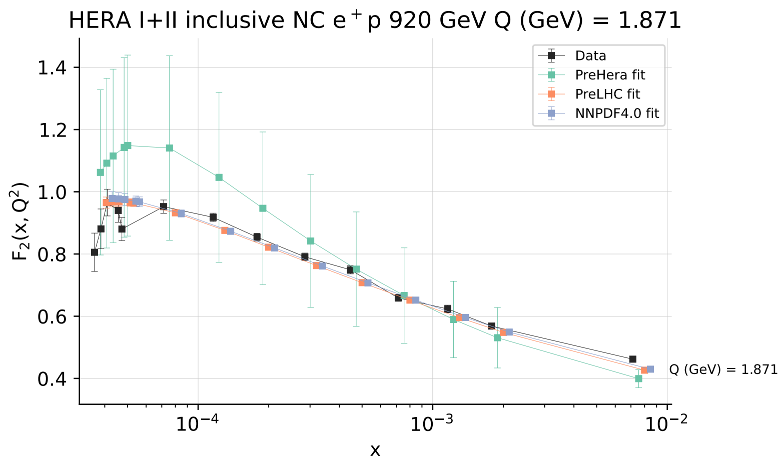
χ^2: FITTED VS EXTRAPOLATED: WITHOUT/WITH PDF UNC.

PROCESS	PRE-HERA	PRE-LHC	NNPDF4.0
FT DIS (NC)	1.05	1.18	1.23
FT DIS (CC)	0.80	0.85	0.87
FT DY	0.92	1.27	1.59
HERA	27.20/1.23	1.22	1.20
COLL. DY (TEV.)	5.52/1.02	0.99	1.11
COLL. DY (LHC)	18.91/1.31	2.63/1.58	1.53
TOP QUARK	20.01/1.06	1.30/0.87	1.01
JETS	2.69/0.98	2.12/1.10	1.26
TOTAL OUT OF SAMPLE	19.48/1.16	2.10/1.15	–

strange PDF



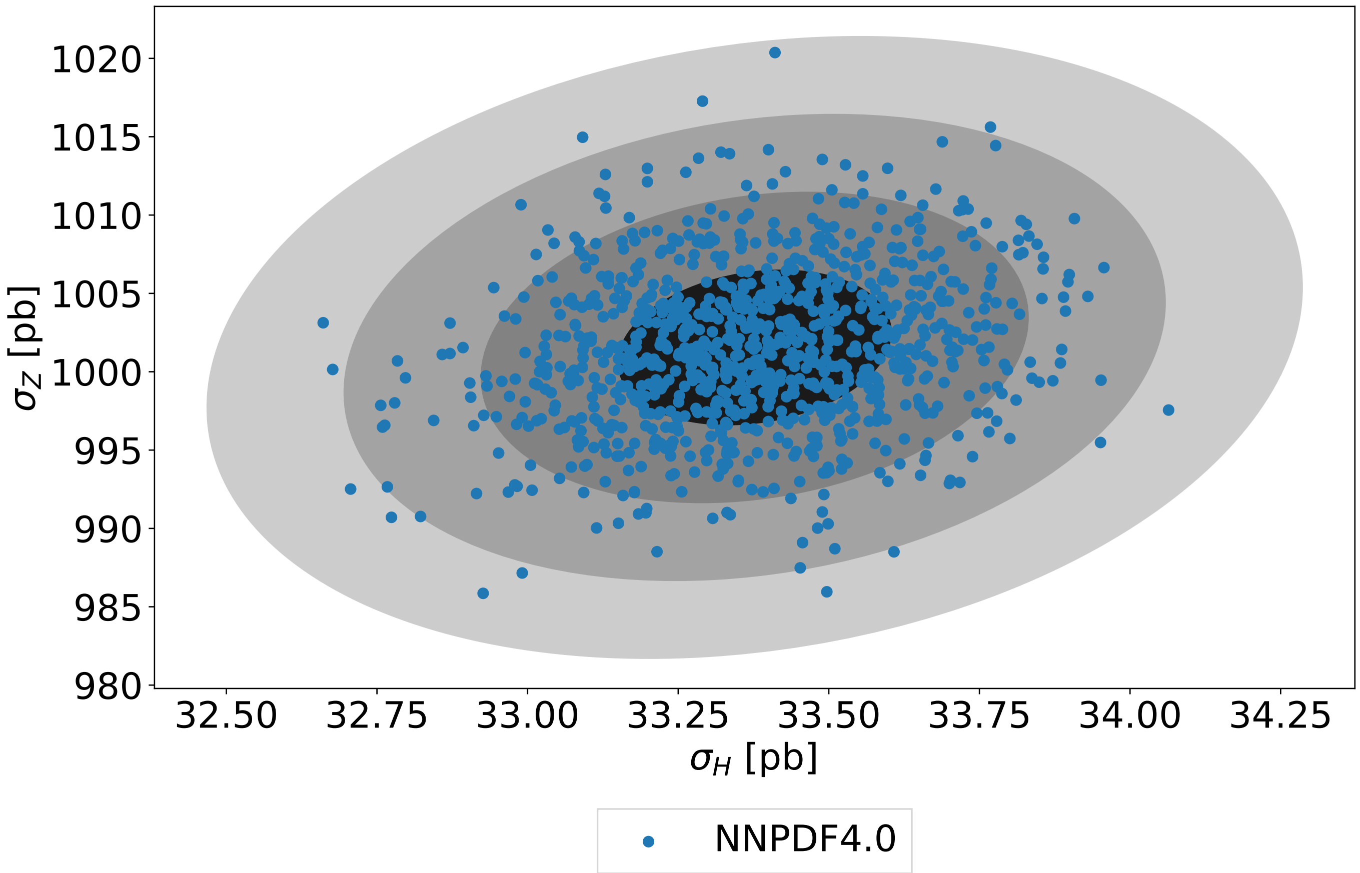
HERA F_2^p



PDF UNCERTAINTIES DO ACCOUNT FOR EXTRAPOLATION UNCERTAINTIES

UNDERSTANDING UNCERTAINTIES
THE REPLICA DISTRIBUTION

- PLOT RESULTS IN (σ_H, σ_Z) PREDICTION SPACE $\Rightarrow$ GAUSSIAN!
- DISTRIBUTION OF REPLICAS $\Rightarrow$ OPTIMAL IMPORTANCE SAMPLING

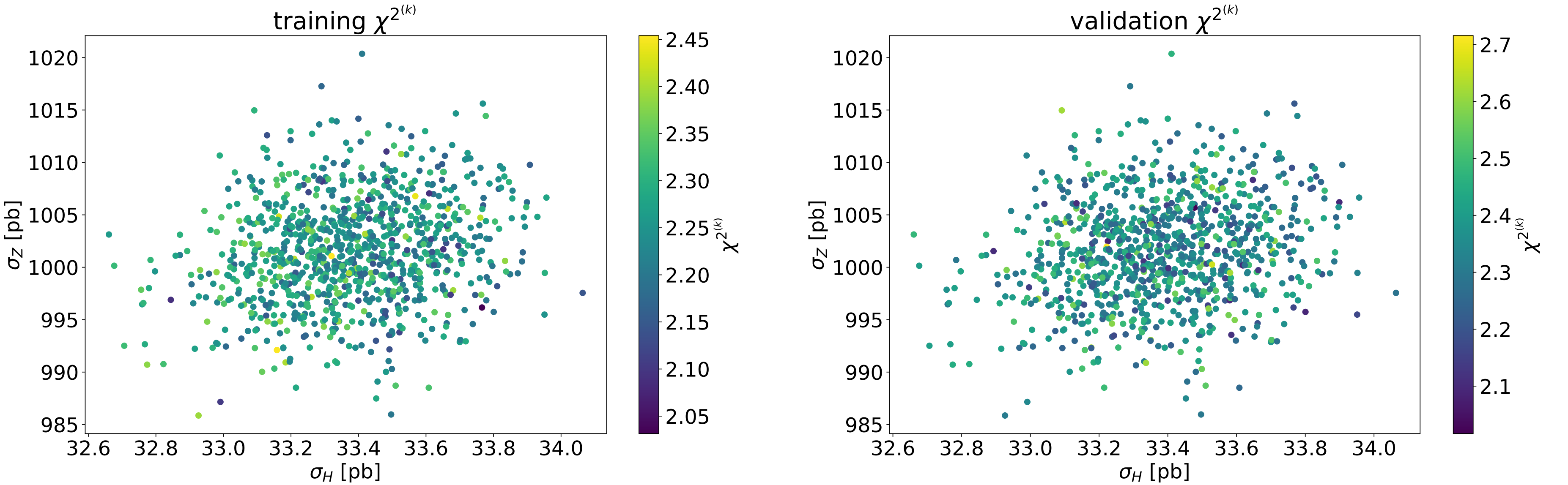


DISTRIBUTION OF REPLICAS DRIVEN BY

- DATA UNCERTAINTIES $\Rightarrow$ DATA REPLICA FLUCTUATION
- INTERPOLATION, EXTRAPOLATION AND FUNCTIONAL UNCERTAINTIES $\Rightarrow$ BEST FIT DEGENERACY

THE REPLICA DISTRIBUTION

ARE ALL FITS EQUALLY GOOD?



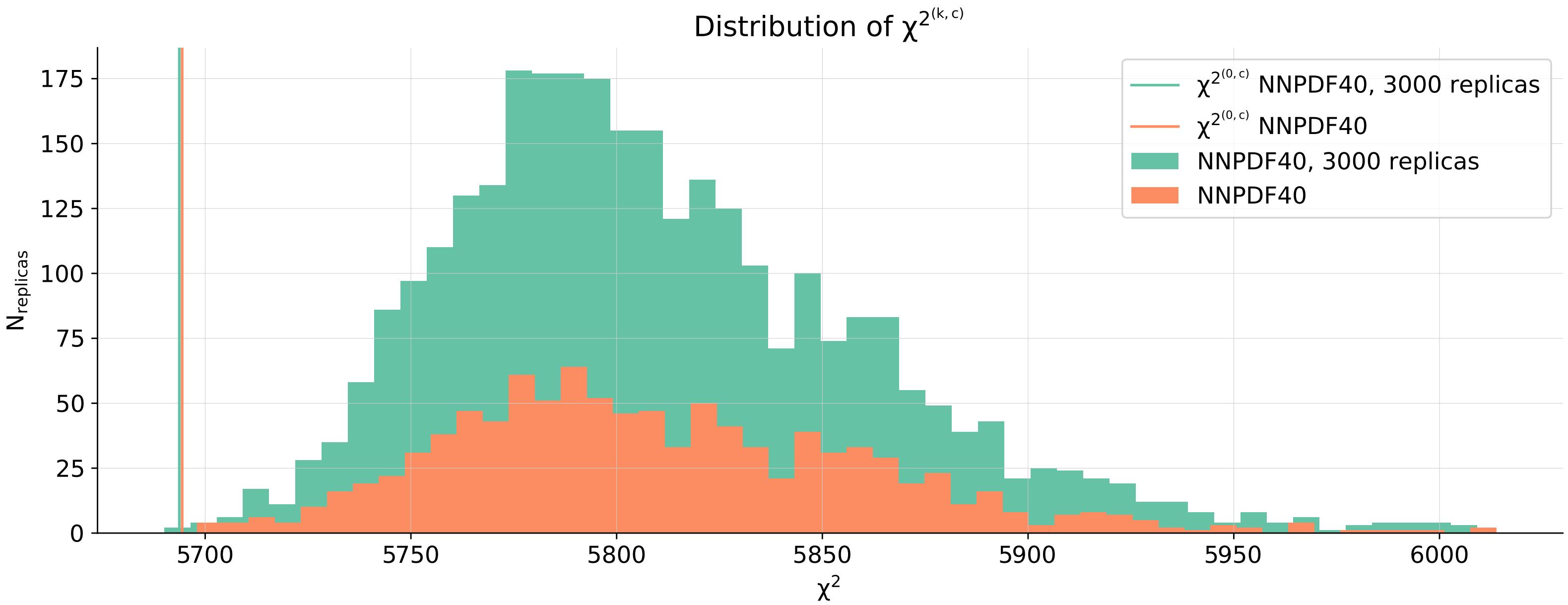
- COMPARE TRAINING AND VALIDATION χ^2 FOR EACH REPLICA
- NO CORRELATION BETWEEN FIT QUALITY AND POSITION IN THE (σ_H, σ_Z) PLANE
- UNIFORM FIT QUALITY

THE REPLICA DISTRIBUTION

COMPARISON TO CENTRAL DATA

- EACH PDF REPLICA FITTED TO A DATA REPLICA
- FIT QUALITY TO CENTRAL DATA STATISTICALLY DISTRIBUTED

1000 REPLICAS VS. 3000 REPLICAS

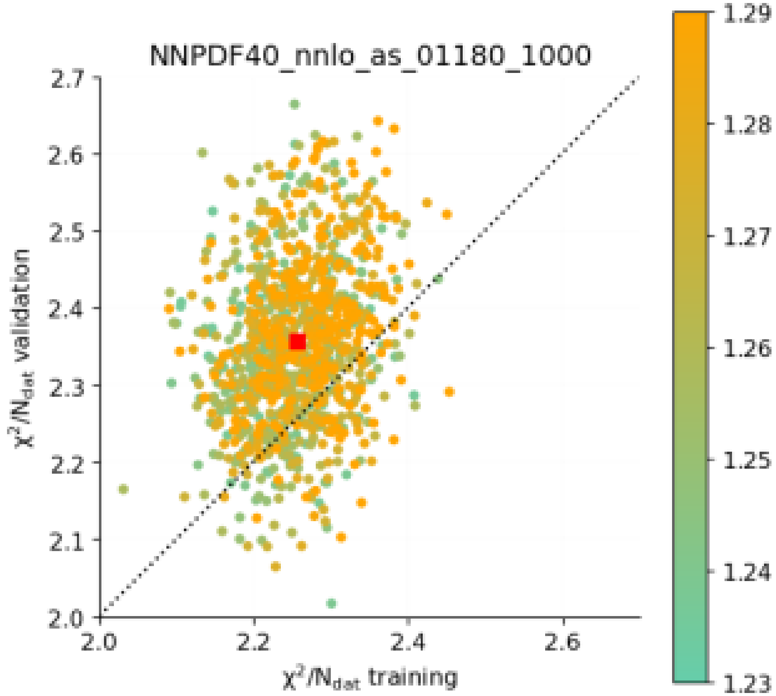


- AVERAGE BEST FIT PDF $\Rightarrow$ LOW χ^2
- NOT NECESSARILY LOWEST

THE REPLICA DISTRIBUTION

COMPARISON TO CENTRAL DATA

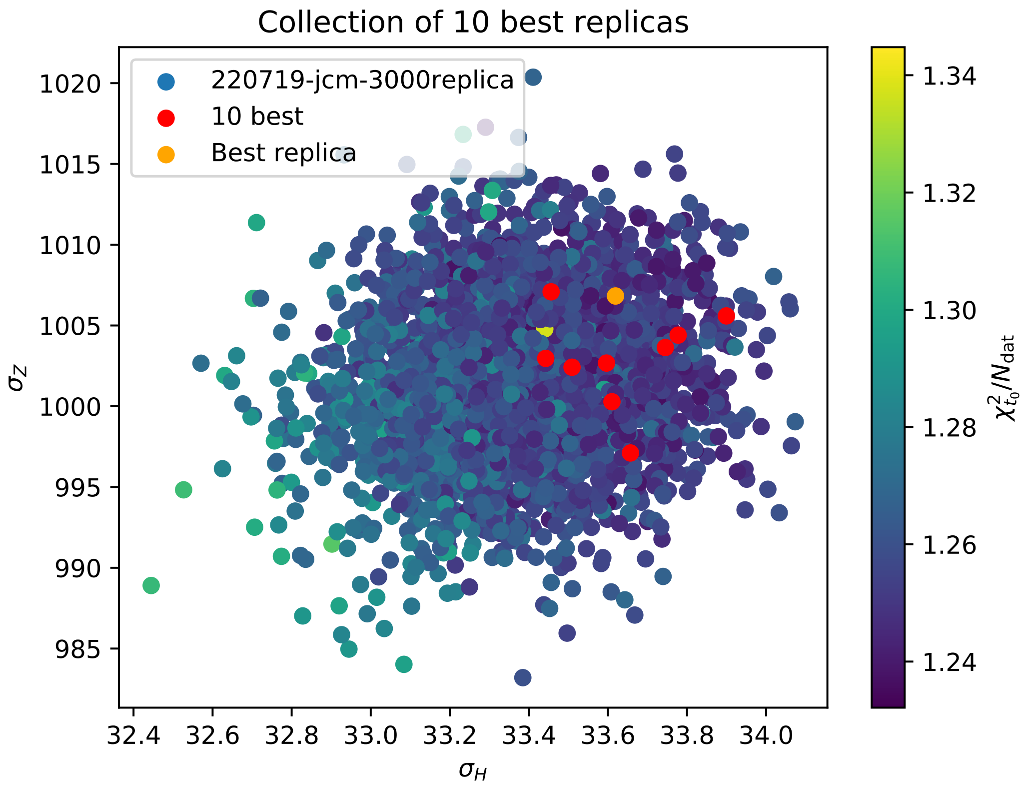
- ARE FITS WITH HIGH χ^2 TO CENTRAL DATA POOR (UNDERLEARNT)?



- NO CORRELATION BETWEEN χ^2 TO CENTRAL DATA AND TRAINING, VALIDATION χ^2
- UNIFORM FIT QUALITY
- DISPERSION DUE
  - DATA REPLICA FLUCTUATION $\Rightarrow$ DATA UNCERTAINTIES
  - BEST FIT DEGENERACY
    $\Rightarrow$ INTERPOLATION, EXTRAPOLATION AND FUNCTIONAL UNCERTAINTIES

REPLICA LOSS DISTRIBUTION

CORRELATION TO FEATURES



χ^2 TO CENTRAL DATA

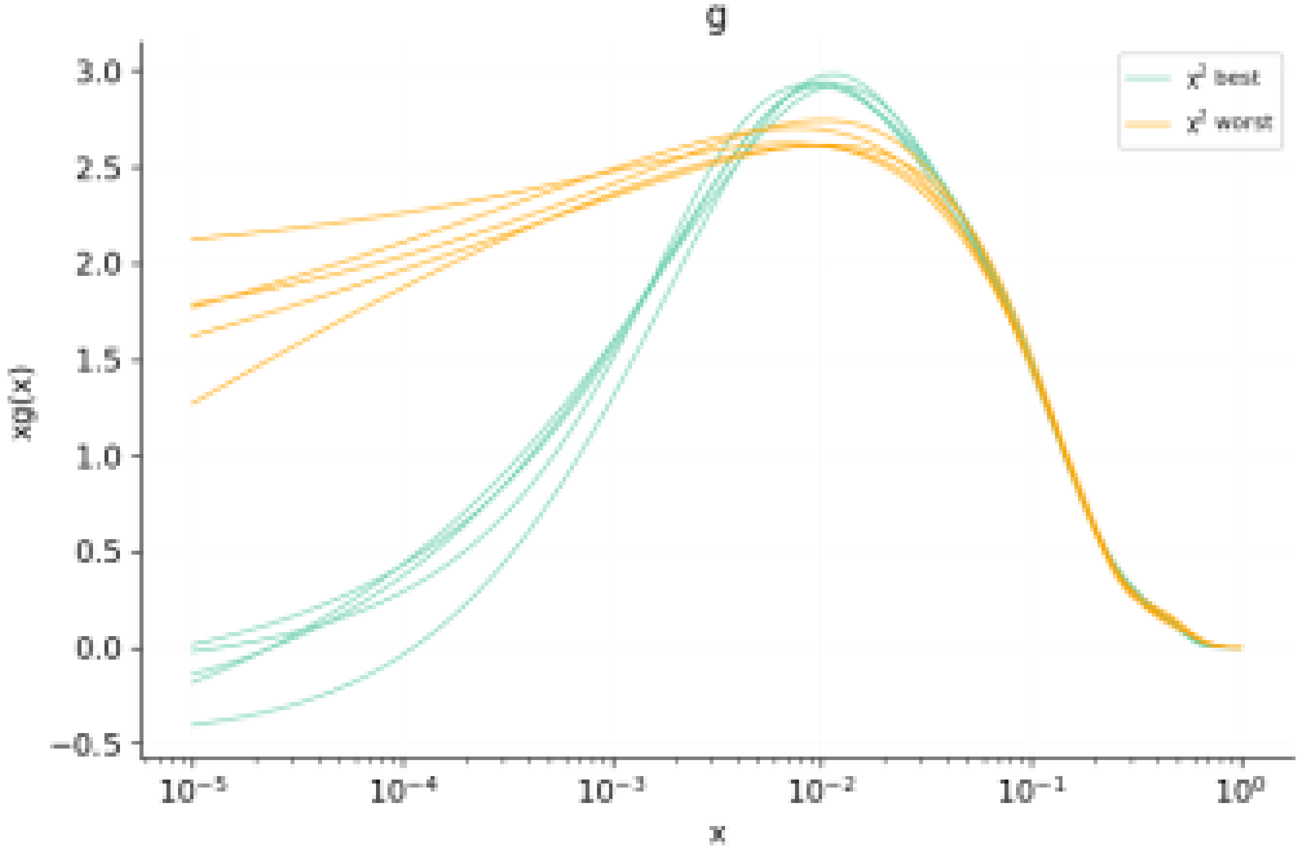
- CORRELATED TO POSITION IN (σ_H, σ_z) PLANE
- CORRELATED TO A FEATURE?

EXPLANATION

LOOKING FOR FEATURES

REPLICAS WITH LOWEST & HIGHEST χ^2 TO CENTRAL DATA

THE GLUON



- REPLICAS CLOSER TO CENTRAL DATA ⇒ MORE STRUCTURE
- CORRELATED TO A FEATURE?

EXPLANATION

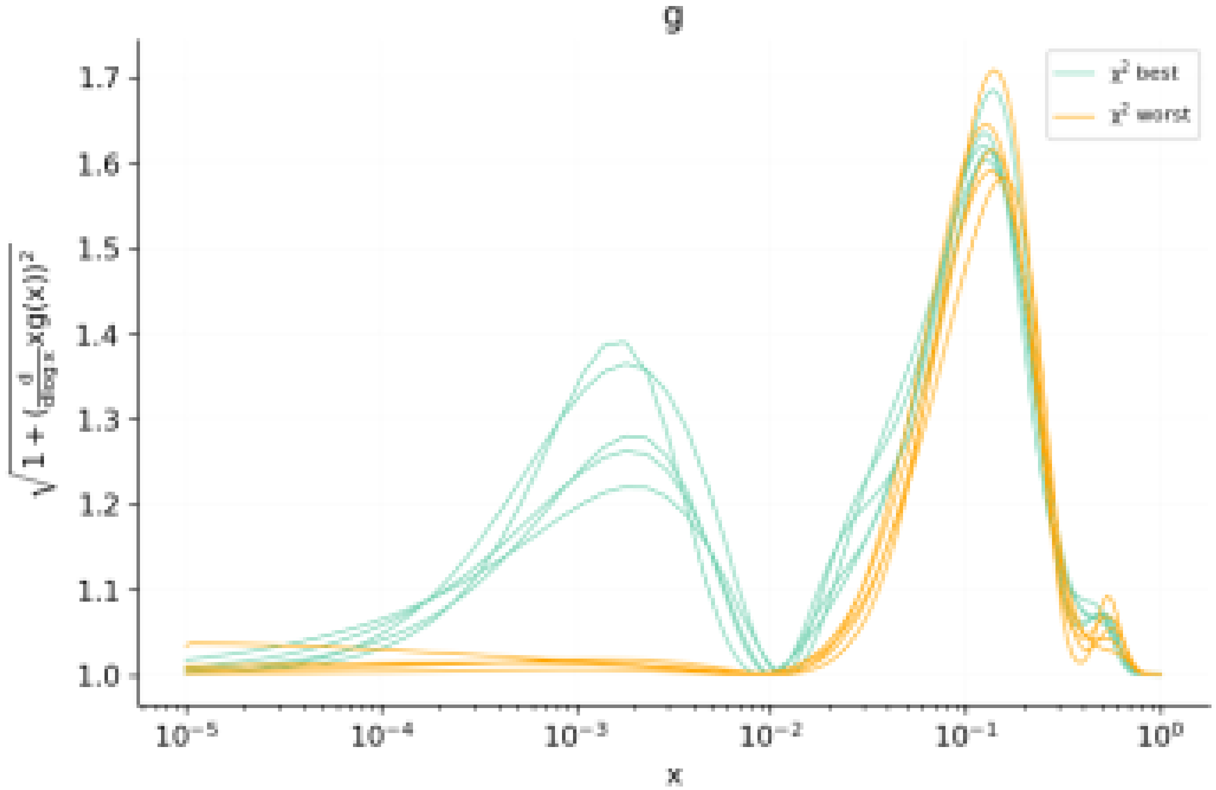
THE PDF KINETIC ENERGY

REPLICAS WITH LOWEST & HIGHEST χ^2 TO CENTRAL DATA

$$\mathrm{KE} = \sqrt{1 + \left(\frac{d}{d\ln x} x f(x, Q^2)\right)^2}$$

ARCLENGTH OF THE NN OUTPUT IN TERMS OF INPUT

THE GLUON



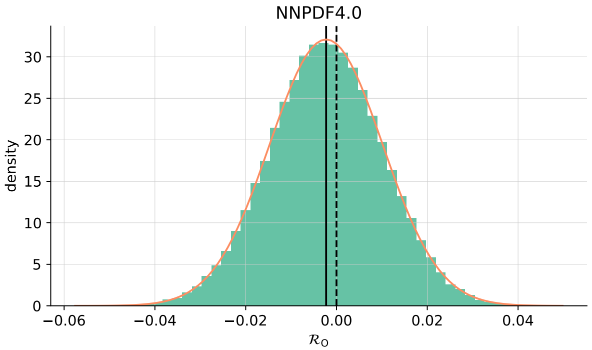
- REPLICAS CLOSER TO CENTRAL DATA ⇒ MORE STRUCTURE
- HIGHER KINETIC ENERGY

EXPLAINING UNCERTAINTIES
OVERLEARNING?

- INDUCE OVERLEARING: DOUBLE TRAINING LENGTH

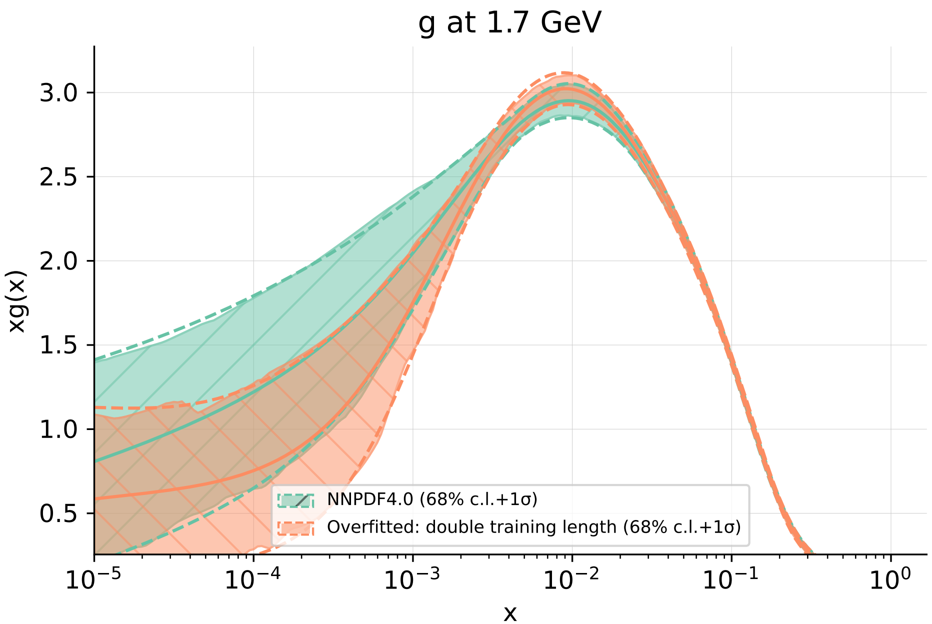
THE OVERFIT METRIC

DEFAULT NNPDF4.0



OVERFIT VARIANT

THE GLUON



- LOOK AT THE OUTPUT ⇒ MORE STRUCTURE IN GLUON

A PARADOX?

- BEST FIT TO CENTRAL DATA CORRELATED TO HIGH ARCLENGTH
- HIGH ARCLENGTH CORRELATED TO OVERLEARNING
- TRAINING/VALIDATION LOSS

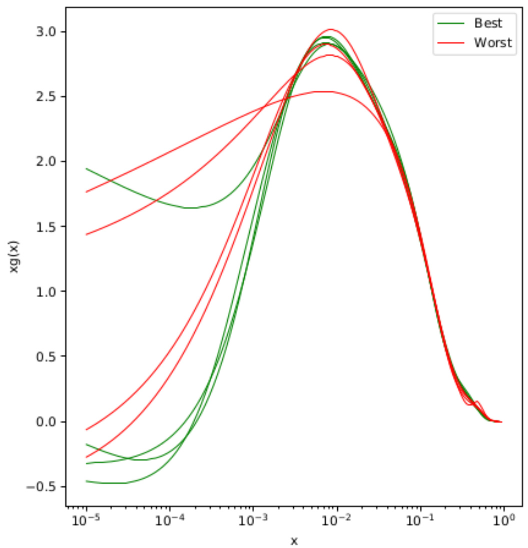
  UNCORRELATED TO QUALITY OF FIT TO CENTRAL DATA

UNDERSTANDING UNCERTAINTIES: TOWARDS XAI
GENERALIZATION!

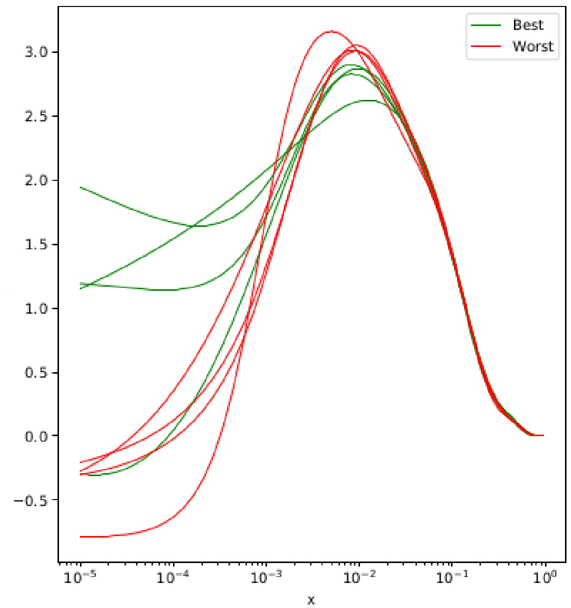
- OVERFITTING CAN MEAN POOR GENERALIZATION
- KEPT IN CHECK BY K-FOLDING (NOT CROSS-VALIDATION)
- LOOK AT BEST χ^2 TO FITTED VS. EXCLUDED FOLDS

THE GLUON

FITTED FOLDS



EXCLUDED FOLD



- BEST VS WORST REVERSED
- HIGH K.E. SOLUTIONS DO NOT GENERALIZE

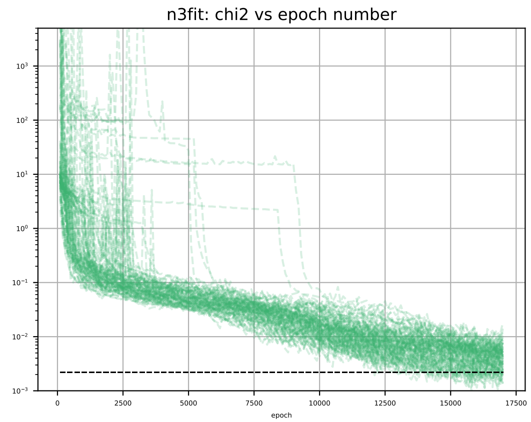
ACT II
ML lessons on PDFs

UNDERSTANDING UNCERTAINTIES: PDFs
DIFFERENT KINDS OF CLOSURE TESTS

- LEVEL 0:
  - EACH DATAPOINT EQUAL TO THE "TRUTH VALUE"; ZERO UNCERTAINTY
  - FIT $\to$ MUST FIND $\chi^2 = 0$ (GET BACK "TRUTH")
  - $\chi^2 \approx 0$ BOTH REPLICA TO REPLICA AND AVERAGE TO TRUTH
  - INTERPOLATION/EXTRAPOLATION UNCERTAINTY
- LEVEL 1:
  - EACH PSEUDO- DATAPOINT IS OBTAINED AS A RANDOM FLUCTUATION WITH GIVEN COVARIANCE MATRIX ABOUT "TRUTH"
    $\Rightarrow$ "RUN OF THE UNIVERSE"
  - FIT DATA OVER AND OVER AGAIN
  - $\chi^2 \approx 1$ BOTH REPLICA TO REPLICA AND AVERAGE TO TRUTH
  - FUNCTIONAL UNCERTAINTY
- LEVEL 2:
  - DATA AS IN LEVEL 1
  - GENERATE DATA REPLICAS OF THESE "DATA"
  - FIT PDF REPLICAS TO DATA REPLICAS
  - $\chi^2 \approx 2$ REPLICA TO REPLICA; $\chi^2 \approx 1$ AVERAGE TO TRUTH
  - DATA UNCERTAINTY

UNCERTAINTIES: TYPE AND SIZE
CLOSURE TEST RESULTS (NNPDF4.0)

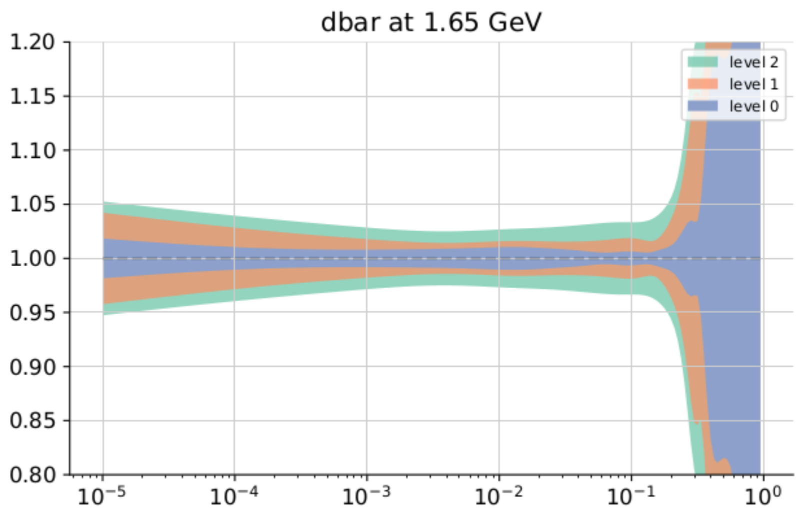
LEVEL 0 χ^2 VS TRAINING



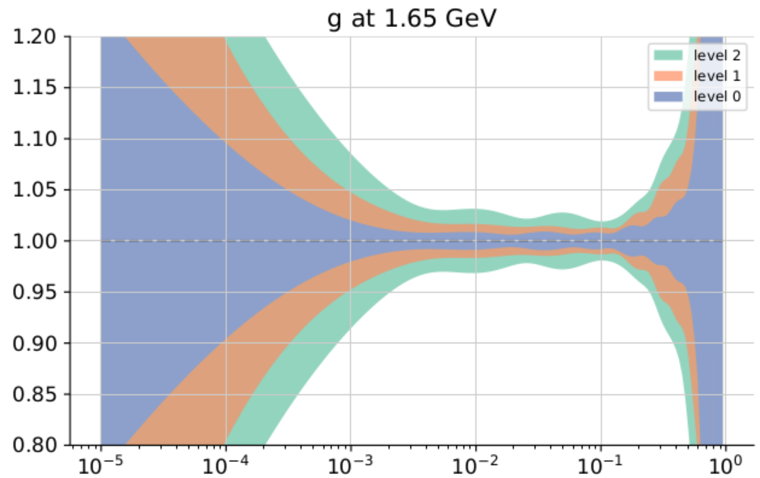
- LEVEL 0 (TRUTH DATA) $\Rightarrow$ $\chi^2 \approx 0$, YET UNCERTAINTY NONZERO
  $\Rightarrow$ NEURAL NETS $\Leftrightarrow$ MANY FUNCTIONAL FORMS
- LEVEL 1 (RUNS OF UNIVERSE) $\Rightarrow$ REPLICAS ALL FITTED TO SAME DATA,
  YET UNCERTAINTY NONZERO
  $\Rightarrow$ DITTO
- LEVEL 0, 1 AND 2 UNCERTAINTIES COMPARABLE IN SIZE

LEVEL 0/1/2 UNCERTAINTIES

ANTIDOWN



GLUON



UNDERSTANDING PDF CORRELATIONS: DATA

example: up vs down PDFs

COVARIANCE: $\text{Cov}[u,d](x,x') = \langle u(x,Q_0^2)d(x',Q_0^2)\rangle - \langle u(x,Q_0^2)\rangle\langle d(x',Q_0^2)\rangle$;

CORRELATION: $\rho[u,d](x,x') = \frac{\text{Cov}[u,d](x,x')}{\sqrt{\text{Var}[u](x)\text{Var}[d](x')}}$

COMPUTATION IN MC APPROACH: $\langle u(x,Q_0^2)d(x',Q_0^2)\rangle = \frac{1}{N}\sum_{r=1}^{N} u^{(r)}(x,Q_0^2)d^{(r)}(x',Q_0^2)$;

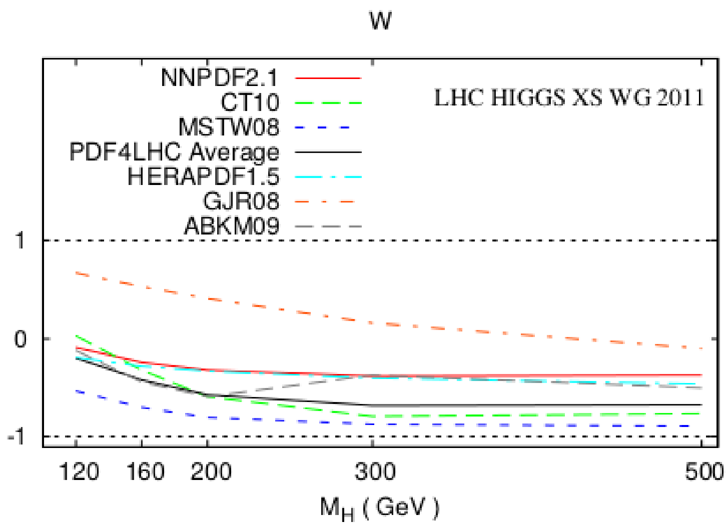
$u^{(r)}(x,Q_0^2)$ REPLICAS

- CORRELATION INDUCED BY DATA, THEORY (E.G. SUM RULES), METHODOLOGY (E.G. ASSUMPTIONS ON EXTRAPOLATION)
- USED E.G. TO ASSESS CORRELATION BETWEEN SIGNAL AND BACKGROUND PROCESSES

PDF-INDUCED CORRELATIONS

BETWEEN HIGGS SIGNAL & BACKGROUND PROCESSES (HXSWG, YR2, 2011)

Higgs in gluon fusion vs. W production



PDF MODEL CORRELATIONS

CORRELATE PDFs IN DIFFERENT SETS

example: up NN model vs down parametric model

$$\mathrm{Cov}[u^N, d^P](x, x') = \langle u^{\mathrm{N}}(x, Q_0^2) d^{\mathrm{P}}(x', Q_0^2)\rangle - \langle u^{\mathrm{N}}(x, Q_0^2)\rangle\langle d^{\mathrm{P}}(x', Q_0^2)\rangle$$

S-CORRELATION VS F-CORRELATION

$\rho[u^N, u^P]$ DIFFERENT SETS, SAME PDF VS. $\rho[u^N, d^N]$ SAME SET, DIFFERENT PDFs

- SAME REPLICA MUST BE USED FOR NONZERO CORRELATION:
  IF REPLICAS UNCORRELATED $\langle u(x, Q_0^2) d(x, Q_0^2)\rangle \stackrel{?}{=} \frac{1}{N}\sum_{r=1}^{N} u^{(r)}(x, Q_0^2) d^{(r')}(x, Q_0^2) = \langle u\rangle\langle d\rangle$
  THEN CORRELATION VANISHES

REPLICA CORRELATION

- FIT PDF REPLICAS $f_i^{(r,\,\mathrm{N})}(x, Q_0^2)$ & $f_i^{(r,\,\mathrm{P})}(x, Q_0^2)$ for all x, i TO SAME DATA REPLICA
- COMPUTE COVARIANCE & CORRELATION USING

$$\langle u(x, Q_0^2) d(x, Q_0^2)\rangle = \frac{1}{N}\sum_{r=1}^{N} u^{(r,\,\mathrm{N})}(x, Q_0^2) d^{(r,\,\mathrm{P})}(x, Q_0^2)$$

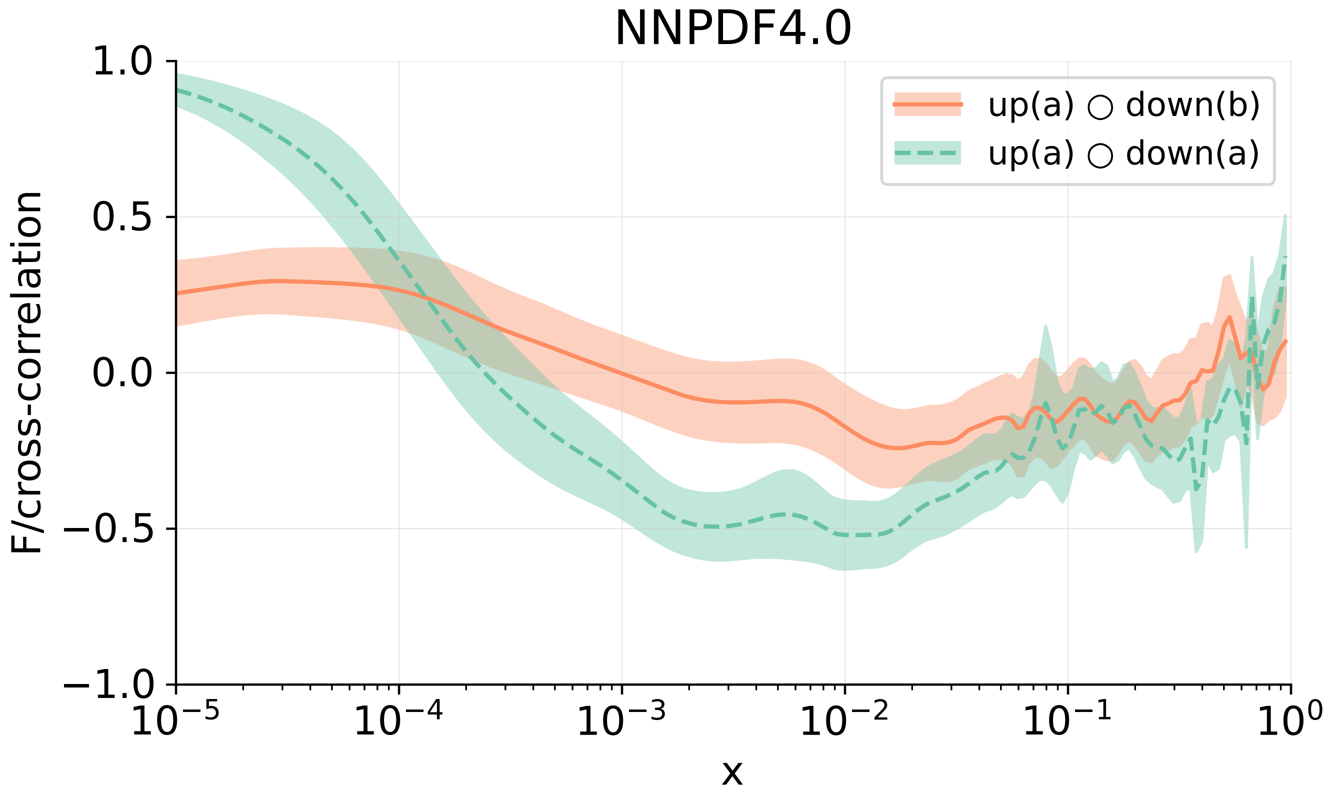
DATA VS MODEL CORRELATION

- NONZERO LEVEL-1 UNCERTAINTY ⇒ DATA REPLICA DOES NOT DETERMINE UNIQUELY THE PDF REPLICA
- IN PRINCIPLE FULL CORRELATION: $r \Leftrightarrow$ DATA REPLICA AND $r' \Leftrightarrow$ LEVEL-1 (METHDOLOLOGY) REPLICAS
  REPLICAS (UP QUARK) $u^{(r,r')}(x, Q_0^2)$;

$$\left| \frac{1}{N} \sum_{r=1}^{N} u^{(r,r')}(x,Q_0^2) d^{(r,r'')}(x,Q_0^2) - \langle u \rangle \langle d \rangle \right| \leq \left| \frac{1}{NM} \sum_{r=1}^{N} \sum_{r'=1}^{M} u^{(r,r')}(x,Q_0^2) d^{(r,r')}(x,Q_0^2) - \langle u \rangle \langle d \rangle \right|$$

- IN PRACTICE METHODOLOGY CORRELATION NOT INCLUDED ⇒ CORRELATION LOSS

FULL VS DATA-INDUCED

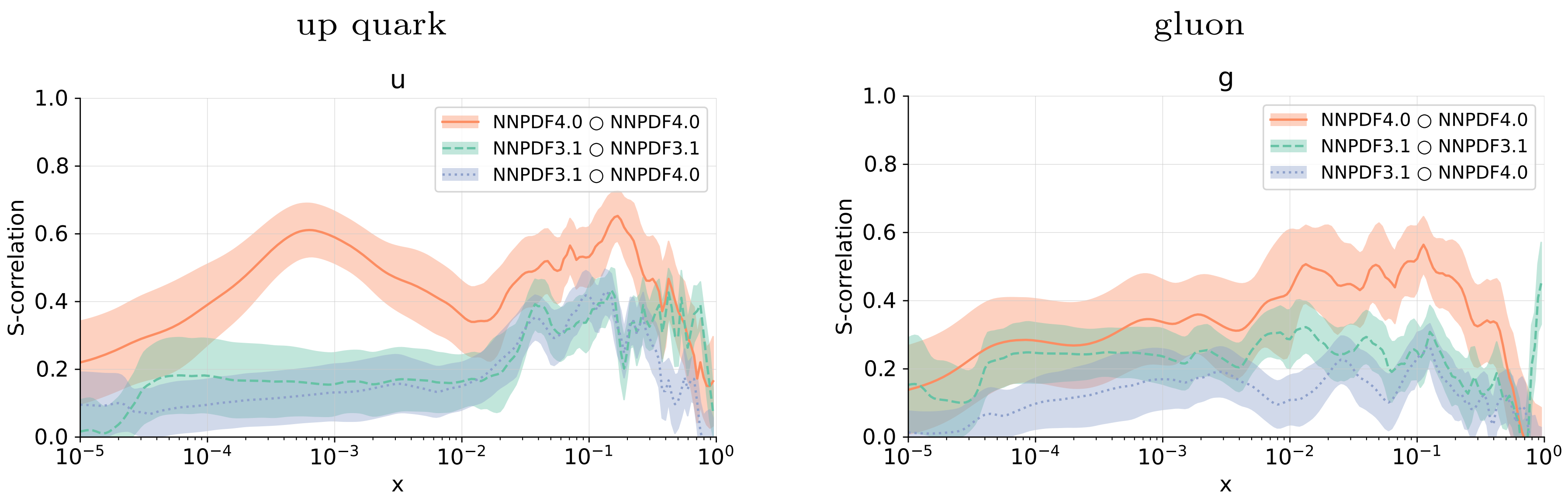


MEASURING MODEL (DE)CORRELATION

- SELF-CORRELATION: S-CORRELATION OF A PDF SET TO ITSELF
  = F-CORRELATION OF A PDF TO ITSELF
- USE TWO DIFFERENT SETS OF PDF REPLICAS FITTED TO THE SAME DATA REPLICAS

$$\langle u(x,Q_0^2)u(x,Q_0^2)\rangle = \frac{1}{N}\sum_{r=1}^{N} u^{(r,r')}(x,Q_0^2)u^{(r,r'')}(x,Q_0^2)$$

- DEVIATION OF CORRELATION FROM 100% MEASURES THE CORRELATION LOSS ⇒ UNCORRELATED FUNCTIONAL UNCERTAINTY
- HIGHER CORRELATION ⇒ MORE EFFICIENT METHODOLOGY

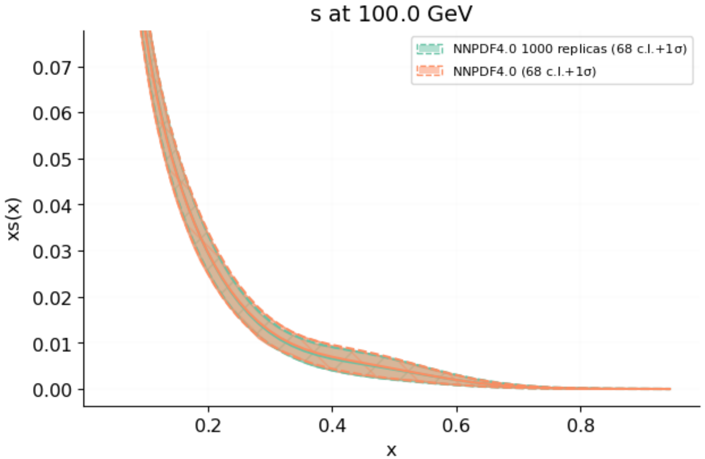


CHARM IN THE PROTON

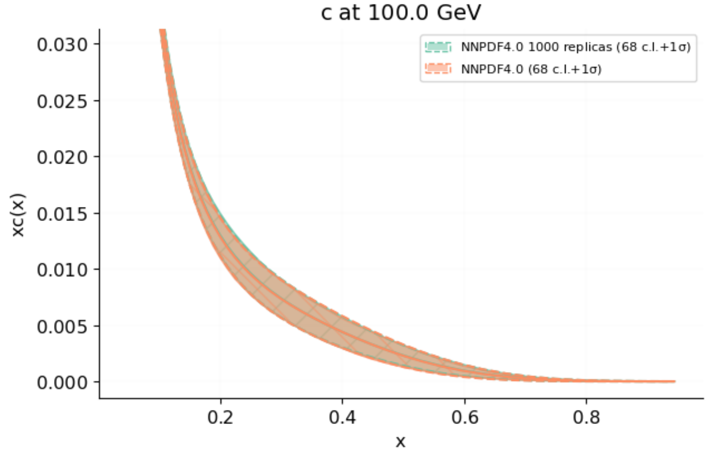
PDFs AT HIGH SCALE

$Q = 100$ GeV

STRANGE



CHARM

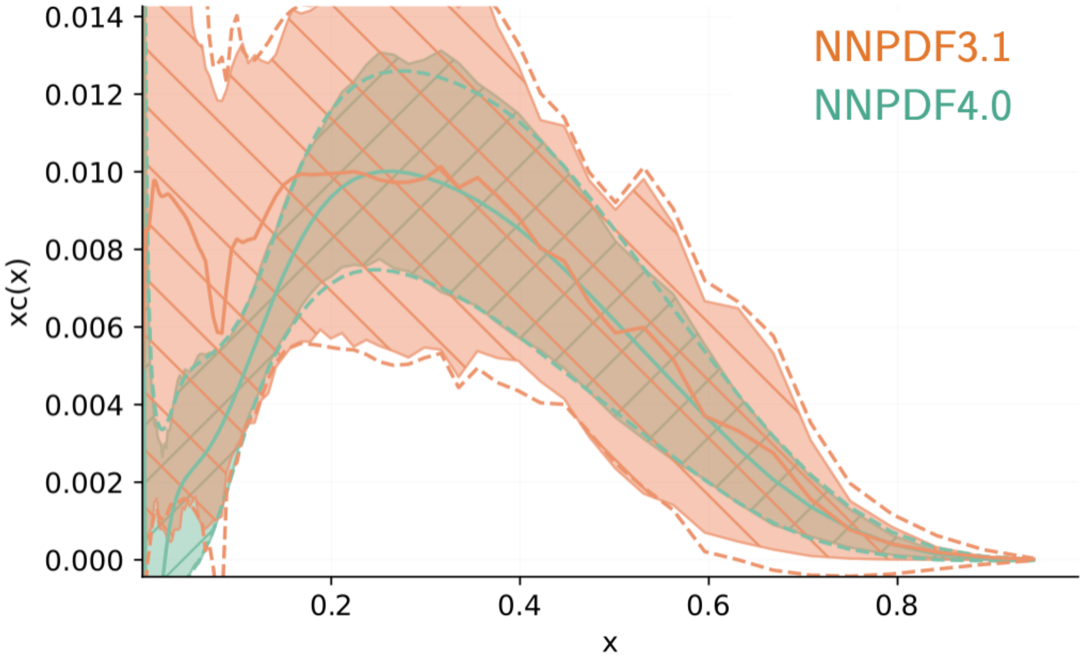


- SEA PDFs AT HIGH SCALE ALL LOOK ALIKE
- IF $Q \gg m_c$, CHARM MASS NEGLIGIBLE: $\ln \frac{Q^2+m_c^2}{m_c^2} \approx \ln \frac{Q^2}{m_c^2}$
- GLUON RADIATION IS FLAVOR BLIND

DECOUPLING

EVOLVE CHARM PDF ($N_f = 4$ SCHEME) DOWN TO $Q \sim m_c$

CHARM PDF (FITTED FROM DATA) AT Q=1.65 GeV



- IF $Q \sim m_c$ ($m_c = 1.51$ GeV), CHARM QUARK DECOUPLES (Collins, Wilczek, Zee, 1978): $\ln \frac{Q^2+m_c^2}{m_c^2} \approx \frac{m_c^2}{Q^2}$
- $N_f = 3$ ACTIVE FLAVORS IN β FUNCTION & EVOLUTION EQUATIONS
- DECOUPLING VS $\overline{\mathrm{MS}}$ $\Leftrightarrow$ DIFFERENT RENORMALIZATION & FACTORIZATION SCHEMES

MATCHING

OME CONTRIBUTING

TO THE CHARM PDF

SOLID $\Rightarrow$ HEAVY; DASHED $\Rightarrow$ LIGHT

- PDFS, α_s IN $N_f = 3$ & $N_f = 4$ RELATED BY MATCHING CONDITIONS

- DETERMINED BY COMPUTING OPERATOR MATRIX ELEMENTS IN EITHER SCHEME AND EQUATING:
  NNLO (Buza, et al., 1998),
  N^{3}LO (Ablinger, Blümlein et al, 2009-2017)

M. Buza et al.: Charm

Fig. 2. $O(\alpha_s^2)$ contributions to the purely-singlet OME $A^{\rm PS}_{q'q}$. Here q and q' are represented by the *dashed* and *solid lines* respectively. In the case of $q' = H$ these graphs contribute to the heavy-quark OME $A^{\rm PS}_{Hq}$

PERTURBATIVE CHARM

- NO CHARM PDF IN $N_f = 3$ SCHEME

- IN $N_f = 4$ SCHEME, CHARM DETERMINED BY PERTURBATIVE MATCHING STARTING AT NNLO (TWO LOOPS) DOES NOT VANISH AT ANY SCALE (HEAVY QUARK LOOPS)

INTRINSIC CHARM

- DEFINE CHARM PDF AS OME:

$$\langle p|\bar{c}\gamma^{\mu_1} D^{\mu_2} \dots D^{\mu_n} c|p\rangle = A_c^n p^{\mu_1} \dots p^{\mu_n} - \text{traces}$$

$$A_c^n = \int_0^1 dx\, x^{n-1} c(x)$$

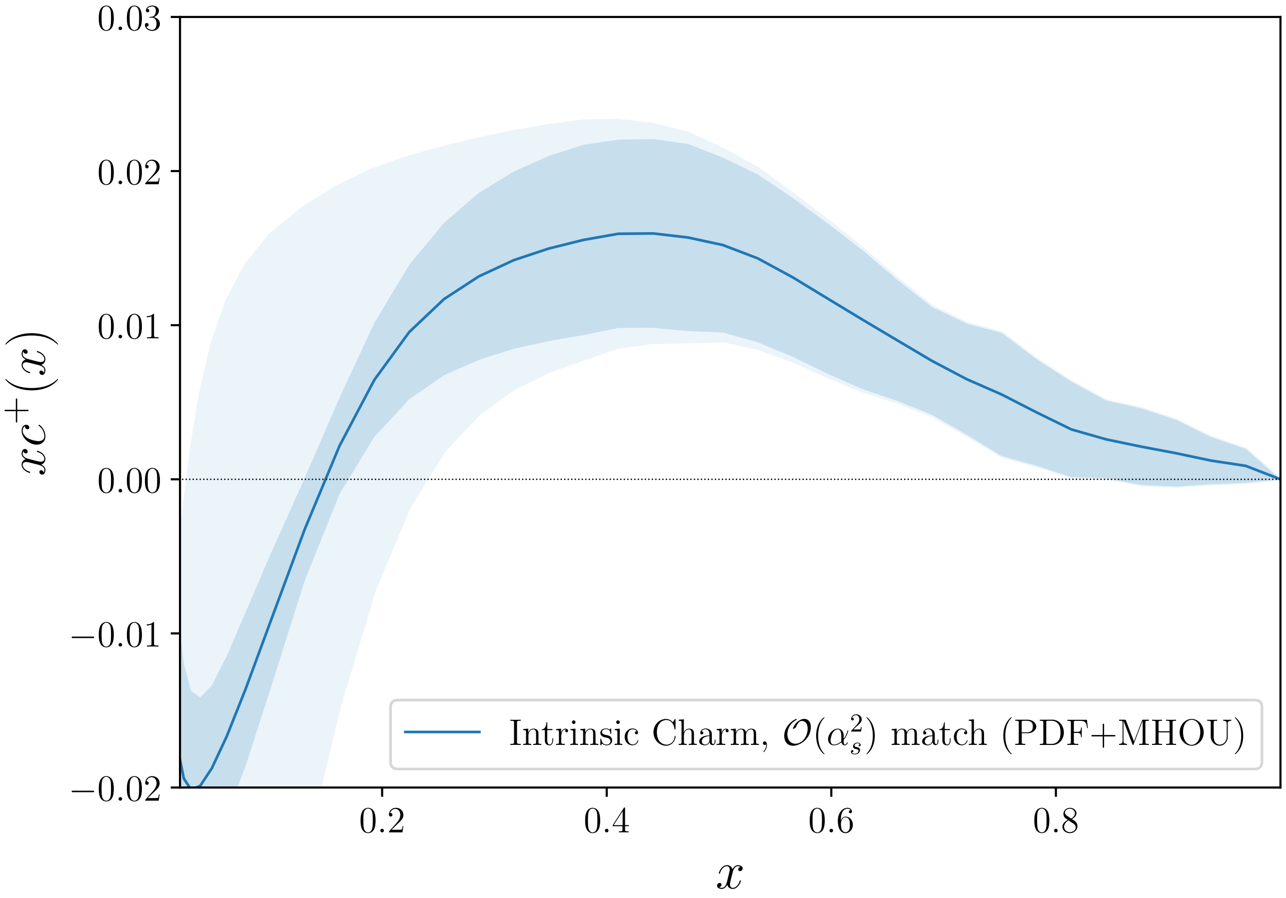
- DO NOT FACTOR CHARM MASS SINGULARITIES INTO OME
- $\Rightarrow$ CHOOSE $n_f = 3$ SCHEME
- CHARM PDF PURELY INTRINSIC, SCALE-INDEPENDENT

INTRINSIC CHARM IS CHARM IN THE $N_F = 3$ (DECOUPLING) SCHEME

INTRINSIC CHARM

- MHOU ESTIMATED FROM N^3LO-NNLO MATCHING DIFFERENCE
  - LARGE UNCERTAINTY AT SMALL x
  - NEGLIGIBLE UNCERTAINTY IN VALENCE REGION
- COMPATIBLE WITH ZERO AT SMALL x
- CLEAR EVIDENCE FOR INTRINSIC VALENCE PEAK

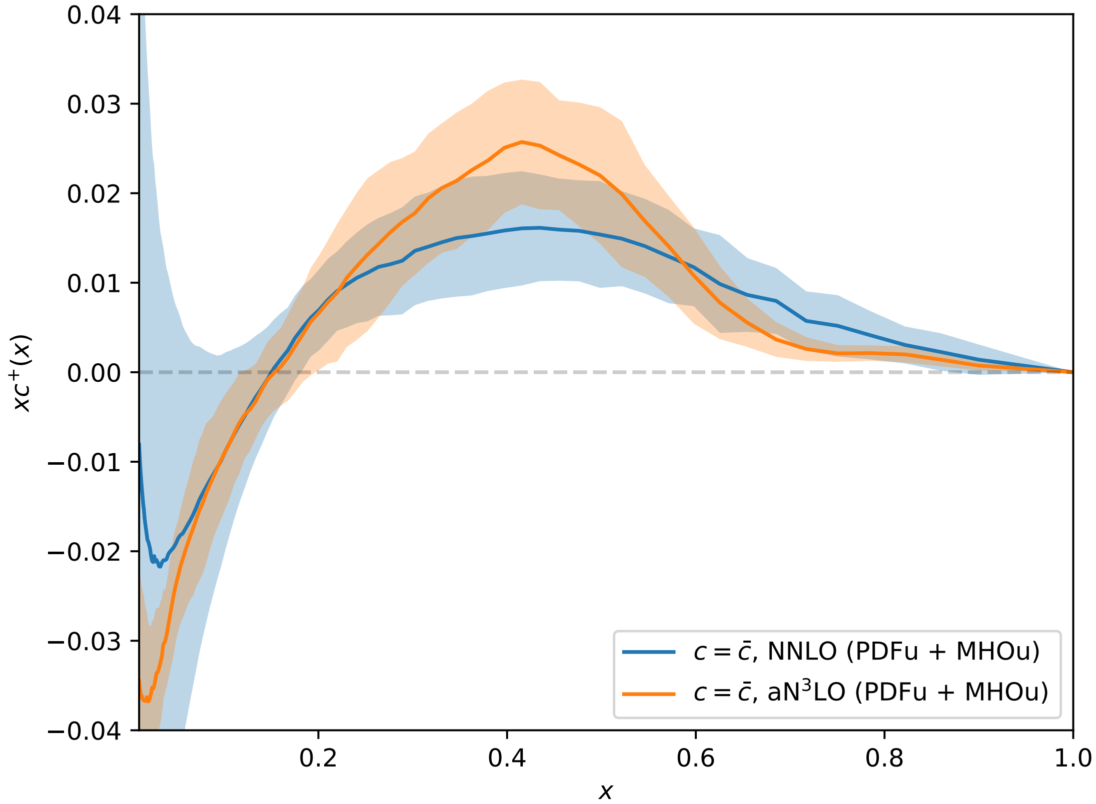
3FNS



CHARM AT AN^{3}LO

- IMPROVED N^{3}LO MATCHING (Blümlein, Ablinger et al., 2023) $\Rightarrow$ SOMEWHAT REDUCED INSTABILITY
- (APPROXIMATE) N^{3}LO PDFs $\Rightarrow$ " TRUE" MHOU
- MHOU (THEORY COVMAT FROM SCALE VARIATION) INCLUDED IN N^{3}LO RESULTS

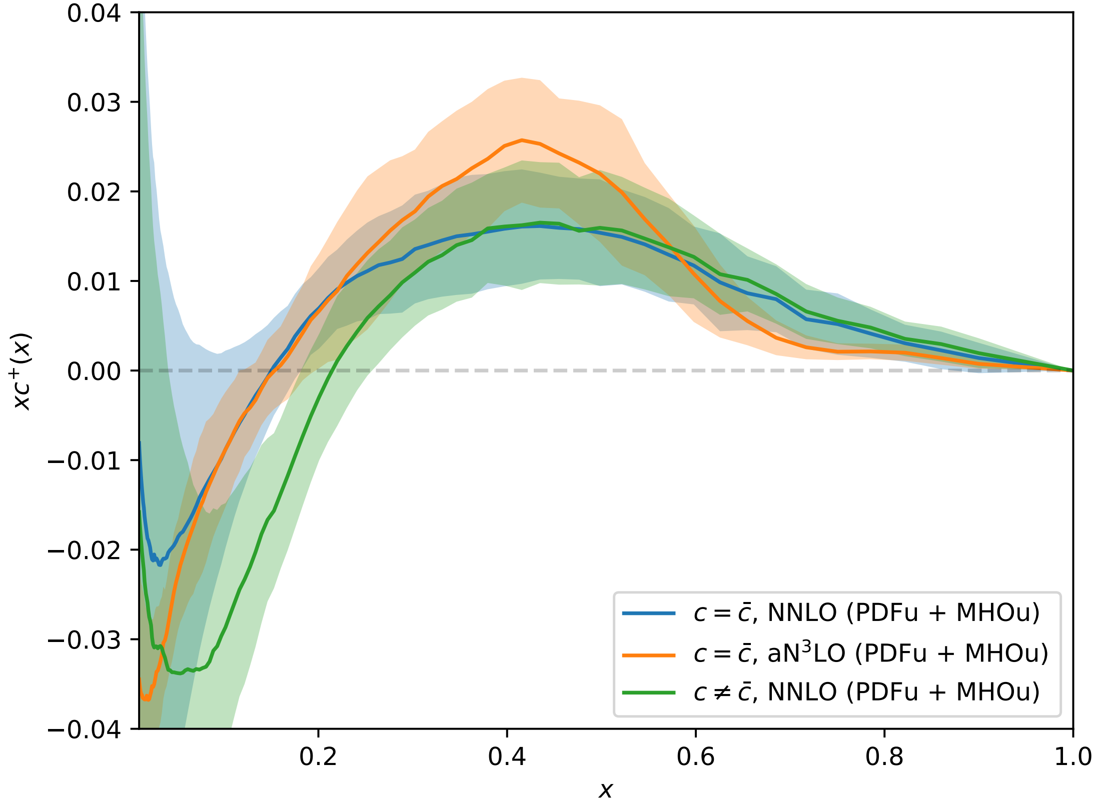
3FNS



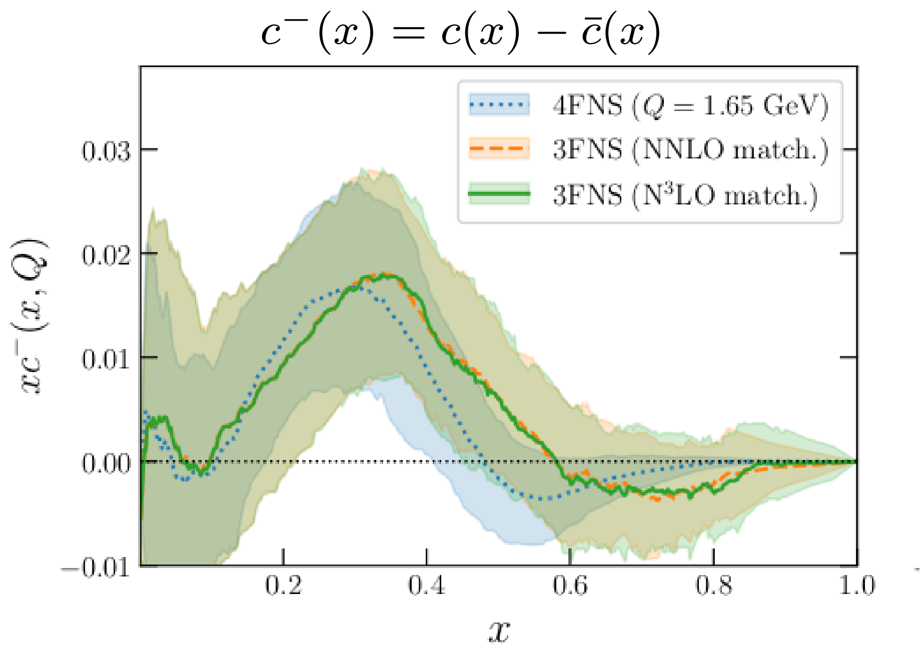
A VALENCE CHARM PDF?

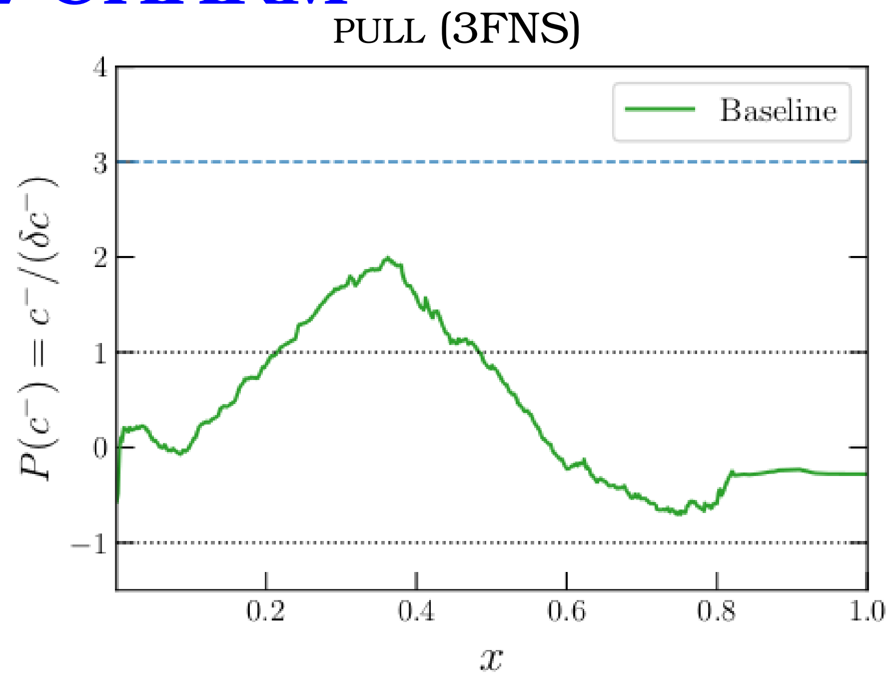
- INDEPENDENT PARAMETRIZATION FOR "SEA" $c^+ = c + \bar{c}$ AND "VALENCE" $c^- = c - \bar{c}$ PDFs
- TOTAL CHARM UNCHANGED

3FNS



VALENCE CHARM





- NNLO $n_f = 4$ VALENCE PDF FROM PERTURBATIVE MATCHING VANISHES
- NONVANISHING VALENCE CHARM PDF IN VALENCE REGION $\Rightarrow$ INTRINSIC CHARM

EPILOGUE
WHAT REMAINS TO BE DONE?

A TO DO LIST

- MACHINE LEARNING: XAI
  - HOW DOES THE ML MODEL RESPOND TO DATA INCONSISTENCIES?
  - AN ON-THE-FLY OVERLEARNING METRIC?
  - NEURAL NETWORKS VS. GAUSSIAN PROCESSES/BAYESIAN INFERENCE?
  - CORRELATION BETWEEN DATA FEATURES AND MODEL FEATURES?
- PDFs: PRECISION AND ACCURACY
  - AUTOMATIC K-FOLDS
  - HYPEROPT BEYOND χ^2 LOSS
  - FULL QCDxEW THEORY BEYOND K-FACTORS
  - $2 \to 2$ PROCESSES (VBS)