THE ANATOMY OF NNPDF

STEFANO FORTE

UNIVERSITÀ DI MILANO & INFN

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI FISICA

Istituto Nazionale di Fisica Nucleare

CMS SMP MEETING

NOVEMBER 3, 2020

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 740006

SUMMARY

UNCERTAINTIES

- UNCERTAINTIES AND PDFS: NOW, AND TOMORROW
- THE PROBLEMS OF PDF UNCERTAINTIES

ARTIFICIAL INTELLIGENCE

- THE NNPDF METHODOLOGY
- CLOSURE TESTS

MACHINE LEARNING

- AI VS. ML
- HYPEROPTIMIZATION
- CLOSURE TESTS REVISITED

LEARNING THE UNKNOWN

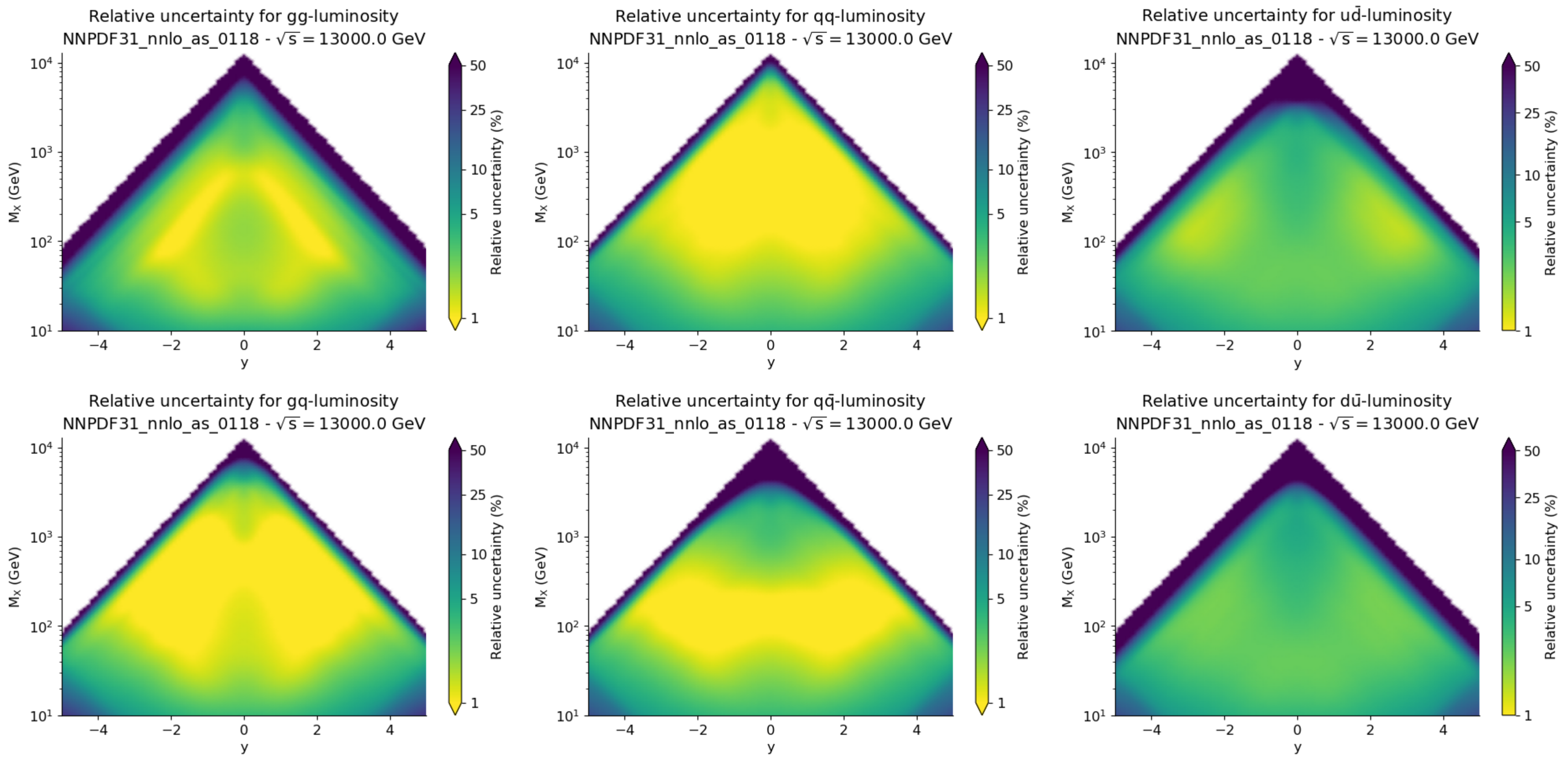
- LEARNING LEARNING
- FUTURE TESTS
- LEARNING THEORY

PDF UNCERTAINTIES

PDF UNCERTAINTIES: NOW

NNPDF3.1 NNLO (2017)

GLUON SINGLET FLAVORS

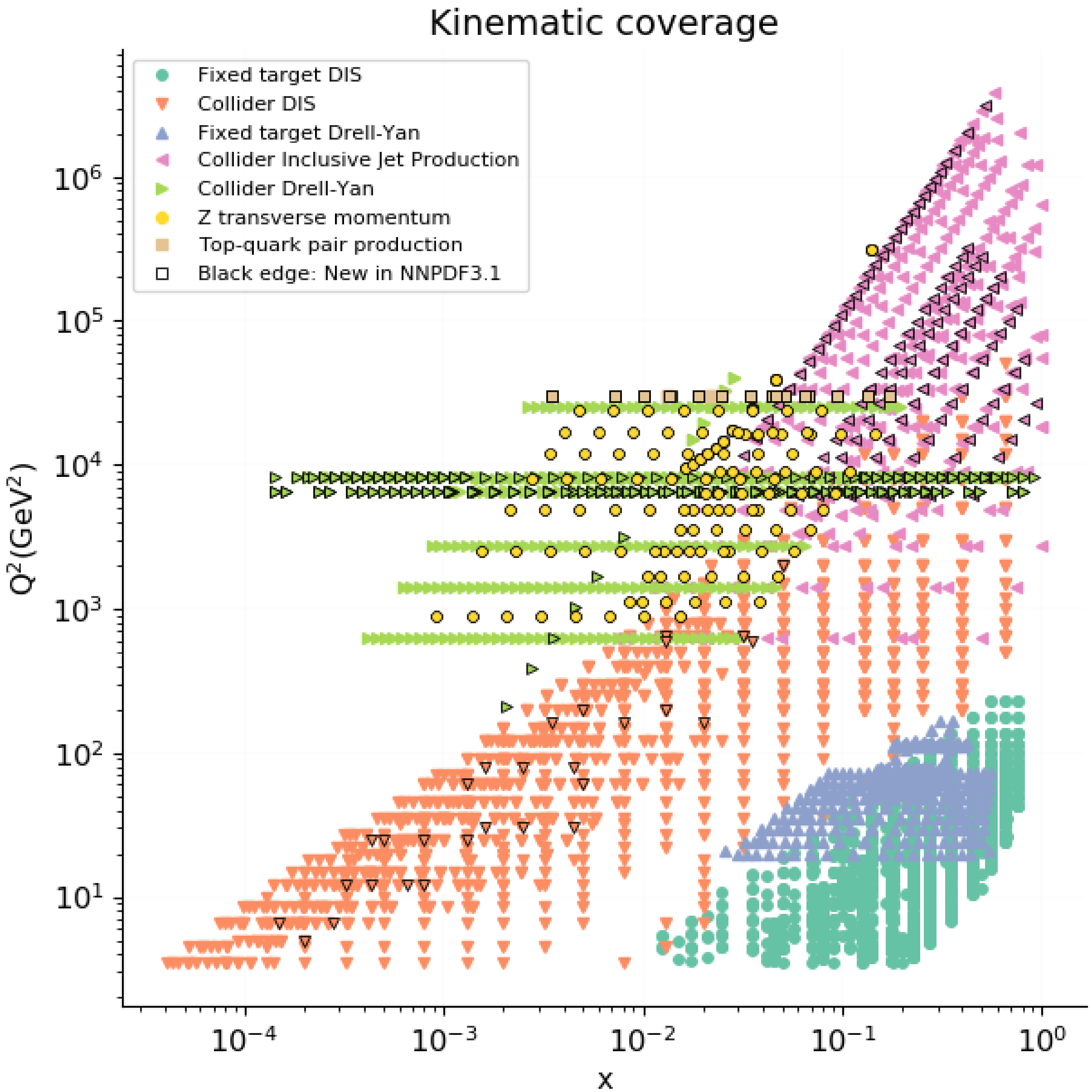


- TYPICAL UNCERTAINTIES IN DATA REGION $\sim 1-3\%$
- SWEET SPOT: VALENCE Q - G; 1% OR BELOW

CT18 (Dec 2019): SOMEWHAT SMALLER DATASET, RATHER LARGER UNCERTAINTIES

DATASET WIDENING

NNPDF3.0 vs NNPDF3.1 (CT14 vs. CT18: similar)



NEW DATA: (black edge)

- HERA combined F_2^b
- D0 W lepton asymmetry
- ATLAS W, Z 2011, high & low mass DY 2011; CMS $W^\pm$ rapidity 8TeV LHCb W, Z 7TeV & 8TeV
- ATLAS 7TeV jets 2011, CMS 2.76TeV jets
- ATLAS & CMS top differential rapidity
- ATLAS Z p_T differential rapidity & invariant mass 8TeV, CMS Z p_T differential rapidity 8TeV

DATASET WIDENING
NNPDF4.0 SUMMARY (EXPECTED IN 2020)

1. OLD DATASETS WITH IMPROVED TREATMENT
   - ASSORTED DEBUGGING
   - CORRELATIONS IN ATLAS TOP DISTRIBUTIONS AT 8 TEV
   - CHOICE OF SCALE AND CORRELATION MODELS FOR SINGLE-JET DATA
   - MASSIVE CORRECTIONS TO NEUTRINO DIS DIMUON CROSS SECTIONS AT NNLO
   - NUCLEAR UNCERTAINTIES IN FIXED-TARGET DIS AND DY
2. NEW DATASETS FOR OLD PROCESSES
   - DIS c AND b PRODUCTION (HERA COMBINED)
   - SINGLE JET PRODUCTION (ATLAS, CMS)
   - TOP PAIR PRODUCTION (ATLAS, CMS)
   - COLLIDER DY/INCLUSIVE VECTOR BOSON PRODUCTION (ATLAS, CMS, LHCB)
   - COLLIDER VECTOR BOSON PRODUCTION IS ASSOCIATION WITH CHARM (CMS)
3. NEW DATASETS FOR NEW PROCESSES
   - ISOLATED PHOTON PRODUCTION (ATLAS)
   - SINGLE TOP PRODUCTION (ATLAS, CMS)
   - COLLIDER DIJET PRODUCTION (ATLAS, CMS)
   - DIS+JET(S) PRODUCTION (H1, ZEUS)
   - COLLIDER VECTOR BOSON PRODUCTION IS ASSOCIATION WITH JETS (ATLAS, CMS)

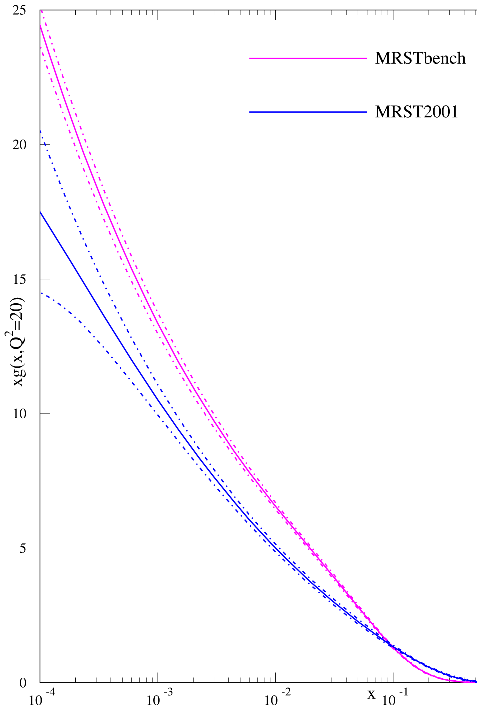
$\mathcal{O}(50)$ NEW/REVISED DATASETS

TOWARDS SUBPERCENT UNCERTAINTIES??!!

THE PDF UNCERTAINTY PROBLEM:
THE HERA-LHC BENCHMARK (2005)

- RESTRICTED AND VERY CONSISTENT DATASET USED
- RESULTS COMPARED TO THEN-BEST RESULT FROM FULL DATASET

BENCHMARK VS DEFAULT GLUON

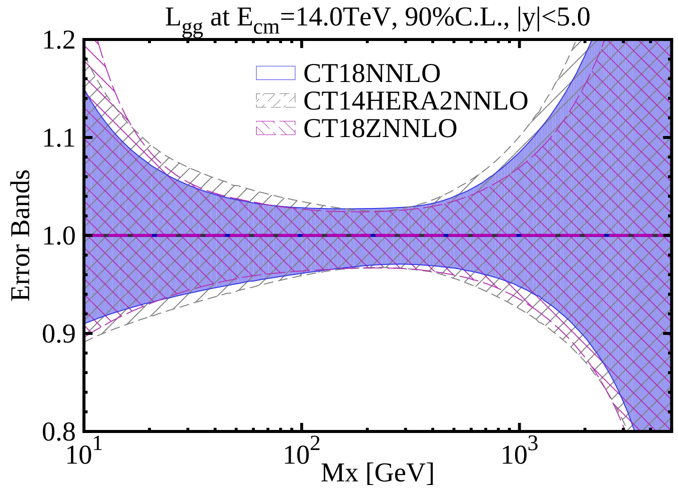


"...the partons extracted using a very limited data set are completely incompatible, even allowing for the uncertainties, with those obtained from a global fit with an identical treatment of errors...The comparison illustrates the problems in determining the true uncertainty on parton distributions." (R.Thorne, HERALHC, 2005)

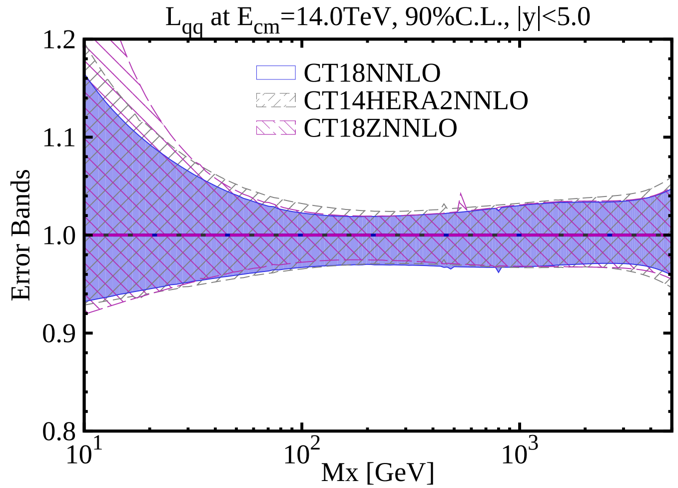
THE PDF UNCERTAINTY PROBLEM: UNCERTAINTY REDUCTION?

CT18 VS. CT14: PARTON LUMINOSITY UNCERTAINTIES

GLUON-GLUON



QUARK-QUARK



MORE DATA $\Rightarrow$ BIGGER UNCERTAINTIES (?!)

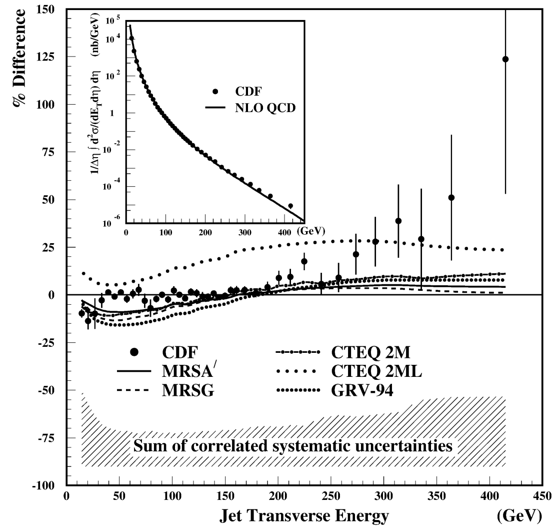
PARTON PARAMETRIZATIONS

- CTEQ5 2002: $xg(x,Q_0^2) = A_0 x^{A_1}(1-x)^{A_2}(1+A_3x^{A_4})$
- 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: $g(x,Q=Q_0) = x^{a_1-1}(1-x)^{a2}\left[a_3(1-y)^3 + a_4 3y(1-y)^2 + a_5 3y^2(1-y) + y^3\right]$;
  $y=\sqrt{x}$; $a_5=(3+2a_1)/3$.

BIAS?

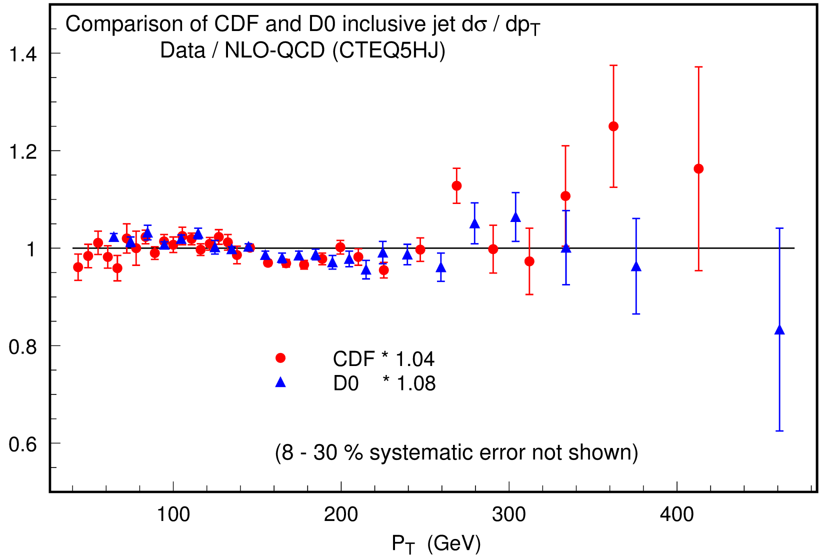
PDF UNCERTAINTIES AND NEW PHYSICS

- DISCREPANCY BETWEEN QCD CALCULATION AND CDF JET DATA (1995)
- EVIDENCE FOR QUARK COMPOSITENESS?
- RESULT STRONGLY DEPENDS ON GLUON AT $x \gtrsim 0.1$
- PDF MUST VANISH AT $x = 0$, BUT (THEN) NO DATA FOR $x \gtrsim 0.05$!



DISCREPANCY REMOVED IF JET DATA USED FOR GLUON DETERMINATION

NEW CTEQ GLUON (1998)

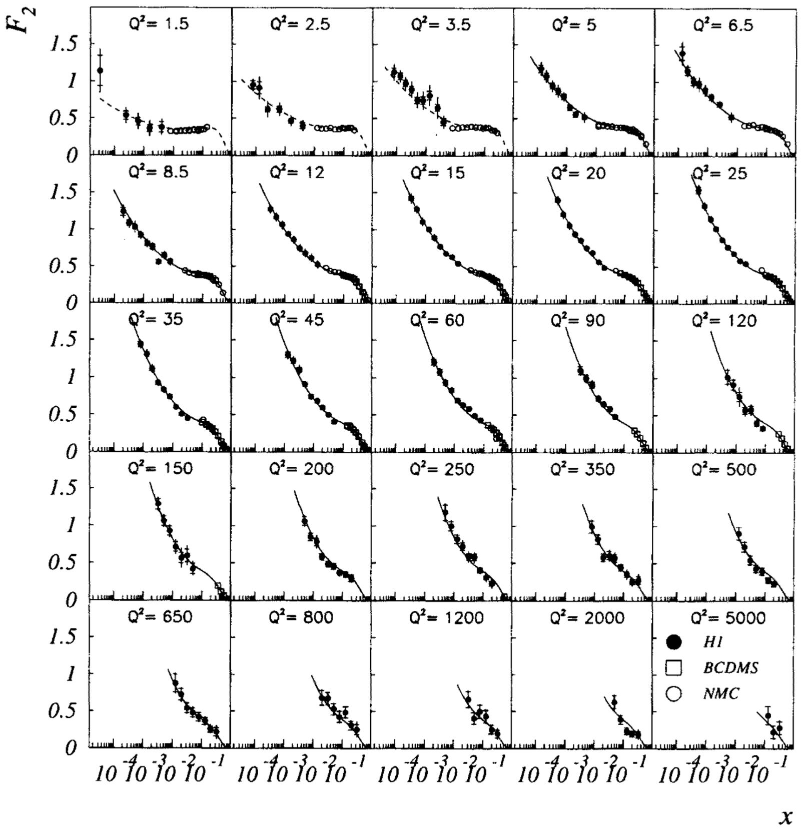


NOW: NO DATA FOR $x \gtrsim 0.5 \Rightarrow$ DISCOVERY (THRESHOLD) REGION!

EXTRAPOLATION AND THEORY BIAS

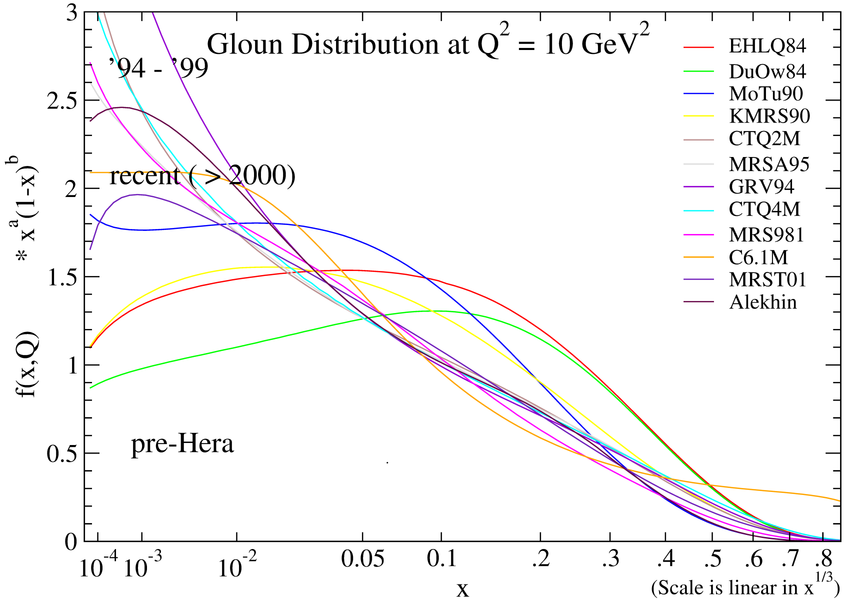
1995: THE RISE OF STRUCTURE FUNCTIONS AT HERA

FIRST HERA DATA VS OLDER DATA



A. de Roeck, Cracow epiphany conf. 1996

HISTORICAL COMPILATION OF GLUON PDFS

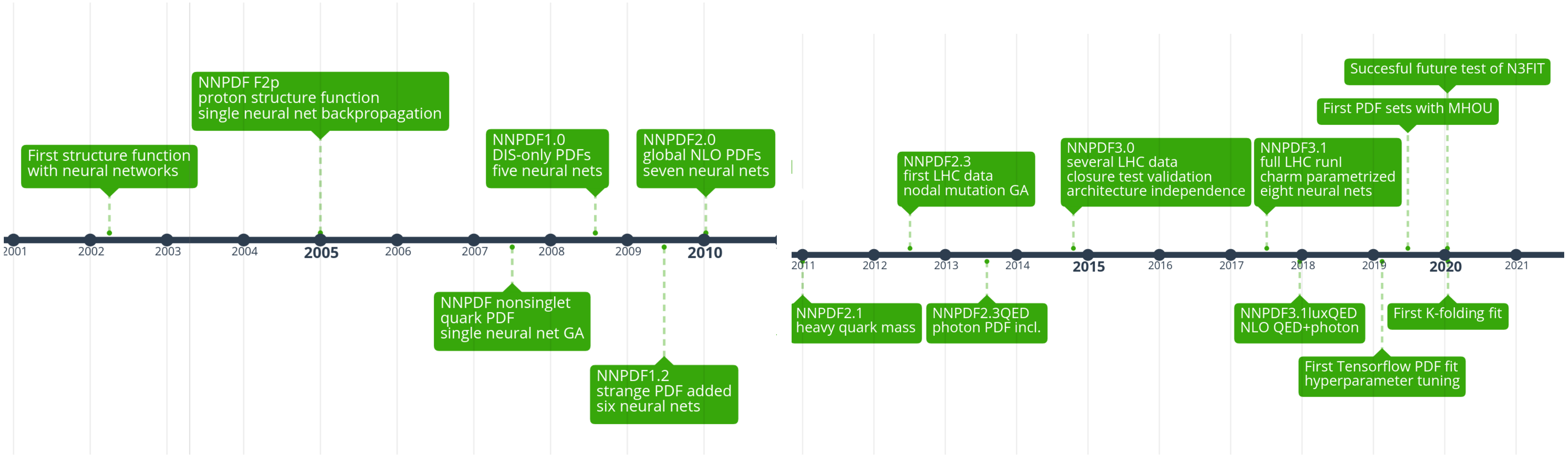


W.K.Tung, DIS 2004

- RISE OF F_2 AT HERA CAME $\Rightarrow$ SURPRIZE
- HINTED BY PRE-HERA DATA; VETOED BY THEORETICAL BIAS

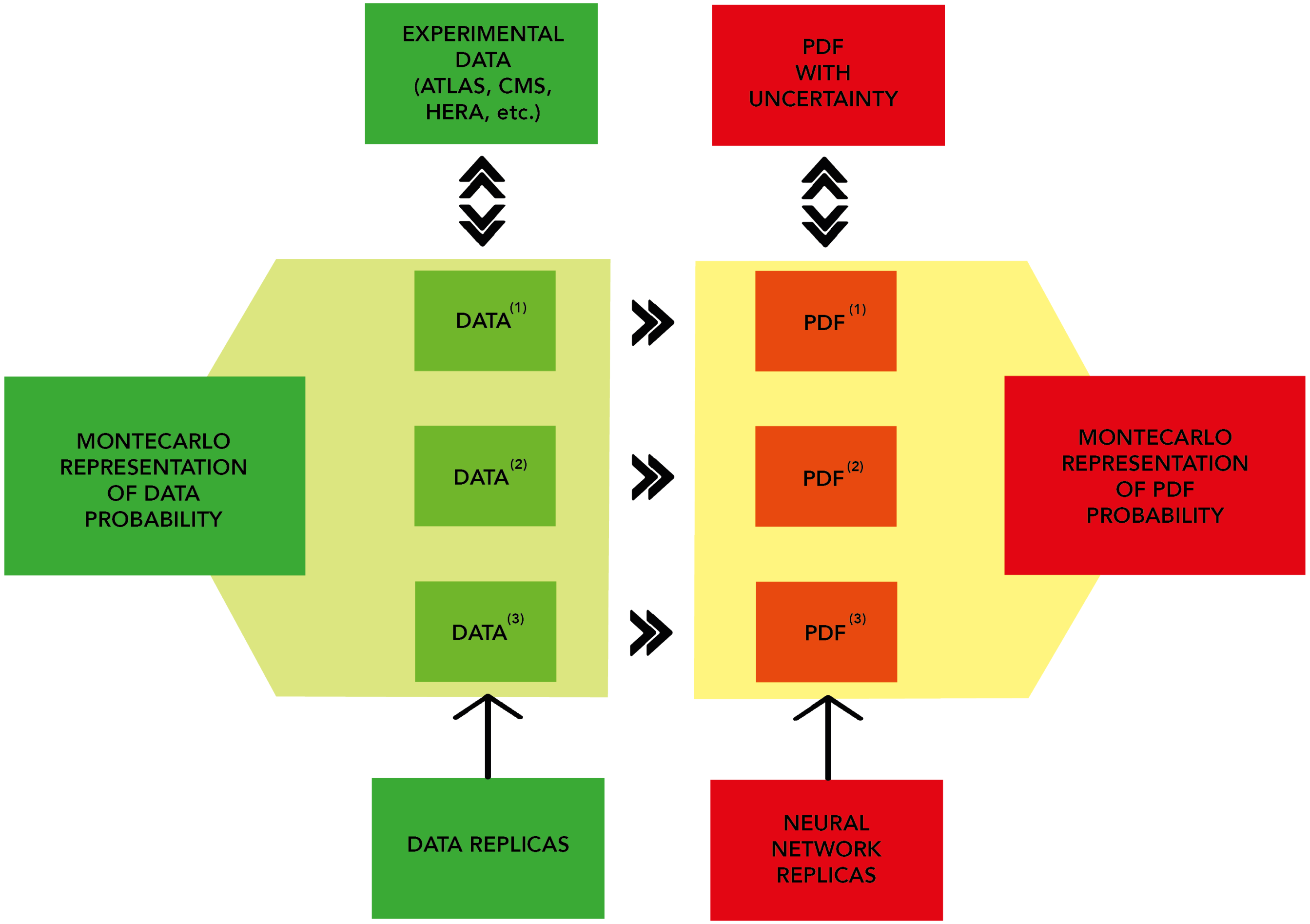
PDFs FROM AI

PROTON STRUCTURE AS AN AI PROBLEM: NNPDF



AI FOR PDFS: THE NNPDF APPROACH
THE FUNCTIONAL MONTE CARLO

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

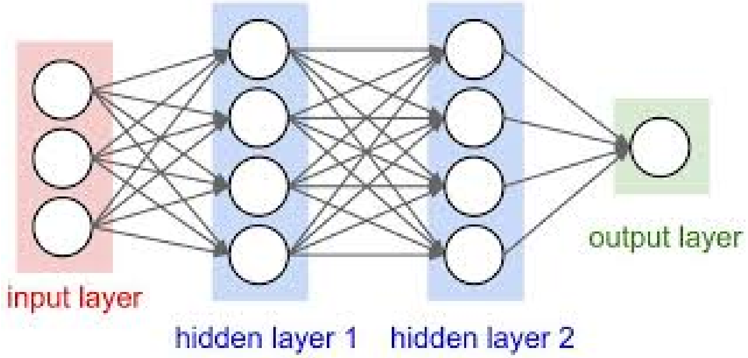


FINAL PDF SET: $f_i^{(a)}(x,\mu)$;
i =up, antiup, down, antidown, strange, antistrange, charm, gluon; $j = 1, 2, \ldots N_{\rm rep}$

ARTIFICIAL INTELLIGENCE

NEURAL NETWORKS

ARCHITECTURE

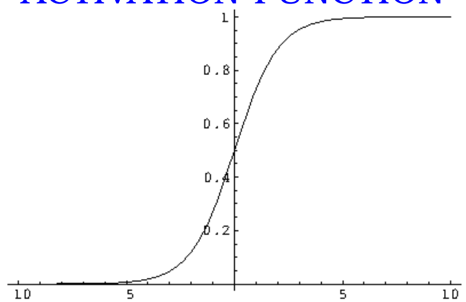


PARAMETERS

- WEIGHTS ω_{ij}
- THRESHOLDS θ_i

$$F_{\text{out}}^{(i)}(\vec{x}_{\text{in}}) = F\left(\sum_j \omega_{ij} x_{\text{in}}^j - \theta_i\right)$$

ACTIVATION FUNCTION



SIMPLEST EXAMPLE
1-2-1

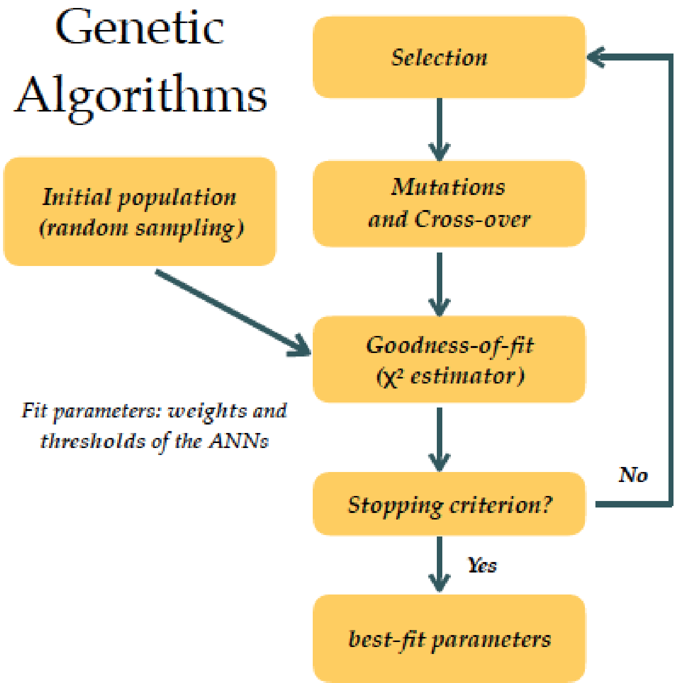
$$f(x) = \frac{1}{1+e^{\theta_1^{(3)} - \frac{\omega_{11}^{(2)}}{1+e^{\theta_1^{(2)} - x\omega_{11}^{(1)}}} - \frac{\omega_{12}^{(2)}}{1+e^{\theta_2^{(2)} - x\omega_{21}^{(1)}}}}}$$

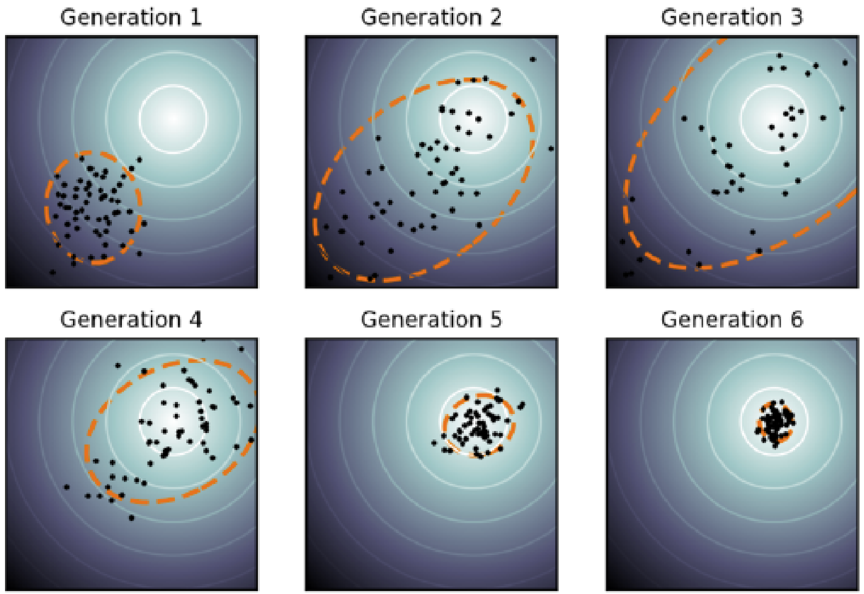
NNPDF: $2-5-3-1$ NN FOR EACH PDF: $37 \times 8 = 296$ PARAMETERS

SUPERVISED LEARNING

GENETIC ALGORITHMS

- BASIC IDEA: RANDOM MUTATION OF THE NN PARAMETER
- SELECTION OF THE FITTEST

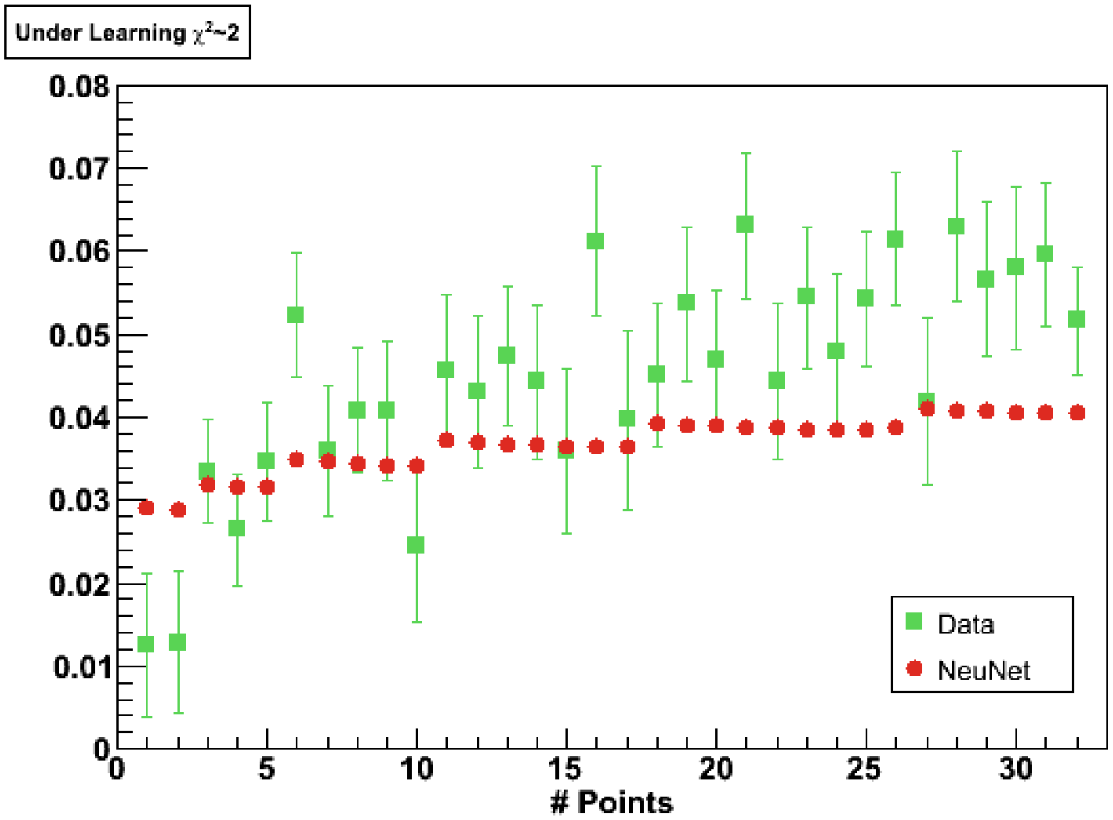




NEURAL LEARNING

- COMPLEXITY INCREASES AS THE FITTING PROCEEDS
- UNTIL LEARNING NOISE
- WHEN SHOULD ONE STOP?

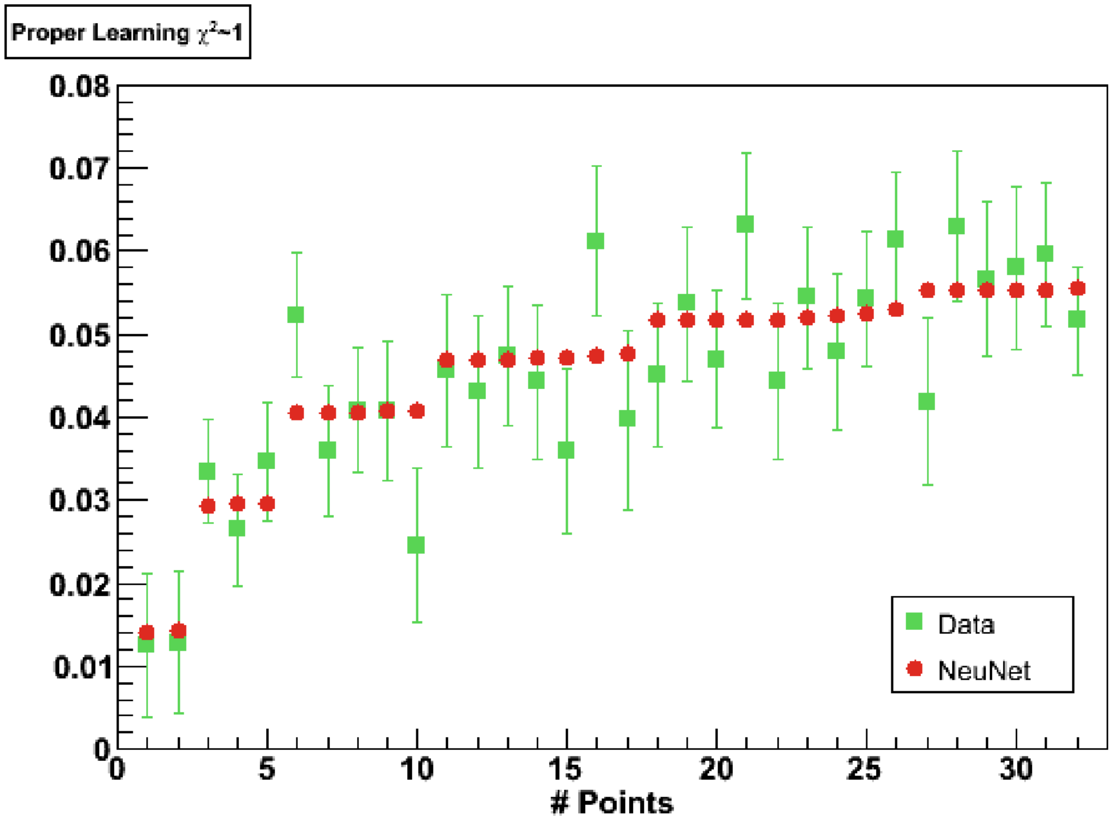
UNDERLEARNING



NEURAL LEARNING

- COMPLEXITY INCREASES AS THE FITTING PROCEEDS
- UNTIL LEARNING NOISE
- WHEN SHOULD ONE STOP?

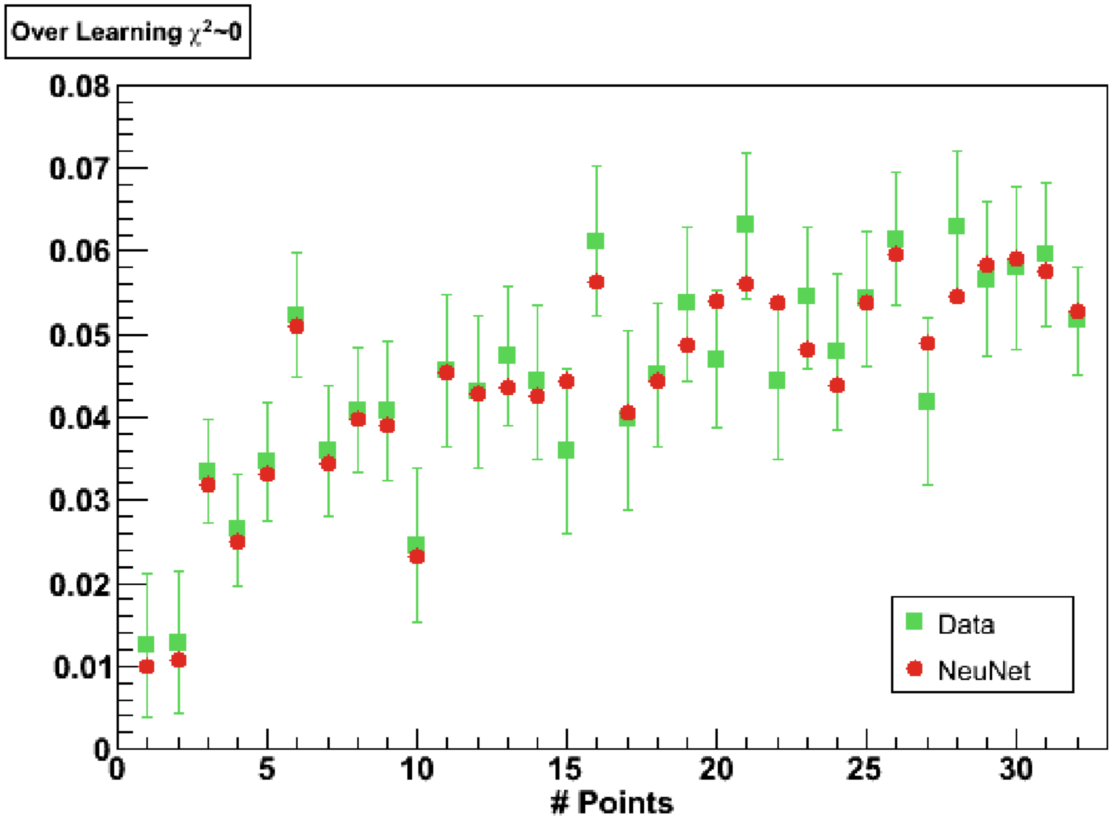
PROPER LEARNING



NEURAL LEARNING

- COMPLEXITY INCREASES AS THE FITTING PROCEEDS
- UNTIL LEARNING NOISE
- WHEN SHOULD ONE STOP?

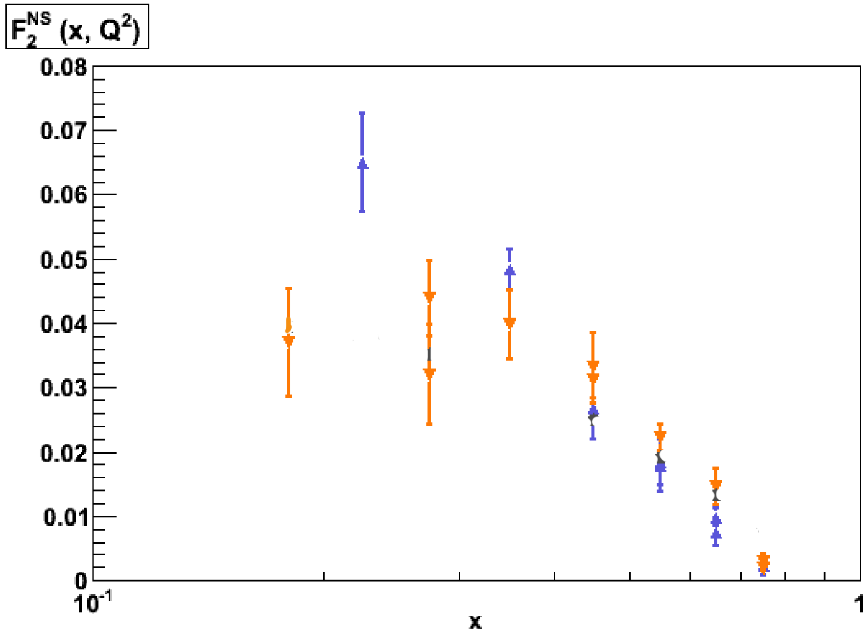
OVERLEARNING



OPTIMAL FIT: CROSS-VALIDATION

GENETIC MINIMIZATION:
AT EACH GENERATION, χ^2 EITHER UNCHANGED OR DECREASING

- DIVIDE THE DATA IN TWO SETS: TRAINING AND VALIDATION
- MINIMIZE THE χ^2 OF THE DATA IN THE TRAINING SET
- AT EACH ITERATION, COMPUTE THE χ^2 FOR THE DATA IN THE VALIDATION SET (NOT USED FOR FITTING)
- WHEN THE VALIDATION χ^2 STOPS DECREASING, STOP THE FIT

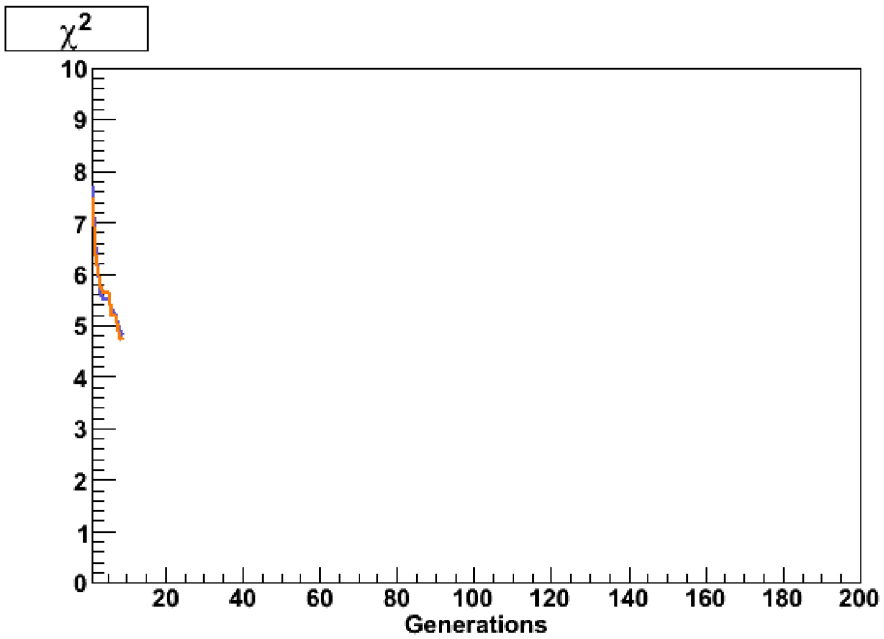


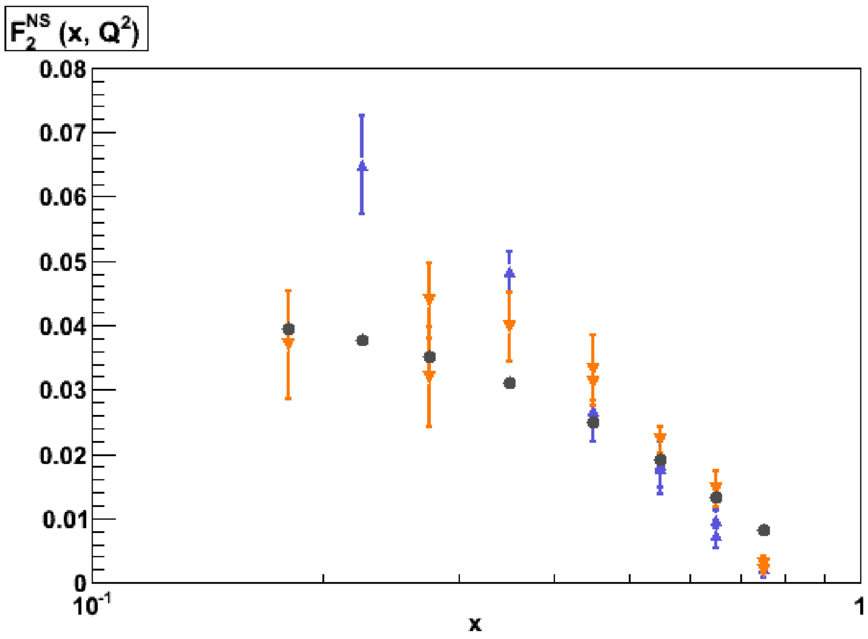
OPTIMAL FIT: CROSS-VALIDATION

GENETIC MINIMIZATION:
AT EACH GENERATION, χ^2 EITHER UNCHANGED OR DECREASING

- DIVIDE THE DATA IN TWO SETS: TRAINING AND VALIDATION
- MINIMIZE THE χ^2 OF THE DATA IN THE TRAINING SET
- AT EACH ITERATION, COMPUTE THE χ^2 FOR THE DATA IN THE VALIDATION SET (NOT USED FOR FITTING)
- WHEN THE VALIDATION χ^2 STOPS DECREASING, STOP THE FIT

GO!



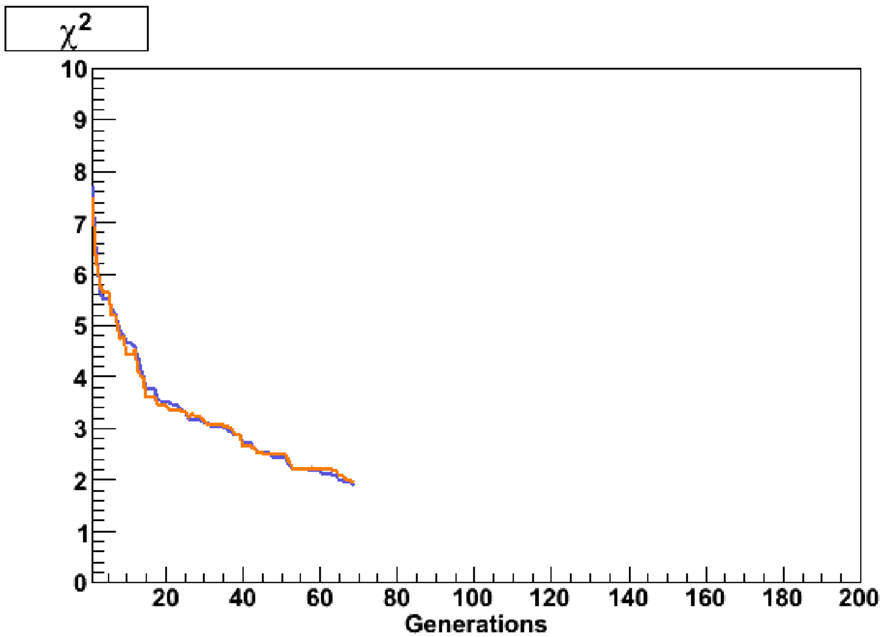


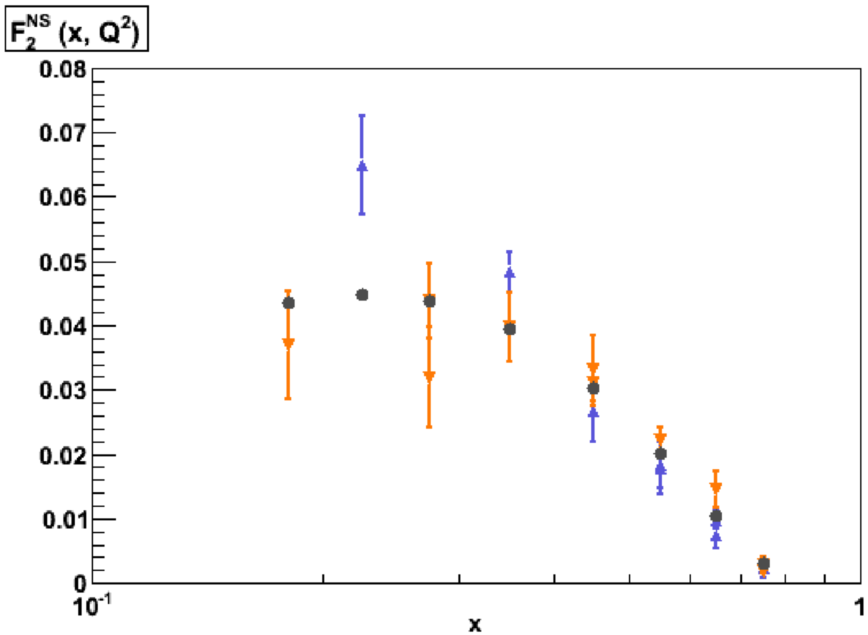
OPTIMAL FIT: CROSS-VALIDATION

GENETIC MINIMIZATION:
AT EACH GENERATION, χ^2 EITHER UNCHANGED OR DECREASING

- DIVIDE THE DATA IN TWO SETS: TRAINING AND VALIDATION
- MINIMIZE THE χ^2 OF THE DATA IN THE TRAINING SET
- AT EACH ITERATION, COMPUTE THE χ^2 FOR THE DATA IN THE VALIDATION SET (NOT USED FOR FITTING)
- WHEN THE VALIDATION χ^2 STOPS DECREASING, STOP THE FIT

STOP!



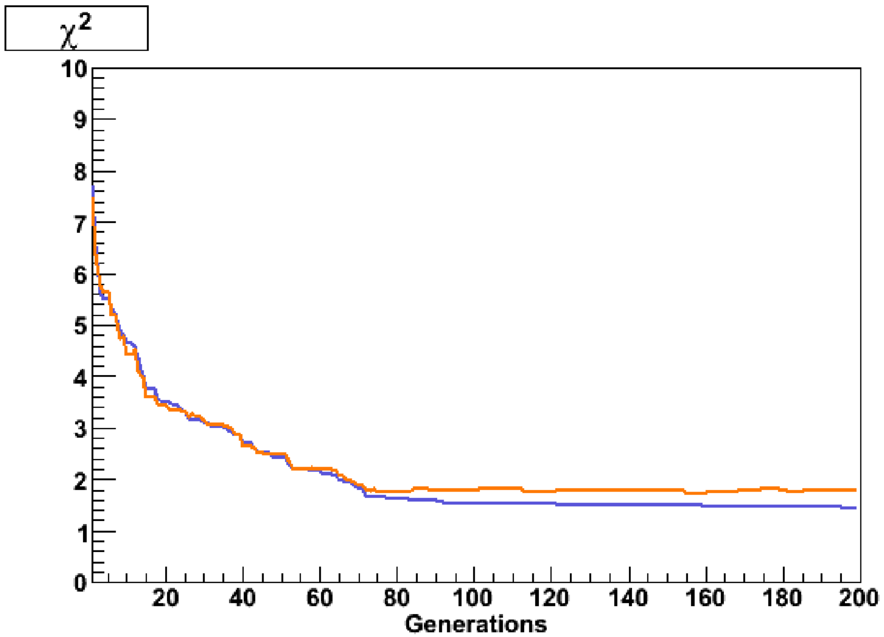


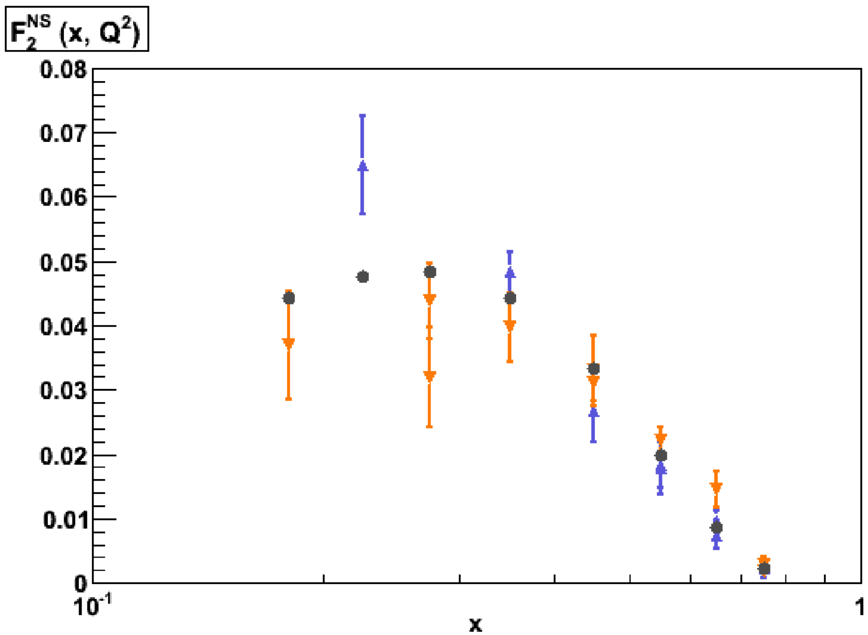
OPTIMAL FIT: CROSS-VALIDATION

GENETIC MINIMIZATION:
AT EACH GENERATION, χ^2 EITHER UNCHANGED OR DECREASING

- DIVIDE THE DATA IN TWO SETS: TRAINING AND VALIDATION
- MINIMIZE THE χ^2 OF THE DATA IN THE TRAINING SET
- AT EACH ITERATION, COMPUTE THE χ^2 FOR THE DATA IN THE VALIDATION SET (NOT USED FOR FITTING)
- WHEN THE VALIDATION χ^2 STOPS DECREASING, STOP THE FIT

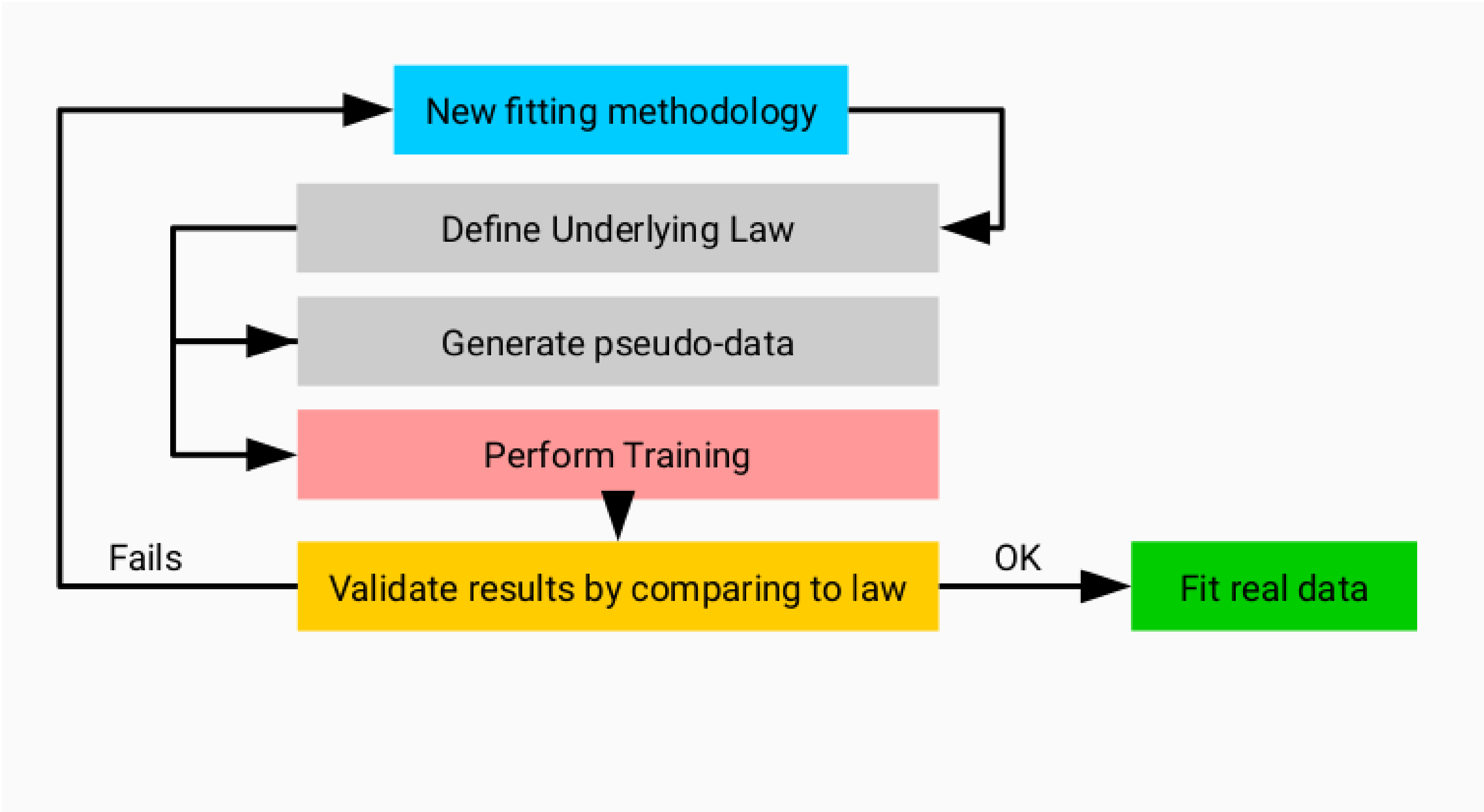
TOO LATE!





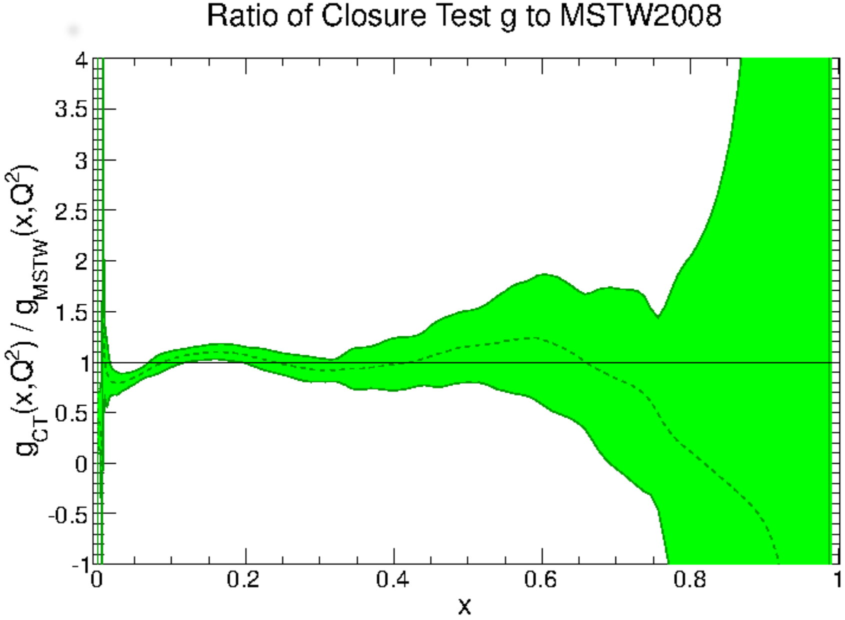
HOW DO WE KNOW THAT WE GOT THE RIGHT ANSWER?

CLOSURE TEST

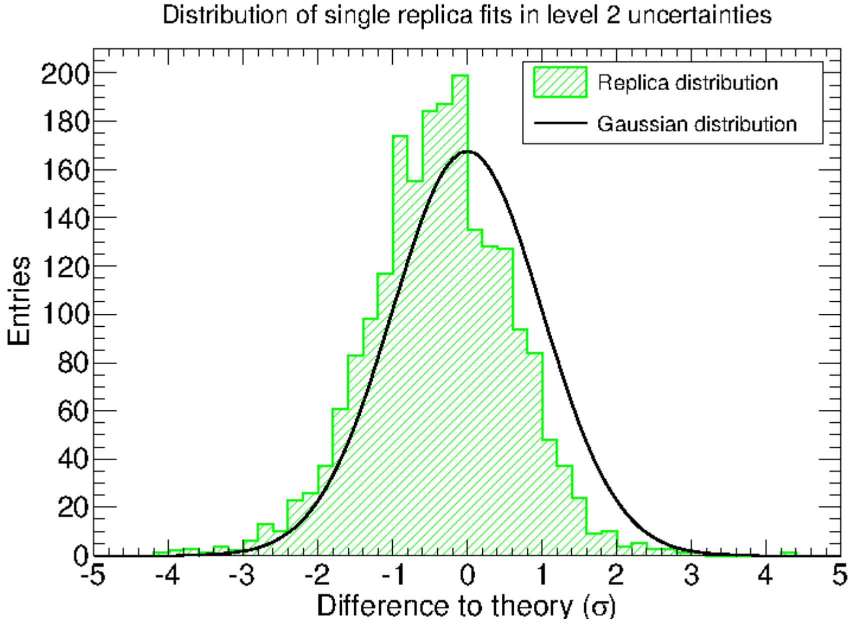


FIRST CLOSURE TEST (NNPDF3.0; 2014)

THE GLUON: RESULT/"TRUTH"



NORMALIZED DISTRIBUTION OF DEVIATIONS



1 σ: 70% (should be 68%)

- THE METHODOLOGY IS FAITHFUL

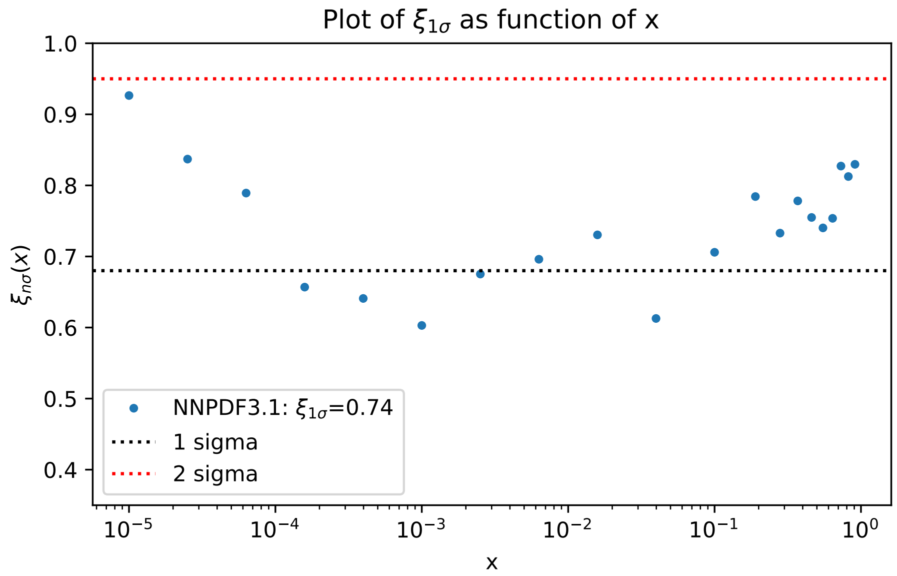
LEARNING THE METHODOLOGY

CLOSURE TEST: A CLOSER LOOK (NNPDF3.1)

ONE σ: ACTUAL/PREDICTED
FOR DATA, BY EXPERIMENT

experiment	NNPDF3.1 ratio
NMC	0.882828
SLAC	0.767063
BCDMS	0.730569
CHORUS	0.698907
NTVDMN	0.991090
HERACOMB	0.847359
HERAF2CHARM	1.867597
F2BOTTOM	1.124157
DYE886	0.655955
DYE605	0.585725
CDF	0.961652
D0	0.881199
ATLAS	0.904127
CMS	1.090241
LHCb	1.092194
Total	0.842168

ONE σ VALUE
FOR PDFS, VS x



- UNCERTAINTIES OVERESTIMATED
- 1 σ>68% AT VERY SMALL AND VERY LARGE x;
  1 σ <68% AT INTERMEDIATE x

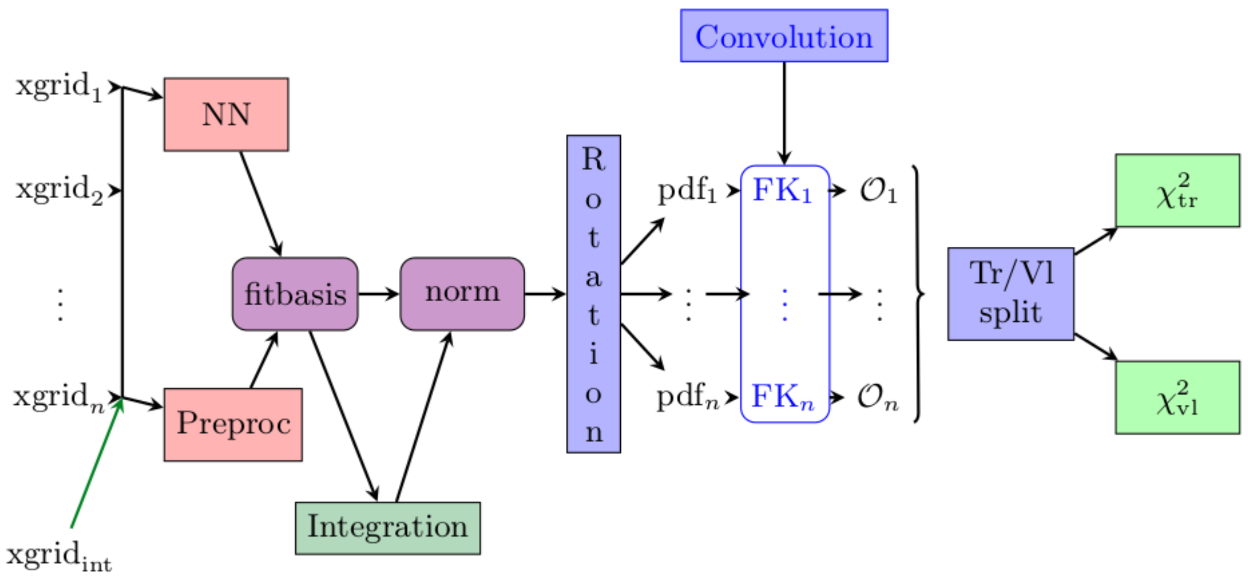
CAN WE DO BETTER?

THE N3FIT PROJECT

N3PDF Machine Learning • PDFs • QCD

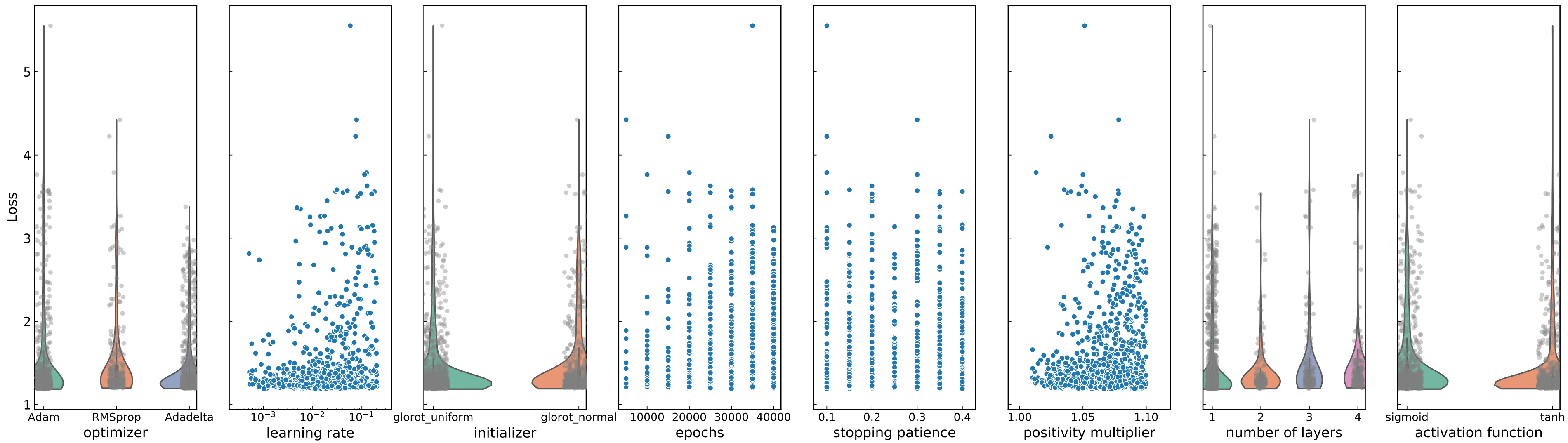
HOW DO WE KNOW THAT THE METHODOLOGY IS THE BEST?
"ACCUMULATED WISDOM" INEFFICIENT AND SLOW

CHANGE OF PHILOSOPHY ⇒ DETERMINISTIC MINIMIZATION (GRADIENT DESCENT)
GO FOR THE ABSOLUTE MINIMUM, AND (HYPER)OPTIMIZE



- PYTHON-BASED KERAS + TENSORFLOW FRAMEWORK
- EACH BLOCK INDEPENDENT LAYER
- CAN VARY ALL ASPECTS OF METHODOLOGY

FITTING THE METHODOLOGY
HYPEROPTIMIZATION SCANS



HYPEROPT PARAMETERS

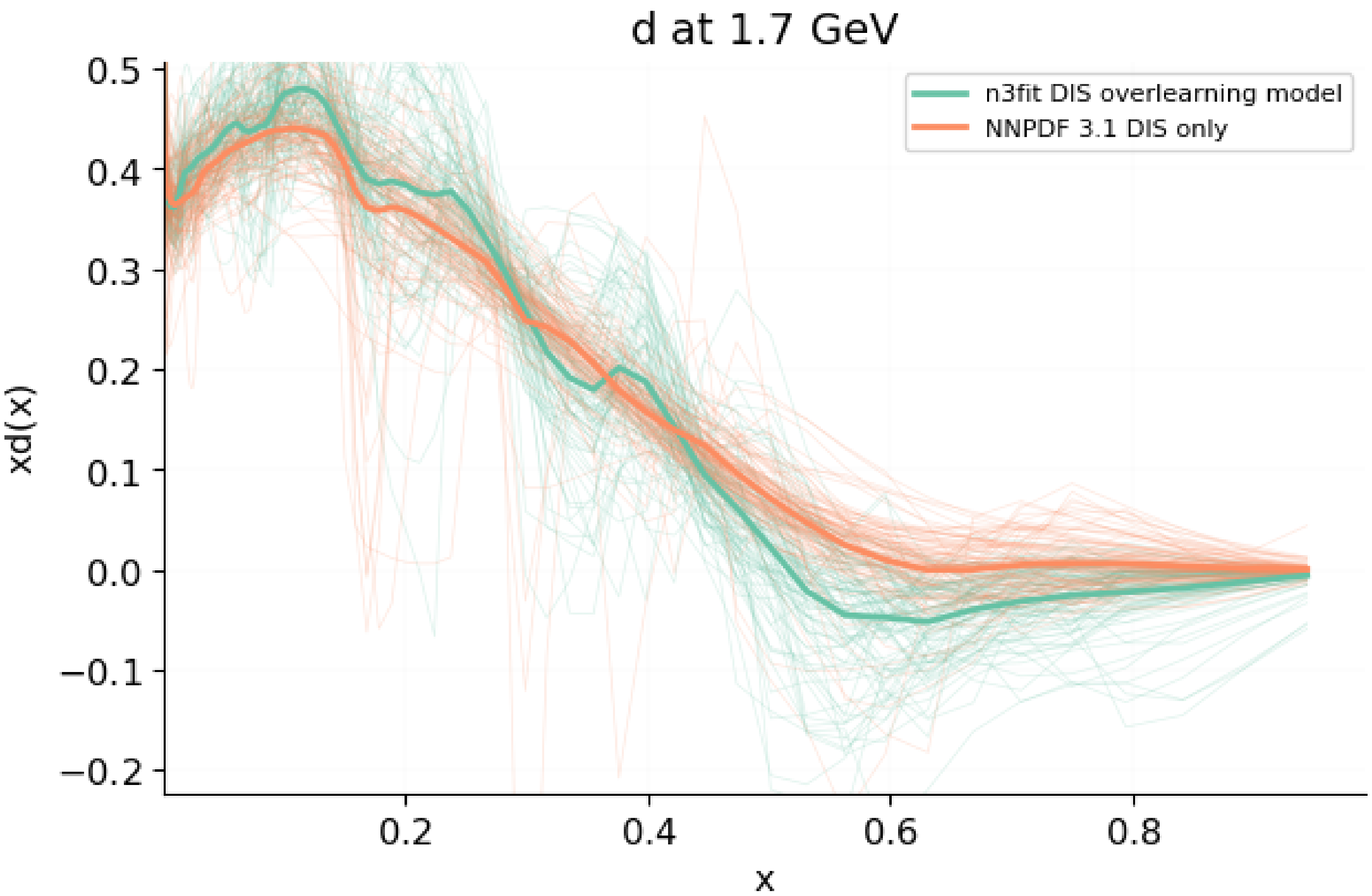
NEURAL NETWORK	FIT OPTIONS
NUMBER OF LAYERS (*)	OPTIMIZER (*)
SIZE OF EACH LAYER	INITIAL LEARNING RATE (*)
DROPOUT	MAXIMUM NUMBER OF EPOCHS (*)
ACTIVATION FUNCTIONS (*)	STOPPING PATIENCE (*)
INITIALIZATION FUNCTIONS (*)	POSITIVITY MULTIPLIER (*)

- SCAN PARAMETER SPACE
- OPTIMIZE FIGURE OF MERIT: VALIDATION χ^2
- BAYESIAN UPDATING

FITTING THE METHODOLOGY

THE OVERFITTING PROBLEM

DOWN QUARK: HYPEROPTIMIZED VS. STANDARD



- NNPDF3.1: WIGGLES: FINITE SIZE $\Rightarrow$ WILL GO AWAY AS N_{rep} GROWS
- N3FIT: WIGGLY PDFS $\Leftrightarrow$ OVERFITTING $\Rightarrow$ WILL NOT GO AWAY ($\chi^2_{\mathrm{train}} \ll \chi^2_{\mathrm{valid}}$!!)

WHAT HAPPENED?

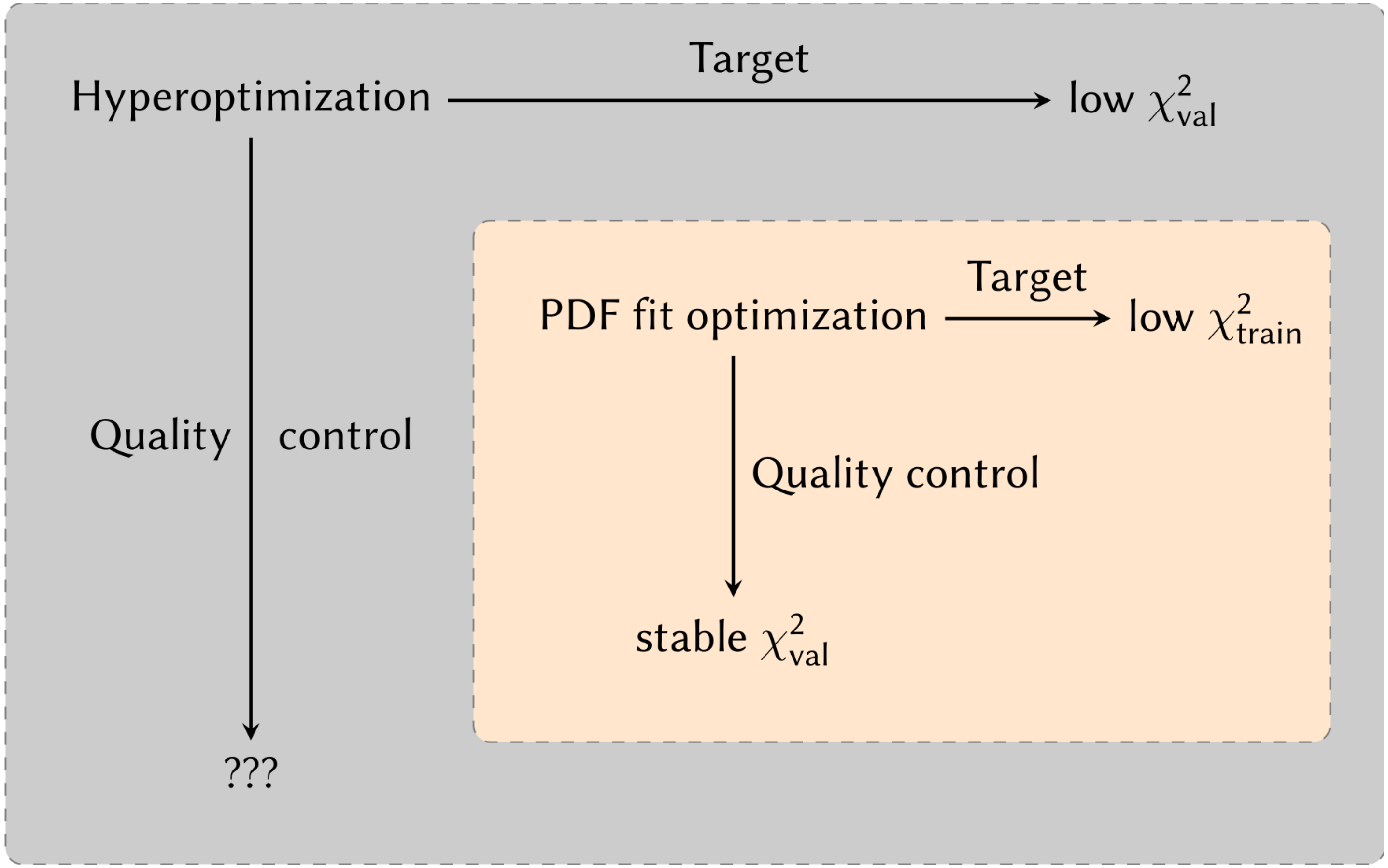
OPTIMIZATION

$$\begin{tikzcd}
\text{PDF fit optimization} \arrow[r, "\text{Target}"] \arrow[d, "\text{Quality control}"] & \text{low } \chi^2_{\text{train}} \\
\text{stable } \chi^2_{\text{val}} &
\end{tikzcd}$$

CROSS-VALIDATION SELECTS THE OPTIMAL MINIMUM

WHAT HAPPENED?

HYPEROPTIMIZATION

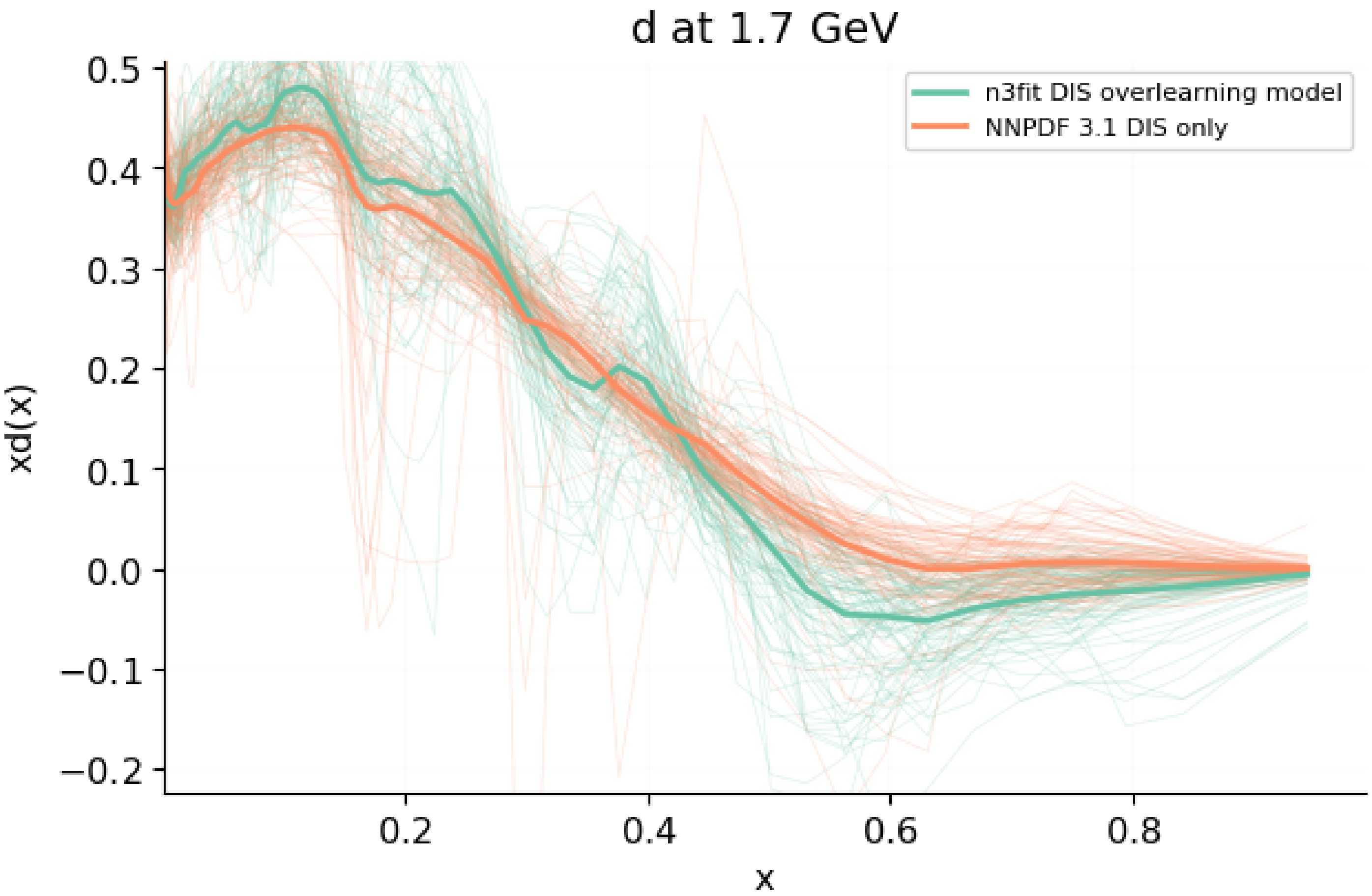


WE ARE MISSING A SELECTION CRITERION

FITTING THE METHODOLOGY

THE OVERFITTING PROBLEM

DOWN QUARK: HYPEROPTIMIZED VS. STANDARD

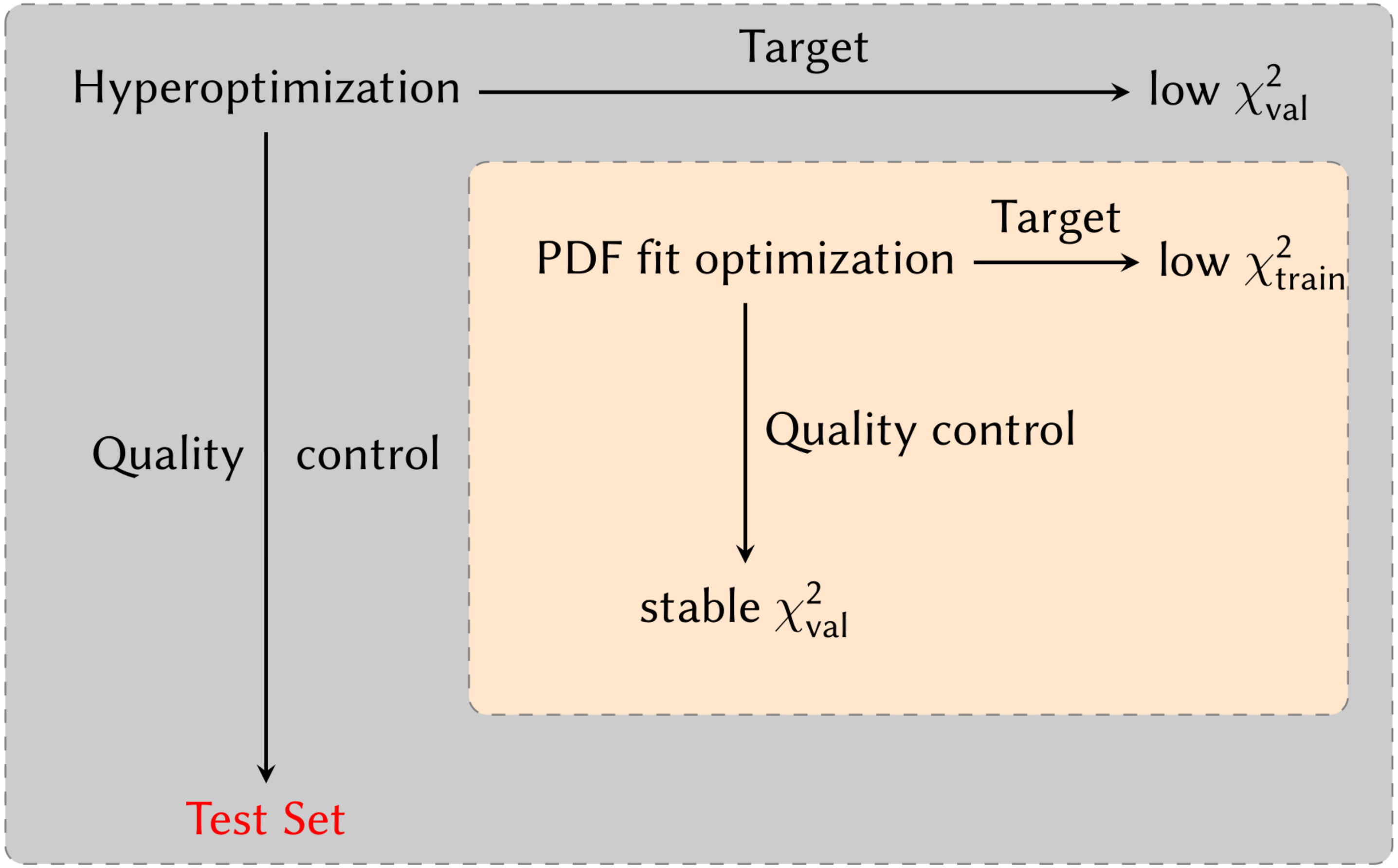


- NNPDF3.1: WIGGLES: FINITE SIZE $\Rightarrow$ WILL GO AWAY AS $N_{\rm rep}$ GROWS
- N3FIT: WIGGLY PDFS $\Leftrightarrow$ OVERFITTING $\Rightarrow$ WILL NOT GO AWAY ($\chi^2_{\rm train} \ll \chi^2_{\rm valid}$!!)
- CORRELATIONS BETWEEN TRAINING AND VALIDATION DATA

MACHINE LEARNING

THE SOLUTION

TUNED HYPEROPTIMIZATION



COMPARE TO A A TEST SET (NEW SET OF DATA PREVIOUSLY NOT USED AT AL)
TESTS GENERALIZATION POWER

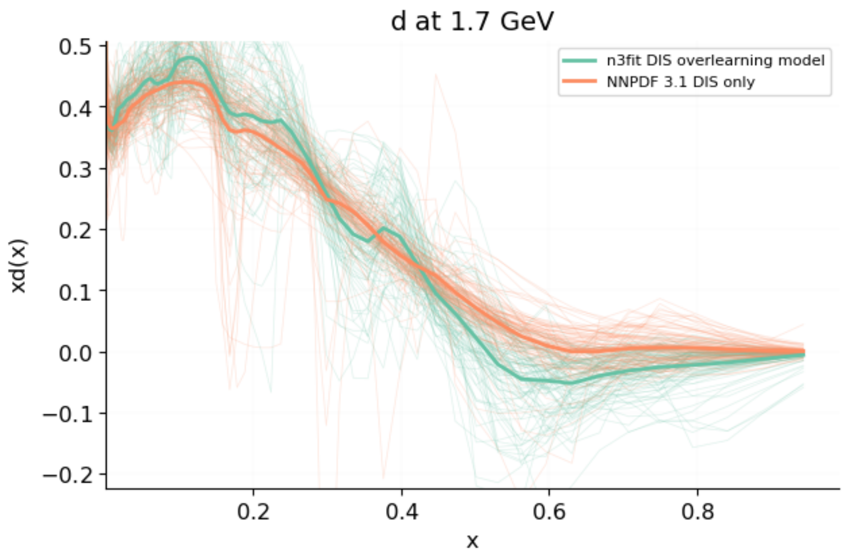
THE TEST SET METHOD

- COMPLETELY UNCORRELATED TEST SET
- OPTIMIZE ON WEIGHTED AVERAGE OF VALIDATION AND TEST
  $\Rightarrow$ NO OVERLEARNING

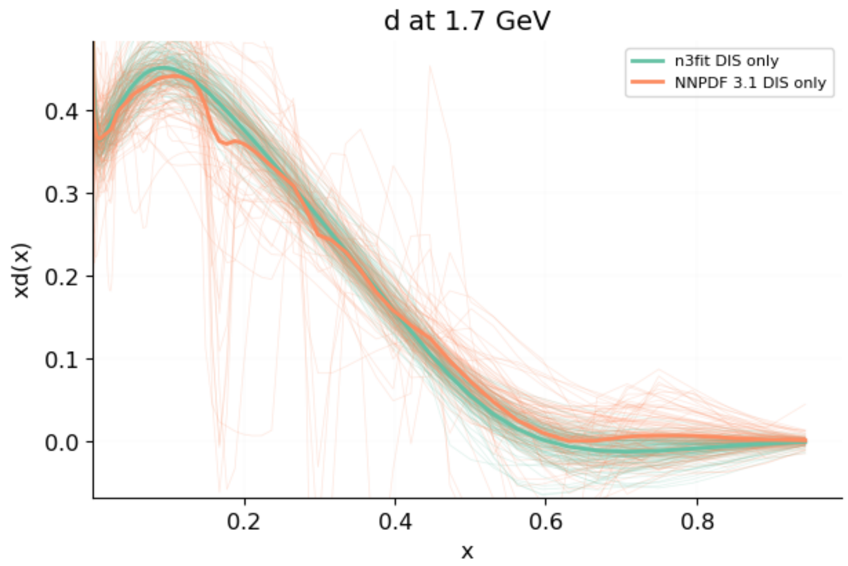
HYPEROPTIMIZED PDFs

DOWN QUARK

N3 OVERFIT VS NNPDF3.1



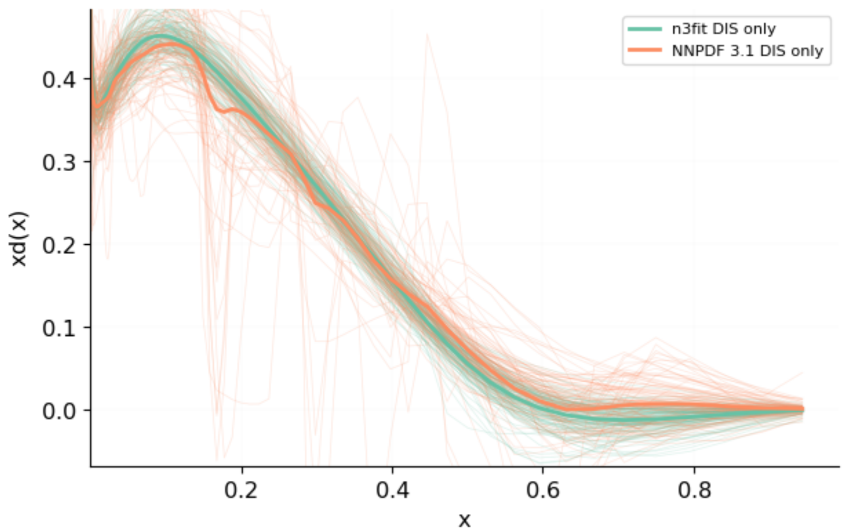
N3FIT VS NNPDF3.1



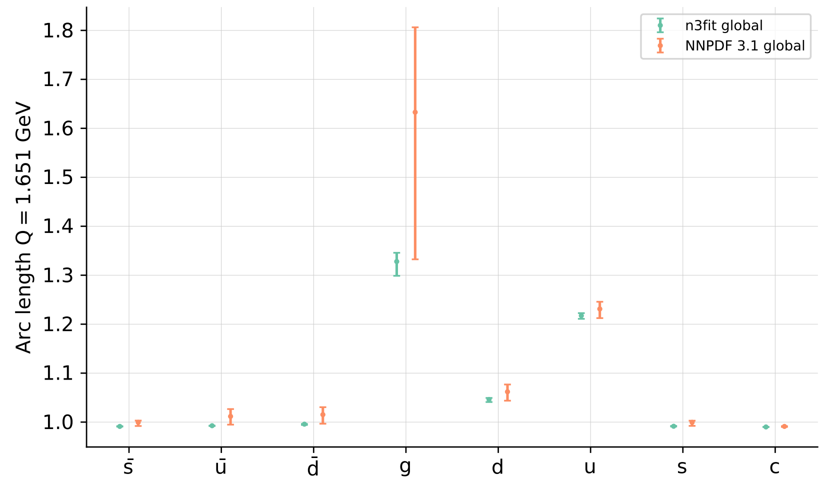
THE TEST SET METHOD

N3FIT VS NNPDF3.1

DOWN PDF



ARCLENGTHS



- NO OVERFITTING
- COMPARED TO NNPDF3.1
  - MUCH GREATER STABILITY ⇒ FEWER REPLICAS FOR EQUAL ACCURACY
  - UNCERTAINTIES SOMEWHAT REDUCED

CLOSURE TESTS AGAIN

NEW METHODOLOGY ⇒ LARGE NUMBER OF "RUNS OF THE UNIVERSE"

- UNCERTAINTIES ON PREDICTIONS: FAITHFUL AT 5% LEVEL
- UNCERTAINTIES ON PDFS σ
  - COMPUTED IN DIAGONAL x-SPACE BASIS IN DATA REGION
  - FAITHFUL AT 10% LEVEL ON AVERAGE, & FOR SINGLET, GLUON, TOTAL AND TRIPLET VALENCE

ONE σ: ACTUAL/PREDICTED

FOR DATA, BY EXPERIMENT

experiment	NNPDF3.1 ratio	n3fit ratio
NMC	0.882828	0.843427
SLAC	0.767063	0.690118
BCDMS	0.730569	0.770704
CHORUS	0.698907	0.734656
NTVDMN	0.991090	0.797017
HERACOMB	0.847359	1.326333
HERAF2CHARM	1.867597	3.566076
F2BOTTOM	1.124157	1.532634
DYE886	0.655955	0.857915
DYE605	0.585725	0.870151
CDF	0.961652	0.779424
D0	0.881199	1.015202
ATLAS	0.904127	1.132229
CMS	1.090241	1.017136
LHCb	1.092194	0.993525
Total	0.842168	0.940737

FOR PDFs, EVOLUTION BASIS

flavour	bootstrap mean $\sqrt{\frac{\mathbf{E}_\eta[\text{bias}]}{\mathbf{E}_\eta[\text{variance}]}}$
Σ	0.90
gluon	0.90
V	1.02
V3	0.99
V8	0.91
T3	0.62
T8	1.31
Total	0.92

INTO THE UNKNOWN

THE CHALLENGE OF MACHINE LEARNING:

- WHAT IS THE UNCERTAINTY WHERE THERE IS NO DATA?
- WHAT IS THE UNCERTAINTY WHERE THERE IS NO THEORY?

THE METHDOLOGY IS AUTOMATICALLY TESTED, BUT....

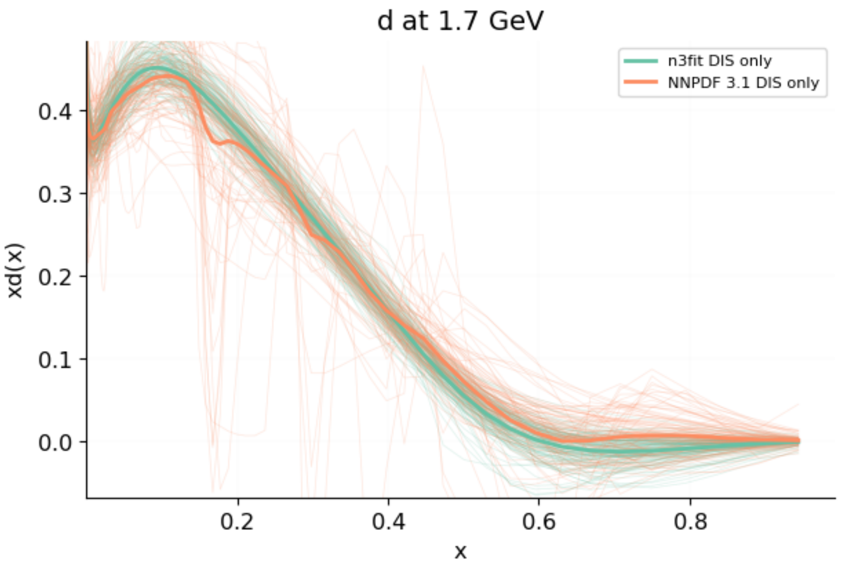
- WHO PICKS THE TEST SET?
- HOW DO WE KNOW THAT THE GENERALIZATION IS FAITHFUL?

AUTOMATIC GENERALIZATION
K-FOLDINGS
THE BASIC IDEA:

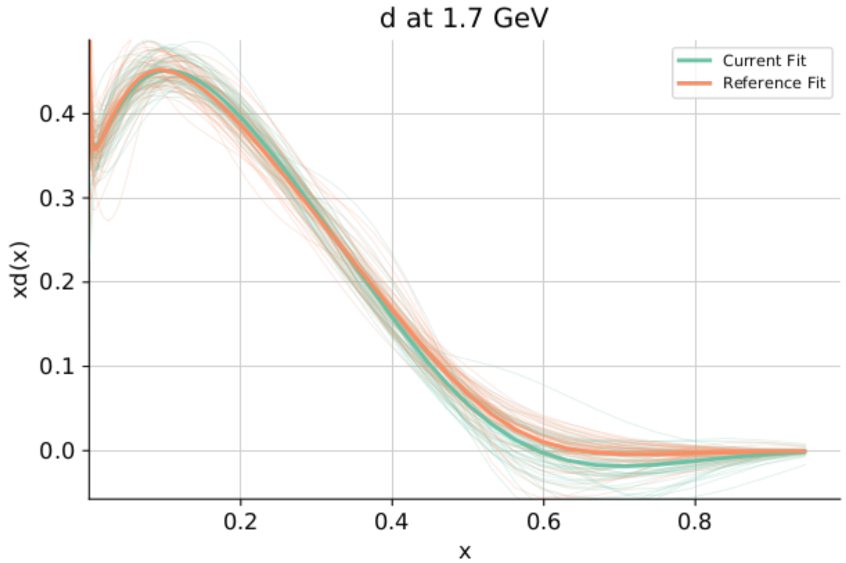
- DIVIDE THE DATA INTO n REPRESENTATIVE SUBSETS
  EACH CONTAINING PROCESS TYPES, KINEMATIC RANGE OF FULL SET
- FIT $n-1$ SETS AND USE n-TH SET AS TEST
  $\Rightarrow n$ VALUES OF $\chi^2_{\text{test, i}}$
- HYPEROPTIMIZE ON MEAN AND STANDARD DEVIATION OF $\chi^2_{\text{test, i}}$
  $\rightarrow$ GOOD & STABLE GENERALIZATION

FOLDED PDFS
DOWN QUARK

N3FIT VS NNPDF3.1



N3FIT-K VS. N3FIT

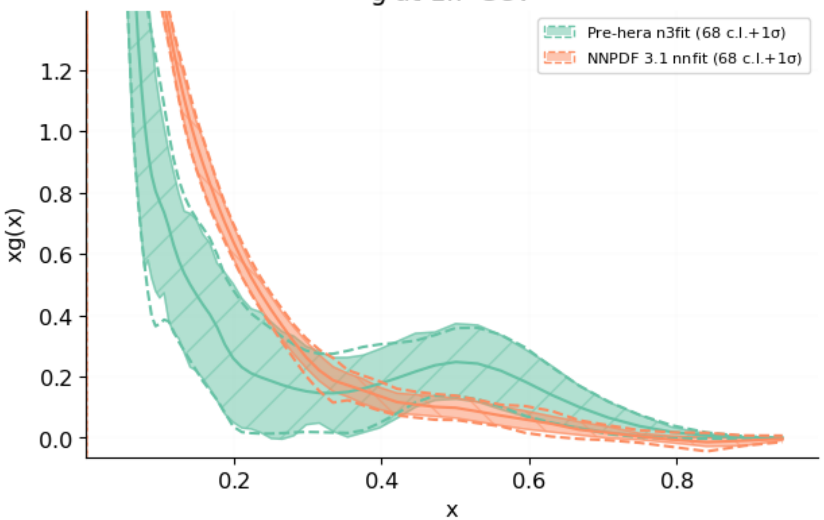


DOES IT WORK?: THE "FUTURE TEST"

COULD WE "PREDICT" THE RISE OF F_2 AT HERA?

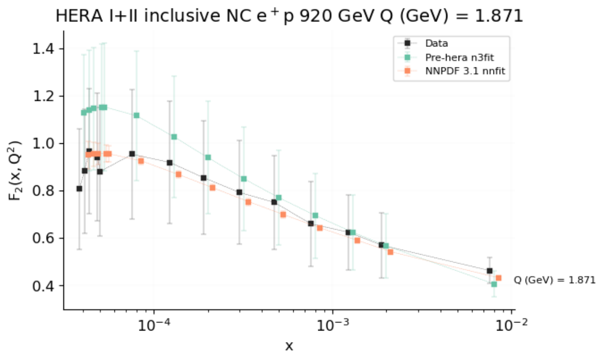
FIT PDFS TO PRE-HERA DATA ONLY

PREDICTED VS TRUE GLUON

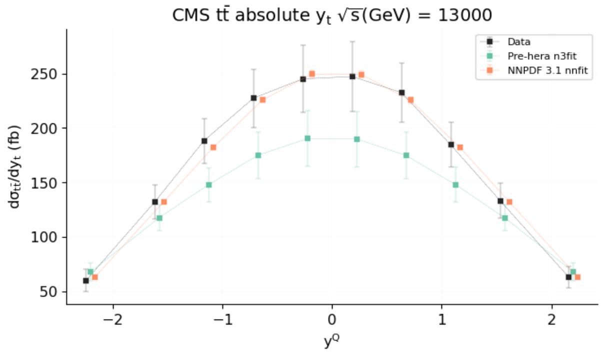


PREDICTION COMPARED TO DATA

HERA F_2



CMS TOP



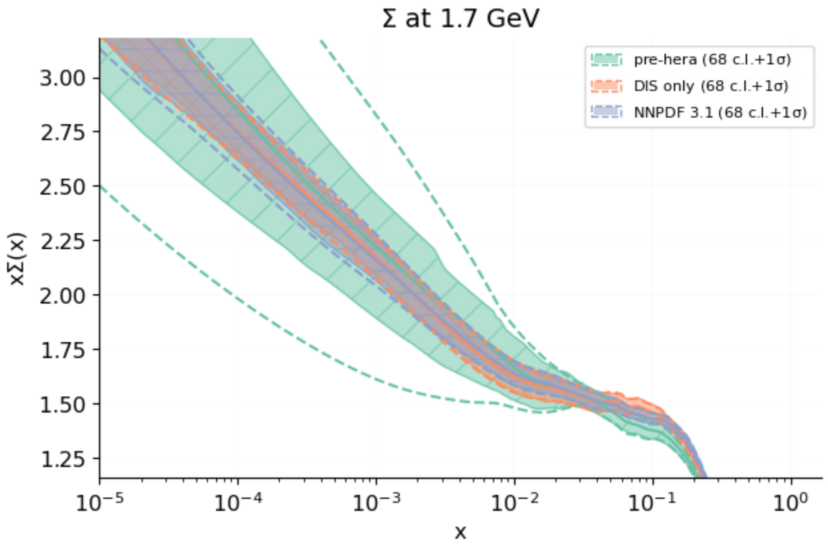
- PDFS ARE FUTURE-COMPATIBLE
- THE DATA ARE WITHIN SHRINKING UNCERTAINTIES
- PREDICTED χ^2/dat=1.20 (WITH PDF UNCERTAINTIES), COMPARE TO FITTED χ^2/dat=1.16 (WITHOUT UNCERTAINTIES)

DOES IT WORK?: THE "FUTURE TEST"

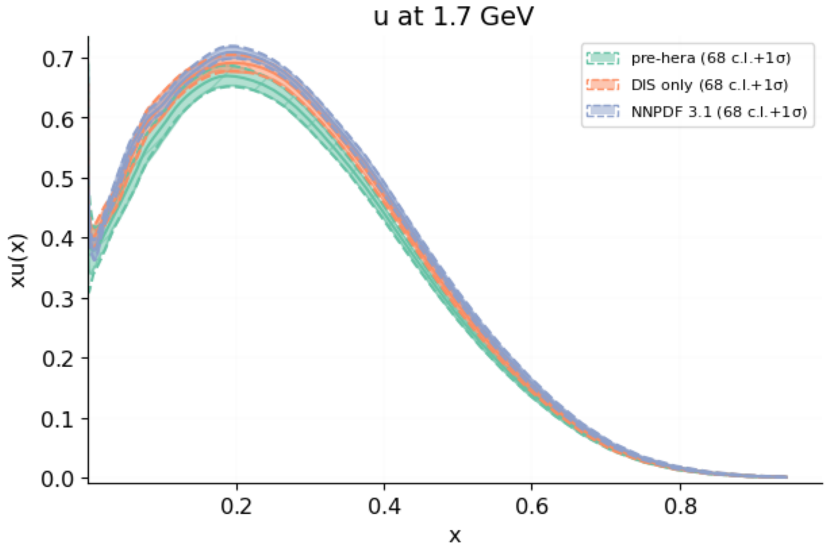
SEQUENTIAL FUTURE TEST DATASETS:

- PRE-HERA
- POST-HERA, PRE-LHC
- LHC RUN I (NNPDF3.1)

QUARK SINGLET



UP QUARK

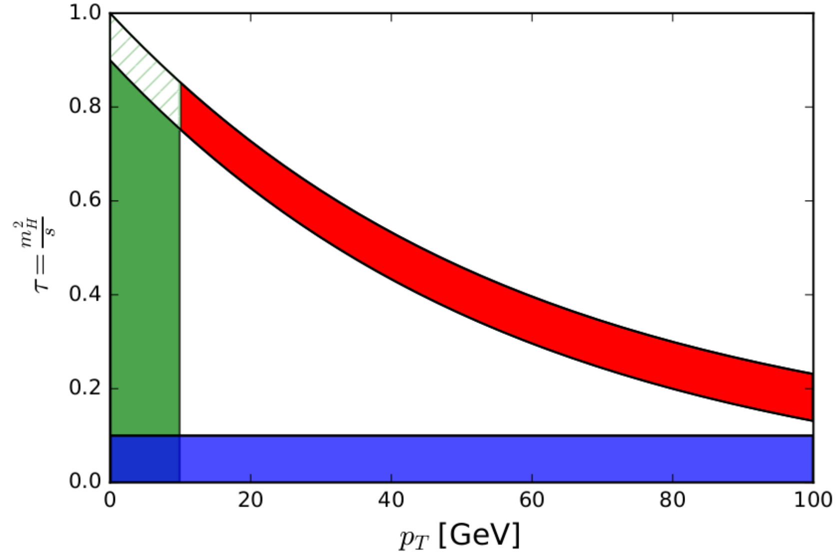


- PDFs ARE FUTURE-COMPATIBLE
- GENERALIZATION FAITHFUL

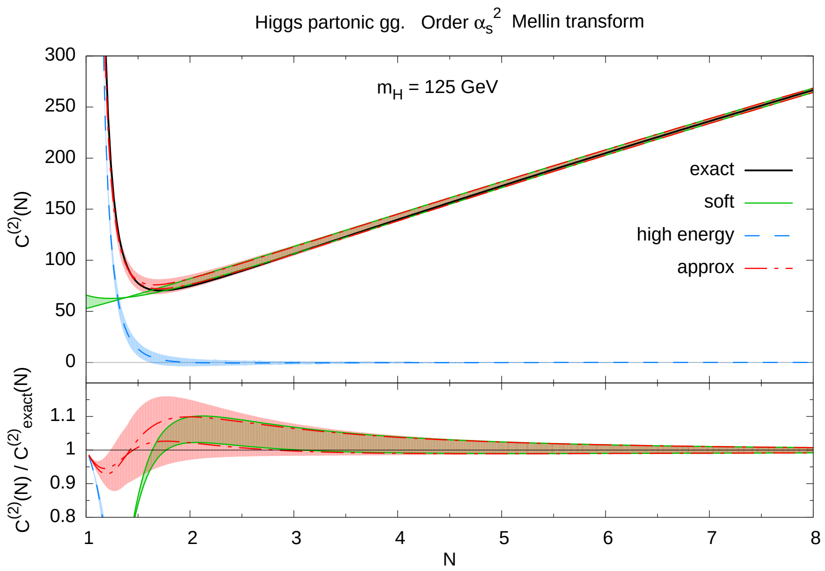
MISSING HIGHER ORDERS FROM ASYMPTOTICS

- HIGHER ORDERS KNOWN IN VARIOUS KINEMATIC LIMITS FROM RESUMMATION
- USED IN THE PAST TO CONSTRUCT ANALYTIC APPROXIMATION TO FULL MHO: E.G. HIGGS IN GLUON FUSION AT N^{3}LO
- MACHINE LEARNING MHO?

(τ, p_T) RESUMMATION REGIONS



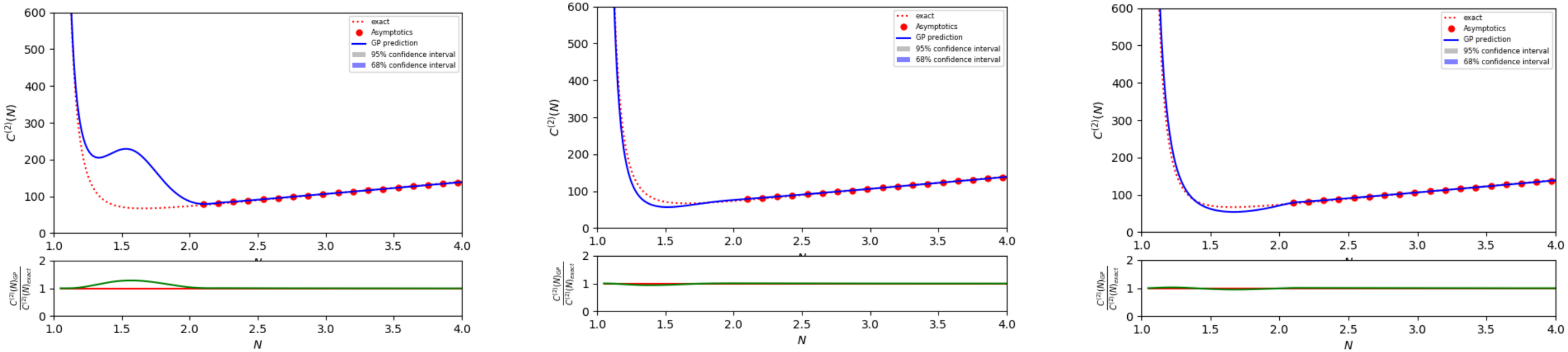
NNLO N-SPACE GGHIGGS
ANALYTIC APPROX VS. EXACT



NAIVE IDEA: GAUSSIAN PROCESS

- PROPAGATE ASYPTOTICS INTO "CENTRAL" REGION USING "GAUSSIAN PROCESS":
  - ASSUME $\sigma(x)$ MULTIGAUSSIAN IN FUNCTION SPACE
  - DETERMINE THE CORRELATION IN KNOWN REGION ASSUMING KERNEL
  - DETERMINE CONDITIONAL DISTRIBUTION IN EXTRAPOLATION
- HYPEROPTIMIZE KERNEL CHOICE AND PARAMETERS BASED ON KNOWN CASES

NNLO N-SPACE GGHIGGS: GAUSSIAN KERNEL INTERPOLATIONS



- TOO FEW DATA $\Rightarrow$ RESULTS UNSTABLE, DEPEND ON CHOICE OF KERNEL

THEORY UNCERTAINTIES

TRANSFER LEARNING?

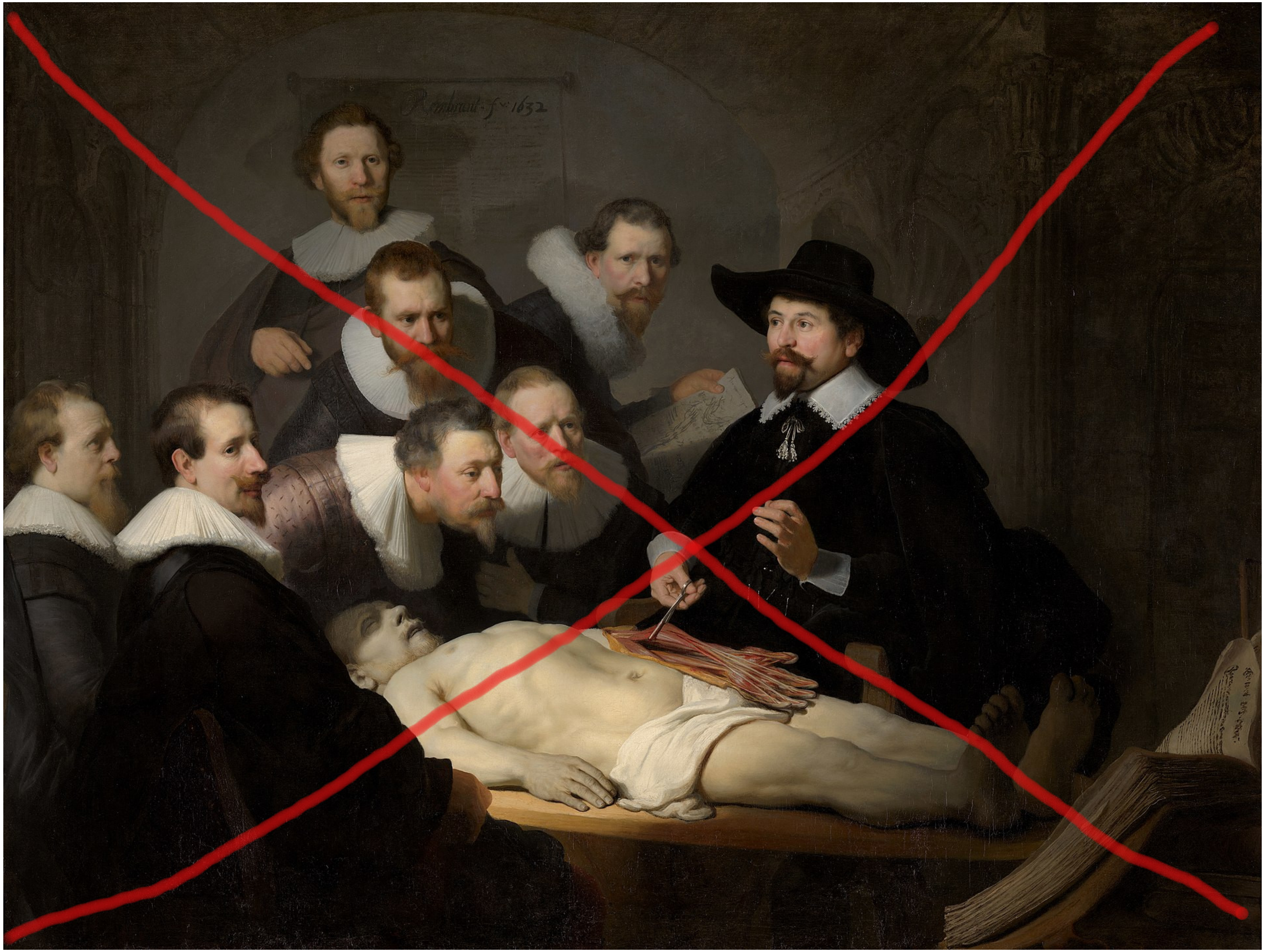
THE BASIC IDEA:

- PERTURBATIVE DEPENDENCE KNOWN UP TO NNLO FOR MANY PROCESSES
- LEARN PERTURBATIVE DEPENDENCE FROM KNOWN CASES
- ADD FINAL LAYER WHICH EXTRAPOLATES FROM ASYMPTOTICS

....STAY TUNED!

ANATOMY?

ANATOMY?



ANATOMY!