

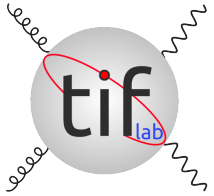



# PDF DETERMINATION AS MACHINE LEARNING

STEFANO FORTE
UNIVERSITÀ DI MILANO & INFN



UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI FISICA







MACHILE LEARNING AT GGI

FLORENCE, SEPTEMBER 22, 2022

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

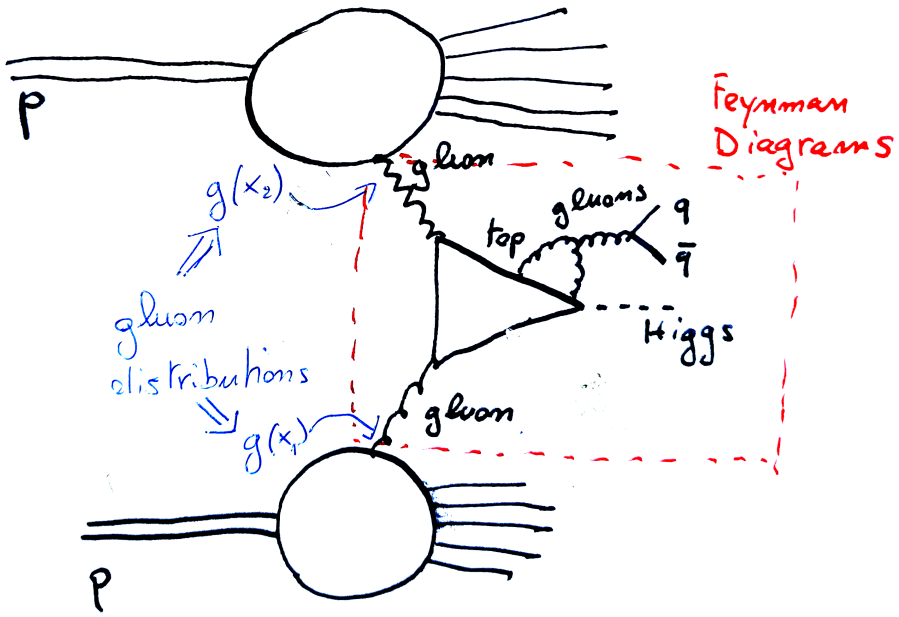
# SUMMARY

- A PDF PRIMER
  - PDFs AND DATA
  - UNCERTAINTIES: DATA AND EXTRAPOLATION
- THE PROBLEM OF PDF UNCERTAINTIES
  - BACKWARD AND FORWARD COMPATIBILITY
  - METHODOLOGY AND THEORY BIAS
  - CLOSURE TESTS
- THE NNPDF METHODOLOGY
  - THE FUNCTIONAL MONTECARLO
  - NEURAL NETS AND GENETIC MINIMIZATION
- MACHINE LEARNINING PDFs
  - CODE STRUCTURE
  - HYPEROPTIMIZATION
  - $K$-FOLDING
- VALIDATION
  - CLOSURE TESTS
  - FUTURE TESTS
- DELIVERY
  - GENETIC COMPRESSION
  - GAN OPTIMIZATION
- FUTURE DEVELOPMENTS
  - FEATURE SCALING
  - OVERLEARNING METRICS

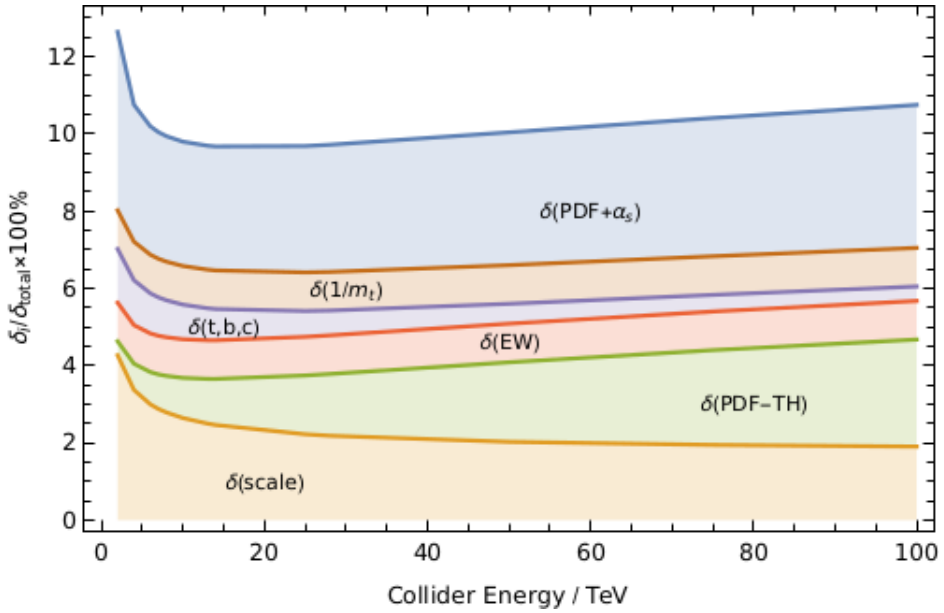
# PDFs: WHAT ARE THEY?

# PDFs AND THEIR UNCERTAINTIES

## QCD FACTORIZATION




## UNCERTAINTIES: HIGGS IN GLUON FUSION

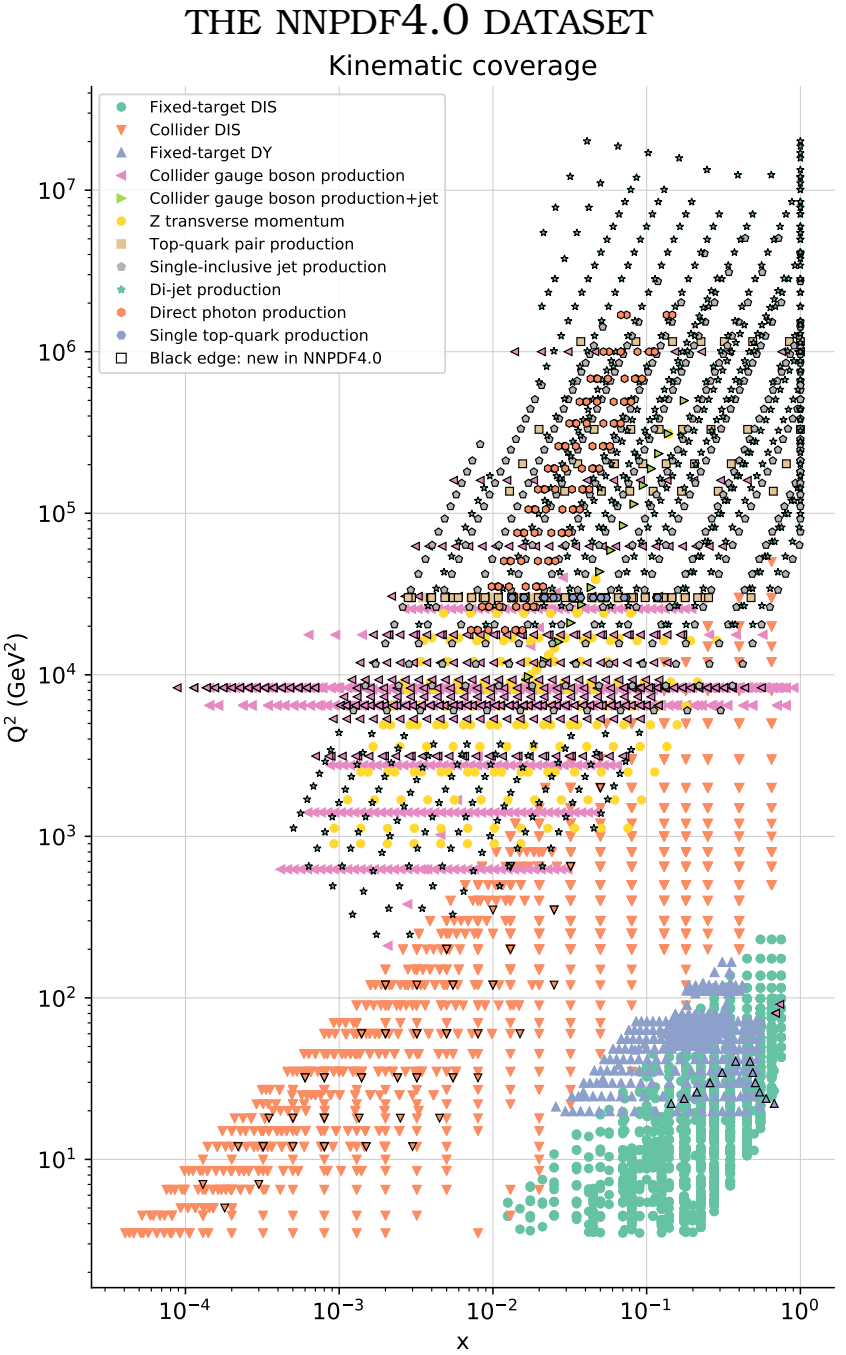



(HL-LHC Higgs WG report, 2019)

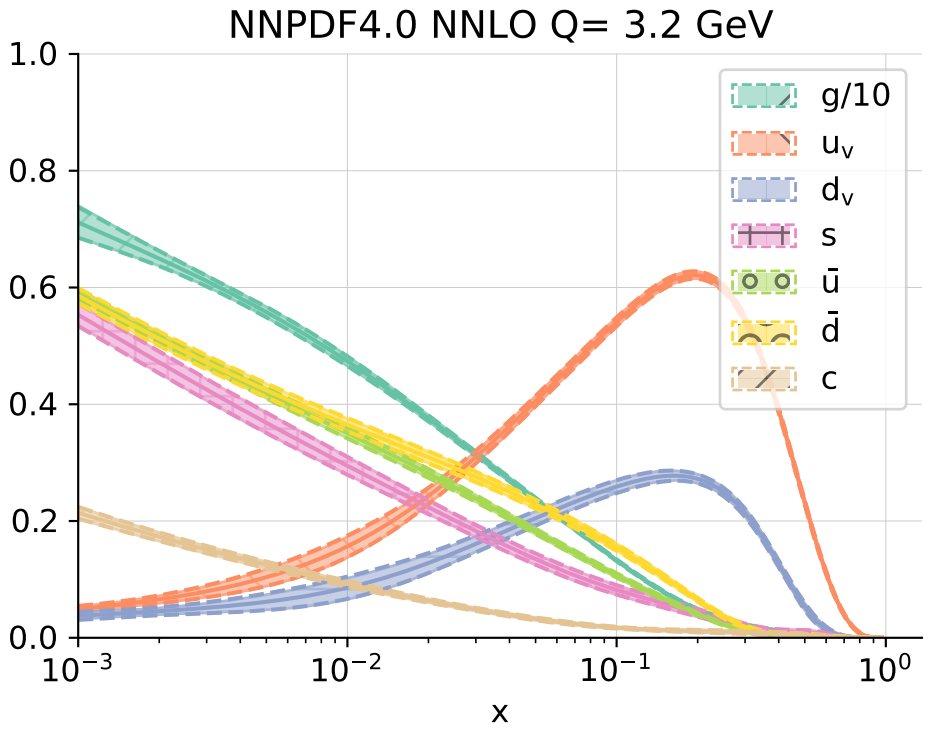
- PDF ESPRESS THE LIKELIHOOD OF A QUARK OR GLUONS (PARTONS) TO ENTER A COLLISION
- THEIR KNOWLEDGE IS REQUIRED FOR THE COMPUTATION OF ANY PROCESS AT THE LHC
- THEIR KNOWLEDGE IS A DOMINANT SOURCE OF UNCERTAINTY

# 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 $\sigma$ TO DATA
  - ABOUT 4600 DATAPOINTS
  - LEPTOPRODUCTION & HADROPRODUCTION, COLLIDER & FIXED-TARGET



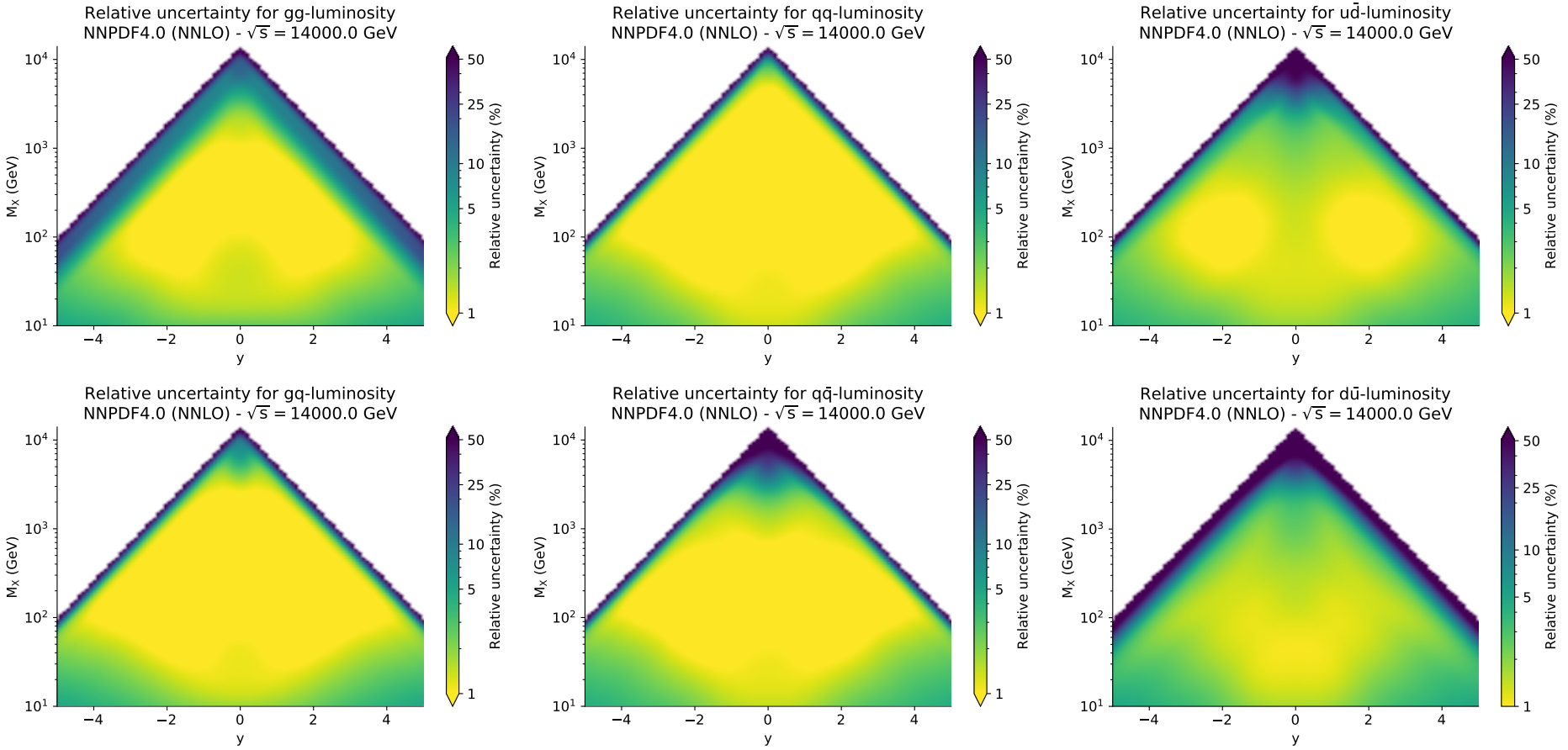
# PDFs: THE STATE OF THE ART (NNPDF4.0, 2021)




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

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

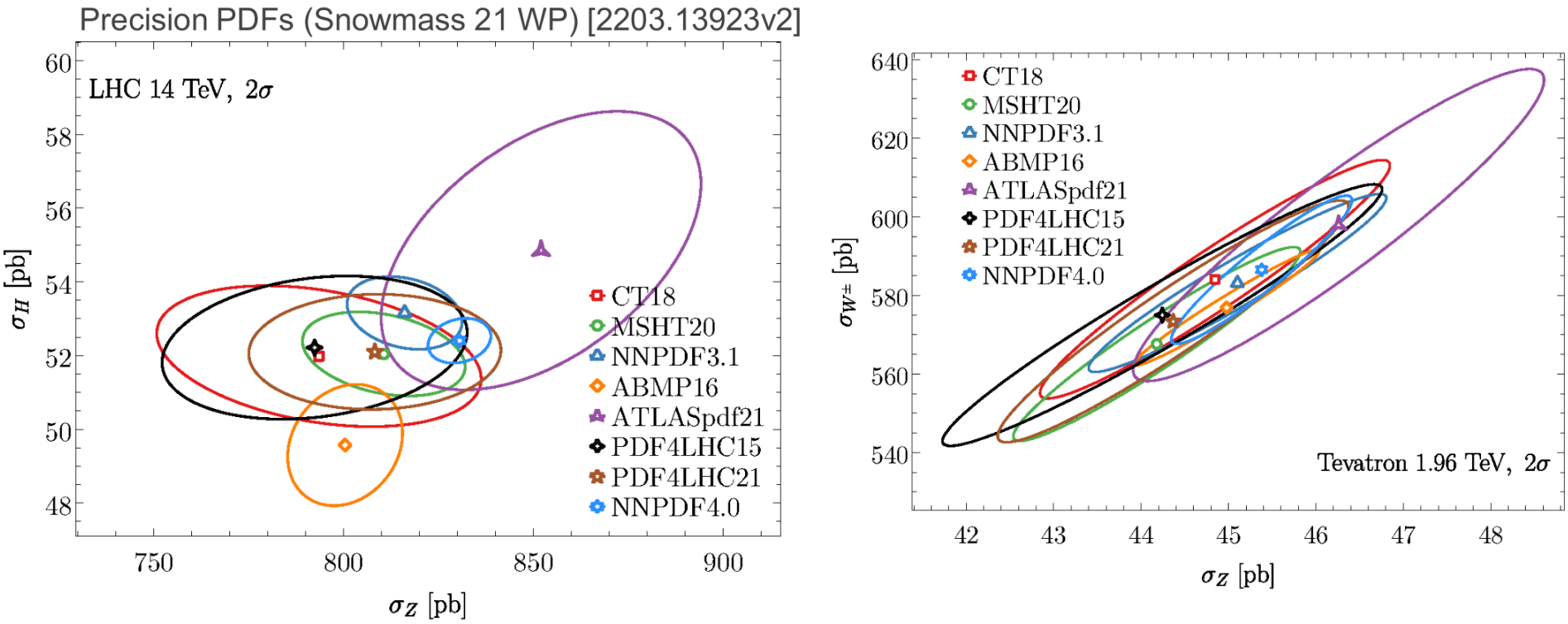
GLUON SINGLET FLAVORS




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

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

Higgs, $W$, $Z$ production

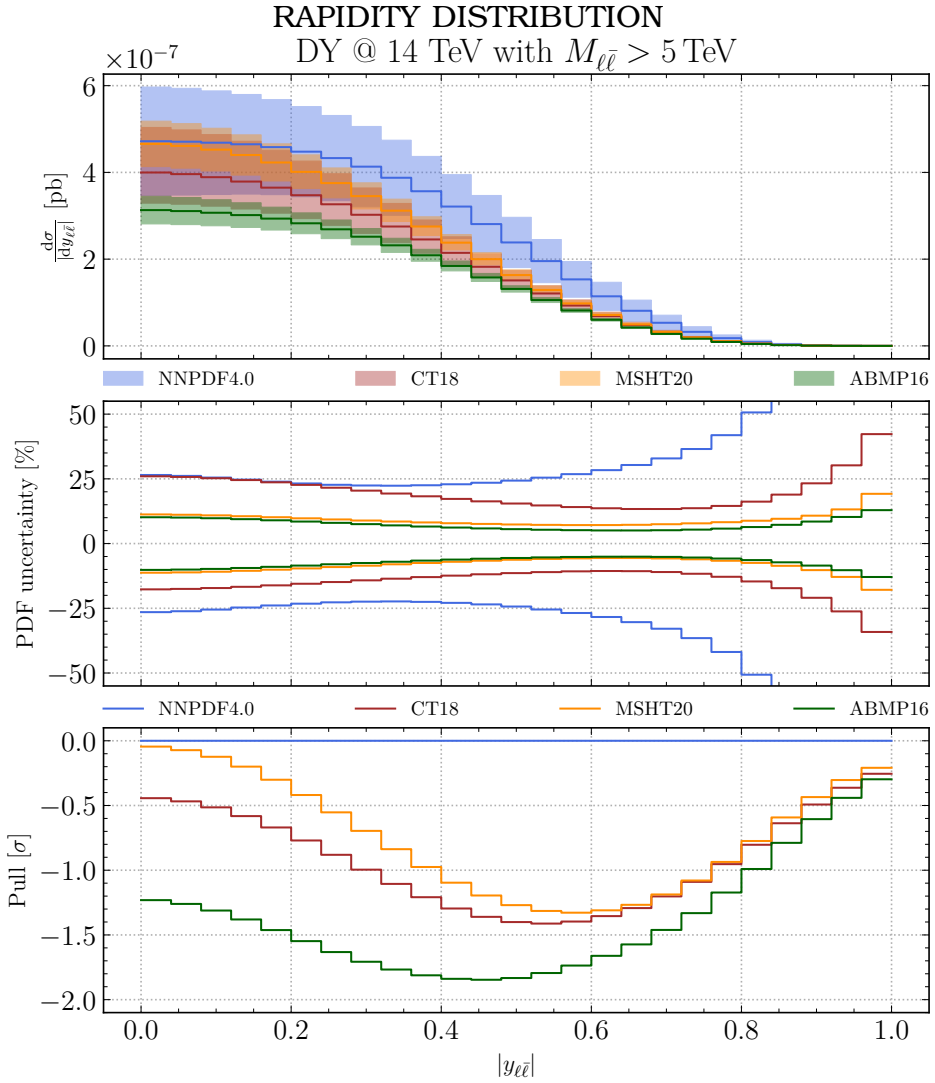


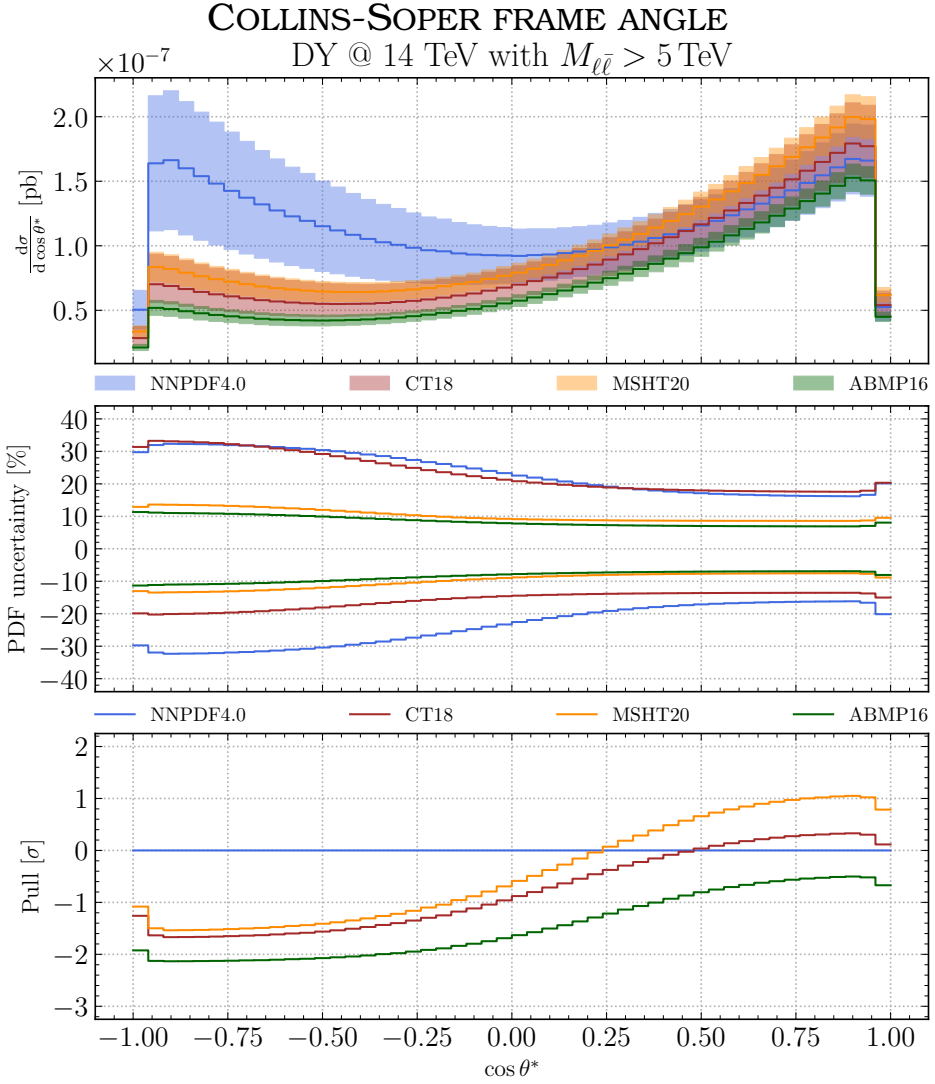

DATA REGION: NNPDF4.0 UNCERTAINTIES $\Rightarrow$ SMALLER THAN OTHER PDF SETS

BACKWARD CONSISTENT WITH NNPDF3.1

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

Z' production




EXTRAPOLATION REGION: NNPDF4.0 UNCERTAINTIES $\Rightarrow$ LARGER THAN OTHER PDF SETS

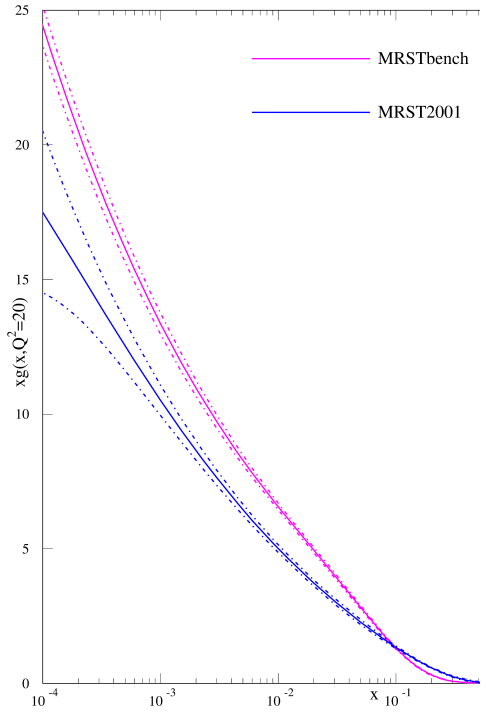
UNBIASED PDF BEHAVIOR

# UNCERTAINTY ESTIMATION

# 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

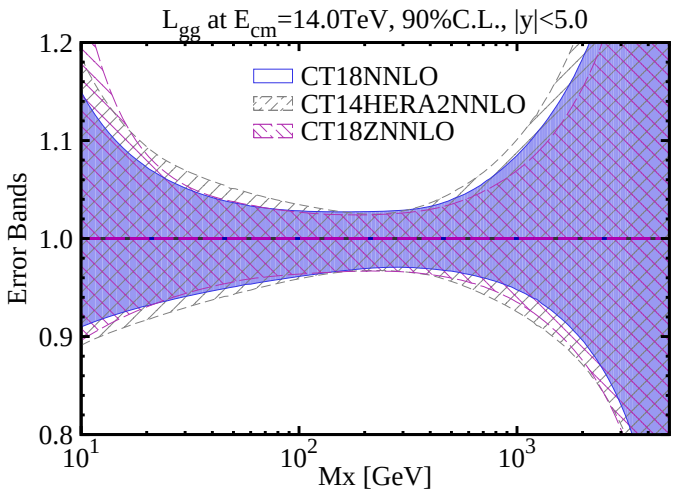



"...the partons extracted using a very limited data set are completely incompatible, even allowing for the uncertainties, with those obtained from a global fit with an identical treatment of errors...The comparison illustrates the problems in determining the true uncertainty on parton distributions."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

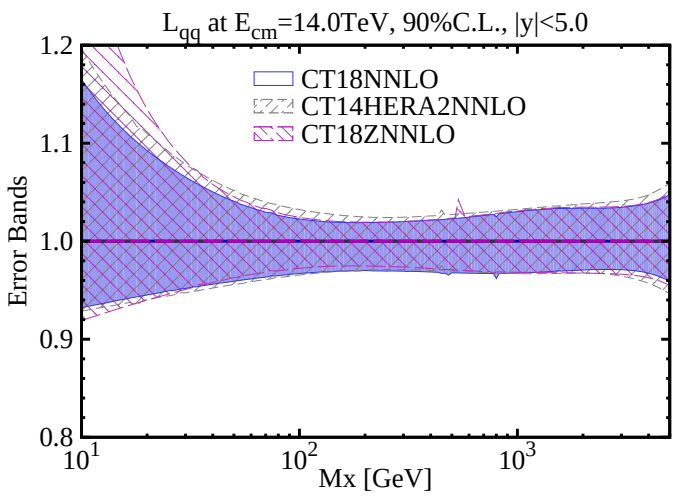
# THE PDF UNCERTAINTY PROBLEM:
# THE PRE-ML APPROACH

## CT18 VS. CT14: PARTON LUMINOSITY UNCERTAINTIES

GLUON-GLUON




QUARK-QUARK



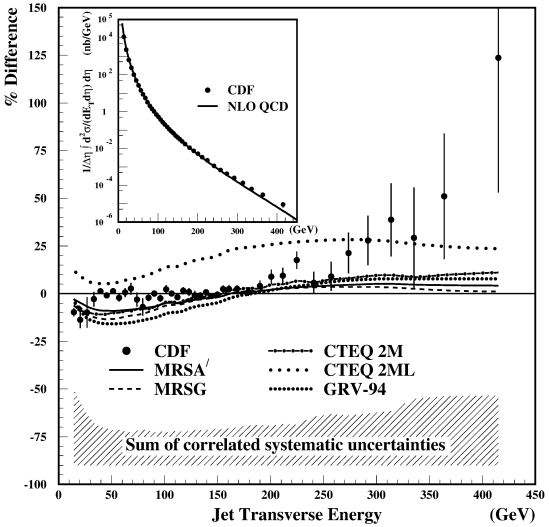

## MORE DATA $\Rightarrow$ BIGGER UNCERTAINTIES (?!)

## PARTON PARAMETRIZATIONS

- CTEQ5 2002: $xg(x,Q_0^2) = A_0 x^{A_1}(1-x)^{A_2}(1+A_3 x^{A_4})$
- 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: $g(x,Q=Q_0) = x^{a_1-1}(1-x)^{a_2}\left[a_3(1-y)^3 + a_4 3y(1-y)^2 + a_5 3y^2(1-y) + y^3\right]$; $y=\sqrt{x}$; $a_5 = (3+2a_1)/3$.

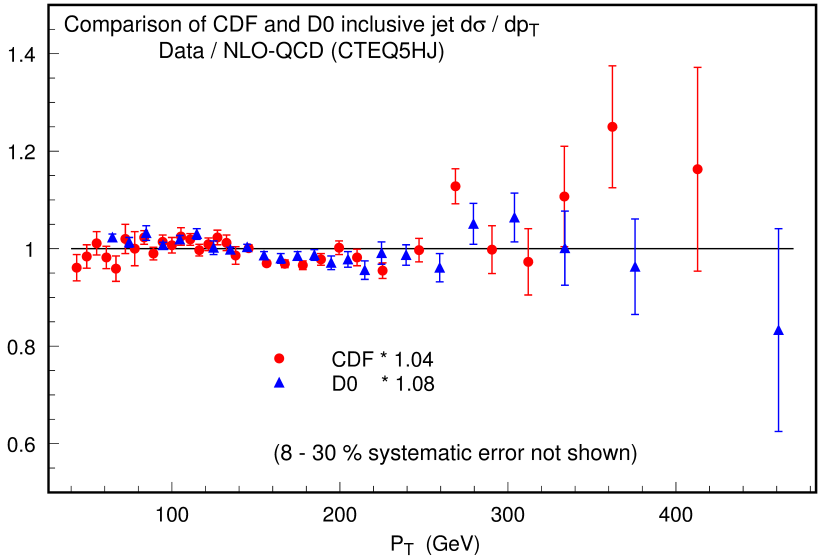
# PDF UNCERTAINTIES AND NEW PHYSICS

- 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)

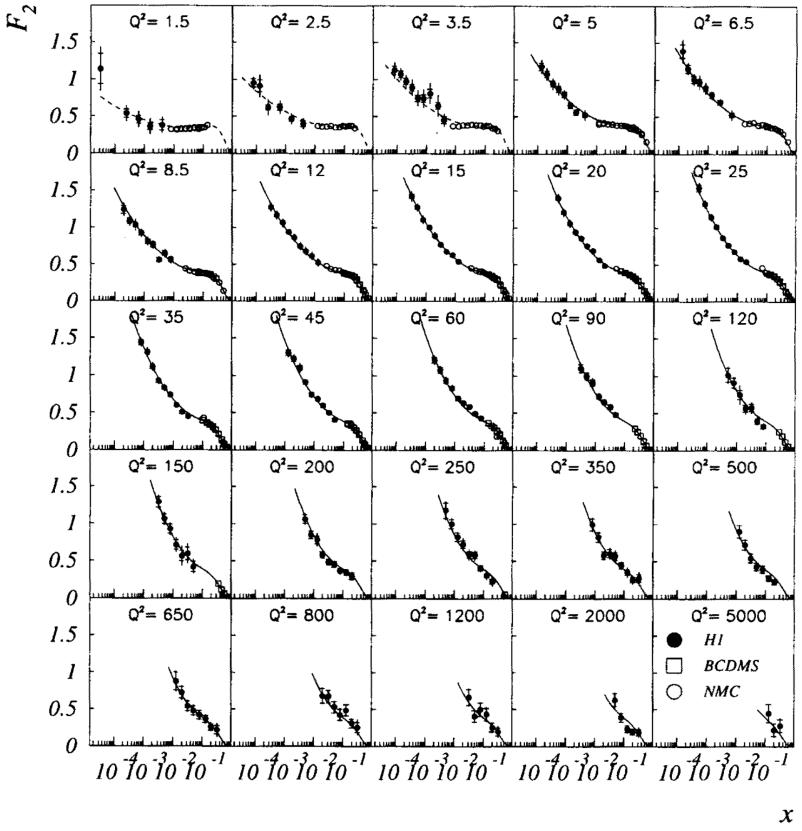



NOW: NO DATA FOR $x \gtrsim 0.5 \Rightarrow$ DISCOVERY (THRESHOLD) REGION!

# EXTRAPOLATION AND THEORY BIAS

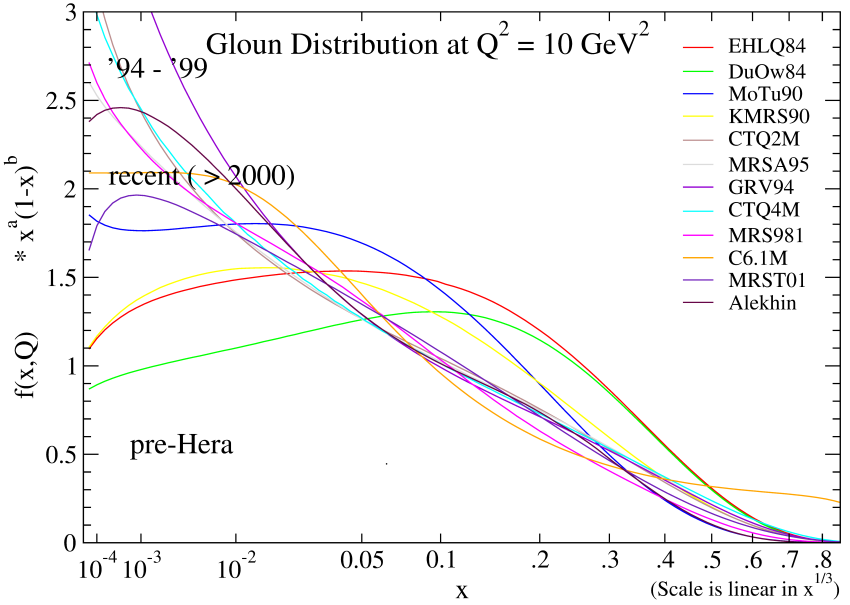
## 1995: THE RISE OF STRUCTURE FUNCTIONS AT HERA

FIRST HERA DATA VS OLDER DATA




A. de Roeck, Cracow epiphany conf. 1996

HISTORICAL COMPILATION OF GLUON PDFS

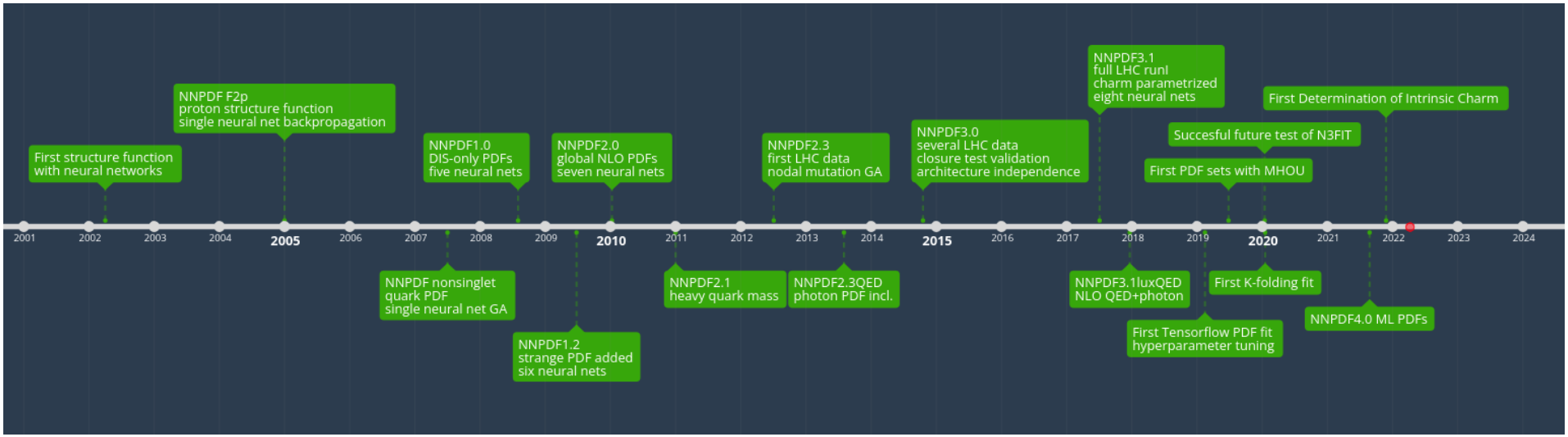



W.K.Tung, DIS 2004

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

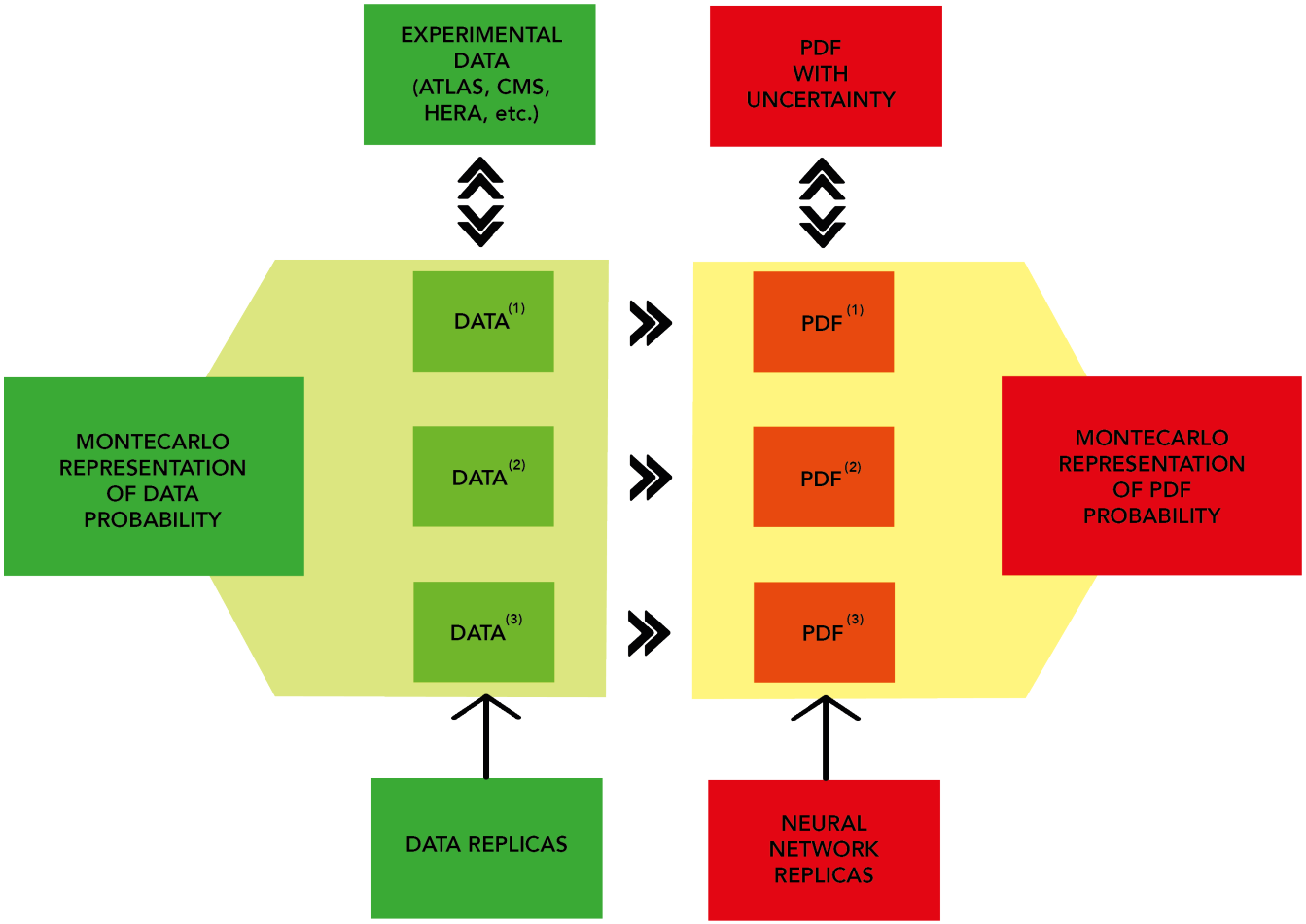
# THE NNPDF METHODOLOGY

# PROTON STRUCTURE AS AN AI PROBLEM:
# NNPDF



# THE FUNCTIONAL MONTE CARLO

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

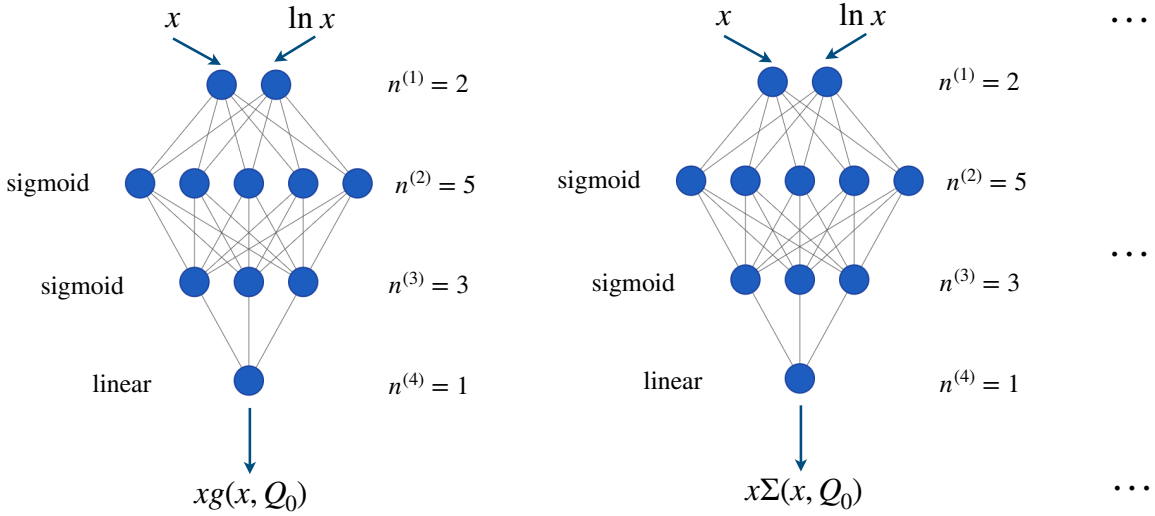



FINAL PDF SET: $f_i^{(a)}(x,\mu)$;
i =up, antiup, down, antidown, strange, antistrange, charm, gluon; $j = 1, 2, \ldots N_{\rm rep}$

# OLD NNPDF (UP TO 2021)

## NEURAL NETWORKS

$$f_i(x, Q_0^2) = x^{\alpha_i}(1-x)^{\beta_i} NN(x, \ln x)$$




- 37 PARAMETERS $\times$ 8 PDFS
- PREPROCESSING EXPONENTS RANDOMIZED REPLICA PER REPLICA, RANGE DETERMINED SELF-CONSISTENTLY

## GENETIC MINIMIZATION

- NODAL MUTATION REPLACED POINT MUTATION
- SINGLE EPOCH WITH NO REWEIGHTING UNLIKE PREVIOUS

## OPTIMAL FIT

- CROSS-VALIDATION WITH 50-50 TRAINING & VALIDATION FRACTIONS
- LOOKBACK STOPPING REPLACED THRESHOLD ON DERIVATIVE

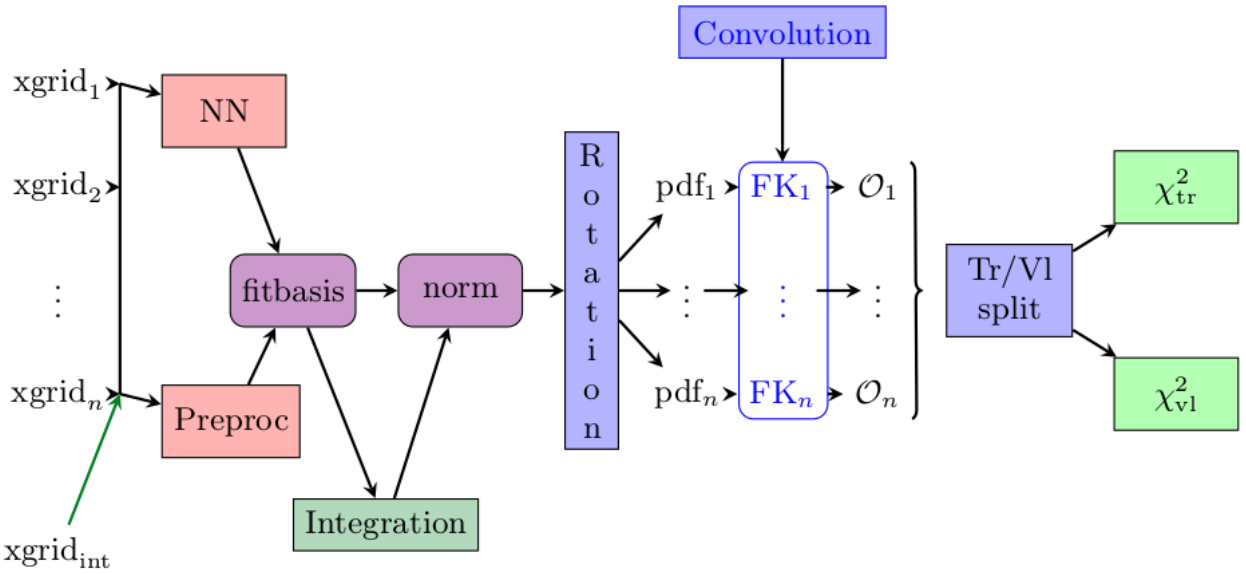
# MACHINE LEARNING PDFs

# LEARNING THE METHODOLOGY

## THE N3FIT PROJECT

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



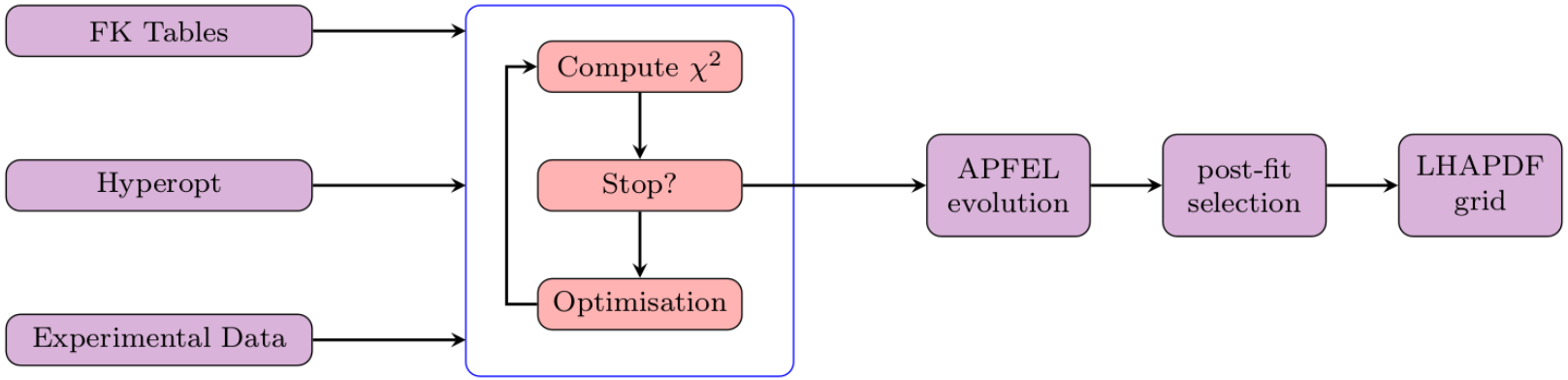

- PYTHON-BASED KERAS + TENSORFLOW FRAMEWORK
- EACH BLOCK INDEPENDENT LAYER
- CAN VARY ALL ASPECTS OF METHODOLOGY

# THE NNPDF CODE STRUCTURE

- MODULAR PYTHON-BASED CODE
- HIGH DEGREE PARALLELIZATION & HARDWARE ACCELERATION

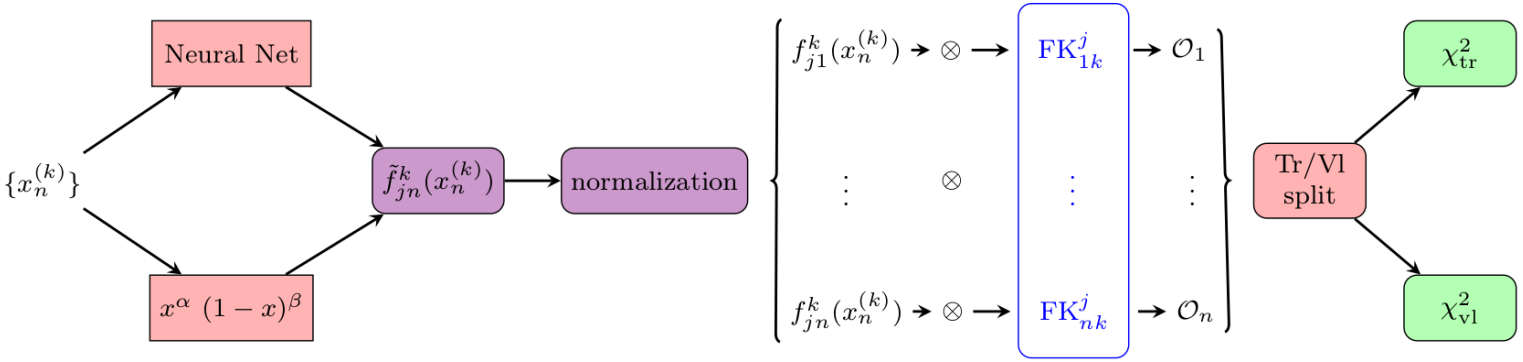
AVERAGE FITTING TIME PER REPLICA AND USE OF RESOURCES
SAME DATASET FOR OLD AND NEW METHODOLOGIES IN CPU AND GPU
CPU: INTEL(R) CORE(TM) I7-4770 AT 3.40GHZ; GPU: NVIDIA TITAN V

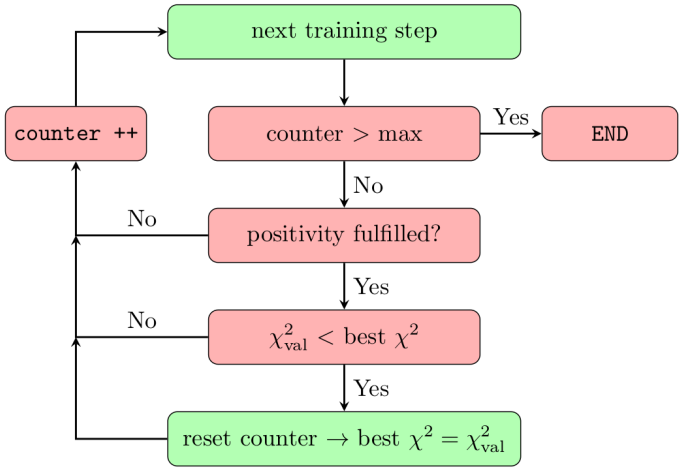
| | NNPDF31 CODEBASE | NNPDF40 CODEBASE IN CPU | NNPDF40 CODEBASE IN GPU |
|---|---|---|---|
| TIME | 15.2 H. | 38 $\pm$ 5 MIN. | 6.6 MIN. |
| RAM USE | 1.5 GB | 6.1 GB | NA |



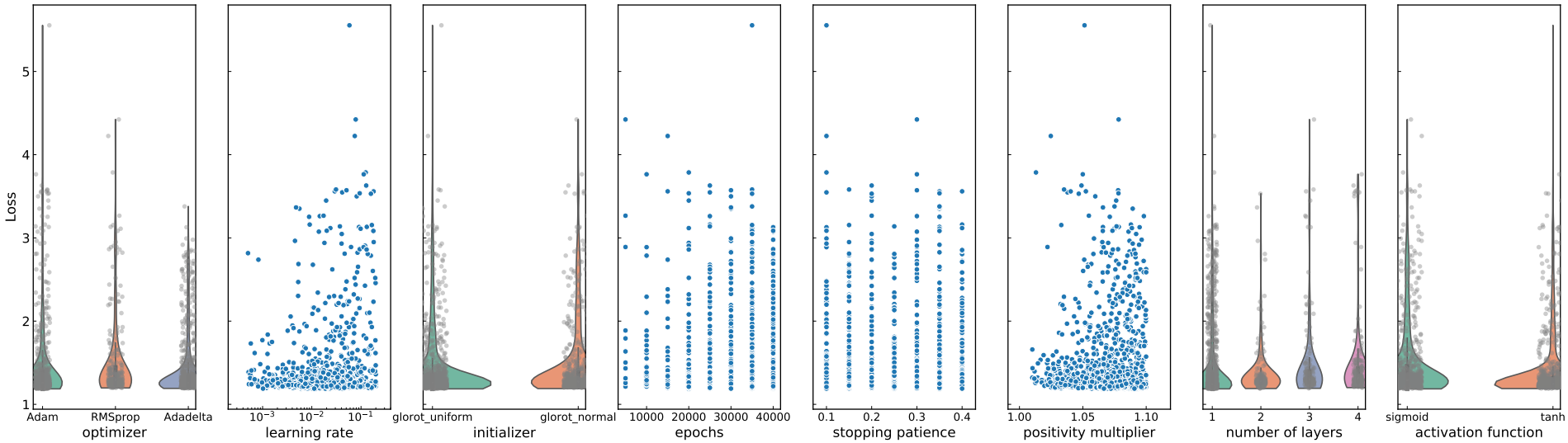
# MINIMIZATION AND CROSS-VALIDATION

- DATA REPLICAS $\Rightarrow$ PDF REPLICAS
- EACH PDF REPLICA: PREPROCESSED NEURAL NET
- NEURAL NET $\Rightarrow$ OBSERVABLES
- RANDOM TRAINING-VALIDATION SPLIT, $\chi^2$ TO TRAINING DATA REPLICAS MINIMIZED
- TRAINING STOPS IF VALIDATION $\chi^2$ GROWS FOR A WHILE (PATIENCE)
- LOWEST VALIDATION $\chi^2$ $\Rightarrow$ OPTIMAL FIT



# FITTING THE METHODOLOGY
# HYPEROPTIMIZATION SCANS



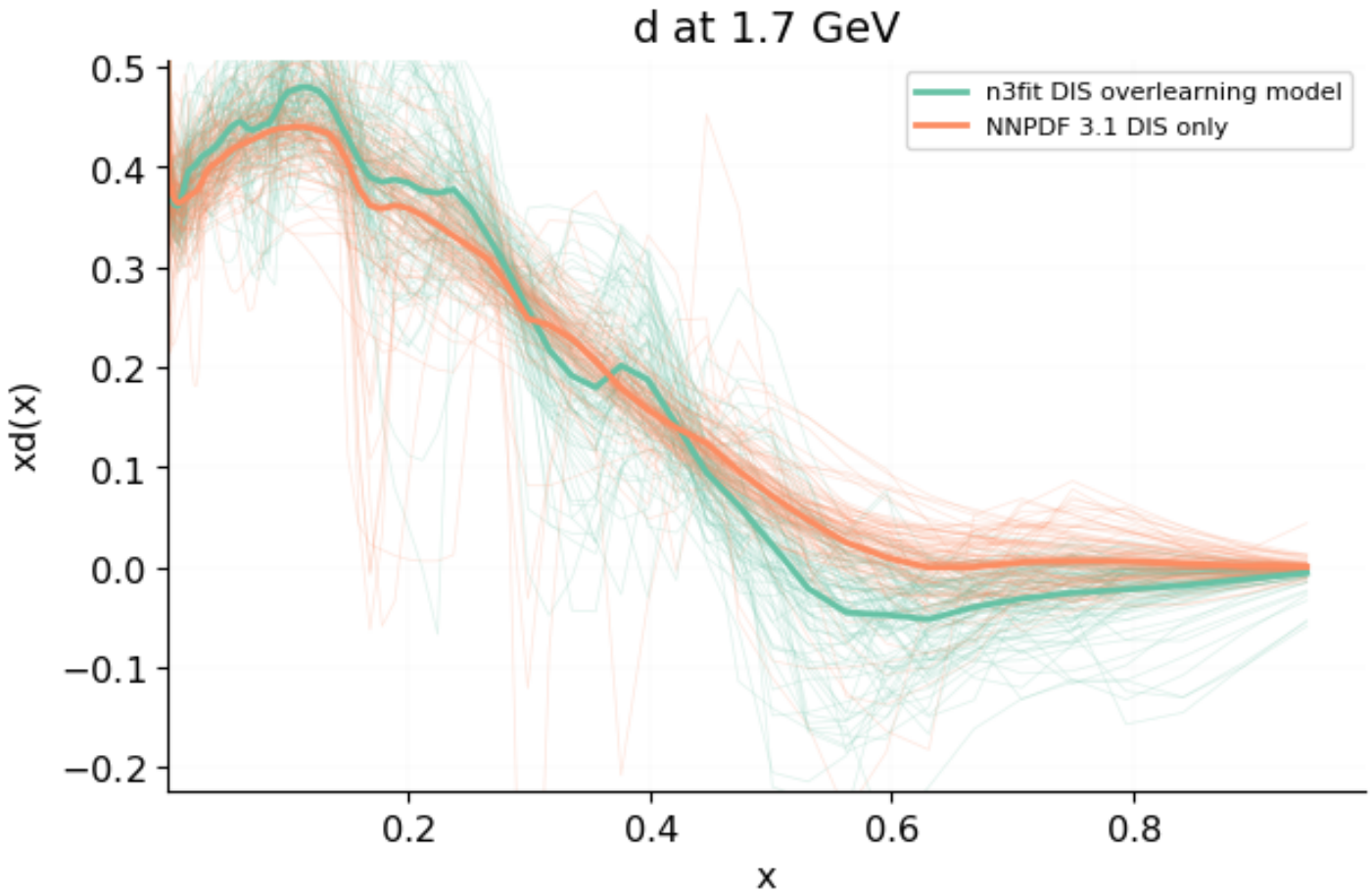

## HYPEROPT PARAMETERS

| 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 $\chi^2$
- BAYESIAN UPDATING

# HYPEROPTIMIZATION: OVERFITTING

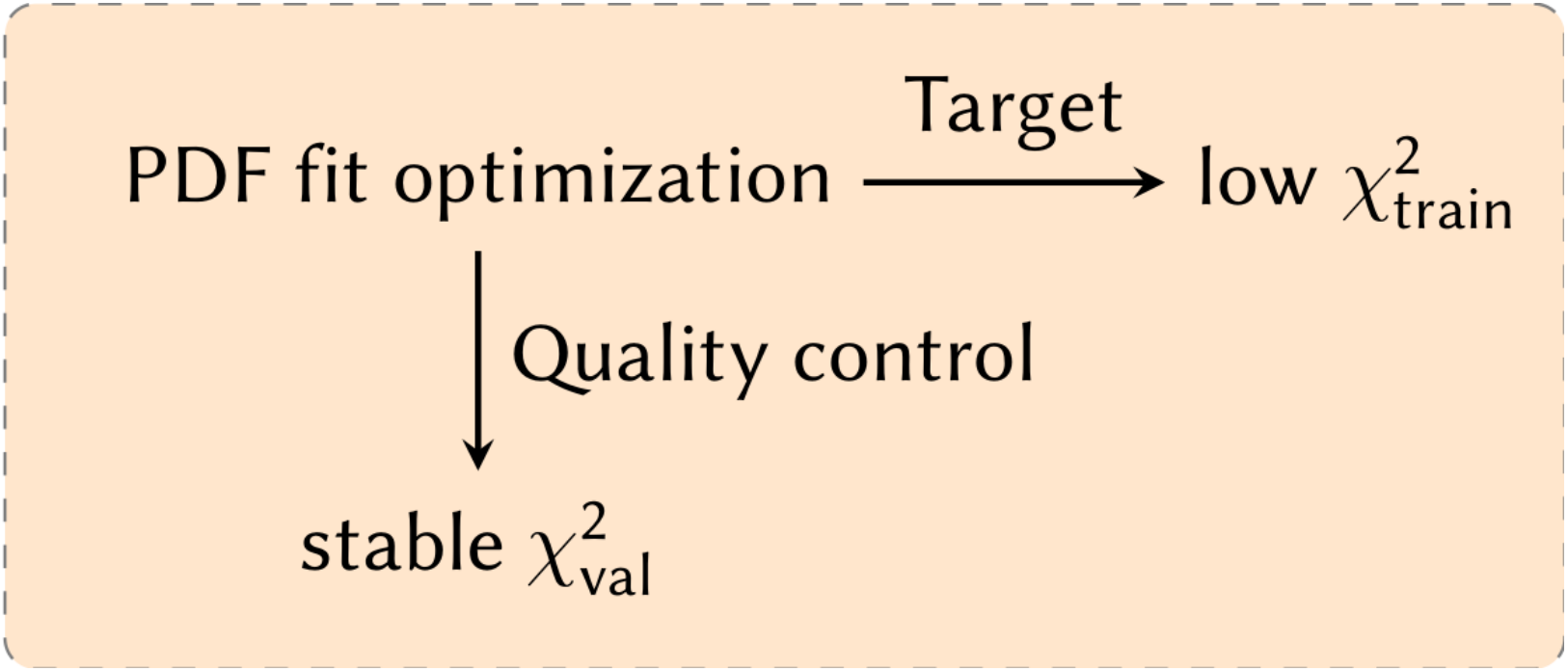
## DOWN QUARK: HYPEROPTIMIZED VS. HAND-PICKED




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

# WHAT HAPPENED?

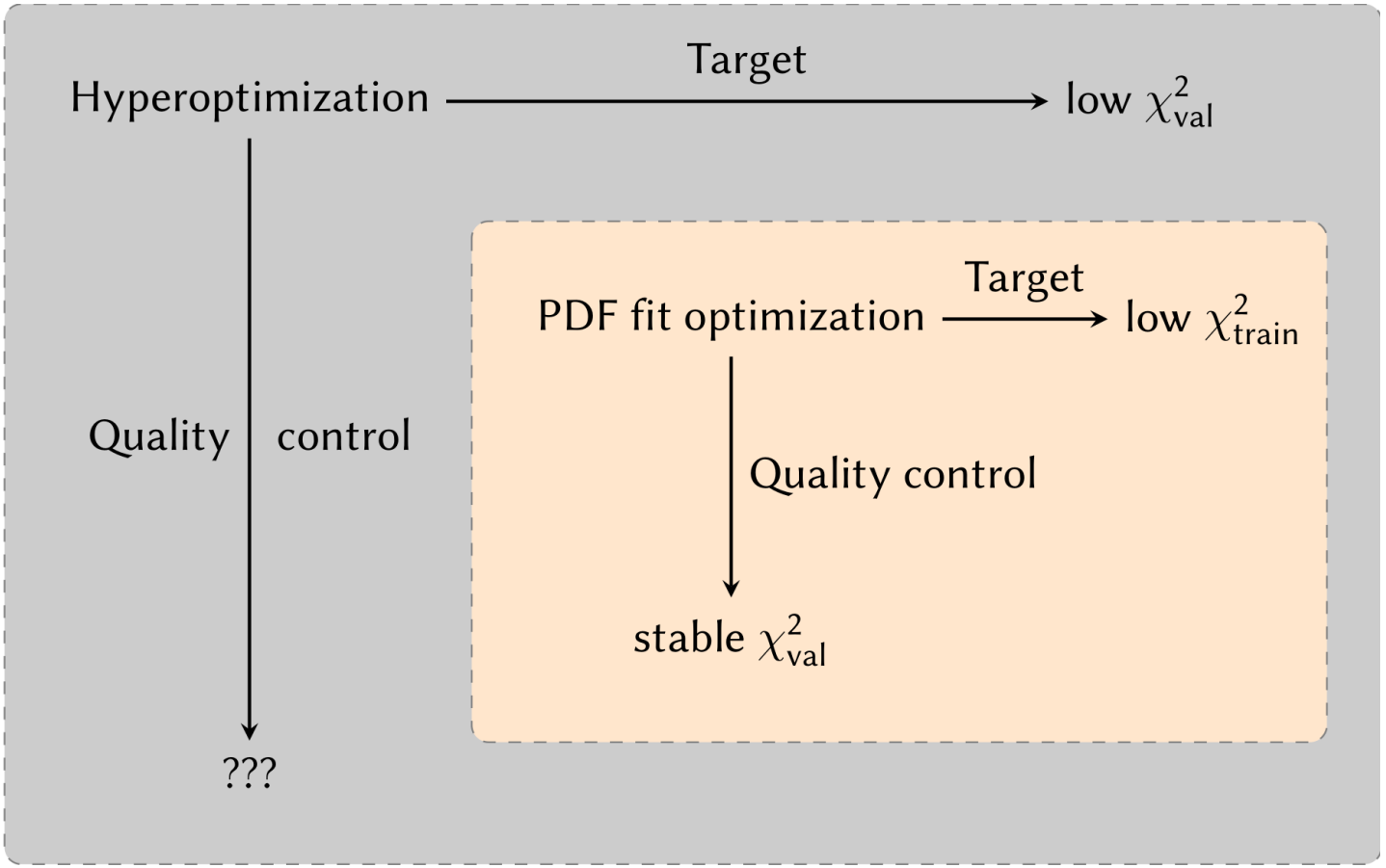
OPTIMIZATION




CROSS-VALIDATION SELECTS THE OPTIMAL MINIMUM

# WHAT HAPPENED?

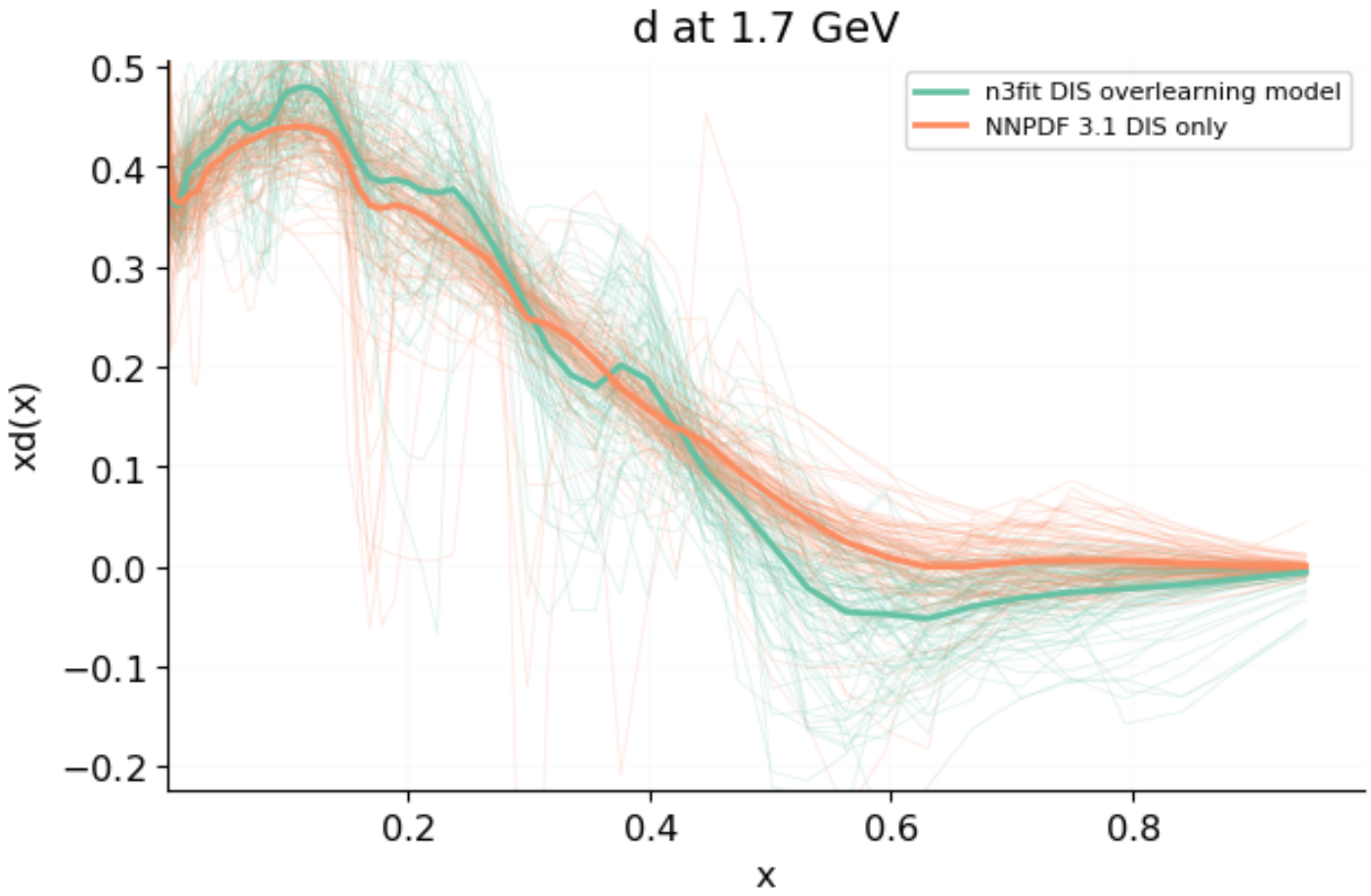
## HYPEROPTIMIZATION




WE ARE MISSING A SELECTION CRITERION

# HYPEROPTIMIZATION: OVERFITTING

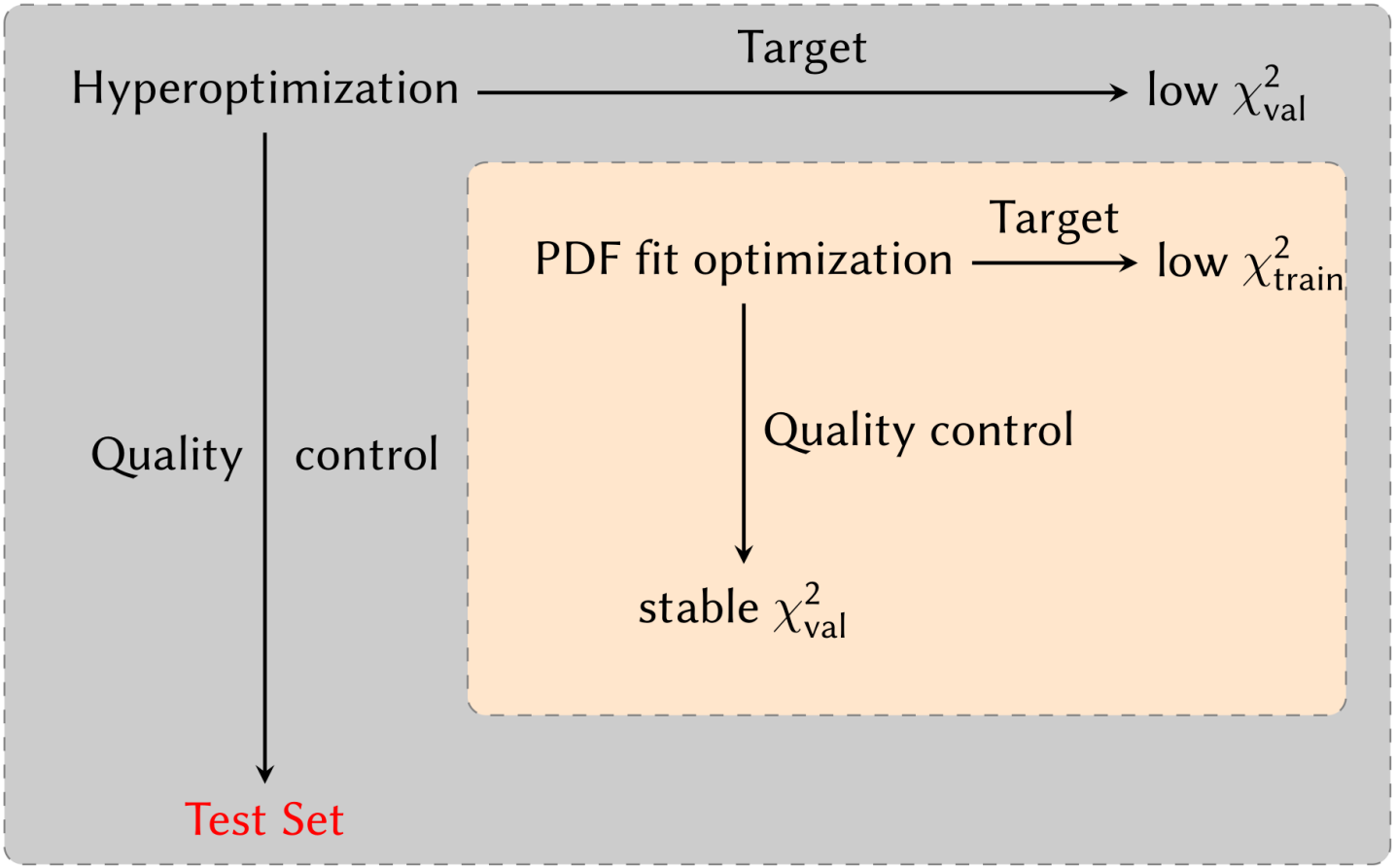
## DOWN QUARK: HYPEROPTIMIZED VS. HANDPICKED




- HANDPICKED: 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

# THE SOLUTION

## TUNED HYPEROPTIMIZATION




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

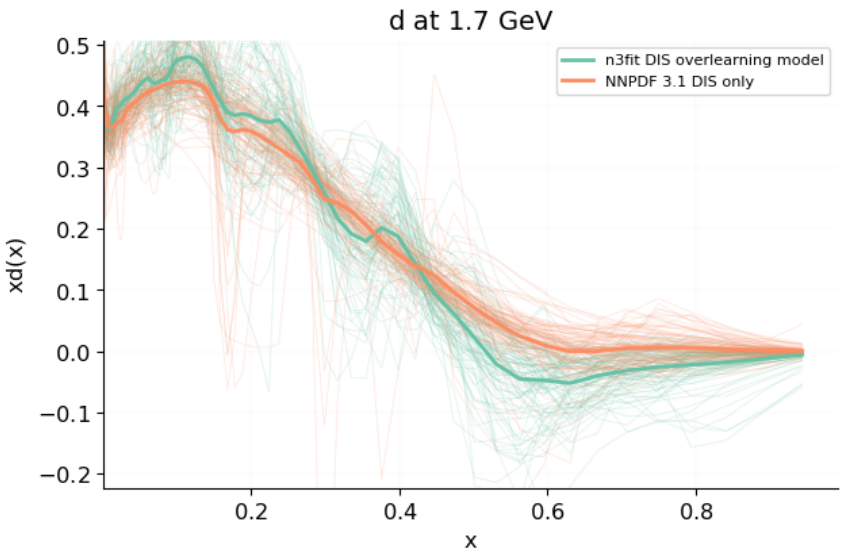
# THE TEST SET METHOD

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

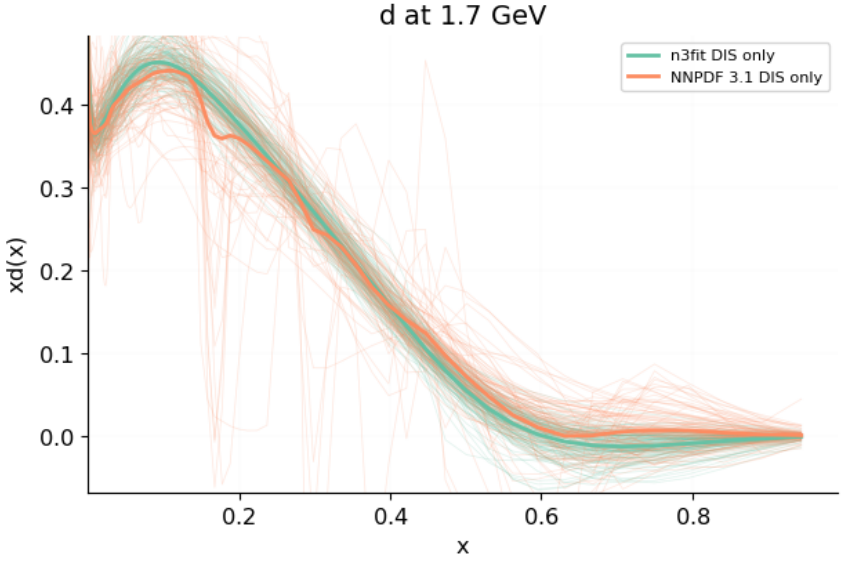
## HYPEROPTIMIZED PDFS

DOWN QUARK

OVERFIT VS HANDPICKED




HYPEROPT VS HANDPICKED




- NO OVERFITTING
- COMPARED TO HANDPICKED
  - MUCH GREATER STABILITY $\Rightarrow$ FEWER REPLICAS FOR EQUAL ACCURACY
  - UNCERTAINTIES SOMEWHAT REDUCED

# $K$-FOLDING

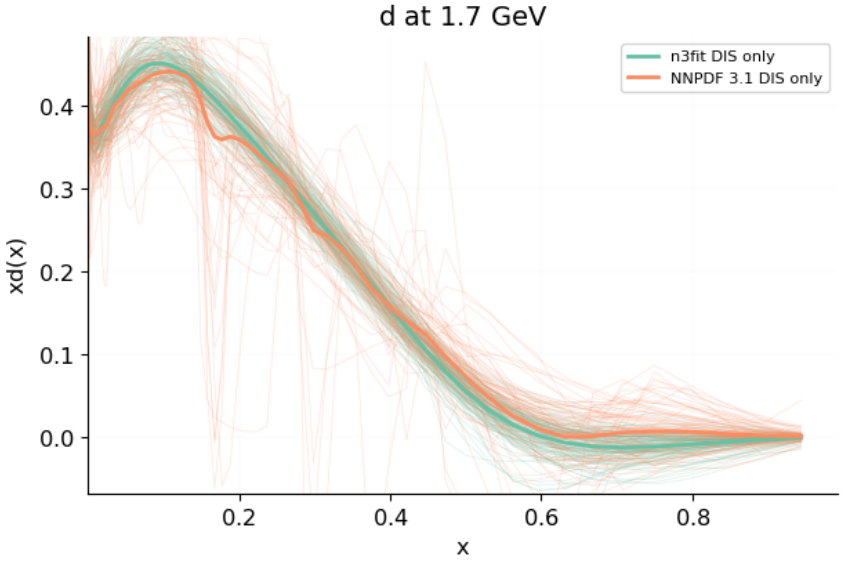
## THE BASIC IDEA:

- 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 NON FITTED $\chi^2_{\text{test, i}}$
  $\rightarrow$ GOOD & STABLE GENERALIZATION

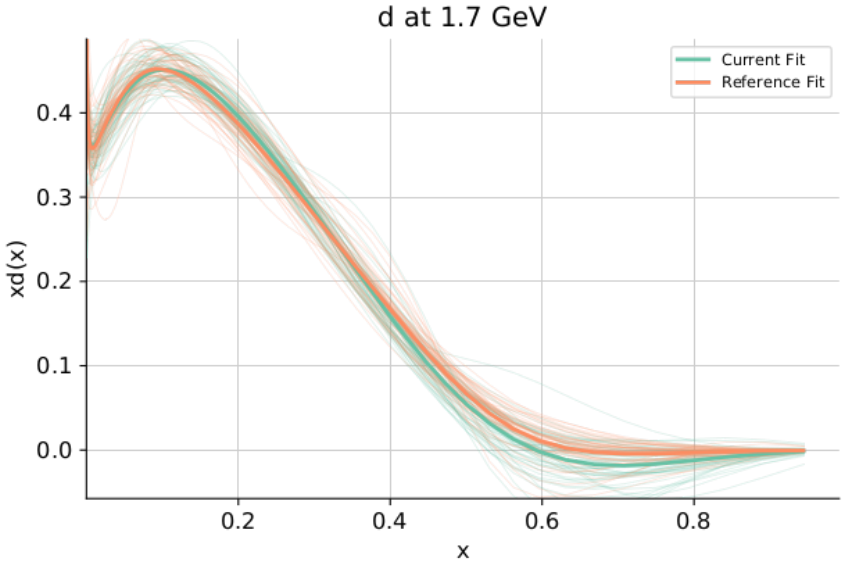
## FOLDED PDFS

DOWN QUARK

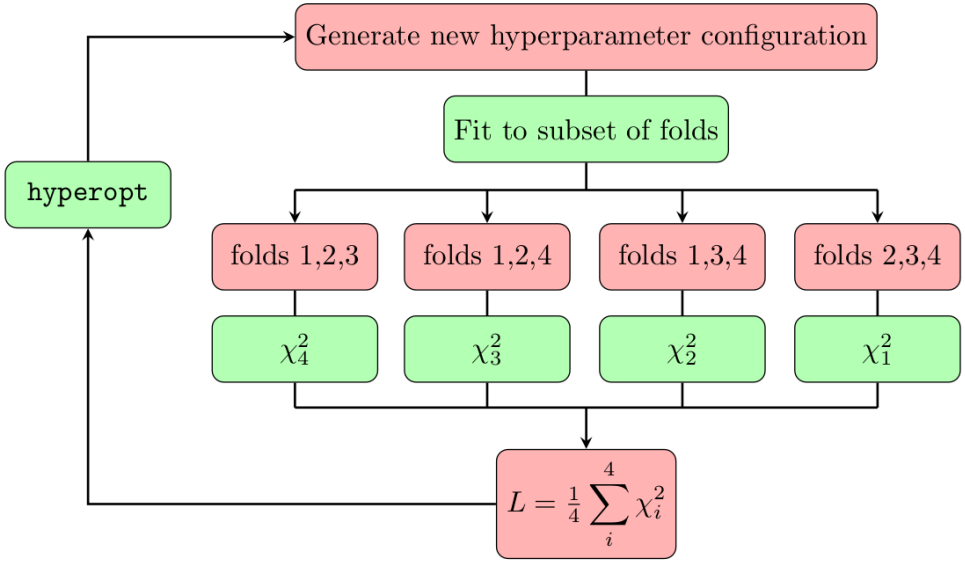
TEST-SET HYPER VS HANDPICKED




K-FOLD HYPER VS. TEST-SET HYPER



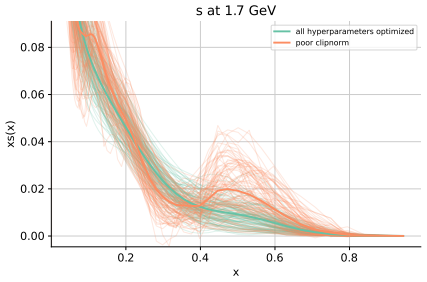
# K-FOLDING IMPLEMENTATION



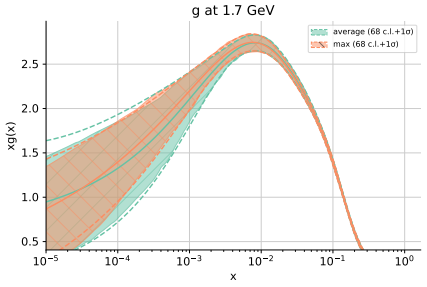

| **Fold 1** | | |
|---|---|---|
| CHORUS $\sigma^{\nu}_{CC}$ | HERA I+II inc NC $e^+p$ 920 GeV | BCDMS $p$ |
| LHCb $Z$ 940 pb | ATLAS $W,Z$ 7 TeV 2010 | CMS $Z$ $p_T$ 8 TeV $(p^{ll}_T, y_{ll})$ |
| DY E605 $\sigma^p_{\rm DY}$ | CMS Drell-Yan 2D 7 TeV 2011 | CMS 3D dijets 8 TeV |
| ATLAS single-$t$ $y$ (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 $\sigma^{\nu}_c$ | LHCb $Z \to ee$ 2 fb |
| CMS $W$ asymmetry 840 pb | ATLAS $Z$ $p_T$ 8 TeV $(p^{ll}_T, M_{ll})$ | 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 $\sigma^{\rm tot}_{t\bar{t}}$ | 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 $\sigma^{\nu}_c$ | 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 $\sigma^{\rm tot}_{t\bar{t}}$ | ATLAS single top $y_t$ (normalised) | CMS $\sigma^{\rm tot}_{t\bar{t}}$ 5 TeV |
| CMS $t\bar{t}$ double diff. $(m_{t\bar{t}}, y_t)$ | | |
| **Fold 4** | | |
| CHORUS $\sigma^{p}_{CC}$ | 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 $(p^{ll}_T, y_{ll})$ | 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 | |

- EACH FOLD REPRODUCES FEATURES OF FULL DATASET
- DIFFERENT CHOICES POSSIBLE FOR LOSS (NON-FITTED)
  - BEST WORST
  - BEST AVERAGE
- RESULTS STABLE

NO K-FOLDING




K-FOLDING VARIATION

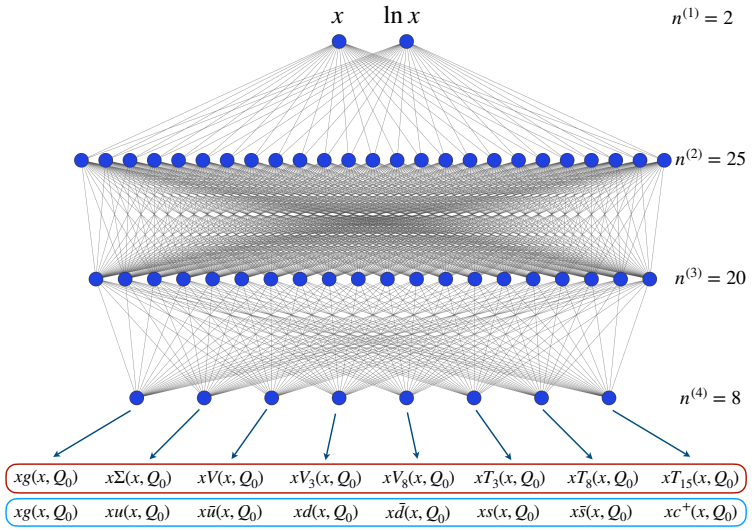


# THE ML METHODOLOGY

HYPEROPTIMIZED PARAMETERS

| Parameter | NNPDF4.0 | $L$ as in Eq. (3.21) | Flavour basis Eq. (3.2) |
|---|---|---|---|
| Architecture | 25-20-8 | 70-50-8 | 7-26-27-8 |
| Activation function | hyperbolic tangent | hyperbolic tangent | sigmoid |
| Initializer | `glorot_normal` | `glorot_uniform` | `glorot_normal` |
| Optimizer | `Nadam` | `Adadelta` | `Nadam` |
| Clipnorm | $6.0\times10^{-6}$ | $5.2\times10^{-2}$ | $2.3\times10^{-5}$ |
| Learning rate | $2.6\times10^{-3}$ | $2.5\times10^{-1}$ | $2.6\times10^{-3}$ |
| Maximum # epochs | $17\times10^{3}$ | $45\times10^{3}$ | $45\times10^{3}$ |
| Stopping patience | 10% of max epochs | 12% of max epochs | 16% of max epochs |
| Initial positivity $\Lambda^{(\mathrm{pos})}$ | 185 | 106 | 2 |
| Initial integrability $\Lambda^{(\mathrm{int})}$ | 10 | 10 | 10 |

NN ARCHITECTURE




- HYPEROPT ADAPTS TO EXTERNAL CHOICES (E.G. PARAMETRIZATION BASIS)
- SIMILAR RESULTS CAN BE OBTAINED WITH RATHER DIFFERENT SETTINGS
- $\sim$ 800 FREE PARAMETERS

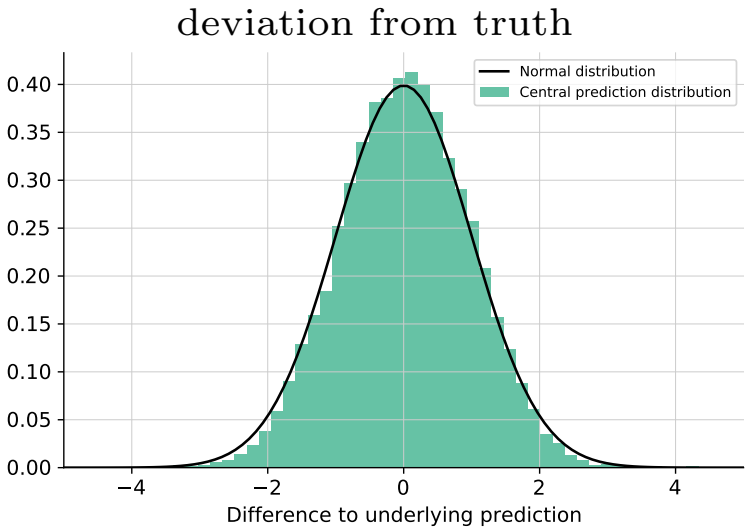
# VALIDATION

# CLOSURE TESTS

## FAITHFUL UNCERTAINTIES IN DATA REGION?

- ASSUME "TRUE" UNDERLYING PDF $\Rightarrow$ 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 PDFS, OR TO TRUE VALUES OF OBSERVABLES (NOT FITTED)
  - BIAS/VARIANCE: MEAN SQUARE DEVIATION WR TO TRUTH VS UNCERTAINTY
  - IS TRUTH WITHIN ONE SIGMA 68% OF TIMES?

## RESULTS



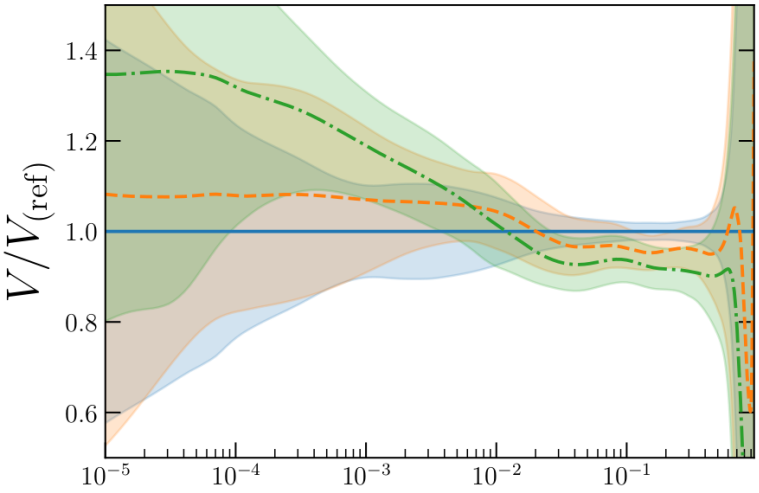

| Dataset | $\sqrt{\text{bias/variance}}$ | $\xi_{1\sigma}^{(\text{data})}$ |
|---|---|---|
| DY | $0.99 \pm 0.08$ | $0.69 \pm 0.02$ |
| Top-pair | $0.75 \pm 0.06$ | $0.75 \pm 0.03$ |
| Jets | $1.14 \pm 0.05$ | $0.63 \pm 0.03$ |
| Dijets | $0.99 \pm 0.07$ | $0.70 \pm 0.03$ |
| Direct photon | $0.71 \pm 0.06$ | $0.81 \pm 0.03$ |
| Single top | $0.87 \pm 0.07$ | $0.69 \pm 0.04$ |
| Total | $1.03 \pm 0.05$ | $0.68 \pm 0.02$ |

# FUTURE TESTS

## FAITHFUL UNCERTAINTIES IN EXTRAPOLATION?

- DETERMINE PDFs FROM A SUBSET OF CURRENT DATA: "PRE-HERA", "PRE-LHC",...
- COMPUTE $\chi^2$ TO THE FULL CURRENT DATASET:
  - WITHOUT PDF UNCERTAINTIES $\Rightarrow$ IF $\gg 1$, MISSING INFORMATION
  - WITH PDF UNCERTAINTY $\Rightarrow$ IF $\sim 1$, MISSING INFO REPRODUCED BY UNCERTAINTY

VALENCE: PRE-HERA VS PRE-LHC VS NNPDF4.0



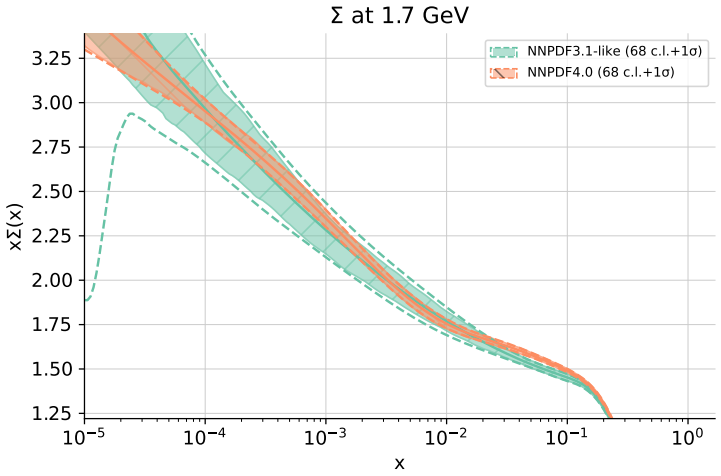

$\chi^2$ WITHOUT/**WITH PDF UNC.**

| PROCESS | PRE-HERA | PRE-LHC | 3.1-LIKE | 4.0-GLOB |
|---|---|---|---|---|
| FT DIS (NC) | 1.04 | 1.17 | 1.17 | 1.26 |
| FT DIS (CC) | 0.80 | 0.86 | 0.88 | 0.90 |
| FT DY | 0.93 | 1.27 | 1.43 | 1.59 |
| HERA | 24.01/**1.12** | 1.22 | 1.21 | 1.21 |
| COLL. DY (TEV.) | 5.31/**1.08** | 0.96 | 0.95 | 1.13 |
| COLL. DY (LHC) | 15.50/**1.37** | 2.64/**1.54** | 1.39 | 1.54 |
| TOP QUARK | 23.35/**1.08** | 1.29/**0.86** | 0.82 | 0.98 |
| JETS | 6.18/**1.21** | 3.66/**1.29** | 2.07/**1.37** | 1.26 |
| TOTAL | 9.70 | 1.44 | 1.22 | 1.17 |

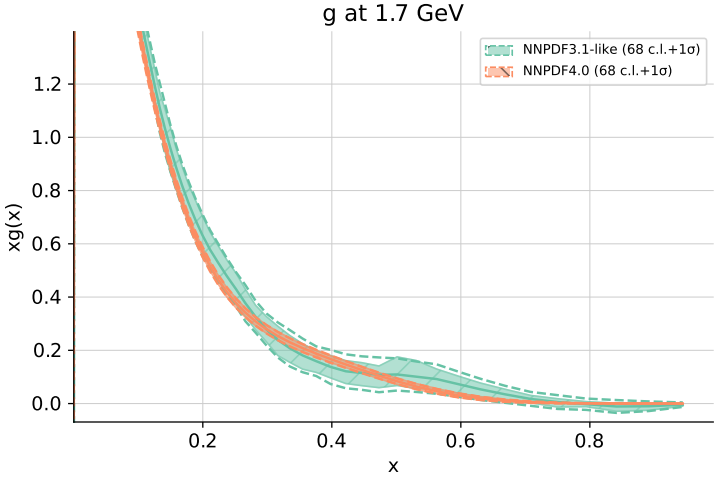
# NNPDF4.0 vs. NNPDF3.1

- FULL BACKWARD COMPATIBILITY
- SUBSTANTIAL REDUCTION IN UNCERTAINTY

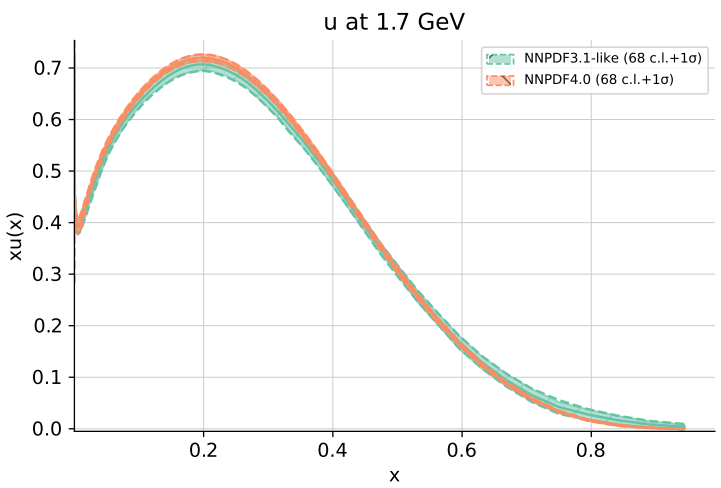
## SINGLET



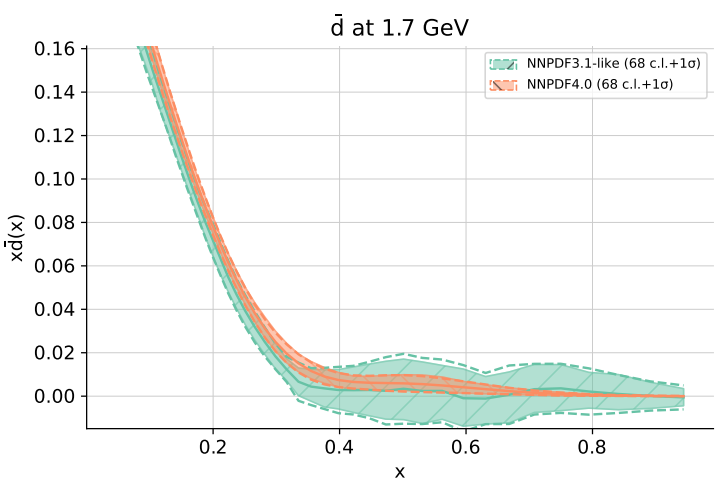

## GLUON




## UP




## ANTIDOWN

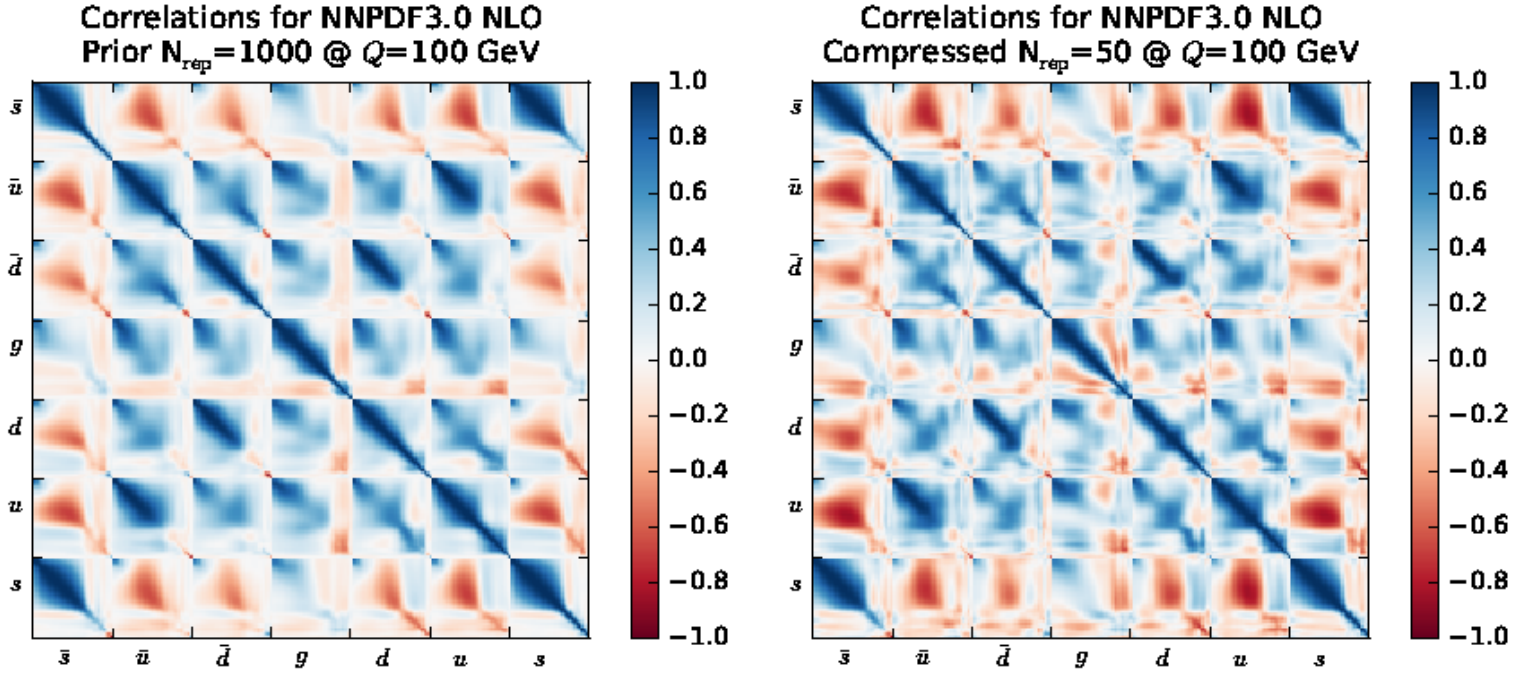


# DELIVERY

# MONTECARLO COMPRESSION

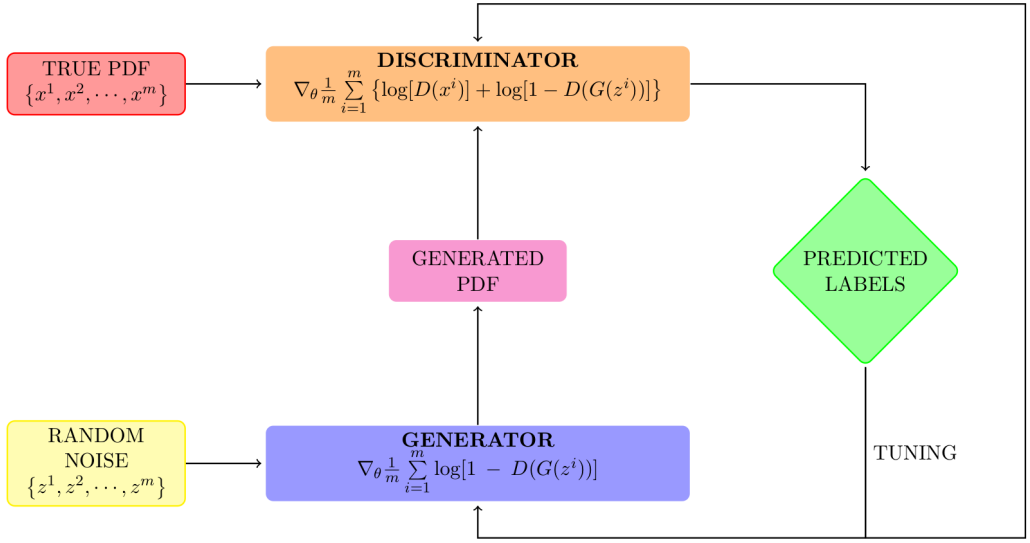
CAN WE REDUCE THE NUMBER OF REPLICAS?

- START WITH LARGE REPLICA SAMPLE
- SELECT BY GENETIC ALGORITHM SUBSET OF REPLICAS $\Rightarrow$ STATISTICAL FEATURES OPTIMIZED TO PRIOR
- MINIMIZE LOSS: DIFFERENCE OF MOMENTS, KL DIVERGENCE, . . .
- 50 COMPRESSED REPLICA REPRODUCE 1000 REPLICA SET TO PRECENT ACCURACY



# GAN ENHANCEMENT

CAN WE FURTHER REDUCE THE NUMBER OF COMPRESSED REPLICAS WITHOUT LOSS OF INFORMATION? GENERATIVE ADVERSARIAL NETWORKS

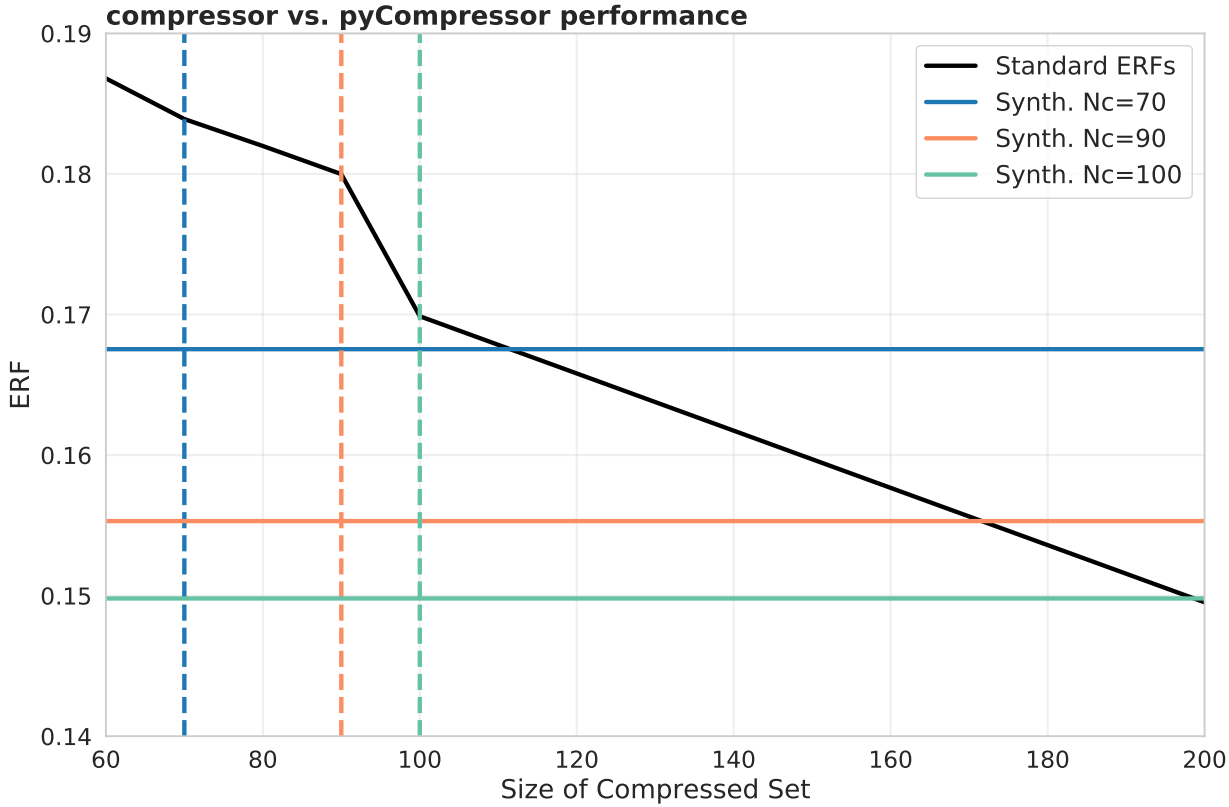



- TRAIN A NETWORK TO SIMULATE THE TRUE DISTRIBUTION (GENERATOR)
- TRAIN A NETWORK TO DISCRIMINATE TRUTH FROM SIMULATION (DISCRIMINATOR)
- TRAIN THE GENERATOR TO TRICK THE DISCRIMINATOR

# GAN ENHANCEMENT

- ENHANCE THE STARTING PDF SET BY ADDING GAN-PDFS TO IT
- PERFORM COMPRESSION OF THE ENHANCED SET

## PERFORMANCE



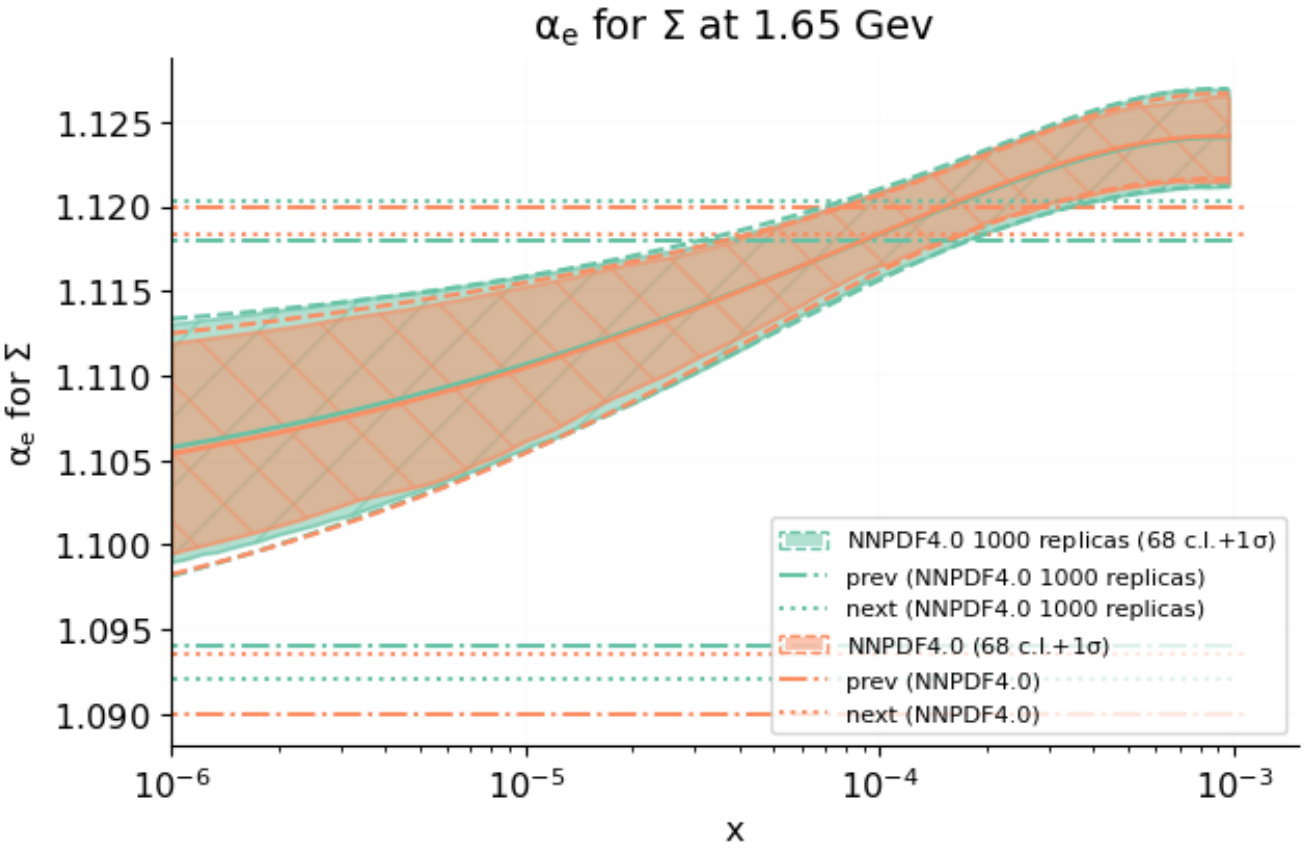

ENHANCED: NUMBER OF REPLICAS CUT IN HALF FOR SAME TARGET ACCURACY

# NEW IDEAS

# PREPROCESSING

- PDF EQUAL TO PREPROCESSED NN: $f_i = x^{\alpha_i}(1-x)^{\beta}_i NN(i, x, \ln x)$
- PREPROCESSING EXPONENTS $\alpha_i$, $\beta_i$ VARIED RANDOMLY REPLICA BY REPLICA
- RANGE OF VARIATION DETERMINED SELF-CONSISTENTLY
- PDF TAKES $x$, $\ln x$ AS INPUTS

QUARK SINGLET EFFECTIVE $\alpha$

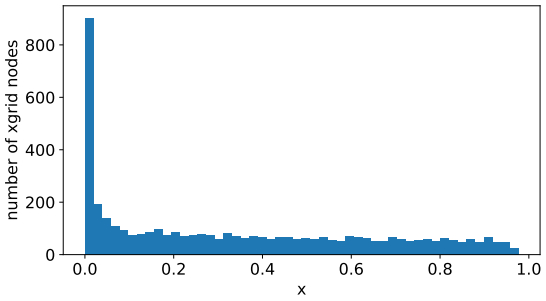



- NEED TO ITERATE FITS
- POSSIBLE SOURCE OF BIAS?

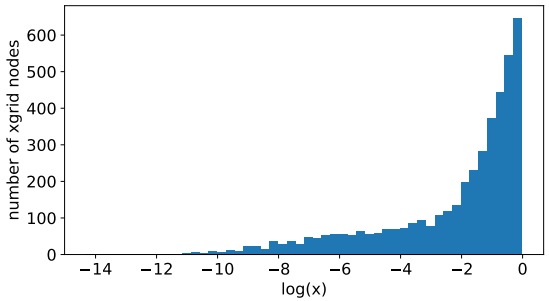
# FEATURE SCALING

## INPUT NODE DISTRIBUTION

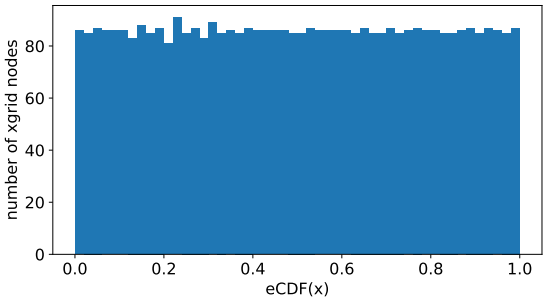
$x$ DISTN




$\ln x$ DISTN



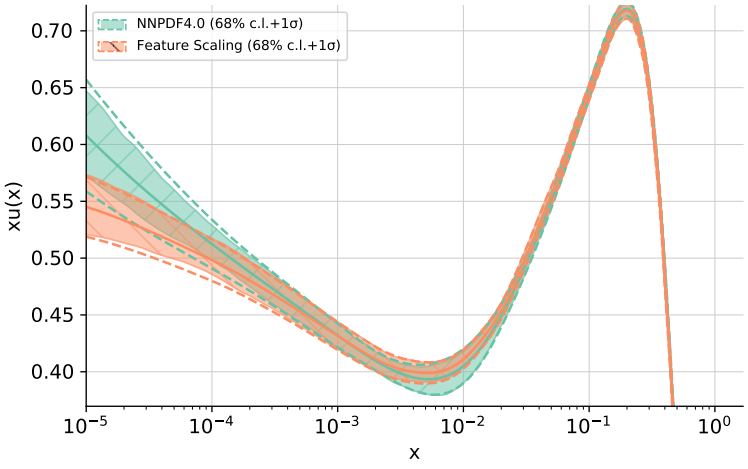

eCDF DISTN



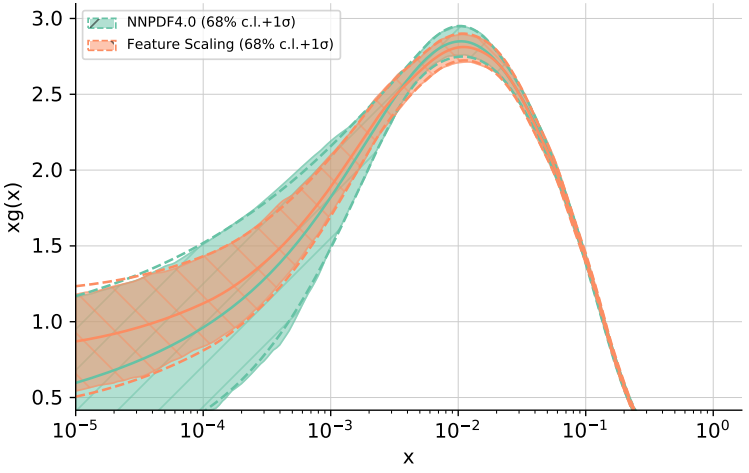

- RESCALE $\Rightarrow$ UNIFORMLY DISTRIBUTED INPUT (eCDF+INTERPOLATION)
- RERUN HYPEROPT
- ONLY ONE INPUT NEEDED

## STANDARD VS. FEATURE SCALING

UP




GLUON

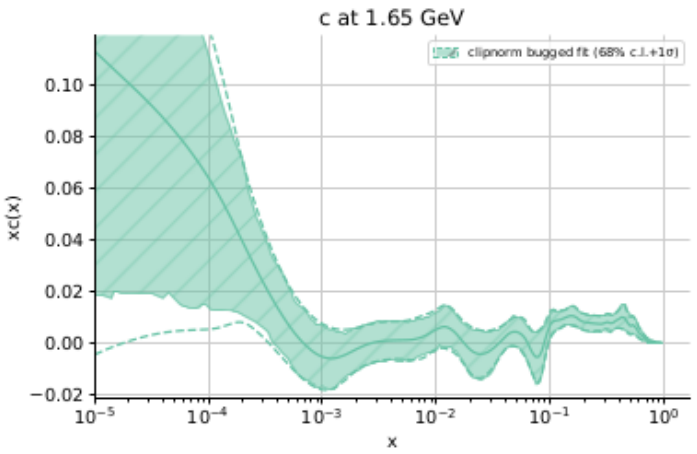


# OVERFITTING IN HYPEROPT

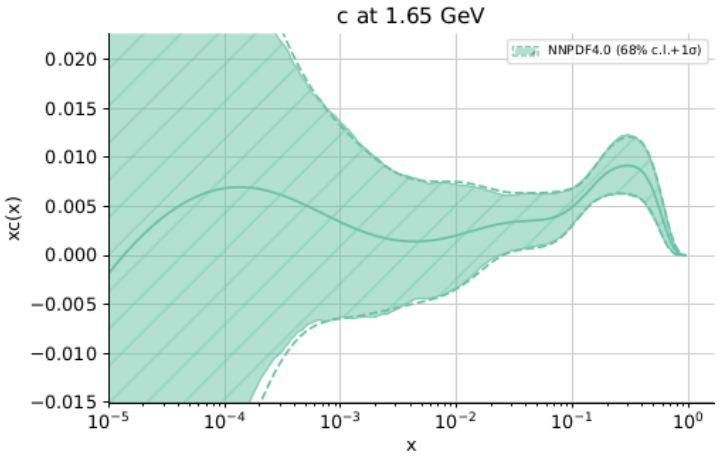
- HYPEROPT $\Rightarrow$ OVERFITTING?
- WHAT IS A "NATURAL" PDF?

CHARM PDF

OVERFIT (NO CLIPNORM)




PROPER FIT (NNPDF4.0)

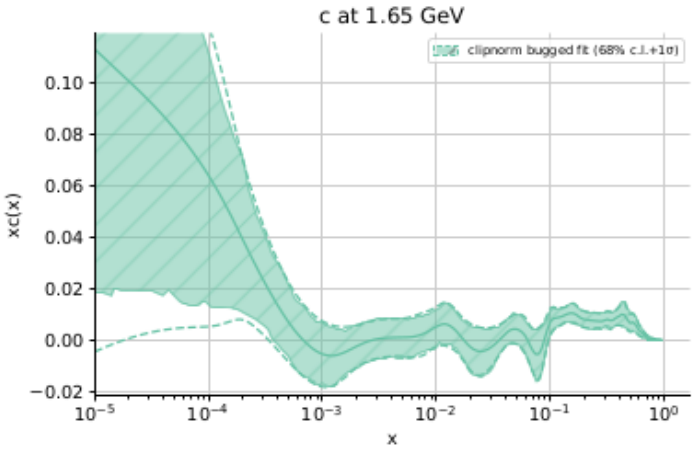


# OVERFITTING METRIC

- RECOMPUTE VALIDATION $\chi^2$
  - SAME TRAINING-VALIDATION SPLIT
  - DIFFERENT FLUCTUATED VALIDATION DATA
- COMPUTE AVERAGE $\chi^2$ & DETERMINE DIFFERENCE TO VALIDATION $\mathcal{R}_O = \langle \chi^2_{\rm val} - \chi^2_{\rm val'} \rangle$ OVERFITNESS
- NEGATIVE OVERFITNESS $\mathcal{R}_O \Rightarrow$ OVERFIT

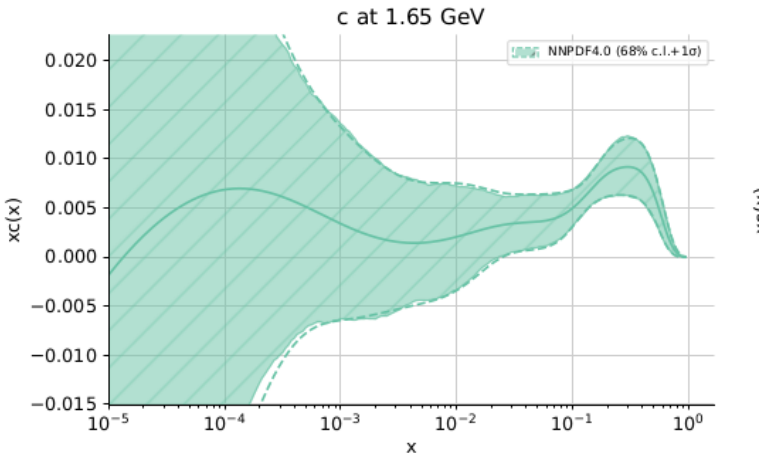
CHARM PDF

OVERFIT (NO CLIPNORM)




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

PROPER FIT (NNPDF4.0)




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

# FOOD FOR THOUGHT

- OPTIMIZED FOLDS
- OVERFITTING AND UNDERFITTING METRICS IN HYPEROPT
- HYPEROPT ON FUTURE TESTS
- BEYOND NEURAL NETS

# AN OPEN SOURCE CODE!

- THE FULL NNPDF CODE IS PUBLIC!
- INCLUDING HYPEROPTIMIZATION, EVOLUTION, THEORY, FITTING, VISUALIZATION
- FULLY DOCUMENTED CODE
- LINKS TO CODE (GITHUB), DOCUMENENTATION, INSTALLATION BINARY PACKAGES AVAILABLE FROM
  `http://nnpdf.mi.infn.it/nnpdf-open-source-code/`

`https://arxiv.org/abs/2109.02671`

**An open-source machine learning framework for global analyses of parton distributions**

**The NNPDF Collaboration: Richard D. Ball · Stefano Carrazza · Juan Cruz-Martinez · Luigi Del Debbio · Stefano Forte · Tommaso Giani · Shayan Iranipour · Zahari Kassabov · Jose I. Latorre · Emanuele R. Nocera · Rosalyn L. Pearson · Juan Rojo · Roy Stegeman · Christopher Schwan · Maria Ubiali · Cameron Voisey · Michael Wilson**

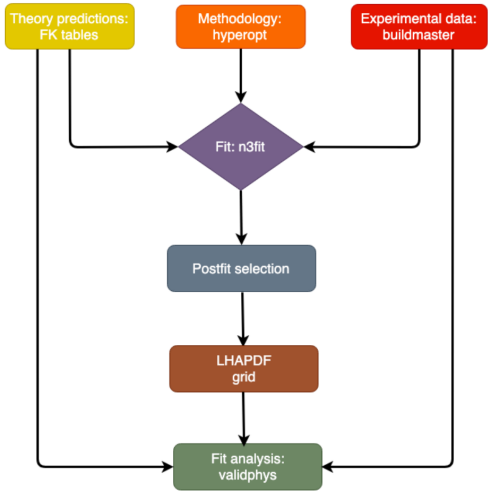



**Fig. 2.1.** Workflow for an NNPDF fit

# THE WORK OF MANY PEOPLE



NNPDF collaboration and N$^3$PDF team meeting,
Gargnano, Italy, August 2022

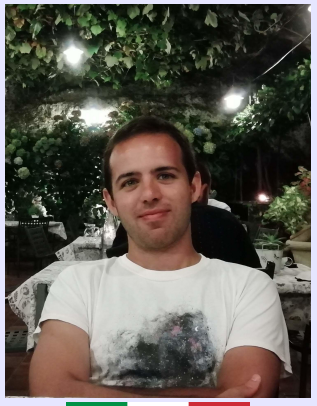
# N$^3$PDF

N3PDF
Machine Learning • PDFs • QCD

A. Barontini



A. Candido



S. Carrazza



J. Cruz Martinez



F. Hekhorn



K. Kudashkin



N. Laurenti



T. Rabemananjara



C. Schwan



R. Stegeman

