CAN WE TRUST MACHINE LEARNING?

PDFS AS A CASE STUDY

STEFANO FORTE

UNIVERSITÀ DI MILANO & INFN

UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI FISICA

IRN TERASCALE MEETING

GRENOBLE, APRIL 26, 2023

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

THE PROBLEM
GENERALIZATION

Search Wikipedia

Machine learning

itents [hide]

Article Talk

From Wikipedia, the free encyclopedia

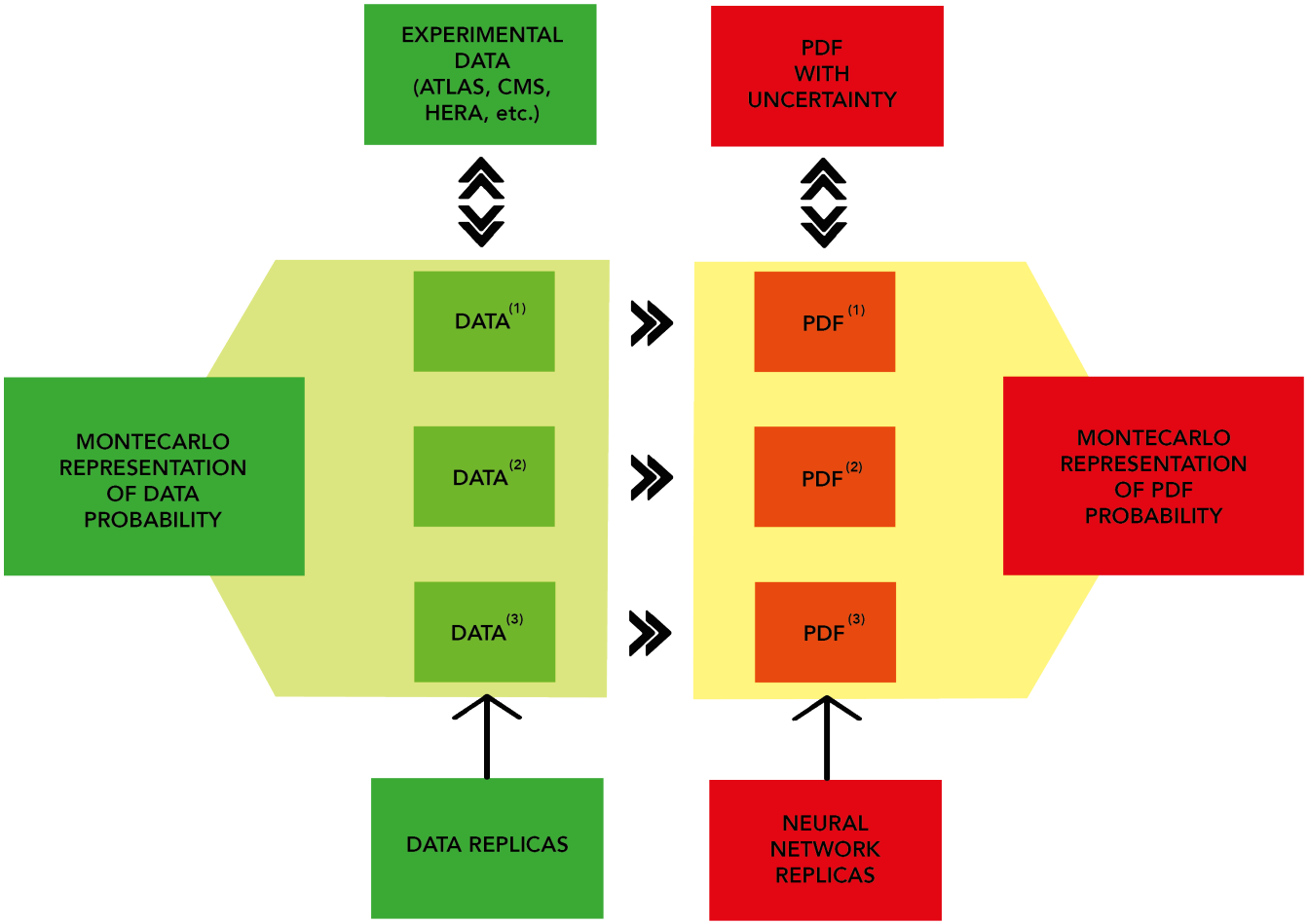
Generalization [edit]

The difference between optimization and machine learning arises from the goal of generalization: while optimization algorithms can minimize the loss on a training set, machine learning is concerned with minimizing the loss on unseen samples. Characterizing the generalization of various learning algorithms is an active topic of current research, especially for deep learning algorithms.

- CAN WE TEST IT? $\Rightarrow$ VALIDATION
- CAN WE UNDERSTAND IT? $\Rightarrow$ EXPLANATION (XML)

PDF/NNPDF RECAP SEQUENCE
THE FUNCTIONAL MONTE CARLO

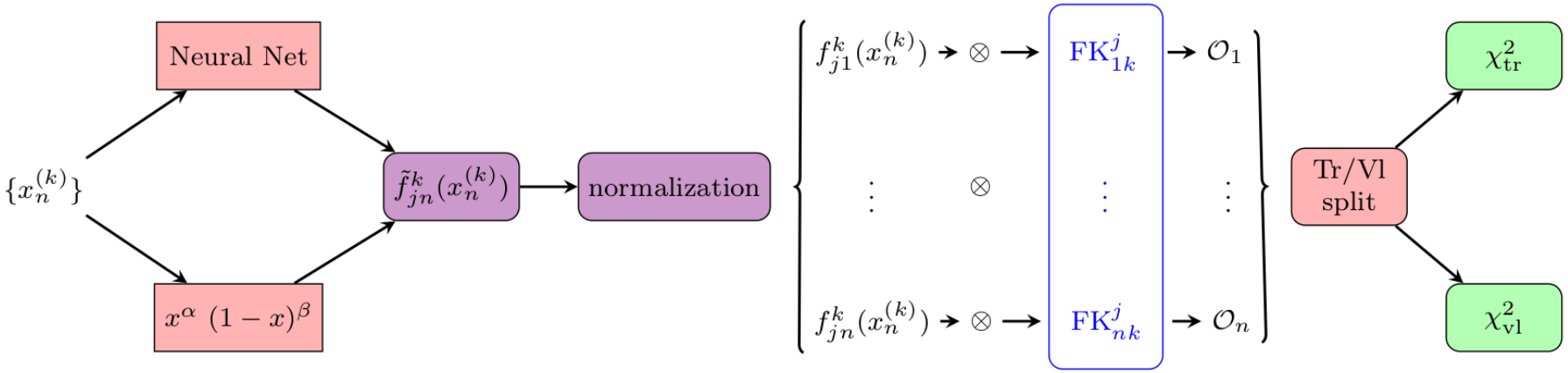
REPLICA SAMPLE OF FUNCTIONS $\Leftrightarrow$ PROBABILITY DENSITY IN FUNCTION SPACE
KNOWLEDGE OF LIKELIHHOD SHAPE (FUNCTIONAL FORM) NOT NECESSARY

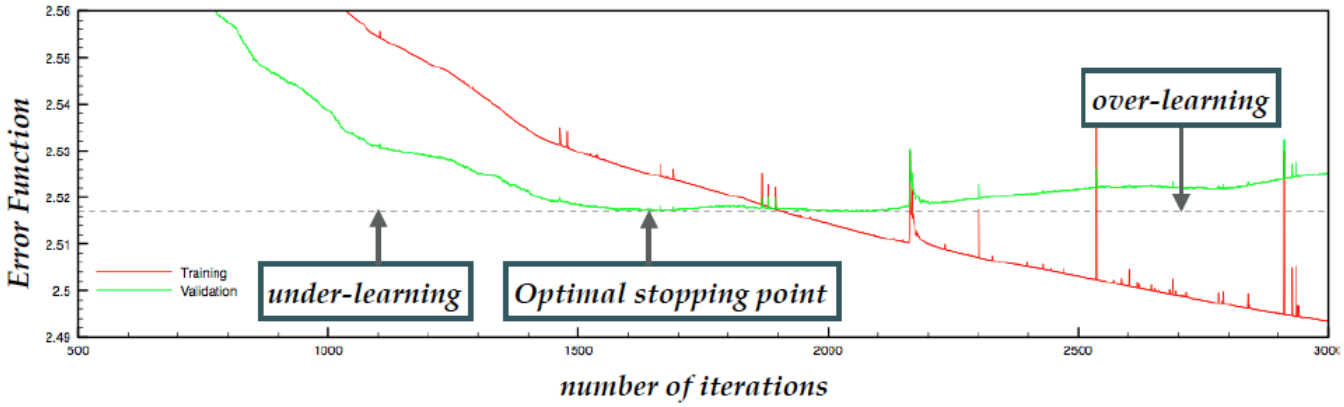


FINAL PDF SET: $f_i^{(a)}(x,\mu)$;
i =up, antiup, down, antidown, strange, antistrange, charm, gluon; $j = 1, 2, \ldots N_{\rm rep}$

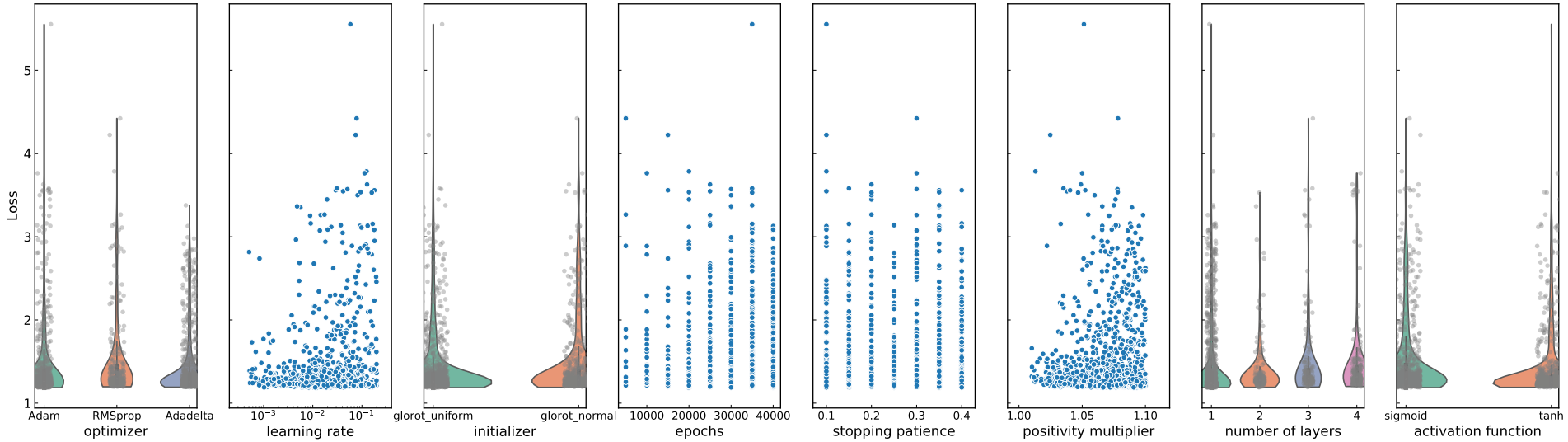
MINIMIZATION AND CROSS-VALIDATION

- 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 $\chi^2 \Rightarrow$ OPTIMAL FIT





FITTING THE METHODOLOGY
HYPEROPTIMIZATION

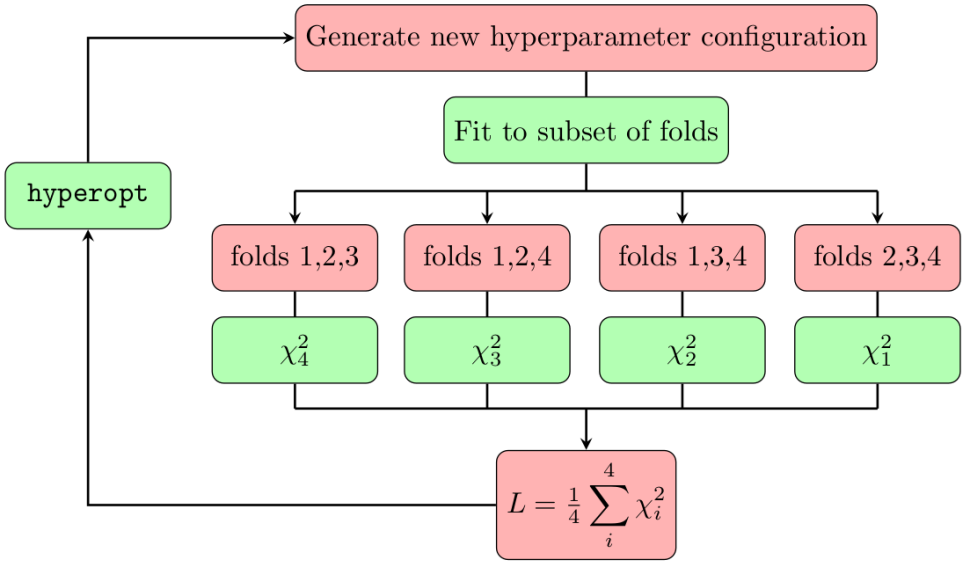


HYPEROPT PARAMETERS

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-FOLDING



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^{ll}_T, 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		

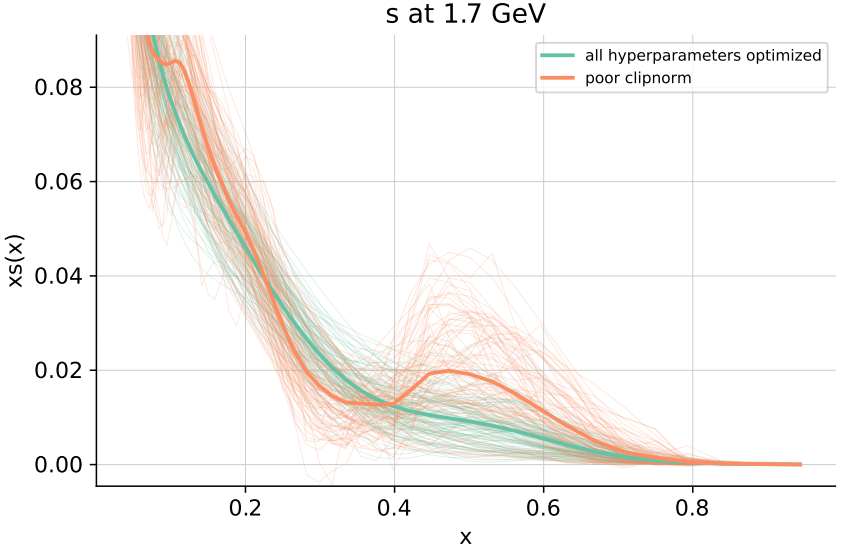
Fold 2		
HERA I+II inc CC e^-p	HERA I+II inc NC e^+p 460 GeV	HERA comb. $\sigma^{\rm red}_{bb}$
NMC p	NuTeV σ^{ν}_c	LHCb $Z \to ee$ 2 fb
CMS W asymmetry 840 pb	ATLAS Z p_T 8 TeV (p^{ll}_T, M_{ll})	D0 $W \to \mu\nu$ asymmetry
DY E886 $\sigma^p_{\rm DY}$	ATLAS direct photon 13 TeV	ATLAS dijets 7 TeV, R=0.6
ATLAS single antitop y (normalised)	CMS $\sigma^{\rm tot}_{tt}$	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 σ^{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
- LOSS: AVERAGE χ^2 OF NON-FITTED FOLDS
- OVERFITTING REMOVED $\Rightarrow$ CORRECT GENERALIZATION

K-FOLDING VS NO K-FOLDING

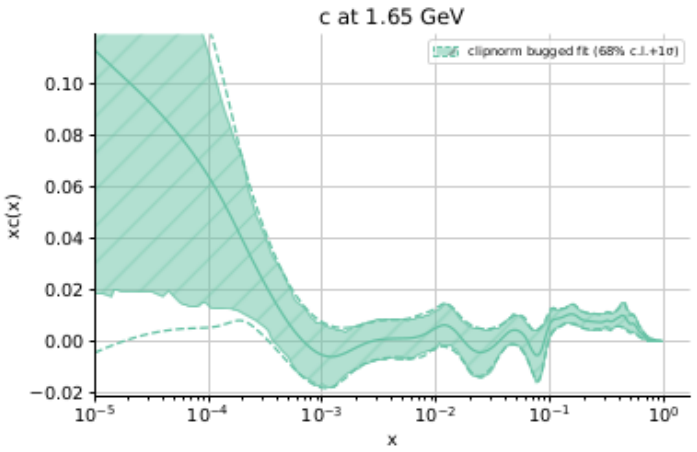


CAN WE TRUST IT?
VALIDATION: OVERFITTING METRIC

- RECOMPUTE VALIDATION χ^2
  - SAME TRAINING-VALIDATION SPLIT
  - DIFFERENT FLUCTUATED VALIDATION DATA
- COMPUTE AVERAGE χ^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

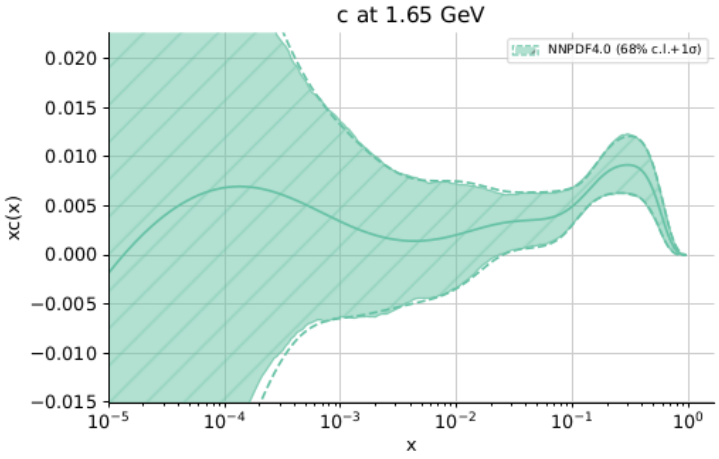
CHARM PDF

OVERFIT (NO CLIPNORM)



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

PROPER FIT (NNPDF4.0)

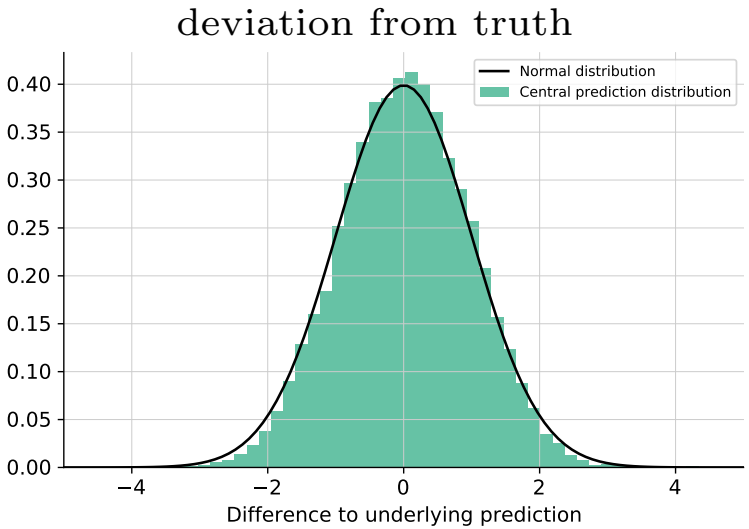


$\mathcal{R}_O = -0.001 \pm 0.013$

CAN WE TRUST IT?
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 VALUES OF OBSERVABLES (NOT FITTED)
  - BIAS/VARIANCE: MEAN SQUARE DEVIATION WR TO TRUTH VS UNCERTAINTY
  - IS TRUTH WITHIN ONE SIGMA 68% OF TIMES?

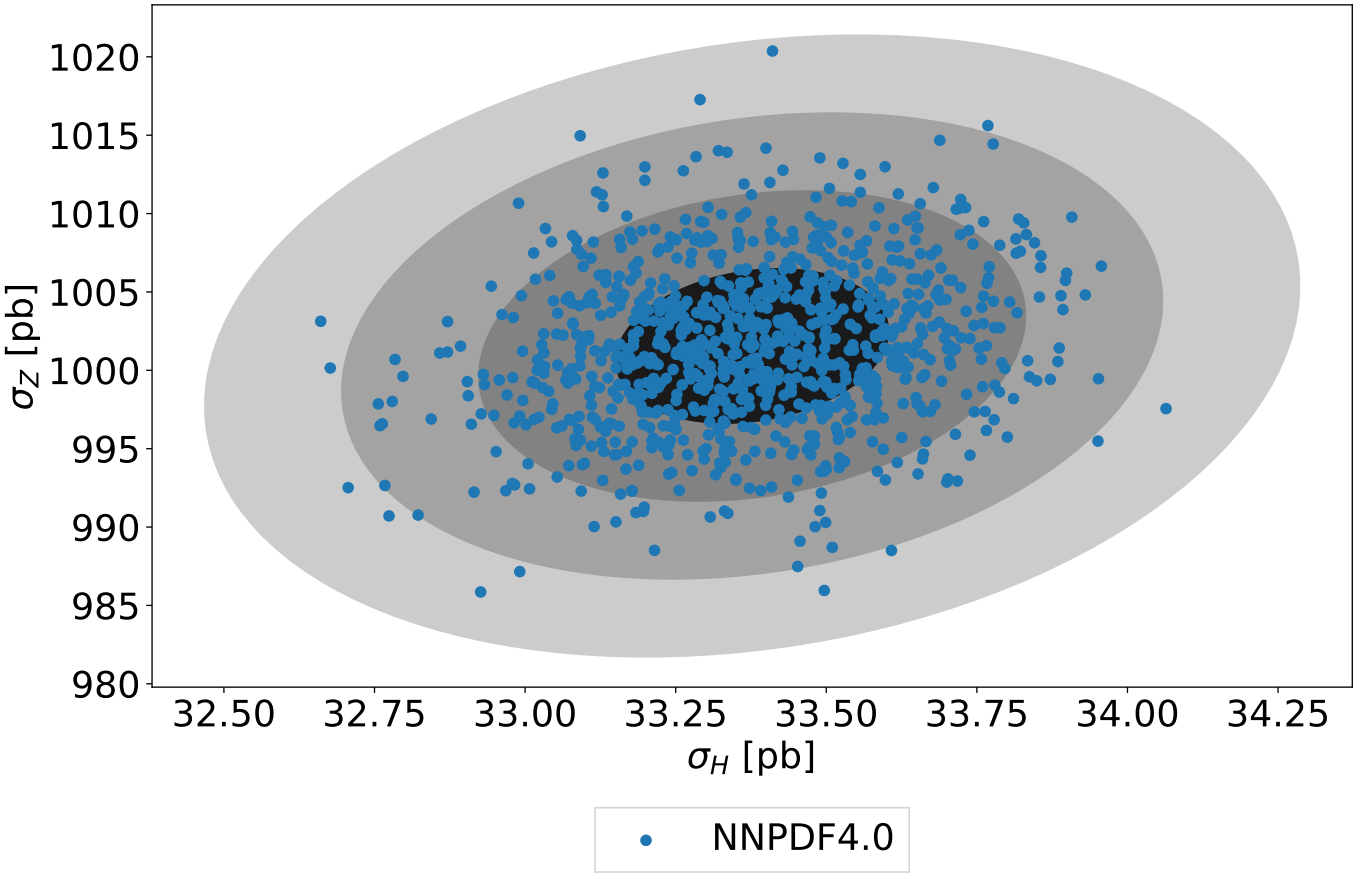
RESULTS



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

CAN WE TRUST IT?
EXPLANATION: HOW DO RESULTS LOOK LIKE?

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



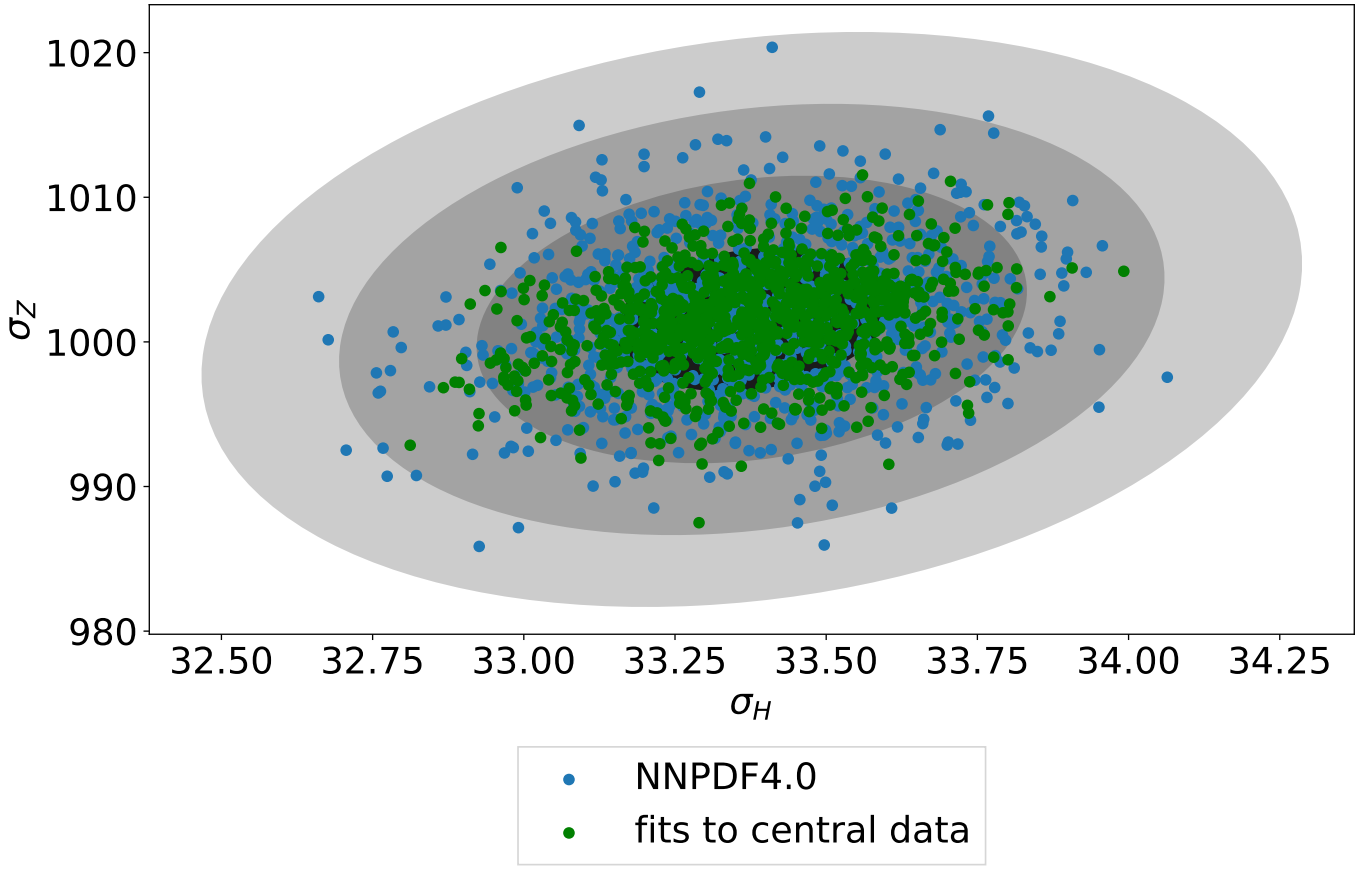
DISTRIBUTION OF REPLICAS DRIVEN BY

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

EXPLANATION

THE REPLICA DISTRIBUTION

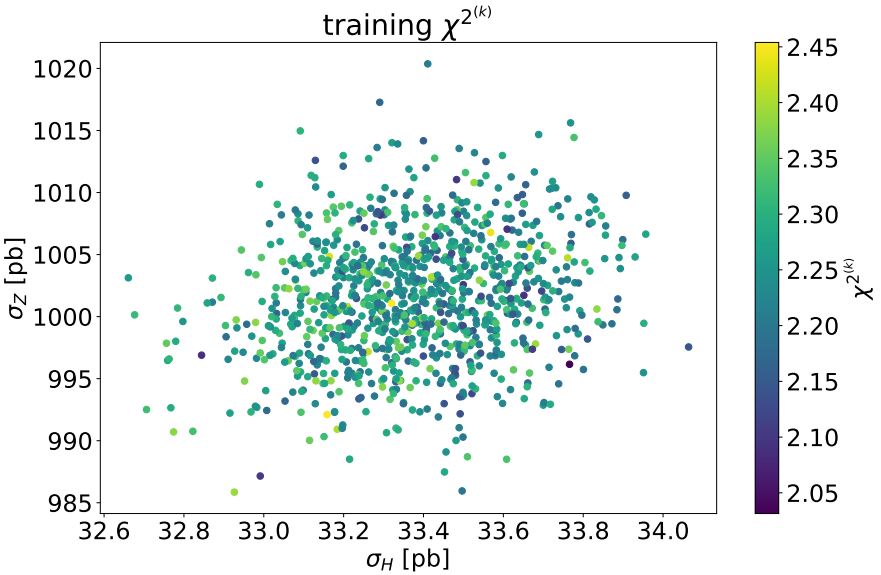
- REPLICA FLUCTUATION $\Rightarrow$ DATA UNCERTAINTIES
- NO REPLICA FLUCTUATION $\Rightarrow$ FIT DEGENERACY

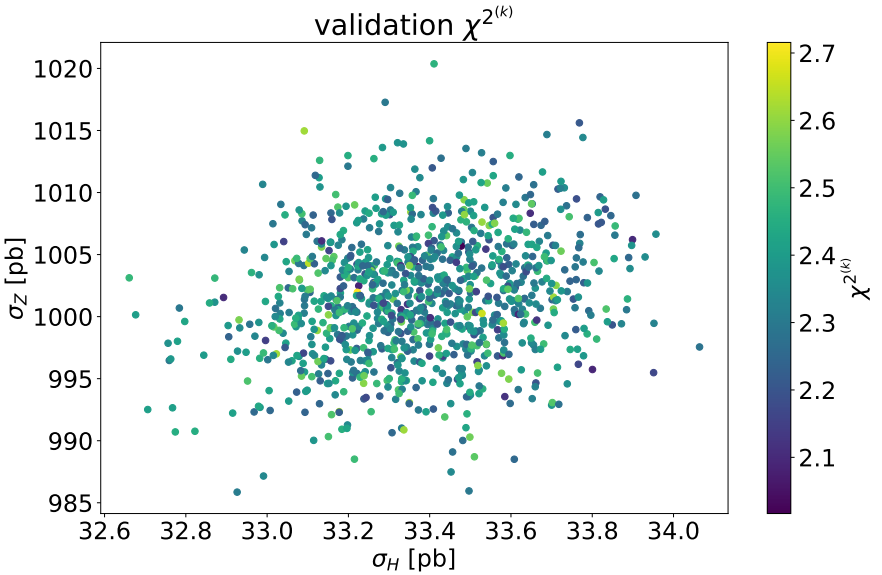


EXPLANATION

THE REPLICA DISTRIBUTION

ARE ALL FITS EQUALLY GOOD?



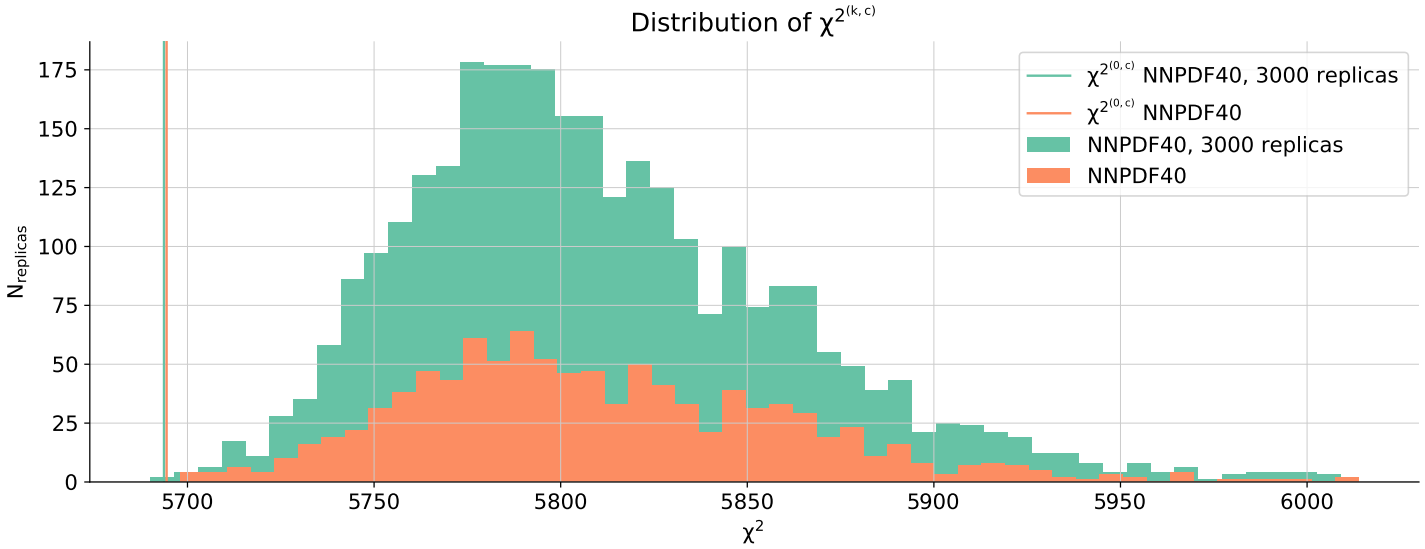


- COMPARE TRAINING AND VALIDATION χ^2 FOR EACH REPLICA
- NO CORRELATION BETWEEN FIT QUALITY AND POSITION IN THE (σ_H, σ_Z) PLANE
- UNIFORM FIT QUALITY

THE REPLICA DISTRIBUTION

COMPARISON TO CENTRAL DATA

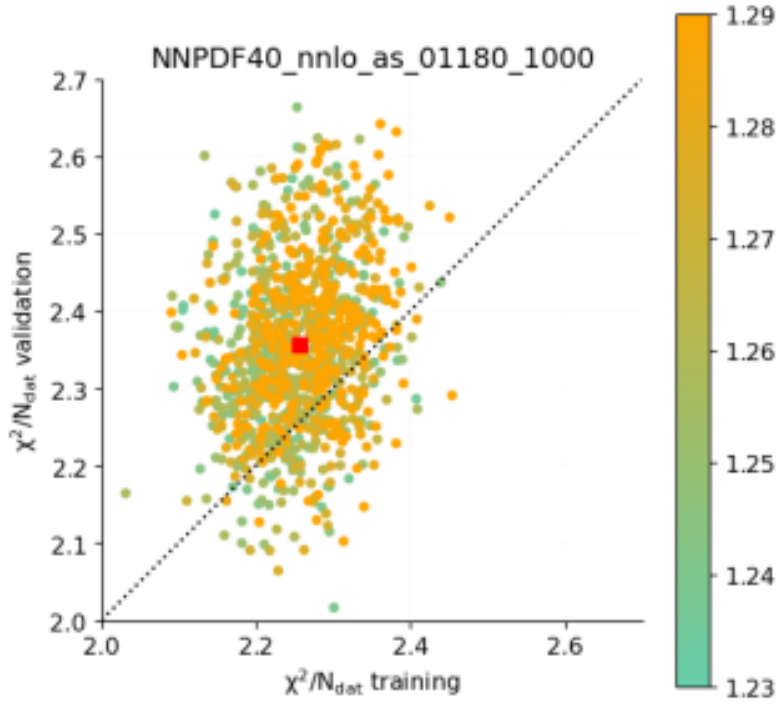
- EACH PDF REPLICA FITTED TO A DATA REPLICA
- FIT QUALITY TO CENTRAL DATA STATISTICALLY DISTRIBUTED



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

THE REPLICA DISTRIBUTION
COMPARISON TO CENTRAL DATA

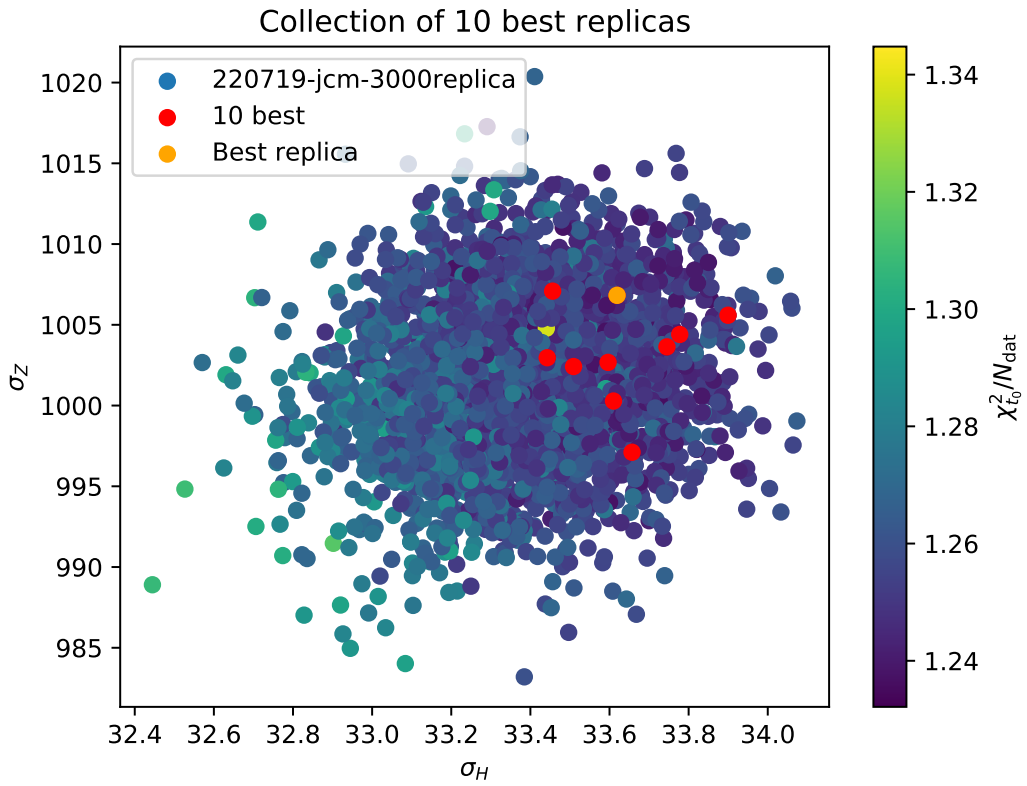
- ARE FITS WITH HIGH χ^2 TO CENTRAL DATA POOR (UNDERLEARNT)?



- 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
- BOTH?

THE REPLICA DISTRIBUTION

COMPARISON TO CENTRAL DATA



χ^2 TO CENTRAL DATA

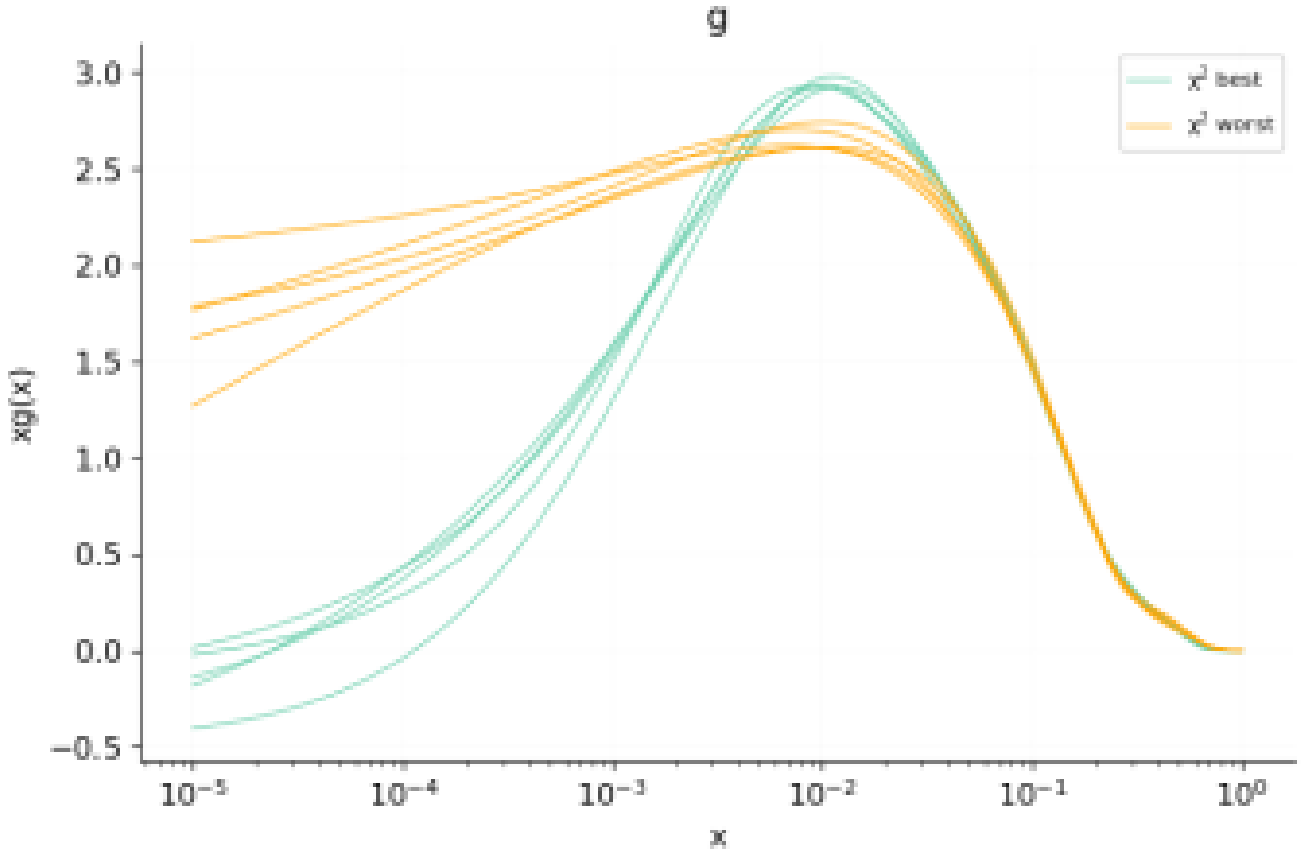
- CORRELATED TO POSITION IN (σ_H, σ_z) PLANE
- CORRELATED TO A FEATURE?

EXPLANATION

LOOKING FOR FEATURES

REPLICAS WITH LOWEST & HIGHEST χ^2 TO CENTRAL DATA

THE GLUON



- REPLICAS CLOSER TO CENTRAL DATA $\Rightarrow$ MORE STRUCTURE
- CORRELATED TO A FEATURE?

EXPLANATION

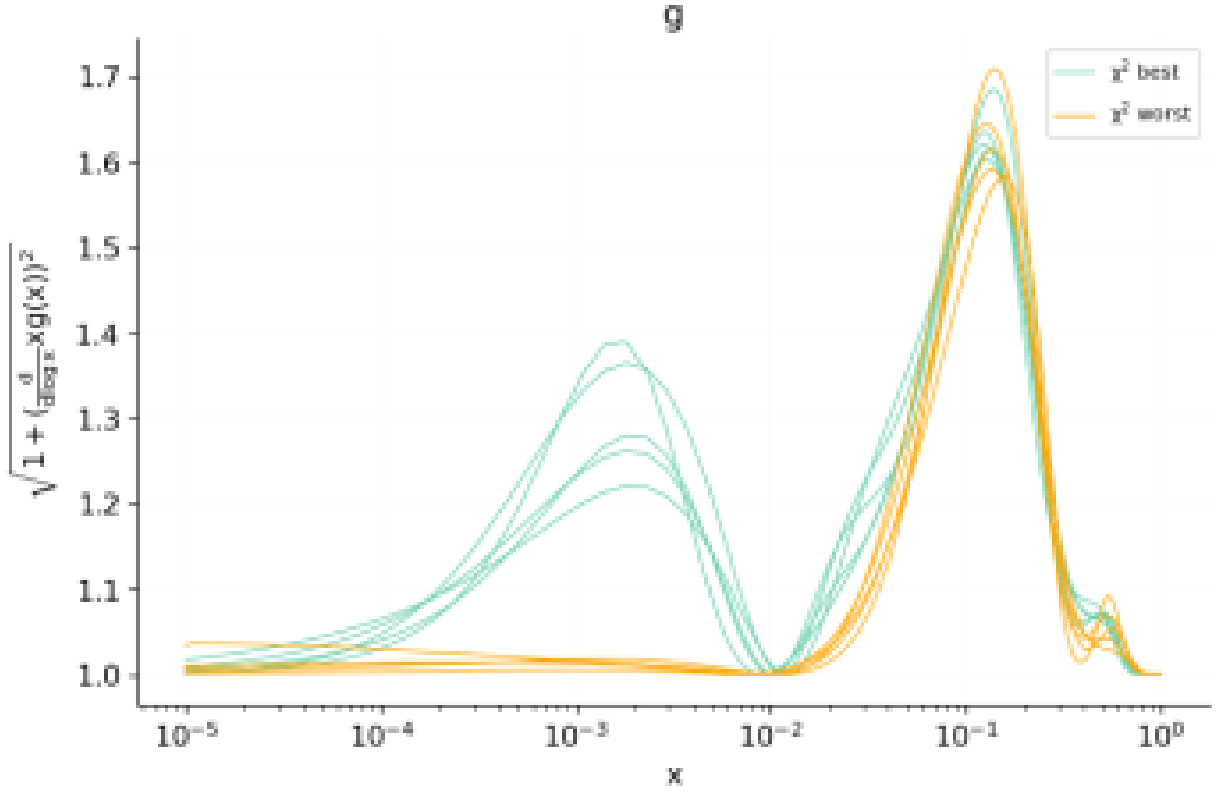
THE PDF KINETIC ENERGY

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 $\Rightarrow$ MORE STRUCTURE
- HIGHER KINETIC ENERGY

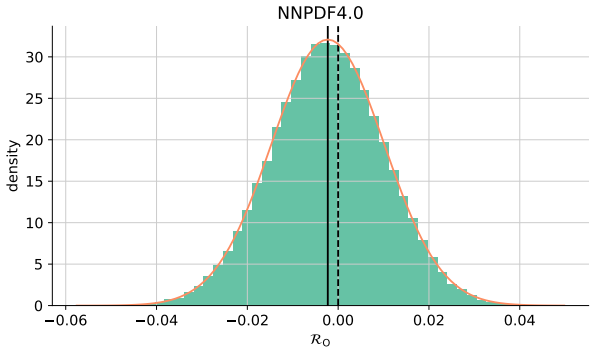
EXPLANATION

OVERLEARNING?

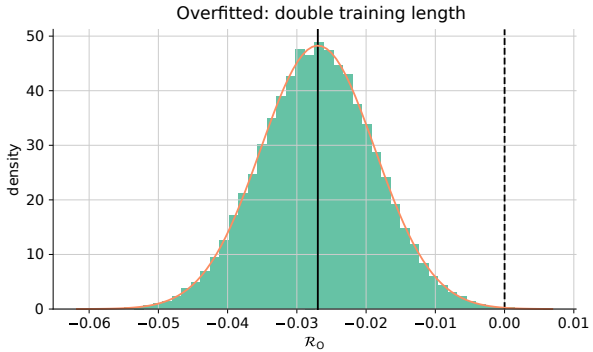
- FORCE OVERLEARING

THE OVERFIT METRIC

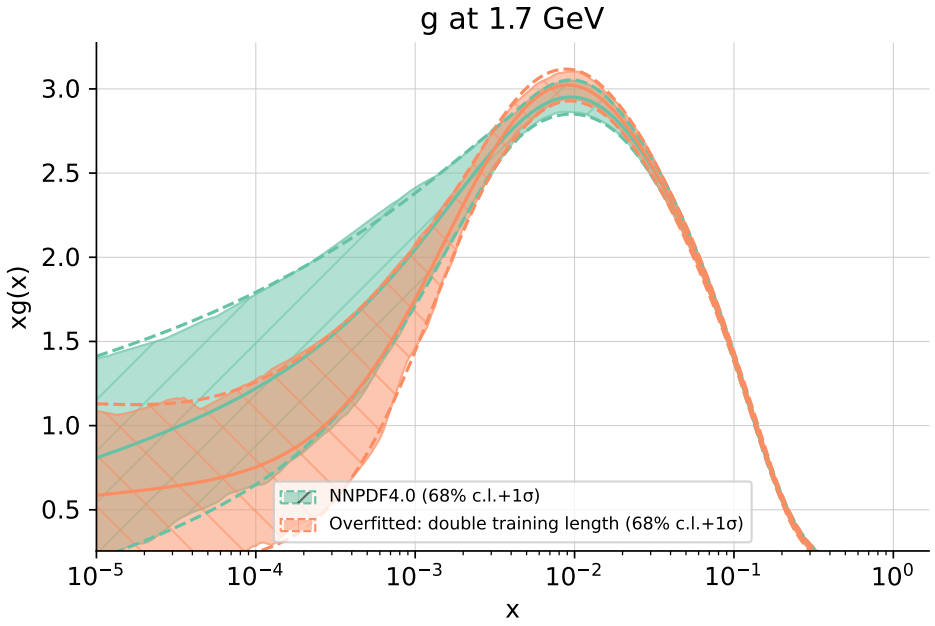
DEFAULT NNPDF4.0



OVERFIT VARIANT



THE GLUON



- LOOK AT THE OUTPUT ⇒ MORE STRUCTURE IN GLUON

EXPLANATION

A PARADOX?

- BEST FIT TO CENTRAL DATA CORRELATED TO HIGH ARCLENGTH
- HIGH ARCLENGTH CORRELATED TO OVERLEARNING
- TRAINING/VALIDATION BALANCE UNCORRELATED TO QUALITY OF FIT TO CENTRAL DATA

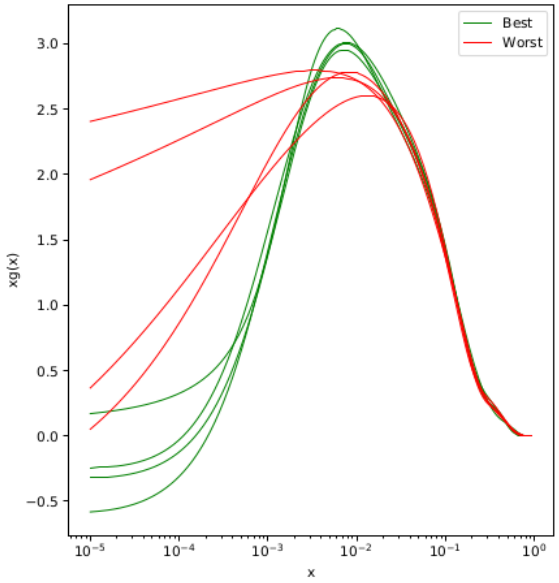
EXPLANATION

GENERALIZATION!

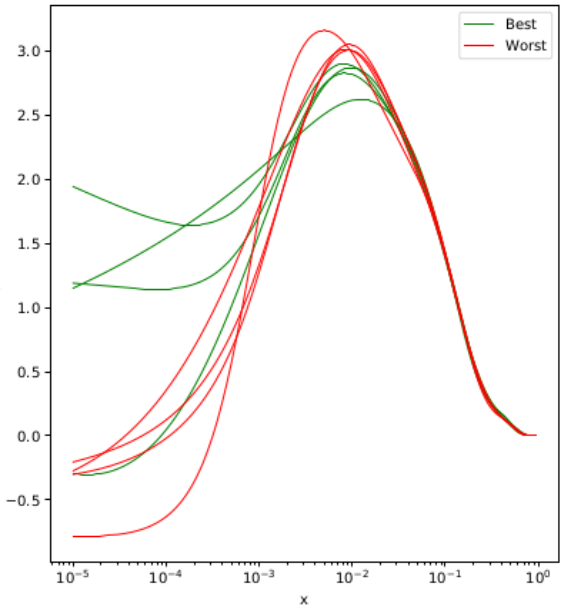
- OVERFITTING CAN MEAN POOR GENERALIZATION
- KEPT IN CHECK BY K-FOLDING (NOT CROSS-VALIDATION)
- LOOK AT BEST χ^2 TO FITTED VS. EXCLUDED FOLDS

THE GLUON

FITTED FOLDS



EXCLUDED FOLD



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

SO, CAN WE TRUST IT?
ASK CHATGPT!

can one trust machine learning for the determination of parton distributions?

Machine learning can be a useful tool for the determination of parton distributions, but whether one can trust the results obtained using this approach depends on various factors.

Parton distribution functions (PDFs) describe the probability distribution of partons, such as quarks and gluons, within a hadron, such as a proton. Determining PDFs from experimental data is a complex and challenging task, and machine learning algorithms can be used to assist in this process. In recent years, several studies have demonstrated the potential of machine learning techniques for PDF determination, and some have shown promising results.

However, it's important to keep in mind that machine learning algorithms are not perfect, and their results should always be interpreted with caution. One potential concern is overfitting, which occurs when a model is trained too well on a particular dataset and ends up being unable to generalize to new data. This can lead to biased or unreliable results. Another concern is the need for a large and diverse dataset to train the machine learning model, as the accuracy of the model depends on the quality and quantity of the training data.

Therefore, while machine learning can be a valuable tool for PDF determination, it's crucial to validate the results obtained using this approach against other methods and to carefully assess the reliability and accuracy of the machine learning models used.