MACHINE LEARNING

PRECISION HIGH-ENERGY PHYSICS

Stefano Forte
Università di Milano & INFN

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI FISICA

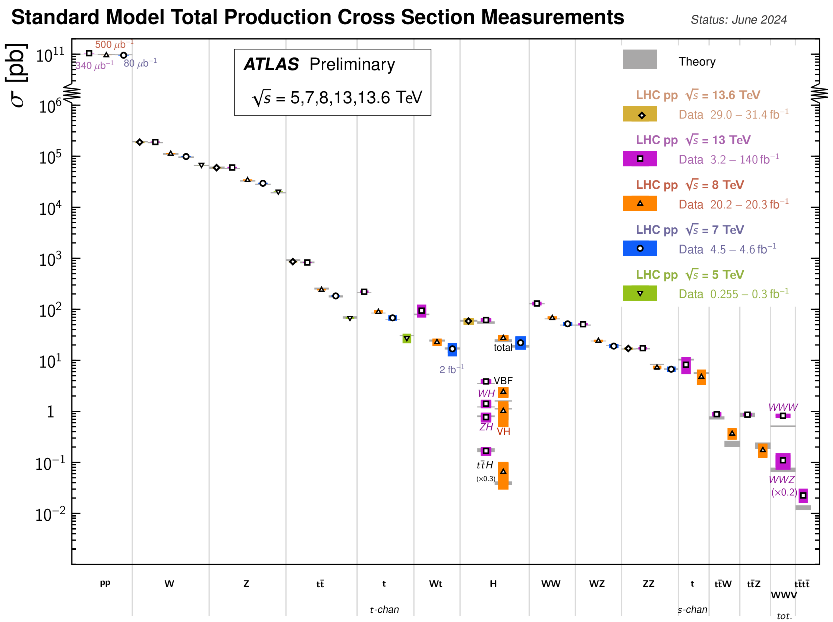
Istituto Nazionale di Fisica Nucleare

Physics in the AI era

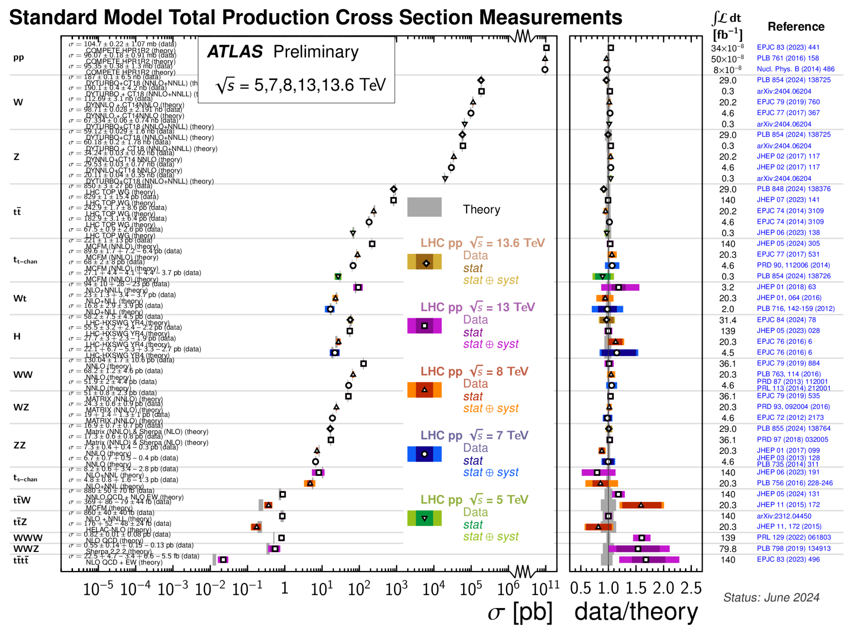
Pisa, September 26, 2024

PRECISION HIGH-ENERGY PHYSICS

PARTICLE PRODUCTION PROCESSES AT LHC



RATIO TO THEORY

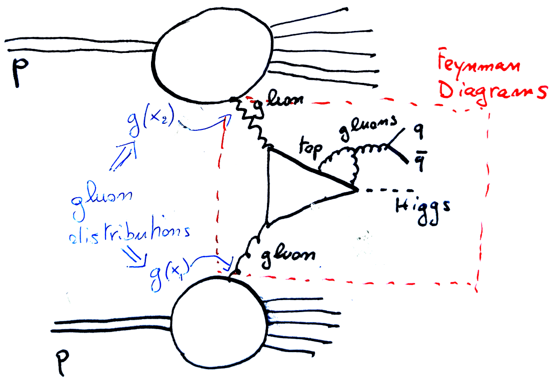


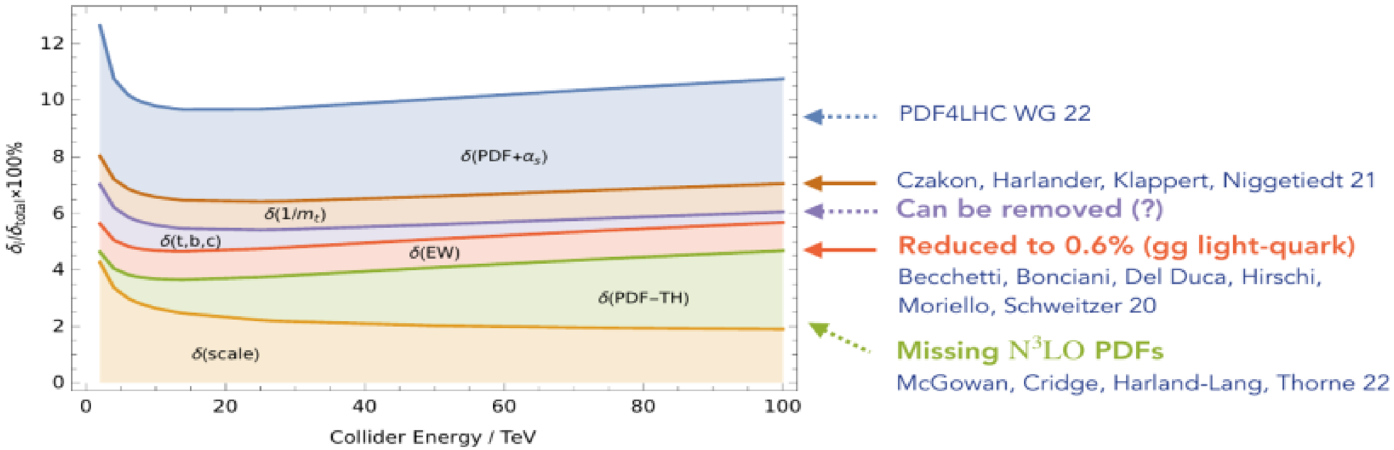
- PRODUCTION RATE PREDICTED OVER ~ 10 ORDERS OF MAGNITUDE
- TYPICAL ACCURACY APPROACHING PERCENT
- LOOKING FOR DEVIATIONS

THE THEORY BOTTLENECK
PROTON STRUCTURE

UNCERTAINTIES:
HIGGS IN GLUON FUSION

QCD FACTORIZATION



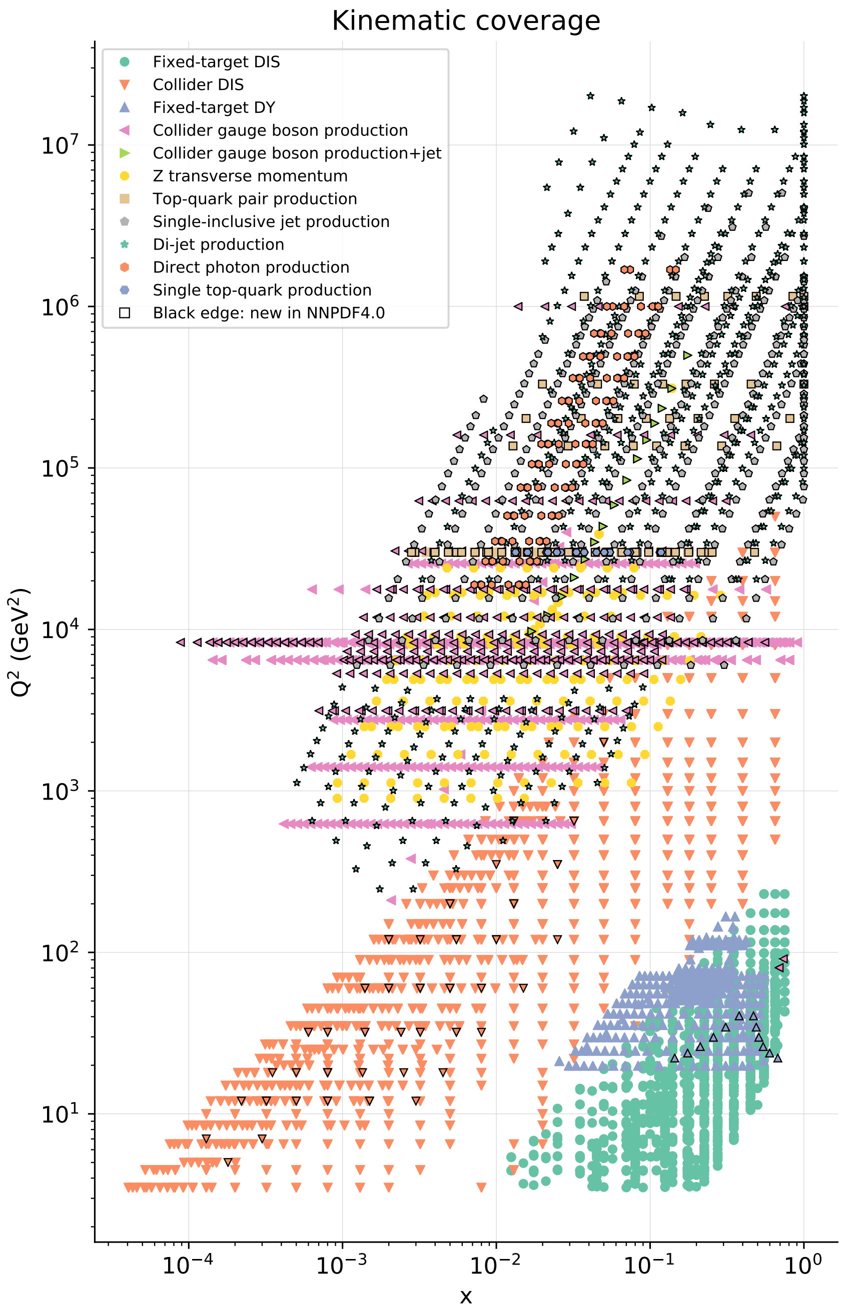


(R. Röntsch, Les Houches 2023)

- PARTON DISTRIBUTIONS (PDF) "PROBABILITY" TO PULL OUT A PROTON CONSTITUENT
- IMPOSSIBLE TO COMPUTE AT PRESENT
- DOMINANT SOURCE OF UNCERTAINTY

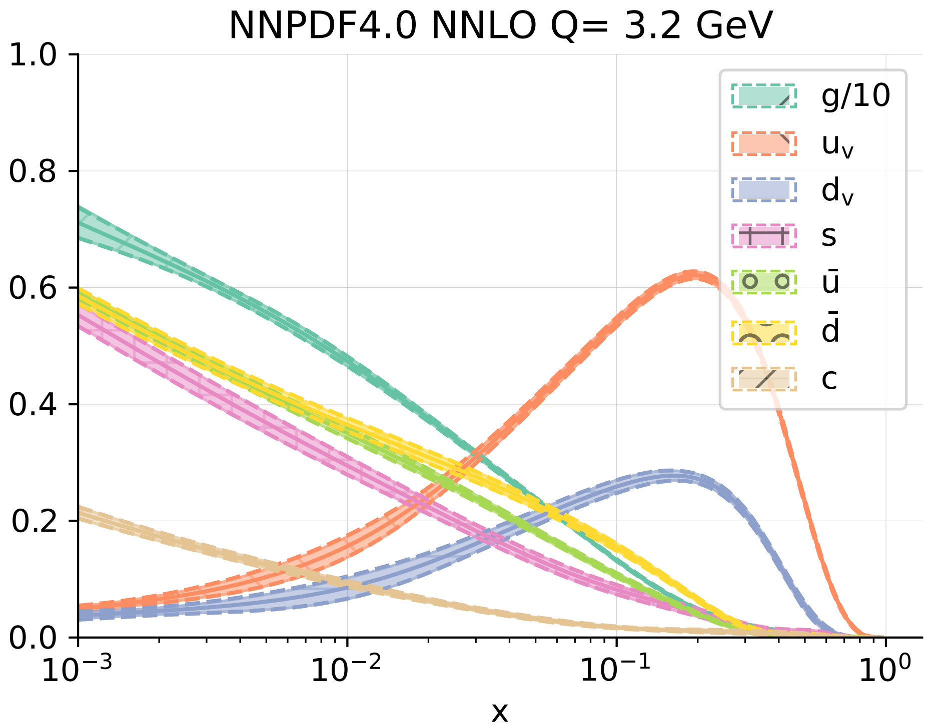
A PATTERN RECOGNITION PROBLEM

A CURRENT DATASET



- COLLISION WITH PROTON(S) $\Rightarrow$ RESULT DEPENDS ON PDF
- COMPUTE RESULT FOR MANY PROCESSES
- COMPARE TO (LOTS) OF DATA

QUALITATIVE BEHAVIOR, QUANTITATIVE PROBLEMS



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

DO WE REALLY NEED MACHINE LEARNING?

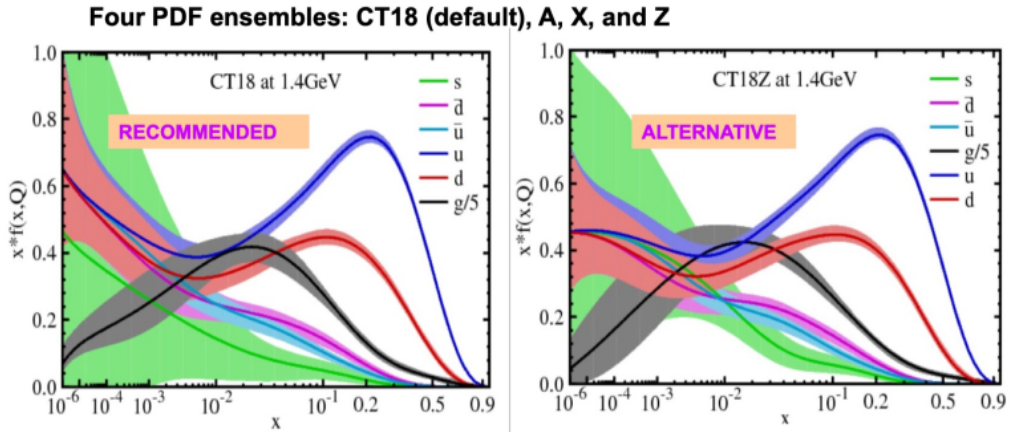
ALTERNATIVE: A MODEL-DEPENDENT APPROACH

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

MORE DATA $\Rightarrow$ BIGGER PARAMETRIZATION (?)

PROLIFERATION OF PDF SETS



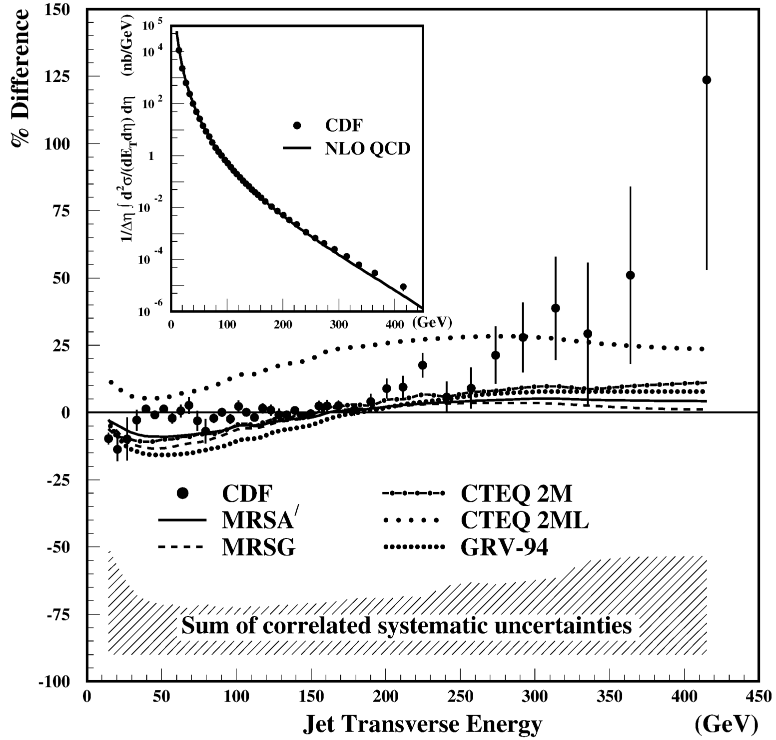
- The CT18 family of PDFs includes LHC data available up to 2018, i.e. mostly 7 and 8 TeV data
- CT18 is the primary PDF; CT18A includes the ATLAS 7 TeV W/Z data (excluded from CT18 due to very poor fit); CT18X includes scale to simulate effects of low x resummation for DIS; CT18Z includes both effects
- 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 (!)

WHAT HAPPENED IN THE PREHISTORY

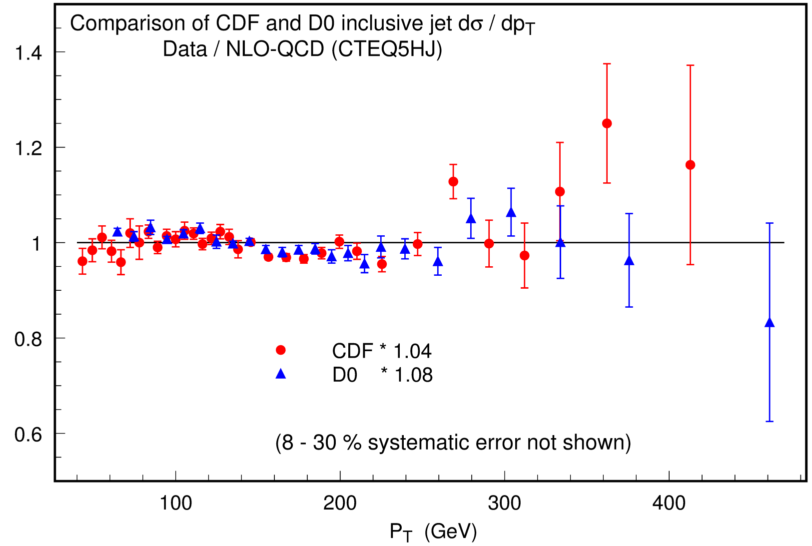
DISCOVERY PHYSICS 1995



- HUGE DATA-THEORY DISCREPANCY
- ~~COMPOSITE QUARKS~~???
- BAD MODELING!

BETTER MODELING ⇒ NO DISCREPANCY

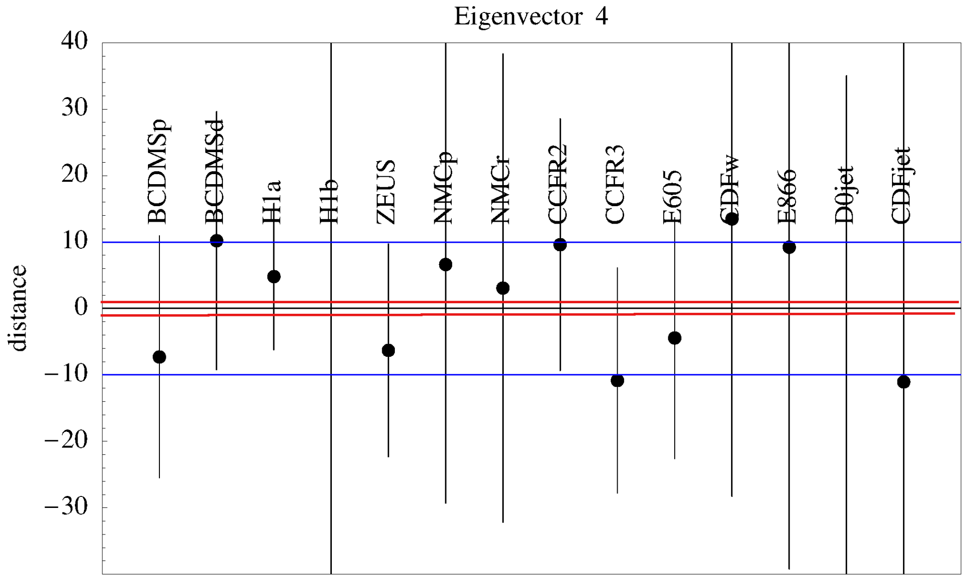
FINAL RESULTS (1998)



WHAT STILL HAPPENS TODAY
"TOLERANCE UNCERTAINTIES"

FIRST PDFS WITH UNCERTAINTIES (2002)

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



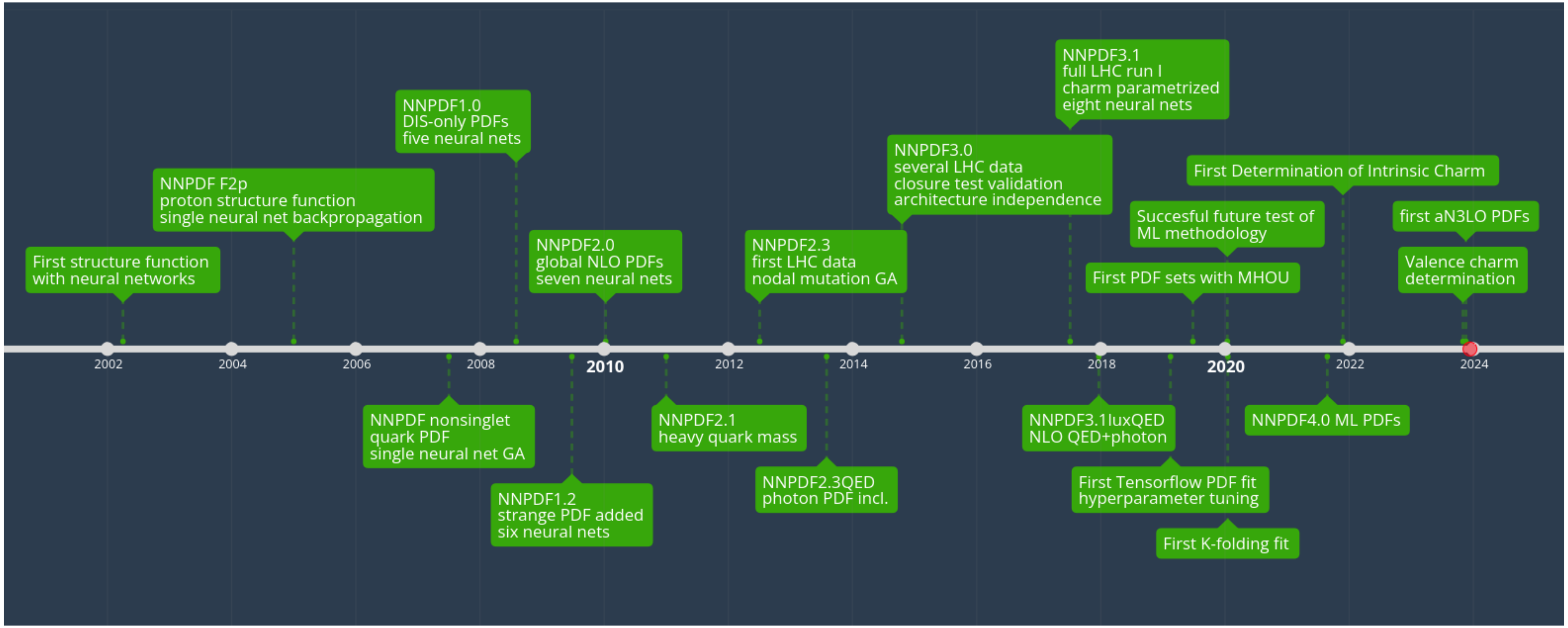
MSHT PDFS (2020)

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

A COOKBOOK RECIPE

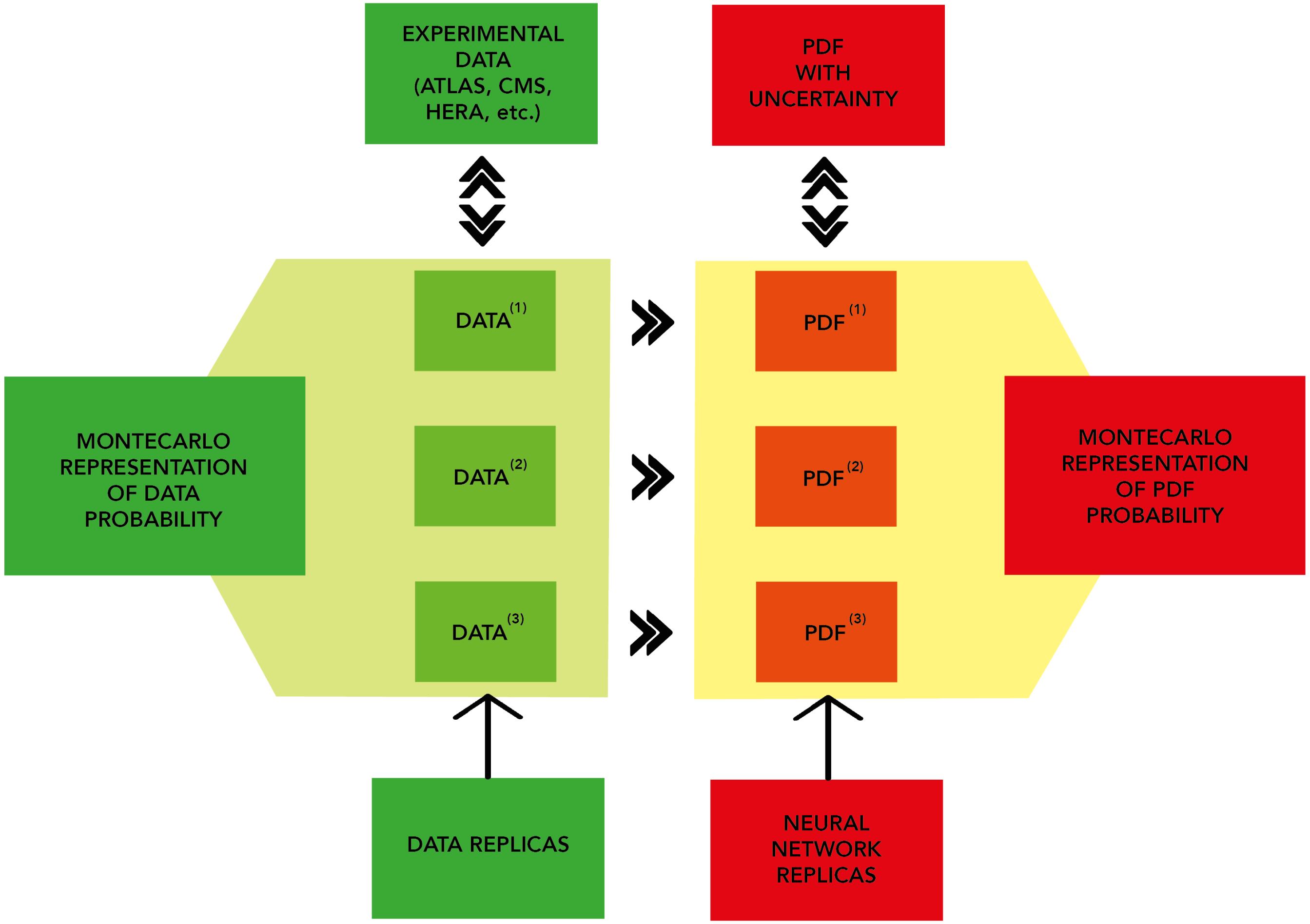
- UNCERTAINTIES RESCALED BY "TOLERANCE" $T \sim 4 \div 10$
- DETERMINED FROM SPREAD OF BEST-FIT FROM DIFFERENT DATA

PROTON STRUCTURE AS A ML PROBLEM
NNPDF



PROBABILITY REGRESSION

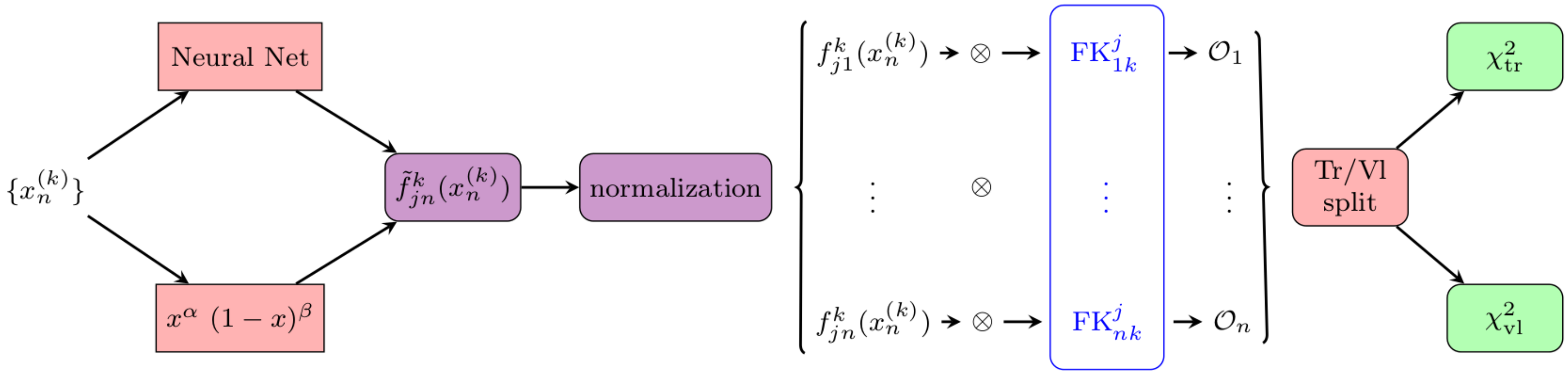
REPLICA SAMPLE OF FUNCTIONS $\Leftrightarrow$ PROBABILITY DENSITY IN FUNCTION SPACE
KNOWLEDGE OF LIKELIHHOD SHAPE (FUNCTIONAL FORM) NOT NECESSARY

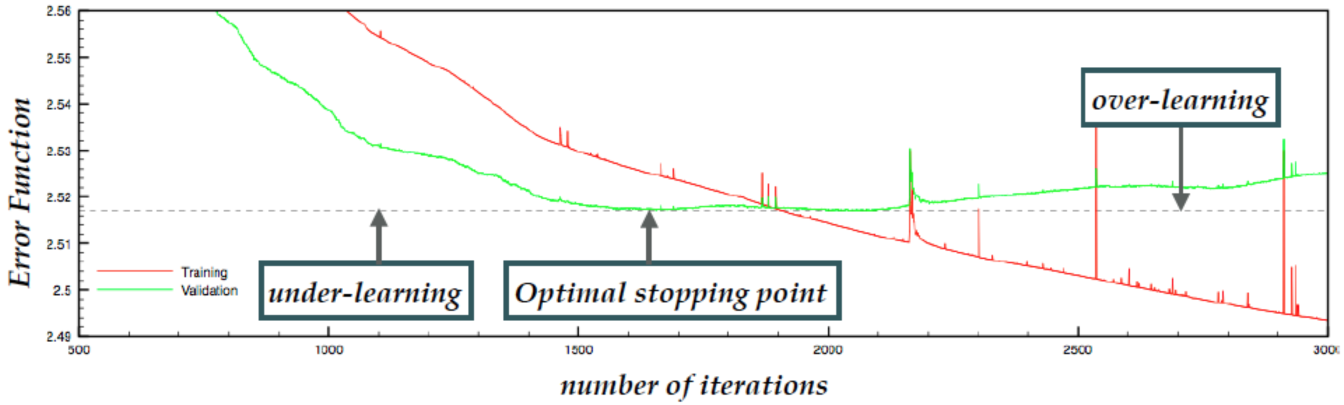


FINAL PDF SET: $f_i^{(a)}(x,\mu)$;
i =up, antiup, down, antidown, strange, antistrange, charm, gluon; $j = 1, 2, \ldots N_{\rm rep}$

CROSS-VALIDATED LEARNING

- MODEL PARAMETERS DETERMINED BY LOSS MINIMIZATION THROUGH GRADIENT DESCENT
- RANDOM TRAINING-VALIDATION SPLIT, LOSS TO TRAINING DATA MINIMIZED
- STOP TRAINING IF VALIDATION LOSS GROWS FOR A WHILE (PATIENCE)
- LOWEST VALIDATION LOSS OPTIMAL LEARNING FIT

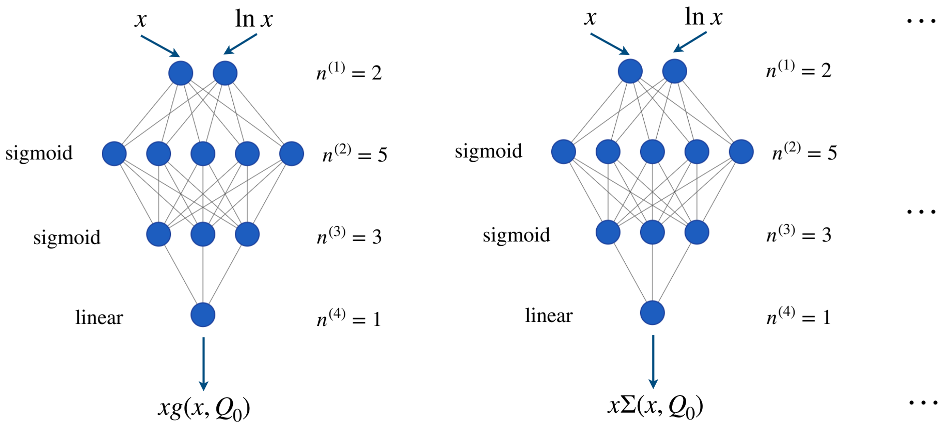




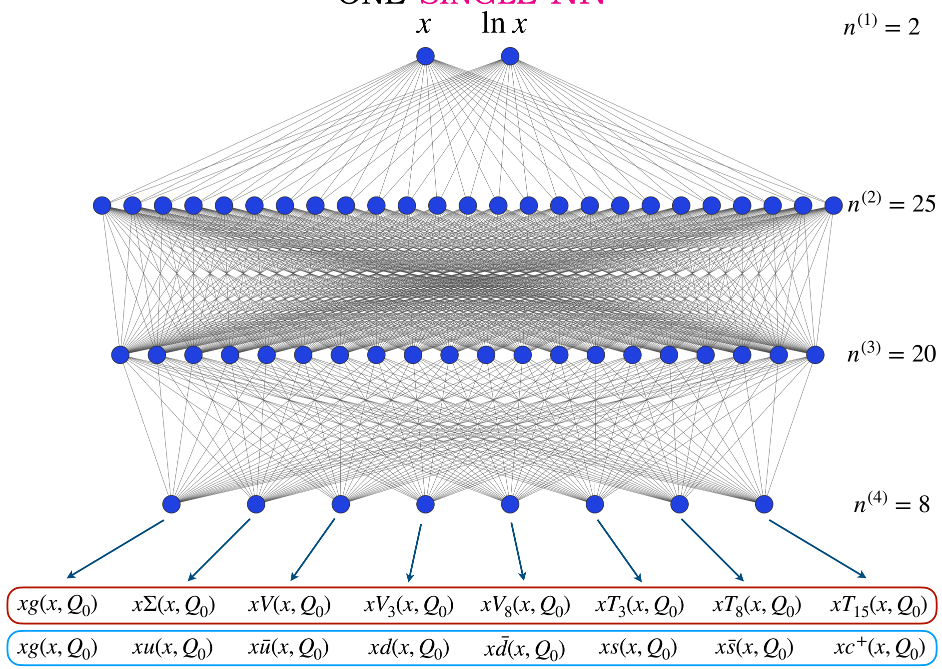
WHICH MODEL?
NEURAL NETWORKS
ARCHITECTURE

- HOW MANY INPUTS?
- HOW MANY INDEPENDENT NNs?

ONE NN PER PDF



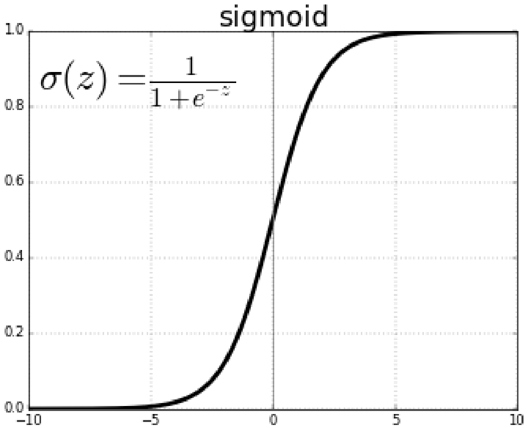
ONE SINGLE NN

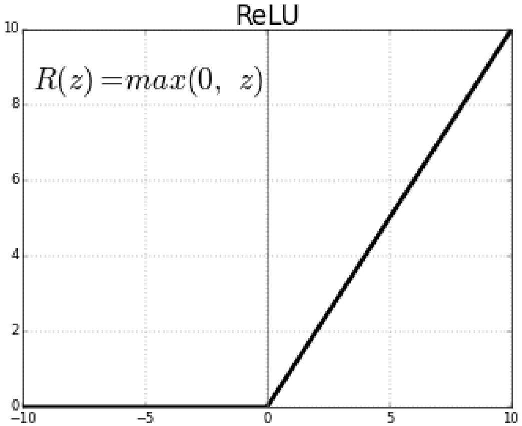


WHICH MODEL?
NEURAL NETWORKS
ACTIVATION FUNCTION

$$F^{(i)}_{\rm out}(\vec{x}_{\rm in}) = F\left(\sum_j \omega_{ij} x^j_{\rm in} - \theta_i\right)$$

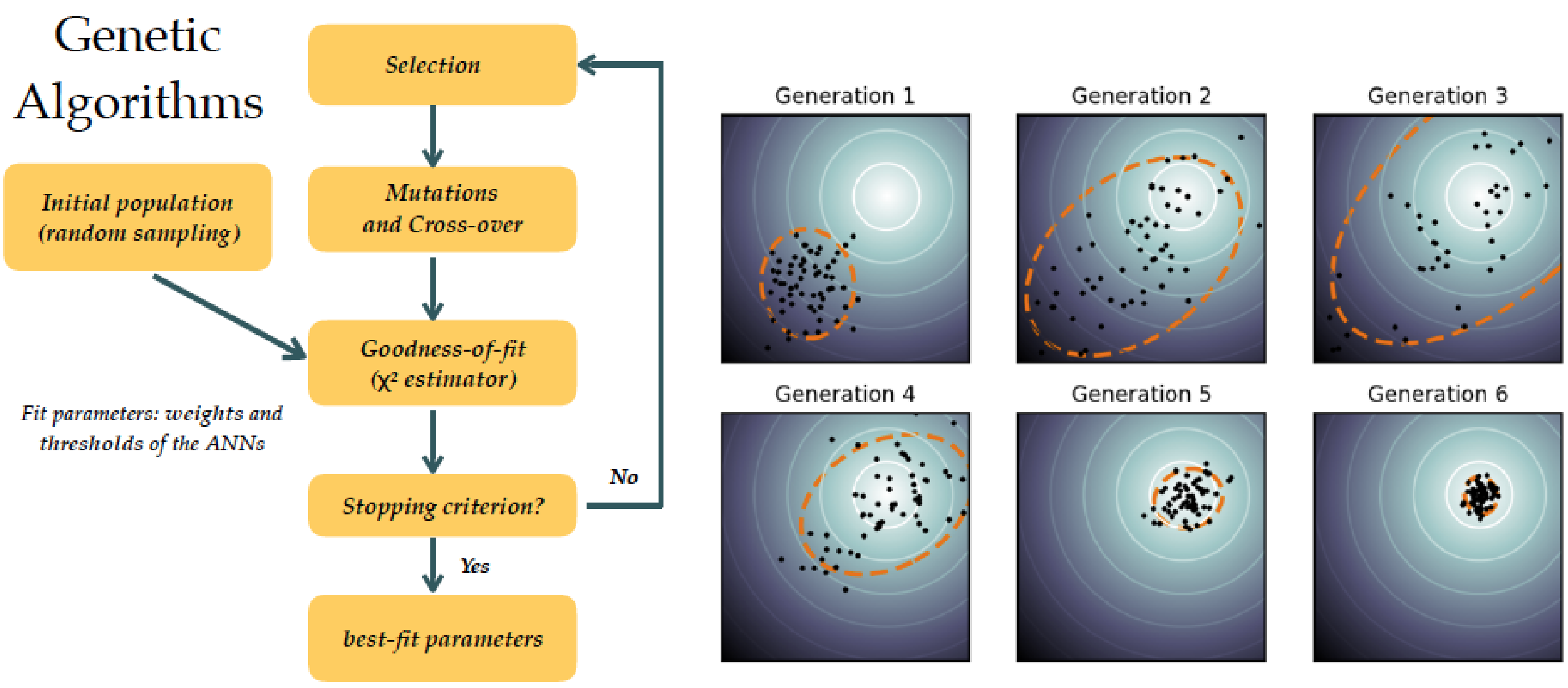
- LINEAR ACTIVATION $\Rightarrow$ MULTILINEAR REGRESSION
- \+ NONLINEAR PROFILE $\Rightarrow$ UNIVERSAL INTERPOL.
  - sigmoid $F(x) = \frac{1}{1+e^{-x}}$
  - arctan $F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan x$
  - RELU $F(x) \begin{cases} 0; & x < 0 \\ x; & x > 0 \end{cases}$





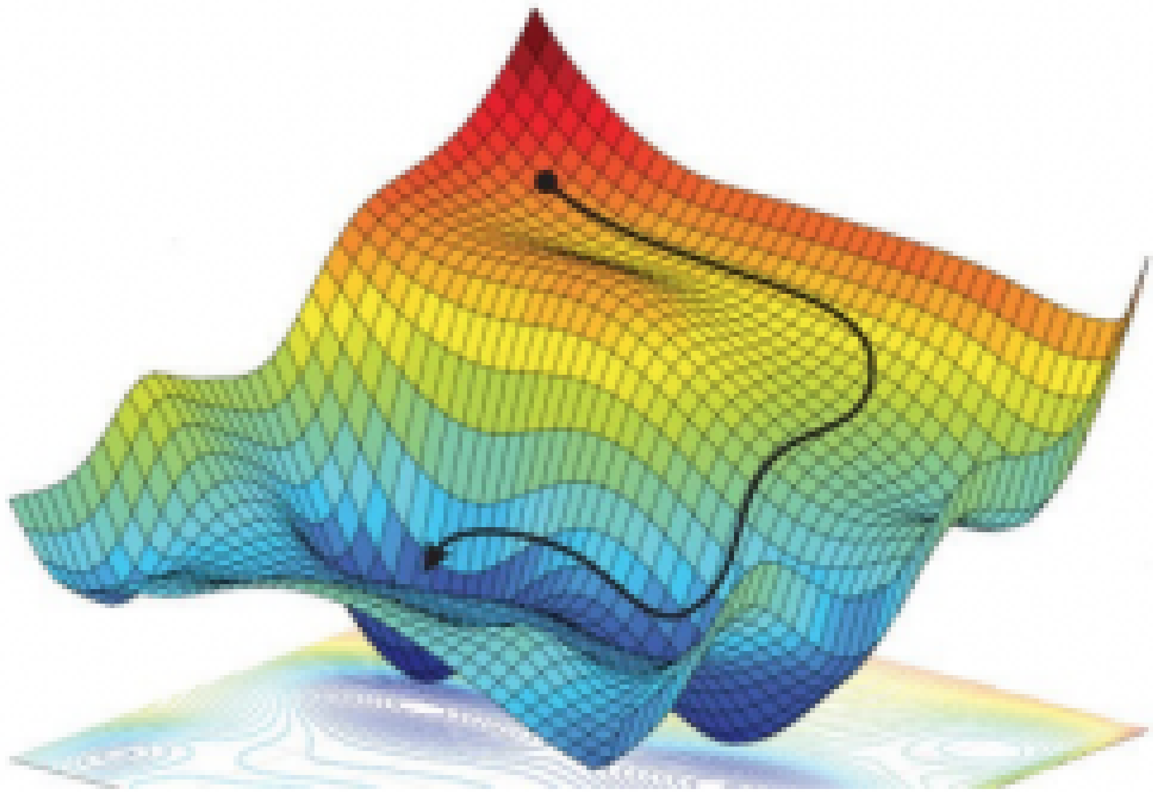
WHICH LEARNING?
GENETIC ALGORITHMS

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



WHICH LEARNING?
GRADIENT DESCENT

- BASIC IDEA: COMPUTE GRADIENT OF LOSS W.R. TO PARAMETERS
- SELECT DIRECTION OF DESCENT



WHICH LEARNING?
DESIDERATA

- FAST CONVERGENCE
- DO NOT STOP ON LOCAL MINIMA
- EXPLORE SPACE OF MINIMA (DEGENERATE CASE)

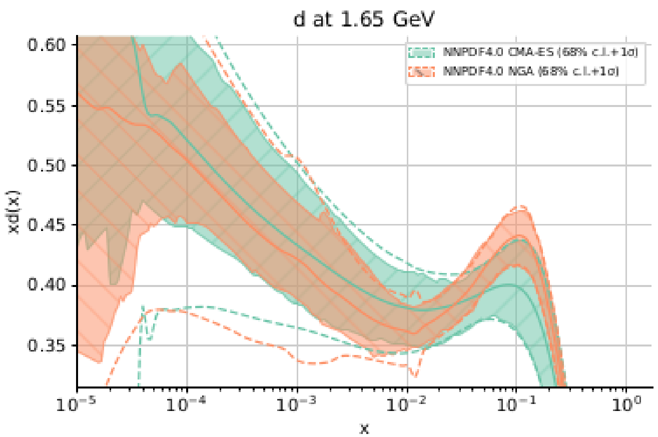
GENETIC ALGORITHMS

- DIFFERENT EPOCHS; VARIABLE MUTATION RATE
- REWEIGHTING DIFFERENT DATA CONTRIBUTIONS TO LOSS
- NODAL MUTATION
- COVARIANCE MATRIX ADAPTATION (CMA)

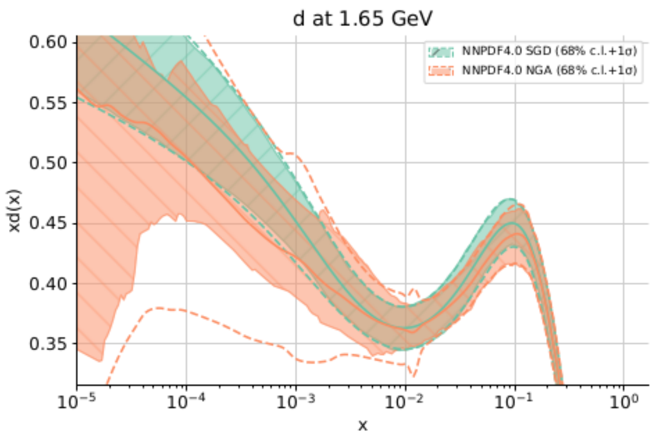
GRADIENT DESCENT

- GLOROT NORMAL/UNIFORM INITALIZATION
- ADAPTIVE GRADIENT / ADAPTIVE MOMENT
- STOCHASTIC GD
- BATCH GD

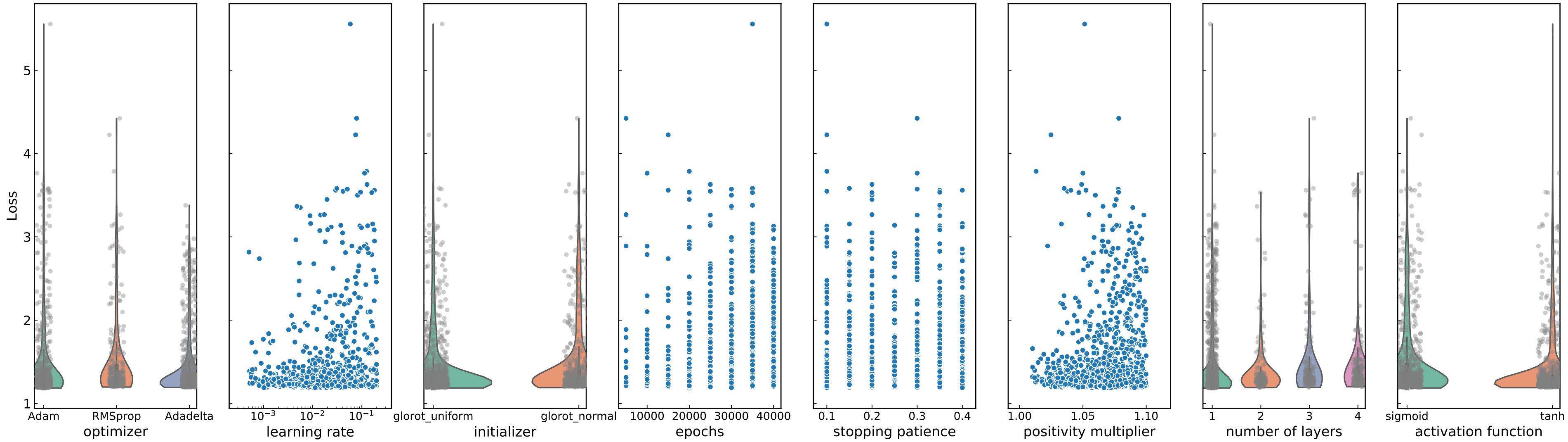
NAIVE GA VS. CMA



GA (NAIVE) VS GD (ADADELTA)



METHODOLOGY HYPEROPTIMIZATION



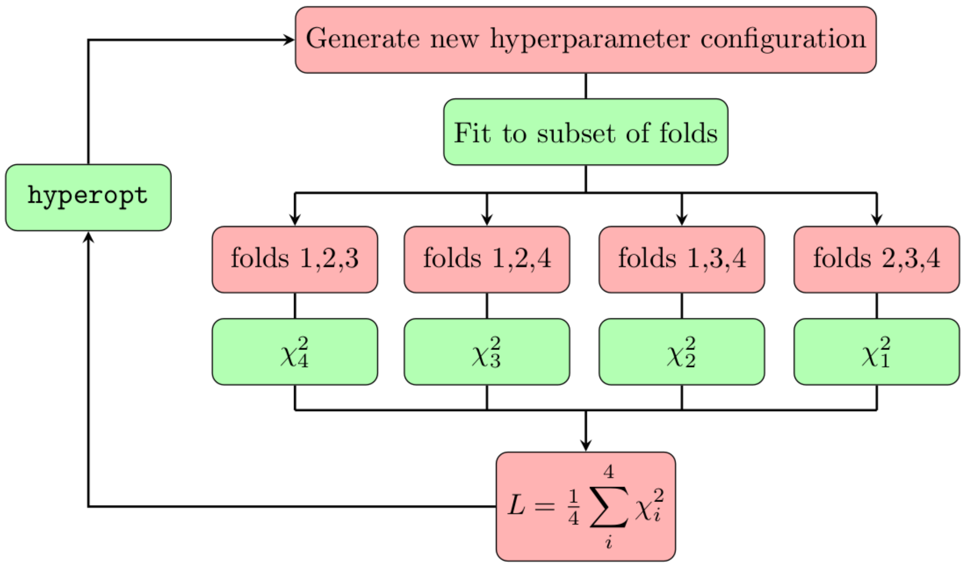
HYPEROPT PARAMETERS

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 LOSS??

BEST RESULT $\Rightarrow$ BEST GENERALIZATION



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		

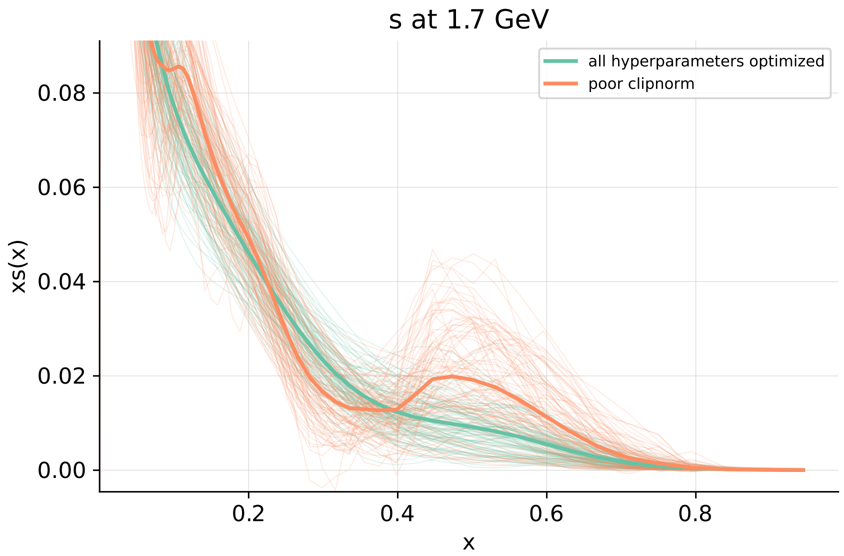
Fold 2		
HERA I+II inc CC e^-p	HERA I+II inc NC e^+p 460 GeV	HERA comb. $\sigma^{\rm red}_{b\bar{b}}$
NMC p	NuTeV $\sigma^{\bar{\nu}}_c$	LHCb $Z \to ee$ 2 fb
CMS W asymmetry 840 pb	ATLAS Z p_T 8 TeV (p^{ll}_T, M_{ll})	D0 $W \to \mu\nu$ asymmetry
DY E886 $\sigma^p_{\rm DY}$	ATLAS direct photon 13 TeV	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 $\sigma^{\bar{\nu}}_{CC}$	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^{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 FIT QUALITY OF NON-FITTED FOLDS
- OVERFITTING REMOVED $\Rightarrow$ CORRECT GENERALIZATION

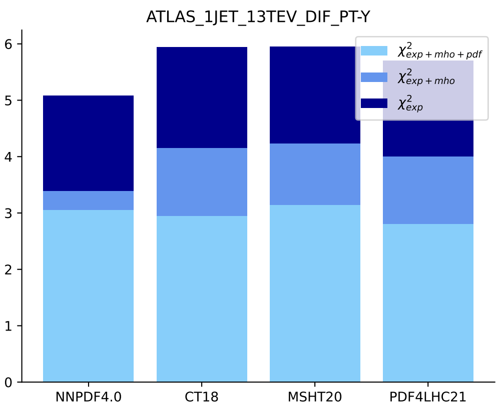
K-FOLDING VS NO K-FOLDING



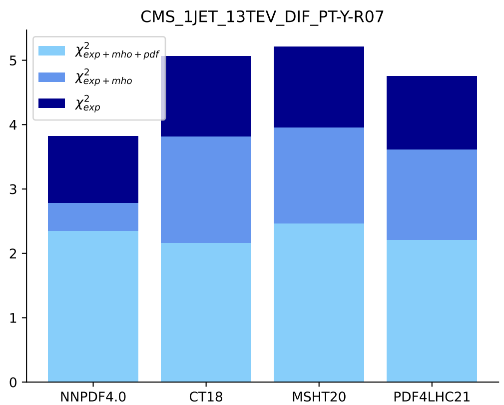
WHAT DOES ML BUY US?
PRECISION + ACCURACY

- AGREEMENT (χ^2) WITH DATA PUBLISHED AFTER PUBLICATION OF NNPDF4.0 PDF SET
- EXP, EXP+TH AND TOTAL (EXP+TH+PDF) UNCERTAINTIES

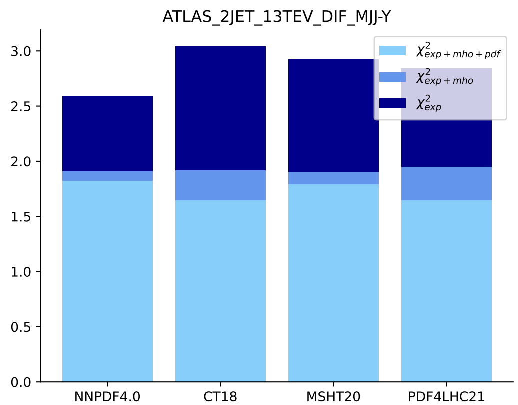
ATLAS JETS



CMS JETS



ATLAS DIJETS



- EXP χ^2 LOWER $\Rightarrow$ NNPDF4.0 AGREES BETTER WITH DATA $\Rightarrow$ MORE PRECISE
- EXP AND TOTAL χ^2 CLOSER $\Rightarrow$ NNPDF4.0 PDF UNCERTAINTIES SMALLER
- AGREEMENT WITH DATA OF ALL PDF SETS COMPARABLE $\Rightarrow$ ALL UNCERTAINTIES FAITHFUL $\Rightarrow$ EQUALLY ACCURATE

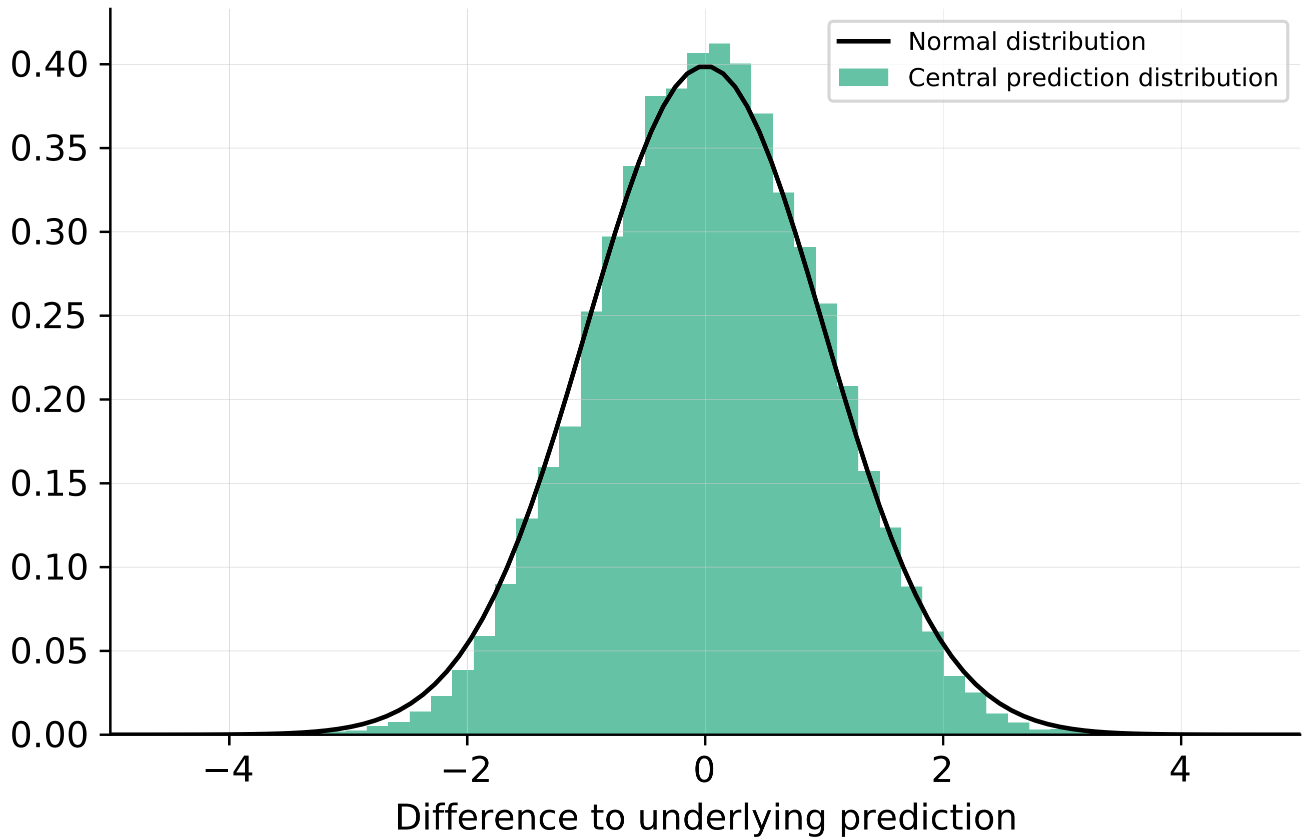
SYSTEMATIC UNCERTAINTY VALIDATION:

CLOSURE TESTS

- 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": IS TRUTH WITHIN ONE SIGMA 68% OF TIMES?

TESTING UNCERTAINTIES

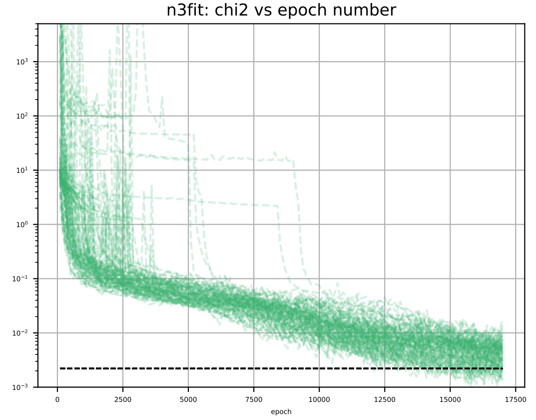
DISTRIBUTION OF DEVIATIONS FROM TRUTH



- COMPARISON OF PREDICTIONS TO TRUTH
- STATISTICS OVER RUNS OF THE UNIVERSE
- CORRECTLY NORMALIZED GAUSSIAN DISTRIBUTION OF OUTCOMES

CLOSURE TEST
UNDERSTANDING UNCERTAINTIES

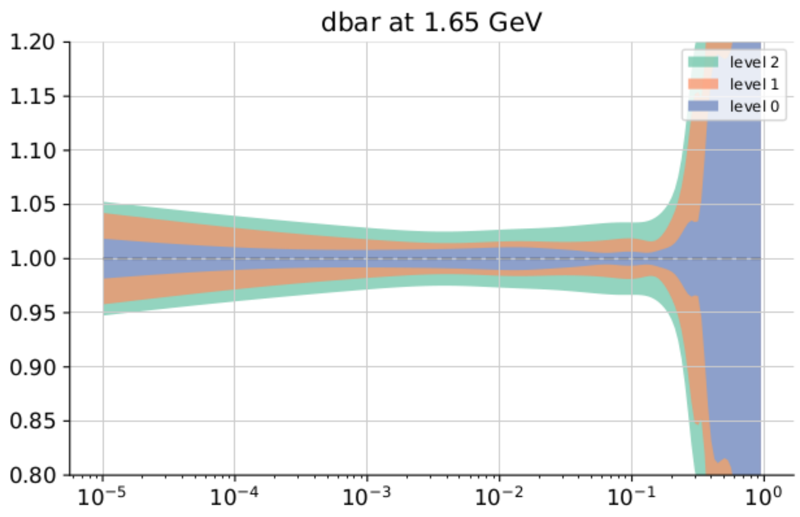
LEVEL 0 LOSS VS TRAINING



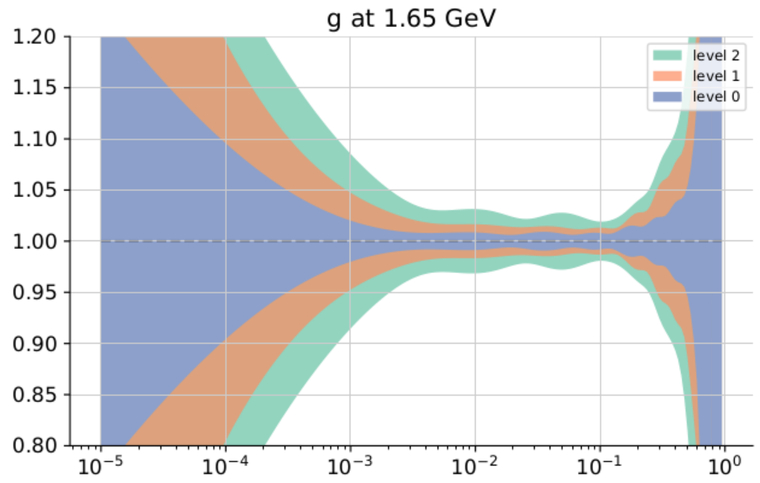
- LEVEL 0 (TRUTH DATA) $\Rightarrow$ PERFECT AGREEMENT ($\chi^2 \approx 0$)
  YET UNCERTAINTY NONZERO
  $\Rightarrow$ NEURAL NETS $\Leftrightarrow$ MANY FUNCTIONAL FORMS
- LEVEL 1 (RUNS OF UNIVERSE) $\Rightarrow$ REPLICAS ALL FITTED TO SAME DATA,
  YET UNCERTAINTY NONZERO
  $\Rightarrow$ DITTO
- LEVEL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

LEVEL 0/1/2 UNCERTAINTIES

ANTIDOWN

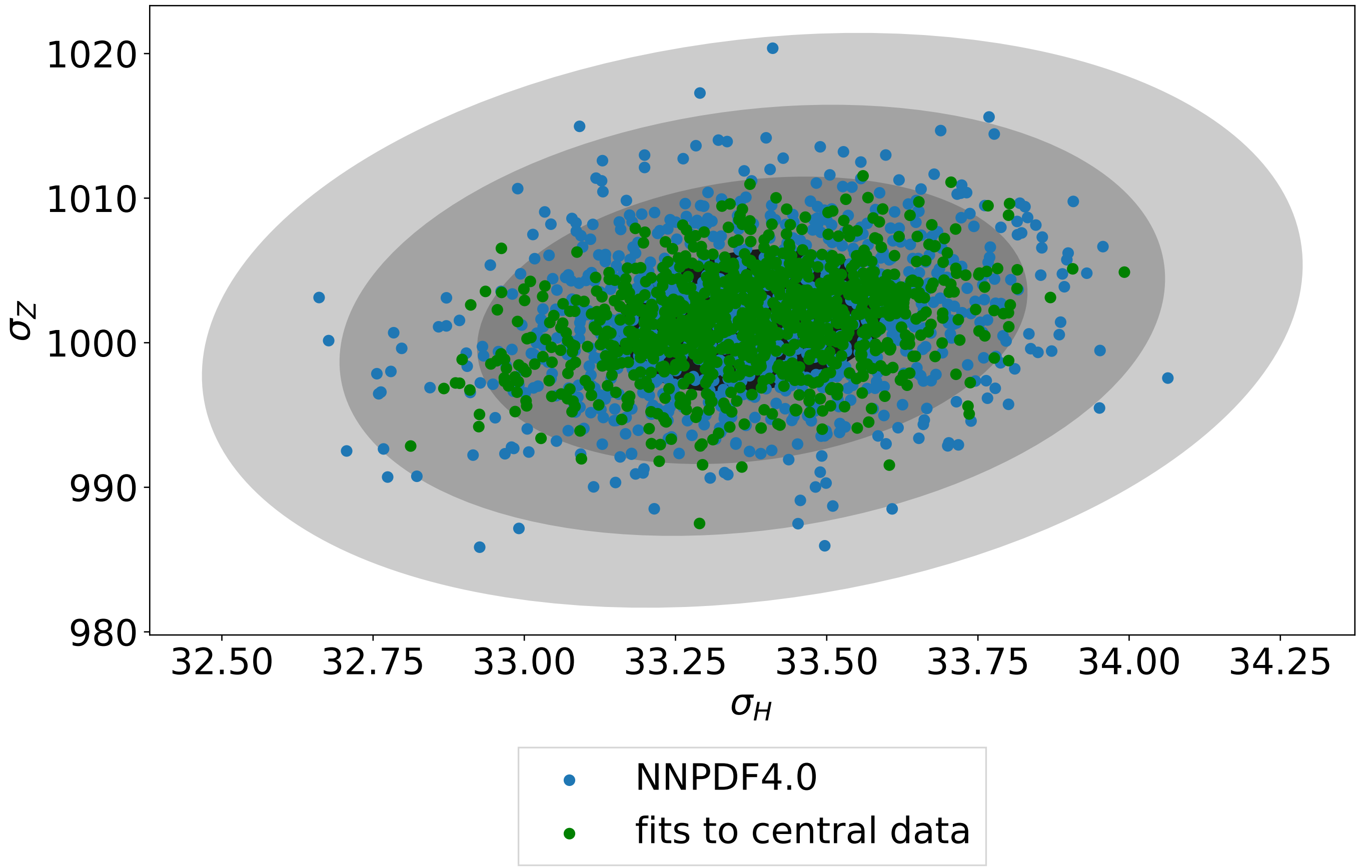


GLUON



UNDERSTANDING UNCERTAINTIES
THE REPLICA DISTRIBUTION

- PLOT RESULTS IN (σ_H, σ_Z) PREDICTION SPACE $\Rightarrow$ GAUSSIAN!
- REPLICA FLUCTUATION $\Rightarrow$ DATA UNCERTAINTIES
- NO REPLICA FLUCTUATION $\Rightarrow$ MODEL UNCERTAINTY

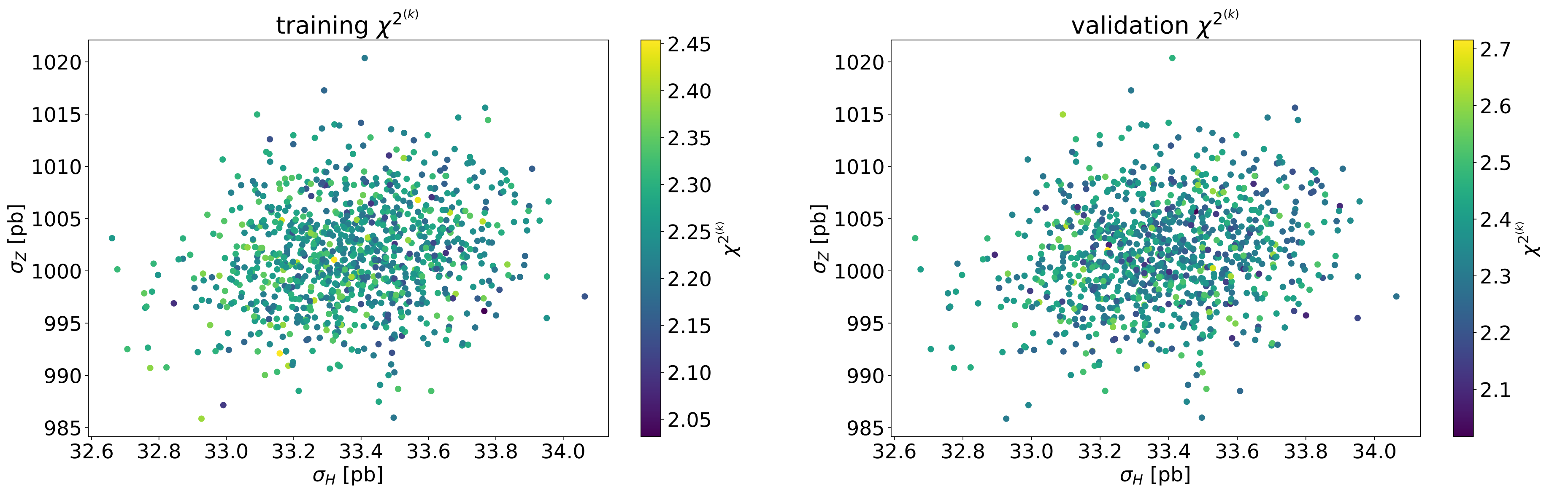


DISTRIBUTION OF REPLICAS DRIVEN BY

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

UNDERSTANDING UNCERTAINTIES
THE REPLICA DISTRIBUTION

ARE ALL FITS EQUALLY GOOD?

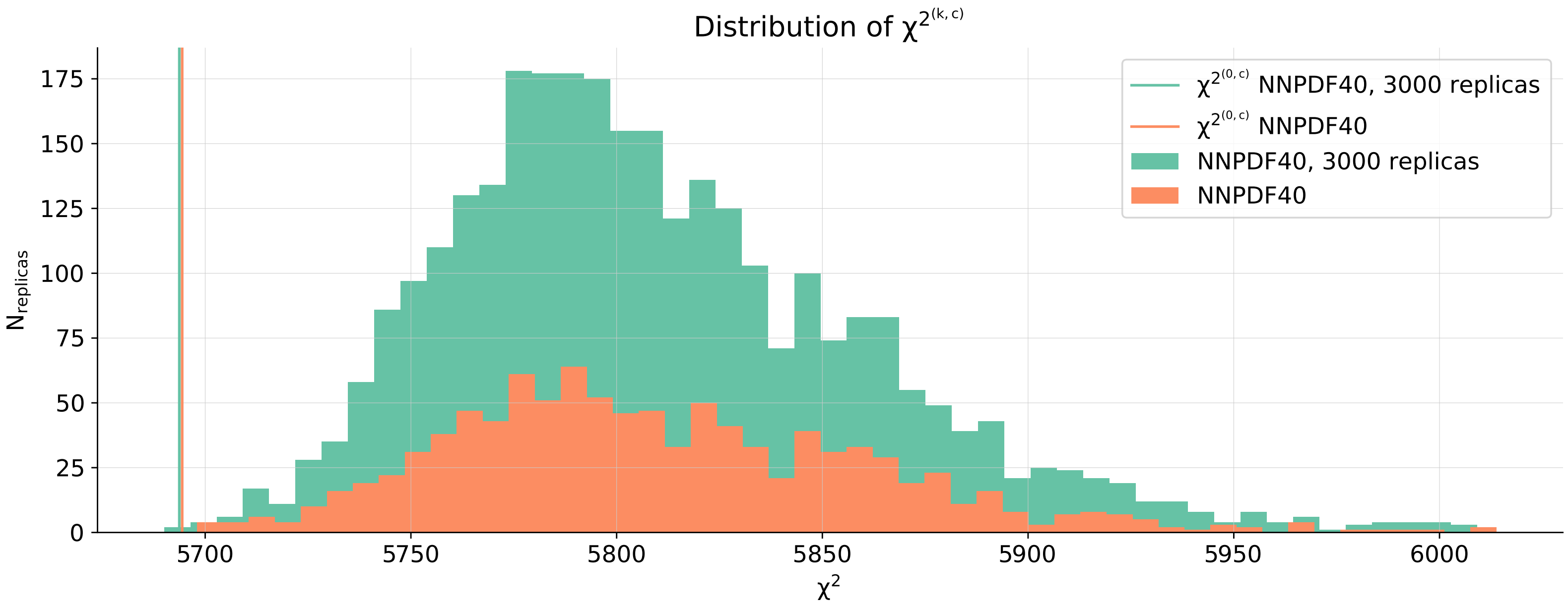


- COMPARE TRAINING AND VALIDATION LOSS FOR EACH REPLICA
- NO CORRELATION BETWEEN FIT QUALITY AND POSITION IN THE (σ_H, σ_Z) PLANE
- UNIFORM FIT QUALITY

UNDERSTANDING UNCERTAINTIES
THE REPLICA DISTRIBUTION
COMPARISON TO CENTRAL DATA

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

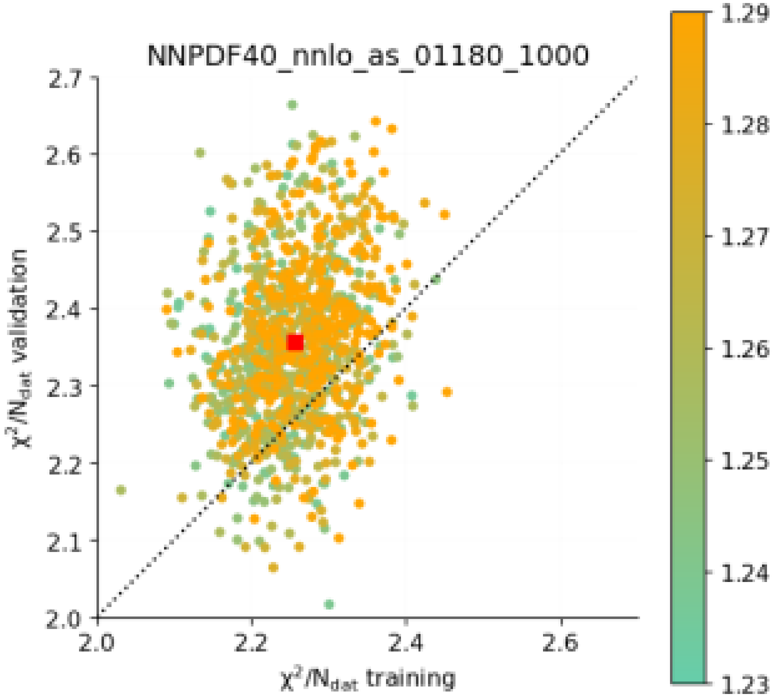
1000 REPLICAS VS. 3000 REPLICAS



- AVERAGE BEST FIT PDF $\Rightarrow$ BETTER AGREEMENT
- NOT NECESSARILY BEST

UNDERSTANDING UNCERTAINTIES
COMPARISON TO CENTRAL DATA

- ARE FITS WITH WORSE AGREEMENT WITH CENTRAL DATA POOR (UNDERLEARNT)?



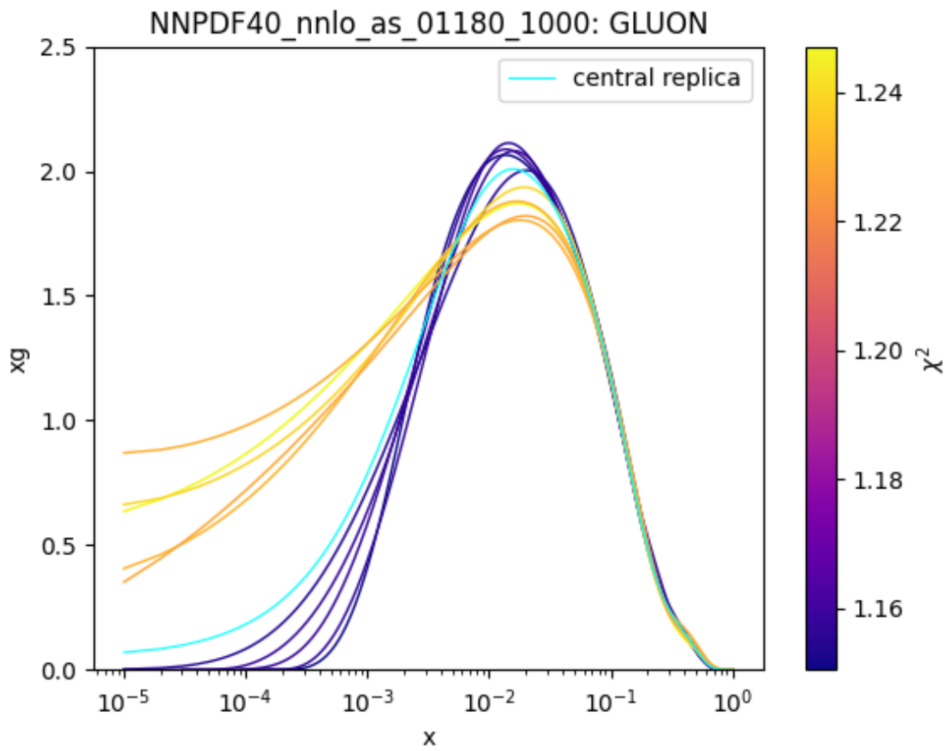
- NO CORRELATION BETWEEN AGREEMENT WITH CENTRAL DATA AND TRAINING, VALIDATION LOSS
- UNIFORM FIT QUALITY
- DISPERSION DUE
  - DATA REPLICA FLUCTUATION $\Rightarrow$ DATA UNCERTAINTIES
  - BEST FIT DEGENERACY
    $\Rightarrow$ INTERPOLATION, EXTRAPOLATION AND FUNCTIONAL UNCERTAINTIES

UNDERSTANDING UNCERTAINTIES

EXPLAINING THE DISTRIBUTION

THE GLUON

REPLICAS WITH BEST & WORST AGREEMENT WITH CENTRAL DATA



- CENTRAL INTERMEDIATE STRUCTURE $\Rightarrow$ OUTLIERS WITH MORE/LESS STRUCTURE
- MORE STRUCTURE $\Rightarrow$ BETTER AGREEMENT WITH (CENTRAL) DATA
- WHY IS MORE STRUCTURE OUTLIER DESPITE BETTER AGREEMENT?

UNDERSTANDING UNCERTAINTIES

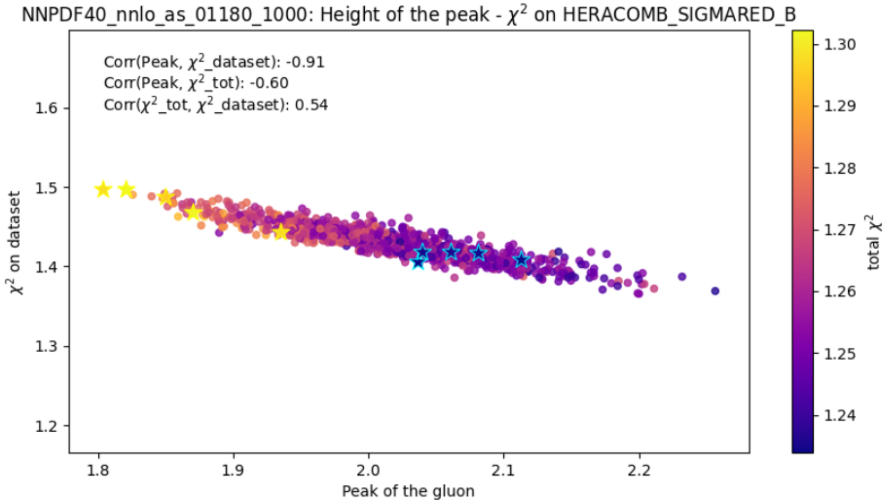
EXPLAINING THE DISTRIBUTION

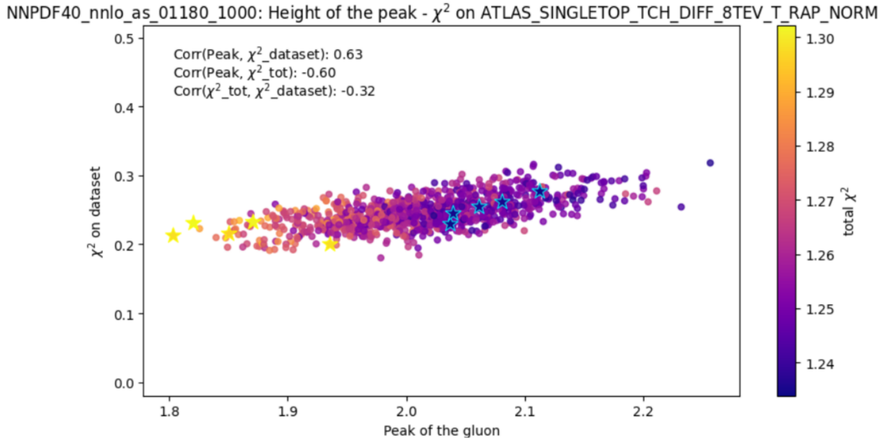
AGREEMENT WITH DATA SUBSET VS HEIGHT OF THE GLUON PEAK

WORST VS BEST AGREEMENT WITH TOTAL DATASET

DATA FAVORING HIGH PEAK (MORE STRUCTURE)

DATA FAVORING LOW PEAK (LESS STRUCTURE)





- MORE OR LESS STRUCTURE (HIGH/LOW PEAK) FAVORED BY
- MORE OR LESS STRUCTURE (HIGH/LOW PEAK) FAVORED BY DIFFERENT DATA SUBSETS
- HIGH PEAK SUBSET MORE NUMEROUS $\Rightarrow$ HIGH PEAK BETTER GLOBAL AGREEMENT
- HIGH PEAK WOULD NOT GENERALIZE $\Rightarrow$ OUTLIER
- MACHINE LEARNING $\Rightarrow$ OPTIMAL MODEL

NO EFFECT THAT REQUIRES MORE THAN 10% ACCURACY IN MEASUREMENT IS WORTH INVESTIGATING

Walther Nernst

~~NO EFFECT THAT REQUIRES MORE THAN 10% ACCURACY IN MEASUREMENT IS WORTH INVESTIGATING~~

Walther Nernst

ACCURACY OF OBSERVATION IS THE EQUIVALENT OF ACCURACY OF THINKING

Wallace Stevens