MACHINE LEARNING
PRECISION
HIGH ENERGY PHYSICS

STEFANO FORTE
UNIVERSITÀ DI MILANO & INFN

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI FISICA

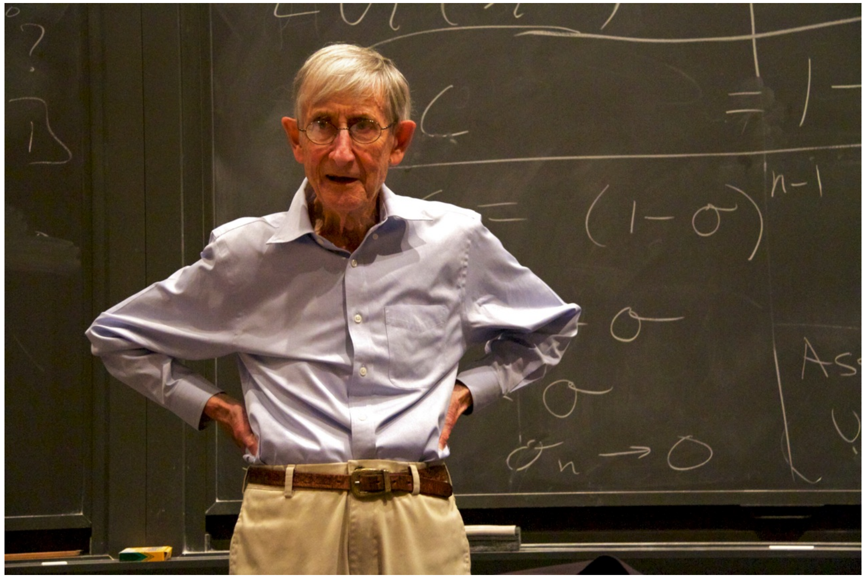
Istituto Nazionale di Fisica Nucleare

MPP COLLOQUIUM

MÜNCHEN, SEPTEMBER 11, 2018

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

ENERGY FRONTIER VS. ACCURACY FRONTIER



"There are historical reasons not to expect too much from the LHC. (...)

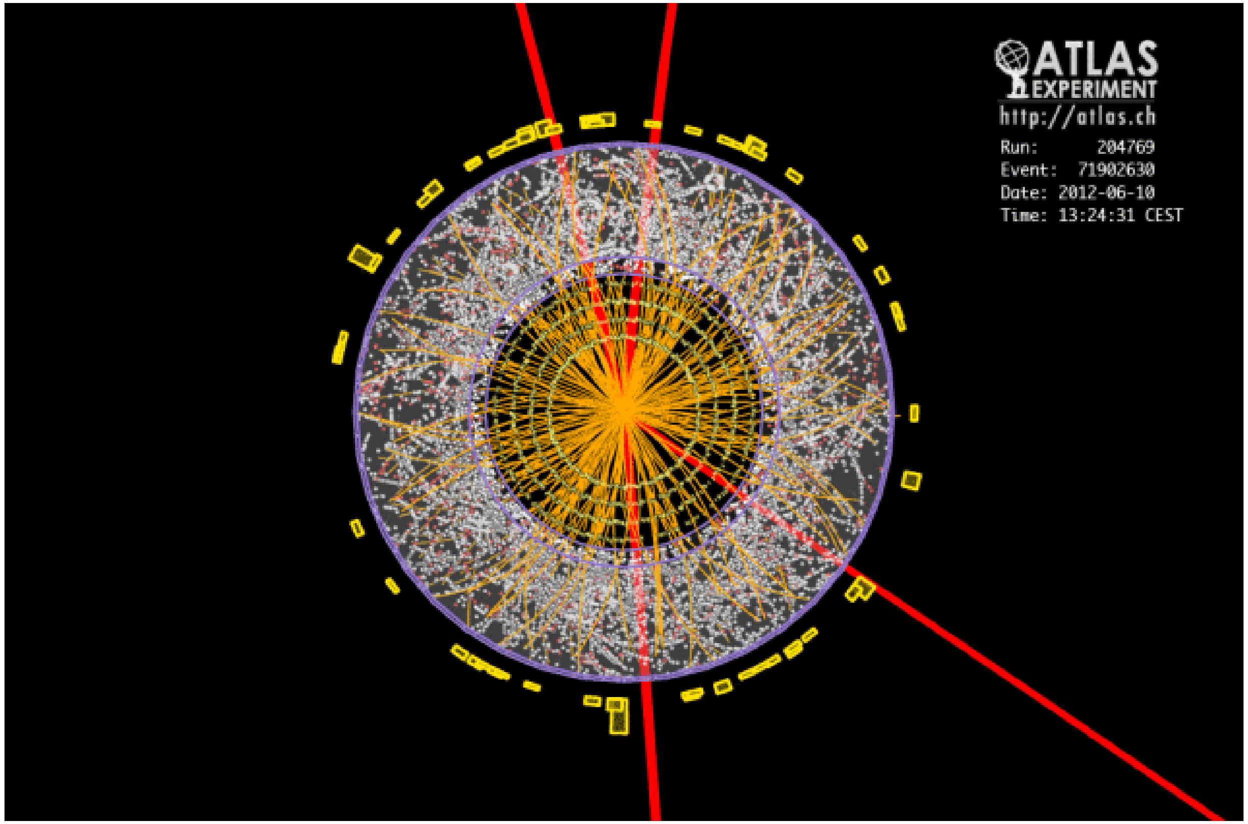
There have been sixteen important discoveries" (in HEP)
"between 1945 and 2008:
four discoveries on the energy frontier,
four on the rarity frontier,
eight on the accuracy frontier"

Freeman Dyson, 2008

PHYSICS AT THE LHC

"There are two reasons to be skeptical about the importance of the LHC: one technical and one historical".

HIGGS BOSON PRODUCED IN pp COLLISION, DECAY INTO TWO $\mu^+\mu^-$ PAIRS

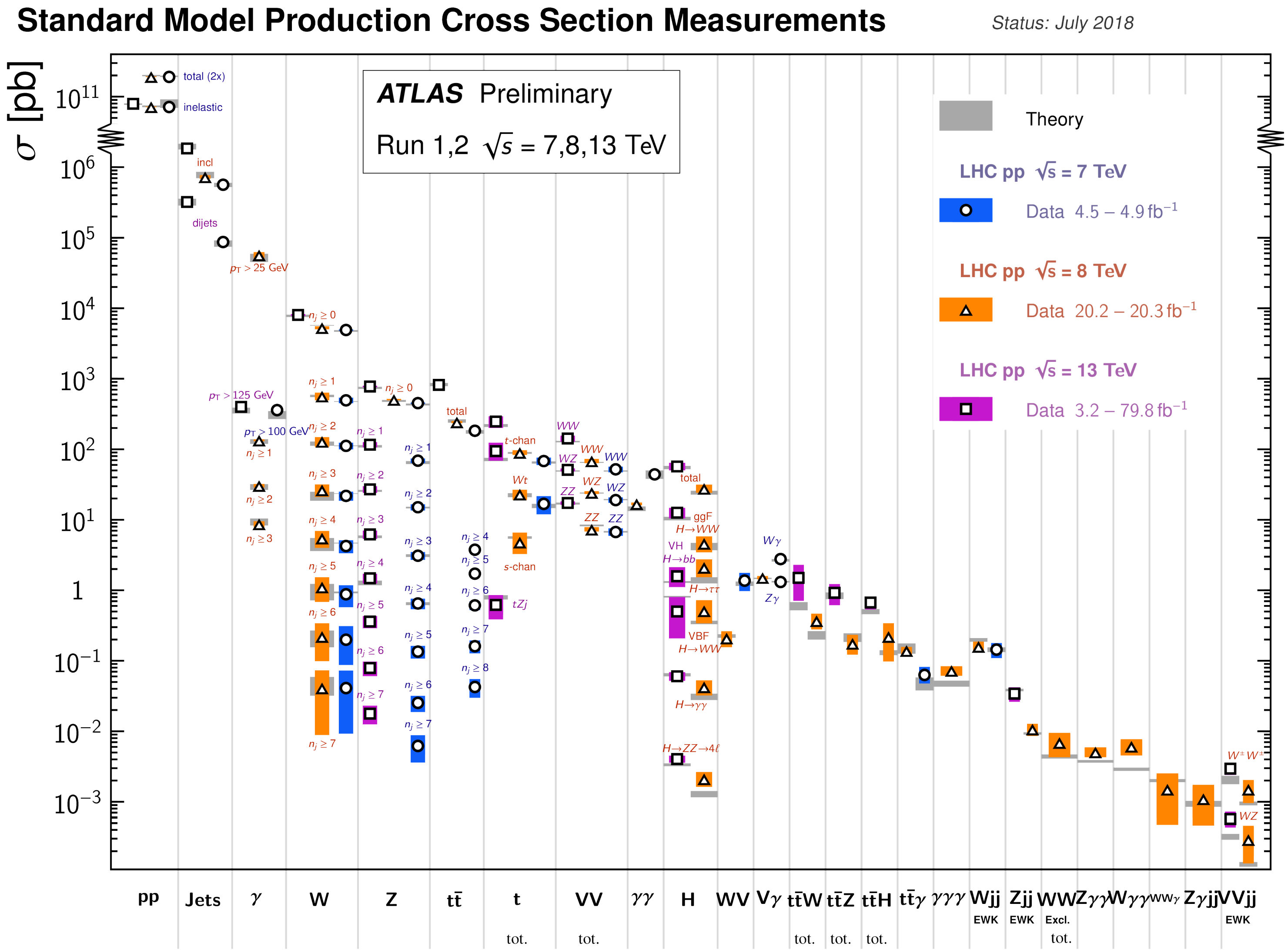


"The technical weakness of the LHC arises from the nature of the collisions that it studies. These are collisions of protons with protons, and they have the unfortunate habit of being messy"

Freeman Dyson, 2008

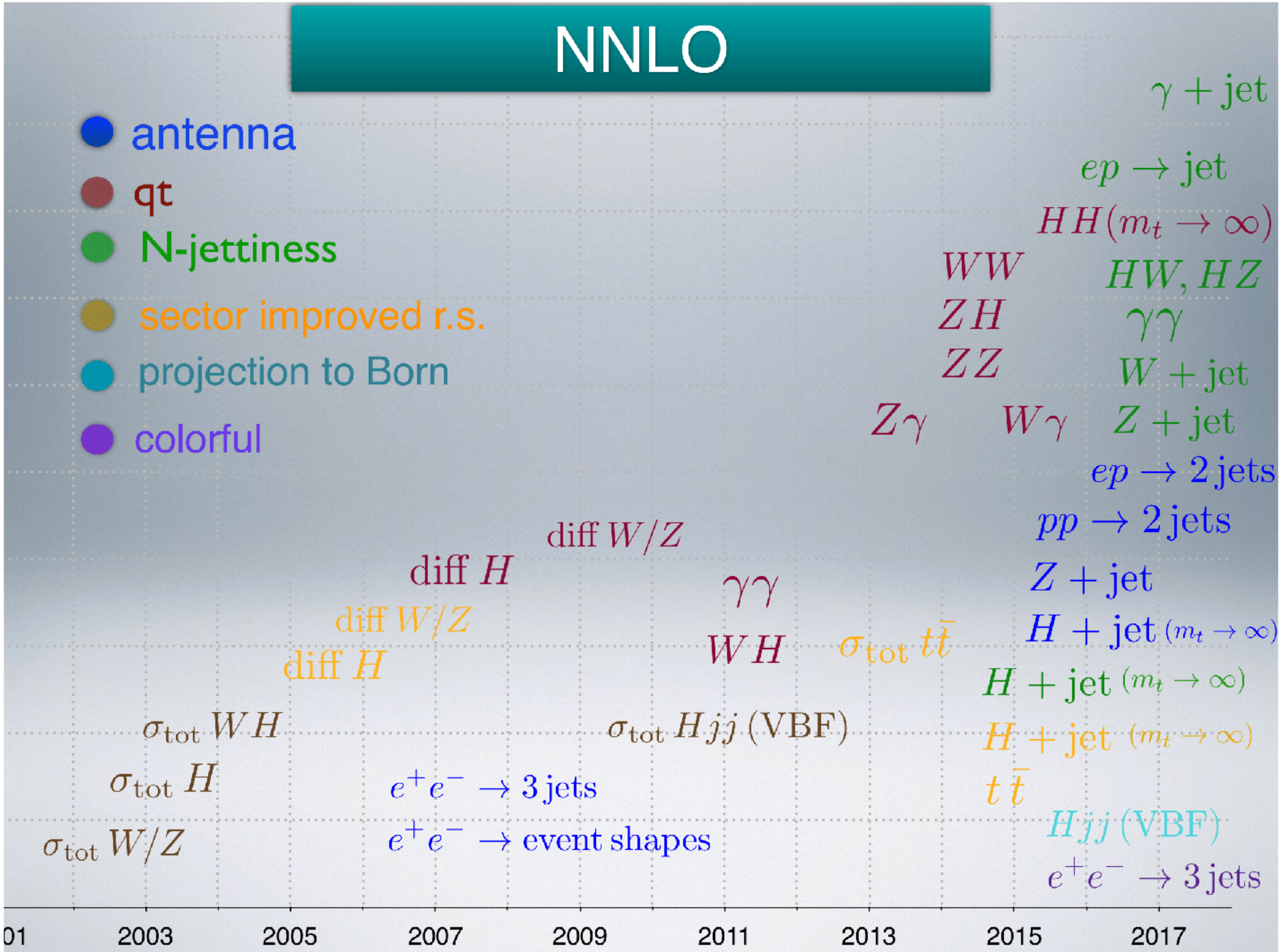
PRECISION PHYSICS AT THE LHC: EXPERIMENT

MEASUREMENT OF STANDARD MODEL PROCESSES



PRECISION PHYSICS AT THE LHC: THEORY

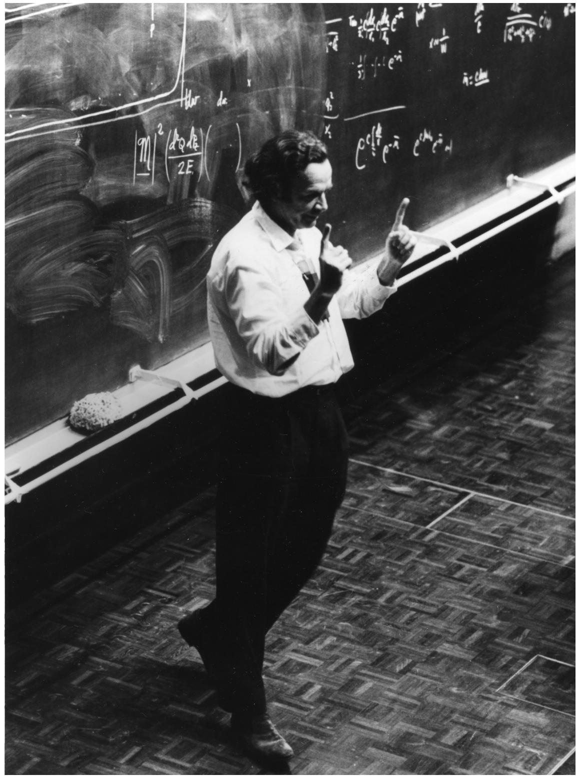
NNLO QCD CALCULATIONS OF COLLIDER PROCESSES



(G. Heinrich, 2017)

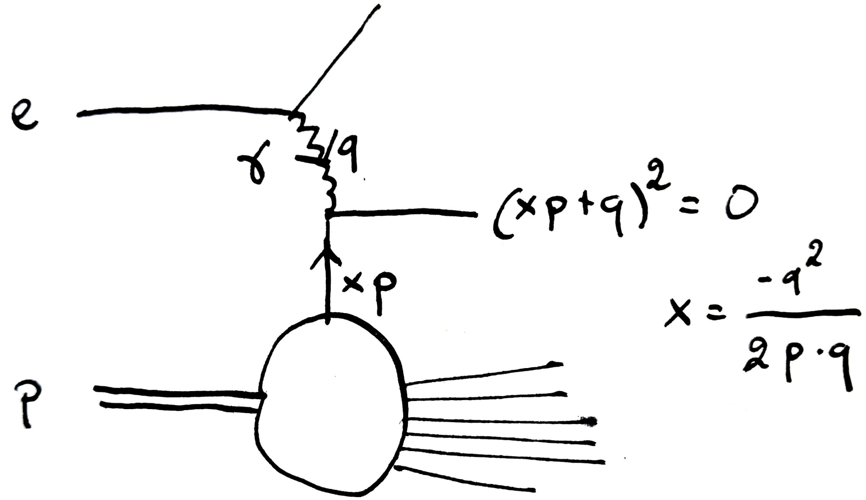
PERTURBATIVE CALCULATIONS?
THE MIRACLE OF FACTORIZATION

THE FEYNMAN PARTON MODEL



R. Feynman explaining the parton model at CERN

PROBE THE PROTON WITH A SHORT-WAVELENGTH PHOTON:

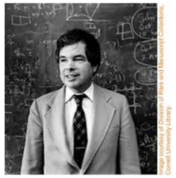


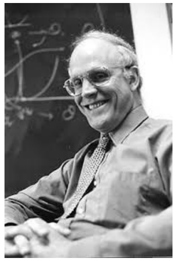
- ASSUME PHOTON STRIKES A FREE MASSLESS “PARTON” (QUARK, GLUON) THAT CARRIES FRACTION x OF ITS PARENT PROTON $p_{\rm quark} = xP_{\rm proton}$
- VALUE OF x FIXED BY FINAL-STATE KINEMATICS
- CROSS-SECTION PROPORTIONAL TO PROBABILITY $q_i(x)$ OF FINDING PARTON OF SPECIES i WITH MOMENTUM-FRACTION x IN TARGET PROTON

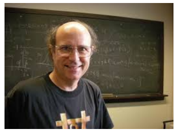
THE MIRACLE OF FACTORIZATION

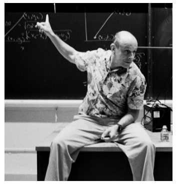
PERTURBATIVE QCD

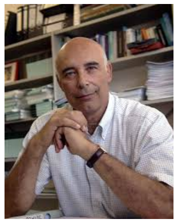
PROBE THE PROTON WITH A SHORT-WAVELENGTH PHOTON:

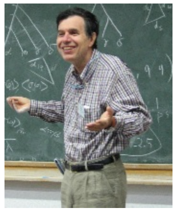

Wilson


Gross

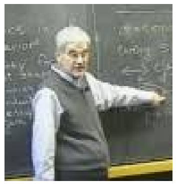

Wilczek

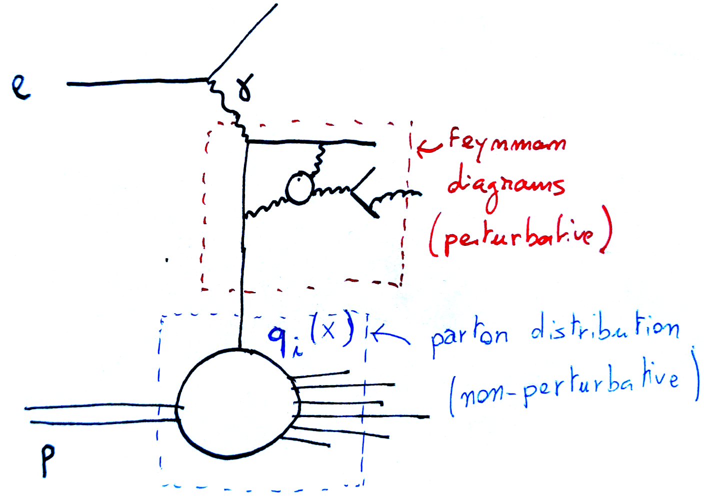

Politzer


Altarelli


Parisi


Collins

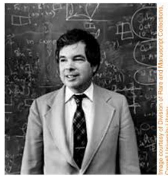

Sterman



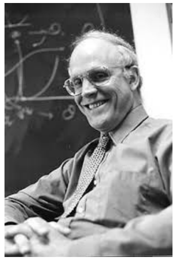
- THE PARTON MODEL IS THE FIRST ORDER OF A PERTURBATIVE EXPANSION
- AT ALL ORDERS, THE CALCULATION FACTORIZES: $\sigma_{\rm experimental} = \hat{\sigma}_{\rm perturbative} \otimes q_i$
- CAN BE PROVEN RIGOROUSLY USING THE OPE (WILSON EXPANSION)

THE MIRACLE OF FACTORIZATION

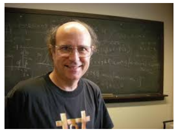
PERTURBATIVE QCD



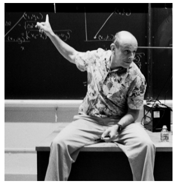
Wilson



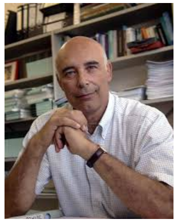
Gross



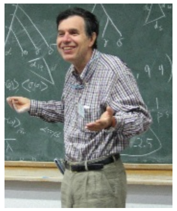
Wilczek



Politzer

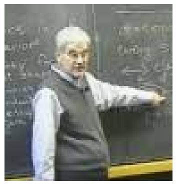


Altarelli



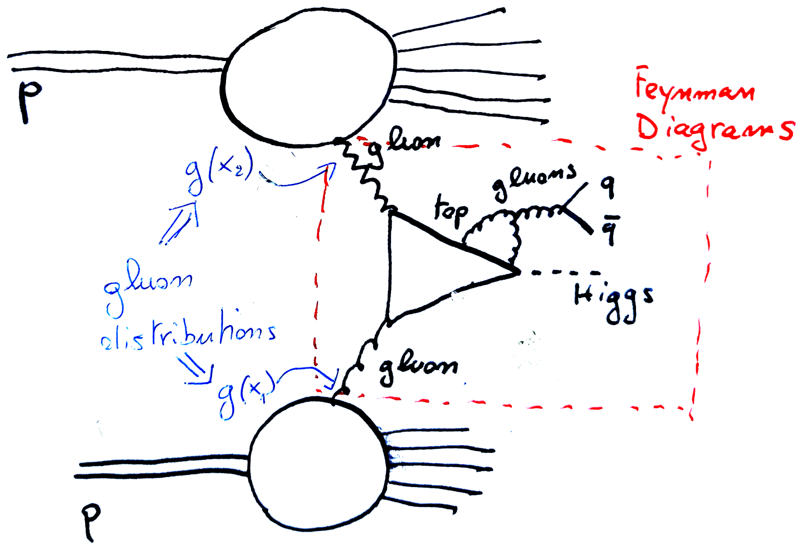
Parisi

Collins



Sterman

PROTON-PROTON COLLISIONS
WITH A SHORT DISTANCE SCALE λ:

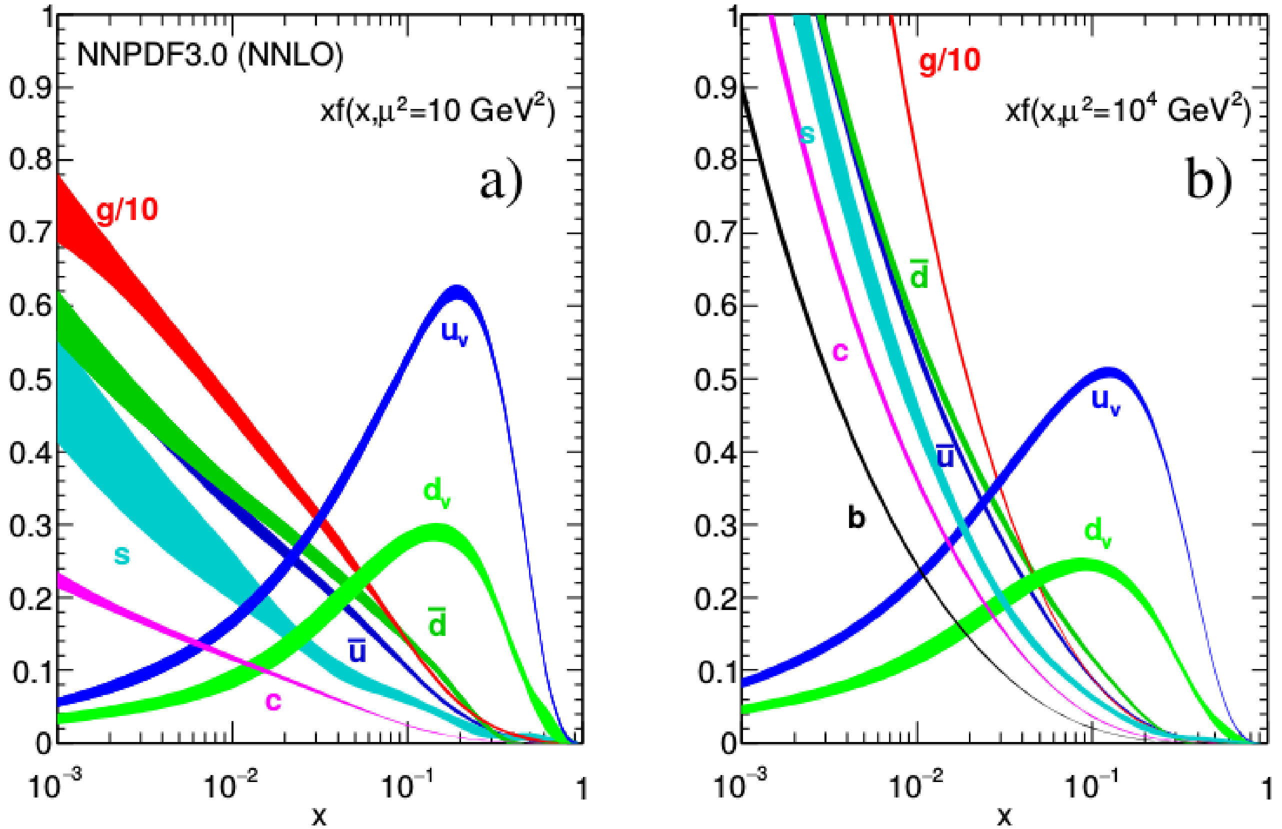


$$\lambda \sim \frac{1}{M_{\text{Higgs}}}$$

Higgs production at LHC at $\mathrm{N}^3\mathrm{LO}$

- THE PARTON MODEL IS THE FIRST ORDER OF A PERTURBATIVE EXPANSION
- AT ALL ORDERS, THE CALCULATION FACTORIZES:
  $\sigma_{\text{experimental}} = \hat{\sigma}_{\text{perturbative}} \otimes q_i \otimes q_j$
- CAN BE SHOWN ORDER BY ORDER USING DIAGRAMMATIC ARGUMENTS

A PORTRAIT OF THE PROTON
AS SEEN FROM A HIGGS BOSON



(PDG 2018)

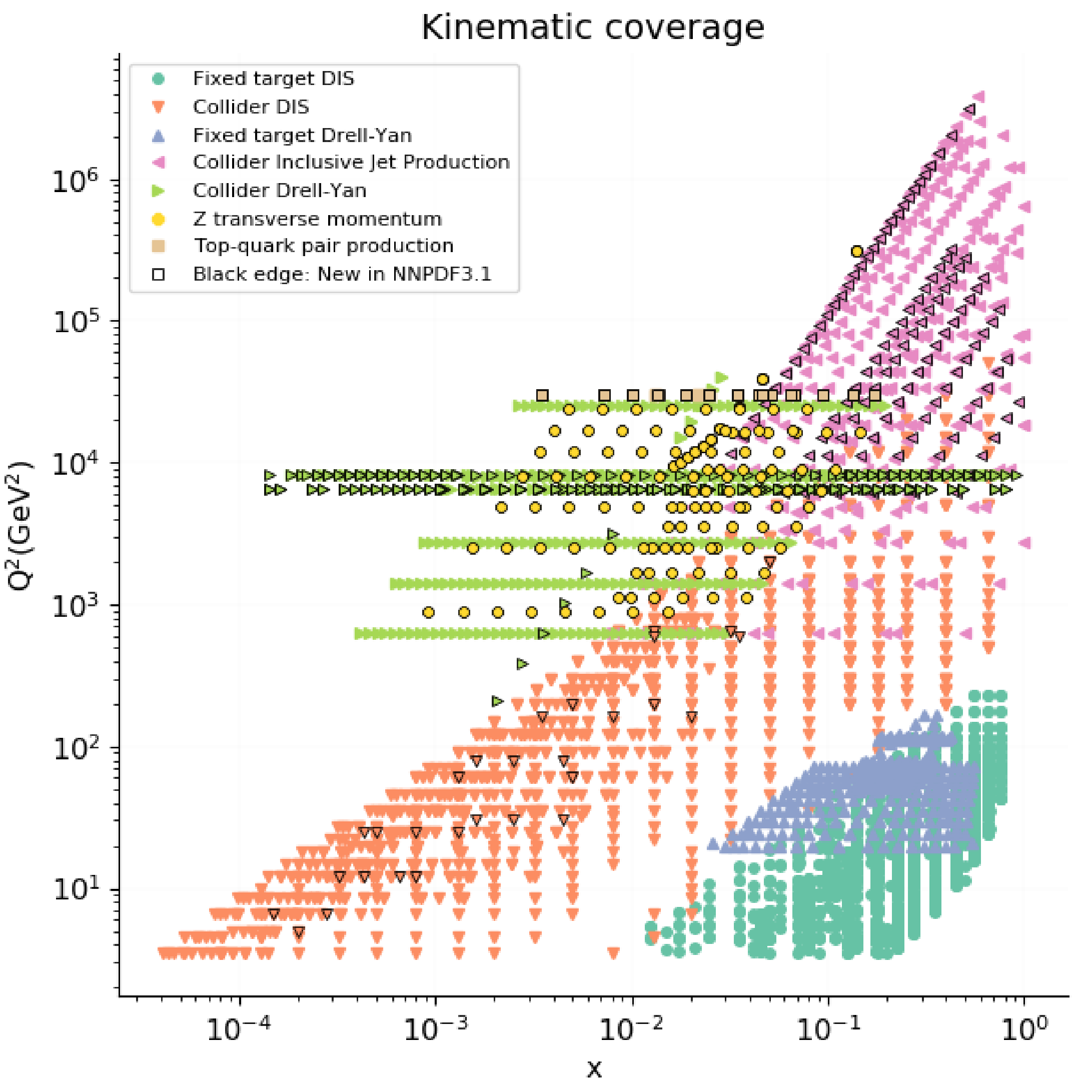
THE "PARTON DISTRIBUTIONS"

- MOMENTUM FRACTION DISTRIBUTIONS
  FOR EACH TYPE OF QUARK, ANTIQUARK & THE GLUON
- $\int_0^1 f_{\rm up}(x)\,dx = N_{\rm up},\ \ldots$
- "VALENCE" $N_{\rm up} - N_{\rm antiup} = 2;\ N_{\rm down} - N_{\rm down} = 2$
- RESULT DEPENDS ON SCALE μ (RESOLUTION); dep. on scale perturbatively computable

HOW IS THE PORTRAIT TAKEN?

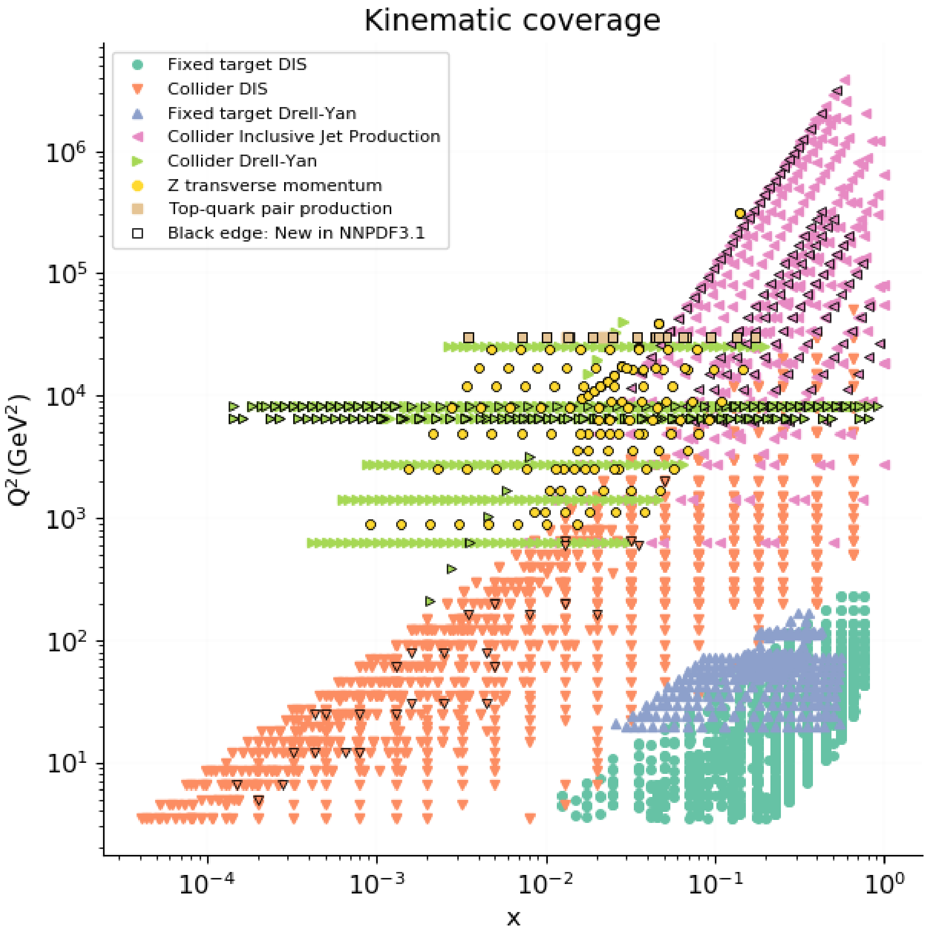
- COMPUTE PHYSICAL PROCESS $\Rightarrow$ DEPENDS ON PDF
- COMPARE TO DATA
- INVERT

A TYPICAL DATASET (NNPDF3.1)



HOW IS THE PORTRAIT TAKEN?

A TYPICAL DATASET (NNPDF3.1)



ISSUES

- EACH PDF IS A FUNCTION actually, a distribution
  $\Rightarrow$ WE ARE DETERMINING A
  PROBABILITY DISTRIBUTION IN A SPACE OF FUNCTIONS

- WE ARE DETERMINING A
  FUNCTIONAL (CONTINUOUS) FROM A DISCRETE SET OF DATA

FLASHBACK: PDFs CIRCA 2004

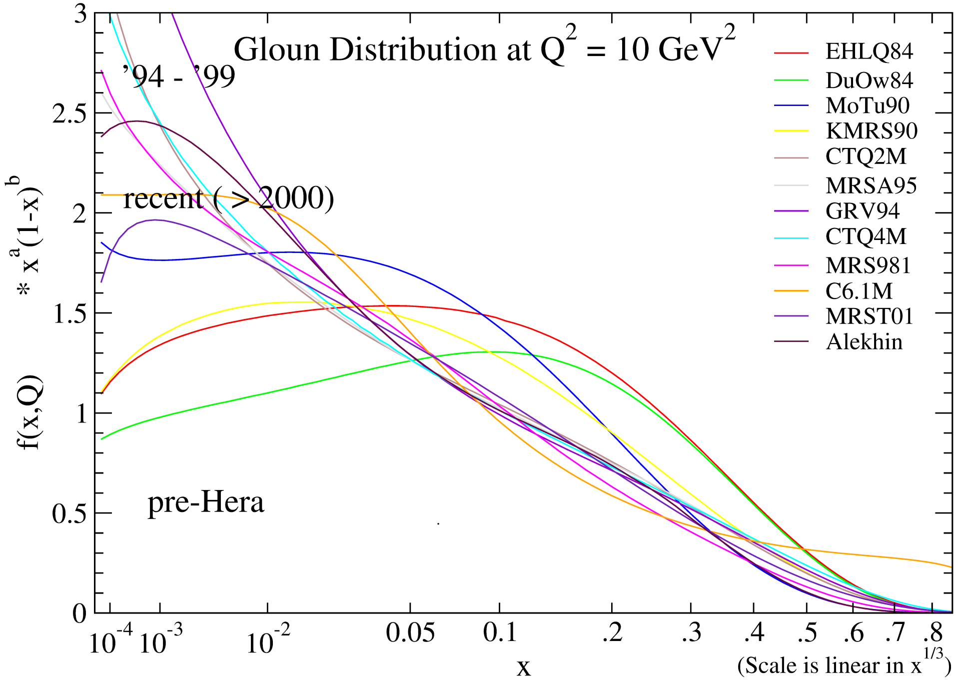
SIMPLE SOLUTION: PICK A FUNCTIONAL FORM

gluon parametrization (MRST 2004)

$$xg(x,Q_0^2) = A_g(1-x)^{\eta_g}(1+\epsilon_g x^{0.5}+\gamma_g x)x^{\delta_g} - A_-(1-x)^{\eta_-}x^{-\delta_-}$$

- PRO: PROBLEM REDUCED TO FINITE-DIMENSIONAL
- CON: WHO PICKS THE FUNCTIONAL FORM?

HISTORICAL COMPILATION OF GLUON PDFs



W.K.Tung, DIS 2004

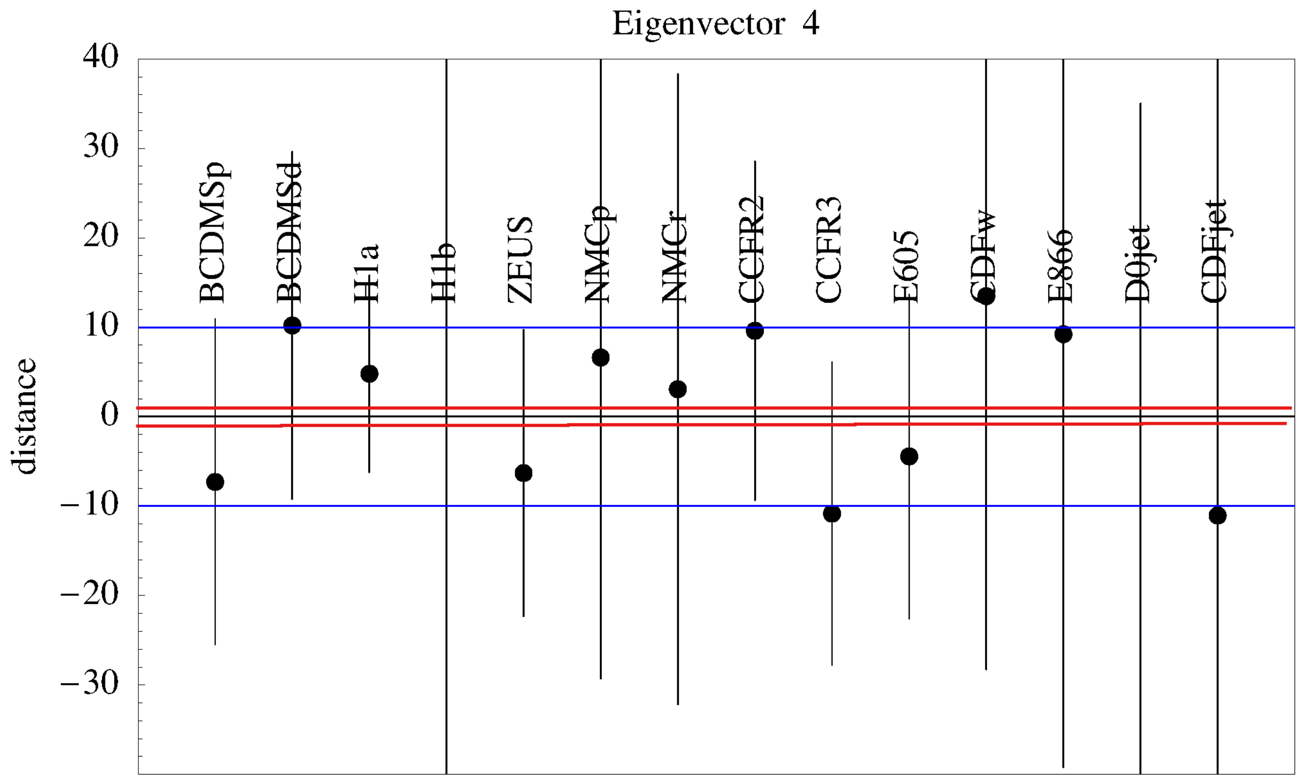
THE HERA-LHC WORKSHOP

...this is when Dyson made his comments!

"TOLERANCE"

2002: FIRST PDFs WITH UNCERTAINTIES

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

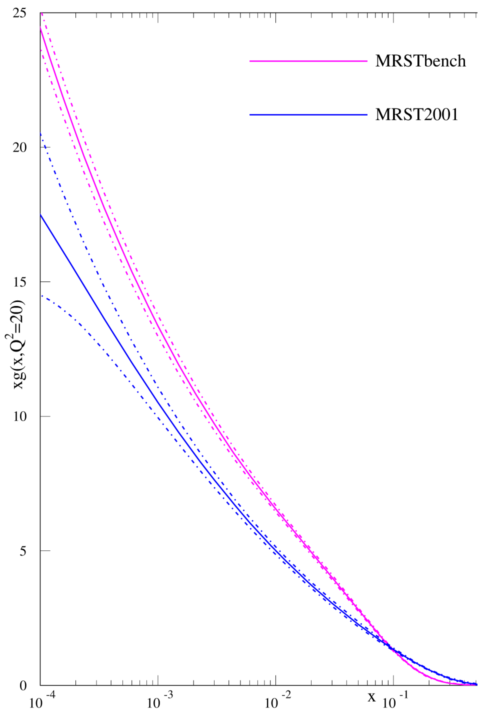


$\Rightarrow$ REALISTICS PDF UNCERTAINTIES NEED SIZABLE "TOLERANCE" RESCALING ($T = 5 - 10$)

THE HERA-LHC BENCHMARK

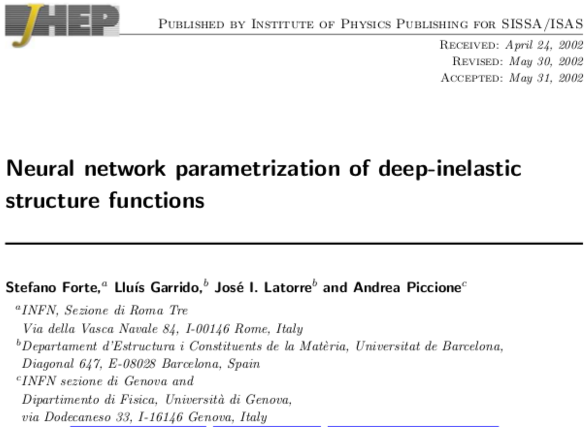
- RESTRICTED AND VERY CONSISTENT DATASET USED
- RESULTS COMPARED TO THEN-BEST RESULT FROM FULL DATASET

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)

A NEW APPROACH?



PUBLISHED BY INSTITUTE OF PHYSICS PUBLISHING FOR SISSA/ISAS

RECEIVED: *April 24, 2002*
REVISED: *May 30, 2002*
ACCEPTED: *May 31, 2002*

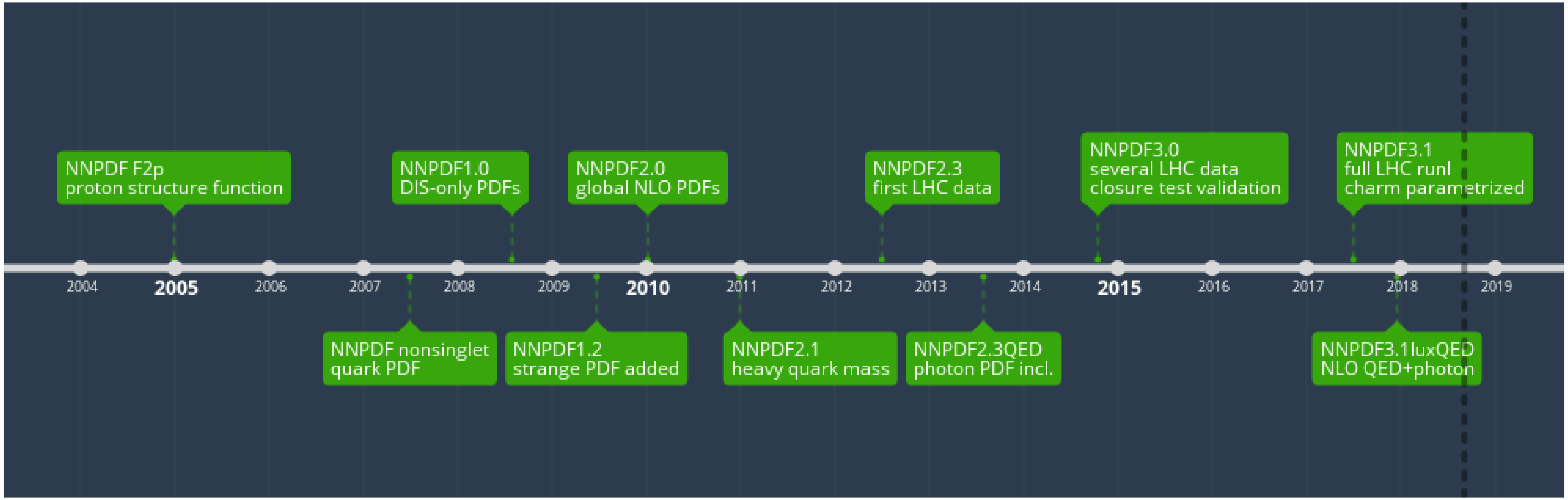
Neural network parametrization of deep-inelastic structure functions

Stefano Forte,[a] Lluís Garrido,[b] José I. Latorre[b] and Andrea Piccione[c]

[a] *INFN, Sezione di Roma Tre*
Via della Vasca Navale 84, I-00146 Rome, Italy
[b] *Departament d'Estructura i Constituents de la Matèria, Universitat de Barcelona,*
Diagonal 647, E-08028 Barcelona, Spain
[c] *INFN sezione di Genova and*
Dipartimento di Fisica, Università di Genova,
via Dodecaneso 33, I-16146 Genova, Italy

FROM THE PROOF OF CONCEPT...

...TO THE NNPDF TIMELINE



COMBINING DATA BY MONTE CARLO

TWO MEASUREMENTS: $\mu_1 \pm \sigma_1; \qquad \mu_2 \pm \sigma_2$

ML COMBINATION: $\bar{\mu} \pm \bar{\sigma}; \quad \bar{\mu} = \frac{\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}; \ \bar{\sigma}^2 = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$

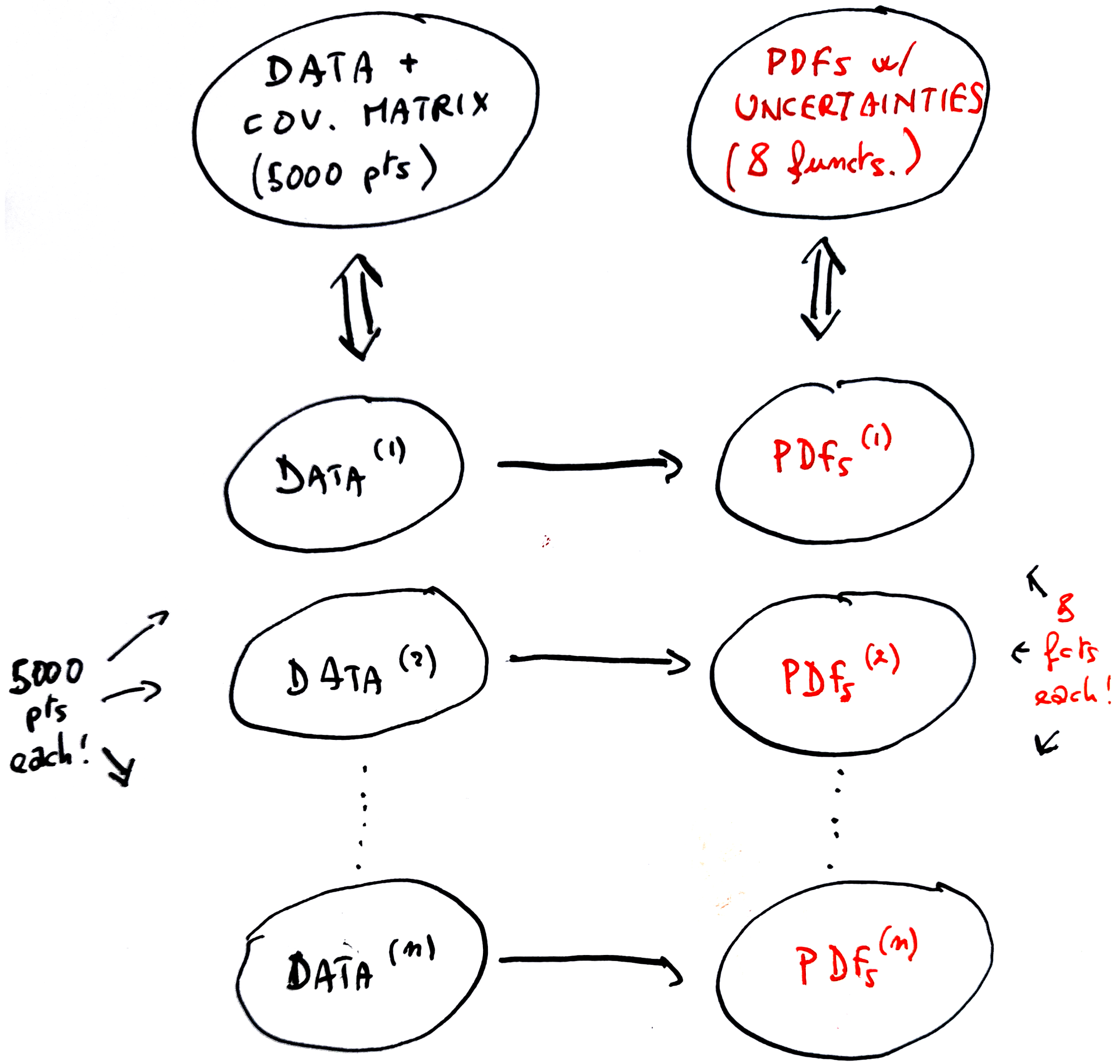
MONTE CARLO REPRESENTATION

$$
\begin{array}{ccccc}
\mu_1 \pm \sigma_1 & & \mu_2 \pm \sigma_2 & \longrightarrow & \bar{\mu} \pm \bar{\sigma} \\
\Updownarrow & & \Updownarrow & & \Updownarrow \\
\mu_1^{(1)} & & \mu_2^{(1)} & \longrightarrow & \bar{\mu}^{(1)} \\
\mu_1^{(2)} & & \mu_2^{(2)} & \longrightarrow & \bar{\mu}^{(2)} \\
\vdots & & \vdots & \vdots & \\
\mu_1^{(n)} & & \mu_2^{(n)} & \longrightarrow & \bar{\mu}^{(n)}
\end{array}
$$

$\mu^{(i)} \Leftrightarrow$ REPLICA SAMPLE $\Leftrightarrow$ REPRESENTATION OF PROBABILITY DISTRIBUTION

THE FUNCTIONAL MONTE CARLO

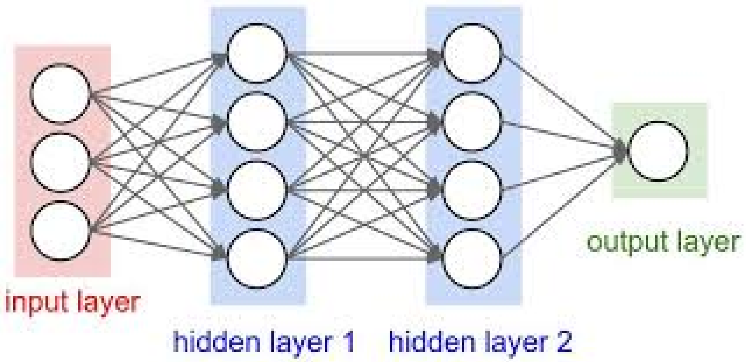
REPLICA SAMPLE OF FUNCTIONS $\Leftrightarrow$ PROBABILITY DENSITY IN FUNCTION SPACE



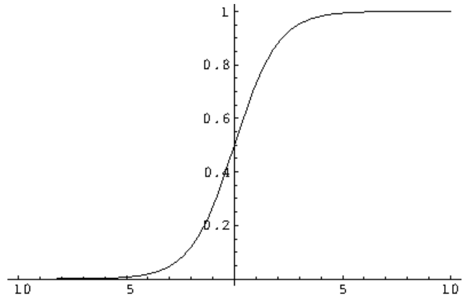
FINAL PDF SET: $f_i^{(a)}(x,\mu)$; i =up, antiup, down, antidown, strange, antistrange, charm, gluon; $j = 1, 2, \ldots N_{\mathrm{rep}}$

UNBIASED INTERPOLANTS: NEURAL NETWORKS

ARCHITECTURE



ACTIVATION FUNCTION



PARAMETERS

- WEIGHTS ω_{ij}
- THRESHOLDS θ_i

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

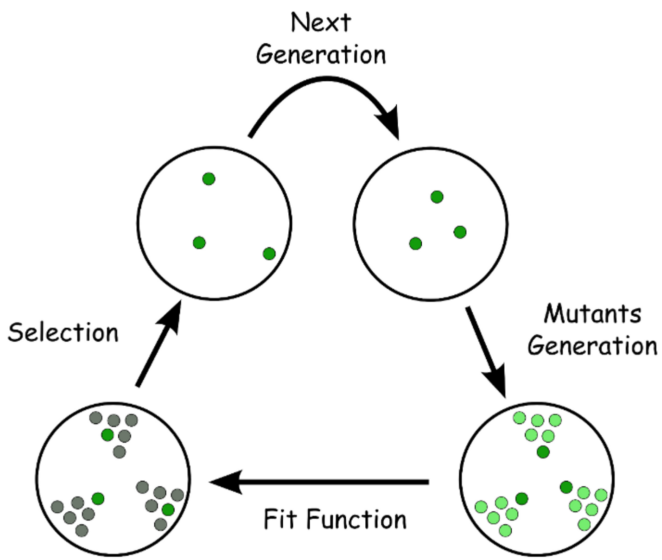
SIMPLEST EXAMPLE
1-2-1

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

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

GENETIC MINIMIZATION

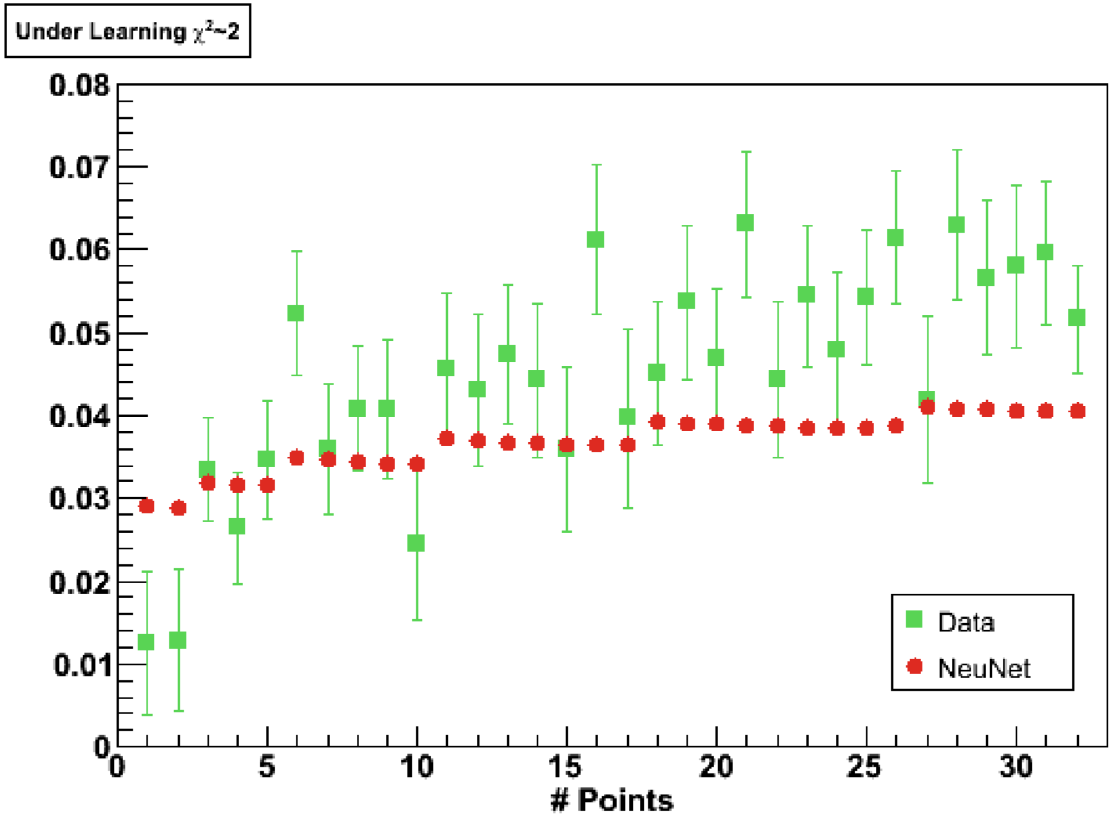
BASIC IDEA: RANDOM MUTATION OF THE NN PARAMETER, SELECTION OF THE FITTEST



NEURAL LEARNING

- CHOOSE HIGHLY REDUNDANT PARAMETRIZATION
- COMPLEXITY INCREASES AS THE FITTING PROCEEDS
- $\Rightarrow$ THE BEST FIT IS NOT THE ABSOLUTE MINIMUM:
  MUST LOOK FOR OPTIMAL LEARNING POINT

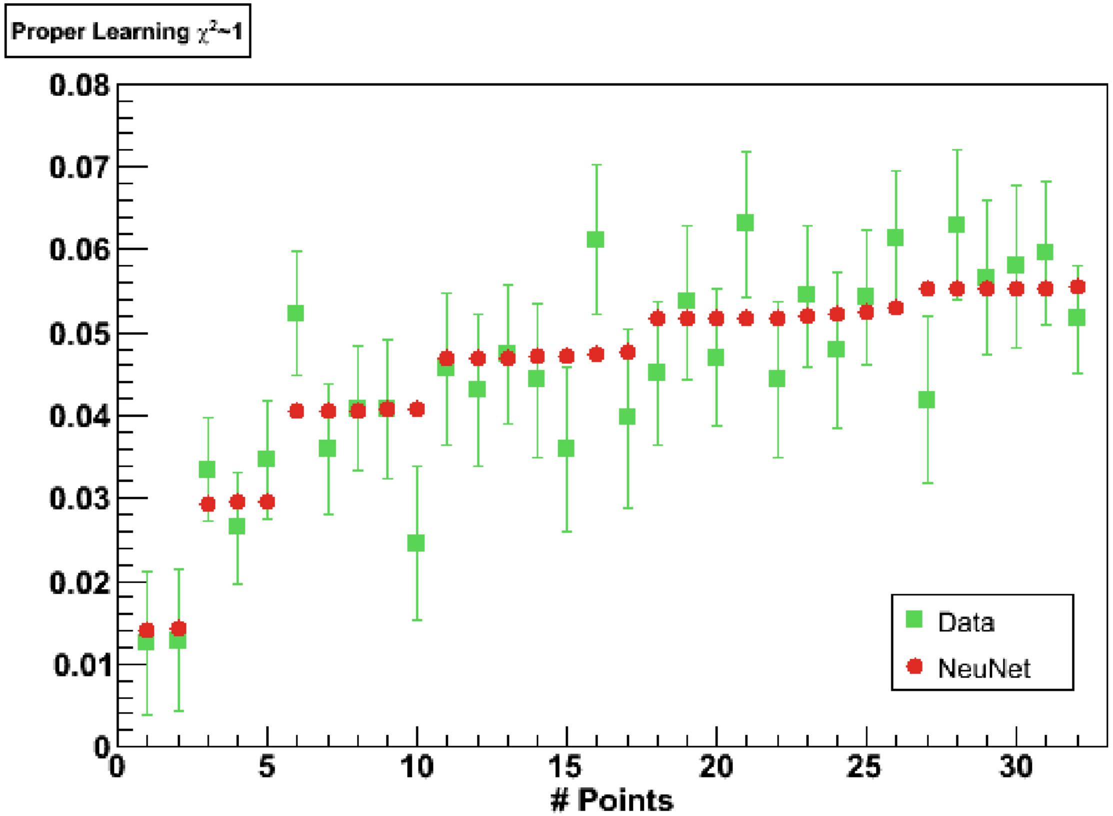
UNDERLEARNING



NEURAL LEARNING

- CHOOSE HIGHLY REDUNDANT PARAMETRIZATION
- COMPLEXITY INCREASES AS THE FITTING PROCEEDS
- $\Rightarrow$ THE BEST FIT IS NOT THE ABSOLUTE MINIMUM:
  MUST LOOK FOR OPTIMAL LEARNING POINT

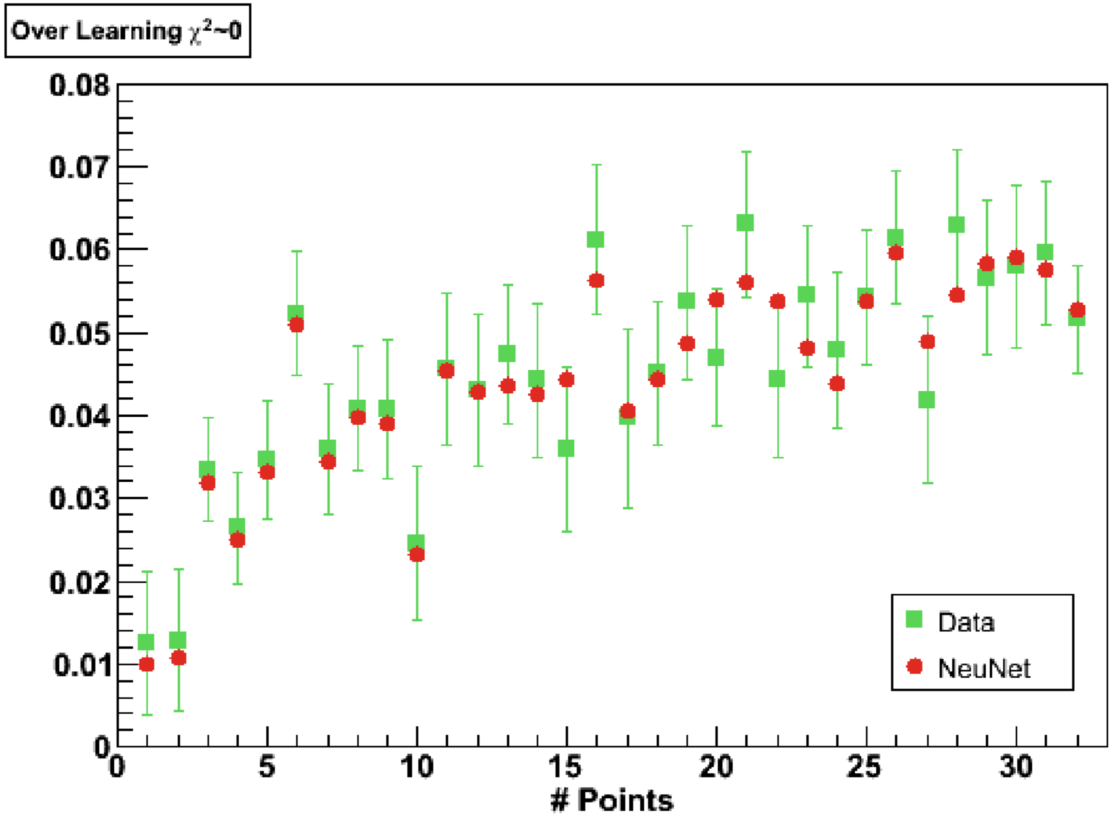
PROPER LEARNING



NEURAL LEARNING

- CHOOSE HIGHLY REDUNDANT PARAMETRIZATION
- COMPLEXITY INCREASES AS THE FITTING PROCEEDS
- $\Rightarrow$ THE BEST FIT IS NOT THE ABSOLUTE MINIMUM:
  MUST LOOK FOR OPTIMAL LEARNING POINT

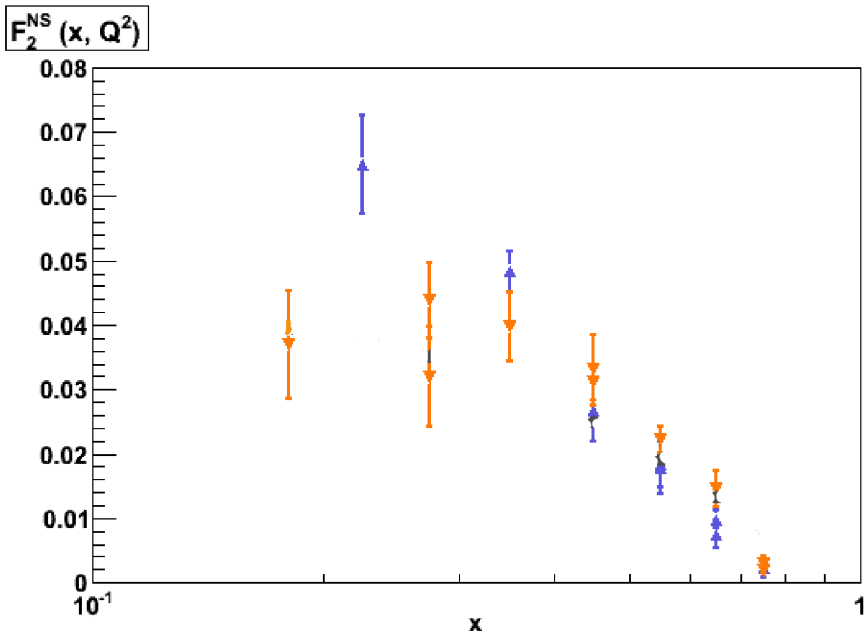
OVERLEARNING



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

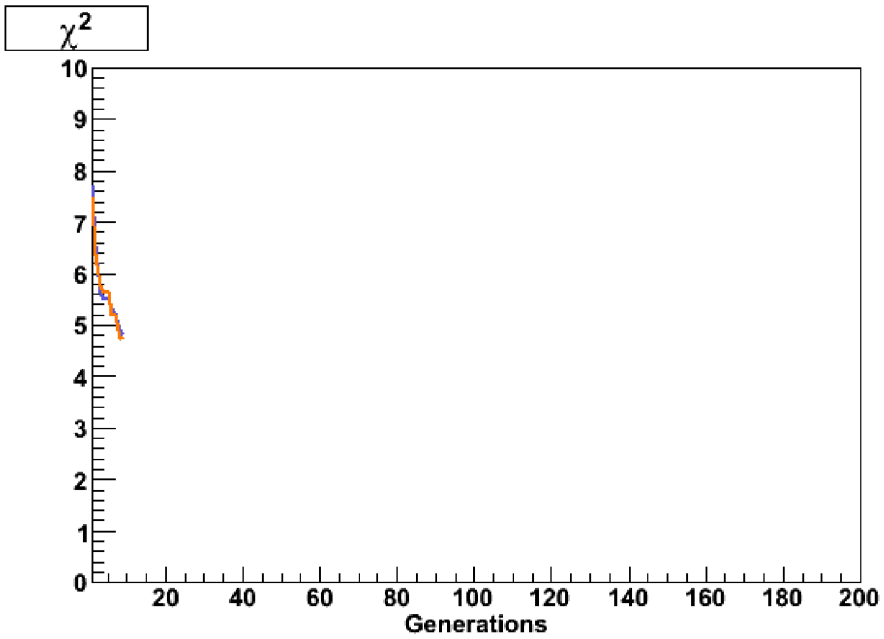


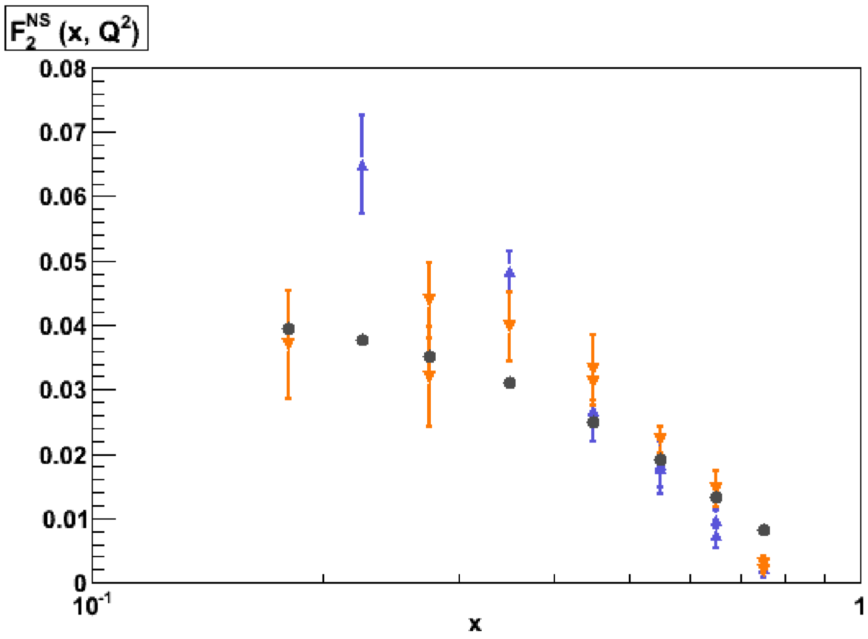
OPTIMAL FIT: CROSS-VALIDATION

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!



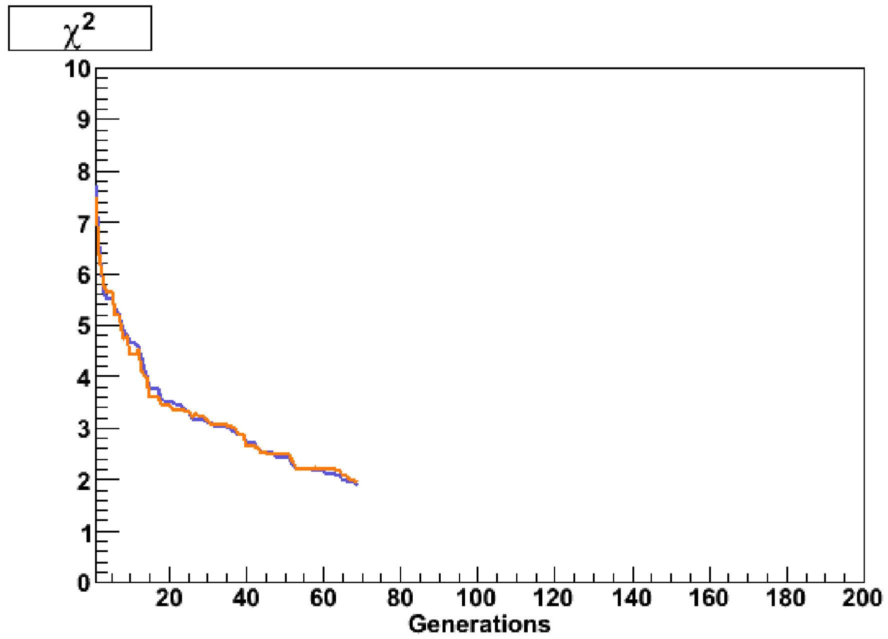


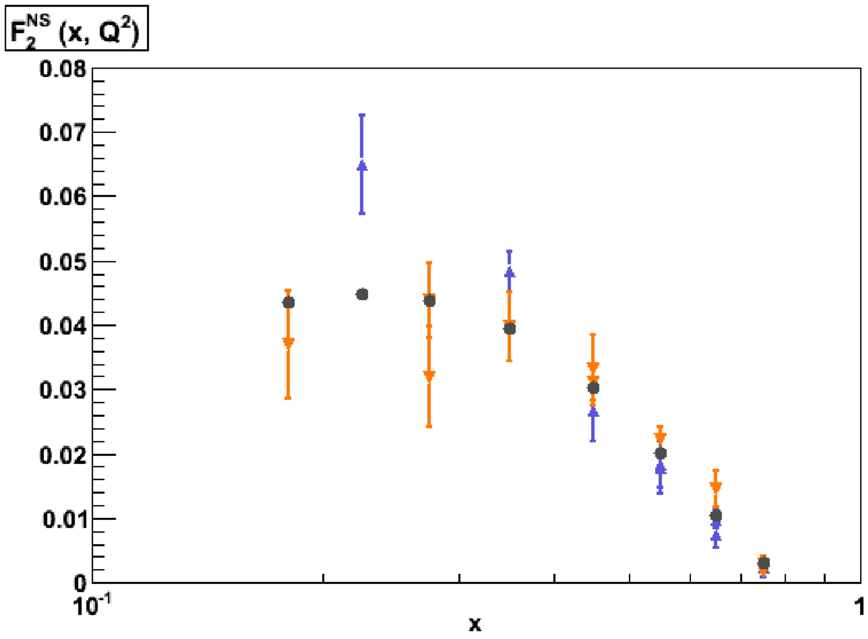
OPTIMAL FIT: CROSS-VALIDATION

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!



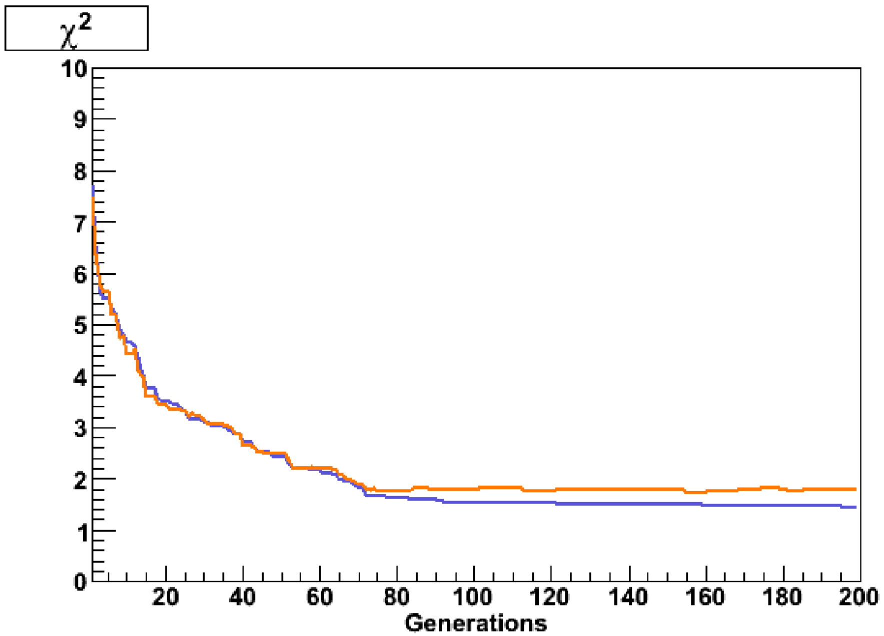


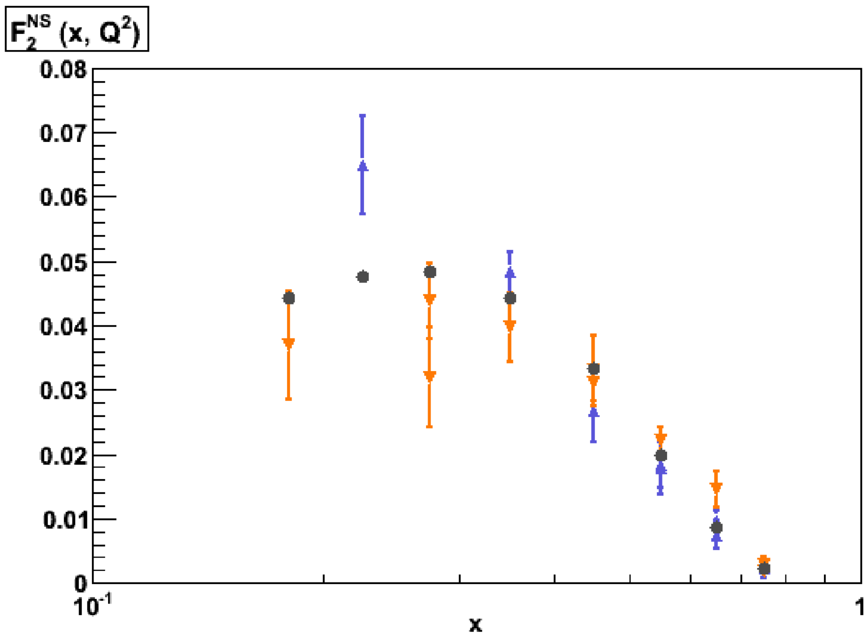
OPTIMAL FIT: CROSS-VALIDATION

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

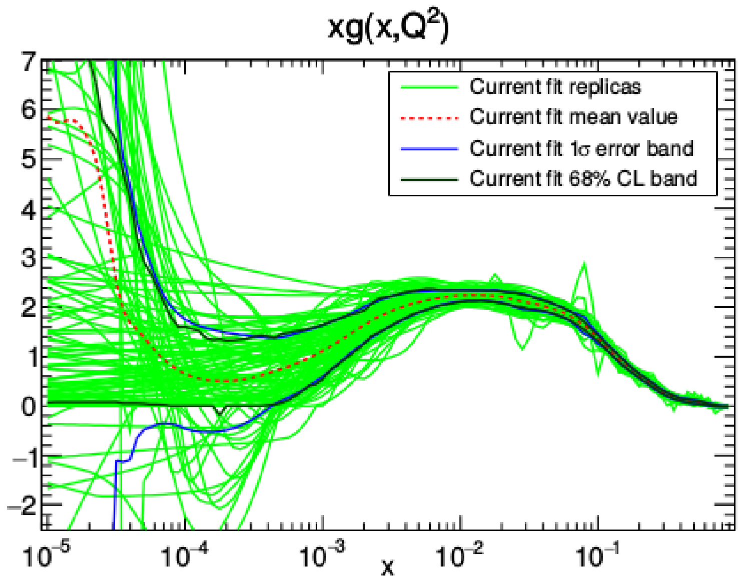
TOO LATE!



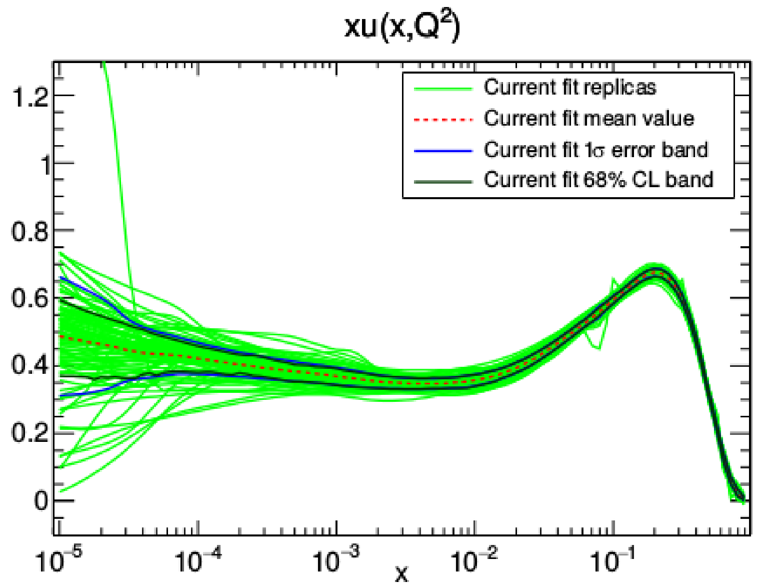


THE RESULT
NNPDF3.1 REPLICAS

GLUON



UP



- PROBABILITY DISTRIBUTION OF PDFs ↔ ENSEMBLE OF REPLICAS
- EXPECTED CENTRAL VALUE ↔ MEAN
- UNCERTAINTY ↔ STANDARD DEVIATION
- 68% C.L. ALSO SHOWN

CLOSURE TESTS

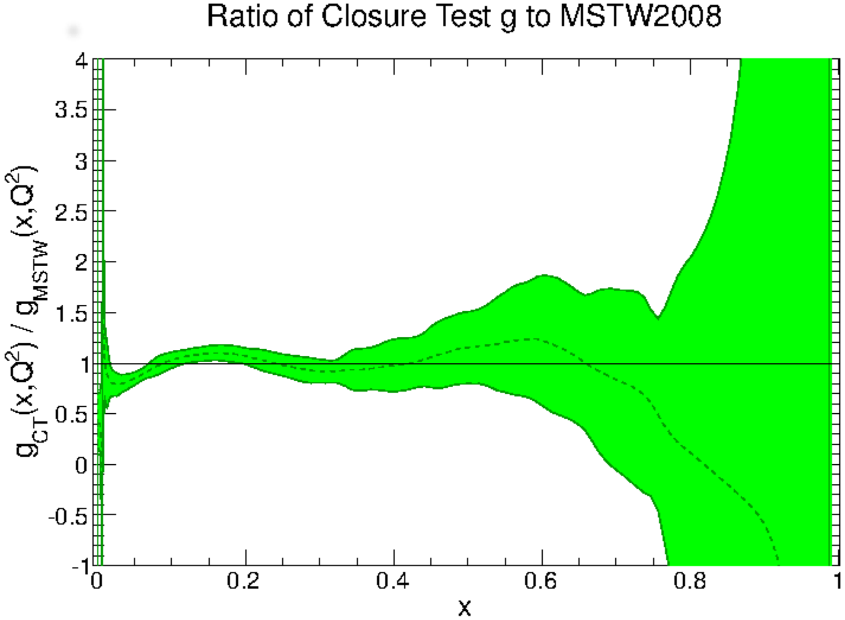
THE IDEA

- ASSUME PDFs KNOWN: GENERATE FAKE EXPERIMENTAL DATA
- CAN DECIDE DATA UNCERTAINTY (ZERO, OR AS IN REAL DATA, OR . . .)
- FIT PDFs TO FAKE DATA
- CHECK WHETHER FIT REPRODUCES UNDERLYING "TRUTH":
  - CHECK WHETHER TRUE VALUE GAUSSIANLY DISTRIBUTED ABOUT FIT
  - CHECK WHETHER UNCERTAINTIES FAITHFUL
  - TRACE DIFFERENT SOURCES OF UNCERTAINTY

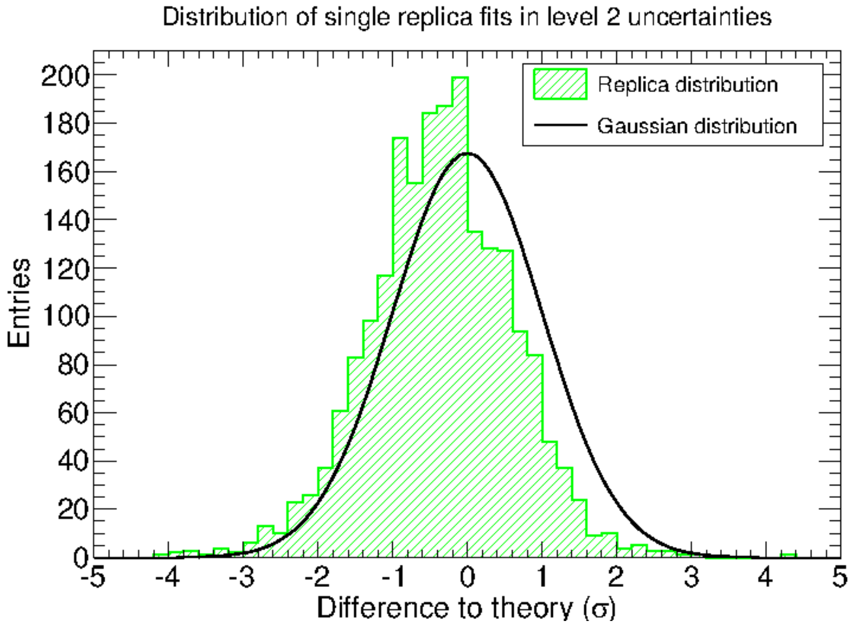
CLOSURE TESTS
RESULTS

- CENTRAL VALUES: FITTED VS. "TRUE" χ^2: $\Delta\chi^2 = 0.001 \pm 0.003$
- UNCERTAINTIES: DISTRIBUTION OF DEVIATIONS BETWEEN FITTED AND "TRUE": 69.9% FOR ONE-SIGMA, 94.8% FOR TWO-SIGMA C.L.

THE GLUON: FITTED/"TRUE"



NORM. DISTRIBUTION OF DEVIATIONS



SUCCESS

NNPDFs INCREASINGLY STANDARD!

inSPIRE 2017 TOPCITE LIST

20. 491 core citations in 2017

Particle Creation by Black Holes
S.W. Hawking (Cambridge U.). Aug 1975. 22 pp.
Published in **Commun.Math.Phys. 43 (1975) 199-220,** Erratum**: Commun.Math.Phys. 46 (1976) 206**
DOI: 10.1007/BF02345020
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
AMS MathSciNet; Project Euclid

---

21. 490 core citations in 2017

Measurements of Omega and Lambda from 42 high redshift supernovae
Supernova Cosmology Project Collaboration (S. Perlmutter (UC, Berkeley, CfPA) *et al.*). Dec 1998. 33 pp.
Published in **Astrophys.J. 517 (1999) 565-586**
LBNL-41801, LBL-41801
DOI: 10.1086/307221
e-Print: **astro-ph/9812133** | **PDF**
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
ADS Abstract Service; OSTI Information Bridge Server

---

22. 487 core citations in 2017

Parton distributions for the LHC Run II
NNPDF Collaboration (Richard D. Ball (U. Edinburgh, Higgs Ctr. Theor. Phys. & CERN) *et al.*). Oct 31, 2014. 138 pp.
Published in **JHEP 1504 (2015) 040**
EDINBURGH-2014-15, IFUM-1034-FT, CERN-PH-TH-2013-253, OUTP-14-11P, CAVENDISH-HEP-14-11
DOI: 10.1007/JHEP04(2015)040
e-Print: **arXiv:1410.8849 [hep-ph]** | **PDF**
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
CERN Document Server; ADS Abstract Service; Link to Article from SCOAP3

---

23. 462 core citations in 2017

Results from a search for dark matter in the complete LUX exposure
LUX Collaboration (D.S. Akerib (Case Western Reserve U. & KIPAC, Menlo Park & SLAC) *et al.*). Aug 26, 2016. 8 pp.
Published in **Phys.Rev.Lett. 118 (2017) no.2, 021303**
DOI: 10.1103/PhysRevLett.118.021303
e-Print: **arXiv:1608.07648 [astro-ph.CO]** | **PDF**
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
ADS Abstract Service; Link to Science News article

---

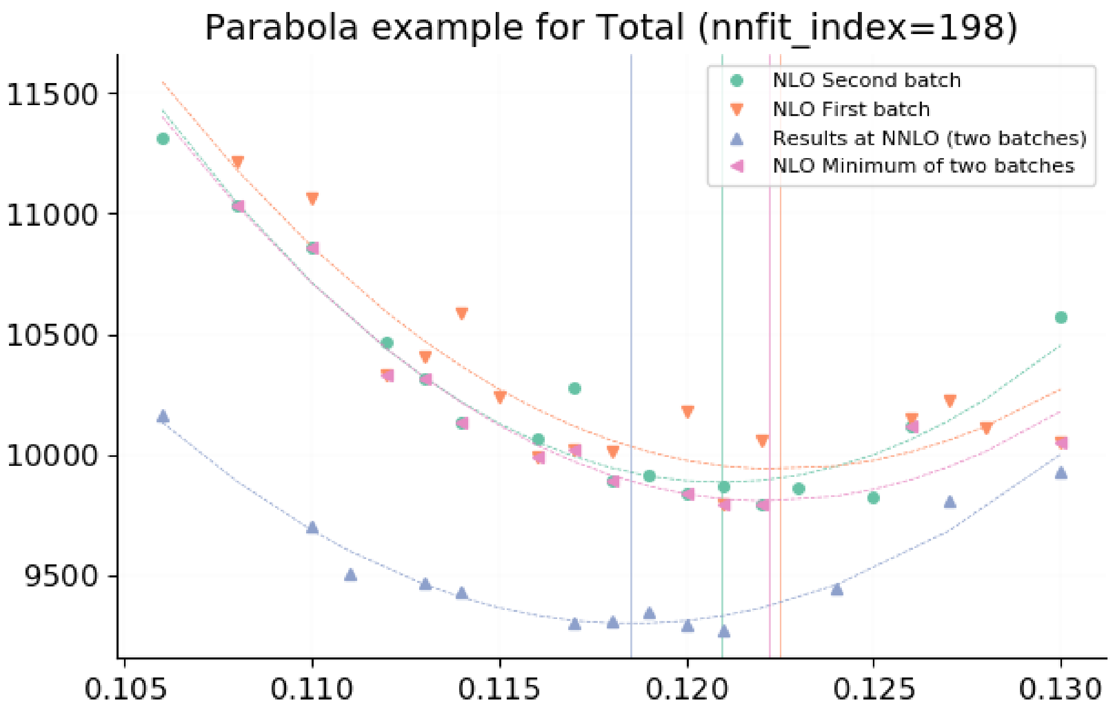
24. 434 core citations in 2017

The Inflationary Universe: A Possible Solution to the Horizon and Flatness Problems
Alan H. Guth (SLAC). Jul 1980. 32 pp.
Published in **Phys.Rev. D23 (1981) 347-356**
SLAC-PUB-2576
DOI: 10.1103/PhysRevD.23.347
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote

ALL IS WELL?

FIT QUALITY VS VALUE OF THE STRONG COUPLING

A REPRESENTATIVE PDF REPLICA



- IF REFIT, FIND DIFFERENT VALUES
- BEST OF TWO MORE STABLE SIGNIFICANT STABILIZATION

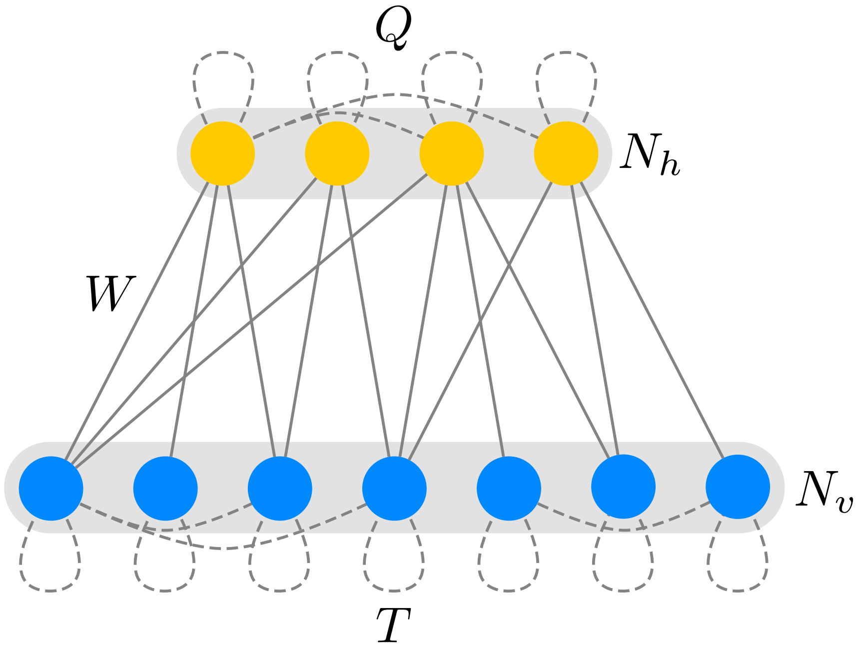
MINIMIZATION INEFFICIENCY!

MACHINE LEARNING

(REALLY!)

THE RIEMANN-THETA BOLTZMANN MACHINE: A NEW ML TOOL

$\vec{v} \Rightarrow$ VISIBLE, $v_i \in \mathbb{R}$; $\vec{h} \Rightarrow$ HIDDEN, $h_i \in \mathbb{Z}$



ONE-DIMENSIONAL

- TUNE PARMS SO $P(v) = \sum_h P(v,h)$ REPRODUCES PROBABILITY
- USE $P(h|v)$ AS NN ACTIVATION FUNCTION

FROM NNPDF…

NNPDF

THE COLLABORATION

…TO THE N^{3}PDF PROJECT

N3PDF

Machine Learning • PDFs • QCD

THE TEAM





STAY TUNED!

NNPDF Collaboration & N3PDF Kickoff Meeting

16-19 September 2018
Palazzo Feltrinelli, Gargnano, Lake Garda
Europe/Rome timezone

Search...

- Overview
- Timetable
- Contribution List
- My Conference
  - My Contributions
- Participants List

…TOWARDS A SINGLE AI AGENT FOR PDF DETERMINATION

EXTRAS

CONTEMPORARY PDF TIMELINE (ONLY PUBLISHED GLOBAL)

	2008		2009		2010		2011	2012		2013		2014		2015	2017	
SET	CTEQ6.6	NNPDF1.0	MSTW	ABKM09	NNPDF2.0	CT10 (NLO)	NNPDF2.1 (NNLO)	ABM11	NNPDF2.3	CT10 (NNLO)	ABM12	NNPDF3.0	MMHT	CT14	ABMP16	NNPDF3.1
MONTH	(02)	(08)	(01)	(08)	(02)	(07)	(07)	(02)	(07)	(02)	(10)	(10)	(12)	(06)	(01)	(06)
F. T. DIS	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
ZEUS+H1-HI	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
COMB. HI	✗	✗	✗	✗	✔	✗	✔	✗	✔	✗	✔	✔	✗	✗	✔	✔
ZEUS+H1-HII	✗	✗	✗	✗	✗	✗	some	✗	✗	some	✗	✔	✗	✗	✔	✔
HERA JETS	✗	✗	✔	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔	✗	✗	✗
F. T. DY	✔	✗	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
TEV W+Z	✔	✗	✔	✗	✔	✔	✔	✗	✔	✔	✗	✔	✔	✔	✗	✔
LHC W+Z	✗	✗	✗	✗	✗	✗	✗	✗	✔	✗	some	✔	✔	✔	some	✔
TEV JETS	✔	✗	✔	✗	✔	✔	✗	✔	✔	✔	✗	✔	✔	✔	✗	✔
LHC JETS	✗	✗	✗	✗	✗	✗	✗	✗	✔	✗	✗	✔	✔	✔	✗	✔
TOP TOTAL	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔	✔	✗	✗	✔	✔
SINGLE TOP TOTAL	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔	✗
TOP DIFFERENTIAL	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔
W p_T	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔	✗	✗	✗	✗
W+C	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔	✗	✗	✗	✗
Z p_T	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔

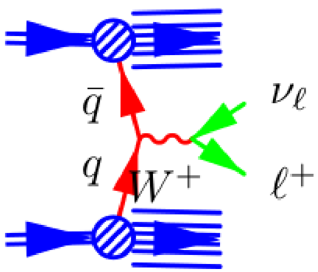
THEORY PROGRESS:

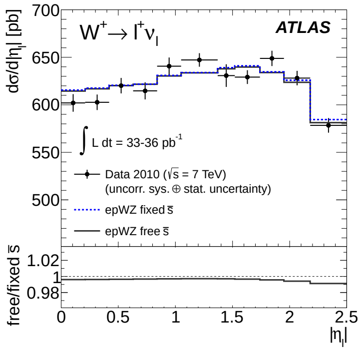
- MSTW, ABKM: all NNLO; NNPDF NNLO since 07/11 (2.1), CT since 02/13 (CT10);
  NNPDF THRESHOLD RESUMMATION (3.0RESUM, 07/15), SMALL x RESUMMATION (3.1SX, 10/17)
- MSTW, CT, NNPDF all GM-VFN; NNPDF since 01/11 (2.1);
  ABM FFN+ZM-VFN since 01/17 (ABMP16)
- NNPDF FITTED CHARM since 05/16 (NNPDF3IC)
- PHOTON PDF: (mrst2004qed), NNPDF2.3QED (08/13), NNPDF3.0QED (06/16), NNPDF3.1LUXQED (12/17)

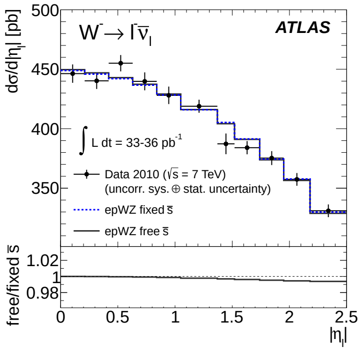
DISENTANGLING PDFs

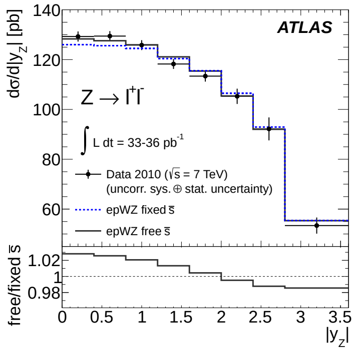
EXAMPLE: $W^\pm$ AND Z PRODUCTION

W AND Z CROSS SECTIONS

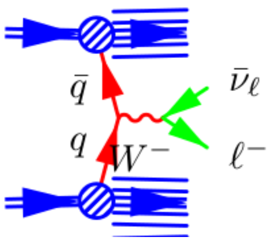






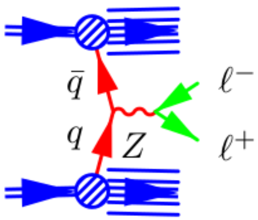


ATLAS (2012)

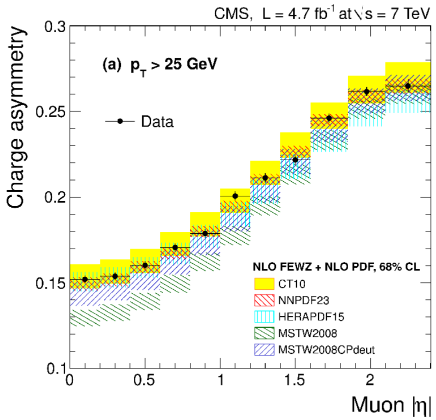


. $\sigma^{p\bar p}_{W^+} = u\bar d + c\bar s$; $\sigma^{p\bar p}_{Z} = u\bar u + d\bar d + s\bar s$: STRANGENESS DETERMINED

BY COMPARISON



W MUON ASYMMETRY



$$\frac{\sigma^{p\bar p}_{W^+}}{\sigma^{p\bar p}_{W^-}} = \frac{u(x_1)\bar d(x_2)+\bar d(x_1)u(x_2)}{d(x_1)\bar u(x_2)+\bar u(x_1)d(x_2)}$$

"VALENCE" $x \Rightarrow$ NEGLECT STRANGENESS

$\Rightarrow$ DETERMINE $\bar u - \bar d$

CMS (2013)

GENETIC MINIMIZATION

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

- LARGE NUMBER OF MUTANT (~ 100) PDF SETS GENERATED FROM PARENT
- FIGURE OF MERIT COMPUTED
- BEST-FIT KEPT & PASSED TO NEXT GENERATION

$$w \to w + \frac{\eta r_\delta}{N_{\rm ite}^{r_{\rm ite}}}$$

CHOICES

- MUTATION RATE η
- POINTLIKE VS. NODAL MUTATION
- NUMBER (POINTLIKE) OR PROBABILITY (NODAL) OF MUTATIONS
- TARGETED WT: WEIGTHS $p_i = E_i/E_i^{\rm targ}$
- GA EPOCHS: $N_{\rm gen}^{\rm mut}$

	$N_{\rm gen}^{\rm wt}$	$N_{\rm gen}^{\rm mut}$	$N_{\rm gen}^{\rm max}$	$E^{\rm sw}$	$N_{\rm mut}^a$	$N_{\rm mut}^b$
NNPDF 2.3	10000	2500	50000	2.3	80	30
NNPDF 3.0	–	-	30000	-	80	-

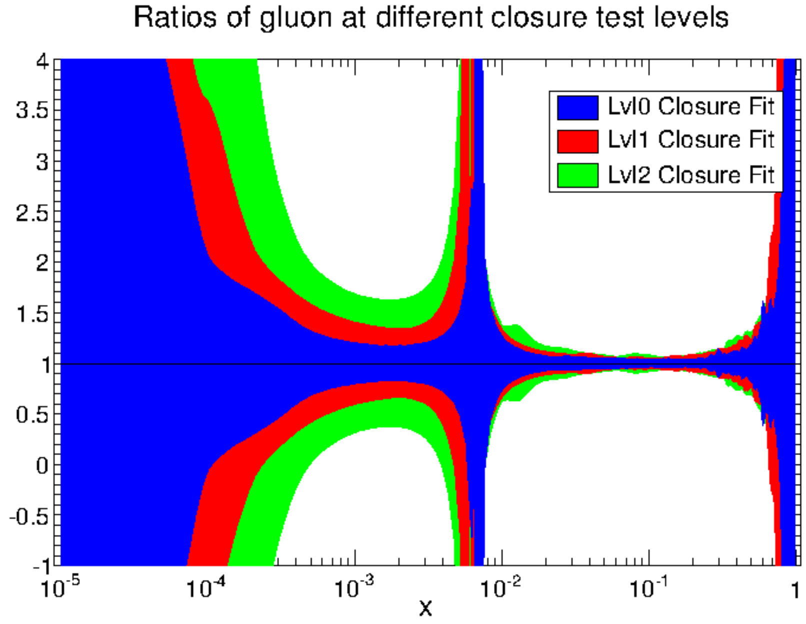
NNPDF2.3		
Single Parameter Mutation		
PDF	$N_{\rm mut}$	η
$\Sigma(x)$	2	10, 1
$g(x)$	3	10, 3, 0.4
$T_3(x)$	2	1, 0.1
$V(x)$	3	8, 1, 0.1
$\Delta_S(x)$	3	5, 1, 0.1
$s^+(x)$	2	5, 0.5
$s^-(x)$	2	1, 0.1

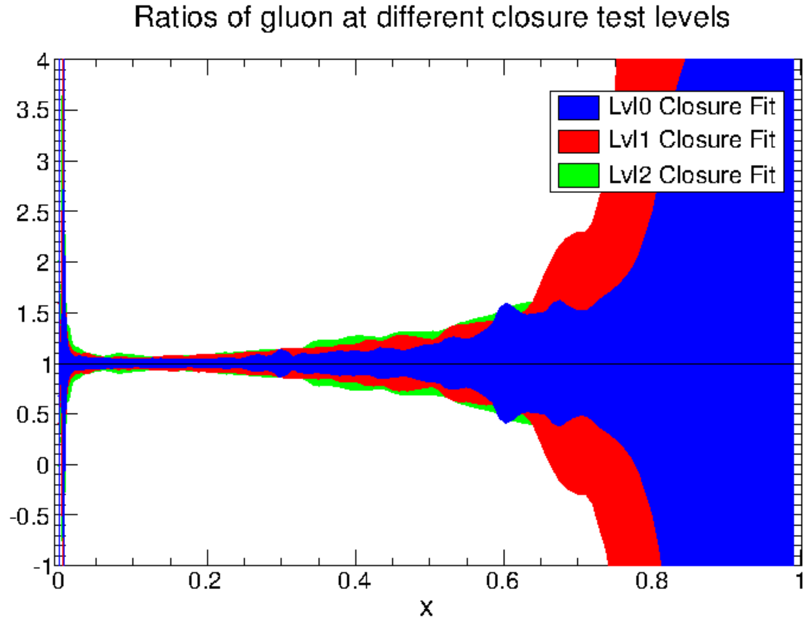
NNPDF3.0		
Nodal Mutation		
PDF	$P_{\rm mut}$	η
$\Sigma(x)$	5% per node	15
$g(x)$	5% per node	15
$V(x)$	5% per node	15
$V_3(x)$	5% per node	15
$V_8(x)$	5% per node	15
$T_3(x)$	5% per node	15
$T_8(x)$	5% per node	15

TRACING SOURCES OF UNCERTAINTY

- LEVEL 0: FAKE DATA GENERATED WITH NO UNCERTAINTY
  → INTERPOLATION AND EXTRAPOLATION UNCERTAINTY
- LEVEL 1-2: FAKE DATA GENERATED WITH SAME UNCERTAINTY AS REAL DATA (INCLUDING CORRELATIONS)
- LEVEL 1: NO PSEUDODATA REPLICAS:
  ⇒ REPLICAS FITTED TO SAME DATA OVER AND OVER AGAIN
  → FUNCTIONAL UNCERTAINTY DUE TO INFINITY OF EQUIVALENT MINIMA
- LEVEL 2: STANDARD NNPDF METHODOLOGY
  ⇒ REPLICAS FITTED TO PSEUDODATA REPLICAS
  → DATA UNCERTAINTY
- THREE SOURCES OF UNCERTAINTY COMPARABLE IN DATA REGION

THE GLUON: LEVEL 0, LEVEL 1 AND LEVEL 2



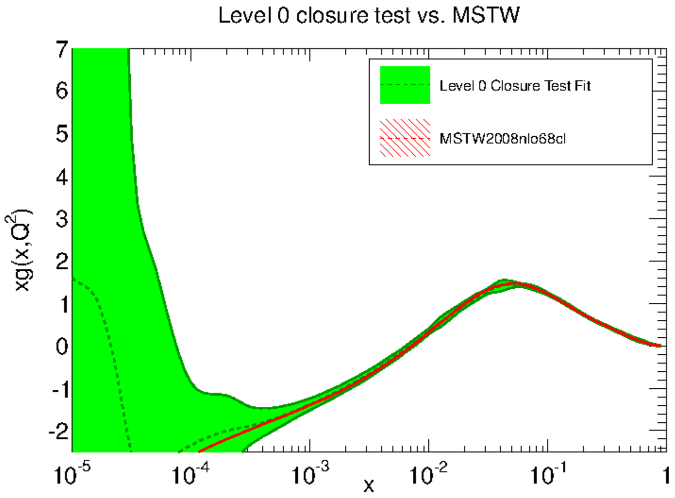


FITTING EFFICIENCY

LEVEL 0

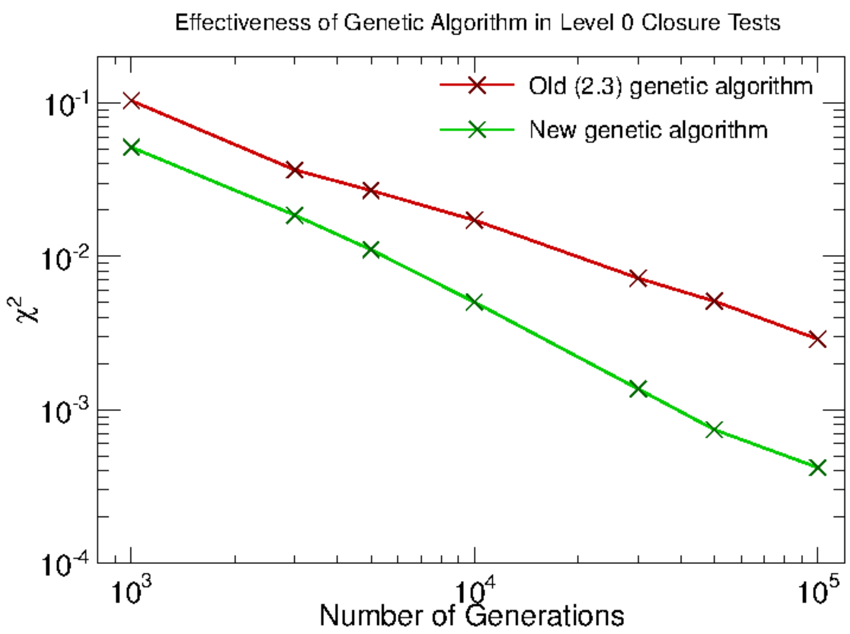
- ASSUME VANISHING EXPERIMENTAL UNCERTAINTY
- MUST BE ABLE TO GET $\chi^2 = 0$
- UNCERTAINTY AT DATA POINTS TENDS TO ZERO (NOT NECESSARILY ON PDF!)
  DEFINE $\phi \equiv \sqrt{\langle \chi^2_{rep} \rangle - \chi^2}$,
  EQUALS FIT UNCERTAINTY/DATA UNCERTAINTY; CHECK $\phi \to 0$
- CAN STUDY EFFICIENCY OF MINIMIZATION

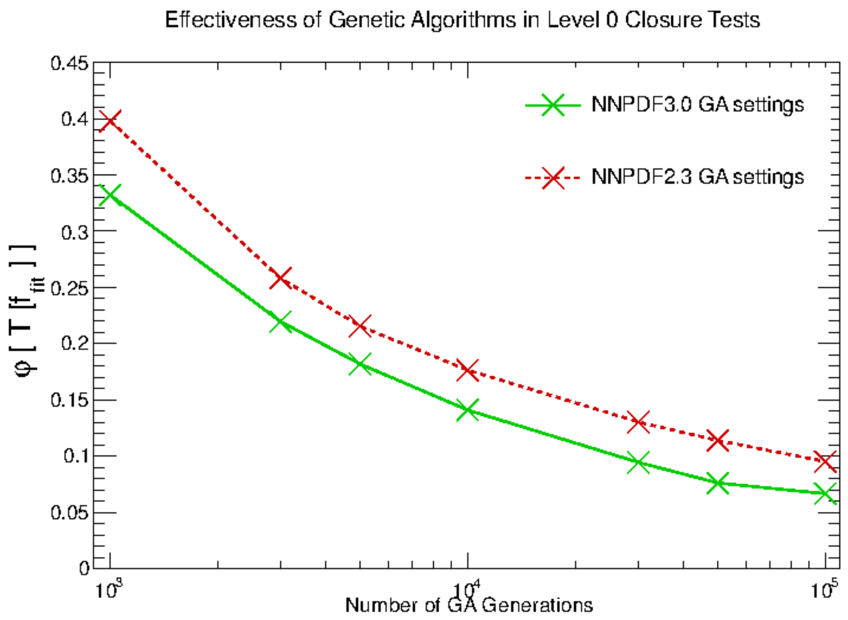
THE GLUON



FRACTIONAL UNCERTAINTY VS TRAINING LENGTH

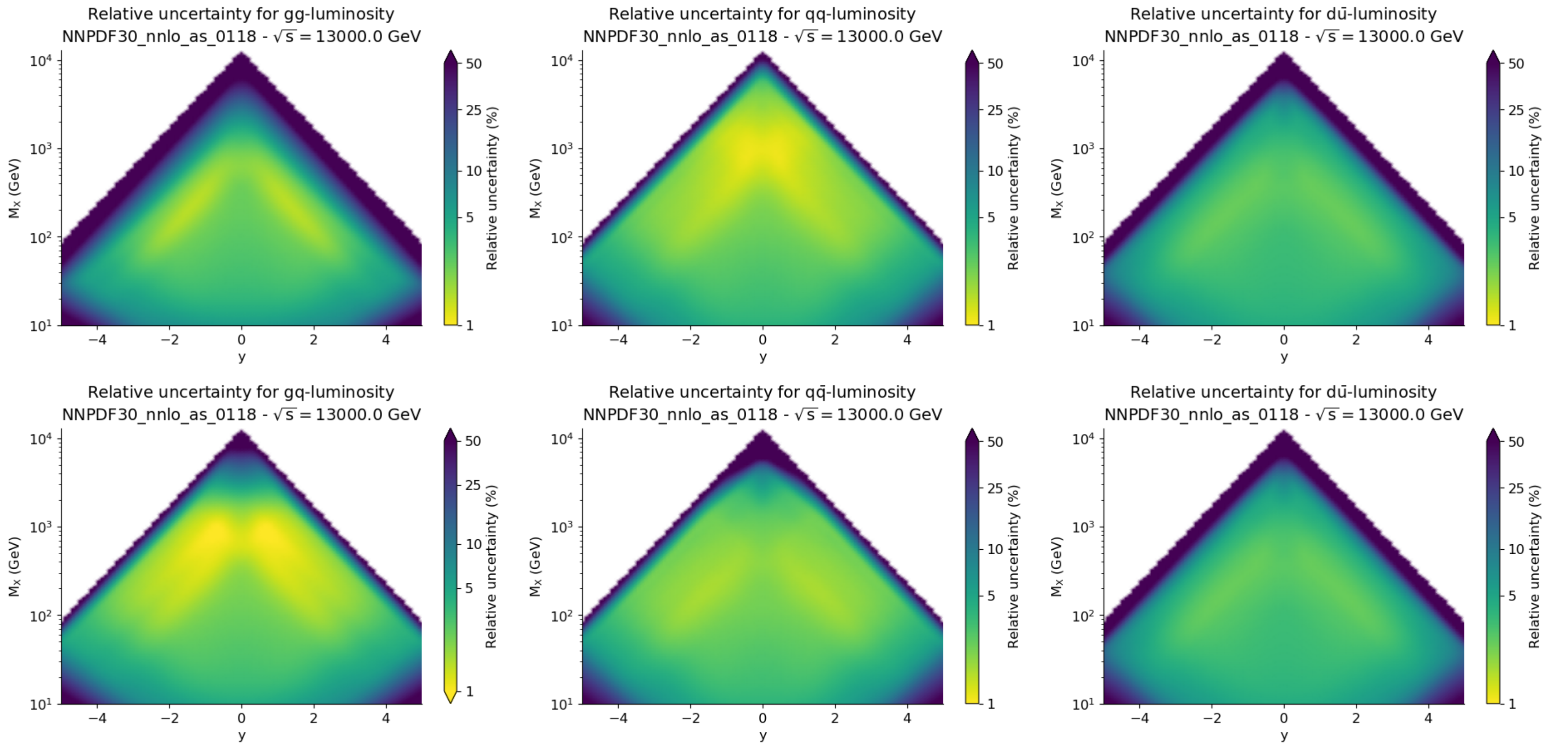
χ^2 VS TRAINING LENGTH





THE IMPACT OF LHC DATA
PDF UNCERTAINTIES: PAST ⇒ PRESENT (NNPDF3.0 NNLO)

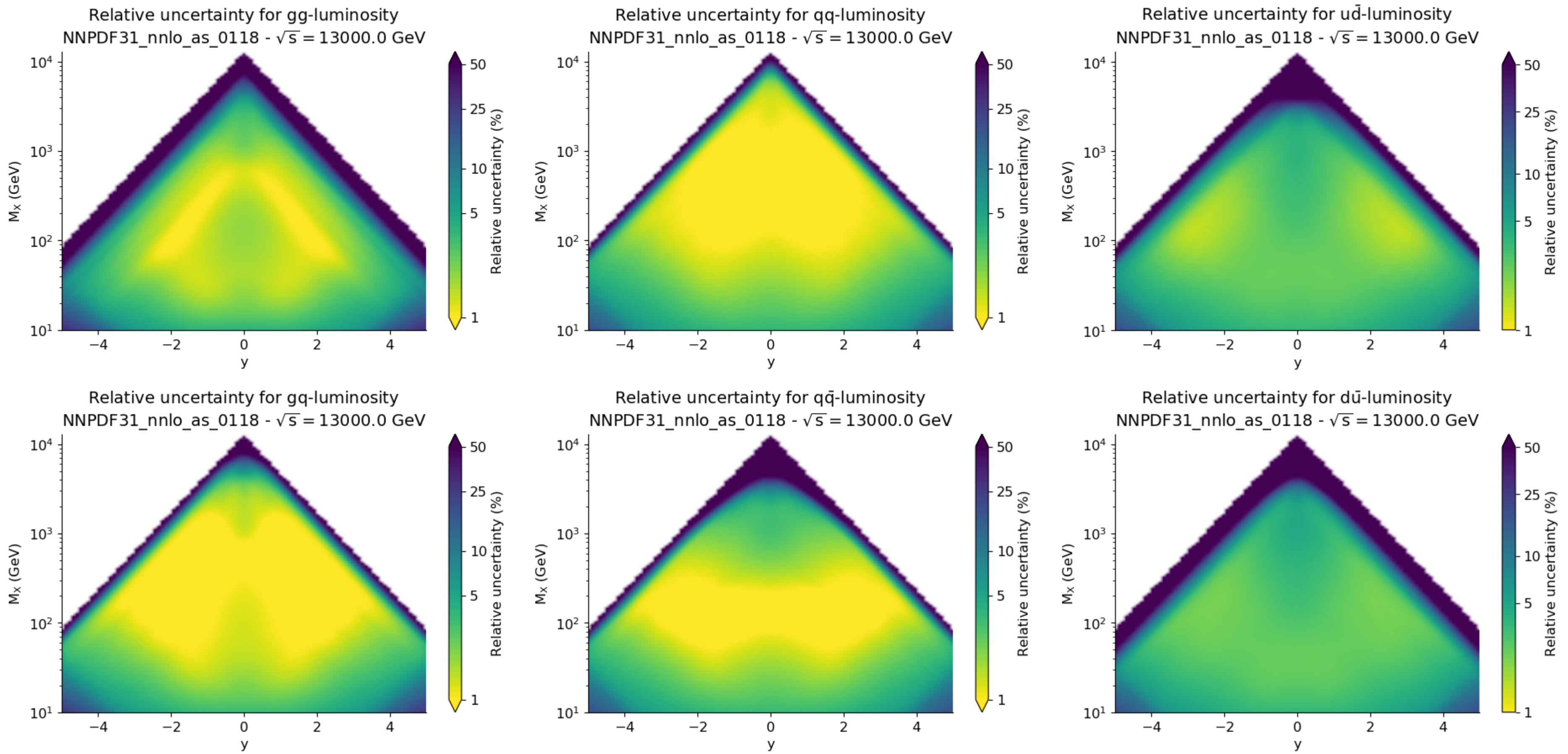
GLUON SINGLET FLAVORS



- GLUON BETTER KNOWN AT SMALL x, VALENCE QUARKS AT LARGE x, SEA QUARKS IN BETWEEN
- SWEET SPOT: VALENCE Q - G; UNCERTAINTIES DOWN TO 1%
- UP BETTER KNOWN THAN DOWN; FLAVOR SINGLET BETTER THAN INDIVIDUAL FLAVORS

THE IMPACT OF LHC DATA
PDF UNCERTAINTIES: PRESENT ⇒ FUTURE (NNPDF3.1 NNLO)

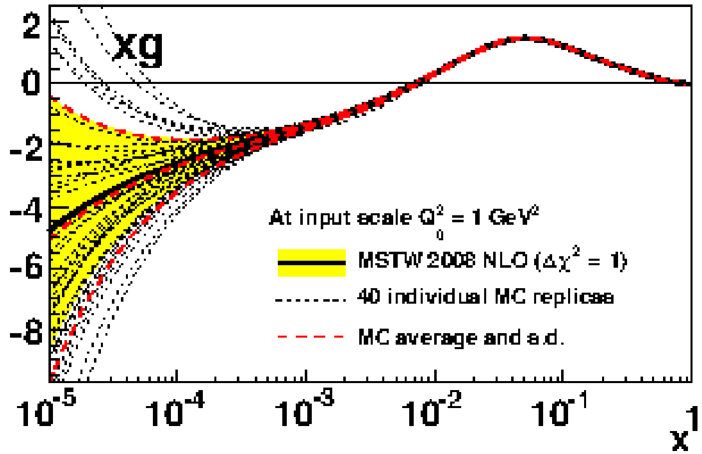
GLUON SINGLET FLAVORS



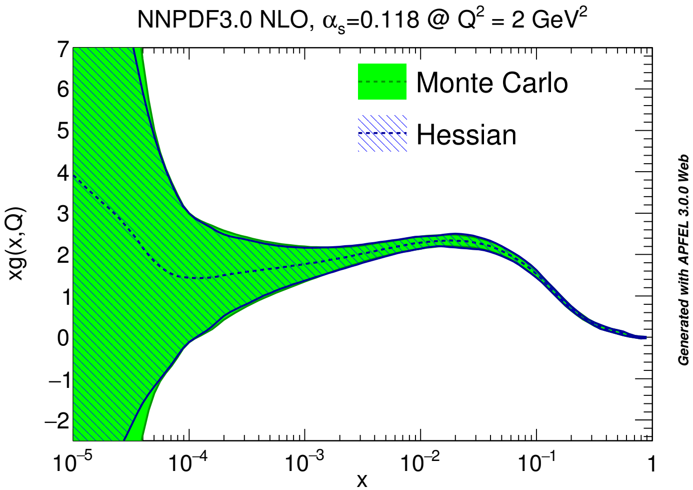
- GLUON BETTER KNOWN AT SMALL x, VALENCE QUARKS AT LARGE x, SEA QUARKS IN BETWEEN
- SWEET SPOT: VALENCE Q - G; UNCERTAINTIES DOWN TO 1%
- UP BETTER KNOWN THAN DOWN; FLAVOR SINGLET BETTER THAN INDIVIDUAL FLAVORS
- NEW LHC DATA ⇒ SIZABLE REDUCTION IN UNCERTAINTIES

PROGRESS I
MC $\Leftrightarrow$ HESSIAN

- TO CONVERT HESSIAN INTO MONTECARLO GENERATE MULTIGAUSSIAN REPLICAS IN PARAMETER SPACE
- ACCURATE WHEN NUMBER OF REPLICAS SIMILAR TO THAT WHICH REPRODUCES DATA



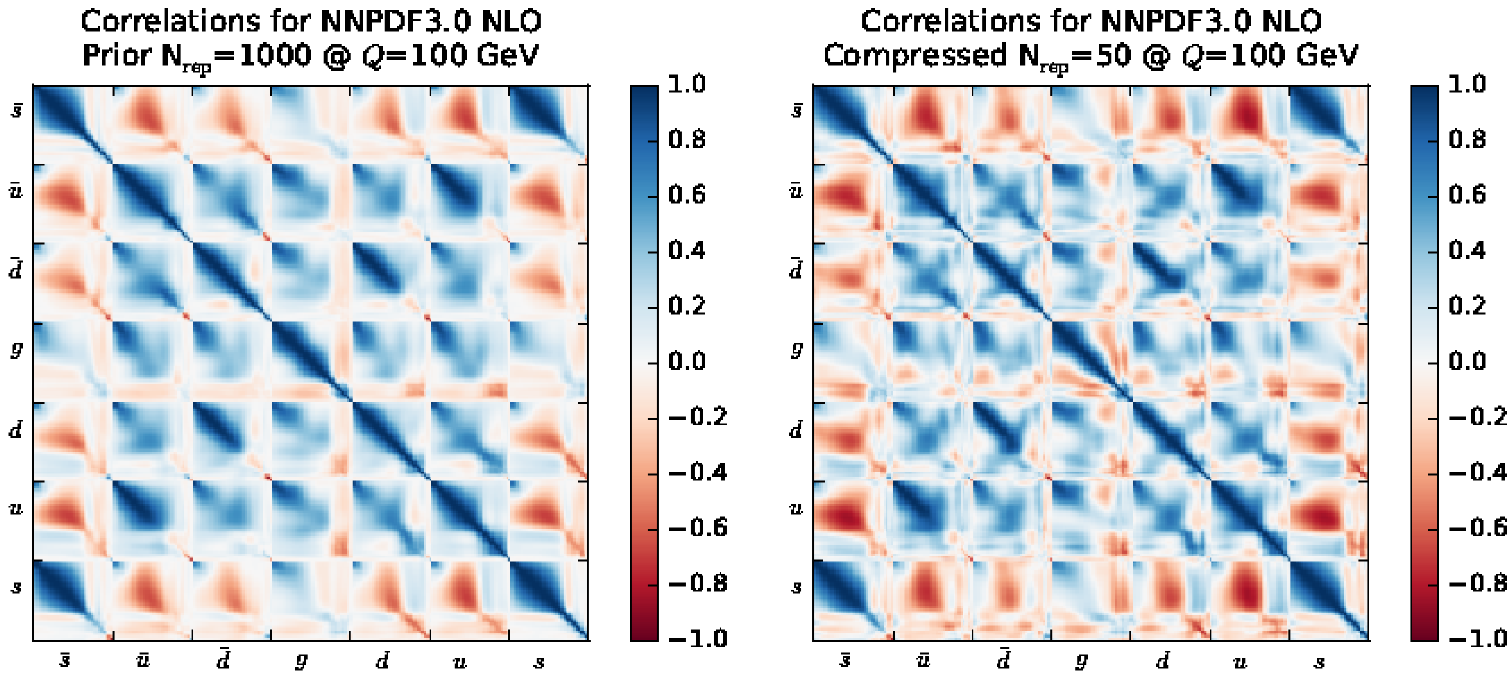
(Thorne, Watt, 2012)



(Carrazza, SF, Kassabov, Rojo, 2015)

- TO CONVERT MONTE CARLO INTO HESSIAN, SAMPLE THE REPLICAS $f_i(x)$ AT A DISCRETE SET OF POINTS & CONSTRUCT THE ENSUING COVARIANCE MATRIX
- EIGENVECTORS OF THE COVARIANCE MATRIX AS A BASIS IN THE VECTOR SPACE SPANNED BY THE REPLICAS BY SINGULAR-VALUE DECOMPOSITION
- NUMBER OF DOMINANT EIGENVECTORS SIMILAR TO NUMBER OF REPLICAS $\Rightarrow$ ACCURATE REPRESENTATION

MONTECARLO COMPRESSION



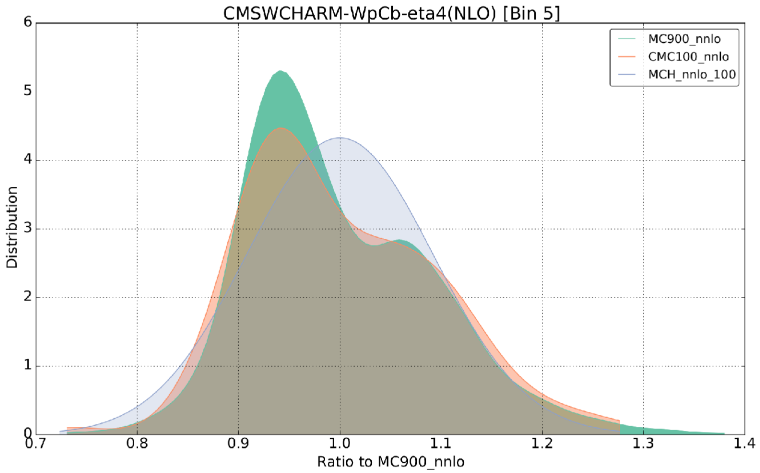
(Carrazza, Latorre, Kassabov, Rojo, 2015)

- CONSTRUCT A VERY LARGE REPLICA SAMPLE
- SELECT (BY GENETIC ALGORITHM) A SUBSET OF REPLICAS WHOSE STATISTICAL FEATURES ARE AS CLOSE AS POSSIBLE TO THOSE OF THE PRIOR
- $\Rightarrow$ FOR ALL PDFs ON A GRID OF POINTS// MINIMIZE DIFFERENCE OF: FIRST FOUR MOMENTS, CORRELATIONS; OUTPUT OF KOLMOGOROV-SMIRNOV TEST (NUMBER OF REPLICAS BETWEEN MEAN AND σ, 2σ, INFINITY)
- 50 COMPRESSED REPLICA REPRODUCE 1000 REPLICA SET TO PRECENT ACCURACY

NONGAUSSIAN BEHAVIOUR

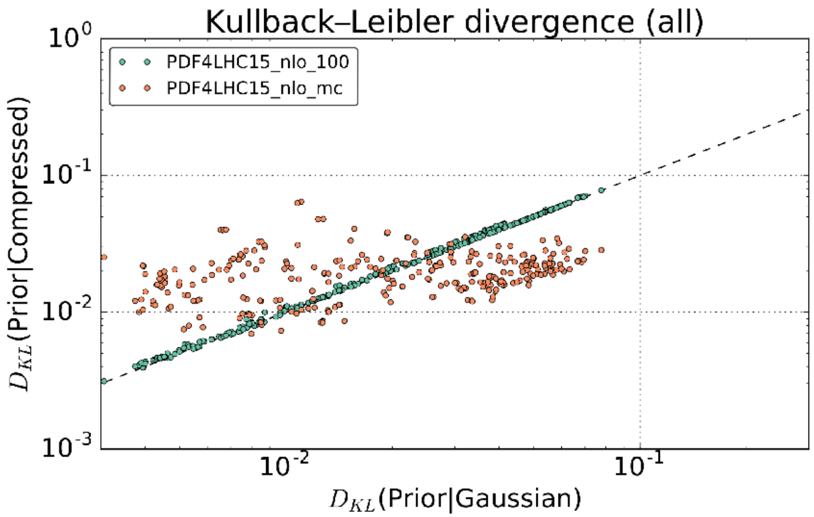
MONTE CARLO COMPARED TO HESSIAN

CMS $W + c$ production



- DEVIATION FROM GAUSSIANITY E.G. AT LARGE x DUE TO LARGE UNCERTAINTY + POSITIVITY BOUNDS
$\Rightarrow$ RELEVANT FOR SEARCHES
- CANNOT BE REPRODUCED IN HESSIAN FRAMEWORK
- WELL REPRODUCED BY COMPRESSED MC

- DEFINE KULLBACK-LEIBLER DIVERGENCE
$D_{\mathrm{KL}} = \int_{-\infty}^{\infty} P(x) \frac{\ln P(x)}{\ln Q(x)}\, dx$
BETWEEN A PRIOR P AND ITS REPRESENTATION Q
- D_{KL} BETWEEN PRIOR AND HESSIAN DEPENDS ON DEGREE OF GAUSSIANITY
- D_{KL} BETWEEN PRIOR AND COMPRESSED MC DOES NOT



CAN (A) GAUGE WHEN MC IS MORE ADVANTAGEOUS THAN HESSIAN;
(B) ASSESS THE ACCURACY OF COMPRESSION

MONTE CARLO DATA GENERATION

- BCDMS+ NMC PROTON & DEUTERON F_2 DATA (FULL CORRELATED SYSTEMATICS AVAILABLE), TAKEN AT 4 BEAM ENERGIES

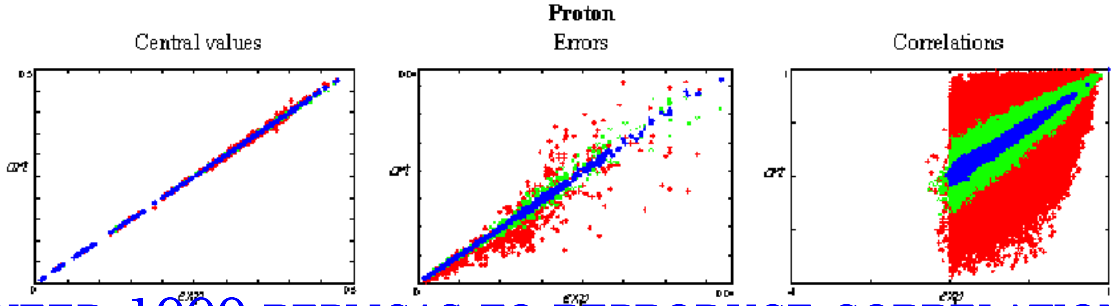
- ON TOP OF STAT. ERRORS, 4 SYSTEMATICS + 1 NORMALIZATION (NMC) OR 6 SYSTEMATICS + 1 ABSOLUTE & 2 RELATIVE NORMALIZATIONS (BCDMS), WITH VARIOUS FORMS OF CORRELATION (FULL, OR FOR EACH TARGET, OR FOR EACH BEAM ENERGY)

GENERATE DATA ACCORDING TO A MULTIGAUSSIAN DISTRIBUTION

$$F_i^{(art)\,(k)} =$$

$$(1 + r_5^{(k)}\,\sigma_N)\sqrt{1 + r_{i,6}^{(k)}\,\sigma_{N_t}}\sqrt{1 + r_{i,7}^{(k)}\,\sigma_{N_b}}\left[F_i^{(exp)} + \frac{r_{i,1}^{(k)}\,f_b + r_{i,2}^{(k)}\,f_{i,s} + r_{i,3}^{(k)}\,f_{i,r}}{100}F_i^{(exp)} + r_{i,s}^{(k)}\,\sigma_s^i\right]$$

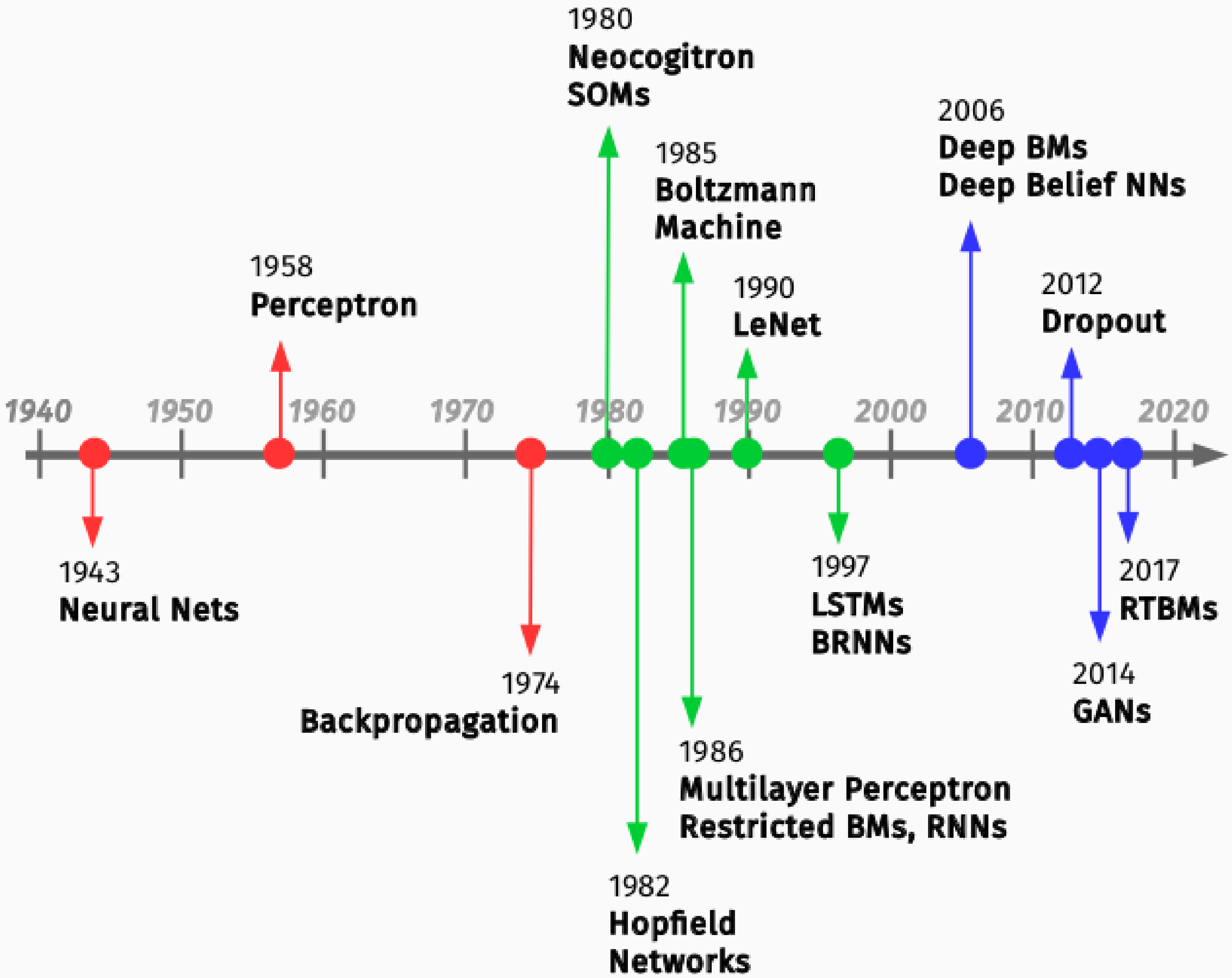
r univariate gaussian random nos., one $r_{i,s}$ for each data, but single $r_{i,j}$ for all correlated data



SCATTER PLOT ART. VS. EXP. FOR 10 (RED) 100 (GREEN) AND 1000 (BLUE) REPLICAS

NEED 1000 REPLICAS TO REPRODUCE CORRELATIONS TO PERCENT ACCURACY

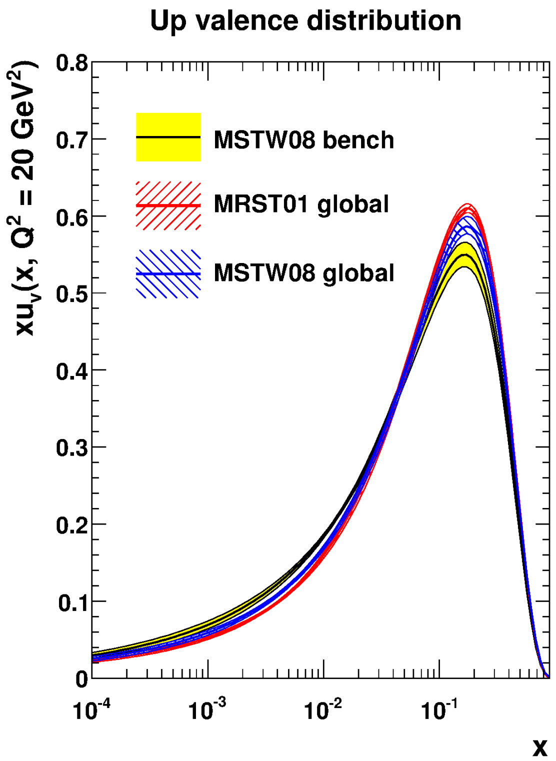
Non-linear models timeline



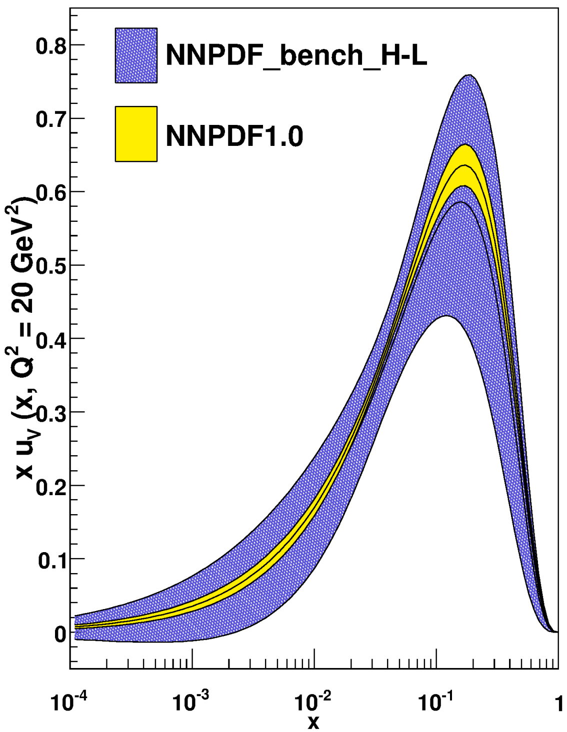
WHAT DETERMINES PDF UNCERTAINTIES?

THE NNPDF SOLUTION (HERALHC 2008)

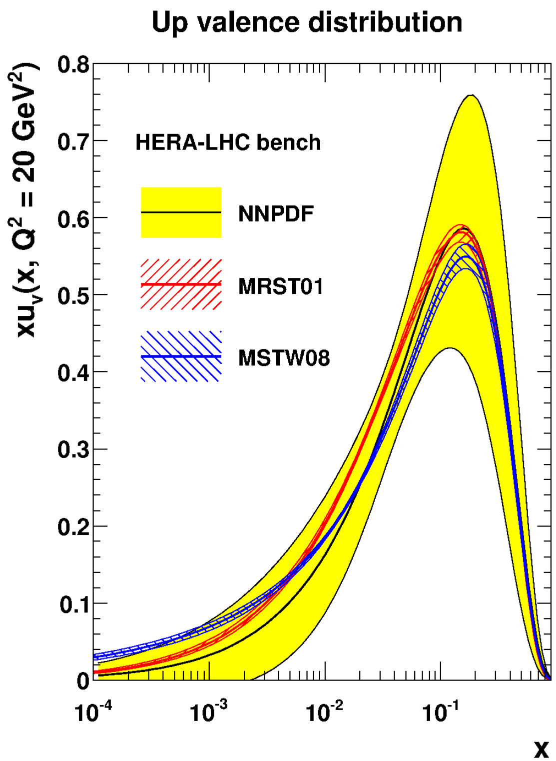
MRST/MSTW: BENCH VS REF



NNPDF: BENCH VS REF



NNPDF BENCH VS MRST/MSTW BENCH



- SINGLE PARAMETRIZATION AND STAT. TREATMENT CAN ACCOMMODATE DIFFERENT DATASETS
- IMPACT OF DATA CAN BE STUDIED INDEPENDENT OF THEORETICAL FRAMEWORK