Determining parton distribution functions accurately and precisely with machine learning

Digital Twins for Nuclear and Particle physics - NPTwins 2024

Emanuele R. Nocera

Università degli Studi di Torino and INFN, Torino

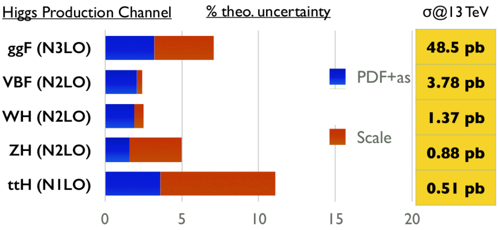
18 December 2024

Parton Distribution Functions at the LHC

$$\sigma(Q^2,\tau,\mathbf{k}) = \sum_{ij}\int_\tau^1 \frac{dz}{z}\mathcal{L}_{ij}(z,Q^2)\hat{\sigma}_{ij}\left(\frac{\tau}{z},\alpha_s(Q^2),\mathbf{k}\right) \qquad \mathcal{L}_{ij}(z,Q^2) = (f_i^{h_1}\otimes f_j^{h_2})(z,Q^2)$$

PDF uncertainty is often the dominant source of uncertainty in LHC cross sections

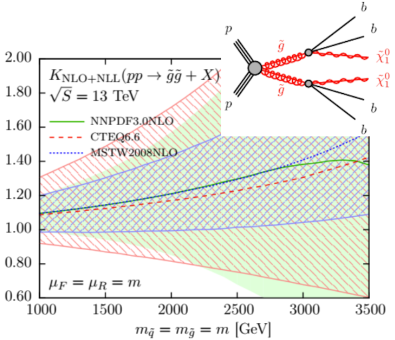
Precision



Unc. [MeV]	Total	Stat.	Syst.	PDF	A_i	Backg.	EW	e	μ	u_T	Lumi	Γ_W	PS
p_T^ℓ	16.2	11.1	11.8	4.9	3.5	1.7	5.6	5.9	5.4	0.9	1.1	0.1	1.5
m_T	24.4	11.4	21.6	11.7	4.7	4.1	4.9	6.7	6.0	11.4	2.5	0.2	7.0
Combined	15.9	9.8	12.5	5.7	3.7	2.0	5.4	6.0	5.4	2.3	1.3	0.1	2.3

[CERN Yellow Report 2016; arXiv:2403.15085]

Discovery



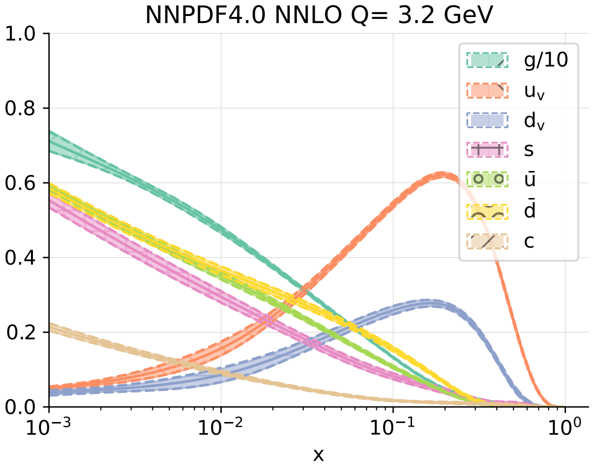
[EPJC 76 (2016) 53]

Parton Distribution Functions

PDFs express the likelihood of a quark or gluon (partons) to enter a collision

That is, $x\times$PDFs are momentum fraction distributions for each parton

Dependence on x is non-perturbative (fit); dependence on Q^2 is perturbative (DGLAP)



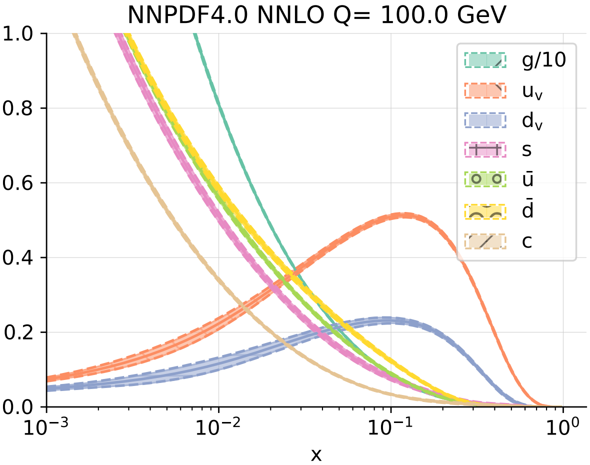
[Plot from the PDG Review of Particle Physics]

Parton Distribution Functions

PDFs express the likelihood of a quark or gluon (partons) to enter a collision

That is, $x\times$PDFs are momentum fraction distributions for each parton

Dependence on x is non-perturbative (fit); dependence on Q^2 is perturbative (DGLAP)



[Plot from the PDG Review of Particle Physics]

PDF determination in statistical language

<u>Inverse problem</u>

Given a set of data D, determine $p(f|D)$ in the space of functions $f : [0,1] \to \mathbb{R}$.

<u>Solution: parametric regression</u>

Approximate $p(f|D)$ with its projection in the space of parameters $p(\boldsymbol{\theta}|D)$

$$xf_i(x, Q_0^2) = A_{f_i}\, x^{a_{f_i}}\, (1-x)^{b_{f_i}}\, \mathscr{F}(x, \{c_{f_i}\})$$

Determine $p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})$ as MAP $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|D)$

$$\chi^2 = \sum_{i,j}^{N_{\rm dat}} [T_i[\boldsymbol{\theta}] - D_i](\mathrm{cov}^{-1})_{ij}[T_j[\boldsymbol{\theta}] - D_j]$$

Use a prescription to compute expectation values and uncertainties of observables

$$E[\mathcal{O}] = \int \mathcal{D}f \mathcal{P}(f|D)\mathcal{O}(f) \qquad V[\mathcal{O}] = \int \mathcal{D}f \mathcal{P}(f|D)[\mathcal{O}(f) - E[\mathcal{O}]]^2$$

Monte Carlo: $\mathcal{P}(f|D) \longrightarrow \{f_k\}$

$E[\mathcal{O}] \approx \frac{1}{N}\sum_k \mathcal{O}(f_k)$

$V[\mathcal{O}] \approx \frac{1}{N}\sum_k [\mathcal{O}(f_k) - E[\mathcal{O}]]^2$

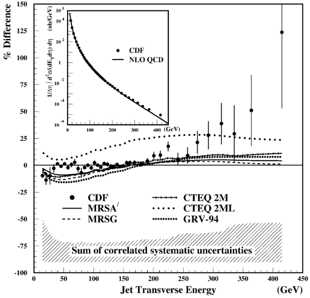
Maximum likelihood: $\mathcal{P}(f|D) \longrightarrow f_0$

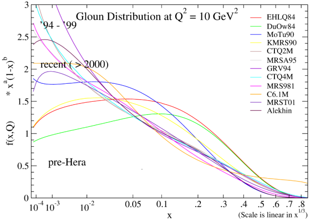
$E[\mathcal{O}] \approx \mathcal{O}(f_0)$

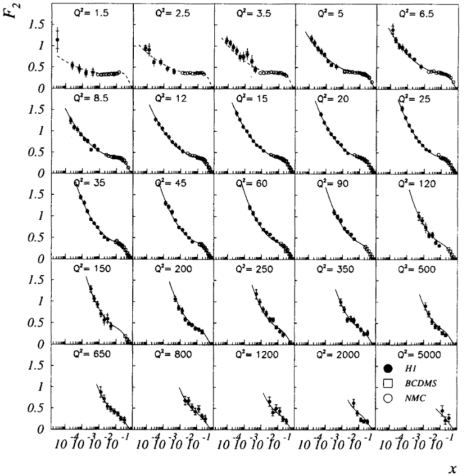
$V[\mathcal{O}] \approx$ Hessian, $\Delta\chi^2$envelope, ...

Interplay between DATA, THEORY, and METHODOLOGY

Why is the methodology important?



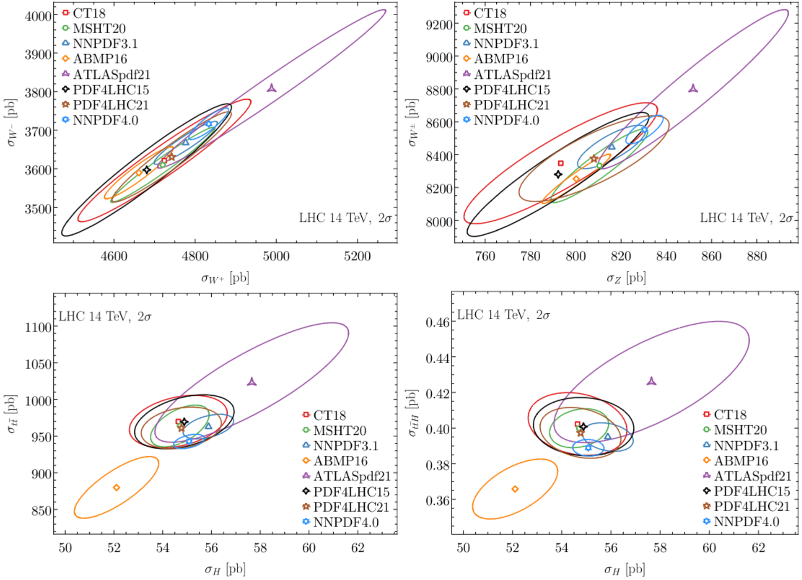




circa 1995: small-x rise of HERA F_2^p and CDF jet discrepancy

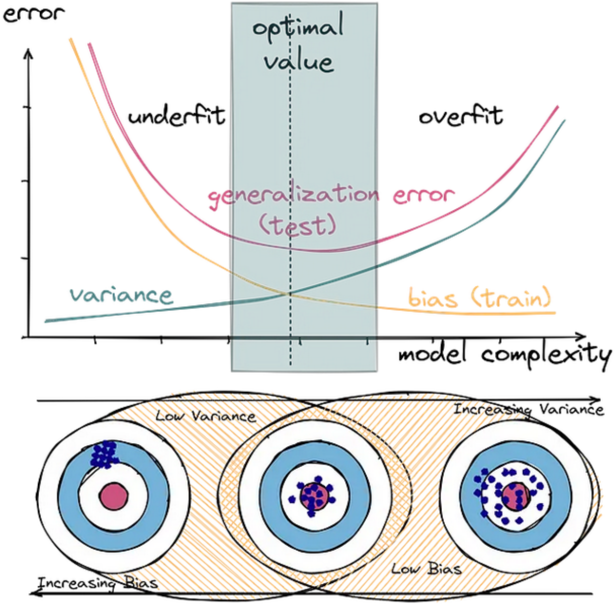
The methodology is crucial if we aim at percent-level accurate PDFs

Making predictions with PDFs



[Acta Phys.Polon.B 53 (2022) 12]

Accuracy vs precision or bias vs variance



A lot of progress in the last (two) years

A Markov Chain Monte Carlo determination of Proton PDF uncertainties at NNLO

Peter Risse [a,*] **Nasim Derakhshanian** [b] **Tomas Ježo** [a] **Karol Kovařík** [a] **and Aleksander Kusina** [b]
[a] *Institut für Theoretische Physik, Universität Münster, Wilhelm-Klemm-Straße 9, D-48149 Münster, Germany*
[b] *Institute of Nuclear Physics Polish Academy of Sciences, PL-31342 Krakow, Poland*
E-mail: risse.p@uni-muenster.de

ANL-186490

Learning PDFs through Interpretable Latent Representations in Mellin Space

Brandon Kriesten and T. J. Hobbs
High Energy Physics Division, Argonne National Laboratory, Lemont, IL 60439
(Dated: June 21, 2024)

Eur. Phys. J. C (2024) 84:716
https://doi.org/10.1140/epjc/s10052-024-13100-1
THE EUROPEAN PHYSICAL JOURNAL C
Regular Article - Theoretical Physics

Bayesian inference with Gaussian processes for the determination of parton distribution functions

Alessandro Candido[1], **Luigi Del Debbio**[2], **Tommaso Giani**[3,4,a], **Giacomo Petrillo**[5]
[1] Theoretical Physics Department, CERN, 1211 Geneva 23, Switzerland
[2] Higgs Centre for Theoretical Physics, School of Physics and Astronomy, Peter Guthrie Tait Road, Edinburgh EH9 3 FD, UK
[3] Department of Physics and Astronomy, Vrije Universiteit, 1081 HV Amsterdam, The Netherlands
[4] Nikhef Theory Group, Science Park 105, 1098 XG Amsterdam, The Netherlands
[5] Dipartimento di Statistica, Informatica, Applicazioni "Giuseppe Parenti" (DISIA), Università di Firenze, Viale Morgagni 59, 50134 Firenze, Italy

Received: 7 May 2024 / Accepted: 1 July 2024 / Published online: 22 July 2024

MSUHEP-24-002

A generalized statistical model for fits to parton distributions

Mengshi Yan,[1] Tie-Jiun Hou,[2] Zhao Li,[3,4] Kirtimaan Mohan,[5,*] and C.-P. Yuan[5]
[1] *Department of Physics and State Key Laboratory of Nuclear Physics and Technology, Peking University, Beijing 100871, China.*
[2] *School of Nuclear Science and Technology, University of South China, Hengyang 421001, China.*
[3] *Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China; School of Physics Sciences, University of Chinese Academy of Sciences, Beijing 100039, China*
[4] *Center for High Energy Physics, Peking University, Beijing 100871, China.*
[5] *Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824, U.S.A.*
(Dated: June 5, 2024)

PUBLISHED FOR SISSA BY SPRINGER
RECEIVED: *April 26, 2024*
REVISED: *September 18, 2024*
ACCEPTED: *November 5, 2024*
PUBLISHED: *December 10, 2024*

A critical study of the Monte Carlo replica method

Mark N. Costantini[a] **Maeve Madigan**[b] **Luca Mantani**[a] **and James M. Moore**[a]
[a] *DAMTP, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WA, U.K.*
[b] *Institut für Theoretische Physik, Universität Heidelberg, Philosophenweg 16, D-69120, Heidelberg, Germany*
E-mail: mnc33@cam.ac.uk, madigan@thphys.uni-heidelberg.de, luca.mantani@maths.cam.ac.uk, jmm232@cam.ac.uk

PUBLISHED FOR SISSA BY SPRINGER
RECEIVED: *August 5, 2024*
ACCEPTED: *October 11, 2024*
PUBLISHED: *November 5, 2024*

Explainable AI classification for parton density theory

Brandon Kriesten[a] **Jonathan Gomprecht**[a,b] **and T.J. Hobbs**[a]
[a] *High Energy Physics Division, Argonne National Laboratory, Lemont, IL 60439, U.S.A.*
[b] *Department of Physics, University of Arizona, Tucson, AZ 85721, U.S.A.*
E-mail: bkriesten@anl.gov, jgomprecht@arizona.edu, tim@anl.gov

[EPJ C84 (2024) 716; JHEP 11 (2024) 007; JHEP 12 (2024) 064; arXiv:2312.02278; arXiv:2406.01664; arXiv:2407.12377]

NNPDF4.0: a machine learning methodology

uncertainty representation

Monte Carlo sampling of experimental uncertainties
what is the statistical meaning of uncertainties?

parametrisation

neural network(s)
is there a bias due to the parametrisation?

optimisation

(adaptive) gradient descent
is the parameter space explored efficiently?

delivery

GAN enhancement and compression
can the number of replicas be reduced?

uncertainty characterisation and validation

closure tests (what happens if I know in advance the underlying law that I am fitting?)
are interpolation and extrapolation uncertainties statistically faithful?

[EPJC 82 (2022) 428]

The NNPDF code is public, see **https://github.com/NNPDF** [EPJ C81 (2021) 958]

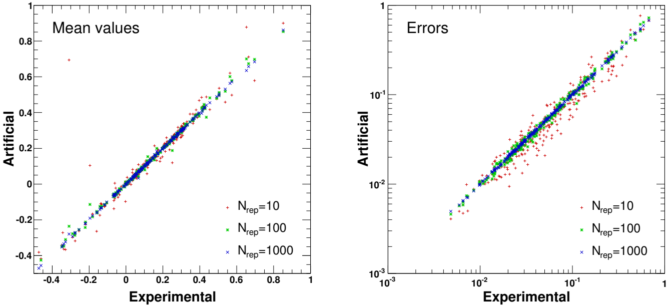
Uncertainty representation

Generate Monte Carlo replicas and perform a fit to each replica

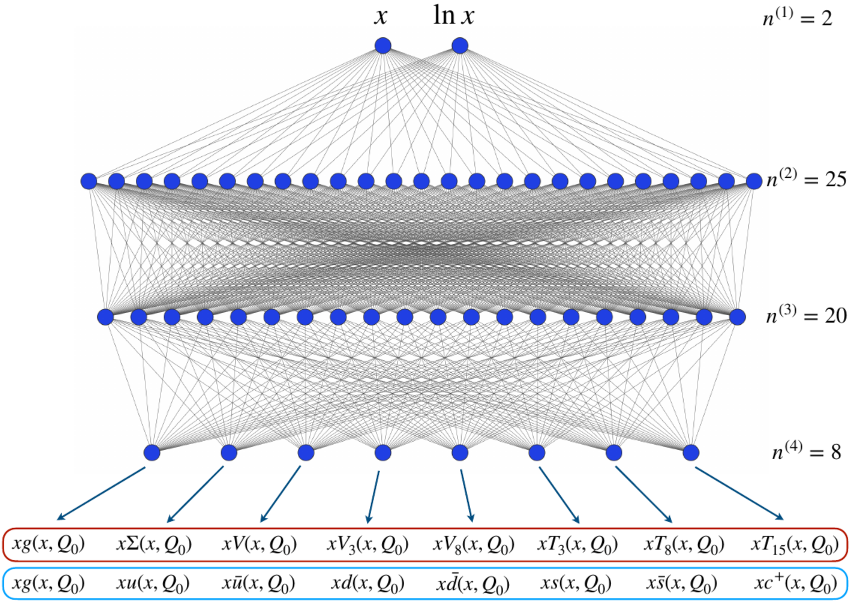
$$\mathcal{O}_i^{(art)(k)} = \mathcal{O}_i^{(exp)} + L_{ij} r_i^{(k)} \quad i,j = 1 \dots N_{\rm dat} \quad k = 1 \dots, N_{\rm rep} \quad {\rm cov} = L \cdot L^T$$

$$\langle \mathcal{O}[f(x,Q^2)] \rangle = \frac{1}{N_{\rm rep}} \sum_{k=1}^{N_{\rm rep}} \mathcal{O}[f^{(k)}(x,Q^2)]$$

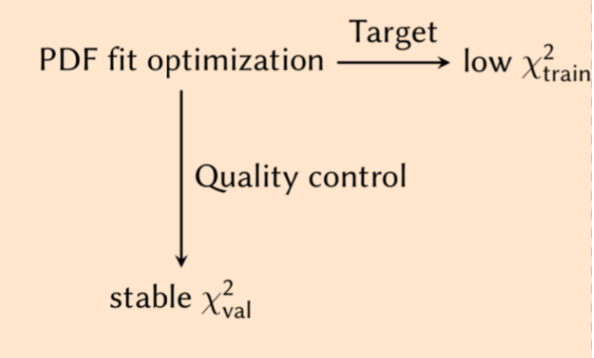
$$\sigma_{\mathcal{O}}[f(x,Q^2)] = \left[\frac{1}{N_{\rm rep}-1} \sum_{k=1}^{N_{\rm rep}} \left(\mathcal{O}[f^{(k)}(x,Q^2)] - \langle \mathcal{O}[f(x,Q^2)] \rangle \right)^2 \right]^{1/2}$$



Parametrisation



Optimisation

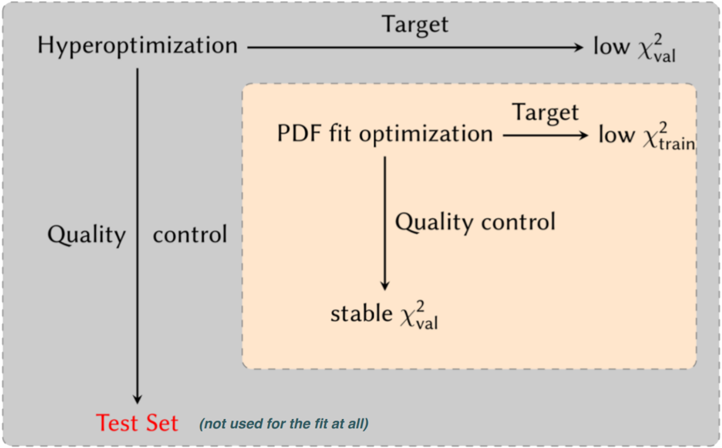


$$\chi^2 = \sum_{i,j}^{N_{\text{dat}}} (T_i[\{\boldsymbol{\theta}\}] - D_i) \left(\text{cov}^{-1}\right)_{ij} (T_j[\{\boldsymbol{\theta}\}] - D_j)$$

$$(\text{cov})_{ij} = \delta_{ij} s_i^2 + \left(\sum_{\alpha}^{N_c} \sigma_{i,\alpha}^{(c)} \sigma_{j,\alpha}^{(c)} + \sum_{\alpha}^{N_{\mathcal{L}}} \sigma_{i,\alpha}^{(\mathcal{L})} \sigma_{j,\alpha}^{(\mathcal{L})} \right) D_i D_j$$

stochastic gradient descent with backpropagation

Hyperoptimisation: fitting the methodology



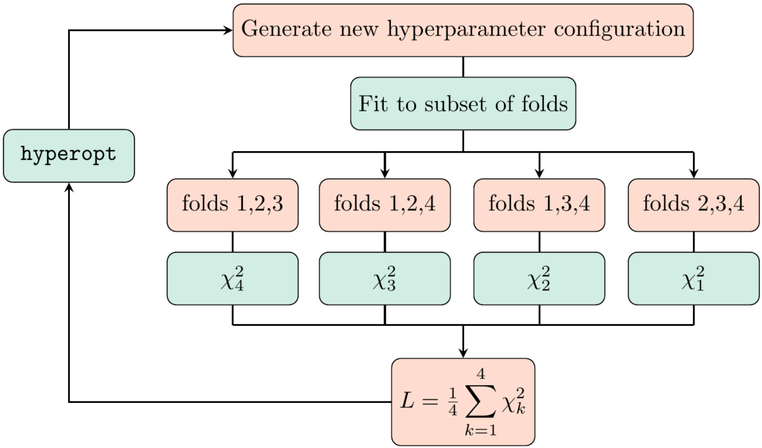
Compare to a Test Set (new set of data previously not used at all)

Who picks the Test Set? Automatic generalisation based on K foldings

Divide the data into n representative sets, fit $n-1$ sets and use the n-th set as test set

Hyperoptimise on mean and standard deviation of $\chi^2_{\text{test},i}$, $i = 1 \ldots n$

Hyperoptimisation: K-folding



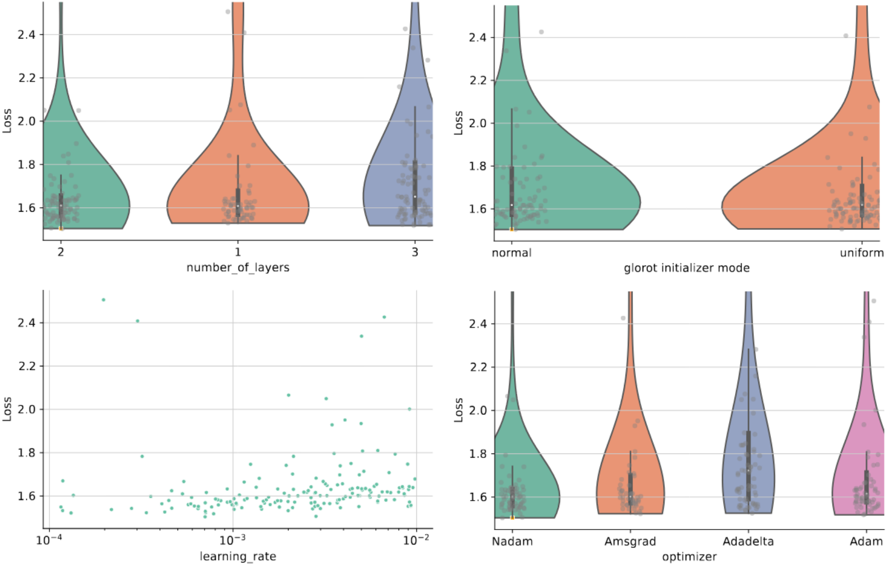
Compare to a Test Set (new set of data previously not used at all)

Who picks the Test Set? Automatic generalisation based on K foldings

Divide the data into n representative sets, fit $n-1$ sets and use the n-th set as test set

Hyperoptimise on mean and standard deviation of $\chi^2_{\mathrm{test},i}$, $i = 1\ldots n$

Hyperparameters

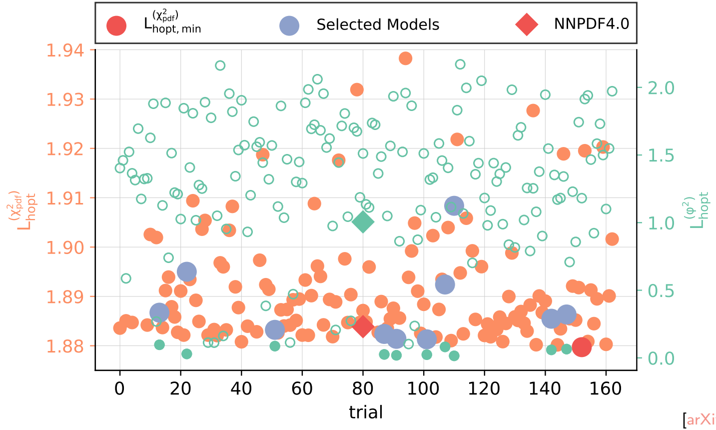


Hyperoptimisation: metrics

$$\varphi^2_{\chi^2} \equiv \left\langle \chi^2 \left[T^{(k)}, D^{(0)} \right] \right\rangle_{\rm rep} - \chi^2 \left[\langle T \rangle_{\rm rep}, D^{(0)} \right] = \frac{1}{n_{\rm dat}} \sum_{i,j}^{n_{\rm dat}} ({\rm cov})^{-1} T_{ij} \qquad L^{(\varphi^2)}_{\rm hopt} = \left(\frac{1}{n_{\rm fold}} \sum_{p}^{n_{\rm fold}} \varphi^{2\,(p)}_{\chi^2} \right)^{-1}$$

Select hyperparameters leading to best χ^2 and largest PDF errors in non-fitted data

Sample over the acceptable hyperparameters displaying comparable performance



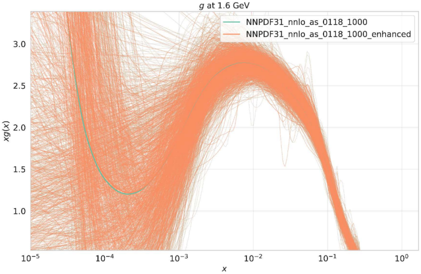
[arXiv:2410.16248]

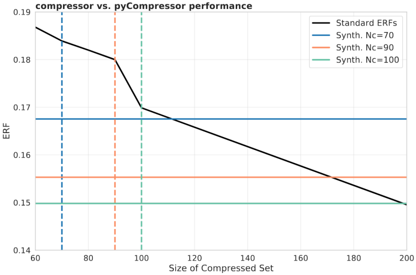
Delivery: compression

Find a subset of replicas that describe the underlying probability distribution as accurately as the original ensemble of replicas

$$\mathrm{ERF} = \frac{1}{N_{\mathrm{EST}}} \sum_k \frac{1}{N_k} \sum_i \left(\frac{C^{(k)}(x_i) - P^{(k)}(x_i)}{P^{(k)}(x_i)} \right)^2$$

Use GANs to enhance the number of replicas before minimising ERF





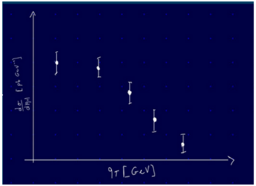
GAN enhancement allows one to achieve the same ERF with fewer replicas

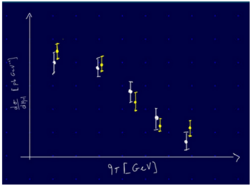
[EPJ C81 (2021) 530]

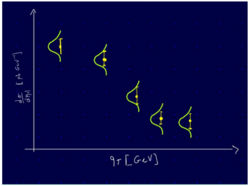
Validation: closure tests

Fit PDFs to pseudodata generated assuming a known underlying law

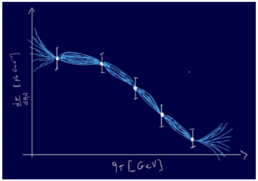
Level 0	Level 1	Level 2
no fluctuations	Gaussian fluctuation	Monte Carlo replicas



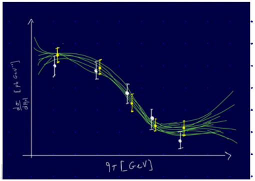




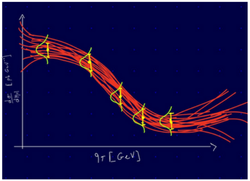
interpolation uncertainty



fuctional uncertainty



data uncertainty



Closure tests at work

Data region: closure tests

Fit PDFs to pseudodata generated assuming a known underlying law

Define bias and variance

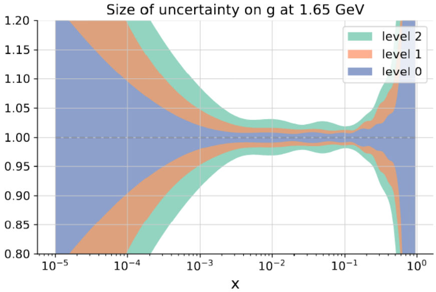
bias difference of central prediction and truth

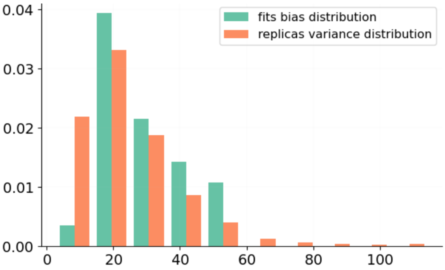
variance uncertainty of replica predictions

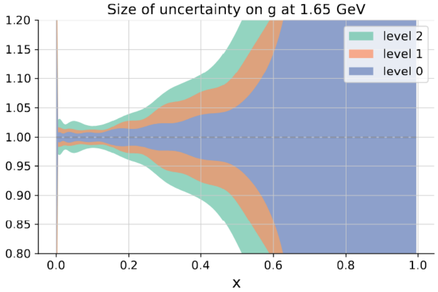
If PDF uncertainty faithful, then

E[bias] = variance

25 fits, 40 replicas each

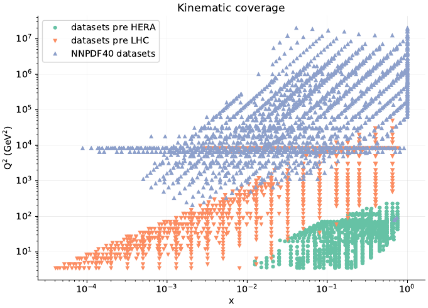


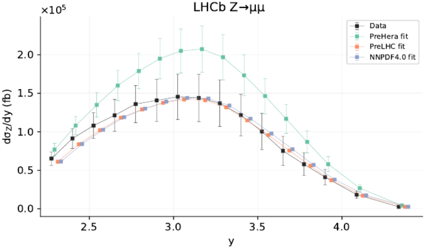




[EPJ C77 (2017) 663; EPJ C82 (2022) 330]

Future tests



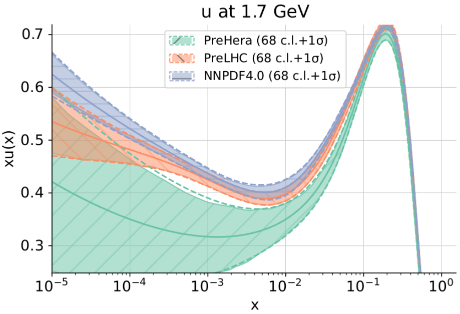


Extrapolation regions: future test

Test PDF uncertainties on data sets
not included in a given PDF fit
that cover unseen kinematic regions

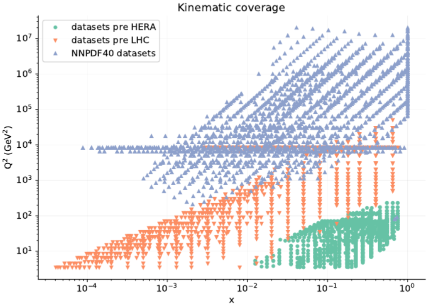
Data set	NNPDF4.0	pre-LHC	pre-HERA
pre-HERA	1.09	1.01	0.90
pre-LHC	1.21	1.20	23.1
NNPDF4.0	1.29	3.30	23.1

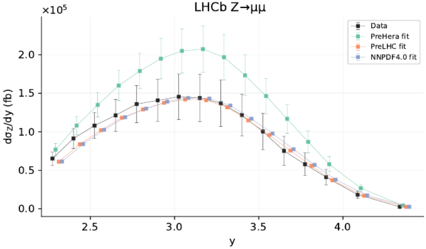
Only exp. cov. matrix



[Acta Phys.Polon. B52 (2021) 243]

Future tests



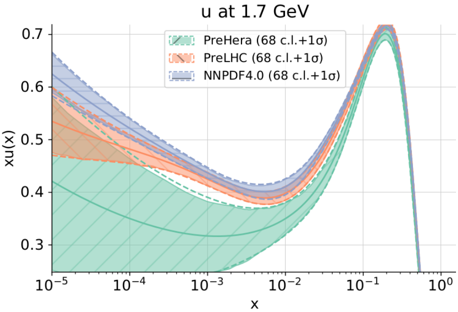


Extrapolation regions: future test

Test PDF uncertainties on data sets
not included in a given PDF fit
that cover unseen kinematic regions

Data set	NNPDF4.0	pre-LHC	pre-HERA
pre-HERA			0.86
pre-LHC		1.17	1.22
NNPDF4.0	1.12	1.30	1.38

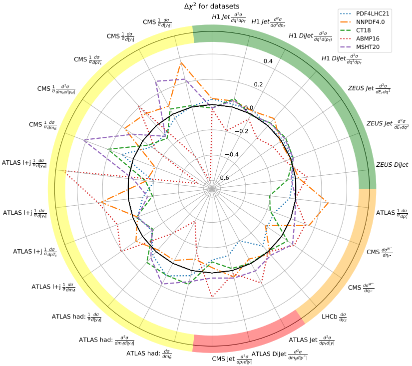
Exp+PDF cov. matrix

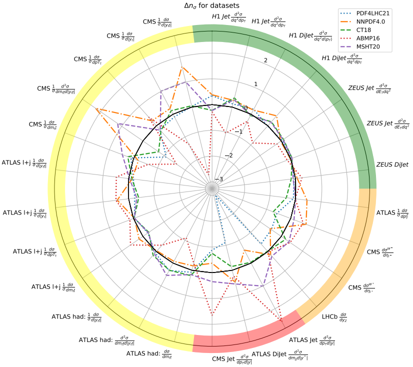


[Acta Phys.Polon. B52 (2021) 243]

To conclude: are all PDF sets equally accurate?

$$\Delta\chi^{2(i)} = \frac{\chi^{2(i)}_{\rm exp+th} - \left\langle \chi^2_{\rm exp+th} \right\rangle_{\rm pdfs}}{\left\langle \chi^2_{\rm exp+th} \right\rangle_{\rm pdfs}} \qquad \Delta n_\sigma^{(i)} = \frac{\chi^{2(i)}_{\rm exp+th} - \left\langle \chi^2_{\rm exp+th} \right\rangle_{\rm pdfs}}{\sqrt{2/n_{\rm data}}}$$

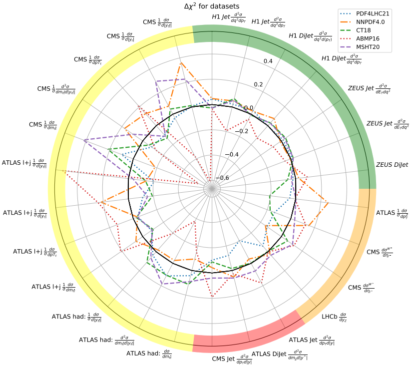


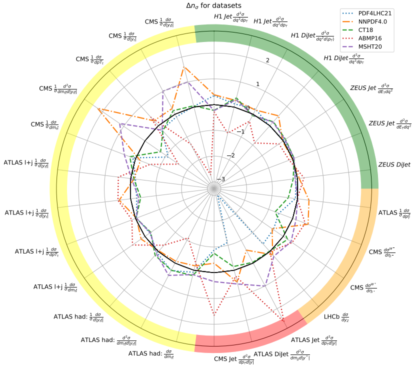


The PDF set obtained with ML is more precise and more accurate than all the others

To conclude: are all PDF sets equally accurate?

$$\Delta\chi^{2(i)} = \frac{\chi^{2(i)}_{\rm exp+th} - \left\langle \chi^2_{\rm exp+th} \right\rangle_{\rm pdfs}}{\left\langle \chi^2_{\rm exp+th} \right\rangle_{\rm pdfs}} \qquad \Delta n_\sigma^{(i)} = \frac{\chi^{2(i)}_{\rm exp+th} - \left\langle \chi^2_{\rm exp+th} \right\rangle_{\rm pdfs}}{\sqrt{2/n_{\rm data}}}$$



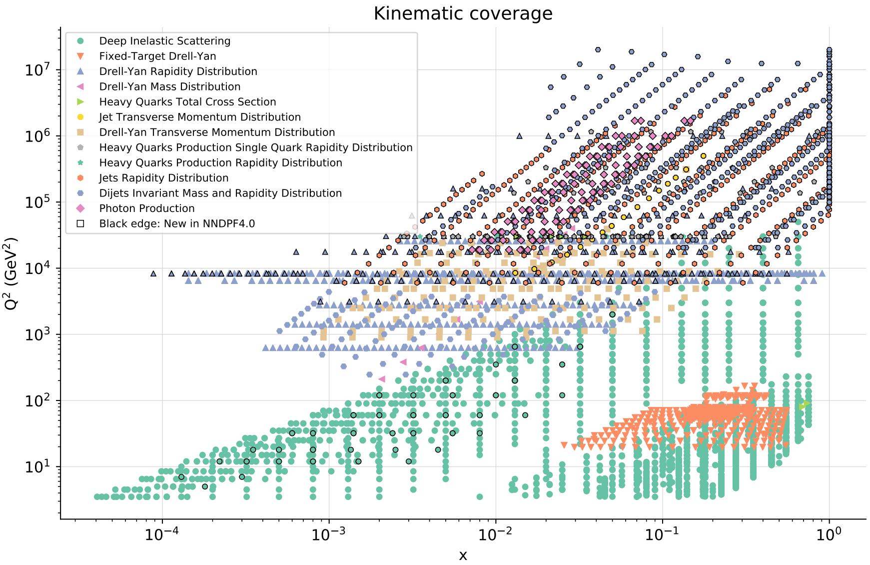


The PDF set obtained with ML is more precise and more accurate than all the others

Thank you

Extra material

Overview of experimental data

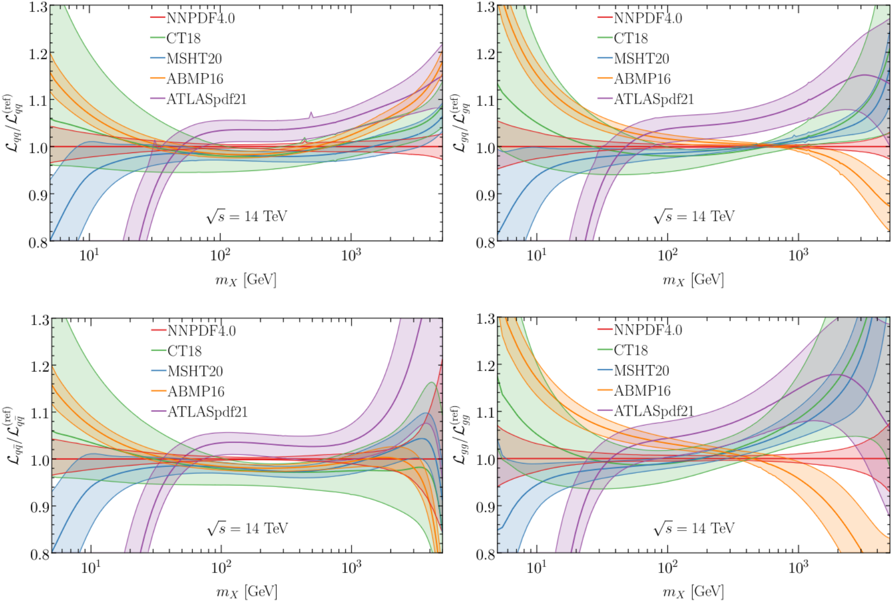


$$N_{\rm dat} = 4618$$

Overview of current PDF determinations

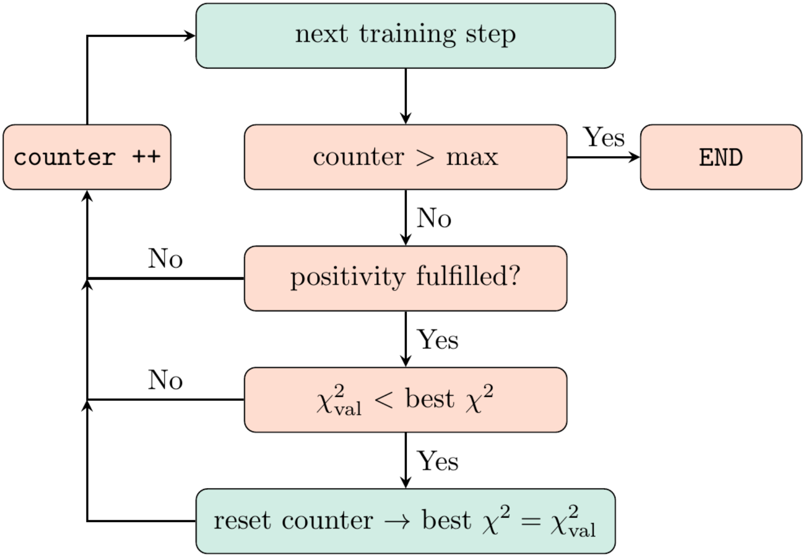
	NNPDF4.0	MSHT20	CT18	HERAPDF2.0	CJ22	ABMP16
Fixed-target DIS	✓	✓	✓	✗	✓	✓
JLAB	✗	✗	✗	✗	✓	✗
HERA I+II	✓	✓	✓	✓	✓	✓
HERA jets	✓	✗	✗	✓	✗	✗
Fixed target DY	✓	✓	✓	✗	✓	✓
Tevatron W, Z	✓	✓	✓	✗	✓	✓
LHC vector boson	✓	✓	✓	✗	✓	✓
LHC $W+c$ $Z+c$	✓	✗	✗	✗	✗	✗
Tevatron jets	✓	✓	✓	✗	✓	✗
LHC jets	✓	✓	✓	✗	✗	✗
LHC top	✓	✓	✗	✗	✗	✓
LHC single t	✓	✗	✗	✗	✗	✗
LHC prompt γ	✓	✗	✗	✗	✗	✗
statistical treatment	Monte Carlo	Hessian $\Delta\chi^2$ dynamical	Hessian $\Delta\chi^2$ dynamical	Hessian $\Delta\chi^2 = 1$	Hessian $\Delta\chi^2 = 1.645$	Hessian $\Delta\chi^2 = 1$
parametrisation	Neural Network	Chebyschev pol.	Bernstein pol.	polynomial	polynomial	polynomial
HQ scheme	FONLL	TR′	ACOT-χ	TR′	ACOT-χ	FFN
accuracy	aN^{3}LO	aN^{3}LO	NNLO	NNLO	NLO	NNLO
latest update	EPJ C82 (2022) 428	EPJ C81 (2021) 341	PRD 103 (2021) 014013	EPJ C82 (2022) 243	PRD 107 (2023) 113005	PRD 96 (2017) 014011

Comparing PDF sets



[Acta Phys.Polon.B 53 (2022) 12]

Optimisation



stochastic gradient descent with backpropagation