17 September 2021

Parton distrubution functions and Neural Networks

An introductory overview and the state of art

Lorenzo Mansi

Structure

Table of contents	
PDFs	definition, properties, why are so interesting.
NN	structure, training algorithms.
NNPDFs	data structure, NN structure and closure test for each version.

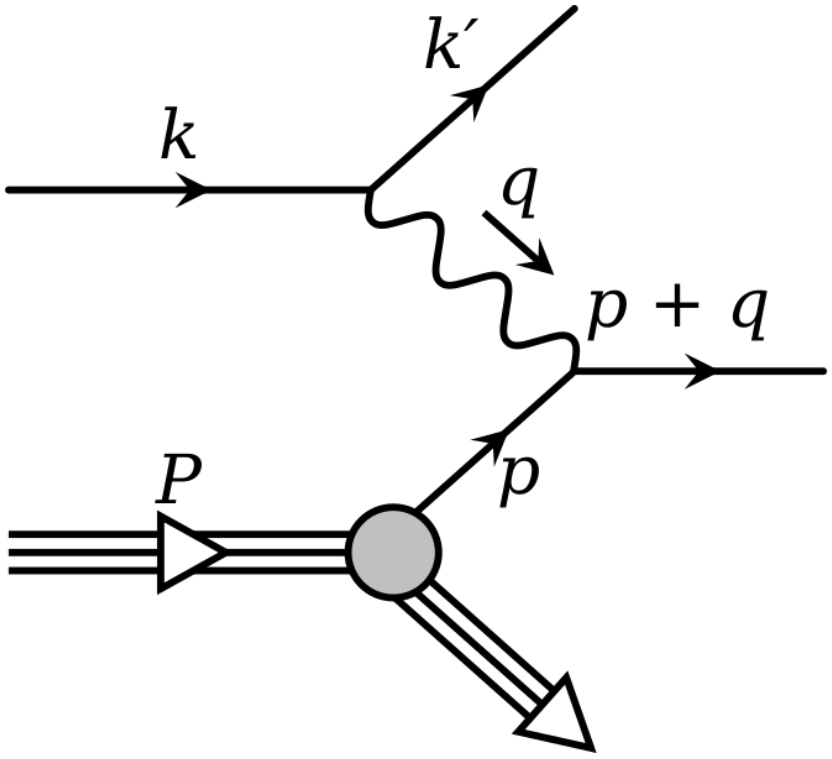
Probing protons

Colliding protons it was observed that produced hadrons had momenta almost collinear with the collision axis: the probability to produce a large transverse momentum hadron falls off exponentially with the transverse momentum.

Parton model

Proton is a loosely bound assemblage of small costituents called partons.

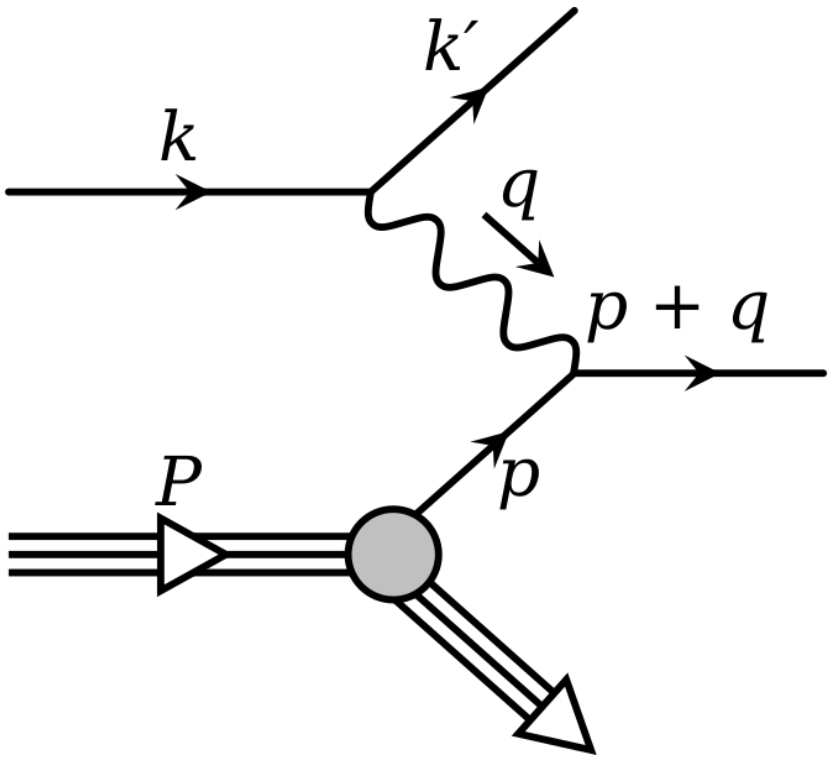
Deep inelastic scattering $ep \to eX$



Hat-variables are the Mandelstam variables for the scattering $e^- q \to e^- q$.

- ξ is the longitudinal fraction of momenta carried by the parton $p = \xi P$
- $\hat{t} = -Q^2 = q^2$
- $\hat{s} = \xi s$
- High energy $\to$ massless particles
- $x = \frac{Q^2}{2P \cdot q}$ and at LO $\xi = x$

Deep inelastic scattering



The unpolarized cross section for the process is:

$$\frac{d^2\sigma}{dxdQ^2} = \sum_i f_i(x) Q_i^2 \cdot \frac{2\pi\alpha^2}{Q^4}\left[1 + \left(1 - \frac{Q^2}{xs}\right)^2\right]$$

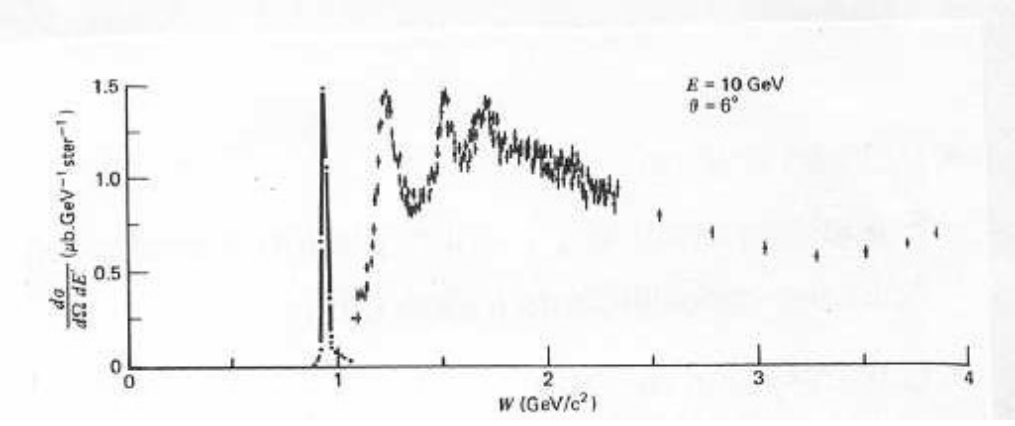
PDFs

The $f_i(\xi)$ are the parton distribution function, at leading order they are the probability to find a parton with longitudinal momentum fraction ξ.

BUT WHY DOES THIS MODEL WORK SO WELL?

Measurement of DIS cross-section

First done at SLAC in 1970s
(Friedman, Kendall & Taylor: Nobel Prize 1990)



Peaks are from proton (elastic) and baryonic resonances

Figure 1: W is the mass of the outgoing hadron X produced, slides taken by proton structure lecture of Edinburgh University.

Renormalizing the theory

The answer lies in the asymptotic safety of QCD, the Renormalization Group Equation for the running couple α_s is:

$$\frac{\partial \alpha_s}{\partial \ln Q^2} = \beta(\alpha_s) = -\left(\frac{\alpha_s}{4\pi}\right)^2 \sum_{n=0} \beta_n \left(\frac{\alpha_s}{4\pi}\right)^n$$

The exact solution is known only at leading order with $\beta_0 = 11 - \frac{2}{3} n_f$ where n_f is the number of active flavor at the Q^2 scale.

$$\alpha_s(Q^2) = \frac{4\pi}{\beta_0 \log\left(Q^2/\Lambda^2\right)} \quad \textit{conventionally } \Lambda \approx \ 200\text{MeV}$$

DGLAP

PDFs emerge from soft QCD processes which cannot be described in a perturbative framework so a non analytical method to obtain them is to fit data which comes from different experiments carried at different energy scales.

Including contribution from collinear gluons and quark emission in PDFs makes them no longer independent from the scale Q (violation of Bjorken scaling), they evolve with scale according to DGLAP:

$$\frac{\partial f_a}{\partial \ln Q^2} = \frac{\alpha_s(Q^2)}{2\pi} \sum_b P_{ab} \otimes f_b \quad with \quad a \otimes b = \int_x^1 dy \frac{a(y)}{y} b(\frac{x}{y})$$

Evolution basis

To solve DGLAP is better to work in the Evolution Basis where the system is minimally coupled rather than in Flavor basis, defined $q^{\pm} = q \pm \bar{q}$, for the proton PDF we will need at most a charm PDF so our evolution basis will be:

$$g \qquad \Sigma = \sum_i q^+ \quad V = \sum_i q^- \quad V_3 = u^- - d^-$$
$$V_8 = u^- + d^- - 2s^- \quad T_3 = u^+ - d^+ \quad V_8 = u^+ + d^+ - 2s^+$$

Neural Network

A neural network basically aims to mimic the behaviour of biological neurons.

Neurons

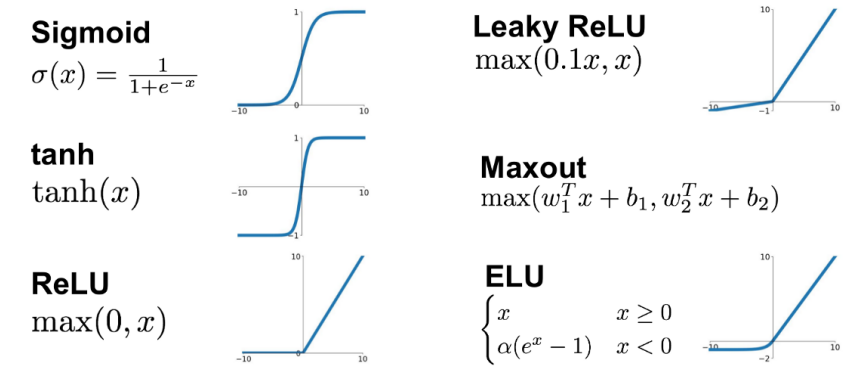
Neurons can be either active or inactive if a certain threshold "θ" is reached or not.

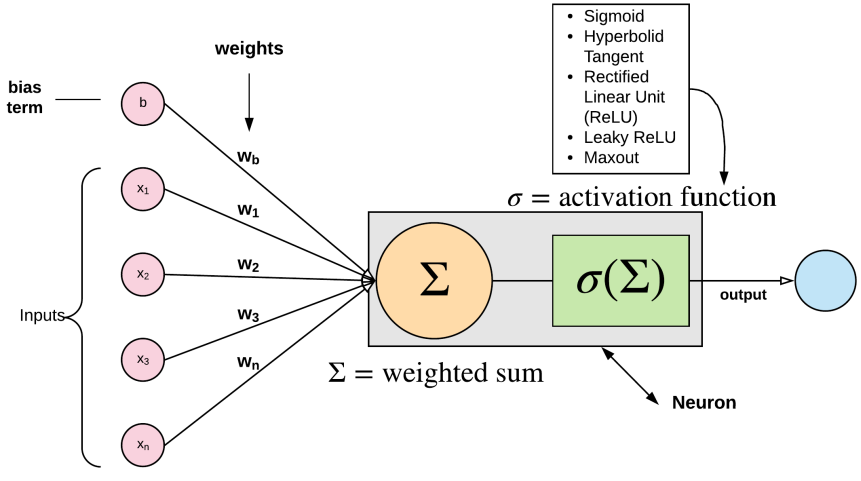
Node

A node is a neuron which is fed by other neurons' output each weighted suitably.

Activation functions

The activation function of a node defines the output of that node given an input or set of inputs, namely $x = \sum_i w_i x_i + bias$.



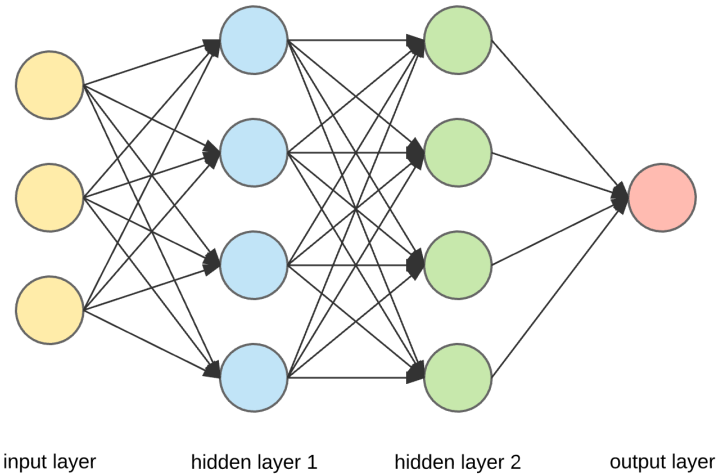


Training a NN

To perform a specific action the weights have to be chosen suitably. In order to do so we split our data in two sets.

Training set

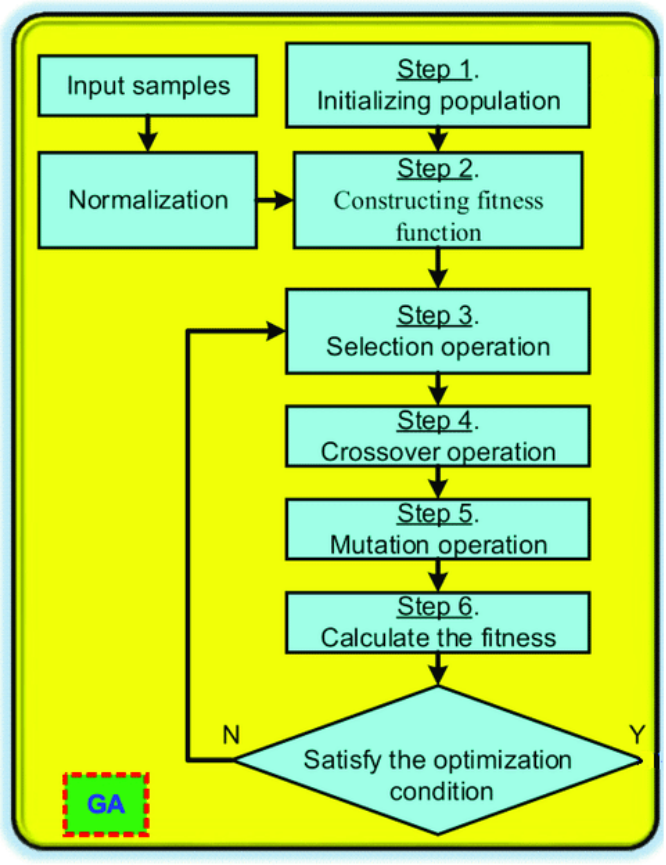
A portion of the data set, which the network is actually trained on, adjusting weights in order to minimize a loss function χ^2.



Validation set

A portion of data set that does not undergo the fitting procedure, but its χ^2 is monitored during the training procedure to avoid overfitting.

Genetic Algorithms GA

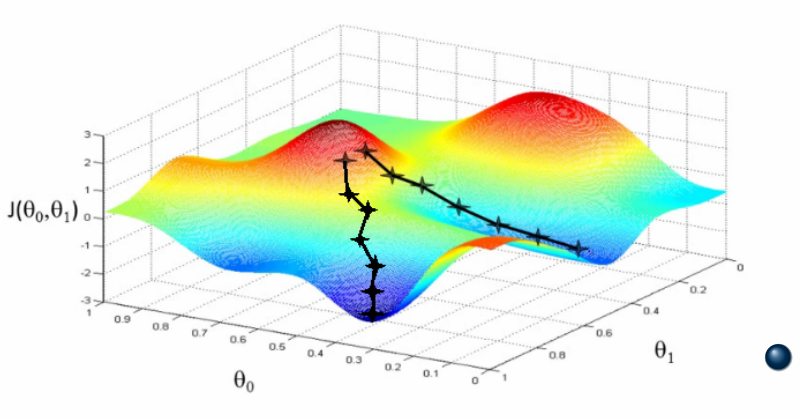


No analytic prior knowedge needed, low computational power but long time needed.

- Choose the number of mutants N_m and generations N_g.
- At each generation vary mutants' weight with a certain probability law.
- Go on until χ^2_{opt} or N_g is reached.

Gradient Descent GD

Need to compute derivatives, high computational power but short time needed.



- Given a starting point x_0 and a function $y(x)$, we move from the starting point of a quantity proportional to the gradient of the function in that point:
  $\Delta x = -a\nabla y|_{x_0} \rightarrow \Delta y = -a\left(\nabla y|_{x_0}\right)^2$
- Choose $a > 0$ to move down the function.
- Choose an error over path.

GOD BLESS TensorFlow

NNPDF

PDF determination is a pattern recognition problem which only less than 20 years ago has been recognized as solvable with AI techniques.

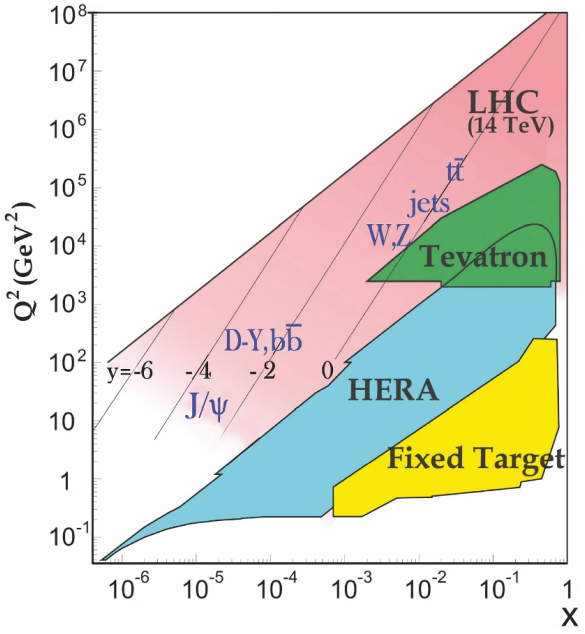
Neural networks were involved because they made possible to determine a function without fixing a predetermined model.

Several aspects have to be taken into account, but first we need to understand the form of our data.

Data Structure

A wide range of experiments has been performed at various Q^2 scale.

A generic observable, for example the cross section σ_X, in the two hadrons colliding case has the form:



10.1103/PhysRevD.98.030001

$$\sigma_X(s, M_X^2) = \sum_{a,b} \int_{x_{min}}^{1} dx_1 dx_2 f_{a/h_1}(x_1, M_X^2) f_{b/h_2}(x_2, M_X^2) \hat{\sigma}_{ab\to X}(x_1 x_2 s, M_X^2)$$

Traditional approach

- Fixed functional form postulated, for instance $f_i = x^{\alpha_i}(1-x)^{\beta_i}$
- From DGLAP expected $\ln x$ for $x \to 0$ and $\ln 1 - x$ for $x \to 1$ corrections.
- Fit error smaller than standard error propagation on bestfit values by one order of magnitude.
- More data cloud result in a more extended parametrization and an increase in uncertainties.

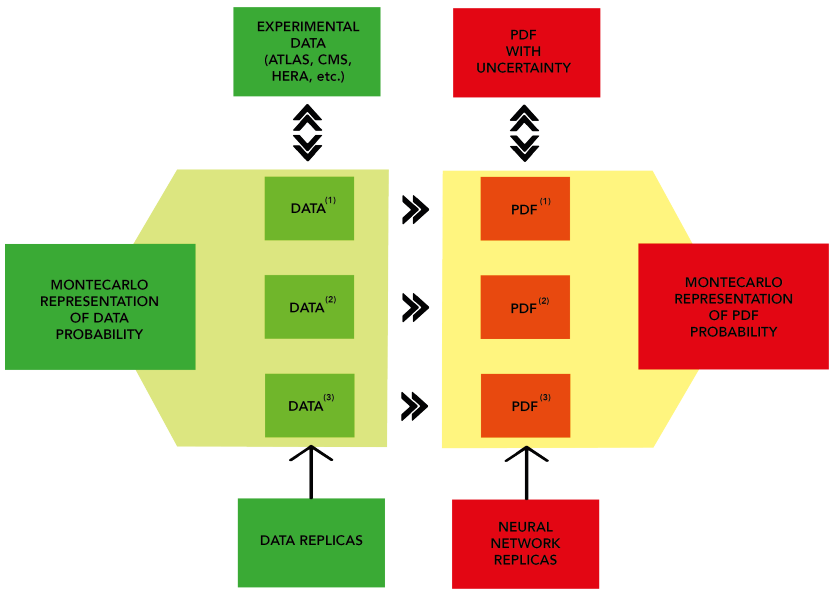
PDF problem peculiarities

- PDFs are probabilities distributions of observables. The goal of PDF determination is to determine the probability distribution of a probability distribution of observables.

- Full knowledge of PDF correlations is needed, from DGLAP each uncertainty on $f_i(x, Q_0^2)$ is correlated to each $f_j = (x', Q_0^2)$, so the determination of a covariance matrix of uncertainties in the space of probability distributions is needed.

NNPDF approach

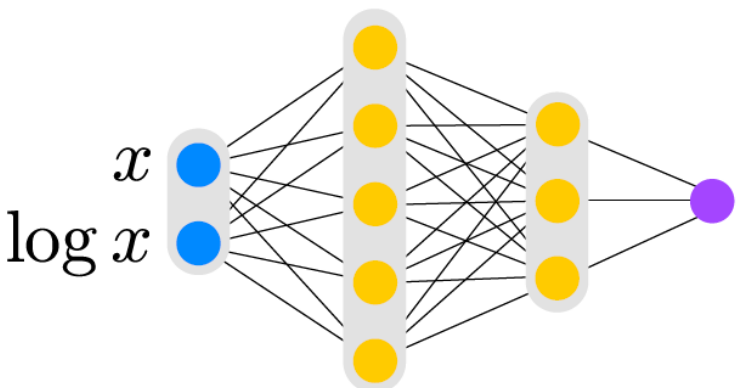
- Turn the input probability distribution in a Monte Carlo representation.
- For each replica extract a PDF by minimization of a certain χ^2.
- NNs with x as input and PDF as output are used, training and validation sets are slip randomly from data.
- Do statistics over PDF replicas.



2008.12305

NN structure

The structure of the NN has chosen to be fixed up to NNPDF 3.1 in a 2-5-3-1 scheme. The number of net used depended upon the considered PDF.



2008.12305

- NNPDF 1.0, 5 nets $(u, \bar{u}, d, \bar{d}, g)$
- NNPDF 1.1, 7 nets $(u, \bar{u}, d, \bar{d}, s, \bar{s}, g)$
- NNPDF 3.1, 8 nets $(u, \bar{u}, d, \bar{d}, s, \bar{s}, c^+, g)$

Parametrization

Each PDF (or PDF combination) is parametrized as:

$$xf_i(x, Q_0^2) = A_i x^{-\alpha_i+1}(1-x)^{\beta_i} NN_i(x)$$

The preprocessing exponents α_i and β_i:

- were chosen to be fixed NNPDF1.0 .
- were extracted under a uniform distribution for each replica in a fixed range and kept fixed during the fit NNPDF1.2 .
- NNPDF2.0 the range is determined computing correlation between the various χ^2 and requiring it remains small.
- NNPDF3.0 range determined self-consistently, 68% C.L. range determined, fit repeated in a range twice the 68% C.L. one.

Minimization Procedure

The loss function is defined as:

$$\chi^2 = \sum_{i,j}^{N_{dat}} (D - P)_i \sigma_{ij}^{-1} (D - P)_j$$

- P_i is the convolution between the PDF model and the Fast Kernel table
- σ_{ij} is thee covariance matrix between data points i and j.
- t_0 method used to handle multiplicative uncertanties only from NNPDF2.0. [0912.2276]

Genetic Minimization

- Weights initialized using random gaussian distribution.
- Copy the initial network and mutate the weights.
- Then keep the one with the lowest χ^2 and repeat.

Weights changed with the law:

$$w_i \to w_i + \eta_i r_i \qquad \eta_i = \frac{\eta_i^{(0)}}{N_{ite}^p}$$

NNPDF1.0, 2 mutations, $N_{mut} = 120$, $p = 1/3$, $N_{max} = 5000$

NNPDF2.0, 2 epoch, $N_{mut}^1 = 80$,change at $N_{ite} = 2500$, $N_{mut}^2 = 10$, p random in $[0, 1]$, $N_{max} = 30000$

NNPDF2.3, 3 mutations

Nodal mutation

Starting from NNPDF3.0 a GA based on nodal mutation was adopted:

- each network is assigned a probability of being mutated
- if a node is selected all its weights change according to the previous equations
- $\eta = 15$ fixed and mutation probability of 15%, values chose after closure test.

Stopping Criterion

- NNPDF1.0 χ^2_{val} stops improving $\rightarrow$ does not require to reach the N_{max} of generations.

- NNPDF3.0 *look* $-$ *back* method: the absolute minimum of χ^2_{val} within the a given number of iteration is stored $\rightarrow$ need to reach the N_{max} of generations.

Proton as seen by Higgs

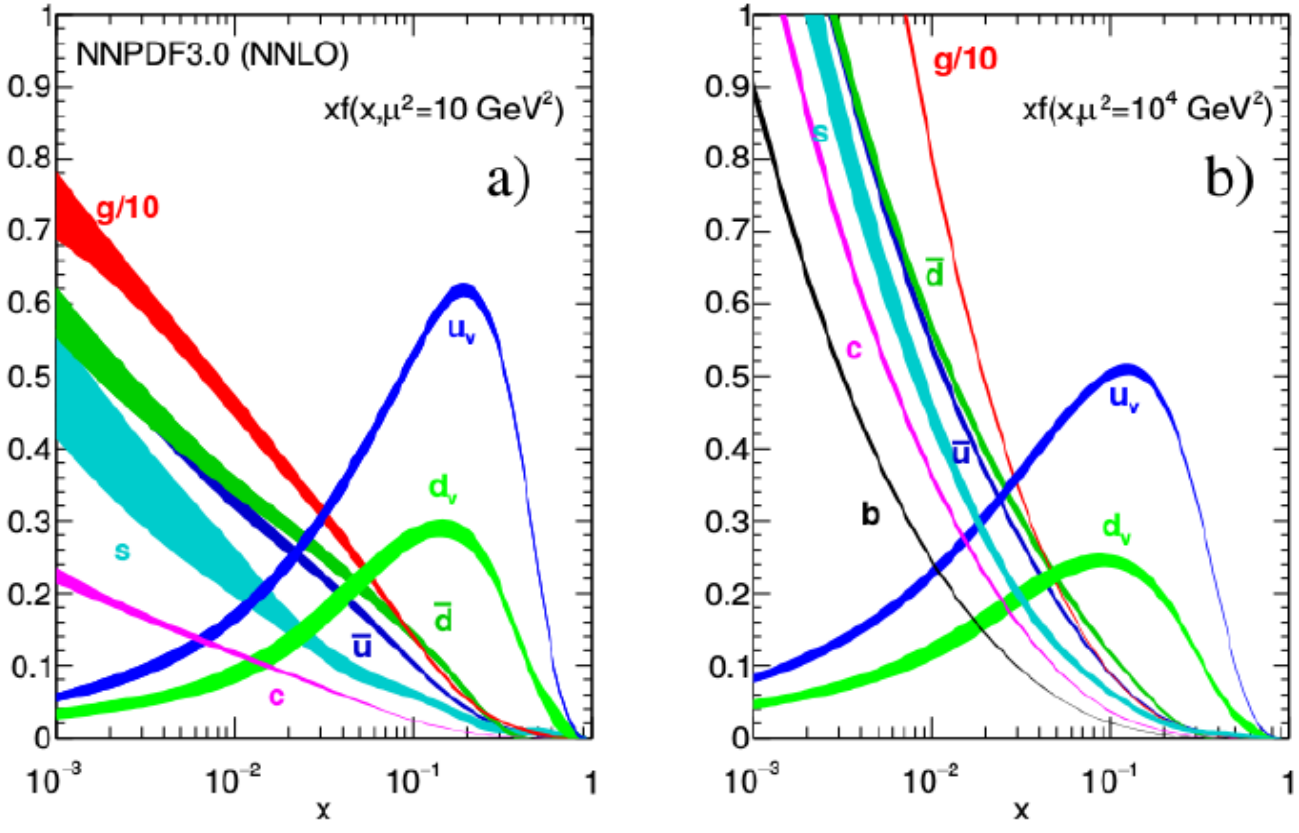


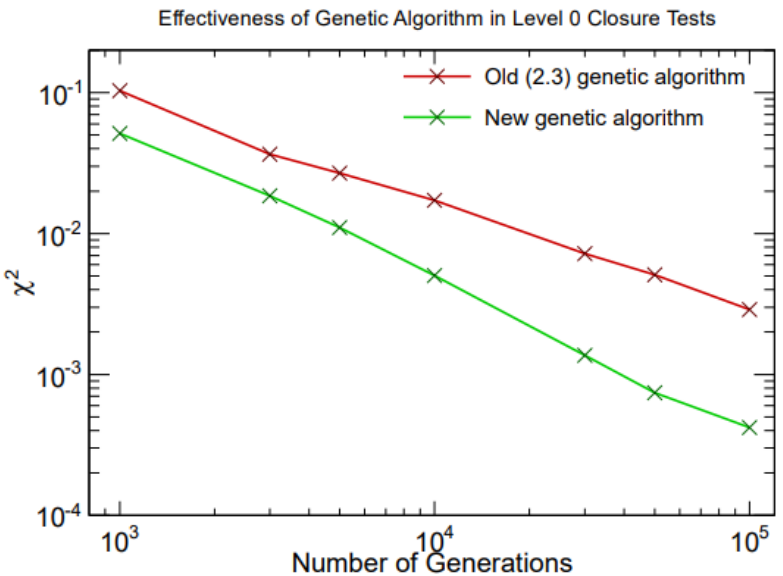
Figure 2: From PDG 2018, PDF obtained via NNPDF3.0

Closure Tests

How can we check if PDF uncertainties are faithful? Via closure tests.

- Level 0, data generated with no uncertanties, a perfect χ^2 is in principle possible.

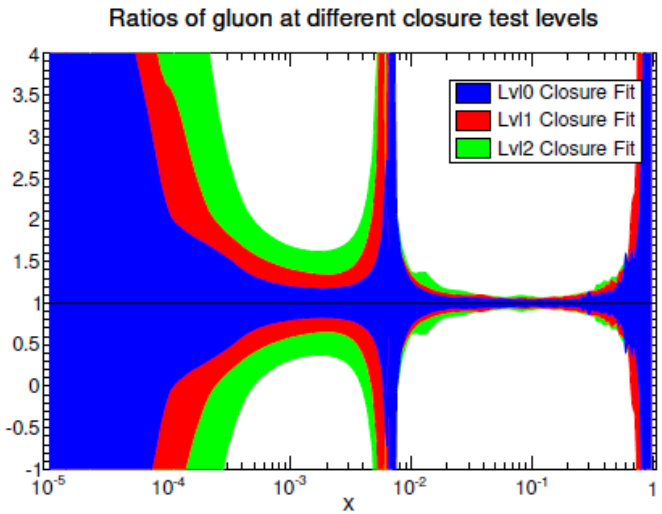
2008.12305

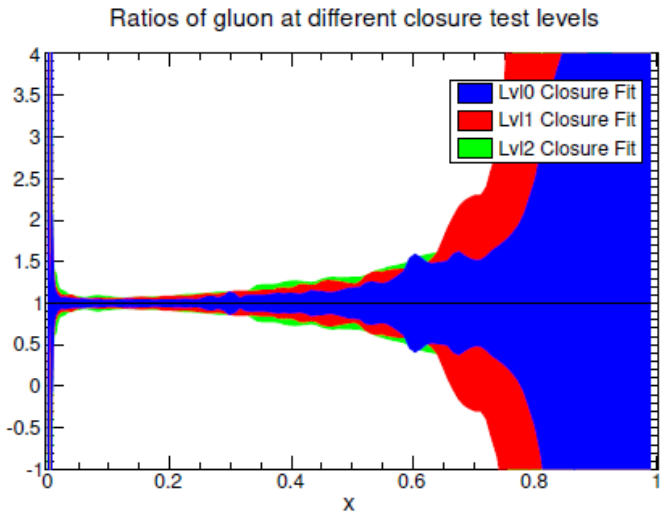


Closure Tests

- Level 1, data generated assuming the probability distribution which corresponds to the published covariance matrix.
- Level 2, taking Level 1 data as if they were actual experimental data and then apply the NNPDF methodology.

2008.12305





Closure Tests

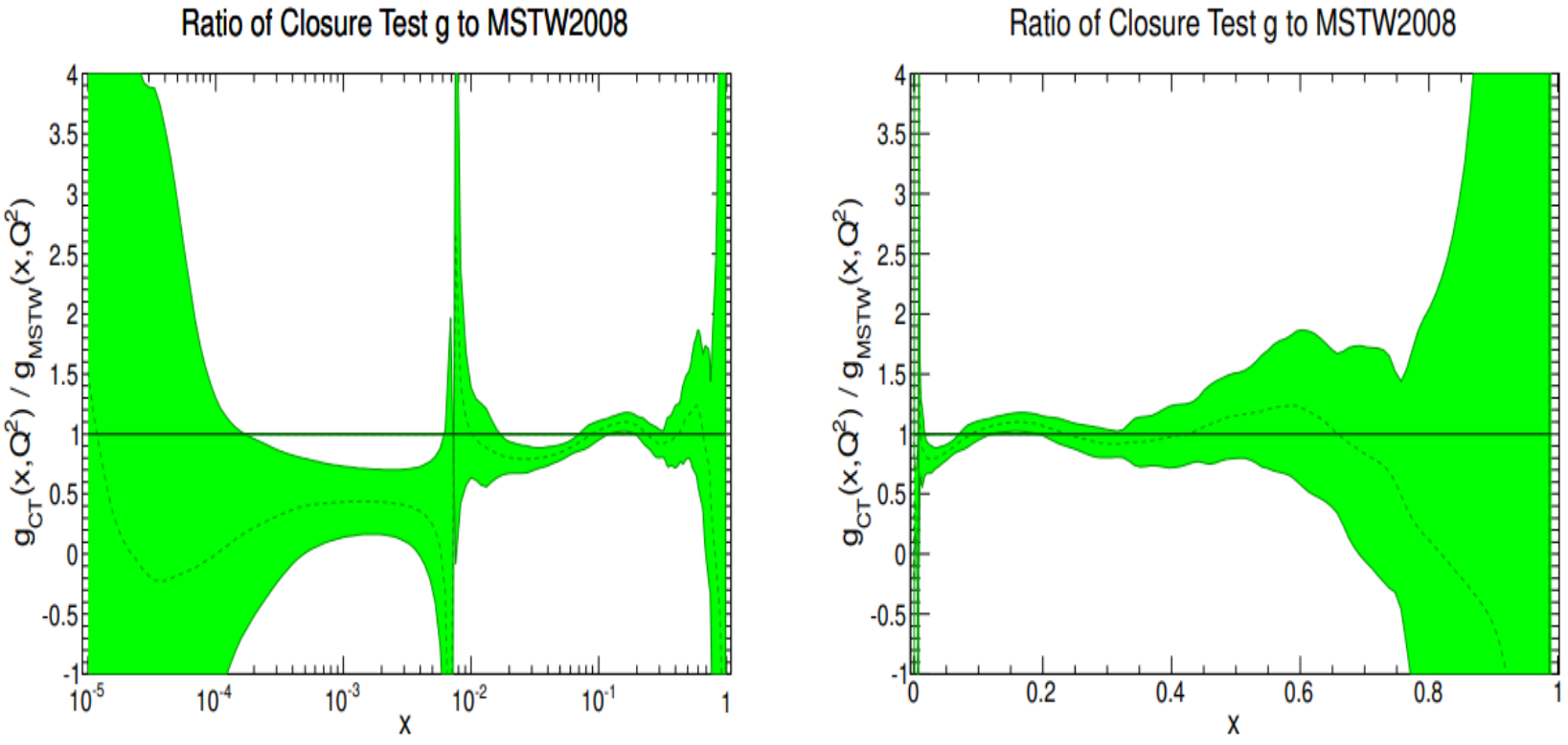


Figure 3: The green band is the one-σ uncertainty (Level 2 data test) and the result is shown as a ratio to the underlying truth. 2008.12305

Closure Tests

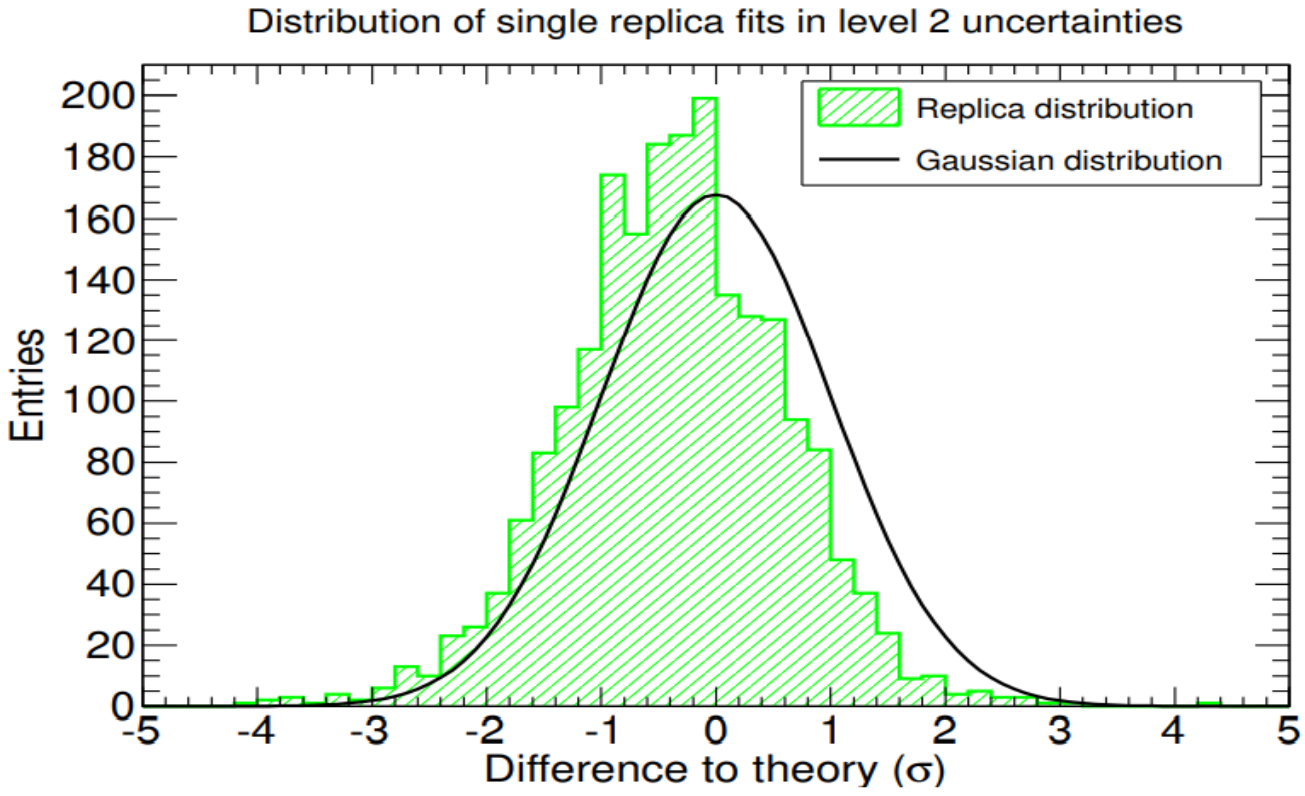


Figure 4: Distribution of deviation between the PDF and the underlying truth normalized to its nominal uncertainty, compared to a univariate Gaussian. 2008.12305

A new approach

How do we improve everything made so far?

- Hyperoptimization to avoid architecture-dependent local minima.
- Use of Gradient Descent methods to reduce the computing cost thanks to the state of art the arte tools.
- We use only one densely connected network with 8 outputs, one per PDF.

n3fit

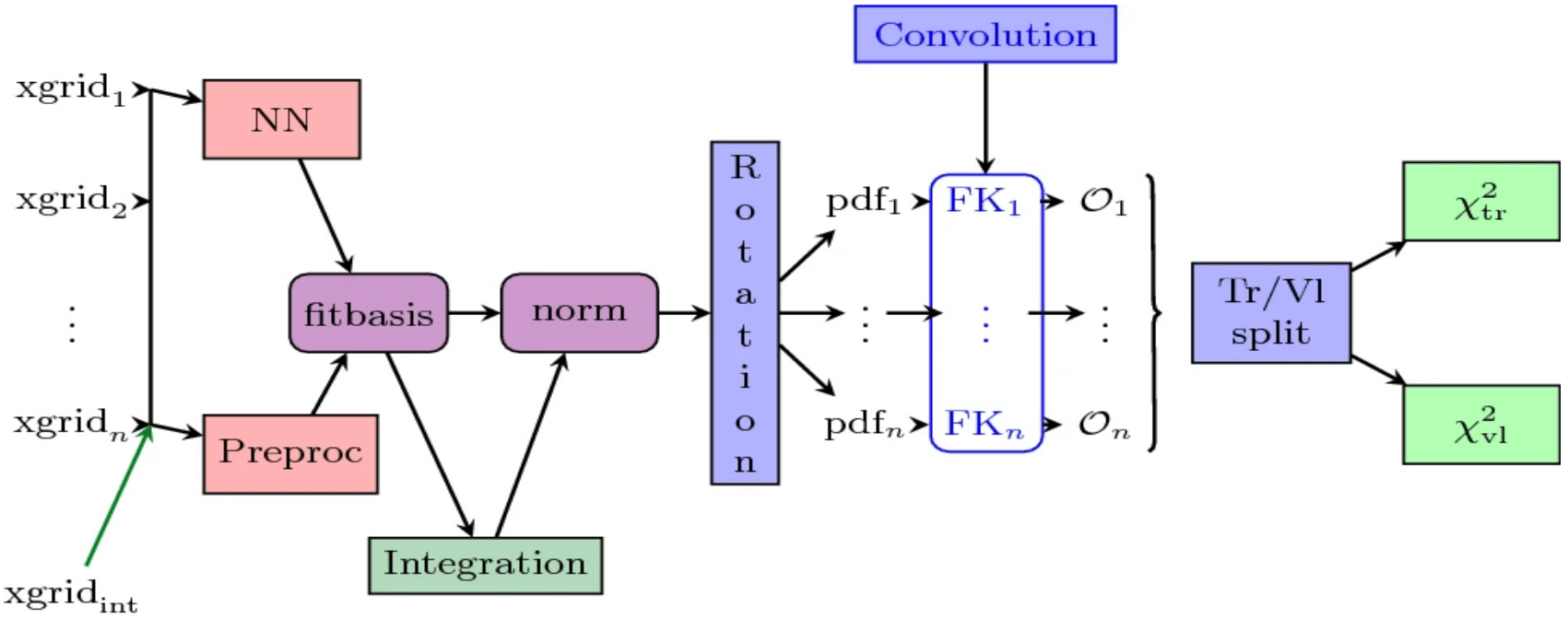


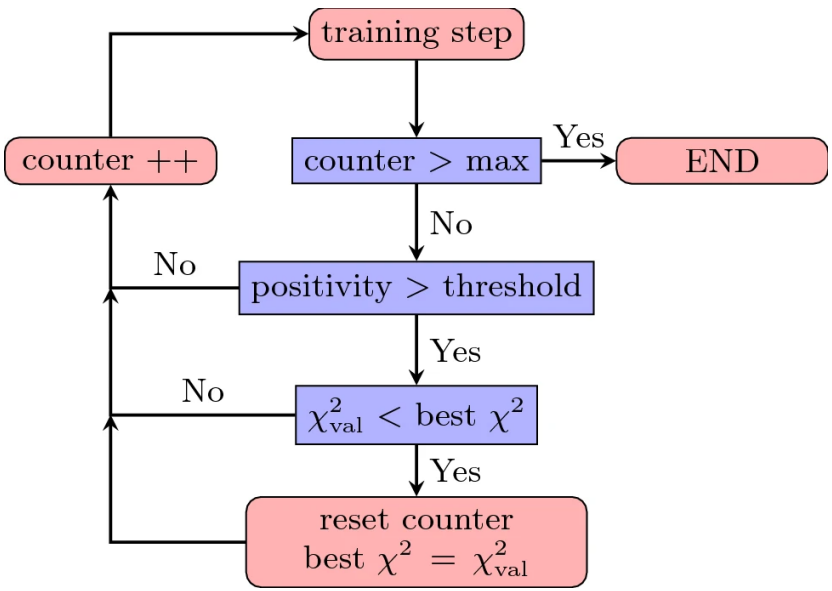
Figure 5: Diagrammatic view of the n3fit code,each block corresponds to a layer.The square red blocks contain all the fittable parameters. 1907.05075

n3fit

- xgrid$_k$ contanis the x data from the k-th experiment entering the fit.
- In the Prepoc the exponents α_i and β_i are now free to vary during the fit.
- Rotation takes PDF from evolution basis to flavor basis
- Convolution first create a rank-4 tensor: $\mathcal{L}_{i\alpha j\beta} = f_{i\alpha} f_{j\beta}$
- Then contracts it with FK Tables: $\mathcal{O}^n = FK^n_{i\alpha j\beta} \mathcal{L}_{i\alpha j\beta}$

Latin letters stand for flavor index, greek letters refer to the number of xgrid.

Stopping criterion



Patience algorithm scheme.
1907.05075

- Before updating the parameters the χ^2 is monitored for both training and validation sets.
- We train the net until χ^2_{val} stops improving.
- A patience algorithms is enabled to avoid false positives.
- Positivity means that the PDF must produce positive prediction for a special set of data.

Hyperopt

Table 2. Parameters on which the hyperparameter scan is performed from.[42]

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

Figure 6: Figures taken from 2008.12305

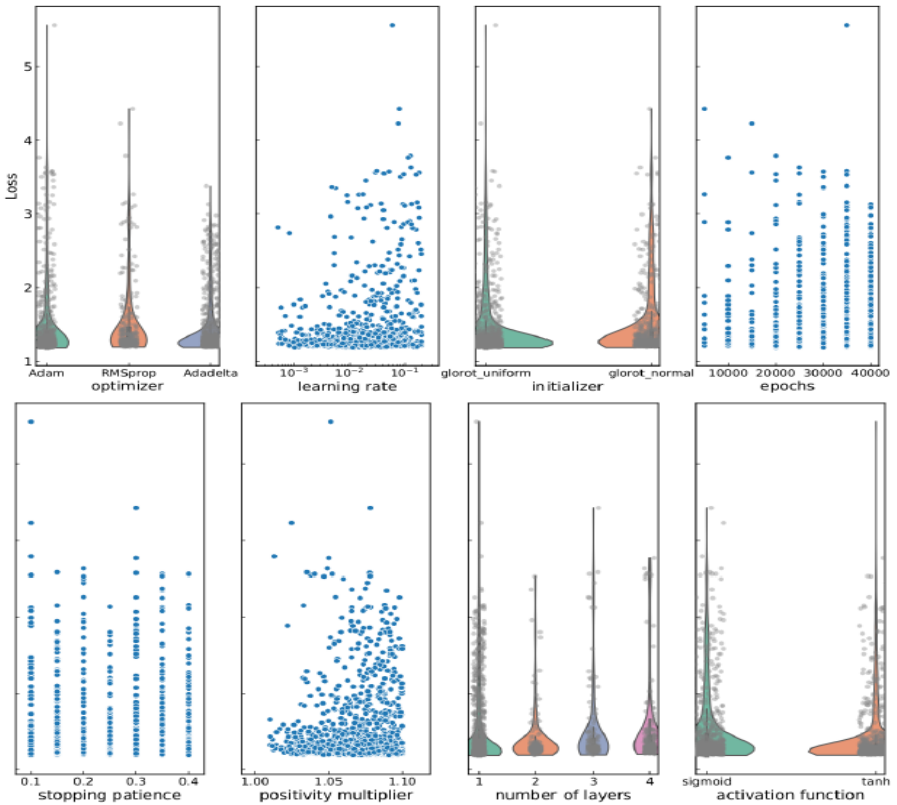


Fig. 10. Graphical representation of a hyperparameter scan for a DIS only fit with 2000 trials (from Ref.[42]). The loss function presented in the y-axis is an average of the validation and testing χ^2. The shape of the violin plots represent a visual aid on the behavior of the fit as a function of the free parameter. Fatter plots represent better stability, i.e., configurations which are less likely to produce outliers.

Benchmark

Table 1. Comparison of the average computing resources consumed by the old and new methodologies for the DIS and Global setups.

DIS fit	CPU h.	Mem. Usage (GB)	Good replicas
`n3fit` (new)	0.2	2	95%
`nnfit` (old)	4	4	70%
Global fit	CPU h.	Mem. Usage (GB)	Good replicas
`n3fit` (new)	1.5	4	95%
`nnfit` (old)	30	5	70%

Parameter	DIS only	Global
Hidden layers	2	3
Architecture	35-25-8	50-35-25-8
Activation	tanh	sigmoid
Initializer	glorot_normal	glorot_normal
Dropout	0.0	0.006
Optimizer	Adadelta	Adadelta
Max epochs	40000	50000
Stopping patience	30%	30 %

Table 4. Best models found by our hyparparameter scan for the DIS and global setups using the new `n3fit` methodology.

Figures taken from 1907.05075

Quality control

A procedure similar to cross validation is needed to ensure faithful results from hyperoptimization.

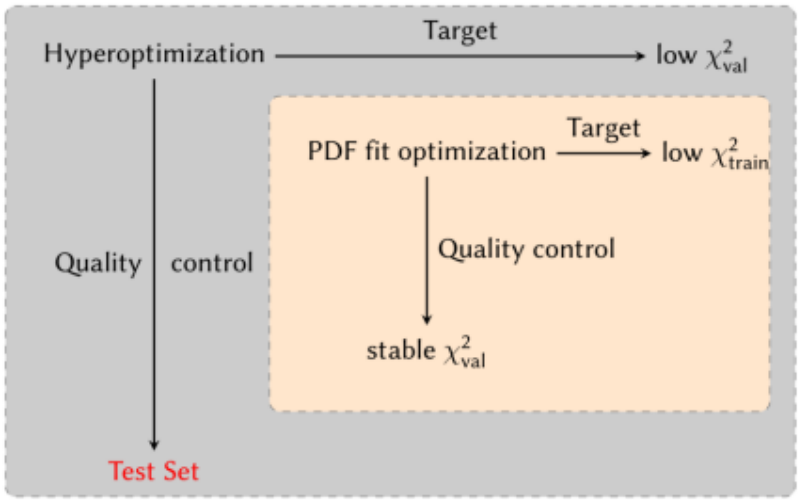


Figure 7: 2008.12305

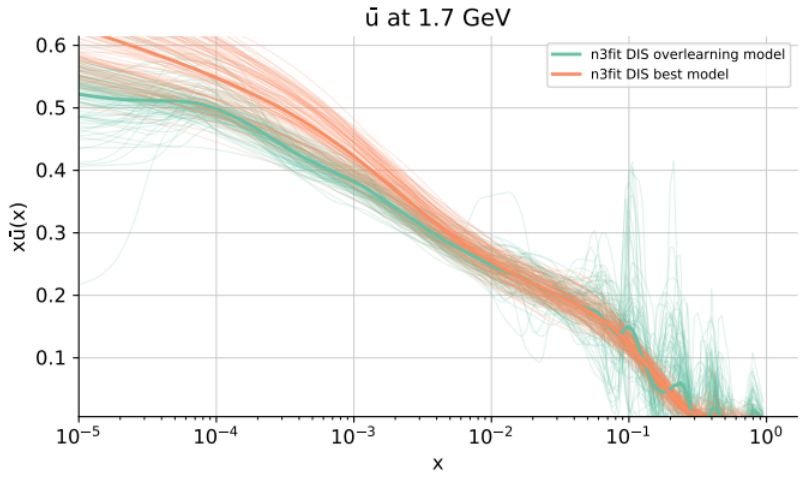


Figure 8: PDF for $\bar{u}$ obtained via n3fit with quality control (orange) and without (green). 2008.12305

Quality control

- Although data set is only fitted and hyperopt is performed on χ^2_{val} there is a correlation between this two.
- A new Test Set is needed which is neither fitted or hyperscanned to be sure to avoid overtraining
- Test Set is constructed by picking the smallest kinematic range experiments for the same process. And excluding this data from the n3fit.
- Check the test loss χ^2_{test}.

Results

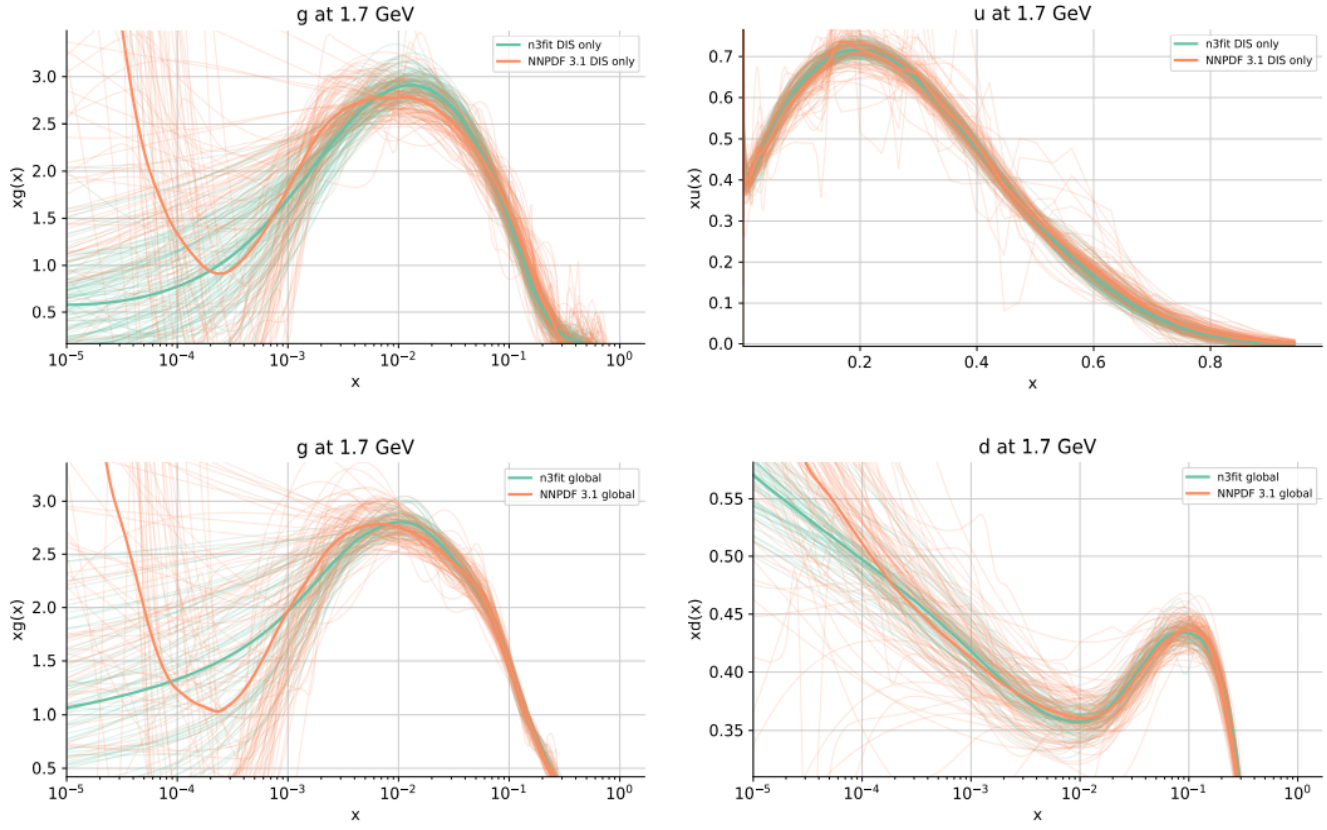


Figure 9: comparison of methodology, DIS fit is shown at top while the down one is the global fit. 2008.12305

PDF arc-length

The stability of the new methodology is clear by looking at the arc-lenght of the PDF at a given scale, in Figure 10 the mean value of arc-length and the one σ band computed by a set of PDF replicas.

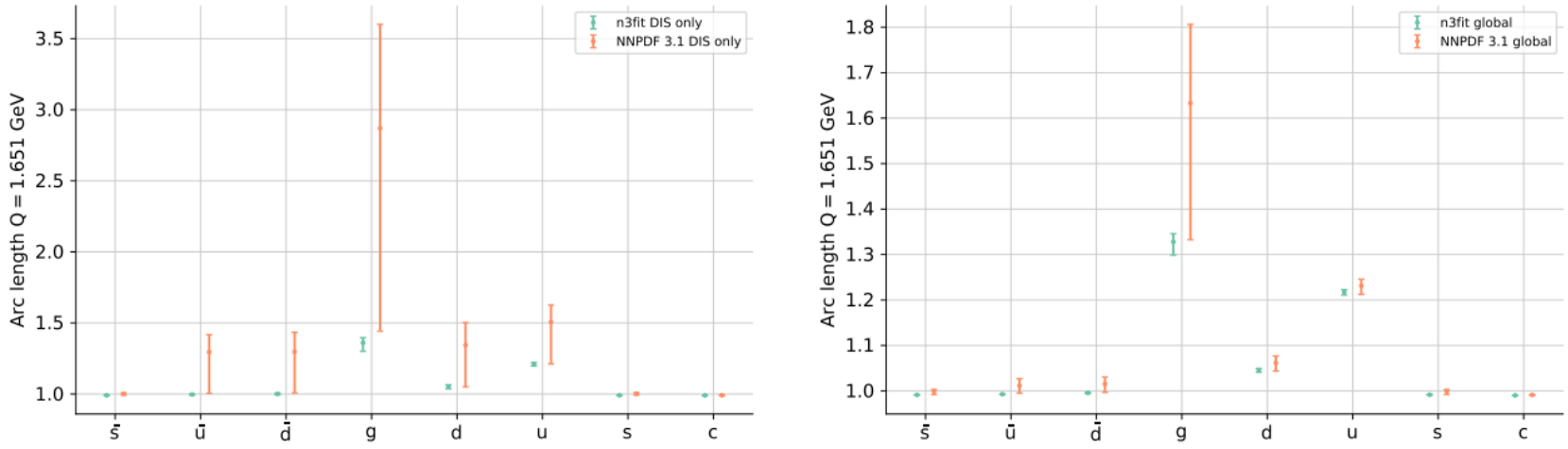
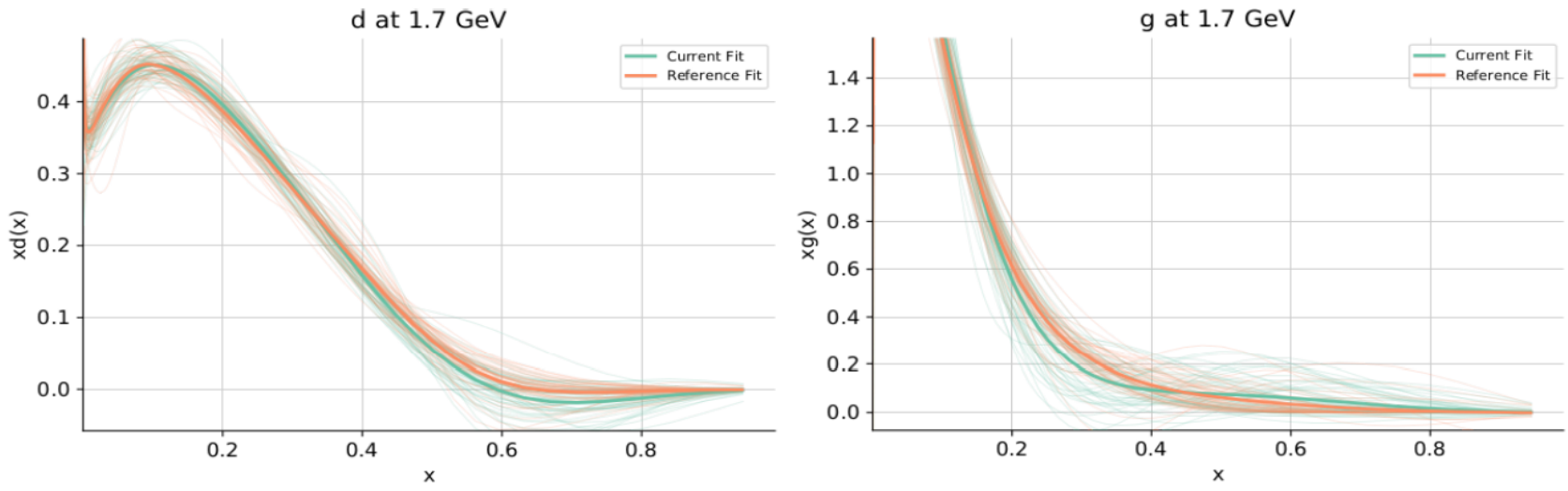


Figure 10: PDF arc-length comparison, DIS fit on the left, global fit on the right. 2008.12305

Test set stability

The choice of an appropriate test set is important and it stands parallel with the need to have as much data as possible in the training set. 2008.12305 in green the fit with the k-fold, in orange without.



k-folding method: spilt the data in *k*-partitions, then each plays in turn the role of testing set, hyperparametrs optimization on the mean value of the loss over excluded partitions.

Closure Test

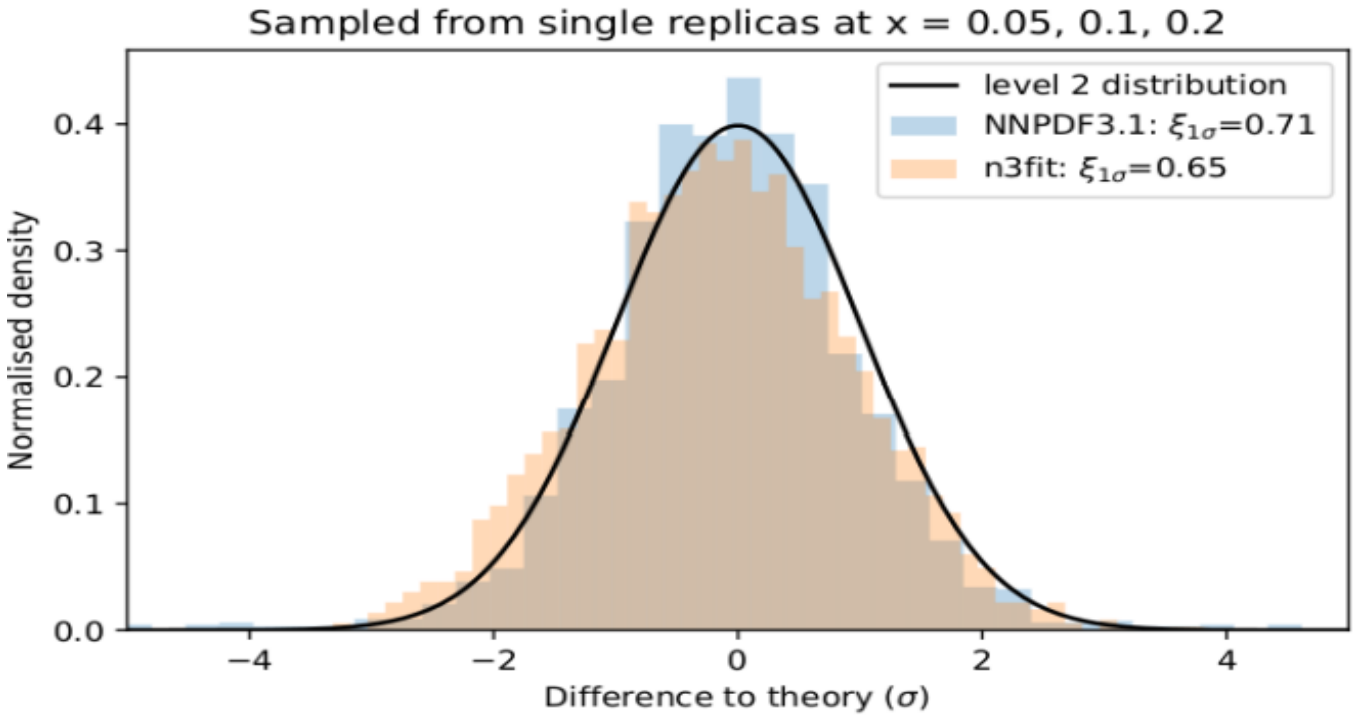


Figure 11: Distribution of deviation between the PDF and the underlying truth normalized to its nominal uncertainty, compared to a univariate Gaussian. Comparison between the old approximated methodology and the new exact one. 2008.12305

Closure Test

The new n3fit made possible to perform further tests as the bias-variance ratio.

$$R_{bv} = \sqrt{\frac{1}{N_{dat}} \sum_{i=1}^{N_{dat}} \frac{(d_i - d_i^{(0)})^2}{\sigma_i^2}}$$

PDF	R_{bv}	one-σ c.l.
Σ	0.9	70%
gluon	0.9	69%
V	1.0	66%
V3	1.0	93%
V8	0.9	71%
T3	0.6	89%
T8	1.3	46%
total	0.9	0.71

Figure 12: The R_{bv} and the one σ level for individual PDFs, computed using four points in x space per PDF along eigenvectors of covariance matrix. 2008.12305

Future Tests

Takes old data and then try to predict the current data with n3fit.

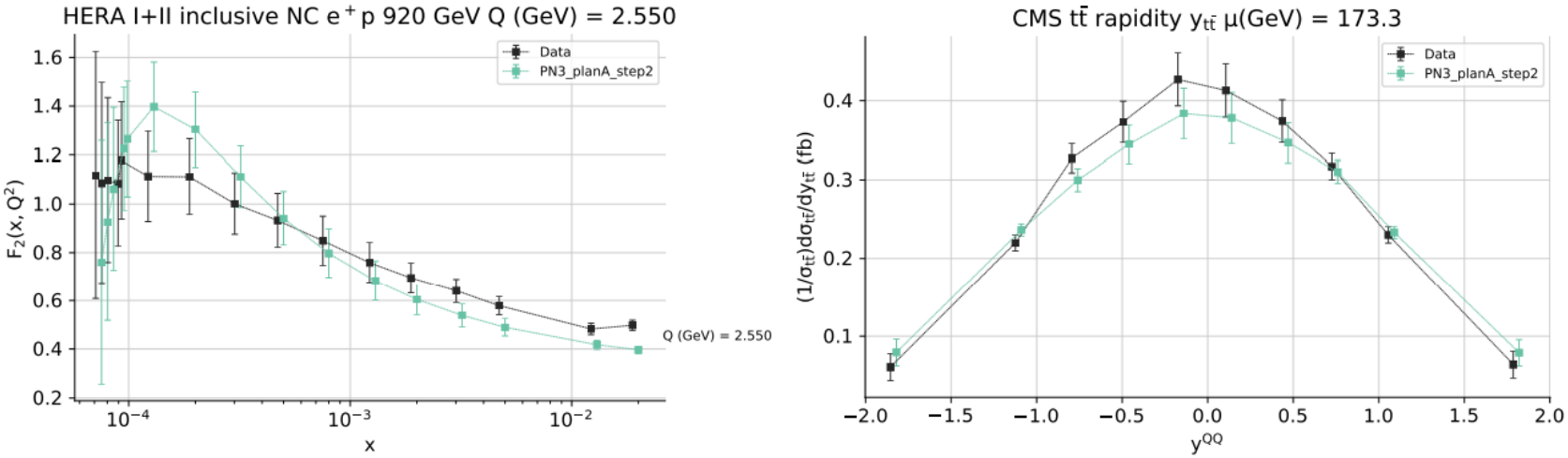


Figure 13: Data for the proton structure function f_2 measured at HERA (left) and top-pair production measured at the LHC (right) compared to a prediction based on PDFs determined from a fit to pre-HERA data.

The end!

Thanks for your attention!