

# Hyperparameter optimisation of neural networks for proton structure analyses

*Based on [2410.16248](#) with NNPDF and eScience center*

**Roy Stegeman**

The University of Edinburgh

*EPS-HEP, Marseille, 7 July 2025*



# Parton Distribution Functions

**Factorization: the LHC master formula**

$$\sigma = \sum_{ij} f_i \otimes f_j \otimes \hat{\sigma}_{ij}$$

Measurable quantity      PDFs      pQCD

Parton Distribution Functions (PDFs) are:

- Roughly speaking the probability of sampling a parton (quark or gluon) from a proton

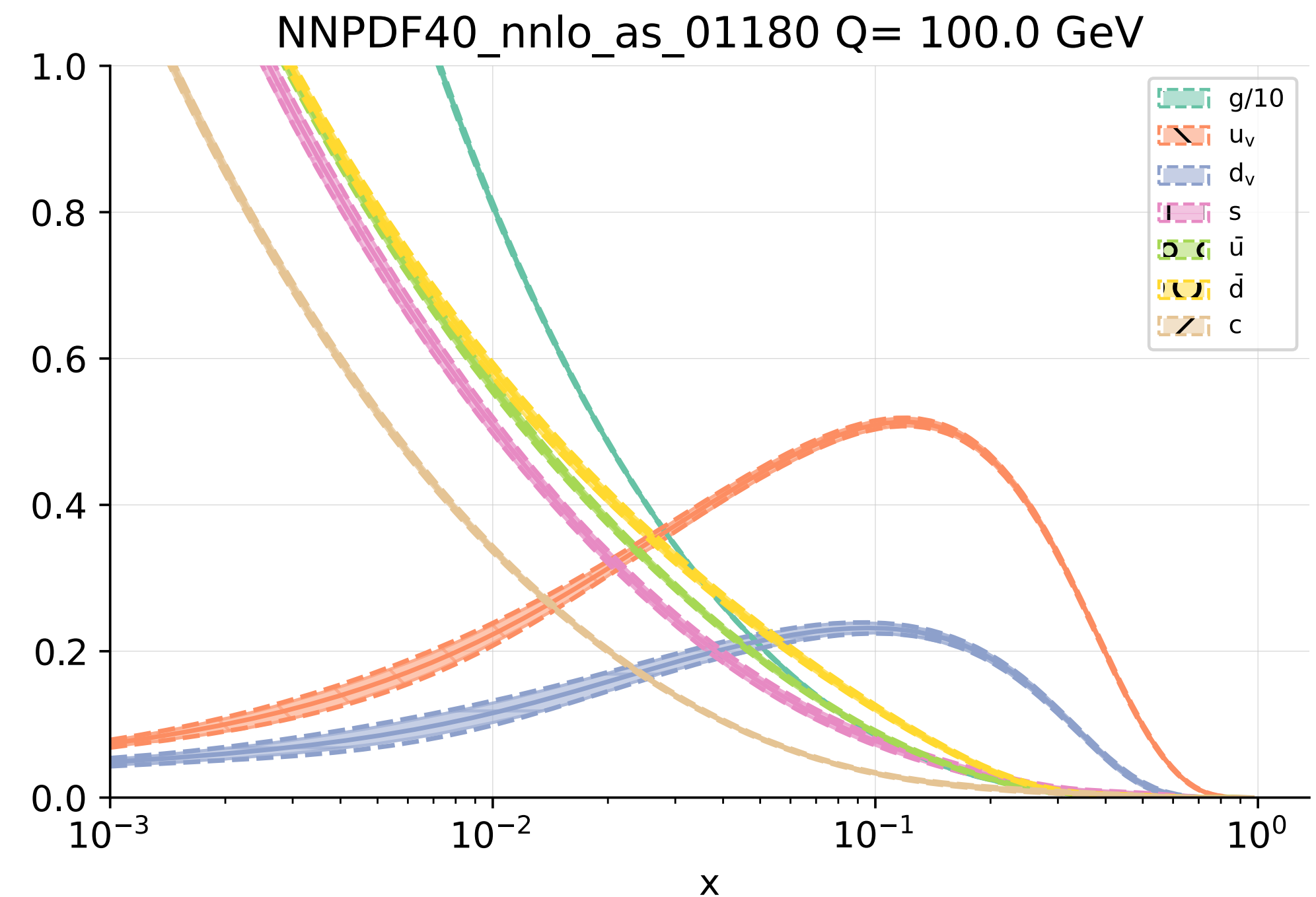
# Parton Distribution Functions

Factorization: the LHC master formula

$$\underbrace{\sigma}_{\text{Measurable quantity}} = \sum_{ij} \underbrace{f_i \otimes f_j}_{\text{PDFs}} \otimes \underbrace{\hat{\sigma}_{ij}}_{\text{pQCD}}$$

Parton Distribution Functions (PDFs) are:

- Roughly speaking the probability of sampling a parton (quark or gluon) from a proton
- 2 variables:
  - Longitudinal momentum fraction of the parton  
 $x = p_{\text{parton}}/p_{\text{proton}}$
  - Energy scale  $Q$
- $Q$  scaling is known from perturbative QCD (DGLAP)
- **ML problem:** find  $f_i(x)$  at a fixed scale  $Q_0$



- ▶ **The NNPDF methodology**
- ▶ Hyperoptimisation
- ▶ Validation of the methodology

# The experimental dataset

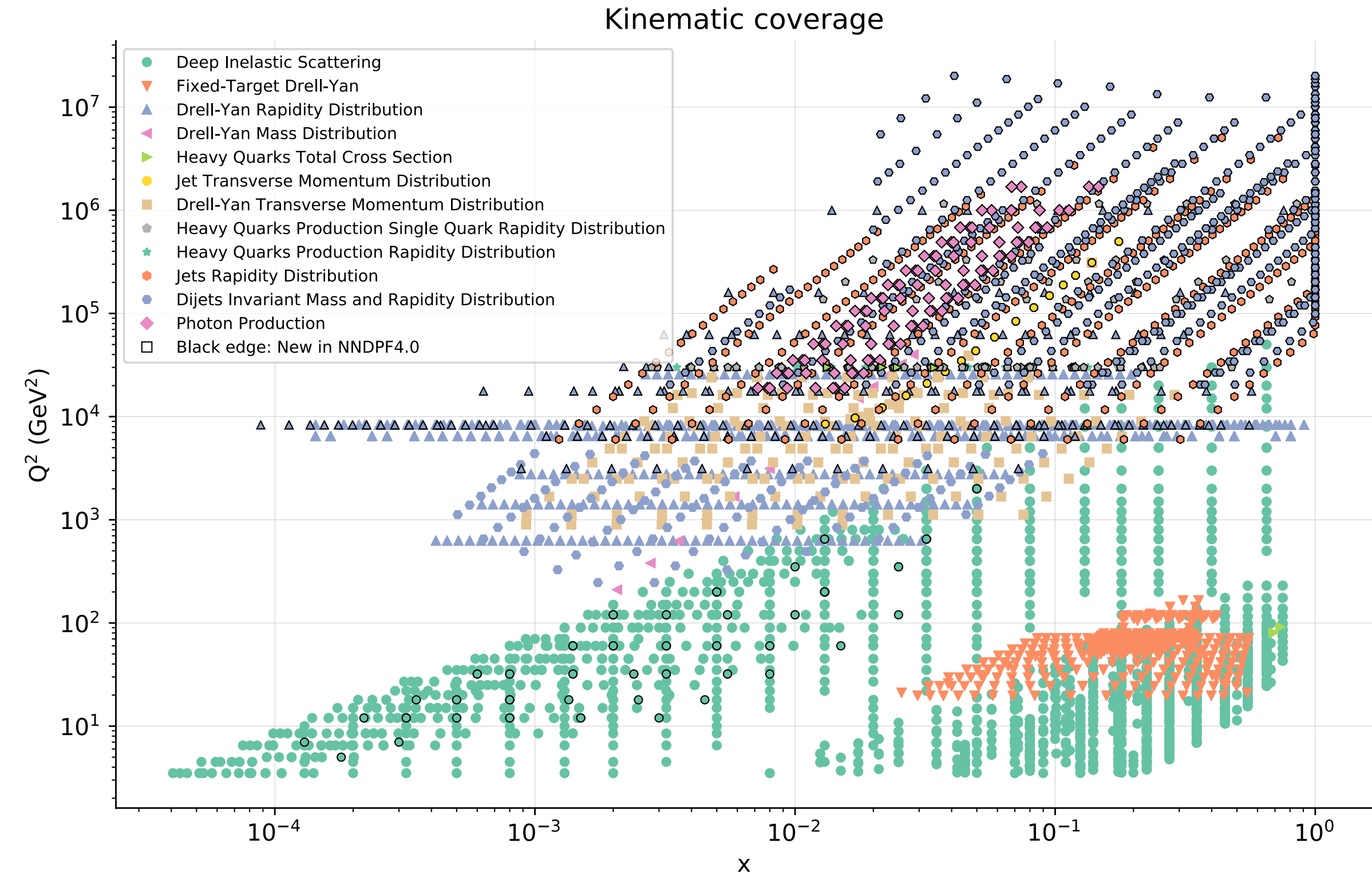
- **~4500 datapoints** across a wide range of kinematics and processes

- Uncertainties are approximated as **Gaussian**

- For Gaussian data the likelihood estimator is

$$P(\text{model} | \text{data}) \propto \exp[-\chi^2/2]$$

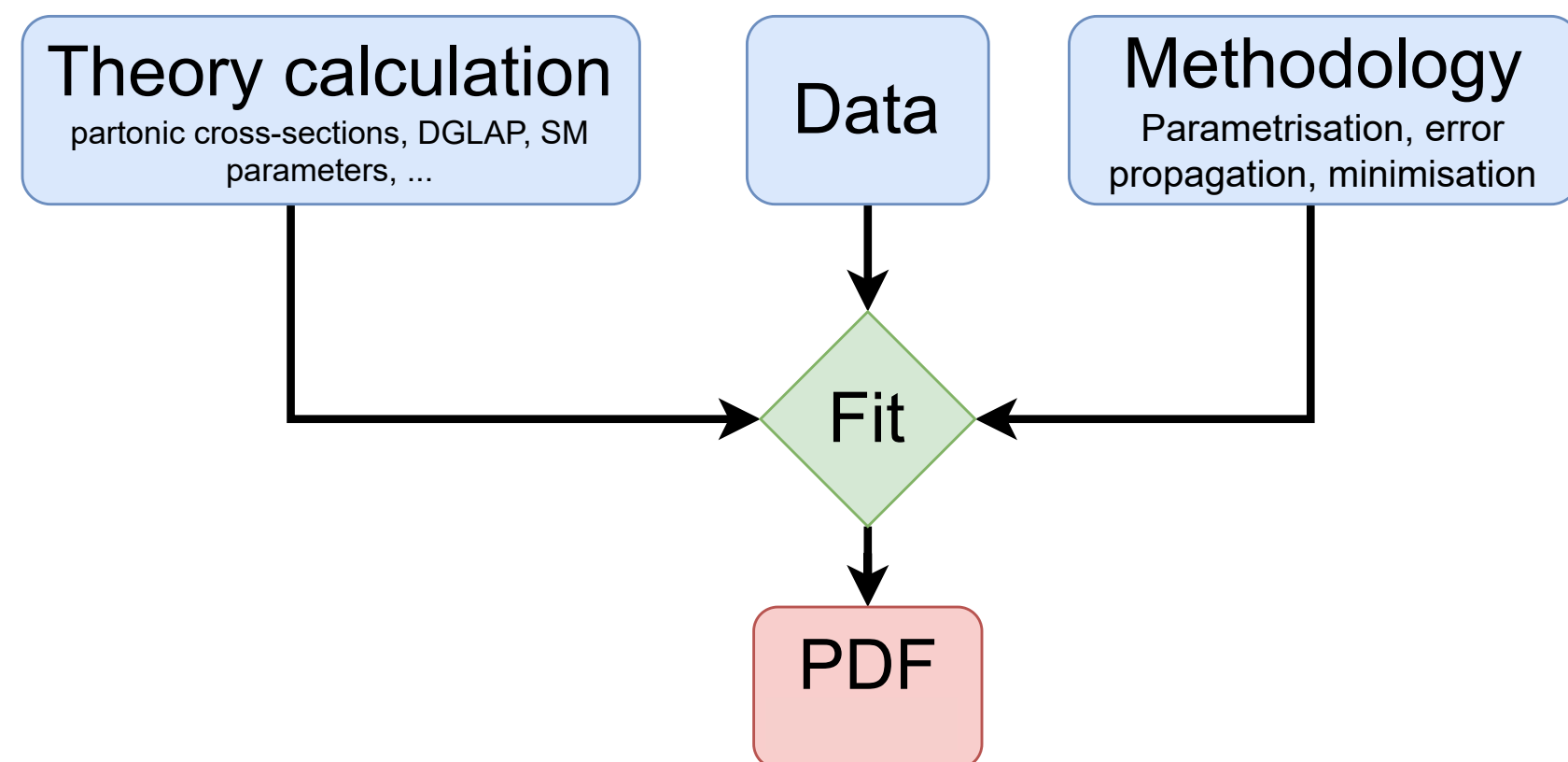
$$\chi^2 = \sum_{i,j}^{N_{\text{dat}}} (\text{data} - \text{prediction})_i \text{cov}_{ij}^{-1} (\text{data} - \text{prediction})_j$$



# PDF determination

Besides data a PDF fit requires theory calculations and a methodology, this talk is about the latter

*Different groups, make different choices for each*



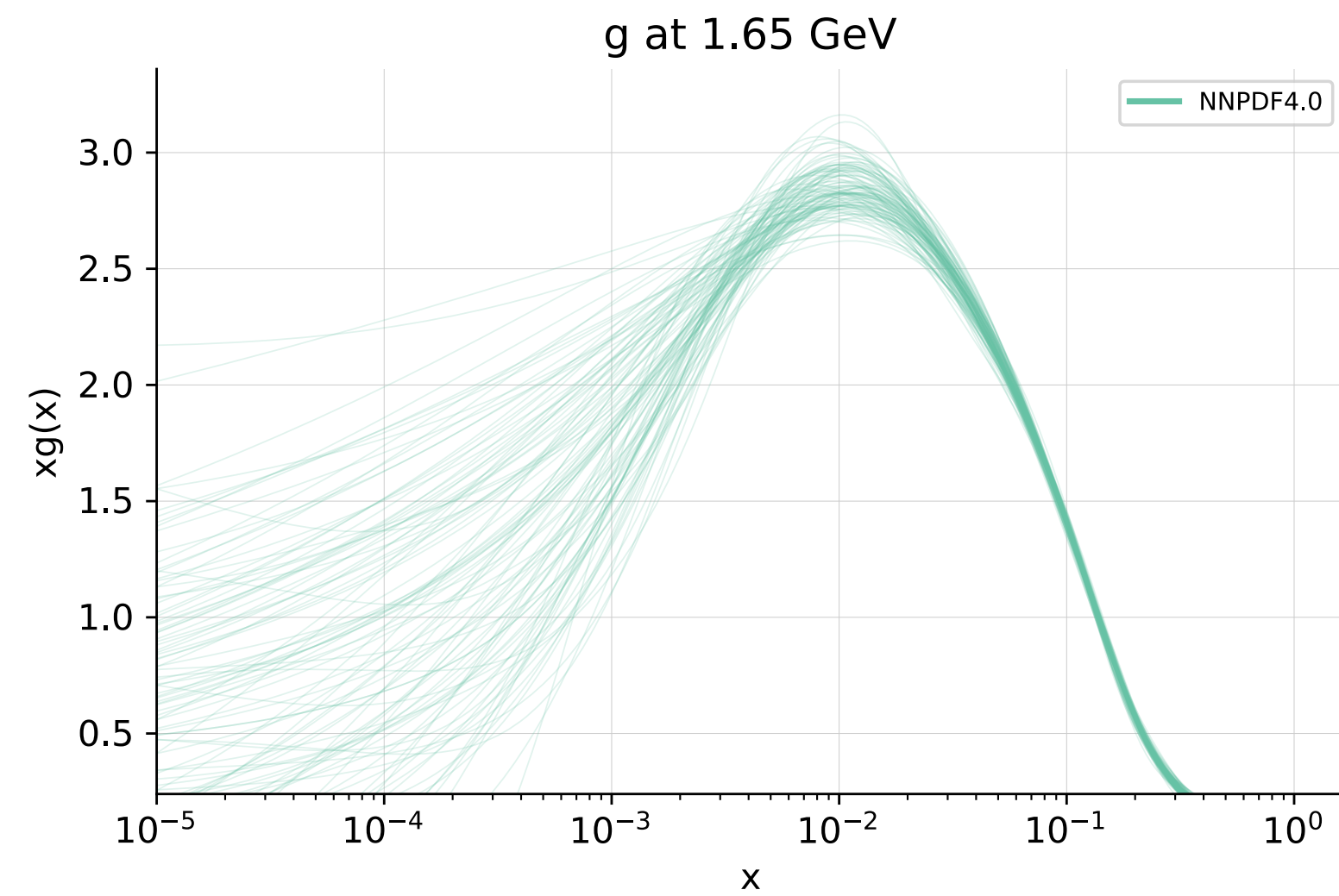
A methodology consists of...

- ...a way to **parametrise** the PDFs -> neural network
- ...a way to **fit parameters** to data -> gradient descent
- ...a way to **propagate uncertainties** from data to functions?

# Uncertainty propagation

Create a Monte Carlo samples of “**synthetic data replicas**”

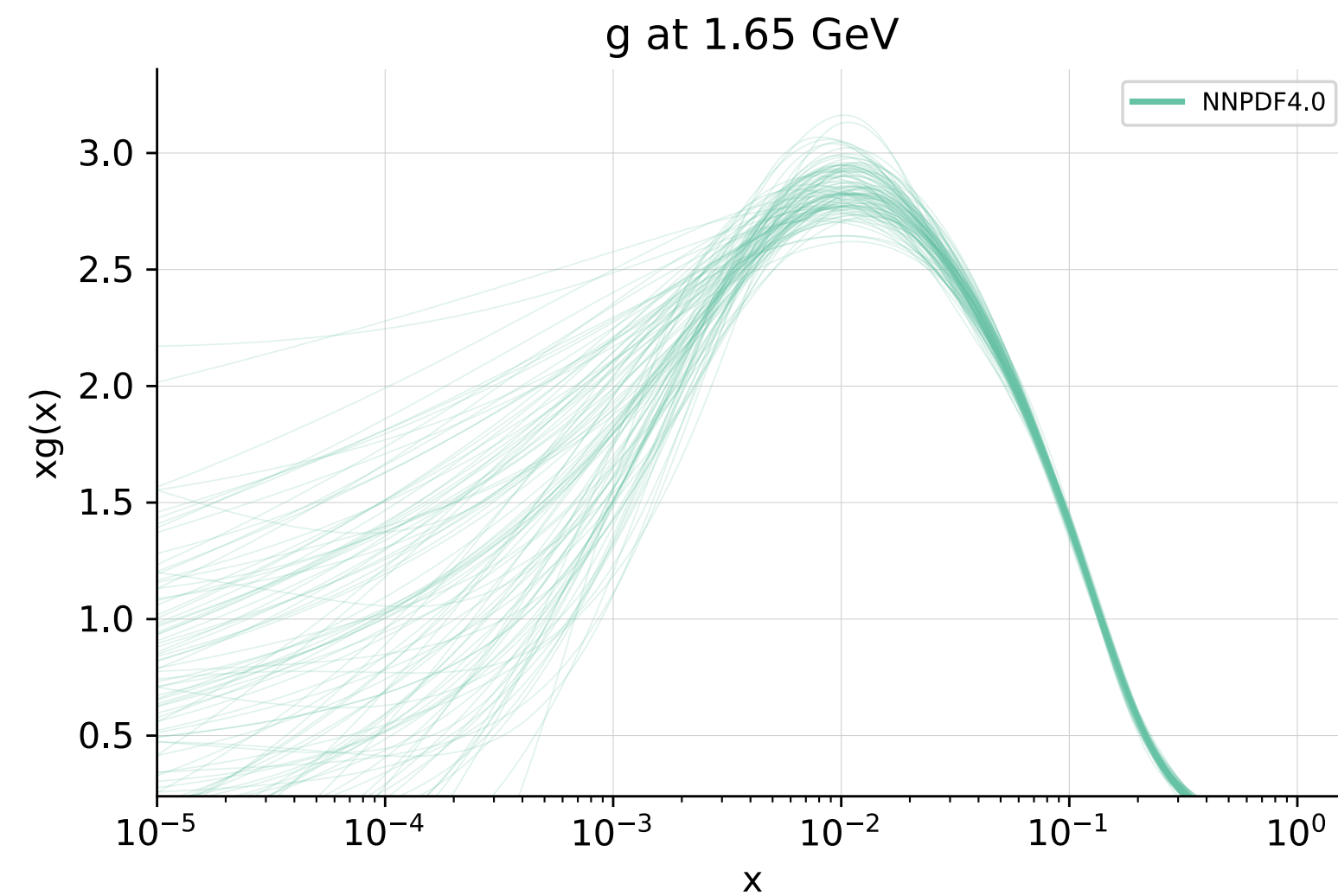
$$D^{(k)} \sim \mathcal{N}(D, \text{Cov}_{\text{exp}})$$





# Uncertainty propagation

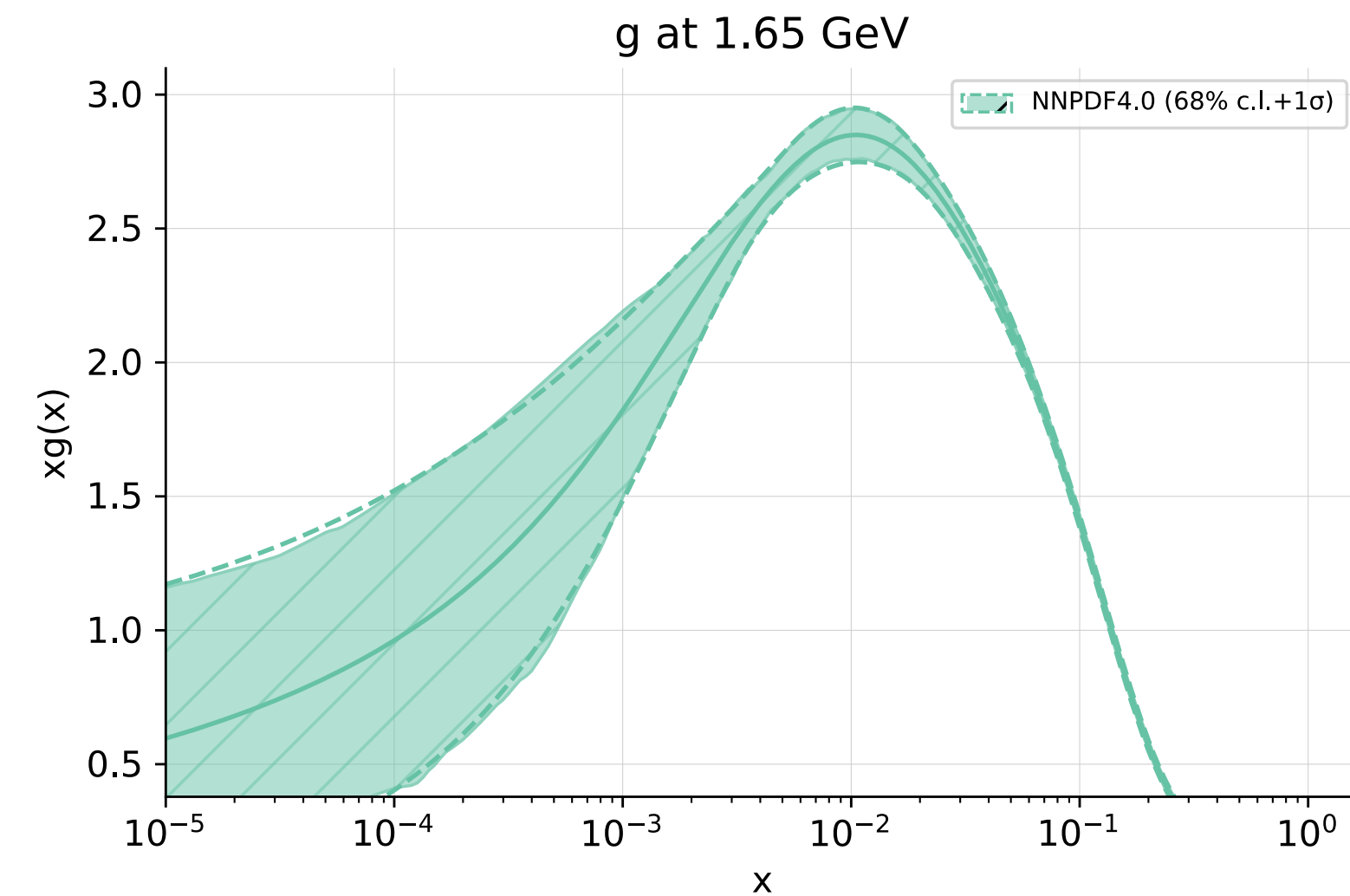
Create a Monte Carlo samples of “**synthetic data replicas**”  
 $D^{(k)} \sim \mathcal{N}(D, \text{Cov}_{\text{exp}})$



Compute mean and variance of PDF-dependent observables

$$\langle \mathcal{O}[f] \rangle \simeq \frac{1}{N} \sum_{k=1}^N \mathcal{O} [f^{(k)}]$$

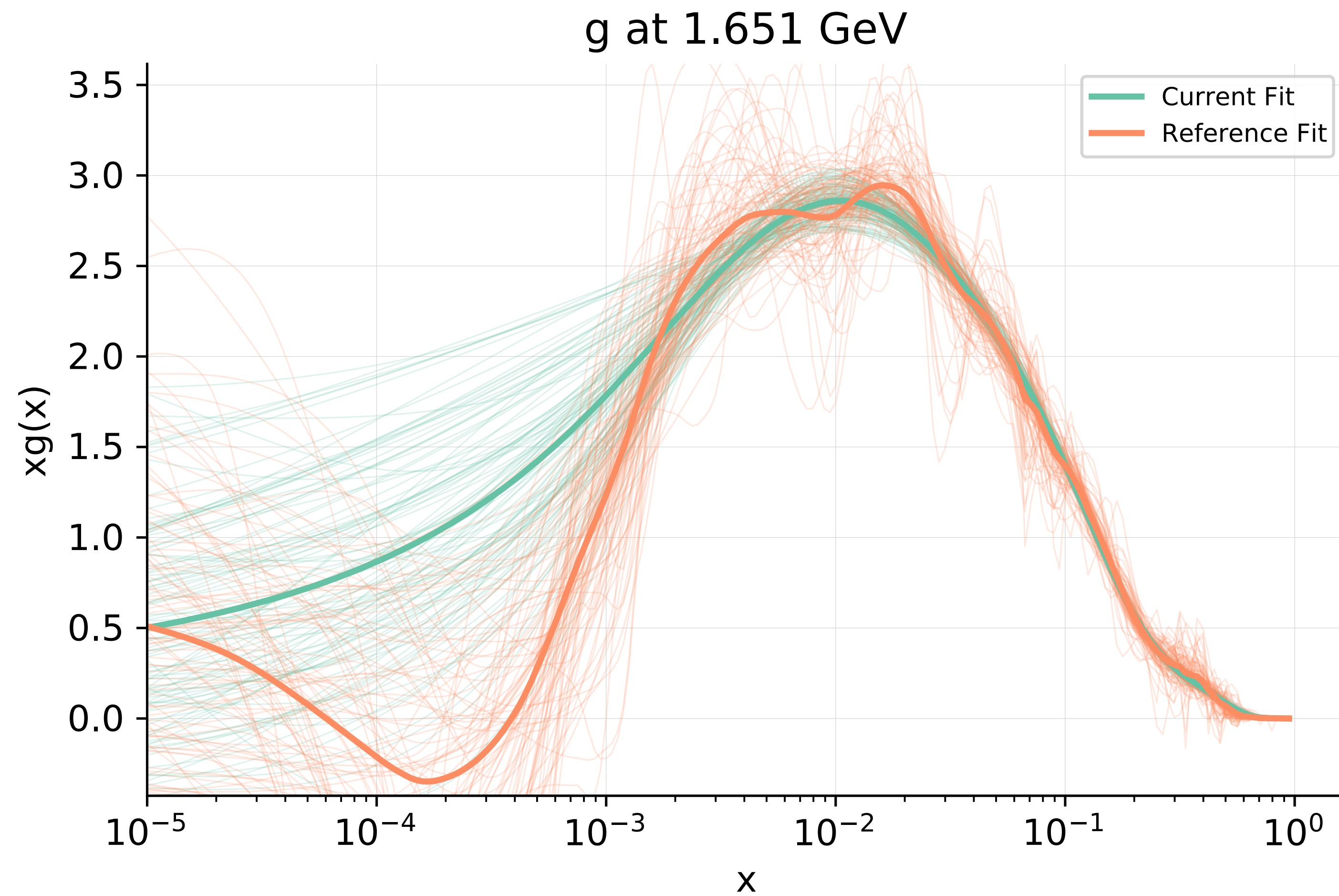
$$\text{Var}[\mathcal{O}] \simeq \frac{1}{N} \sum_{k=1}^N (\mathcal{O} [f^{(k)}] - \langle \mathcal{O} \rangle)^2$$





- ▶ The NNPDF methodology
- ▶ **Hyperoptimisation**
- ▶ Validation of the methodology

# Fitting PDFs



Setting the methodology hyperparameters requires care  
The wrong choice may lead to over- or under-fitting

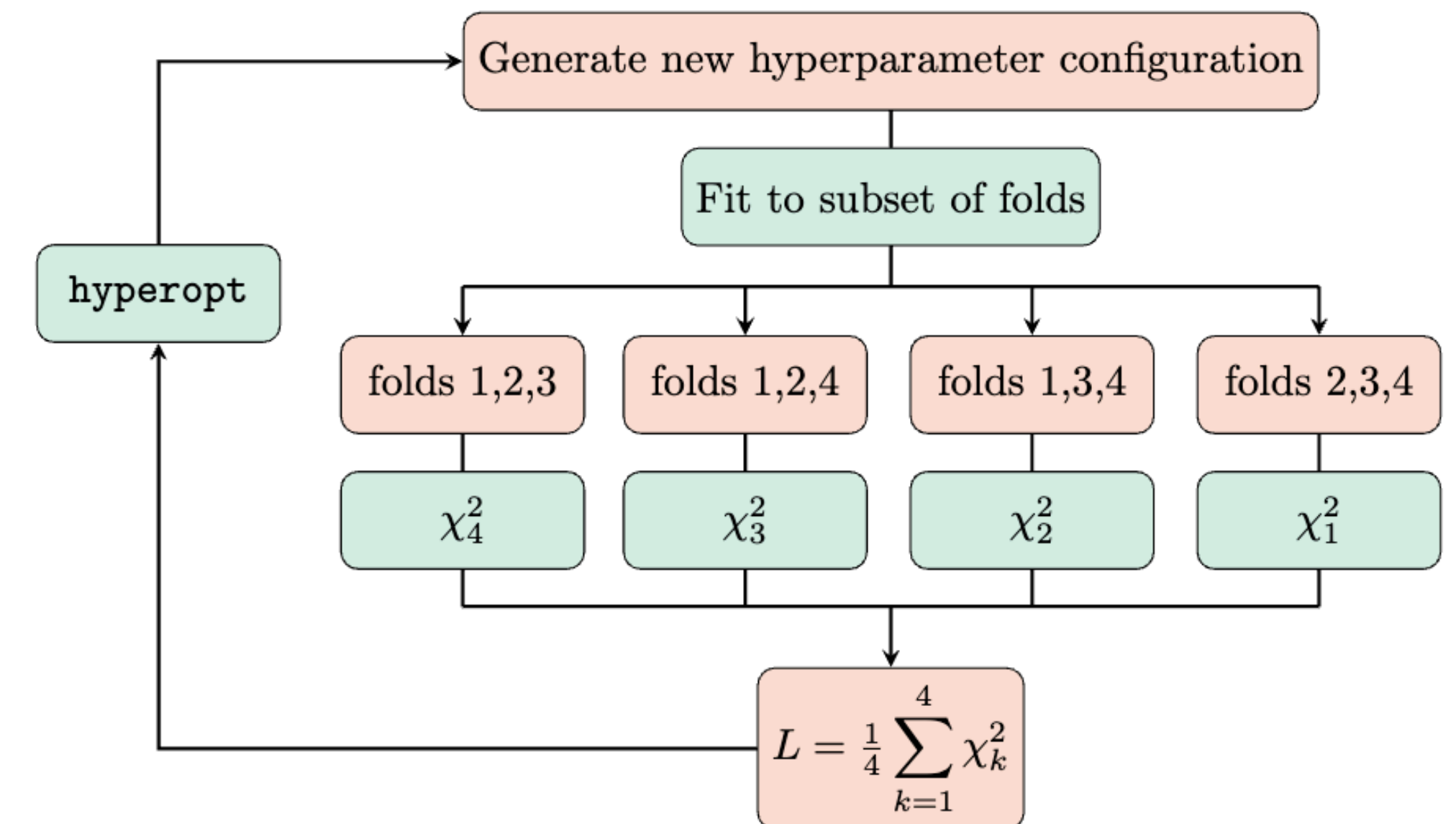
# Hyperparameter optimisation

## NNPDF4.0 used k-folds cross-validation

1. Partition the dataset into 4 folds
2. Exclude one at a time, perform 4 fits
3. Hyperoptimization metric: best average  $\chi^2$  to non-fitted data

Ideally include the PDF uncertainty in the hyperoptimization

- **computationally heavy:**  
4 cpu hours x 4 folds = 16 hours at 16 GB of memory
  - This had to be reduced to use higher moments in hyperoptimization
- ➡ solution: GPUs!



# Hyperopt using GPUs

# Replicas	10	50	100
Energy reduction	78%	87%	91%
Cost reduction	−45%	47%	55%

*NVIDIA H100 GPU vs 16 AMD EPYC Genoa CPU on SURF's SNELLIUS cluster*

## Technical changes:

- Single NN model for all samples
- Share memory-heavy objects
- Single hyperopt database shared by GPUs

## Results:

- Memory usage scales only weakly with number of replicas, enabling a 100 replica fit in a single GPU
- 90% energy reduction: faster and more affordable fits!

# Model selection

**Difficult question:** what metric should be used to define a “good fit”?

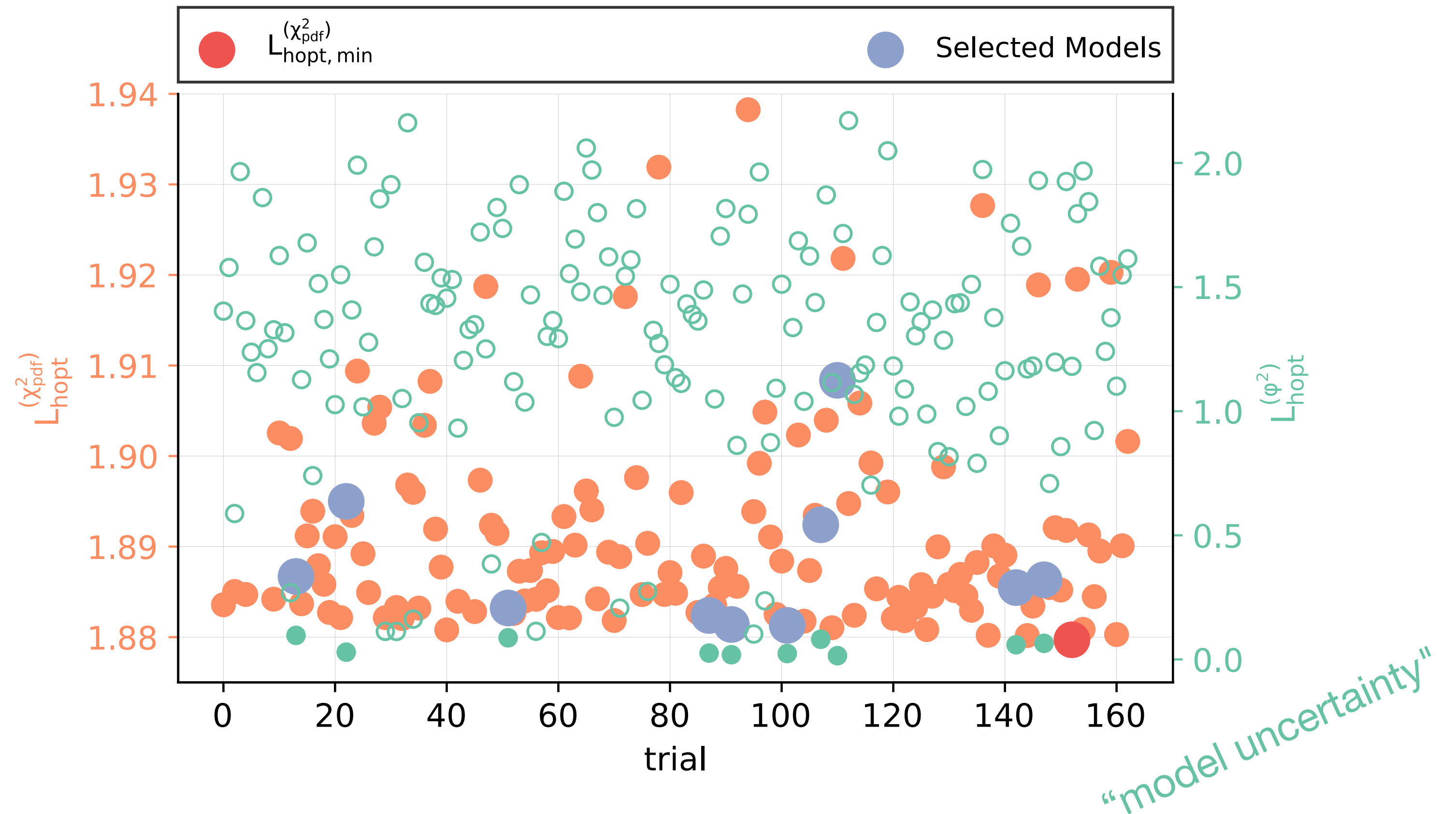
This is actively worked on, but as a first attempt we used the strategy:

- 1) look at configurations that describe data equally well
- 2) Pick the ones with the largest uncertainty

In a fit: select not a single setup but randomly **sample over all acceptable configurations**

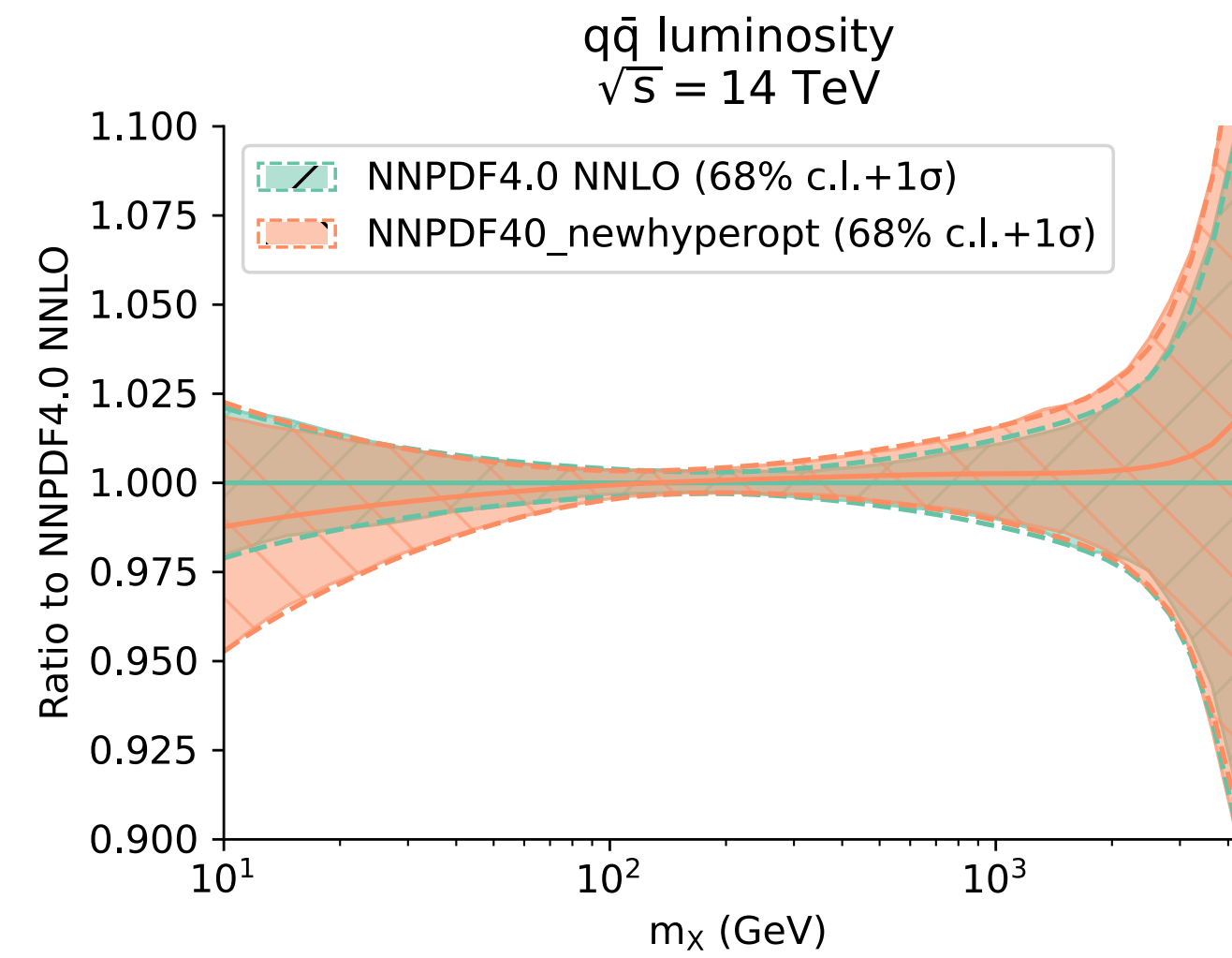
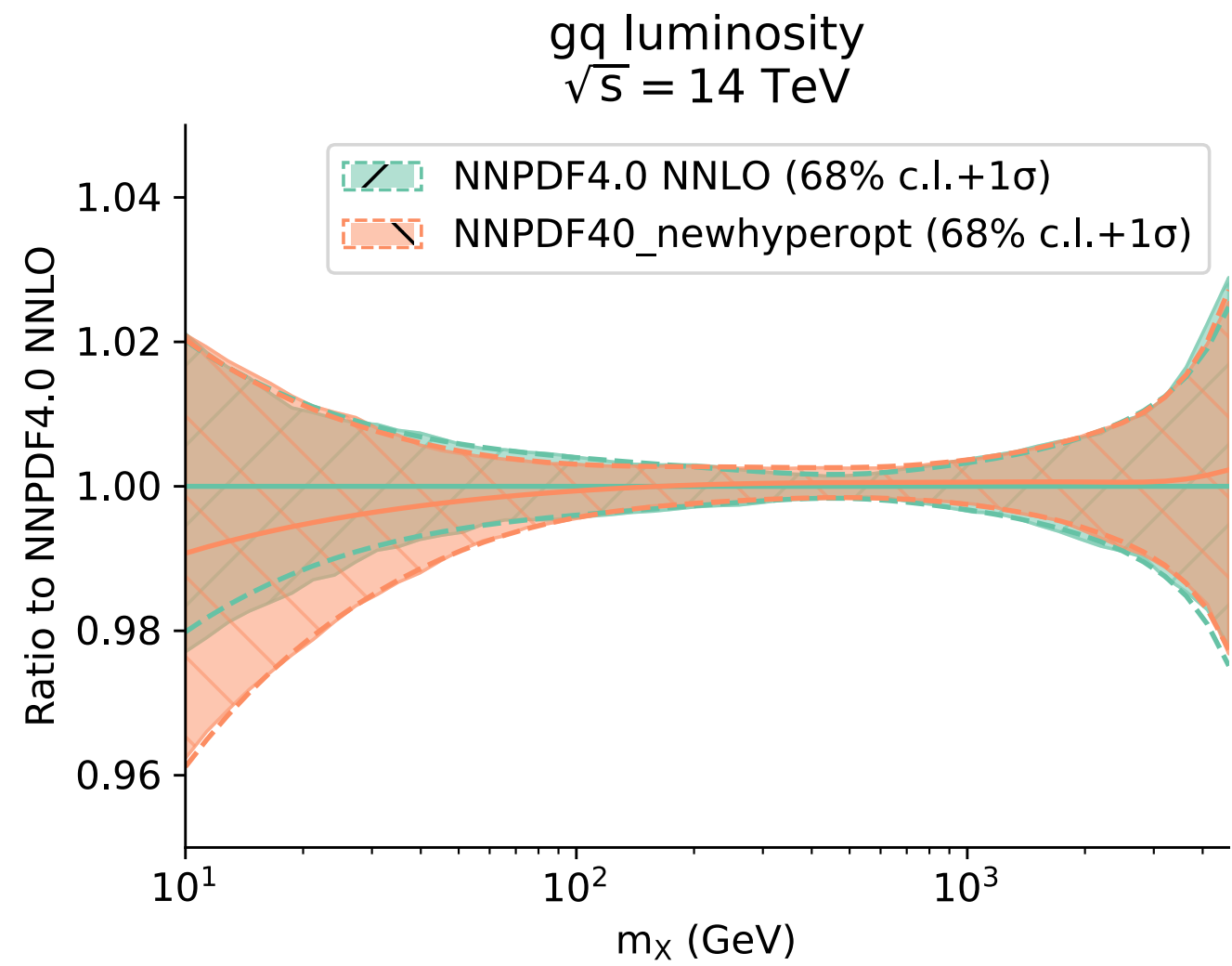
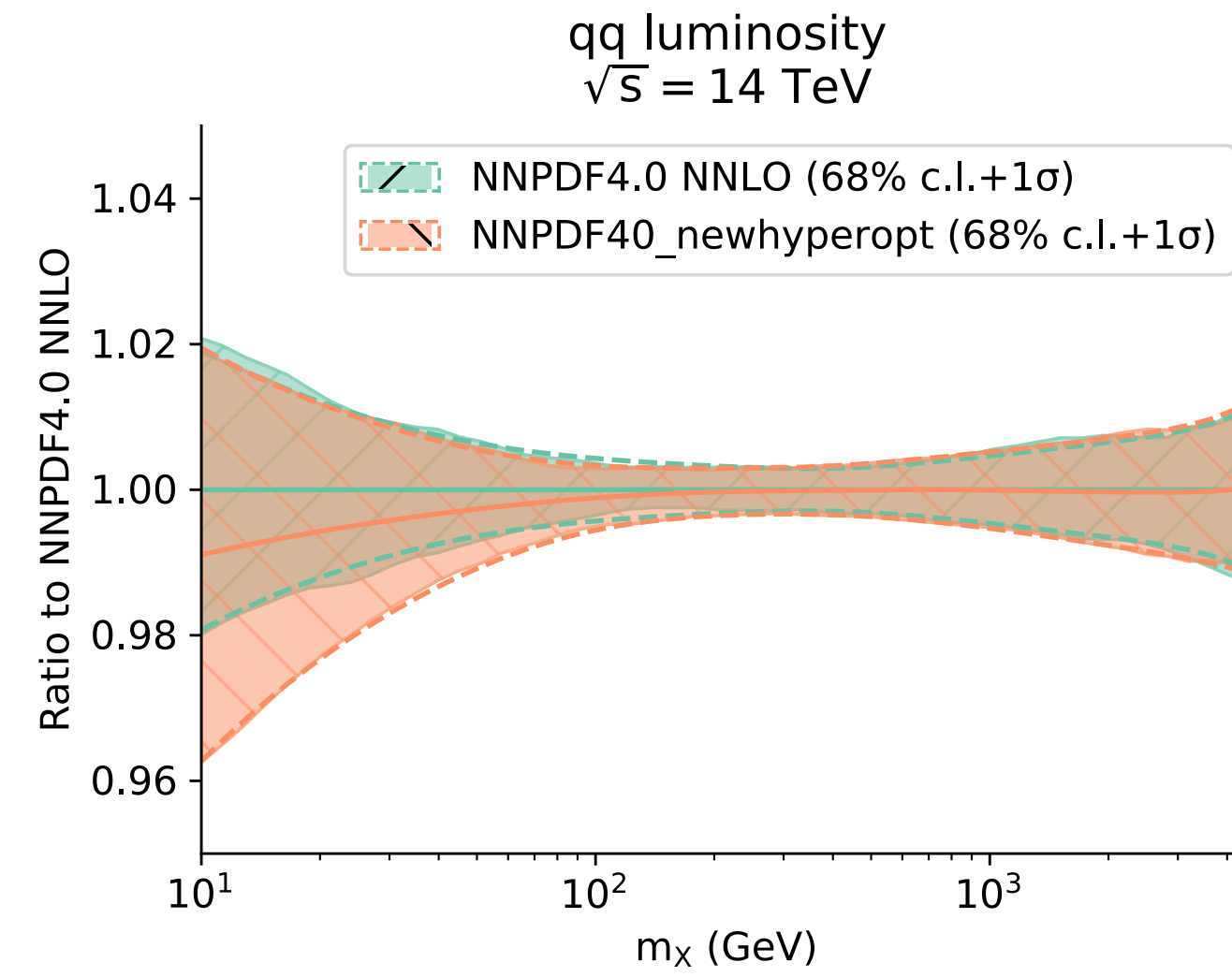
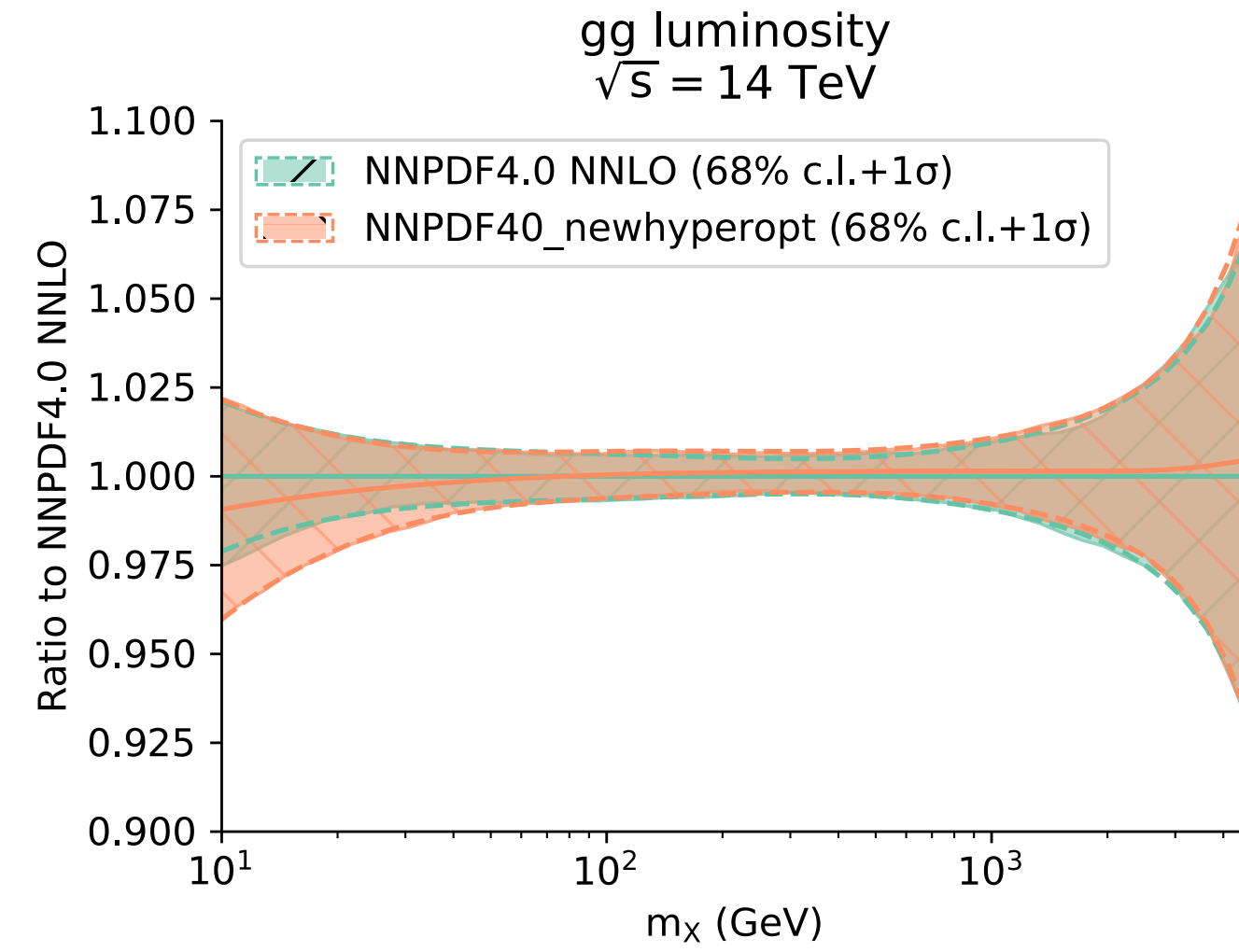


How to define the figure of merit?





# Results



Large changes to the hyper parameter determination methodology,  
but results still in **good agreement with NNPDF4.0**

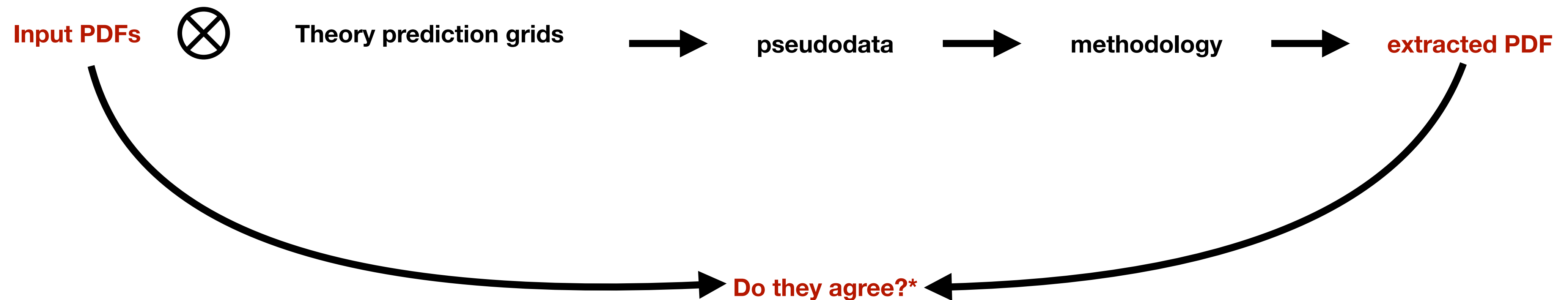


- ▶ The NNPDF methodology
- ▶ Hyperoptimisation
- ▶ **Validation of the methodology**

# Uncertainty validation: closure tests

[Del Debio, Giani, Wilson, 2111.05787 ]

**Basic idea:** generate a global pseudo dataset from theory predictions and extract the PDFs from this



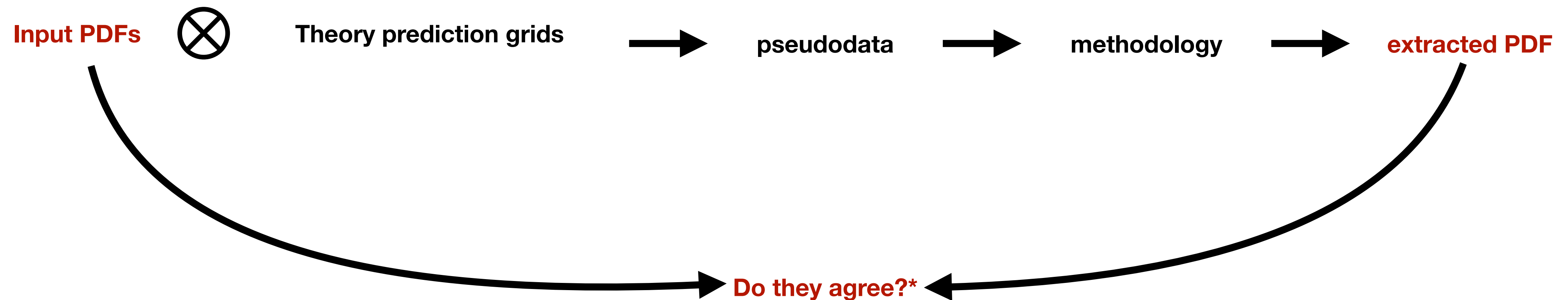
# Uncertainty validation: closure tests

[Del Debio, Giani, Wilson, 2111.05787 ]

**Basic idea:** generate a global pseudo dataset from theory predictions and extract the PDFs from this

\*

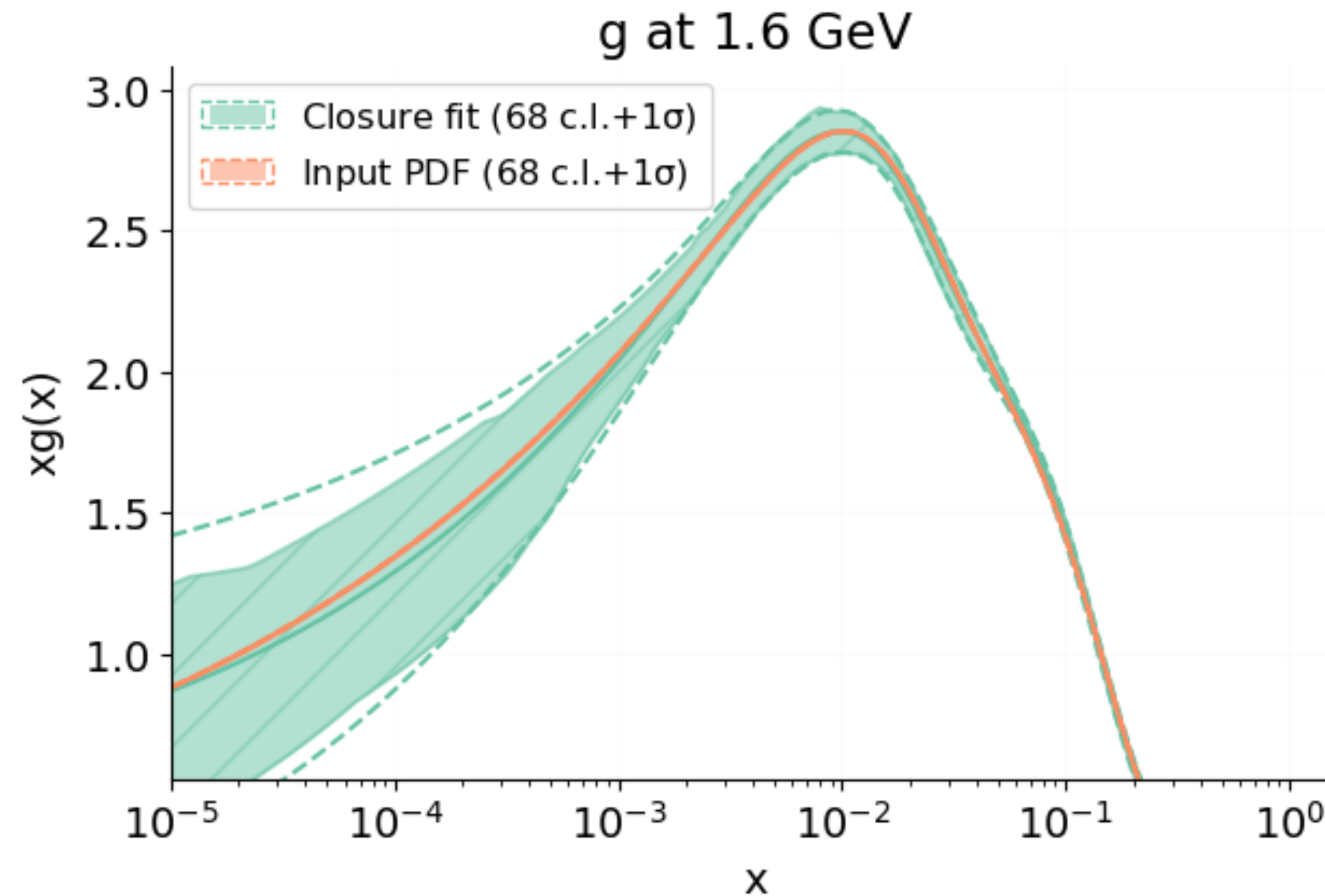
Experimental data is sampled from a distribution, therefore  
 $\text{pseudodata} = \text{prediction} + \text{noise}$



# Uncertainty validation: closure tests

[Del Debio, Giani, Wilson, 2111.05787 ]

Look at PDF: this seems okay



More quantitative: is the input data within 1 sigma of the prediction 68% of the time?

Use statistical measures to answer this

Recently the impact of **inconsistent data** was studied in a closure test

[Barontini et al., 2503.17447]

# Everything is open source!

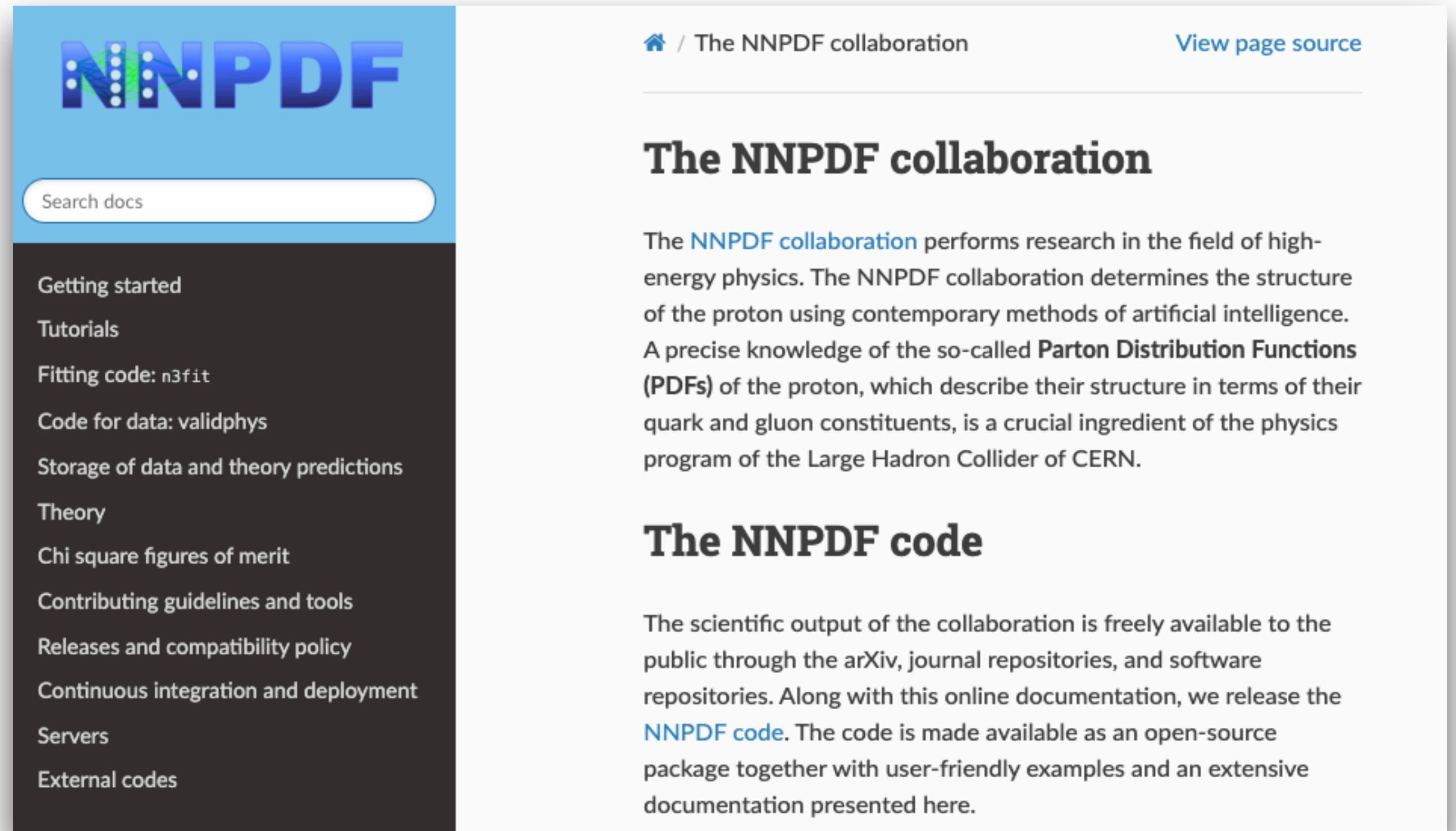
The NNPDF code is developed in a public repository

Everything to do your own PDF fit is open source:

- Data
- Theory grids
- Fitting methodology
- Analysis

GitHub: [github.com/NNPDF/nnpdf](https://github.com/NNPDF/nnpdf)

Documentations: [docs.nnpdf.science/](https://docs.nnpdf.science/)



The screenshot shows the NNPDF website. The top navigation bar includes the NNPDF logo, a search bar labeled 'Search docs', and links for 'The NNPDF collaboration' and 'View page source'. The left sidebar contains a list of links: 'Getting started', 'Tutorials', 'Fitting code: n3fit', 'Code for data: validphys', 'Storage of data and theory predictions', 'Theory', 'Chi square figures of merit', 'Contributing guidelines and tools', 'Releases and compatibility policy', 'Continuous integration and deployment', 'Servers', and 'External codes'. The main content area displays the title 'The NNPDF collaboration' followed by a paragraph describing the collaboration's research in high-energy physics and the determination of the proton's structure using artificial intelligence. Below this is the title 'The NNPDF code' followed by a paragraph explaining that the scientific output is freely available to the public through arXiv, journal repositories, and software repositories, and that the NNPDF code is released as an open-source package with user-friendly examples and extensive documentation.

🏠 / The NNPDF collaboration [View page source](#)

## The NNPDF collaboration

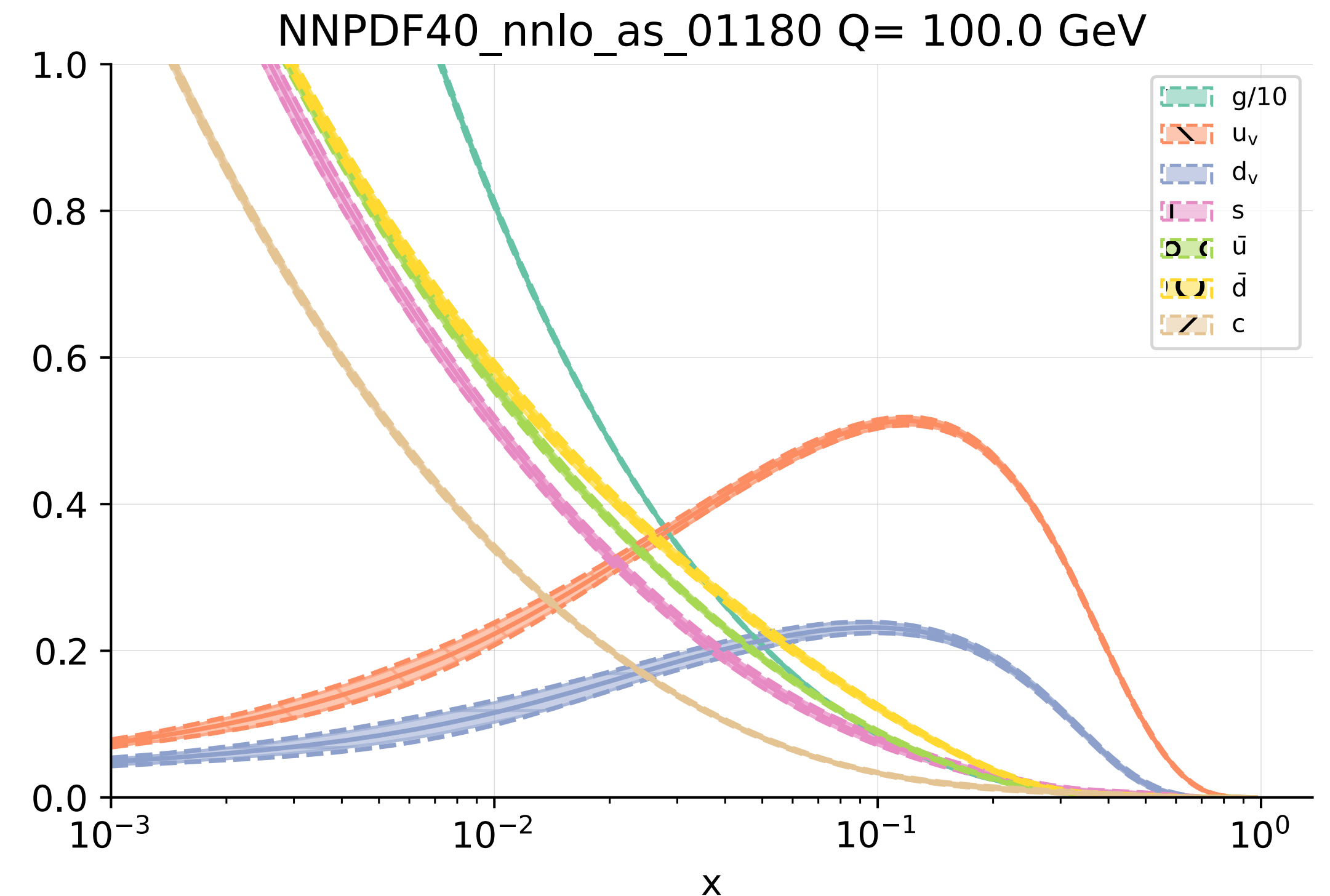
The [NNPDF collaboration](#) performs research in the field of high-energy physics. The NNPDF collaboration determines the structure of the proton using contemporary methods of artificial intelligence. A precise knowledge of the so-called **Parton Distribution Functions (PDFs)** of the proton, which describe their structure in terms of their quark and gluon constituents, is a crucial ingredient of the physics program of the Large Hadron Collider of CERN.

## The NNPDF code

The scientific output of the collaboration is freely available to the public through the arXiv, journal repositories, and software repositories. Along with this online documentation, we release the [NNPDF code](#). The code is made available as an open-source package together with user-friendly examples and an extensive documentation presented here.

# Summary and Outlook

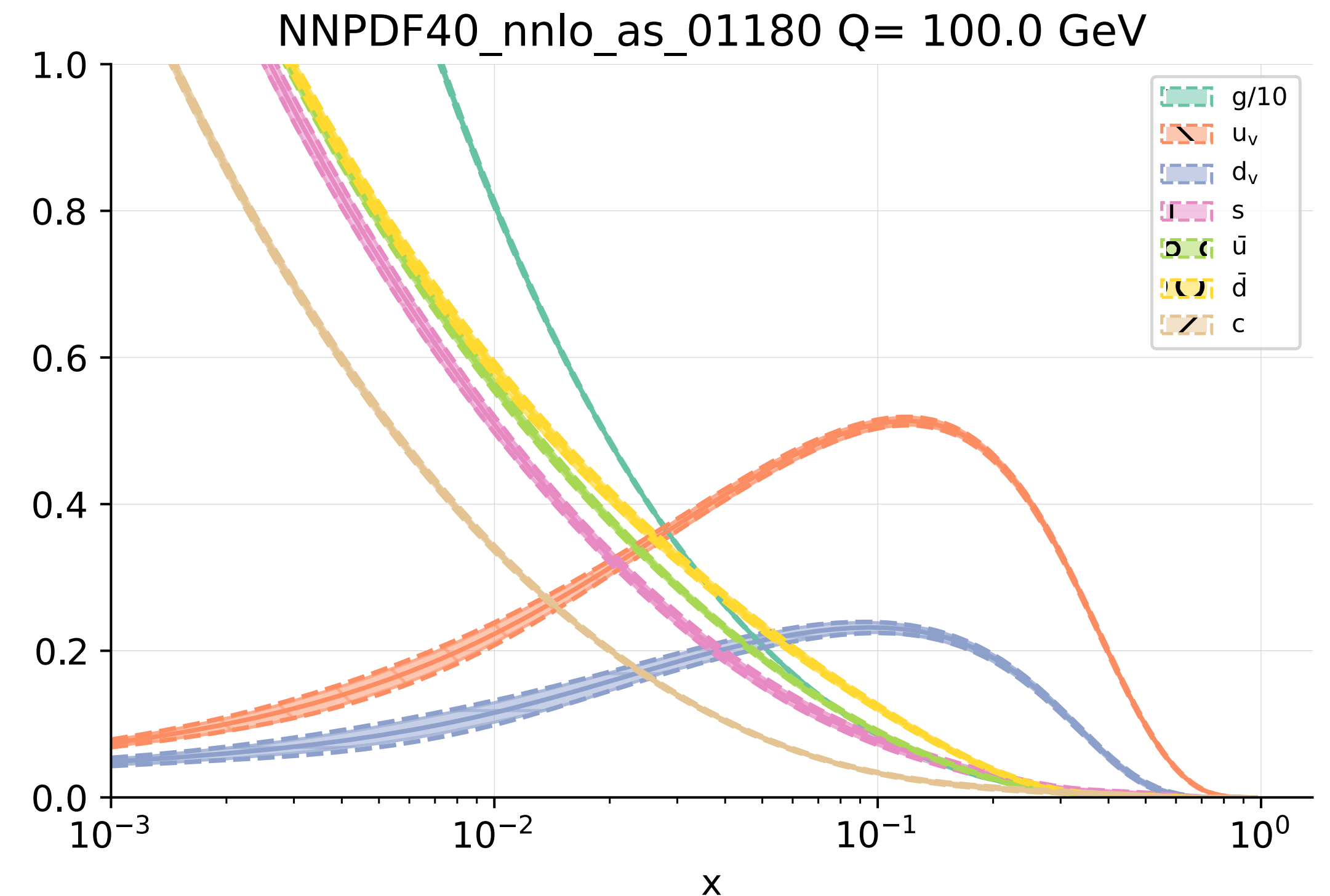
- PDF determination is a Machine Learning challenge
- Hyperparameter tuning is an important step in selecting good ML models
- GPU optimisation has led to 90% reduction in energy cost ...
- ... and enables us to do hyperoptimisation based on PDF distributions rather than a single replica





# Summary and Outlook

- PDF determination is a Machine Learning challenge
- Hyperparameter tuning is an important step in selecting good ML models
- GPU optimisation has led to 90% reduction in energy cost ...
- ... and enables us to do hyperoptimisation based on PDF distributions rather than a single replica



**Thank you for your attention!**

**Backup slides**