On the extraction of collinear PDFs

Valerio Bertone

IRFU, CEA, Université Paris-Saclay

June 23, 2021, Kickoff meeting - GDR-QCD / WG1

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement № 824093

Disclaimer

- When talking about PDFs we often mean ***unpolarised collinear PDFs of the proton***.
  - They are relevant in processes that involve *unpolarised* protons in the initial state and in which *all hard scales are of the same order* and much larger than Λ_{QCD}.
  - Typical situation at the **LHC**.
- **Many other kinds of PDFs** exist:
  - collinear *nuclear* PDFs,
  - collinear *longitudinally* polarised PDFs,
  - collinear *transversely* polarised PDFs,
  - transverse-momentum-dependent (*TMD*) PDFs,
  - *unintegrated* PDFs
  - *diffractive* PDFs,
  - **...**
- All these different PDFs are appropriate in **other contexts** different from that mentioned above.

Everything starts from...

A **collinear factorisation theorem**:

$$d\sigma_{\rm had} = W_{\{i\}} \otimes \mathcal{L}_{\{i\}}\, d\Phi$$

Everything starts from...

A **collinear factorisation theorem**:

$$d\sigma_{\text{had}} = W_{\{i\}} \otimes \mathcal{L}_{\{i\}} d\Phi$$

Hard cross sections:

- process dependent,
- high-energy dominated,
- computable in perturbation theory.

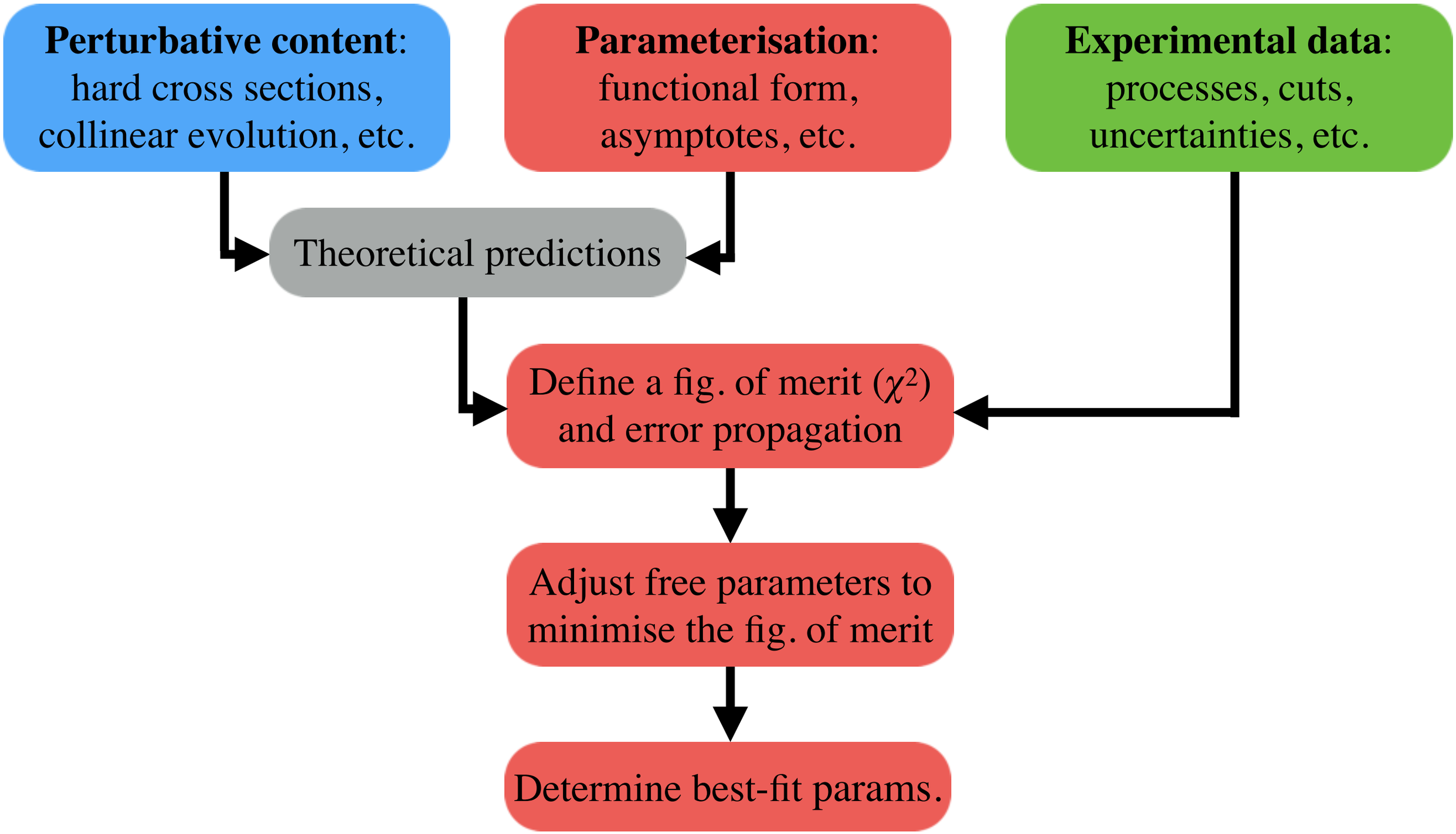
Parton distribution functions (PDFs):

- universal,
- low-energy dominated,
- perturbation theory inapplicable.

How do we determine PDFs?

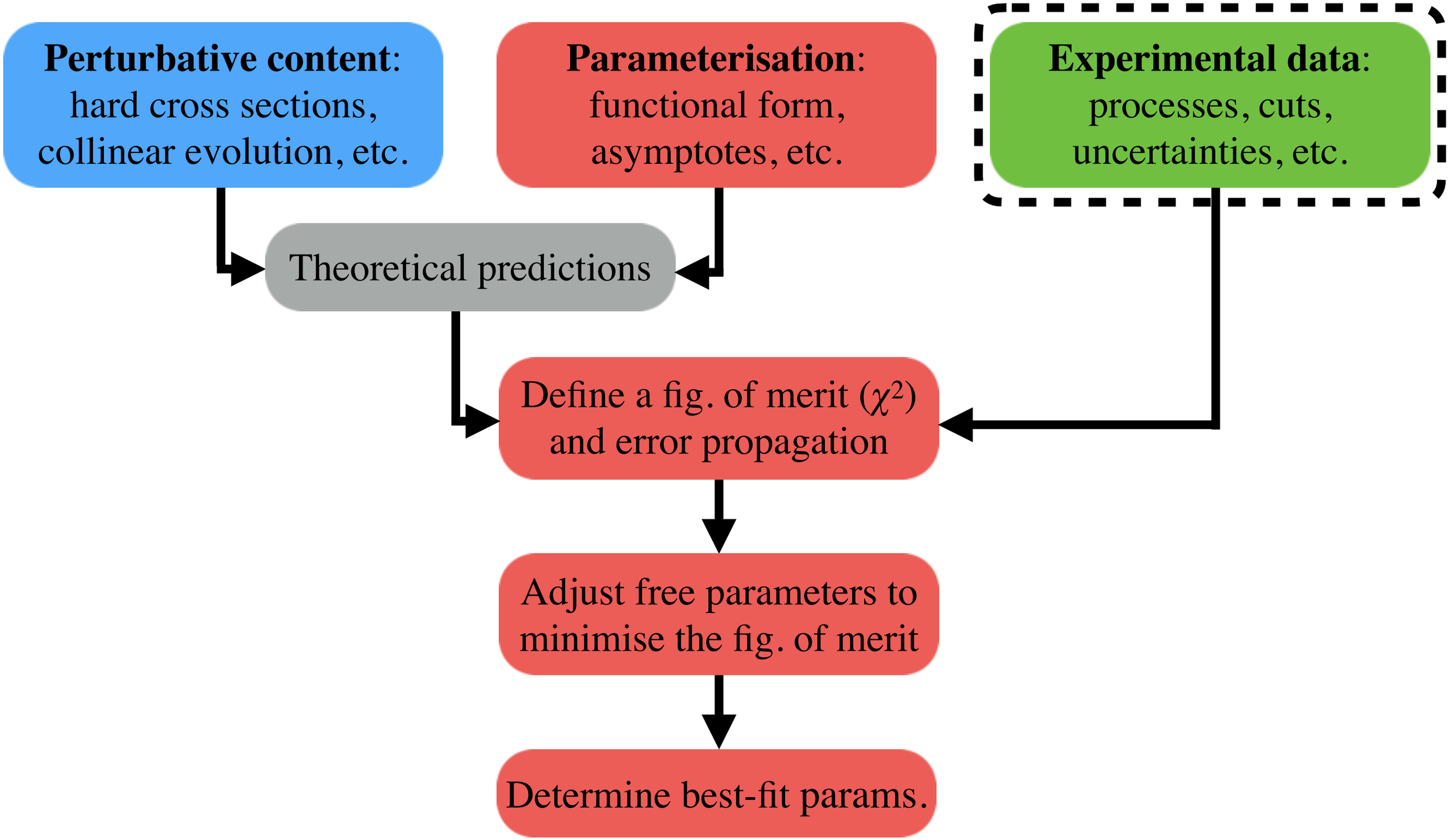
Currently, the most accurate and reliable way is through **fits to data**.

The general strategy

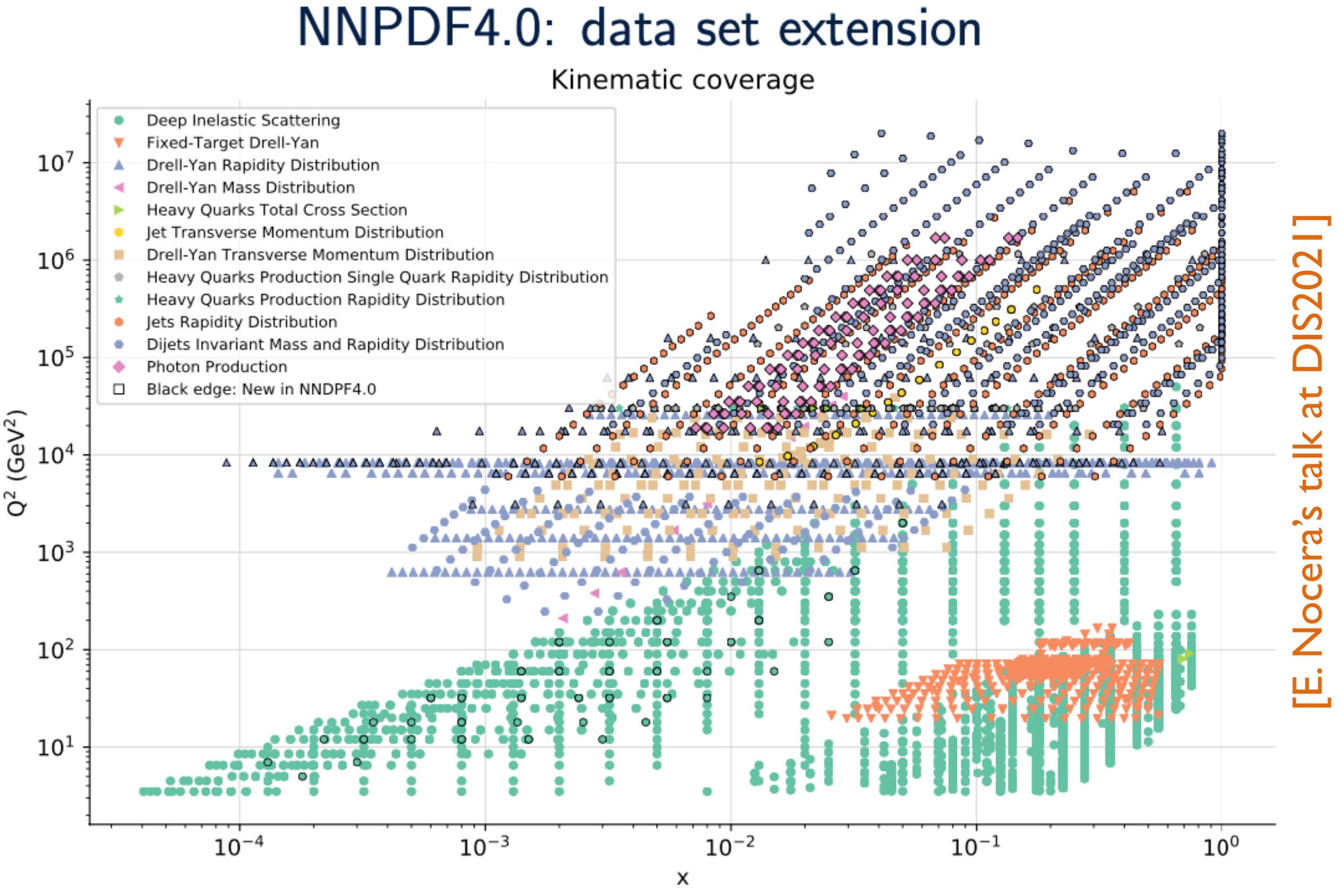


Each box requires a choice. **Different choices** lead to **different determinations**.

The general strategy

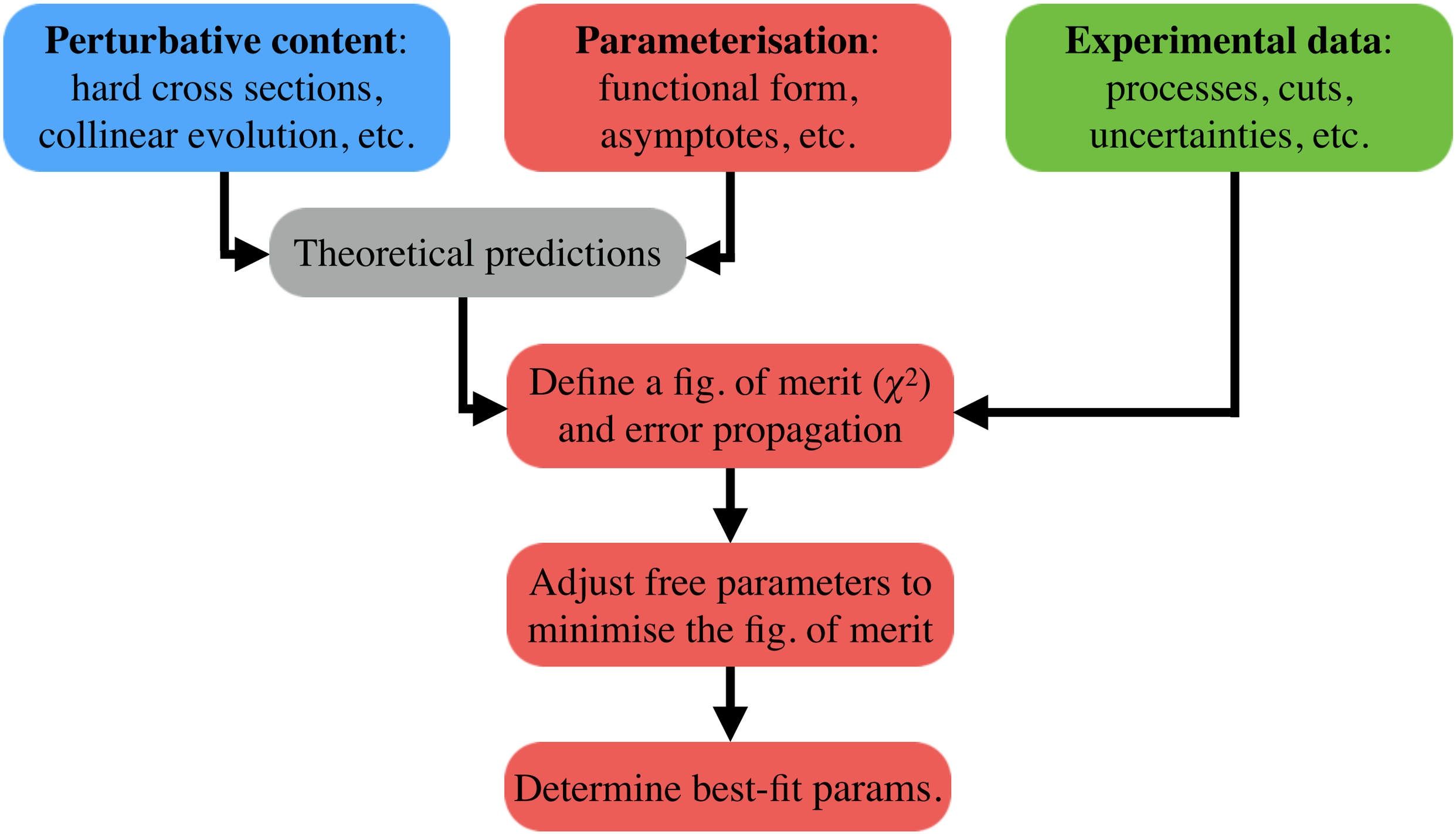


Experimental data

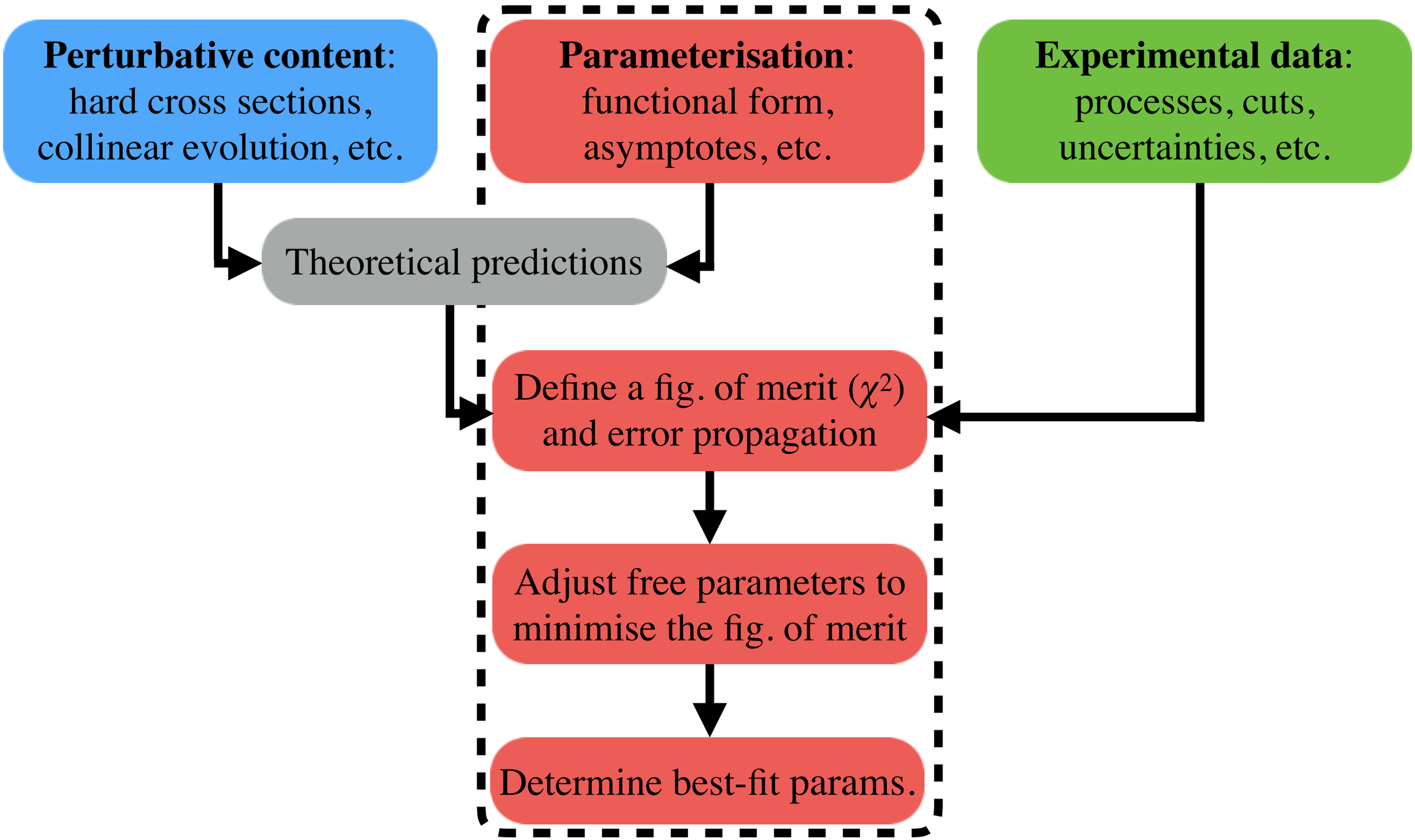


[E. Nocera's talk at DIS2021]

The general strategy



The general strategy



Fit methodologies

Parameterisation: the "standard" approach

- Distributions are parametrised by means of the functional form:

$$f_i(x) = A_i\, x^{\alpha_i} (1-x)^{\beta_i} P_i(x)$$

with:

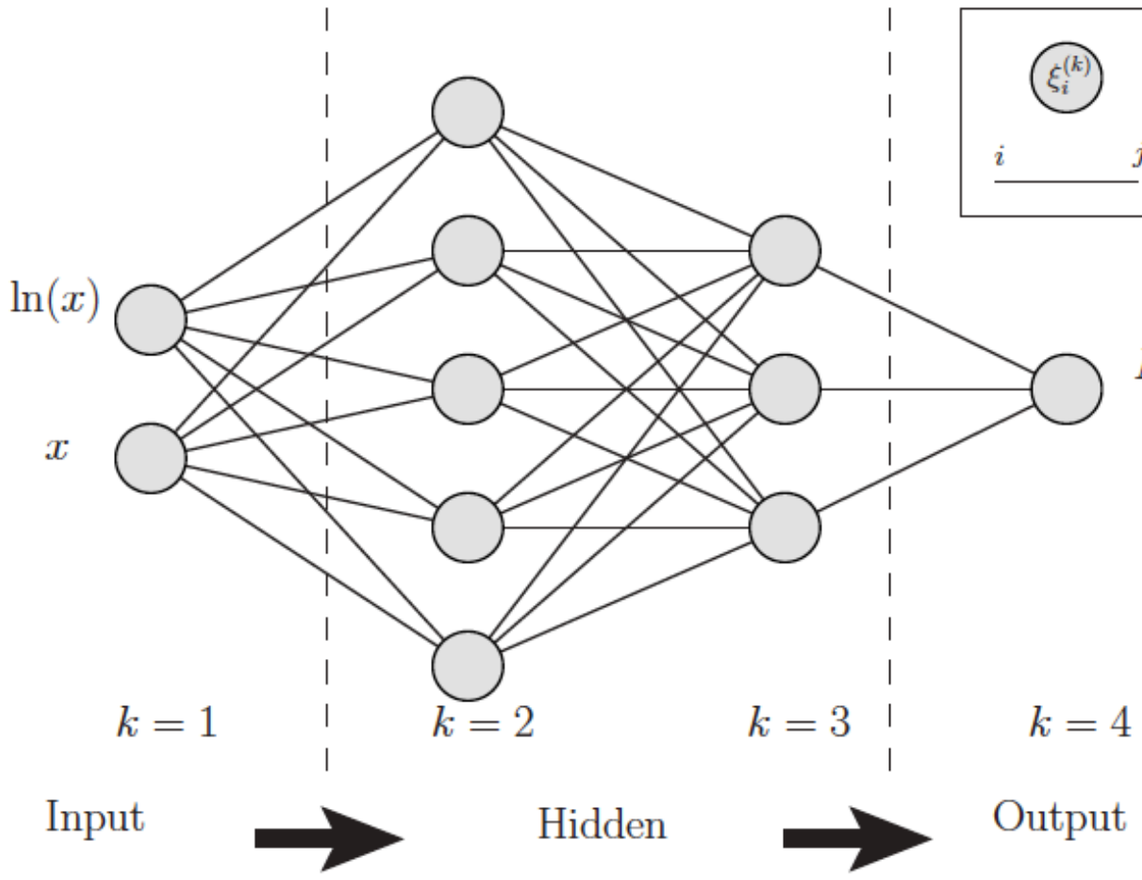
$$P_i(x) = \begin{cases} 1 \\ 1 + \gamma_i x \\ 1 + \gamma_i x + \delta_i \sqrt{x} \\ \dots \end{cases}$$

- **O(3-5) free parameters** for each distribution.
- **Asymptotic behaviour** defined by the exponents α_i and β_i.
- Typically easy to transform analytically in **Mellin space**.
- **Easy to handle** in a fit thanks to its simplicity.
- Potential **source of bias**.

Fit methodologies

Parameterisation: neural networks

- Distributions are parametrised in terms of artificial NNs:



$$\xi_i^{(j)} = g\left(\sum_k^{\text{(j-1)th layer}} \xi_k^{(j-1)}\omega_{ki}^{(j)} - \theta_i^{(j)}\right)$$

Activation function
(sigmoid, tanh, ReLu, …)

- Each NN has a **large number free parameters**.
- NNs are usually augmented with constraints in the extrap. regions:

$$f_i(x) = A_i x^{\alpha_i}(1-x)^{\beta_i}\mathrm{NN}_i(x) \quad \text{or} \quad f_i(x) = \mathrm{NN}_i(x) - \mathrm{NN}_i(1)$$

- NNs are **flexible** and thus limit biases but are **harder to handle**.

Fit methodologies

Figure of merit: the χ^2 definition

- A crucial aspect in the determination of PDFs is the definition of the **figure of merit** to be minimised/maximised.

- A popular choice is the χ^2 but **many variants** are possible:

  - No correlation, no normalisation unc.: $\chi^2 = \sum_{i=1}^{N_{\rm dat}} \frac{(T_i - D_i)^2}{\sigma_i^2}$

  - No correlation, with normalisation unc.: $\chi^2 = \sum_{j=1}^{N_{\rm exp}} \left[\left(\frac{1-\mathcal{N}_j}{\delta \mathcal{N}_j} \right) + \sum_{i=1}^{N_{\rm dat}^j} \frac{(\mathcal{N}_j T_i - D_i)^2}{\sigma_i^2} \right]$

  - Nuisance parameters: $\chi^2 = \sum_i \frac{\left[T_i \left(1 - \sum_j \gamma_j^i b_j \right) - D_i \right]^2}{\delta_{i,\rm unc}^2 T_i^2 + \delta_{i,\rm stat}^2 D_i T_i} + \sum_j b_j^2$

  - Covariance matrix: $\chi^2 = \sum_{ij} (T_i - D_i) \sigma_{ij}^{-2} (T_j - D_j)$

- Due to the **D'Agostini bias**, a sound treatment of normalisation uncertainties requires particular care (*e.g.* the t_0 prescription).

Fit methodologies

Error propagation

- A faithful determination implies a solid estimate of the **uncertainty** on PDFs propagating from the **experimental** dataset.

1. **Hessian** method: the χ^2 is **expanded** around its minimum $\mathbf{a}_0$:

$$\chi^2(\{\mathbf{a}\}) \simeq \chi^2(\{\mathbf{a}_0\}) + \underbrace{\frac{1}{2}\left.\frac{\partial\chi^2}{\partial a_i \partial a_j}\right|_{\mathbf{a}_0}}_{H_{ij}} (a_i - a_{0i})(a_j - a_{0j})$$

The Hessian matrix H_{ij} is **diagonalised** and an uncertainty along each eigenvector is defined as $\Delta\chi^2 = 1$ (often a larger **tolerance** is introduced).

2. **Monte Carlo** sampling: artificial **replicas** of the dataset generated as:

$$D_i^{(k)} = D_i + r_i^{(k)} \sigma_i \,, \qquad \begin{matrix} k = 1, \ldots, N_{\text{rep}} \\ i = 1, \ldots, N_{\text{dat}} \end{matrix}$$

$r_i^{(k)}$ is a *normally distributed* and *univariate* random number. A fit is performed to each replica to produce N_{rep} sets of distributions $\{f_k\}$, such that:

$$\langle \mathcal{O} \rangle = \frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} \mathcal{O}[f_k] \quad \text{and} \quad \sigma_{\mathcal{O}} = \sqrt{\langle \mathcal{O}^2 \rangle - \langle \mathcal{O} \rangle^2}$$

Fit methodologies

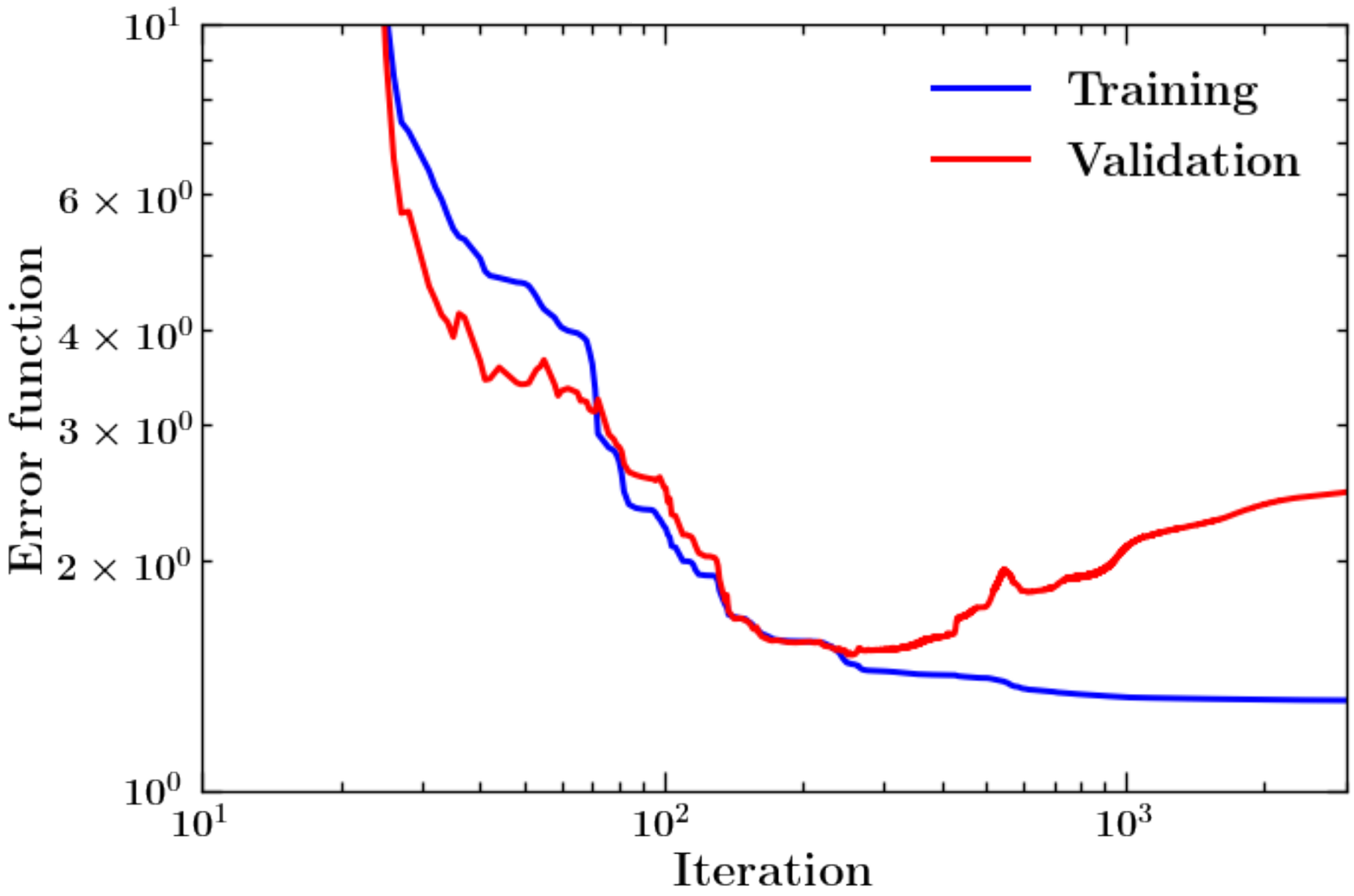
Minimisation and stopping

- Simple parameterisations (**O(20) free parameters**) are usually fitted using **MINUIT** (or similar):
  - the absolute minimum of the χ^2 is found *deterministically* by computing (numerically or analytically) the first derivative and moving **downhill**.
- A NN parameterisation (**O(200) free parameters**) generates a complex parameter space impossible to explore with MINUIT:
  - a **genetic algorithm** is often used to explore the parameter space,
  - this avoids getting trapped into **local minima** of the χ^2.
  - Algorithms inspired by **machine-learning** techniques are being explored,
  - **gradient-descent** based algorithms are recently also being used.
- The extreme flexibility of NNs may cause **overfitting**, *i.e.* statistical fluctuations of the data sample may be unwillingly fitted:
  - the **cross-validation** method allows one to overcome this problem.

Fit methodologies

Cross validation

- Split the dataset into **training** and **validation** subsets.
- Minimise the training χ^2 while monitoring the validation χ^2.
- Stop the fit when the validation χ^2 reaches its absolute minimum.



Main PDF collaborations

Unpolarised proton PDFs

- **CTEQ** collaboration:
  - standard parameterisation (**Bernstein** polynomials),
  - **Hessian** method (with dynamical tolerance) for error propagation.
- **NNPDF** collaboration:
  - neural network parameterisation (feed forward NN with preprocessing),
  - **Monte Carlo** method for error propagation.
- **MSHT** collaboration:
  - standard parameterisation (**Chebyshev** polynomials),
  - **Hessian** method (with dynamical tolerance) for error propagation.
- Other collaborations exist (e.g. ABMP, HERAPDF, CJ, etc.) but they are typically less inclusive in terms of data.

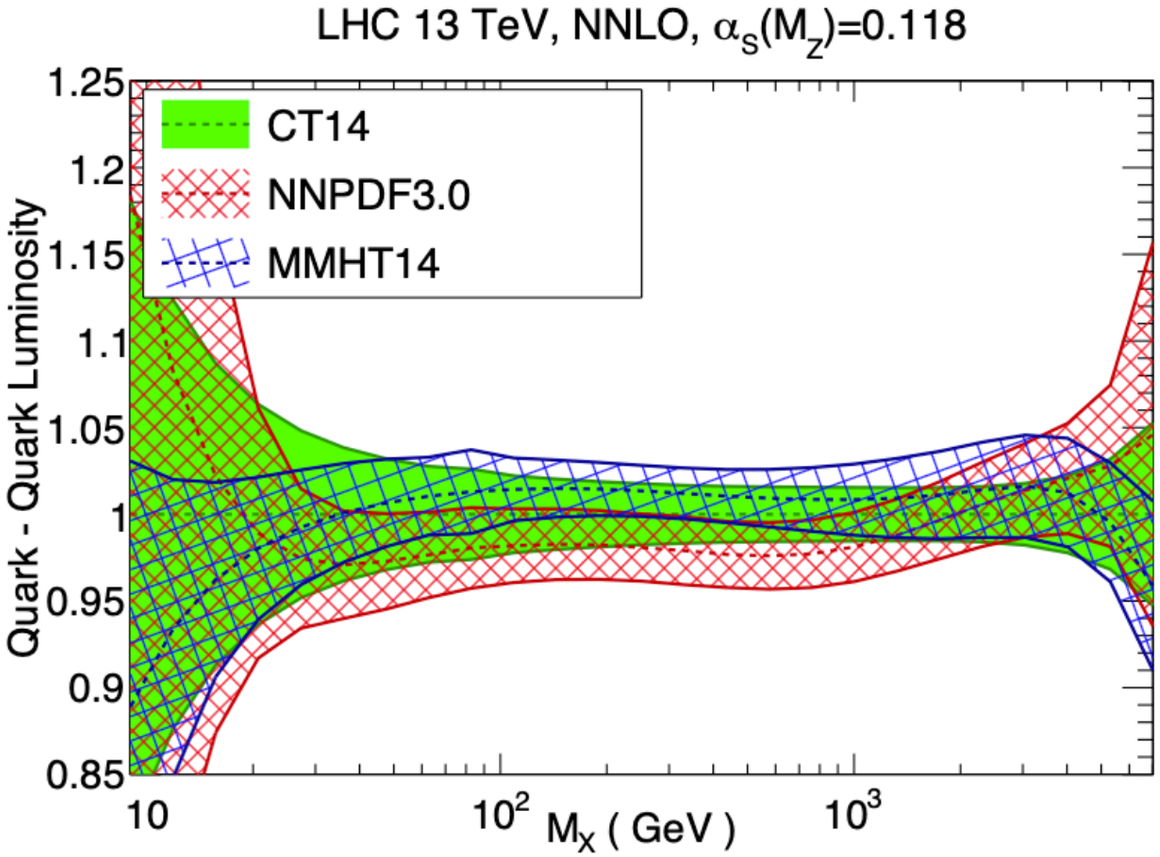
Parton luminosities

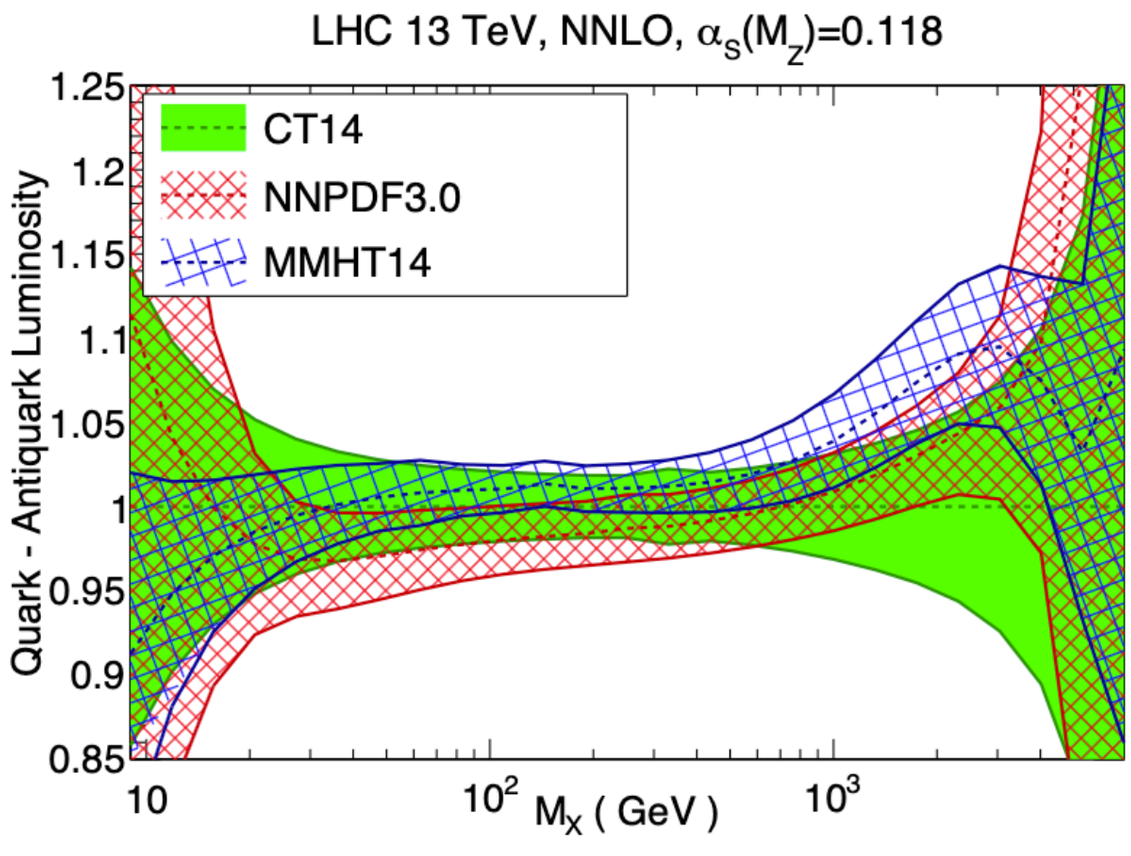
- Interesting quantities are the so-called **parton luminosities**:

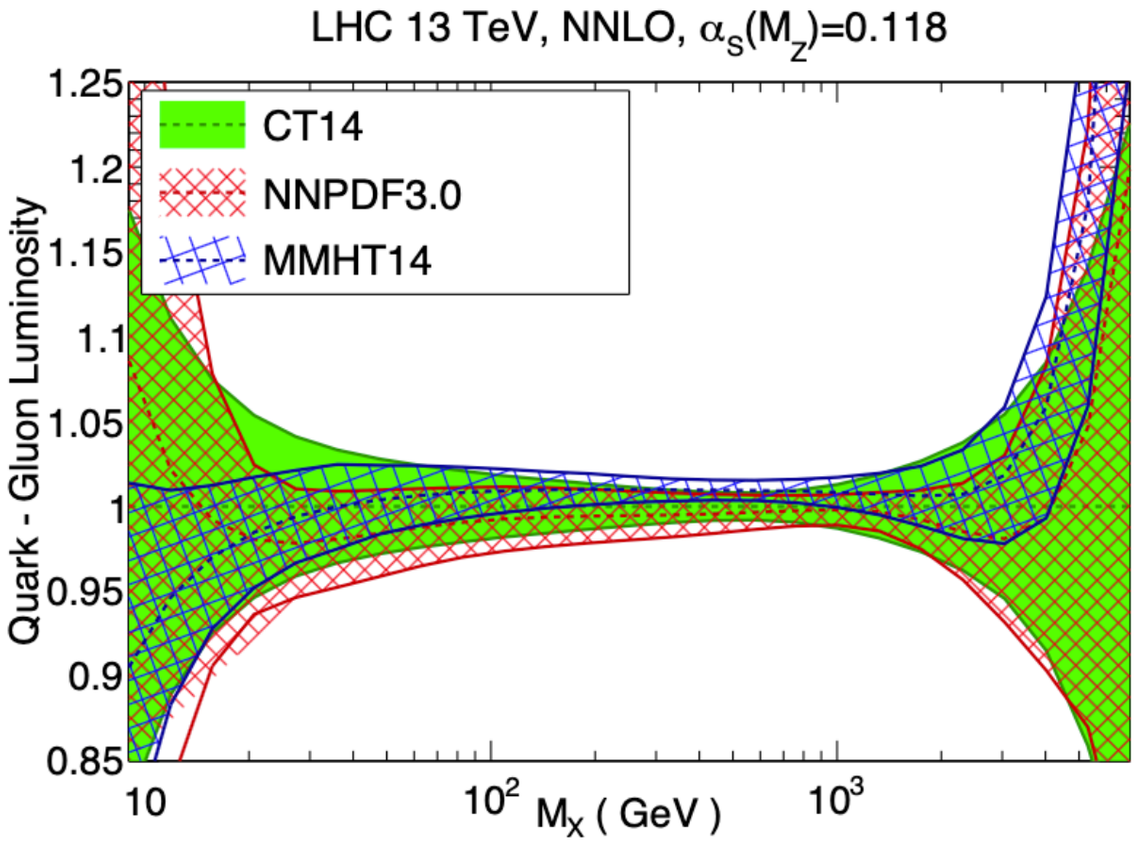
$$\mathcal{L}_{ij} = \frac{1}{s}\int_{M_X^2/s}^1 \frac{dy}{y} f_i(y, M_X) f_j\left(\frac{M_X^2}{ys}, M_X\right)$$

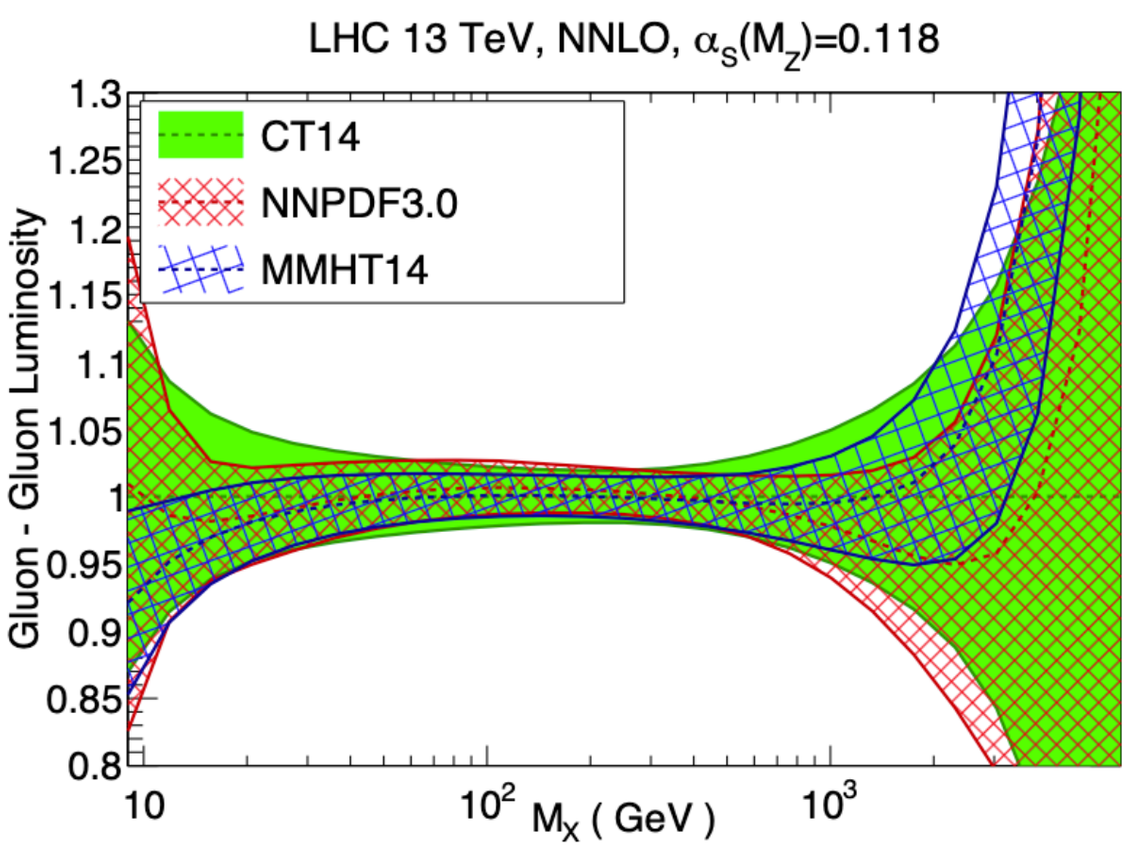
- Relevant for invariant mass distributions of the final state in pp collision processes, *e.g.*:
  - **Drell-Yan** mostly sensitive to $\mathscr{L}_{q\bar{q}}$,
  - **Higgs** production in gluon fusion mostly sensitive to $\mathscr{L}_{gg}$,
  - ***W* + charm** mostly sensitive to $\mathscr{L}_{sg}$,
  - …

A snapshot back in 2015



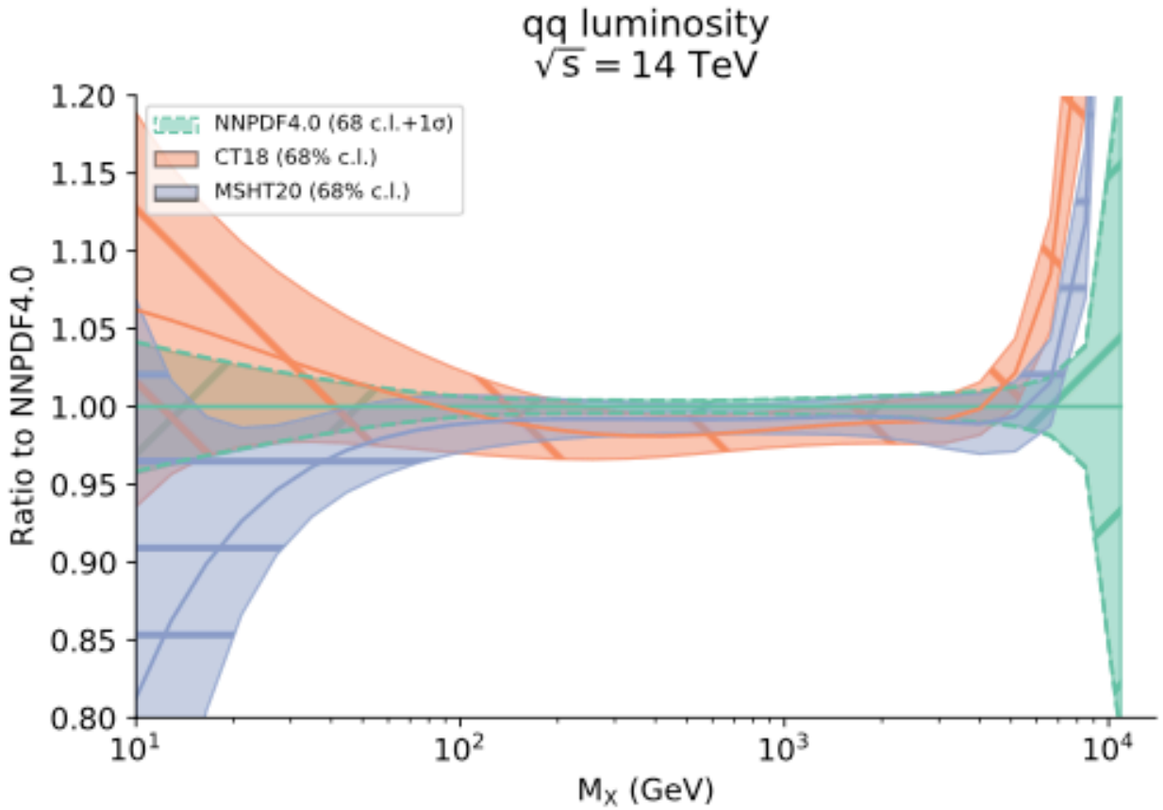


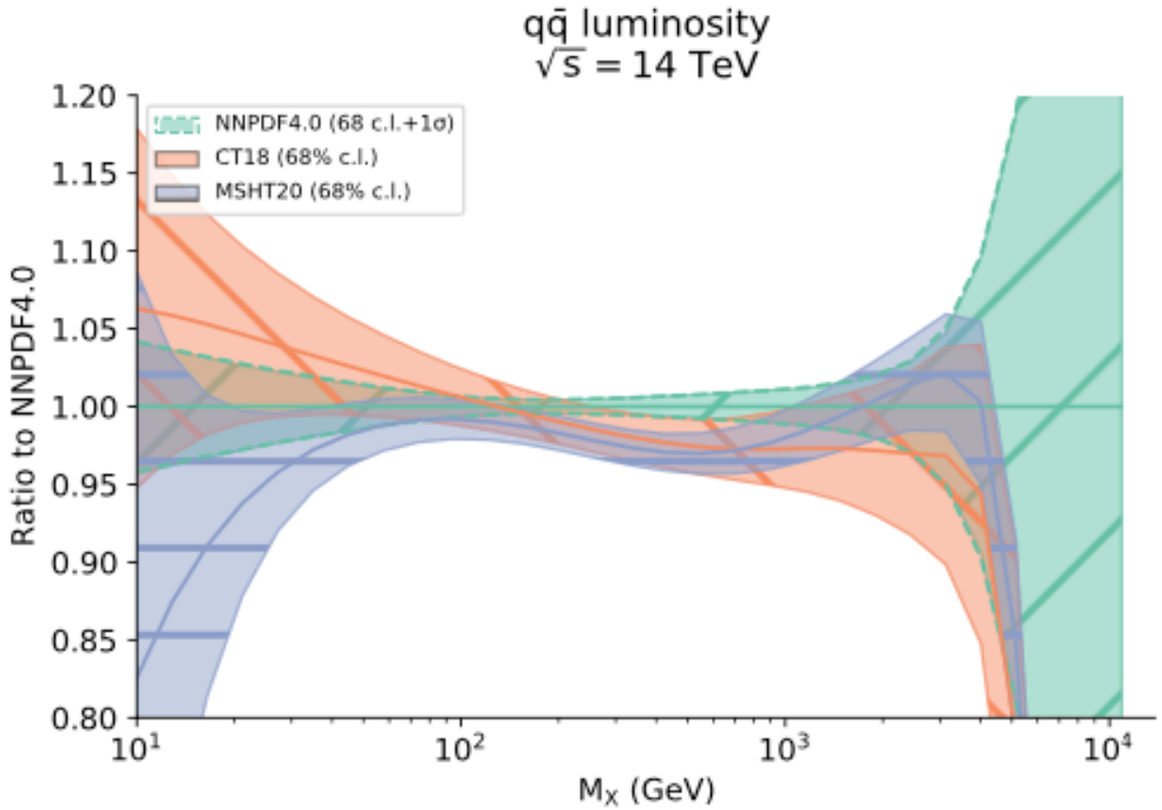


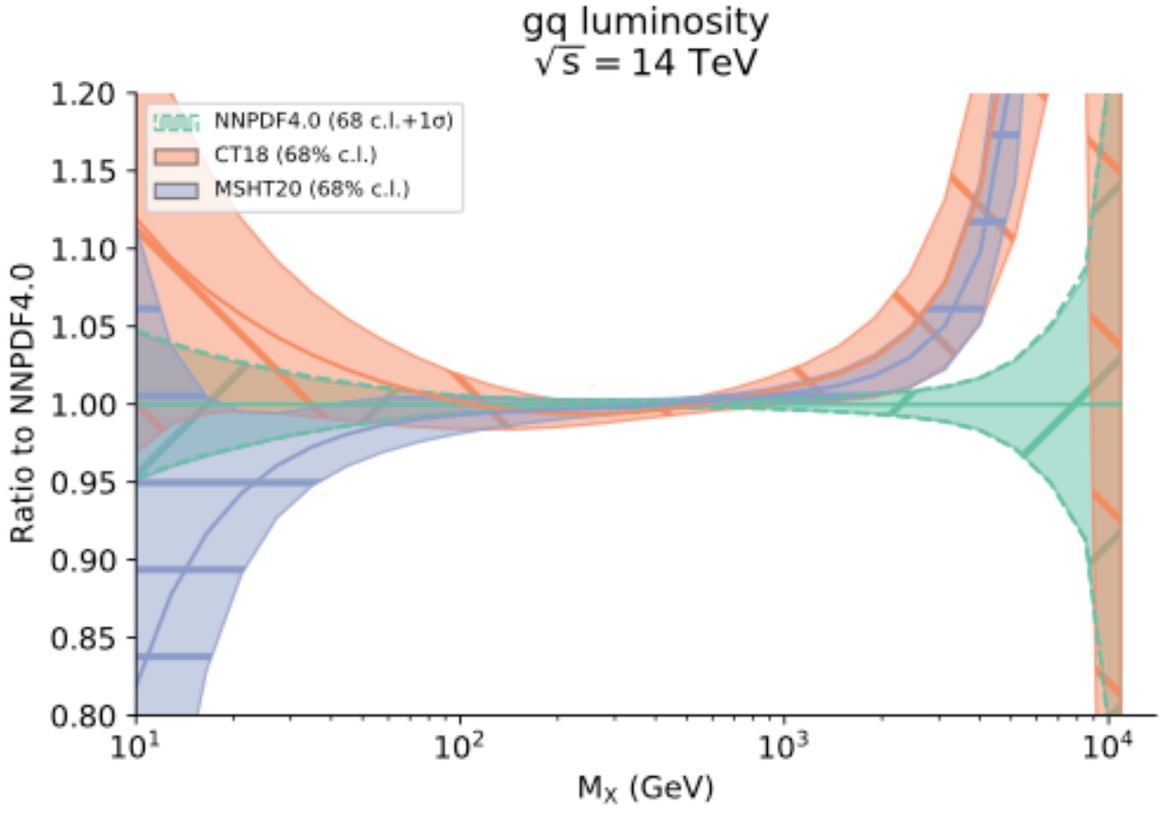


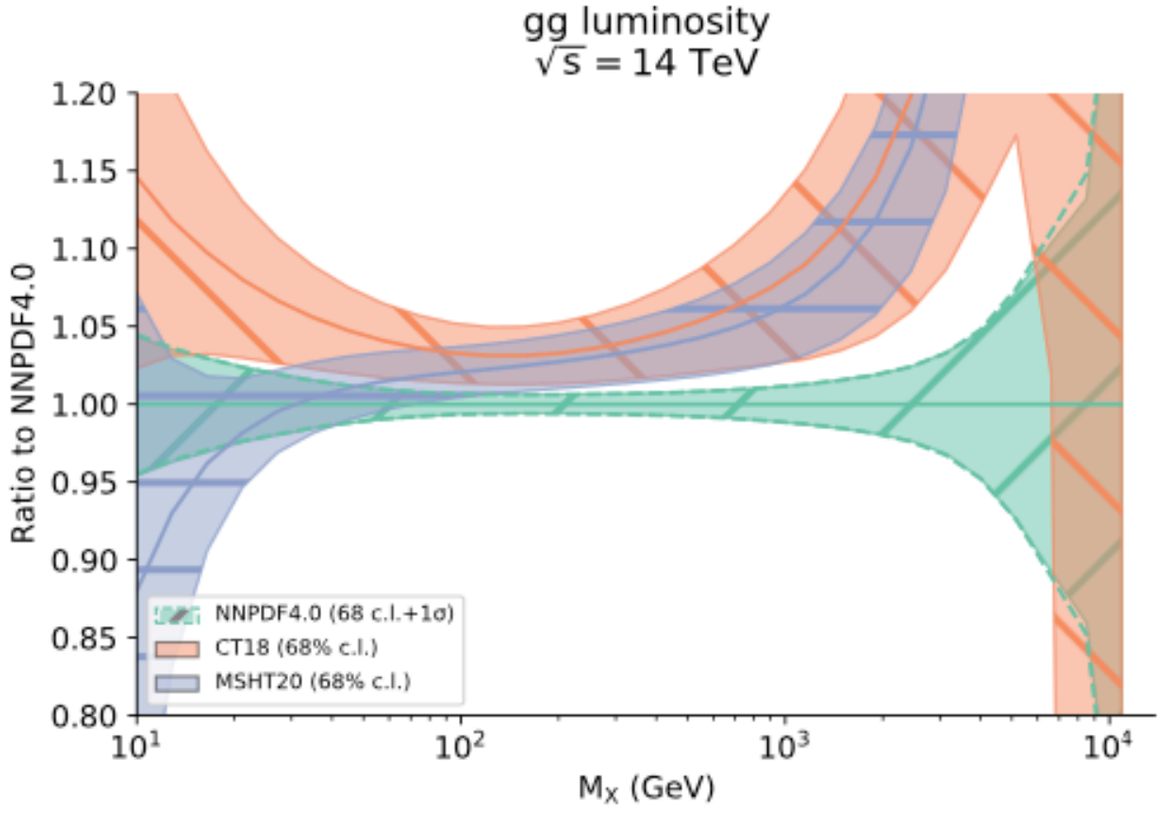
[https://arxiv.org/pdf/1510.03865.pdf]

A snapshot today



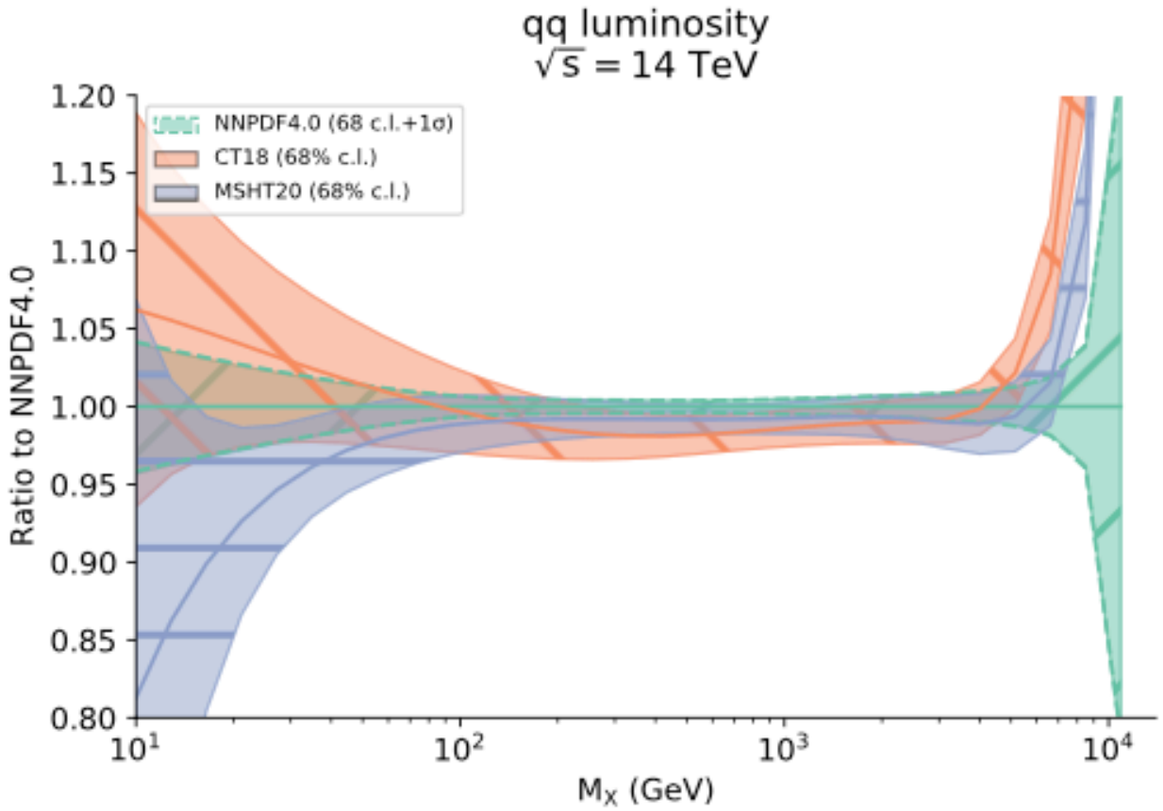


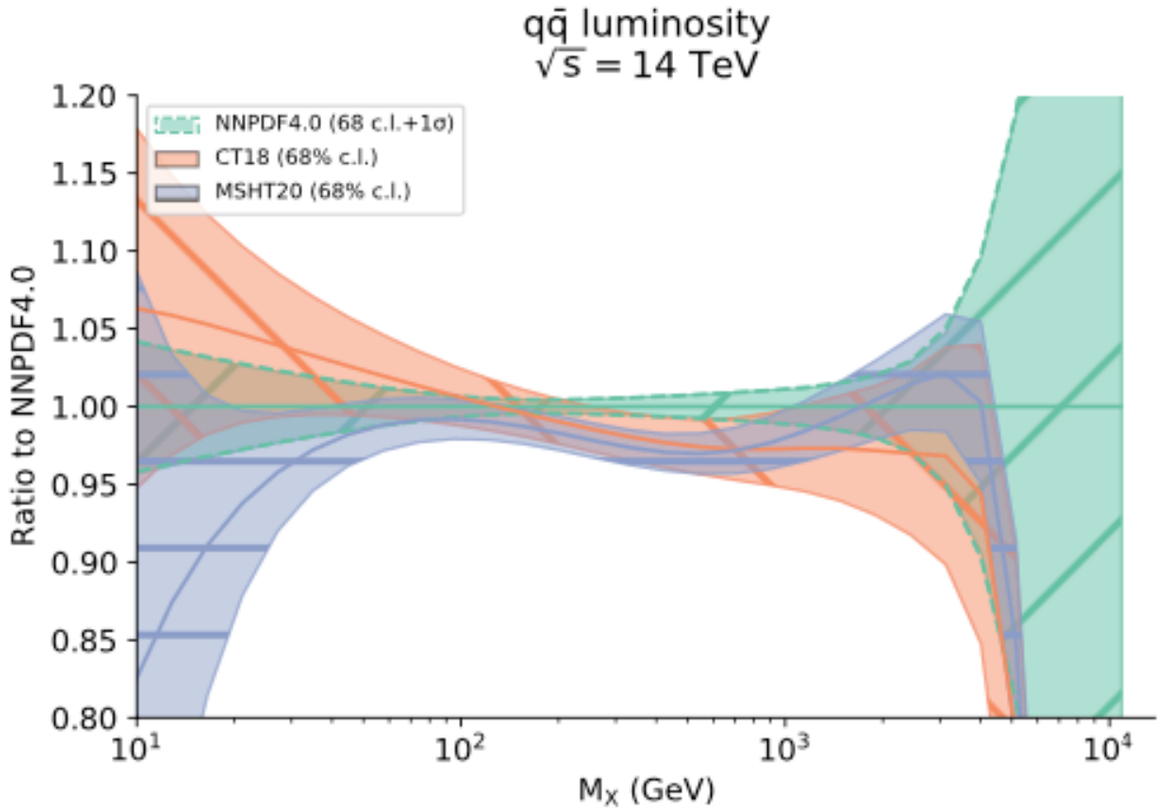


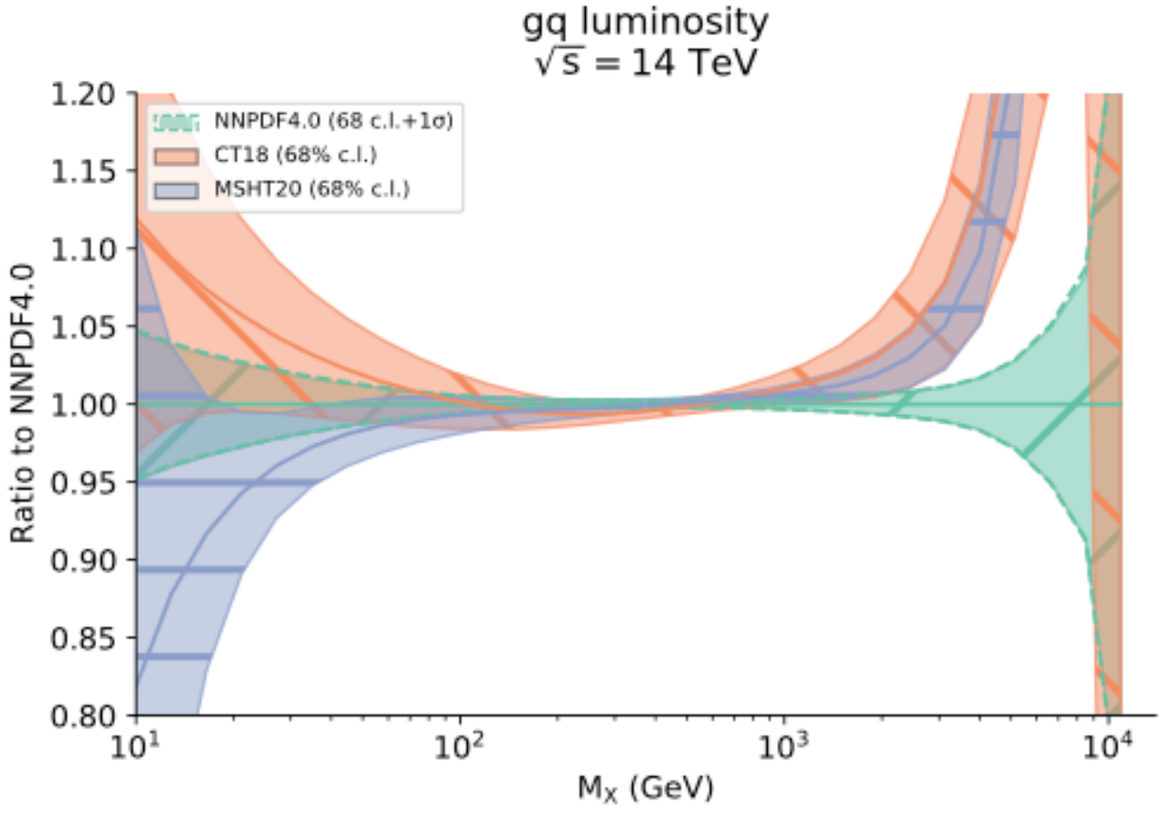


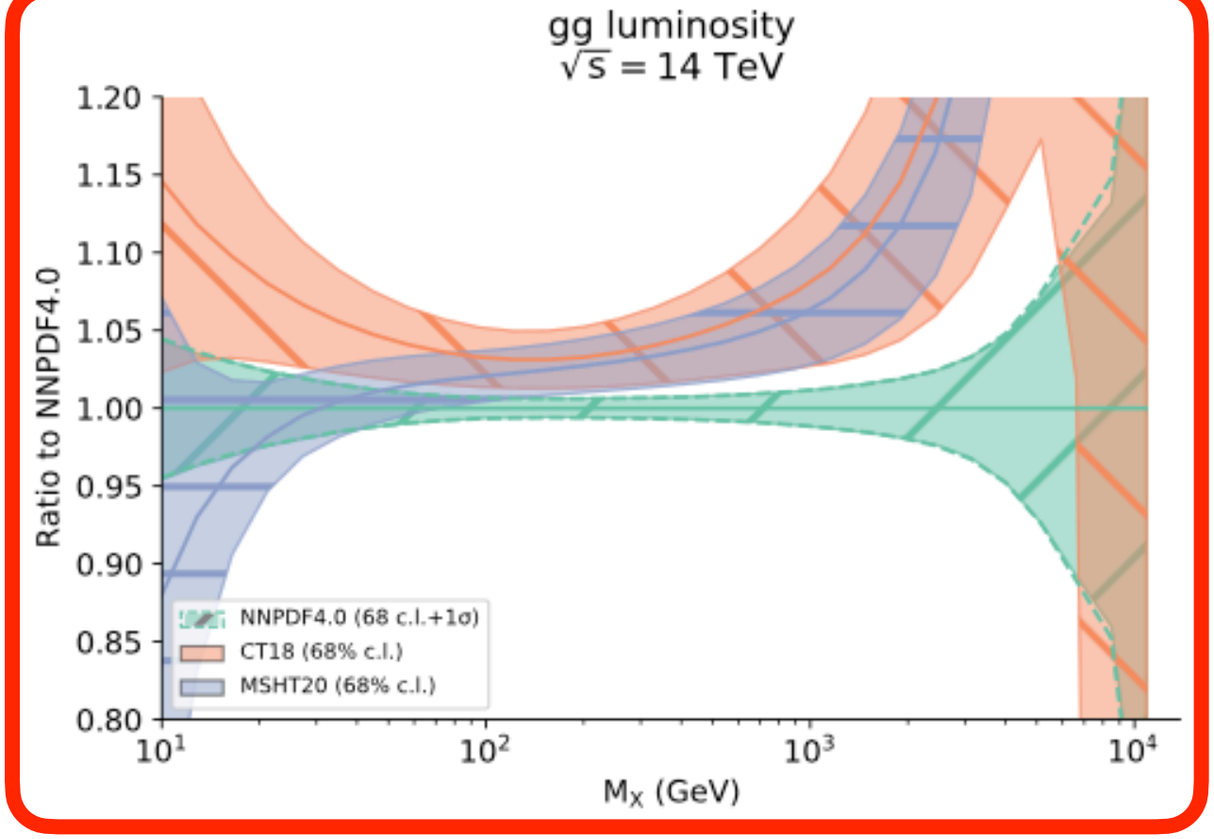
[E. Nocera's talk at DIS2021]

A snapshot today









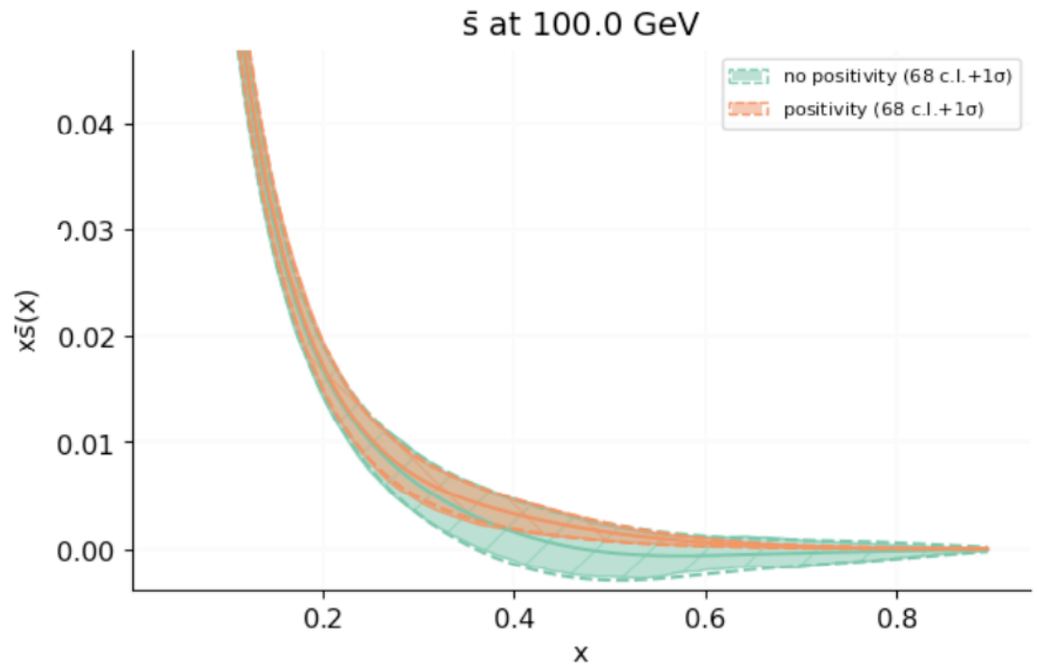
[E. Nocera's talk at DIS2021]

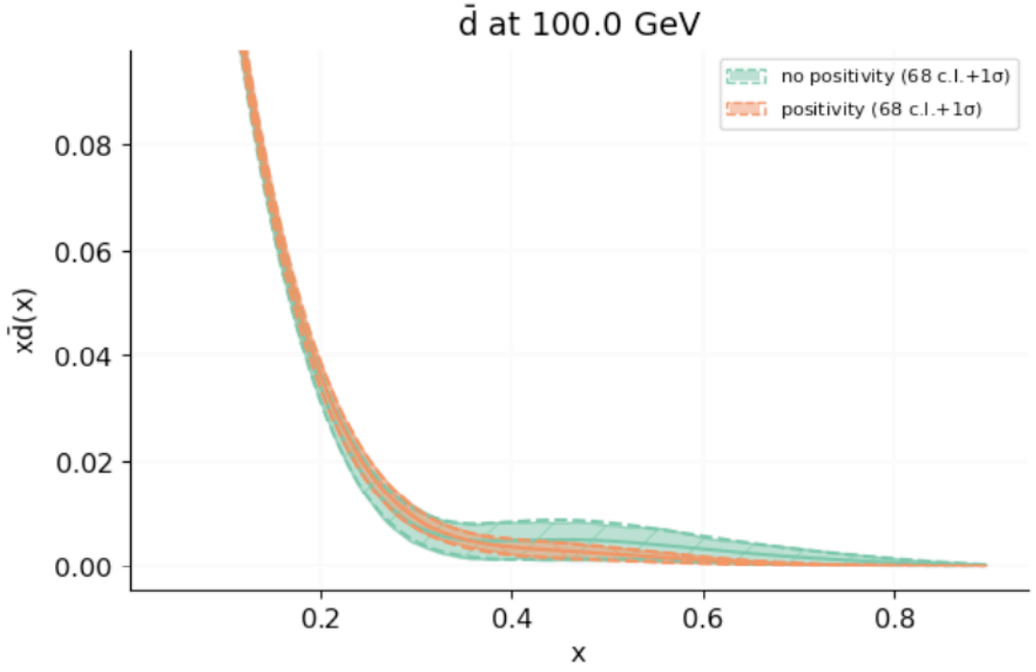
Positivity and PDFs

- PDFs have to such to guarantee the **positivity of cross sections**:
  - cross sections can be interpreted as probabilities $\Longrightarrow$ must be **positive**.

- Possible ways to enforce positivity are:
  1. determine PDFs enforcing that a **specific set of observables** is positive:
     [NNPDF, *JHEP* 04 (2015) 040]
     - does not guarantee *all* possible observables to be positive.
     - allows PDFs to be negative (sometimes unwanted, *e.g.* MC generators).
  2. Assume **PDFs** to be **positive definite** from the start:
     [CTEQ, *Phys.Rev.D* 103 (2021) 1, 014013]
     - does it really guarantee positivity of the observables?

- Positivity has a strong impact of PDFs:





[E. Nocera's talk at DIS2021]

Positivity and PDFs

- Recently it has been proposed that **PDFs** in $\overline{\text{MS}}$ ***are*** **positive**:

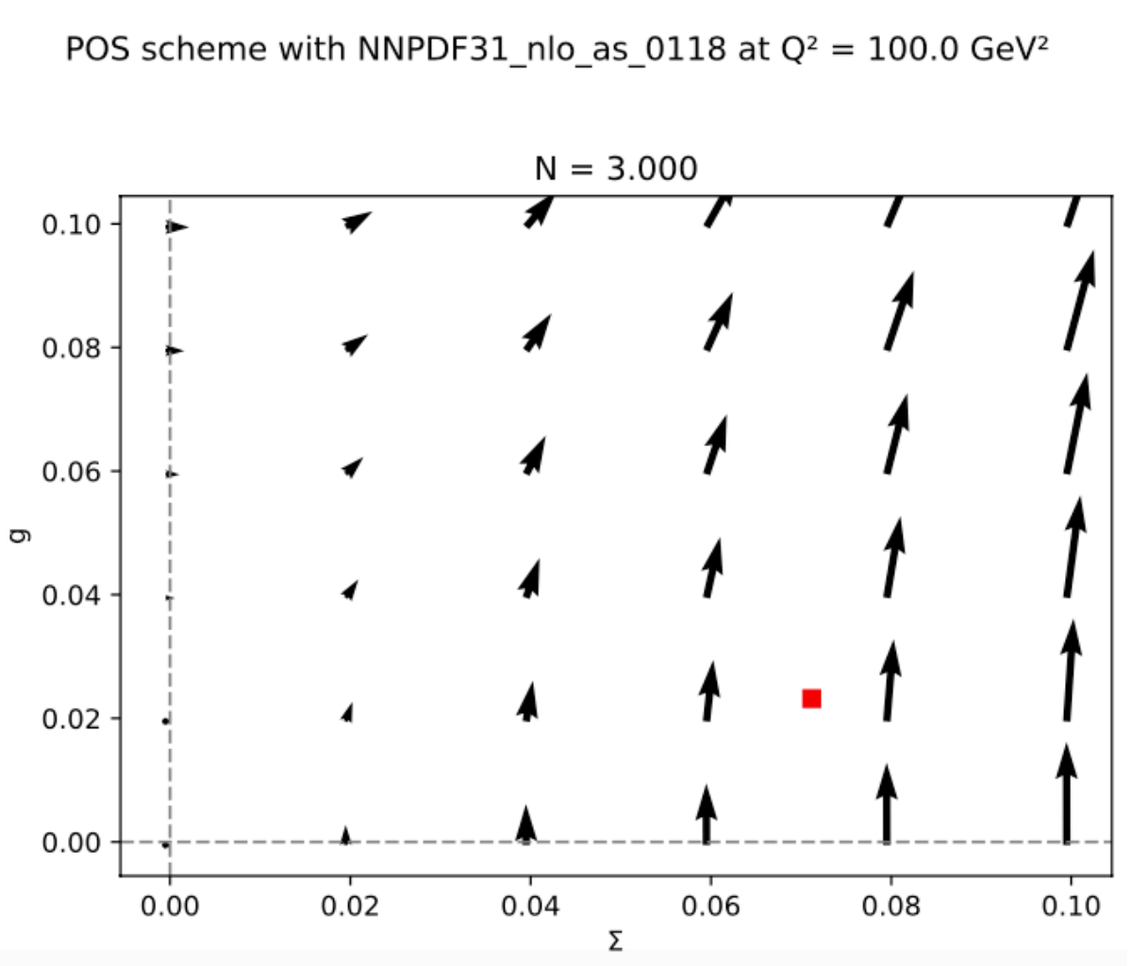
[Candido, Forte, Hekhorn, *JHEP* 11 (2020) 129]

> It is common lore that Parton Distribution Functions (PDFs) in the $\overline{\text{MS}}$ factorization scheme can become negative beyond leading order due to the collinear subtraction which is needed in order to define partonic cross sections. We show that this is in fact not the case and next-to-leading order (NLO) $\overline{\text{MS}}$ PDFs are actually positive in the perturbative regime. In order to prove this, we modify the subtraction prescription, and perform the collinear subtraction in such a way that partonic cross sections remain positive. This defines a factorization scheme in which PDFs are positive. We then show that positivity of the PDFs is preserved when transforming from this scheme to $\overline{\text{MS}}$, provided only the strong coupling is in the perturbative regime, such that the NLO scheme change is smaller than the LO term.

- Define an *ad hoc* factorisation scheme (for DIS) in which PDFs are positive (POS scheme).
- Find the transformation that gives $\overline{\text{MS}}$ PDFs in terms of the POS ones:

$$f^{\overline{\text{MS}}}(Q^2) = \left[\mathbb{I} + \frac{\alpha_s}{2\pi} K^{\text{POS}} \otimes\right]^{-1} f^{\text{POS}}(Q^2)$$

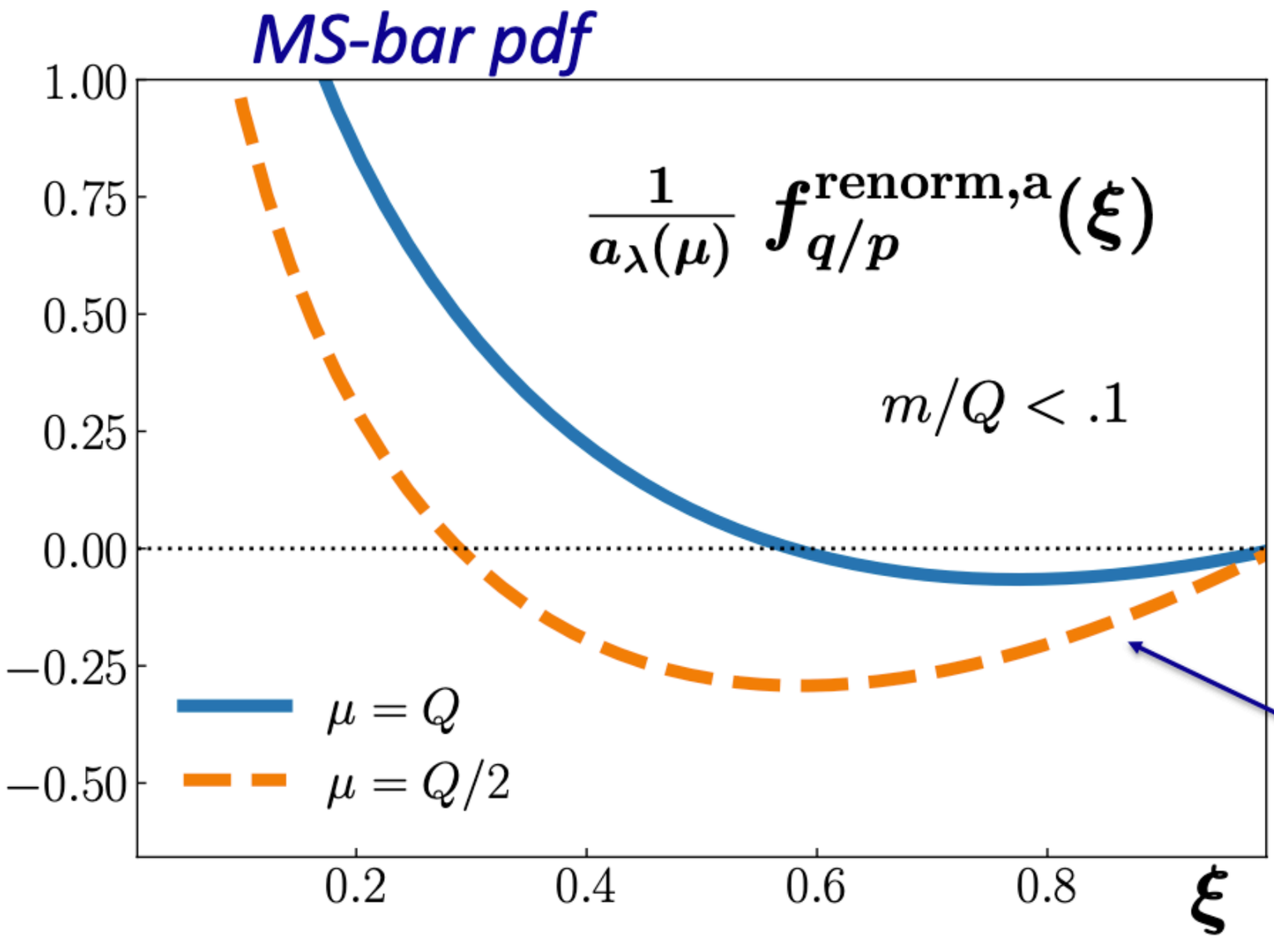
- The authors find that this transformation tends to make PDFs more positive.
- If POS PDFs are positive (by definition) $\Longrightarrow$ $\overline{\text{MS}}$ are to be even more positive.



Positivity and PDFs

- More recently though Collins, Rogers, and Sato have found an opposite result:
  - by direct computation of the PDF using its operator definition focusing on the removal of the **UV divergence**.

[Rogers, talk at QCD Evolution 2021]



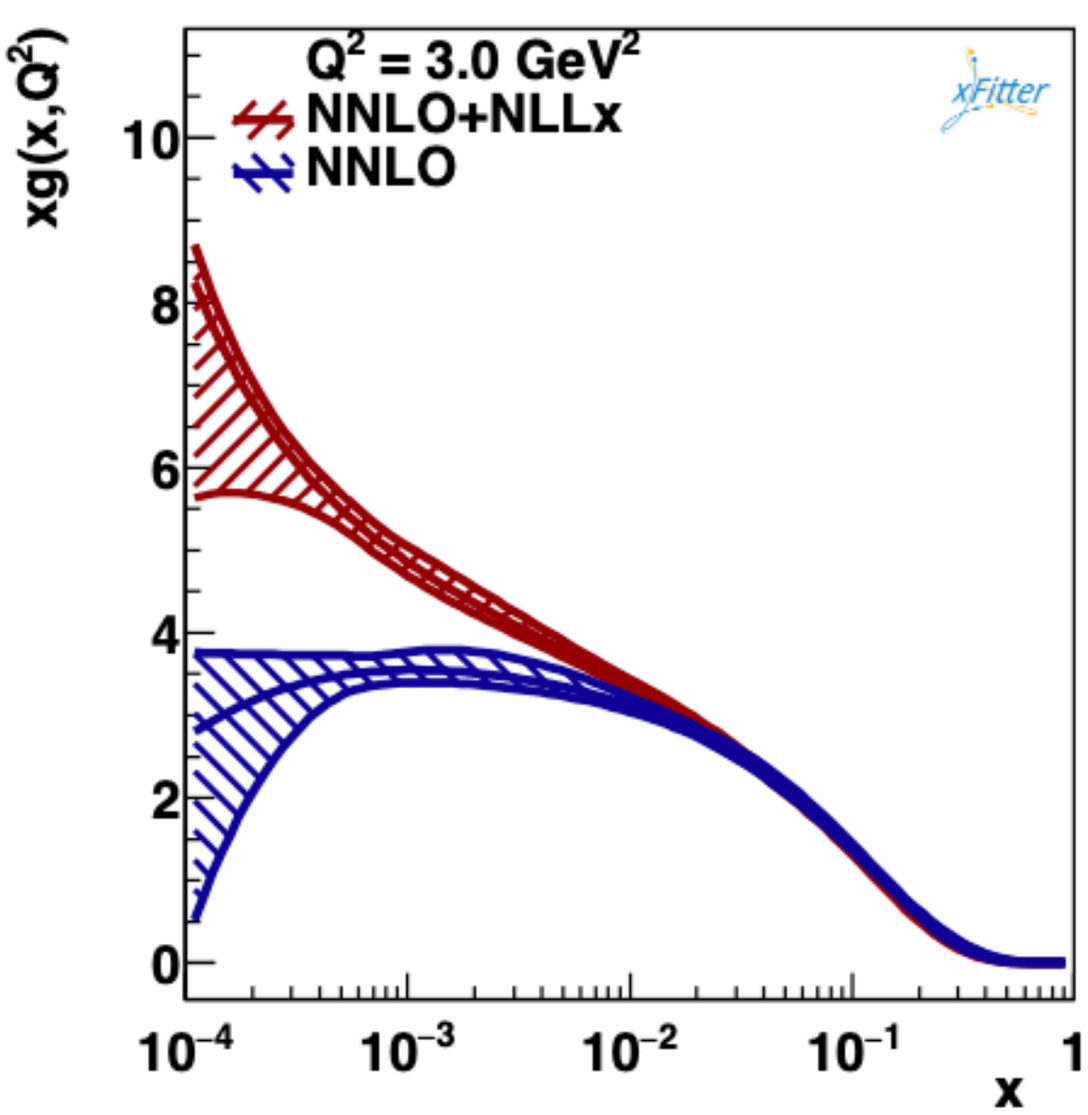
- Nothing forces the pdf to be strictly positive, even for relatively large Q.

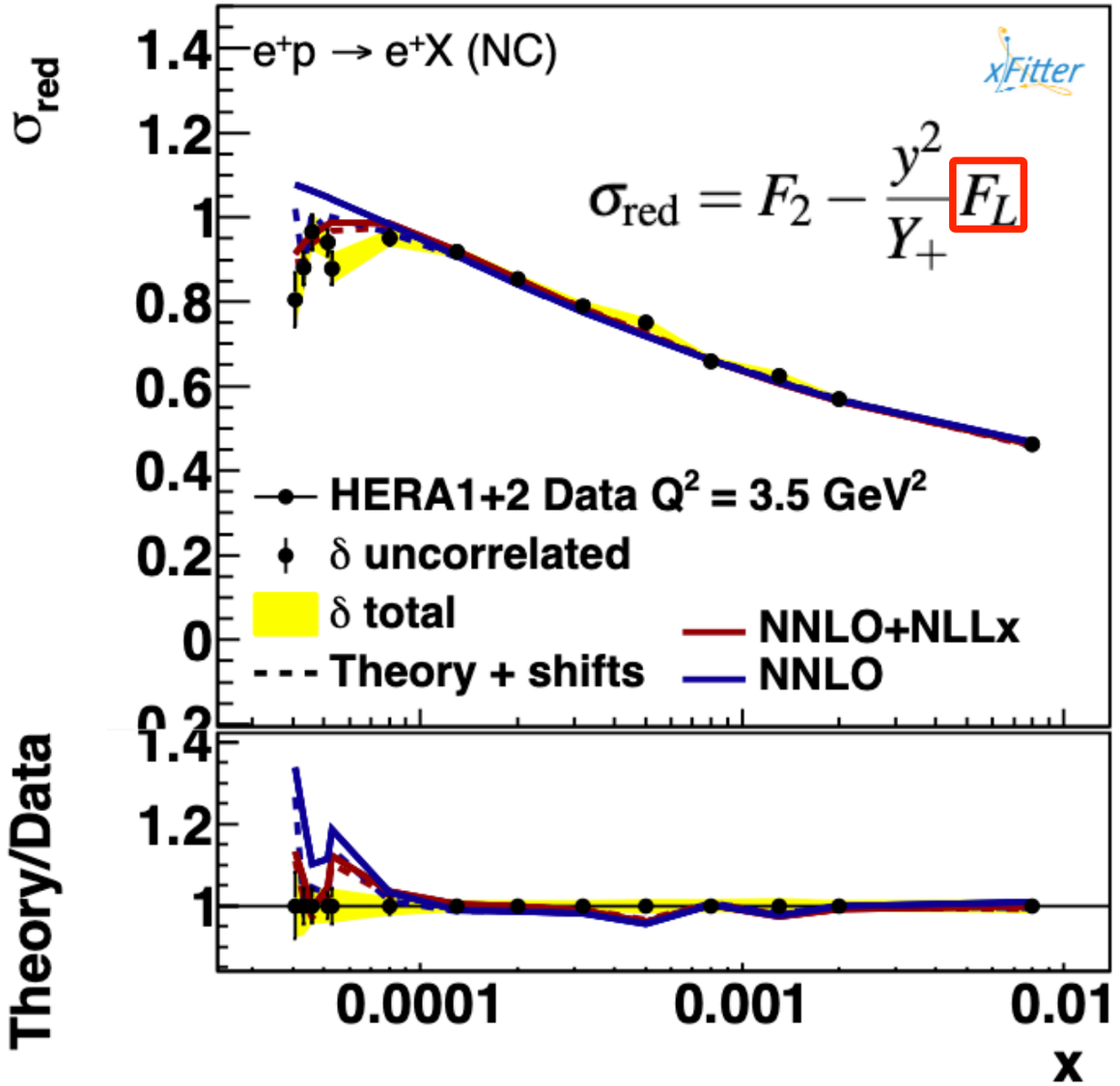
- The question remains open: are $\overline{\text{MS}}$ PDFs allowed to go negative?

Small-x resummation

[NNPDF, *Eur.Phys.J.C* 78 (2018) 4, 321]
[xFitter, *Eur.Phys.J.C* 78 (2018) 8, 621]

- The issue of the NLO low-x gluon PDF going negative at low scales is greatly mitigated by including **small-x (BFKL) resummation** effects in PDF fits:
  - relevant for **quarkonium** production at the LHC, [Lansberg, Ozcelik, *Eur.Phys.J.C* 81 (2021) 6, 497]
- Small-x resummation makes the DGLAP evolution is **less steep** and thus allows for a larger small-x gluon PDF that behaves as a sea-like distribution.





Thank you