On the extraction of collinear distributions

Valerio Bertone

IRFU, CEA, Université Paris-Saclay

January 13, 2023, Aussois

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

Everything starts with...

A **collinear factorisation theorem**:

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

Everything starts with...

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.

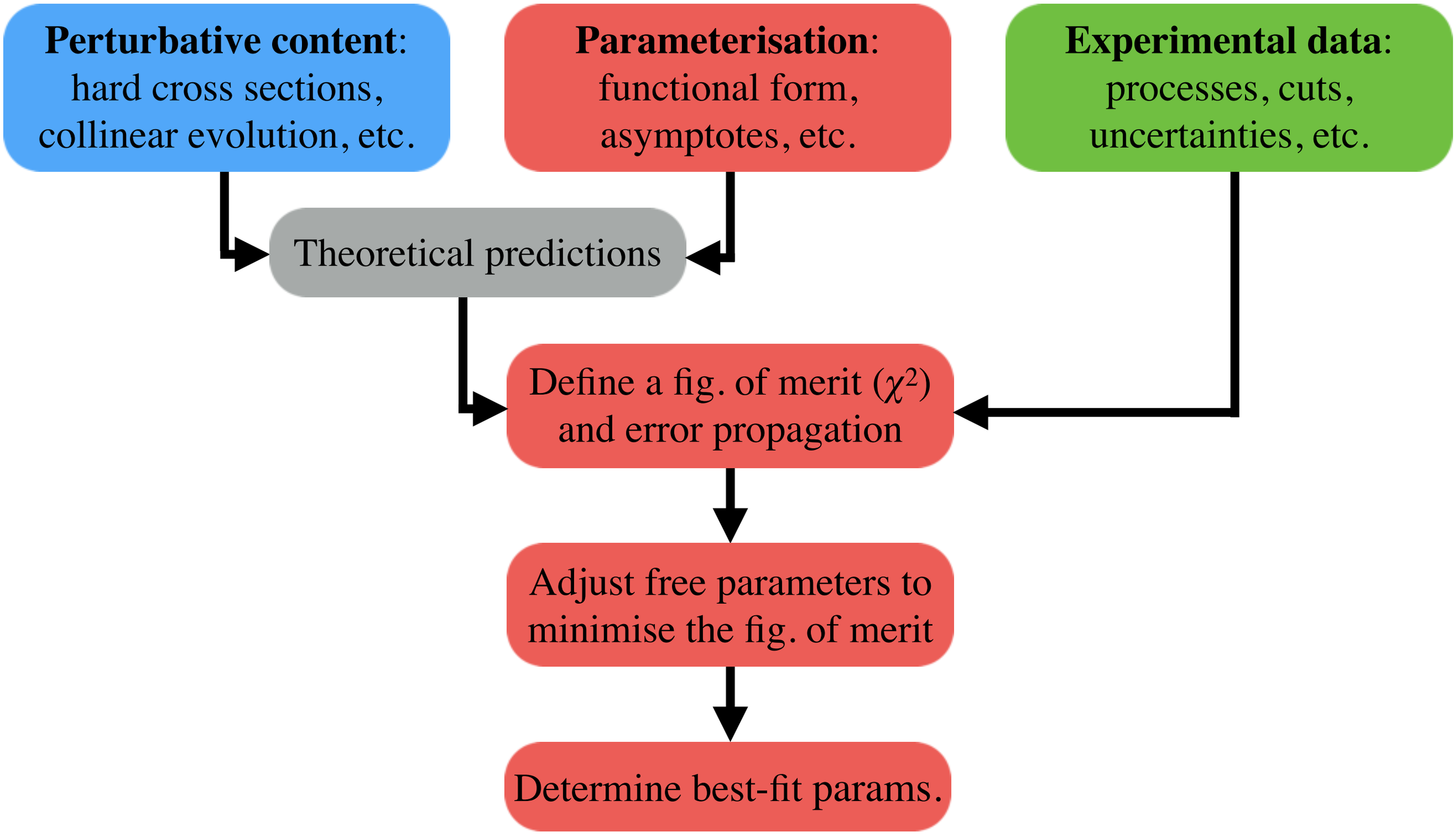
Collinear distributions (PDFs/FFs):

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

How do we determine collinear distributions?

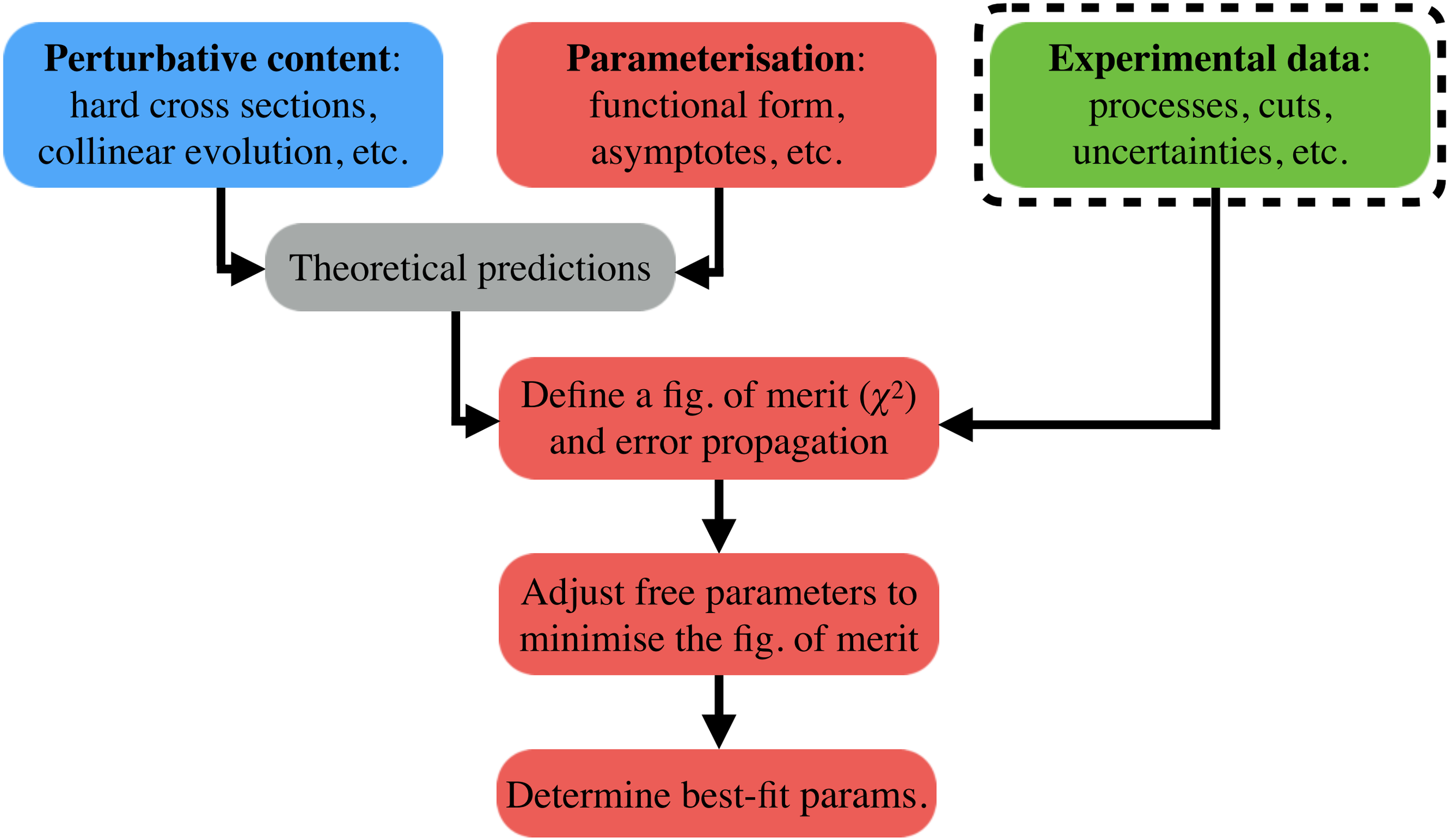
Currently, the most accurate and reliable way is through **fits to data**.

The general fit strategy

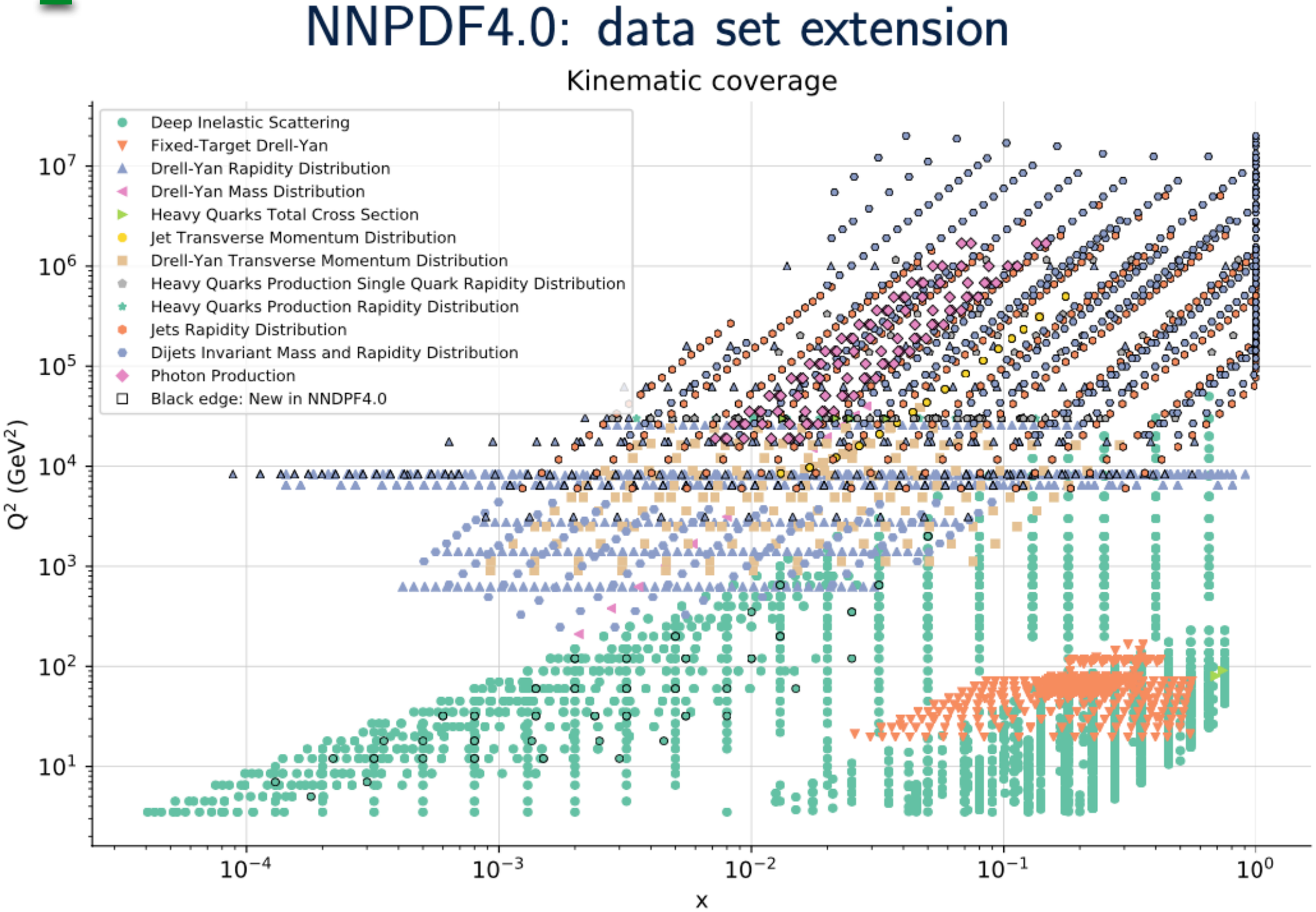


Each box requires a choice. **Different choices** lead to **different determinations**.

The general fit strategy



Experimental data

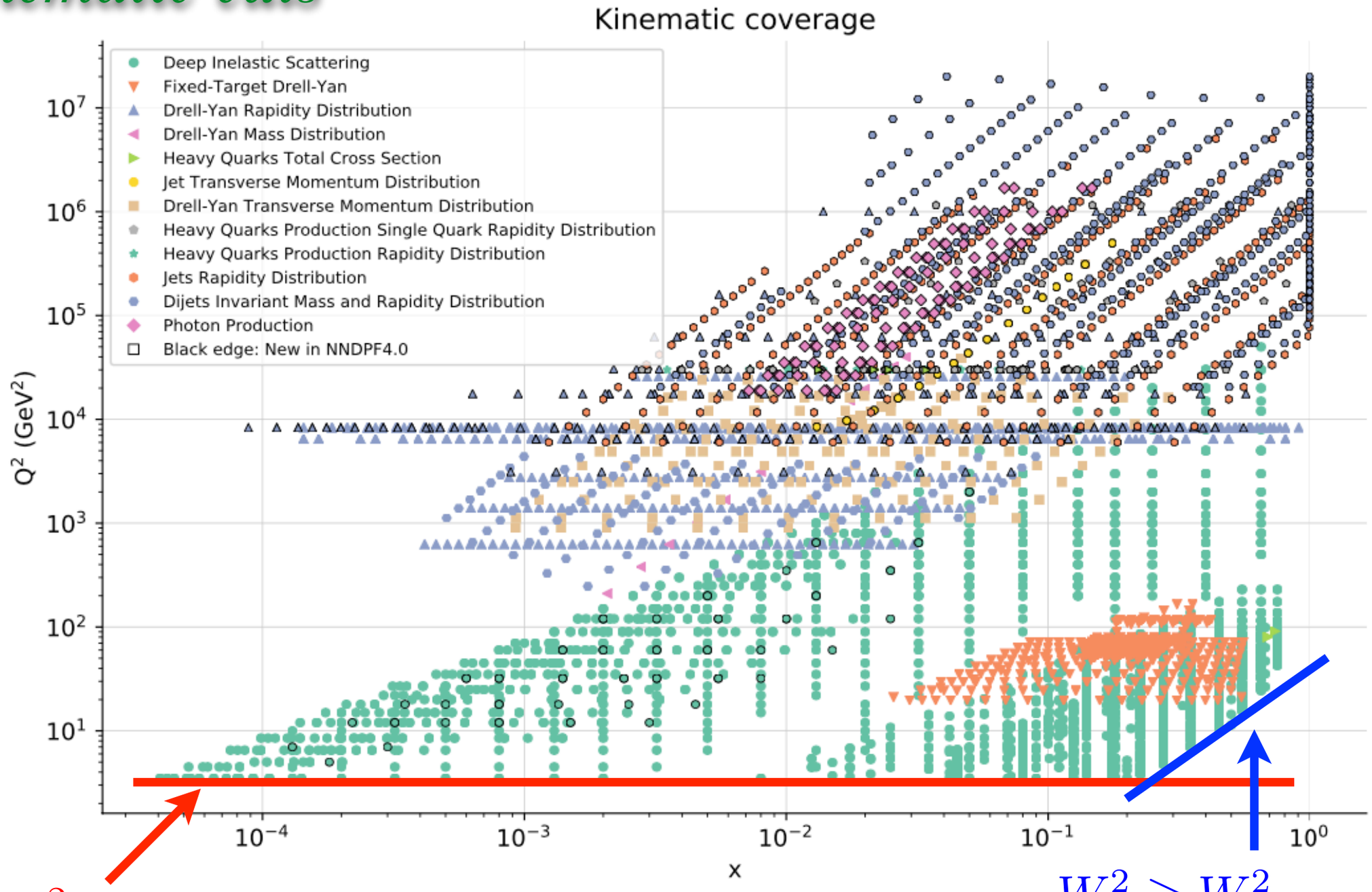


[NNPDF4.0, arXiv:2109.02653]

- Impressive number of processes and data points (O(4500)):
  - *ep* data, *i.e.* DIS (one PDF involved),
  - *pp* data (two PDFs involved),
  - very wide kinematic coverage: $10^{-5} \lesssim x \lesssim 1$ and 2 GeV $\lesssim Q \lesssim$ 5 TeV.

Experimental data

Kinematic cuts



[NNPDF4.0, arXiv:2109.02653]

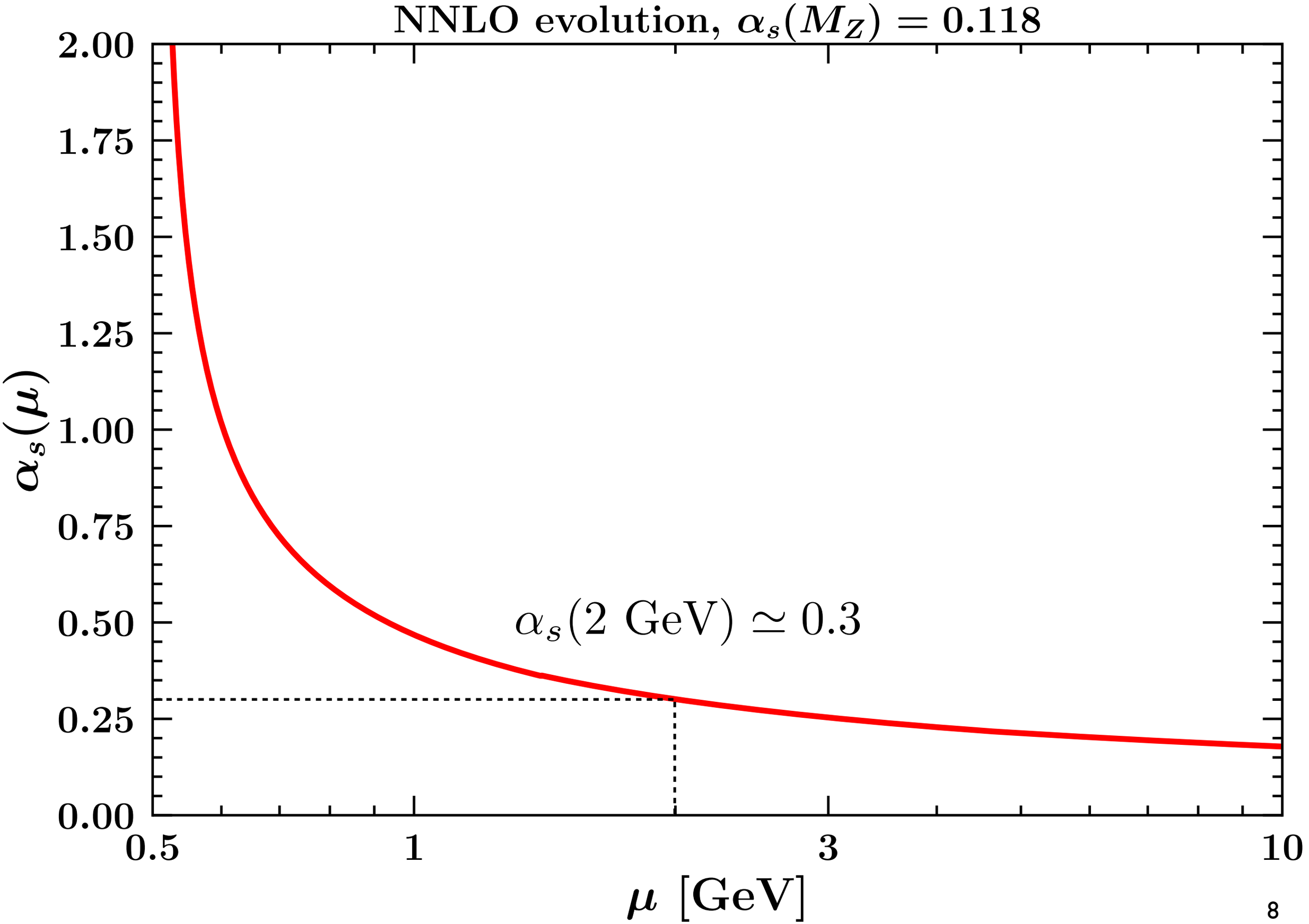
$Q^2 \geq Q^2_{\rm min}$ $(Q_{\rm min} \simeq 2$ GeV$)$

Ensure that perturbation theory is applicable

$W^2 \geq W^2_{\rm min}$

$(W^2 = \frac{Q^2(1-x)}{x}, \quad W_{\rm min} \simeq 4$ GeV$)$

Remove higher-twist contributions (mostly for DIS and Fixed-target DY data)


NNLO evolution, $\alpha_s(M_Z) = 0.118$
$\alpha_s(\mu)$
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.5
1
3
10
μ [GeV]
$\alpha_s(2\ \mathrm{GeV}) \simeq 0.3$

Experimental data

Kinematic cuts

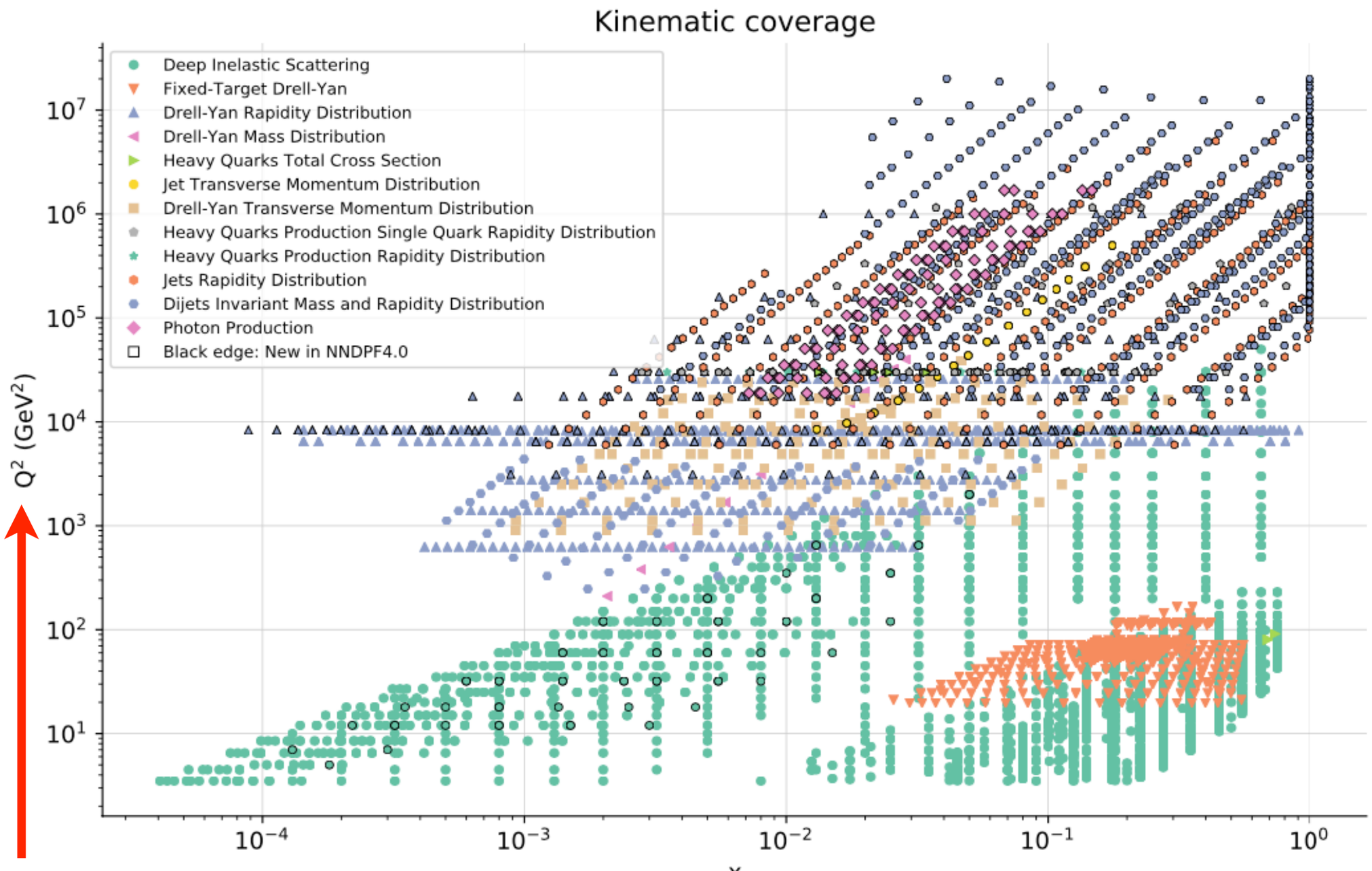
- Usually, many more cuts are enforced…

Dataset	NLO	NNLO
DIS measurements	$W^2 \geq 12.5$ GeV^2; $Q^2 \geq 3.5$ GeV^2	$W^2 \geq 12.5$ GeV^2; $Q^2 \geq 3.5$ GeV^2
HERA I+II $\sigma^c_{\rm NC}$ (in addition to the above)	—	$Q^2 \geq 8$ GeV^2 (fitted charm)
E866/E605 σ^p	$\tau \leq 0.08$; $\lvert y/y_{\rm max}\rvert \leq 0.0663$	$\tau \leq 0.08$; $\lvert y/y_{\rm max}\rvert \leq 0.0663$
D0 W electron/muon asymmetry	—	$\lvert A_\ell\rvert \geq 0.03$
ATLAS low-mass DY 7 TeV	$m_{\ell\ell} > 22$ GeV	—
ATLAS high-mass DY 7 TeV	$m_{\ell\ell} < 210$ GeV	$m_{\ell\ell} < 210$ GeV
CMS DY 2D 7 TeV	$30 \leq m_{\ell\ell} \leq 200$ GeV; $\lvert y_{\ell\ell}\rvert \leq 2.2$	$m_{\ell\ell} \leq 200$ GeV; $\lvert y_{\ell\ell}\rvert \leq 2.2$
LHCb $W, Z \to \mu$ 7 TeV	—	$\lvert\eta_\mu\rvert/\lvert y_{\mu\bar{\mu}}\rvert \geq 2.25$
ATLAS low-mass DY 2D 8 TeV	$m_{\ell\ell} \leq 116$ GeV	$m_{\ell\ell} \leq 116$ GeV
LHCb $W, Z \to \mu$ 8 TeV	—	$\lvert\eta_\mu\rvert/\lvert y_{\mu\bar{\mu}}\rvert \geq 2.25$
LHCb $Z \to ee/Z \to \mu\mu$ 13 TeV	—	$\lvert y_{\ell\ell}\rvert \geq 2.25$
ATLAS $W^\pm$+jet 8 TeV	$p_T^W \geq 25$ GeV	$p_T^W \geq 25$ GeV
ATLAS Z p_T 8 TeV $(p_T, m_{\ell\ell})$	$p_T^Z \geq 30$ GeV	$p_T^Z \geq 30$ GeV
ATLAS Z p_T 8 TeV (p_T, y_Z)	$30 \leq p_T^Z \leq 150$ GeV	$30 \leq p_T^Z \leq 150$ GeV
CMS Z p_T 8 TeV	$30 \leq p_T^Z \leq 170$ GeV; $\lvert y_Z\rvert \leq 1.6$	$30 \leq p_T^Z \leq 170$ GeV; $\lvert y_Z\rvert \leq 1.6$
CMS incl. jets 8 TeV	$p_T^{\rm jet} \geq 74$ GeV	$p_T^{\rm jet} \geq 74$ GeV

[NNPDF4.0, arXiv:2109.02653]

Experimental data

How to read the scatter plot



[NNPDF4.0, arXiv:2109.02653]

Hard scale $\simeq$ factorisation scale
(*i.e.* scale at which PDFs are computed)

Kinematic lower bound

Experimental data

How to read the scatter plot

- Inclusive DIS: $e(k) + p(P) \rightarrow e(k') + X$

$$q = k - k' , \qquad Q = \sqrt{-q^2} , \qquad x_{\rm B} = \frac{Q^2}{2q \cdot P}$$

$$\frac{d\sigma}{dx_{\rm B} dQ} \propto \int_{x_{\rm B}}^1 \frac{dy}{y} C\left(\frac{x_{\rm B}}{y}, \alpha_s(Q)\right) f(y, Q)$$

- Drell-Yan: $p(P_1) + p(P_2) \rightarrow e^+(k) + e^-(k') + X$

$$\begin{aligned} q &= k + k' \\ s &= (P_1 + P_2)^2 \\ y &= \tfrac{1}{2} \ln \tfrac{q_0 + q_3}{q_0 - q_3} \end{aligned} \quad , \quad M = \sqrt{q^2} , \quad \tau = \frac{M^2}{s} , \quad x_{1,2} = \sqrt{\tau}\, e^{\pm y}$$

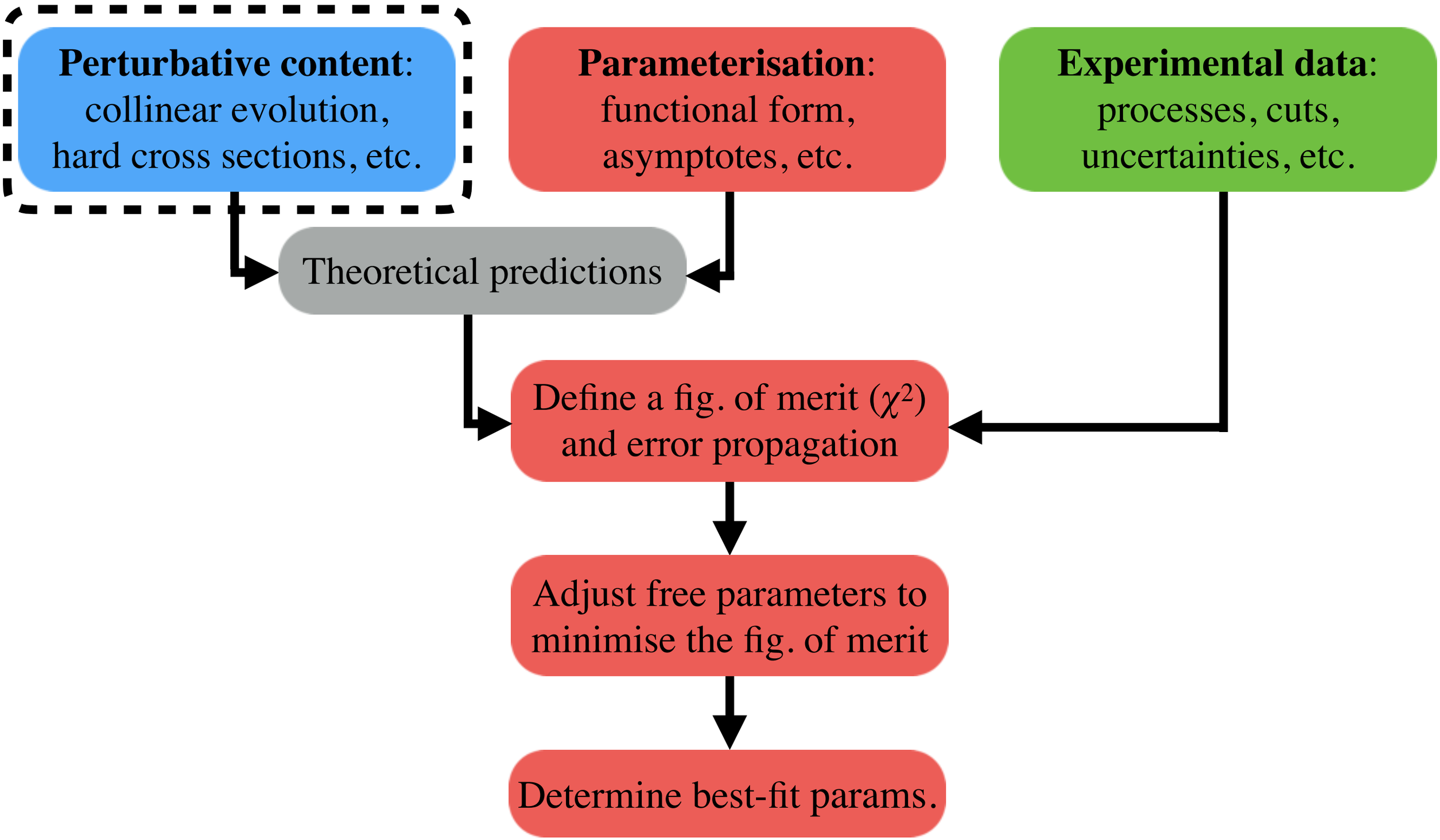
- Mass-differential distribution:

$$\frac{d\sigma}{dM} \propto \int_{\tau}^1 \frac{dt}{t} C\left(\frac{\tau}{t}, \alpha_s(Q)\right) \int_t^1 \frac{dy}{y} f_1(y, Q) f_2\left(\frac{t}{y}, Q\right)$$

- Rapidity-differential distribution:

$$\frac{d\sigma}{dy dM} \propto \int_{x_1}^1 \frac{dy_1}{y_1} \int_{x_2}^1 \frac{dy_2}{y_2} C\left(\frac{x_1}{y_1}, \frac{x_2}{y_2}, \alpha_s(Q)\right) f_1(y_1, Q) f_2(y_2, Q)$$

The general fit strategy



Perturbative content

DGLAP evolution

- PDFs obey a set of **coupled integro-differential equations**: the DGLAP equations

$$\frac{df_i(x,\mu)}{d\ln\mu^2} = \sum_j \int_x^1 \frac{dy}{y} P_{ij}\left(y, \alpha_s(\mu)\right) f_j\left(\frac{x}{y}, \mu\right)$$

$$P_{ij}\left(y, \alpha_s(\mu)\right) = \sum_{n=0}^{\infty} \alpha_s^{n+1}(\mu) P_{ij}^{(n)}(y)$$

- Unpolarised splitting functions fully known up to NNLO.
  Moch, Vermaseren, Vogt [*Nucl.Phys.B* 688 (2004) 101-134] [*Nucl.Phys.B* 691 (2004) 129-181]
- Several numerical implementations:
  - x-space approach:
    - QCDNUM, Botje [arXiv:1602.08383]
    - HOPPET, Salam [arXiv:0804.3755]
    - APFEL(++), Bertone, Carrazza, Rojo [arXiv:1310.1394], Bertone [arXiv1708.00911]
  - $\mathcal{N}$-space approach:
    - PEGASUS, Vogt [hep-ph/0408244]
    - MELA, Bertone, Carrazza, Nocera [arXiv:1501.00494]
    - EKO. Candido, Hekhorn, Magni [arXiv:2202.02338]

Perturbative content

DGLAP evolution: x-space approach

- The DGLAP equations reduce to a set of coupled **linear ordinary differential equations** (ODEs) that in vectorial notation read:

$$\frac{d\mathbf{f}(\mu)}{d\ln\mu^2} = \mathbf{\Gamma}(\mu)\cdot\mathbf{f}(\mu) \equiv \mathbf{F}(\mu,\mathbf{f})$$

- The functions to be determined are the partonic distribution f as a function of μ on each node of the grid knowing $\mathbf{f}(\mu_0)$ on the grid.

- Many algorithms to solve ODEs exist. A particularly popular choice is the **4th order Runge-Kutta** (RK4):

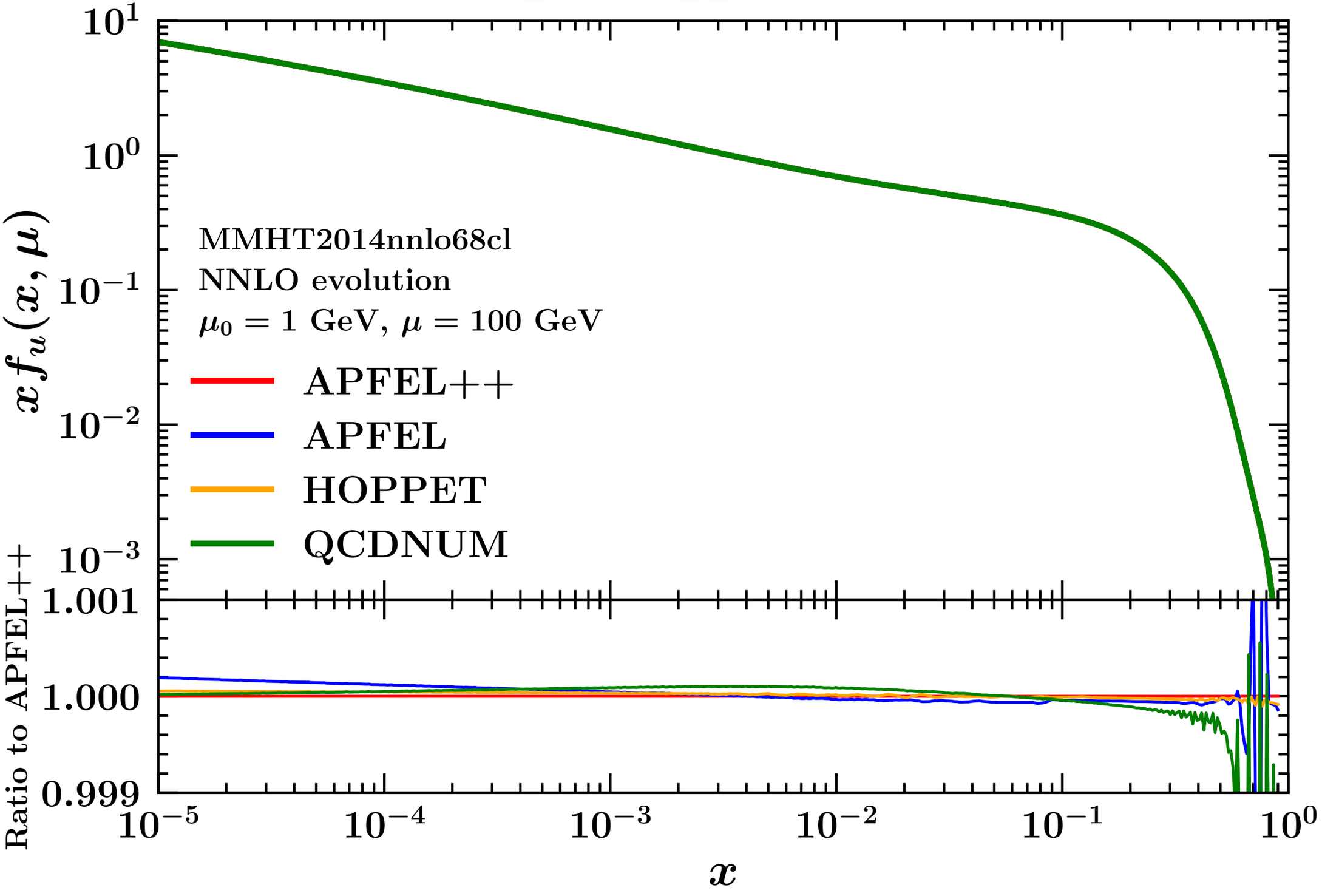
$$\begin{aligned}
\mathbf{f}_{n+1} &= \mathbf{f}_n + \tfrac{1}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4) + \mathcal{O}(h^5)\\
\mathbf{k}_1 &= h\mathbf{F}(\mu_n, \mathbf{f}_n)\\
\mathbf{k}_2 &= h\mathbf{F}(\mu_n + \tfrac{h}{2}, \mathbf{f}_n + \tfrac{\mathbf{k}_1}{2})\\
\mathbf{k}_3 &= h\mathbf{F}(\mu_n + \tfrac{h}{2}, \mathbf{f}_n + \tfrac{\mathbf{k}_2}{2})\\
\mathbf{k}_3 &= h\mathbf{F}(\mu_n + h, \mathbf{f}_n + \mathbf{k}_3)
\end{aligned}$$

- Tabulate $\mathbf{f}$ on a **grid in μ** that can successively interpolated.

- Final result: $f(x,\mu)$ is know on a 2D grid in x and μ.

Perturbative content

DGLAP evolution: x-space approach



- Independent implementations agree well below the per-mil level!

Perturbative content

DGLAP evolution: N-space approach

- Mellin tranform:

$$f(N) \equiv \mathcal{M}\left[f(x)\right](N) = \int_0^1 dx\, x^{N-1} f(x)$$

- turns a Mellin convolution:

$$[f \otimes g](x) \equiv \int_x^1 \frac{dy}{y} f(y) g\left(\frac{x}{y}\right) = \int_x^1 \frac{dz}{z} f\left(\frac{x}{z}\right) g(z) = \int_0^1 dy \int_0^1 dz \delta(x - yz) f(y) g(z)$$

- into a simple product:

$$[f \otimes g](N) = f(N) g(N)$$

- The DGLAP in Mellin space becomes an ODE:

$$\frac{df(N,\mu)}{d\ln\mu^2} = P(N, \alpha_s(\mu)) f(N,\mu)$$

- A fully analytic solution can be found changing the evolution variable:

$$\frac{d\alpha_s(\mu)}{d\ln\mu^2} = \beta(\alpha_s(\mu)) \quad \Rightarrow \quad \frac{df(N,\mu(\alpha_s))}{d\alpha_s} = \frac{P(N,\alpha_s)}{\beta(\alpha_s)} f(N,\mu(\alpha_s))$$

Perturbative content

DGLAP evolution: N-space approach

- Also the computation of some observables benefits of this simplification in Mellin space.

- Inclusive DIS is a typical example:

$$F(x,Q) = [C \otimes f](x,Q) \quad \Rightarrow \quad F(N,Q) = C(N,\alpha_s(Q))f(N,Q)$$

- This reduces the computation of an observable to a simple product that is thus very fast to compute numerically.

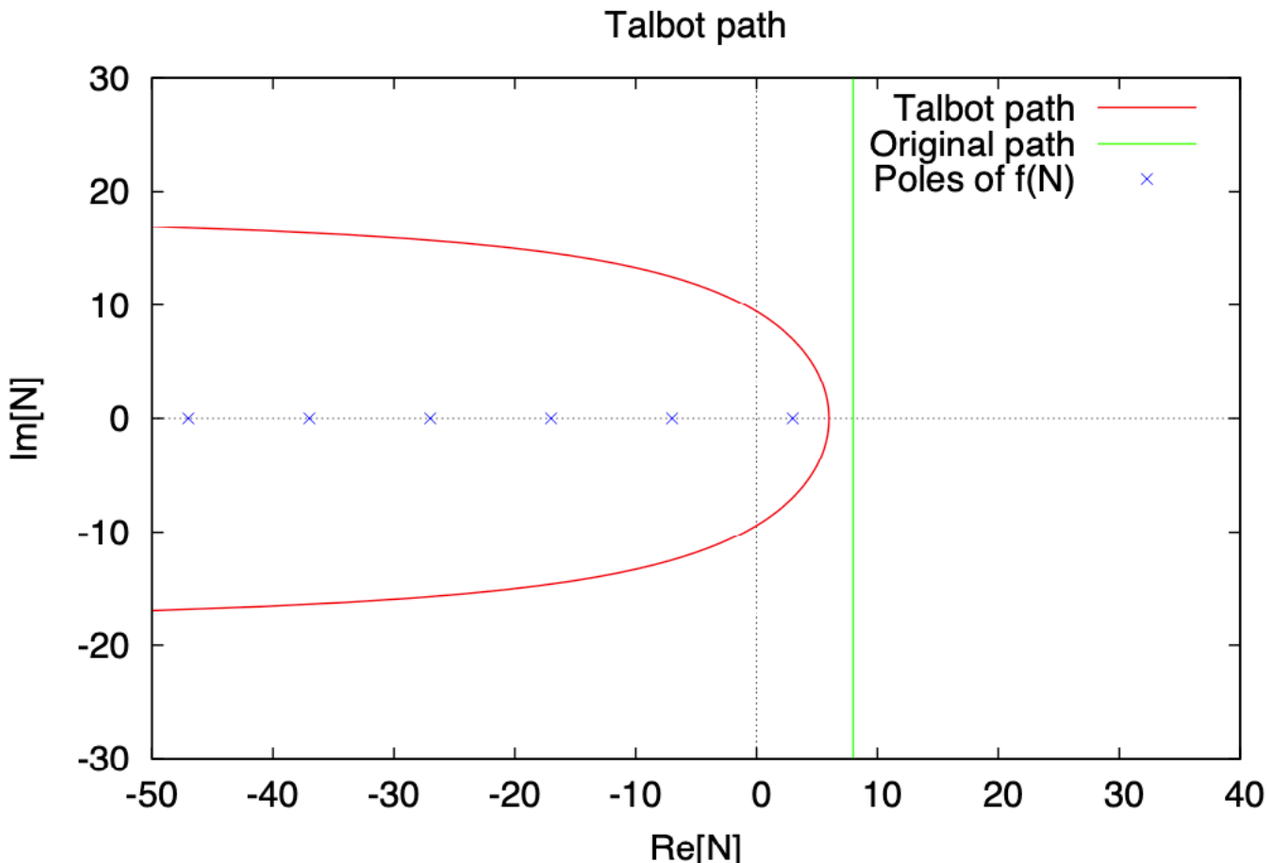
Perturbative content

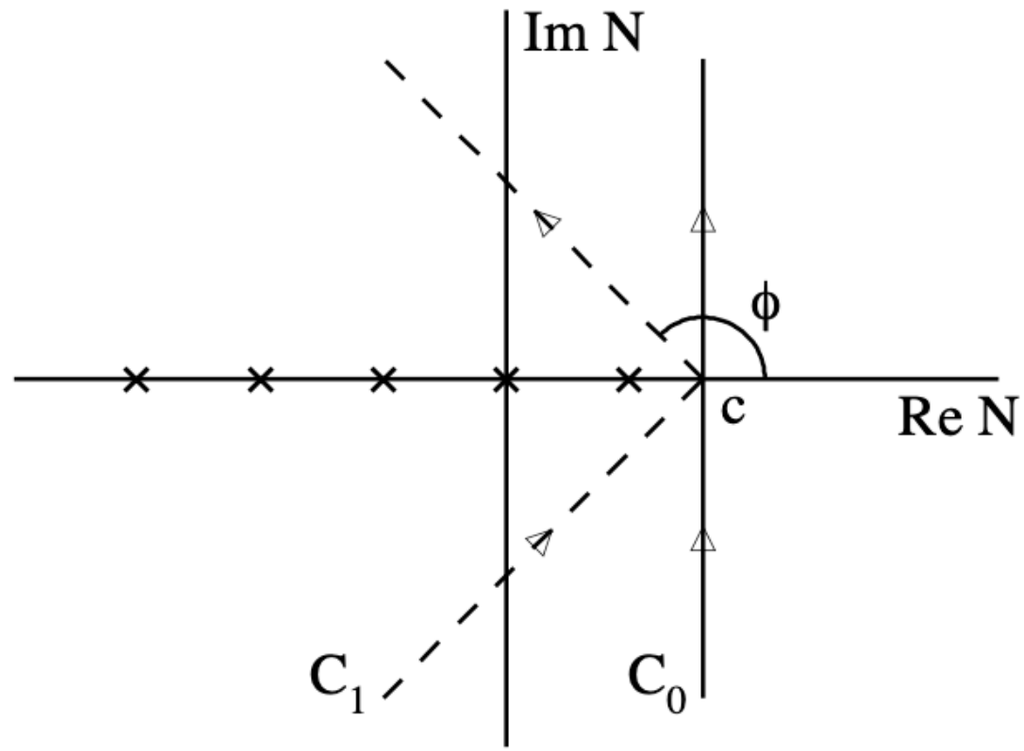
DGLAP evolution: N-space approach

- Data is however delivered in x space and it is therefore needed to transform back from N to x-space:

$$f(x) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} dN\, x^{-N} f(N)$$

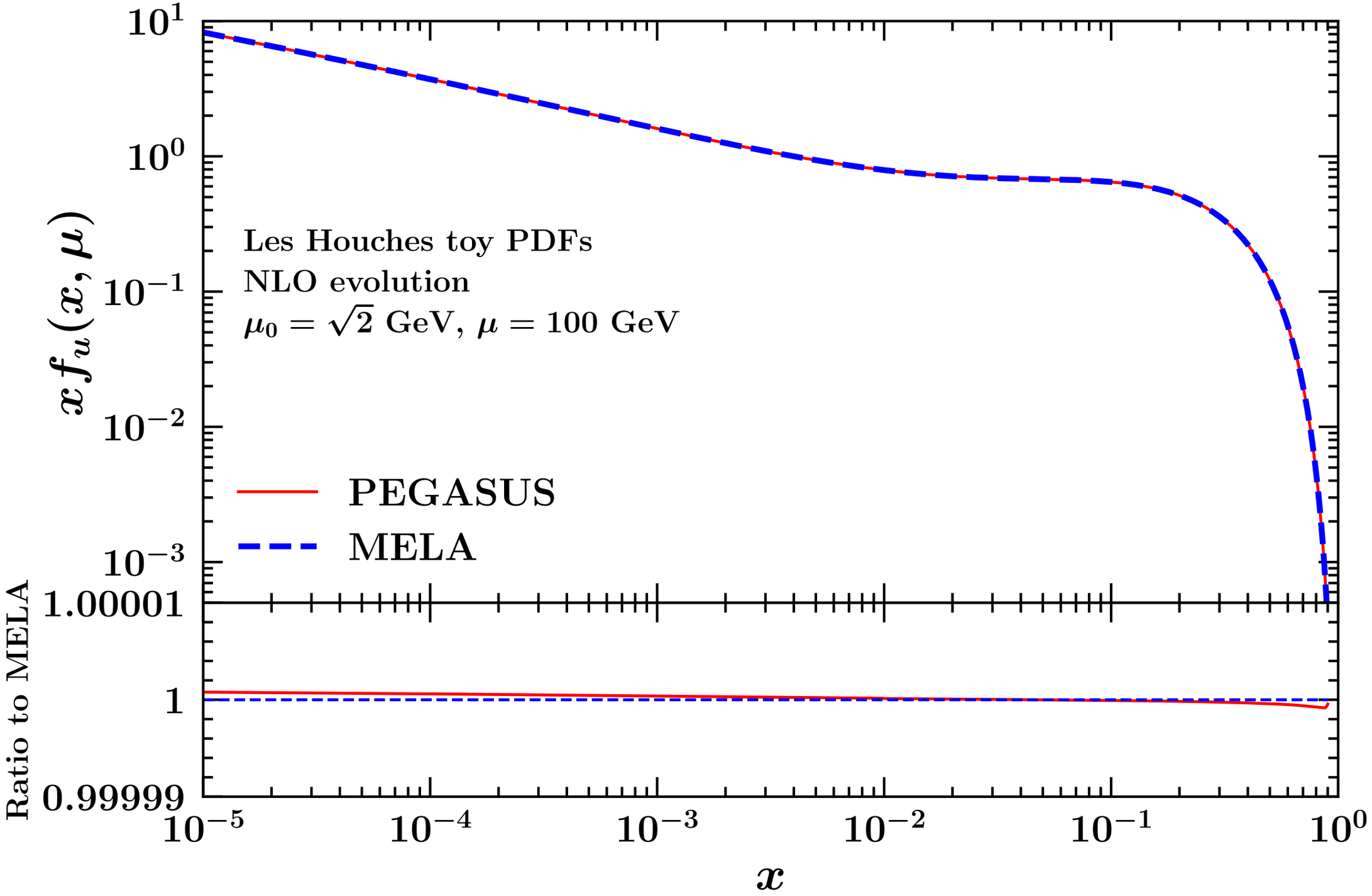
- where c is a real constant such that the integration contour lies to the right of all the singularities of the integrand:
- the integration contour is typically deformed for numerical implementations.





Perturbative content

DGLAP evolution: N-space approach



- Independent implementations agree extremely well!

Perturbative content

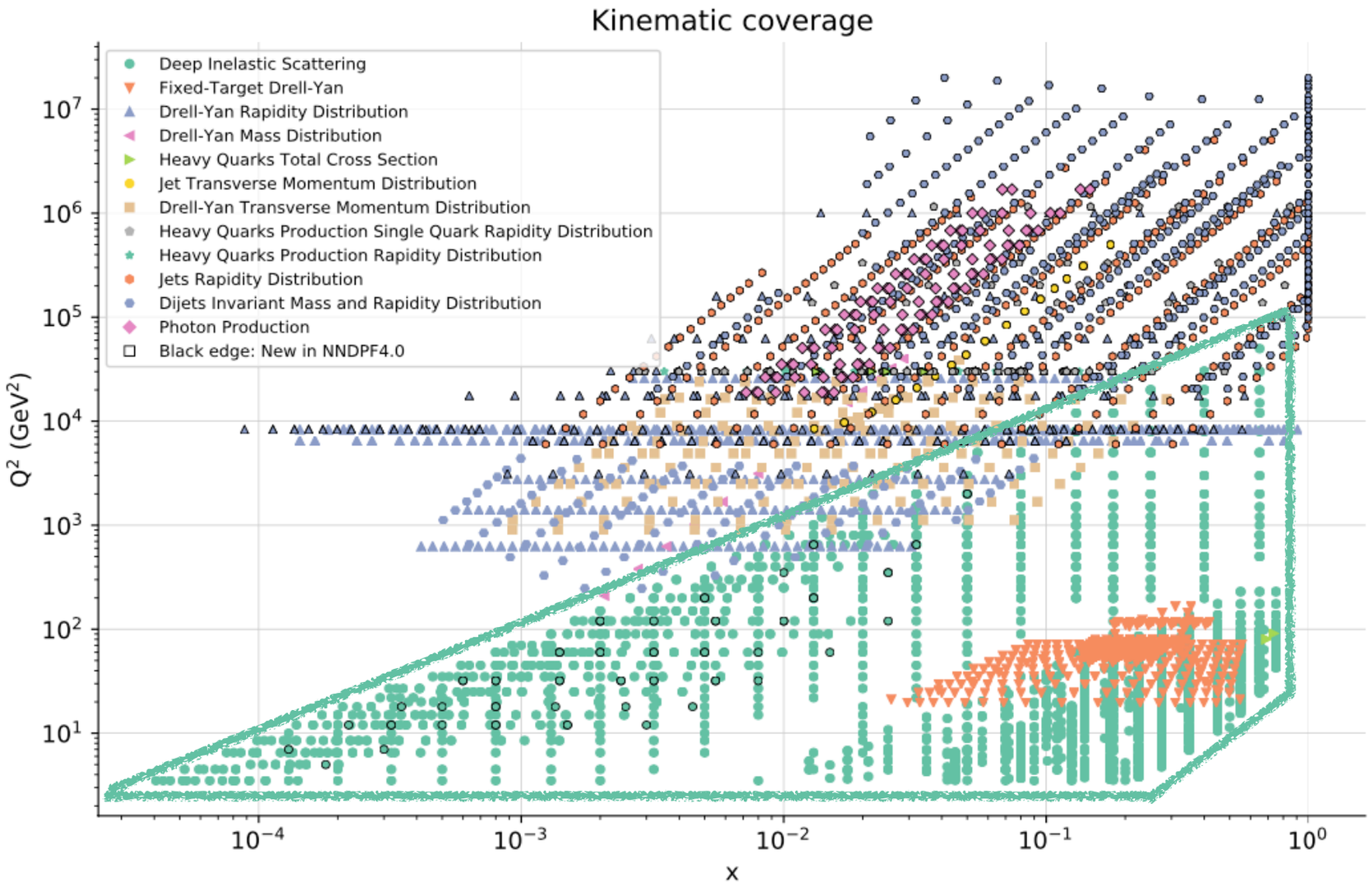
DGLAP evolution: N-space approach

- While being so numerically appealing, the N-space approach is affected by a few shortcomings:
  - the PDF parameterisation needs to be analytically transformable in Mellin space:
    - this drastically restricts the range of possible parameterisations.
  - The partonic cross sections need to be analytically known in N space:
    - this is the case only for a small number of inclusive processes and observables (*e.g.* DIS, transverse-momentum-integrated DY, and not much more).
    - experimental cuts typically make the computation of analytic Mellin transforms unfeasible.
- As of today, no large PDF collaboration uses the N-space approach as discussed above:
  - NNPDF has adopted in the past and is now planning to adopt again a "hybrid" technology that combines the N-space approach with the x-space one.

Perturbative content

Hard cross sections: inclusive DIS

- A large part of the dataset of modern PDF fits is still made of DIS data:



- Very broad kinematic coverage:
  - **low values of Q** comparable with m_c and m_b → **heavy-quark mass corrections**,
  - **low values of x → small-x resummation corrections**.

Perturbative content

Hard cross sections: inclusive DIS

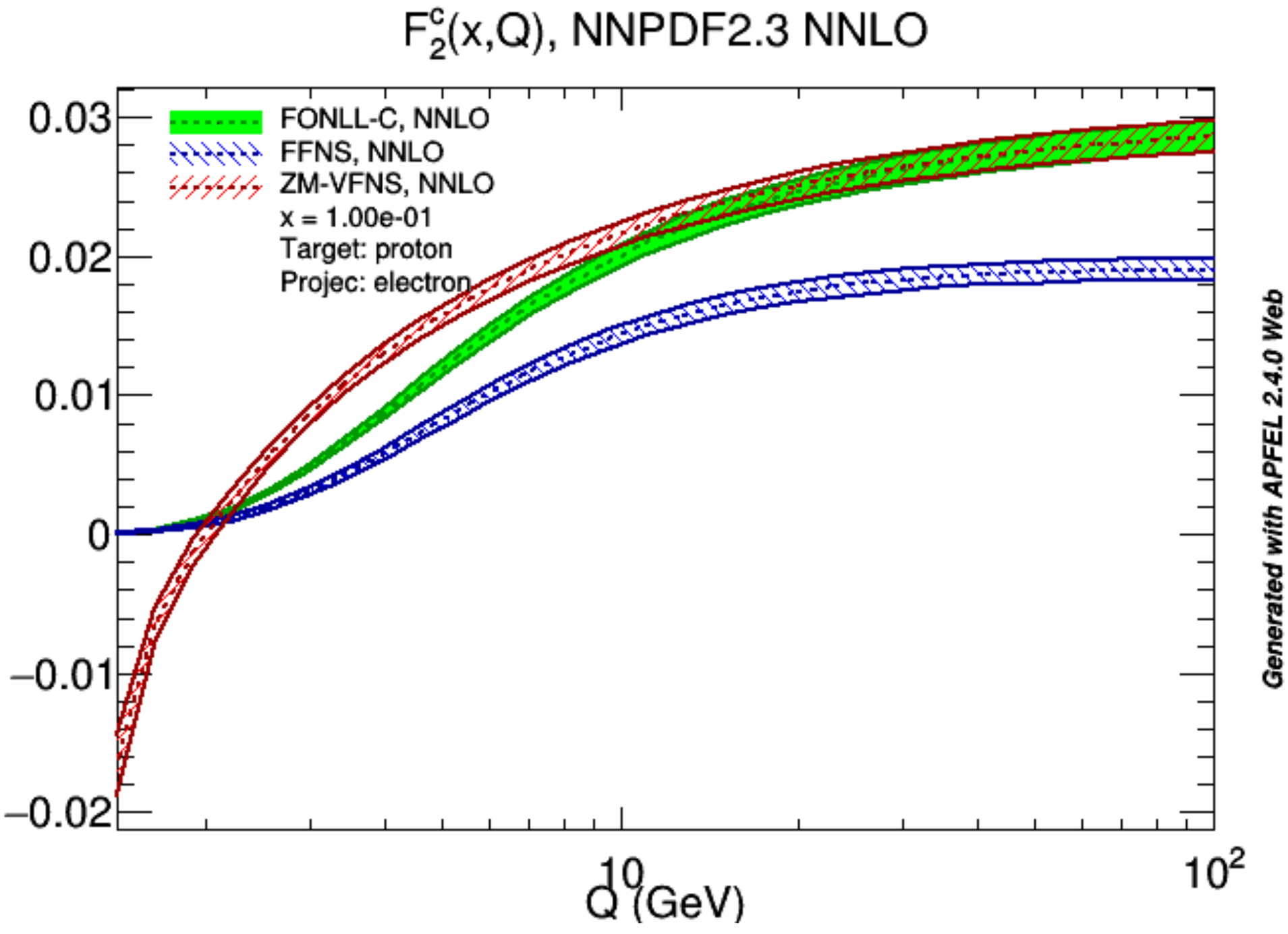
- The inclusive DIS cross section is a combination of **structure functions**.
- The inclusion of **heavy-quark** mass corrections in the structure functions needs to be reconciled with the resummation of **collinear logarithms**.
- This is achieved by **matching** different schemes in a General-Mass scheme:
  - the fixed-flavour (FF) scheme that includes mass power corrections m_h/Q, valid for $m_h \sim Q$,
  - the zero-mass (ZM) scheme that resums $\ln(m_h/Q)$, valid for $m_h \ll Q$,
  - subtraction of the double counting (DC) needed.

$$F^{\mathrm{GM}}(x,Q) = F^{\mathrm{FF}}(x,Q) + F^{\mathrm{ZM}}(x,Q) - F^{\mathrm{DC}}(x,Q)$$

- Several implementations exist, *e.g.*:
  - ACOT [Phys.Rev., vol. D50, pp. 3102–3118, 1994],
  - FONLL [JHEP, vol. 9805, p. 007, 1998],
  - RT [Phys.Rev., vol. D57, pp. 6871–6898, 1998],
  - BMSN [Eur.Phys.J., vol. C1, pp. 301–320, 1998].
- They all differ by subleading power corrections in the intermediate region.

Perturbative content

Hard cross sections: inclusive DIS



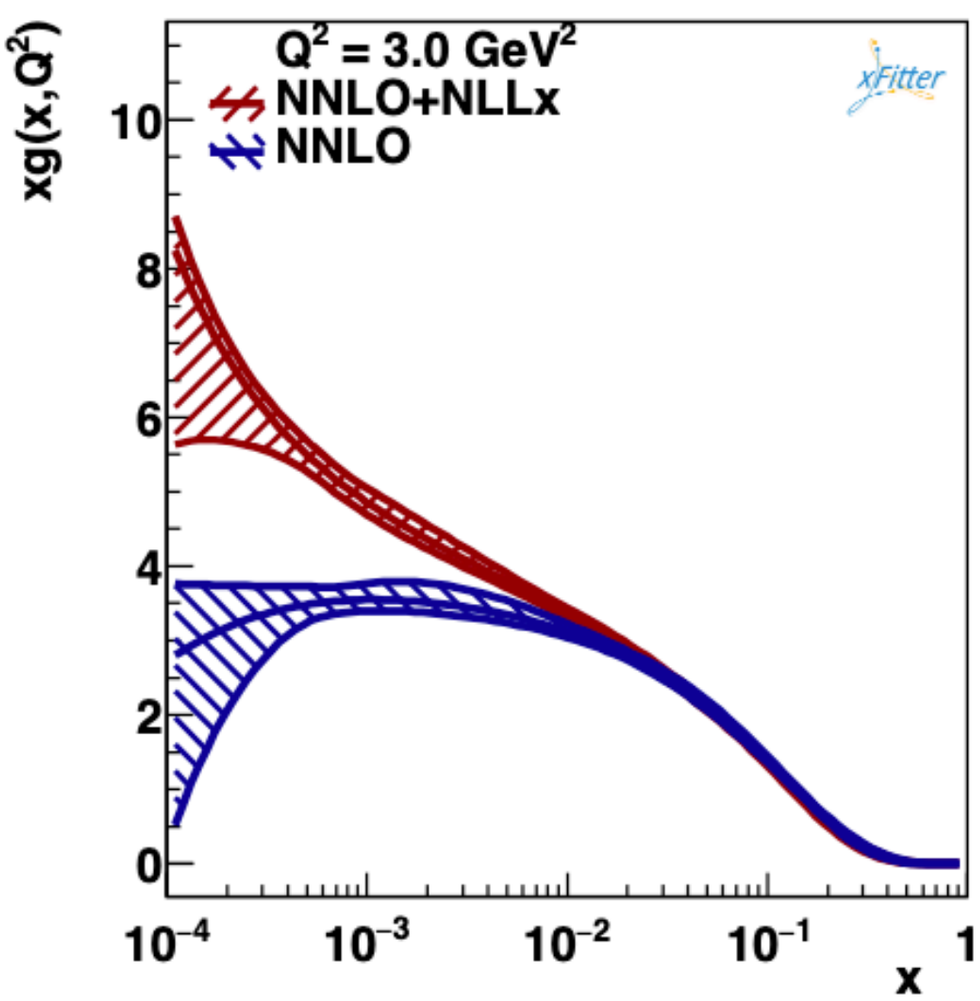
- This example makes evident how a GM scheme interpolates between the FF scheme at $m_h \sim Q$ and the ZM scheme for $m_h \ll Q$.

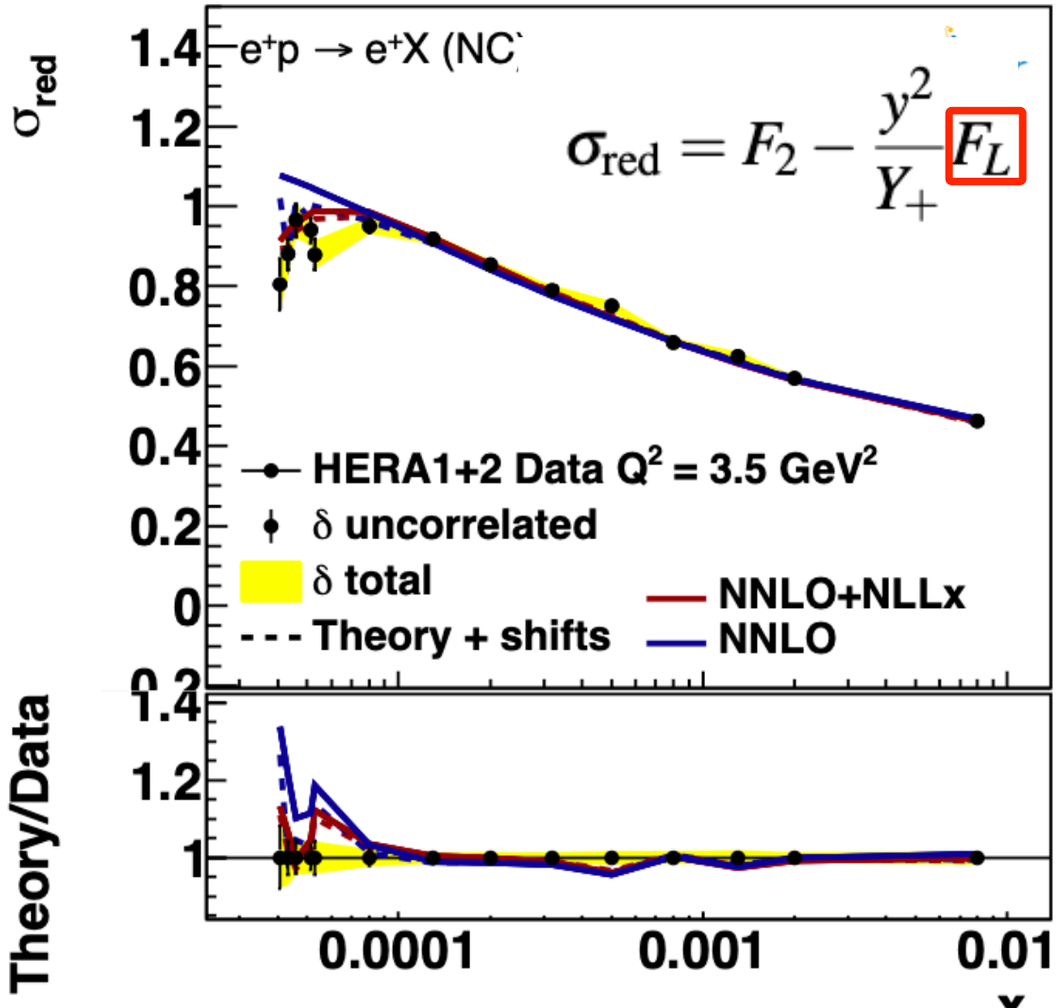
Perturbative content

Hard cross sections: inclusive DIS

[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 **less steep** and thus allows for a larger small-x gluon PDF that behaves as a sea-like distribution.



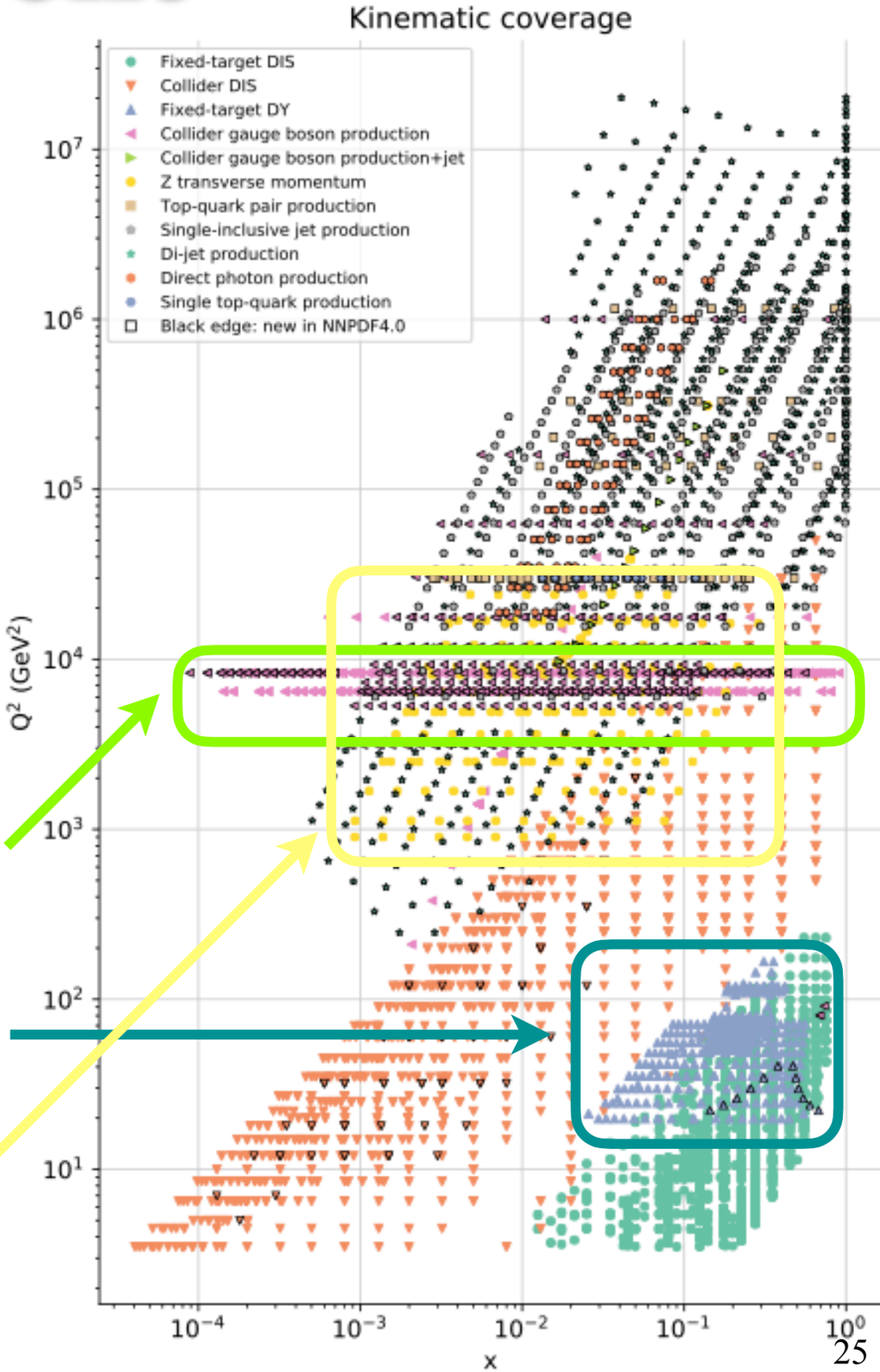


Perturbative content

Hard cross sections: Drell-Yan

Drell-Yan (both Z and W) is a very important process in **PDF determinations**:

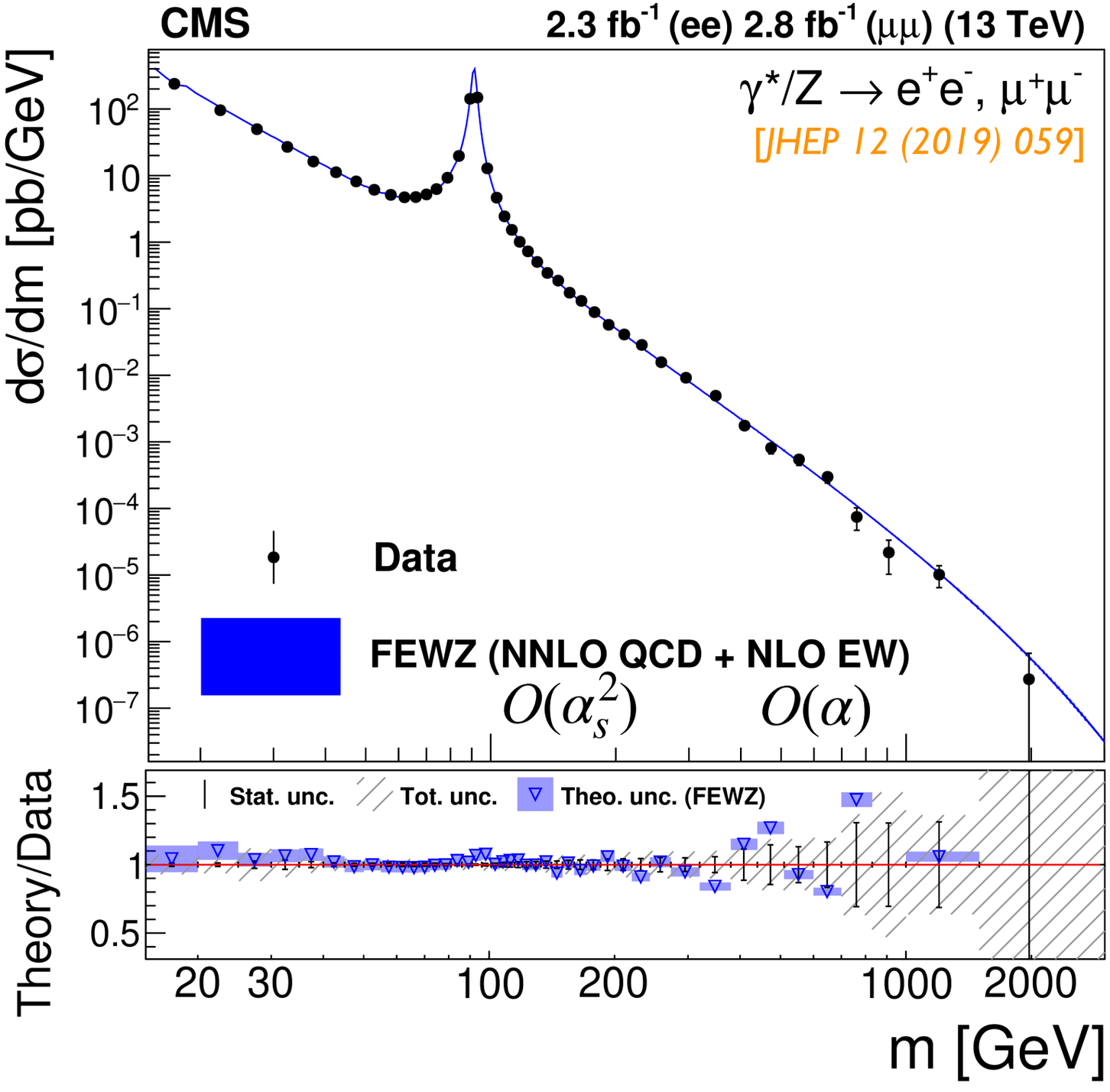
- as opposite to DIS, Drell-Yan gives access to a larger variety of quark PDF combinations:
  - this enables **flavour/anti-flavour separation**.
- very wide **kinematic coverage**:
  - collider data, placed at higher energies can reach values of x as low as 10^{-4},
  - fixed-target data is placed at lower scales and probes quark PDFs at higher values of x.
- q_T distribution gives access to the **gluon PDF**.



Perturbative content

Hard cross sections: Drell-Yan

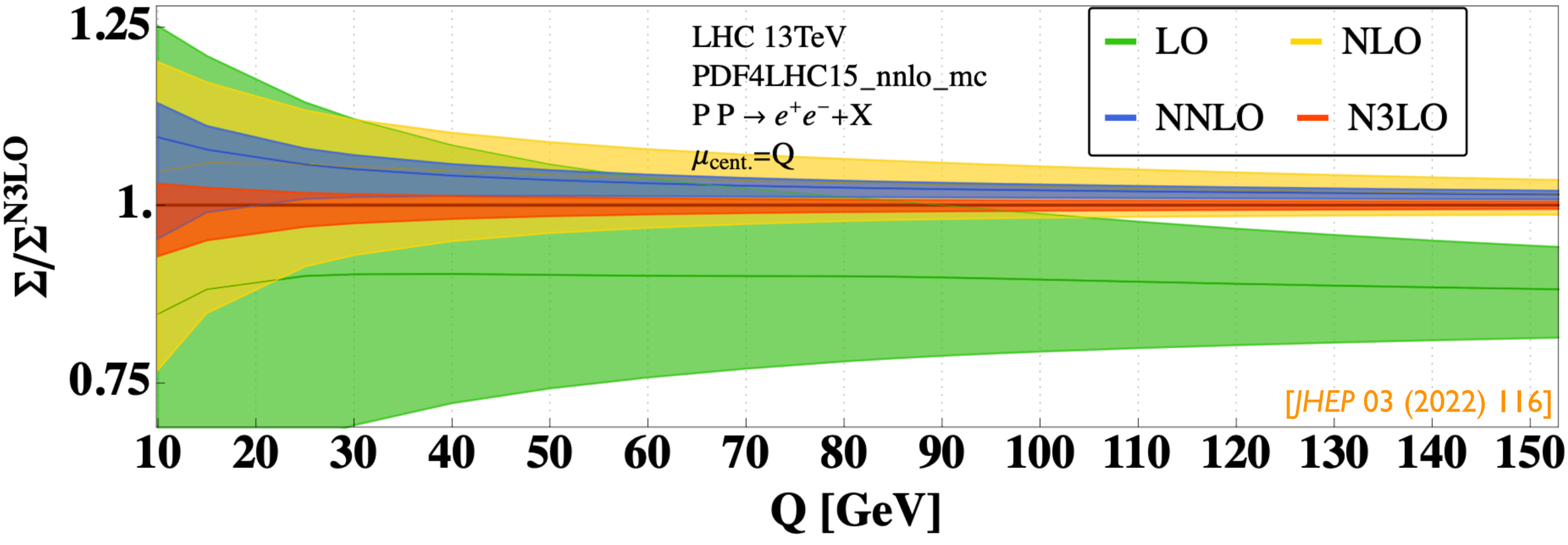
Our current understanding of the **invariant mass spectrum** in Drell-Yan is very good.



Perturbative content

Hard cross sections: Drell-Yan

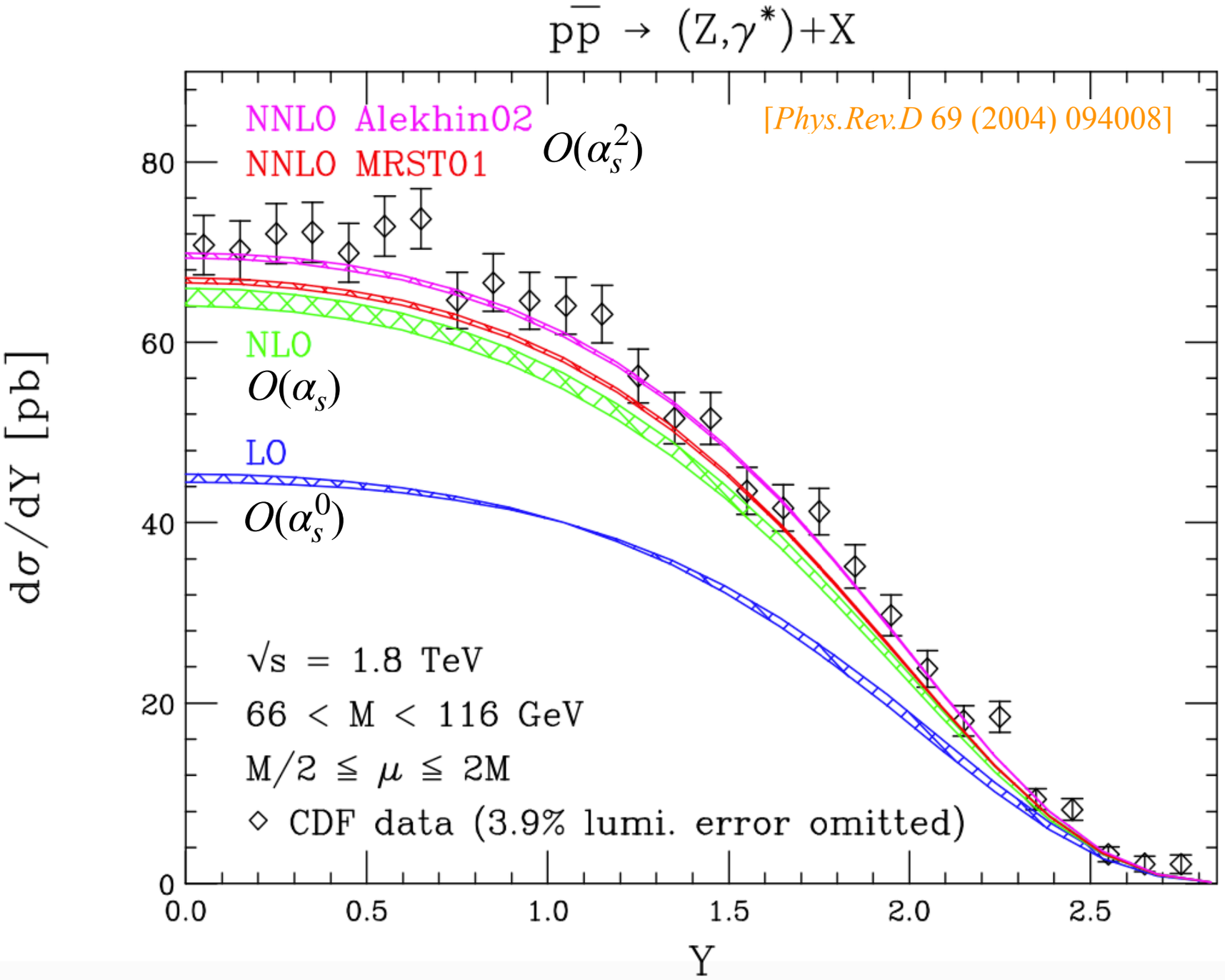
Our current understanding of the **invariant mass spectrum** in Drell-Yan is very good.



Perturbative content

Hard cross sections: Drell-Yan

Also the rapidity spectrum is particularly well know.

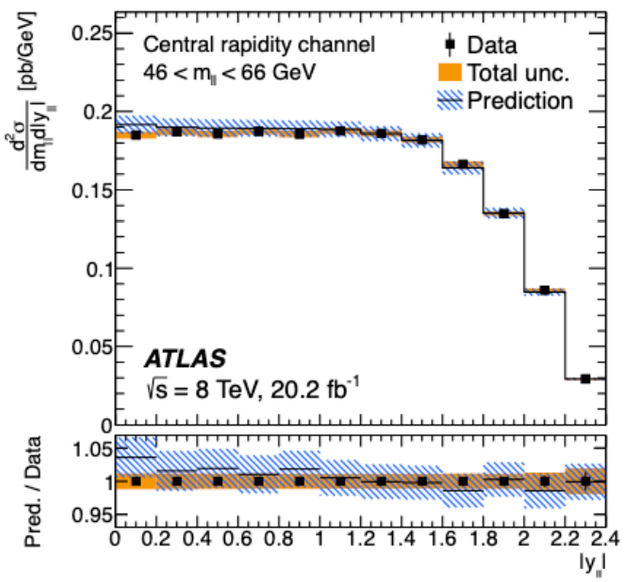


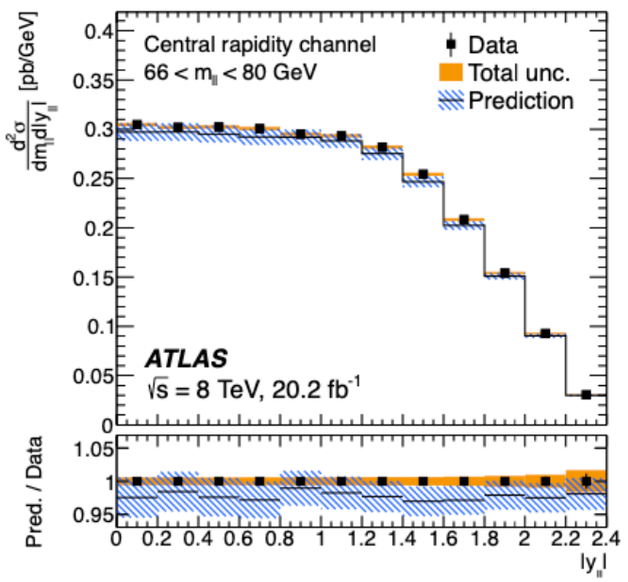
NNLO corrections significantly improve the agreement with data.

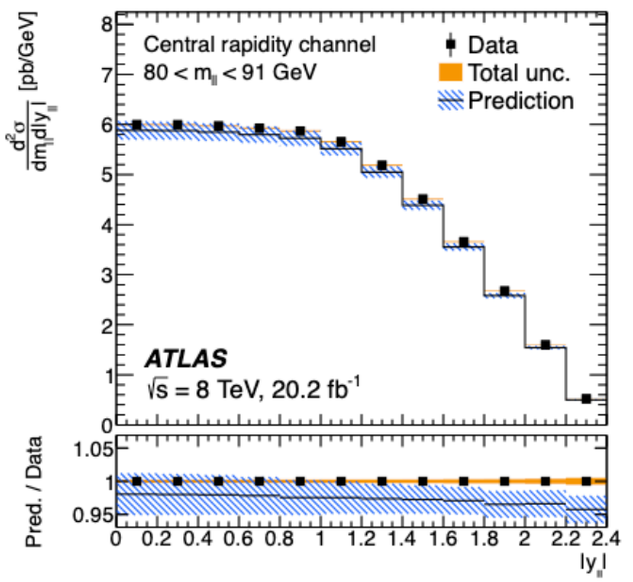
Perturbative content

Hard cross sections: Drell-Yan

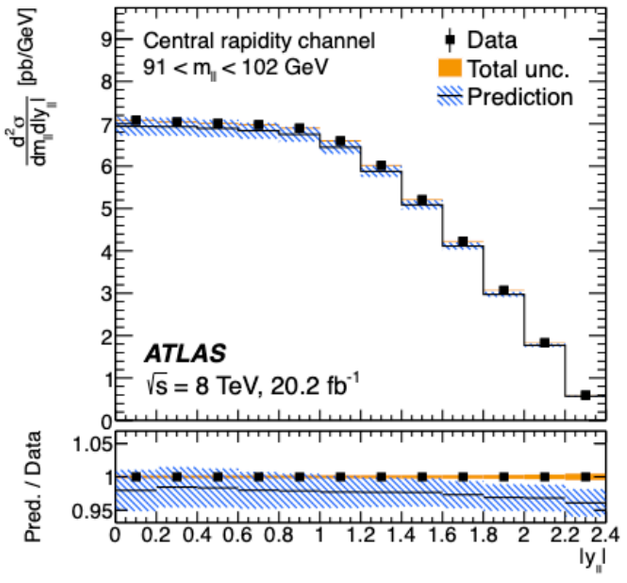
Also the rapidity spectrum is particularly well know.

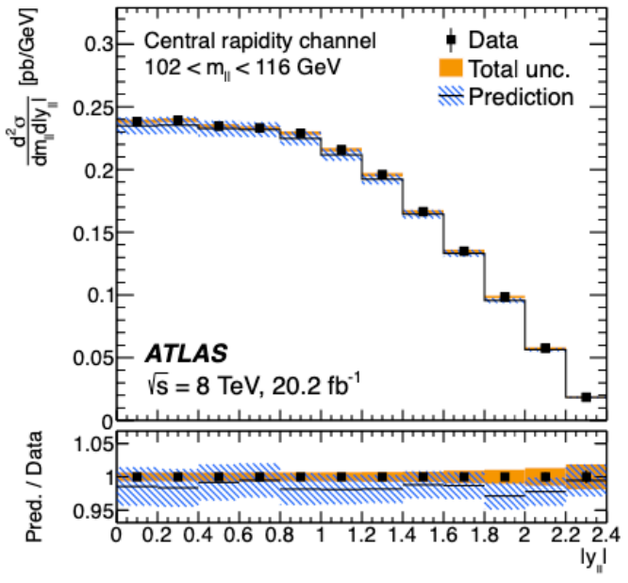


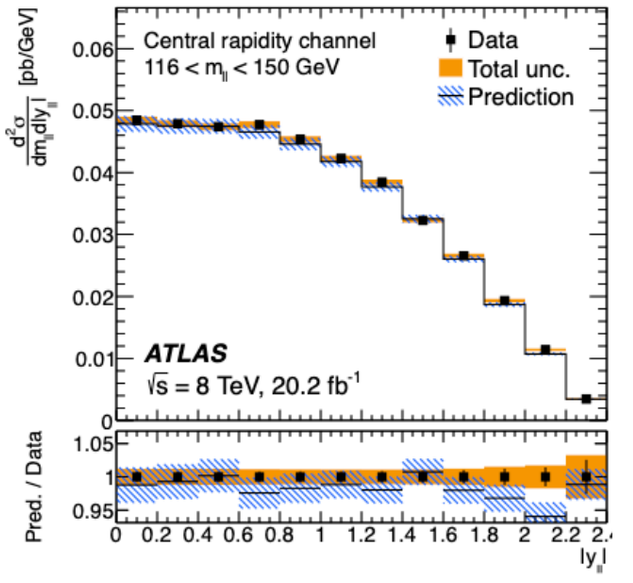


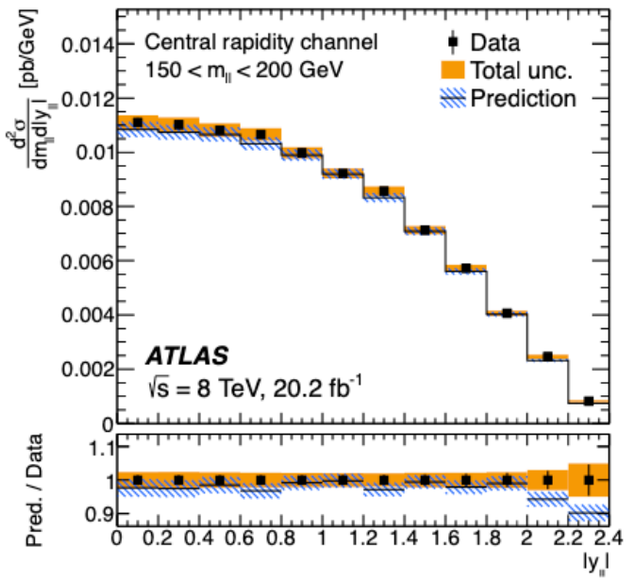


[JHEP 12 (2017) 059]







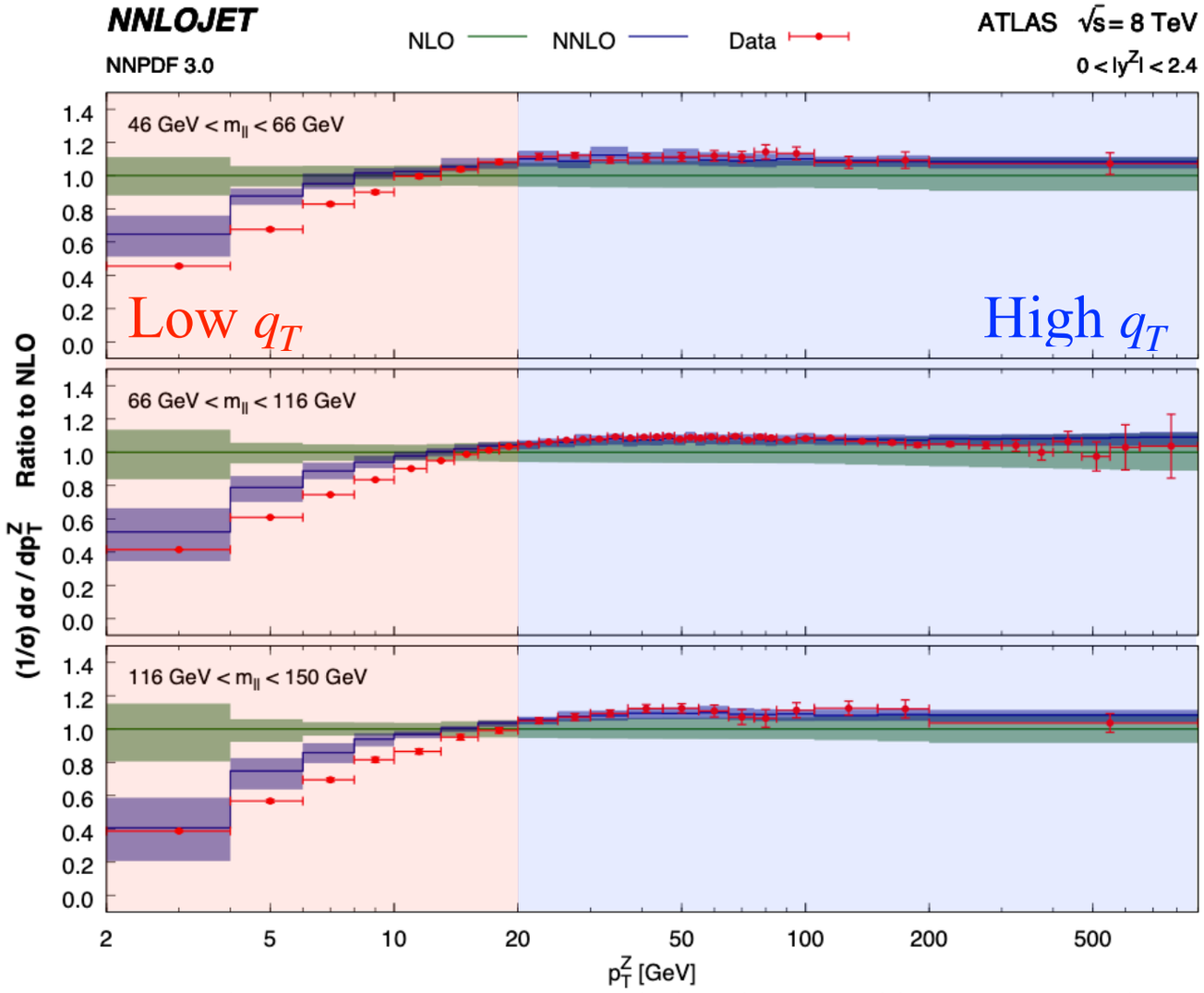


- NNLO + NLO EW predictions.
- Very good data/theory agreement also in lower and higher invariant mass bins.

Perturbative content

Hard cross sections: Drell-Yan

The fully differential NNLO (*i.e.* $O(\alpha_s^3)$) corrections to the cross section for $pp \to Z$+jet was presented in [*Phys. Rev. Lett.* 117 (2016) 2, 022001] and [*Phys.Rev.Lett.* 116 (2016) 15, 152001]. These calculations allow us to compute the q_T spectrum of the Z to NNLO accuracy.



NNLO corrections improve the agreement with data all across the board:

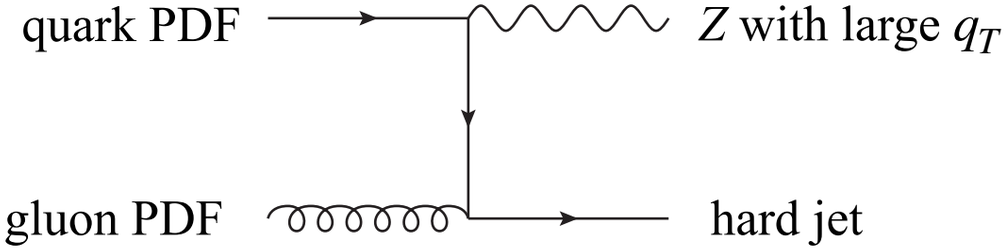
- for $q_T \sim Q$ the agreement with data is now excellent,
- for $q_T \ll Q$, NNLO partly captures the double-log behaviour and provides qualitative improvements in the description of the shape of the data: **resummation still needed**.

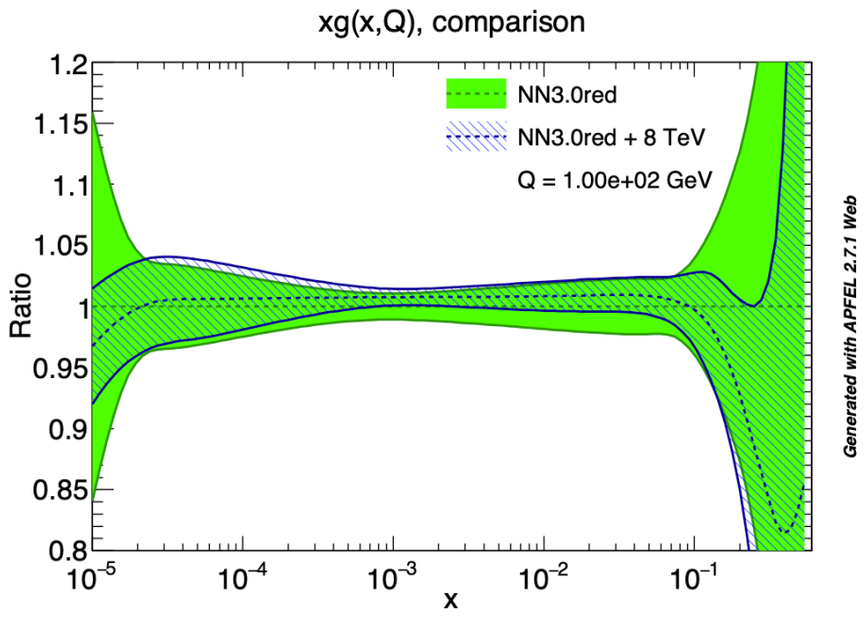
Perturbative content

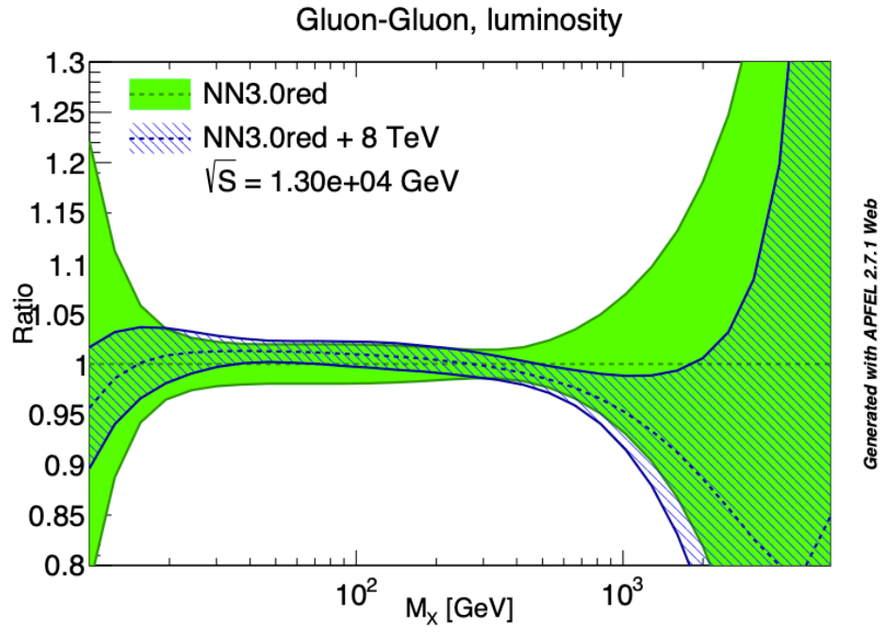
Hard cross sections: Drell-Yan

The q_T of the Z boson allows us to constraint the collinear gluon PDF:

- collinear factorisation is reliable for $q_T \simeq Q$.
- In order for the Z to have a large q_T, it needs an object to recoil against. This is typically a jet. As a consequence, the relevant process is $pp \to Z + j + X$.
- One of the leading-order partonic cross sections contributing to this process is:







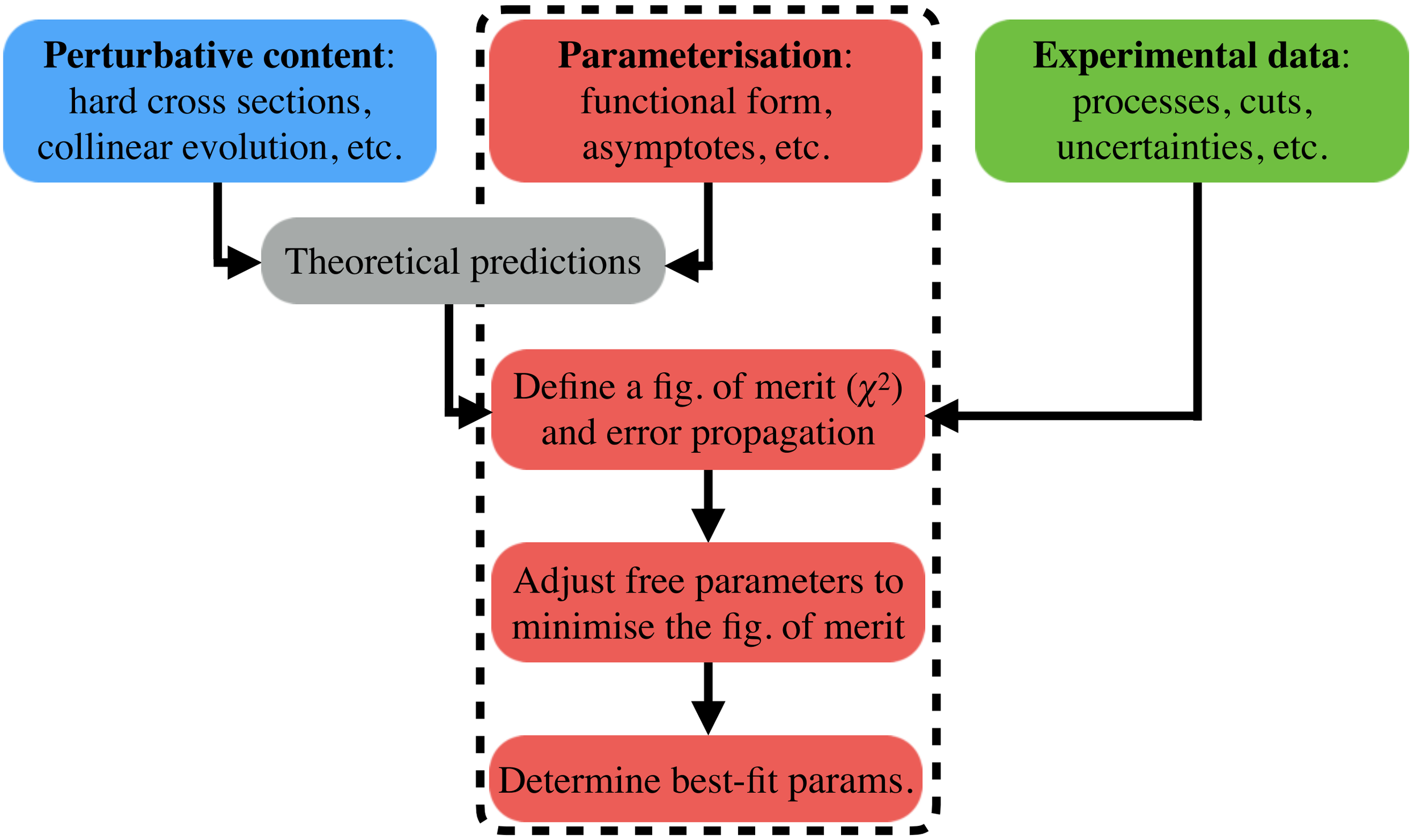
The impact on the gluon PDF is **significant**.

Perturbative content

Hard cross sections: fast interfaces

- Very often, a direct "on-line" computation of physical observables is not a viable option when fitting PDFs:
  - computing a single distribution may take days and this needs to be iterated a large number (thousands) of times during a fit.
- Fast interfaces have been devised to overcome this problem.
- As of today, different interfaces exist:
  - APPLgrid Carli *et al.*,[*Eur.Phys.J.C* 66 (2010) 503-524],
  - FastNLO FastNLO Collaboration [DIS 2012, 217-221],
  - APFELgrid Bertone, Carrazza, Hartland [Comput.Phys.Commun. 212 (2017) 205-209],
  - aMCfast Bertone, Frederix, Frixione, Rojo, Sutton [*JHEP* 08 (2014) 166],
  - PineAPPL Carrazza, Nocera, Schwan, Zaro [*JHEP* 12 (2020) 108].
- They are all based on the same idea: interpolating PDFs.

The general fit strategy



Fit methodology

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:

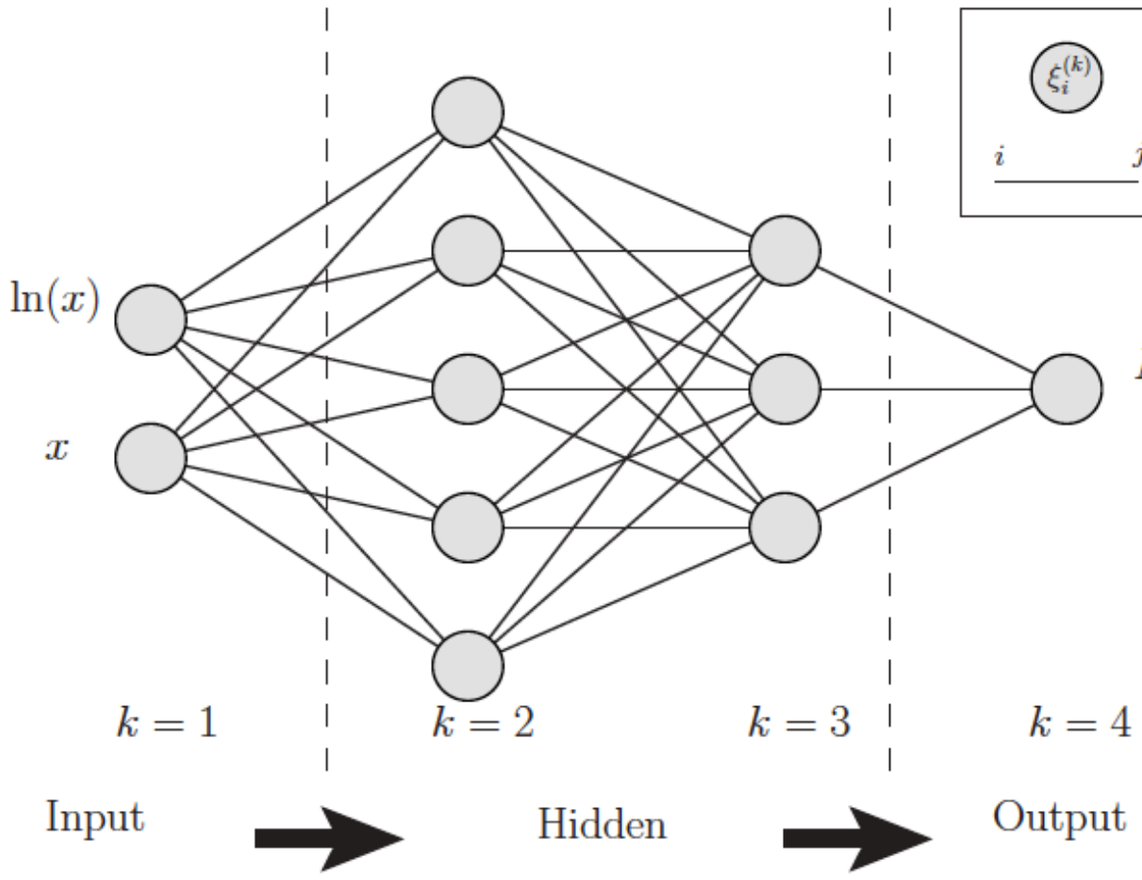
$$P_i(x) = \begin{cases} 1 \\ 1 + \gamma_i x \\ 1 + \gamma_i x + \delta_i \sqrt{x} \\ \dots \end{cases}$$

- **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: $\chi^2 = \sum_{ij} (T_i - D_i) V_{ij}^{-1} (T_j - D_j)$

- 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:

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

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

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

Fit methodologies

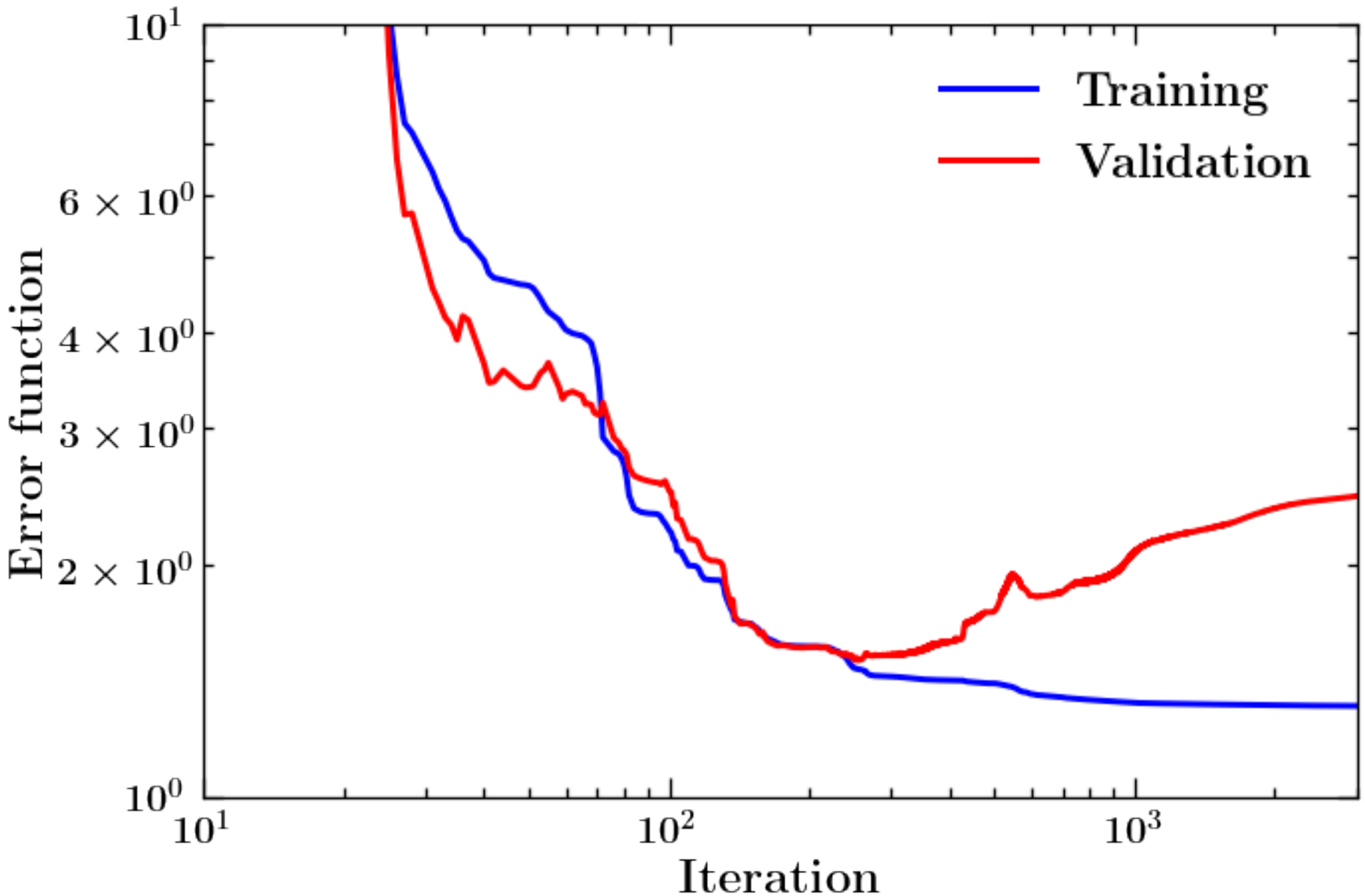
Minimisation and stopping

- 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 more 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, JAM, etc.) but they are typically less inclusive in terms of data.

Parton luminosities

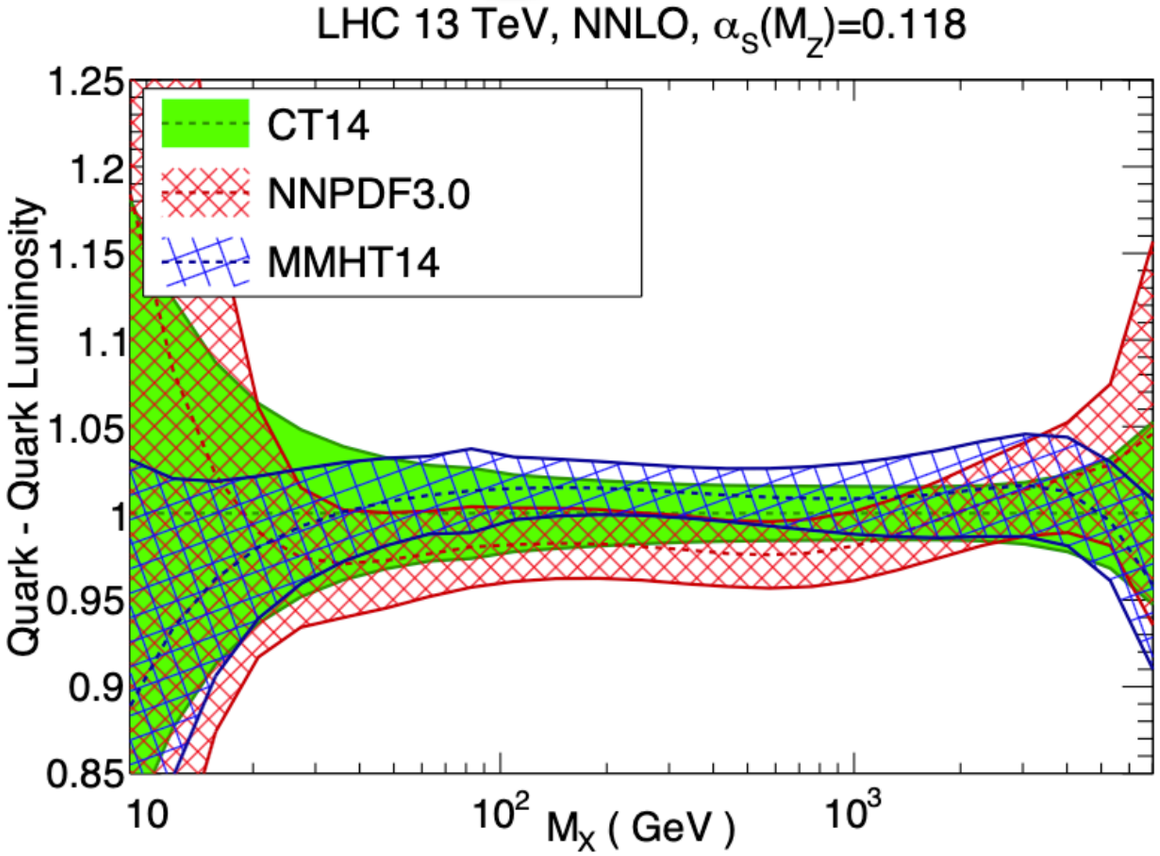
- Interesting quantities are the so-called **parton luminosities**:

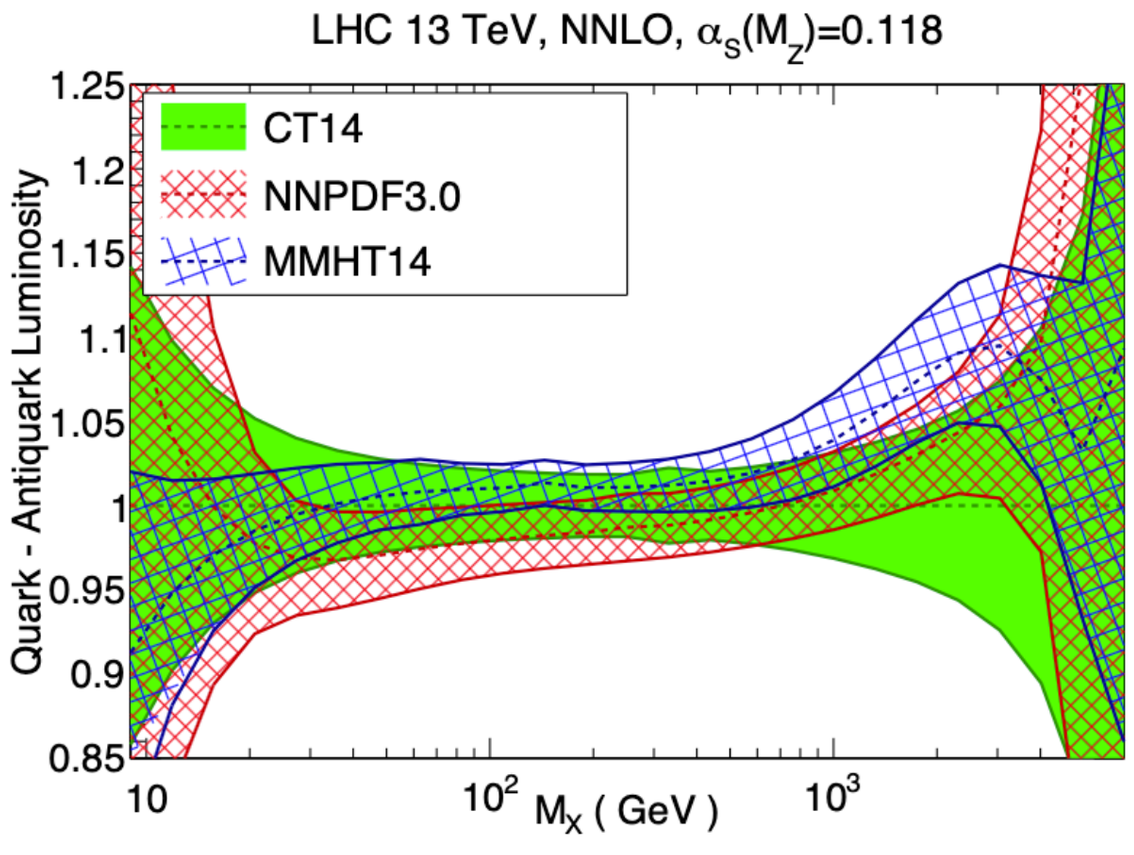
$$\mathcal{L}_{ij} = \frac{1}{s}\int_{M_X^2/s}^1 \frac{dy}{y} f_i(y, M_X) f_j\left(\frac{M_X^2}{ys}, M_X\right)$$

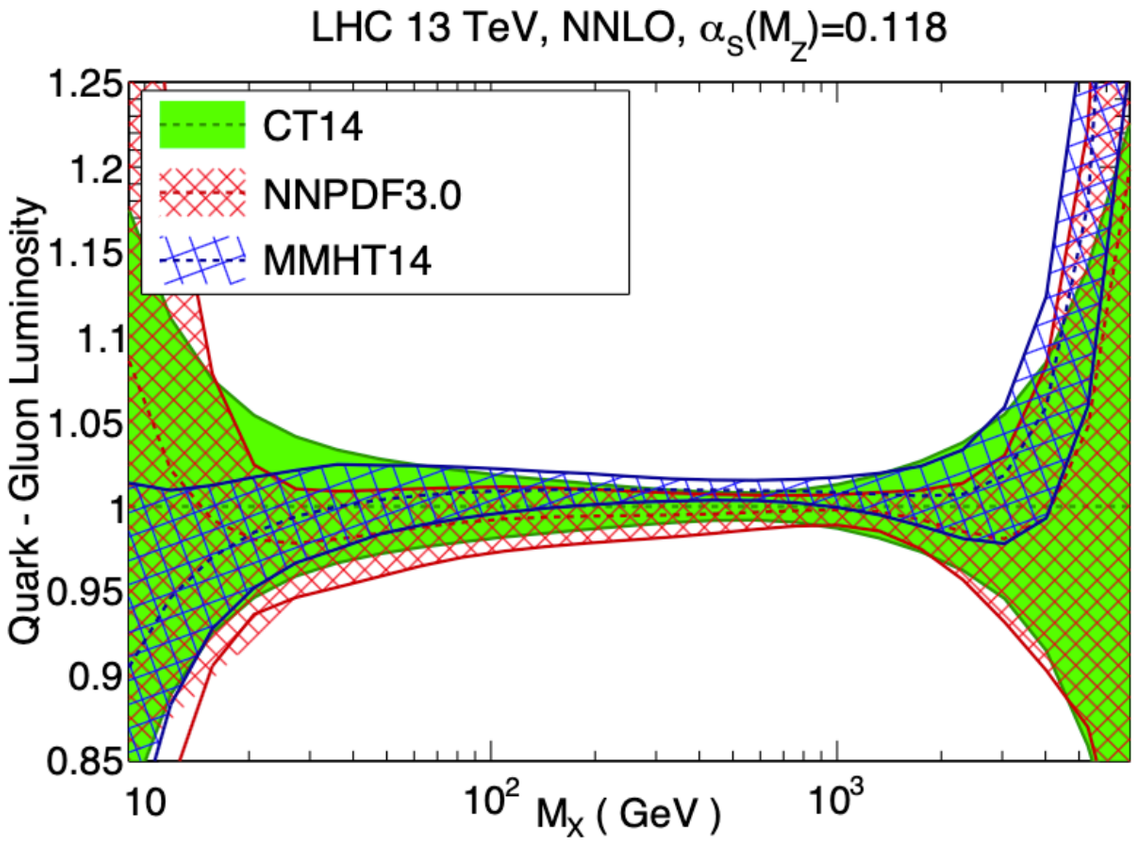
- Relevant for invariant mass distributions of the final state in pp collision processes, *e.g.*:

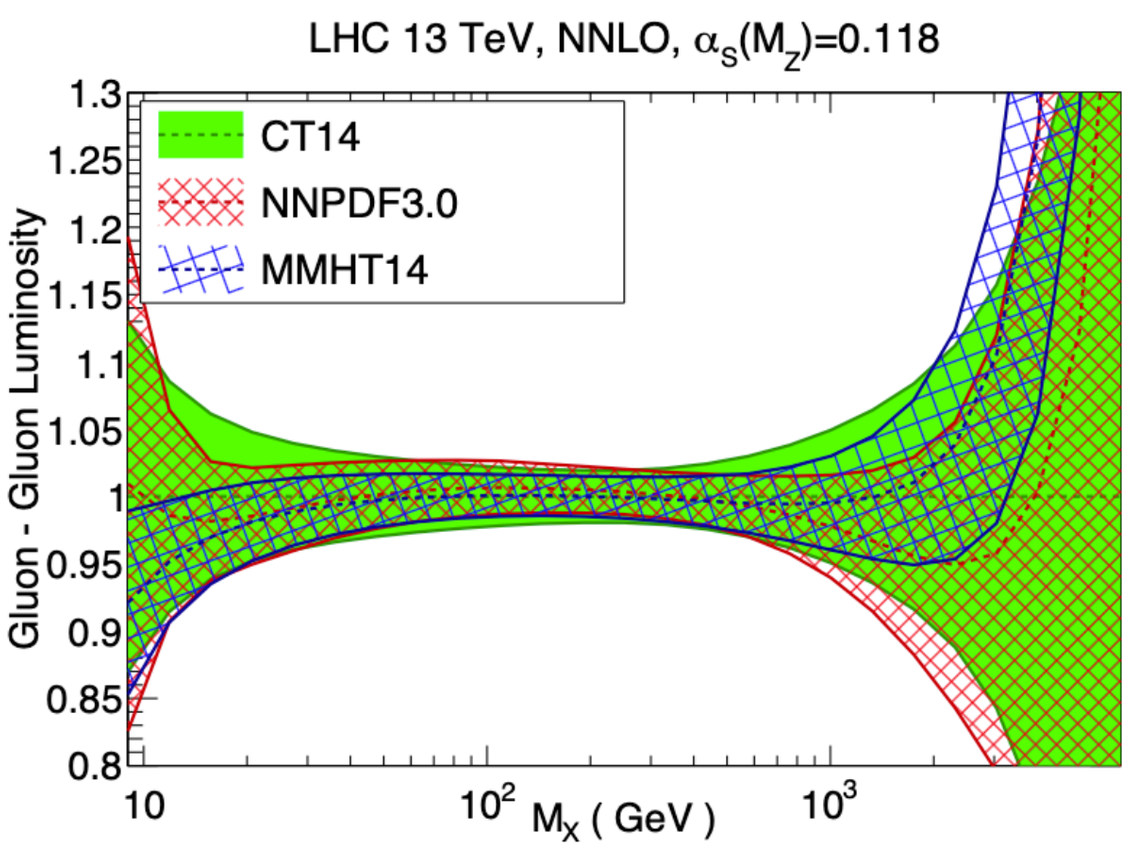
- **Drell-Yan** mostly sensitive to $\mathscr{L}_{q\bar{q}}$,
- **Higgs** and $t\bar{t}$ production in gluon fusion mostly sensitive to $\mathscr{L}_{gg}$,
- ***W* + charm** mostly sensitive to $\mathscr{L}_{sg}$,
- …

A snapshot back in 2015



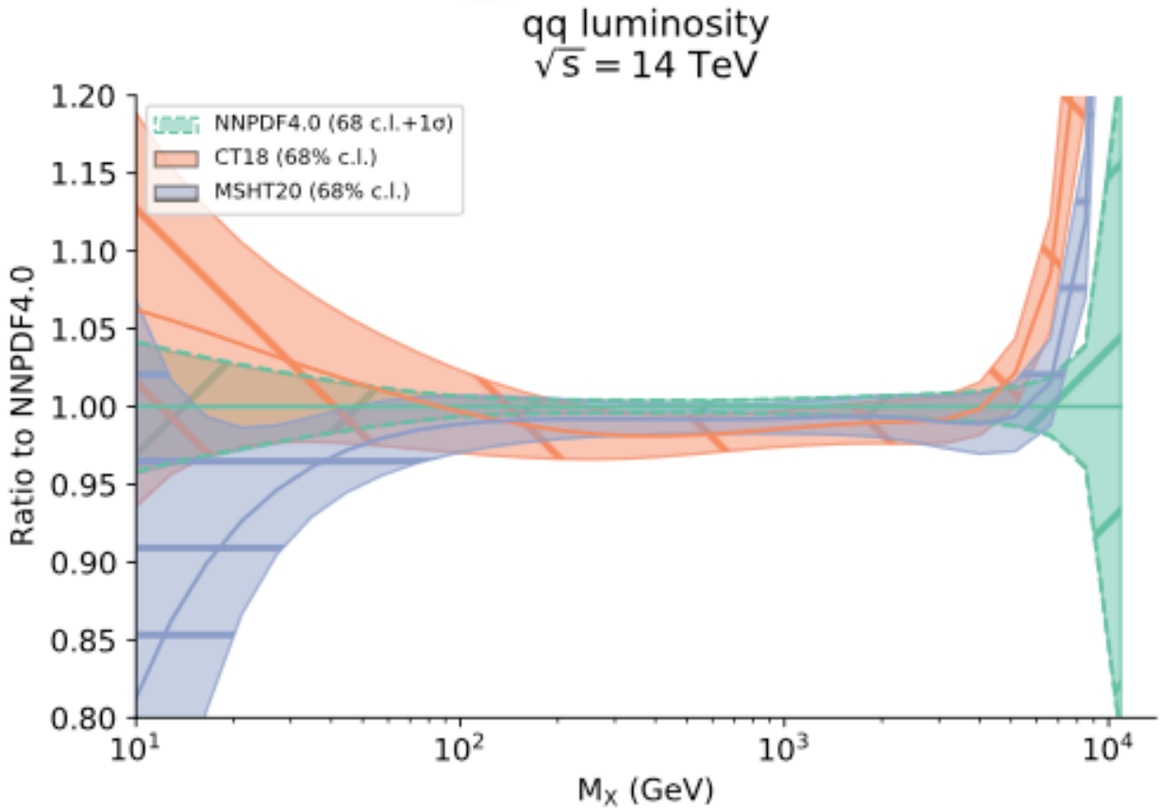


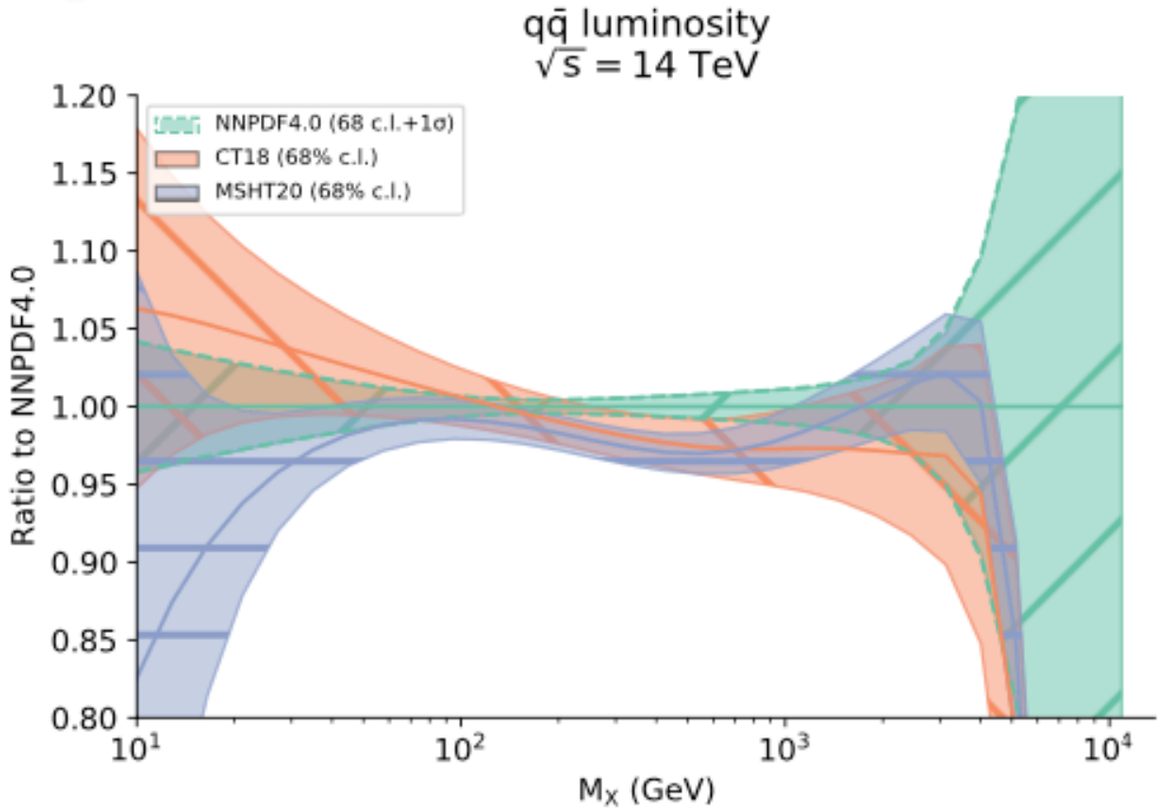


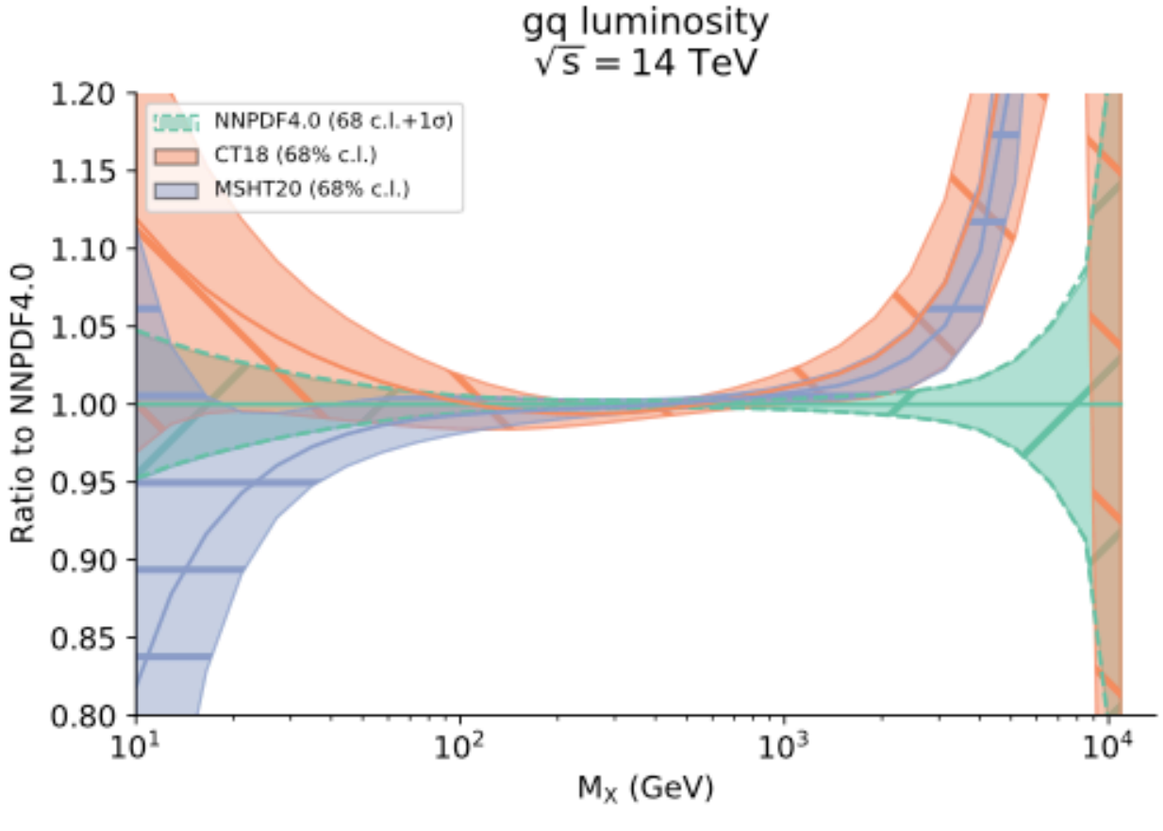


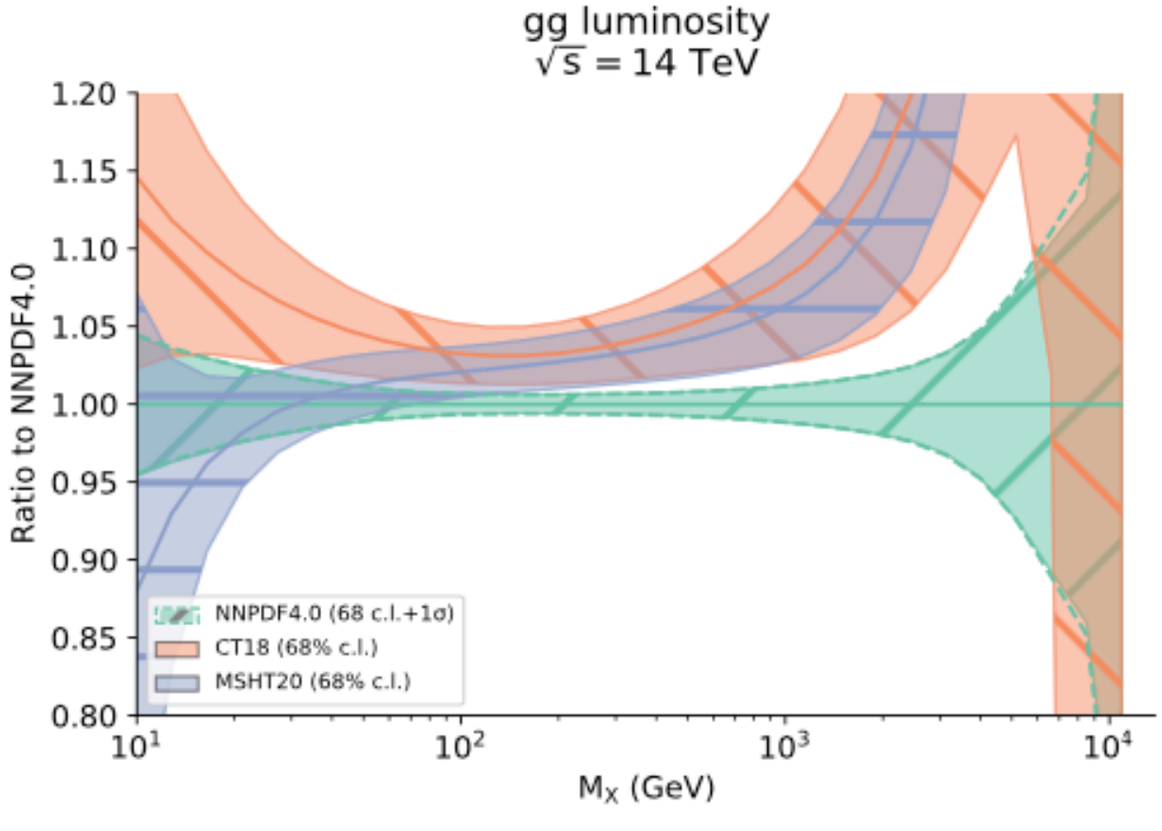
[https://arxiv.org/pdf/1510.03865.pdf]

A snapshot today



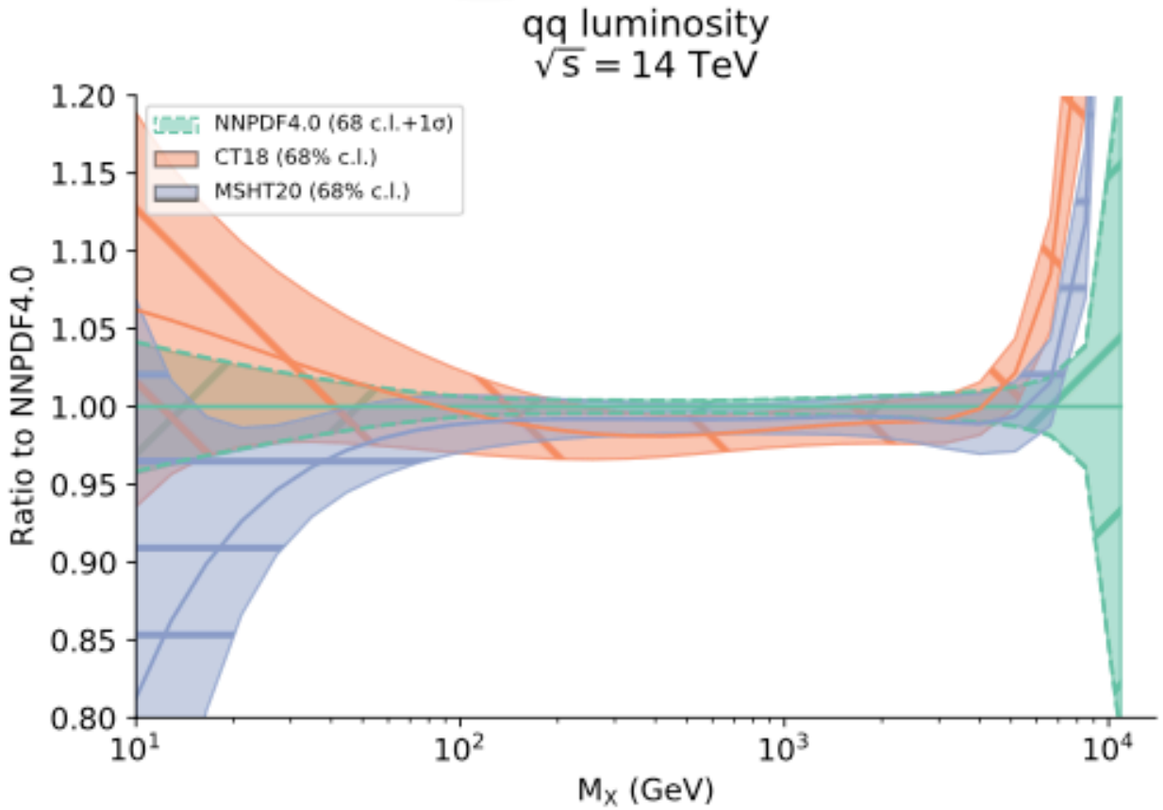


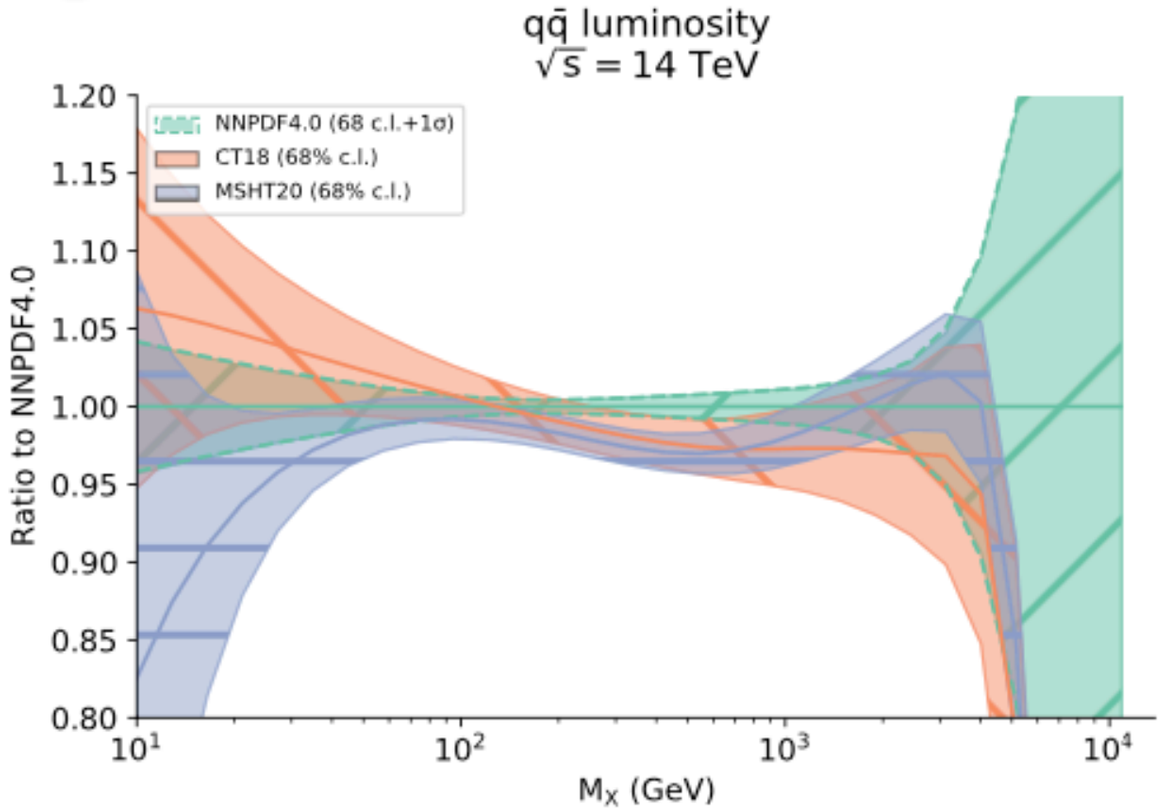


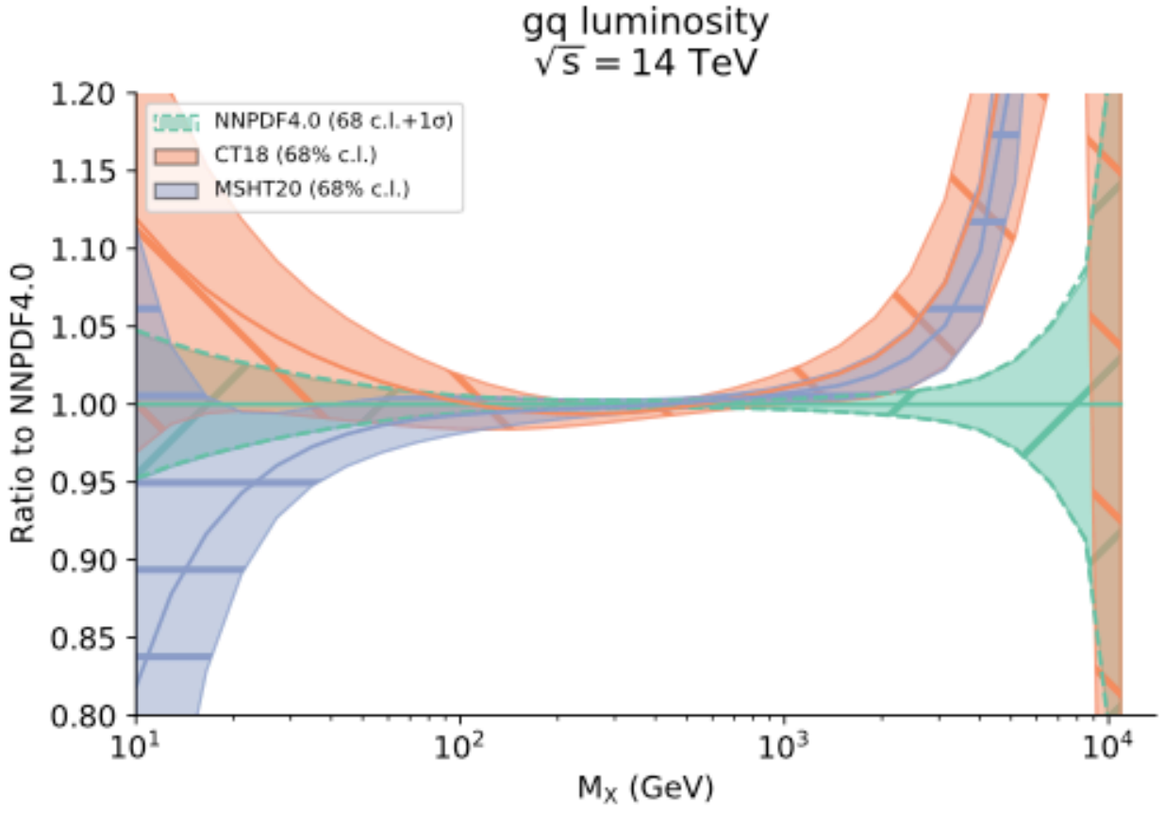


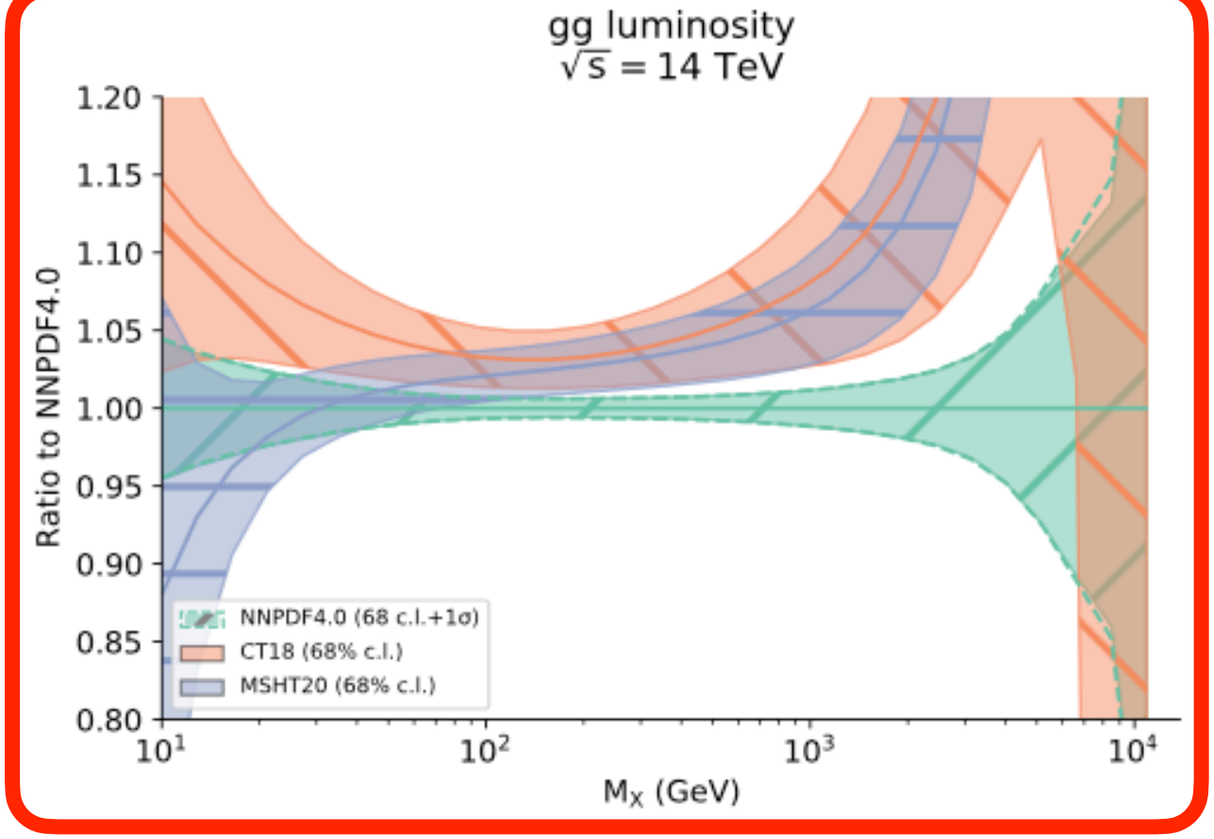
[E. Nocera's talk at DIS2021]

A snapshot today









[E. Nocera's talk at DIS2021]

Conclusions

- Fitting collinear distributions requires making many choices:
  - Data set.
  - Perturbative content:
    - DGLAP evolution,
    - hard cross sections.
  - Fitting methodology:
    - PDF parameterisation,
    - $\chi 2$ and error propagation,
    - minimisation and stopping.
- Different choices may lead to different results:
  - "Global" fitters have reached a nice degree of agreement,
    - even though most recent fits present a larger departure as compared to the past.

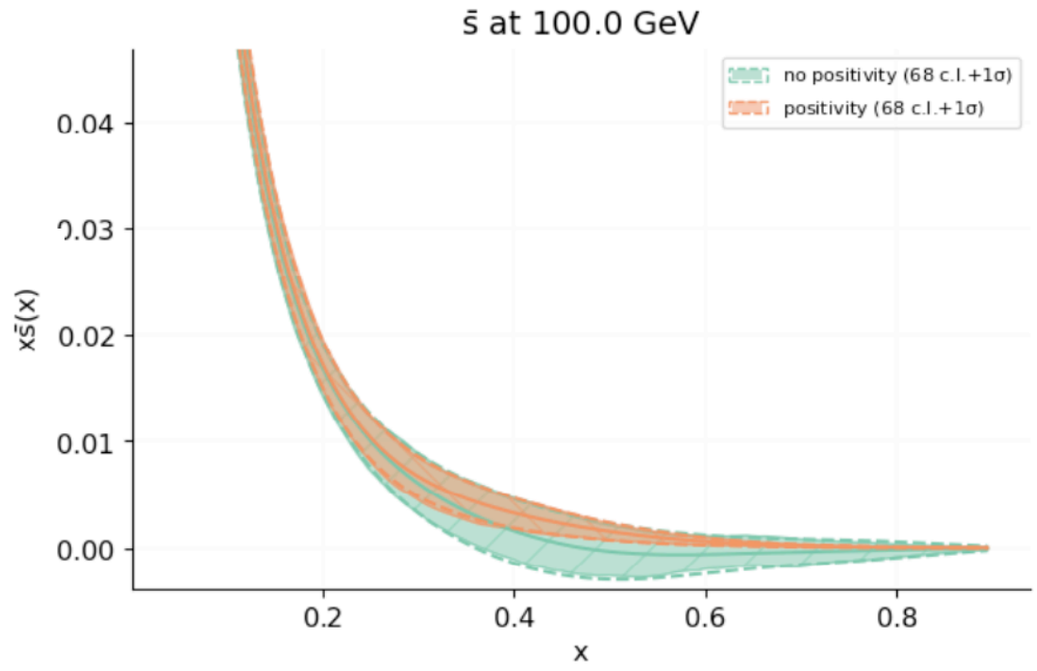
Backup

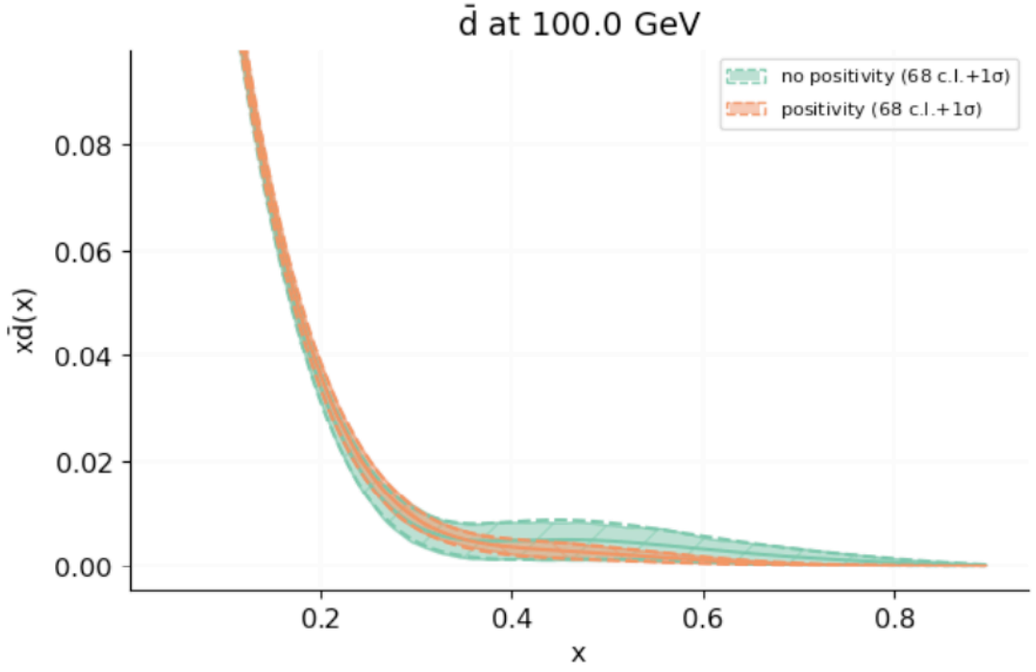
Positivity and PDFs

- PDFs have to be 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**:

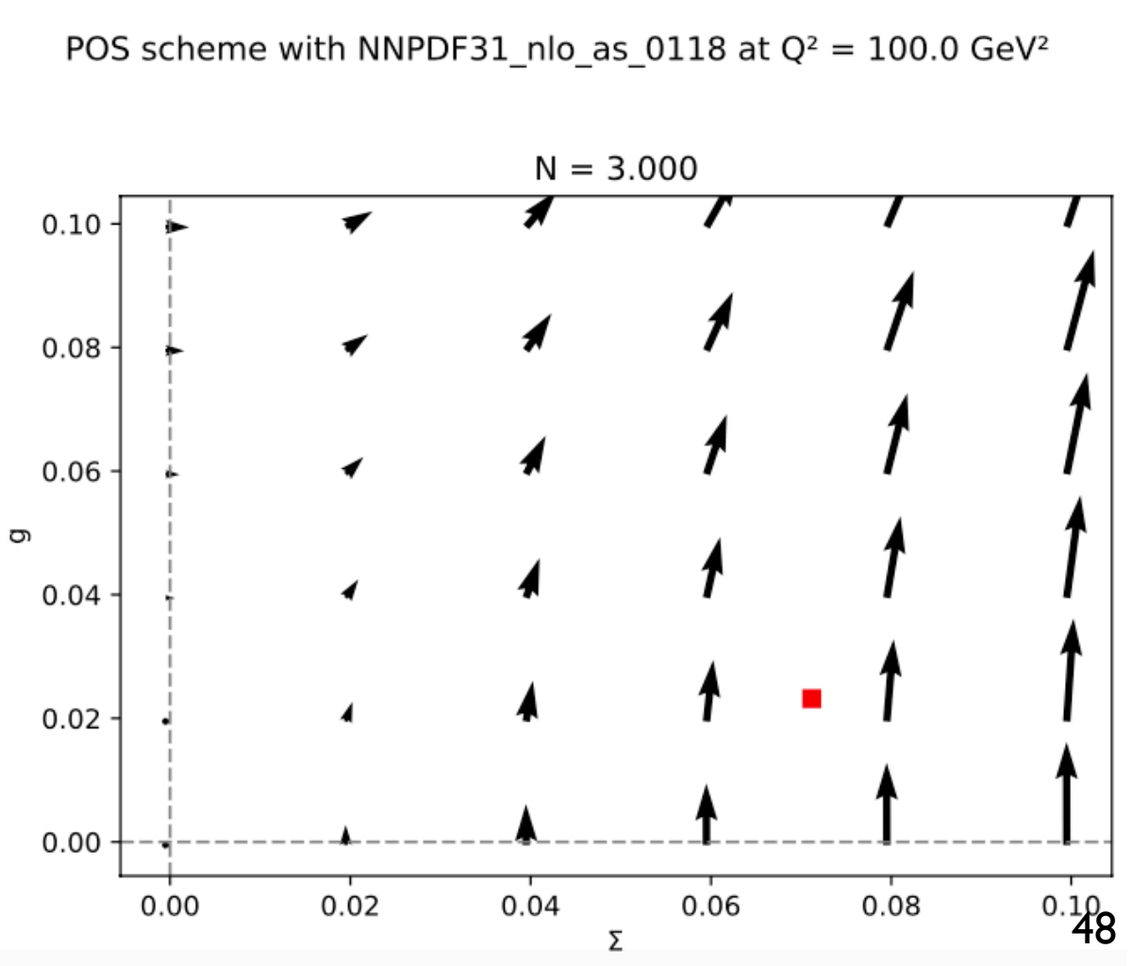
[Candido, Forte, Hekhorn, *JHEP* 11 (2020) 129]

> 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:

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

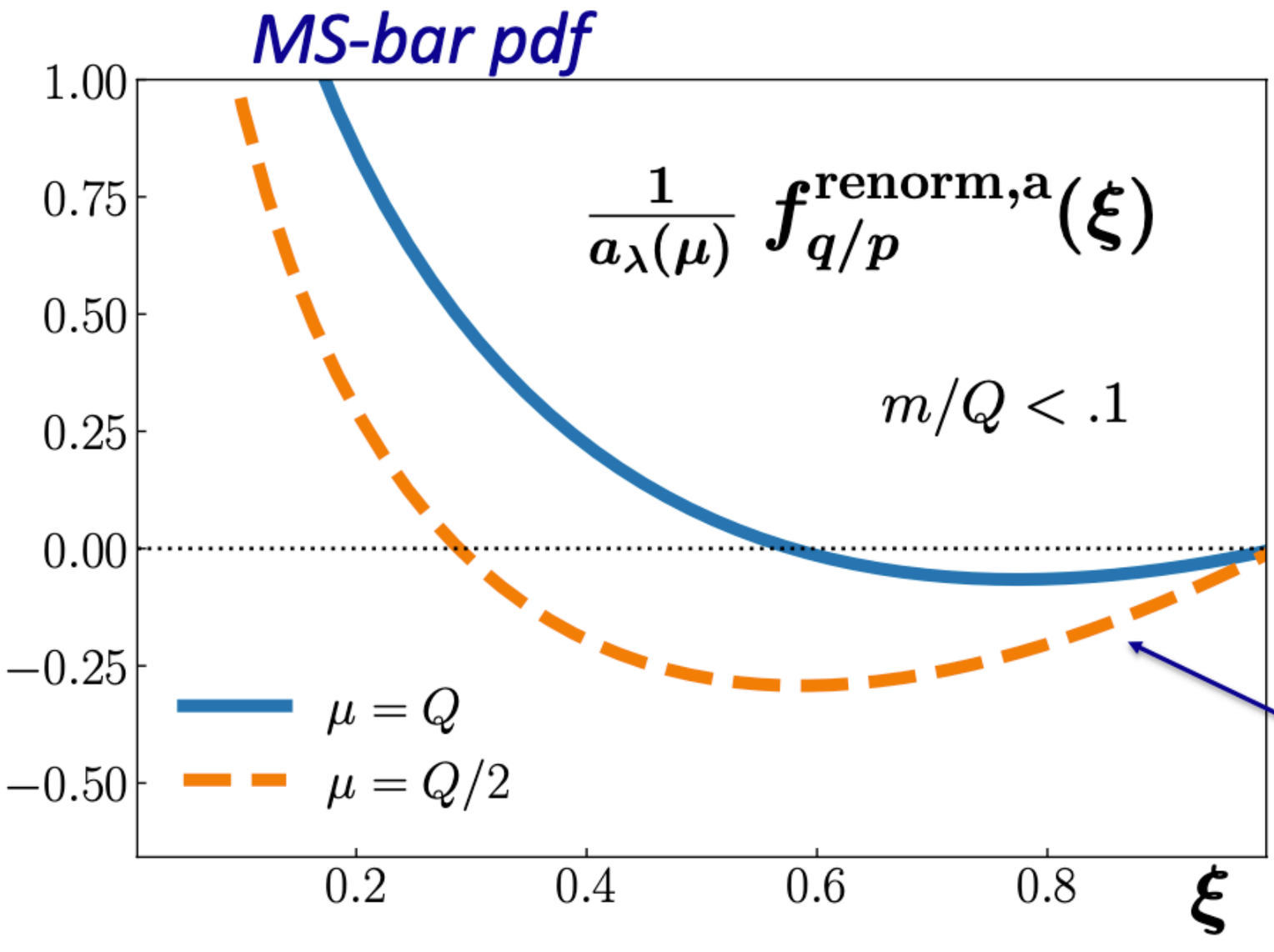
- The authors find that this transformation tends to make PDFs more positive.
- If POS PDFs are positive (by definition) $\Longrightarrow$ $\overline{\text{MS}}$ are to be even more positive.



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]



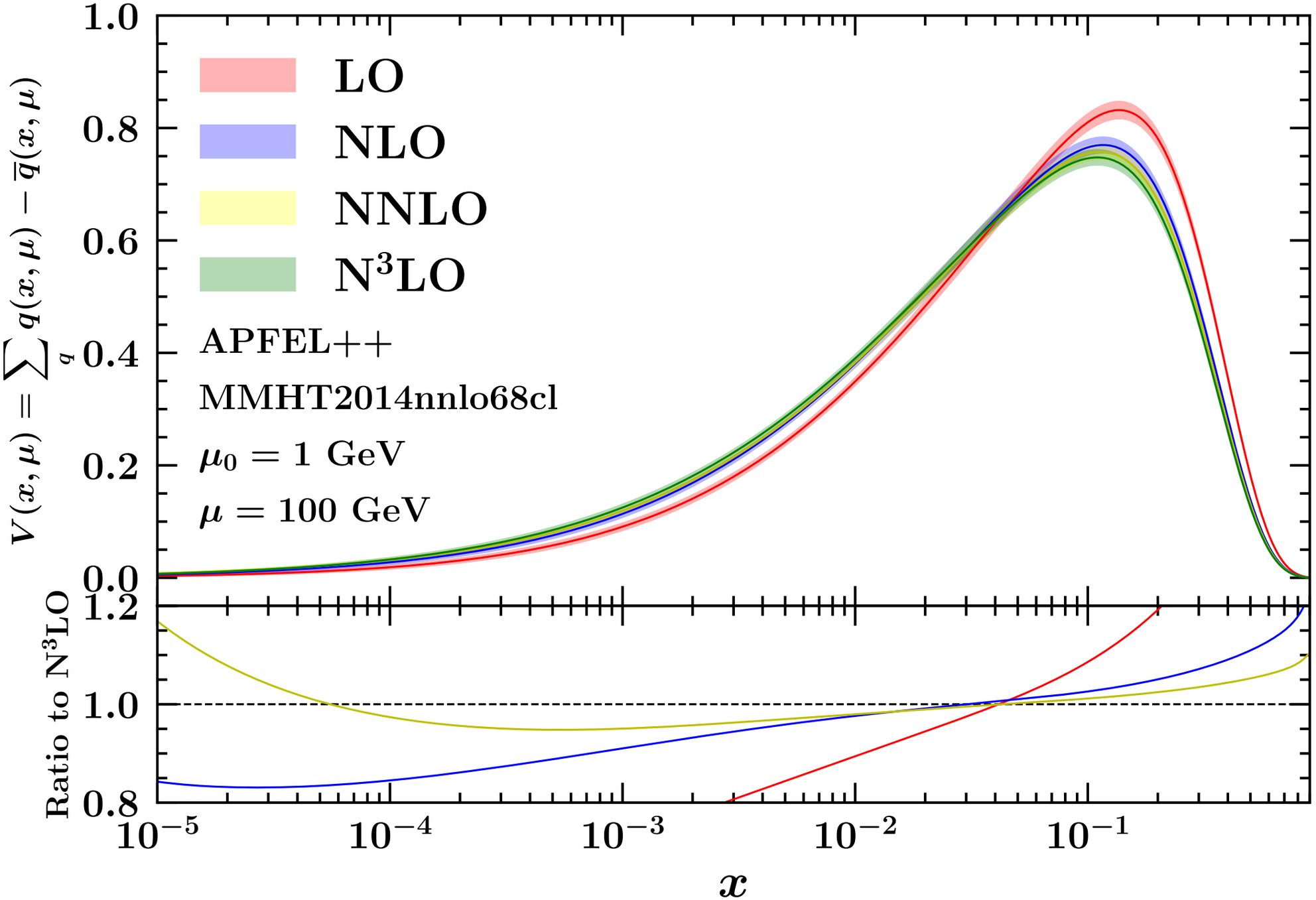
- Nothing forces the pdf to be strictly positive, even for relatively large Q.

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

Perturbative content

DGLAP evolution: x-space approach

- $O(\alpha_s^4)$ **non-singlet** splitting functions in the planar limit recently computed [Moch *et al.*, arXiv:1707.08315] enabling partial N^{3}LO evolution. Approximations available for the **Singlet**.



Perturbative content

Hard cross sections: Drell-Yan

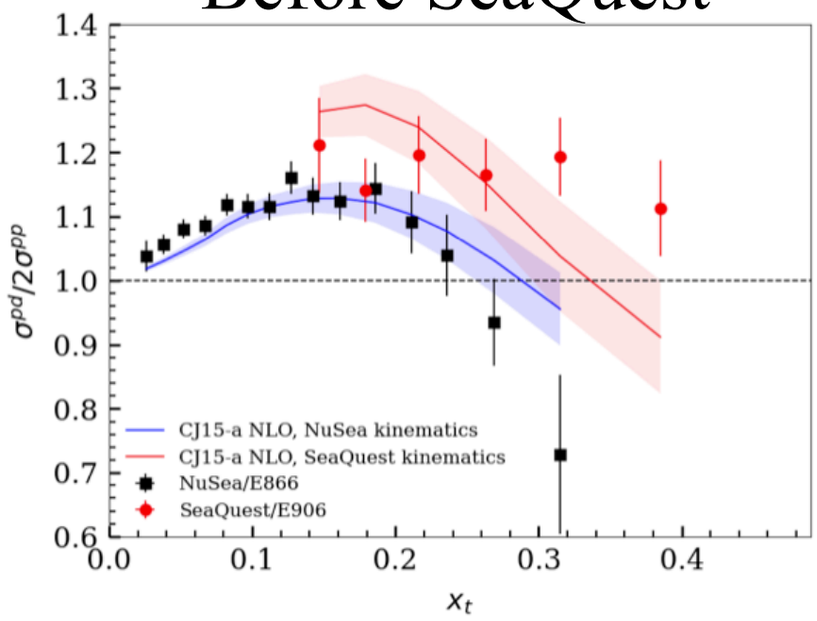
Recently the SeaQuest (E906) experiment at Fermilab has released data for the ratio of cross sections σ_{pd}/σ_{pp} [*Nature* 590 (2021) 7847, 561-565].

This ratio is **sensitive to the ratio of sea quark PDFs**:

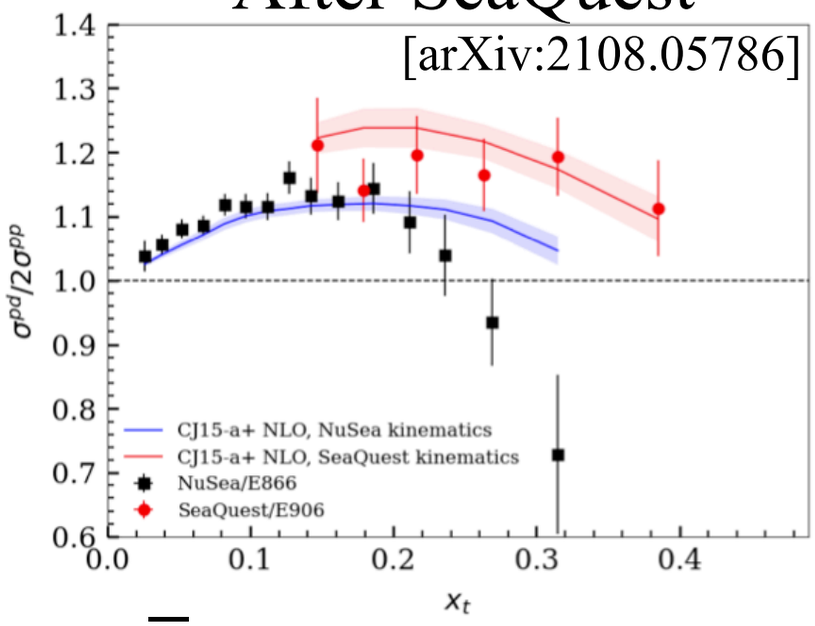
$$\frac{\sigma_{pd}}{\sigma_{pp}} \simeq 1 + \frac{\bar{d}(x)}{\bar{u}(x)}$$

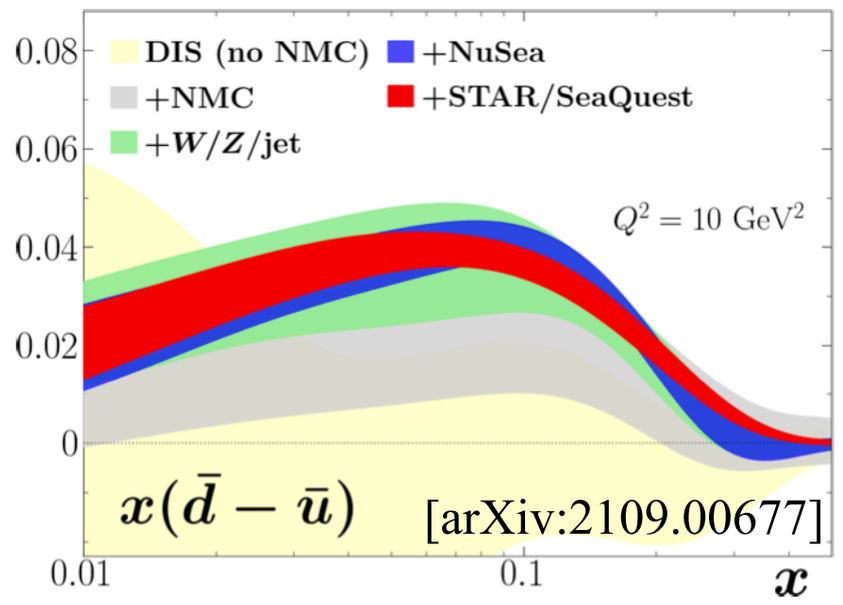
Being a fixed-target experiment, **large values of *x*** are probed giving us access to the sea quark PDFs in a region that is presently **poorly known**.

Before SeaQuest



After SeaQuest

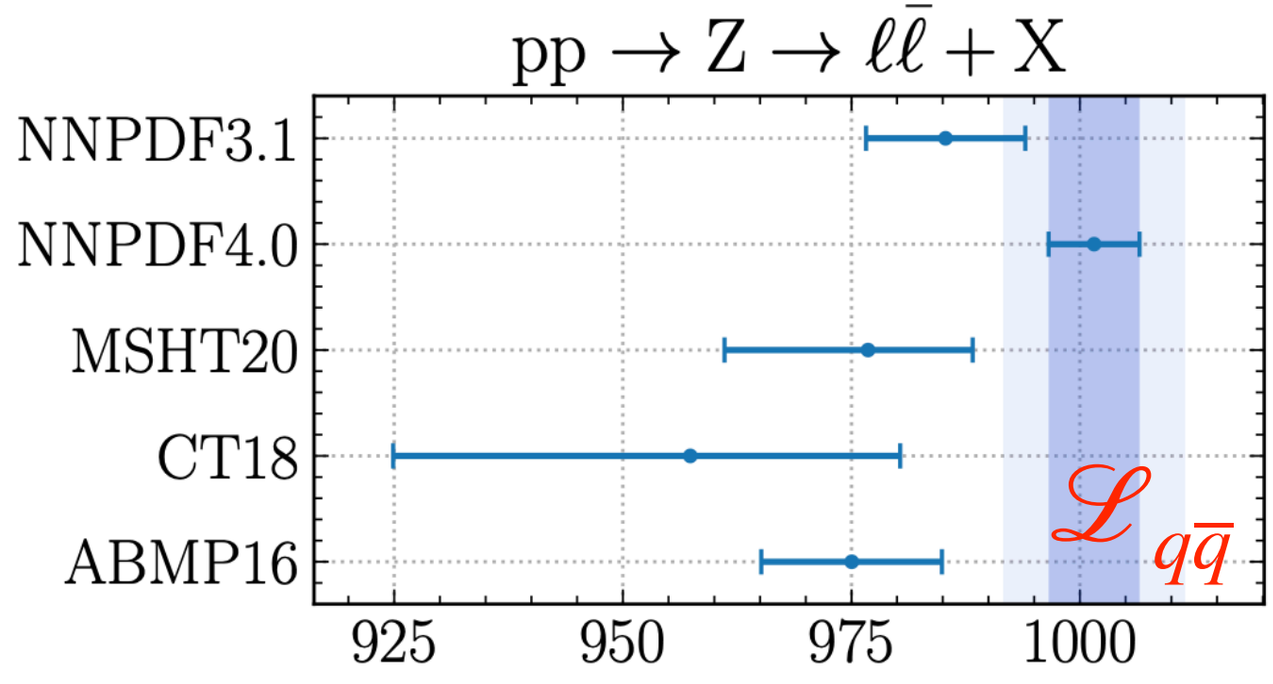


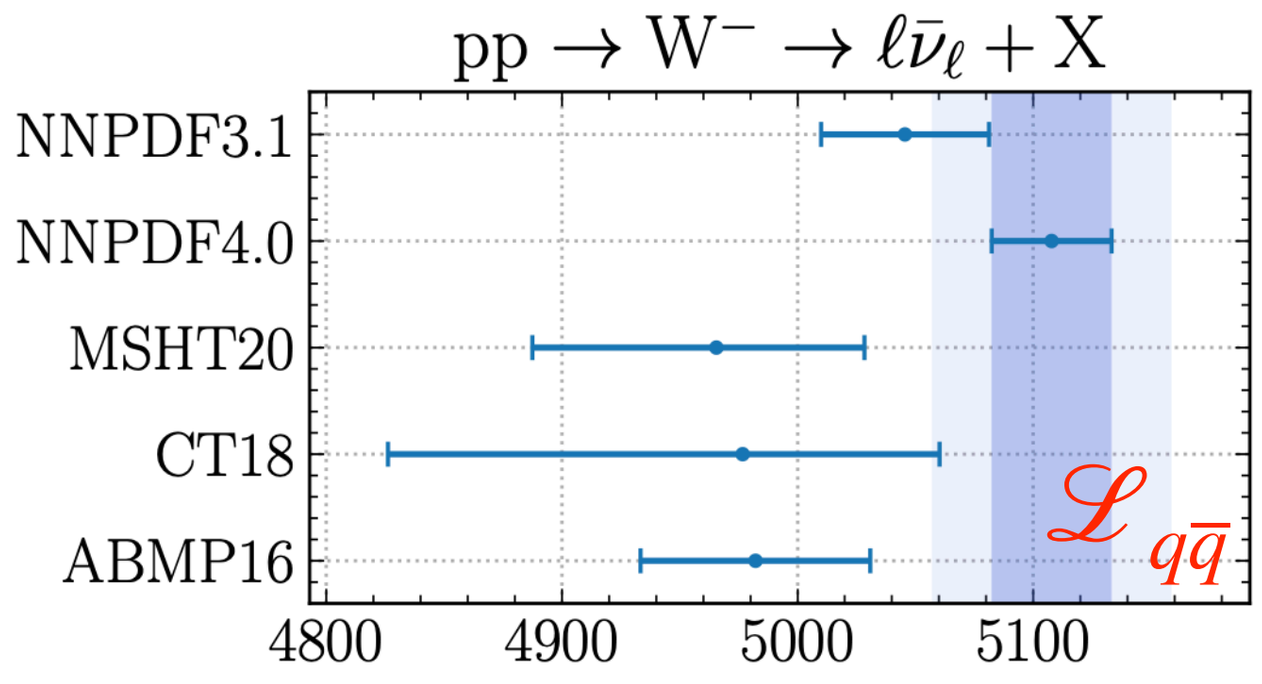


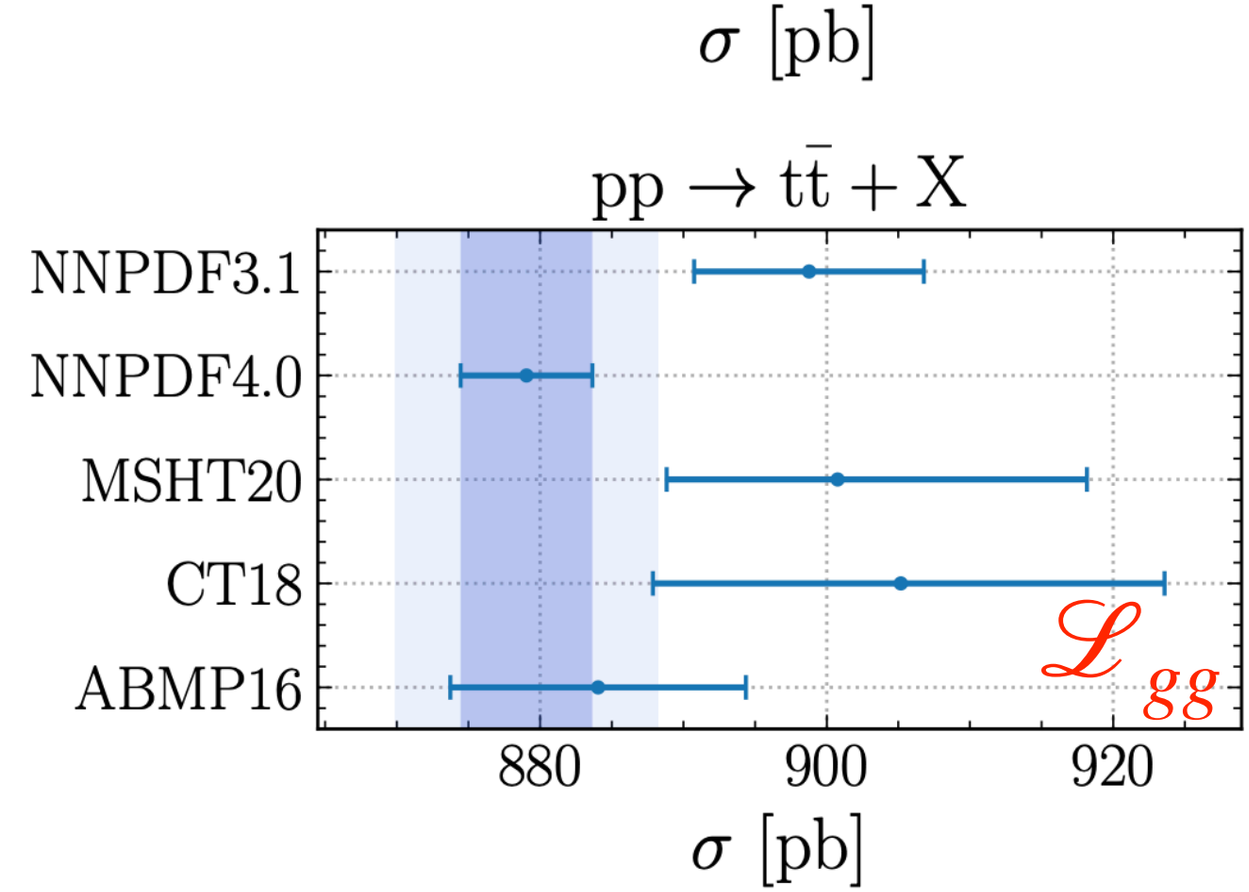
Significant impact on the $\bar{u}$ and $\bar{d}$ PDFs at large x.

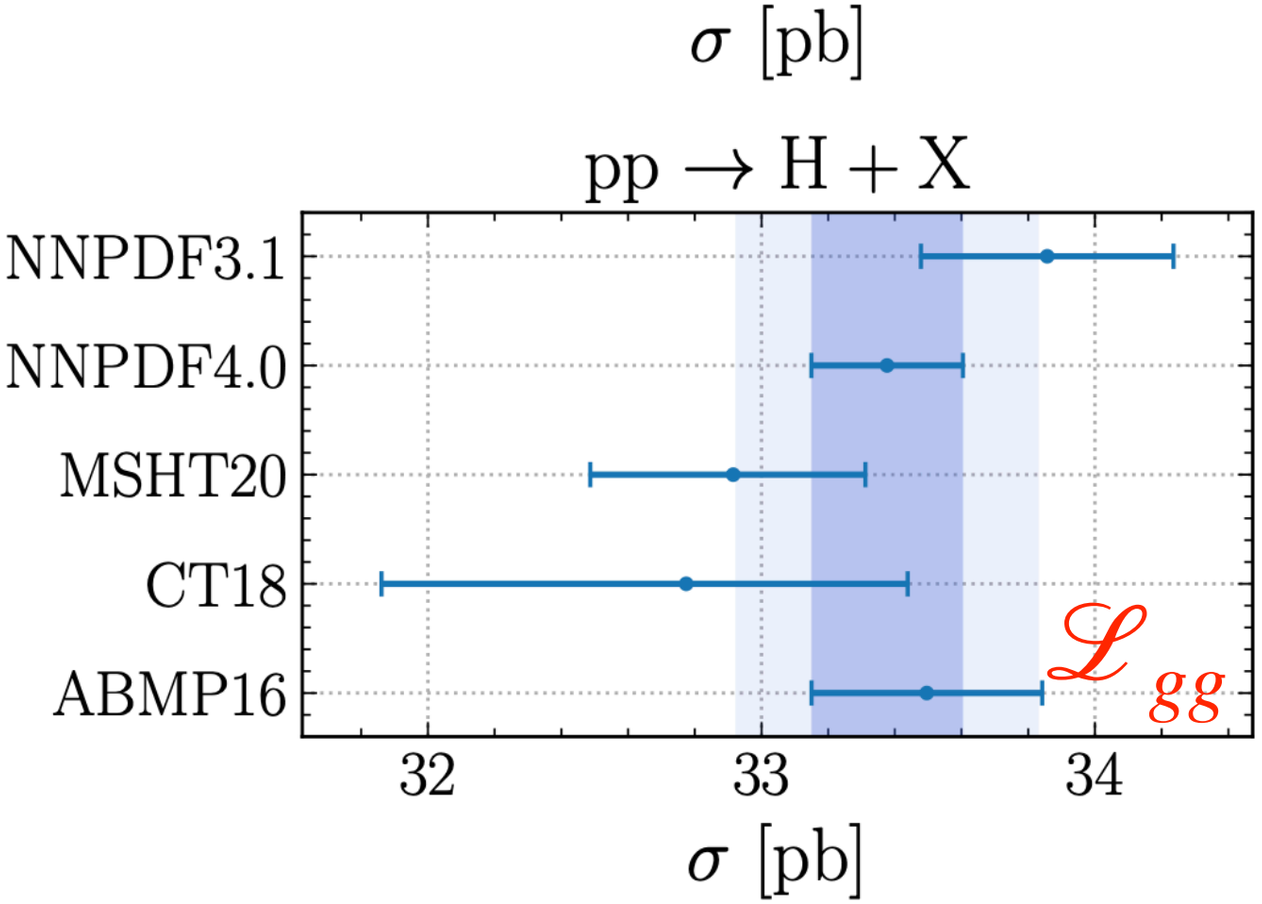
Currently unresolved tension with the older NuSea (E866) data.

A snapshot today









NNPDF4.0 [*Eur.Phys.J.C* 82 (2022) 5, 428]