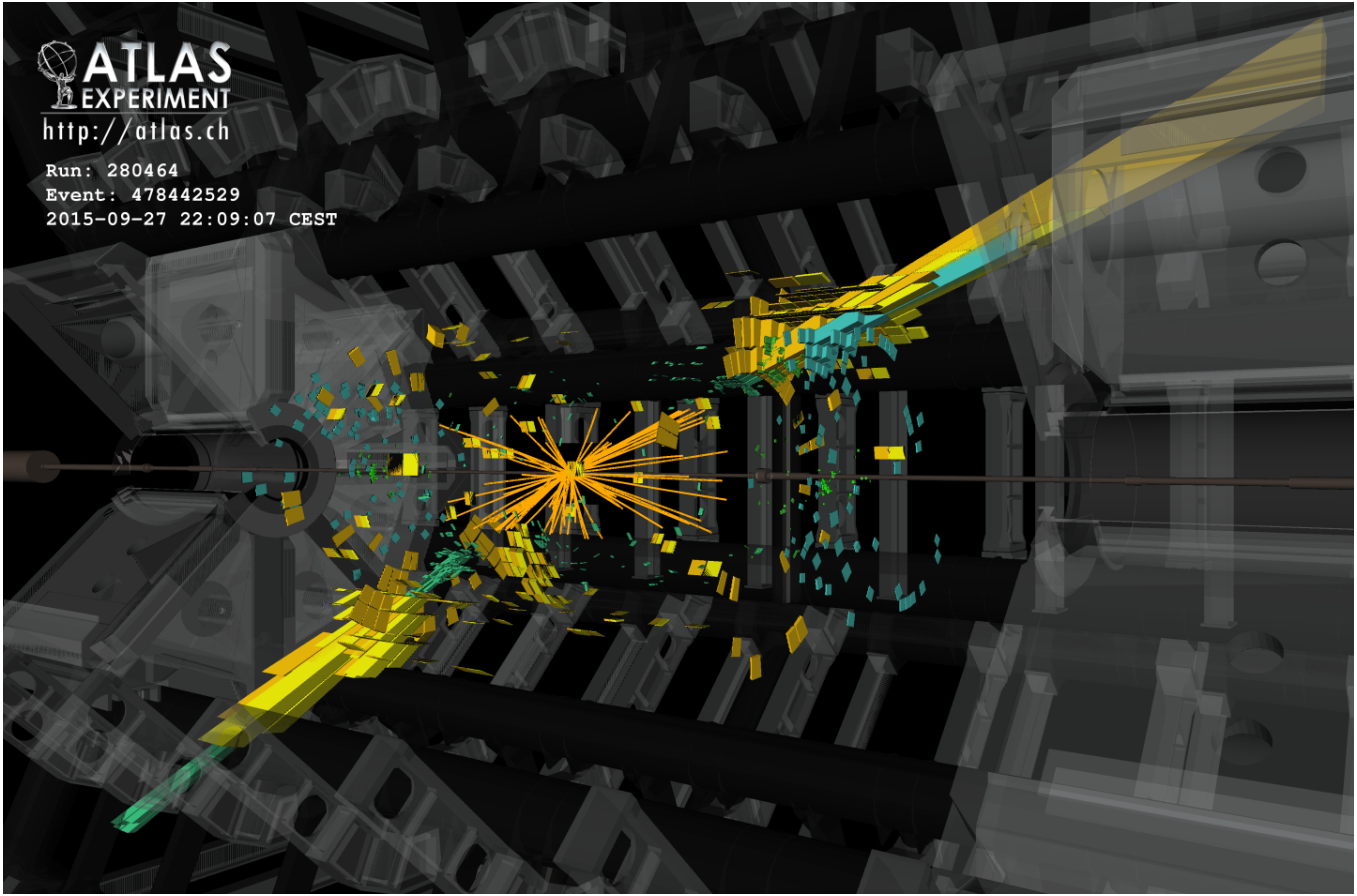
Parton Distribution Functions + Some Fun ML

Sanmay Ganguly
(sanmay@iitk.ac.in)

Machine Learning for High Energy Physics, 2026
TIFR, Mumbai

06/07/2026

How a collider event looks like?



What is our best explanation?

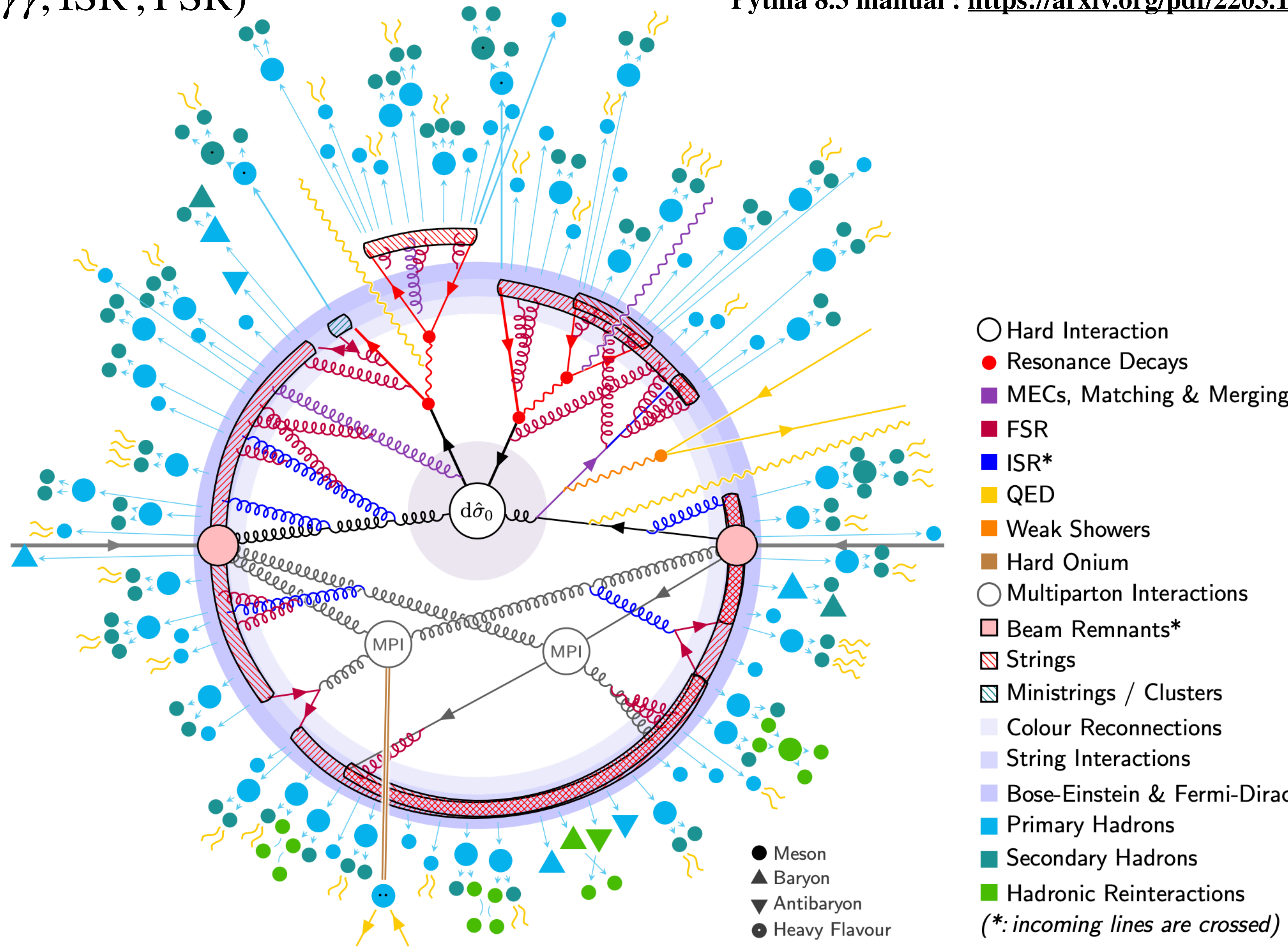
$\gamma\ (\pi^0 \rightarrow \gamma\gamma, \mathrm{ISR}, \mathrm{FSR})$

K_L, K_S

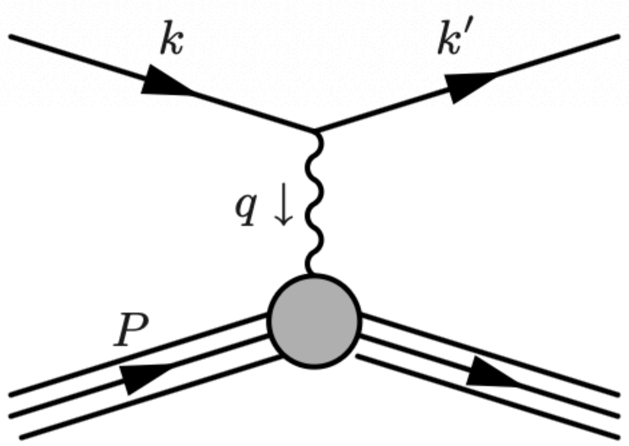
$\pi^{\pm}$

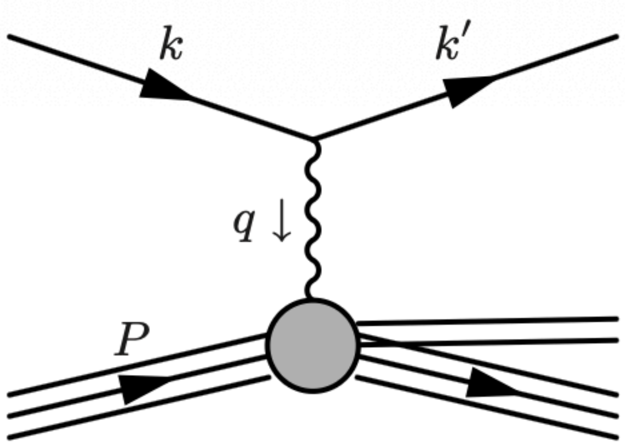
$l^{\pm}$

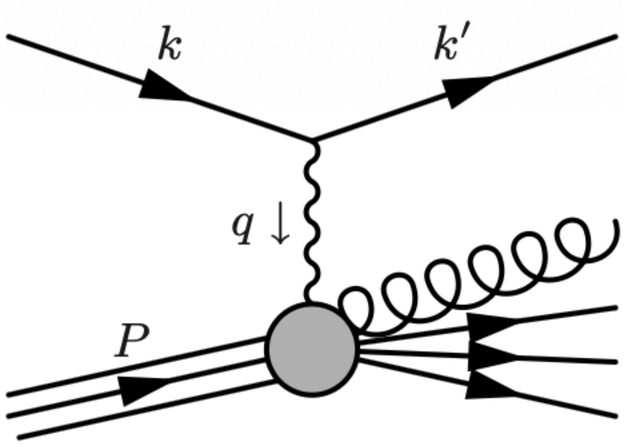
Pythia 8.3 manual : https://arxiv.org/pdf/2203.11601



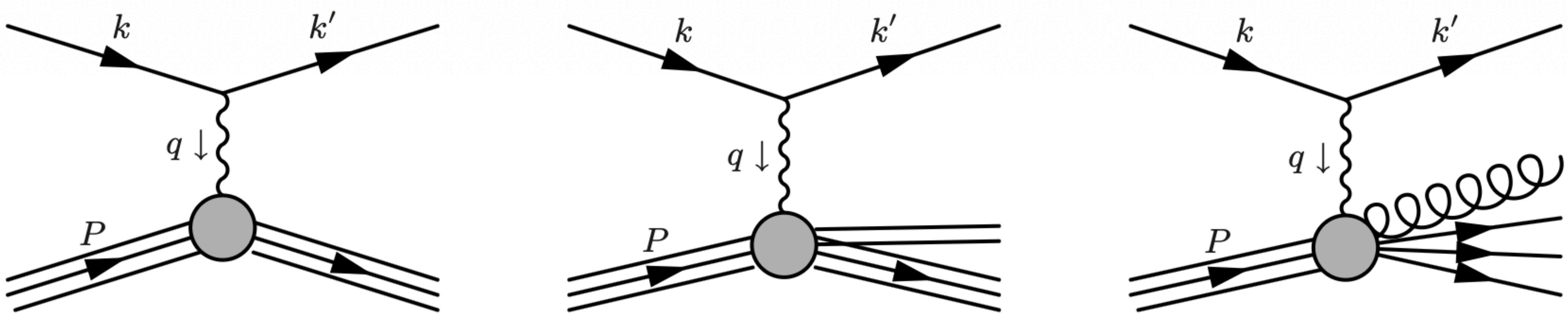
How do we try to make sense?





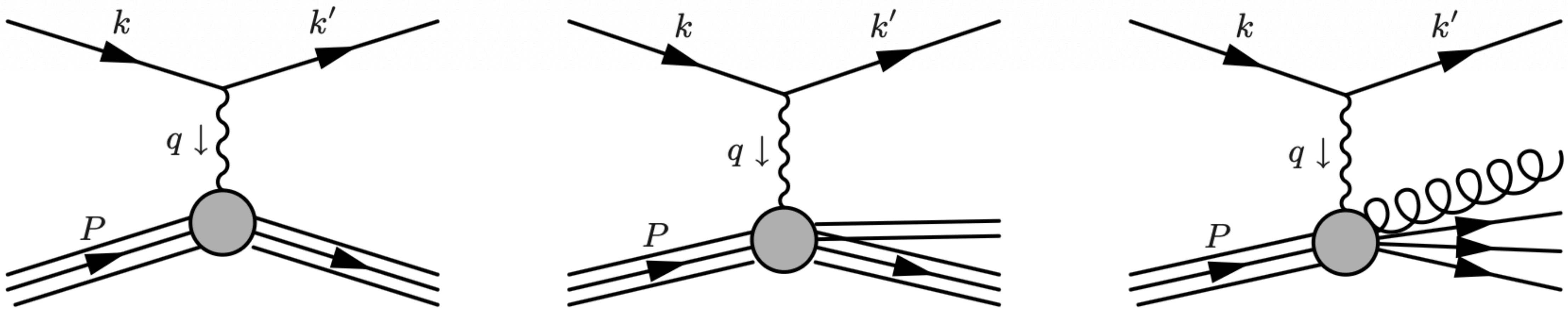


How do we try to make sense?



$$\boxed{\left(\frac{d\sigma}{d\Omega\, dE'}\right)_{\text{lab}} = \frac{\alpha_e^2}{8\pi E^2 \sin^4\frac{\theta}{2}}\left[\frac{m_p}{2} W_2(x,Q)\cos^2\frac{\theta}{2} + \frac{1}{m_p} W_1(x,Q)\sin^2\frac{\theta}{2}\right]}$$

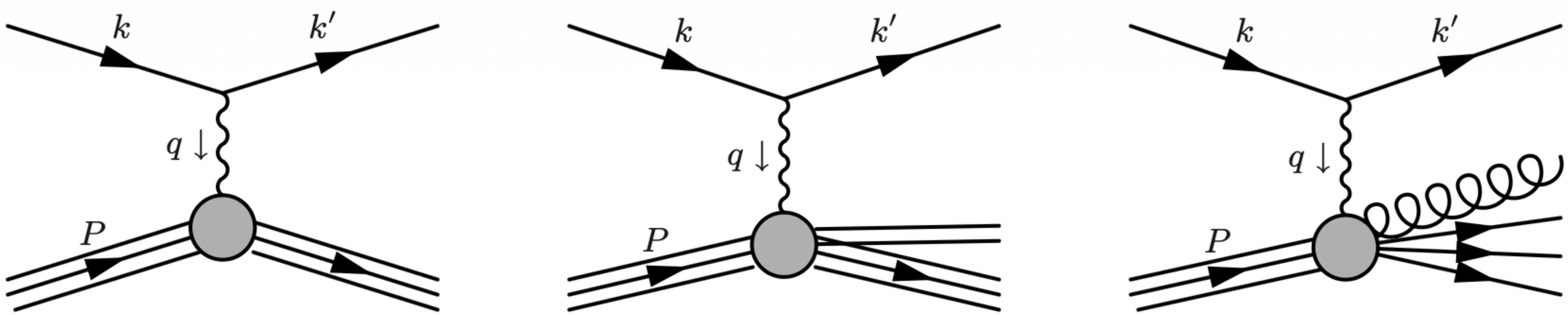
How do we try to make sense?



$$\boxed{\left(\frac{d\sigma}{d\Omega\, dE'}\right)_{\text{lab}} = \frac{\alpha_e^2}{8\pi E^2 \sin^4\frac{\theta}{2}}\left[\frac{m_p}{2} W_2(x,Q)\cos^2\frac{\theta}{2} + \frac{1}{m_p} W_1(x,Q)\sin^2\frac{\theta}{2}\right]}$$

$$\boxed{\sigma(e^- P^+ \to e^- X) = \sum_i \int_0^1 d\xi\, f_i(\xi)\,\hat{\sigma}(e^- p_i \to e^- X)}$$

How do we try to make sense?

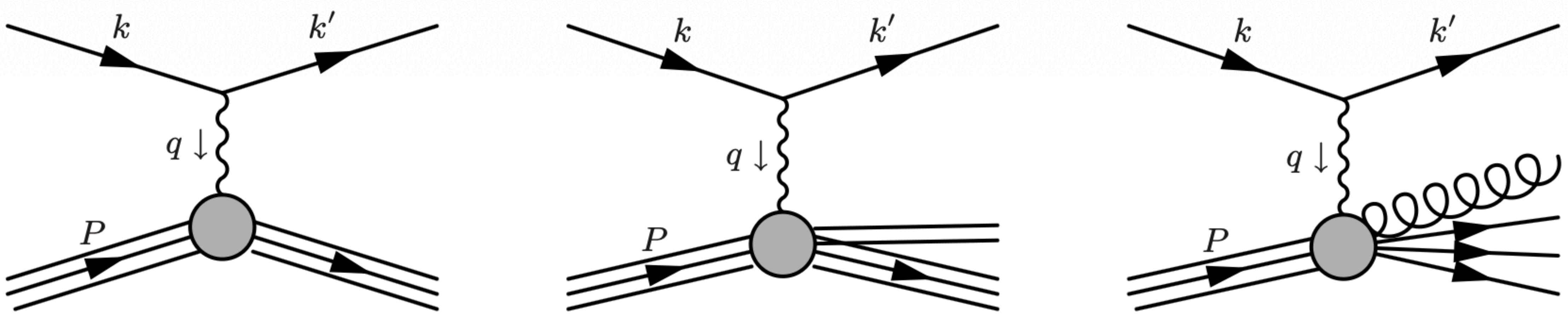


$$\left(\frac{d\sigma}{d\Omega\, dE'}\right)_{\text{lab}} = \frac{\alpha_e^2}{8\pi E^2 \sin^4\frac{\theta}{2}}\left[\frac{m_p}{2} W_2(x,Q)\cos^2\frac{\theta}{2} + \frac{1}{m_p} W_1(x,Q)\sin^2\frac{\theta}{2}\right]$$

$$\sigma(e^- P^+ \to e^- X) = \sum_i \int_0^1 d\xi\, f_i(\xi)\hat{\sigma}(e^- p_i \to e^- X)$$

$$\left(\frac{d\sigma(e^- P \to e^- X)}{d\Omega\, dE'}\right)_{\text{lab}} = \sum_i f_i(x)\frac{\alpha_e^2 Q_i^2}{4E^2\sin^4\frac{\theta}{2}}\left[\frac{2m_p}{Q^2}x^2\cos^2\frac{\theta}{2} + \frac{1}{m_p}\sin^2\frac{\theta}{2}\right]$$

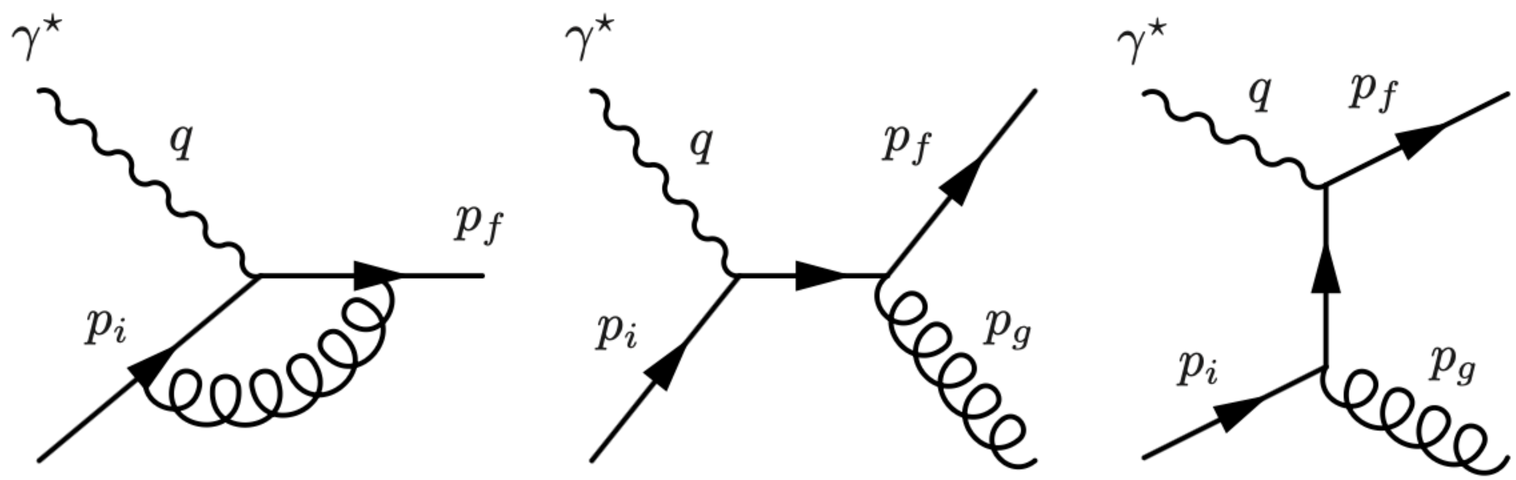
There are non-perturbative prescriptions



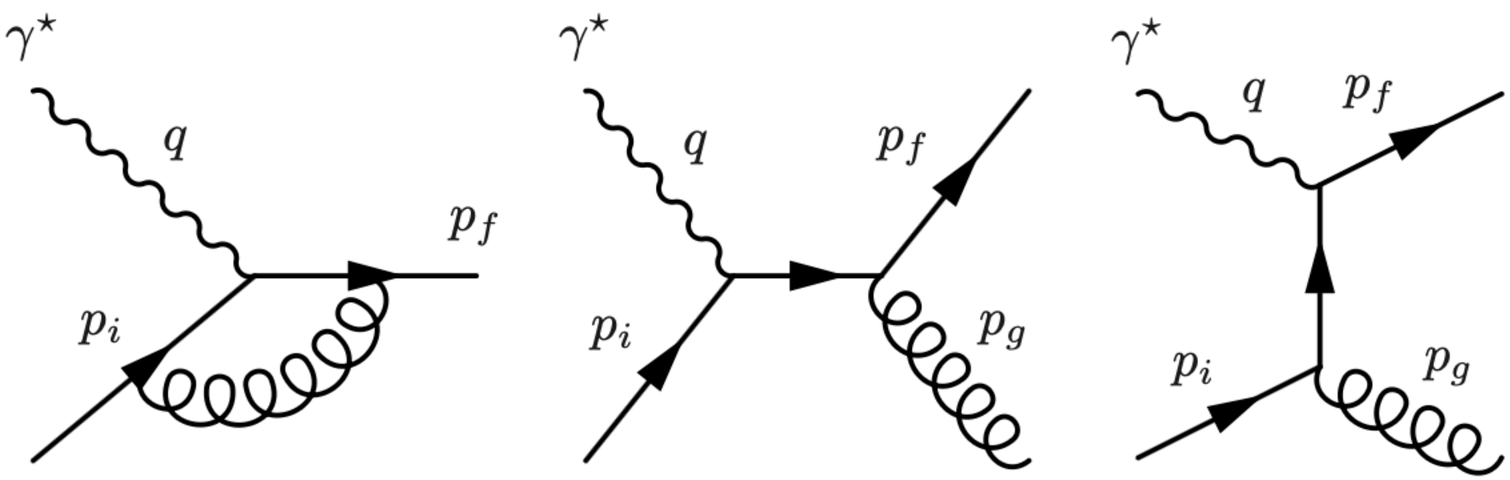
$$\boxed{W_n = P\exp\left\{ig_s n_\nu \int_0^t ds\, A^\nu(sn^\mu)\right\}}$$

$$f_q(\xi) = \int_{-\infty}^{\infty} \frac{dt}{2\pi} e^{-it\xi(n\cdot P)} \langle P|\bar{\psi}_q(tn^\mu)\, \frac{\not{n}}{2} W_n \psi_q(0)|P\rangle.$$

Some non-trivial properties of PDFs

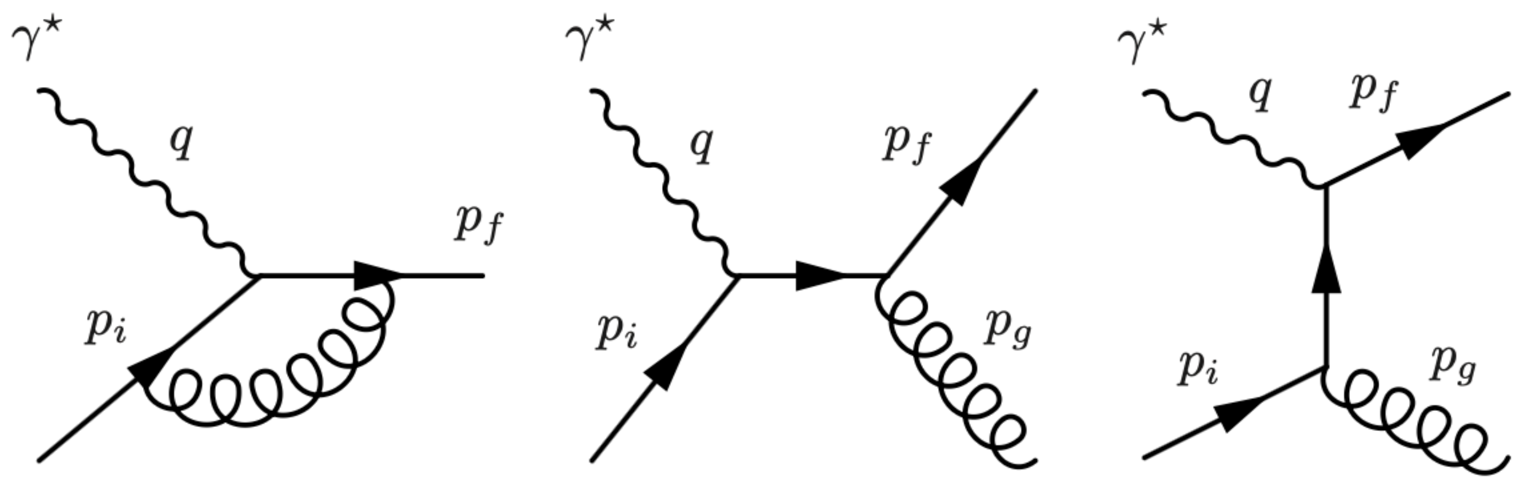


Some non-trivial properties of PDFs



$$\boxed{\mu\frac{d}{d\mu}\begin{pmatrix} f_i(x,\mu) \\ f_g(x,\mu)\end{pmatrix} = \sum_j \frac{\alpha_s}{\pi}\int_x^1 \frac{d\xi}{\xi}\begin{pmatrix} P_{q_iq_j}(\frac{x}{\xi}) & P_{q_ig}(\frac{x}{\xi}) \\ P_{gq_j}(\frac{x}{\xi}) & P_{gg}(\frac{x}{\xi})\end{pmatrix}\begin{pmatrix} f_j(\xi,\mu) \\ f_g(\xi,\mu)\end{pmatrix}}$$

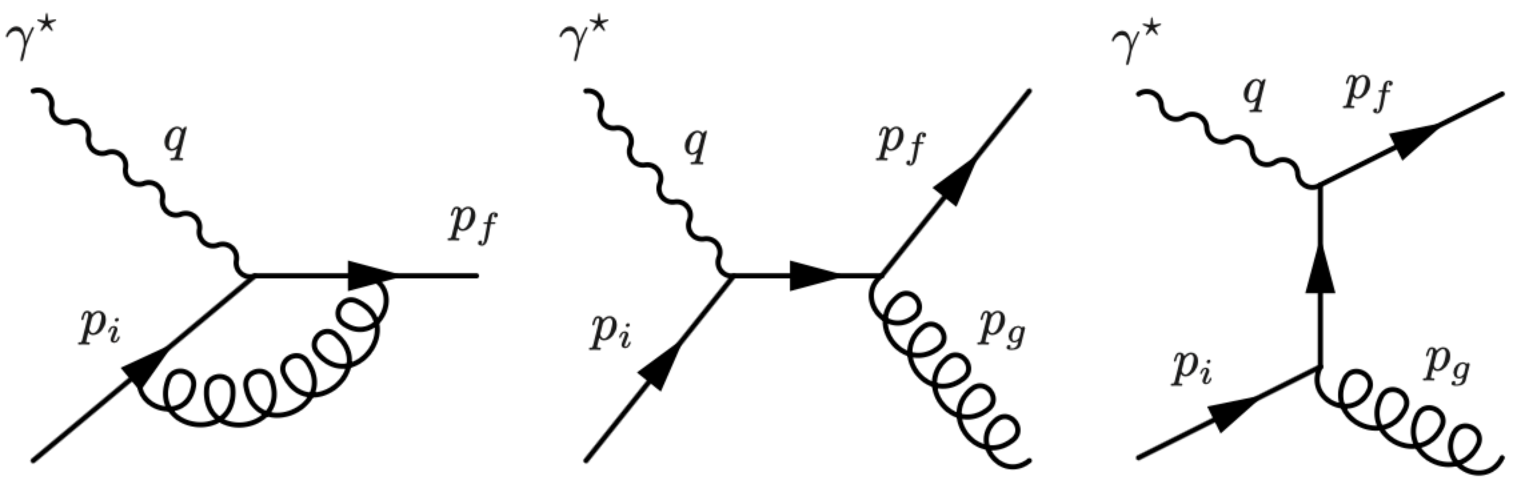
Some non-trivial properties of PDFs



$$\boxed{\mu \frac{d}{d\mu}\begin{pmatrix} f_i(x,\mu) \\ f_g(x,\mu) \end{pmatrix} = \sum_j \frac{\alpha_s}{\pi} \int_x^1 \frac{d\xi}{\xi} \begin{pmatrix} P_{q_i q_j}(\frac{x}{\xi}) & P_{q_i g}(\frac{x}{\xi}) \\ P_{g q_j}(\frac{x}{\xi}) & P_{gg}(\frac{x}{\xi}) \end{pmatrix} \begin{pmatrix} f_j(\xi,\mu) \\ f_g(\xi,\mu) \end{pmatrix}}$$

$$\boxed{P_{qq}(z) = C_F\left[\frac{1+z^2}{[1-z]_+} + \frac{3}{2}\delta(1-z)\right]}$$

Some non-trivial properties of PDFs

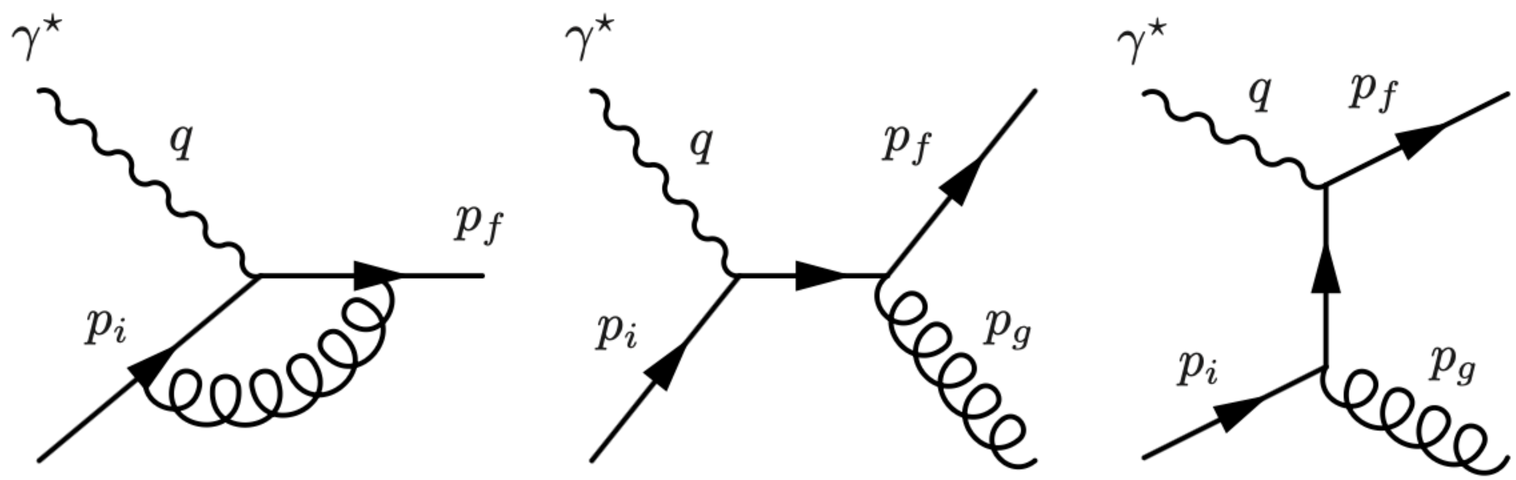


$$\boxed{\mu\frac{d}{d\mu}\begin{pmatrix} f_i(x,\mu) \\ f_g(x,\mu)\end{pmatrix} = \sum_j \frac{\alpha_s}{\pi}\int_x^1 \frac{d\xi}{\xi}\begin{pmatrix} P_{q_iq_j}(\frac{x}{\xi}) & P_{q_ig}(\frac{x}{\xi}) \\ P_{gq_j}(\frac{x}{\xi}) & P_{gg}(\frac{x}{\xi})\end{pmatrix}\begin{pmatrix} f_j(\xi,\mu) \\ f_g(\xi,\mu)\end{pmatrix}}$$

$$\boxed{P_{qq}(z) = C_F\left[\frac{1+z^2}{[1-z]_+} + \frac{3}{2}\delta(1-z)\right]}$$

$$\boxed{P_{qg}(z) = T_F\left[z^2 + (1-z)^2\right]}$$

Some non-trivial properties of PDFs



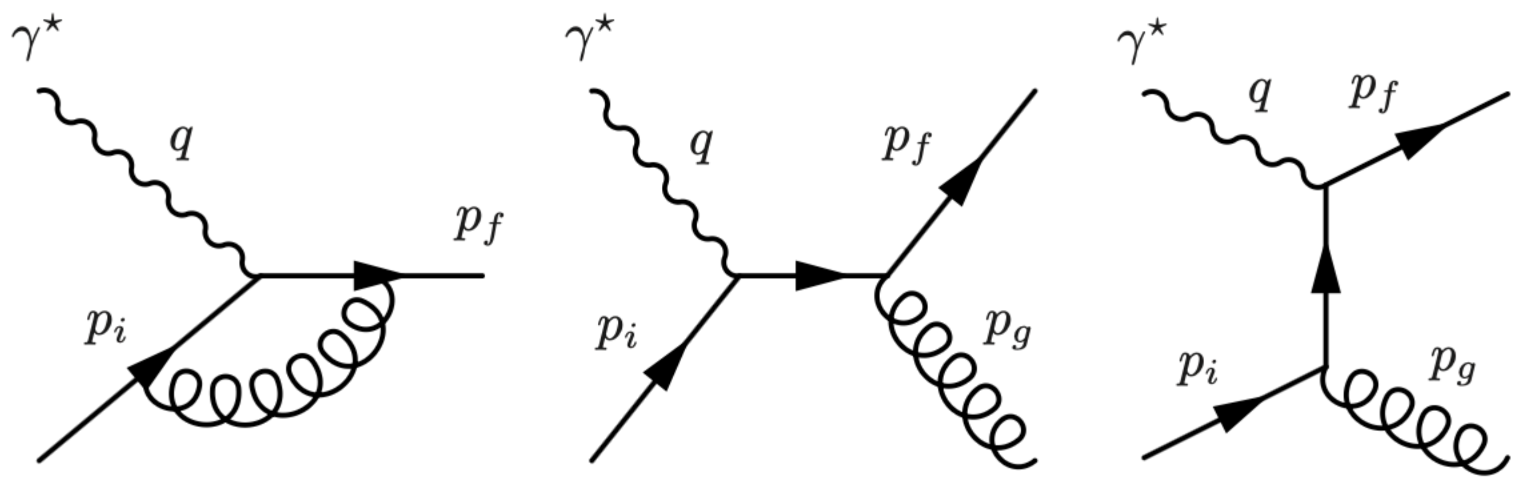
$$\boxed{\mu \frac{d}{d\mu} \begin{pmatrix} f_i(x,\mu) \\ f_g(x,\mu) \end{pmatrix} = \sum_j \frac{\alpha_s}{\pi} \int_x^1 \frac{d\xi}{\xi} \begin{pmatrix} P_{q_i q_j}(\frac{x}{\xi}) & P_{q_i g}(\frac{x}{\xi}) \\ P_{g q_j}(\frac{x}{\xi}) & P_{gg}(\frac{x}{\xi}) \end{pmatrix} \begin{pmatrix} f_j(\xi,\mu) \\ f_g(\xi,\mu) \end{pmatrix}}$$

$$\boxed{P_{qq}(z) = C_F \left[\frac{1+z^2}{[1-z]_+} + \frac{3}{2}\delta(1-z) \right]}$$

$$\boxed{P_{qg}(z) = T_F \left[z^2 + (1-z)^2 \right]}$$

$$\boxed{P_{gq}(z) = C_F \left[\frac{1+(1-z)^2}{z} \right]}$$

Some non-trivial properties of PDFs



$$\mu \frac{d}{d\mu}\begin{pmatrix} f_i(x,\mu) \\ f_g(x,\mu) \end{pmatrix} = \sum_j \frac{\alpha_s}{\pi}\int_x^1 \frac{d\xi}{\xi}\begin{pmatrix} P_{q_i q_j}(\frac{x}{\xi}) & P_{q_i g}(\frac{x}{\xi}) \\ P_{g q_j}(\frac{x}{\xi}) & P_{gg}(\frac{x}{\xi}) \end{pmatrix}\begin{pmatrix} f_j(\xi,\mu) \\ f_g(\xi,\mu) \end{pmatrix}$$

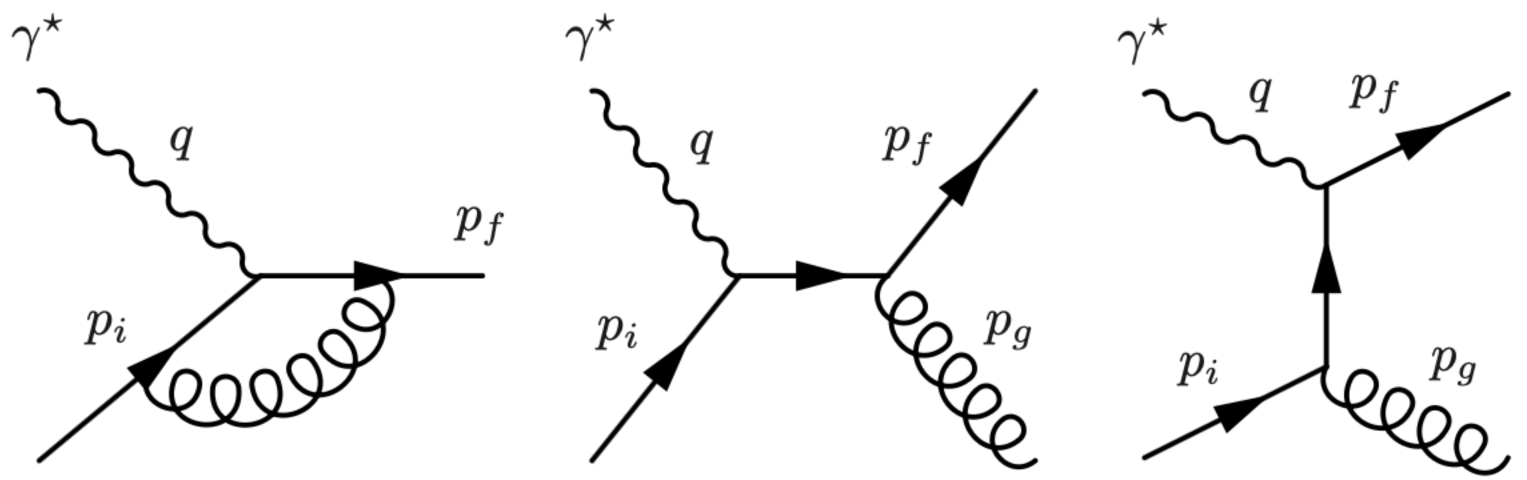
$$P_{qq}(z) = C_F\left[\frac{1+z^2}{[1-z]_+} + \frac{3}{2}\delta(1-z)\right]$$

$$P_{qg}(z) = T_F\left[z^2 + (1-z)^2\right]$$

$$P_{gq}(z) = C_F\left[\frac{1+(1-z)^2}{z}\right]$$

$$P_{gg}(z) = 2C_A\left[\frac{z}{[1-z]_+} + \frac{1-z}{z} + z(1-z)\right] + \frac{\beta_0}{2}\delta(1-z)$$

Some non-trivial properties of PDFs



$$\mu \frac{d}{d\mu}\begin{pmatrix} f_i(x,\mu) \\ f_g(x,\mu)\end{pmatrix} = \sum_j \frac{\alpha_s}{\pi}\int_x^1 \frac{d\xi}{\xi}\begin{pmatrix} P_{q_i q_j}(\frac{x}{\xi}) & P_{q_i g}(\frac{x}{\xi}) \\ P_{g q_j}(\frac{x}{\xi}) & P_{gg}(\frac{x}{\xi})\end{pmatrix}\begin{pmatrix} f_j(\xi,\mu) \\ f_g(\xi,\mu)\end{pmatrix}$$

$$P_{qq}(z) = C_F\left[\frac{1+z^2}{[1-z]_+} + \frac{3}{2}\delta(1-z)\right]$$

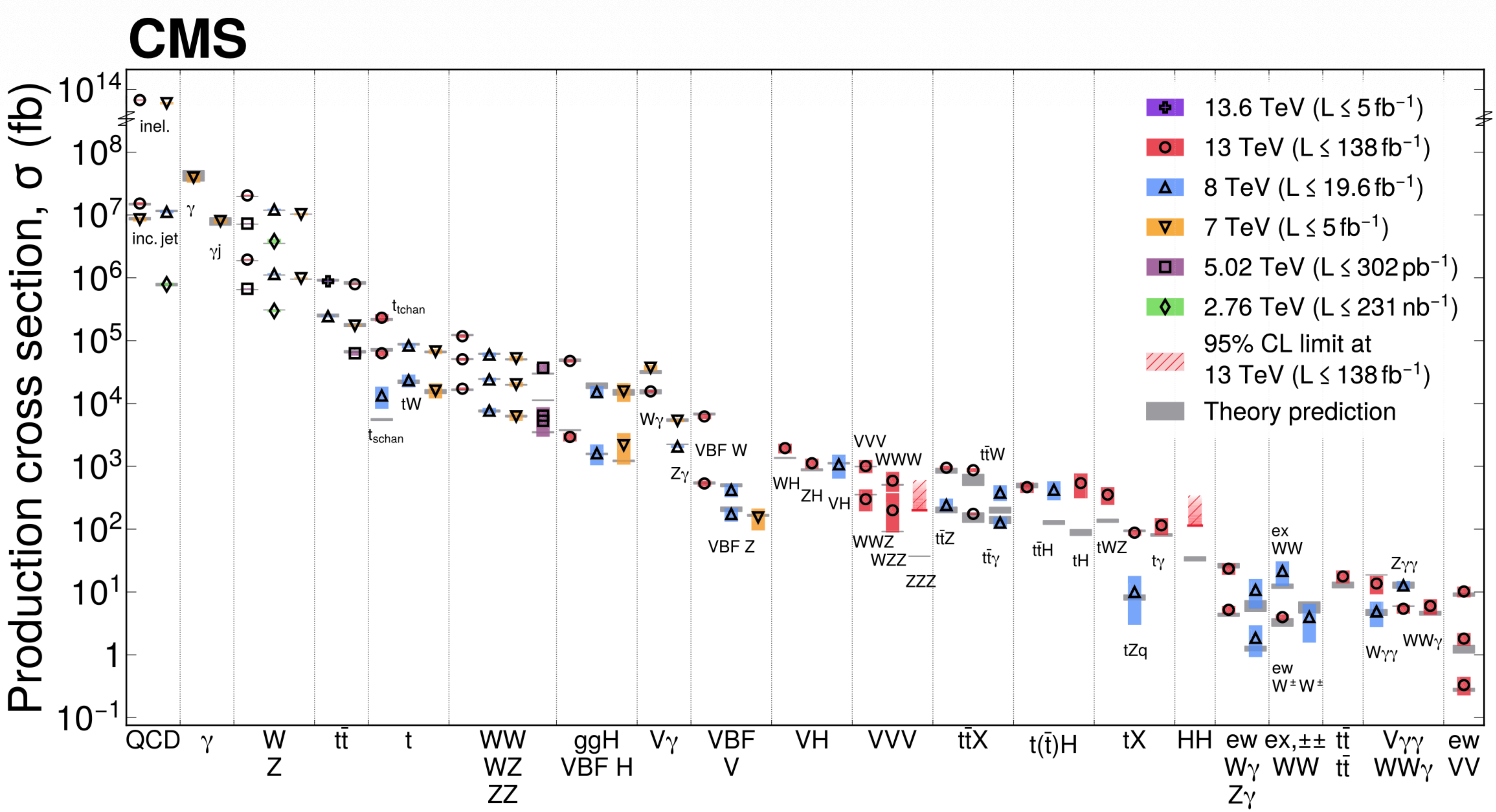
$$P_{qg}(z) = T_F\left[z^2 + (1-z)^2\right]$$

$$P_{gq}(z) = C_F\left[\frac{1+(1-z)^2}{z}\right]$$

$$P_{gg}(z) = 2C_A\left[\frac{z}{[1-z]_+} + \frac{1-z}{z} + z(1-z)\right] + \frac{\beta_0}{2}\delta(1-z)$$

$$\beta_0 = \tfrac{11}{3}C_A - \tfrac{4}{3}T_F n_f .$$

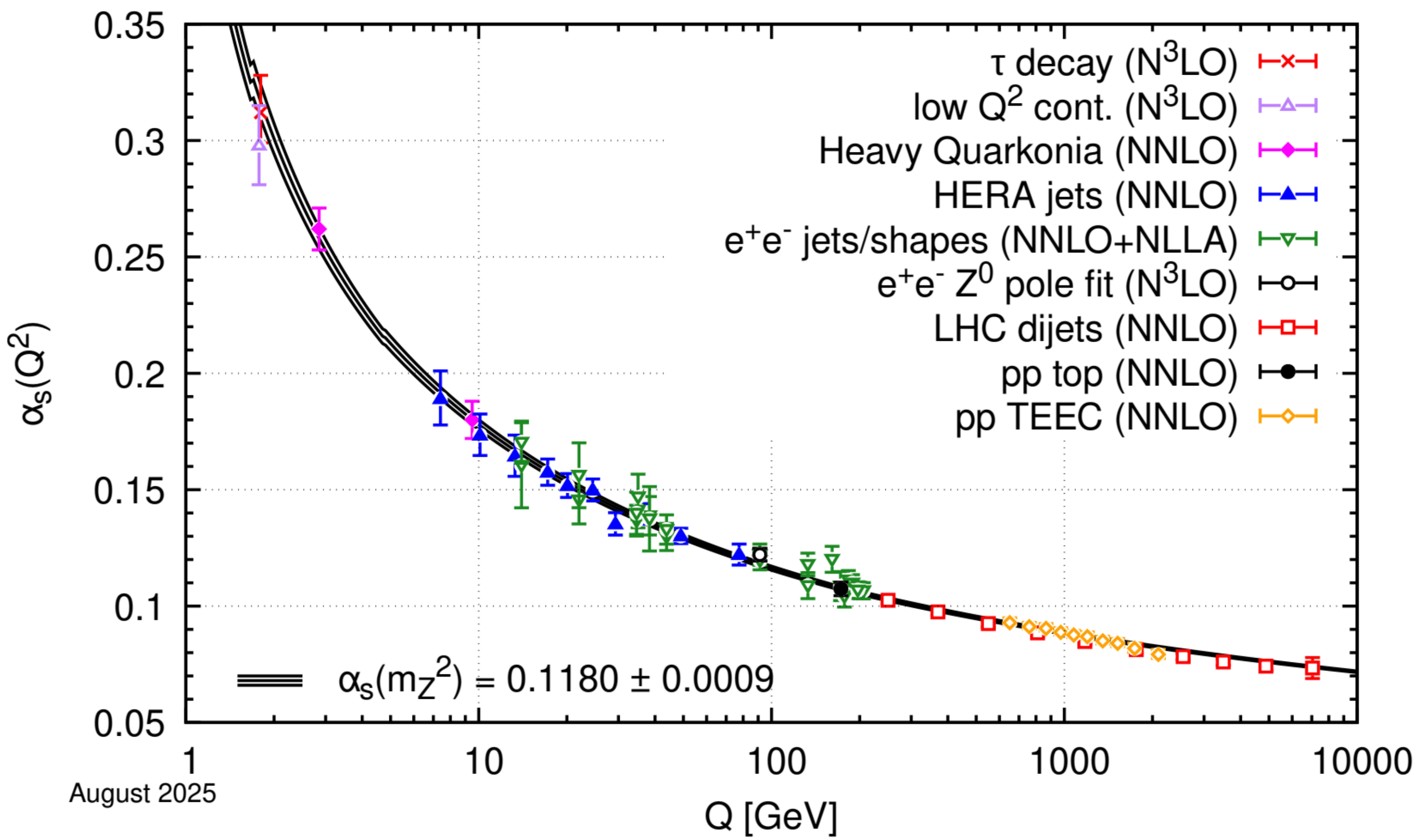
The utility of PDF's

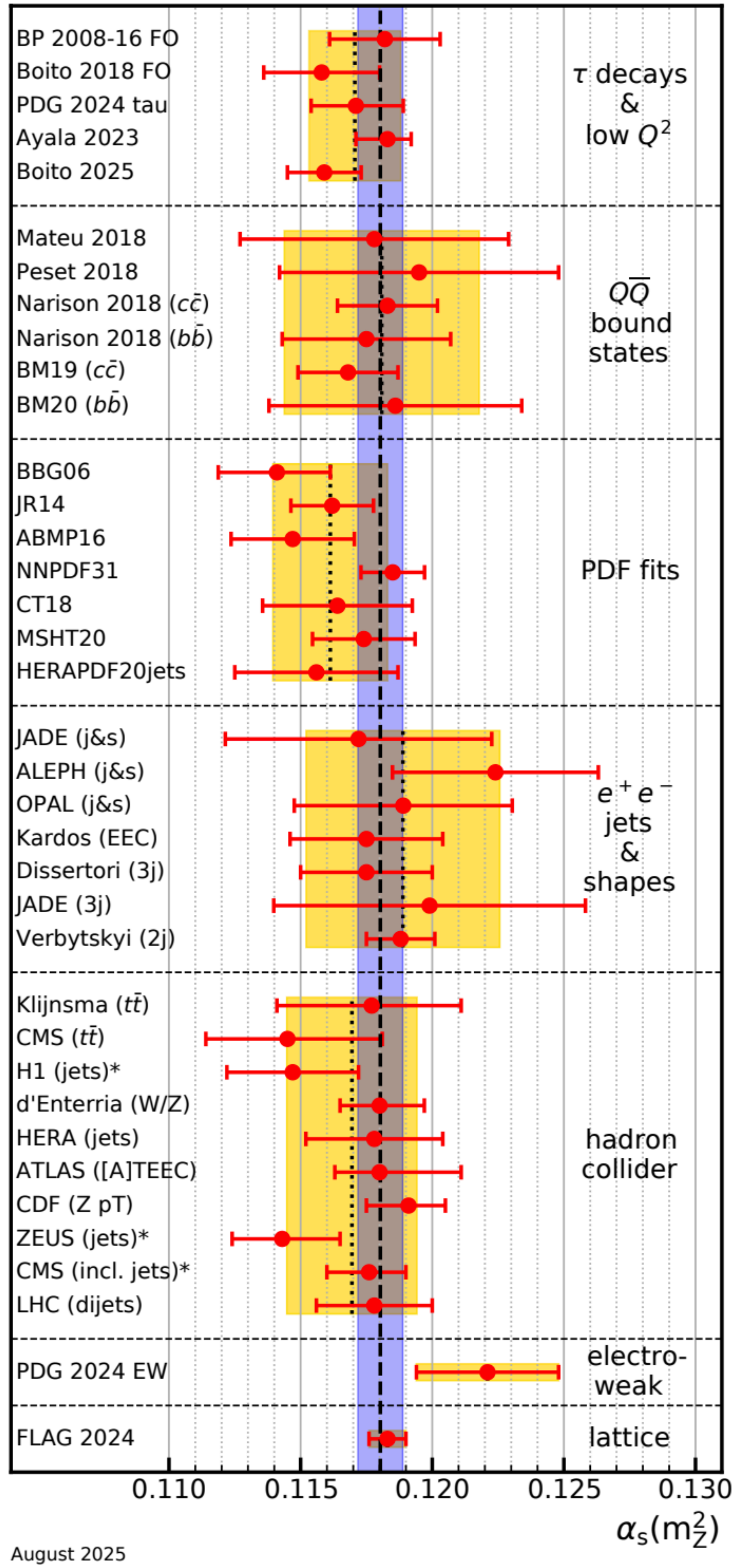


$$\sigma(h_1 h_2 \to W + X) = \sum_{n=0}^{\infty} \alpha_s^n(\mu_R^2) \sum_{i,j} \int dx_1 dx_2\, f_{i/h_1}\left(x_1, \mu_F^2\right) f_{j/h_2}\left(x_2, \mu_F^2\right)$$

$$\times\, \hat{\sigma}^{(n)}_{ij\to W+X}\left(x_1 x_2 s, \mu_R^2, \mu_F^2\right) + \mathcal{O}\left(\frac{\Lambda^2}{m_W^4}\right),$$

What is required?





What does PDF fitting mean?

- choose a factorisation scheme (e.g. $\overline{\mathrm{MS}}$), an order in perturbation theory (LO, NLO, NNLO) and a 'starting scale' Q_0 where pQCD applies (e.g. 1-2 GeV)
- parametrise the quark and gluon distributions at $Q_{0,}$, e.g.

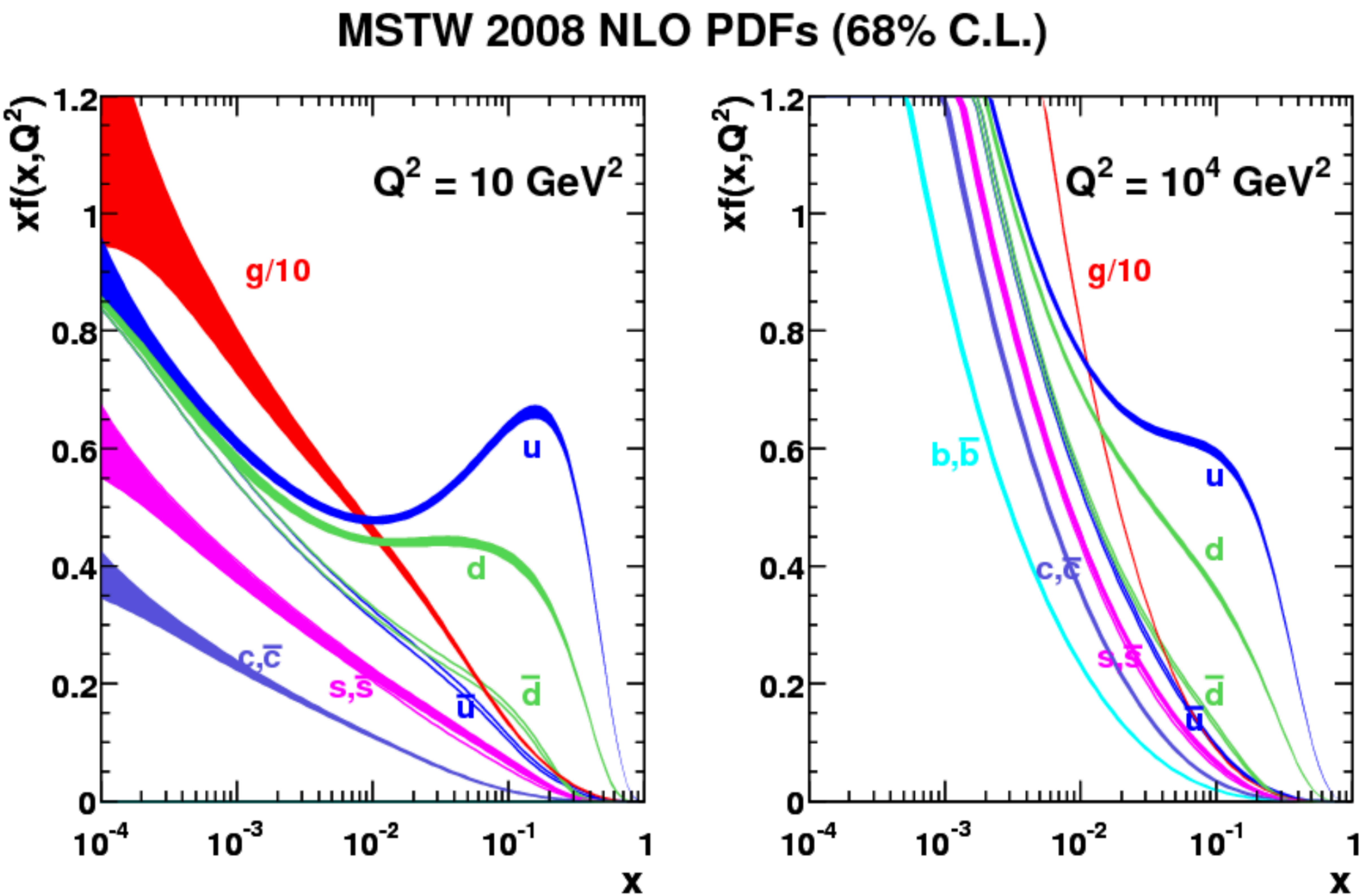
$$f_i(x, Q_0^2) = A_i x^{a_i}[1 + b_i\sqrt{x} + c_i x](1-x)^{d_i}$$

- solve DGLAP equations to obtain the pdfs at any x and scale $Q > Q_0$; fit data for parameters $\{A_i, a_i, \ldots \alpha_S\}$
- approximate the exact solutions (e.g. interpolation grids, expansions in polynomials etc) for ease of use; thus the output 'global fits' are available 'off the shelf", e.g.

```
SUBROUTINE PDF(X,Q,U,UBAR,D,DBAR,…,BBAR,GLU)
```

input | output

How do the PDF's look like?



PDF with uncertainties

- most groups produce ‘pdfs with errors’
- typically, 20-40 ‘error’ sets based on a ‘best fit’ set to reflect 1σ variation of all the parameters* $\{A_i, a_i, \ldots, \alpha_S\}$ inherent in the fit
- these reflect the uncertainties on the data used in the global fit (e.g. $\delta F_2 \approx 3\% \rightarrow \delta u \approx 3\%$)
- however, there are also systematic pdf uncertainties reflecting theoretical assumptions/prejudices in the way the global fit is set up and performed (see earlier slide)

* e.g. $f_i(x, Q_0^2) = A_i x^{a_i}[1 + b_i\sqrt{x} + c_i x](1-x)^{d_i}$

Global PDF fit parametrization

- Two distinct methodologies on the market to parameterising PDFs: **Neural Nets** (NNPDF) or **Explicit Parameterisation** (CT, MSHT).

$$f_i(x, Q_0): \ A_f x^{a_f}(1-x)^{b_f} \times \ \longrightarrow \ \sum_{i=1}^{n} \alpha_{f,i} P_i(y(x)) \,, \ \textbf{CT, MSHT...}$$

$$\searrow \ \mathrm{NN}_i(x) \quad \textbf{NNPDF}$$

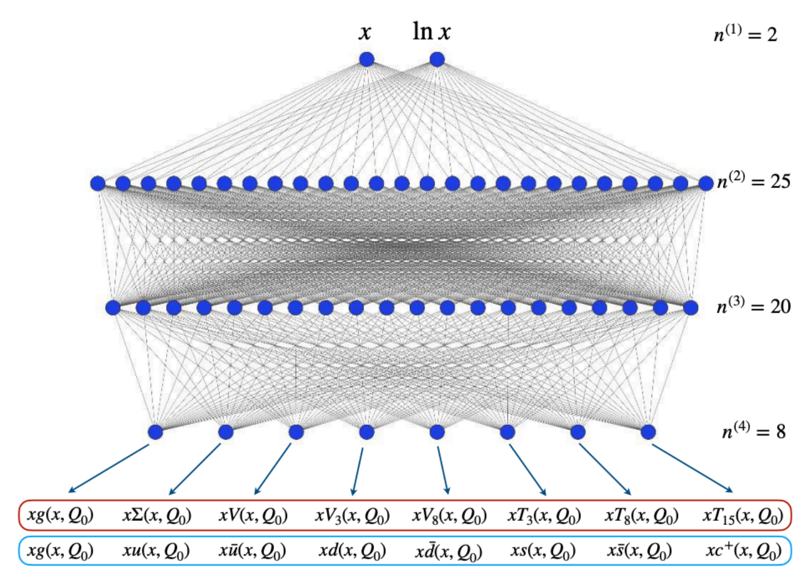
  - ✦ MSHT20: **52** free parameters in terms of Chebyshev polynomials.

$$xf(x, Q_0) = Ax^{\delta}(1-x)^{\eta}\left(1 + \sum_{i=1}^{6} a_i T_i(y(x))\right) \,,$$

  - ✦ CT18: **29** free parameters in term of Bernstein polynomials. Variations considered and PDF uncertainty expanded to account for these.

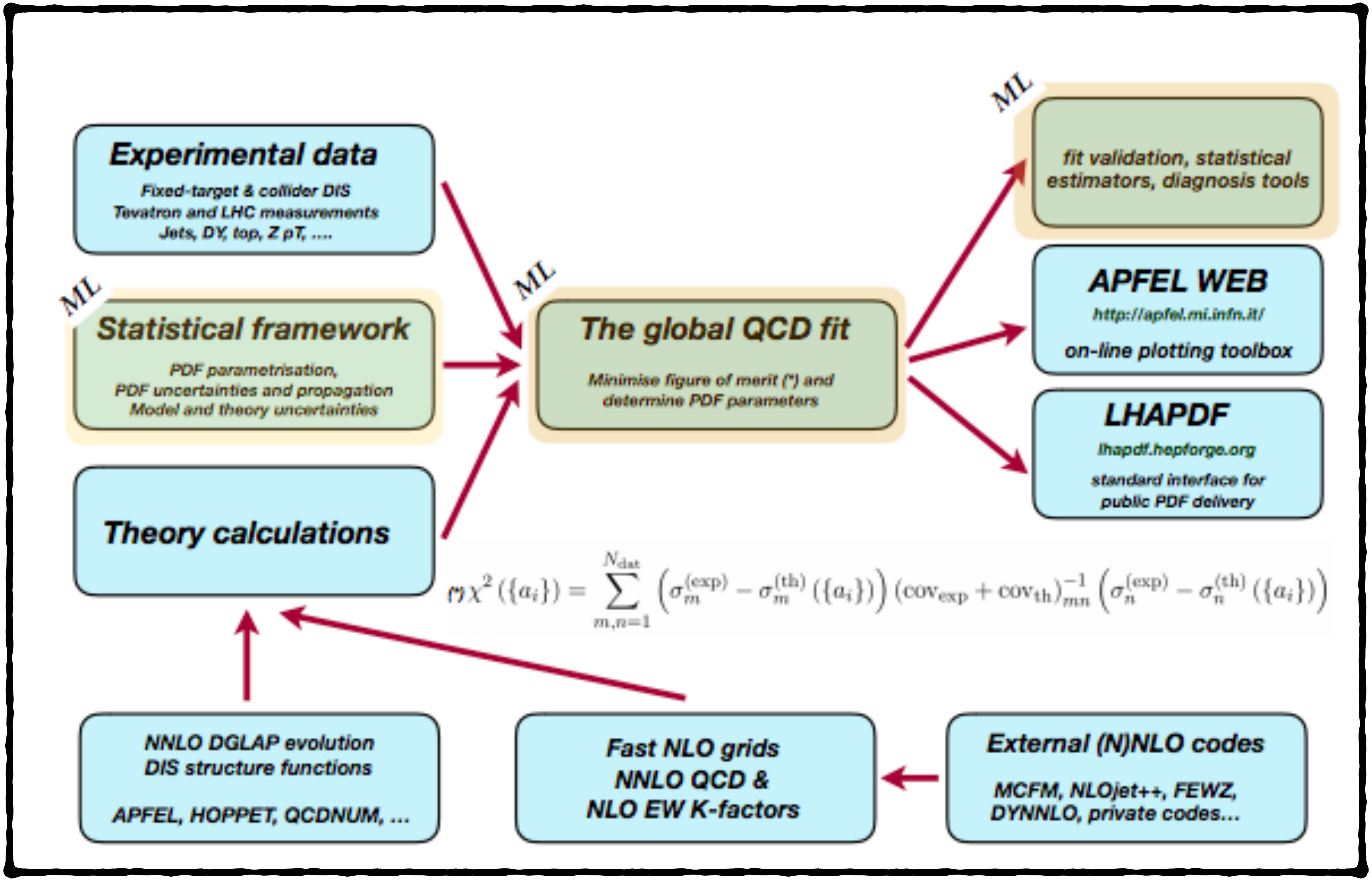
  - ✦ Less flexible in general - need to be sure flexible enough! Allows direct handle on uncertainties in Hessian framework.

- ★ NNPDF4.0: ***763*** free parameter Neural Net.



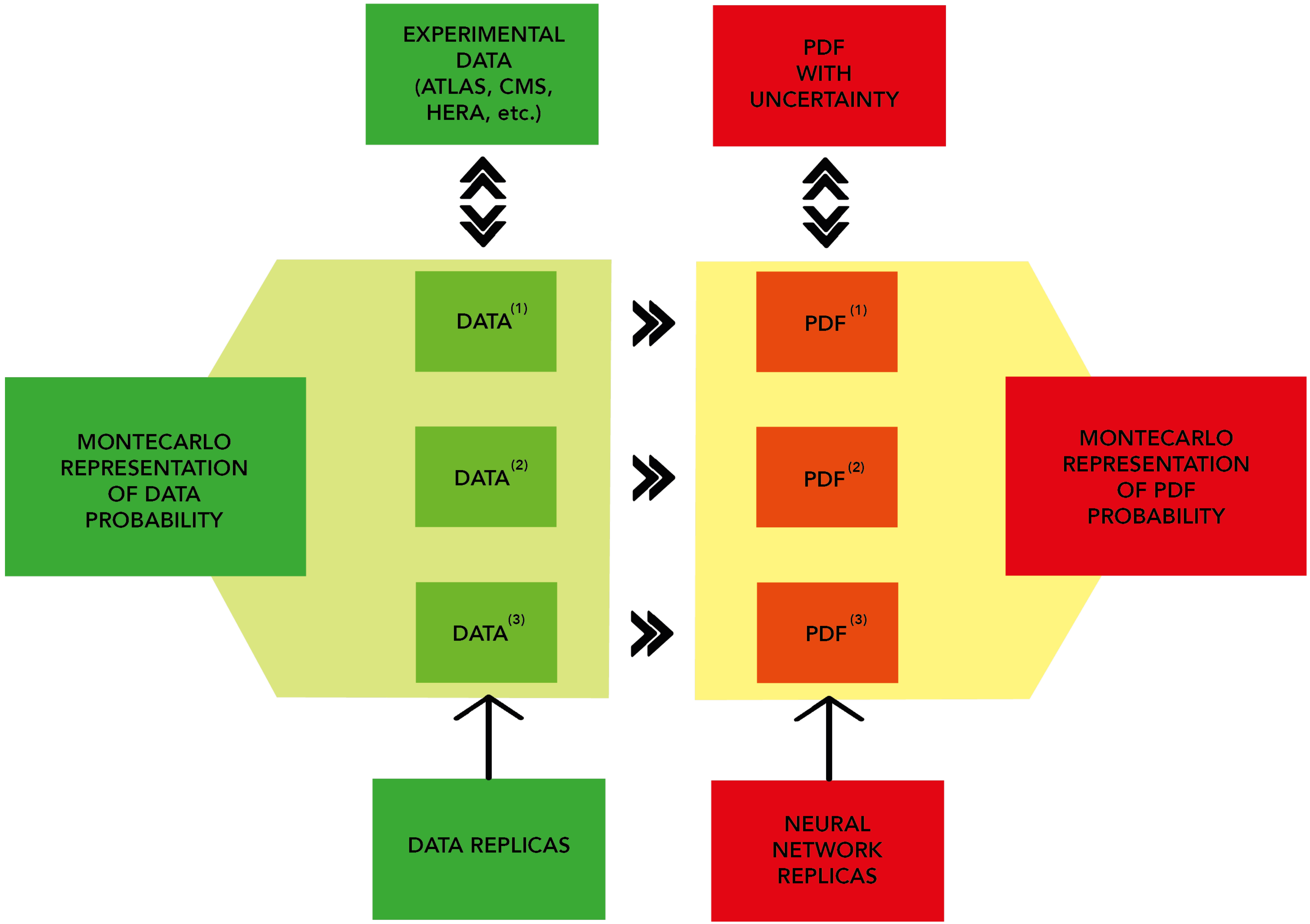
- ★ Increased flexibility, but needs robust optimisation + stopping (avoid over and under fitting).

Steps adapted with NNPDF - 1



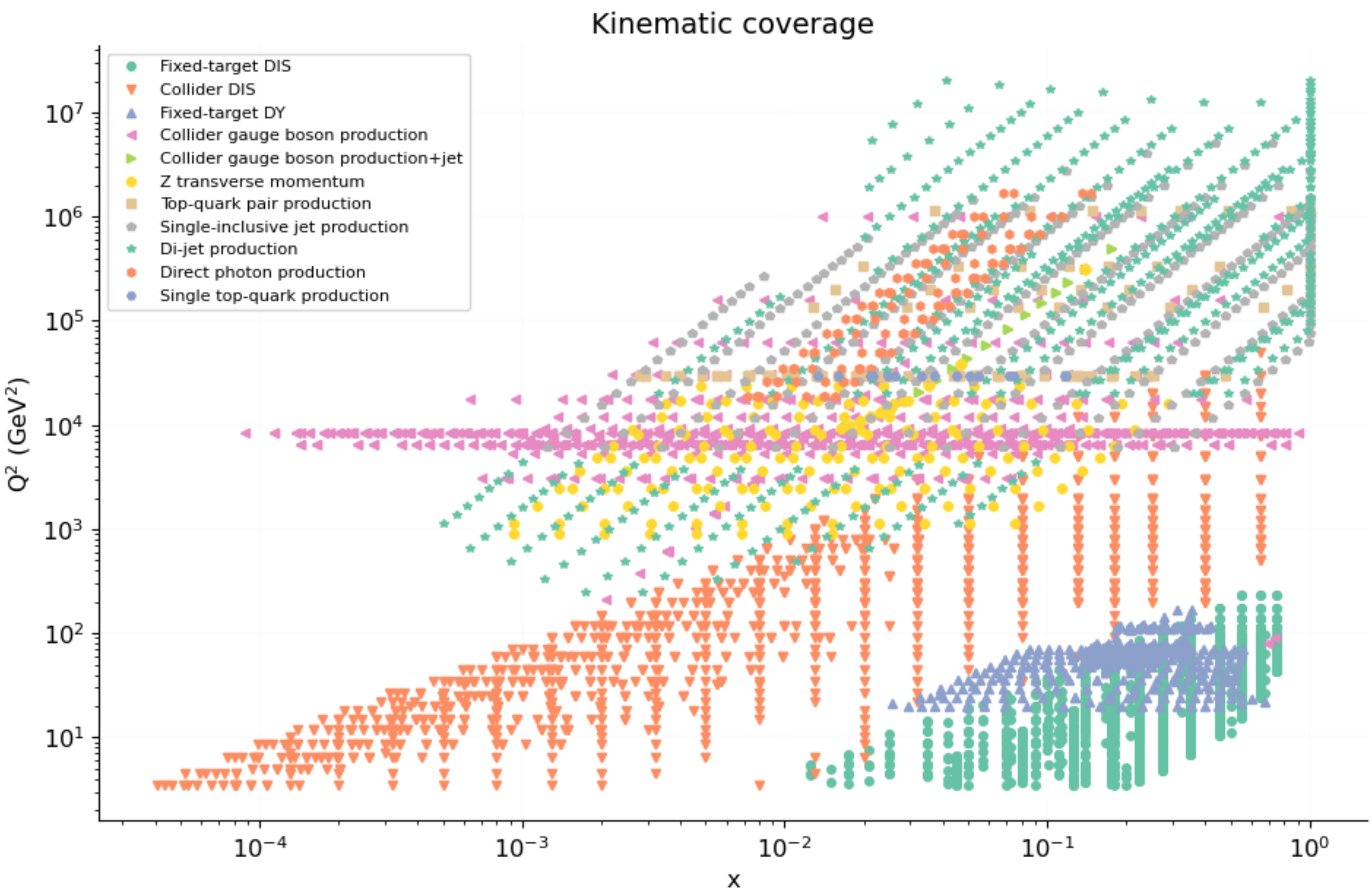
The NN based approach is cross verified with parametric fits.

Steps adapted with NNPDF - 2



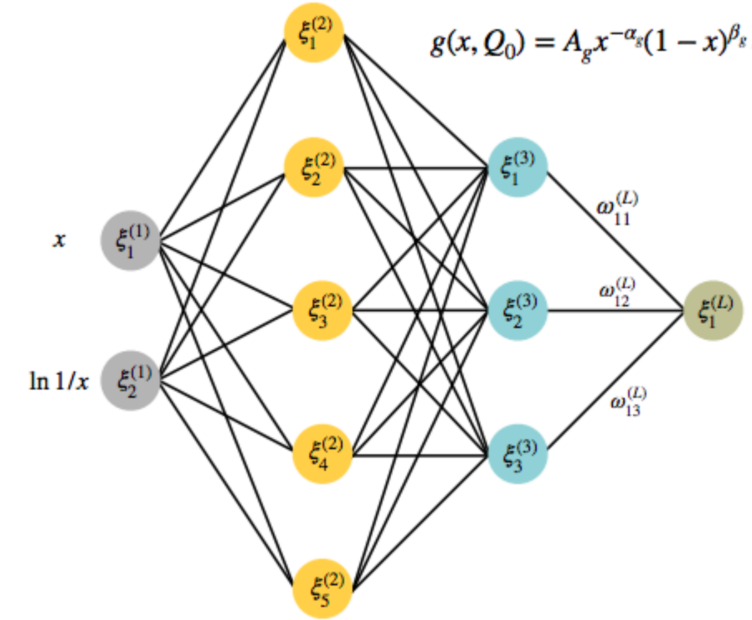
The data replicas are generated from experimental mean & variance.

Datasets in use



NNPDF optimization

$$g(x, Q_0) = A_g x^{-\alpha_g}(1-x)^{\beta_g}\, \xi_L^{(1)}(x)$$



Try harder!

New Fitting Methodology

Define Underlying Physical Law
ie **input PDFs from MSTW08, CT10, NNPDF2.3...**

Generate random pseudo-data for the NNPDF3.0 dataset from info of experimental uncertainties and correlations

Perform (NN)PDF fit

Validate resulting PDF set:
- ☑ **Reproduce input PDFs**
- ☑ **Both central values and uncertainties**
- ☑ **Expected values of χ^2 are determined by pseudo-data**
- ☑ **PDF reweighting equal to refitting (Bayesian inference)**

Fail?

OK!

Closure Test successful!

Now you can fit real exp data!

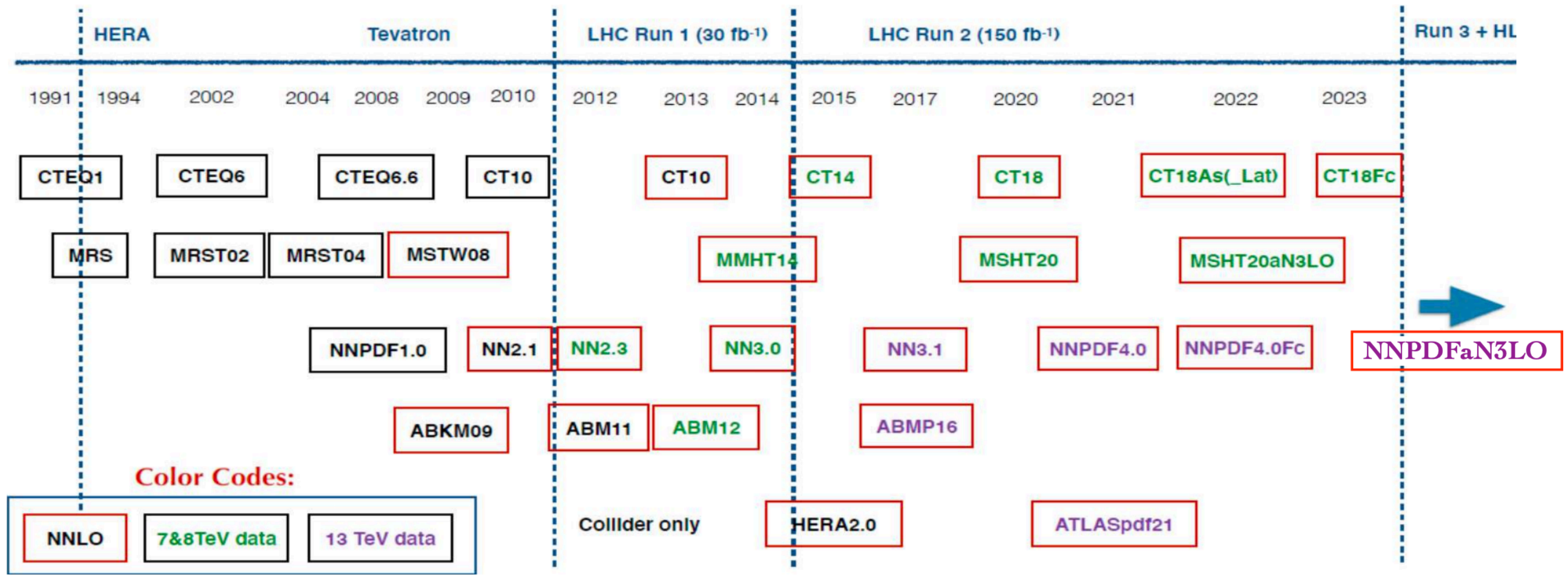
Our current hold on PDF sets

- Multiple PDF analyses, with different methodogies and datasets.
- Major releases from 3 global fitters (**CT**, **MSHT**, **NNPDF**) ~ 2 or more years ago. But they have been busy:

★ Major push to approximate N3LO + theoretical uncertainties

★ QED/EW corrections standard

★ Many dedicated studies



Where do we stand?