THE UNIVERSITY of EDINBURGH

Proton structure from NSBI at the LHC

Based on [2604.13157] - R. Barrue, L. Benato, A. K. Guven, E. Hammou, **J. ter Hoeve**, C. Krause, A. Li, L. Mantani, J. Rojo, S.S. Cruz, R. Schoefbeck, M Ubiali and D. Wang

ML4Jets 2026

Jaco ter Hoeve
17/09/2026

NNPDF

Parton distribution functions for precision measurements

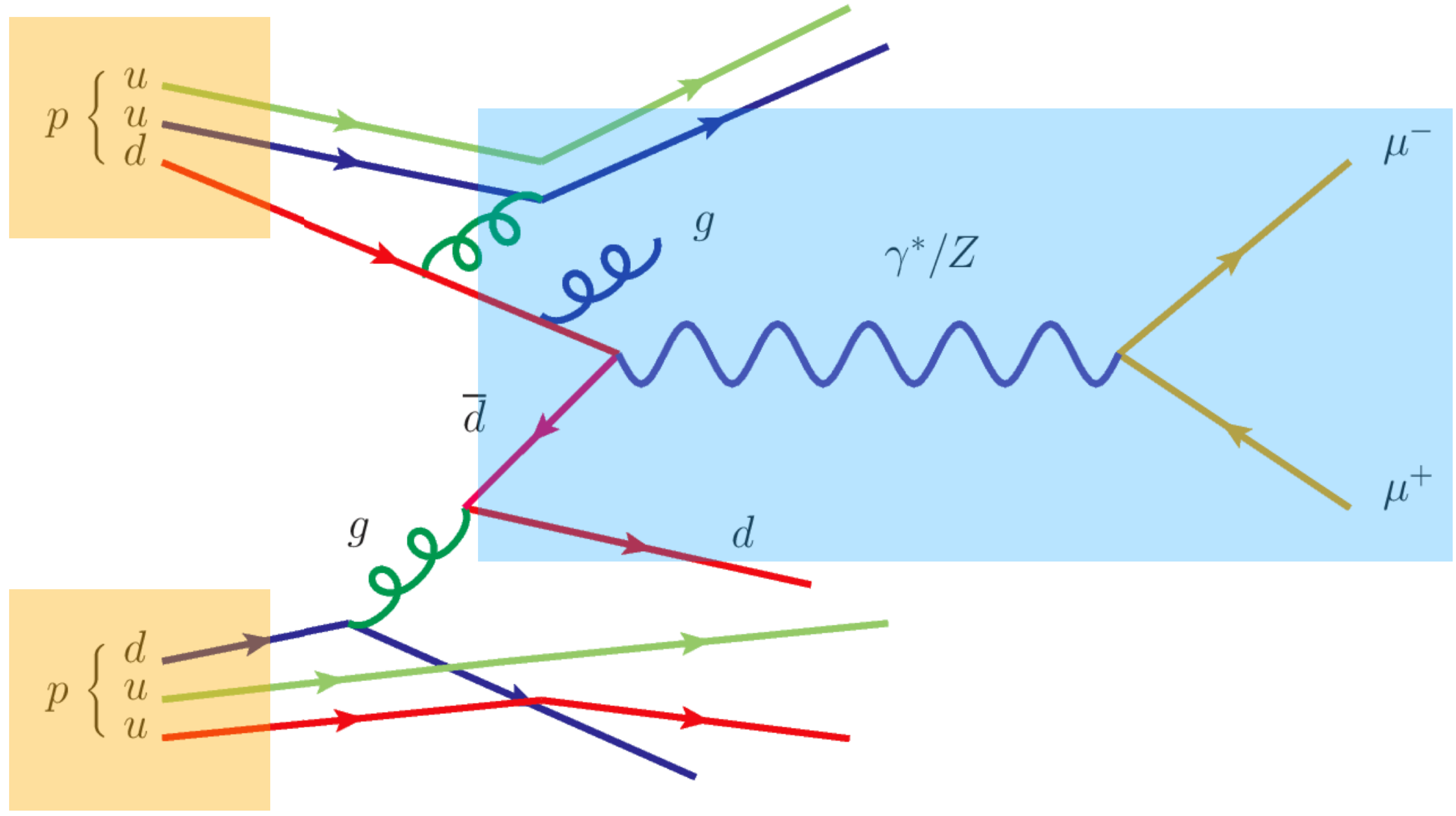
PDFs are key input to many SM and BSM calculations at the LHC

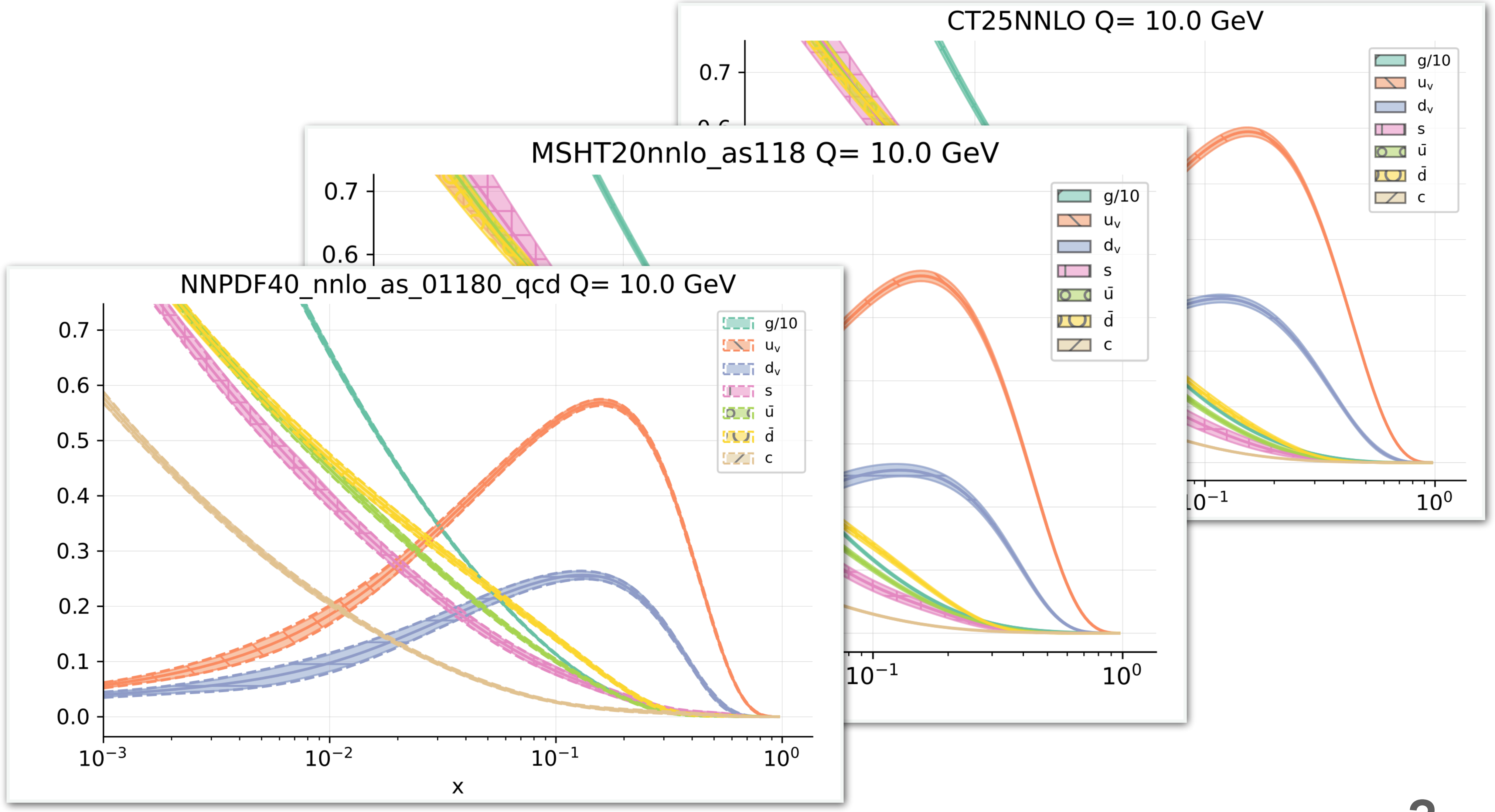
$$d\sigma^{pp\to ab} = \sum_{i,j} f_i \otimes f_j \otimes d\hat{\sigma}^{ij\to ab} + \mathcal{O}\left(\Lambda^2/Q^2\right)$$

Flavours

Partonic cross-section

Higher twist corrections





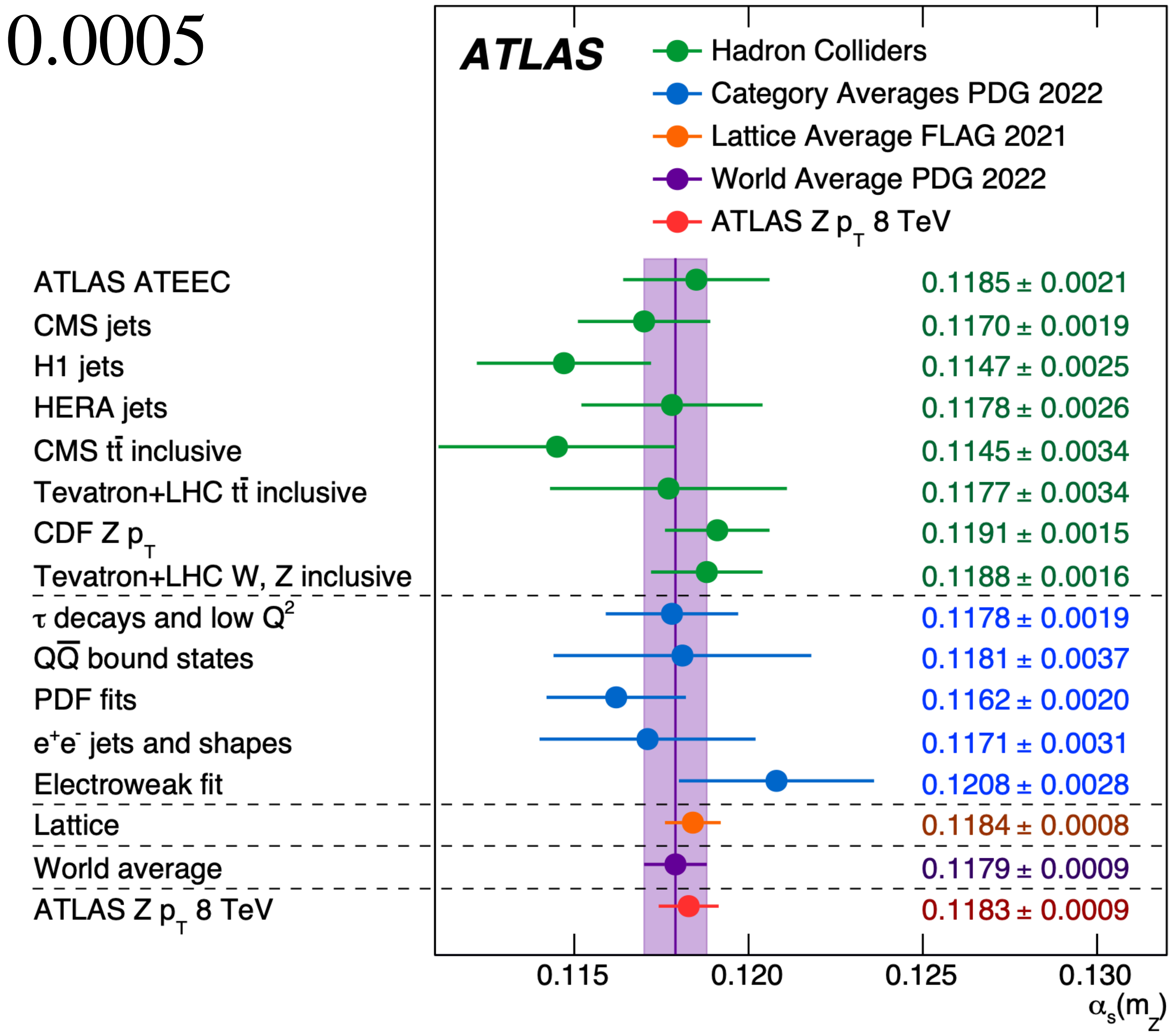
PDFs for precision measurements

Moreover, PDFs often make up the dominant source of uncertainty

ATLAS α_s determination

$\Delta_{\text{tot}} = 0.0009$

$\Delta_{\text{PDF}} = 0.0005$



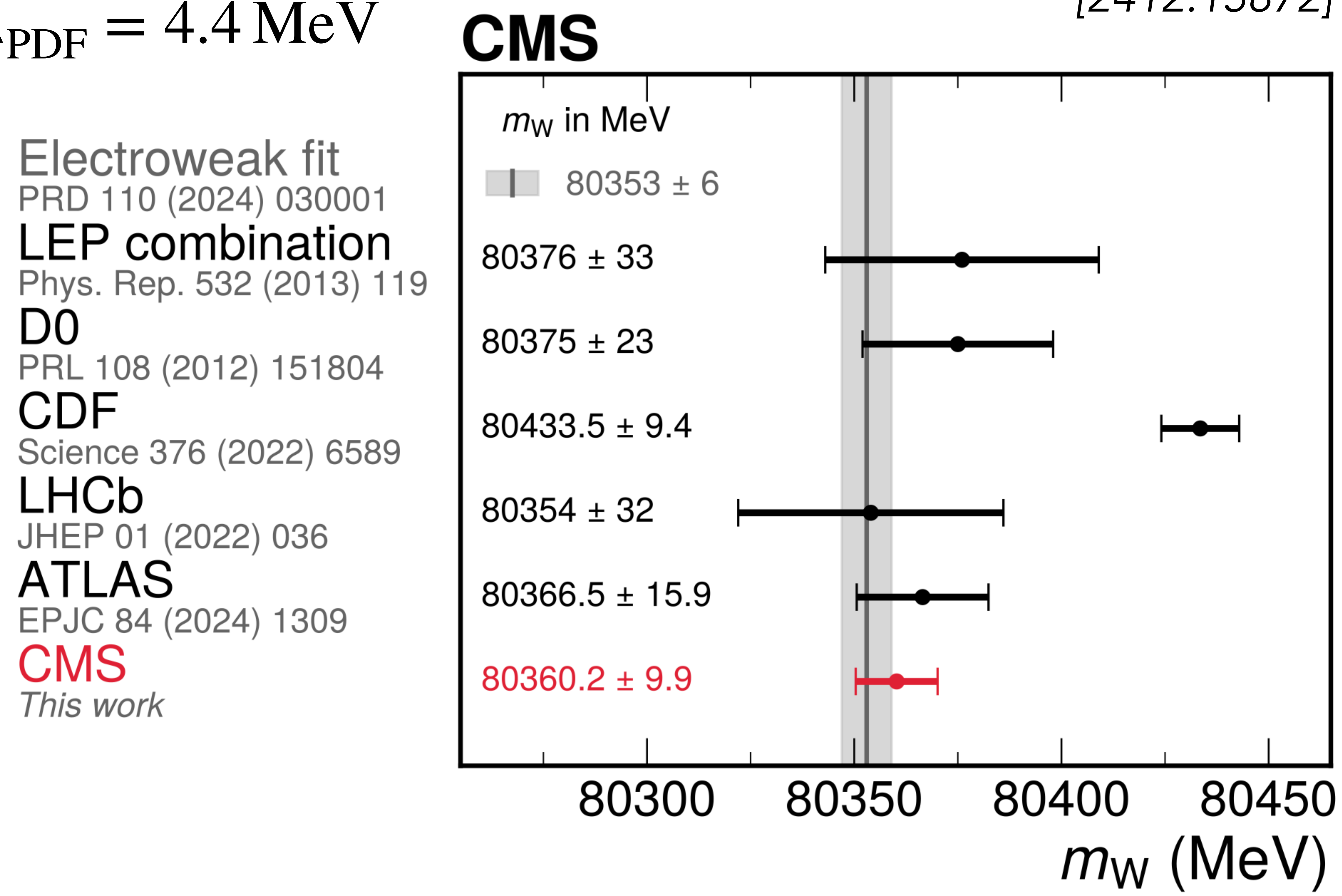
[2309.12986]

CMS M_W determination

$\Delta_{\text{tot}} = 9.9\,\text{MeV}$

$\Delta_{\text{PDF}} = 4.4\,\text{MeV}$

[2412.13872]



The (HL)-LHC precision era clearly calls for more precise and accurate PDFs. What is the bottleneck?

Observables for PDF fits

So far, major PDF fitting collaborations consider **binned, parton-level, low-dimensional observables**. Why?

1. Most PDF-sensitive measurments are released as **binned multi-Gaussian distributions** on HepData

2. Parton level observables admit **high theoretical precision** in QCD/EWK with fast interpolation grids, e.g. at NNLO or N3LO QCD

3. Their likelihood is particularly **simple**

$$\chi^2 = \frac{1}{n_{\rm dat}} \sum_{i,j=1}^{n_{\rm dat}} \left(D_i - T_i\right)(\mathrm{cov})^{-1}_{ij}(D_j - T_j)$$

Observables for PDF fits

So far, major PDF fitting collaborations consider **binned, parton-level, low-dimensional observables**. Why?

1. Most PDF-sensitive measurements are released as **binned multi-Gaussian** d

Binning, unfolding and multiplicity reduction lead to information loss that increase the PDF uncertainty.

Can we do better?

2. P ... t
   in

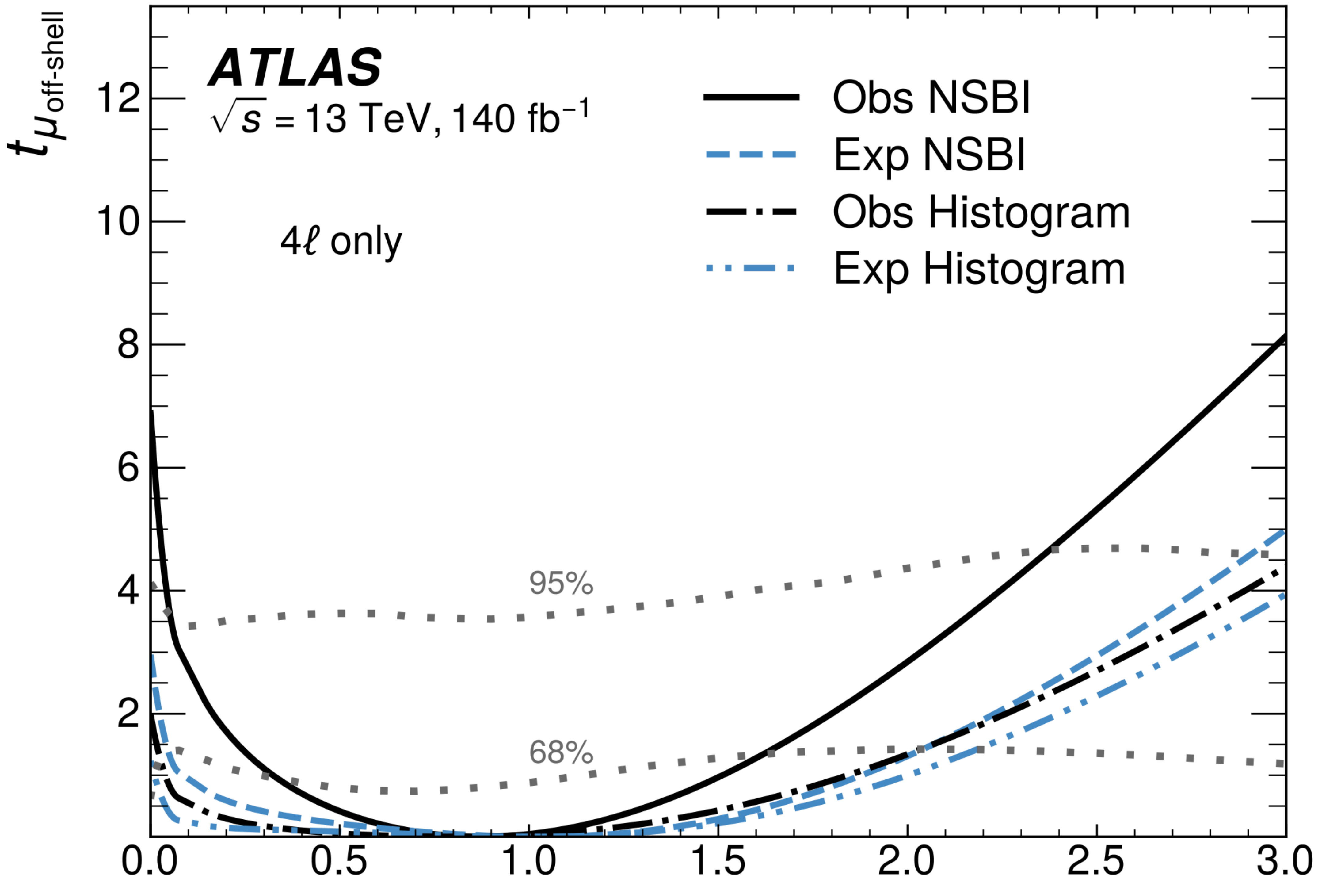
3. Their likelihood is particularly **simple**

$$\chi^2 = \frac{1}{n_{\rm dat}} \sum_{i,j=1}^{n_{\rm dat}} \left(D_i - T_i\right)(\mathrm{cov})^{-1}_{ij}(D_j - T_j)$$

Neural Simulation Based Inference at the LHC

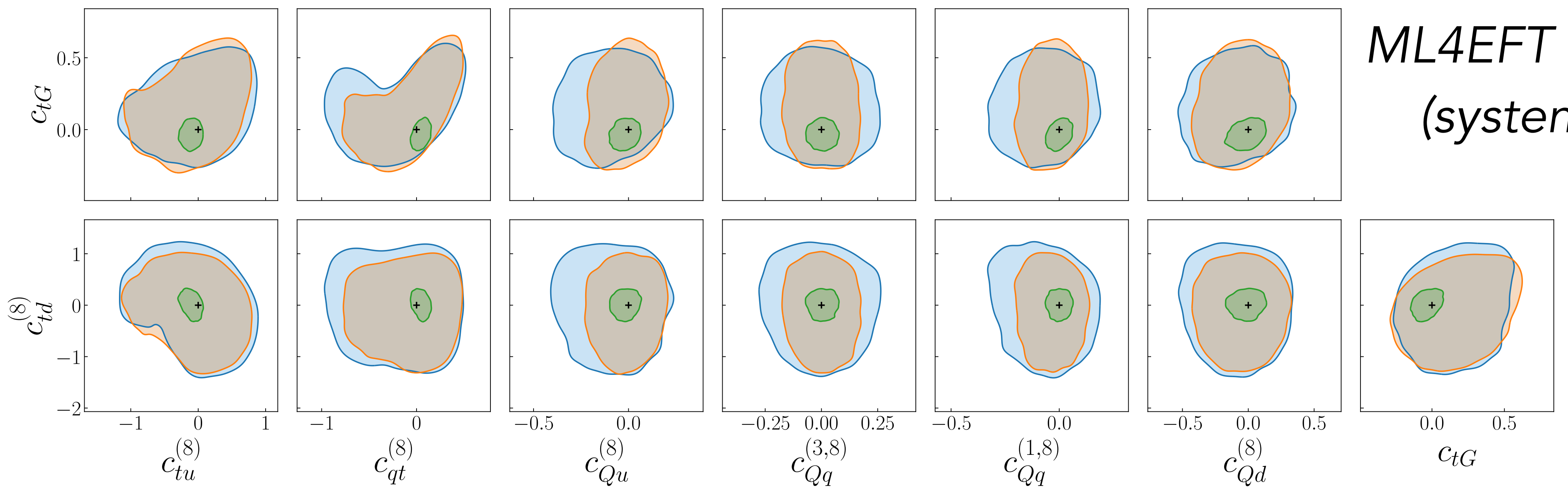
Higgs measurements: 30% increase in precision on the Higgs width using NSBI

ATLAS collaboration: [2412.01548]

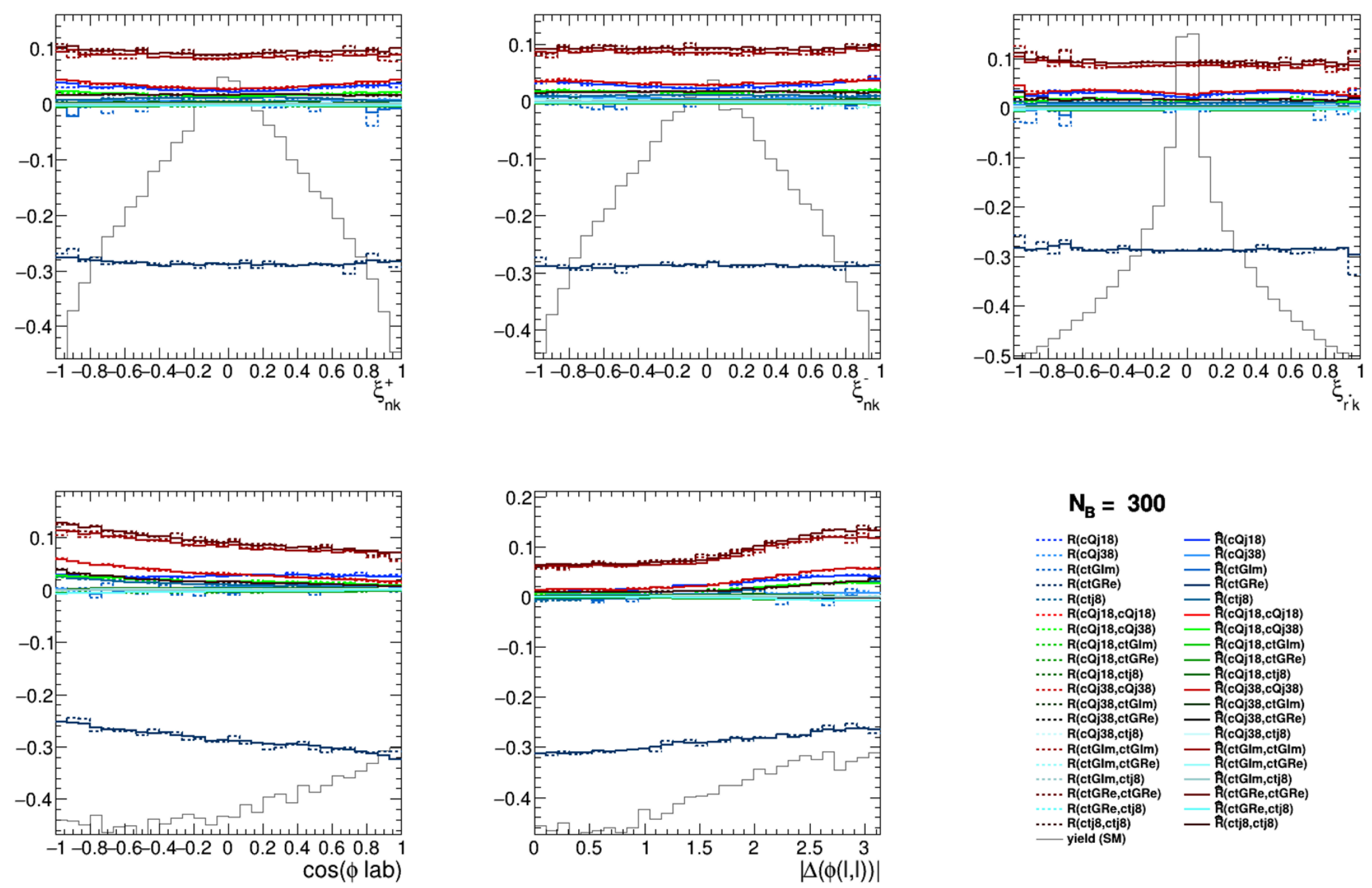


See Jay Sandesara's talk from Tuesday

SMEFT: bounds on Wilson coefficients improve up to a factor of ~5



ML4EFT [2211.02058]
(systematics free)



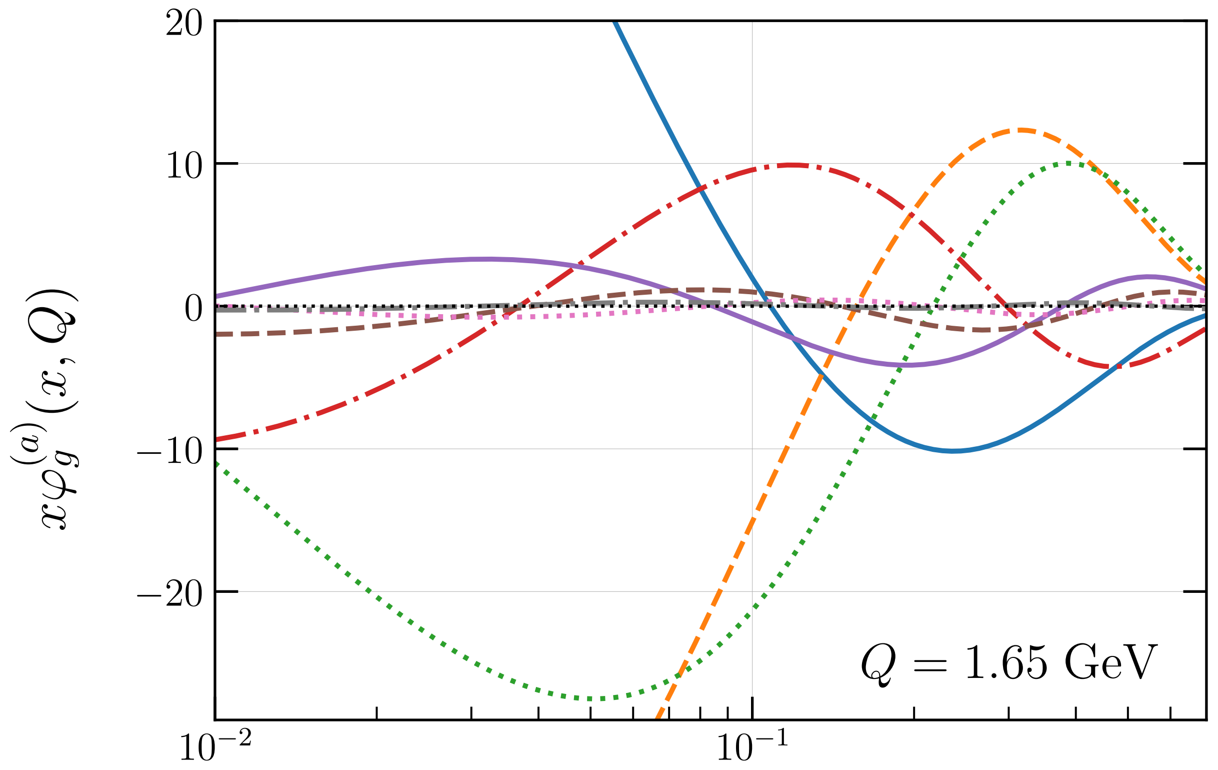
R. Schöfbeck [2406.19076]
(with systematics)

From SMEFT to PDF fits with NSBI

PDFs can be represented in terms of a linear expansion of basis functions

JHEP04 (2026) 068
M. Costantini et al.

$$f_{\mathbf{w}}(x) = \mathbf{w}^T \boldsymbol{\varphi}(x) = \varphi_0(x) + \sum_{k=1}^{N} w_k \varphi_k(x)$$



[2604.13157]

This resembles the SMEFT where SM predictions receive linear corrections

$$g(x, Q_0) = g_{\text{ref}}(x, Q_0) + \sum_j c_j \left(g_j(x, Q_0) - g_{\text{ref}}(x, Q_0) \right)$$

"EFT gluon"

"SM gluon"

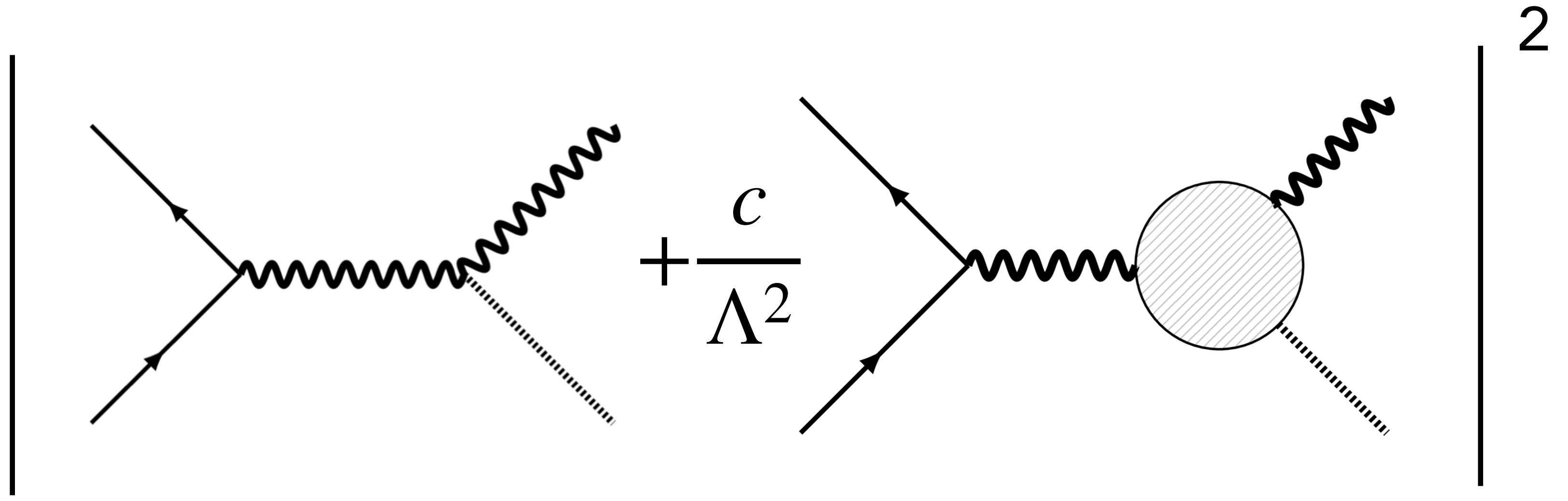
"Wilson coefficients" / POIs

Known basis functions

From SMEFT to PDF fits with NSBI

In the SMEFT, the differential cross-section ratio to the SM is characterised by a **quadratic** polynomial

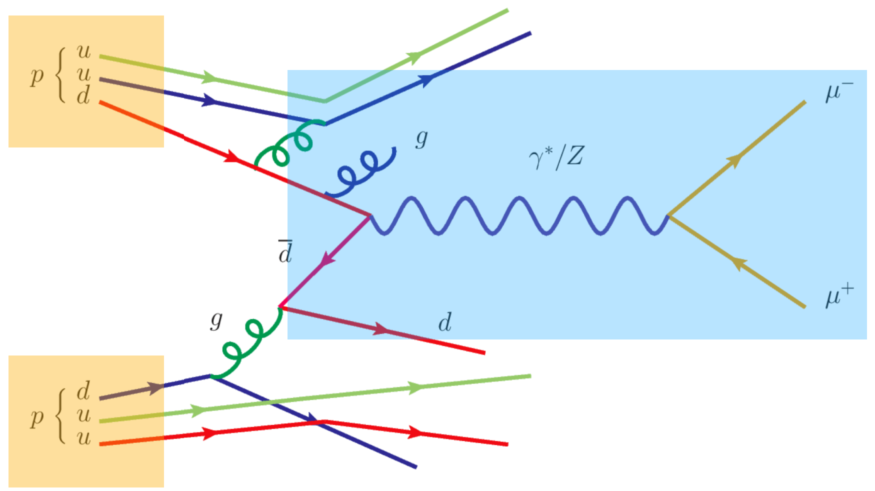
$$r_\sigma(\boldsymbol{x}, \boldsymbol{c}) \equiv \frac{f_\sigma(\boldsymbol{x}, \boldsymbol{c})}{f_\sigma(\boldsymbol{x}, \boldsymbol{0})} = 1 + \sum_{j=1}^{n_{\rm eft}} r_\sigma^{(j)}(\boldsymbol{x}) c_j + \sum_{j=1}^{n_{\rm eft}} \sum_{k \geq j}^{n_{\rm eft}} r_\sigma^{(j,k)}(\boldsymbol{x}) c_j c_k$$



PDFs enter detector-level hadronic cross-sections **quadratically**

$$\frac{d\sigma(\boldsymbol{x}|\boldsymbol{c})}{d\boldsymbol{x}} = \sum_{m,n} \int dx_1\, dx_2 \int d\Phi\, f_m(x_1; \mu_F, \boldsymbol{c})\, f_n(x_2; \mu_F, \boldsymbol{c})\, \frac{d\hat{\sigma}^{mn}(\Phi; \mu_F, \mu_R)}{d\Phi}\, p_{X|\Phi}(\boldsymbol{x}|\Phi)$$

Detector level features



A linear PDF model thus gives a DCR **quadratic** in the POIs (next slides)

$$\hat{R}(\boldsymbol{x}, \boldsymbol{c}) = 1 + \sum_a c_a\, \hat{R}_a(\boldsymbol{x}) + \sum_{a, b \leq a} c_a c_b\, \hat{R}_{ab}(\boldsymbol{x})$$

$$f_g(x, Q, \boldsymbol{c}) \;=\; \varphi_g^{(0)}(x, Q) + \sum_{a=1}^{N} c_a \varphi_g^{(a)}(x, Q)$$

ML surrogates

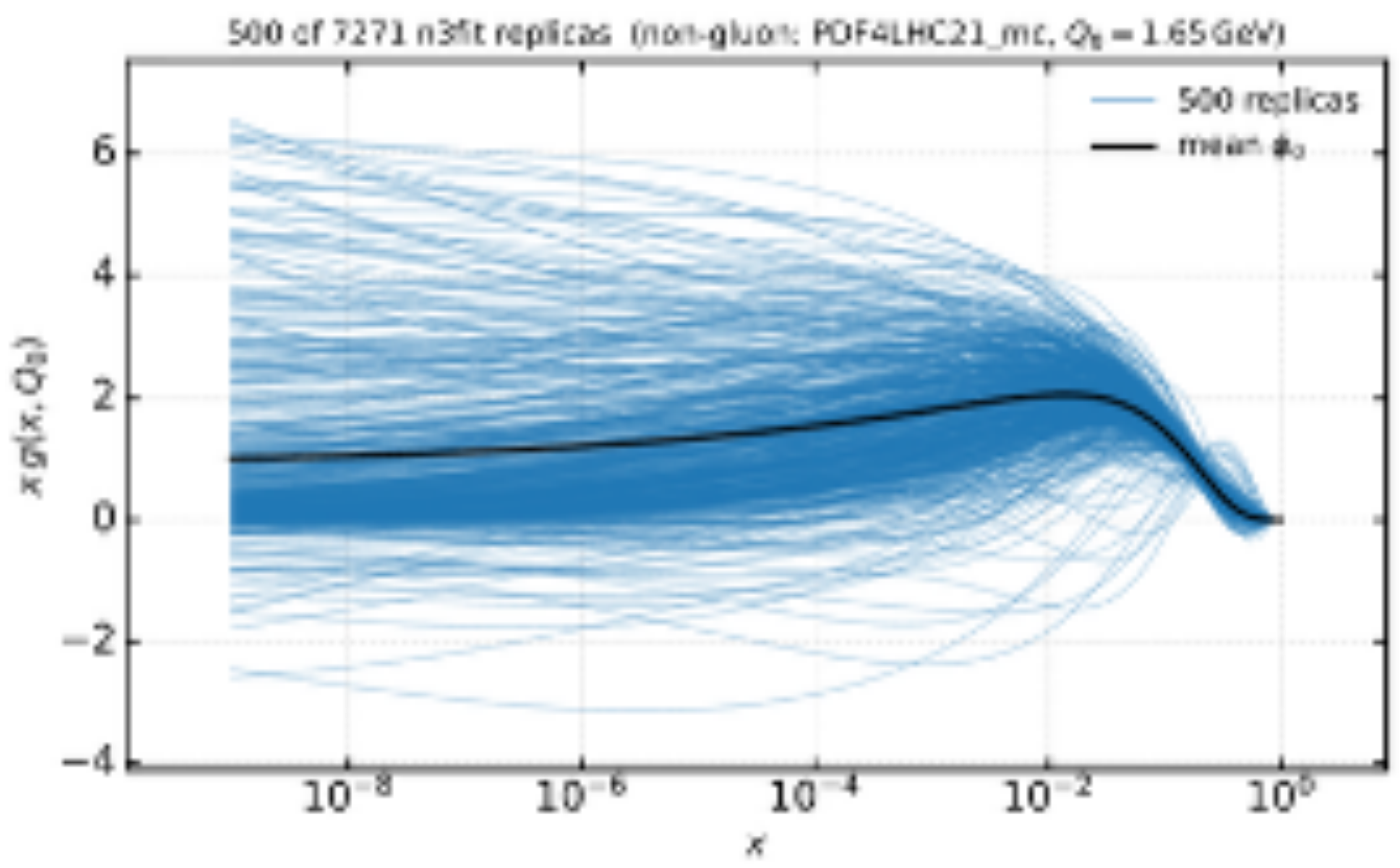
ML surrogates

A linear model for the gluon PDF

1. Construct space of $M = 2 \times 10^4$ **randomly initialised gluon** PDFs with NNPDF4.0 parameterisation

$$x\widetilde{f}_g^{(m)}(x, Q_0) \quad = \quad A_m x^{1-\alpha_m}(1-x)^{\beta_m}\mathrm{NN}_m(x;\boldsymbol{\theta})$$

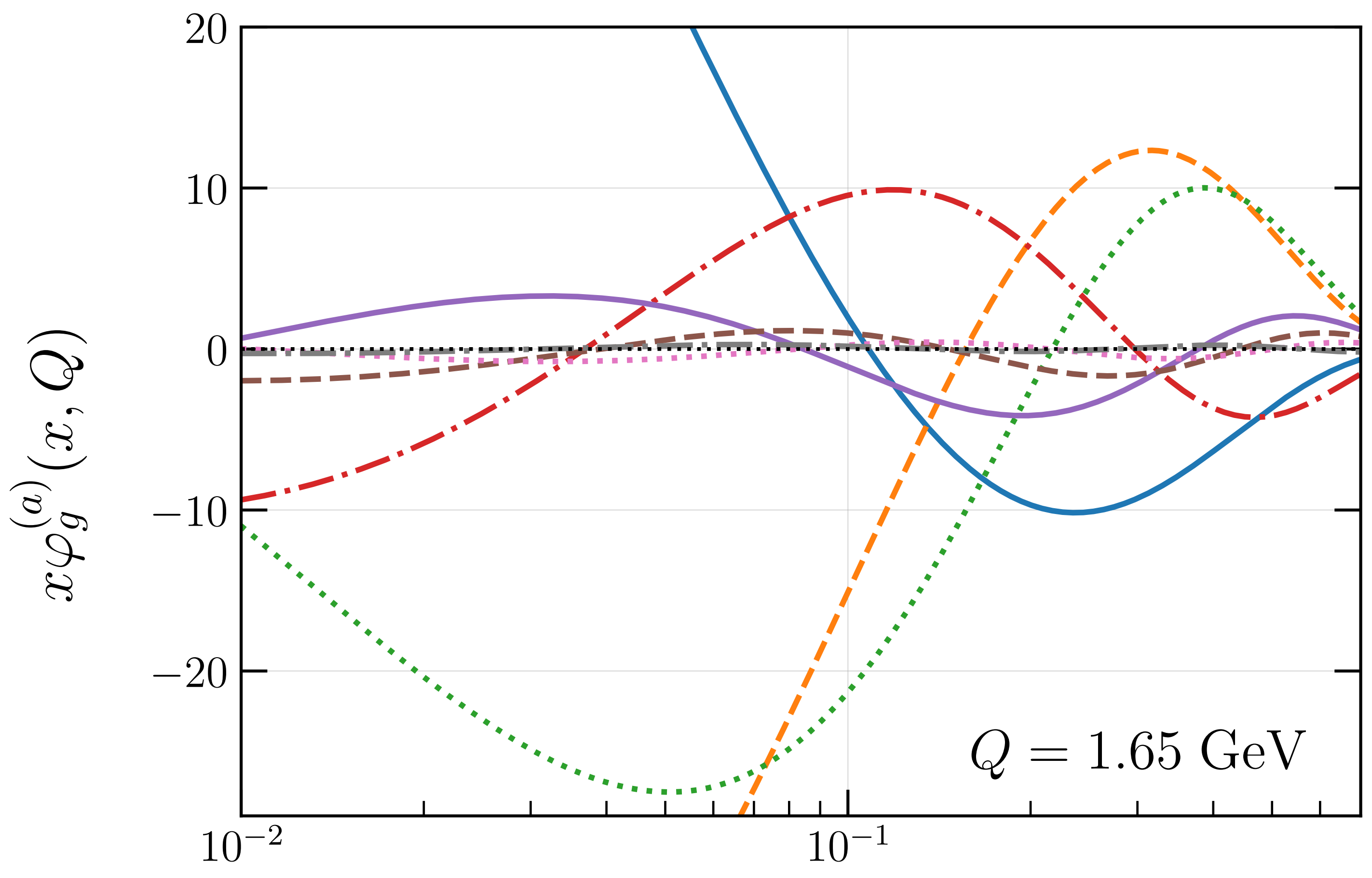
keep quark PDF fixed at PDF4LHC21



2. **Identify leading modes** through Proper Orthogonal Decomposition

$$A = \sum_{m=1}^{M} \widehat{\boldsymbol{f}}_g^{(m)} \widehat{\boldsymbol{f}}_g^{(m)T}$$

Gluon basis functions up to N=7



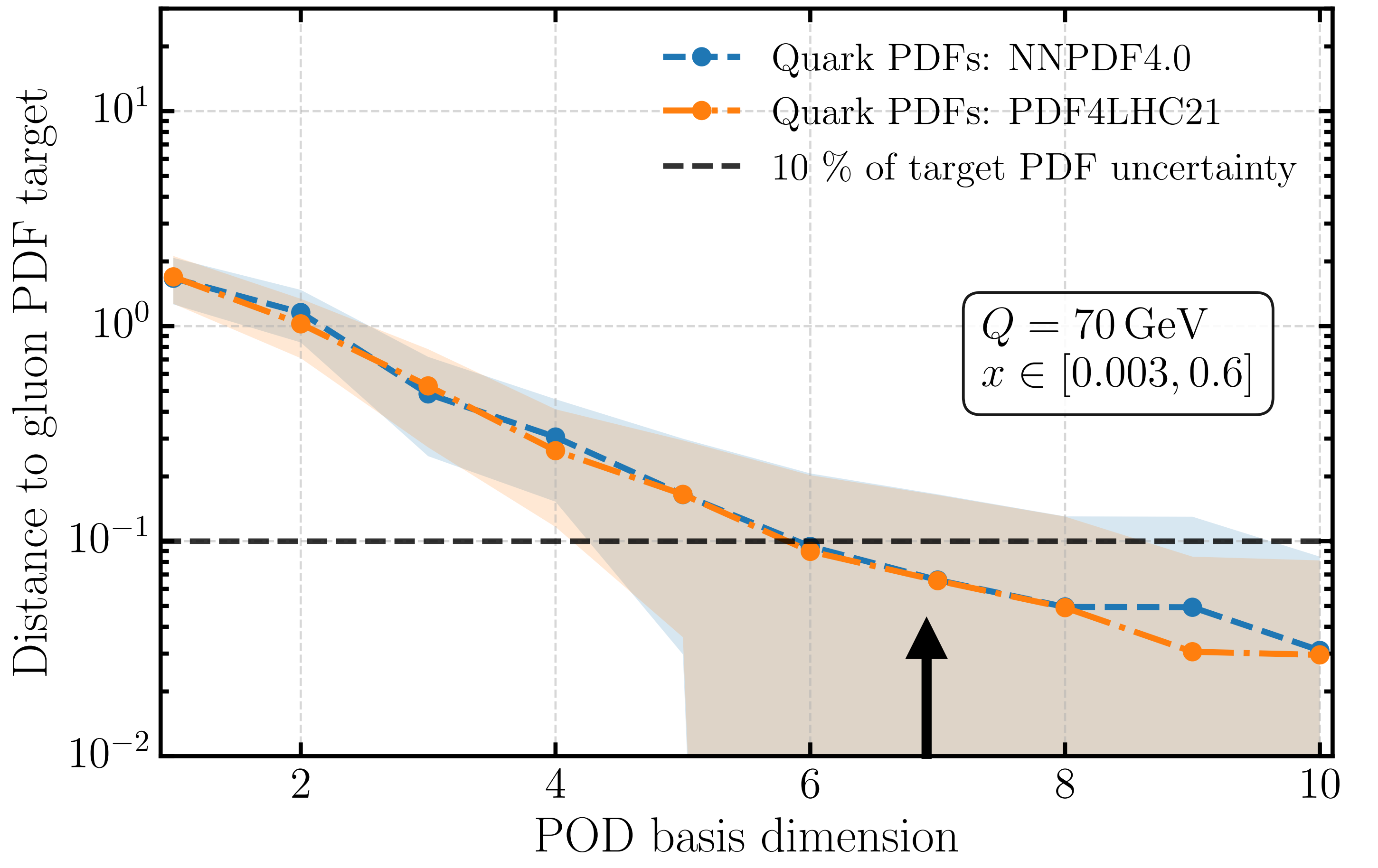
3. Linear gluon PDF

$$f_g(x, Q, \boldsymbol{c}) \quad = \quad \varphi_g^{(0)}(x, Q) + \sum_{a=1}^{N} c_a \varphi_g^{(a)}(x, Q)$$

Mean gluon

POIs

Respect theory constraints + DGLAP automatic by linearity



Learning the PDF likelihood with NSBI

Extended likelihood for an unbinned data set of $N_{\rm obs}$ events

$$L(\mathcal{D}|\boldsymbol{c},\boldsymbol{\nu}) = \mathrm{Pois}(N_{\rm obs}\,|\,\mathcal{L}(\boldsymbol{\nu})\,\sigma(\boldsymbol{c},\boldsymbol{\nu}))\,\mathcal{N}(\boldsymbol{\nu}|\boldsymbol{0},\mathbb{1})\prod_{j=1}^{N_{\rm obs}} p(\boldsymbol{x}_j|\boldsymbol{c},\boldsymbol{\nu})$$

POIs

Nuisances

Auxiliary constraints

Detector-level features

$$p(\boldsymbol{x}|\boldsymbol{c},\boldsymbol{\nu}) = \frac{1}{\sigma(\boldsymbol{c},\boldsymbol{\nu})}\frac{d\sigma(\boldsymbol{x}|\boldsymbol{c},\boldsymbol{\nu})}{d\boldsymbol{x}}$$

Learning the PDF likelihood with NSBI

Extended likelihood for an unbinned data set of $N_{\rm obs}$ events

$$L(\mathcal{D}|\boldsymbol{c},\boldsymbol{\nu}) = \mathrm{Pois}(N_{\rm obs}\,|\,\mathcal{L}(\boldsymbol{\nu})\,\sigma(\boldsymbol{c},\boldsymbol{\nu}))\,\mathcal{N}(\boldsymbol{\nu}|\boldsymbol{0},\mathbb{1})\prod_{j=1}^{N_{\rm obs}} p(\boldsymbol{x}_j|\boldsymbol{c},\boldsymbol{\nu})$$

POIs

Nuisances

Auxiliary constraints

Detector-level features

$$p(\boldsymbol{x}|\boldsymbol{c},\boldsymbol{\nu}) = \frac{1}{\sigma(\boldsymbol{c},\boldsymbol{\nu})}\frac{\mathrm{d}\sigma(\boldsymbol{x}|\boldsymbol{c},\boldsymbol{\nu})}{\mathrm{d}\boldsymbol{x}}$$

Likelihood ratio to the reference hypothesis $(c,\nu=0)$ only needs the **differential cross-section ratio**

$$\log\frac{L(\mathcal{D}|\boldsymbol{c},\boldsymbol{\nu})}{L(\mathcal{D}|\boldsymbol{0},\boldsymbol{0})} = -\Big[\mathcal{L}(\boldsymbol{\nu})\sigma(\boldsymbol{c},\boldsymbol{\nu}) - \mathcal{L}(\boldsymbol{0})\sigma(\boldsymbol{0},\boldsymbol{0})\Big] + \boxed{\sum_{j=1}^{N_{\rm obs}}\log\left[\frac{\mathcal{L}(\boldsymbol{\nu})}{\mathcal{L}(\boldsymbol{0})}\frac{\mathrm{d}\sigma(\boldsymbol{x}_j|\boldsymbol{c},\boldsymbol{\nu})}{\mathrm{d}\sigma(\boldsymbol{x}_j|\boldsymbol{0},\boldsymbol{0})}\right]} - \frac{1}{2}|\boldsymbol{\nu}|^2$$

Factorise PDF dependence c from the nuisances ν

$$\boxed{\frac{\mathrm{d}\sigma(\boldsymbol{x}|\boldsymbol{c},\boldsymbol{\nu})}{\mathrm{d}\sigma(\boldsymbol{x}|\boldsymbol{0},\boldsymbol{0})} \approx S(\boldsymbol{x},\boldsymbol{\nu})\,R(\boldsymbol{x},\boldsymbol{c})}$$

$$\hat{R}(\boldsymbol{x},\boldsymbol{c}) = 1 + \sum_a c_a\,\hat{R}_a(\boldsymbol{x}) + \sum_{a,\,b\leq a} c_a c_b\,\hat{R}_{ab}(\boldsymbol{x})$$

$$\hat{S}(\boldsymbol{x},\boldsymbol{\nu}) = \exp\Big(\nu_A\,\hat{\Delta}_A(\boldsymbol{x})\Big)$$

Learning the PDF likelihood with NSBI

Machine-learn surrogates with a **parametric** ansatz using the **likelihood ratio trick** *[2107.10859], [2205.12976]*

$$\frac{d\sigma(\boldsymbol{x}|\boldsymbol{c},\boldsymbol{\nu})}{d\sigma(\boldsymbol{x}|\mathbf{0},\mathbf{0})} \approx S(\boldsymbol{x},\boldsymbol{\nu})\, R(\boldsymbol{x},\boldsymbol{c})$$

$$\hat{R}(\boldsymbol{x},\boldsymbol{c}) = 1 + \sum_a c_a\, \hat{R}_a(\boldsymbol{x}) + \sum_{a,b\leq a} c_a c_b\, \hat{R}_{ab}(\boldsymbol{x})$$

$$\hat{S}(\boldsymbol{x},\boldsymbol{\nu}) = \exp\left(\nu_A\, \hat{\Delta}_A(\boldsymbol{x})\right)$$

$$L_{\mathrm{MSE}} = \langle \hat{R}(\mathbf{x}_i, c) - \omega(\mathbf{z}_i, c)\rangle^2$$

$$\omega(\boldsymbol{z}_i,\boldsymbol{c}) = \frac{f_{m_i}(x_{1i};Q_i,\boldsymbol{c})\, f_{n_i}(x_{2i};Q_i,\boldsymbol{c})}{f^{(\mathrm{gen})}_{m_i}(x_{1i};Q_i)\, f^{(\mathrm{gen})}_{n_i}(x_{2i};Q_i)} \sim c^2$$

Detector level features

Latent space (parton level quantities)

PDF reweighing factor

Plug surrogates into profiled likelihood-ratio test statistic (MINUIT) evaluated on **Asimov** data

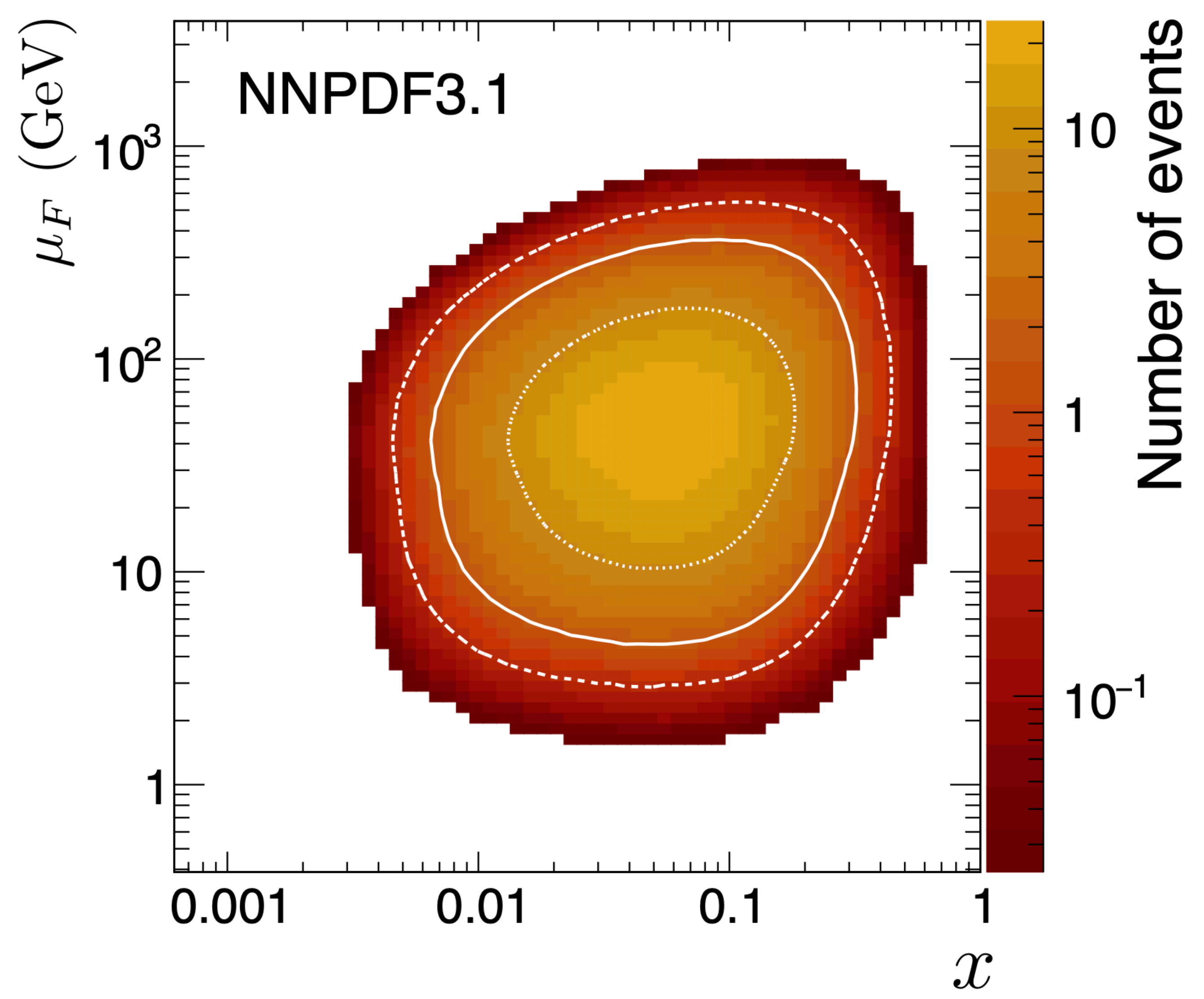
$$q_{\boldsymbol{c}}(\mathcal{D}) = -2\left[\max_{\boldsymbol{\nu}} \log \frac{L(\mathcal{D}|\boldsymbol{c},\boldsymbol{\nu})}{L(\mathcal{D}|\mathbf{0},\mathbf{0})} - \max_{\boldsymbol{c}',\boldsymbol{\nu}'} \log \frac{L(\mathcal{D}|\boldsymbol{c}',\boldsymbol{\nu}')}{L(\mathcal{D}|\mathbf{0},\mathbf{0})}\right]$$

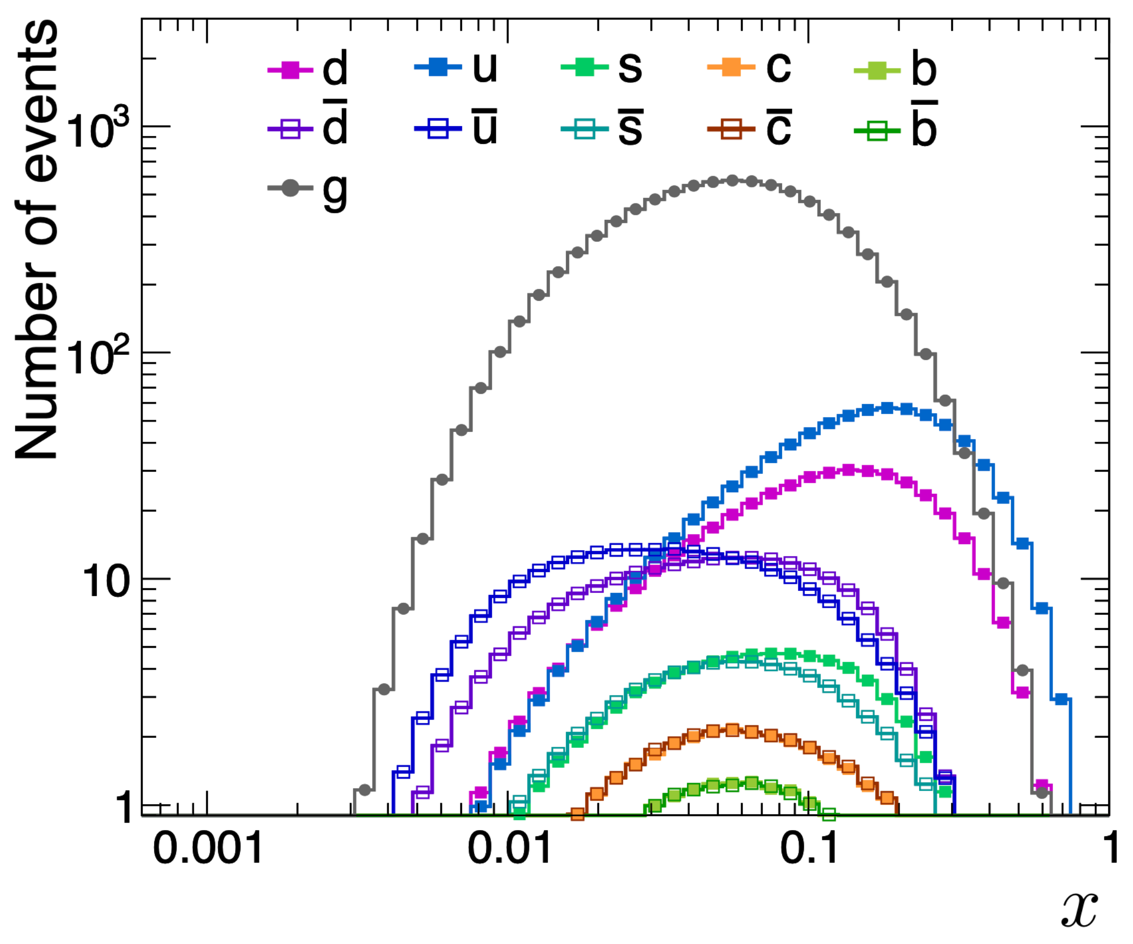
Application: unbinned top-quark pair production at 13 TeV

Unbinned top-quark pair production at 13 TeV at the detector level

$$pp \to t\bar{t} \to b\ell^+\nu_\ell\bar{b}\ell^-\bar{\nu}_\ell$$

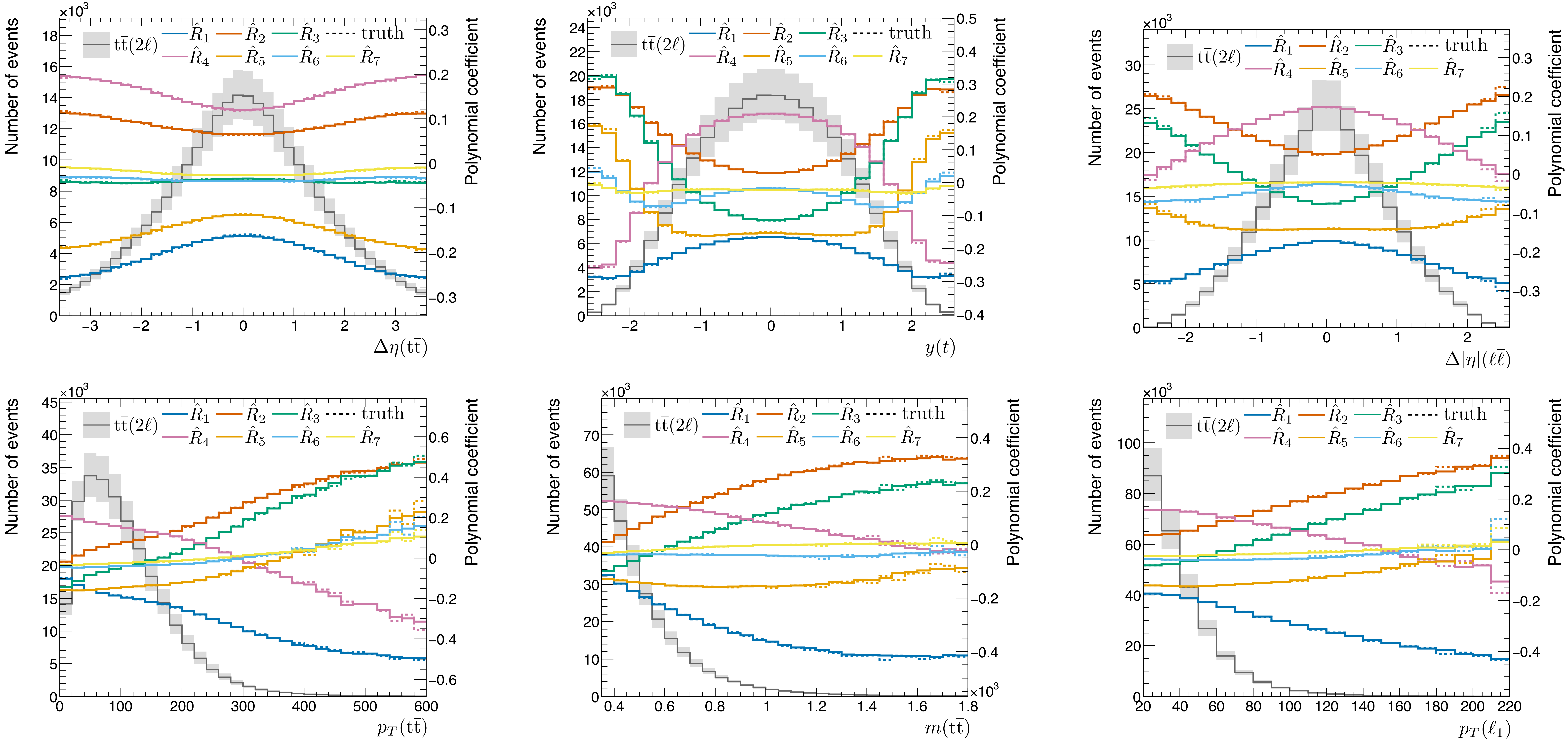
- **Simulation**: POWHEG + PYTHIA 8
  - Delphes CMS detector model, 10M events
  - NNPDF3.1 nominal, $\mu = H_T/4$, normalised to NNLO + NNLL (Top++)
- **Dilepton channel** with L=137/fb (legacy Run II): 2 isolated leptons ($p_T > 20$ GeV, $|\eta|$<2.5), Z-mass veto, > 2 jets ($p_T > 30$ GeV, $|\eta| < 2.4$), ≥ 2 b-tags
- **Kinematic coverage:** $10^{-2} \leq x \leq 0.5$
- **16 event-level features** ($m_{t\bar{t}}, p_T, \eta, \ldots$) with detector-level top quark reconstruction
- **20 nuisances** (Ren. & fac. Scale, b-tagging, JES, JER, Lepton scale, α_s, lumi)





Machine Learning the POIs

$\mathbb{R}^{16} \to \mathbb{R}^{35}$



+ many more that did not fit on the slide!

Binned reference

Bin the **same events** in 16 bins double-differential in $m_{t\bar{t}}$ and $|y_{t\bar{t}}|$ from CMS

$$m(\mathrm{t\bar{t}}) \in [300,\ 400,\ 500,\ 650,\ 1500]\ \ \mathrm{GeV} \qquad\qquad |y(\mathrm{t\bar{t}})| \in [0,\ 0.4,\ 0.8,\ 1.2,\ 2.5]$$

Binned

$$\log \frac{L_{\mathrm{binned}}(\mathcal{D}|\boldsymbol{c},\boldsymbol{\nu})}{L_{\mathrm{binned}}(\mathcal{D}|\mathbf{0},\mathbf{0})} = \sum_{i=1}^{N_{\mathrm{bins}}} \left[-\big(\lambda_i(\boldsymbol{c},\boldsymbol{\nu}) - \lambda_i(\mathbf{0},\mathbf{0})\big) + N_{i,\mathrm{obs}} \log \frac{\lambda_i(\boldsymbol{c},\boldsymbol{\nu})}{\lambda_i(\mathbf{0},\mathbf{0})} \right]$$

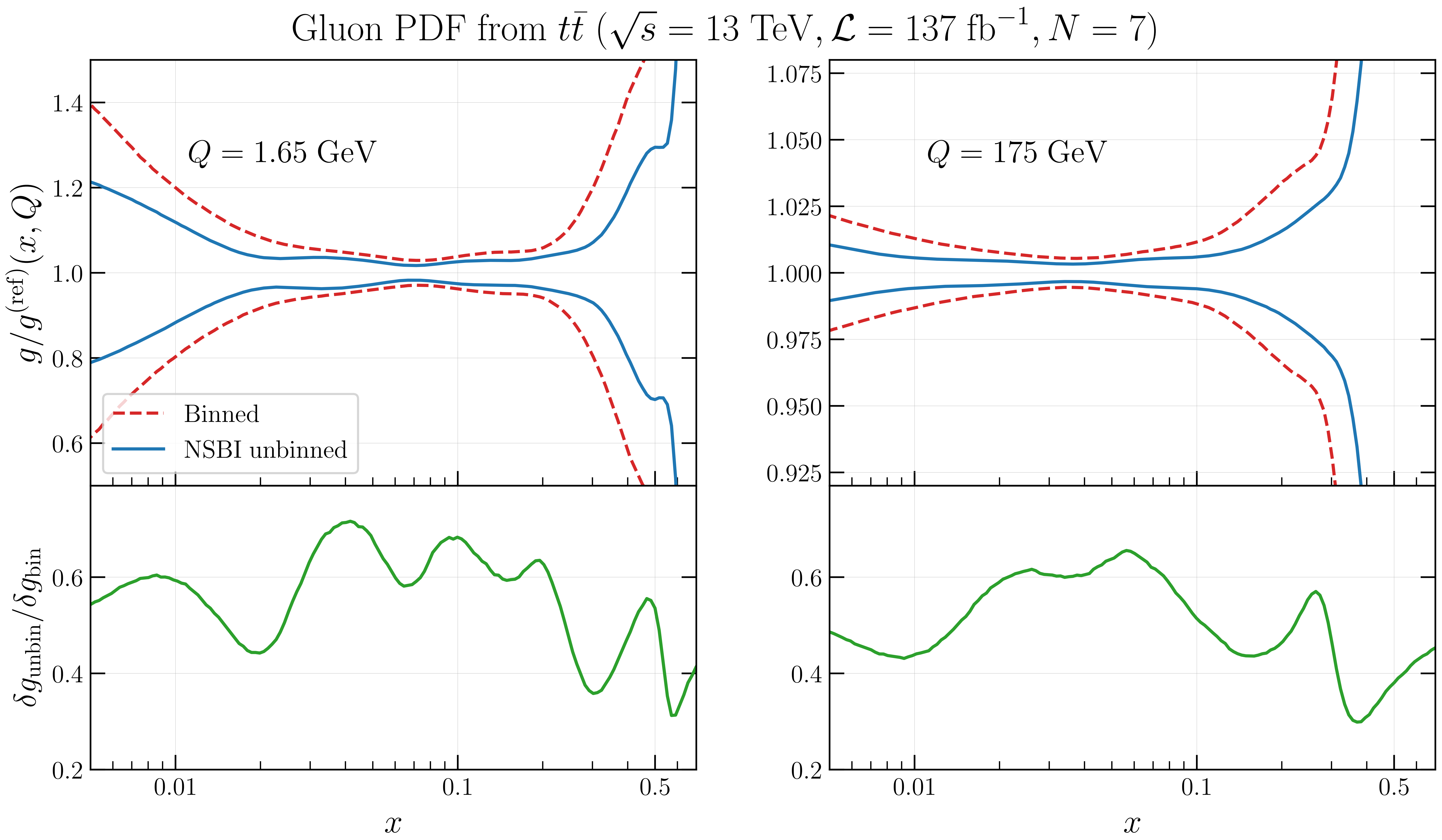
Unbinned

$$\log \frac{L(\mathcal{D}|\boldsymbol{c},\boldsymbol{\nu})}{L(\mathcal{D}|\mathbf{0},\mathbf{0})} = -\mathcal{L}(\mathbf{0}) \int \mathrm{d}\sigma(\boldsymbol{x}|\mathbf{0},\mathbf{0})\, \hat{T}(\boldsymbol{x};\boldsymbol{c},\boldsymbol{\nu}) + \sum_{j=1}^{N_{\mathrm{obs}}} \log\Big(1 + \hat{T}(\boldsymbol{x}_j;\boldsymbol{c},\boldsymbol{\nu})\Big)$$

Identical setup, same POIs and nuisances, so any difference between binned and unbinned is a measure of information loss

Results: the gluon PDF from unbinned top-quark pairs

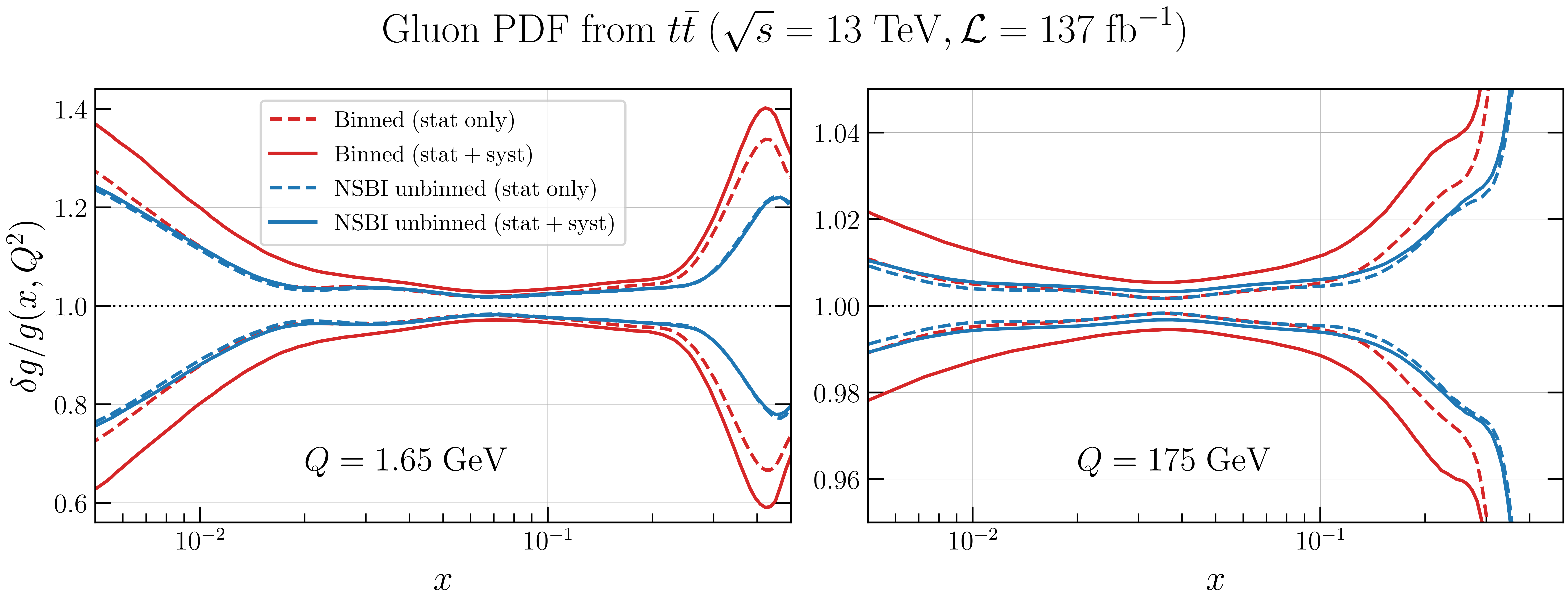
Sensitivity Improvement between 30% and 70% compared to the binned analysis



Low scale **High scale**

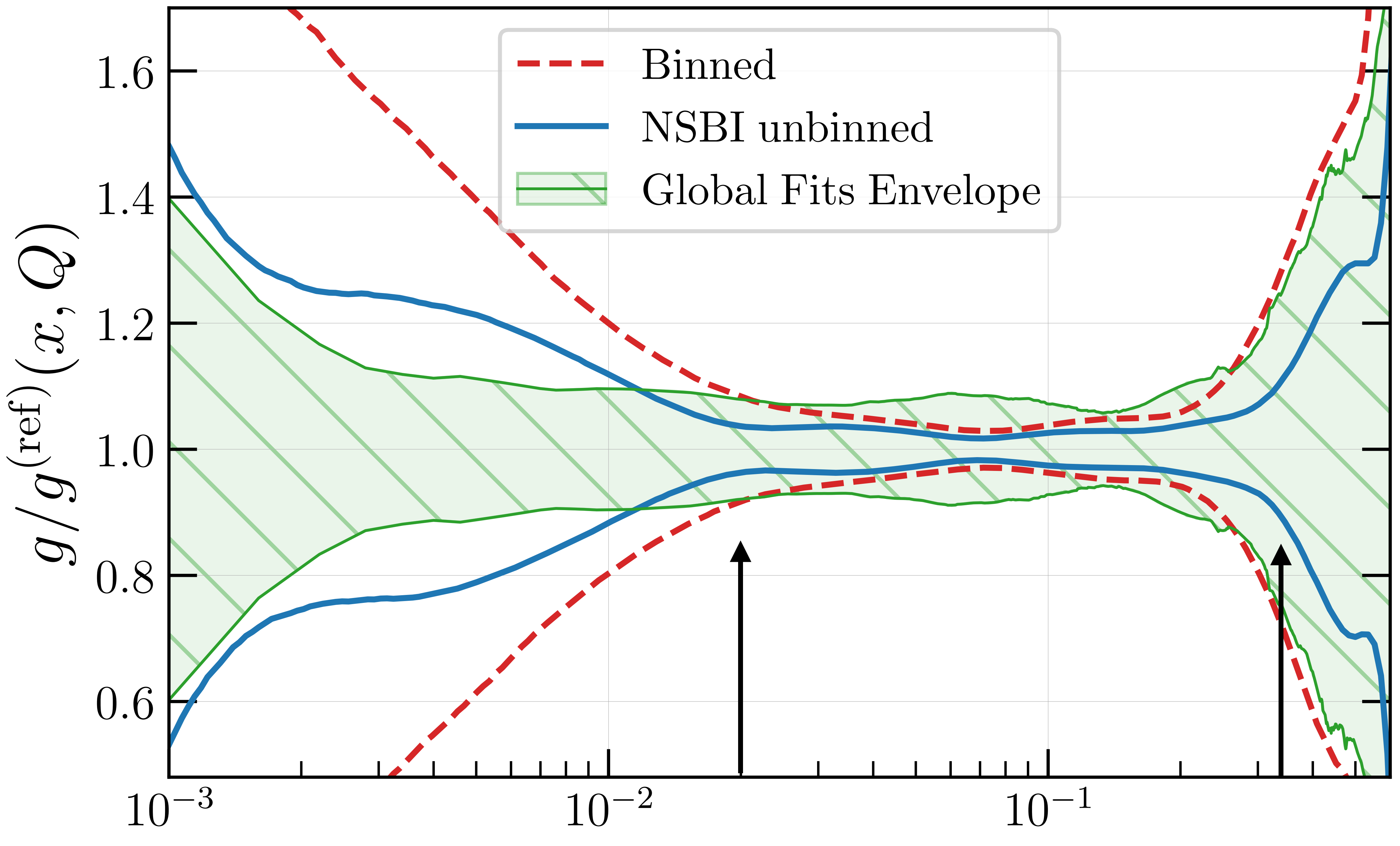
Results: systematics

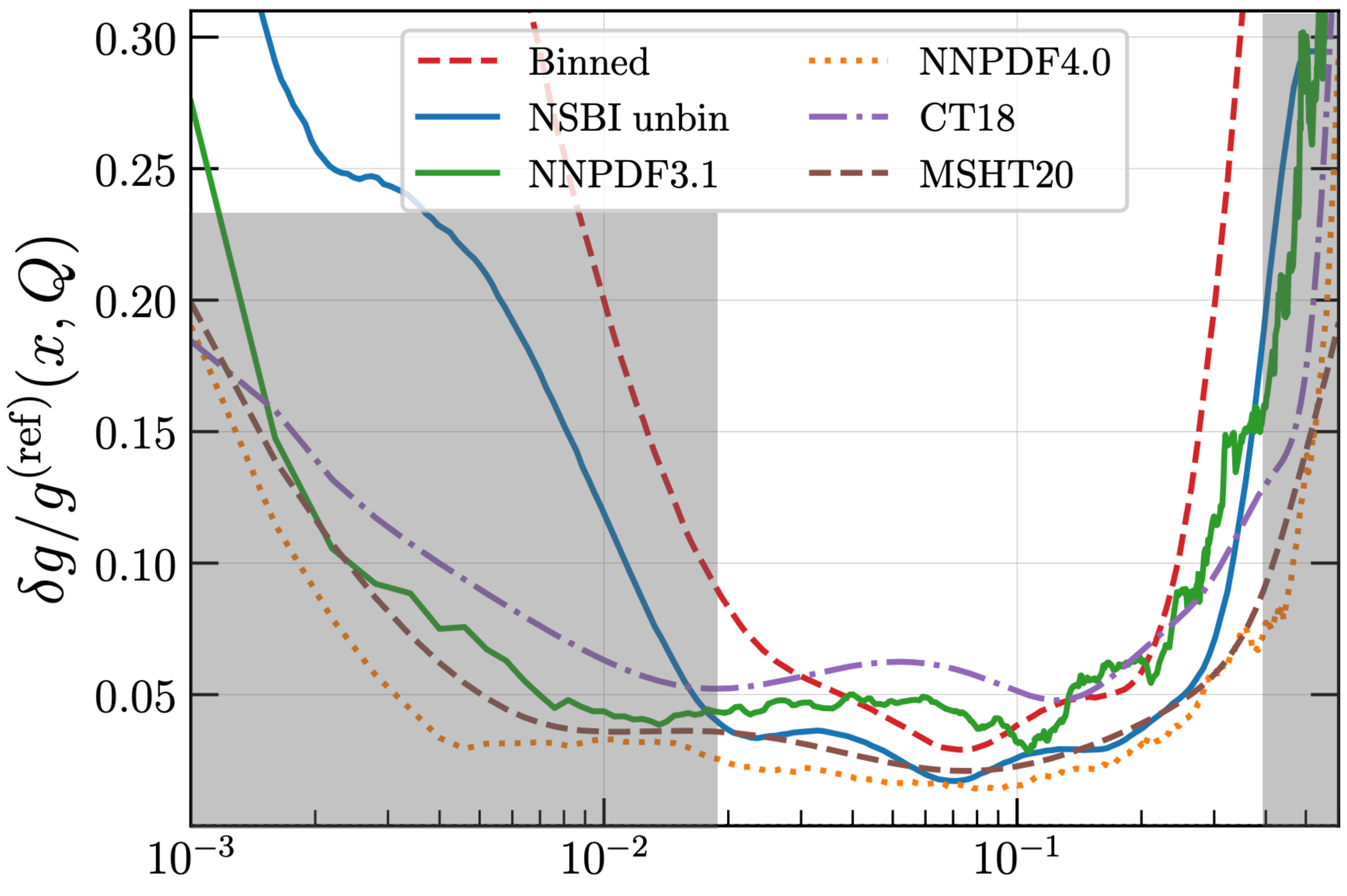
Unbinned measurements are more efficient to constrain systematic uncertainties from the same data than binned analyses



Results: global fits

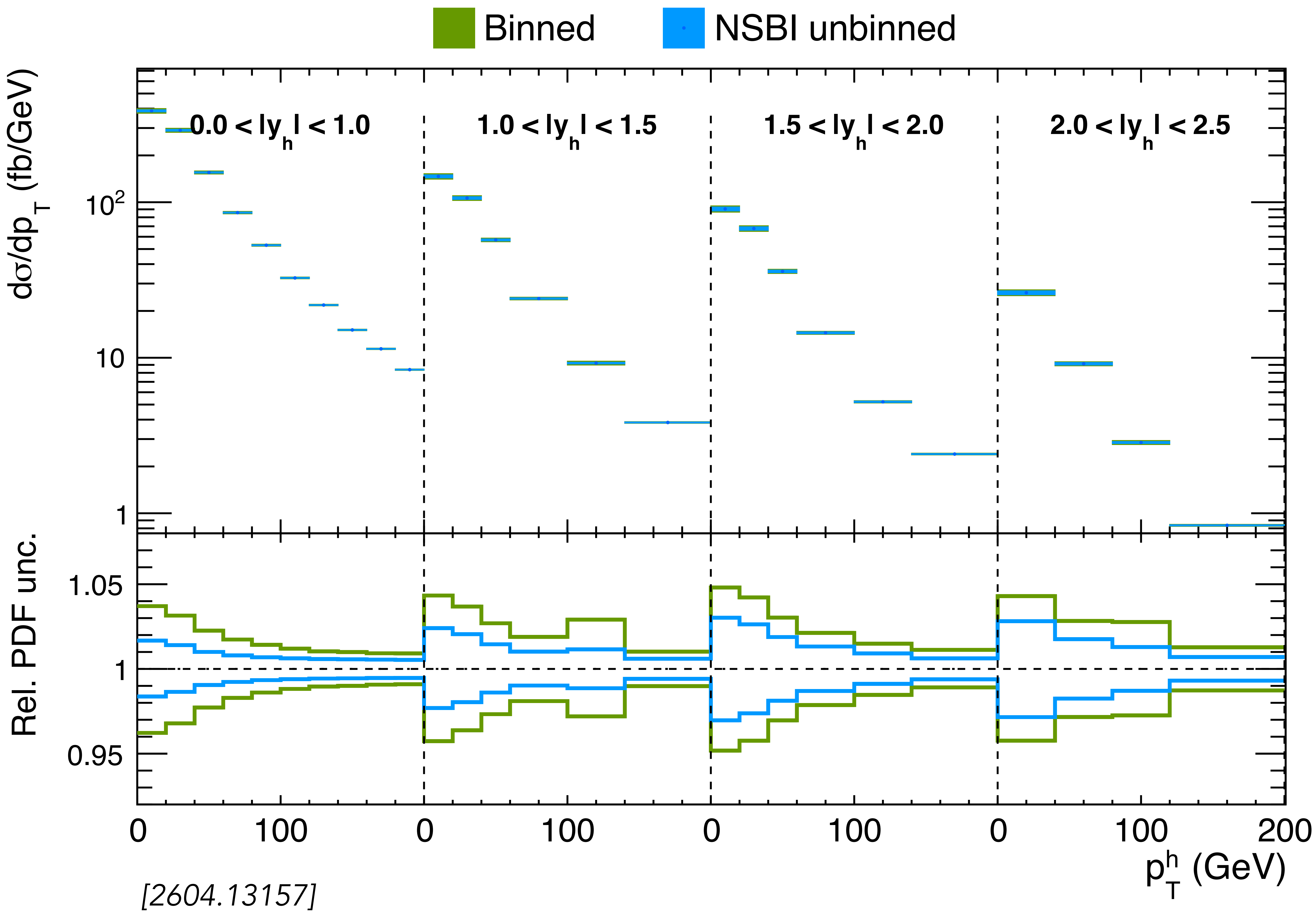
The CMS top-quark measurement, binned or unbinned, adds sensitivity to the gluon PDF beyond existing global PDF sets



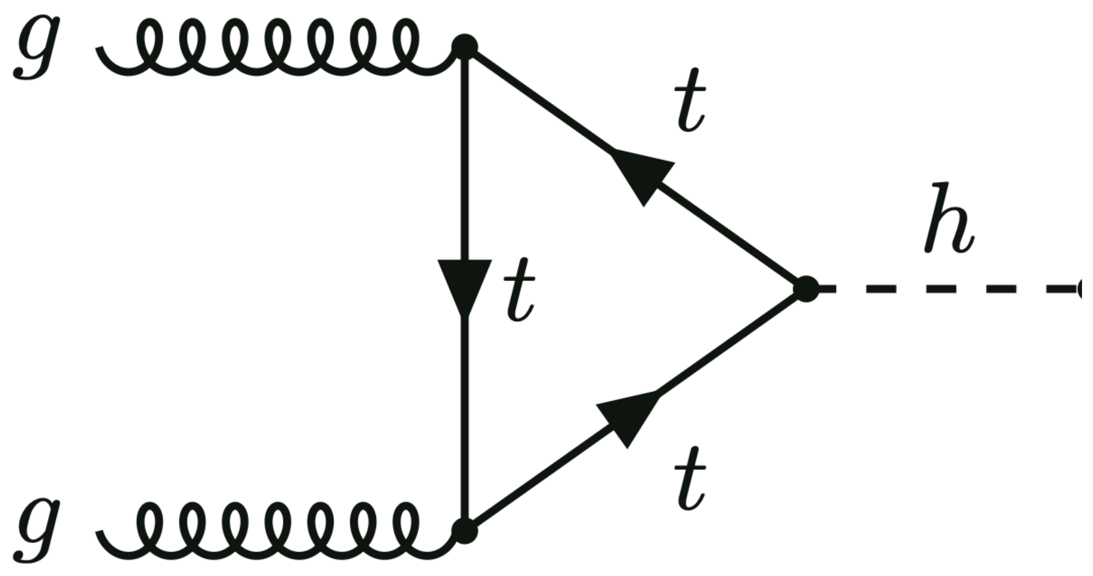


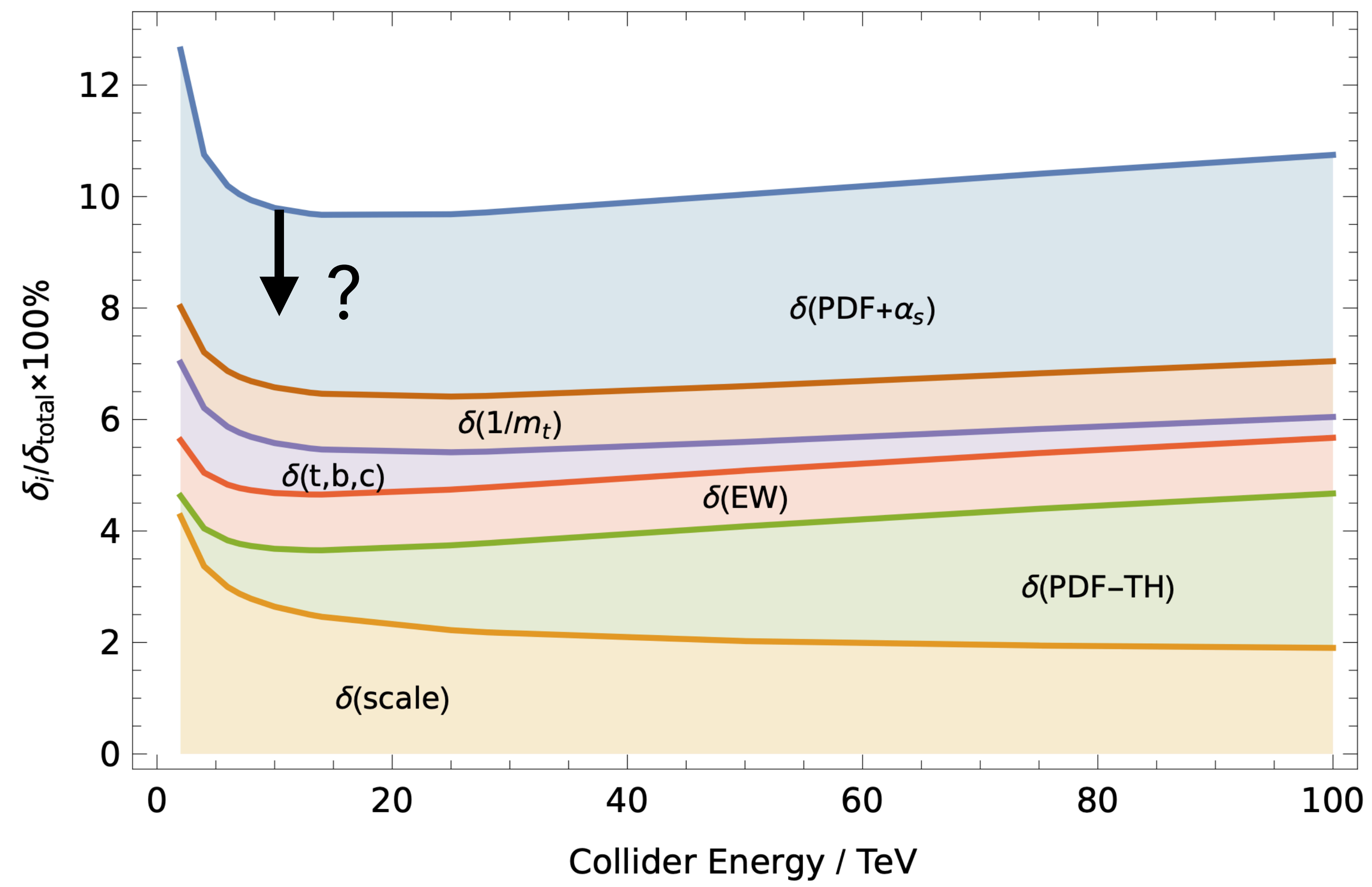
Implications for Higgs production in gluon fusion

Large reduction of PDF uncertainty in the dominant Higgs production mode



[2604.13157]





LHC XSWG [1802.00827]

Summary

- Interdisciplinary project between experts from CMS, ML/AI and Theory
- First NSBI determination of the gluon PDF from high dimensional unbinned data on Asimov data
- The linear PDF model moved the inference problem to a SMEFT like problem
- Full treatment of experimental and theoretical unbinned systematic uncertainties
- Around 50% sensitivity gain on the gluon PDF and its Higgs production uncertainty
- Done: Asimov. Next: real data

Summary

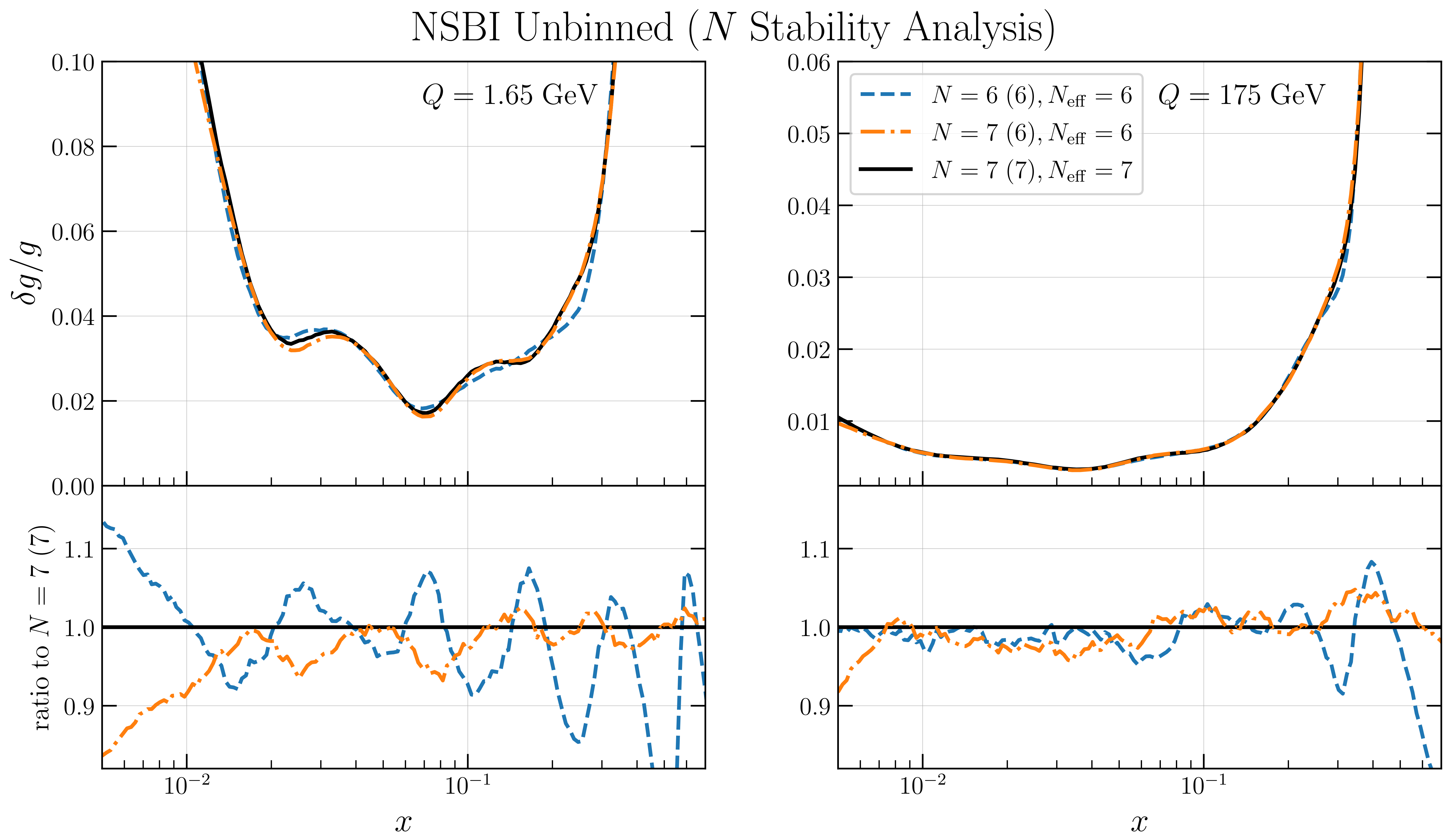
- Interdisciplinary project between experts from CMS, ML/AI and Theory
- First NSBI determination of the gluon PDF from high dimensional unbinned data on Asimov data
- The linear PDF model moved the inference problem to a SMEFT like problem
- Full treatment of experimental and theoretical unbinned systematic uncertainties
- Around 50% sensitivity gain on the gluon PDF and its Higgs production uncertainty
- Done: Asimov. Next: real data

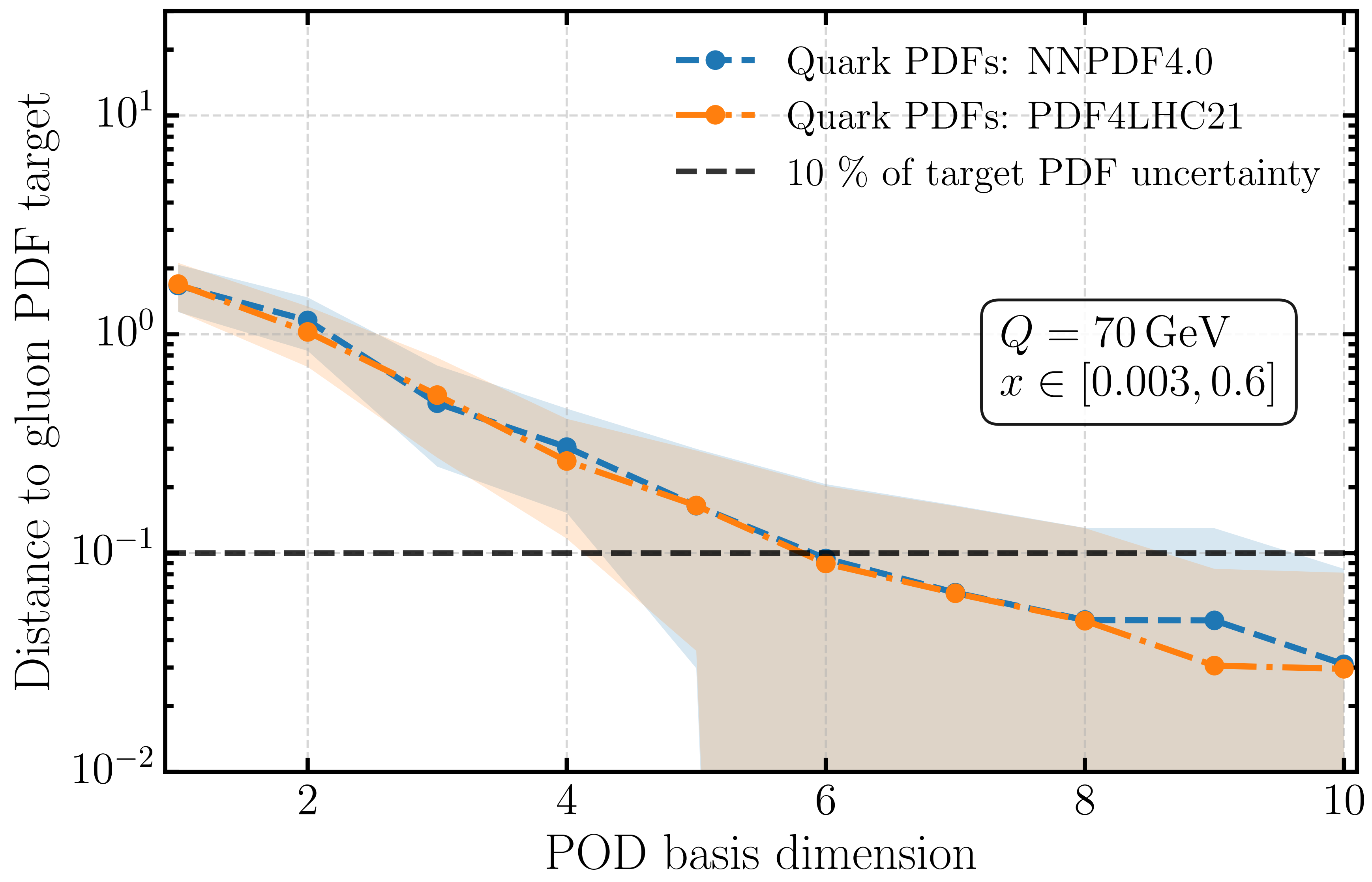
Thank you for your attention!

e-mail: jaco.ter.hoeve@ed.ac.uk

Backup

Stability analysis



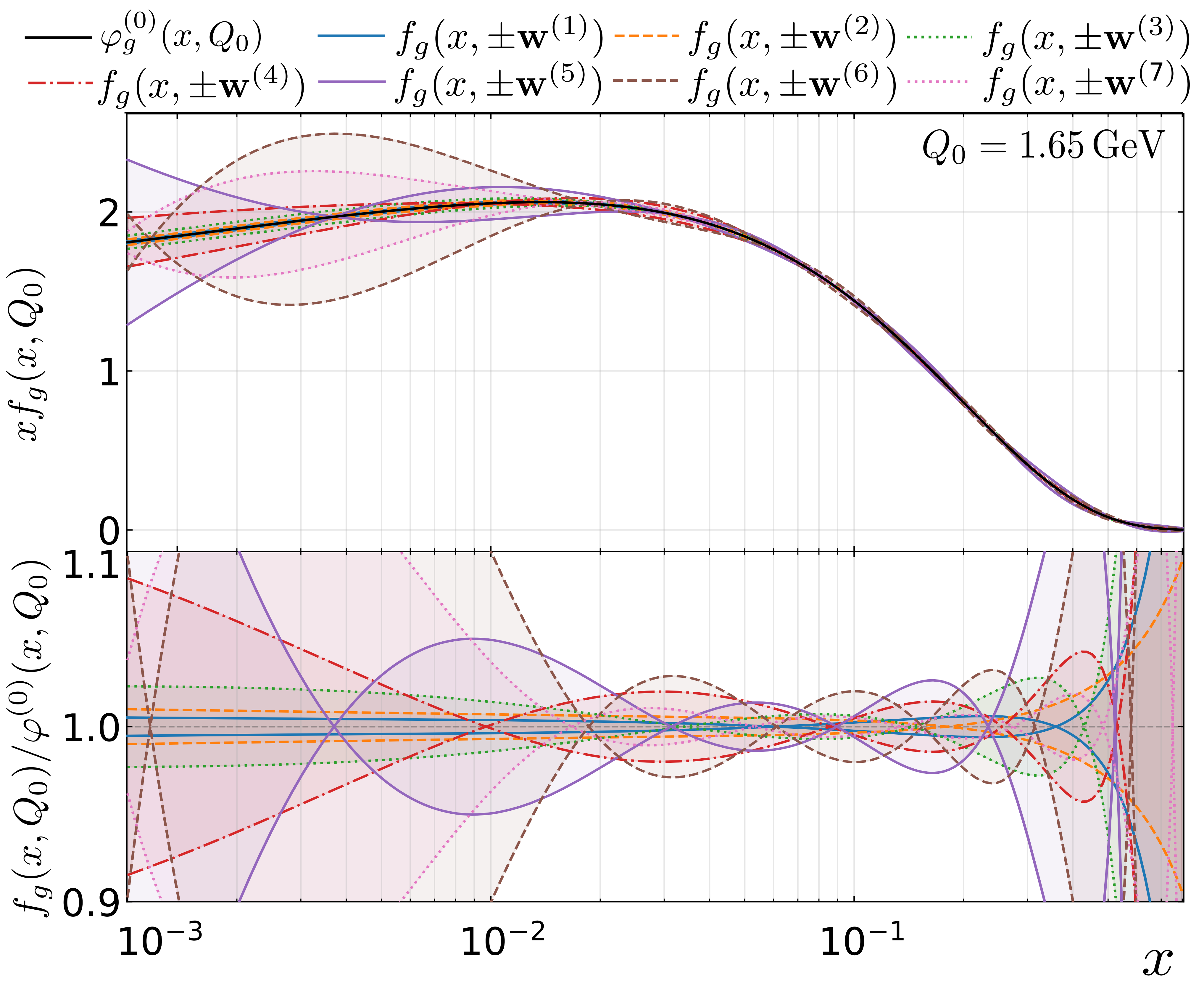


Feature vectors

Observable	Description	Observable	Description
$m(\mathrm{t\bar{t}})$	invariant mass of the top quark pair	$p_T(\mathrm{t\bar{t}})$	p_{T} of the top quark pair
$y(\mathrm{t\bar{t}})$	rapidity of the top quark pair	$\Delta\eta(\mathrm{t\bar{t}})$	difference of η of the $\mathrm{t\bar{t}}$ system
$\Delta\lvert\eta\rvert(\mathrm{t\bar{t}})$	absolute rapidity difference of the $\mathrm{t\bar{t}}$ system	$p_T(t)$	p_{T} of the top quark
$p_T(\bar{t})$	p_{T} of the anti-top quark	$y(t)$	rapidity of the top quark
$y(\bar{t})$	rapidity of the anti-top quark	$p_T(\ell_0)$	p_{T} of the leading lepton
$p_T(\ell_1)$	p_{T} of the subleading lepton	$p_T(\ell\bar{\ell})$	p_{T} of the $\ell\bar{\ell}$ system
$m(\ell\bar{\ell})$	invariant mass of the dilepton system	$\eta(\ell\bar{\ell})$	pseudorapidity of the $\ell\bar{\ell}$ system
$\Delta\eta(\ell\bar{\ell})$	difference of η of the $\ell\bar{\ell}$ system	$\Delta\lvert\eta\rvert(\ell\bar{\ell})$	difference of $\lvert\eta\rvert$ of the $\ell\bar{\ell}$ system

Flat directions

Modes wiggle around the reference gluon



Mode decomposition