Updates from NNPDF

Juan Rojo
VU Amsterdam & Theory group, Nikhef

CMS Physics Generator meeting
CMS Collaboration meeting
Bangkok, 17/12/2019

Outline

- ☑ Towards NNPDF4.0: new data, new theory, new methodology
- ☑ Parton Distributions with theory uncertainties
- ☑ Combined PDF+SMEFT fits
- ☑ Nuclear NNPDFs and the impact of LHC p+Pb data

Towards NNPDF4.0

The NNPDF timeline

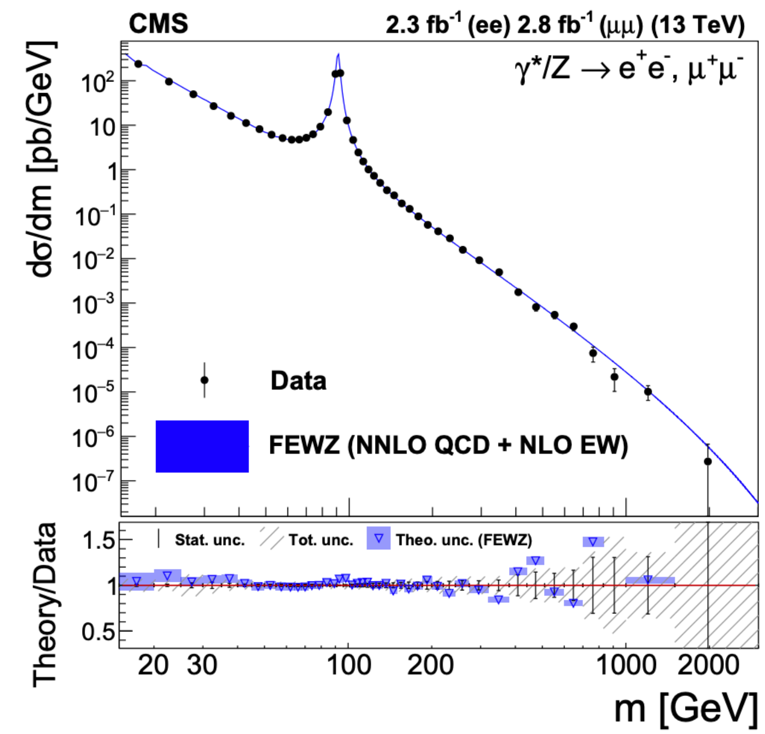
07/12

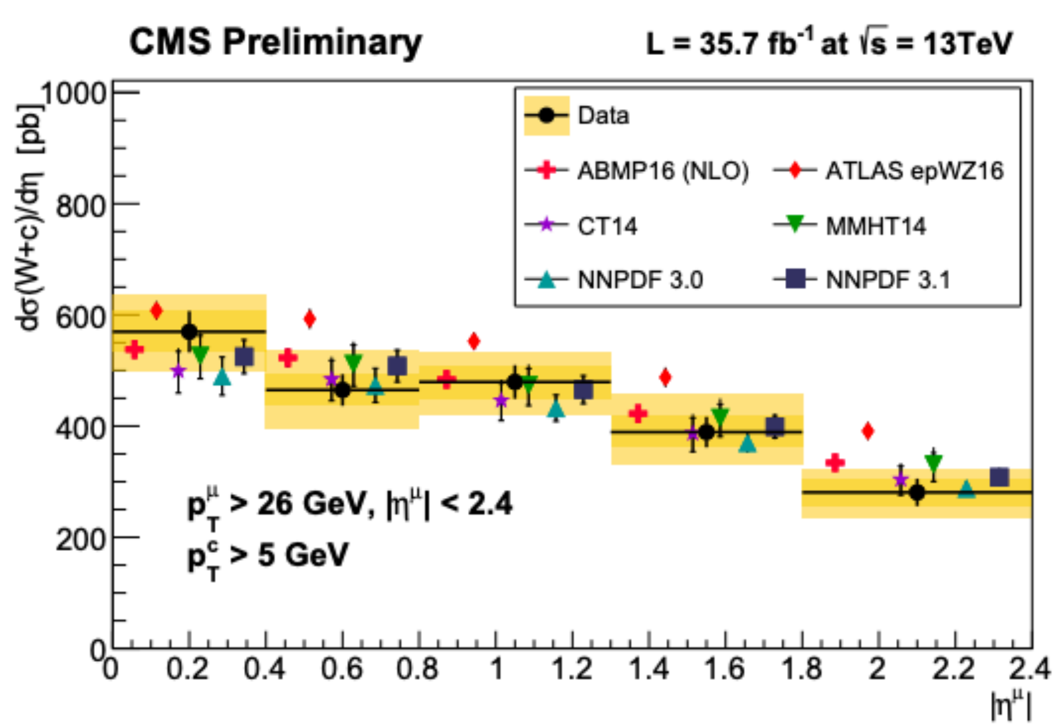
- **NNPDF2.3**: first PDF set with LHC data. LO set used in Monash 2013 Tune of Pythia8. Default (internal) PDF sets in MadGraph_aMC@NLO
- **NNPDF2.3QED**: model-independent determination of the photon PDF 04/13
- **NNPDF3.0**: large amounts of LHC data, methodological improvements closure-tested. Baseline PDF set in CMS MC event generation. 10/14
- **NNPDF3.1**: most LHC Run I data included (several new processes: ttbar differential, high-mass Drell-Yan, Z pT spectra, …). NNLO fit markedly superior than NLO one 06/17
- **NNPDF3.1QED**: LuxQED prescription for photon PDF 12/17
- **NNPDF3.1smallx**: BFKL-improved PDFs for small-x physics (HERA, forward physics) 10/17
- **NNPDF3.1TH**: first set of parton distributions with theory (MHO) uncertainties 05/19
- **NNPDF4.0**: work on progress, expected release in summer 2020 2020

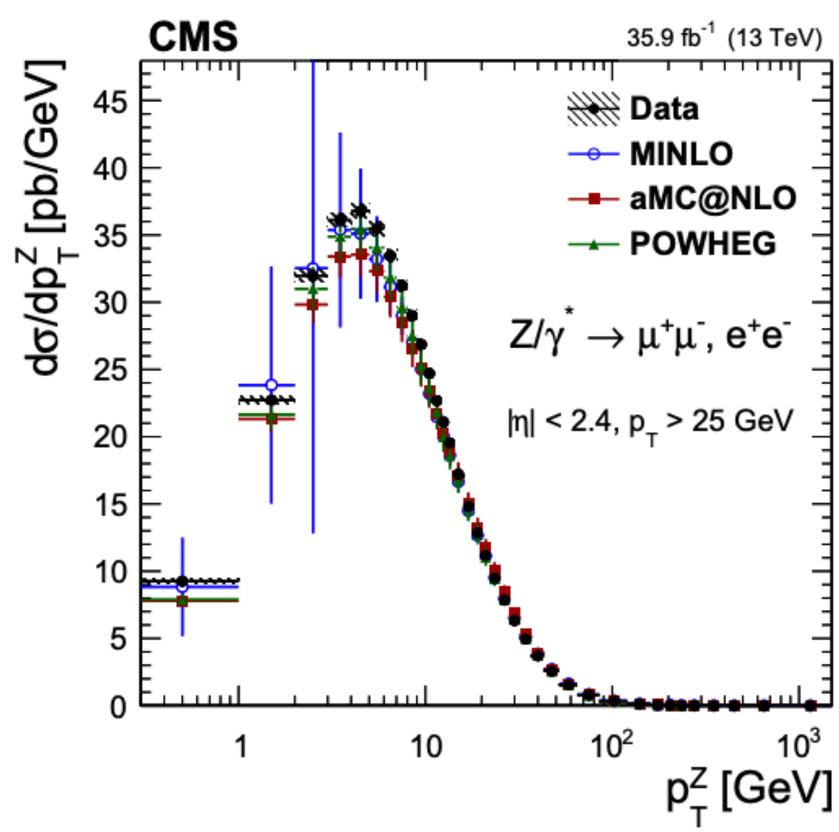
NNPDF4.0: new experimental data

Many new LHC Run I & II data included, e.g. from CMS

- Differential Drell-Yan cross-sections at 13 TeV
- Z and W pT spectra at 8 and 13 TeV **New**
- W+c at 8 and 13 TeV (strangeness)
- Dijet production at 7 and 8 TeV **New**
- Top-quark pair production at 8 and 13 TeV
- Single-top cross-sections **New**

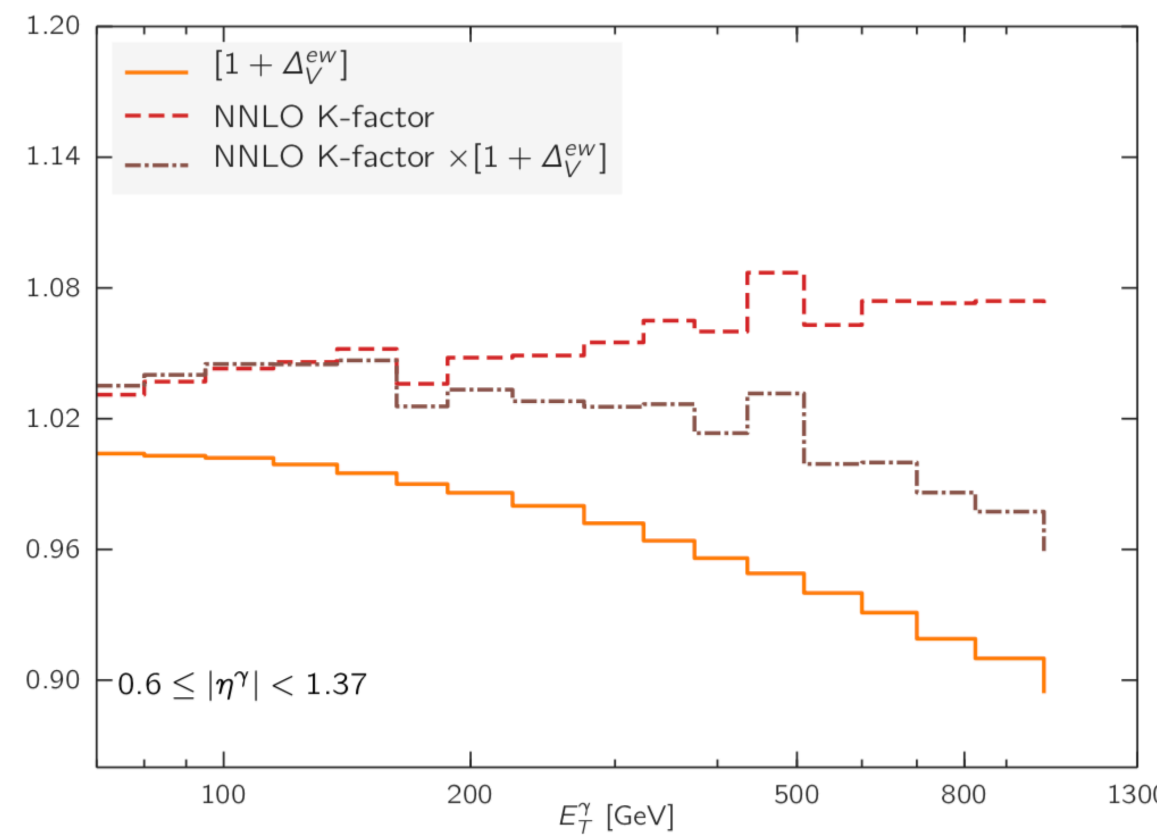




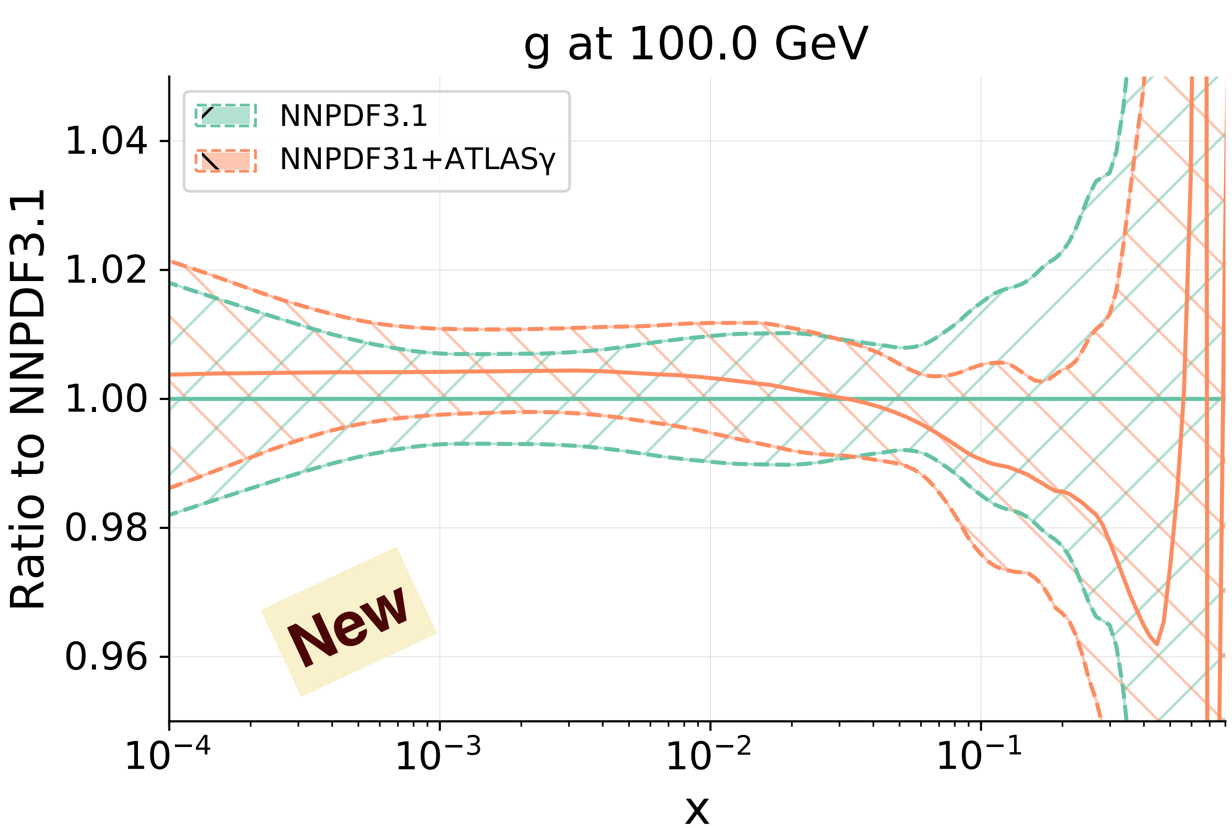


Direct photon production

- Revisited the impact of LHC **direct photon data** into the global PDF fit
  Campbell, JR, Slade, Williams 18
- Theory based on **NNLO QCD** and **LL electroweak** calculations
- Moderate impact on medium-*x* gluon
- Good **consistency** with other gluon-sensitive experiments in NNPDF3.1



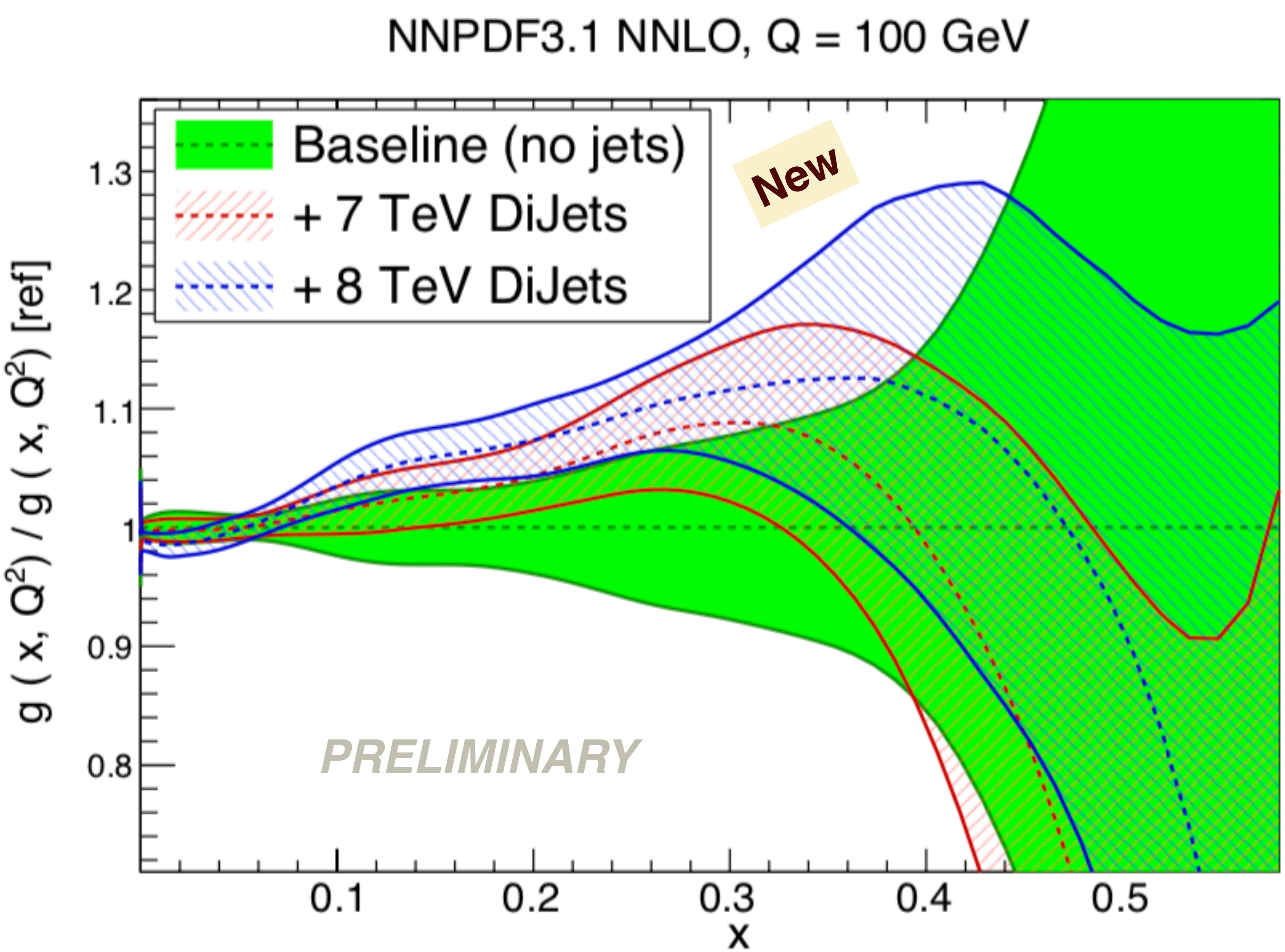
	NNPDF3.1	NNPDF3.1+ATLASγ
Fixed-target lepton DIS	1.207	1.203
Fixed-target neutrino DIS	1.081	1.087
HERA	1.166	1.169
Fixed-target Drell-Yan	1.241	1.242
Collider Drell-Yan	1.356	1.346
Top-quark pair production	1.065	1.049
Inclusive jets	0.939	0.915
Z p_T	0.997	0.980
Total dataset	**1.148**	**1.146**



Dijet production

- Added several new measurements of **inclusive jet** and **dijet production** at 7 and 8 TeV, including the CMS 8 TeV 3D dijet distributions, using NNLO QCD theory

- Explore sensitivity to gluon PDF and robustness wrt experimental correlation models



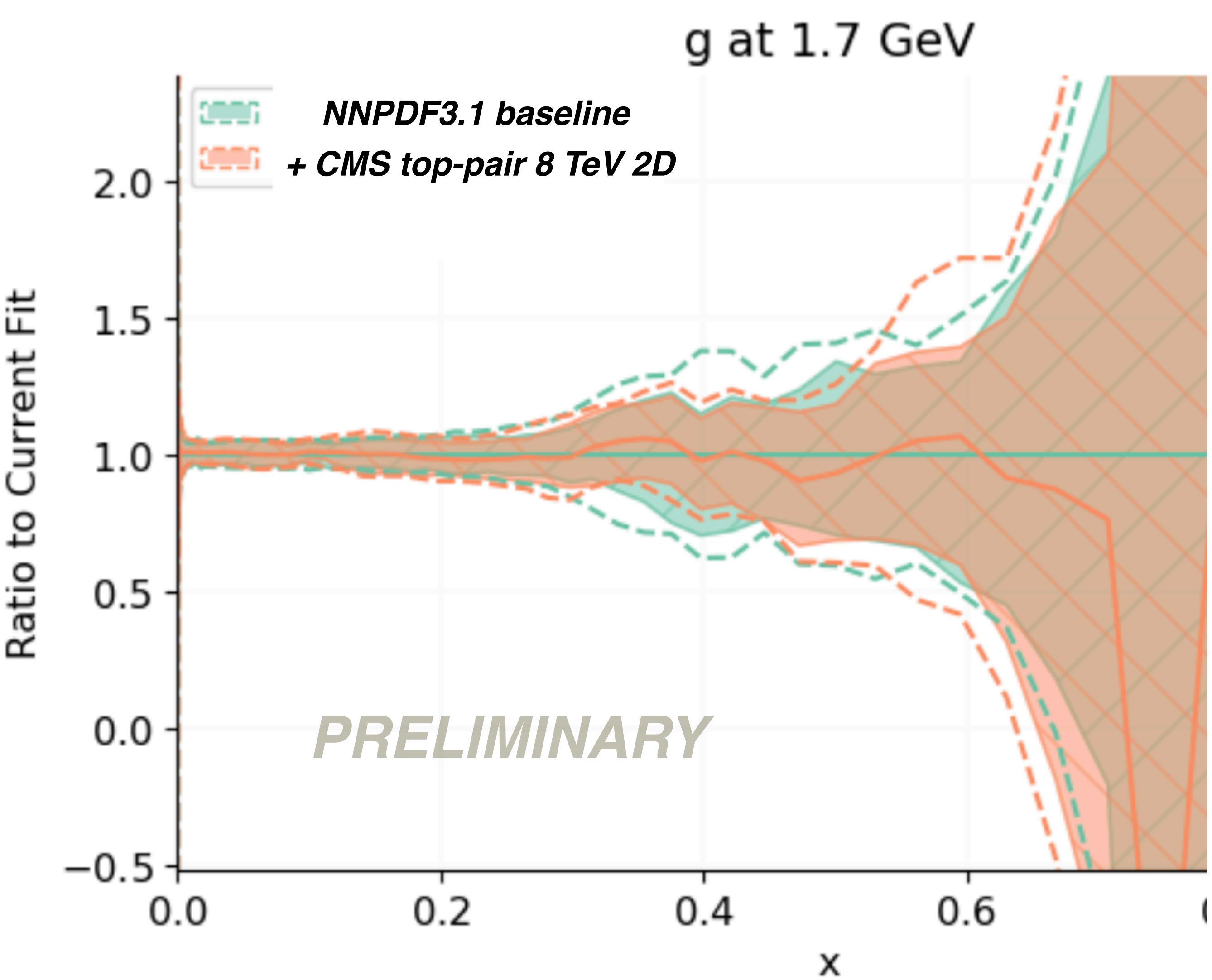
satisfactory description of dijet data within NNLO global PDF fit

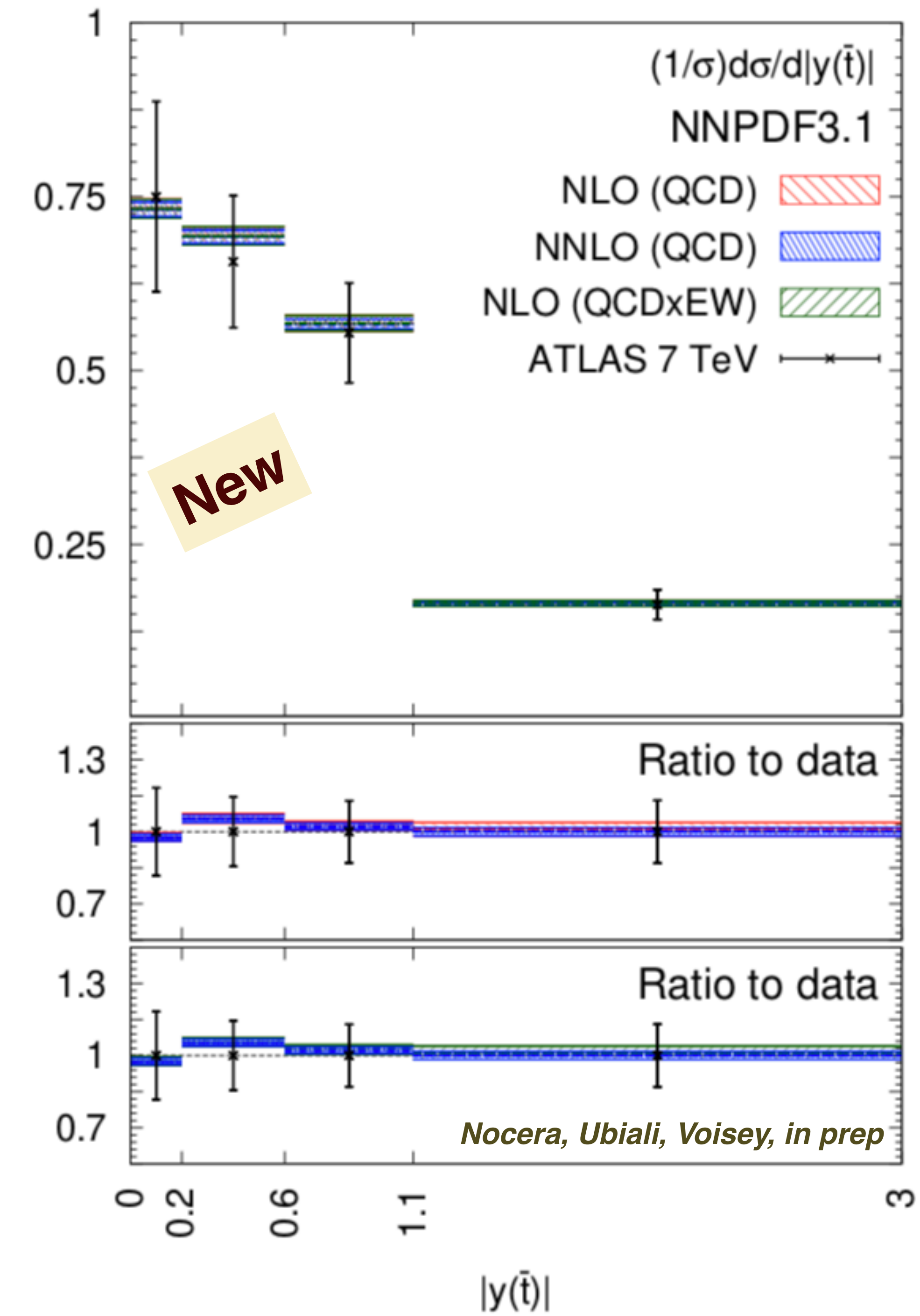
13 TeV Run II data currently restricted to low-statistics datasets

Top-quark production

- Added several new measurements of **top quark pair production at 8 and 13 TeV**, including the CMS 8 TeV 2D dilepton distributions, using NNLO QCD theory
- Also including **single-top t-channel production** (d/u ratio, bottom PDF, mass schemes)

several new handles on large-x gluon





NNPDF4.0: methodology improvements

	NNPDF3.1	***NNPDF4.0***
Random numbers	main seed, closure filter seed	multi seed
Data management	libnnpdf	same as nnfit
Neural net	fixed architecture, per flavour	**single net, flexible architecture**
Preprocessing	random fixed	**fitted in range**
Integration	a posteriori per iteration	**buildin in the model**
Optimizer	genetic optimizer	**gradient descent**
Stopping	lookback	**patience**
Positivity	penalty and threshold	**dynamic penalty, PDF must fulfill positi**
Postfit	4-sigma chi2 and arclenght	same as nnfit
Fine tuning	manual	**semi-automatic**
Model selection	closure test	closure test, **hyper optimization**

NNPDF4.0: methodology improvements

In most Machine Learning applications, the model has several parameters which are typically **adjusted by hand** (trial and error) rather than algorithmically:

- Network architecture: number of layers of neurons per layer, activation functions, …
- Choice of minimiser (which of the Gradient Descent variants?)
- Learning rate, momentum, memory, size of mini-batches, ….
- Regularisation parameters, stopping, dropout rate, patience, …

one can avoid the need of subjective choice by means of **an hyperoptimisation procedure**, where all model and training/stopping parameters are determined algorithmically

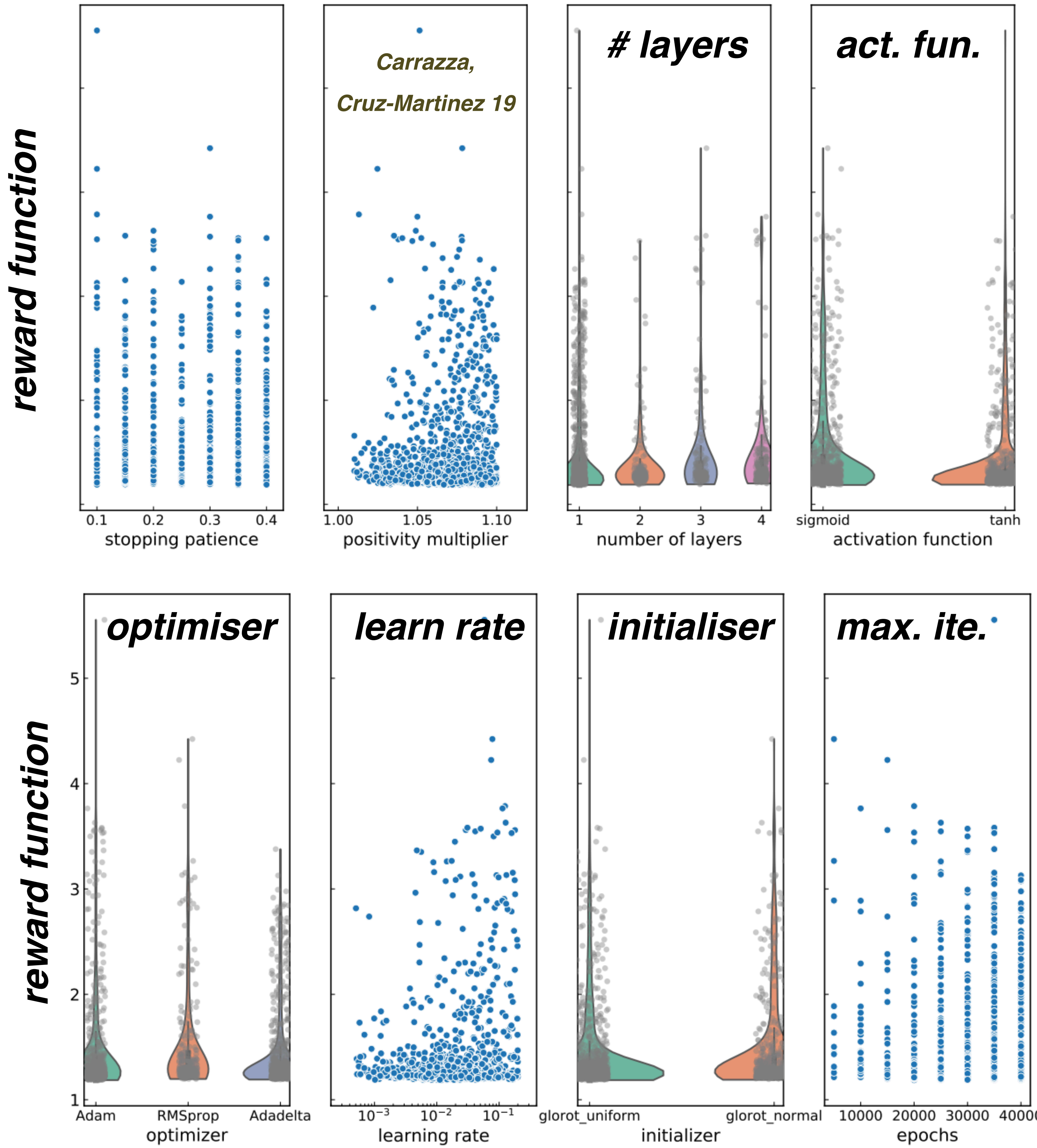
Such hyperoptimisation requires introducing a **reward function** to grade the model. Note that this is different from the **cost function:** the latter is optimised separately model by model (e.g. for each NN architecture) while the former compares between all optimised models

e.g. cost function $C = E_{\rm tr}$

$R = E_{\rm val}$ *reward function*

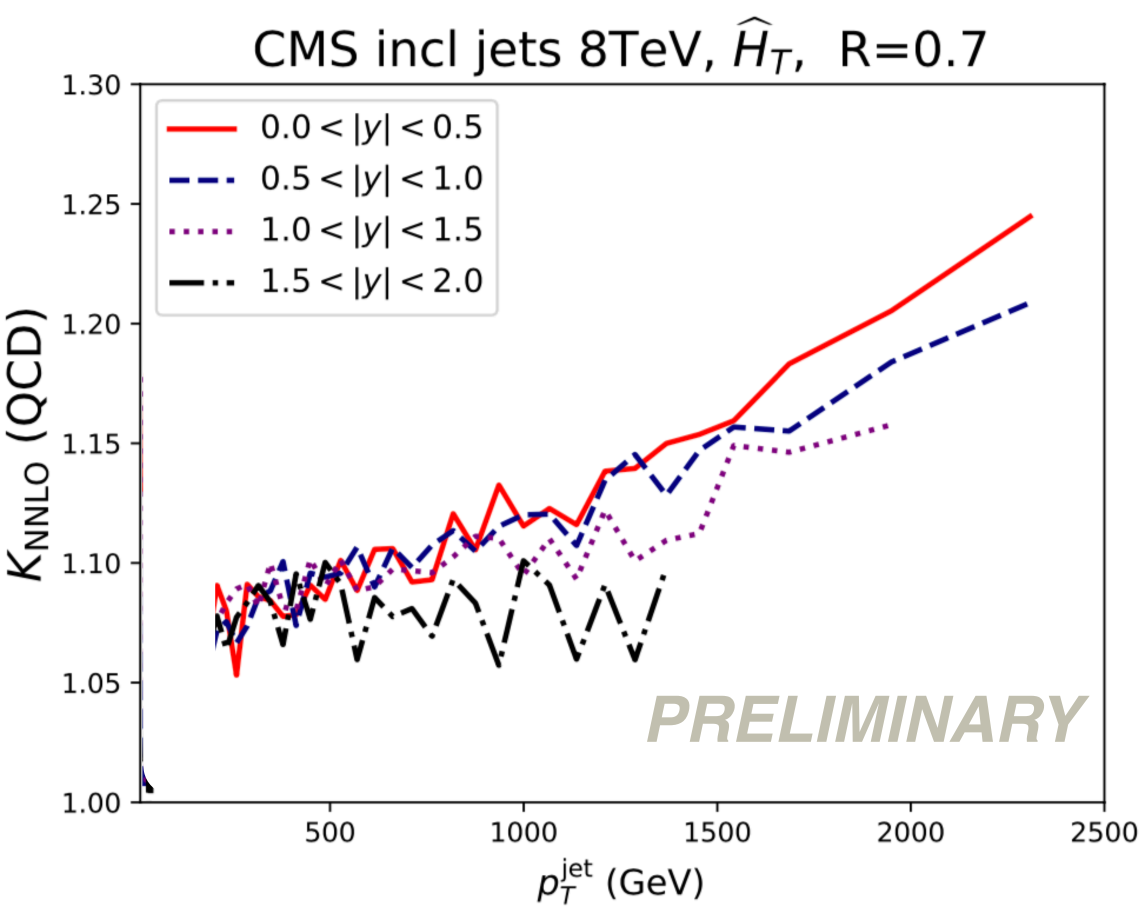
NNPDF4.0: methodology improvements

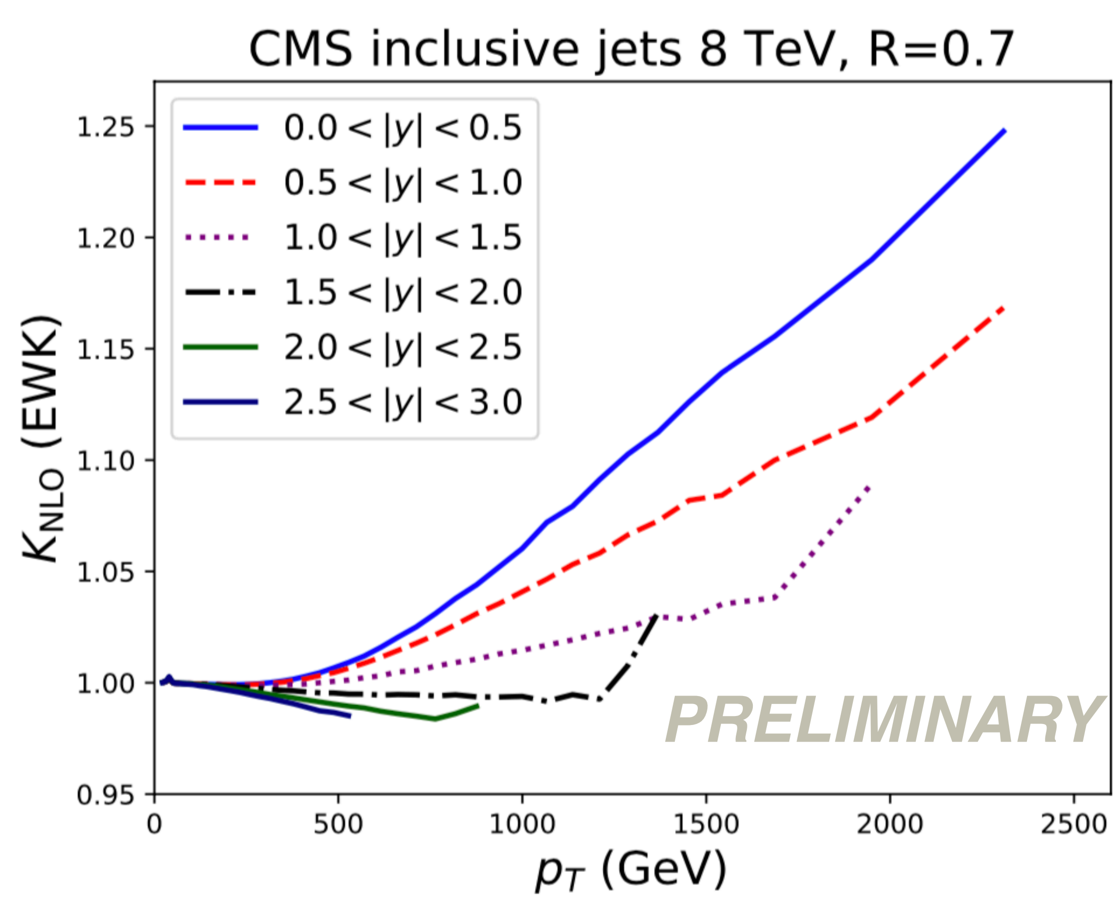
- In a hyperparameter scan one can compare the performance of **hundreds or thousands** of parameter combinations
- Some choices are **discrete** (type of minimiser, # of layers) others are **continuous** (learning rate)
- One can also **visualise** which choices are more crucial and which ones less important
- The violin plots are the **KDE-reconstructed probability distributions** for the hyper-parameters



NNPDF4.0: new theory

- For all datasets considered we use **NNLO QCD** calculations, supplemented by **NLO electroweak corrections** (with photon-induced processes) when relevant
- Improved treatment of charged-current DIS scattering with **NNLO heavy quark mass corrections** (neutrino data, strangeness)
- **Missing Higher Order uncertainties** (MHOUs) and nuclear uncertainties included systematically in the PDF fits





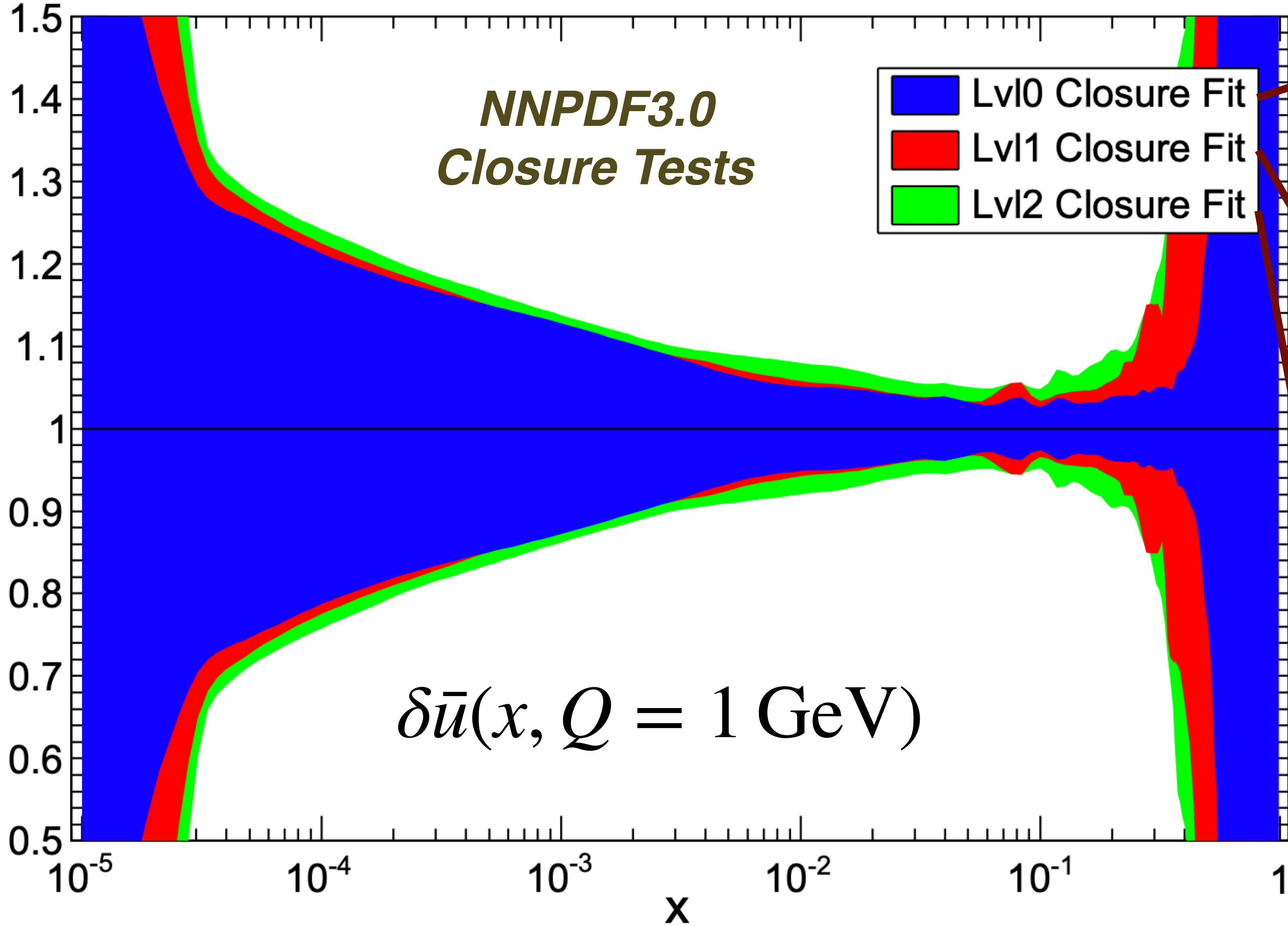
Parton Distributions with Theoretical Uncertainties

Based on ***NNPDF Collaboration:*** *R. Abdul Khalek, R. D. Ball, S. Carrazza, S. Forte, T. Giani, Z. Kassabov, R. L. Pearson, E. R. Nocera, J. Rojo, L. Rottoli, M. Ubiali, C. Voisey, M. Wilson*

arXiv:1905.04311 *and* ***arXiv:1906.10698*** *(EPJC)*

PDF uncertainties

PDF uncertainties receive contributions from **different sources**:



extrapolation/ degeneracy

input functional form (parametrisation)

experimental data uncertainties

What about **theory uncertainties**? Parametric theory errors from $\delta\alpha_S$, δm_c are routinely accounted for, but other relevant contributions are missing ….

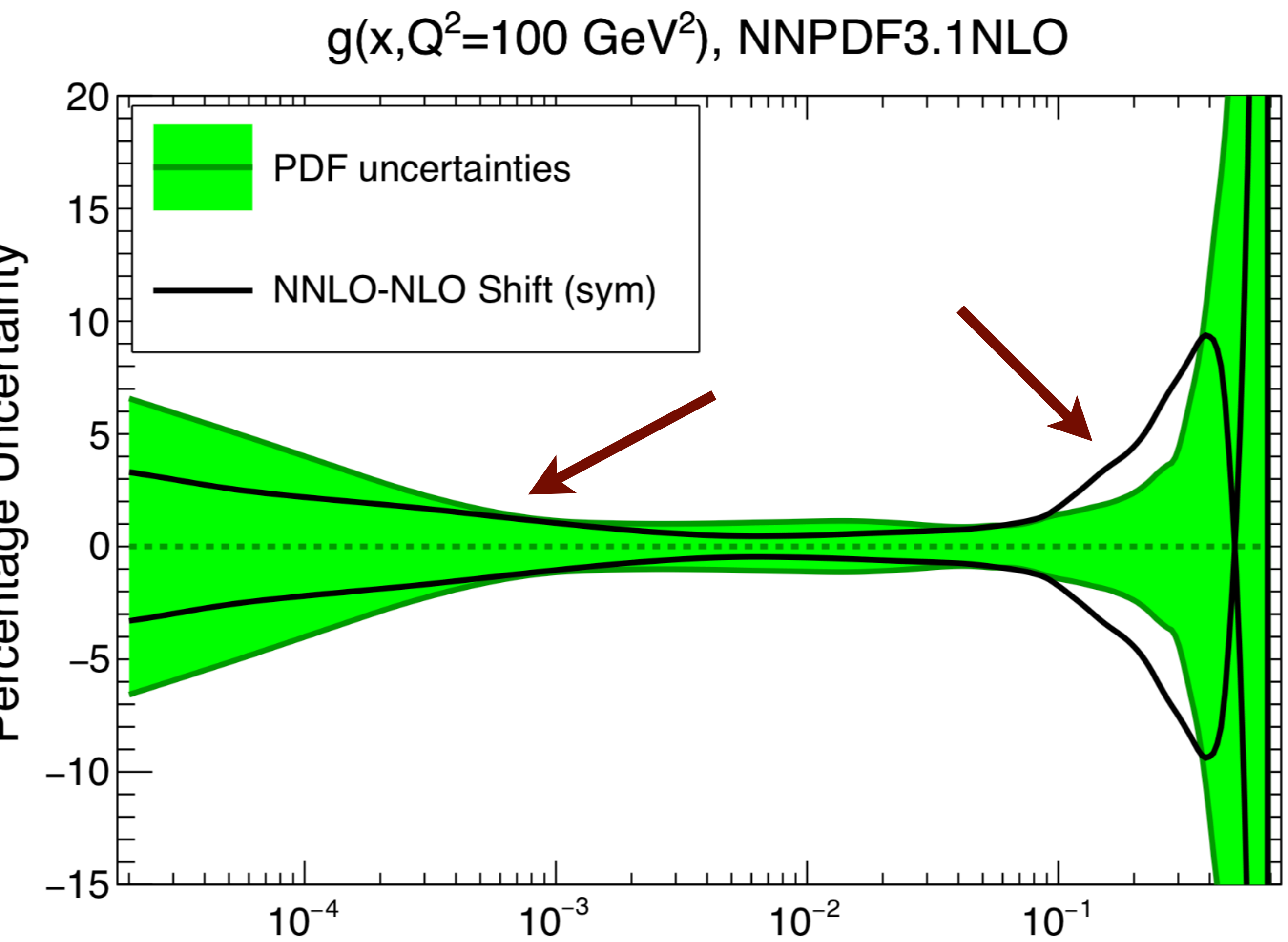
QCD uncertainties in PDF fits

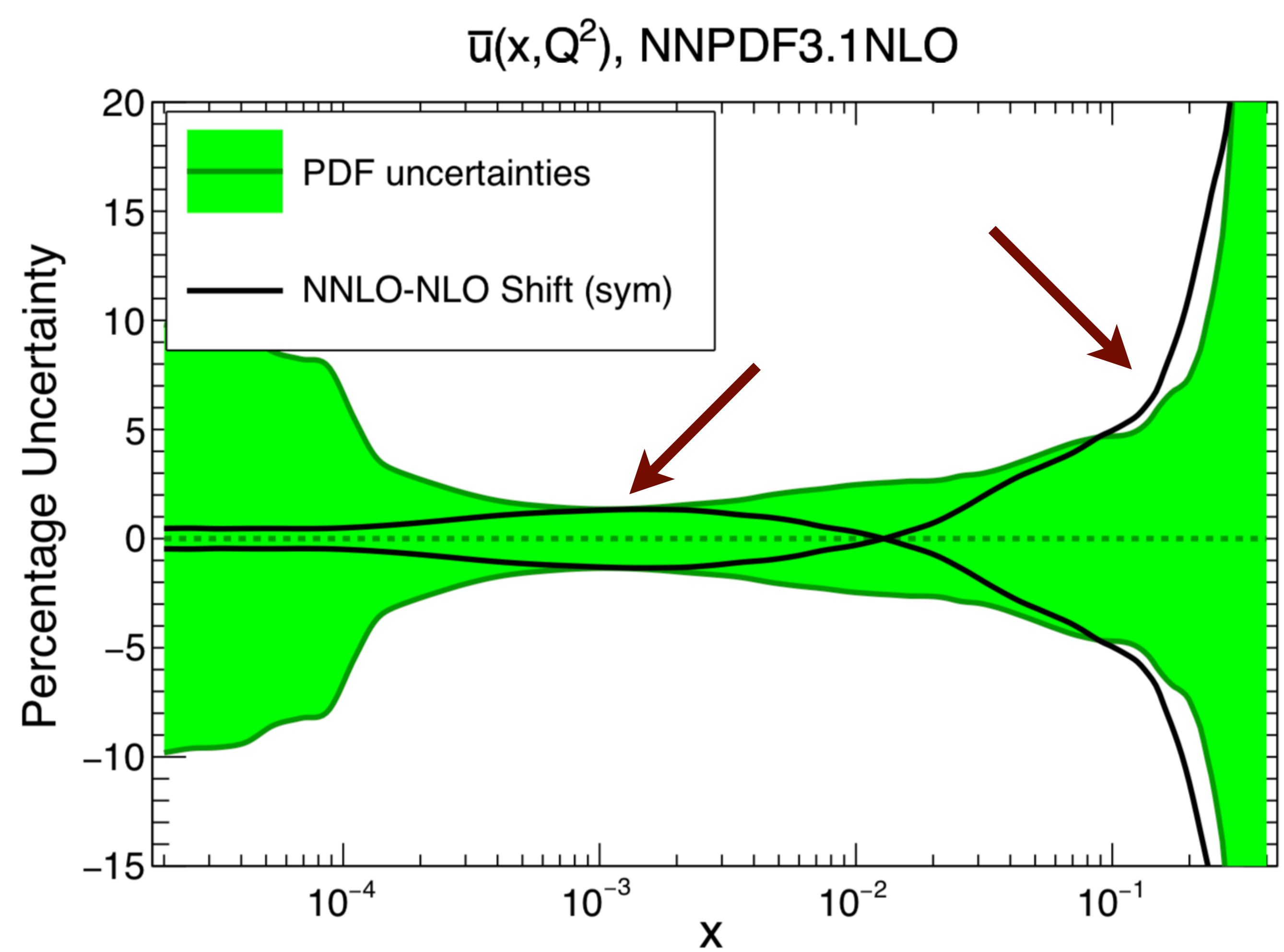
Standard global PDF fits are based on **fixed-order QCD calculations**

$$\sigma = \alpha_s^{p}\sigma_0 + \alpha_s^{p+1}\sigma_1 + \alpha_s^{p+2}\sigma_2 + \mathcal{O}(\alpha_s^{p+3})$$

The truncation of the perturbative series has associated a theoretical uncertainty: **Missing Higher Order (MHO)** uncertainty

How severe is **ignoring MHOUs** in modern global PDFs fits?





A theoretical covariance matrix

Construct a **theory covariance matrix** from **scale-varied cross-sections** and combine it with the experimental covariance matrix

$$\chi^2 = \frac{1}{N_{\rm dat}} \sum_{i,j=1}^{N_{\rm dat}} \left(D_i - T_i\right) \left(C + S\right)^{-1}_{ij} \left(D_j - T_j\right)$$

experimental ***theoretical***

assumption: theory errors are **Gaussianly distributed** around **true value**

Formally the theory covariance matrix is defined as

$$S_{ij} = \left\langle (\mathcal{T}_i - T_i)(\mathcal{T}_j - T_j) \right\rangle \equiv \left\langle \Delta_i \Delta_j \right\rangle$$

true result *actual calculation*

How to estimate these **theory systematic shifts**?

A theoretical covariance matrix

Here we use **scale variations** to estimate the MHOUs

$$\Delta_{i_a}(\kappa_f, \kappa_{r_a}) \equiv T_{i_a}(\kappa_f, \kappa_{r_a}) - T_{i_a}(0,0)$$

theory shifts

labels data points in process a

$$\kappa_{f(r_a)} = \ln \mu^2_{f(r_a)}/Q^2$$

central prediction

note: renormalisation scale variations are only correlated within the *same* process

$$S_{ij} = n_m \sum_{V_m} \Delta_{i_a}(\kappa_f, \kappa_{r_a})\Delta_{i_b}(\kappa_f, \kappa_{r_b})$$

theory covariance matrix

normalisation factor

sum over scale variations

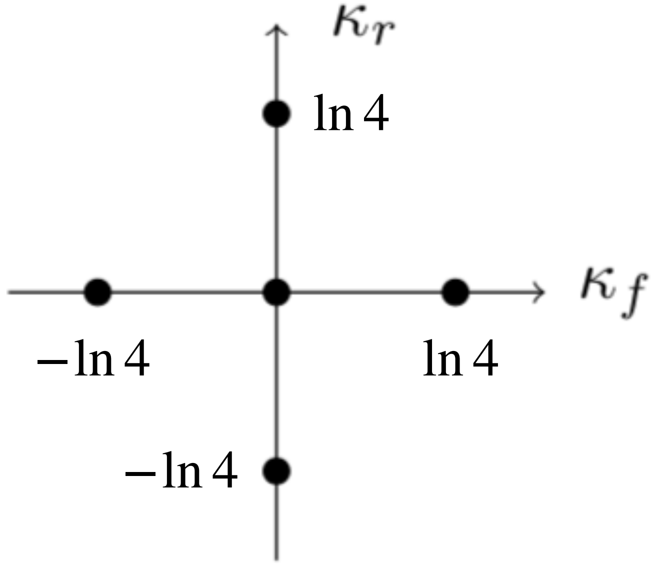
Different prescriptions for scale variations possible:

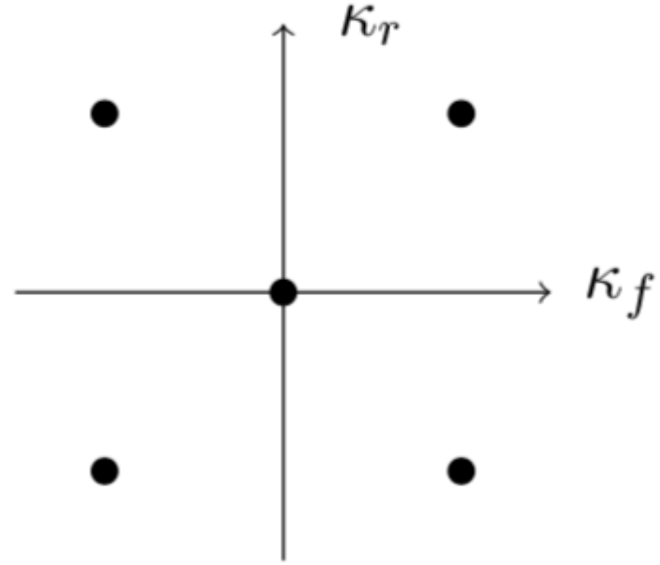
Need to validate which ones exhibit the best performance

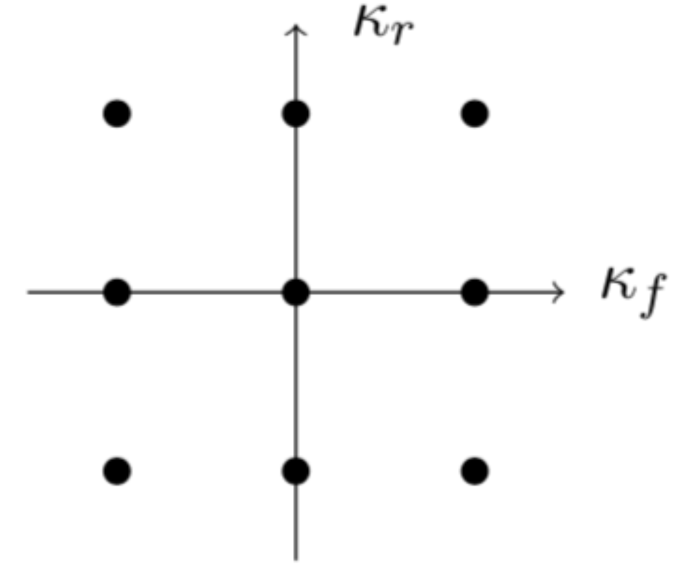
Point prescriptions

single process

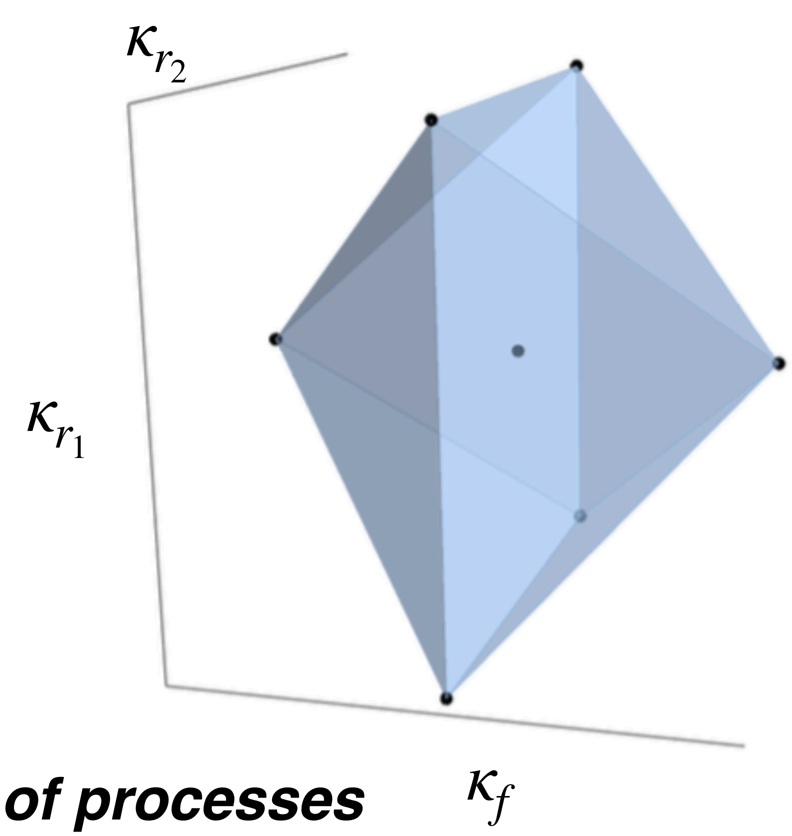
symmetric prescriptions

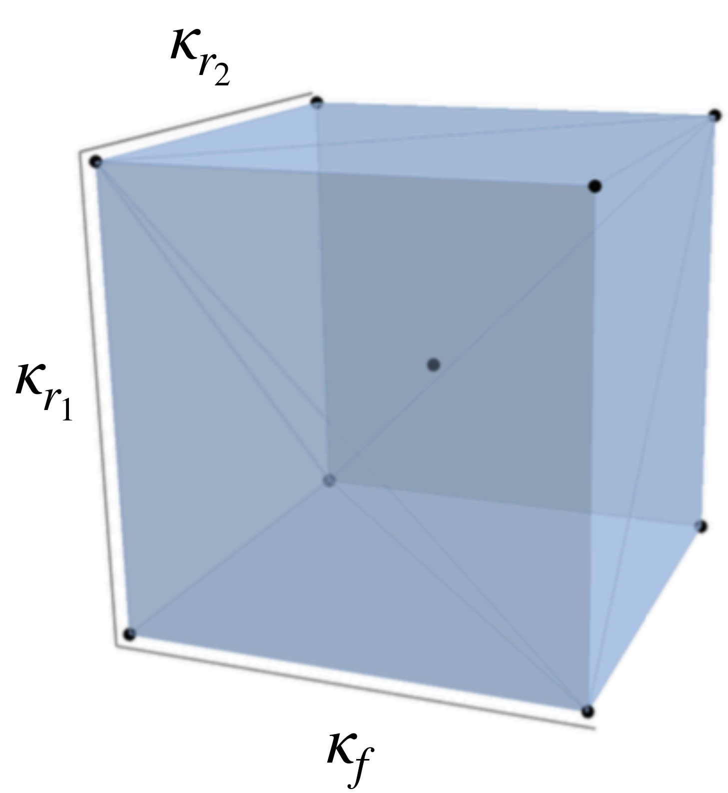


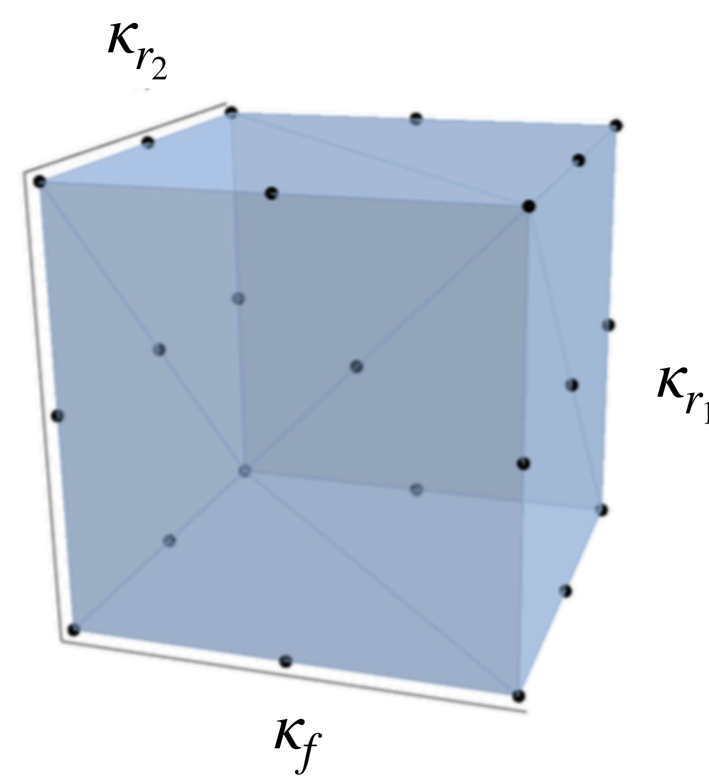




5-points ***5bar-points*** ***9-points***



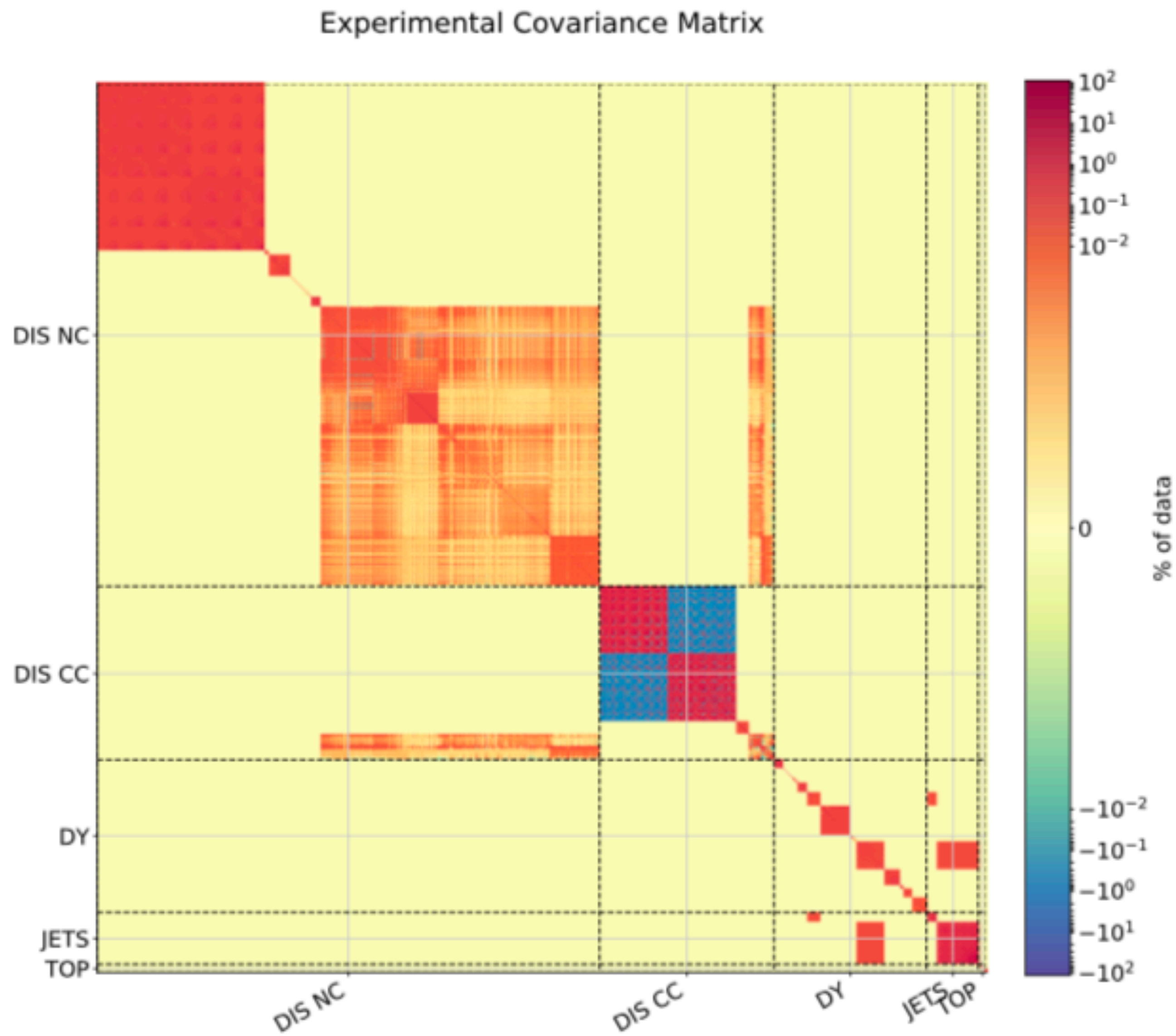


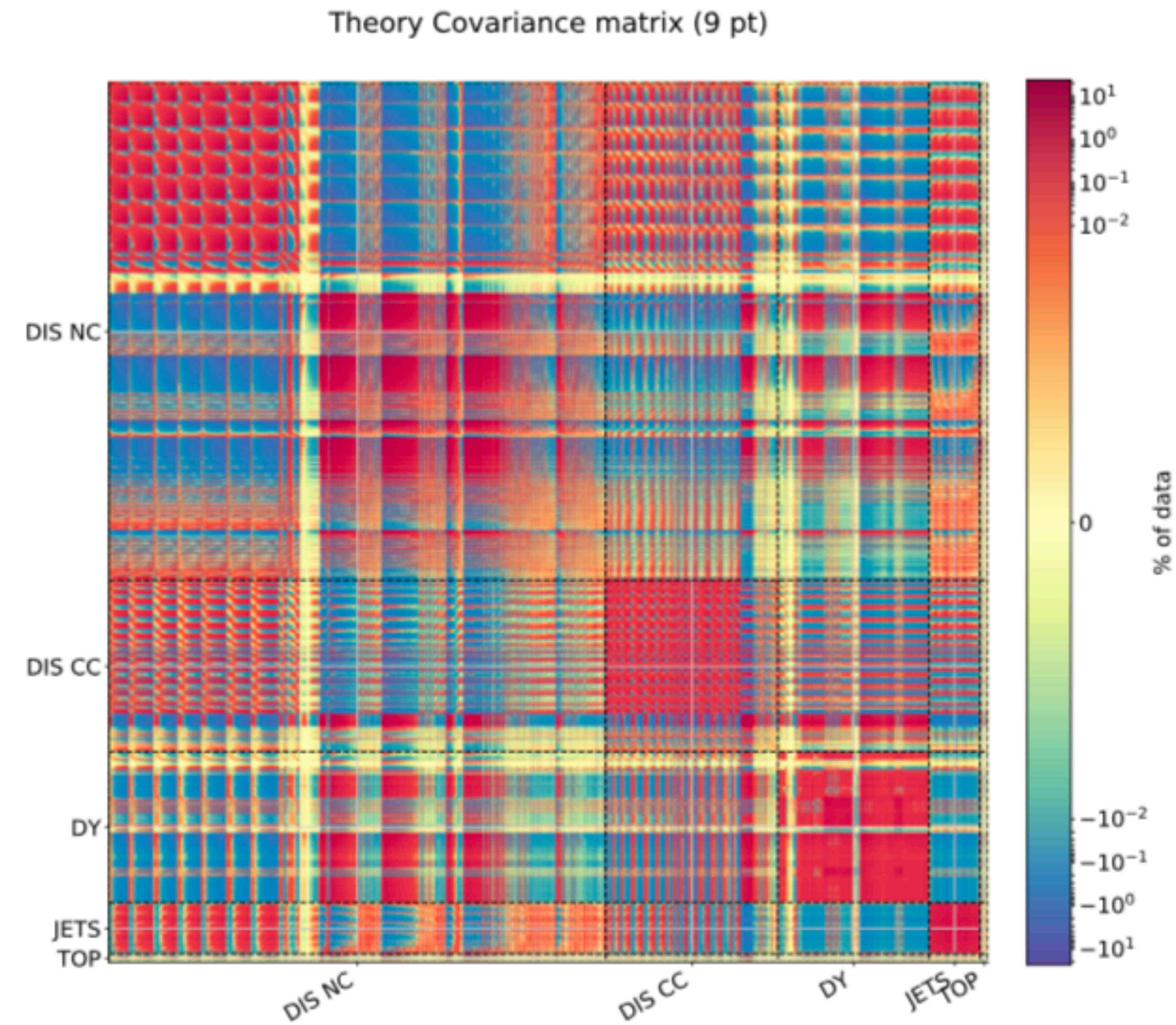


Pair of processes

The theory covariance matrix

covariance matrices

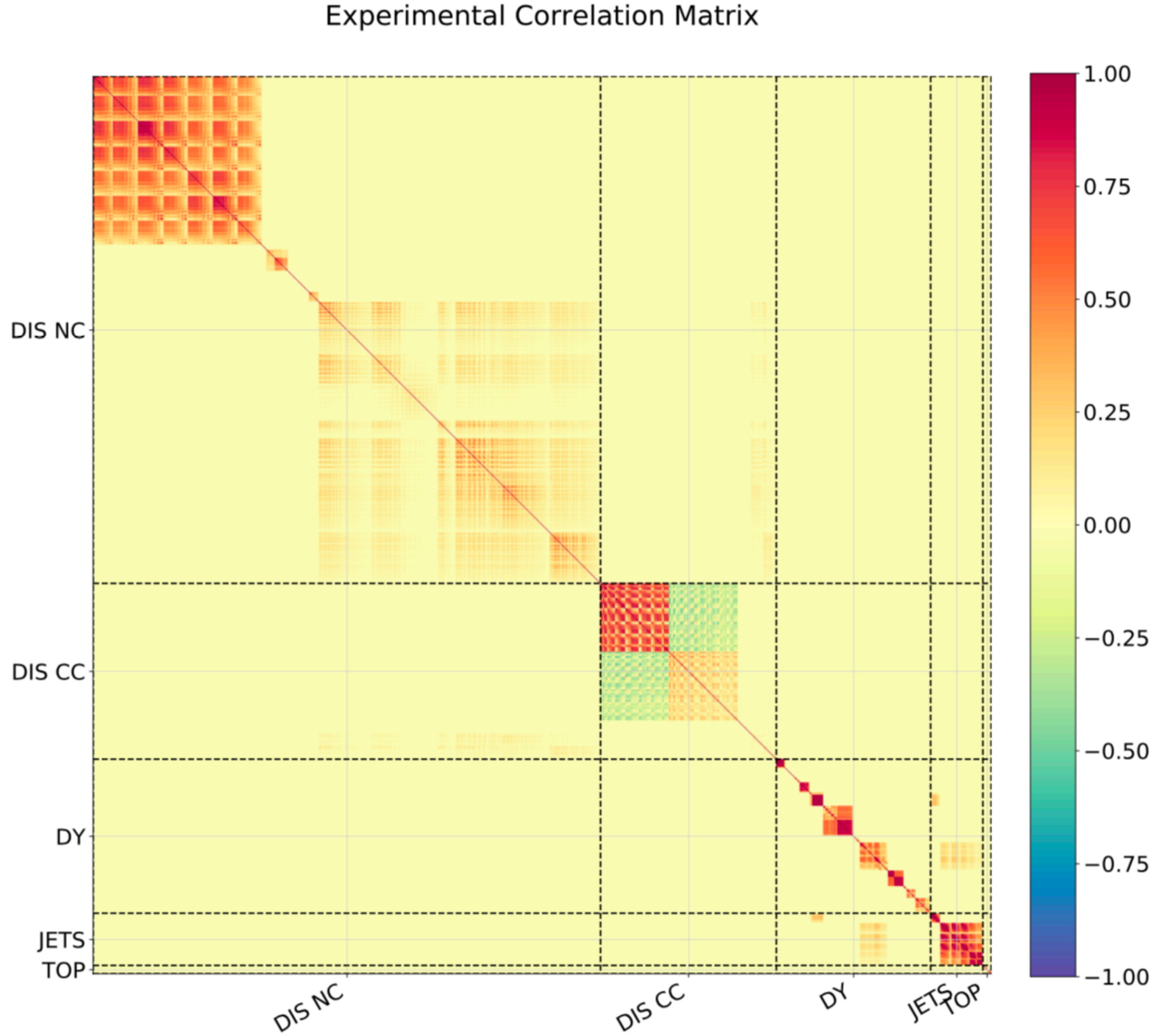


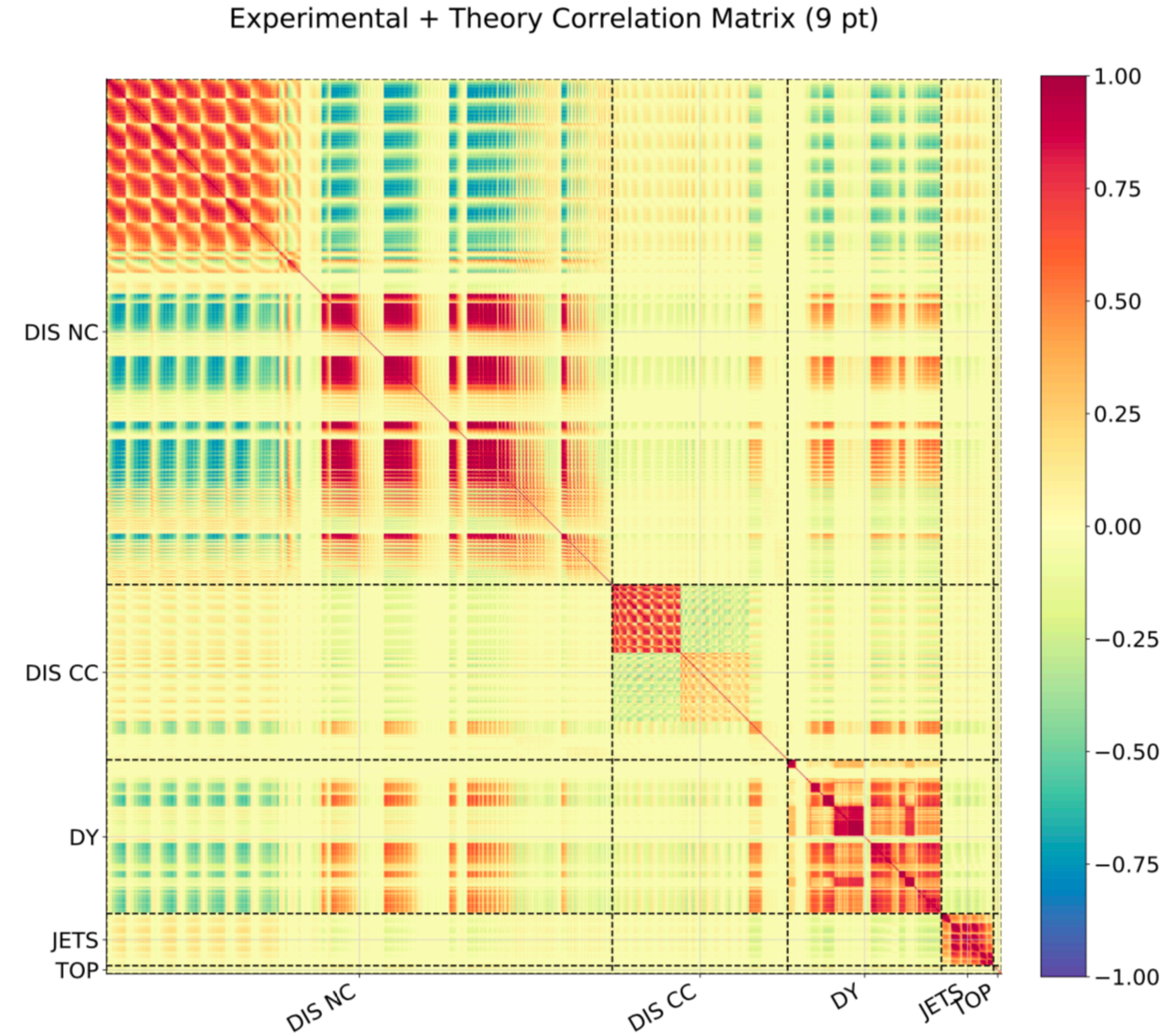


Rich pattern of **theory-induced correlations:**
Absent if only experimental errors considered

The theory covariance matrix

correlation matrices



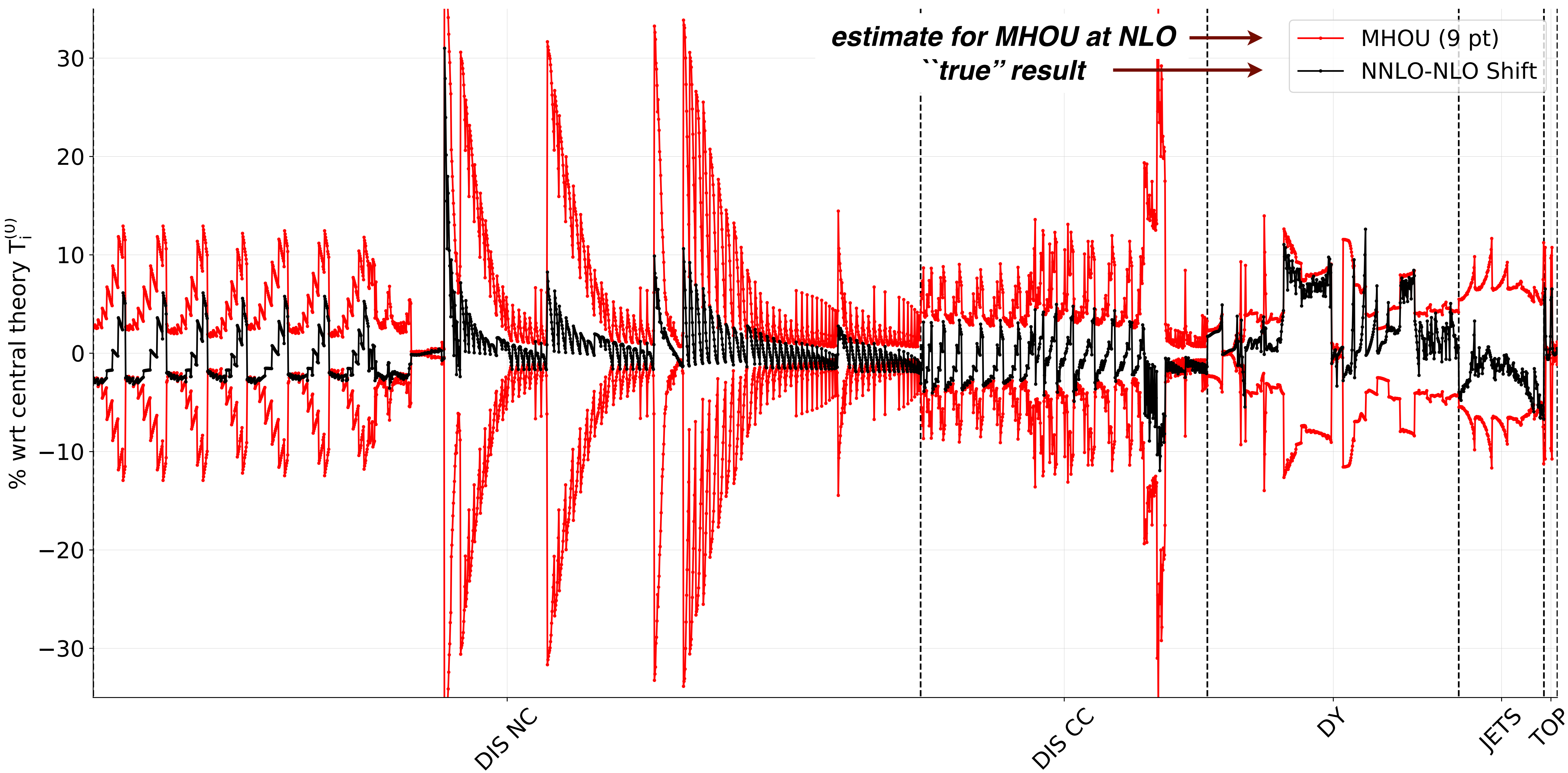


Rich pattern of **theory-induced correlations:**

Absent if only experimental errors considered

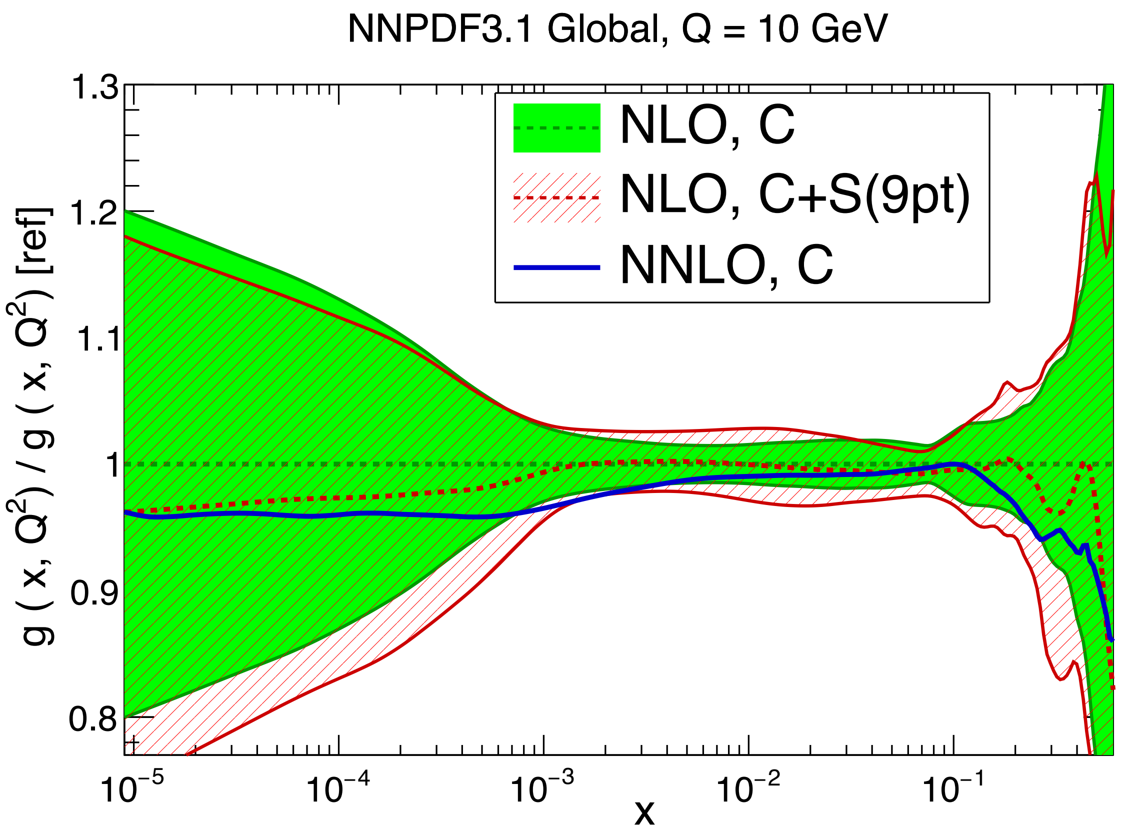
Validation

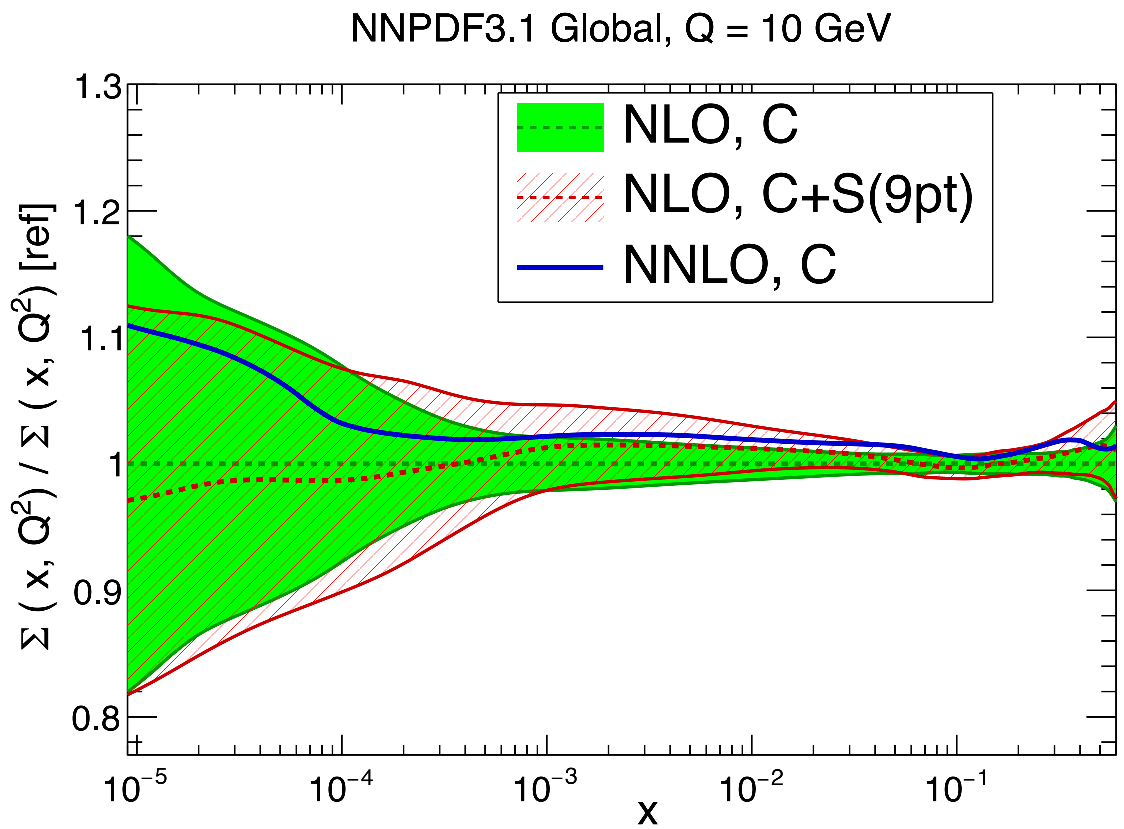
Systematic validation of NLO theory covariance matrix on the `exact' result, the **NLO=>NNLO shift**, for **O(3000) data points** of the global PDF fit



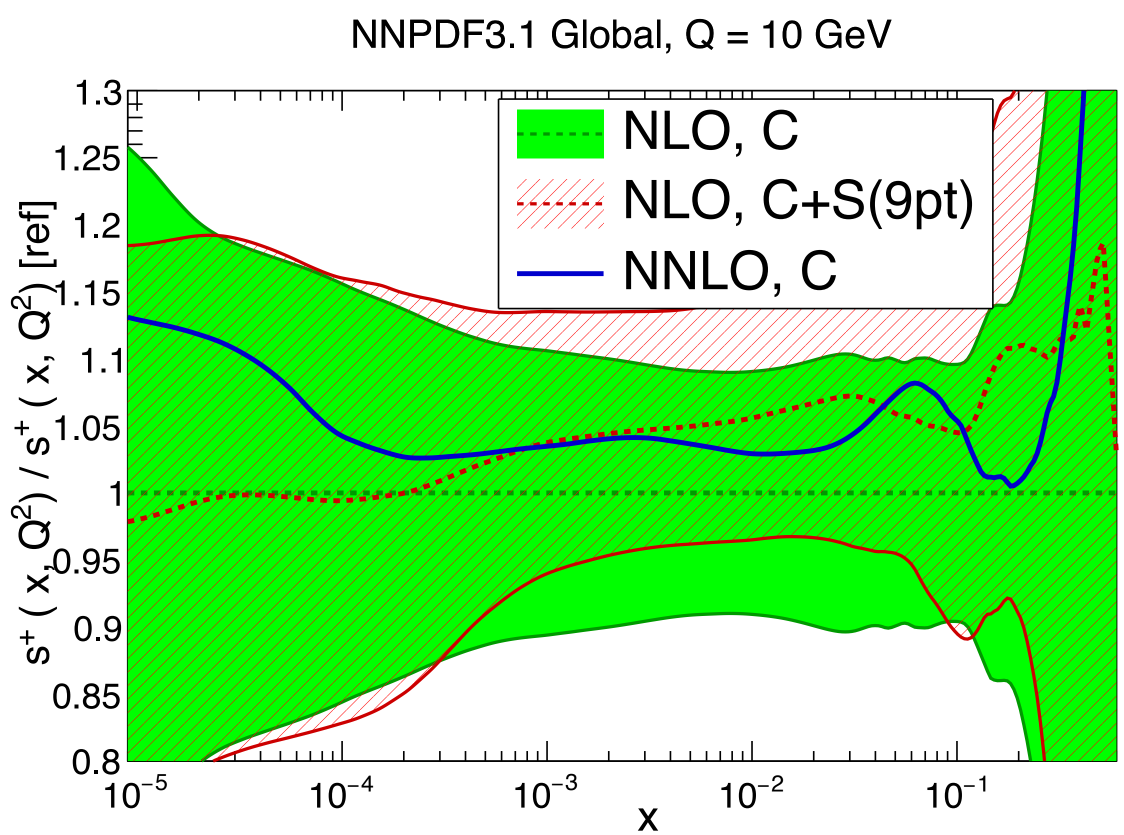
Scale variations: good estimate of MHOU for processes of relevance in PDF fits

Impact on PDFs

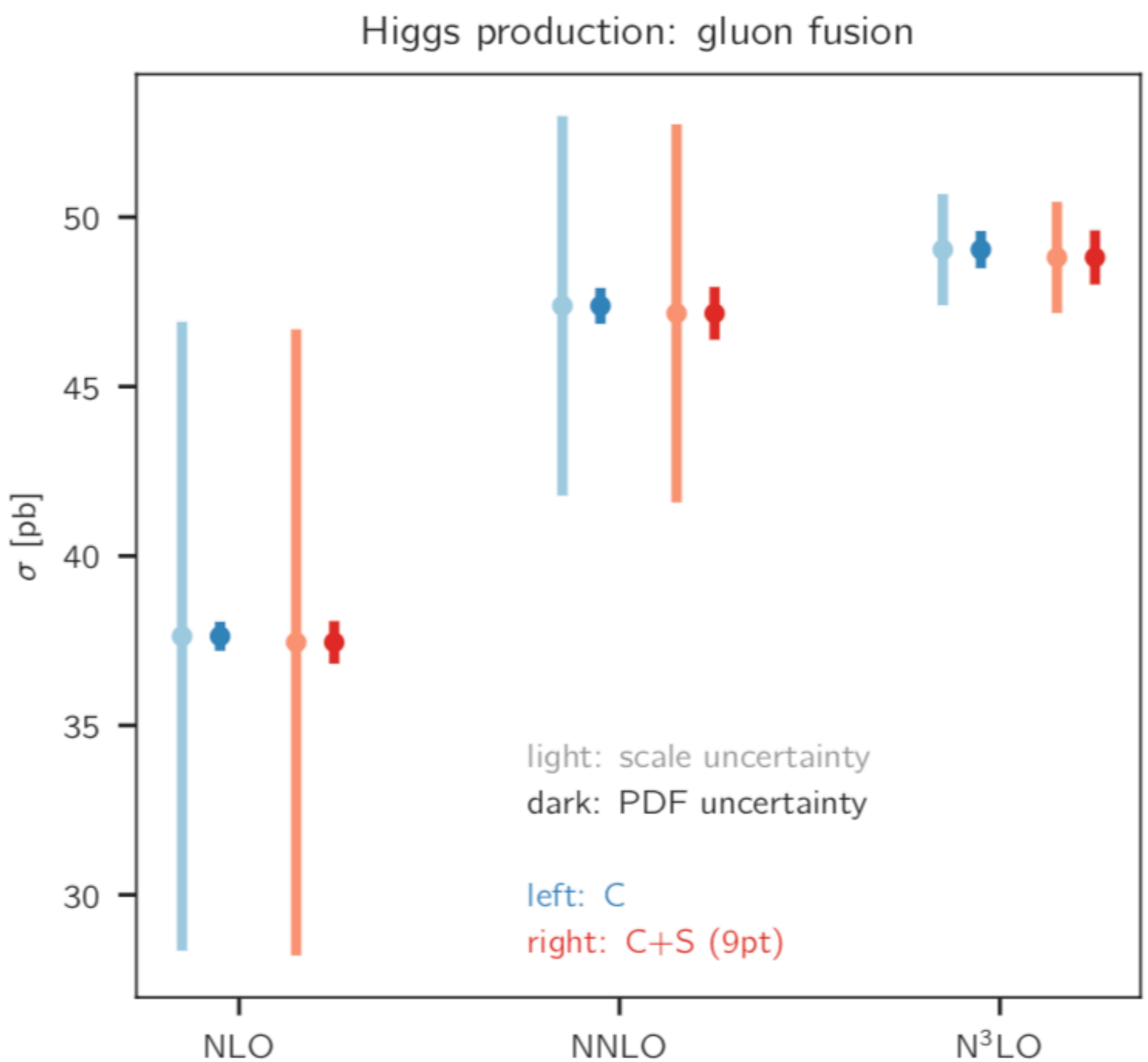


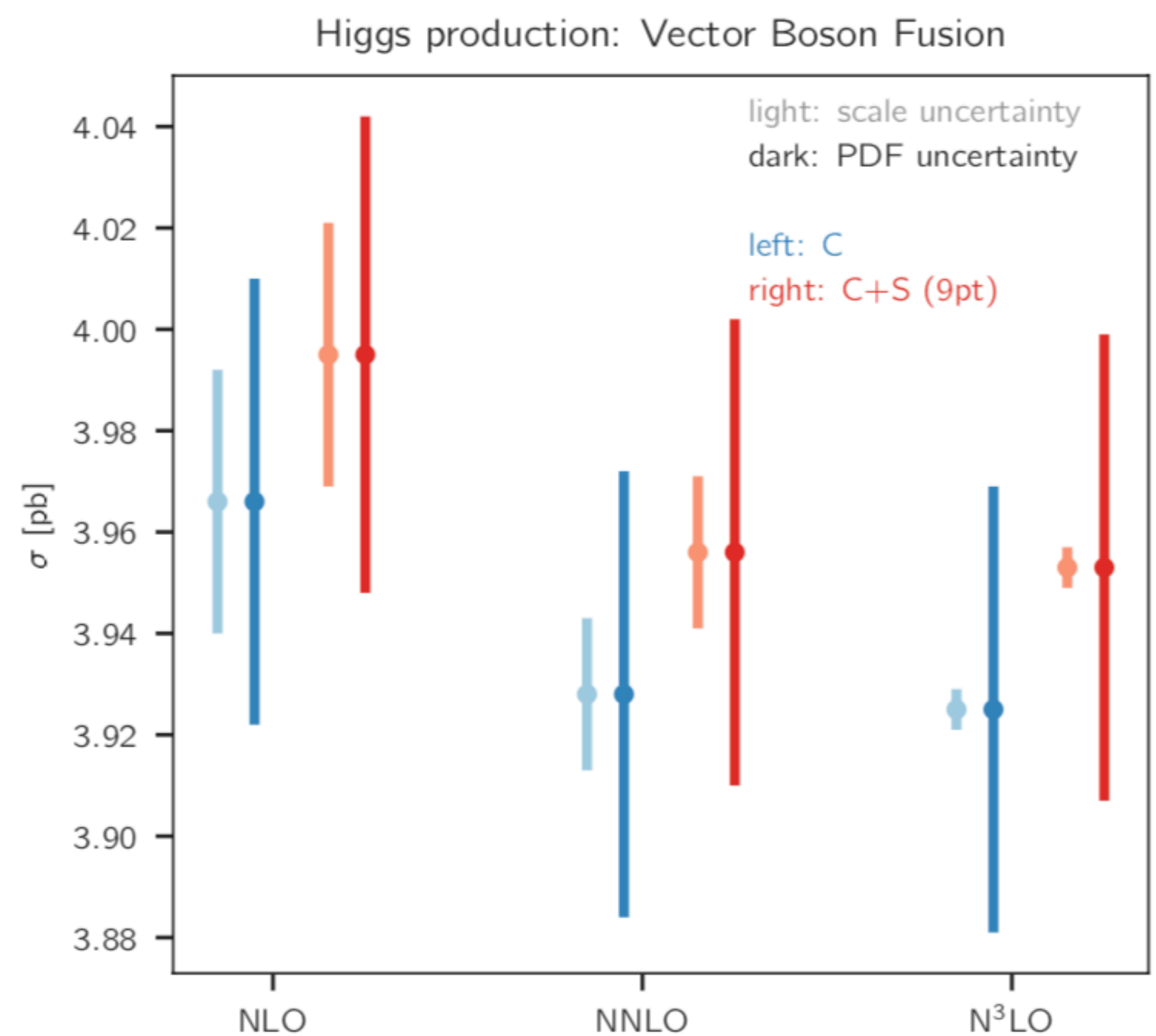


- **True result** (NLO to NNLO shift) contained within the NLO C+S band
- Central value of NLO C+S **shifts towards true result** in most cases
- Moderate (but not dramatic) **increase** in overall PDF uncertainties



Impact for LHC phenomenology

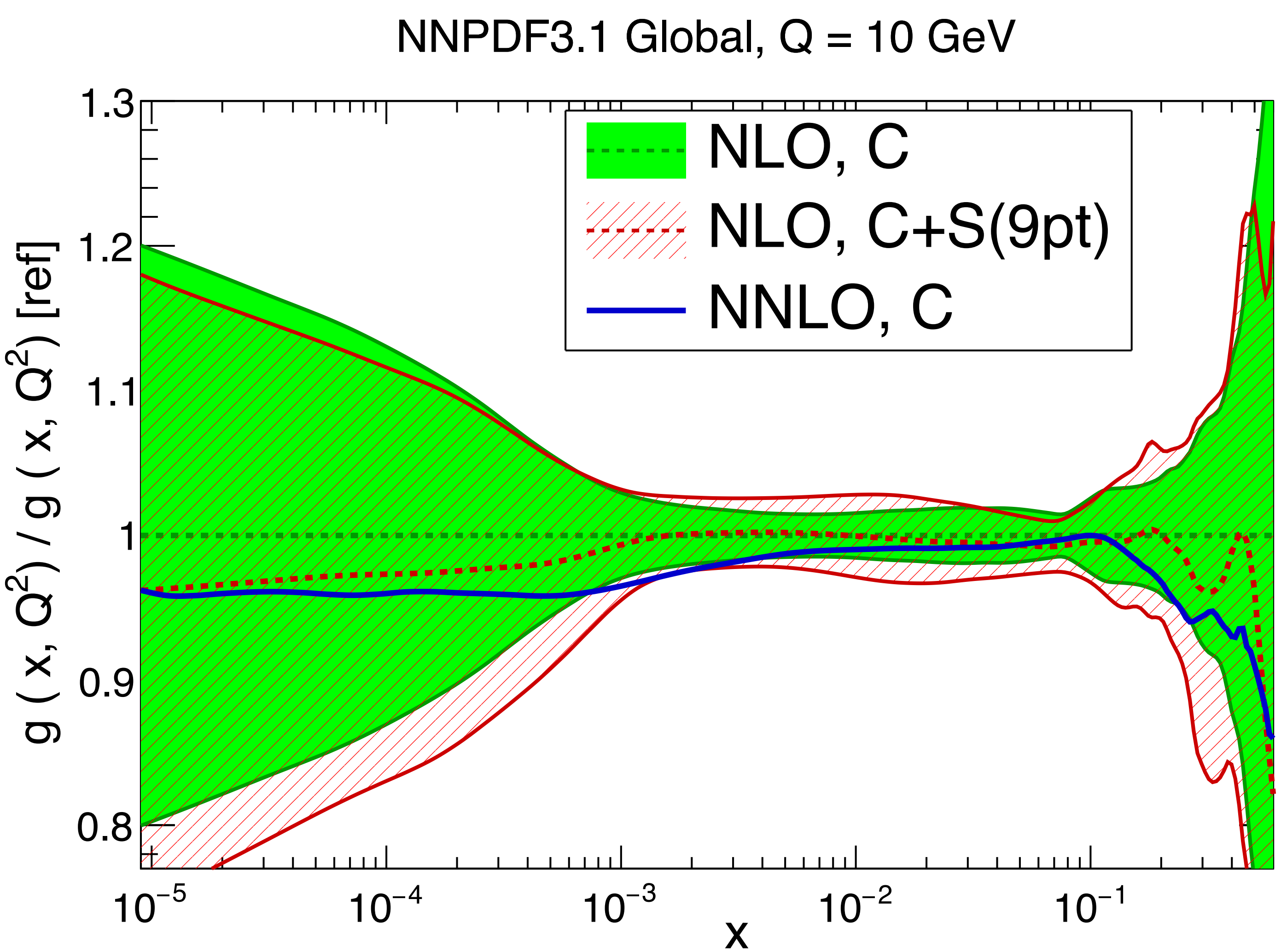




Depending on process, main consequence of **MHOUs in PDF fit for LHC pheno** is shift in central values, increase in overall PDF uncertainties, or both

Next steps

- Extend formalism to **NNLO hadronic cross-sections**
- Assess role of MHOUs in a global NNLO fit: NNPDF4.0
- Study pheno implications, e.g. Higgs production, strong coupling fits, …



Simultaneous fits of PDFs + BSM physics

Based on S. Carrazza, C. Degrande, S. Iranipour, J. Rojo, M. Ubiali

arXiv:1905.05215 *(PRL)*

The SM as an Effective Field Theory

$$\mathscr{L}_{\rm SMEFT} = \mathscr{L}_{\rm SM} + \sum_i^{N_{d6}} \frac{c_i}{\Lambda^2}\mathcal{O}_i^{(6)} + \sum_j^{N_{d8}} \frac{b_j}{\Lambda^2}\mathcal{O}_i^{(8)} + \ldots$$

- ☑ The SMEFT is the **low-energy limit** of generic UV-complete theories at high energies
- ☑ **Complete basis** at given mass-dimension: systematic parametrisation of BSM effects
- ☑ **Fully renormalizable**, full-fledged QFT: can compute higher orders in QCD and EW
- ☑ Can be matched to **any BSM model** that reduces to the SM at low energies: exploits the full power of SM ``measurements'' for model-independent BSM searches

The SMEFT is not some new model: **it is the SM** once we remove the **theoretical prejudice** of its validity up to arbitrarily large scales

SMEFT & PDF fits

- General heavy bSM physics beyond the direct reach of the LHC can be **parametrised in a model-independent** in terms of a **complete** basis of higher-dimensional operators

$$\mathcal{L}_{\rm SMEFT} = \mathcal{L}_{\rm SM} + \sum_i^{N_{d6}} \frac{c_i}{\Lambda^2}\mathcal{O}_i^{(6)} + \sum_j^{N_{d8}} \frac{b_j}{\Lambda^4}\mathcal{O}_j^{(8)} + \ldots ,$$

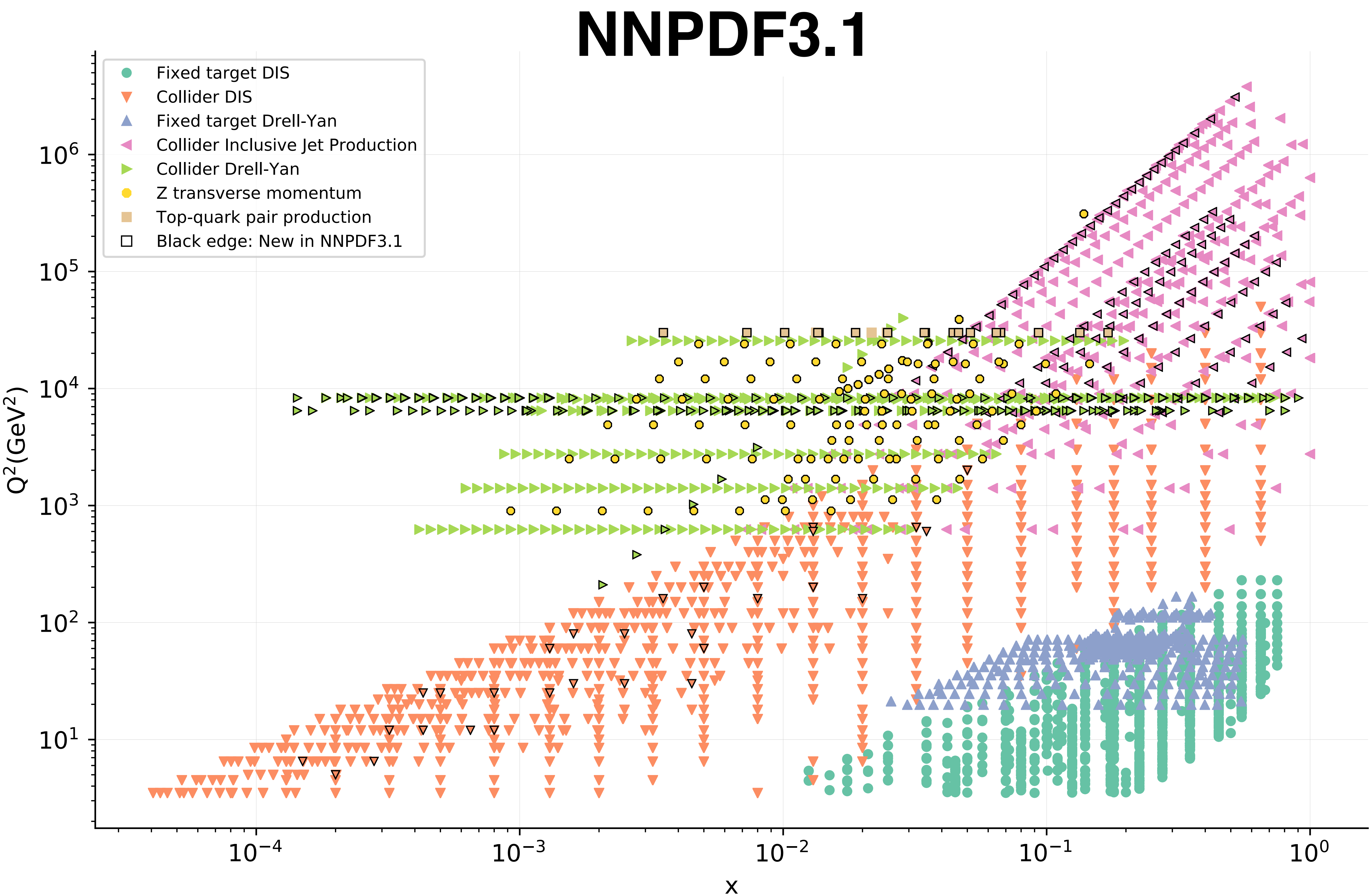
- Some operators induce **growth with the partonic centre-of-mass energy**: increased sensitivity in LHC cross-sections in the TeV region

$$\sigma(E) = \sigma_{\rm SM}(E) \times \left(1 + \sum_i^{N_{d6}} \omega_i \frac{c_i\, m_{\rm SM}^2}{\Lambda^2} + \sum_i^{N_{d6}} \widetilde{\omega}_i \frac{c_i\, E^2}{\Lambda^2} + \mathcal{O}\left(\Lambda^{-4}\right) \right)$$

many BSM/SMEFT studies interpret same data as in PDF fits: should one worry?

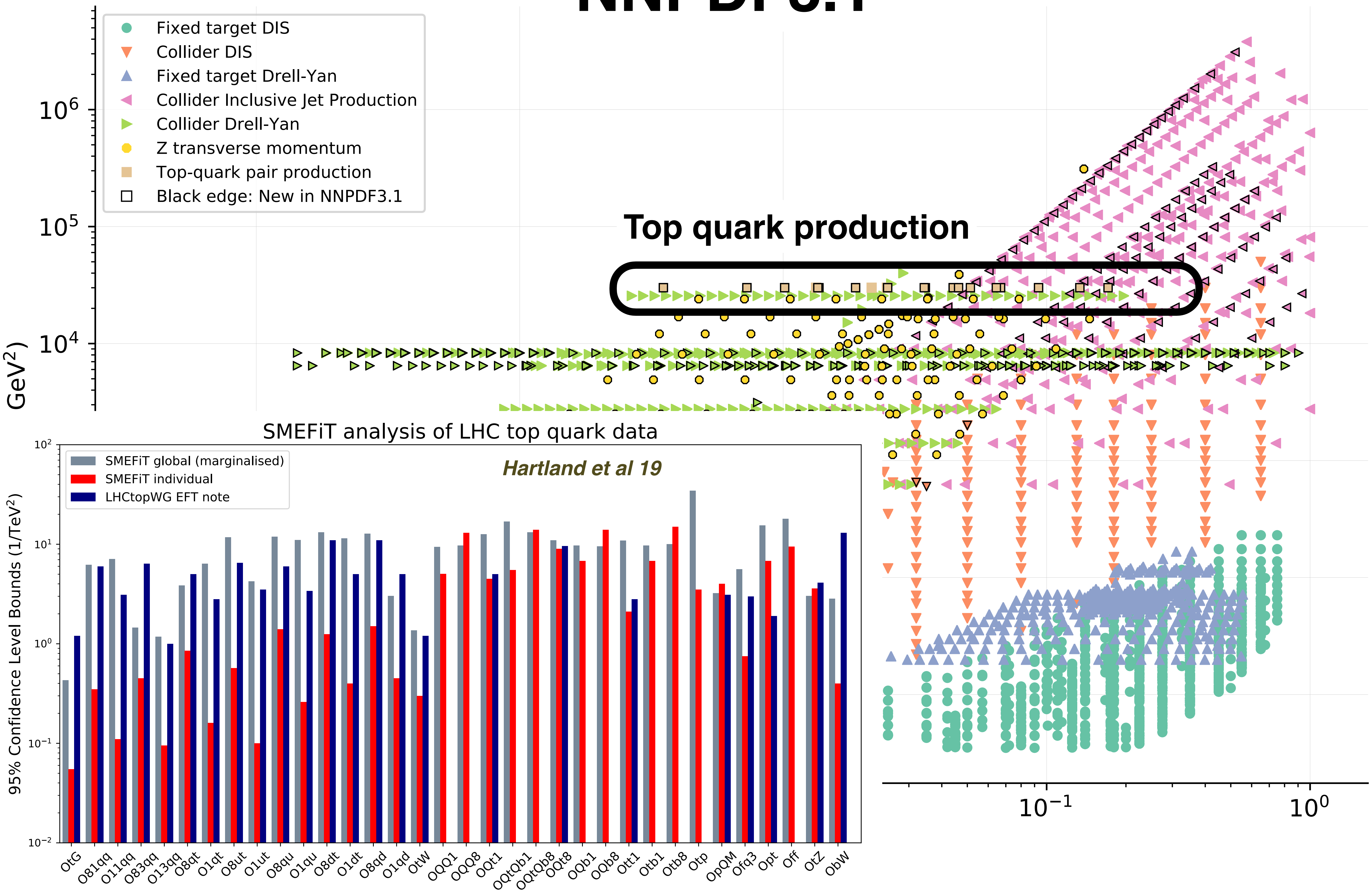
enhanced sensitivity from ***TeV-scale processes:*** *unique feature of LHC*

SMEFT & PDFs

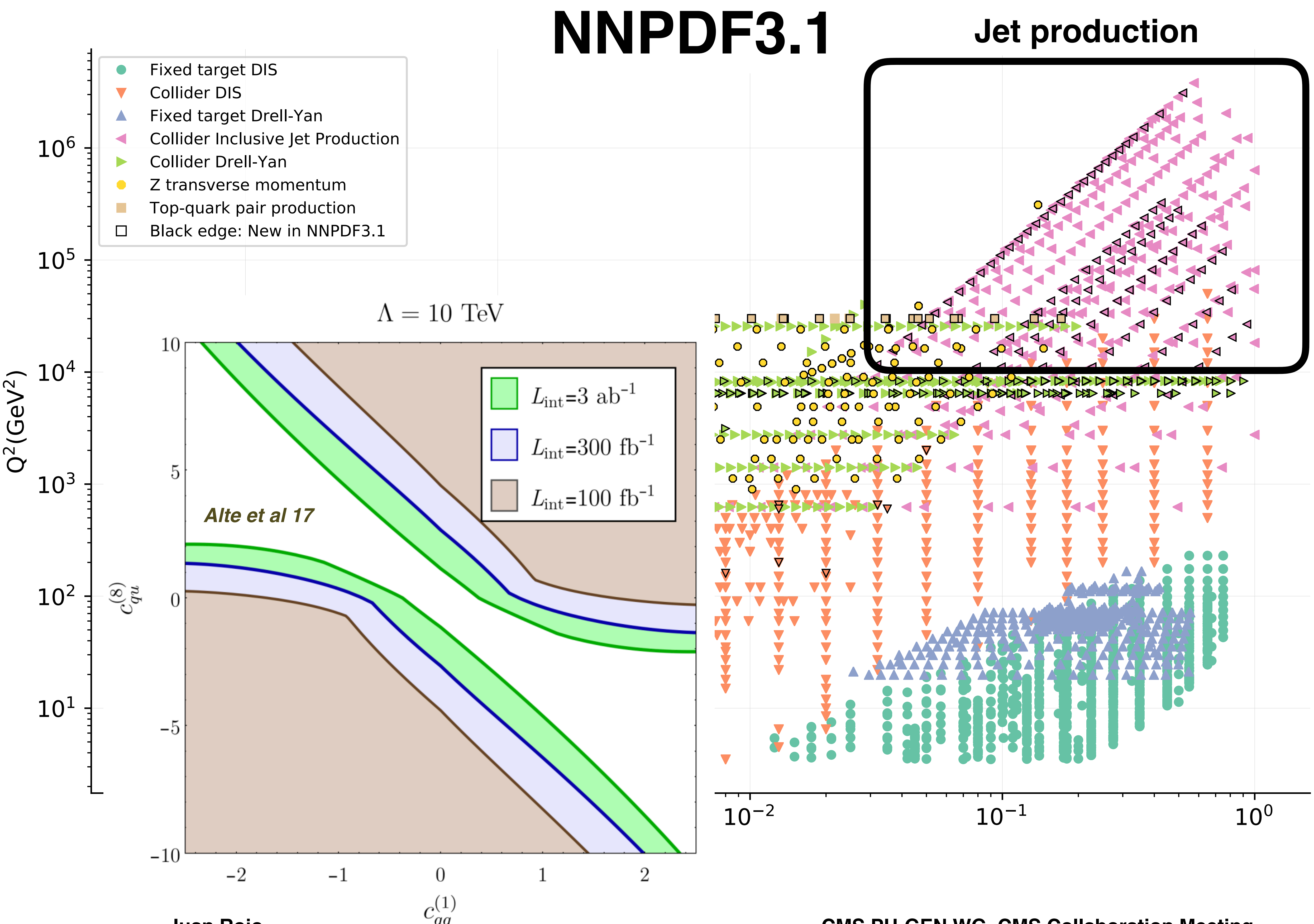

NNPDF3.1
Fixed target DIS
Collider DIS
Fixed target Drell-Yan
Collider Inclusive Jet Production
Collider Drell-Yan
Z transverse momentum
Top-quark pair production
Black edge: New in NNPDF3.1
$Q^2(\mathrm{GeV}^2)$
10^1
10^2
10^3
10^4
10^5
10^6
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^0
x

SMEFT & PDFs

NNPDF3.1

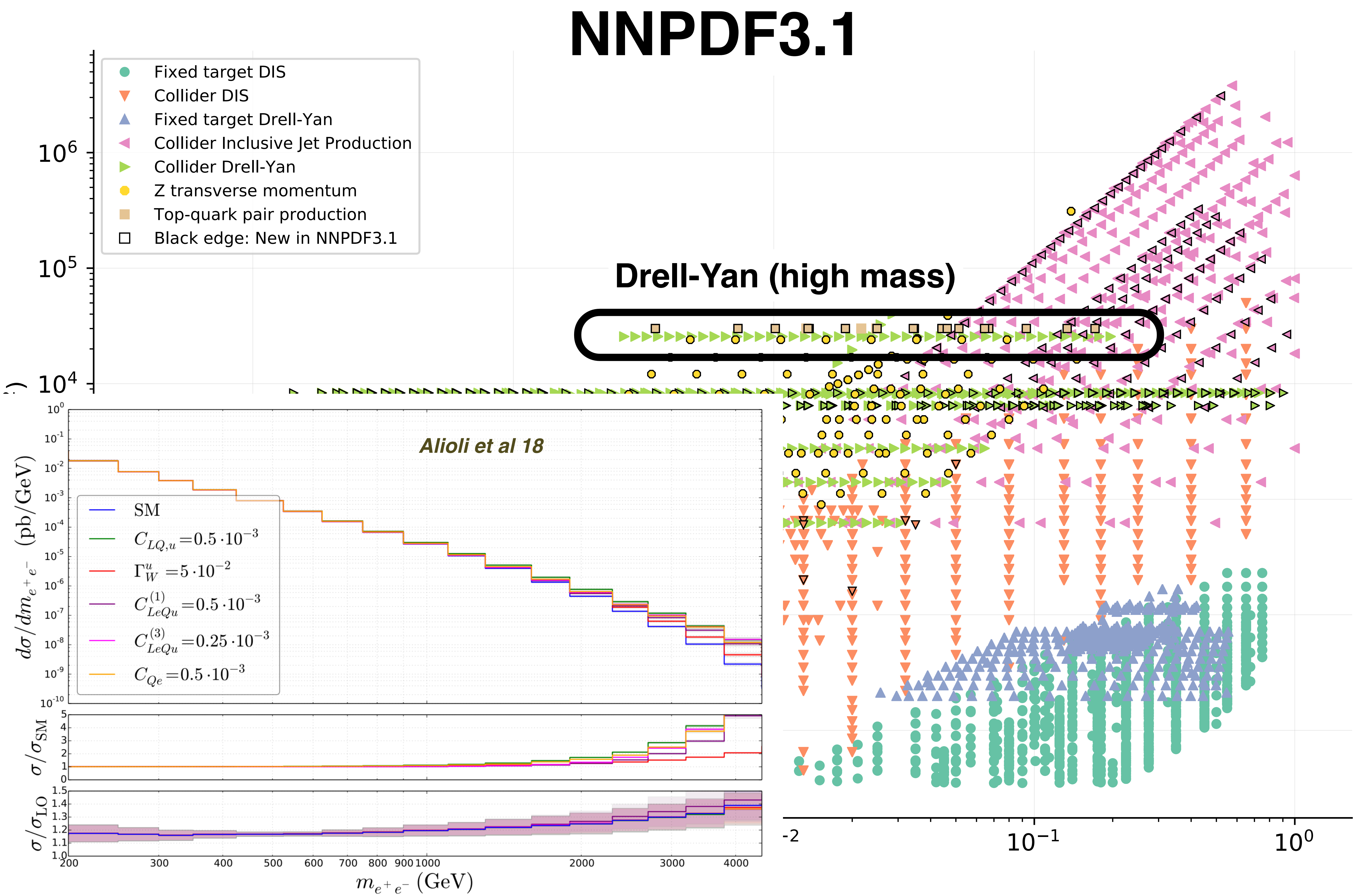


SMEFT & PDFs



NNPDF3.1
Jet production
Fixed target DIS
Collider DIS
Fixed target Drell-Yan
Collider Inclusive Jet Production
Collider Drell-Yan
Z transverse momentum
Top-quark pair production
Black edge: New in NNPDF3.1
$Q^2(\mathrm{GeV}^2)$
10^1
10^2
10^3
10^4
10^5
10^6
10^{-2}
10^{-1}
10^0
$\Lambda = 10$ TeV
Alte et al 17
L_{int}=3 ab^{-1}
L_{int}=300 fb^{-1}
L_{int}=100 fb^{-1}
$c_{qu}^{(8)}$
$c_{qq}^{(1)}$
10
5
0
−5
−10
−2
−1
0
1
2

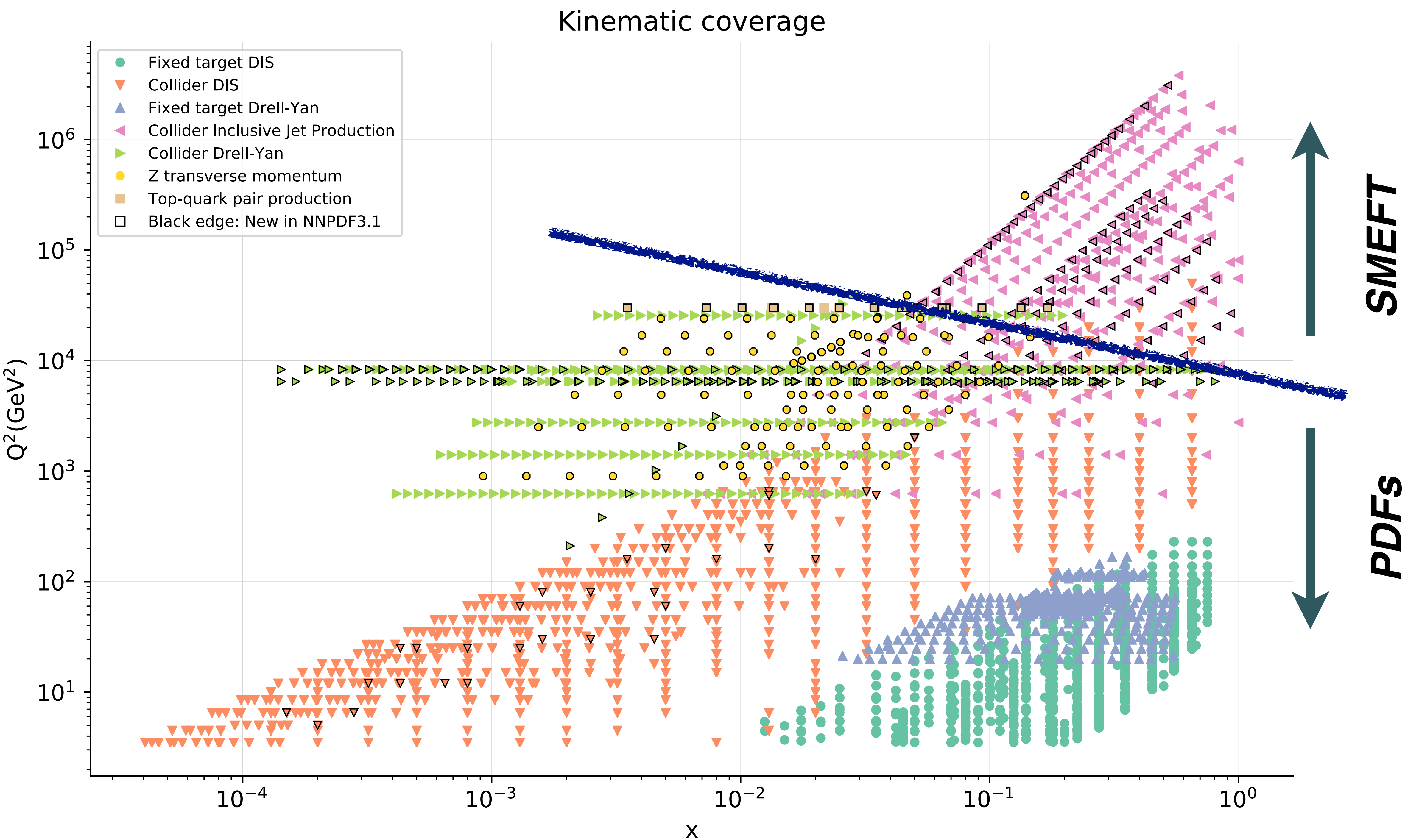
SMEFT & PDFs



NNPDF3.1
Fixed target DIS
Collider DIS
Fixed target Drell-Yan
Collider Inclusive Jet Production
Collider Drell-Yan
Z transverse momentum
Top-quark pair production
Black edge: New in NNPDF3.1
Drell-Yan (high mass)
Alioli et al 18
SM
$C_{LQ,u} = 0.5 \cdot 10^{-3}$
$\Gamma_W^u = 5 \cdot 10^{-2}$
$C_{LeQu}^{(1)} = 0.5 \cdot 10^{-3}$
$C_{LeQu}^{(3)} = 0.25 \cdot 10^{-3}$
$C_{Qe} = 0.5 \cdot 10^{-3}$
$d\sigma/dm_{e^+e^-}$ (pb/GeV)
σ/σ_{SM}
σ/σ_{LO}
$m_{e^+e^-}$ (GeV)

Naive approach

Separate LHC data into **input for PDF fits** and **input for SMEFT studies?**



Can we do better?

Simultaneous PDF+SMEFT fits

Our goal: constrain **simultaneously** both the PDFs and SMEFT degrees of freedom

Proof of concept: DIS-only fits where SM **augmented** by four *d=6* SMEFT operators

$$\mathscr{L}_{\rm SMEFT} = \mathscr{L}_{\rm SM} + \sum_{q=u,d,s,c} \frac{a_q}{\Lambda^2} \left(\bar{l}_R \gamma^\mu l_R\right) \left(\bar{q}_R \gamma_\mu q_R\right)$$

which can arise *e.g.* from a ***Z' boson*** with non-universal couplings to quarks

These SMEFT operators modify the DIS structure functions and thus **affect the PDF fit**

$$\Delta F_2^{\rm SMEFT} \supset \frac{x}{12 e^4} \left(4 a_u e^2 \frac{Q^2}{\Lambda^2} (1 + 4 K_Z \sin^4 \theta_W) + 3 a_u^2 \frac{Q^4}{\Lambda^4} \right) \left(u + \bar{u} \right)$$

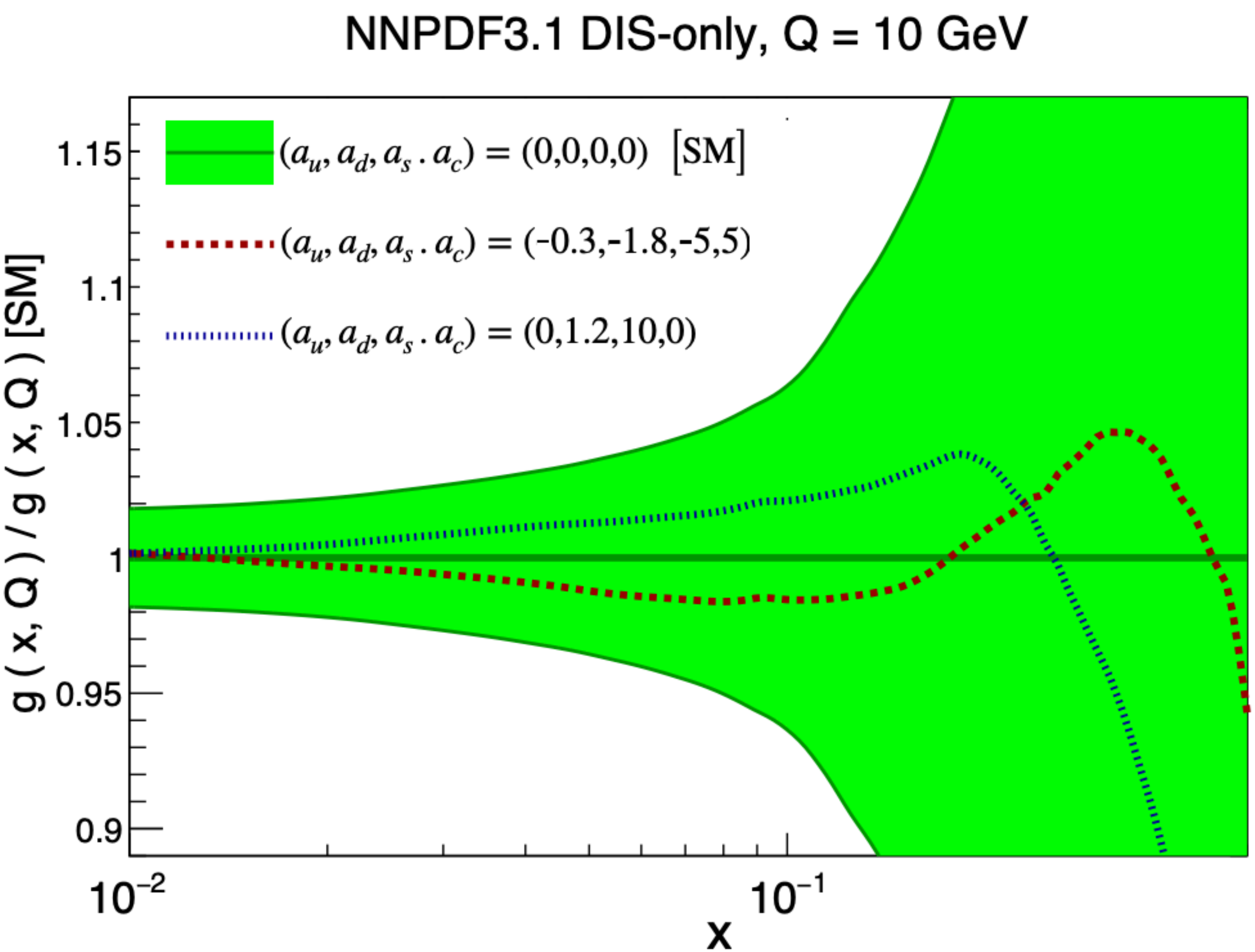
SMEFT effects enhanced by Q² : constrain from HERA data!

from interference with SM

from squared amplitude

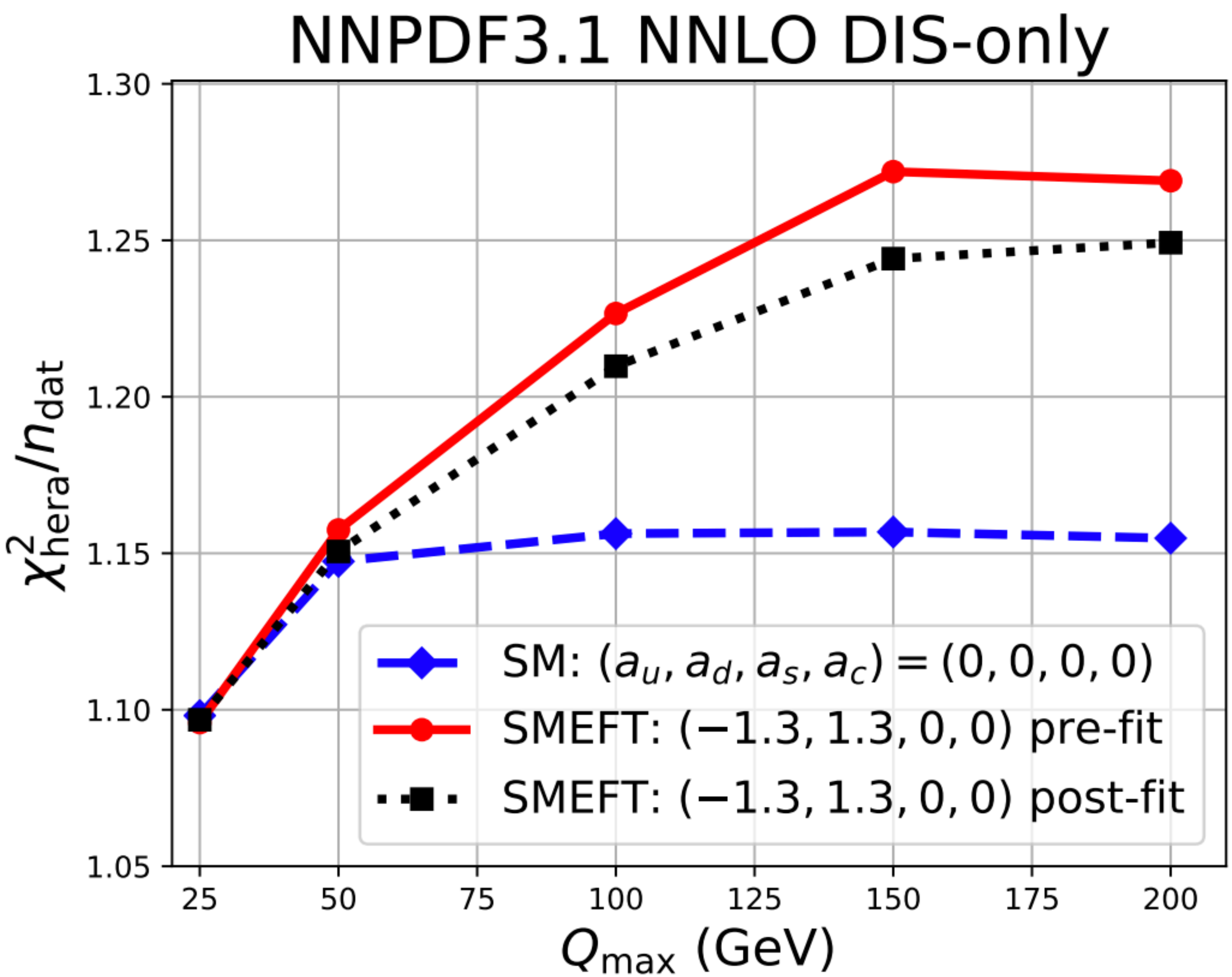
Impact on the PDFs

For a large region of the allowed parameter space,
SMEFT effects can be partially (but not completely) **reabsorbed into the PDFs**



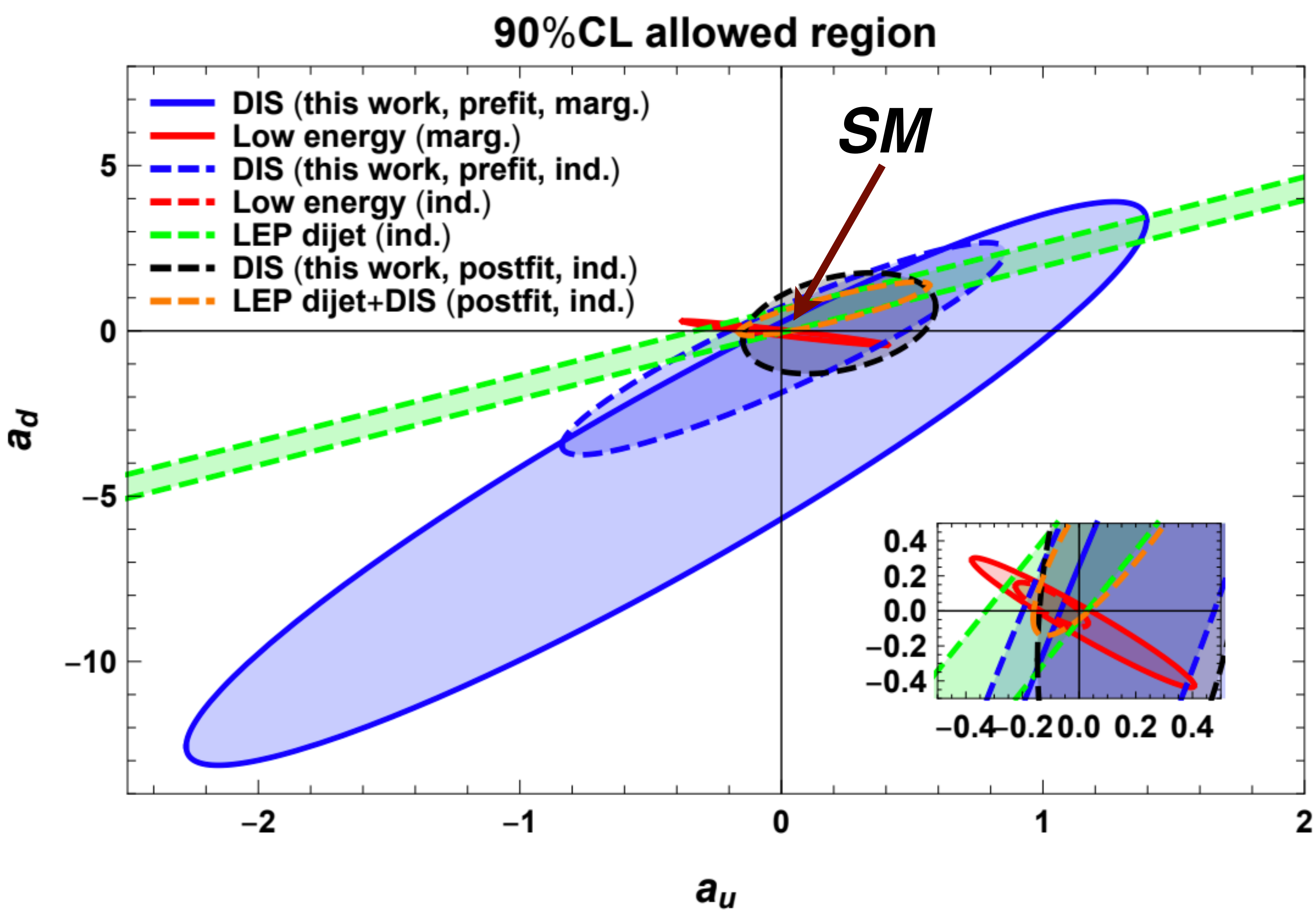
Fingerprinting BSM effects

Tell-tale sign of SMEFT effects: **rapid variation with *Q*** (DGLAP evolution slower)



Fingerprinting BSM effects

We can compare bounds on SMEFT degrees of freedom in the joint fit as compared to the usual approach where PDFs are kept fixed



Ultimate goal: **simultaneous PDF+SMEFT global analyses**

Nuclear NNPDF fits and the impact of LHC data

Based on R. Abdul-Khalek, J. Ethier, J. Rojo

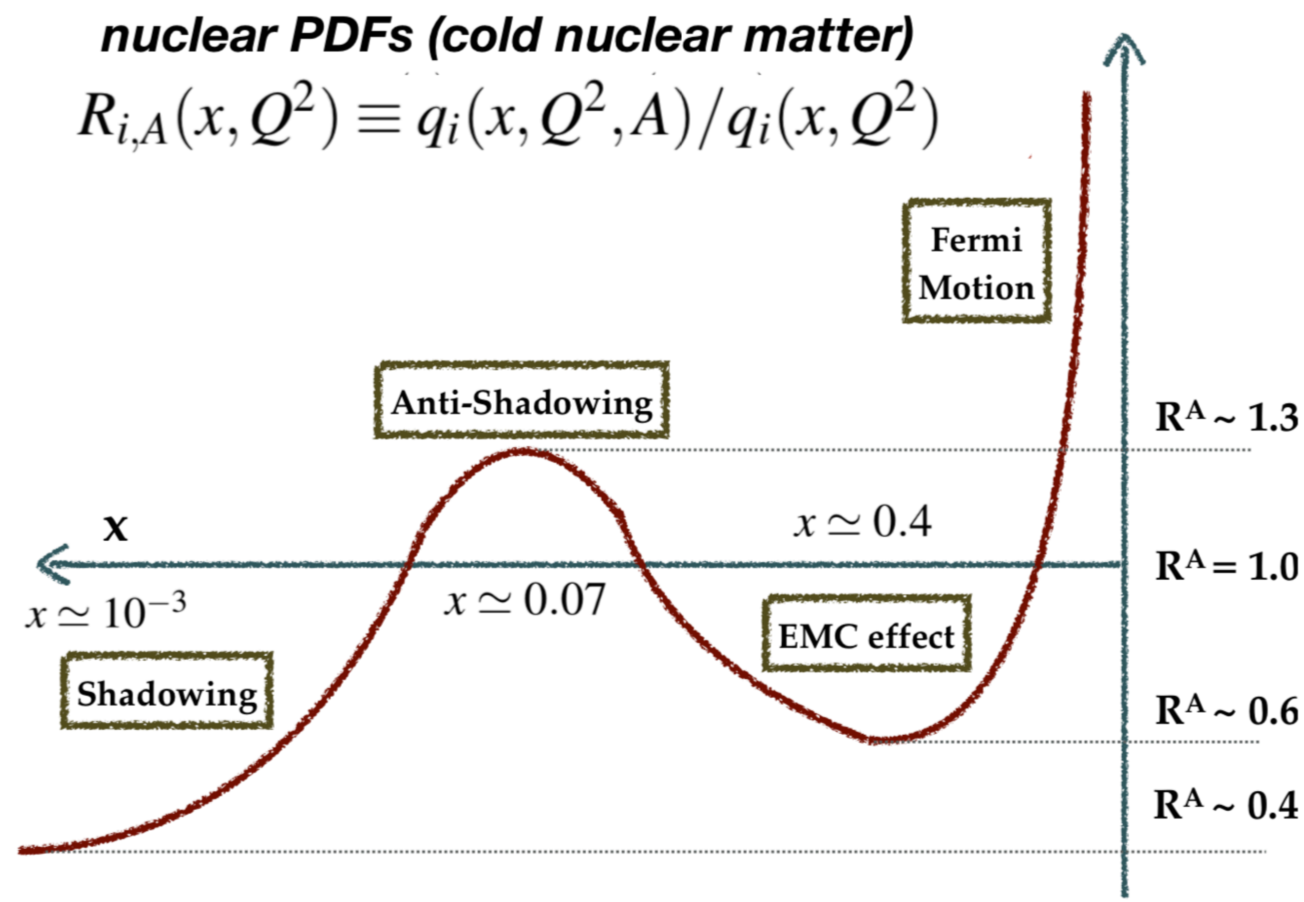
arXiv:1904.00018 *(EPJ) + work in progress*

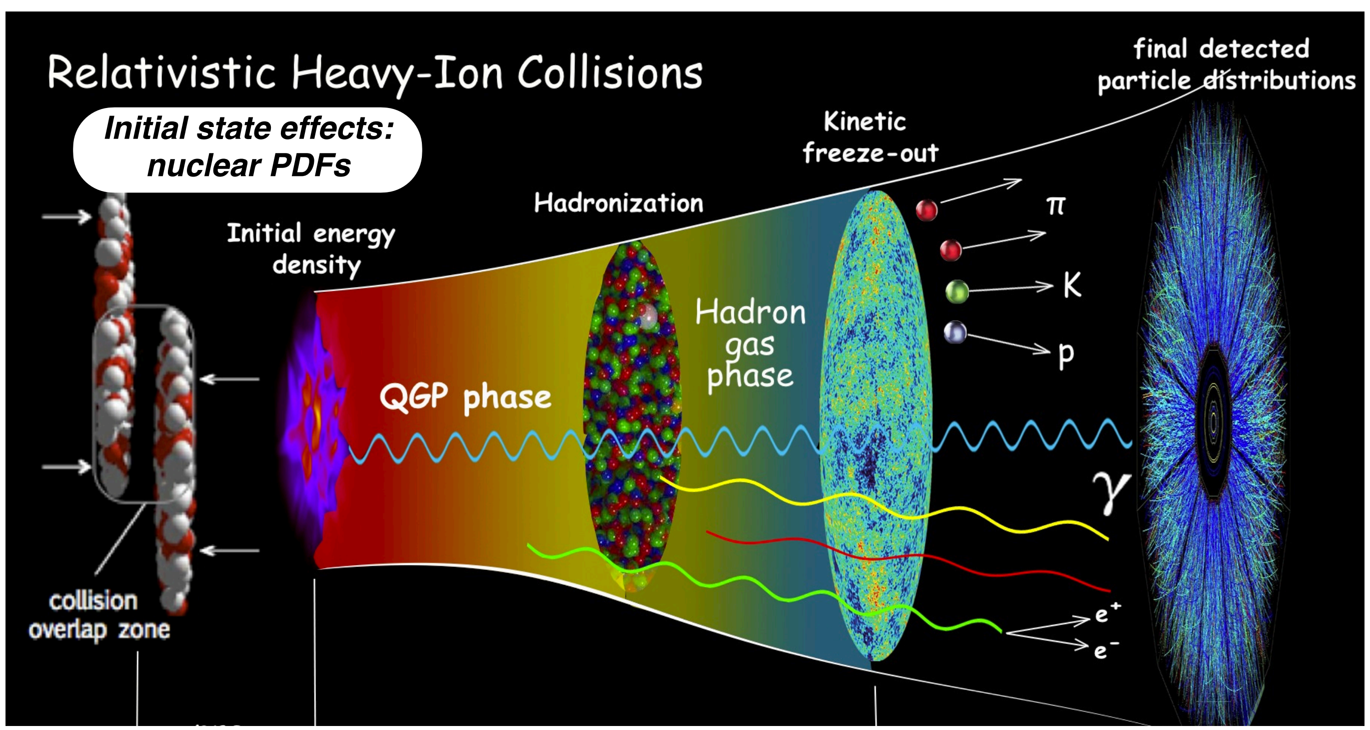
Why nuclear PDFs?

- Cold nuclear matter effects modify the **PDFs of bound nucleons** as compared to the free-proton case

- Rich connection with **nuclear calculations**: EMC effect, shadowing

- Non-linear gluon interactions enhanced in heavy nuclei: **Color Glass Condensate**?

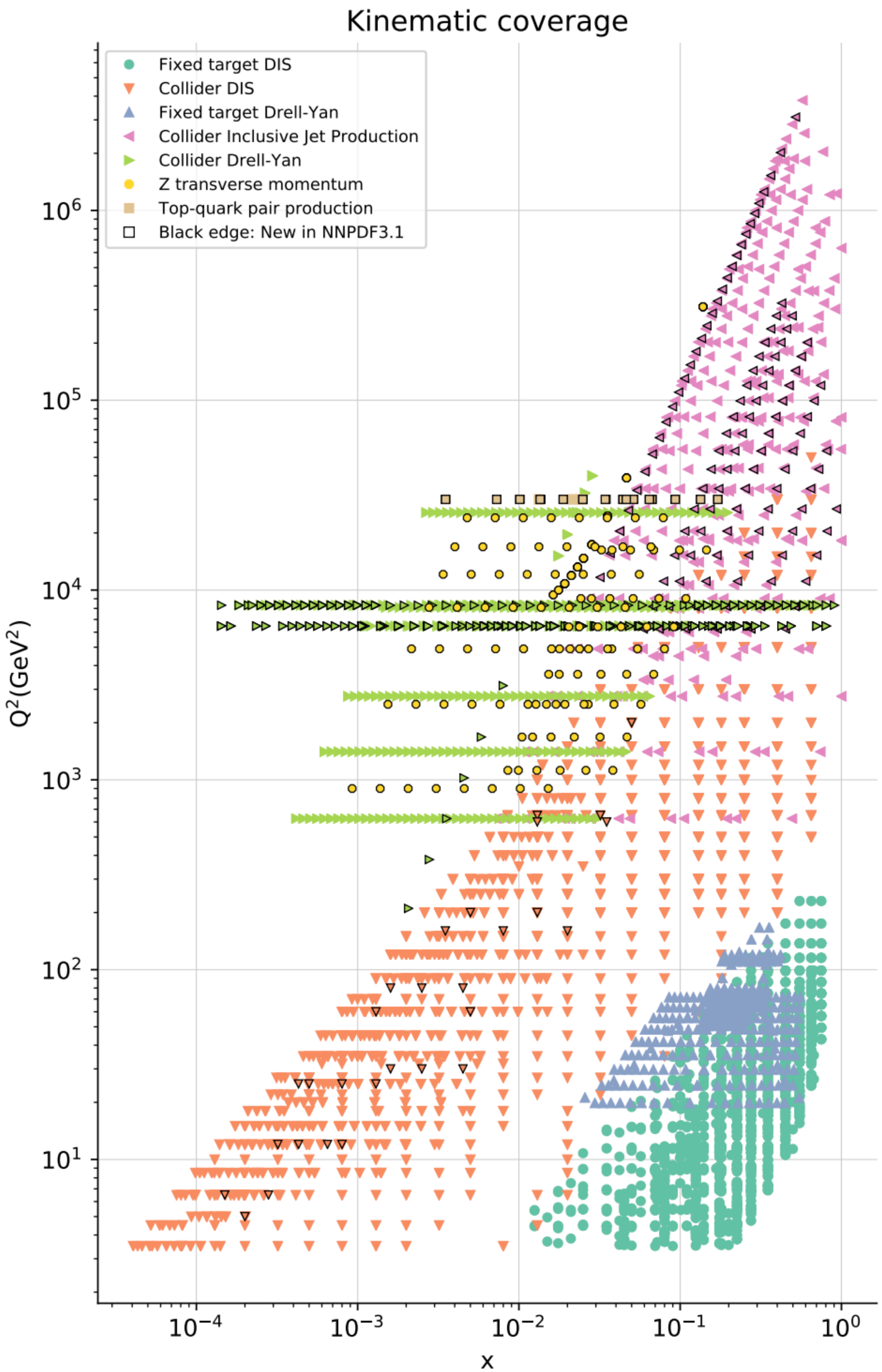




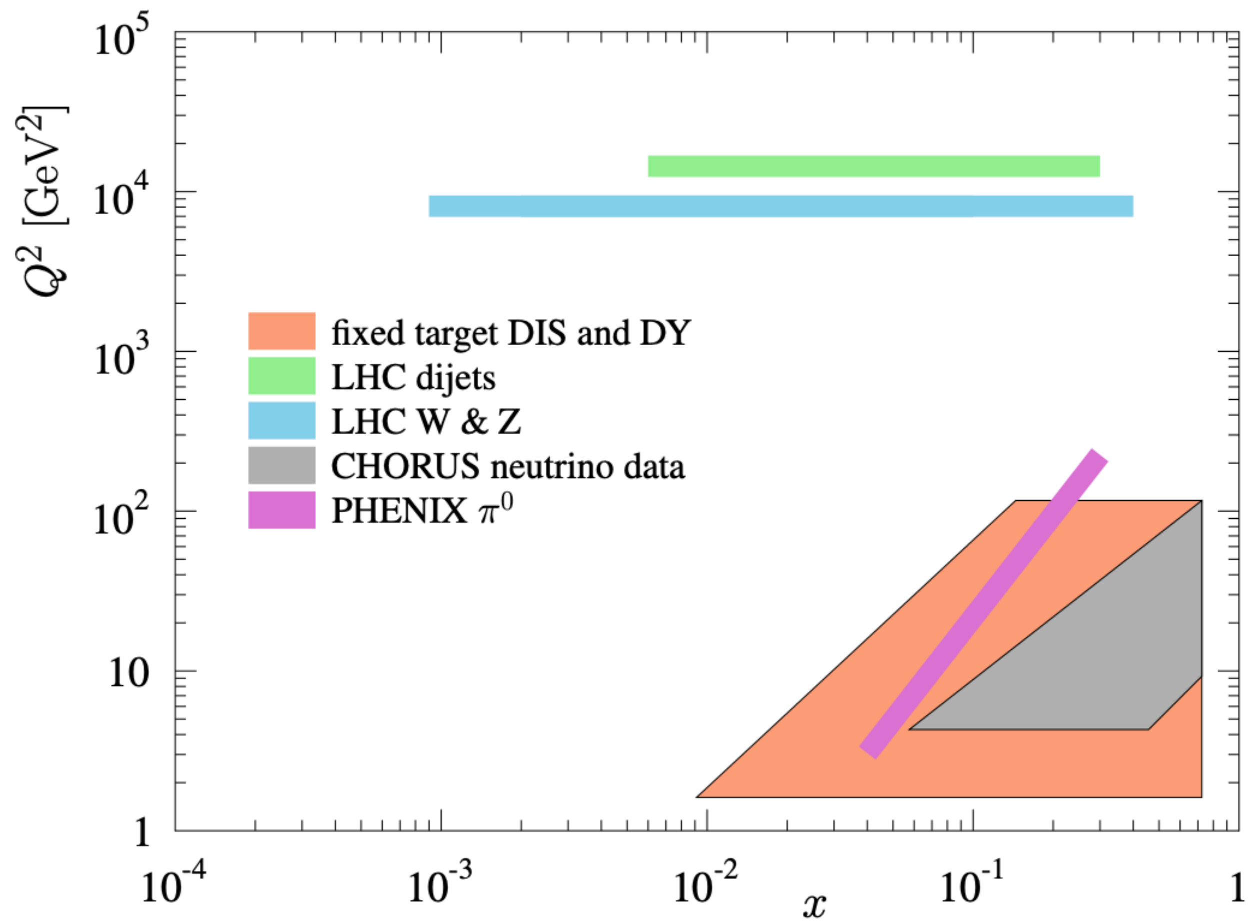
- nPDFs relevant for the **initial state of heavy-ion collisions**: benchmarks for Quark-Gluon Plasma characterisation

- nPDFs also required for **ultra-high-energy astrophysics** e.g. neutrino telescopes such as IceCube

From protons to heavy nuclei



- Nuclear dataset << proton dataset
- Limited info on **nuclear gluon and quark sea**, few constraints for $x < 10^{-2}$
- Recently: info from **p+Pb collisions** at the LHC provides novel opportunities to pin down nPDFs



Nuclear NNPDF fits

- Parametrize nPDFs with ANNs with ***x, ln(x), A*** as input: **fully model-independent**

$$q_i(x, Q_0, A) = B_i x^{-\alpha_i}(1-x)^{\beta_i}\mathrm{NN}(x, A)\,, \quad i = g, \Sigma, T_8$$

- Gluon normalisation (*A*-dependent) fixed by the **momentum sum rule**

$$B_g(A) = \left(1 - \int_0^1 dx\, x\Sigma(x, Q_0, A)\right) \Big/ \int_0^1 dx\, xg(x, Q_0, A)$$

- **Proton boundary condition** implemented as a **penalty** in the figure of merit

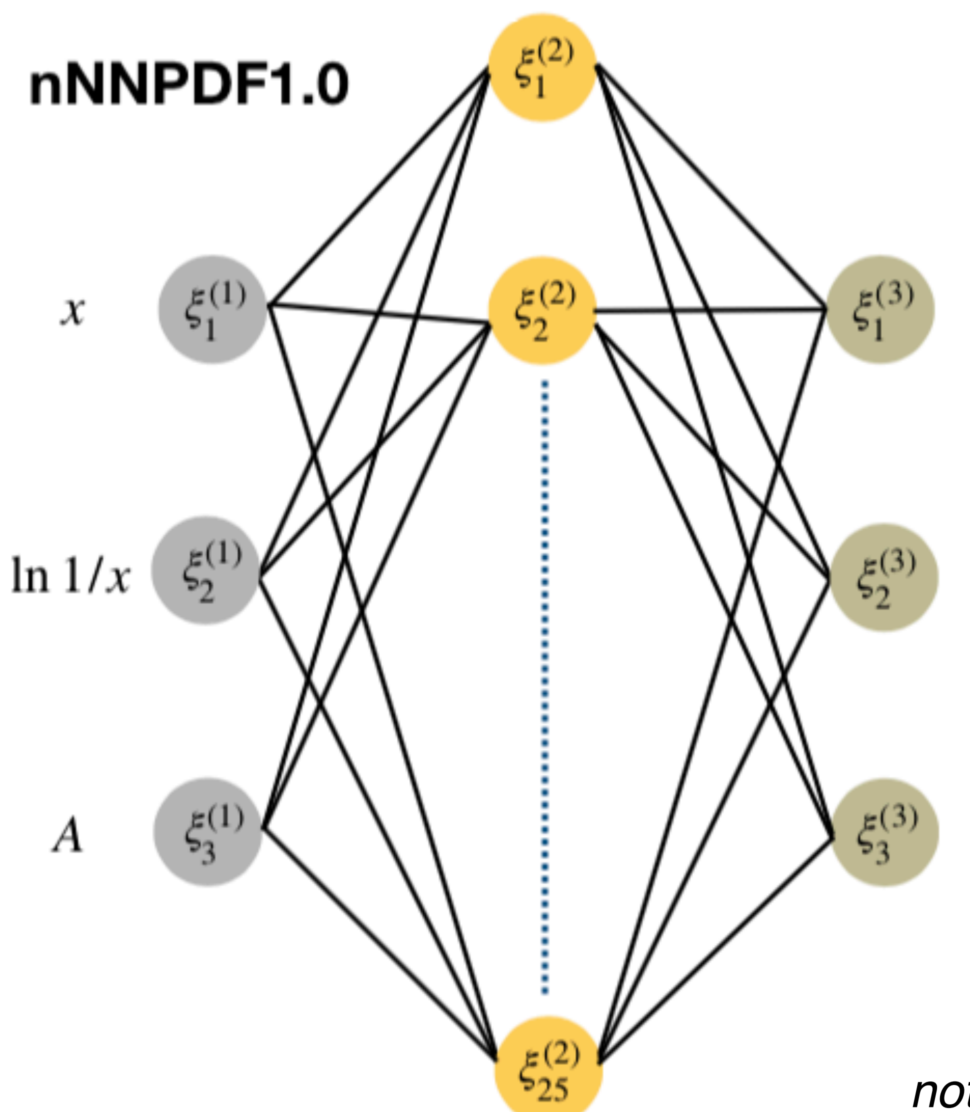
$$\chi^2 = \sum_{j=1}^{n_{\rm dat}} \frac{\left(F_j^{(\rm exp)} - F_j^{(\rm th)}\right)^2}{\sigma_j^{(\rm exp)2}} + \lambda \sum_{i=g,\Sigma,T_8} \sum_{k=1}^{n_x} \frac{\left(q_i(x_k, Q_0, A) - q_i^{(\rm ref)}(x_k, Q_0, A=1)\right)^2}{\left(\delta q_i^{(\rm ref)}(x_k, Q_0, A=1)\right)^2}$$

$q_i^{(\rm ref)}(x_k, Q_0, A=1)$ Isoscalar **NNPDF3.1** NNLO global fit

Nuclear NNPDF fits

- Parametrize nPDFs with ANNs with ***x, ln(x), A*** as input: **fully model-independent**

$$q_i(x, Q_0, A) = B_i x^{-\alpha_i}(1-x)^{\beta_i}\mathrm{NN}(x, A)\,, \quad i = g, \Sigma, T_8$$



$$g(x, Q_0, A) = B_g x^{-\alpha_g}(1-x)^{\beta_g}\,\xi_1^{(3)}$$

$$\Sigma(x, Q_0, A) = x^{-\alpha_\Sigma}(1-x)^{\beta_\Sigma}\,\xi_2^{(3)}$$

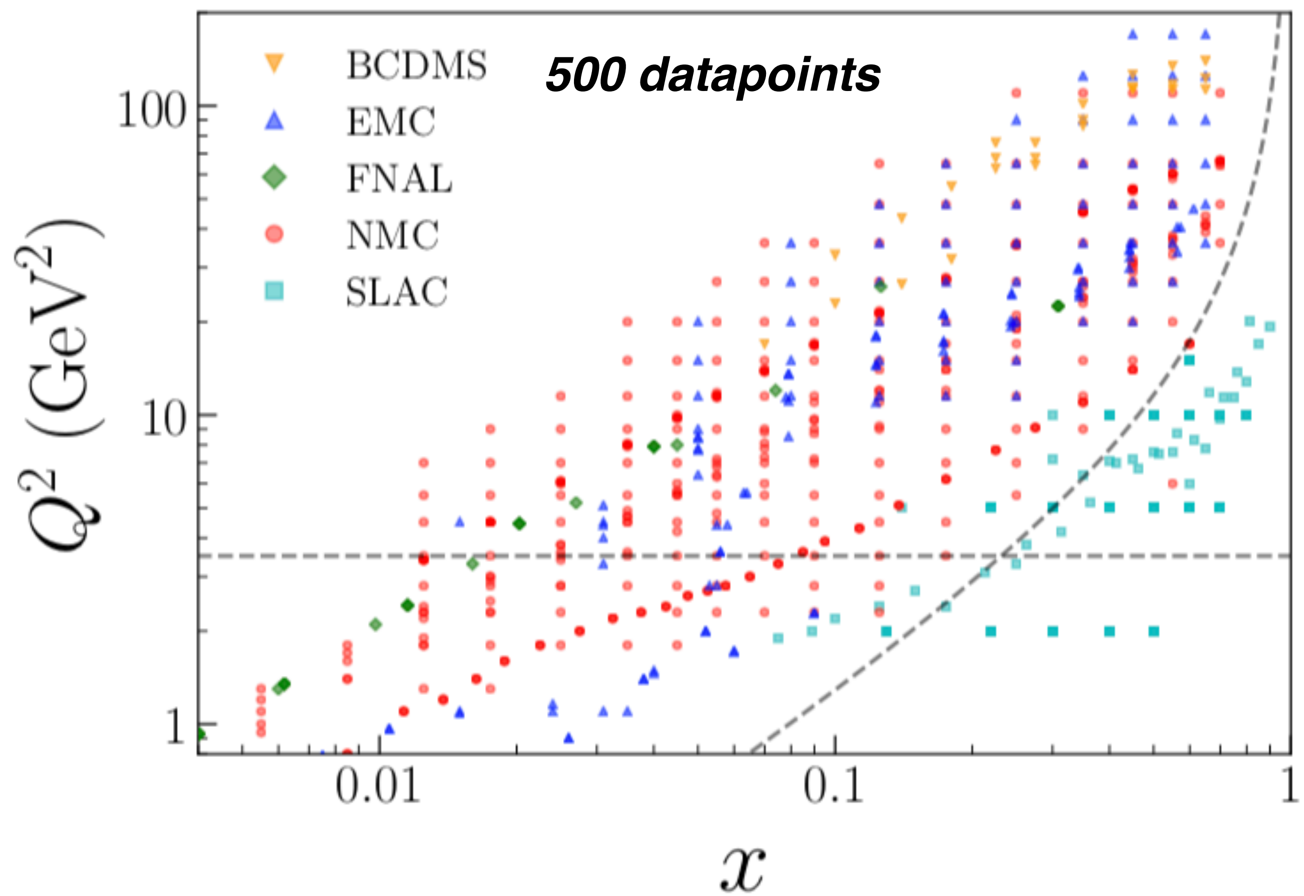
$$T_8(x, Q_0, A) = x^{-\alpha_{T_8}}(1-x)^{\beta_{T_8}}\,\xi_3^{(3)}$$

note: no assumption on A-dependence, purely data-driven fit

- Optimisation of NN parameters using ADAM Stochastic Gradient Descent with **TensorFlow**
- Fitting methodology validated with **closure tests**
- Flexibility to vary flavour assumptions, proton boundary condition etc

nNNPDF1.0

- Fit to fixed-target neutral-current **nuclear DIS structure functions**

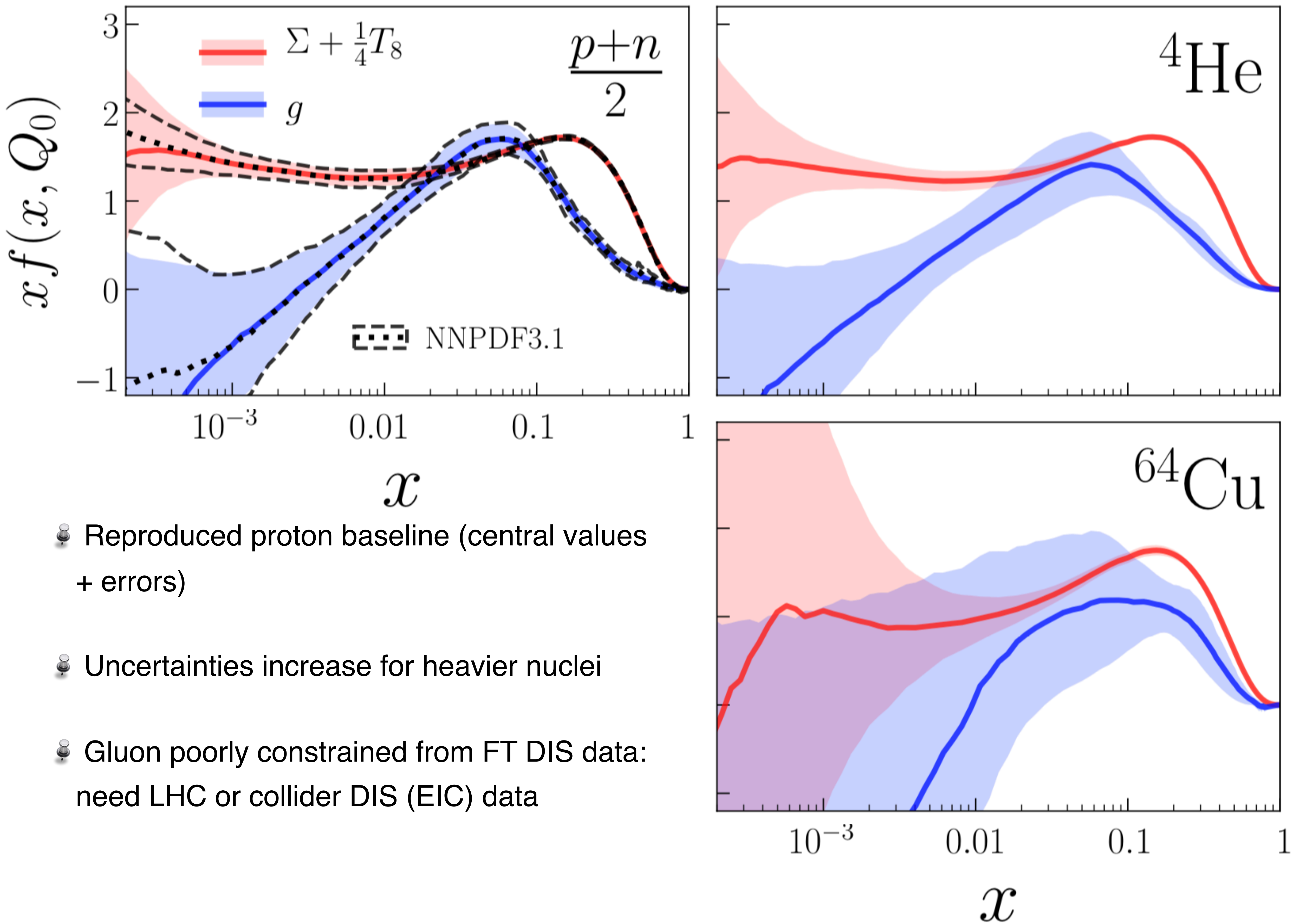


- Only the gluon and one quark combination can be extracted from this dataset

$$\Sigma(x, Q_0, A) + \frac{1}{4} T_8(x, Q_0, A)$$

$$T_8 = u^+ + d^+ - 2s^+ \qquad T_3 = u^+ - d^+ = 0$$ *(isoscalarity)*

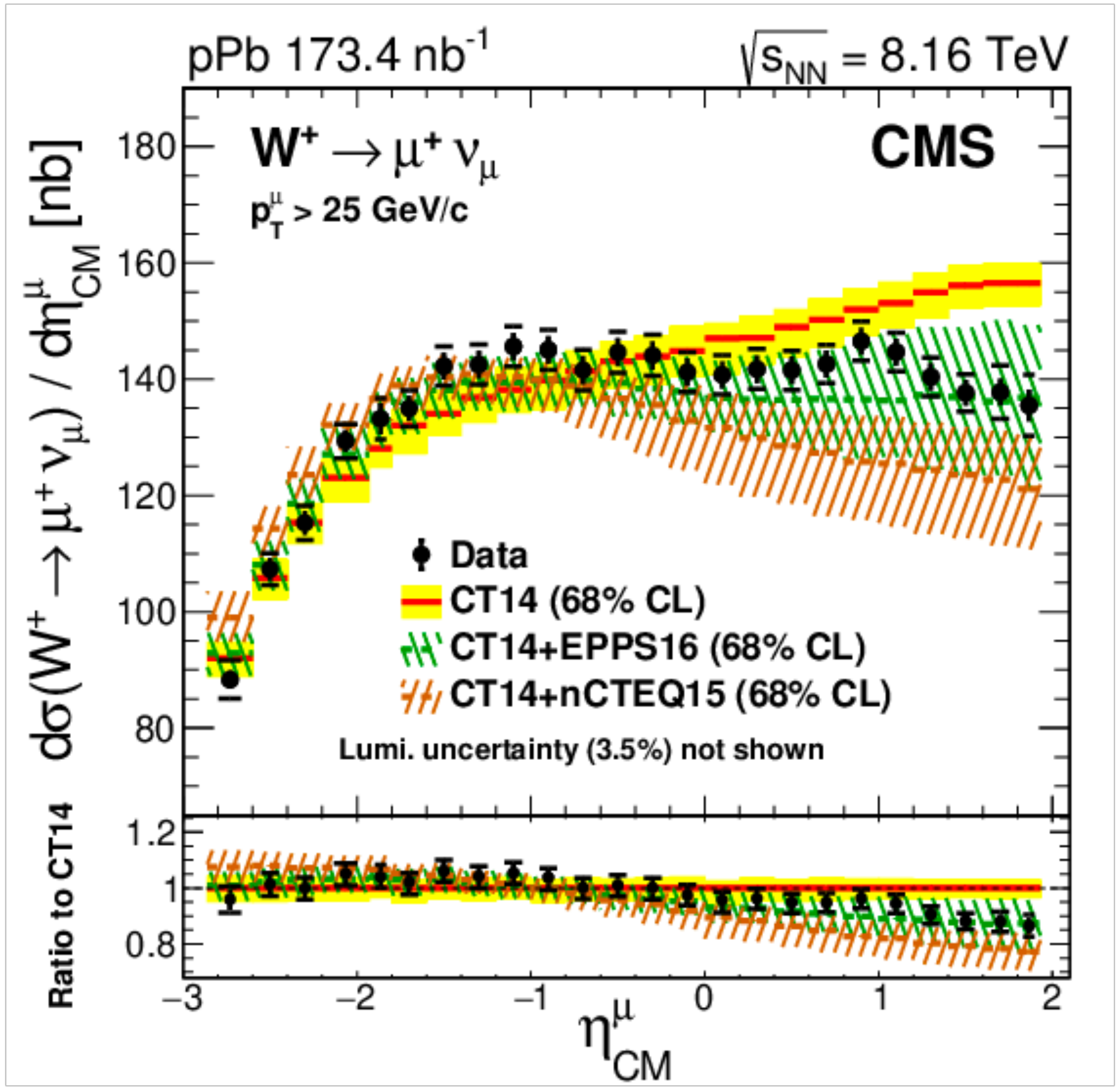
nNNPDF1.0

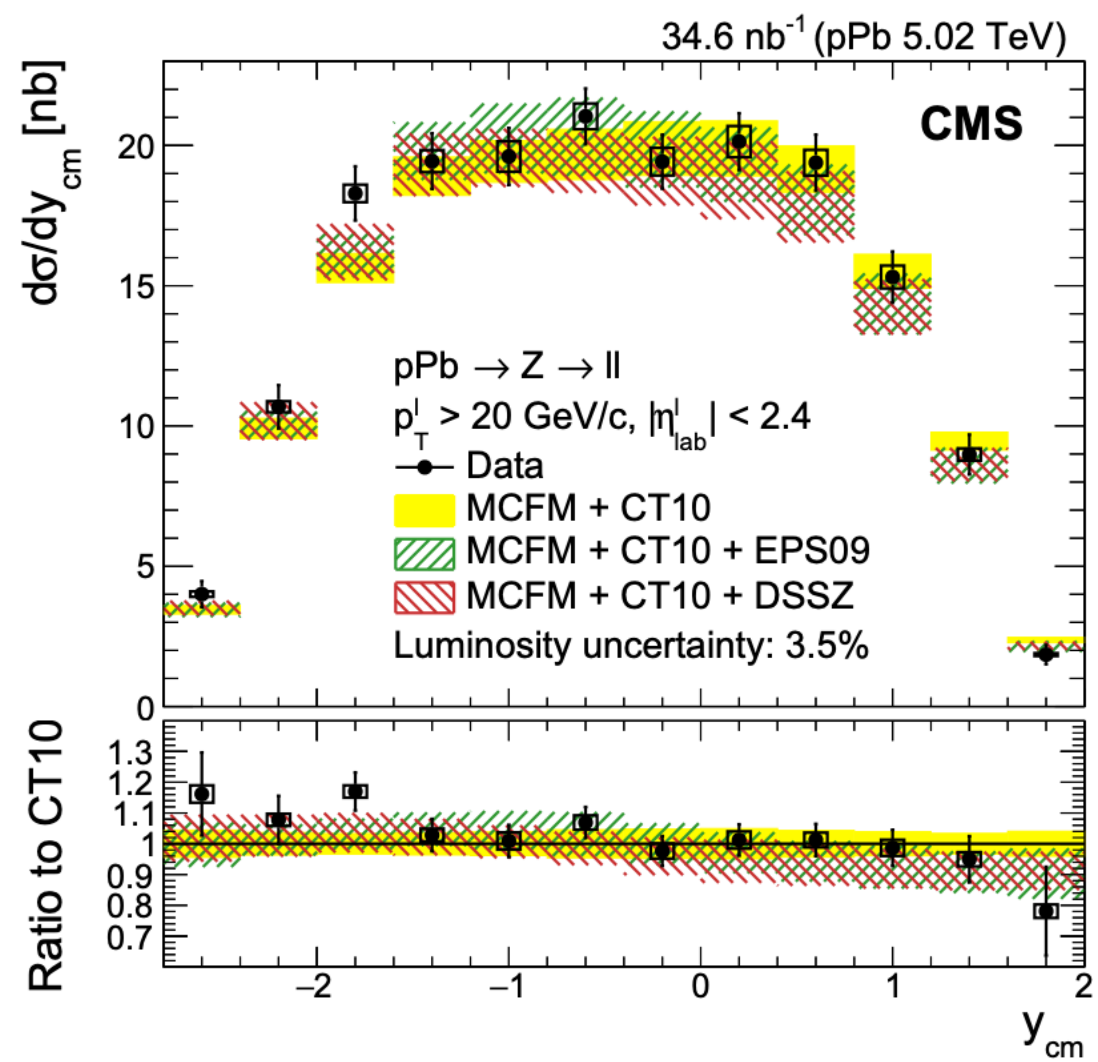


- Reproduced proton baseline (central values + errors)
- Uncertainties increase for heavier nuclei
- Gluon poorly constrained from FT DIS data: need LHC or collider DIS (EIC) data

Towards nNNPDF2.0

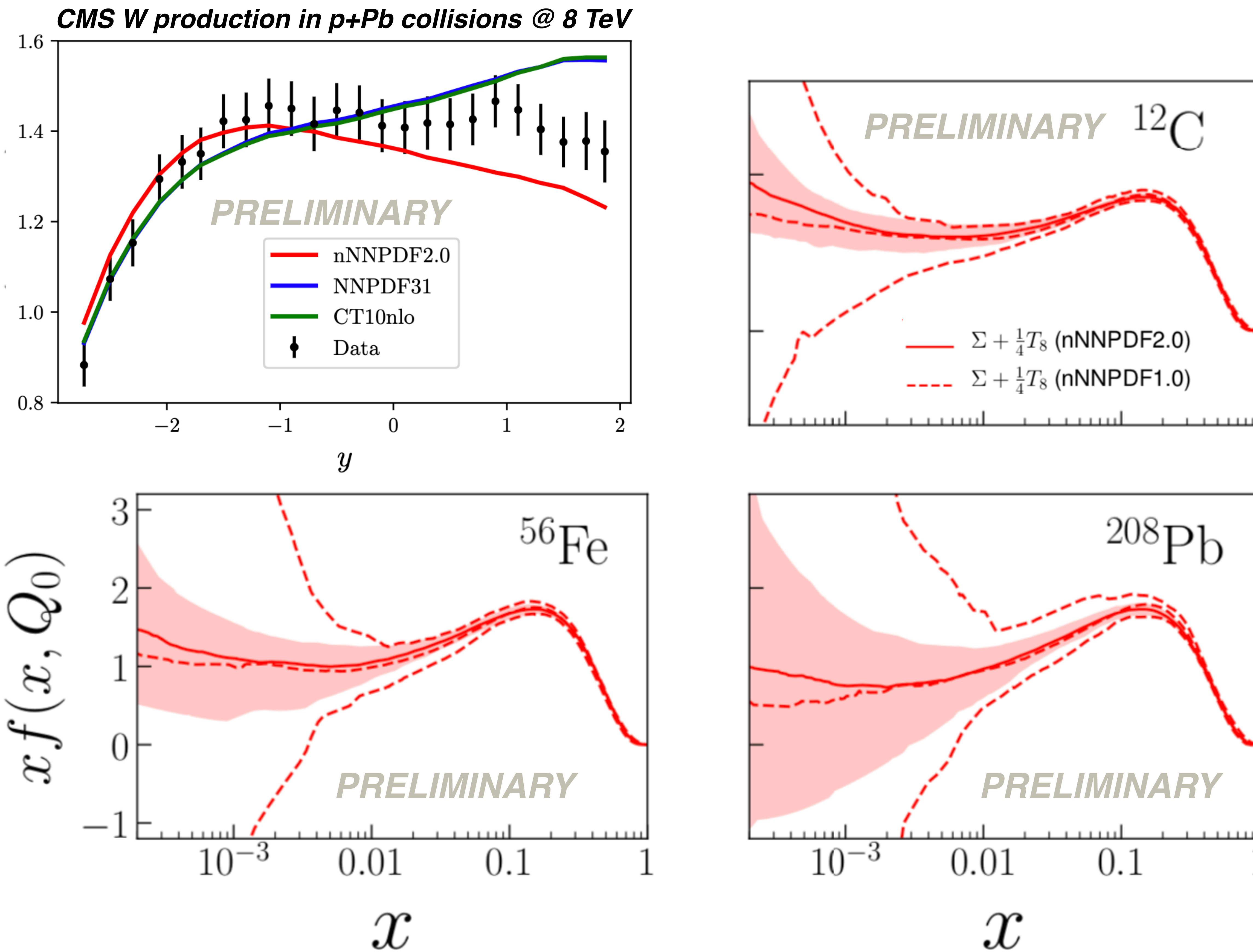
- Inclusion of **neutrino CC nuclear structure function** data and several LHC measurements of hard-probes in **proton-lead collisions**





- General quark-flavour decomposition: assess **flavour dependence of nuclear effects**
- Study also role that non-isoscalar effects play
- Phenomenological implications for the **LHC heavy ion program**

Impact of LHC p+Pb data



Summary and outlook

The accurate determination of the **quark and gluon structure of the proton** is an essential ingredient for **LHC phenomenology** and **beyond**

- Working towards a next major release, **NNPDF4.0**, with significant improvements from the theoretical, data, and methodological aspects

- **MHOUs** (theory errors) can now be systematically included within a global PDF analysis

- The robust interpretation of high-energy measurements from the LHC benefits from **simultaneous extraction of PDFs and BSM effects**, *e.g.* within the SMEFT framework

- NNPDF methodology used to produce **nuclear PDF fits** constrained by LHC p+Pb measurements and accounting from the information provided by the proton baseline fits