MARIA UBIALI
UNIVERSITY OF CAMBRIDGE

NEURAL NET ENSEMBLES FOR BAYESIAN INFERENCE OF PDFS

ARTIFICIAL INTELLIGENCE FOR THE ELECTRON-ION COLLIDER (AI4EIC) 2025

28TH OCTOBER 2025

European Research Council
Established by the European Commission

OUTLINE

➡ Part I

- Motivation: Parton Distribution Functions (PDFs) for precision physics and new physics searches
- Statistical formulation of the inverse problem of determining PDFs from a finite number of collider data

➡ **Part II**

(in collaboration with M. Costantini, L. Mantani, J. Moore, V. Schütze Sanchez)

- A linear model from a NN ensemble for robust Bayesian inference of PDFs
  (based on arXiv: 2507.16913)
- Colibri: a flexible general tool for PDF and parameter Bayesian inference
  (based on arXiv: 2510.03391)

PART I

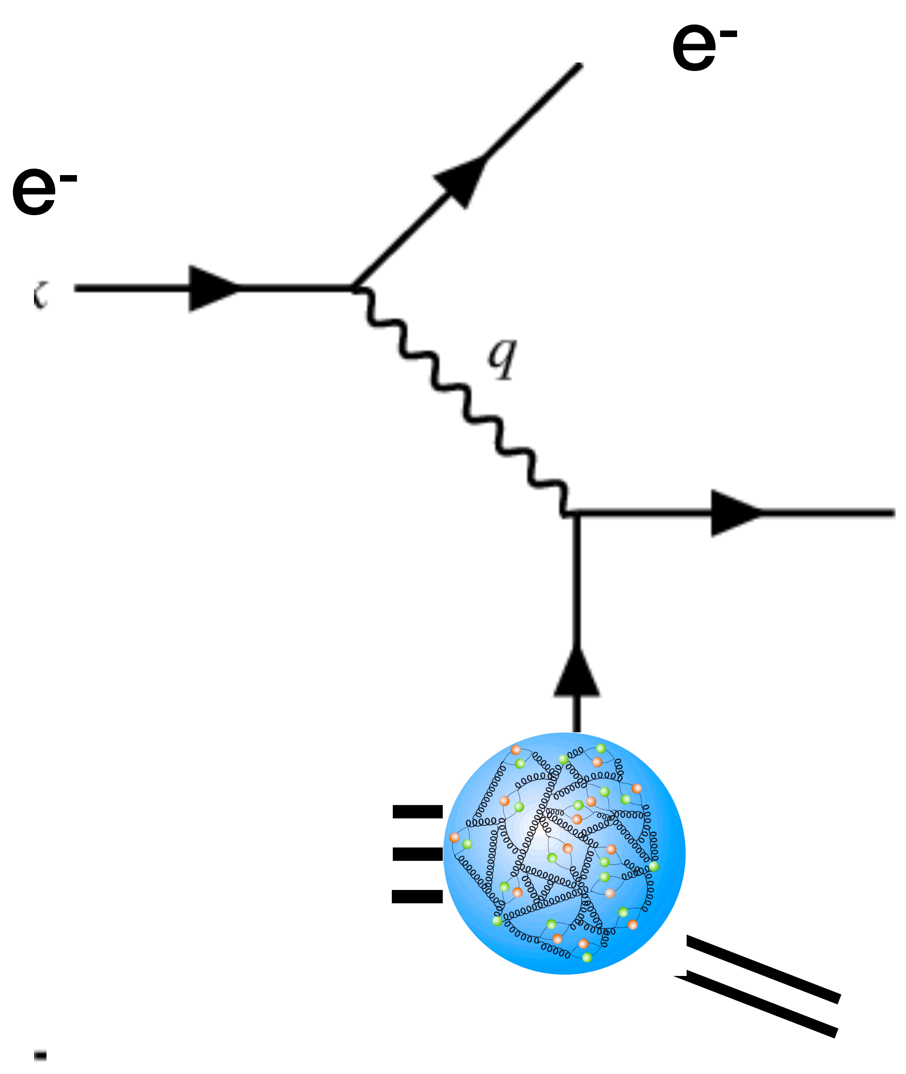
COLLINEAR FACTORIZATION AND THEORY PREDICTIONS

- Longitudinal unpolarised PDFs are **universal,** partonic momentum distribution is independent of the process that partons undergo (DIS, DY, H production..)
- PDF **evolution with the factorisation** scale predicted by perturbative QCD up to approximate N3LO

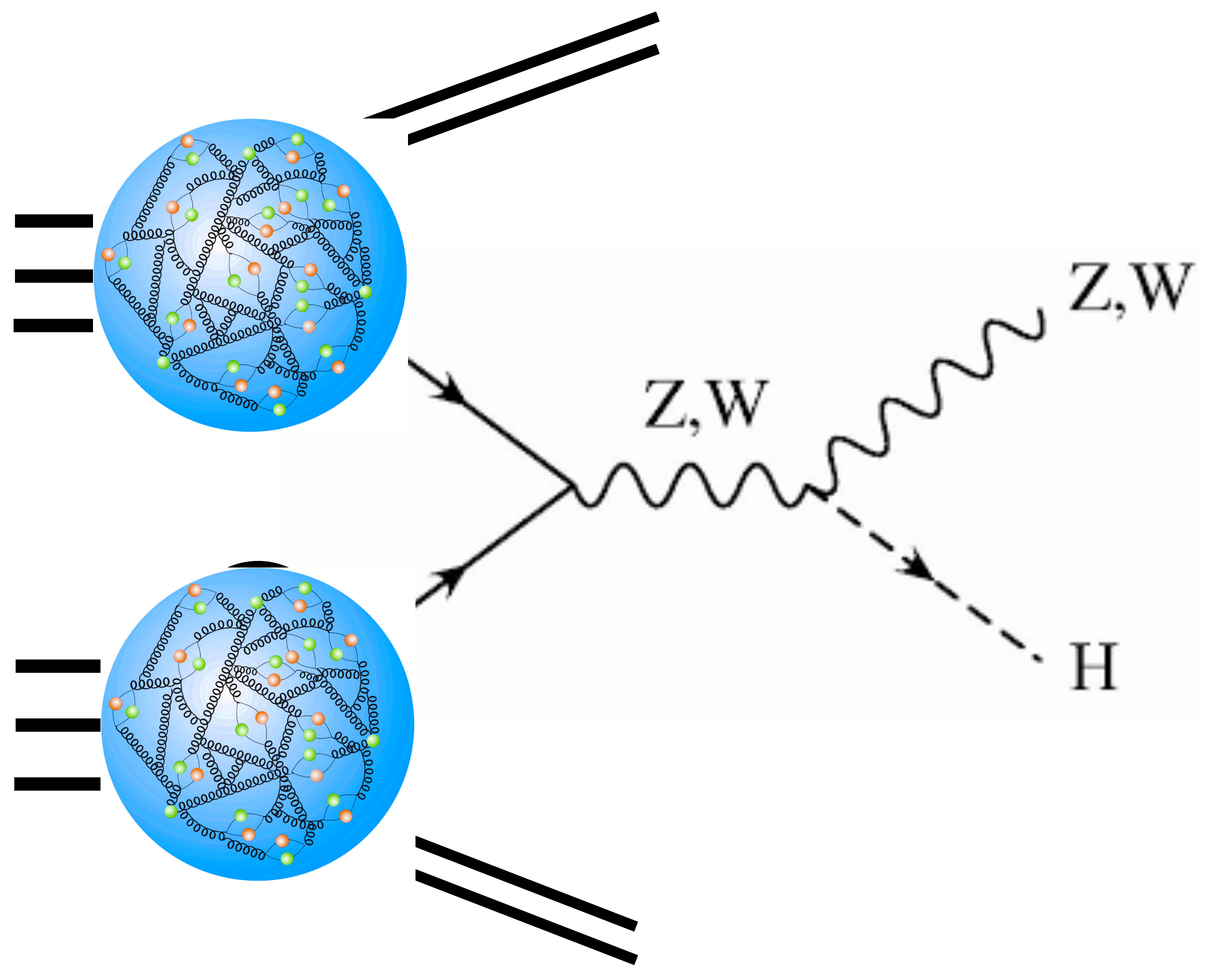
$$\sigma_{p,e} = \sum_i \boxed{f_i(x,\mu_F)} \otimes \boxed{\hat{\sigma}_{i,e}(xP,\mu_F,\alpha_s(\mu_R))}$$

$$\sigma_{p,p} = \sum_{i,j} \boxed{f_i(x_1,\mu_F)} \otimes \boxed{f_j(x_2,\mu_F)} \otimes \boxed{\hat{\sigma}_{i,j}(x_1P_1,x_2P_2,\mu_F,\alpha_s(\mu_R))}$$

Partonic Cross Sections
calculable using perturbative QED and QCD



EIC: Deep Inelastic Scattering of electrons off protons/nucleons

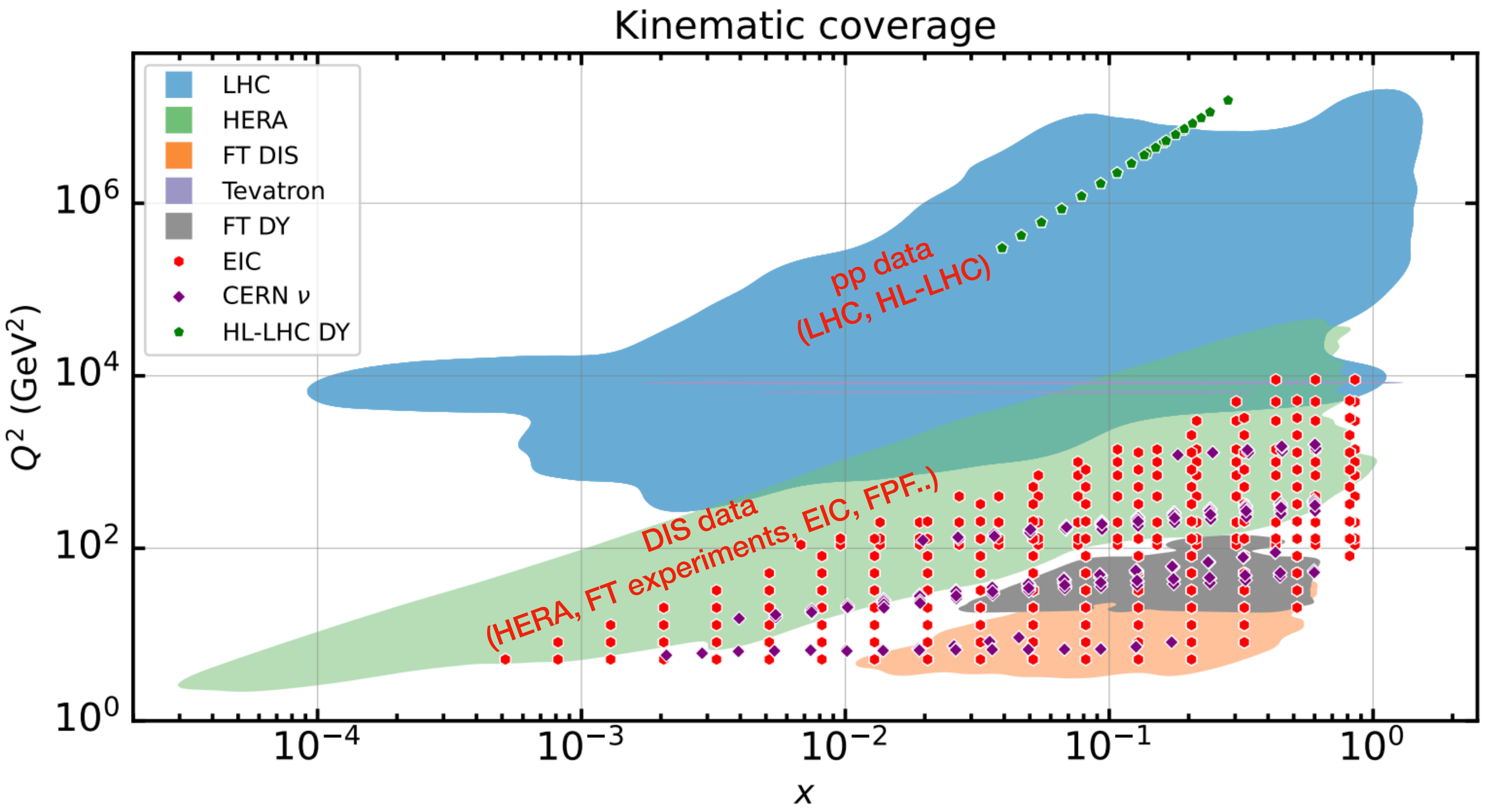


LHC: Proton-Proton scattering producing Higgs and vector boson W/Z

A WEALTH OF INPUT INFORMATION

$$f_i(x, \mu)$$

Data

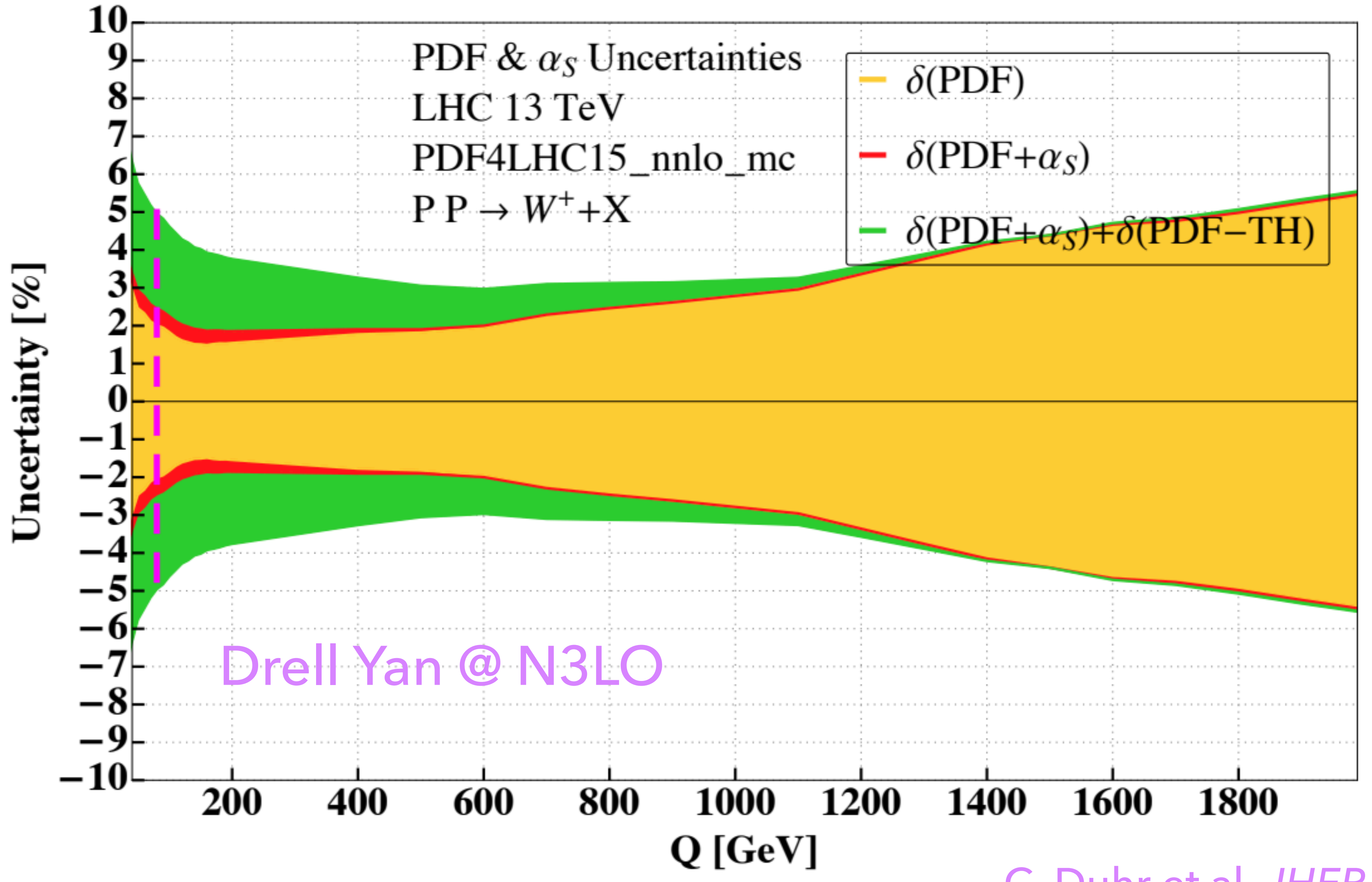


- Different datapoints constrain different PDF combinations convoluted with different partonic cross sections (forward map) in different x regions
- We want to determine 8 different functions and their uncertainty (which reflects to a first approximation data uncertainty)

Hammou, MU, 2410.00963

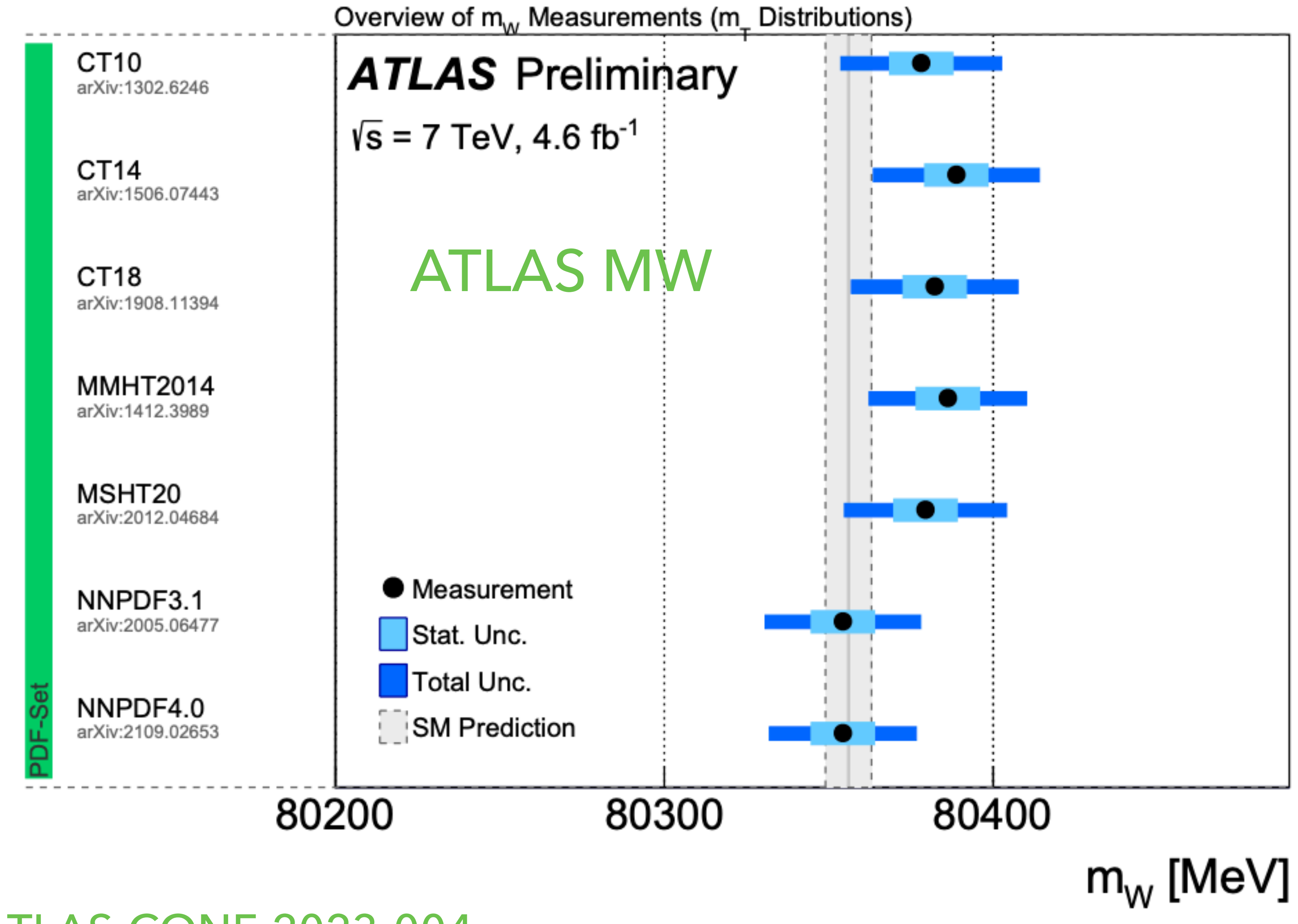
PDFS ARE A CRUCIAL INPUT FOR PRECISION PHYSICS…

#1: Theory uncertainty of SM predictions



C. Duhr et al, *JHEP* 11 (2020) 143

#2: Determination of SM parameters



ATLAS-CONF-2023-004

. . . AND NEW PHYSICS SEARCHES

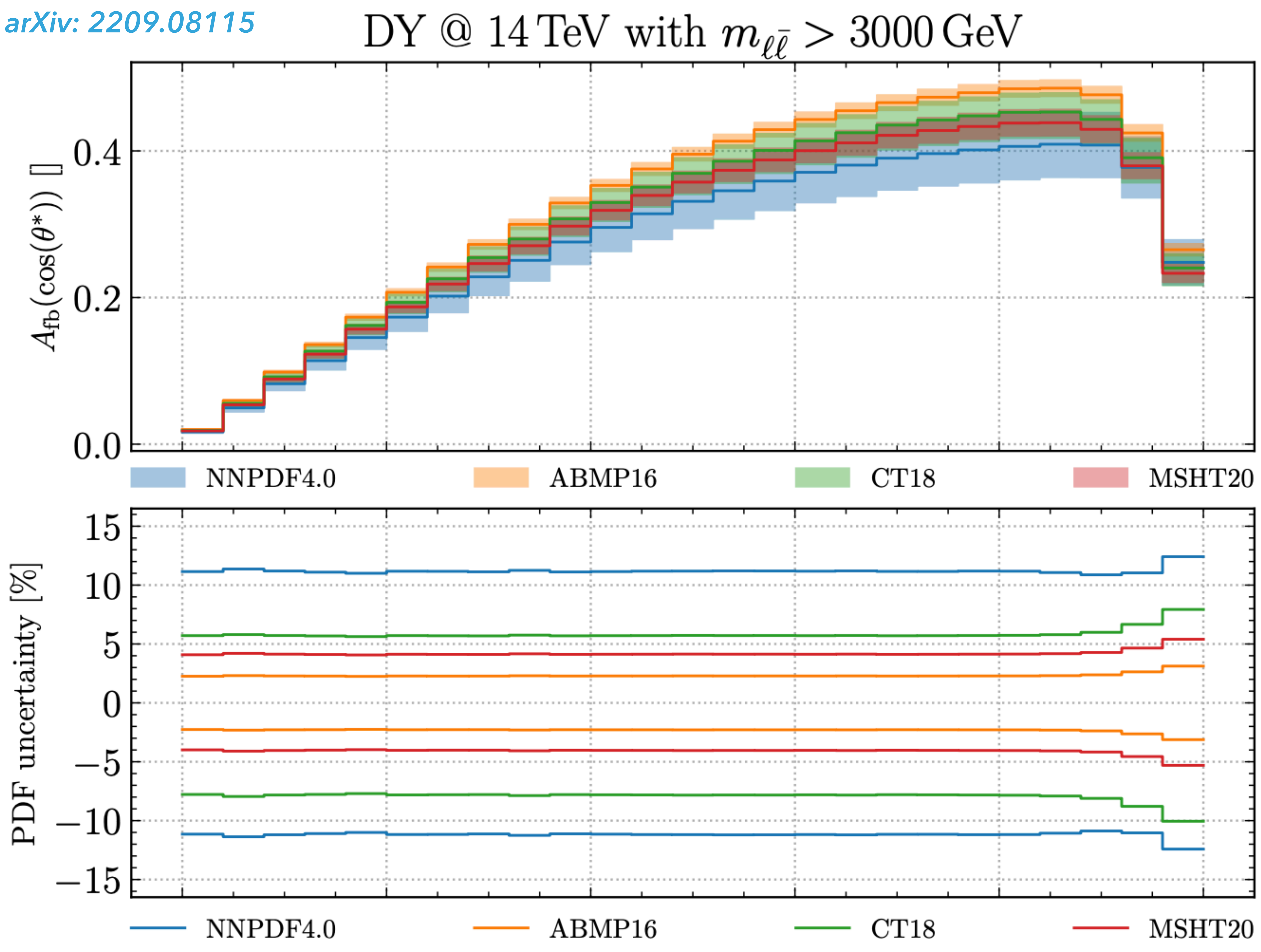
$$x \approx \frac{M}{\sqrt{S}}$$

High-mass final states

⇔

Large-x PDFs

Ball et al arXiv: 2209.08115



#1: Direct searches: heavy Z′ search in DY Forward-Backward asymmetry

. . . AND NEW PHYSICS SEARCHES

$$x \approx \frac{M}{\sqrt{S}}$$

High-mass final states

⇔

Large-x PDFs

- Potential BSM effects might be absorbed in PDFs if only high-mass HL-LHC data are included in fit
- Adding EIC data minimises the risk of BSM-induced bias in PDF fits and allow for more consistent identification of BSM effects in high energy data

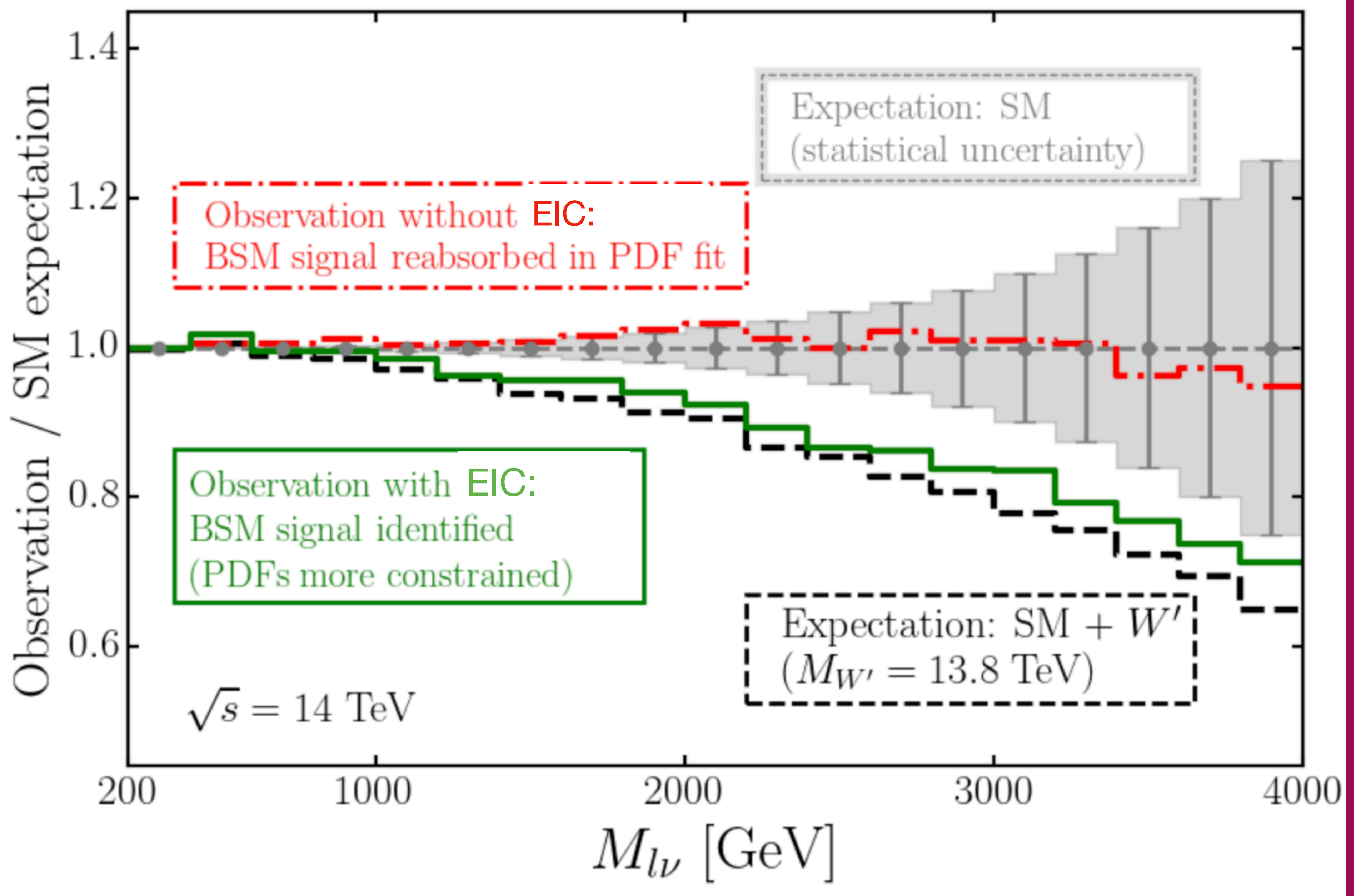
[Hammou, MU, 2410.00963]

If high-energy data (HL-LHC) dominant constrain the large-x region

$$\hat{\sigma}_{\rm BSM} \otimes f_{\rm SM} \approx \hat{\sigma}_{\rm SM} \otimes f_{\rm BSM}$$

If high- (HL-LHC) and low- (EIC) energy constrain the large-x region

$$\hat{\sigma}_{\rm BSM} \otimes f_{\rm SM} \neq \hat{\sigma}_{\rm SM} \otimes f_{\rm BSM}$$



.. more in Tim Hobbs's talk

A STATISTICAL FORMULATION OF THE PROBLEM

- Given some model constraints and given O(5000) experimental data points **D** want to determine the set of functions **{f}** (with i = 1, …, $n_{partons}$) and **estimate their uncertainty**
- Want to find a infinite-dimensional object from a finite number of information.
- Hence, PDF inference is an example of a **non-linear infinite-dimensional** inverse problem.
- Mapping **D** into **f** is mathematically ill defined, nobody knows the true **f = f***
- Best we can do it to find best **f** given the data **D**.

$$P(f|D) \propto P(D|f)\, P(f)$$

✓ Likelihood

? Prior

$$P(\mathbf{D}|f) \sim \exp(-\chi^2/2)$$

$$\chi^2 = \sum_{a,b=1}^{N_{\rm dat}} (T_a(\{f\},\vec{c}) - D_a)\, \Sigma^{-1}_{ab}\, (T_b(\{f\},\vec{c}) - D_b)$$

A STATISTICAL FORMULATION OF THE PROBLEM

- Given some model constraints and given O(5000) experimental data points **D** want to determine the set of functions **{f}** (with i = 1, …, n_{partons}) and **estimate their uncertainty**
- Want to find a infinite-dimensional object from a finite number of information.
- Hence, PDF inference is an example of a **non-linear infinite-dimensional** inverse problem.
- Mapping **D** into **f** is mathematically ill defined, nobody knows the true **f = f***
- Best we can do it to find best **f** given the data **D**.

$$\boxed{P(f|D) \propto P(D|f)\, P(f)}$$

✓ Likelihood

? Prior

$$P(\mathbf{D}|f) \sim \exp(-\chi^2/2)$$

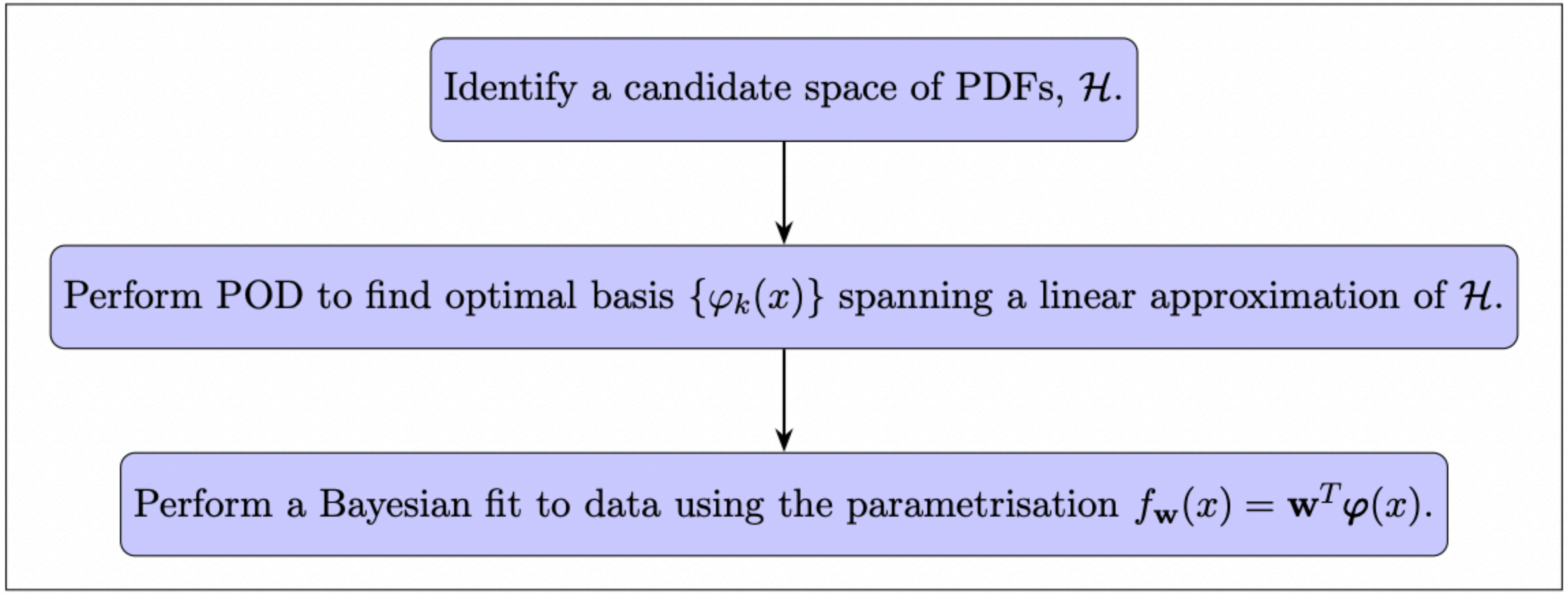
$$\chi^2 = \sum_{a,b=1}^{N_{\text{dat}}} (T_a(\{f\},\vec{c}) - D_a)\, \Sigma_{ab}^{-1}\, (T_b(\{f\},\vec{c}) - D_b)$$

Two ways of finding the **prior so far**:

- Explicit parametrization (parametrical modelling)
  Function is projected on a vector space of parameters **θ**
- Non-parametrical inference (NNs, functional space sampling,…): use data to also infer probable 'smoothness' of f(x) and thus infer probability function P(f) throughout the space of functions

PART II

A LINEAR PDF PARAMETRISATION



Desirable qualities for PDF parametrisation

- Should respect known theoretical constraints such as **small- and large-x scaling** and **sum rules**

$$f(x) \sim A\, x^{1-\alpha}(1-x)^{\beta}$$

- Should be flexible enough to explore the space of candidate PDFs amongst $C^1[0,1]^{n_f}$
- Should have a fair degree of complexity to efficiently fit model parameters

M. Costantini, L. Mantani, J. Moore, MU, arXiv:2507.16913

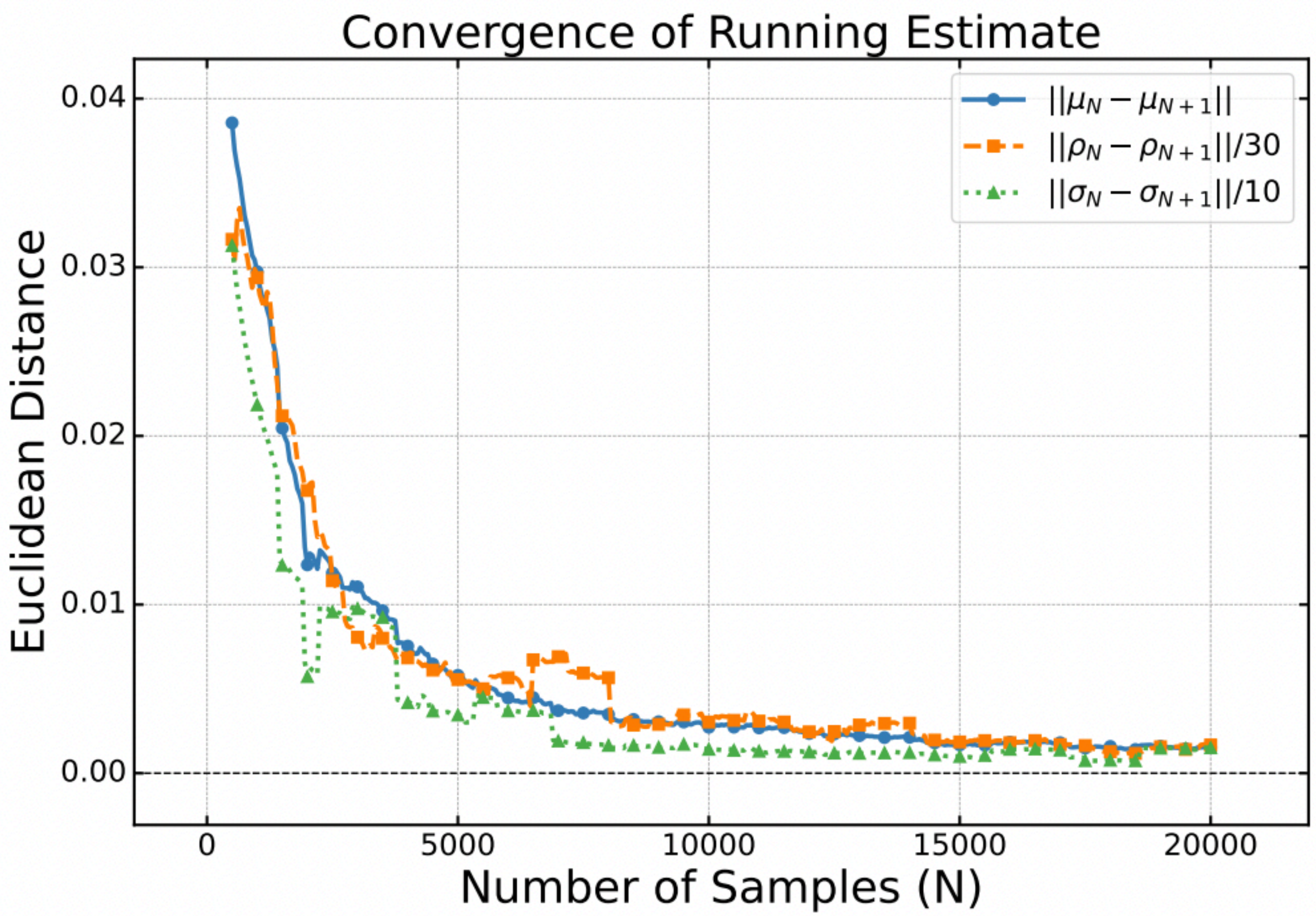
A NEURAL NETWORK ENSEMBLE TO SPAN PDF SPACE

$$xg^j_m(x, Q_0; \boldsymbol{\theta}) = A^j_m x^{1-\alpha^j_m}(1-x)^{\beta^j_m}\mathrm{NN}_m(x; \boldsymbol{\theta}), \qquad m = 1, \ldots, M$$

Pre-processing exponents randomly picked in a suitable range (uniformly)

Output of a **randomly initialised** NN
Here NNPDF4.0 architecture (fully connected feedforward NN with input layer of 2 neurons, two hidden layers with 25 and 20 neurons & n_f outputs).
NN weights randomly initialised with Glorot normal distribution

M = 20K reproduces first momenta of NN ensembles to sufficient accuracy



M. Costantini, L. Mantani, J. Moore, MU, arXiv:2507.16913

POD FOR A LINEAR PDF PARAMETRISATION

- Proper Orthogonal Decomposition (POD): decompose NN ensemble in its principal modes, from M=20K samples to N = O(10 - 100) basis functions → dimensional reduction of $\mathscr{H}$ to a linear functional space, where orthogonal basis functions $\{\boldsymbol{\varphi}_k\}$ with k = 1, …, N are organised in order of importance

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

Basis functions

$$\boldsymbol{\varphi} = (\varphi_0(x), \varphi_1(x), \ldots, \varphi_N(x))$$

Weights = parameters $\mathbf{w} = (1, w_1, \ldots, w_N)$

- POD provides most efficient way of capturing key features of an infinite-dimensional ensemble of functions using only a finite, often surprisingly small, number of 'modes'.

- Theoretical constraints (sum rules, integrability..) can be imposed directly on the weights **w**

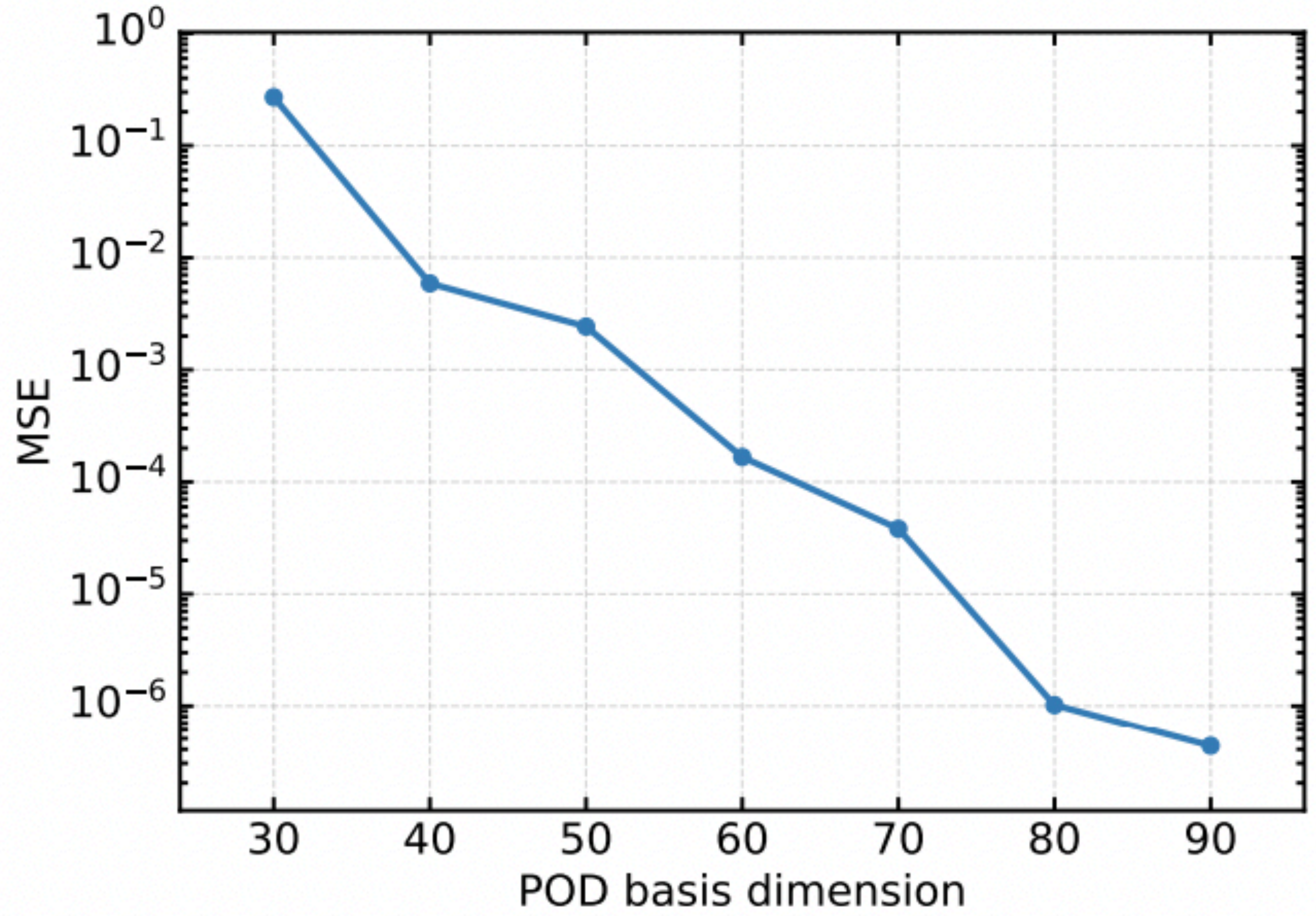
ASSESSING COMPLETENESS AND GENERALIZATION

- Completeness: any PDF produced using a random initialisation of the NN can be well-approximated by the linear mode in POD basis
- Generalization: any PDF parametrisation (from all available in LHAPDF) can be well-approximated by the linear model in POD basis

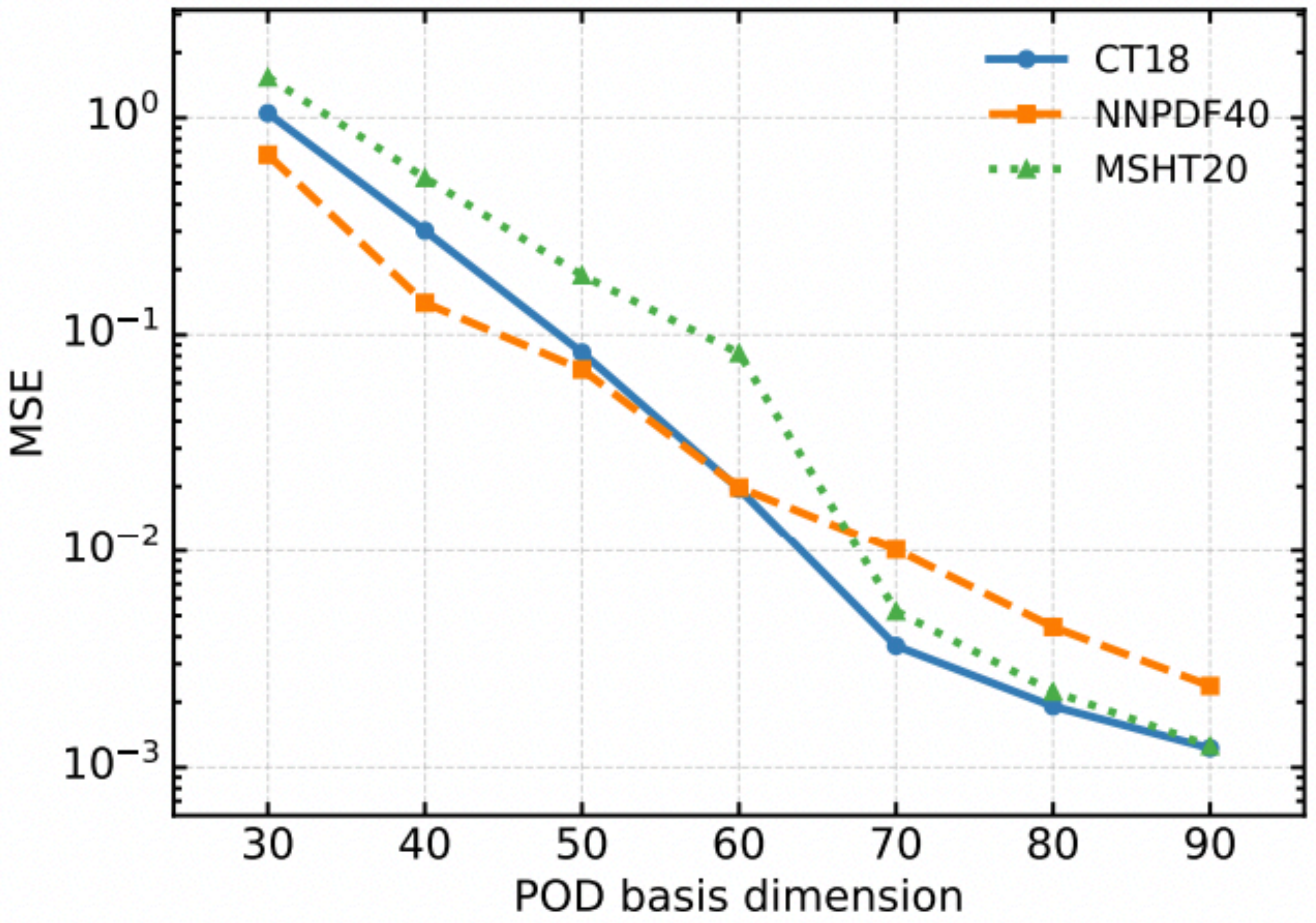
$$\mathrm{MSE} = \frac{1}{N_{\mathrm{targ}}} \sum_{i=1}^{N_{\mathrm{targ}}} \left\| \mathbf{f}_T^{(i)} - \mathbf{f}_{\mathbf{w}^{(i)}} \right\|_2^2 = \frac{1}{N_{\mathrm{targ}}} \sum_{i=1}^{N_{\mathrm{targ}}} \left\| \mathbf{f}_T^{(i)} - \left(\varphi_0 + \sum_{n=1}^{N} w_n^{(i)} (\varphi_n - \varphi_0) \right) \right\|_2^2$$

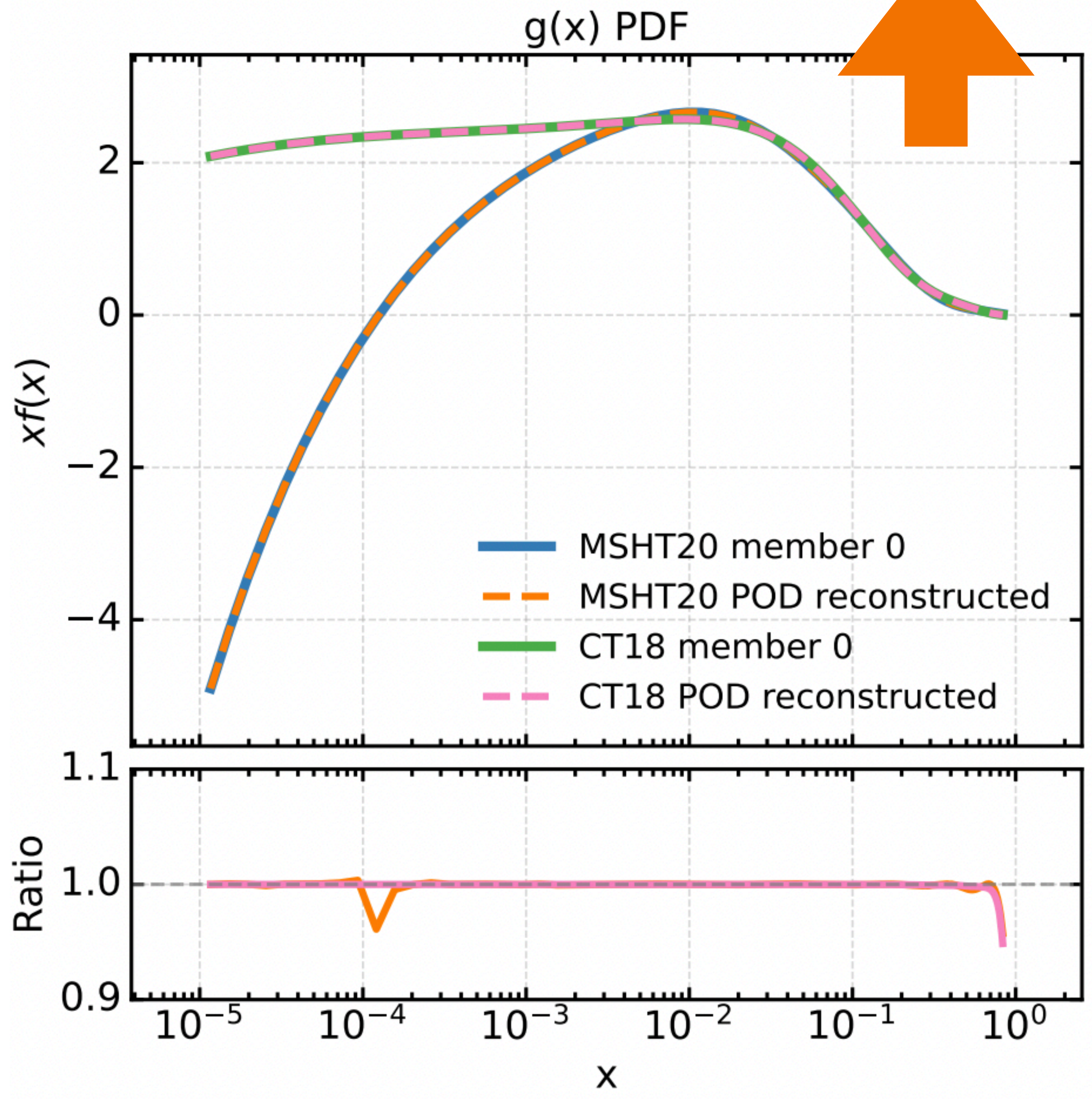
80 weights enough to reproduce other parametrisation

COMPLETENESS



GENERALIZATION





M. Costantini, L. Mantani, J. Moore, MU, arXiv:2507.16913

BAYESIAN METHODOLOGY

- We can now turn to the problem of inferring the PDF parameters **w** from the available data via numerical sampling

$$p(\mathbf{w}|\mathbf{D}) = \frac{p(\mathbf{D}|\mathbf{w})p(\mathbf{w})}{\mathcal{Z}}$$

Likelihood $\quad p(\mathbf{D}|\mathbf{w}) \propto \exp\left(-\frac{1}{2}\left[\chi^2_{\rm data}(\mathbf{D},\mathbf{w}) + \chi^2_{\rm pos}(\mathbf{w}) + \chi^2_{\rm integ}(\mathbf{w})\right]\right)$

Prior

Evidence or marginal likelihood

- How many parameters do we need?

$$Z = p(\mathbf{D}\,|\,\mathscr{M}_N) = \int d\mathbf{w}\; p(\mathbf{w}\,|\,\mathbf{D},\mathscr{M}_N)p(\mathbf{w})$$

Model with N weights = N vectors in the POD basis

$$\ln Z = -\frac{1}{2}\chi^2 + \frac{N}{2}\ln(2\pi) + \ln\left(\frac{\sqrt{|(X^T\Sigma^{-1}X)^{-1}|}}{\prod_i (b_i - a_i)}\right)$$

- Determine model(s) with largest evidence (balance between max likelihood and complexity)
- Bayesian average of models with best evidence
- Bayesian posterior update: start from large uniform prior and compute posterior analytically on linear data and use posterior as a prior to do numerical sampling of final posterior

APPLICATION: A CLOSURE TEST OF DIS FIT

Dataset	N_{data}	Experiment	Model
Numerical fit			
NMC F_2^d/F_2^p [55]	121	Deuteron	Ratio
SLAC F_2^p [56]	33	Deuteron	Linear
SLAC F_2^d [56]	34	Deuteron	Linear
BCDMS F_2^p [57]	333	Deuteron	Linear
BCDMS F_2^d [57]	248	Deuteron	Linear
Analytical fit			
(*) NMC $\sigma^{\text{NC,p}}$ [58]	204	NMC	Linear
CHORUS σ_{CC}^{ν} [59]	416	Nuclear	Linear
CHORUS $\sigma_{CC}^{\bar{\nu}}$ [59]	416	Nuclear	Linear
NuTeV dimuon σ_{CC}^{ν} [60]	39	Nuclear	Linear
(*) NuTeV dimuon $\sigma_{CC}^{\bar{\nu}}$ [60]	37	Nuclear	Linear
HERA I+II $\sigma_{CC}^{e^+p}$ [61]	39	HERA	Linear
HERA I+II charm σ_{NC} [62]	37	HERA	Linear
HERA I+II $\sigma_{NC}^{e^-p}$ (320 GeV)	159	HERA	Linear
(*) HERA I+II $\sigma_{NC}^{e^+p}$ (460 GeV)	204	HERA	Linear
HERA I+II $\sigma_{NC}^{e^+p}$ (575 GeV)	254	HERA	Linear
HERA I+II $\sigma_{NC}^{e^+p}$ (820 GeV)	70	HERA	Linear
HERA I+II $\sigma_{NC}^{e^+p}$ (920 GeV)	377	HERA	Linear
(*) HERA I+II $\sigma_{CC}^{e^-p}$ [61]	42	HERA	Linear
HERA I+II bottom σ_{NC} [62]	26	HERA	Linear
Total	3089		

- Underlying law: model obtained with N=40 weights

$$\mathbf{f}_{\text{in}} = \boldsymbol{\varphi}_0 + \sum_{k=1}^{N} w_k^{\text{in}} \left(\boldsymbol{\varphi}_k - \boldsymbol{\varphi}_0\right)$$

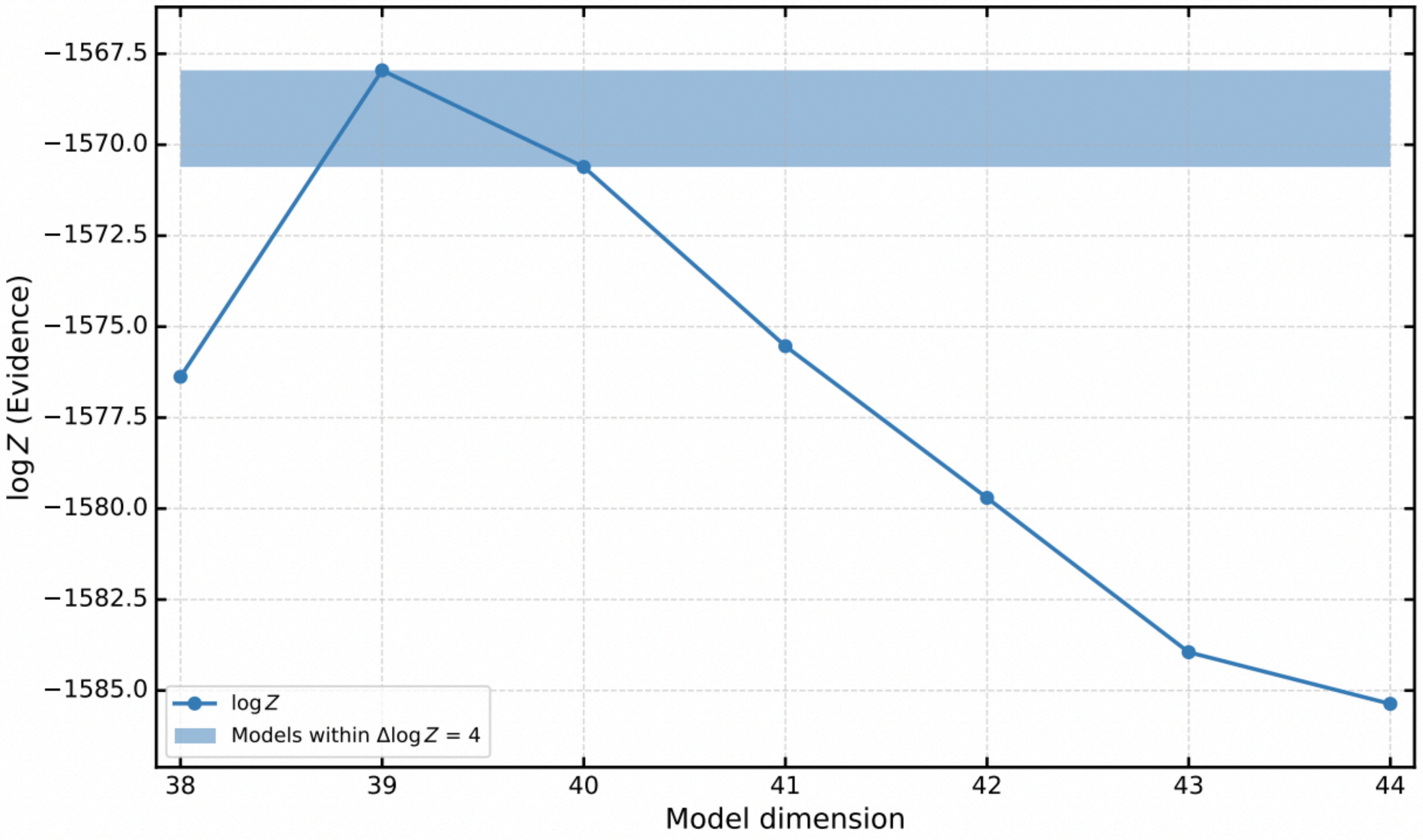


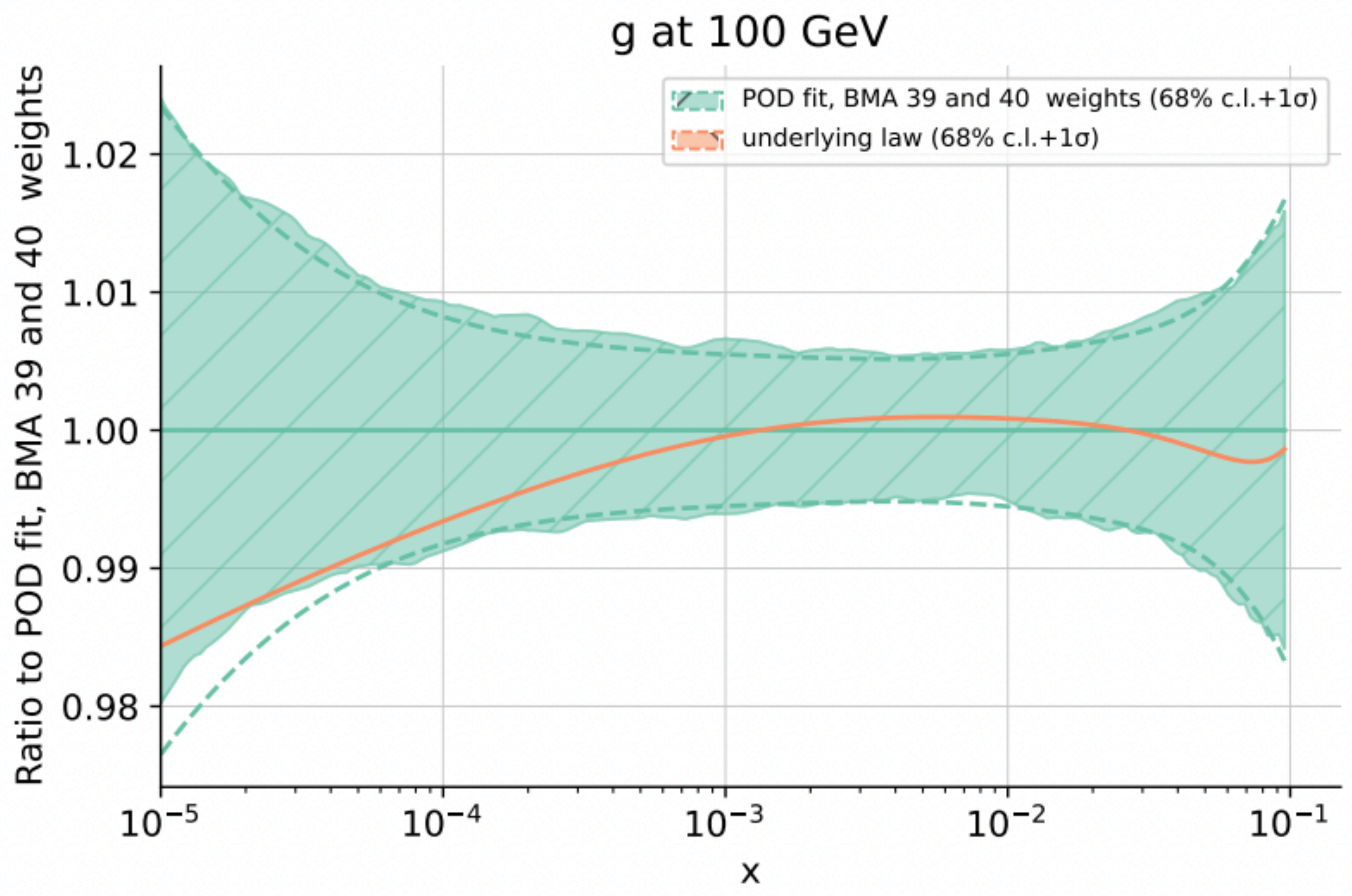
Table 3: Goodness-of-fit and model posterior probabilities for the two most likely models.

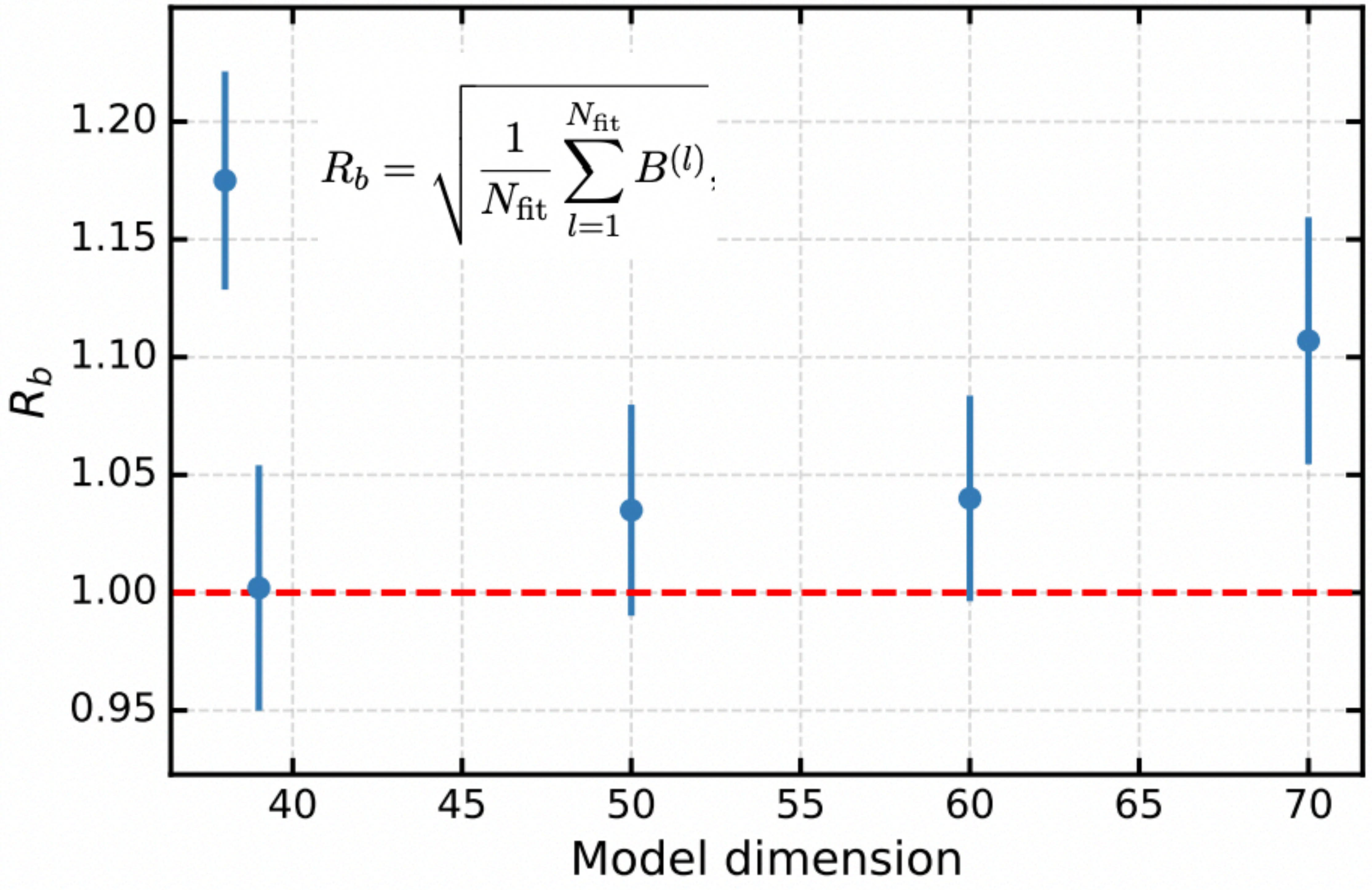
Model Complexity	χ^2/N_{data}	$p(\mathcal{M}_k\|D)$
39	1.033	0.934
40	1.032	0.066

M. Costantini, L. Mantani, J. Moore, MU, arXiv:2507.16913

APPLICATION: A CLOSURE TEST OF DIS FIT

M. Costantini, L. Mantani, J. Moore, MU, arXiv:2507.16913



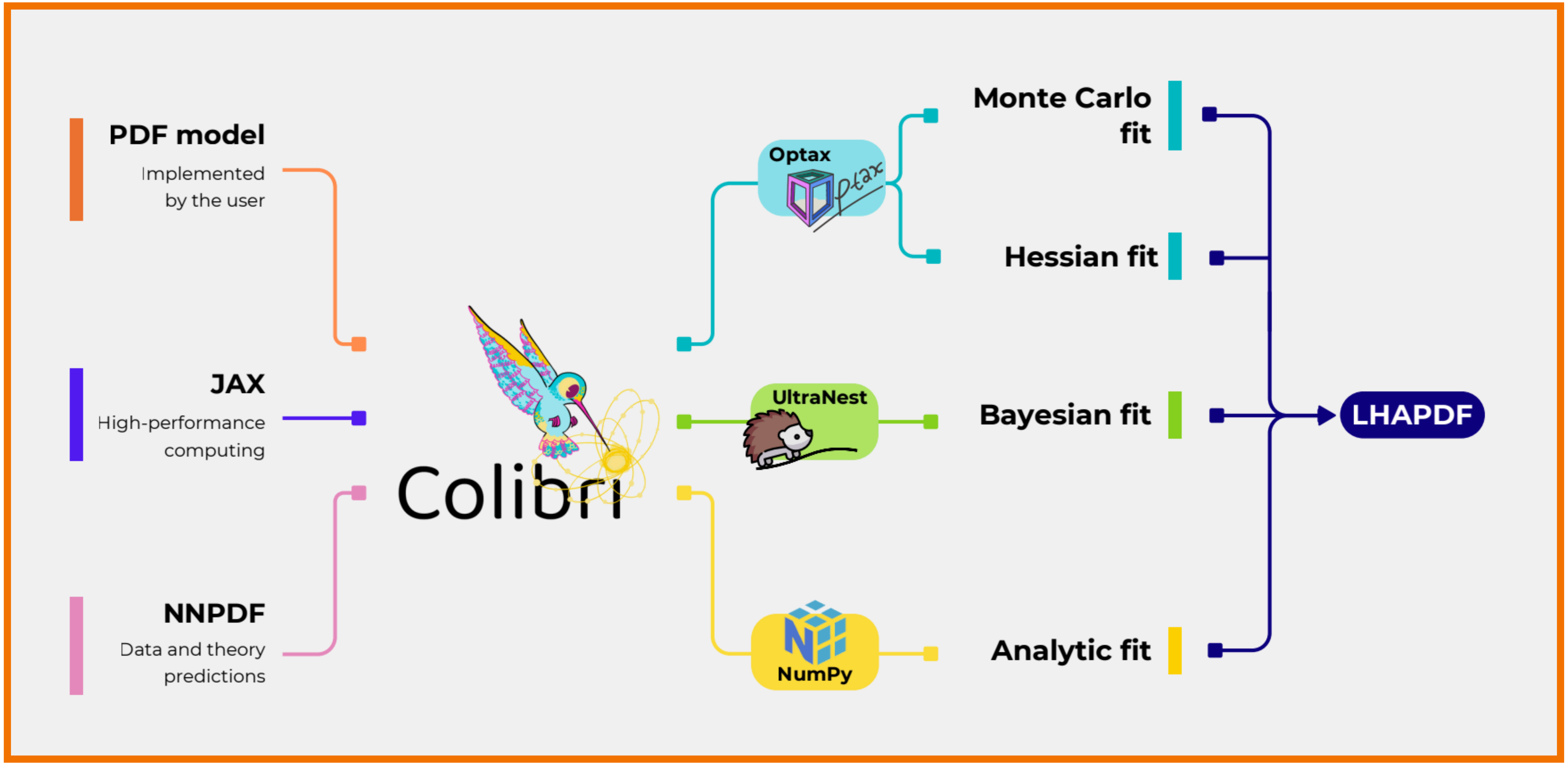


$$B^{(l)} = \frac{1}{N_{\rm data}} \sum_{i,j=1}^{N_{\rm data}} \left(\mathbb{E}_\epsilon T_i[f_k^{(l)}] - D_i^{L_0} \right) (C_{\rm PDF})_{ij}^{-1} \left(\mathbb{E}_\epsilon T_j[f_k^{(l)}] - D_j^{L_0} \right)$$

- Model-averaged posterior successfully recovers underlying law within estimated uncertainties.
- Posterior from average of models with best evidence faithfully reproduces statistical distribution of the data
- Under- and over-parametrised models systematically underestimate uncertainties in data region.

COLIBRI: A NEW TOOL FOR FAST-FLYING PDFS

- COLIBRI: open-source Python code for users to implement PDF model, using the built-in functionalities for fast forward map computation and easy access to data
- Several error propagation methodologies built in (Hessian, Monte Carlo bootstrapping and Bayesian)



M. Costantini, L. Mantani, J. Moore, V. Schütze Sanchez, MU, arXiv:2510.03391

CONCLUSIONS

- NN ensembles have broad applications [S. Beneveded, J. Thaler, arXiv:2506.00113]
- Linear model simplest and most appealing model & with a carefully chosen basis becomes a complete, flexible parametrisation [M. Costantini, L. Mantani, J. Moore, MU, arXiv:2507.16913]
- Bayesian evidence picks the "right" model and lets us perform principled model averaging across top candidates.
- Bias tests show the method selects appropriate complexity (Occam's razor) and yields faithful uncertainty estimation
- This model (and many more) are publicly available in Colibri [M. Costantini, L. Mantani, J. Moore, V. Schütze Sanchez, MU, arXiv:2510.03391] https://github.com/HEP-PBSP/colibri & doc https://hep-pbsp.github.io/colibri/
  - ✓ Polynomial parametrisation (LH), deterministic NN parametrisation and linear model available
  - ✓ Work in progress on Bayesian NN, Gaussian Processes
  - ✓ Work in progress for simultaneous Bayesian fit of PDFs with SM/BSM parameters

THANK YOU FOR YOUR ATTENTION!