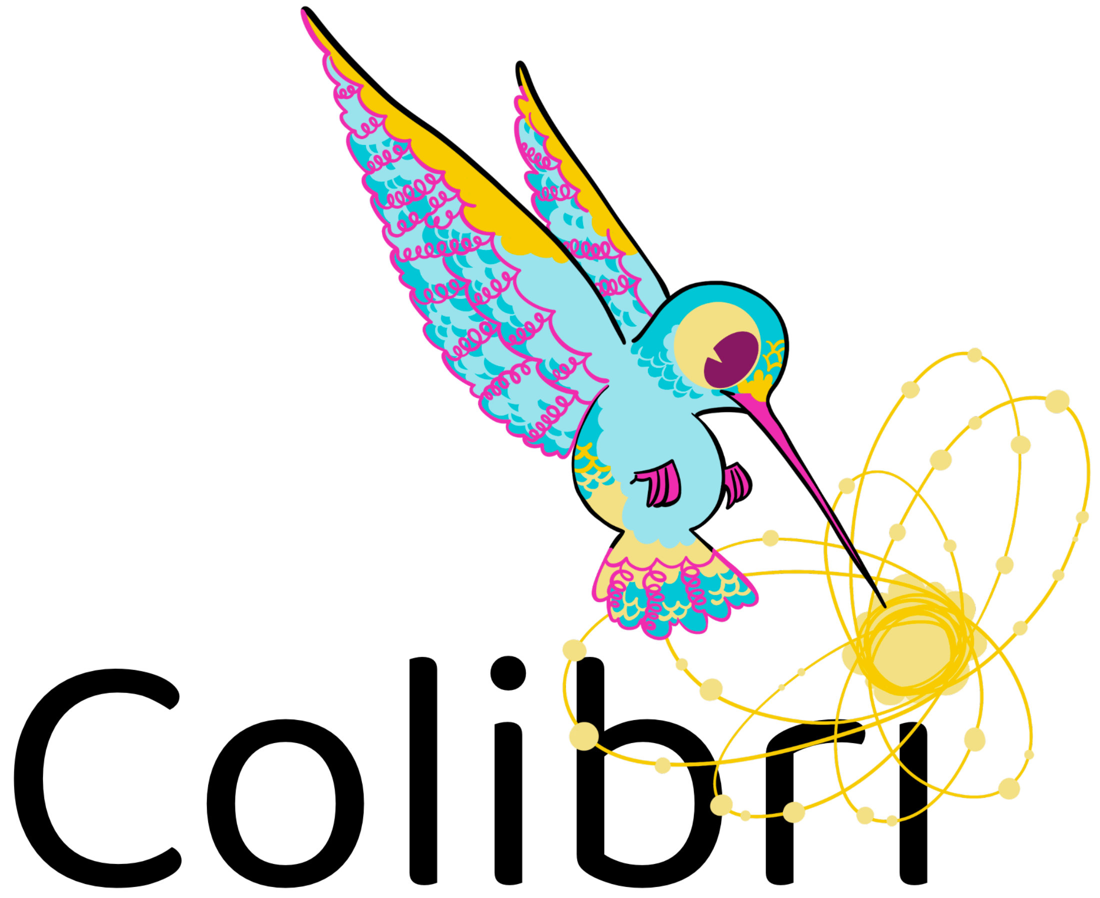
Linear models for Bayesian PDF fits

Mark N. Costantini

In collaboration with L. Mantani, J. Moore and M. Ubiali

PDF4LHC, CERN December 2024



Artwork by: @qftoons

UNIVERSITY OF CAMBRIDGE

Funded by the European Union

European Research Council
Established by the European Commission

Outline

- Introduction / Motivation
- POD Parametrisation of PDFs
- Bayesian Workflow
- Benchmark of the methodology: closure tests
- Conclusions / Outlook

Introduction

PDF parametrisation(s)

What are desirable qualities that a good PDF parametrisation should possess?

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

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

2. It should be flexible enough to explore the space of candidate PDFs amongst $C^1[0,1]$

3. It should be straightforward to fit the model parameters

PDF parametrisation(s)

What are desirable qualities that a good PDF parametrisation should posses?

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

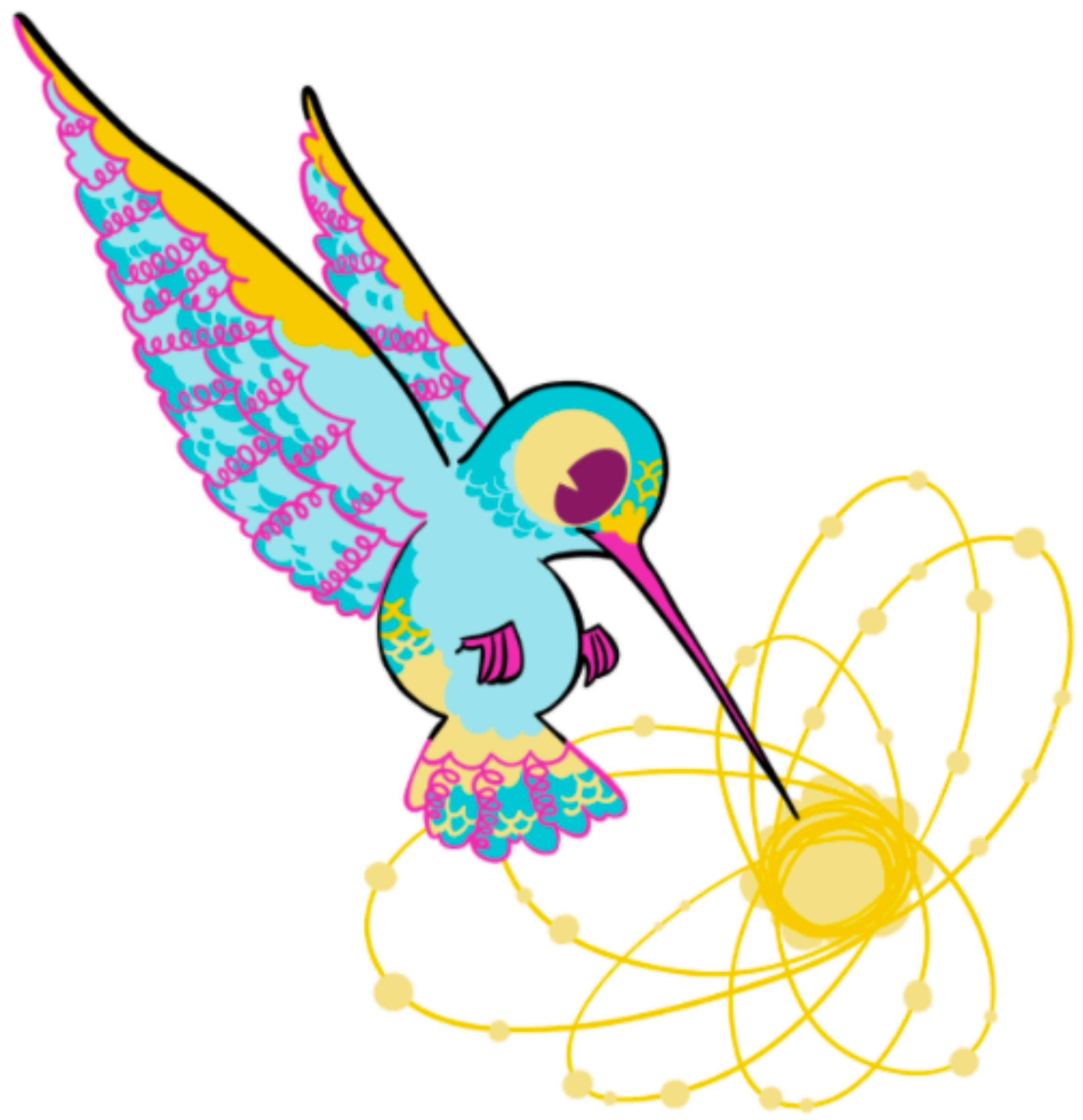
$$f(x) \sim Ax^{\alpha}(1-x)^{\beta}$$

2. It should be flexible enough to explore the space of candidate PDFs amongst $C^1[0,1]$

3. It should be straightforward to fit the model parameters $\Rightarrow$ Facilitate realistic PDF fit using fully Bayesian methodology

Fitting Framework

Colibri



Artwork by @qftoons

colibri

Tests passing | code style black | codecov 96%

A reportengine app to perform PDF fits using arbitrary parametrisations.

→Backbone: reportengine and validphys

→Makes use of Jax for high performance array computing (GPUs, JIT)

→Compatible with OpenMPI

→ Allows flexible implementation of any PDF parametrisation

→ Bayesian (Nested Sampling and now PYMC) and MC fits possible

POD Parametrisation

Linear PDF parametrisation

$$f_{POD}(x, Q^2) = \xi_0(x, Q^2) + \sum_{i=1}^{N} w_i \left(\xi_i(x, Q^2) - \xi_0(x, Q^2)\right)$$

$\{\xi_1, \ldots, \xi_N\}$ is a collection of basis functions

$\rightarrow$ If ξ_i satisfy Sum Rules (SRs), then f_{POD} also does (same holds for Integrability and small- large-x scaling)

$\rightarrow f_{POD}$ is a linear model, linear in w_i

Proper Orthogonal Decomposition

→ Combine multiple LHAPDF sets and perform a POD

$$X_{lk} \equiv f_\alpha^{(k)}(x_i, Q) - f_\alpha^{(0)}(x_i, Q) \qquad \alpha \in \{1,\ldots,N_f\}$$

$$l \in N_x(\alpha - 1) + i \quad , k \in \{1,\ldots,N_{rep}\} \qquad i \in \{1,\ldots,N_x\}$$

POD: explore the principal directions in a space of functions, ordering them from most important direction, to least important direction.

In the finite-dimensional case POD reduces to the Singular Value Decomposition (SVD) +Principal Component Analysis (PCA) of the given set

Proper Orthogonal Decomposition

Construction of the basis

Combine multiple LHAPDF sets and perform a POD

PDF Sets	Number of Replicas in MC representation
MSHT20nnlo_as118	112
CT18NNLO	200
CT10nnlo	320
MMHT2014nnlo68cl	235
CT14nnlo	320
MSTW2008nnlo90cl	197
NNPDF23_nnlo_as_0118	65

→ Impose exact SRs

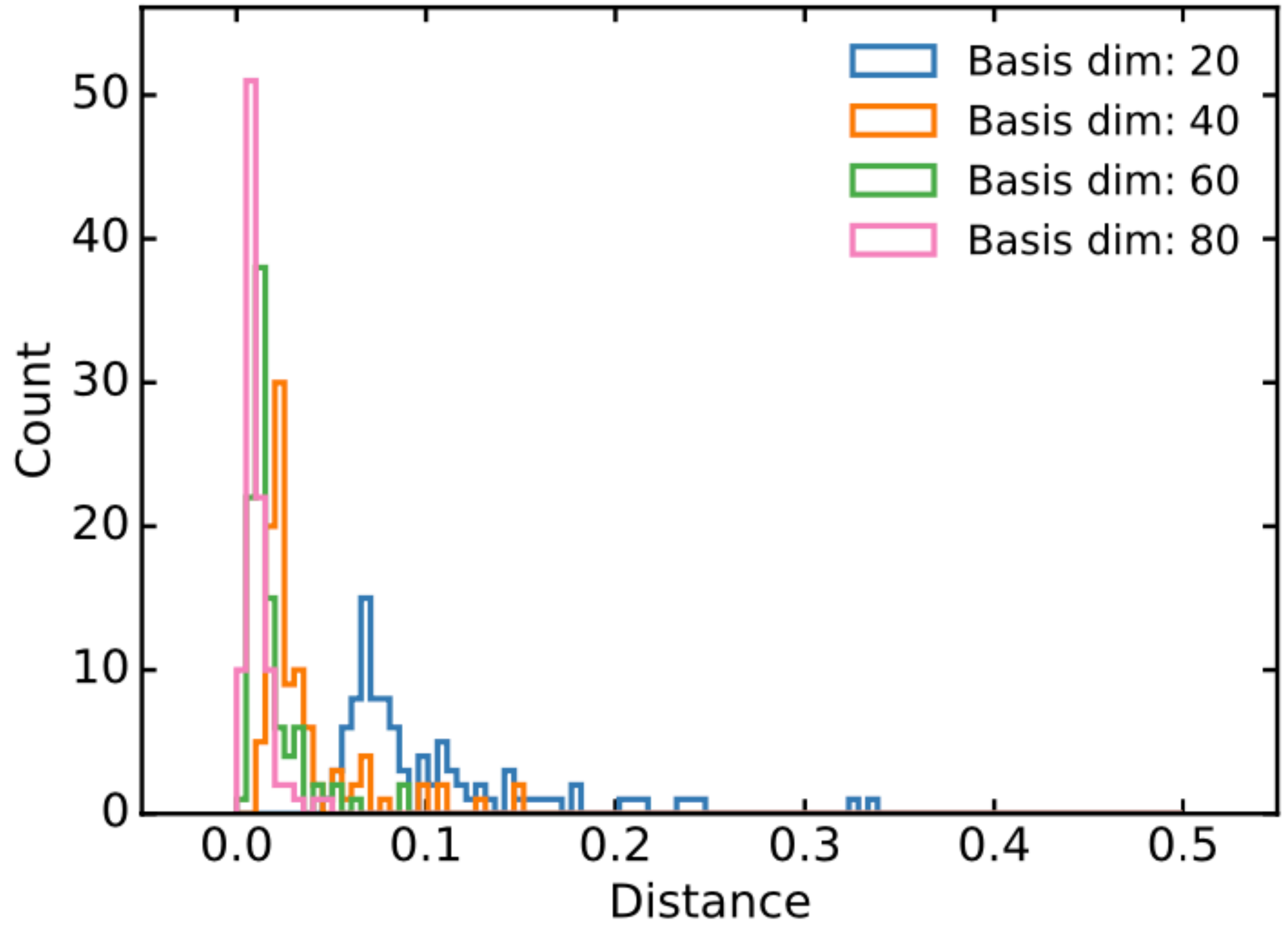
→ Impose basis consistency, eg, for Intrinsic Charm basis at $Q = 1.65$ GeV

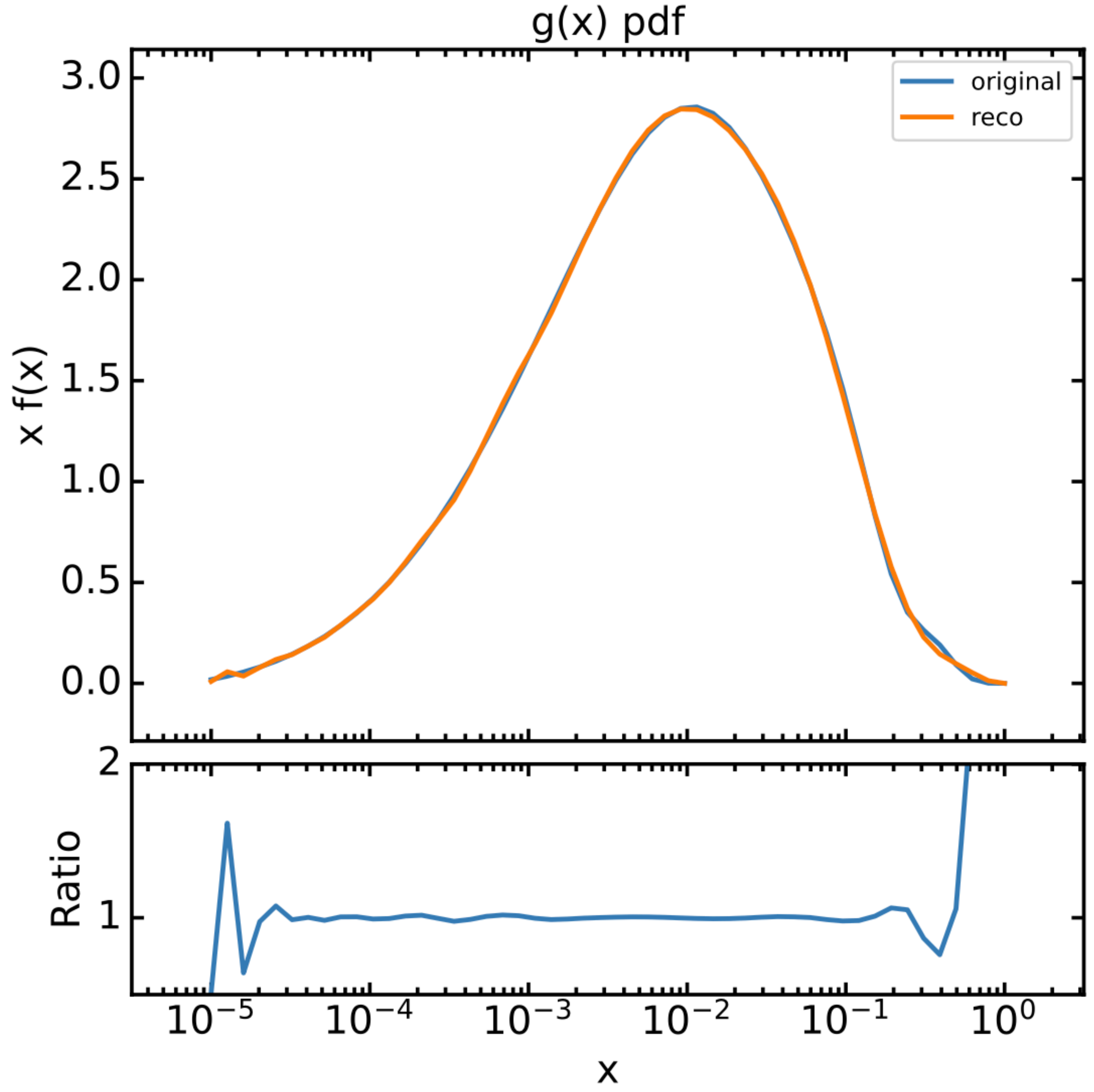
$$V = V_{15} = V_{24} = V_{35}, \quad \Sigma = T_{24} = T_{35}$$

Completeness of the basis

Check performance of the basis on target PDF set: eg NNPDF4.0

$$d = ||f_T(x) - f_{POD}(x)||^2$$





Evidence "tells us" what the required flexibility of the parametrisation needs to be given the data

Bayesian Workflow

Bayesian linear regression

"Linear Data" (DIS) $\mathbf{y} \sim \mathcal{N}(\mathbf{y}_0, \Sigma)$

→ forward model is linear in the parameters **w**

$$\mathbf{t}(\mathbf{w}) = X\mathbf{w} + \epsilon$$

Analytic posterior distribution

$$p(\mathbf{w} \mid \mathbf{y}_0) \sim \mathcal{N}(\hat{\mathbf{w}}, (X^T\Sigma^{-1}X)^{-1}), \qquad \hat{\mathbf{w}} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\mathbf{y}_0$$

Bayesian linear regression

Given a model $\mathscr{M}_k$ the evidence is defined as

$$Z = p(\mathbf{y}_0 \,|\, \mathscr{M}_k) = \int d\mathbf{w} p(\mathbf{w} \,|\, \mathbf{y}_0, \mathscr{M}_k) p(\mathbf{w})$$

For a Gaussian posterior we can use the Laplace approximation

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

Bayesian linear regression

Given a model $\mathscr{M}_k$ the evidence is defined as

$$Z = p(\mathbf{y}_0 | \mathscr{M}_k) = \int d\mathbf{w} p(\mathbf{w} | \mathbf{y}_0, \mathscr{M}_k) p(\mathbf{w})$$

For a Gaussian posterior we can use the Laplace approximation

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

Favours models that fit well data

Bayesian linear regression

Given a model $\mathscr{M}_k$ the evidence is defined as

$$Z = p(\mathbf{y}_0 | \mathscr{M}_k) = \int d\mathbf{w} p(\mathbf{w} | \mathbf{y}_0, \mathscr{M}_k) p(\mathbf{w})$$

For a Gaussian posterior we can use the Laplace approximation

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

Favours models that fit well data

penalises models with too many parameters

Non-linear regression

Eg ratio of DIS observables

MCMC to sample from the parameter space (and compute the evidence integral)

Fit convergence can be sped up massively by updating the analytical posterior when experiments are uncorrelated

$$\mathbf{y} \sim \mathcal{N}(\mathbf{t}(\mathbf{w}), \Sigma), \quad \text{with} \quad \Sigma = \Sigma_1 \oplus \Sigma_2, \, \mathbf{y}_0^T = (\mathbf{y_1}^T, \mathbf{y_2}^T)$$

$$p(\mathbf{w}\,|\,\mathbf{y_0}) = \frac{p_{\mathbf{y_1}}(\mathbf{w}\,|\,\mathbf{y_1}) \quad \exp(-\frac{1}{2}\,||\,\mathbf{y_2} - t_2(\mathbf{w})\,||_{\Sigma_2}^2)}{\int d\mathbf{w}\; p_{\mathbf{y_1}}(\mathbf{w}\,|\,\mathbf{y_1}) \quad \exp(-\frac{1}{2}\,||\,\mathbf{y_2} - t_2(\mathbf{w})\,||_{\Sigma_2}^2)}$$

Bayesian model average

Having fixed a POD basis we can explore multiple models $\mathscr{M}_k, k \in \{1, \ldots, N\}$ with different number of basis elements

At the end we can average over all of them as

$$p(\mathbf{f}_{POD} \mid \mathbf{y}_0) = \sum_k p(\mathbf{y}_0 \mid \mathbf{f}_{POD}, \mathscr{M}_k) p(\mathscr{M}_k \mid \mathbf{y}_0)$$

And probability of the model given by

$$p(\mathscr{M}_k \mid \mathbf{y}_0) = \frac{p(\mathbf{y}_0 \mid \mathscr{M}_k)}{\sum_l p(\mathbf{y}_0 \mid \mathscr{M}_l)}$$

Closure Tests

Settings of the fit

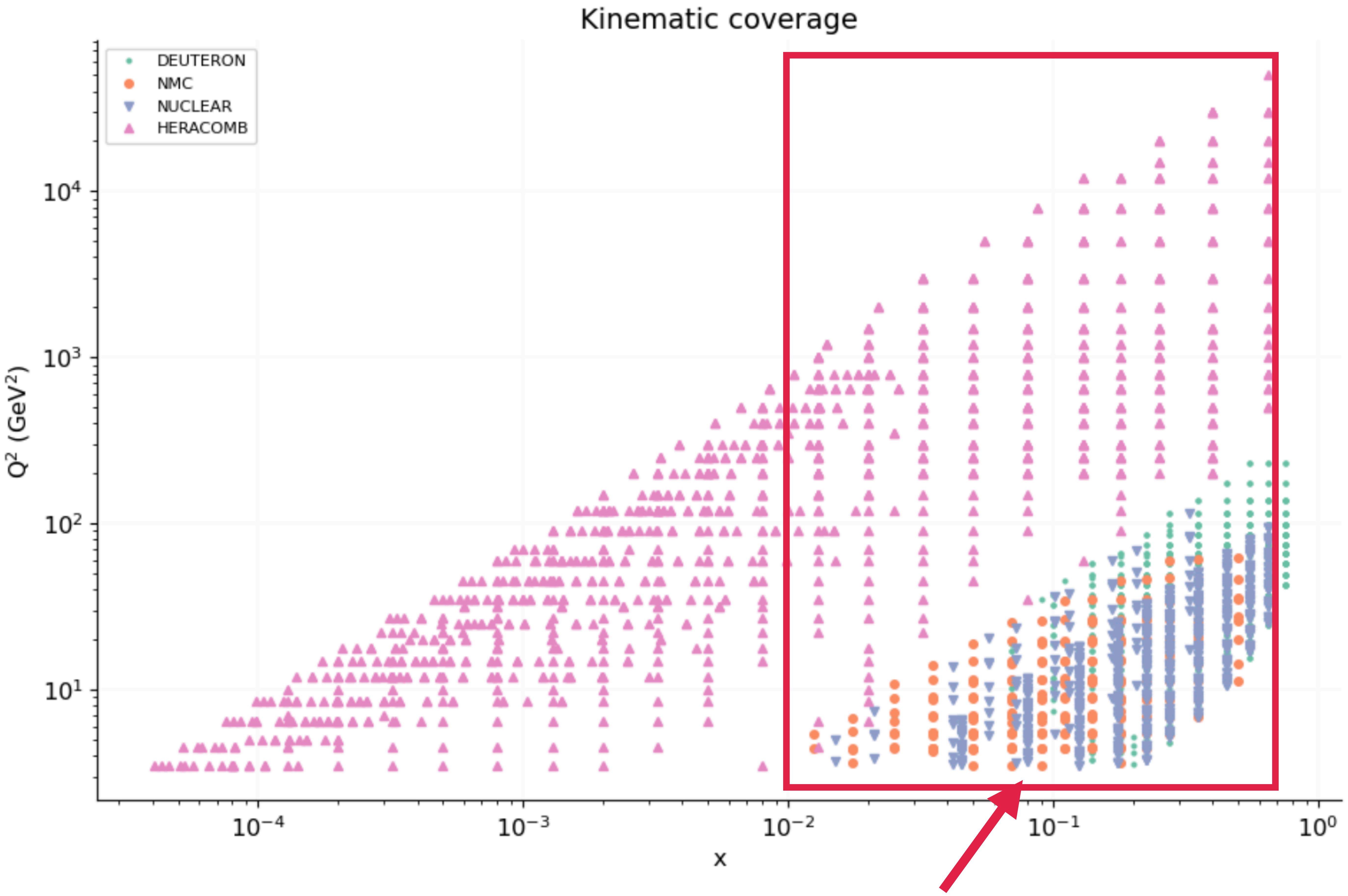
Data

Full NNPDF4.0 DIS dataset,

$N_{dat} = 3084$

$N_{dat}(x > 0.01) = 2463$

$N_{dat}(x < 0.01) = 621$



Data Region

Model specific closure tests

Start from known underlying law

$$\mathbf{f}_{in} = \xi_0 + \sum_{i=1}^{N} (\xi_i - \xi_0)\tilde{w}_i$$

With 15 active parameters

Generate data as

$$\mathbf{d} \sim FK(\mathbf{f}_{in}) + \epsilon,\ \epsilon \sim \mathcal{N}(0,\Sigma)$$

Level 1 closure test

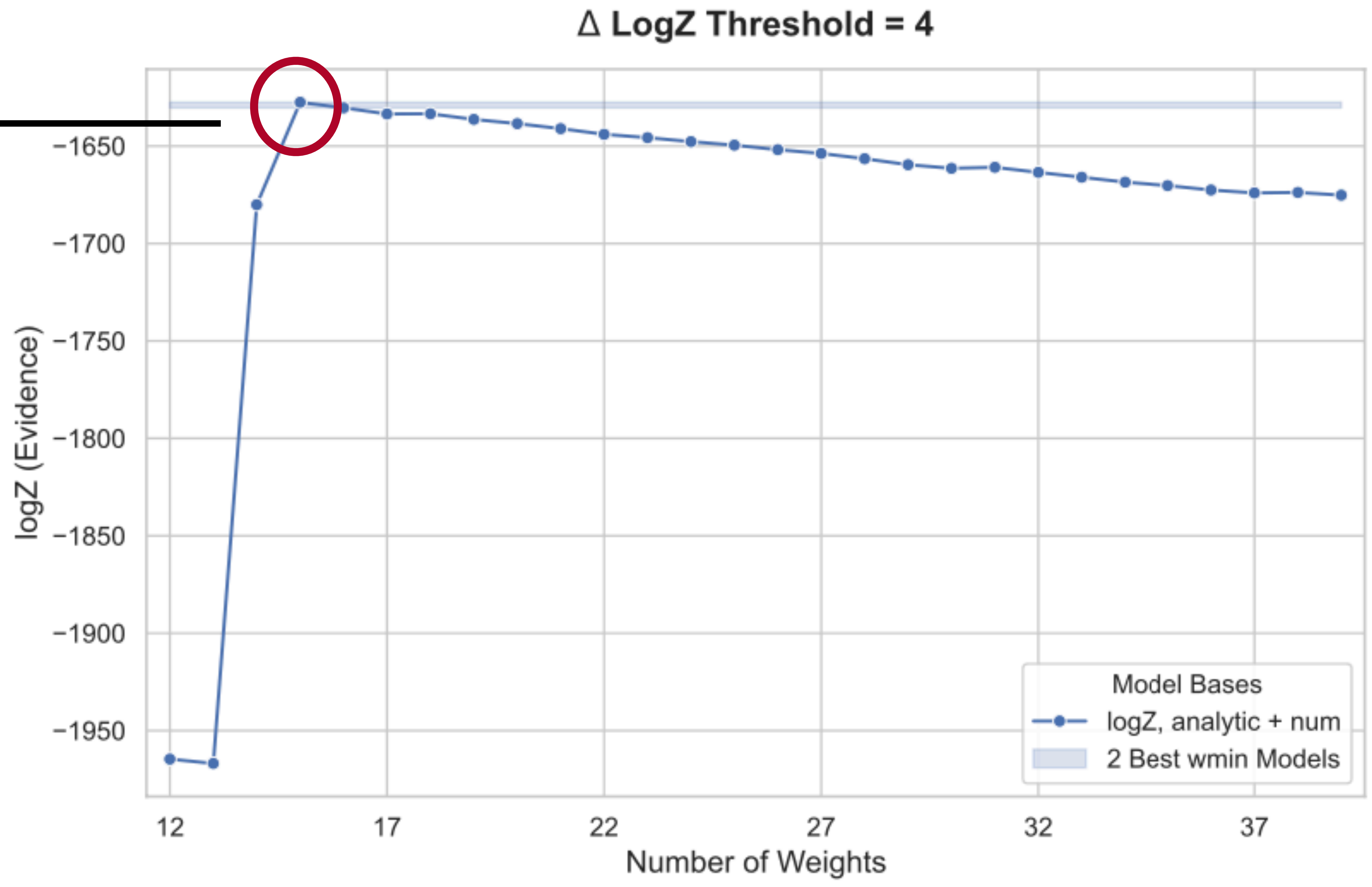
Scan of models, given a fixed POD basis

Data ~ Theory + Gaussian Noise

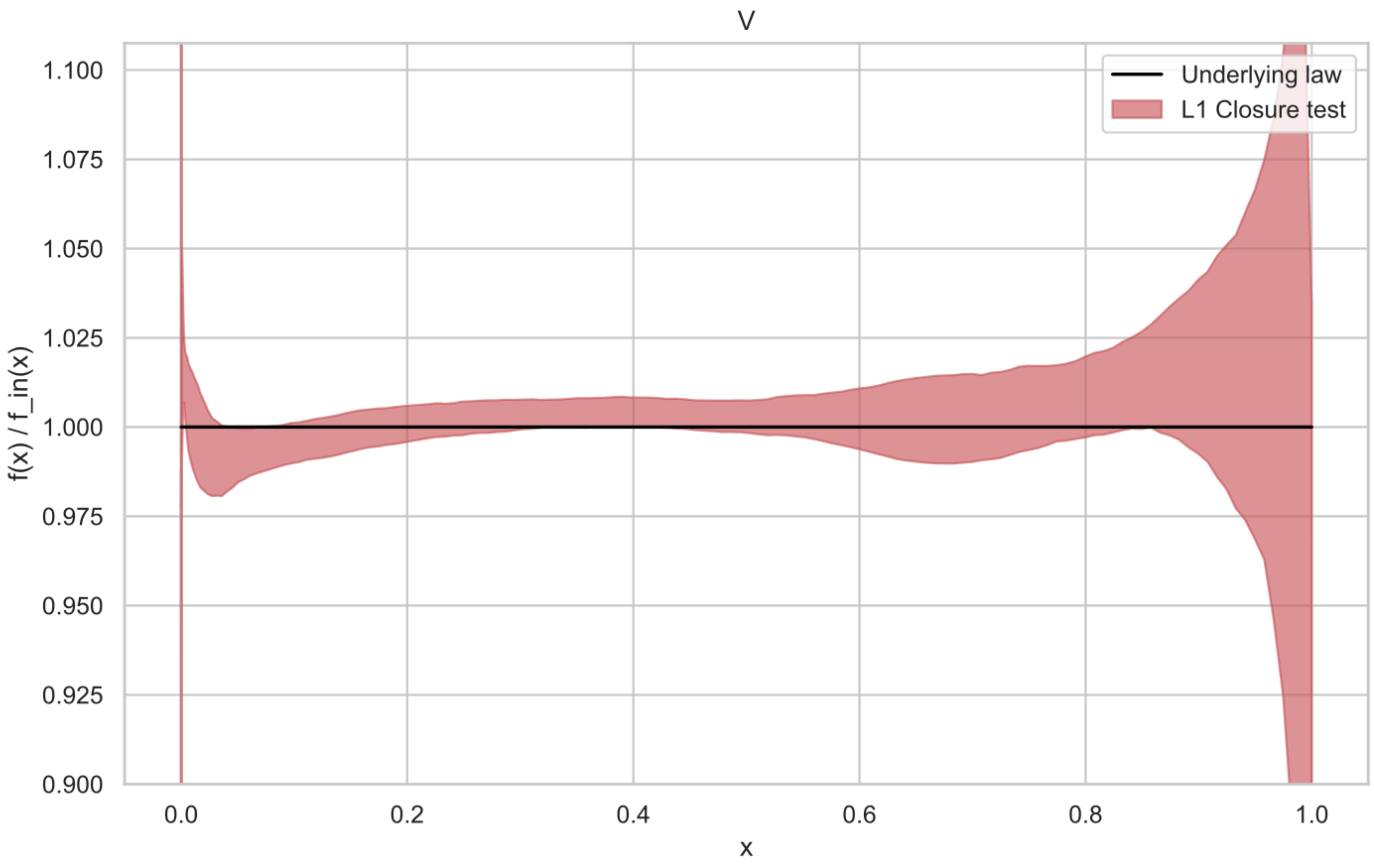
15 parameters strike balance between goodness of fit and Occam penalty

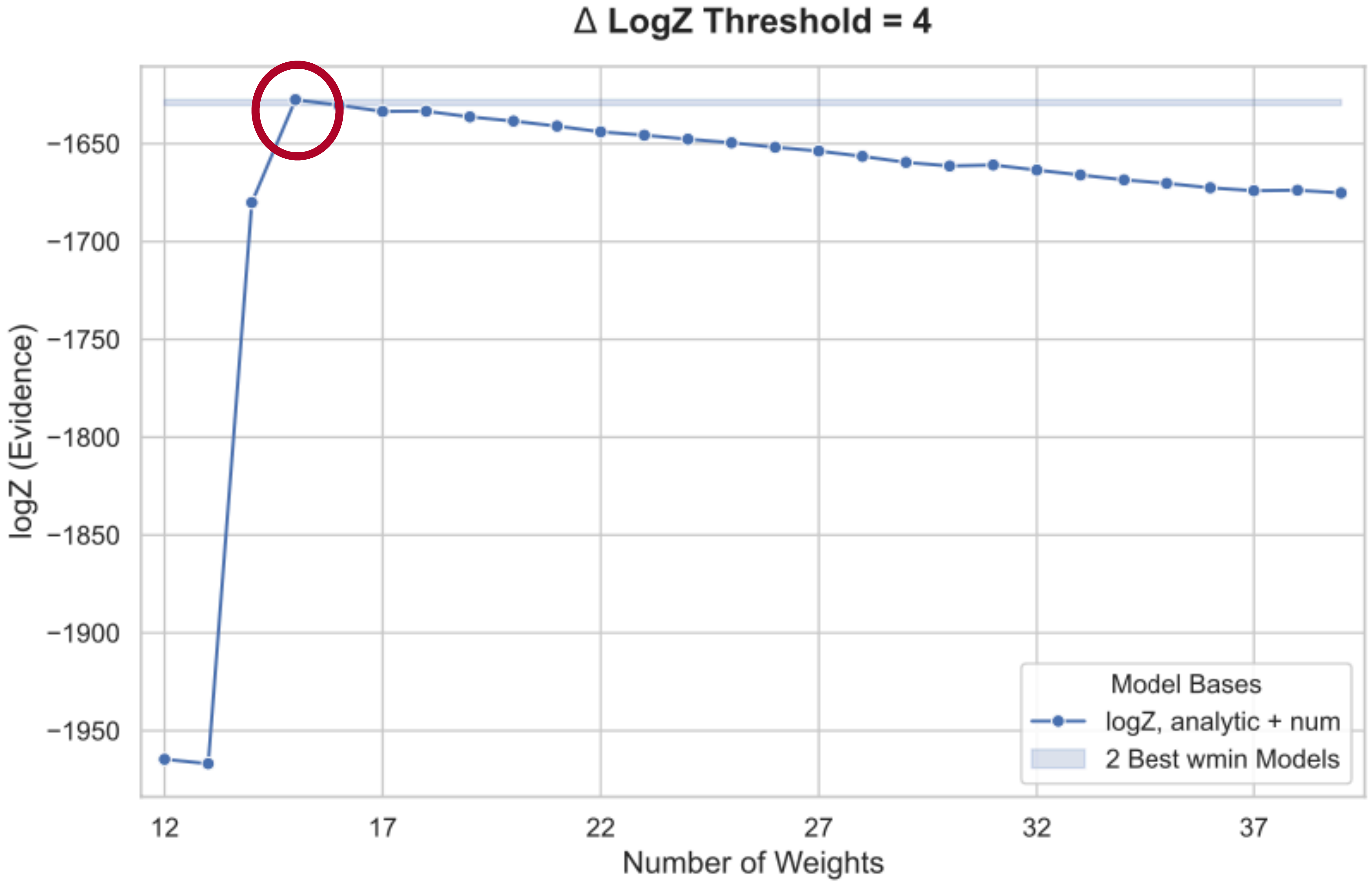
Models with N < 15 struggle to fit data

Over-parametrised models with N > 15 are penalised by the Occam volume factor



Level 1 closure test





Evidence “tells us” what the required flexibility of the parametrisation needs to be given the data

Model Selection

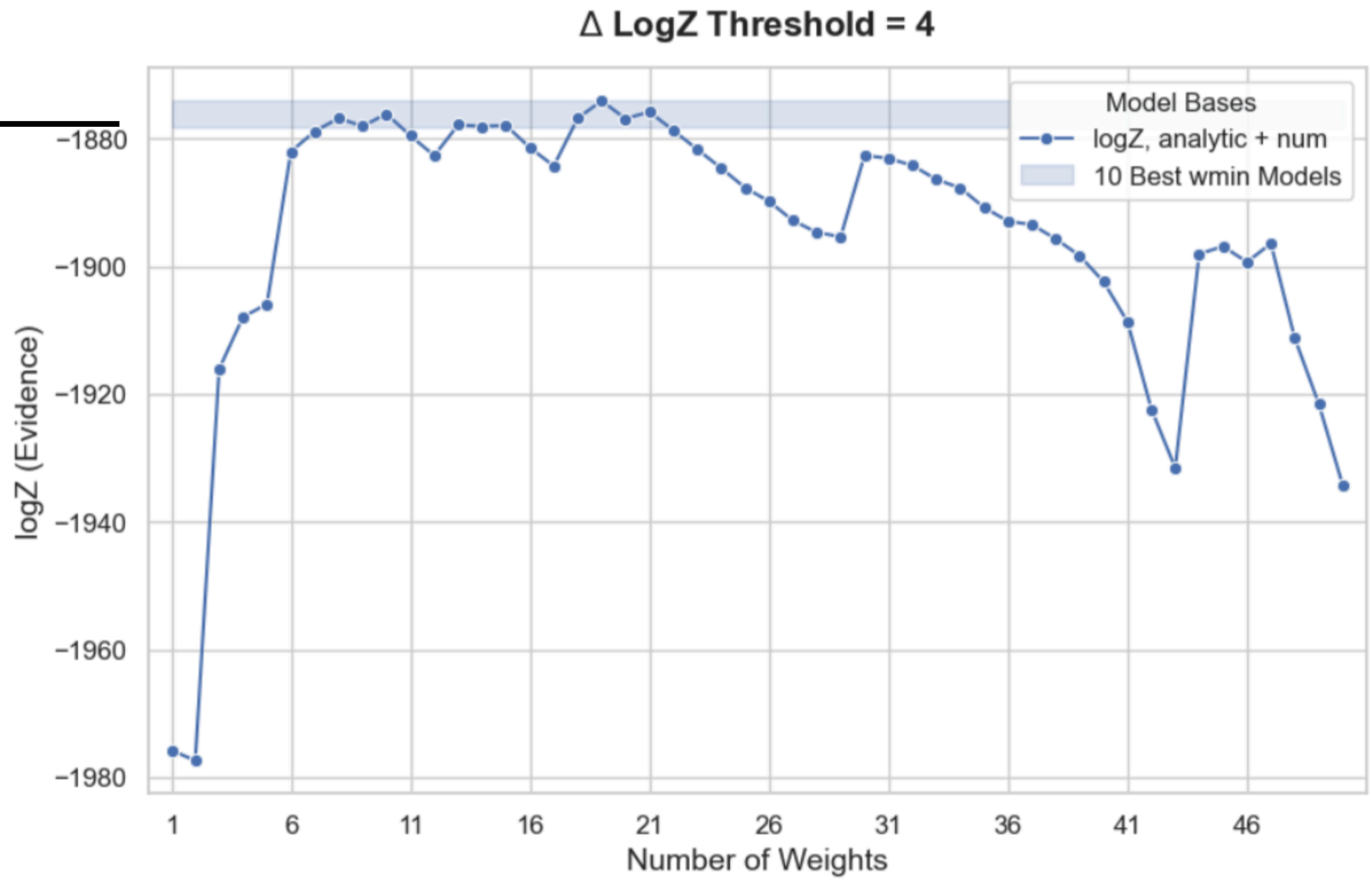
Analytic fit with uniform prior U[-0.6, 0.6]

Results are POD basis-dependent

BMA on 10 models within

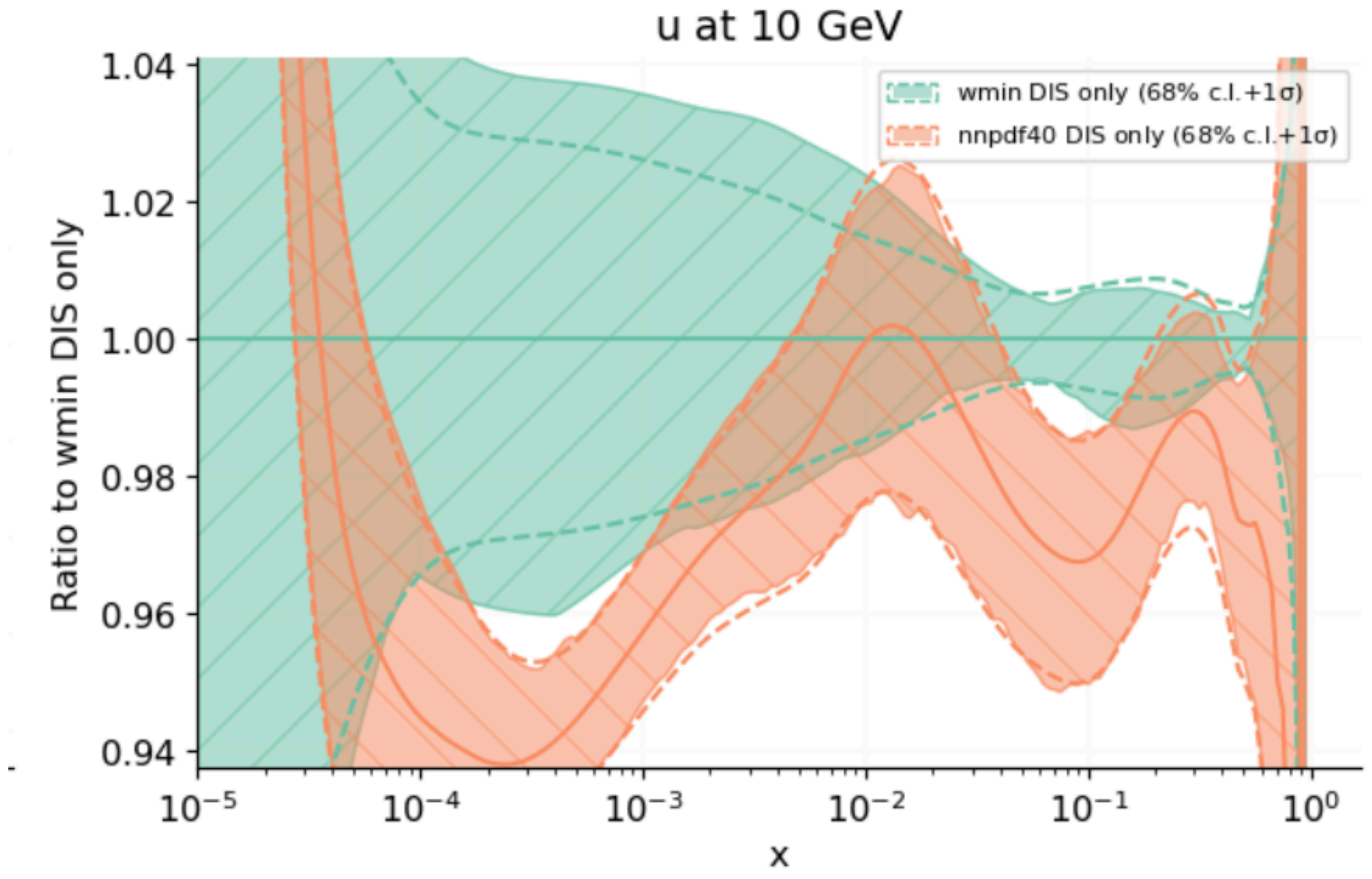
$\Delta \ln Z = 4.$

Model with highest evidence has 19 parameters

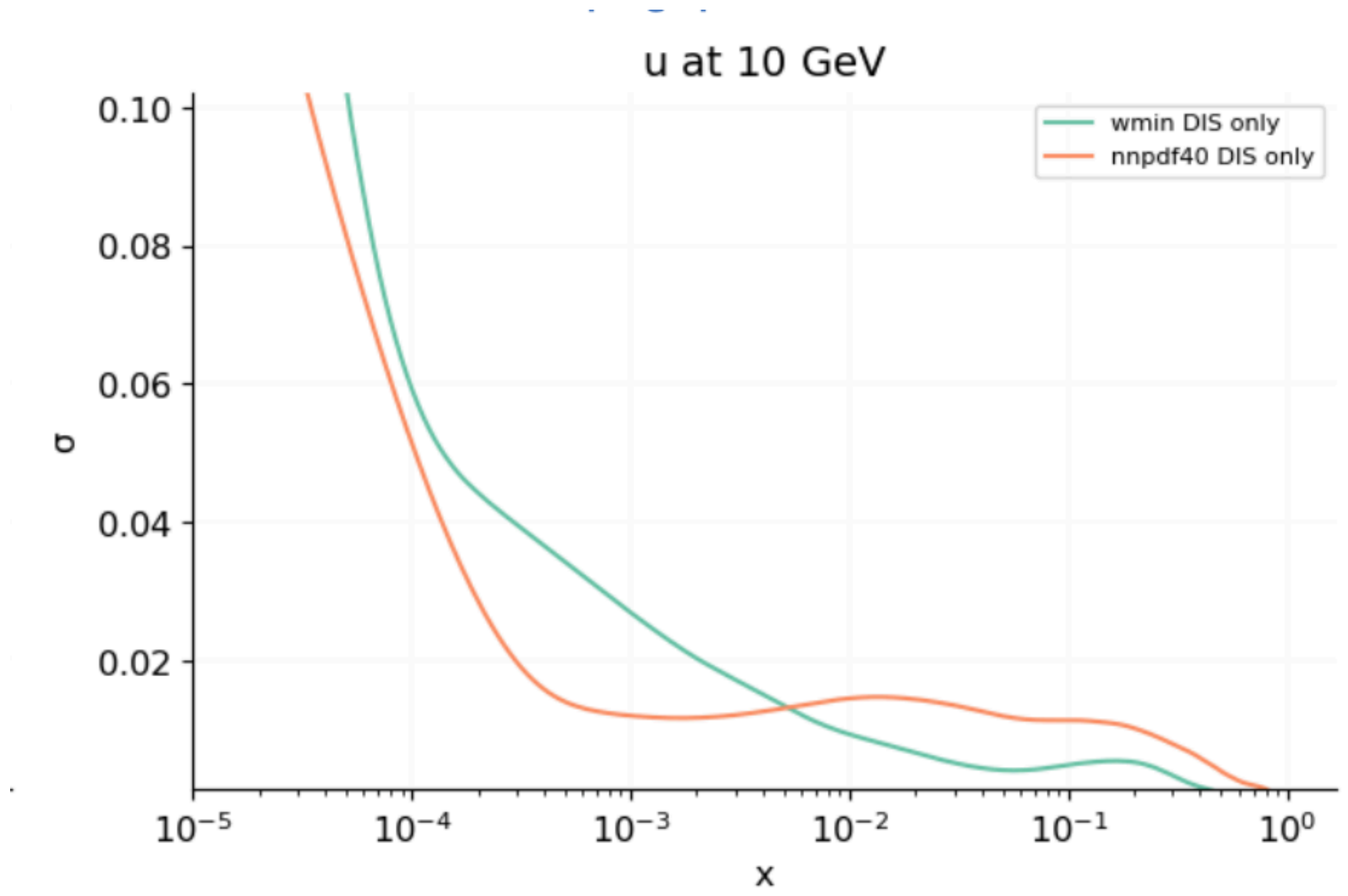


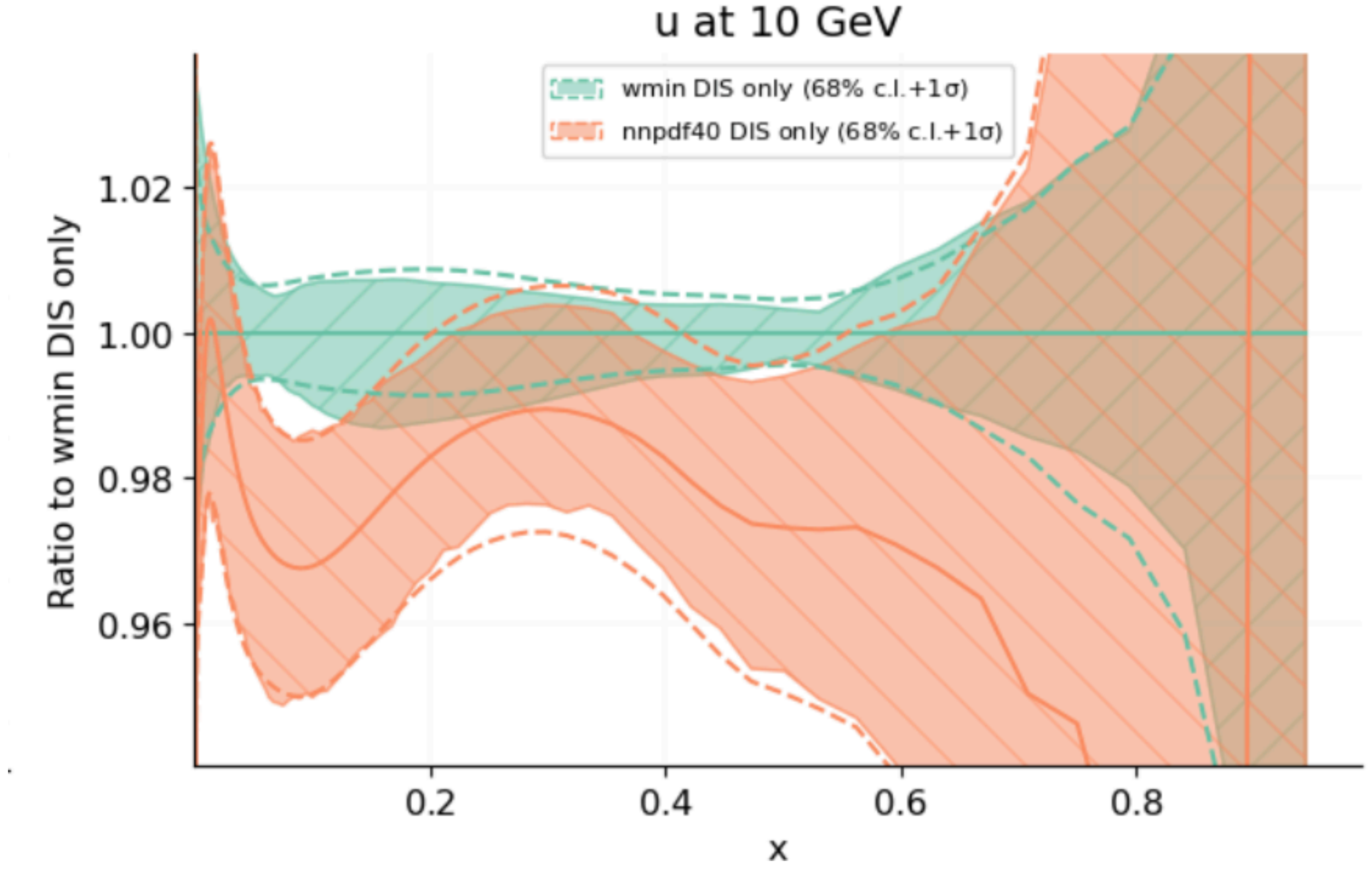
PDFs and Data-Theory

Comparison with NNPDF4.0 DIS-only

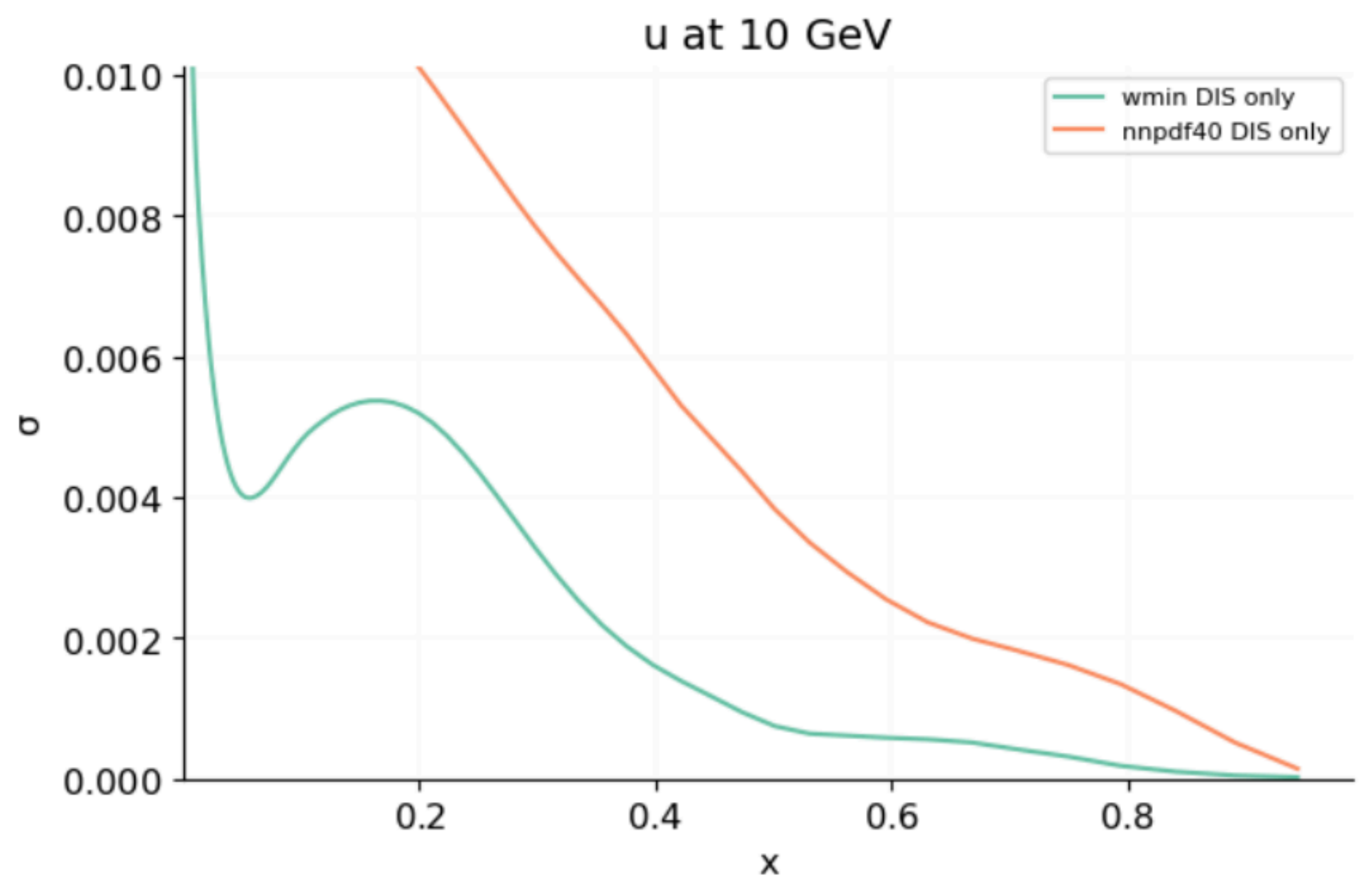


POD has similar uncertainties at small-x





POD has smaller uncertainties at large-x



Conclusions/Outlook

Conclusions / Outlook

- POD Parametrisation of PDFs: simple but effective parametrisation
- Bayesian Workflow: Bayesian model selection and average
- Benchmark of the methodology: closure tests

- Study better the dependence of the results on the POD basis
- Find alternative, data - independent, methods to construct an efficient POD basis

Backup

Proper Orthogonal Decomposition

Finite-Dimensional case, Singular Value Decomposition (SVD)

Construct a “dataset" that is supposed to represent well the space of all possible PDFs

E.g. Given a MC replica $f_\alpha^{(k)}(x_i, Q)$ set such as NNPDF4.0

$$X_{lk} \equiv f_\alpha^{(k)}(x_i, Q) - f_\alpha^{(0)}(x_i, Q)$$

$\alpha \in \{1,\ldots,N_f\}$

$i \in \{1,\ldots,N_x\}$

$l \in N_x(\alpha - 1) + i \;\; , k \in \{1,\ldots,N_{rep}\}$

In the finite-dimensional case POD reduces to the SVD+PCA of the given “dataset"

Same procedure as the one used to find a Hessian representation of an MC set, except that we don’t need the normalisation term $\sqrt{N_{rep} - 1}$ [1602.00005]

Nested Sampling

General Idea

- Monte Carlo algorithm for computing an integral over a model parameter space
- Nested Sampling provides both the posterior samples as well as the marginalised likelihood Z

Bayes Rule

$$P(\Theta\,|\,D) = \frac{L(D\,|\,\Theta)\pi(\Theta)}{Z}$$

Marginalised Likelihood

$$Z = \int L(D\,|\,\Theta)\pi(\Theta)d\Theta$$

Nested Sampling

Algorithm

1. Initialisation: sample randomly from the prior N live points and compute the Likelihood at each point
2. Shrinkage: remove point with the lowest likelihood L_1
3. Likelihood Restricted Prior Sampling: sample new point from prior with Likelihood $> L_1$
4. Iterate

Iteration i reduces integration volume by a factor $\delta V_i \approx \left(1 - \frac{1}{N}\right)^i \frac{1}{N}$,

The integral Z is simply $Z \approx \sum_i \delta V_i \times L_i$

Termination: when $\delta V_i \times L_i$ contributions to Z are negligible

Nested Sampling

Summary

1. It explores the parameter space globally;
2. it handles multi-modal distributions well;
3. it initialises and terminates at a well defined point -> no supervision;
4. it provides both marginal likelihood and posterior samples, hence allowing for Bayesian model selection

Choice of Prior

Uniform prior

$$f_{POD}(x, Q^2) = \xi_0(x, Q^2) + \sum_{i=1}^{N} w_i \left(\xi_i(x, Q^2) - \xi_0(x, Q^2)\right)$$

f_{POD} is linear in the w_i parameters $\rightarrow$ uniform prior in w_i results in uniform prior in f_{POD} !

However, in certain cases we have a much better choice

Choice of Prior

Bayesian Update

However, in certain cases we have a much better choice

$$d \sim \mathcal{N}(t(c), \Sigma), \quad \text{with} \quad \Sigma = \Sigma_1 \oplus \Sigma_2, \, d = (d_1, d_2)$$

→ Fit on d_1 yields a conditional distribution: $p(c \,|\, d_1)$

$$p(c \,|\, d_1) = \frac{\pi(c) \;\exp(-\frac{1}{2}||d_1 - t_1(c)||^2_{\Sigma_1})}{\int dc \;\pi(c) \;\exp(-\frac{1}{2}||d_1 - t_1(c)||^2_{\Sigma_1})}$$

→ Fit model to d_2 using $p(c \,|\, d_1)$ as prior

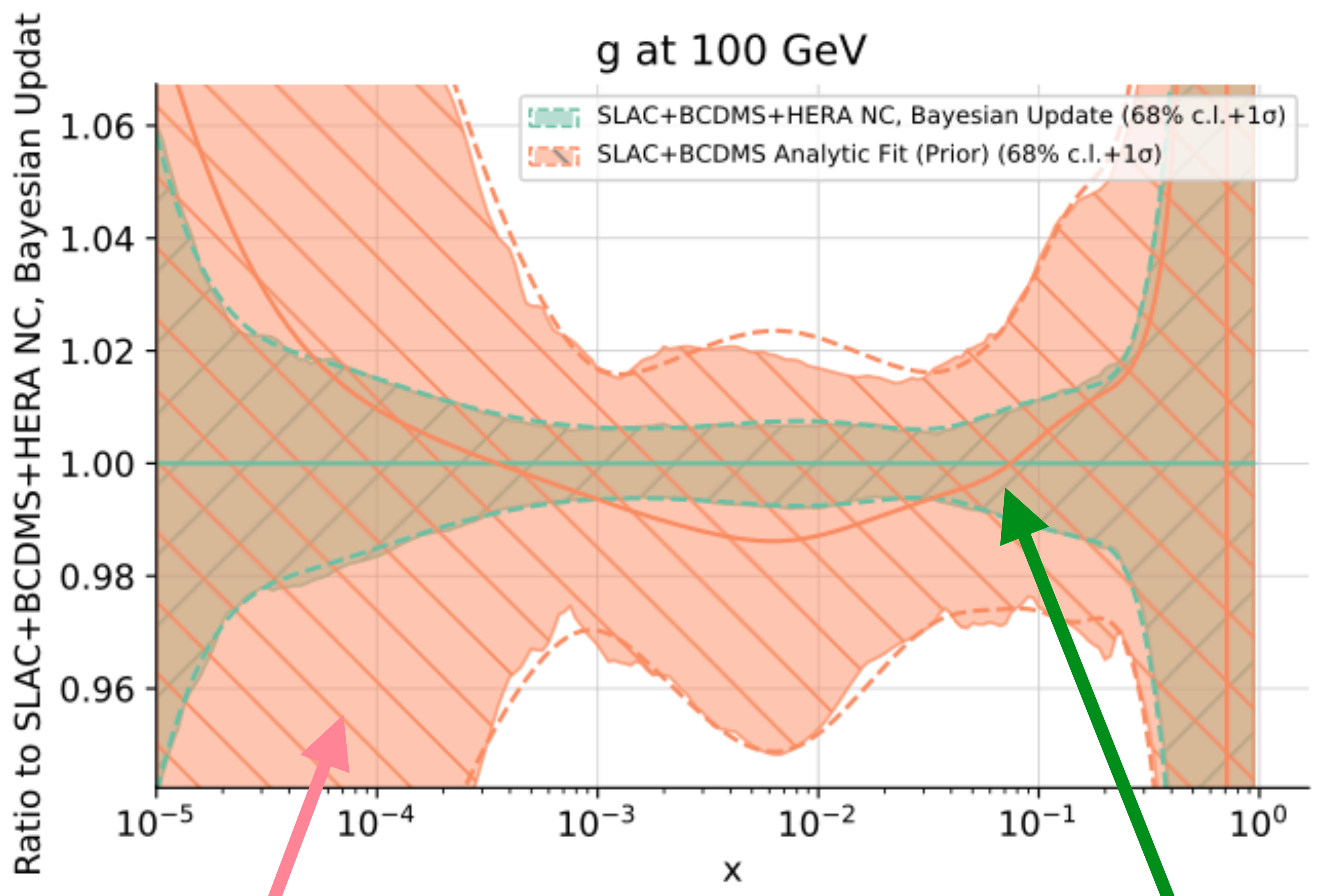
$$p(c \,|\, d_0) = \frac{p_{d_1}(c \,|\, d_1) \;\; \exp(-\frac{1}{2}||d_2 - t_2(c)||^2_{\Sigma_2})}{\int dc \;p_{d_1}(c \,|\, d_1) \;\; \exp(-\frac{1}{2}||d_2 - t_2(c)||^2_{\Sigma_2})}$$

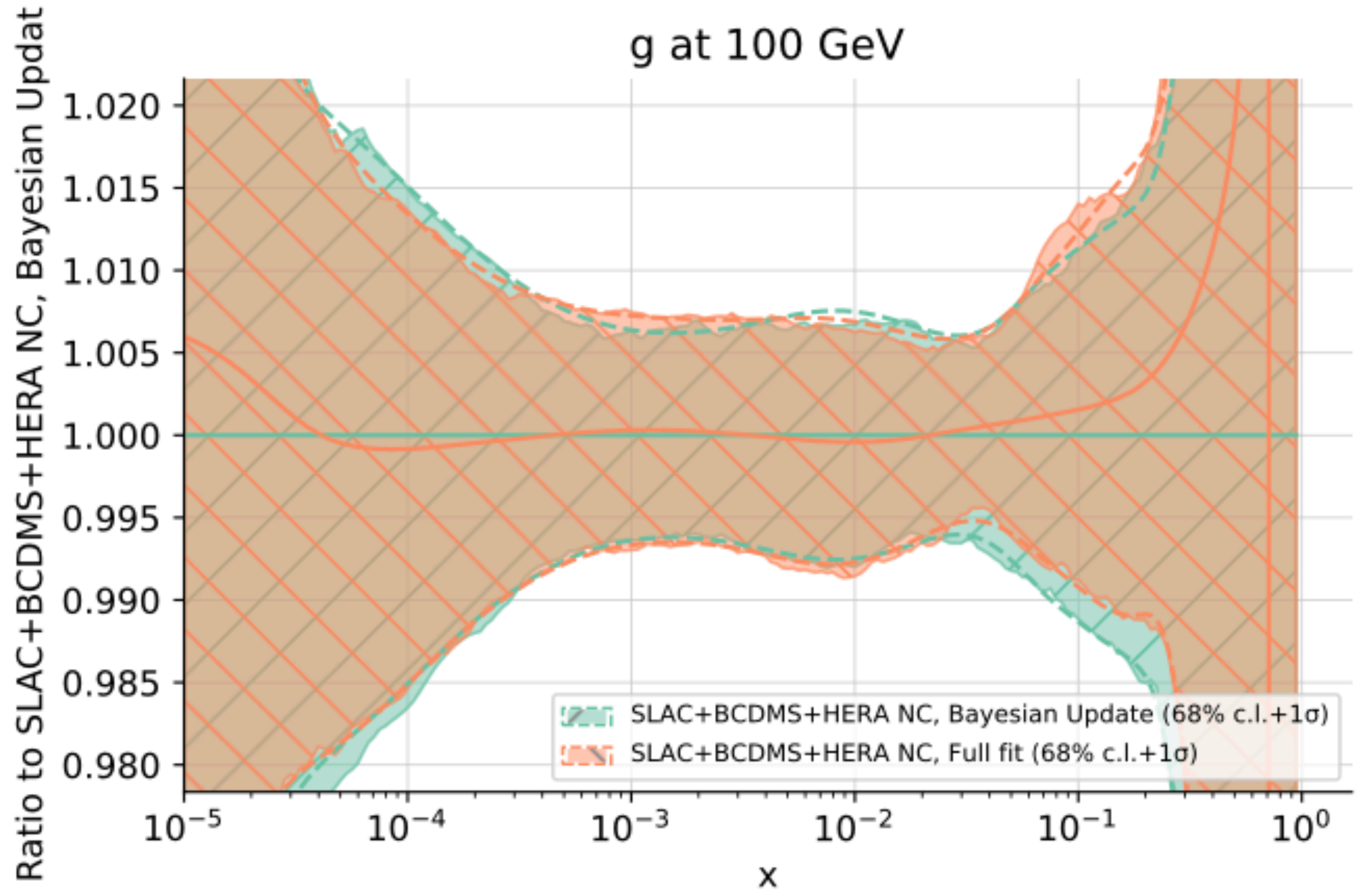
Choice of Prior

Bayesian Update

Example: consider data from (SLAC, BCDMS) + (HERA NC), (uncorrelated)

Fit SLAC + BCDMS first then use it as prior





Prior distribution (SLAC+BCDMS fit)

Full fit including HERA NC data

Comparison between bayesian update and uniform prior

Positivity Constraints

```
added_filter_rules:
  - process_type: POS_XPDF
    rule: "x > 0.1 and x < 0.74"

  - process_type: POS_DIS # affects structure functions
    rule: "x > 3e-05"
```

NNPDF40 Positivity

Fixed penalty term (Λ) set to ~ 3000

Impose cuts on DIS Pos sets: $x > 3e-05$

Impose cuts on MS bar PDFs Pos sets: $x > 0.1$, $x < 0.74$ ($N_{dat} = 1805$)

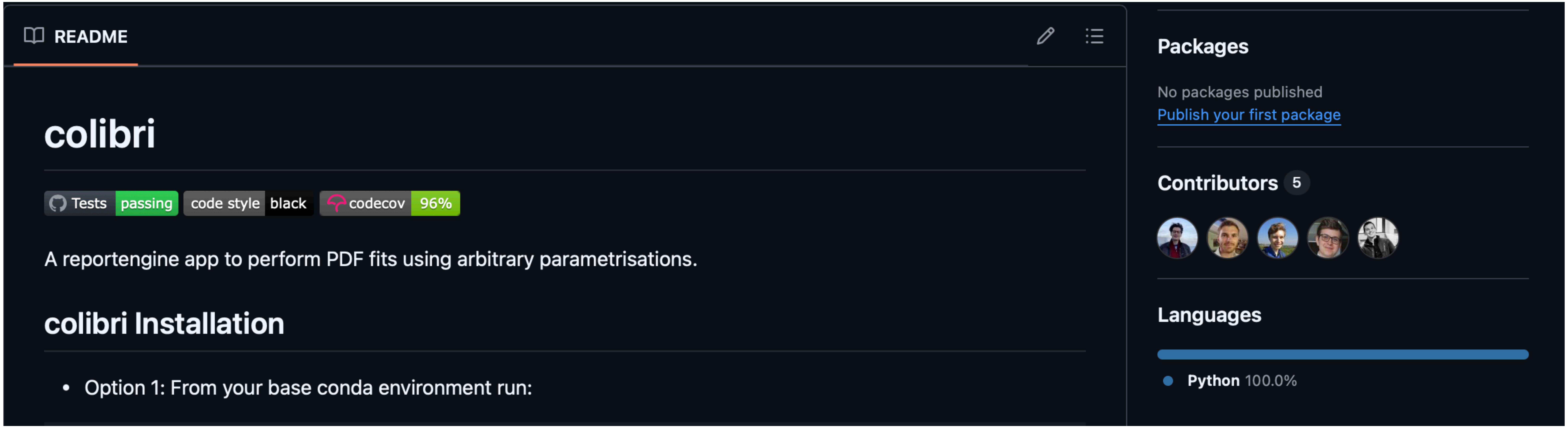
$N_{dat}(x > 0.74) = 55$

do not to impose any conditions in extrapolation region

$N_{dat}(x < 0.1) = 1210$

PDF MSbar POS argument breaks down at low x

Colibri



→Backbone: reportengine and validphys

→Makes use of Jax for high performance array computing (GPUs, JIT)

→Compatible with OpenMPI

Colibri
Fitting Routines

Analytic

- Only linear, very fast
- Computes bayesian metrics

Nested Sampling

- flexible prior choices
- bayesian model comparison
- use posterior as priors

Monte Carlo

Colibri

PDF Models

→Subpackage of Colibri inheriting all features (also form reportengine and validphys)

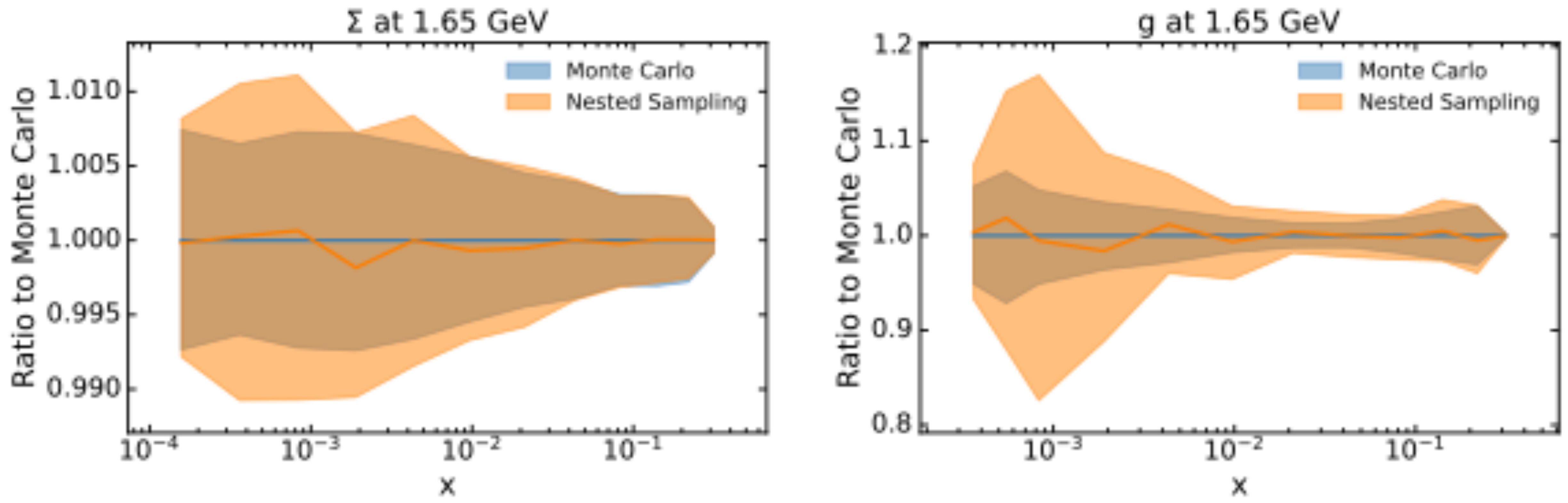
→Very flexible implementation of PDF Model (Abstract class in Colibri)

→A PDF model is a map F : params $\rightarrow PDF(N_{fl}, N_x)$

Colibri

PDF Models

→ grid pdf model used for the study [2404.10056]



→Gaussian Process, work in progress