Gaussian processes for PDF determination

Tommaso Giani

Based on Eur.Phys.J.C 84 (2024)
+
WIP with Maria Ubiali, Mark Constantini, Luigi Del debbio, Luca Mantani

PDF4LHC24, Cern, 03/12/2024

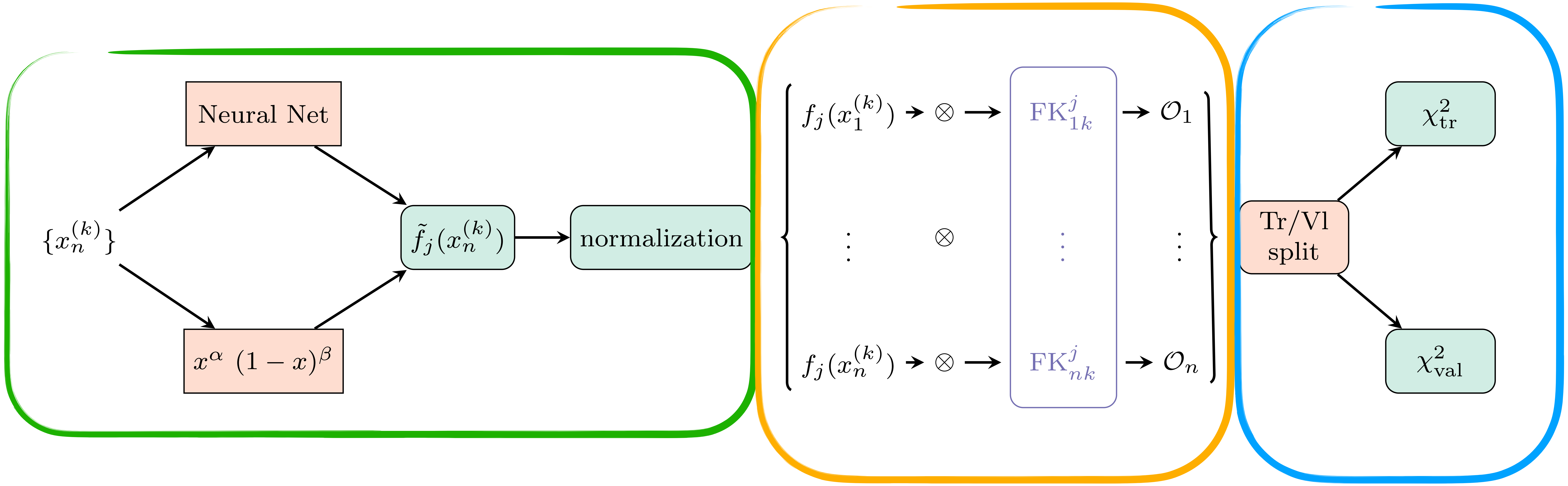
MAX-PLANCK-INSTITUT FÜR PHYSIK

Parametric regression

- PDFs are parametrised at some initial scale Q_0
- Theory prediction are computed as function of the free parameters
- Use data to build χ^2 and determine best fit

Error propagation:

- Montecarlo replicas
- Hessian approach



Non-parametric regression and Bayesian approach

- Start from a prior on the model $p(f)$
- Look at the data
- Get the posterior $p(f|D)$

Prior on the model

$$p(f|D) = \frac{p(D|f)\,p(f)}{p(D)}$$

Posterior of model given the data

- → Introduce probability distribution on a space of functions
- → Build a suitable prior
- → Use Bayes' theorem

Gaussian Processes

$$\mathbf{f} = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_N) \end{pmatrix} \in \mathbb{R}^N$$

Parameters $\mathbf{f}$: stochastic variables representing values of the PDF on a grid of points

Kernel $\mathbf{K}$ and mean function $\mathbf{m}$: functions modelling the correlation between parameters

$$m\left(x_i;\theta\right) = \mathrm{E}\left(f\left(x_i\right)\right)$$

$$k\left(x_i, x_j;\theta\right) = \mathrm{cov}\left(f\left(x_i\right), f\left(x_j\right)\right)$$

Hyperparameters θ: set of parameters entering the definition of the kernel (they control some specific feature of the prior)

Joint probability distribution of $\mathbf{f}$ and θ: target of the analysis

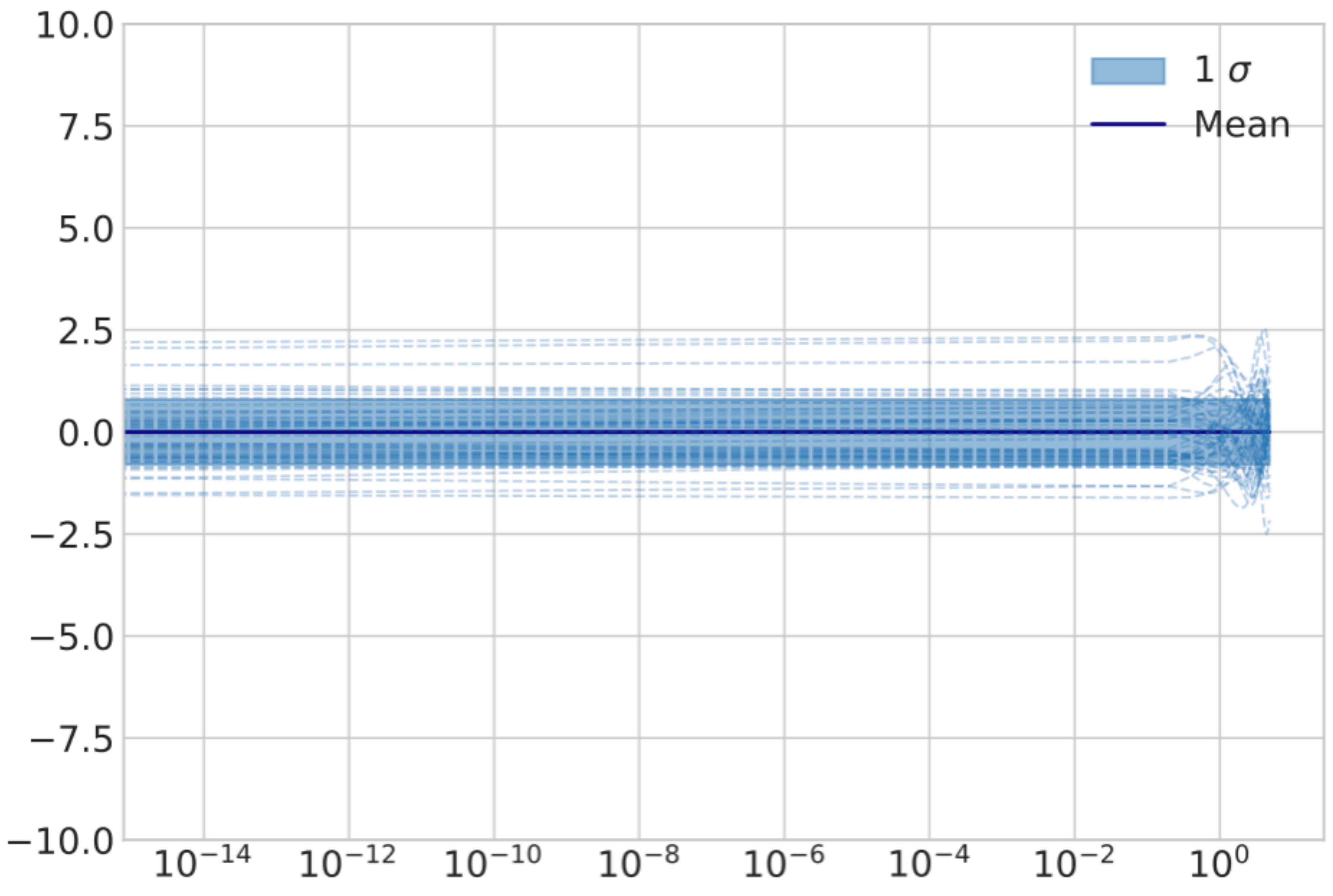
$$p\left(\mathbf{f}, \theta \,|\, \text{data}\right)$$

Prior for PDF

$$m(x) = 0$$

Exponential quadratic

$$k(x, y) = \sigma^2 \exp\left[-\frac{(x-y)^2}{l^2}\right]$$



Prior for PDF

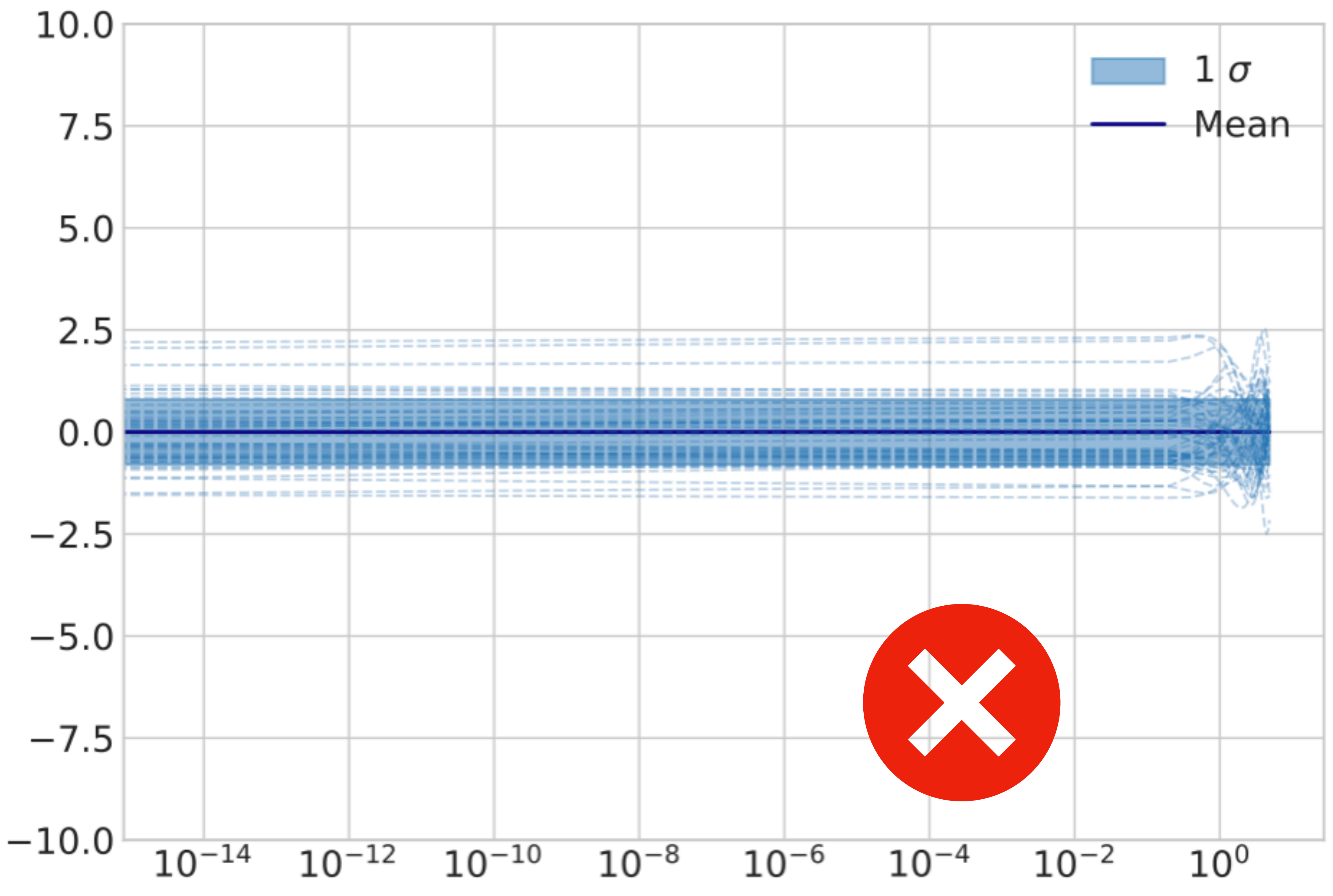
$$m(x) = 0$$

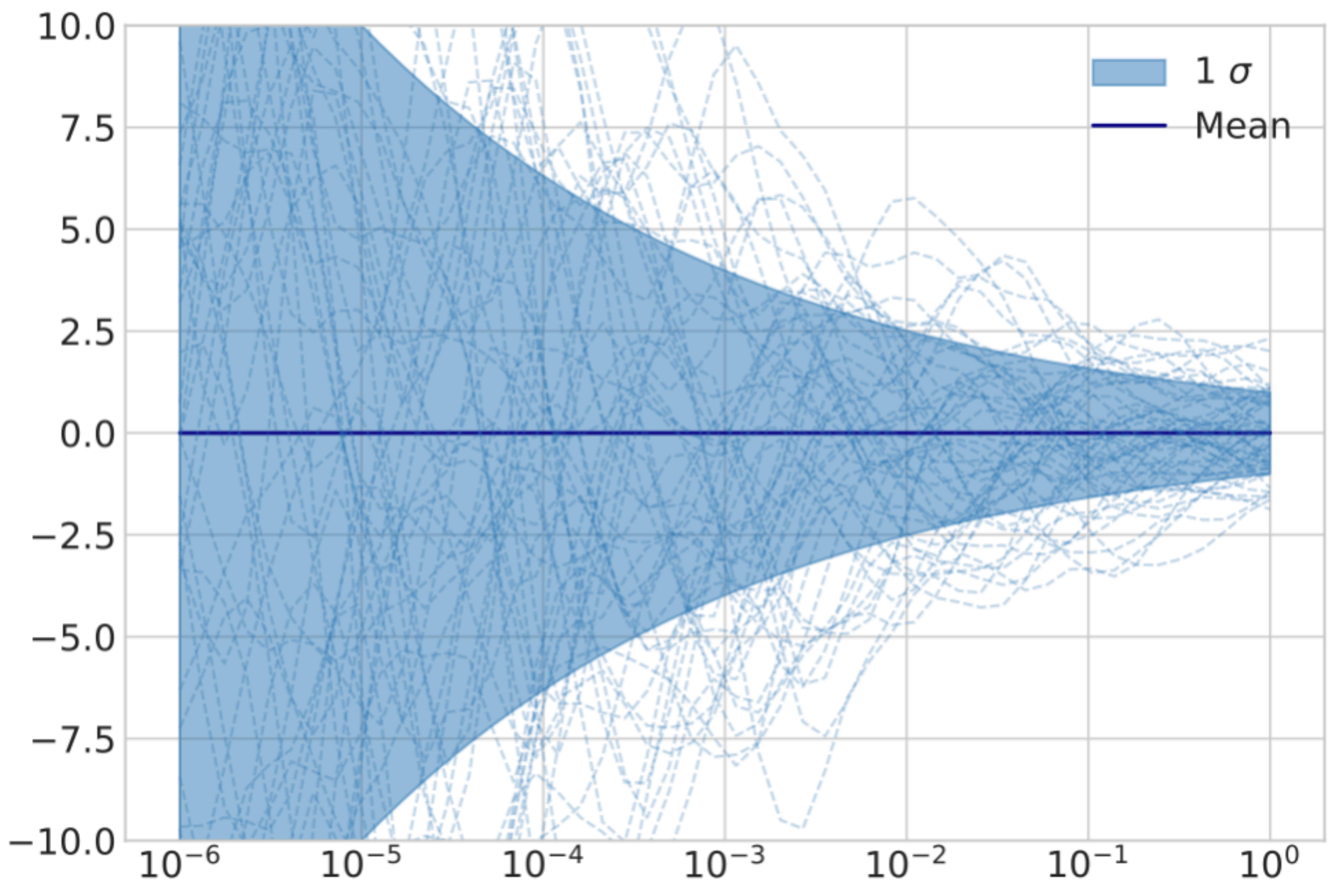
Exponential quadratic

$$k(x, y) = \sigma^2 \exp\left[-\frac{(x-y)^2}{l^2}\right]$$

Gibbs Kernel

$$\tilde{k}(x, y) \propto x^{\alpha} y^{\alpha} \quad \sigma^2 \exp\left[-\frac{(x-y)^2}{l^2(x) + l^2(y)}\right] \quad \text{with} \quad l(x) = l_0 x$$





Prior for PDF

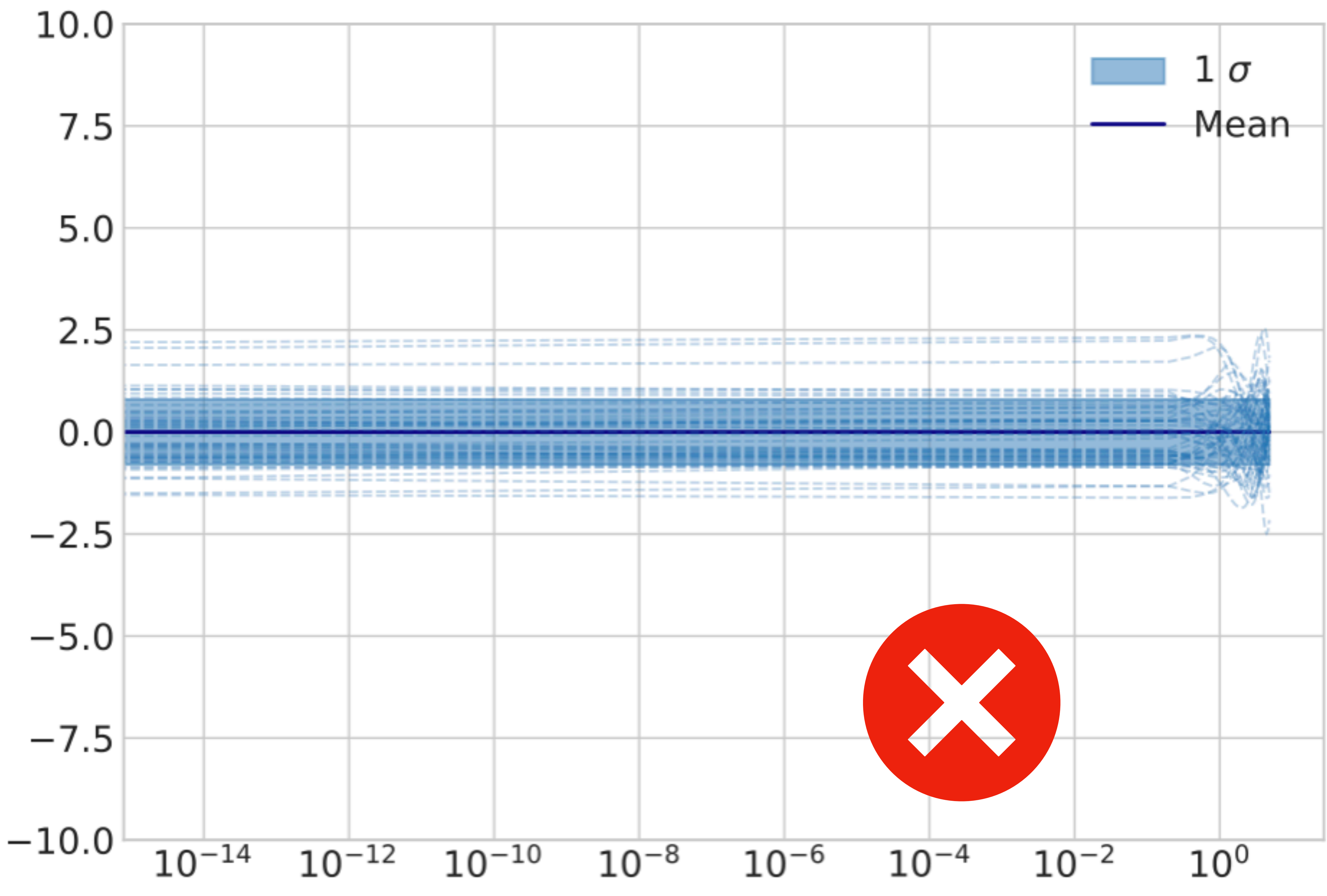
$$m(x) = 0$$

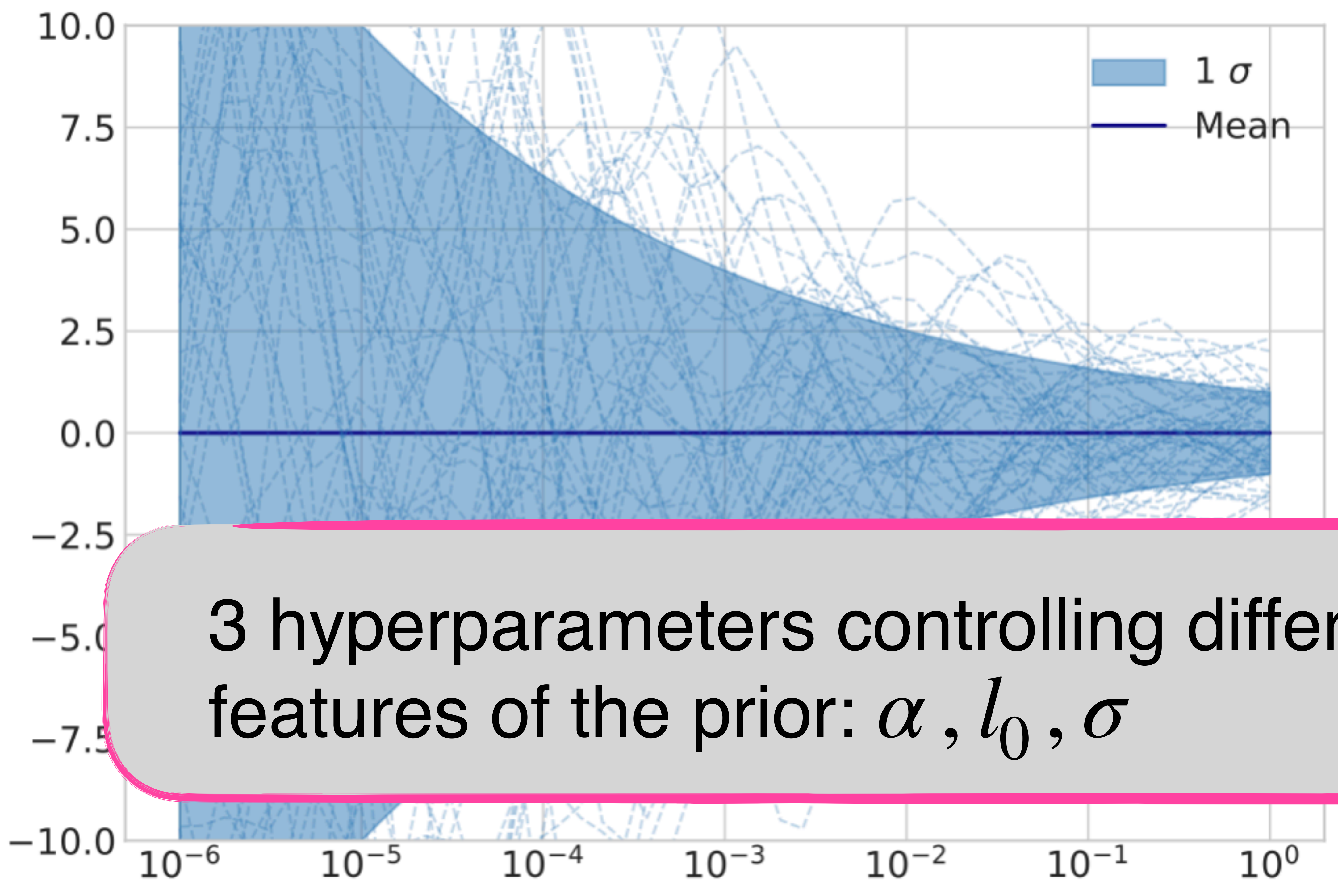
Exponential quadratic

$$k(x,y) = \sigma^2 \exp\left[-\frac{(x-y)^2}{l^2}\right]$$

Gibbs Kernel

$$\tilde{k}(x,y) \propto x^\alpha y^\alpha \quad \sigma^2 \exp\left[-\frac{(x-y)^2}{l^2(x)+l^2(y)}\right]$$ with $l(x) = l_0\, x$





3 hyperparameters controlling different features of the prior: α, l_0, σ

Example: PDFs from DIS

Introduce an interpolation basis for f

$$F\left(x, Q^2\right) = \sum_i \int_x^1 dy\, C_i\left(\frac{x}{y}, \frac{Q}{\mu}, \alpha_s\right) f_i\left(y, \mu\right) \rightarrow F_i = \sum_\alpha (FK)_{i\alpha} f\left(x_\alpha\right) = FK\,\mathbf{f}$$

Gaussian variable representing PDF on interpolation points $\mathbf{x}$ $\mathcal{O} = FK\,\mathbf{f}$	$\mathbf{f}^*$ Gaussian variable representing PDF on any set of points $\mathbf{x}^*$	$K\left(x, y; \theta\right)$ Function modelling correlation	$y, \quad \epsilon \sim N(0, C_y)$ Data and corresponding experimental error

Gaussian inference (fixed hyperparameters)

Consider stochastic variable

$$\begin{pmatrix} \mathbf{f}^* \\ FK\,\mathbf{f} \end{pmatrix} \sim \mathcal{N}\left(0\,, \begin{pmatrix} K_{\mathbf{x}^*\mathbf{x}^*} & K_{\mathbf{x}^*\mathbf{x}}\,FK^T \\ FK\,K_{\mathbf{x}\mathbf{x}^*} & FK\,K_{\mathbf{x}\mathbf{x}}\,FK^T \end{pmatrix}\right)$$

Compute the distribution

$$p\left(\mathbf{f}^* \,|\, FK\,\mathbf{f} + \epsilon = y, \theta\right)$$

This is a gaussian distribution. Its mean and covariance can be computed analytically

$$\tilde{\mathbf{m}}^* = \mathbf{m} + K_{\mathbf{x}^*\mathbf{x}}FK^T\left(FK\,K_{\mathbf{x}\mathbf{x}}FK^T + C_y\right)^+(\mathbf{y} - \mathbf{m})$$

$$\tilde{K}^* = K_{\mathbf{x}^*\mathbf{x}^*} - K_{\mathbf{x}^*\mathbf{x}}FK^T\left(FK\,K_{\mathbf{x}\mathbf{x}}FK^T + C_y\right)^+FK\,K_{\mathbf{x}\mathbf{x}^*}$$

- We generate pseudo-data for NNPDF4.0 DIS datasets
(NNPDF4.0 as underlying law)

- Gibbs Kernel with fixed hyperparameters

$$\tilde{\mathbf{m}}^* = \mathbf{m} + K_{\mathbf{x}^*\mathbf{x}} FK^T \left(FK\, K_{\mathbf{x}\mathbf{x}} FK^T + C_y \right)^+ (\mathbf{y} - \mathbf{m})$$

$$\tilde{K}^* = K_{\mathbf{x}^*\mathbf{x}^*} - K_{\mathbf{x}^*\mathbf{x}} FK^T \left(FK\, K_{\mathbf{x}\mathbf{x}} FK^T + C_y \right)^+ FK\, K_{\mathbf{x}\mathbf{x}^*}$$

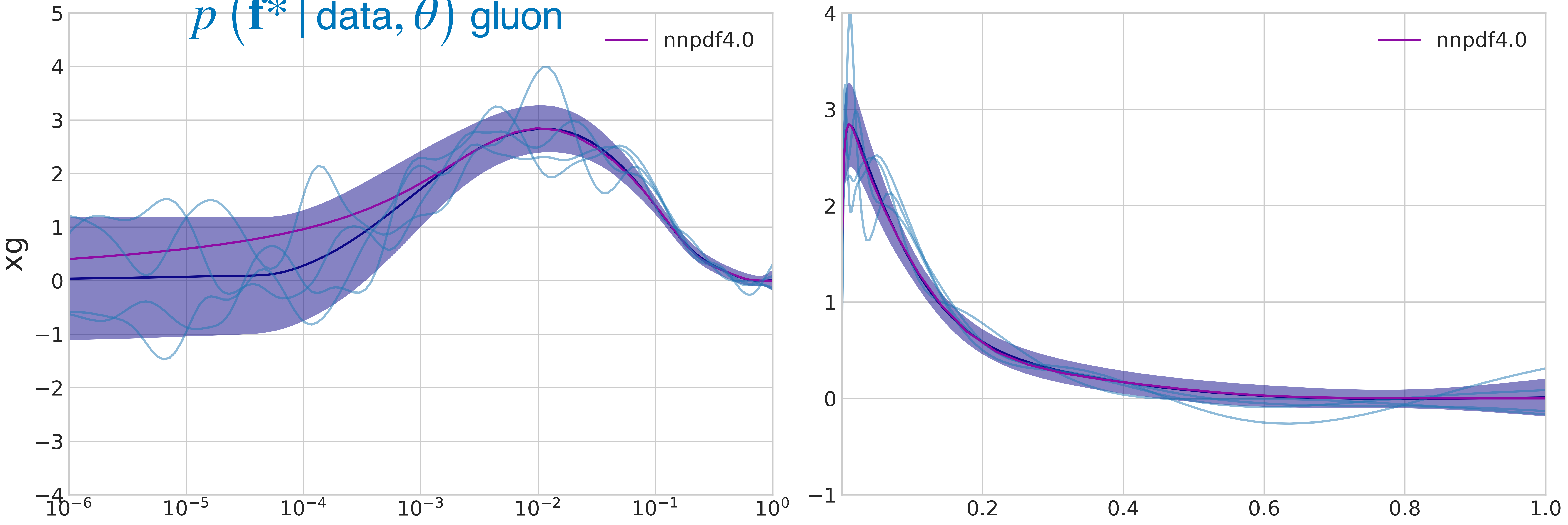
- We generate pseudo-data for NNPDF4.0 DIS datasets
(NNPDF4.0 as underlying law)

- Gibbs Kernel with fixed hyperparameters

$$\tilde{\mathbf{m}}^* = \mathbf{m} + K_{\mathbf{x}^*\mathbf{x}} FK^T \left(FK\, K_{\mathbf{xx}} FK^T + C_y \right)^+ (\mathbf{y} - \mathbf{m})$$

$$\tilde{K}^* = K_{\mathbf{x}^*\mathbf{x}^*} - K_{\mathbf{x}^*\mathbf{x}} FK^T \left(FK\, K_{\mathbf{xx}} FK^T + C_y \right)^+ FK\, K_{\mathbf{xx}^*}$$

$p\left(\mathbf{f}^* \,|\, \text{data}, \theta\right)$ gluon

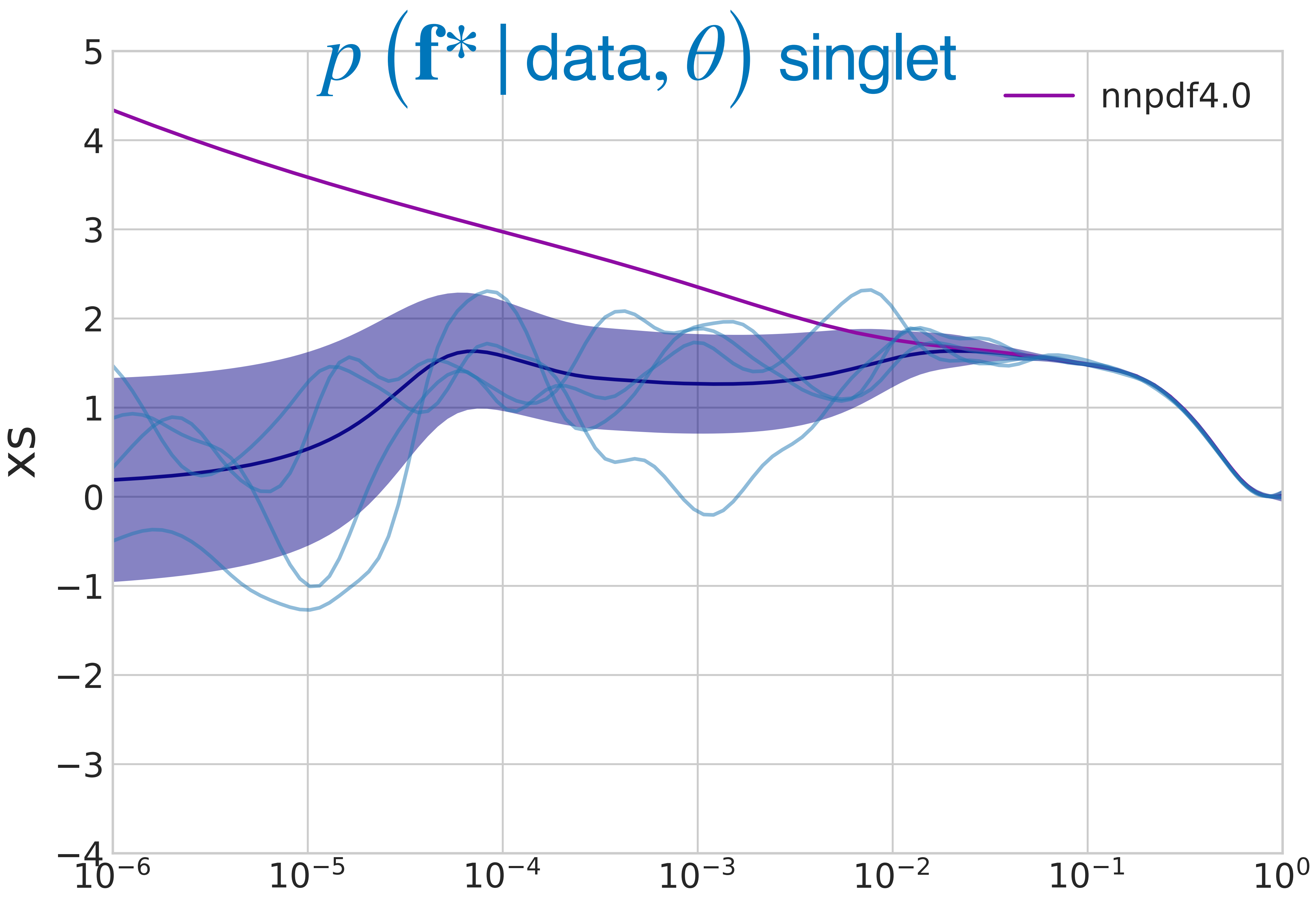


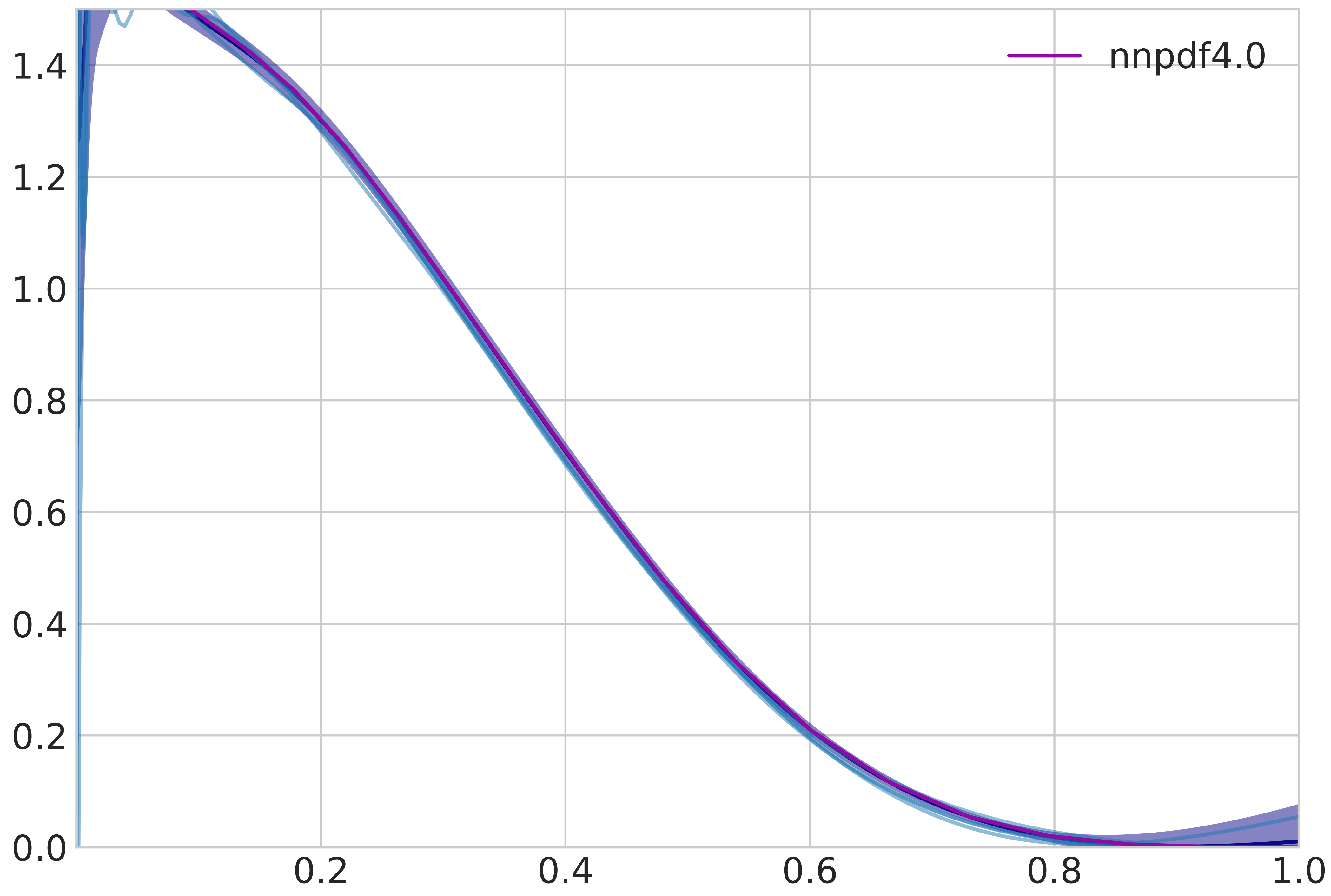
- We generate pseudo-data for NNPDF4.0 DIS datasets
  (NNPDF4.0 as underlying law)

- Gibbs Kernel with fixed hyperparameters

$$\tilde{\mathbf{m}}^* = \mathbf{m} + K_{\mathbf{x}^*\mathbf{x}} FK^T \left(FK\, K_{\mathbf{x}\mathbf{x}} FK^T + C_y \right)^+ (\mathbf{y} - \mathbf{m})$$

$$\tilde{K}^* = K_{\mathbf{x}^*\mathbf{x}^*} - K_{\mathbf{x}^*\mathbf{x}} FK^T \left(FK\, K_{\mathbf{x}\mathbf{x}} FK^T + C_y \right)^+ FK\, K_{\mathbf{x}\mathbf{x}^*}$$





Inference on the hyperparameters

Hyperparameters are also stochastic variables

Posterior on the hyperparamters given the data

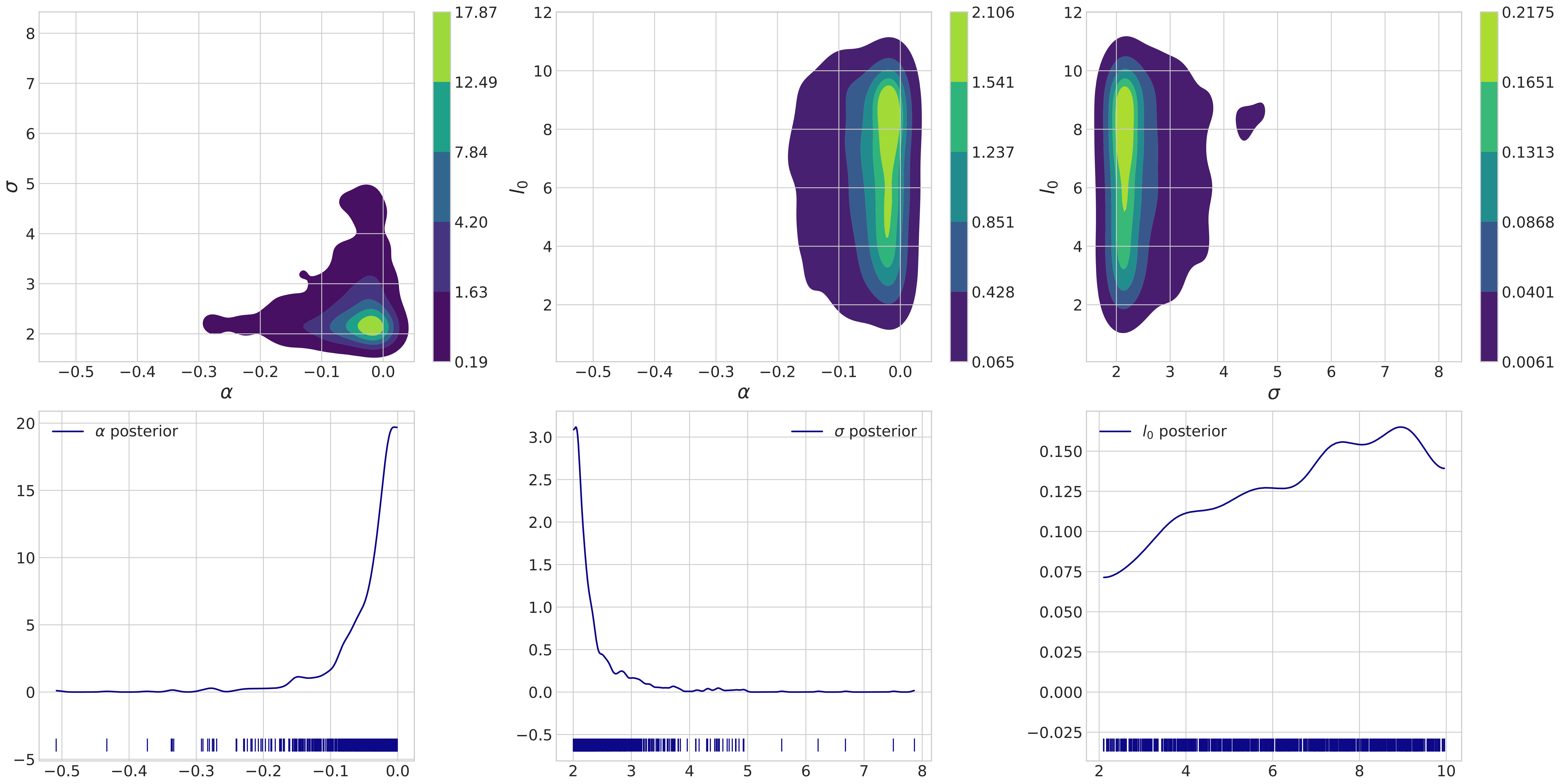
$$p\left(\mathbf{f}^{*}, \theta \,|\, \text{data}\right) = p\left(\mathbf{f}^{*} \,|\, \theta, \text{data}\right) p\left(\theta \,|\, \text{data}\right)$$

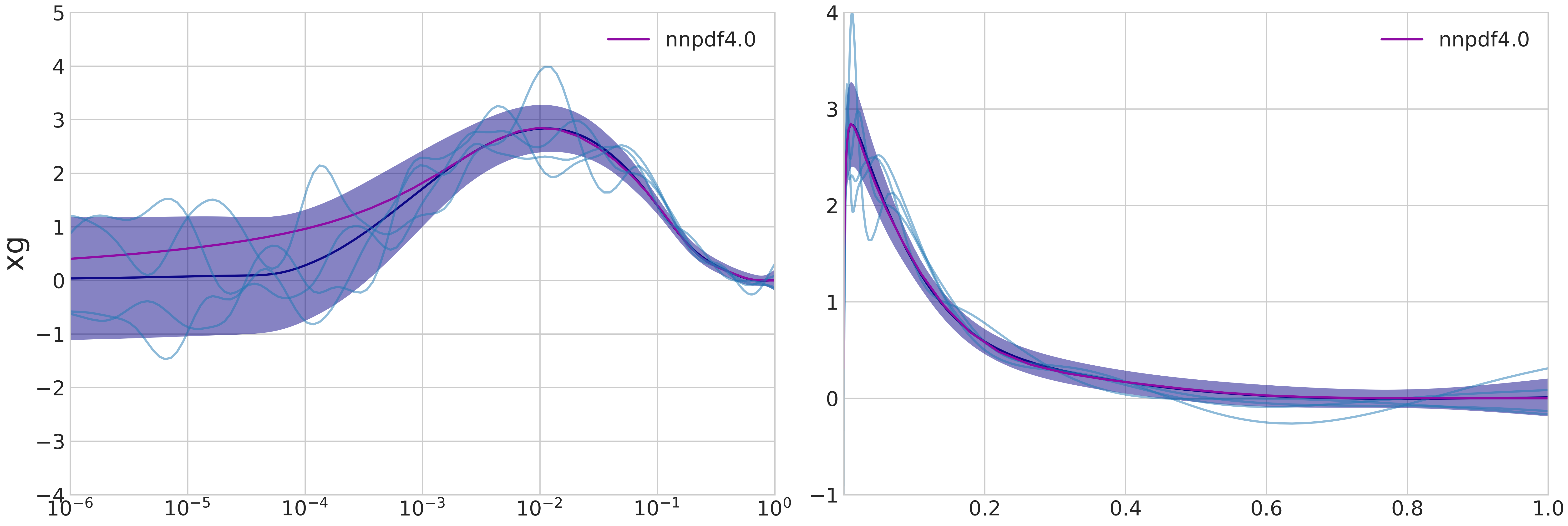
Joint probability distribution of $\mathbf{f}^{*}$ and θ

$$\propto p\left(\text{data} \,|\, \theta\right) p_{\theta}(\theta)$$

We can sample from $p\left(\theta \,|\, \text{data}\right)$ running a MCMC algorithm

Posterior for hyperparameters (gluon)

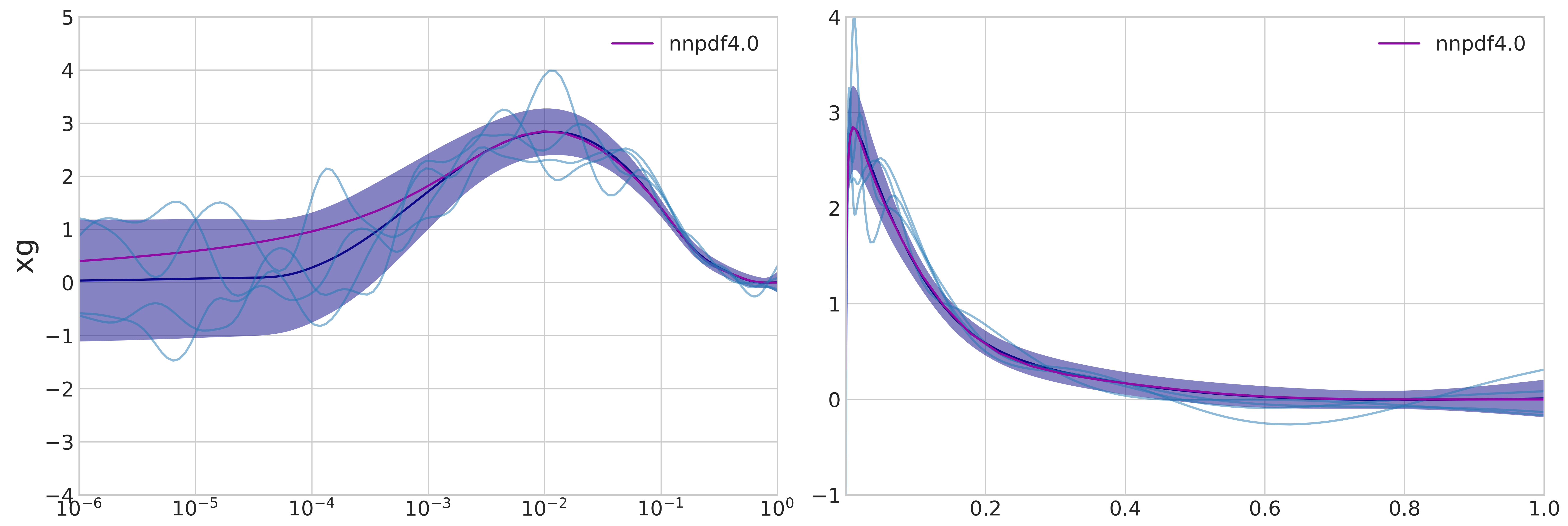




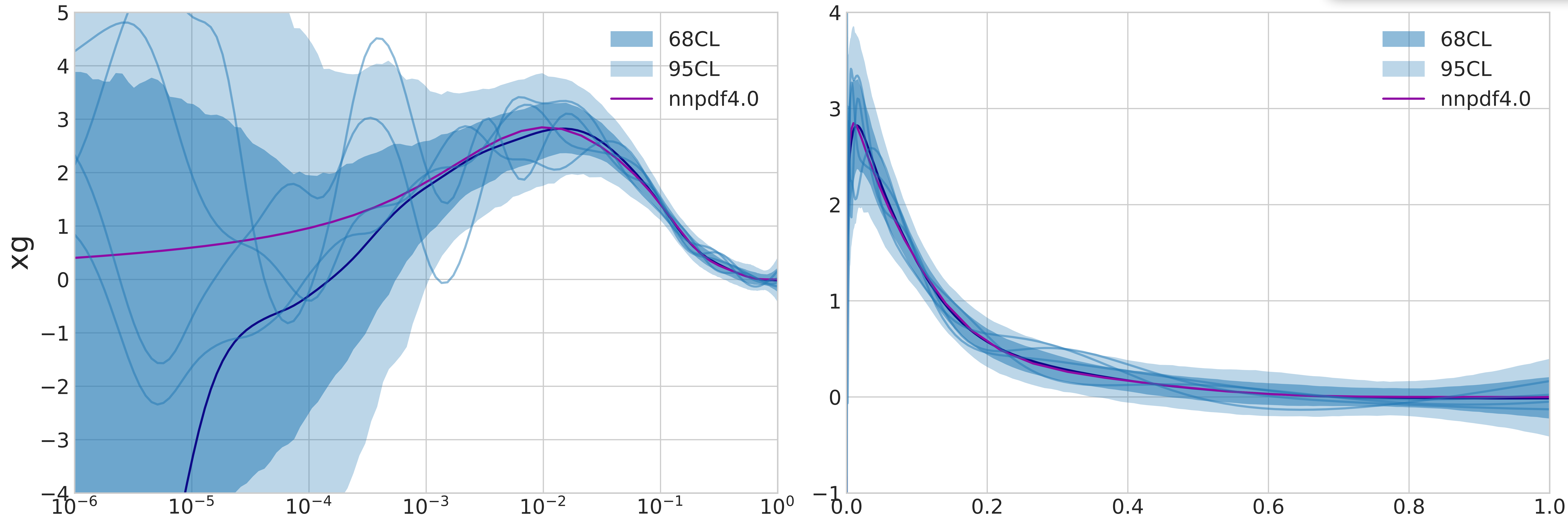
Fixed hyperparameters

$$p\left(\mathbf{f}^{*} \,|\, \text{data}, \theta\right)$$

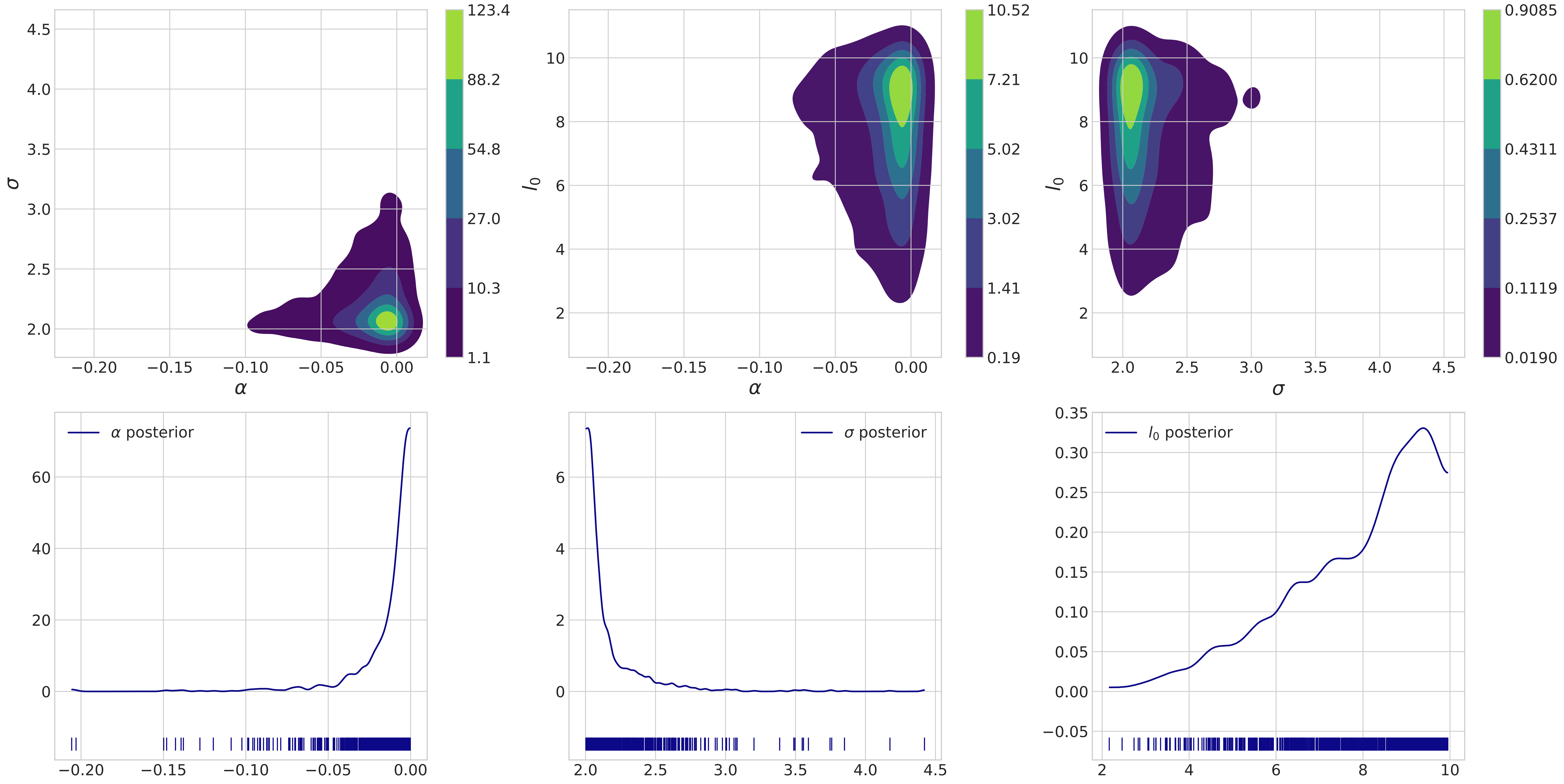
Fixed hyperparameters $p\left(\mathbf{f}^* \,|\, \text{data}, \theta\right)$

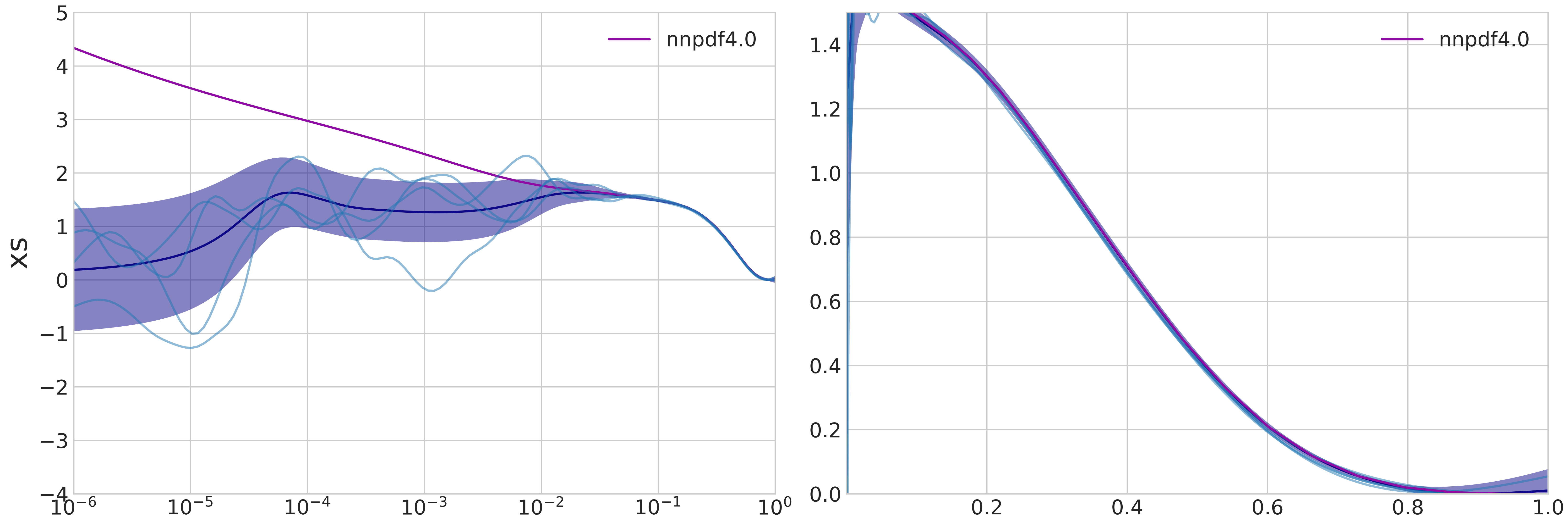


Full posterior $p\left(\mathbf{f}^*, \theta \,|\, \text{data}\right)$



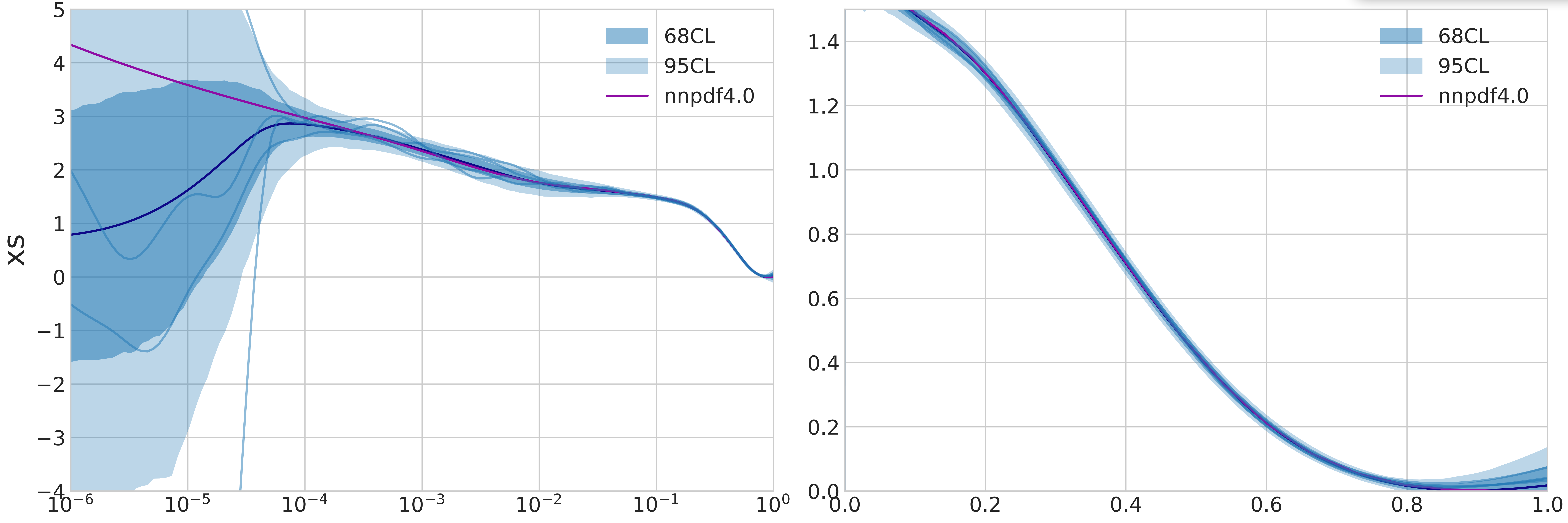
Posterior for hyperparameters (singlet)





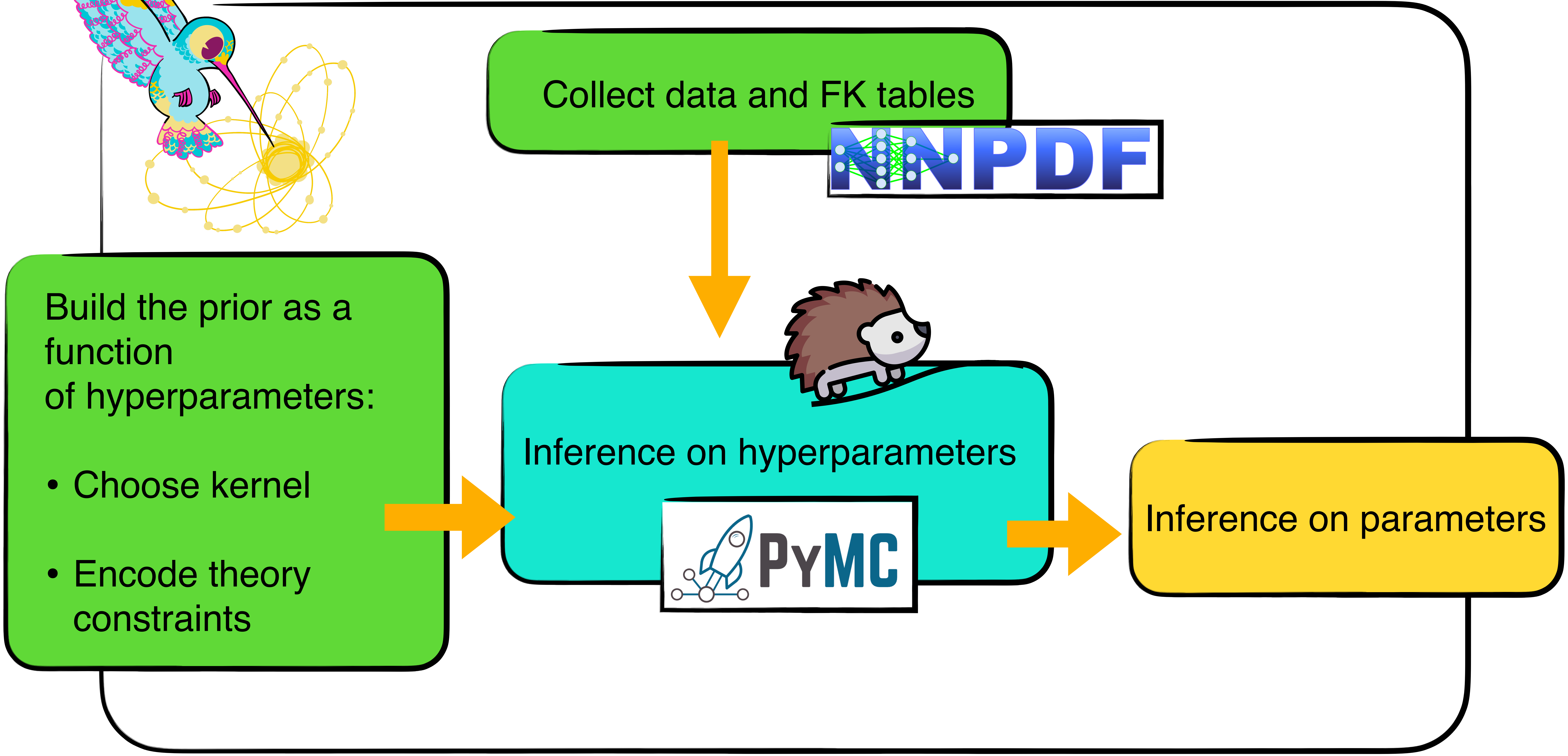
Fixed hyperparameters
$p\left(\mathbf{f}^* \,|\, \text{data}, \theta\right)$

Full posterior $p\left(\mathbf{f}^*, \theta \,|\, \text{data}\right)$



Workflow within colibri

WIP with Maria Ubiali, Mark Constantini, Luigi Del debbio, Luca Mantani



Theory constraints

Kinetic limit

$$f(1) = 0$$

Additional linear constrain on the PDF: can be implemented directly in the FK table

Sum rules

Sum rules can be implemented as additional linear constrains on the primitive of the PDF

$$g \sim GP\left(0, K\left(x, y\right)\right) \longrightarrow \frac{dg}{dx} \sim GP\left(0, \partial_x \partial_y K\left(x, y\right)\right)$$

Positivity, integrability ….

Penalty terms in the likelihood

Summary and future work

- Alternative methodology to fit PDFs, orthogonal to the ones currently used
- Well defined uncertainties
- Assumptions clearly defined in the prior
- Analytical understanding of what s going on during NN fits (NN dynamic, see Luigi's talk)

- Systematic study of different possible kernels
- Comparison with non Bayesian methodologies. Are there any differences?
- Implementation of a full global analysis

Artwork by @qftoons

Work in progress within colibri…

Backup slides

Global fits

$$\sigma = \sum_{i,j} \int dx_1 dx_2\, f_i\left(x_1, \mu\right) f_j\left(x_2, \mu\right) \hat{\sigma}\left(x_1, x_2, \frac{Q}{\mu}\right) \times \left(1 + \mathcal{O}\left(\Lambda/M\right)^p\right)$$

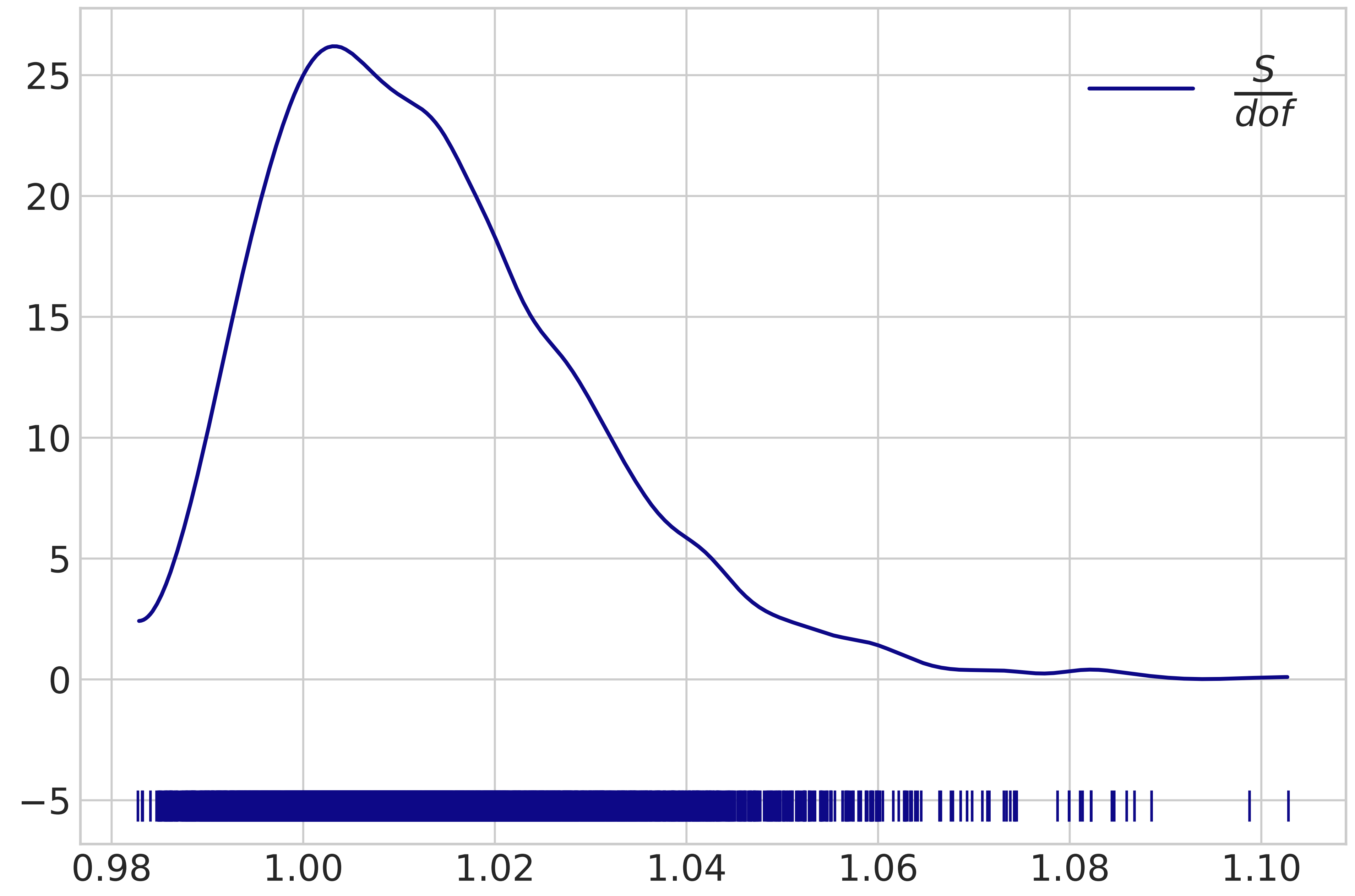
$$p\left(\mathbf{f}, \theta \,|\, \mathsf{data}\right) = p\left(\mathbf{f} \,|\, \mathsf{data}, \theta\right) p\left(\theta \,|\, \mathsf{data}\right)$$

This bit is not a gaussian distribution any longer

To access the posterior we have to run a MCMC having dimension $\dim \mathbf{f} + \dim \theta$

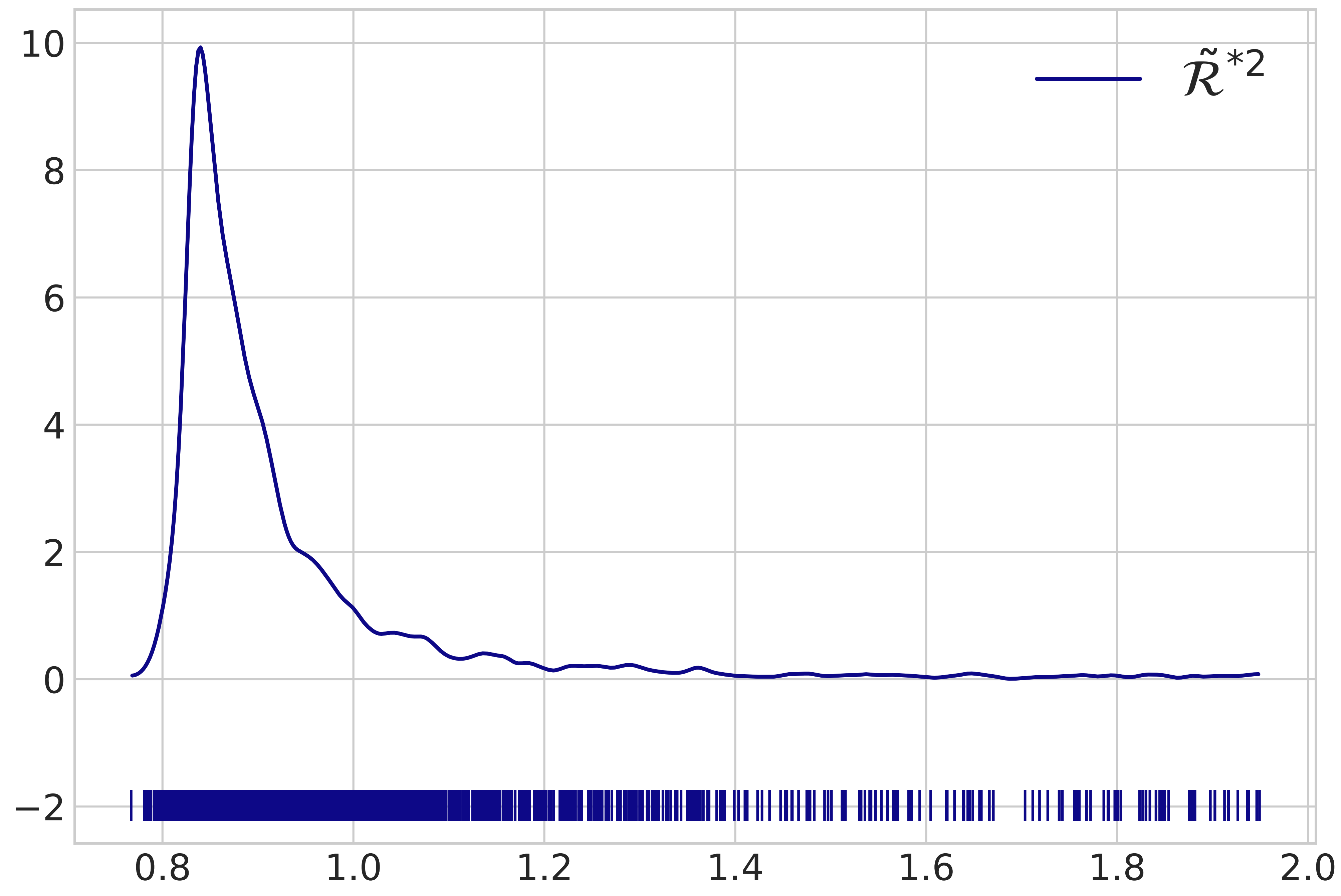
Fit quality

$$\frac{S}{dof} = \frac{1}{N_{\text{data}}}\Big((\mathbf{m}-\tilde{\mathbf{m}})^T K_{xx}^{-1}(\mathbf{m}-\tilde{\mathbf{m}}) + (y - FK\,\tilde{\mathbf{m}})^T C_Y^{-1}(y - FK\,\tilde{\mathbf{m}})\Big)$$



Generalisation on unseen data

$$\tilde{\mathscr{R}}^{*2} = \frac{1}{\dim(y^* \,|\, y)}(FK^*\,\tilde{\mathbf{m}} - y^*)^T\Big(FK^*\,\tilde{K}_{xx}FK^{*T} + C_Y^*\Big)^+(FK^*\,\tilde{\mathbf{m}} - y^*)$$



Decomposition of PDF uncertainty

$$\tilde{K} = \left(I - R_{xx}\right) K_{\mathbf{xx}} \left(I - R_{xx}\right)^T + a_{xx}^T C_y a_{xx}$$

Methodology — Experimental error

$$a_{xx}^T = K_{\mathbf{xx}} FK^T \left(FK\, K_{\mathbf{xx}} FK^T + C_y \right)^+$$

$$R_{xx} = a_{xx}^T\, FK$$