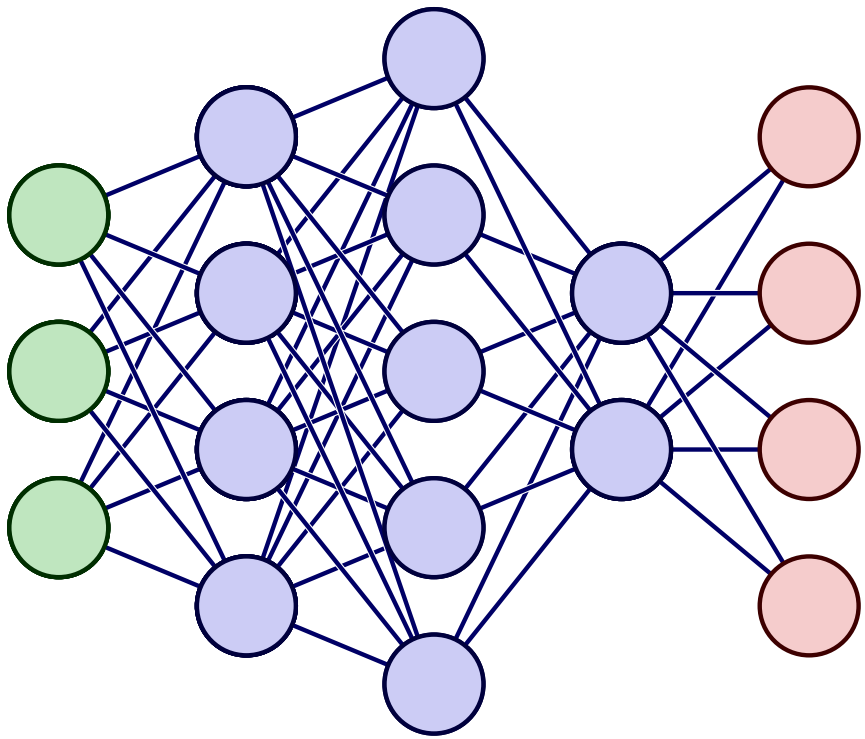
Stochastic Models of ML

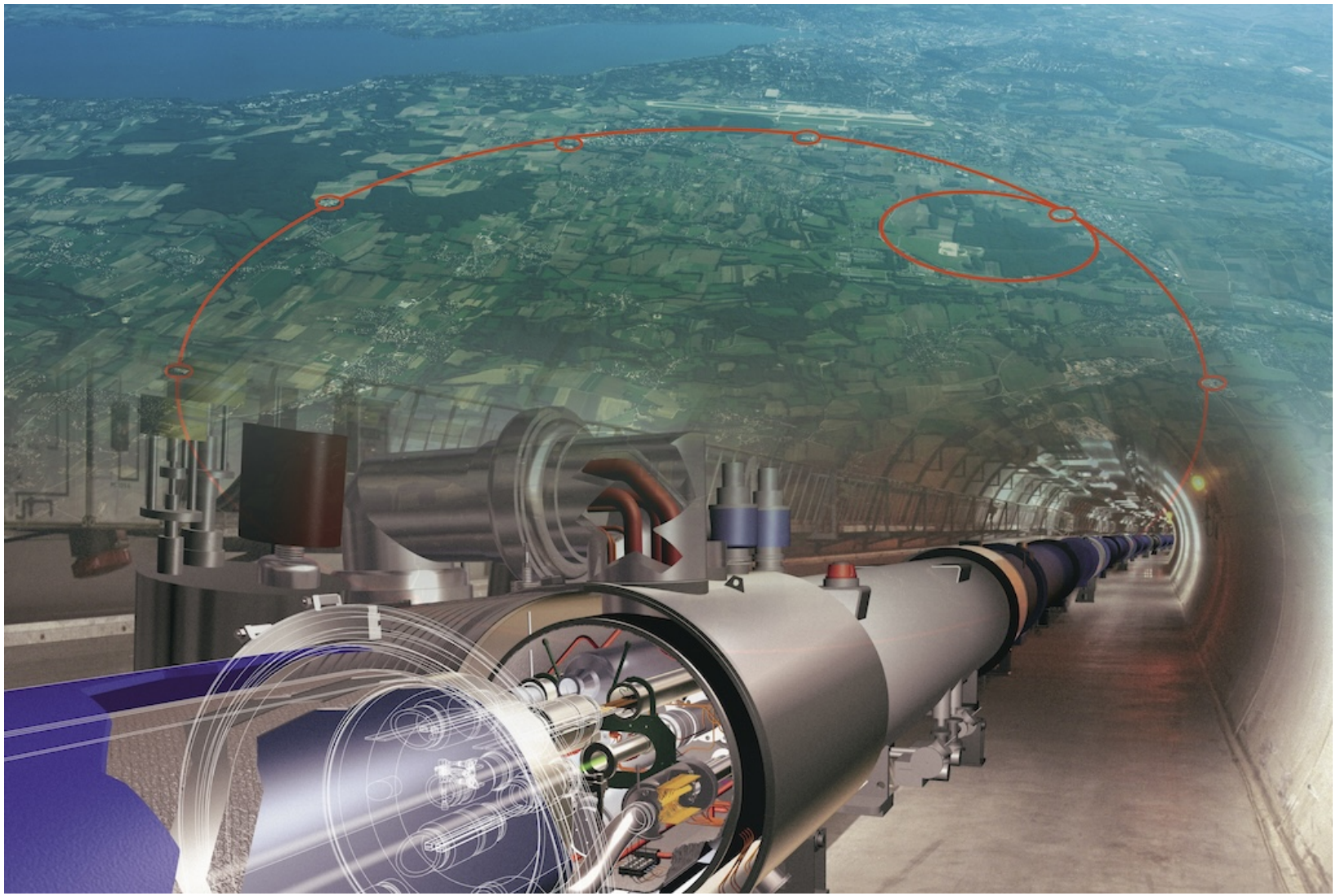


L Del Debbio

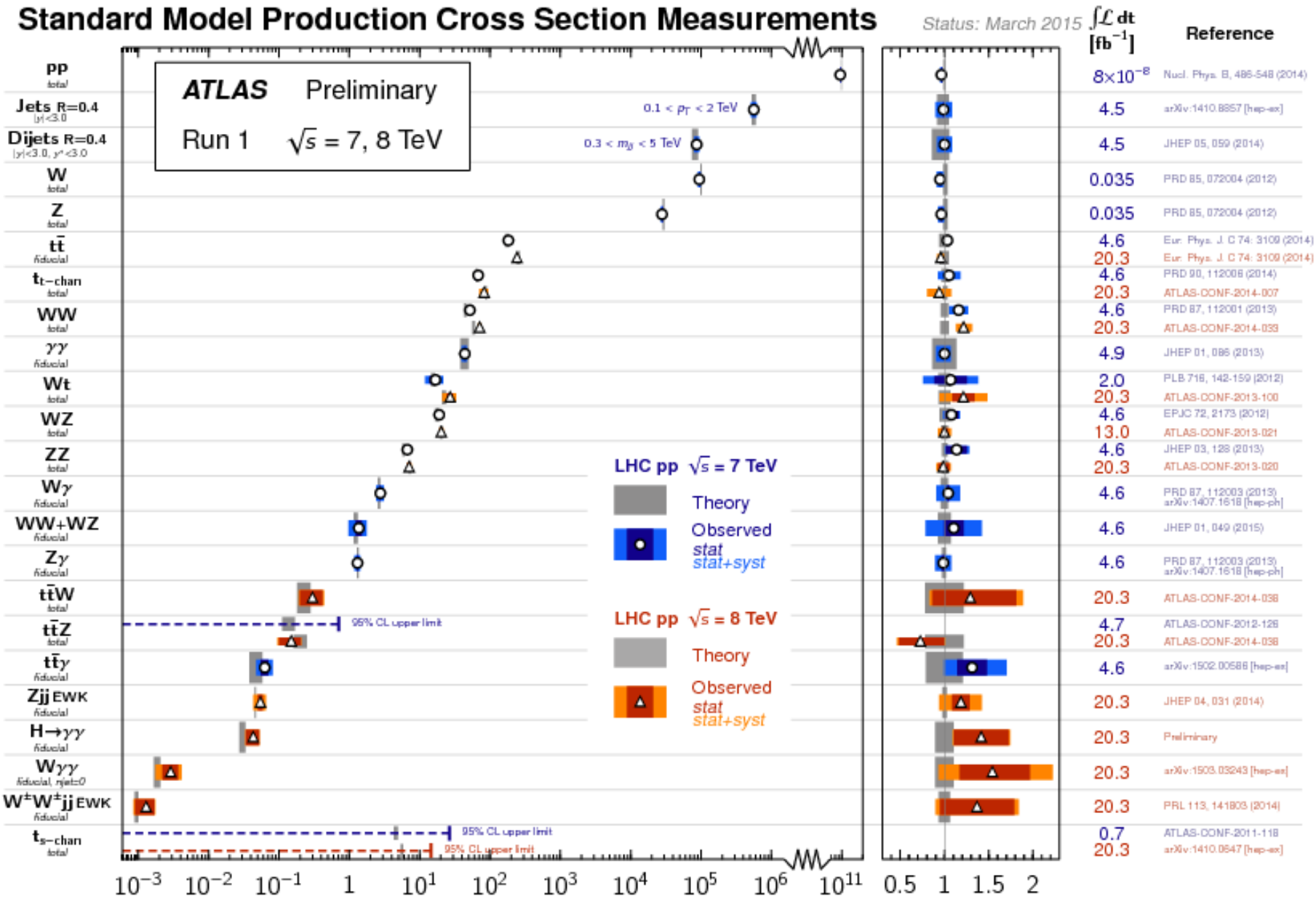
Higgs Centre for Theoretical Physics
University of Edinburgh

work in collaboration with A Chiefa, R Kenway, T Giani, A Candido, G Petrillo, M Wilson

the LHC & the status of the SM

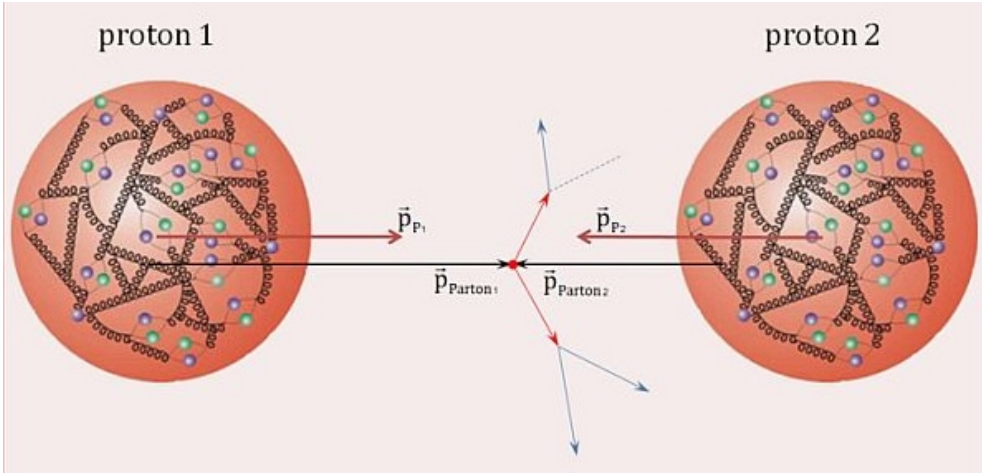


the LHC & the status of the SM



precision physics at the LHC

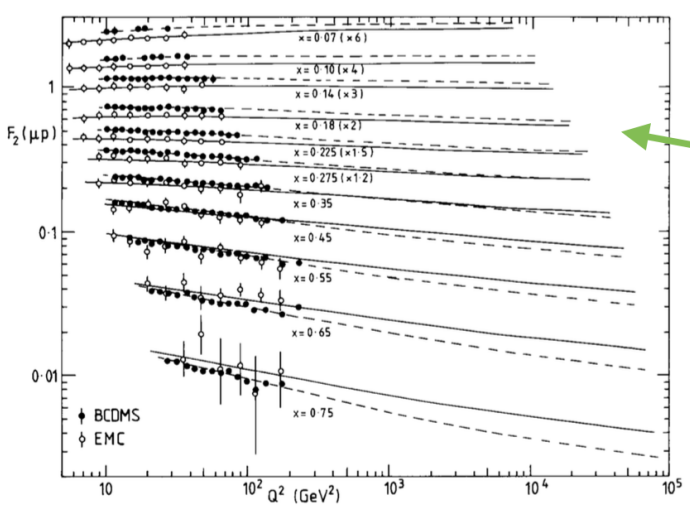
protons are not elementary particles



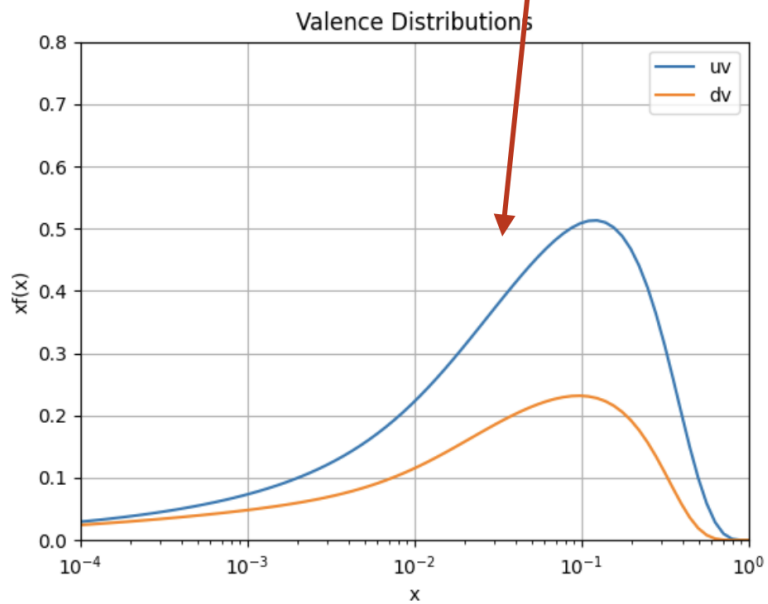
theoretical description in terms of parton distribution functions (PDFs)

$$T_I = \int dx\, C_I(x) f(x)$$

inverse problem



$$y_I = \int dx\, C_I(x) f(x)$$

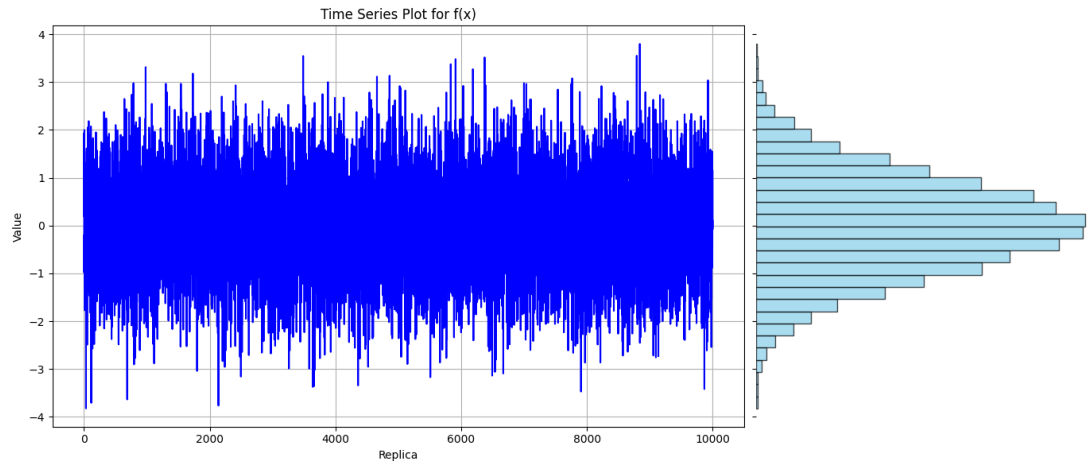


[NNPDF4.0]

parametrization of $f_i(x)$: bias vs variance of the results

Bayesian framework

- promote f to a stochastic process, $f(x)$ are stochastic variables
- consider a grid of points $f_\alpha = f(x_\alpha)$, $\alpha = 1, \ldots, N_{\text{grid}}$
- **choose** a prior distribution $p(f)$ - prior knowledge about f
- probability distributions are represented by ensembles of *replicas*

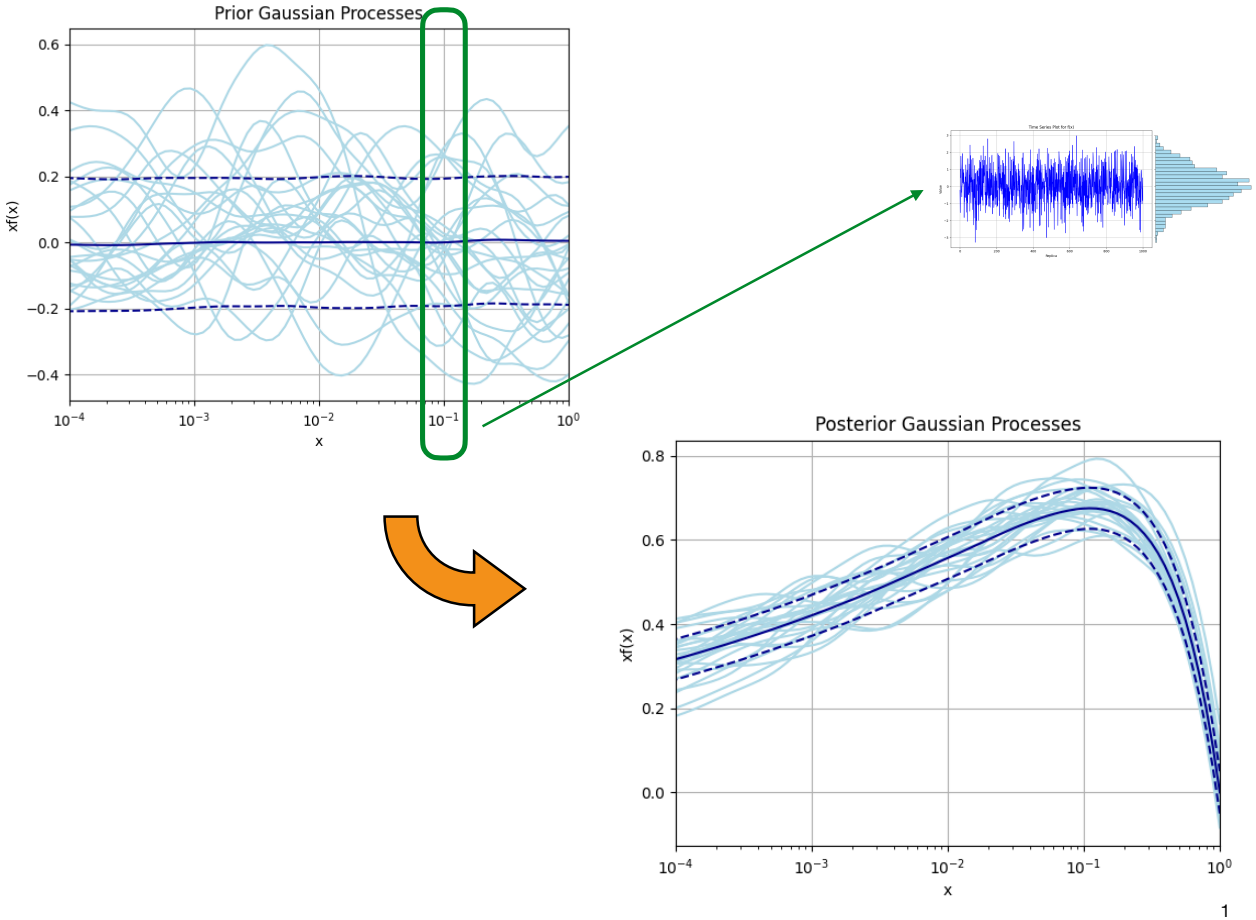


- observables are computed from averages over replicas

$$\bar{O} = E_p\left[O(f)\right] = \int df\, p(f)\, O(f) = \frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} O\left(f^{(k)}\right)$$

posterior distribution

- use the data to infer a posterior probability $\tilde{p}(f)$

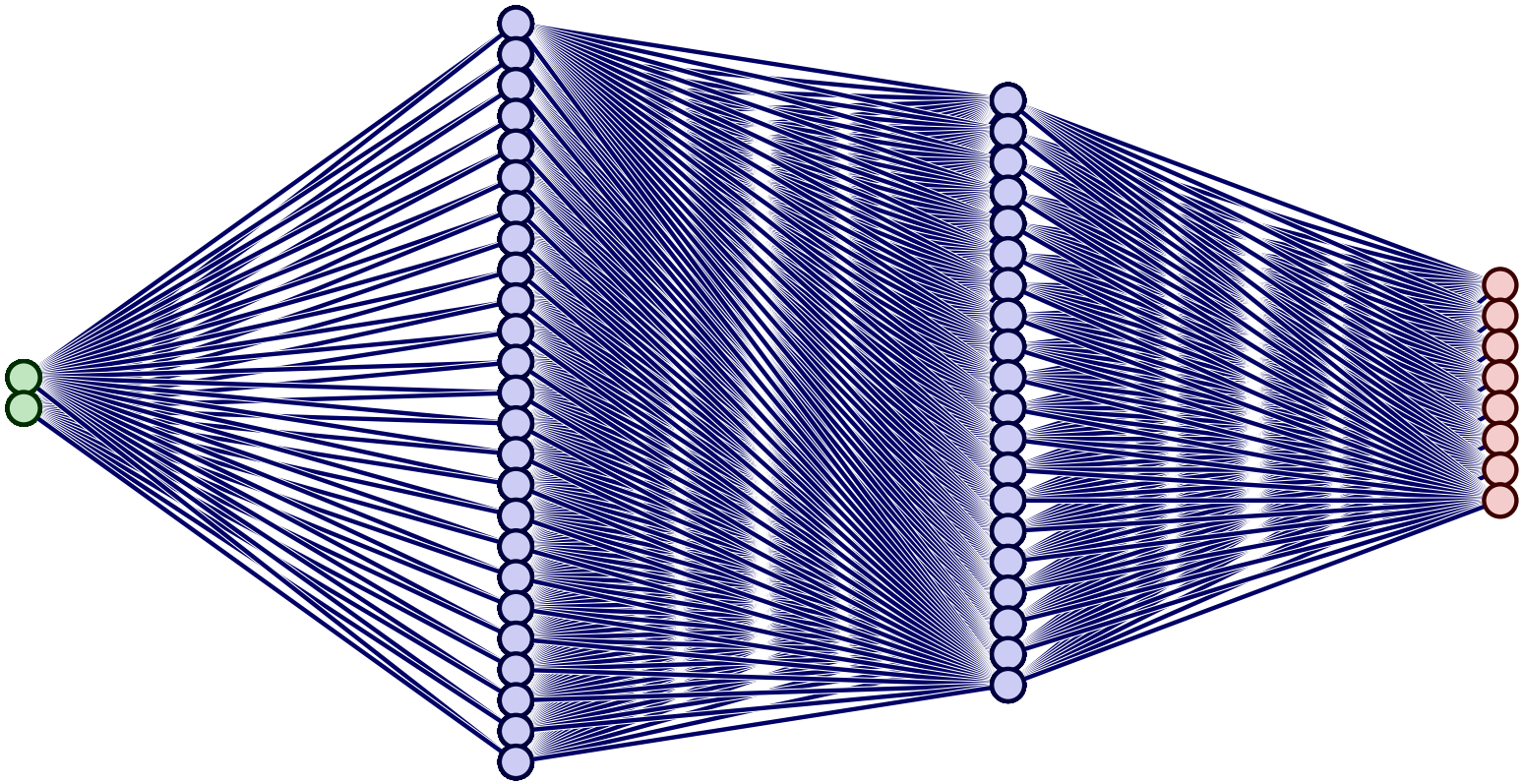


$$\bar{f}(x) = E_{\tilde{p}}\left[f(x)\right] , \quad \delta f(x) = E_{\tilde{p}}\left[\left(f(x) - \bar{f}(x)\right)^2\right]$$

NNPDF paradigm

- parametrize the unknown functions f using neural networks (MPL)
- initialize $N_{\rm rep}$ NN replicas
- train each NN on a replica of the dataset
- the ensemble of trained NNs provides the posterior distribution

NNPDF parametrization - conventions



$$\rho_i^{(\ell)} = \rho(\phi_i^{(\ell)}), \quad \phi_i^{(\ell)} = \sum_{j=1}^{n_{\ell-1}} w_{ij}^{(\ell)} \rho_j^{(\ell-1)} + b_i^{(\ell)}$$

$$\boxed{f(x) = \phi^{(L)}(x;\theta)}$$

finite-dimensional, yet flexible

NN prior distribution

parameters θ are initialized using a Glorot-Normal distribution

initialize weights and biases using Gaussians

$$\langle b_i^{(\ell)} \rangle = 0\,, \quad \langle b_{i_1}^{(\ell)} b_{i_2}^{(\ell)} \rangle = \delta_{i_1 i_2} C_b^{(\ell)}$$

$$\langle w_{ij}^{(\ell)} \rangle = 0\,, \quad \langle w_{i_1 j_1}^{(\ell)} w_{i_2 j_2}^{(\ell)} \rangle = \delta_{i_1 i_2} \delta_{j_1 j_2} \frac{C_w^{(\ell)}}{n_{\ell-1}}$$

parameters/functions duality

$$p(\phi^{(\ell)}) = \int [dw\, p(w)]\, [db\, p(b)] \prod_{i,\alpha} \delta\left(\phi_{i\alpha}^{(\ell)} - \sum_j w_{ij}^{(\ell)} \rho\left(\phi_{j\alpha}^{(\ell-1)} \right) - b_i^{(\ell)} \right)$$

EFT approach

symmetry and n counting yield the probability distribution for ϕ

$$
\begin{aligned}
p(\phi^{(\ell)}) &= \frac{1}{Z}\,\exp\left[-S(\phi^{(\ell)})\right] \\
&= \frac{1}{Z}\,\exp\left[-\frac{1}{2}\gamma^{(\ell)}_{\alpha_1\alpha_2}\phi^{(\ell)}_{\alpha_1}\cdot\phi^{(\ell)}_{\alpha_2}\right. \\
&\qquad\left.-\frac{1}{8n_{\ell-1}}\gamma^{(\ell)}_{\alpha_1\alpha_2,\alpha_3\alpha_4}\phi^{(\ell)}_{\alpha_1}\cdot\phi^{(\ell)}_{\alpha_2}\,\phi^{(\ell)}_{\alpha_3}\cdot\phi^{(\ell)}_{\alpha_4}+O(1/n_{\ell-1}^2)\right]
\end{aligned}
$$

correlators can be computed using Feynman diagrams

$$
\begin{aligned}
\langle\phi^{(\ell)}_{i_1,\alpha_1}\phi^{(\ell)}_{i_2,\alpha_2}\rangle &= \delta_{i_1 i_2}K^{(\ell)}_{\alpha_1\alpha_2}+O(1/n_{\ell-1}) \\
\langle\phi^{(\ell)}_{i_1,\alpha_1}\phi^{(\ell)}_{i_2,\alpha_2}\phi^{(\ell)}_{i_3,\alpha_3}\phi^{(\ell)}_{i_4,\alpha_4}\rangle_c &= O(1/n_{\ell-1})
\end{aligned}
$$

ϕ are approximately Gaussian processes, corrections are $O(1/n)$

going deep – recursion relations

two-pt function at leading order

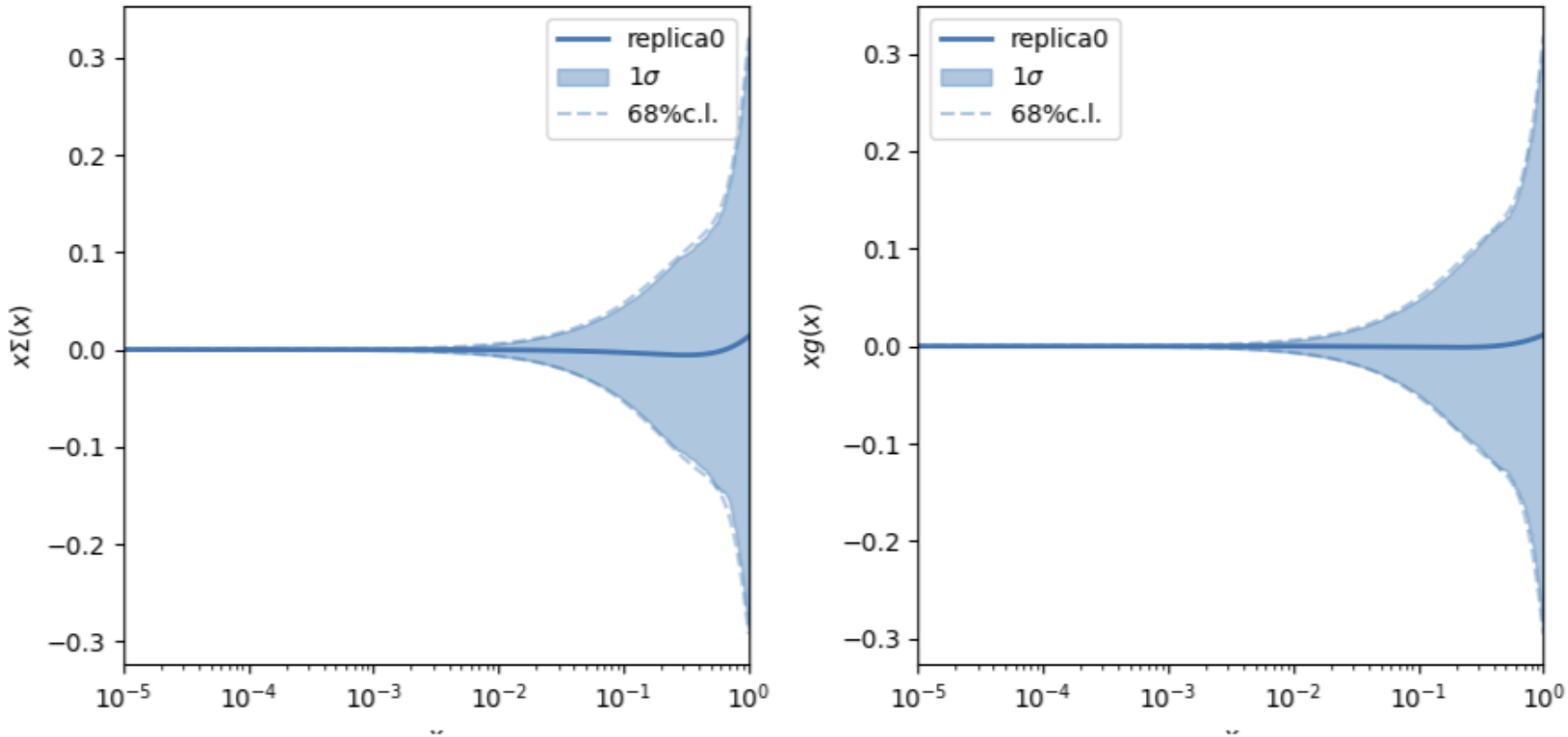
$$K^{(\ell+1)}_{\alpha_1\alpha_2} = C_b^{(\ell+1)} + C_w^{(\ell+1)} \frac{1}{n_\ell} \langle \vec{\rho}^{(\ell)}_{\alpha_1} \cdot \vec{\rho}^{(\ell)}_{\alpha_2} \rangle \Big|_{O(1)}$$
$$= C_b^{(\ell+1)} + C_w^{(\ell+1)} \frac{1}{n_\ell} \langle \vec{\rho}^{(\ell)}_{\alpha_1} \cdot \vec{\rho}^{(\ell)}_{\alpha_2} \rangle_{K^{(\ell)}}$$

where

$$\frac{1}{n_\ell} \langle \vec{\rho}^{(\ell)}_{\alpha_1} \cdot \vec{\rho}^{(\ell)}_{\alpha_2} \rangle_{K^{(\ell)}} = \int \prod_\alpha d\phi_\alpha \, \frac{e^{-\frac{1}{2}\left(K^{(\ell)}\right)^{-1}_{\beta_1\beta_2}\phi_{\beta_1}\phi_{\beta_2}}}{\left|2\pi K^{(\ell)}\right|^{1/2}} \rho(\phi_{\alpha_1})\rho(\phi_{\alpha_2})$$

increasing ℓ, the couplings *evolve*, exactly RG evolution

NNPDF initialization



very good agreement with the analytical predictions from EFT

training

gradient descent – for all parametrizations

$$\frac{d}{dt}\theta_\mu = -\nabla_\mu \mathcal{L}\,, \quad \mathcal{L} = \frac{1}{2}\left(y - T[f_t]\right)^T C_Y^{-1}\left(y - T[f_t]\right)$$

$$\nabla_\mu \mathcal{L} = -\left(\nabla_\mu f_t\right)^T \left(\frac{\partial T}{\partial f}\right)_t C_Y^{-1}\epsilon_t\,, \quad \epsilon_t = \left(y - T[f_t]\right)$$

$$\frac{d}{dt} f_t = \left(\nabla_\mu f_t\right)\frac{d}{dt}\theta_\mu = \Theta_t \left(\frac{\partial T}{\partial f}\right)_t C_Y^{-1}\epsilon_t$$

where

$$\Theta_t = \left(\nabla_\mu f_t\right)\left(\nabla_\mu f_t\right)^T$$

is the Neural Tangent Kernel (NTK)

linear data & NN parametrization

for linear data

$$T = (\mathrm{FK})f \quad \Longrightarrow \quad \left(\frac{\partial T}{\partial f}\right) = (\mathrm{FK})$$

for wide neural networks

$$\Theta_t = \Theta + O(1/n) \quad \text{(lazy training)}$$

hence we obtain a linear equation for f_t

$$\begin{aligned}\frac{d}{dt} f_t &= \Theta(\mathrm{FK})^T C_Y^{-1} \left(y - (\mathrm{FK}) f_t\right) \\ &= -\Theta M f_t + b\end{aligned}$$

flat directions & NTK spectrum

eigenvalues of the NTK

$$\Theta z^{(k)} = \lambda^{(k)} z^{(k)}\,, \quad f_{t,k} = \left(z^{(k)}, f_t\right)$$

evolution equations in the NTK eigenbasis

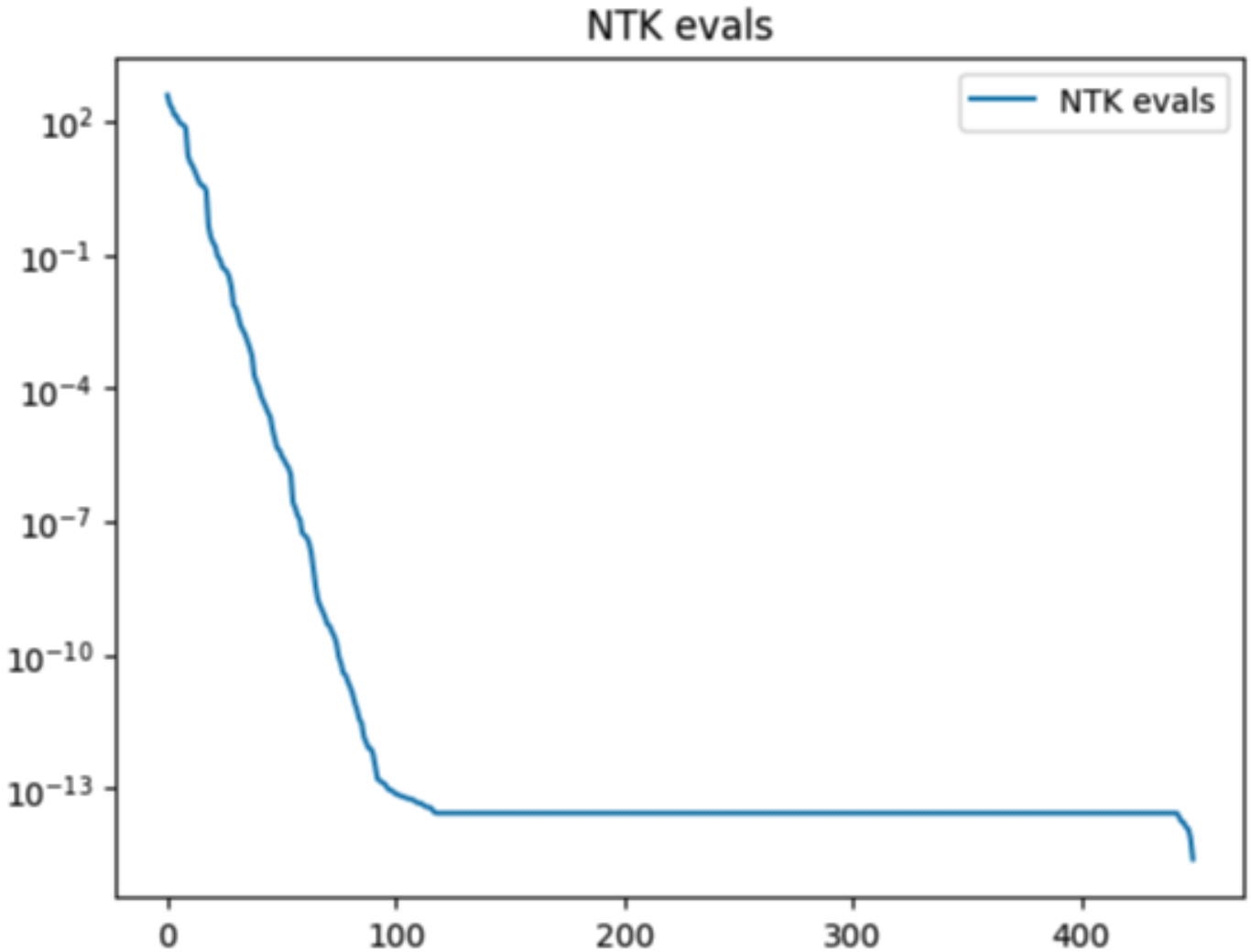
$$\frac{d}{dt} f_{t,k} = \lambda^{(k)} \left(z^{(k)}, (\mathrm{FK})^T C_Y^{-1} \left(y - (\mathrm{FK}) f_t\right)\right)$$

NTK kernel: directions that do NOT evolve during training

$$f_{t,\|} = f_{0,\|}$$

no bias, but irreducible noise dictated by the prior

NTK spectrum



integrating the flow equation

$$\frac{d}{dt} f_t = -\Theta M f_t + b\,, \quad M = (\mathrm{FK})^T C_Y^{-1} (\mathrm{FK}) = R D R^T$$

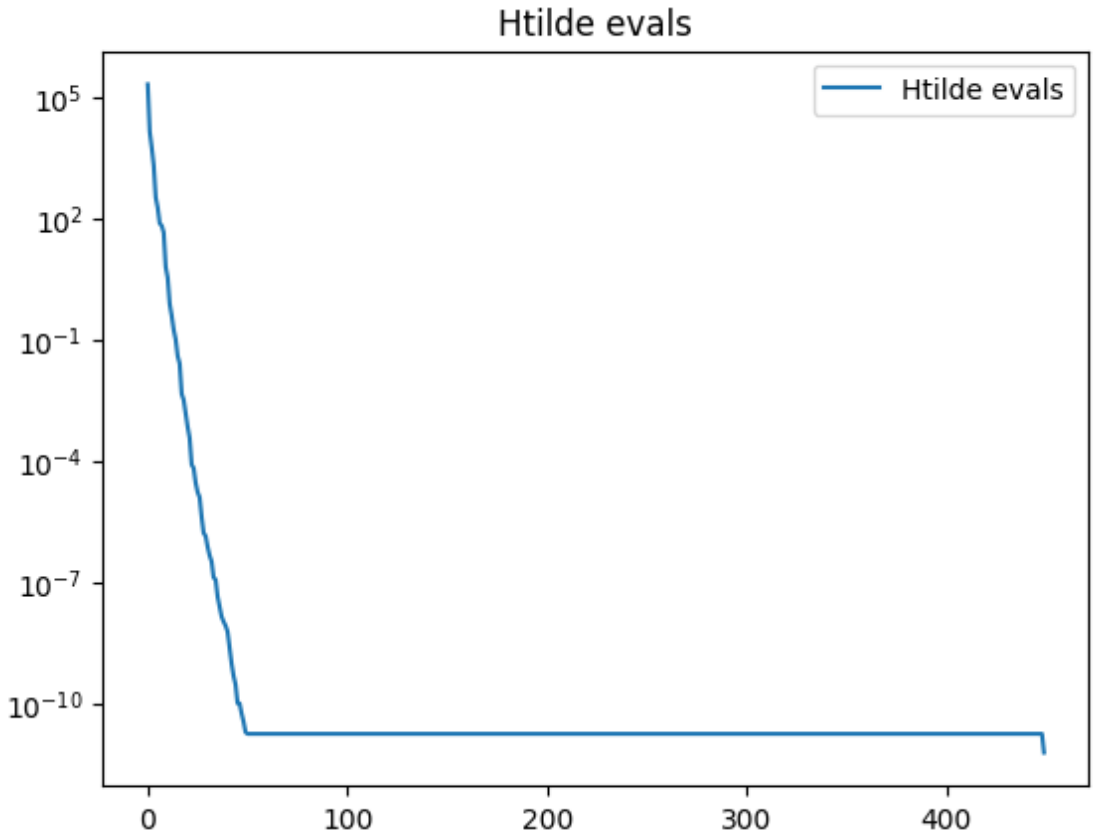
introducing new variables

$$\tilde{f}_t = D^{1/2} R^T f_t\,, \quad \tilde{b} = D^{1/2} R^T b$$

$$\frac{d}{dt} \tilde{f}_t = -\tilde{H} \tilde{f}_t + \tilde{b}\,, \quad \tilde{H} = D^{1/2} R^T \Theta R D^{1/2}$$

interplay between data and architecture

spectrum of the flow hamiltonian



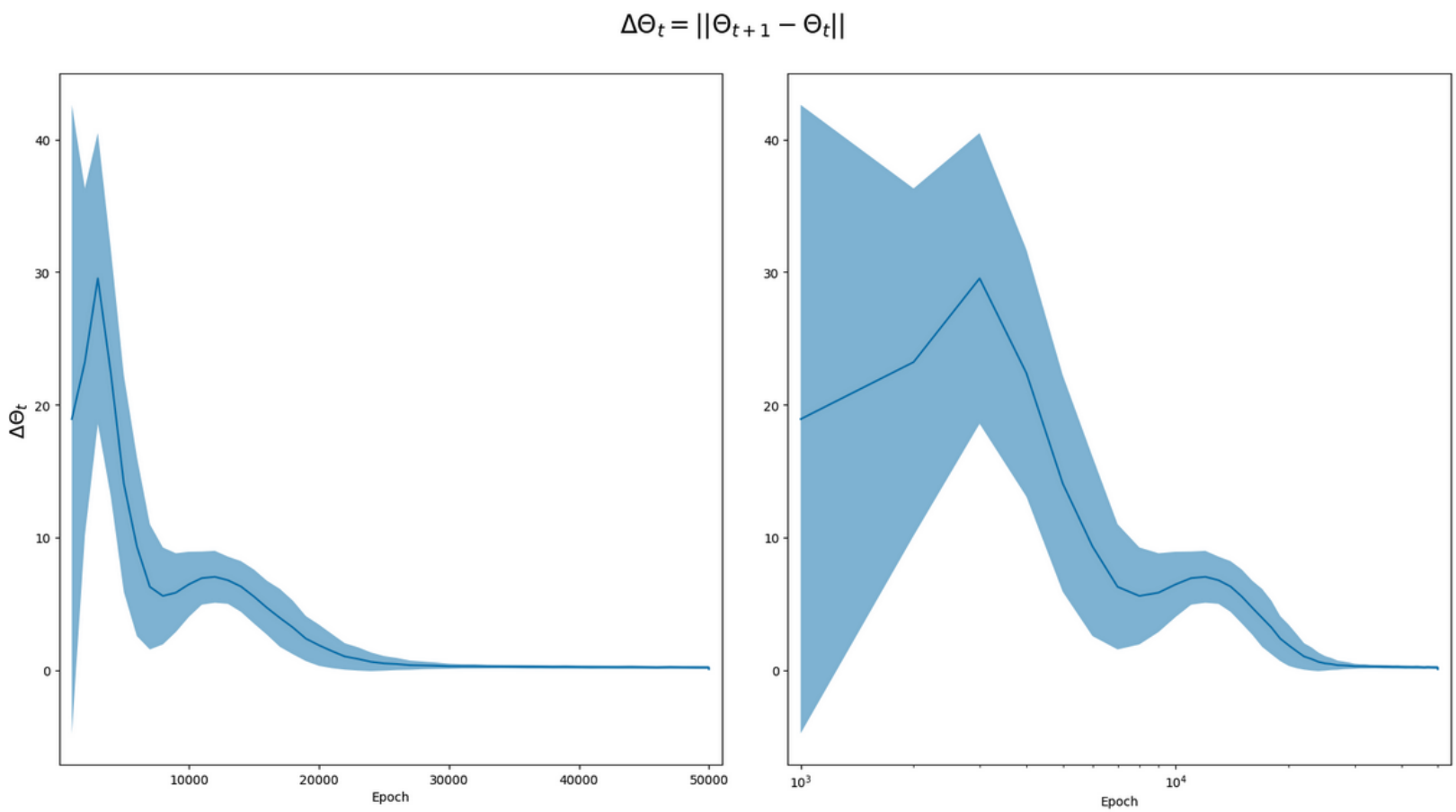
many flat directions!

- the rate at which features are learned is dictated by the evals of $\tilde{H}$
- there is a strong hierachy in the evals (spectral bias)
- solution of the flow equation

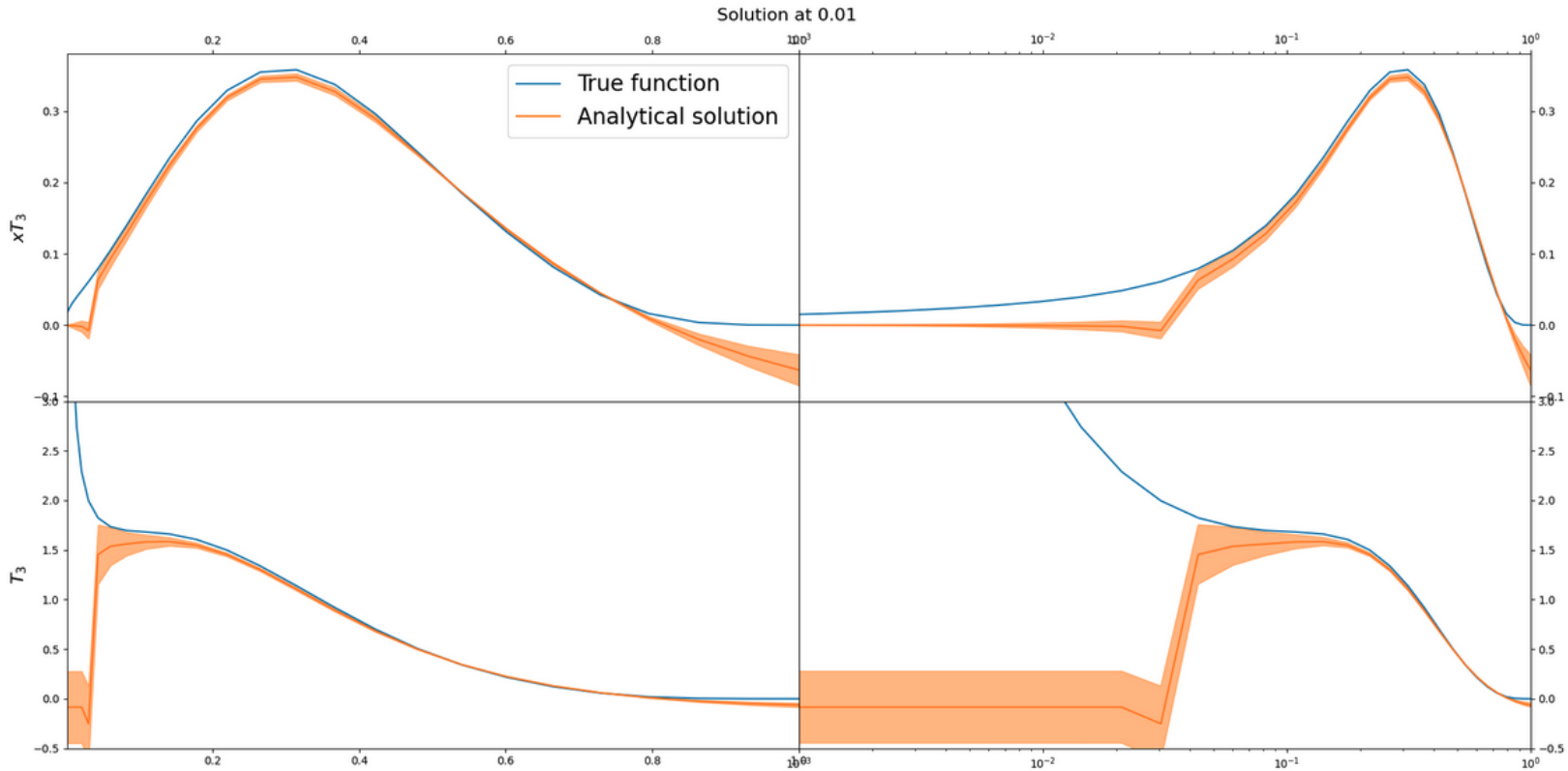
$$f_t = \mathcal{A} e^{-\tilde{H}t} f_0 + \mathcal{A} \left(1 - e^{-\tilde{H}t}\right) \mathcal{A}^T (\mathrm{FK})^T C_Y^{-1} y$$

- analytical framework to understand stopping criteria $\longrightarrow$ bias and variance

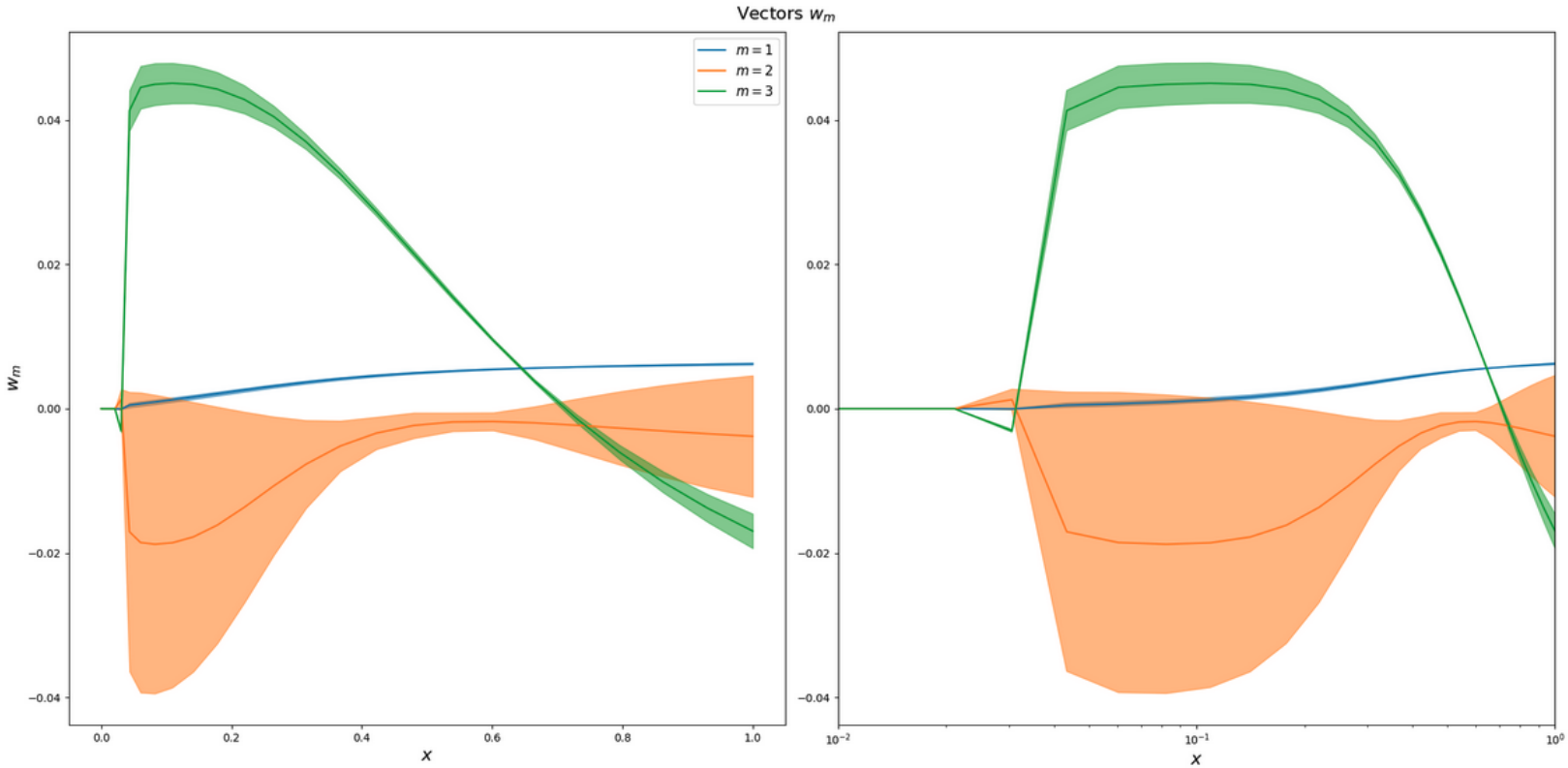
approaching the lazy training regime



integrated analytical solution



high modes of the flow Hamiltonian



outlook

- PDFs are a central ingredient for LHC analyses
- Bayesian approach is convenient to solve these problems
- all hypotheses are clearly spelled in the prior
- analytical control of the training process is key to quantify the bias/variance budget