ML/AI activities at Nikhef Theory

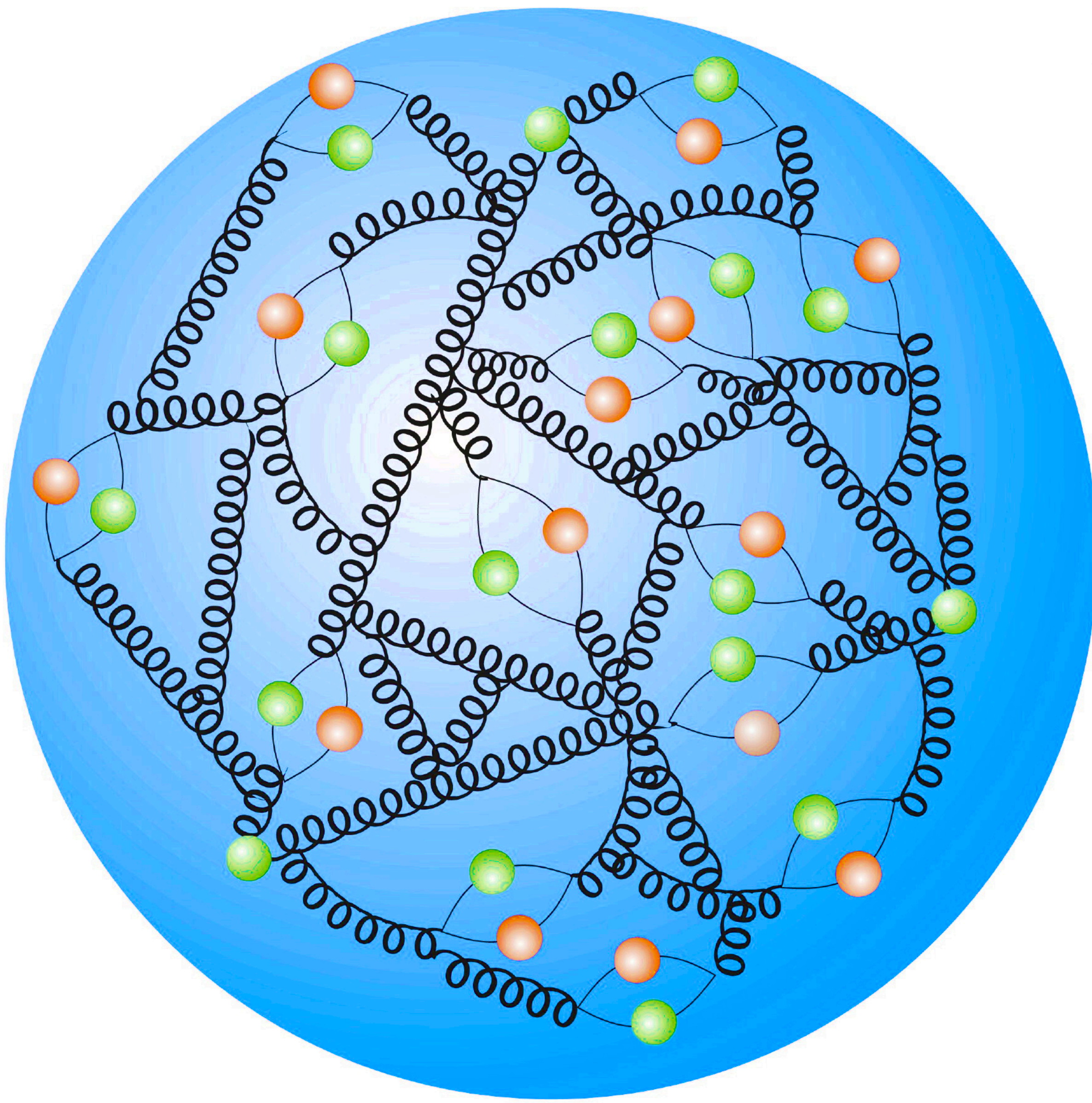
Juan Rojo, VU Amsterdam & Nikhef

Nikhef ML/AI Working Group - Kick-off Meeting

Nikhef, 6th September 2024

The many faces of the proton

QCD bound state of **quarks** and **gluons**



- **Valence quarks (up and down)** give the proton its quantum numbers (e.g. electric charge)

$$|\Psi\rangle \approx |uud\rangle$$

$$Q_p = +1 \qquad \begin{array}{l} Q_u = +2/3 \\ Q_d = -1/3 \end{array}$$

- **Sea quarks (antiup, antidown, strange, …)** arise from quantum fluctuations

- Tightly held together by **gluons,** can only be broken in extremely energetic collisions

Parton Distributions

$$g(x, Q)$$

Probability of **finding a gluon inside a proton**, carrying a fraction ***x*** of the proton momentum, when probed with energy ***Q***

x: fraction of proton momentum carried by gluon

Energy of hard-scattering reaction**:** inverse of resolution length

Dependence on *x* fixed by **non-perturbative QCD dynamics**: extract from experimental data

$$g(x, Q_0, \{a_g\}) = f_g(x, a_g^{(1)}, a_g^{(2)}, \ldots)$$

constrain from data

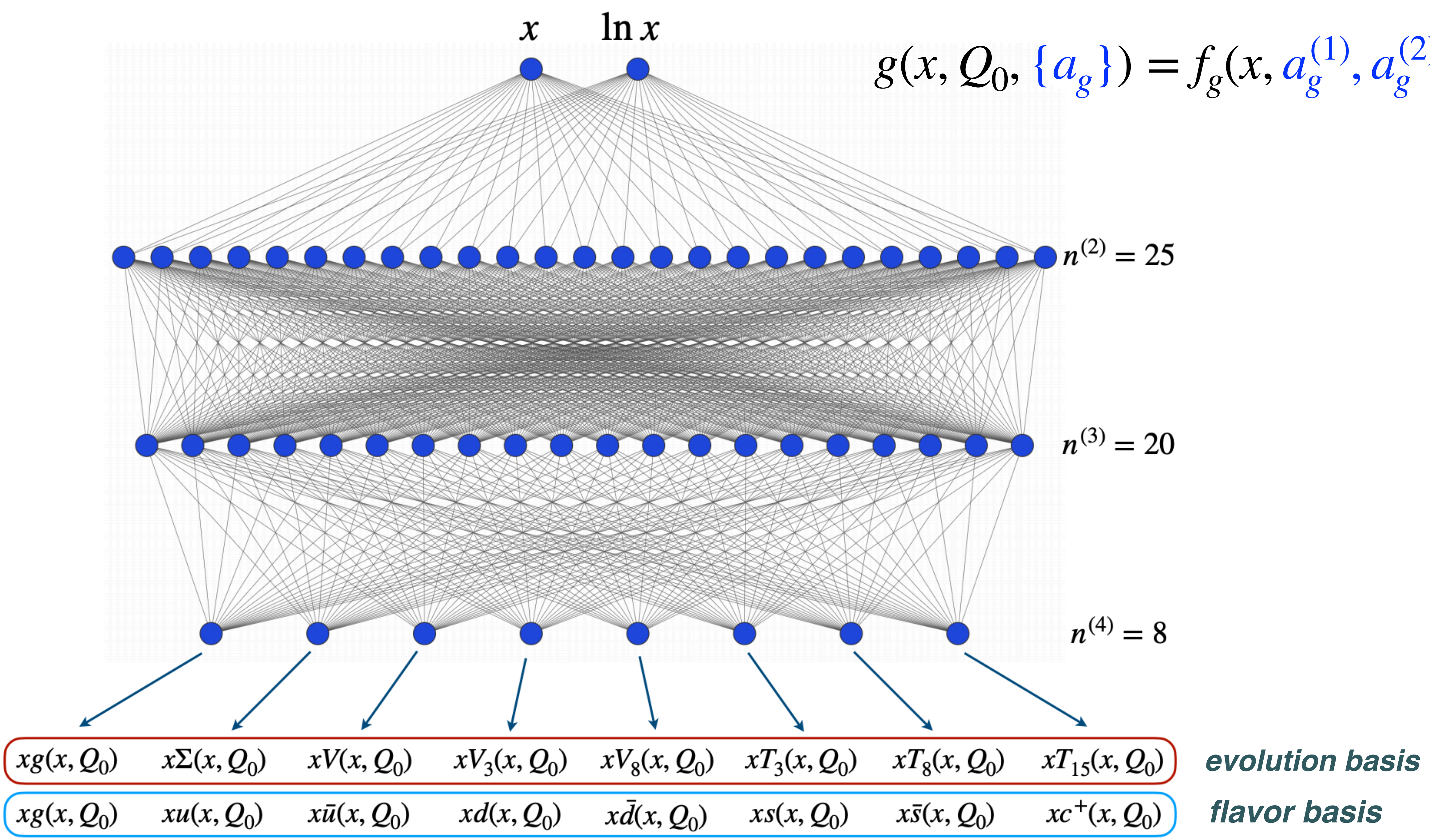
- Dependence with **resolution scale *Q***: DGLAP evolution, computable from first principles

- **Energy conservation** and **quark number conservation** are fixed boundary conditions

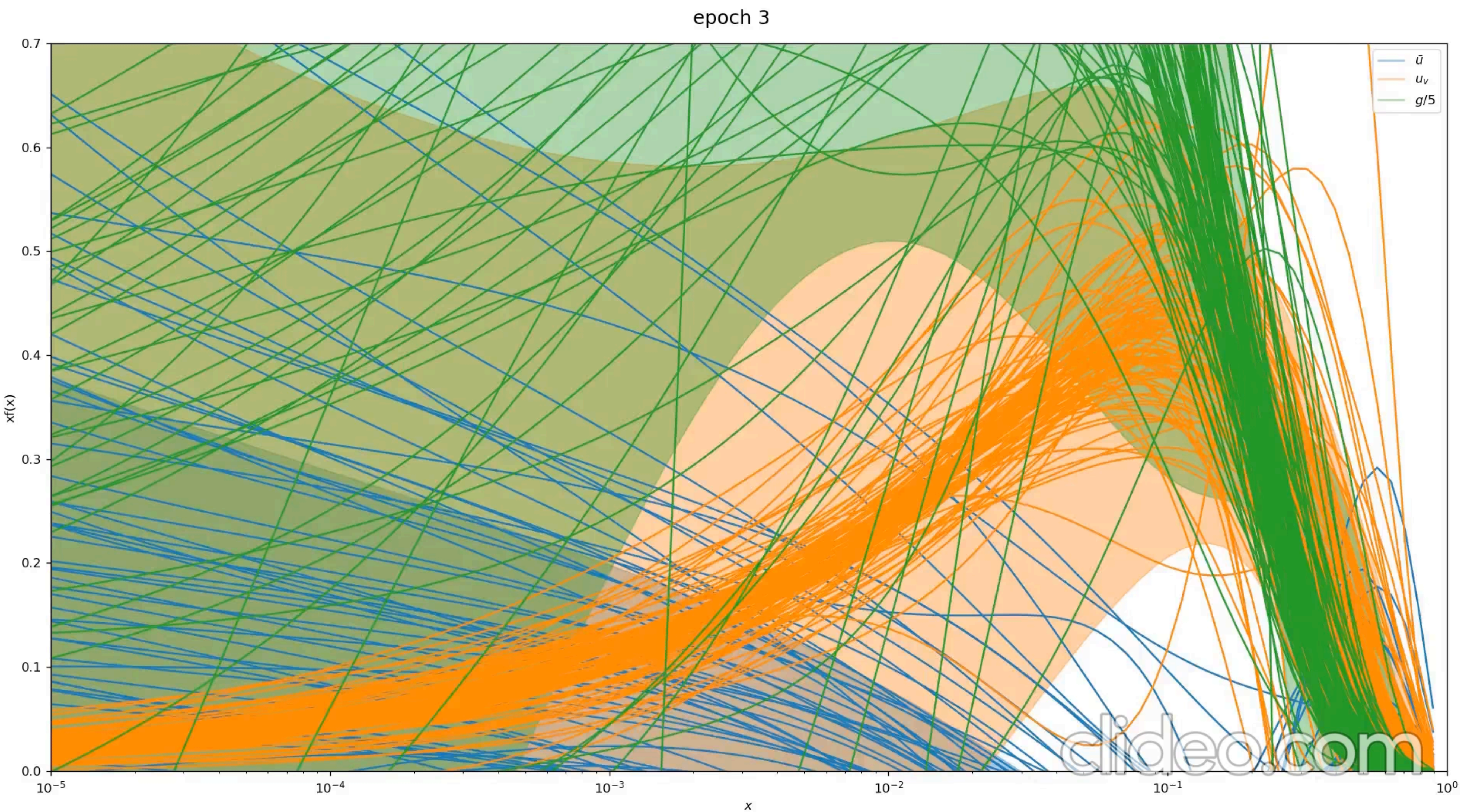
ML Proton Structure

- Model-independent PDF parametrisation with neural networks as **universal unbiased interpolants**
- **Stochastic Gradient Descent** via TensorFlow for neural network training
- Automated model **hyperparameter optimisation:** NN architecture, minimiser, learning rates …

$$g(x, Q_0, \{a_g\}) = f_g(x, a_g^{(1)}, a_g^{(2)}, \ldots)$$



Machine Learning PDFs



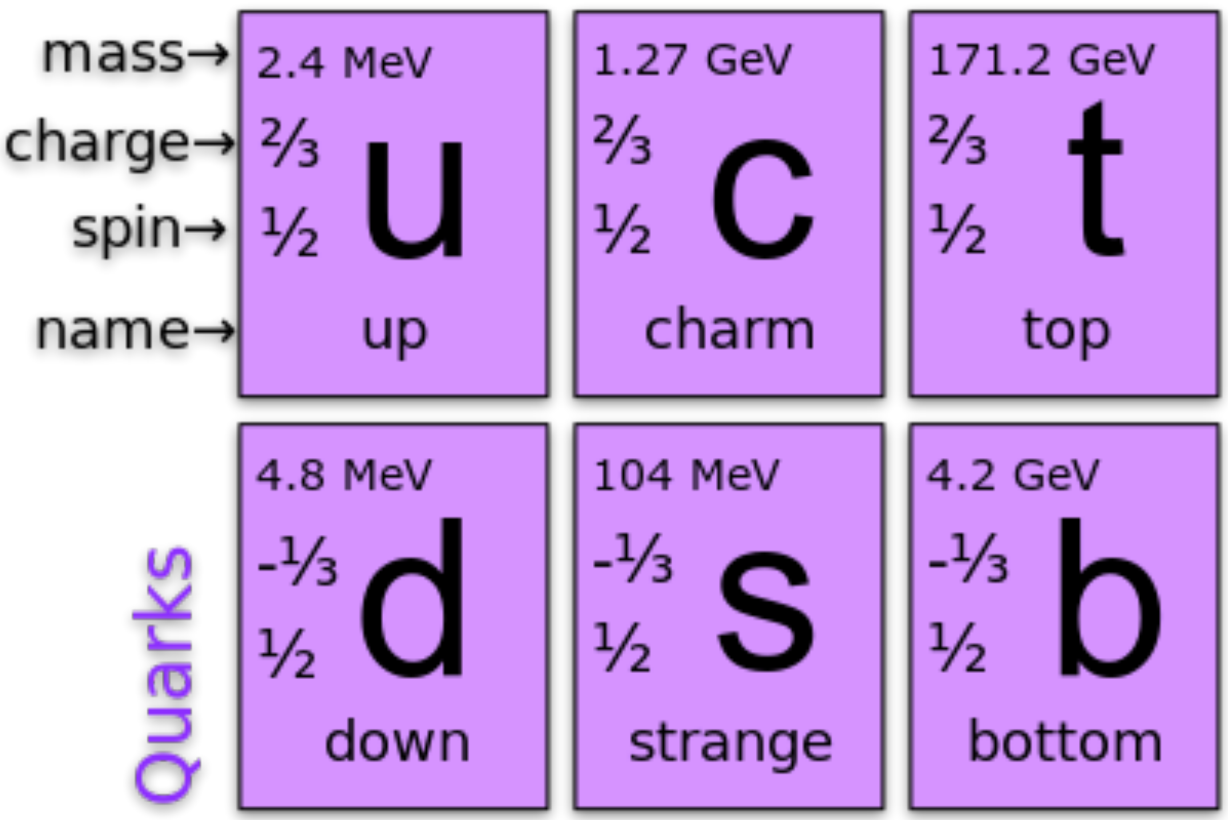
Error estimate based on **Monte Carlo replica method** (band: standard deviation over the MC replicas)

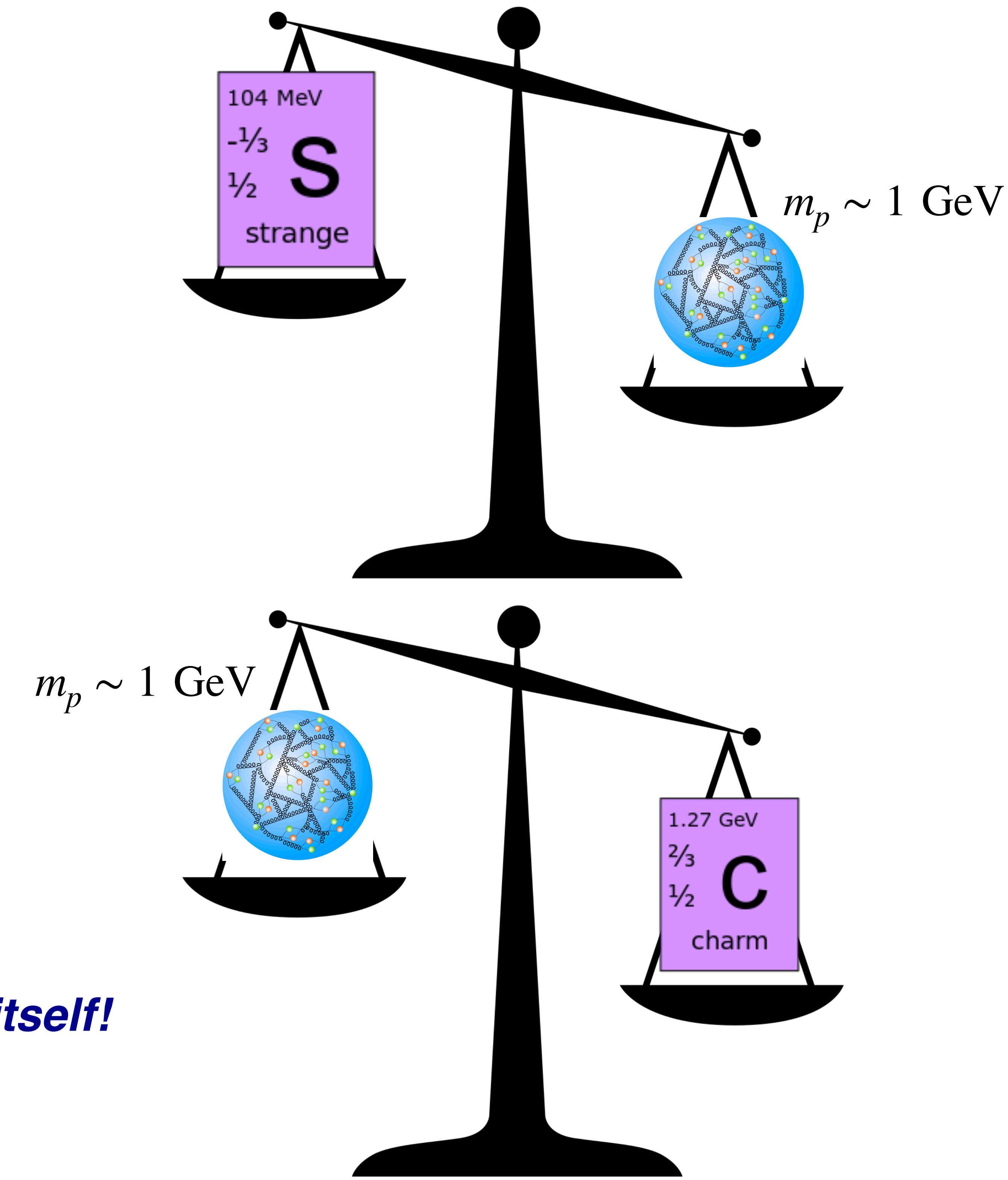
each curve is a separately trained neural network

The charm content of the proton

common assumption: the proton wave function does not contain charm quarks

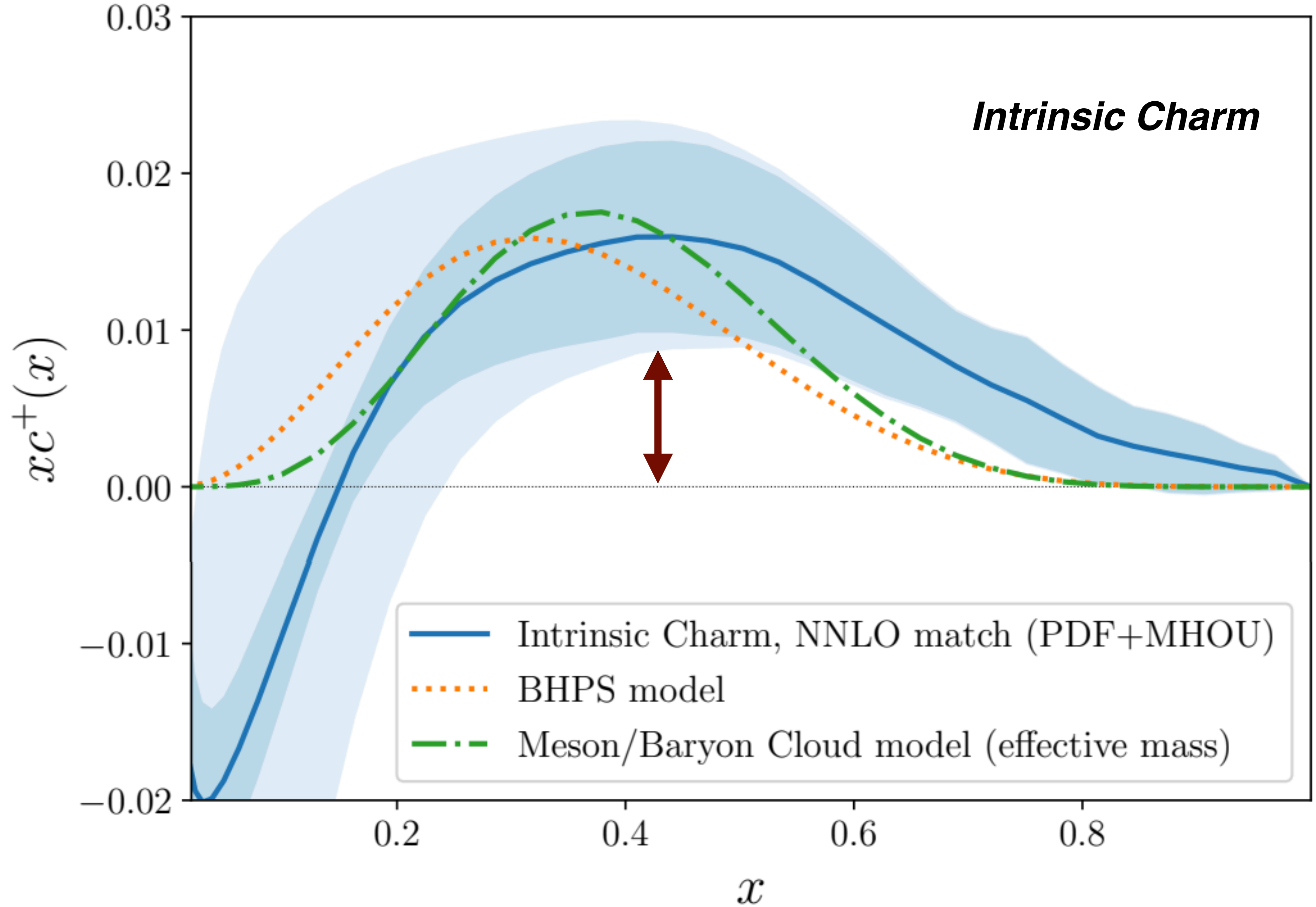
the proton contains **intrinsic up, down, strange (anti-)quarks** but **no intrinsic charm quarks**





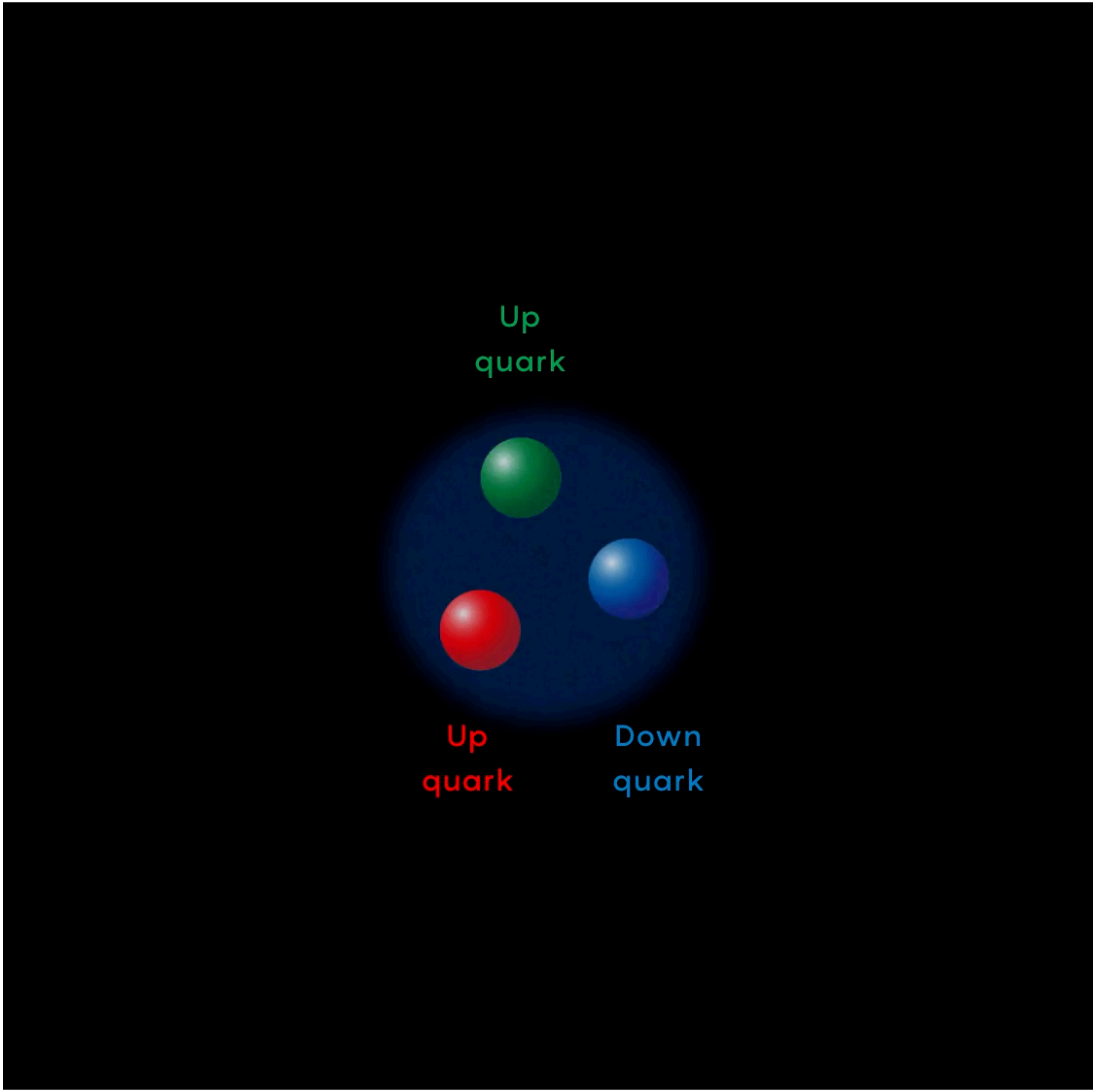
charm quarks heavier than the proton itself!

Intrinsic Charm in the Proton



The 3FNS charm PDF displays **non-zero component** peaked at large-*x* which can be identified with **intrinsic charm**

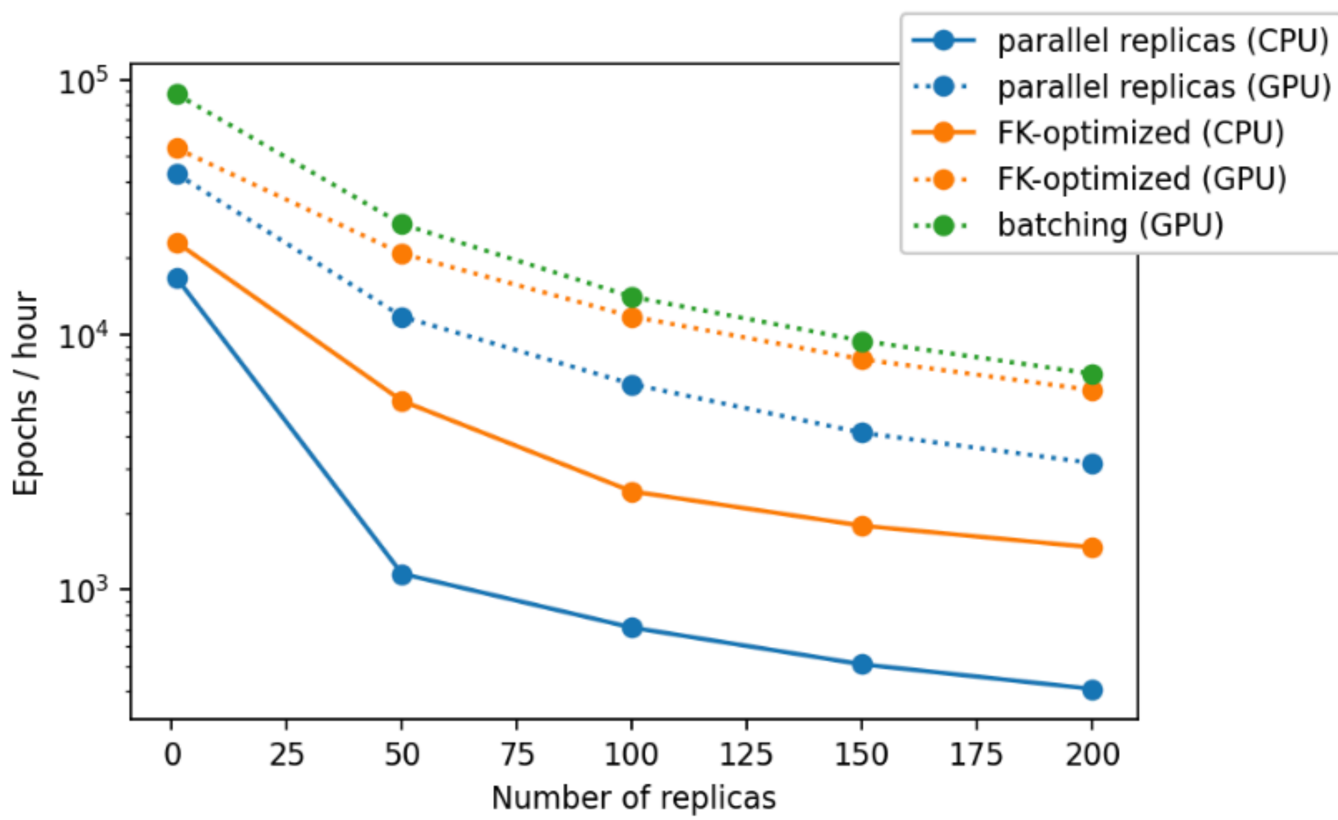
JR et al, Evidence for intrinsic charm quarks in the proton, ***Nature* (2022)**

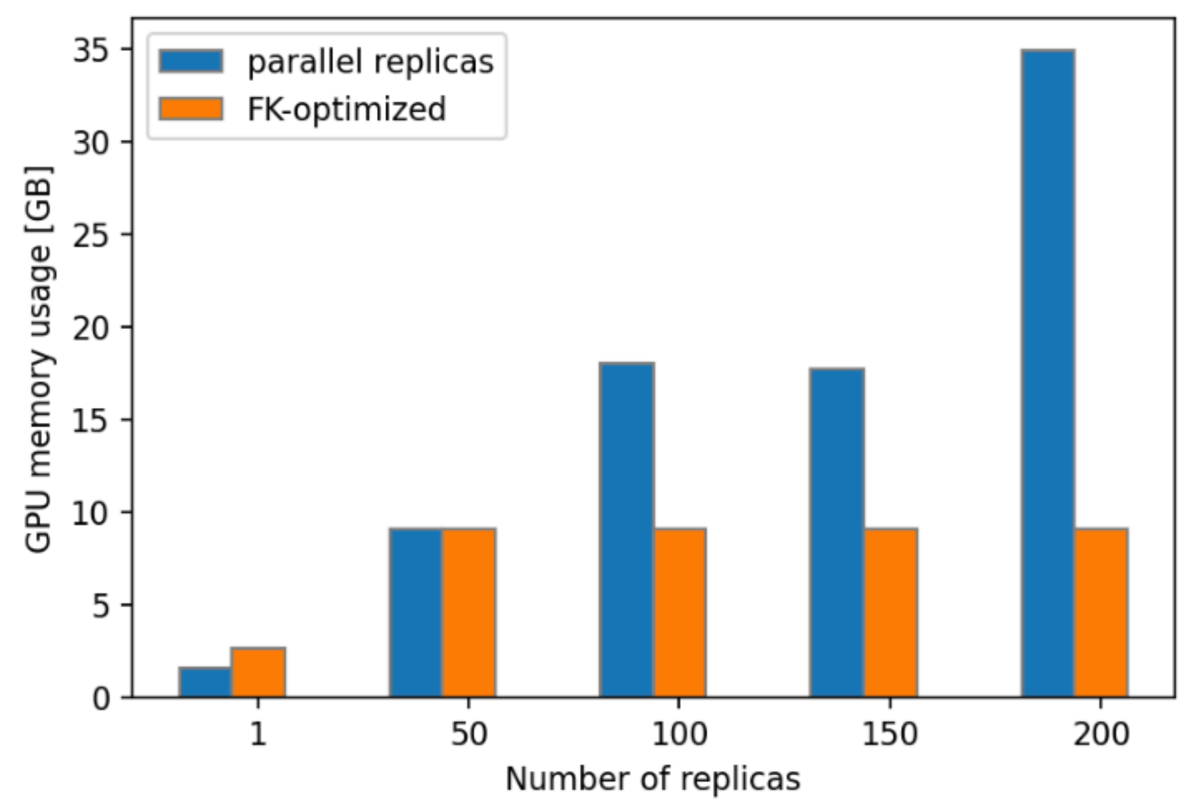


credits: https://www.quantamagazine.org/inside-the-proton-the-most-complicated-thing-imaginable-20221019/

GPU & Hyperparameter Optimisation

- Deploy NNPDF machinery on GPUs & optimise performance: **factor speed 200 improvement!**
- Also ensure CPU memory consumption keep reasonable
- Develop new strategies for **hyperparameter optimisation** based on the full posterior probability distribution, not only on first moment as most approaches



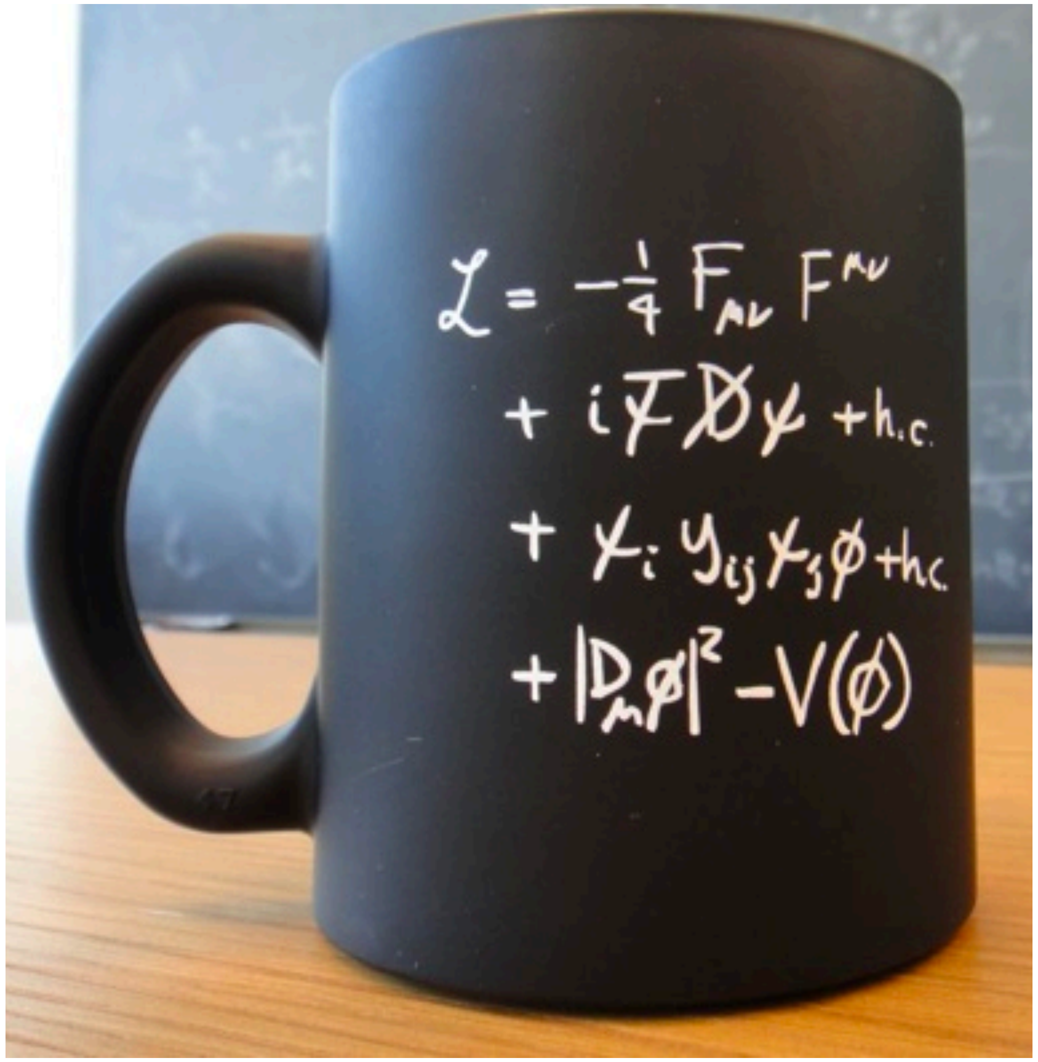


netherlands eScience center

The Standard Model as an Effective Theory

The Standard Model EFT is defined by:

- **Particle (matter) content**: quarks and leptons
- **Gauge** (local) symmetries and their eventual breaking mechanisms
- **Lorentz** invariance and other global symmetries
- Linearly realised $SU(2)_L$ EW symmetry breaking
- **Validity only up to certain energy scale *Λ***



$$\mathcal{L}_{\text{SMEFT}}(\{c_i\}, \Lambda) = \mathcal{L}_{\text{SM}} + \sum_{d=5}^{\infty} \sum_{i=1}^{N_d} c_i^{(d)} \frac{\mathcal{O}_i^{(d)}}{\Lambda^{d-4}}$$

EFT coupling constants, to be determined from **data**

All possible operators of **mass-dimension *d*** consistent with above requirements

Statistically optimal observables from ML

Optimal observables depend on **all kinematic variables** and **all EFT coefficients**

$$r_\sigma(\boldsymbol{x}, \boldsymbol{c}) \equiv \frac{f_\sigma(\boldsymbol{x}, \boldsymbol{c})}{f_\sigma(\boldsymbol{x}, \boldsymbol{0})} = 1 + \sum_{j=1}^{n_{\rm eft}} r_\sigma^{(j)}(\boldsymbol{x}) c_j + \sum_{j=1}^{n_{\rm eft}} \sum_{k \geq j}^{n_{\rm eft}} r_\sigma^{(j,k)}(\boldsymbol{x}) c_j c_k$$

parametrised with **neural networks** trained to Monte Carlo simulations & benchmarked with exact calculations

$$\hat{r}_\sigma(\boldsymbol{x}, \boldsymbol{c}) = 1 + \sum_{j=1}^{n_{\rm eft}} \mathrm{NN}^{(j)}(\boldsymbol{x}) c_j + \sum_{j=1}^{n_{\rm eft}} \sum_{k \geq j}^{n_{\rm eft}} \mathrm{NN}^{(j,k)}(\boldsymbol{x}) c_j c_k$$

extendable to **arbitrary number** of kinematic variables and EFT coefficients: training can be parallelised

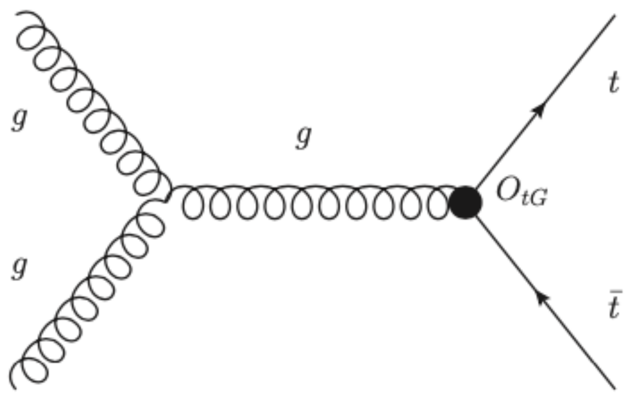
methodological uncertainties (e.g. finite training samples) assess with the **replica method**

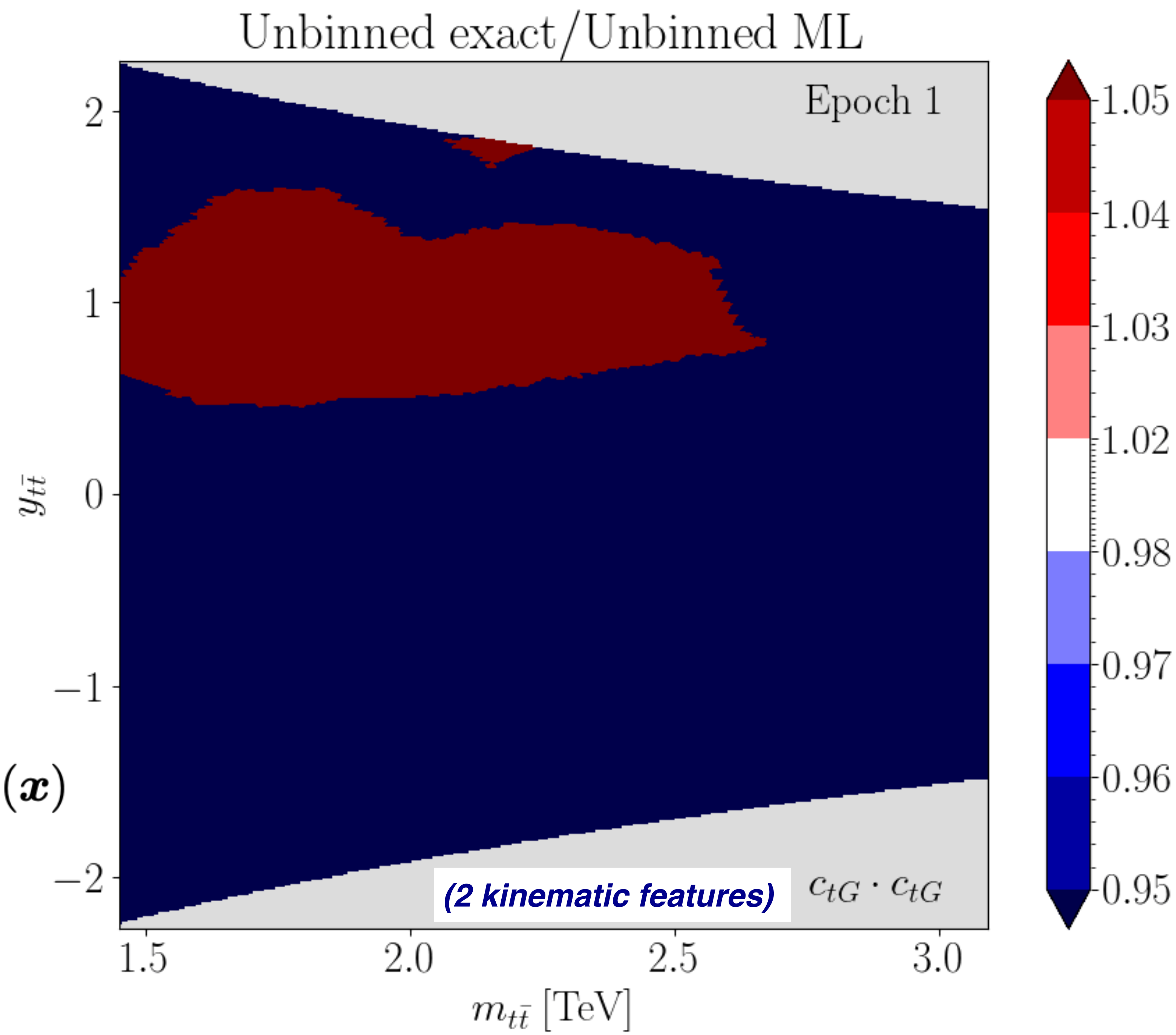
$$\hat{r}_\sigma^{(i)}(\boldsymbol{x}, \boldsymbol{c}) \equiv 1 + \sum_{j=1}^{n_{\rm eft}} \mathrm{NN}_i^{(j)}(\boldsymbol{x}) c_j + \sum_{j=1}^{n_{\rm eft}} \sum_{k \geq j}^{n_{\rm eft}} \mathrm{NN}_i^{(j,k)}(\boldsymbol{x}) c_j c_k \,, \qquad i = 1, \ldots, N_{\rm rep}$$

each replica trained to an independent set of MC events

representation of the probability distribution in the space of ML models

Neural network training





$$r_\sigma(\boldsymbol{x}, c_j^{(\mathrm{tr})}, c_k^{(\mathrm{tr})}) = 1 + c_j^{(\mathrm{tr})} c_k^{(\mathrm{tr})} \mathrm{NN}^{(j,k)}(\boldsymbol{x})$$

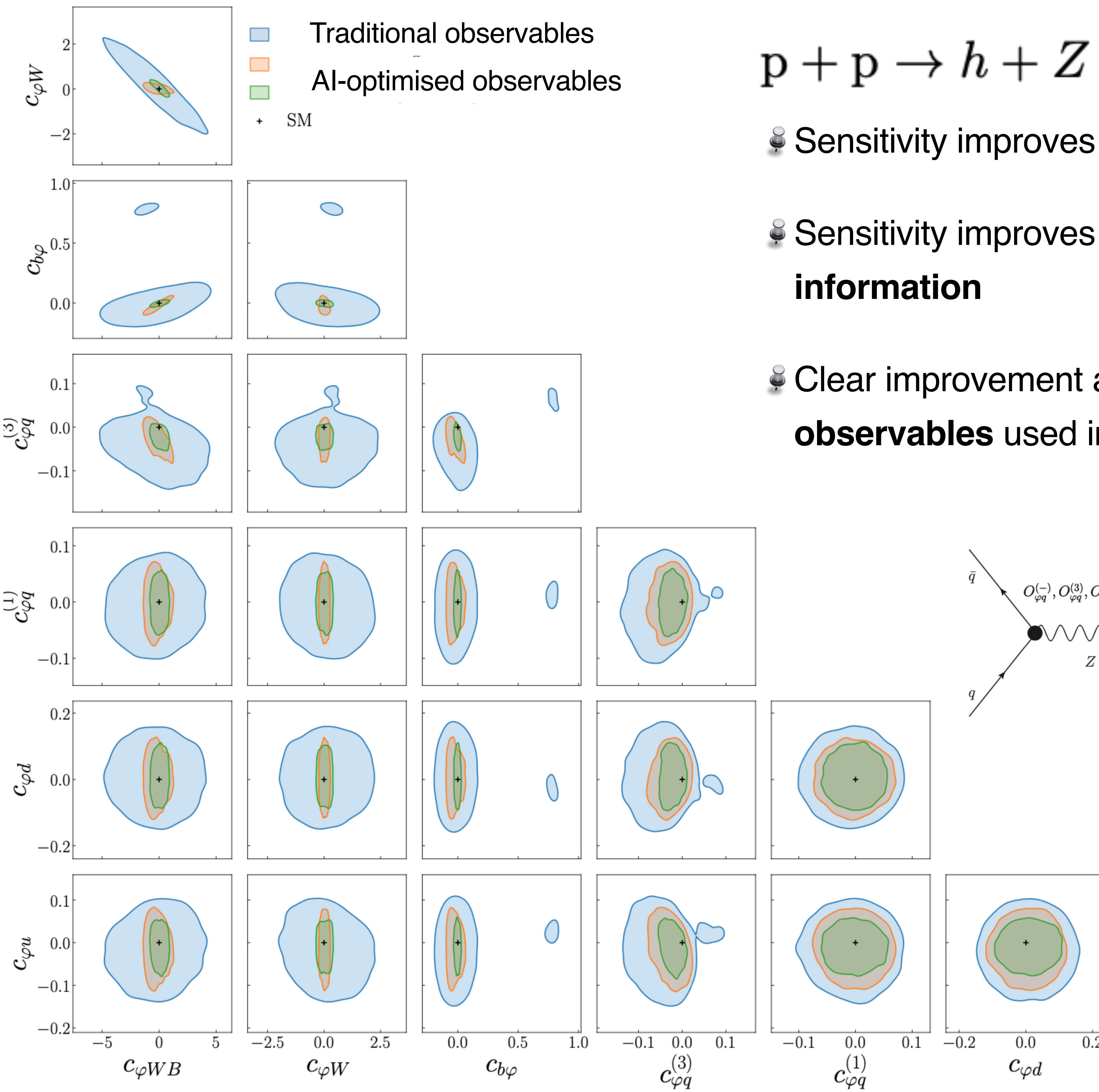
$$\boldsymbol{x} = (m_{t\bar{t}}, y_{t\bar{t}})$$

NN training by minimising cross-entropy loss function

$$L[g(\boldsymbol{x}, \boldsymbol{c})] = -\sigma_{\mathrm{fid}}(\boldsymbol{c}) \sum_{i=1}^{N_{\mathrm{ev}}} \log(1 - g(\boldsymbol{x}_i, \boldsymbol{c})) - \sigma_{\mathrm{fid}}(\boldsymbol{0}) \sum_{j=1}^{N_{\mathrm{ev}}} \log g(\boldsymbol{x}_j, \boldsymbol{c}) \qquad g = (1 + r_\sigma)^{-1}$$

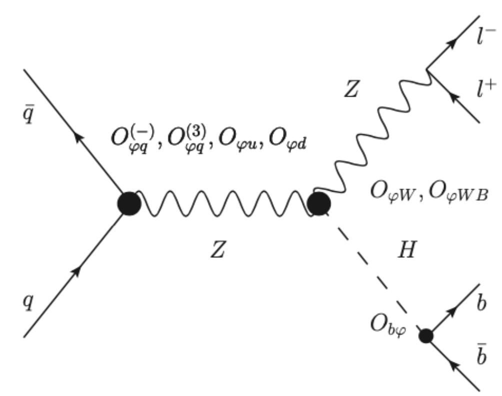
Results: Higgs+Z production

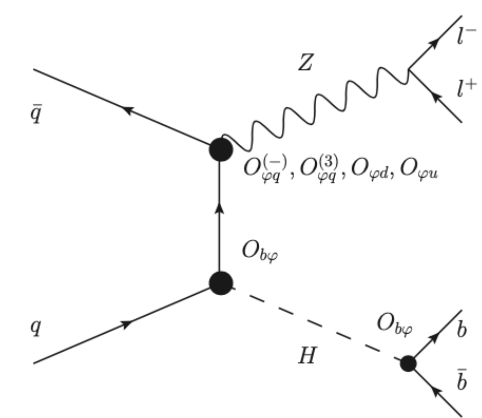
Marginalised 95 % C.L. intervals, $\mathcal{O}\left(\Lambda^{-4}\right)$ at $\mathcal{L} = 300\ \text{fb}^{-1}$



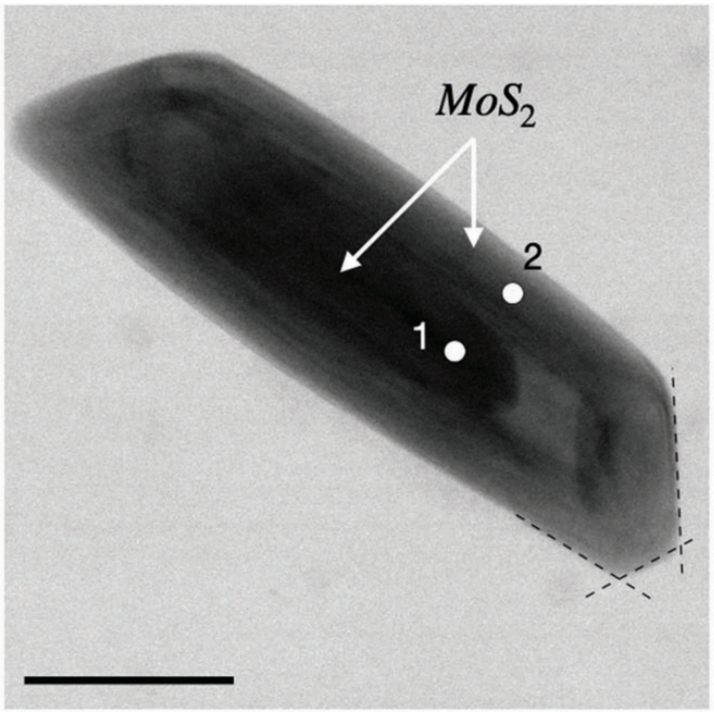
$$\mathrm{p} + \mathrm{p} \to h + Z \to b + \bar{b} + \ell^{+} + \ell^{-}$$

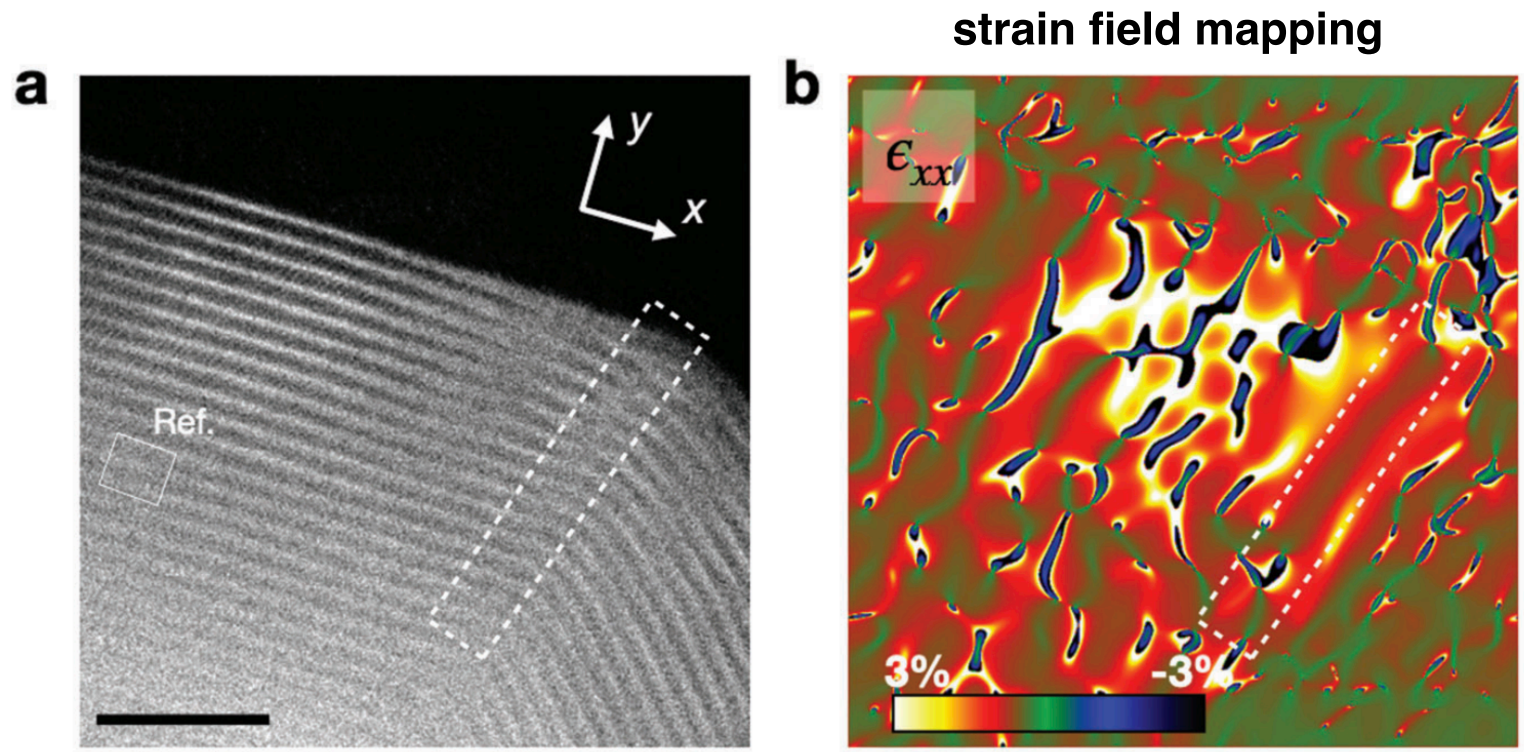
- Sensitivity improves in **unbinned analysis**
- Sensitivity improves when **using all kinematic information**
- Clear improvement as compared to **traditional observables** used in EFT fits

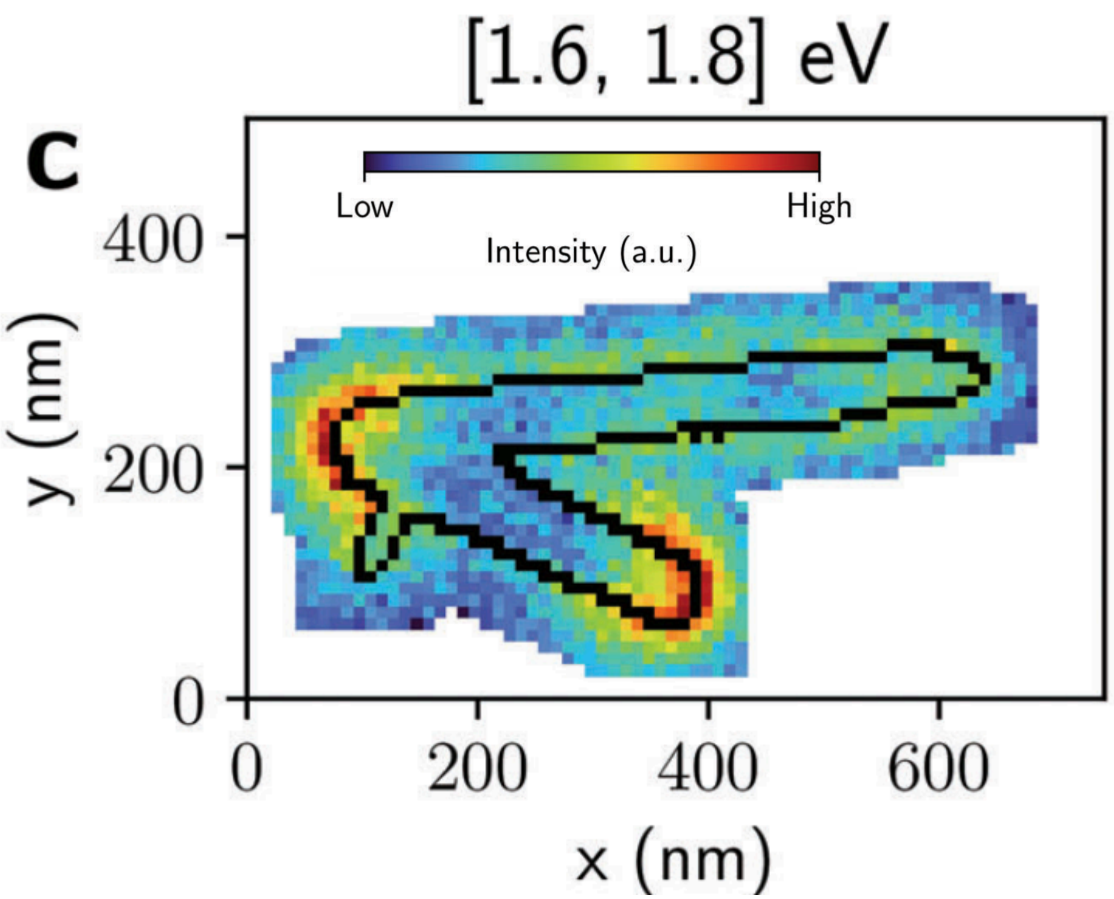




Applications to Material Science







- ☑ Background subtraction & anomaly detection problems from HEP arise also in material spectroscopy
- ☑ Direct correlation of **strain fields, band gap modulation, and exciton localisation** in 1D-MoS_2 nanostructures with different morphologies
- ☑ Developed EELSfitter ML framework together with TU Delft researchers

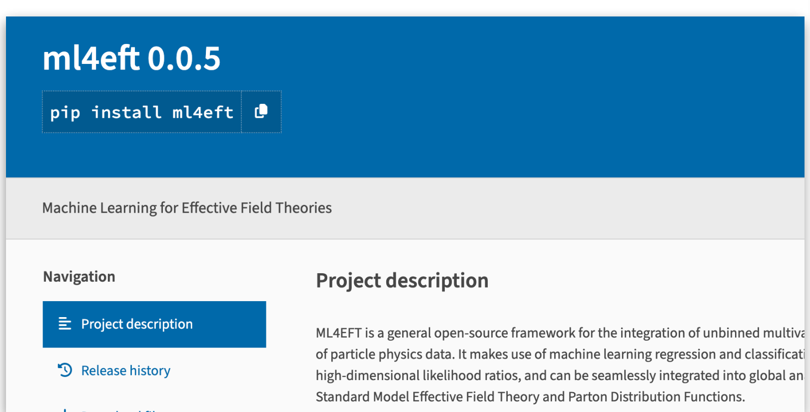
Summary and outlook

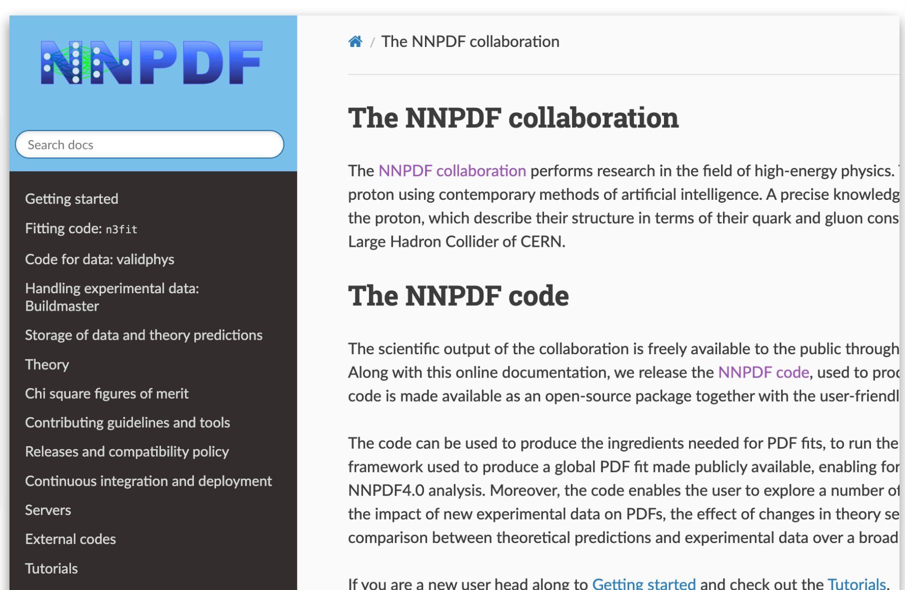
- Machine learning makes possible **identifying patterns in the data** whereby one can efficiently solve problems which are difficult of intractable with traditional approaches

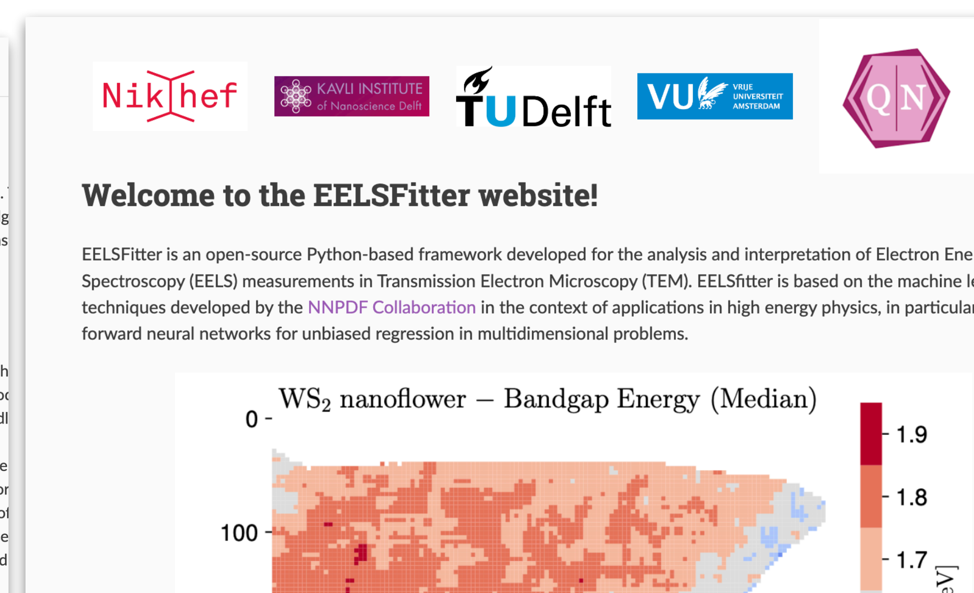
- **Enable discoveries** such as intrinsic charm quarks in the proton & make possible to **optimise the sensitivity of searches** for interesting phenomena hidden in the data

- Our technology is portable to many other problems, as demonstrated for their applicability to data analysis in **electron microscopy of quantum materials**

- Codes are **open source** and extensively documented, and have benefitted from contributions as well from BSc and MSc students in our groups







Summary and outlook

- Machine learning makes possible **identifying patterns in the data** whereby one can efficiently solve problems which are difficult of intractable with traditional approaches

- **Enable discoveries** such as intrinsic charm quarks in the proton & make possible to **optimise the sensitivity of searches** for interesting phenomena hidden in the data

- Our technology is portable to many other problems, as demonstrated for their applicability to data analysis in **electron microscopy of quantum materials**

- Codes are **open source** and extensively documented, and have benefitted from contributions as well from BSc and MSc students in our groups

Peter

Jaco

Giacomo

Tanjona

Tommaso