European Research Council

Established by the European Commission

MARIA UBIALI
UNIVERSITY OF CAMBRIDGE

STATISTICAL ASPECTS IN THE DETERMINATION OF THE PROTON'S STRUCTURE (PART II)

PHYSTAT SEMINAR SERIES

28TH MAY 2025

OUTLINE

➡ Physics background:
The Standard Model, the LHC, factorisation theorem & parton distributions functions

➡ Statistical formulation of the inverse problem of determining the proton structure from the LHC data and **the CT solution**

PAVEL

➡ **Part I** Statistical formulation of the inverse problem of determining the proton structure from the LHC data and **the NNPDF solution**

➡ **Part II** Statistical closure, experimental inconsistencies and simultaneous parameter determination

ME

➡ **Part III** (Some) open questions and discussion

➡ Part I Statistical formulation of the inverse problem of determining the proton structure from the LHC data and **the NNPDF solution**

A STATISTICAL FORMULATION OF THE PROBLEM

- Data **D** vector of dimension N_{dat} ~ O(5000), is obtained from many correlated collider measurements: **(ep)** with one proton or nucleus in the initial state and **(pp)** with two protons in the initial state.

- Experimentalists tell us that results are approximately distributed as a multivariate normal distribution, and provide us with a **mean value** and an experimental **covariance matrix Σ_{exp}** for the data.

$$D \sim \mathcal{N}(D_0, \Sigma_{\rm exp})$$

- Model predictions for the data **T** depend on a vector of physics parameters **c** (coupling constants, heavy quark and boson masses…) plus some continuous functions $\mathbf{f_i}$ that parametrise the proton nuclear substructure in terms of PDFs. Note that PDFs themselves depend on physics parameters.

$$T^{(\rm ep)}(\{f\}, \vec{c}) = \sum_{i=1}^{n_{\rm partons}} \int_x^1 dy\, f_i(y; \vec{c})\, \sigma_i\left(\frac{x}{y}; \vec{c}\right) \equiv \sum_{i=1}^{n_{\rm partons}} \left[f_i \otimes \sigma_i\right](x; \vec{c})$$

$$T^{(\rm pp)}(\{f\}, \vec{c}) = \sum_{i,j=1}^{n_{\rm partons}} \left[f_i \otimes f_j \otimes \sigma_{ij}\right](x; \vec{c})$$

A STATISTICAL FORMULATION OF THE PROBLEM

- For now let's leave the parameters **c = c*** fixed to some optimal value (e. g. PDG value).

$$T(\{f\}, \vec{c} = \vec{c}^{*}) \equiv T(\{f\})$$

- What theory constraints/indications on the functions **{f}** do we have from the theory?

✳ Observable predictions are necessarily positive, so $T(\{f\}) \geq 0$

✳ Parton distributions are **smooth** functions $f \in C^{\infty}[0,1]$.

✳ Parton distributions obey the '**Regge' limits**: as $x \to 0$ and as $x \to 1$, the **scaling behaviour** of the PDFs is a **power law**, $f(x) \sim x^{\alpha}$ and $f(x) \sim (1-x)^{\beta}$ respectively, where α, β are unknown.

✳ When we consider all flavours of PDFs, there exist **sum rules** which constrain integrals of linear combinations of the PDFs (e.g. $f_u - f_{\bar{u}}$ must integrate to 2).

A STATISTICAL FORMULATION OF THE PROBLEM

- Given these 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

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

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

A STATISTICAL FORMULATION OF THE PROBLEM

- Given these 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

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

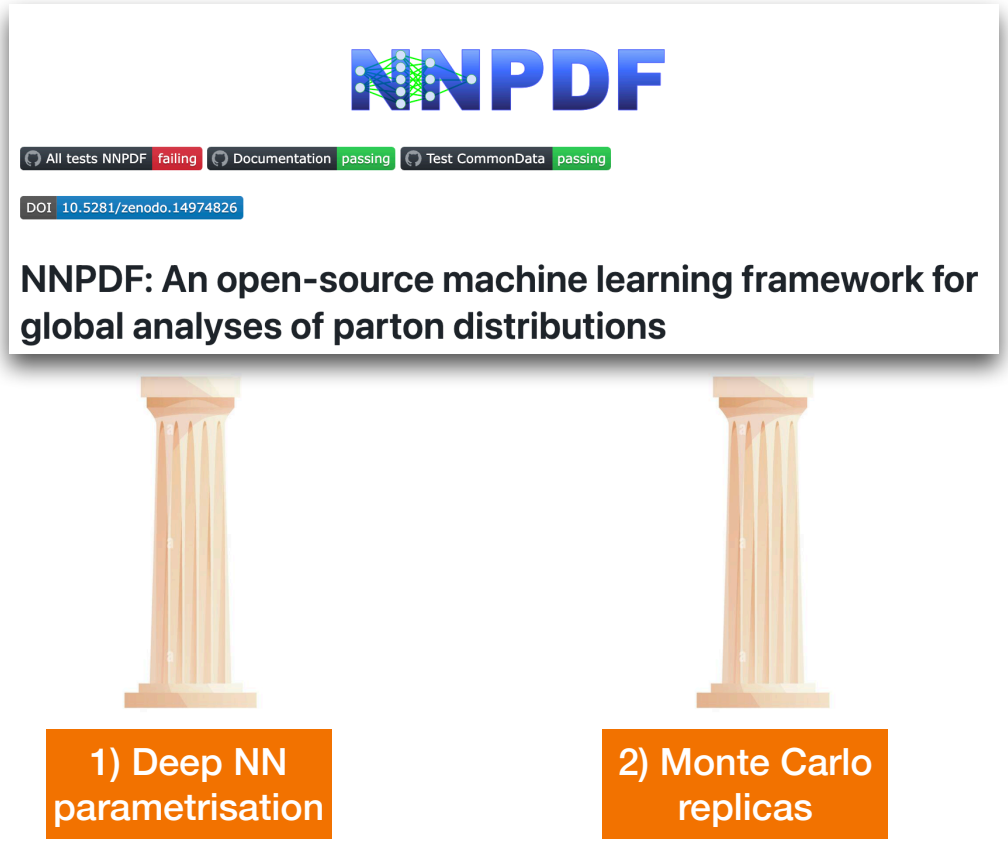
$$\chi^2 = \sum_{a,b=1}^{N_{\rm dat}} \left(T_a(\{f\},\vec{c}) - D_a\right) \Sigma_{ab}^{-1} \left(T_b(\{f\},\vec{c}) - D_b\right)$$

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

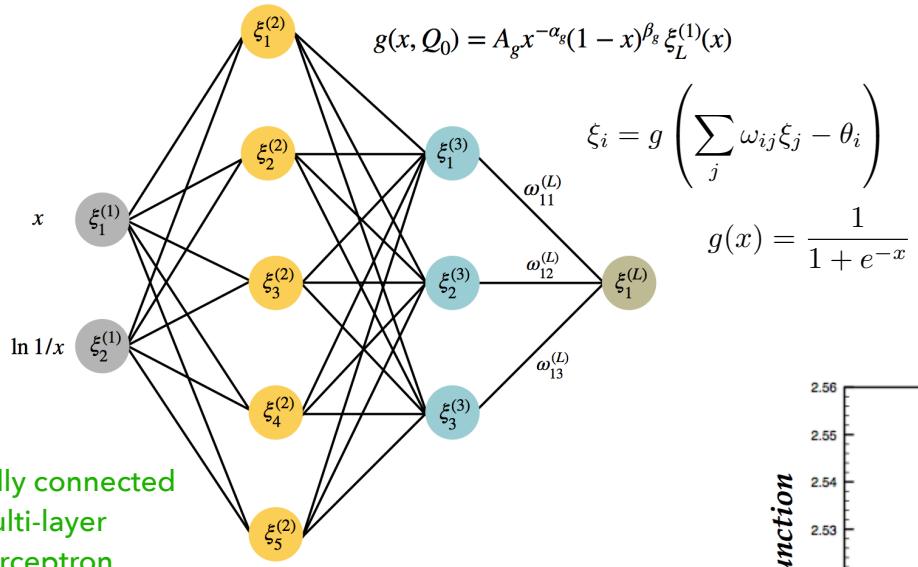
NNPDF APPROACH

- NNPDF produces regular fits of PDFs, since the NNPDF4.0 global fit in 2021 public code with up-to-date documentation that includes all theory and experimental inputs to reproduce all results and produce new ones https://github.com/NNPDF/nnpdf & https://docs.nnpdf.science/



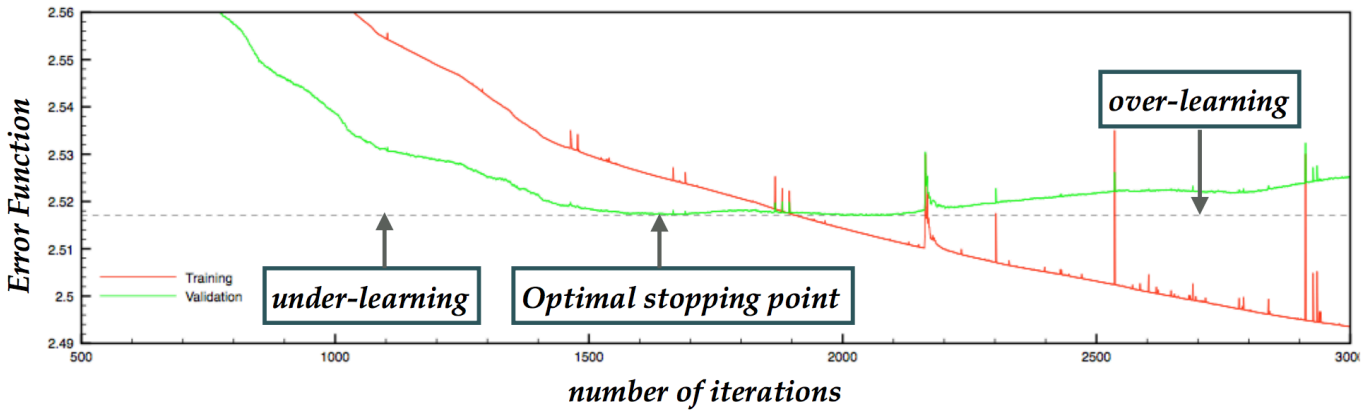
(1) NN PARAMETRIZATION

- Instead of using fixed parametrisation, choose parametrisation so large that in principle can fit any conceivable function **f**.
- One option is to use NNs [Forte, Latorre 2002]
- First PDF fit using NN to parametrise full set of PDFs [Ball, Del Debbio, Forte, Guffanti, Piccione, Rojo, MU, 2008]



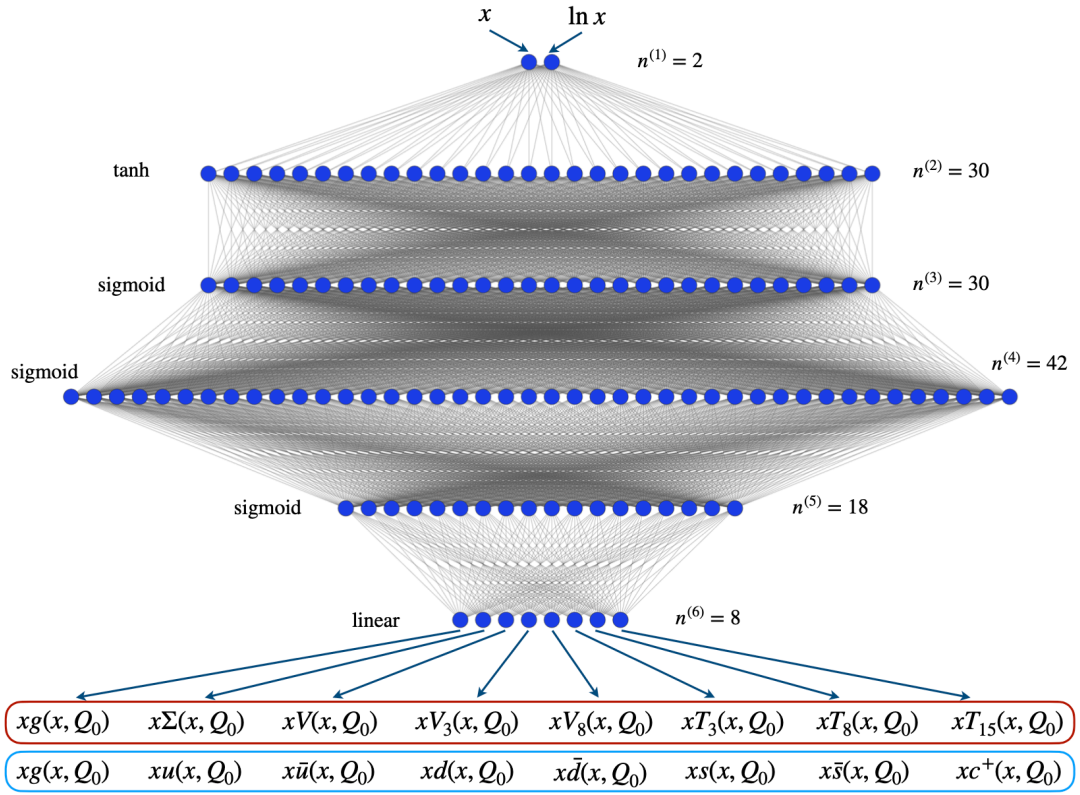
Fully connected multi-layer perceptron

• First NNPDF fits: each independent PDF associated to an unbiased and flexible parametrisation: O(300) parameters of seven (2-5-3-1) fully connected multi-layer perceptron versus O(30) in polynomial parametrisation
• Training and validation splitting to avoid over learning

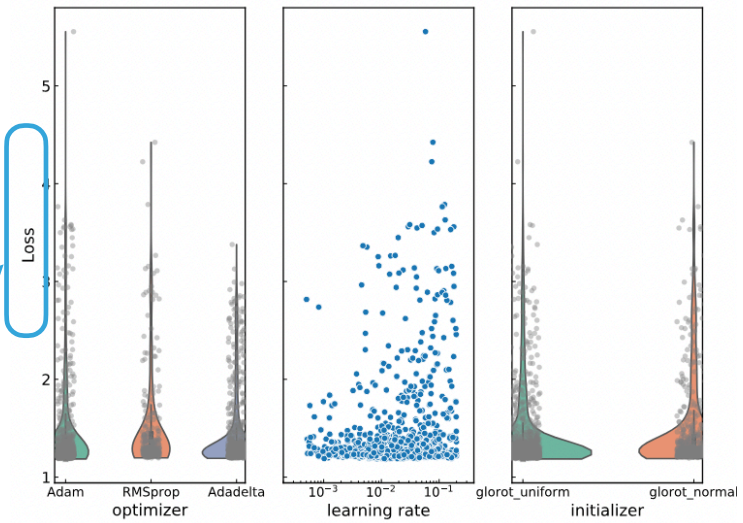


(1) DEEP NN PARAMETRIZATION

- NNPDF4.0: Single deep Neural Network (763 parameters) [Ball et al, 2021 arXiv:2109.02653]
- Hyper-parameter optimisation via K-fold procedure: scan parameter space and optimise K-fold loss.



K-fold loss



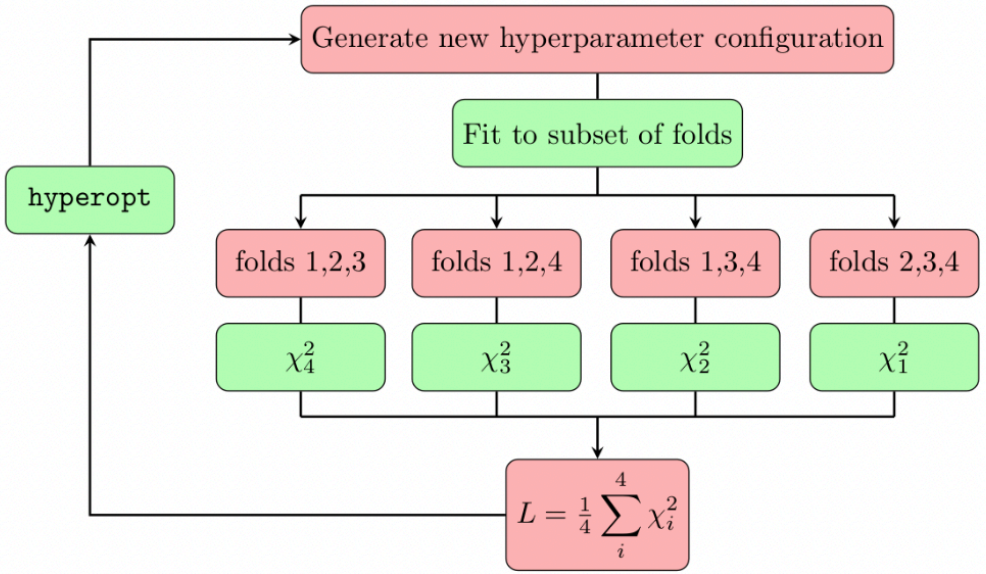
List of hyper-parameter

Neural Network	Fit options
Number of layers (*)	Optimizer (*)
Size of each layer	Initial learning rate (*)
Dropout	Maximum number of epochs (*)
Activation functions (*)	Stopping Patience (*)
Initialization functions (*)	Positivity multiplier (*)

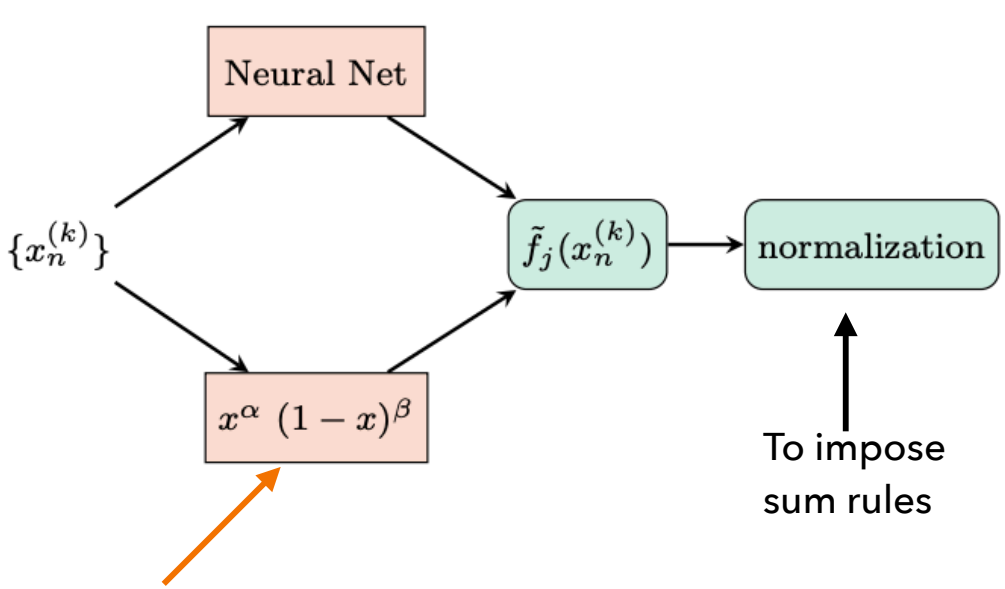
(1) DEEP NN PARAMETRIZATION

- NNPDF4.0: Single deep Neural Network (763 parameters) [Ball et al, 2021 arXiv:2109.02653]
- Hyper-parameter optimisation via K-fold procedure: scan parameter space and optimise K-fold loss.
- Each fold reproduces features of the full dataset, loss = average of the non-fitted folds => good generalisation

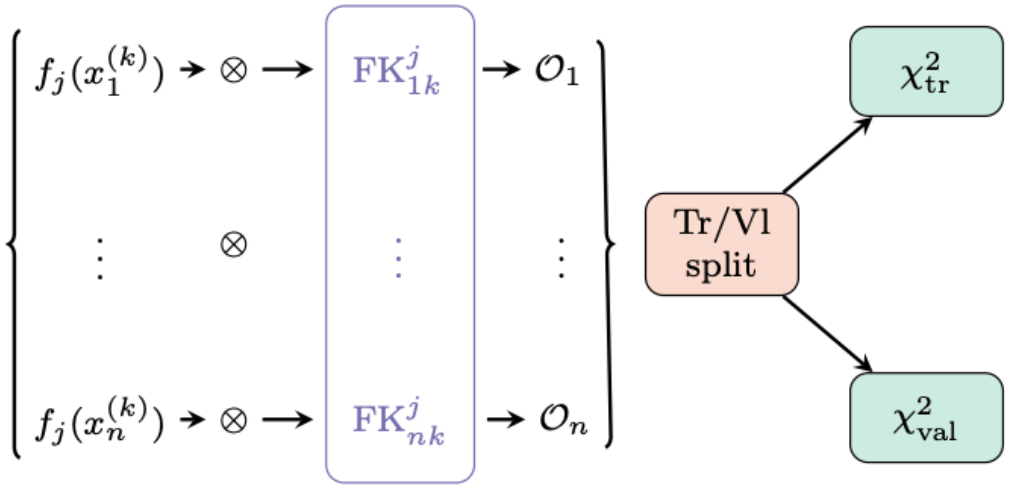
Fold 1		
CHORUS σ^{ν}_{CC}	HERA I+II inc NC e^+p 920 GeV	BCDMS p
LHCb Z 940 pb	ATLAS W,Z 7 TeV 2010	CMS Z p_T 8 TeV (p_T^{ll}, y_{ll})
DY E605 $\sigma^{p}_{\rm DY}$	CMS Drell-Yan 2D 7 TeV 2011	CMS 3D dijets 8 TeV
ATLAS single-$\bar{t}$ y (normalised)	ATLAS single top R_t 7 TeV	CMS $t\bar{t}$ rapidity $y_{t\bar{t}}$
CMS single top R_t 8 TeV		
Fold 2		
HERA I+II inc CC e^-p	HERA I+II inc NC e^+p 460 GeV	HERA comb. $\sigma^{\rm red}_{bb}$
NMC p	NuTeV σ^{p}_{c}	LHCb $Z \to ee$ 2 fb
CMS W asymmetry 840 pb	ATLAS Z p_T 8 TeV (p_T^{ll}, M_{ll})	D0 $W \to \mu\nu$ asymmetry
DY E886 $\sigma^{p}_{\rm DY}$	ATLAS direct photon 13 TeV	ATLAS dijets 7 TeV, R=0.6
ATLAS single antitop y (normalised)	CMS $\sigma^{\rm tot}_{tt}$	CMS single top $\sigma_t + \sigma_{\bar{t}}$ 7 TeV
Fold 3		
HERA I+II inc CC e^+p	HERA I+II inc NC e^+p 575 GeV	NMC d/p
NuTeV σ^{ν}_{c}	LHCb $W,Z \to \mu$ 7 TeV	LHCb $Z \to ee$
ATLAS W,Z 7 TeV 2011 Central selection	ATLAS W^++jet 8 TeV	ATLAS HM DY 7 TeV
CMS W asymmetry 4.7 fb	DYE 866 $\sigma^{d}_{\rm DY}/\sigma^{p}_{\rm DY}$	CDF Z rapidity (new)
ATLAS $\sigma^{\rm tot}_{tt}$	ATLAS single top y_t (normalised)	CMS $\sigma^{\rm tot}_{tt}$ 5 TeV
CMS $t\bar{t}$ double diff. $(m_{t\bar{t}}, y_t)$		
Fold 4		
CHORUS σ^{p}_{CC}	HERA I+II inc NC e^+p 820 GeV	LHCb $W,Z \to \mu$ 8 TeV
LHCb $Z \to \mu\mu$	ATLAS W,Z 7 TeV 2011 Fwd	ATLAS W^-+jet 8 TeV
ATLAS low-mass DY 2011	ATLAS Z p_T 8 TeV (p_T^{ll}, y_{ll})	CMS W rapidity 8 TeV
D0 Z rapidity	CMS dijets 7 TeV	ATLAS single top y_t (normalised)
ATLAS single top R_t 13 TeV	CMS single top R_t 13 TeV	



FROM DNN TO THEORY PREDICTIONS



α, β are pre-processing exponents randomised within a range for each parton combination and iterated until range is stable



The convolution of the PDFs with the theory functions σ is done by computing PDFs on a grid of $x^{(k)}$ points and multiplying the by a precomputed FK table

(2) THE MONTE CARLO SAMPLING FOR ERROR PROPAGATION

- In such a large parameter space Hessian approach not applicable.
- Use Monte Carlo - or bootstrap - error propagation by importance sampling

[Giele, Kosover 1993] [Forte et al, 2006]

$$D_0 = t + \eta$$

$$\eta \sim \mathcal{N}(0, \Sigma_{\rm exp})$$

Vector of central experimental values

Vector of "true", thus unknown experimental values

Observational noise

For now let's only consider experimental covariance matrix. In the NNPDF approach a theory covariance matrix is also included and added to the experimental covariance matrix

$$\epsilon^{(k)} \sim \mathcal{N}(0, \Sigma_{\rm exp})$$

$$\mu^{(k)} = D_0 + \epsilon^{(k)} = t + \eta + \epsilon^{(k)}$$

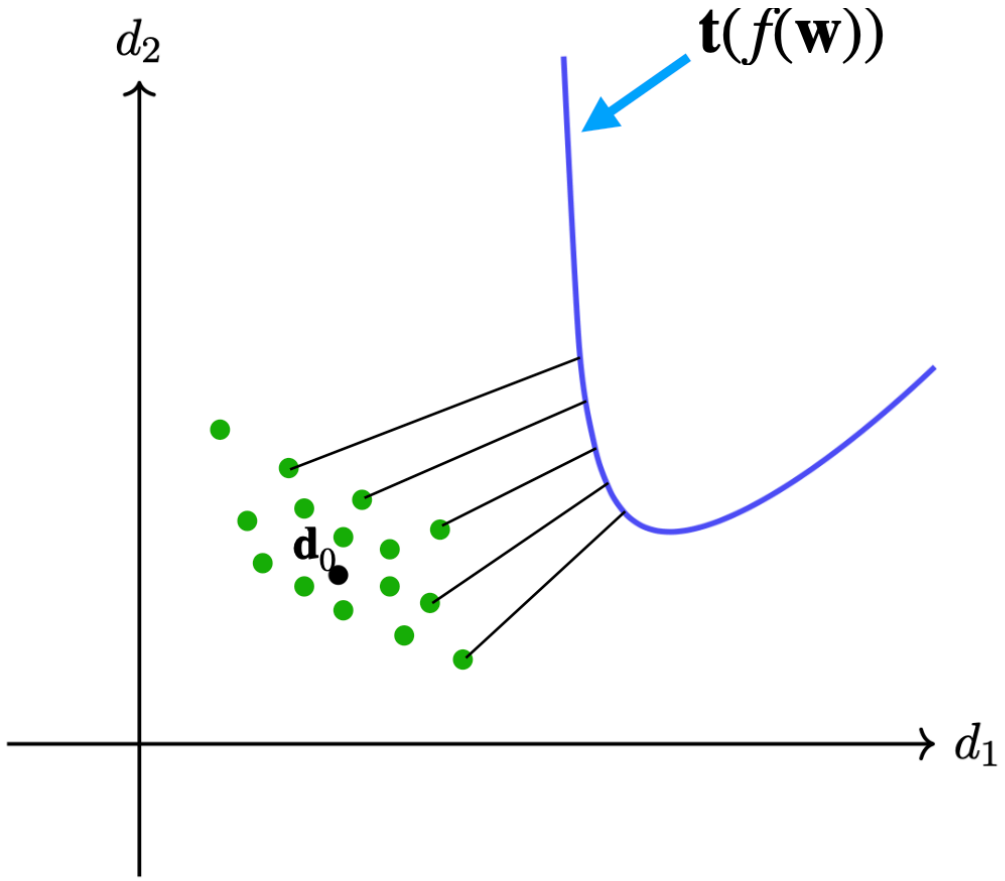
Pseudo-data replicas, k = 1, ..., N_{rep}

(2) THE MONTE CARLO SAMPLING FOR ERROR PROPAGATION

- Visually, fitting pseudo data = throwing random pseudo data points about the experimental data $\mathbf{D_0}$, according to a multivariate normal distributions centred on the experimental data $\mathbf{D_0}$ with experimental covariance matrix $\boldsymbol{\Sigma}_{\text{exp}}$.

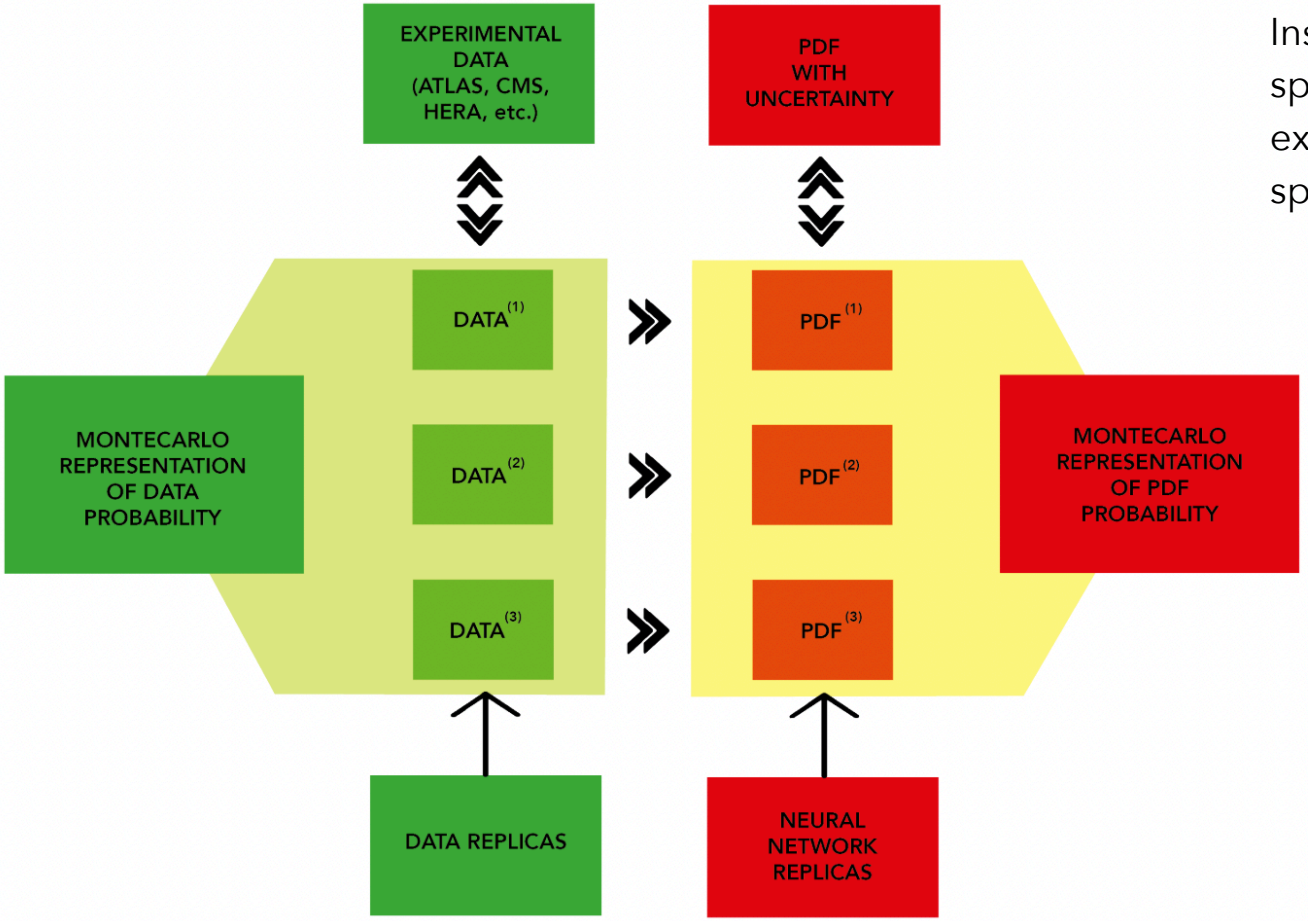
- For each pseudo data compute the **optimal point on the theory surface** based on training-validation splitting and minimisation stopping procedure and obtain associated parameter values.

- Repeating gives an approximation to the **parameter distribution** by importance sampling.

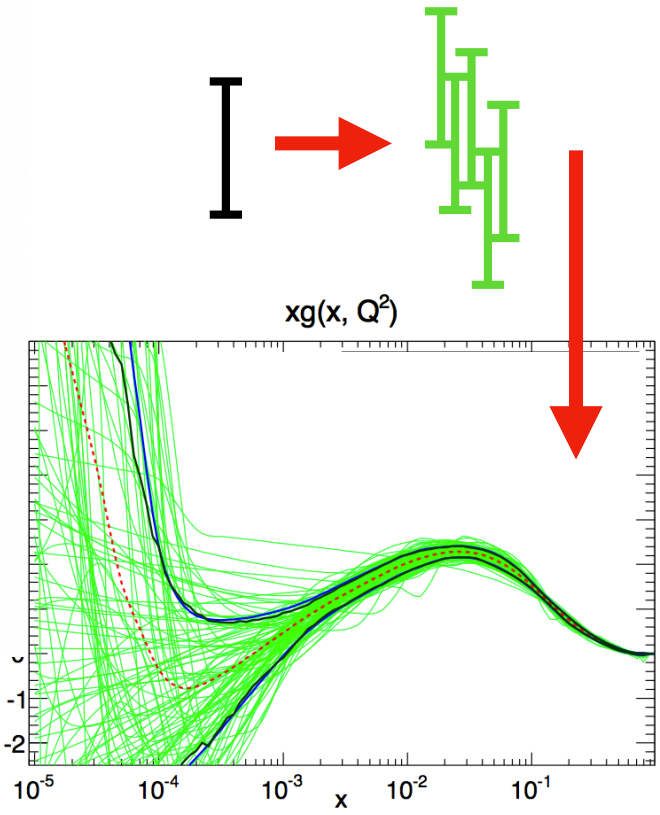


Costantini et al, arXiv:2404.10056

(2) THE MONTE CARLO SAMPLING FOR ERROR PROPAGATION



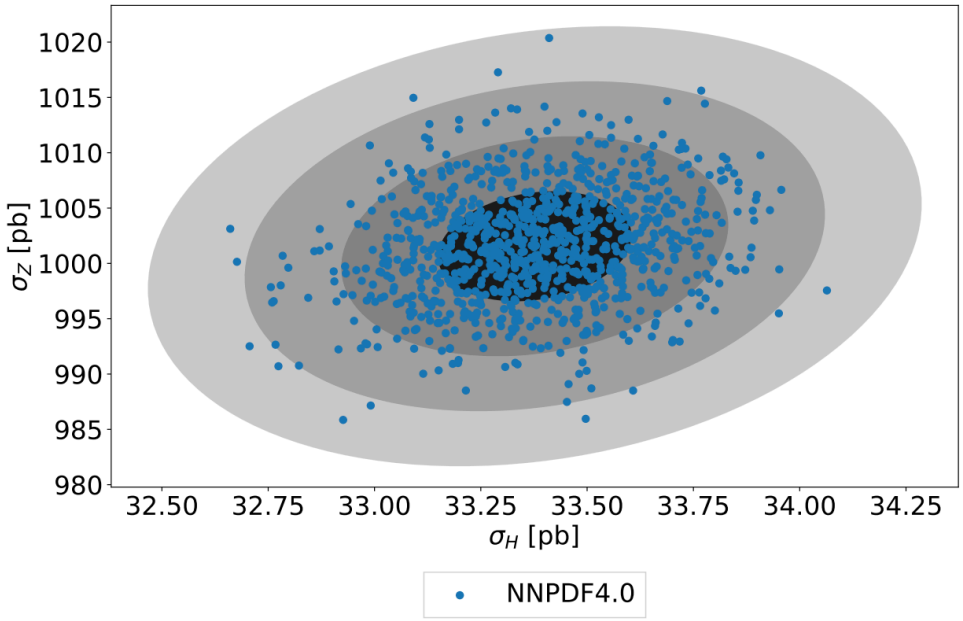
Instead of sampling directly in PDF functional space, sampling is done at the level of experimental error and "projected" in the PDF space.

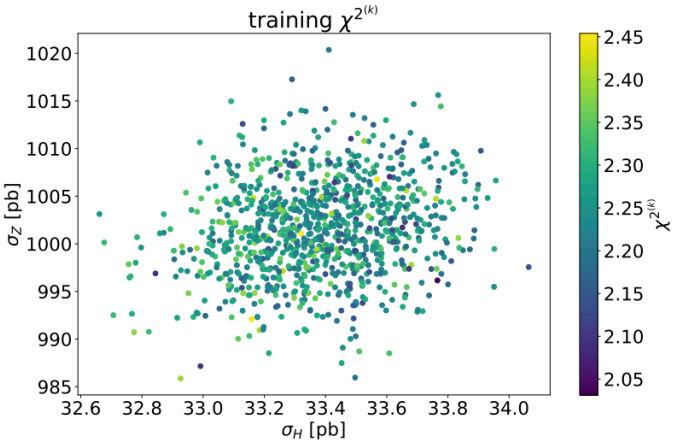


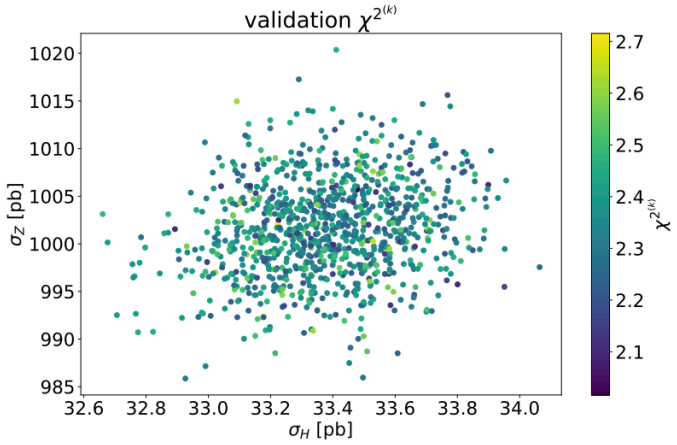
(2) THE REPLICA DISTRIBUTION

- Using results of the optimisation for each replica (PDF replicas) we can make predictions for observables that depend on PDFs. In prediction space: Gaussian distributions.
- Data uncertainty => PDF replica fluctuations
- Interpolation, extrapolation and functional uncertainties => best fit degeneracy
- No correlation between fit quality and position in the Z-H plane => uniform fit quality

Ball et al, 2211.12961



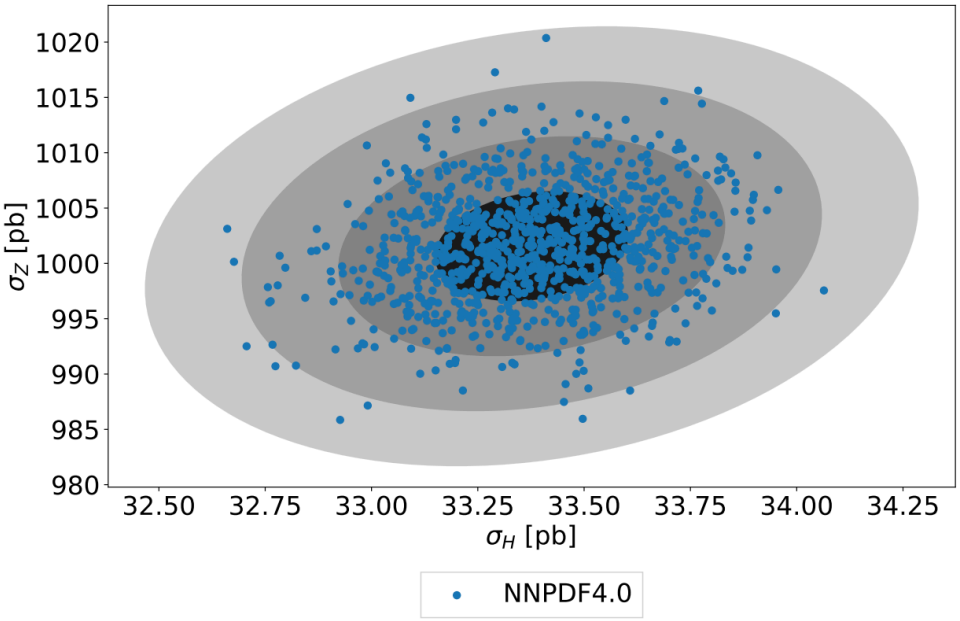




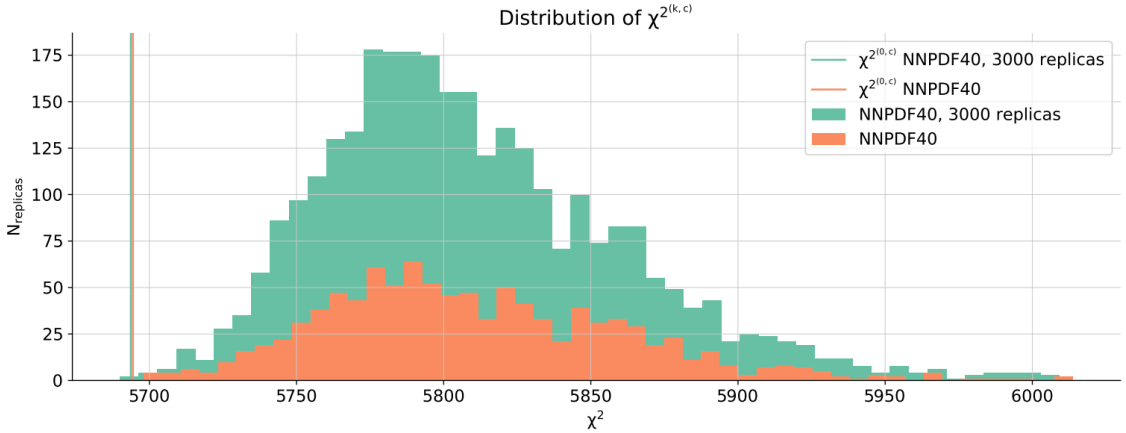
(2) THE REPLICA DISTRIBUTION

- Using results of the optimisation for each replica (PDF replicas) we can make predictions for observables that depend on PDFs. In prediction space: Gaussian distributions.
- Data uncertainty => PDF replica fluctuations
- Interpolation, extrapolation and functional uncertainties => best fit degeneracy
- No correlation between fit quality and position in the Z-H plane => uniform fit quality

Ball et al, 2211.12961



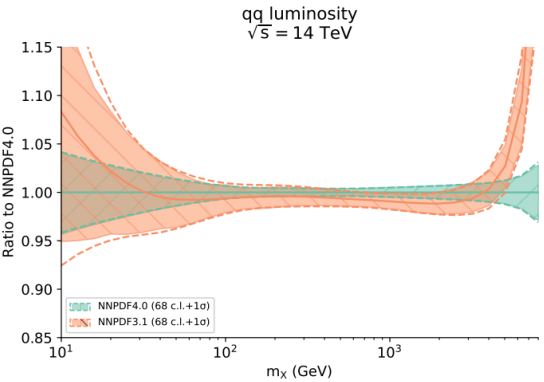
1000 REPLICAS vs. 3000 REPLICAS

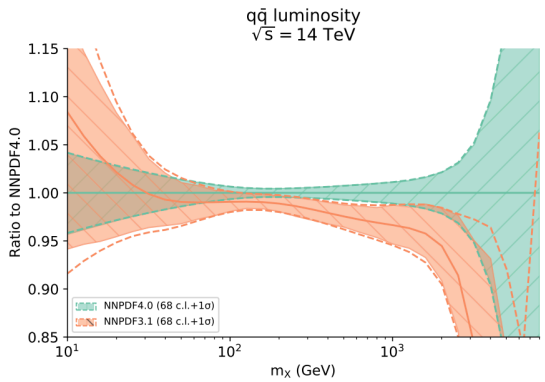


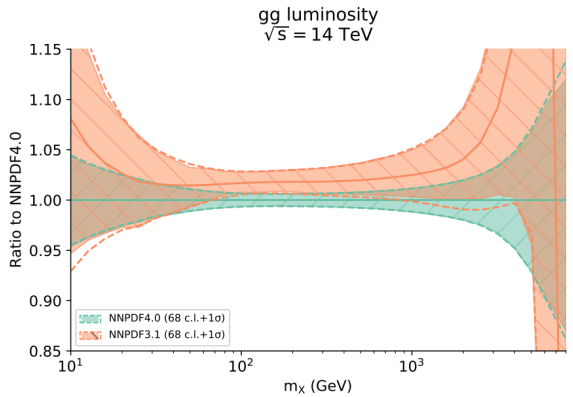
Fit quality of each replica to the central data D_0 statistically distributed.
Average replica (best fit PDF) not necessarily the one with lowest chi2

THE ROLE OF DATA AND METHODOLOGY IN A PDF FIT

NNPDF3.1 vs NNPDF4.0

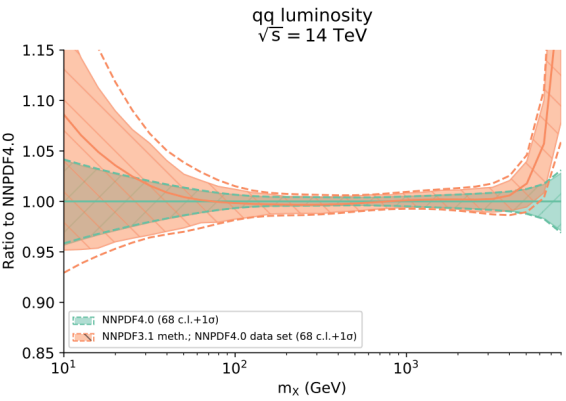


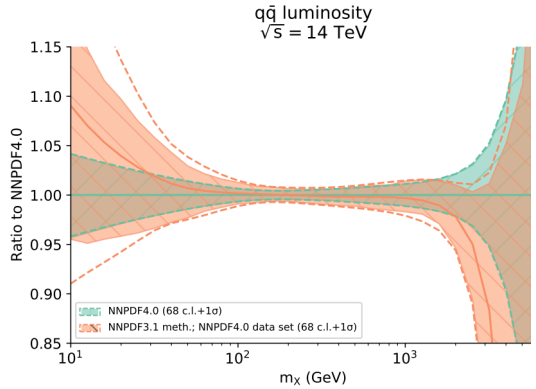


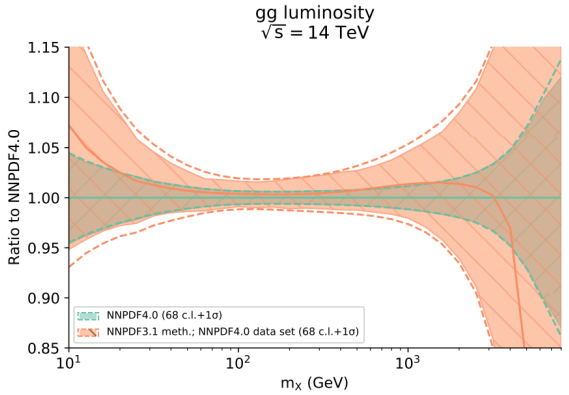


METHODOLOGY / DATA	NNPDF3.1	NNPDF4.0
NNPDF3.1 (4093)	1.19	1.12
NNPDF4.0 (4491)	1.25	1.17

NNPDF4.0 data - NNPDF3.1 vs NNPDF4.0 methodology







METHODOLOGY / DATA	NNPDF3.1	NNPDF4.0
NNPDF3.1 (4093)	1.19	1.12
NNPDF4.0 (4491)	1.25	1.17

Ball et al, 2021 arXiv:2109.02653

- Shift in parton luminosities mostly due to inclusion of O(500) more data points
- Parton luminosities based on same dataset are consistent with each other but 4.0 methodology displays smaller uncertainty than 3.1 methodology

➡ **Part II** Statistical closure test, experimental inconsistencies and simultaneous parameter fits

(3) CLOSURE TESTS

- What checks are done on the methodology?
- Assume a given underlying law of Nature: e.g. NNLO predictions for partonic cross section and a given PDF set (for example a random NNPDF replica)
- Generate data with central values given by the "true" law of Nature, and distributed according to experimental covariance matrix.
- Run a fit with NNPDF methodology
- Do statistics on "runs of the universe": generate D_0 N_{fit} times with different random noise drawn from experimental distribution

L1 data, each corresponding to a "run of the universe"

$$D_0 = t + \eta$$

$$t = G(w^0)$$

Experimental noise

Pretend know the "true" value of observables, unknown in real life but identified with G in clos. test

$$\hat{\sigma}_{\rm NNLO} \otimes (f \otimes f)_{\text{"true" PDF}}$$

L2 pseudo-data (to propagate uncertainties in Monte Carlo fits)

$$\mu^{(k)} = D_0 + \epsilon^{(k)} = t + \eta + \epsilon^{(k)}$$

$$k = 1, ..., N_{\rm rep}$$

(3) STATISTICAL ESTIMATORS

$$u_{*,k} = \underset{u_k}{\text{argmin}} \left[\chi^{2(k)}_{\text{val}} \middle| \underset{u_k}{\text{argmin}}\, \chi^{2(k)}_{\text{tr}} \right], \quad k = 1, ..., N_{\text{rep}}$$

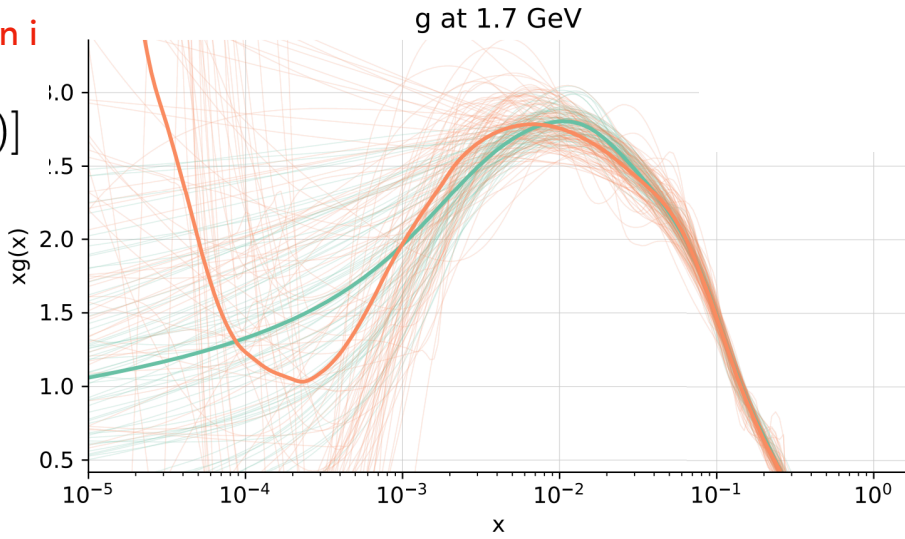
Best fit for each data replica in Monte Carlo approach

$$\left(\bar{\mathcal{G}}\right)_i = \mathbb{E}_\epsilon \left[\mathcal{G}(u_{*,k})_i\right]$$ → Central value of prediction i

$$\left(\Delta_{\text{PDF}}\right)_i = \mathbb{E}_\epsilon \left[\mathcal{G}(u_{*,k})_i - \mathbb{E}_\epsilon \mathcal{G}(u_{*,k})_i\right]$$ → PDF uncertainty of prediction i

$$\left(C_{\text{PDF}}\right)_{ij} = \frac{N_{\text{reps}}}{N_{\text{reps}}-1} \mathbb{E}_\epsilon \left[\left(\mathcal{G}(u_{*,k})_i - \mathbb{E}_\epsilon \mathcal{G}(u_{*,k})_i\right)\left(\mathcal{G}(u_{*,k})_j - \mathbb{E}_\epsilon \mathcal{G}(u_{*,k})_j\right)\right]$$

PDF covariance matrix of predictions i and j
(PDF induced correlation among observables)



- Key estimate is the normalised **bias**:
  measure mean square deviation of predictions from the "truth" in units of predicted standard deviation

$$B^{(l)}(C_{\text{PDF}}) = \frac{1}{N_{\text{data}}} \sum_{i,j=1}^{N_{\text{data}}} \left(\mathbb{E}_\epsilon \mathcal{G}(u^{(l)}_{*,k})_i - f_i\right) \left(\overline{C}_{\text{PDF}}\right)^{-1}_{ij} \left(\mathbb{E}_\epsilon \mathcal{G}(u^{(l)}_{*,k})_j - f_j\right)$$

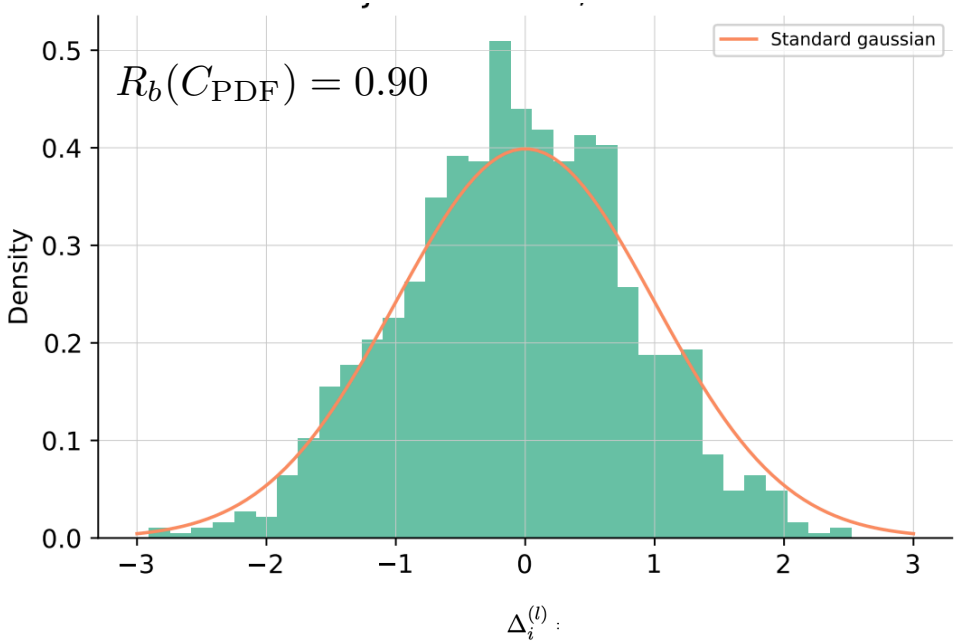
$$R_b = \sqrt{\mathbb{E}_\eta B^{(l)}(\overline{C}_{\text{PDF}})}$$

Bias averaged over
N_{fit} L1 data, i.e. N_{fit} runs of
the universe, l = 1, ..., N_{fit}

$R_b \sim 1$ faithfully estimated PDF uncertainties
$R_b < 1$ overestimated PDF uncertainties
$R_b > 1$ underestimated PDF uncertainties

(3) CONSISTENT CLOSURE TEST

NNPDF4.0 global fit: distributions of Δi



Barontini et al, 2503.17447

- Normalised **bias** in the basis of the eigenvectors of the PDF covariance matrix

$$B^{(l)}(\overline{C}_{\rm PDF}) = \frac{1}{N_{\rm data}} \sum_{i=1}^{N_{\rm data}} \frac{(\Delta_i^{(l)})^2}{(\sigma_{{\rm PDF}i})^2}$$

- With Δi being the projection of the Bias along each normalised eigenvector of C_{PDF}

$$\Delta_i^{(l)} = \sum_{j=1}^{N_{\rm data}} \left(\mathbb{E}_\epsilon \mathcal{G}(u_{*,k}^{(l)})_j - f_j \right) v_j^{(i)}$$

$$B^{(l)}(C_{\rm PDF}) = \frac{1}{N_{\rm data}} \sum_{i,j=1}^{N_{\rm data}} \left(\mathbb{E}_\epsilon \mathcal{G}(u_{*,k}^{(l)})_i - f_i \right) (\overline{C}_{\rm PDF})_{ij}^{-1} \left(\mathbb{E}_\epsilon \mathcal{G}(u_{*,k}^{(l)})_j - f_j \right)$$

(4) MODELLING EXPERIMENTAL INCONSISTENCIES

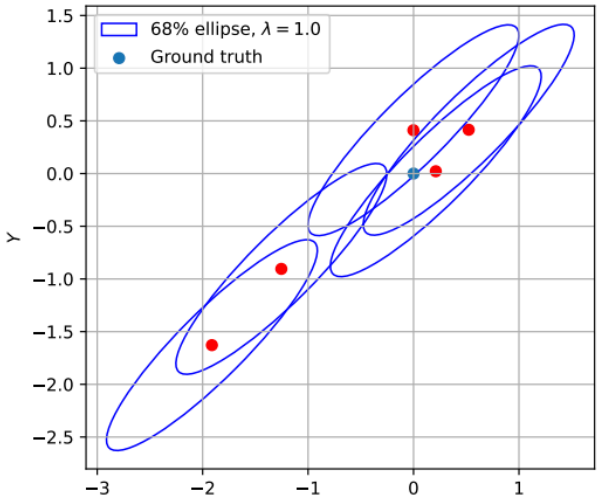
- How to model an inconsistency of experimental origin = some underestimated experimental systematics?

- Generate L_1 data with "true" experimental covariance matrix

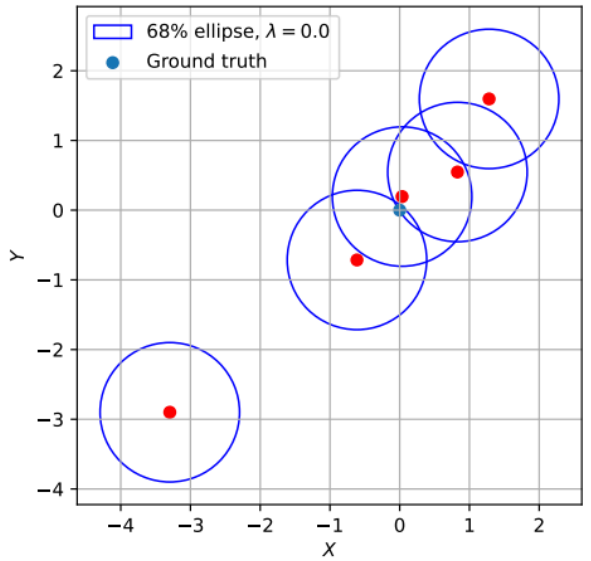
$$(C_{\text{exp}})_{ij} = \delta_{ij}\sigma_i^{(\text{uncorr})}\sigma_j^{(\text{uncorr})} + \sum_{k=1}^{N_{\text{corr}}} \sigma_{i,k}^{(\text{corr})}\sigma_{j,k}^{(\text{corr})}$$

- Fit the data using the rescaled experimental covariance matrix (both in pseudo data generation and in the loss function)

$$(C_{\text{exp}}^{\lambda})_{ij} = \delta_{ij}\sigma_i^{(\text{uncorr})}\sigma_j^{(\text{uncorr})} + \sum_{k=1}^{N_{\text{corr}}} \lambda_{i,k}\sigma_{i,k}^{(\text{corr})}\lambda_{j,k}\sigma_{j,k}^{(\text{corr})}$$



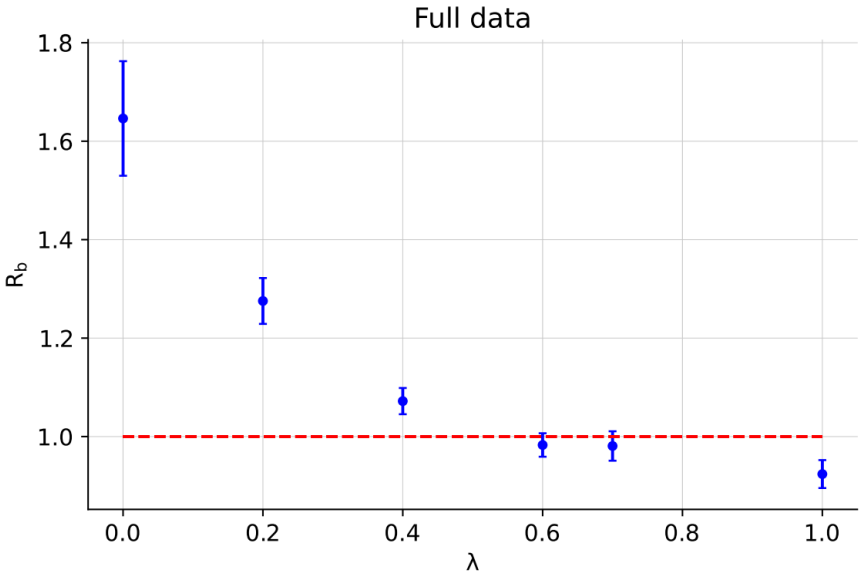
Consistent
$\lambda = 1$

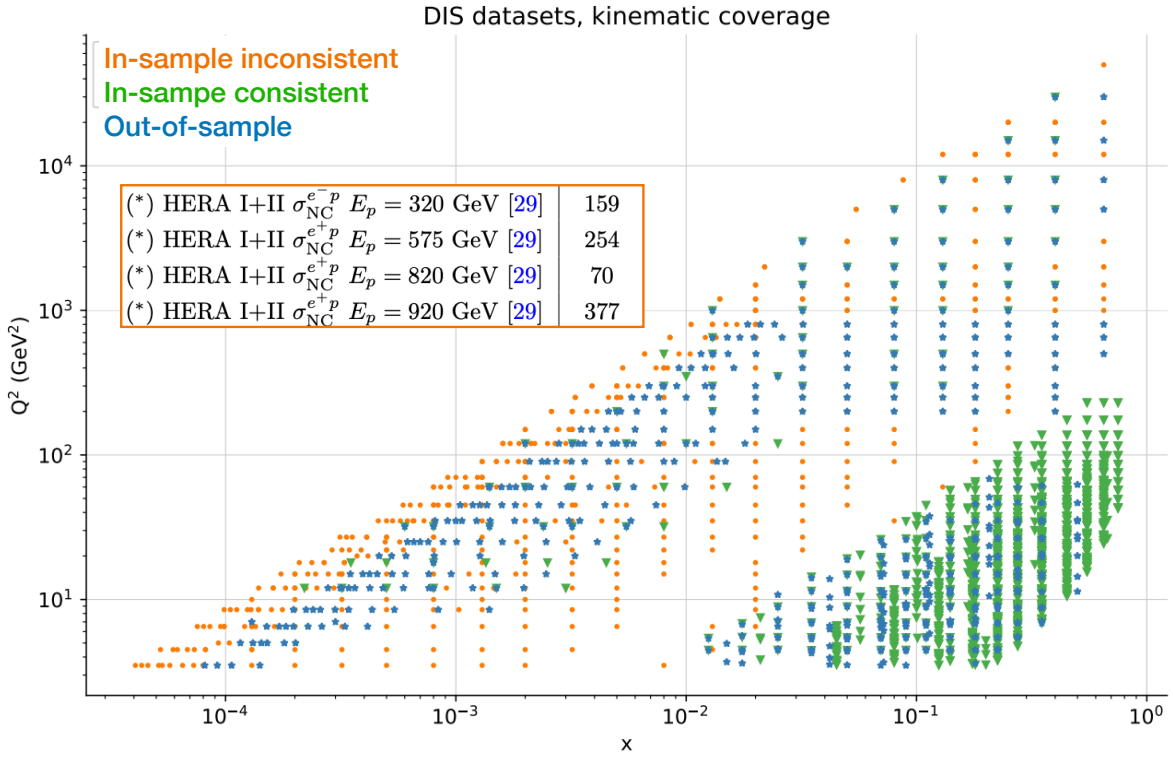


Extreme inconsistency
$\lambda = 0$

(4) BULK INCONSISTENCY

- DIS only fit, in-sample HERA NC data are inconsistent.
- **860 out of 2576** inconsistent datapoints, with all systematic uncertainties underestimated by the same factor λ.

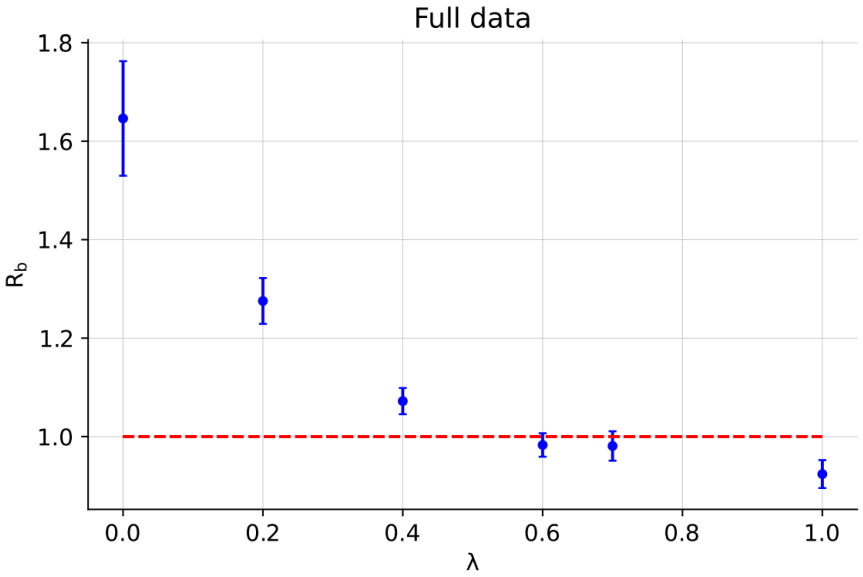


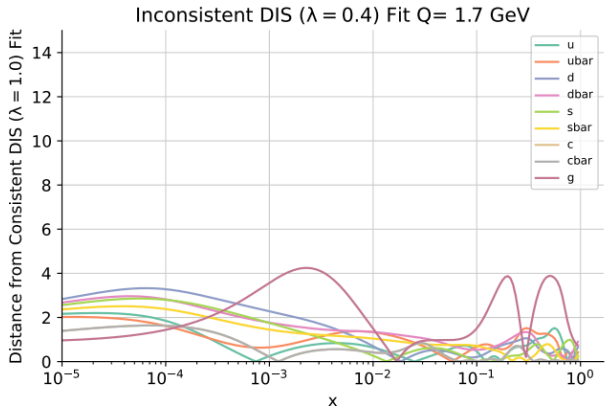


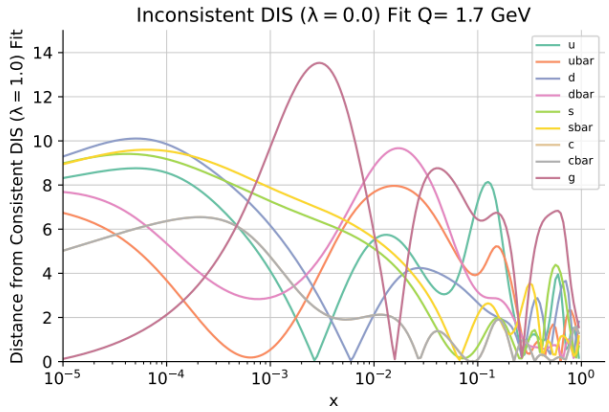
- Highly **non linear behaviour**: for $\lambda \geq 0.4$ despite inconsistency, PDF uncertainties remain faithful, but then sharply rises.
- In-sample and out-of-sample datasets behave in a similar way => NN model effective at generalising.

(4) BULK INCONSISTENCY

- DIS only fit, in-sample HERA NC data are inconsistent.
- **860 out of 2576** inconsistent datapoints, with all systematic uncertainties underestimated by the same factor λ.





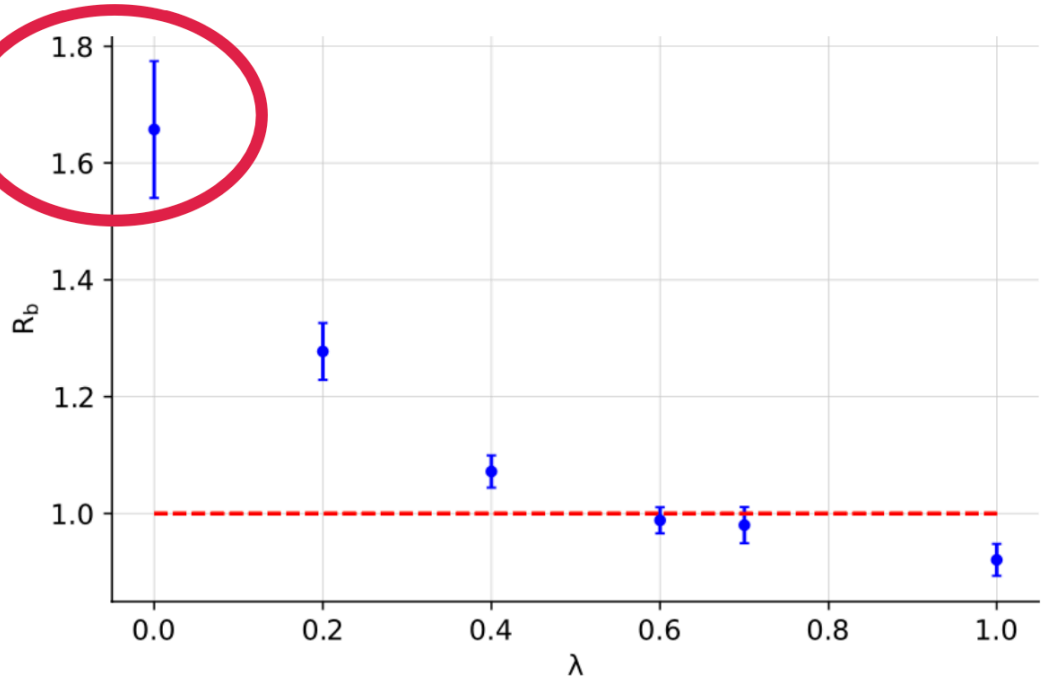


$\lambda \geq 0.4$: Model corrects for underestimated uncertainty in the inconsistent data, PDF uncertainties do not decrease despite the reduced data uncertainty.

$\lambda < 0.4$: PDF uncertainty shrinks => underestimated PDF uncertainties, largest shifts in the gluon and quark singlet combinations.

- Highly non linear behaviour: for $\lambda \geq 0.4$ despite the inconsistency, PDF uncertainties remain faithful, but then sharply rises.
- In-sample and out-of-sample datasets behave in a similar way => NN model effective at generalising.

(4) A CRITERION TO SPOT EXPERIMENTAL INCONSISTENCIES



- So far learn that NN model cures experimental inconsistencies until they are not too big
- If strong inconsistencies model fails, as either PDF uncertainties shrink or PDFs shift far from underlying law.
- In real life no access to the “truth”, normalised bias cannot be computed and only a single run of the Universe.
- How to spot inconsistencies in a actual PDF fit?

(4) A CRITERION TO SPOT EXPERIMENTAL INCONSISTENCIES

Large R_b → Large χ^2 of the inconsistent dataset and of consistent datasets correlated to it by the PDFs

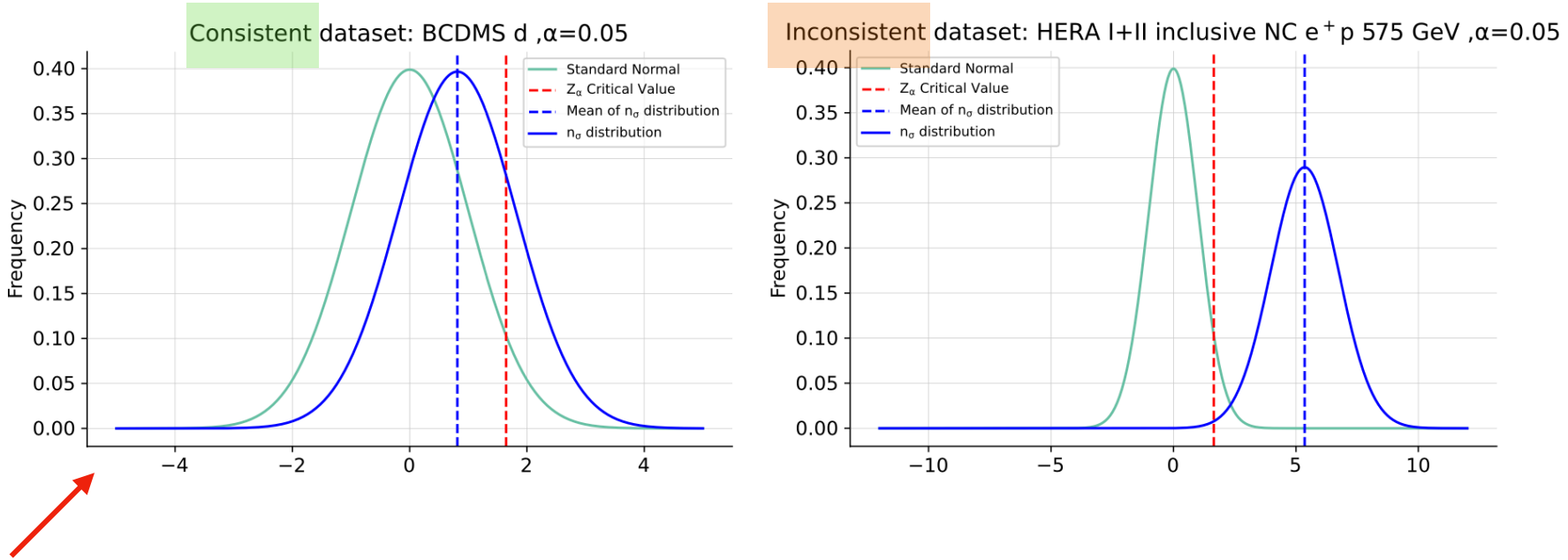
$$n_\sigma^{(i)} = \frac{\chi_i^2 - N_{\text{data}}^{(i)}}{\sqrt{2/N_{\text{data}}^{(i)}}}$$

Normalised deviation of χ^2 from its mean

$$\mu_i = \left\langle n_\sigma^{(i)} \right\rangle_{N_{\text{fit}}}$$

$$\mu_i > Z$$

Z: Critical threshold



Distribution of n_σ values across N_{fits} in DIS case with maximal inconsistency ($\lambda = 0$)

(4) A CRITERION TO SPOT EXPERIMENTAL INCONSISTENCIES

Large R_b → Large χ^2 of the inconsistent dataset and of consistent datasets correlated to it by the PDFs

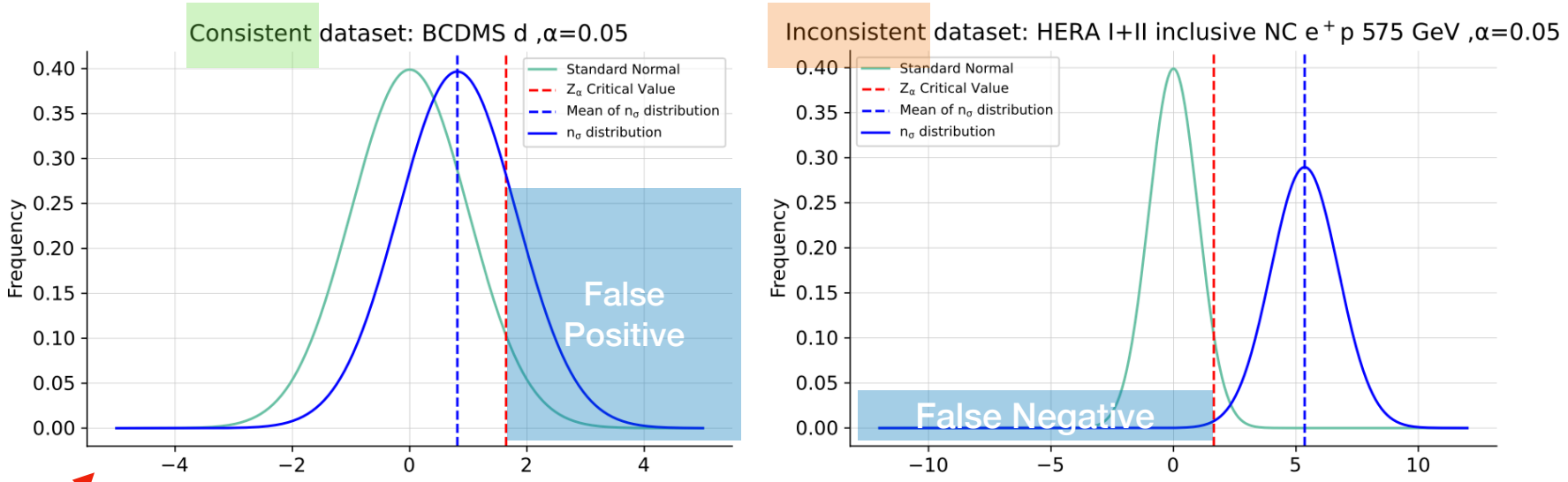
$$n_\sigma^{(i)} = \frac{\chi_i^2 - N_{\text{data}}^{(i)}}{\sqrt{2/N_{\text{data}}^{(i)}}}$$

Normalised deviation of χ^2 from its mean

$$\mu_i = \left\langle n_\sigma^{(i)} \right\rangle_{N_{\text{fit}}}$$

$$\mu_i > Z$$

Z: Critical threshold



Distribution of n_σ values across N_{fits} in DIS case with maximal inconsistency ($\lambda = 0$)

(4) A CRITERION TO SPOT EXPERIMENTAL INCONSISTENCIES

A single criterion S_1 given by n_σ threshold Z not enough to minimise false positive and false negative

$$n_\sigma^{(i)} = \frac{\chi_i^2 - N_{\text{data}}^{(i)}}{\sqrt{2/N_{\text{data}}^{(i)}}}$$

$$\chi^{2,(i)}_{\text{weighted}} = \frac{1}{N_{\text{data}} - N_{\text{data}}^{(j)}} \sum_{j=1, j\neq i}^{N_{\text{exp}}} N_{\text{data}}^{(j)} \chi_j^2 + w^{(i)} \chi_i^2$$

$$w^{(i)} = N_{\text{data}}/N_{\text{data}}^{(i)}$$

Extra criterion: what happens if a large weight is given to inconsistent/consistent dataset?
Expect that the inconsistent dataset, if given extra weight in the fit, will either fail to improve or will improve but spoil $\chi 2$ of other consistent datasets, while the consistent dataset will not.

S_1 $\mu_i > Z$

S_2 $n_\sigma^{\text{weighted},(i)} > Z$

S_3 $n_\sigma^{\text{weighted},(j)} - n_\sigma^{(j)} > Z \qquad \forall j \neq i$

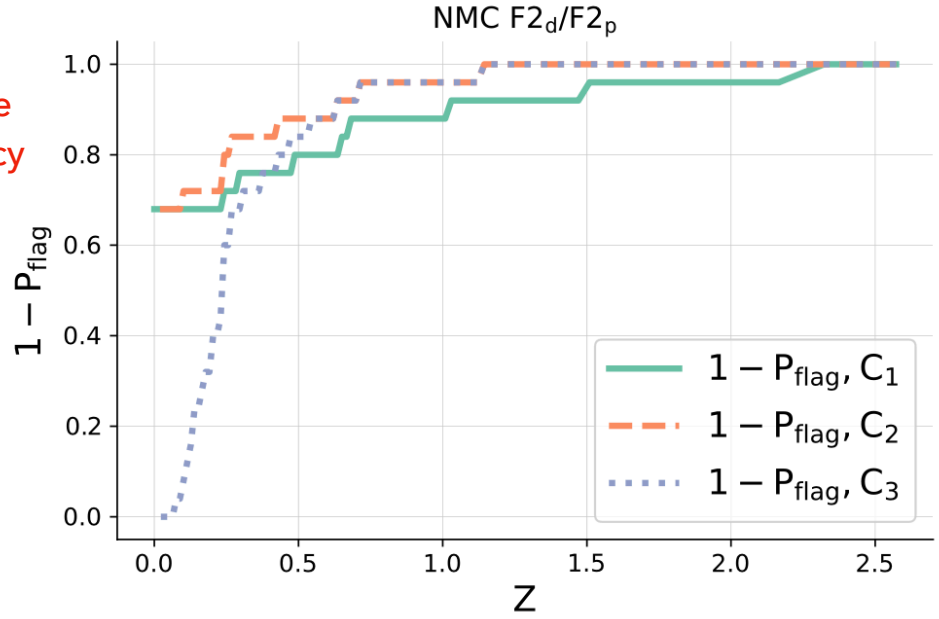
(4) OPTIMAL INCONSISTENCY DETECTION

C_1: condition S_1 satisfied

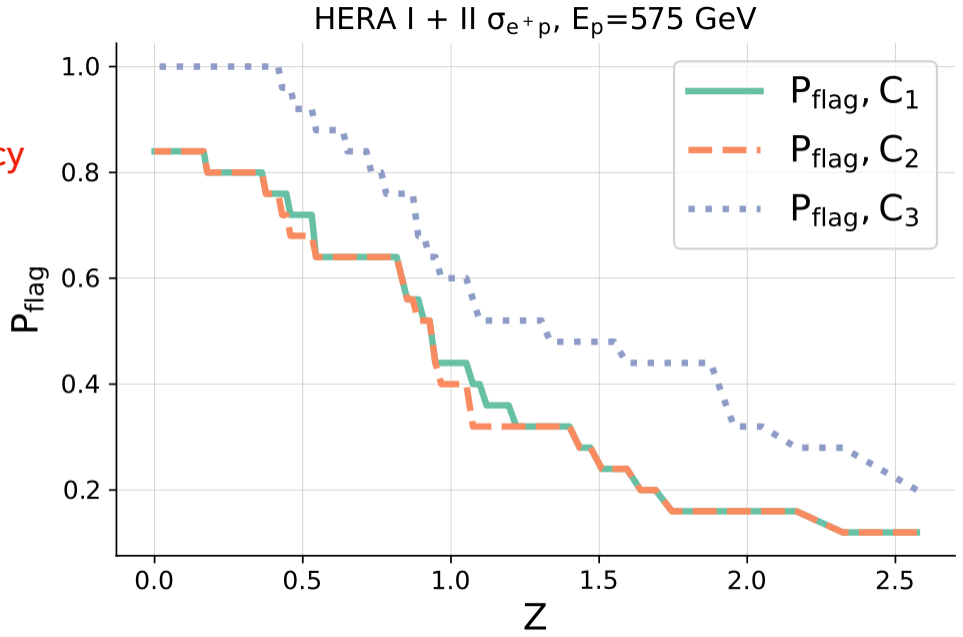
C_2: condition S_1 satisfied, and in a weighted fit either S_2 or S_3 are satisfied (NNPDF4.0 criterion)

C_3: in weighted fit either S_2 or S_3 are satisfied

True negative efficiency



True positive efficiency



S_1 $\mu_i > Z$

S_2 $n_\sigma^{\text{weighted},(i)} > Z$

S_3 $n_\sigma^{\text{weighted},(j)} - n_\sigma^{(j)} > Z \qquad \forall j \neq i$

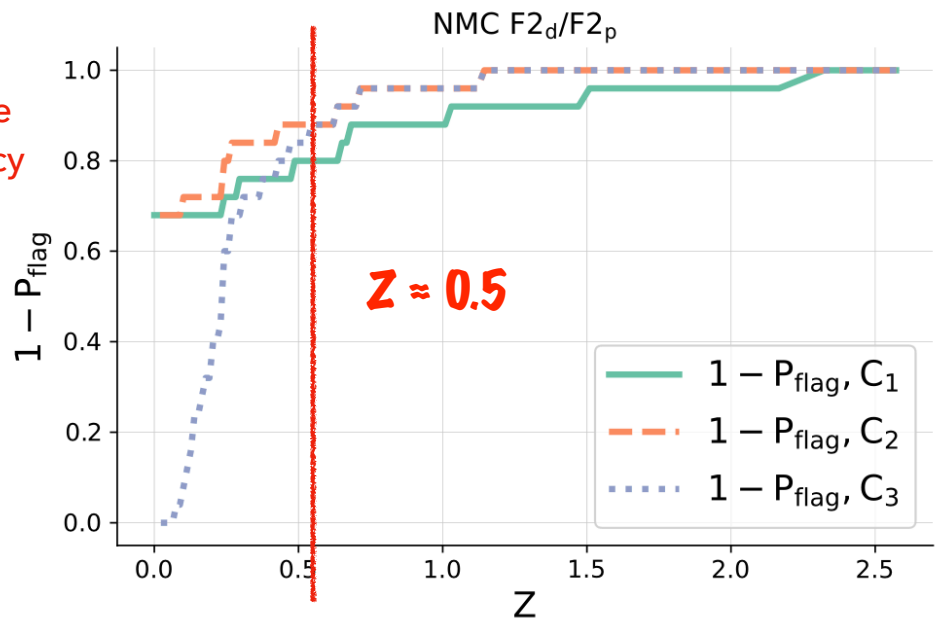
(4) OPTIMAL INCONSISTENCY DETECTION

C_1: condition S_1 satisfied

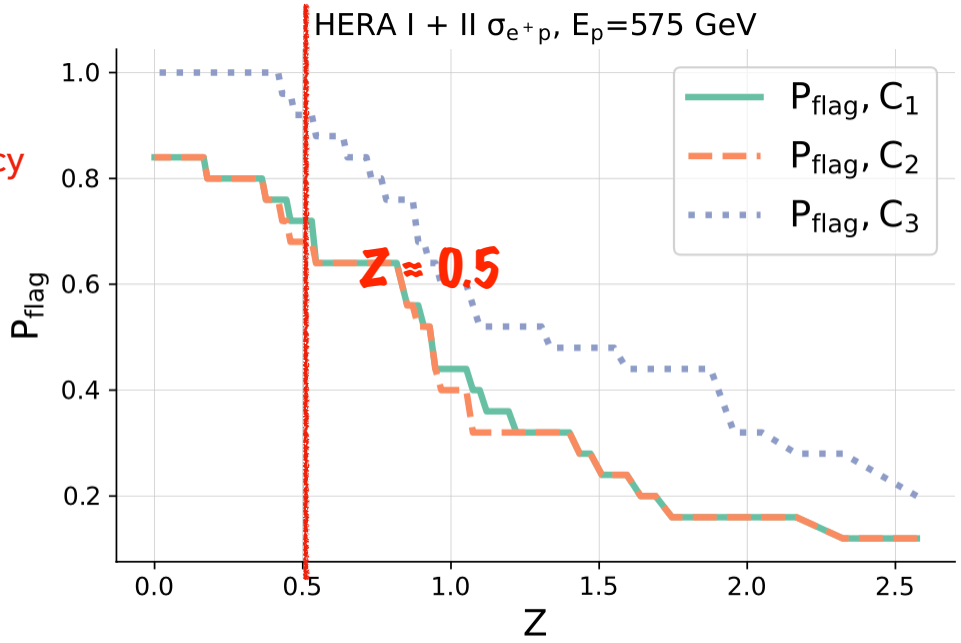
C_2: condition S_1 satisfied, and in a weighted fit either S_2 or S_3 are satisfied (NNPDF4.0 criterion)

C_3: in weighted fit either S_2 or S_3 are satisfied

True negative efficiency



True positive efficiency



In this case for Z ≈ 0.5, 90% probability of NOT flagging a consistent dataset as inconsistent and 95% probability of flagging an inconsistent dataset as inconsistent.

(5) EXTRACTING PARAMETERS FROM DATA

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

$$T^{(\rm ep)}(\{f\},\vec{c}) = \sum_{i=1}^{n_{\rm partons}} \int_x^1 dy\, f_i(y;\vec{c})\, \sigma_i\left(\frac{x}{y};\vec{c}\right) \equiv \sum_{i=1}^{n_{\rm partons}} \left[f_i \otimes \sigma_i\right](x;\vec{c})$$

PDFs

SM parameters (αs(Mz), Mw, θw …), or BSM parameters (SMEFT WCs …)

$$T(\{f\},c) = T^{\rm SM}(\{f\},c=0)(1 + k_{\rm lin}\cdot c + k_{\rm quad}\cdot c^2)$$

✓ In a PDF fit typically

$$T(\{f\}) = T^{\rm SM}(\{f\},c=0)$$

✓ In a fit of SMEFT Wilson Coefficients

$$T(c) = T^{\rm SM}(\{f\}=\{f\}^*)(1 + k_{\rm lin}\cdot c + k_{\rm quad}\cdot c^2)$$

(5) SIMULTANEOUS EXTRACTION OF PARAMETERS AND PDF

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

$$T^{(\rm ep)}(\{f\},\vec{c}) = \sum_{i=1}^{n_{\rm partons}} \int_x^1 dy\, f_i(y;\vec{c})\, \sigma_i\left(\frac{x}{y};\vec{c}\right) \equiv \sum_{i=1}^{n_{\rm partons}} \left[f_i \otimes \sigma_i\right](x;\vec{c})$$

PDFs

SM parameters (αs(Mz), Mw, θw …), or BSM parameters (SMEFT WCs …)

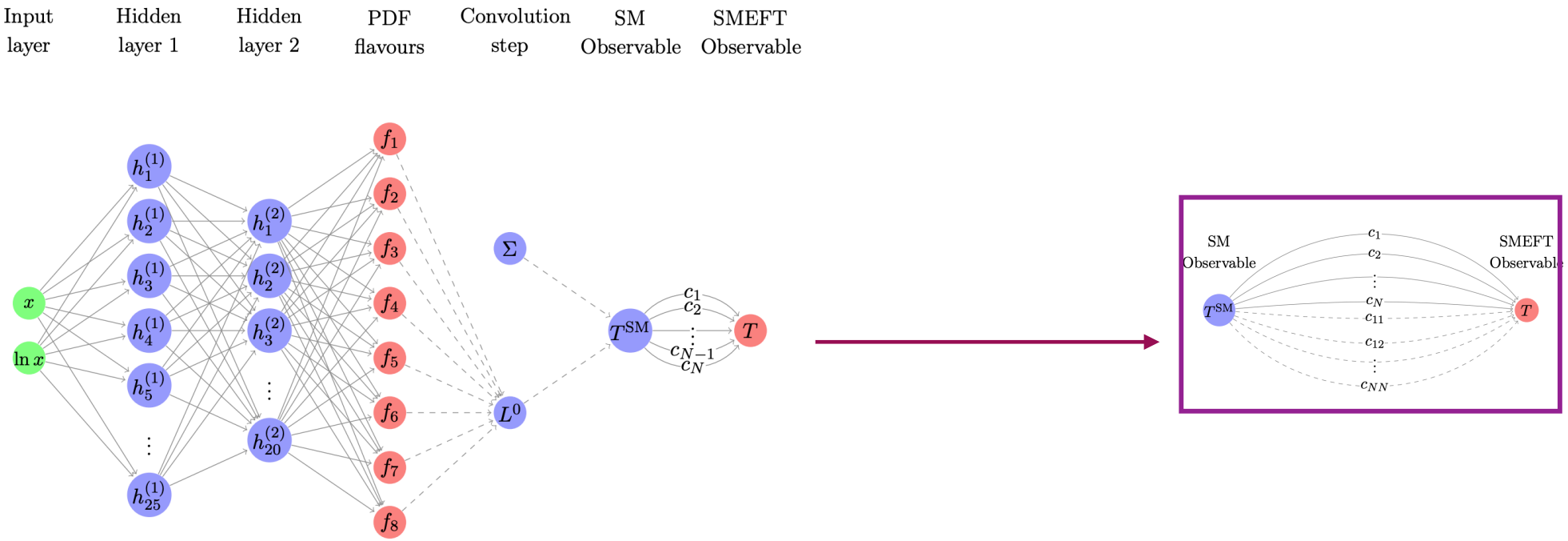
$$T(\{f\},c) = T^{\rm SM}(\{f\},c=0)(1 + k_{\rm lin}\cdot c + k_{\rm quad}\cdot c^2)$$

✓ But PDFs and parameters are correlated and PDFs can absorb shifts in the parameters (absorb new physics) [E. Hammou, et al 2307.10370]: can we do simultaneous fits of PDFs and parameters?

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

(5) THE SIMUNET SOLUTION

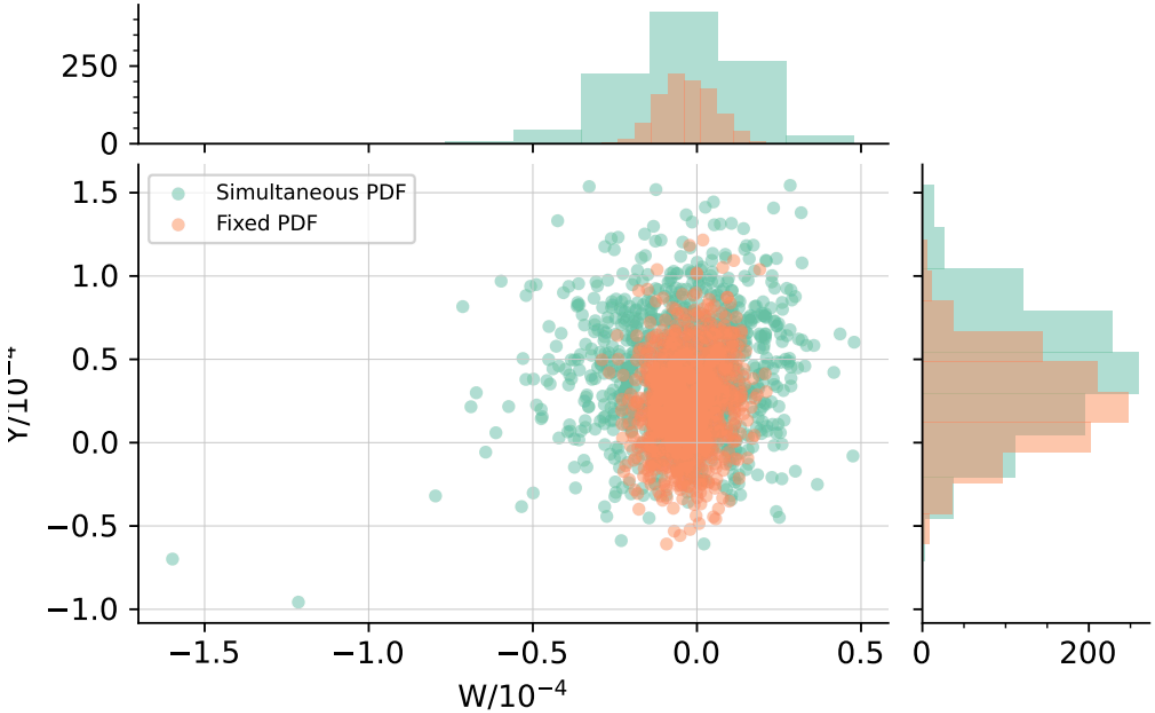
- The idea: take a PDF fit based on NNPDF4.0 methodology and make dependence of observables on physics parameters $\{c_i\}$ explicit before computing the loss function (e.g. adding SMEFT corrections, or expanding observables in terms of SM precision parameters)
- Perform minimisation of loss function over $\hat{\theta} = \theta \bigcup \{c_i\}$ by adding new layer to the deep neural network used in NNPDF4.0



S. Iranipour, MU - arXiv: 2201.07240

(5) EXAMPLE: DRELL-YAN DATA @HL-LHC

S. Iranipour, MU - arXiv: 2201.07240



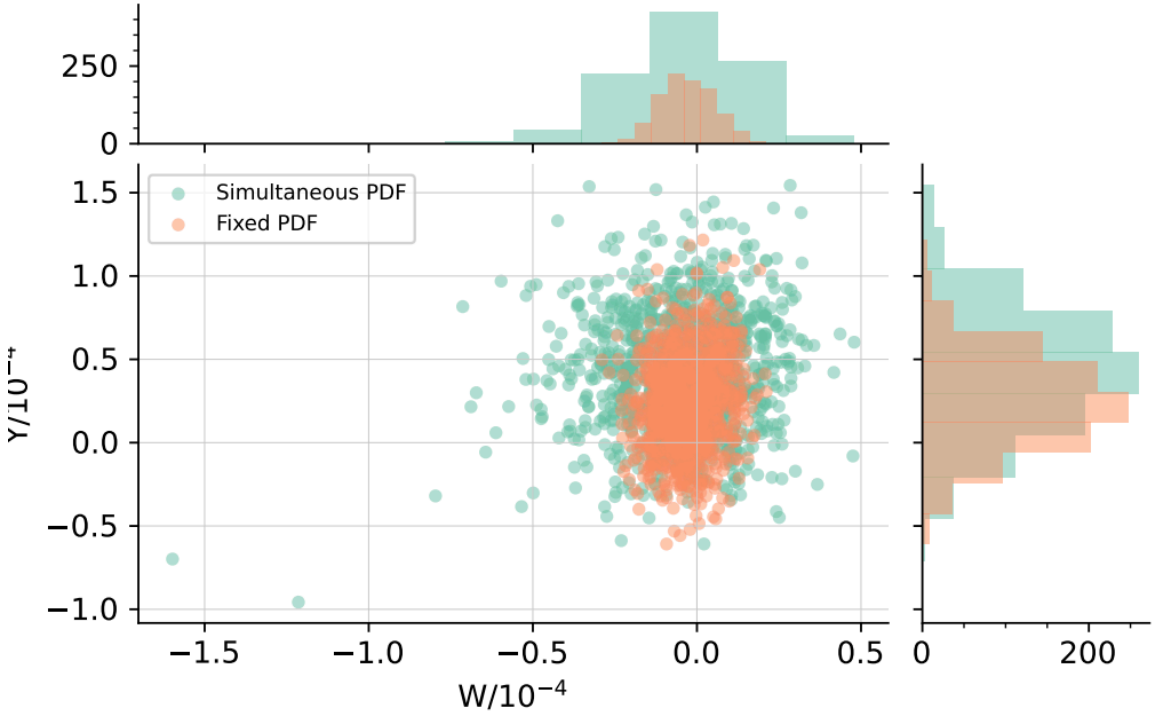
	SM PDFs	SMEFT PDFs
$W \times 10^5$ (68% CL)	$[-1.1, 0.5]$	$[-2.4, 1.5]$
$W \times 10^5$ (95% CL)	$[-2.0, 1.4]$	$[-4.3, 3.4]$
$Y \times 10^5$ (68% CL)	$[-0.4, 5.2]$	$[0.6, 8.0]$
$Y \times 10^5$ (95% CL)	$[-3.2, 8.1]$	$[-3.1, 11.7]$

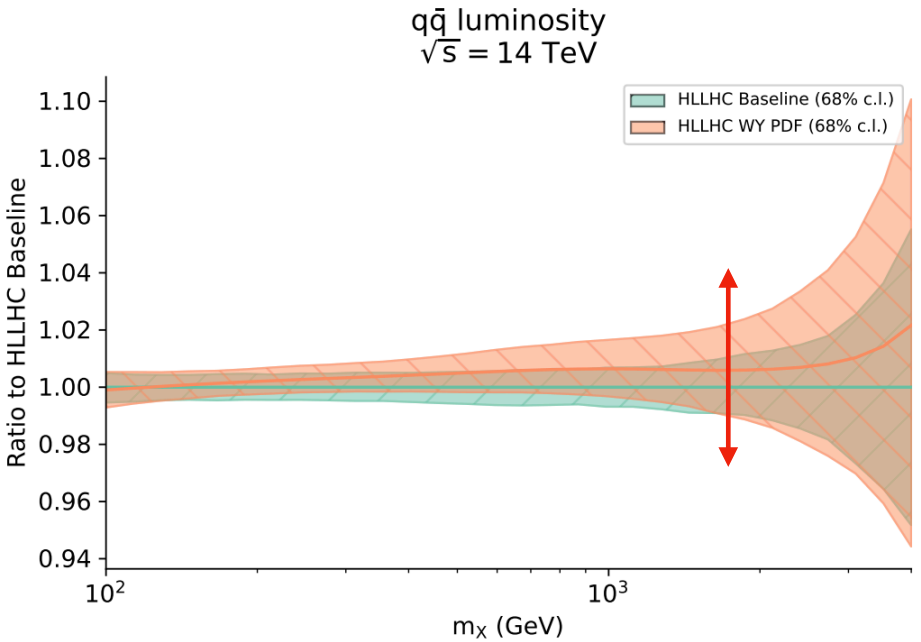
x 2.3 broadening of bounds for W
x 1.3 broadening of bounds for Y

✓ Simultaneous analysis of PDFs and W&Y SMEFT coefficient using simuNET method shows that at HL-LHC the effect of interplay becomes important as WCs bounds broaden and PDF uncertainties change significantly once SMEFT effects allowed in theory predictions entering PDF fit

(5) EXAMPLE: DRELL-YAN DATA @HL-LHC

S. Iranipour, MU - arXiv: 2201.07240

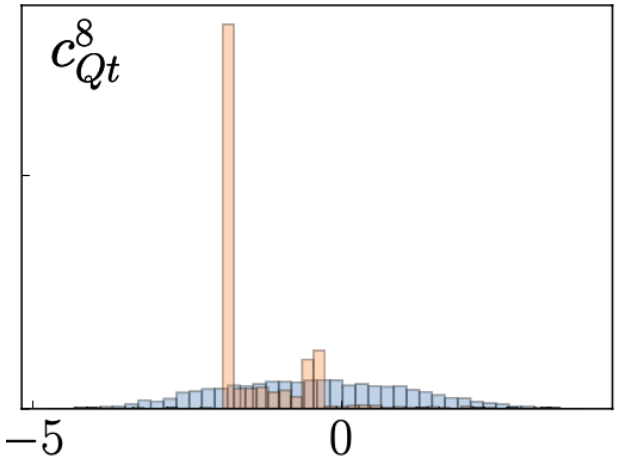


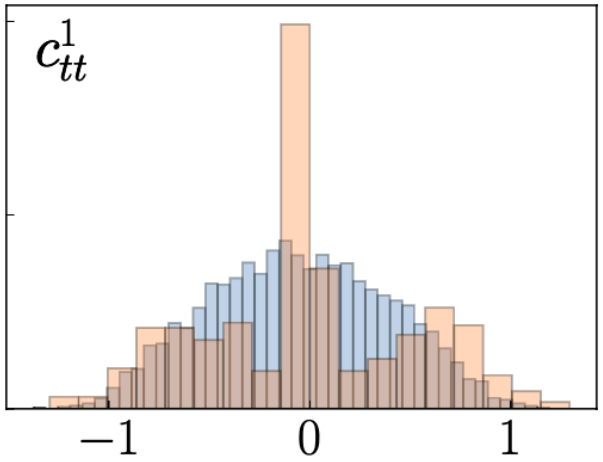


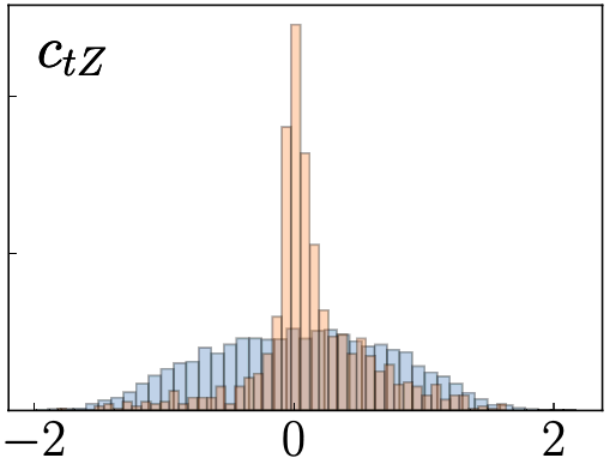
✓ Simultaneous analysis of PDFs and W&Y SMEFT coefficient using simuNET method shows that at HL-LHC the effect of interplay becomes important as WCs bounds broaden and PDF uncertainties change significantly once SMEFT effects allowed in theory predictions entering PDF fit

✓ In this study the dependence on c was **linear**

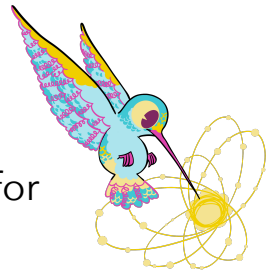
(5) ISSUE OF MONTE CARLO METHOD WITH QUADRATIC DEPENDENCE







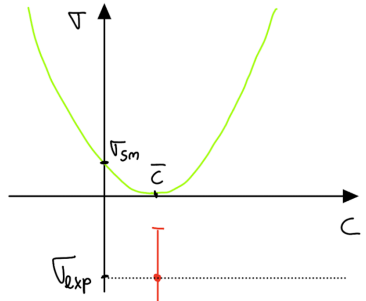
- In the quadratic SMEFT fit observed disagreement between MC method and Bayesian sampling method. Very different posterior (hence different CLs)
- Mathematical analysis of the problem in [Costantini, Madigan, Mantani, Moore arXiv 2404.10056]



- Towards a general Bayesian methodology for simultaneous fits
  [Colibri: Costantini, Mantani, Moore, Sanchez, MU - in progress]

Let's consider a simple scenario: 1 operator, 1 datapoint

$$\chi^2 = \frac{(\sigma(c) - \sigma_{exp})^2}{\delta\sigma^2} \qquad \Delta\chi^2 = \chi^2 - \chi_{min} = 1 \quad \rightarrow \quad [c_-, c_+]$$



CONCLUSIONS AND OUTLOOK

- In an era of precision, need careful assessment of PDF uncertainties and rigorous claims about faithfulness of PDF uncertainties. Key is to strive for deeper understanding.

- Some random pressing questions from me:

  - In Bayesian terms, do we really have control of the prior probability that we assume in our PDF fits and on how the outcome depends on it?
  - Can we have a truly Bayesian determination of PDFs?
  - And a Bayesian determination of PDFs and SM/BSM parameters?
  - What can be used as a complete basis of the PDF space? Dimensionality is a problem for sampling?
  - Hyper-parameter determination using K-loss: should one use a single configuration or several?
  - Closure tests can be used as a controlled test set where to test faithfulness of PDF uncertainties: is the analysis of the normalised bias enough? Is the inconsistency modelled in a sound way? How to determine a set of maximally consistent datasets?
  - At LHC we have huge amount of data and, while lots of progress has been done in determining individual quantities from these (PDFs, alphaS, …) we should strive for simultaneous fits, thus including the correlations between parameters and PDFs. How can this be done in a clever and robust way?