KRAKOW SCHOOL OF INTERDYSCIPLINARY PHD STUDIES

nCTEQ
nuclear parton distribution functions

Markov Chain Monte Carlo Methods For Determination of Nuclear PDFs

Nasim Derakhshanian

IFJ PAN
Department of Theoretical Particle Physics

Peresenting at IFJ PAN young researcher seminars, Krakow, March 27 2025

Markov Chain Monte Carlo Methods For Determination of Nuclear PDFs

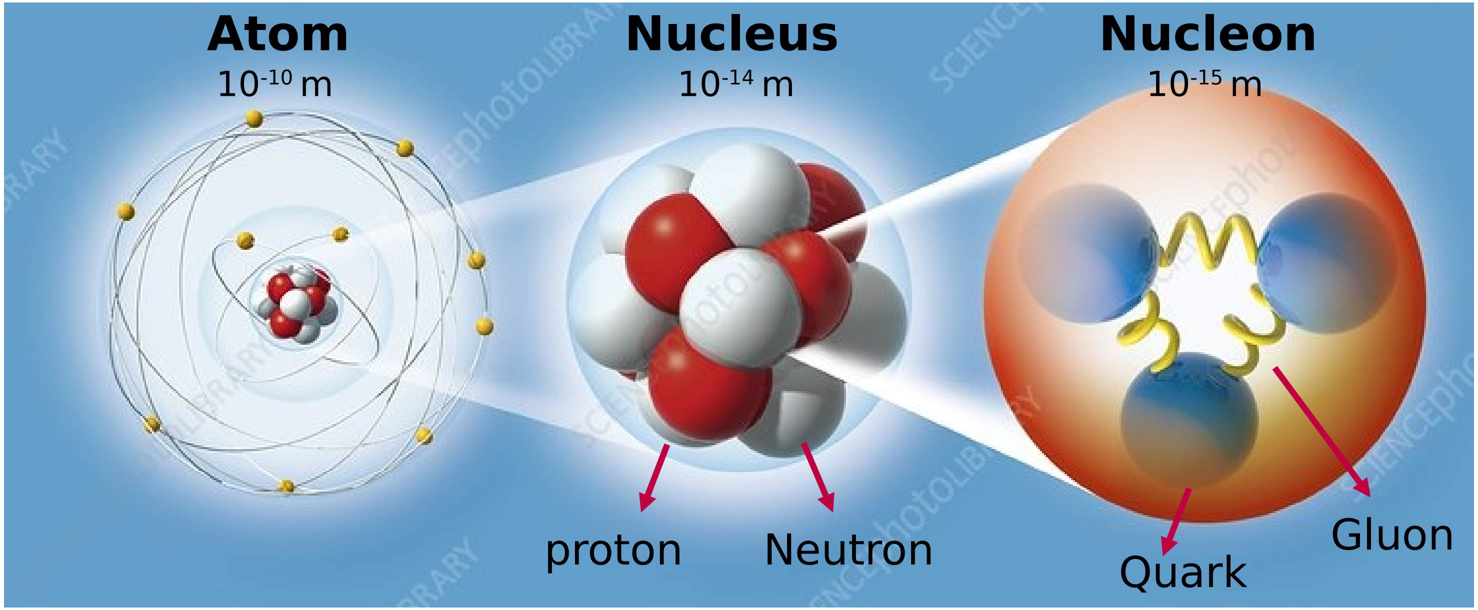
Nasim Derakhshanian

IFJ PAN
Department of Theoretical Particle Physics

Peresenting at IFJ PAN young researcher seminars, Krakow, March 27 2025

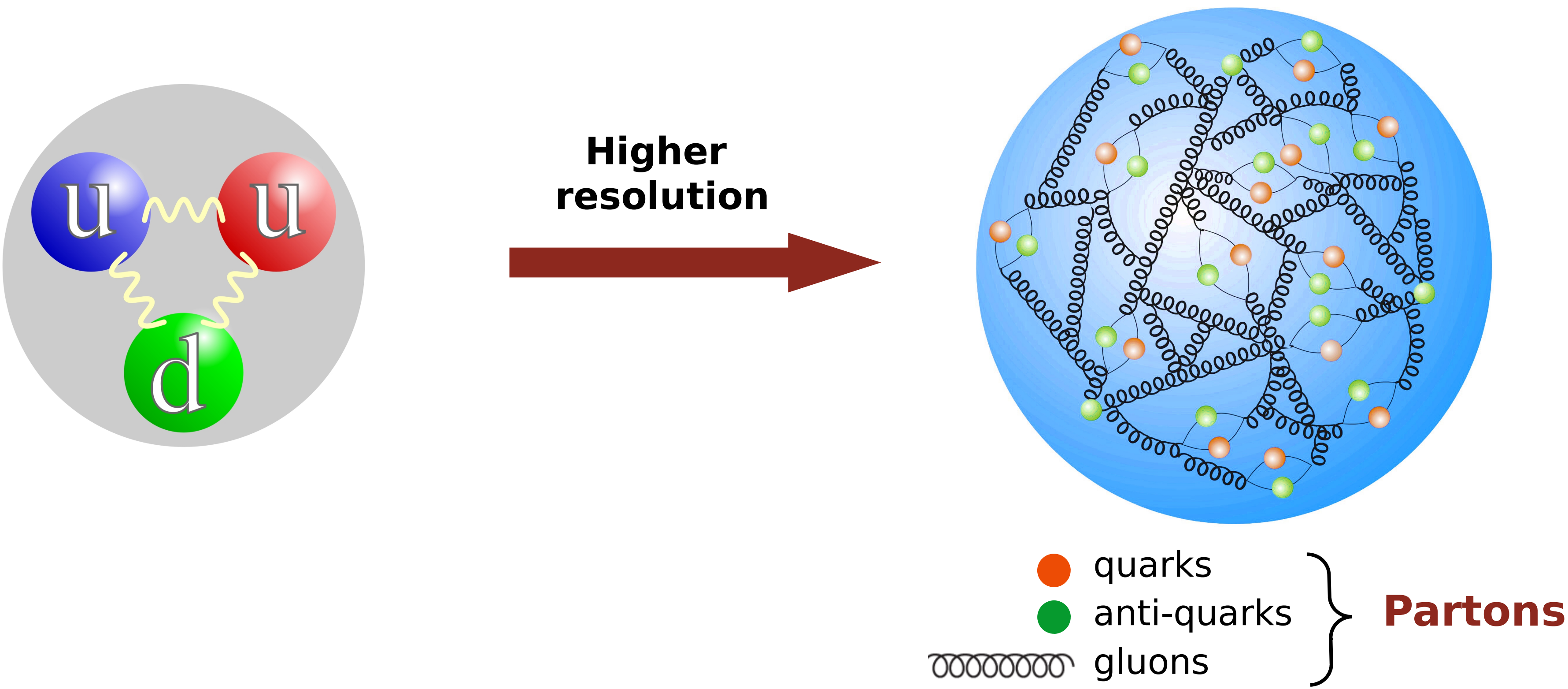
Structure of Matter:

- Structure of matter depends on the resolution scale at which it is observed!



Structure of Matter:

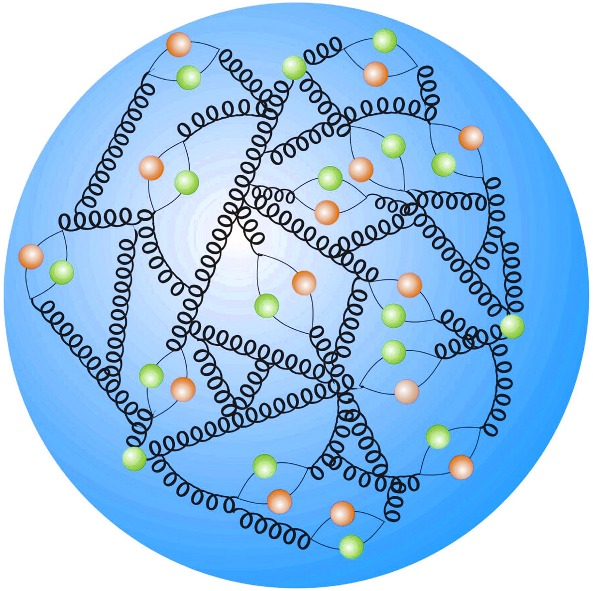
- Structure of matter depends on the resolution scale at which it is observed!



Structure of Matter:

- Structure of matter depends on the resolution scale at which it is observed!

The complex behavior of partons, including their momentum distributions, is governed by the strong interaction dynamics described by **Quantum Chromodynamics (QCD)** theory



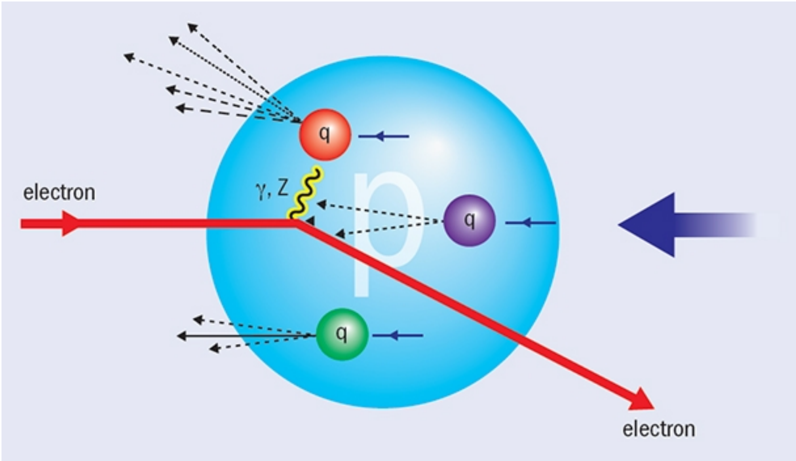
Parton Distribution Function (PDF):

The probability $\mathbf{f_{a/p}(x,Q)}$ that a parton **a** carries fraction **x** of the proton's momentum

Q: energy scale

x: momentum fraction

DIS process: e + p → e + X



QCD Factorization in case of DIS:

$$\frac{d^2\sigma}{dx dQ^2} = \sum_{i=q,\bar{q},g} \int_x^1 \frac{dz}{z} f_i(z,\mu) d\hat{\sigma}_{il \to l'X}\left(\frac{x}{z}, \frac{Q}{\mu}\right)$$

PDFs

Partonic scattering cross-section

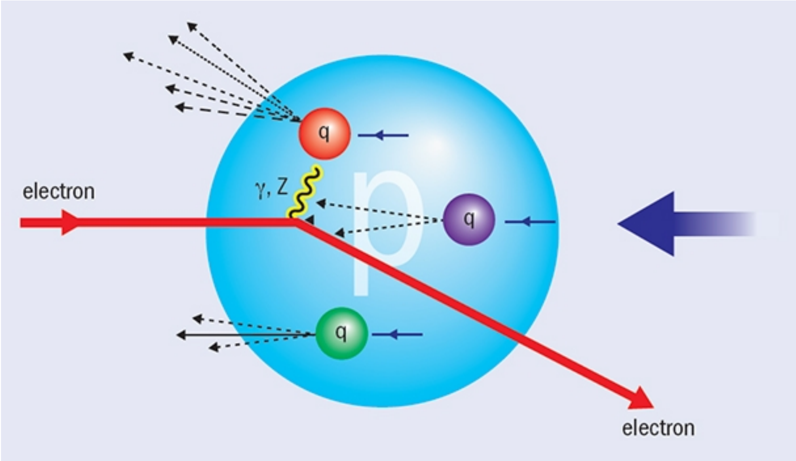
Parton Distribution Function (PDF):

The probability $\mathbf{f_{a/p}(x,Q)}$ that a parton **a** carries fraction **x** of the proton's momentum

Q: energy scale

x: momentum fraction

DIS process: e + p ⟶ e + X



QCD Factorization in case of DIS:

$$\frac{d^2\sigma}{dx dQ^2} = \sum_{i=q,\bar{q},g} \int_x^1 \frac{dz}{z} f_i(z,\mu) d\hat{\sigma}_{il\to l'X}\left(\frac{x}{z}, \frac{Q}{\mu}\right)$$

PDFs — $f_i(z,\mu)$

Partonic scattering cross-section — $d\hat{\sigma}_{il\to l'X}\left(\frac{x}{z}, \frac{Q}{\mu}\right)$

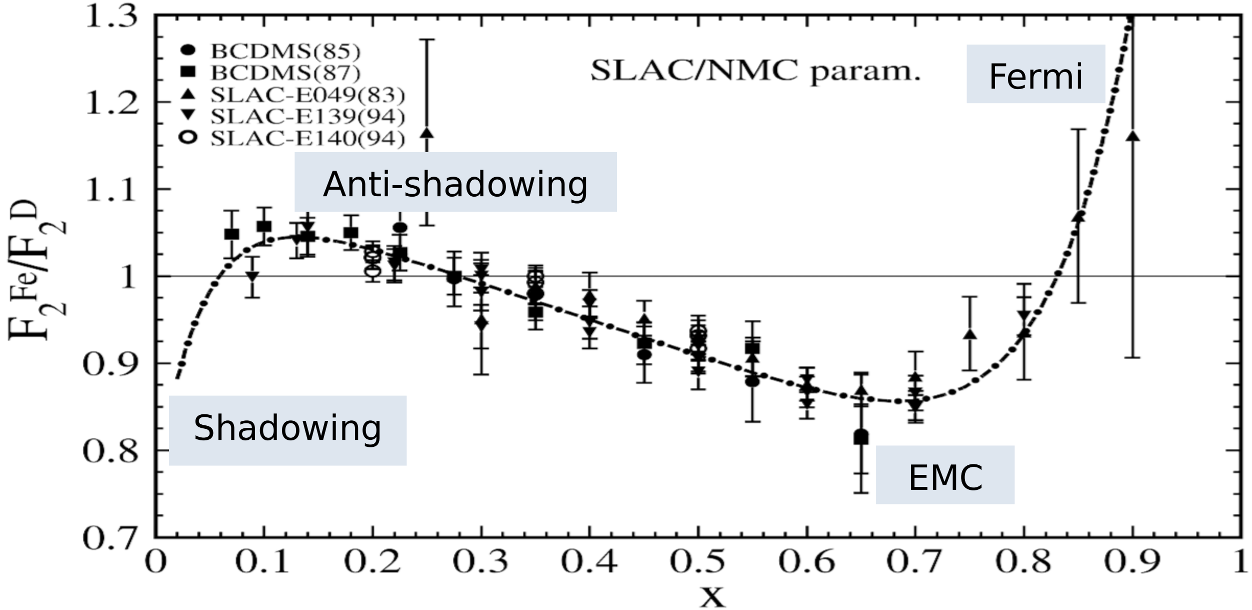
PDF properties:

- Universal (independent of the process)
- Q-dependence governed by DGLAP evolution equations
- **Non-perturbative**: x-dependence of PDF is NOT calculable in pQCD

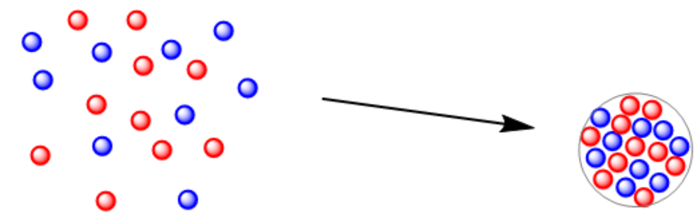
Nuclear PDFs (nPDFs):

nPDF describes the momentum distribution of partons (quarks and gluons) inside a nucleus

$$F_2^A(x) \neq Z F_2^p(x) + N F_2^n(x)$$



Schienbein, et al., arXiv: 0907.2357



Nuclear correction ratio:

$$R_A(x) \equiv \frac{F_2^A(x)}{F_2^D(x)}$$

▶ we can incorporate these modifications into **universal nuclear PDFs** under certain theoretical assumptions and kinematic conditions.

Nuclear PDFs (nPDFs):

nPDF describes the momentum distribution of partons (quarks and gluons) inside a nucleus

Where are nPDFs useful?

- **High-Energy Collider Physics (LHC & RHIC)**
  essential for predicting the outcomes of collisions involving nuclear targets
- **Neutrino Physics**
  Nuclei are used as targets in neutrino scattering experiments to increase the interaction probability
- **Nuclear Structure**
  provide a deeper insights into our understanding of nuclear matter.

Global Analysis of nPDF

- Q dependence is governed by PQCD (DGLAP evolution equations)
- **x dependence** of PDF is **NOT** calculable in pQCD

Global PDF fits:

$$\text{Data} = \text{PDF} \otimes \sigma_H$$

measure — measured cross section from diverse experiments

fit — Methodology

nCTEQ, EPPS, NNPDF, ect

↓

Standard library (**LHAPDF**)

predict — Precision Calculations: NLO & NNLO

Harland-Lang DIS2024

Global Analysis of nPDF

Input function at Q_0

Parameterize nuclear PDF at initial scale: 1.3GeV

DGLAP evolution

Compute theory predictions at scale Q by solving DGLAP equations

Construct χ^2 function

Calculate the goodness of fit in terms of theory predictions, data and uncertainties

Minimization

Minimize χ^2 function with respect to nPDF parameters

Experimental data

Choose experimental data (e.g. DIS, DY, W/Z, etc.) and apply kinematical cuts

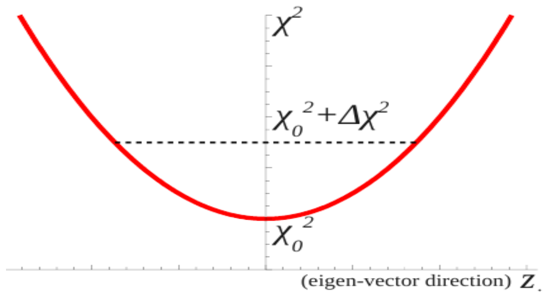
$$\chi^2(\{a_i\}) = \sum_j^N \left(\frac{(\mathrm{data}_j - \mathrm{theory}_j(\{a_i\}))}{\sigma_j} \right)^2$$

Uncertainties estimation

nPDF uncertainties estimation

The **Hessian** method is widely used for error estimation in both proton and nuclear PDFs.

It relies on the quadratic behavior of the χ^2 function near the minimum.



Shortcomings:

- Non-gaussian errors
- Global minima judgment
- Choice of χ^2 tolerance

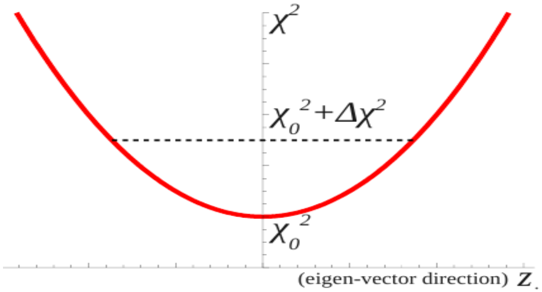
nPDF difficulties:

- Lacking data (range and precision of data for nuclei are generally lower than for proton)
- Complexity and nature of nuclear effects

nPDF uncertainties estimation

The **Hessian** method is widely used for error estimation in both proton and nuclear PDFs.

It relies on the quadratic behavior of the χ^2 function near the minimum.



Shortcomings:

- Non-gaussian errors
- Global minima judgment
- Choice of χ^2 tolerance

nPDF difficulties:

- Lacking data (range and precision of data for nuclei are generally lower than for proton)
- Complexity and nature of nuclear effects

Markov Chain Monte Carlo method

advanced statistical method as an alternative for Hessian

Global Analysis of nPDF

Input function at Q_0

Parameterize nuclear PDF at initial scale: 1.3GeV

DGLAP evolution

Compute theory predictions at scale Q by solving DGLAP equations

Construct χ^2 function

Calculate the goodness of fit in terms of theory predictions, data and uncertainties

Experimental data

Choose experimental data (e.g. DIS, DY, W/Z, etc.) and apply kinematical cuts

$$\sum_j \left(\frac{(\text{data}_j - \text{theory}_j(\{a_i\}))}{\sigma_j} \right)^2$$

MCMC method

Minimization

Minimize χ^2 function with respect to nPDF parameters

Uncertainties estimation

Markov Chain Monte Carlo (MCMC)

A sequence of random variables where the current value is dependent on the value of the prior variable (Memory-less property)

A technique for randomly sampling a probability distribution and approximating a desired quantity.

Bayes theorem:

Posterior, Likelihood, Prior

$$p(\theta \mid \text{data}) = \frac{p(\text{data} \mid \theta) \cdot p(\theta)}{p(\text{data})}$$

Normalization

Prior: initial belief about the parameter before considering the data.
Likelihood: probability of observing the data given a specific value of the parameter.
Posterior: updated belief about the parameter given the data.

- We aim to find the set of nPDF parameters that maximizes the posterior probability distribution given the experimental data.

Likelihood: $p(data|\theta) \propto \exp\left(-\frac{\chi^2}{2}\right)$

$$\chi^2(\{a_i\}) = \sum_j^N \left(\frac{(\mathrm{data}_j - \mathrm{theory}_j(\{a_i\}))}{\sigma_j} \right)^2$$

Statistical error
Correlated and uncorrelated
systematic errors

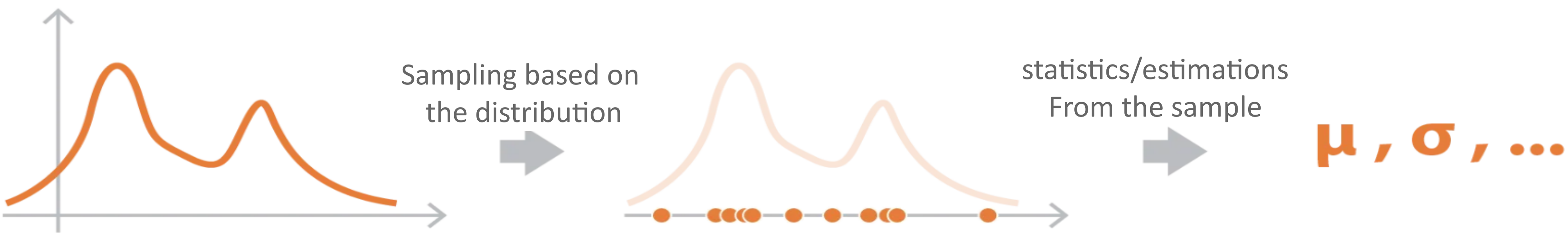
➢ We aim to find the set of nPDF parameters that maximizes the posterior probability distribution given the experimental data.

Likelihood: $p(data|\theta) \propto \exp\left(-\frac{\chi^2}{2}\right)$

$$\chi^2(\{a_i\}) = \sum_j^N \left(\frac{(\text{data}_j - \text{theory}_j(\{a_i\}))}{\sigma_j}\right)^2$$

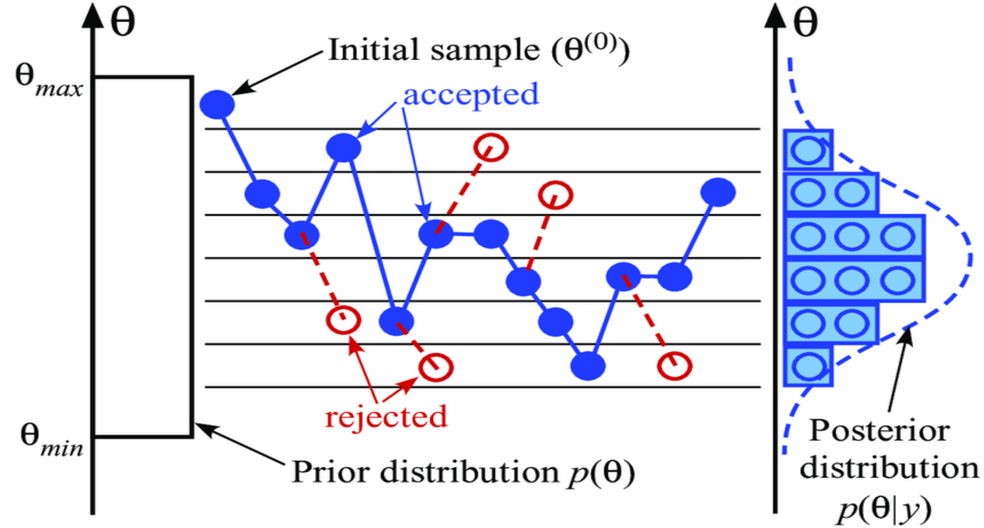
Statistical error
Correlated and uncorrelated systematic errors

Bayesian inference → **MCMC algorithms**



Posterior distribution **samples**

Metropolis algorithm:



Initialize parameters

for i=1 to i=N:

multiplicity =1

Proposing new parameters $\theta^* \sim q(\theta^*|\theta)$

Compute acceptance probability

$$\alpha = \min(p(\theta^*|D)/p(\theta|D), 1)$$

Sample from uniform distribution $u \sim \mathbf{U}(0,1)$

If $u < \min(1,\alpha)$ then $\theta_{i+1} = \theta^*$

Else $\theta_{i+1} = \theta$ (multiplicity +=1)

- **Multiplicity**: the number of consecutive rejections of proposed points before an acceptance occurs.
- Each point in the chain represents a vector of the posterior parameter values.

nPDF fit setup

Fit properties:

- fit **NLO** QCD predictions
- Kinematic cuts: Q > 2GeV, W > 3.5GeV, p_T > 3.0 GeV
- NC & CC DIS, W/Z boson and Heavy Quark →
  - **Pb PDF fit**
  - Multiple nuclei PDF fit
- 10 free parameters: 2 gluon, 6 valence, 2 sea
- Parameterization:

CJ15

Functional form for bound protons at Q_0: $$x f_i^{p/A}(x, Q_0) = c_0 x^{c_1} (1-x)^{c_2} (1 + c_3 \sqrt{x} + c_4 x)$$

$$f_i^{(A,Z)} = \frac{Z}{A} f_i^{p/A} + \frac{A-Z}{A} f_i^{n/A}$$

Atomic number dependence: $$c_k \to p_k + a_k \ln(A) + b_k \ln^2(A).$$

Accardi et al., arXiv:1602.03154

MCMC setup:

Adaptive MH algorithm setup:

- The algorithm starts with a normal random-walk MH phase until N_0 samples have been generated

  Proposal distribution: Multivariate Gaussian with fixed covariance C_0

  $$\mathbf{X}_{i+1} = \mathcal{N}(\mathbf{X}_i, C_0)$$

- Then it switches to a self-learning proposal distribution

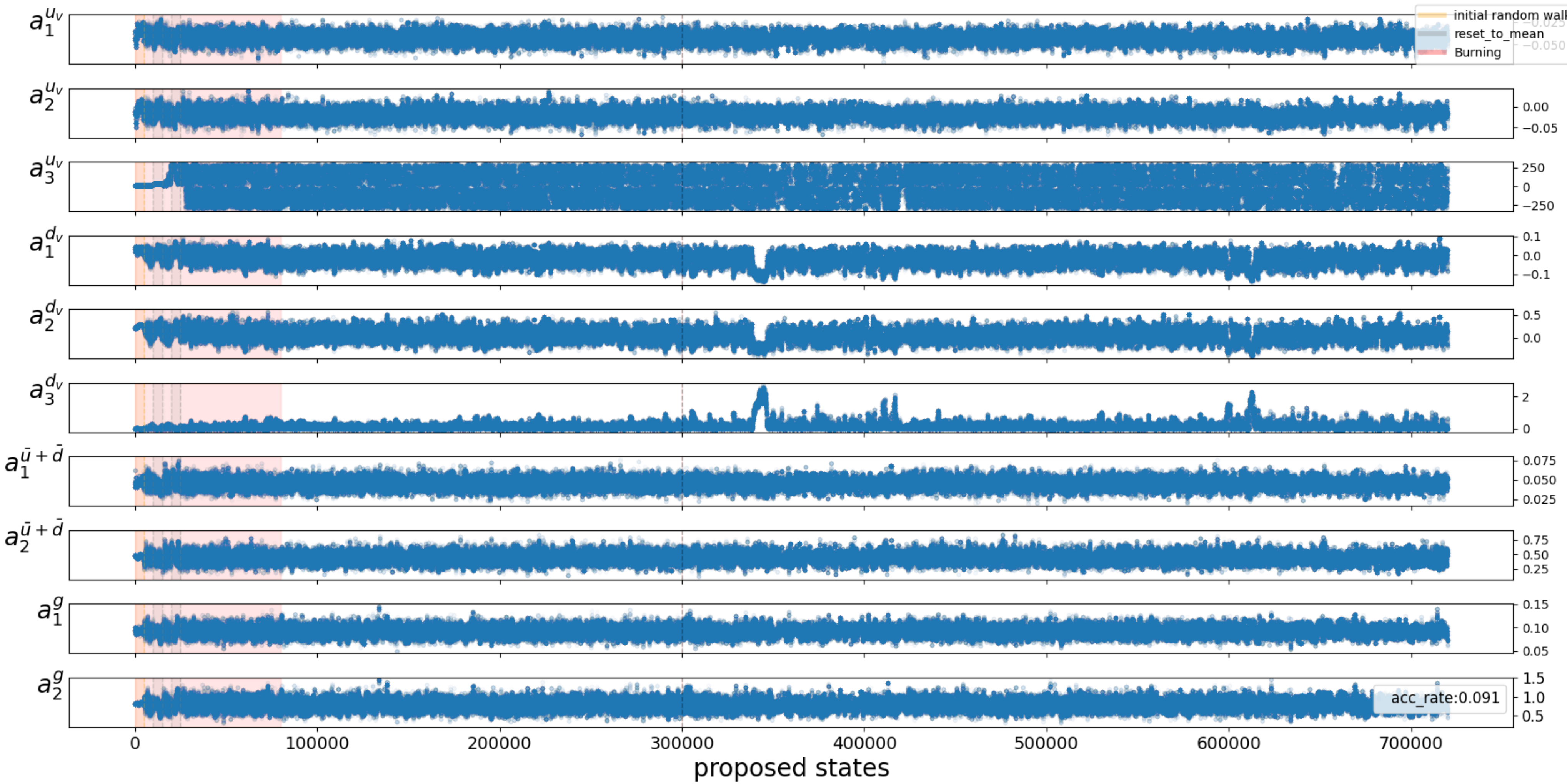
  Adaptive proposal distribution: Multivariate Gaussian with self learned covariance C_i (covariance from collected samples so far)

  $$\mathbf{X}_{i+1} = (1-\beta)\mathcal{N}(\mathbf{X}_i, \frac{(2.4)^2}{d}.C_i) + \beta\mathcal{N}(\mathbf{X}_i, C_0)$$

- To boost the convergence, the algorithm restarts from its current mean value*

*The fixed covariance matrix is first given by a fraction of initial parameter values and then after restarting, it adjusts to the fraction of diagonal elements in the current self-learned covariance C_i

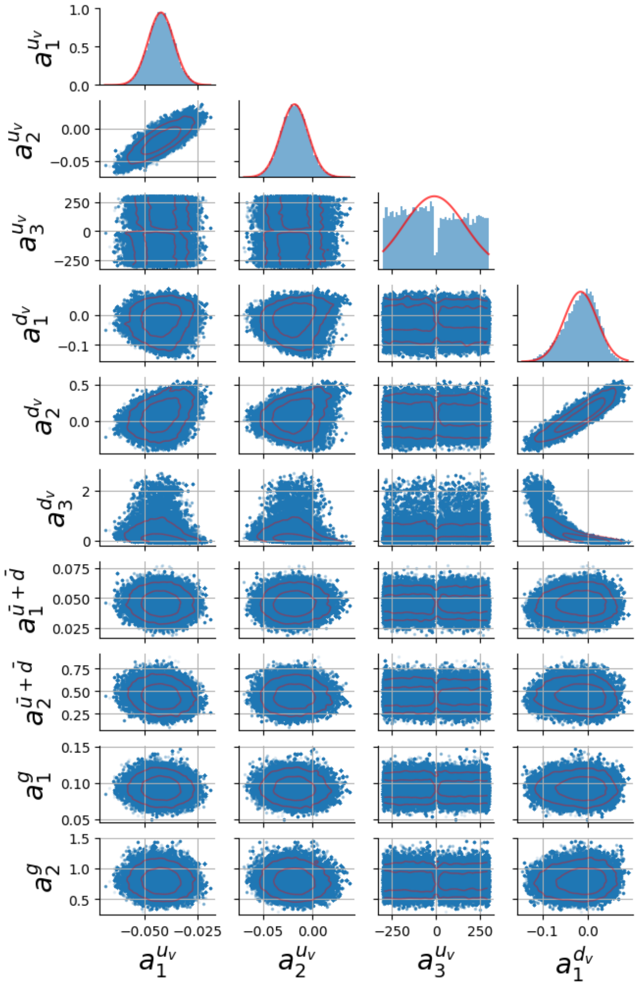
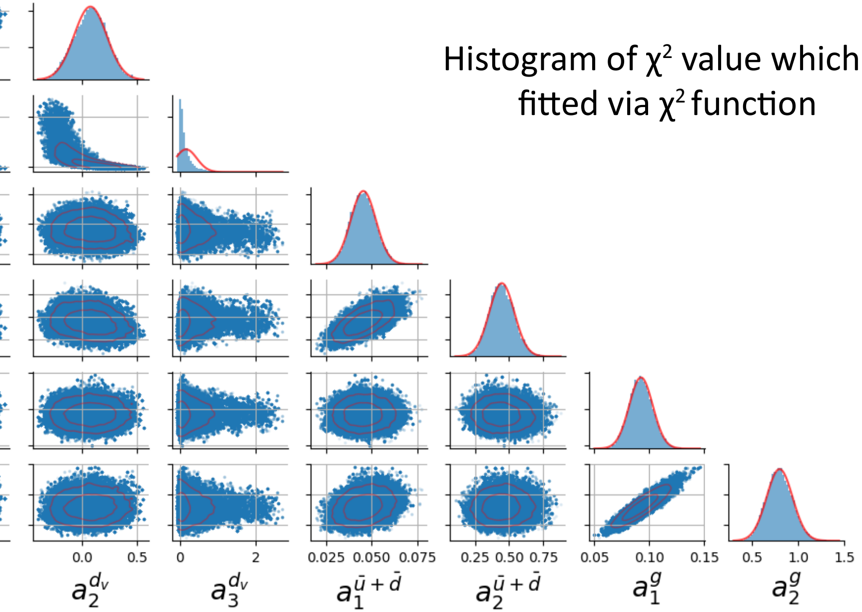
Markov chain generated for Pb PDF parameters (W/Z and Heavy Quark and ν-DIS(chorus); 1448 data)

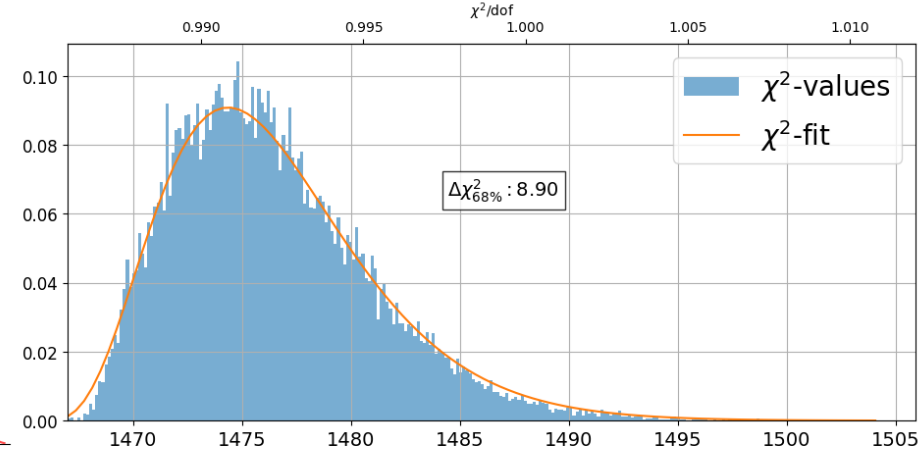


Pairwise plot

diagonal: histogram of each parameter
off-diagonal: 2D correlation plots between parameters

MCMC can reveal non-Gaussian features of the underlying distribution



Histogram of χ² value which is fitted via χ² function

Error estimation:

Autocorrelation function (ACF): $\rho(k) = \frac{\mathrm{Cov}(k)}{\mathrm{Cov}(0)}$ $\qquad \mathrm{Cov}(k) = \frac{1}{n}\sum_{t=1}^{n-k}(x_{t+k} - \bar{x})(x_t - \bar{x}),$

measures the correlation between samples separated by a certain lag k

Integrated autocorrelation time: $\tau \approx \frac{1}{2} + \sum_{t=1}^{\infty} \rho(k)$ → **Gamma-method** Estimating by analyzing the sum of autocorrelation up to a certain lag W_{opt}

measures how many steps it takes for the samples in the chain to become effectively independent

Monte Carlo error estimation (uncorrelated)

$$\sigma^2_{MC} = \frac{1}{n-1}\sum_{t=1}^{n}(X_t - \hat{\mu})^2$$

MCMC error estimation (correlated)

$$\sigma^2_{MCMC} = 2\,\tau_{int}\,\sigma^2_{MC}$$

Thinning method:

keep only every k-th sample in the Markov chain and discard the rest

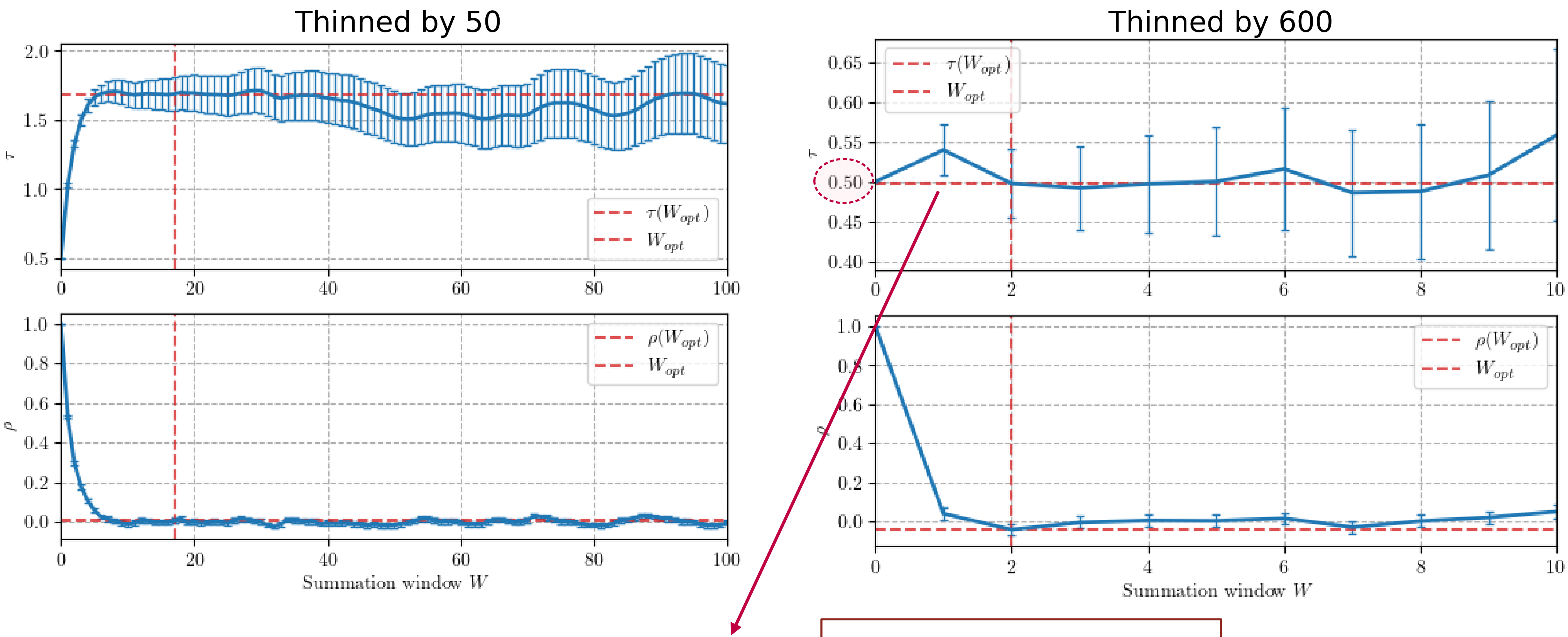
Why Thinning?

- It provides an **uncorrelated** chain so we can use Monte-Carlo error estimation:

$$\sigma^2_{MCMC} = 2\,\tau_{int}\,\sigma^2_{MC} \quad \Rightarrow \quad \sigma^2_{MC} = \frac{1}{n-1}\sum_{t=1}^{n}(X_t - \hat{\mu})^2$$

- We aim to generate a set of PDF grids corresponding chain’s units. Thinning the chain makes it more applicable.

ACF and integrated autocorrelation time: $\tau \approx \frac{1}{2} + \sum_{t=1}^{\infty} \rho(t)$



For uncorrelated samples: $\tau_{\text{optimal}} = 0.5 \; (\rho = 0)$

Methodology:

- **Generating Multiple Chains**
  Each chain starts with random values from the Hessian fit results. Use different random seeds

- **Removing Burn-In Phase**
  Discard the initial segment of each chain, known as the burn-in or thermalization phase, which represents the period before the chain converges to the target distribution

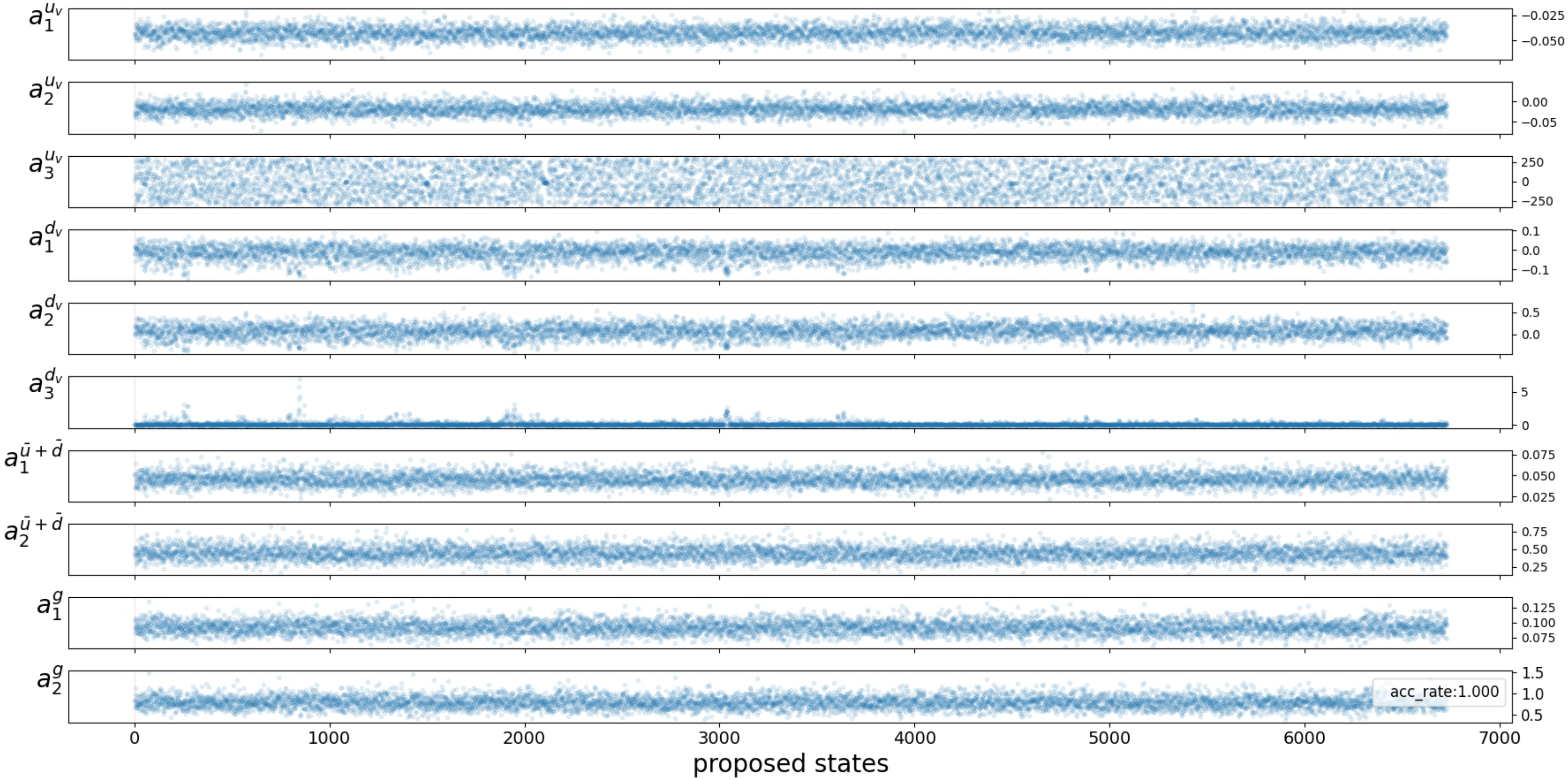
- **Thinning Each Chain**
  Apply thinning to each chain to reduce the autocorrelation, aiming to retain only uncorrelated samples

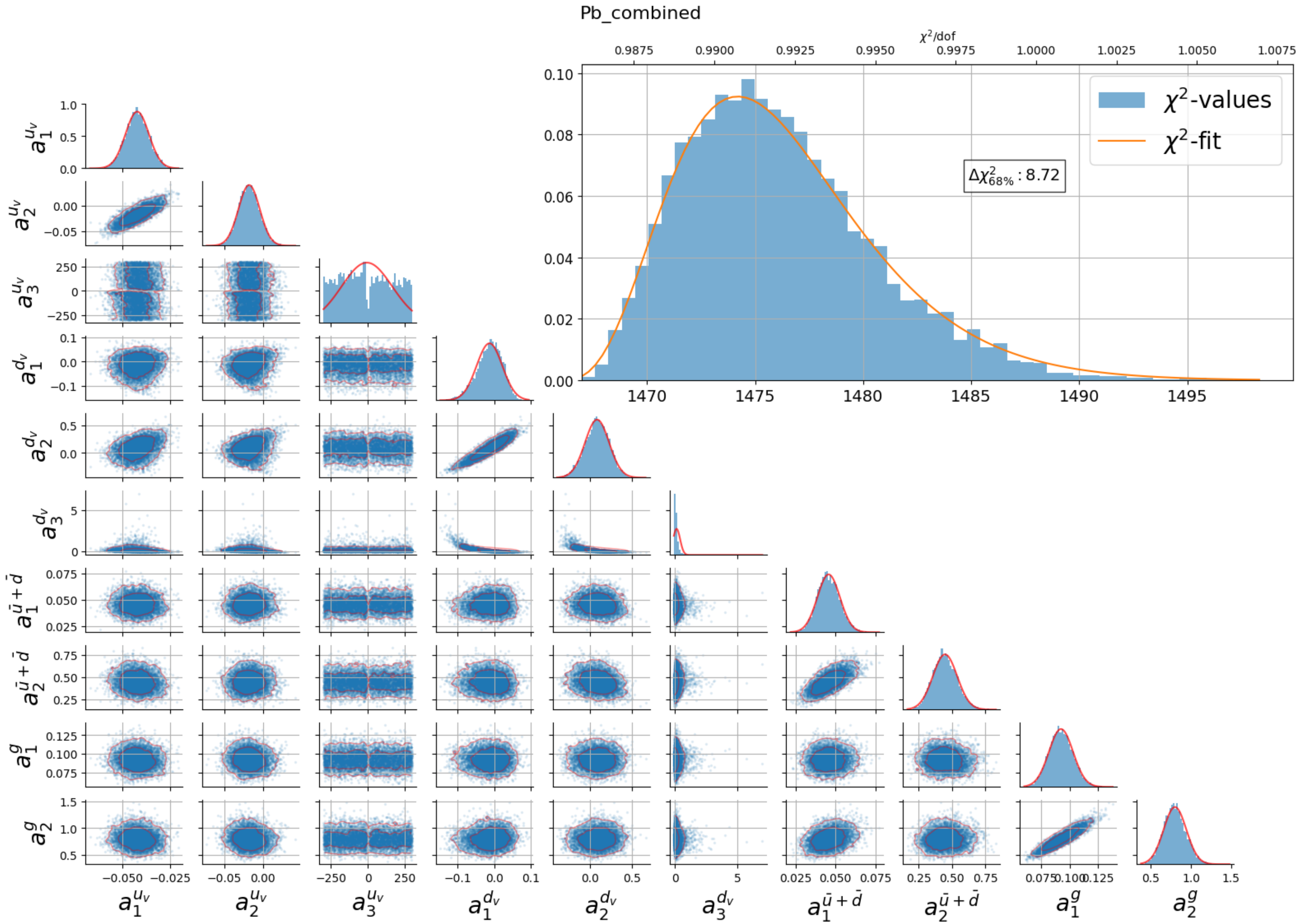
- **Combining Uncorrelated Samples**
  Merge all the thinned, uncorrelated samples from the different chains into a single chain

- **Estimating Parameters and Uncertainties**
  Use the combined set of uncorrelated samples to estimate the values of nPDF parameters and their uncertainties.

- **generating an LHAPDF set**
  Construct nPDF corresponding to each unit of the combined chain and perform error estimation in the level of nPDF (Saving them in the standard LHAPDF format so that anyone can use such nPDFs)

Final Chain (combined):



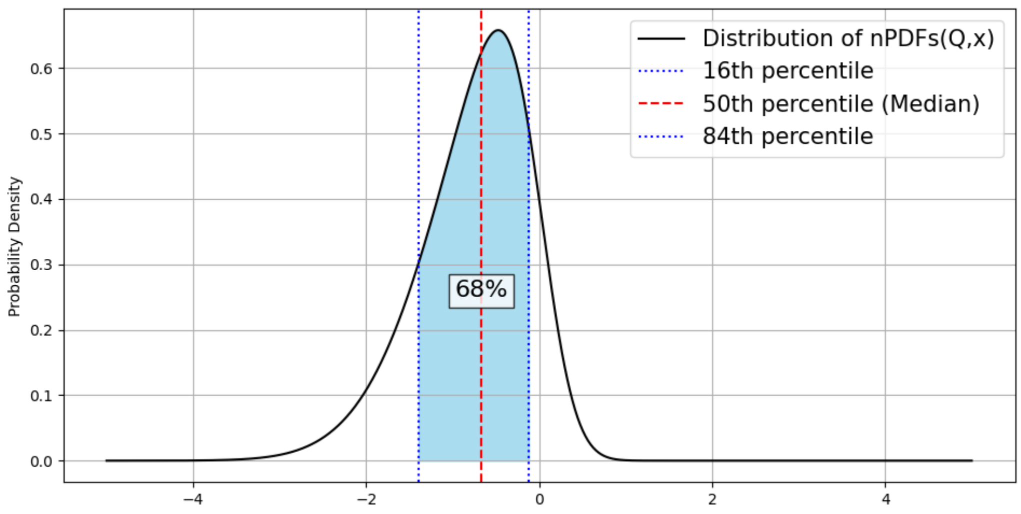


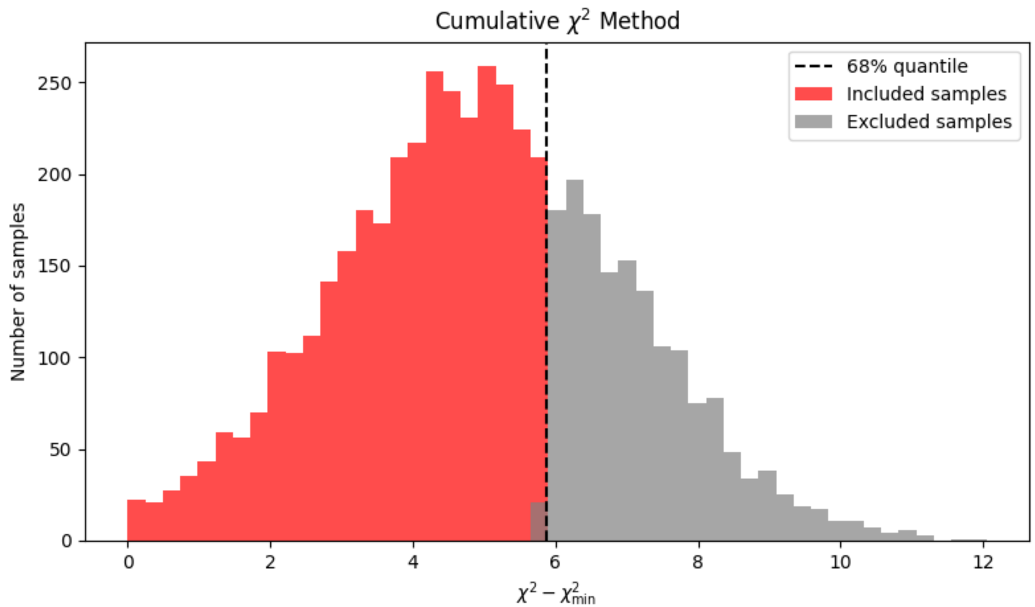
Pb_combined
χ^2/dof
χ^2-values
χ^2-fit
$\Delta\chi^2_{68\%}$: 8.72
$a_1^{u_v}$
$a_2^{u_v}$
$a_3^{u_v}$
$a_1^{d_v}$
$a_2^{d_v}$
$a_3^{d_v}$
$a_1^{\bar{u}+\bar{d}}$
$a_2^{\bar{u}+\bar{d}}$
a_1^{g}
a_2^{g}

nPDFs uncertainties:

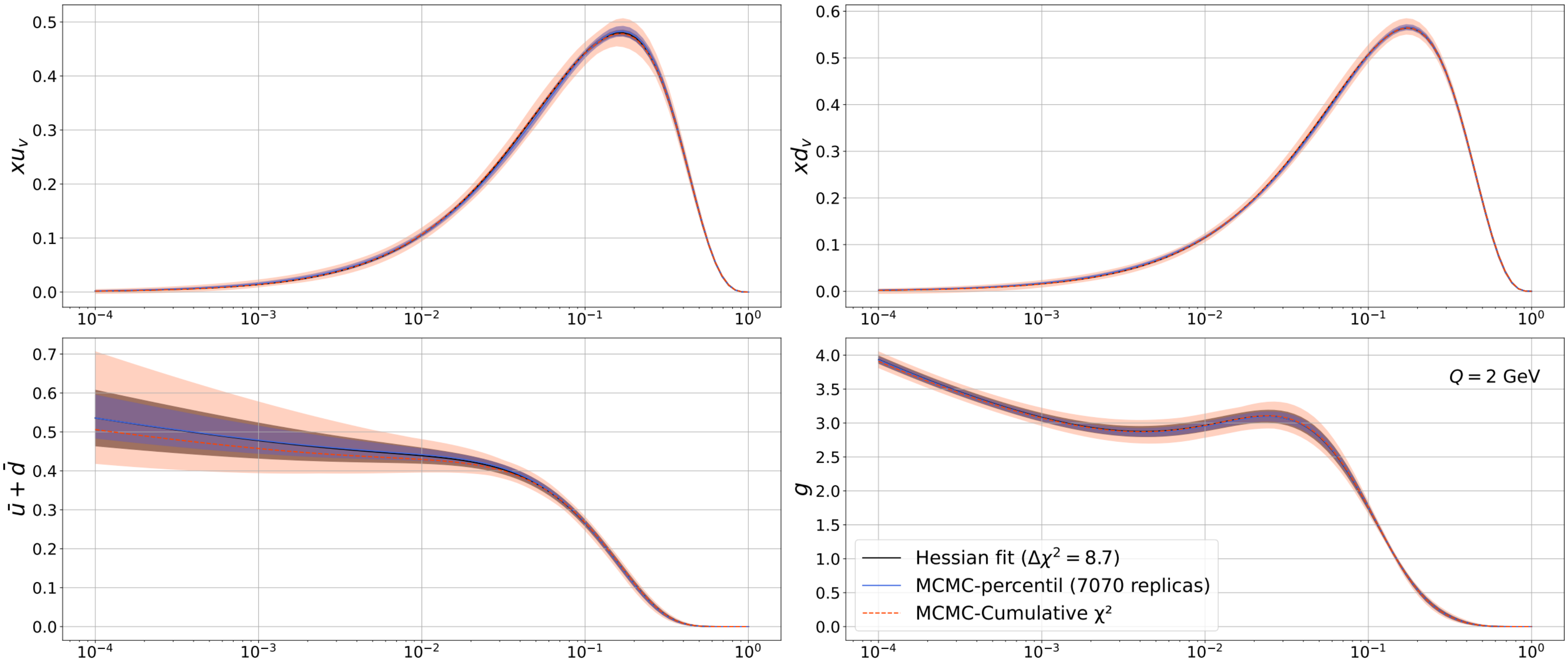
▶ **Percentile method (68% CI asymmetric)**
- central value: 50th percentile of distribution of samples
- lower (upper) bound: 16th (84th) percentile of distribution of samples

▶ **Cumulative χ²** [A. Putze et al., arXiv: 0808.2437]
- central value: the best-fit sample with the minimum χ2 value
- lower (upper) bound: minimum (maximum) value of the the samples found within this 68% χ2 quantile range





Pb^{208} PDF resulting from **MCMC** (percentile & cumulative χ^2 methods for uncertainty estimation) and **Hessian** methods



Conclusion:

- Despite the MCMC challenges (mainly computational cost), this method has become a powerful tool for determining nPDFs and so far we have obtained promising results (comparing with Hessian) for Pb PDF fit

- We would like to extend this approach for multiple nuclei PDF fits and investigate additional statistical methods for estimating Markov Chains uncertainty.

Acknowledgment:
This work was supported by Narodowe Centrum Nauki under grant no.\ 2019/34/E/ST2/00186.

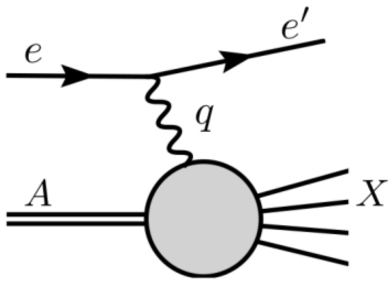
Backup

DIS variables for **nucleus**

$$q \equiv k' - k\,,\ Q^2 \equiv -q^2 \qquad x_A \equiv \frac{Q^2}{2p_A.q}$$

p_A : nucleus momentum

$x_A \in (0,1)$: fraction of the nucleus momentum carried by a nucleon



$$e(k) + A(p_A) \to e'(k') + X$$

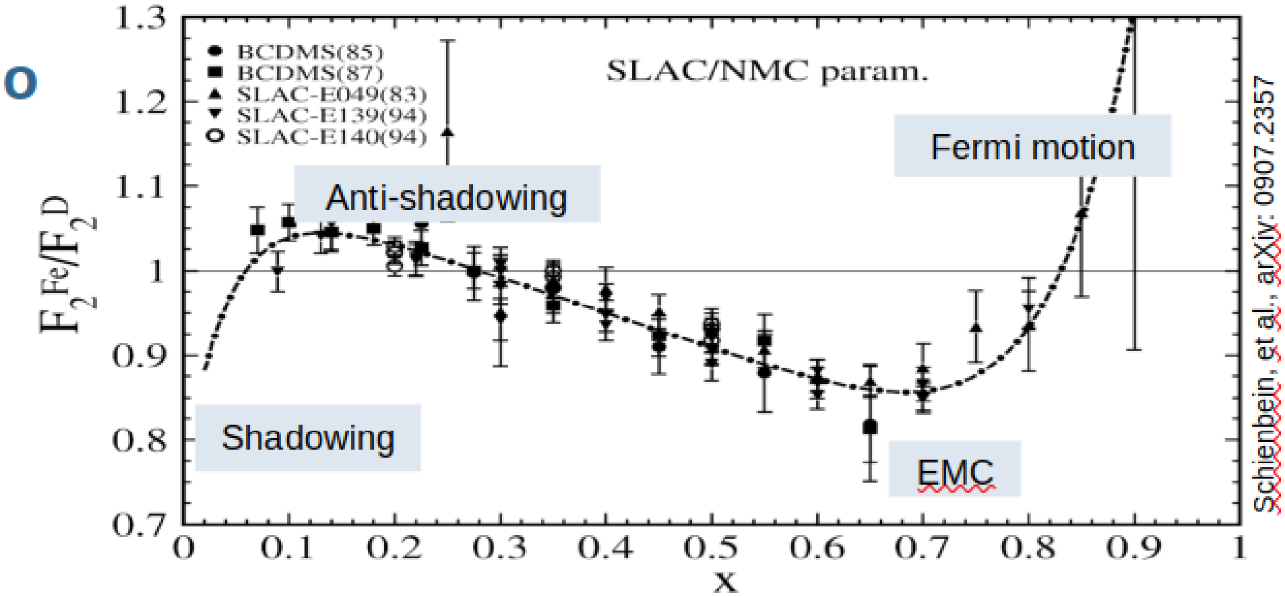
DIS variables for **parton**

$x_N = A x_A$: parton momentum fraction with respect to the average nucleon momentum p_N

$$p_N = \frac{p_A}{A}$$

$x_N \in (0, A)$

Nuclear correction ratio



- **Shadowing**: a suppression due to the overlap of partons from different nucleons at low x which reduce the chance of interacting with the probe
- **Anti-Shadowing**: an enhancement of parton densities, compensates for shadowing based on the momentum sum rule.
- **EMC effect**: a reduction in parton densities due to nuclear binding, Pion Excess, quark clusters, Short-Range Correlations, etc.
- **Fermi motion**: an increase at high x, attributed to the intrinsic motion of nucleons within the nucleus

The underlying dynamics are still to be fully theoretically understood!

Sum rules:

$$\int_0^1 \mathrm{d}x_A\, \tilde{u}_v^A(x_A, Q^2) \quad = \quad 2Z + N\,,$$
$$\int_0^1 \mathrm{d}x_A\, \tilde{d}_v^A(x_A, Q^2) \quad = \quad Z + 2N\,,$$

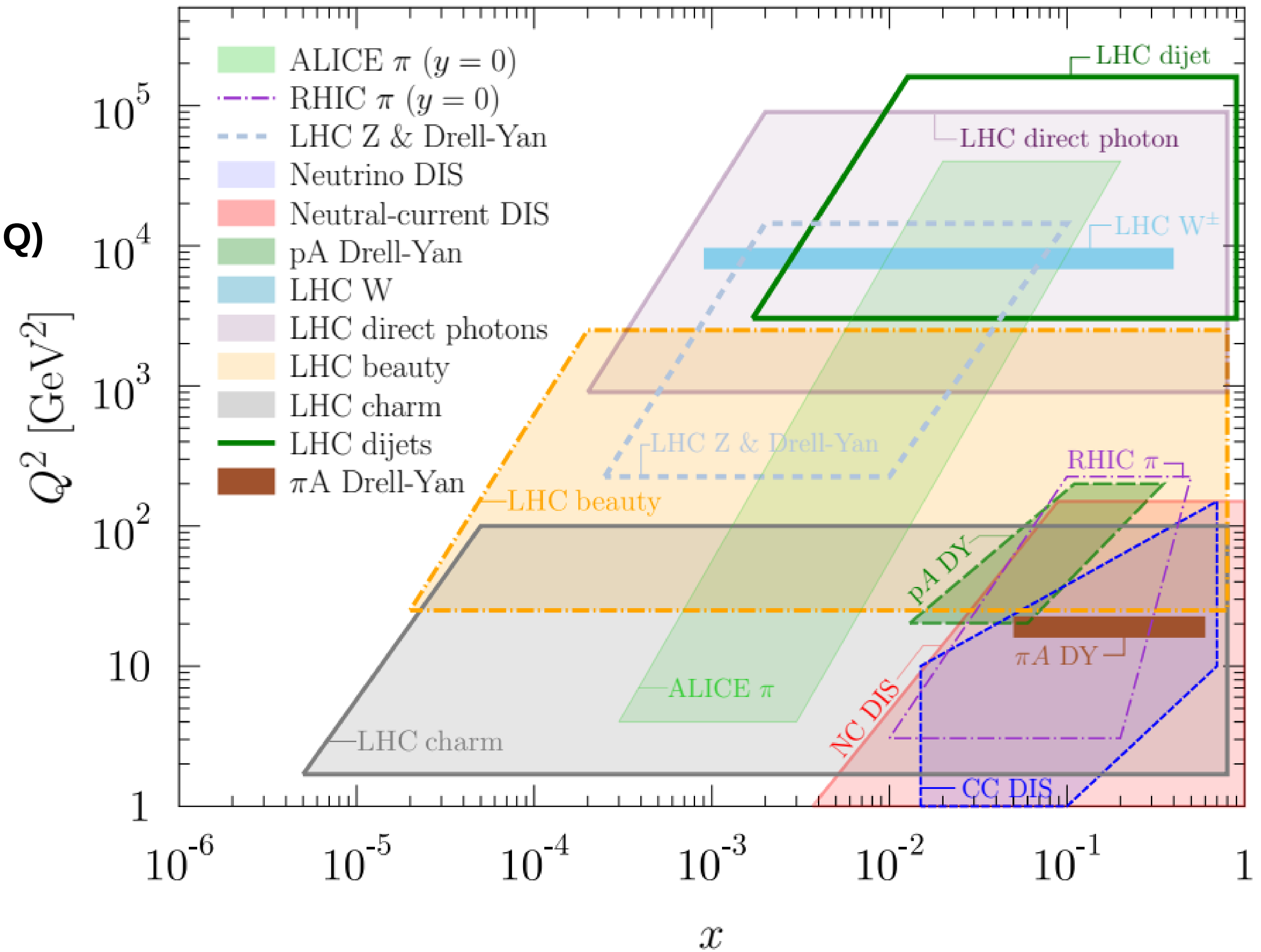
and the momentum sum rule

$$\int_0^1 \mathrm{d}x_A\, x_A \left[\tilde{\Sigma}^A(x_A, Q^2) + \tilde{g}^A(x_A, Q^2)\right] = 1\,,$$

where $N = A - Z$ and $\tilde{\Sigma}^A(x_A) = \sum_i (\tilde{q}_i^A(x_A) + \tilde{\bar{q}}_i^A(x_A))$

Experimental data:

- NC & CC DIS
- LHC W/Z production
- Heavy Quark production (HQ)



nPDF fit setup

$$x f_i^{p/A}(x, Q_0) = c_0 x^{c_1}(1-x)^{c_2}(1 + c_3\sqrt{x} + c_4 x)$$

$$c_k \rightarrow p_k + a_k \ln(A) + b_k \ln^2(A).$$

$$xu_v \rightarrow a_1,\ a_2,\ a_3$$

$$xd_v \rightarrow a_1,\ a_2,\ a_3$$

$$x(\bar{d} + \bar{u}) \rightarrow a_1,\ a_2$$

$$xg \rightarrow a_1,\ a_2$$

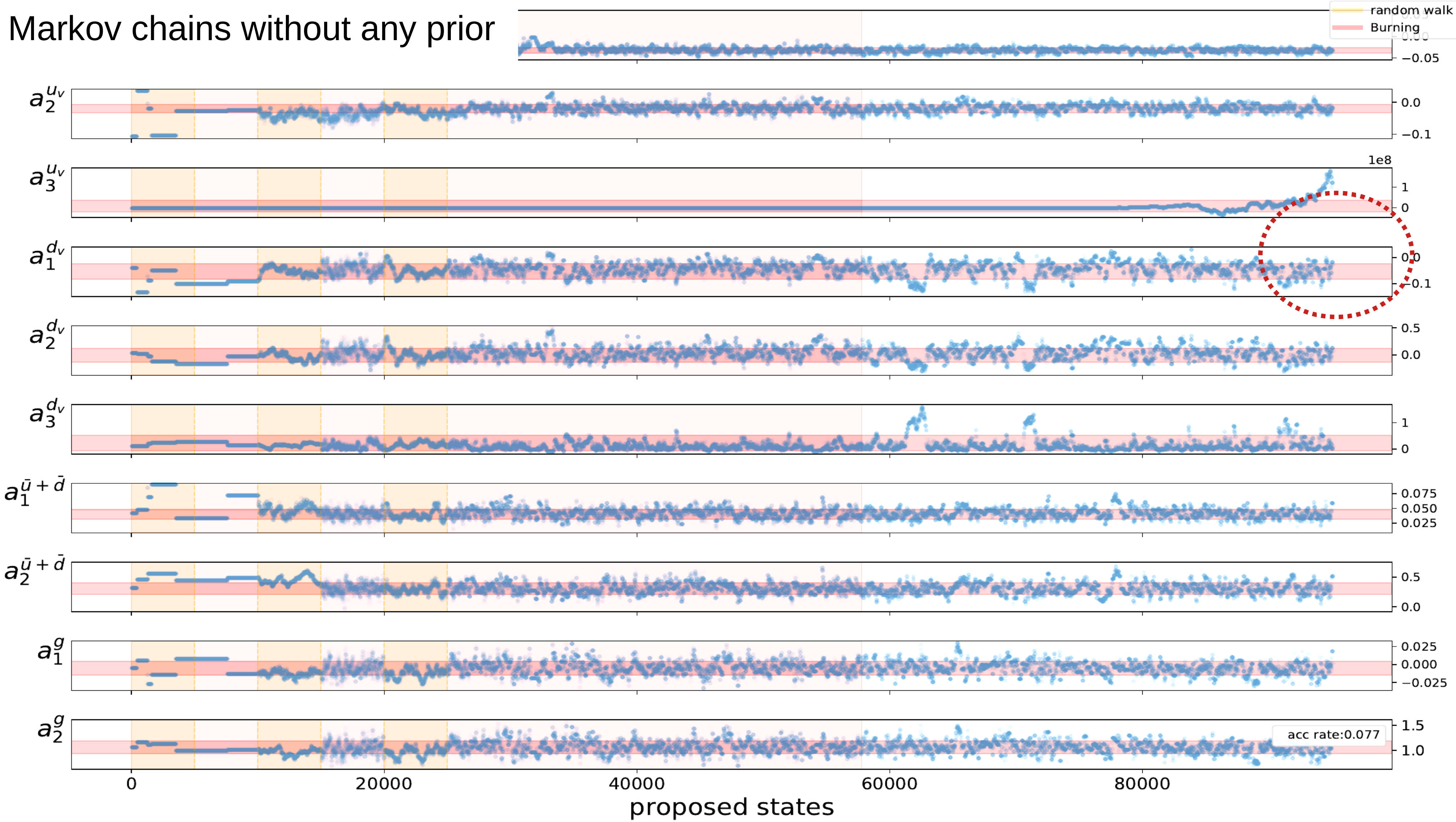
CJ15

Functional form for bound protons at Q_0: $x f_i^{p/A}(x, Q_0) = c_0 x^{c_1}(1-x)^{c_2}(1 + c_3\sqrt{x} + c_4 x)$

Atomic number dependence: $c_k \rightarrow p_k + a_k \ln(A) + b_k \ln^2(A).$

$$f_i^{(A,Z)} = \frac{Z}{A} f_i^{p/A} + \frac{A-Z}{A} f_i^{n/A}$$

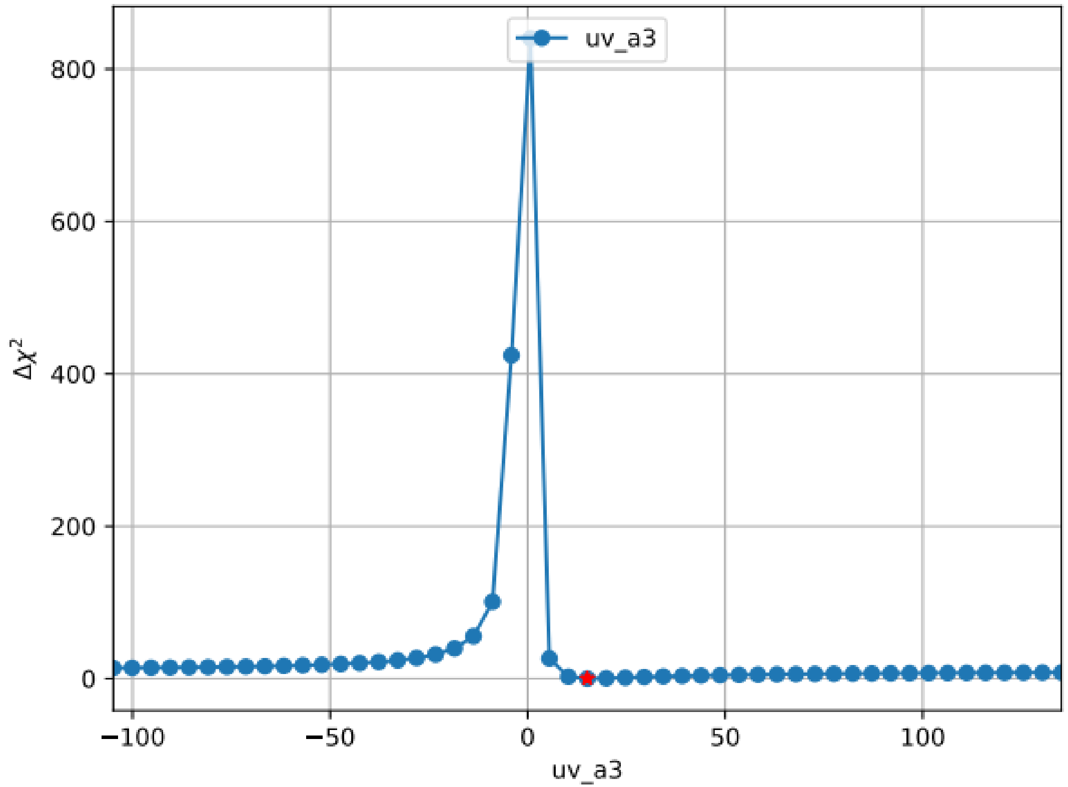
Markov chains without any prior



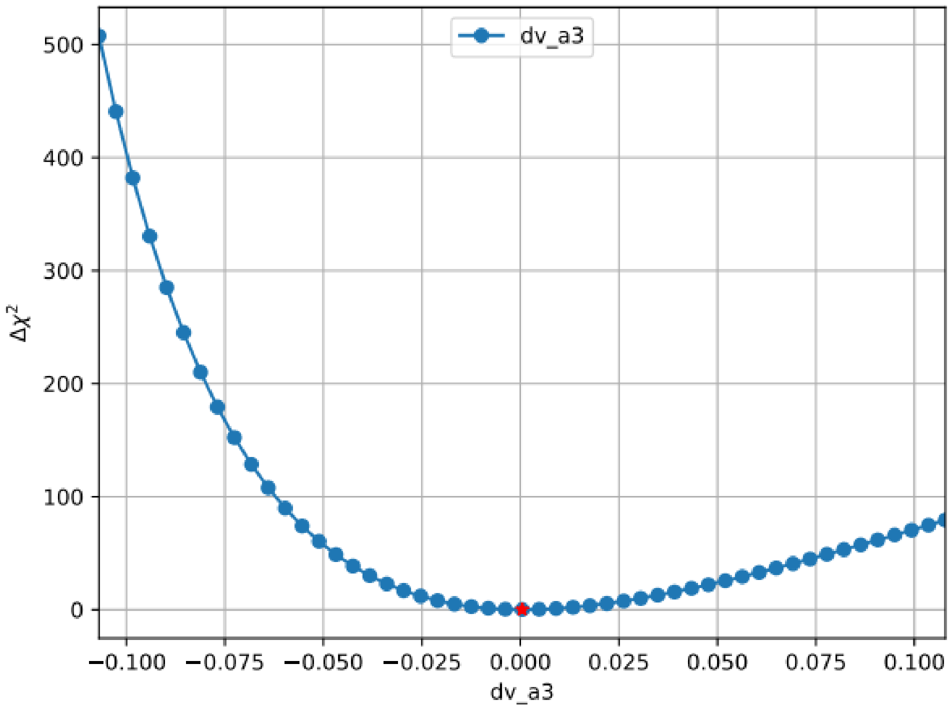
Prior setup:

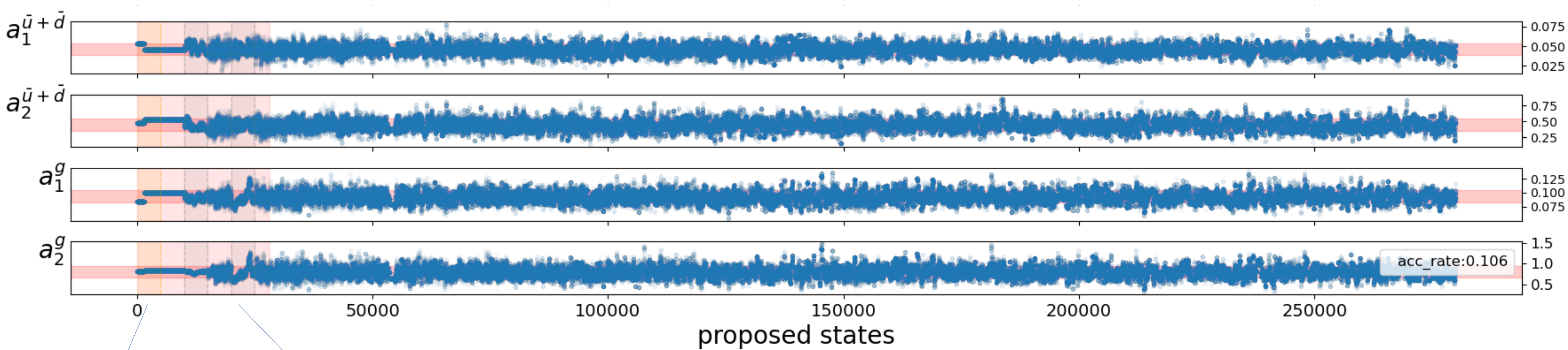
Prior → we just use a uniform prior for the parameter: $a_3{}^{u_v} : U(-300, 300)$

Scan of the χ^2 function along the nPDF parameters:
(varying always one free parameter at a time while other parameters were left fixed at the global minimum)



Scan of the χ^2 function along dv-a3 parameter

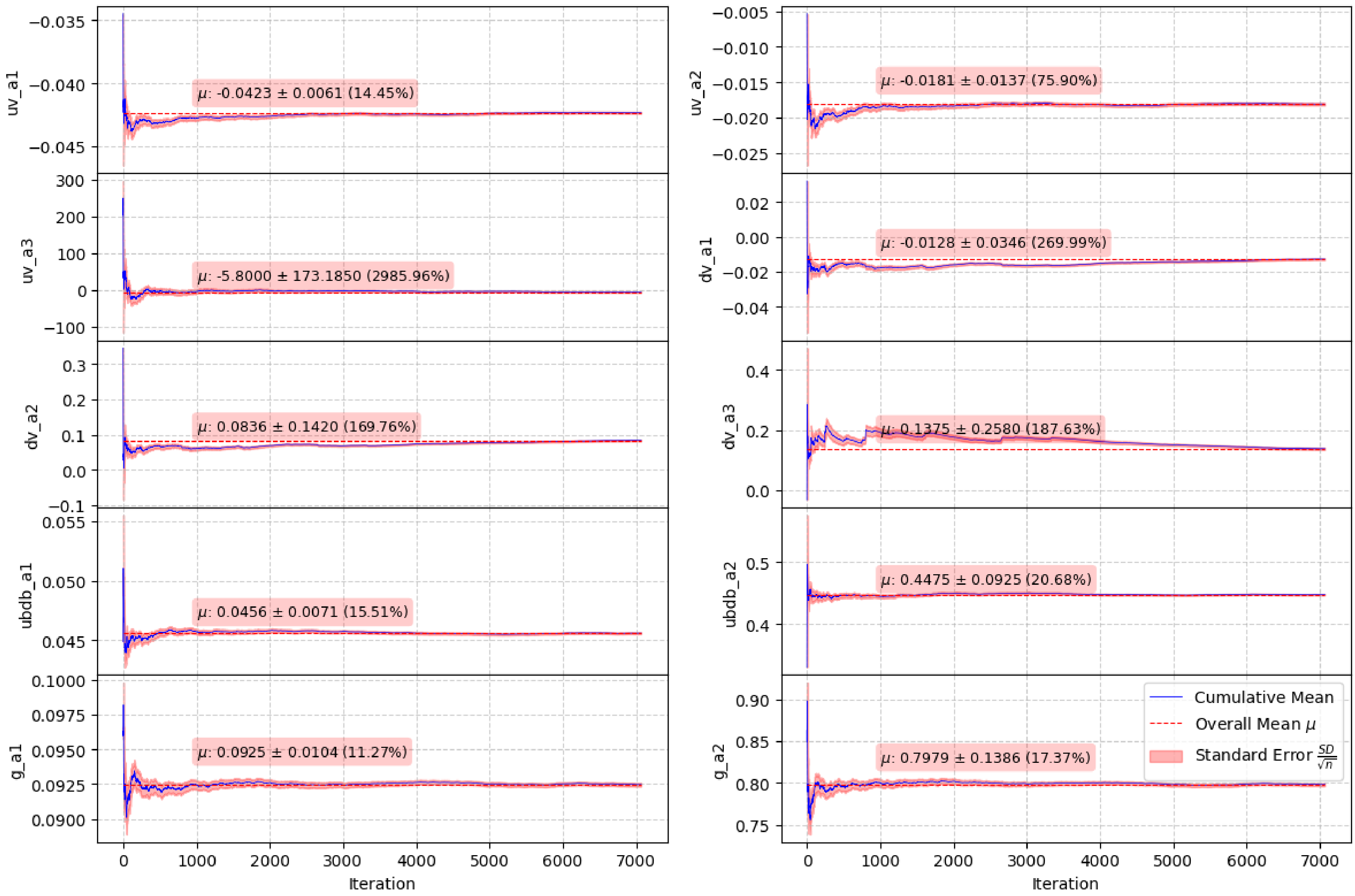




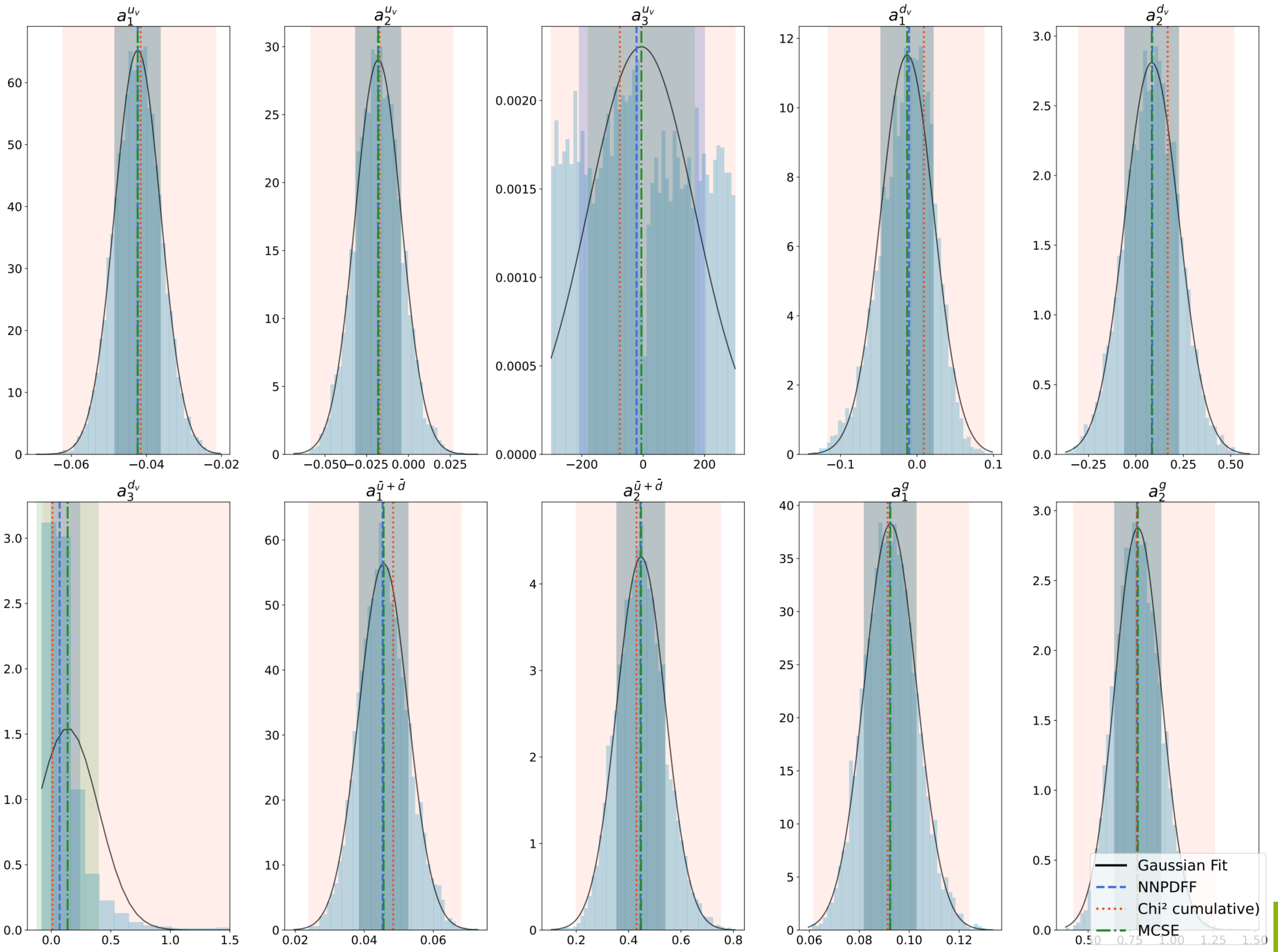
$N_0 = 5000$

Restarting the chain at 10,000 and 20,000

Starting point: global minimum from Hessian fit + Gaussian noise (width= 20 % of minimum value)
Thermalization (burn-in phase): removing first 8000 accepted points



uv_a1
μ: -0.0423 ± 0.0061 (14.45%)
uv_a2
μ: -0.0181 ± 0.0137 (75.90%)
uv_a3
μ: -5.8000 ± 173.1850 (2985.96%)
dv_a1
μ: -0.0128 ± 0.0346 (269.99%)
dv_a2
μ: 0.0836 ± 0.1420 (169.76%)
dv_a3
μ: 0.1375 ± 0.2580 (187.63%)
ubdb_a1
μ: 0.0456 ± 0.0071 (15.51%)
ubdb_a2
μ: 0.4475 ± 0.0925 (20.68%)
g_a1
μ: 0.0925 ± 0.0104 (11.27%)
g_a2
μ: 0.7979 ± 0.1386 (17.37%)
Cumulative Mean
Overall Mean μ
Standard Error SD/√n
Iteration
Iteration



$a_1^{u_v}$
$a_2^{u_v}$
$a_3^{u_v}$
$a_1^{d_v}$
$a_2^{d_v}$
$a_3^{d_v}$
$a_1^{\bar{u}+\bar{d}}$
$a_2^{\bar{u}+\bar{d}}$
a_1^{g}
a_2^{g}
Gaussian Fit
NNPDFF
Chi² cumulative)
MCSE

MH vs adaptive MH