KRAKOW SCHOOL OF INTERDYSCIPLINARY PHD STUDIES

nCTEQ
nuclear parton distribution functions

Determination of Nuclear PDFs using Markov Chain Monte Carlo

[Based on arXiv:2603.13150]

Nasim Derakhshanian

IFJ PAN
Department of Theoretical Particle Physics
25 June 2026

This work was supported by Narodowe Centrum Nauki under grant no.\

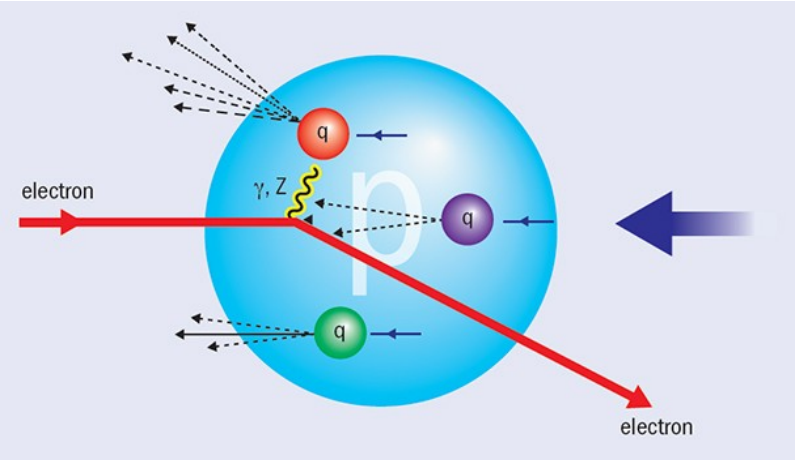
Parton Distribution Function (PDF):

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:

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

PDFs

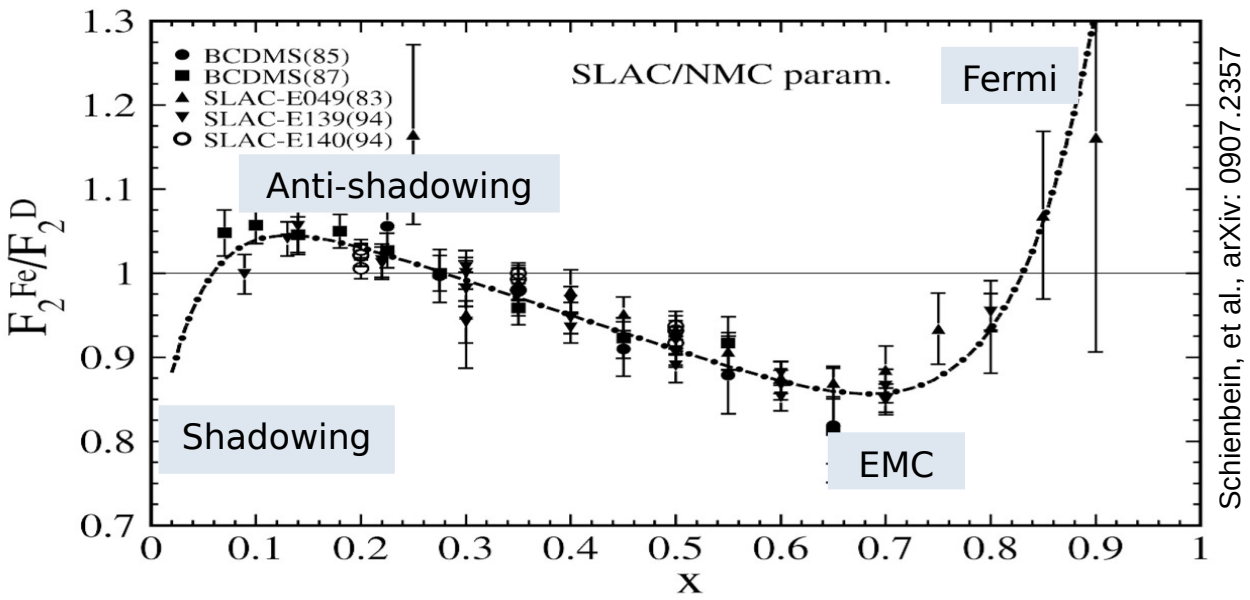
Partonic scattering cross-section

- PDFs encode the non-perturbative structure of hadrons.

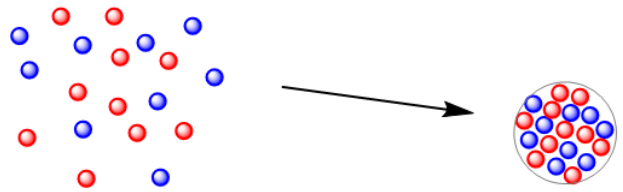
Nuclear PDFs (nPDFs):

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

$$F_2^A(x) \neq ZF_2^p(x) + NF_2^n(x)$$



Schienbein, et al., arXiv: 0907.2357



Nuclear correction ratio:

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

- Nuclear effects can be encoded into universal nPDFs.

Global Analysis of nPDF

- The Q dependence is governed by pQCD (DGLAP evolution equations)
- The x dependence must be parametrized at an initial scale Q_0

$$x f_i^A(x, Q_0^2, \mathbf{a}) = x^{a_1}(1-x)^{a_2} P(x, \mathbf{a})$$

Global PDF fits: PDFs are extracted by fitting theory predictions to experimental data.

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

measure — fit — predict

measured cross section from diverse experiments

Methodology

nCTEQ, EPPS, NNPDF, ect

Precision Calculations: NLO & NNLO

Standard library (**LHAPDF**)

Global Analysis of nPDF

Input function at Q_0

Parameterize nuclear PDF at initial scale: 1.3GeV

$$xf(x, Q_0, \mathbf{a}) = x^{a_1}(1-x)^{a_2} P(x, \mathbf{a})$$

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

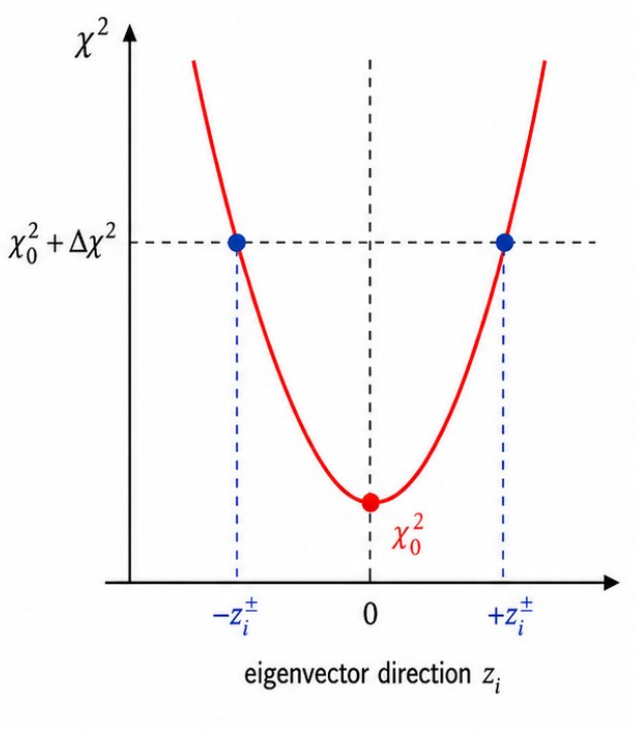
$$\chi^2(\mathbf{a}) = \sum_{i,j} (D_i - T_i(\mathbf{a}))(C^{-1})_{ij}(D_j - T_j(\mathbf{a}))$$

Uncertainties estimation

Traditional methods to nPDF uncertainty estimation

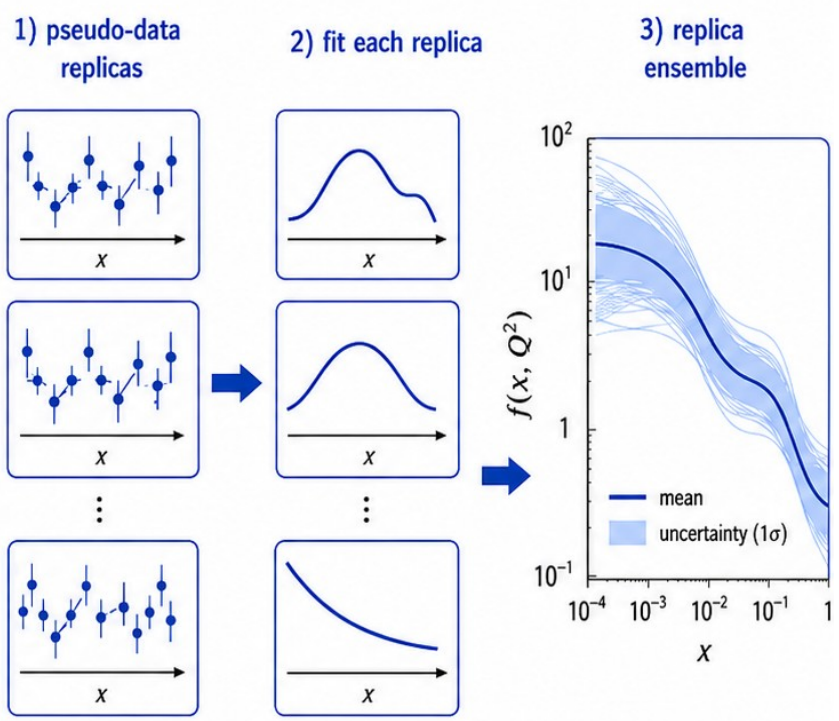
Hessian method (nCTEQ / EPPS)

- expand χ^2 around the minimum
- assume local quadratic behavior
- uncertainties from Hessian eigenvectors



Monte Carlo replica method (NNPDF)

- generate pseudo-data replicas
- fit each replica independently
- uncertainties from the replica ensemble



Challenges in nPDF uncertainty estimation

nPDF-specific challenges:

- Limited nuclear data → weakly constrained uncertainties
- Dataset tensions complicate global fits
- Proton–neutron isospin symmetry introduces parameter correlations

Limitations of traditional methods:

Hessian method

- Assumes near-Gaussian uncertainties
- Depends on tolerance choice
- Not sensitive to local minima

Monte Carlo replica method [arXiv:2404.10056]

- Replicas may not reproduce the exact posterior
- Uncertainties can be underestimated in nonlinear regions

Toward a Full Posterior Description

We would like to:
sample the full posterior distribution of nPDF parameters, including nonlinear correlations, multiple minima, and statistically consistent uncertainties.

Bayes theorem: $$p(a|D) \propto \mathcal{L}(D|a)\, p(a)$$

Posterior — Likelihood — Prior

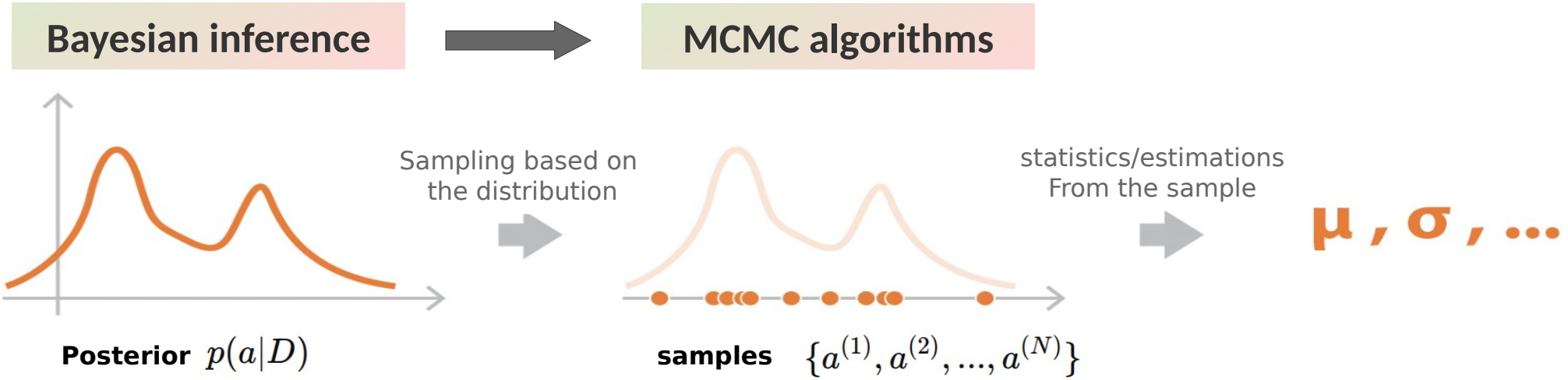
- Likelihood given by the χ^2 function: $$\mathcal{L}(D|a) \propto \exp\left(-\frac{1}{2}\chi^2(\boldsymbol{a})\right)$$

Problem: the posterior is too complex to sample analytically.

Solution: **MCMC** constructs a Markov chain whose samples follow the posterior distribution.

Markov Chain Monte Carlo (MCMC)

- MCMC constructs a Markov chain whose stationary distribution is the posterior p(a|D).
- Samples are used to estimate posterior quantities such as means, variances, and credible intervals.



Key idea: high-posterior-probability regions are sampled more frequently.

MCMC Algorithm — Adaptive Metropolis

[H. Haario et al.: "An adaptive Metropolis algorithm", Bernoulli 7.2 (Apr. 2001)]

Standard Metropolis:

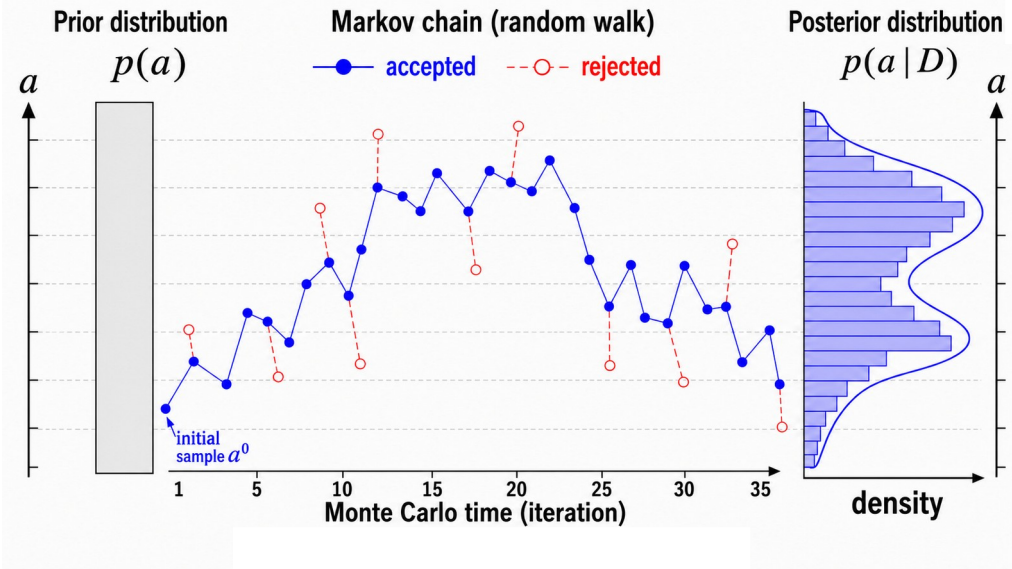
1) Initialize parameters $\boldsymbol{a}^0$

2) Propose a new sample: $\tilde{\boldsymbol{a}}^{i+1} \sim \mathcal{N}(\boldsymbol{a}^i, C_0)$

3) Compute acceptance probability:

$$\alpha = \min\left(1, \frac{p(\tilde{\boldsymbol{a}}^{i+1}|D)}{p(\boldsymbol{a}^i|D)}\right)$$

4) $$a^{i+1} = \begin{cases} \tilde{a}^{i+1}, & \mathcal{U}(0,1) < \alpha \quad \textbf{accepted} \\ a^i, & \textbf{rejected} \end{cases}$$



Adaptive proposal:

After N_0 samples, the proposal covariance is learned from the chain history to improve sampling efficiency.

$$\tilde{\boldsymbol{a}}^{i+1} \sim (1-\beta)\,\mathcal{N}(\boldsymbol{a}^i, s_d.\bar{C}_i) + \beta\,\mathcal{N}(\boldsymbol{a}^i, C_0)$$

- $\bar{C}_i$ is a self-learn covariance and $0 < \beta < 1$ controls mixing with the fixed proposal

Fitting setup: data and theory inputs

Theory inputs:

NLO theory prediction | Q>2 GeV, W>3.5 GeV, p_T>3 GeV | **10** free parameters

Pb-only fit (primary):

1488 data points

- CHORUS neutrino DIS on Pb
- W/Z production in pPb at the LHC
- heavy-quark production in pPb at the LHC

Multi-nuclei fit:

1995 data points

- Pb-only dataset
- neutral-current DIS on multiple nuclei
- Drell–Yan process

Fitting setup: parameterization

1)Bound-proton PDF at Q_0

[Accardi et al., arXiv:1602.03154]

$$x f_i^{p/A}(x, Q_0) = c_0^i \, x^{c_1^i} (1-x)^{c_2^i} \left(1 + c_3^i \sqrt{x} + c_4^i x\right) \qquad i = \{u_v, \bar{u}+\bar{d}, g, s+\bar{s}\}$$

2)Nuclear A-dependence

$$c_j(A) = p_j + a_j \ln A + b_j \ln^2 A$$

$$\begin{cases} u_v & \to & a_1 & a_2 & a_3 \\ d_v & \to & a_1 & a_2 & a_3 \\ \bar{u}+\bar{d} & \to & a_1 & a_2 \\ g & \to & a_1 & a_2 \end{cases}$$

3)Nuclear PDF combination

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

- The **MCMC** samples the free 10 parameters of this nPDF parameterization.

Fitting setup: MCMC sampling

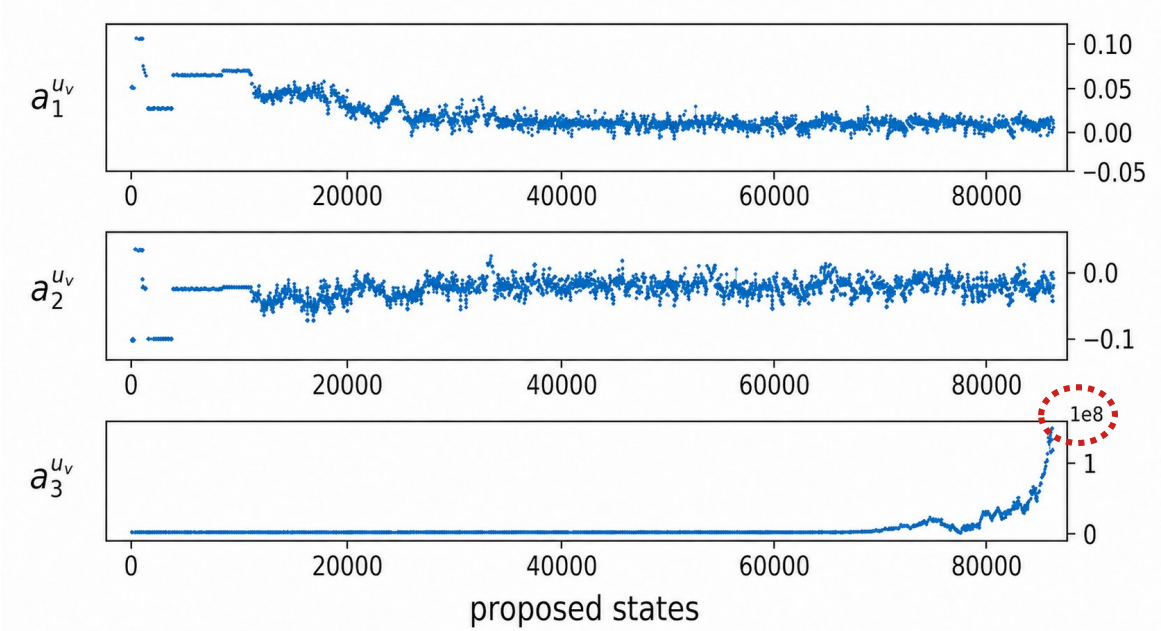
- Adaptive Metropolis-Hasting algorithm
- Multiple independent chains with different random seeds
- Initial parameters generated around the Hessian best fit

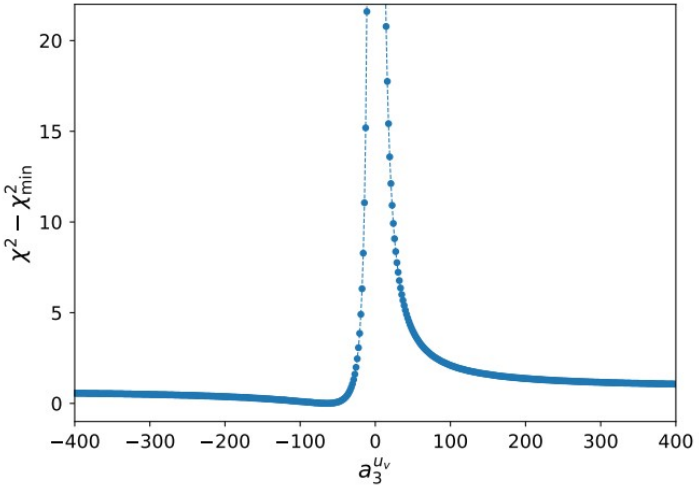
$$a_i^{(0)} = a_{i,\mathrm{Hessian}}(1 \pm 20\%)$$

- Fixed proposal during the first N=5000 steps
- **Pb-only** fit: 20 000 iterations/day/core
- **Multi-nuclei** fit: 10 000 iterations/week/core

Priors and weakly constrained parameter

- Most fitted nPDF parameters are assigned non-informative flat priors.
- In the Pb-only fit, $a_3{}^{u_v}$ is weakly constrained by the data.
- We therefore use a **bounded uniform prior** $a_3{}^{u_v} : U(-300, 300)$
- This regularizes the flat direction without fixing the parameter.



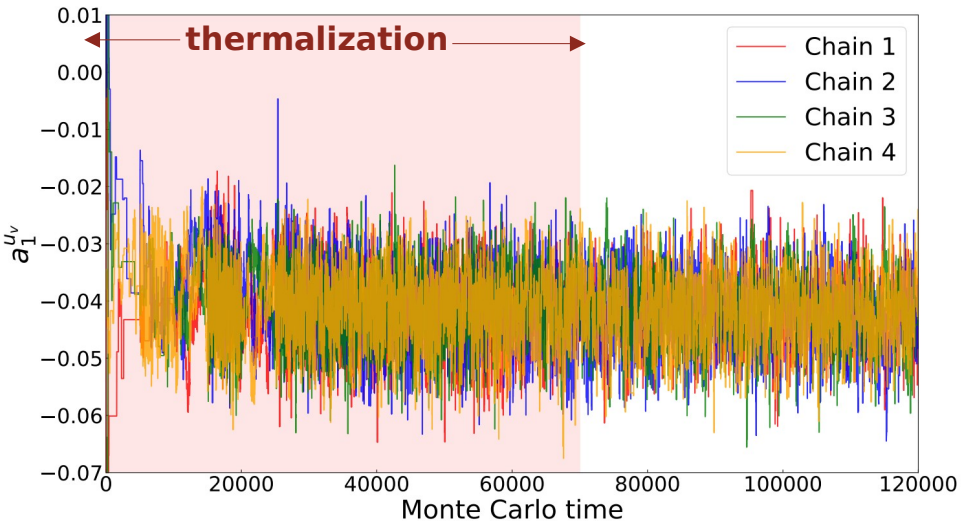


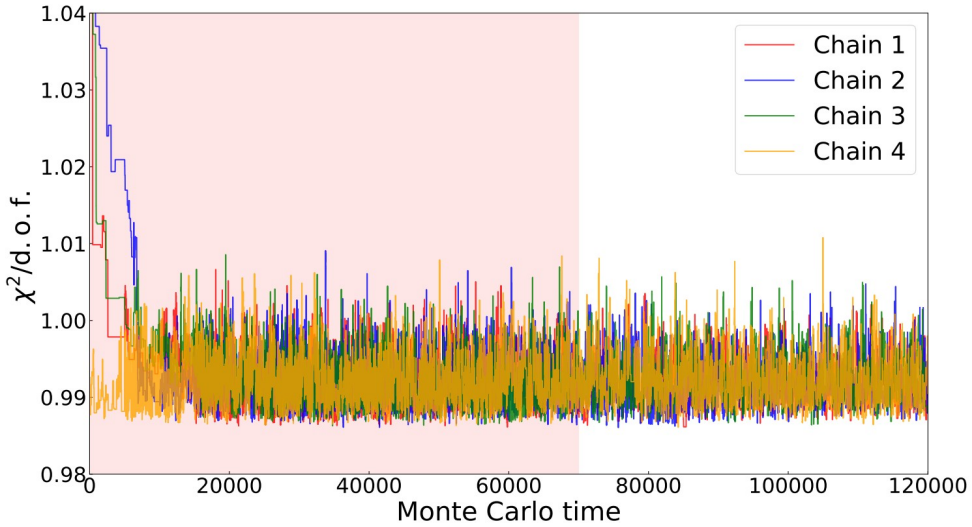
χ^2 scan along $a_3{}^{u_v}$

MCMC diagnostics

Thermalization and convergence:

- Initial samples depend on the starting point.
- The early phase is removed as **burn-in** / **thermalization**.
- **Convergence** is checked using trace plots and χ^2/d.o.f. stability.
- Only stationary post-burn-in samples are used for posterior analysis.

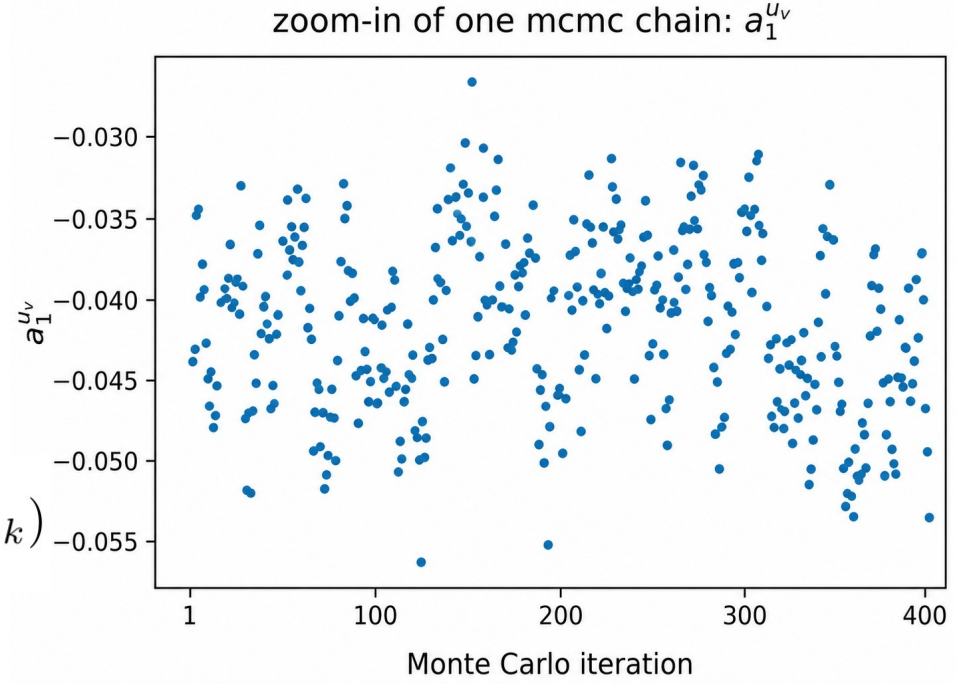




Note: convergence diagnostics are heuristic; they provide evidence of stability, not a mathematical proof.

Autocorrelation

- Successive MCMC samples are correlated.
- Correlation reduces the number of effectively independent samples.
- This affects uncertainty estimation.



Autocorrelation function(AFC):

$$\rho(k) = \frac{\Gamma(k)}{\Gamma(0)}, \qquad \Gamma(k) = \mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{X}_{i+k})$$

Autocorrelation time: $\tau_{\mathrm{int}} \approx \frac{1}{2} + \sum_{k=1}^{\infty} \rho(k)$

Estimated with the Γ-method

[U. Wolff, CPC 156 (2004) 143]

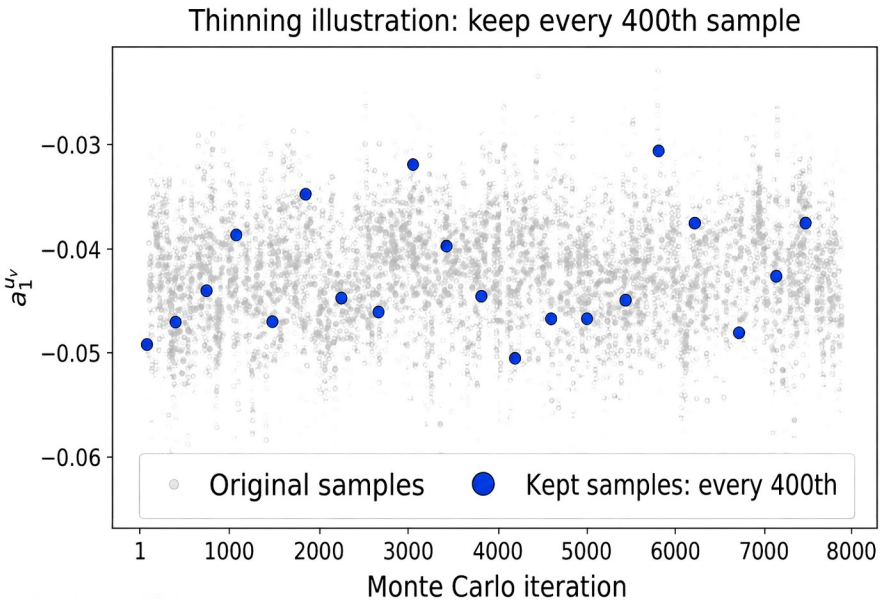
▶ uncertainty is underestimated if correlation is ignored

Thinning

- Keep one sample every η iterations and discard the rest in between.

Why Thinning?

- Reduces autocorrelation, so the remaining samples are approximately independent.
- Allows standard Monte Carlo error estimation.
- Reduces the number of PDF grids, making the final LHAPDF set practical.



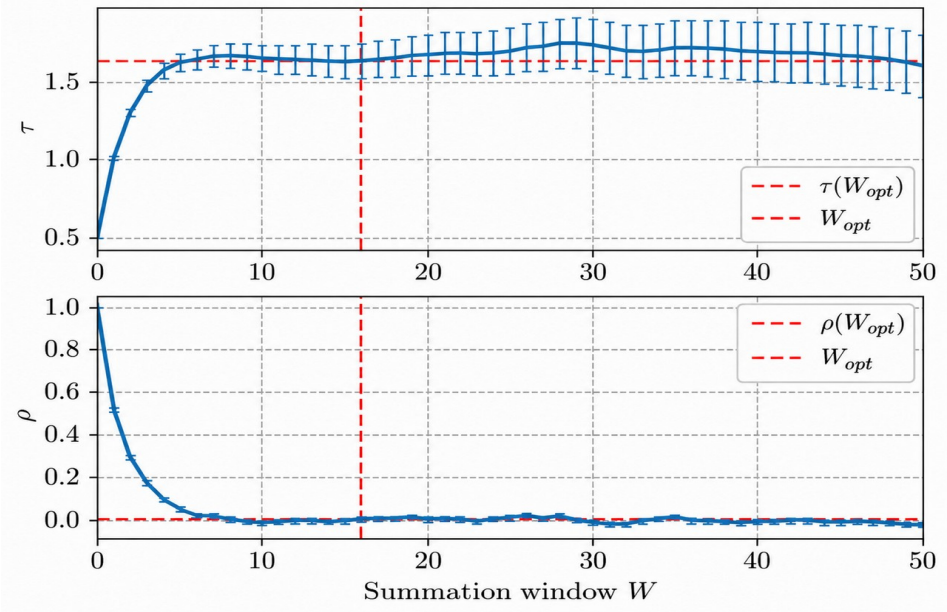
$$\sigma^2_{MCMC} = 2\,\tau_{int}\,\sigma^2_{MC} \quad \Rightarrow \quad \sigma^2_{MC} = \frac{1}{n-1}\sum_{t=1}^{n}(X_t - \hat{\mu})^2$$

Estimating autocorrelation time with the Γ-method

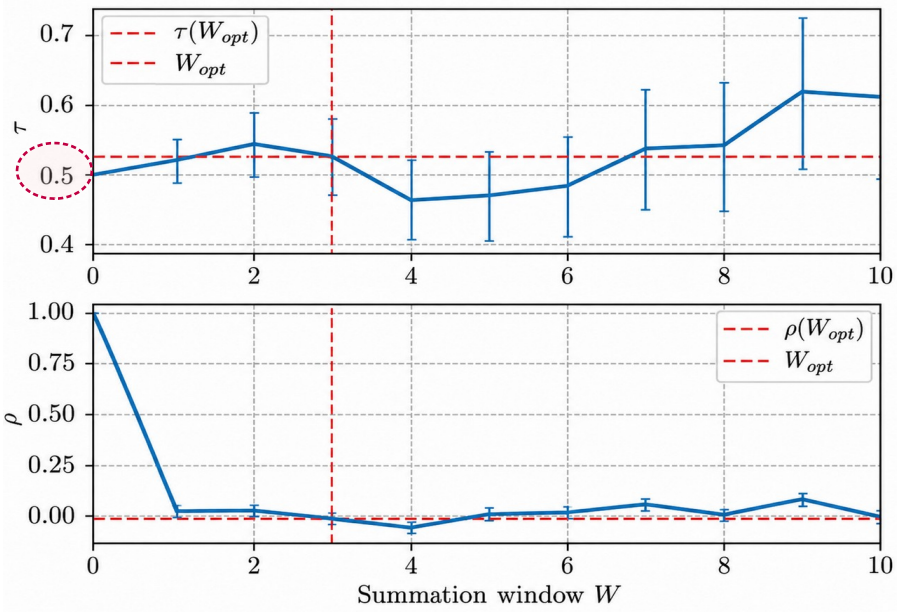
- The **Γ-method** estimates τ_{int} by summing the ACF up to an optimal window W_{opt}
- If samples are still correlated, $\tau_{int}>0.5$
- For effectively independent samples, $\rho(k)\approx 0$ and $\tau_{int}\approx 0.5$

$$\tau_{\mathrm{int}} \approx \frac{1}{2} + \sum_{k=1}^{W_{\mathrm{opt}}} \rho(k)$$

Thinned by 50



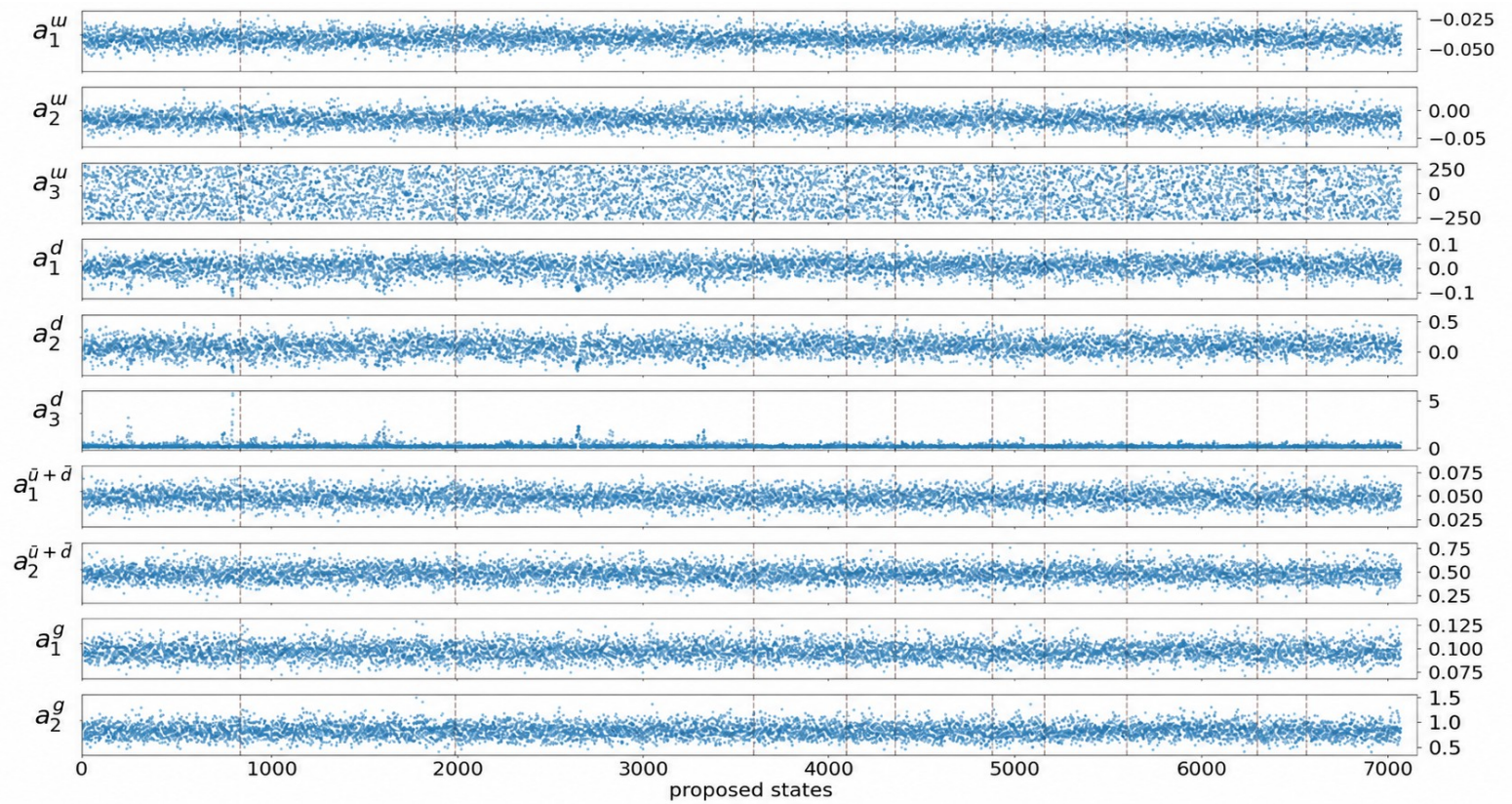
Thinned by 600



Results: Pb-only analysis

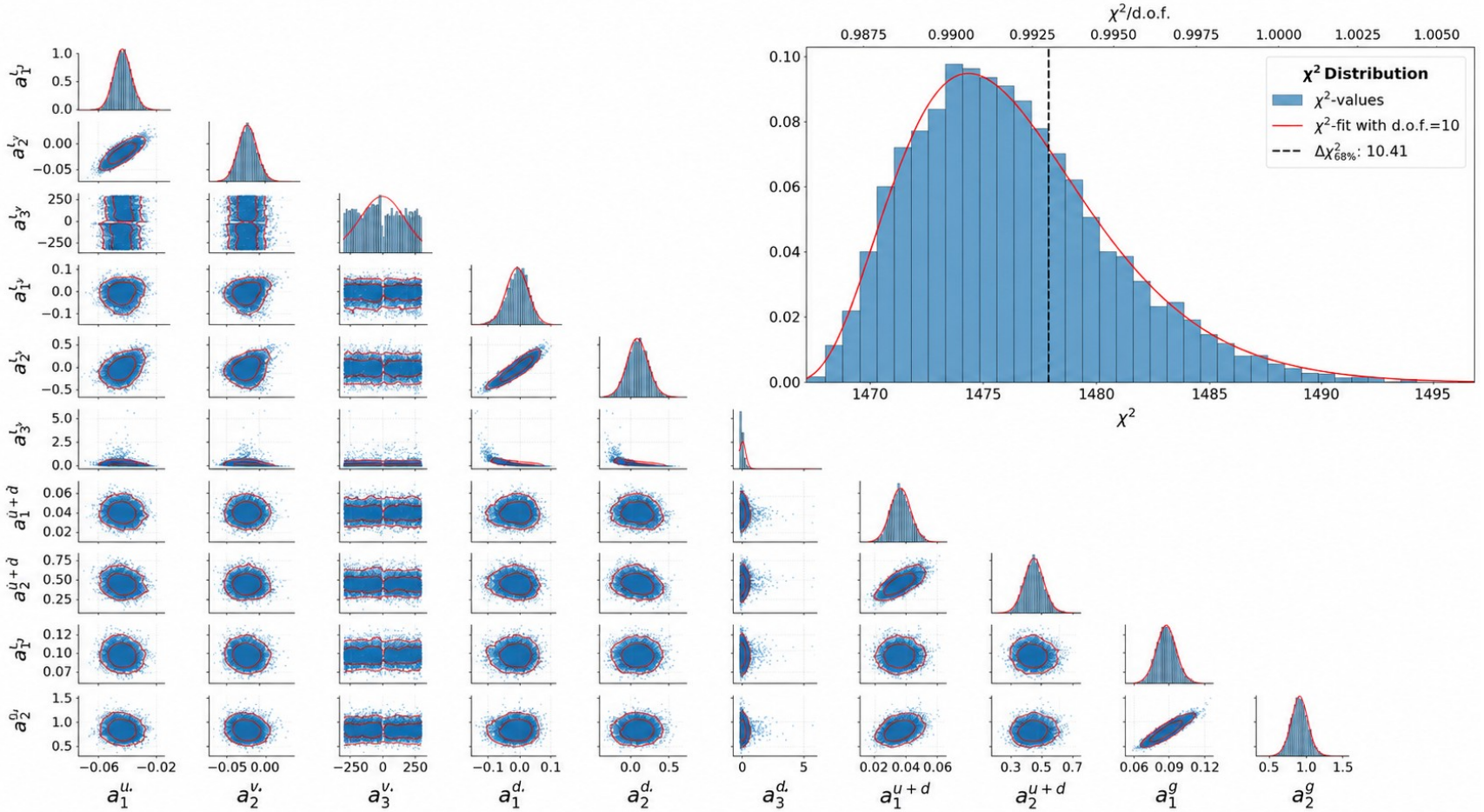
Final posterior ensemble:

- **11** converged Pb-only chains
- burn-in removed: typically 60k–80k samples
- post-burn-in samples: about **3.8 million**
- thinning factor: η~400–600
- final independent ensemble: **7069** samples



Posterior structure: pairwise plot

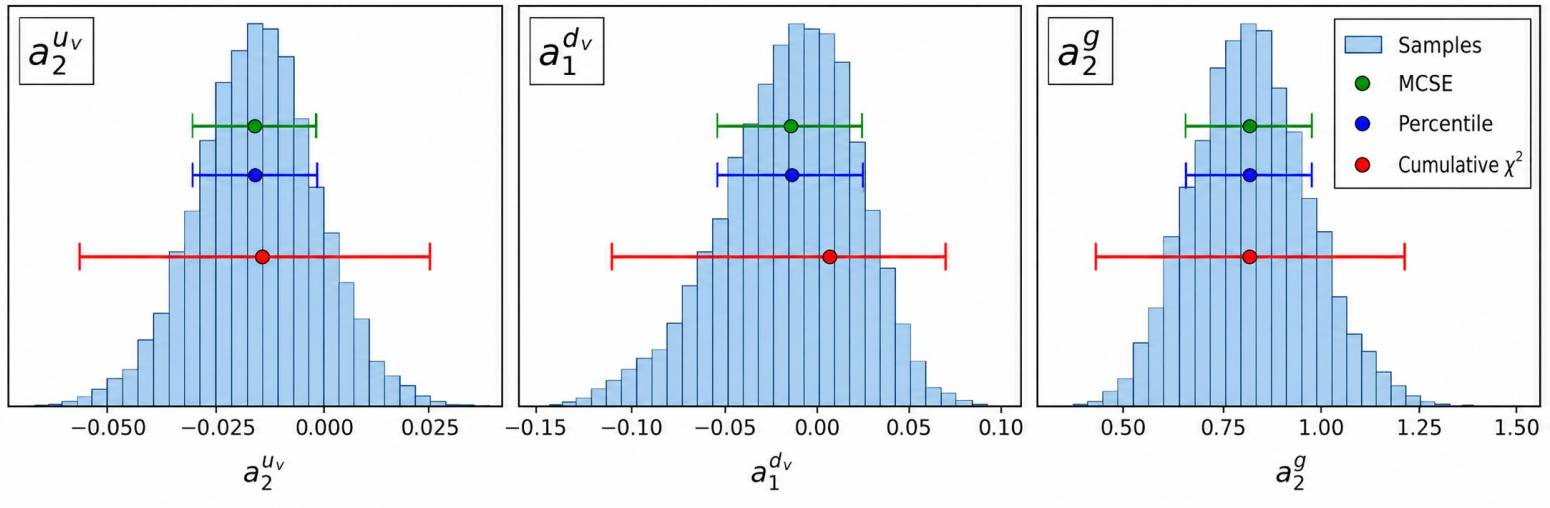
- Diagonal panels: marginal posterior distributions of fitted parameters.
- Off-diagonal panels: parameter correlations.
- MCMC reveals non-Gaussian shapes, correlations, and possible multi-modal structures.



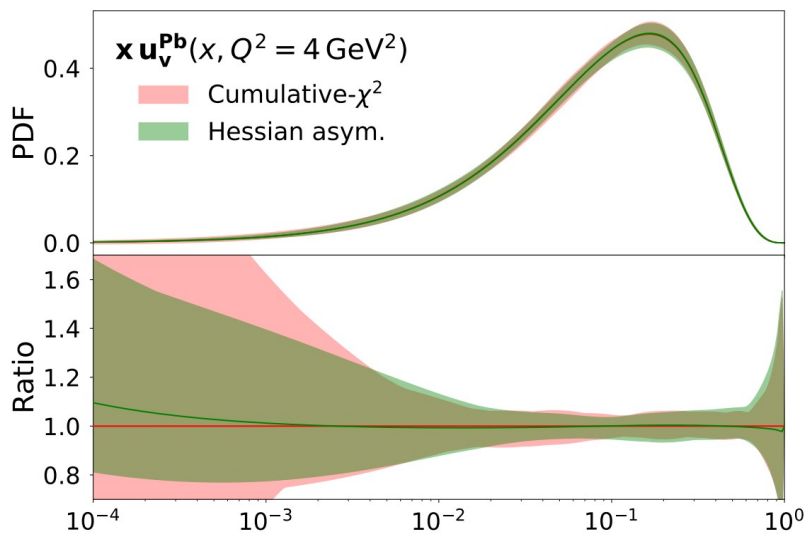
Uncertainty estimation from posterior samples

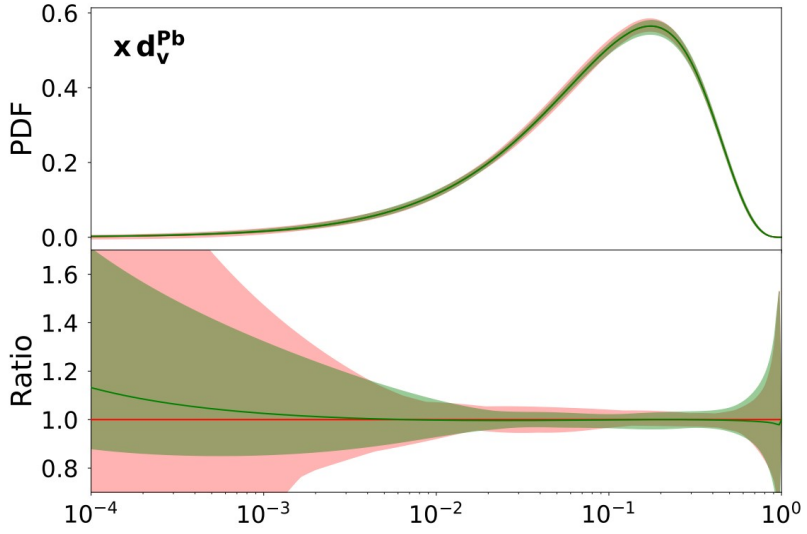
- Once the chains are processed, each remaining sample corresponds to one possible nPDF set.

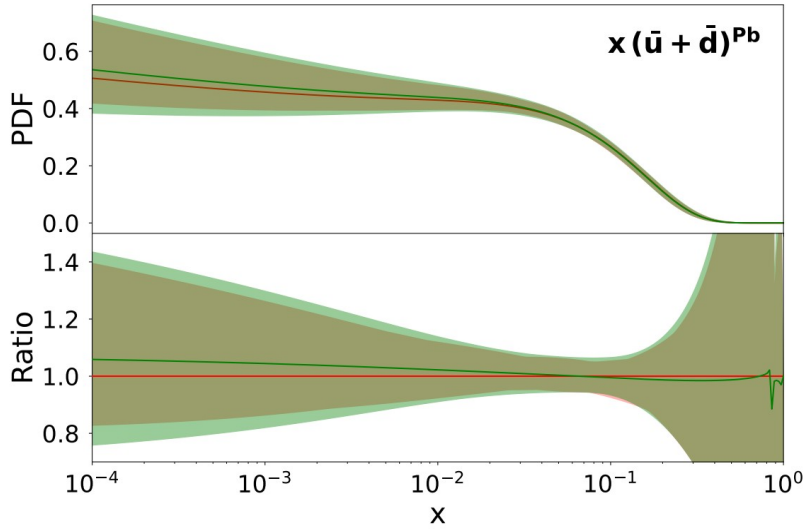
Method	Central value	68% uncertainty
MCSE	Mean (μ)	$\mu \pm$ S
Percentile	Median	16th–84th percentiles
Cumulative-χ^2	Minimum-χ^2 sample	68% lowest-χ^2 region

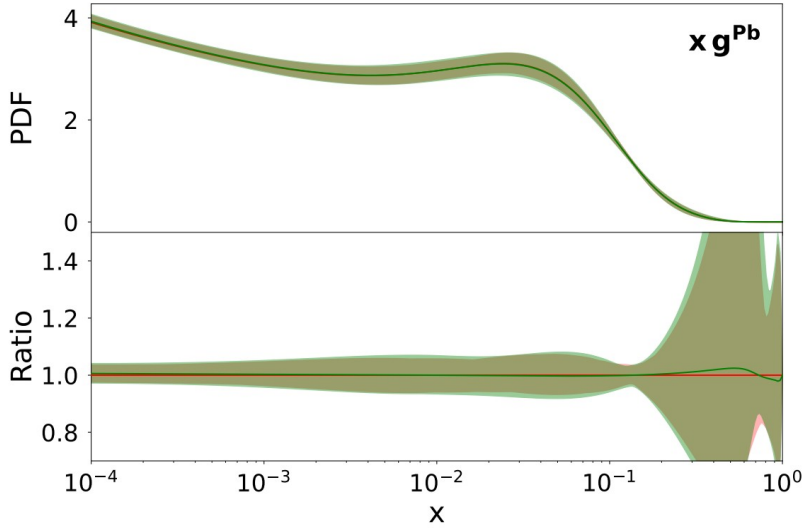


Pb PDFs: Hessian vs MCMC

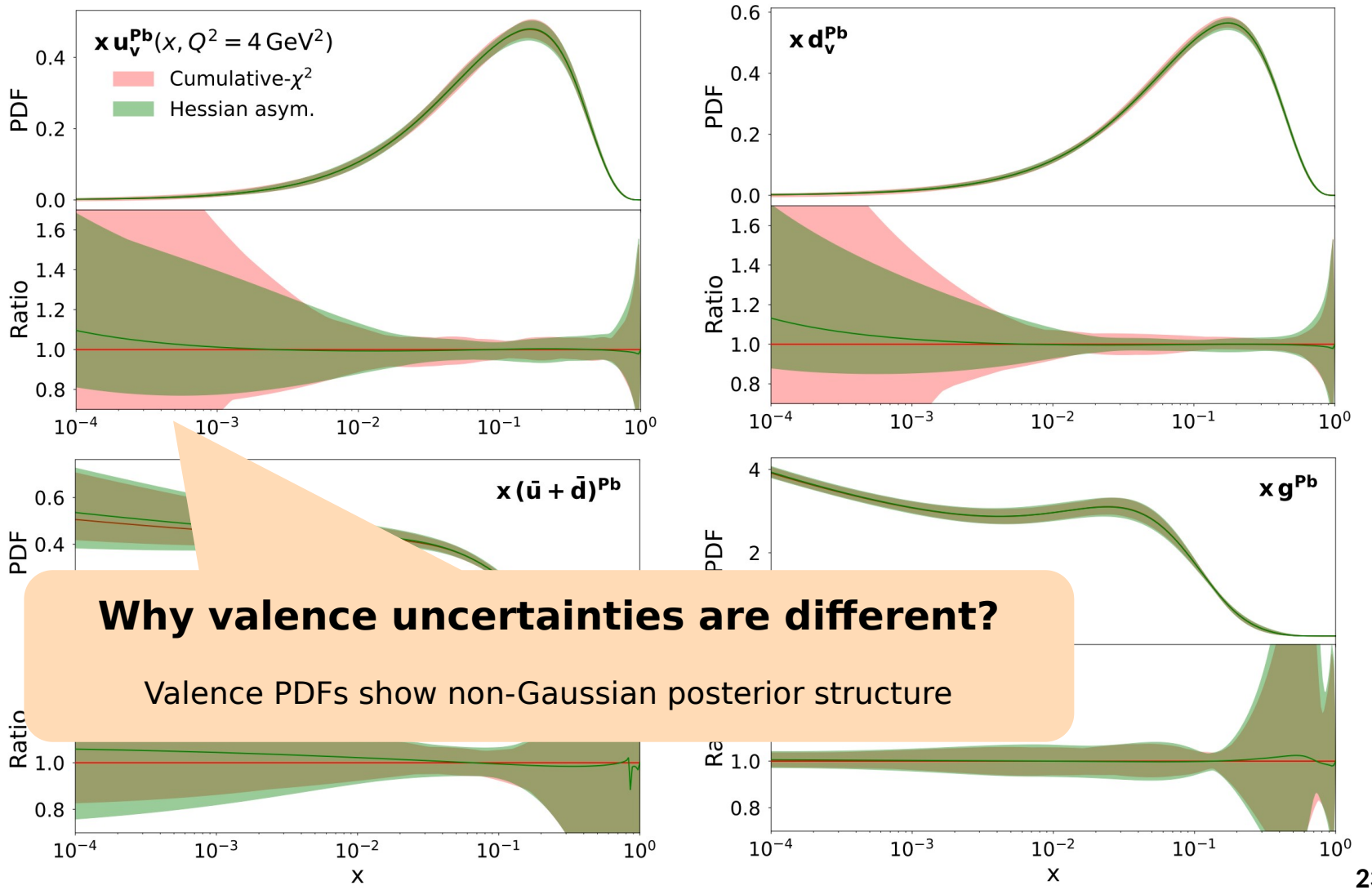








Pb PDFs: Hessian vs MCMC

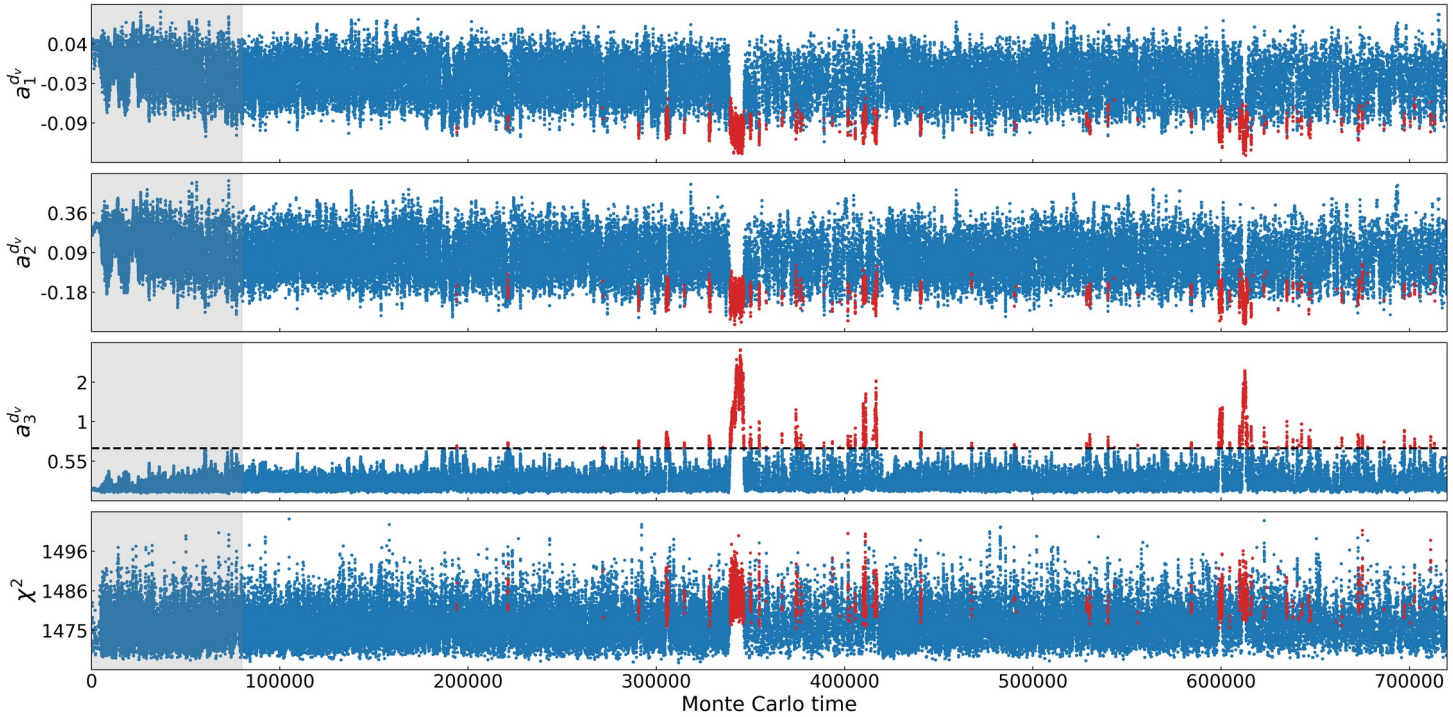


Why valence uncertainties are different?

Valence PDFs show non-Gaussian posterior structure

Secondary structure in d-valence

- For the parameter $a_3^{d_v}$ the chain occasionally moves to a region with larger value $a_3^{d_v} > 0.8$



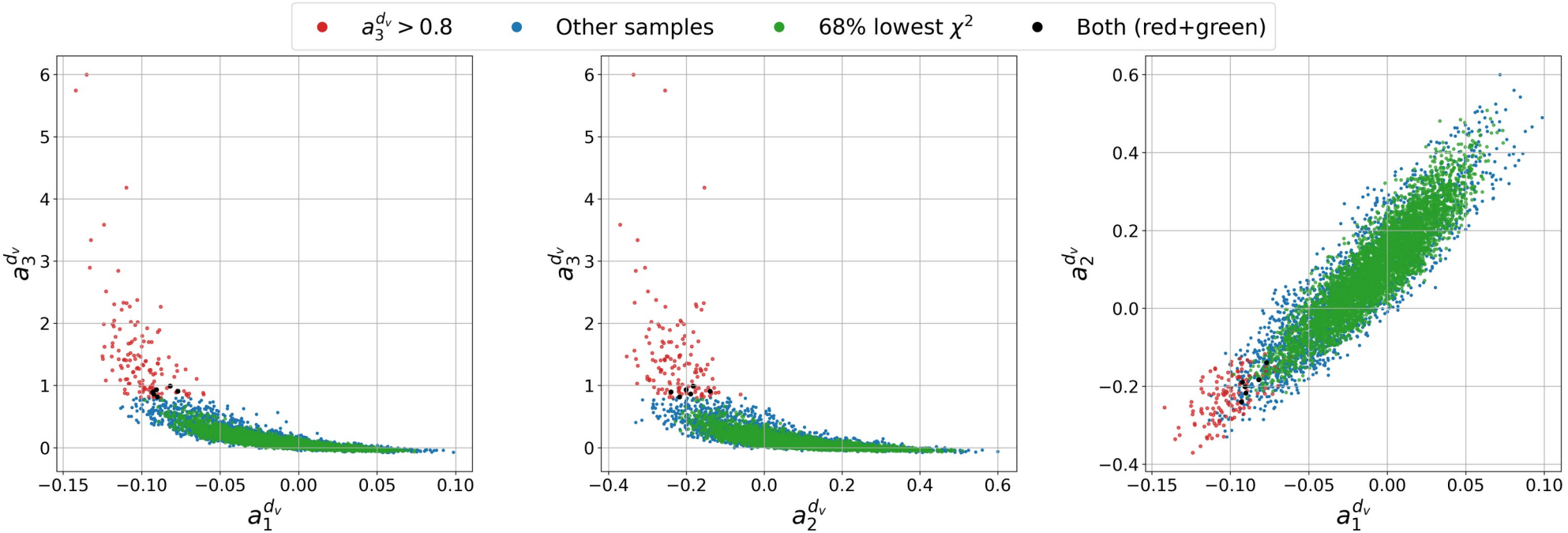
- The secondary region is less probable, but still well sampled and relevant for the uncertainty.

$$N_{\text{blue}}/N_{\text{red}} = 37.02$$

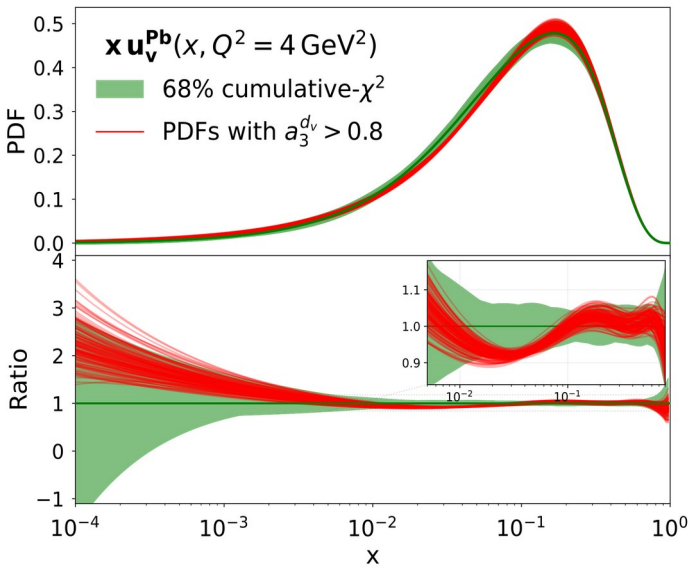
$$\mathcal{L}_{\text{blue}}/\mathcal{L}_{\text{red}} = 42.57$$

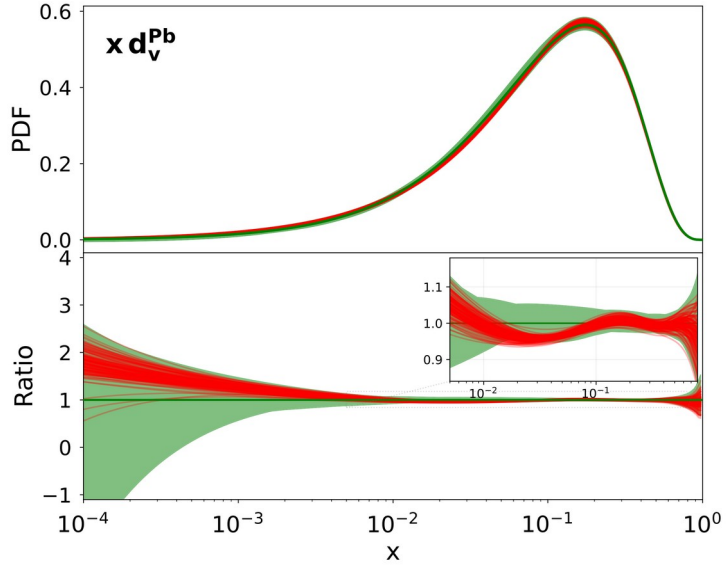
Secondary structure in d-valence

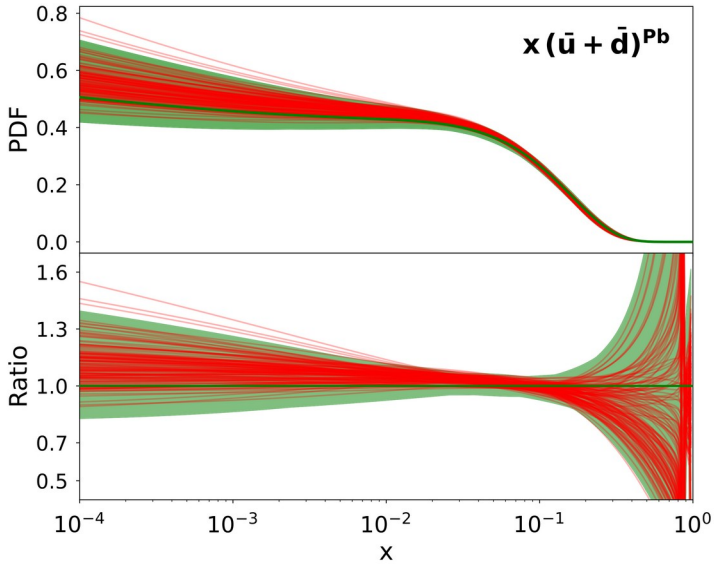
- Red samples ($a_3^{d_v} > 0.8$) populate a distinct region of the d-valence subset.
- MCMC includes these points in the posterior uncertainty; Hessian would mostly miss them.
- Important question: **does this alternative configuration produce different PDFs?**

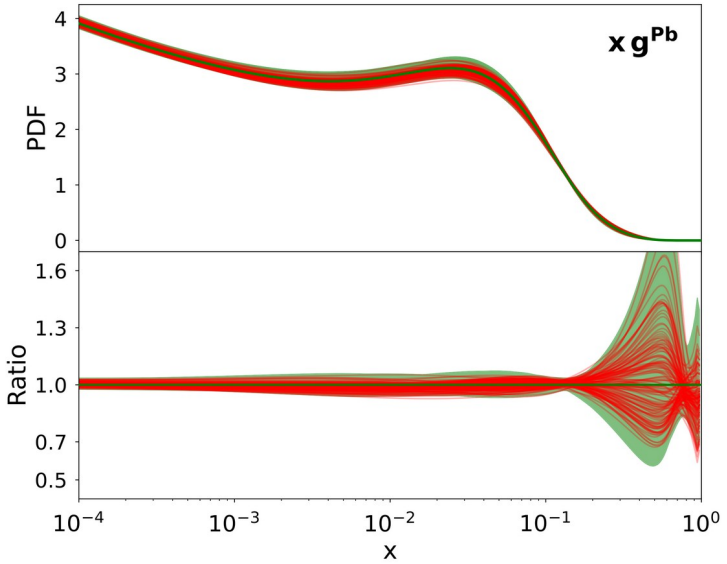


Secondary structure in d-valence



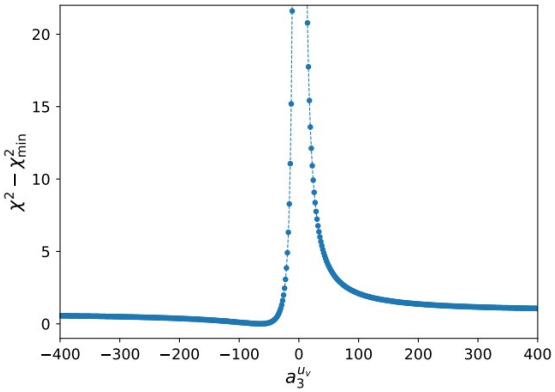


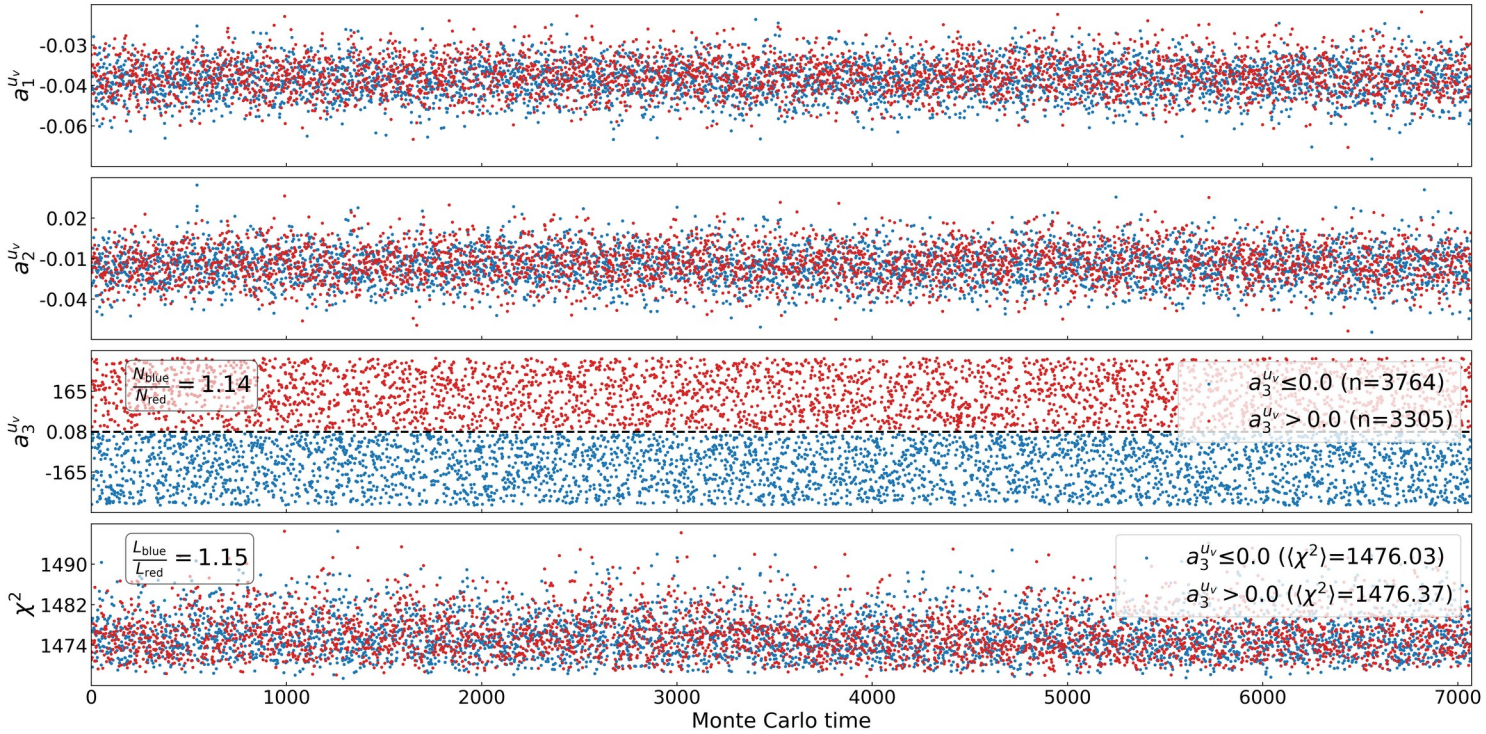




Secondary minima in u-valence

- $a_3^{u_v}$ separates into positive and negative regions.
- Values near $a_3^{u_v} \approx 0$ are strongly disfavored.
- This gives stronger evidence for separated posterior regions.
- The effect mainly appears in low-x valence PDFs.



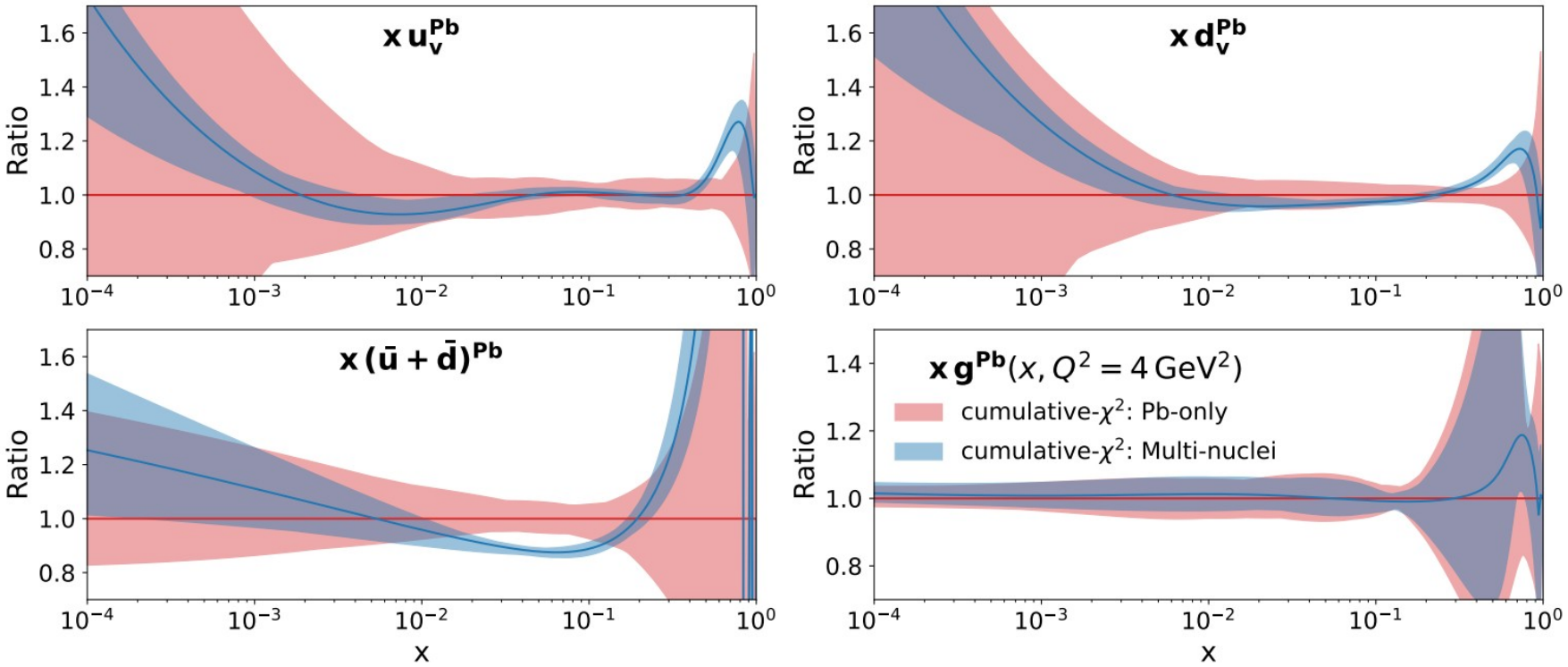


Results: Multi-nuclei analysis

Multi-nuclei fit: role of A-dependence

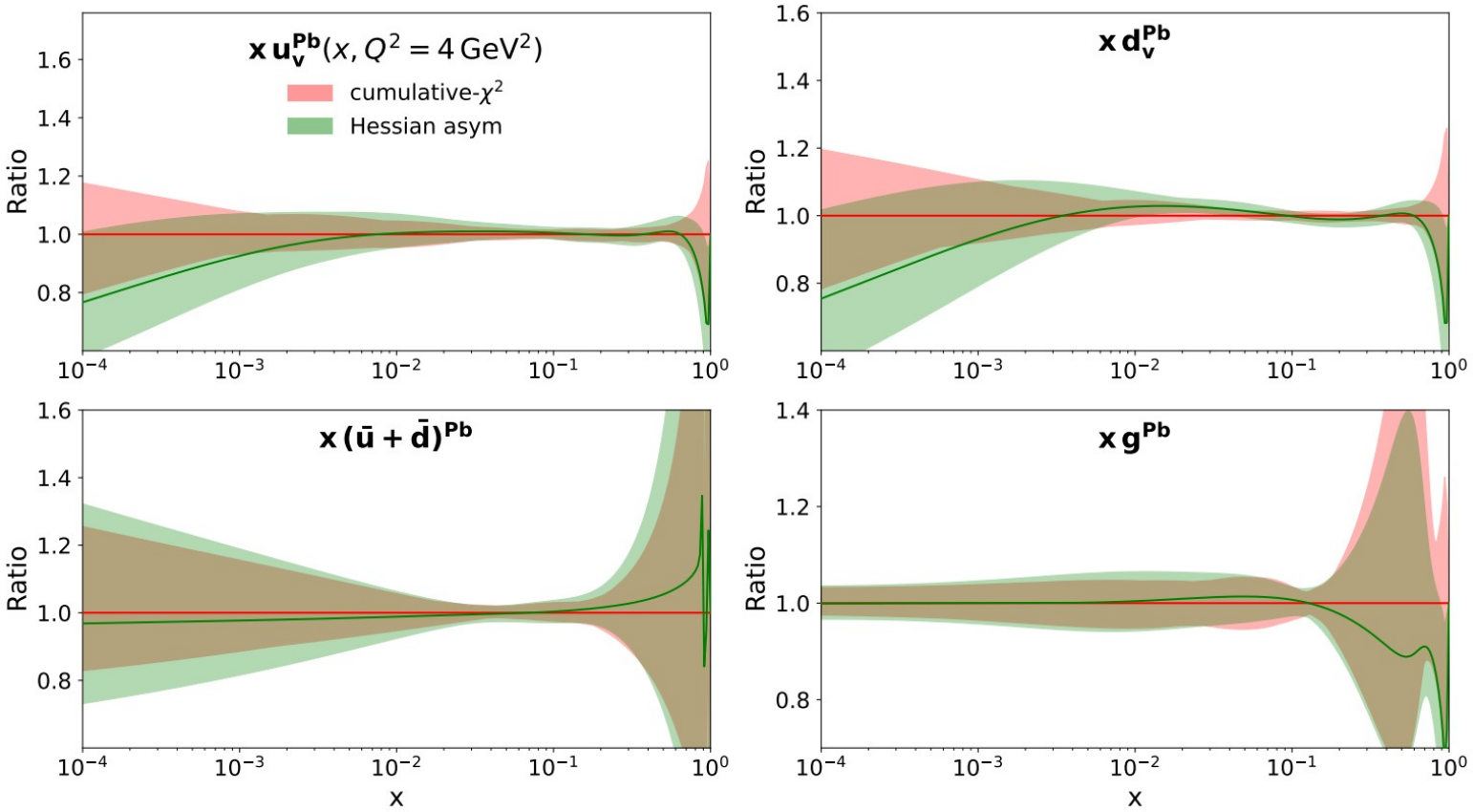
- Multi-nuclei fit: same framework, but includes data from different nuclei.
- Additional nuclei constrain Pb PDFs indirectly through the A-dependence parameterization.

$$c_j(A) = p_j + a_j \ln A + b_j \ln^2 A$$



Multi-nuclei fit: MCMC vs Hessian

- Agreement between MCMC and Hessian improves compared to the Pb-only fit.
- Remaining differences are still driven mainly by the valence sector.
- MCMC still captures residual non-Gaussian correlations that Hessian misses.



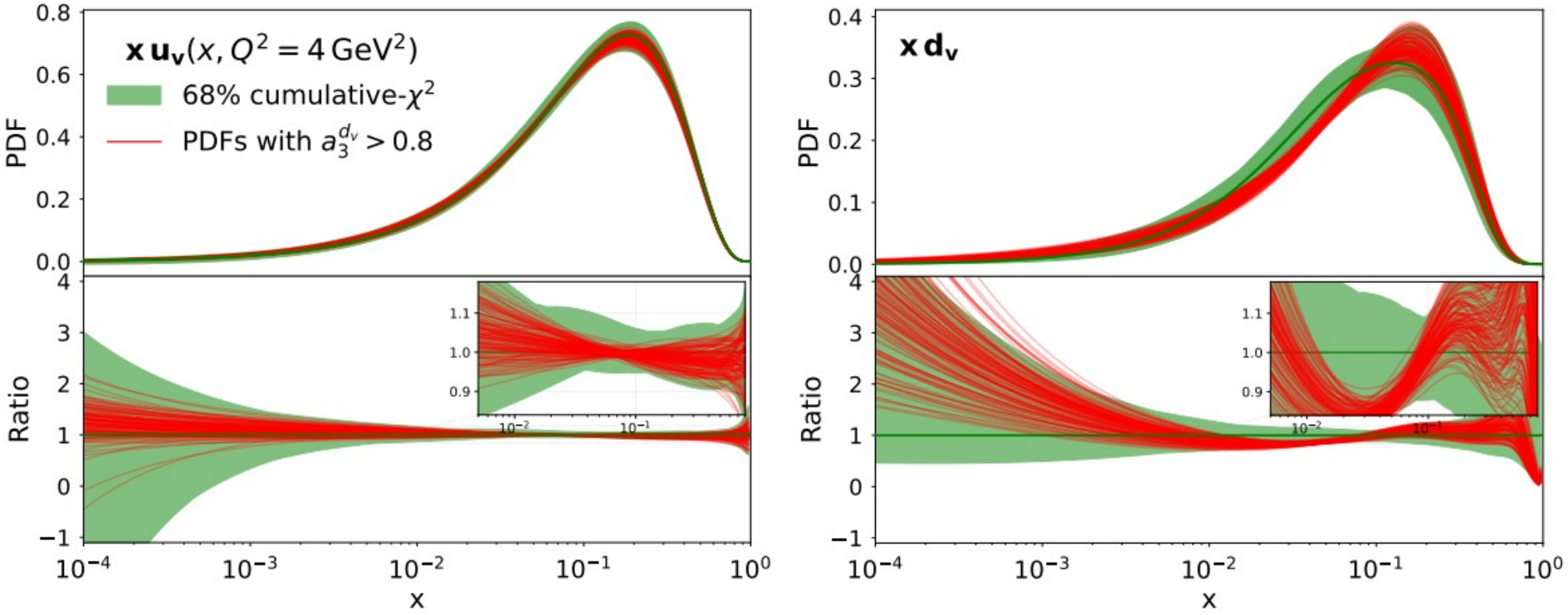
Conclusion

- Developed a fully Bayesian framework for nPDF determination using MCMC sampling
- Performed two complementary analyses: **Pb-only** and **multi-nuclei** fits
- Found good agreement with Hessian uncertainties for well-constrained Gaussian parameters
- Observed significant differences for weakly constrained and non-Gaussian parameters
- First estimate of systematic uncertainty from nuclear A-dependence assumptions

Key Takeaway:

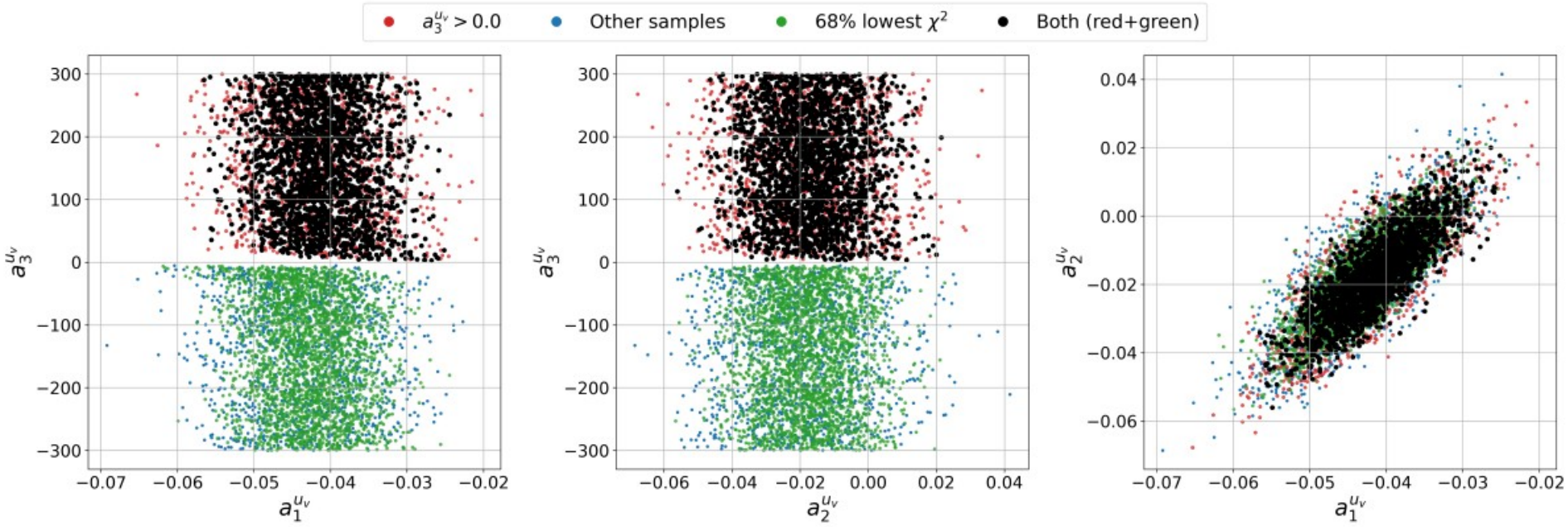
MCMC provides the full posterior distribution without Gaussian assumptions, but at a higher computational cost. For sufficiently constrained datasets, the Hessian method still remains an efficient and reliable approximation.

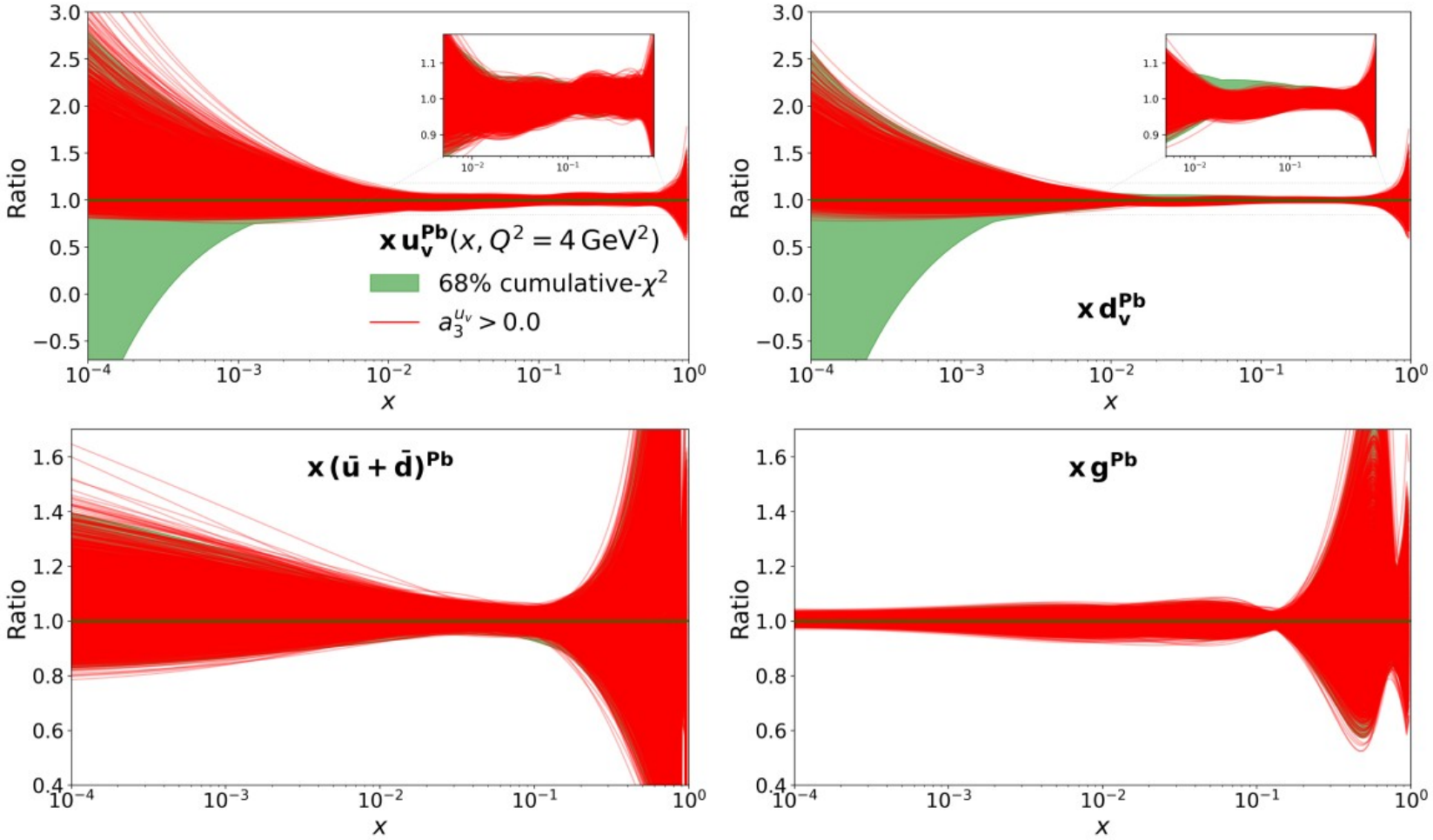
Backup



(b) Bound-proton Pb PDFs

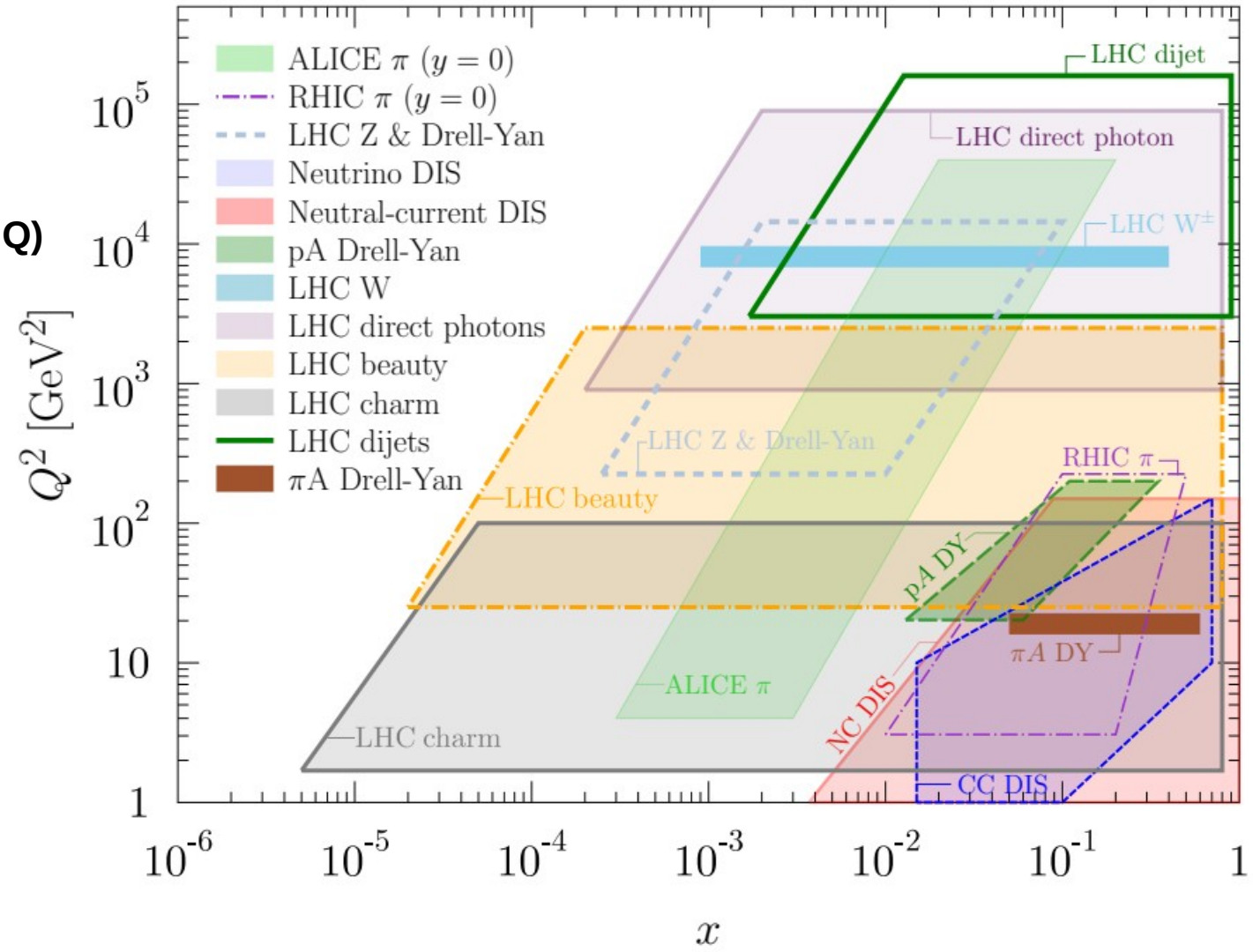
$\chi^2_{\min}$	Pb-only	Multi-nuclei
Hessian (per d.o.f.)	1466.57 (0.985)	2056.10 (1.035)
MCMC (per d.o.f.)	1467.30 (0.992)	2051.82 (1.028)




$\mathbf{x\,u_v^{Pb}}(x, Q^2 = 4\,\mathrm{GeV}^2)$
68% cumulative-χ^2
$a_3^{u_v} > 0.0$
$\mathbf{x\,d_v^{Pb}}$
$\mathbf{x\,(\bar{u} + \bar{d})^{Pb}}$
$\mathbf{x\,g^{Pb}}$
Ratio
x

Experimental data:

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

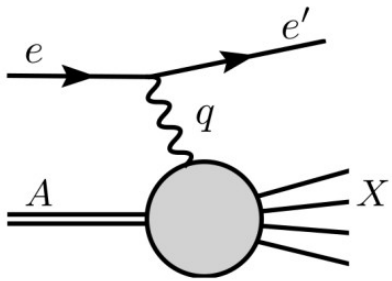


DIS variables for **nucleus**

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

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

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

$x_N \in (0, A)$

Sum rules:

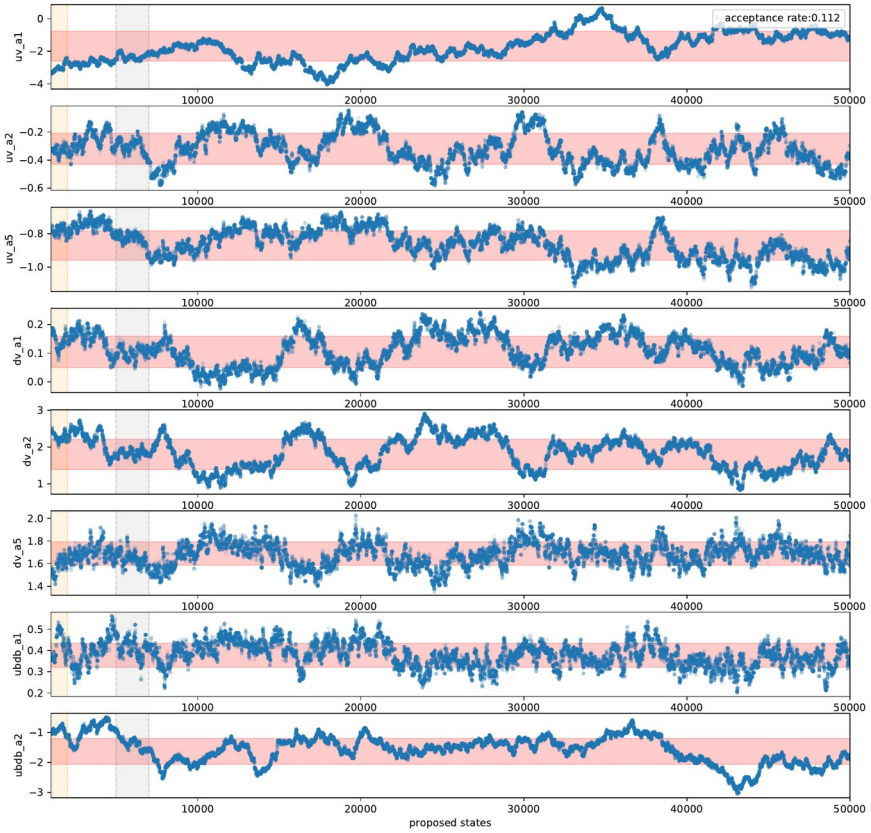
$$\int_0^1 \mathrm{d}x_A \; \tilde{u}_v^A(x_A, Q^2) = 2Z + N \,,$$
$$\int_0^1 \mathrm{d}x_A \; \tilde{d}_v^A(x_A, Q^2) = Z + 2N \,,$$

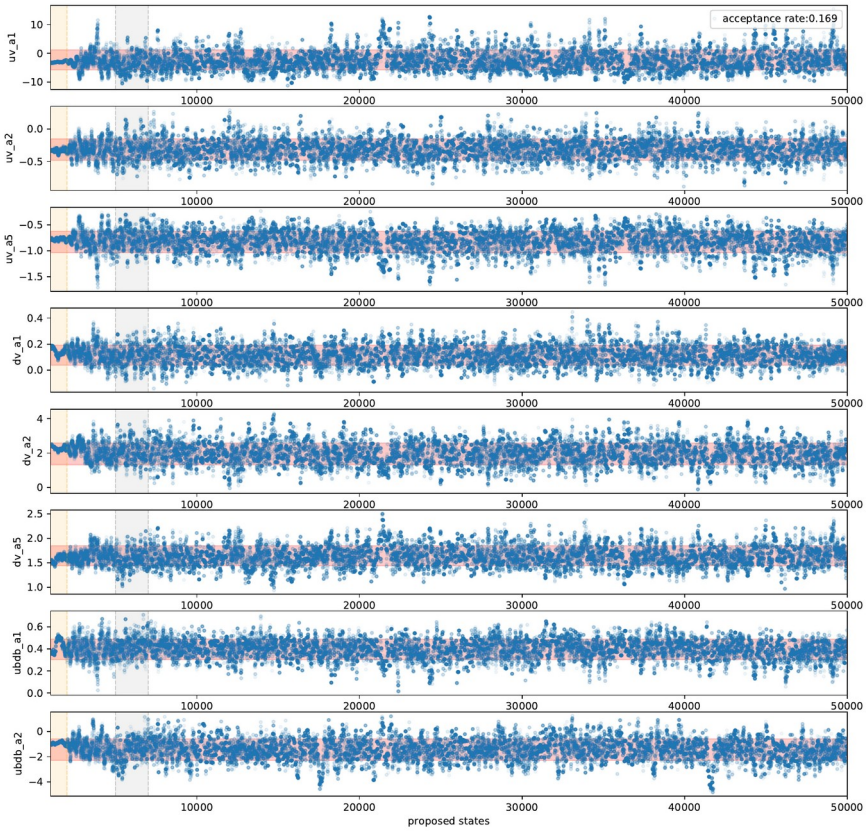
and the momentum sum rule

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

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

MH vs adaptive MH

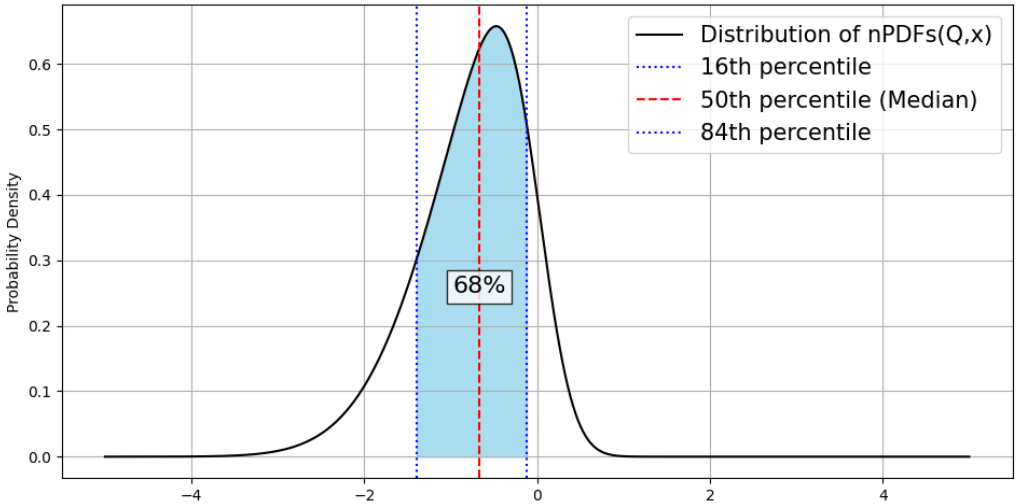


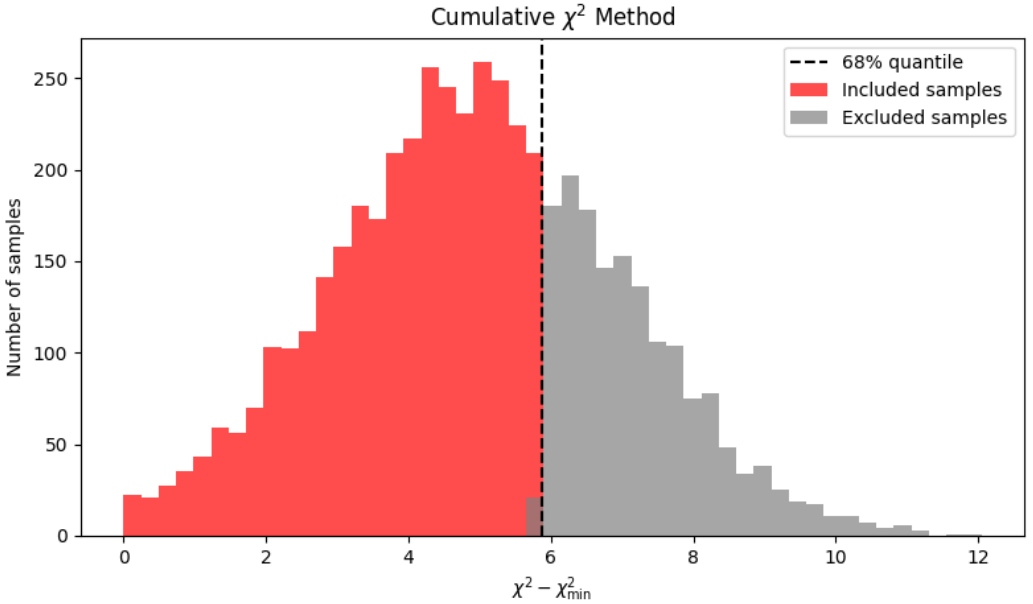


nPDFs uncertainties:

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

- **Cumulative χ^2** [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





Methodology:

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

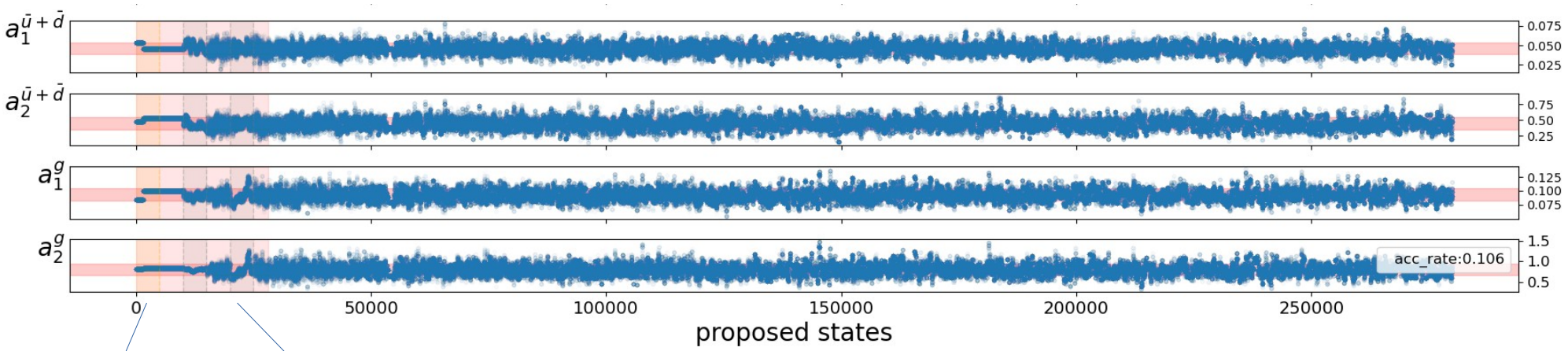
- **Removing Burn-In Phase**
  Discard the initial segment of each chain, known as the burn-in or thermalization phase, which represents the period before the chain converges to the target distribution

- **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)

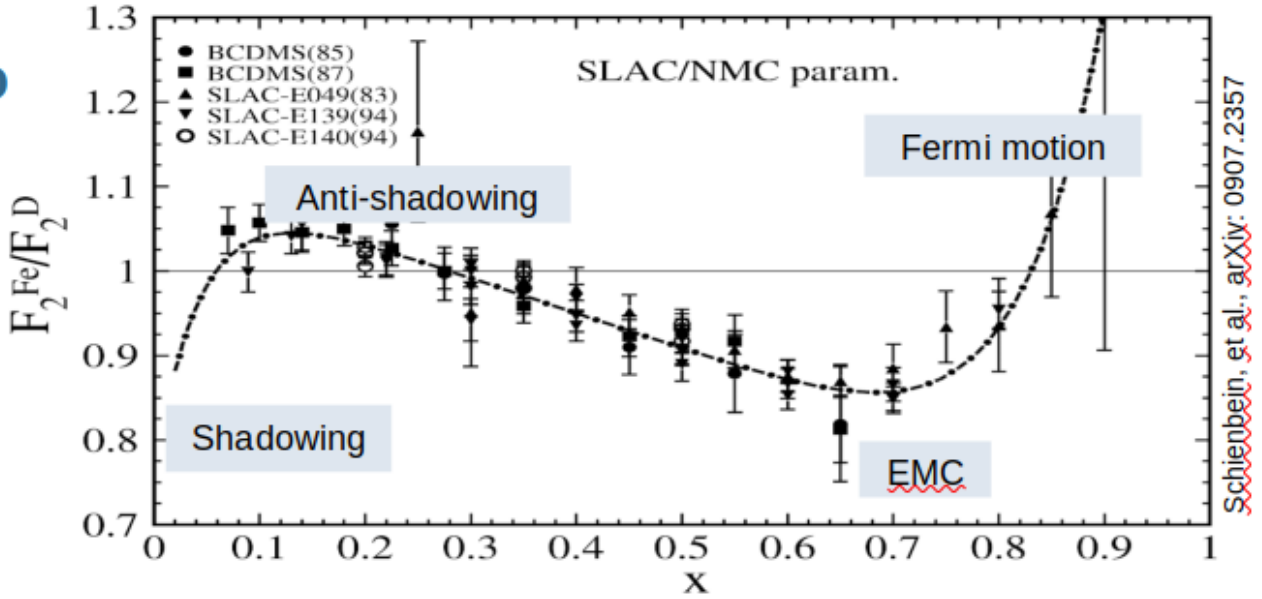


N_0 = 5000

Restarting the chain at 10,000 and 20,000

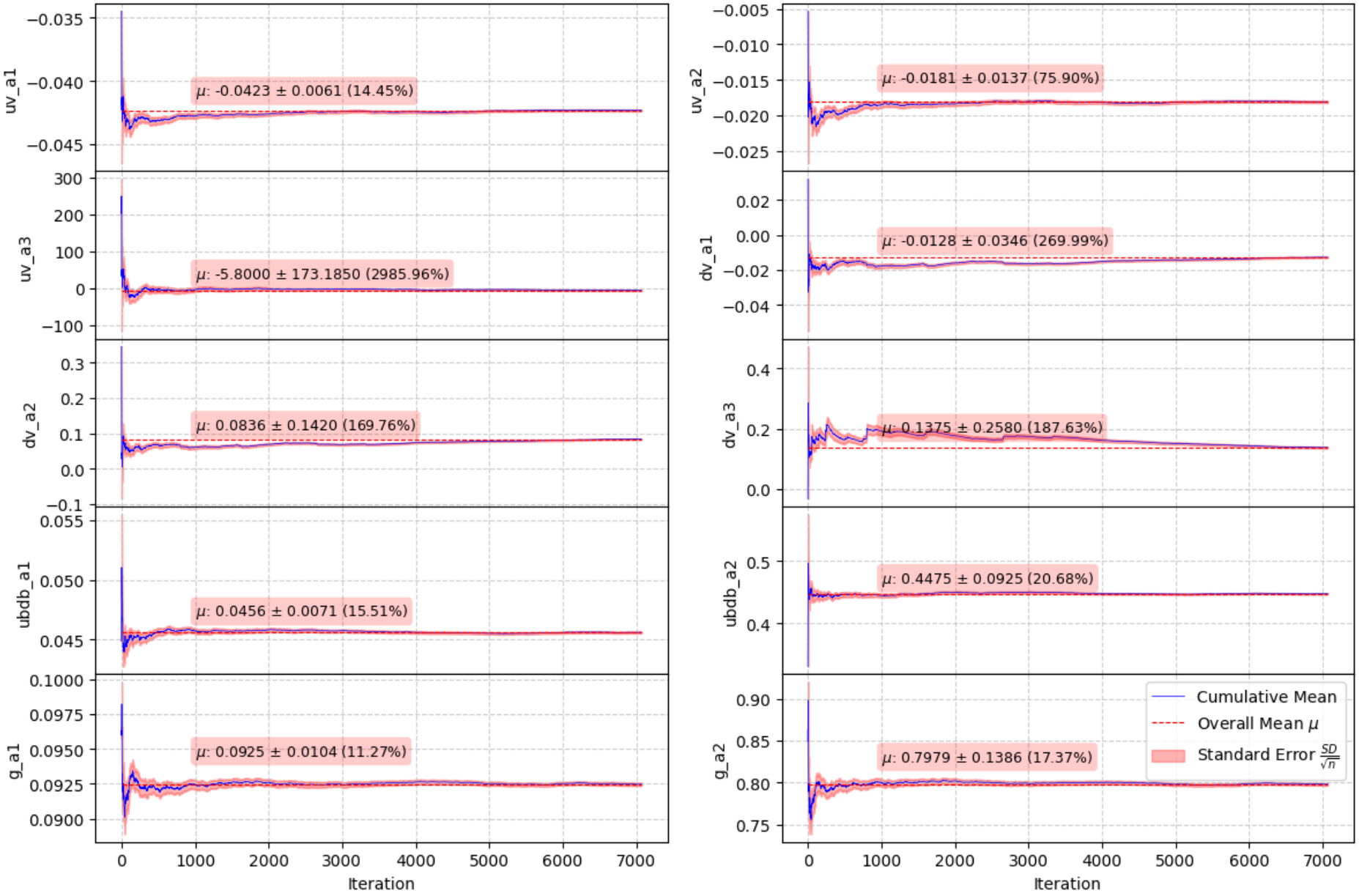
Starting point: global minimum from Hessian fit + Gaussian noise (width= 20 % of minimum value)
Thermalization (burn-in phase): removing first 8000 accepted points

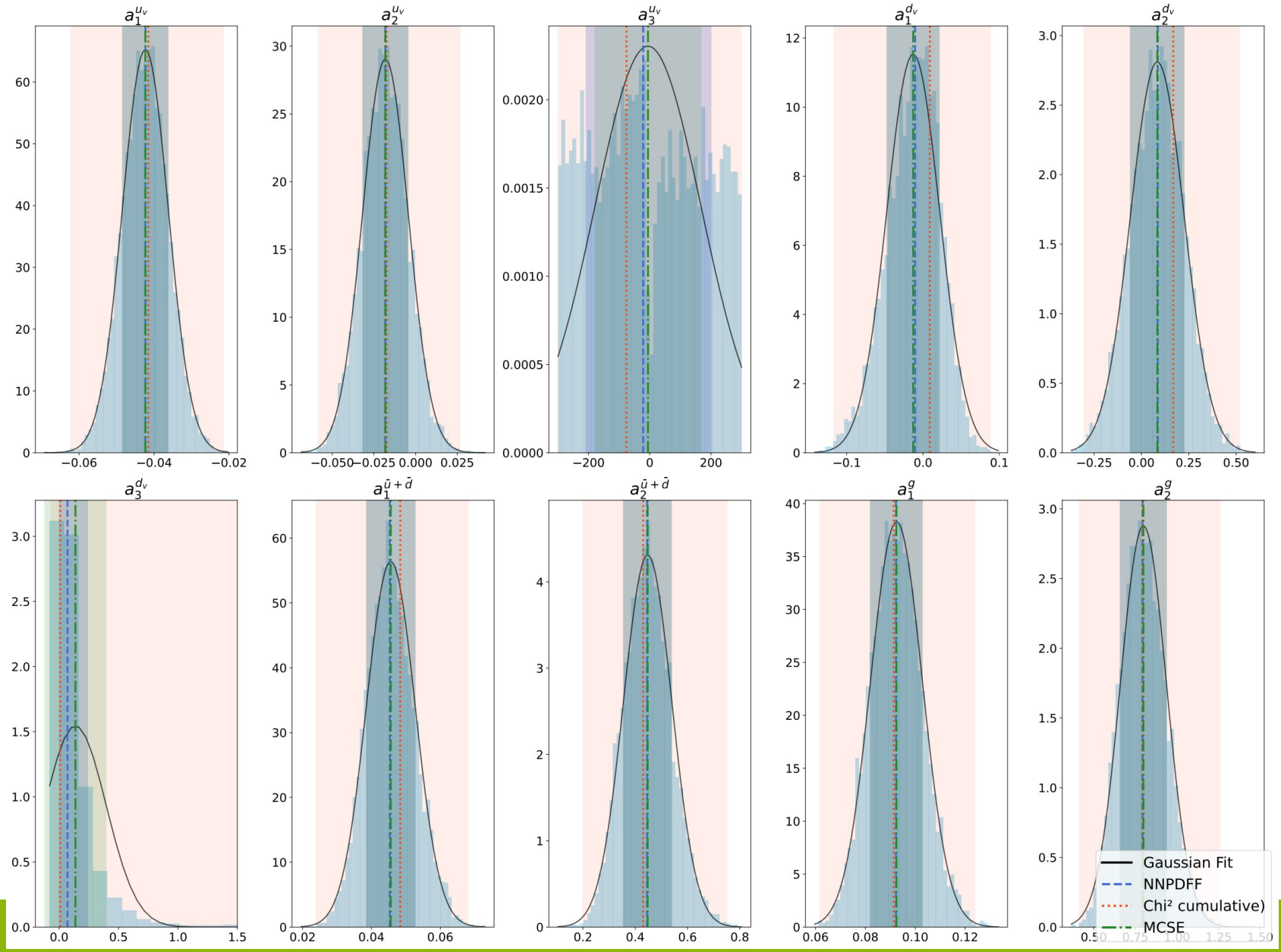
Nuclear correction ratio



- **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!


uv_a1
μ: -0.0423 ± 0.0061 (14.45%)
uv_a3
μ: -5.8000 ± 173.1850 (2985.96%)
dv_a2
μ: 0.0836 ± 0.1420 (169.76%)
ubdb_a1
μ: 0.0456 ± 0.0071 (15.51%)
g_a1
μ: 0.0925 ± 0.0104 (11.27%)
uv_a2
μ: -0.0181 ± 0.0137 (75.90%)
dv_a1
μ: -0.0128 ± 0.0346 (269.99%)
dv_a3
μ: 0.1375 ± 0.2580 (187.63%)
ubdb_a2
μ: 0.4475 ± 0.0925 (20.68%)
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