

Reconstructing hadron structure using AI/ML

Wally Melnitchouk

Jefferson Lab

— JAM Collaboration —



$$\mathcal{L}_{\text{QCD}} = \sum_q \bar{\psi}_q (i\gamma_\mu D^\mu - m_q) \psi_q - \frac{1}{2} \text{Tr}[G_{\mu\nu} G^{\mu\nu}]$$

<http://www.jlab.org/jam>

The central inference problem

Forward map

PDF shape and nuisance parameters enter factorized cross sections,

$$\theta \longmapsto t(\theta)$$

Evolution, convolutions, cuts, and correlated systematics make each evaluation expensive.

Inverse problem

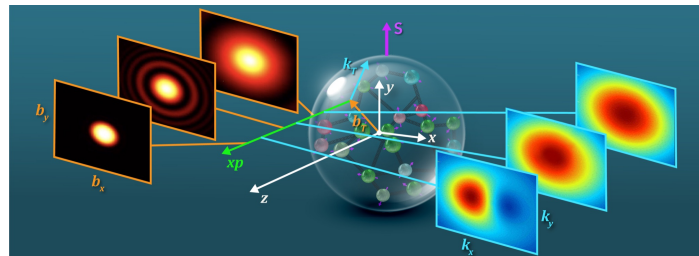
Global analysis reconstructs a high-dimensional posterior,

$$p(\theta \mid \text{data}) \propto \exp\left[-\frac{1}{2}\chi^2(\theta)\right] \pi(\theta)$$

Result must represent correlations, degeneracies, and weakly constrained directions



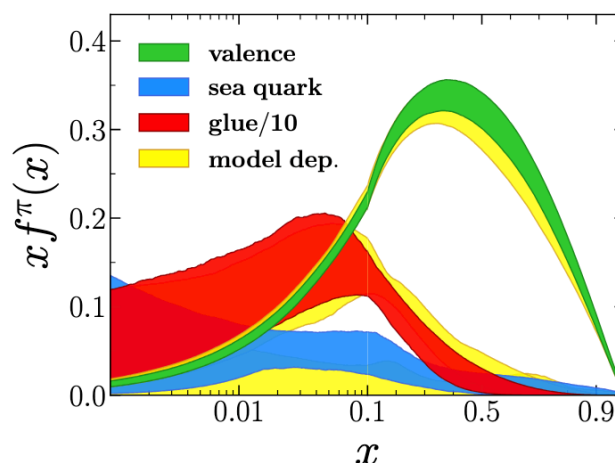
AI/ML can accelerate posterior exploration, relax assumptions about functional forms



Outline

■ Traditional approach — parametrizations, MC ensembles

→ *e.g.*, JAM global
QCD analysis



■ Beyond ensemble replicas — normalizing flows

■ Beyond parametric forms — neural networks, pixels / finite elements

→ Bayesian pixel approach to TMDs

→ *Poonam Choudhary*
Thursday 3:50pm

Traditional approach to global analysis

- Use fixed parametric form for input scale PDFs
- Generate data replicas by shuffling within experimental uncertainties (“data resampling”), use ensemble approach to build posteriors
- Minimize χ^2 for every replica
- Multistep strategy to minimize cost of χ^2 evaluations
- Propagate accepted ensemble to observables

Assumption: fitted ensembles approximate samples from the Bayesian posterior

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$



Traditional approach to global analysis

■ Likelihood function

$$\mathcal{L}(\text{data} \mid \theta) = \exp \left[-\frac{1}{2} \chi^2(\theta) \right]$$

with χ^2 defined as

$$\chi^2 = \sum_e \left\{ \sum_i \left[\frac{d_i^e - t_i^e(\theta)/n_e - \sum_k r_k^e \beta_{ki}^e}{\alpha_i^e} \right]^2 + \sum_k (r_k^e)^2 + \left(\frac{n_e - 1}{\delta n_e} \right)^2 \right\}$$

where data d_i^e and theory t_i^e for experiment e and data point i ,
with α_i^e uncorrelated uncertainties, β_{ki}^e effective correlated-error coefficients,
 n_e normalization factors with uncertainties δn_e , and r_k^e correlated shifts

→ 2nd & 3rd terms penalize departures from nominal systematic shifts & normalizations

■ Z-score metric $Z = \sqrt{2} \operatorname{erf}^{-1}(2p - 1)$, where p is Pearson's χ^2 statistic

→ reduced χ^2 does not tell entire story — number of data points also relevant

→ Z-score gives statistical significance (in units of normal distribution)

Traditional approach to global analysis

- A common template for the functional form at the input scale is

$$T(x; \theta) = N x^\alpha (1 - x)^\beta P(x)$$



$P(x) = (1 + \gamma\sqrt{x} + \eta x)$, Bernstein polynomials, neural net, ...

- Advantages:

- compact and easily interpretable
- sum rules are straightforward
- stable with sparse data

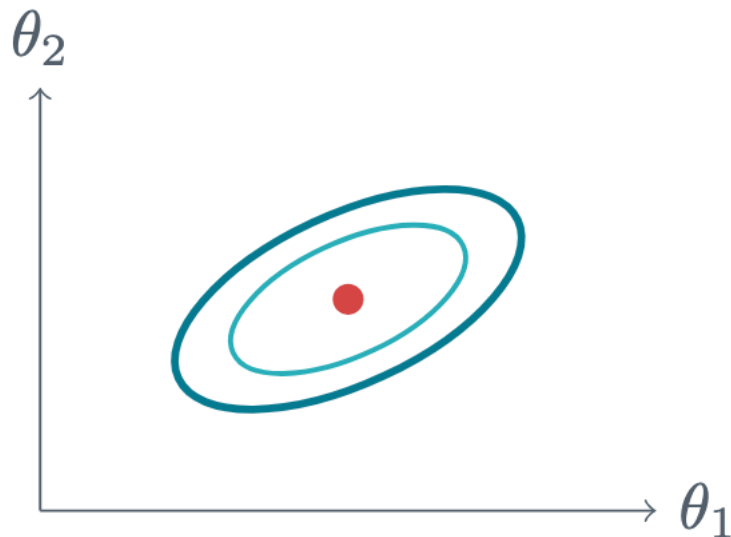
- Disadvantages:

- potential functional bias
- uncertainties depend on allowed shapes
- flexible forms can enter unphysical regions

Hessian vs. ensembles

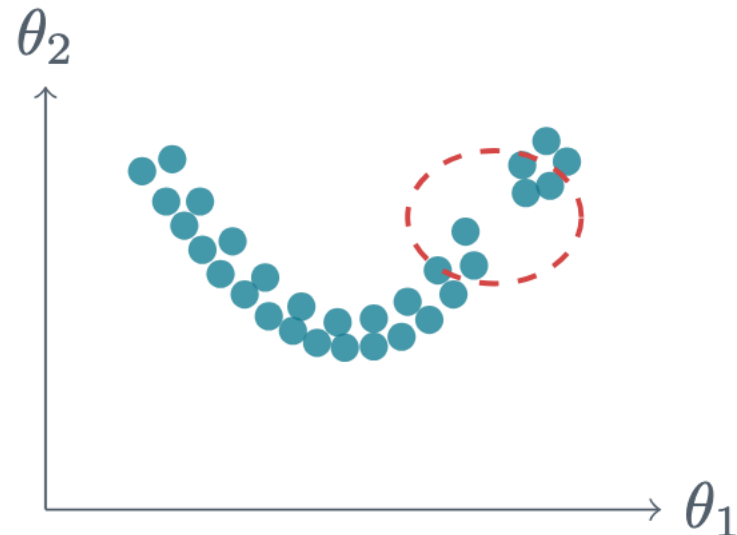
Hessian approximation

Gives maximum likelihood, explores potential solutions only around that maximum



Replica ensemble

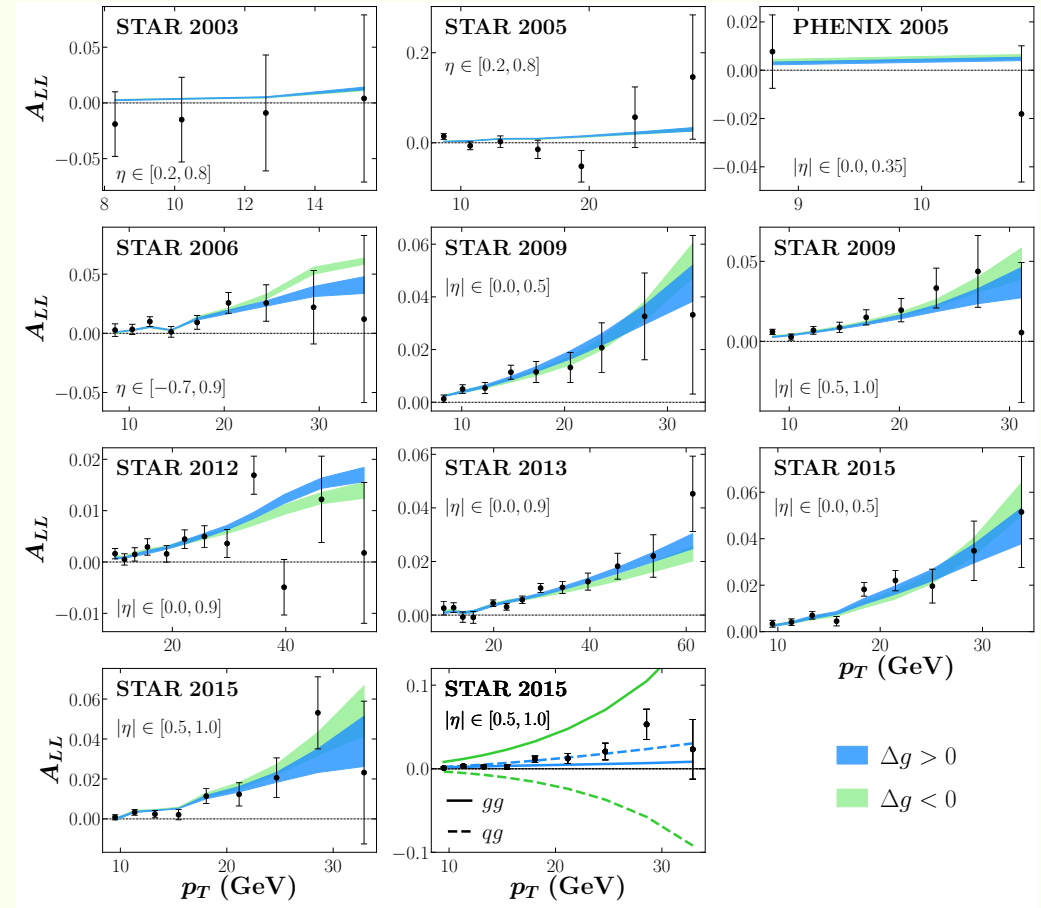
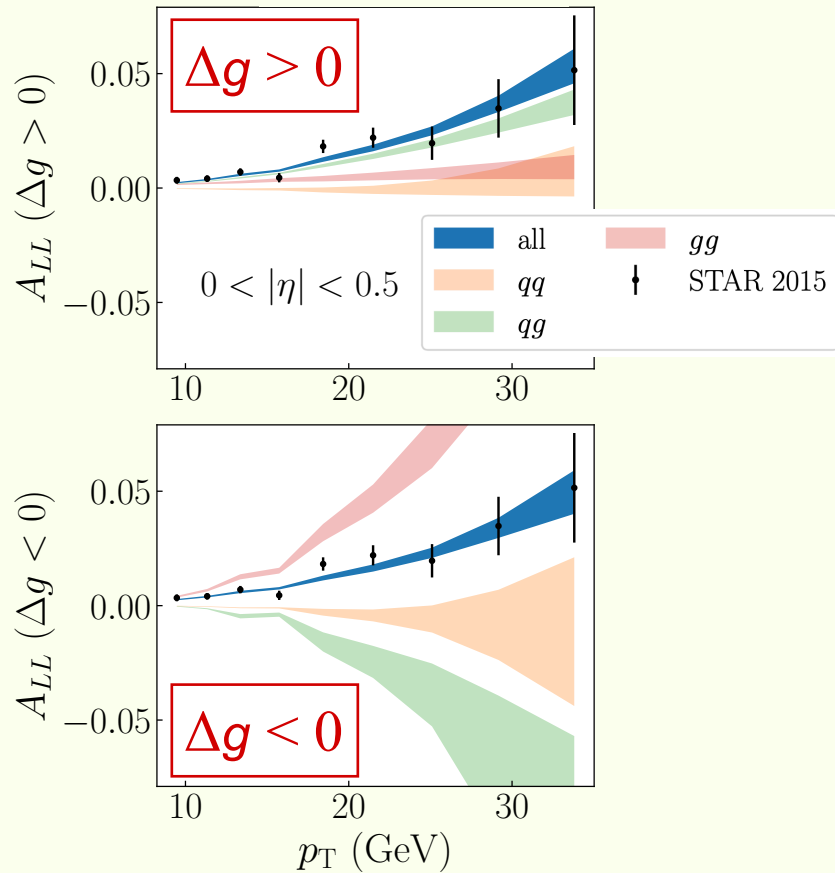
Can retain skewness, nonlinear correlations, boundaries, and multiple admissible regions



- global fits often contain directions where positivity, sum rules, or sparse data make the posterior strongly *non-Gaussian*
- explore more thoroughly parameter space to find all solutions admissible by the data

Real world example: gluon polarization in the proton

Double longitudinal polarization asymmetry for inclusive jet production



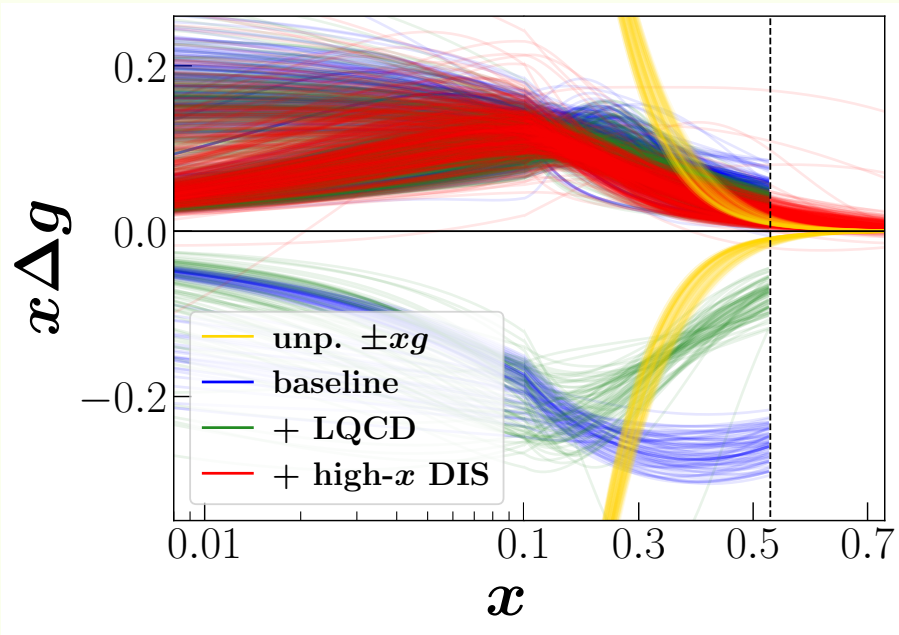
→ significant cancellations between (positive) $\Delta g \otimes \Delta g$ and (negative) $\Delta q \otimes \Delta g$ terms

→ positive and negative Δg solutions describe jet data equally well

→ negative Δg also implies negative $\Delta \Sigma$

Real world example: gluon polarization in the proton

- Need additional constraints from lattice QCD and high- x DIS data to determine the sign of gluon polarization



Reaction	$\chi^2_{\text{red}}(\Delta g > 0)$			$\chi^2_{\text{red}}(\Delta g < 0)$			N
	Baseline	+LQCD	+ high- x DIS	Baseline	+LQCD	+ high- x DIS	
Polarized	0.89	0.90	1.18	0.92	1.06	1.24	2067
DIS	0.95	0.96	1.21	0.98	1.12	1.25	1735*
SIDIS	0.85	0.84	1.08	0.84	0.96	1.11	231
Jets	0.84	0.89	0.90	0.88	1.10	1.44	83
$W^\pm/Z$	0.60	0.60	0.99	0.83	0.84	1.32	18
Unpolarized	1.14	1.14	1.14	1.15	1.15	1.15	5954
SIA	0.86	0.86	0.89	0.90	0.90	0.92	564
LQCD	...	0.57	0.58	...	1.18	3.92	48
Total	1.08	1.10	1.13	1.10	1.12	1.17	8633

Hunt-Smith et al., *Phys. Rev. Lett.* **133**, 161901 (2024)

- including lattice QCD and high- x DIS data strongly disfavors negative $\Delta\Sigma$ solutions at $x > 0.5$
- in data-driven approach, $\Delta g < 0$ ruled out only with inclusion of polarized jet, lattice, and high- x DIS data

Sampling full χ^2 landscape

Maximum-likelihood analysis

χ^2 minimum identifies one preferred solution in parameter space

→ a local approximation can miss asymmetric uncertainties, curved correlations, and separated solutions

Bayesian analysis

Markov-chain samples represent the full probability distribution over the model degrees of freedom,

$$p(\boldsymbol{\theta} \mid \text{data}) \propto \exp\left[-\frac{1}{2}\chi^2(\boldsymbol{\theta})\right] \pi(\boldsymbol{\theta})$$

- Generate samples from full posterior with Markov chain Monte Carlo (MCMC)
 - capture non-Gaussian uncertainties & correlations
 - cost of MCMC is prohibitive, however, for real world applications
 - use AI/ML tools such as *normalizing flow* for practical global QCD analysis

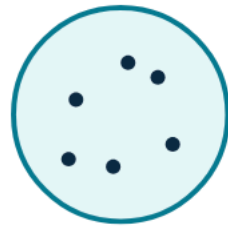
Normalizing flow

- A *normalizing flow* (NF) is a trainable, reversible map that *transforms* a simple probability distribution into a complicated one

Reference distribution

Start with a simple random input drawn from a reference distribution,

$$z \sim p_0(z)$$



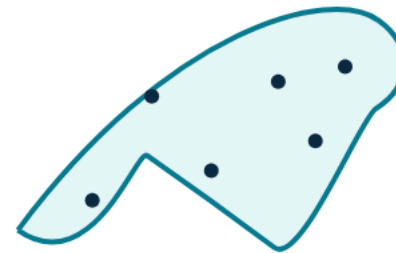
Gaussian reference

f_ϕ

Invertible transformation

A trainable, invertible map sends latent variables into QCD parameter space,

$$\theta = f_\phi(z)$$



target posterior

→ NF learns coordinates in which a complex target distribution can be represented and sampled efficiently

Normalizing flow

- For every parameter point, the inverse map identifies its latent point,

$$\mathbf{z} = f_{\phi}^{-1}(\boldsymbol{\theta})$$

and changing variables leads to the parameter probability density,

$$q_{\phi}(\boldsymbol{\theta}) = p_0\left(f_{\phi}^{-1}(\boldsymbol{\theta})\right) \left| \det \frac{\partial f_{\phi}^{-1}}{\partial \boldsymbol{\theta}} \right|$$


Jacobian corrects probability density for local stretching or compression under the transformation such that the total probability is preserved

$$\int q_{\phi}(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int p_0(\mathbf{z}) d\mathbf{z} = 1.$$

→ NF can both generate samples and evaluate their probability density

Normalizing flow

- Training changes ϕ to minimize Kullback-Leibler divergence (measure of mismatch)

$$\mathbb{E}_{\theta \sim q_\phi} \left[\log \frac{q_\phi(\theta)}{p(\theta \mid \text{data})} \right]$$

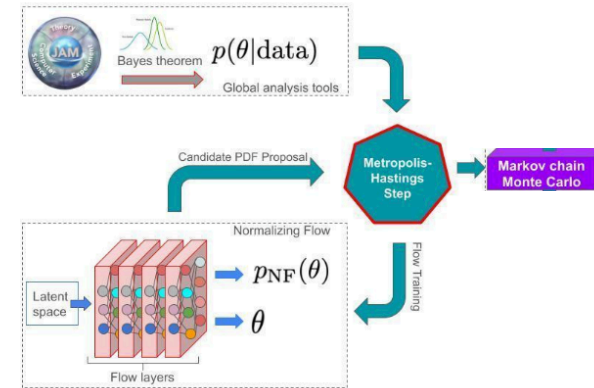
- reaches zero when NF density $q_\phi(\theta)$ matches target posterior $p(\theta \mid \text{data})$
- NF learns joint probability distribution of all parameters θ and their correlations (incl. weakly constrained directions, separated solutions)
- NF sampling allows to switch from serialized to parallelized MCMC

- Note that the trained flow is ultimately an approximation

- Metropolis-Hastings corrects the errors, with each proposal tested using exact posterior
- accepting/rejecting proposals removes bias from learned approximation
- NF accelerates sampling; M-H step guarantees correct posterior

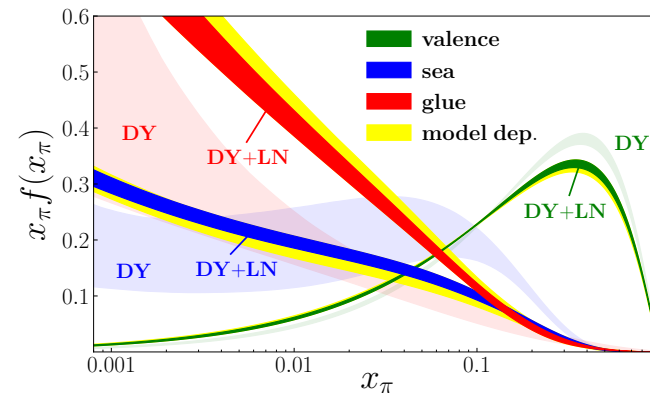
Application to pion PDFs

- As a demonstration, apply the workflow to the (simpler) case of pion PDFs



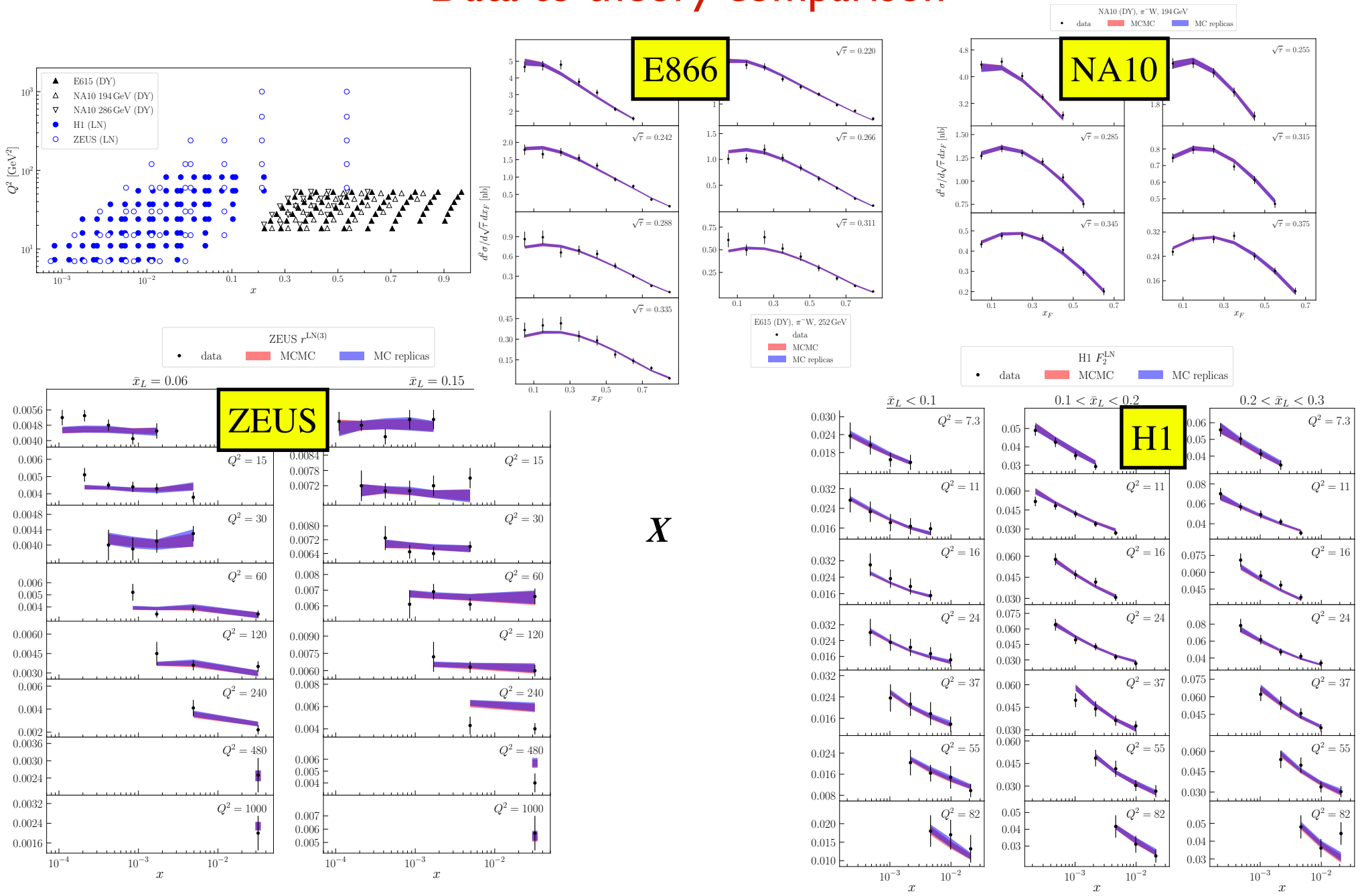
- Pion PDFs constrained by π -A Drell-Yan data (E615, NA10) at intermediate & large x and by leading-neutron electroproduction data (H1, ZEUS) at small x (Sullivan process)
 - JAM Bayesian analysis using data resampling
 - fit valence, sea-quark, gluon PDFs, data normalizations, and chiral splitting functions simultaneously (≈ 15 parameters total)

Process	Experiment (observables)	N_{dat}	χ^2_{dat}	n_e
DY	E615 (x_F, Q)	61	0.85	1.08
	NA10 (194 GeV) (x_F, Q)	36	0.52	0.88
	NA10 (286 GeV) (x_F, Q)	20	0.78	0.83
	E615 (Q, p_T)	34	1.08	0.83
	E615 (x_F, p_T)	49	0.85	0.50
LN	H1	58	0.38	1.26
	ZEUS	50	1.51	0.95
Total		308	0.85	



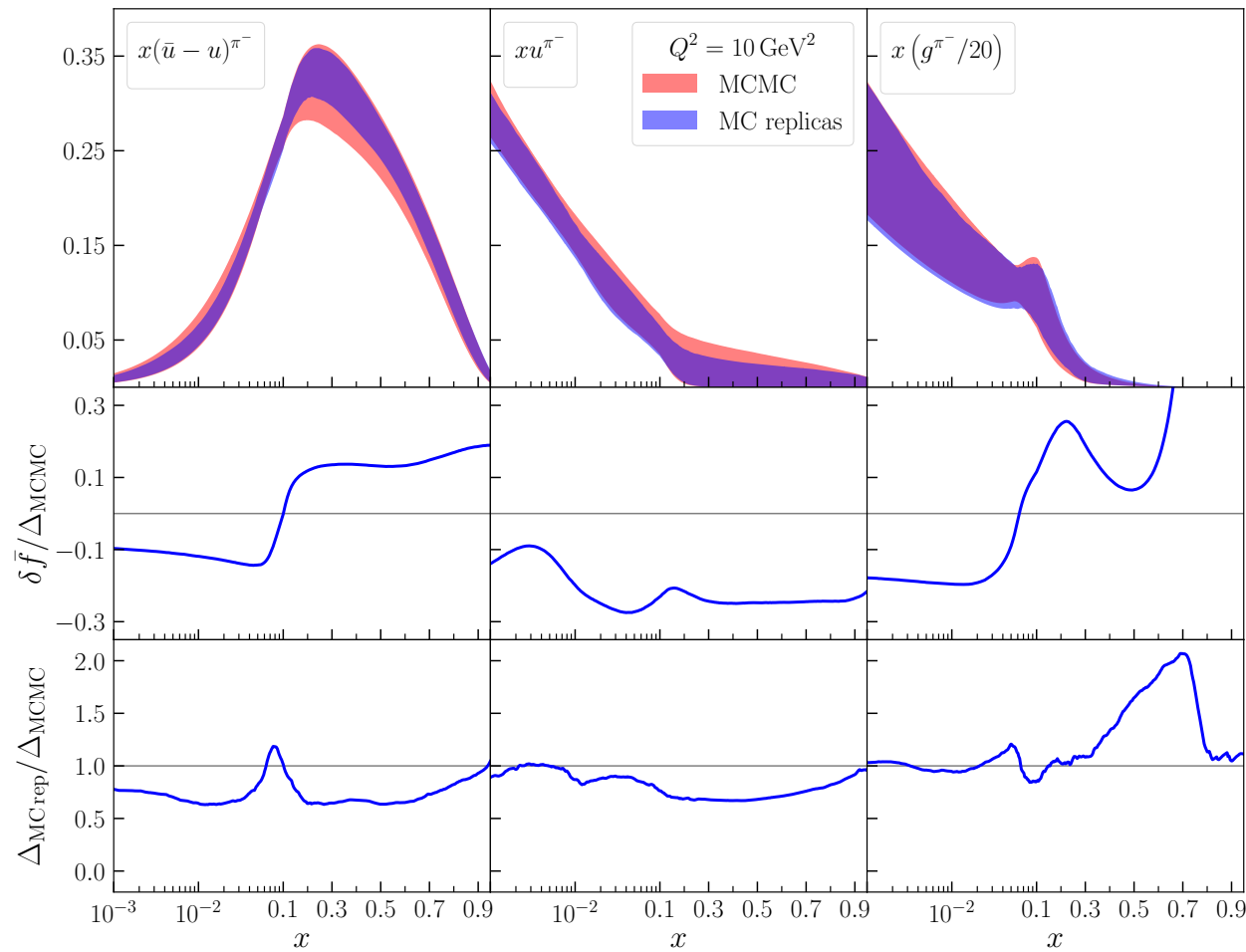
Barry, Sato, WM, Ji
Phys. Rev. Lett. **121**, 152001 (2018)

Data to theory comparison



→ good description of data for both JAM/MC replicas and MCMC+NF

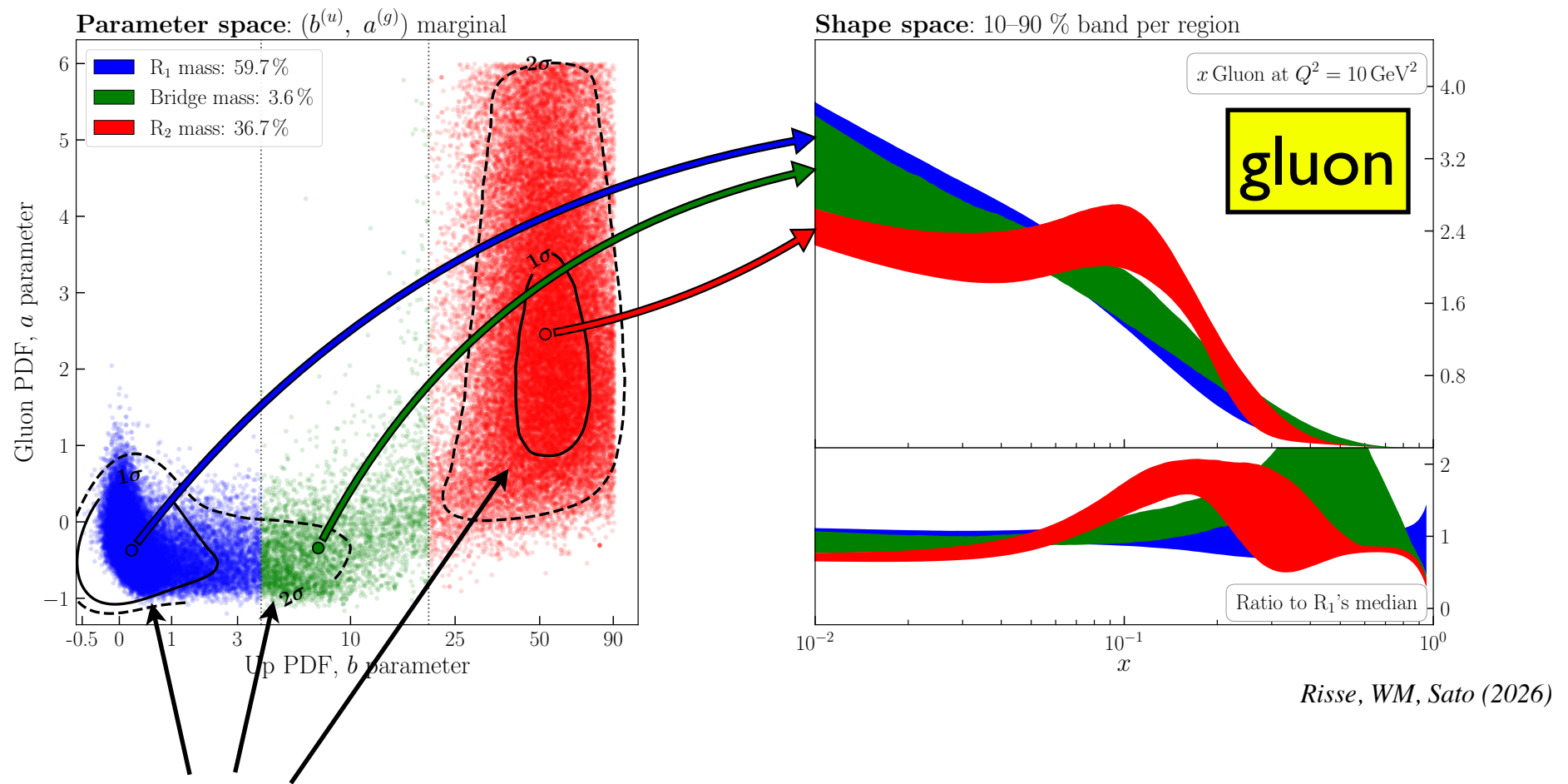
Extracted pion PDFs



Risse, WM, Sato (2026)

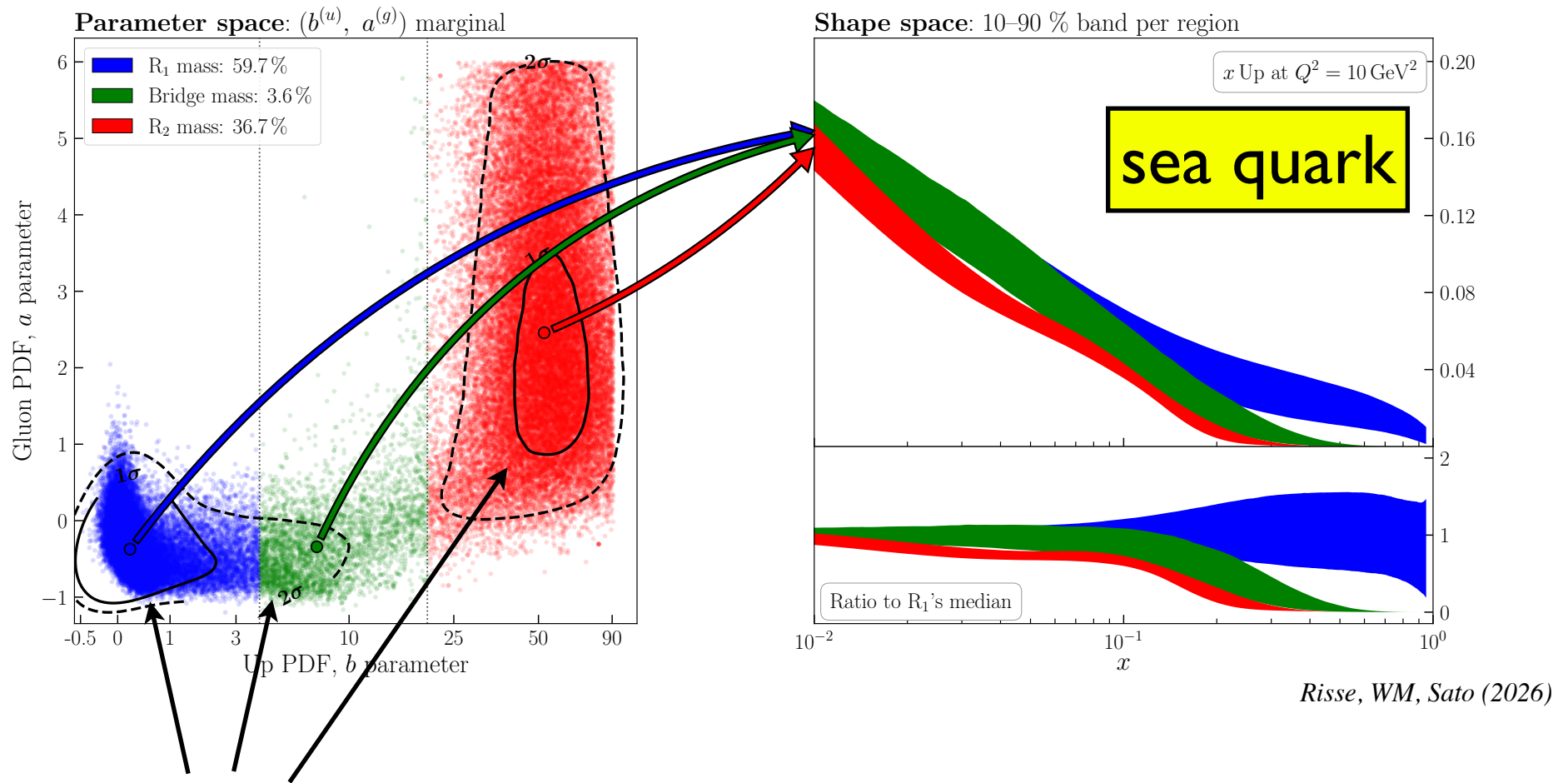
- generally similar shapes of PDFs
- MC replicas underestimate valence & sea quark PDF uncertainties by $\sim 20\% - 30\%$ compared with MCMC

Parameter correlations



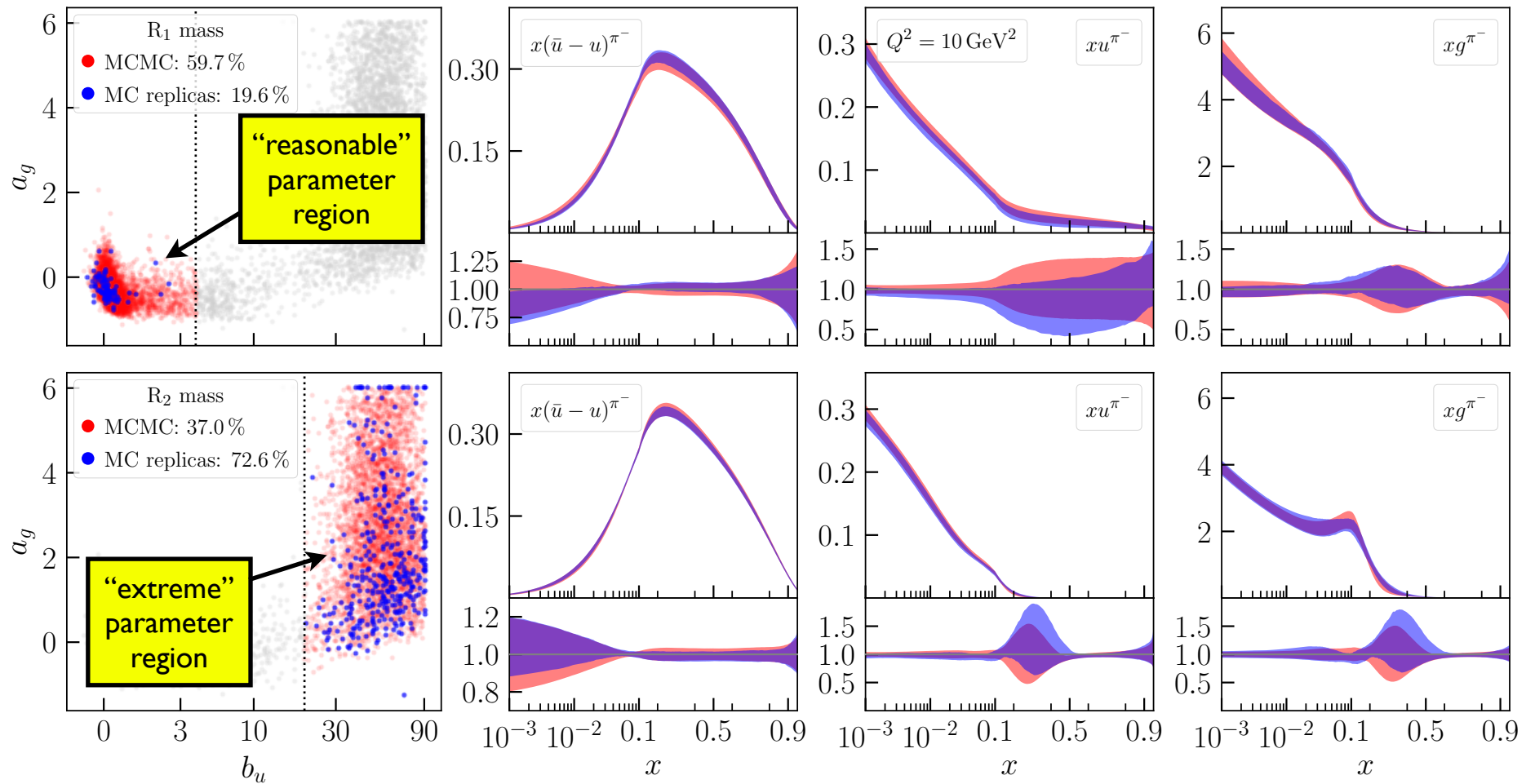
different regions of parameter space
correspond to different PDF shapes

Parameter correlations



different regions of parameter space
correspond to different PDF shapes
(Hessian approach would only find 1 minimum)

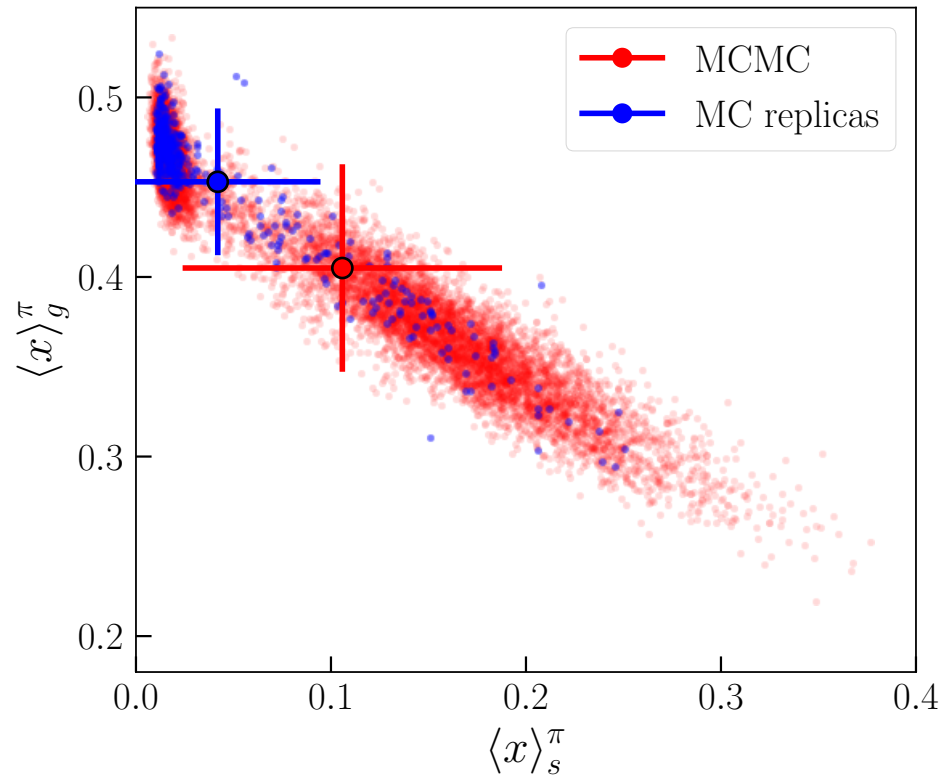
Parameter correlations



Risse, WM, Sato (2026)

→ traditional MC replica method has higher density of solutions in more “extreme” parameter region compared with MCMC+NF

Momentum fractions



$$\langle x \rangle_g^{\text{MCMC}} = 0.41(6) \quad \langle x \rangle_g^{\text{rep}} = 0.45(4)$$

$$\langle x \rangle_s^{\text{MCMC}} = 0.11(8) \quad \langle x \rangle_s^{\text{rep}} = 0.04(5)$$

Risse, WM, Sato (2026)

- momentum sum rule: valence + sea + glue = 1
- MC replicas underestimate uncertainty on sea quark momentum fraction
- next: apply to (unpolarized and polarized) proton PDFs ($\sim O(200)$ parameters)

Beyond parametric forms

- Neural networks have been proposed (*e.g.* NNPDF Collaboration) to remove parametrization bias in traditional global QCD analyses

→ replace rigid analytic form with more flexible class of functions

- **BUT** NN architecture and “preprocessing” still restrict functional space

$$T(x; \theta) = N x^\alpha (1 - x)^\beta \text{NN}(x)$$

preprocessing coefficients

- regularization and stopping rules favor some shapes over others
- where data are sparse, choices can dominate inferred PDFs
- NN can reduce parametrization bias, but does not eliminate assumptions about function shape

Beyond parametric forms

■ Pixel representation of a distribution function

→ represent unknown function by nodal values $\mathbf{h} = (h_1, \dots, h_N)$

$$f(x) = \sum_{j=1}^N h_j \varphi_j(x)$$

interpolation or finite-element
basis functions

■ Advantages

- local, transparent degrees of freedom
- resolution can follow data sensitivity
- linear constraints easy to encode

■ Requirements

- smoothness (hierarchical priors)
- positivity and sum rule constraints
- scalable posterior sampling

Beyond parametric forms

- For many observables, discretization yields

$$y = Rh + \epsilon$$

response matrix
mapping pixels
into observables

measurement noise

- response matrix represents convolution & evolution such that data constrain correlated combinations of pixels
- some parameter combinations tightly constrained, others only weakly
- local random-walk sampler explores correlated posterior very slowly
- NF learns correlated geometry of high-dimensional posterior and proposes simultaneous changes of correlated parameters

Pixel-based analysis of TMDs

■ Flexible representation

→ represent unknown TMD by $O(30)$ nodal values (instead of fixed analytic shape)

■ Bayesian target

→ evolution, convolution, data, and priors define joint posterior over all pixels and auxiliary parameters

■ AI proposal

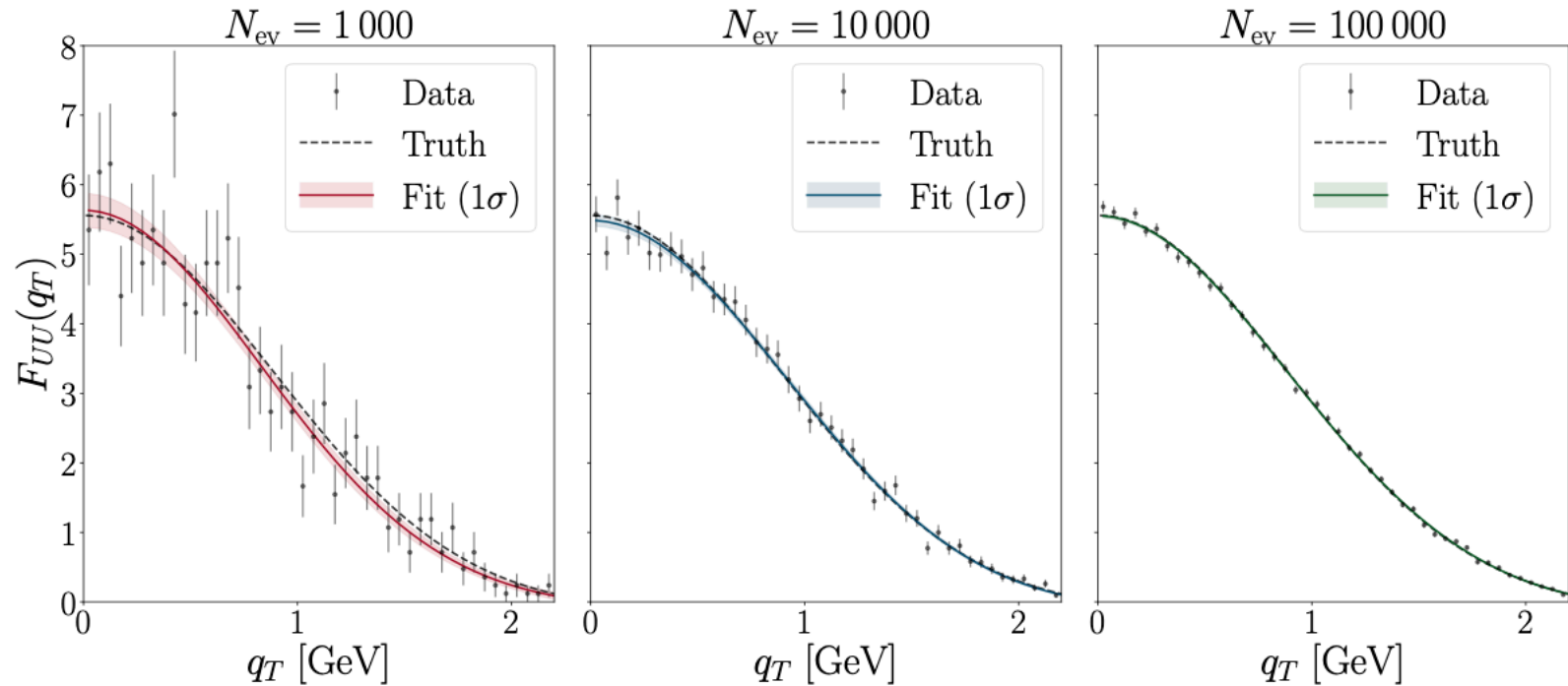
→ NF learns correlations and generates global proposals for all parameter simultaneously

■ Exact correction

→ Metropolis-Hastings accepts or rejects each proposal so final ensemble follows exact posterior

Pixel-based analysis of TMDs

- Comparison between pseudodata for SIDIS structure function $F_{UU,T}(q_T)$ and Bayesian fit in q_T space

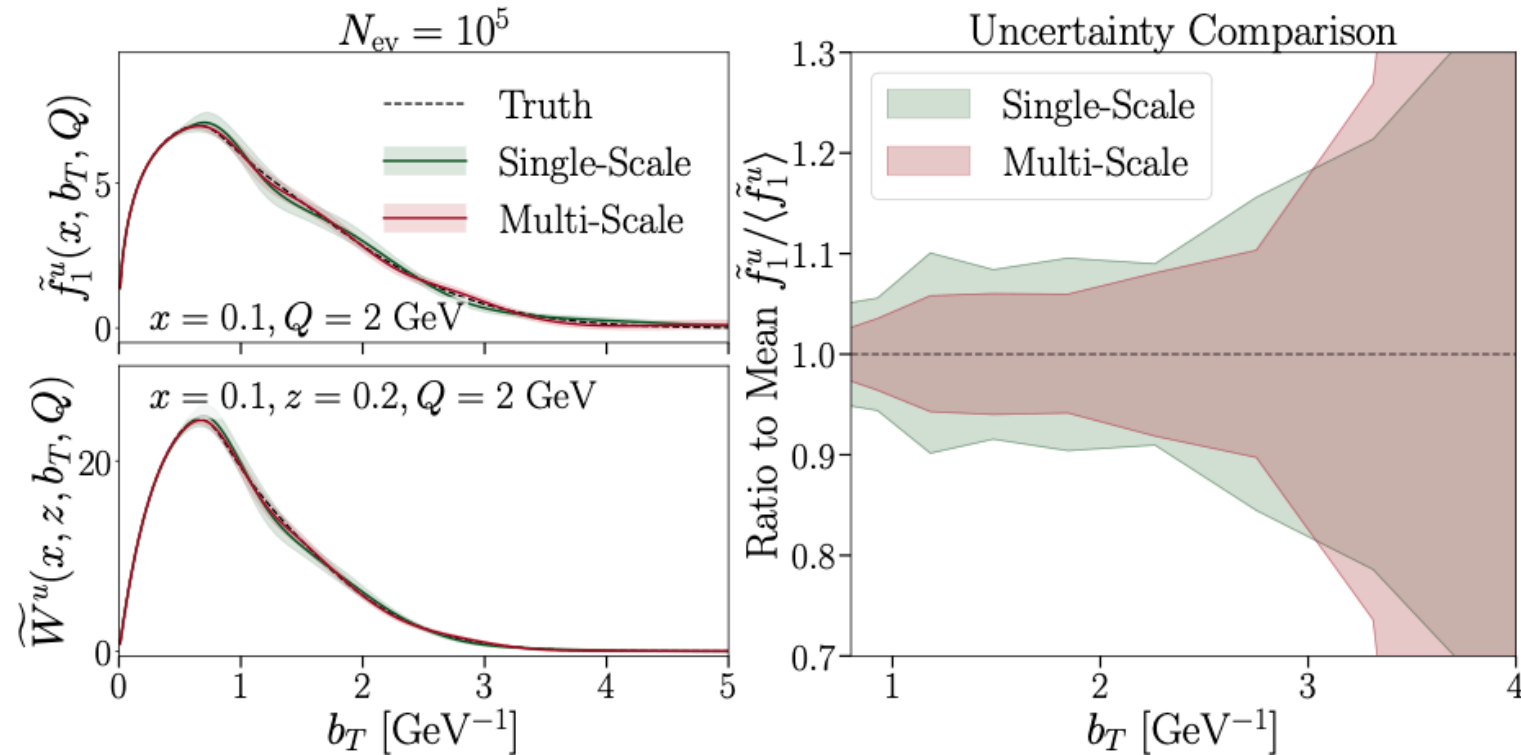


*Zaccheddu, Gamberg, WM, Pitonyak, Prokudin, Qiu, Sato
arXiv:2605.06606*

→ framework successfully deconvolutes the signal, and describes structure function across all statistical regimes

Pixel-based analysis of TMDs

■ Closure test for reconstructed TMDs in impact parameter space



*Zaccheddu et al.,
arXiv:2605.06606*

- correctly recover hadronic structure from convoluted data
- establishes robust foundation for future 3D imaging programs at JLab & EIC

Summary

- New approach to global QCD analysis — move from fitting parameters to minimize χ^2 function to fitting χ^2 function itself
 - practical use of MCMC sampling requires AI/ML methods
 - first application of MCMC + normalizing flow to pion structure
 - next: apply to (unpolarized and polarized) proton structure (PDFs, FFs)
- Reduce parametrization bias using pixels instead of functional forms, but at the expense of complicated high-dimensional posterior
 - NF allows practical sampling of posterior
- New methodology establishes foundation for future 3D imaging programs (Jefferson Lab, EIC)
 - next: apply to simultaneous extraction of collinear and TMD PDFs and FFs
- AI/ML will play a growing role in global analysis efforts to reconstruct hadron structure from experiment with rigorous uncertainty quantification and significantly reduced bias

Special thanks to:



Peter Risse
(JLab)



Nobuo Sato
(JLab)



Marco Zaccheddu
(ANL)