GPDs from Symbolic Regression

Zaki Panjsheeri

20 July 2026

Towards improved femtography workshop

University of Virginia

GPDs and angular momentum

- Access to the proton's intrinsic angular momentum structure

$$\frac{1}{2} = J_q + J_g = \frac{1}{2}\int dx x\Big(H_q(x,0,0) + E_q(x,0,0) + H_g(x,0,0) + E_g(x,0,0)\Big)$$

Xiangdong Ji, PRL 78 (1997) 610-613

- Compare with the Jaffe-Manohar picture

$$\frac{1}{2}\Delta\Sigma_q + L_q^{\mathrm{JM}} + \Delta G + L_g^{\mathrm{JM}} = \frac{1}{2}$$

$$\frac{1}{2}\Delta\Sigma_q + L_q^{\mathrm{Ji}} + J_g^{\mathrm{Ji}} = \frac{1}{2}$$

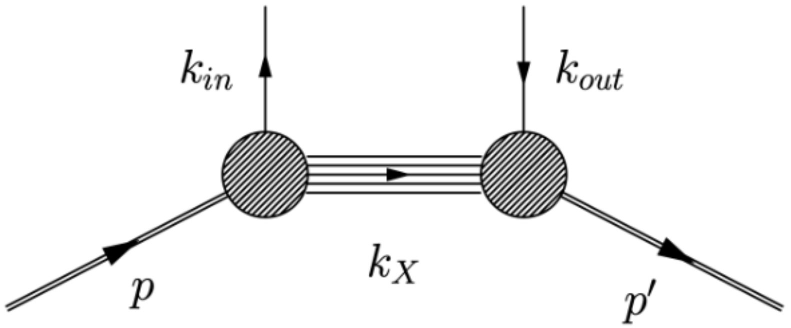
Jaffe and Manohar, *Nucl.Phys.B* 337 (1990) 509-546

- The two pictures are reconcilable through equations of motion and Lorentz invariance relations

A. Rajan, M. Engelhardt, & S. Liuti, PRD 98, 074022 (2018); M. Engelhardt PRD 95, 094505 (2017)

Spatial distribution

- Access to the proton's internal spatial structure
- Momentum fraction *x* dependence, different from electromagnetic form factors



D. Soper, PRD **15**, 1141 (1977)

M. Burkardt, PRD **62**, 071503(R) (2000)

$$H_q(X,0,t) = \int \frac{dz^-}{2\pi}\, e^{iXp^+z^-} \langle p-\Delta \mid \bar{\psi}_+(0)\,\psi_+(z^-) \mid p\rangle$$

$$H_q(X,0,t) = \int d^2\mathbf{k}_{T,in}\, \phi^*(X,\mathbf{k}_{T,in} - \mathbf{\Delta})\phi(X,\mathbf{k}_{T,in}).$$

$$H_q(X,0,t) = \int d^2\mathbf{b}\, e^{i\mathbf{b}\cdot\mathbf{\Delta}}\, \tilde{\phi}^*(X,\mathbf{b})\; \tilde{\phi}(X,\mathbf{b})$$

$$\rho_q(X,\mathbf{b}) = \tilde{\phi}^*(X,\mathbf{b})\; \tilde{\phi}(X,\mathbf{b})\,.$$

$$\langle b_T^2(X)\rangle^{1/2} = \frac{\int d^2b_T b_T^2 \rho^{q,g}(X,b_T)}{\int d^2b_T \rho^{q,g}(X,b_T)}$$

Outline

1. Traditional phenomenology: parameterizations based on a physics model (UVA2)
2. Neural networks: replacing physics model parameterizations with "super flexible" parameterizations (NNGPD)
3. Symbolic regression: discovering the parameterization *from the data*

Rationale behind UVA2

- To be able to make predictions and get a handle on what GPDs look like in their four-dimensional variable space, start *modeling/parameterizing*
- We do so with all available information:
  - Inclusive scattering data (PDFs, LO NNPDF21)
  - Electromagnetic form factors (Cates et al., PRL106, 252003 (2011))
  - LQCD M_2 results (Hackett et al., PRL 132, 251904 (2024))
  - Boundary conditions and symmetries around x = 0
- Hessian based statistics – sufficient for us to get *a* solution that satisfies all the constraints

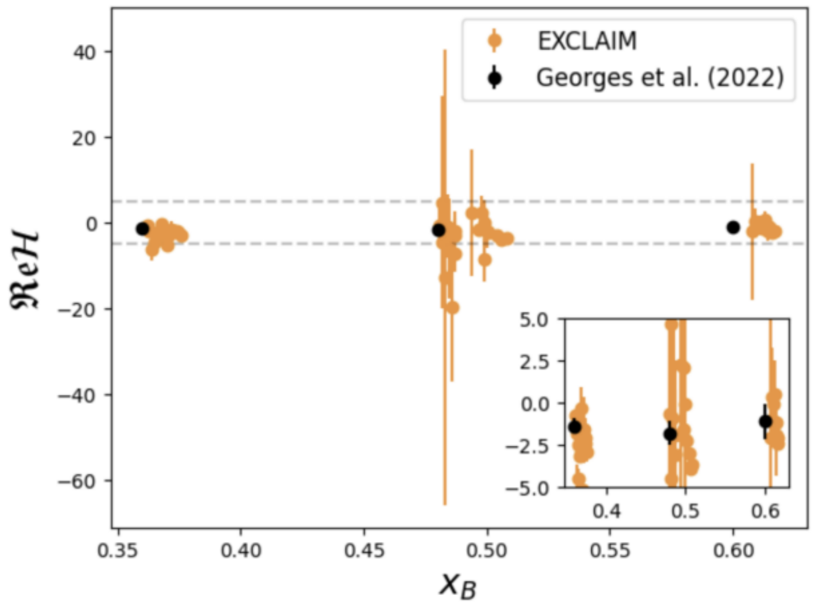
$$\chi_E^2(a,\lambda) = \sum_{k=1}^{N_{pt}} \frac{1}{s_k^2}\left(\boxed{D_k - T_k(a)} - \sum_{\alpha=1}^{N_\lambda} \lambda_\alpha \beta_{k\alpha}\right)^2 + \sum_{\alpha=1}^{N_\lambda} \lambda_\alpha^2$$

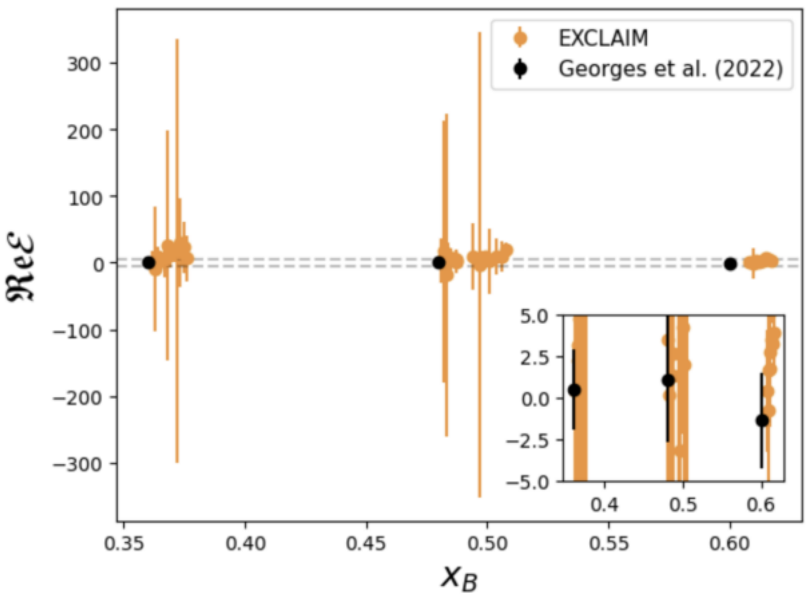
arXiv:2410.23469

Submitted to PRD

Rationale behind UVA2

- Challenging to apply presently available DVES data in a Hessian approach, due to issues of the degeneracy of the solution space





Likelihood & Correlation Analysis of Compton Form Factors for Deeply Virtual Exclusive Scattering on the Nucleon

Douglas Q. Adams,[1,*] Joshua Bautista,[1] Marija Čuić,[1] Adil Khawaja,[1] Saraswati Pandey,[1] Zaki Panjsheeri,[1] Gia-Wei Chern,[1,†] Yaohang Li,[2] Simonetta Liuti,[1,‡] Marie Boër,[3] Michael Engelhardt,[4] Gary R. Goldstein,[5] Huey-Wen Lin,[6] and Matthew D. Sievert[4]
(EXCLAIM Collaboration)

See
D. Q. Adams'
talk from this
morning

Parameterizing GPDs (UVA2)

arXiv:2511.03065

Accepted to EPJC

Grids available on GitHub: https://github.com/Exclaim-Collaboration/UVA2

- Incorporates LO off-forward DGLAP evolution and preserves the symmetries of the ERBL region

$$F^{q,g}(X,\zeta,t;Q^2) = F^{DGLAP}_{q,g}\left[\theta(X-\zeta)+\theta(-X)\right] + F^{ERBL}_{q,g}\left[\theta(-X+\zeta)+\theta(X)\right]$$

- Includes valence, sea quark, and gluon degrees of freedom

$$\begin{aligned} H_g^{DGLAP} &= \mathcal{A}_H^g\, X^{-(\alpha_g+\alpha'_g(1-X)^{p_g}t)}\,(a_0^g I_0 + a_1^{g+} I_1 + a_2^{g+} I_2) \\ H_g^{ERBL} &= a_g^- X^2 + b_g^- X + c_g^- \end{aligned}$$

Updated flexible global parametrization of generalized parton distributions from elastic and deep inelastic inclusive scattering data

Zaki Panjsheeri,[1,*] Douglas Q. Adams,[1,†] Adil Khawaja,[1,‡]
Saraswati Pandey,[1,§] Kemal Tezgin,[2,¶] and Simonetta Liuti[1,**]

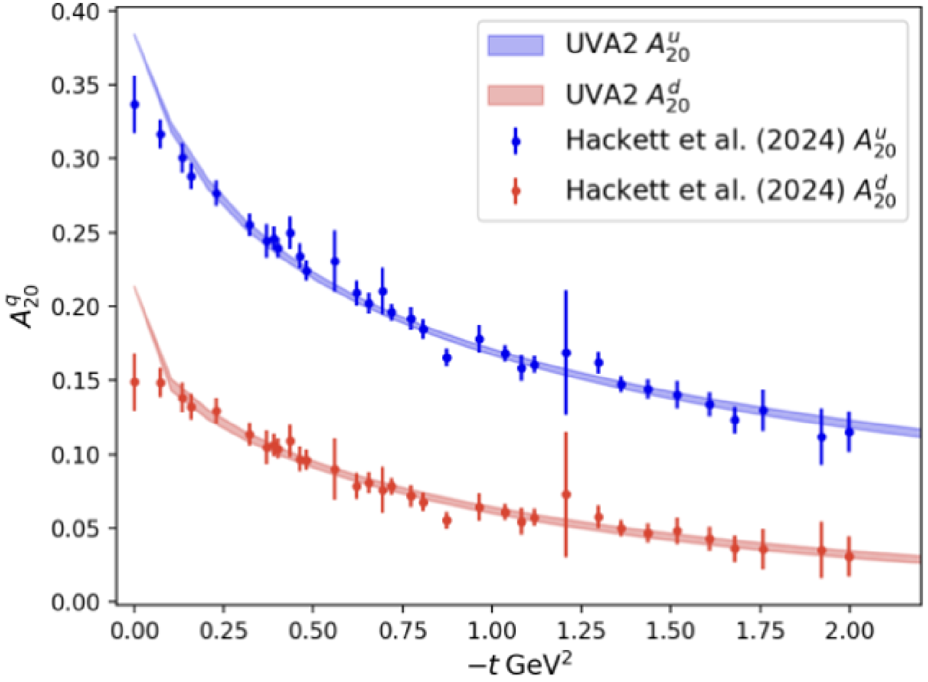
Reggeized spectator model

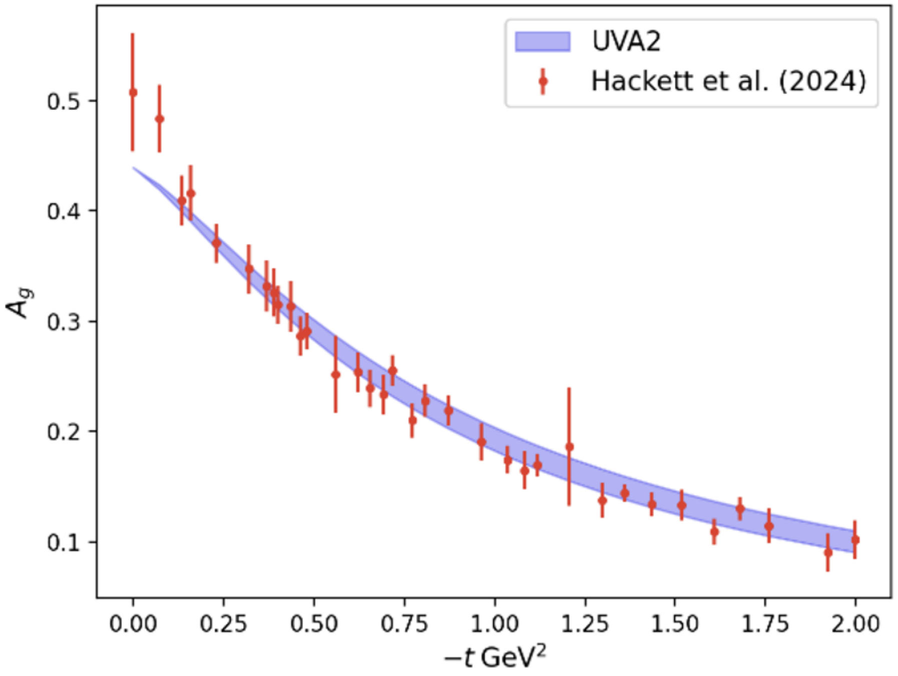
$$M_2^{q,F} = \frac{1}{2}\int_{-1+\zeta}^{1}\frac{dX}{1-\zeta/2}X(H^+ + H^-) = \int_{\zeta/2}^{1}\frac{dX}{1-\zeta/2}XH^+ = M_2^{q_v} + 2M_2^{\bar{q}}.$$

$$A_{10}^{g} = \int_{\zeta/2}^{1}\frac{dX}{1-\zeta/2}H_g$$

$$\boxed{R_p^{\alpha,\alpha'}(X,t) = X^{-(\alpha+\alpha'(1-X)^p t)}}$$

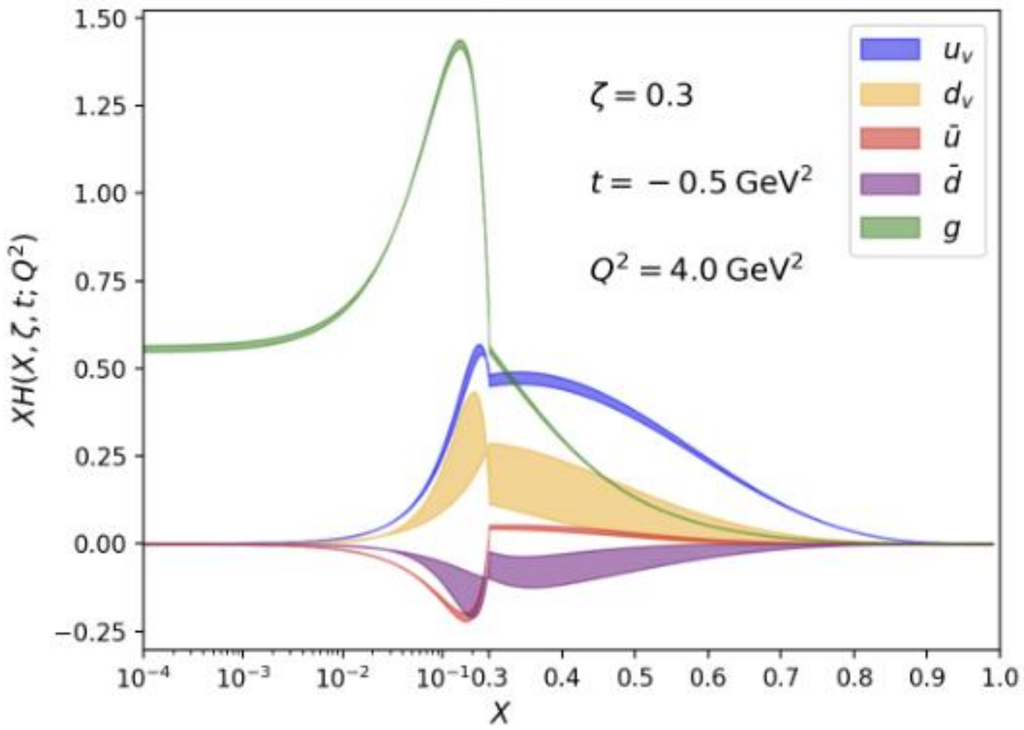
J. O. Gonzalez-Hernandez, S. Liuti, G. Goldstein, K. Kathuria, PRC 88, 065206 (2013)



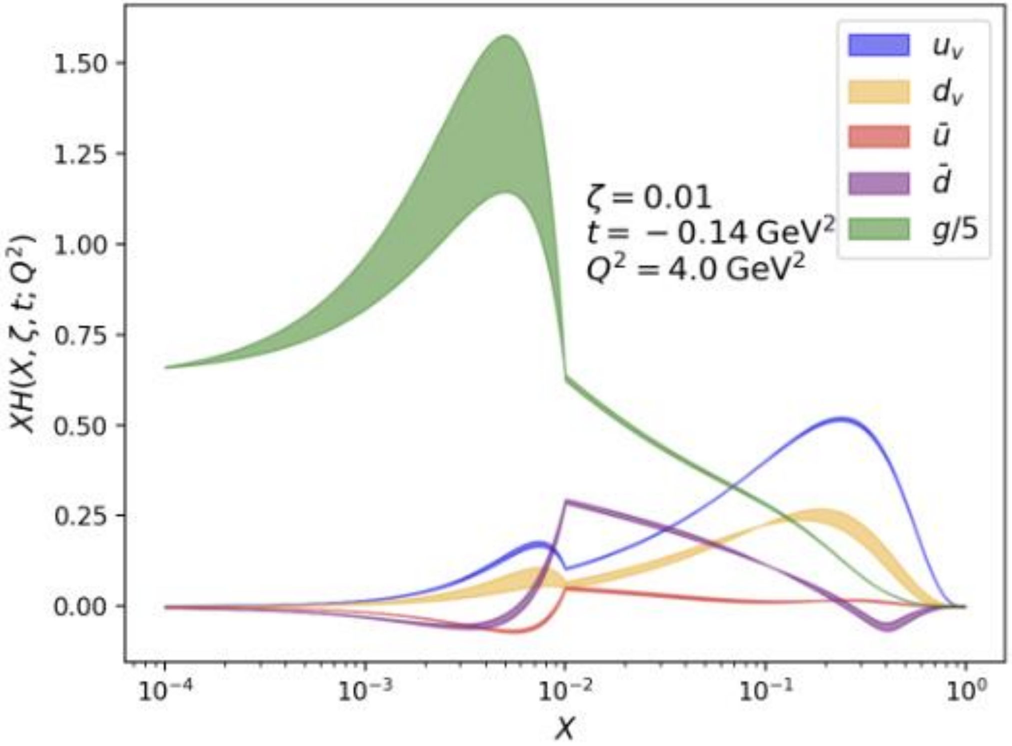


Sea quarks and gluons constrained with lattice data from Hackett et al., PRL 132, 251904 (2024) in UVA2

Predictions



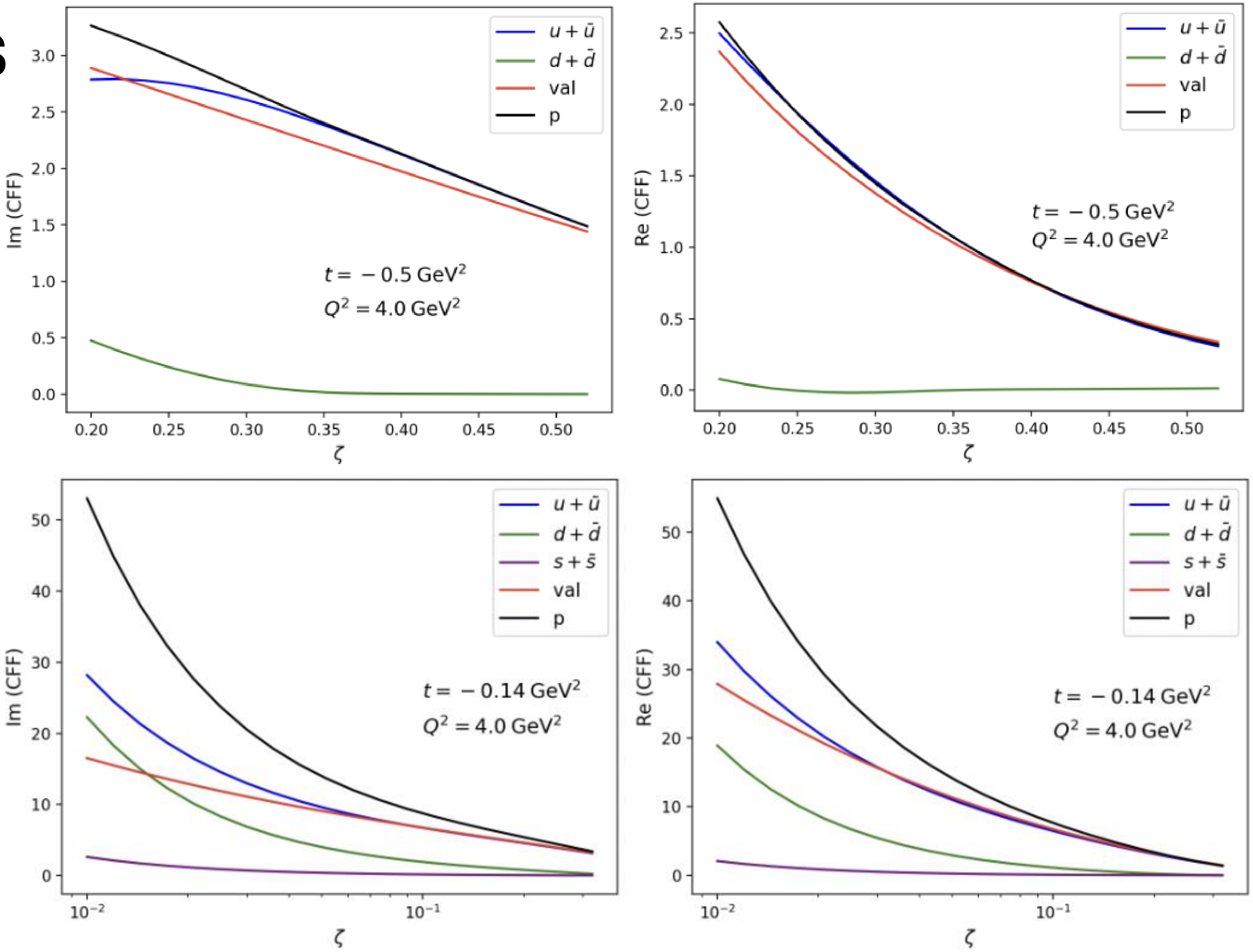
JLab



EIC

Emphasis on uncertainty quantification in UVA2

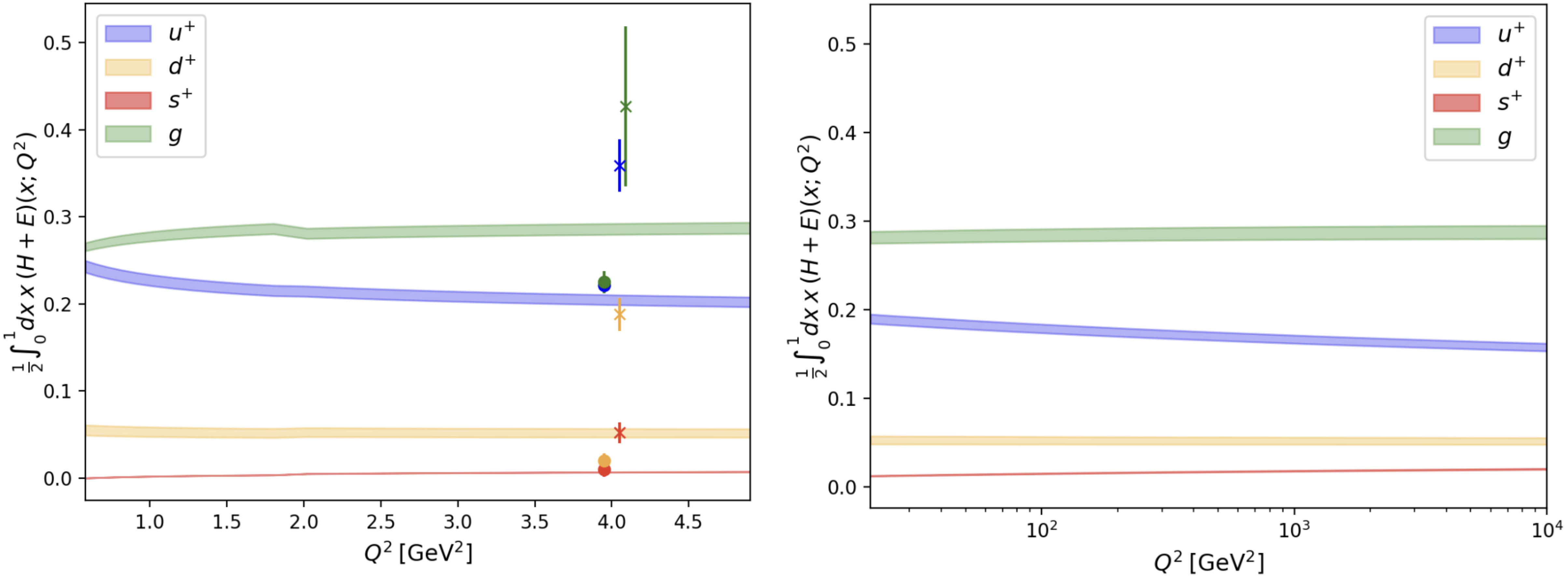
CFFs



JLab

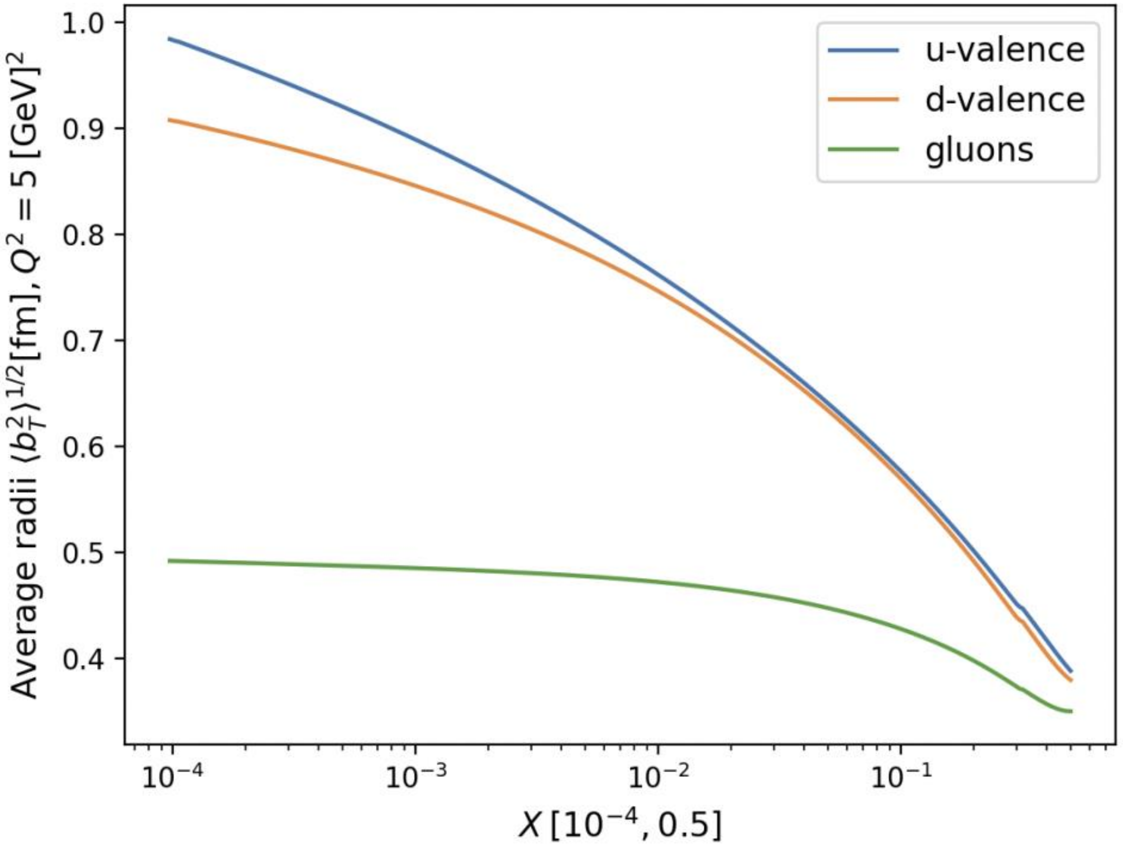
EIC

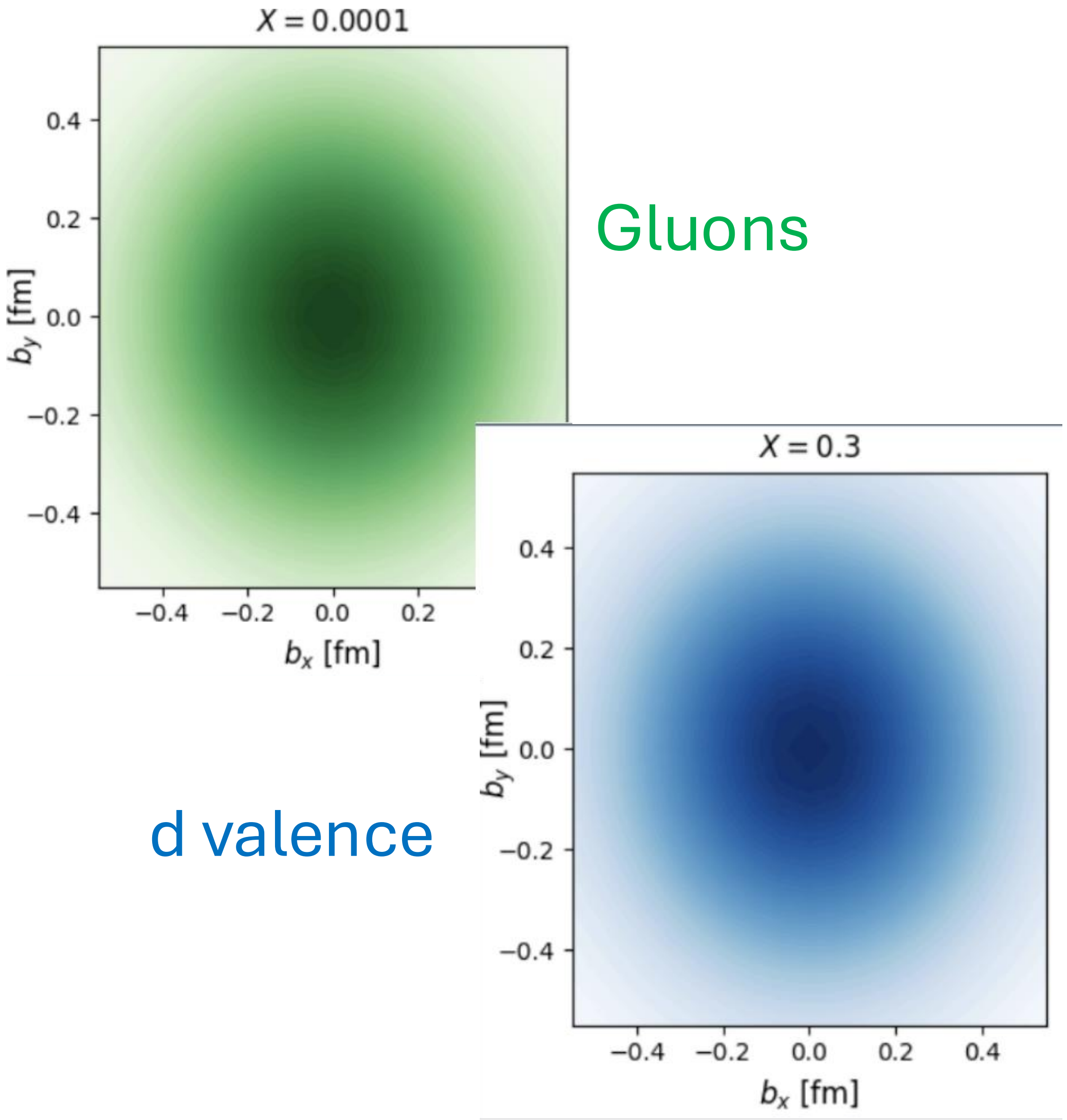
Angular momentum



Compare to (x markers) C. Alexandrou et al. PRD 101, 094513 (2020) and (dot markers) D. Hackett, PRL 132, 251904 (2024), and A. V. Thomas, PRL 101, 102003 (2008)

Spatial structure

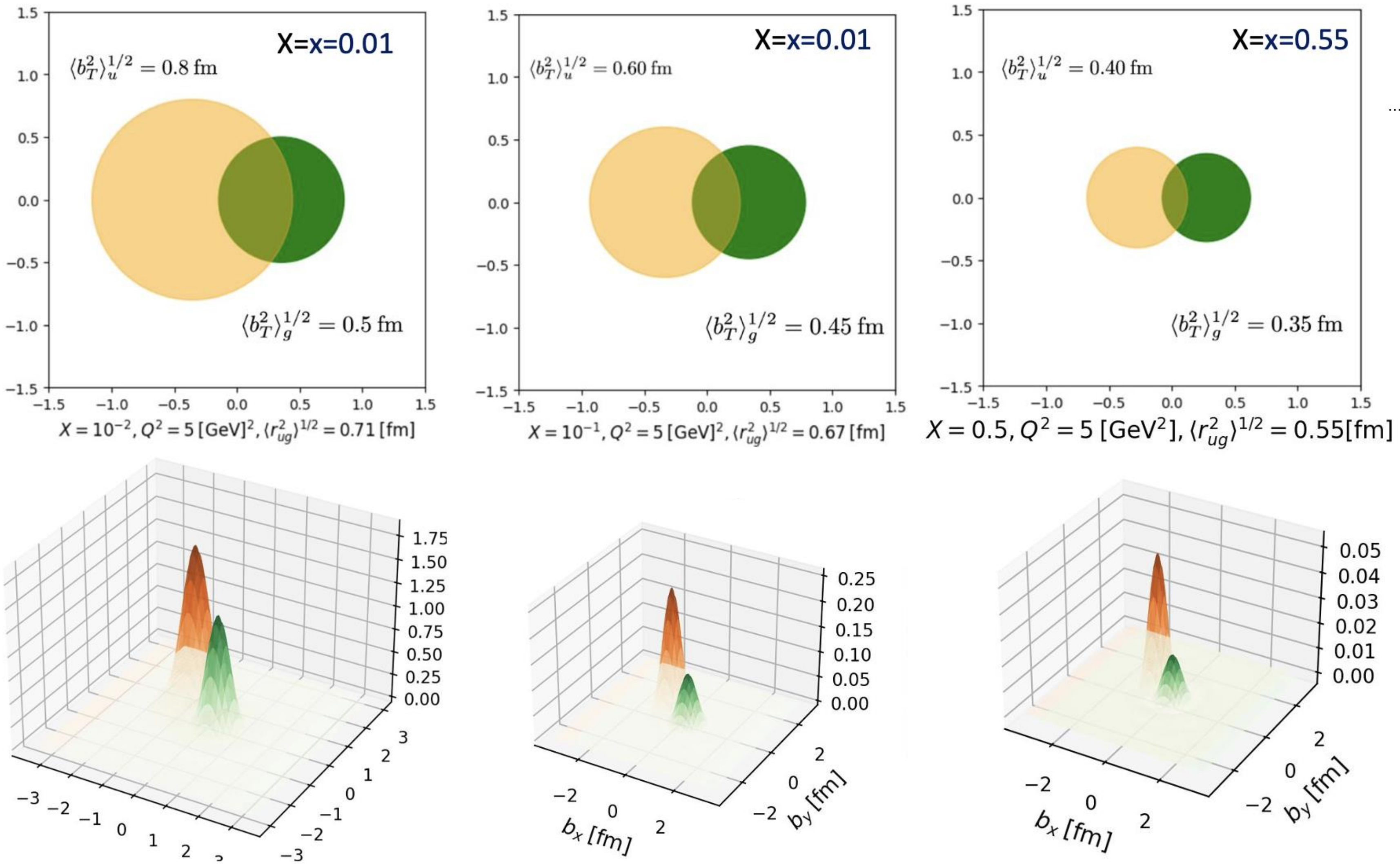




Gluons

d valence

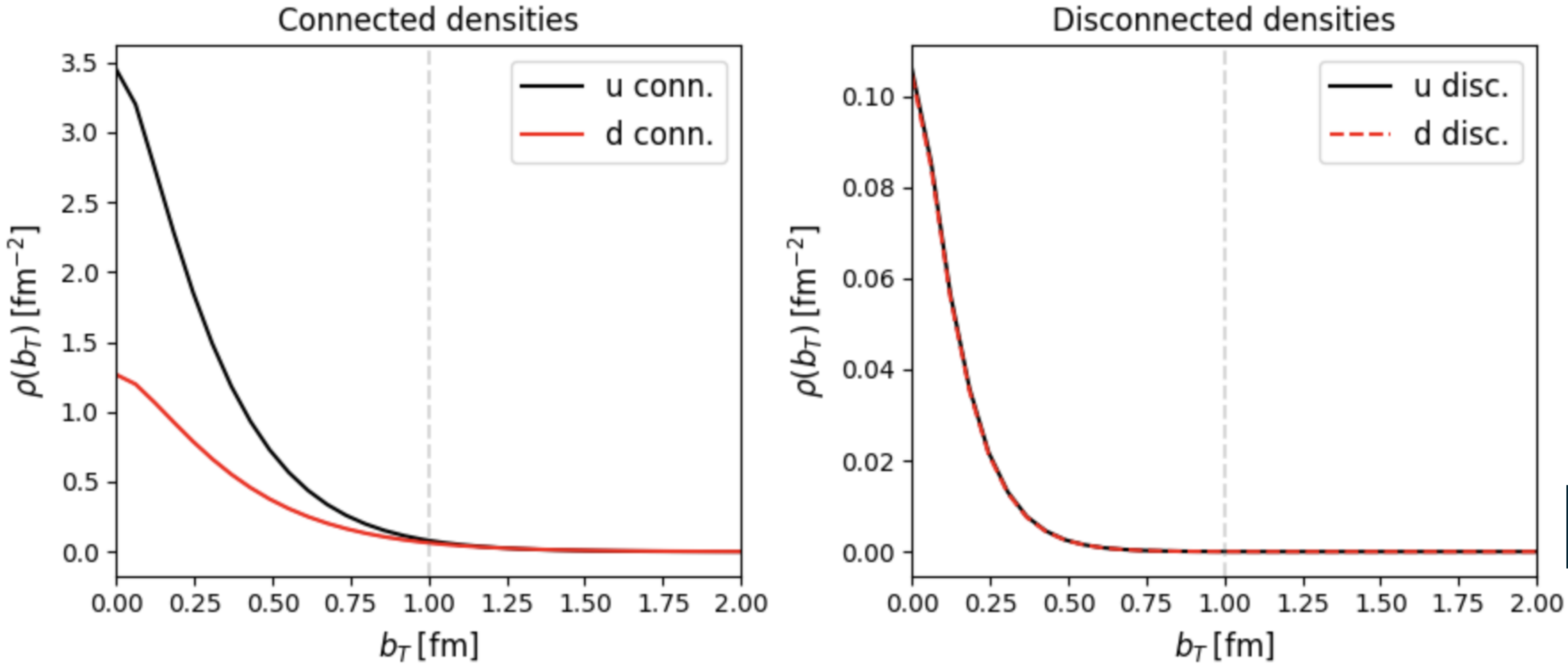
Beyond the usual one-body picture



More on this in Z.P. and S. Liuti, “The Correlated Spatial Structure of the Proton: a framework for dynamical imaging,” forthcoming.

Beyond the usual one-body picture

arXiv:2512.20853



See
S. Pandey's talk
on Friday

$\langle \mathbf{b}_\perp^2 \rangle^{1/2}$ [fm]	u	d
total	0.596	0.721
connected	0.601	0.734
disconnected	0.290	0.289

Connected and disconnected contributions to nucleon form factors and parton distributions

Zaki Panjsheeri,[1,*] Saraswati Pandey,[1] Brannon Semp,[1] and Simonetta Liuti[1,†]

[1] *Department of Physics, University of Virginia, Charlottesville, VA 22904, USA.*

Why AI/ML?

- Two central questions guide the Exclusives in AI/ML (ExclAIm) Collaboration:
  - How can AI/ML be used as a tool for elucidating proton structure?
  - How can we help, with the expertise and mindset of physicists, explain or interpret the working of ML algorithms?

***Position:* Is machine learning good or bad for the natural sciences?**

David W. Hogg [1 2 3] **Soledad Villar** [3 4 5] arXiv: 2405.18095

Why do we do astrophysics?

by **David W. Hogg**[1] arXiv:2602.10181

I conclude with a discussion of two possible (extreme and bad) policy recommendations related to the use of LLMs in astrophysics, dubbed "let-them-cook" and "ban-and-punish." I argue strongly against both of these; it is not going to be easy to develop or adopt good moderate policies.

Rationale behind NNGPD

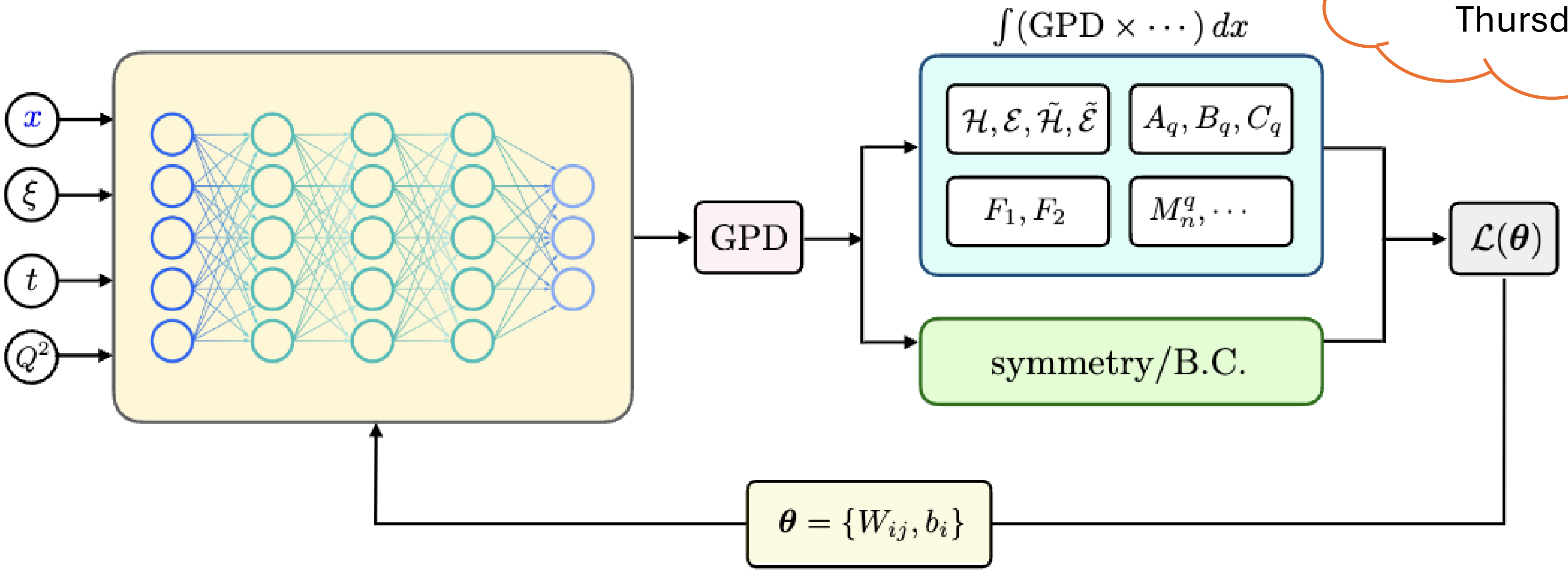
- All of this was the extraction of GPDs from data (electromagnetic form factors, LQCD generalized form factors, inclusive scattering data, etc.) within *particular* parameterizations (Hessian-based)
- Can we make a parameterization that is unbiased towards the underlying physics, but captures the constraints from symmetries, experimental data, and lattice results?
  - Need a "super flexible" parameterization to do so

Scheme of NNGPD

arXiv:2605.06994

Submitted to JHEP

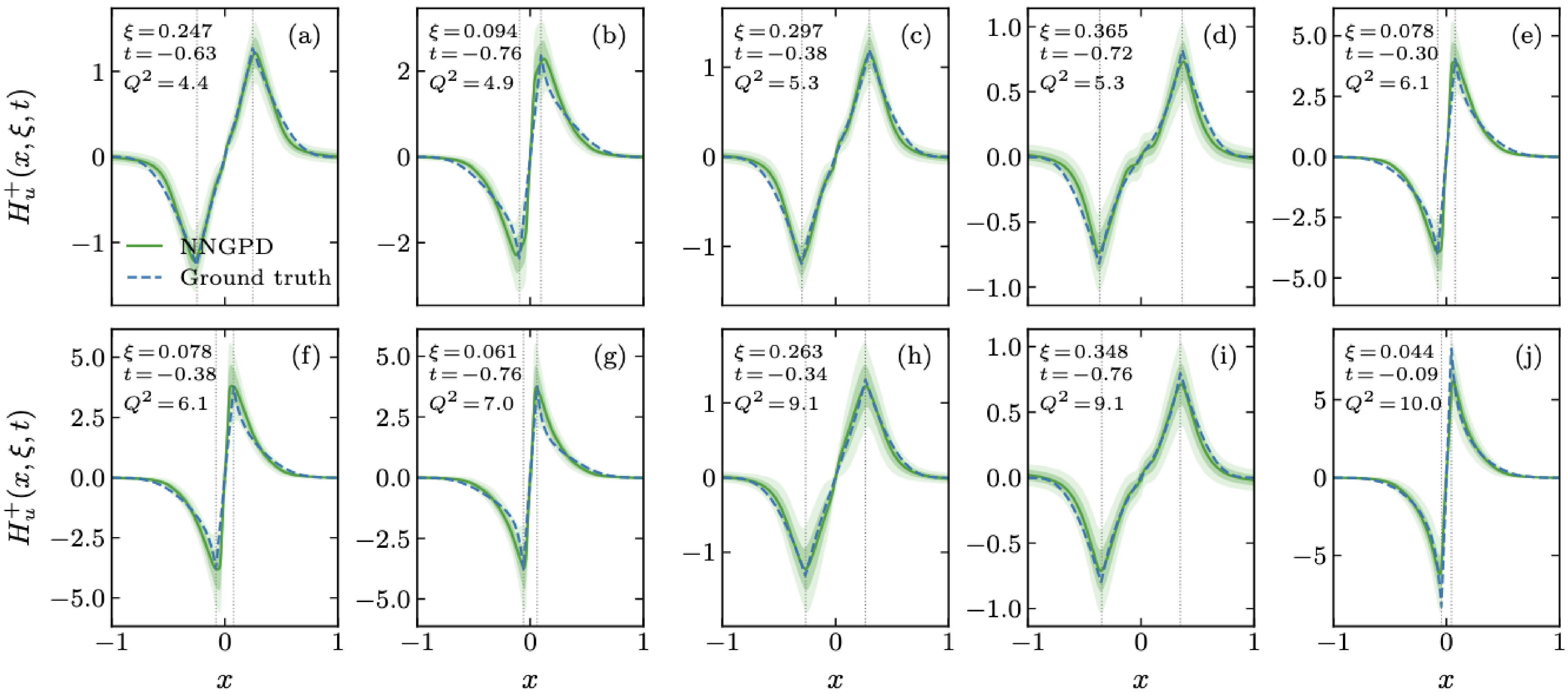
See
G.-W. Chern's
talk on
Thursday



Neural Network Representation of Generalized Parton Distributions (NNGPD)

Jitao Xu,[1,*] Ho Jang,[2,*] Zaki Panjsheeri,[2,*] Gia-Wei Chern,[2,†] Yaohang Li,[1,‡]
Simonetta Liuti,[2,§] Douglas Q. Adams,[2] Michael Engelhardt,[3] Gary R. Goldstein,[4]
Adil Khawaja,[2] Huey-Wen Lin,[5] Saraswati Pandey,[2] and Kemal Tezgin[6]

EXCLAIM Collaboration



- Closure test performed on UVA2
- The "+" distribution is well-captured by the network over a wide range of kinematics, inclusive of JLab and EIC

$$\mathcal{L}^+_{\text{total}} \;=\; w_{CFF}\mathcal{L}_{CFF} + w^+_{MM}\mathcal{L}^+_{MM} + w^+_{EP}\mathcal{L}^+_{EP} + w_{KL}\mathcal{L}_{KL}$$

Rationale behind GPDs from SR

- We want to do more than use a neural network “black box”
- Let’s take a step beyond the NN: Uncover the parameterization *itself* from the data!

GPDs from SR + LQCD

arXiv:2504.13289

Accepted to PRD

- What if we tried to *learn* the x and *t* dependence directly from data?
- How well can we extract physics content from LQCD GPD results obtained in a limited range of the kinematics? Can we extrapolate?

Generalized Parton Distributions from Symbolic Regression

Andrew Dotson,[1,*] Zaki Panjsheeri,[2,†] Anusha Reddy Singireddy,[3] Douglas Q. Adams,[2] Marija Čuić,[2] Emmanuel Ortiz-Pacheco,[4] Marie Boër,[5] Gia-Wei Chern,[2] Michael Engelhardt,[1] Gary R. Goldstein,[6] Yaohang Li,[3] Huey-Wen Lin,[4] Simonetta Liuti,[2,‡] and Matthew D. Sievert[1]

EXCLAIM Collaboration

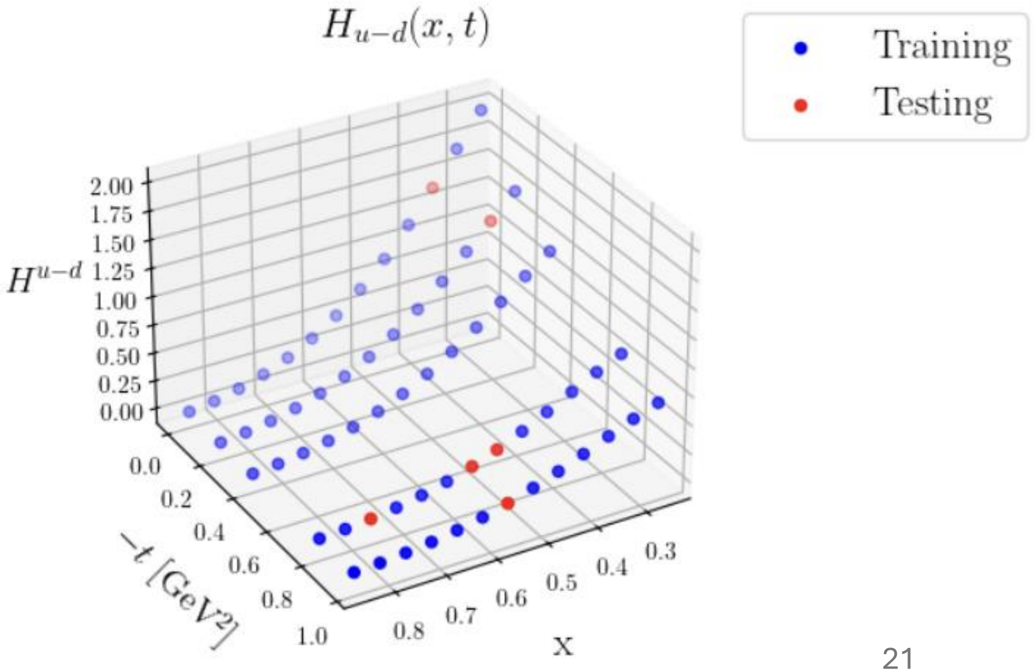
First result

LQCD results calculated in the LaMET, quasi-PDF method, H.W. Lin, PRL 127, 182001 (2021)

- The first time we ran PySR, we happened to obtain a result that is factorized in the *x* and *t* variables!
- Significant consequences in impact-parameter space

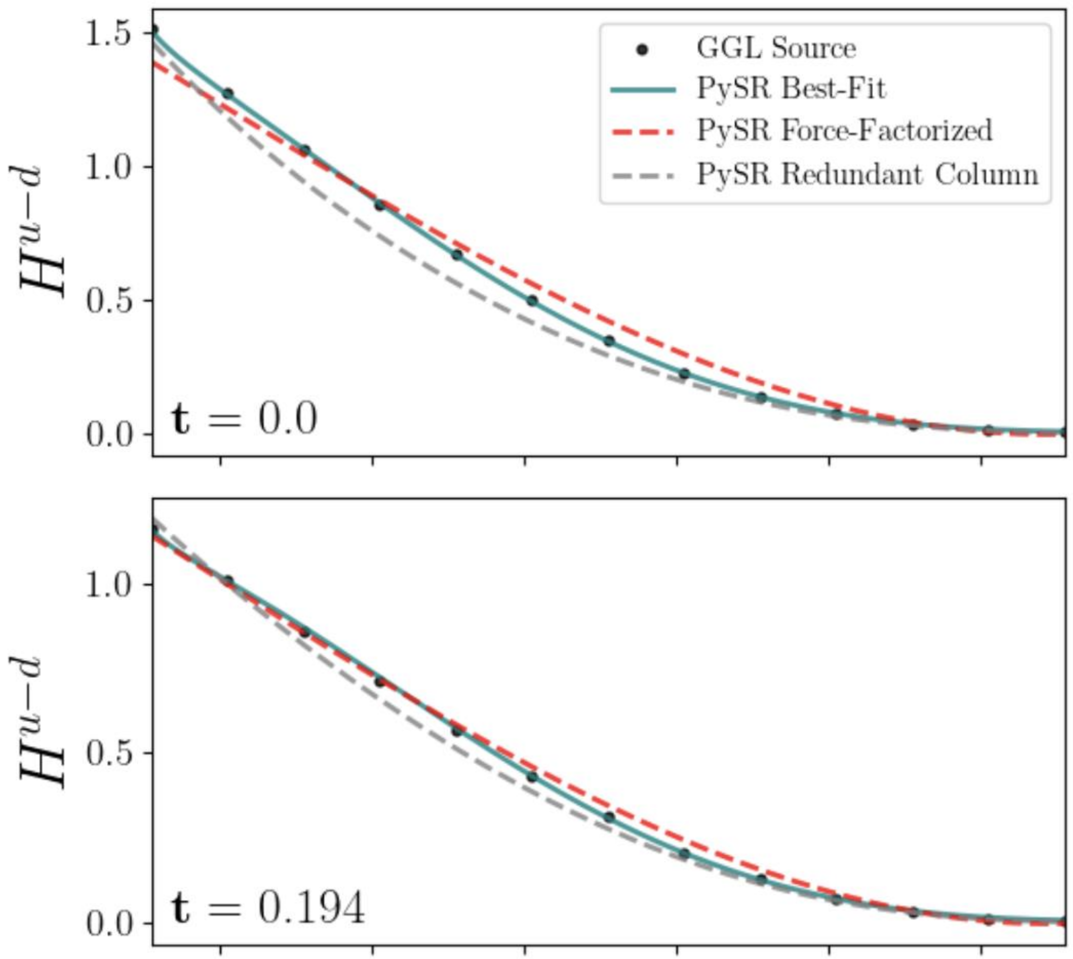
$$H^{u-d}(X,t) = \frac{\alpha(\beta - X)(\gamma - X)}{-t + \delta}$$

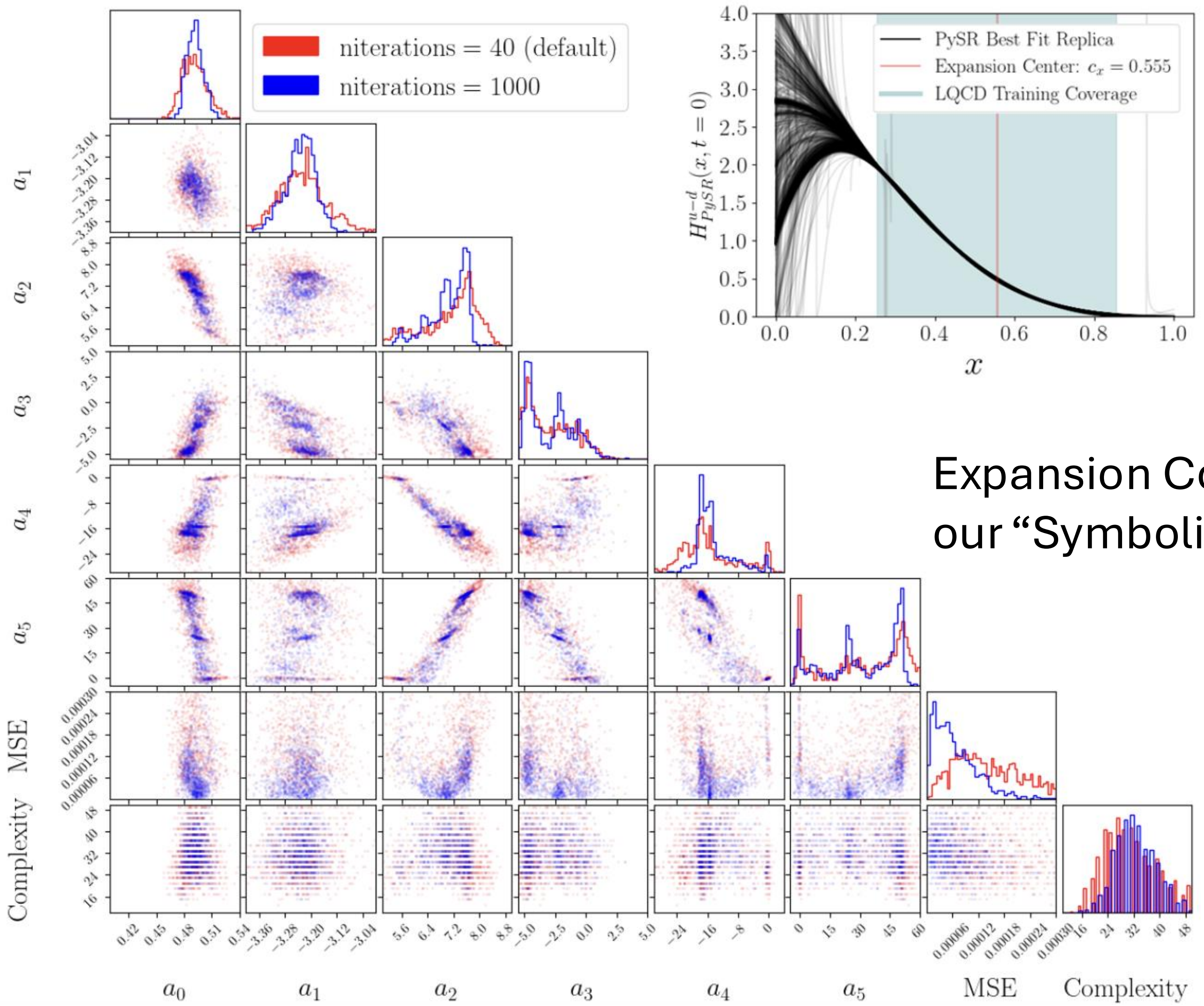
$$\begin{aligned} \alpha &= 2.019 \\ \beta &= 0.832 \\ \gamma &= 1.040 \\ \delta &= 0.495 \end{aligned}$$



Benchmarking

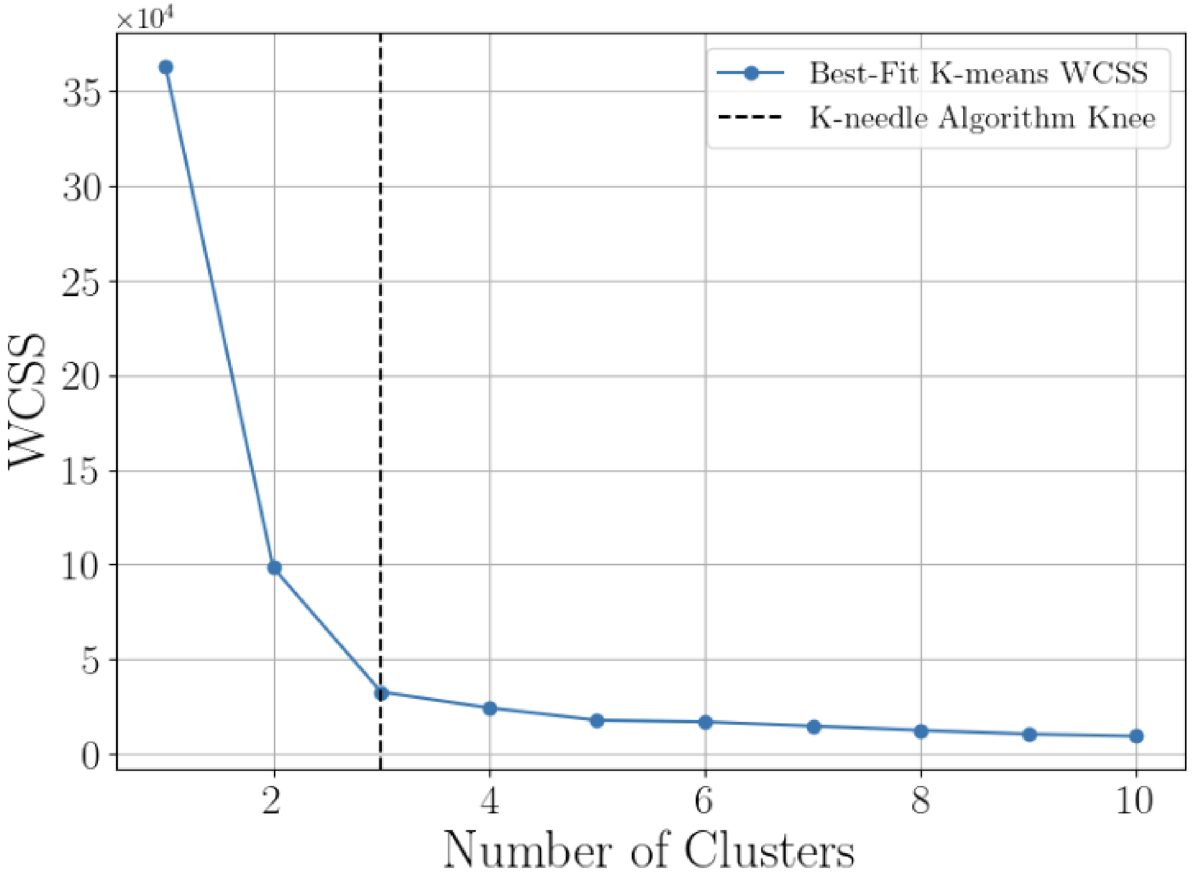
- Force-factorized versus best-fit
- Finite moment filter
- Force-Reggeized form
- Compared the result on LQCD data with three phenomenological models, GK, VGG, and GGL/UVA2
- Clustered the solutions obtained by PySR
- Compared to a NN fit
- Compared to experimental extraction of electromagnetic form factors





Expansion Coefficient Clustering (ECC), our “Symbolic Covergence” method

From 1,000 to 3 solutions

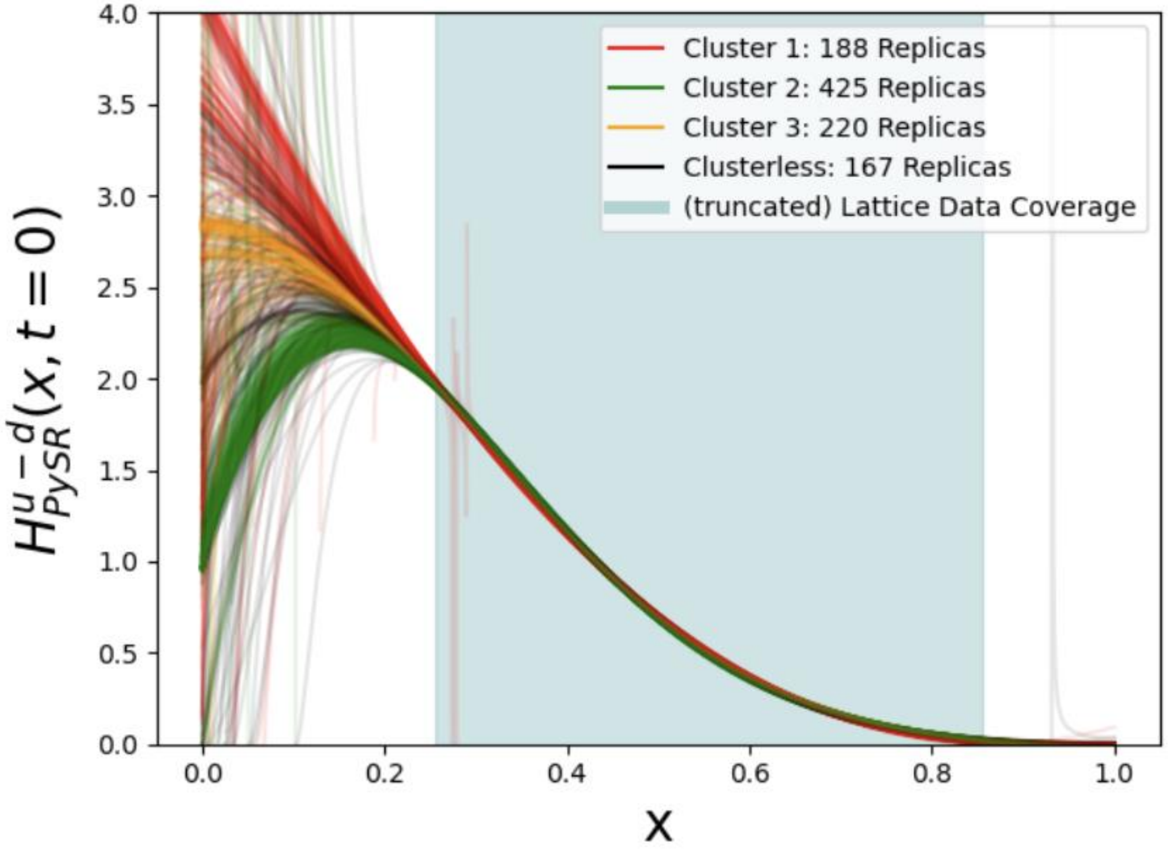


Within-cluster sum of squares:

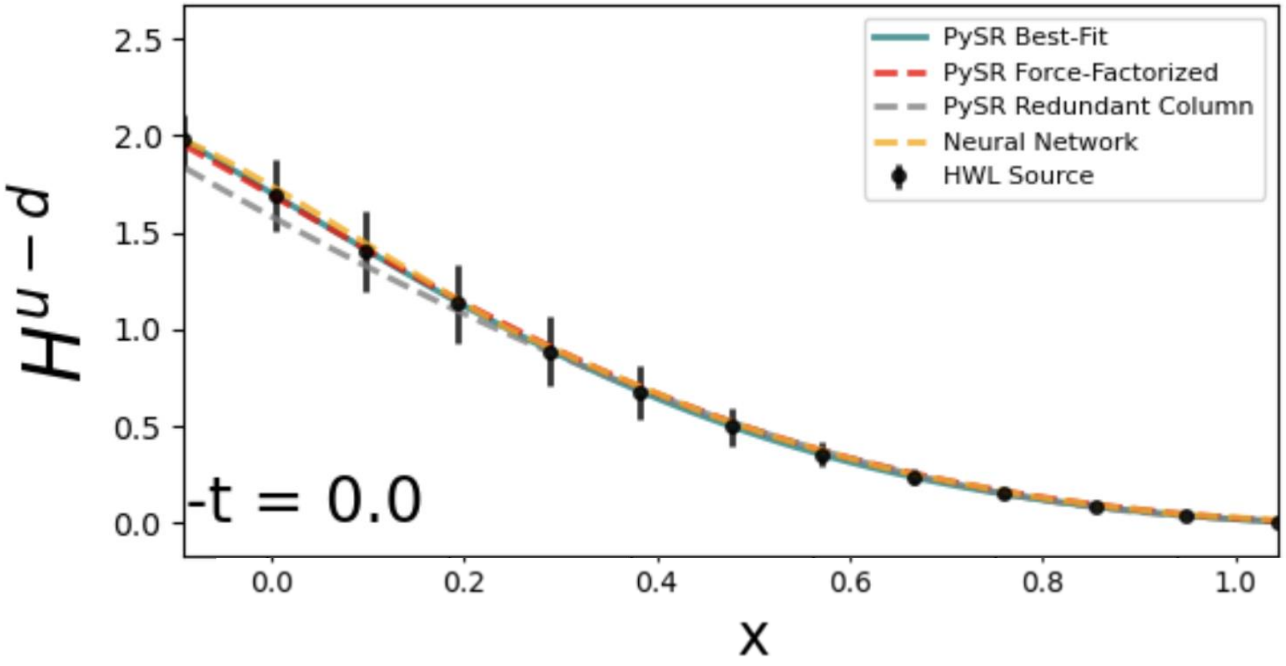
$$\mathrm{WCSS}(k) = \underset{C_1,\ldots C_k}{\arg\min} \sum_{j=1}^{k} \sum_{x_i \in C_j} |x_i - \mu_j|^2$$

$$K_f(k) = \frac{f''(k)}{(1 + f'(k)^2)^{3/2}}$$

Clustering and NN comparison

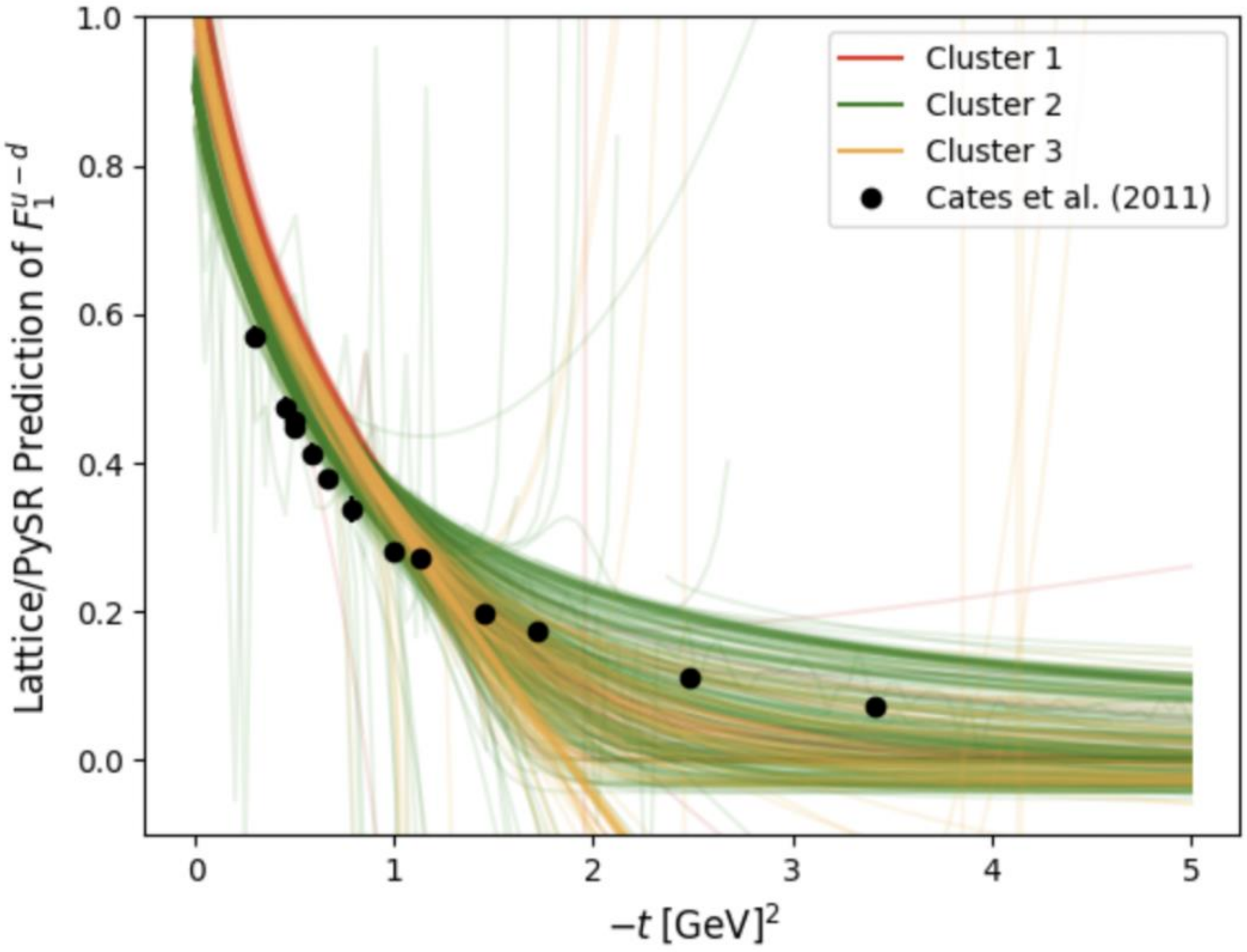


Clustering with HDBSCAN

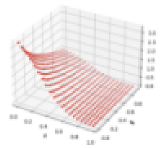
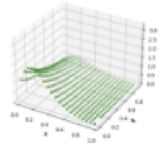
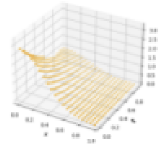
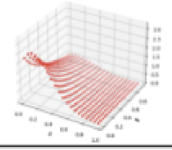
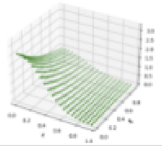
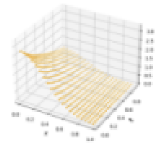
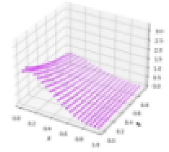


PySR allows for *human-interpretable ML output*.

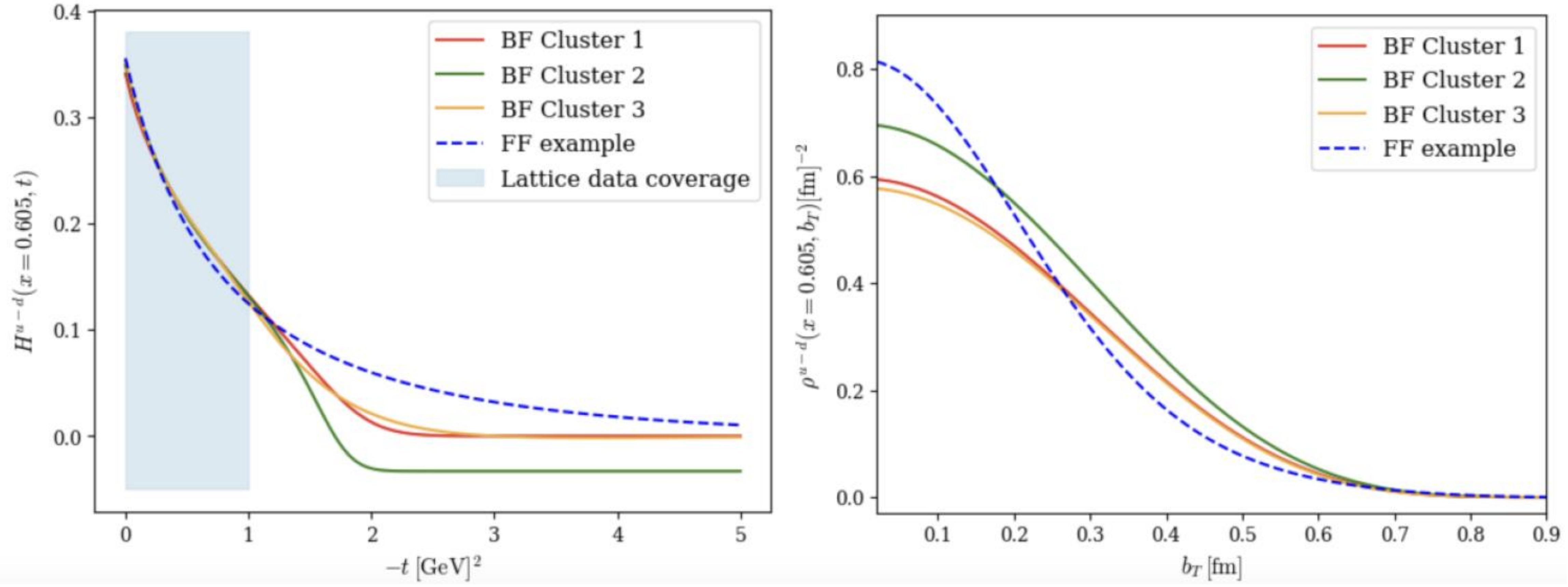
Comparison to experimental data



- Comparison to Cates et al. (2011) form factor flavor separation as a benchmark against our SR results
- Serves as a step towards understanding the uncertainties of the SR algorithm

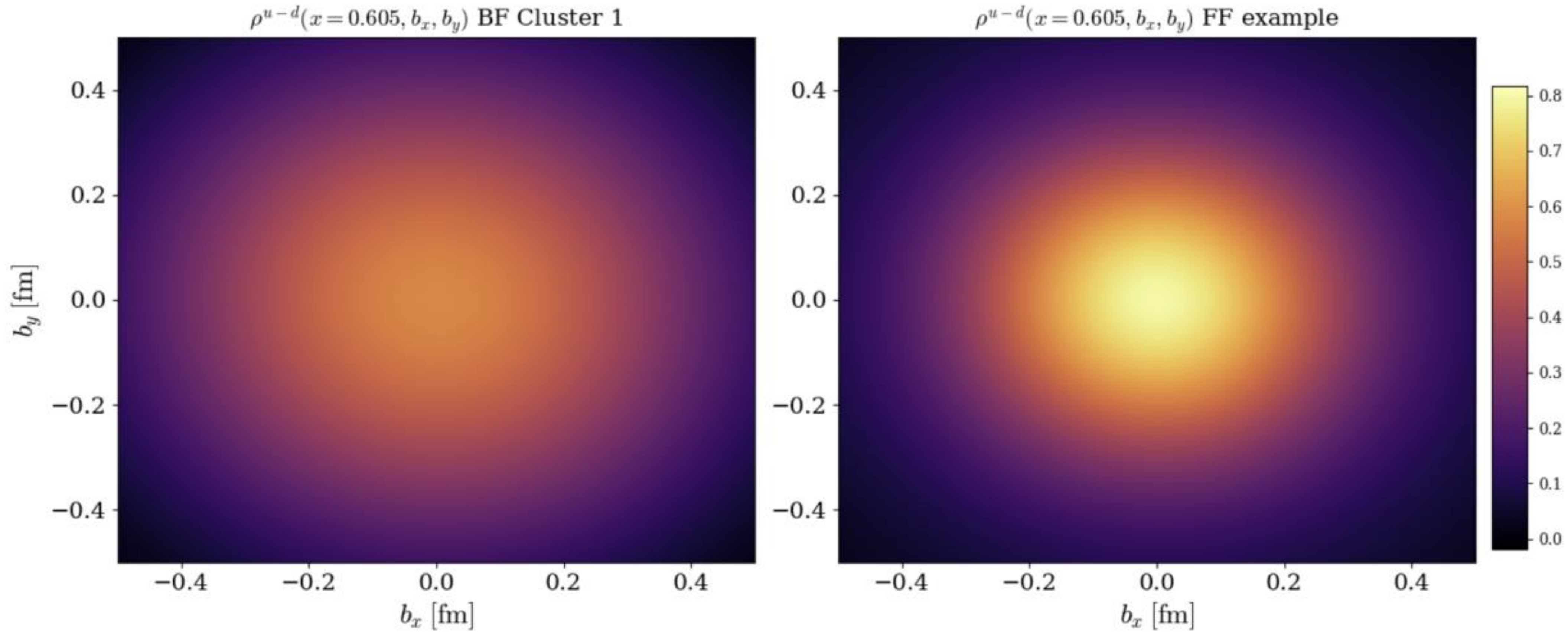
Selection criteria	Predictions over full phase space	Symbolic Forms	Fitness metrics WMSE MSE	Complexity
H_{BF}^{1}		$\frac{3.3(1-0.8\,x^{2})^{5.2}}{x+\,2\,\boldsymbol{t}+(1.52\,\boldsymbol{t}^{t})^{t}}$	3.3×10^{-3} 2.1×10^{-6}	21
H_{BF}^{2}		$\frac{(0.034\,x^{-4.52})^{x}}{1.89\boldsymbol{t}+0.9451x+0.94\{[(0.86(\boldsymbol{t}+1)-x)^{\boldsymbol{t}}]^{\boldsymbol{t}}\}^{\boldsymbol{t}}}-0.031$	2.5×10^{-3} 2.3×10^{-6}	31
H_{BF}^{3}		$\frac{1.02\{0.0047[[x+(\boldsymbol{t}-0.82)(\boldsymbol{t}-0.29)]^{t}]^{x}\}^{x}-0.018}{1.77^{t}-0.64}$	1.9×10^{-3} 2.1×10^{-6}	29
H_{FF}^{1}		$1.55\left(-0.483+\frac{1}{(0.42+\boldsymbol{t})^{0.75}}\right)\left(\frac{-0.0019+0.0043^{x}}{x}\right)^{x}$	6.6×10^{-2} 4.4×10^{-5}	23
H_{FF}^{2}		$\frac{-0.01+1.45(323^{x}x^{x})^{-x}}{0.56+\boldsymbol{t}^{1.14}}$	4.3×10^{-2} 5.5×10^{-5}	19
H_{FF}^{3}		$\frac{1.19\{-0.048+2.34[(0.063^{x})^{x}]^{1.85}\}}{(0.913+\boldsymbol{t}^{\boldsymbol{t}})^{t}+\boldsymbol{t}}$	7.1×10^{-3} 2.0×10^{-5}	29
H_{Regge}		$2.74\;e^{[-\boldsymbol{t}-0.67\;\ln^{2}(1-x)]}x^{0.087}(1-x)^{1.57}$	1.5×10^{-2} 1.9×10^{-4}	26

Coordinate space results

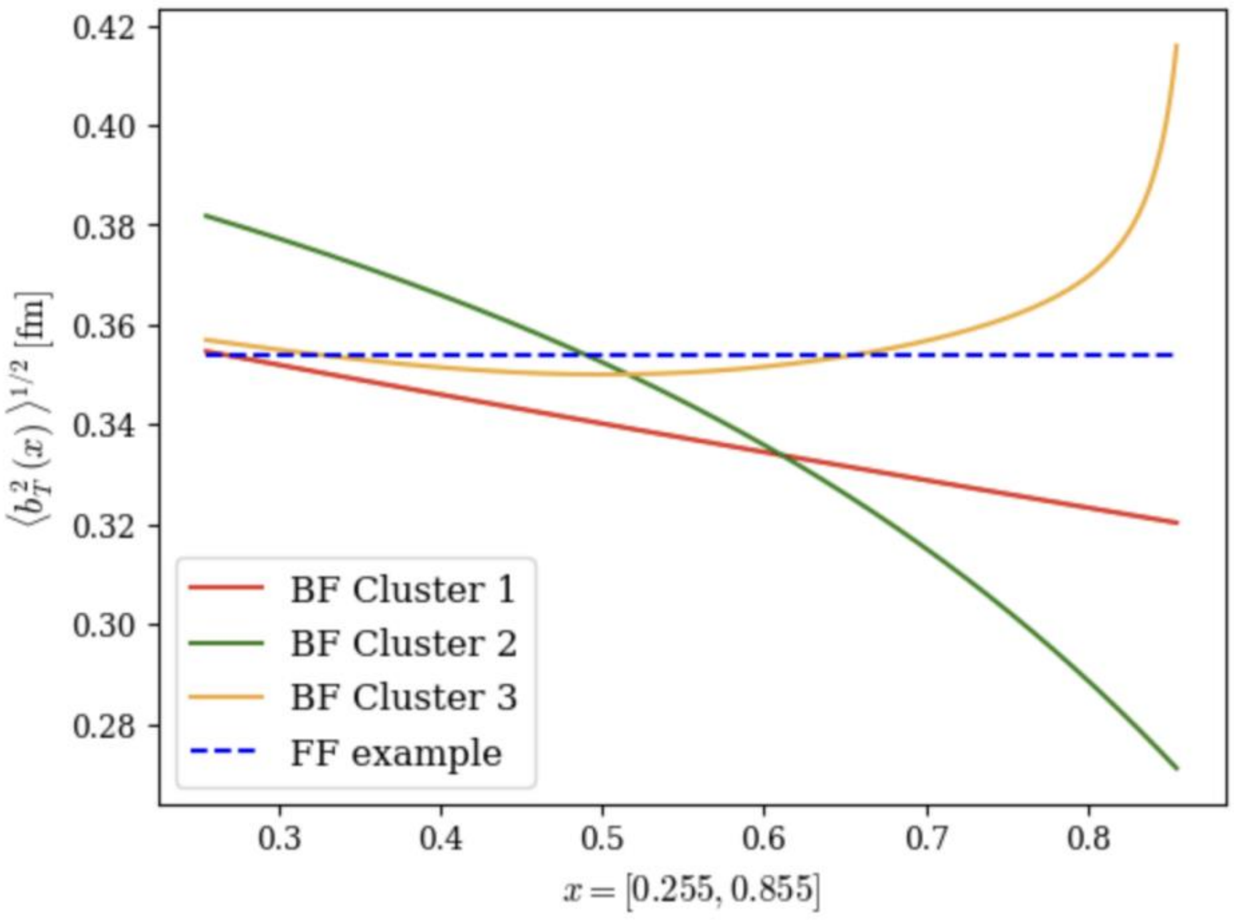


The extrapolation to large momentum transfer Δ_T affects the small coordinate b_T

Coordinate space results

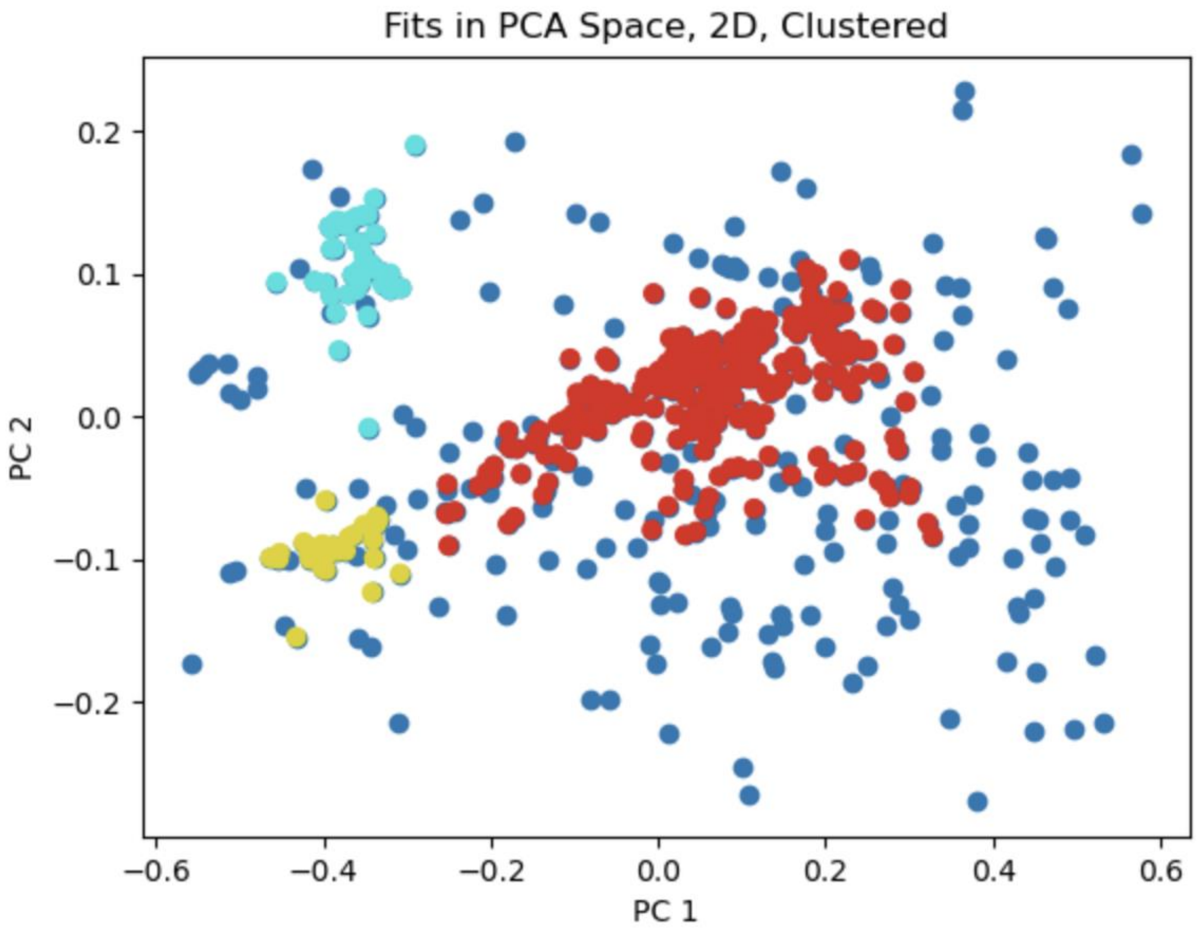


Coordinate space results



- The expectation from standard QCD phenomenology is that as x goes to 1, the average radius goes to 0, but notice how the factorized result is a constant in x, and BF 3 increases with increasing x.
- We have made minimal physics assumptions here!
  - Future work will incorporate more constraints

Improvements in the Symbolic Convergence



Brannon Semp

- Expand each function to 20th order in a Legendre basis

$$f(x) = \sum_{n=0}^{\infty} c_n L_n(2x-1)$$

$$c_n = \frac{2n+1}{2} \int_0^1 f(x) L_n(2x-1)\, \mathrm{d}x$$

- Principal component analysis is performed to further reduce dimensionality
- Fits are clustered in PCA space using HDBSCAN algorithm

Conclusions and outlook

- ML is an exciting tool for extracting GPDs from LQCD and experimental data, and phenomenology provides a much-needed benchmark for our ML results
- Future work includes improving the symbolic convergence criteria and applying these to other data
    - Gluon Ioffe time distributions from the SR with lattice data, Emmanuel Ortiz-Pacheco, ZP, et al. (EXCLAIM)

Backup

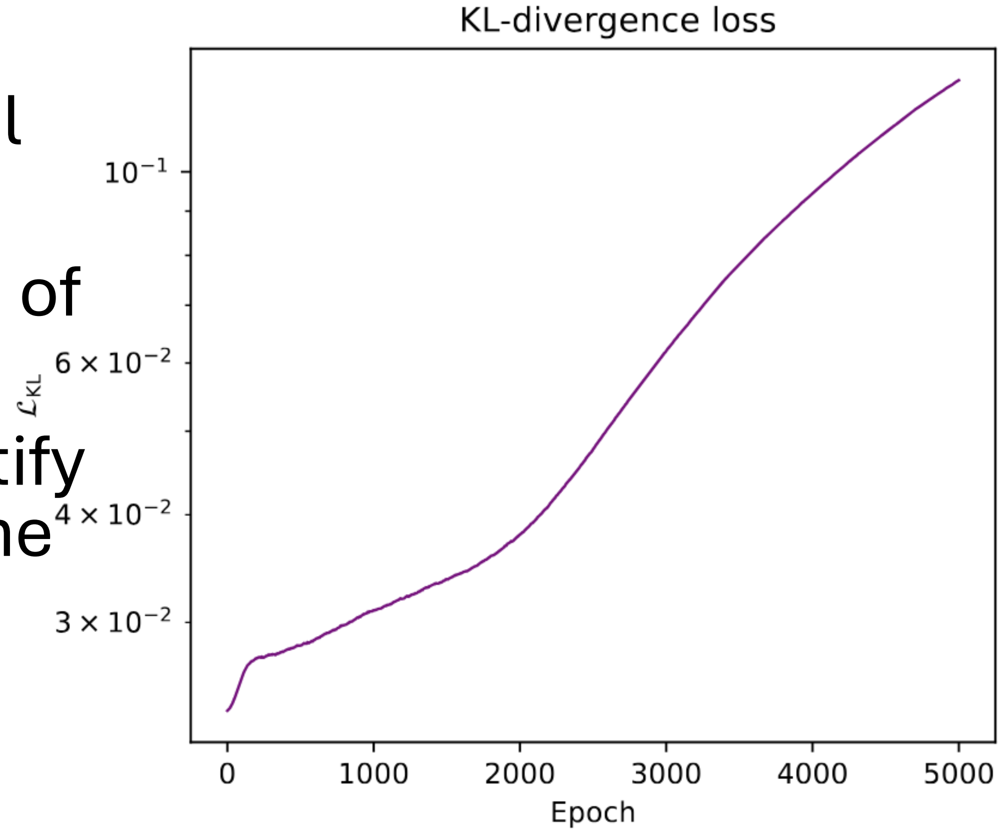
Statistics of the Bayesian Neural Network (BNN)

Epistemic uncertainty – error on the model

- In a BNN, the weights and biases of the network treated as distributions, instead of as fixed points
- Introduce Kullback-Leibler term to quantify the uncertainty due to degeneracies in the solution space

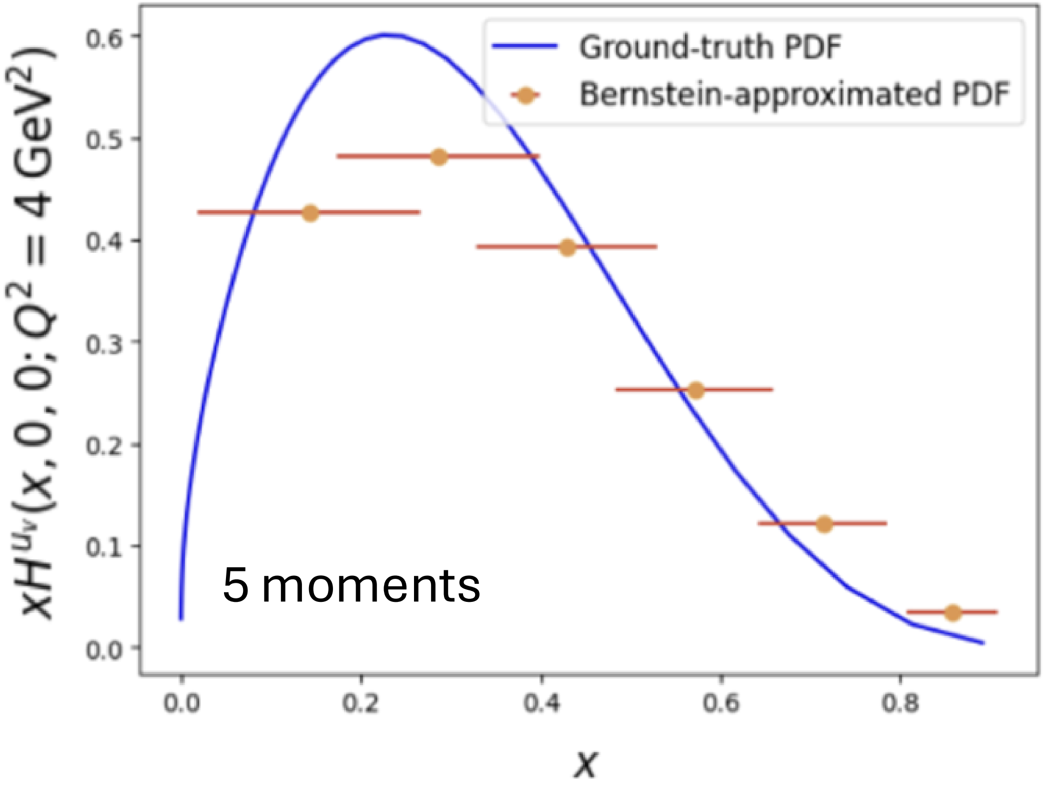
$$\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}\big(q_{\boldsymbol{\theta}}(\mathbf{W}) \,\|\, p(\mathbf{W})\big)$$

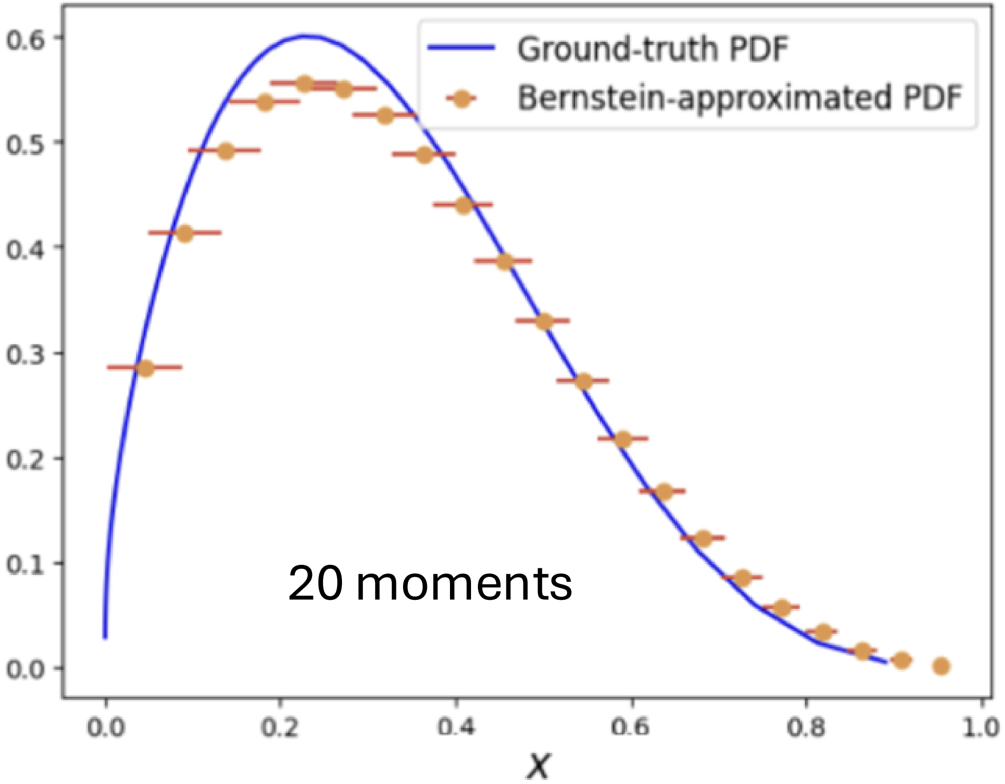
- Aleatoric uncertainty – error on the data
  - In the loss, the errors on the data points are sampled, which can easily include a covariance matrix



Why jump to ML?

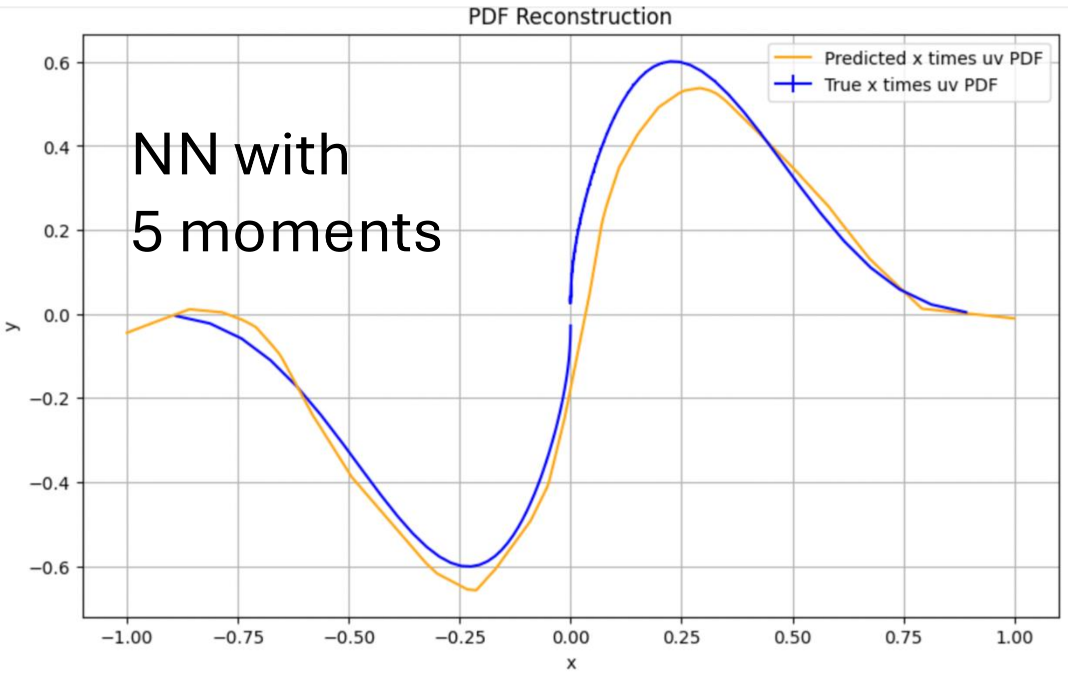
See, e.g., Ahmad et al., *Eur. Phys. J. C* (2009) 63: 407–421 for non-zero skewness GPDs, and Yndurain, PLB (1978) for DIS structure functions

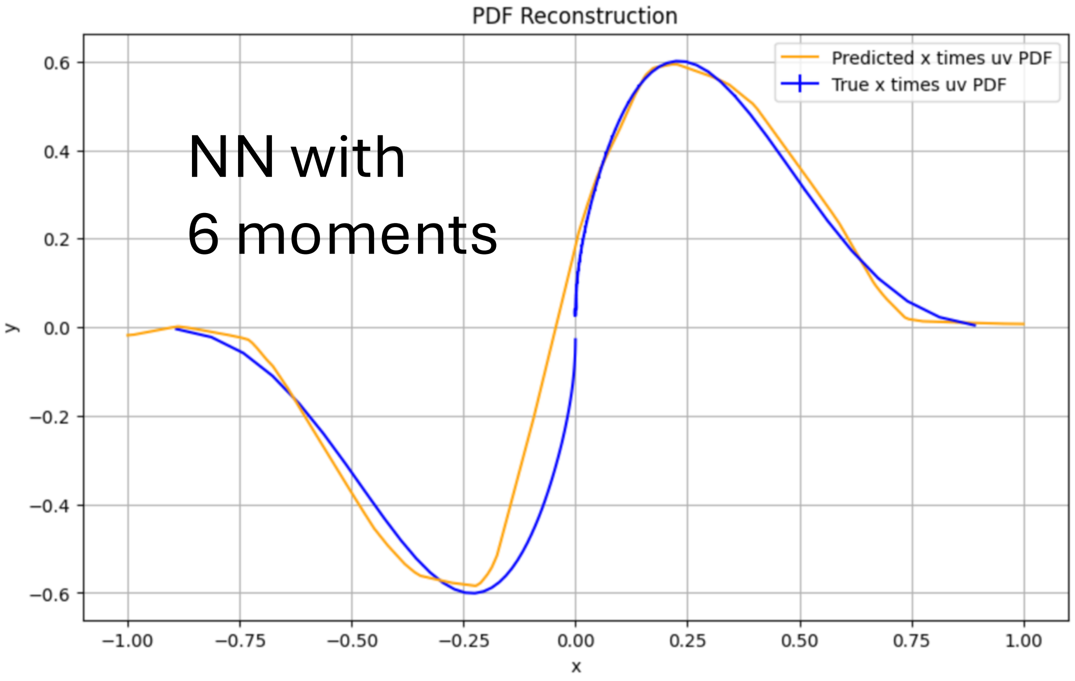


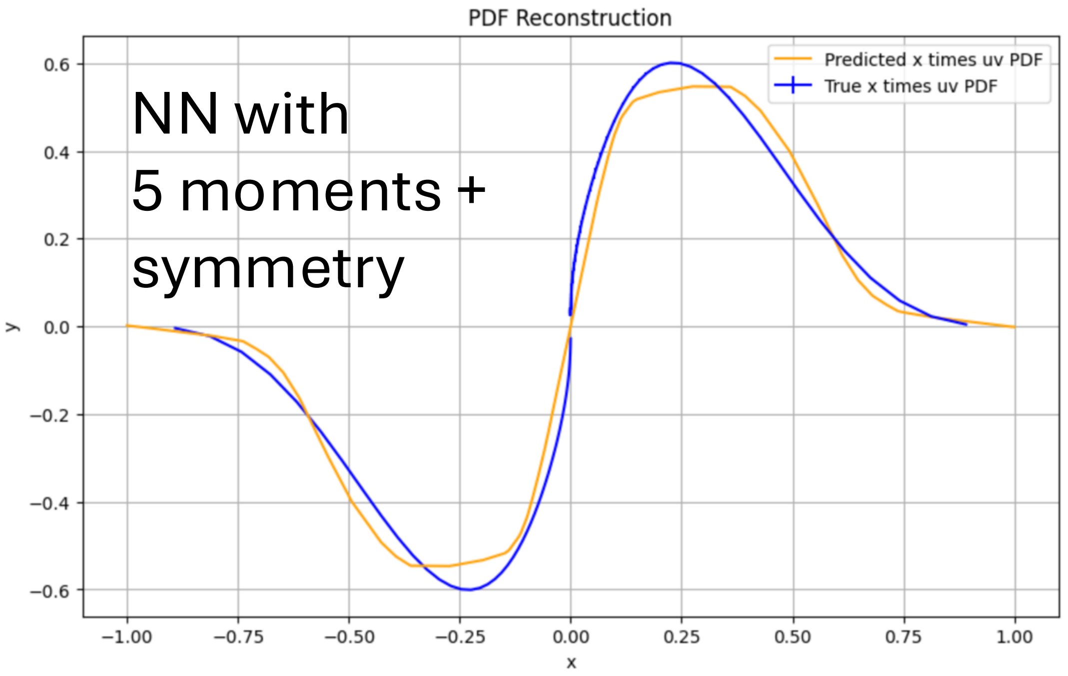


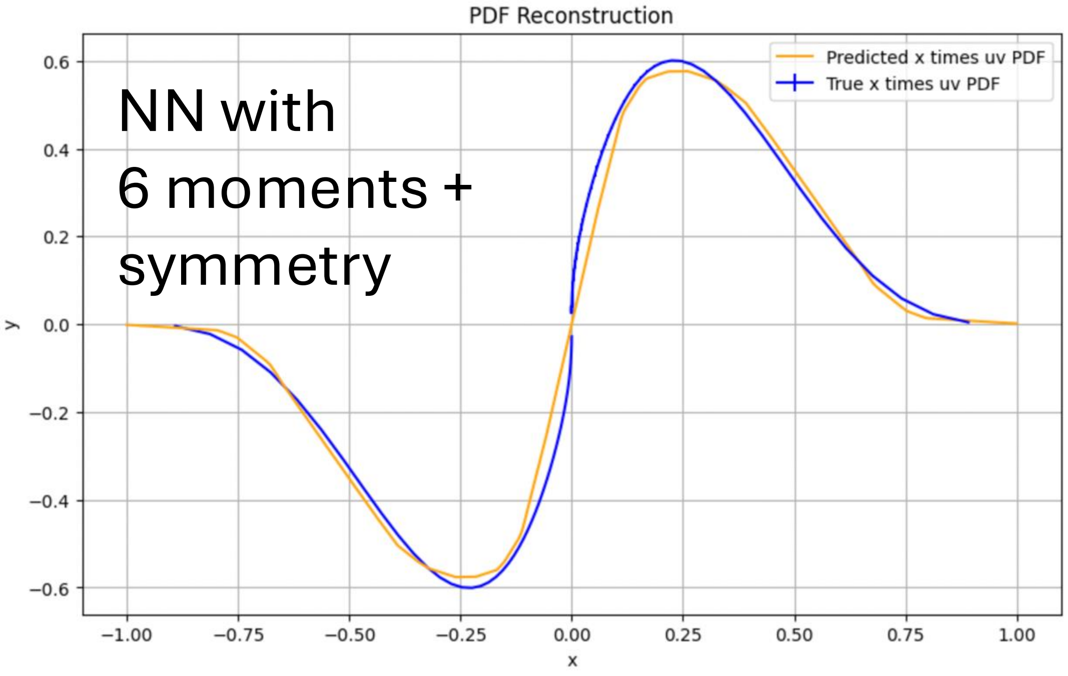
- To get a proper resolution of the x dependence, many moments are required. But state-of-the-art lattice calculations only go up to around 5 or 6 moments!
- Has potential for use as a prior in our network

E.g., Bhattacharya et al., PRD 108, 014507 (2023)

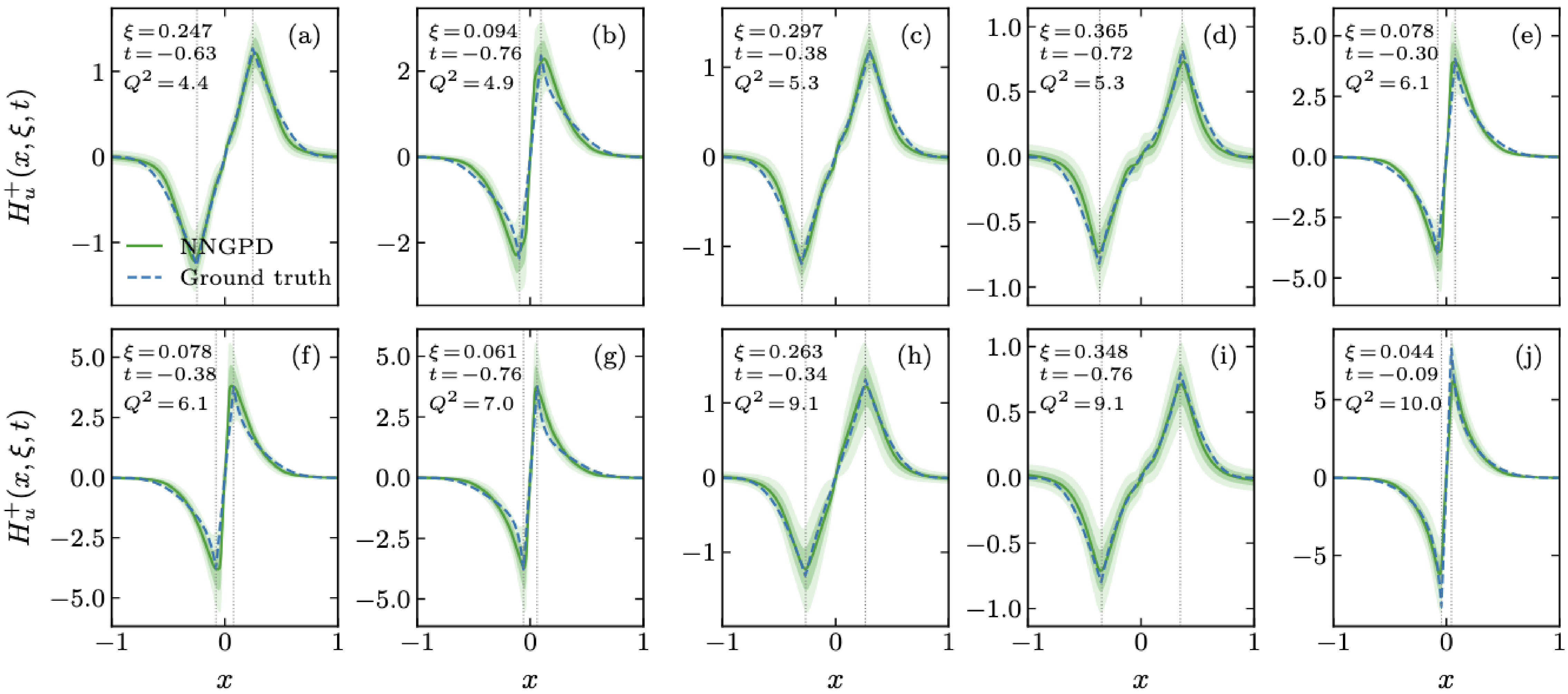






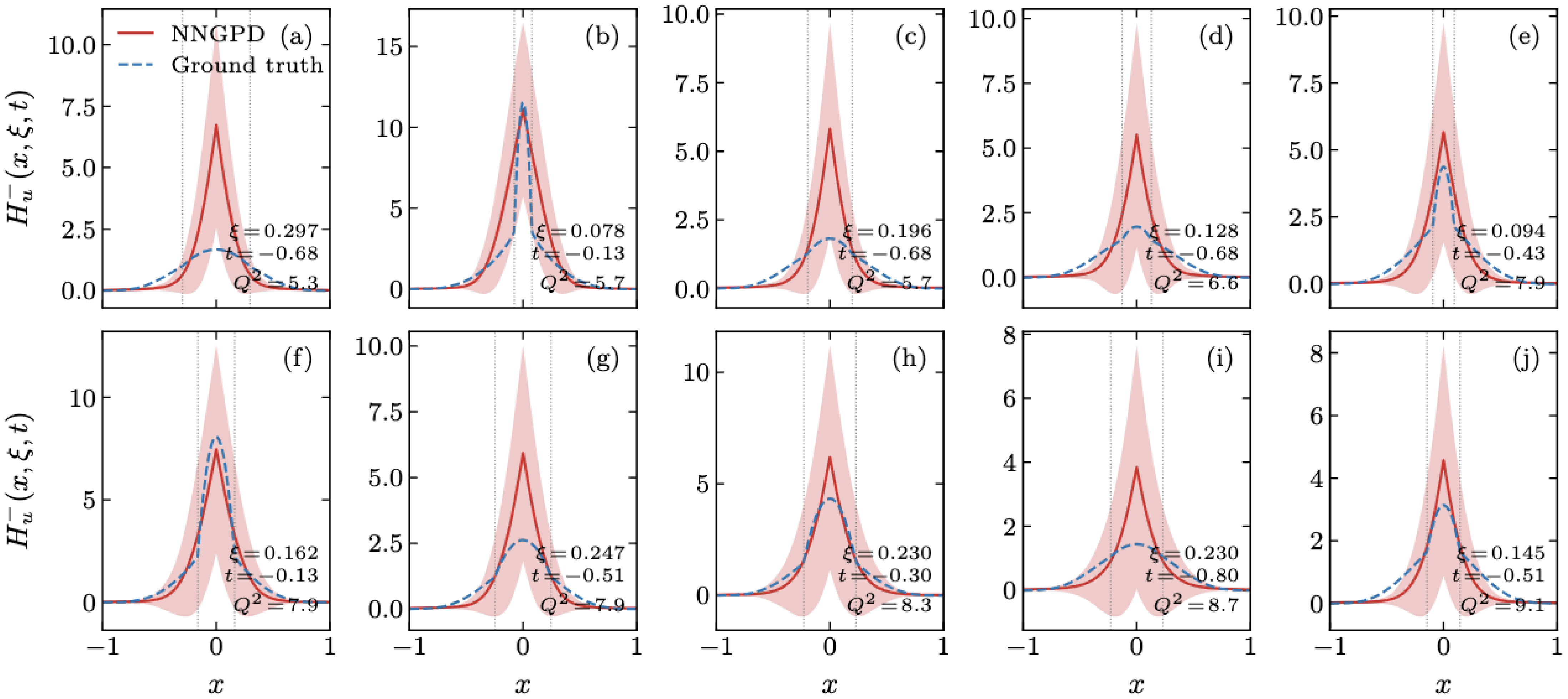


Extracting a signal in the PDF case



- Closure test performed on UVA2
- The “+” distribution is well-captured by the network over a wide range of kinematics, inclusive of JLab and EIC

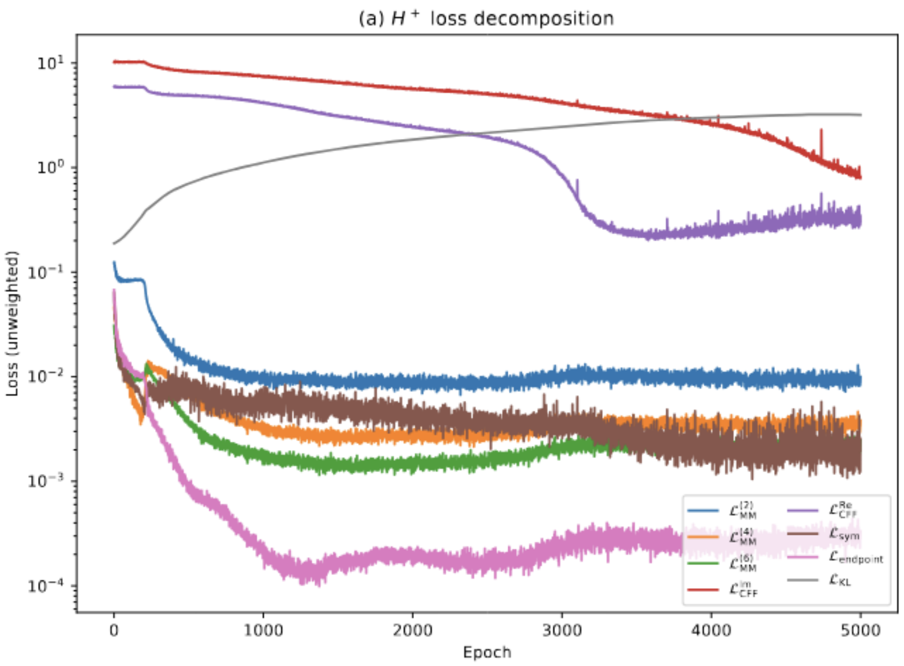
$$\mathcal{L}^{+}_{\text{total}} \;=\; w_{CFF}\mathcal{L}_{CFF} + w^{+}_{MM}\mathcal{L}^{+}_{MM} + w^{+}_{EP}\mathcal{L}^{+}_{EP} + w_{KL}\mathcal{L}_{KL}$$

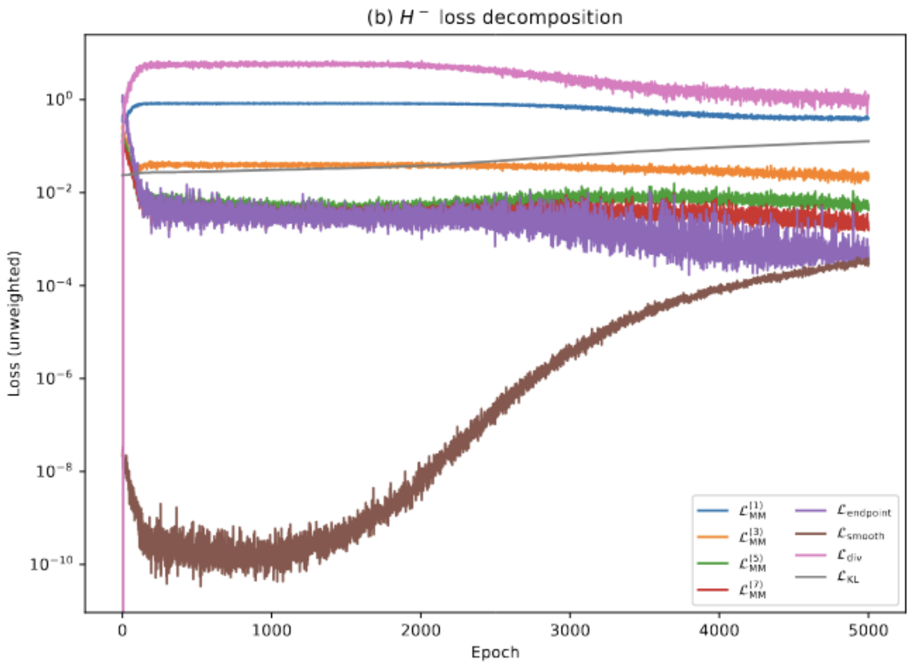


- Closure test performed on UVA2
- The “-” distribution is less well-captured, since the CFF projects out only the ”+”

$$\mathcal{L}_{\text{total}}^{-} \;=\; w_{MM}^{-}\mathcal{L}_{MM}^{-} + w_{EP}^{-}\mathcal{L}_{EP}^{-} + w_{\text{smooth}}\mathcal{L}_{\text{smooth}} + w_{\text{diversity}}\mathcal{L}_{\text{div}} + w_{KL}\mathcal{L}_{KL}$$

Loss decomposition of NNGPD





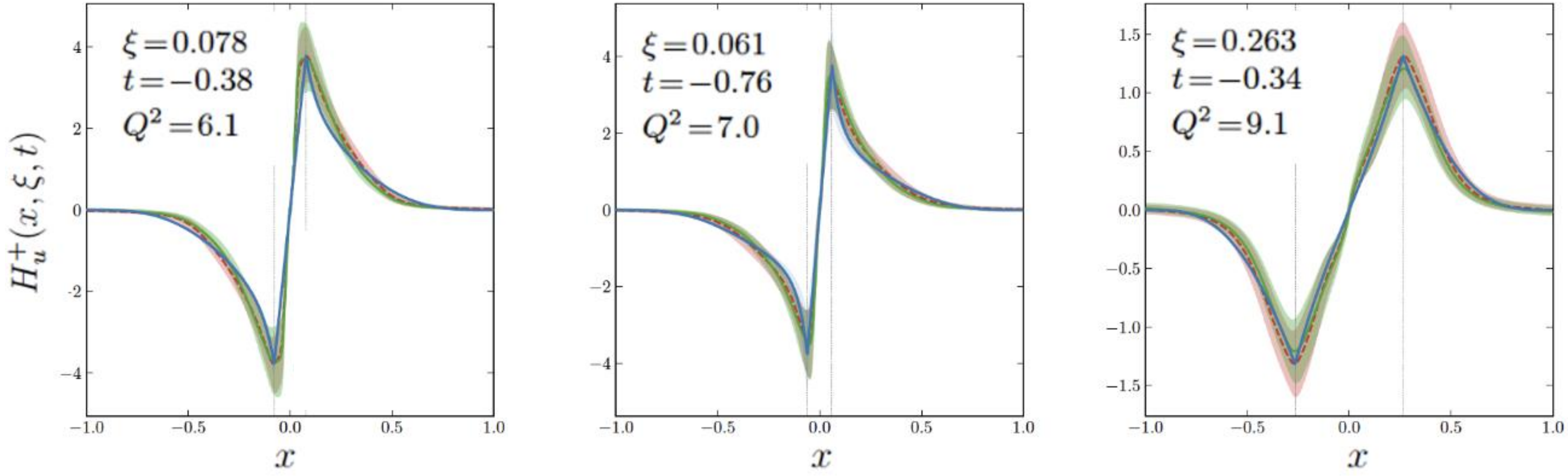
$$\mathcal{L}_{\mathrm{smooth}} = \sum_m \int dx \left| \frac{\partial \hat{H}_q^-}{\partial x} \right|^2$$

$$\mathcal{L}_{\mathrm{div}} = -\mathbb{E}\left[\log \mathrm{Var}(\hat{H}_q^-)\right]$$

Need the diversity term since the CFF doesn't enter the optimization of the H^-

- Challenging to manage the various sources of loss in the network
- Requires the selection of weights in the loss proportional to magnitude of the data

Uncertainty quantification



- Blue band: UVA2 parameterization error estimate
- Green band: BNN trained on errorless model (epistemic)
- Red band: BNN trained on error-full model (epistemic+aleatoric)