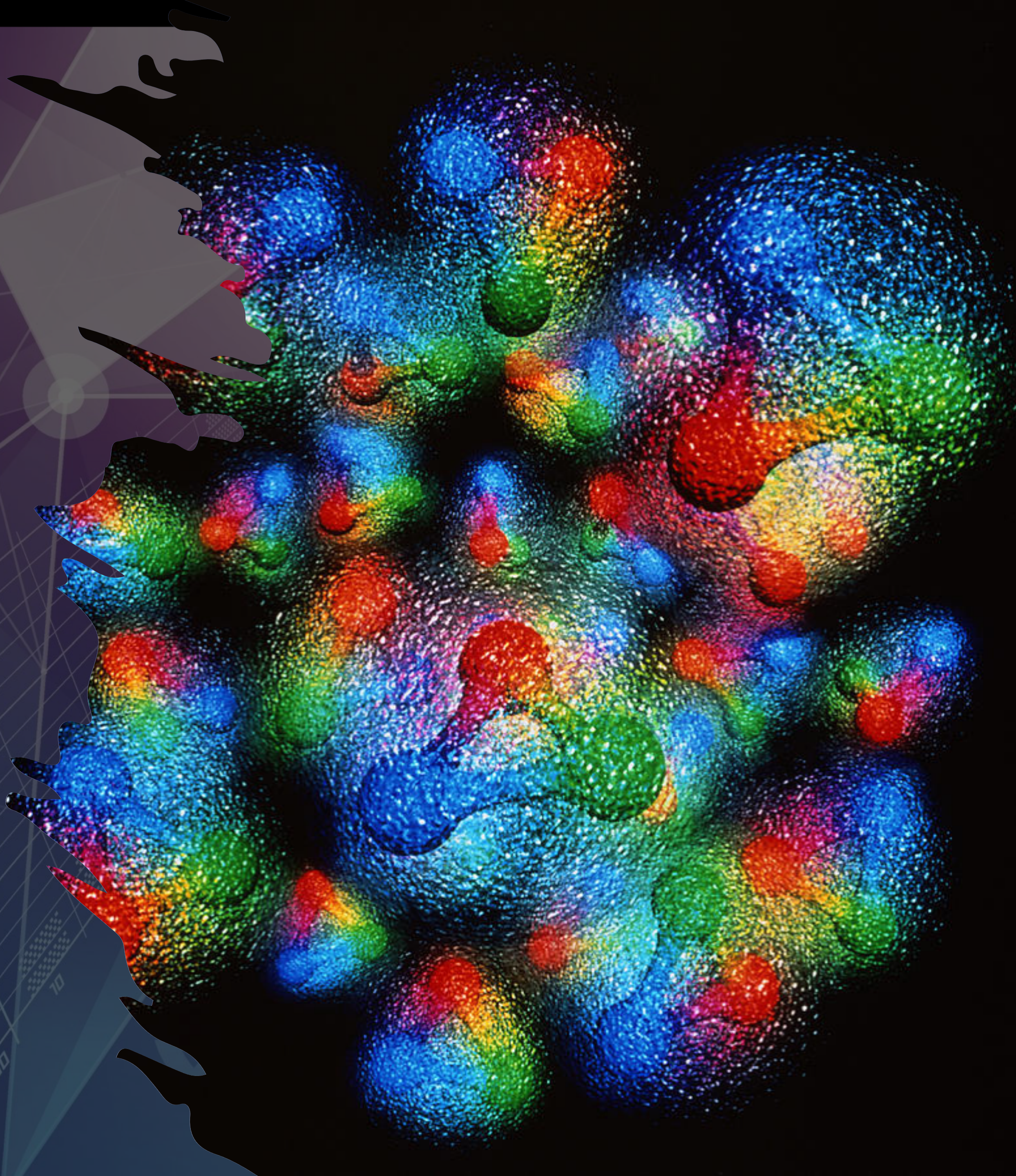
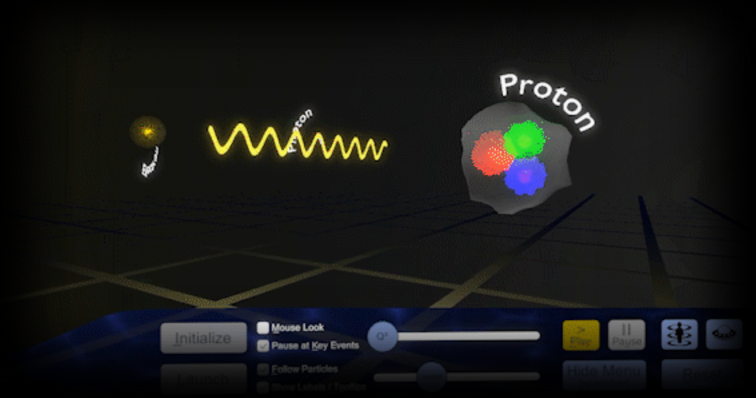
Proton 3D structure from AI : Highlights from the EXCLAIM collaboration

Simonetta Liuti



"The EIC will be a particle accelerator that collides electrons with protons and nuclei to produce snapshots of those particles' internal structure—like a CT scanner for atoms. The electron beam will reveal the arrangement of the quarks and gluons that make up the protons and neutrons of nuclei."

https://www.bnl.gov/eic/

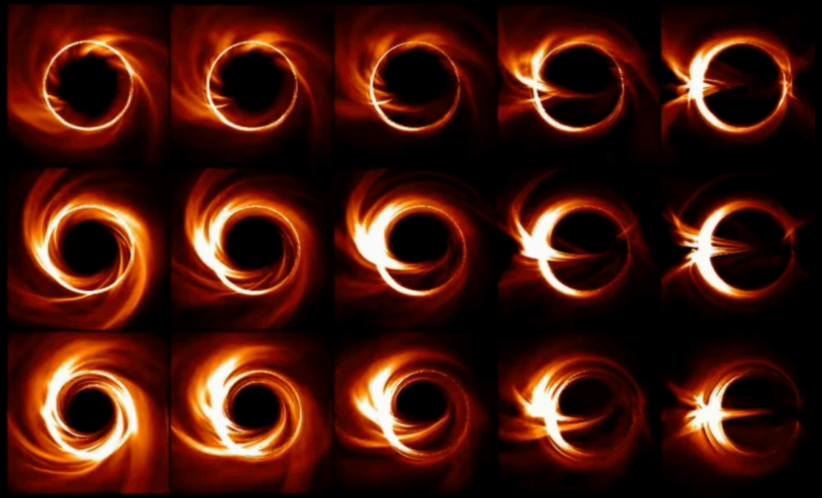


The importance of imaging

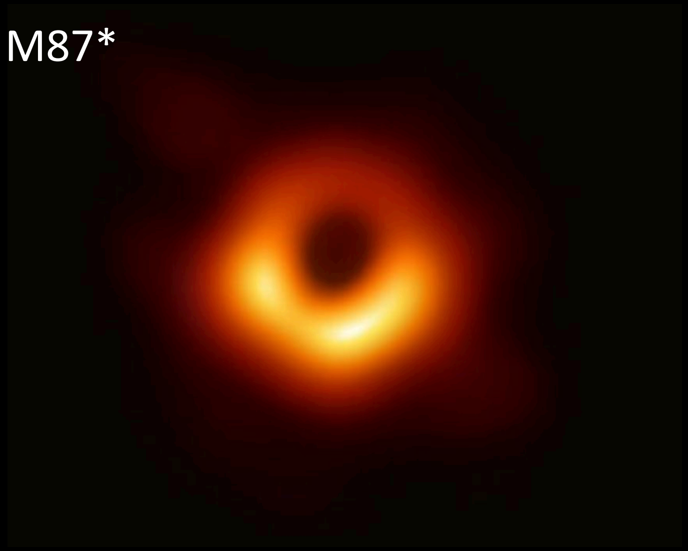
- **<u>One instance that we are well aware of</u>**: The Event Horizon Telescope (EHT) imaged and object 55 M light-years away= 5×10^{23} m

- But what is the science that goes into imaging the proton, observing its spatial structure at 10^{-15} m?

SgsA*



M87*



Horton does not use Uncertainty Quantification

The EXCLAIM collaboration

PIs: Marie Boer, Gia-Wei Chern , Michael Engelhardt, Gary Goldstein, Yaohang Li, Huey-Wen Lin, SL, Matt Sievert, Dennis Sivers

Current Postdocs: Douglas Adams, Marija Cuic,Liam Hockely, Saraswati Pandey, Emanuel Ortiz, Kemal Tegzin

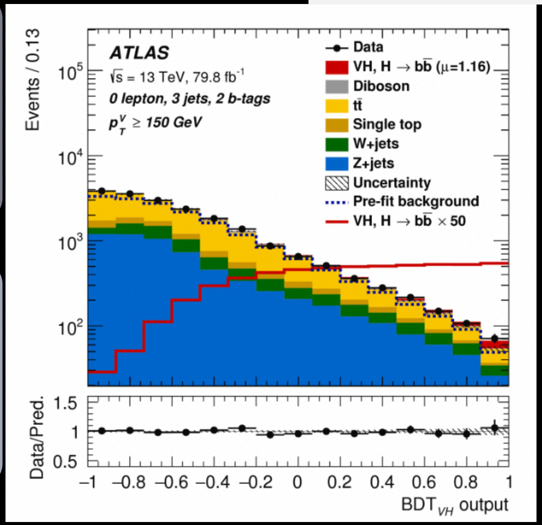
Current Students: Andrew Dotson, Carter Gustin, Jang (Jason) Ho, Fayaz Hossen, Adil Khawaja, Zaki Panjsheeri, Anusha Singireddy

Thanks to the EXCLAIM collaborators, Douglas Adams and Yaohang Li

$$H \to b\bar{b}$$

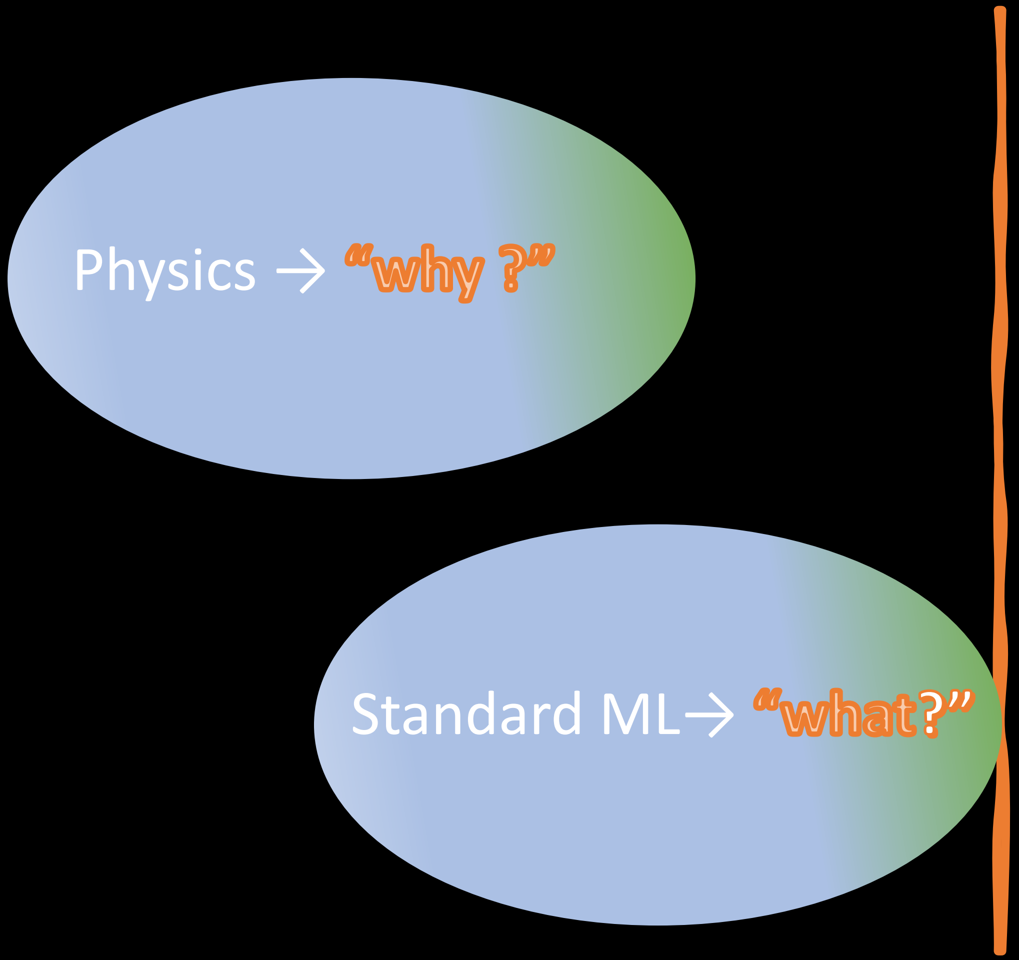
<u>Standard Approaches</u>: Industrial Machine Learning (ML) tools are used/adapted to aid computation in Nuclear/Particle Physics

<u>Example</u>: Ensemble learning methods such as Boosted Decision Trees (BDT) invented for image recognition/object detection used in self-driving cars are used to identify b-hadrons



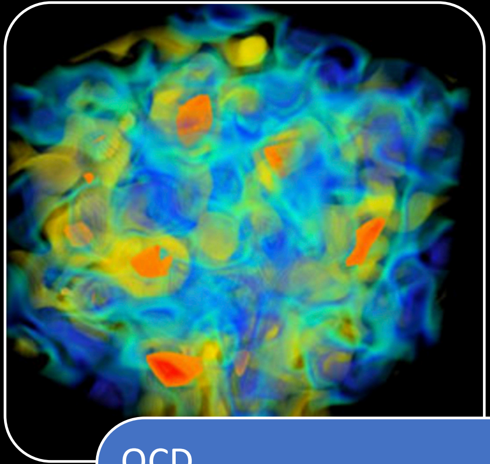
https://atlas.cern/

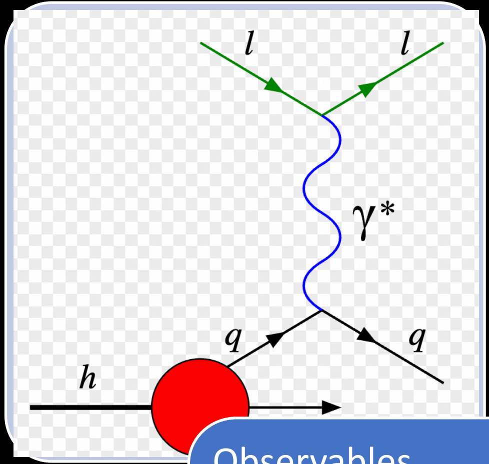
To address the “why”?

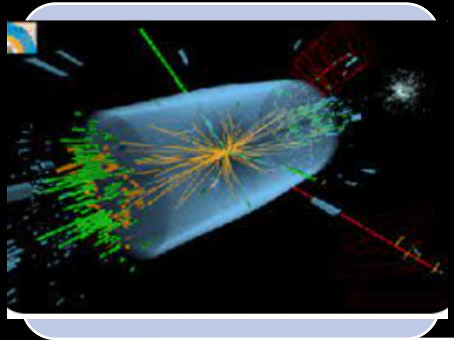


- We introduce physics aware NNs as explainable ML models: C-VAIM
- Symbolic Regression: ML algorithm where data are modeled directly with analytic expressions. Direct interpretability
- Explainable and interpretable models are necessary for the 3D nuclear problem directly enabling discovery laws in Nuclear and Particle Physics
- Not just a set of advanced computational tools: It is about finding a common language between physics and AI

Forward Problem







QCD

- quark field
- gluon field
- correlation functions

Observables

- Form Factors
- Structure Functions
- Compton Form Factors
- Fragmentation functions
- …

Data/Measurements

Inverse Problem

- **Interpretability**

The goal in physics is to extract information as accurately as possible from data

- **Predictivity**

The goal of ML is to obtain statistical models that can make predictions from the data

An immense potential!

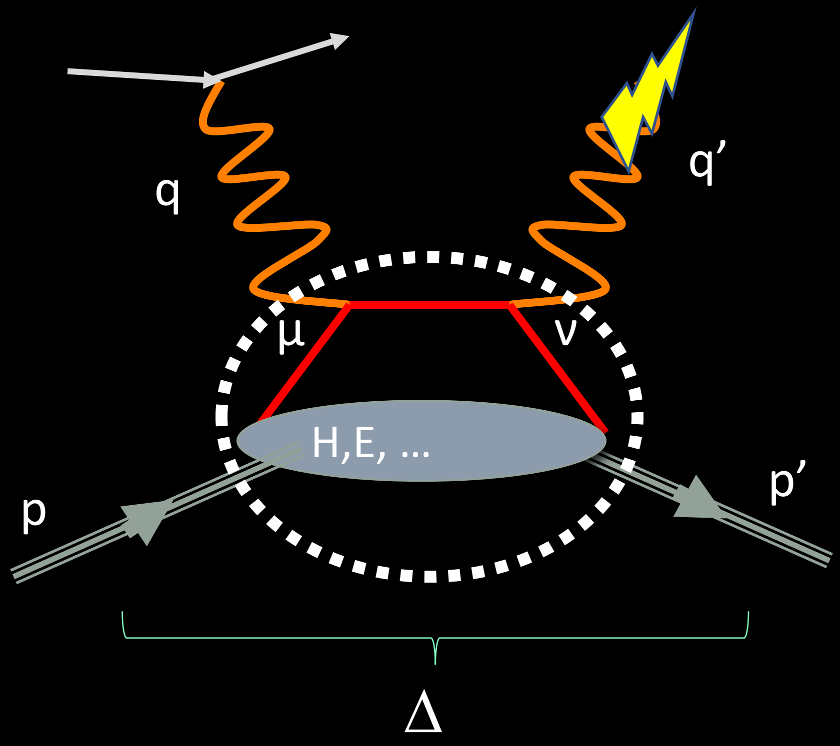
Through ML we will be able to see the emergence of **new physics relations/laws**

- **Inverse problem**

To address this we need to define a bridge between CS experts and physicists that is centered on how we define and treat the respective data uncertainty and correlations

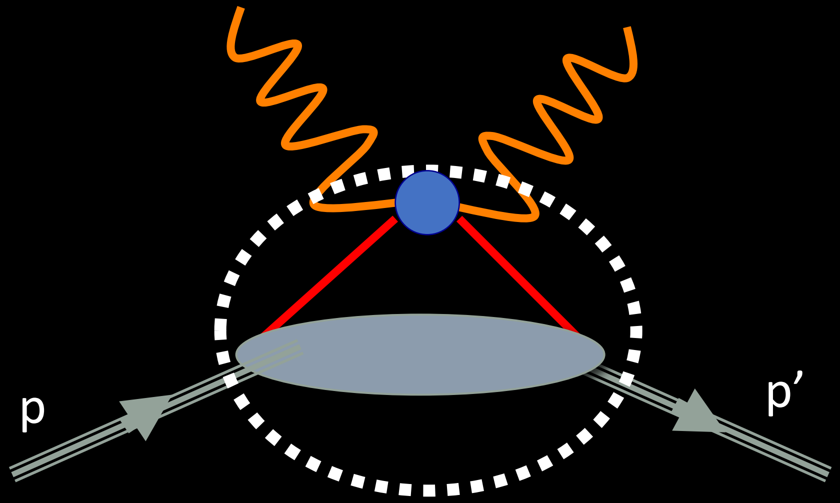
Physics Case: Extracting information from exclusive deeply virtual scattering

QCD matrix element



Generalized Parton Distribution

Local Operator



GPD Moments → (EMT) Form Factors

Twist three GPD Physical interpretation at the core of spin puzzle

$$\begin{array}{ccccc} J_L & = & L_L & + & S_L \\ \frac{1}{2}\int dx\, x(H+E) & = & \boxed{\begin{array}{c}\int dx\, x(\widetilde{E}_{2T}+H+E) \\ -\int dx\, F_{14}^{(1)}\end{array}} & + & \frac{1}{2}\int dx\, \widetilde{H} \\ & = & & + & \frac{1}{2}\int dx\, \widetilde{H} \end{array}$$

A. Rajan, M. Engelhardt and S. Liuti, Phys. Rev. D98, 074022 (2018)
A. Rajan, A. Courtoy, M. Engelhardt and S. Liuti, Phys. Rev. D94, 034041 (2016)
M. Rodekamp, M. Engelhardt, J.R. Green, S. Krieg, S. Liuti, S. Meinel, J.W. Negele, A. Pochinsky and S. Syritsyn, Phys. Rev. D 109, 074508 (2024)

*Twist 3 GPD notation from Meissner, Metz and Schlegel, JHEP(2009)

Transverse Angular Momentum Sum Rule

O. Alkassasbeh, M. Engelhardt, SL and A. Rajan, https://arxiv.org/abs/2410.21604

$$\underset{J_T}{\underbrace{\frac{1}{2}\int dx\, x(H+E) - \frac{1}{2}\int dx\, \mathcal{M}_T}} = \underset{L_T}{\underbrace{\int dx\, x\Big(\widetilde{E}_{2T} + H + E + \frac{H_{2T}}{\xi}\Big)}} + \underset{S_T}{\underbrace{\frac{1}{2}\int dx\, g_T - \frac{1}{2}\int dx\, x\mathcal{A}_T}}$$

How do we separate twist two and twist three components?

Twist 3 GPDs Physical Interpretation

	GPD	$P_q P_p$	TMD	Ref. 1
	$H^\perp$	UU	$f^\perp$	$2\widetilde{H}_{2T} + E_{2T}$
	$\widetilde{H}_L^\perp$	LL	$g_L^\perp$	$2\widetilde{H}'_{2T} + E'_{2T}$
J_L	$H_L^\perp$	UL	$f_L^{\perp\,(*)}$	$\widetilde{E}_{2T} - \xi E_{2T}$
J_L	$\widetilde{H}^\perp$	LU	$g^{\perp\,(*)}$	$\widetilde{E}'_{2T} - \xi E'_{2T}$
J_T	$H_T^{(3)}$	UT	$f_T^{(*)}$	$H_{2T} + \tau \widetilde{H}_{2T}$
J_T	$\widetilde{H}_T^{(3)}$	LT	g'_T	$H'_{2T} + \tau \widetilde{H}'_{2T}$

- B. Kriesten and S. Liuti, *Phys.Rev. D105 (2022),* arXiv 2004.08890

1/Q correction to H

1/Q correction to $\widetilde{\mathrm{H}}$

NEW!! Orbital Angular Momentum L

NEW!! Spin Orbit correlation $L \cdot S$

NEW!! Transverse OAM L_T

Transverse spin

(*) T-odd

[1] Meissner, Metz and Schlegel, JHEP(2009)

8/8/23

A. Rajan, A. Courtoy, M. Engelhardt, S.L., PRD (2016)

20

A. Rajan, M. Engelhardt, S.L., PRD (2018)

A. Rajan, O. Alkassasbeh, M. Engelhardt, S.L., (2023)

Extract Compton form factors from Leading order parametrization of DVCS cross section

$$|T_{UU}^{BH}|^2 = \frac{\Gamma}{t}\left[A_{UU}^{BH}\boxed{(F_1^2+\tau F_2^2)} + B_{UU}^{BH}\boxed{\tau G_M^2(t)}\right]$$

$$|T_{UU}^{\mathcal{I}}|^2 = \frac{\Gamma}{Q^2 t}\left[A_{UU}^{\mathcal{I}}\Re e\boxed{(F_1\mathcal{H}+\tau F_2\mathcal{E})} + B_{UU}^{\mathcal{I}}G_M\Re e\boxed{(\mathcal{H}+\mathcal{E})} + C_{UU}^{\mathcal{I}}\boxed{G_M\Re e\widetilde{\mathcal{H}}}\right]$$

$$|T_{LU}^{\mathcal{I}}|^2 = \frac{\Gamma}{Q^2 t}\left[A_{LU}^{\mathcal{I}}\Im m\boxed{(F_1\mathcal{H}+\tau F_2\mathcal{E})} + B_{LU}^{\mathcal{I}}G_M\Im m\boxed{(\mathcal{H}+\mathcal{E})} + C_{LU}^{\mathcal{I}}\boxed{G_M\Im m\widetilde{\mathcal{H}}}\right]$$

Azimuthal angle ϕ dependent coefficients

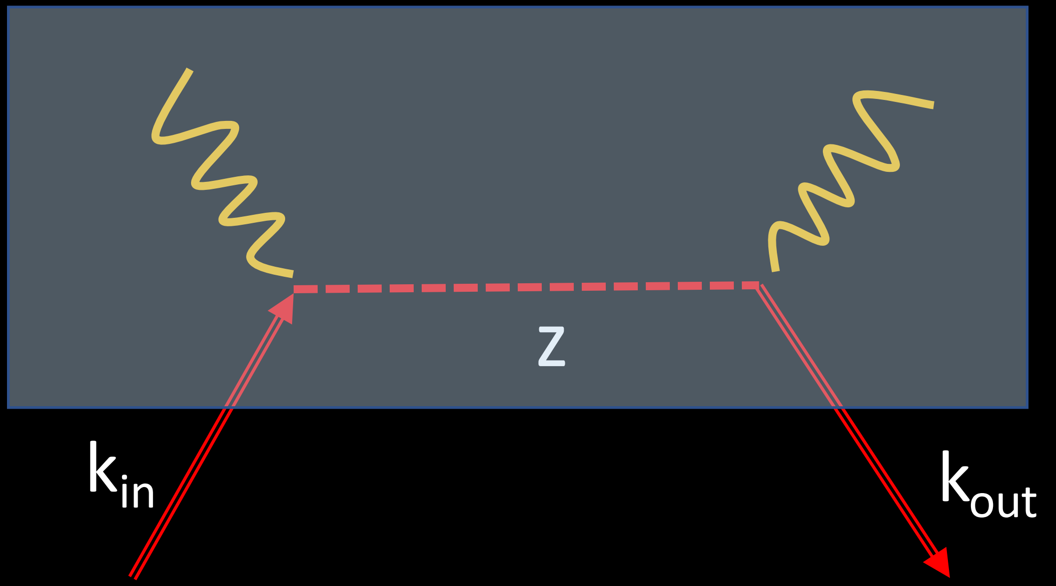
$$\begin{aligned}|T_{UU}^{DVCS}|^2 = \frac{\Gamma}{Q^2}\frac{2}{1-\epsilon}\Bigg[&(1-\xi^2)\left[(\Re e\mathcal{H})^2+(\Im m\mathcal{H})^2+(\Re e\widetilde{\mathcal{H}})^2+(\Im m\widetilde{\mathcal{H}})^2\right] \\ &+\frac{t_o-t}{4M^2}\left[(\Re e\mathcal{E})^2+(\Im m\mathcal{E})^2+\boxed{\xi^2(\Re e\widetilde{\mathcal{E}})^2+\xi^2(\Im m\widetilde{\mathcal{E}})^2}\right] \\ &-2\xi^2\left(\Re e\mathcal{H}\,\Re e\mathcal{E}+\Im m\mathcal{H}\Im m\mathcal{E}+\Re e\widetilde{\mathcal{H}}\,\Re e\widetilde{\mathcal{E}}+\Im m\widetilde{\mathcal{H}}\Im m\widetilde{\mathcal{E}}\right)\Bigg]\end{aligned}$$

spin non-flip

spin flip

mixed

- B. Kriesten et al, *Phys.Rev. D* 101 (2020)
- B. Kriesten and S. Liuti, *Phys.Rev. D105 (2022),* arXiv 2004.08890
- B. Kriesten and S. Liuti, Phys. Lett. B829 (2022), arXiv:2011.04484



At leading order in pQCD

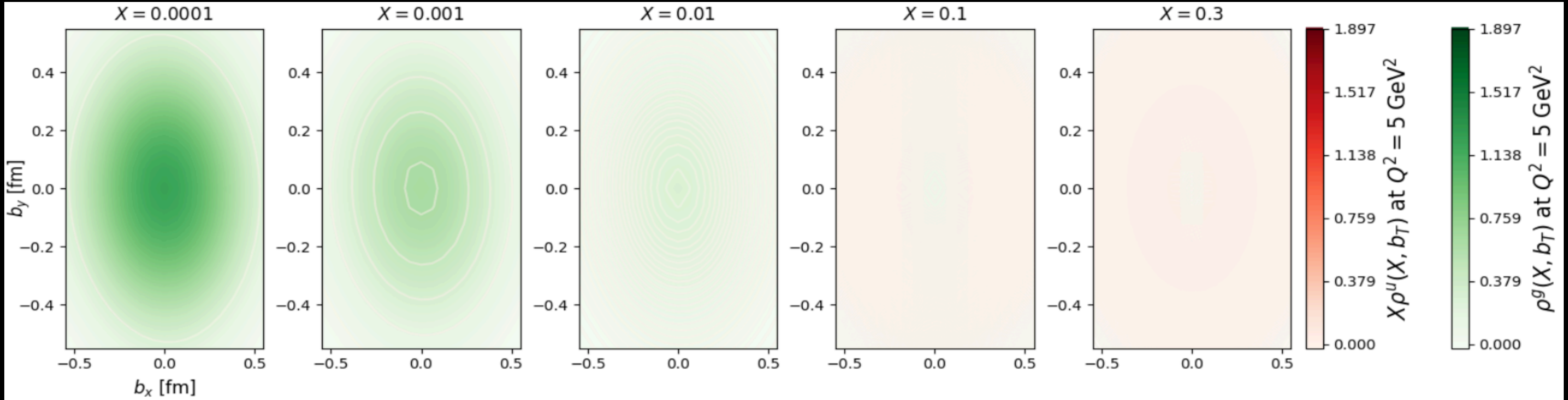
$$\int_{-1}^{1} dX \, \frac{1}{X - \zeta + i\epsilon} = P.V. \int_{-1}^{1} dX \, \frac{1}{X - \zeta} - i\pi\delta(X - \zeta)$$

3D Coordinate Space Representation

Observables from DVES matrix elements can be Fourier transformed from momentum space into coordinate space, providing insight into the spatial distributions of quarks and gluons inside the proton, besides matter and charge distributions.

$$\overbrace{\boxed{\mathcal{H}^q(X,0,b_T)}}^{\text{Wigner phase space distribution}} = \int \frac{d^2\Delta_T}{(2\pi)^2} \underbrace{\boxed{H^q(X,0,\Delta_T)}}_{\text{GPD}} e^{-i\Delta_T\cdot b_T}$$

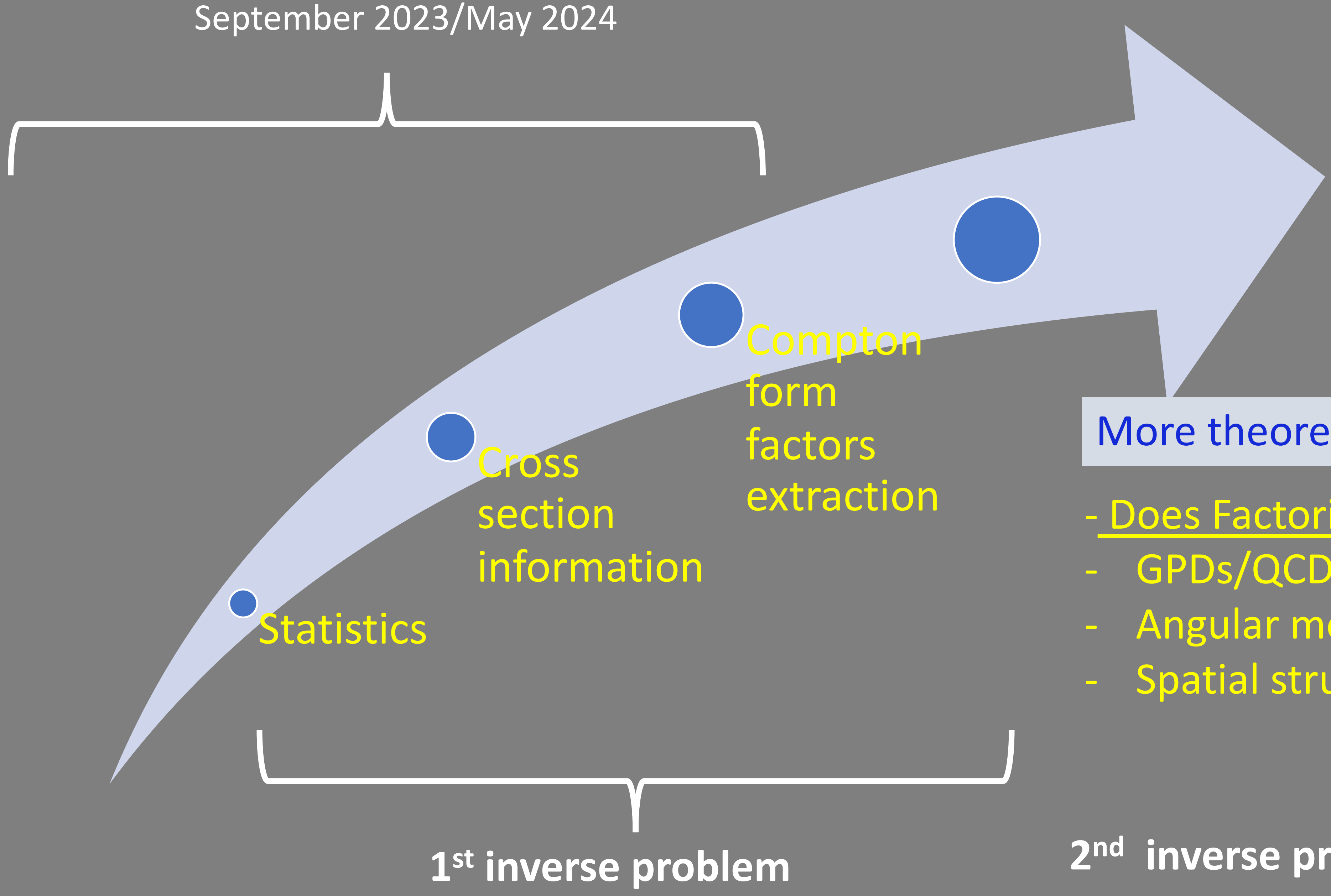
Z. Panjsheeri



GGL

UVA gluon GPD parametrization (from lattice QCD and experiment)

B. Kriesten. P. Velie, E. Yeats, F. Y. Lopez, & S. Liuti, *Phys.Rev.D* 105 (2022) 5, 056022



More theoretical questions

- Does Factorization work?
- GPDs/QCD matrix element
- Angular momentum
- Spatial structure

2nd inverse problem

1. Fully constraining Likelihood analysis

2. Inverse Problem techniques: Variational Autoencoder Inverse Mapper (VAIM)

3. Symbolic Regression for Partonic Observables

All these methods share the common goal of going beyond simple regression by understanding the underlying correlations of the system

Graduate Students:

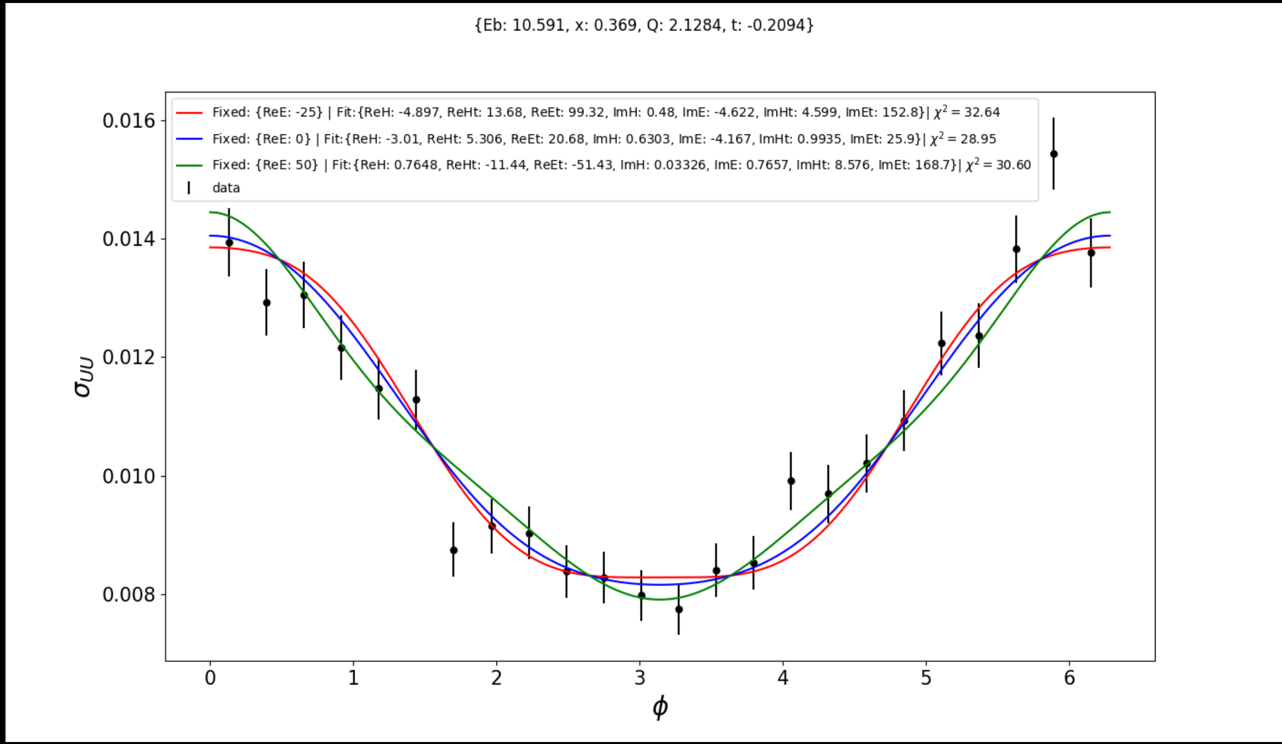
Joshua Bautista, Adil Khawaja, Zaki Panjsheeri

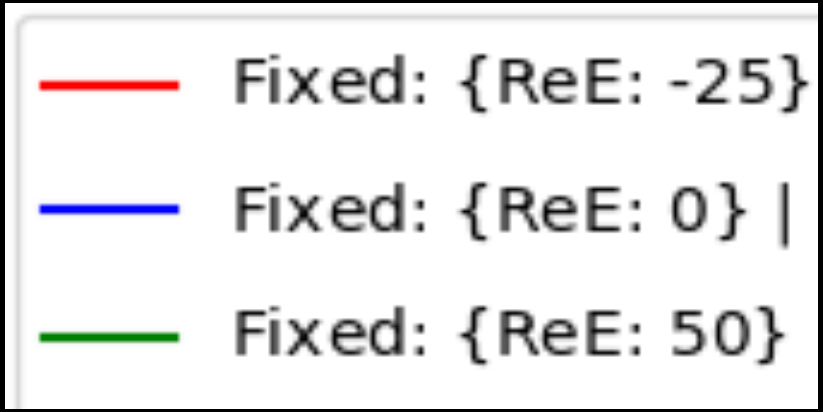
1. Fully Constraining CFFs : Likelihood Analysis

GOAL: Use DVCS data and comparing to cross section model to find CFFs

- We find a CFF result using VAIM: Got some valid CFFs
- Curve fit: A really bad result: Encounter a problem 1!
- Definition of the likelihood: Try to fix the problem
- Canonical Likelihood: Reproduces the problem in explainable way
- Canonical Likelihood Modified: Fix the problem in 2 ways
    - Difference method Likelihood
    - Canonical Likelihood
- Encounter a problem 2!: Poll the audience
- Some results: Table of CFFs and errors

Try a curve fit for a kinematic setup forcing one CFF





Likelihood function: Bayes law

$$\underbrace{\boxed{\vec{X}_{all}\&\vec{\Theta}}_{pdf}(\vec{v}_{x\,all},\vec{v}_{\theta})}_{\underbrace{Joint}_{\text{Omited In Textbooks}}} = \underbrace{\boxed{\vec{\Theta}|\vec{X}_{all}}_{pdf}(\vec{v}_{\Theta}|\vec{v}_{x\,all})}_{=\underbrace{Posterior}_{\text{(all data)}}} \times \underbrace{\boxed{\vec{X}_{all}}_{pdf}(\vec{v}_{x\,all})}_{\times\underbrace{Evidence}_{(=1)}} = \underbrace{\boxed{\vec{X}_{all}|\vec{\Theta}}_{pdf}(\vec{v}_{x\,all},\vec{v}_{\Theta})}_{=\underbrace{Likelihood}_{\text{(all data)}}} \times \underbrace{\boxed{\vec{\Theta}}_{pdf}(\vec{v}_{\Theta})}_{\times\underbrace{Prior}_{\text{(no data)}}}$$

For frequentists: Prior = 1

$$\text{Likelihood} = \boxed{\vec{X}_{all}|\vec{\Theta}}_{pdf} = \prod_i \boxed{\vec{X}_{single}|\vec{\Theta}}_{pdf}(\vec{v}_{x\,i},\vec{v}_{\Theta}) = \text{ is model and experiment error determined}$$

(Canonical)

Canonical Likelihood Derivation

$$\mathcal{L}_{canonical}(\text{parameters}) = \prod_i \text{Gaussian}(x = \sigma_{obs}(\phi_i), \mu = \sigma_{model}(\phi_i), \sigma = Err(\sigma_{obs}))$$

Each data point's error bar:

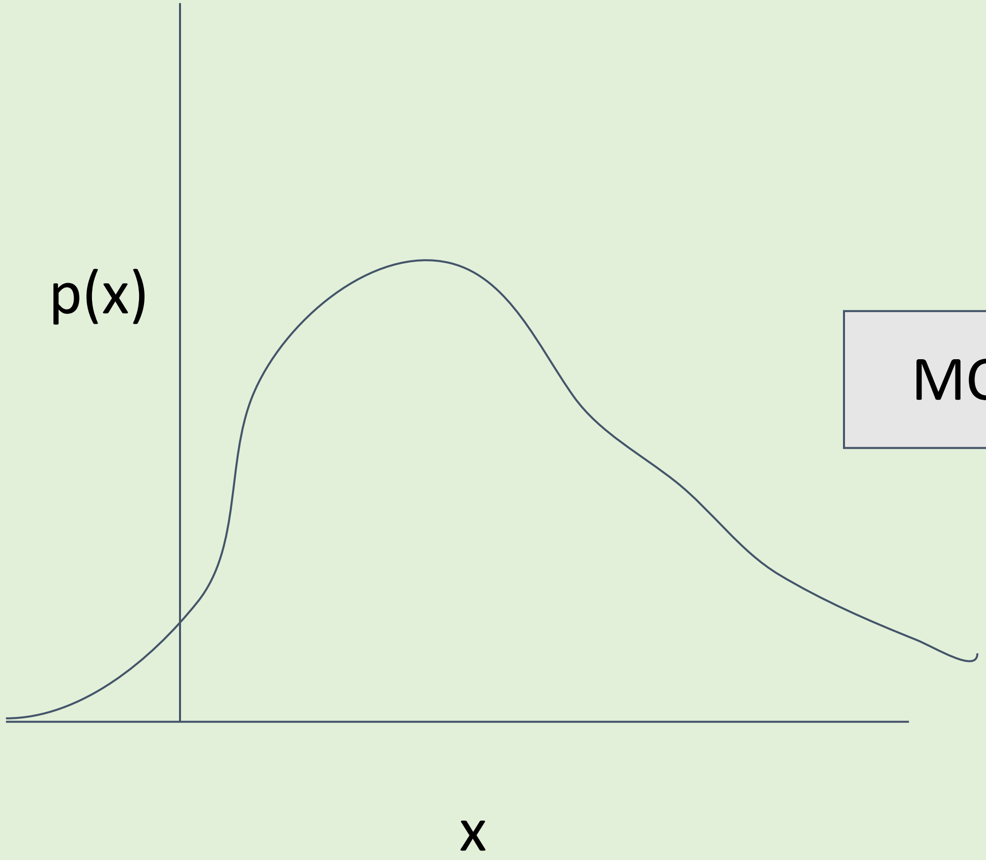
- defines a gaussian
- should explain why the data does not match the model exactly
- (canonically) multiplies to derive a total likelihood function

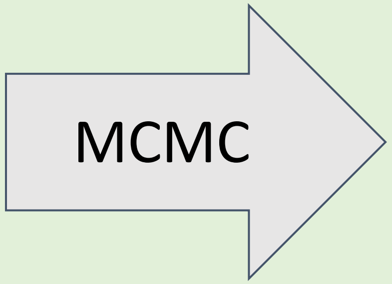
The total likelihood function and a choice of prior:

- uniquely defines a posterior probability density function
- can be used to generate samples (MCMC)

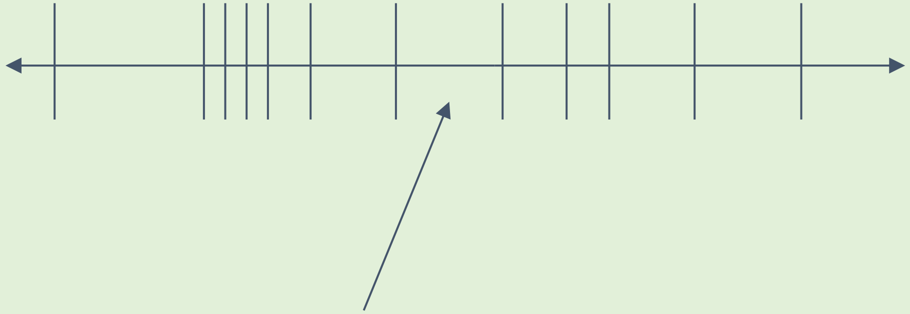
Reminder: What is MCMC?

Start with a probability distribution





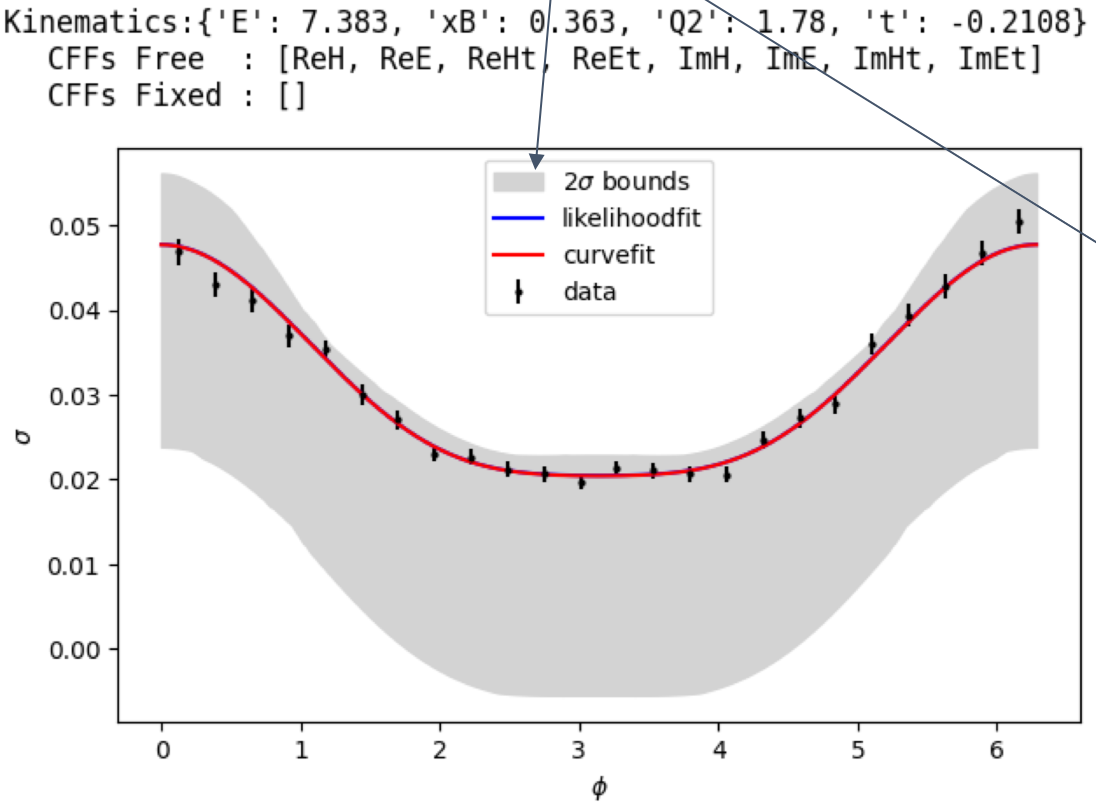
Generate samples which represent that distribution

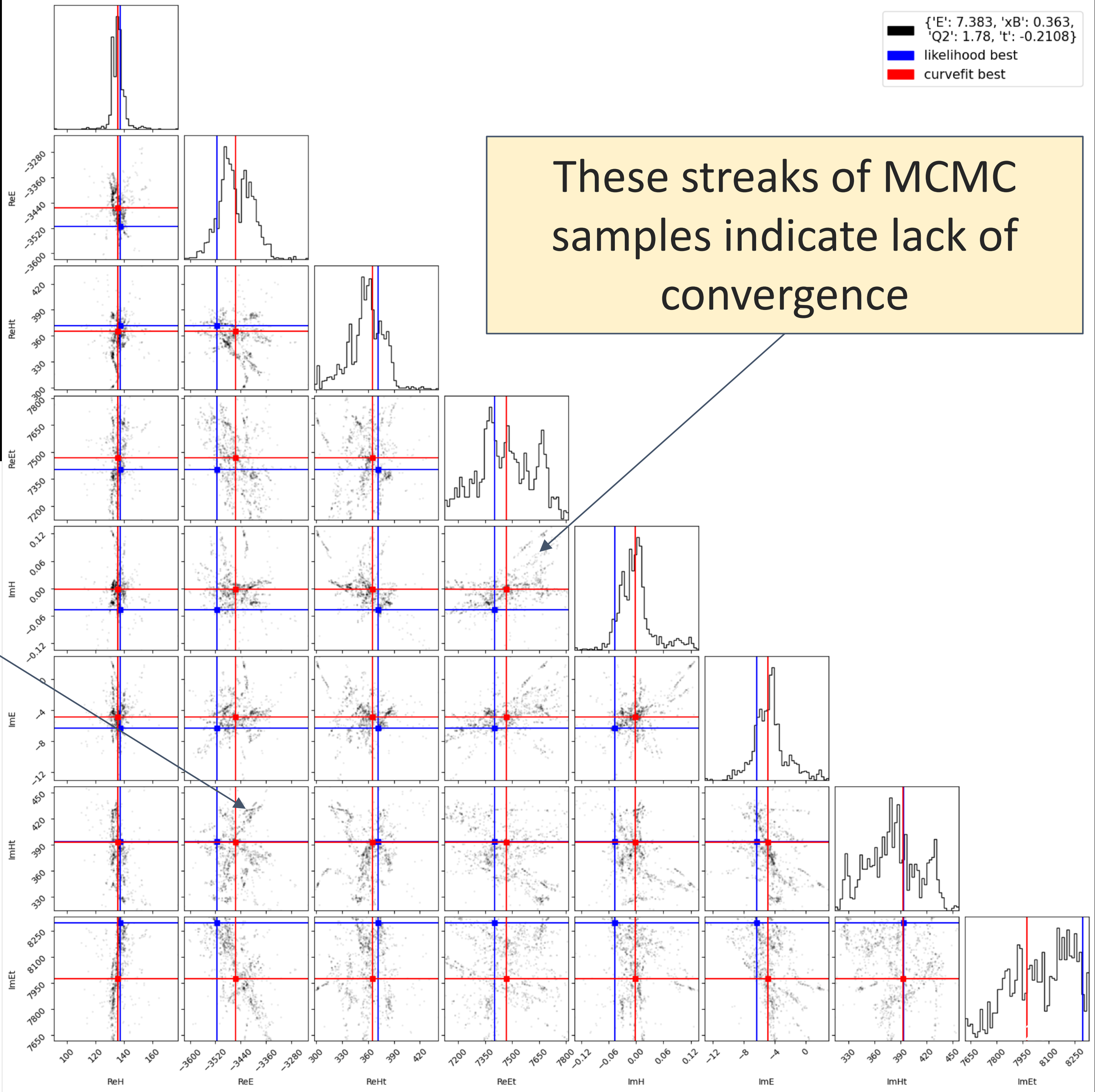


Good MCMC algorithms generate samples which would reproduce the distribution as a histogram

Naive MCMC 1

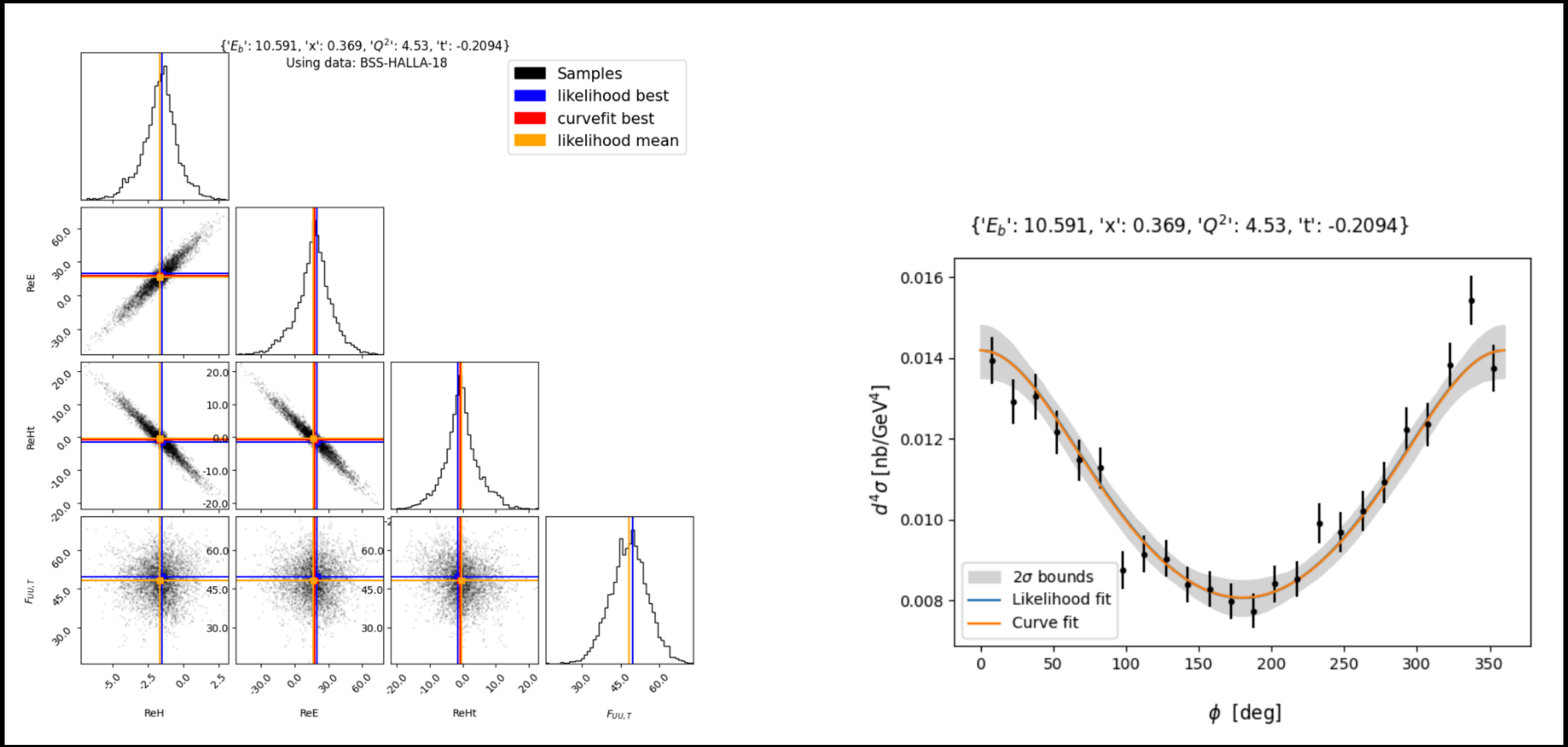
Fitting $\sigma_{TOT}(\phi_A)$ directly with all 8 CFFs
Provides a highly degenerate result
(nonsense bounds)





These streaks of MCMC samples indicate lack of convergence

(1) A likelihood analysis of the DVCS cross section model vs. deeply virtyual



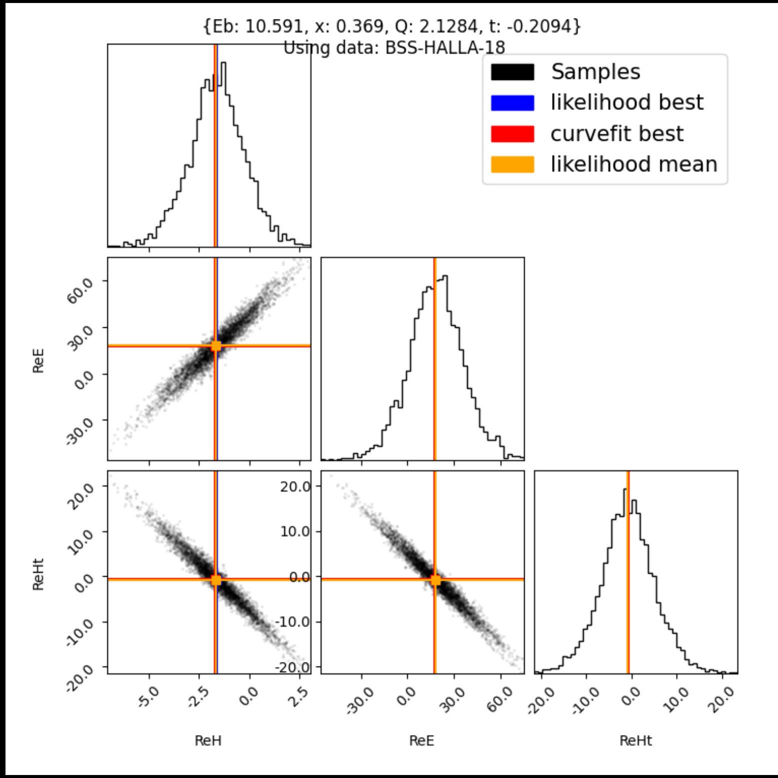
Only three CFFs are non degenerate!

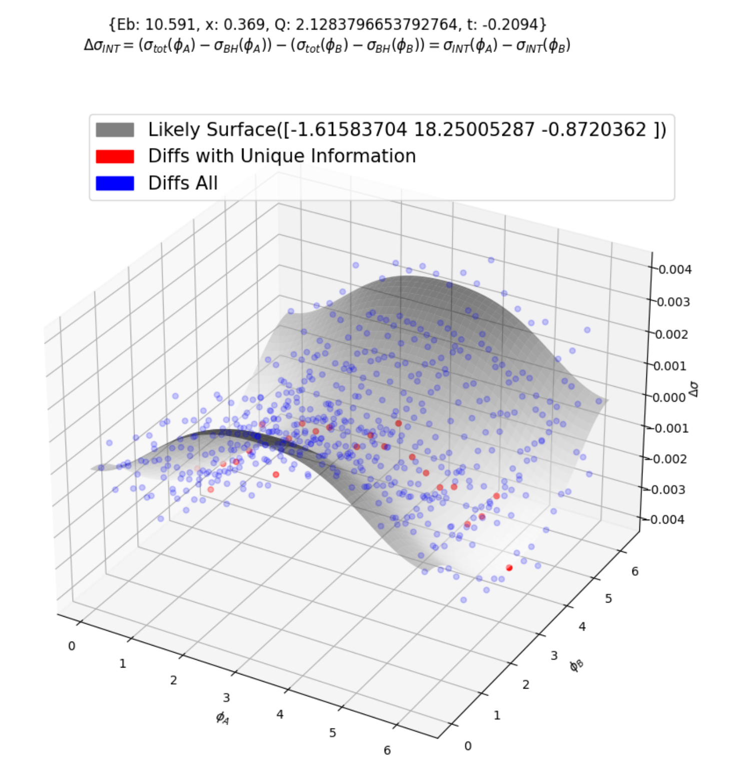
CFFs cannot be extracted from unp DVCS
x-sec

Outliers analysis (not shown) improves results

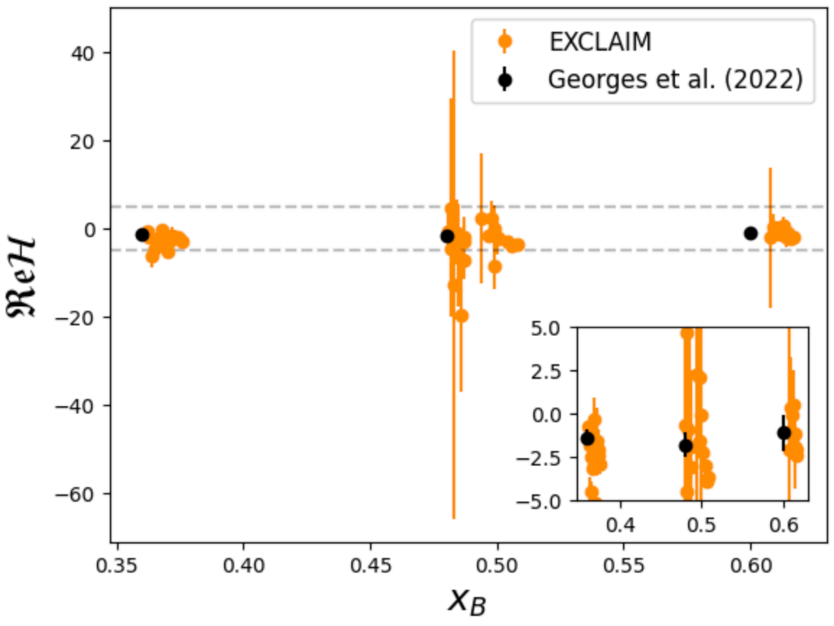
Difference Likelihood Result

- Here the maximum likelihood is achieved allowing 3CFFs to vary.
Only 23 combinations of 2 angles are used.

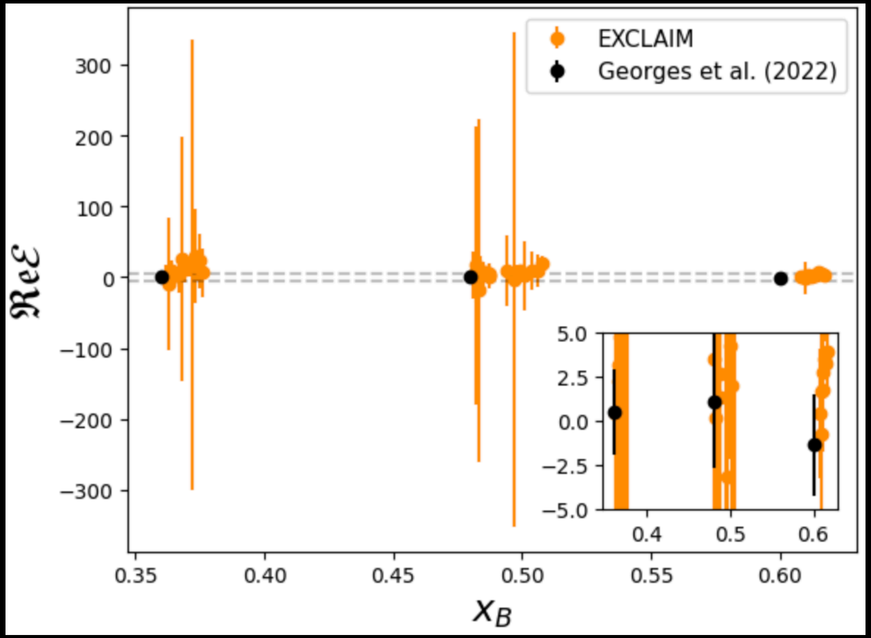




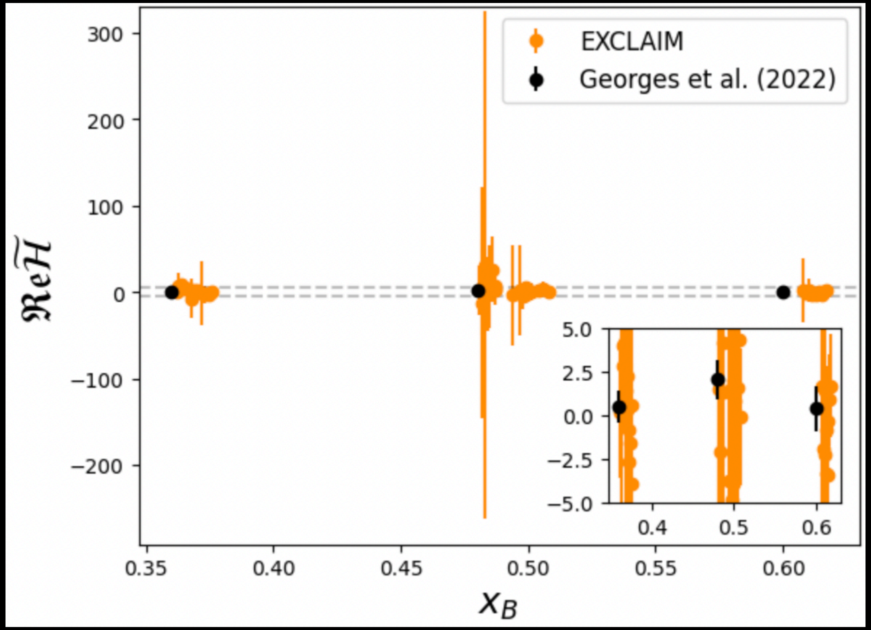
$\mathcal{H}$



$\mathcal{E}$



$\widetilde{\mathcal{H}}$



CFF Likelihood Result Summary https://arxiv.org/abs/2410.23469

- Using the UVA DVCS twist 2 unpolarized cross section model ($\sigma_{TOT,\ UVA,\ UU}$) assuming:
  - Cross section model is True
  - Cross section model has 8 CFFs only
  - Each CFF is independent of phi, but dependent on other kinematics
- Using Hall-A DVCS Data from Georges thesis
  - doi:10.1038/s41567-019-0774-3
  - each kinematic bin has 24 rows of (ϕ, σ_{TOT}) data
- Naively one would assume we can use the model to produce 24 equations and 8 unknowns to fully constrain the unknowns (as an overdetermined system).
  - However 5 CFFs are degenerate because σ_{DVCS} has no phi dependence.
  - Thus only the other 3 CFFs can be fully constrained using σ_{INT}
- We produced a table of CFF results for 45 kinematic bins

2. Inverse Problem Techniques: VAIM, C-VAIM, MCMC

Approaches to find parameters statistically in an underdetermined system:

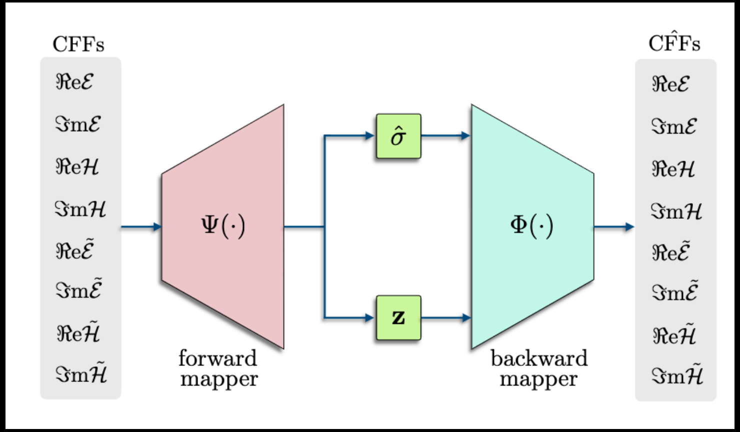
- Can quantify parameter uncertainty when more parameters than data

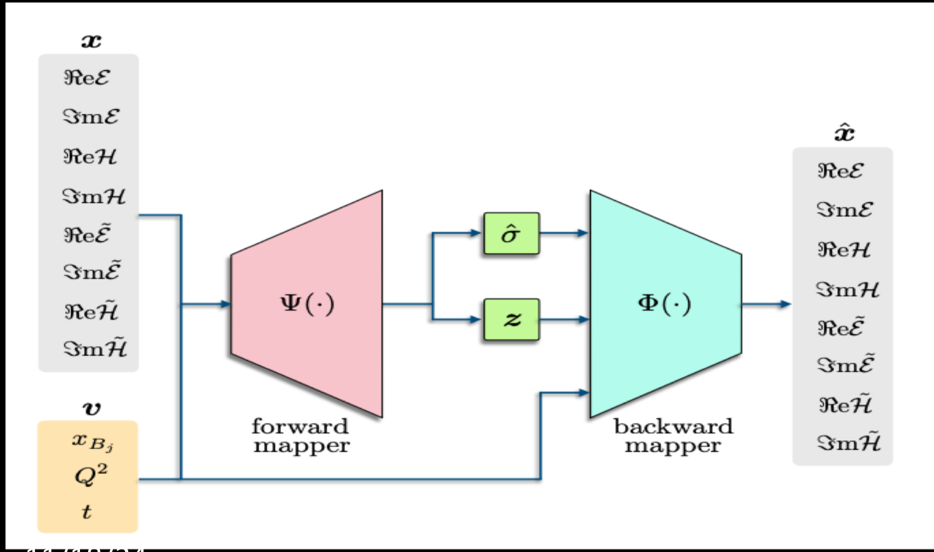
- Techniques highly dependent on bounded parameter priors

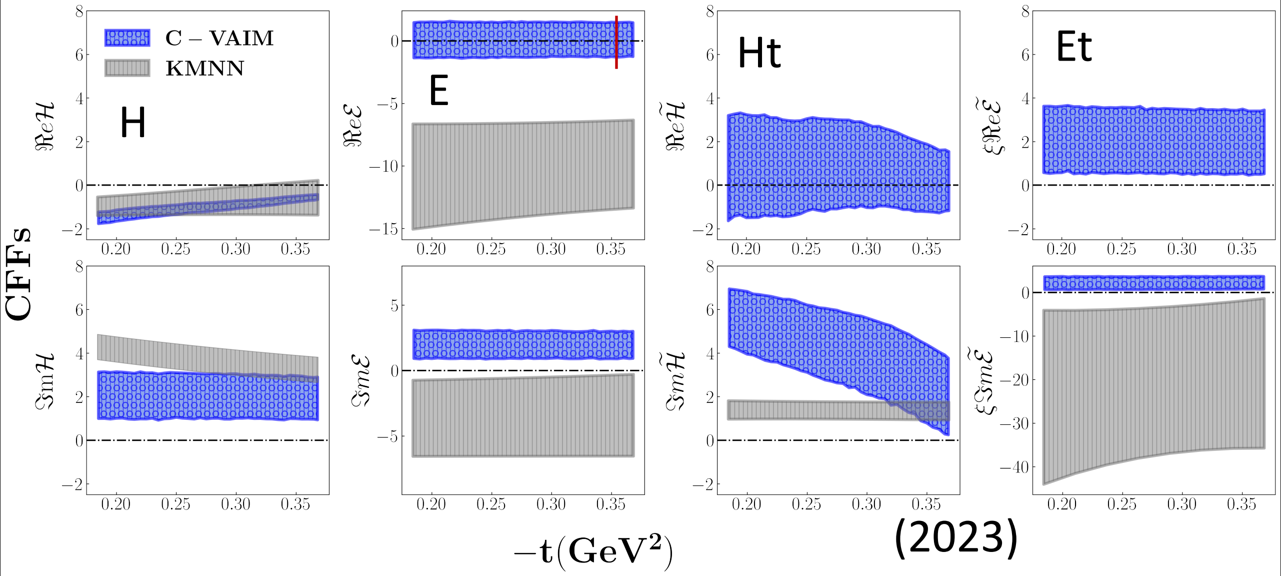
- These methods give us an initial way to perceive:
    - the correlation between parameters on a complicated model
    - what information is missing (latent space)

(2)

- **A variational autoencoder inverse mapper solution to Compton form factor extraction from deeply virtual exclusive reactions** arXiv: 2405.05826





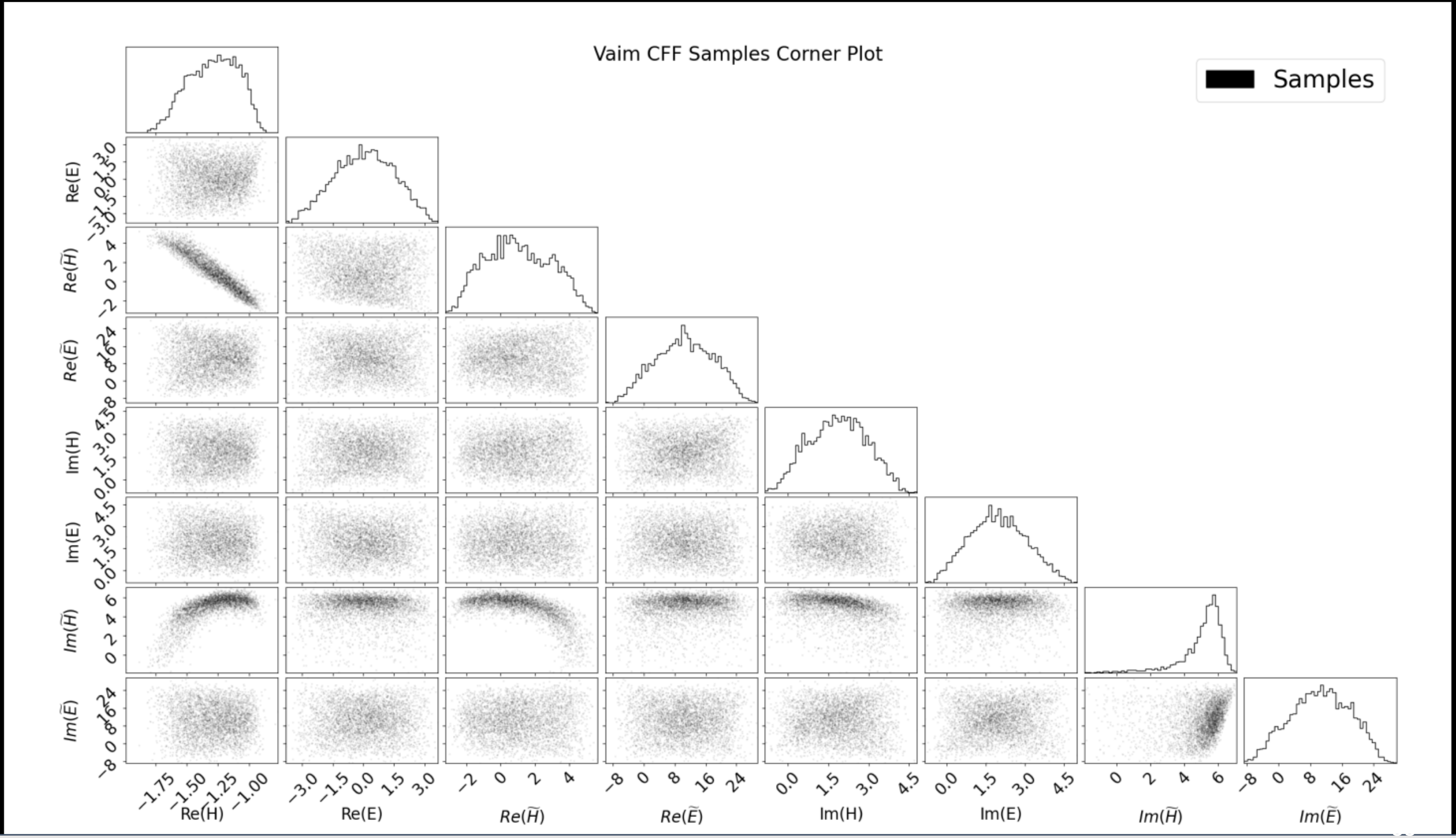


- KMNN, https://arxiv.org/abs/2007.00029

VAIM Result Using Prior for CFFs (2)

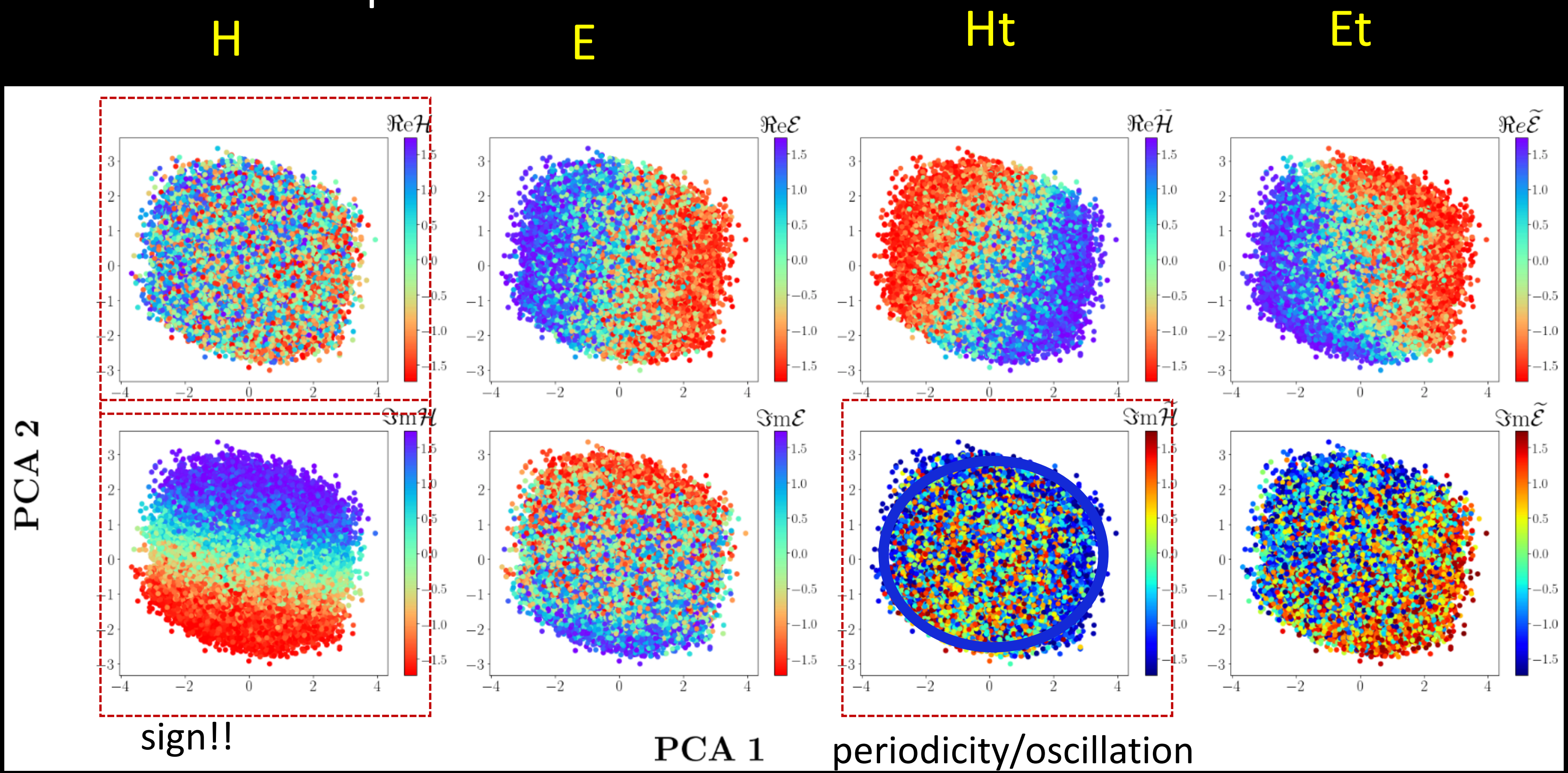
https://arxiv.org/pdf/2405.05826

- Apply cross section equation as constraint with observed data
- Include a prior
- Generate random but viable CFFs which try to satisfy the constraint



https://arxiv.org/pdf/2405.05826

CFFs Analysis of Latent Space



VAIM Results Motivate a likelihood analysis

- Requires a prior for the CFFs
- Assume the same CFFs work for many different kinematic bins
- Approximated the error bars on the data

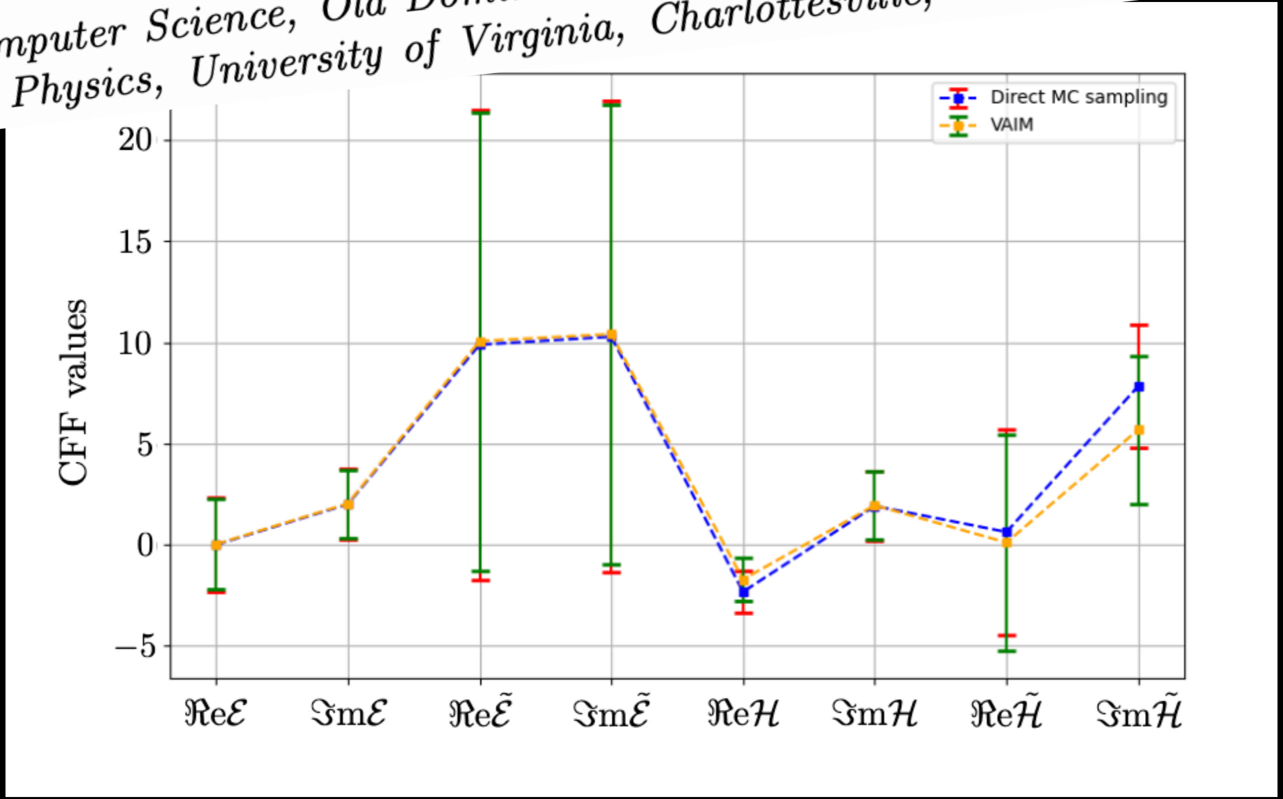
It would be nice to reduce assumptions required

https://arxiv.org/pdf/2405.05826

Comparison of MCMC sampling/inverse mapper and VAIM

Variational autoencoder inverse mapper for extraction of Compton form factors: Benchmarks and conditional learning

Fayaz Hossen,[1] Joshua Bautista,[2] Douglas Adams,[2] Yaohang Li,[1,*] Gia-Wei Chern,[2,†] and Simonetta Liuti[2,‡]
[1] *Department of Computer Science, Old Dominion University, Norfolk, VA 23529, USA.*
[2] *Department of Physics, University of Virginia, Charlottesville, VA 22904, USA.*



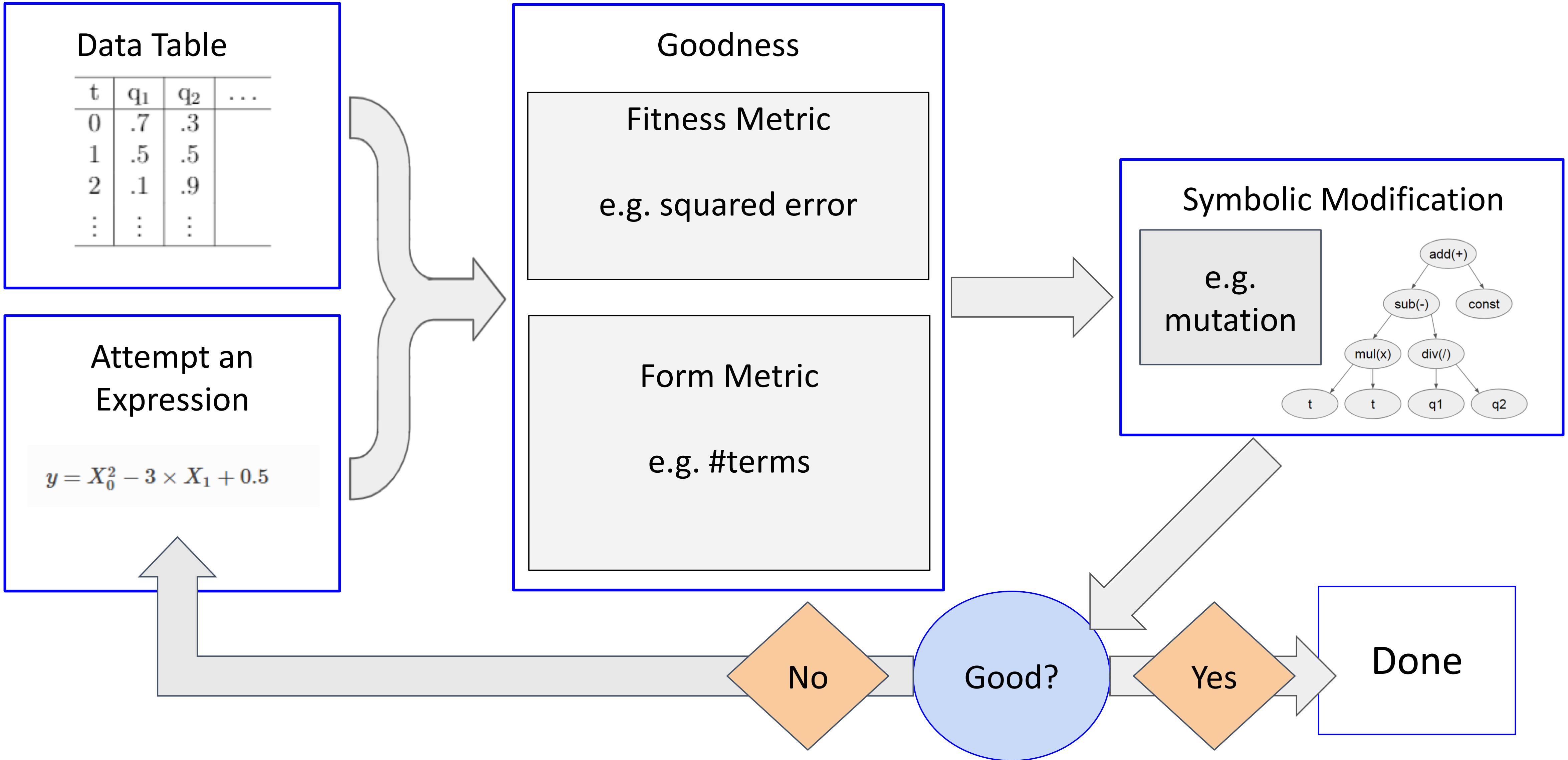
3. Symbolic Regression for Parton Research

1) What is symbolic regression and why do we care?
    a) Spoiler: because then humans can read the answer

2) What are the existing tools out there?
    a) Eureka
    b) Gplearn
    c) AI Feynman
    d) PySr
    e) RL-SR
    f) *Meijer-G-Function (very preliminary)

Graduate Students:
Andrew Dotson (NMSU)
Anusha SingiReddy (ODU)
Zaki Panjsheeri (UVA)

What is symbolic regression (SR)?



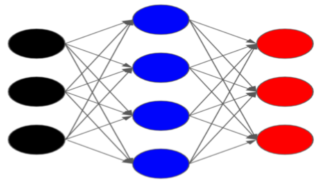
Why bother with SR when we have neural networks?

Which is easier to read? (a.k.a interpretability of AI)

$$y = X_0^2 - 3 \times X_1 + 0.5$$

vs

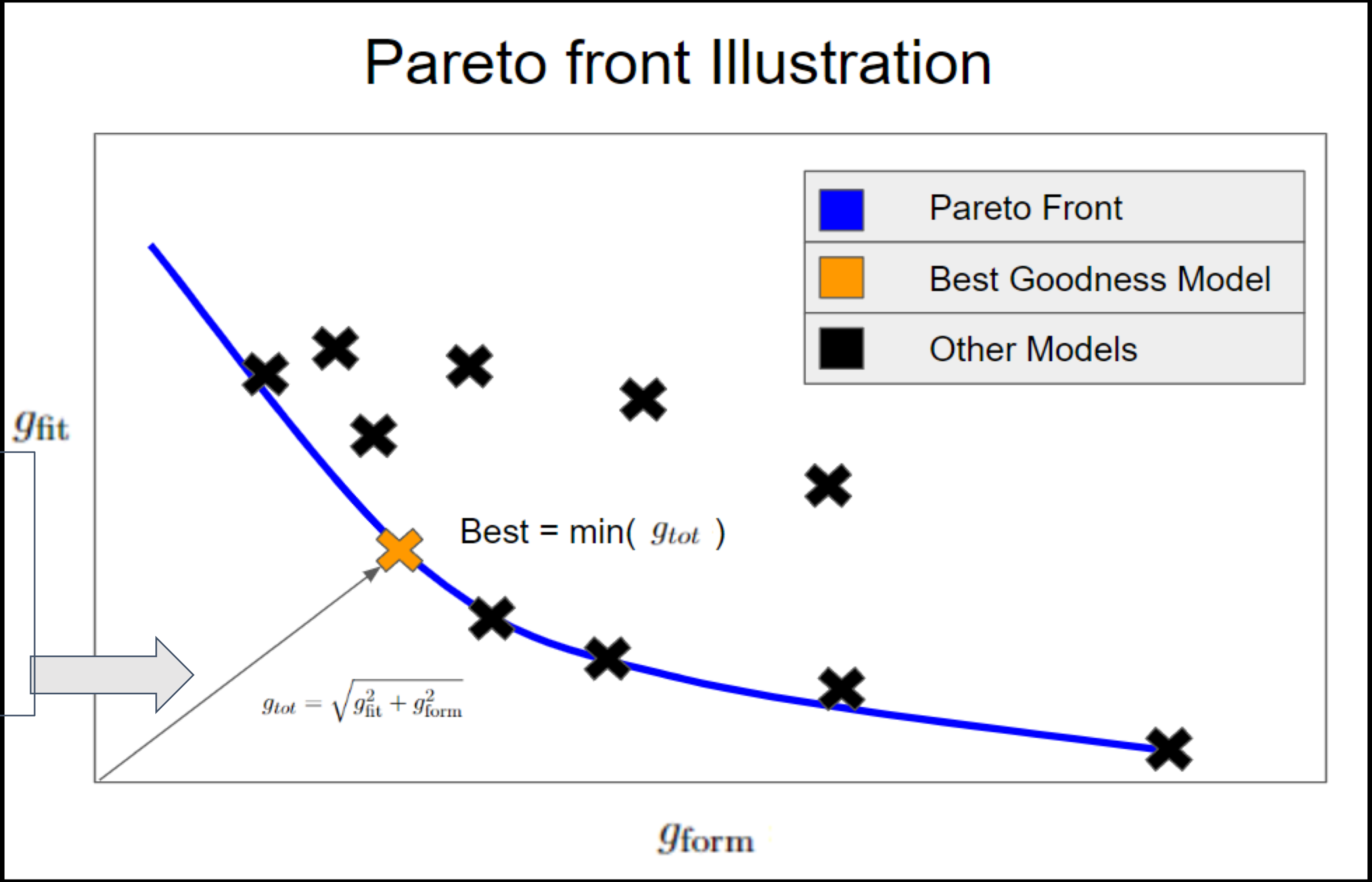
Black Box Regression



$$q_{i,BB}(t) \approx \ldots$$

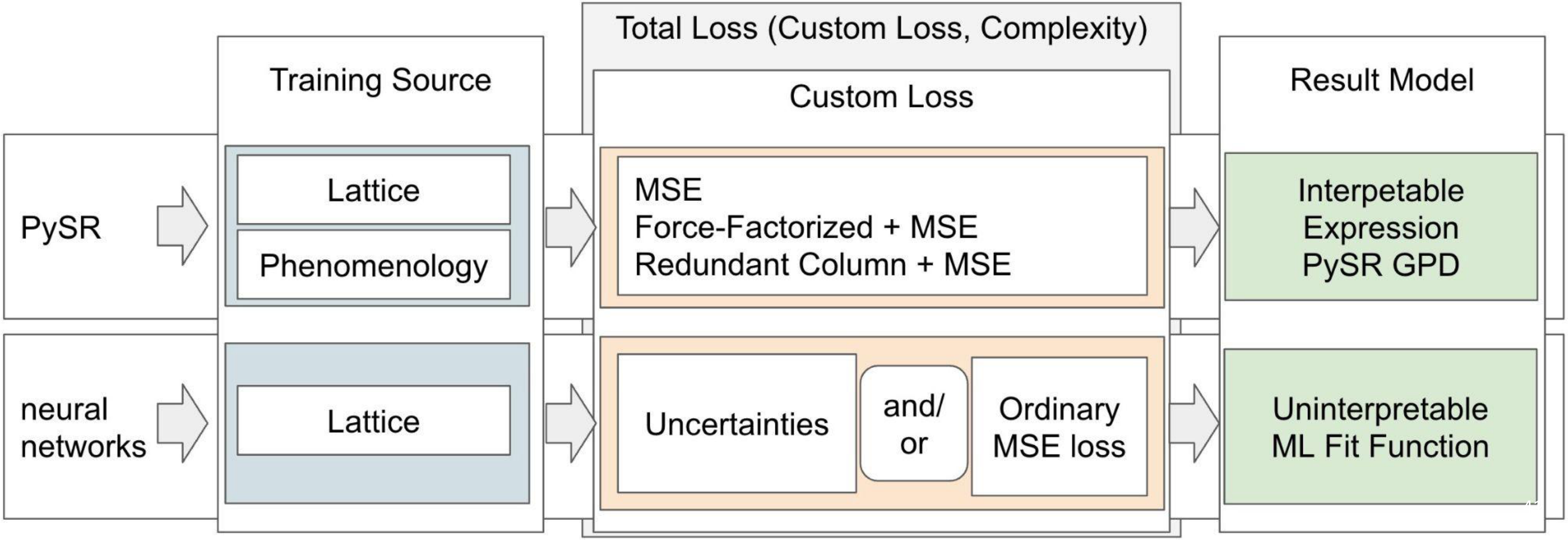
Using a pareto front to choose amongst forms

One must define a "metric" which balances fit vs. form to choose a best

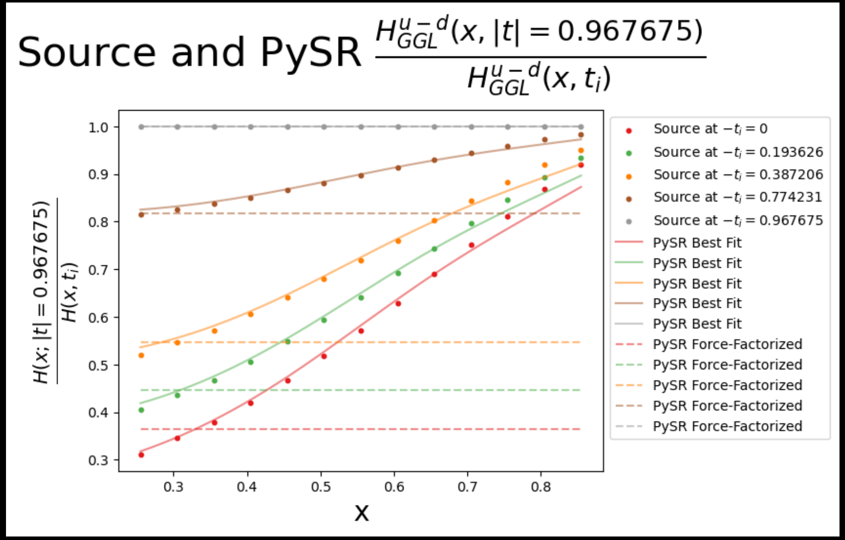


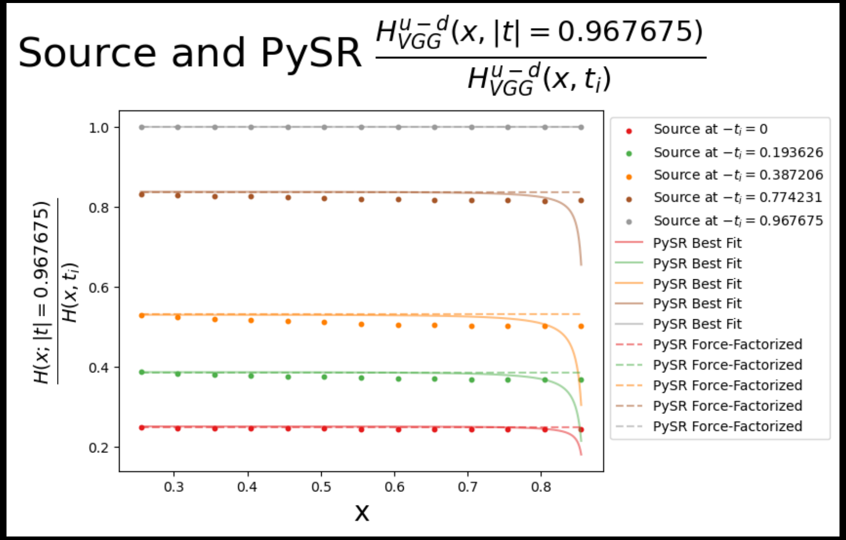
Generalized Parton Distributions from Symbolic Regression

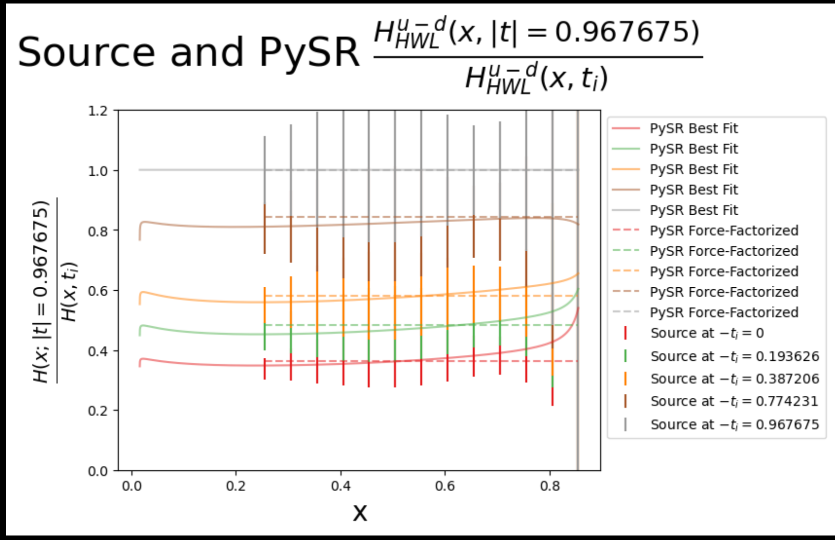
We have a lattice simulation of a GPD as a function of x, t, Q^2
The goal is to find a closed form expression for that GPD



Testing x and t factorization (important for spatial configurations!)







phenomenology

lattice

J. Holligan and H-W Lin, Phys.Rev. D110, 034503 (2023)
H-W Lin, Phys. Lett. B824, 136821 (2022)

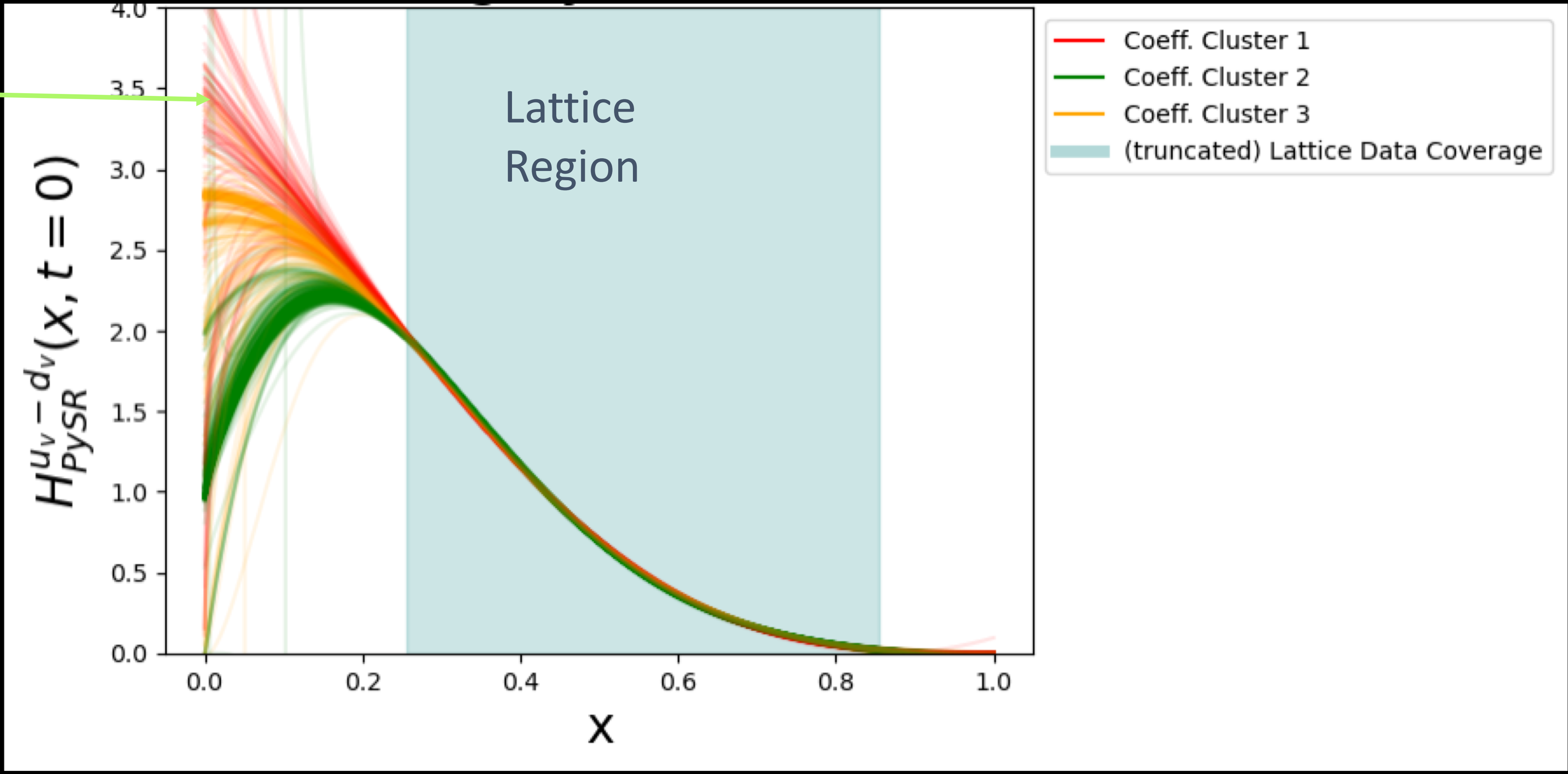
Example of a factorized in x and t form

(MSE constant power)

$$\frac{1.07 \cdot \left(2.78\,(1 - 0.766x)^{3.83} - 0.0437\right)}{-t + 0.603}$$

Novel SR Convergence Clustering

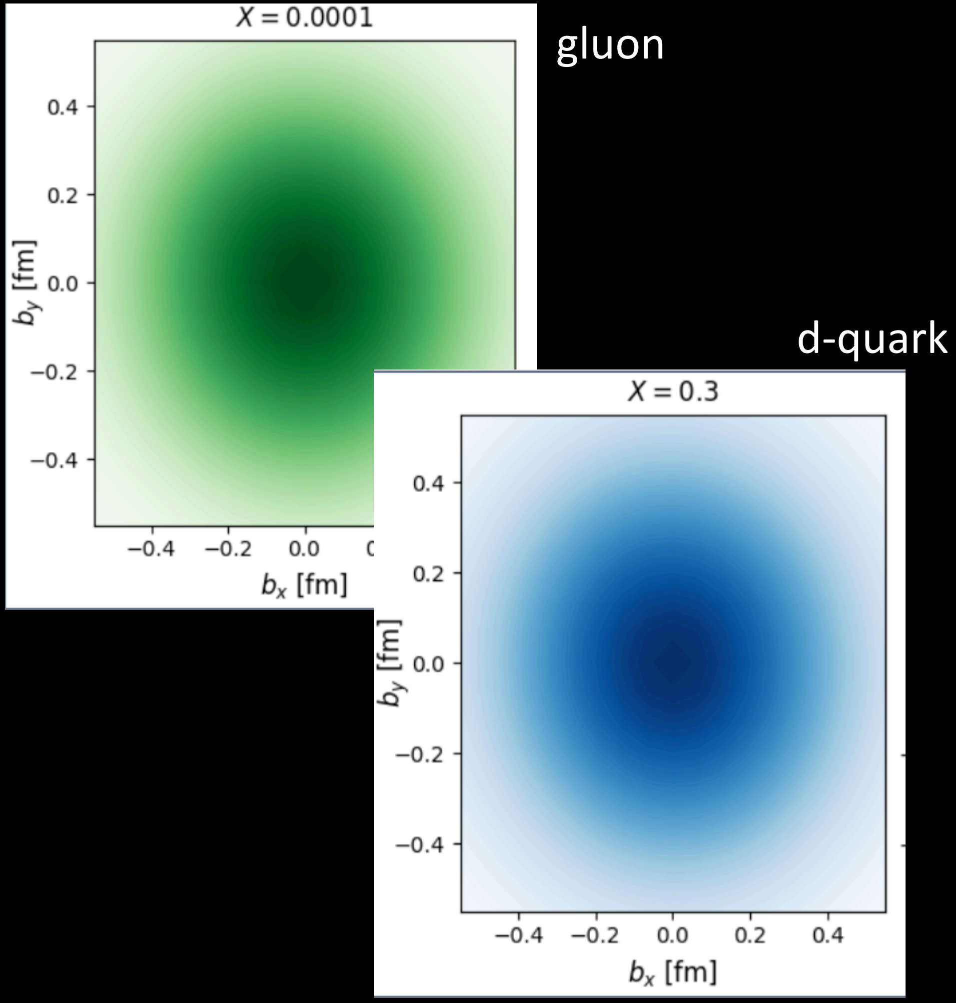
Extrapolation



Whether the x and t dependences factorize has consequences on the 3D Coordinate Space picture

GPDs can be Fourier transformed from momentum space into coordinate space, providing insight into the spatial distributions of quarks and gluons inside the proton, besides matter and charge distributions.

$$\overbrace{\mathcal{H}^q(X,0,b_T)}^{\text{Slice of Wigner phase space distribution}} = \int \frac{d^2\Delta_T}{(2\pi)^2}\, \underbrace{H^q(X,0,\Delta_T)}_{\text{GPD}}\, e^{-i\Delta_T\cdot b_T}$$



With Z. Panjsheeri and J. Bautista

Gluon and quark matter density radius

$$< b_T^2 >^q (X) = \frac{\int_0^\infty d^2 b_T b_T^2 \mathcal{H}^q(X,0,b_T)}{\int_0^\infty d^2 b_T \mathcal{H}^q(X,0,b_T)}$$

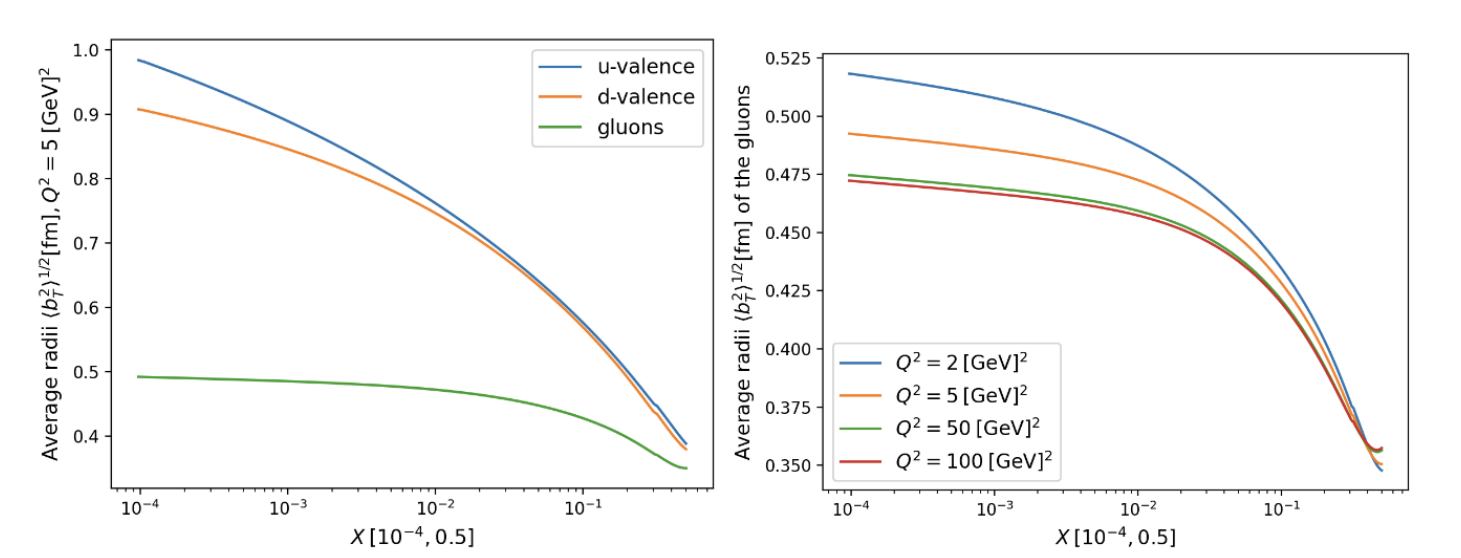
Bautista, Panjsheeri, SL (2024)

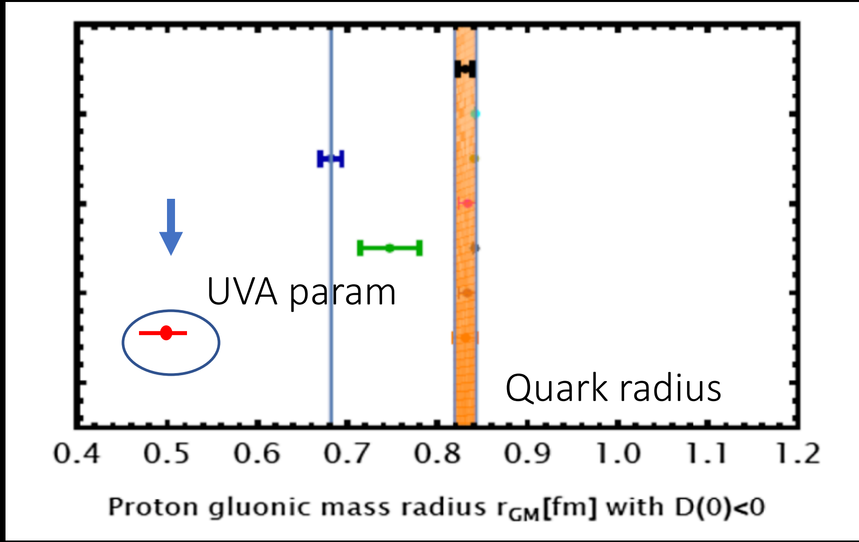
Compare to lattice and AdS/CFT integrated value

K. Mamo and I. Zaeed

PRD106, 086004 (2022)

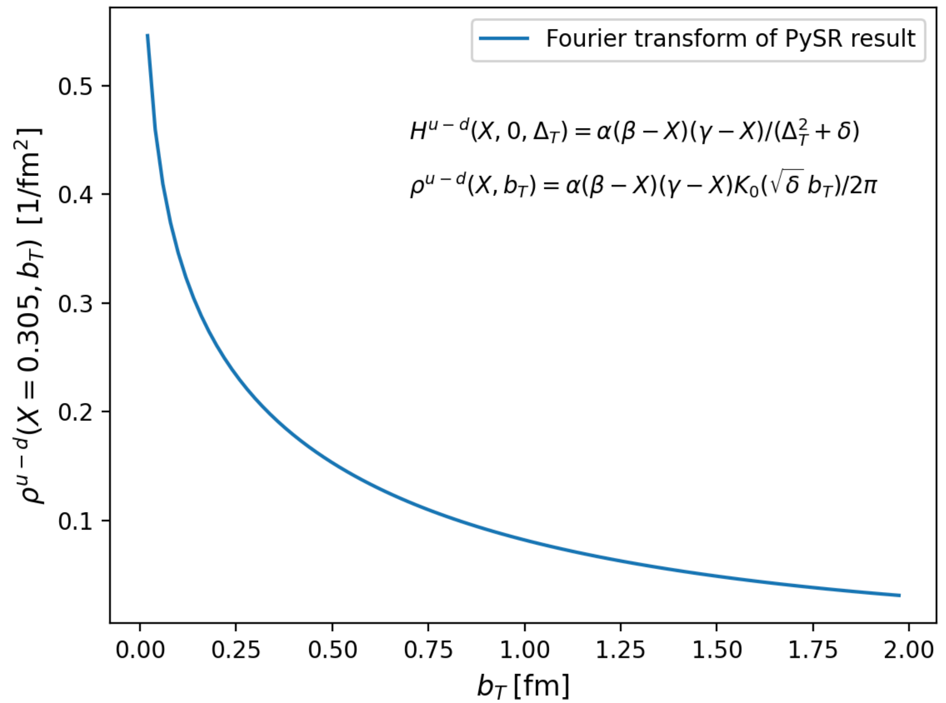
LQCD: Detmold and Shanahan

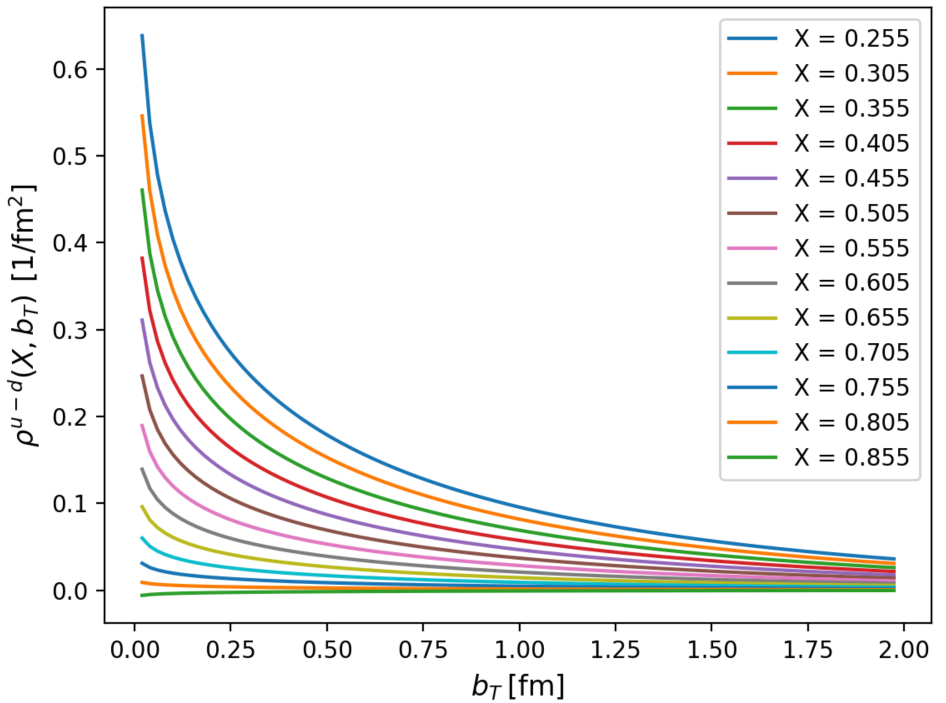




arXiv:2405.05842

From SR Analysis





Papers Recent & In Preparation:

Variational autoencoder inverse mapper for extraction of Compton form factors: Benchmarks and conditional learning
https://arxiv.org/abs/2408.11681

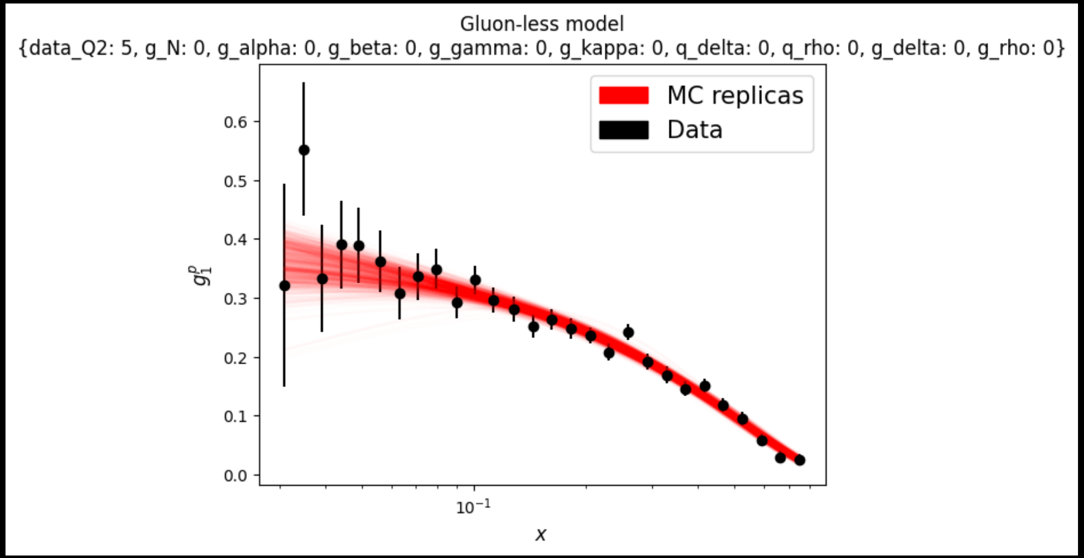
VAIM-CFF: A variational autoencoder inverse mapper solution to Compton form factor extraction from deeply virtual exclusive reactions
https://arxiv.org/abs/2405.05826

Likelihood and Correlation Analysis of Compton Form Factors for Deeply Virtual Exclusive Scattering on the Nucleon
https://arxiv.org/abs/2410.23469

Generalized Parton Distributions from Symbolic Regression
(in preparation)

What I did not talk about:

- Epistemic and aleatoric uncertainty through BNN (Fayaz Bin Hossen, ODU)
- Analysis of latent space
- Δg extraction (Saraswati Pandey, UVA see poster)
- NNGPD Project (Yang (Jason) Ho, Adil Khawaja, Zaki Panjsheeri, UVA, see poster)



Conclusions

1. A successful reconstruction of the **spatial structure of the proton** (and all of its mechanical properties) relies on our ability to understand the **cross section** for **all the various DVES processes**

2. This implies solving **multiple inverse problems**

3. We have defined a path to extract the **observables** from experiment that allows us to fully take into account UQ from data and ab initio QCD calculations

4. Bringing interpretability and benchmarking to AI tools is a necessity for us to progress faster towards understanding the 3D picture of the proton

5. Obtaining spatial images of the proton including UQ is feasible using AI/ML to extend the momentum transfer reach for an accurate Fourier transformation

Back up

PySR convergence

Andrew Dotson