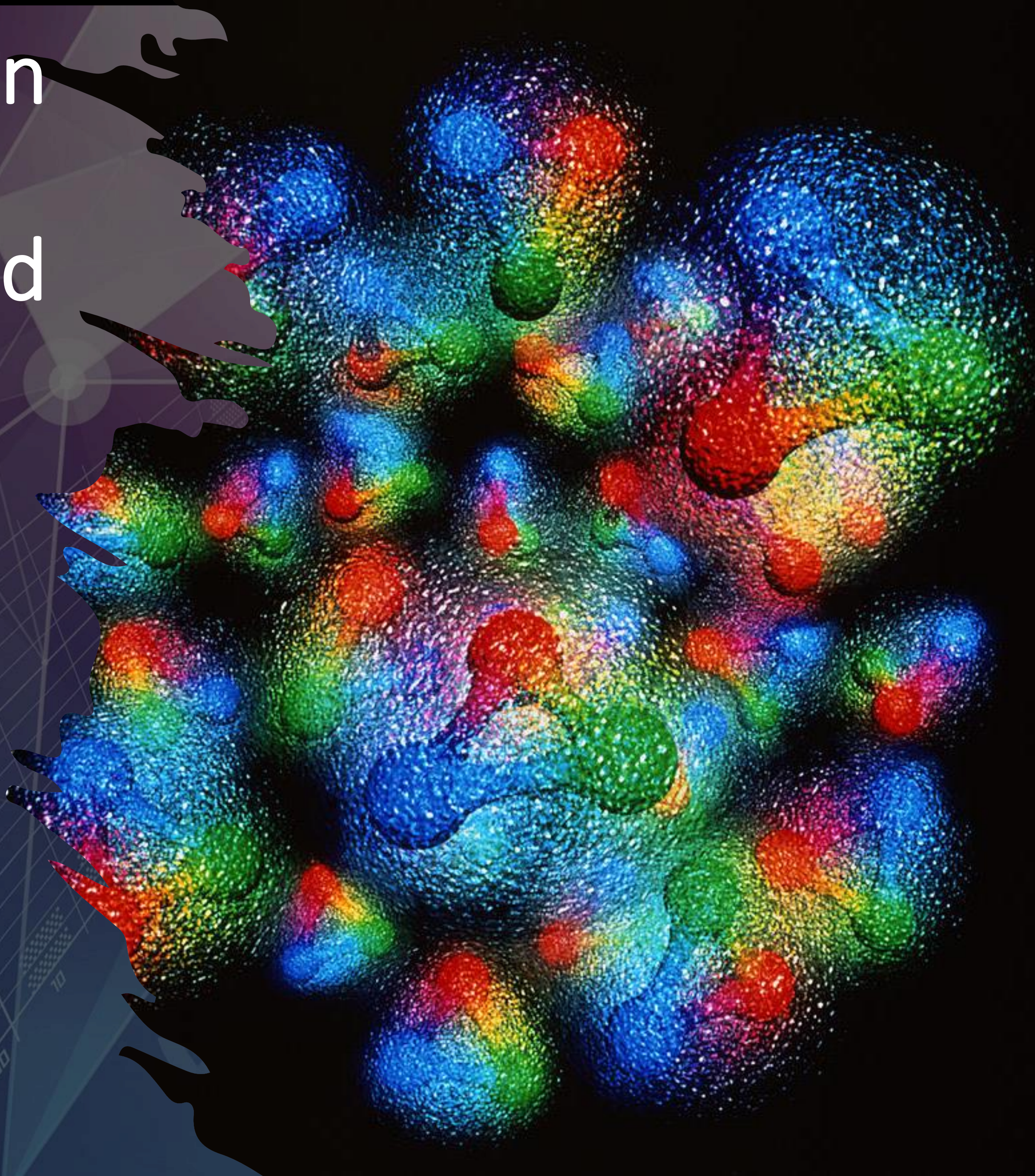
What physical information is in deeply virtual exclusive experiments and how do we take a snapshot of the proton?

Simonetta Liuti



UNIVERSITY OF VIRGINIA 1819

DEPARTMENT OF ENERGY UNITED STATES OF AMERICA

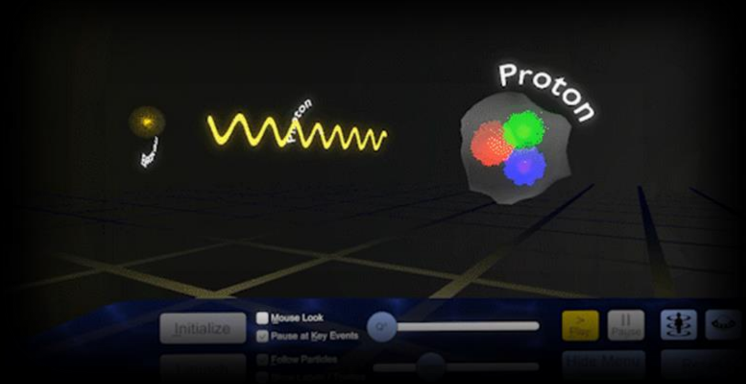
EXCLAIM

The **Electron-Ion Collider (EIC)** is a next-generation particle accelerator planned to be built at Brookhaven National Laboratory in the U.S., with construction expected to start in the mid-2020s and first collisions later in the decade.

“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 EXCLAIM collaboration

https://exclaimcollab.github.io/web.github.io/

PIs:

CS Methods: Prasanna Balachandran (UVA), Gia-Wei Chern (UVA), Yaohang Li (ODU)

Lattice QCD: Michael Engelhardt (NMSU), Huey-Wen Lin (MSU)

Phenomenology: Gary Goldstein (Tufts), SL (UVA), Matt Sievert (NMSU), (Dennis Sivers)

Experiment: Marie Boer (VT)

Currently Funded Postdocs: Douglas Adams, Liam Hockley, Saraswati Pandey, Kemal Tezgin

Graduate Students: Andrew Dotson, Jang (Jason) Ho, Adil Khawaja, Zaki Panjsheeri, Brannon Semp, Jitao Xu

Undergraduates: Will Faircloth, Lionel Strauss

UVA grad students/postdocs sub-group

Douglas Q. Adams

Saraswati Pandey

Jang (Jason) Ho

Zaki Panjsheeri

Adil Khawaja

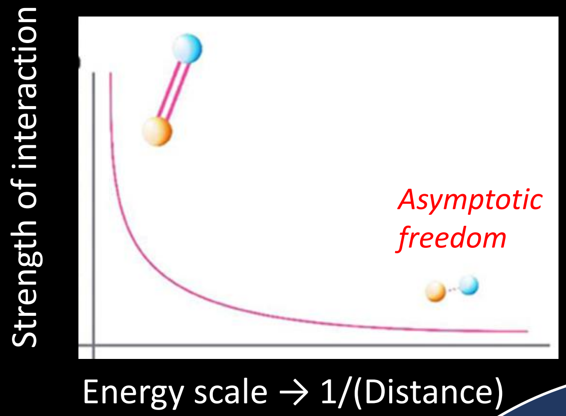
ODU sub-group

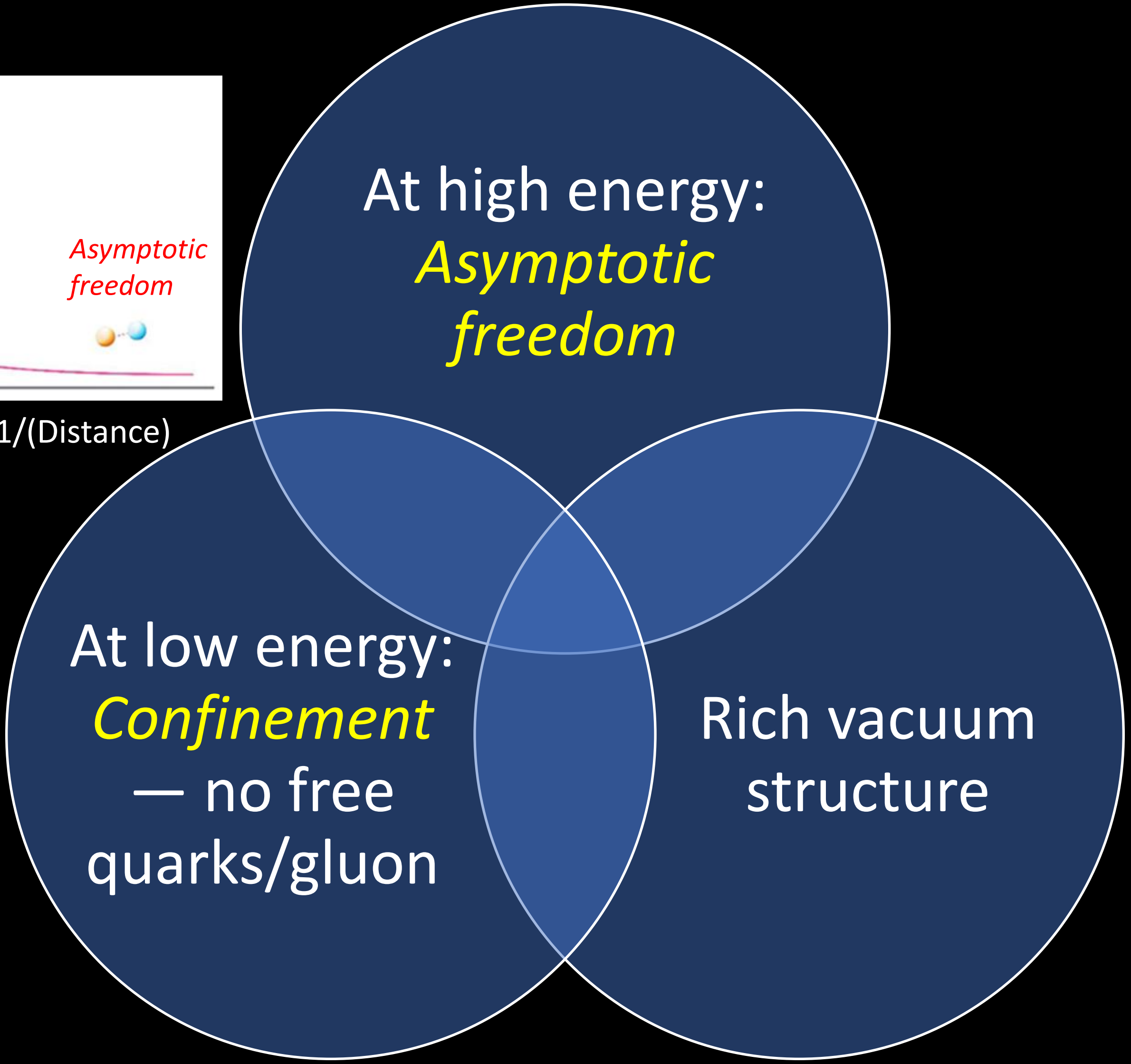
Jitao Xu

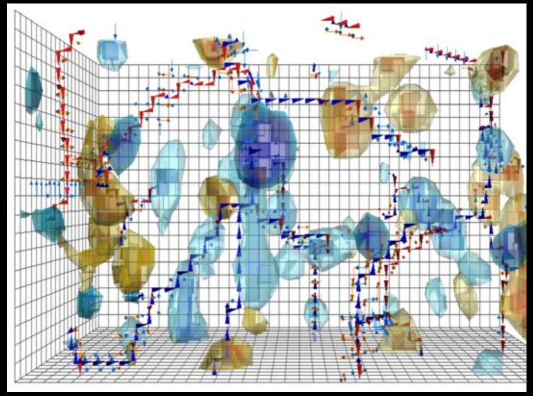
NMSU sub-group

Andrew Dotson

Quantum Chromodynamics and the structure of the proton

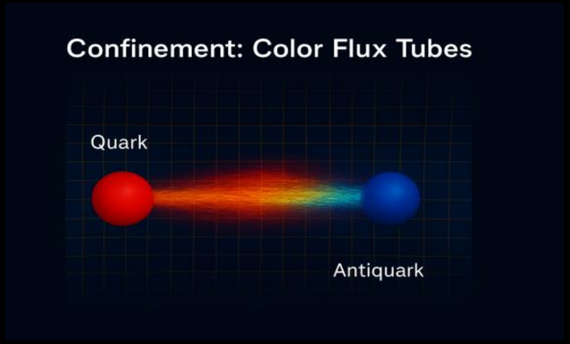






D. Leinweber's visualizations

gluon and quark fluctuations/ condensates described in terms of instantons and/or color vortices



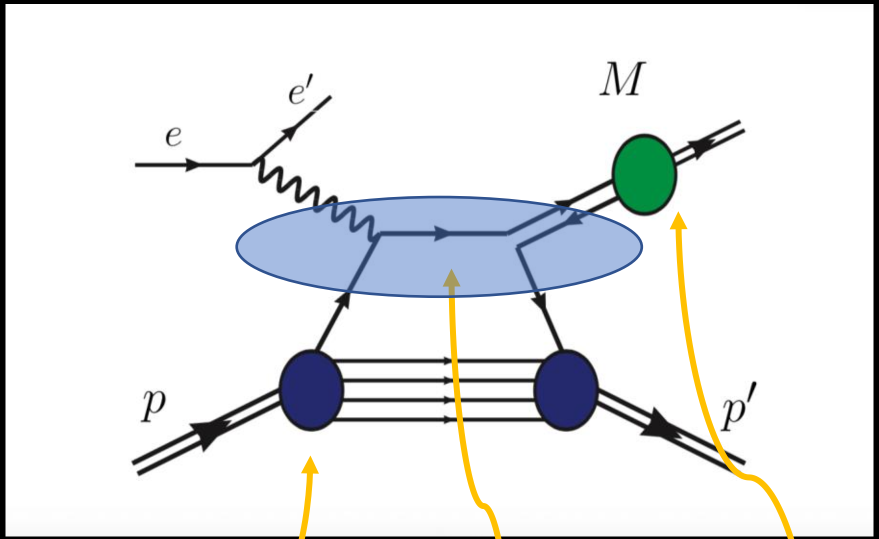
Questions in phenomenology/connecting theory to data

- understanding the ***factorization properties*** of the theory

- Role of the non perturbative regime of QCD

- understanding the non perturbative-perturbative connection

Factorization



Parton distribution function

Fragmentation function

$$\sigma = \sum_{i,j,k} f_{i/p}(x,\mu) \otimes \hat{\sigma}_{ij\to k}(Q,\mu) \otimes D_{h/k}(z,\mu)$$

Hard scattering process

But.....even in a framework where factorization occurs at a given value of a hard scale

- We cannot predict where the demarcation between high and low and low energy happens (value of the factorization scale)

- There are many largely unknown layers that **separate the observable from the dynamical and topological properties of the underlying QCD vacuum**

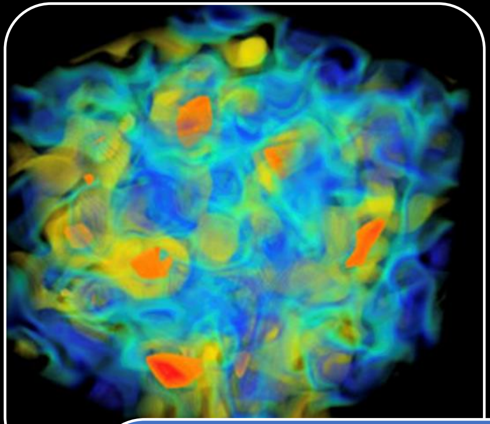
- Lattice QCD provides only partial answers to these questions

Where AI comes in: solving the Inverse Problem in QCD

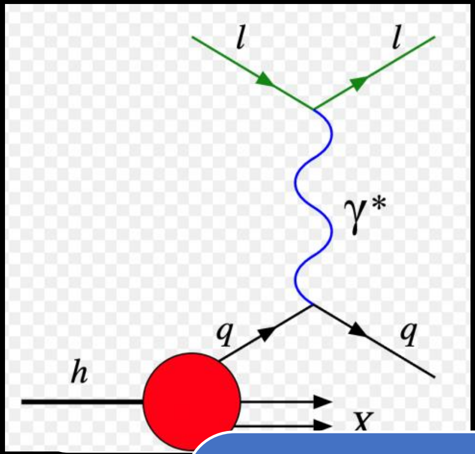
Forward

Inverse

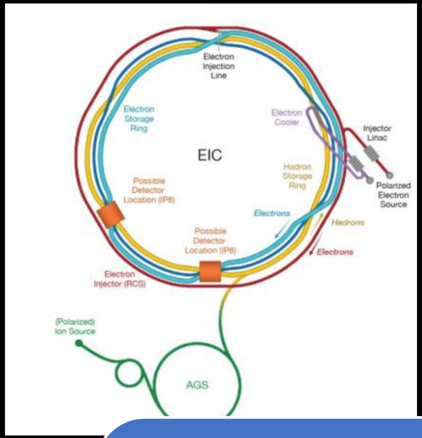
Theory



Pheno



Data



QCD
- quark field
- gluon field
- correlation functions

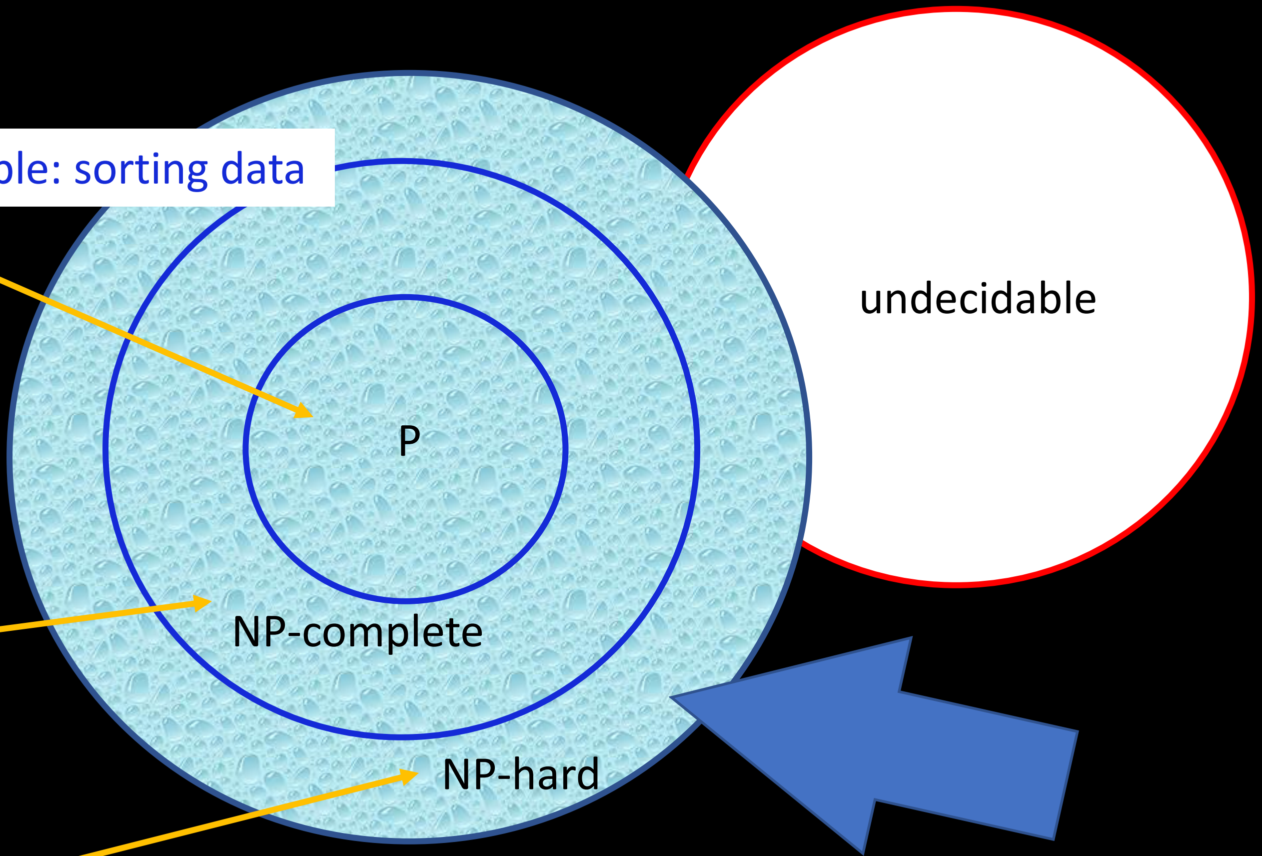
Observables
- Form Factors
- Structure Functions
- Compton Form Factors
- Fragmentation functions
- …

Data/Measurements

SOME THEORETICAL BACKROUND…

P=based on fast polynomial algorithm

Example: sorting data



NP-complete = based on slow algorithms, exponential or slower

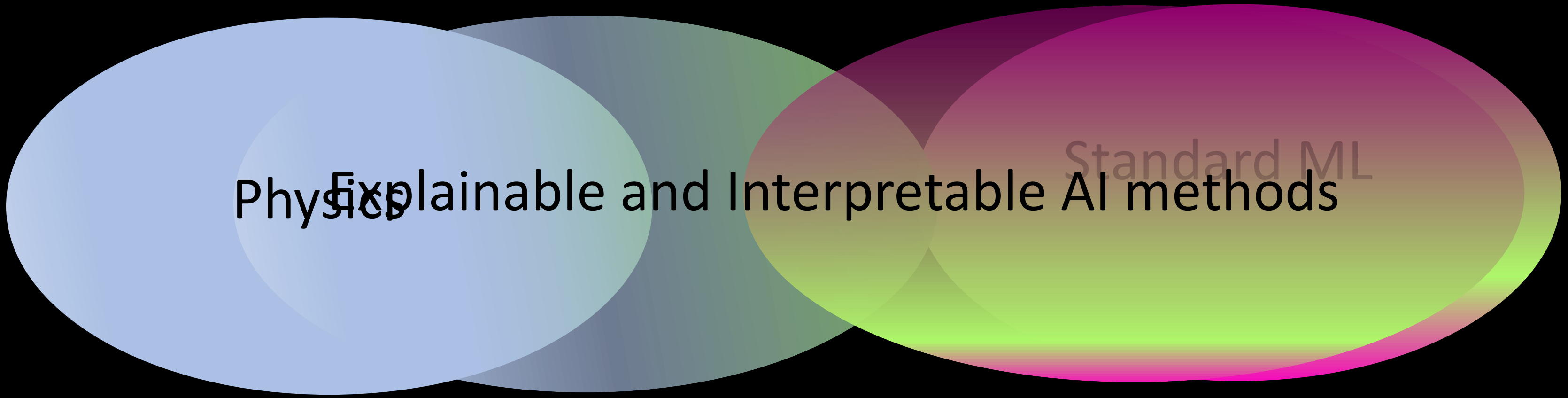
NP-hard = the hardest problems

Example: Traveling Salesman Problem

To find a solution need to address tension between computation **complexity** and physical existence of a solution

Interpretability: directly <u>enabling discovery of laws</u>

(beyond computing)

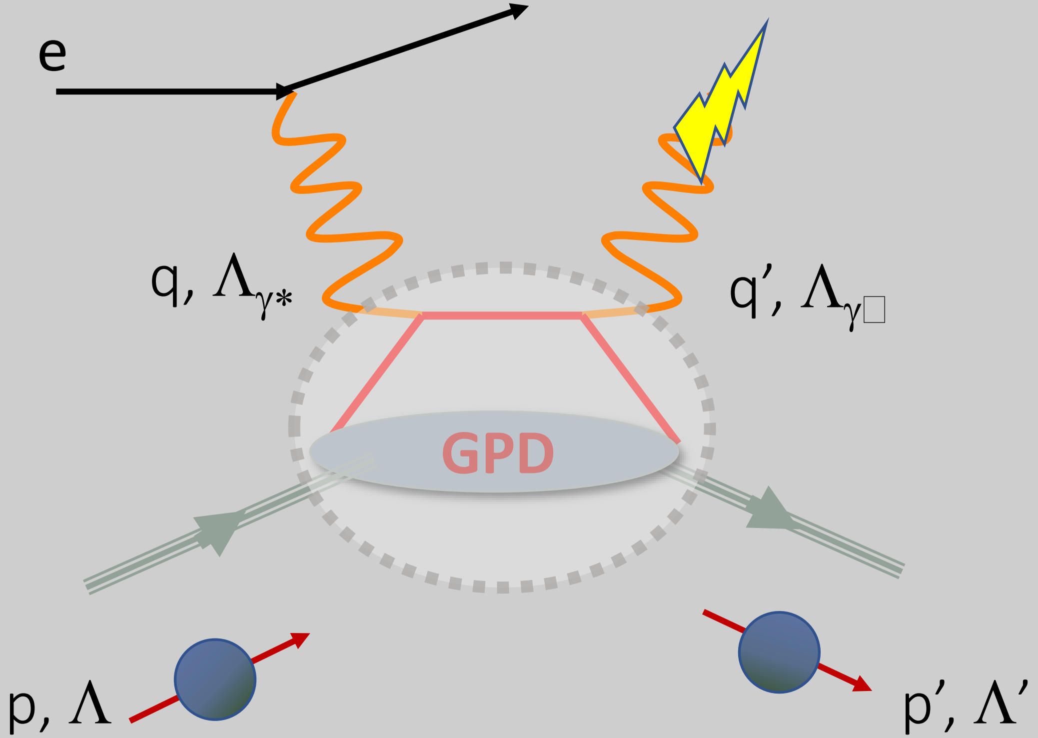


Finding a common language needs to be taken very seriously and is a research topic in itself

Our Physics Case

Extracting information on the femtoscale structure of the proton and the atomic nuclei from deeply virtual exclusive scattering experiments

In a QCD factorized scenario:
quark and gluon spatial configurations from deeply virtual exclusive scattering



Deeply virtual Compton scattering (DVCS)
Generalized Parton Distribution (GPDs)

At least one photon is hard

helicity amplitude

$$T^{h\Lambda'_\gamma}_{DVCS,\Lambda\Lambda'} = \sum_{\Lambda_{\gamma^*}} A_h^{\Lambda_{\gamma^*}}(k,k',q)\,\boxed{f_{\Lambda\Lambda'}^{\Lambda_{\gamma^*}\Lambda'_\gamma}(q,p,q',p')}$$

e γ* -> e'

γ* p -> γ' p'

From Helicity Amplitudes to DVCS Cross Section

$$\frac{d^5\sigma_{DVCS}}{dx_{Bj}dQ^2d|t|d\phi d\phi_S} = \Gamma\left|T_{DVCS}\right|^2$$

$$= \frac{\Gamma}{Q^2(1-\epsilon)}\Big\{F_{UU,T} + \epsilon F_{UU,L} + \epsilon\cos 2\phi F_{UU}^{\cos 2\phi} + \sqrt{\epsilon(\epsilon+1)}\cos\phi F_{UU}^{\cos\phi}$$
$$+ (2h)\sqrt{2\epsilon(1-\epsilon)}\sin\phi F_{LU}^{\sin\phi}$$

unpolarized

$$+ (2\Lambda)\Big[\sqrt{\epsilon(\epsilon+1)}\cos\phi F_{UL}^{\cos\phi} + \epsilon\sin 2\phi F_{UL}^{\sin 2\phi}$$
$$+ (2h)\left(\sqrt{1-\epsilon^2}F_{LL} + 2\sqrt{\epsilon(1-\epsilon)}\cos\phi F_{LL}^{\cos\phi}\right)\Big]$$

long. polarized

$$+ (2\Lambda_T)\Big[\sin(\phi-\phi_S)\left(F_{UT,T}^{\sin(\phi-\phi_S)} + \epsilon F_{UT,L}^{\sin(\phi-\phi_S)}\right)$$
$$+ \quad \epsilon\sin(\phi+\phi_S)F_{UT}^{\sin(\phi+\phi_S)} + \epsilon\sin(3\phi-\phi_S)F_{UT}^{\sin(3\phi-\phi_S)}$$
$$+\sqrt{2\epsilon(1+\epsilon)}\left(\sin\phi_S F_{UT}^{\sin\phi_S} + \sin(2\phi-\phi_S)F_{UT}^{\sin(2\phi-\phi_S)}\right)\Big]$$
$$+(2h)(2\Lambda_T)\Big[\sqrt{1-\epsilon^2}\cos(\phi-\phi_S)F_{LT}^{\cos(\phi-\phi_S)} + \sqrt{2\epsilon(1-\epsilon)}\cos\phi_S F_{LT}^{\cos\phi_S}$$
$$+\sqrt{2\epsilon(1-\epsilon)}\cos(2\phi-\phi_S)F_{LT}^{\cos(2\phi-\phi_S)}\Big]\Big\}$$

transv. polarized

First inverse problem!

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

Assuming a QCD factorized scenario → CFFs → GPDs

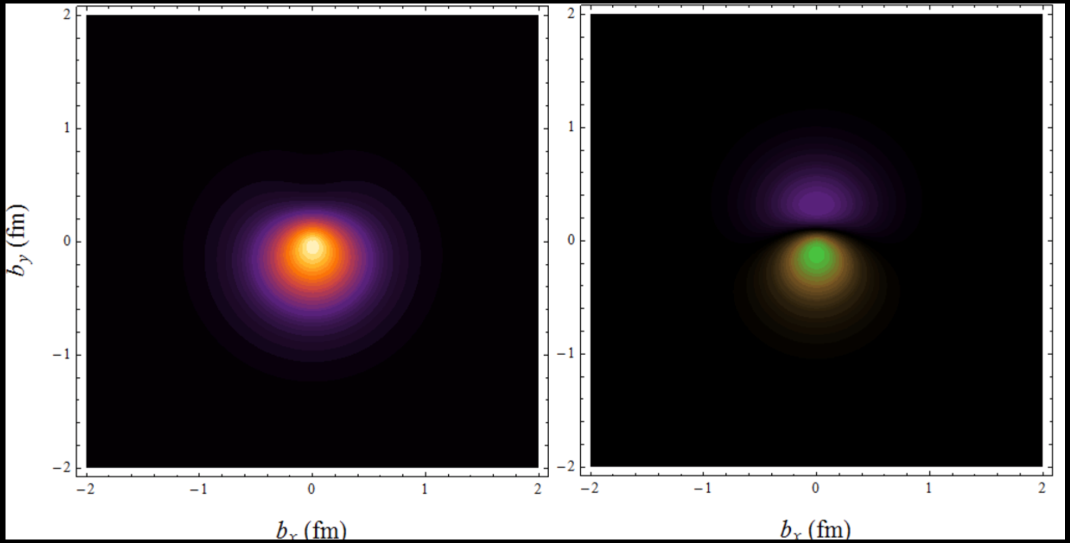
$$\mathrm{CFF} = \int \underbrace{(\mathrm{QCDKernel})}_{\text{Assume known from PQCD}} \times \mathrm{GPD}$$

Second Inverse Problem!

3D Coordinate Space Representation requires a Fourier transform of a GPD

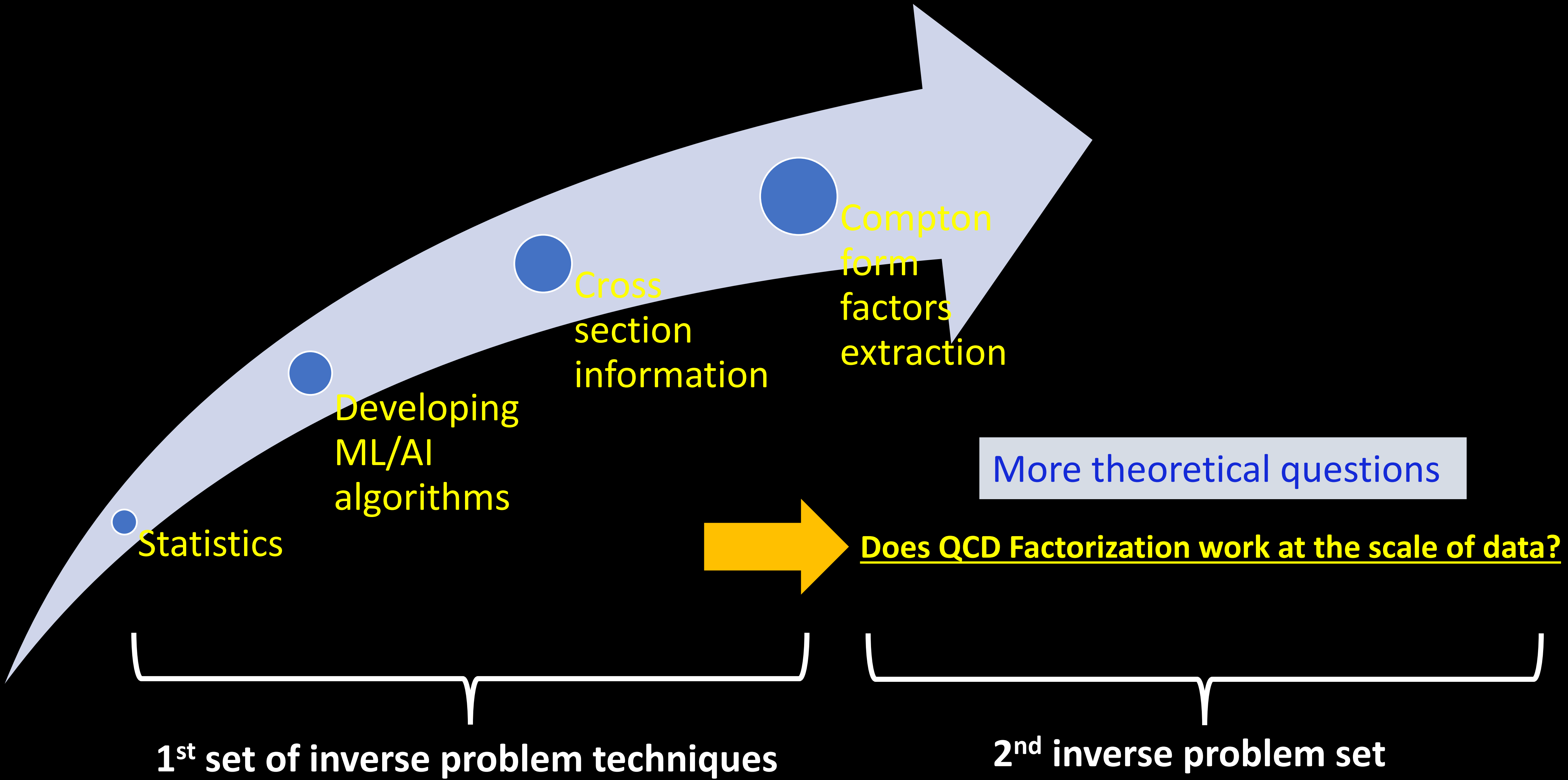
Observables from DVES matrix elements can be Fourier transformed from momentum space into coordinate space, providing insight into the spatial distributions of quarks and gluons inside the proton, besides matter and charge distributions.

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



H-W Lin,

https://web.pa.msu.edu/people/hwlin/research.html


Statistics
Developing ML/AI algorithms
Cross section information
Compton form factors extraction
More theoretical questions
Does QCD Factorization work at the scale of data?
1st set of inverse problem techniques
2nd inverse problem set

Introducing Maximum Likelihood Analysis of deeply virtual exclusive scattering data

Douglas Adams et al., arXiv:2410.23469

Working Framework: Bayesian statistical inference

- parameters are determined from **probability distribution functions (PDFs)** that quantify the degree of **uncertainty of our knowledge** of those parameters

- **PDFs** quantify the randomness inherent in a random variable and are in principle unknown but related to one another in a measurement process through Bayes theorem

- Bayes theorem

Likelihood : PDF of Data given Model (parameters θ)

$$P(\theta|D) = \frac{P(D|\theta) \times P(\theta)}{P(D)}$$

Posterior : PDF of Model (parameters θ) given Data (D)

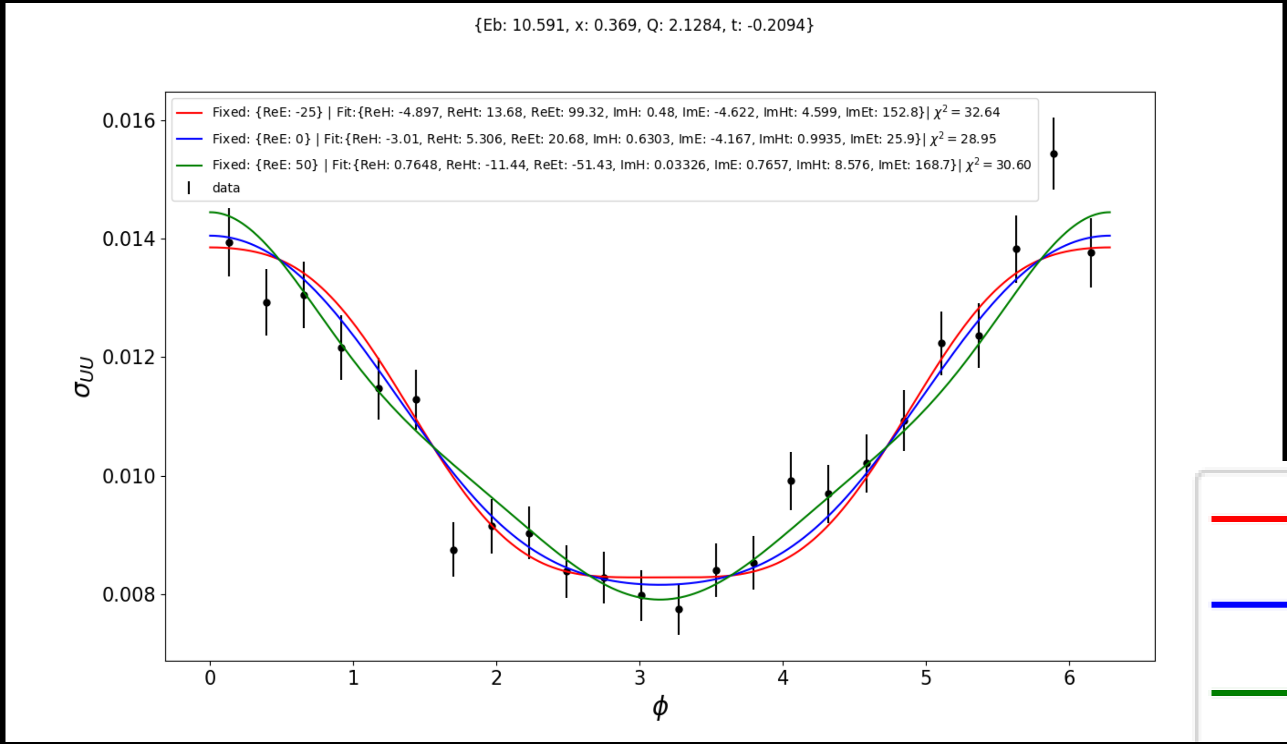
Prior: PDF of Model

Evidence/Normalization: PDF of Data

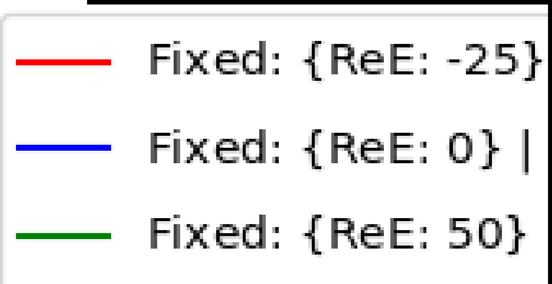
$$Posterior\left(CFFs\right) \approx \frac{Likelihood \times Prior}{Evidence} \approx \frac{Likelihood \times 1}{1}$$

Uninformative prior

Performing CURVE FIT with “standard methods” results in degeneracy



The extracted CFFs are significantly sensitive to the way the fit is done!



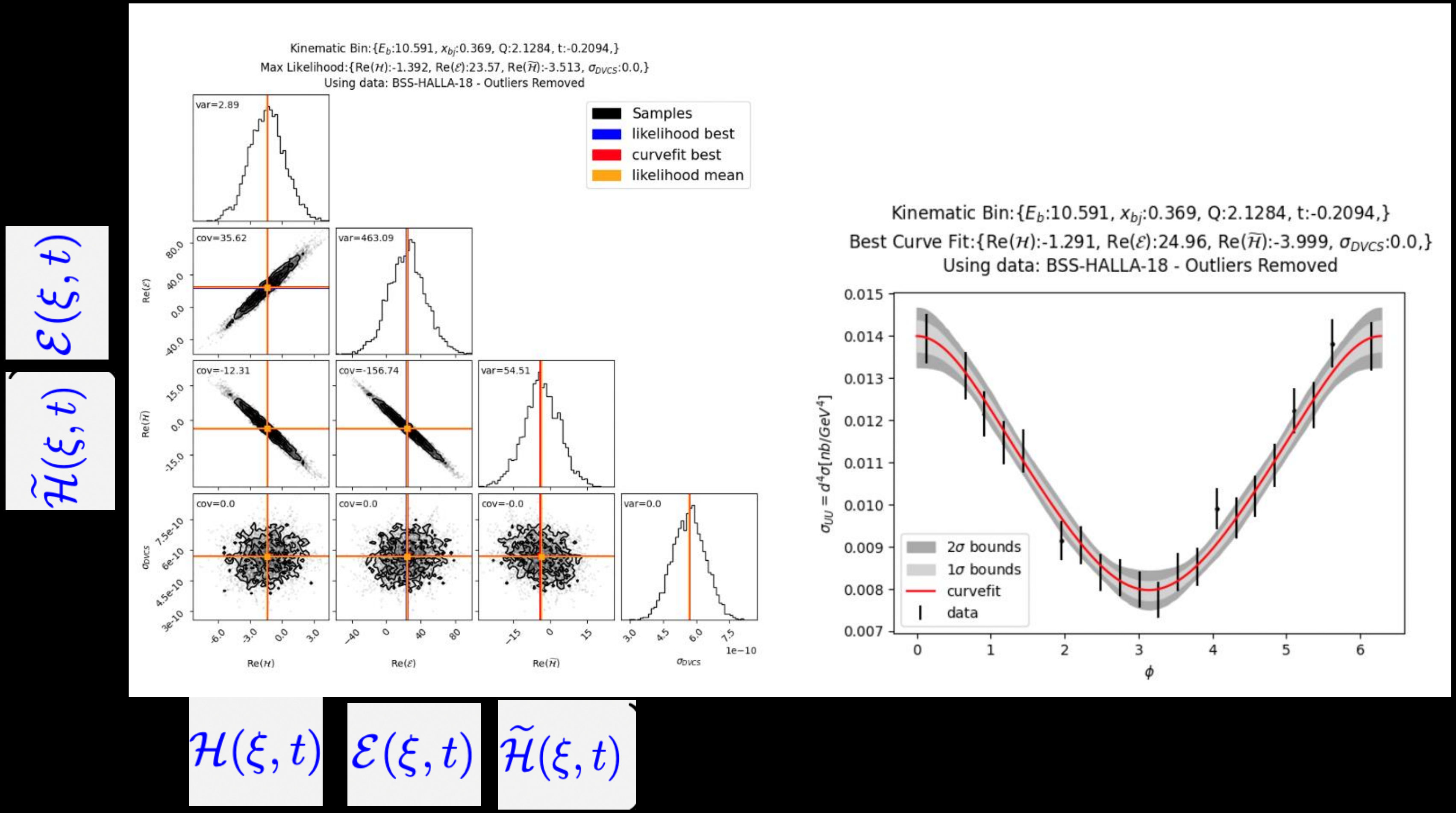
Working Framework: Bayesian statistical inference

$$Posterior\left(CFFs\right) \approx \frac{Likelihood \times Prior}{Evidence} \approx \frac{Likelihood \times 1}{1}$$

Uninformative prior

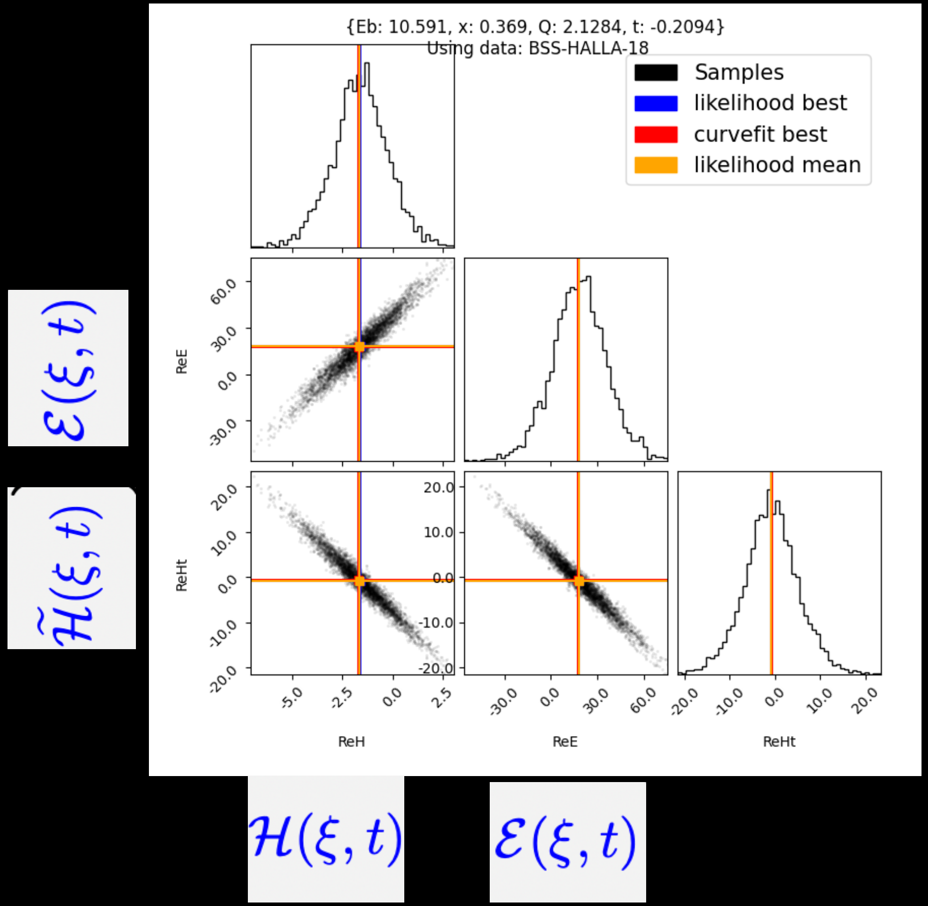
Evaluated Likelihood with MCMC/Metropolis Hasting Algorithm

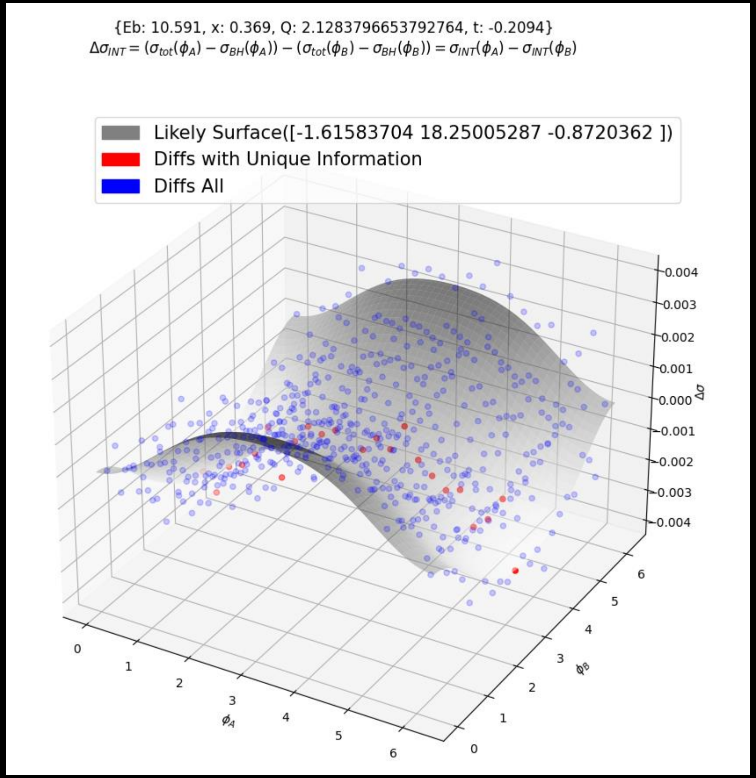
Deeply Virtual Compton Scattering Results



Covariance of CFF Results

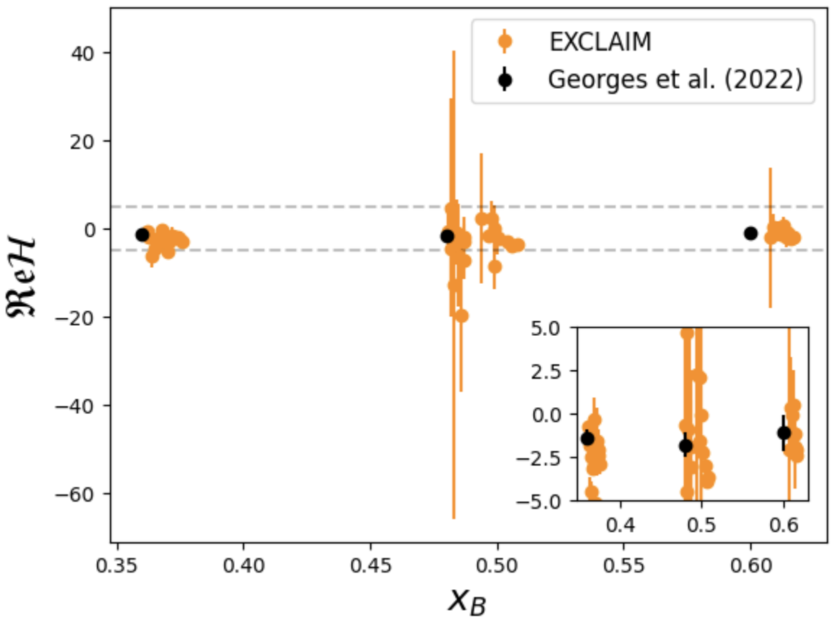
- Here the maximum likelihood is achieved allowing 3 CFFs to vary.
- Only 23 combinations of 2 angles are used.



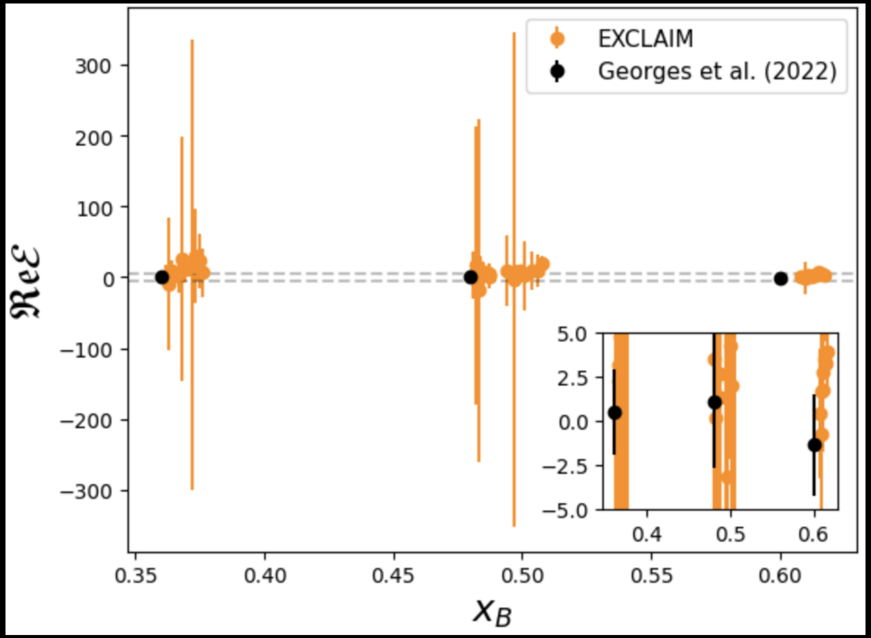


First fully quantitative/statistically sound extraction of CFFs from experiment

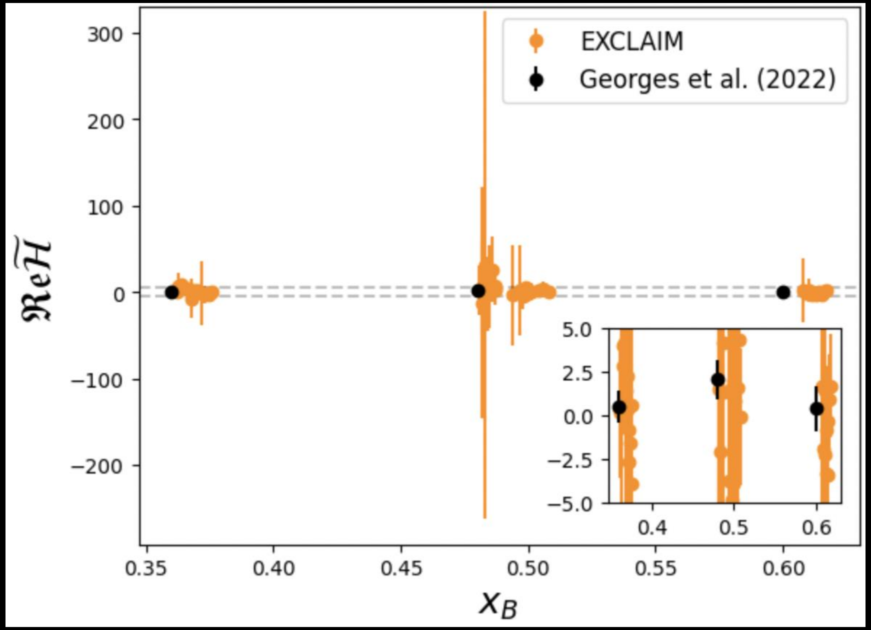
$\mathcal{H}$



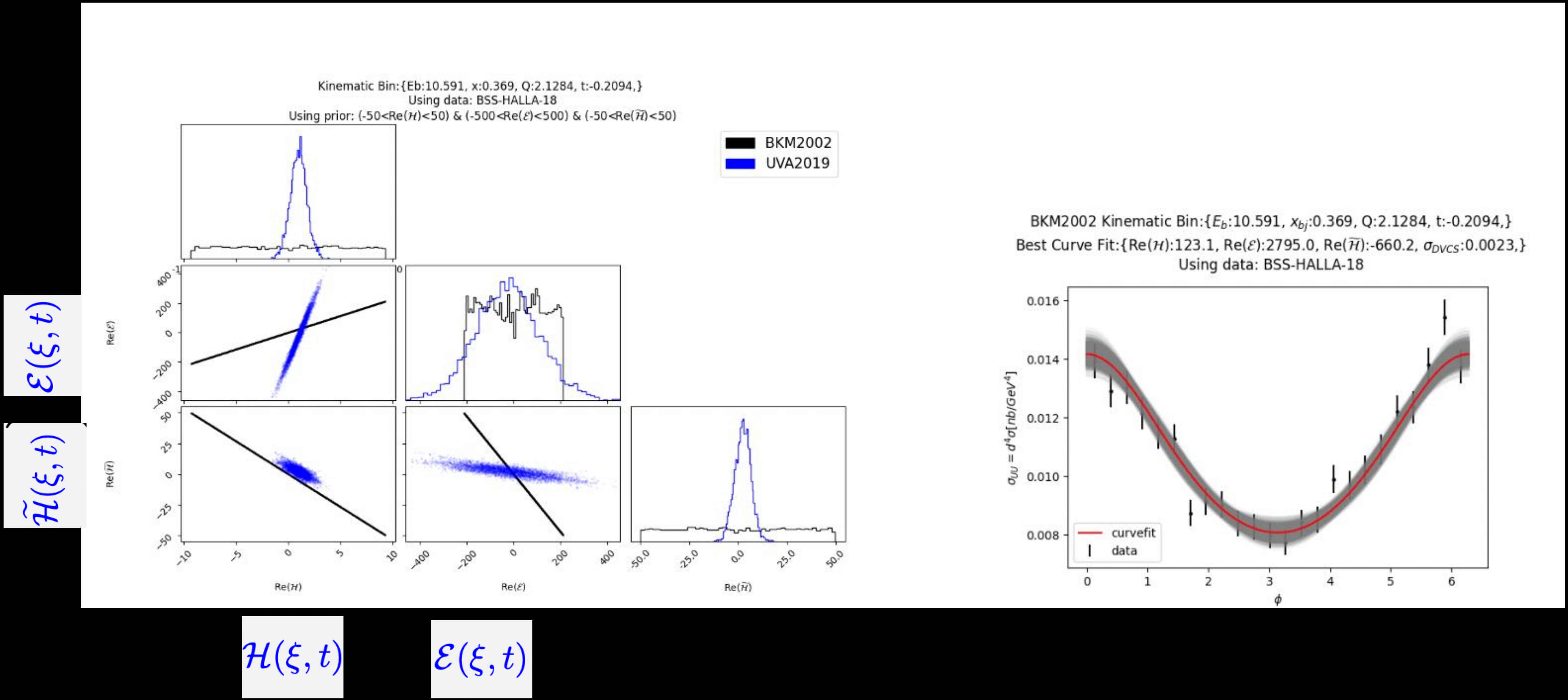
$\mathcal{E}$



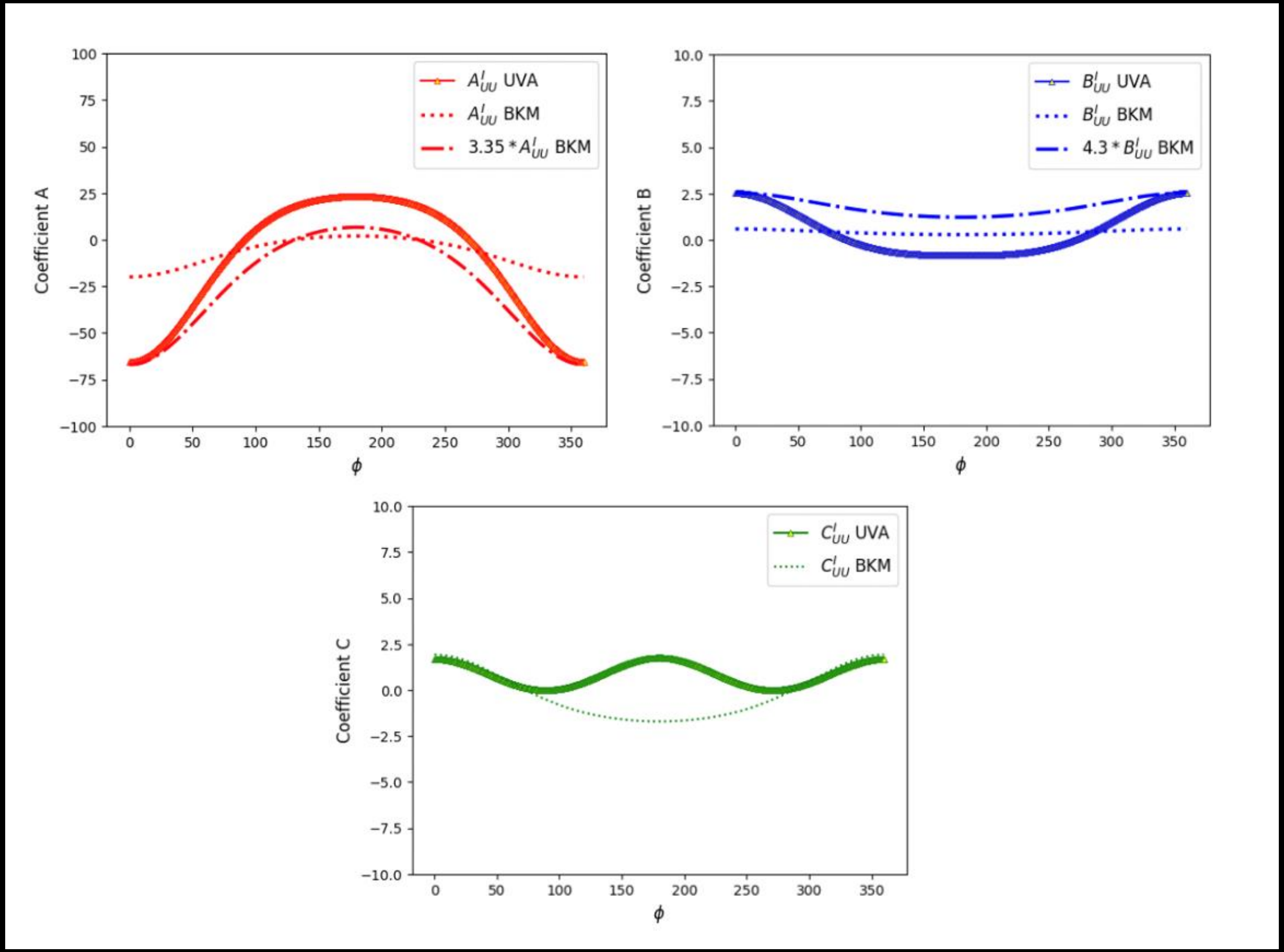
$\widetilde{\mathcal{H}}$



Accurate description of phase dependence of cross section is crucial: formalism of BKM, Braun et al. leads to degeneracy



Understanding the origin of degeneracy of BKM results



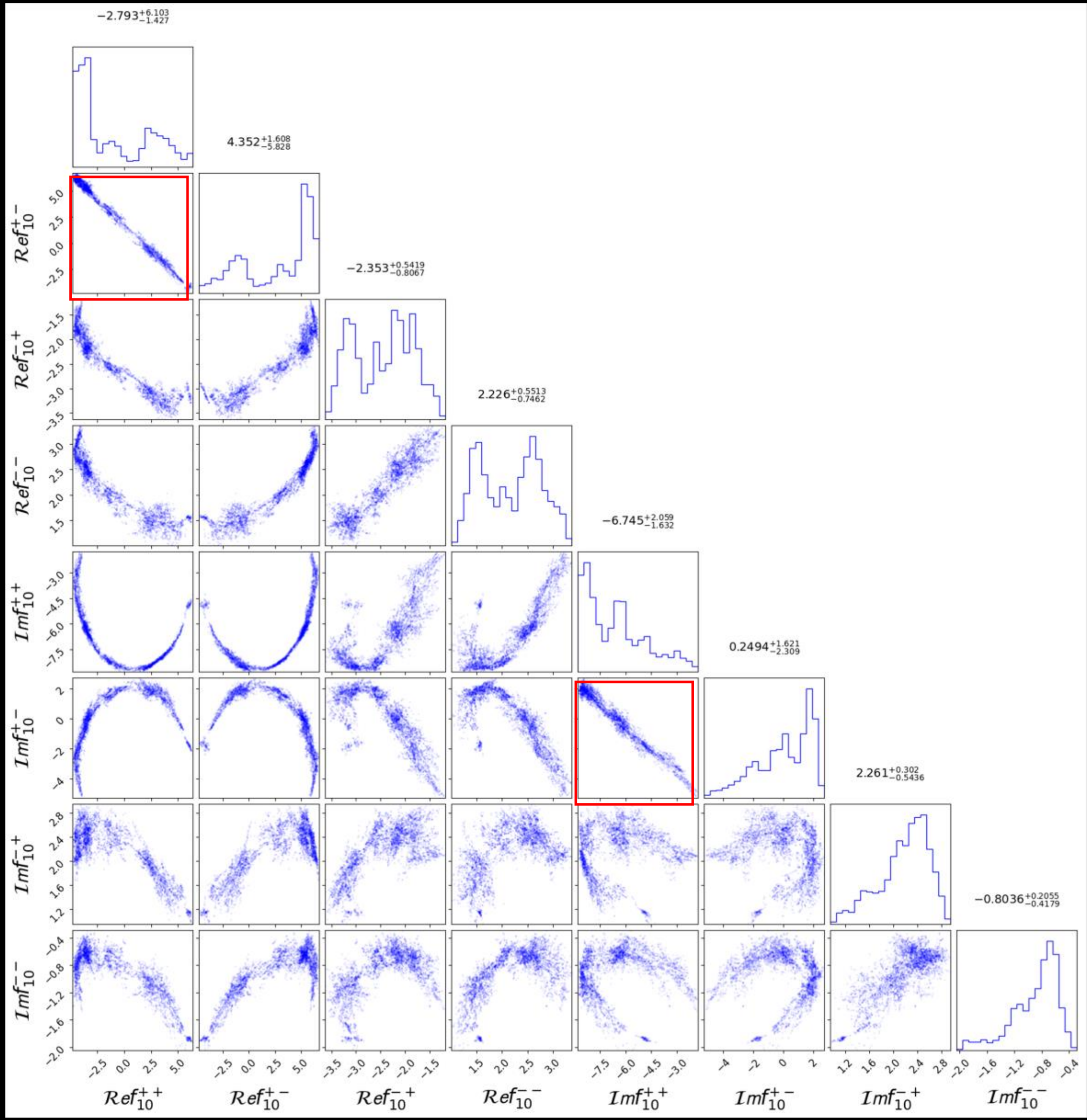
What's next? To reduce the error our analysis points at use of other observable keeping kinematics the same, before expanding phase space

(Polarization is a crucial handle)

$$F_{LL}^{\cos\phi} = A_{LL}^{\mathcal{I}} \Re\mathrm{e}\Big(F_1 (\widetilde{\mathcal{H}} - \xi \widetilde{\mathcal{E}}) + \tau F_2 \widetilde{\mathcal{E}} \Big) + B_{LL}^{\mathcal{I}} G_M \Re e \widetilde{\mathcal{H}} + C_{LL}^{\mathcal{I}} G_M \Re\mathrm{e} \left(\mathcal{H} + \mathcal{E} \right)$$

LL type data: E.Seder et al. [CLAS], PRL114 (2015)

Study role of QCD factorization in π^0 electroproduction a window on transversity GPDs



$Re\, f^{+-}$

$Im\, f^{+-}$

$Re\, f^{++}$ $Im\, f^{++}$

Saraswati Pandey

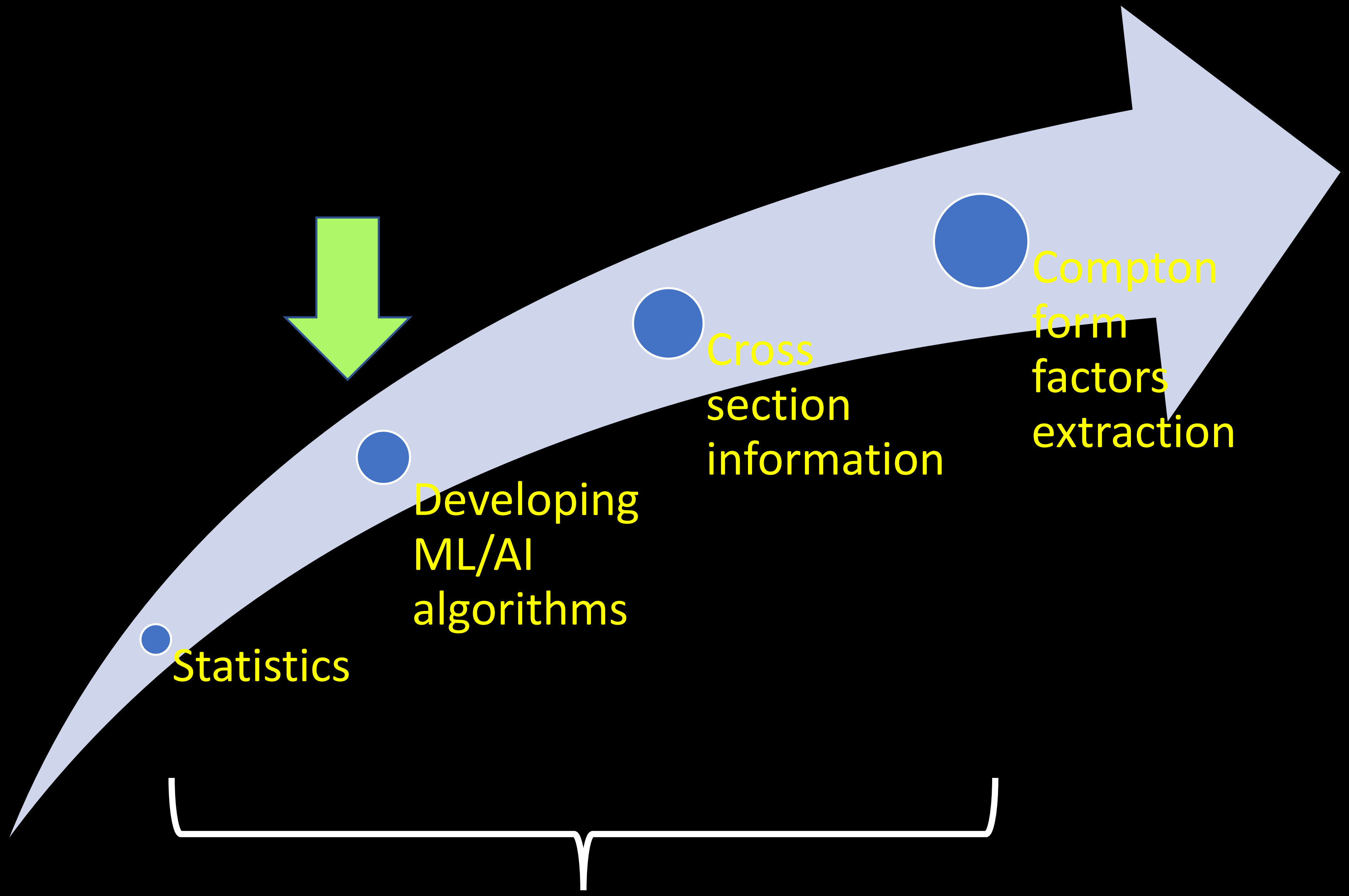
$$f_{10}^{++} \propto \Delta\left(2\widetilde{\mathcal{H}}_T + (1-\xi)\mathcal{E}_T - (1-\xi)\widetilde{\mathcal{E}}_T\right)$$

$$f_{10}^{+-} \propto \mathcal{H}_T - \frac{t_0 - t}{4M^2}\widetilde{\mathcal{H}}_T + \frac{\xi^2}{1-\xi^2}\mathcal{E}_T + \frac{\xi}{1-\xi^2}\widetilde{\mathcal{E}}_T$$

$$f_{10}^{-+} \propto \Delta^2\, \widetilde{\mathcal{H}}_T$$

$$f_{10}^{--} \propto \Delta\left(2\widetilde{\mathcal{H}}_T + (1+\xi)\mathcal{E}_T + (1+\xi)\widetilde{\mathcal{E}}_T\right),$$

using Hall A data, Dlamini et al.



1st set of inverse problem techniques

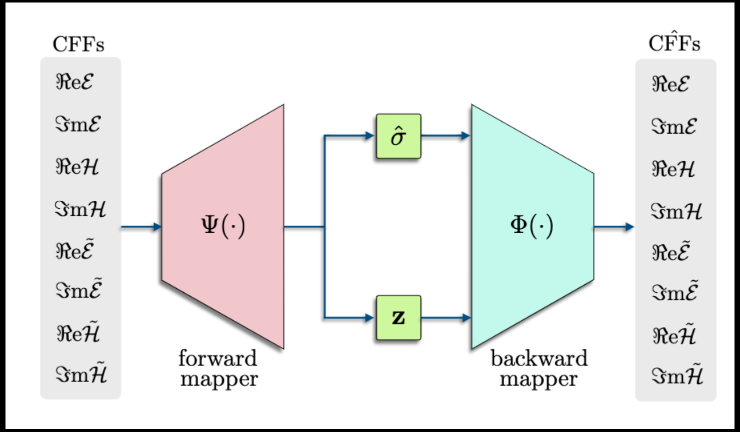
2. Inverse Problem Techniques: VAIM, C-VAIM, MCMC

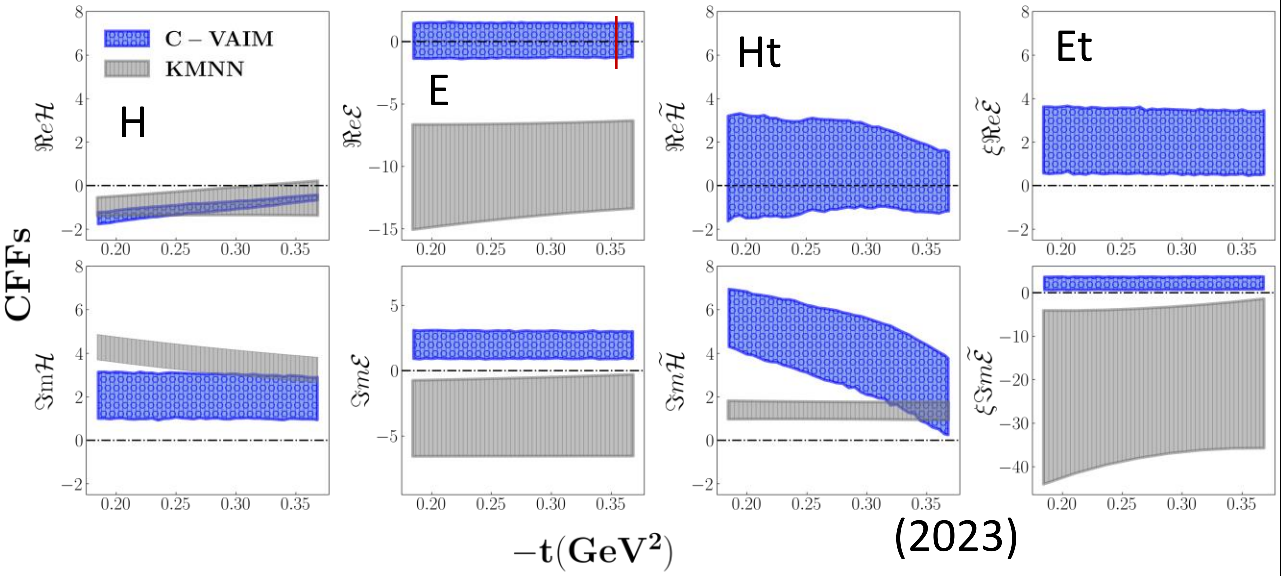
Approaches to find parameters statistically in an underdetermined system

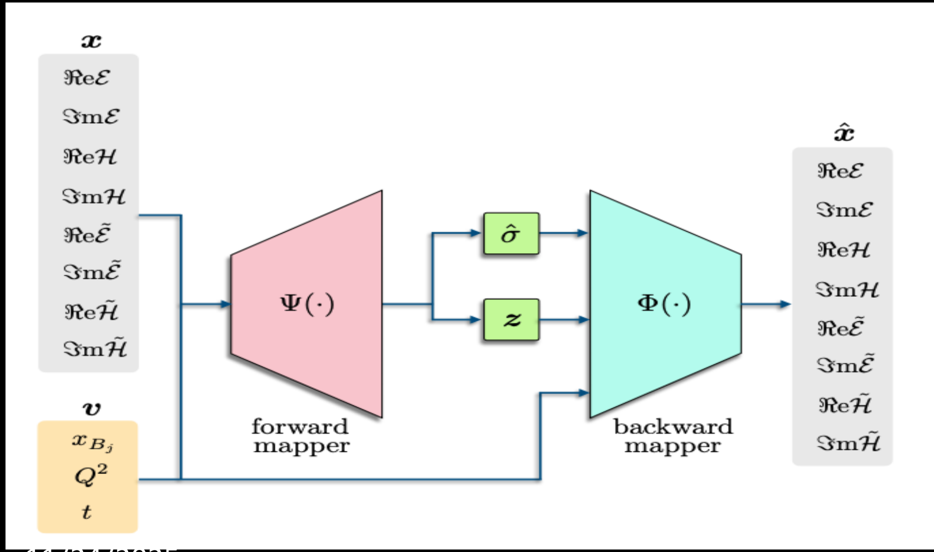
- Can quantify parameter uncertainty when more parameters than data
- Techniques highly dependent on bounded parameter priors
- These methods give us an initial way to perceive:
  1. → the correlation between parameters on a complicated model
  2. → what information is missing (latent space)

A variational autoencoder inverse mapper solution to Compton form factor extraction from deeply virtual exclusive reaction

arXiv: 2405.05826 (published in EPJC 2025)

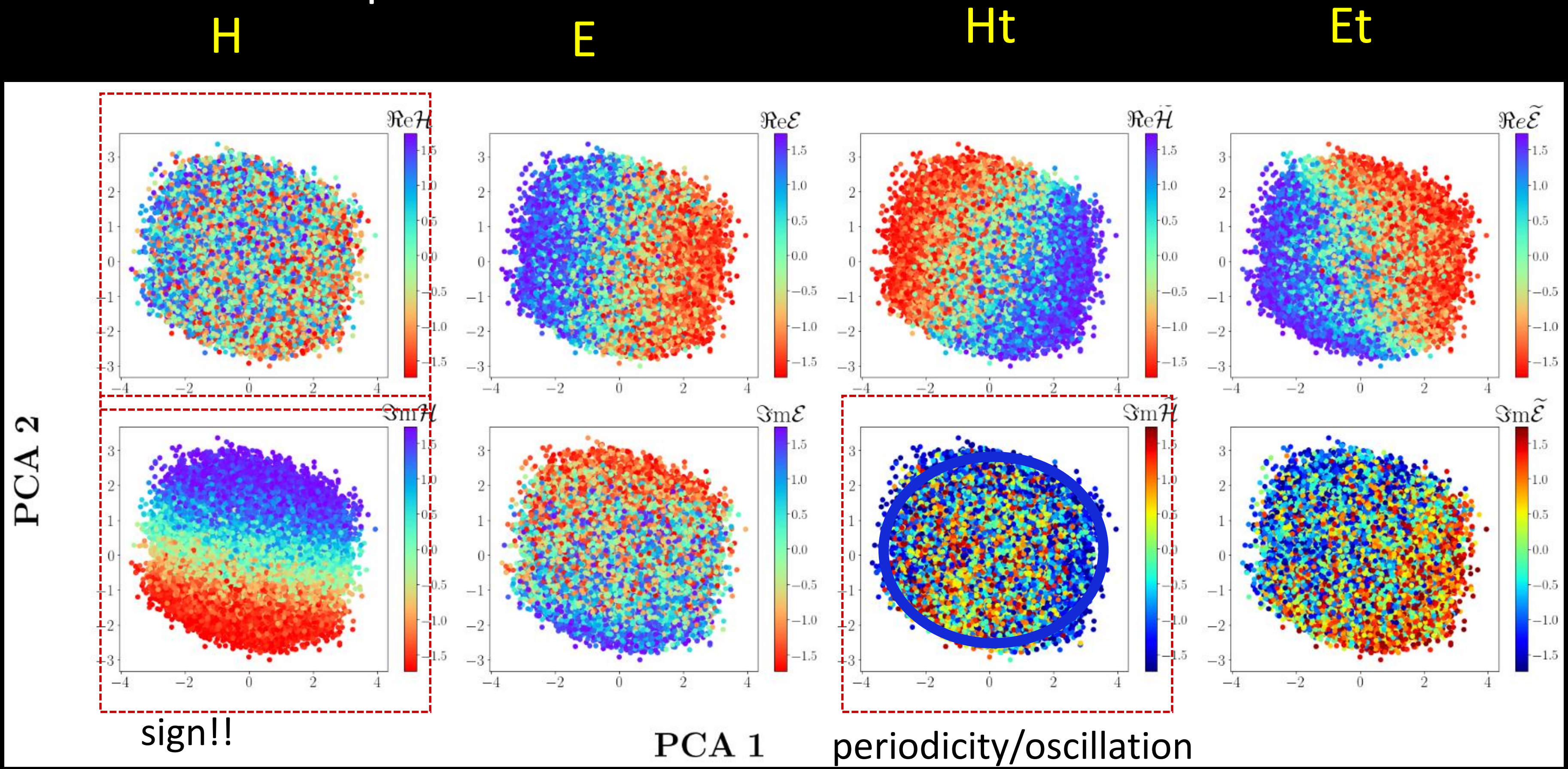






- KMNN, https://arxiv.org/abs/2007.00029

CFFs Analysis of Latent Space



3. Symbolic Regression for Parton Research

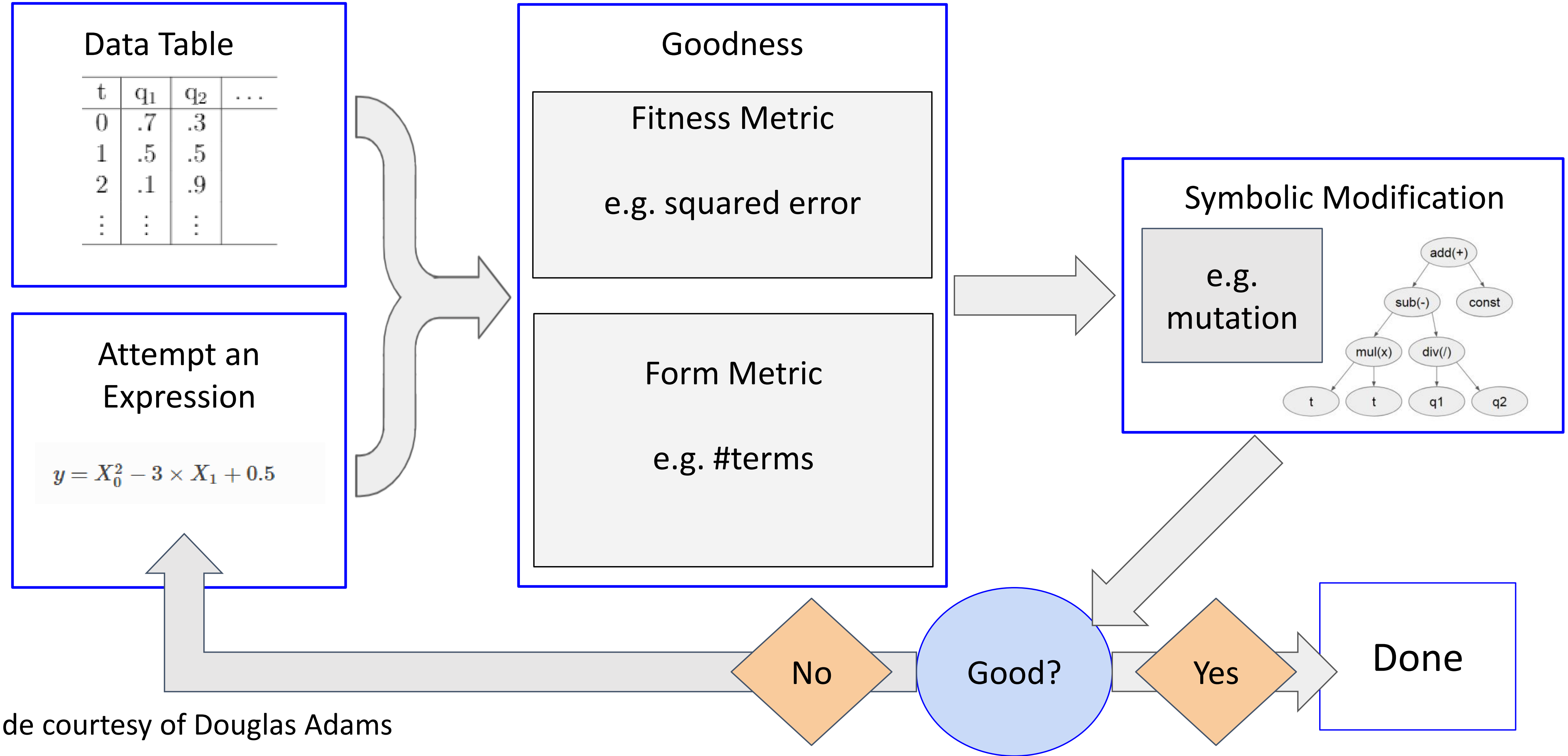
1) What is symbolic regression and why do we care?
   a) Spoiler: because then humans can read the answer

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)

Slide courtesy of Douglas Adams

Symbolic regression (SR) in a nutshell:



Slide courtesy of Douglas Adams

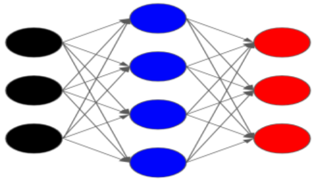
Why bother with SR when we have neural networks?

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

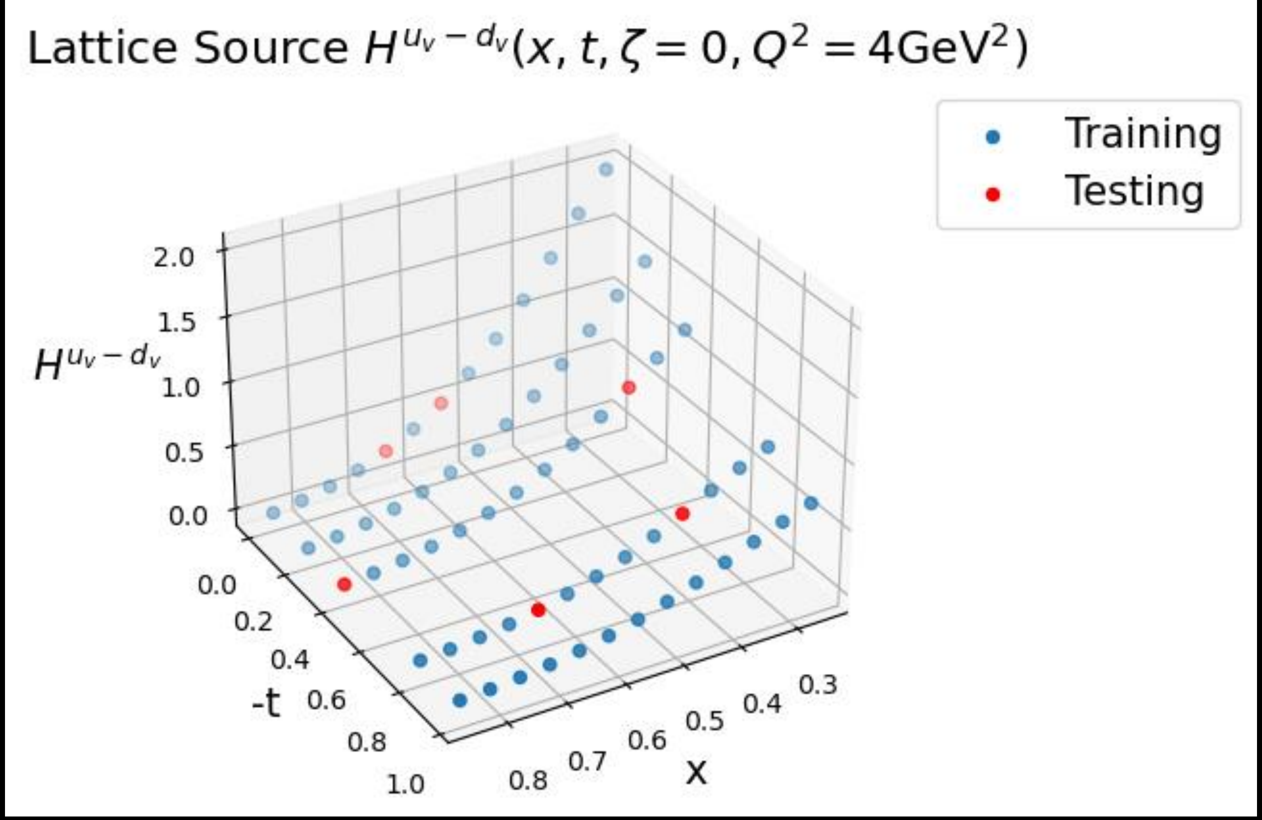


$$q_{i,BB}(t) \approx \ldots$$

Slide courtesy of Douglas Adams

Lattice QCD can predict the x dependence of GPDs for various t:

1) Can SR help us find an interpretation of the numerical x dependence?
2) Do the x and t dependences factorize?



H.-W. Lin, Phys. Rev. Lett. 127, 182001 (2021)

Does x and t dependence of GPDs factorize? Why is it important to check x and t factorization?

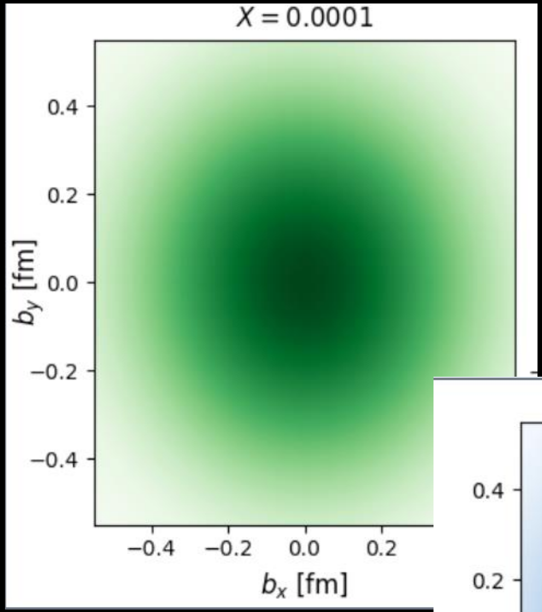
3D Coordinate Space picture

$t = -\Delta_T^2$

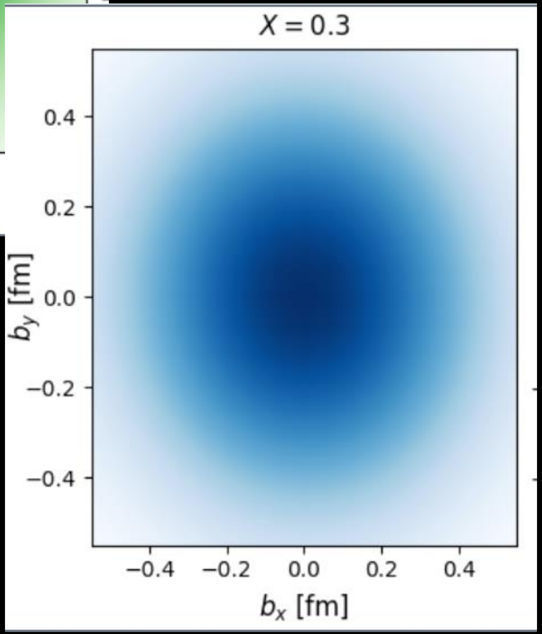
$$\underbrace{\boxed{\mathcal{H}^q(X,0,b_T)}}_{\text{3D Coordinate Space picture}} = \int \frac{d^2\Delta_T}{(2\pi)^2} \underbrace{\boxed{H^q(X,0,\Delta_T)}}_{\text{GPD}} e^{-i\Delta_T \cdot b_T}$$

GPD

gluon



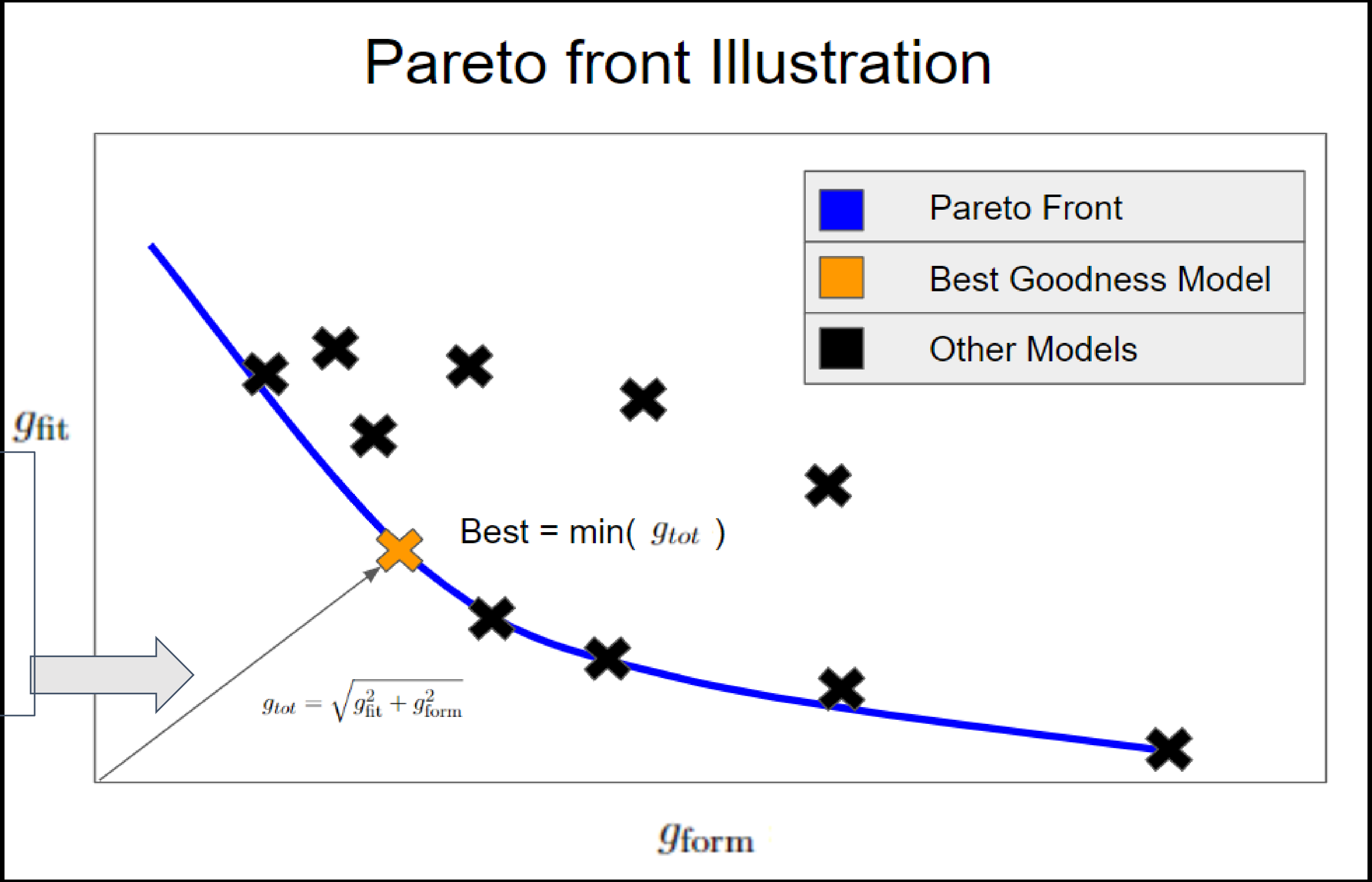
d-quark



If x and t dependences factorize, the radius of the partonic distributions is constant throughout the proton volume!

Z. Panjsheeri

Using a pareto front to choose amongst forms

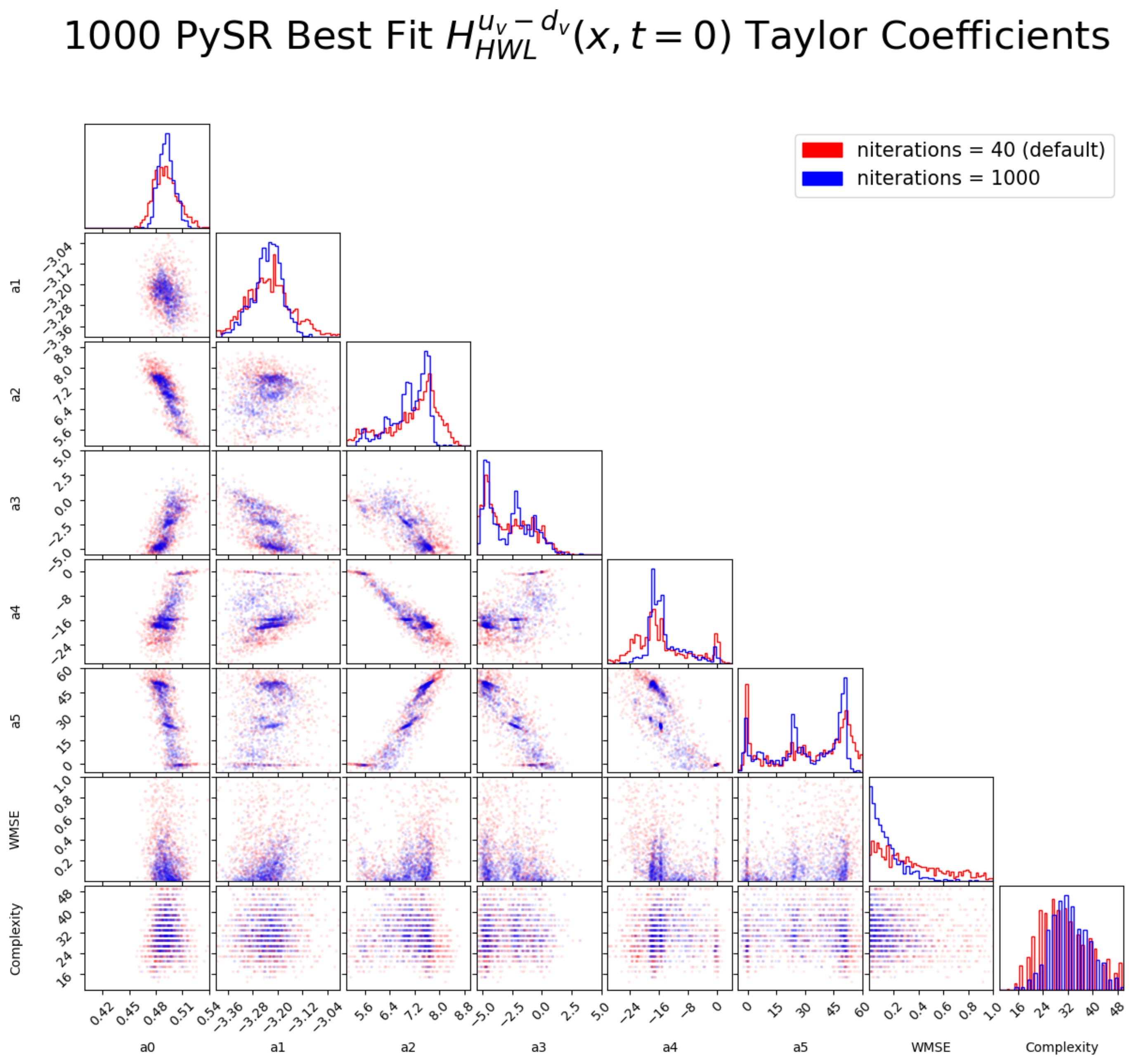


define a "metric" which balances fit vs. form to choose a best

“Convergence” criterion using Taylor expansion around points in x

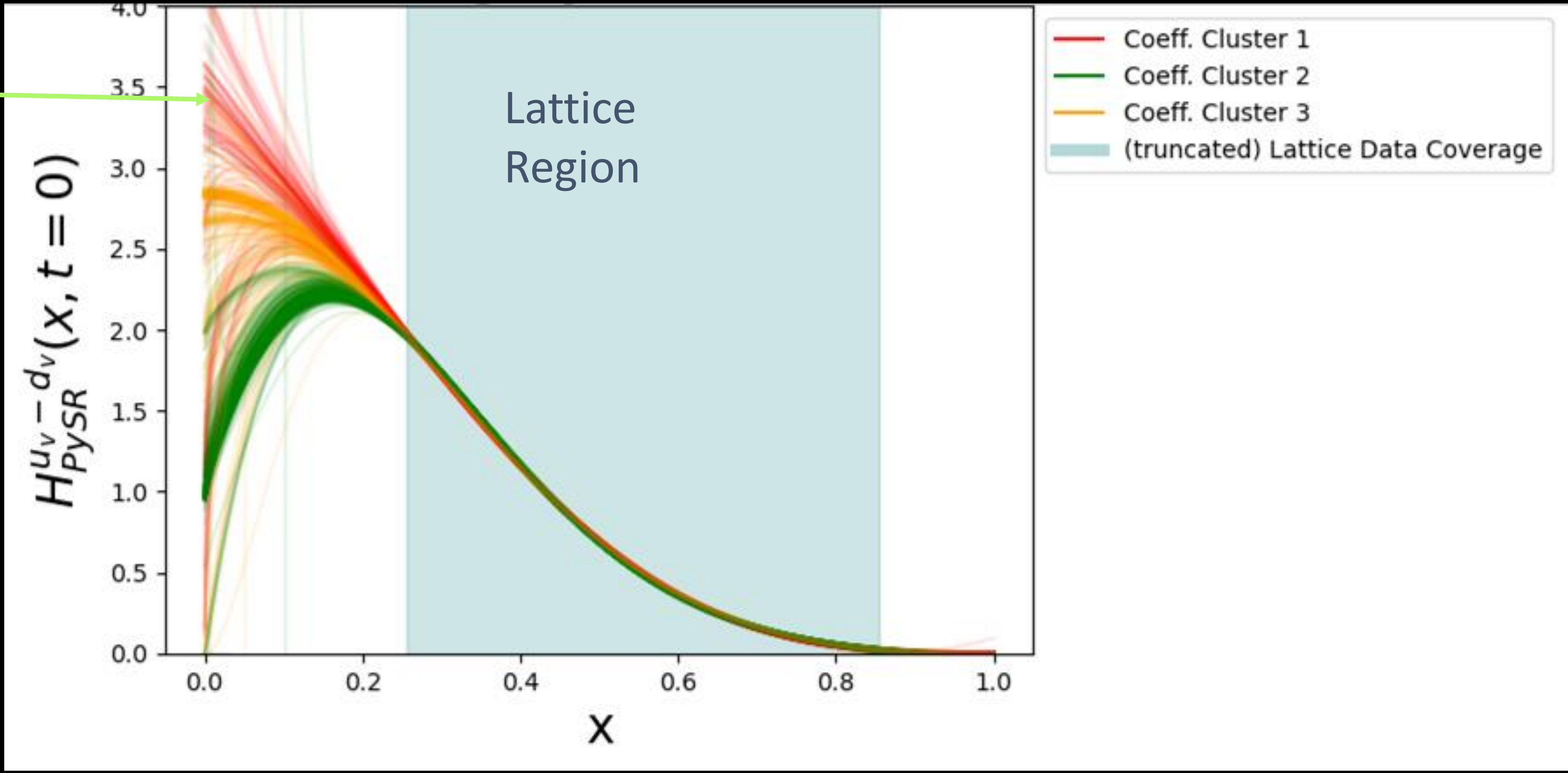
PySR convergence

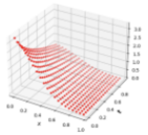
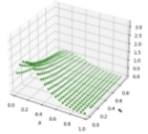
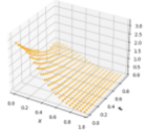
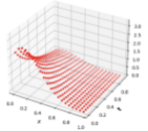
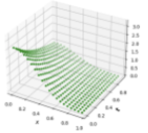
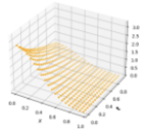
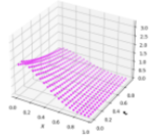
Andrew Dotson



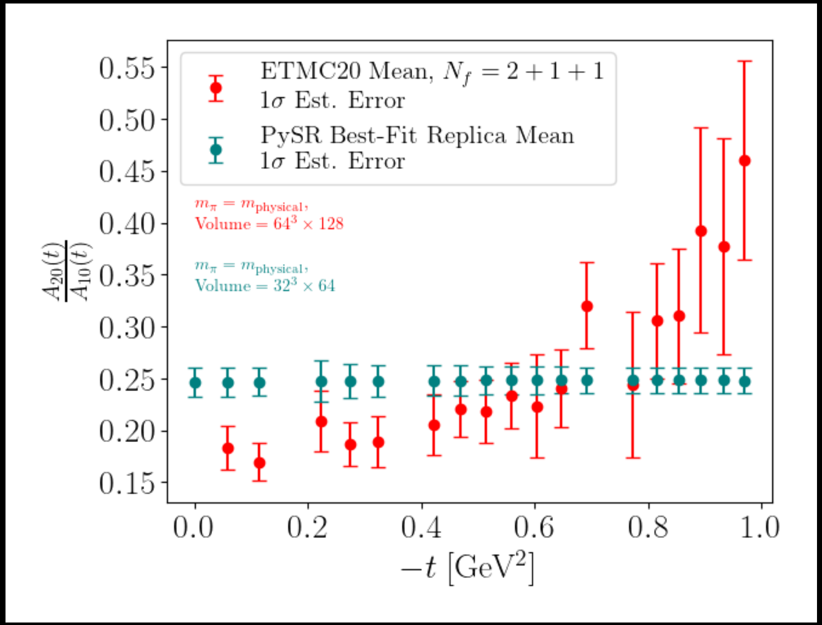
Novel SR Convergence Clustering

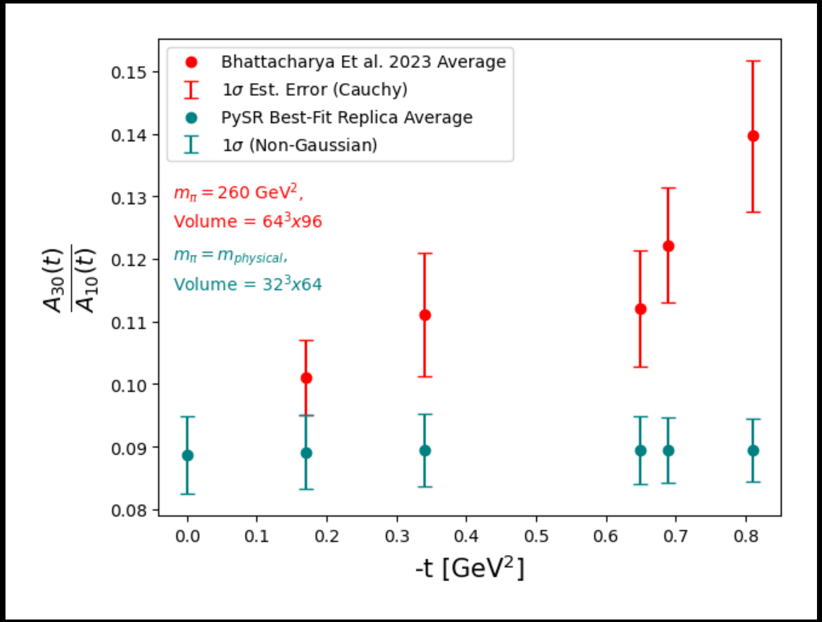
Extrapolation



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

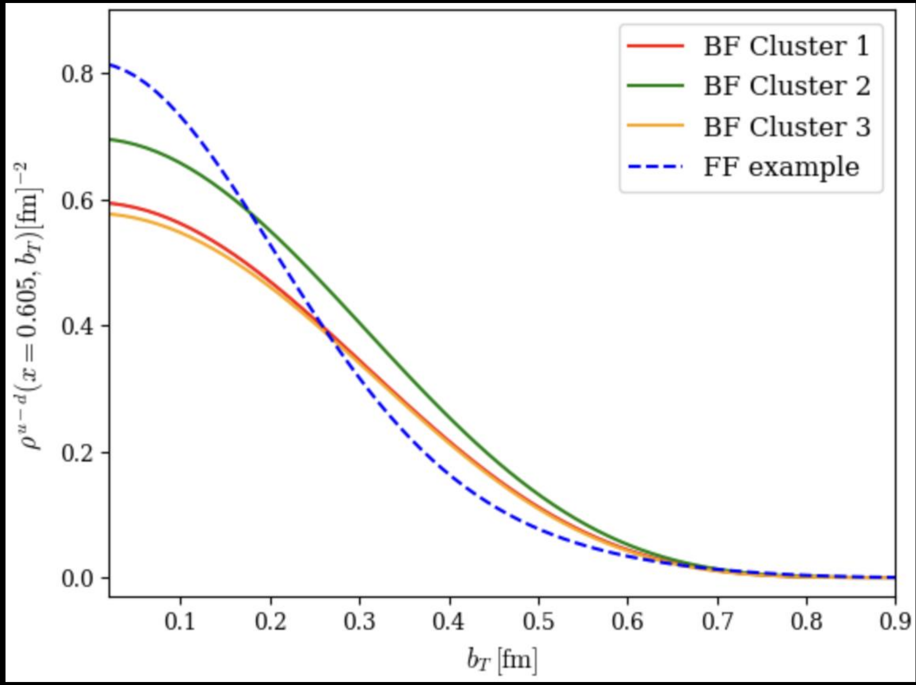
Comparison with lattice QCD Moments: testing x,t factorization



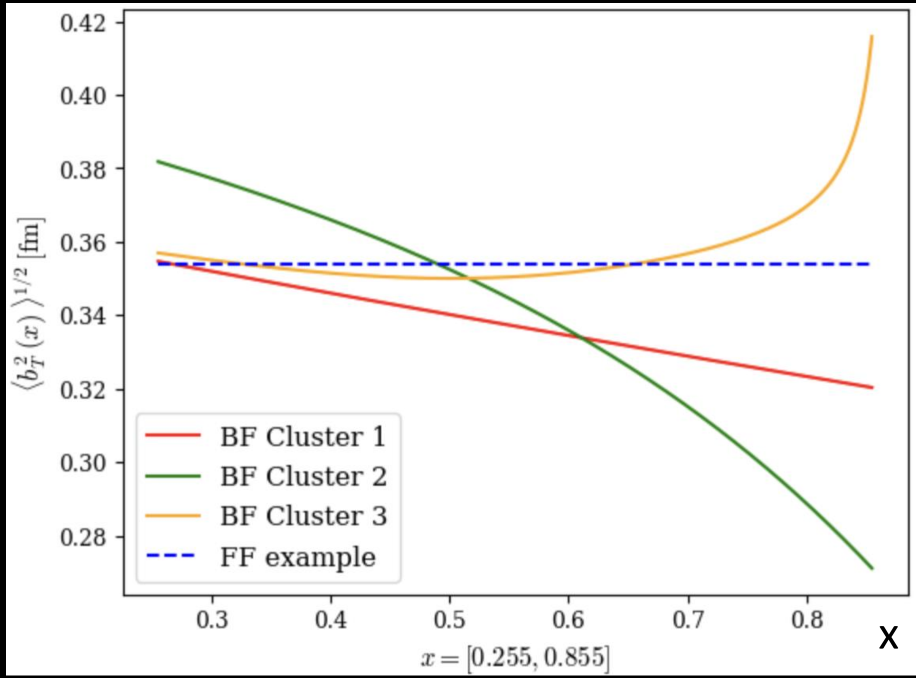


Fourier transform

Density at fixed x



rms Radius of partonic configurations vs x



4. NNGPD and End to End problem

Gia-Wei Chern, Yaohang Li

$$M_n = \int x^{n-1} \times GPD$$

angular momentum, calculable in lattice QCD

$$\mathrm{CFF} = \int (\mathrm{QCDKernel}) \times \mathrm{GPD}$$

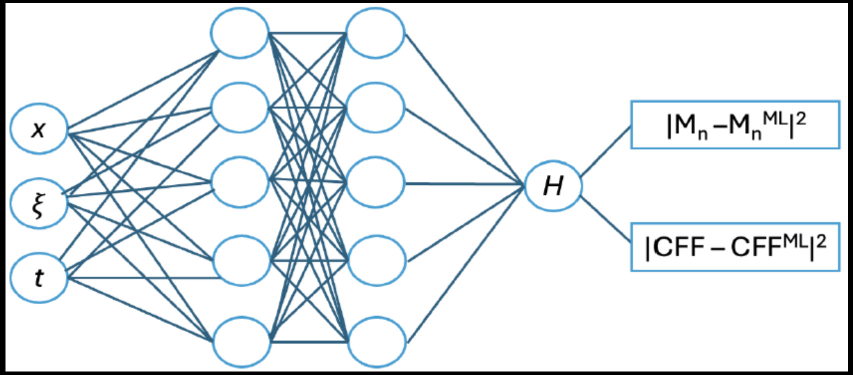
experimental observable

GPD

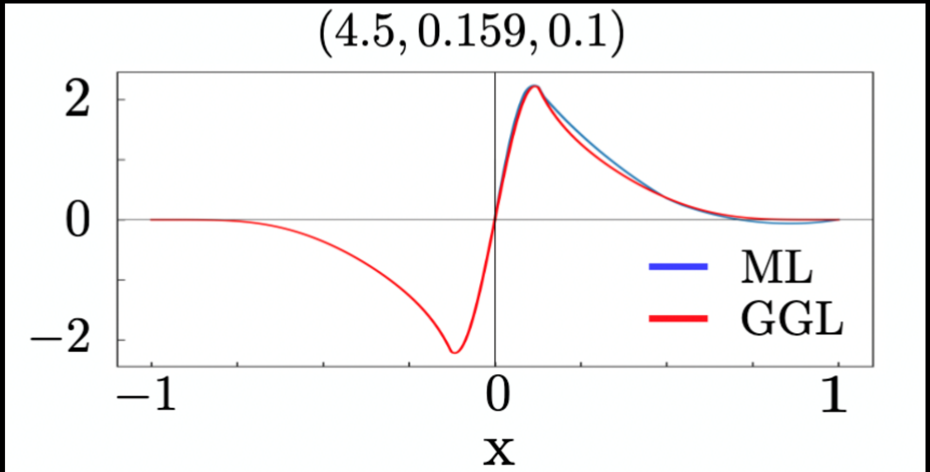
Is information on angular momentum directly extractable from data?

NNGPD

Standard feed forward neural network for the extraction of GPDs from data, lattice QCD and symmetries



Q^2 ξ, t



Preliminary results from Jason Ho (UVA)

Diffusion model

Yaohang Li

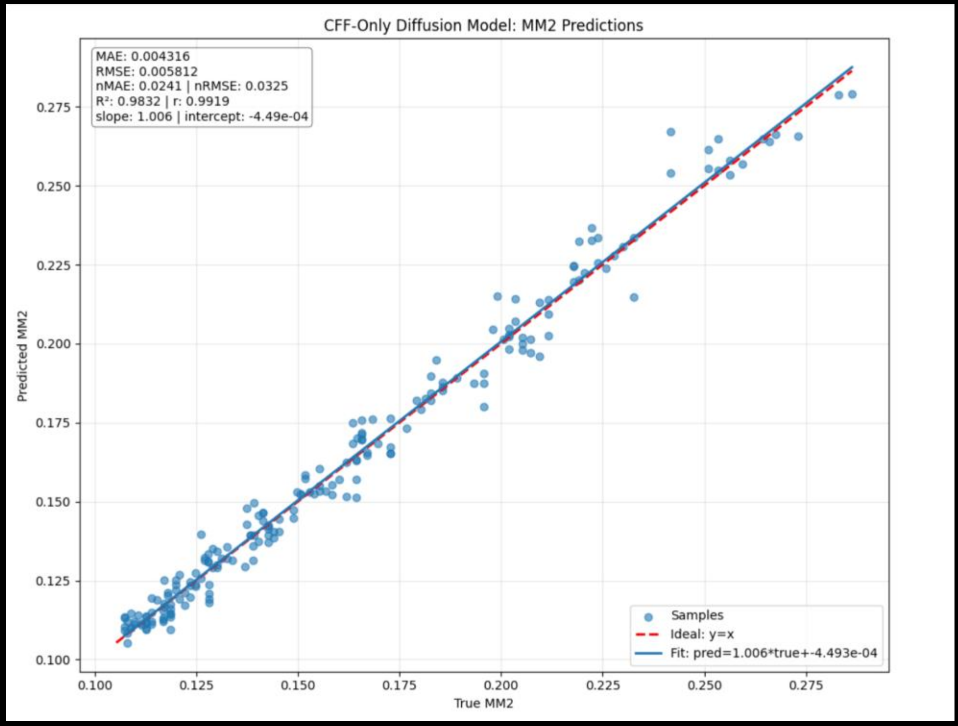
$$\mathcal{L} = \mathbb{E}_{x_0,t,\varepsilon}[||\varepsilon - \hat{\epsilon}_\theta(x_t, t, CFF)||^2]$$

Standard noise-prediction loss without any additional prior term:

- x_0: the true MM2 value (target)
- t: the diffusion time step, randomly sampled from {1,...,T} each iteration
- ε: Gaussian noise sampled from N(0,I), which is directly linked to the target x_0 because it is the exact noise we added to x_0 in the forward process to obtain x_t
- x_t: the noisy MM2 obtained from the forward process

- ε noise predicted by the neural network, conditioned on the CFF values (ReCFF, ImCFF)

predicted



true

preliminary results from Jitao Xu

Future work

We are developing a novel approach to foundation models (FMs) for theoretical hadronic physics

This is an exciting time for NP&AI, but what are the challenges and how to address them?

A few suggestions up for discussion

- Forming/Developing the NP/AI community

  More topical meetings
  (IAIFI, AI4EIC are very necessary and even need to be expanded but “broad”, workshops at DNP are great but isolated episodes)

- Dissemination of results

  -- New subsection of Physical Review (reviewrs are often not up to the task)

  -- Common “repositories”/websites modeled after e.g. LHAPDF (existing ones are based in Europe and hard to use)

- Jobs/retention of workforce

Conclusions

1. **Extracting QCD-based information from data**

    1. relies on our ability to understand the **cross section** for **all the various DVES processes**
    2. requires solving **multiple inverse problems**

2. **Key example: spatial structure of the proton** (and all of its mechanical properties including angular momentum)
    - We have defined a path to extract the **observables** from experiment that allows us to fully take into account UNCERTAINTY QUANTIFICATION (UQ) from data using likelihood analysis and MCMC
    - We have detected inconsistencies leading to degeneracy in approximated formalism of deeply virtual exclusive processes
    - On one side: our path to **AI&theoretical physics** brings **interpretability** features in the computation and benchmarking of AI tools
    - **On the other: diffusion models provide a reliable method to extract angular momentum from data**