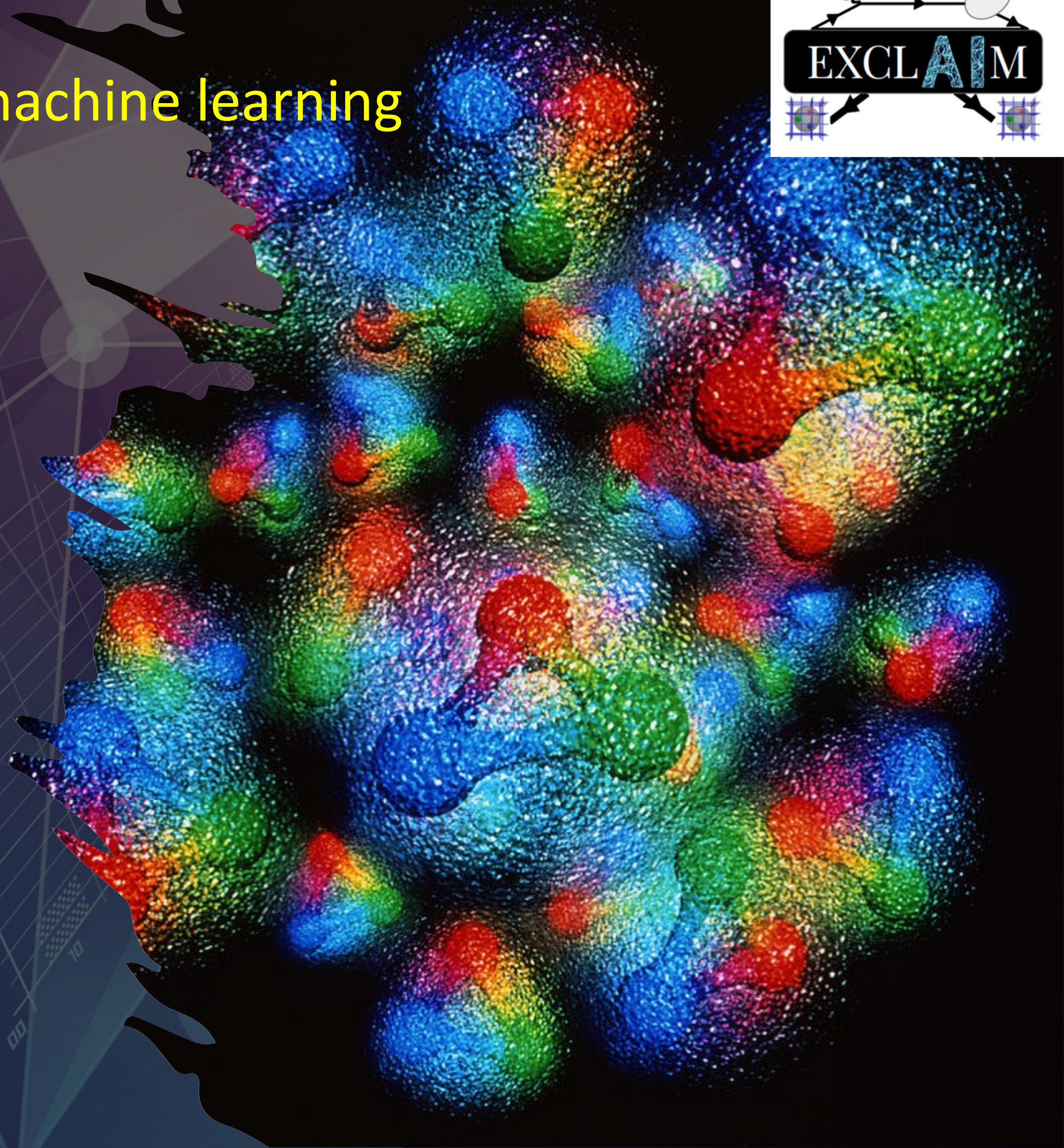
Proton 3D structure from leading-edge machine learning

Highlights from the EXCLAIM project

REF 2024
Paris, Saclay
October 14th -18h 2024

Simonetta Liuti



I am grateful for the contributions of

Douglas Adams (postdoc), Jang (Jason) Ho, Adil Khawaja, Saraswati Pandey, Zaki Panjsheeri (graduate students)

A new community is emerging: AI and Theoretical Nuclear and Particle Physics

2024 workshops/meetings

Inverse Problem @ UQ	INT, July 2024: Inverse Problems and Uncertainty Quantification in Nuclear Physics
IAIFI@MIT	Institute for Artificial Intelligence and Fundamental Interactions August 2024 : IAIFI@MIT Summer Workshop
DNP Symposium	APS DNP Meeting, October 2024 : Mini-Symposium -- From Data to Discovery: Machine Learning8, 2024 Hilton Boston Park Plaza Room: Arlington, Mezzanine Hilton Boston Park Plaza Room: Arlington, Mezzanine Level

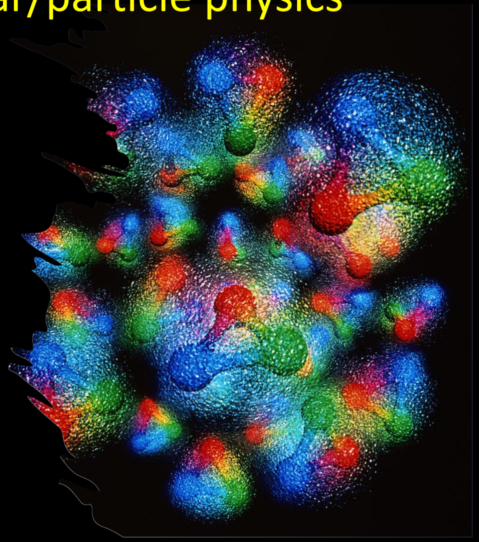
https://exclaimcollab.github.io/web.github.io/#/

The EXCLAIM collaboration is part of this community

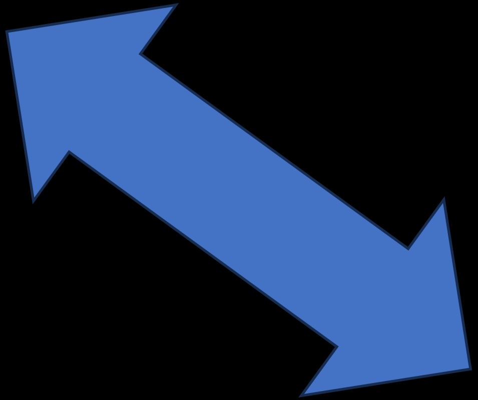
CoPIs: Marie Boer, Gia-Wei Chern , Michael Engelhardt, Gary Goldstein, Yaohang Li, Huey-Wen Lin, SL, Matt Sievert, Dennis Sivers

Postdocs: Douglas Adams, Marija Cuic, Saraswati Pandey, Emanuel Ortiz, Kemal Tegzin, William Hockley

Nuclear/particle physics



"Leverage AI to understand the unsolved theoretical questions in nuclear/particle physics"



Generative AI → Foundational AI

Use physics concepts to inform the NN

EXCLAIM: physics aware deep learning models with theory constraints

- ML is not treated as a set of “black boxes” whose working is not fully controllable

- Utilize concepts in *information theory and quantum information theory* to interpret the working of ML algorithms necessary to extract information from data

- At the same time, use ML methods as a testing ground for the working of quantum information theory in a large class of deeply virtual scattering processes

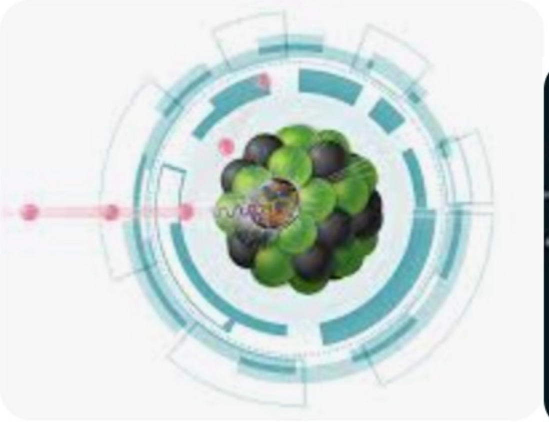
Key Questions in Hadronic Physics

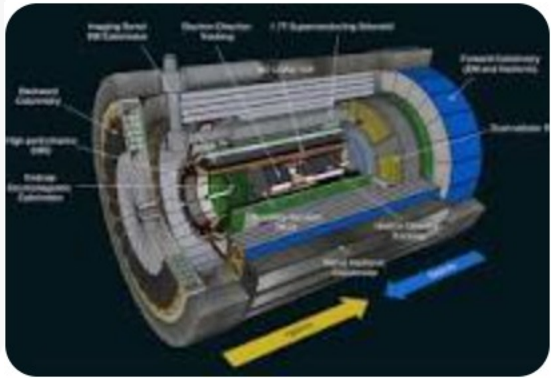
- How do quarks and gluons generate the strongly interacting particles and their properties (mass, spin)?
- What are the phases of strongly interacting matter?
- How does the structure of nuclei emerge from the strong interactions?
- How does hadronic physics contribute to BSM searches?

How do we address these questions in experiment?

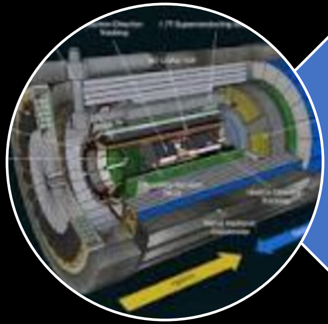
Jefferson Lab, LHC and EIC

- Exploring 3D tomography of the nucleon
  - Going beyond one-dimensional PDFs to map out the proton spin in terms of quark and gluon spins, including their orbital motion
- Observables:
  - Nucleon transverse structure in momentum space: TMDs
  - Nucleon transverse structure in coordinate space: GPDs

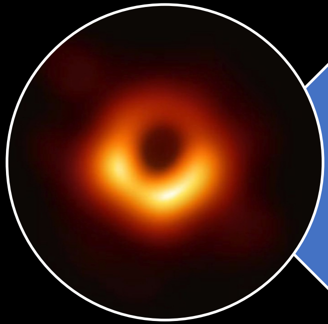




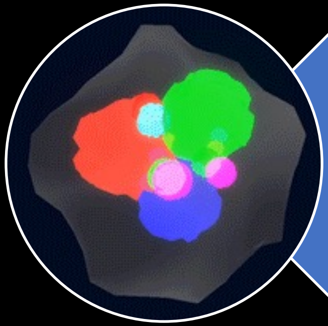
Epic detector https://www.bnl.gov/eic/



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**."



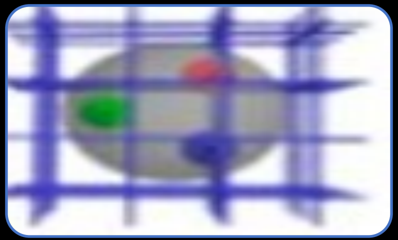
An example that we are well aware of:
The Event Horizon Telescope (EHT) image of blackhole at 55 M lightyears = 5 x 10^{23} m



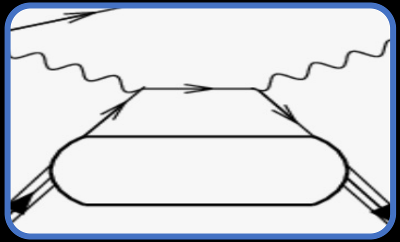
Our NP challenge: can we image the proton by observing its spatial structure at 1 fm = 10^{-15} m?

What do physicists do?

QCD Theory

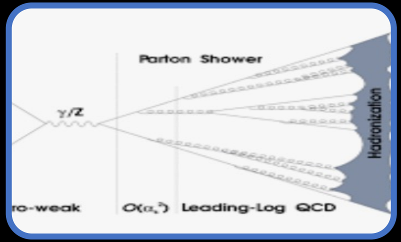


Factorization

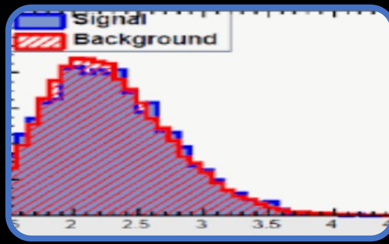


pQCD corrections

Hadronization



Events reconstruction



simulation simulation simulation simulation

Forward process

Inverse process: inference

Most analyses so far give

“catchy punchlines and compact, approximate representations, even when those are unjustified and unnecessary”

Hogg, Bovy and Lang, astro-ph: 1008.4686

Experimental Data, Simulations and Inference

1

In the analysis of the 3D structure of the proton, can we define a generative model (statistical procedure that could have generated the data set) that will reproduce data samples from a distribution?

2

Can one deduce a high-dimensional/multi-variable *prior* directly from data?

3

How does one sample and model the density distribution?

Likelihood Analysis

- **Likelihood**: probability of the observed data given our cross section model
- **Problem**: we want to find an N-dimensional surface in parameter space that maximizes the likelihood

$$\mathcal{L} = \prod_i \mathrm{Gaussian}(x(\phi_i), \mu(\phi_i), \sigma)$$

$$\sigma \rightarrow uncertainty$$

$$\mathrm{Gaussian}(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right]$$

The likelihood uniquely defines the **posterior** probability density distribution:

$$p(\theta = [H, E...]|data\&prior) = p(data|\theta, prior) \times \frac{p(\theta|prior)}{p(data|prior)}$$

posterior likelihood "normalized" prior probability

Likelihood Analysis: Bayes Theorem

The likelihood uniquely defines the **posterior** probability density distribution:

$$\underbrace{p(\theta = [H, E...]|data\&prior)}_{\text{posterior}} = \underbrace{p(data|\theta, prior)}_{\text{likelihood}} \times \underbrace{\frac{p(\theta|prior)}{p(data|prior)}}_{\text{“normalized” prior probability}}$$

prior probability

evidence

Posterior: gives the best model/parameters (θ) given the data

Likelihood: is the probability of the data given the model/parameters (θ)

Prior probability: probability obtained from previous/prior, independent information

Evidence: probability of data regardless of any attempt at modeling

Marginal probabilities

The goal is to update our prior knowledge/model with new data/likelihood to obtain the posterior knowledge

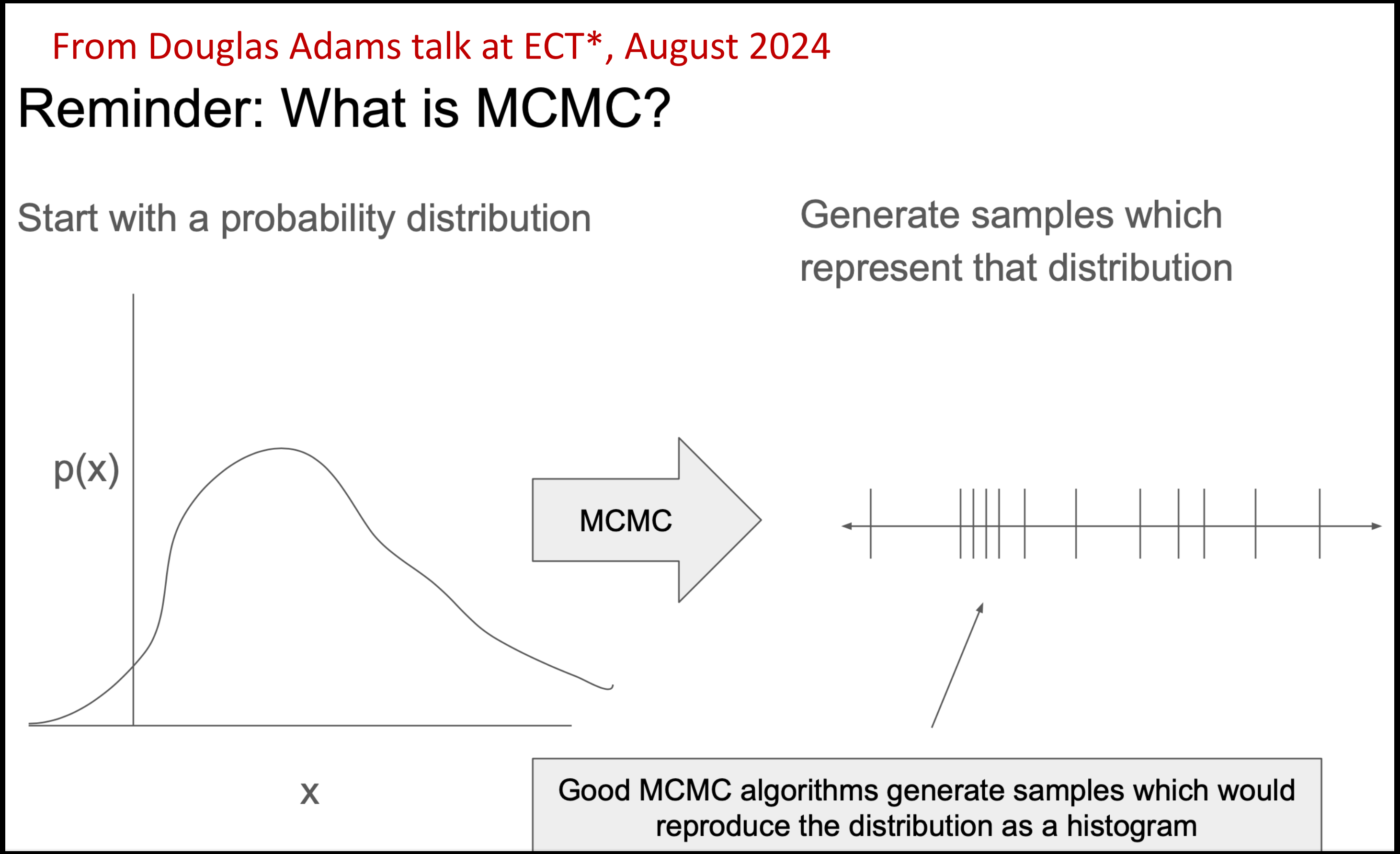
How this is practically implemented: Markov Chain MonteCarlo (MCMC)

- MCMC generates samples of a given distribution

- Our problems involve multi-parameter/multi-dimensional quantities.

- To obtain posterior distributions for each parameter (marginals) one integrates over all the other parameters.

- With many parameters to estimate, these multi-dimensional integrals become intractable.



More reasons:

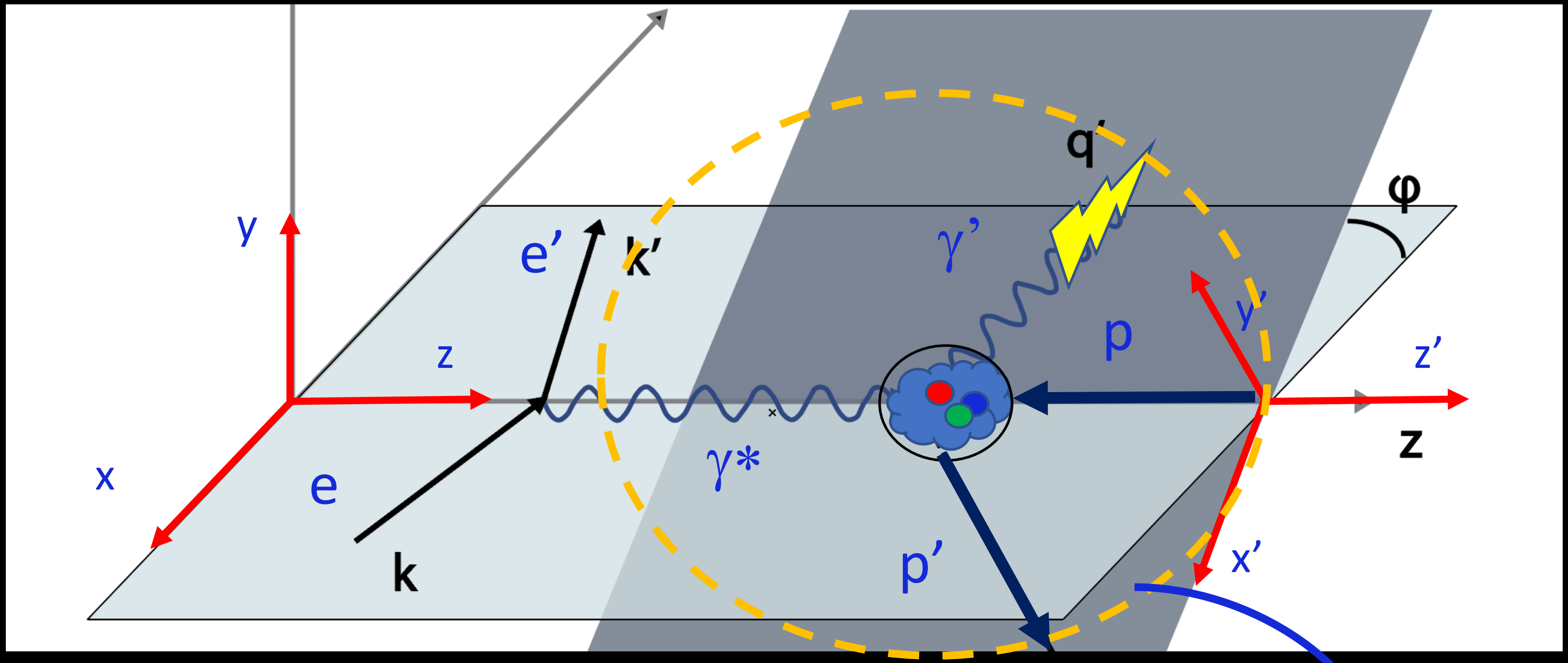
1) Versatility: to use the samples in further calculations substitute each row of generated samples into some other function of the samples which returns a new number

2) Visualization: to visualize the samples which we can look at to discern our own human belief in the parameters.

From Douglas Adams talk at ECT*, August 2024

Our problem

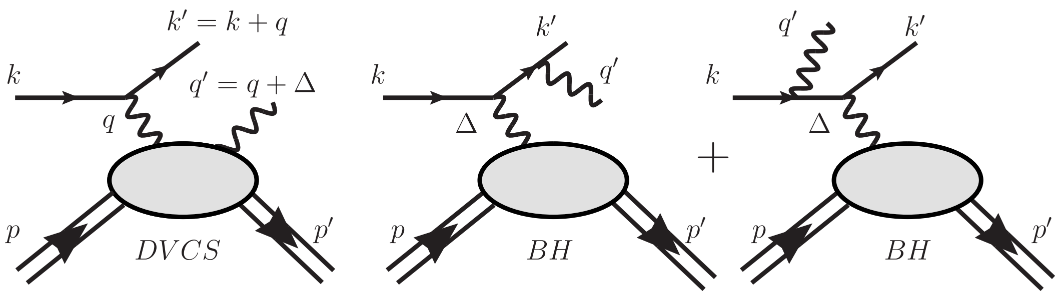
Deeply virtual exclusive processes (a unique class of probes)



Virtual Compton

$\gamma^* p \rightarrow \gamma' p'$

Experimental Data, Simulations and Inference



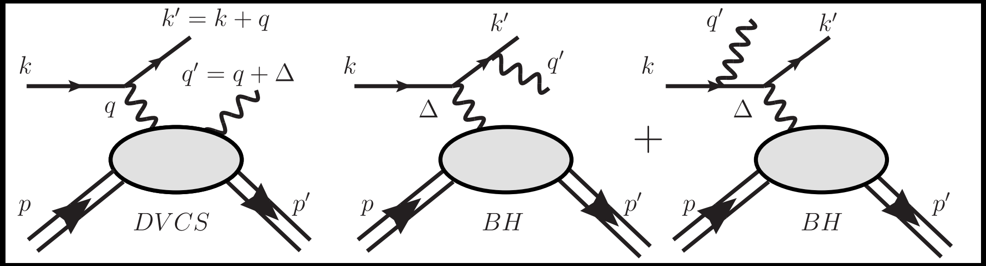
$$x = \frac{d^4\sigma_{exp}}{dQ^2 dt\, dE\, dx_{Bj}} \rightarrow measured\, cross\, section$$

$$\mu = \frac{d^4\sigma_{th}}{dQ^2 dt\, dE\, dx_{Bj}} \rightarrow model\, cross\, section$$

in one specific (Q^2 , x_{Bj} , E, t) bin

If the model was perfect, x and μ would coincide

1st Inverse Problem: extracting Compton form factors from cross section



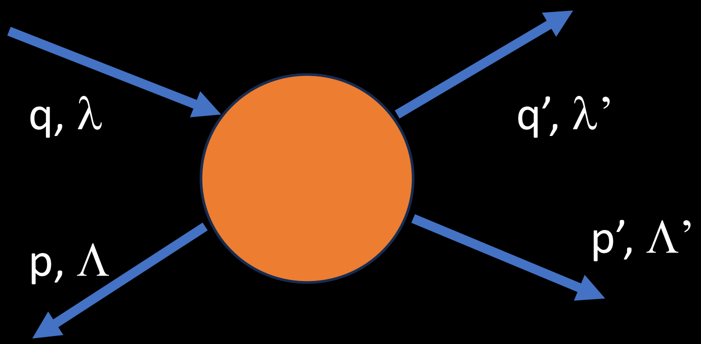
$$\frac{d^4\sigma}{dx_{Bj}dQ^2dtd\phi} = \sigma_{BH} + \sigma_{DVCS} + \sigma_{\mathcal{I}}$$

Relativistic two-body scattering in CoM frame

M. Jacob and G. Wick, Annals Phys. 7, 404 (1959)

helicity amplitudes

$$f^{\lambda\lambda'}_{\Lambda\Lambda'}(\theta,\phi)$$



Role of helicity amplitudes in polarized and unpolarized DIS, DY, SIDIS

A. Manohar, hep-ph/9204208
R. Jaffe, hep-ph/9602236

(1) DVCS

B. Kriesten, S.L. et al. , Phys. Rev. D 101, 054021 (2020)

The cross section observables are obtained by squaring the coherent sum of the helicity amplitudes

$$F^{11}_{++} = (1-\xi^2)\,|\,\mathcal{H}+\widetilde{\mathcal{H}}\,|^2 - \frac{2\xi^2}{1-\xi^2}\,\Re e\left[(\mathcal{H}+\widetilde{\mathcal{H}})(\mathcal{E}+\widetilde{\mathcal{E}})\right]$$

$$F^{11}_{--} = (1-\xi^2)\,|\,\mathcal{H}-\widetilde{\mathcal{H}}\,|^2 - \frac{2\xi^2}{1-\xi^2}\,\Re e\left[(\mathcal{H}-\widetilde{\mathcal{H}})(\mathcal{E}-\widetilde{\mathcal{E}})\right]$$

$$F^{11}_{+-} = \frac{t_0-t}{4M^2}\,|\,\mathcal{E}+\xi\widetilde{\mathcal{E}}\,|^2$$

$$F^{11}_{-+} = \frac{t_0-t}{4M^2}\,|\,\mathcal{E}-\xi\widetilde{\mathcal{E}}\,|^2$$

1. No cancellation of different polarization configurations as in DIS happens

2. **No direct connection between polarization observables and quark-proton polarization**

The UVA Formalism:
“good use of helicity amplitudes”

(alternative to BKM)

The DVCS cross section is expressed in a compact, meaningful form (that fits in half a page compared to previous formalism that takes several pages, and where the "harmonics" dependence has no underlying principle behind it)

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

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

$$+ (2h)\sqrt{2\epsilon(1-\epsilon)}\sin\phi F_{LU}^{\sin\phi}$$

$$+ (2\Lambda)\Big[\sqrt{\epsilon(\epsilon+1)}\sin\phi F_{UL}^{\sin\phi} + \epsilon\sin 2\phi F_{UL}^{\sin 2\phi}$$

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

$$+ (2\Lambda_T)\Big[\sin(\phi-\phi_S)\left(F_{UT,T}^{\sin(\phi-\phi_S)} + \epsilon F_{UT,L}^{\sin(\phi-\phi_S)}\right)$$

$$+ \quad \epsilon\sin(\phi+\phi_S)F_{UT}^{\sin(\phi+\phi_S)} + \epsilon\sin(3\phi-\phi_S)F_{UT}^{\sin(3\phi-\phi_S)}$$

$$+\sqrt{2\epsilon(1+\epsilon)}\left(\sin\phi_S F_{UT}^{\sin\phi_S} + \sin(2\phi-\phi_S)F_{UT}^{\sin(2\phi-\phi_S)}\right)\Big]$$

$$+(2h)(2\Lambda_T)\Big[\sqrt{1-\epsilon^2}\cos(\phi-\phi_S)F_{LT}^{\cos(\phi-\phi_S)} + \sqrt{2\epsilon(1-\epsilon)}\cos\phi_S F_{LT}^{\cos\phi_S}$$

$$+\sqrt{2\epsilon(1-\epsilon)}\cos(2\phi-\phi_S)F_{LT}^{\cos(2\phi-\phi_S)}\Big]\Big\}$$

1. Eq.(36) (1)
2. cf. SIDIS
3. cf. pseudo-scalar meson production (Ahmad, Goldstein, SL, Gonzalez Hernandez, Goldstein, SL)

Important for QCD interpretation

- Twist 2: $F_{UU,T}$, F_{LL}, $F_{UT,T}^{\sin(\phi-\phi_S)}$, $F_{LT}^{\cos(\phi-\phi_S)}$
- Twist 3: $F_{UU}^{\cos\phi}$, $F_{UL}^{\sin\phi}$, $F_{LU}^{\sin\phi}$, $F_{LL}^{\cos\phi}$, $F_{UT}^{\sin\phi_S}$, $F_{UT}^{\sin(2\phi-\phi_S)}$, $F_{LT}^{\cos\phi_S}$, $F_{LT}^{\cos(2\phi-\phi_S)}$
- Twist 4: $F_{UU,L}$, $F_{UT,L}^{\sin(\phi-\phi_s)}$
- Transverse gluons: $F_{UU}^{\cos 2\phi}$, $F_{UL}^{\sin 2\phi}$, $F_{UT}^{\sin(\phi+\phi_s)}$, $F_{UT}^{\sin(3\phi-\phi_s)}$.

At leading twist: 4 different Compton Form Factors because of the 4 possible different helicity configurations

$$\sigma^{unp}_{DVCS} = \frac{\Gamma}{Q^2(1-\epsilon)} \Big\{ (1-\xi^2)\Big[(\Re e\mathcal{H})^2 + (\Im m\mathcal{H})^2 + (\Re e\widetilde{\mathcal{H}})^2 + (\Im m\widetilde{\mathcal{H}})^2\Big] + \frac{t_o - t}{2M^2}\Big[(\Re e\mathcal{E})^2 + (\Im m\mathcal{E})^2 + \xi^2(\Re e\widetilde{\mathcal{E}})^2 + \xi^2(\Im m\widetilde{\mathcal{E}})^2\Big] - 2\xi^2\left(\Re e\mathcal{H}\,\Re e\mathcal{E} + \Im m\mathcal{H}\Im m\mathcal{E} + \Re e\widetilde{\mathcal{H}}\,\Re e\widetilde{\mathcal{E}} + \Im m\widetilde{\mathcal{H}}\Im m\widetilde{\mathcal{E}}\right)\Big\}$$

spin non-flip

spin flip

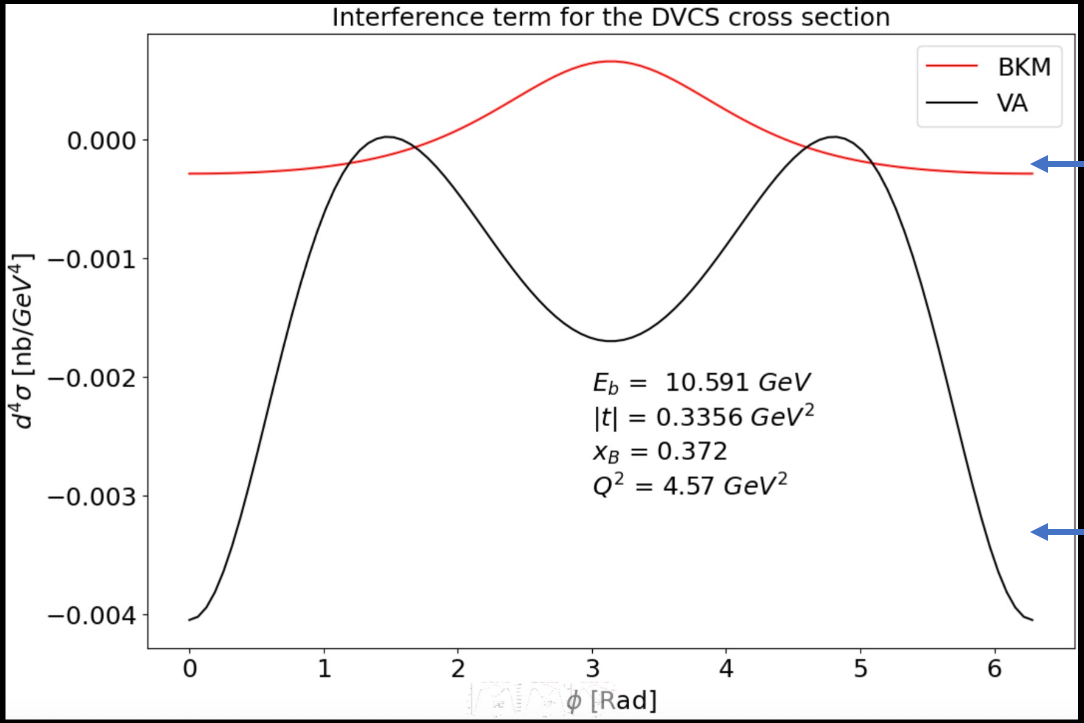
mixed

$F_1^2 + \tau F_2^2$ $\quad (F_1+F_2)^2$ $\quad (F_1+F_2)G_A$

$$\sigma^{unp}_{\mathcal{I}} = \frac{\Gamma}{Q^2(1-\epsilon)}\Big\{A^{\mathcal{I}}_{UU}\,(F_1\Re e\mathcal{H} + \tau F_2\Re e\mathcal{E}) + B^{\mathcal{I}}_{UU}(F_1+F_2)(\Re e\mathcal{H} + \Re e\mathcal{E}) + C^{\mathcal{I}}_{UU}(F_1+F_2)\Re e\widetilde{H}\Big\}$$

Azimuthal angle ϕ dependent coefficients

- 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

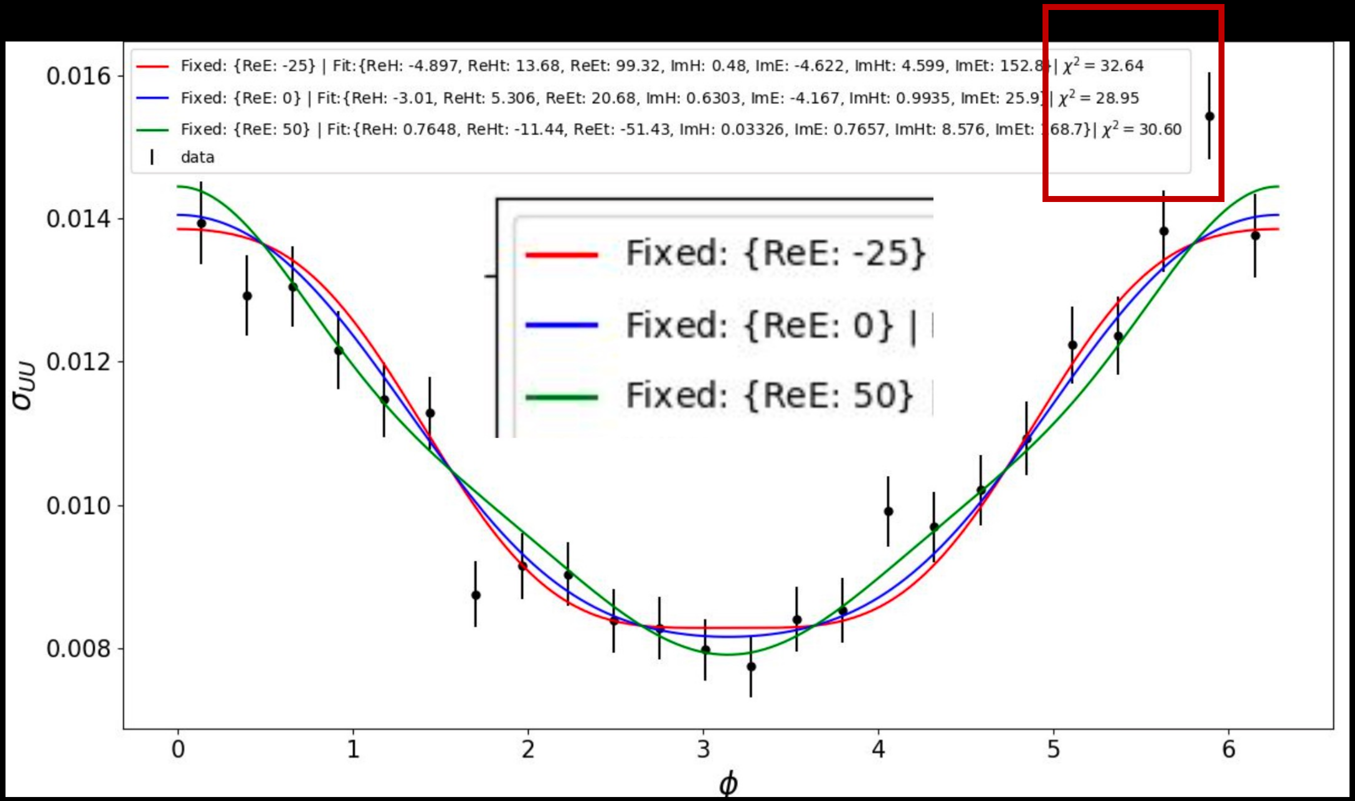


BKM formalism

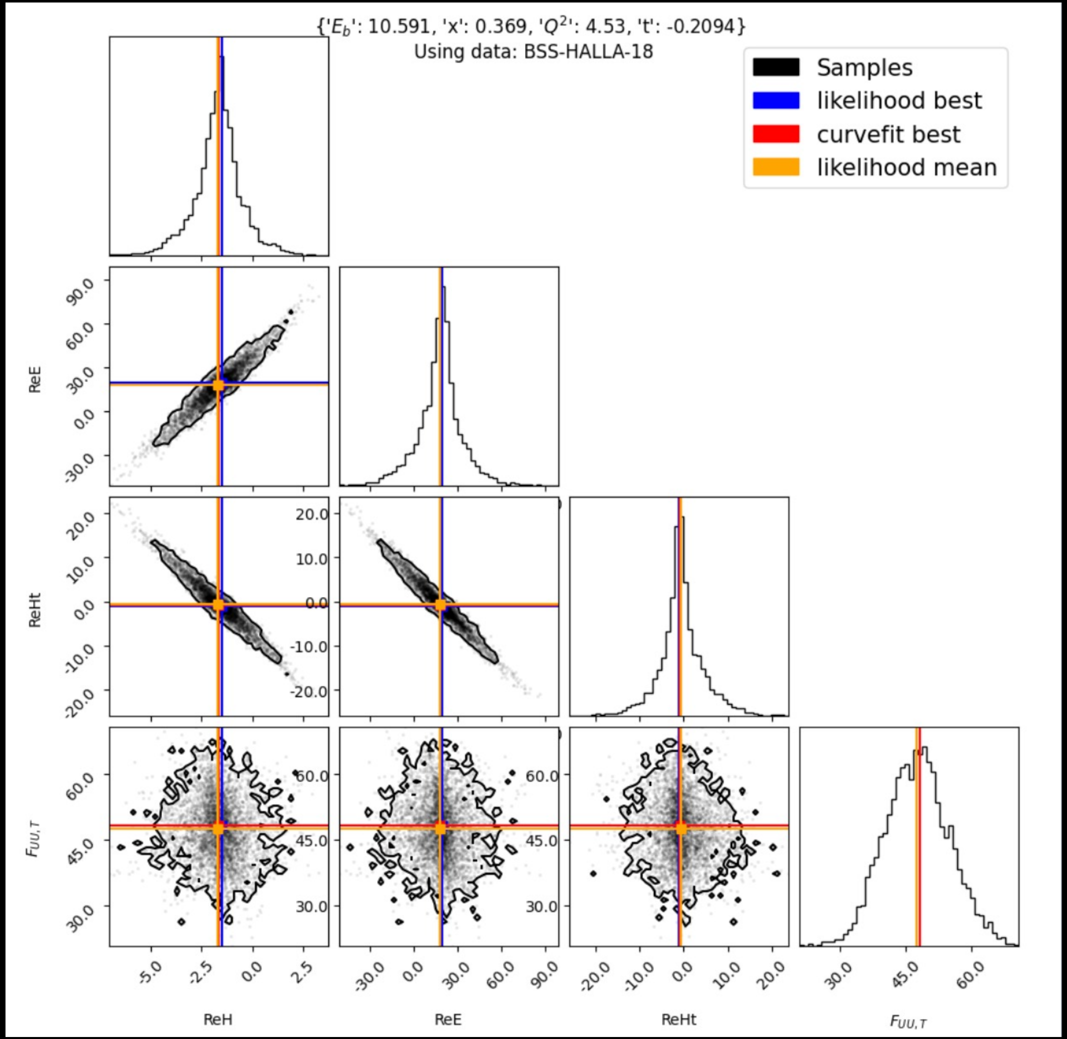
with helicity amplitudes

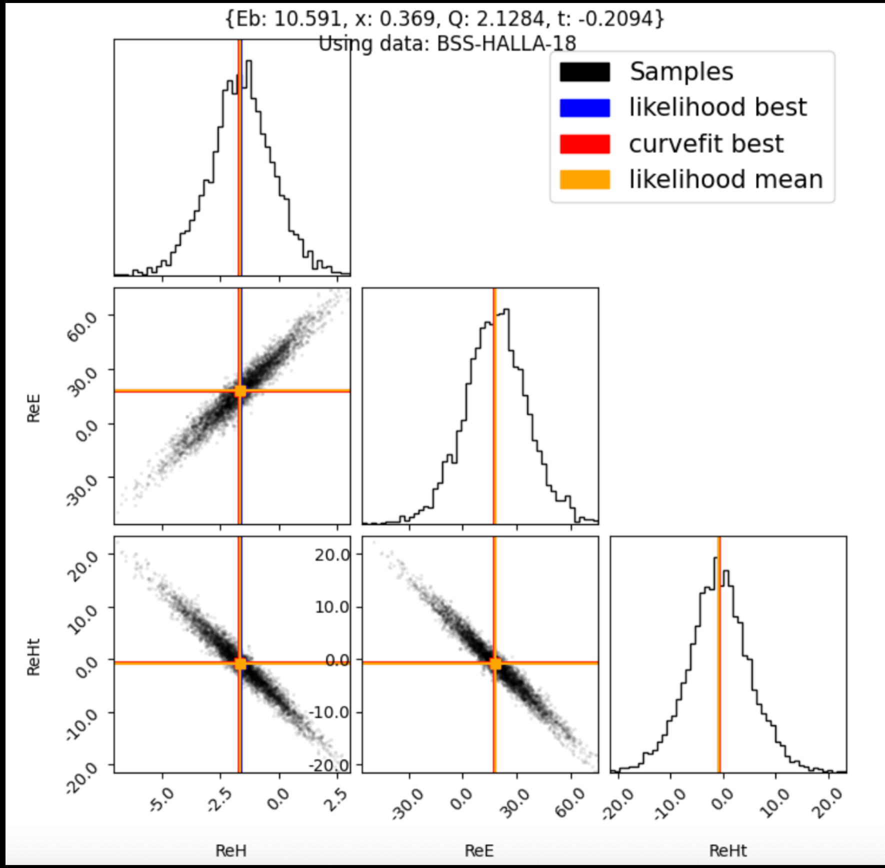
Adil Khawaja (preliminary)

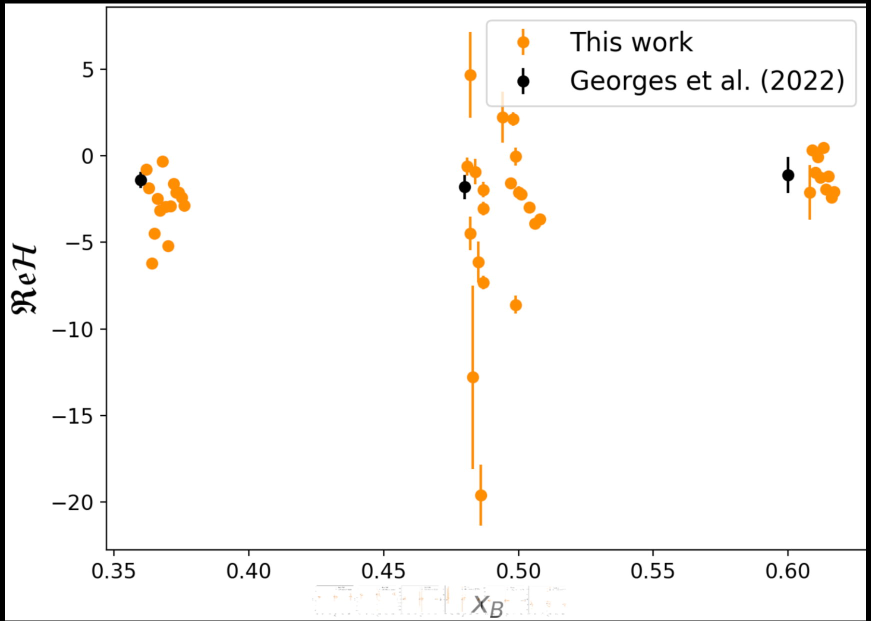
We performed a likelihood analysis of unpolarized DVCS.
Starting from a standard curve fit, we encountered the following situation

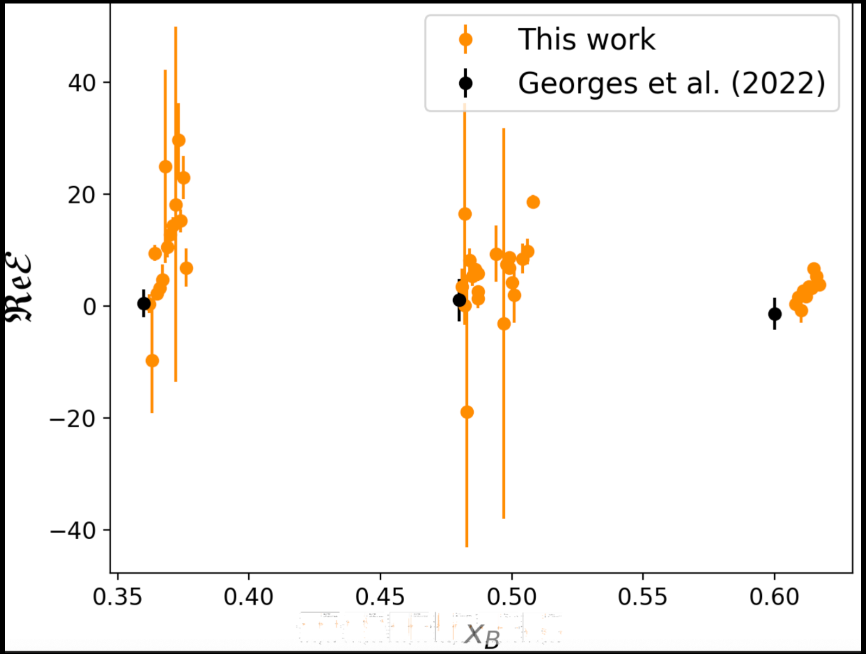


Likelihood analysis with MCMC sampling results with two different methods









Are we in the sector of:

<u>Most analyses rely on “catchy punchlines and compact, approximate representations, even when those are unjustified and unnecessary”</u>

<u>(Hogg, Bovy and Lang, astro-ph: 1008.4686</u>

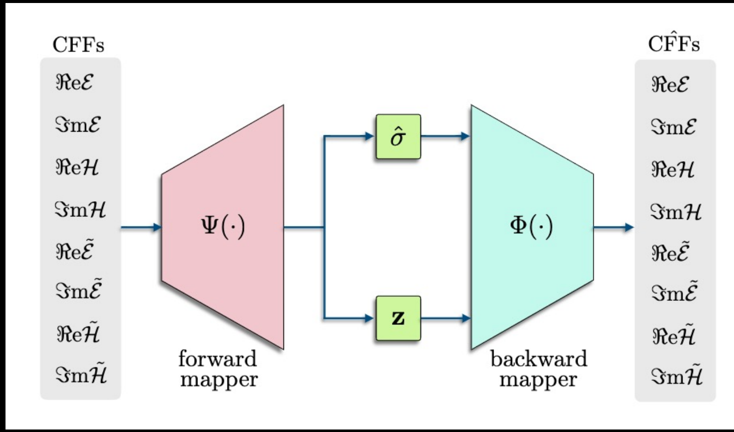
Or can we sit down and solve this discrepancy?

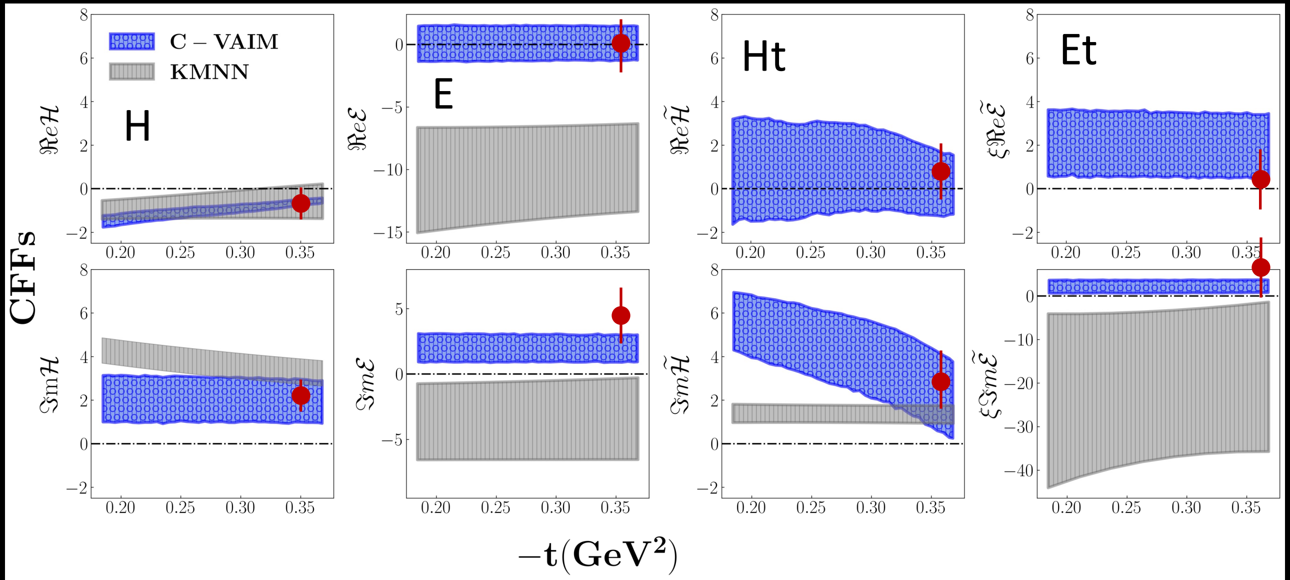
Graphs by Zaki Panjsheeri

Another method to extract the CFFs is to use a
Conditional –Variational Autoencoder Inverse Mapper (C-VAIM)

(work with Yaohang Li, ODU and Gia-Wei Chern, UVA)

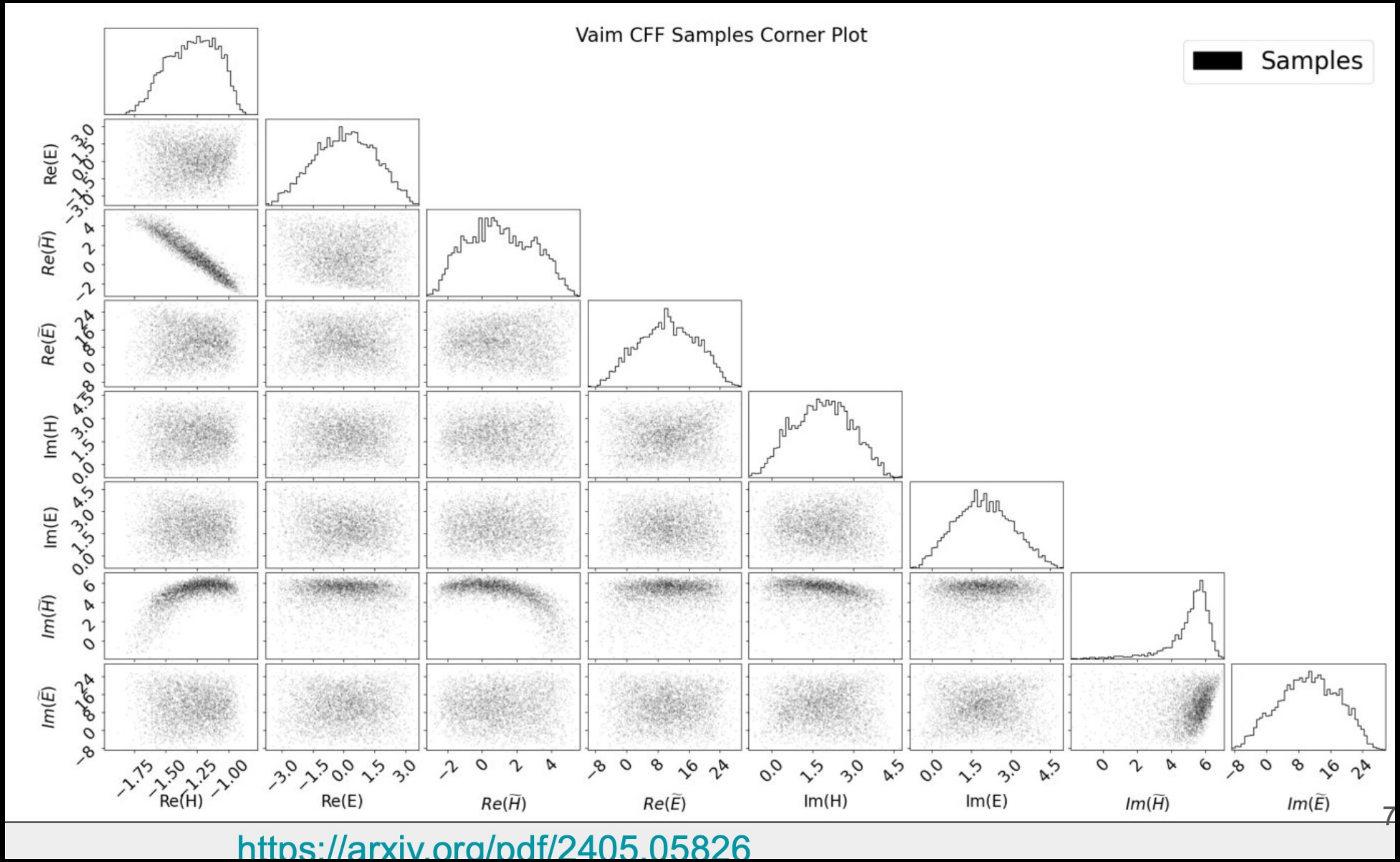
- **A variational autoencoder inverse mapper solution to Compton form factor extraction from deeply virtual exclusive reactions** arXiv: 2405.05826





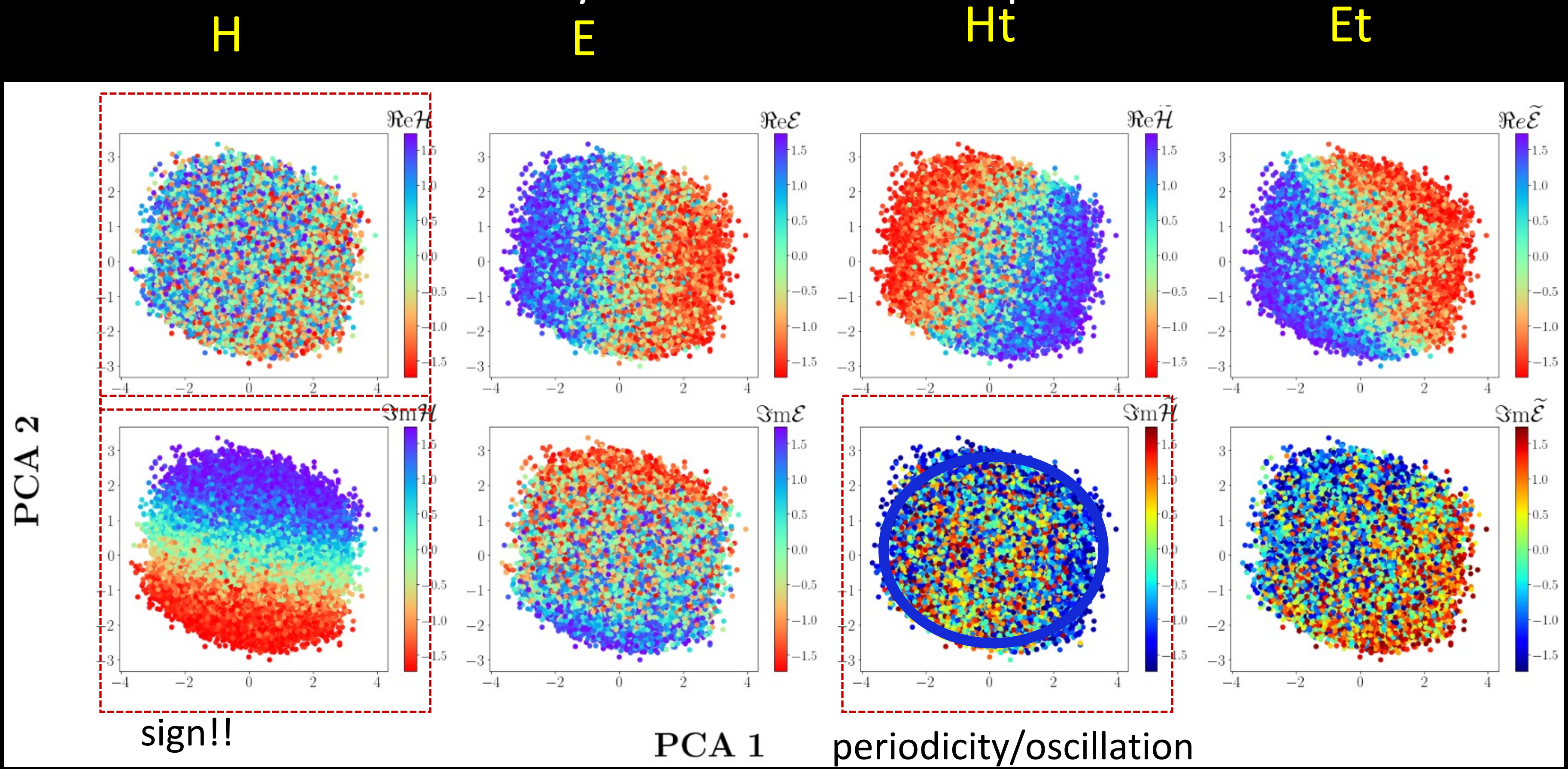
- KMNN, Kumericki et al. https://arxiv.org/abs/2007.00029

Study correlations: 2. CFFs Analysis of Latent Space



Generate viable CFFs which satisfy the constraint from the cross section

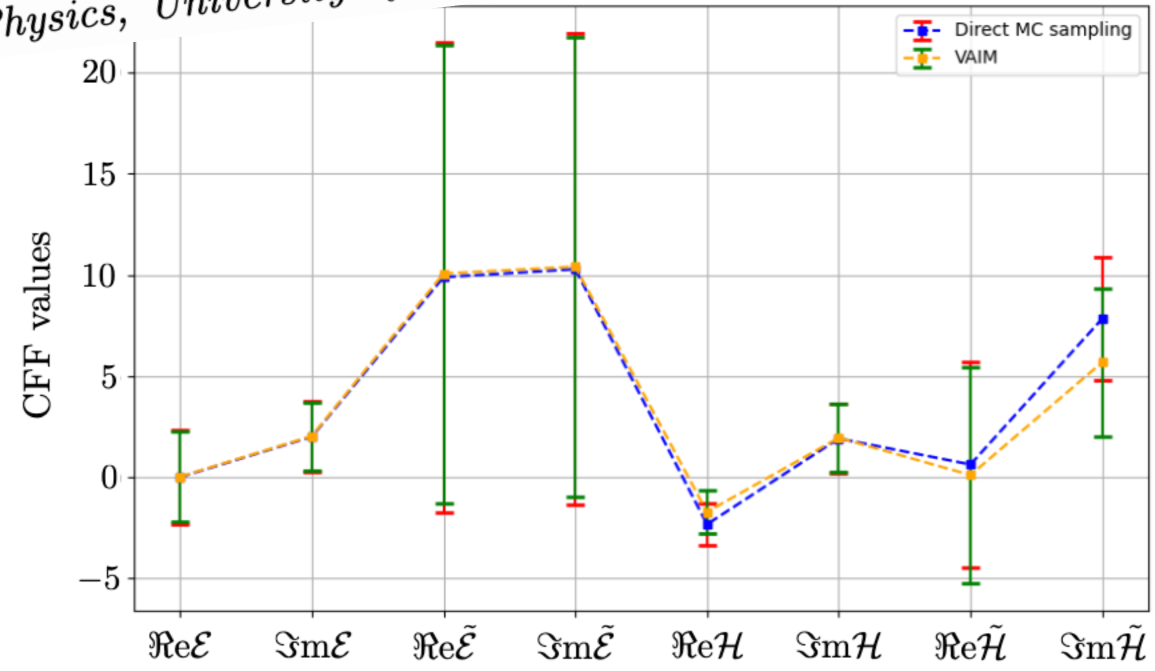
Study correlations: 1. CFFs Analysis of Latent Space



Comparison of MC sampling 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.*



Conclusions

- We presented a new paradigm for the analysis of deeply virtual exclusive experiments using a model for likelihood-based inference in high dimensional data settings.

- We obtained joint covariant results for Compton form factors using a twist-two cross section model
for the unpolarized process for which the largest amount of data with all the kinematic dependences are available from corresponding datasets from Jefferson Lab.

- Based on the observation that the unpolarized twist-two cross section likelihood fully constrains only three of the CFFs

- The twist-two difference-model likelihoods were explored using Markov chain Monte Carlo (MCMC) methods.
Error bars, covariances and associated joint $n-\sigma$ confidence contours were evaluated.

- Using a more ``well behaved" set of data does not substantially help.

- However: higher twists are important. This casts doubts on the validity of factorization and motivates both further investigati EIC kinematics and large Q^2 span (analysis of H1 and ZEUS data as precursor of EIC), as well as including constraints from othe sets of data (PDFs), longitudinally polarized, transversely polarized.

- Our analysis lays out the statistics foundation of our new framework which leverages state of the art ML algorithms to carry o efficiently Bayesian inference in high-dimensional settings.

Outlook: The Next Questions…

1. Interpretability of AI generated results

2. Where is the information on angular momentum and spatial dependence?

- The problem we are dealing with in physics is to extract information from data: **interpretability**

- The goal of ML is to obtain statistical model that has **predictivity** from the data

- Both sides define an **inverse problem**: more cross talking is needed between CS experts and physicists to explore all the synergies, the common aspects, focusing on **why** any given method works

- An immense potential: through ML we will be able to see the emergence of **new physics relations/laws: SYMBOLIC REGRESSION**