Towards benchmarking of the partonic structure of the proton by the EXCLAIM collaboration

Marija Čuić

University of Virginia

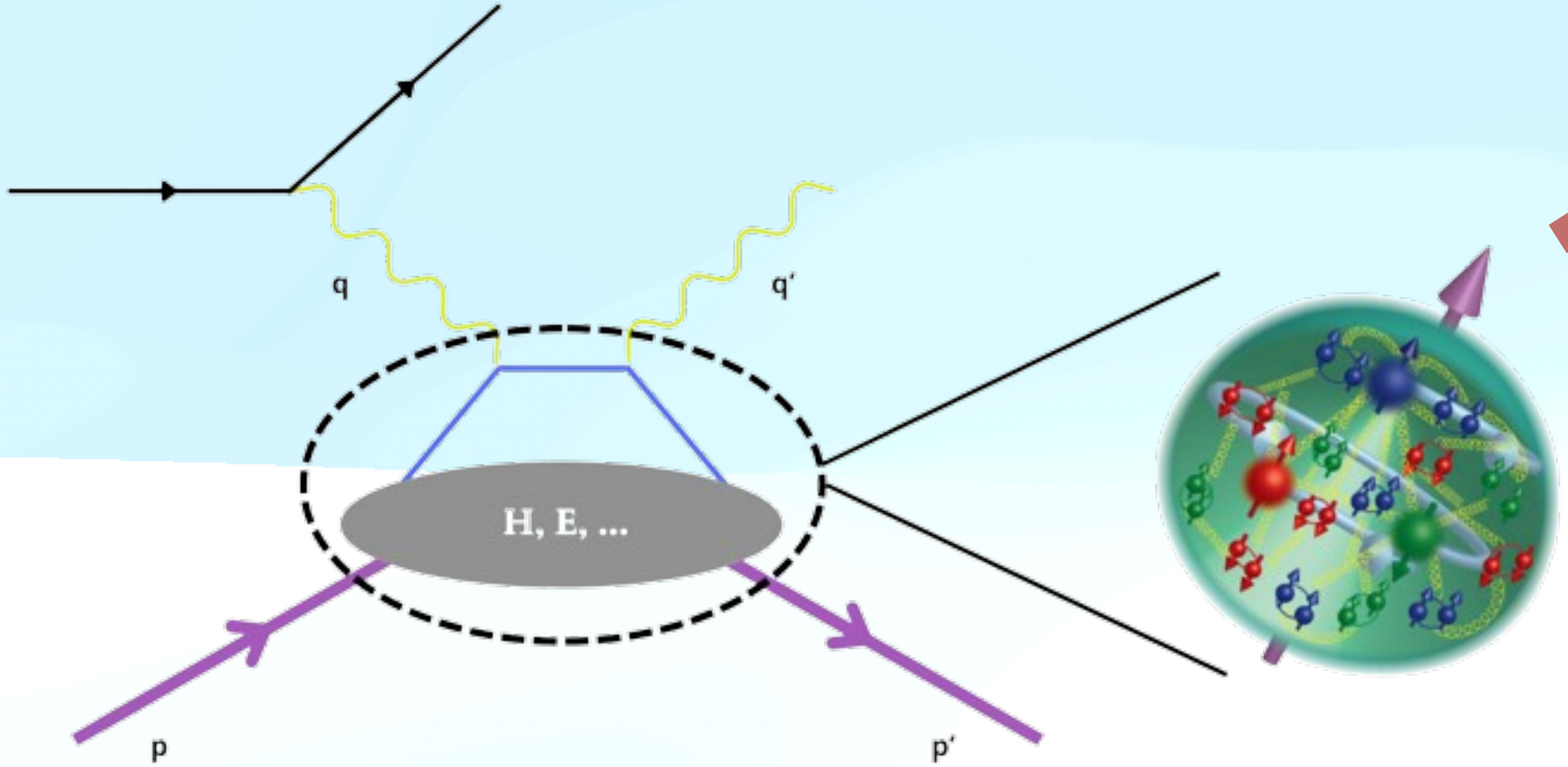
July 22nd 2024

EXCLAIM (Exclusives with Artificial Intelligence and Machine Learning)

- **ML/AI**: Yaohang Li, Gia-Wei Chern, Douglas Adams, Tareq Alghamdi, Md. Fayaz Bin-Hossen, Anusha Reddy Singireddy, Siwen Liao
- **Experiment**: Marie Boër, Debaditya Biswas, Kemal Tezgin
- **Lattice QCD**: Michael Engelhardt, Huey-Wen Lin, Emanuel Ortiz Pacheco, Liam Hockley
- **Phenomenology/Theory**: Simonetta Liuti, Gary Goldstein, Matt Sievert, Dennis Sivers, MČ, Saraswati Pandey, Joshua Bautista, Adil Khawaja, Zaki Panjsheeri, Carter Gustin, Andrew Dotson, Kiara Ruffin

Nucleon structure

DVCS: $\ell + p \to \ell + \gamma + p$

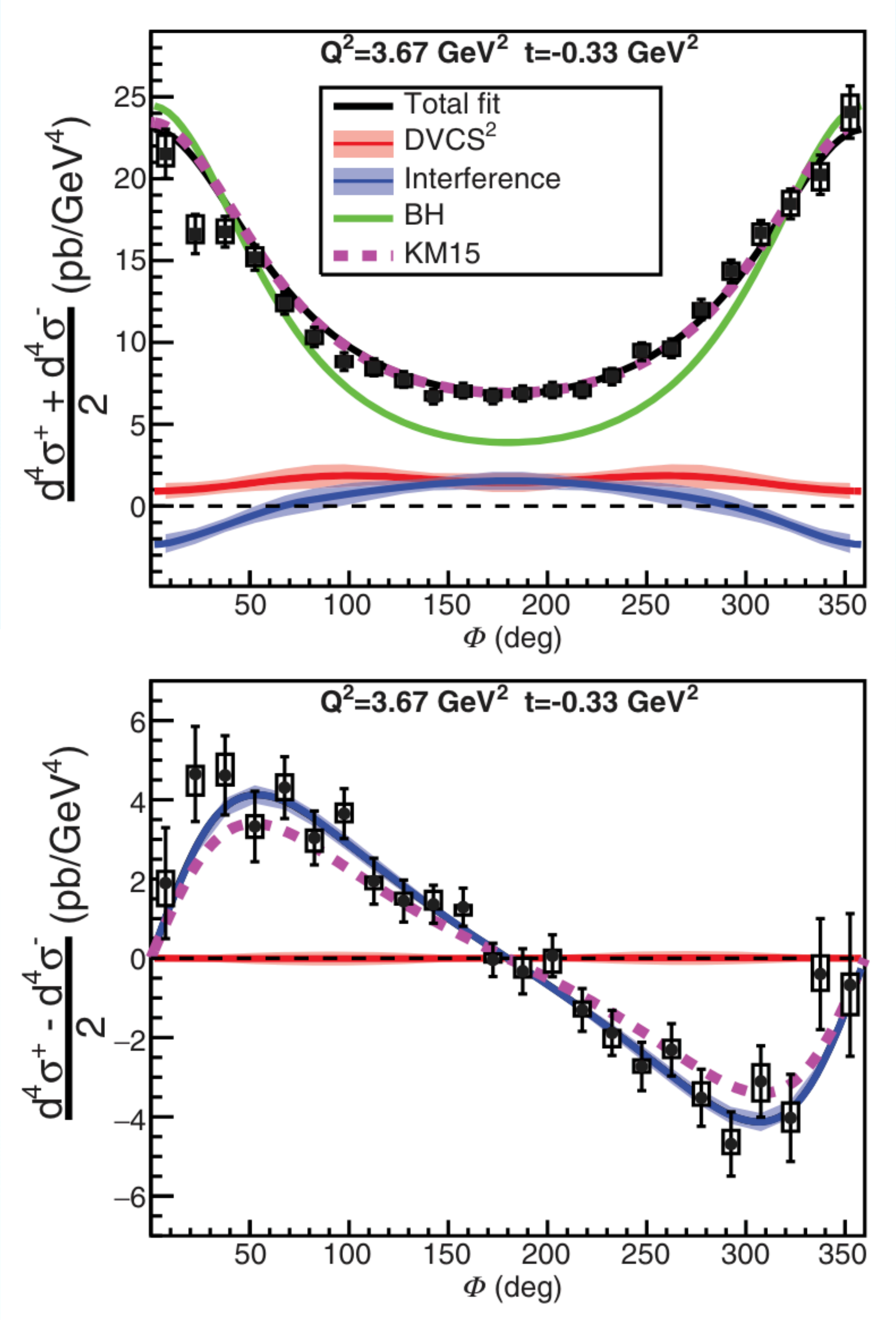


$$\frac{d^5\sigma_{DVCS}}{dx_B\,dy\,dt\,d\phi\,d\varphi} \propto 4\,(1-x_B)\left(|\mathcal{H}|^2 + |\widetilde{\mathcal{H}}|^2\right) + \dots$$

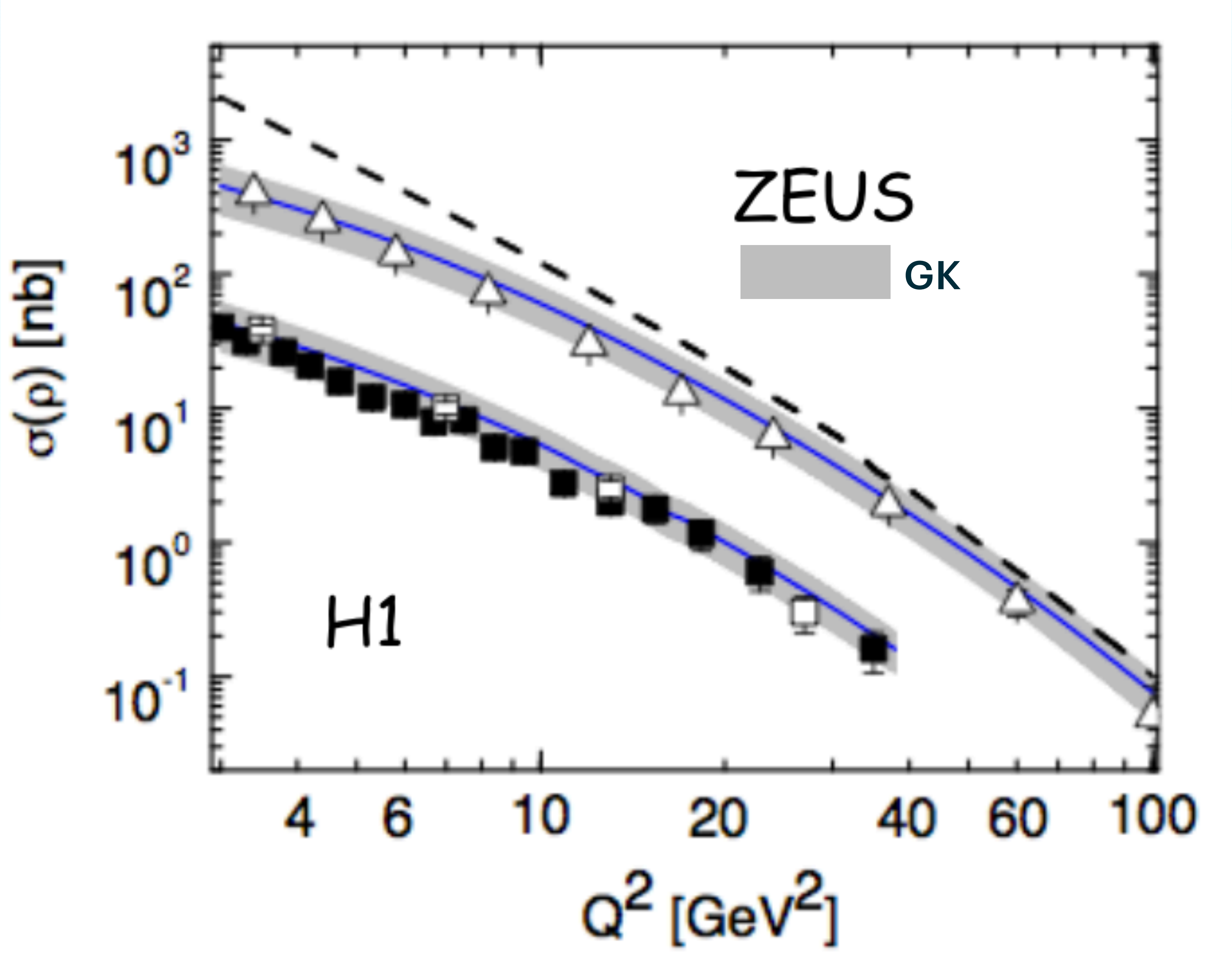
$$\mathcal{H}^A\left(\xi,\Delta^2,Q^2\right) = \int_{-1}^{1}\frac{dx}{2\xi}\,\underbrace{{}^{A}T\left(x,\xi \mid \alpha_s\left(\mu_R\right),\left\{\frac{Q^2}{\mu^2}\right\}\right)}_{\text{hard scale}}\underbrace{H^A\left(x,\xi,\Delta^2,\mu_F^2\right)}_{\text{soft scale}}$$

Inverse problem: observables → GPDs inside convolution
→ ill-posed problem!

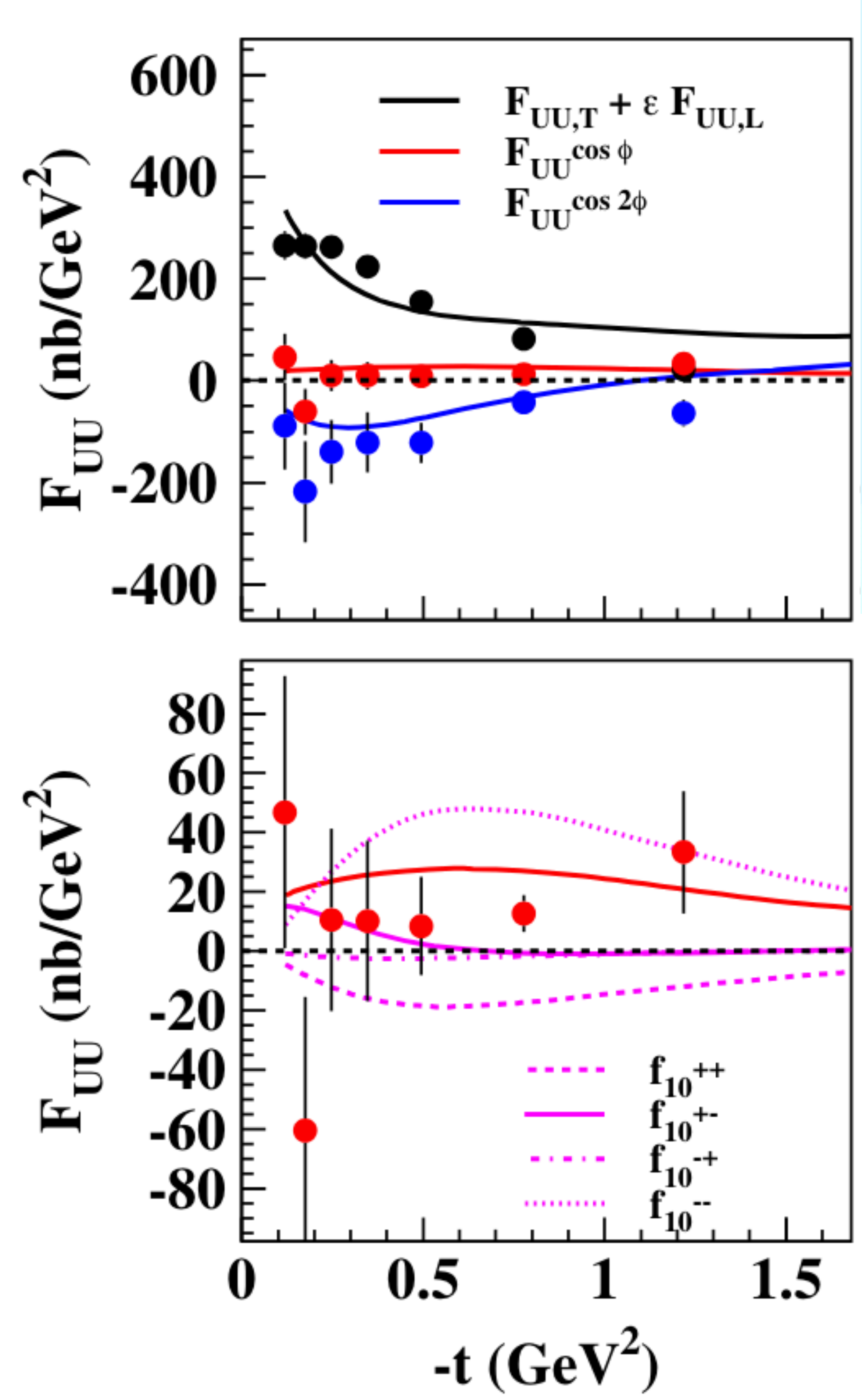
GPD models



Kumerički-Muller model

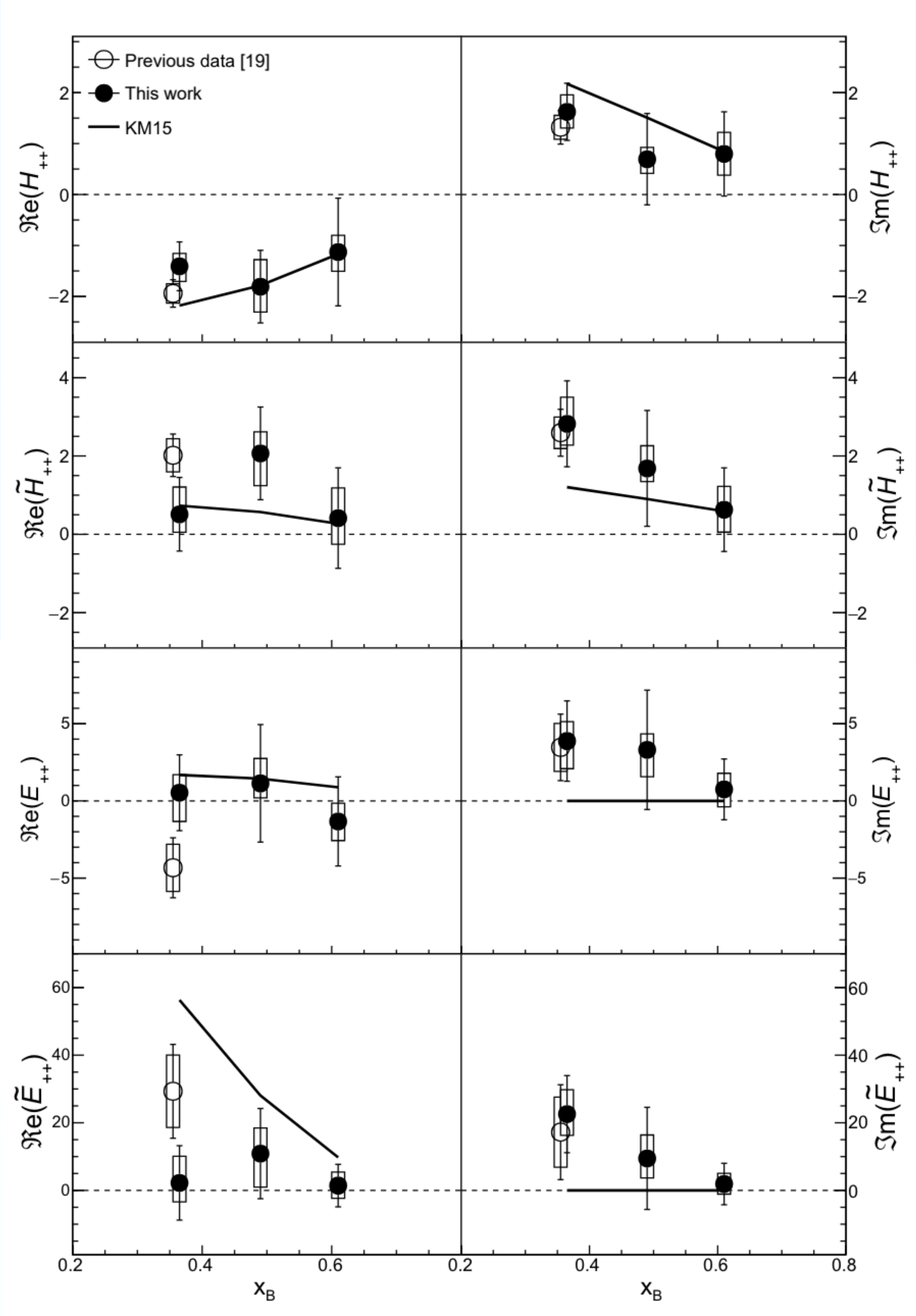


Goloskokov-Kroll model

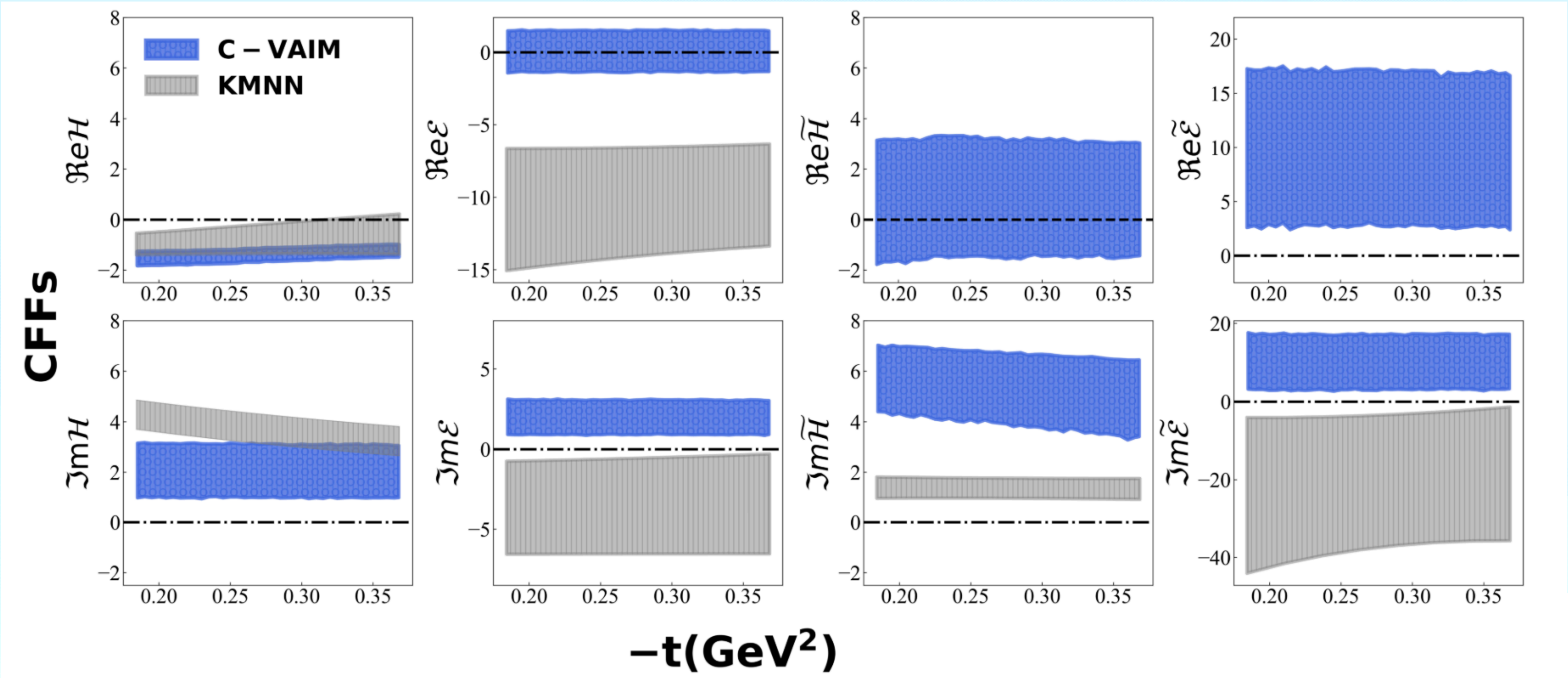


Goldstein-Gonzales-Liuti

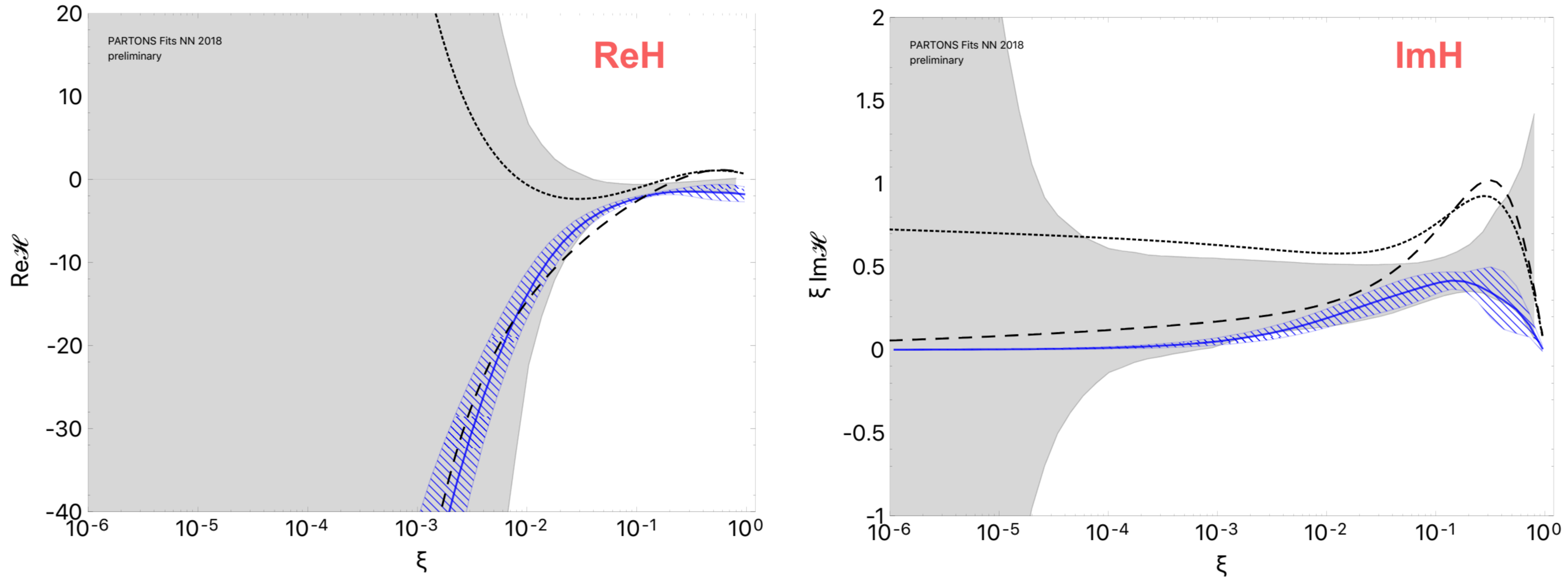
CFF extraction



Hessian based approach, Hall A



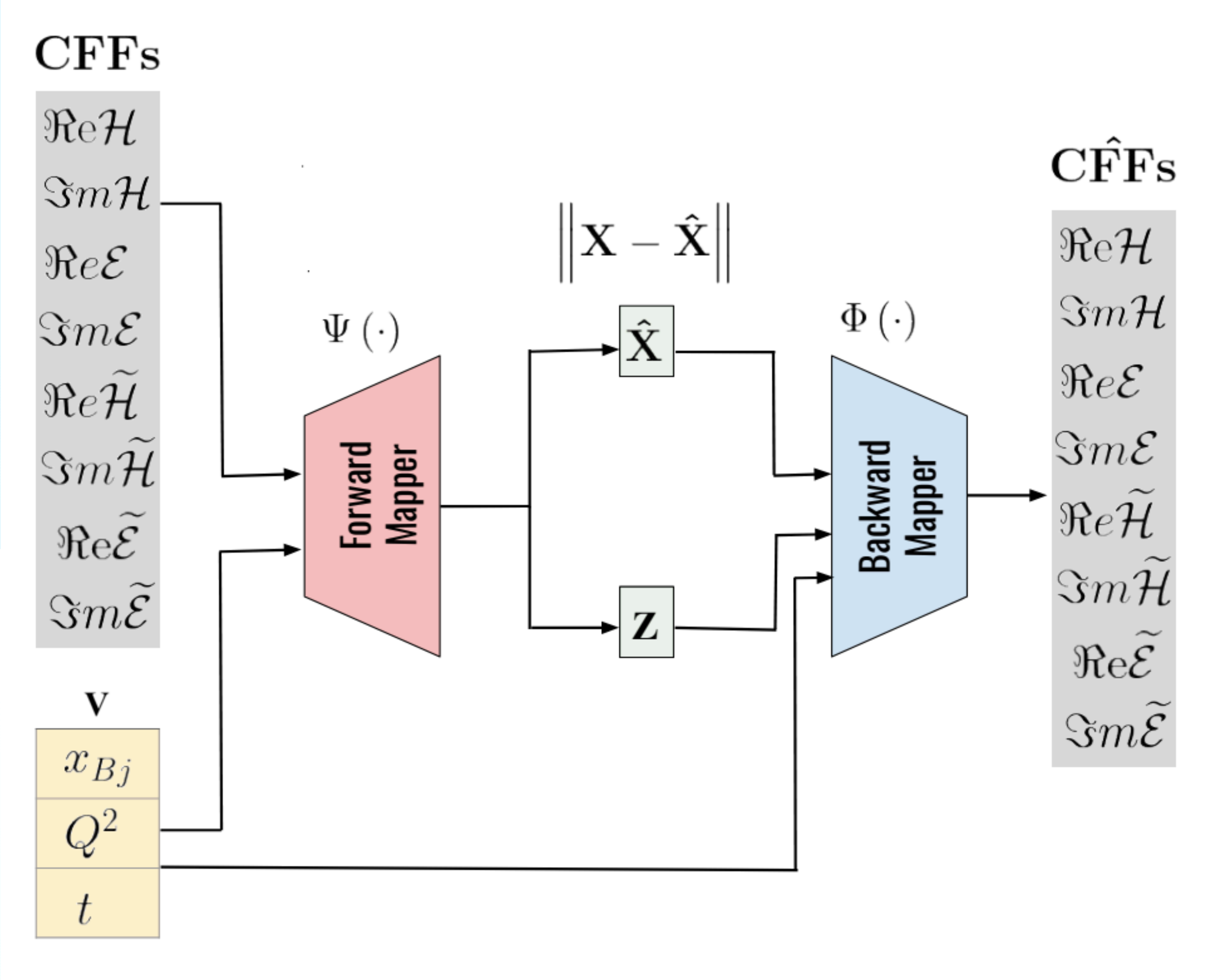
ML calculations, Kumerički vs VAIM



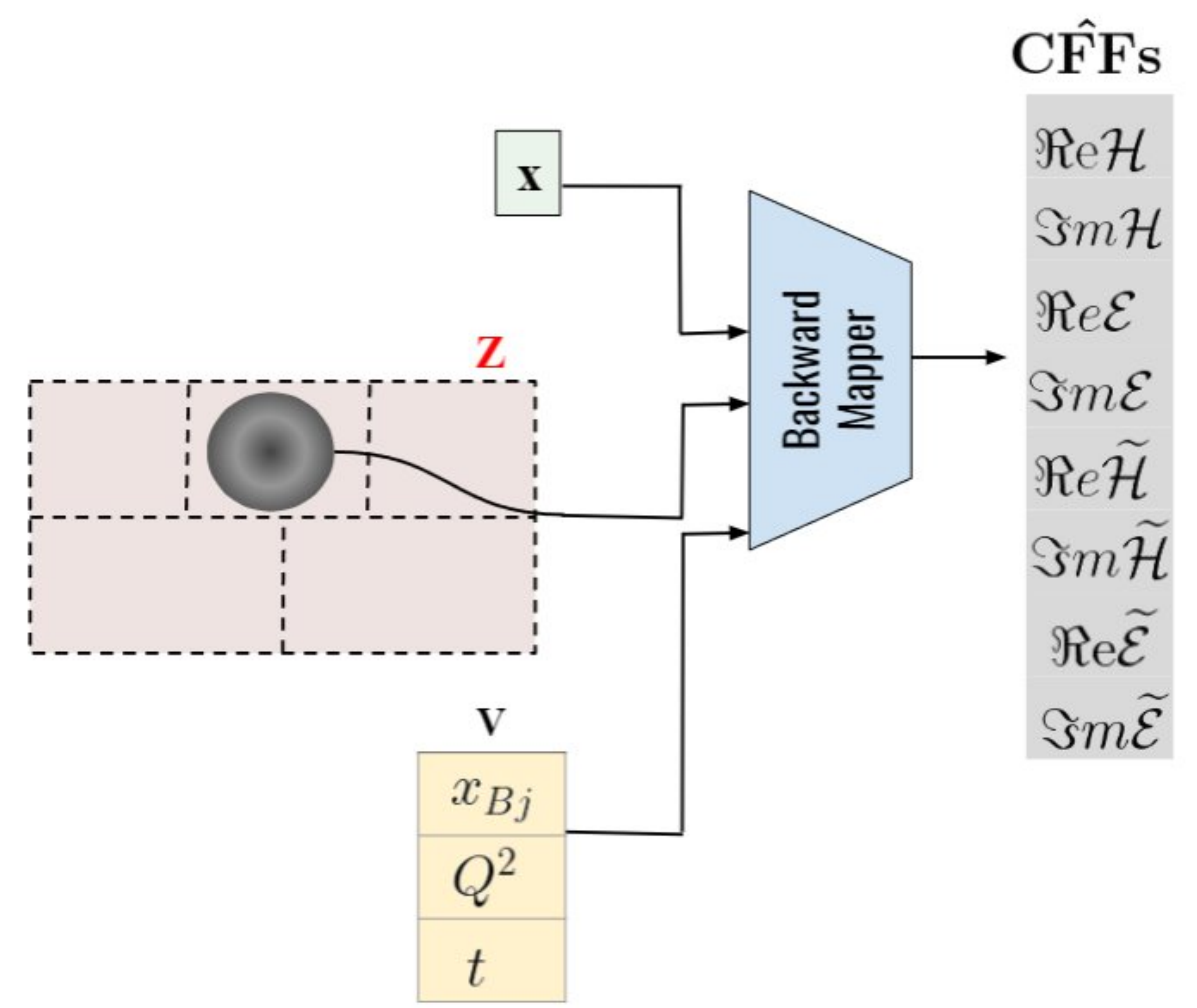
Partons 2018

Can we benchmark?

VAIM (Variational Autoencoder Inverse Mapper)

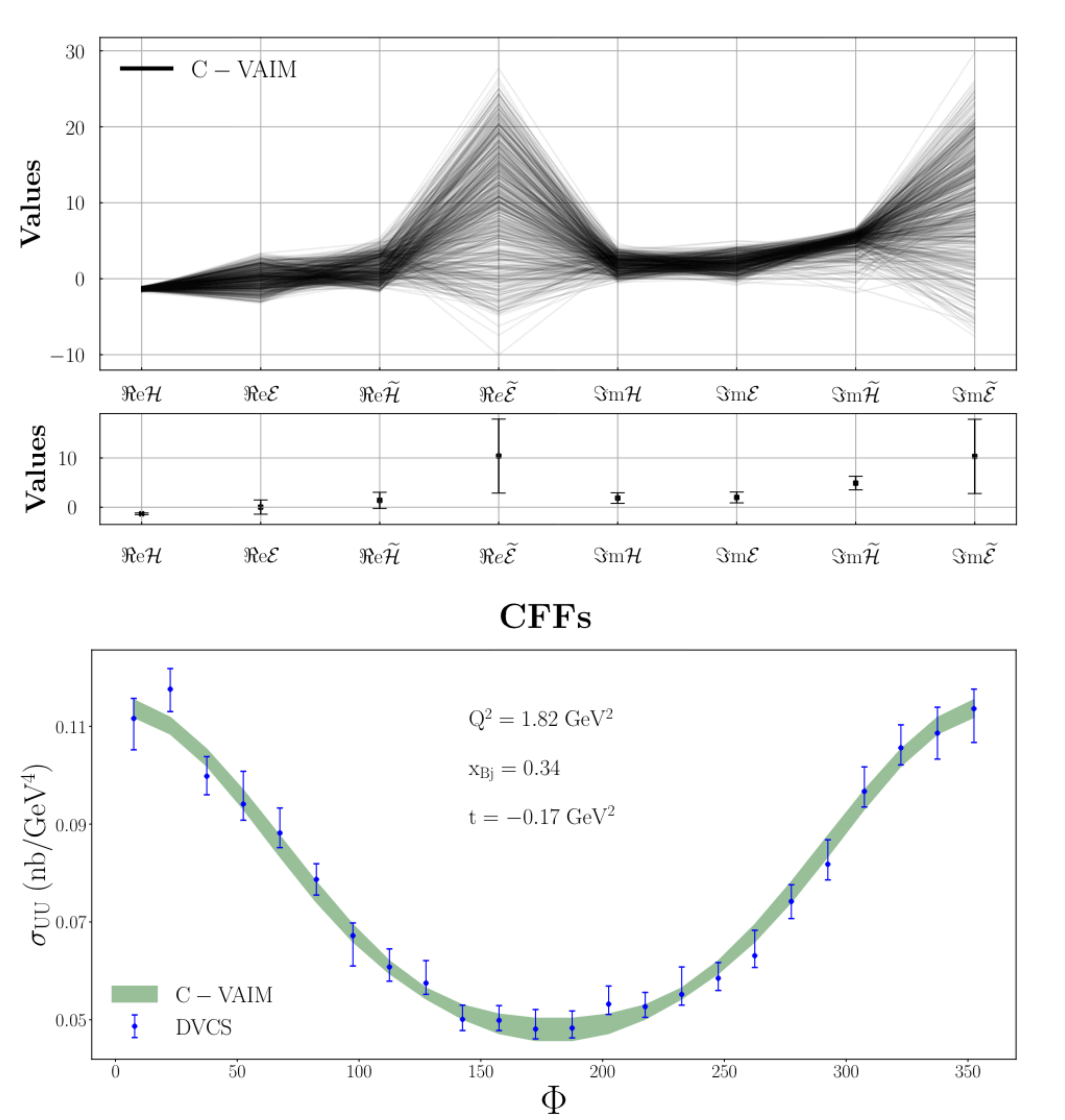


C-VAIM architecture to extract CFFs

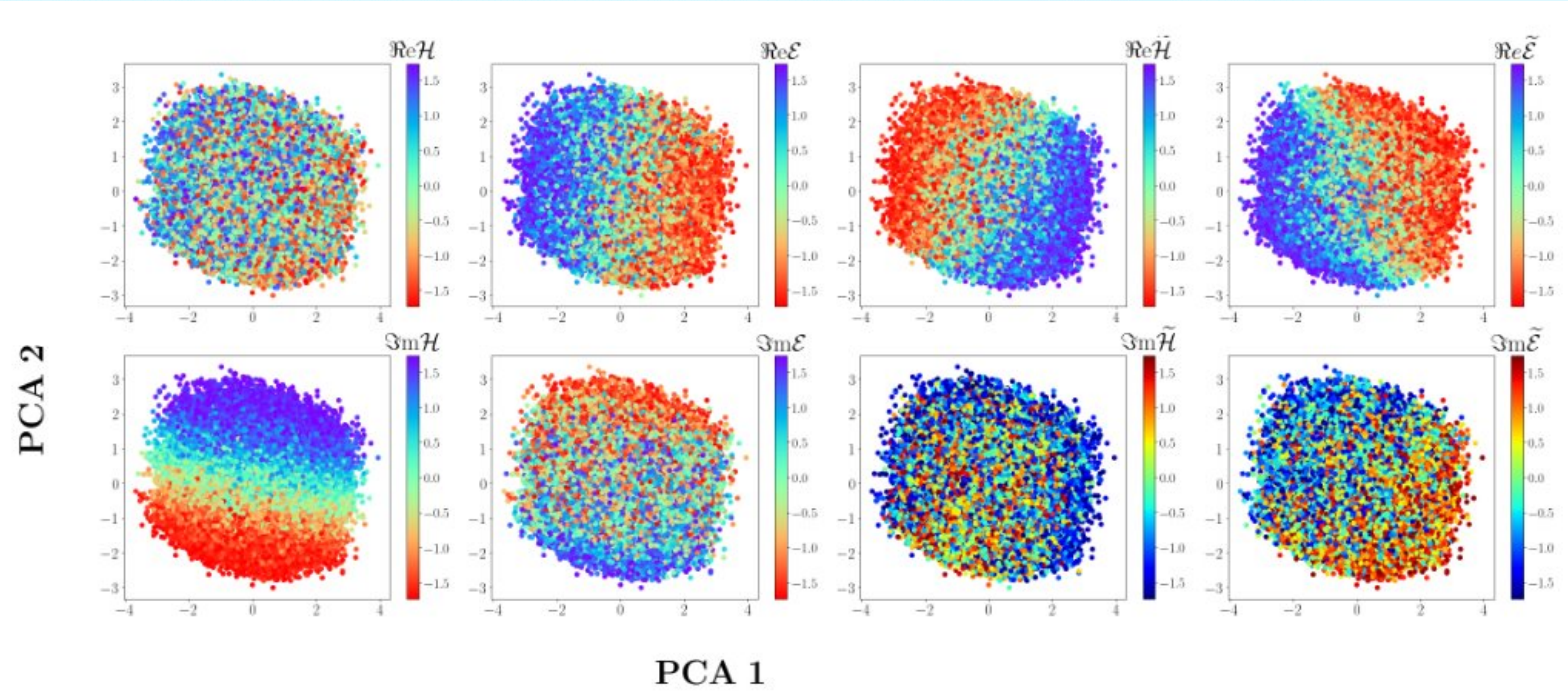


Decoder after VAIM is trained

cVAIM - CFF results



Latent space



We are not losing information on $\mathfrak{Re}\mathscr{H}$, we are losing sign information for $\mathfrak{Im}\mathscr{H}$, we cannot extract $\mathfrak{Im}\tilde{\mathscr{E}}$, etc…

[arXiv:2405.05826]

Symbolic regression (PySR)

Data table

t	Q^2	ϕ
-0.1	2	0
-0.2	4	20
⋮	⋮	⋮

Attempted expression

$y + ax^2 + bx + c$

Goodness

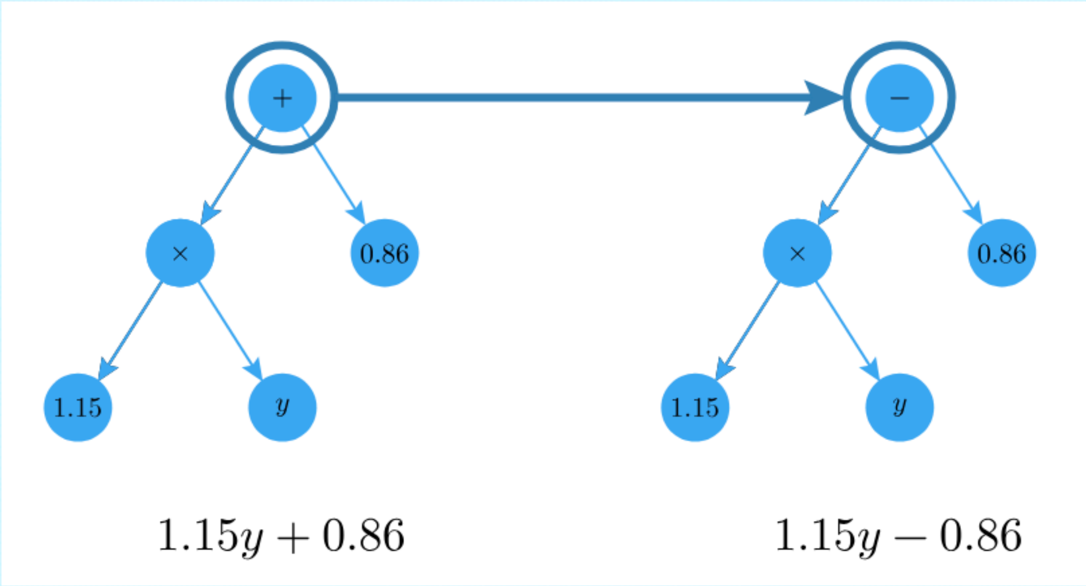
Fitness metric

e.g. squared error

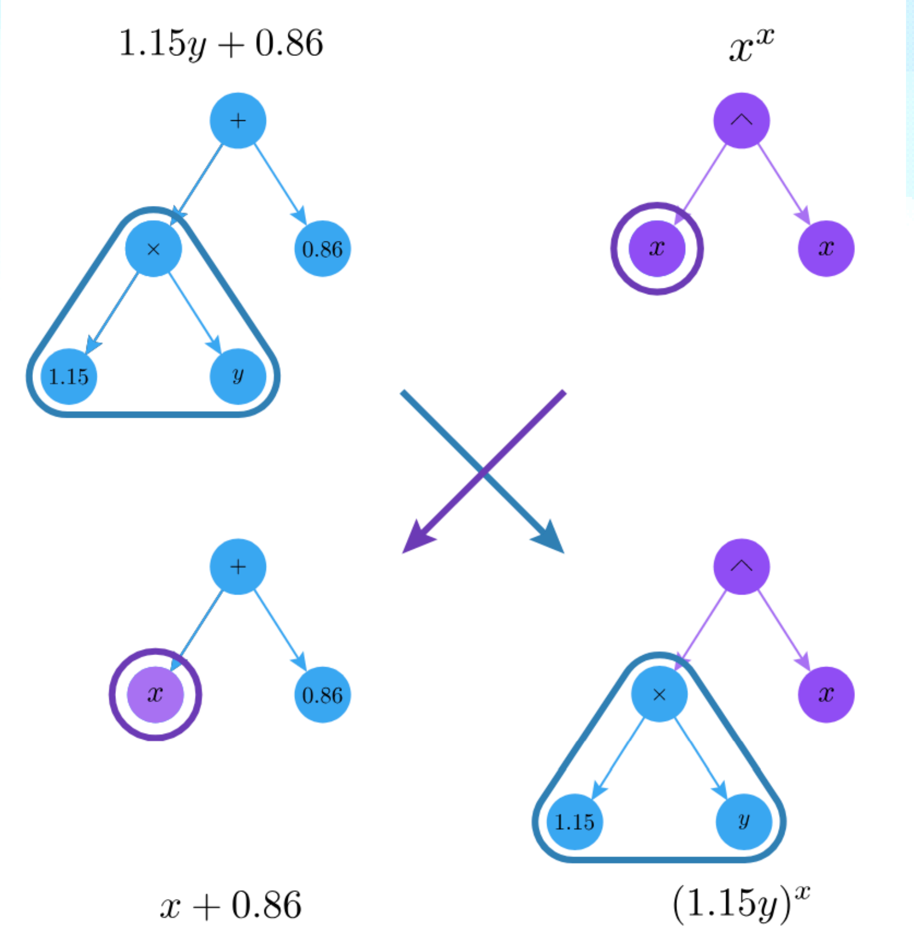
Form metric

e.g. #terms

Symbolic modification



Mutation



Crossover

Good?

No

Yes

Done

Symbolic regression on lattice and phenomenological models

- Numerically inexpensive and highly customizable
- Are the results stable and physical?
- Does it compare to neural networks?

```
Expressions evaluated per second: 7.760e+04
Head worker occupation: 12.8%
Progress: 172 / 200 total iterations (86.000%)
==============================================================================================
Hall of Fame:
---------------------------------------------------------------------------------------------
Complexity  Loss       Score     Equation
3           1.708e-02  5.314e+00  y = -0.073884 * -0.77004
5           5.798e-03  5.401e-01  y = 1.0953 * (0.82268 - x0)
7           3.690e-03  2.259e-01  y = (0.83205 - x0) * (1.5454 - x0)
9           2.859e-03  1.276e-01  y = (0.88501 - x0) * (1.8414 - (x0 + x0))
10          2.834e-03  8.746e-03  y = custom_function(0.88676) * ((x0 - 0.88676) * (0.88676 - x0...
                                  ))
11          2.809e-03  8.867e-03  y = (x0 - 0.90462) * (((-1.7693 + x0) + x0) * 1.0997)
12          2.804e-03  2.049e-03  y = ((x0 - 0.90212) * 1.0593) * ((custom_function(0.20822) + x...
                                  0) + x0)
17          2.800e-03  2.783e-04  y = ((x0 - 1.1832) * (custom_function(0.37198) + (x0 + x0))) +...
                                   ((-0.052411 * custom_function(x0)) * x0)
19          2.777e-03  4.149e-03  y = ((x0 - 0.97894) * (x0 - 0.97894)) * ((1.5362 - (x0 * ((x0 ...
                                  + -0.48992) + x0))) + 0.17568)
20          2.762e-03  5.127e-03  y = (((x0 * ((0.52577 - -0.12462) + (custom_function(0.12709) ...
                                  * x0))) + 1.331) * (0.88898 - x0)) * (1.32 - x0)
---------------------------------------------------------------------------------------------
==============================================================================================
```

We encourage parametrizations:

1) $H^{u-d}(x,t) = P(x)F(t)$
2) $H^{u-d}(x,t) = x^{\alpha}(1-x)^{\beta}P(x,t)$
3) $H^{u-d}(x,t) = x^{\alpha}(1-x)^{\beta}P(x)F(t)$

with custom loss:

- MSE + Non-Factorization Penalty
- MSE + Integral Constraint $\int_0^1 dx H^{u-d}(x,0,0) = \ldots,$

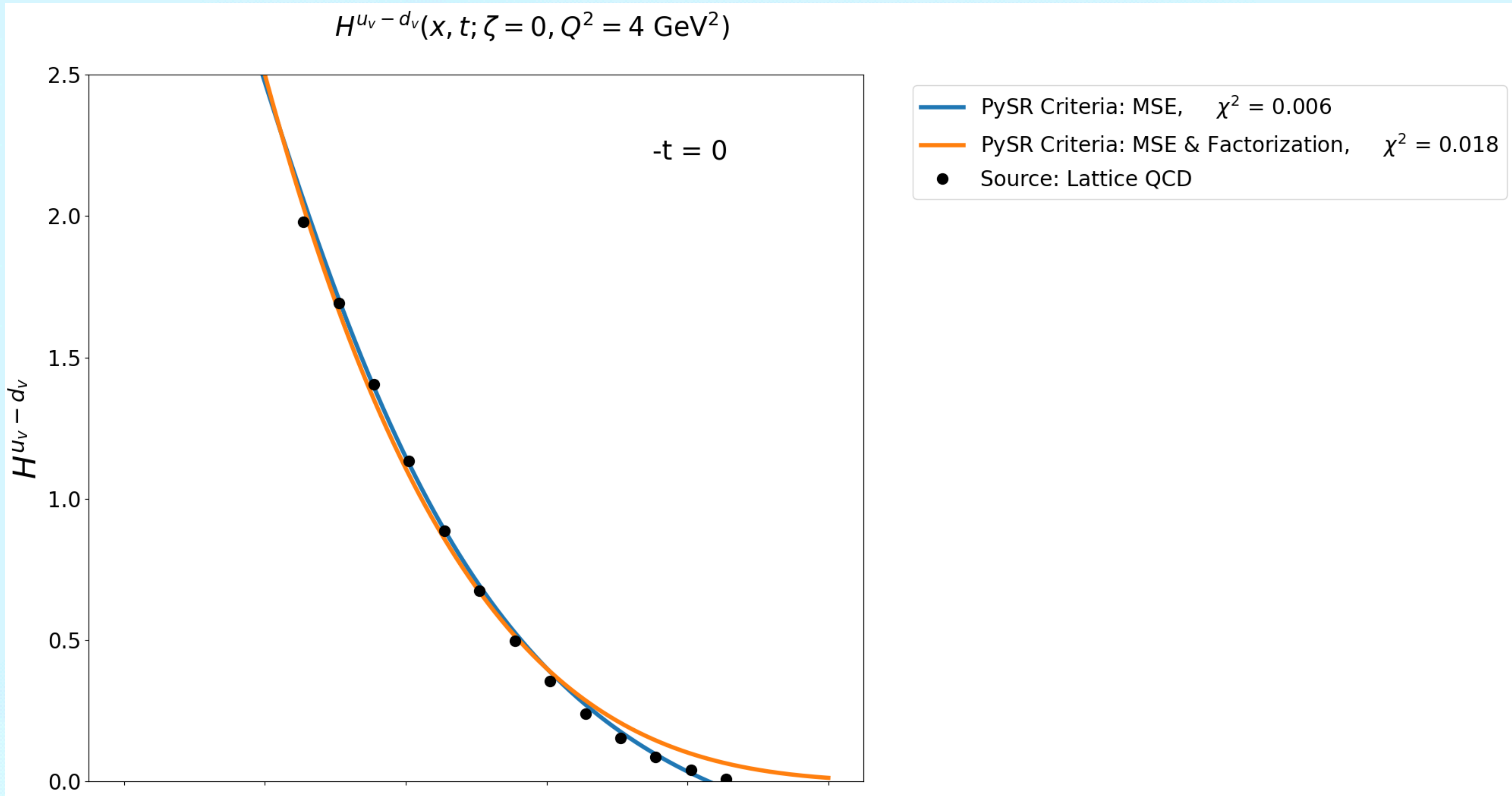
complexity:

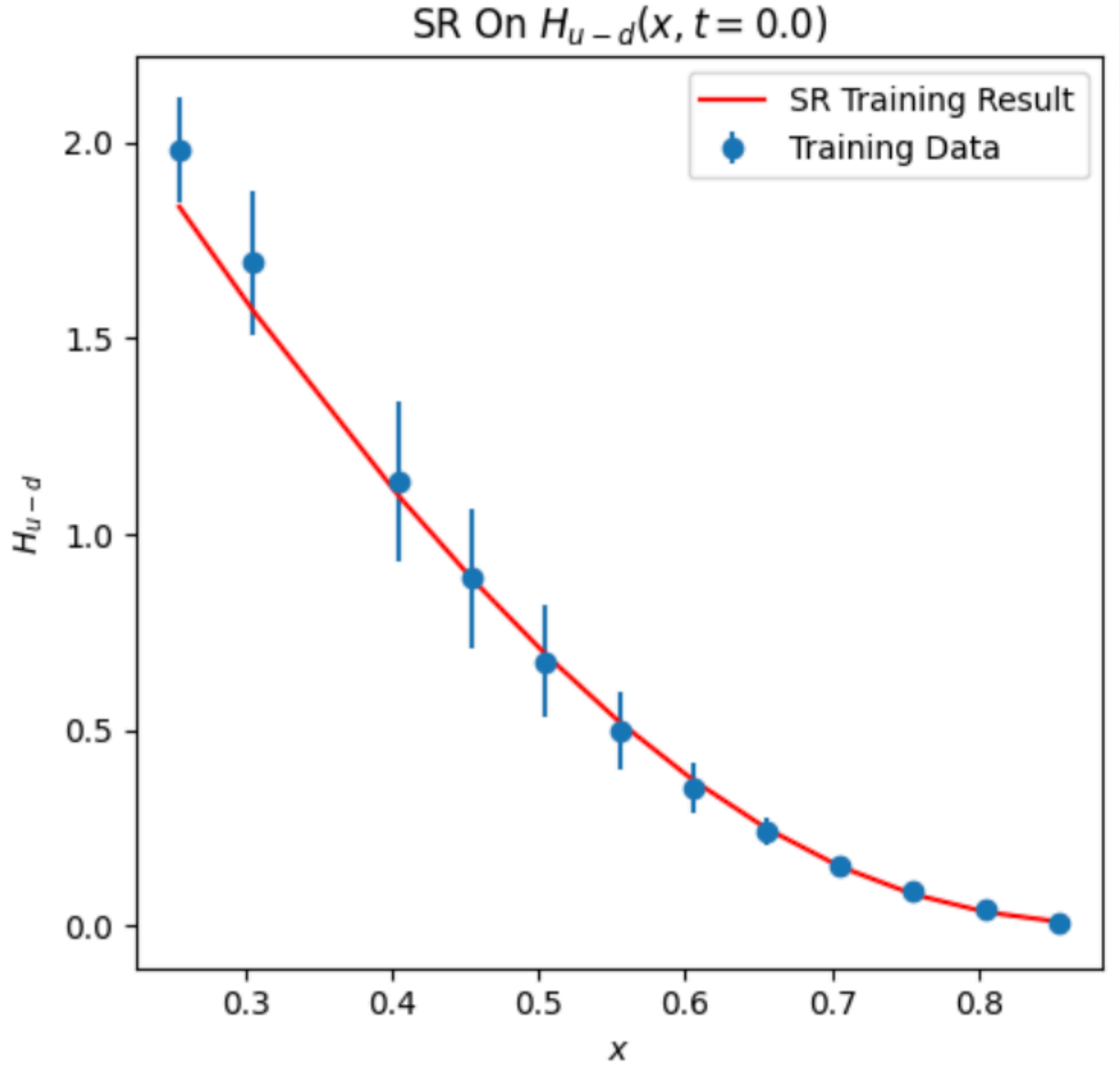
- Maximum complexity allowed for power operator (a,b) = (4,1) $(1-x)^{\beta}$,

and forcing the PDF form:

- replace eq. 2) with $H^{u-d}(x,y,t) = x^{\alpha}y^{\beta}P(x,t)$, where y = 1-x, and do 3-D fit.

Results on lattice

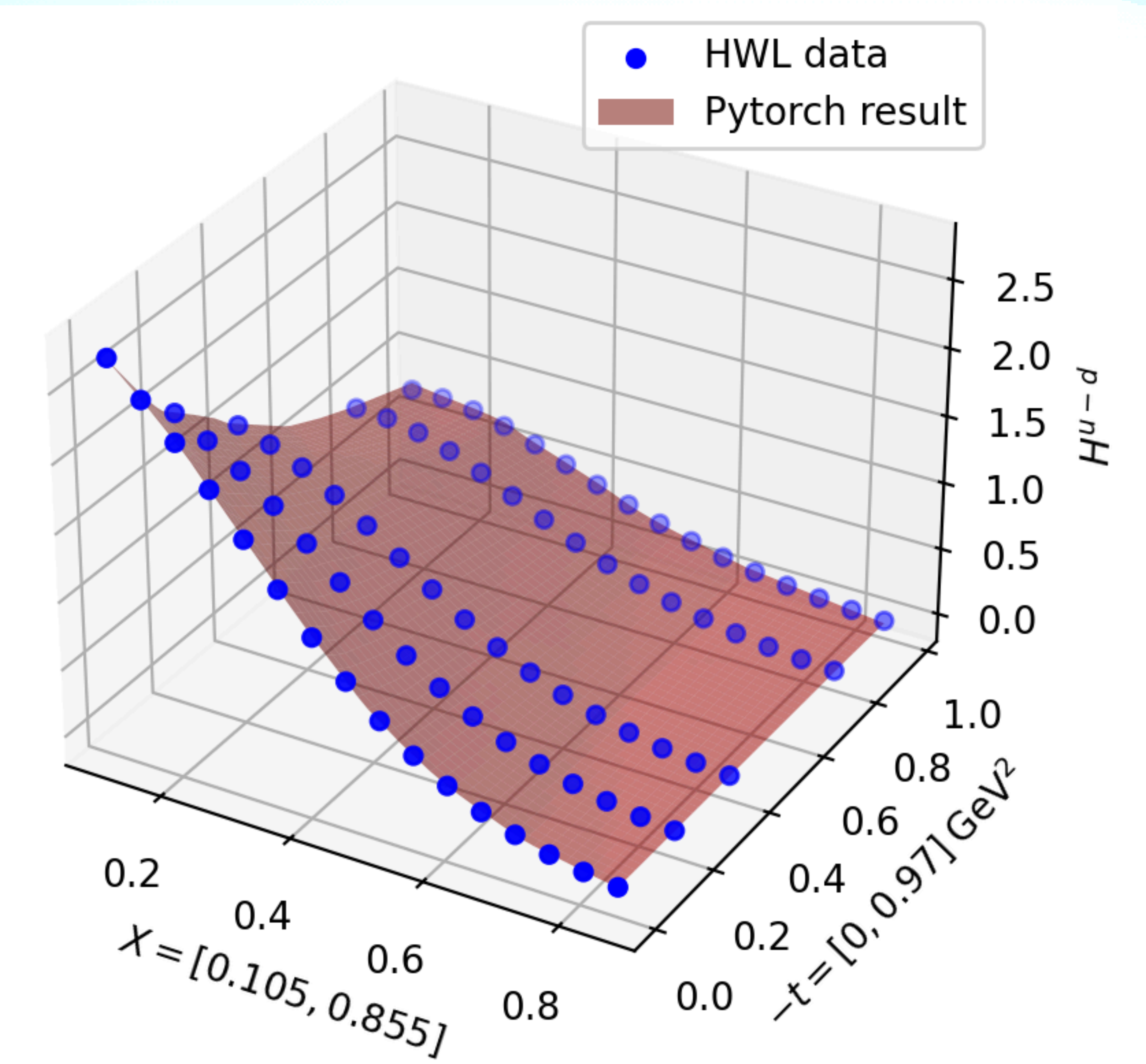




MSE: $H^{u-d}(x,0,t) = \dfrac{0.831 - x}{-t(x + 0.24) + x^2 + 0.21}$

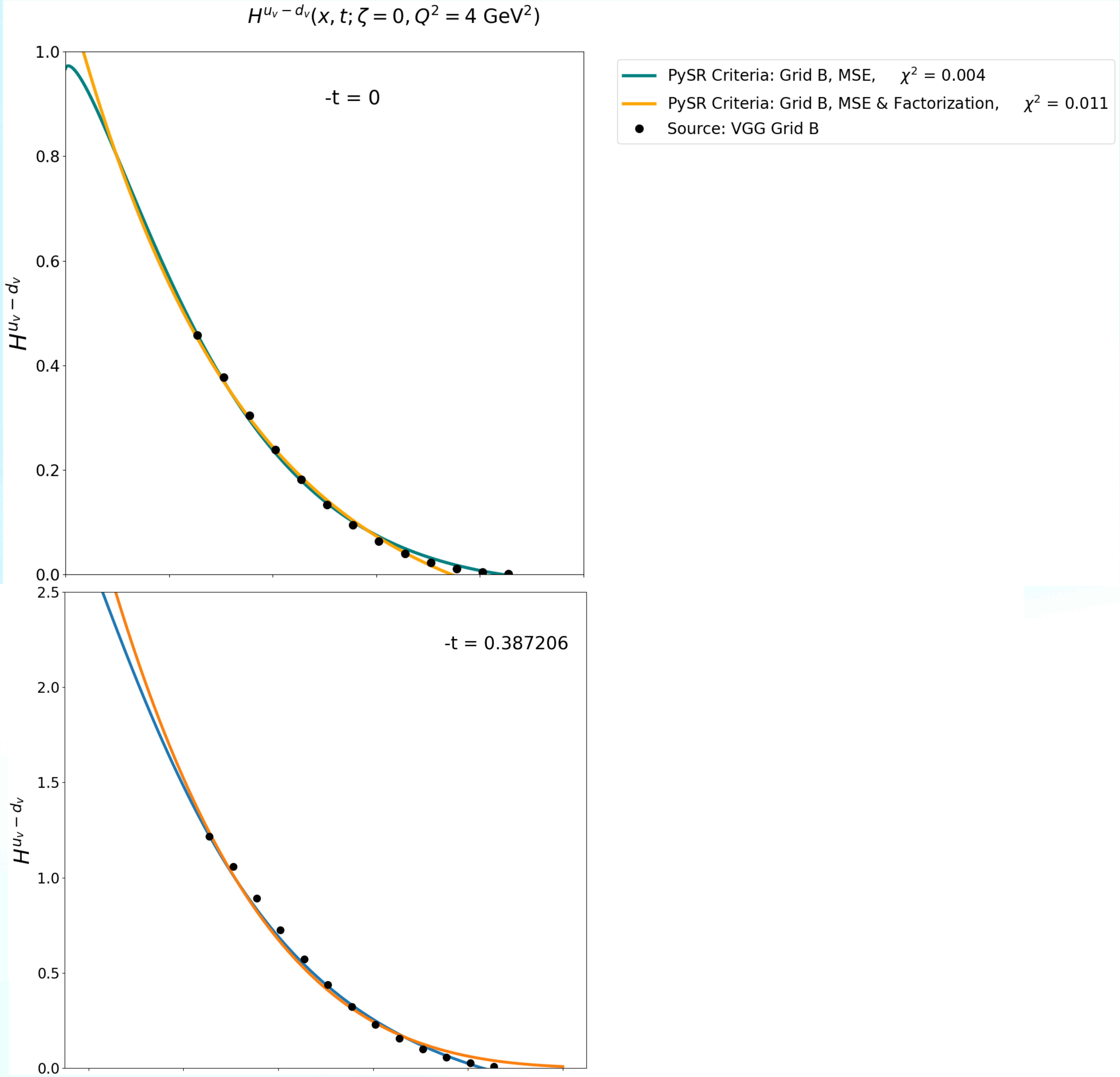
MSE+factorization: $H^{u-d}(x,0,t) = \dfrac{3(1 - 0.77x)^{4.1}}{-t + 0.6}$

forced PDF form: $H^{u-d}(x,0,t) = x^{-0.09}(1-x)^{0.43}e^{-0.68\ln^2(1-x)}e^{-t+1.01}$



Comparison to PyTorch

Figures by A. Dotson, A. Reddy and Z. Panjsheeri



$$\text{MSE: } H^{u-d}(x,0,t) = \frac{0.02^x - 0.03}{-3t + x^x}$$

$$\text{MSE+factorization: } H^{u-d}(x,0,t) = \frac{0.05^x - 0.11}{-2.03t + 0.79}$$

Factorized form has worse χ^2,
even though VGG has implemented factorization!

Can we benchmark?

- We need uncertainty quantification for all results and phenomenological models!
- We need to understand if interpolation and extrapolation are reliable
- Likelihood analysis as a means of assessing the information content in the data