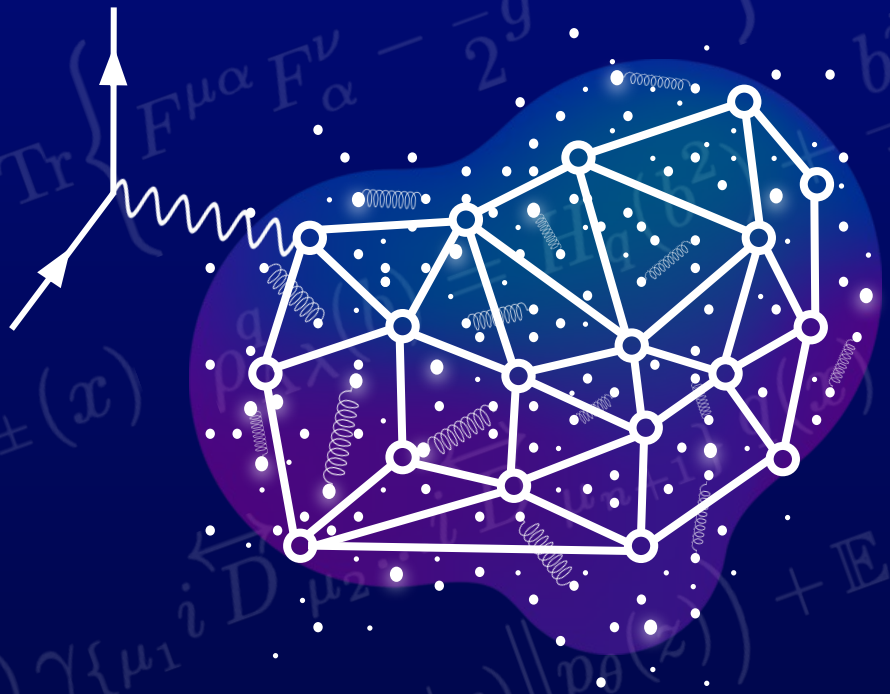


Explainable AI classification for parton density theory

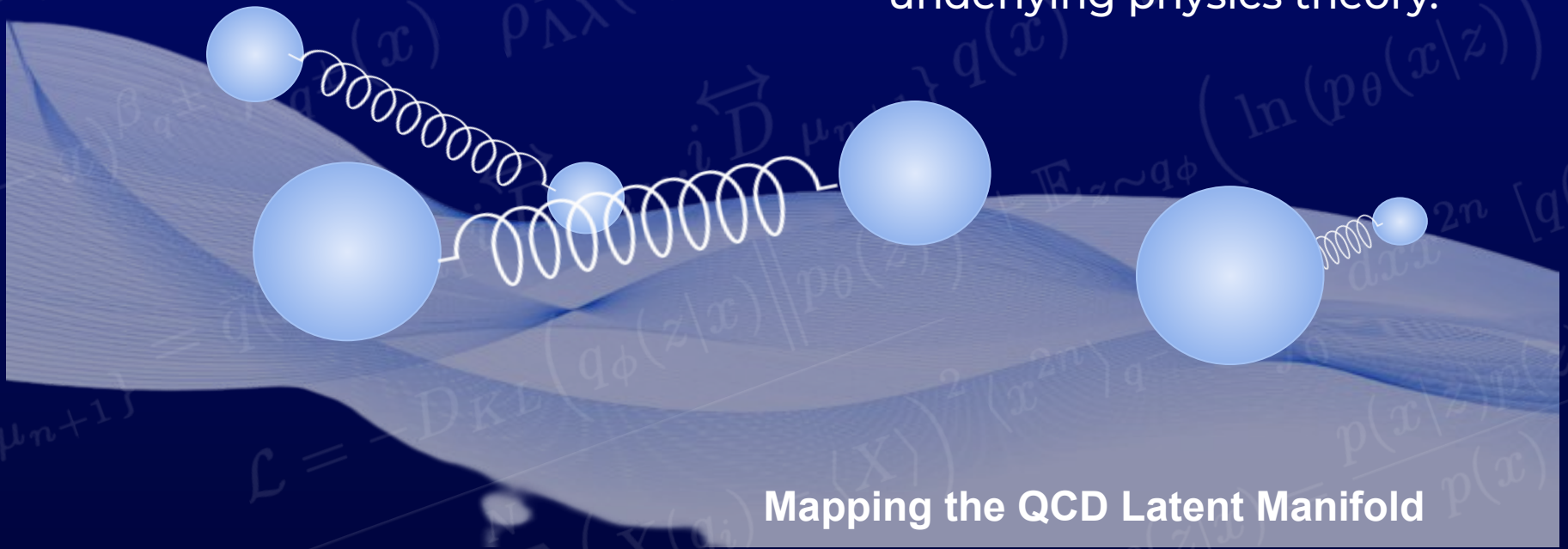
Brandon Kriesten • 9 August 2024 • Exclusive Reactions III

Motivation

- Neural network representations of quantum correlation functions
  - reformatting a phenomenological fit as an *inverse problem*
  - physics constraints (lattice QCD inputs / theory constraints / exp. data)
  - possible many solutions
- Raises a jumble of questions with neural networks
  - How do we quantify uncertainties? aleatoric / epistemic / distributional separation?
  - Can we interpret the ‘black-box’?
  - How do we trust neural networks for physics research?

Machine learning for Theoretical Physics

Leveraging novel methods to unravel aspects of the underlying physics theory.



Mapping the QCD Latent Manifold

Outline

- Physics constrained models of collinear hadron structure
- Explainable fits for Parton Distribution Functions
- Connection to 3D Hadronic Structure
- Conclusions & Outlooks

Outline

- **Physics constrained models of collinear hadron structure**
- Explainable fits for Parton Distribution Functions
- Connection to 3D Hadronic Structure
- Conclusions & Outlooks

Parton Distribution Functions from latent space Mellin moments

If we have an infinite amount of information, we can construct the PDF exactly from classical methods

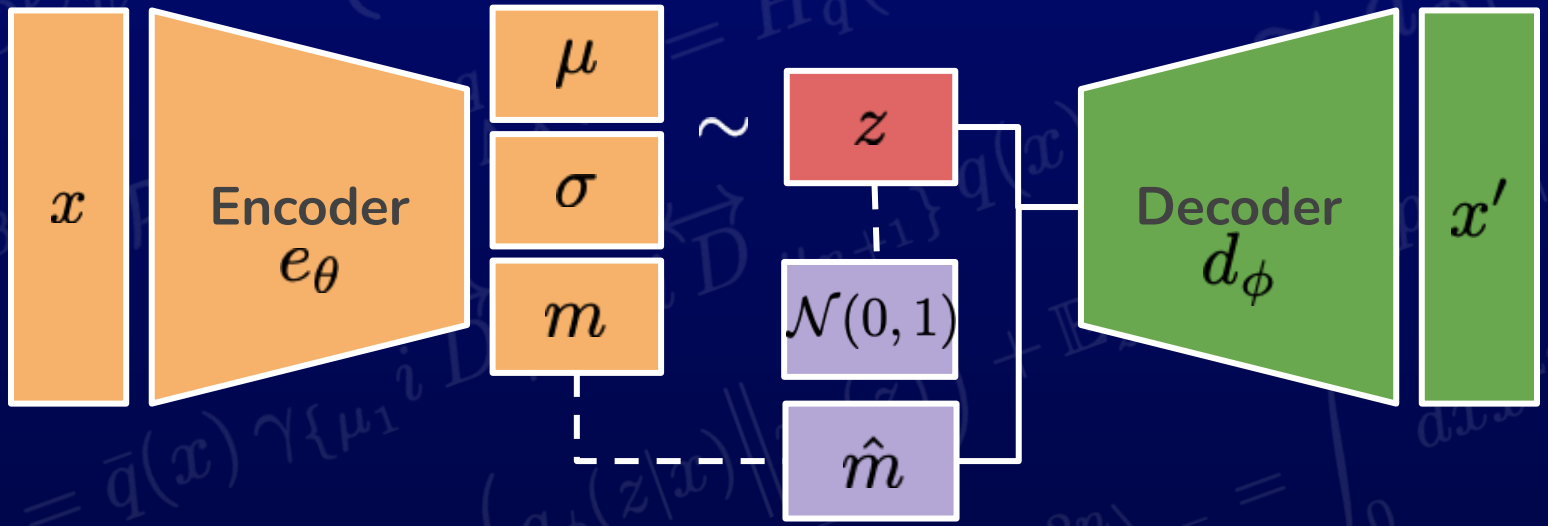
$$q(x) + (-1)^{n+1}\bar{q}(x) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} dn x^{-n-1}\langle x^n \rangle_q$$

Well motivated problem with many explorers!

We don't have an infinite amount of information, typically we have just a few of these moments from the lattice.

BK, T.J. Hobbs **arXiv: 2312.02278 (submitted to PRD)**

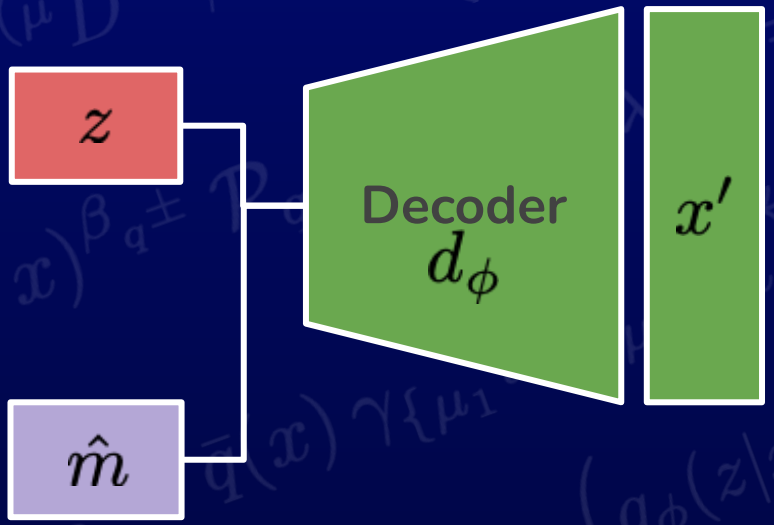
Variational Autoencoder
Inverse Mapper



We utilize variational autoencoders as a powerful tool to dissect inverse problems!

M. Almaeen, Y. Alanazi, N. Sato, W. Melnitchouk, M.P. Kuchera, Y. Li **IJCNN (2021)**

Variational Autoencoder
Inverse Mapper



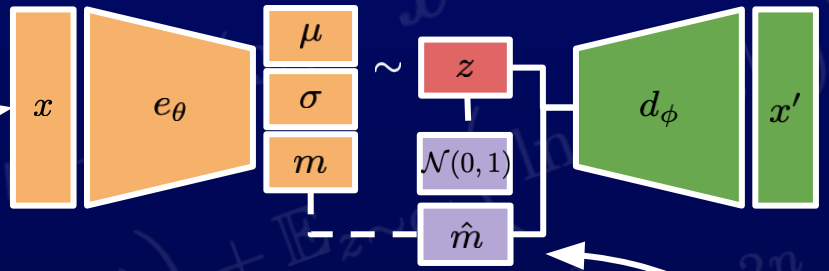
The goal is to train a decoder model to accept latent information and an observable to generate a never before seen input!

Generative AI provides access to new algorithms beyond neural network interpolation.

M. Almaeen, Y. Alanazi, N. Sato, W. Melnitchouk, M.P. Kuchera, Y. Li **IJCNN (2021)**

Parton Distribution Functions from latent space Mellin moments

Question: How well can we determine the full x-dependence of PDFs from a finite number of Mellin moments?



Model Inputs

Randomly generated PDFs with 5 parameters.

$$q(x) \pm \bar{q}(x) = \mathcal{N}_{q^\pm} x^{\alpha_{q^\pm}} (1-x)^{\beta_{q^\pm}} \mathcal{P}_{q^\pm}(x)$$

$$\mathcal{P}_{q^\pm}(x) = 1 + \gamma_{q^\pm}\sqrt{x} + \delta_{q^\pm} x$$

Latent Observable

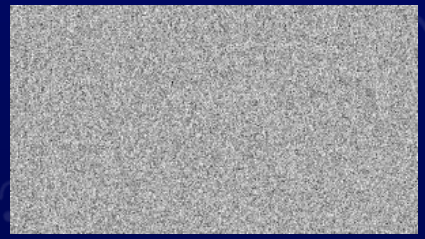
Organized and interpretable as a series of moments

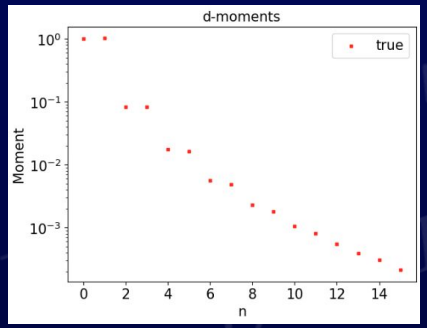
$$\langle 1 \rangle_{q^-}, \langle x \rangle_{q^+}, \langle x^2 \rangle_{q^-}, \langle x^3 \rangle_{q^+}, \ldots$$

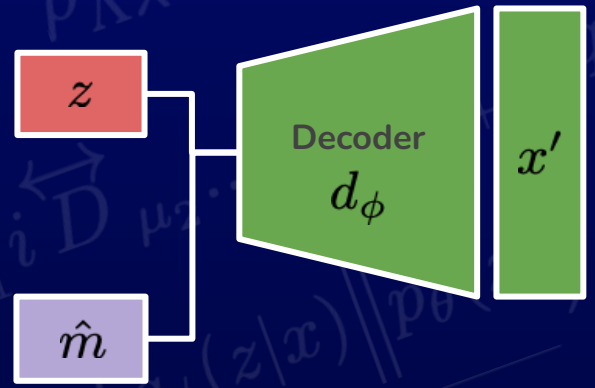
BK, T.J. Hobbs **arXiv: 2312.02278 (submitted to PRD)**

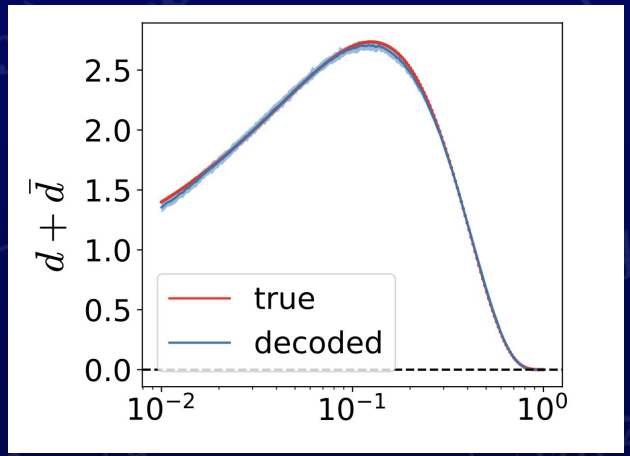
Parton Distribution Functions from latent space Mellin moments

Making predictions from a trained decoder model.









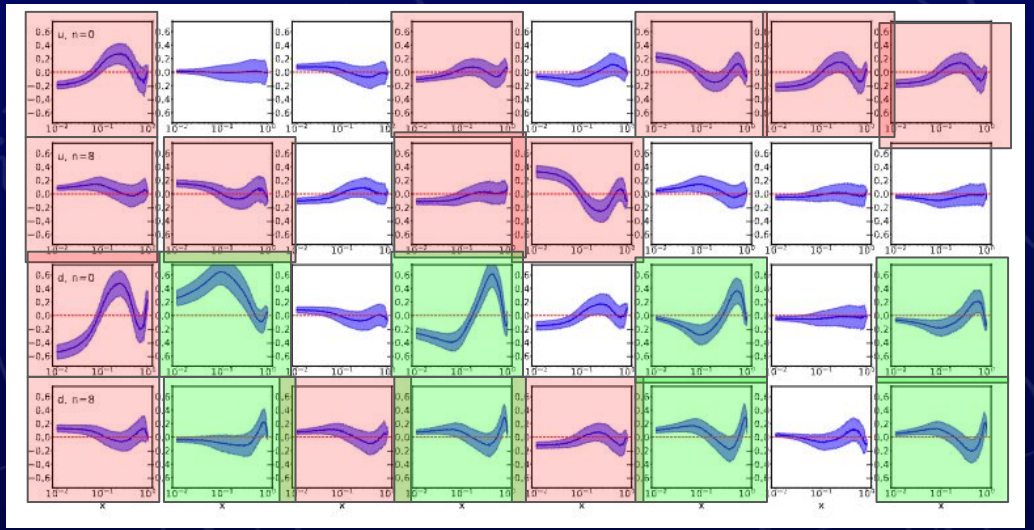
BK, T.J. Hobbs **arXiv: 2312.02278 (submitted to PRD)**

Parton Distribution Functions from latent space Mellin moments

We utilize the Pearson correlation between the learned moments from the encoder and the decoded PDF (d^+) as an explainability technique. One can see spurious correlated effects as well as consistent correlations.

$$\mathrm{Corr}(X,Y) = \frac{\langle XY\rangle - \langle X\rangle\langle Y\rangle}{\Delta X \Delta Y}$$

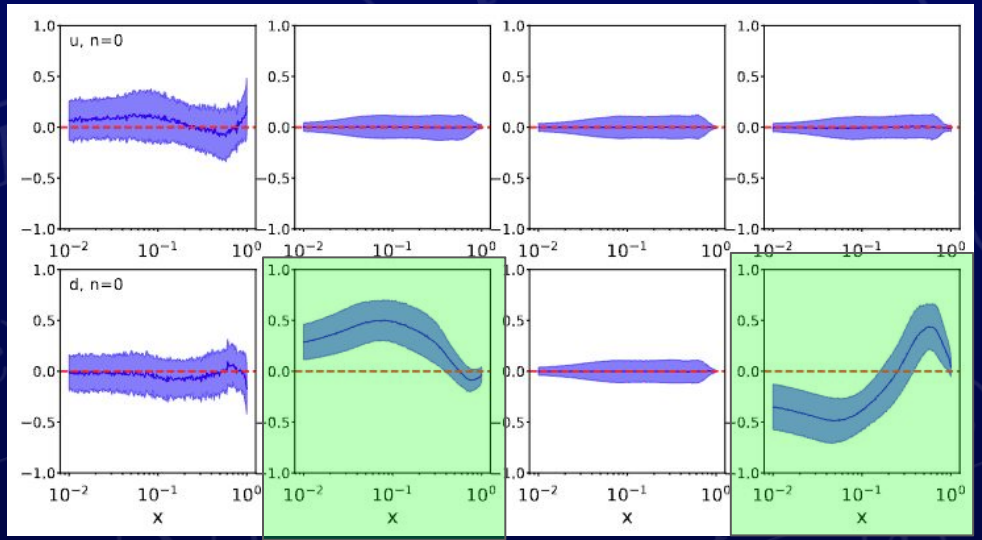
$$\mathrm{Corr}[d^+(x), \langle x^n\rangle_{u^\pm, d^\pm}]$$



BK, T.J. Hobbs **arXiv: 2312.02278 (submitted to PRD)**

Parton Distribution Functions from latent space Mellin moments

$$\mathrm{Corr}[d^+(x), \langle x^n \rangle_{u^\pm, d^\pm}]$$



With a more dramatically undercomplete autoencoder architecture, the correlations are statistically consistent with 0 everywhere except for the d^+ moments. Obvious spurious correlations seem to disappear.

BK, T.J. Hobbs **arXiv: 2312.02278 (submitted to PRD)**

Outline

- Physics constrained models of collinear hadron structure
- **Explainable fits for Parton Distribution Functions**
- Connection to 3D Hadronic Structure
- Conclusions & Outlooks

Explainability vs. interpretability

Explainability

Explainability techniques are tools that are used to describe why a neural network made the decisions it did in human readable terms.

Such tools are hooks in the model to inspect gradients, or specialized backpropagation tools.

Interpretability

Interpretable models are human readable by construction and don't require any other tools. Such models include linear models, decision trees, etc.

Often, highly interpretable models are not the most accurate models for complex datasets.

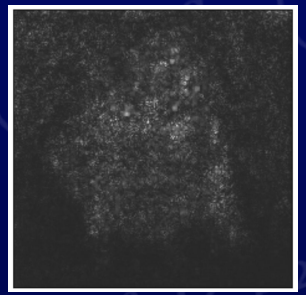
BK, J. Gomprecht, T.J. Hobbs **arXiv: 2407.03411 (submitted to JHEP)**

Explainability ... a fun example!

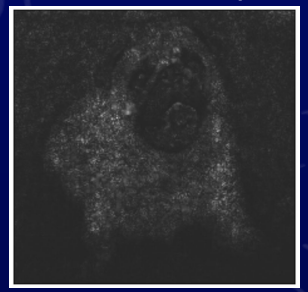
A survey of techniques

Input

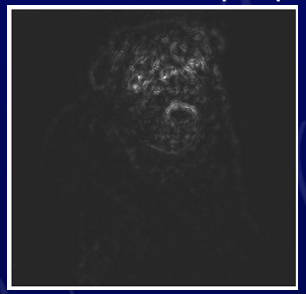
Gradients



Gradients ⊙ Input

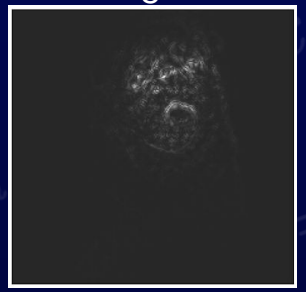


Guided Backprop



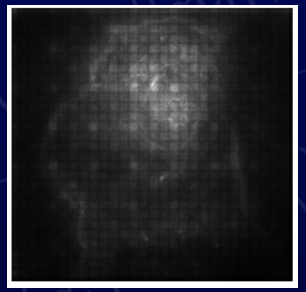
gradCAM

Guided gradCAM



Integrated Gradients

smoothGrad



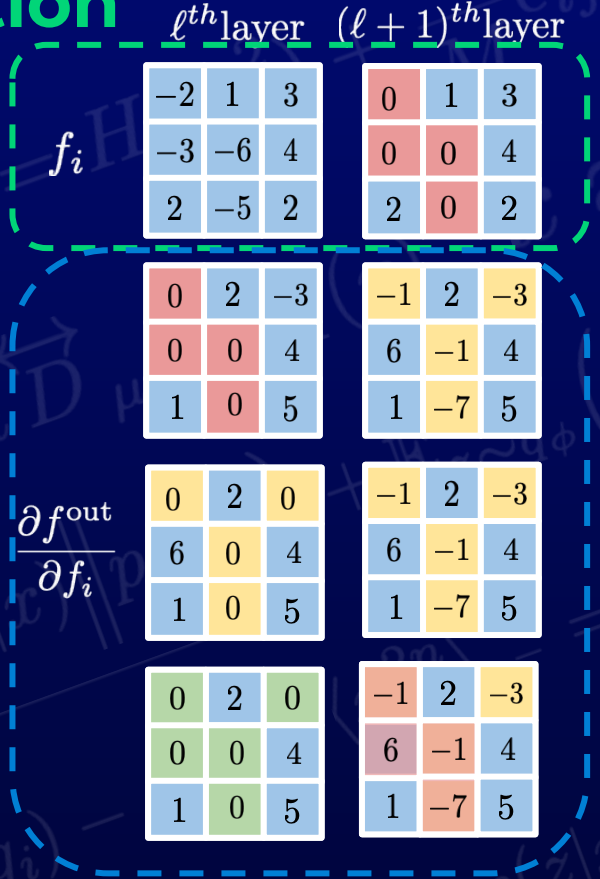
Occlusion

Edge Detection

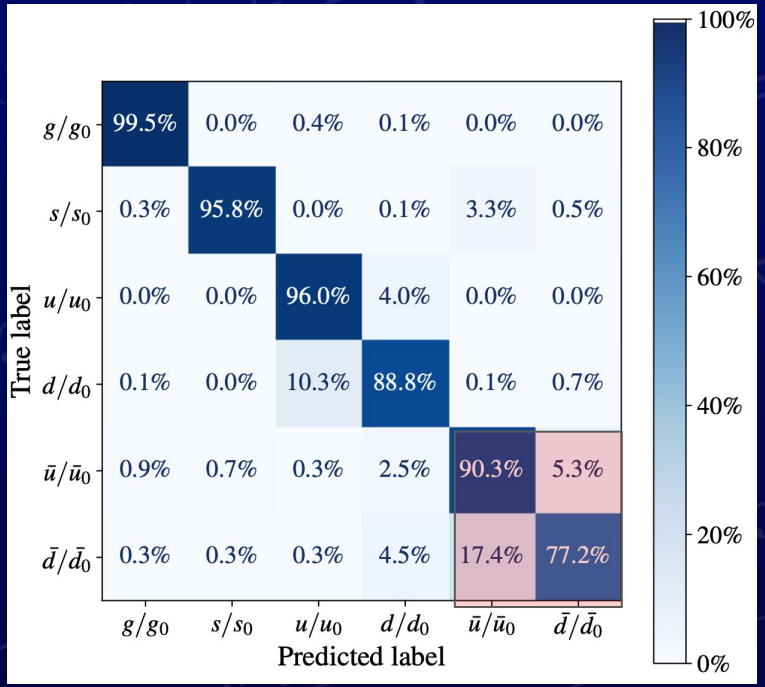
Guided Backpropagation

$$\frac{\partial f^{\text{out}}}{\partial f_i^{\ell}} = (f_i^{\ell} > 0) \cdot \left(\frac{\partial f^{\text{out}}}{\partial f_i^{\ell+1}} > 0 \right) \cdot \frac{\partial f^{\text{out}}}{\partial f_i^{\ell+1}}$$

Guided backprop is a technique in which the gradients of a neural network layer are masked during backpropagation holding the weights fixed to determine which input features positively affect the classification outcome the most.



Explainability within fitted PDFs



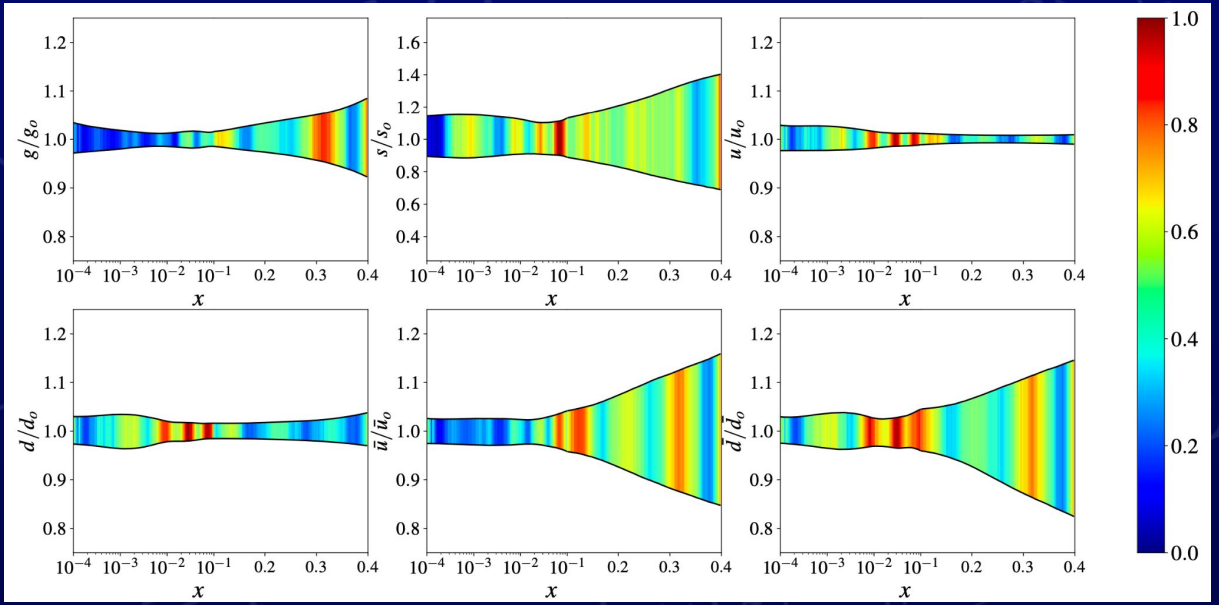
We Monte-Carlo sample the fitted PDF error set to generate training data for a parton flavor classifier.

The confusion between the u-bar and d-bar ratios is related to flavor-separation challenges in phenomenological fits of the sea-quark densities;

Ex. highlights the importance of measurements of the d-bar/u-bar asymmetry in experiments such as SeaQuest.

BK, J. Gomprecht, T.J. Hobbs **arXiv: 2407.03411 (submitted to JHEP)**

Explainability within fitted PDFs



The large variances in the gradients (depicted by red bands) occur in regions where there are more significant shape changes amalgamated over the set of MC replicas as compared to the other PDF flavors.

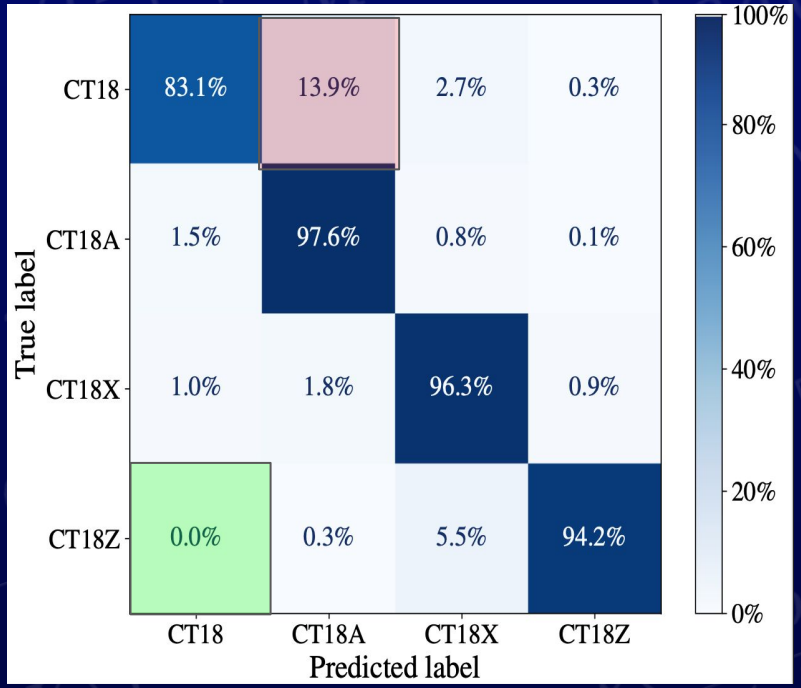
The regions where the gradients vary the most are regions where there are shape changes that are unique to that particular PDF.

BK, J. Gomprecht, T.J. Hobbs **arXiv: 2407.03411 (submitted to JHEP)**

XAI4PDF: Explainability across fitted PDFs

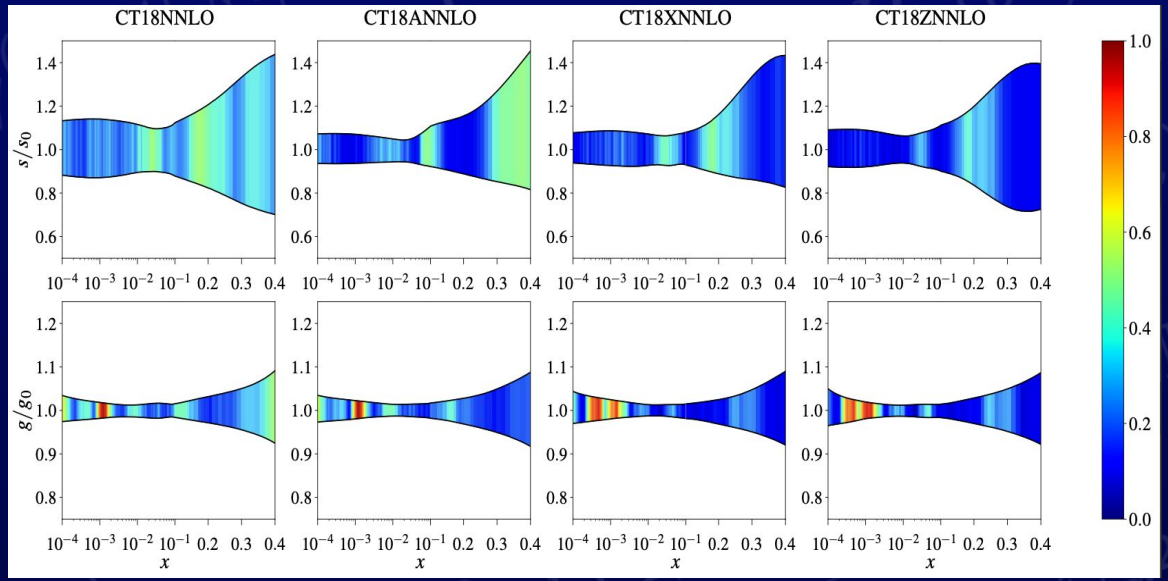
PDF fits	Factorization scale in DIS	ATLAS 7 TeV W/Z data included?	CDHSW $F_2^{p,d}$ data included?	Pole charm mass, GeV
CT18	$\mu^2_{F,DIS} = Q^2$	No	Yes	1.3
CT18A	$\mu^2_{F,DIS} = Q^2$	Yes	Yes	1.3
CT18X	$\mu^2_{F,DIS} = 0.8^2\left(Q^2 + \frac{0.3\ \mathrm{GeV}^2}{x_B^{0.3}}\right)$	No	Yes	1.3
CT18Z	$\mu^2_{F,DIS} = 0.8^2\left(Q^2 + \frac{0.3\ \mathrm{GeV}^2}{x_B^{0.3}}\right)$	Yes	No	1.4

The two analyses which are “furthest” from each other (CT18 and CT18Z) are also the least confused, confirming that the shift in theory assumptions drives the statistical distinguishability as inferred by the XAI calculation.



BK, J. Gomprecht, T.J. Hobbs **arXiv: 2407.03411 (submitted to JHEP)**

XAI4PDF: Explainability across fitted PDFs



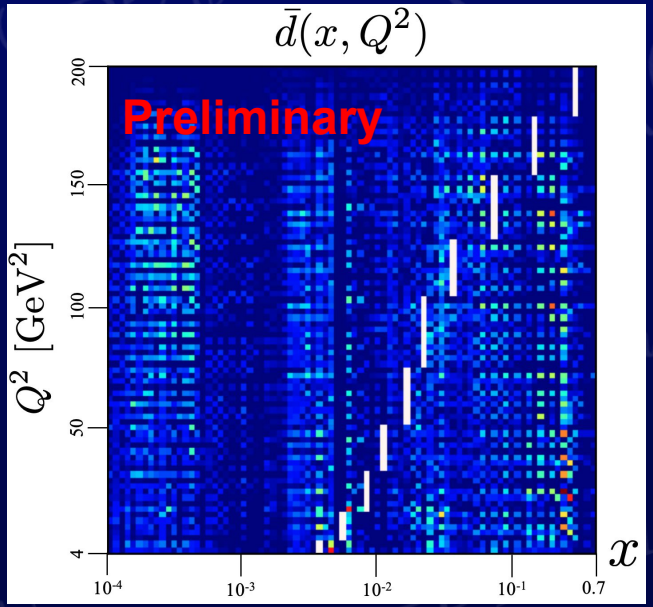
The strange and gluon PDFs stand out while discerning between different theory fits!

The gluon replicas have a dominant role in the classification among the CT18 series with highly localized gradients.

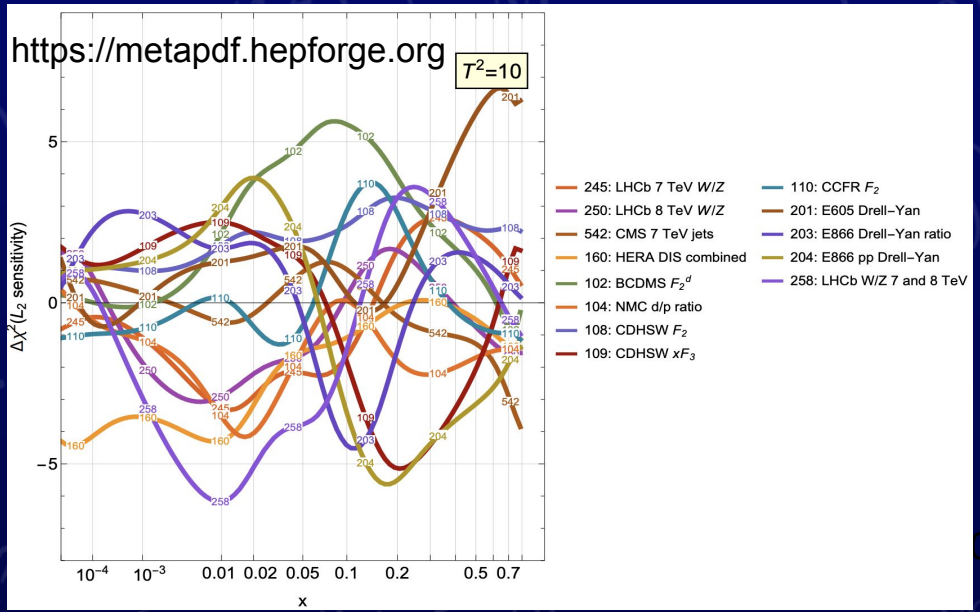
The strange replicas have smoother gradients indicating a weaker role.

BK, J. Gomprecht, T.J. Hobbs **arXiv: 2407.03411 (submitted to JHEP)**

XAI4PDF: Explainable AI for PDFs



BK, T.J. Hobbs **(in progress)**



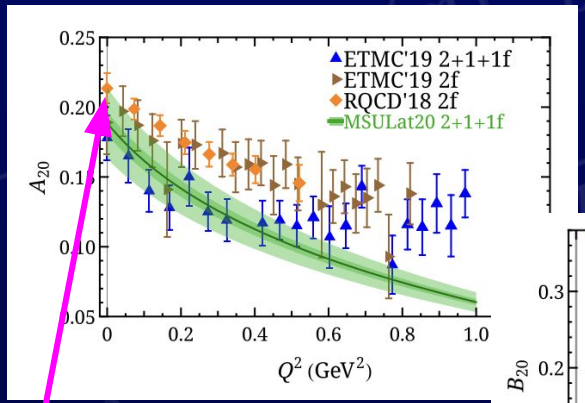
χ^2 on the CDHSW F_2 data (neutrino-iron CC DIS) traces back to regions in the phase space of the fitted PDF. Connections to PDF sensitivities??

Outline

- Physics constrained models of collinear hadron structure
- Explainable fits for Parton Distribution Functions
- **Connection to 3D Hadronic Structure**
- Conclusions & Outlooks

Connections to 3D Hadron Structure: Lattice

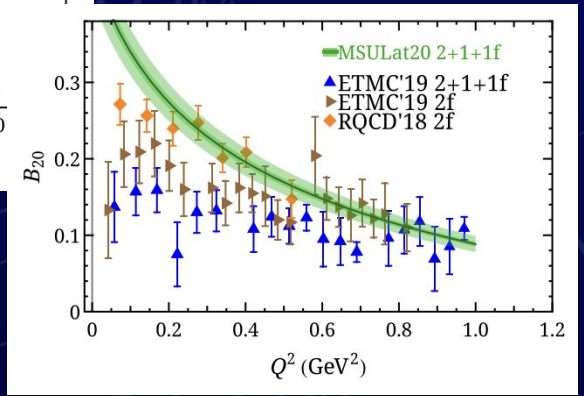
Incorporating Lattice QCD calculated moments with full t-dependence in GPD pheno fits.



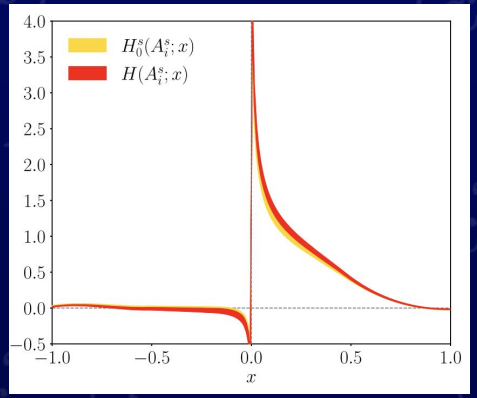
H-W Lin 2022

BK, T.J. Hobbs 2023

A good space to test the VAIM with higher dimensional object.



Constraining phenomenological fits of GPDs with x-dependence calculated from the Lattice.



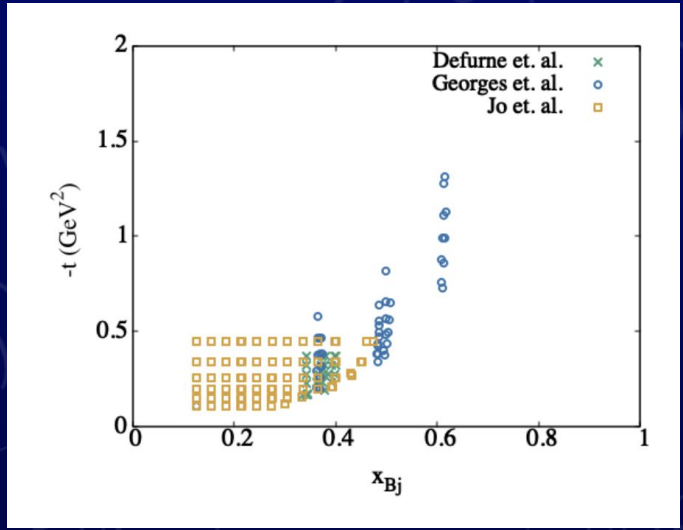
Bhattacharya et. al. 2022

Connections to 3D Hadron Structure: Experiment

There are many experimental DVCS programs, it is crucial to understand which experiment has what effect on the statistical fit of the GPDs.

Use XAI to investigate how pulls on the total χ^2 of a phenomenological fit of GPDs depends on the experiment.

Identify high-impact regions in the phase space that have significant pulls on GPD fit.



Grigsby et. al.Phys. Rev. D. 104 (2021)

Outline

- Physics constrained models of collinear hadron structure
- Explainable fits for Parton Distribution Functions
- Connection to 3D Hadronic Structure
- **Conclusions & Outlooks**

Conclusions

ML in HEP / NP is nascent, not to mention theory applications, but is expected to grow substantially.

The research I have discussed here can translate from PDFs to GPDs - solving large scale inverse problems for robust and explainable insights into hadron structure.

The activity fills an urgent need for more direct collaboration between theorists, data scientists, and computational experts.

Thank you for your attention!