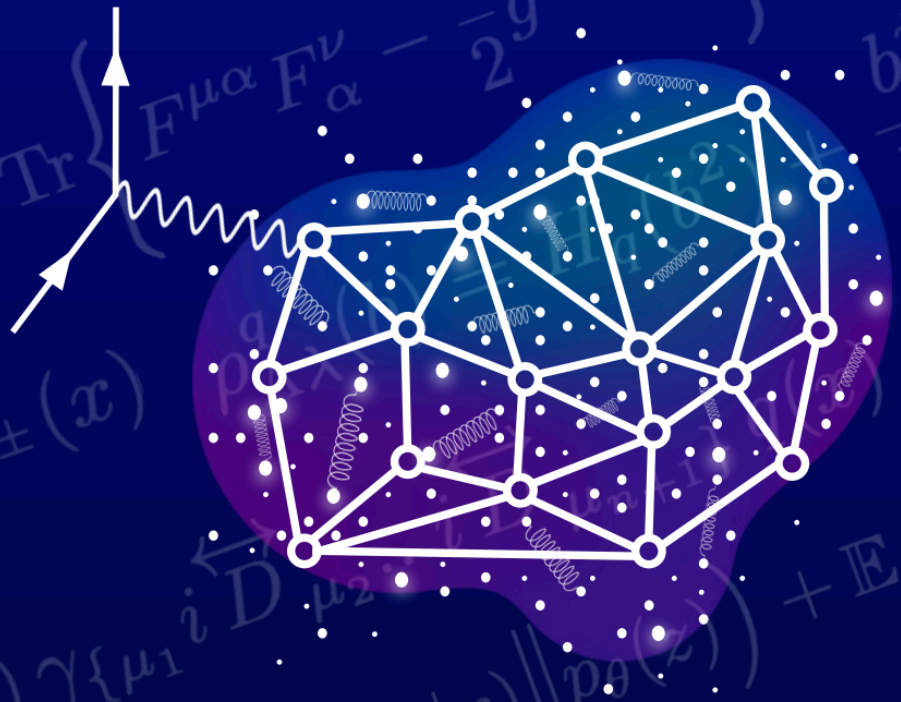


Progress in Explainable AI for Parton Density Theory

Brandon Kriesten • 7 October 2024 • QCD@LHC 2024

Argonne NATIONAL LABORATORY

Argonne NATIONAL LABORATORY

Motivation / Outline

- **Challenges in PDF determination and precision theory**
  - **reformatting a phenomenological PDF fit as an *inverse problem***
  - **physics constraints (lattice QCD inputs, theory constraints)**
  - **Uncertainty quantification - major limitation in physics searches**
- **A jumble of questions with machine learning**
  - **How do we quantify uncertainties?**
    - **Aleatoric / epistemic (/ distributional OOD) separation?**
  - **Can we dissect and explain the 'black-box'?**
  - **Repurpose standard ML tools for physics discovery …**
- **<u>Works</u>:**
  - reconstruct PDFs from their Mellin moments — BK, T.J. Hobbs **arXiv: 2312.02278**
  - explore explainability techniques — BK, J. Gomprecht, T.J. Hobbs **arXiv: 2407.03411**
  - uncertainty quantification studies — BK, T.J. Hobbs (in progress)

Argonne NATIONAL LABORATORY

Precision Theory and uncertainty

Dall-E: “What does QCD at the LHC look like?”

Many measurements at the LHC to test the SM and search for new physics involve colliding protons and critically rely on the subsequent understanding of the partonic level interactions with their uncertainties - PDFs.

So,what Dall-E should have showed me is something like this …

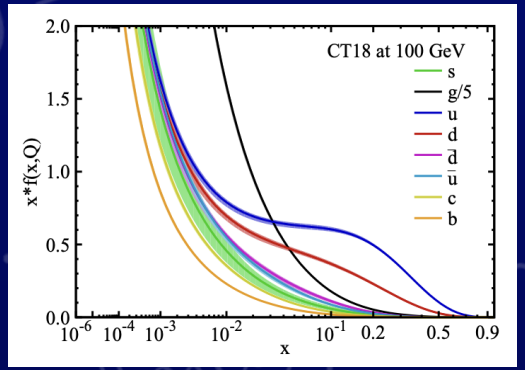


Image credit: Phys.Rev.D 103 (2021)

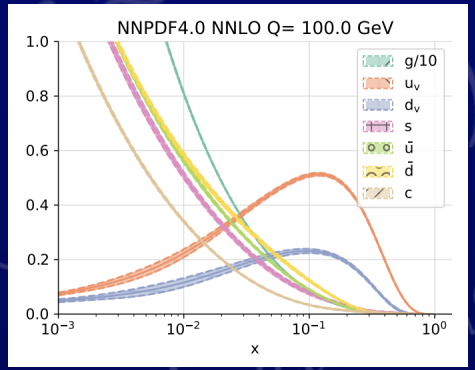


Image credit: Eur.Phys.J.C 82 (2022)

Global fits resulting from an expansive and diverse dataset representative of two powerful methodologies. Analytic and Neural Network parameterizations.

Argonne NATIONAL LABORATORY

Precision Theory and uncertainty

PDF uncertainties place limitations on scope of phenomenology both at the LHC and beyond.

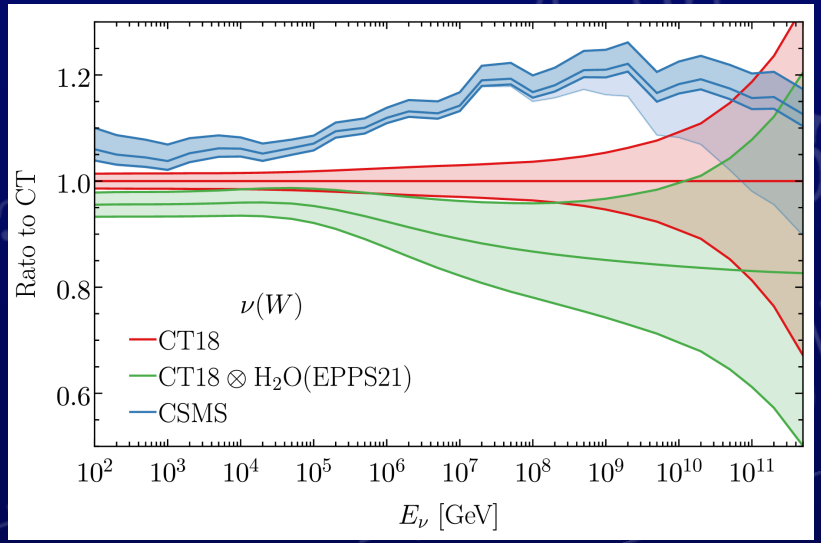


Image credit: 2303.13607

Neutrino DIS cross sections to astrophysical scales

Impacts of small-x extrapolations, nuclear corrections, and scale uncertainties on the neutrino-nucleus νDIS cross section.

Relevant for LHC forward physics facilities such as FASERν and neutrino experiments / telescopes such as KM3NET, ICECube, and DUNE.

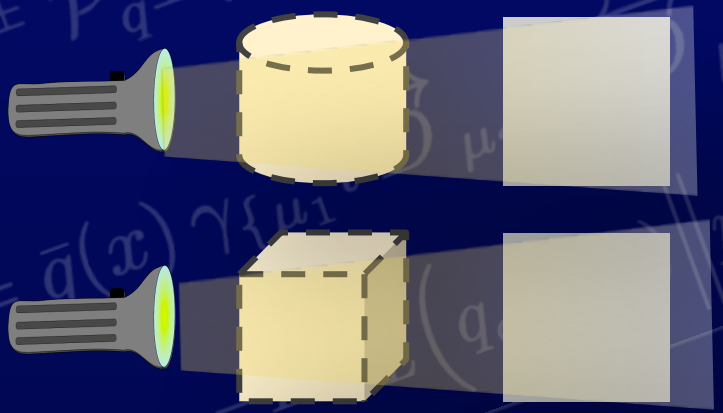
Key takeaways: Inverse problems are fundamental and ubiquitous contributing to uncertainty - in which there could be signs of new physics.

Argonne NATIONAL LABORATORY

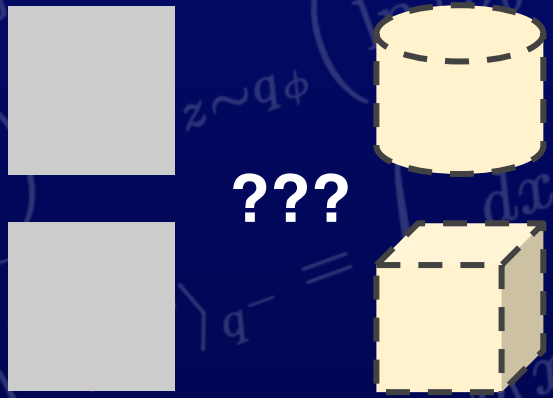
Ill-posed inverse problems

The class of inverse problems in QCD that are most common in theory are the non-uniqueness of the inverse mapping.

Forward Mapping



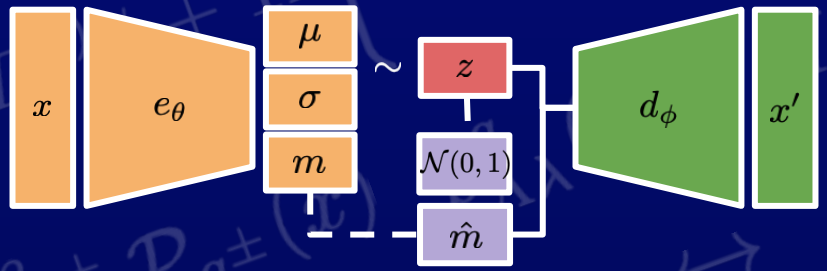
Inverse Mapping



Q: Which shape is it? A: It's both!

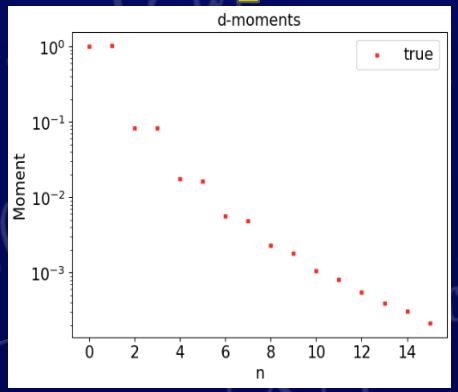
Argonne NATIONAL LABORATORY

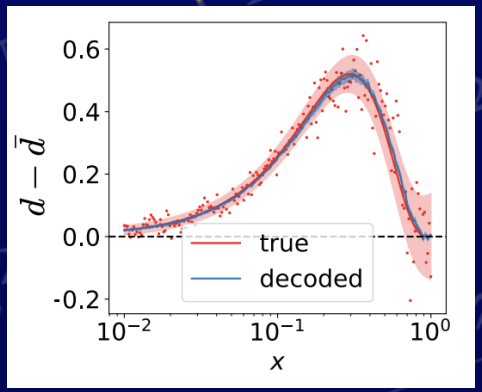
Generative AI for Inverse Problems



Using variational autoencoder as powerful generative model to generate solutions to inverse problems.

The latent variables are organized into interpretable physics constraints such as Mellin moments calculated on the lattice.



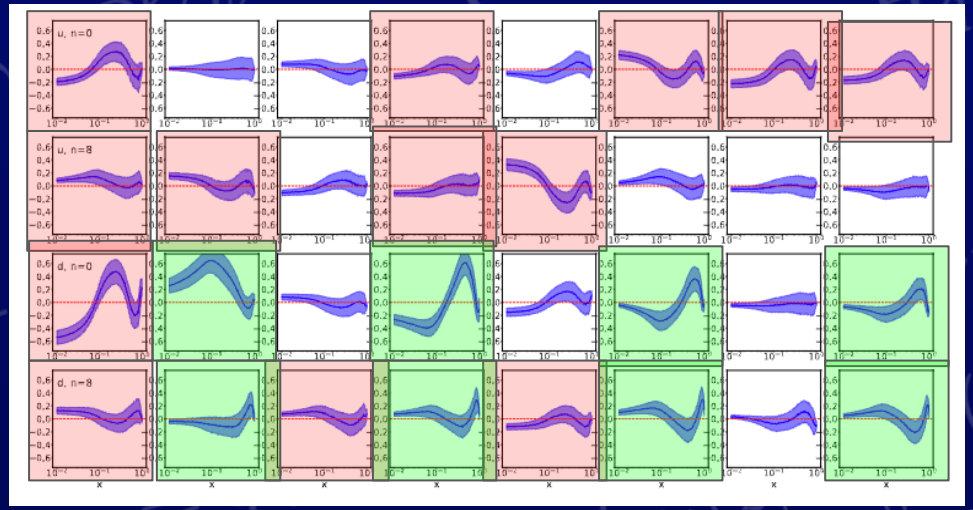


BK, T.J. Hobbs **arXiv: 2312.02278**

Argonne NATIONAL LABORATORY

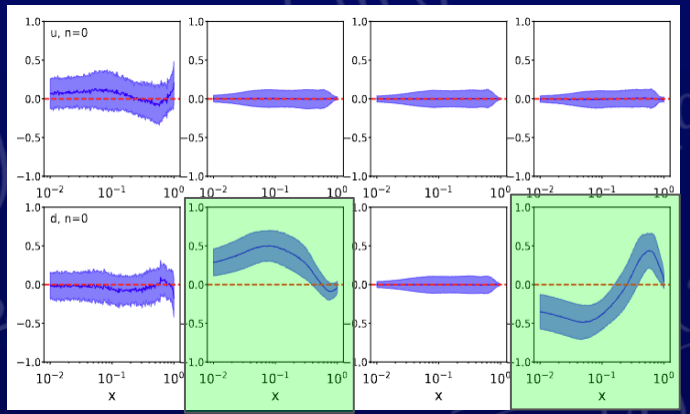
Generative AI for Inverse Problems

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



With a large latent space, information is too free to create spurious correlations between moments and PDFs – not physics.

By constraining the latent dimensions, squeezing the bottleneck, one can force the AI to generate physics-like properties.



BK, T.J. Hobbs **arXiv: 2312.02278**

Argonne NATIONAL LABORATORY

Explainability vs. interpretability

Explainability

Toolkit for providing human-readable reasons for a model's decisions or outputs.

Post-hoc tools are hooks in the model to inspect gradients, or specialized backpropagation tools, specifically for complex "black box" models.

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**

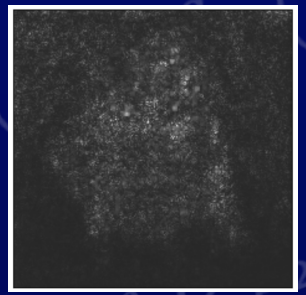
Argonne NATIONAL LABORATORY

Explainability … a fun example!

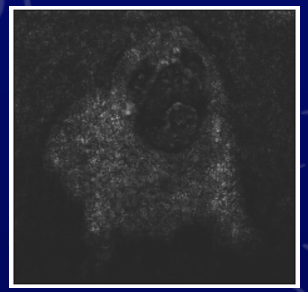
A survey of techniques

Input

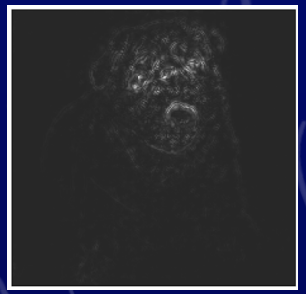
Gradients



Gradients ⊙ Input

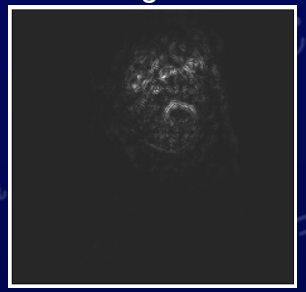


Guided Backprop



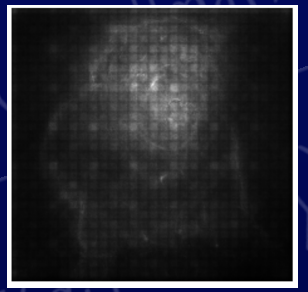
gradCAM

Guided gradCAM

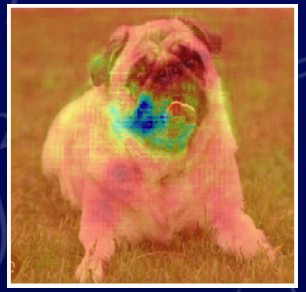


Integrated Gradients

smoothGrad



Occlusion



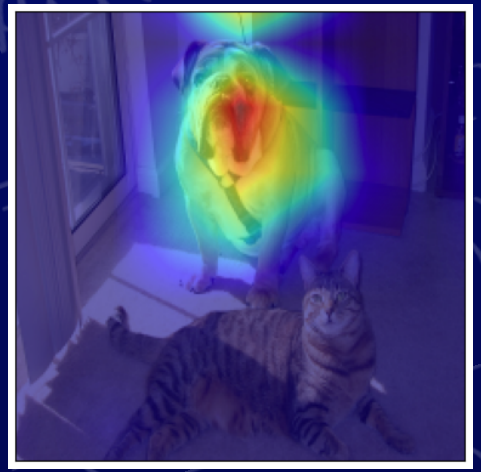
Edge Detection

Argonne NATIONAL LABORATORY

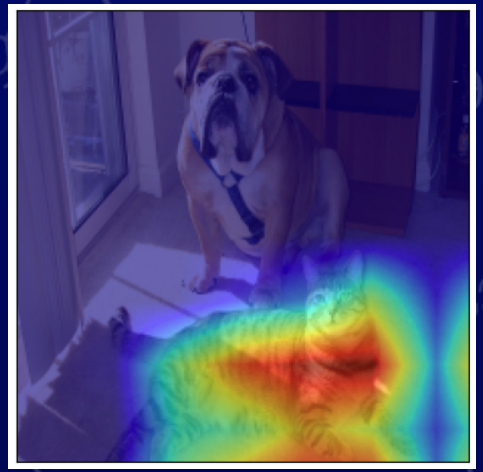
Picking out features from image with multiple possible labels

Original

Dog



Cat



Argonne NATIONAL LABORATORY

Guided backpropagation

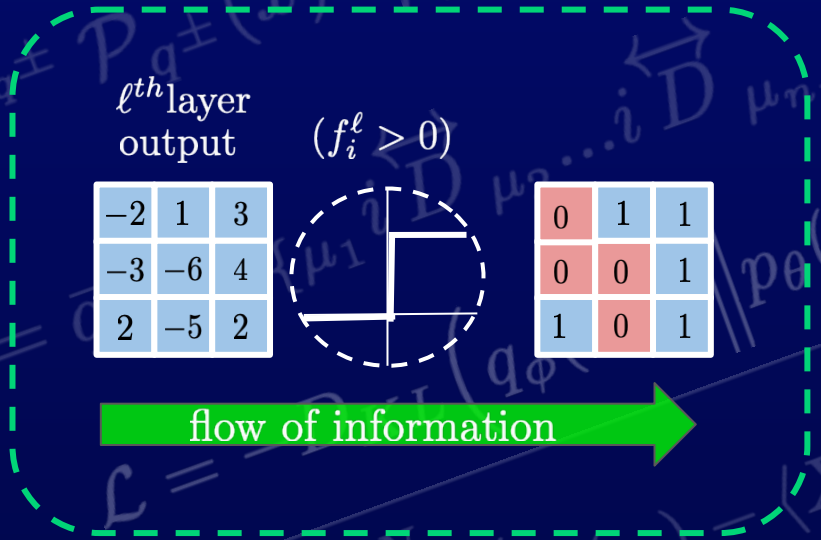
$$\frac{\partial f^{\text{out}}}{\partial f_i^{\ell}} = (f_i^{\ell} > 0) \cdot \left(\frac{\partial f^{\text{out}}}{\partial f_i^{\ell+1}} > 0 \right) \cdot \frac{\partial f^{\text{out}}}{\partial f_i^{\ell+1}}$$

Guided backprop is a “re-purposing” of the auto differentiation process in ML in which the gradients of a neural network layer are masked during a single backpropagation pass holding the weights fixed post-learning to determine which input features positively affect the classification outcome the most.

Simonyan et. al. **arXiv: 1312.6034**

Guided backpropagation

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



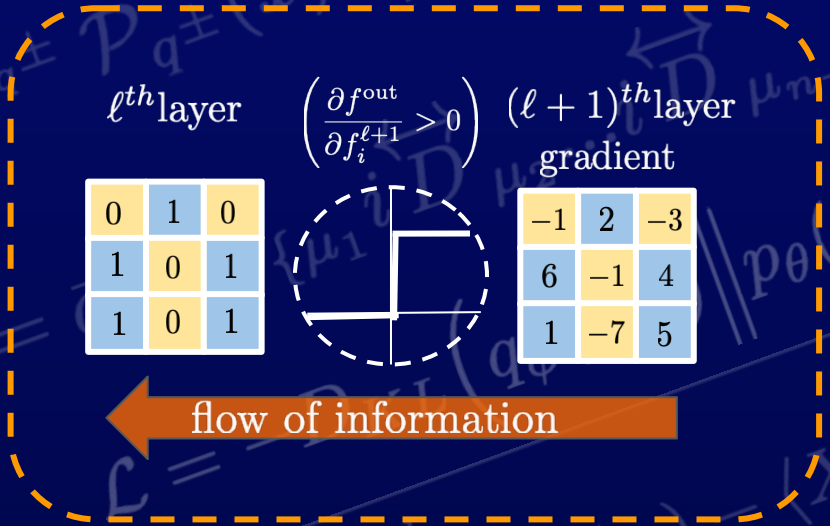
Forward Pass Logic Gate

Ensures only positive activations in the ℓ^{th} layer are considered when backpropagating the gradient. Prevents input from negative activations in the forward flow of information.

Argonne NATIONAL LABORATORY

Guided backpropagation

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



Backward Gradient Logic Gate

Ensures only positive gradients from the $(\ell+1)^{\text{th}}$ layer are considered when backpropagating the gradient. Prevents negative gradients in the backward flow of information.

Argonne NATIONAL LABORATORY

Guided backpropagation

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

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} -1 & 2 & -3 \\ 6 & -1 & 4 \\ 1 & -7 & 5 \end{bmatrix} = \begin{bmatrix} 0 & 2 & 0 \\ 0 & 0 & 4 \\ 1 & 0 & 5 \end{bmatrix}$$

The double-masking procedure during backpropagation generates highly detailed saliency maps, effectively highlighting fine-grained input features that most influence the network's output.

Argonne NATIONAL LABORATORY

Classifying PDFs from salient features

There are many open questions in phenomenological fitting of PDFs, many of which boil down to the open question of parameterization dependence: “How to effectively capture the associated effects of underlying theory assumptions on the fitted shape of the PDFs”

Model discrimination among classes of parton densities - a classification problem. We can therefore trace-back the classification score to the x-dependence of the PDF.

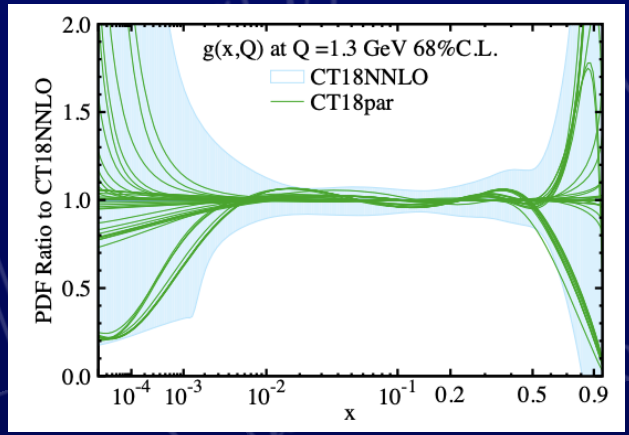


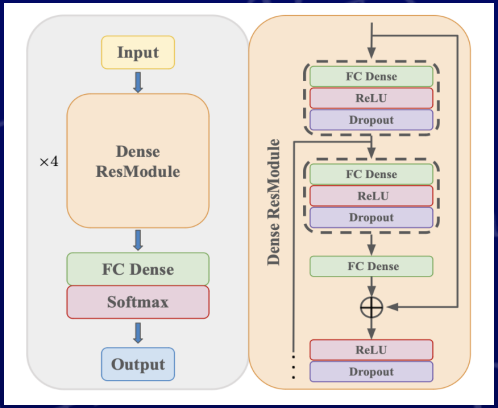
Image credit: PhysRevD.103.014013

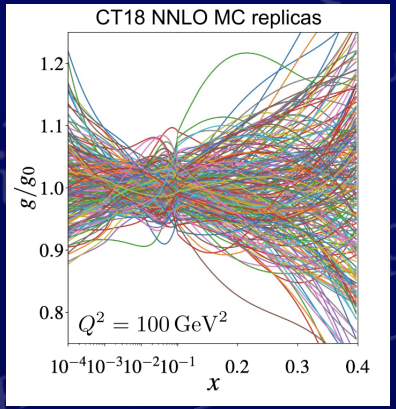
BK, J. Gomprecht, T.J. Hobbs **arXiv: 2407.03411**

Argonne NATIONAL LABORATORY

Classifying PDFs from salient features

Train a ResNet-like model on PDF MC replicas to identify salient features in x-dependence for classification tasks.





Proof-of-principle: what makes a particular flavor PDF unique amongst the others? A fingerprint to identify PDF flavors.

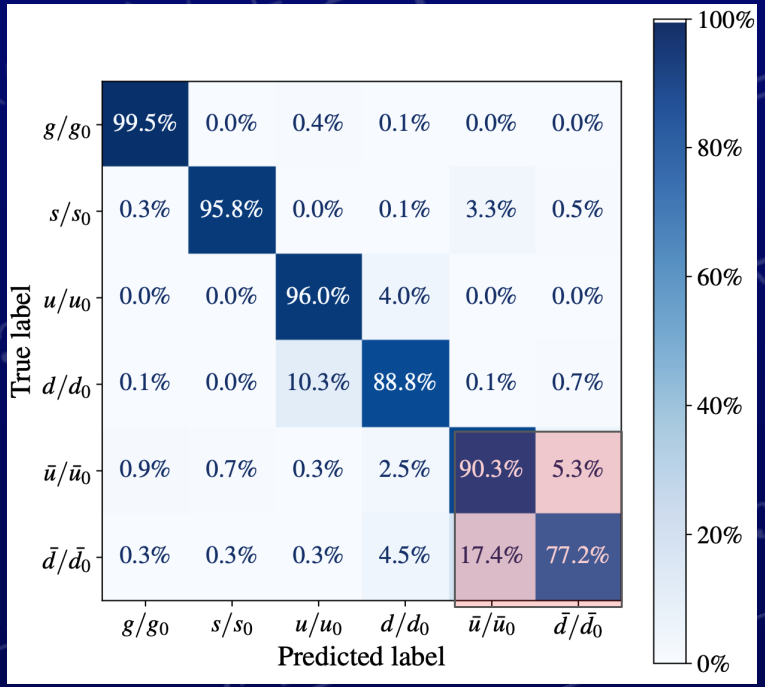
Fitting methodology: can we trace effects from the underlying theory back to the x-dependence of the PDF?

Regression: possible extensions to regression to correlate with L2 sensitivity studies.

BK, J. Gomprecht, T.J. Hobbs **arXiv: 2407.03411**

Argonne NATIONAL LABORATORY

Explainability within fitted PDFs



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

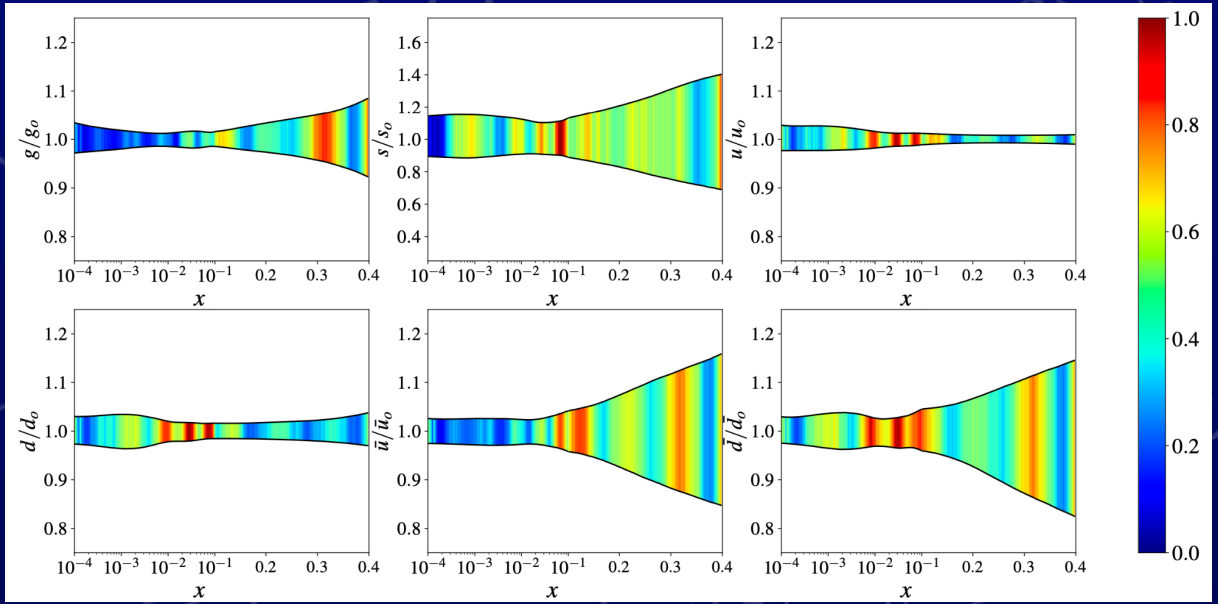
The confusion between the u-bar and d-bar ratios is related to flavor-separation challenges in phenomenological fits of the sea-quark densities;

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**

Argonne NATIONAL LABORATORY

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.

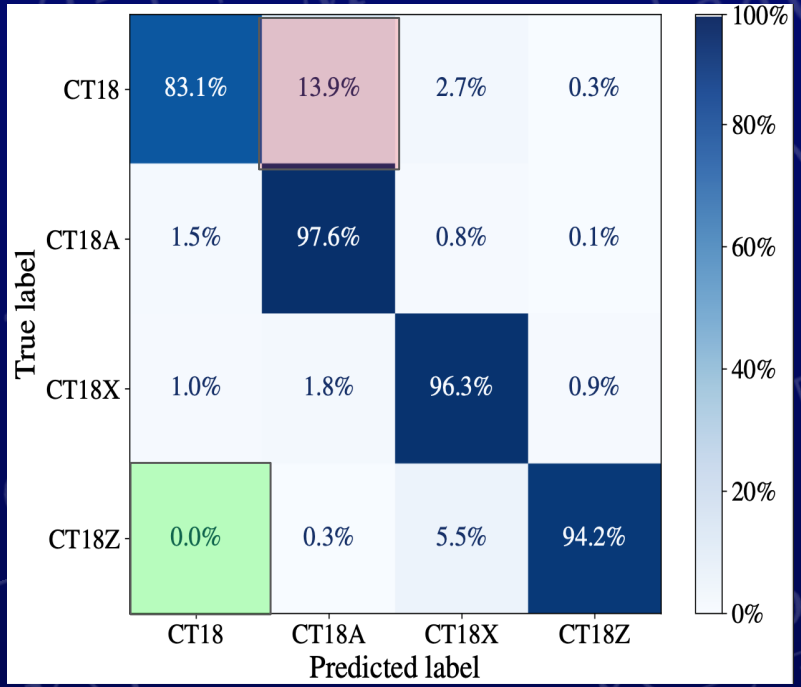
BK, J. Gomprecht, T.J. Hobbs **arXiv: 2407.03411**

Argonne NATIONAL LABORATORY

XAI4PDF: Explainability for 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

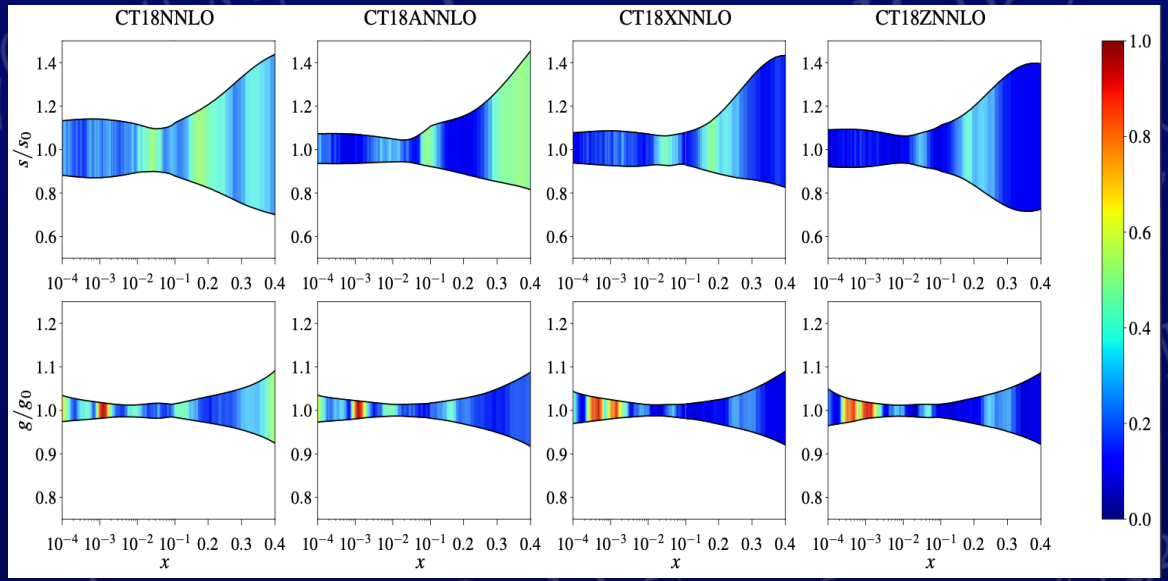
The two analyses which are “furthest” from each other (CT18 and CT18Z) are also the least confused, confirming that the shift in theory assumptions drives the statistical distinguishability as inferred by the XAI calculation.



BK, J. Gomprecht, T.J. Hobbs **arXiv: 2407.03411**

Argonne NATIONAL LABORATORY

XAI4PDF: Explainability for fitted PDFs



The strange and gluon PDFs demonstrate greater salience in discriminating among various CT alternative PDF fits."

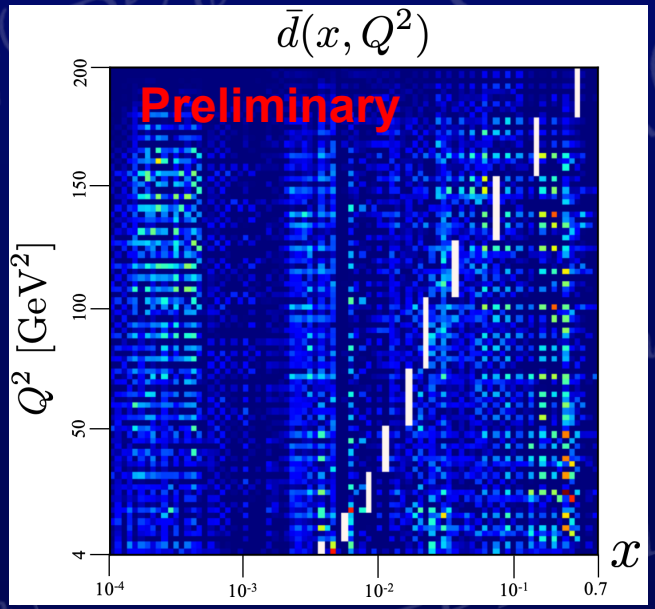
The gluon replicas have a dominant role in the classification among the CT18 series with highly localized gradients.

The strange replicas have smoother gradients indicating a weaker role.

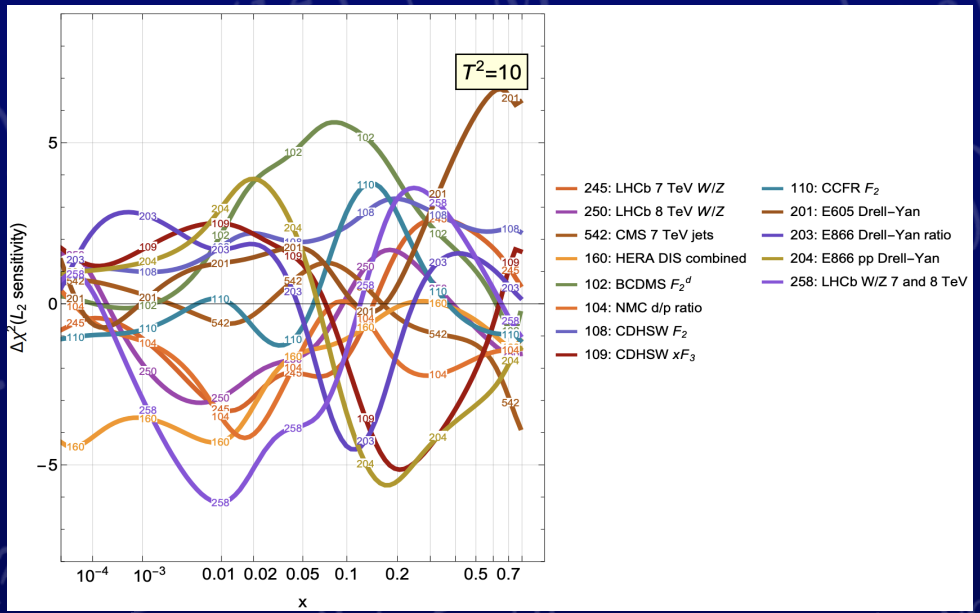
BK, J. Gomprecht, T.J. Hobbs **arXiv: 2407.03411**

Argonne NATIONAL LABORATORY

XAI4PDF: Explainable AI for PDFs



BK, T.J. Hobbs **(arXiv:2410.XXXXX)**



χ^2 on the CDHSW F_2 data (neutrino-iron CC DIS) is related to the PDF behaviors at specific values of x, Q^2; can be quantified via L_2 sensitivities (arXiv: 2306.03918).

Argonne NATIONAL LABORATORY

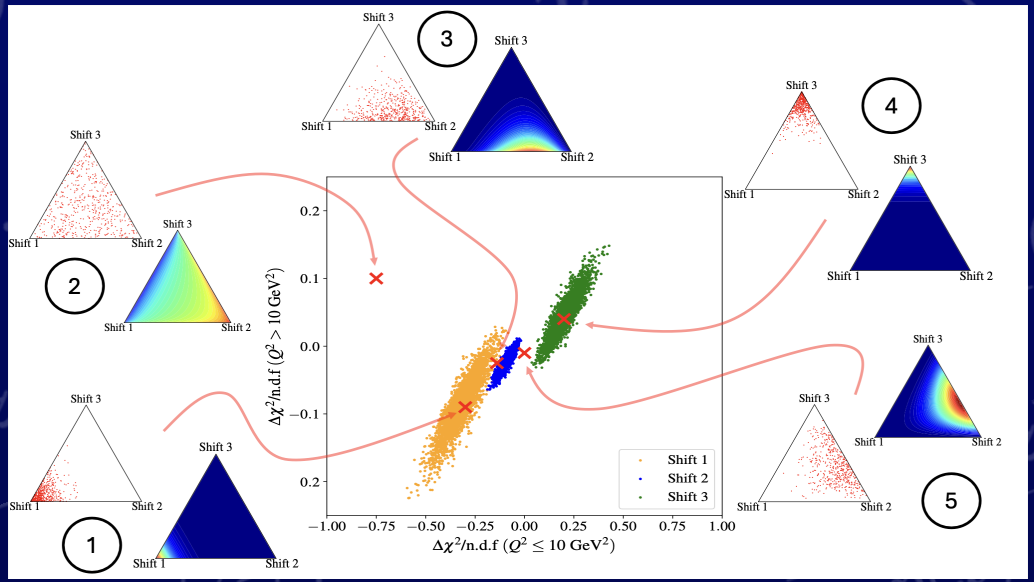
Uncertainty Quantification with prior networks

Mapping parametric theory space of BSM models

We construct mock BSM scenarios corresponding to non-standard neutrino interactions by shifting CKM matrix elements in CC ν-DIS cross section.

Construct latent space and use UQ metrics to study how these models overlap.

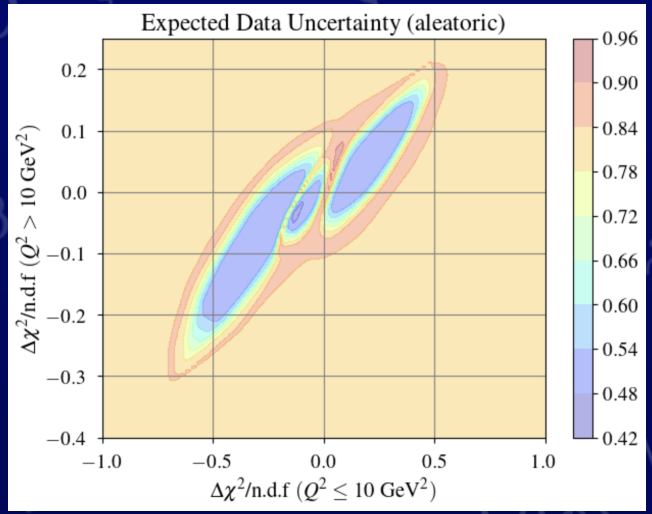
Dirichlet prior networks and classification for model discrimination.

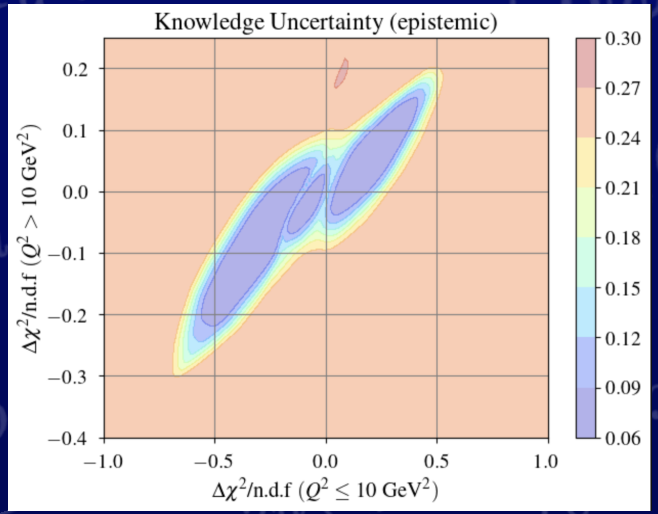


BK, T.J. Hobbs **(arXiv:2410.XXXXX)**

Argonne NATIONAL LABORATORY

Uncertainty Quantification with prior networks





We can measure how models overlap with aleatoric uncertainty.

… as well as quantify the boundaries of extrapolation into out of distribution sampling.

BK, T.J. Hobbs **(arXiv:2410.XXXXX)**

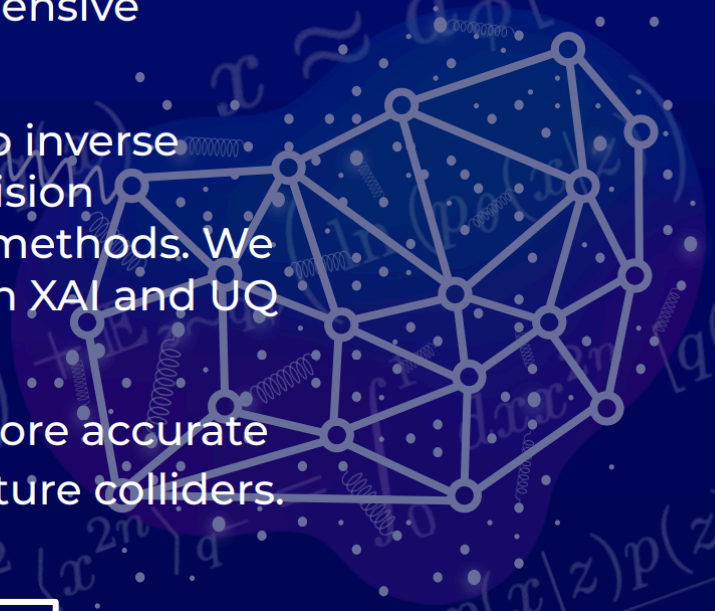
Argonne NATIONAL LABORATORY

Conclusions and Future Work

The research I have discussed here is the nucleus of a wide-reaching program culminating in comprehensive phenomenological fits with new physics.

Generative AI offers a transformative approach to inverse problem solutions, driving the next wave of precision phenomenology by pushing beyond traditional methods. We need to understand how these models work with XAI and UQ methods.

Goal: Frontier discoveries for BSM physics and more accurate predictions for high impact measurements at future colliders.



This work at Argonne National Laboratory was supported by the U.S. Department of Energy under contract DE-AC02-06CH11357.