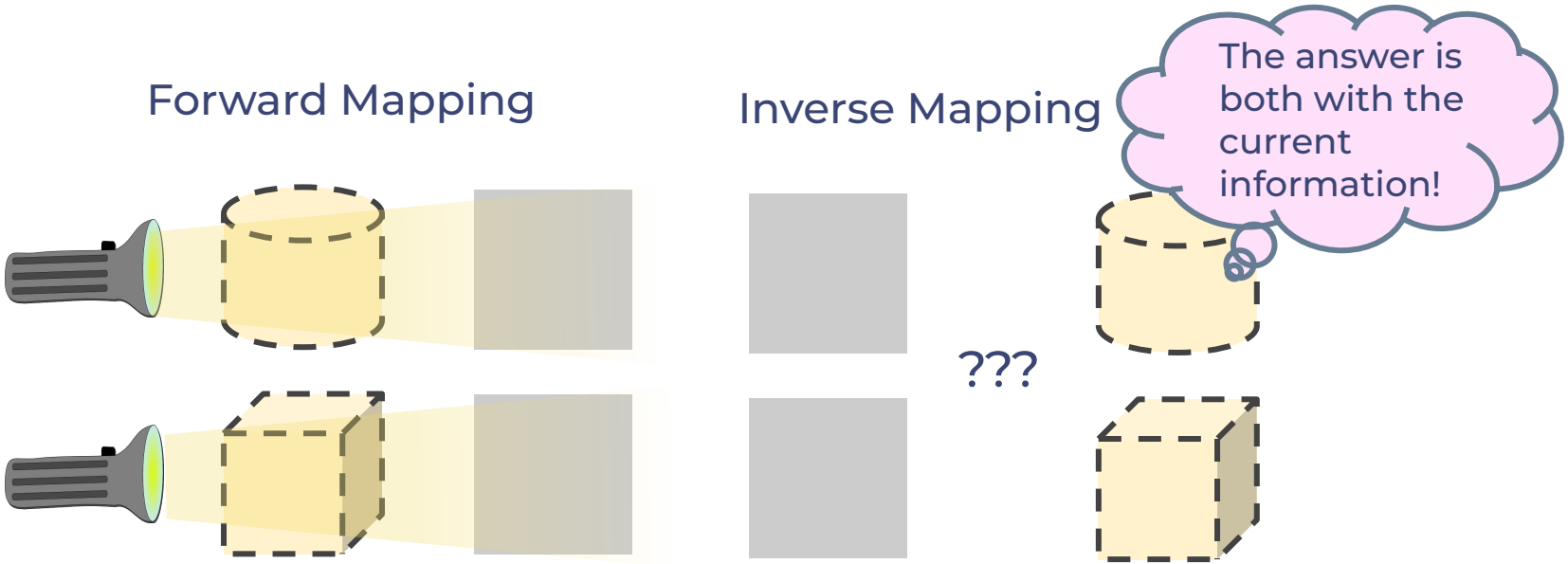
Argonne NATIONAL LABORATORY

AI for Global Fitting

B. Kriesten

Global fitting is an ill-posed inverse problem

In global fitting, the forward mapping from matrix elements to experimental observables is well-defined. The inverse problem is generally ill-defined: multiple parameter configurations can be compatible with the available data. Identifying these multiple solutions (flat directions), and ensembling them into **a definition of uncertainty**.



Translating theory into AI-enabled discovery

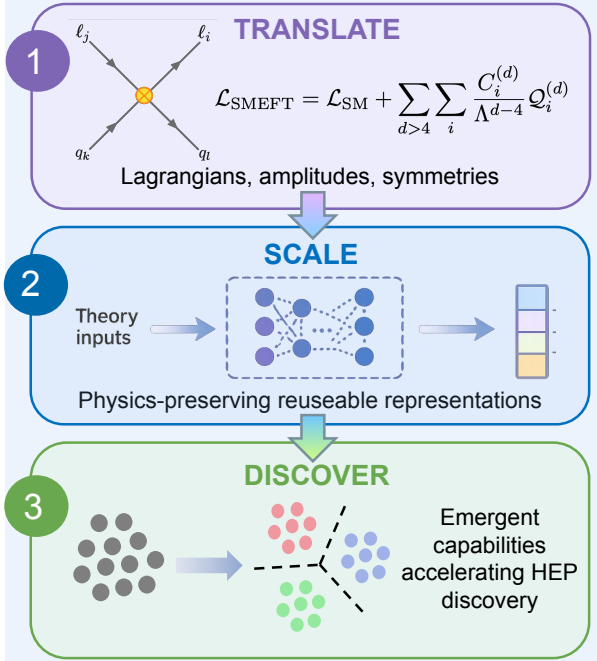
We are seeing the emergence of a new science, one based in the design and validation of physics-preserving representation layers for AI/ML integration.

Theory	AI/ML
• Amplitudes	• Arrays/tensors
• Symmetries	• Floats
• Analytic equations	• Graphs
• Fitting assumptions	• Tokens
• Simulations	• Distributions
• Matrix elements	• Metadata

Choices in representation can drastically affect what information the model can/cannot learn.

The scientific interface

Translate physics structures into machine-learnable representations that foundation models can ingest, preserving the underlying science.



AI/ML workflows: task-specific → global learning

Task-specific models probe particular physics mechanisms; agentic systems / foundation models scale to discover emergent learning behavior.

Task-specific: precision predictions → **Scaling: emergent behavior**

Bespoke Models

Inverse inference	arXiv: 26XX.XXXX arXiv: 2507.17810 arXiv: 2312.02278
UQ + Anomaly	arXiv: 2412.16286
Interpretability	arXiv: 2407.03411
BSM Matching	arXiv: 26XX.XXXX

Foundation Models

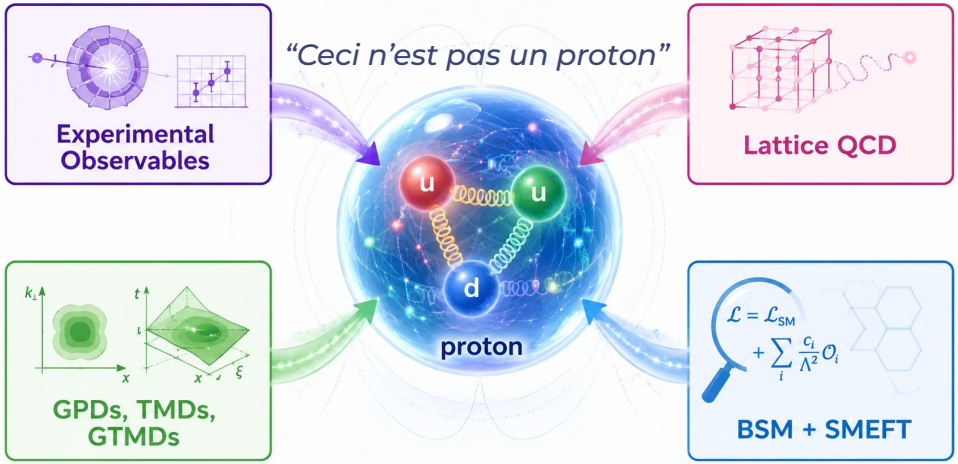
SMEFT	arXiv: 2512.15862
PDF-MORPH	arXiv: 2610.XXXXX

Agentic Workflows

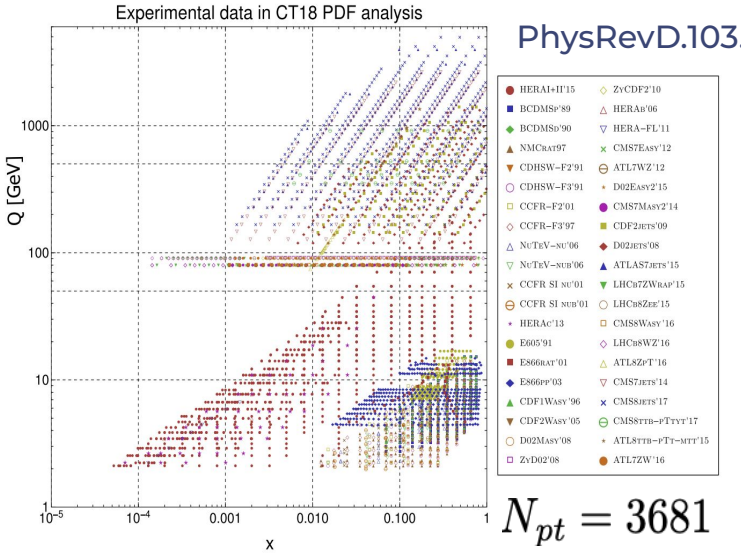
ArgoLOOM	arXiv: 2510.02426
PDFAgent	arXiv: 26XX.XXXXX

Hadronic structure lives in a multimodal ecosystem

Modalities: are distinct ways of representing information about the same underlying system.



Distinct views of the same underlying proton



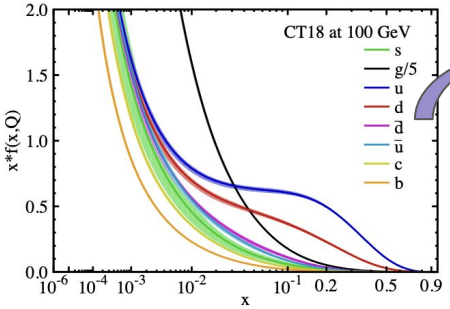
PhysRevD.103.014013

$N_{pt} = 3681$

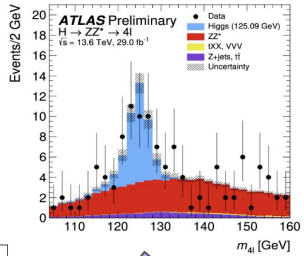
Global PDF fits unify data across all sectors of the SM, building a robust picture of hadron structure and effectively mediating our knowledge / or lack thereof of fundamental interactions.

PDFs (QCFs) + Lattice + SMEFT/BSM all belong together

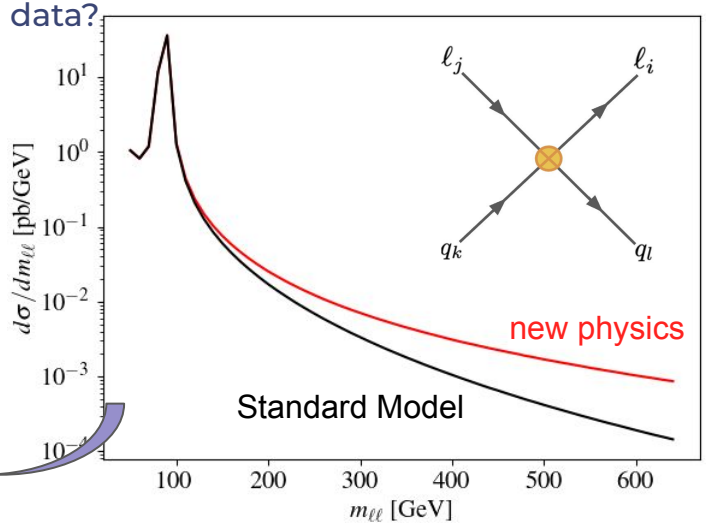
PDFs determine partonic luminosities at hadron colliders and govern baseline LHC predictions and BSM constraints.



$gg \to H$

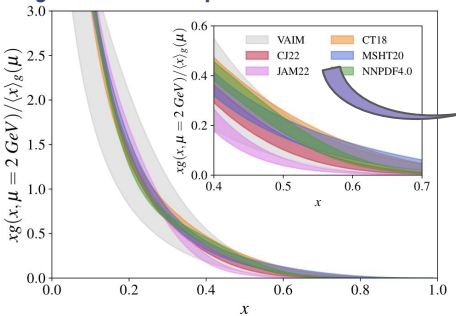


BSM effects in experimental observables can be absorbed into PDF uncertainties … how much do we trust that we can assume SM-only in data?



$$\mathcal{L}_{\rm SMEFT} = \mathcal{L}_{\rm SM} + \sum_{d>4}\sum_{i}\frac{C_i^{(d)}}{\Lambda^{d-4}}\mathcal{Q}_i^{(d)}$$

Recent developments in LQCD provide complementary views of partonic densities.

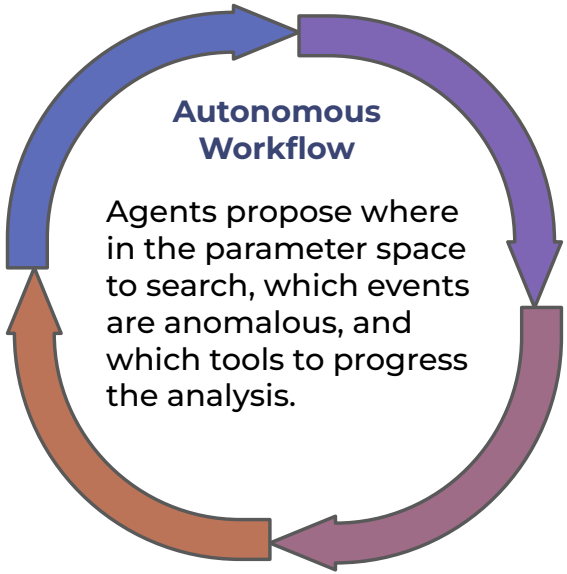


A universal representation must attempt to capture (and explain) everything.

The big picture: autonomous workflows for discovery

Theory simulations of hard scatter, collider pheno, PDFs/structure functions, Mellin moments, lattice QCD

Data/Tools



Foundation Models

Semantic spaces encoding physics, shared representation across potentially many physics domains.

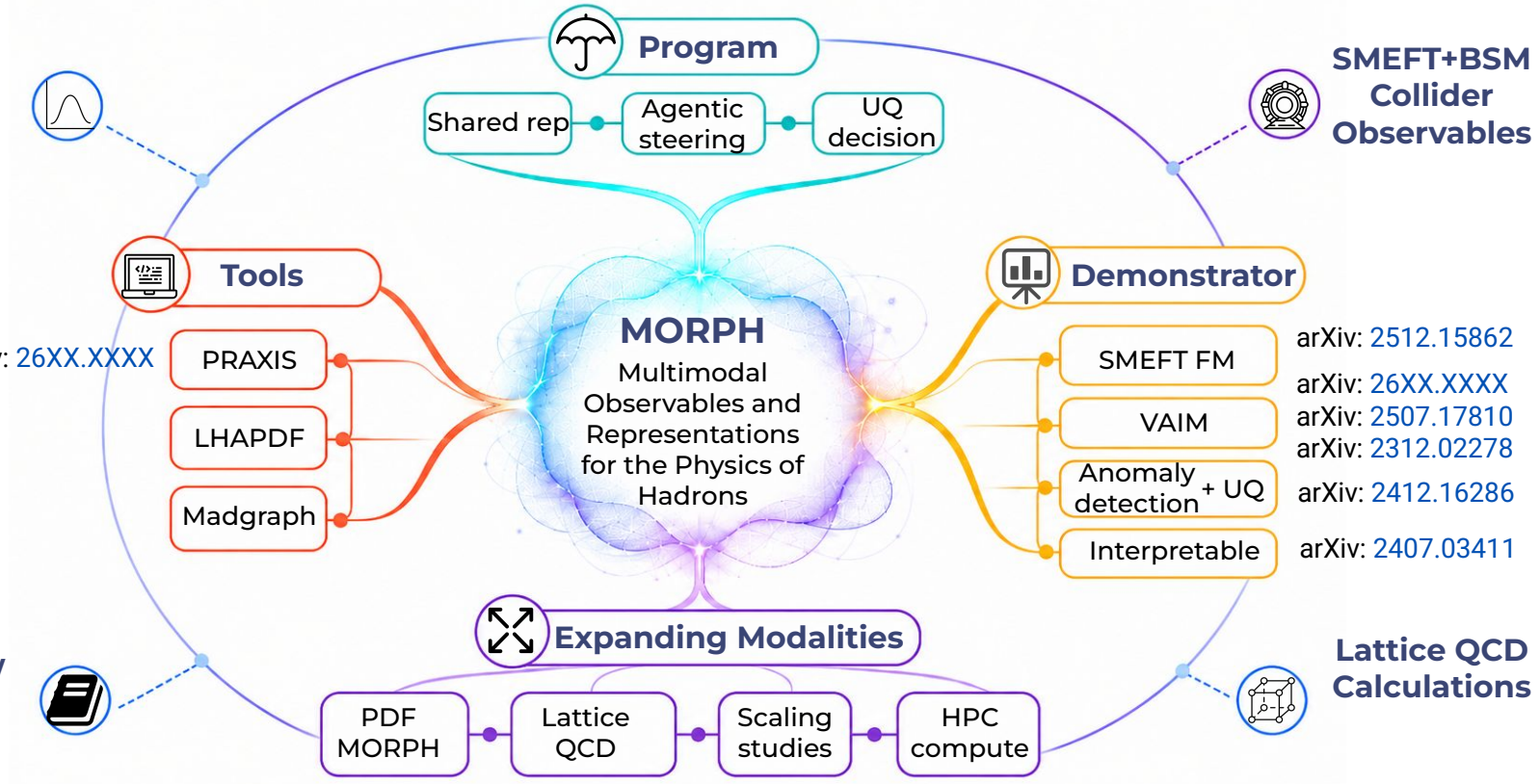
Inverse mapping, event generation, classification, anomaly detection, event retrieval, active decision making.

Downstream Tasks

Uncertainty Quantification

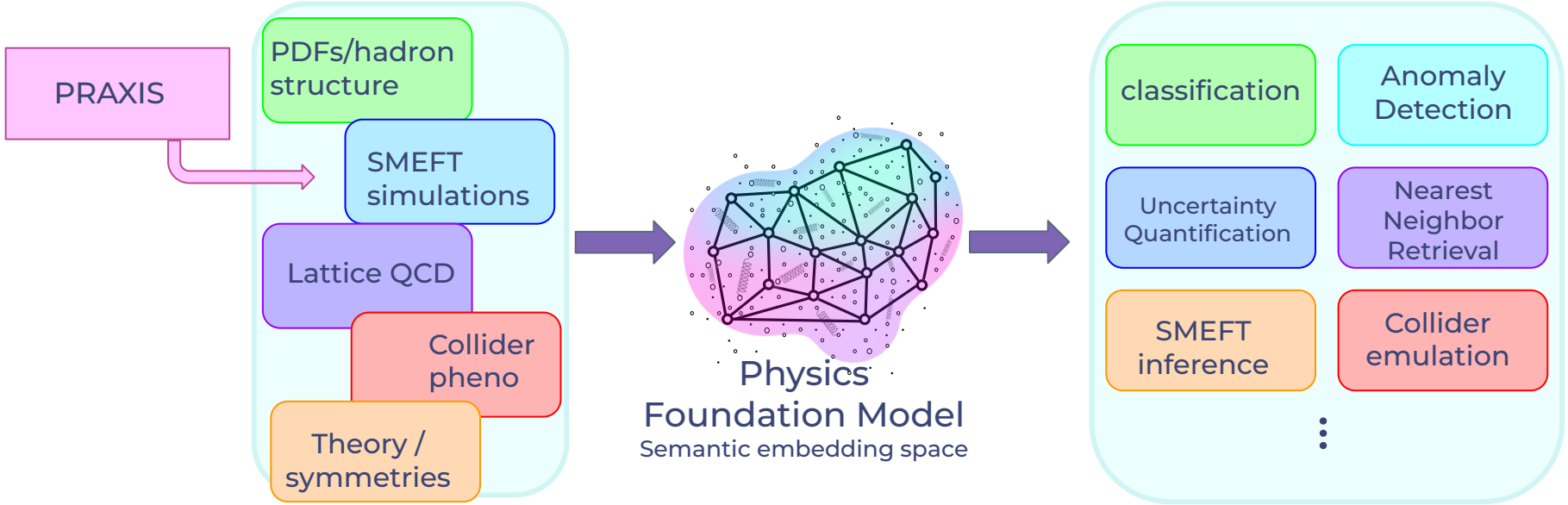
Replace single answer with a distribution over many solutions … decomposition of uncertainty between data and the modeling choice.

Agent-steered multimodal physics



Foundation models: the science representation layer

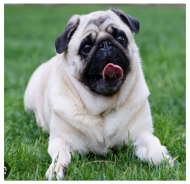
The foundation model is a shared representation layer compressing physics distributions into semantic embeddings that can be reused across downstream tasks and potentially steered by agents.

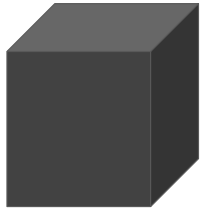


The representation tells us where an input lives; **uncertainty** tells us how much we should trust that location and the conclusions drawn from it. A reusable representation is useful only if it knows where its knowledge ends.

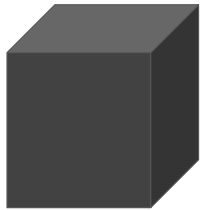
When should we trust the learned representation?

Traditional Classification

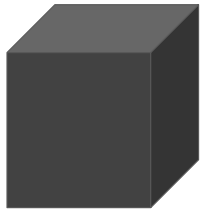




Dog: 98 %
Cat: 1.8 %
Bird: 0.2 %



Dog: 23 %
Cat: 65 %
Bird: 12 %



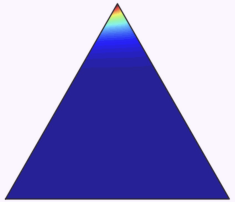
Dog: 0.3 %
Cat: 0.2 %
Bird: 0.5 %
Idk?: 99%

Malinin and Gales arXiv:1802.10501

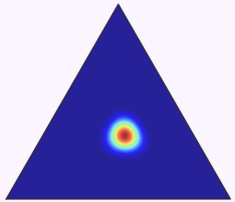
Dirichlet Prior Networks

Teaching an ML algorithm to say "I don't know" ...

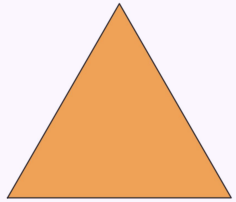
$$p(\mu;\alpha) = \frac{\Gamma(\alpha_o)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K} p_k^{\alpha_k - 1}$$



Confident



Uncertain



Anomalous

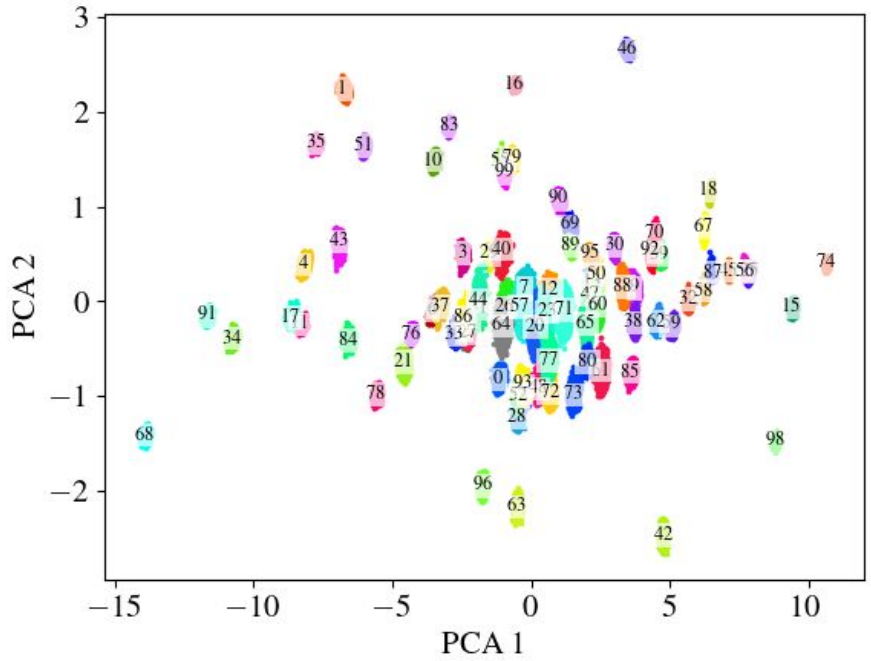
SMEFT demonstrator: learning a semantic space of BSM signatures

We generate collider-distribution samples from many SMEFT Wilson coefficient-configurations and train an encoder with a contrastive objective function. Each "blob" is a collection of 10,000 samples per configuration.

The learned latent space organizes the SMEFT universes by phenomenological similarity – nearby points have similar distributional signatures / shapes, while separated regions correspond to distinguishable new physics effects.

The embedding learns distance in **observable-distribution space**, not in Wilson-coefficient space.

Once learned, the same latent space supports plug-and-play downstream heads: classification, anomaly detection, nearest neighbor retrieval, and Wilson coefficient inference.

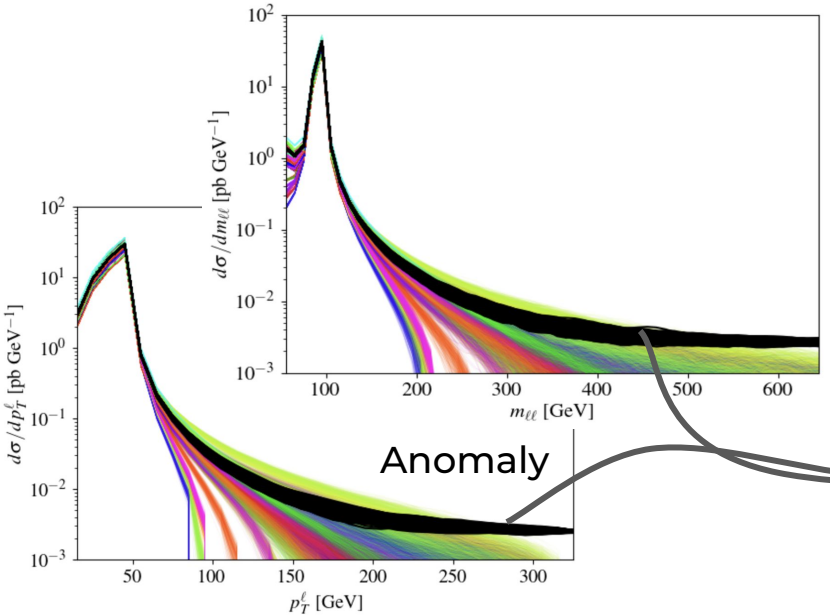


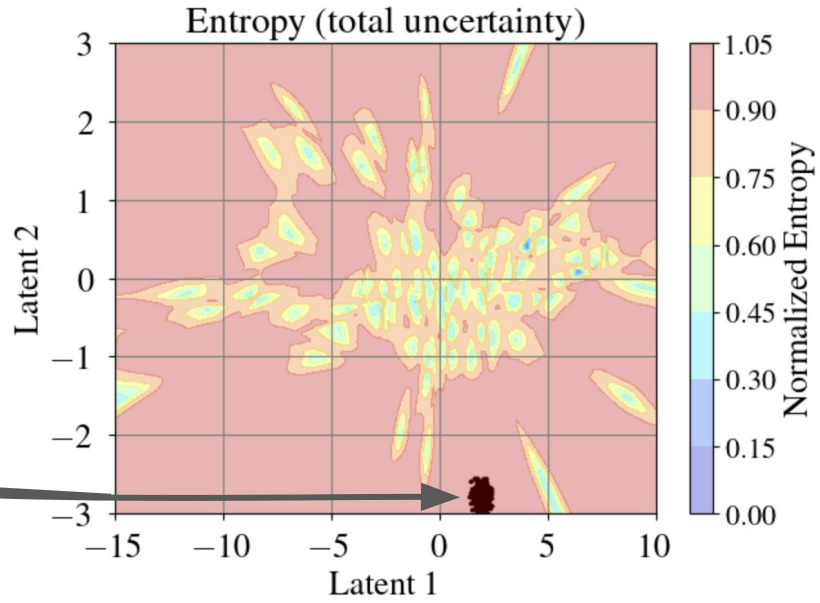
S. Das Bakshi, T.J. Hobbs, BK 2512.15064

Classifying distributions + flagging anomalies

A classifier head assigns samples to known SMEFT universes, while the entropy of the predictive distribut identifies regions where the model is uncertain, including OOD or anomalous signatures.

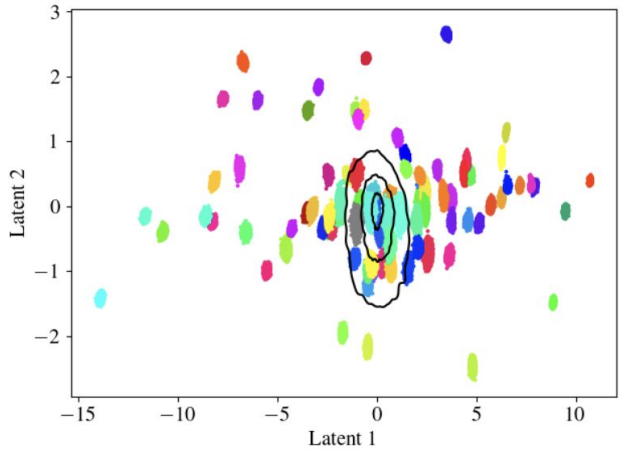
The embedding contains information on the model's predictive uncertainty, teaching the model to say "I don't know".

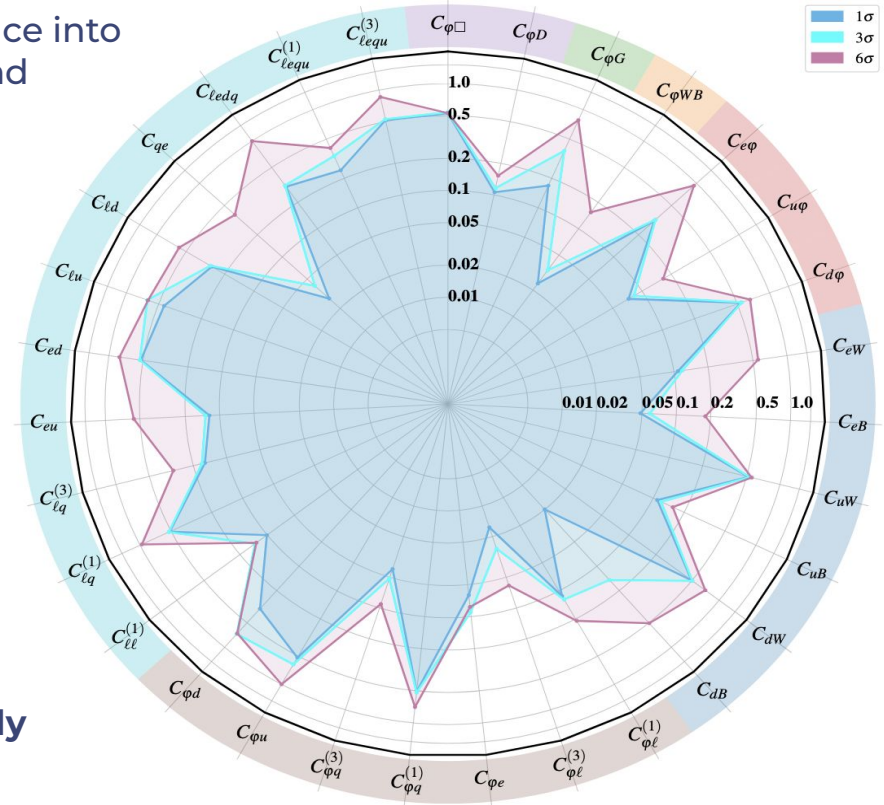




Toward Wilson-coefficient inference from embedding retrieval

Nearest neighbor retrieval turns the embedding space into a filtering tool by identifying SM-like distributions and weakly constrained WC-directions.





Rapid collapse of the radial component indicates sensitivity to selected WC-directions. **Embedding geometry becomes a fast proxy for computationally heavy SMEFT parameter-space inference.**

PDF global fitting ecosystem

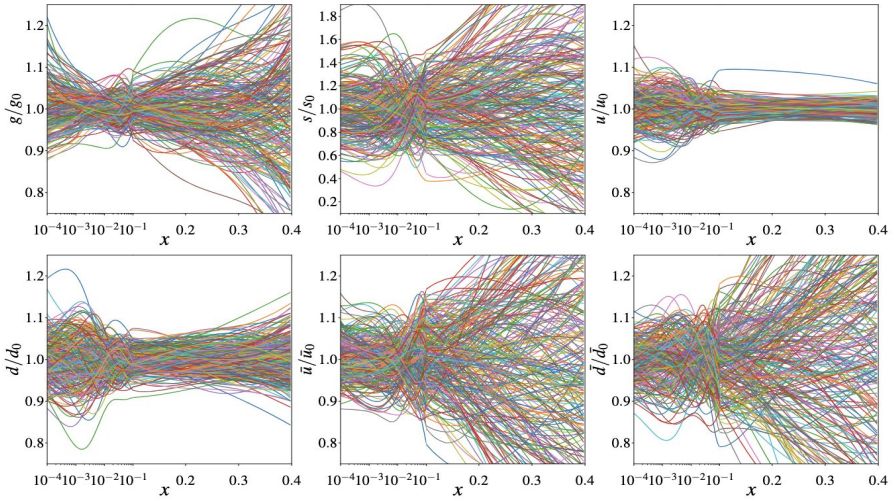
PDF global fits are expensive requiring large amounts of software, large amounts of compute, and must be repeated for different fitting assumptions.

Different global analyses encode different data selections, theory assumptions, parametrizations/analytic forms, and uncertainty prescriptions.

It is therefore a challenge to navigate this landscape … what does it mean to interpolate between fits with different fitting assumptions?

Can a single model learn the shared structure connecting this ecosystem? Can we then connect this landscape to the observable space bypassing (complementing) global fits.

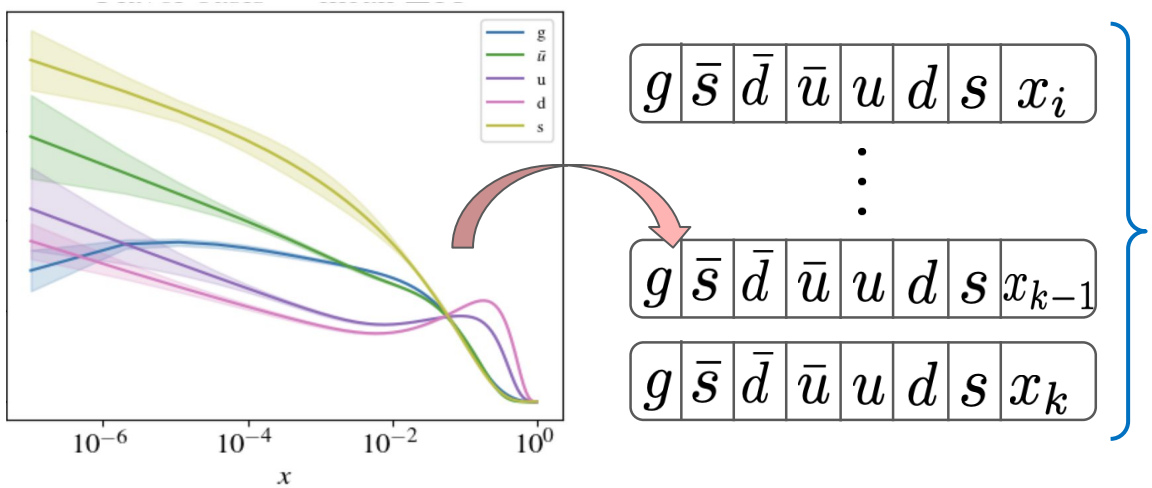
PDF fits	Factorization scale in DIS	ATLAS 7 TeV W/Z data included?	CDHSW $F_2^{p,d}$ data included?	Pole charm mass, GeV
CT18 NNLO	$\mu_{F,DIS}^2 = Q^2$	No	Yes	1.3
CT18A NNLO	$\mu_{F,DIS}^2 = Q^2$	Yes	Yes	1.3
CT18X NNLO	$\mu_{F,DIS}^2 = 0.8^2\left(Q^2 + \frac{0.3\ \mathrm{GeV}^2}{x_B^{0.3}}\right)$	No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



Building the input language — tokenization a PDF ensemble

A **token** is a machine-readable unit of information.

Tokenization decomposes a complex concept into a sequence of smaller units that a transformer can compare and attend across to learn correlations. What goes into the token must be chosen carefully!



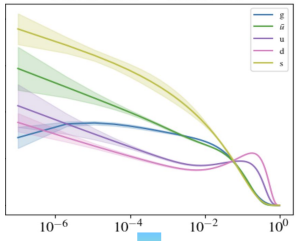
For a first demonstrator we compile **all flavor information** at a particular-x. A PDF replica is then represented as a sequence along x.

What information is placed in each token — and how that information is normalized, embedded, and conditioned — determines which physical structures the model can learn.

Learning a shared representation of global fits

1 Input tokens

PDF Ensembles



Discretized Tokens

$$\begin{matrix} g & \bar{s} & \bar{d} & \bar{u} & u & d & s & x_i \\ \vdots & & & & & & & \\ g & \bar{s} & \bar{d} & \bar{u} & u & d & s & x_{k-1} \\ g & \bar{s} & \bar{d} & \bar{u} & u & d & s & x_k \end{matrix}$$

Sequence $[N_x]$

$$X \in \mathbb{R}^{N_x \times (2+N_f)}$$

2 Transformer Encoder Stack

Embedding Projection + CLS token append $W_{in} : \mathbb{R}^{N_x \times (2+N_f)} \to \mathbb{R}^{(N_x+1) \times d_{model}}$

Transformer Block x4

Input $X \in \mathbb{R}^{(N_x+1) \times d_{model}}$

Layer Norm $X \in \mathbb{R}^{(N_x+1) \times d_{model}}$

W_Q W_K W_V

Distribute over heads $X \in \mathbb{R}^{(N_x+1) \times n_{heads} \times \frac{d_{model}}{n_{heads}}}$

Q K V

Scaled dot-product attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$X \in \mathbb{R}^{(N_x+1) \times n_{heads} \times \frac{d_{model}}{n_{heads}}}$

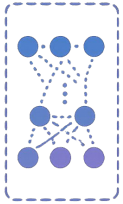
Concatenate + Project + Dropout $X \in \mathbb{R}^{(N_x+1) \times d_{model}}$

$X \in \mathbb{R}^{(N_x+1) \times d_{model}}$

Output

Dropout

FFN/MLP GELU Activation



Layer Norm

3 Latent

Encoder Hidden States

$X \in \mathbb{R}^{(N_x+1) \times d_{model}}$

Pool CLS/Attention

$X \in \mathbb{R}^{2d_{model}}$

Layer Norm

Dense

Latent Representation

$X \in \mathbb{R}^{d_z}$

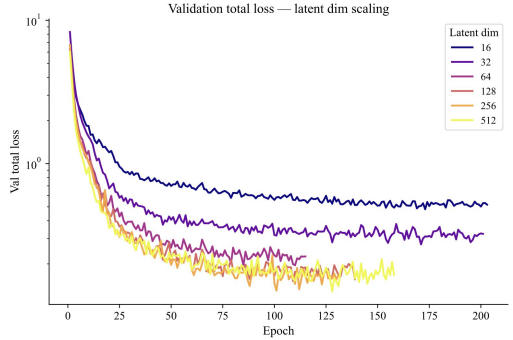
The latent representation is shared among all inputs and is constrained by the training heads.

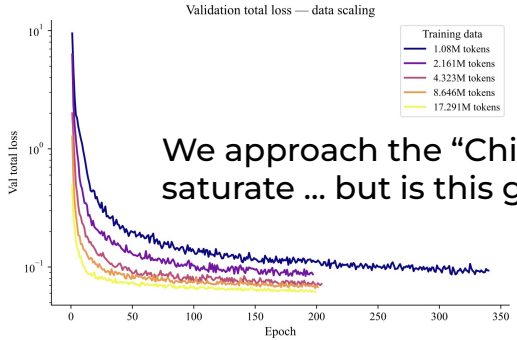
4 Heads

- Contrastive Alignment
- Classification
- Masked Flavor Modeling
- Reconstruction
- Denoising

Neural scaling locates the statistical performance frontier

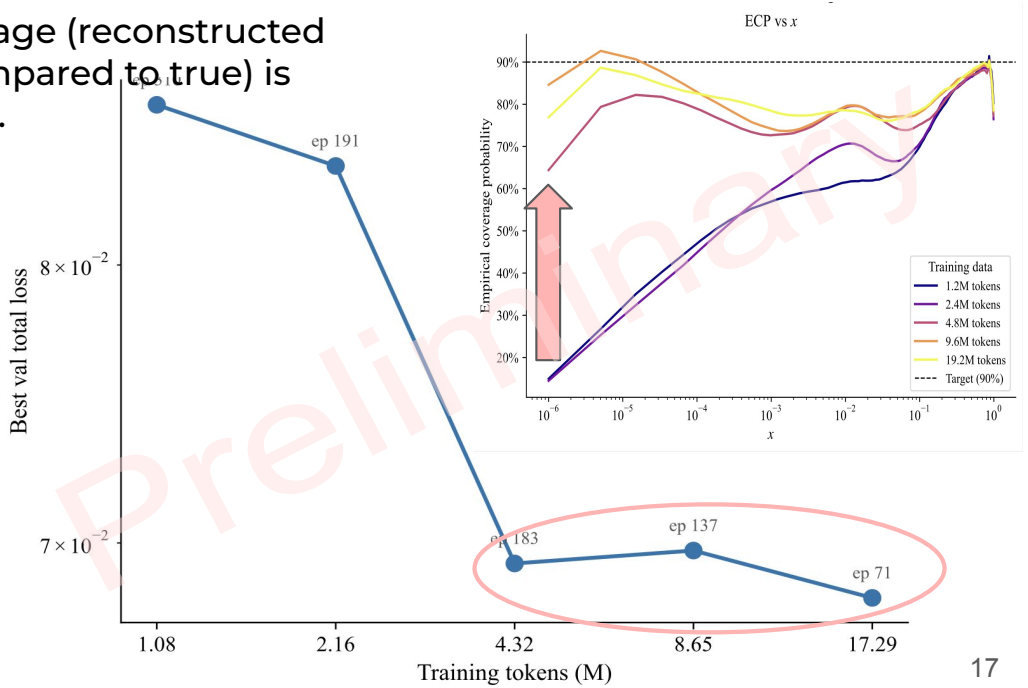
Neural scaling laws — mainly evaluated for NLP-based models — tell us about expected behavior of the model as a function of various components such as datasize, compute time, and model size.





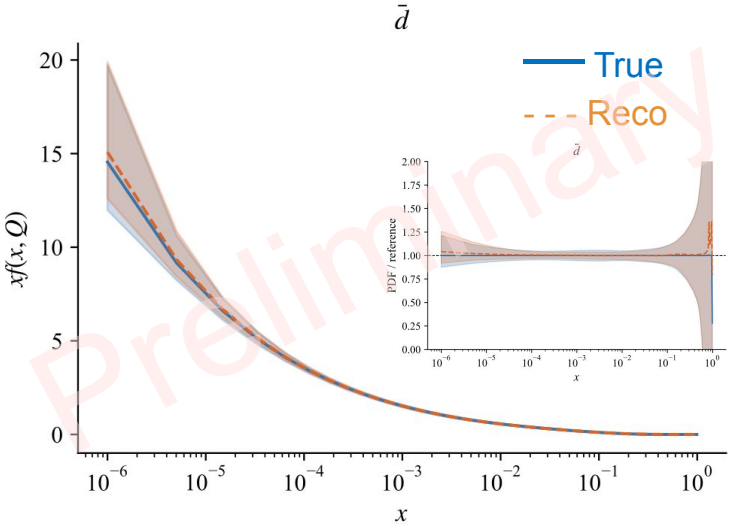
We approach the “Chinchilla limit” and saturate … but is this good enough?

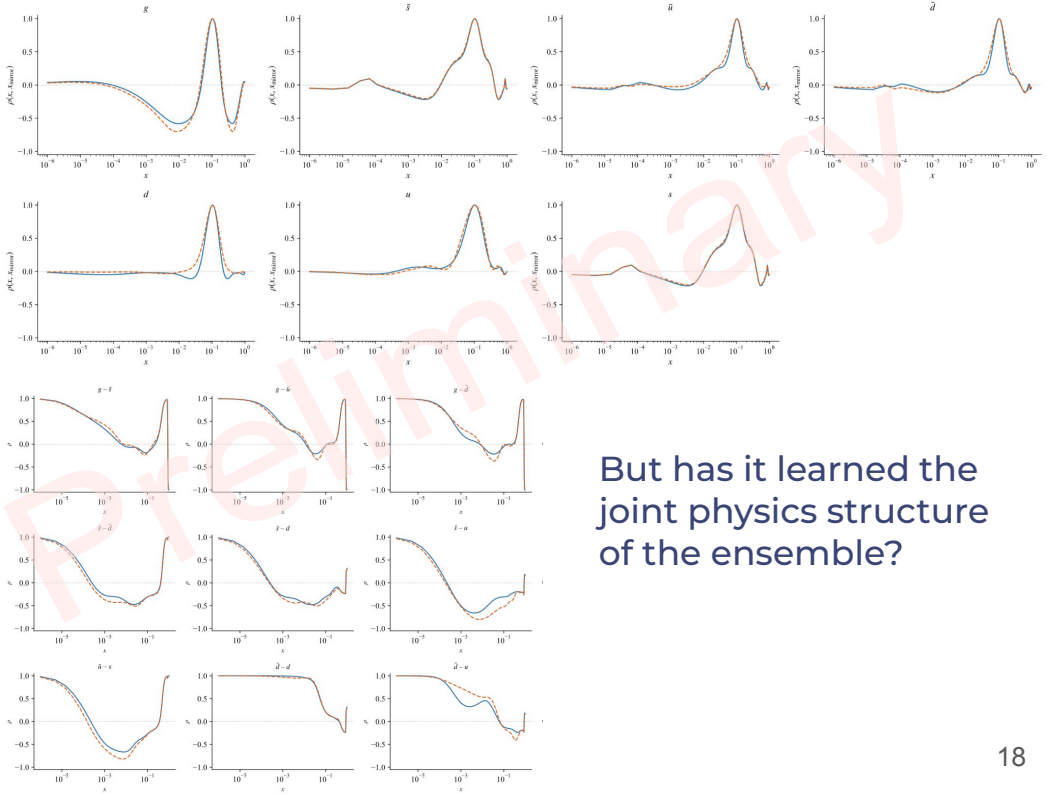
Expected coverage (reconstructed uncertainty compared to true) is learned at scale.



At scale, conventional metrics suggest success

If I just look at the reconstructed error, things look pretty good. In fact, when I look at basic PDF diagnostics they look great!

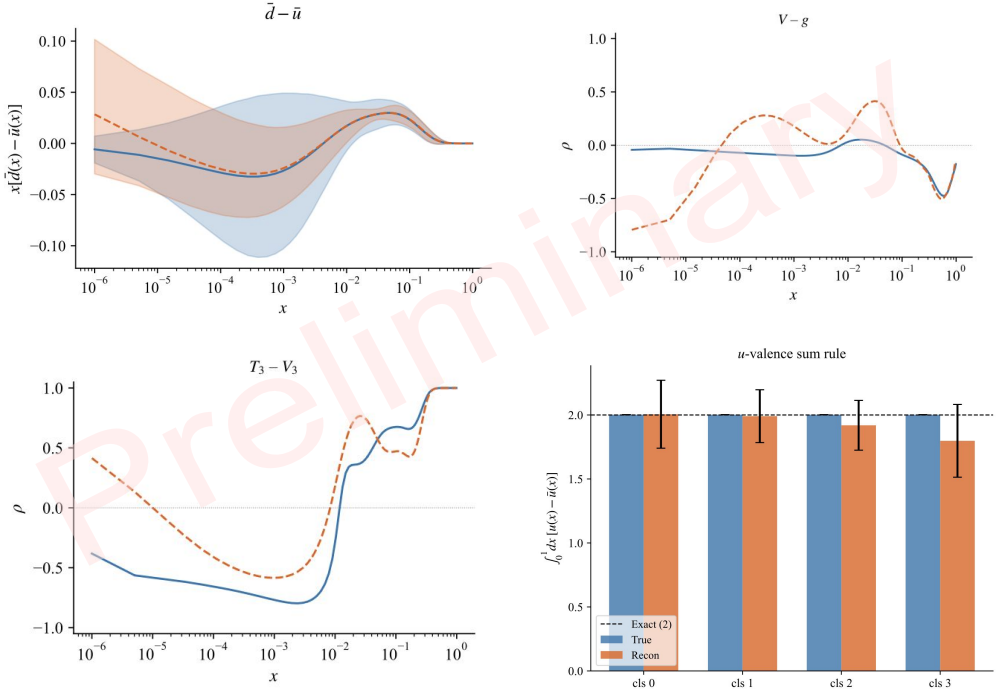




But has it learned the joint physics structure of the ensemble?

Physics benchmarks reveal a representation gap

The best scaled model reproduces the marginal PDFs but does not yet preserve all of the joint statistical structure carried by the replica ensemble.



The performance plateau may indicate a representation bottleneck rather than insufficient scale. This motivates richer tokenization, architectural changes, and more rigorous physics-based validation.

Scaling optimizes the loss we specify; physics benchmarks determine whether that loss captured the science we care about.

More work to be done!

Conclusions

Global fitting is an ill-posed, multi-modal inverse problem and we are tackling it with AI/ML.

Agents can organize and steer calculations; foundation models learn reusable representations across fitting choices and modalities.

Neural scaling is necessary, but physics-aware benchmarks determine whether the learned representation is scientifically valid and robust.

The next frontier is physics-aware scaling: more data, richer representations, advanced architectures, calibrated uncertainty, and objectives that preserve the joint structure required for global inference.

Argonne NATIONAL LABORATORY

Thank you for your attention!

This work was supported by the Argonne Laboratory Directed Research and Development (LDRD) project, MORPH: Multimodal Observables and Representations for the Physics of Hadrons. Argonne, a U.S. Department of Energy Office of Science Laboratory, is operated by UChicago Argonne LLC under contract no. DE-AC02-06CH11357.