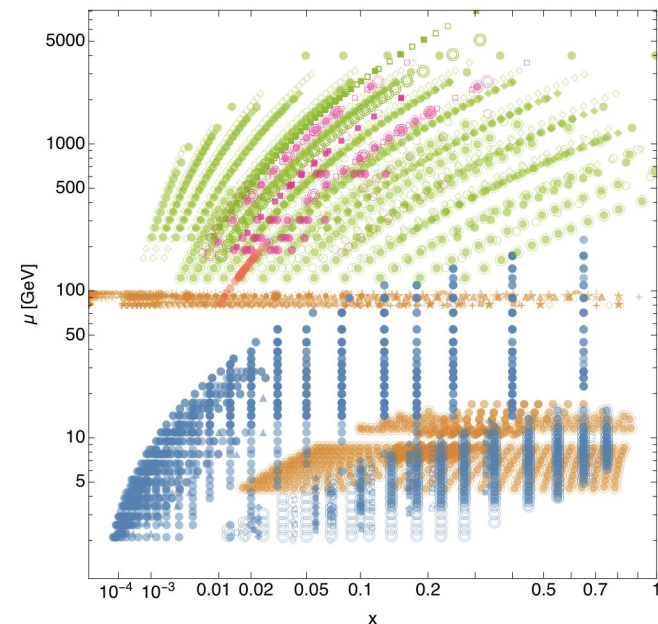


# MORPH: Multimodal Observables and Representations for the Physics of Hadrons

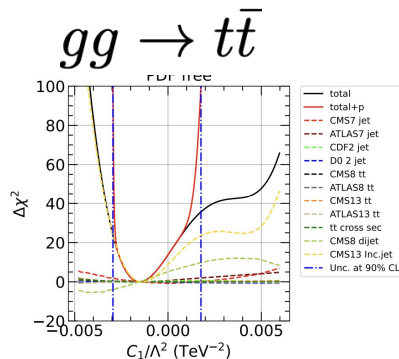
Brandon Kriesten · Argonne National Laboratory

QCD@LHC · 1 October 2026

## PDFs set the baseline for discovery

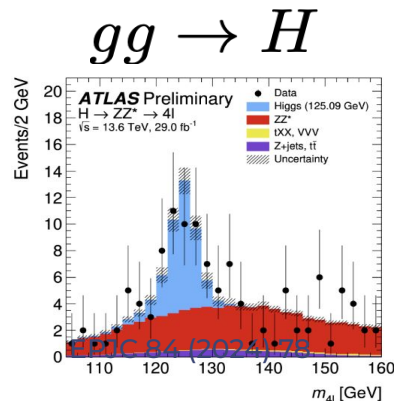


PDF fits unify thousands of measurements across all sectors of the SM, build a consistent picture of hadron structure and mediate our knowledge / or lack thereof of fundamental interactions.



JHEP 05 (2023) 003

**Uncertainty in proton structure propagates into collider observables and potentially what we call “new physics.”**

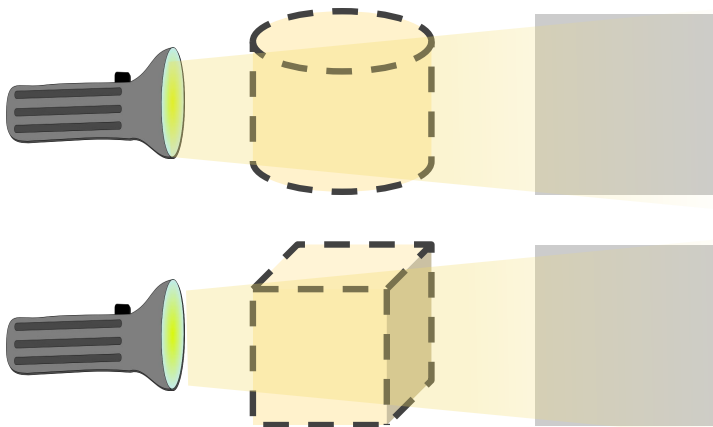


*“Why, sometimes I’ve believed as many as six [New Physics] things before breakfast.” – Lewis Carroll Through the Looking Glass*

## Global fitting is an ill-posed inverse problem

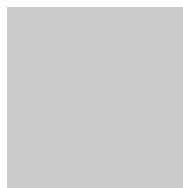
### Forward Mapping

proton structure  $\rightarrow$  observable

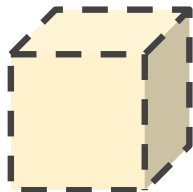


### Inverse Mapping

observable  $\rightarrow$  ???  $\rightarrow$  proton structure(s)



???



The data do not choose a unique answer!

The goal is not to find one answer, but to characterize the full space of answers consistent with the evidence.

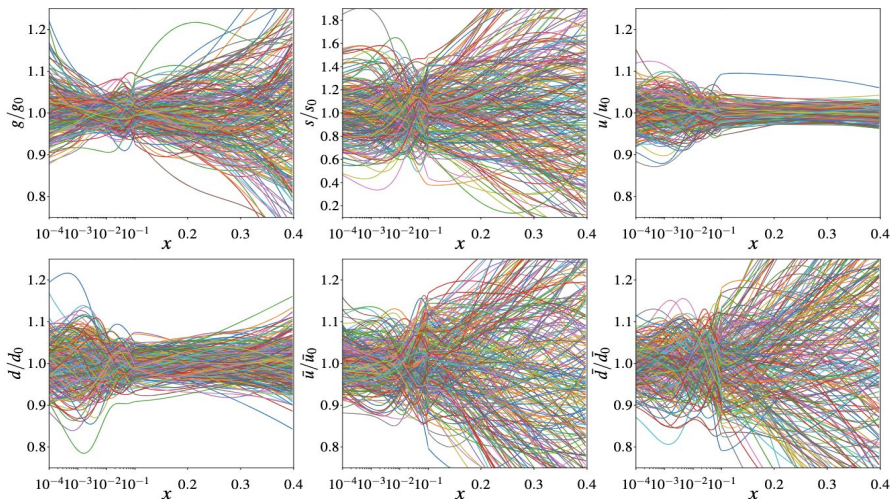
# PDF global fitting ecosystem

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

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

**Can a single model learn the shared structure connecting this ecosystem?**

| 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 | $\mu_{F,DIS}^2 = 0.8^2 \left( Q^2 + \frac{0.3 \text{ GeV}^2}{x_B^{0.3}} \right)$ | No                             | Yes                              | 1.3                  |
| CT18Z NNLO | $\mu_{F,DIS}^2 = 0.8^2 \left( Q^2 + \frac{0.3 \text{ GeV}^2}{x_B^{0.3}} \right)$ | Yes                            | No                               | 1.4                  |



## The information structure behind PDF inference

Every PDF fit is a sample of a large conditional probability distribution

$$p(f|D, T, M) \propto p(D|f, T) \cdot p(f|M)$$

**Data**  
**D** What constrains the PDF;  
 measurements, kinematics

**Theory**  
**T** The forward map; order,  
 evolution, lattice,  $\alpha_s$ ,  $m_c$

**Methodology**  
**M** Parameterization  
 dependence, uncertainty

Amanda Sarkar — Mon  
 Stephen Maple — Mon  
 James Ferrando — Mon  
 William Barter — Tue  
 Laura Fabbri — Tue  
 Cedric Verstege — Tue  
 ...

Lucian Harland-Lang — Mon  
 Chirara Signorile — Tue  
 Marina Krstic Marinkovic — Tue  
 $\alpha_s$  panel  
 ...

Stefano Forte — Mon  
 Kamil Laurent — Mon  
 Tim Hobbs — Mon  
 Robert Shoenbeck — Tue  
 ...

**Can one model learn a useful, reusable space encapsulating this probability distribution?**

## Foundation models for PDFs

A scientific foundation model learns a **shared representation** from a large training corpus of scientific data — in this case PDFs — so that the same learned space can support **many downstream tasks**.

### Training inputs

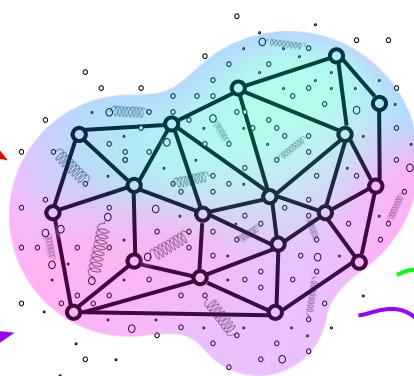
PDF replicas

Flavor structure

Theory metadata

Lattice QCD

Collider observables



PDF Foundation Model

### Downstream tasks

Organize PDF ecosystem

Uncertainty-aware generation

Condition on theory metadata

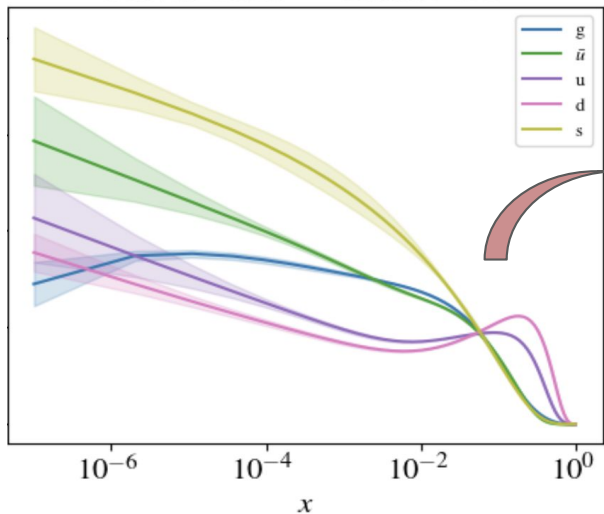
Anomaly Detection

Phenomenology derived quantities

PDF-MORPH - a first demonstrator foundation model representation for PDFs.

## Building the input language — tokenizing a PDF ensemble

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



$g \ \bar{s} \ \bar{d} \ \bar{u} \ u \ d \ s \ x_i$

$\vdots$

$g \ \bar{s} \ \bar{d} \ \bar{u} \ u \ d \ s \ x_{k-1}$

$g \ \bar{s} \ \bar{d} \ \bar{u} \ u \ d \ s \ x_k$

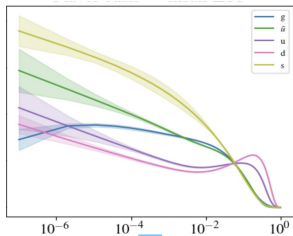
For a first demonstration we compile **all flavor information** at a particular- $x$ . A PDF replica is then represented as a sequence along  $x$ .

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

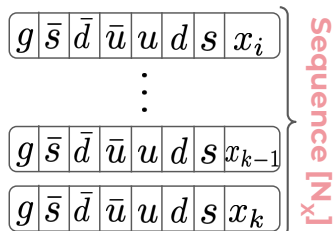
# Learning a shared representation of global fits

## 1 Input tokens

### PDF Ensembles



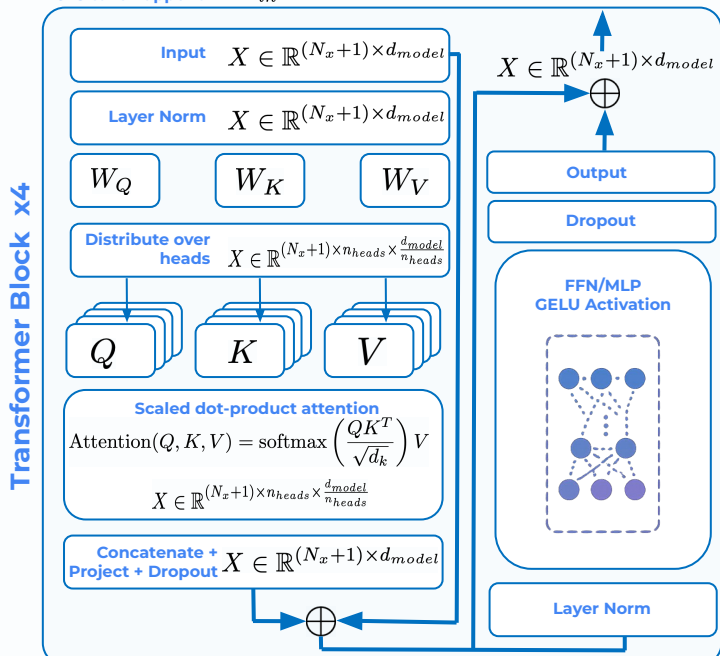
### Discretized Tokens



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

## 2 Transformer Encoder Stack

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



## 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

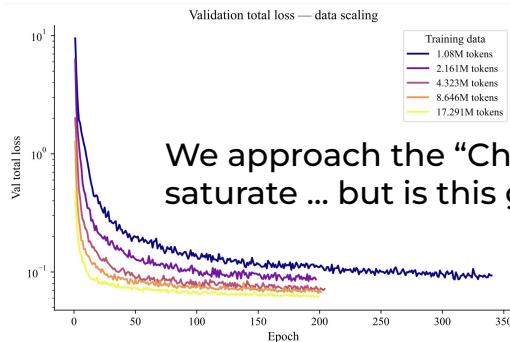
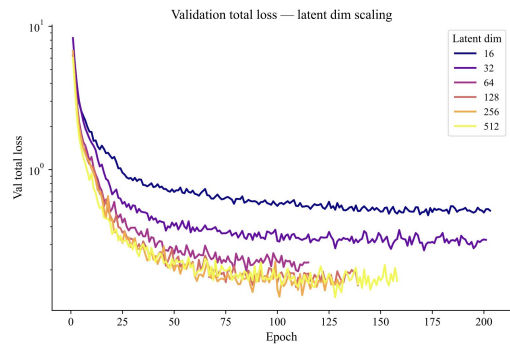
Masked Flavor Modeling

Reconstruction

Denoising

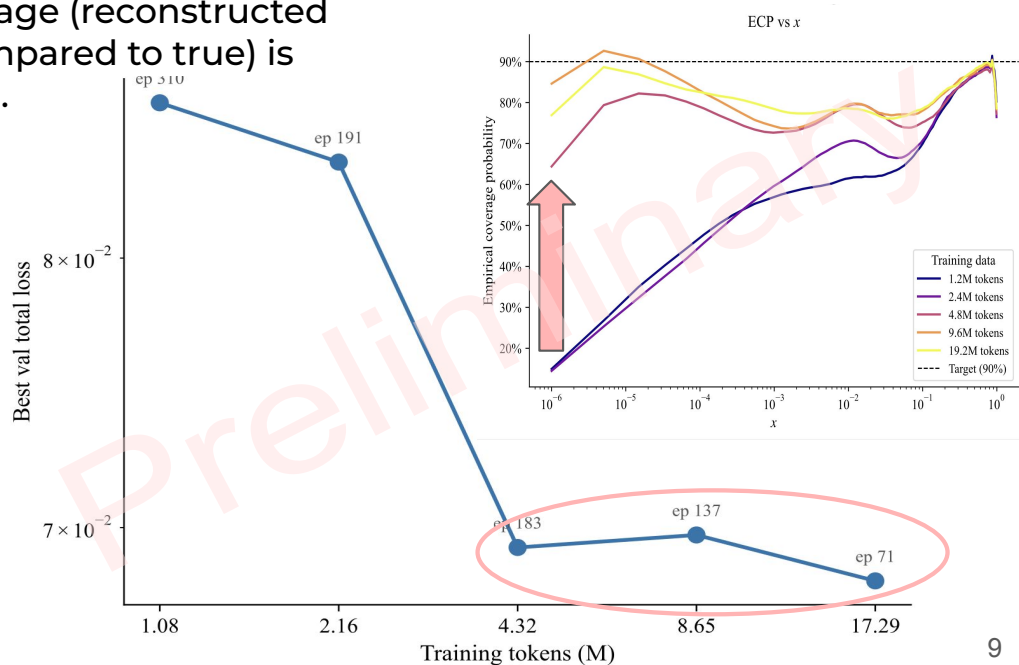
# Neural scaling locates the statistical performance frontier

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



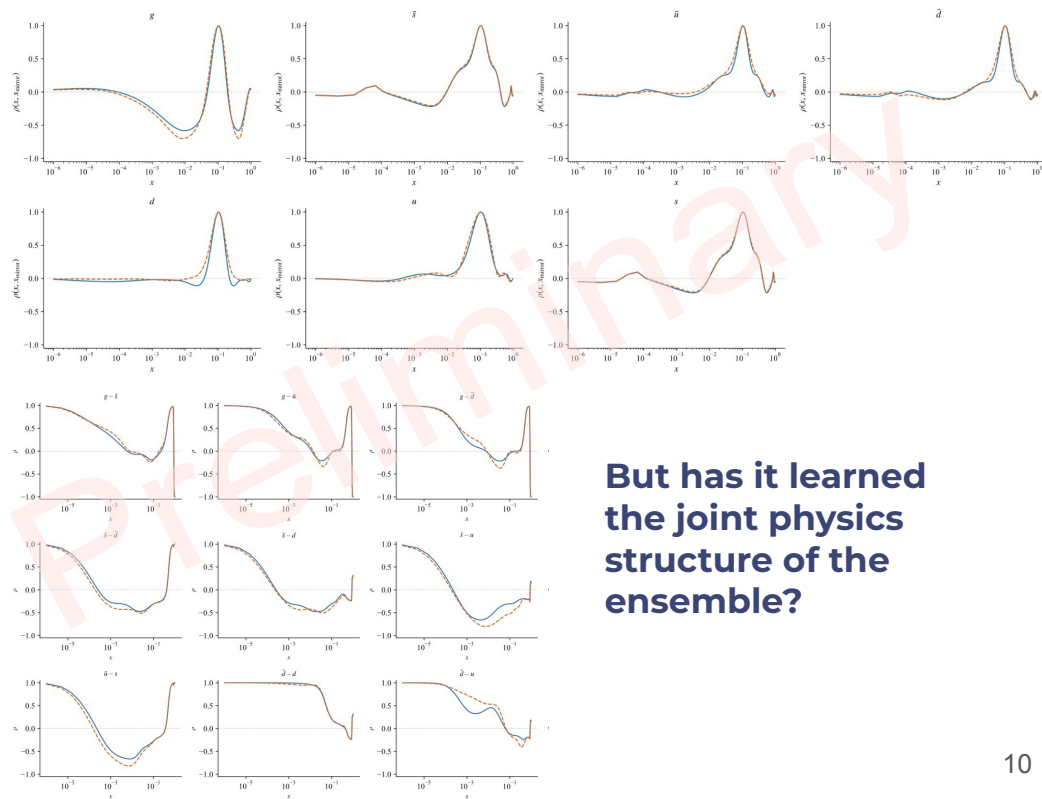
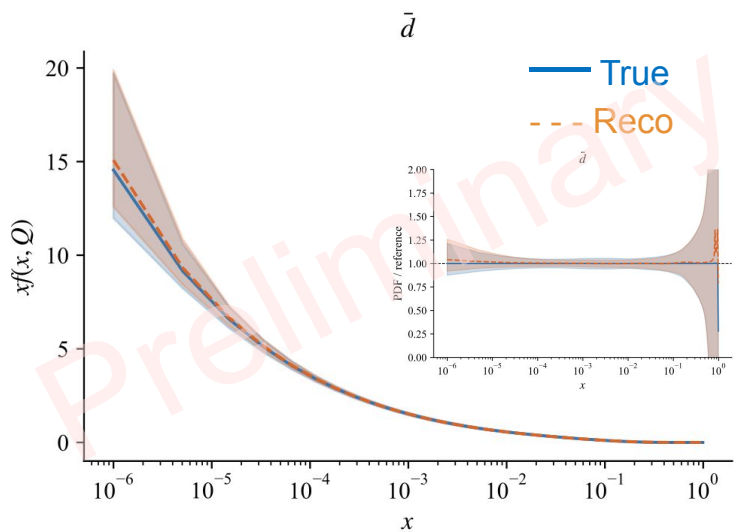
We approach the “Chinchilla limit” and saturate ... but is this good enough?

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



## At scale, conventional metrics suggest success

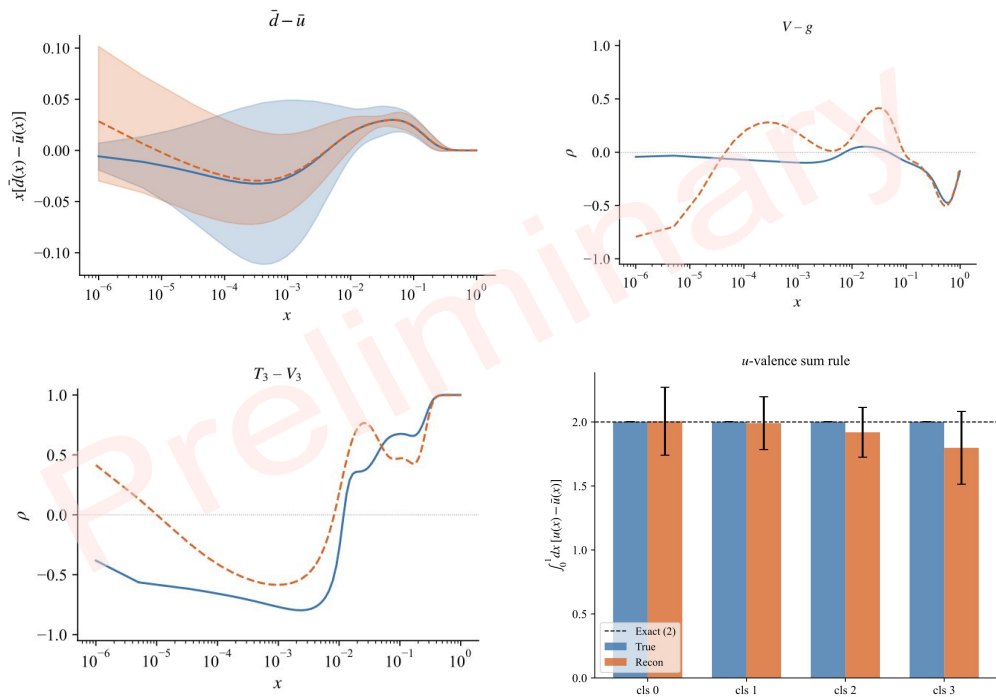
If I just look at the reconstructed error, things look pretty good. In fact, when I look at basic PDF diagnostics they look great!



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.



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

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

# Changing the language changes what the model can learn

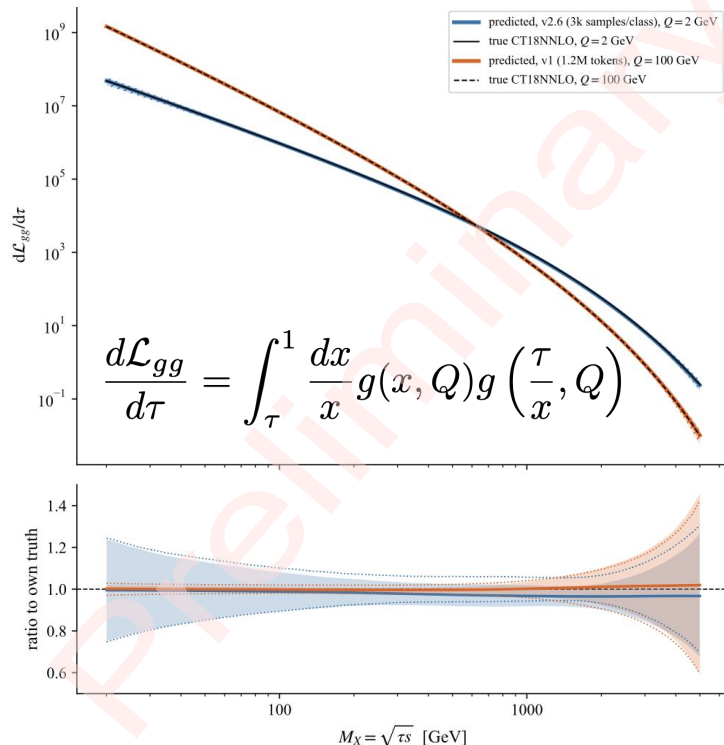
**Baseline:** x-sequence tokens

$$T_i = [x_i, g_i, u_i, \dots]$$

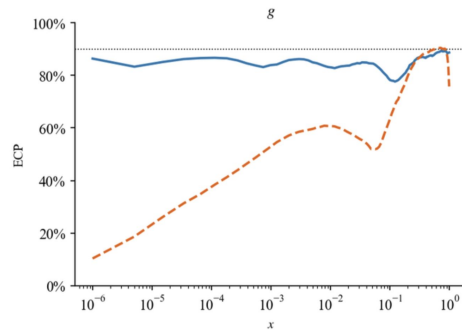
$$T_1 \rightarrow T_2 \rightarrow \dots \rightarrow T_{N_x}$$

**Physics structured:** x+flavor attention

|           | $x_1$                    | $x_2$                    | $x_3$                    | $\dots$ | $x_{N_x}$                |
|-----------|--------------------------|--------------------------|--------------------------|---------|--------------------------|
| $g$       | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | $\dots$ | <input type="checkbox"/> |
| $u$       | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | $\dots$ | <input type="checkbox"/> |
| $\bar{u}$ | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | $\dots$ | <input type="checkbox"/> |
| $d$       | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | $\dots$ | <input type="checkbox"/> |
| $\bar{d}$ | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | $\dots$ | <input type="checkbox"/> |
| $s$       | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | $\dots$ | <input type="checkbox"/> |
| $\bar{s}$ | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | $\dots$ | <input type="checkbox"/> |

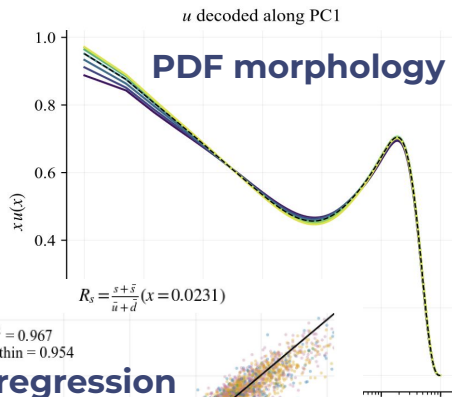
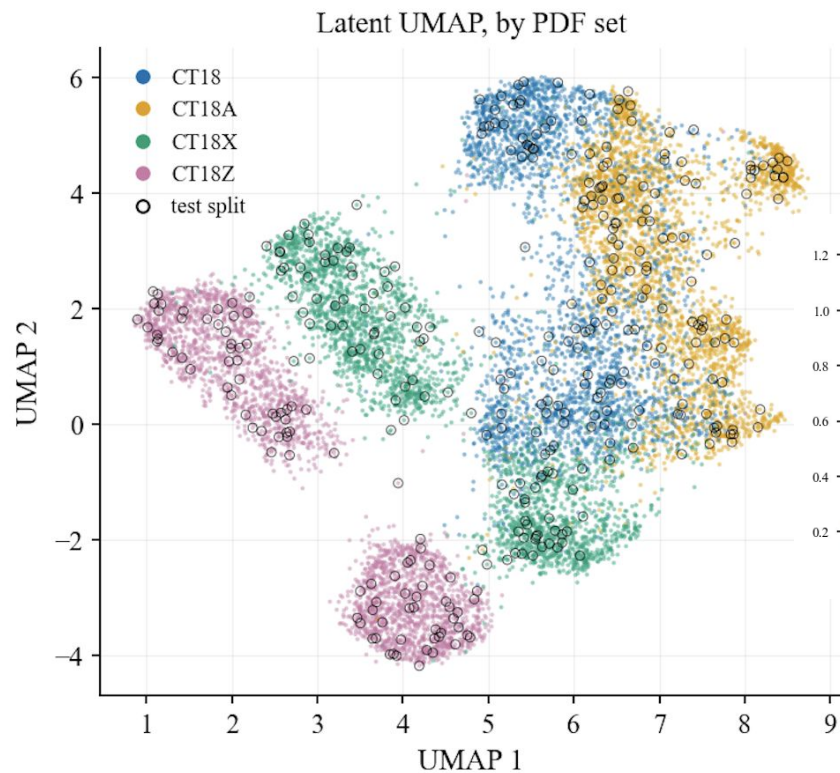


Exposing the physical structure of the input through tokenization improves calibration and that improvement survives into a collider-relevant derived quantity.

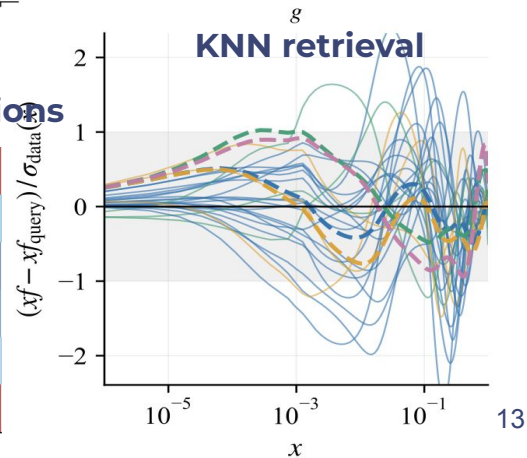
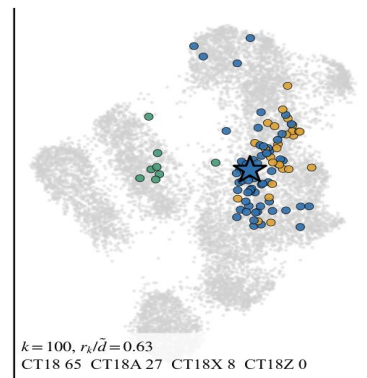
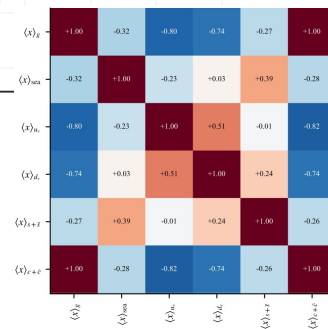


**Same training samples. Better uncertainty coverage. Better preservation of the gg luminosity!**

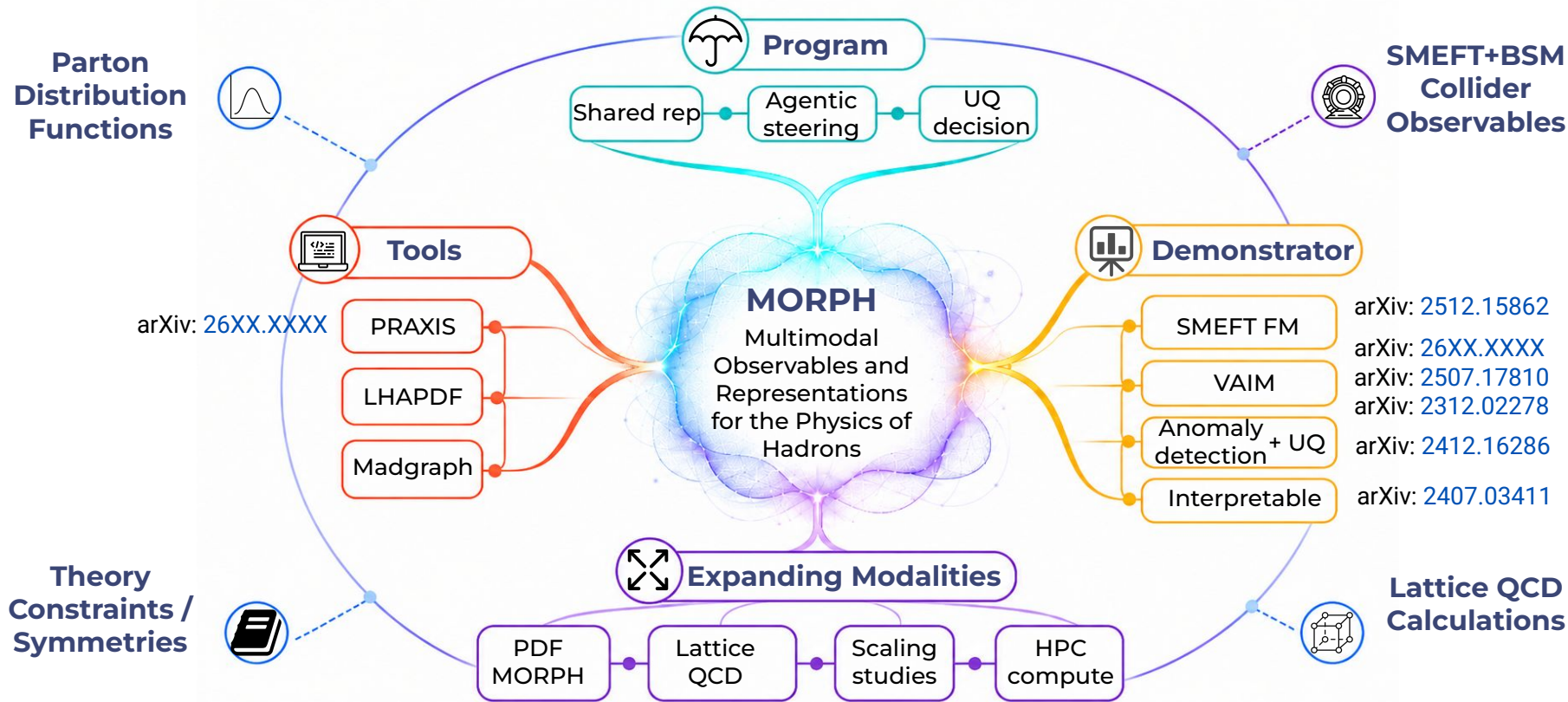
# MORPH and QCD phenomenology

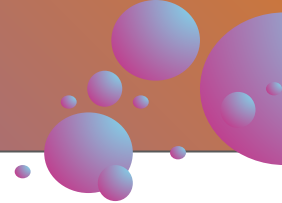


## Physics correlations



# Agent-steered multimodal physics





## Conclusions

**The PDF ecosystem contains structure worth learning** — data, theory, and methodology jointly define a global inferred ensemble.

**Reconstruction alone is not a good enough benchmark** — complex physics benchmarks expose missing statistical information that standard ML metrics miss.

**How we represent information matters** — tokenization schemes, x + flavor + axial attention improve data efficiency and preserve meaningful phenomenological information.

**Multimodality is the next step** — combining heterogeneous information of PDFs through lattice, collider, and BSM observables is the goal.

**The objective is a representation that preserves the physics needed for phenomenology.**

**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.