

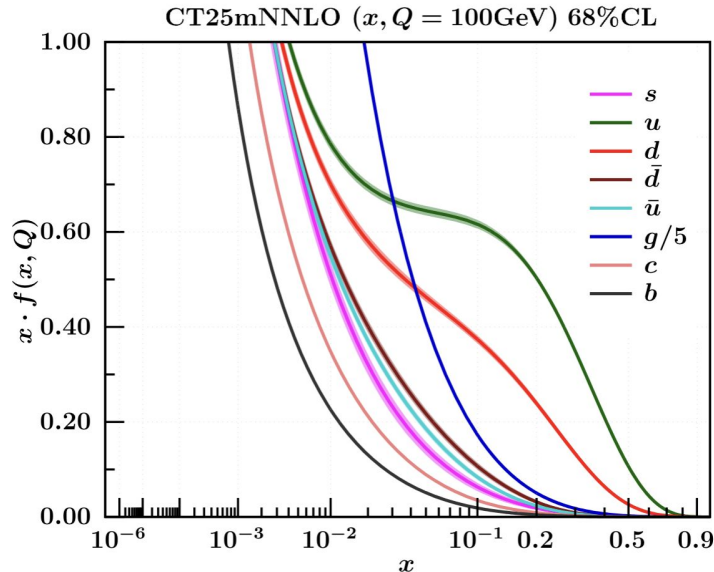
From inverse problems to foundation models in hadron structure

Brandon Kriesten · Argonne National Laboratory

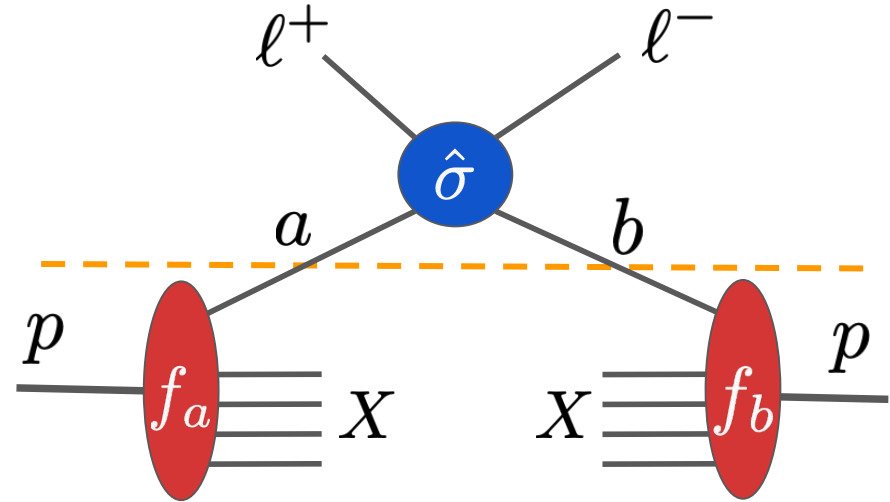
MENU26 · 22 October 2026

The structure of hadrons — Parton Distribution Functions

PDFs describe how the hadron's momentum is carried by the constituents (quarks and gluons), and are defined through factorization theorems.

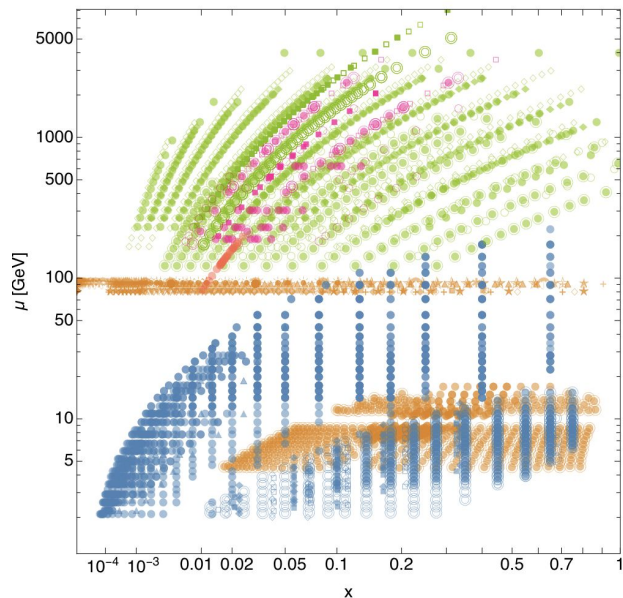


Neutral Current Drell-Yan

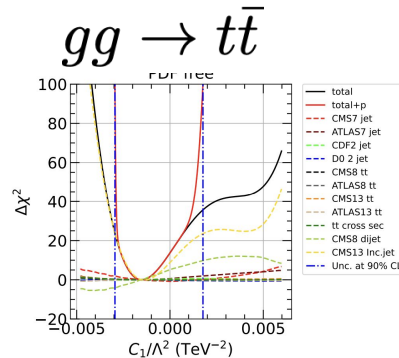


$$\sigma = \sum_{a,b} \int dx_a dx_b f_a(x_a) \hat{\sigma}_{ab \rightarrow \ell^+ \ell^- + X} f_b(x_b)$$

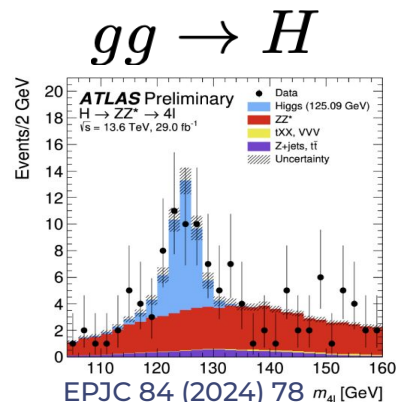
PDFs set the baseline for discovery



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.

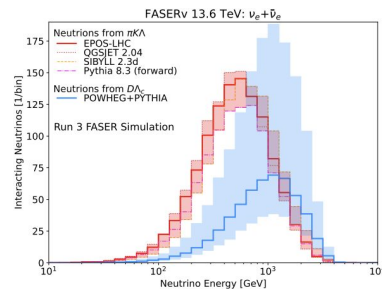


JHEP 05 (2023) 003



EPJC 84 (2024) 78

$$\nu N \rightarrow \ell X$$



2402.13318

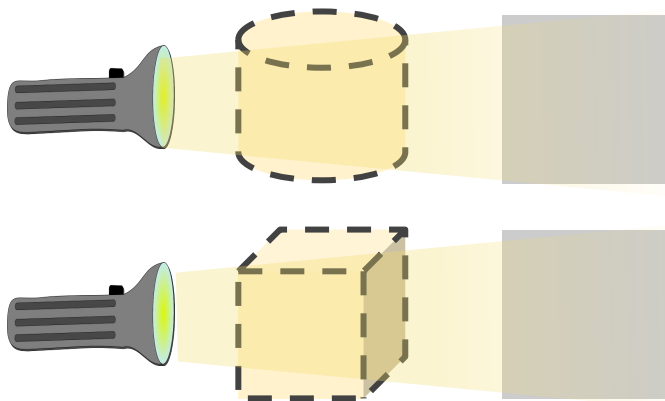
Motivating studies relevant for current and future science programs like the (HL) LHC, EIC, FCC, and DUNE.

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

Global fitting is an ill-posed inverse problem

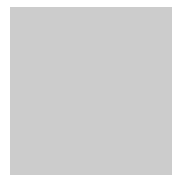
Forward Mapping

proton structure $\rightarrow$ observable

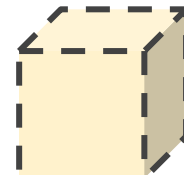


Inverse Mapping

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



???



The data do not choose a unique answer!

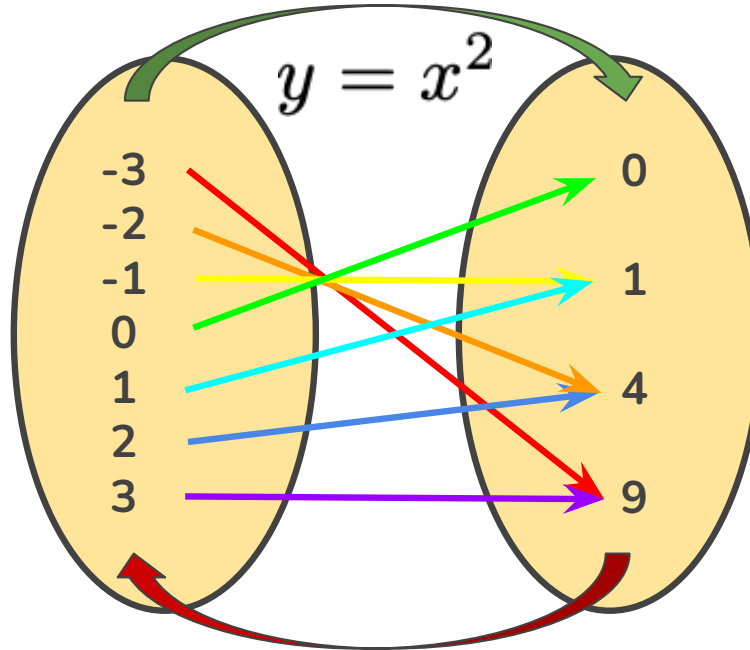
The goal is not to find one answer, but to characterize the full space of answers consistent with the evidence.

Inverse inference: when observables lose information

$$f : X \rightarrow Y$$

$$\mathbb{Z}_2$$

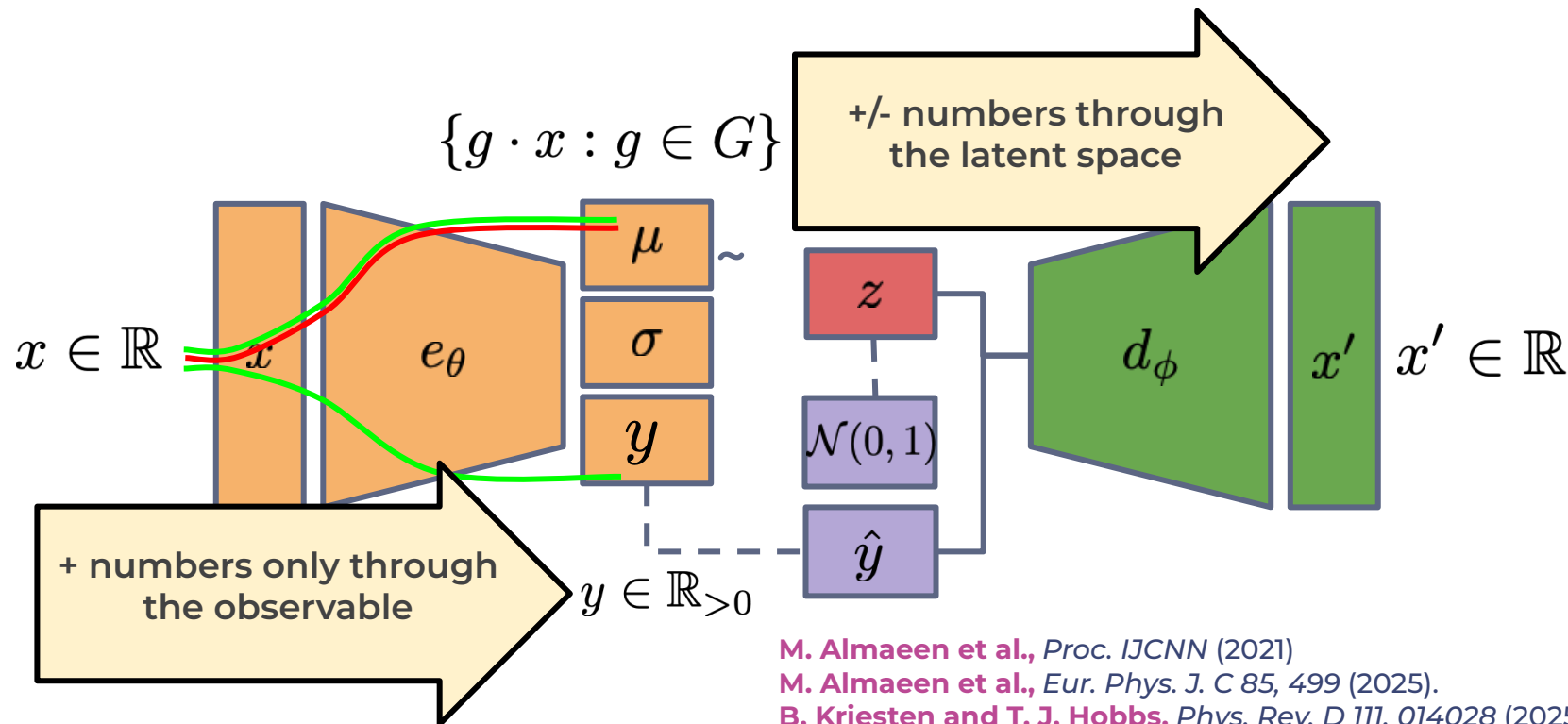
$$x \mapsto -x$$



The forward mapping is many-to-one: y does not contain enough information to uniquely recover x .

VAIM in action — a flow of information

$\mathbb{Z}_2$



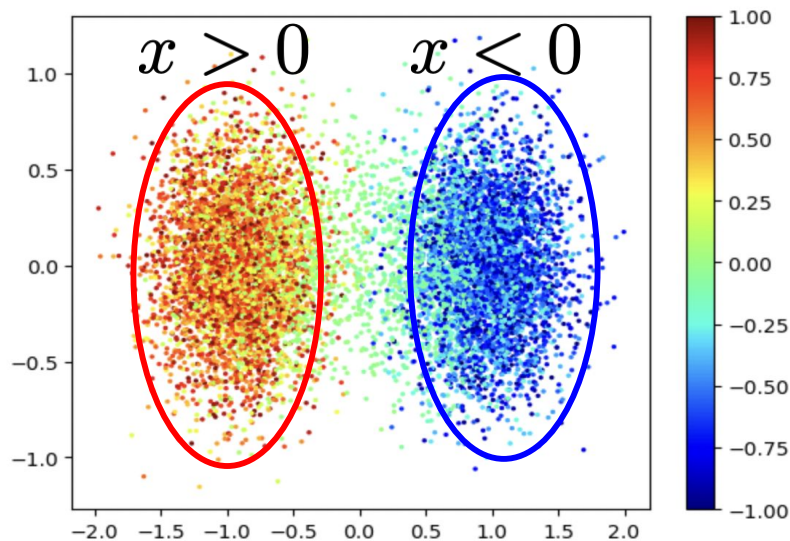
M. Almaeen et al., *Proc. IJCNN* (2021)

M. Almaeen et al., *Eur. Phys. J. C* 85, 499 (2025).

B. Kriesten and T. J. Hobbs, *Phys. Rev. D* 111, 014028 (2025).

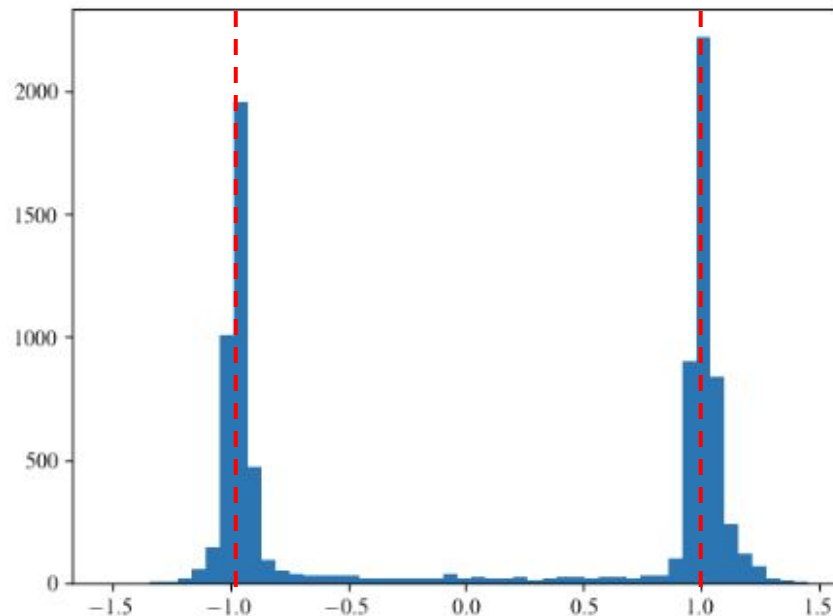
B. Kriesten et al., *Phys. Lett. B* 874, 140132 (2026).

The latent space recovers the hidden degeneracy



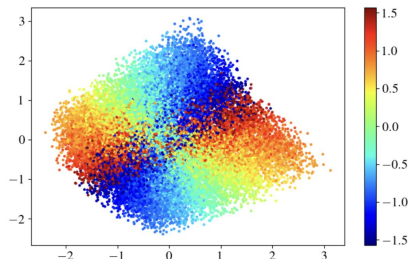
The degeneracy of the inverse problem becomes emergent structure in the learned representation.

The model reconstructs each solution that is consistent with the evidence.

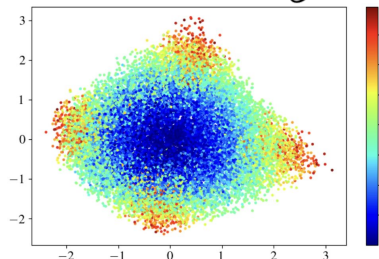


The latent space learns the group structure

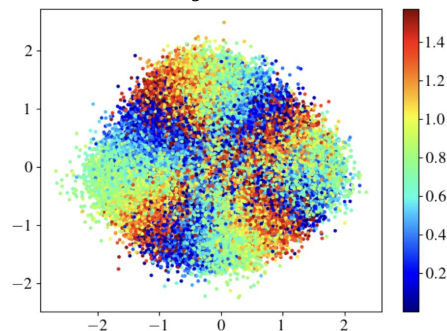
$SO(2)$



$$o = x^2 + y^2$$

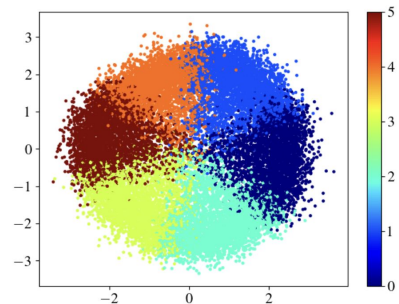


$$\begin{aligned} D_4 \quad o &= r^2 & o &= r^4 \cos(4\theta) \\ r &= x^2 + y^2 & \theta &= \arctan(y/x) \end{aligned}$$

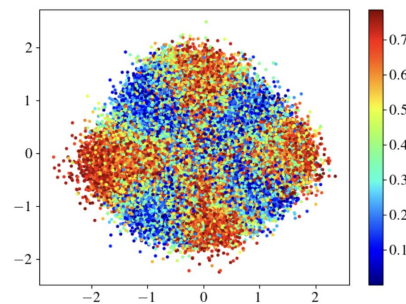
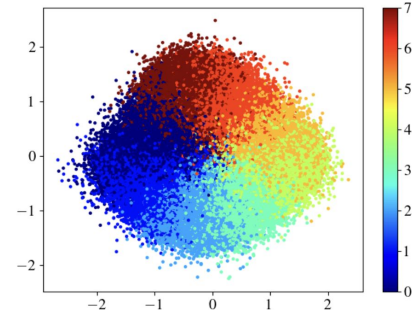
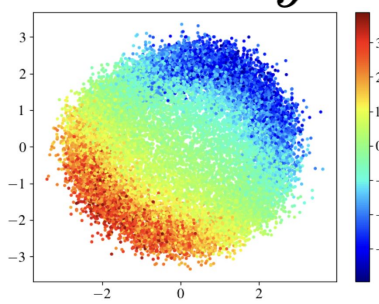


In these mathematical examples, we can directly interpret the symmetries from the latent distribution.

S_3



$$o = x + y + z$$



Lattice QCD inverse problem

Lattice QCD innovation allows for new calculations of moment-like objects that contain PDF information.

Radyushkin, PRD 96 (2017)
Balitsky, et al., PLN 808 (2020)
Fan, et al., PRD 108 (2023)

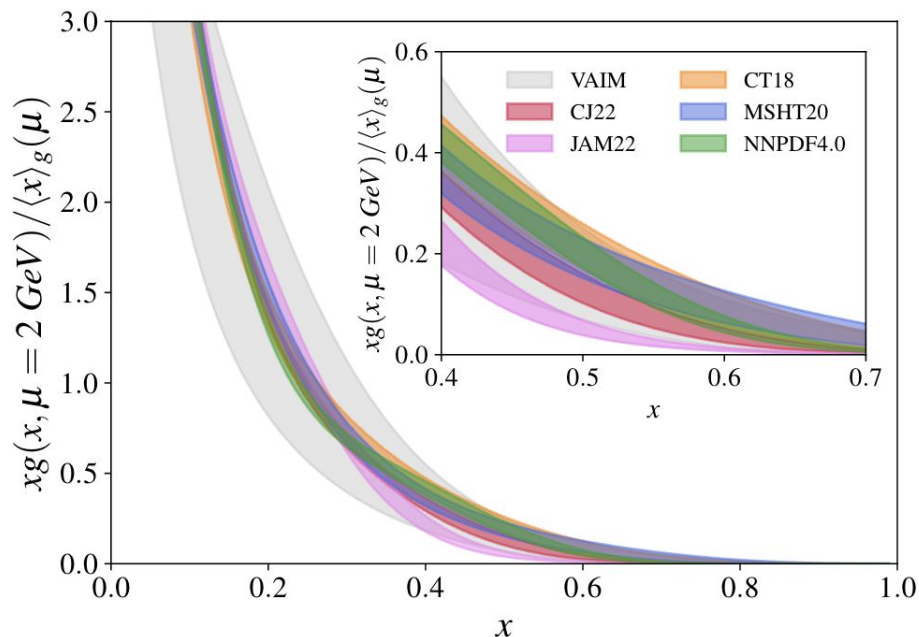
$$\boxed{\mathcal{M}(\nu, z^2)} = \int_0^1 dx \frac{xg(x, \mu^2)}{\langle x \rangle_g(\mu)} R_{gg}(x\nu, z^2\mu^2) + \mathcal{O}(z^2m^2, z^2\Lambda_{QCD})$$

Lattice calculated reduced pseudo loffe time distributions (RpITDs) = $\int_0^1 dx$ PDF $\times$ perturbative known kernel

Ensemble	a09m310
a (fm)	0.0888(8)
$L^3 \times T$	$32^3 \times 96$
M_π^{val} (GeV)	0.313(1)
$M_{\eta_s}^{\text{val}}$ (GeV)	0.698(7)
P_z (GeV)	[0, 3.05]
N_{cfg}	1009
$N_{\text{meas}}^{2\text{pt}}$	387,456
t_{sep}	[6, 10]

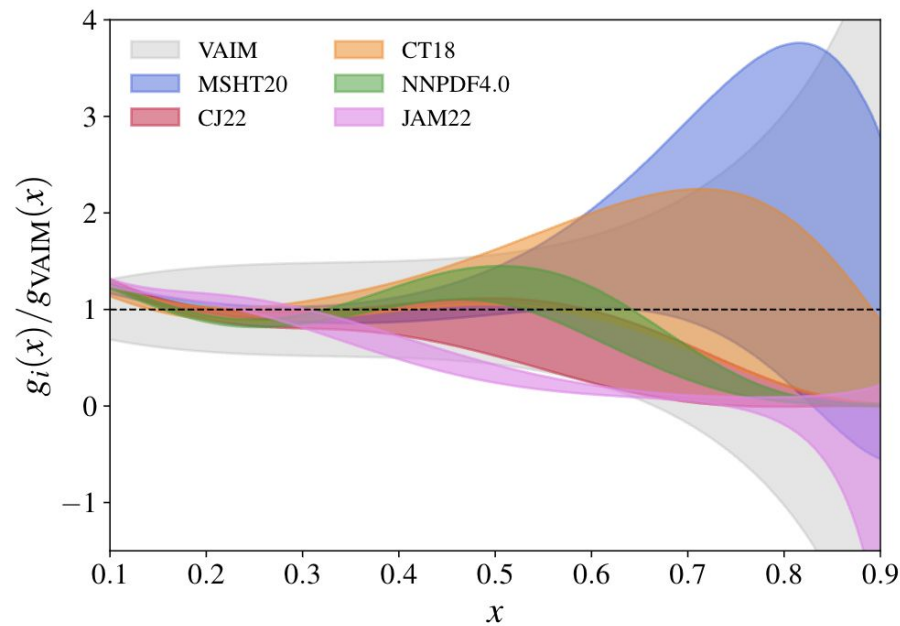
Recovering the underlying x-dependence is an inverse problem.

Gluon PDFs from the Lattice



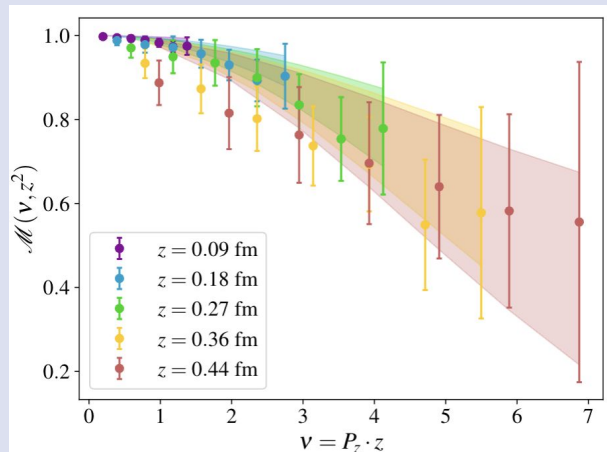
VAIM reconstructs a family of gluon PDFs consistent with both lattice information and phenomenological expectations from data.

B. Kriesten et al., *Phys. Lett. B* 874, 140132 (2026).

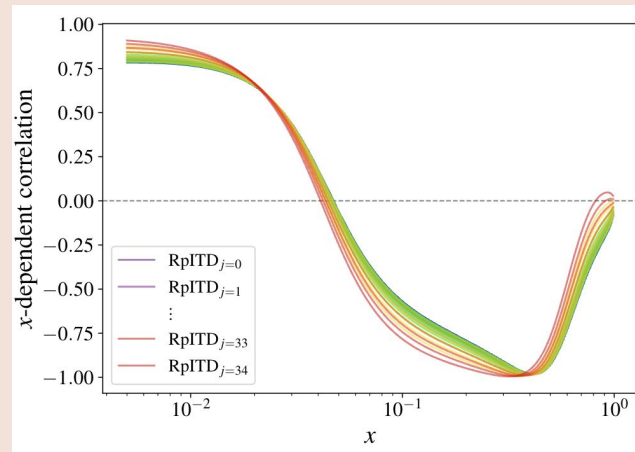


Did the model learn the right physics?

Closure test — does the result produce the observable?



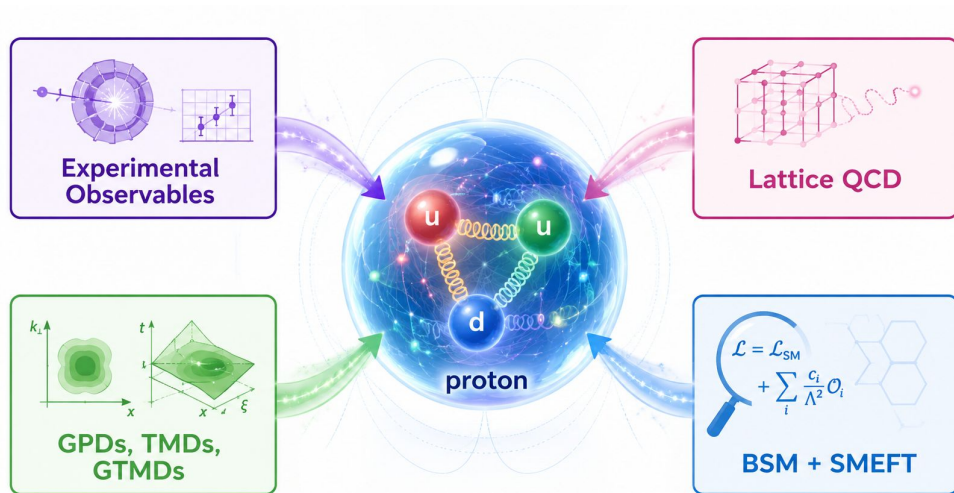
Interpretability — which lattice information constrains which part of the PDF?



Validation requires both predictive agreement and an interpretable dependence on the collected evidence.

Hadronic structure lives in a multimodal ecosystem

Modalities: different experiments and calculations provide complementary views of the same underlying proton structure.



Different observables constrain complementary directions in the space of hadron structures!

The challenge is connecting heterogeneous information while preserving the physics encoded within each source.

The information structure behind PDF inference

Every global PDF fit is conditioned on data, theory, and methodological assumptions.

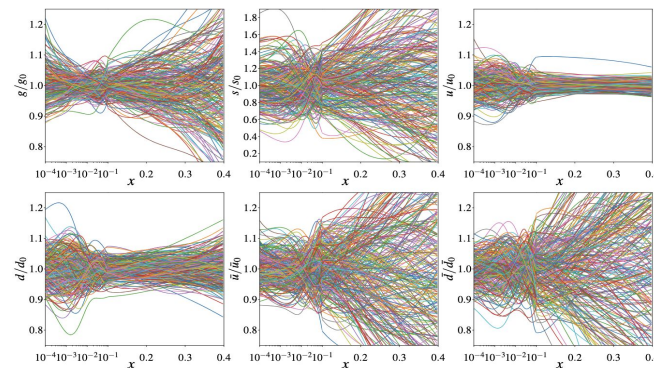
$$p(f|D, T, M) \propto p(D|f, T) \cdot p(f|M)$$

D Data
What constrains the PDF;
measurements, kinematics

T Theory
The forward map; order,
evolution, lattice, α_s , m_c

M Methodology
Parameterization
dependence, uncertainty

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



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

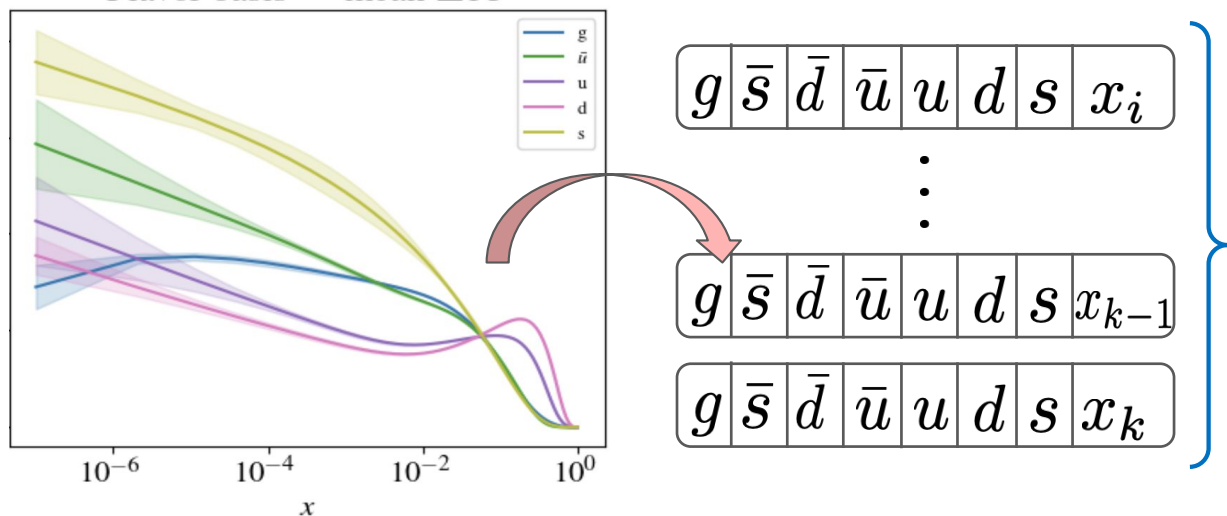
Anomaly Detection

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.



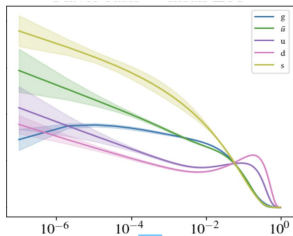
For a first demonstration we compile **all flavor information** at a particular- x . A PDF replica is then represented as a sequence along x .

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

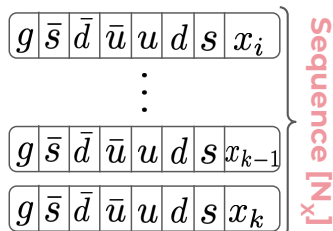
Learning a shared representation of global fits

1 Input tokens

PDF Ensembles



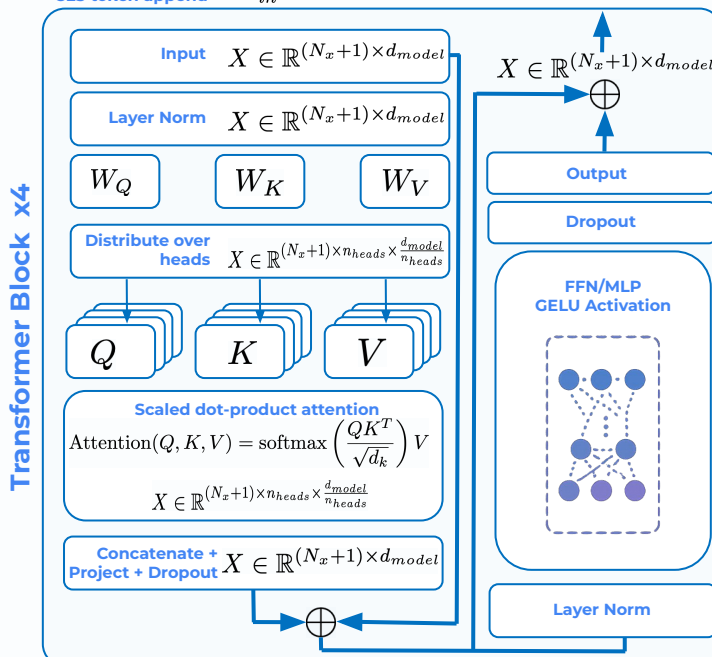
Discretized Tokens



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

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

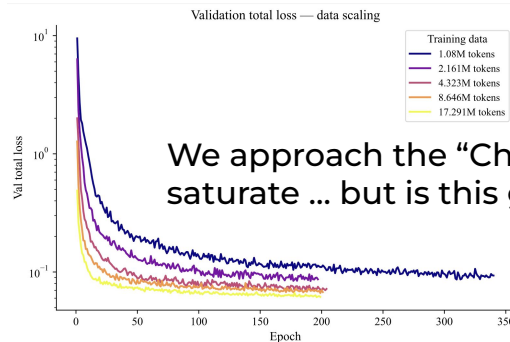
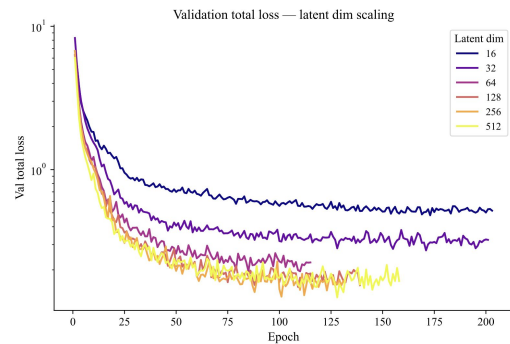
Masked Flavor Modeling

Reconstruction

Denoising

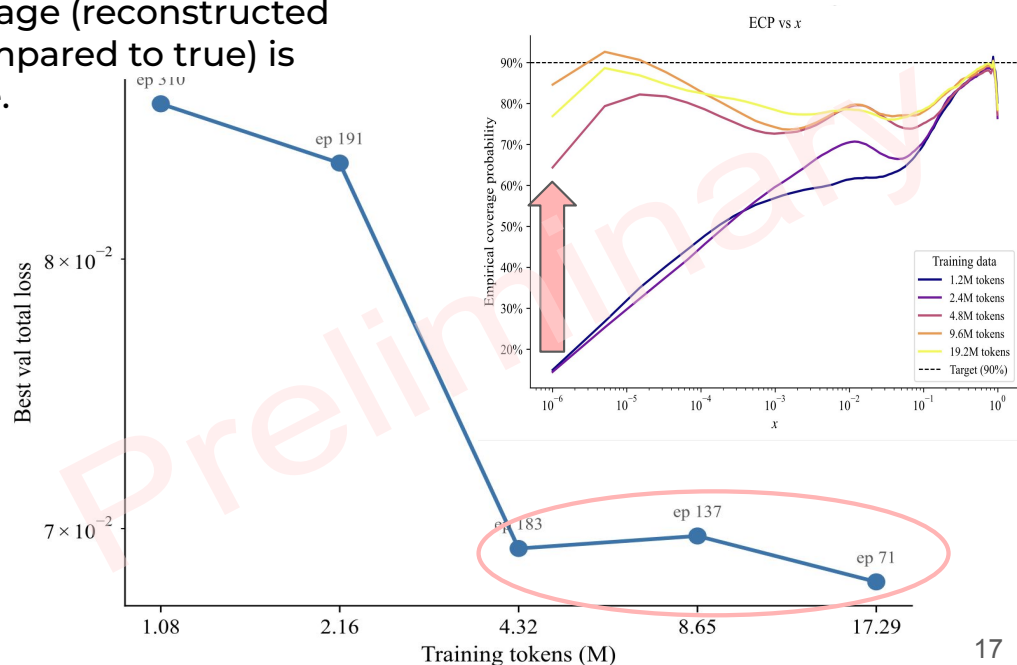
Neural scaling locates the statistical performance frontier

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.



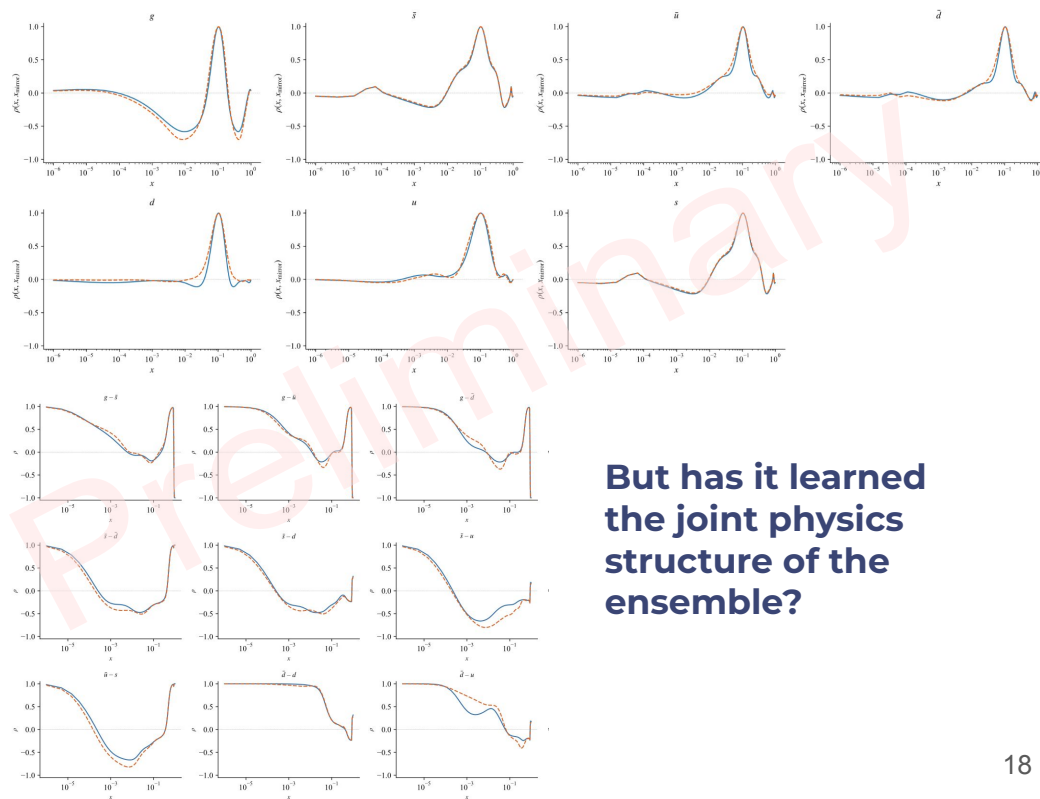
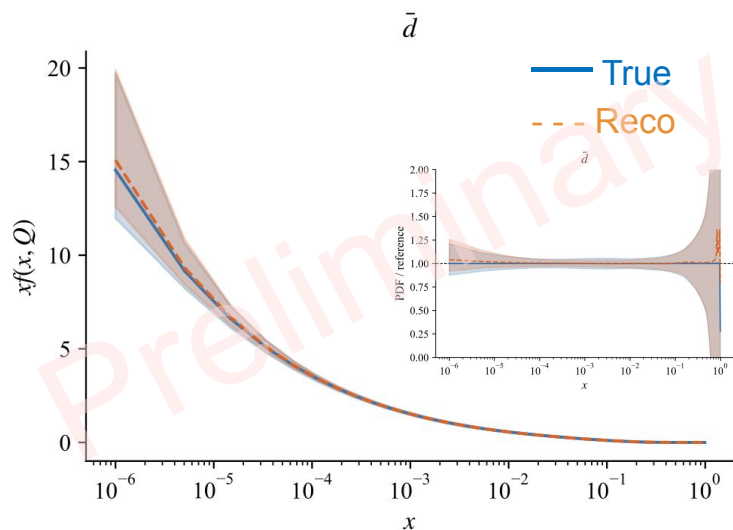
We approach the “Chinchilla limit” and saturate ... but is this good enough?

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



At scale, conventional metrics suggest success

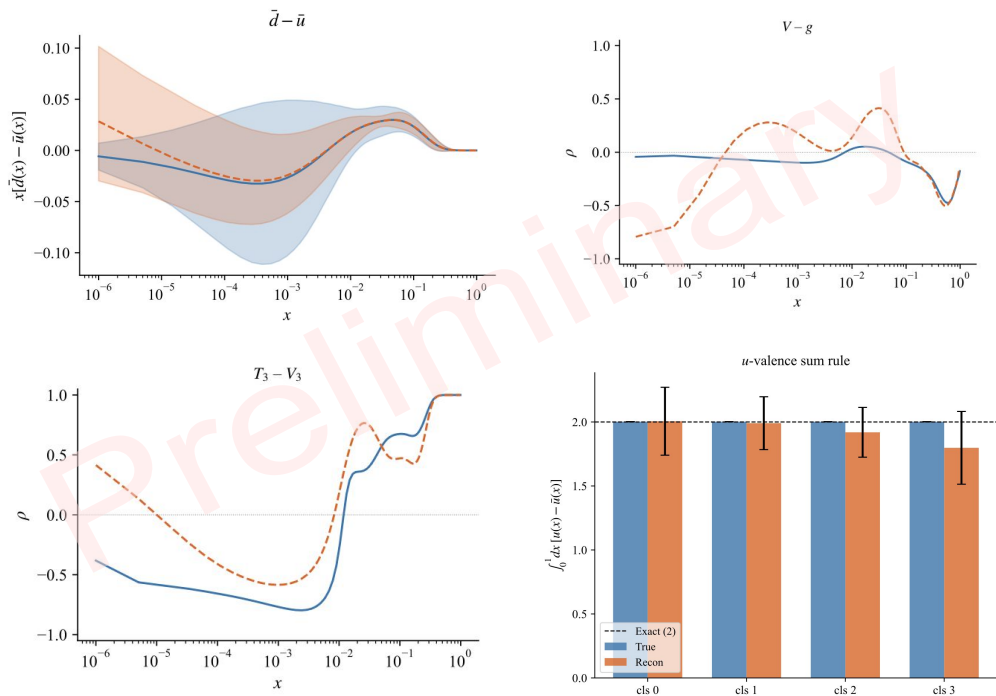
If I just look at the reconstructed error, things look pretty good. In fact, when I look at basic PDF diagnostics they look great!



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

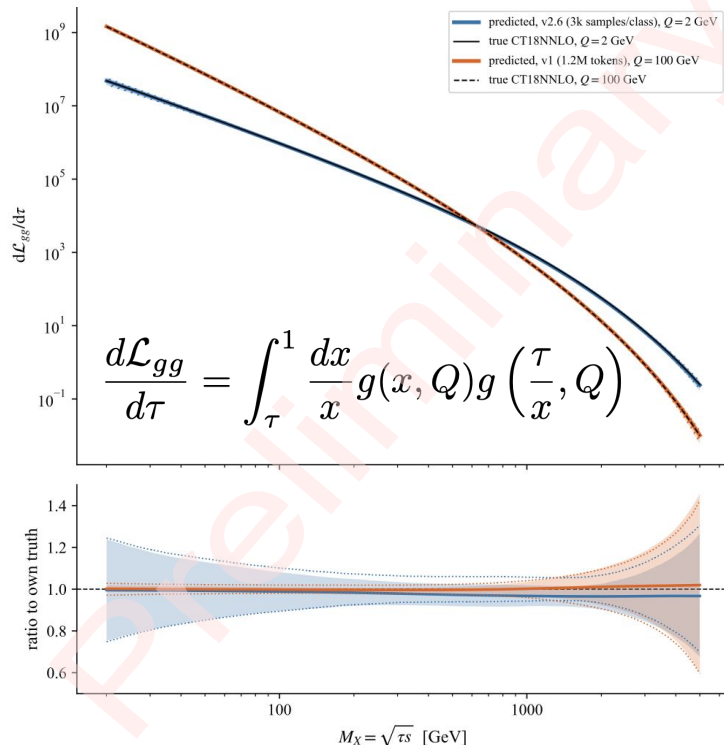
Baseline: x-sequence tokens

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

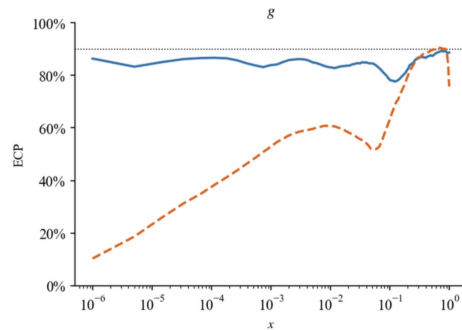
$$T_1 \rightarrow T_2 \rightarrow \dots \rightarrow T_{N_x}$$

Physics structured: x+flavor attention

	x_1	x_2	x_3	$\dots$	x_{N_x}
g	☐	☐	☐	$\dots$	☐
u	☐	☐	☐	$\dots$	☐
$\bar{u}$	☐	☐	☐	$\dots$	☐
d	☐	☐	☐	$\dots$	☐
$\bar{d}$	☐	☐	☐	$\dots$	☐
s	☐	☐	☐	$\dots$	☐
$\bar{s}$	☐	☐	☐	$\dots$	☐

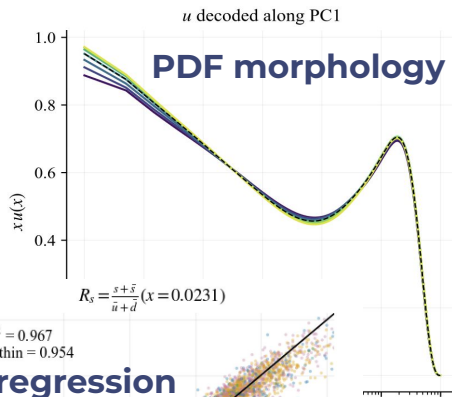
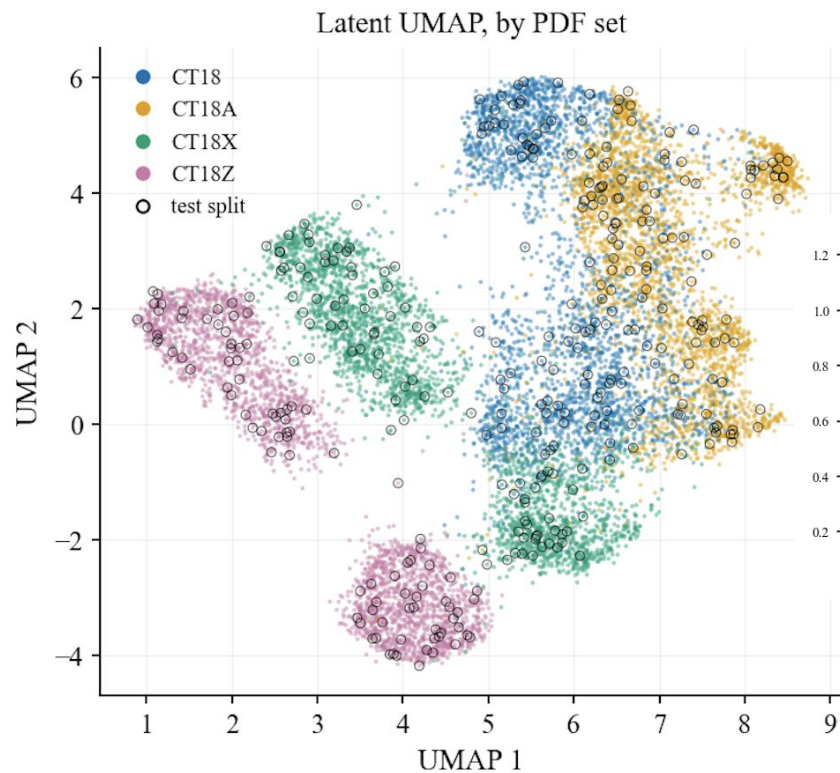


Exposing the physical structure of the input through tokenization improves calibration and that improvement survives into a collider-relevant derived quantity.

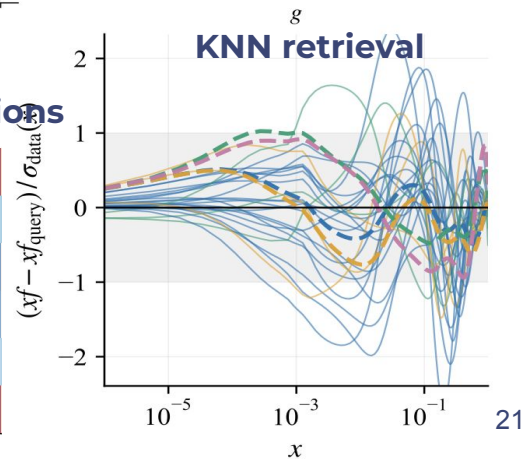
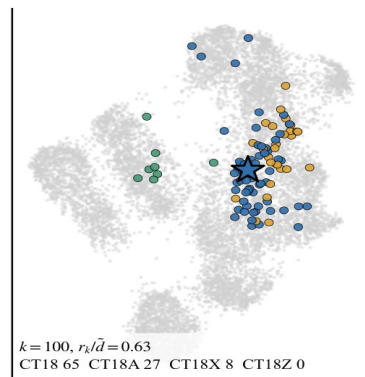
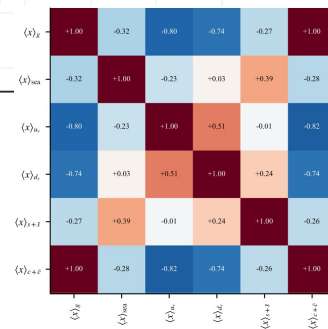


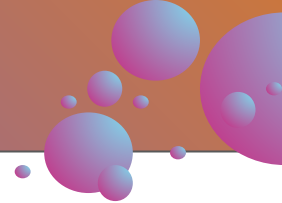
Same training samples. Better uncertainty coverage. Better preservation of the gg luminosity!

MORPH and QCD phenomenology



Physics correlations





Conclusions

Ill-posed inverse problems have structure — symmetries of the forward map organize degeneracies into equivalence classes, and a latent space can learn them ($\mathbb{Z}_2$, $SO(2)$, S_3 , D_4).

Generative inverse solvers separate what observables fix from what they leave free — lattice RpITDs
→ gluon PDF families, validated by closure and interpretability.

A reusable representation must preserve physically meaningful statistical structure — accurate reconstruction alone is not enough.

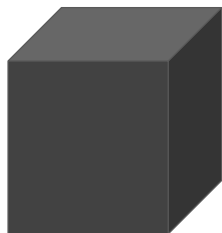
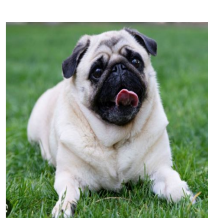
How we represent information matters — tokenization schemes, x + flavor + axial attention improve data efficiency and preserve meaningful phenomenological information.

The opportunity for frontier AI is not only to accelerate physics workflows, but to help make scientific reasoning itself scalable.

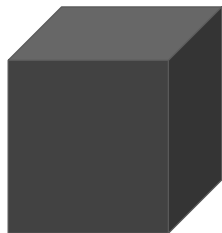
Thank you for your attention!

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.

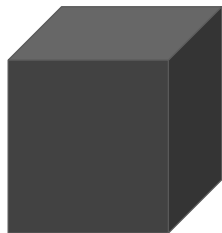
When should we trust the learned representation?



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%



Even if a model has learned physically meaningful structure, that does not mean every point in the learned representation space should be treated as certain.

A scientifically useful model must know when it does not know!

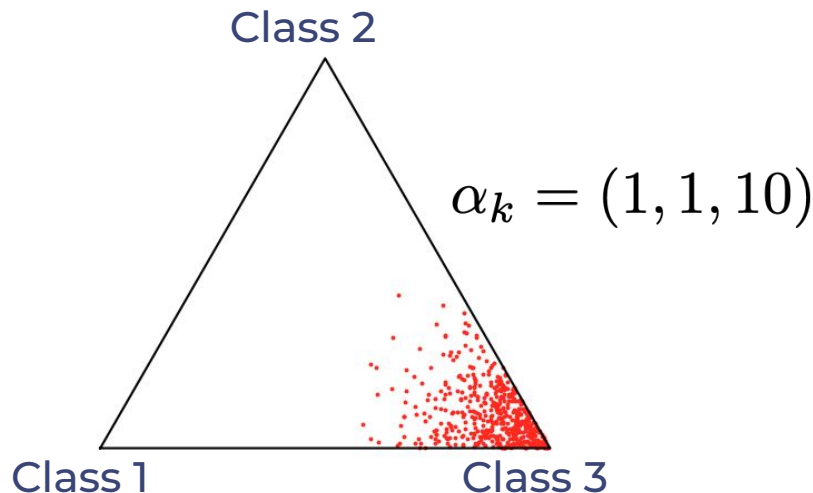
Dirichlet Prior Networks

A categorical distribution models individual probabilities of specific categories $(p_1, p_2, \dots, p_k)$. The Dirichlet distribution models the prior beliefs over each category's probabilities, designated by parameters $(\alpha_1, \alpha_2, \dots, \alpha_k)$.

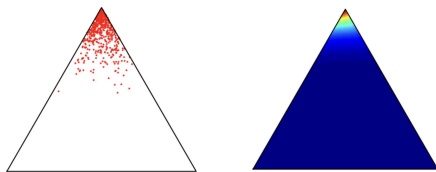
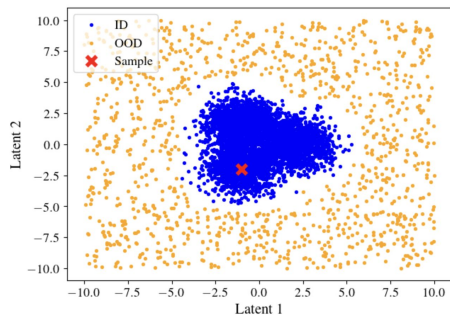
$$p(\mu; \alpha) = \frac{\Gamma(\alpha_o)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K p_k^{\alpha_k - 1}$$

We model evidence of the prediction, not just the prediction itself.

Malinin and Gales arXiv:1802.10501

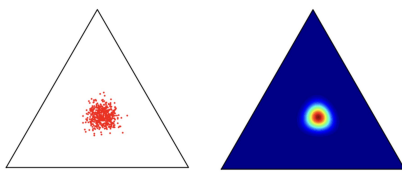
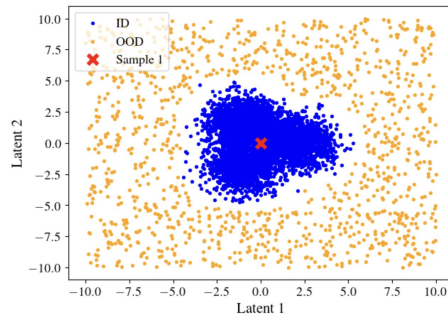


DPNs: an overlapping Gaussian example



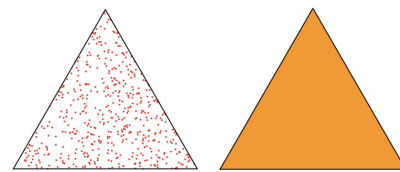
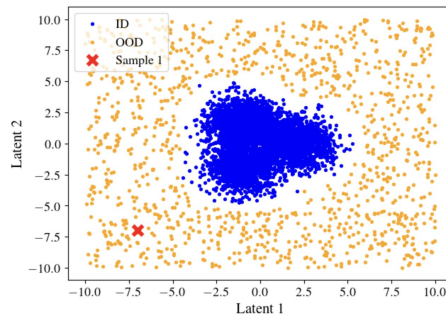
$$\alpha_k = (1, 10, 1)$$

In-distribution
low ambiguity



$$\alpha_k = (8, 8, 8)$$

In-distribution
high ambiguity



$$\alpha_k = (1, 1, 1)$$

out-of-distribution

Uncertainty can arise for fundamentally different reasons — all statistically valid!

Aleatoric and Epistemic Uncertainty Decomposition

Through training with maximum likelihood estimation (MLE) we can rework the expression to factorize into two distinct components. One is epistemic - the KL divergence term; and the other is aleatoric - the entropy term.

$$\mathbb{E}_{p_{true}(x,y)} [\mathcal{L}^{NLL}(y, x, \theta)] = \mathbb{E}_{p_{true}(x)} \left[D_{KL} (p_{true}(y|x) || p(y|x, \theta)) - \mathbb{H}(p_{true}(y|x)) \right]$$

Epistemic

Epistemic uncertainty is reducible by training procedures or algorithmic development. Also can be influenced by choice of distribution to model the data.

Aleatoric

Aleatoric uncertainty is not reducible because it is related to the true underlying data distribution.

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

- Amplitudes
- Symmetries
- Analytic equations
- Fitting assumptions
- Simulations
- Matrix elements

AI/ML

- Arrays/tensors
- Floats
- Graphs
- Tokens
- Distributions
- Metadata

What a model can learn is constrained by how we choose to represent the physics.

The scientific interface

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

