Developments in Interpretable and Explainable AI/ML for PDFs

14 June 2024

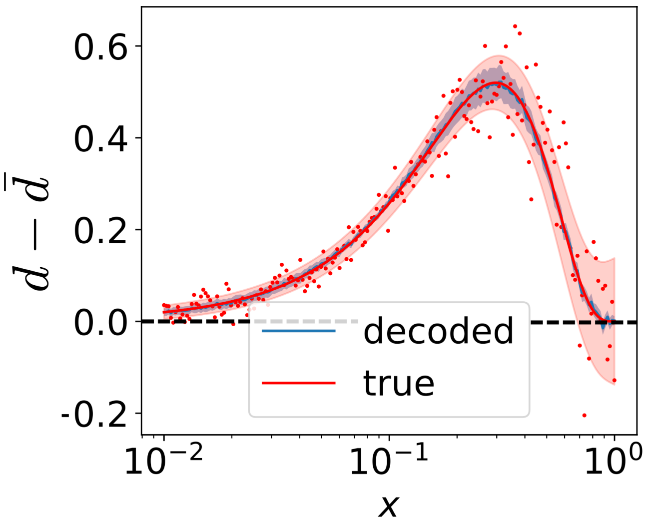
or, can we understand what ML models are actually doing in the quest to quantify PDFs and their uncertainties…

DALL-E: "A confused, despondent robot painted in the style of Matisse."

T. J. Hobbs

Argonne HEP Division

Theory Group



thanks to… Brandon Kriesten, Jon Gomprecht; CTEQ colleagues

(views are my own..)

Argonne NATIONAL LABORATORY

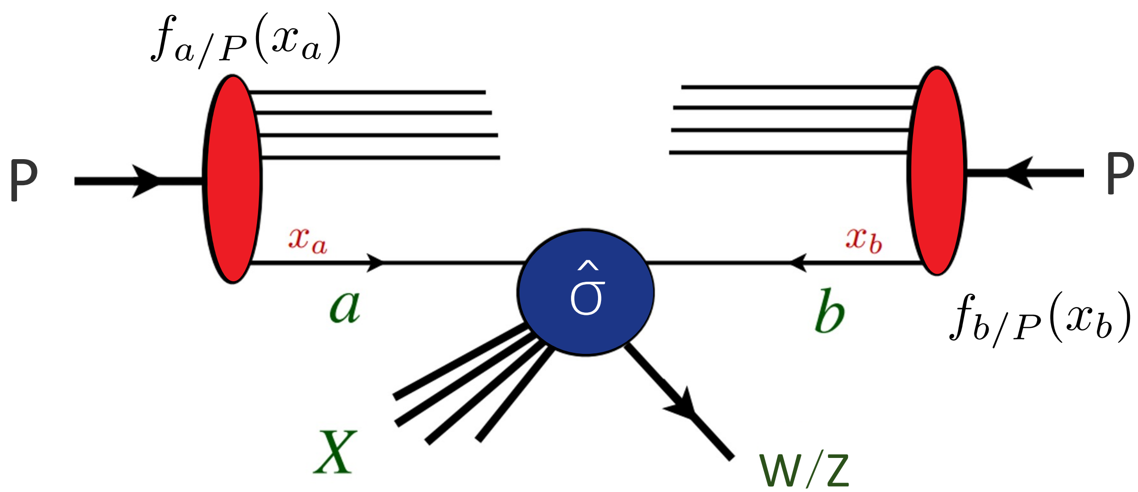
CTEQ

HEP motive: PDFs are ubiquitous in precision theory

$$\sigma(PP \to W/Z+X) \; = \; \sum_n \alpha_s^n \sum_{a,b} \int dx_a dx_b$$

QCD factorization theorem(s)

$$\times \; f_{a/P}(x_a)\, \hat{\sigma}^{(n)}_{ab\to W/Z+X}(\hat{s})\, f_{b/P}(x_b)$$



MANY measurements to test the SM involve colliding protons

→ protons scatter through parton-level interactions: quark-quark, quark-gluon, …

→ precision means accurate theory + **parton distribution functions (PDFs)**

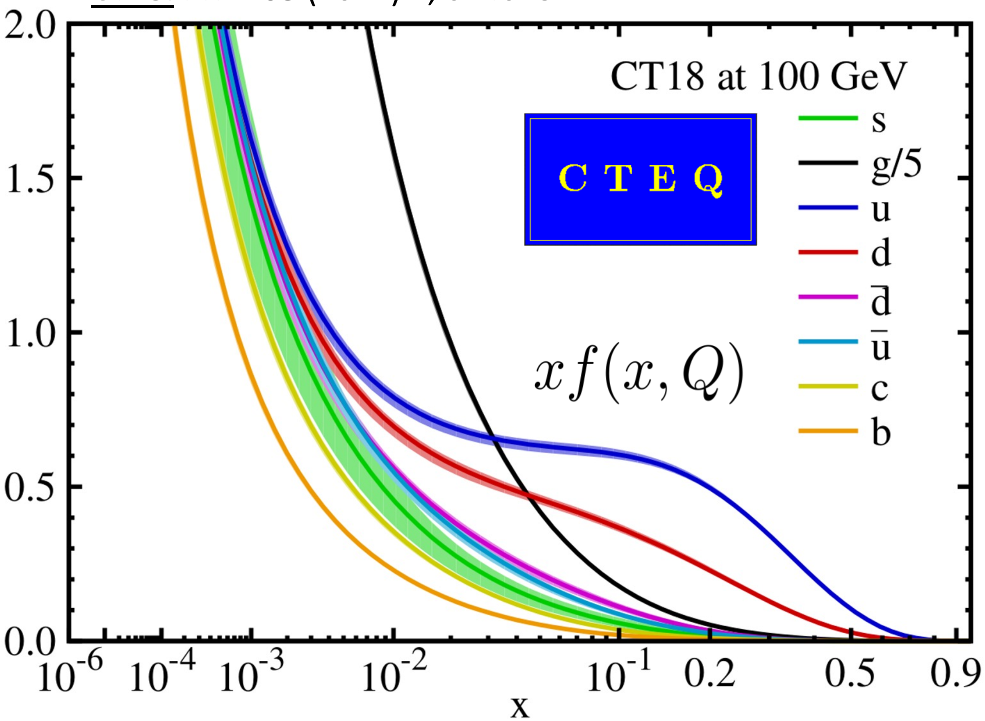
[**inverse problem**: *extract* from data...]

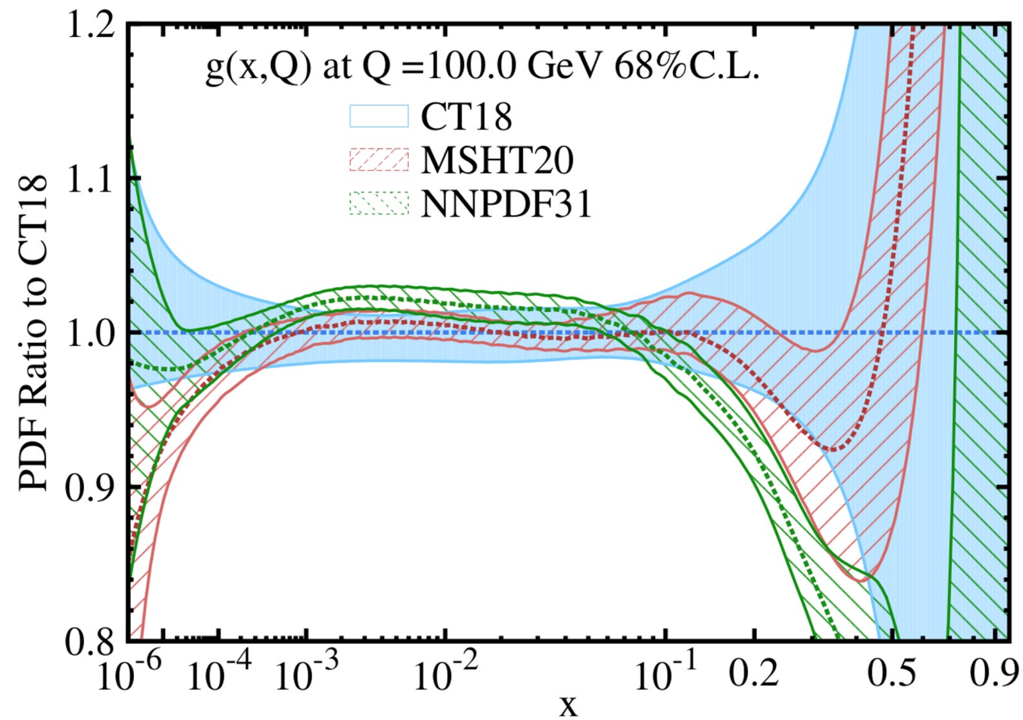
result: QCD global fits – precision from diverse data

pheno needs high-precision → (reproducible) reductions to PDF uncertainties

→ necessary to push theory accuracy; (N)NNLO QCD, NLO EW, …

CT18: PRD**103** (2021) 1, 014013





→ HEP pheno requirements impose stringent demands on PDFs

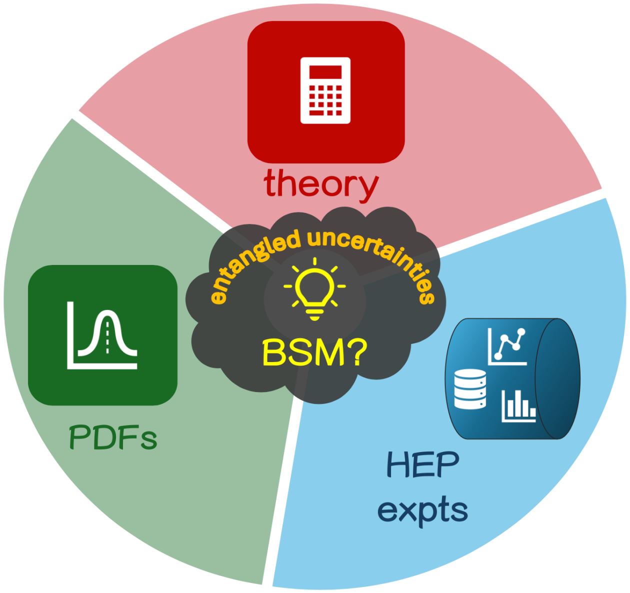
→ extensive benchmarking for high-stakes channels; UQ

J. Phys. G 49 (2022) 8, 080501

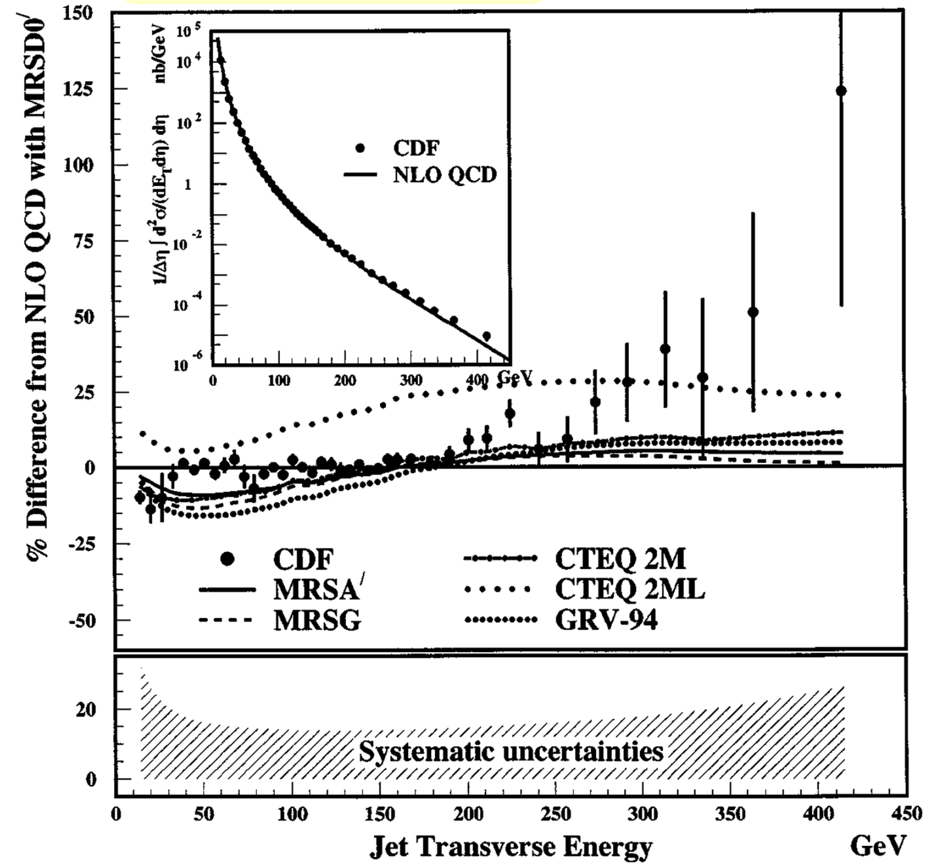
discovery reach closely tied to PDF precision

- SM baselines at colliders depend on PDFs

$$\sigma \sim f_a(x) \otimes \hat{\sigma}^{\rm pQCD} \otimes f_b(x)$$



PRL77, 438 (1996)

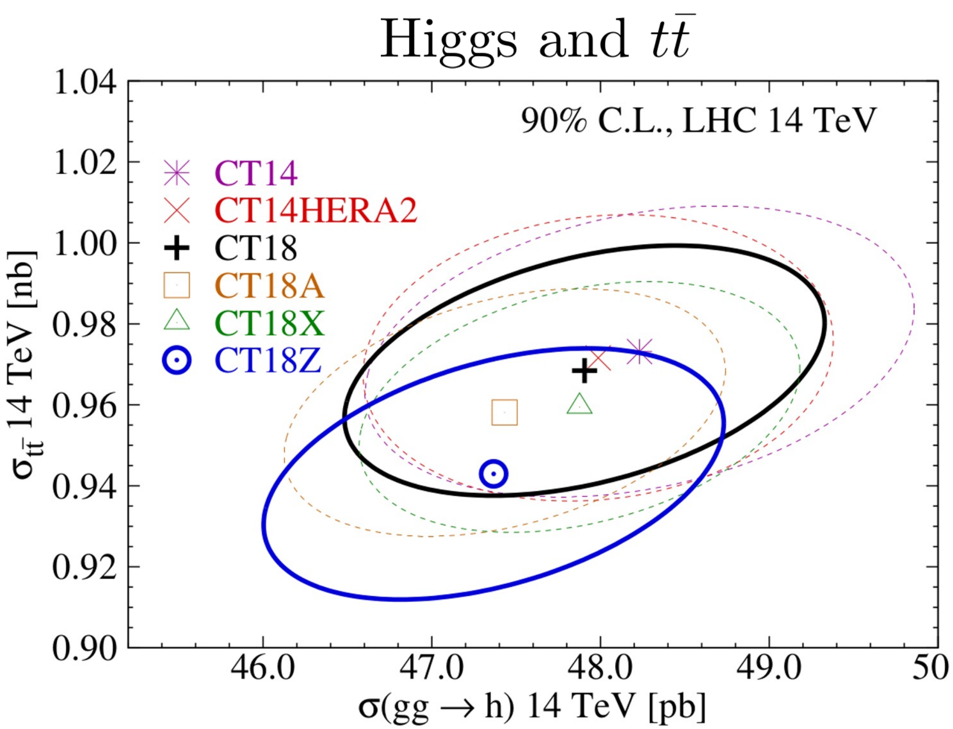


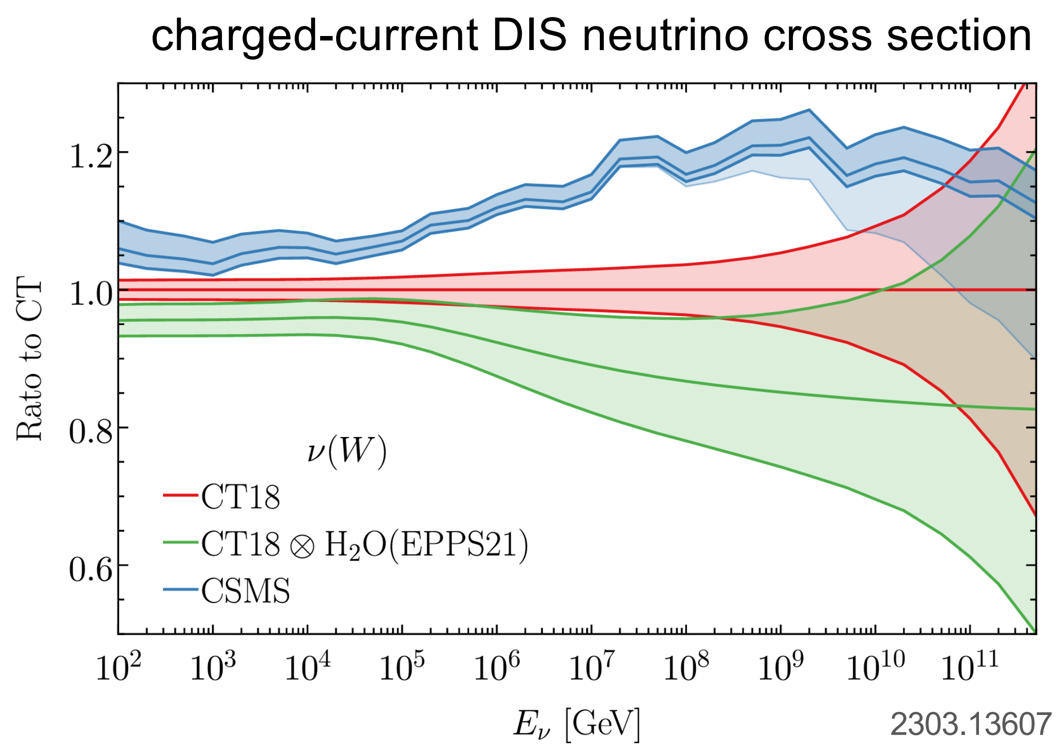
- cautionary e.g., CDF incl-jet E_T anomaly
  - → quark compositeness?
  - → …**no**: mismodeling of gluon PDF

high theory accuracy; PDF precision; faithful uncertainties essential to BSM interpretations

PDF errors translate into phenomenological limitations

- from PDF analysis, state-of-the-art predictions for fundamental LHC observables → *e.g.*, **total cross sections at 14 TeV**





- pervasive issue beyond LHC: neutrino cross sections similarly PDF-limited

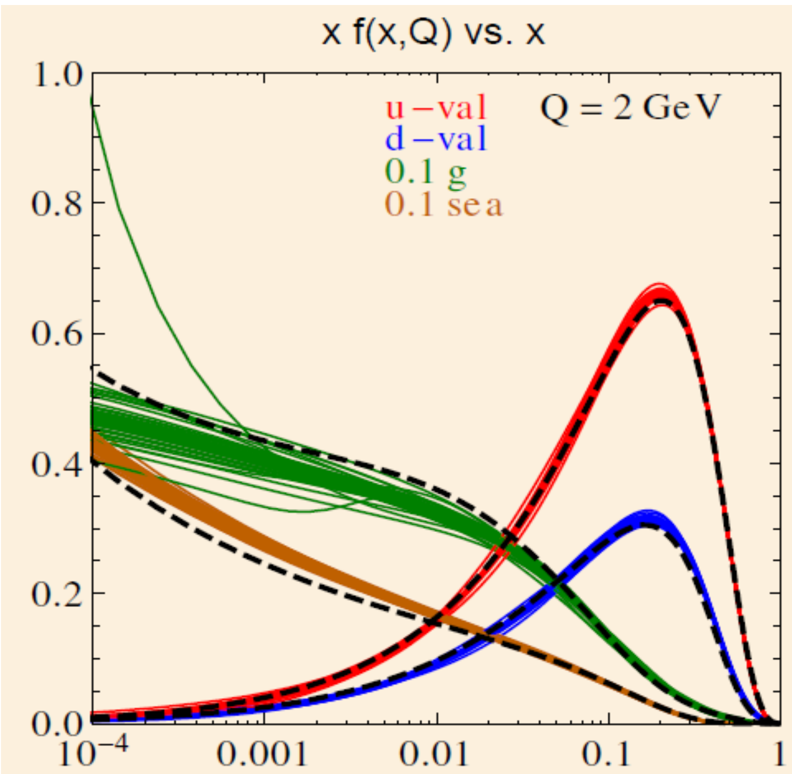
…above, for ν telescopes; analogous PDF uncertainties at low energies relevant for DUNE

interface with MC event generators, experimental interpretation

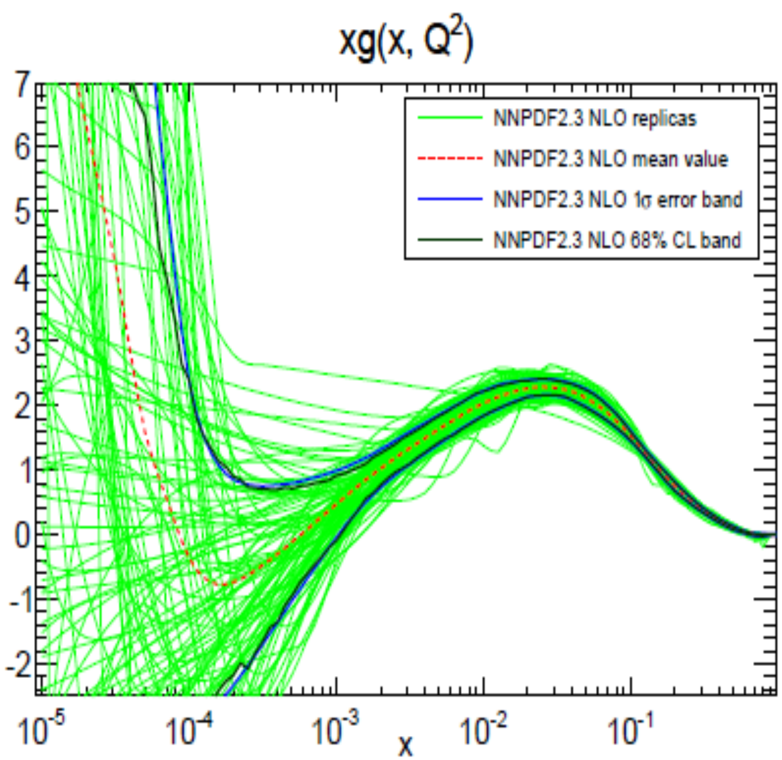
two types of modern PDF analysis approaches

Two powerful, complementary representations.

- Analytic parametrizations +
- Hessian PDF eigenvector sets **(ABM, CTEQ, HERA, MMHT,…)**



Neural network parameterizations + Monte Carlo PDF replicas **(NNPDF)**



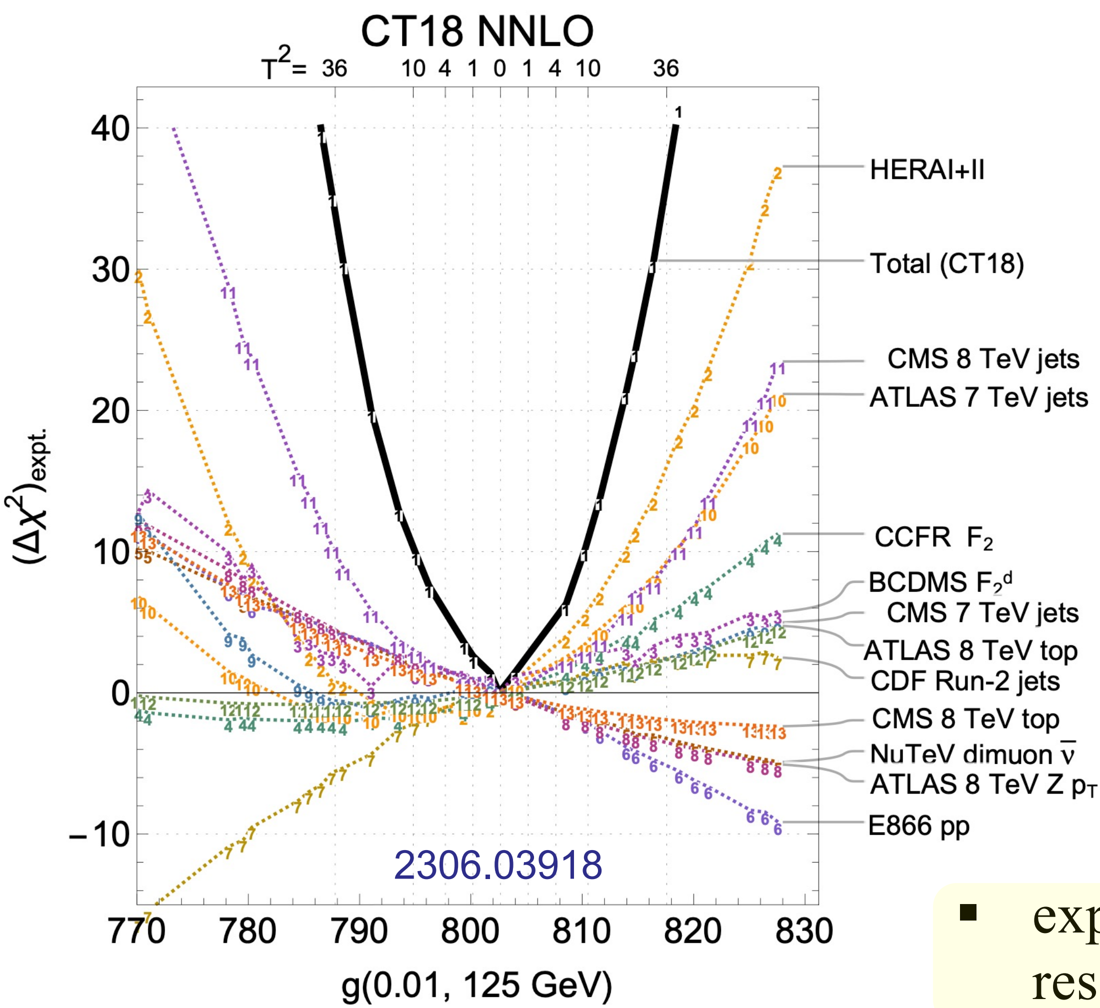
Hessian PDFs can be converted into MC ones, and vice versa.

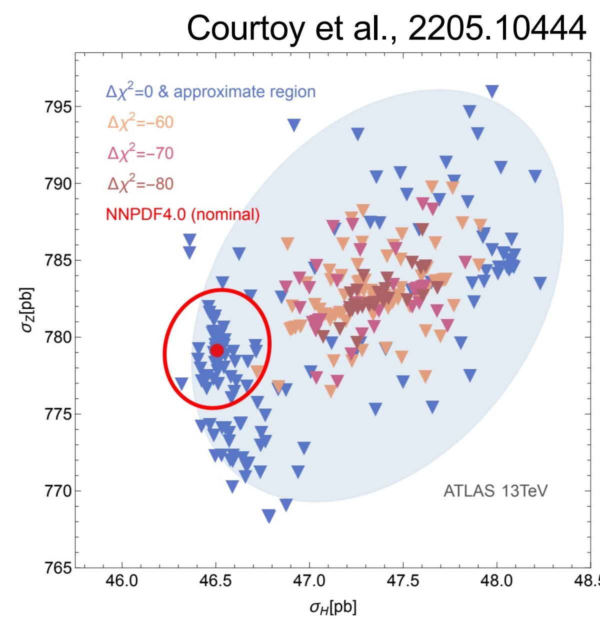
reproducible, robust PDF uncertainties

[see talk, Nadolsky]

- community-wide interest in quantifying PDF uncertainties

→ quandaries in stat theory (e.g., MC sampling challenges); explore large model spaces

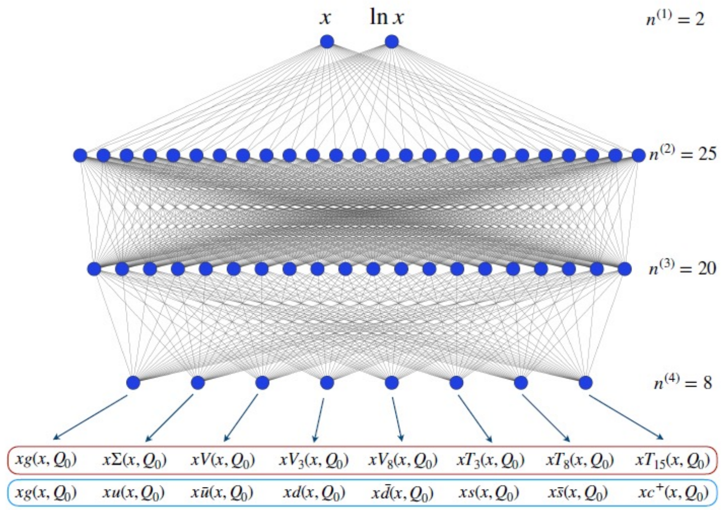




- need first-principles research in PDF uncertainties

- exploration of relation(s) among resampling, Hessian, ML, …, methods

neural network status and open questions



NNPDF, arXiv: 2109.02653

- various potential pitfalls exist

→ systematics of inference and UQ can differ relative to traditional methods

- dissect PDF uncertainties?
  aleatoric, epistemic, distributional, …

- <u>more broadly</u>: can ML tools be used to quantify correspondence among PDF parametrizations?

→ might such models be used generatively, to produce PDFs?

→ relation to previously studied PDF-lattice analyses? 1904.00022 [hep-ph]

question: can we 'stress test' ML models for PDFs?

specifically, ability to parametrize and interpolate..

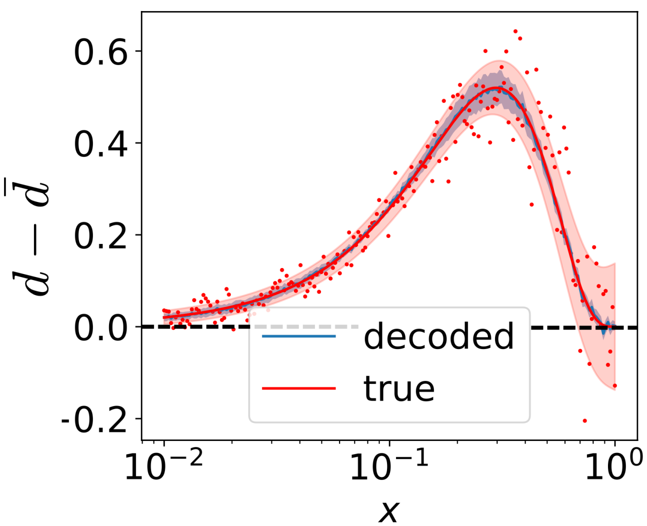
arXiv:2312.02278

ANL-186490

Learning PDFs through Interpretable Latent Representations in Mellin Space

Brandon Kriesten and T. J. Hobbs

High Energy Physics Division, Argonne National Laboratory, Lemont, IL 60439
(Dated: December 6, 2023)



→ this work: **toy demonstration**;

→ many unanswered questions to explore

→ introduces numerical platform (`PDFdecoder`)

[public code available shortly]

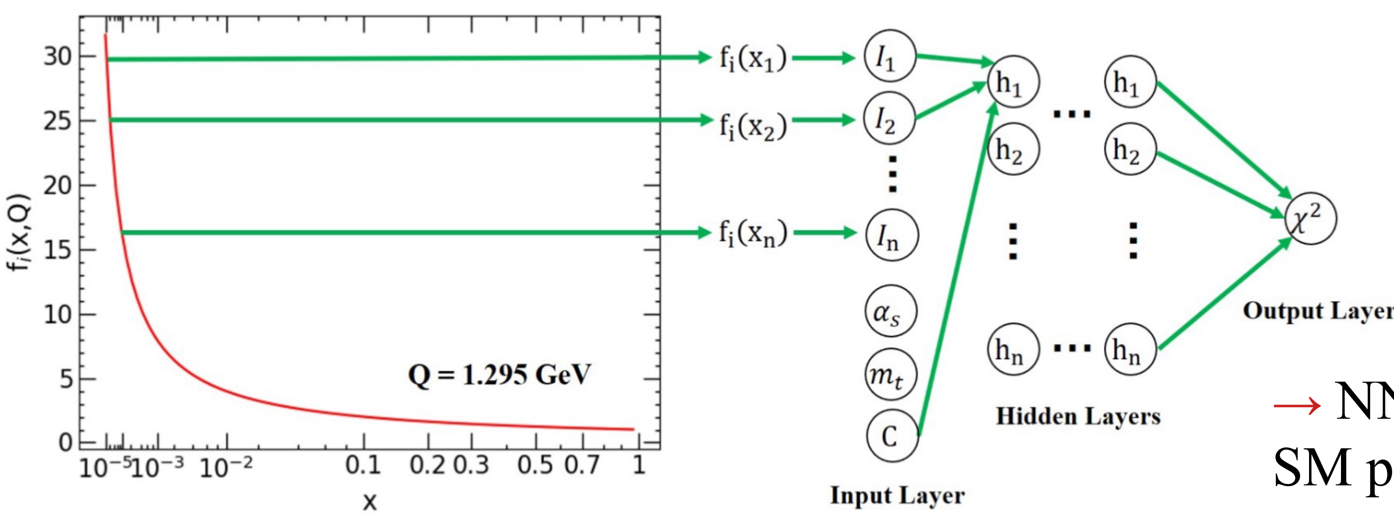
learning likelihoods with NNs

- quantify agreement of theory/data through χ^2 :

$$\chi^2(\{a_\ell\},\{\lambda\}) = \sum_{k=1}^{N_{\rm pt}} \frac{1}{s_k^2}\left(D_k - T_k(\{a_\ell\}) - \sum_{\alpha=1}^{N_\lambda}\beta_{k,\alpha}\lambda_\alpha\right)^2 + \sum_{\alpha=1}^{N_\lambda}\lambda_\alpha^2$$

$\rightarrow$ train a feed-forward neural network (NN) on PDF replicas

Liu, Sun, and Gao; arXiv: 2201.06586



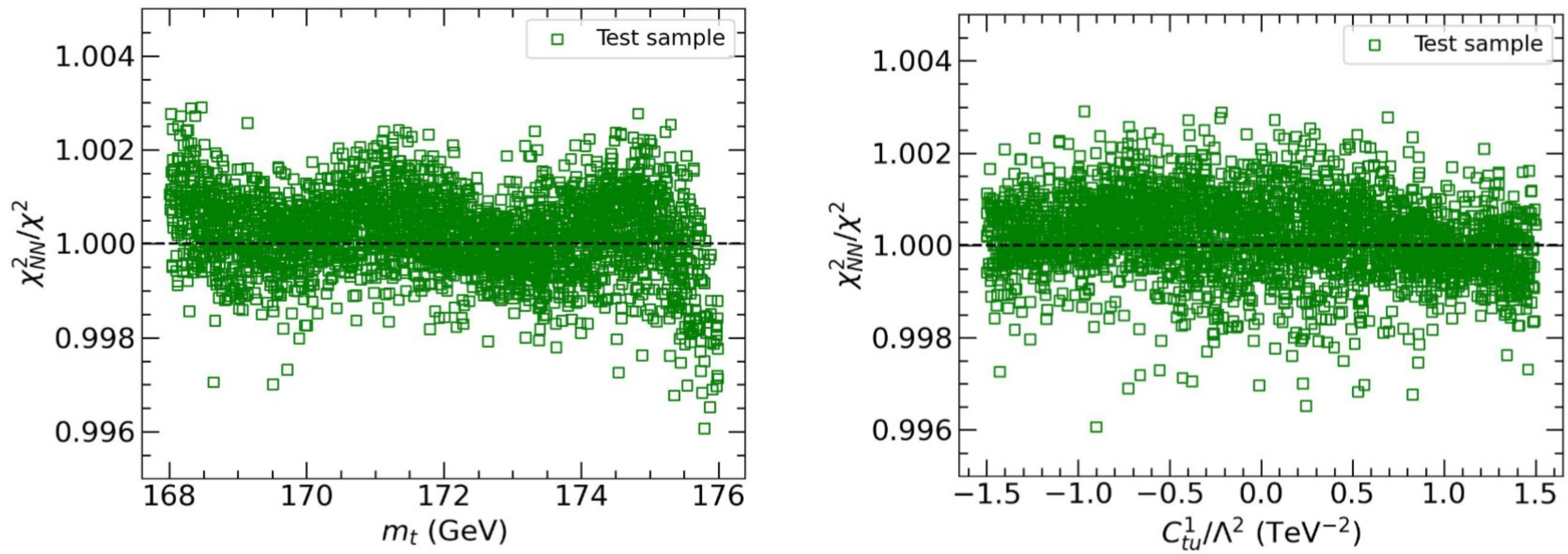
$\rightarrow$ NN associates PDFs, SM parameters, SMEFT coefficients with χ^2

NNs effectively learn (PDF-SMEFT) likelihood function

- generate 1.2×10^4 replicas over PDFs, SM parameters, SMEFT coeffs.

  $\rightarrow$ validate performance on 4×10^3 test set

2211.01094 [hep-ph]



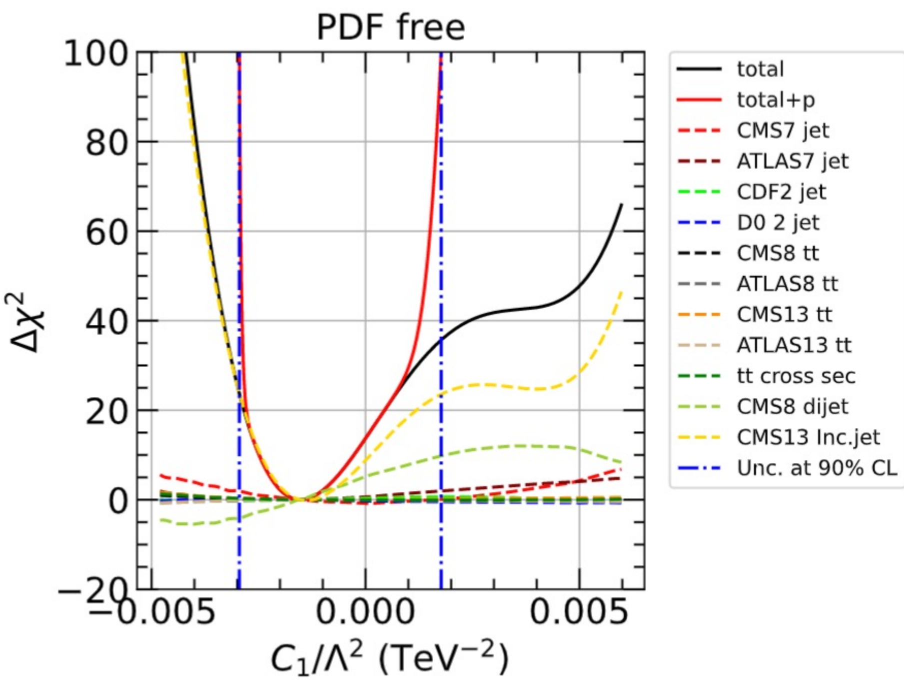
$\rightarrow$ strong, permille-level agreement achieved!

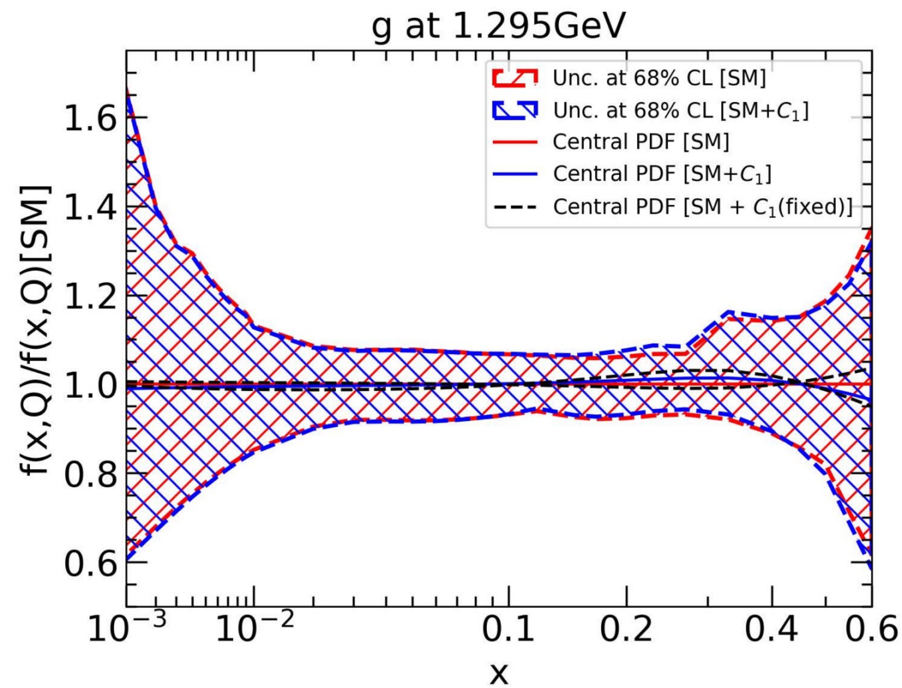
(NB: perfect agreement corresponds to $\chi^2_{\rm NN}/\chi^2 = 1$)

- allows *rapid* exploration of combined PDF-SMEFT uncertainties

(aside) NN parametrization allows fast PDF+BSM fits

- jet data modestly sensitive to C_1 (4-fermion contact interaction)





- evidence of very weak correlations between EFT parameter, high-x gluon PDF

- fixing PDFs: slight underestimate of EFT (Wilson coeff.) uncertainty

TeV^{-2}	nominal	CMS 8 dijet	CMS 8 jet	CMS 13 jet
PDF free	$-0.0015^{+0.0033}_{-0.0014}$	$-0.0022^{+0.0187}_{-0.0054}$	$-0.0009^{+0.0138}_{-0.0045}$	$-0.0013^{+0.0059}_{-0.0016}$
PDF fixed	$-0.0015^{+0.0024}_{-0.0014}$	$-0.0022^{+0.0180}_{-0.0051}$	$-0.0009^{+0.0131}_{-0.0049}$	$-0.0013^{+0.0026}_{-0.0015}$

NNs for PDF parametrization

- previous learning task was associational

  $\rightarrow$ *i.e.*, connecting parameter (or PDF) values to a target function, χ^2

  $\rightarrow$ can then bypass challenging, expensive theory calculations

- testbed to explore NN-based parametrizations of PDFs (?)

  $\rightarrow$ instead of NN proceeding to a target, telescope back to original input

  $\rightarrow$ this is a *reconstruction* task

  $\rightarrow$ large class of ML models for this

  $\rightarrow$ simultaneously provide arena to explore systematics of training, etc.

  hyperparameter optimization, training procedures, network structure, activation function choices, ….

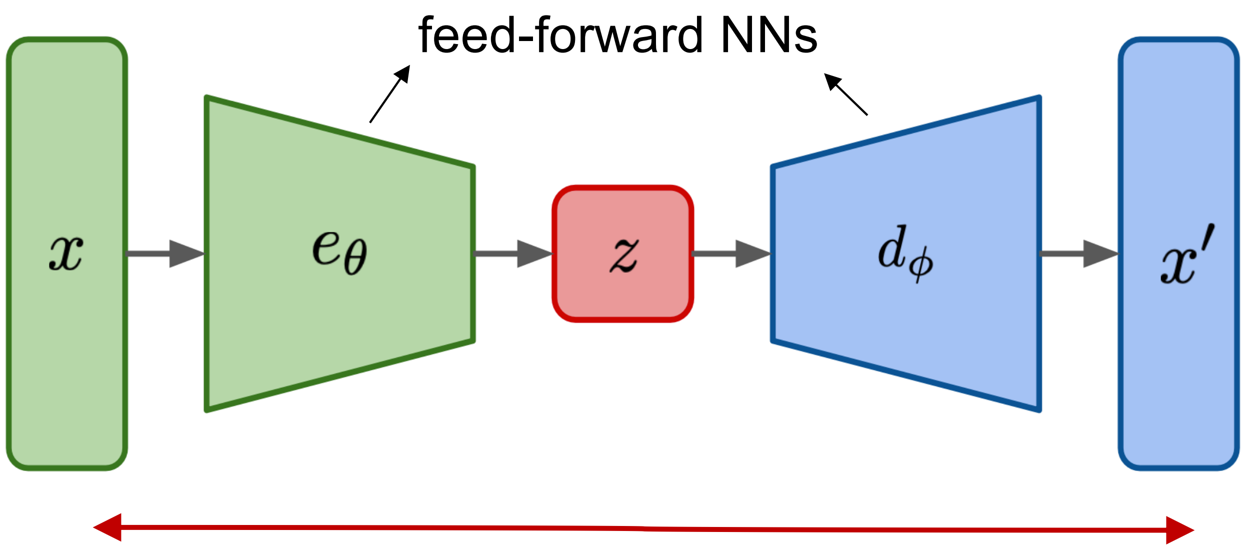
$$x_{k+1} = A_k\big(W_k\, x_k + b_k\big)$$

(rough NN analogue of Hessian parametrization dependence)

PDF reconstruction: autoencoder

- basic structure: *encoder* takes input space, *x*, to latent vector, *z*

  → corresponding *decoder* maps latent, z, to decoded output, *x*'



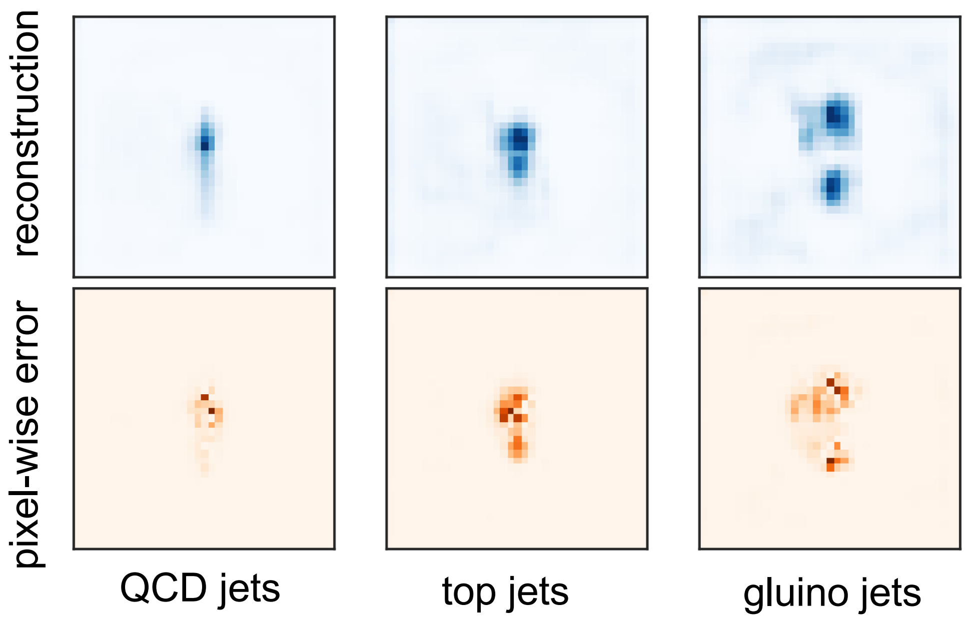
$$\mathcal{L} = \|x - d_\phi(e_\theta(x))\|_2^2$$

- *undercomplete* network structure

  → latent space of lesser dimensional size than input (dimensionality reduction)

(aside) autoencoders in HEP phenomenology

- ML dimensionality reduction --- applicable in 'big data' contexts
- anomaly detection
  - → distinguish between in- *vs* out-of-distribution (ID vs OOD) behavior
  - → *e.g.*, train model on SM-only baseline events; BSM appears OOD



Farina, Nakai, Shih: 1808.08992

- QCD analysis inverse problems

Almaeen, Alanazi, Sato, Melnitchouk, Li (2022)

base PDF reconstruction problem

- how might such an ML model encapsulate PDFs?

- generate a sizable training validation set for SU(2) toy problem

$$q(x) \pm \bar{q}(x) \;=\; \mathcal{N}_{q^\pm}\, x^{\alpha_{q^\pm}} (1-x)^{\beta_{q^\pm}}\, \mathcal{P}_{q^\pm}(x) \qquad q = u, d$$
$$\mathcal{P}_{q^\pm}(x) \;=\; 1 + \gamma_{q^\pm}\sqrt{x} + \delta_{q^\pm} x$$

- sample this basic form to obtain 10,000 MC PDF replicas

…impose number sum rules, sample parameters over uniform distribution, …

→ evaluate each replica a 196 x-values per flavor/charge combination (784 total)

$$x \in [10^{-2}, 0.999]$$

→ the resulting set of x-dependent PDF values are the inputs to reconstruct

split 10^4-member set 70/15/15 for training, validation, testing

imposing physics logic on ML model

- could stop at previous slide, explore autoencoder PDF reconstruction
- but what about interpretability of the model?

  → typically requires imposing some structure, constraint on intermediate latent

  → idea: use PDFs' x-integrated Mellin moments to organize latent space,

$$\langle x^{2n} \rangle_{q^-} = \int_0^1 dx x^{2n} \; [q(x) - \bar{q}(x)]$$

$$\langle x^{2n+1} \rangle_{q^+} = \int_0^1 dx x^{2n+1} \; [q(x) + \bar{q}(x)]$$

  → inspired by recent physics-informed (e.g., equivariant) NN development(s)

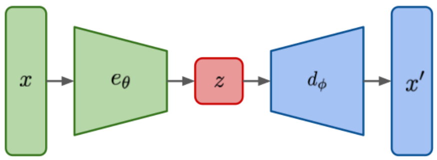
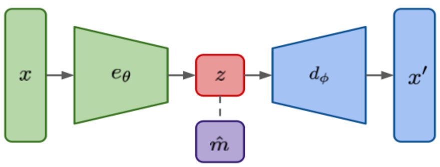
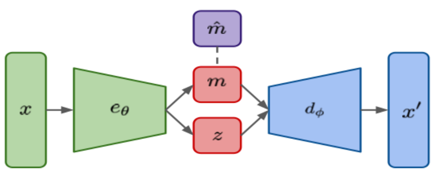
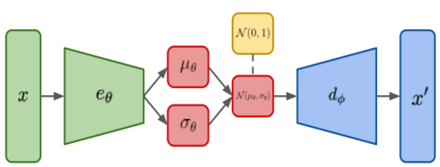
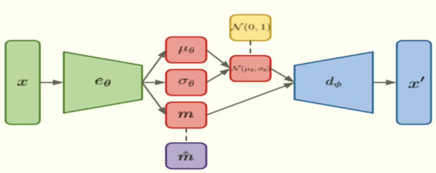
  → in addition to (PDF) reconstruction loss, there is a moment reconstruction loss

- during training, build into network constrained behavior for:

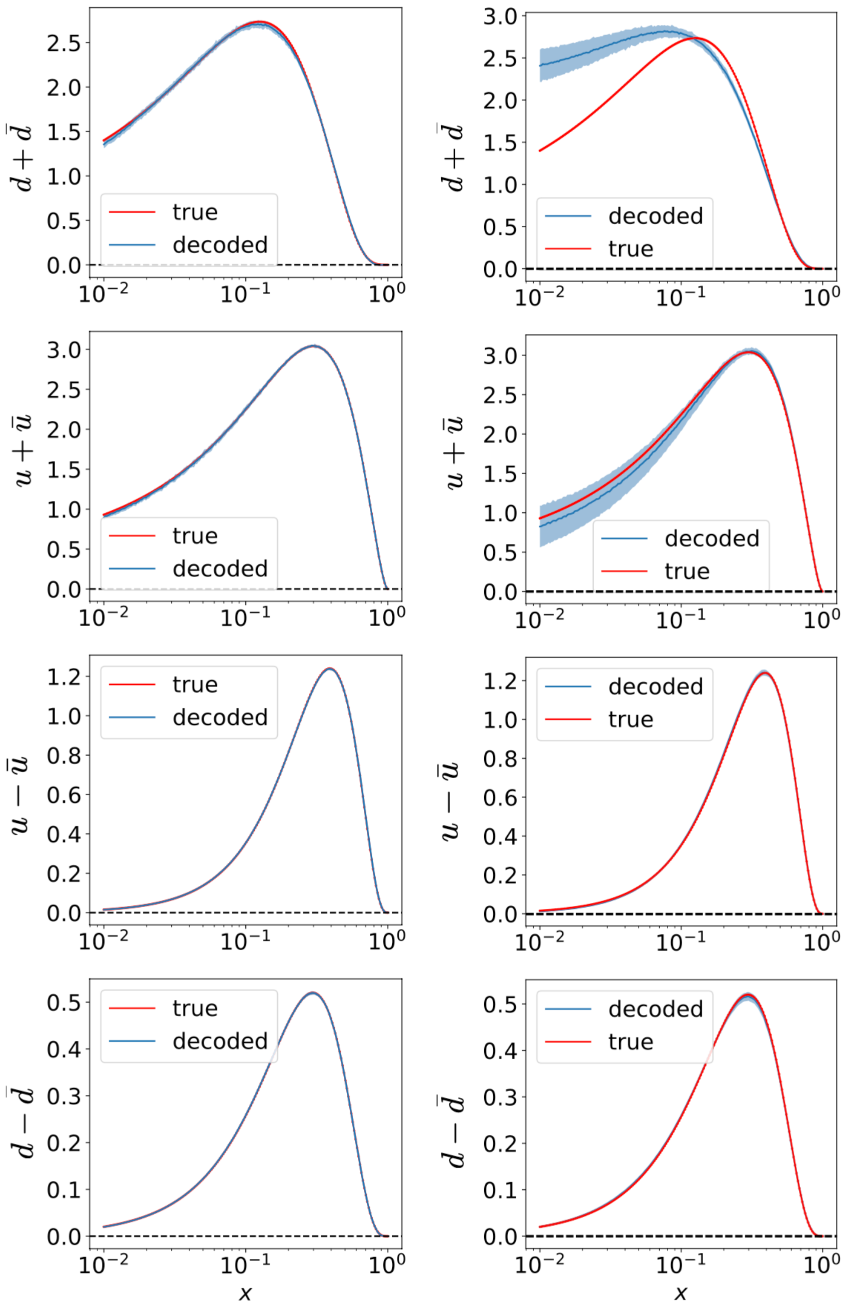
$$\langle 1 \rangle_{u^-}, \langle x \rangle_{u^+}, \langle x^2 \rangle_{u^-}, \langle x^3 \rangle_{u^+}, \cdots$$

many alternative network models possible

- variations on encoder-decoder structure; different internal topologies

Name	Diagram	Loss	Recreates PDFs	Tractable Latent	Free Latent Dimension	Moment Constraint
AE		$\mathcal{L} = \lVert x - d_\phi(e_\theta(x))\rVert_2^2$	✓	✗	✓	✗
AE-CL		$\mathcal{L} = \lVert x - d_\phi(e_\theta(x))\rVert_2^2 + \lVert z - \hat{m}\rVert_2^2$	✓	✓	✓	✓
AE-WC		$\mathcal{L} = \lVert x - d_\phi(e_\theta(x))\rVert_2^2 + \lVert m - \hat{m}\rVert_2^2$	✓	✗	✓	✓
VAE		$\mathcal{L} = \lVert x - d_\phi(e_\theta(x))\rVert_2^2 + KL(\mathcal{N}(\mu_\theta, \sigma_\theta \vert \mathcal{N}(0,1))$	✓	✓	✓	✗
VAIM		$\mathcal{L} = \lVert x - d_\phi(e_\theta(x))\rVert_2^2 + \lVert m - \hat{m}\rVert_2^2 + KL(\mathcal{N}(\mu_\theta, \sigma_\theta \vert \mathcal{N}(0,1))$	✓	✓	✓	✓

trained model performance: VAIM

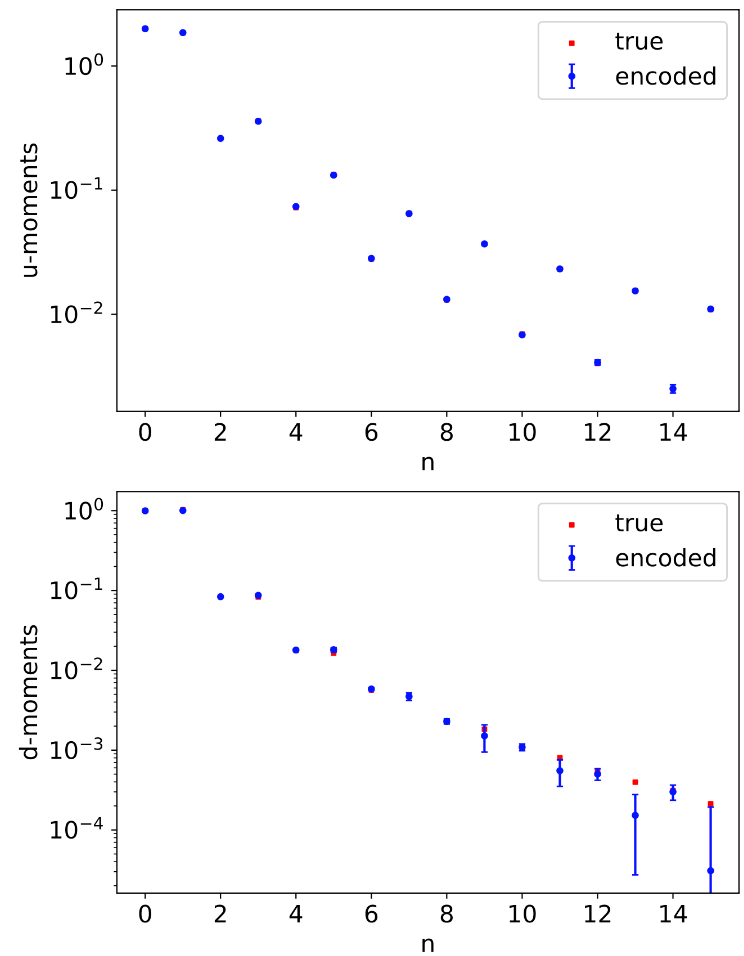


- as default, illustrate for VAIM: consistently robust reconstructions

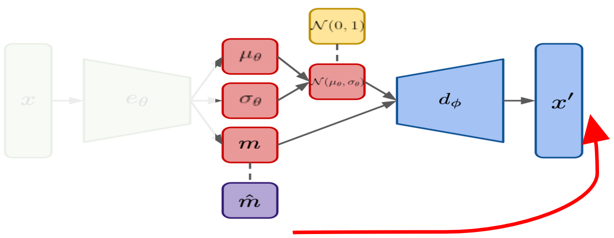
nb, open questions in UQ; ensembling [left] vs latent sampling [right];

(more in later study)

moment reconstructions, by order

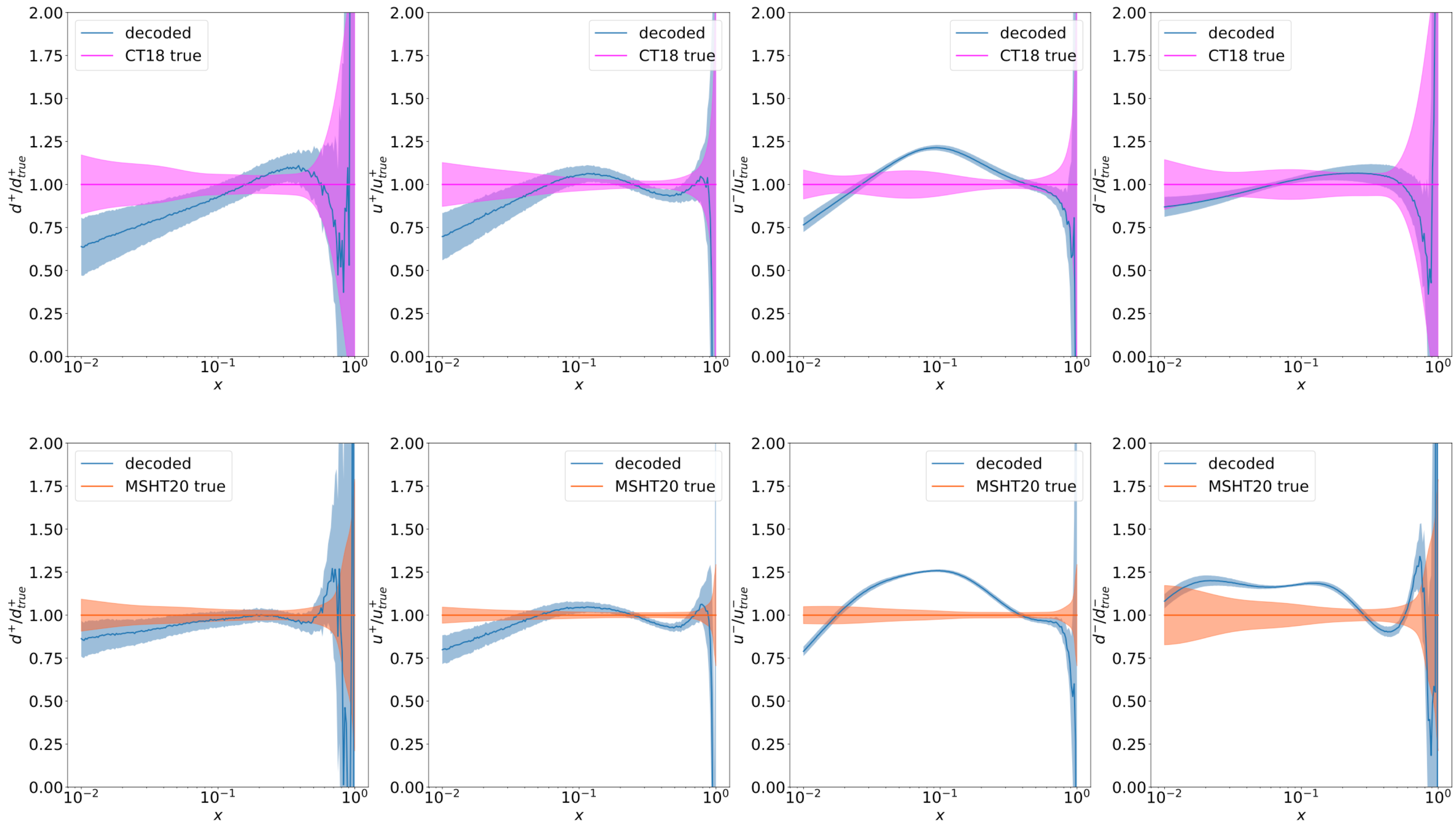


exercise: compare ML model against pheno PDFs (i)



- given imposed tractability of the latent, amputate the VAIM: [moments] → [PDFs]

→ predict pheno PDFs from their moments

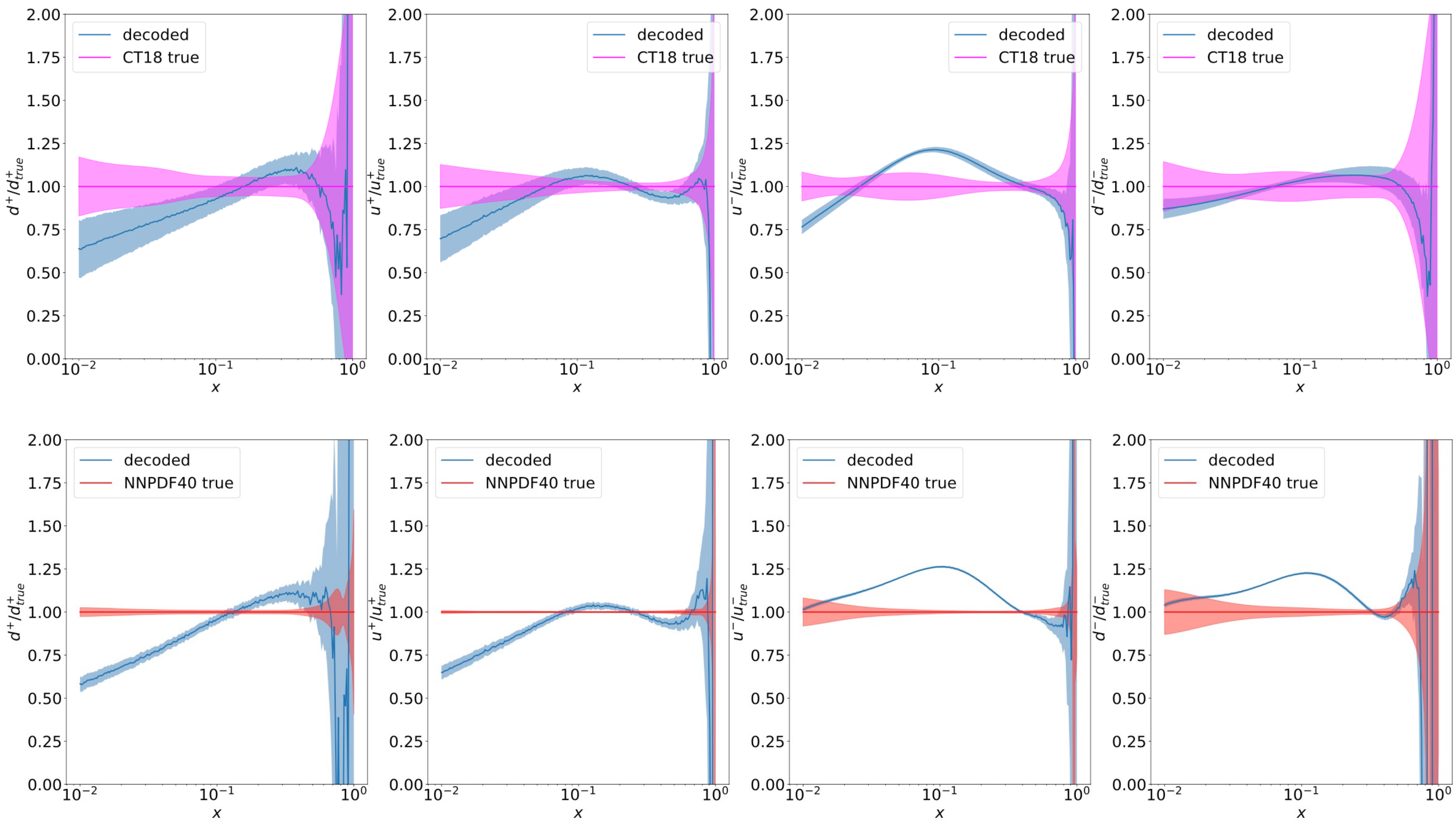


exercise: compare ML model against pheno PDFs (ii)

- the resulting framework generally results in concordance at highest x

  → can also be reimagined as an out-of-distribution detector

  (*i.e.*, quantify the parametric dissimilarity of NNPDF)

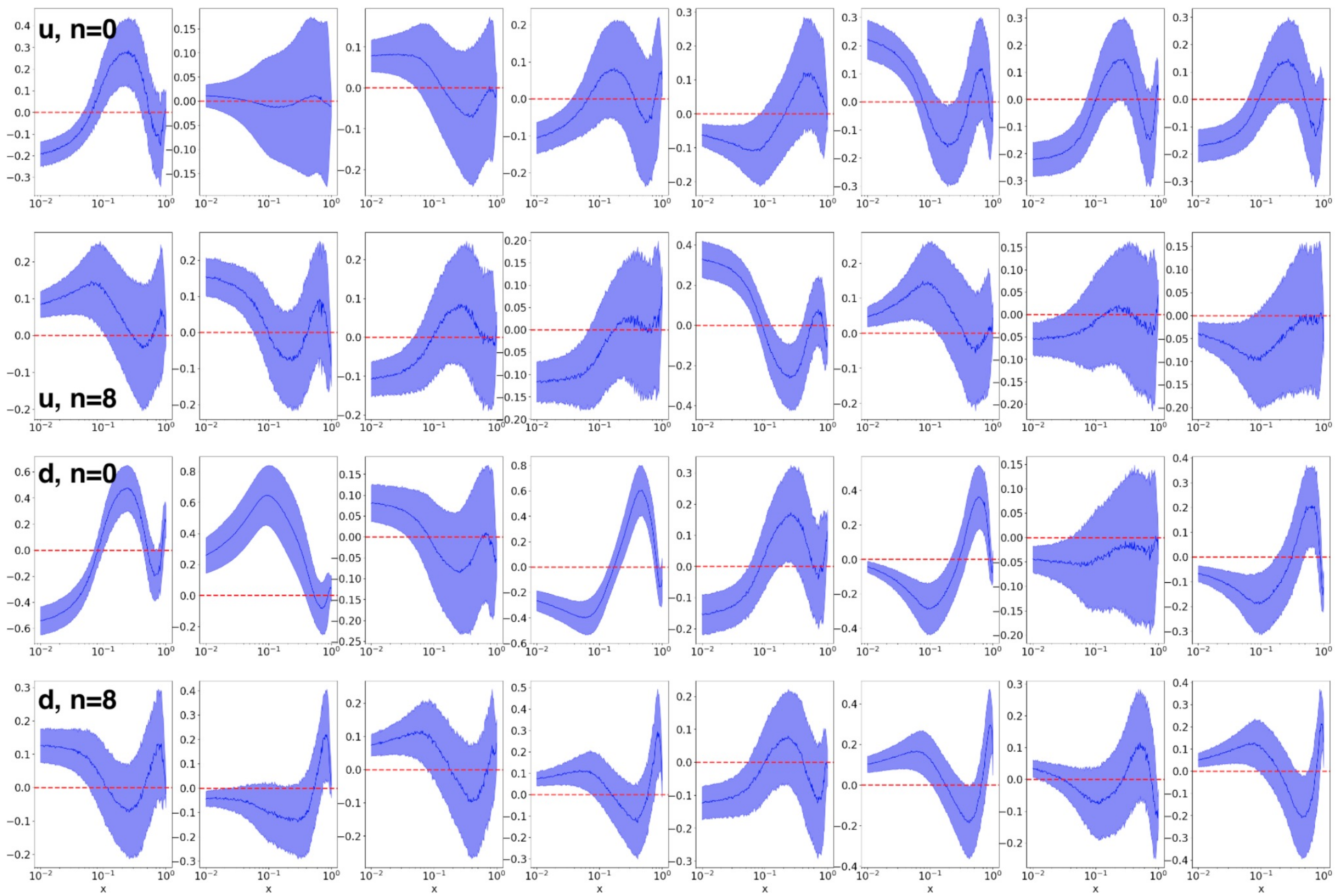


towards interpretable learning through correlations

→ tractable latent allows cross-network MC correlations (*e.g.*, latent to decoded PDF)

- default VAIM: some expected pattern of correlations; also spurious effects

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



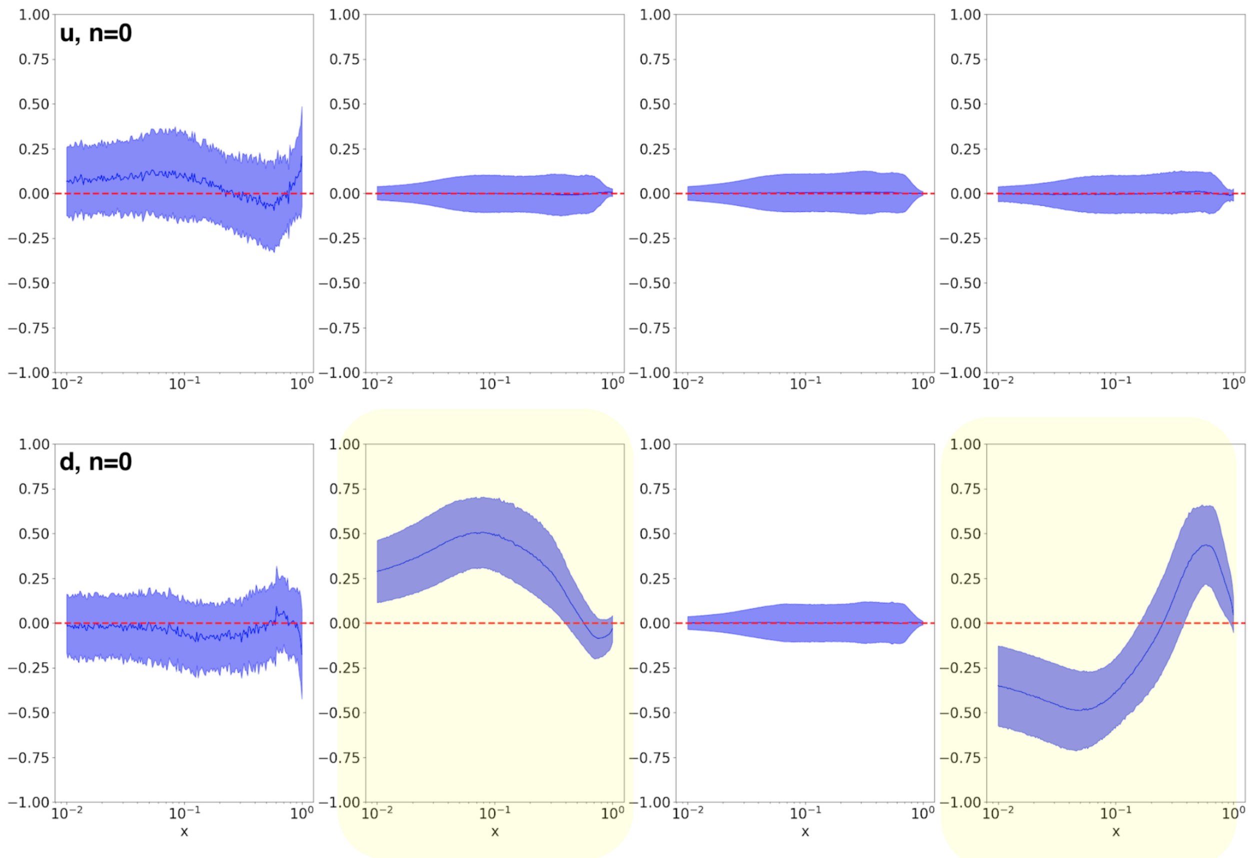
VAIM

more compressed latent space

- can demonstrate via more undercomplete network (8-dim latent)

→ greater dimensionality reduction: tame spurious moment encodings of x dependence

(statistically nonzero correlations only with d^+ moments)



VAIM

more compressed latent space

- variational AE, VAIM methods encapsulate ‘realistic’ PDF model

  → illustrated in generalizable (toy) problem

  → can be improved and extended systematically

- as end-to-end framework, can incorporate more theory ingredients

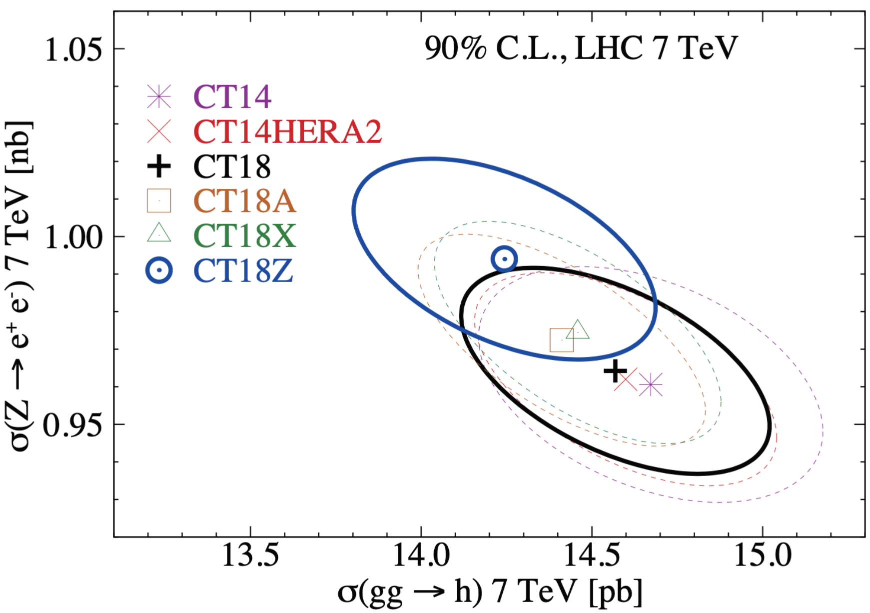
- interpretability allows model, dimensionality to be optimized

- What about “explainability” --- discussed during Week 1?

XAI: connect assumed theory to fitted PDFs

- (QCD/EW) theory settings have subtle downstream PDF implications

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	$\mu^2_{F,DIS} = Q^2$	No	Yes	1.3
CT18A NNLO	$\mu^2_{F,DIS} = Q^2$	Yes	Yes	1.3
CT18X NNLO	$\mu^2_{F,DIS} = 0.8^2 \left(Q^2 + \frac{0.3\ \mathrm{GeV}^2}{x_B^{0.3}}\right)$	No	Yes	1.3
CT18Z NNLO	$\mu^2_{F,DIS} = 0.8^2 \left(Q^2 + \frac{0.3\ \mathrm{GeV}^2}{x_B^{0.3}}\right)$	Yes	No	1.4

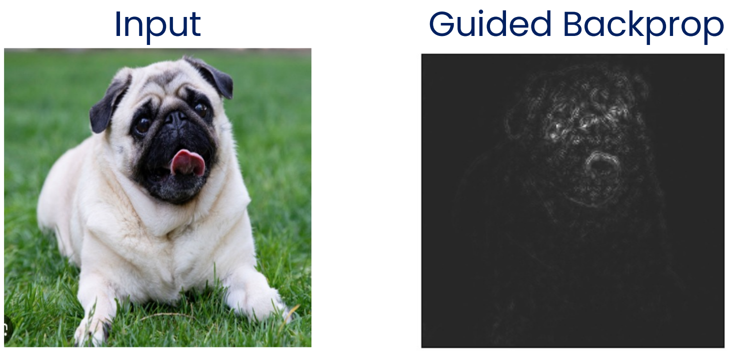


- simultaneous variations of multiple theory/analysis settings influence PDFs and pheno in hard-to-disentangle ways

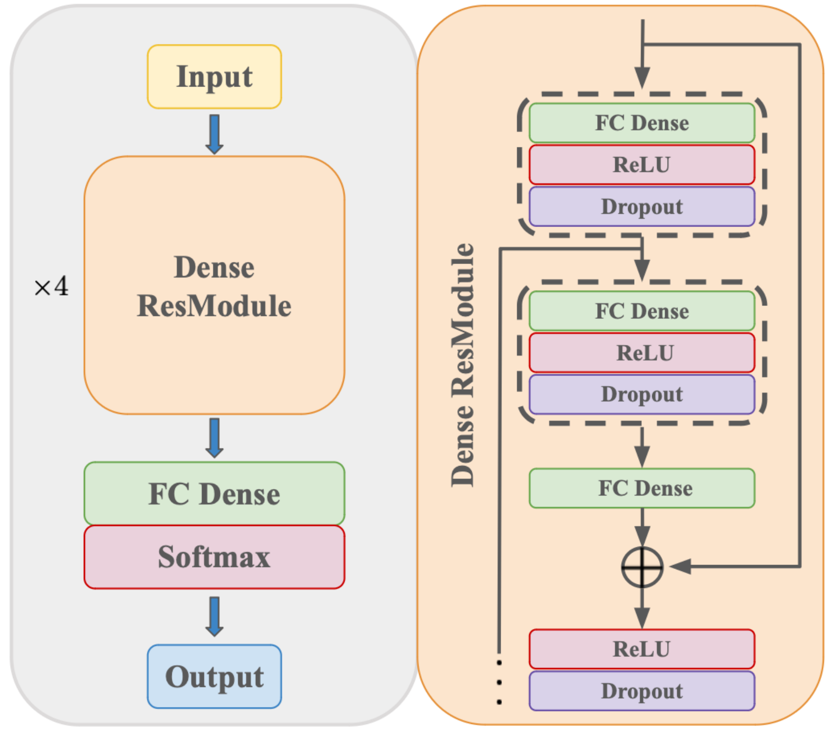
  [classical methods: L_2 sensitivity, 2306.03918]

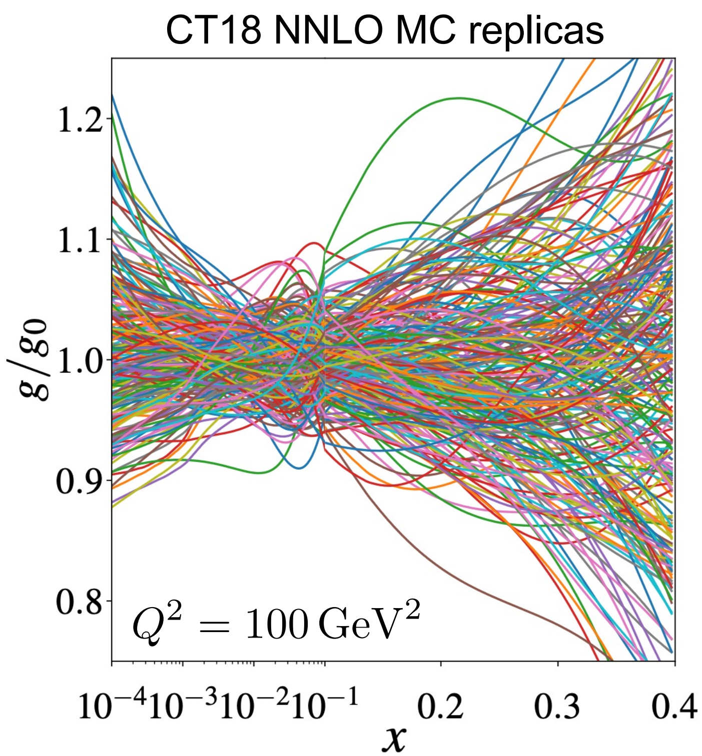
→ might AI methods provide some useful (complementary) guidance?

identify salient features: guided backpropagation



appearing soon, with Kriesten, Gomprecht



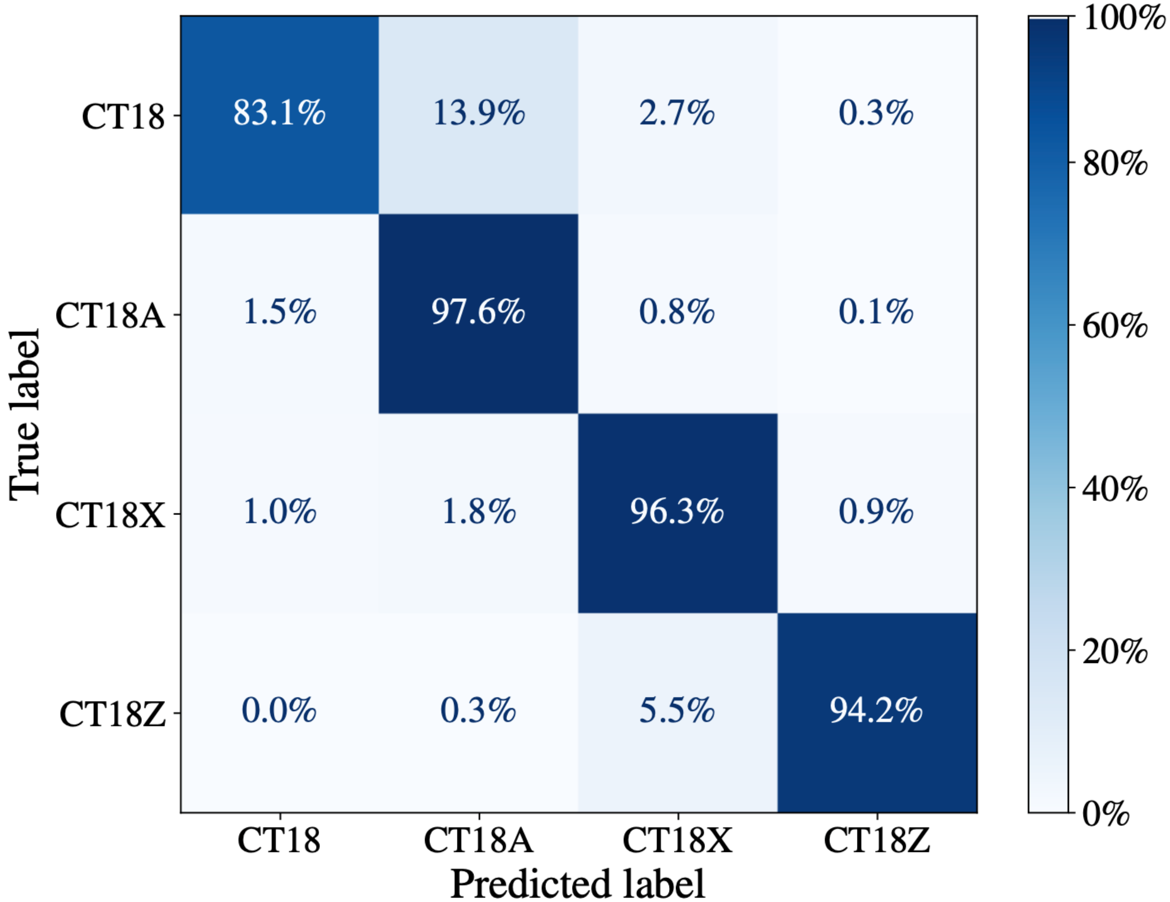


$$\frac{\partial f^{\text{out}}}{\partial f_i^\ell} = (f_i^\ell > 0) \cdot \left(\frac{\partial f^{\text{out}}}{\partial f_i^{\ell+1}} > 0 \right) \cdot \frac{\partial f^{\text{out}}}{\partial f_i^{\ell+1}}$$

train a ResNet-like model on MC PDF replicas;
backpropagate classification scores to PDF shapes

theory classification: ~few-percent accuracy

- evaluate a confusion matrix on test set of MC replicas (normalized PDF ratios)



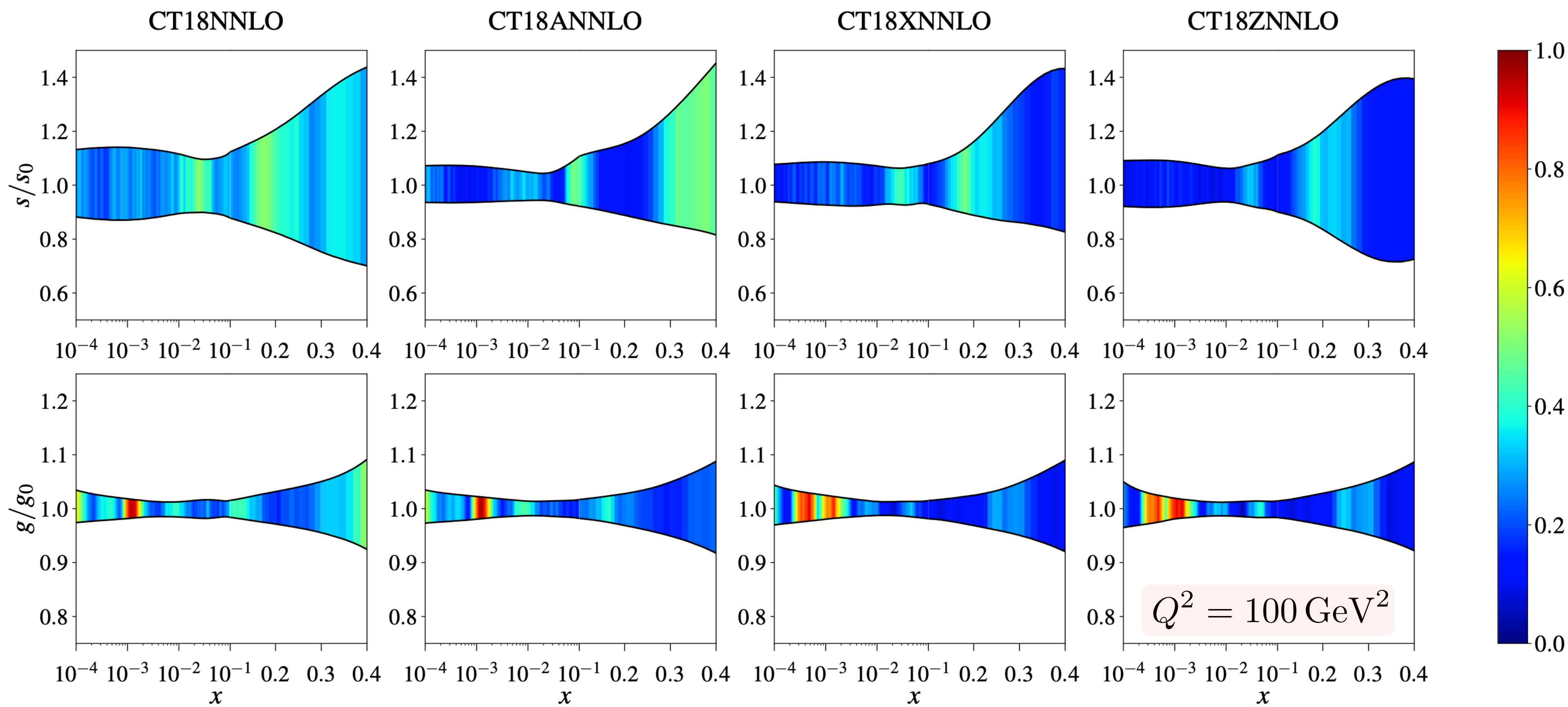
- strongly diagonal matrix at several-percent level

  → weak tendency to misidentify CT18 as CT18A

- can we *explain* the origin of these classifications in (input) feature space?

GBP: local PDF *x* dependence → classification score

- classification score gradients relative to input PDF(s); highlight salient features



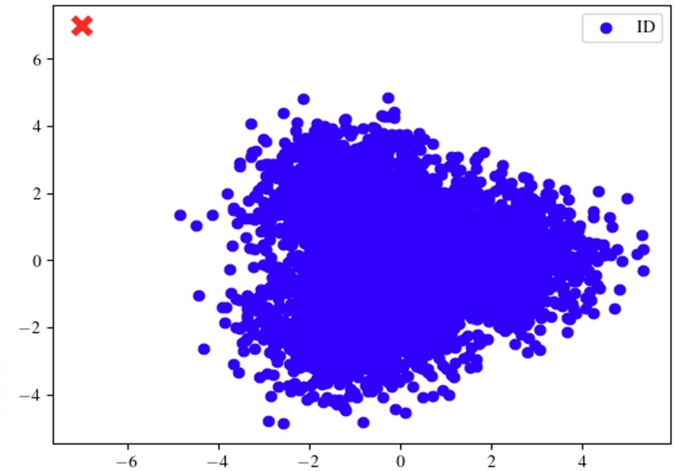
- in line with CT18 main study

→ high-, low-*x* gluon and strangeness provide most model discrimination

stay tuned: evidential learning & prior networks

(Dirichlet) Prior Networks, DPNs Malinin and Gales (2018)

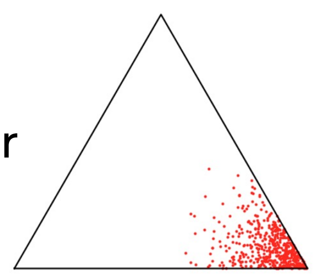
- factorize predictive posterior to model dependence on prior ensemble



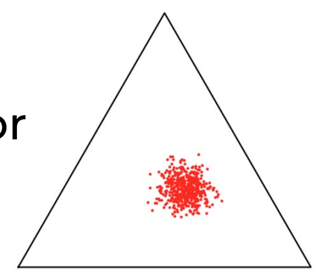
$$P(\omega_c|\mathbf{x}^\star,\mathcal{D}) = \int d\mu d\theta\; \underbrace{p(\omega_c|\mu)}_{\text{Aleatoric}}\; \underbrace{p(\mu|\mathbf{x}^\star,\theta)}_{\text{Distributional}}\; \underbrace{p(\theta|\mathcal{D})}_{\text{Epistemic}}$$

- forward pass: map aleatoric, epistemic, and distributional errors on low-dimensional simplex

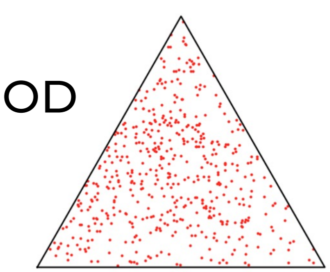
ID
low error



ID
high error

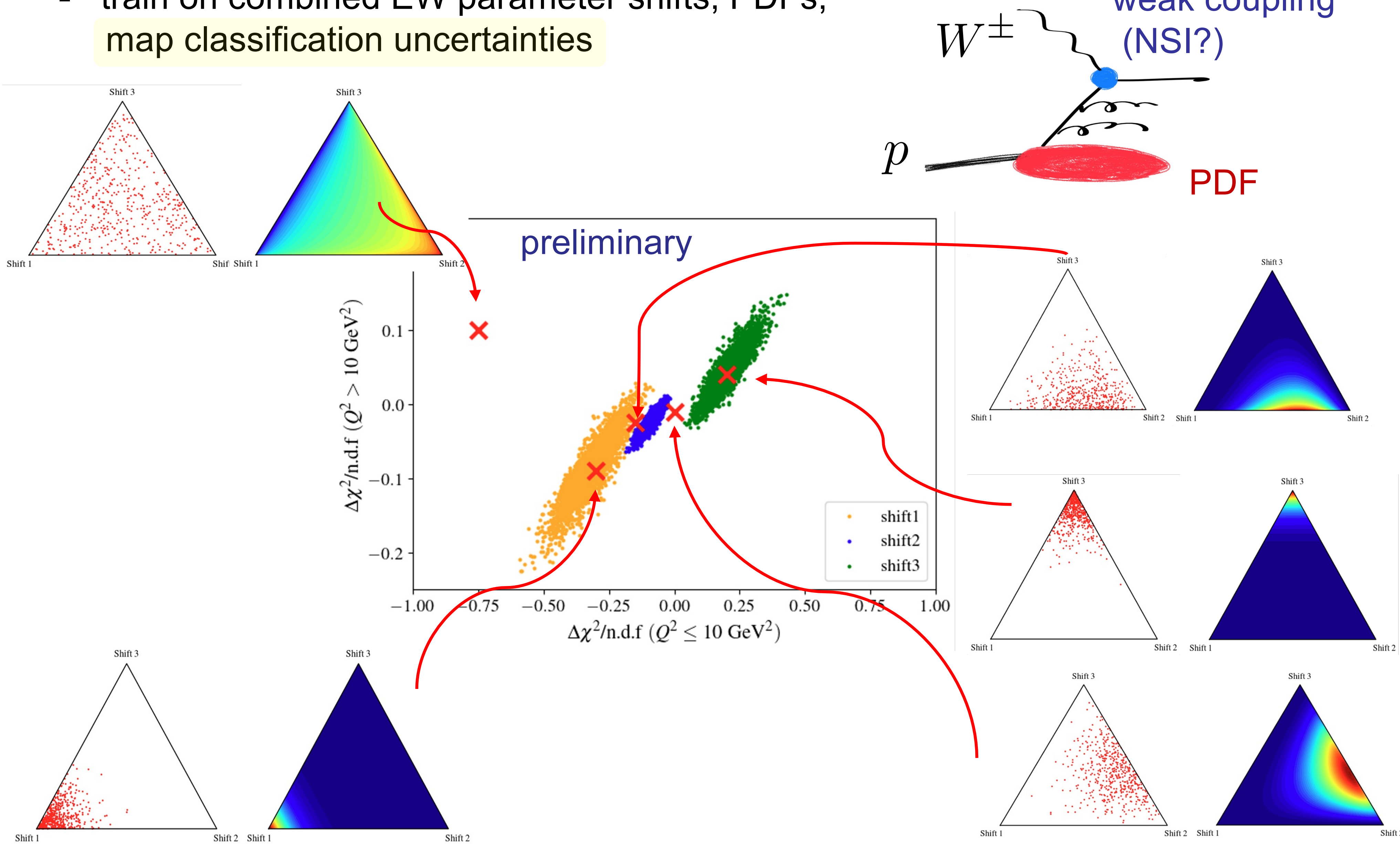


OOD



stay tuned: evidential learning & prior networks

- train on combined EW parameter shifts, PDFs;
  map classification uncertainties



conclusions, outlook

- *initial* studies on ML-model PDF realizations

  → encoder networks: flexibility with many possible variations

  → advantages to 'hardwiring' physics assumptions, symmetries

  (introduces notion of tractability; use as generative model)

- calculations, toolset mesh with precision QCD PDF theory program

- various pitfalls, some familiar

  → challenges in UQ (forthcoming analyses), <u>mode collapse</u>

- extensions: PDF interpolation, query pheno PDFs via OOD behavior

- early XAI/PDF methods show promise; complement classical approaches

  → both offer mutual lever arm to intercompare PDF analysis methods