PDF parametrization and model uncertainty

Aurore Courtoy

Instituto de Física

National Autonomous University of Mexico (UNAM)

Based on works within the Fantômas4QCD and
CTEQ-TEA collaborations

Towards Improved Hadron Femtography

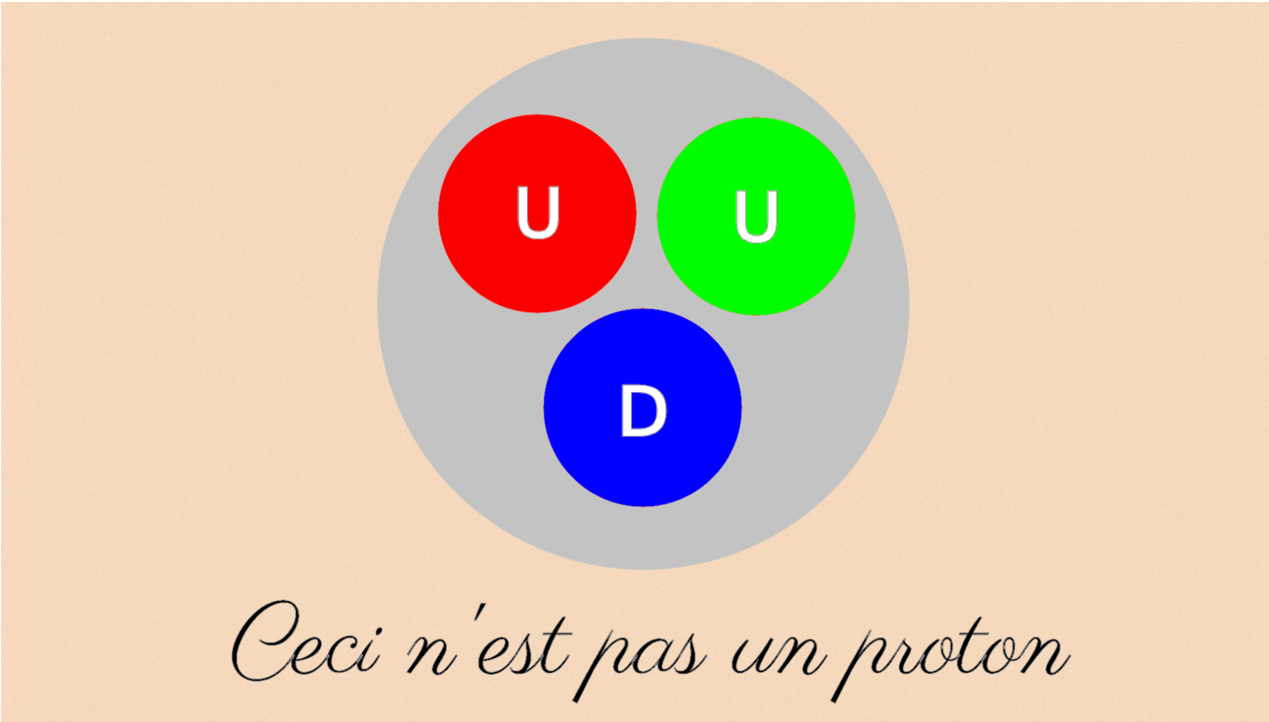
UVA

07/23/26

Dirección General de Asuntos
del Personal Académico

Ciencia y Tecnología
Secretaría de Ciencia, Humanidades, Tecnología e Innovación

Instituto de Física
UNAM



[Towards Improved Hadron Femtography, 2026]

[René Magritte, 1929]

20–24 Jul 2026
University of Virginia

Belgium turned 195 years during the workshop
21 Jul 2026 $\in [20, 24]$ Jul 2026

High(er) than usual participation rate from Belgians to the workshop;).

[I wonder about the correlation between the logo and that high rate.]

Context for model uncertainty

I'll care about **unpolarized PDFs from global analyses.**

Those PDFs are determined through global QCD analyses,

up to NNLO (and approximate N3LO now) for the nucleon
& up to NLO for the mesons

Global analyses involve **various processes**, e.g.

Hadron-hadron collision:

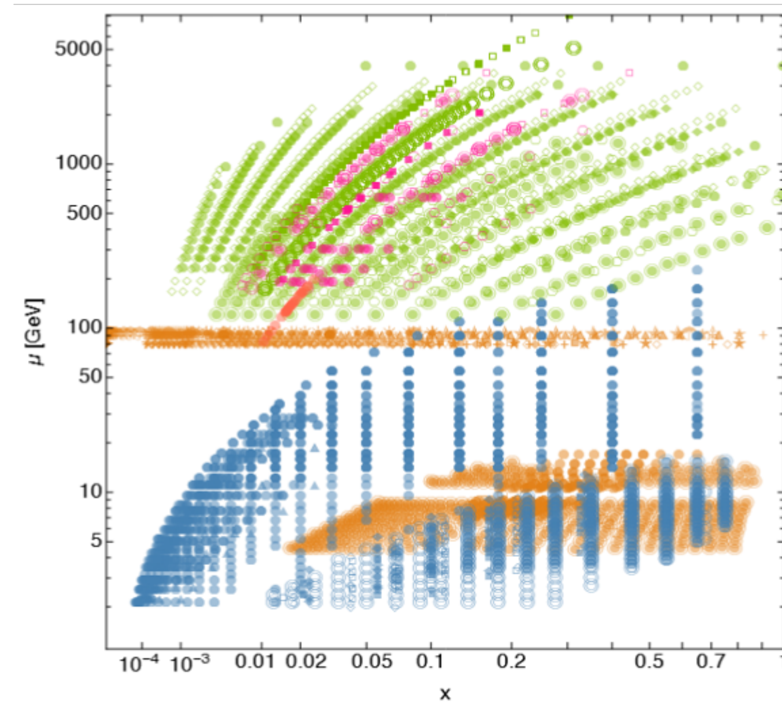
$$\sigma = \sum_{a,b} \int dx_a \int dx_b \, f_{a/A}(x_a, \mu_F^2) f_{b/B}(x_b, \mu_F^2) \, H_{a,b,x_a,x_b,\mu_F^2} + \mathcal{O}(M/Q)$$

DIS structure functions:

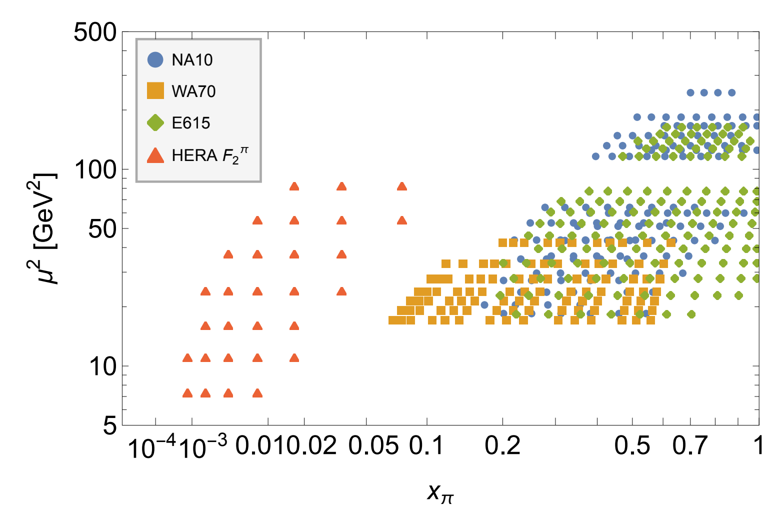
$$F(x_B, Q^2) = \sum_a \int_{x_B}^1 \frac{dx}{x} \, f_{a/p}(x, \mu^2) \, H_a\left(\frac{x_B}{x}, \frac{\mu^2}{Q^2}\right) + \mathcal{O}(M/Q)$$

precision QCD predictions,

and a procedure to tackle the **inverse problem**.



Nucleon PDFs [CT25, soon]



Pion PDFs [Fantômas, PRD109]

Context for model uncertainty

I'll care about **unpolarized PDFs from global analyses.**

Those PDFs are determined through global QCD analyses,

up to NNLO (and approximate N3LO now) for the nucleon
& up to NLO for the mesons

Global analyses involve **various processes**, e.g.

Hadron-hadron collision:

$$\sigma = \sum_{a,b} \int dx_a \int dx_b\, f_{a/A}(x_a, \mu_F^2) f_{b/B}(x_b, \mu_F^2)\, H_{a,b,x_a,x_b,\mu_F^2} + \mathcal{O}(M/Q)$$

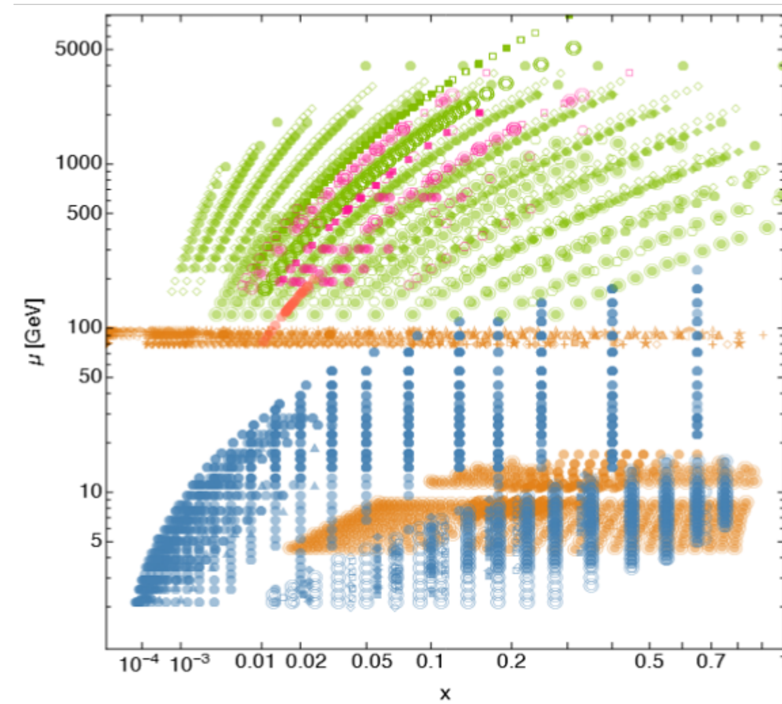
DIS structure functions:

$$F(x_B, Q^2) = \sum_a \int_{x_B}^1 \frac{dx}{x}\, f_{a/p}(x, \mu^2)\, H_a\left(\frac{x_B}{x}, \frac{\mu^2}{Q^2}\right) + \mathcal{O}(M/Q)$$

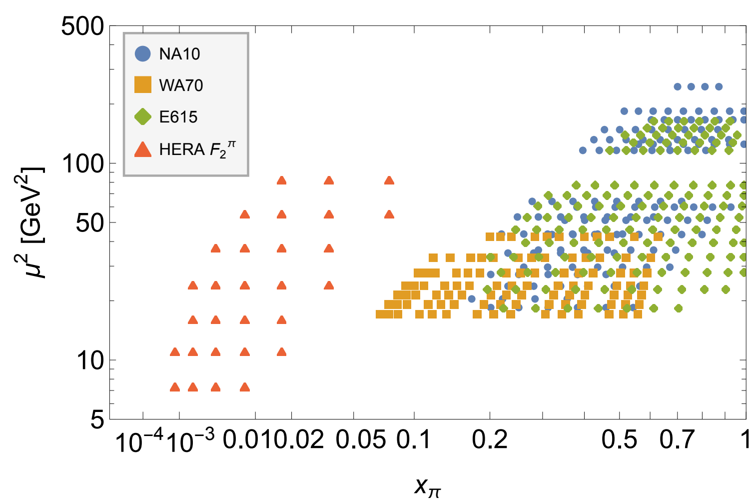
precision QCD predictions,

and a procedure to tackle the **inverse problem.**

involves model/(hyper)parameters/…



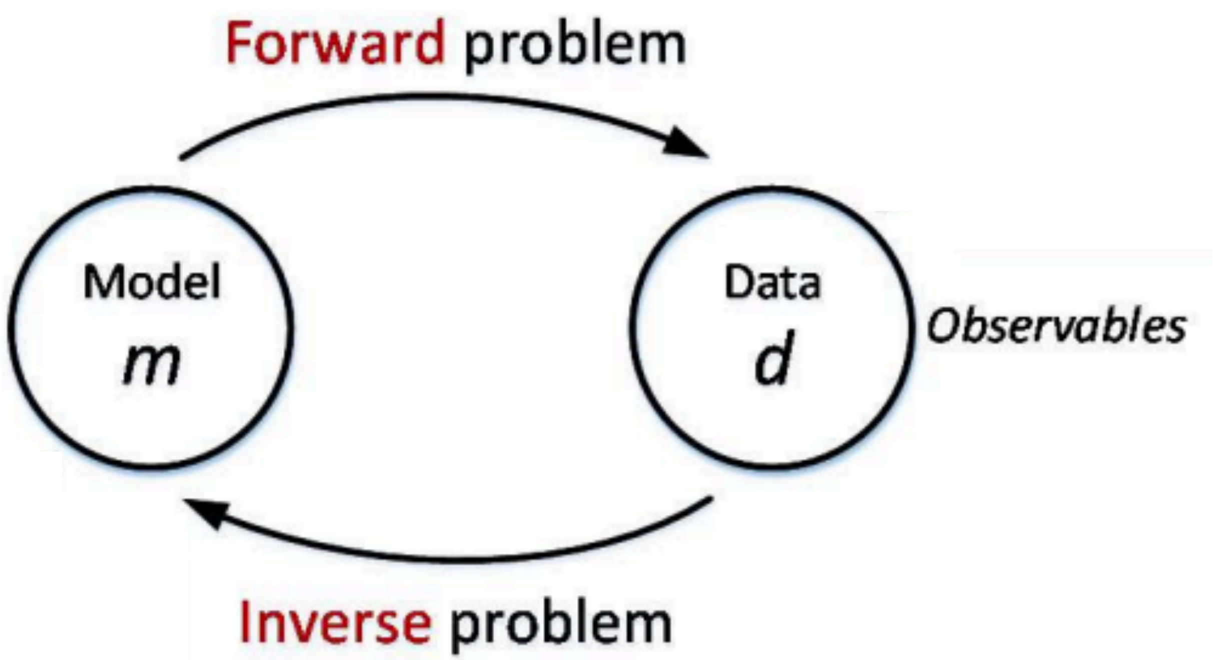
Nucleon PDFs [CT25, soon]



Pion PDFs [Fantômas, PRD109]

Inverse problems

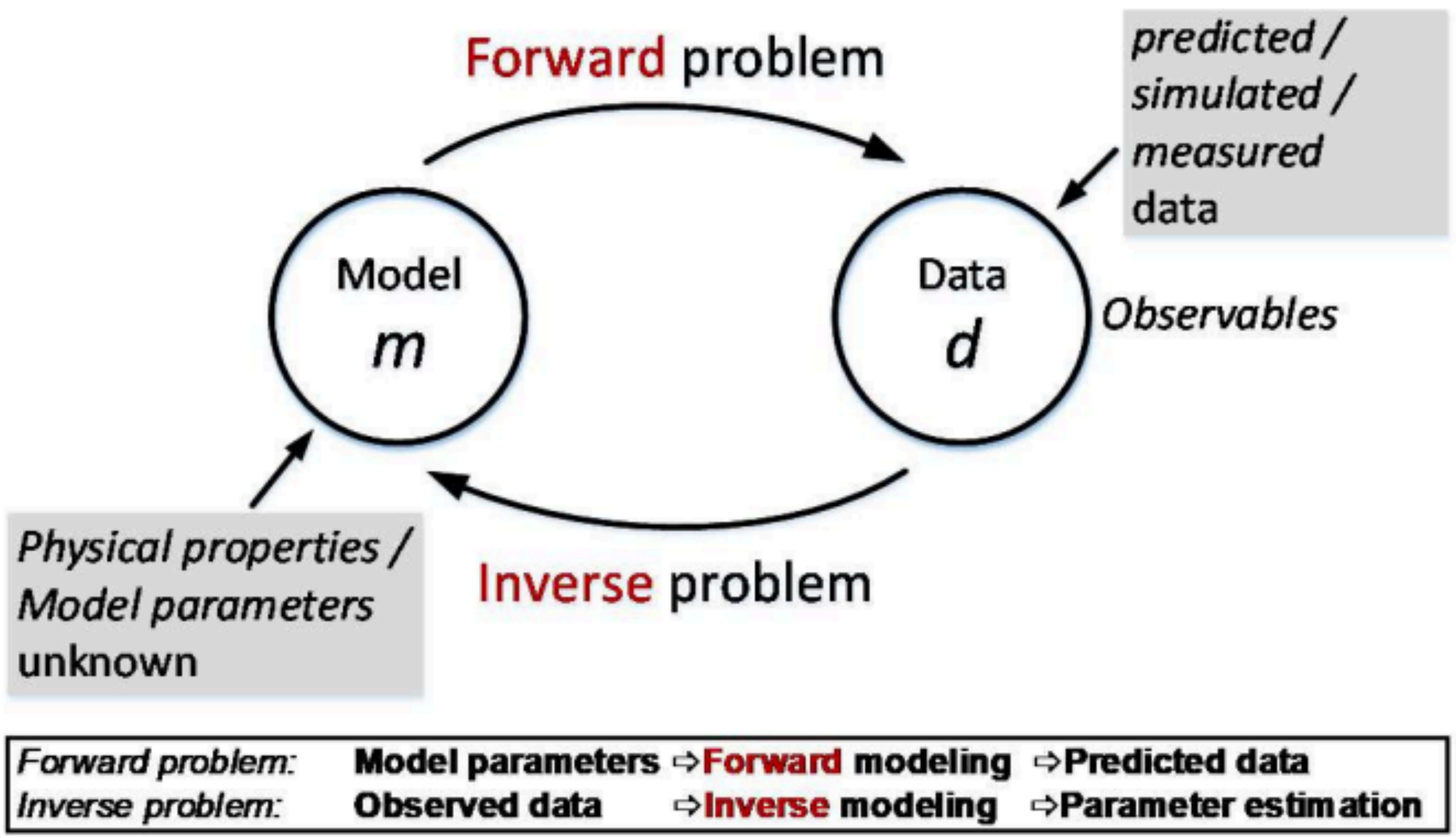
A bridge between <u>measure of information and first principles</u>



Adapted image from
[Goran D. Putnik et al. / Procedia CIRP 109]

Inverse problems

A bridge between <u>measure of information and first principles</u>



Adapted image from
[Goran D. Putnik et al. / Procedia CIRP 109]

Well-posedness of inverse problems

Hadamard conditions:

1. The problem has a solution ;
2. The solution is unique ;
3. The solution's behavior changes continuously with the initial conditions.

Well-posedness of inverse problems

Hadamard conditions:

1. The problem has a solution ;
2. The solution is unique ;
3. The solution's behavior changes continuously with the initial conditions.

Reasons for complexity of inverse problems

- direct products — $d[f] = \sigma \times f(x)$
- Kernel/convolutional problems — $d[f] = \sigma \otimes f(x)$
- Hausdorff moment problem — $< x^n > = \int dx\, x^n f(x)$

Well-posedness of inverse problems

Hadamard conditions:

1. The problem has a solution ;
2. The solution is unique ;
3. The solution's behavior changes continuously with the initial conditions.

Reasons for complexity of inverse problems

- direct products — $d[f] = \sigma \times f(x)$
- Kernel/convolutional problems — $d[f] = \sigma \otimes f(x)$ → PDF determination is often an ill-posed problem
- Hausdorff moment problem — $< x^n > = \int dx\, x^n f(x)$

Glossary for parameterizations

$f_{a/p}(x, \mu^2)$ **is the unknown function on $x \in [0,1]$ that we would like to infer.**

We choose models on a basis that (may) lead to solutions:

Basis

(orthogonal) polynomials $P_n[g(x)]$ —defined through its generating function,
neural networks $\mathscr{F}_\theta(x)$ —defined by activation functions

Models

Choice of structure, defined by **hyper parameters**
For polynomials: degree N and function $g(x)$
For neural networks: number of layers, design choices,…

Solutions

Found through "trainable parameters"$\Longrightarrow$ the parameter inference part

Glossary for parameterizations

$f_{a/p}(x, \mu^2)$ **is the unknown function on $x \in [0,1]$ that we would like to infer.**

We choose models on a basis that (may) lead to solutions:

Basis

(orthogonal) polynomials $P_n[g(x)]$ —defined through its generating function,
neural networks $\mathscr{F}_\theta(x)$ —defined by activation functions

Models

Choice of structure, defined by **hyper parameters**
For polynomials: degree N and function $g(x)$
For neural networks: number of layers, design choices,…

Solutions

Found through "trainable parameters"$\Longrightarrow$ the parameter inference part

Example

Basis: Tchebyshev polynomials
One model: degree $N = 6$ and $g(x) = x$
Solutions for this model: all N_{flavor} sets of parameters $\{A'_q, B_q, C_q, \underline{c}\}$ with $\underline{c}$ the Tchebyshev coefficients

Regression

Entails the definition of an optimization framework with

- a loss/objective function
- models
- criteria for goodness-of-fit
- criteria for uncertainty quantification (UQ)

Regression

Entails the definition of an optimization framework with

- a loss/objective function
  (log-) likelihood and priors or penalties, treatment of systematic uncertainties
- models
  Parametric forms, model sampling, architecture definition, *first principle constraints*
- criteria for goodness-of-fit
  metric, closure tests
- criteria for uncertainty quantification (UQ)
  spelling out the error budget

Regression

Entails the definition of an optimization framework with

- a loss/objective function

  (log-) likelihood and priors or penalties, treatment of systematic uncertainties
- models

  Parametric forms, model sampling, architecture definition, *first principle constraints*
- criteria for goodness-of-fit

  metric, closure tests
- criteria for uncertainty quantification (UQ)

  spelling out the error budget

Interpret the space of solutions ⇒ implication on the result

Regression

Entails the definition of an optimization framework with

- a loss/objective function
  (log-) likelihood and priors or penalties, treatment of systematic uncertainties
- models
  Parametric forms, model sampling, architecture definition, *first principle constraints*
- criteria for goodness-of-fit
  metric, closure tests
- criteria for uncertainty quantification (UQ)
  spelling out the error budget

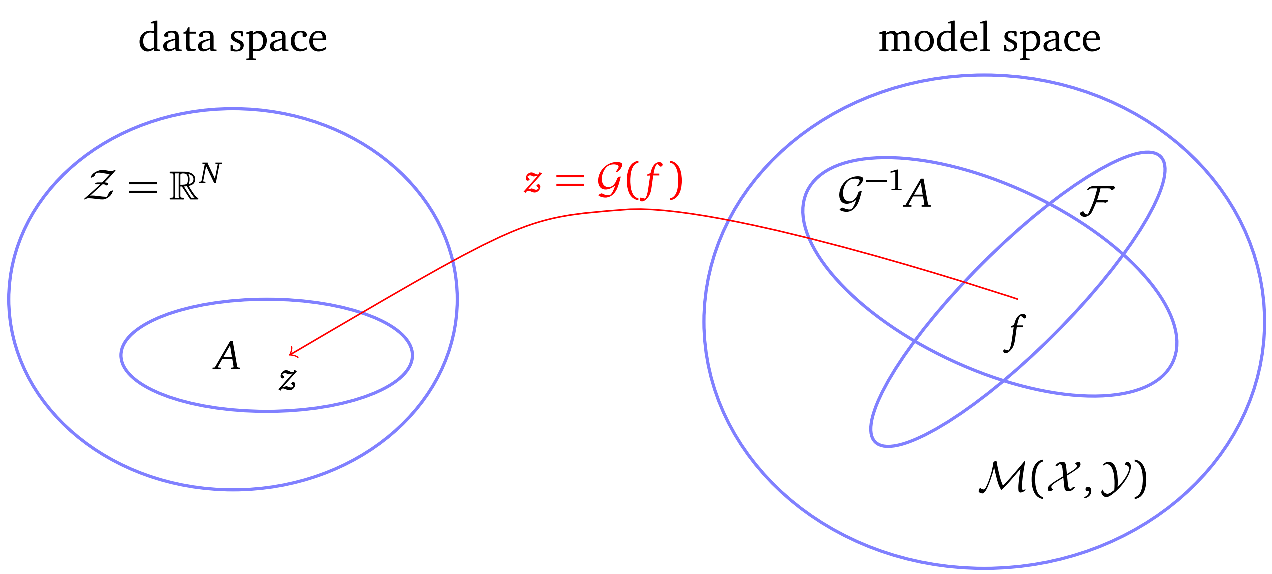
~~Interpret the space of solutions~~ ⇒ implication on the result

More realistically:

find the *support of the ensemble of admissible solutions* in that space

Model bias in inverse problems

Parton Distribution Functions: are determined from data through a mapping of spaces



[Del Debbio, SciPost Phys. Proc. 15]

⟹ data as projected functional $\mathscr{G}$ of a model f

⟹ f represents the underlying truth, but is not uniquely determined by current data

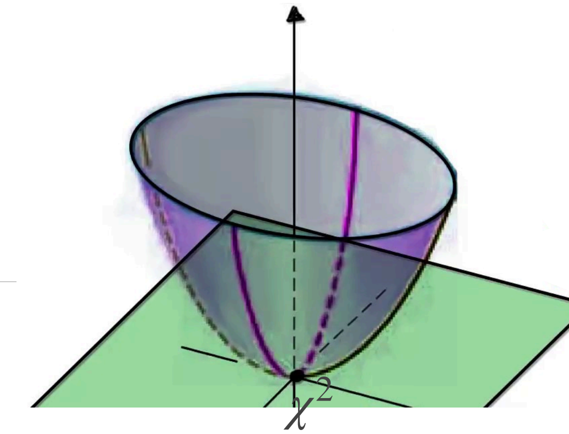
The posterior for a PDF model f given the data is the product of the likelihood and a prior, $\pi\left(f\right)$,

$$p\left(f\,|\,A\right) \propto \boxed{p\left(A\,|\,f\right)}\,\boxed{\pi\left(f\right)}$$

Likelihood of data given f

Is there any constraint or bias, as prior knowledge, to be imposed to our function f?

Phenomenological PDFs



Those inverse problems involve searching for extrema of a (log-)likelihood function.

Uncertainty propagates from data and methodology to the PDF determination

$$\chi^2 = \sum_i^{N_{\exp}} \frac{\left(z_i - \langle \mathcal{G}[f(\{\underline{x},\underline{a}\})]\rangle_i\right)^2}{\sigma_i^2} + \text{penalty terms}$$

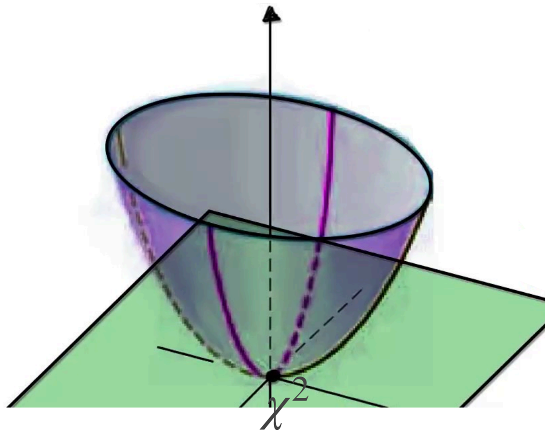
Aleatoric uncertainties

Epistemic uncertainties

Statistical uncertainty propagated from experiments— irreducible

Uncertainty due to lack of knowledge —bias (may be reduced)

Phenomenological PDFs



Those inverse problems involve searching for extrema of a (log-)likelihood function.

Uncertainty propagates from data and methodology to the PDF determination

$$\chi^2 = \sum_i^{N_{\rm exp}} \frac{\left(z_i - \langle \mathcal{G}[f(\{\underline{x}, \underline{a}\})]\rangle_i\right)^2}{\sigma_i^2} + \text{penalty terms}$$

Aleatoric uncertainties

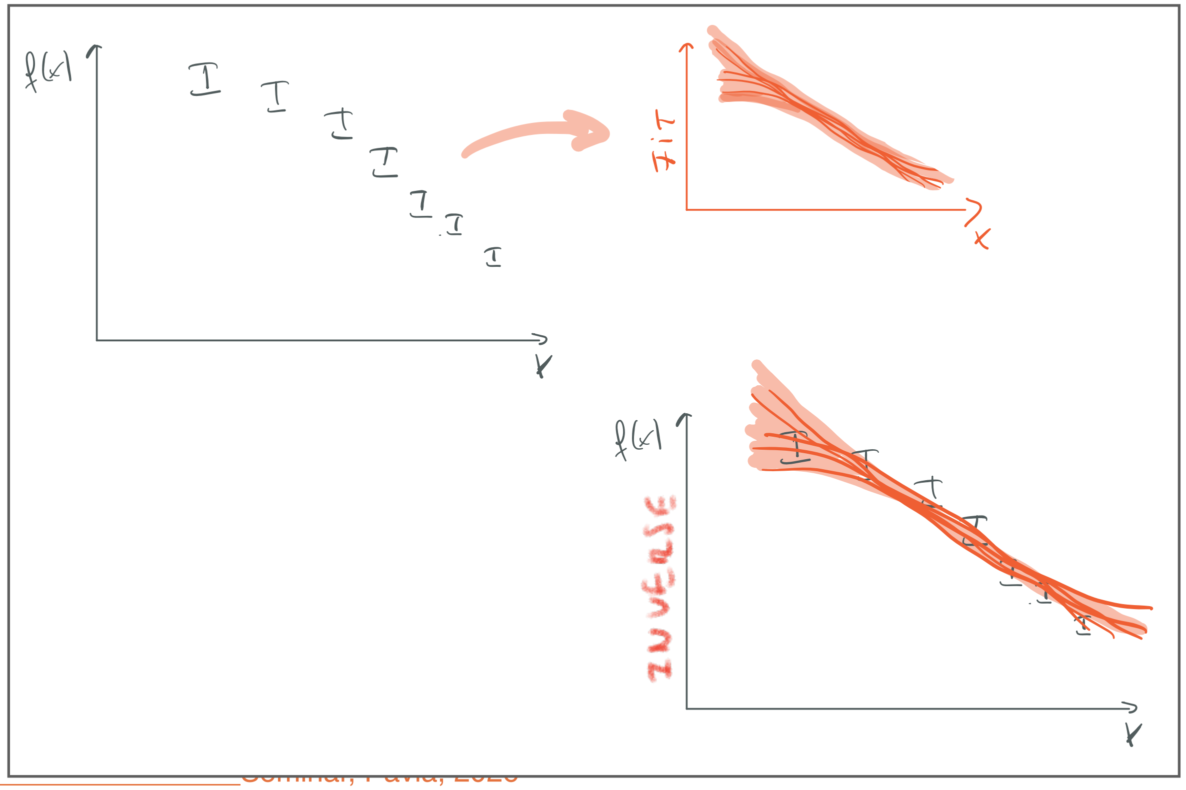
Epistemic uncertainties

Statistical uncertainty propagated from experiments— irreducible

Uncertainty due to lack of knowledge —bias (may be reduced)

The theory input depends on the PDFs, which parametrization is an input to the minimization procedure.

The comparison to data for *various parametrizations* can lead to equally good χ^2 values.



State-of-the-art

Goal: find the shape of PDFs on $x \in [0,1]$ in N_{flavor} space

⟹ Data-driven *global QCD analyses* are inverse problems that aim to infer PDFs

⟹ PDFs as access to underlying non-perturbative QCD dynamics

⟹ Many models are possible

⟹ Approaches:

- Explicit on the parameters: use polynomial approximation — (model+) parameter inference
  use neural networks — architecture infers the behavior of the system

  Flexibility, bias-variance tradeoff

- Very non-parametric: kernel methods, Hausdorff inverse problem, …— still exploratory

State-of-the-art

Goal: find the shape of PDFs on $x \in [0,1]$ in N_{flavor} space

⟹ Data-driven *global QCD analyses* are inverse problems that aim to infer PDFs

⟹ PDFs as access to underlying non-perturbative QCD dynamics

⟹ Many models are possible

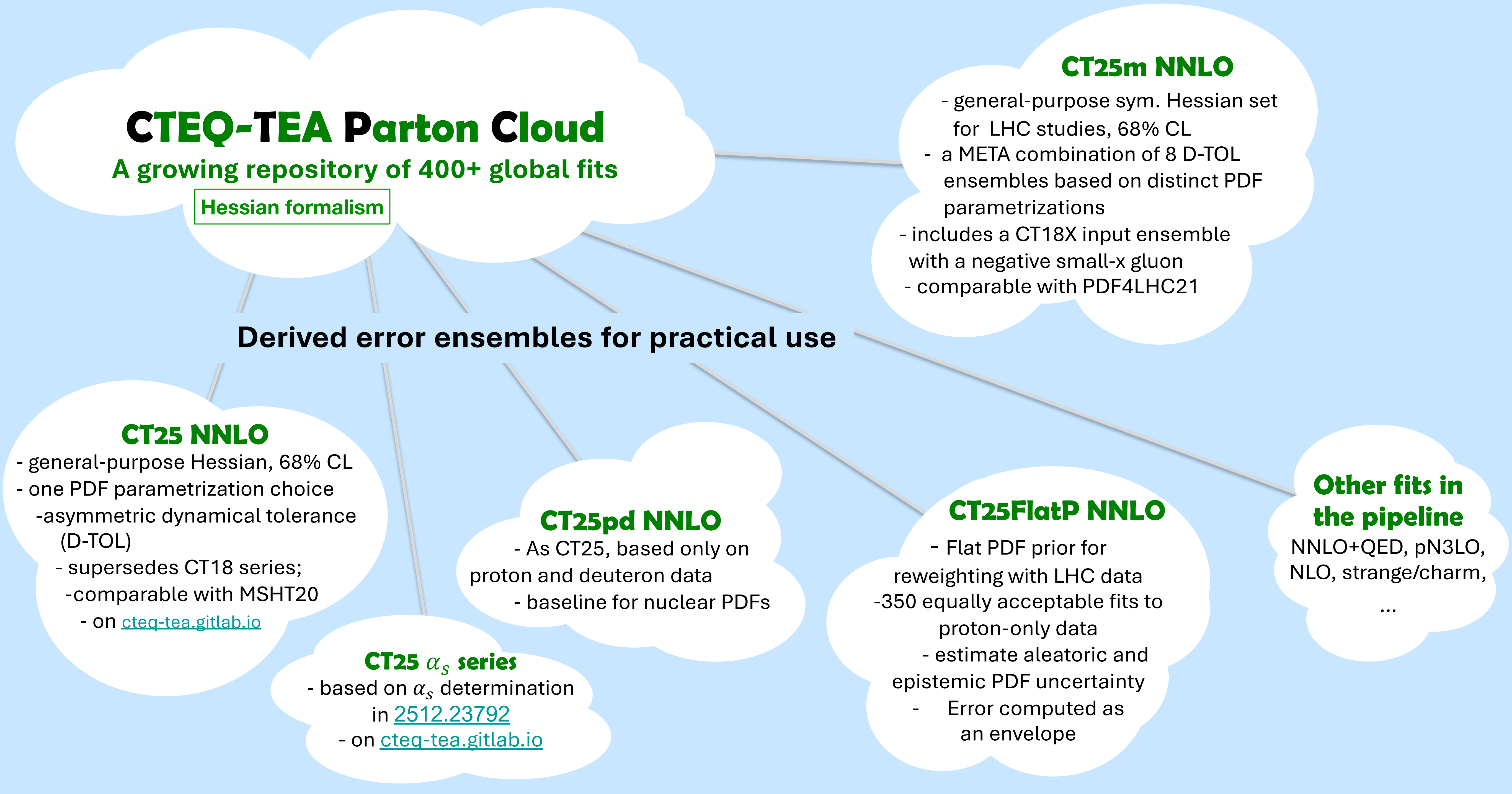
⟹ Approaches:

- Explicit on the parameters: use polynomial approximation — (model+) parameter inference
  use neural networks — architecture infers the behavior of the system

  Flexibility, bias-variance tradeoff

- Very non-parametric: kernel methods, Hausdorff inverse problem, …— still exploratory

I want to show that this, too, is potentially non-parametric, with an extended statistical interpretation.



Many of those sets are simply a different projection of the space of solutions, other imply different data selection.

e.g.,

CT25m combines 8 diverse solutions selected from those 400+ fits

CT25FlatP implies a pruning of redundant solutions

…

Exploration of solution space for CT nucleon PDFs

CTEQ (Wu-Ki Tung Et Al.) group

Epistemic vs. aleatory uncertainties

Uncertainty due to lack of knowledge —bias (may be reduced)

Uncertainty propagated from experiments— irreducible

Exploration of solution space for CT nucleon PDFs

CTEQ (Wu-Ki Tung Et Al.) group

Epistemic vs. aleatory uncertainties

Uncertainty due to lack of knowledge —bias (may be reduced)

Uncertainty propagated from experiments— irreducible

Uncertainty Quantification in CT is a bottom-up construction in the Hessian formalism

a. Experiment-based penalty
   — D-TOL (aleatory)
b. Multiple parametrizations are combined
   — building up an emergent effective $\Delta\chi^2 = T^2$ (epistemic)

Sampling over parametrization contributes to the (epistemic) uncertainty budget.
[AC et al, PRD107, 2205.10444]

Exploration of solution space for CT nucleon PDFs

CTEQ (Wu-Ki Tung Et Al.) group

Epistemic vs. aleatory uncertainties

Uncertainty due to lack of knowledge —bias (may be reduced)

Uncertainty propagated from experiments— irreducible

Uncertainty Quantification in CT is a bottom-up construction in the Hessian formalism

a. Experiment-based penalty
   — D-TOL (aleatory)

b. Multiple parametrizations are combined
   — building up an emergent effective $\Delta\chi^2 = T^2$ (epistemic)

Two key steps:

I. Generating the models that lead to the solutions

I. b. Selecting the representative models

II. Combining the solutions meaningfully

Sampling over parametrization contributes to the (epistemic) uncertainty budget.
[AC et al, PRD107, 2205.10444]

I. Generating the models

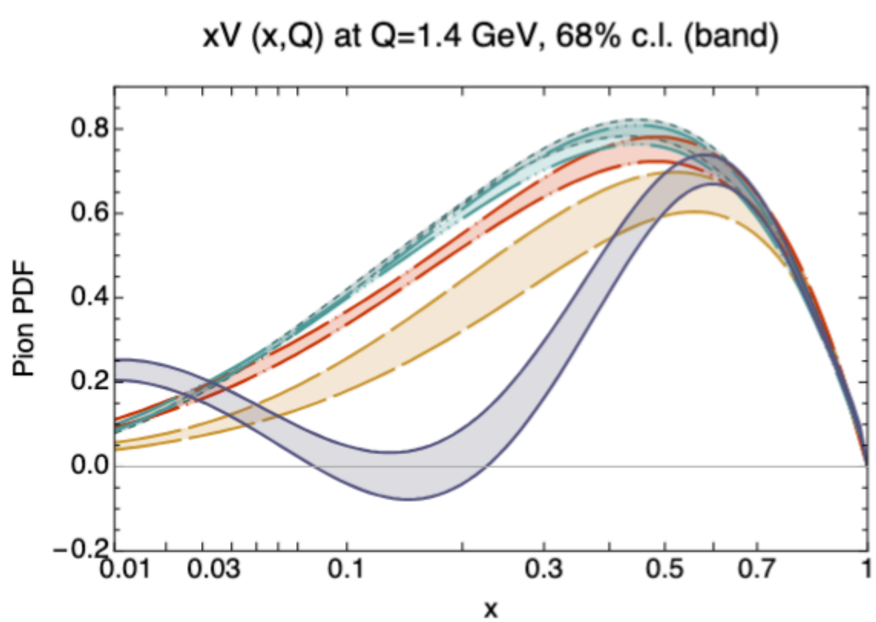
Uncertainty Quantification in CT is partially inspired by Fantômas's sandbox for pion PDFs

b. With Fantômas, multiple parametrizations are generated from Bézier curves** (epistemic)

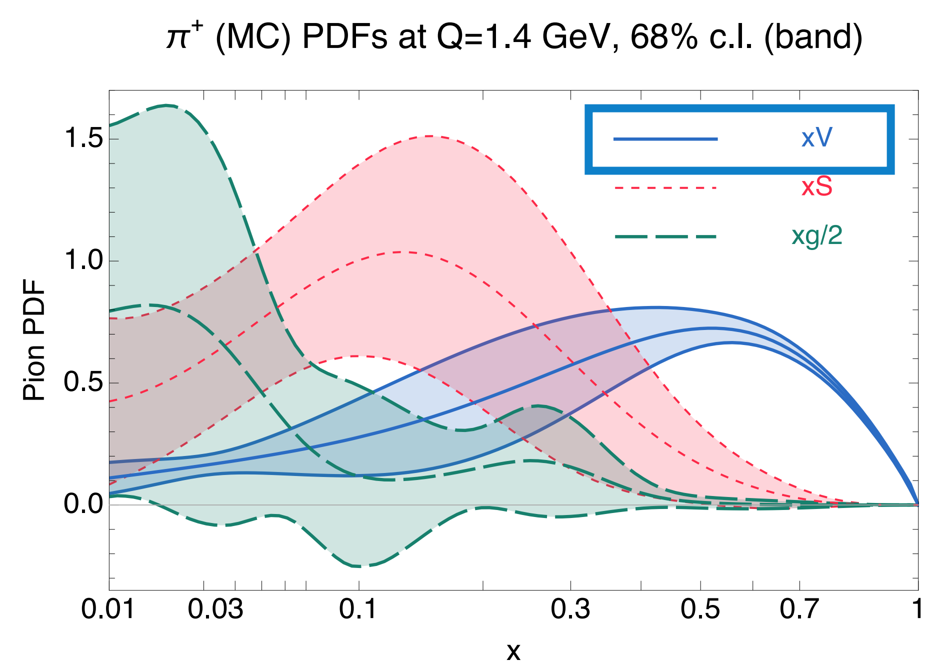
→*these polynomial functional forms can approximate any arbitrary PDF shape*

I. Exploration of solutions space for Fantômas pion PDFs

II. Combination of most diverse solutions



Bundled models →



[Fantômas, PRD109]

**CT25 still uses Bernstein polynomials

Fantômas in a nutshell

`metamorph` is based on Bézier curves — polynomial on a Bernstein basis

$$B_{n,l}(x) \equiv \binom{n}{l} x^l (1-x)^{n-l}$$

$$\boxed{\mathcal{B}^{(n)}(x) = \sum_{l=0}^{n} c_l \; B_{n,l}(x)}$$

The Bézier curve can be expressed as a product of matrices:

$$\mathcal{B} = \boxed{\underline{T}} \cdot \boxed{\underline{\underline{M}}} \cdot \boxed{\underline{C}}$$

vector of coefficients c_l

vector of x^l

matrix of binomial coefficients

Bézier curve characterized by **control points**, vector of $\mathscr{B} \to \underline{P}$:

$$\underline{P} = \boxed{\underline{\underline{T}}} \cdot \underline{\underline{M}} \cdot \underline{C}$$

matrix of x^l at $\{x_{\mathrm{CP}}\}$

PDF shape:

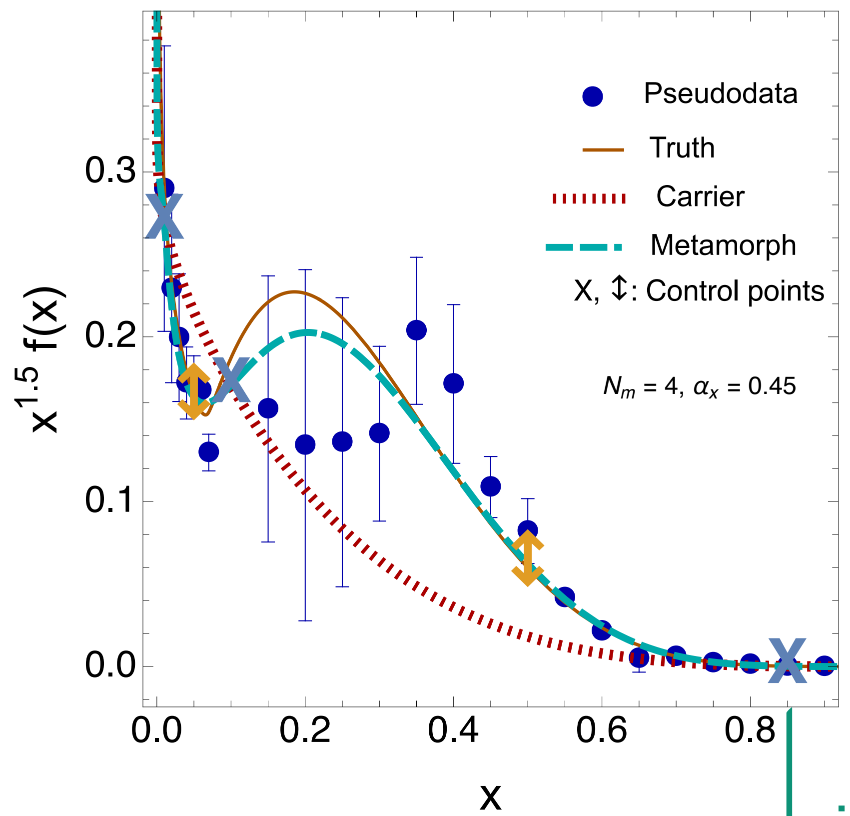
⟹ asymptotics usually ensured by a *carrier function*

⟹ sum rules imposed through normalization

for q = PDF type (flavor, combination or gluon)

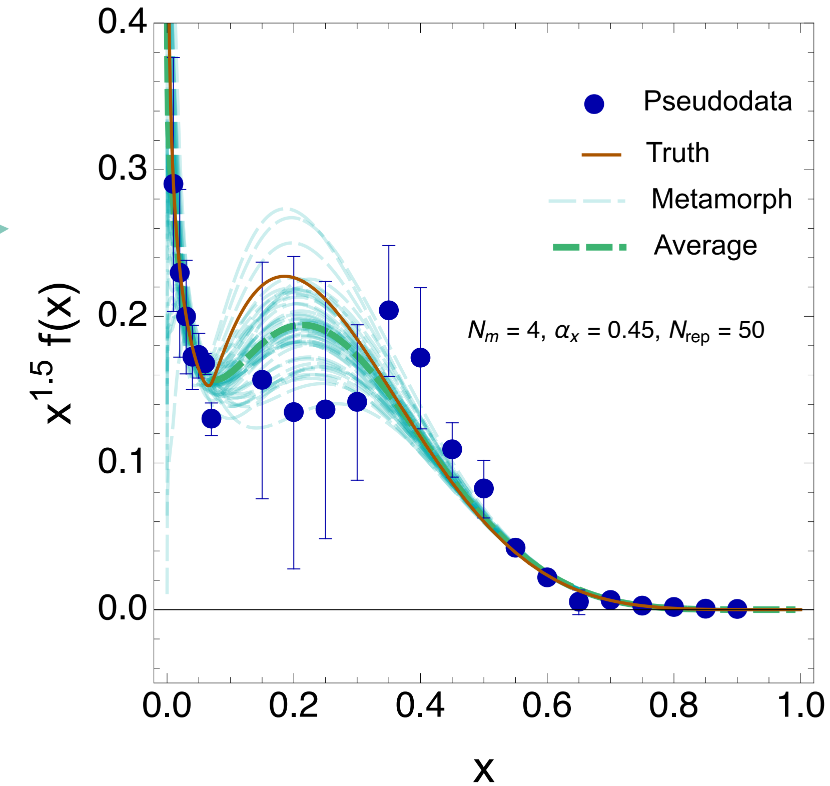
$$\boxed{x\, q(x, Q_0^2) = \underline{A'_q}\, \underline{x^{B_q}(1-x)^{C_q}} \times \left(1 + \mathcal{B}^{(N_m)}(x^{\alpha_x}, Q_0^2; \underline{v})\right)}$$

Fantômas methodology— toy model



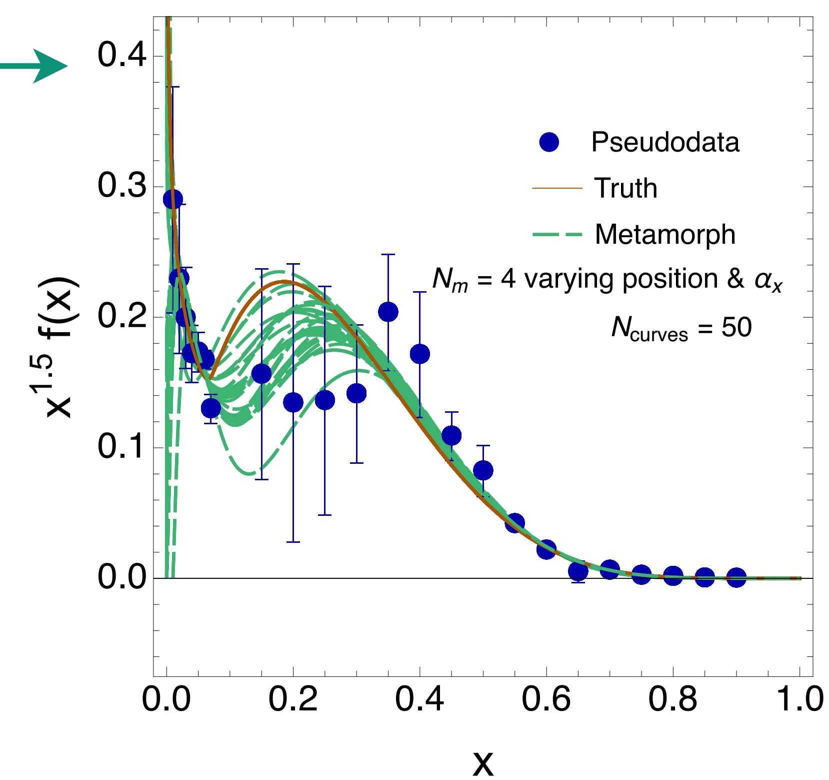
if bootstrapped

sampling on the distribution of data uncertainties



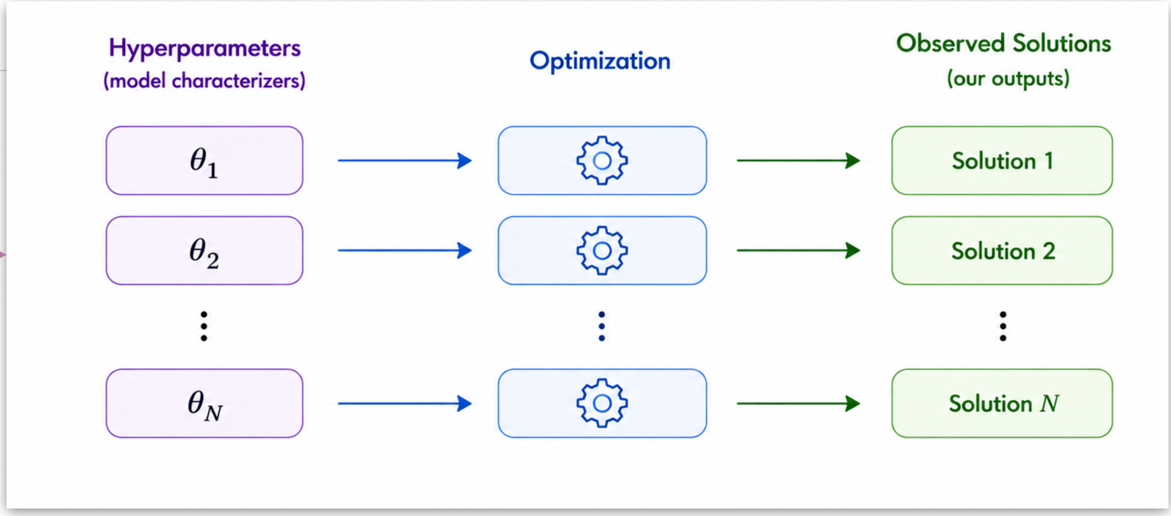
if sampled over `metamorph` settings

sampling over parametrizations



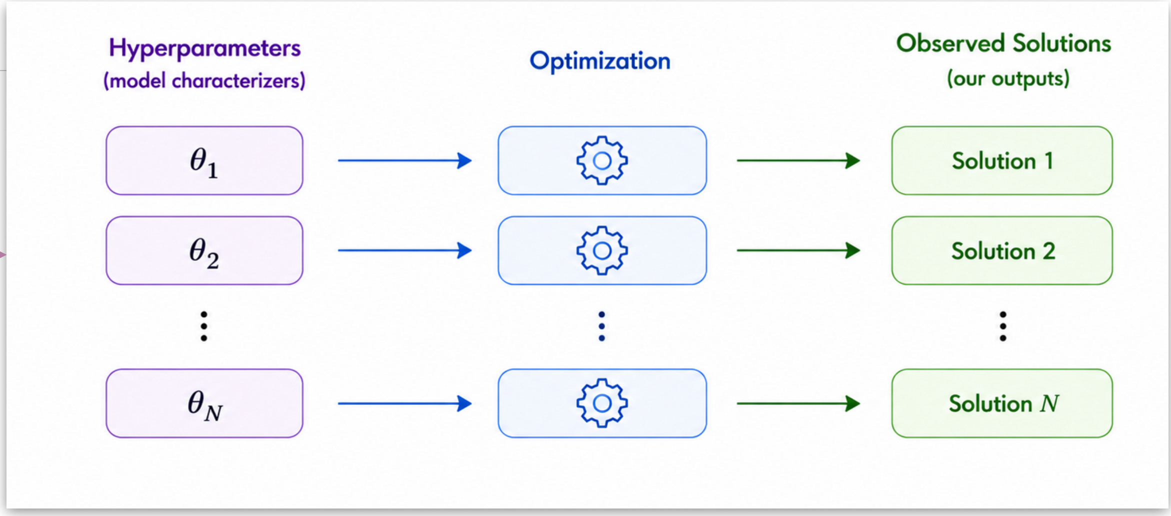
[Kotz et al, Comput.Phys.Commun. 320]

Model's schemes



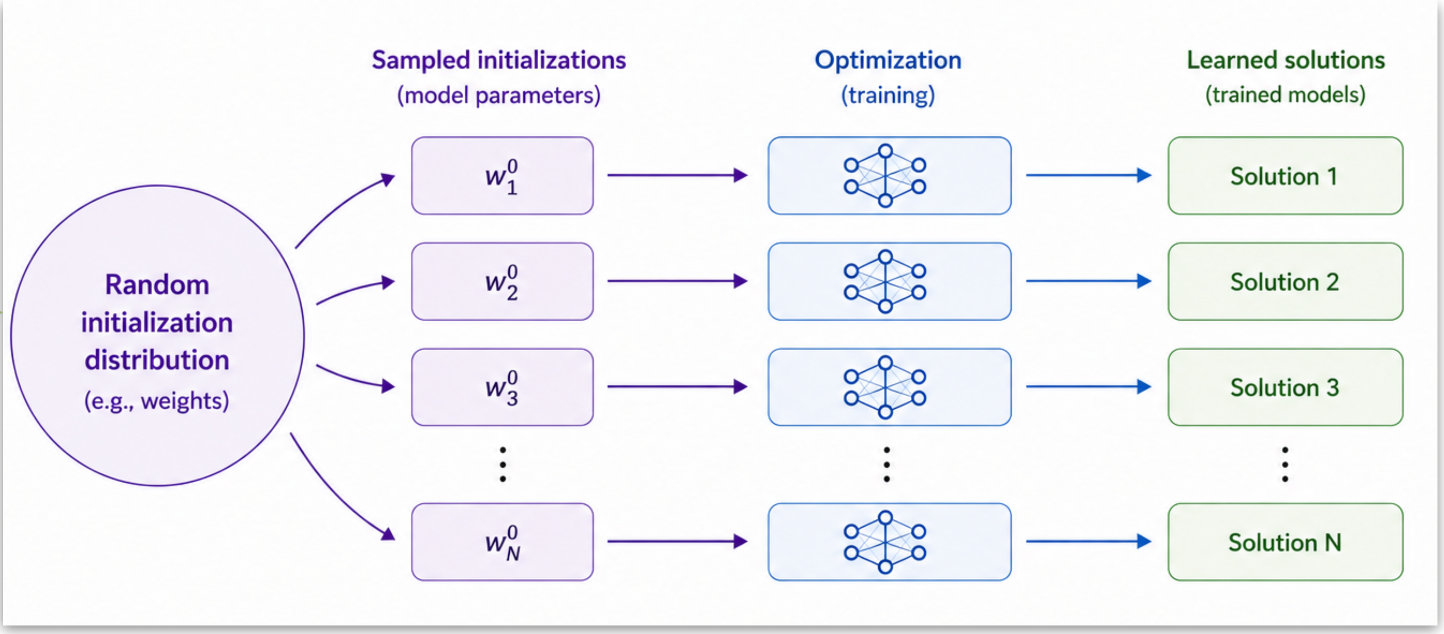
Fantômas's scheme leads to N solutions to the optimization problem

Model's schemes



Fantômas's scheme leads to N solutions to the optimization problem

just like NNPDF's scheme leads to N' replicas.



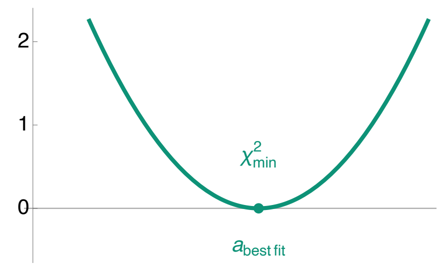
Fantômas reduces the computational time compared to the approaches utilizing traditional polynomial parametrizations.

Exploration of space of solutions

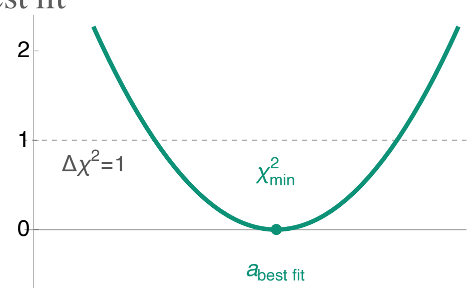
How well do we understand our landscape?

Hessian formalism:

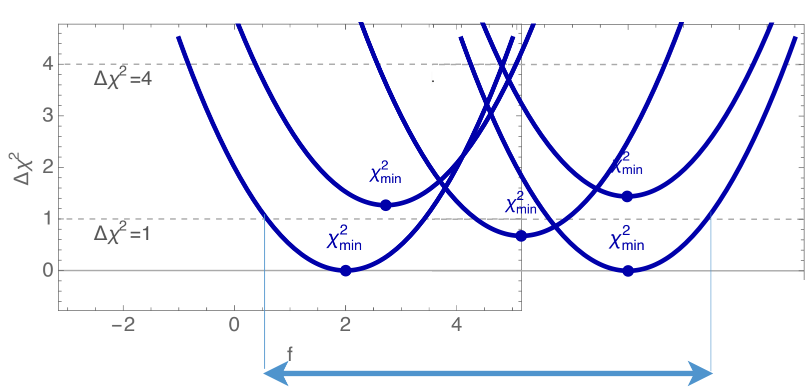
One model (one set of hyper parameters) gives one minimum

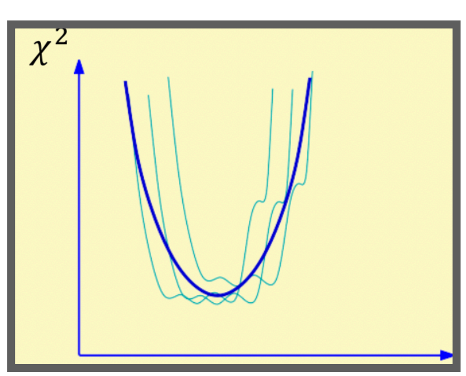


One solution with definite propagation of uncertainty for $a_{\text{best fit}}$



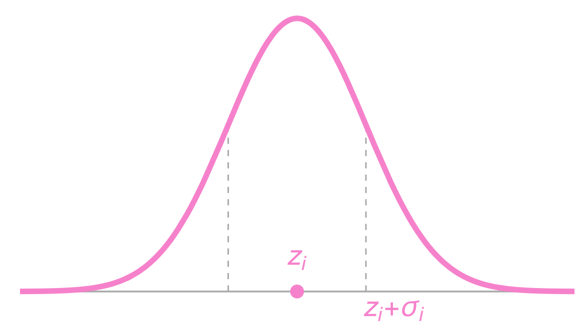
Various sets of hyper parameters gives various minima

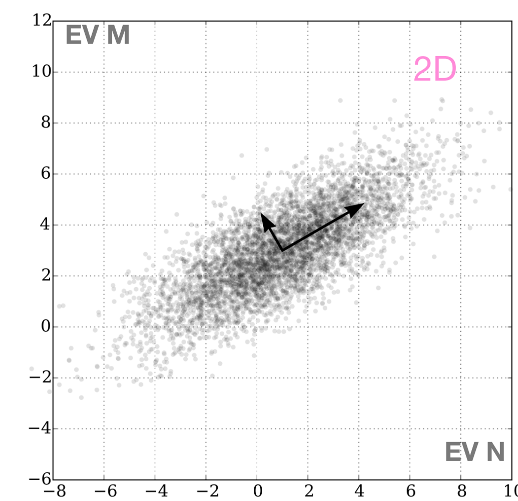




Monte Carlo/bootstrap formalism:

One model (one set of hyper parameters) gives a distribution of the minimum





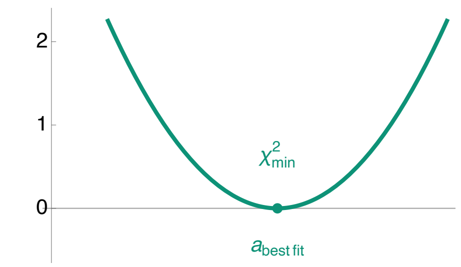
Various sets of hyper parameters gives various distributions of minima

Exploration of space of solutions

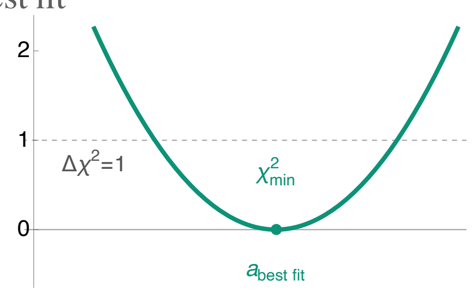
How well do we understand our landscape?

Hessian formalism:

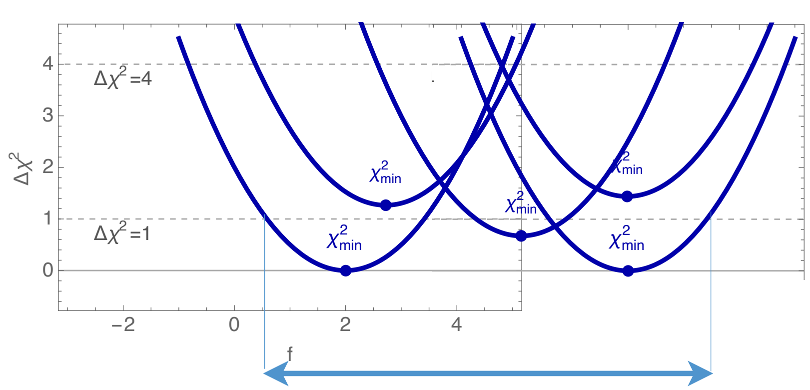
One model (one set of hyper parameters)
gives one minimum



One solution with definite propagation
of uncertainty for $a_{\text{best fit}}$

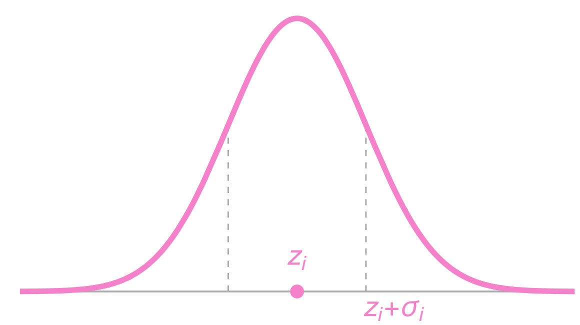


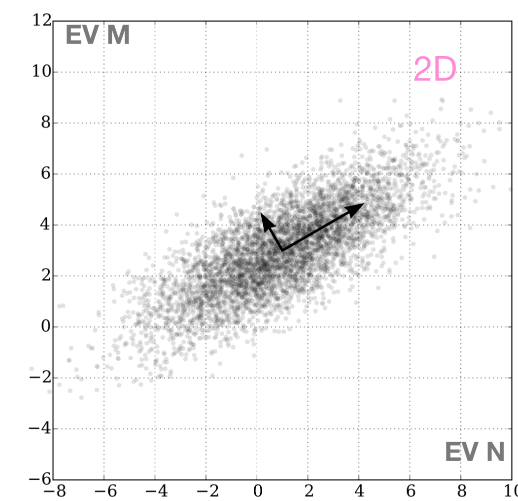
Various sets of hyper parameters
gives various minima



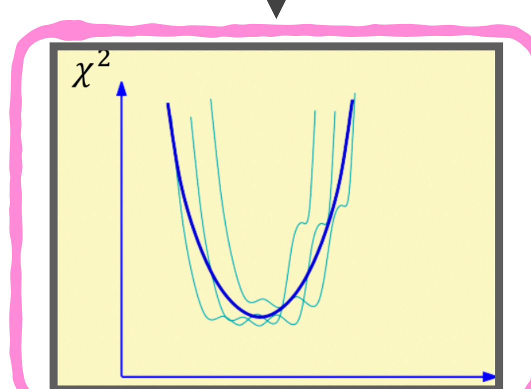
Monte Carlo/bootstrap formalism:

One model (one set of hyper parameters)
gives a distribution of the minimum





Various sets of hyper parameters
gives various distributions of minima



Can we explore the boundaries of the landscape based on complementary aspects to the χ^2?

Ib. Selection of representative solutions

In the Hessian formalism, there is no density concept for the solutions at this point.
We characterize the functions on $[0,1]$ by defining shape evaluators

— “absolute” evaluator through ***information-theoretic criteria***: entropies

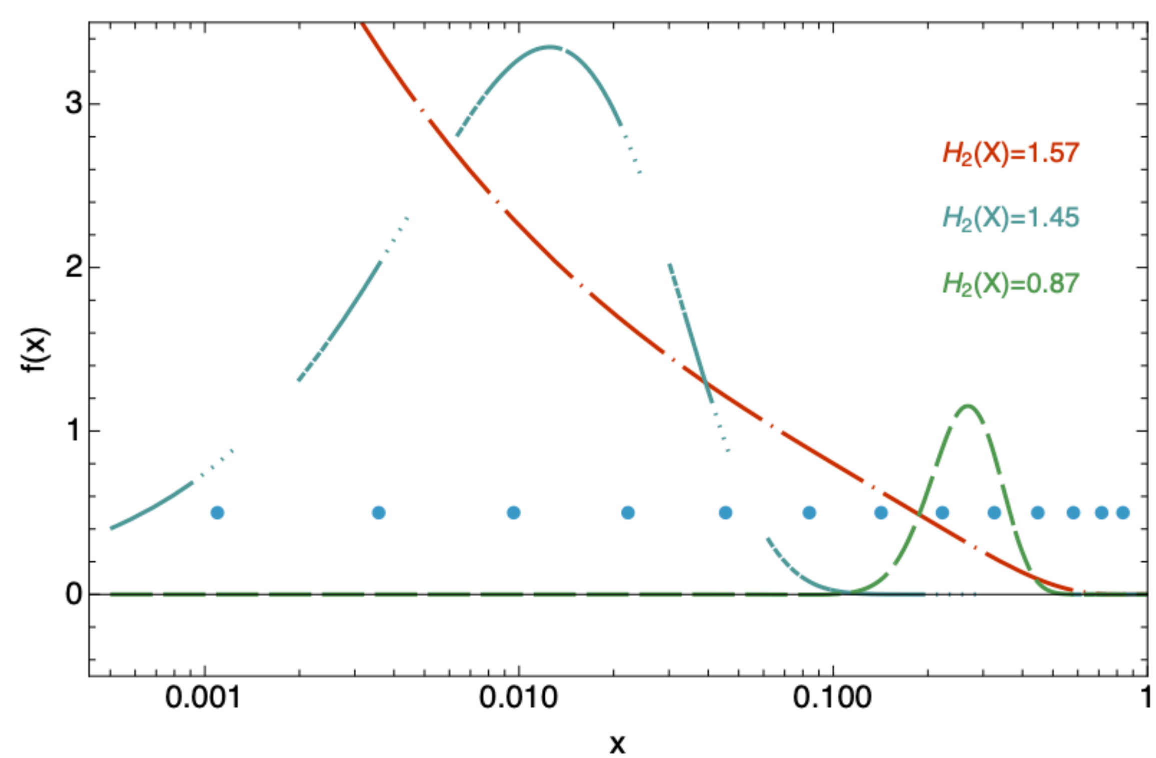
$$H_\alpha(p) = \frac{1}{\alpha - 1} \ln\left(\sum_{i=1}^{n} p(x_i)^\alpha\right), \quad \alpha \neq 1$$

Shannon entropy for $\alpha = 1$
Rényi entropy for $\alpha \neq 1$

How bumpy (certain) is a solution on $x \in [0,1]$?

Flat curve is uncertain → high entropy
— bounded by $\ln n$

Bumpy/peaked curve is certain. → low entropy
— bounded by 0



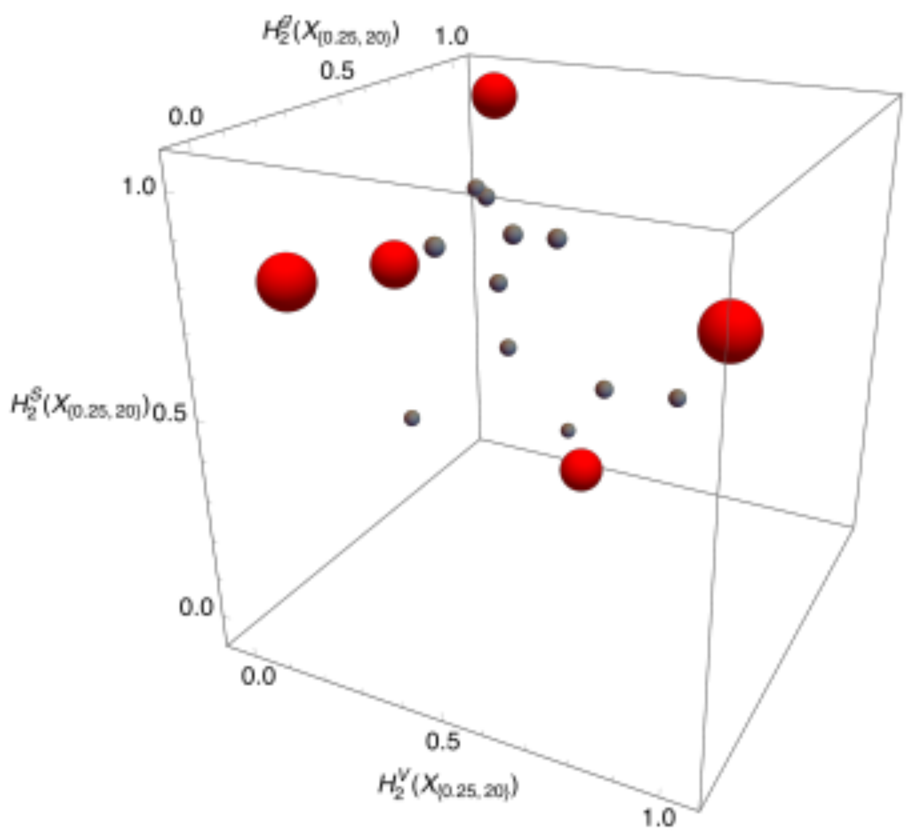
Characterization of functions on $[0,1]$ — the pion PDFs

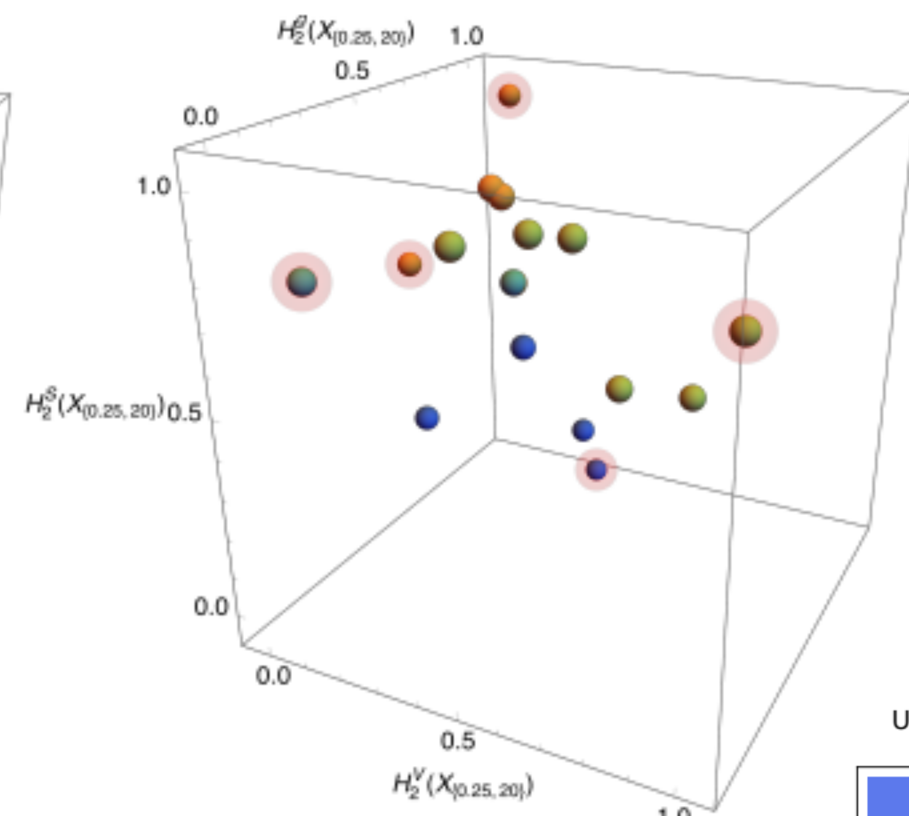
[AC & Ibsen, EPJC86 (2026)]

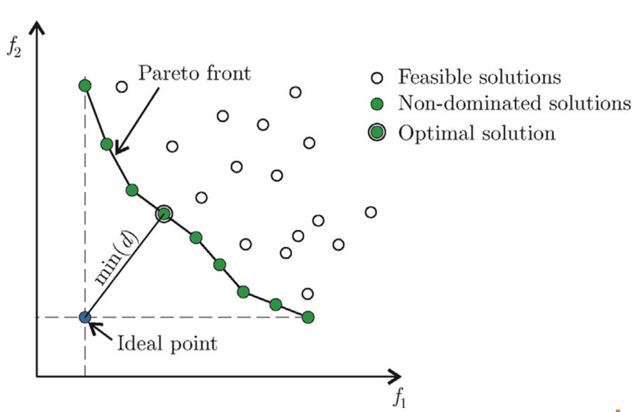
— ***coordinates*** of Rényi entropy (absolute) in 3-dimensional Rényi space.

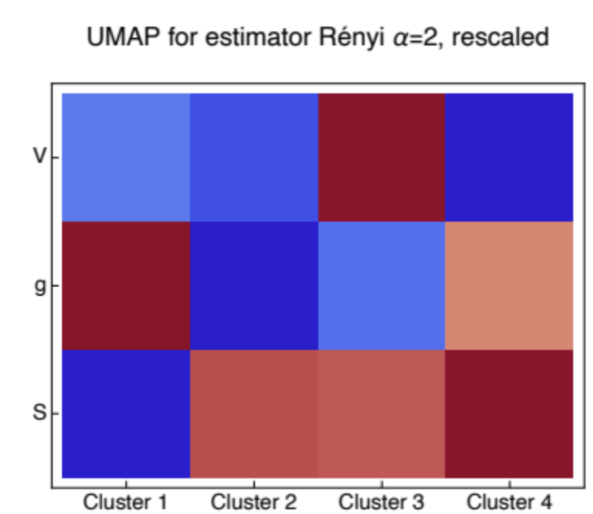
— ***Pareto floor and roof*** to select most diverse shapes

☑ ***Compare to UMAP clustering***: Pareto vertices belong to different clusters









CT25 selection and model uncertainty

To select the **most representative solutions**, we study:

— how they are distributed

— their χ^2 value ($\chi_0^2 + \delta\chi^2$ cut)

I. Clustering by shapes:

Visualizing the structure of the neighborhood of χ_0^2 based on an estimator

— PDF at x-vectors

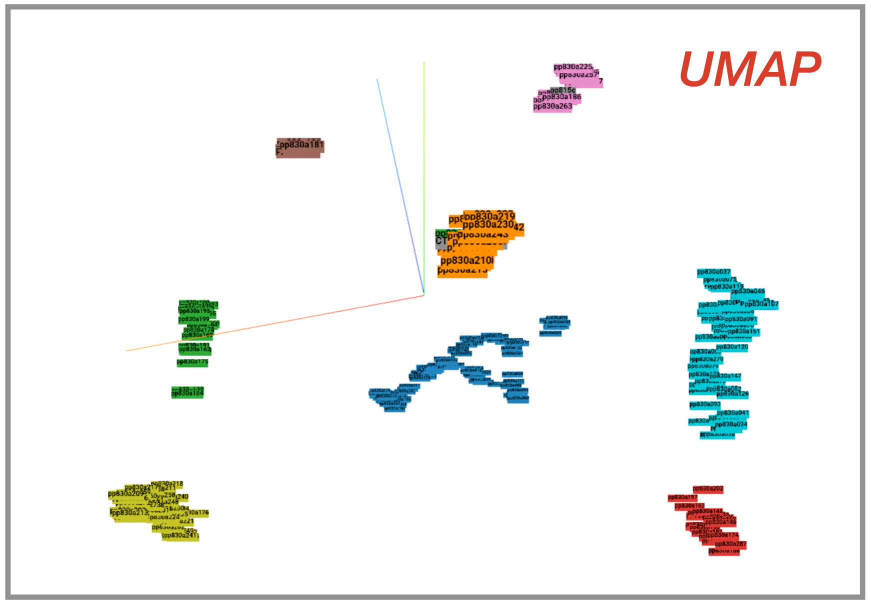
— absolute shape estimator (information criterion)

CT25 selection and model uncertainty

To select the **most representative solutions**, we study:
— how they are distributed
— their χ^2 value ($\chi_0^2 + \delta\chi^2$ cut)

I. <u>Clustering by shapes</u>:

Visualizing the structure of the neighborhood of χ_0^2 based on an estimator
— PDF at x-vectors
— absolute shape estimator (information criterion)



Plot by Y. Fu

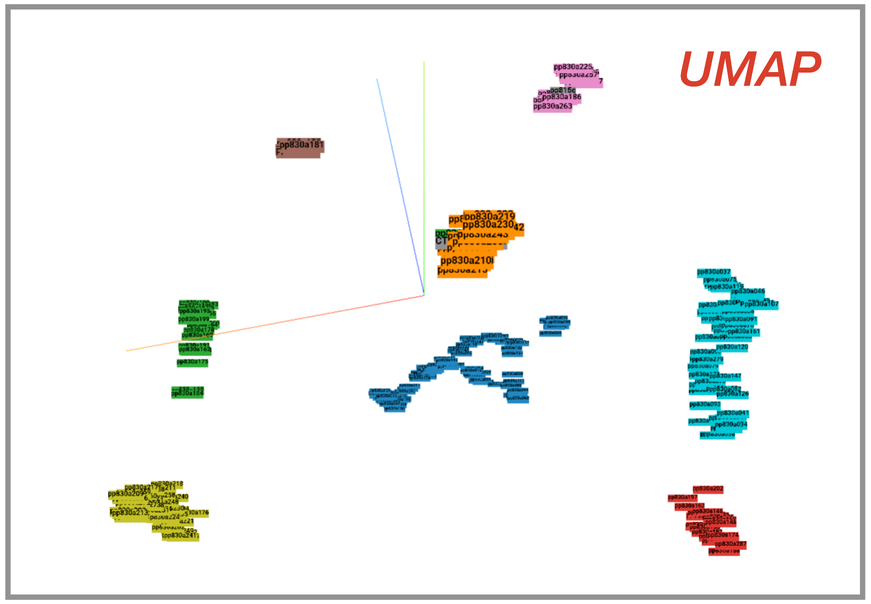
CT25 selection and model uncertainty

To select the **most representative solutions**, we study:
— how they are distributed
— their χ^2 value ($\chi_0^2 + \delta\chi^2$ cut)

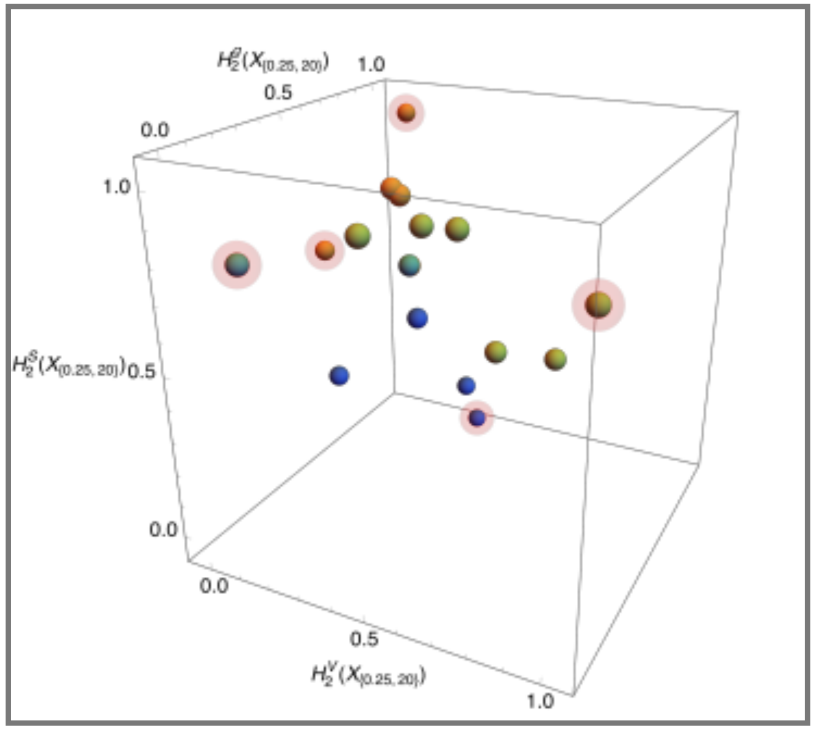
I. Clustering by shapes:

Visualizing the structure of the neighborhood of χ_0^2 based on an estimator

— PDF at x-vectors
— absolute shape estimator (information criterion)



Plot by Y. Fu



Rényi space, illustrated for pion PDF

CT25 selection and model uncertainty

To select the **most representative solutions**, we study:

— how they are distributed

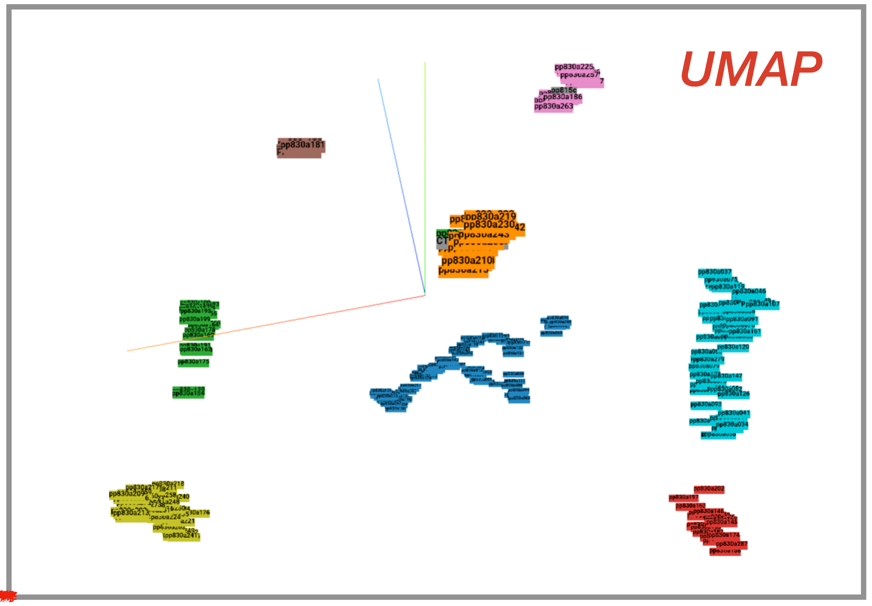
— their χ^2 value ($\chi_0^2 + \delta\chi^2$ cut)

I. <u>Clustering by shapes</u>:

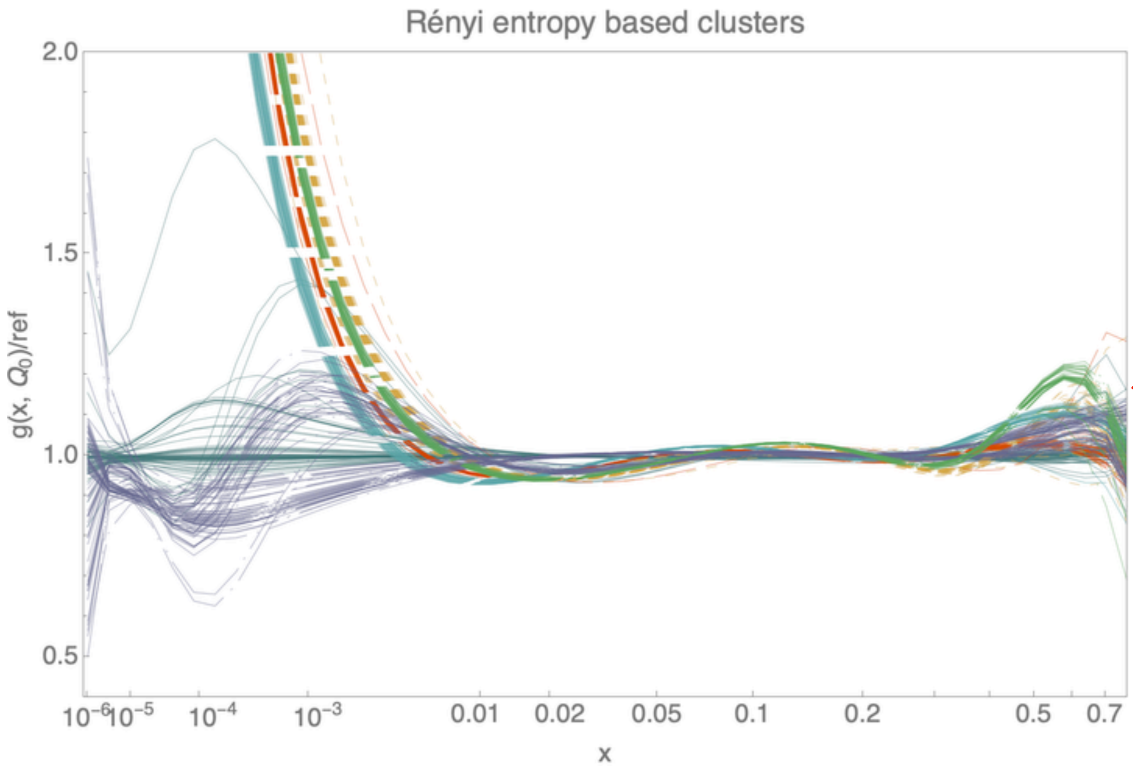
Visualizing the structure of the neighborhood of χ_0^2 based on an estimator

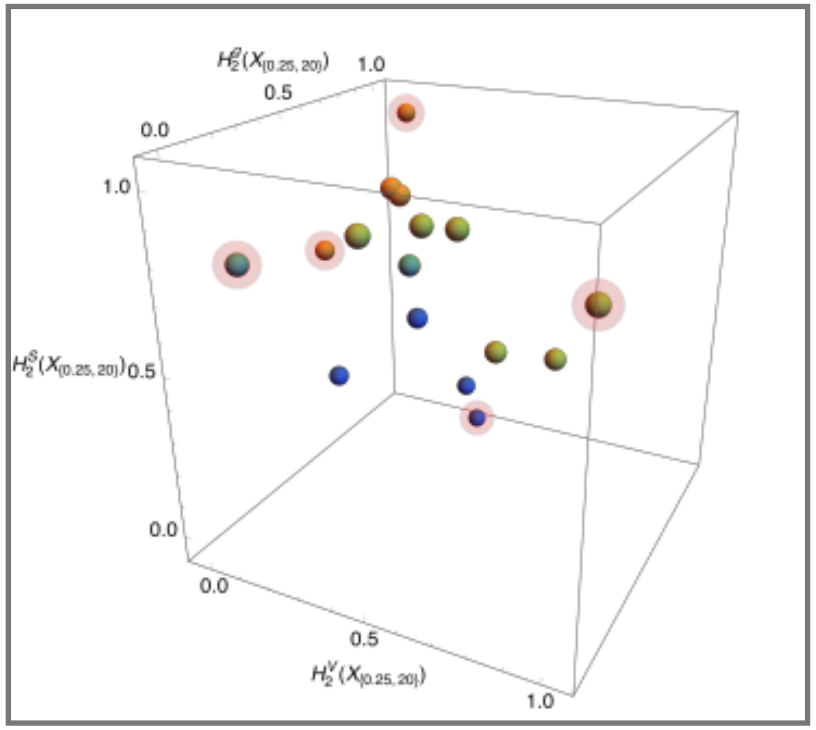
— PDF at x-vectors

— absolute shape estimator (information criterion)



Plot by Y. Fu





Rényi space, illustrated for pion PDF

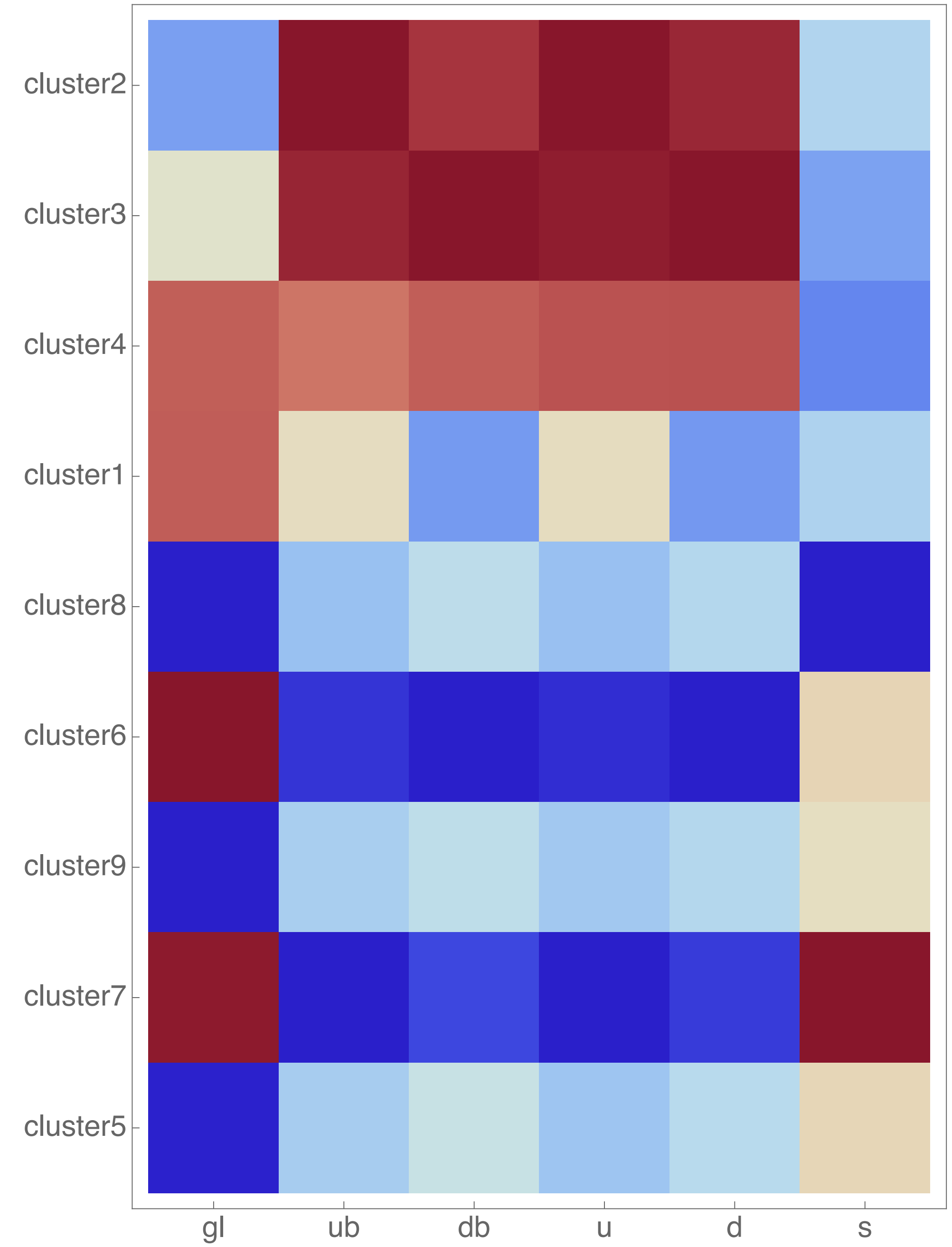
Characteristics of the clusters

Rényi entropy: as an absolute shape estimator

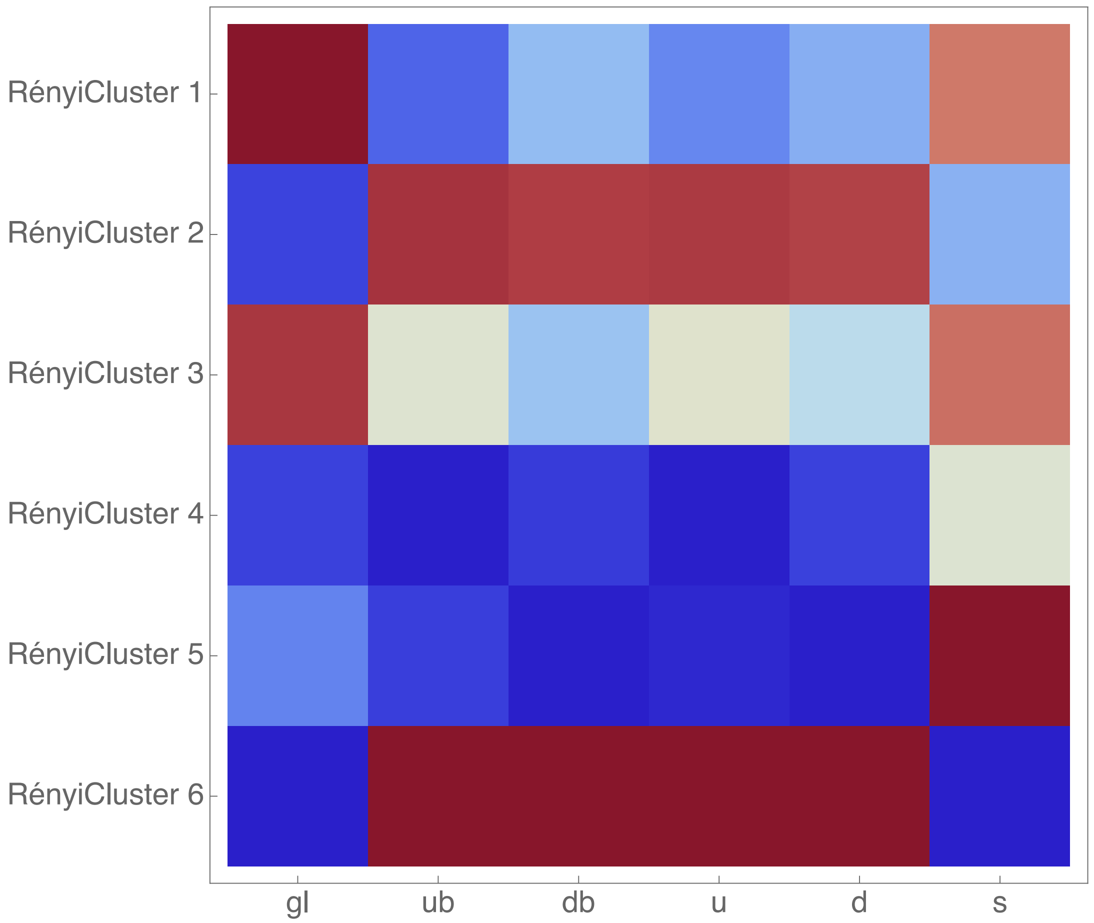
Patterns of entropy per flavor varies cluster to cluster

— 9 clusters with PDF vector as estimator
— 6 clusters with Rényi entropy as estimator

UMAP for estimator PDF vector, Gaussian pts



UMAP for estimator Rényi α=2



2. Combining the solutions

Uncertainty quantification from sampling over models

⟹ representative sampling

⟹ sample over models, translates into sampling over hyper parameters

⟹ if models ~ prior distribution and data ~ aleatoric uncertainty distribution $\mathcal{N}(\mu, \sigma)$,

epistemic uncertainty

- propagation of two distributions (full marginalization over prior and data draws)

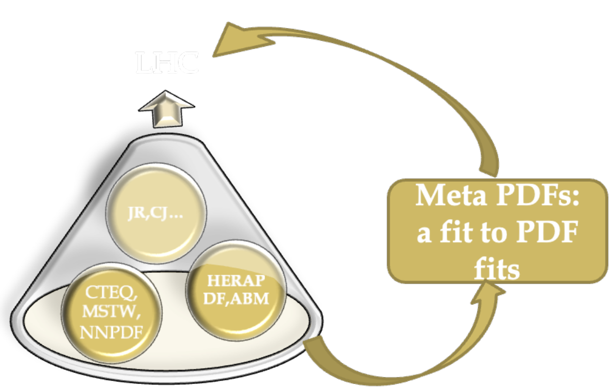
Hessian formalism — *applied by CT (explicit with Fantômas [2505.13594])*
NN — *applied recently by NNPDF [2605.31482]*

- one-on-one (assumes either the prior or the data is represented by each draw)

Typical of MC formalism — *justified by importance sampling*

⟹ combine solutions à la PDF4LHC, Gaussian Mixture Model, …

METAPDF— Gao & Nadolsky [1401.0013]
GMM — Yan et al [2406.01664]



CT25 selection and model uncertainty

II. Building the uncertainty:

Building the Hessian uncertainty accounting for epistemic sources

— explore the set of χ^2 paraboloids (landscape)
— select the solution with best χ^2 per cluster
— combine the solutions through MetaPDF

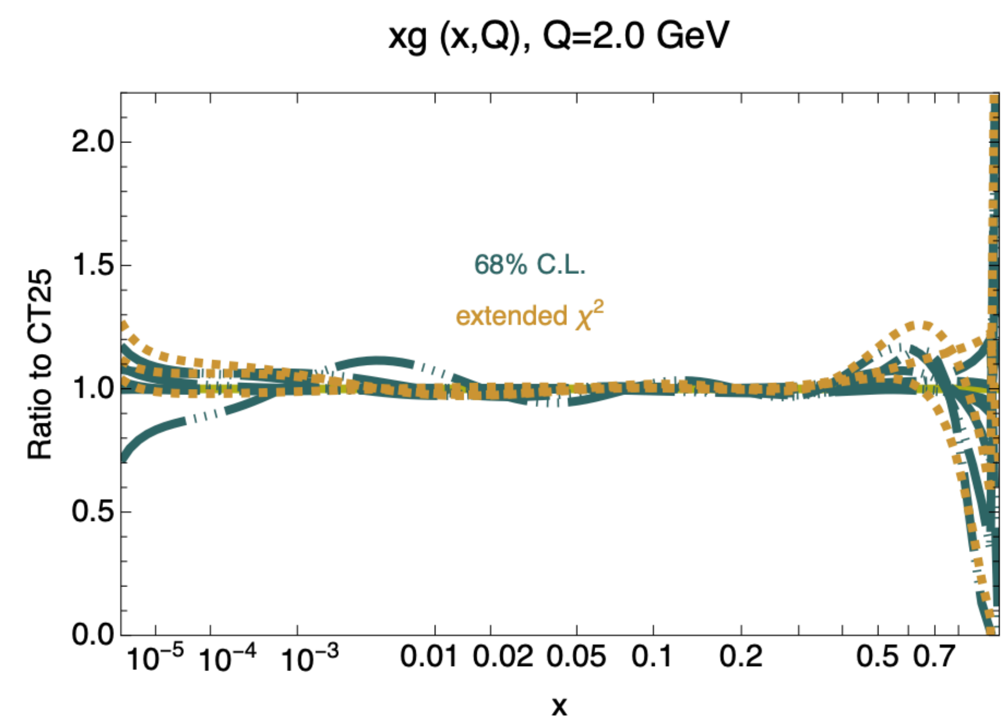
[1401.0013]

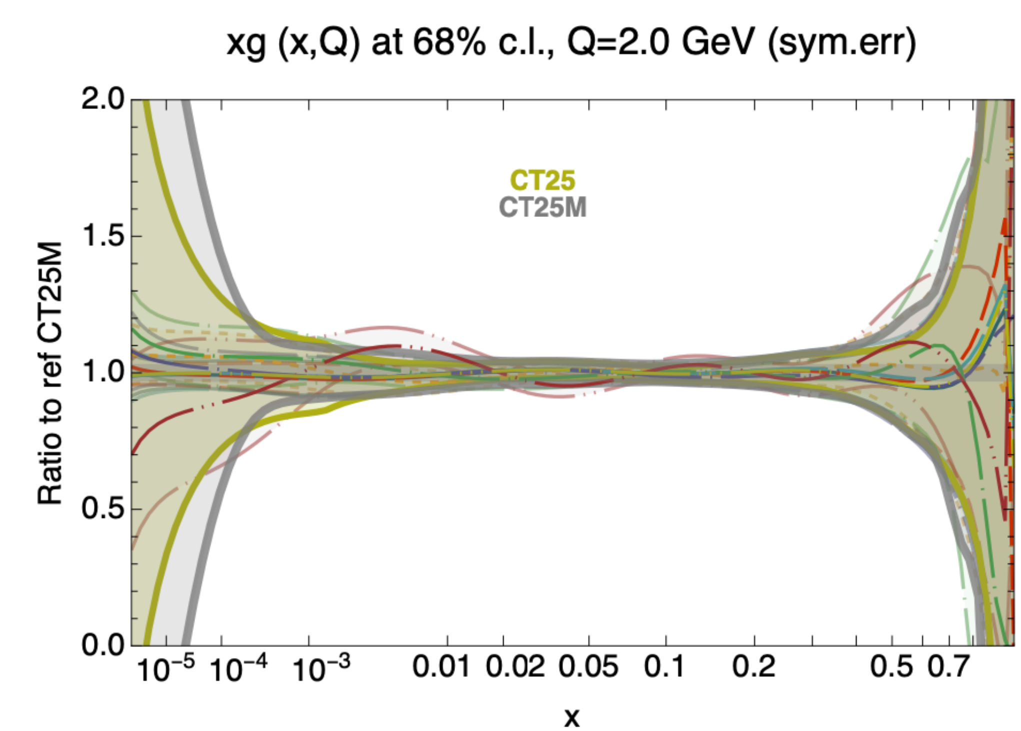
*statistically justified by de Finetti's theorem***

[2505.13594]

Implementation-wise, close to the PDF4LHC prescription

[2203.05506]





*** bear with me for 2 slides*

CTEQ-TEA Parton Cloud

A growing repository of 300+ global fits

CT25m NNLO

- general-purpose sym. Hessian set for LHC studies, 68% CL
- a META combination of 8 D-TOL ensembles based on distinct PDF parametrizations
- includes a CT18X input ensemble with a negative small-x gluon
- comparable with PDF4LHC21

CT25 NNLO

- general-purpose Hessian, 68% CL
- one PDF parametrization choice
- asymmetric dynamical tolerance (D-TOL)
- supersedes CT18 series;
- comparable with MSHT20
- on cteq-tea.gitlab.io

CT25 α_s series

- based on α_s determination in 2512.23792
- on cteq-tea.gitlab.io

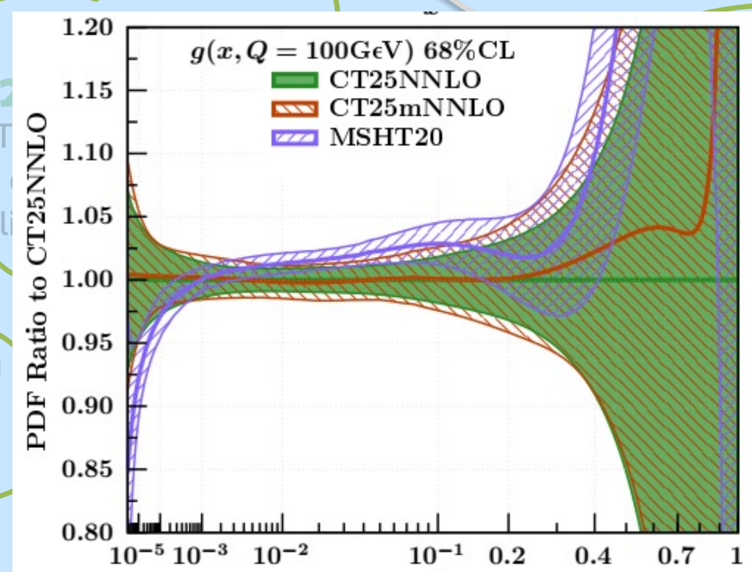
CT2

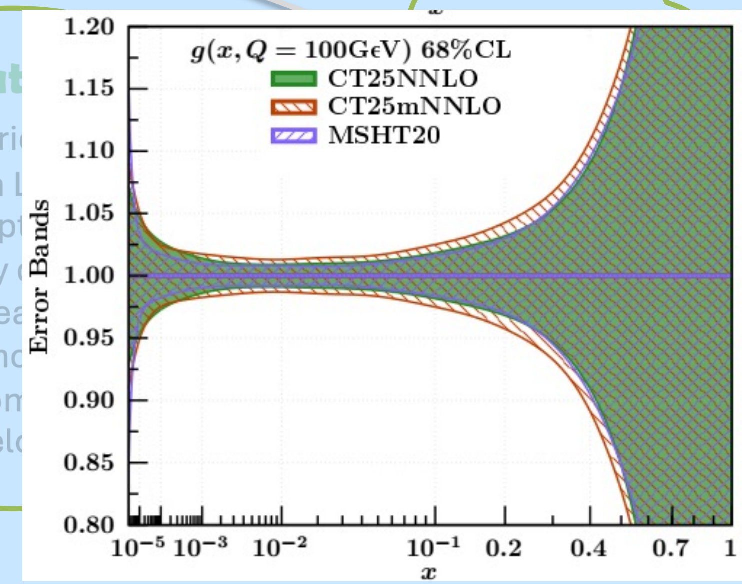
- As CT

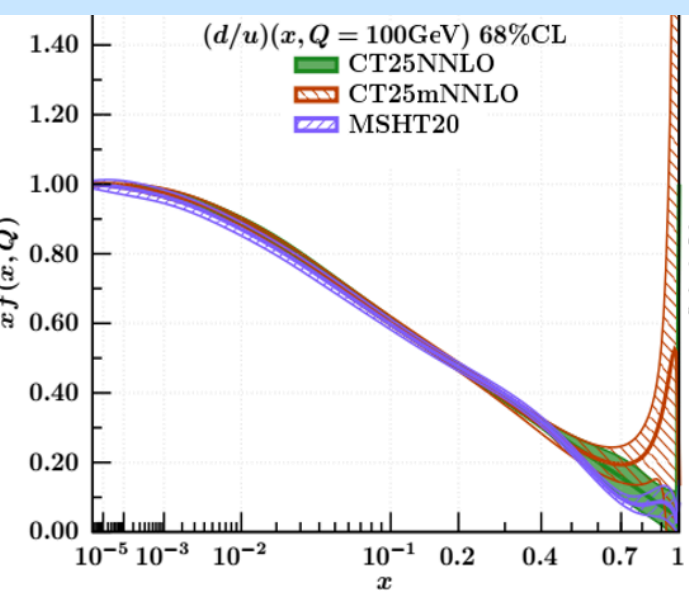
proton and

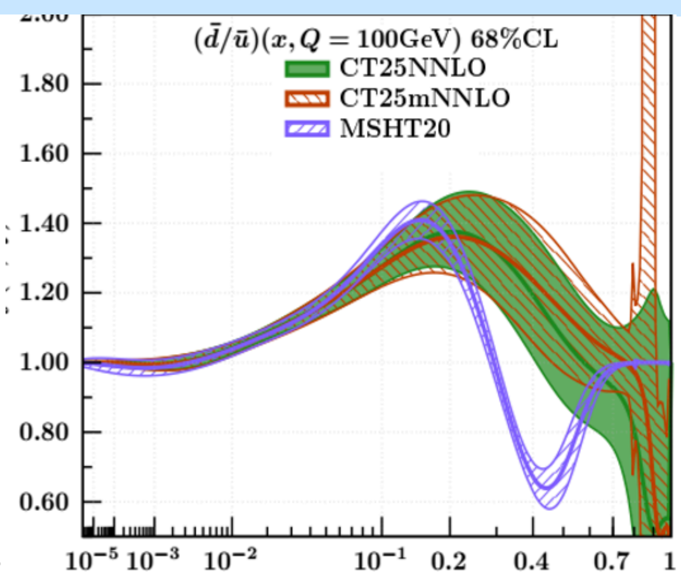
- baseli

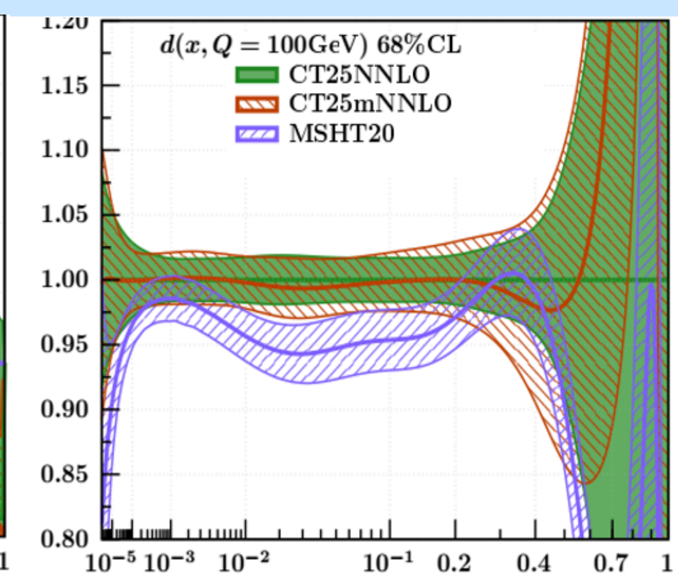
2026-05-05

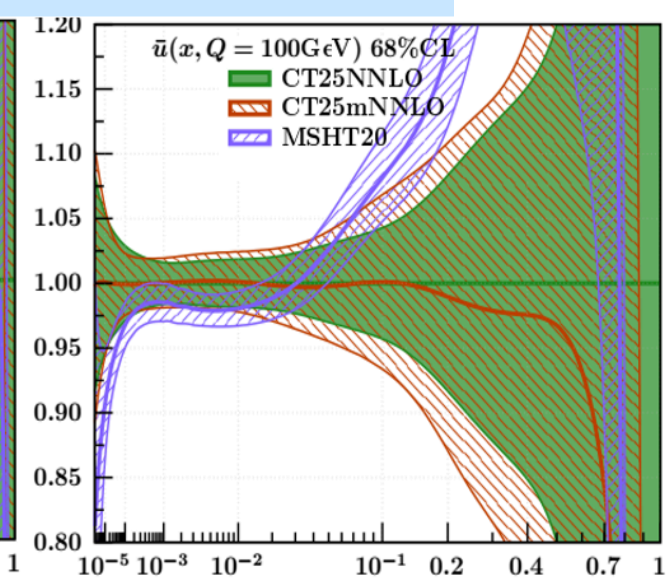












Work-in-progress within the CT25 framework

In the aisle of model uncertainty, we are working on

- **Implementation of the Bézier-curve formalism for the new CT code**

  — exploit the analytical and numerical benefits of Bézier curves to generate models on the fly in a controlled and interpretable way.

- **CT25 Parton Cloud thorough analysis**

  — explore more projections of the space of solutions, complementary to the goodness-of-fit criteria

Work-in-progress within the CT25 framework

In the aisle of model uncertainty, we are working on

- **Implementation of the Bézier-curve formalism for the new CT code**

  — exploit the analytical and numerical benefits of Bézier curves to generate models on the fly in a controlled and interpretable way.

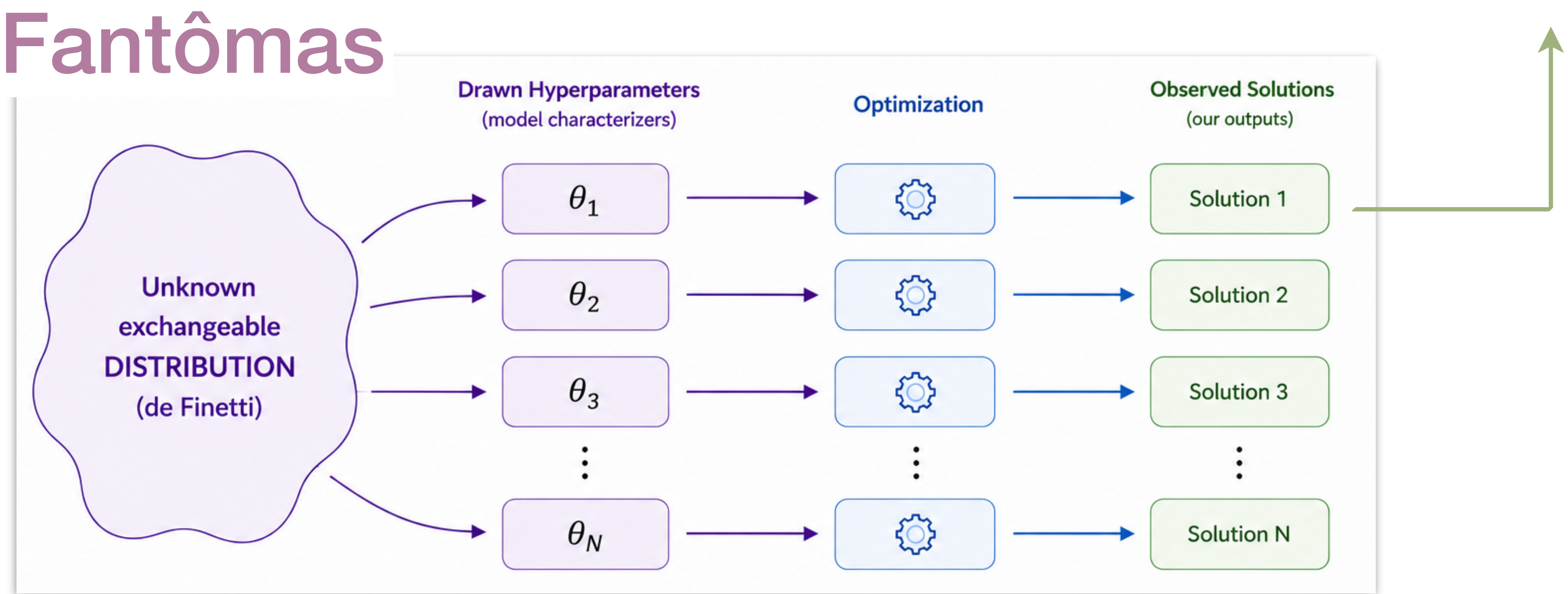
- **CT25 Parton Cloud thorough analysis**

  — explore more projections of the space of solutions, complementary to the goodness-of-fit criteria

⇒ both are motivated by the following statistical interpretation

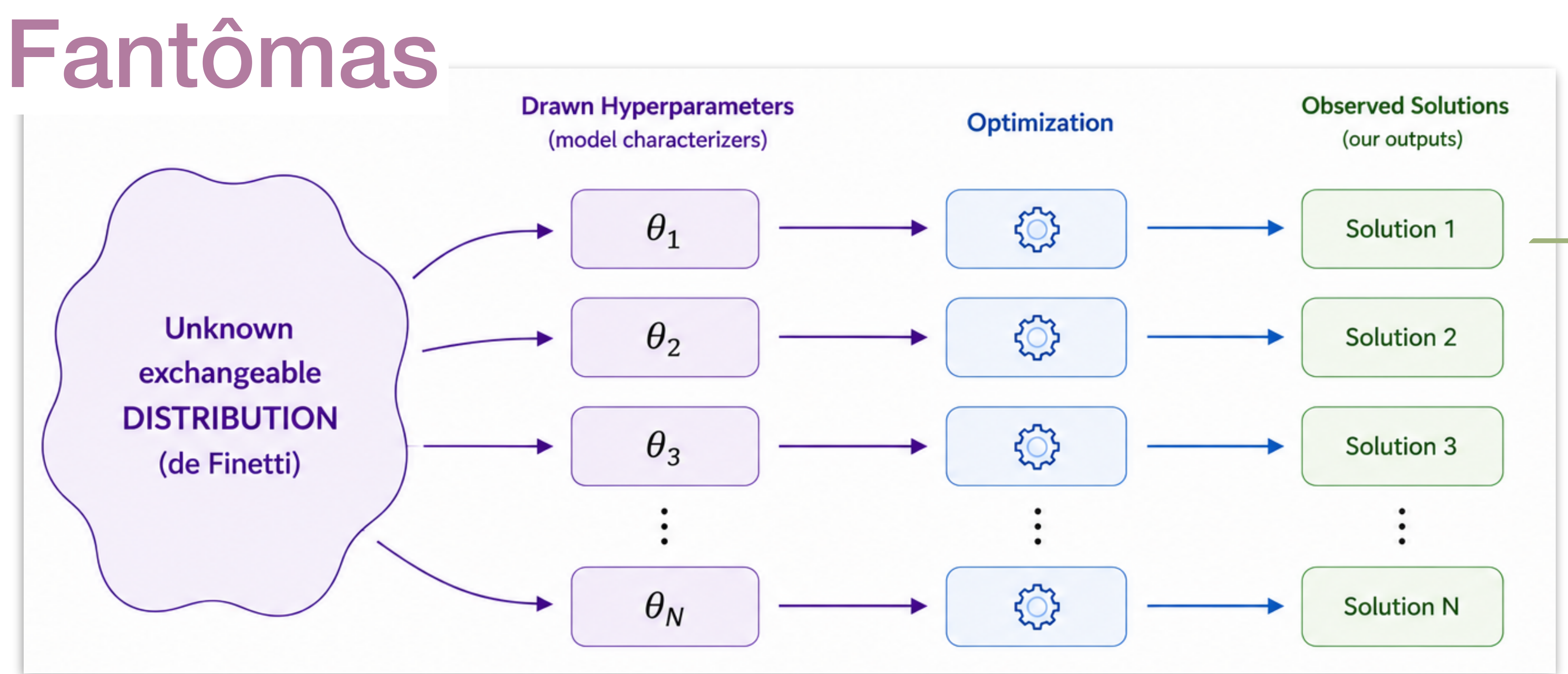
Statistical interpretation

To each solution corresponds a Hessian or MC ensemble uncertainty



Statistical interpretation

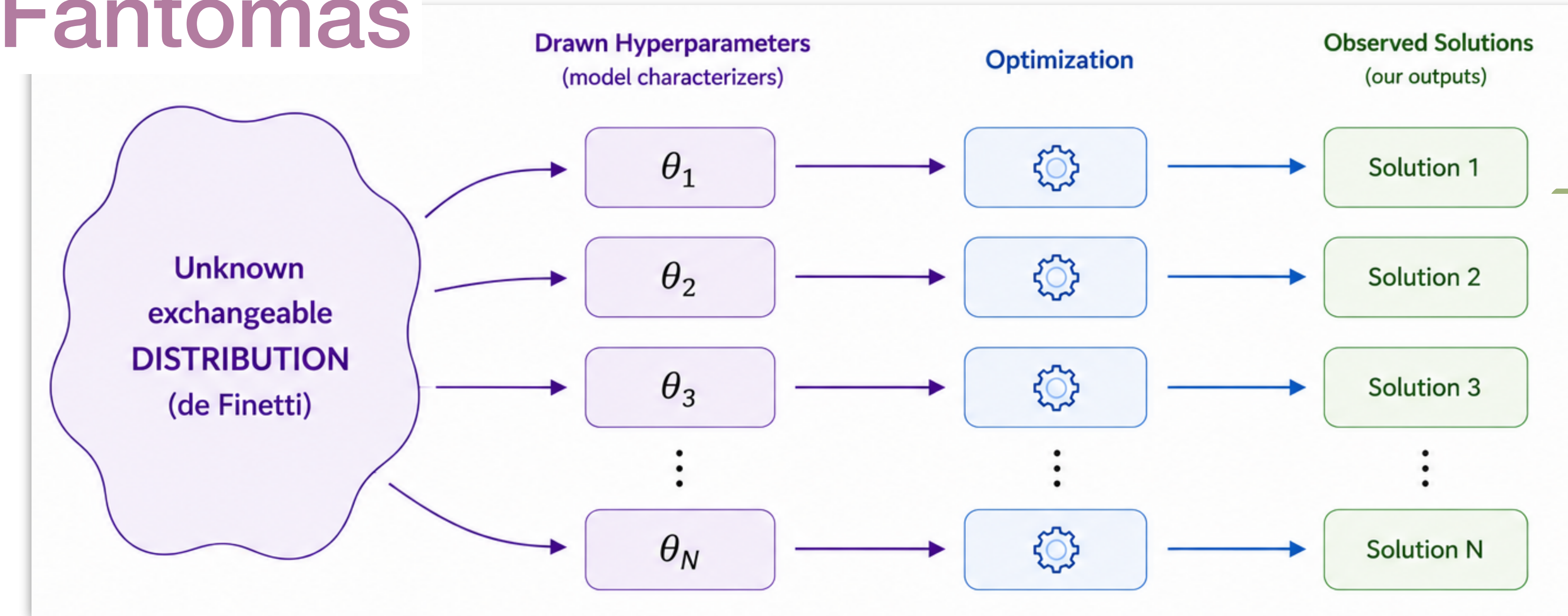
To each solution corresponds a Hessian or MC ensemble uncertainty



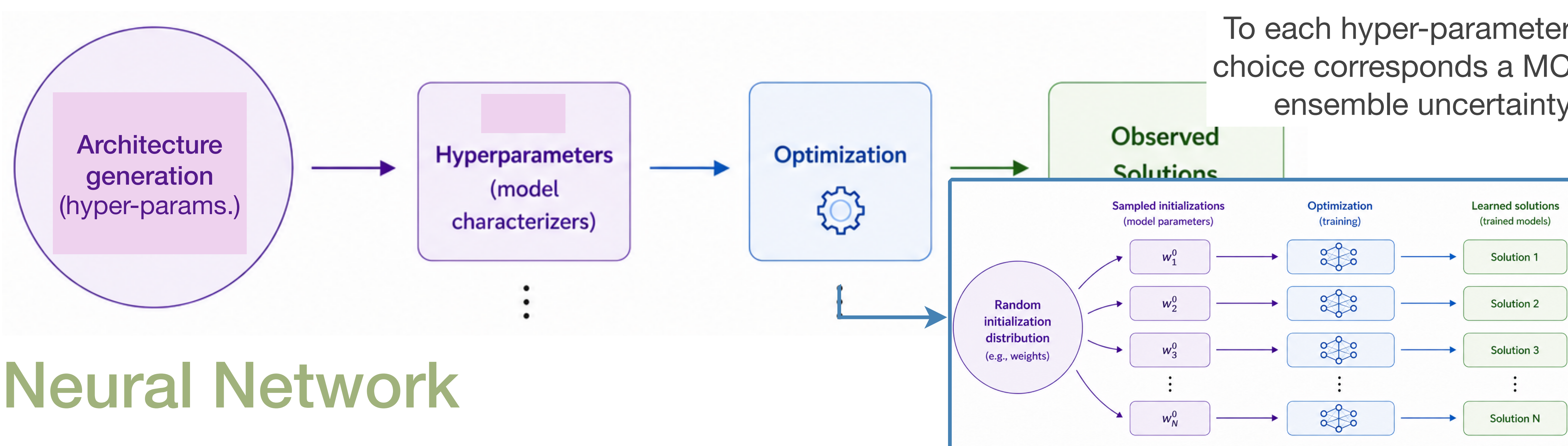
This step is not known yet.

Statistical interpretation

Fantômas



To each solution corresponds a Hessian or MC ensemble uncertainty



To each hyper-parameter choice corresponds a MC ensemble uncertainty

Neural Network

Closure tests

MC formalism:

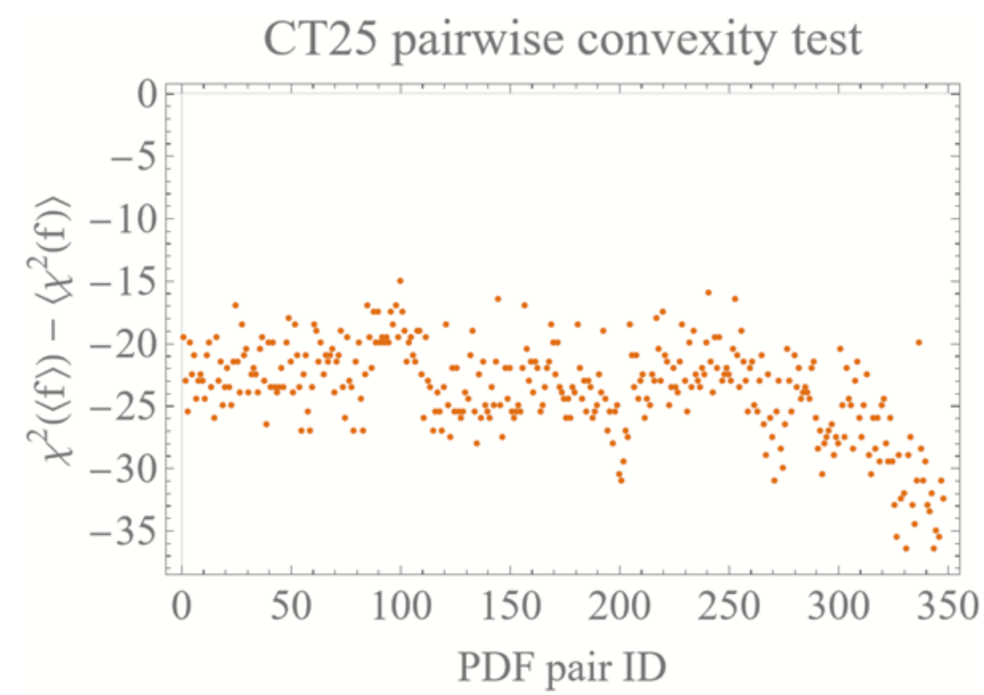
— exploits the concept of distribution in χ^2

— generate pseudo-data from a known result, reapply the fitting procedure, check whether underlying nominal distribution is recovered.

[*e.g.,* NNPDF, Eur.Phys.J.C 82]

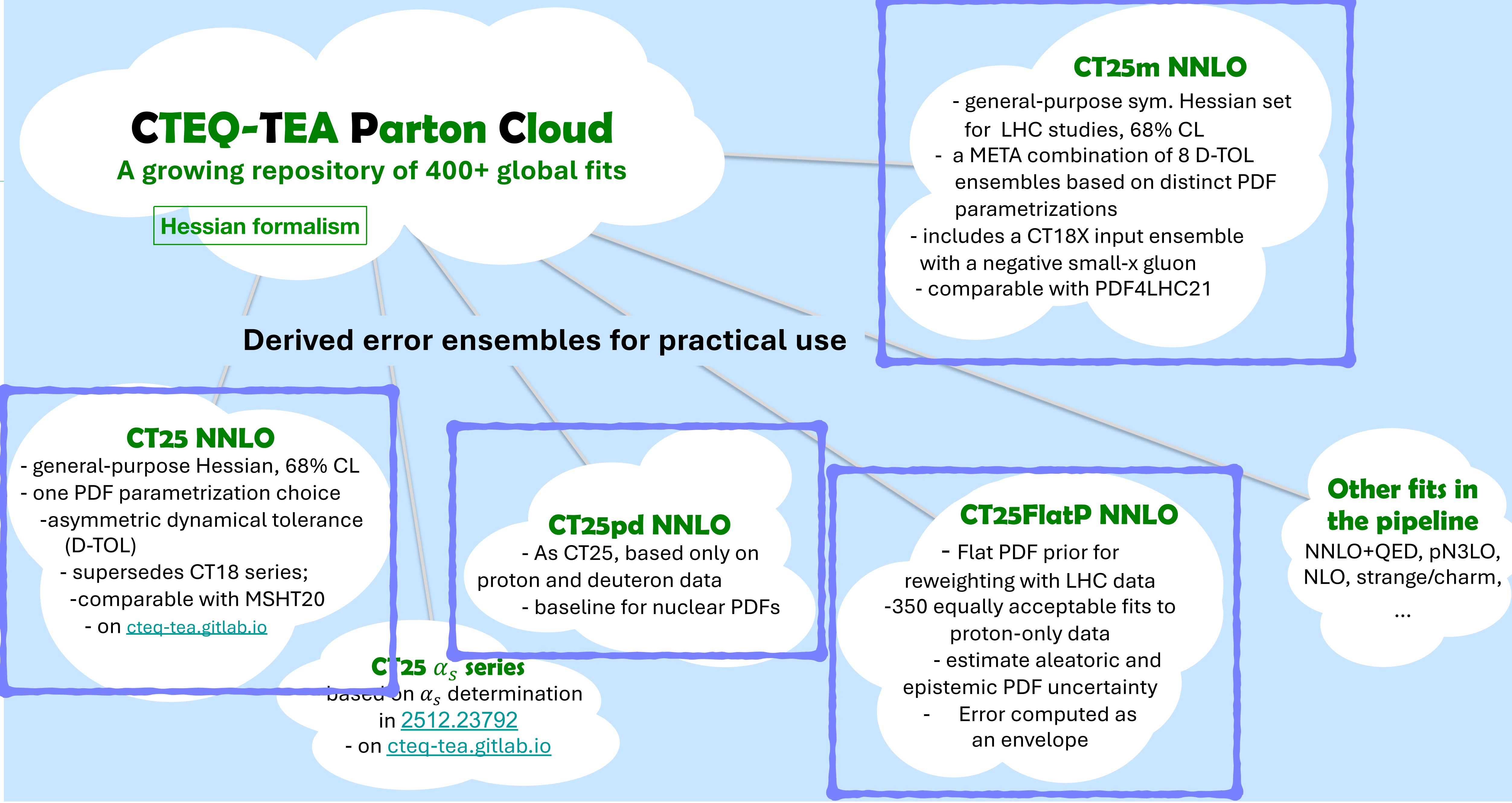
Hessian formalism:

— explore the χ^2 paraboloid (for CT25, it's concave with minimum at χ^2_0)



— MSHT made the exercise of implementing NNPDF-like closure tests for the Hessian formalism

[Harland-Lang, Eur.Phys.J.C 85]



CT25 is a family of 4

with two sets considering **epistemic uncertainty** through the choice of **parametrization**

— CT25m and CT25FlatP

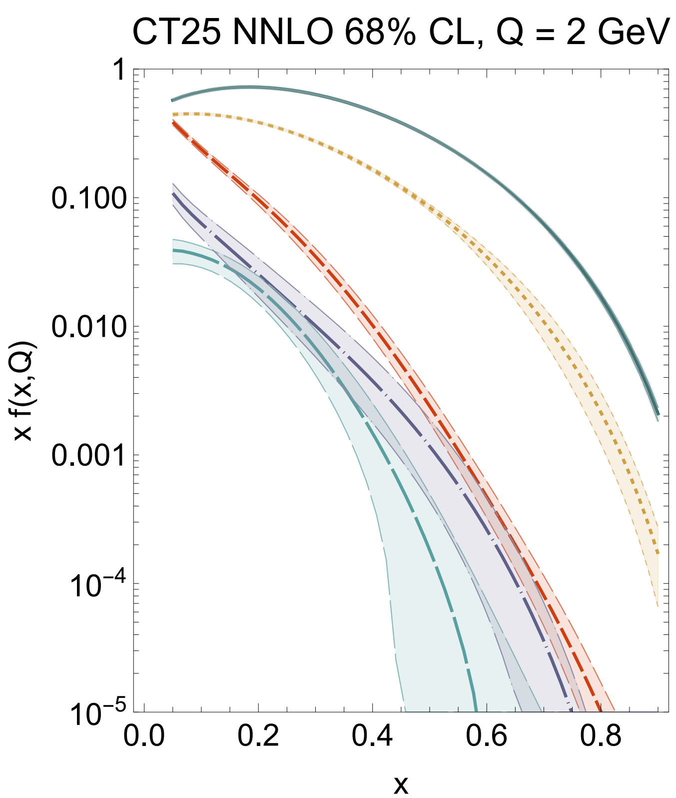
exploration of the **space of solutions**

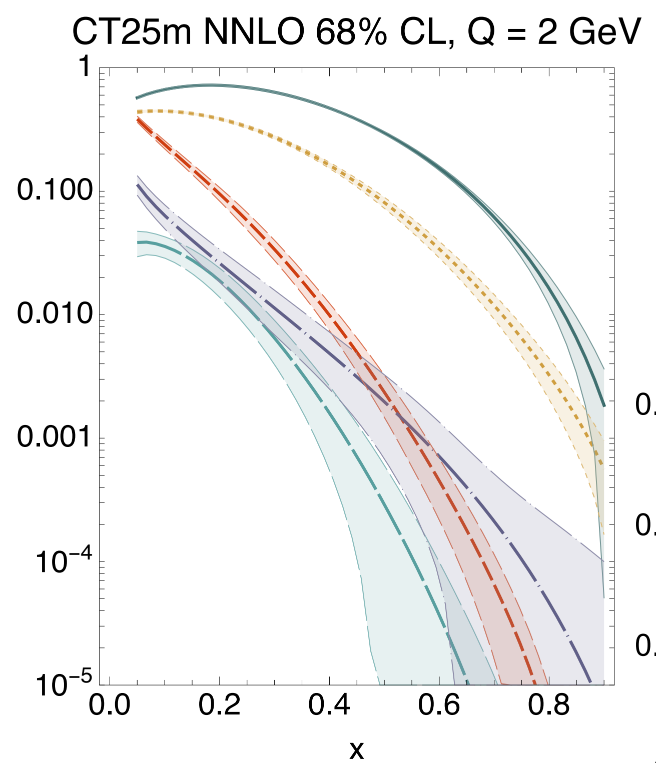
— defining its *boundaries* (CT25m) and populating the *landscape* (CT25FlatP)

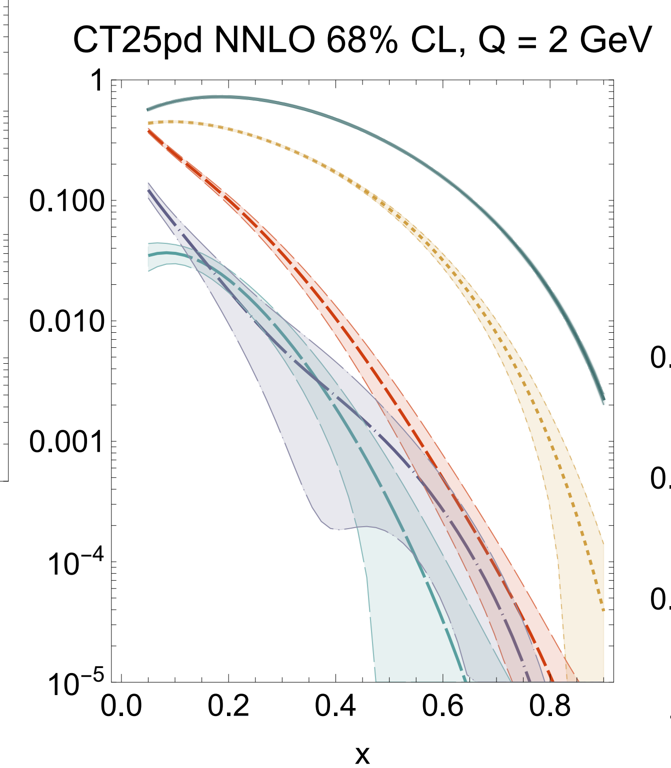
CT15 is a family of 4

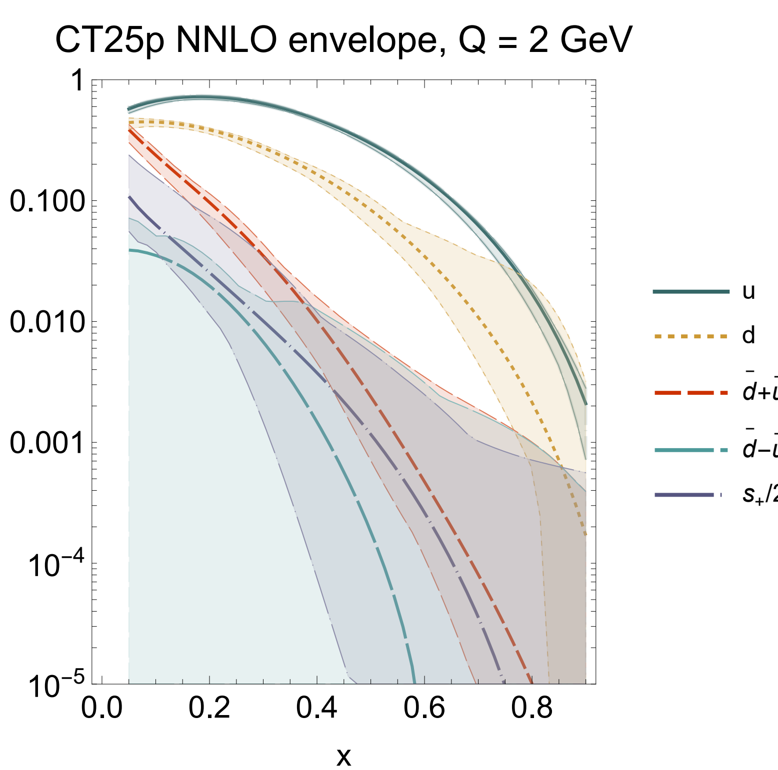
The large-x behavior for the sea is not well determined.

Each set reflects different non-perturbative inputs,
either from PDF model or choice on inclusion of deuteron data.









Nominal CT25 — D-Tol
CT25m — includes model uncertainty
CT25pd — deuteron corrections (less data)
CT25p_flat — flat prior w/o deuteron data

Challenges in quantifying model uncertainty

How well do we understand our landscape?

Tools for exploring the space of solutions

— moving towards non-parametric

— enhanced rôle of priors:

for model distribution, $p(f \mid \theta)$, in exploring the space ; on the parameters, $p(\theta)$, in reducing the space

— propagation of prior distribution?

Classification of solutions beyond goodness-of-fit

— information criteria

Combination of solutions

— statistical meaning of combining solutions

meta-analysis PDF4LHC-like, Gaussian mixture models, Gaussian Processes…

Conclusions

Where do we stand now in terms of epistemic uncertainty quantification? A CT view.

Sampling over models can be cast into sampling over hyperparameters.

We've designed Fantômas4QCD to facilitate the unbiased estimates of parametrization dependence
— Bézier curves to systematize polynomial approximators to functions in $x \in [0,1]$

The statistical interpretation of the "many models" path can be paralleled to a neural-network procedure
— if the latter samples over hyperparameters, too.

CT25 deliverables

Epistemic uncertainty-inclusive sets — building a bottom-up uncertainty prescription

Parton Cloud to explore the boundaries and the topology of the CT25 landscape

Conclusions

Where do we stand now in terms of epistemic uncertainty quantification? A CT view.

Sampling over models can be cast into sampling over hyperparameters.

We've designed Fantômas4QCD to facilitate the unbiased estimates of parametrization dependence
— Bézier curves to systematize polynomial approximators to functions in $x \in [0,1]$

The statistical interpretation of the "many models" path can be paralleled to a neural-network procedure
— if the latter samples over hyperparameters, too.

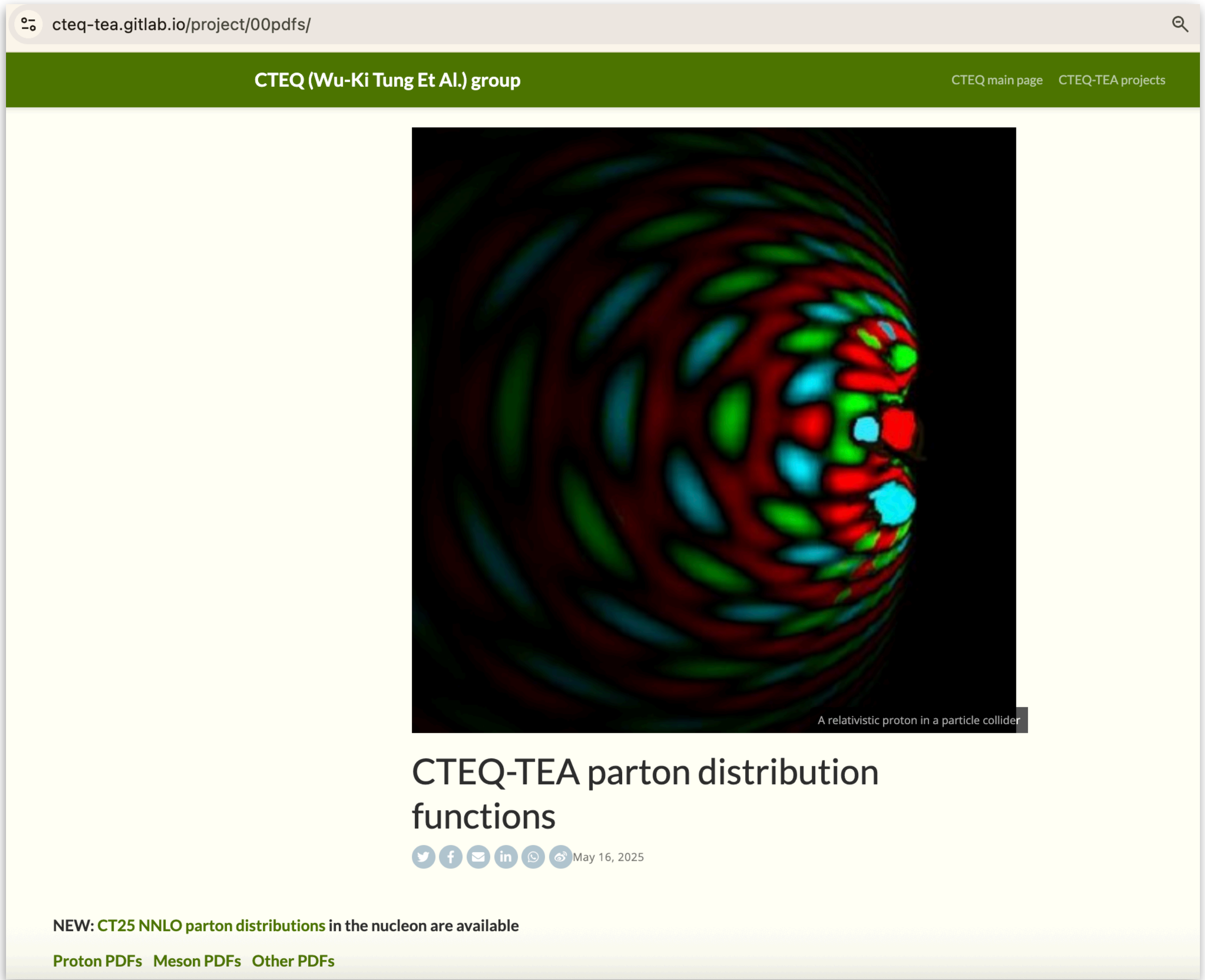
CT25 deliverables

Epistemic uncertainty-inclusive sets — building a bottom-up uncertainty prescription

Parton Cloud to explore the boundaries and the topology of the CT25 landscape

Usage of CT25 epistemic-inclusive sets:

- ☑ CT25m: extrapolation regions, both large- and small-x, important for BSM searches
- ☑ CT25FlatP: studies that need to be free of deuteron-related biases — Drell-Yan,…
  — LHC profiled analyses,…



CTEQ-TEA (Tung Et Al.) group

A. Ablat, A. Courtoy, S. Dulat, Y. Fu, M. Guzzi, T.J. Hobbs, J. Huston, E. Lopez, K. Mohan, P. Nadolsky, M. Ponce Chavez, T. Sharma, D. Stump, K. Xie, C.-P. Yuan and collaborators

Fantômas team

A. Courtoy, T.J. Hobbs, L. Kotz*, P. Nadolsky, F. Olness, M. Ponce Chavez

CT25 PDFs and extended family:
https://cteq-tea.gitlab.io/project/00pdfs/

Backup

State-of-the-art

Goal: find the shape of PDFs on $x \in [0,1]$ in N_{flavor} space

⟹ Data-driven *global QCD analyses* are inverse problems that aim to infer PDFs

⟹ PDFs as access to underlying non-perturbative QCD dynamics

⟹ Many models are possible — more in a few slides

⟹ Approaches:

- Explicit on the parameters: use polynomial approximation — (model+) parameter inference
  use neural networks — architecture infers the behavior of the system

MSHT
CT
…
Fantômas

NNPDF

Flexibility, bias-variance tradeoff

More in C. Schafer's response

- Very non-parametric: kernel methods, Hausdorff inverse problem, …— still exploratory

Candido et al
Yndurain, Hobbs et al
…

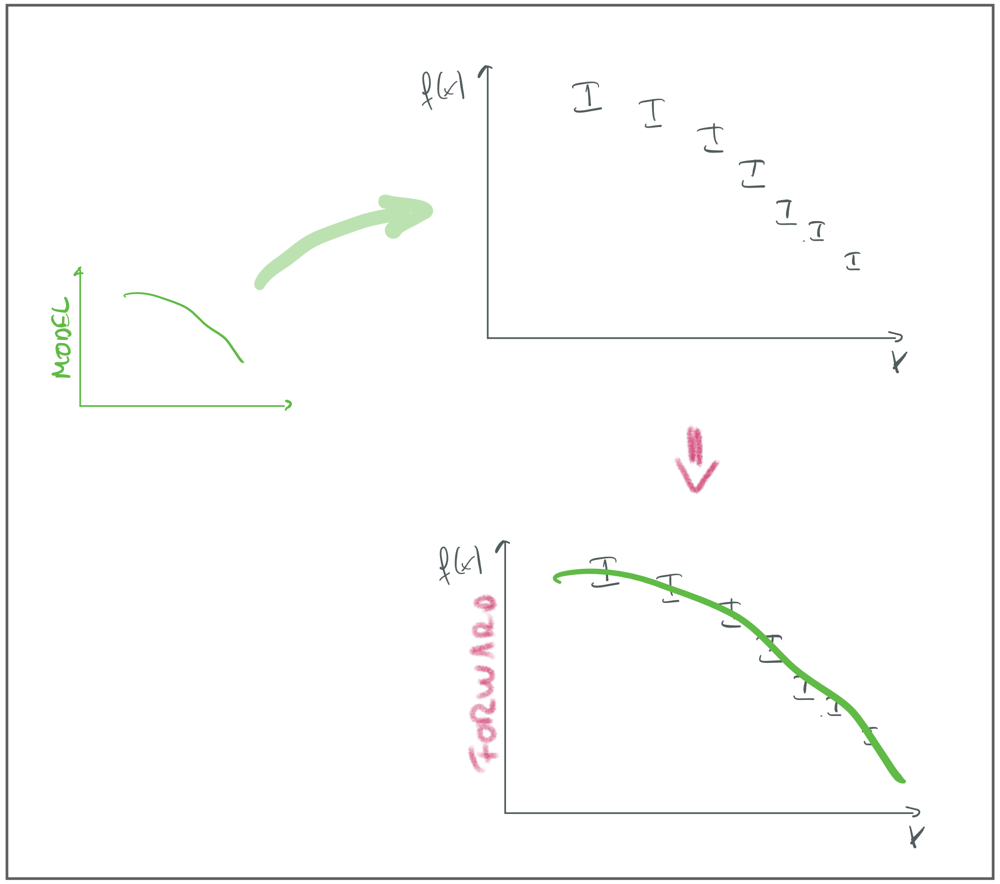
⟹ Parametrization bias new to the game —it contributes to the **epistemic uncertainty**

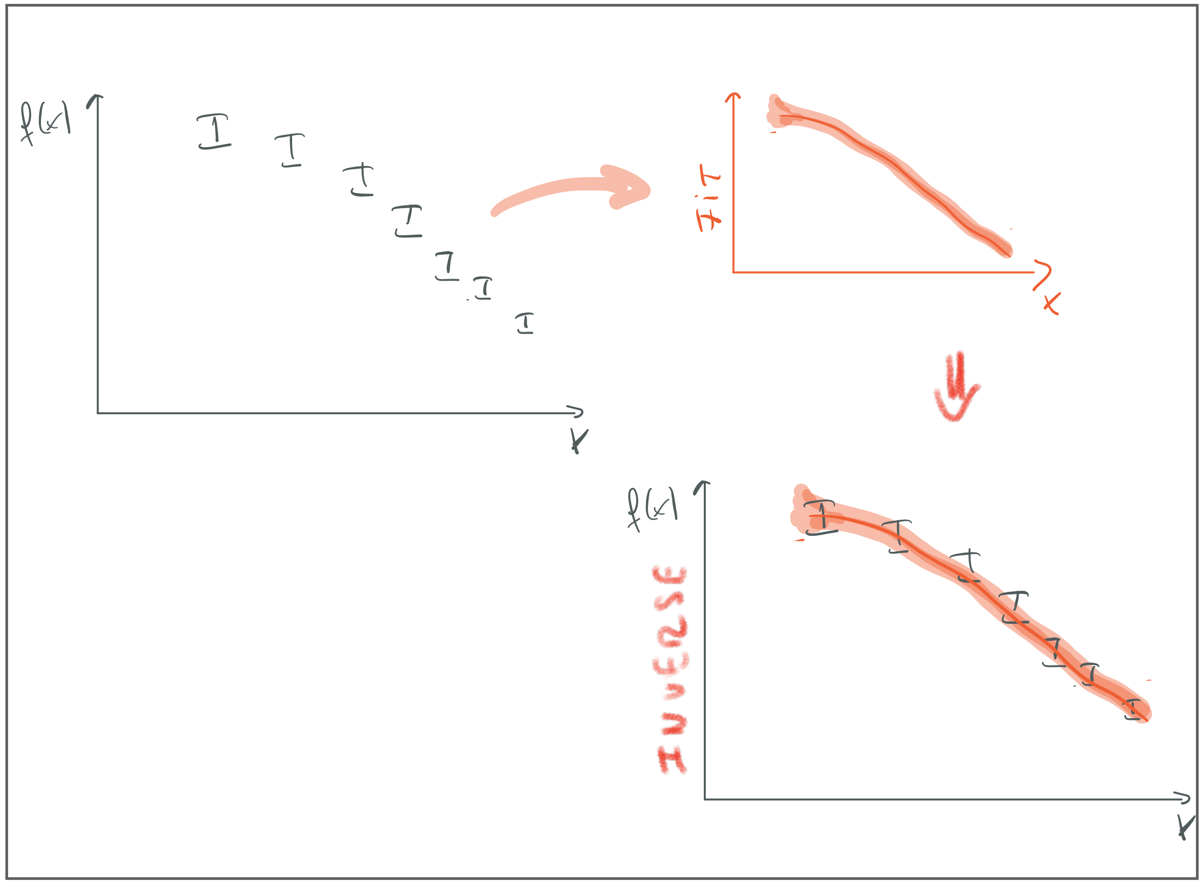
⟹ **Model uncertainty is mostly Bayesian-like:**

CT

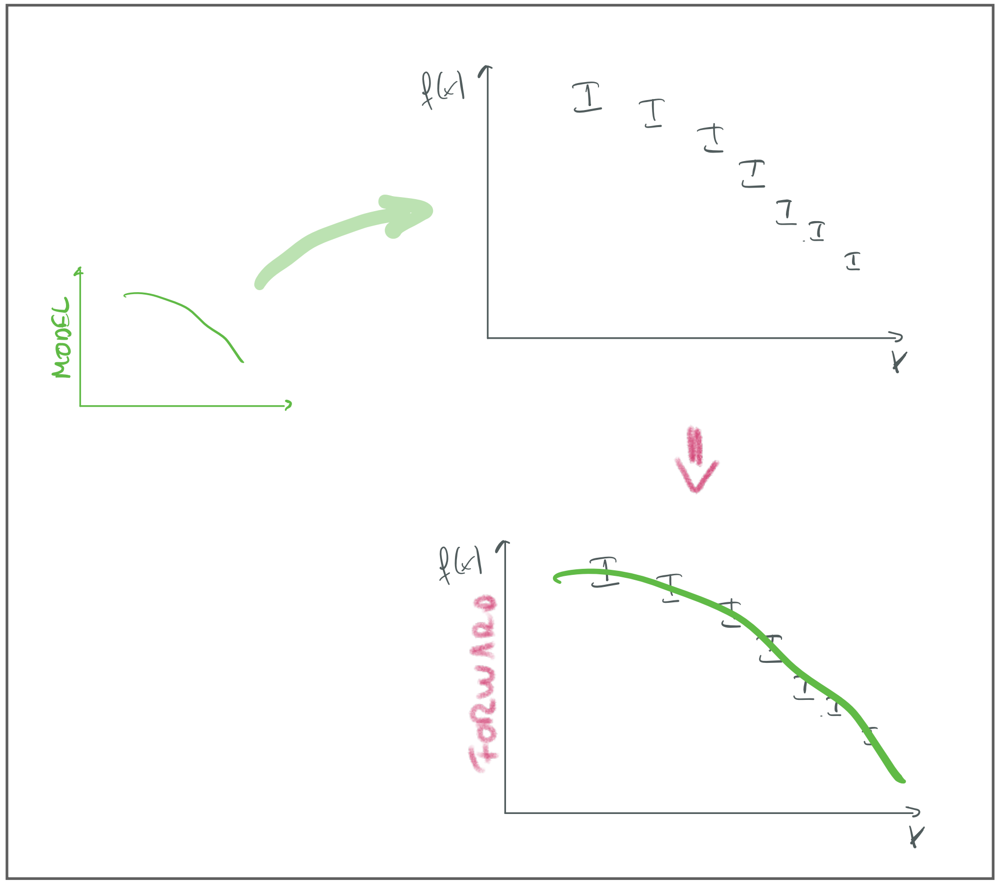
generate sets of approximating functions
combine solutions — encompassing the underlying truth

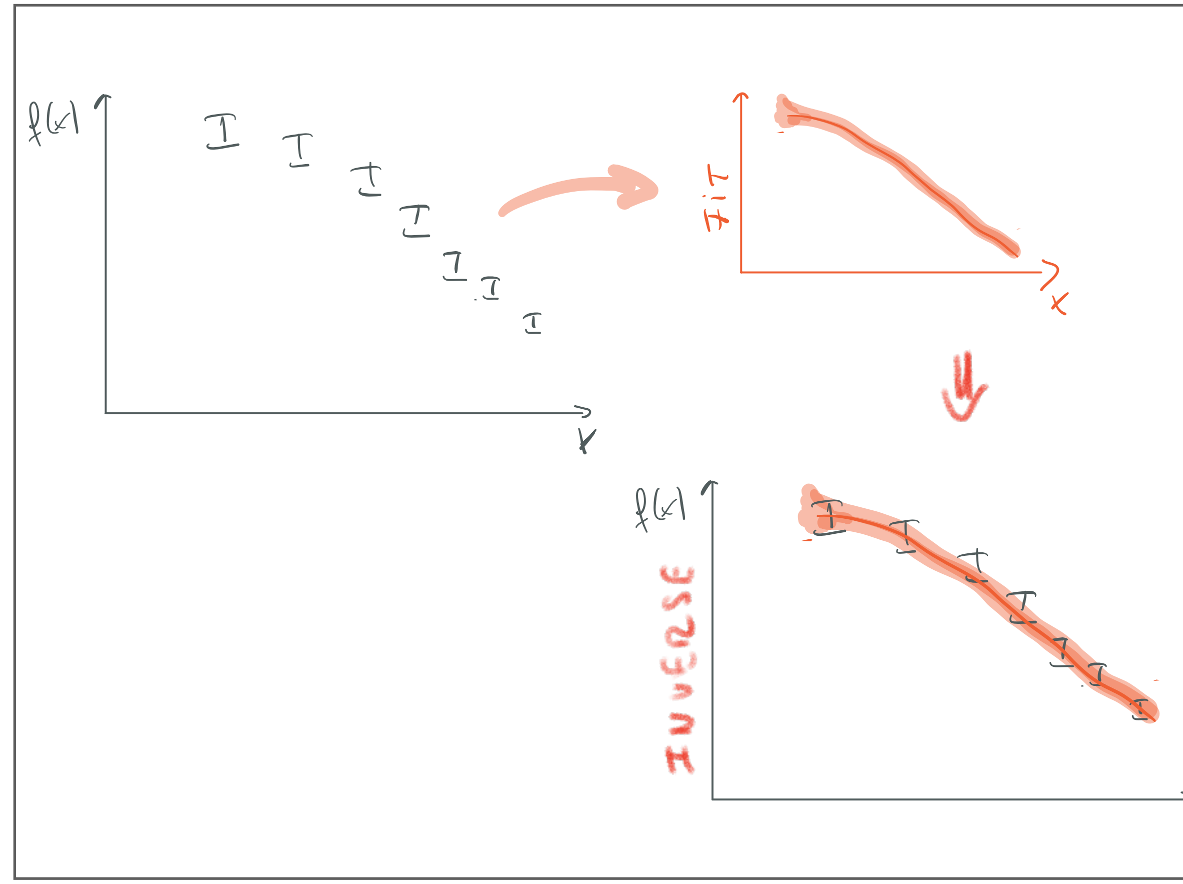
Forward and inverse problems illustrated





Forward and inverse problems illustrated





Sufficient and necessary conditions (⇔)

- Sufficient — a model agrees with data via a measure of distance, typically χ^2
- Necessary — the data imply that the model is the truth,
  not possible in convolutional inverse problems,
  not possible with sizable data uncertainty.

e.g. [AC & Nadolsky, PRD 103]

Leading/established models

⟹ polynomial approximators (~ 30 (hyper) params) —Stone-Weierstrass theorem

⟹ neural networks (dimension of hyper parameter space+…) — universal approximation theorem

More in C. Schafer's response

(Approximate) linearity of the PDFs in the parameters

On (non) linearity of the observables in the parameters
Costantini et al [2404.10056]

Little explored analyticity (integrability, smoothness) and Vandermonde matrices

NNPDF
Fantômas

<u>**Caveats:**</u> Can we define overfitting from that analytical view point?

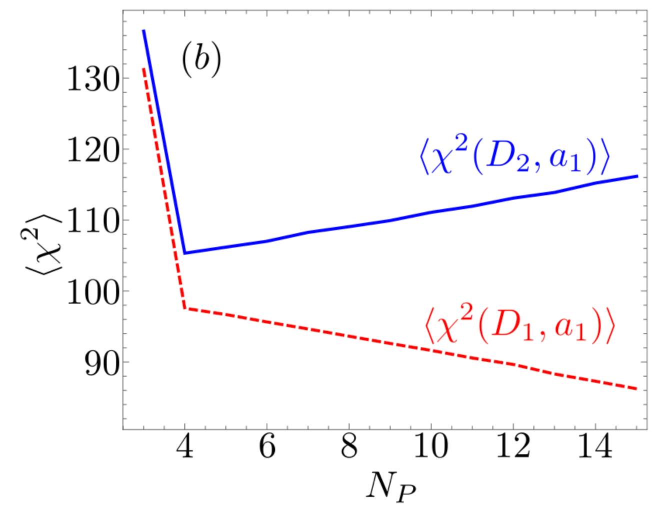
⟹ subjective wiggliness, kinkiness,…

⟹ subjective tradeoff between flexibility and overfitting

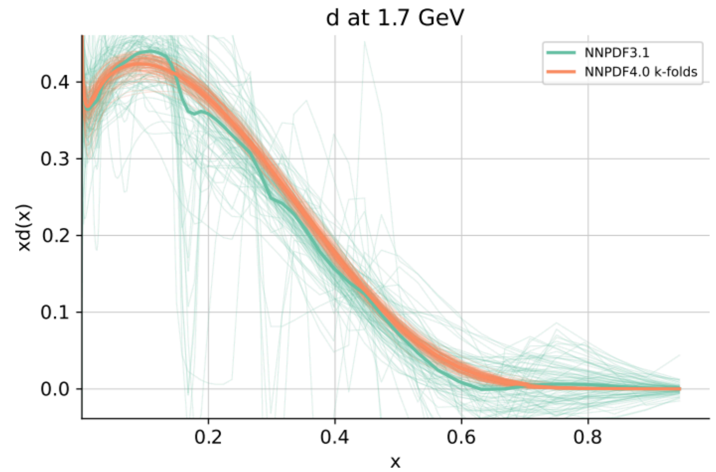
Kovařík et al [1905.06957]
NNPDF [2211.12961]
[WIP]

Dependence on the number of parameters

Stopping criteria, cross-validation
Goodness-of-fit criteria



Kovařík et al [1905.06957]



NNPDF4.0

"Parametric" approaches: Identify flat directions in the space of solutions
"Non-parametric" approaches: regularization methods (?)

$\Big\}$ Effective dimensions

Definition of the **number of degrees of freedom**

Information criteria can penalize models whose degrees of freedom are heavily affected by the number of params — AIC,…

Mainly used by the lattice community, based on *[Jay & Neil] [WIP]*

PDF models are not expansions, there is no truncation over parameters. Instead, the # of params is inherent to each model, with possible mapping among models (polynomial mimicry, Bézier curves).

More tests of consistency

Closure tests

Assuming a true PDF distribution, test of consistency of the methodology

Do the fits to pseudo data from assumed distribution reproduce said distribution?

Introduced by NNPDF, justified by the complexity of their landscape ;
explored for Hessian formalism by MSHT in defining confidence intervals**

NNPDF4.0
MSHT [2407.07944]

Convexity tests

What is the shape of the figure-of-merit (log-likelihood) close to the minimum?

Convexity of χ^2 built from multiple models tested, pursued in Hessian formalism by CT

More in Huston and Mohan's talks tomorrow

** see session 1 this morning

New directions

Exploratory stage, non-parametric

⟹ **Gaussian Processes** — generation of models from the prior

Candido et al [2404.07573]

⟹ **Fantômas** — generation of models, and combination from the posterior

Fantômas, Kotz et al [2505.13594, 2507.22969]

⟹ **Flexible bases** — “hopscotch” algorithm, proper orthogonal decomposition

Courtoy et al [2205.10444], Costantini et al [2507.16913]

⟹ **Hausdorff problem** — learning x-dependence from Mellin moments

Kriesten & Hobbs [2312.02278],…

Key role played by priors

Priors have been identified to reduce the phase space

Constraints fit exploits the benefits of well-controlled priors

"Constrained curve fitting," [Lepage et al., Nucl.Phys.B Proc.Suppl.106(2002) 12-20] — lattice oriented

⟹ similarly for polarized PDF analysis [Benel et al, EPJC]

Solutions may be prejudiced by strong priors

Some publications show how strong priors have affected results (that has led to important claims)

⟹ Proton structure: "Parton distributions need representative sampling"+ communication with NNPDF

[us]

⟹ Neutrino physics: "Neutrino mass and mass ordering: no conclusive evidence for normal ordering"

[Stefano Gariazzo et al JCAP10(2022)010]

Definition of relative/pairwise evaluators

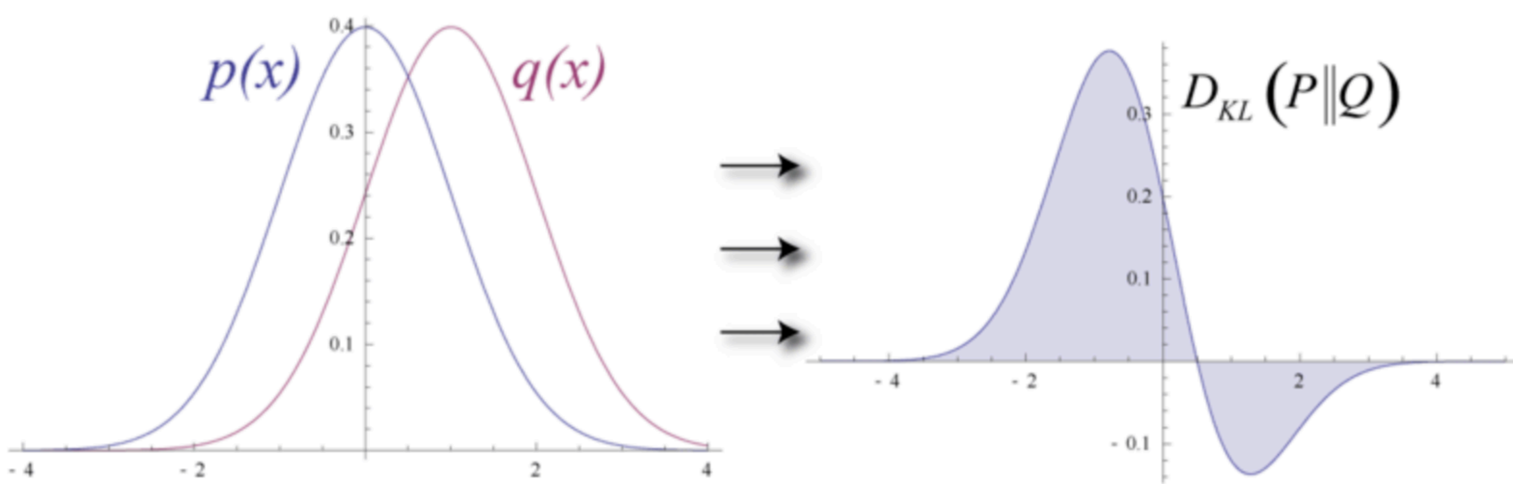
<u>Relative evaluators rely on a reference shape or distribution</u>

— ***information-theoretic criteria***: divergences are *relative* or cross entropies

$$D_\alpha(p\,||\,q) = \frac{1}{\alpha - 1} \ln\left(\sum_{i=1}^{n} \frac{p(x_i)^\alpha}{q(x_i)^{\alpha-1}}\right), \quad \alpha \neq 1$$

KL divergence for $\alpha = 1$ — *relative* version of Shannon entropy.

Rényi divergence for $\alpha \neq 1$.



Original Gaussian PDF's

KL Area to be Integrated

Definition of relative/pairwise evaluators

Relative evaluators rely on a reference shape or distribution

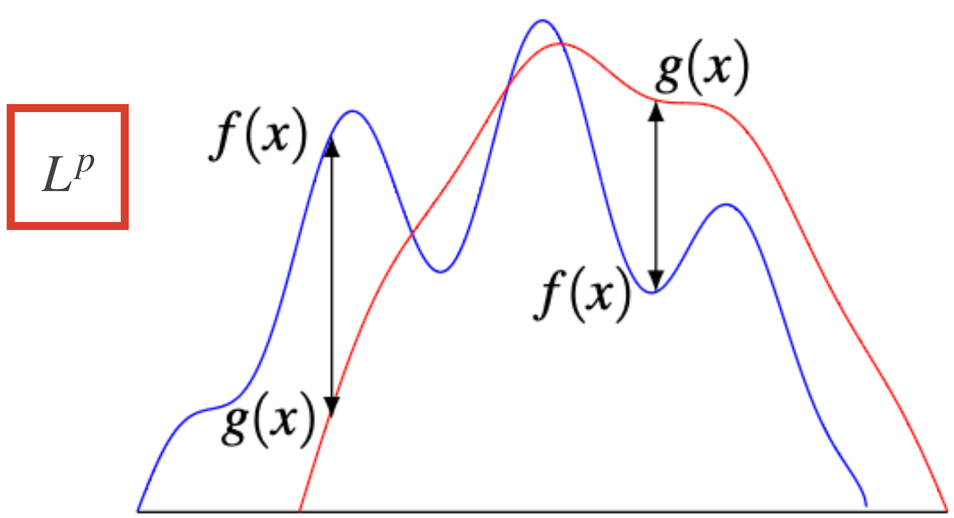
— ***metrics*** like L^p are often used to reflect vertical displacements,
the optimal-transport Wasserstein distance W_p captures horizontal displacements

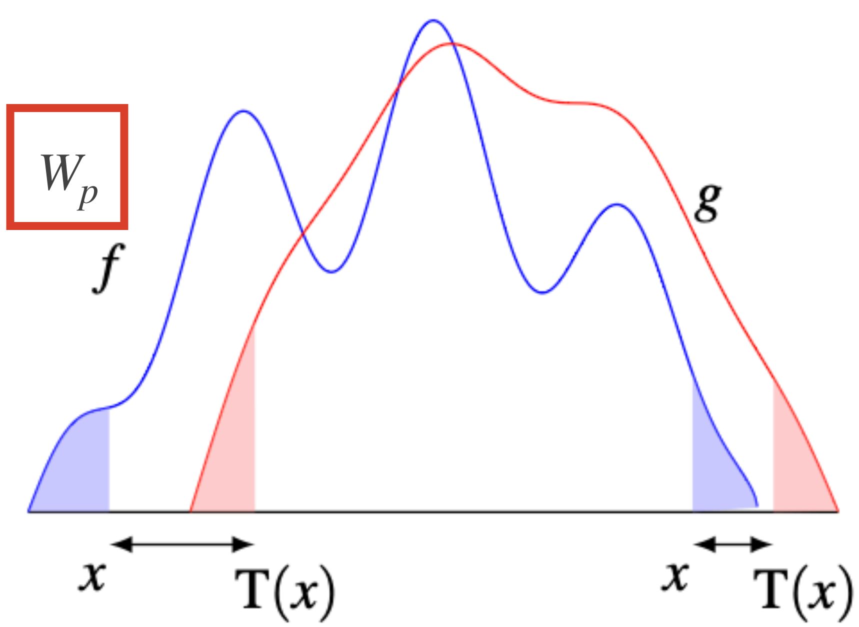
For a translational transport in 1D, $\gamma(x, y) = Q(x)\delta(y - T(x))$,

$$W_p(P, Q) = \left(\int_\Omega |x - T(x)|^p \, dQ(x) \right)^{\frac{1}{p}}$$

e.g. , F. Santambrogio, *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*

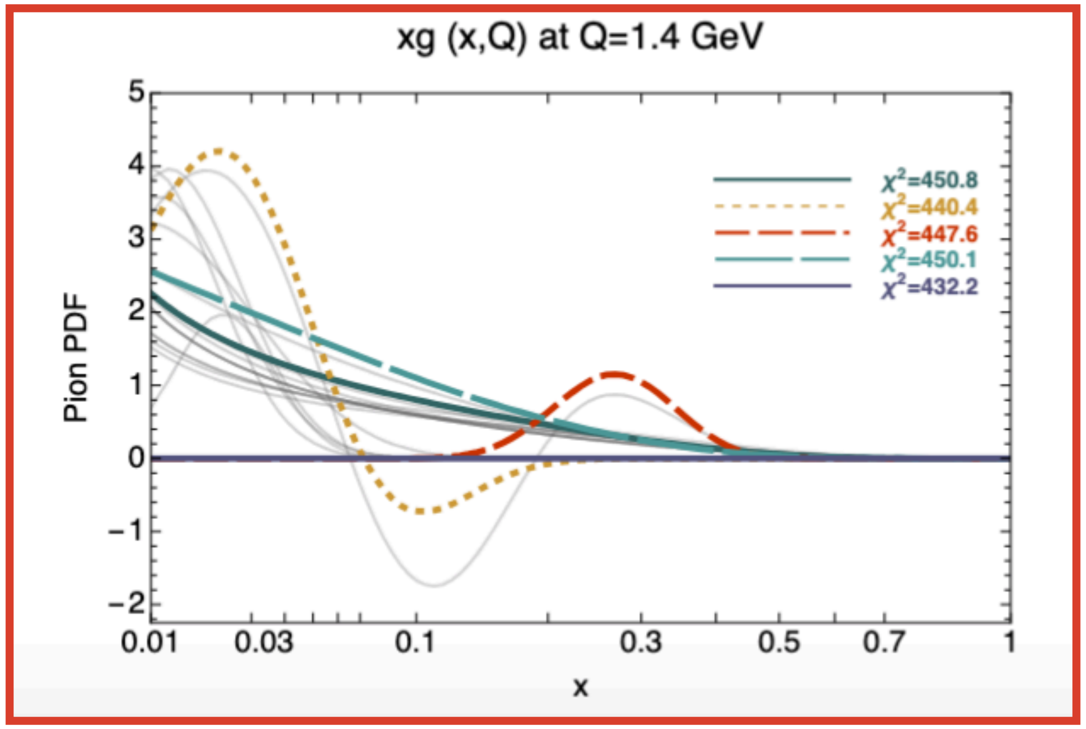
C. Villani, *Optimal Transport: Old and New*





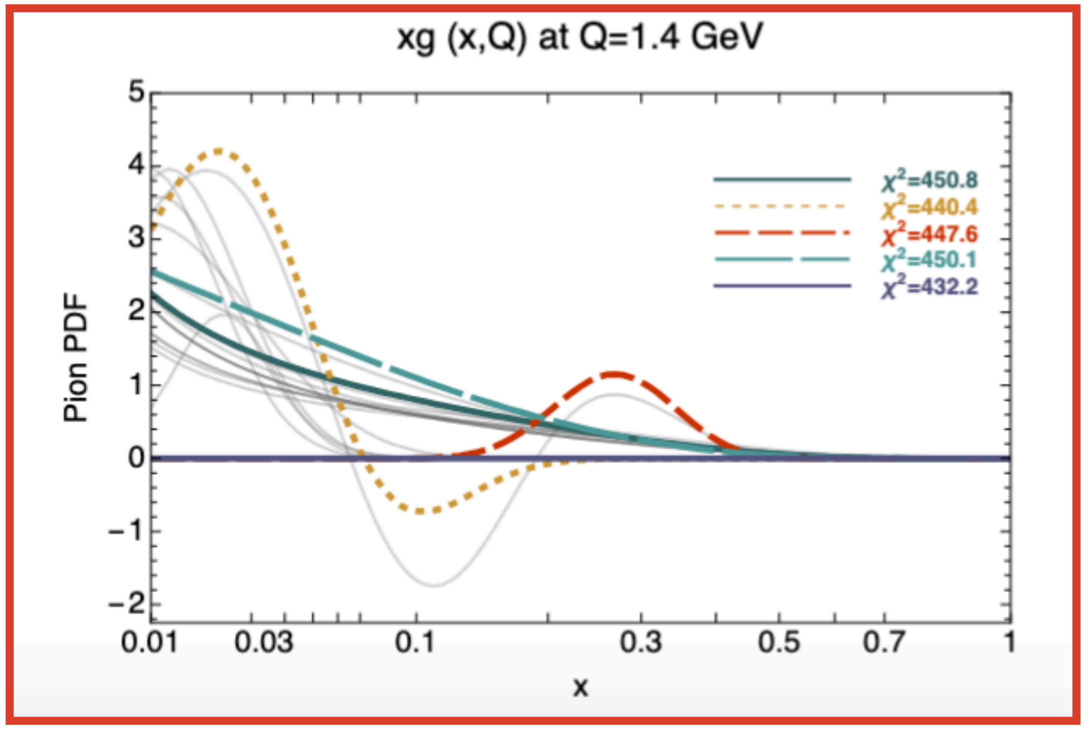
Pion PDF selection

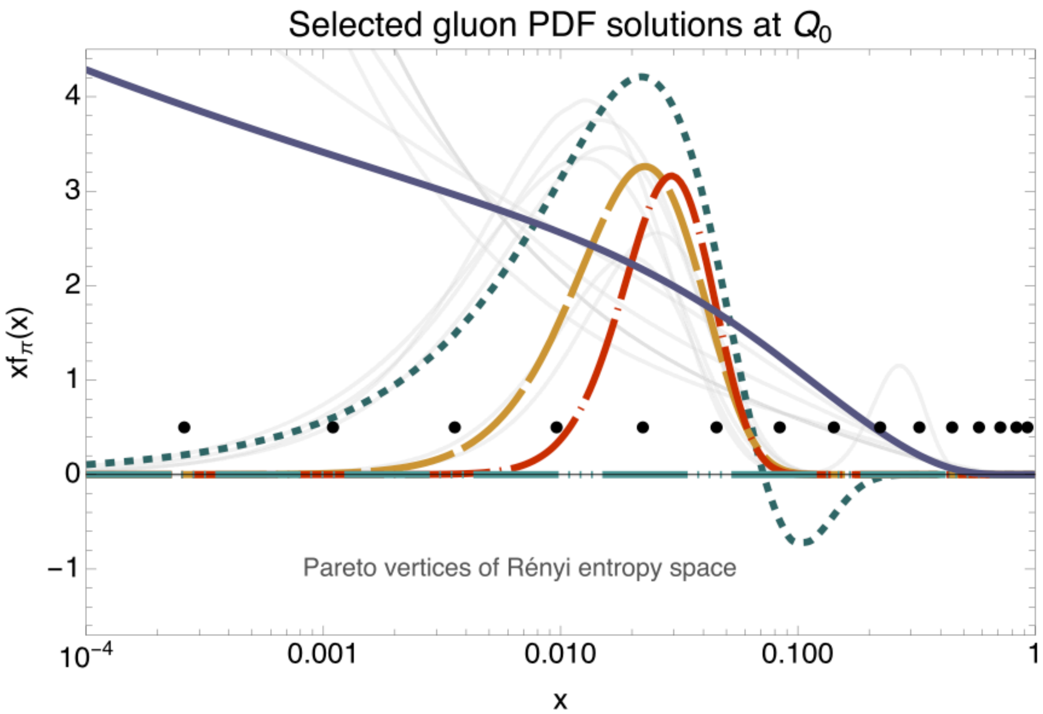
Baseline / published



Pion PDF selection

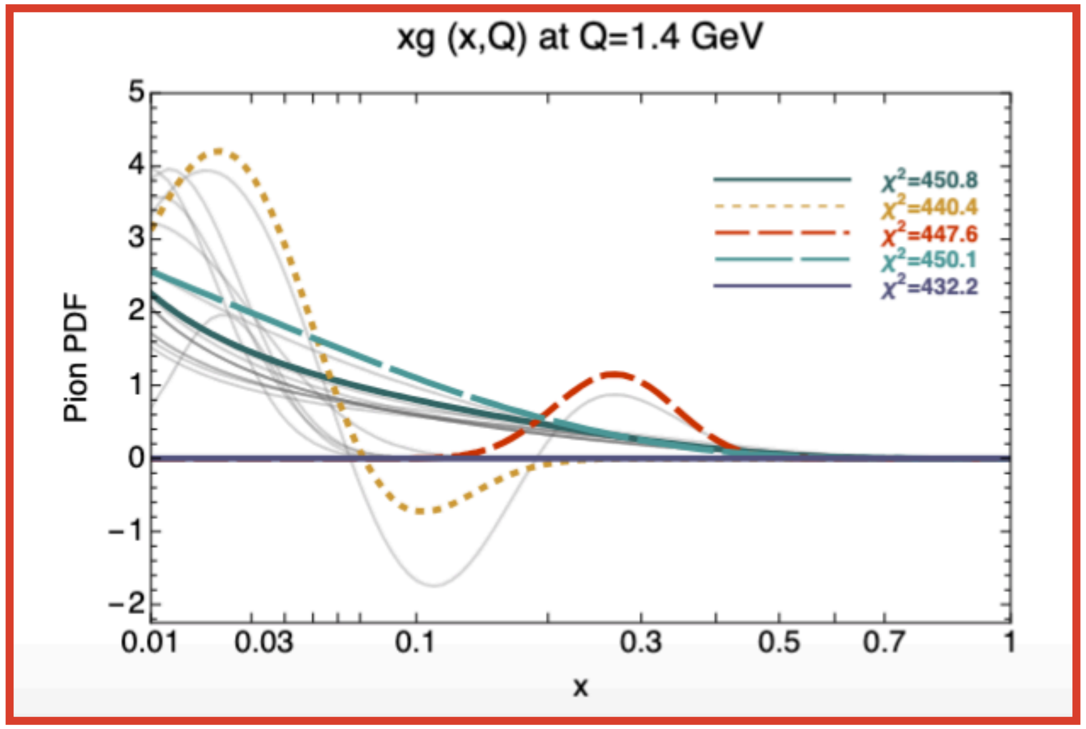
Baseline / published

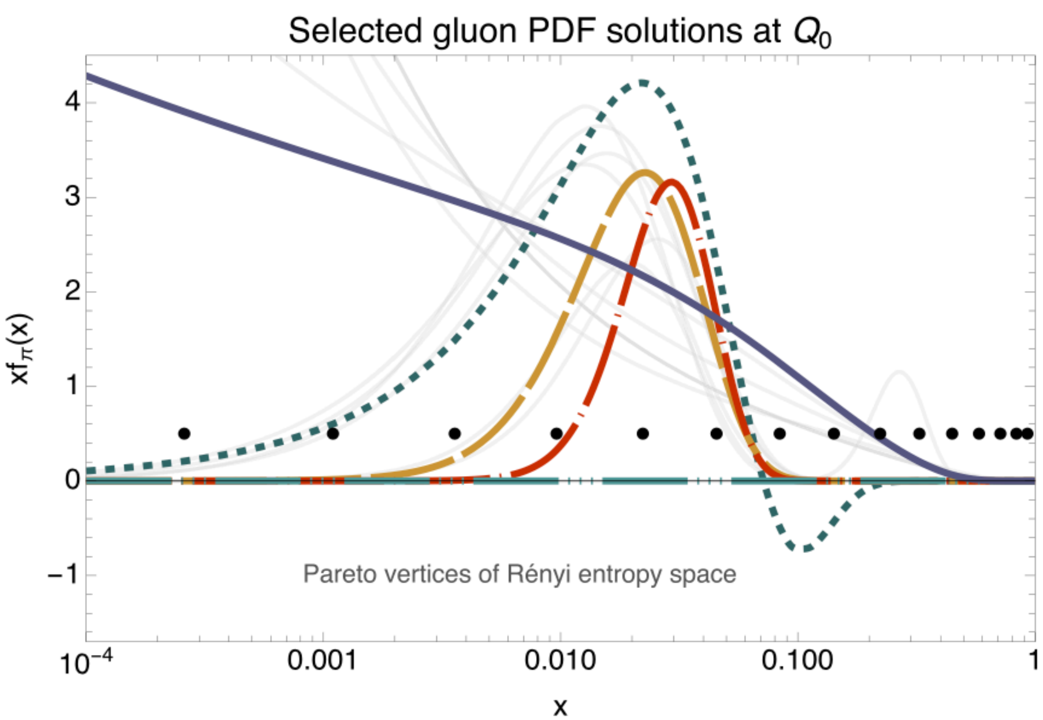


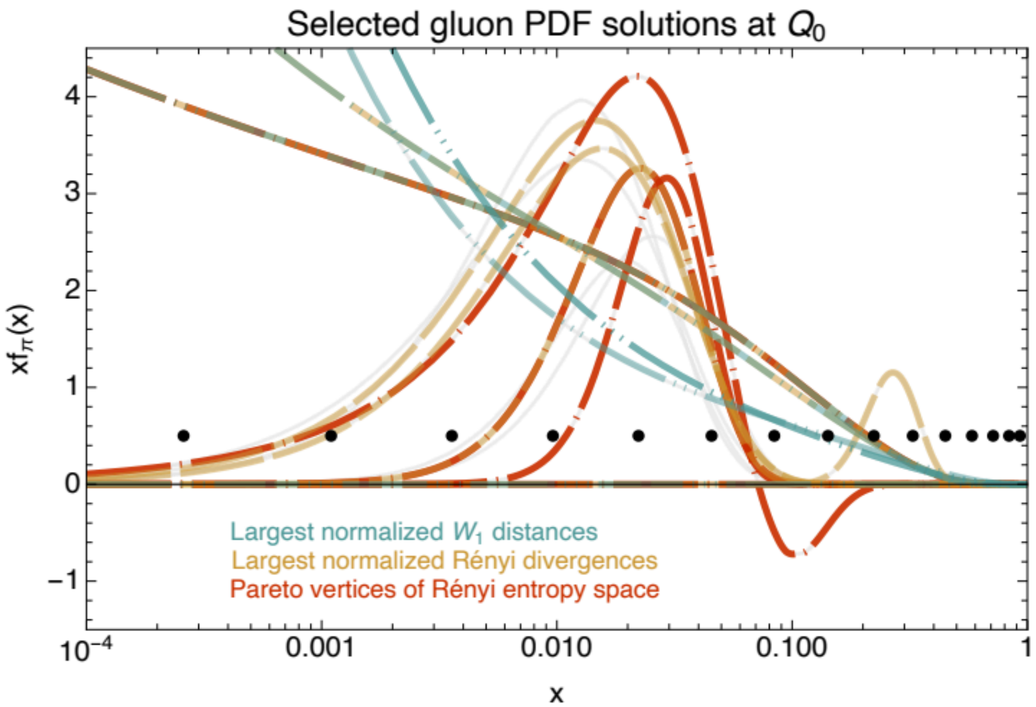


Pion PDF selection

Baseline / published

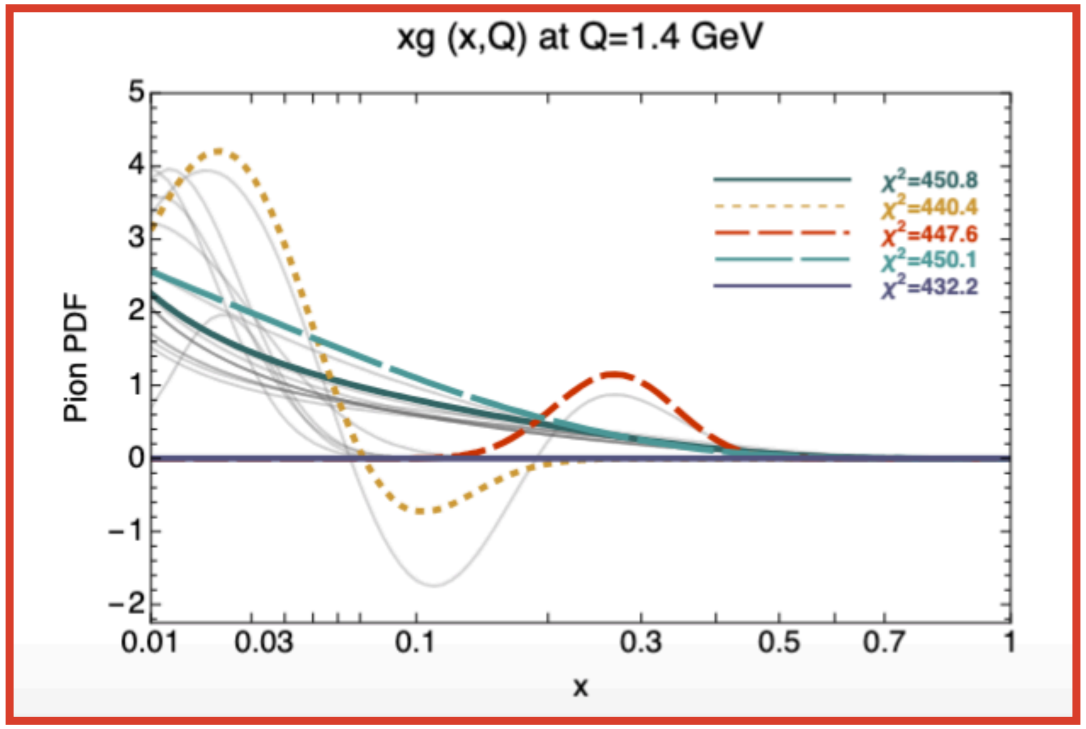


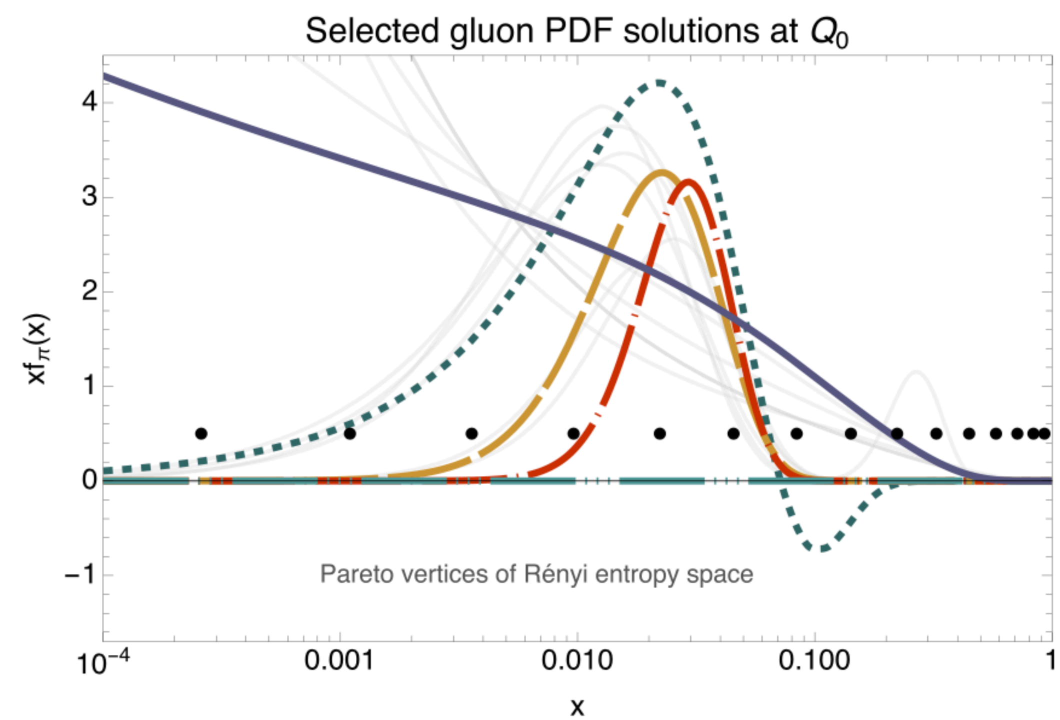


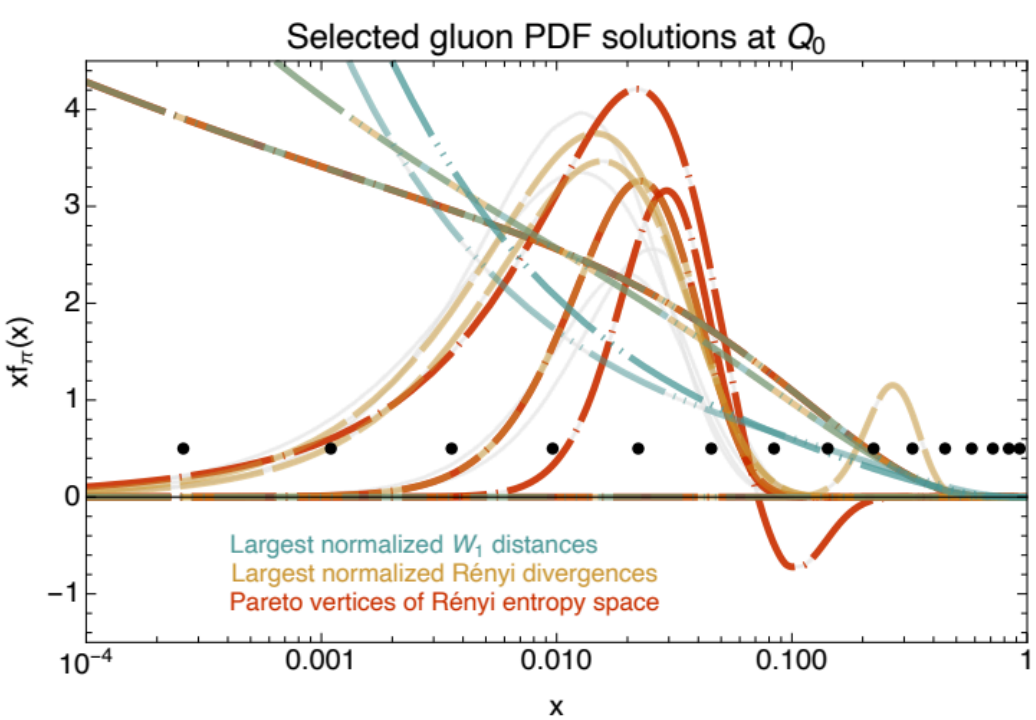


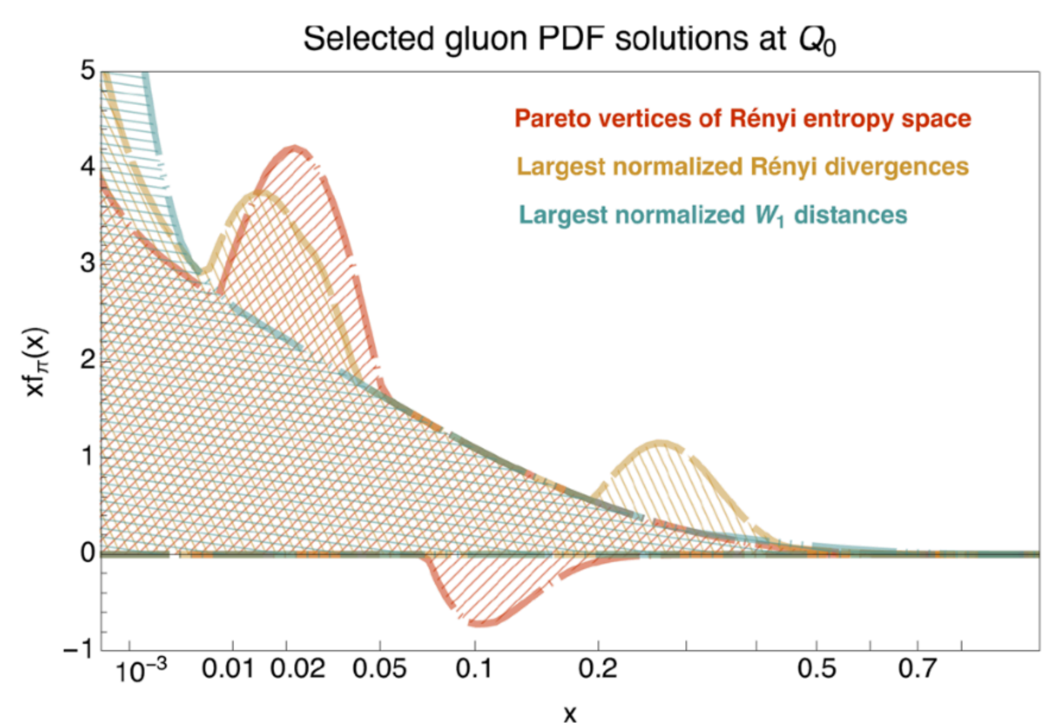
Pion PDF selection

Baseline / published







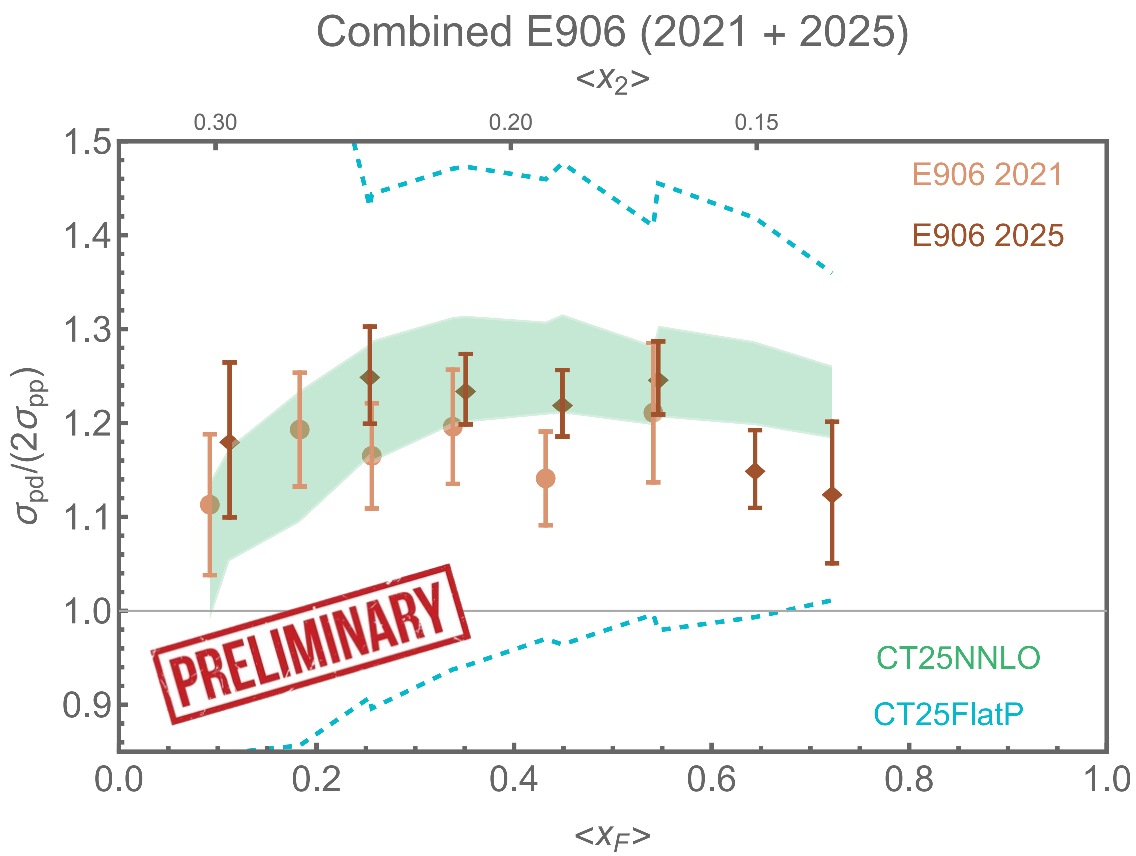


New SeaQuest data

December 2025:

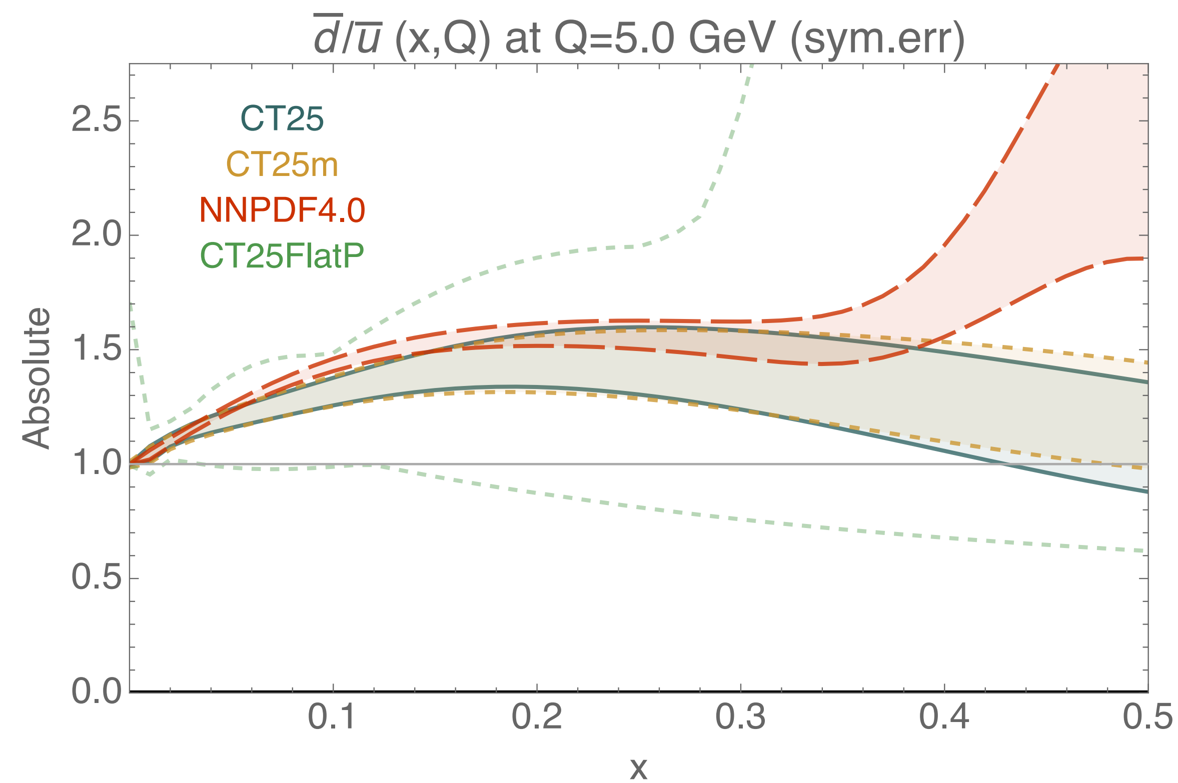
CT25 includes E906 (2021)

Postdictions for E906 (2025).



$\bar{d}/\bar{u}$ ratio:

Expected leading sensitivity.

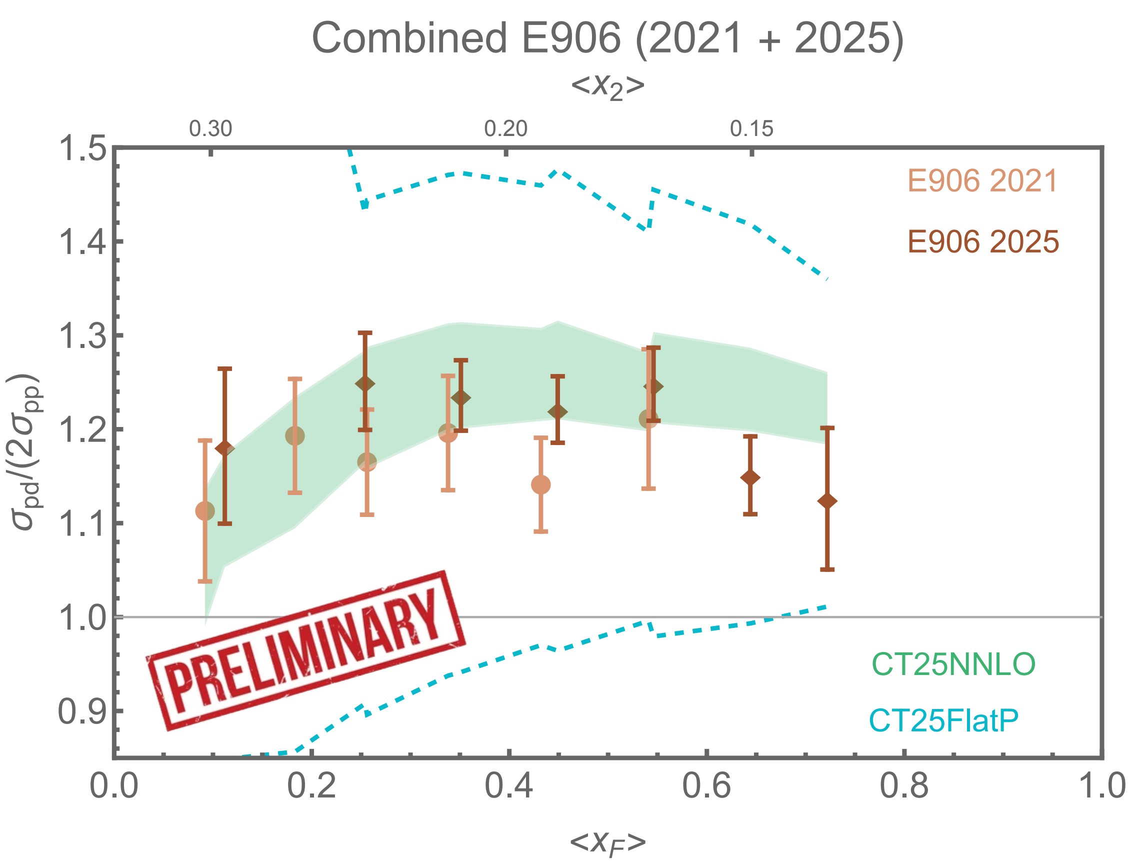


New SeaQuest data

December 2025:

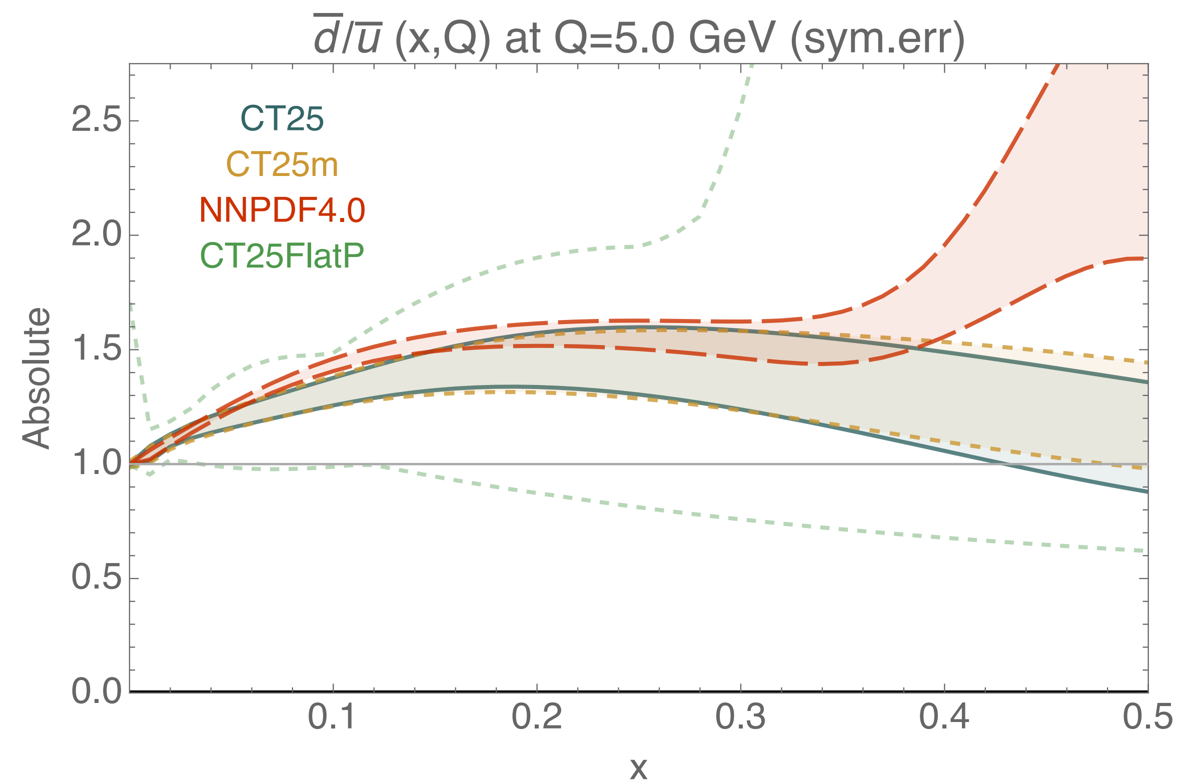
CT25 includes E906 (2021)

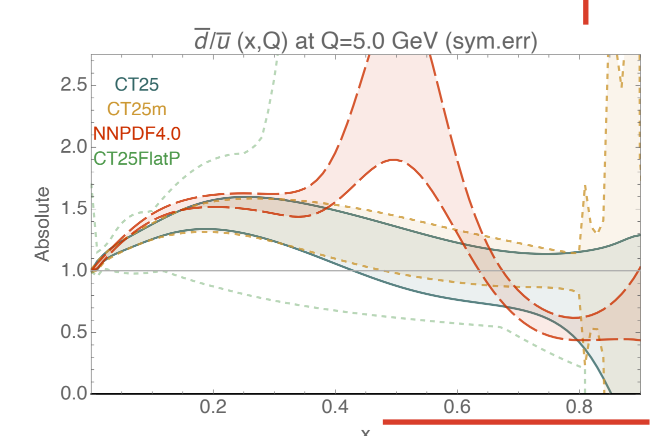
Postdictions for E906 (2025).



$\bar{d}/\bar{u}$ ratio:

Expected leading sensitivity.





Impact on larger Q^2 data

Forward-backward asymmetry at large invariant mass A_{FB}:

Sensitive to PDF ratios $\bar{u}/u, \quad \bar{d}/d$

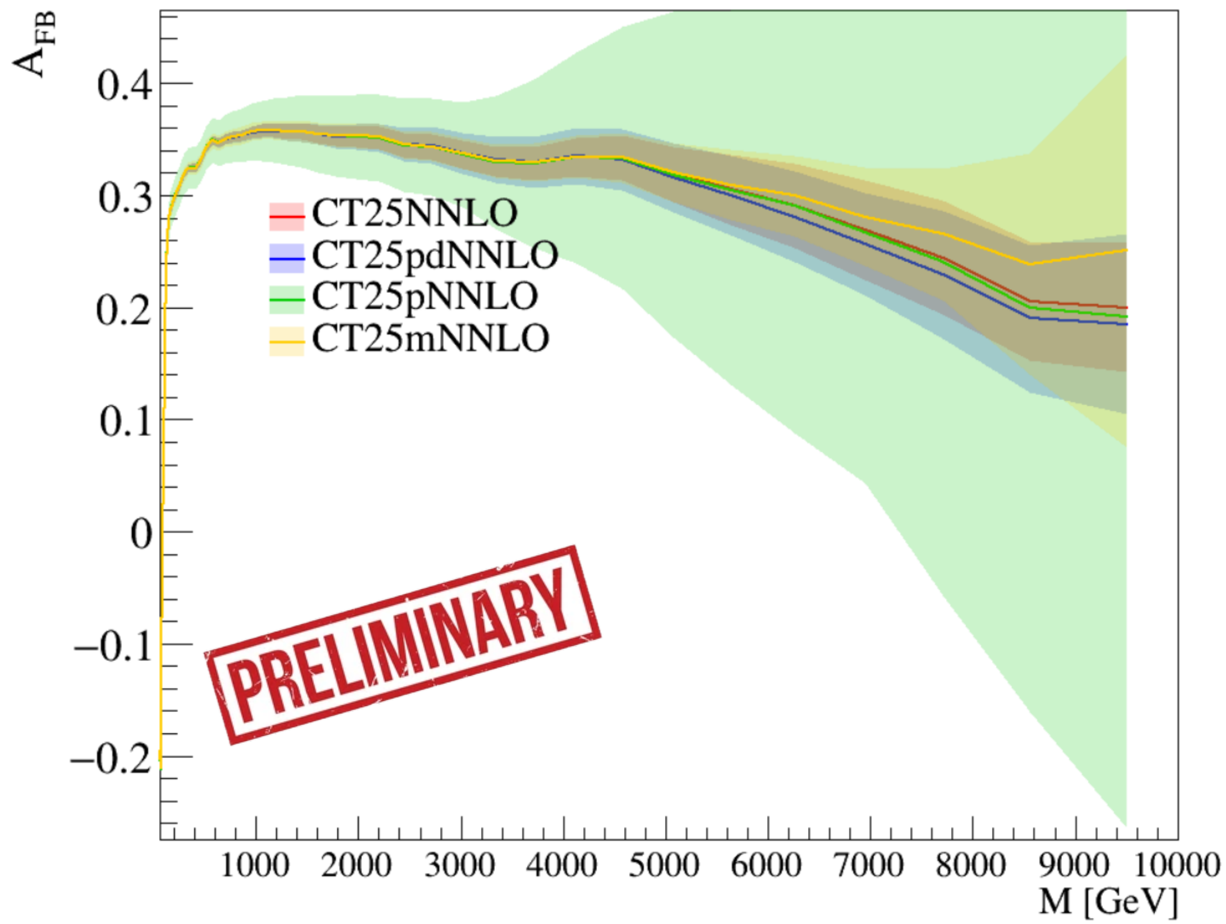
[2307.07839]

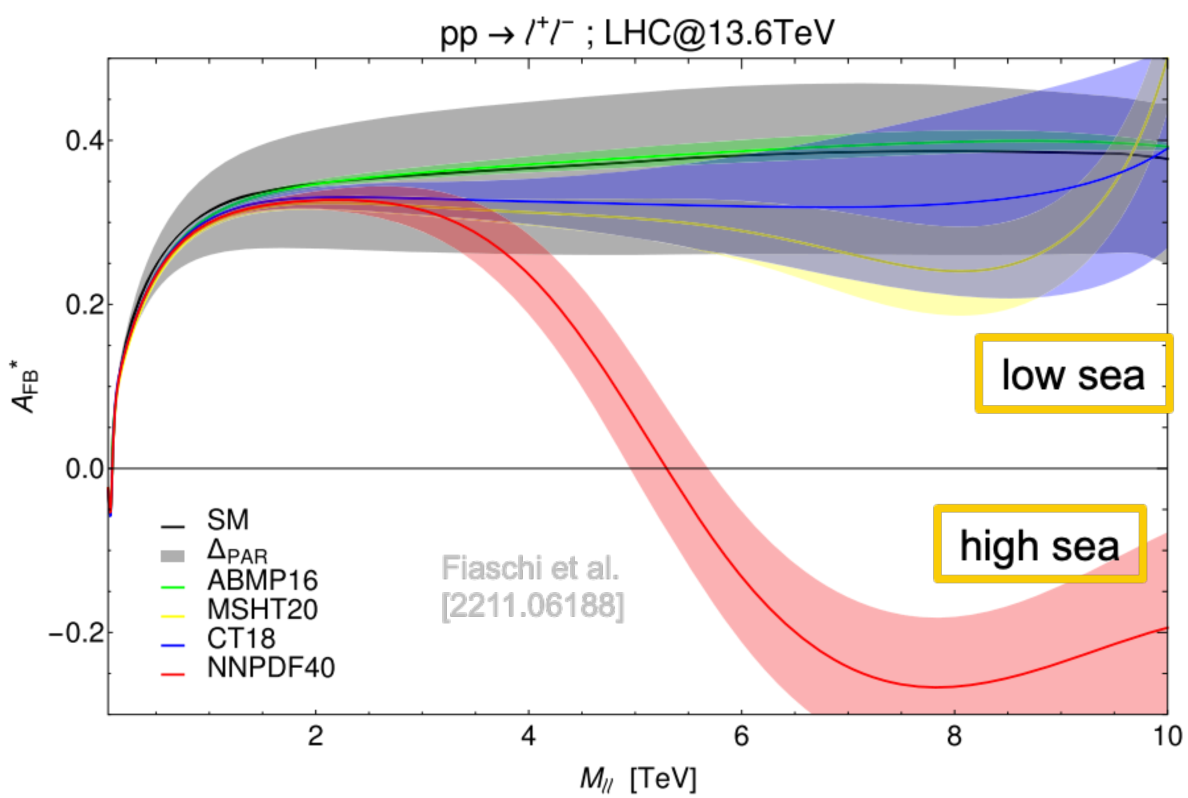
CT family — low-sea scenarios ;

compatible with positive A_{FB} at large mass

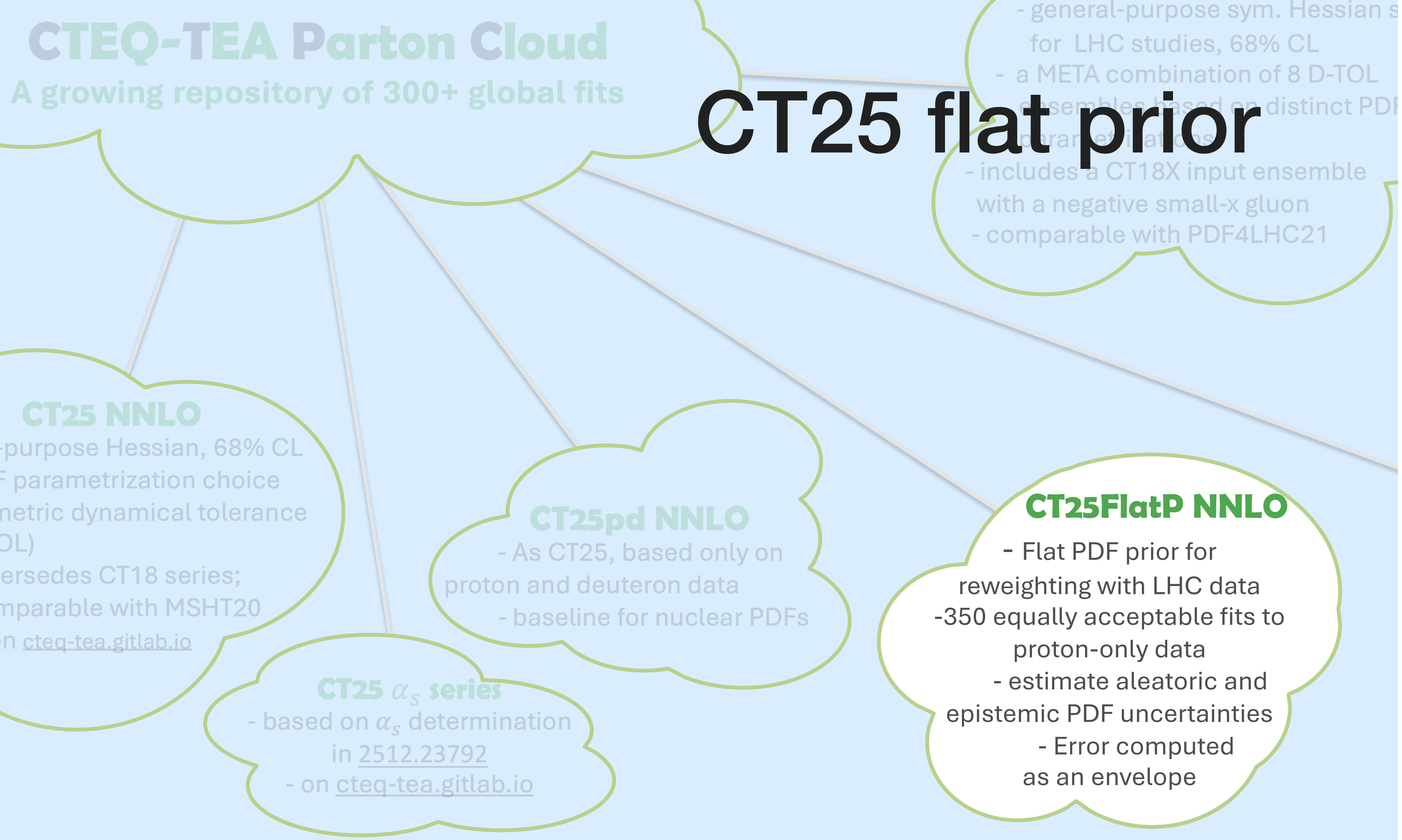
Extrapolation region — PDFs are still very underdetermined

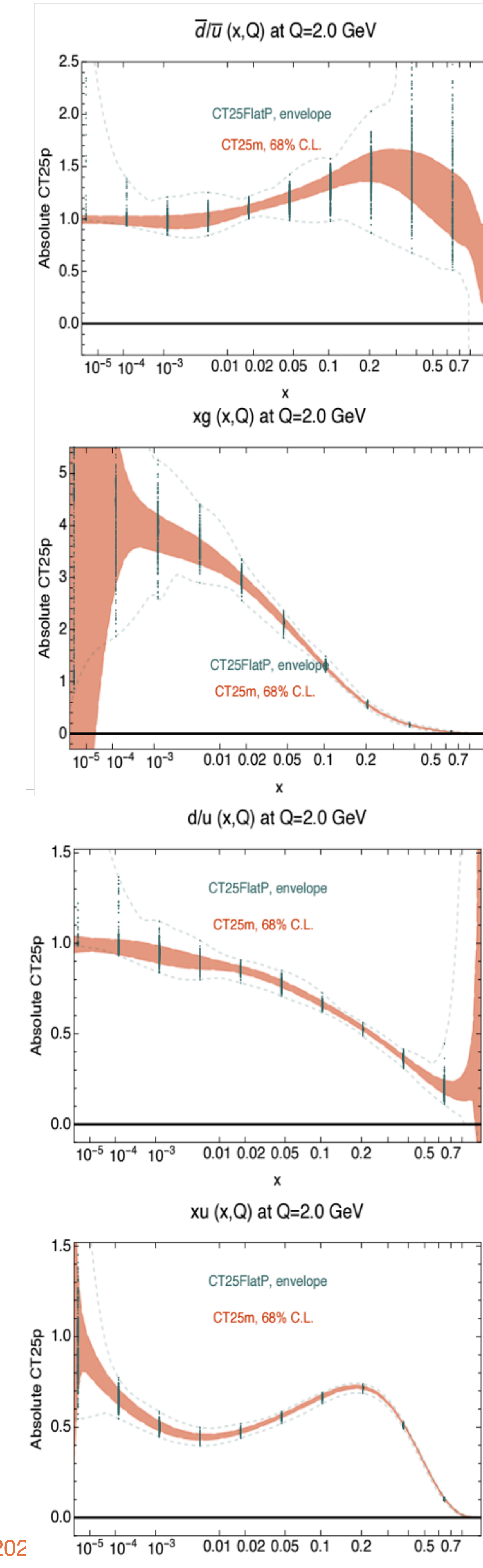
at such large (x, Q^2).





CT25 flat prior

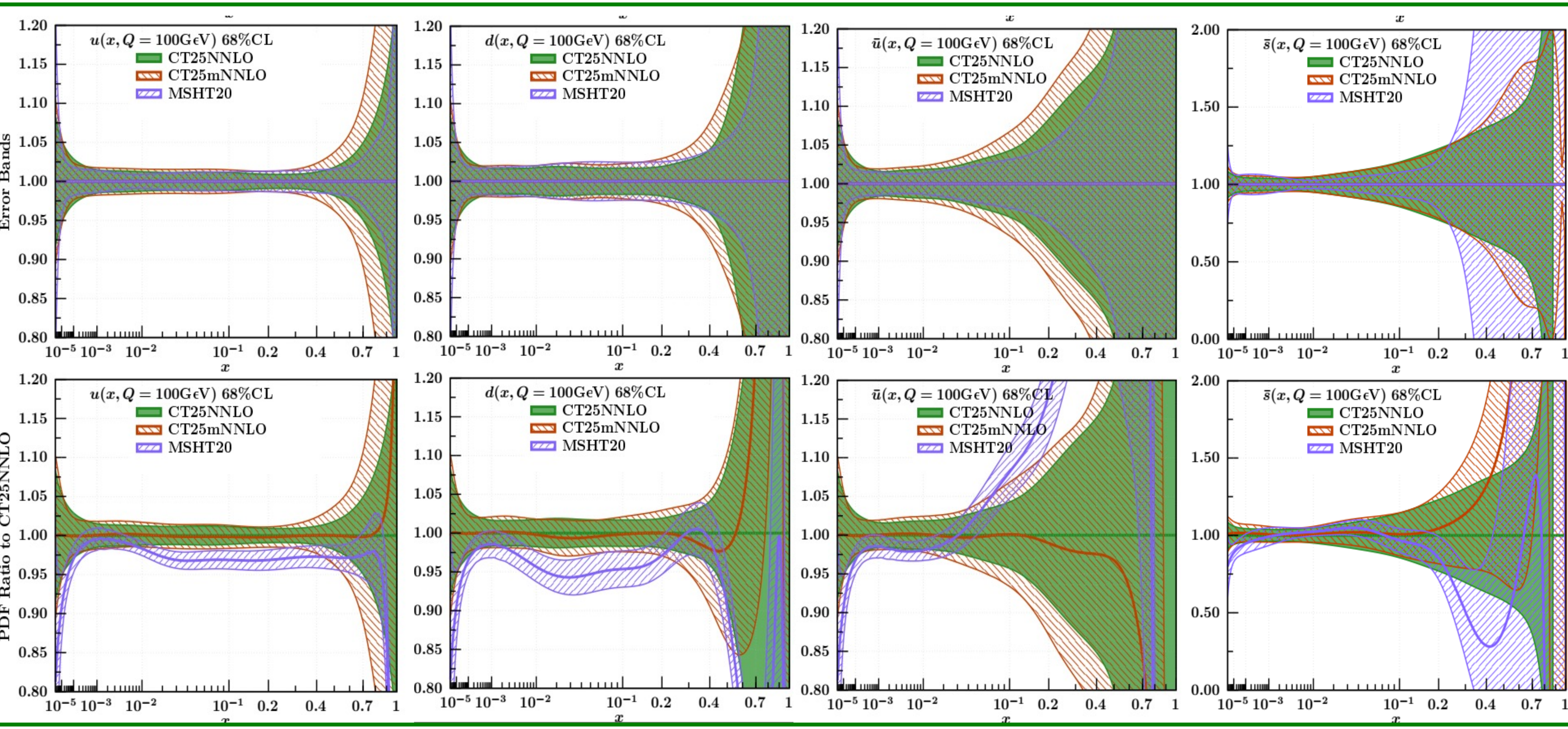




Proton-only data and ~150 *as-good*, pruned central solutions on top of CT25-like and small-x (with negative gluon at Q_0) behavior. Less flavor-separation power.

The $N_s = 350$ solutions form an envelop known as CT25FlatP
— building up a prior for those predictions that need be free of deuteron biases.

CT25m vs CT25 and MSHT20 PDFs, (anti)quark sector



CT25m vs CT25 and MSHT20 PDFs, ratios and quark singlet