Pion structure in global analyses
focus on uncertainty quantification

Aurore Courtoy

Instituto de Física
Universidad Nacional Autónoma de México (UNAM)

CONACYT
Consejo Nacional de Ciencia y Tecnología

Dirección General de Asuntos
del Personal Académico

Non-Perturbative Physics: Tools and Applications

UMSNH, Morelia
08/09/23

"Fantômas For QCD: parton distributions in a pion with Bézier parametrizations"

DIS23 proceedings [2309.00152]
Full manuscript to be released (very) soon

Fantômas team

A. Courtoy, L. Kotz, P. Nadolsky, F. Olness, D.M. Ponce-Chávez

Towards epistemic parton distributions

Mainly based in the following publications

***"Parton distributions need representative sampling"** [Phys.Rev.D 107]*

CTEQ-TEA collaboration

China: S. Dulat, J. Gao, T.-J. Hou, I. Sitiwaldi, M. Yan, and collaborators
Mexico: A. Courtoy
USA: T.J. Hobbs, M. Guzzi, J. Huston, P. Nadolsky, C. Schmidt, D. Stump, K. Xie, C.-P. Yuan

and forthcoming studies.

***"Fantômas: global analysis of the pion PDF with Bézier curves"** [upcoming]*

Fantômas team

Mexico: A. Courtoy, D.M. Ponce-Chávez
USA: L. Kotz, P. Nadolsky, F. Olness

based on [AC & Nadolsky, Phys.Rev.D 103].

PDF phenomenology — a large population analysis

Precision and accuracy in PDF global analyses are possible thanks to a wealth of data, computer power and the correct statistical methods.
Multivariate problem that involves a high-dimensional space.

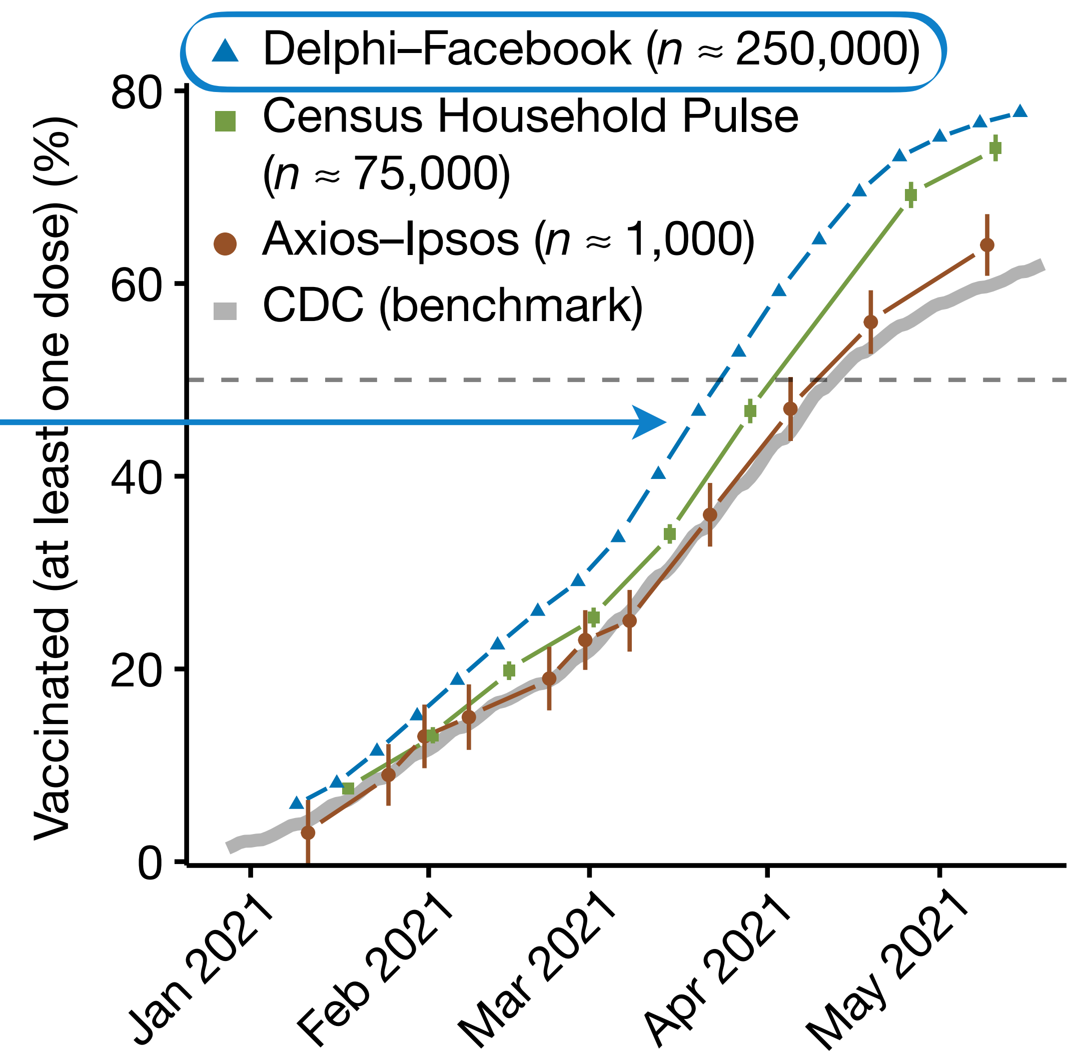
Unrepresentative big surveys significantly overestimated US vaccine uptake

Valerie C. Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng & Seth Flaxman

Nature **600**, 695–700 (2021) | Cite this article

Within some hypotheses, data uncertainty decreases as the number of data/events increases.

That's the law of large numbers.



PDF phenomenology — a large population analysis

Precision and accuracy in PDF global analyses are possible thanks to a wealth of data, computer power and the correct statistical methods.

Multivariate problem that involves a high-dimensional space.

Unrepresentative big surveys significantly overestimated US vaccine uptake

Valerie C. Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng & Seth Flaxman

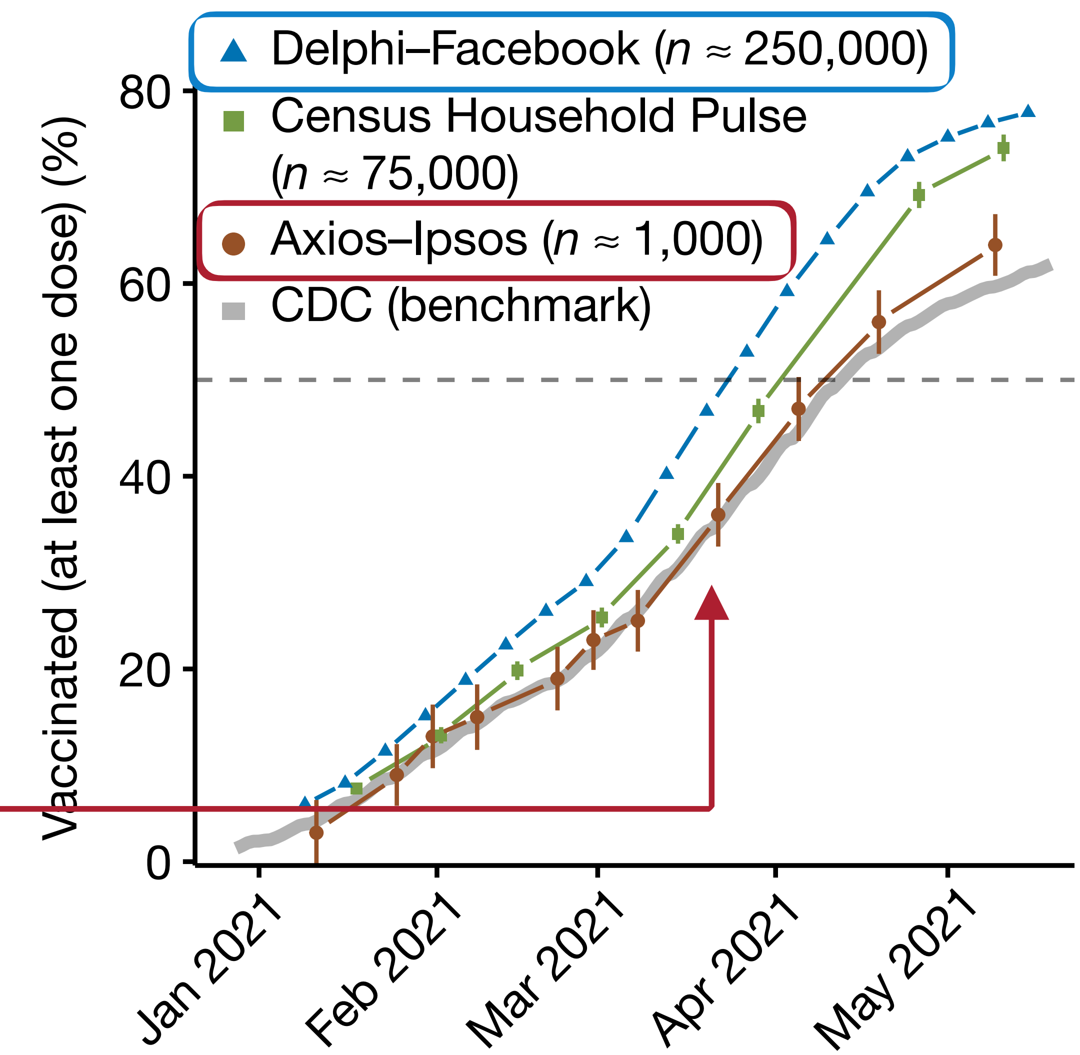
Nature **600**, 695–700 (2021) | Cite this article

Within some hypotheses, data uncertainty decreases as the number of data/events increases.

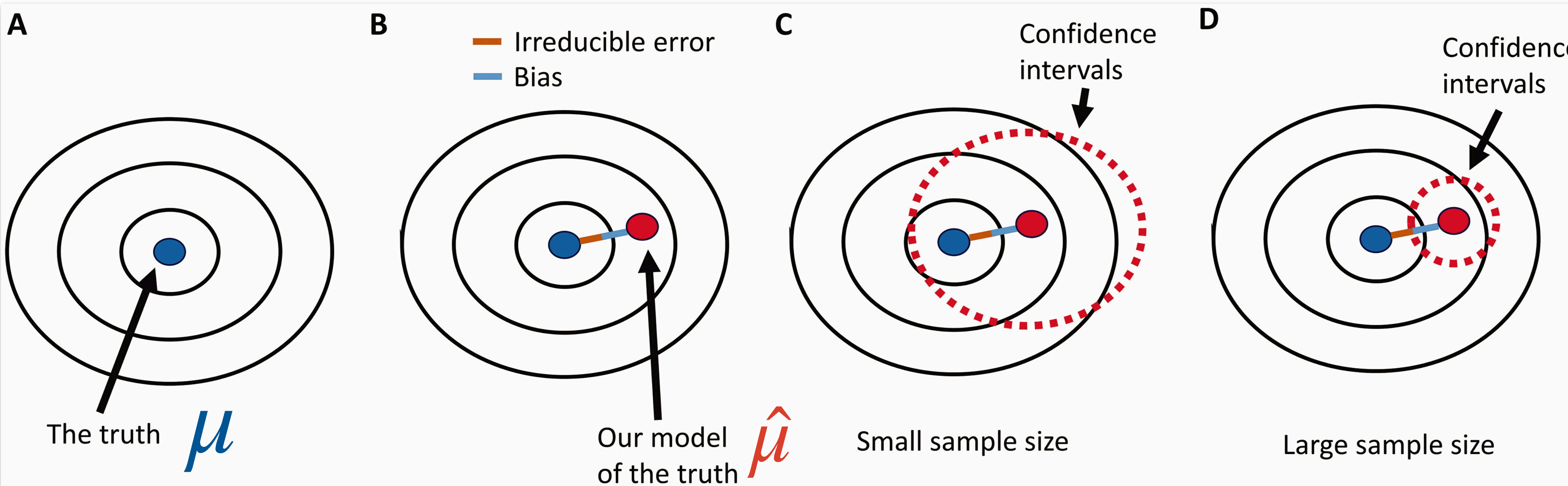
That's the law of large numbers.

Smaller sample collected through specific methods display larger uncertainties but are closer to the benchmark.

There are other factors that determine the distance to the truth: big-data paradox.

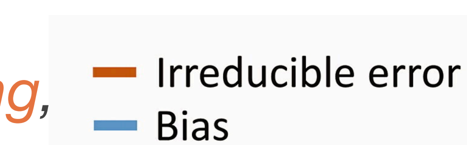


Sampling bias and big-data paradox



Pavlos Msaouel (2022)
Cancer Investigation, 40:7, 567-576

What uncertainties keep us from including *the truth*, μ?

The law of large numbers disregards the *quality of the sampling*, 

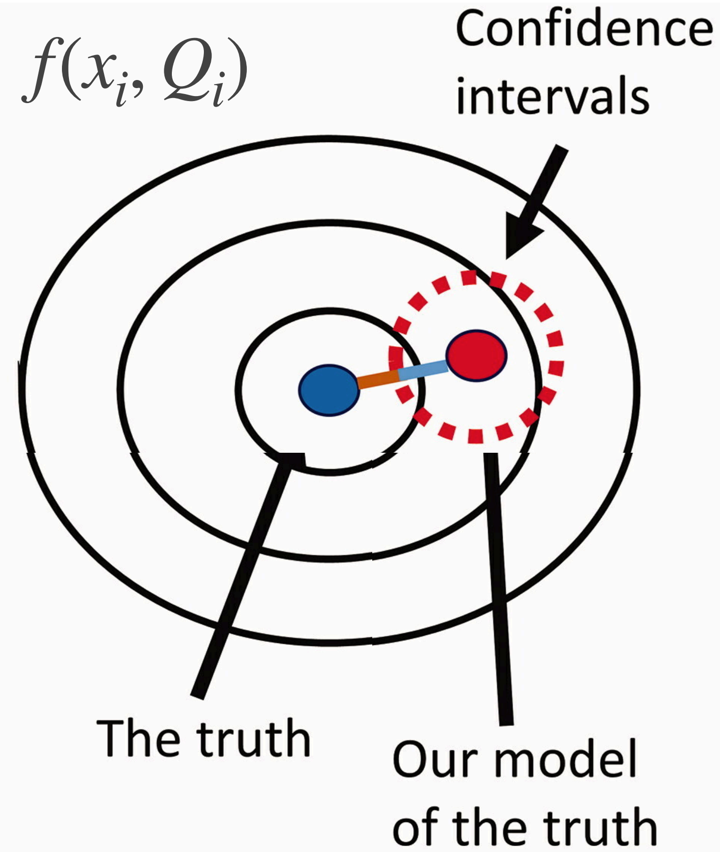
.

Xiao-Li Meng
The Annals of Applied Statistics
Vol. 12 (2018), p. 685

Physics phenomenology and accuracy

Suppose we know the true parton distribution function at given (x_i, Q_i).

We want our determination from global analysis to encompass it.

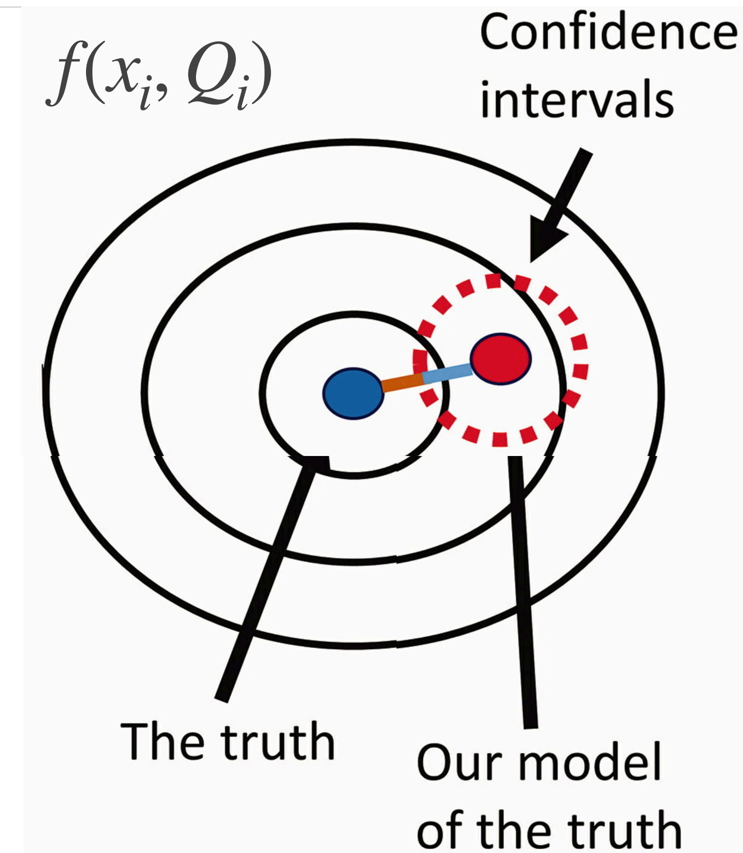


Large sample size

Physics phenomenology and accuracy

Suppose we know the true parton distribution function at given (x_i, Q_i).

We want our determination from global analysis to encompass it.



Large sample size

$$\mu - \hat{\mu} = \text{(data+sampling defect)} \times \text{(measure discrepancy)} \times \text{(inherent problem difficulty)}$$

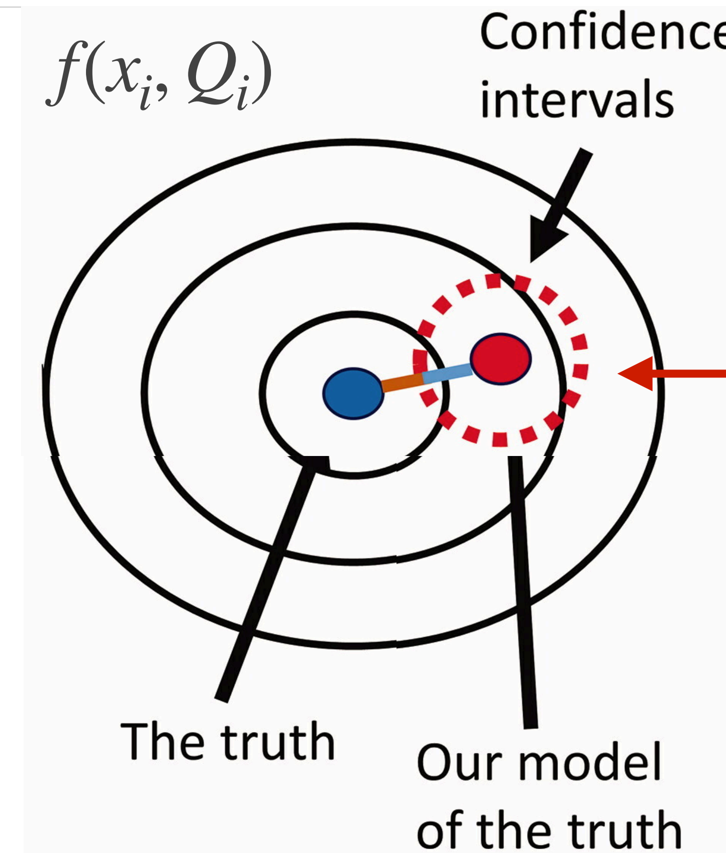
depends on the sampling algorithm

Irreducible error

Bias

$\equiv$ statistical model, quality of data,…

Physics phenomenology and accuracy



Large sample size

Suppose we know the true parton distribution function at given (x_i, Q_i).

We want our determination from global analysis to encompass it.

$$\boxed{\mu - \hat{\mu} = (\text{data+sampling defect}) \times (\text{measure discrepancy}) \times (\text{inherent problem difficulty})}$$

depends on the sampling algorithm

can tend to $\left(\sqrt{n}\right)^{-1}$ for random sampling

Irreducible error
Bias

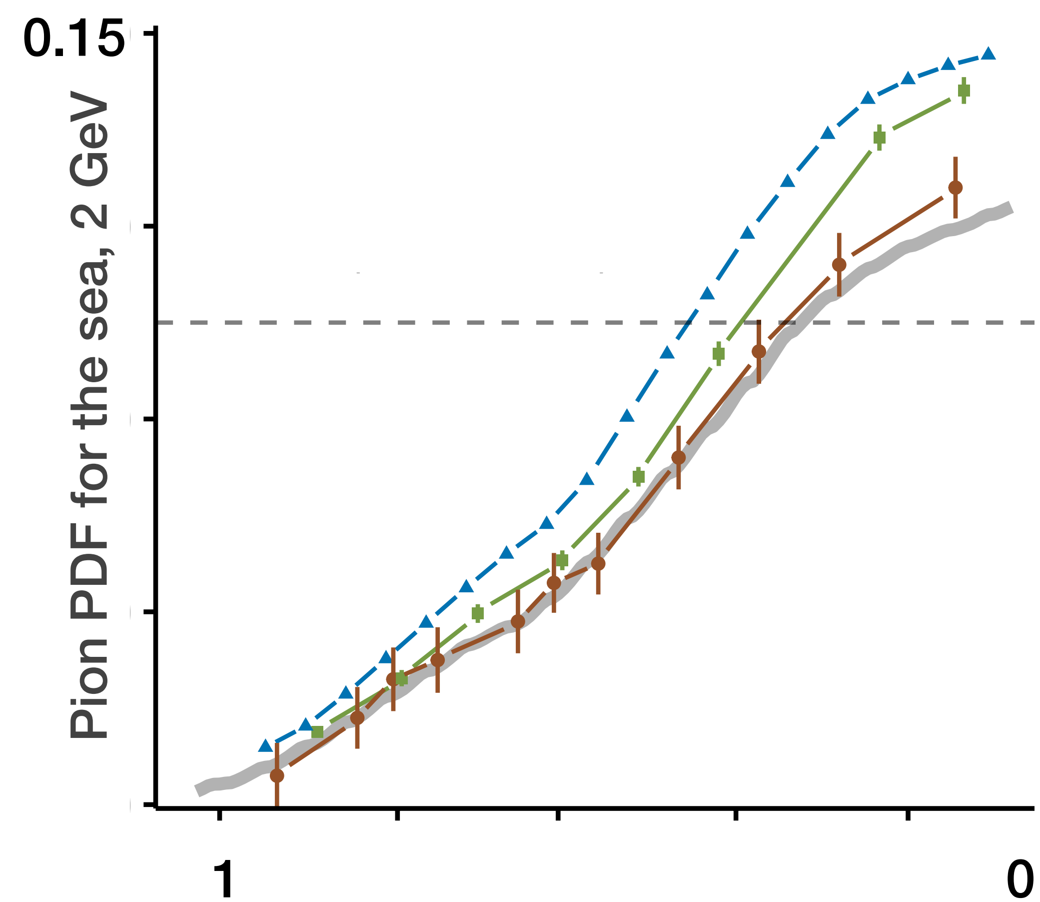
$\equiv$ statistical model, quality of data,…

Only by comparing with data may QCD phenomena be revealed

very fanciful!!

If the light-gray curve is the truth, hence its shape would reveal information on the underlying non-perturbative mechanisms.

e.g., in our fairy-tale example, that the sea quarks freeze about $x = 0.1$ at $Q = 2$ GeV or the slope at which it falls down at $x \to 1$.



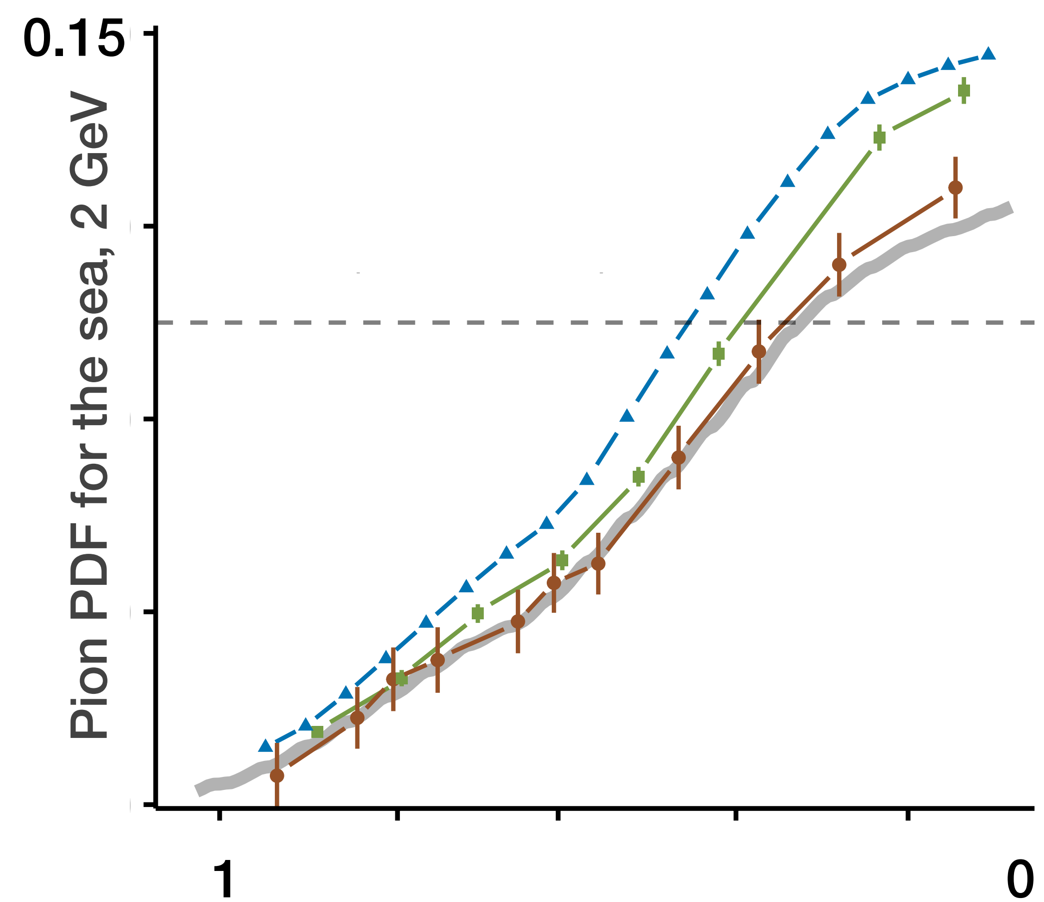
Disclaimer:
I purposedly adapted to vaccination plot for illustration. It's not a true PDF.

Only by comparing with data may QCD phenomena be revealed

very fanciful!!

If the light-gray curve is the truth, hence its shape would reveal information on the underlying non-perturbative mechanisms.

e.g., in our fairy-tale example, that the sea quarks freeze about $x = 0.1$ at $Q = 2$ GeV or the slope at which it falls down at $x \to 1$.



If sampling only focuses on the quantity of the data/replicas/parameters/…, we risk to reproduce the blue curve and its tiny uncertainty [that misses the truth].

and wrongly deduce that there are, e.g., more sea quarks at small x than there is.

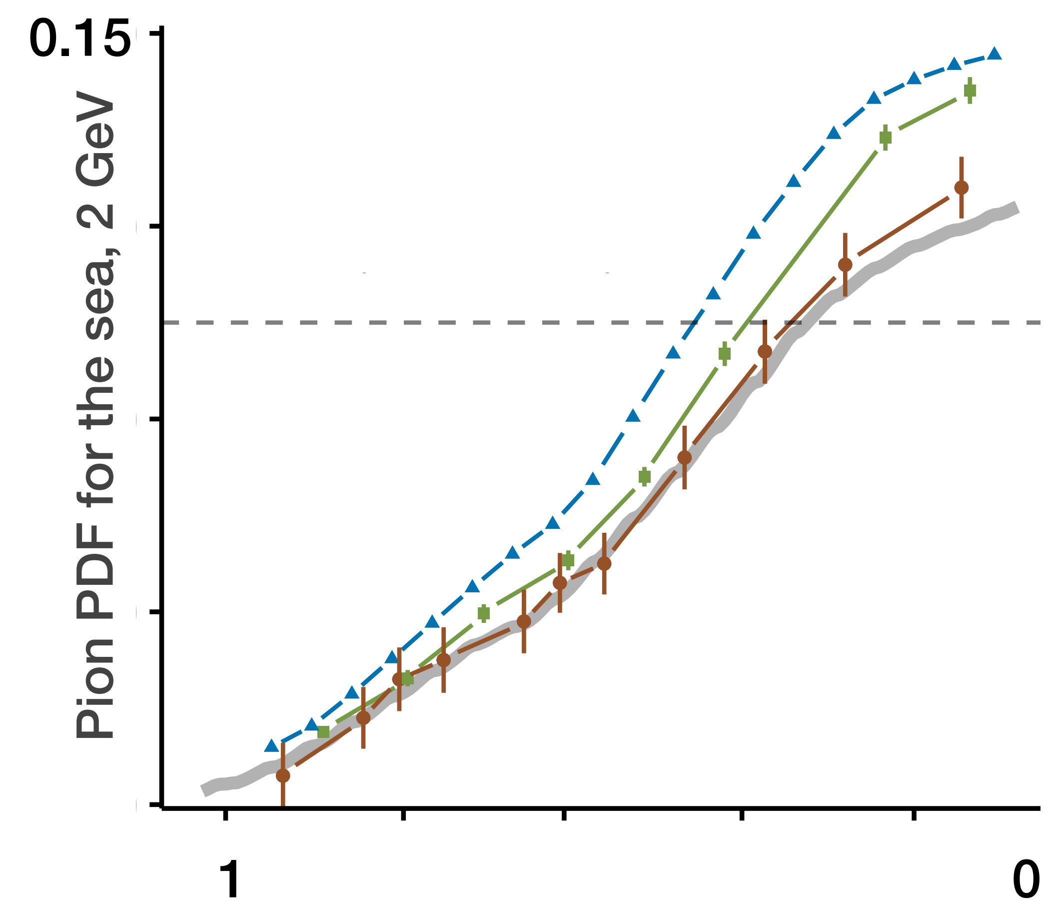
Disclaimer:
I purposedly adapted to vaccination plot for illustration. It's not a true PDF.

Only by comparing with data may QCD phenomena be revealed

very fanciful!!

If the light-gray curve is the truth, hence its shape would reveal information on the underlying non-perturbative mechanisms.

e.g., in our fairy-tale example, that the sea quarks freeze about $x = 0.1$ at $Q = 2$ GeV or the slope at which it falls down at $x \to 1$.



If sampling only focuses on the quantity of the data/replicas/parameters/…, we risk to reproduce the blue curve and its tiny uncertainty [that misses the truth].

and wrongly deduce that there are, e.g., more sea quarks at small x than there is.

If sampling is inclusive of various factors [optimization of the space exploration], we might reproduce the maroon curve together with a larger uncertainty [and include the truth].

Disclaimer:
I purposedly adapted to vaccination plot for illustration. It's not a true PDF.

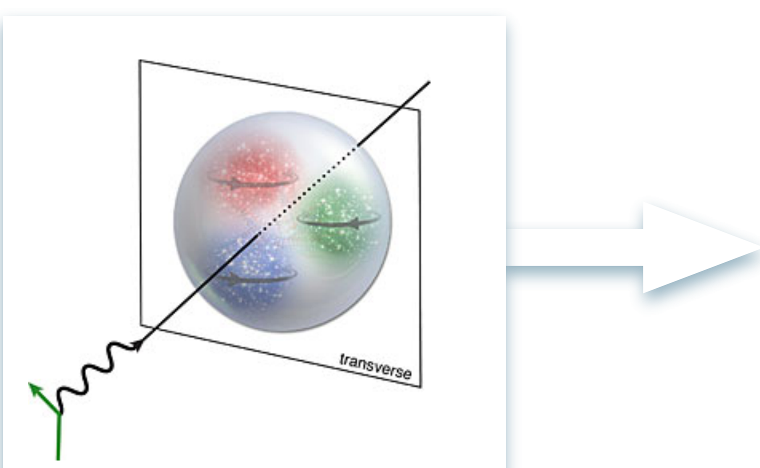
PDFs in QCD phenomenology

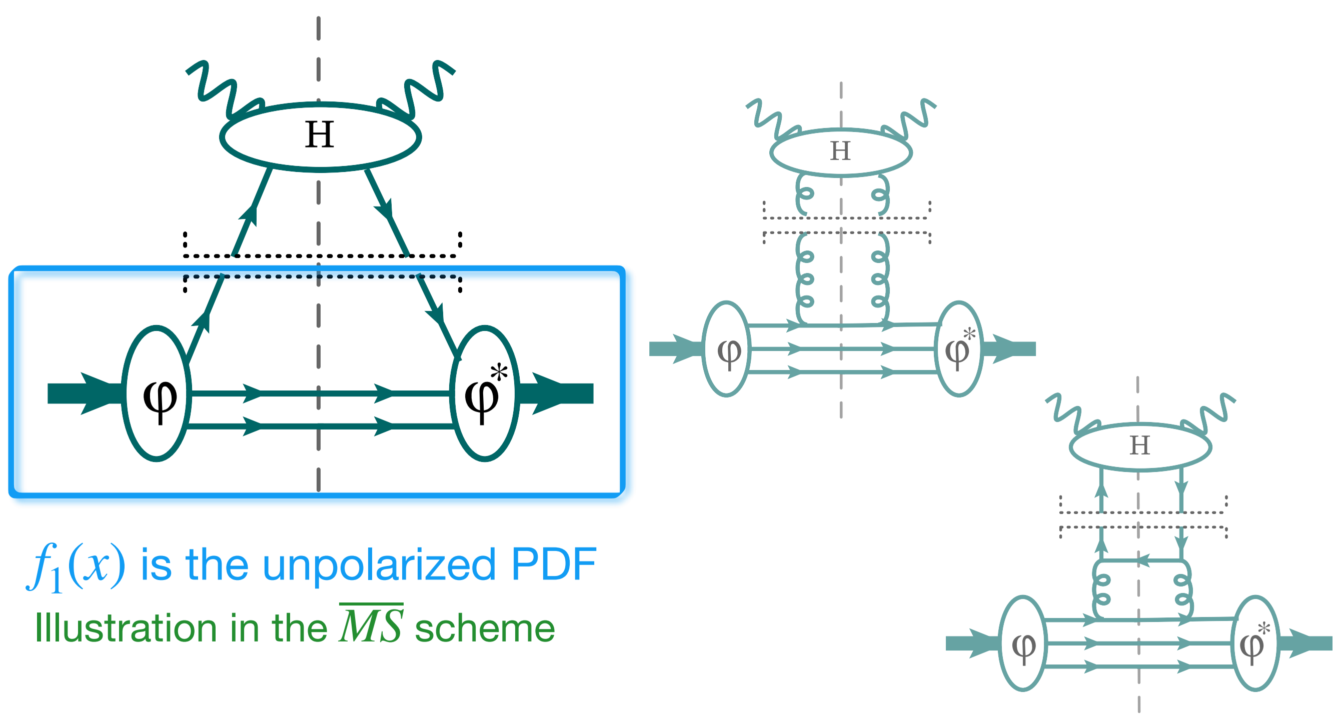
In the realm of QCD…

The two regimes of QCD demand that non-perturbative objects are accessed through *factorization theorems*,

i.e., theory is expressed through a convolution of hard and soft part to which corrections are added.

⇒ the data is not reproduced by the PDF only





$f_1(x)$ is the unpolarized PDF

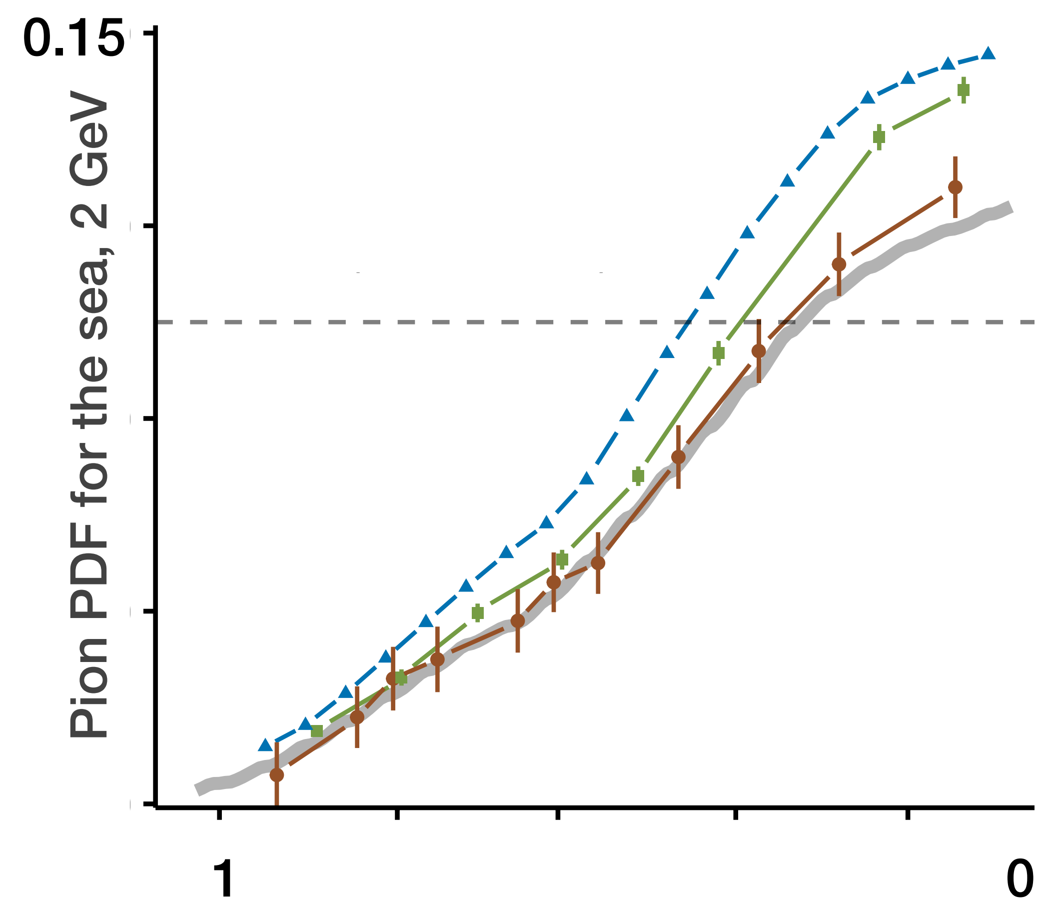
Illustration in the $\overline{MS}$ scheme

Only by comparing with data may QCD phenomena be revealed

What are the conventions on “QCD phenoma”? [first principles may need adaptation to the present understanding of perturbative QCD].

[AC & Nadolsky, PRD103]
[Candido et al, *JHEP* 11& 2308.00025]
[Collins et al, PRD105]

E.g., what is the smoking-gun sign for role of chiral symmetry in the emergence of hadronic mass? On how many and which parameters does that smoking gun depend?

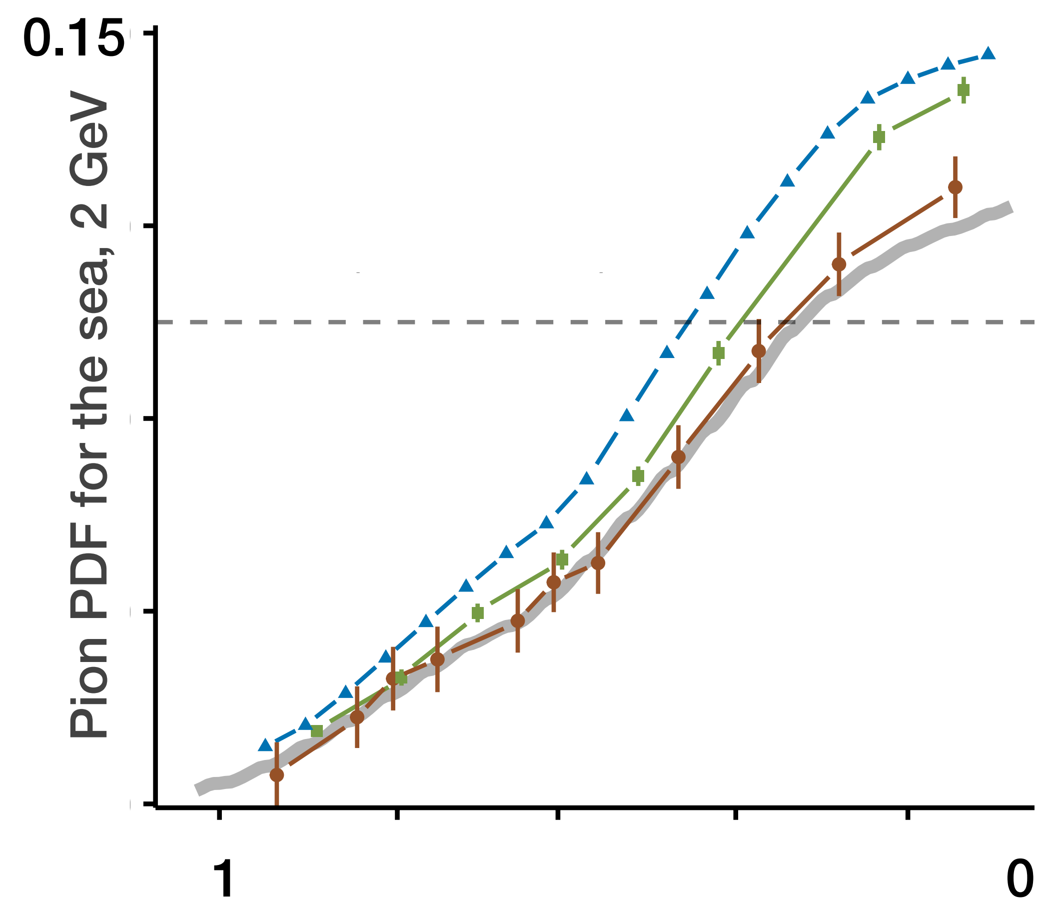


Only by comparing with data may QCD phenomena be revealed

What are the conventions on “QCD phenoma”?
[first principles may need adaptation to the present understanding of perturbative QCD].

[AC & Nadolsky, PRD103]
[Candido et al, *JHEP* 11& 2308.00025]
[Collins et al, PRD105]

E.g., what is the smoking-gun sign for role of chiral symmetry in the emergence of hadronic mass? On how many and which parameters does that smoking gun depend?



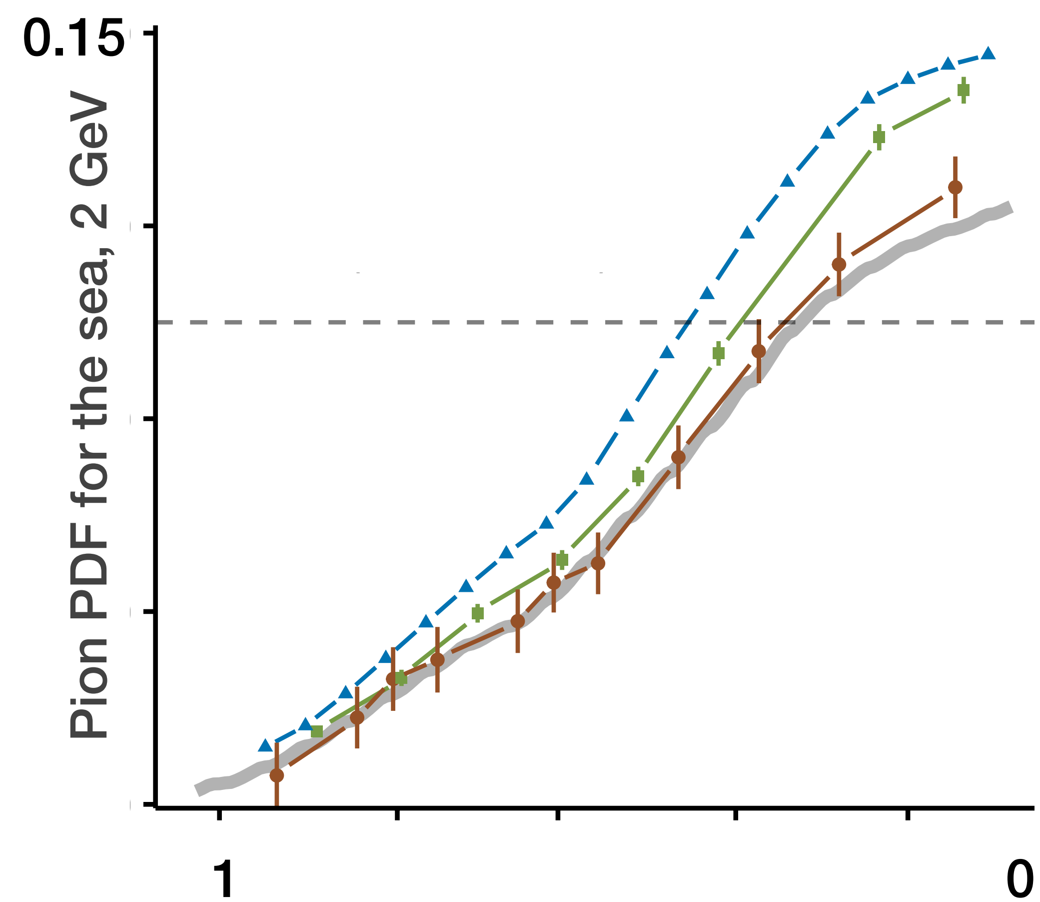
Poor sampling can sometimes be due to over-constrained space where solutions are deemed acceptables — i.e., through priors or penalties.

For the Hessian-methodology based global analyses, a functional form is required. A parametrization is a prior-like penalty.

Only by comparing with data may QCD phenomena be revealed

What are the conventions on “QCD phenoma”? [first principles may need adaptation to the present understanding of perturbative QCD].

[AC & Nadolsky, PRD103]
[Candido et al, *JHEP* 11& 2308.00025]
[Collins et al, PRD105]



E.g., what is the smoking-gun sign for role of chiral symmetry in the emergence of hadronic mass? On how many and which parameters does that smoking gun depend?

Poor sampling can sometimes be due to over-constrained space where solutions are deemed acceptables — i.e., through priors or penalties.

For the Hessian-methodology based global analyses, a functional form is required. A parametrization is a prior-like penalty.

To purposely rectify the sampling over parametrization, we have designed `metamorph`.

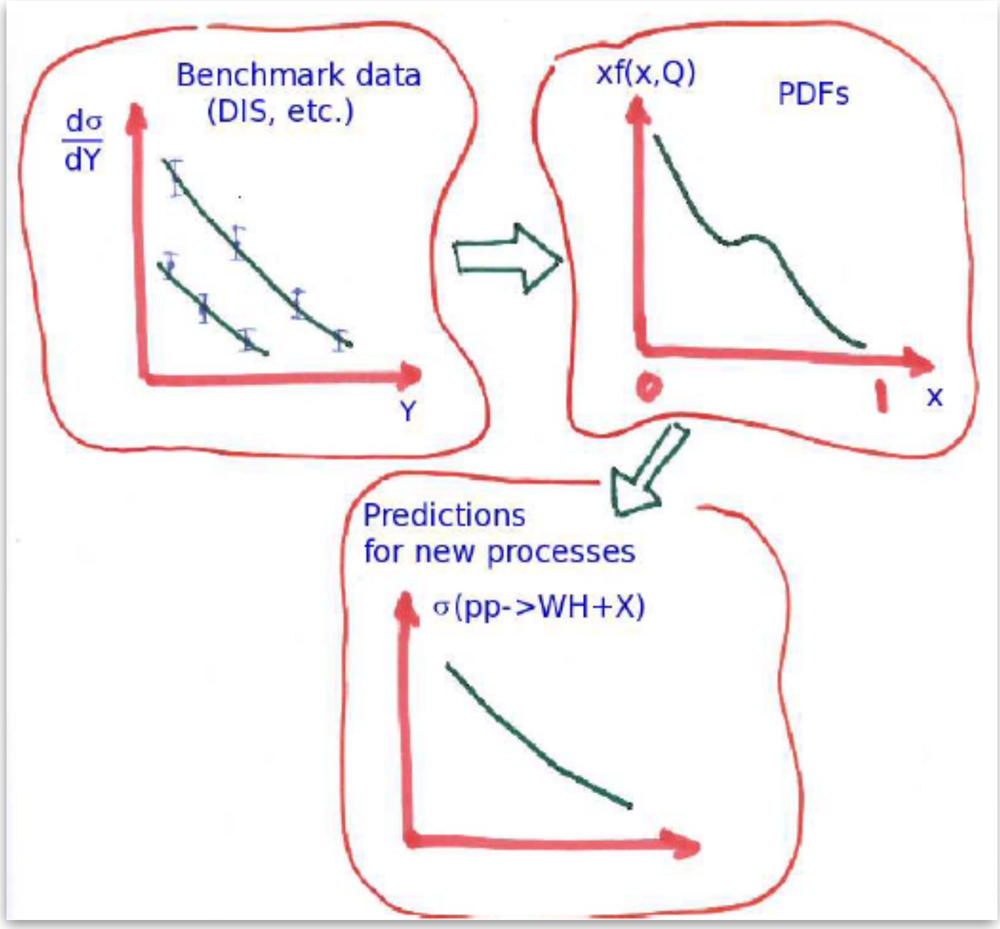
The shape of parton distributions

Low-energy QCD dynamics, encapsulated in PDFs, are learned from experimental data.

Shape in x extracted from data that are sensitive to specific PDF flavors, etc.

I. hints of behavior of partons at low scales

II. predictions for other (new) processes



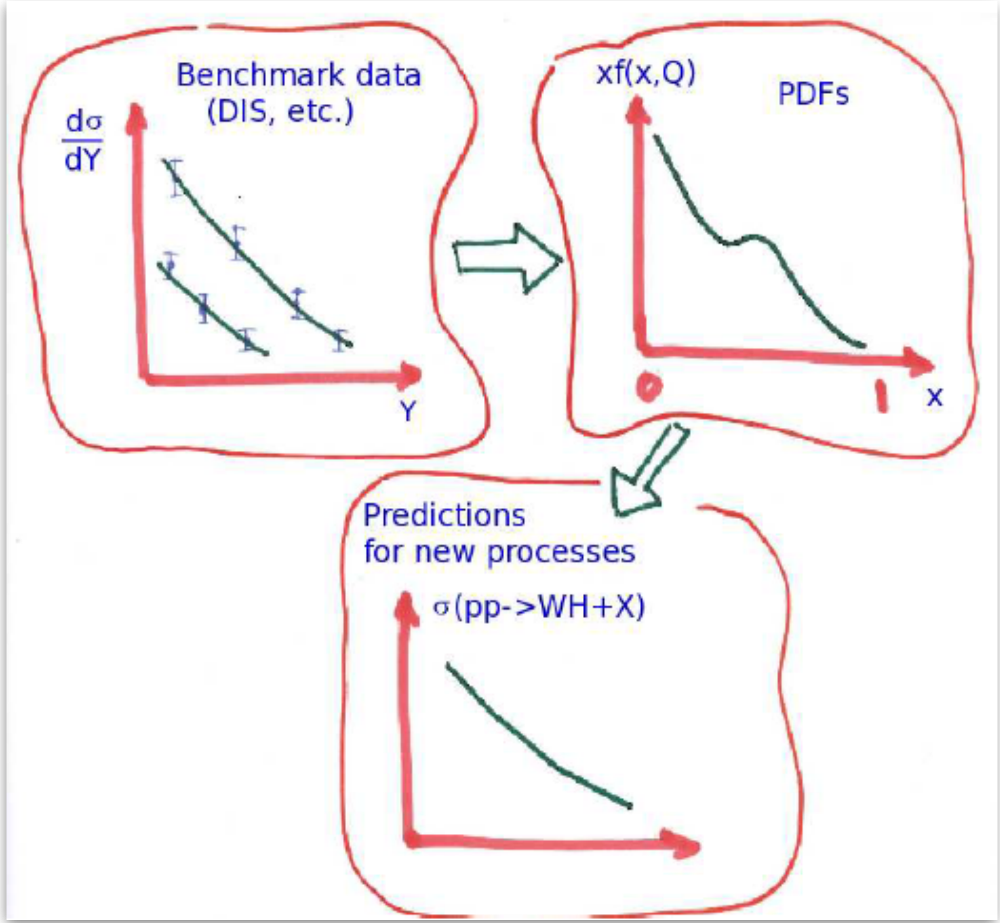
The shape of parton distributions

Low-energy QCD dynamics, encapsulated in PDFs, are learned from experimental data.

Shape in x extracted from data that are sensitive to specific PDF flavors, etc.

I. hints of behavior of partons at low scales

II. predictions for other (new) processes



Classes of *first principle* constraints for *x*-dependence

- positivity of cross sections
- support in $x \in [0,1]$
- end-point: $f(x=1)=0$
- sum rules: $< x >_n = \int_0^1 dx\, x^{n-1} f(x)$

Model evaluation of *x*-dependence (in parallel to data learning)

- use QFT description of $f(x)$ together with model description of hadron wave function (non trivial to define)
- ensure symmetries are fulfilled

The shape of parton distributions

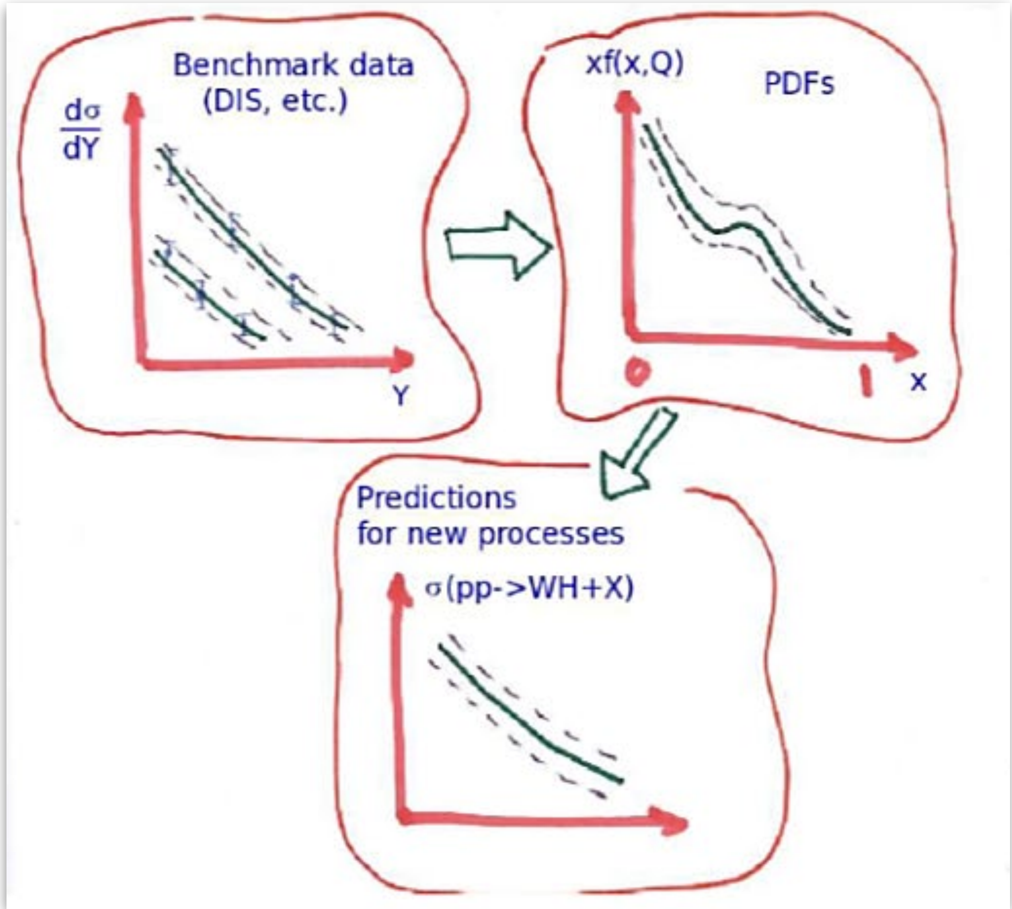
Low-energy QCD dynamics, encapsulated in PDFs, are learned from experimental data.

Uncertainty propagates from data and methodology to the PDF determination

I. assessment of uncertainty magnitude is key

II. advanced statistical problem

III. evolving topic in the era of AI/ML



The shape of parton distributions

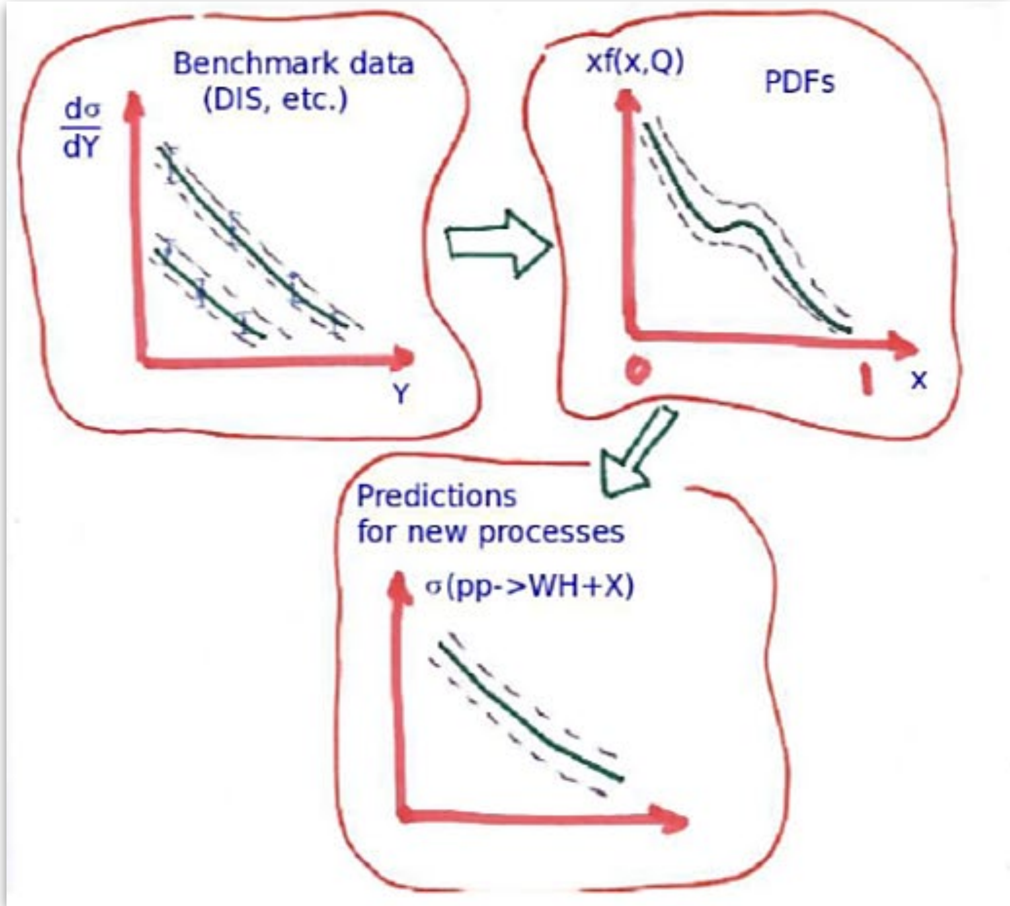
Low-energy QCD dynamics, encapsulated in PDFs, are learned from experimental data.

Uncertainty propagates from data and methodology to the PDF determination

I. assessment of uncertainty magnitude is key

II. advanced statistical problem

III. evolving topic in the era of AI/ML



Epistemic vs. aleatory uncertainties

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

Statistical uncertainty propagated from experiments — irreducible

Hypothesis testing and parton distributions

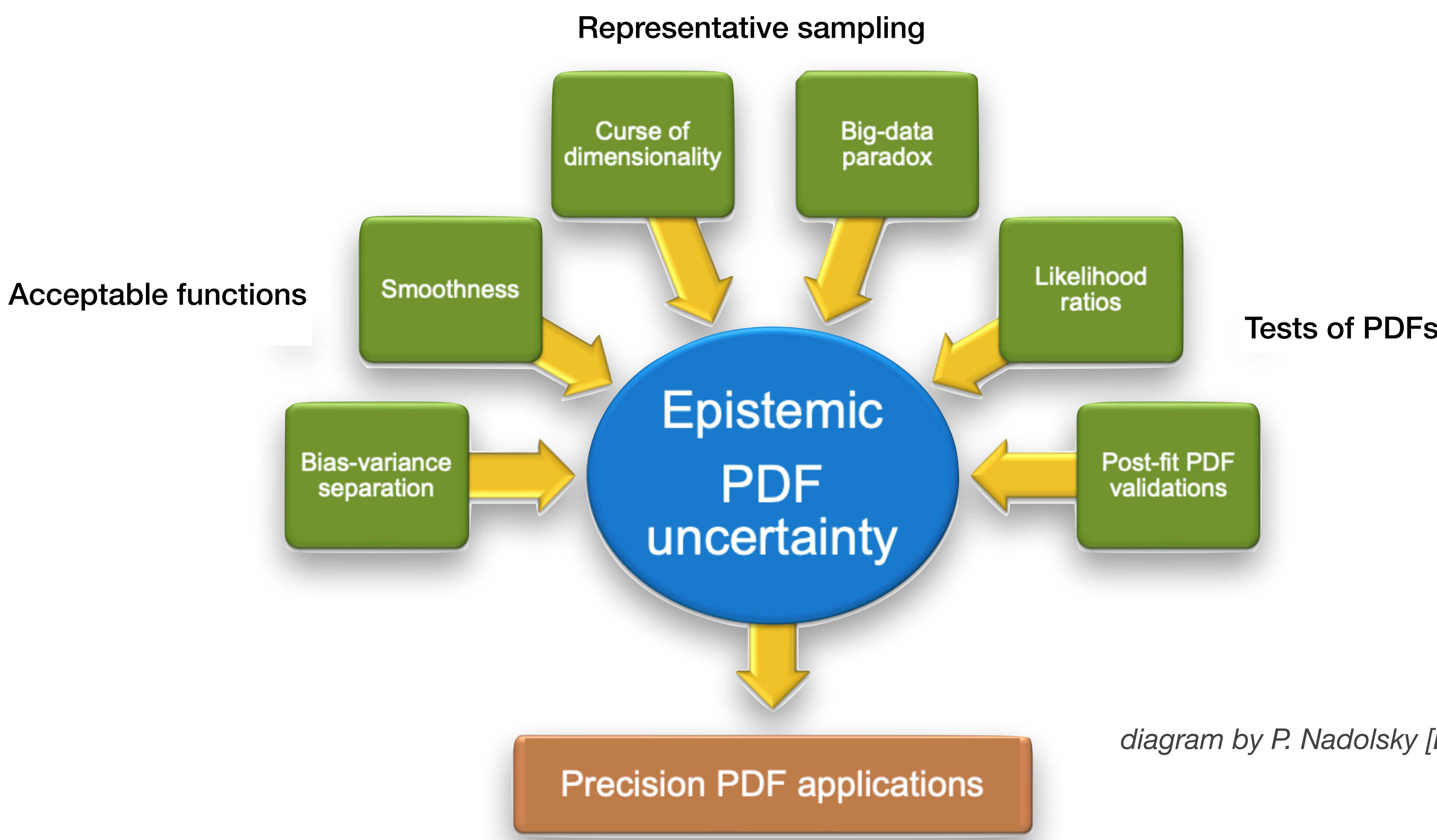
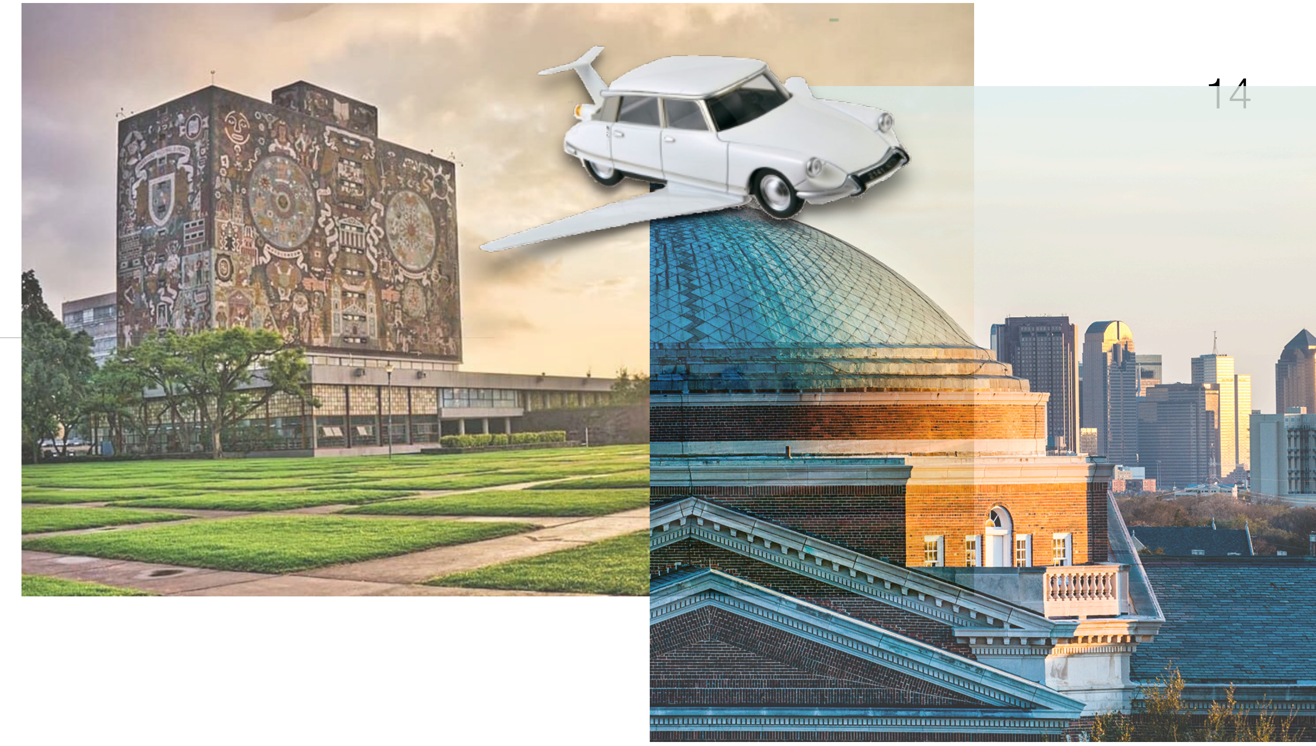


diagram by P. Nadolsky [DIS2023]

Fantômas4QCD



Fantômas4QCD: systematize the role of functional form in global analyses, alternative to neural networks.

CTEQ-TEA and xFitter members + students
Collaboration between Southern Methodist University (SMU) and Institute of Physics at UNAM.

Supported by CONACyT (Mexico) and IANN-QCD.

CONACYT
Consejo Nacional de Ciencia y Tecnología

Inter-American Network of Networks of QCD challenges

Polynomial mimicry

Testing specific x-shapes:

Polynomial mimicry=Mathematical equivalence of polynomials of different orders.

Bézier curves give an example of mathematical equivalence of polynomials of different orders

The interpolation through Bézier curves is unique if the polynomial degree= (# points-1), there's a closed-form solution to the problem,

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

with the Bernstein pol.

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

Polynomial mimicry

Testing specific x-shapes:

Polynomial mimicry=Mathematical equivalence of polynomials of different orders.

Bézier curves give an example of mathematical equivalence of polynomials of different orders

The interpolation through Bézier curves is unique if the polynomial degree= (# points-1), there's a closed-form solution to the problem,

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

with the Bernstein pol.

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

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

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

- $\underline{T}$ is the vector of x^l
- $\underline{\underline{M}}$ is the matrix of binomial coefficients
- $\underline{C}$ is the vector of Bézier coefficient, c_l, to be determined

Polynomial mimicry

We can evaluate the Bézier curve at chosen **control points**, to get a vector of $\mathscr{B} \rightarrow \underline{P}$

- $\underline{\underline{T}}$ is now a matrix of x^l expressed at the control points.

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

Such that the coefficients can be expressed in terms of known matrices

$$\boxed{\underline{C} = \underline{\underline{M}}^{-1} \cdot \underline{\underline{T}}^{-1} \cdot \underline{P}}$$

Polynomial mimicry

We can evaluate the Bézier curve at chosen **control points**, to get a vector of $\mathscr{B} \to \underline{P}$

- $\underline{\underline{T}}$ is now a matrix of x^l expressed at the control points.

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

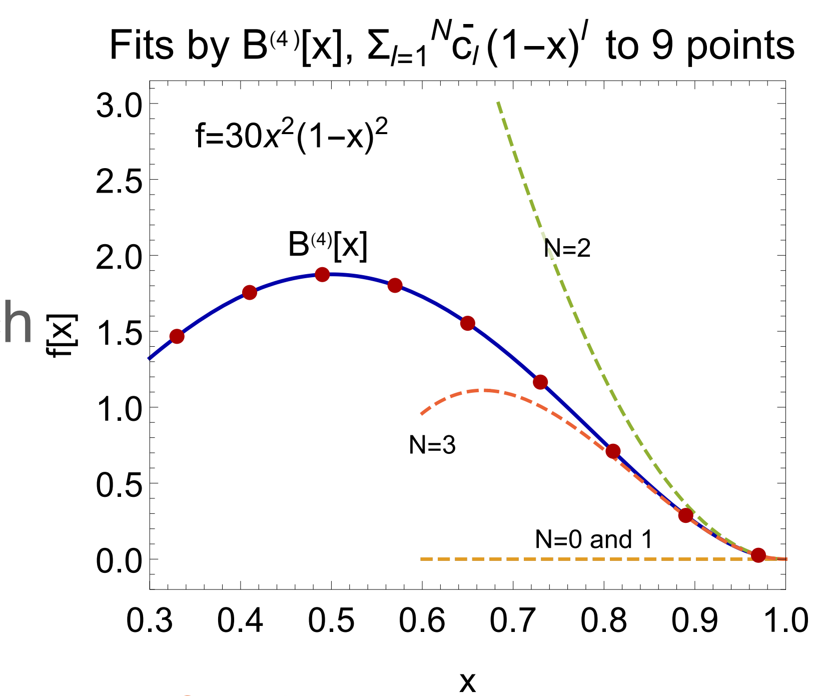
Such that the coefficients can be expressed in terms of known matrices

$$\boxed{\underline{C} = \underline{\underline{M}}^{-1} \cdot \underline{\underline{T}}^{-1} \cdot \underline{P}}$$

To test a $(1-x)$ behavior, we expand the interpolation through Bézier curves about $x = 1$:

$$u_\pi(x \to 1) = \sum_{i=0}^{n} \bar{c}_i \, (1-x)^i$$

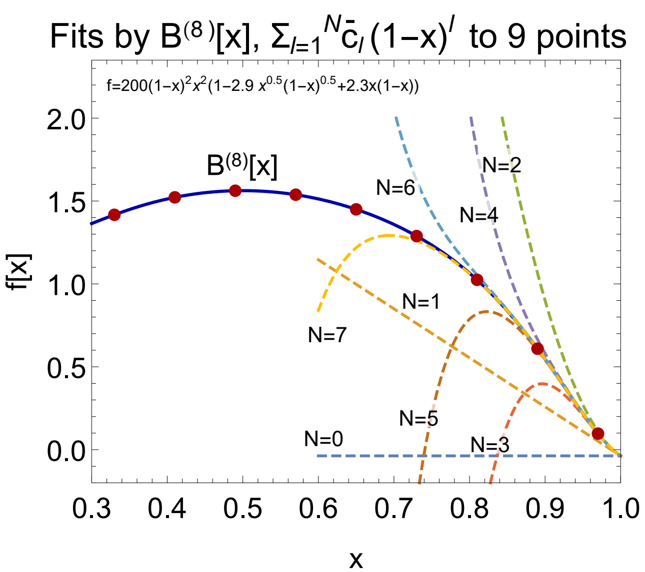
The red points represent the control points, the number of which is related to the degree of the polynomial.

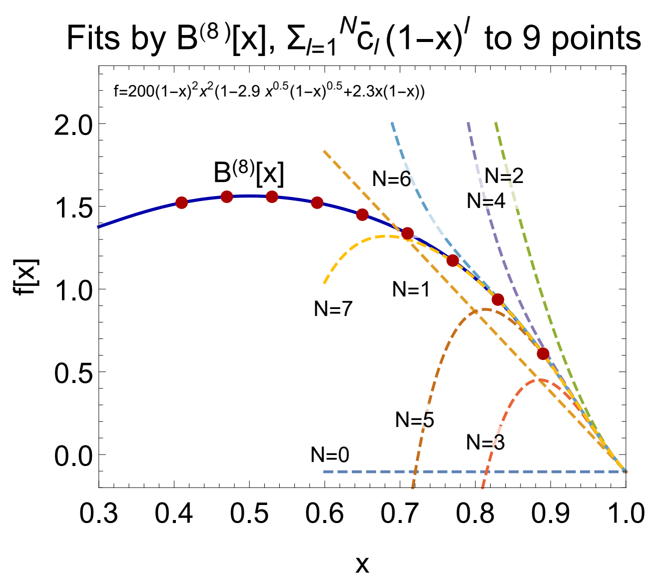


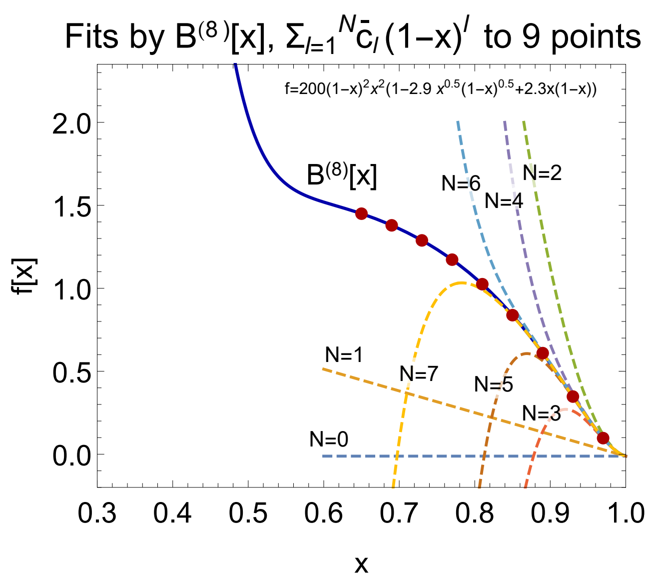
Bézier-curve methodology for global analyses

Reconstruction of a parametrization $f=200(1-x)^2x^2(1-2.9\,x^{0.5}(1-x)^{0.5}+2.3x(1-x))$

⇨ The lowest powers of the expansion cannot be meaningfully reconstructed.



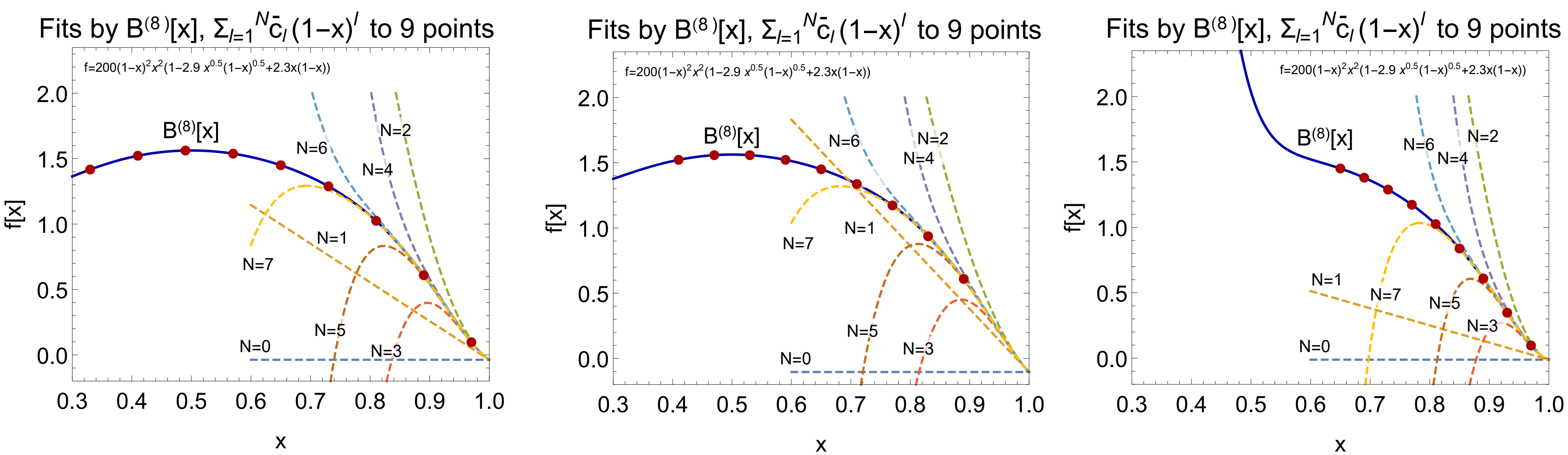




Bézier-curve methodology for global analyses

<u>Reconstruction of a parametrization</u> $f=200(1-x)^2x^2(1-2.9\,x^{0.5}(1-x)^{0.5}+2.3x(1-x))$

⇨ The lowest powers of the expansion cannot be meaningfully reconstructed.



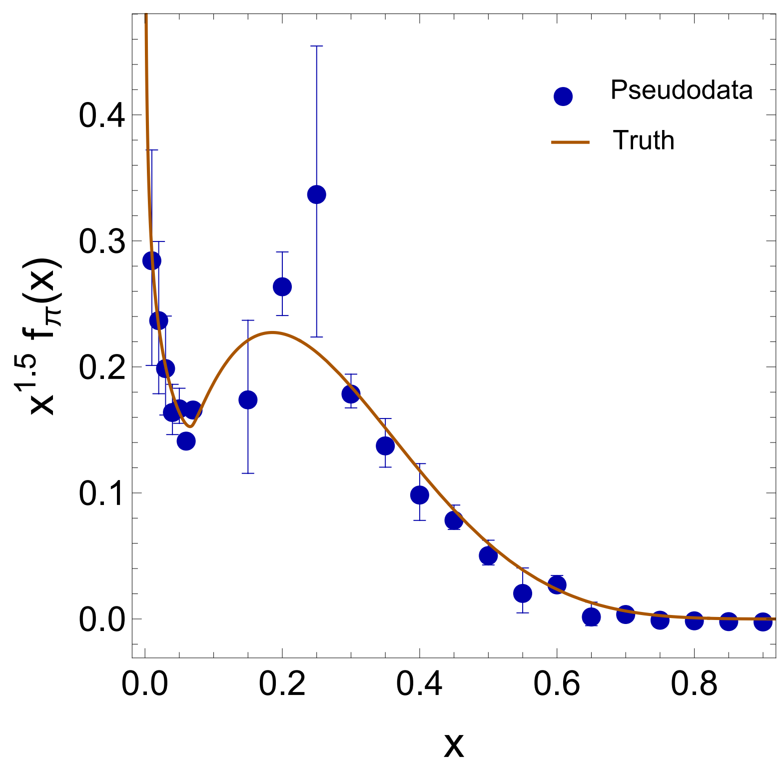
⇨ The reconstructed function depends on the position and number of control points.

This property can be exploited in favor of global analyses: by varying settings of the Bézier curves, we generate a variety of curves, beyond reconstruction.

Bézier-curve methodology for global analyses — the pion

Fantômas4QCD program

- **From interpolation to minimization over parameters through $\mathscr{B}$**
- Exploit polynomial mimicry to systematically improve and flexibilize parametrization of PDFs.



Classical fit:

$$x\, q(x, Q_0^2) = A_q'\, x^{B_q} (1-x)^{C_q} \times \left(1 + \mathcal{B}^{(N_m)}(x, Q_0^2)\right)$$

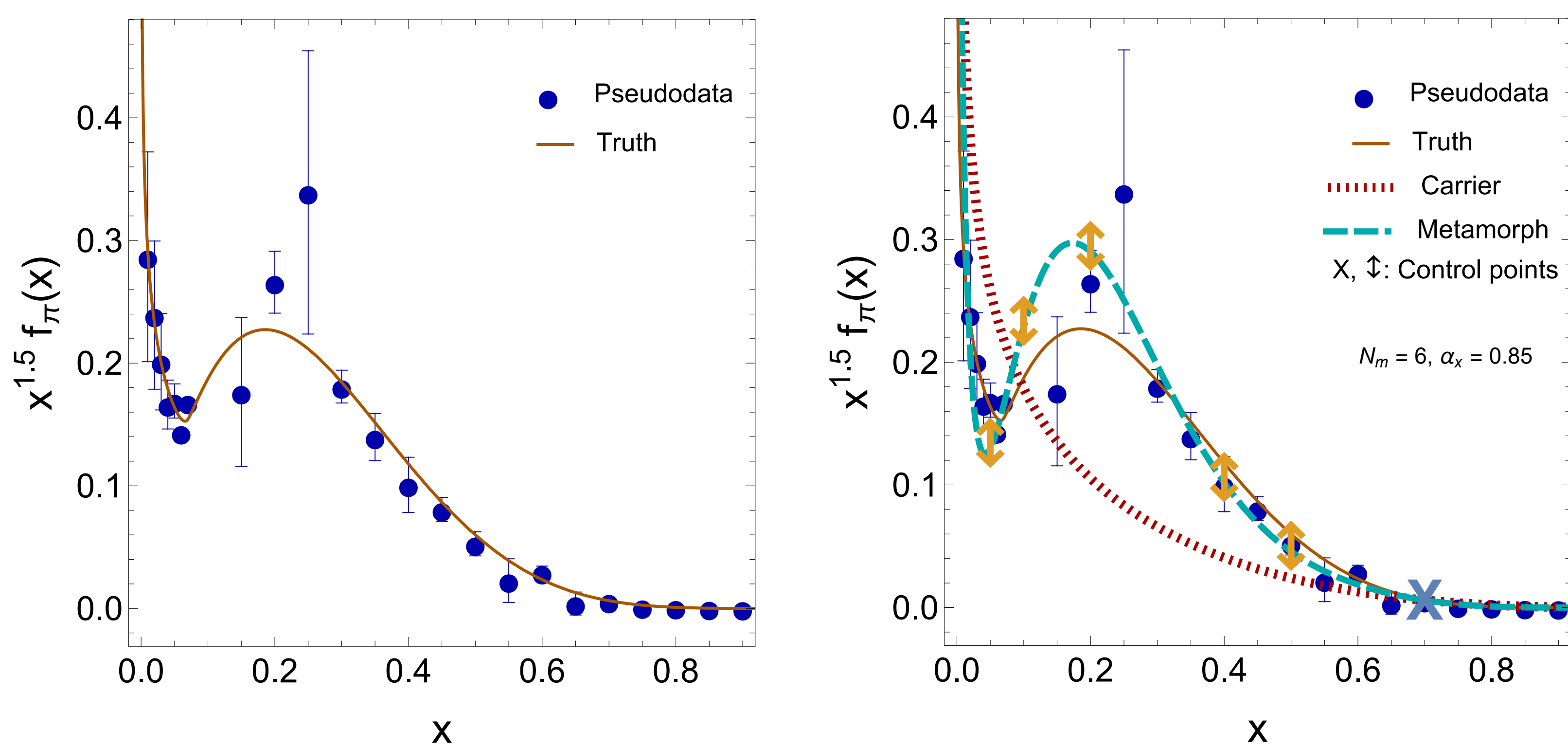
`metamorph` fit:

$$x\, q^{\text{Fanto}}(x, Q_0^2) = a_q\, x^{B_q + \delta B_q} (1-x)^{C_q + \delta C_q} \times \left(1 + \mathcal{B}^{(N_m)}(x, Q_0^2; \{\delta D_q, \delta E_q, \cdots\})\right)$$

We parametrize the Bézier coefficients as the shifts of the position of the **control points:**

$$\begin{aligned} P_i = \mathcal{B}(x_i) \to P_i' &= \mathcal{B}(x_i) + \delta\mathcal{B}(x_i) \\ \to \underline{P}' &= (\mathcal{B}_0(x_1) + \delta D, \mathcal{B}_0(x_2) + \delta E, \cdots) \end{aligned}$$

Bézier-curve methodology for global analyses — the pion



Shift of the control points $(\delta D_q, \ldots)$ replace free parameters

N_m= degree of polynomial can vary

δB_q & δC_q allow the carrier to vary

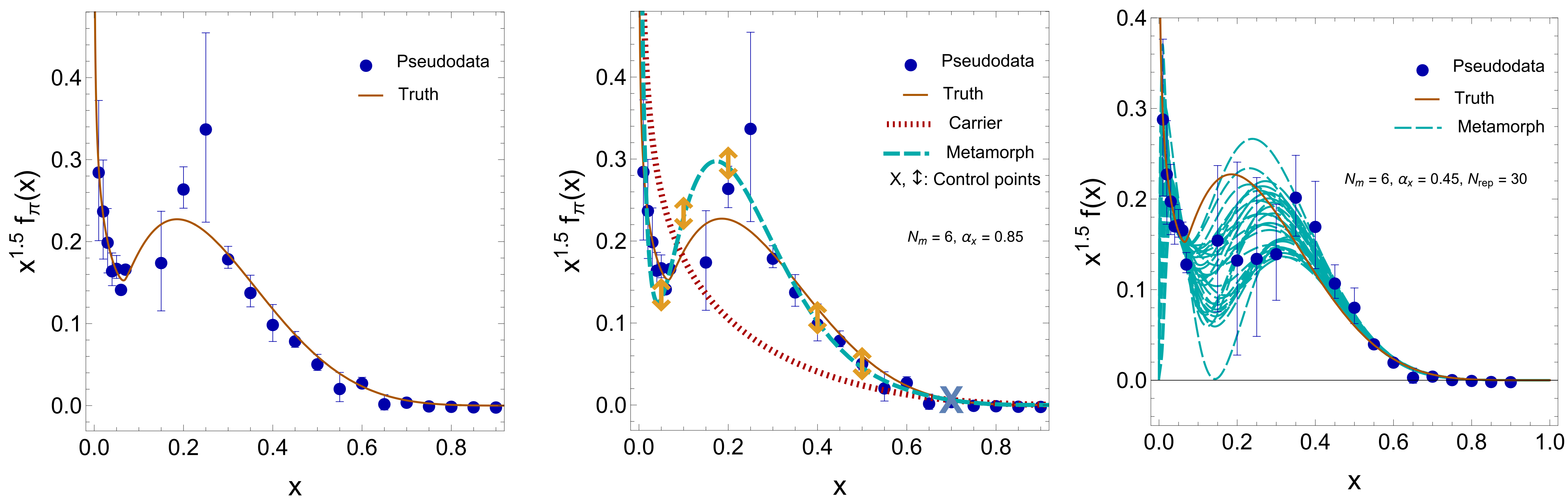
Classical fit: $$x\,q(x,Q_0^2) = A'_q\,x^{B_q}(1-x)^{C_q} \times \left(1+\mathcal{B}^{(N_m)}(x,Q_0^2)\right)$$

`metamorph` fit: $$x\,q^{\text{Fanto}}(x,Q_0^2) = a_q\,x^{B_q+\delta B_q}(1-x)^{C_q+\delta C_q} \times \left(1+\mathcal{B}^{(N_m)}(x,Q_0^2;\{\delta D_q,\delta E_q,\cdots\})\right)$$

Bézier-curve methodology for global analyses — the pion

Exploit polynomial mimicry to systematically improve and flexibilize parametrization of PDFs.

⇨ Fantômas4QCD program



Classical fit: $x\, q(x, Q_0^2) = A'_q\, x^{B_q} (1-x)^{C_q} \times \left(1 + \mathcal{B}^{(N_m)}(x, Q_0^2)\right)$

`metamorph` fit: $x\, q^{\text{Fanto}}(x, Q_0^2) = a_q x^{B_q+\delta B_q} (1-x)^{C_q+\delta C_q} \times \left(1 + \mathcal{B}^{(N_m)}(x, Q_0^2; \{\delta D_q, \delta E_q, \cdots\})\right)$

Various fits created on the fly

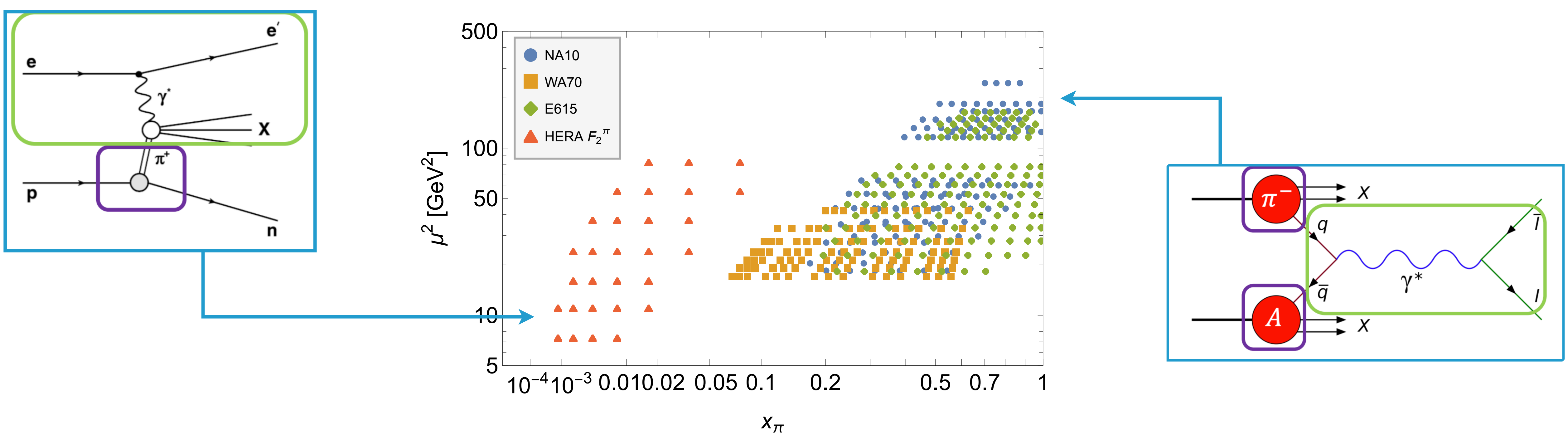
⇨ sampling of parameter space

Bézier-curve methodology for global analyses — the pion

We use the xFitter framework, in which `metamorph` was implemented as an independent parametrization.

We also extend the xFitter data:

- pion-induced Drell-Yan → constraints valence PDF at large x
- prompt photons → may constrain gluon PDF at largish x
- leading neutron (Sullivan process) → only constraints on sea and gluon at $x \lesssim 0.1$

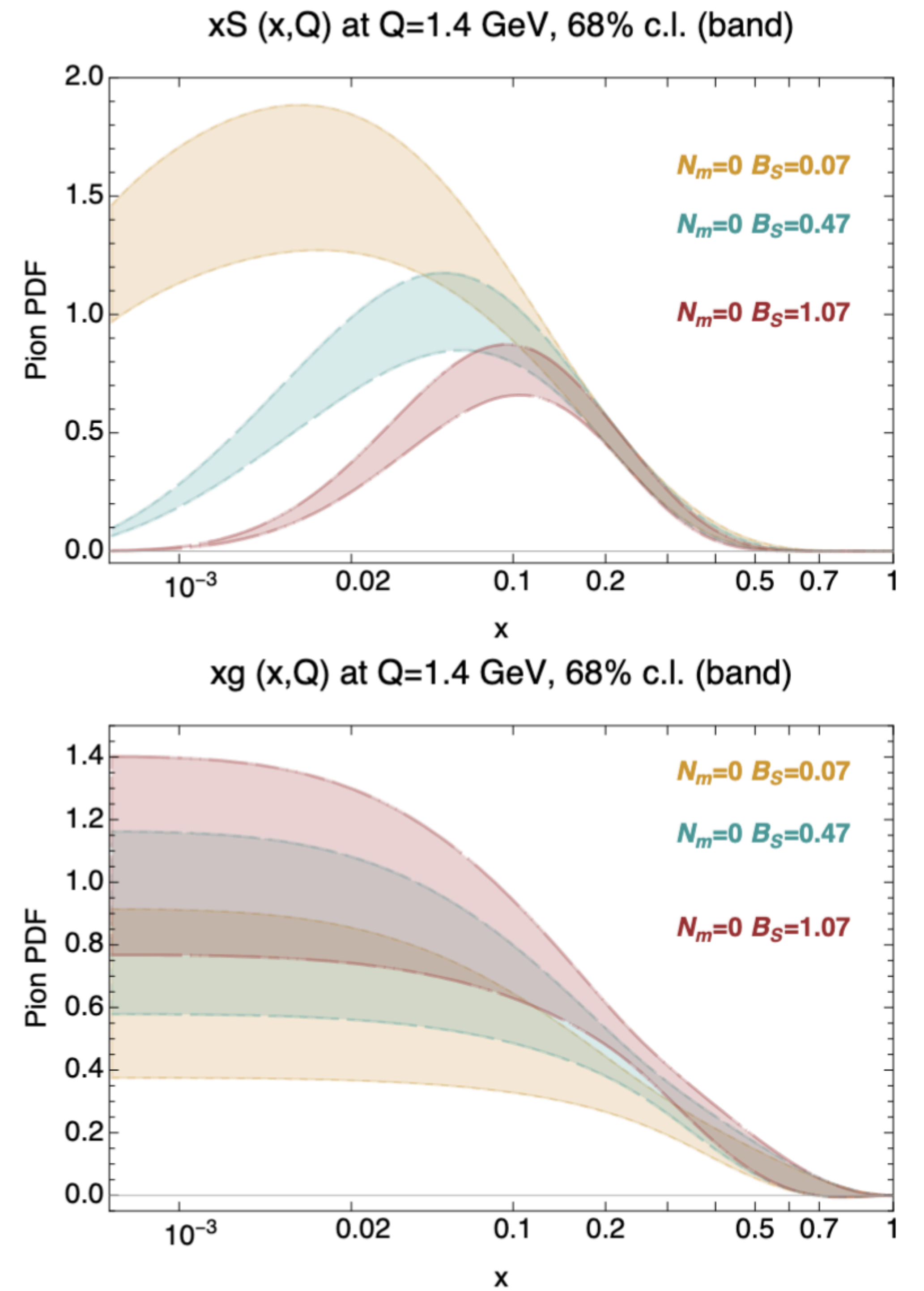


Drell-Yan only analysis

With a rigid parametrization, in Drell-Yan only analysis, the sea and gluon pion distributions are not well determined.

We can achieve equally good or better fits by varying the small x behaviour within xFitter uncertainty.

[Need for complementary processes— universality and flavor separation — EIC and JLab22(?)]()

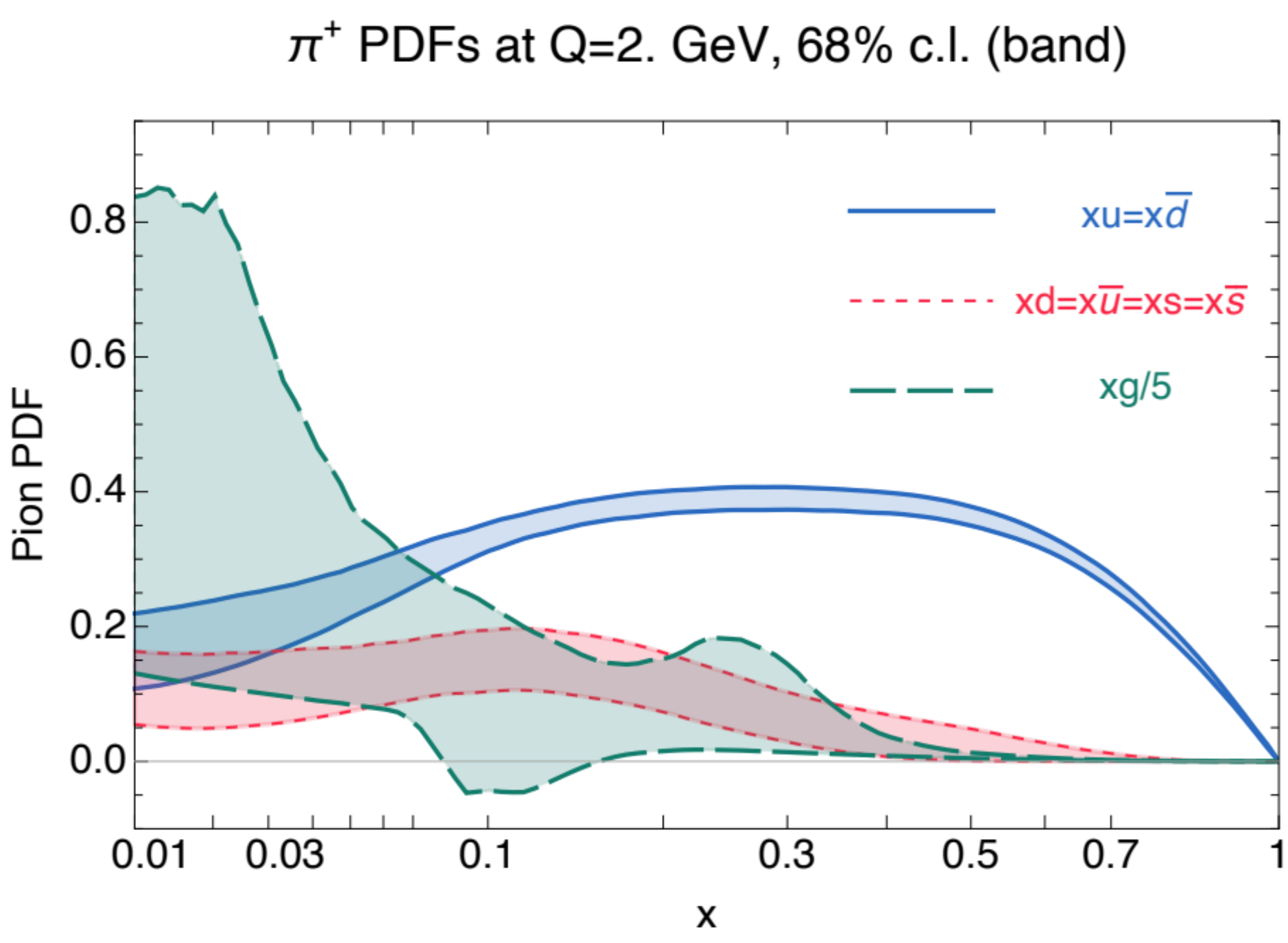


Bézier-curve methodology for global analyses — the pion

We use the xFitter framework, in which `metamorph` was implemented as an independent parametrization.

We also extend the xFitter data:

- pion-induced Drell-Yan → constraints valence PDF at large x
- prompt photons → may constrain gluon PDF at largish x
- leading neutron (Sullivan process) → only constraints on sea and gluon at $x \lesssim 0.1$

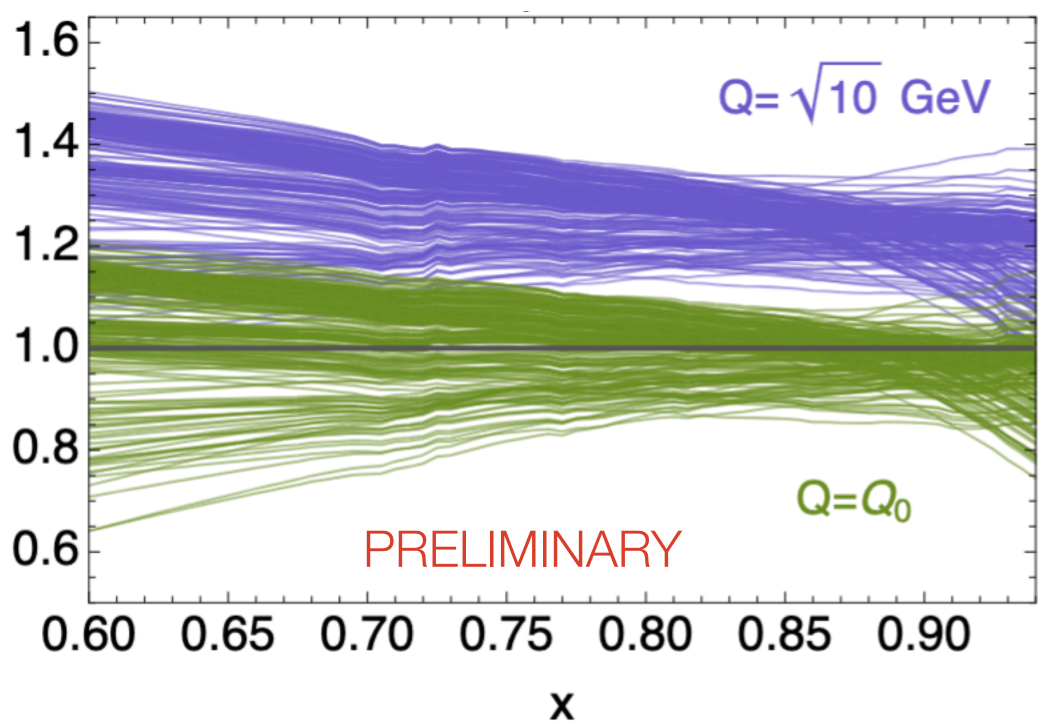


For a selection of $\{N_m, CP\}$sets.
Bundled uncertainty with `mcgen`
[Gao & Nadolsky, JHEP07]

*[Kotz, Ponce-Chávez, **AC,** Nadolsky & Olness, soon]*

Proceedings in 2309.00152.

Bézier-curve methodology for global analyses — the pion



At NLO (MSbar), the valence PDF is well determined at large x

⇨ doesn't fall very much like $(1-x)^2$

⇨ very similar to JAM and xFitter at large x

Corrective terms might need to be taken into account [large-x resummation].
JAM did and found an exponent between 1 to ~2.5, depending on the prescription [JAM, PRL127].

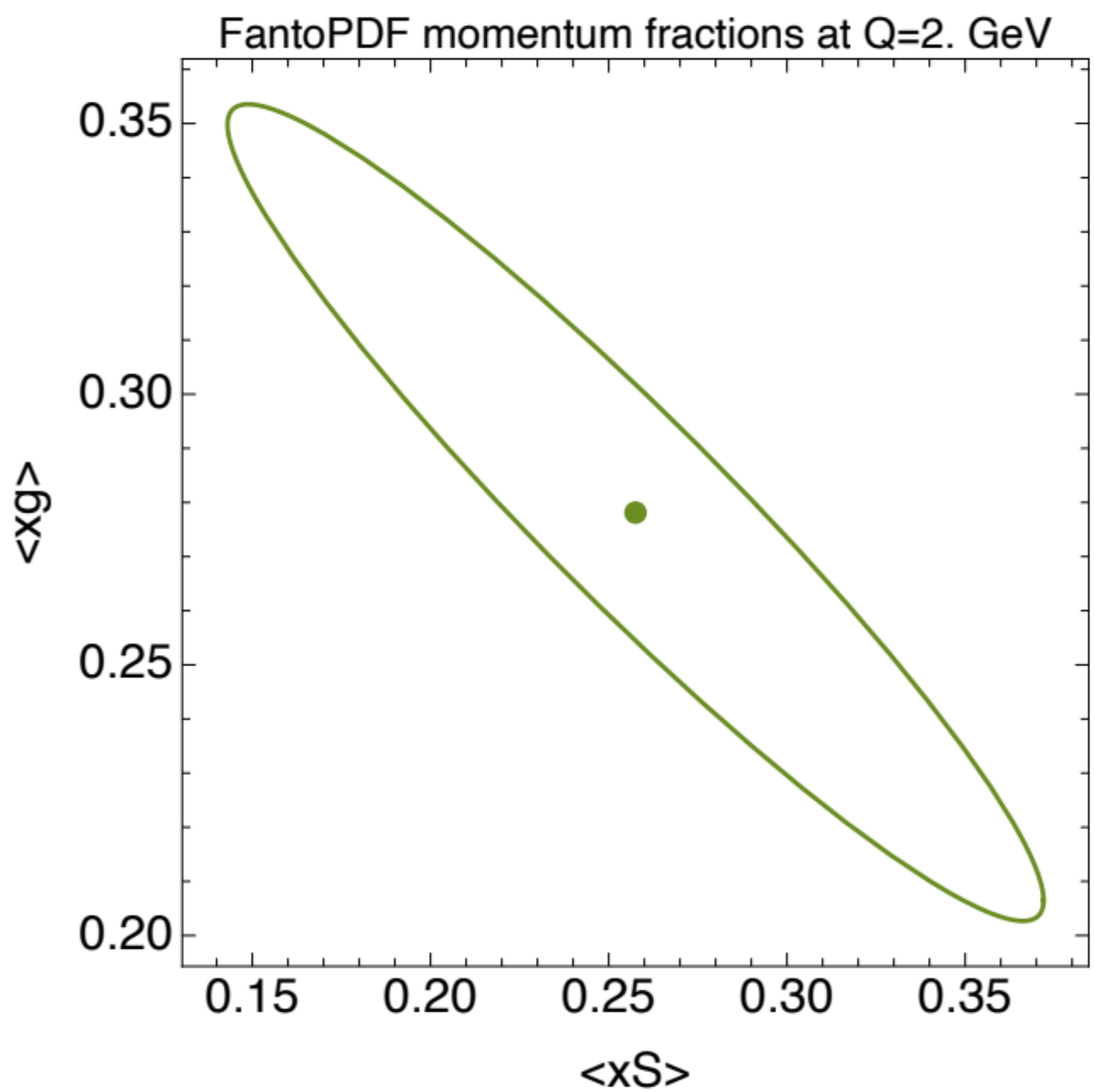
Lattice studies contribute to the information on hadron structure. Mindful analysis of the determination of the effective exponent of the PDF fall-off on the lattice [Gao et al., PRD102].

⇨ inverse problem

The Fantômas pion PDF

Towards epistemic uncertainty: sampling over parameter space more representative

Momentum-fraction distributions for gluon and sea are largely (anti)-correlated.



In contrast with the findings of JAM:
The inclusion of LN data does not drastically change the momentum fractions.

Conclusions

➪ Uncertainties come from various sources in global analyses.
Extension to sampling accuracy, here sampling occurs over parametrization forms.

➪ Rôle of the parametrization in the sampling accuracy: we make use of Bézier-curve methodology

Fantômas4QCD framework *[to appear very soon]*
`metamorph` can be used to study many functions

Reliable uncertainty on the PDF analysis (to NLO)
re: larger where no data constrains $q^\pi(x, Q^2)$

➪ End-point behavior of pion distributions seems to follow the trend given by mass generation vs. quark-counting rules.

Uncertainty quantification in non-perturbative calculations?
At what Q will pQCD (constraints) take over?

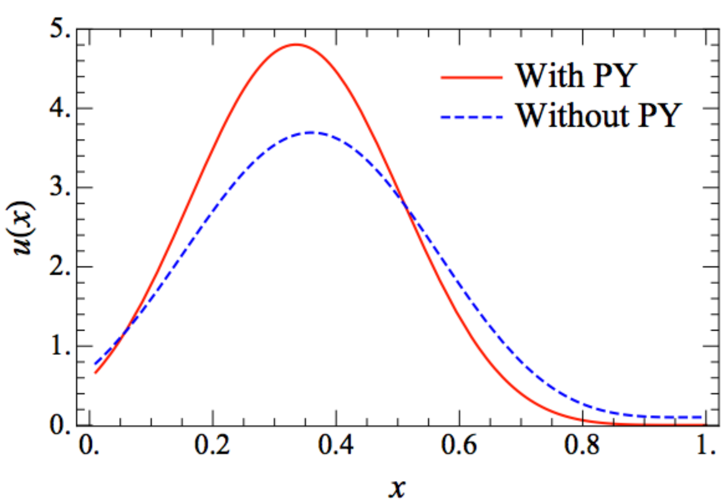

LA AMENAZA
ELEGANTE

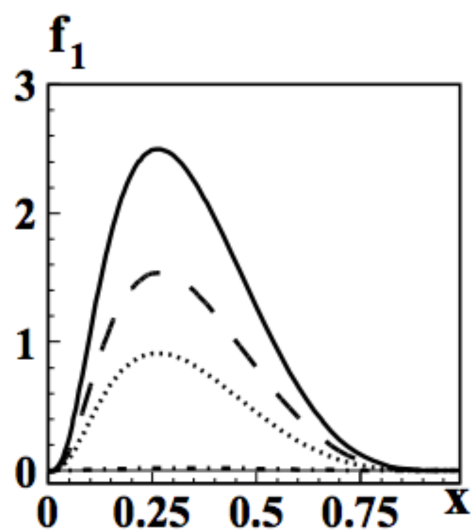
Back up

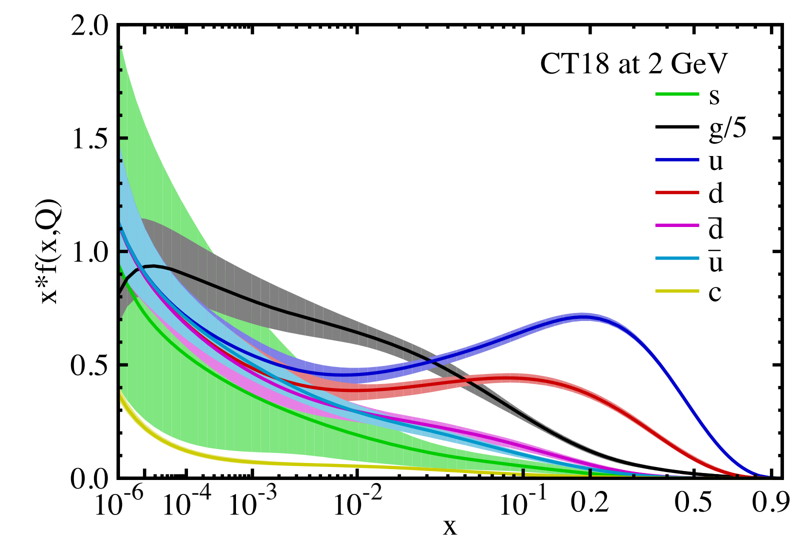
Model inputs and connection to phenomenology

How to compare phenomenological distribution functions to nonperturbative manifestations or behaviors that are characteristic of models for hadron structure?

But pheno PDFs cannot validate those specifics so easily







Behaviors in e.g. MIT bag model, light-cone constituent quark model, …

Global analysis groups: CT (illustrated), MSHT, NNPDF, JAM,…

Hypothesis testing from phenomenological PDF:

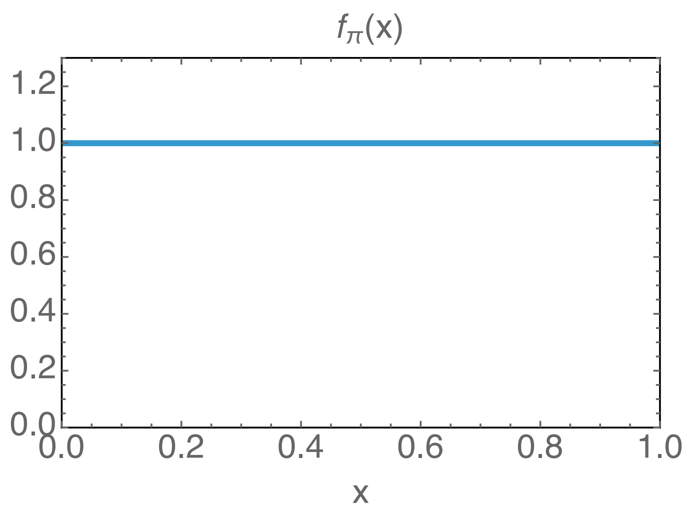
- ⇨ for local true or false statements
- ⇨ for functional behavior constraints

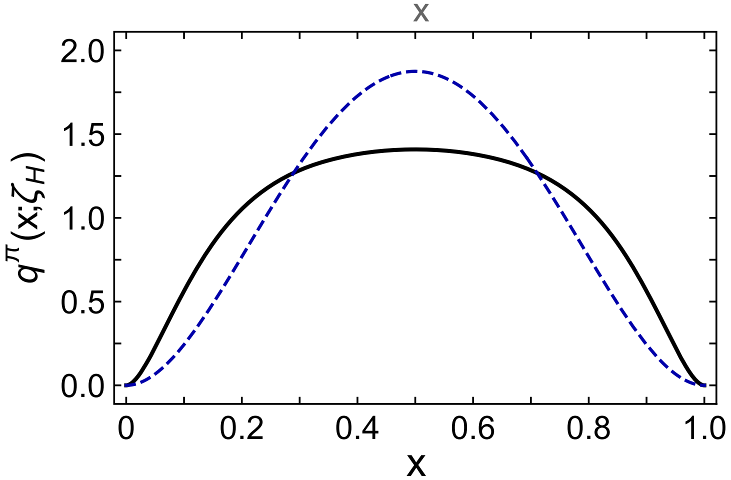
*[**AC** & P. Nadolsky, Phys.Rev.D 103 (2021)]*

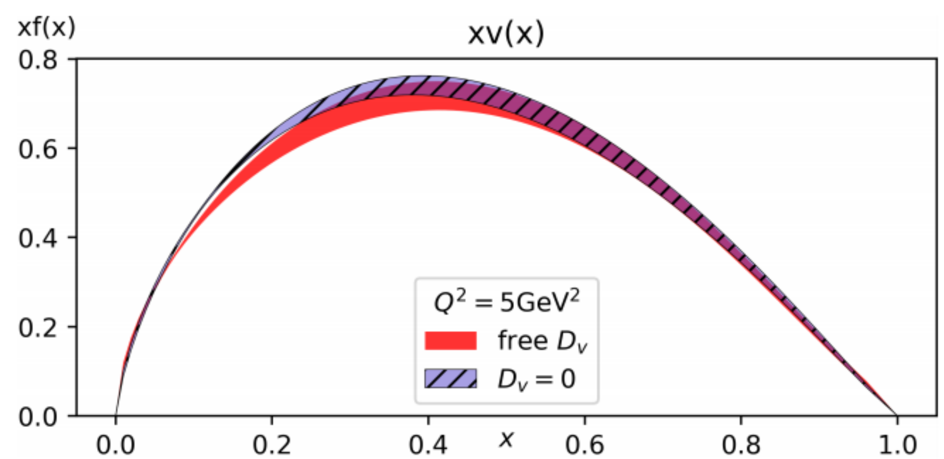
Model inputs and connection to phenomenology: the pion

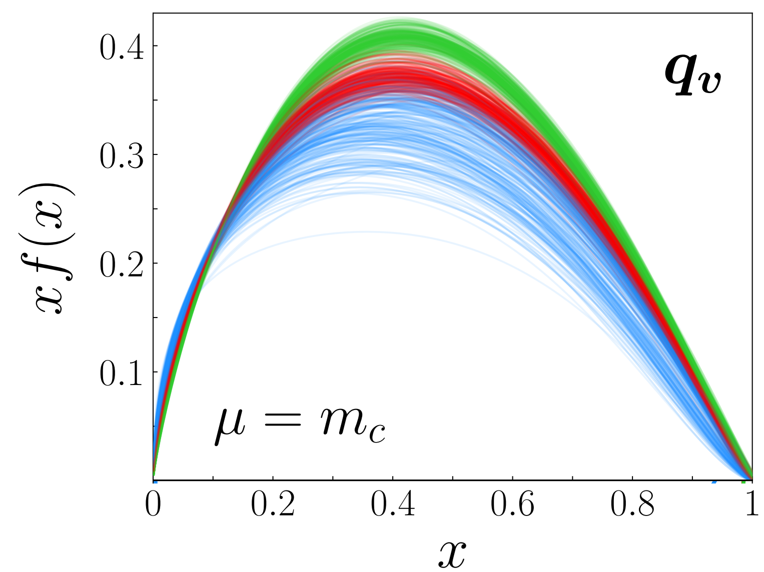
Pion PDFs are closely related to the dynamics of QCD in non-perturbative regime.

Trickier interpretation due to its pseudo-Goldstone nature and ansatze for exclusive-to-inclusive relations.









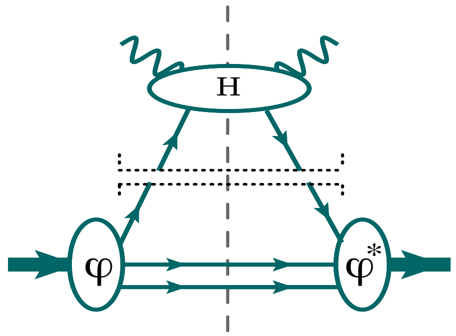
e.g. Nambu—Jona-Lasinio model, Schwinger-Dyson approaches, …

Global analysis groups: xFitters, JAM,…

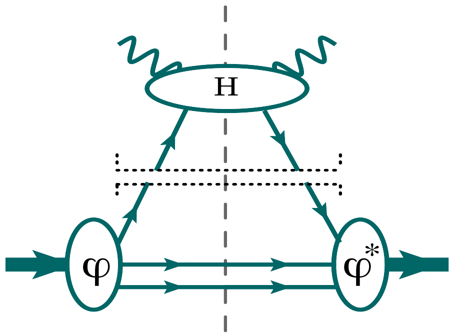
Can we test quark counting rules with pheno PDFs?

Early-QCD predicted behavior for structure functions when one quark carries almost all the momentum fraction:

$$f_{q_v/P}(x) \xrightarrow[x\to 1]{} (1-x)^3, \qquad f_{q_v/\pi}(x) \xrightarrow[x\to 1]{} (1-x)^2$$



Can we test quark counting rules with pheno PDFs?



Early-QCD predicted behavior for structure functions when one quark carries almost all the momentum fraction:

$$f_{q_v/P}(x) \xrightarrow[x\to 1]{} (1-x)^3, \qquad f_{q_v/\pi}(x) \xrightarrow[x\to 1]{} (1-x)^2$$

Evidence of polynomial form

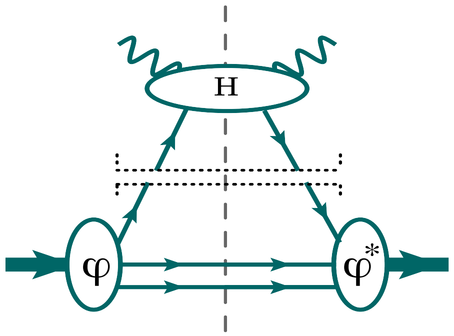
There is more than one possible solution to the choice of functional form.

$\Longleftrightarrow$

Agreement of model with data

Uncertainties needed for a faithful conclusion.

Can we test quark counting rules with pheno PDFs?



Early-QCD predicted behavior for structure functions when one quark carries almost all the momentum fraction:

$$f_{q_v/P}(x) \xrightarrow[x\to 1]{} (1-x)^3, \qquad f_{q_v/\pi}(x) \xrightarrow[x\to 1]{} (1-x)^2$$

Evidence of polynomial form

There is more than one possible solution to the choice of functional form.

$\Longleftrightarrow$

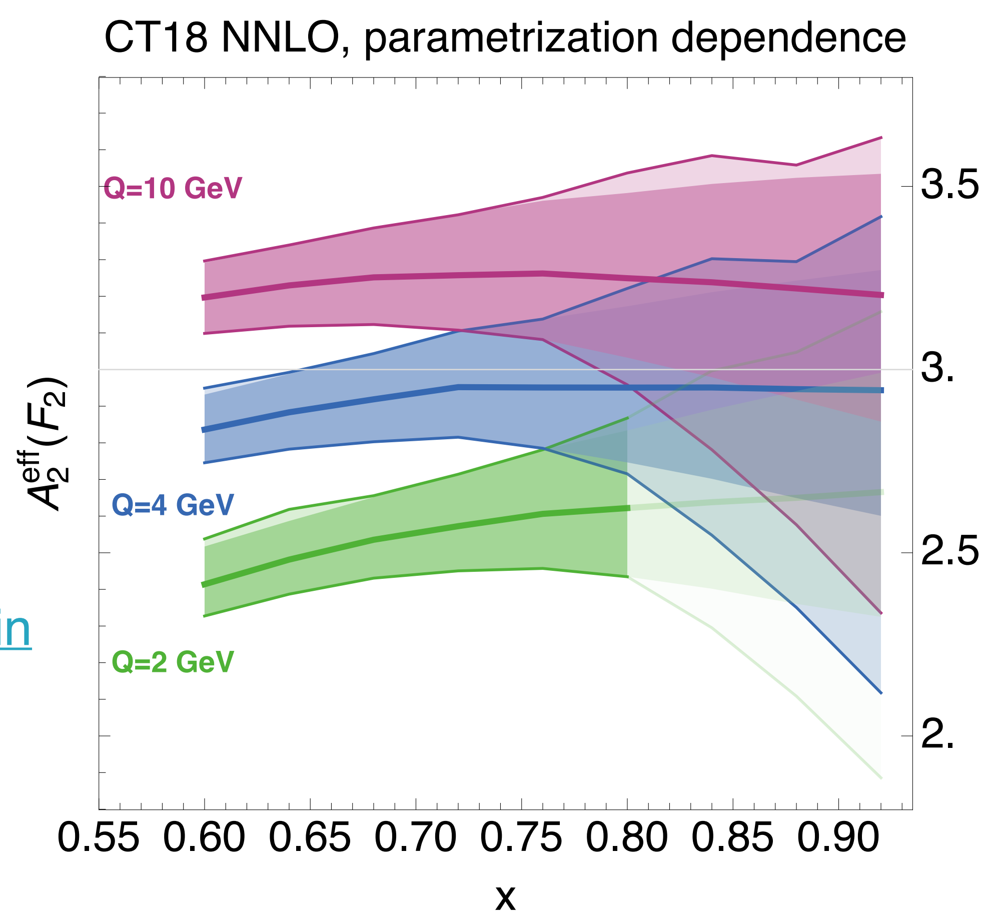
Agreement of model with data

Uncertainties needed for a faithful conclusion.

Effective exponent for x →1

$$A_2^{\mathrm{eff}}(F) \equiv \frac{\partial \ln\left(F(x,Q)\right)}{\partial \ln\left(1-x\right)}$$

Structure Function follows QCRs <u>within uncertainties</u>, dominated by parametrization dependence.

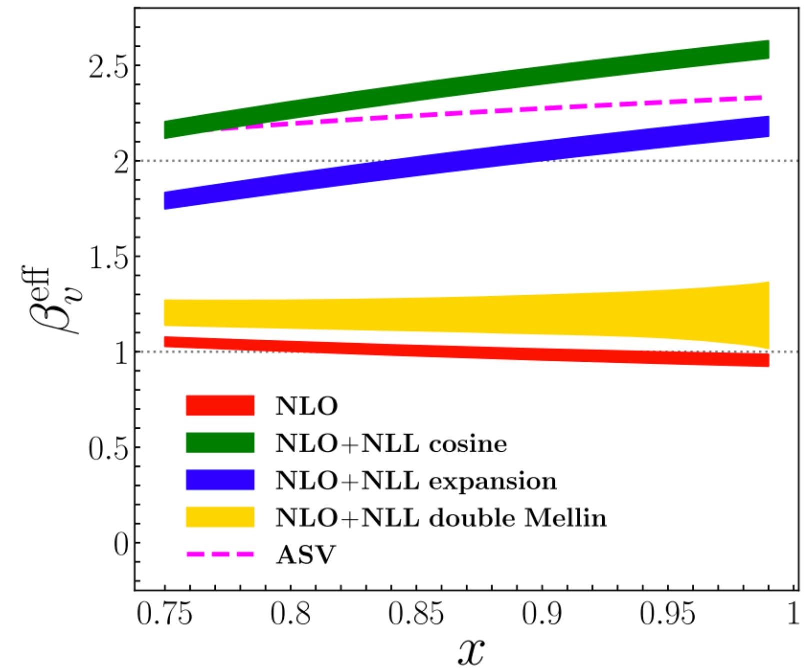


State-of-the-art of the pion at large x

Polynomial mimicry prevents functional behaviors from being validated as *if and only if* conditions.

Mathematical equivalence of polynomials of different orders can be illustrated with Bézier curves.
QCD corrections, at low and large Q^2, also inhibit the $(1-x)^\beta$ power to be tested.

[JAM, PRL127]



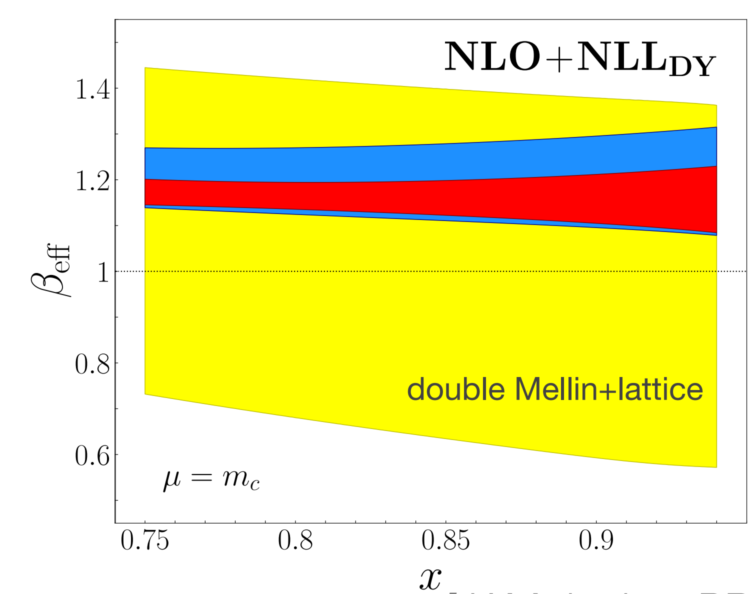
Most global analyses find $A_{2,\text{eff}}$ (β_{eff}) ~1 [xFitter, JAM].

Corrections from threshold resummution allow for $A_{2,\text{eff}}$ (β_{eff}) from 1 to ~2.5.

Lattice studies contribute to the information on hadron structure. Mindful analysis of the determination of the effective exponent of the PDF fall-off on the lattice [Gao et al., PRD102].

⇨ inverse problem

Pheno and lattice PDF of the pion compatible with QCRs within uncertainties.



[JAM+lattice, PRD105]

PDFs in nonperturbative QCD

- **at hadronic scale** $\mu_0^2 < 1$ **GeV**^{2}
  - ⇨ prefactorization picture
  - ⇨ nonperturbative dynamics
  - ⇨ model's degrees of freedom

Phenomenological PDFs

- **at factorization scale** $\mu^2 > 1$ **GeV**^{2}
  - ⇨ quasi-free partonic degrees of freedom
  - ⇨ defined in the MSbar scheme
  - ⇨ leading-power approximation to full dynamics

PDFs in nonperturbative QCD

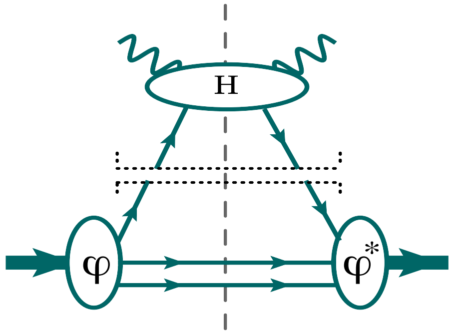
- **at hadronic scale $\mu_0^2 < 1$ GeV^2**
  - ⇨ prefactorization picture
  - ⇨ nonperturbative dynamics
  - ⇨ model's degrees of freedom

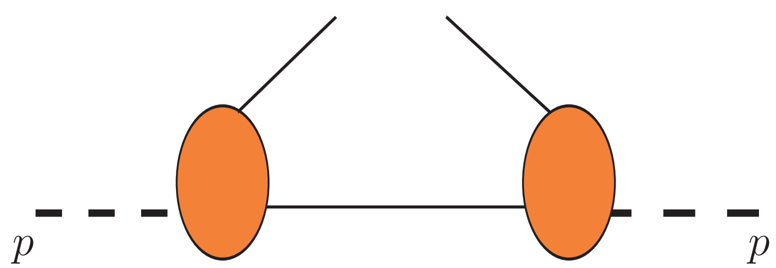
Phenomenological PDFs

- **at factorization scale $\mu^2 > 1$ GeV^2**
  - ⇨ quasi-free partonic degrees of freedom
  - ⇨ defined in the MSbar scheme
  - ⇨ leading-power approximation to full dynamics

How to relate the x dependence of the perturbative and nonperturbative pictures?

$$\int \frac{dz^-}{2\pi} e^{i\,(x-\frac{1}{2})\,z^- p^+} \langle \pi^+(p)|\bar{q}\left(-\frac{z}{2}\right)\, \not{n}\, \frac{1}{2}(1+\tau^3)\, q\left(\frac{z}{2}\right)|\pi^+(p)\rangle = \frac{1}{p^+}\, q(x)$$



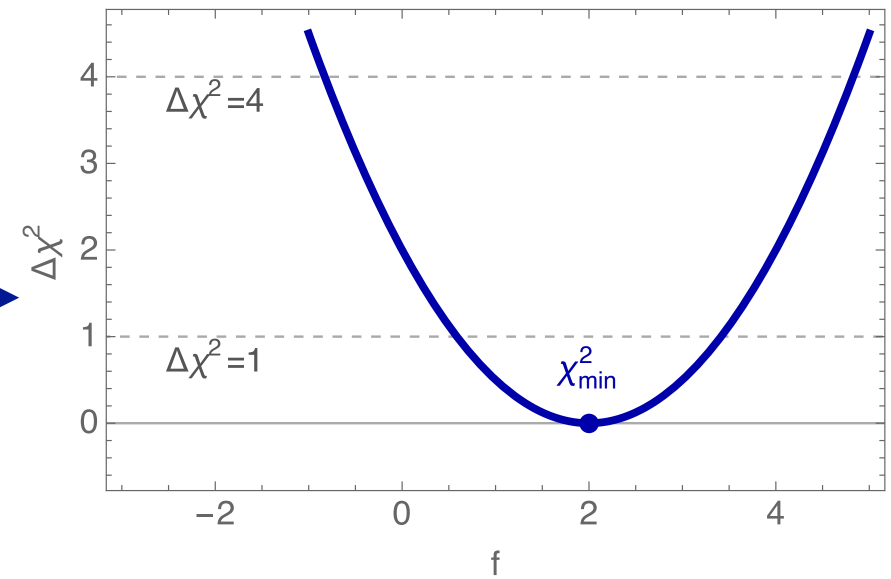


$$F(x_B, Q^2) = \sum_a \int_{x_B}^1 \frac{dx}{x} f_{a/p}(x, \mu^2) H_a\left(\frac{x_B}{x}, \frac{\mu^2}{Q^2}\right) + \mathcal{O}(M/Q),$$

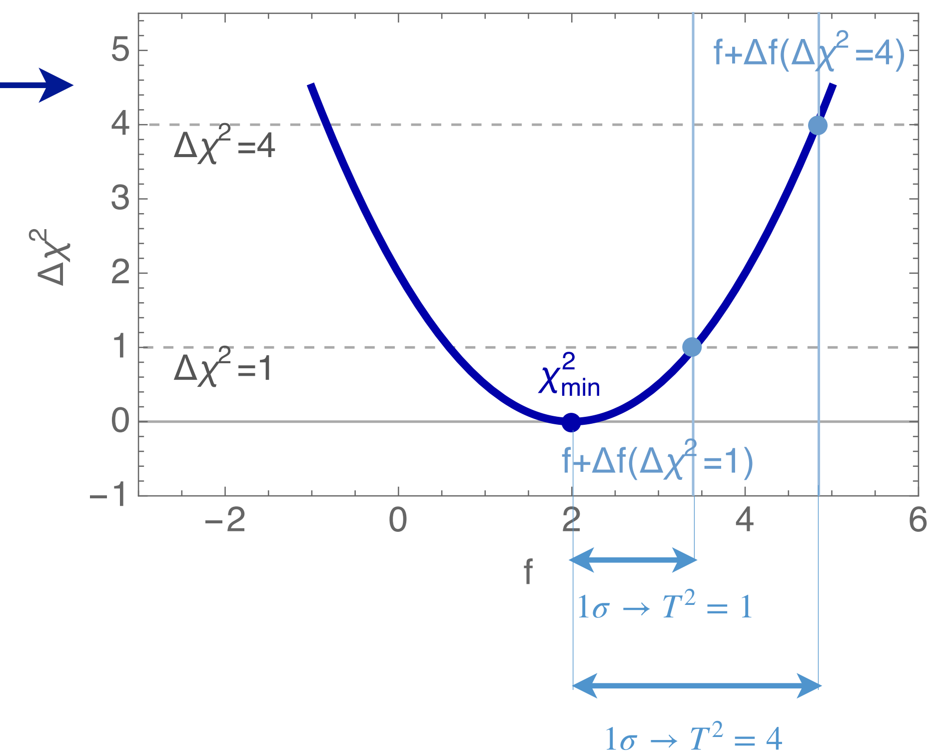
$$\sigma = \sum_{a,b} \int dx_a \int dx_b f_{a/A}(x_a, \mu_F^2) f_{b/B}(x_b, \mu_F^2) H_{a,b,x_a,x_b,\mu_F^2} + \mathcal{O}(M/Q),$$

IV. Towards a thorough understanding of uncertainties in global analyses

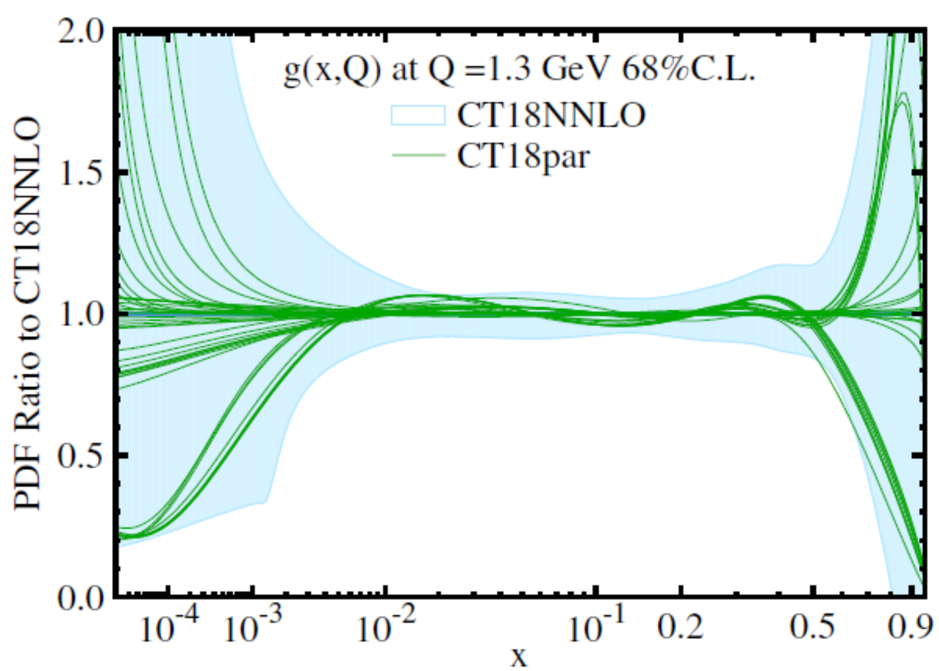
The χ^2 is a paraboloid in N_{par} dimensions.
We can project each dimension as



The criteria of $\Delta\chi^2 = 1$ does not account for various sources of uncertainties. A tolerance criteria can be defined to accommodate for those.



For $T^2 \sim 36$, the CT uncertainties at $68\ \%$ CL encompass up to the variations in the choice of parametrization.



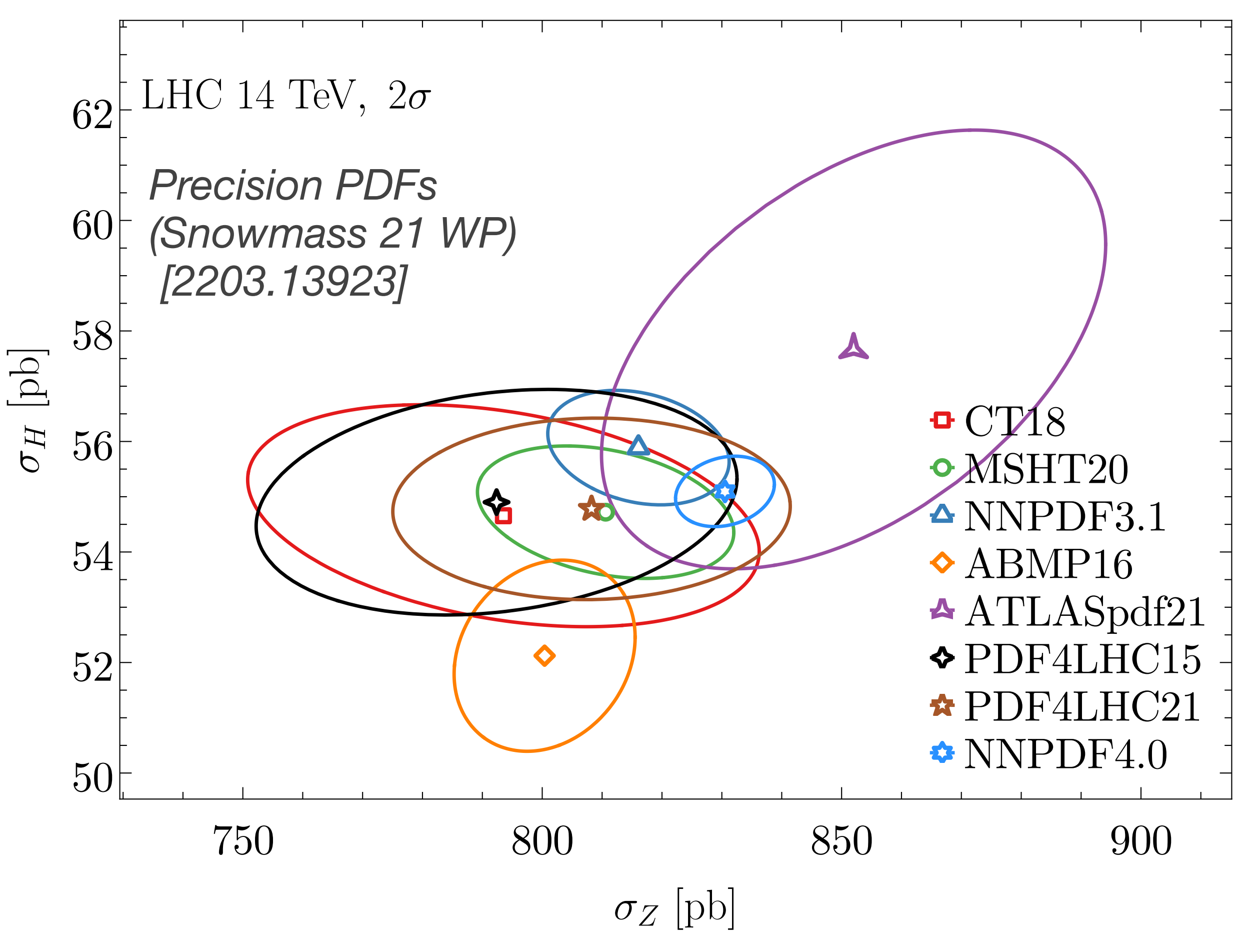
IV. Towards a thorough understanding of uncertainties in global analyses

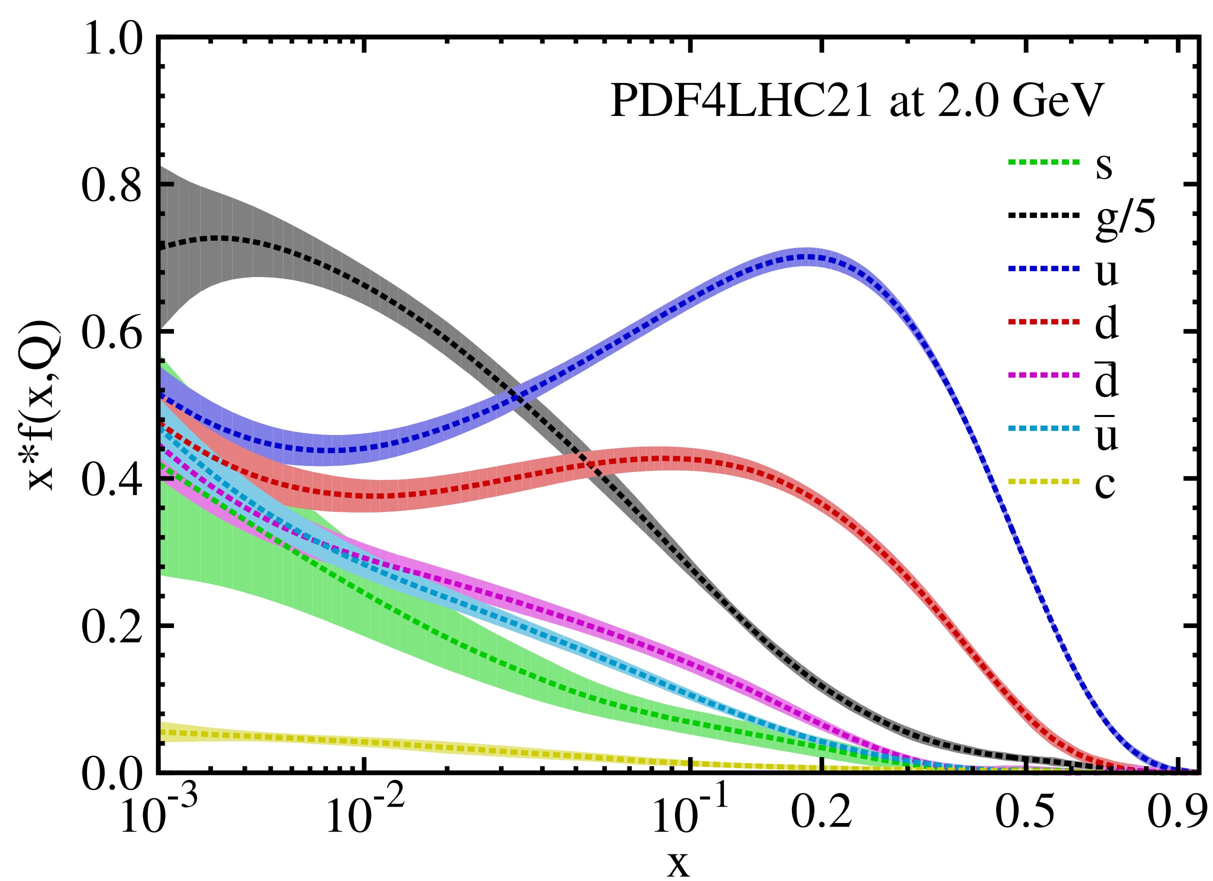
Recent advancements in the determination of unpolarized PDFs:
CT18, MSHT20, NNPDF4.0, ATLASpdf21 as well as PDF4LHC21.

PDF4LHC21:

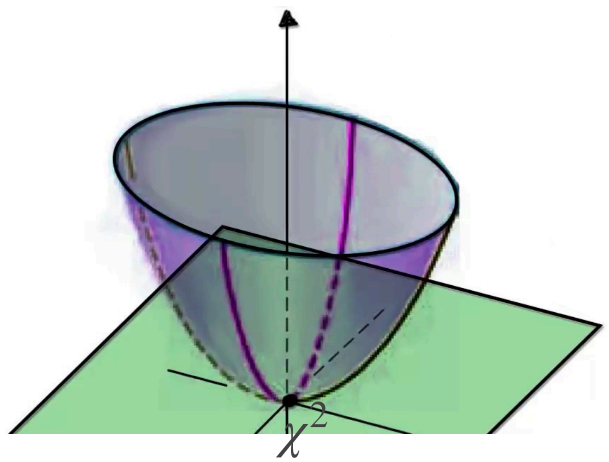
benchmarking and combination of the leader PDF sets, CT, MSHT & NNPDF, for the run III of the LHC.

[Ball, […], AC, et al, J.Phys.G 49 (2022)]

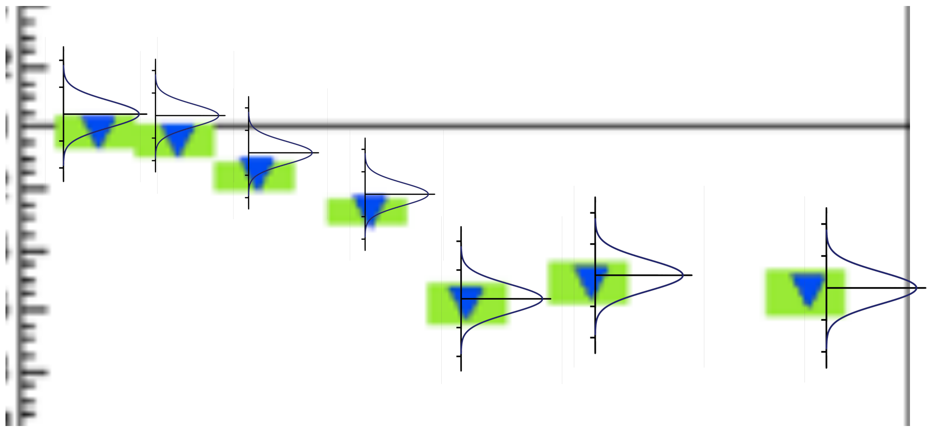




IV. Towards a thorough understanding of uncertainties in global analyses

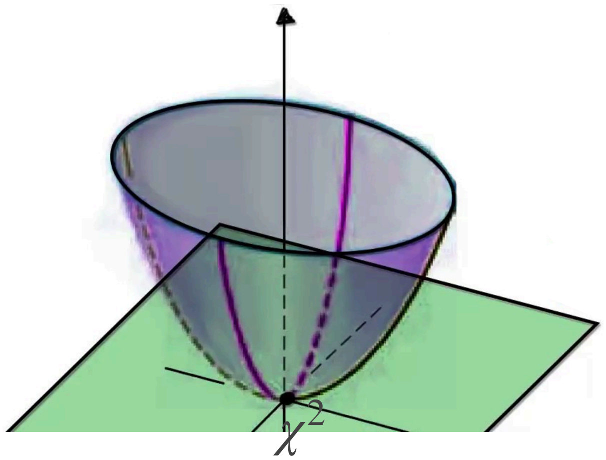


Hessian methodology finds the global minimum and explores the parameter space.

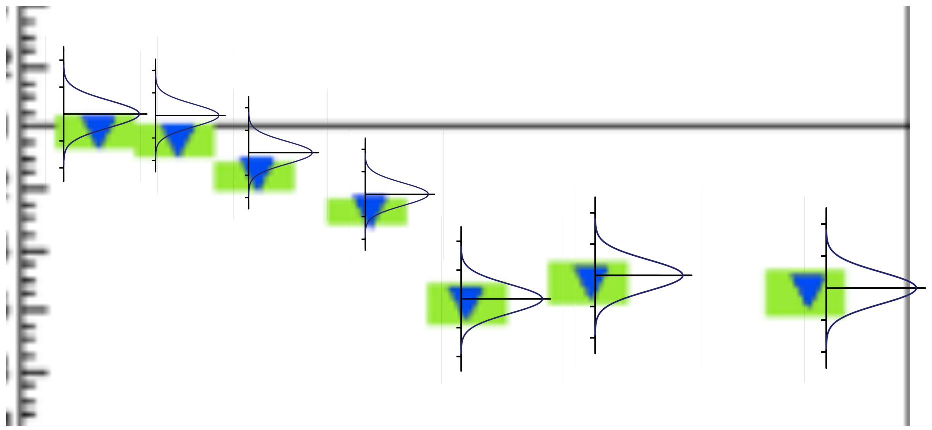


Monte-Carlo methodology (neural network, AI/ML) replicates fluctuated data, then optimizes each replica (up to training).

IV. Towards a thorough understanding of uncertainties in global analyses



Hessian methodology finds the global minimum and explores the parameter space.



Monte-Carlo methodology (neural network, AI/ML) replicates fluctuated data, then optimizes each replica (up to training).

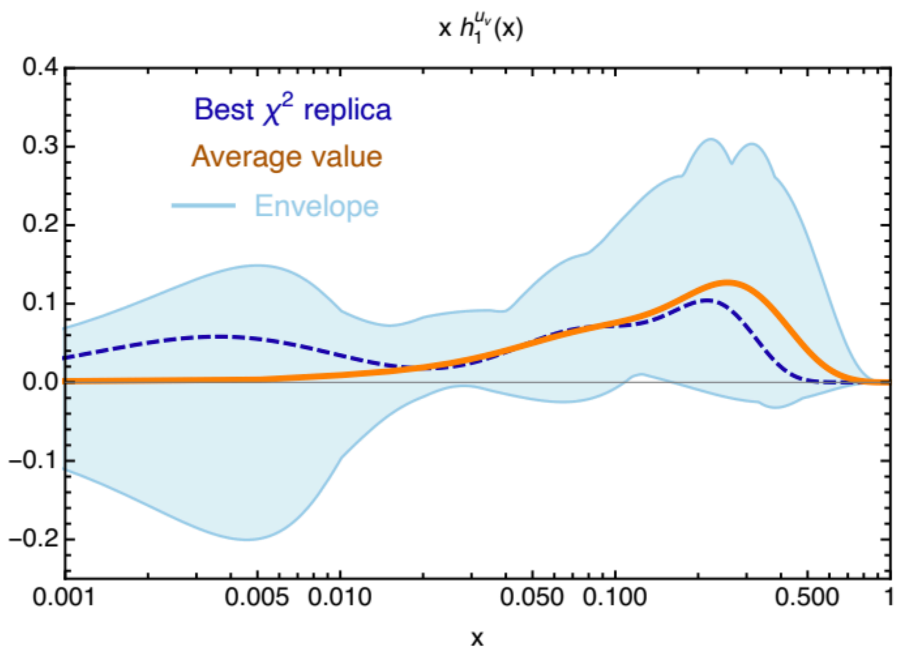


Illustration on *bootstrap* probability distribution with average value vs. the best replica for $N_{par} \sim 7$.

*[**AC**, SciPost Phys.Proc.8 (2022)]*

In multivariate analyses, sampling occurs at various levels — parameter space, bootstrap but also priors, … In large-dimensional problems, sampling is complex.

The tolerance puzzle and the big-data paradox

Outside of HEP, there is significant interest in statistical problems that are similar to the PDF tolerance problem. These studies introduce a fundamental distinction between the fitting uncertainty and sampling uncertainty, often overlooked in the PDF fits.

Article

Unrepresentative big surveys significantly overestimated US vaccine uptake

Nature **v. 600** (2021) 695

https://doi.org/10.1038/s41586-021-04198-4

Received: 18 June 2021

Valerie C. Bradley[1,6], Shiro Kuriwaki[2,6], Michael Isakov[3], Dino Sejdinovic[1], Xiao-Li Meng[4] & Seth Flaxman[5,12]

SCIENCE ADVANCES | RESEARCH ARTICLE

MATHEMATICS

Models with higher effective dimensions tend to produce more uncertain estimates

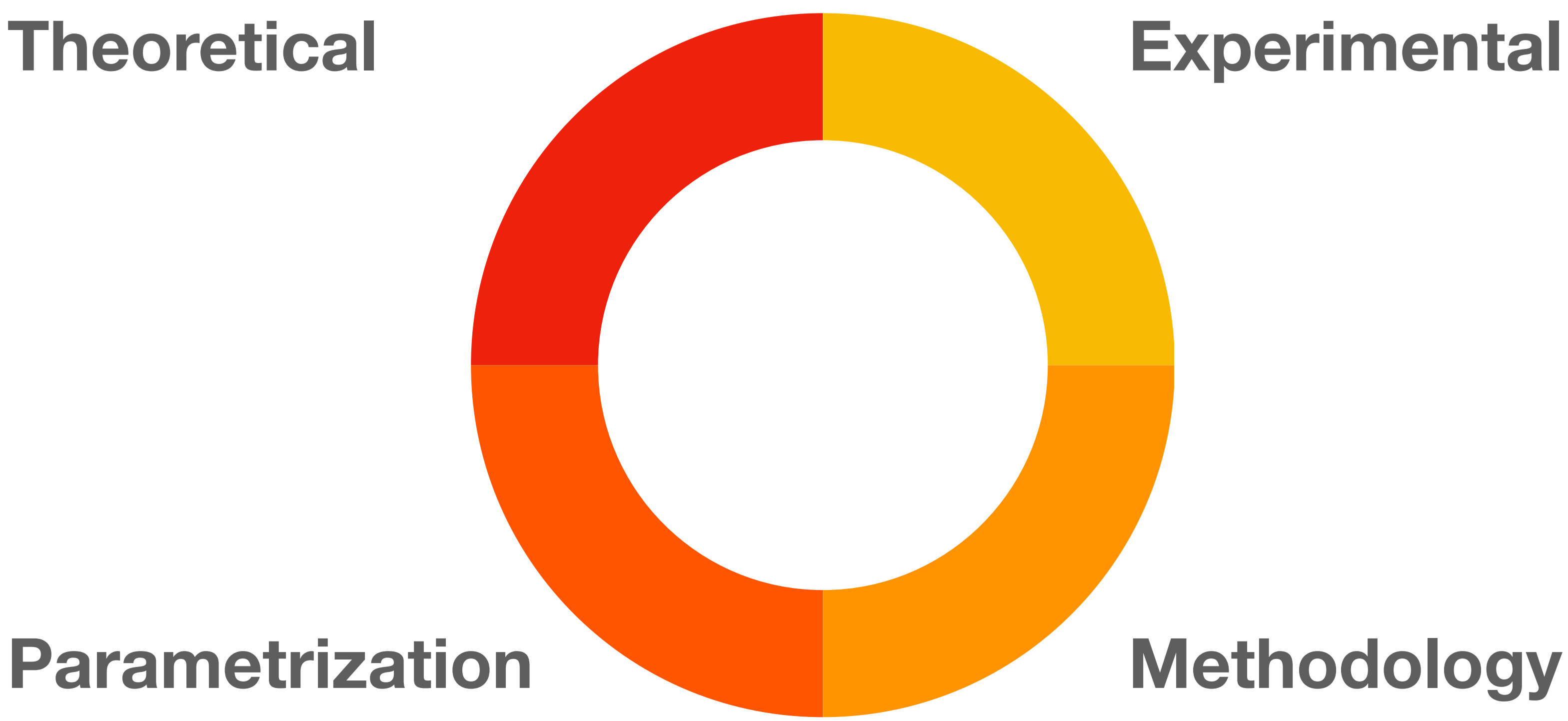
Arnald Puy[1,2,3*], Pierfrancesco Beneventano[4], Simon A. Levin[2], Samuele Lo Piano[5], Tommaso Portaluri[6], Andrea Saltelli[3,7]

The Big Data Paradox in Clinical Practice

Pavlos Msaouel

To cite this article: Pavlos Msaouel (2022) The Big Data Paradox in Clinical Practice, Cancer Investigation, 40:7, 567-576, DOI: 10.1080/07357907.2022.2084621

A new avenue to understand PDF tolerance



In all four categories of uncertainties, we can further distinguish *PDF fitting accuracy* from *PDF sampling accuracy*.

Goodness-of-fit applies to an individual best fit.

[Kovarik et al, Rev.Mod.Phys. 92 (2020)]

Sampling accuracy applies either to the tolerance or the number of error sets in a PDF ensemble.

This talk.

Sampling bias in PDF global analyses—I

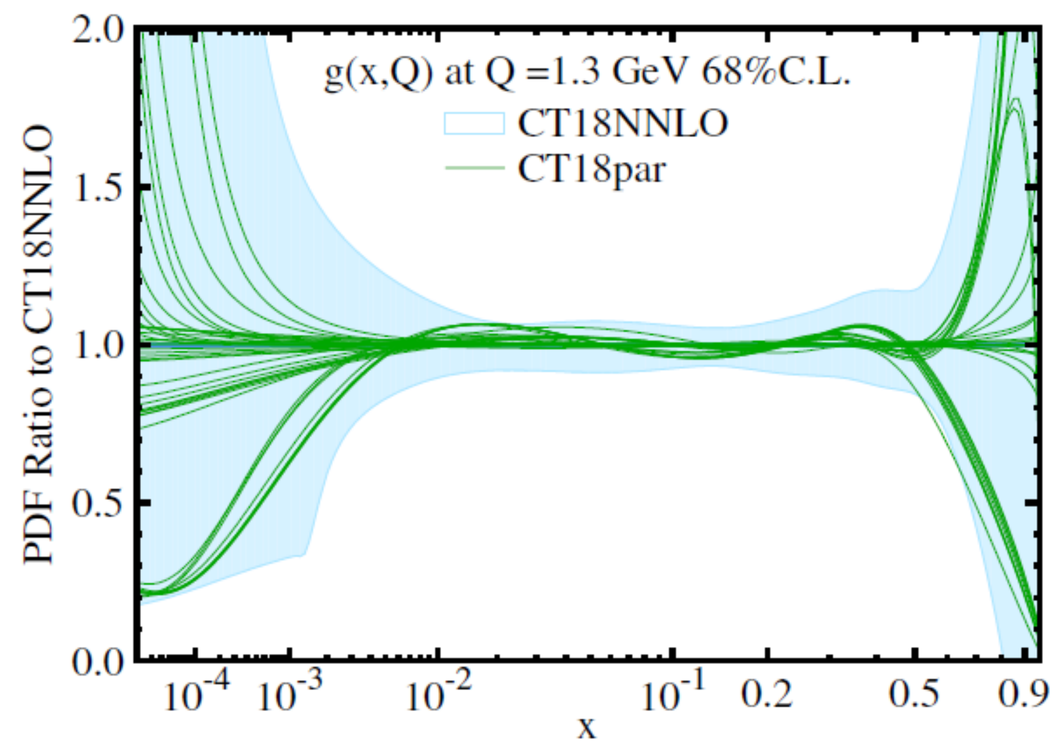
How do we know the “data+sampling defect=confounding correlation” of our analysis?

Hessian-based analysis:
objective function includes penalties, establishing the **tolerance criteria.**

Size of uncertainties reflect a series of confounding sources —selection of fitted experiments, treatment of correlated systematic errors, functional forms of PDFs, …

Verification that proper spanning of parameter space is compatible with total uncertainties (*a posteriori*).
>300 functional forms are tested in CT18.

Dimensions of the problem given by the number of parameters=eigenvector (EV) directions.



Hou et al, Phys.Rev.D 103 (2021)

Sampling bias in PDF global analyses—II

How do we know the “data+sampling defect=confounding correlation” of our analysis?

Monte Carlo-based analysis:
optimization implies selection of hyperparameters

The usage of Neural Networks had as primary goal eliminating the biases associated with the choice of a specific functional form.

However, there are still many choices associated with the optimization:

- Number and with of the layers
- Activation functions and initialization
- Optimization algorithm (and associated parameters)
- Training length, stopping patience, etc.
- Strength of lagrange multipliers (positivity, integrability)

Collectively called “hyperparameters”, usually selected manually.

CERN QCD Seminar
Cruz Martínez, 11/2022

Do we understand sampling for QCD global analyses?

Sampling of multidimensional spaces ($d \gg 20$) is exponentially inefficient and may require $n > 2^d$ replicas to obtain a convergent expectation value.

In general, an intractable problem.

[Hickernell, MCQMC 2016, 1702.01487]
[Sloan, Wo´zniakowski, 1997]

Do we understand sampling for QCD global analyses?

Sampling of multidimensional spaces ($d \gg 20$) is exponentially inefficient and may require $n > 2^d$ replicas to obtain a convergent expectation value.

In general, an intractable problem.

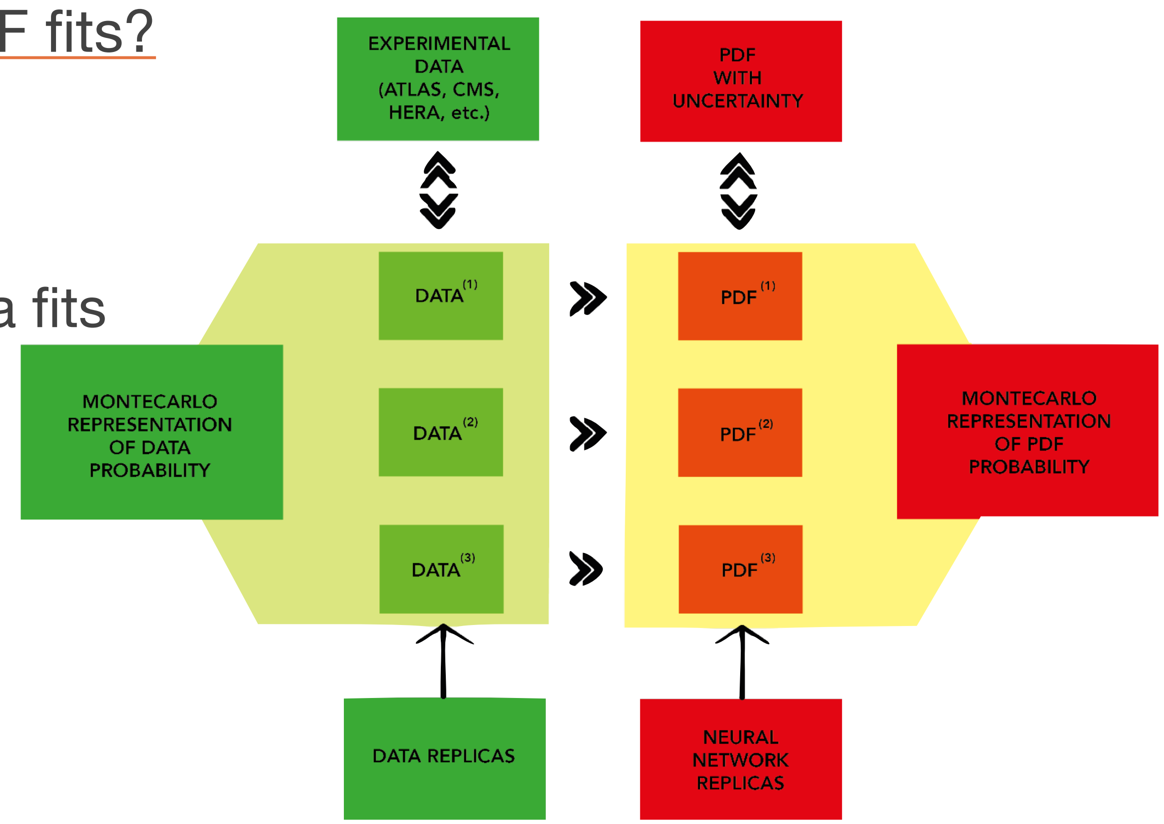
[Hickernell, MCQMC 2016, 1702.01487]
[Sloan, Wo´zniakowski, 1997]

1. Justification for tolerance criteria for Hessian-based PDF fits

2. How is sampling achieved in Monte Carlo-based PDF fits?

Importance sampling, as defined by NNPDF

- =bootstrap/resampling of random fluctuations in data
- expectations are then unweighted averages over replica fits



Such sampling does not include sampling over hyperparameters and priors.

Do we understand sampling for QCD global analyses?

Sampling of multidimensional spaces ($d \gg 20$) is exponentially inefficient and may require $n > 2^d$ replicas to obtain a convergent expectation value.

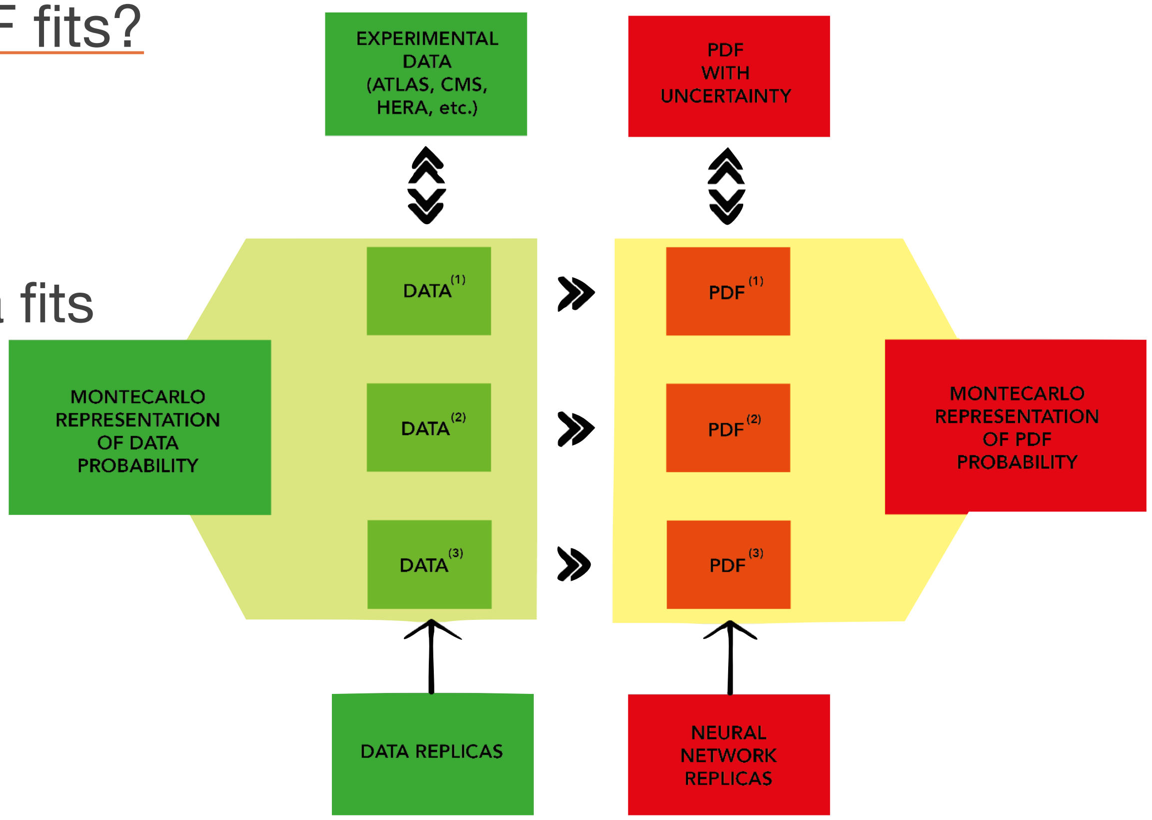
In general, an intractable problem.

[Hickernell, MCQMC 2016, 1702.01487]
[Sloan, Wo´zniakowski, 1997]

1. Justification for tolerance criteria for Hessian-based PDF fits

2. How is sampling achieved in Monte Carlo-based PDF fits?

Importance sampling, as defined by NNPDF
- =bootstrap/resampling of random fluctuations in data
- expectations are then unweighted averages over replica fits



Such sampling does not include sampling over hyperparameters and priors.

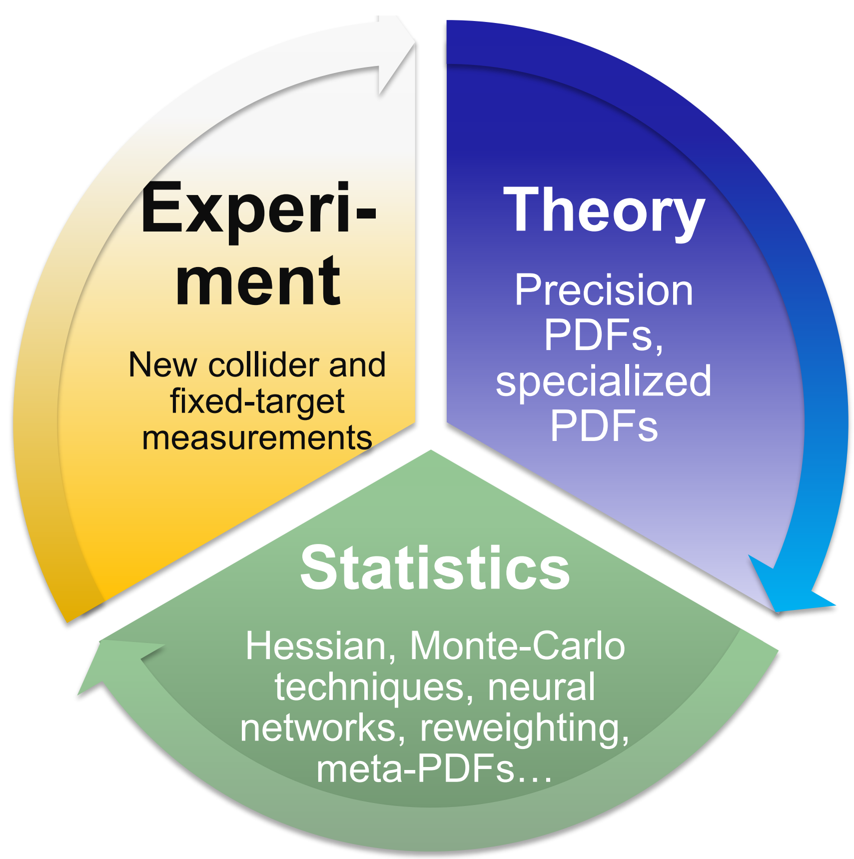
Global analyses of unpolarized PDF — CT

The **C**oordinated **T**heoretical-**E**xperimental project on **Q**CD (CTEQ) is a high-energy physics collaboration whose efforts include *fits* of unpolarized PDFs. This is done in the ***C**TEQ-**T**ung et al* (CT) group.

CT is a renown fitting group, whose PDF sets are widely used in colliders, …

Leading the characterization of uncertainties in PDF analyses and on the connection to nuclear physics (relevant for EIC and JLab physics).

Present data sets (HERA,…)
LHC data
Future: EIC, …



Perturbative QCD (to various loop precision, QCD corrections, …)
First principles
QCD-based ansatze
Low-energy corrections
…

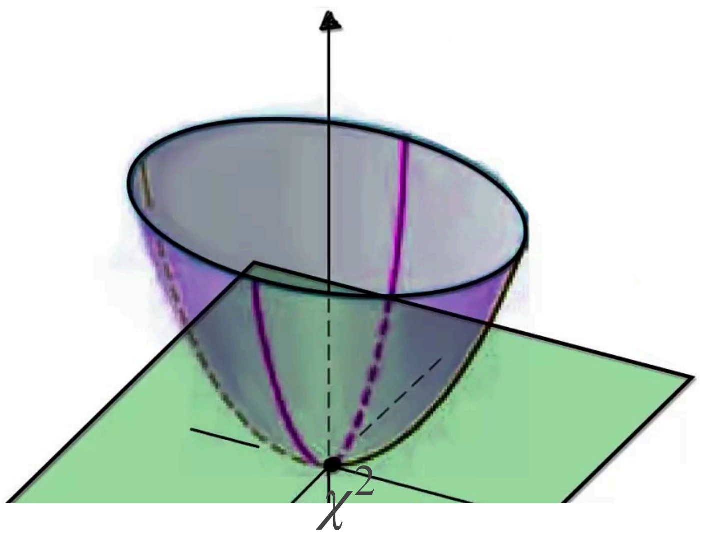
<u>Methodology</u>
Objective function
Sampling
Criteria for uncertainties
…

Global analyses of unpolarized PDF — CT

The ***C**TEQ-**T**ung et al* (CT) group.

CTEQ-TEA members (as of 2023)

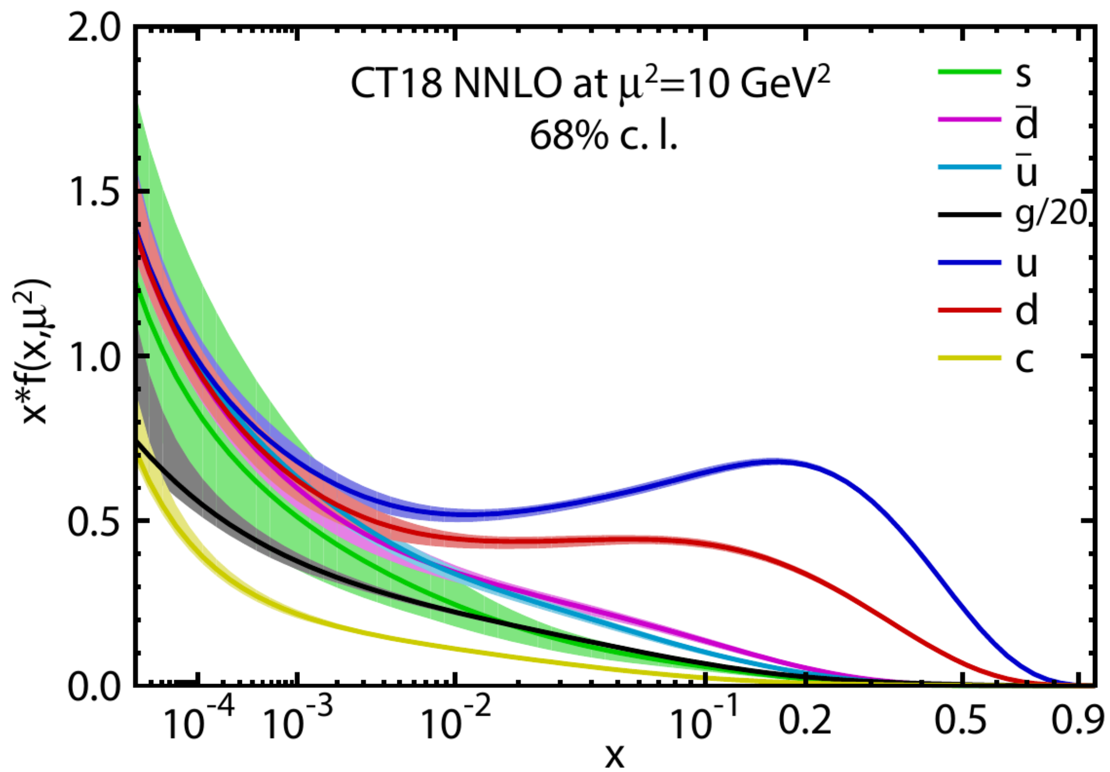
China: S. Dulat, J. Gao, T.-J. Hou, I. Sitiwaldi, M. Yan, and collaborators
Mexico: A. Courtoy
USA: T.J. Hobbs, M. Guzzi, J. Huston, P. Nadolsky, C. Schmidt, D. Stump, K. Xie,



*The **C**TEQ-**T**ung et al [(CT) PDF set.](#)*

CT18 is the latest released PDF set.

CT methodology is based on minimizing a χ^2 expressed in terms of *parametrizations* for the PDFs, finding the global minimum and propagating the uncertainty through the Hessian formalism.



[Hou et al, Phys.Rev.D 103 (2021)]

Hypothesis testing: role of uncertainties

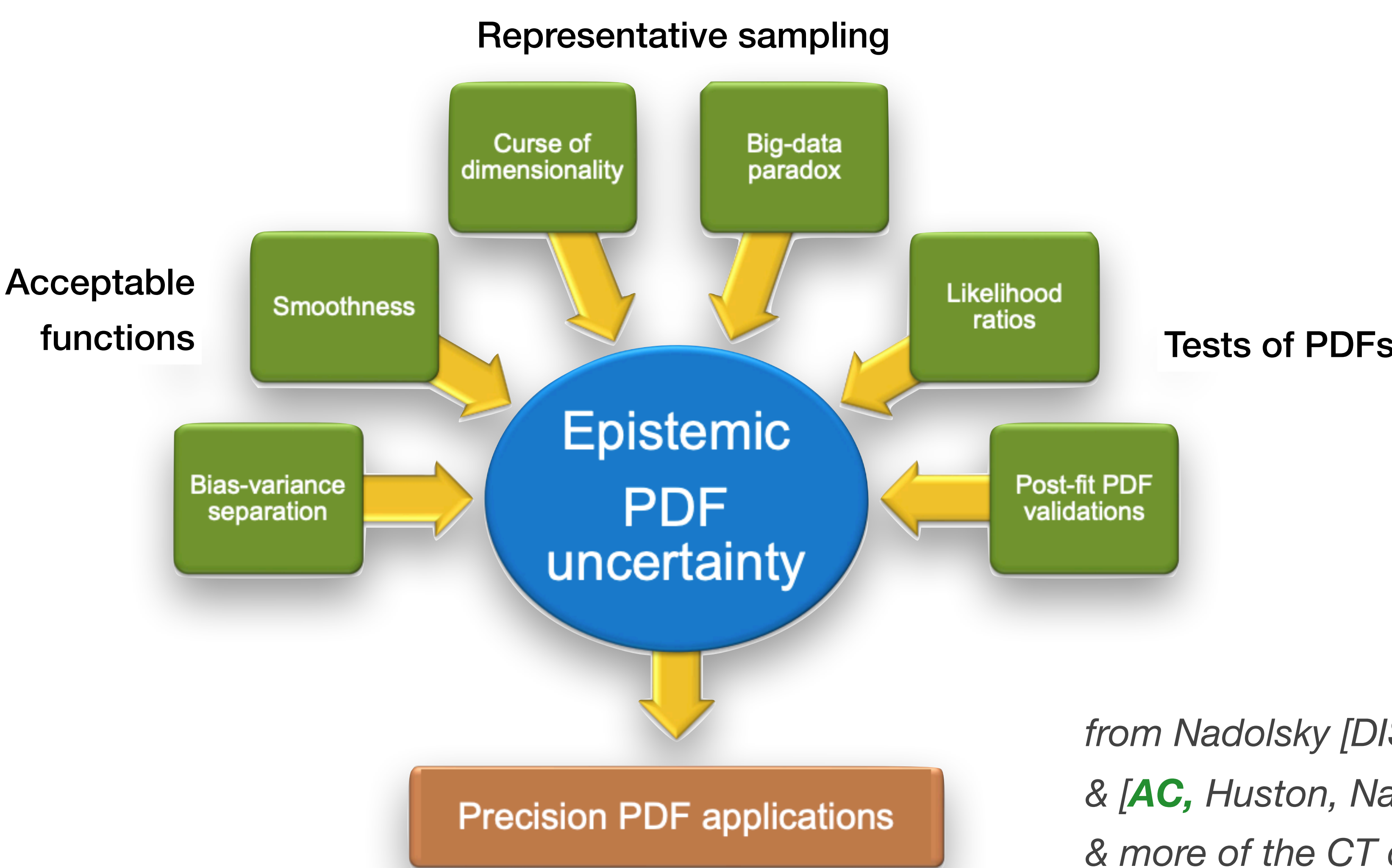
Hypothesis testing of theoretical predictions relies on

1. available data in x range, as well as value of Q,
2. sensitivity of data to the hypothesis,
3. quality of the data,
4. **uncertainties found in the fits.**

Hypothesis testing: role of uncertainties

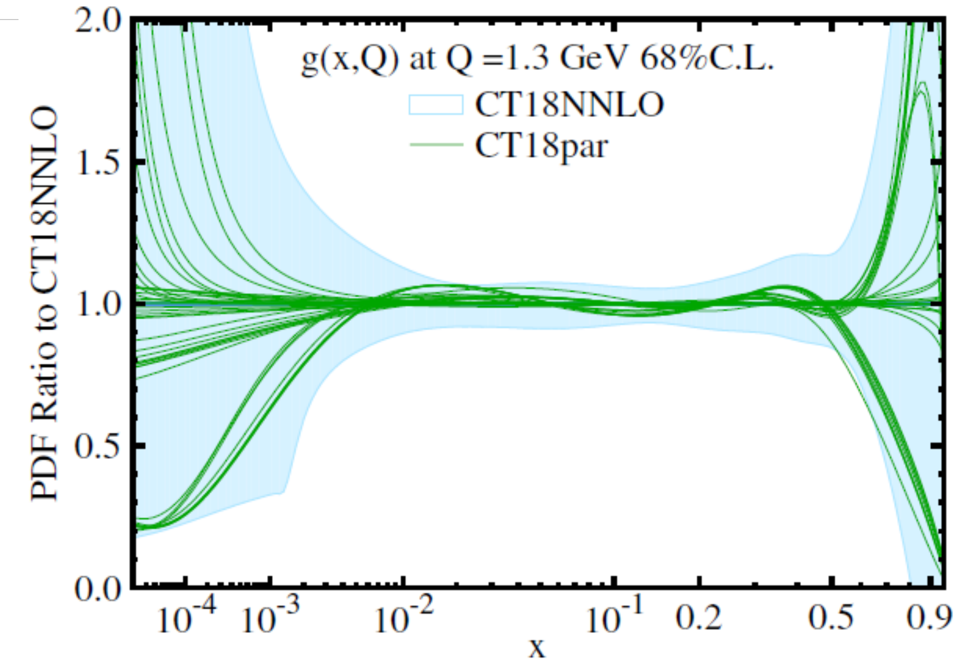
Hypothesis testing of theoretical predictions relies on

1. available data in x range, as well as value of Q,
2. sensitivity of data to the hypothesis,
3. quality of the data,
4. uncertainties found in the fits.



from Nadolsky [DIS2023]
*& [**AC,** Huston, Nadolsky, Xie, Yan & Yuan, Phys.Rev.D 107]*
& more of the CT coll. in preparation.

Hypothesis testing: role of uncertainties

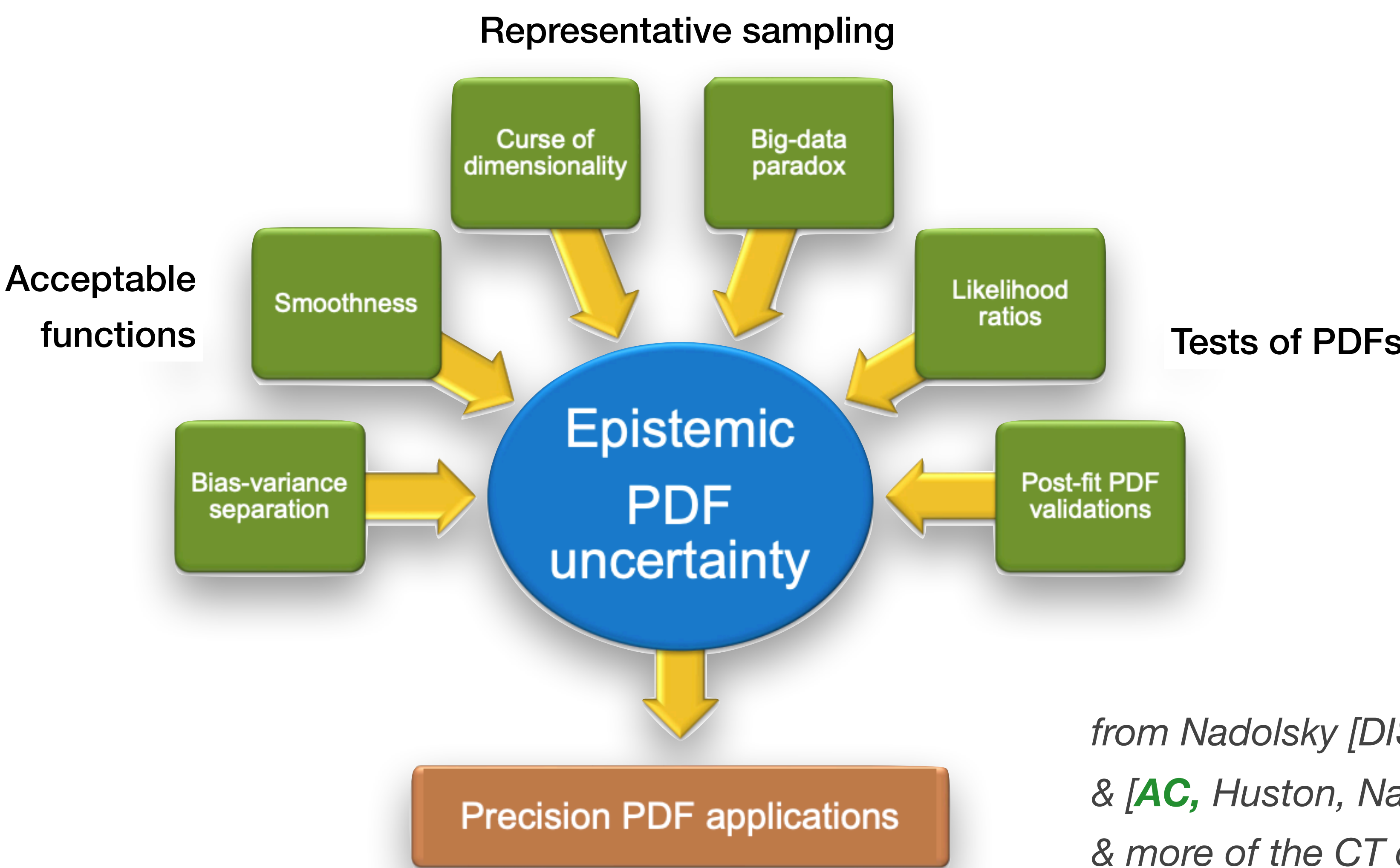


CT18 PDF uncertainty:

*Accounts for the **sampling over 250-350 parametrization** forms and possible choices of fitted experiments and fitting parameters.*

Hypothesis testing of theoretical predictions relies on

1. available data in x range, as well as value of Q,
2. sensitivity of data to the hypothesis,
3. quality of the data,
4. uncertainties found in the fits.

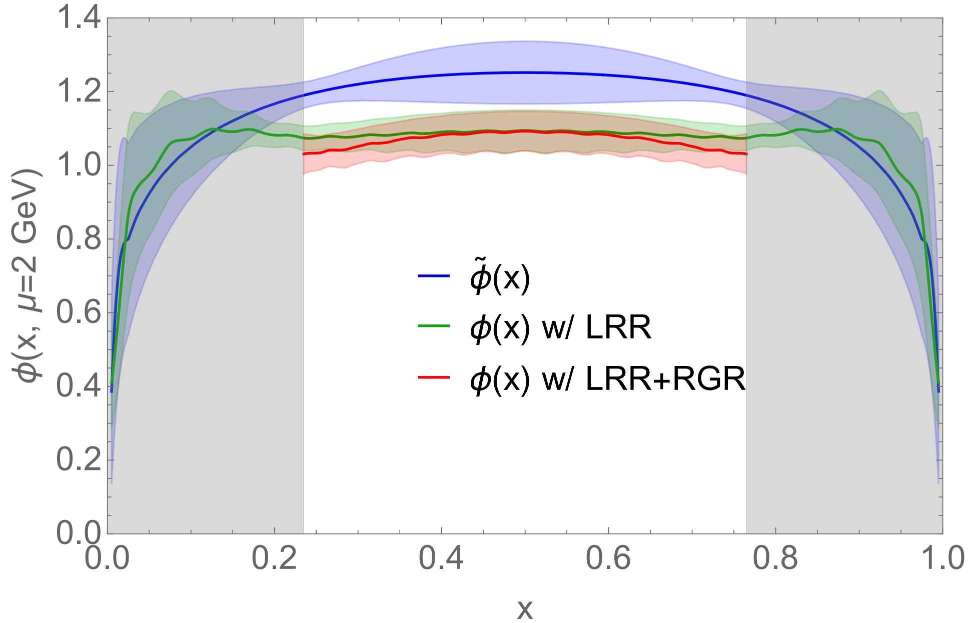


from Nadolsky [DIS2023]

*& [**AC,** Huston, Nadolsky, Xie, Yan & Yuan, Phys.Rev.D 107]*

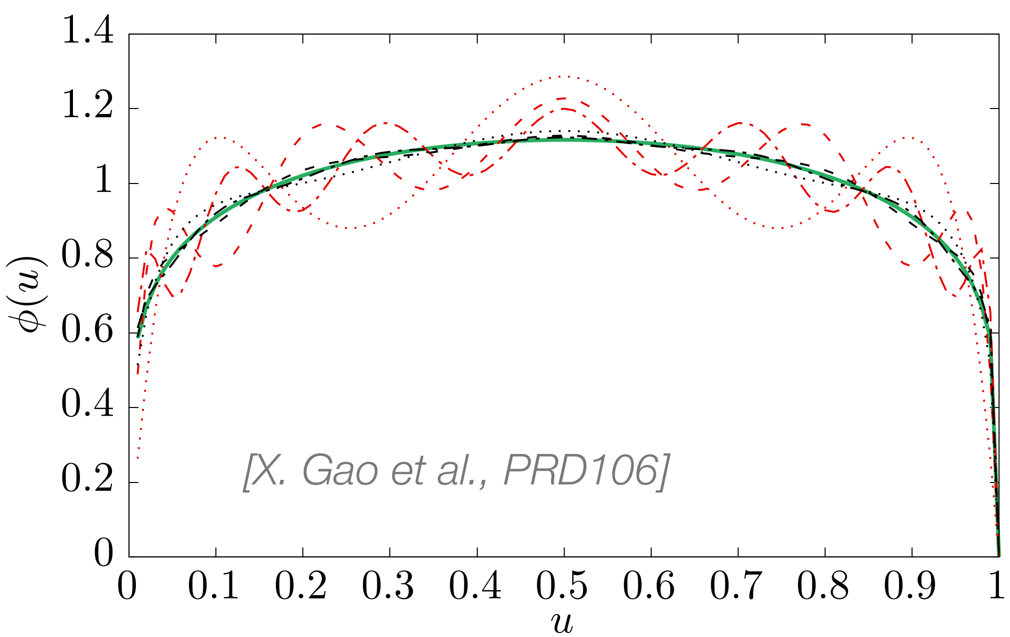
& more of the CT coll. in preparation.

Pion objects and the inverse problem

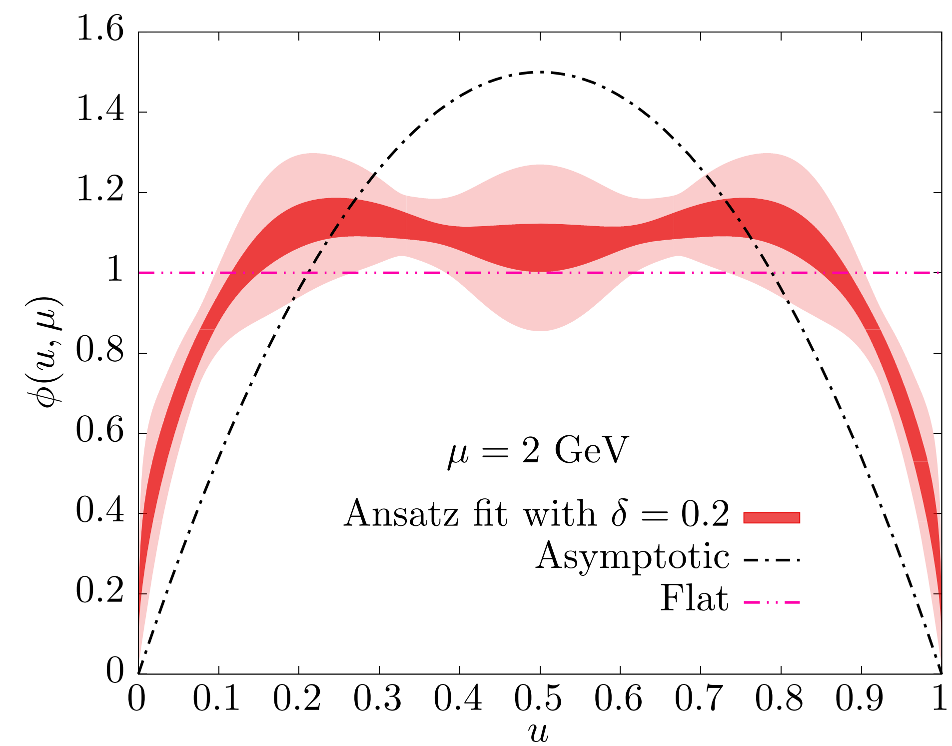


LaMET gives results at $Q = 2$ GeV for $0.23 \lesssim x \lesssim 0.77$
End-point behavior imposed (gray areas).
Very flat DA otherwise…

[Holligan et al., 2301.10372]



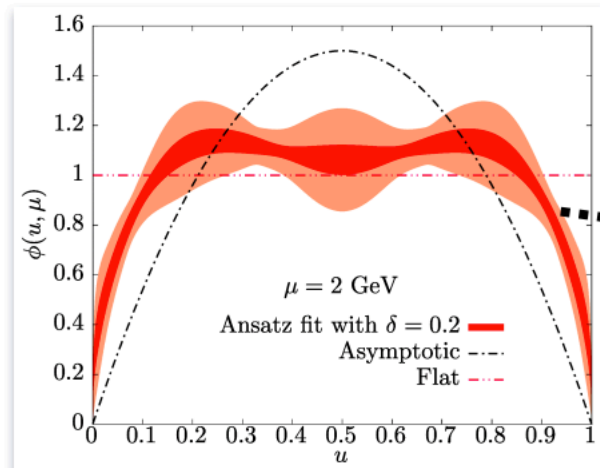
Gegenbauer expansion imposed to reconstruct DA from 2nd moment



Wiggly and flat DA at $Q = 2$ GeV.
Second moment at $Q = 2$ GeV similar to NJL at Q_0 !

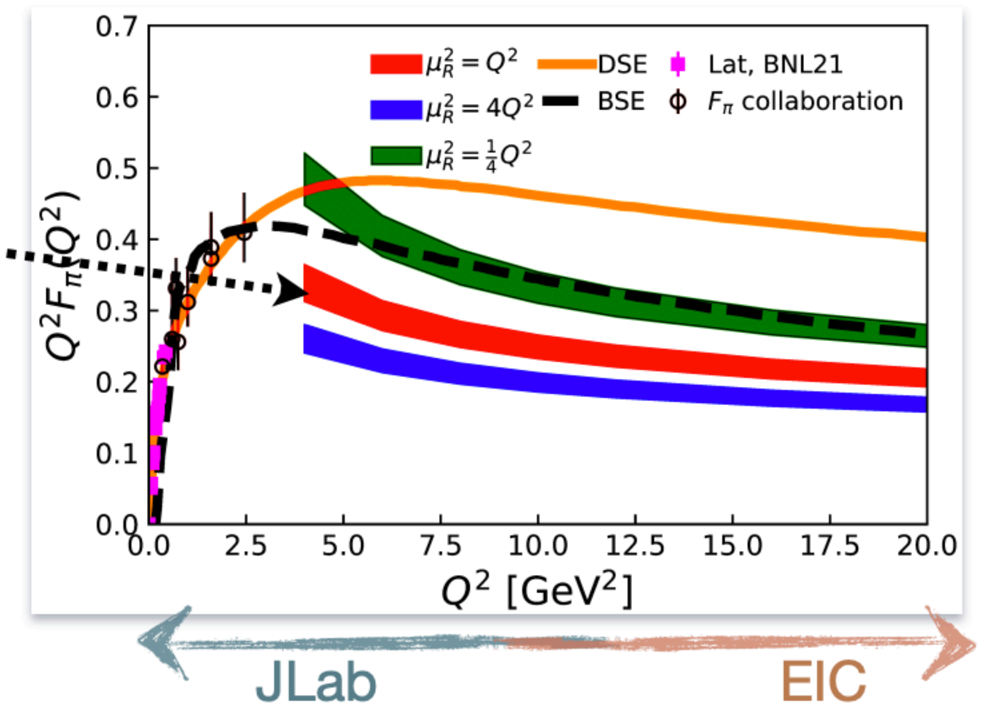
Pion objects and the inverse problem: convergence at end point.

Chiral symmetry seems to control the pion DA well over the Q spectrum.

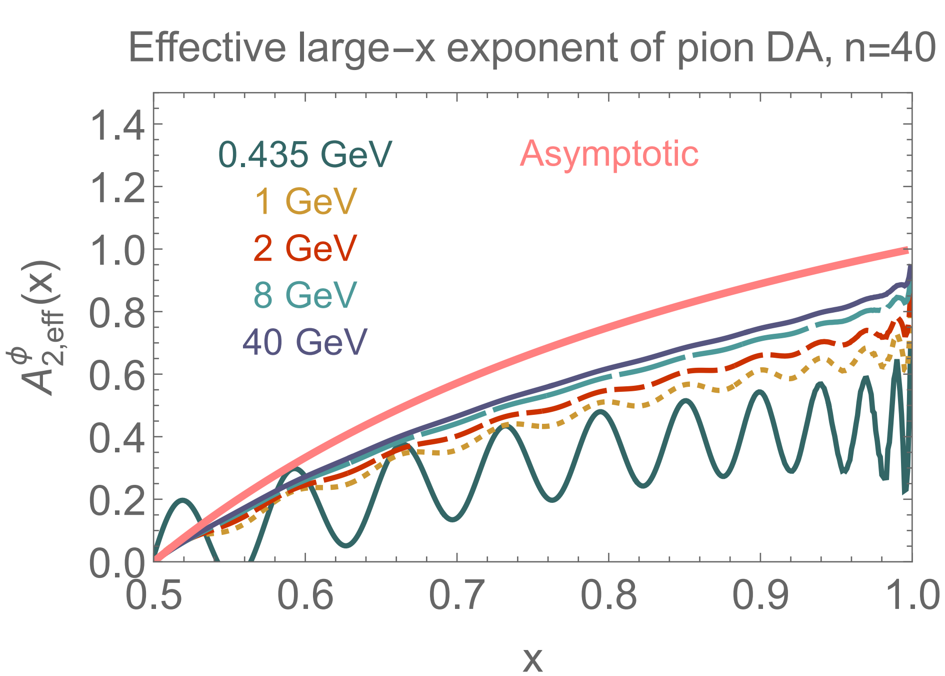


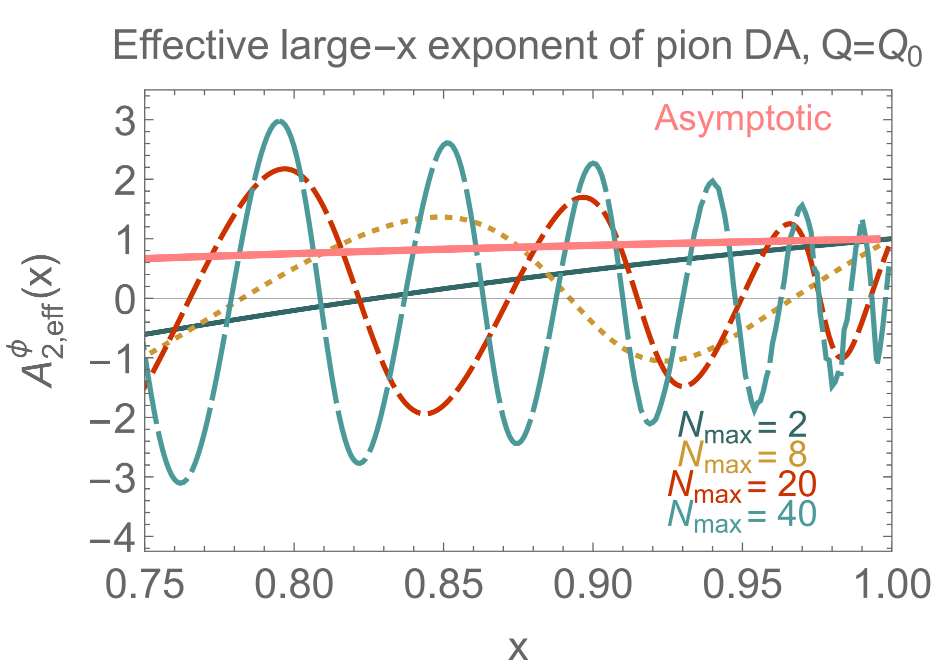
[X. Gao et al., PRD106]

& Swagato Mukherjee at DIS23.



Large-x convergence of the evolved pion DA seems to be key to problem.





*[**AC** & students, in progress]*

D. Navarro, undergrad thesis