Classification and combination of PDFs using polynomial approximators

Aurore Courtoy

Instituto de Física
National Autonomous University of Mexico (UNAM)

PDF4LHC meeting 2024

CERN

03/12/24

In collaboration with
L. Kotz, P. Nadolsky, F. Olness, M. Ponce-Chavez
and the CTEQ-TEA collaboration

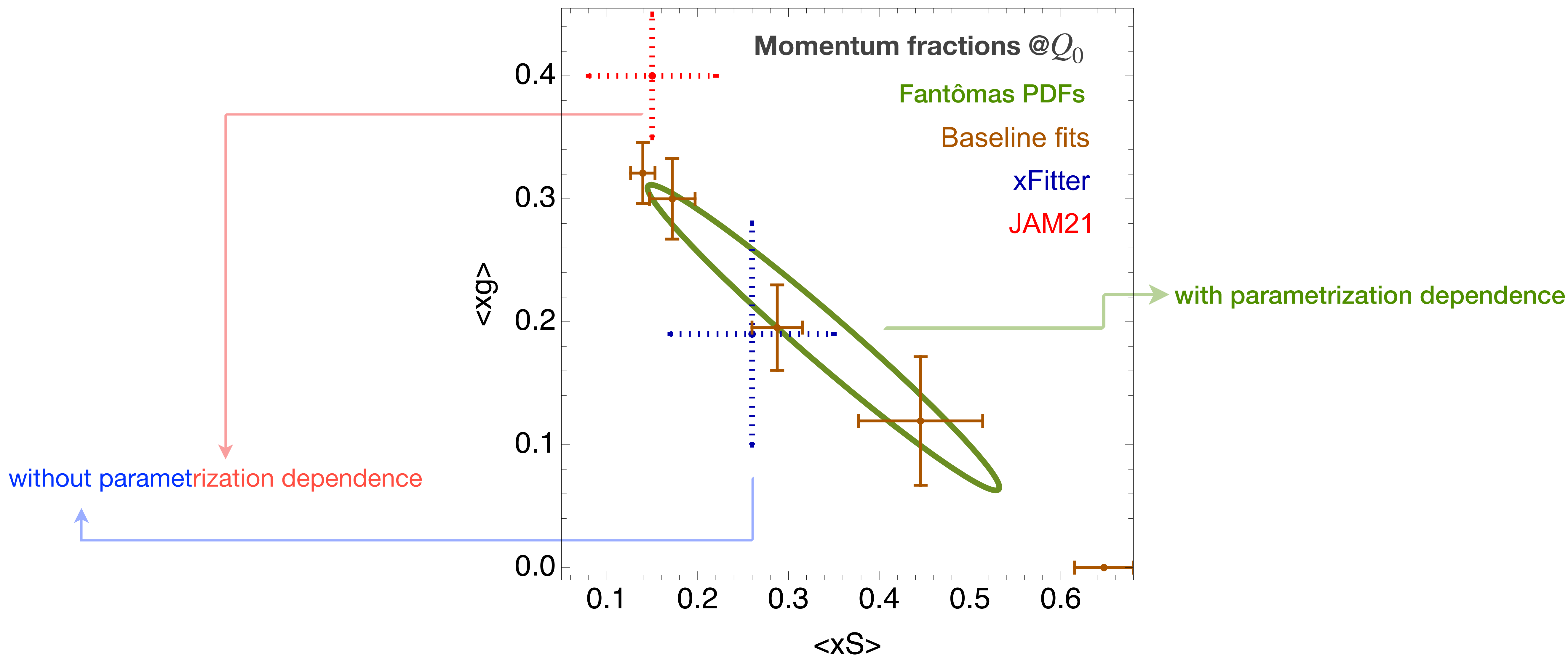
Dirección General de Asuntos
del Personal Académico

Conahcyt

Instituto de Física
UNAM

Fantômas charged-pion PDFs

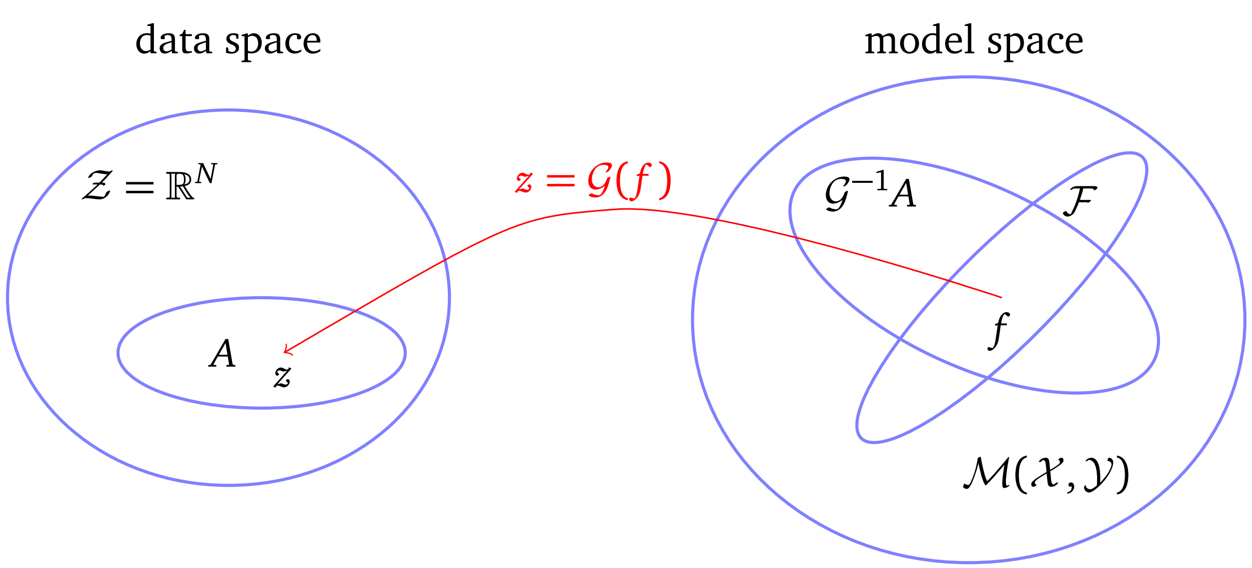
The first physics application of the Bézier curve-based fitting methodology



[Kotz, AC, Nadolsky, Olness, Ponce-Chavez, PRD109]

Global QCD analysis is an inverse problem

Parton Distribution Functions: are determined from data through solving an inverse problem.



del Debbio, SciPost Phys. Proc. 15, 028 (2024)

- ⇨ data as functional $\mathscr{G}$ of a model f
- ⇨ f represents the underlying truth, but is not uniquely determined by current data

Two main approaches to model f in global analyses:

- use an explicit parametrization
- use neural networks

Polynomial approximators enhance the usage of explicit parametrizations

Universality

Just like neural networks, these polynomial functional forms can represent any arbitrary PDF shape.

Interpretability

The shape of PDFs is controlled by PDF values at specific x (*control points*) and asymptotic limits ($x \to 0{,}1$), reframing the role of parameters. It allows for more stability in the optimization.

Controllable framework

They provide a controllable framework, including features like invariance under the initialization of higher degree polynomials.

Representative sampling

They facilitate exploration of the model space from the perspective of parametrization choice.

Separation of independent uncertainty contributions

By isolating uncertainty contributions from parametrization and other priors, these forms facilitate the use of information criteria.

Generation of parametrizations

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

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

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

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

Generation of parametrizations

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

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

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

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 $\{x_{\mathrm{CP}}\}$

AC & Nadolsky, Phys.Rev.D103 (2021)

Kotz, AC, Nadolsky, Olness & Ponce-Chavez, Phys.Rev.D109 (2024)

Generation of parametrizations

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

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

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

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 x^l matrix of binomial coefficients vector of coefficients c_l

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 $\{x_{\rm CP}\}$

AC & Nadolsky, Phys.Rev.D103 (2021)

Kotz, AC, Nadolsky, Olness & Ponce-Chavez, Phys.Rev.D109 (2024)

PDF shape:

⇨ asymptotics ensured by a *carrier function*

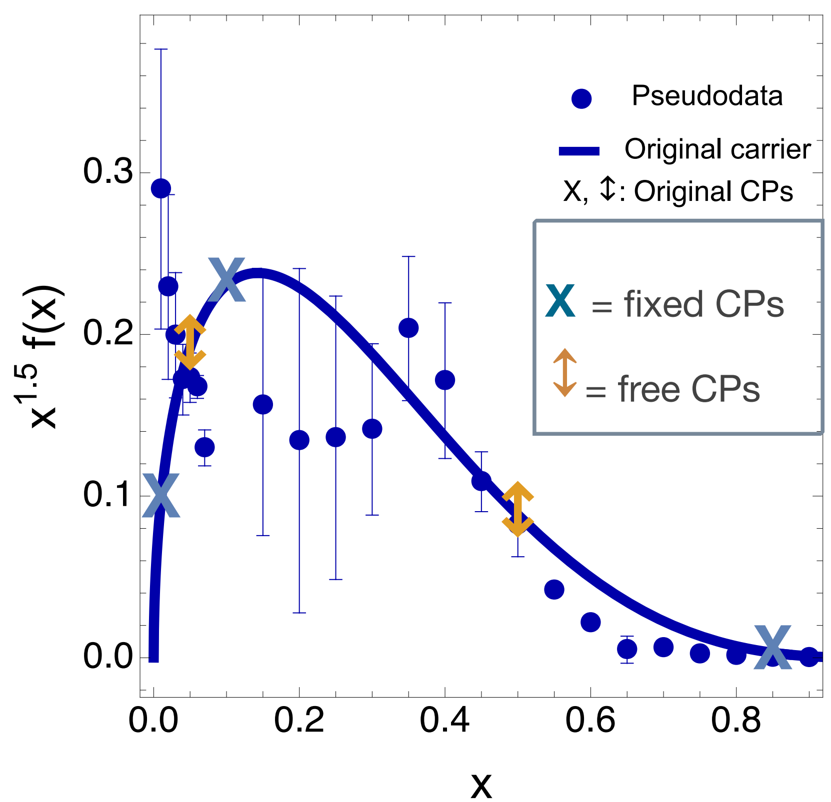
⇨ sum rules imposed through normalization

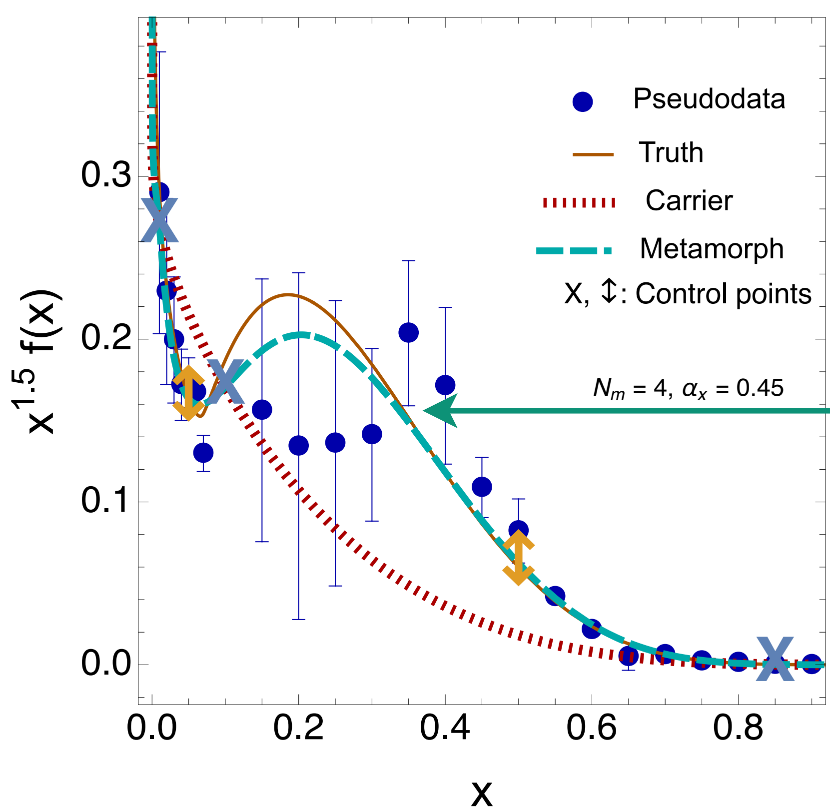
for q = PDF type (flavor, combination or gluon)

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

$\underline{v} = \{\underline{C}, \underline{P}\}$

Bézier-curve methodology— toy model





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

δB_q & δC_q allow the carrier to vary

N_m= degree of polynomial can vary

α_x can vary

`metamorph` fit:

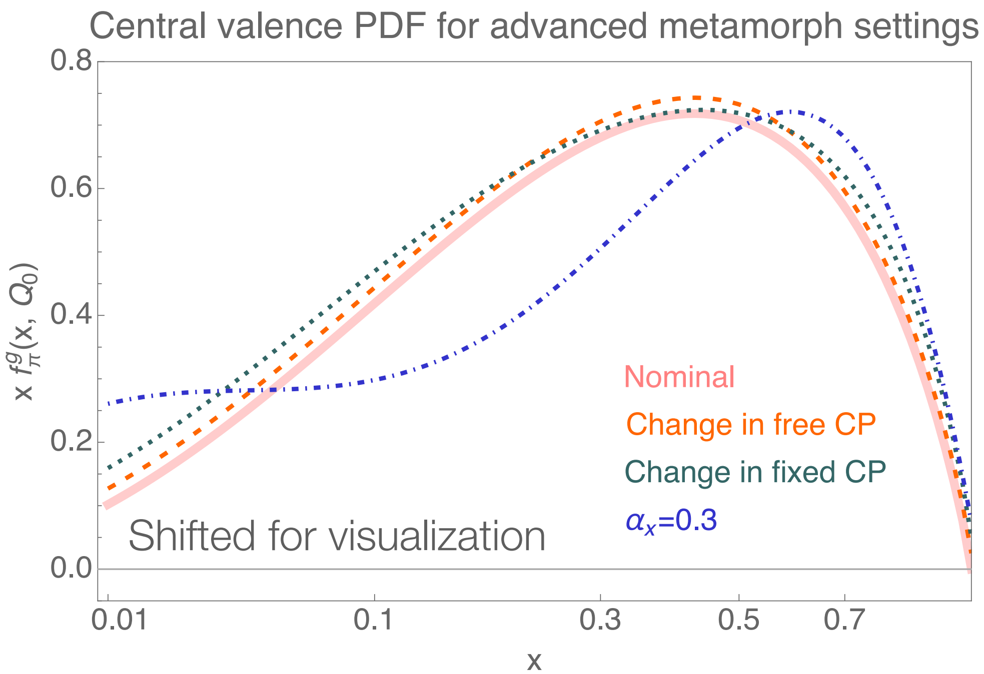
$$x\, q(x, Q_0^2) = A'_q\, x^{B_q} (1-x)^{C_q} \times \left(1 + \mathcal{B}^{(N_m)}(x^{\alpha_x}, Q_0^2; \underline{v})\right)$$

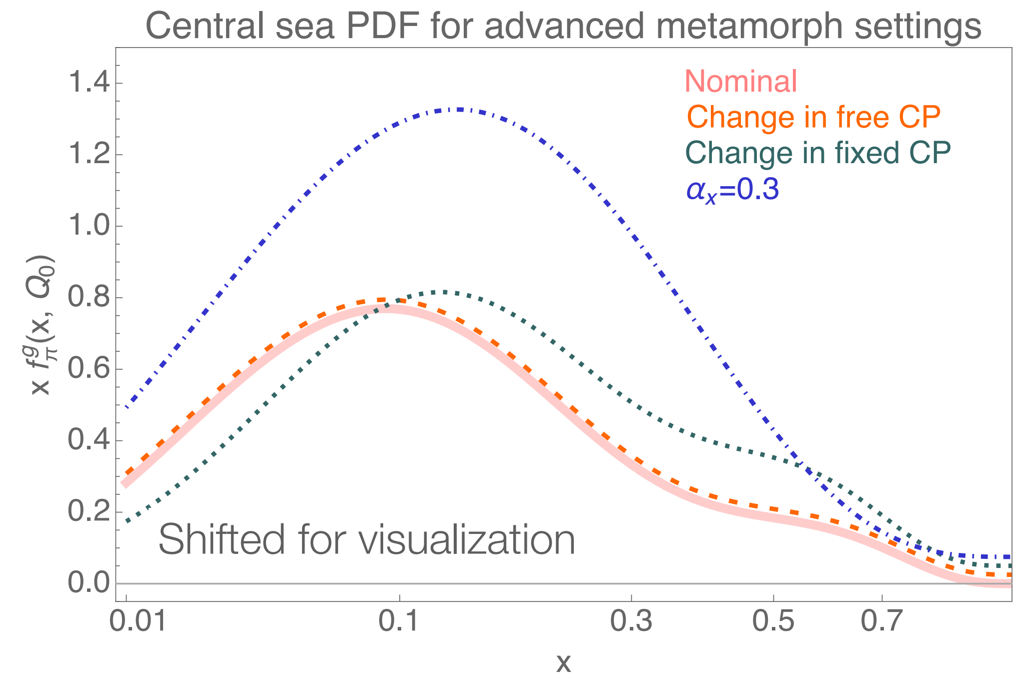
Unisolvent systems for N_m= # CPs-1.

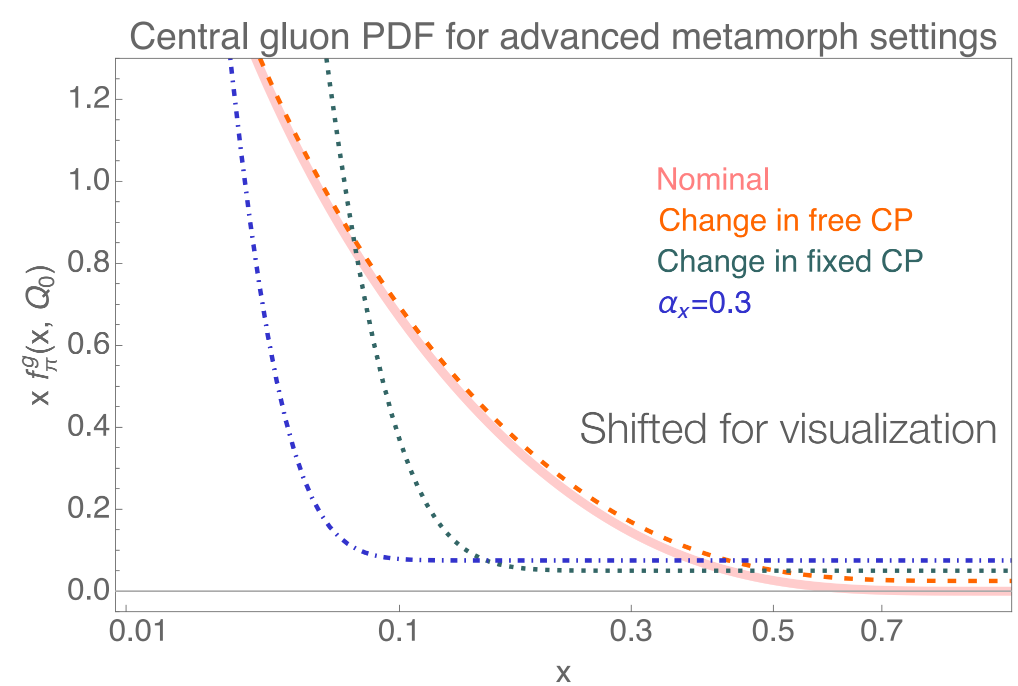
Code to be released soon!

Fantômas unleashed

Kotz, AC, Hobbs, Nadolsky, Olness ,
Ponce-Chavez & Purohit, 2412.XXXXX)







Three features:

invariance under change of free control points

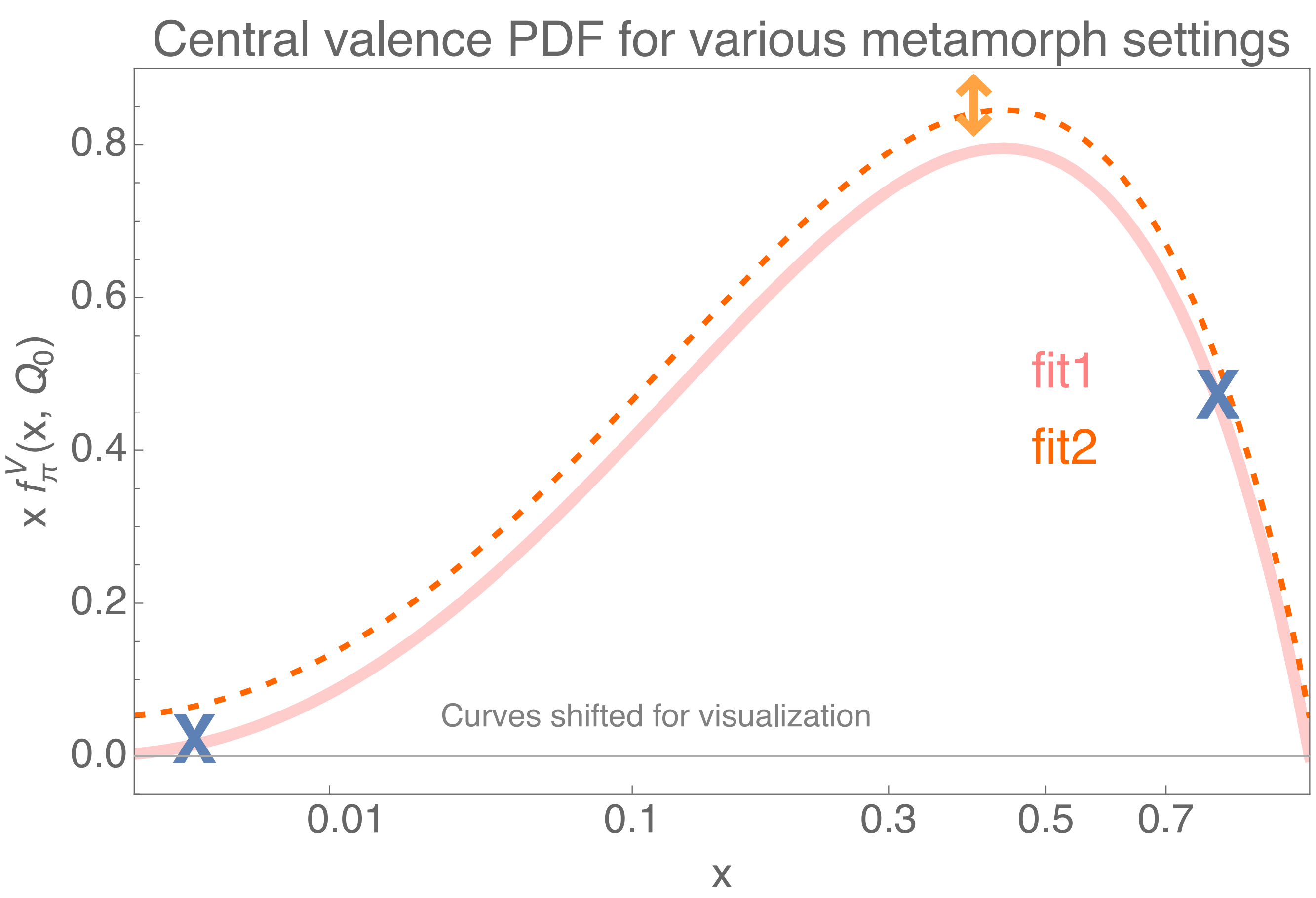
variation of the metamorph for change of fixed control points

variation of the metamorph for change of stretching exponent

Reparametrization invariance

The `metamorph`'s polynomial degree can be increased initially without modifying the shape of the PDF.

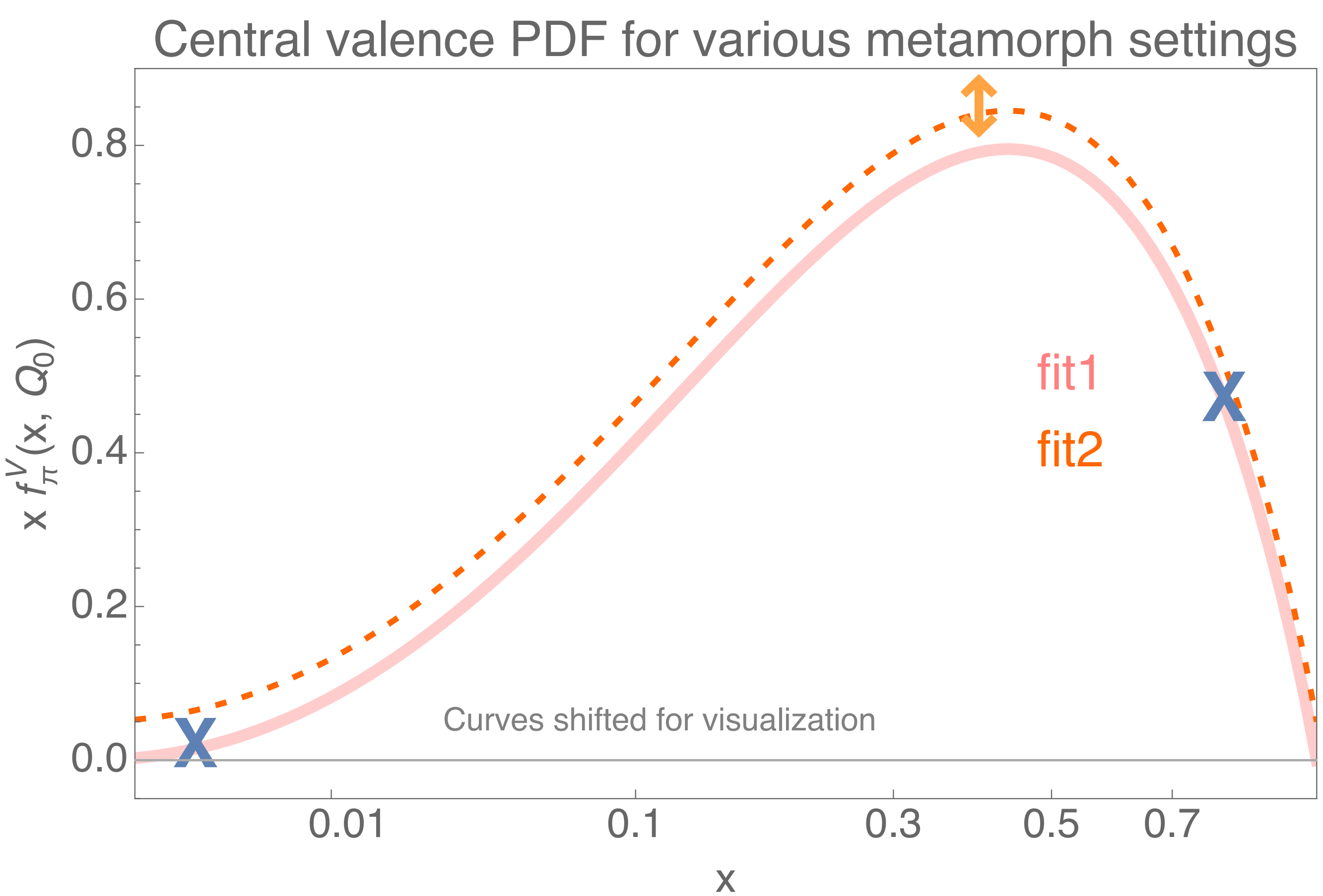
[AC & Nadolsky, PRD103]



Reparametrization invariance

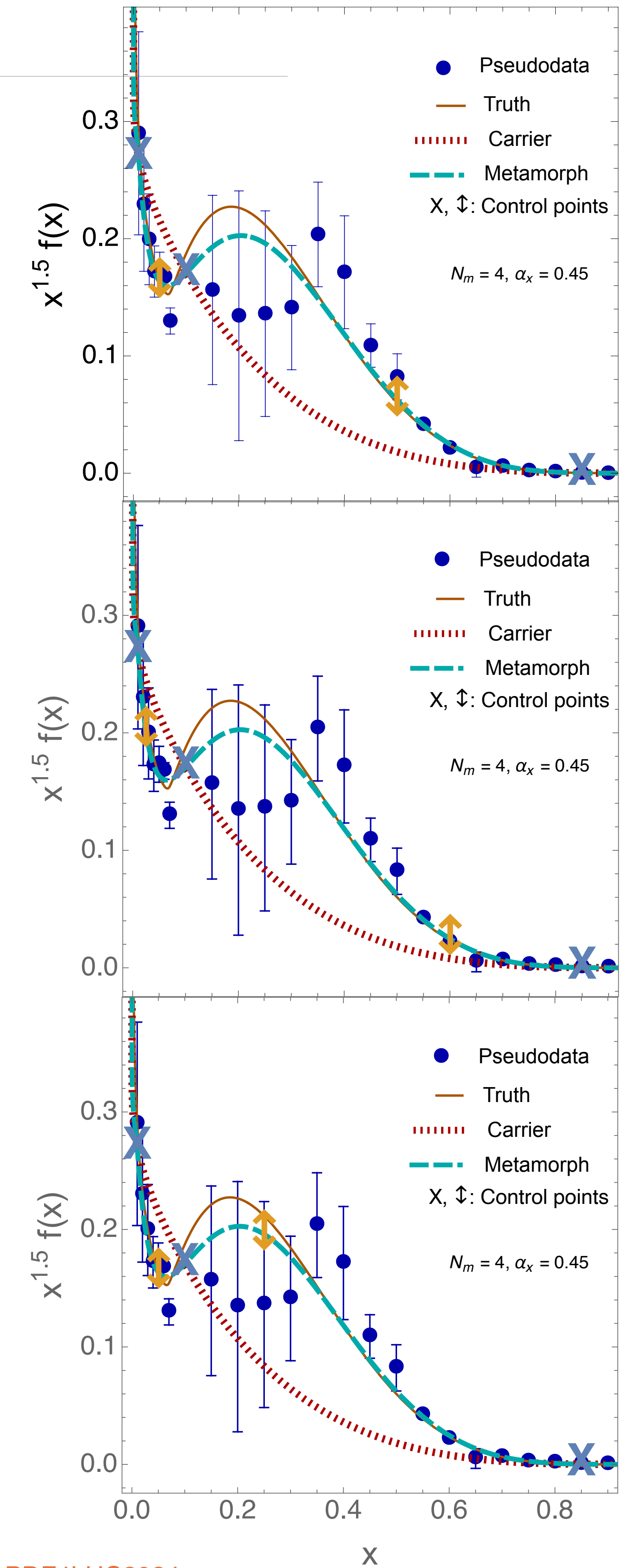
The `metamorph`'s polynomial degree can be increased initially without modifying the shape of the PDF.

[AC & Nadolsky, PRD103]



X = fixed CPs

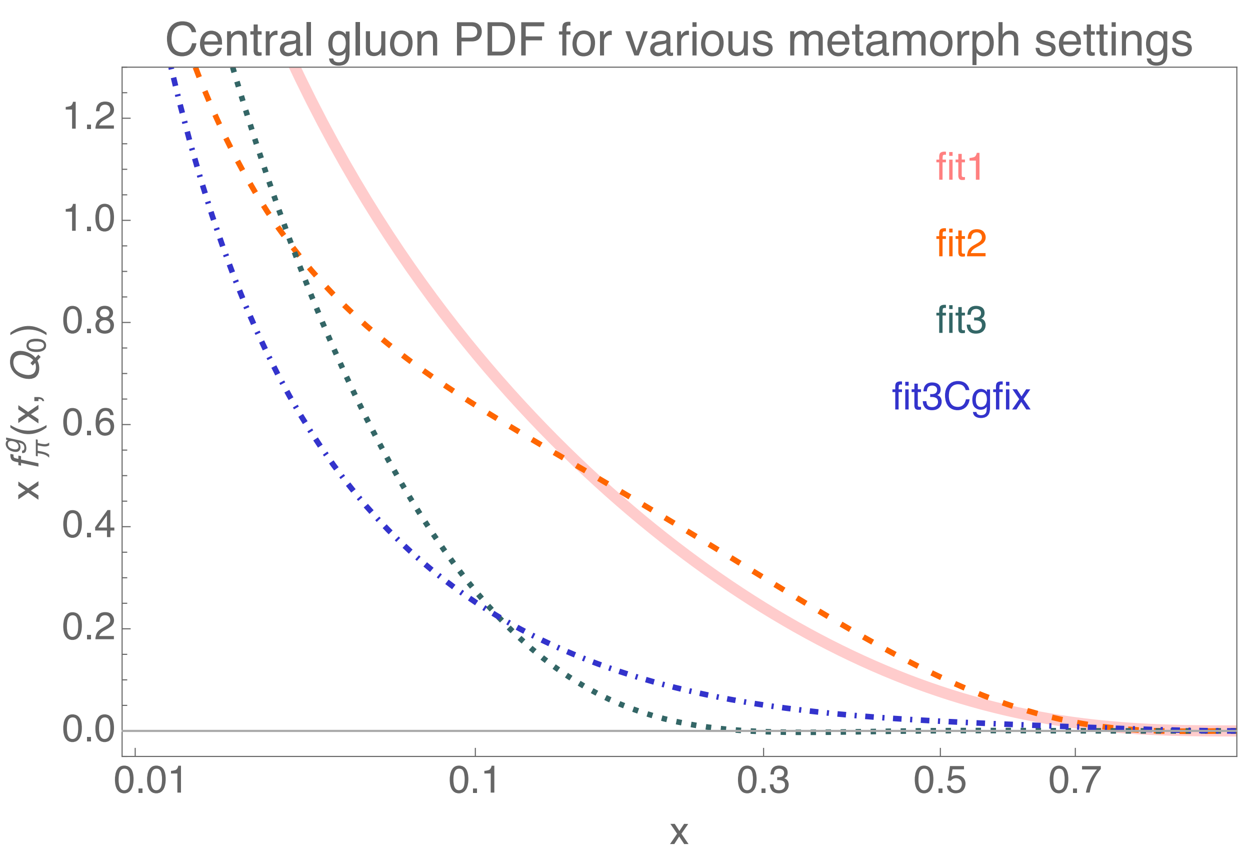
↕ = free CPs

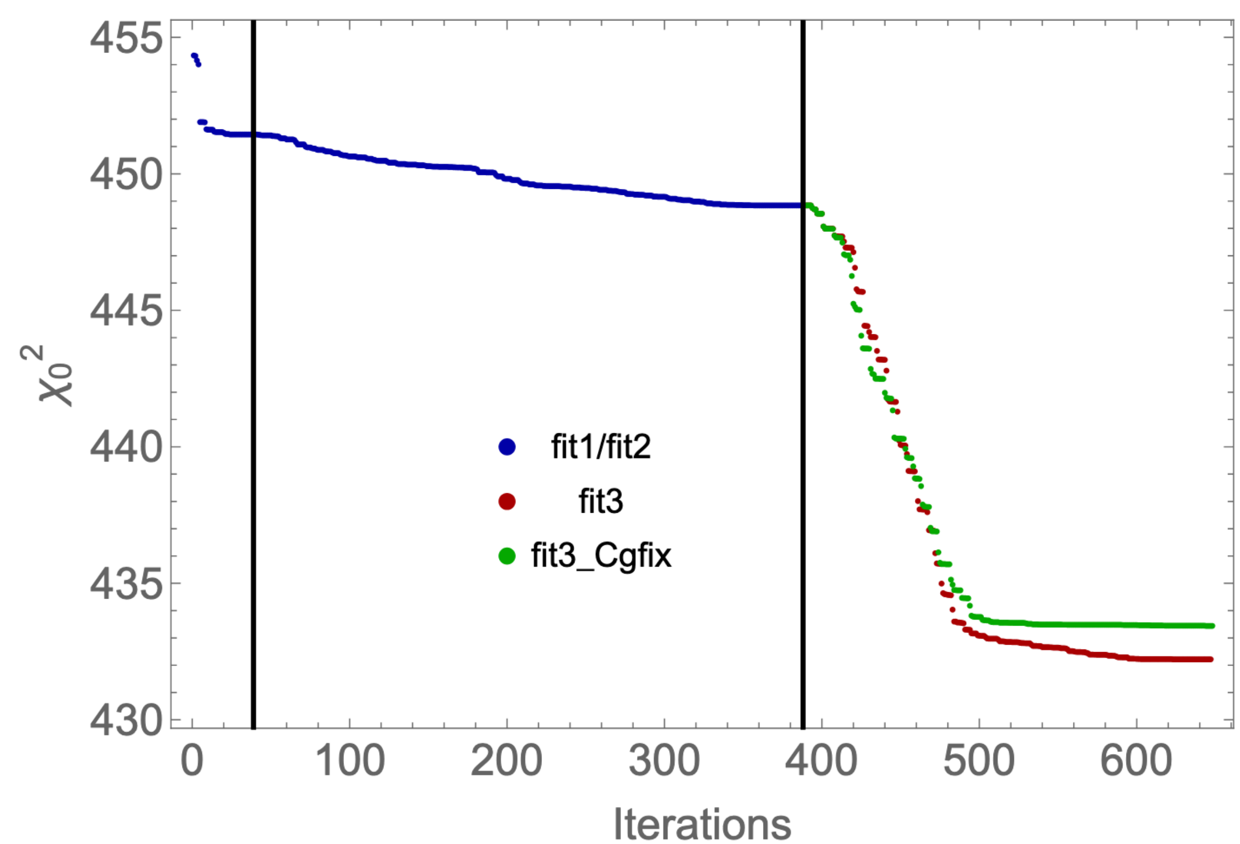


The fixed CPs are the intersection points between the carrier and the final `metamorph`. The fixed CPs set the shape of the curve ; the free CPs act through the minimization procedure.

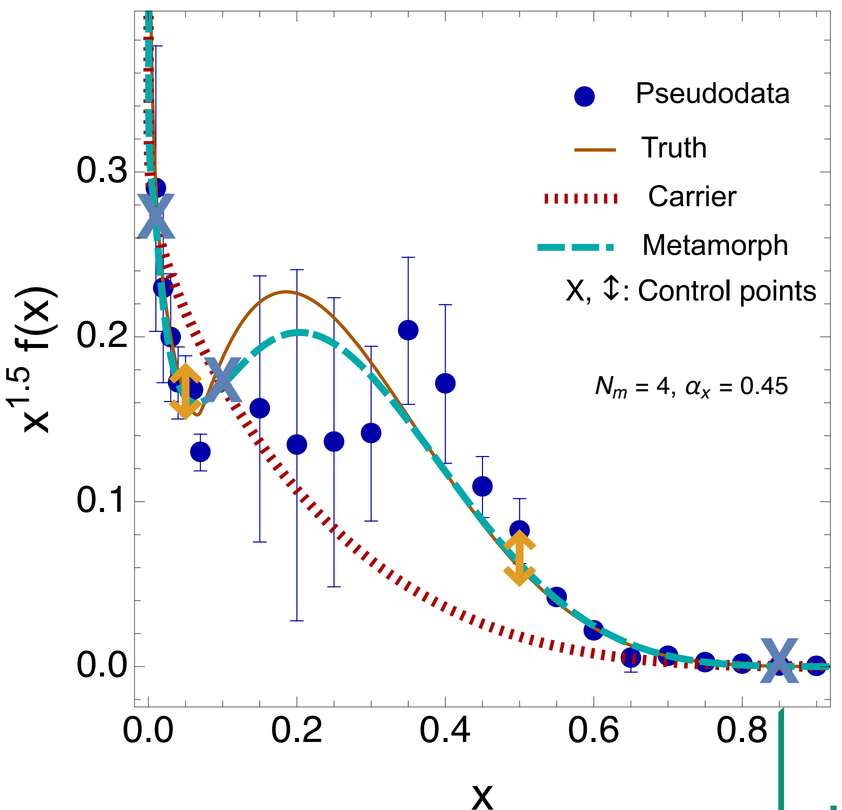
Stability of the framework

Starting with a low polynomial degree, we can add free control points one by one and check the convergence of the minimization procedure.
Incremental addition of CPs can only decrease the chisquare from the previous step.



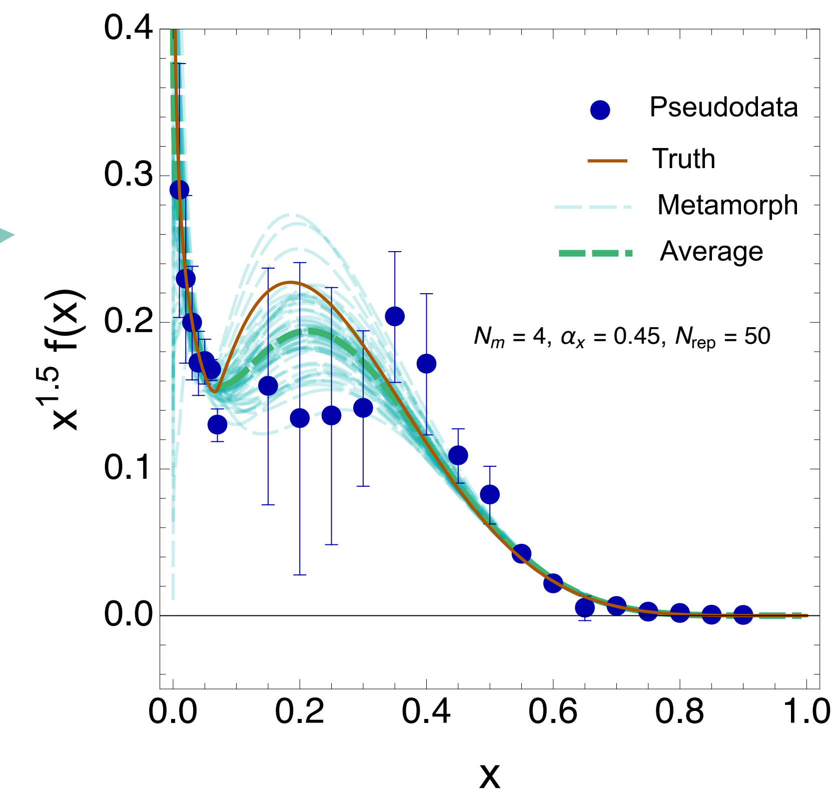


Bézier-curve methodology— toy model



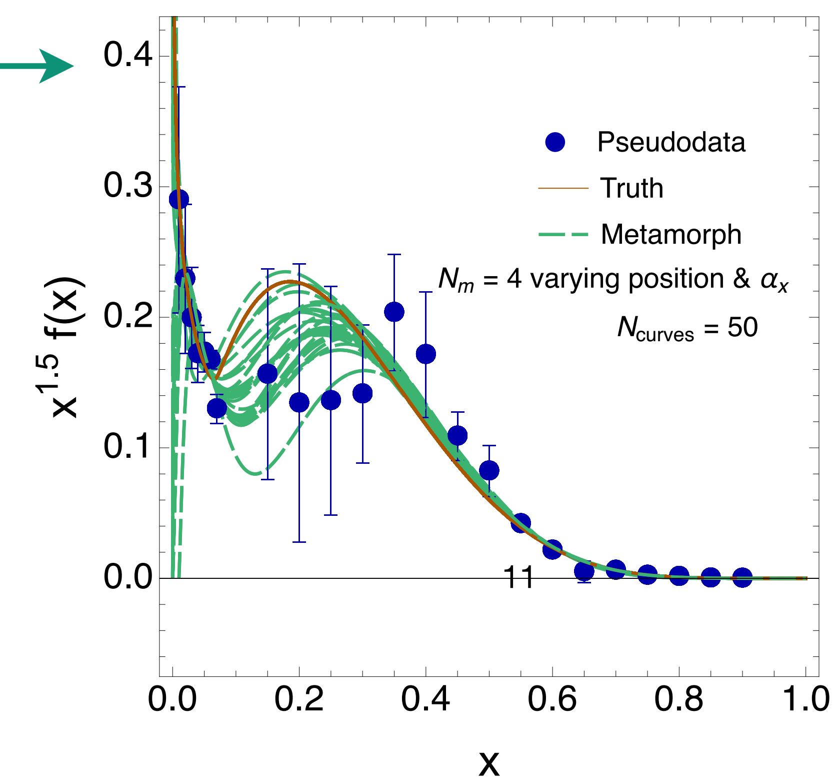
if bootstrapped

sampling on the distribution of data uncertainties



if sampled over `metamorph` settings

sampling over parametrizations



Both the statistic (e.g. bootstrap) and the systematic (e.g. `metamorph` sampling) uncertainties should be accounted for.

[AC et al, PRD107, 2205.10444]

Classification

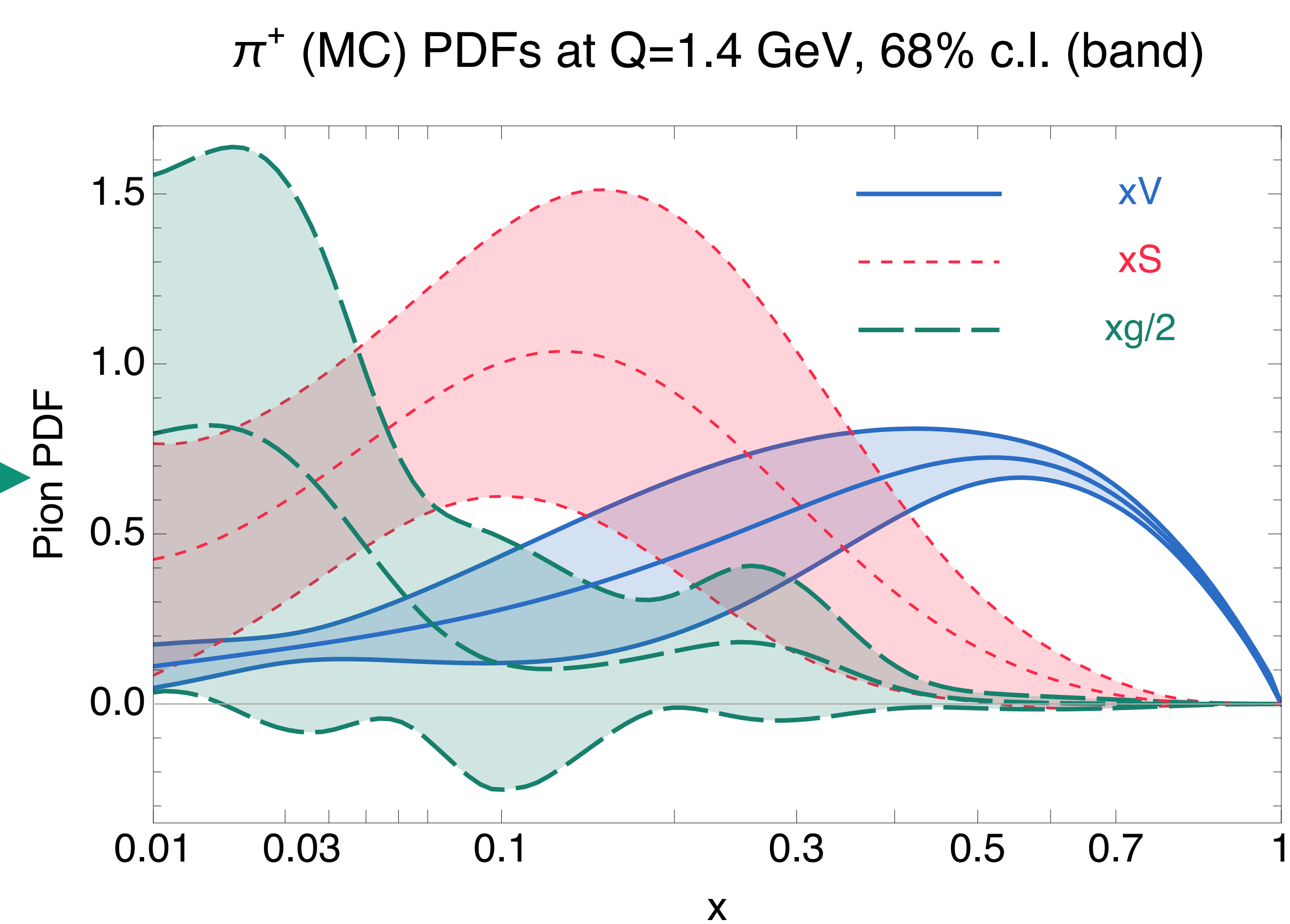
Fantômas πPDFs

⇨ We generated $N \sim 100$ fits corresponding to N sets for $\{N_m, \underline{P}, \alpha_x\}$.

⇨ Well-behaved (convergence + fixed soft constraints) fits are kept.

⇨ Fits within $\chi^2 + \delta\chi^2 = \chi^2 + \sqrt{2(N_{\mathrm{pts}} - N_{\mathrm{par}})}$ are kept.

The final bundle is generated from the 5 most diverse shapes at Q_0.



In progress: automatize the selection based on shapes [UNAM's group] and use of information criteria — likelihood-ratio test and quantititative criteria [see K.. Mohan's talk]

Likelihood-ratio test

Independent contributions to uncertainty:
the parametrization contributes to the (log)-likelihood but constraints on the parameters, …, contribute to the prior.

$$\chi^2_{\rm tot} = \chi^2 + \chi^2_{\rm prior}$$

$$P(a|D) \propto P(D|a)\, P(a)$$

$$\Leftrightarrow \quad \exp(-\chi^2_{\rm tot}) \propto \exp(-\chi^2) \exp(-\chi^2_{\rm prior})$$

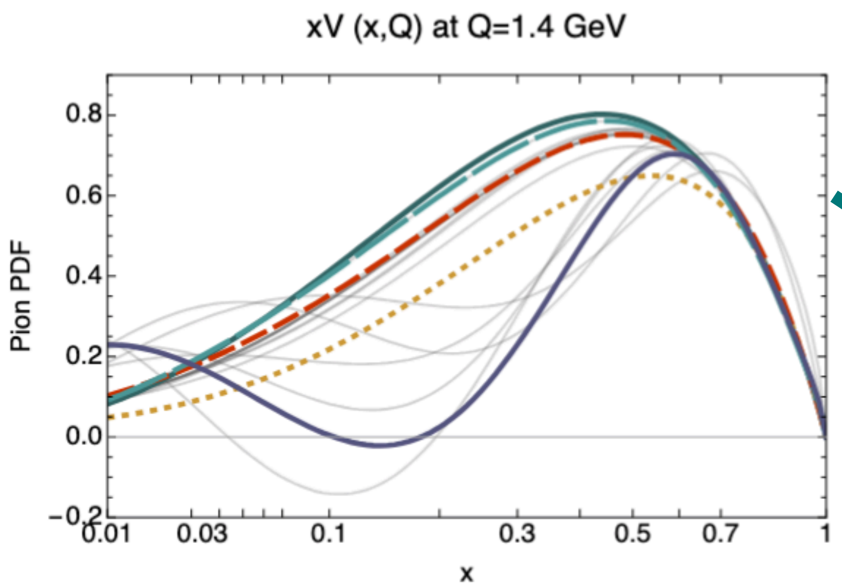
$\boxed{\text{On which basis are PDFs accepted or rejected?}}$

Likelihood ratios:
two replicas can be ordered according to their relative likelihood or relative prior.

$$\underset{\equiv r_{\rm posterior}}{\frac{P(T_2|D)}{P(T_1|D)}} = \underset{\equiv r_{\rm likelihood}}{\frac{P(D|T_2)}{P(D|T_1)}} \times \underset{\equiv r_{\rm prior}}{\frac{P(T_2)}{P(T_1)}}$$

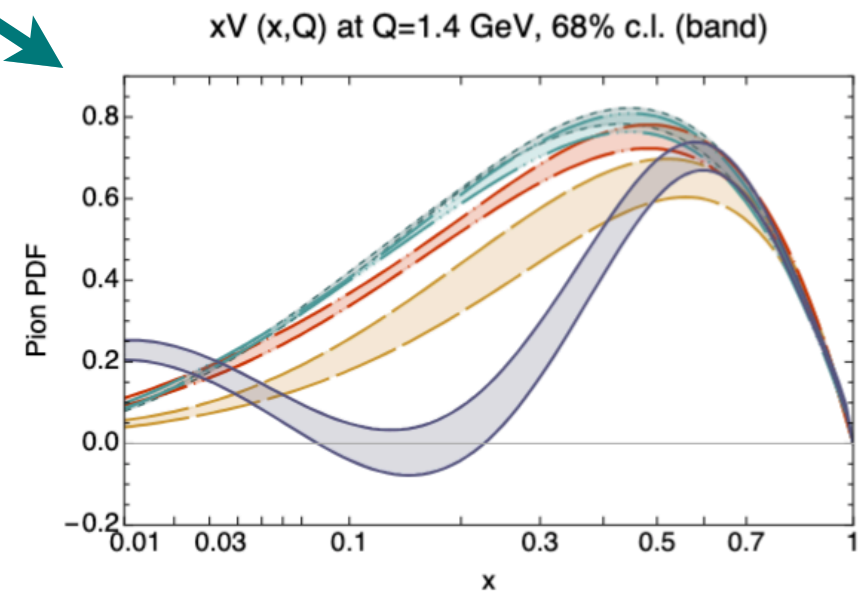
aleatory epistemic + aleatory probabilities

Combination

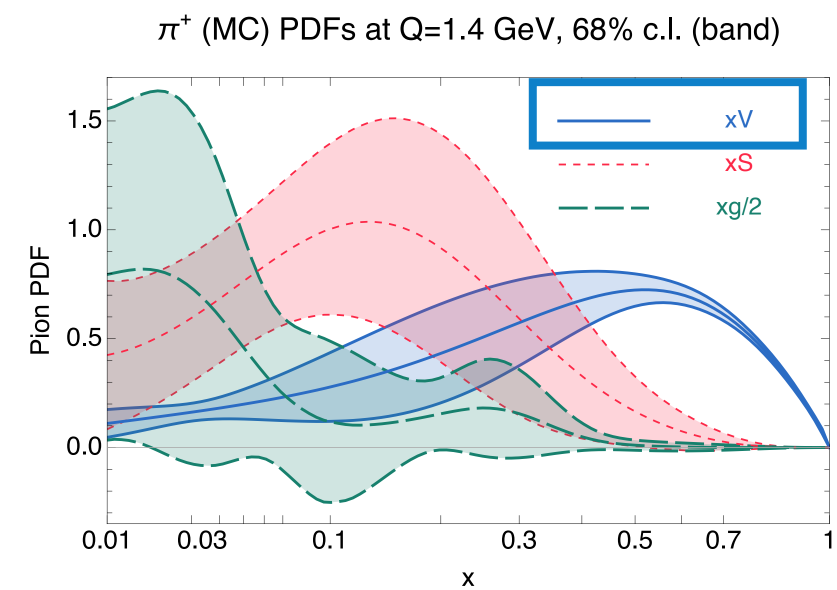


$\Delta\chi^2 = 1$ criterion

For π^+ PDFs,
$q = V = 2(u - \bar{u})$,

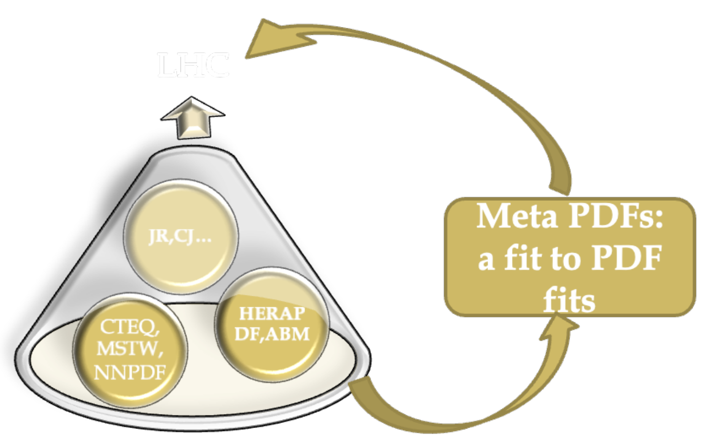


Bundled models



Models combined using METAPDF

Update on the `mp4lhc` and `mcgen` codes in the context of Fantômas
[Kotz et al, in progress]



[Gao & Nadolsky, JHEP07]

Tolerance

We (CT) are looking into information criteria to quantify the tolerance encompassing multiple sources of uncertainties.

Fantômas unlocks the concept of tolerance:

- multiple parametrizations with respective $\Delta\chi^2 = 1$ uncertainty can be bundled into a $\sim \Delta\chi^2 > 1$ error band.
- separation of constraints'contributions

Epistemic vs. aleatory uncertainties

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

Statistical uncertainty propagated from experiments— irreducible

metamorph routine in xFitter

Code to be released soon!

Kotz, AC, Hobbs, Nadolsky, Olness , Ponce-Chavez & Purohit, 2412.XXXXX)

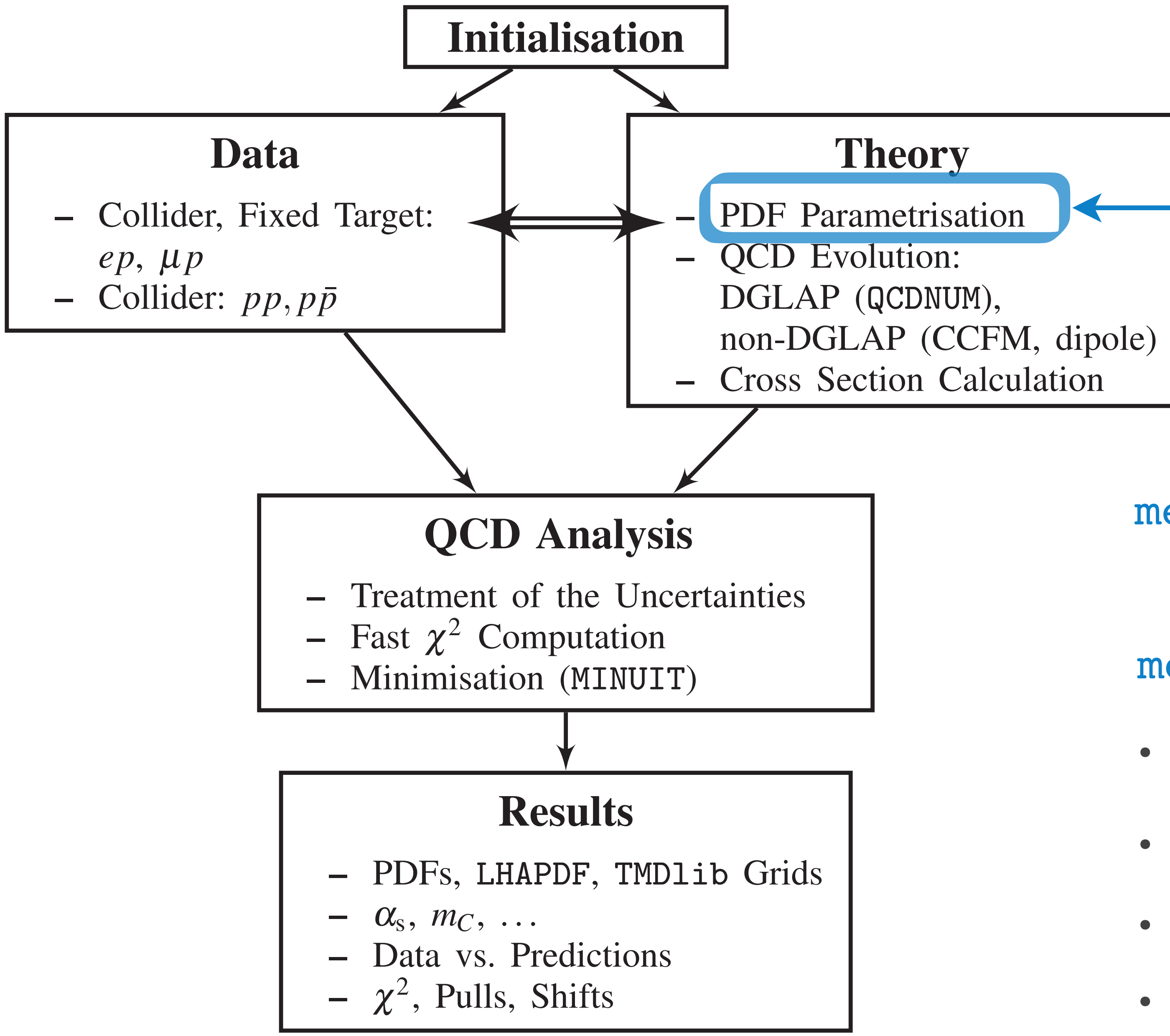


Figure 1: Schematic structure of the `xFitter` program.

`metamorph` routine — PhD thesis of L. Kotz (SMU)

`metamorph` requires inputs from the user:

- N_m — degree of polynomial
- $\{x, f_{\text{in}}(x)\}$ of control points
- fixed or free control points
- stretching parameter

Conclusions

Towards epistemic PDF uncertainties with Fantômas4QCD.

Towards augmenting the aleatory $\Delta\chi^2 = 1$ uncertainties with the uncertainty due to parametrization.

Bézier-curve methodology

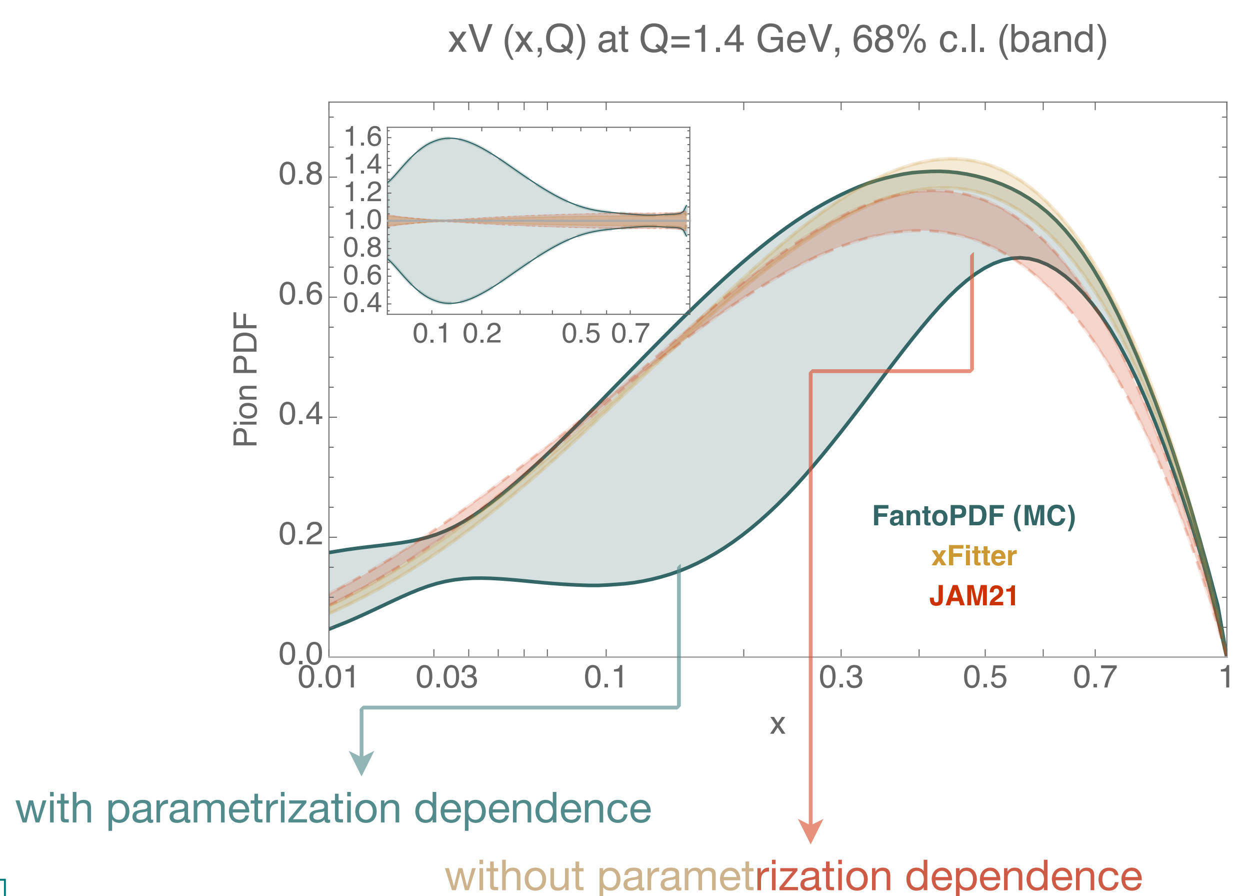
⇨ Universality

⇨ Interpretability

⇨ Controllable framework

⇨ Representative sampling

⇨ Separation of independent uncertainty contributions



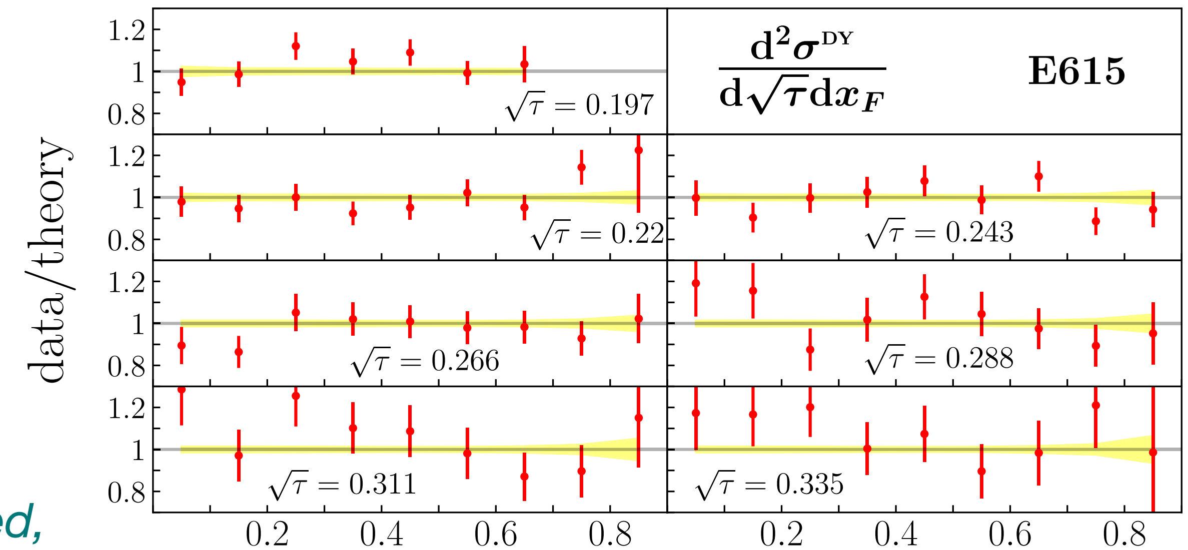
[Kotz, AC, Nadolsky, Olness, Ponce-Chavez, PRD109]
[AC, Hobbs, Kotz, Nadolsky, Olness, Ponce-Chavez, Purohit, soon]

Back up

Regression for data-based analyses

Global analyses involve searching for extrema of a (log-)likelihood function.

[JAM21]



(Very) simplified:

$$\chi^2 = \sum_i^{N_{\text{exp}}} \frac{\left(D_i - \langle T(\{\mathbf{x}, \mathbf{a}\})\rangle_i\right)^2}{\sigma_i^2} + \text{penalty terms}$$

discrete data point

theory prediction averaged,

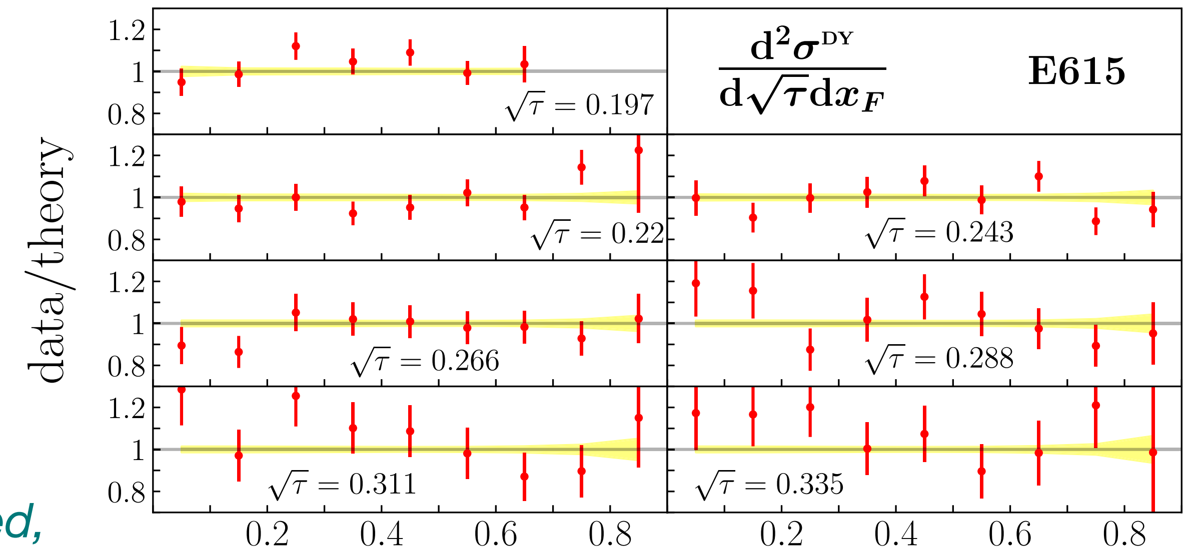
*as a function of the variables {**x**} and free parameters {**a**}*

The theory input depends on the PDFs, whose parametrization is an input to the minimization procedure. The comparison to data for various parametrizations can lead to equally good χ^2 values.

Regression for data-based analyses

Global analyses involve searching for extrema of a (log-)likelihood function.

[JAM21]



<u>(Very) simplified:</u>

$$\chi^2 = \sum_i^{N_{\rm exp}} \frac{\left(D_i - \langle T(\{\mathbf{x},\mathbf{a}\})\rangle_i\right)^2}{\sigma_i^2} + \text{penalty terms}$$

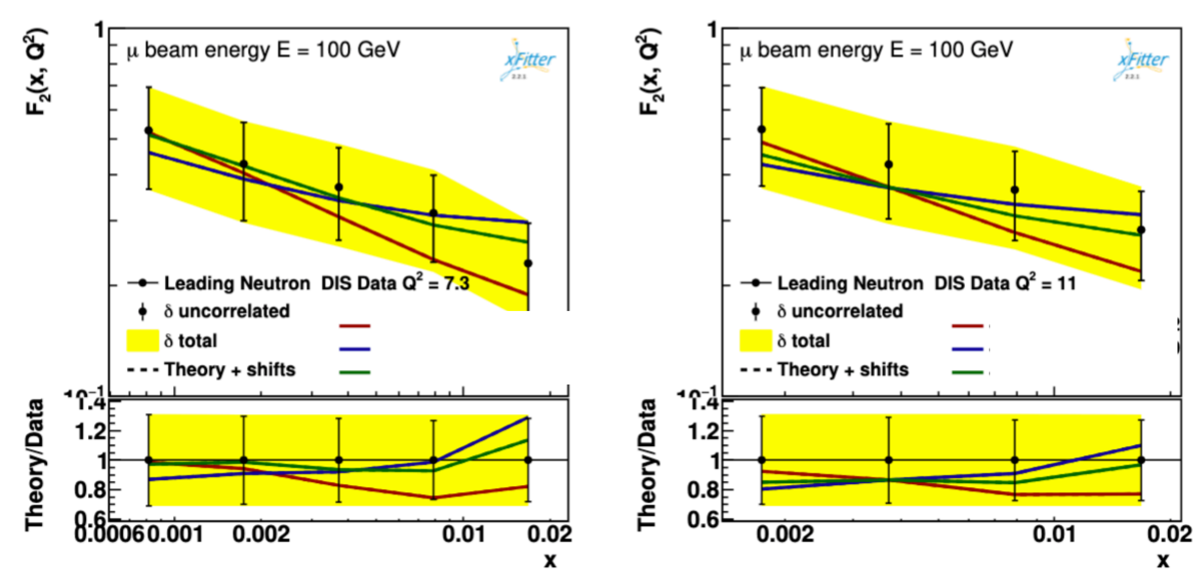
discrete data point

theory prediction averaged,

*as a function of the variables {**x**} and free parameters {**a**}*

The theory input depends on the PDFs, whose parametrization is an input to the minimization procedure.
The comparison to data for various parametrizations can lead to equally good χ^2 values.

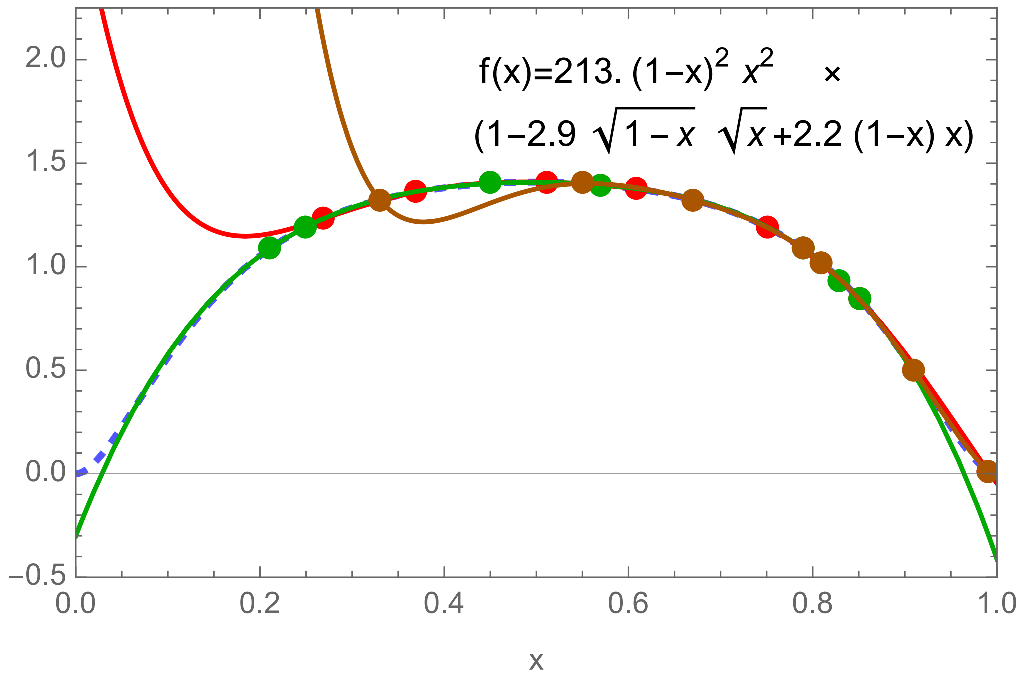
[Fantômas]



That's fine in the data region,
but the results may vary greatly outside
— extrapolation region.

Why not adopt more than one form?

Bézier-curve methodology for global analyses



The reconstructed function may depend on the position and number of control points.

Global analyses can exploit this property to generate many functional forms.

$\Rightarrow$ polynomial mimicry

Behaviour on top of asymptotics is embedded into a Bézier curve

⇨ asymptotics usually ensured by a *carrier function*

⇨ sum rules imposed through normalization

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

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

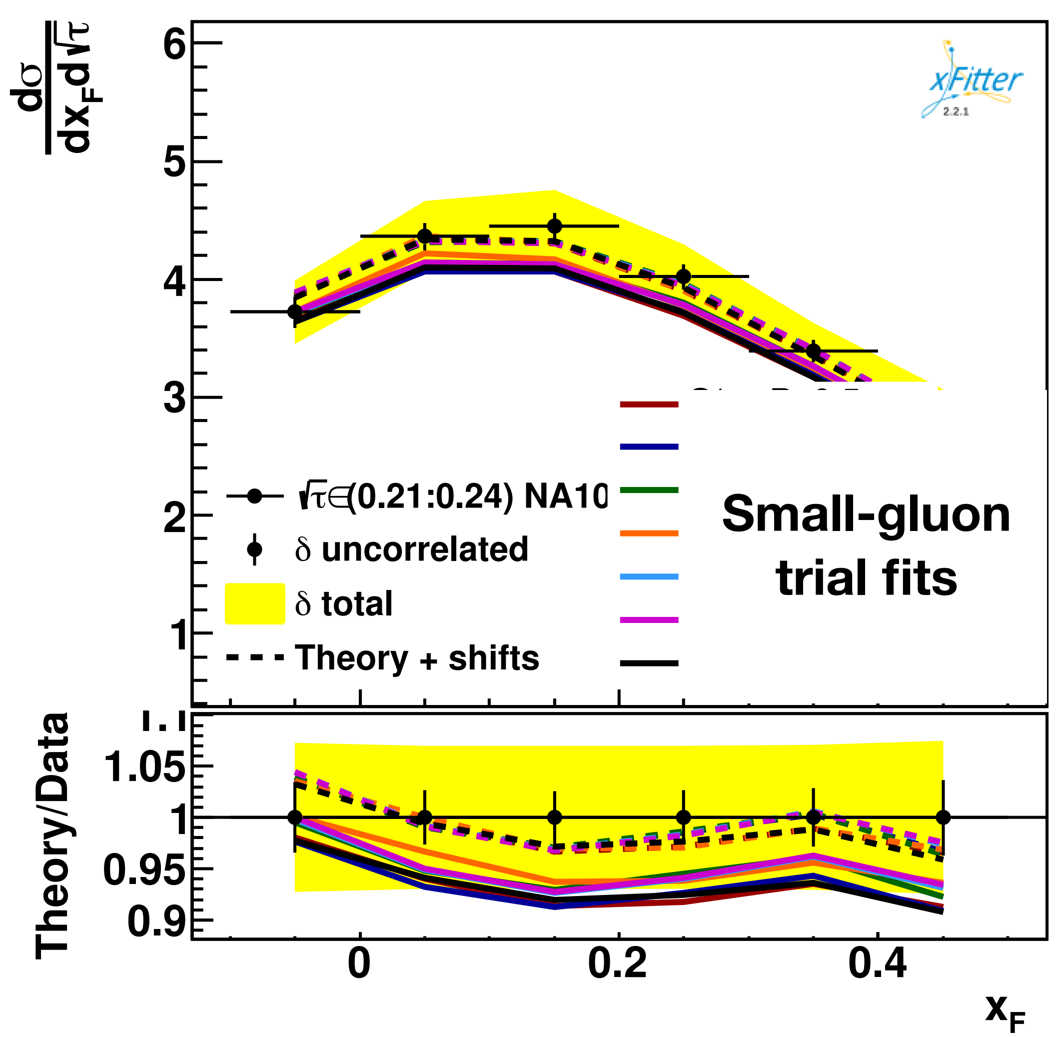
Fantômas4QCD program

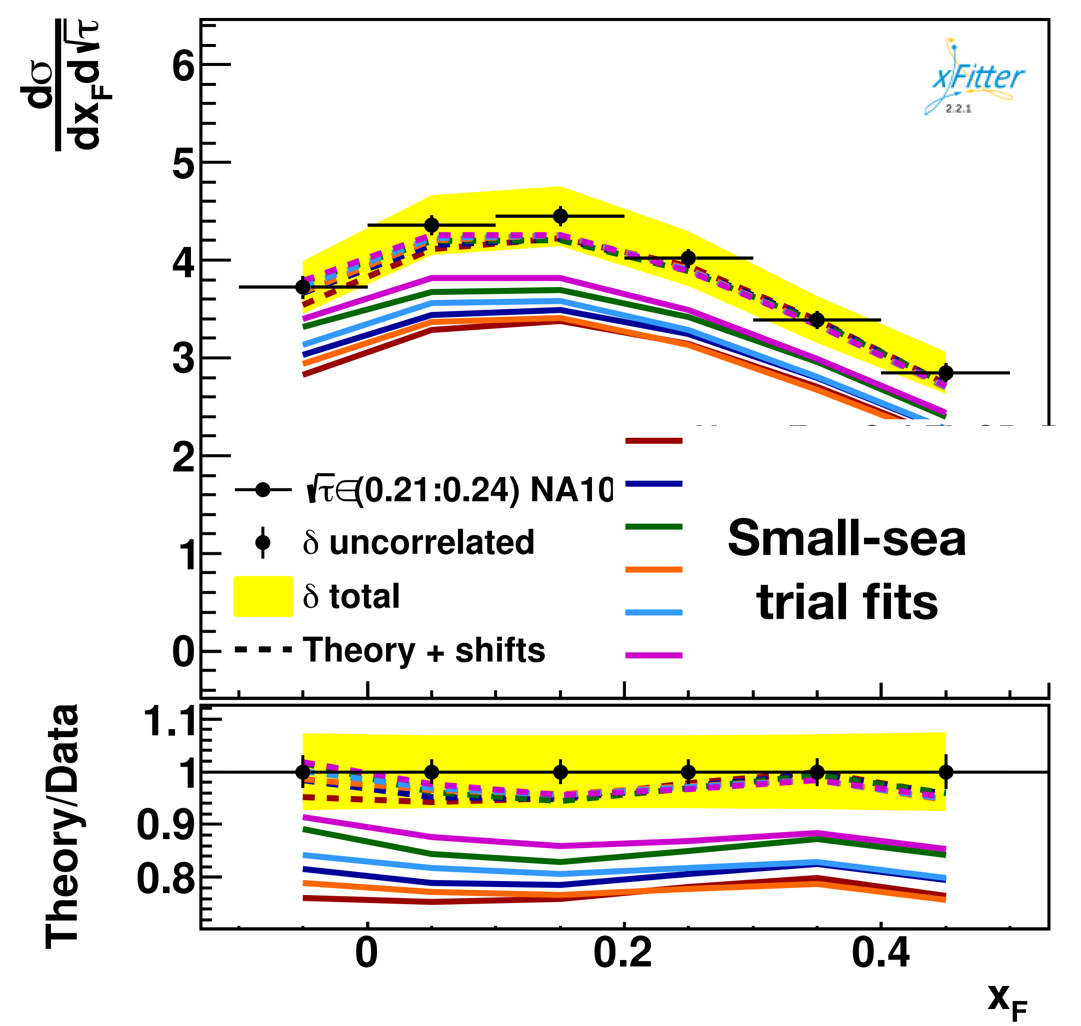
⇨ $\mathscr{B}$ can modulate the PDFs in flexible ways at intermediate x using a set of free and fixed control points

Sea and gluon behavior

Data sets vary between JAM and Fantômas: higher number of NA10 data points for us.

We explored small gluon and small sea scenarios:
zero-gluon solutions are allowed; zero-sea ones are unfavored.



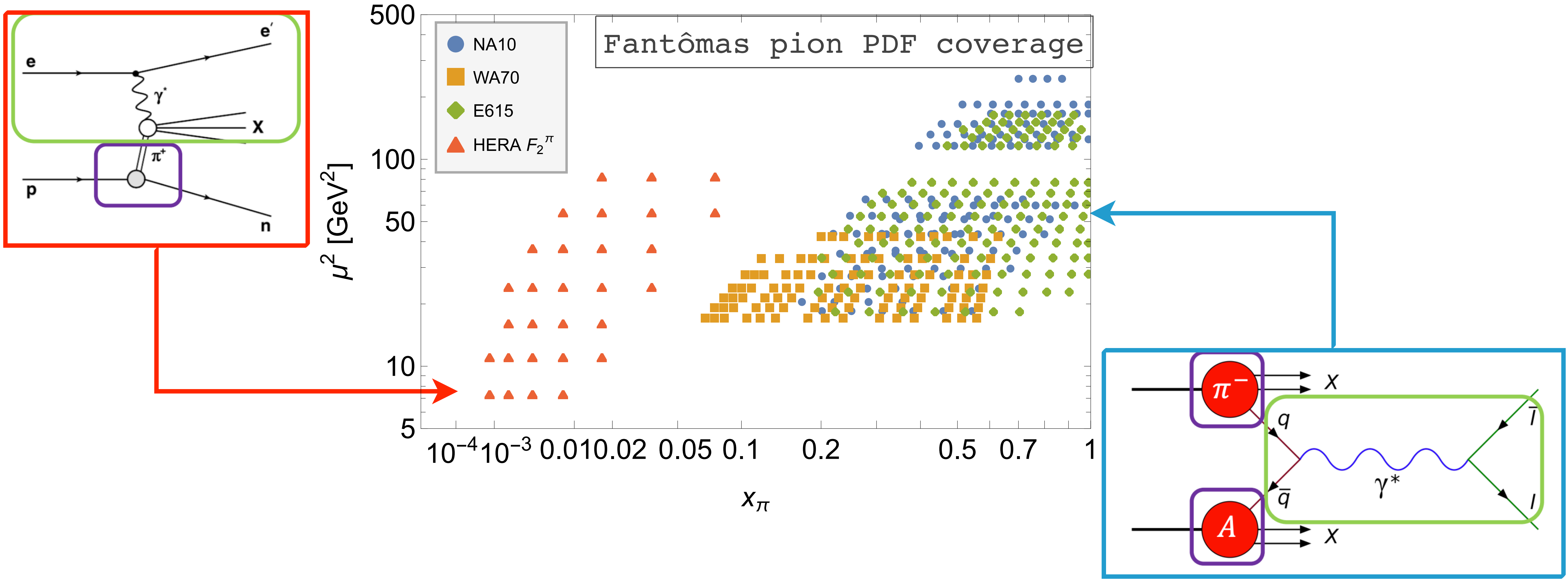


Pion PDFs at NLO — a convolution problem

Previous (modern) pion analyses:

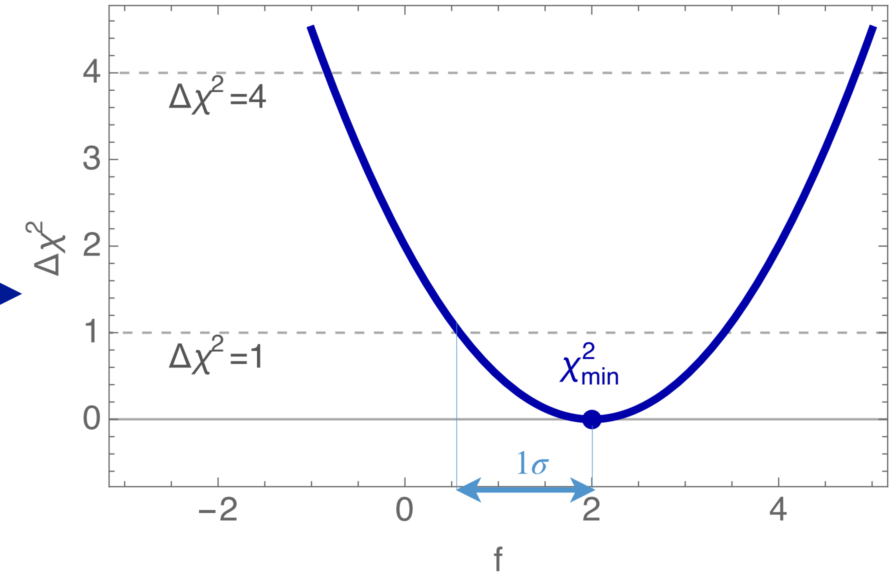
xFitter [PRD102]
JAM [PRL121, PRD103, PRL127]

We use the xFitter framework for pion PDFs.
We also extend the xFitter data by adding leading neutron (Sullivan process) data
— minimal small-x coverage [model-dependence in describing the pion as a target].

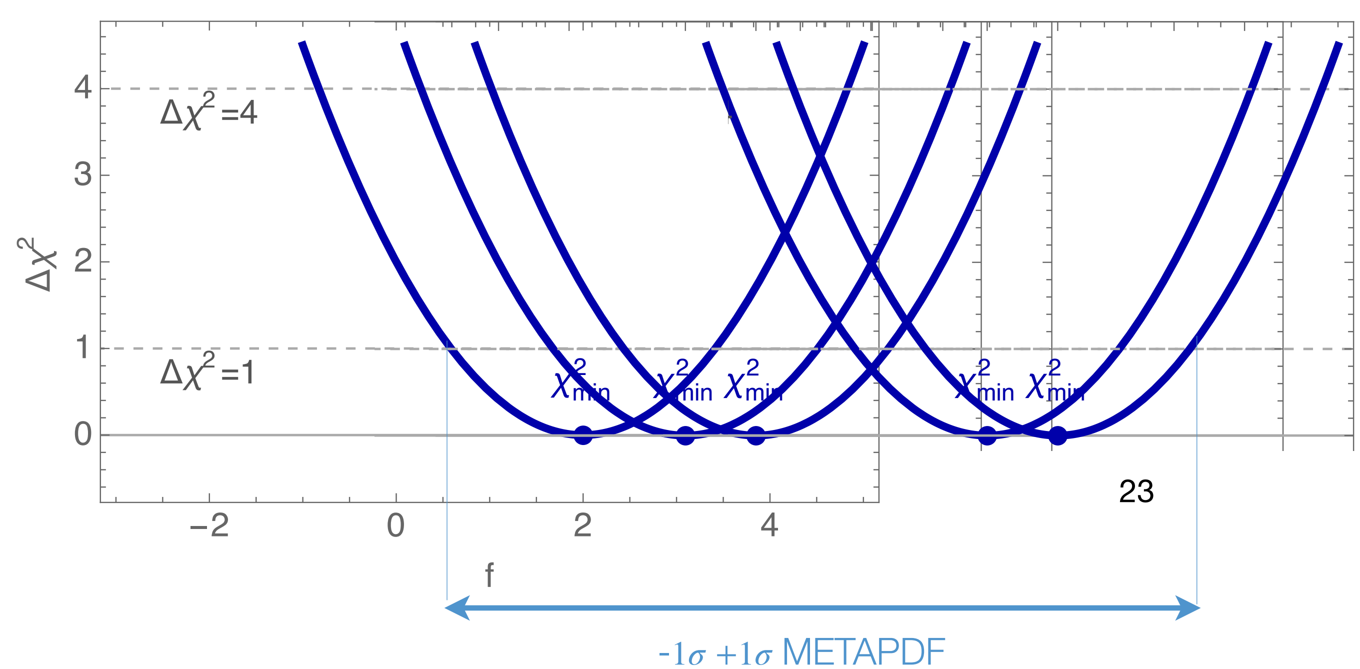


Uncertainties in global analyses

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



The $\Delta\chi^2 = 1$ criterion accounts for the 68% experimental uncertainty for the fixed settings of the fit. Additionally, we account for the uncertainty due to the PDF functional form using the METAPDF method.



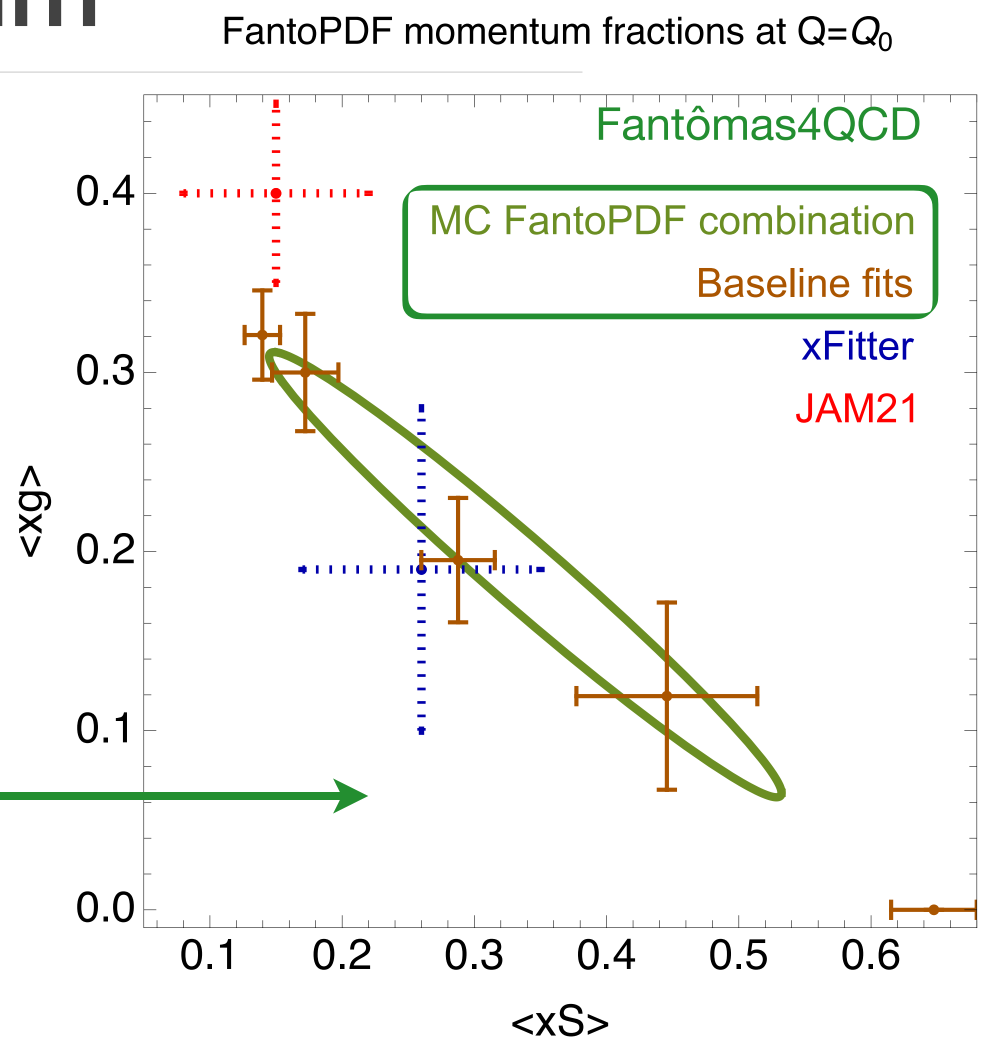
Distribution of the pion momentum

Momentum fraction x weighted by the PDF for $q = V, S, g$

$$\langle xq(Q^2)\rangle = \int_0^1 dx\, x\, f^q_{1,\pi}(x, Q^2)$$

Highlight on the separation of sea and gluon distributions.

The addition of leading-neutron data does not dramatically change the momentum fractions once the uncertainty appropriately include representative sampling.



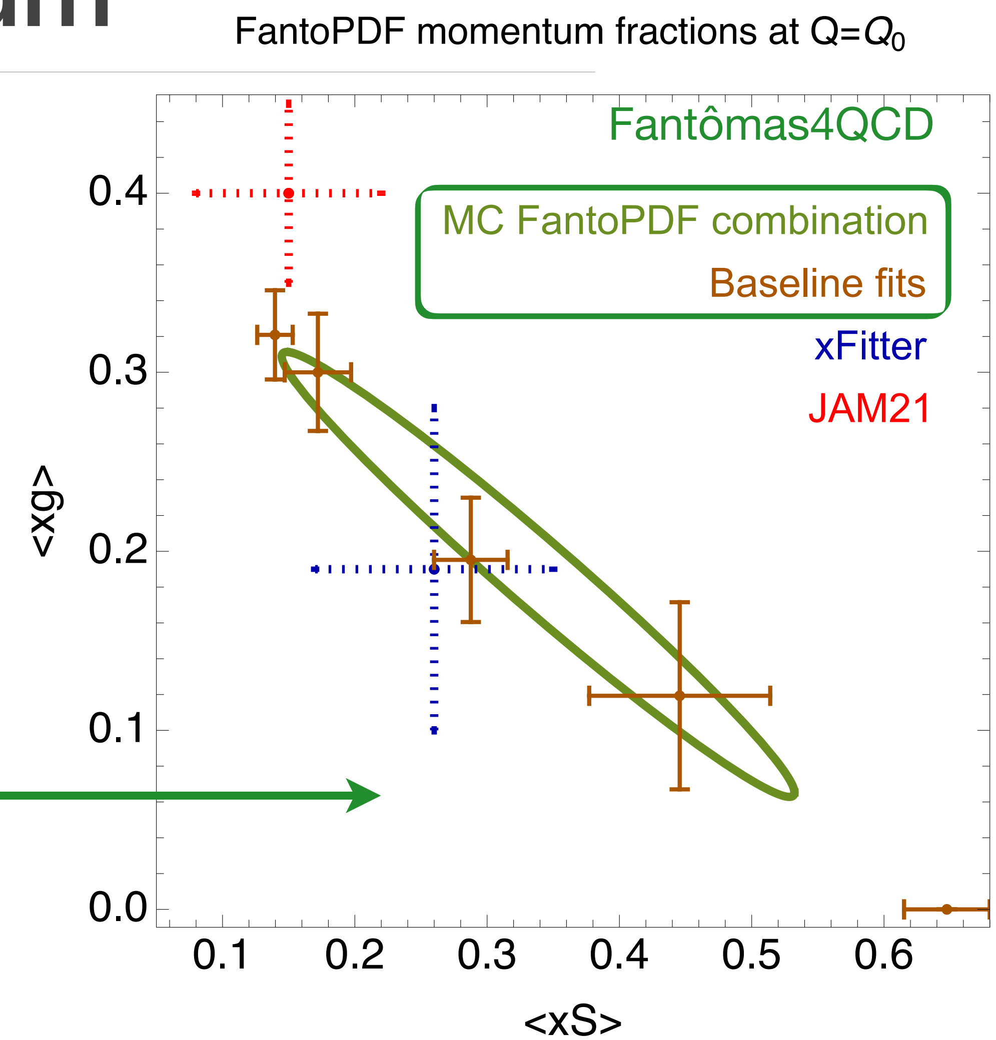
Distribution of the pion momentum

Momentum fraction x weighted by the PDF for $q = V, S, g$

$$\langle xq(Q^2)\rangle = \int_0^1 dx\, x f^q_{1,\pi}(x, Q^2)$$

Highlight on the separation of sea and gluon distributions.

The addition of leading-neutron data does not dramatically change the momentum fractions once the uncertainty appropriately include representative sampling.



Name	Q[GeV]	$\langle x(u+\bar{u})_{\pi^+}\rangle$	$\langle xg\rangle$
FantoPDF	2	0.331(25)	0.24(10)
HadStruct [19]	2	0.2541(33)	–
[Gao et al., PRD102]	3.2	0.216(19)(8)	–
ETM [46]	2	0.261(3)(6)	–
ETM [91]	2	$0.601(28)\|_{u+d}$	0.52(11)
[Meyer et al., PRD77]	2	–	0.37(8)(12)
[Shanahan et al., PRD99]	2	–	0.61(9)
[MSU, 2310.12034]	2	–	0.364(38)(36)
ZeRo Coll. [95]	2	0.245(15)	–
[Martinelli et al., PLB196]	7	0.02	–

Lattice provides complementary access to momentum fractions— only the recent ETM coll. results have both.

All lattice results are work with different ensemble settings.

Fantômas4QCD

The rôle of parametrization form in global analyses can be quantified

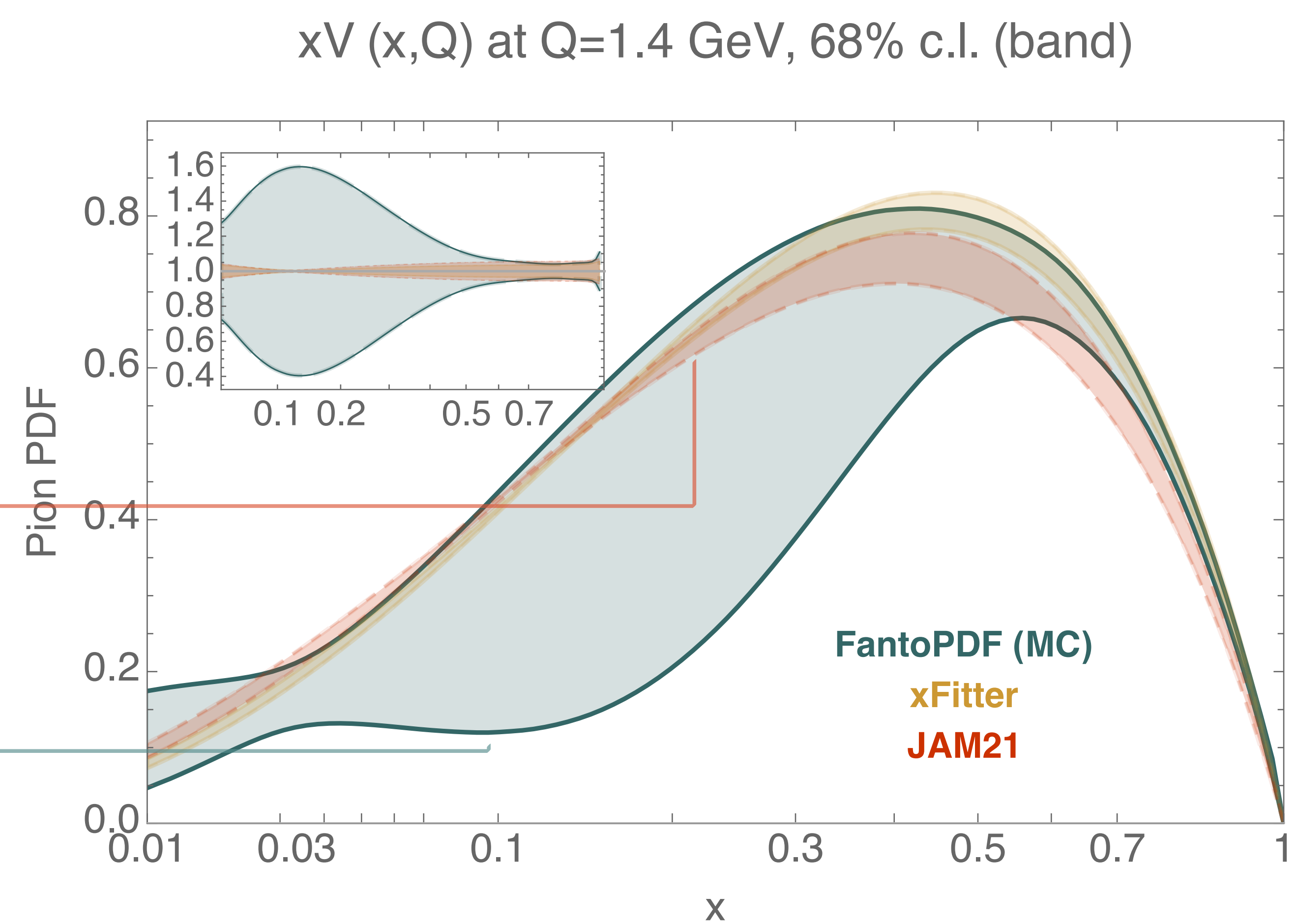
A new `c++` code automates series of fits using multiple functional forms, called `metamorph`.

[Kotz, AC, Nadolsky, Olness, Ponce-Chavez, PRD109]
[AC, Hobbs, Kotz, Nadolsky, Olness, Ponce-Chavez, Purohit, soon]

Fantômas πPDFs

without parametrization dependence

with parametrization dependence



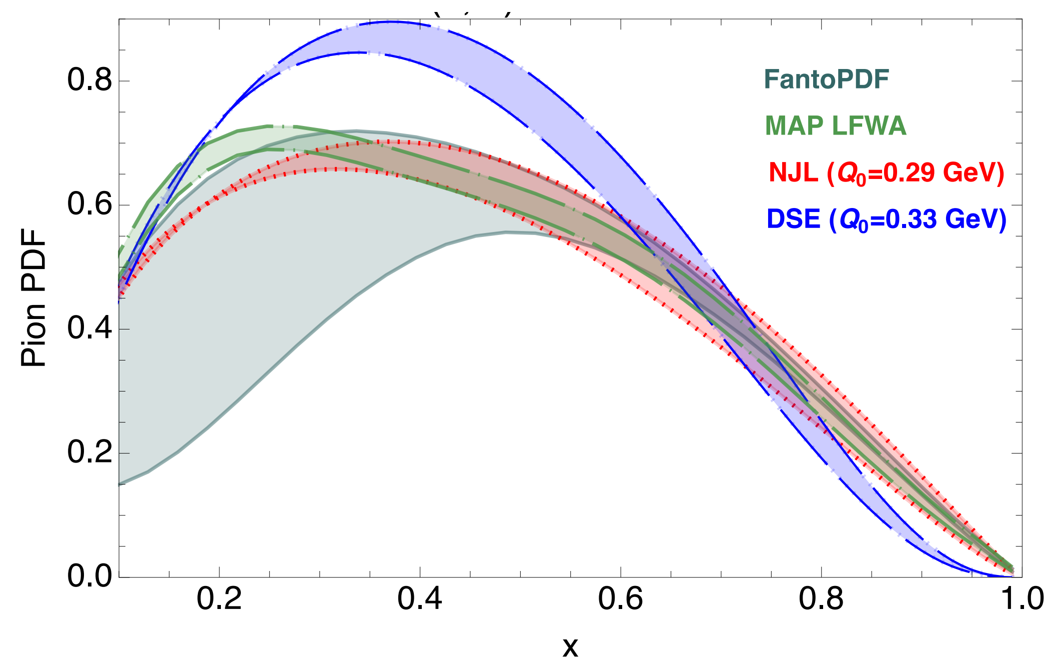
Hints of the mechanism that drives the pion structure

When testing polynomial shapes predicted from models, polynomial mimicry affects any interpretation.
No *if and only* conditions are possible given the state-of-the-art. [A.C. & Nadolsky, PRD103]

Contact-like kernel (NJL) and momentum-dependent kernel @ all order (DSE) calculations prescribe different initial conditions (Q_0^2 & *shape*), that evolve to different predictions at the scale of the data.
Light-front quark model with data-inferred parameters finds a similar large-x behavior.

[Ruiz-Arriola; Ding et al, PRD101]
[Pasquini et al, PRD107]

$2x(u-\bar{u})$ at $Q^2 = 10\,\mathrm{GeV}^2$



Comparing shapes, by evolving models from dangerously small scales.

Hints of the mechanism that drives the pion structure

When testing polynomial shapes predicted from models, polynomial mimicry affects any interpretation.
No *if and only* conditions are possible given the state-of-the-art. [A.C. & Nadolsky, PRD103]

Contact-like kernel (NJL) and momentum-dependent kernel @ all order (DSE) calculations prescribe different initial conditions (Q_0^2 & *shape*), that evolve to different predictions at the scale of the data.
Light-front quark model with data-inferred parameters finds a similar large-x behavior.

[Ruiz-Arriola; Ding et al, PRD101]
[Pasquini et al, PRD107]

Quark-counting rules: $f_{q_v/\pi}(x) \xrightarrow[x\to 1]{} (1-x)^2$

All pheno analyses find $f_{q_v/\pi}(x, Q_0^2) \xrightarrow[x\to 1]{} (1-x)^{\beta_{\rm eff}=1}$