IAIFI Workshop 2025

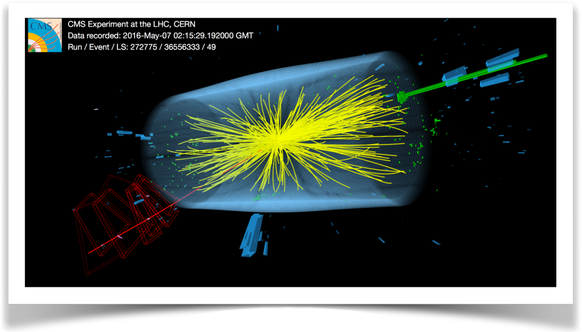
Symbolic regression and precision LHC physics

Manuel Morales-Alvarado

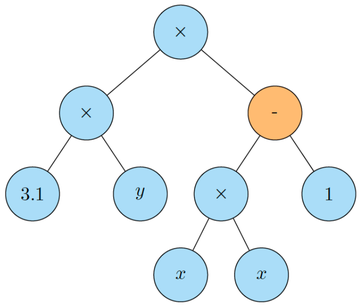
INFN, Sezione di Trieste

[2508.00989] in collaboration with J. Bendavid, D. Conde, V. Sanz, M. Ubiali

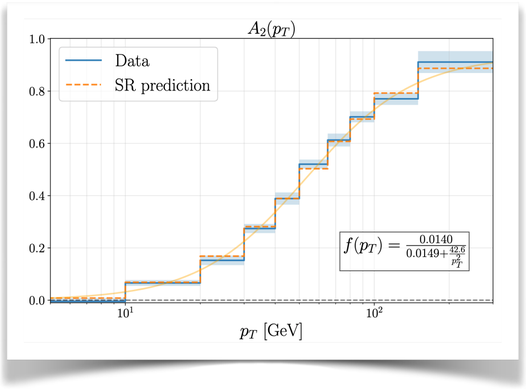
Outline



Introduction

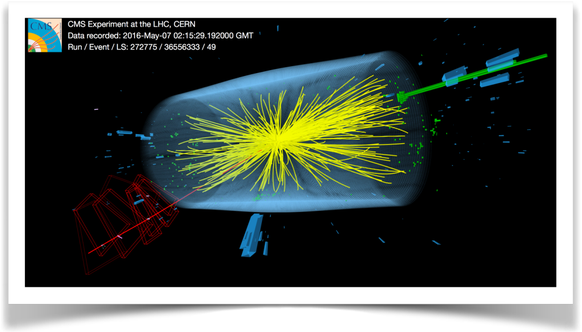


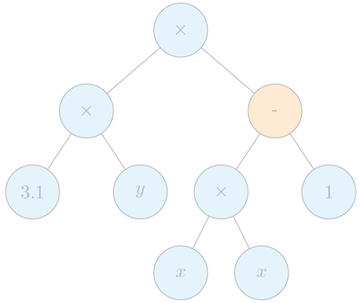
Symbolic regression (SR)

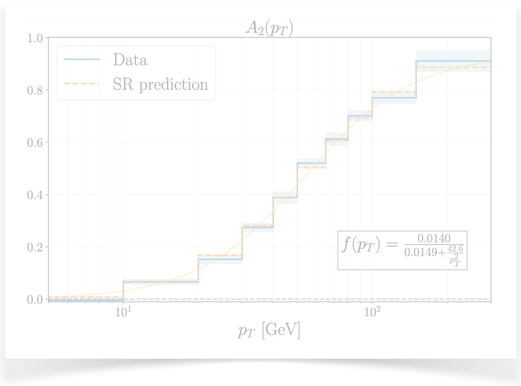


Collider observables via SR

Outline



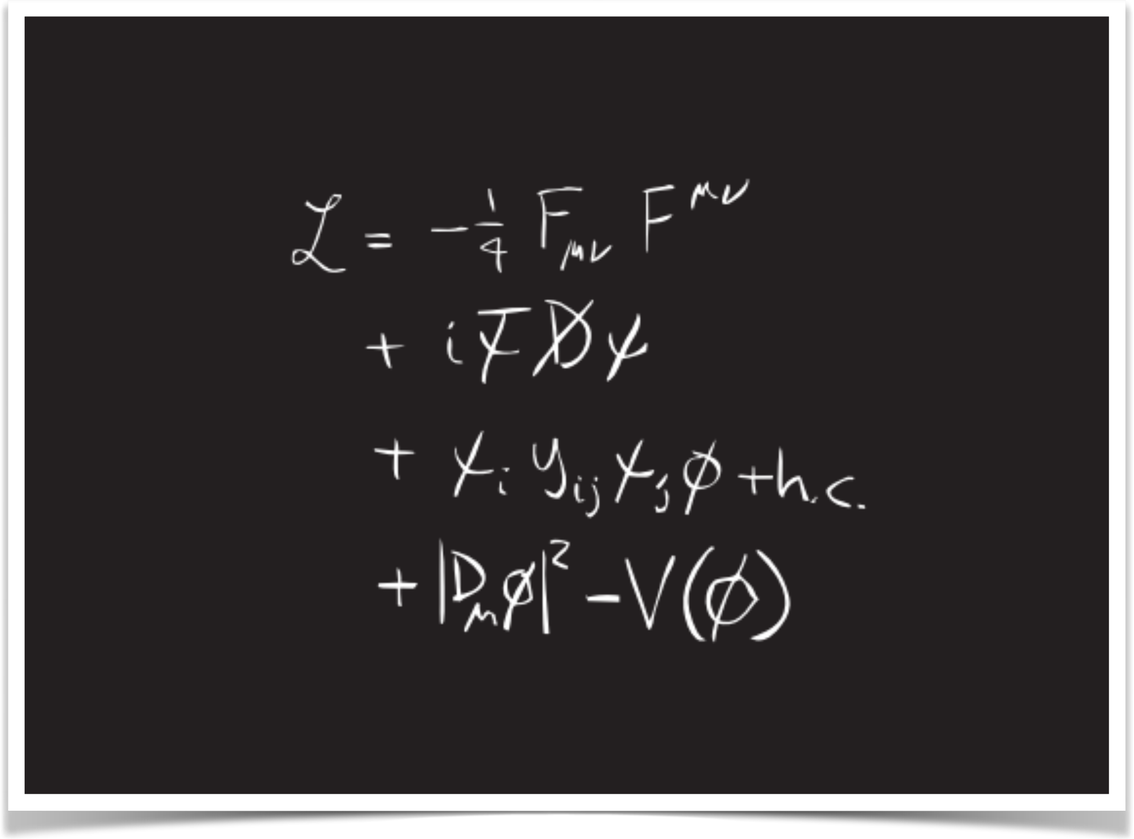




Introduction

High energy physics

The best fundamental theories that we have so far are written in terms of equations, symbolic representations.



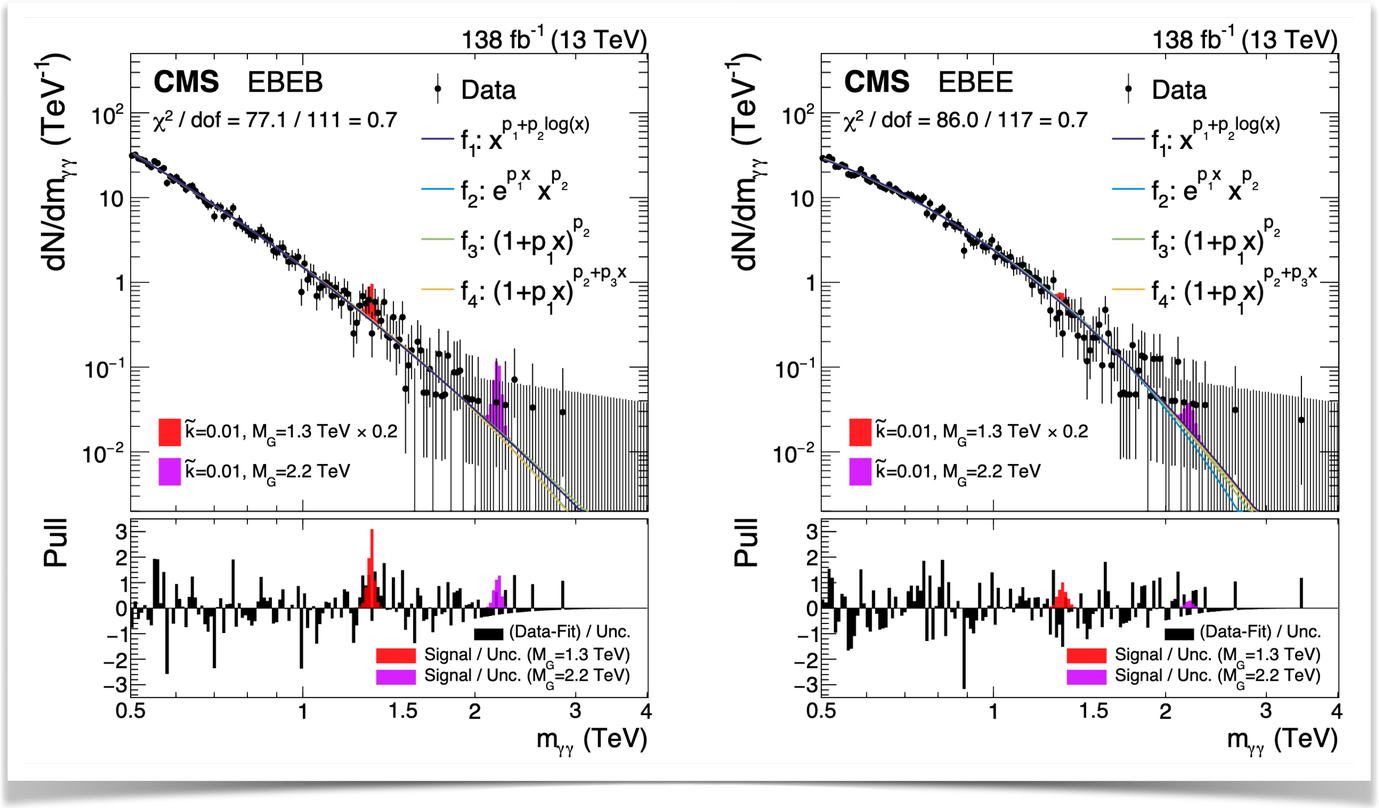
Symbolic representations in HEP

They are useful for several reasons:

1. Bridge the gap between black-box models and physical intuition.
2. Simplify machine-human interface.
3. Concrete examples: S/B modelling, and proton structure.

Symbolic representations in HEP

Example 1: model signal/background distributions in collider data (e.g. [2411.09851]).

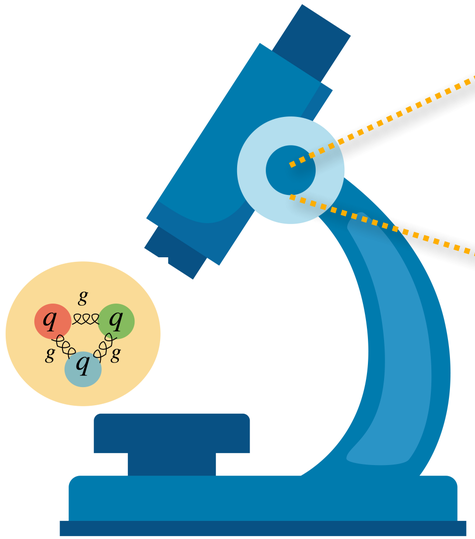


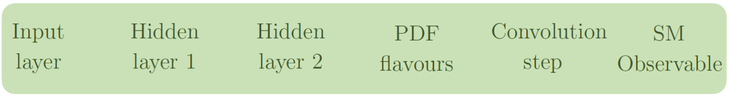
2405.09320

$$f_1(x) = p_0 x^{p_1 + p_2 \log(x)},$$
$$f_2(x) = p_0 e^{p_1 x} x^{p_2},$$
$$f_3(x) = p_0 (1 + x p_1)^{p_2},$$
$$f_4(x) = p_0 (1 + x p_1)^{p_2 + p_3 x}$$

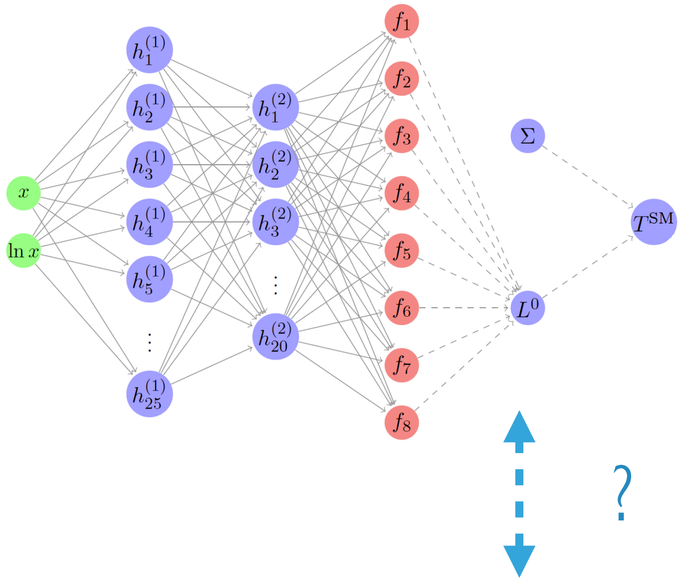
Symbolic representations in HEP

Example 2: study the structure of protons.





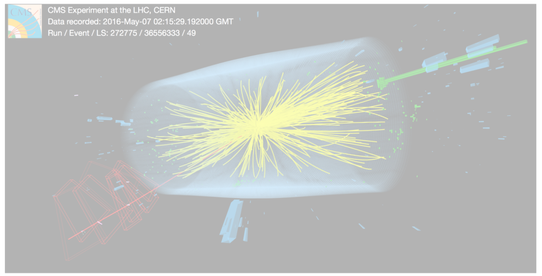
2109.02653

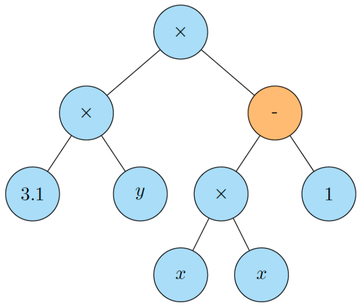


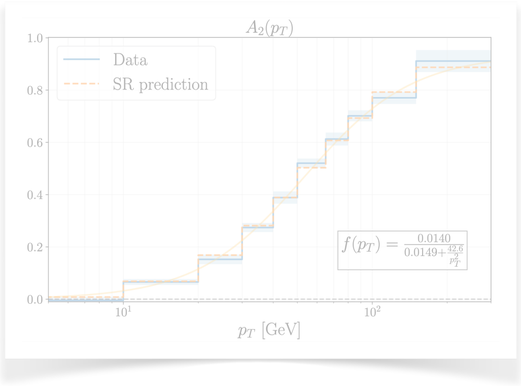
$$xf(x,Q_0^2) = A(1-x)^\eta x^\delta \left(1+\sum_{i=1}^{n} a_i T_i^{\rm Ch}(y(x))\right)$$

2012.04684

Outline







Symbolic regression (SR)

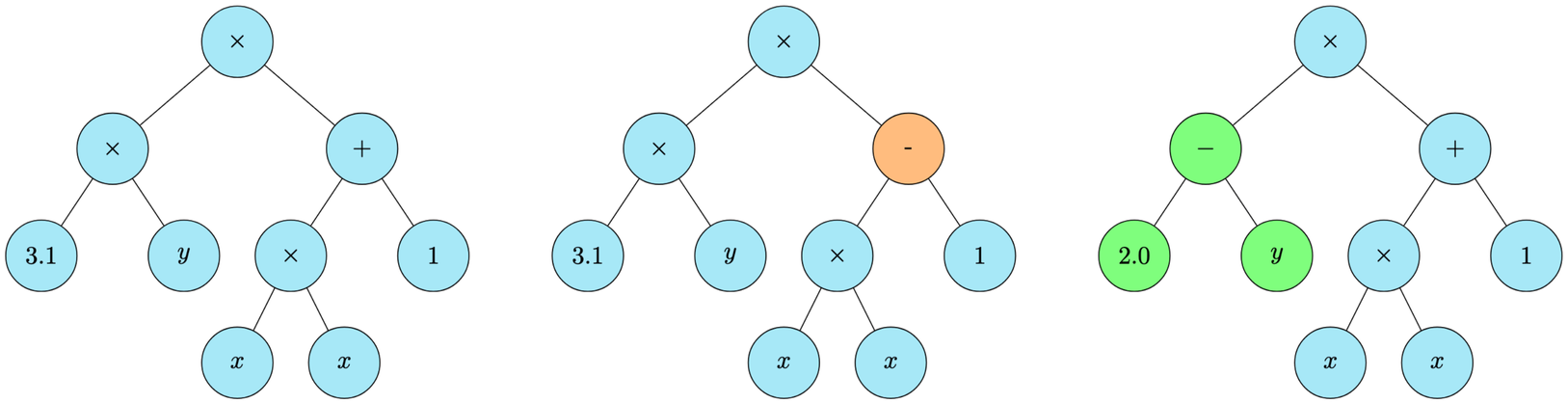
Symbolic regression

- Symbolic regression (SR) is a supervised learning technique whose aim is to automatically discover human-interpretable mathematical expressions that best fit the data.

- This is done without fixing completely the functional form, but allowing a set of operators: e.g. +, -, *, /.

- We use the PySR library [2305.01582], a multipopulation, multi-objective (accuracy/simplicity) optimisation algorithm.

$$3.1y\ (x^2+1)$$



Mutation Crossover

Symbolic regression

- The algorithm can optimise different selection criteria:

Accuracy:

Minimise

$$L = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Score:

Maximise

$$-\frac{\partial \log(L)}{\partial c}$$

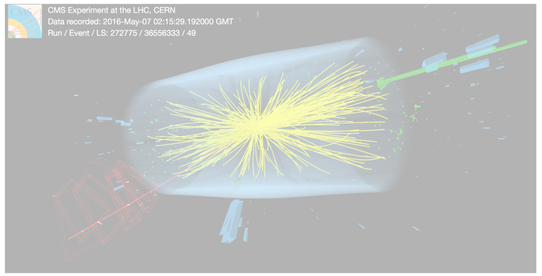
Best:

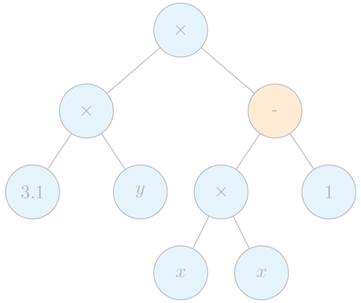
Highest score with

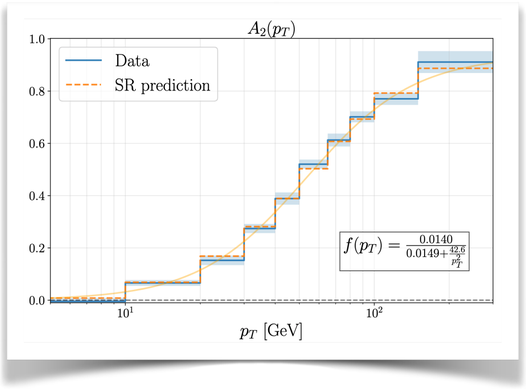
$$L \leq 1.5 \times L_{\min}$$

Equation	Complexity	Loss	Score
$f(p_T) = 0.313$	1	0.0795	0.0
$f(p_T) = 0.00640\,p_T$	3	0.0129	0.908
$f(p_T) = \frac{p_T}{p_T+67.5}$	5	0.00962	0.148
$f(p_T) = \frac{p_T}{p_T+\frac{3.30\cdot10^3}{p_T}}$	7	$1.48\cdot10^{-4}$	2.09
$f(p_T) = \frac{p_T}{1.05\,p_T+\frac{3.08\cdot10^3}{p_T}}$	9	$3.56\cdot10^{-5}$	0.712
$f(p_T) = \frac{0.955\,p_T}{p_T+\frac{2.84\cdot10^3}{p_T}} - 0.00748$	11	$1.89\cdot10^{-5}$	0.315
$f(p_T) = \frac{0.955\,p_T}{p_T+\frac{2.84\cdot10^3}{p_T}} - 0.00748$	13	$1.89\cdot10^{-5}$	$1.99\cdot10^{-5}$
$f(p_T) = \frac{0.953\,p_T}{p_T+\frac{2.84\cdot10^3}{p_T}} - 0.00634 - \frac{0.0102}{p_T}$	15	$1.87\cdot10^{-5}$	0.00562
$f(p_T) = \frac{0.953\,p_T}{p_T-0.0158+\frac{2.84\cdot10^3}{p_T}} - 0.00634 - \frac{0.0102}{p_T}$	17	$1.87\cdot10^{-5}$	$1.23\cdot10^{-4}$
$f(p_T) = \frac{0.953\,p_T}{p_T-0.0158+\frac{2.84\cdot10^3}{p_T-0.0424}} - 0.00634 - \frac{0.0102}{p_T}$	19	$1.87\cdot10^{-5}$	$8.02\cdot10^{-4}$

Outline

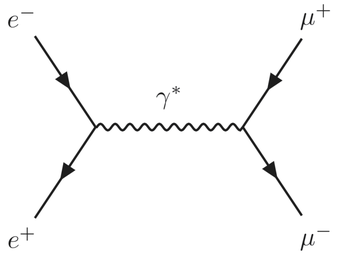






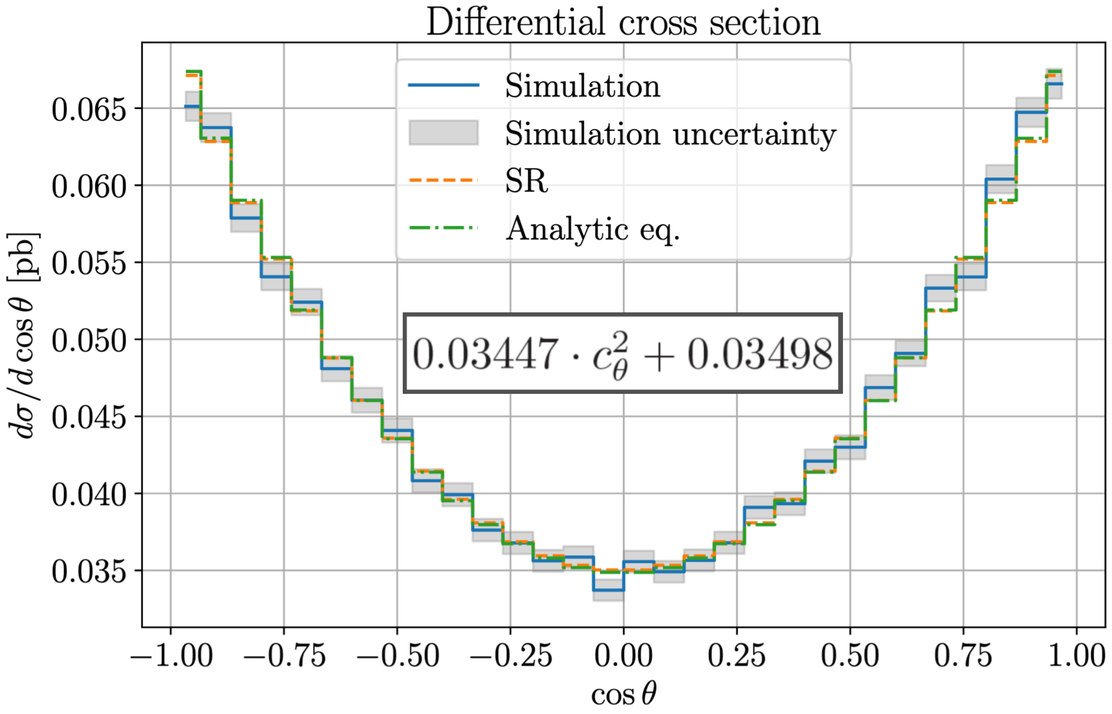
Collider observables via SR

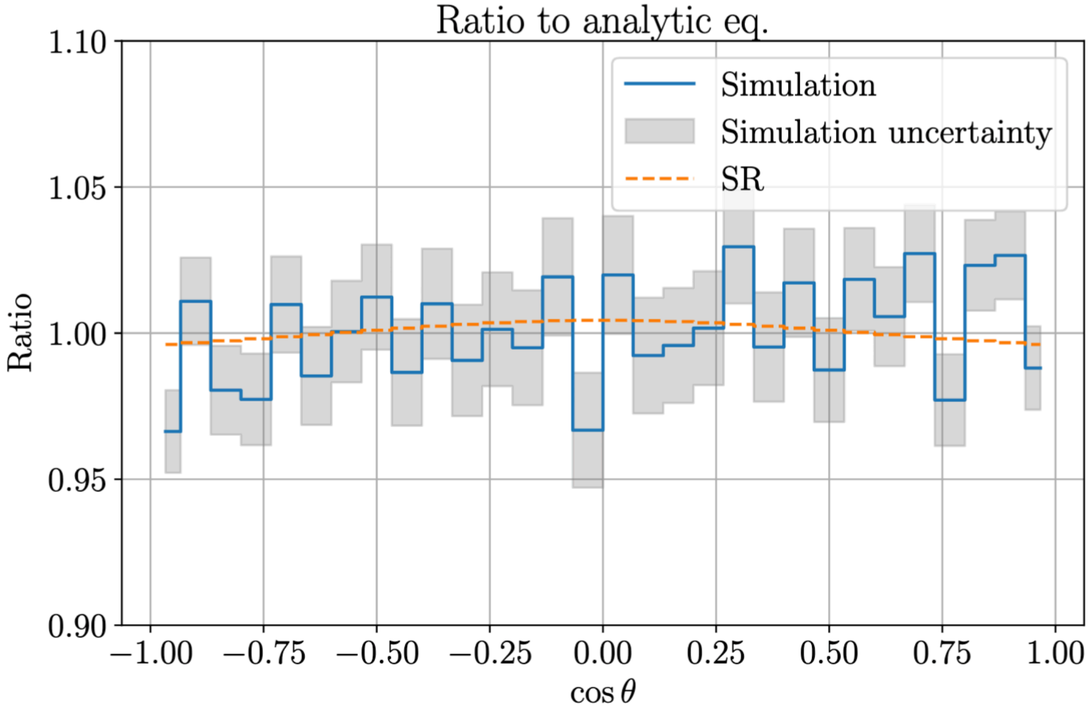
Scenario 1: equation *recovery* in QED



$$e^- e^+ \to \gamma^* \to \mu^- \mu^+$$

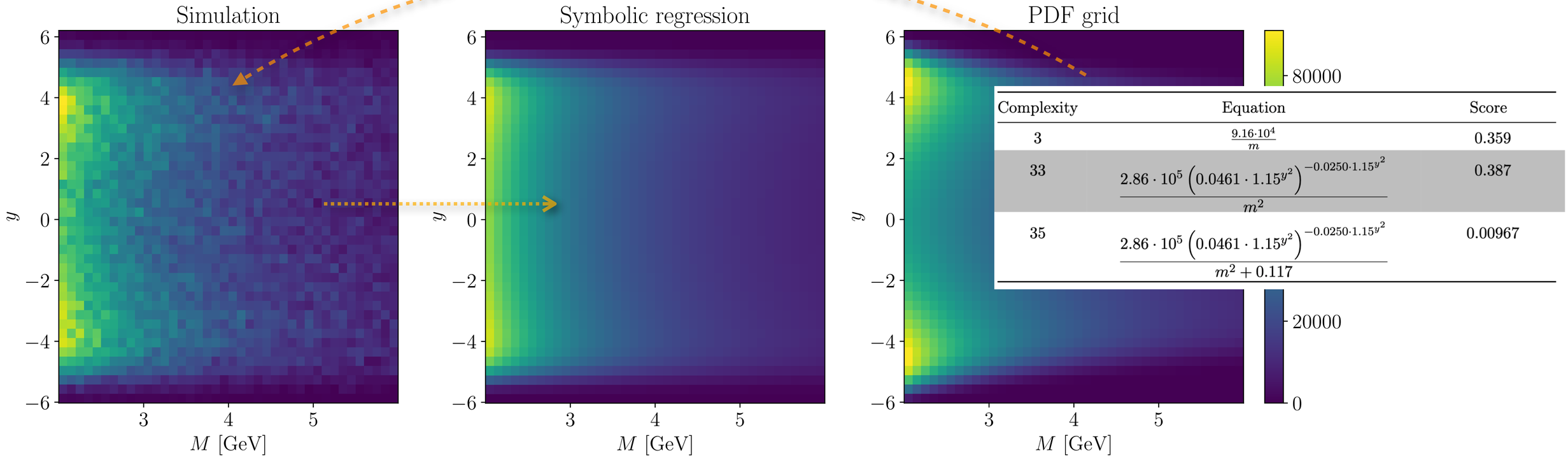
$$\frac{d\sigma}{d\cos\theta} = \frac{\pi\alpha^2}{2s}(1+\cos^2\theta)$$





Scenario 2: equation *discovery* in QCD

$$p\ p \to \gamma^* \to \mu^-\mu^+ \qquad \frac{d^2\sigma}{dMdy} = \frac{8\pi\alpha^2}{9sM} F(M,y)$$

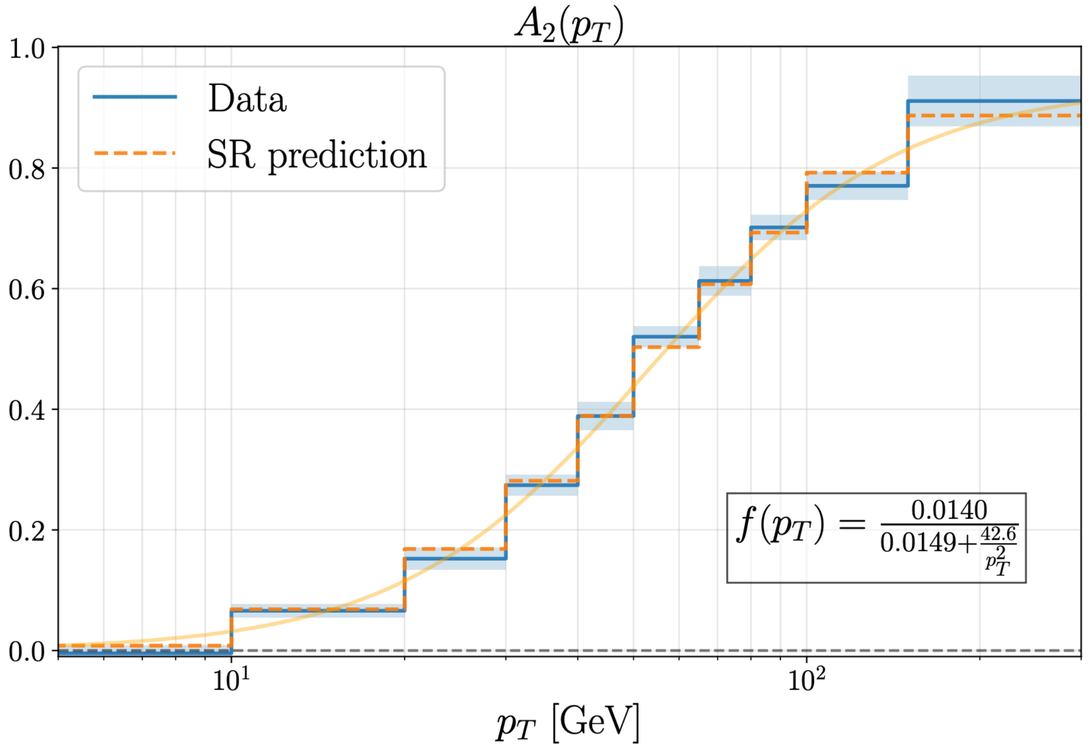


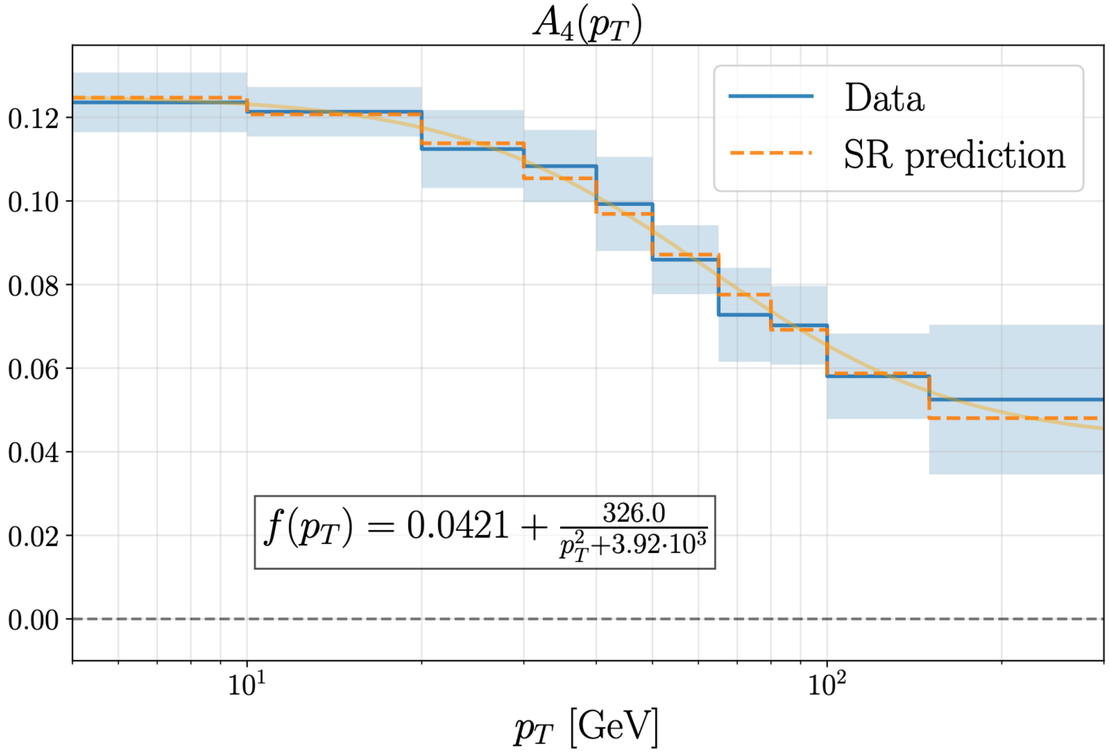
Complexity	Equation	Score
3	$\frac{9.16\cdot10^4}{m}$	0.359
33	$\frac{2.86\cdot10^5\left(0.0461\cdot1.15^{y^2}\right)^{-0.0250\cdot1.15^{y^2}}}{m^2}$	0.387
35	$\frac{2.86\cdot10^5\left(0.0461\cdot1.15^{y^2}\right)^{-0.0250\cdot1.15^{y^2}}}{m^2+0.117}$	0.00967

Scenario 3: 1D eq. *discovery* for angular coeffs.

$$\frac{d^5\sigma}{dp_T\,dy\,dm\,d\cos\theta\,d\phi} = \frac{3}{16\pi}\frac{d^3\sigma^{U+L}}{dp_T\,dy\,dm}\left[(1+\cos^2\theta) + \sum_{i=0}^{7} P_i(\theta,\phi)\,A_i\right]$$

$A_i(p_T, y, m)$

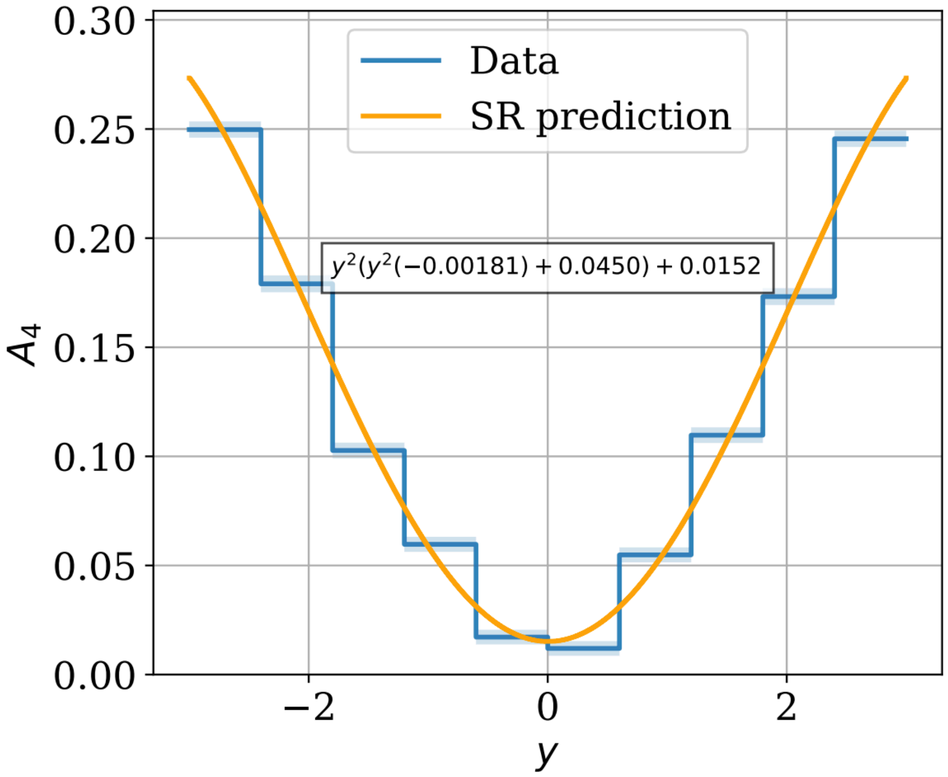


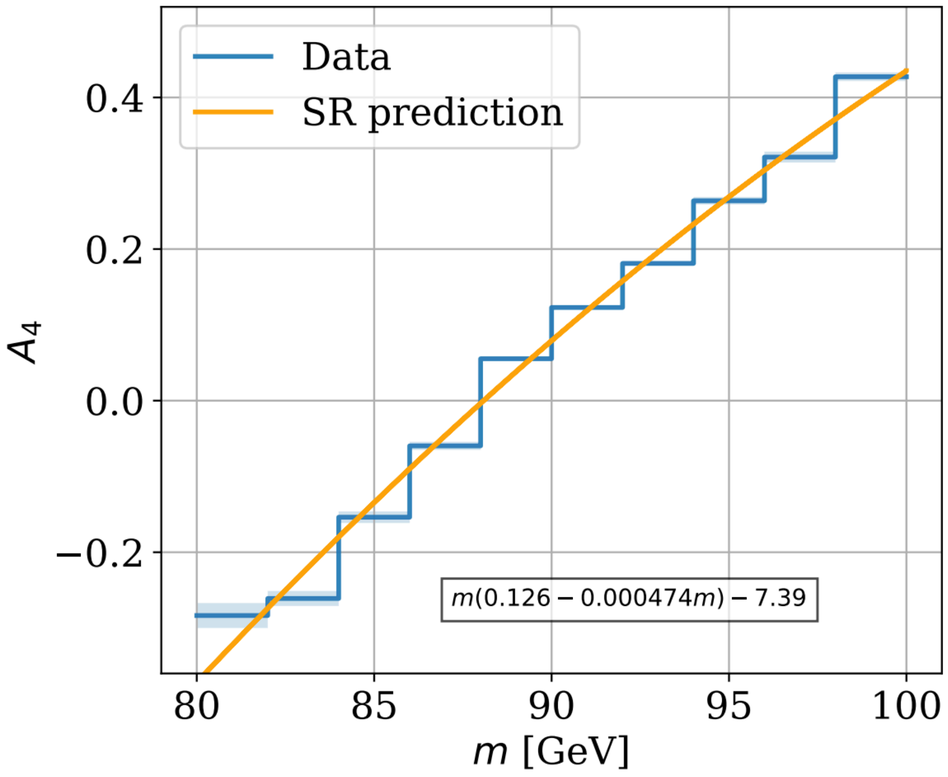


Scenario 3: 1D eq. *discovery* for angular coeffs.

$$\frac{d^5\sigma}{dp_T\,dy\,dm\,d\cos\theta\,d\phi} = \frac{3}{16\pi}\frac{d^3\sigma^{U+L}}{dp_T\,dy\,dm}\left[(1+\cos^2\theta) + \sum_{i=0}^{7} P_i(\theta,\phi)\,A_i\right]$$

$A_i(p_T, y, m)$

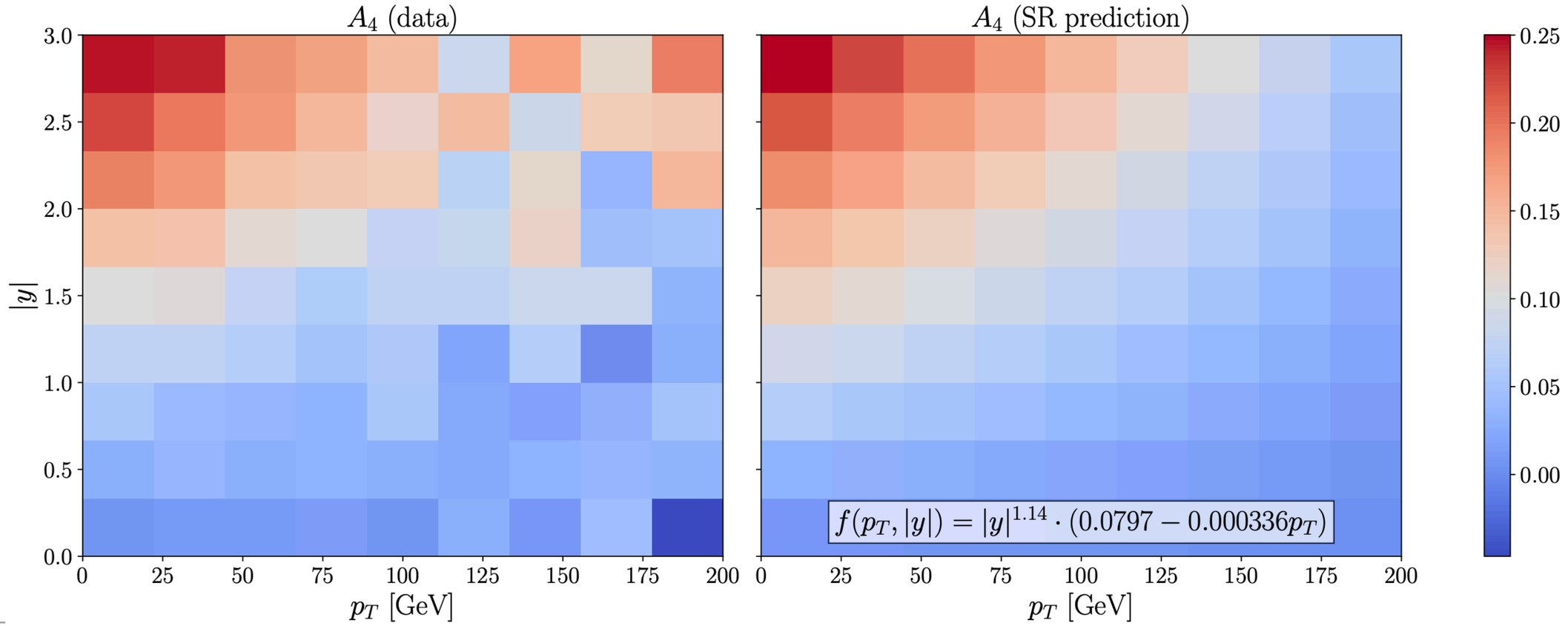




Scenario 3: 2D eq. *discovery* for angular coeffs.

$$\frac{d^5\sigma}{dp_T\,dy\,dm\,d\cos\theta\,d\phi} = \frac{3}{16\pi}\frac{d^3\sigma^{U+L}}{dp_T\,dy\,dm}\left[(1+\cos^2\theta) + \sum_{i=0}^{7} P_i(\theta,\phi)\,A_i\right]$$

$A_i(p_T, y, m)$



Conclusion

- Symbolic representations can be useful in different contexts.
- We can express quantities of interest in collider physics in compact, symbolic ways.
- Further work on the propagation of systematics in SR, hyperparam. optimisation, robustness, etc.

Thank you for your attention!

Backup slides

Halls of fame

Lepton angular distribution:

Bins	Accuracy	Best	Score
30	$x_0^2(x_0(0.00098 - 0.00257 \cdot x_0) + 0.03659) + 0.03477$	$0.03439 \cdot x_0^2 + 0.03499$	$0.03439 \cdot x_0^2 + 0.03499$
50	$x_0(x_0+0.00096)(-0.00119 \cdot x_0^2 + 0.00096 \cdot x_0 + 0.03547) + 0.03486$	$0.03445 \cdot x_0^2 + 0.03496$	$0.03445 \cdot x_0^2 + 0.03496$
100	$x_0^2(-0.00125 \cdot x_0(x_0^2 + x_0 - 1.60285) + 0.03553) + 0.03485$	$0.03446 \cdot x_0^2 + 0.03496$	$0.03446 \cdot x_0^2 + 0.03496$

Halls of fame

Structure function in Drell-Yan

Complexity	Equation	Score
3	$\frac{9.16 \cdot 10^4}{m}$	0.359
33	$\frac{2.86 \cdot 10^5 \left(0.0461 \cdot 1.15^{y^2}\right)^{-0.0250 \cdot 1.15^{y^2}}}{m^2}$	0.387
35	$\frac{2.86 \cdot 10^5 \left(0.0461 \cdot 1.15^{y^2}\right)^{-0.0250 \cdot 1.15^{y^2}}}{m^2 + 0.117}$	0.00967

Halls of fame

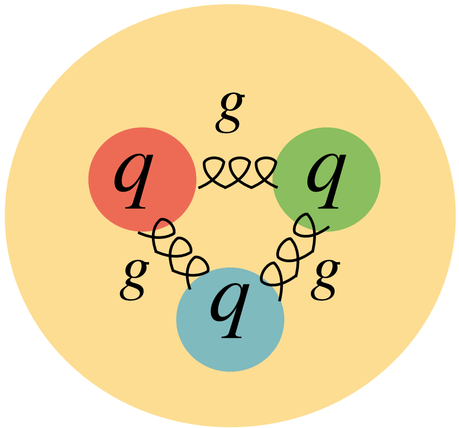
A0 1D distribution in pT.

Equation	Complexity	Loss	Score
$f(p_T) = 0.313$	1	0.0795	0.0
$f(p_T) = 0.00640\,p_T$	3	0.0129	0.908
$f(p_T) = \frac{p_T}{p_T+67.5}$	5	0.00962	0.148
$f(p_T) = \frac{p_T}{p_T+\frac{3.30\cdot10^3}{p_T}}$	7	$1.48\cdot10^{-4}$	2.09
$f(p_T) = \frac{p_T}{1.05\,p_T+\frac{3.08\cdot10^3}{p_T}}$	9	$3.56\cdot10^{-5}$	0.712
$f(p_T) = \frac{0.955\,p_T}{p_T+\frac{2.84\cdot10^3}{p_T}} - 0.00748$	11	$1.89\cdot10^{-5}$	0.315
$f(p_T) = \frac{0.955\,p_T}{p_T+\frac{2.84\cdot10^3}{p_T}} - 0.00748$	13	$1.89\cdot10^{-5}$	$1.99\cdot10^{-5}$
$f(p_T) = \frac{0.953\,p_T}{p_T+\frac{2.84\cdot10^3}{p_T}} - 0.00634 - \frac{0.0102}{p_T}$	15	$1.87\cdot10^{-5}$	0.00562
$f(p_T) = \frac{0.953\,p_T}{p_T-0.0158+\frac{2.84\cdot10^3}{p_T}} - 0.00634 - \frac{0.0102}{p_T}$	17	$1.87\cdot10^{-5}$	$1.23\cdot10^{-4}$
$f(p_T) = \frac{0.953\,p_T}{p_T-0.0158+\frac{2.84\cdot10^3}{p_T-0.0424}} - 0.00634 - \frac{0.0102}{p_T}$	19	$1.87\cdot10^{-5}$	$8.02\cdot10^{-4}$

Protons and partons

- At the LHC we collide *protons*.
- Proton are not elementary particles. They are QCD bound states of elementary particles called *partons* (quarks, gluons, …).
- The parton model postulates that interactions between hadrons (protons, in particular) are interactions of point-like parton convolved with functions that parametrise the structure of the hadron: *parton distribution functions* (PDFs).

Proton



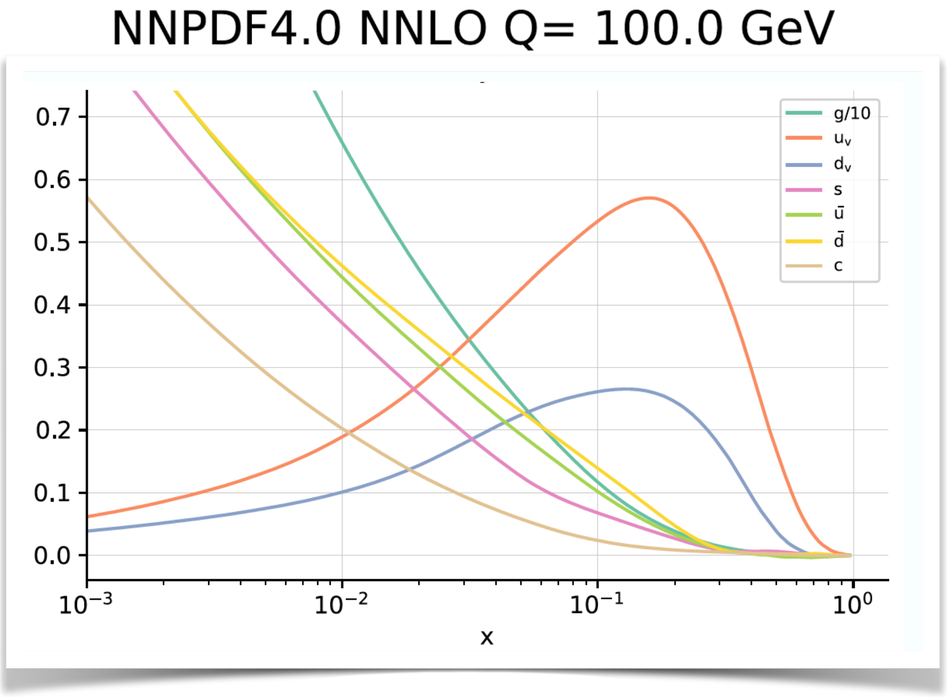
Parton distribution functions

- Some intuition. Consider the PDF of the up quarks:

$$f_u(x) = u(x)$$

$u(x)$ → PDF of the up-quarks: the 'probability' of finding an up-quark in the proton carrying a fraction x < 1 of the momentum of the proton.

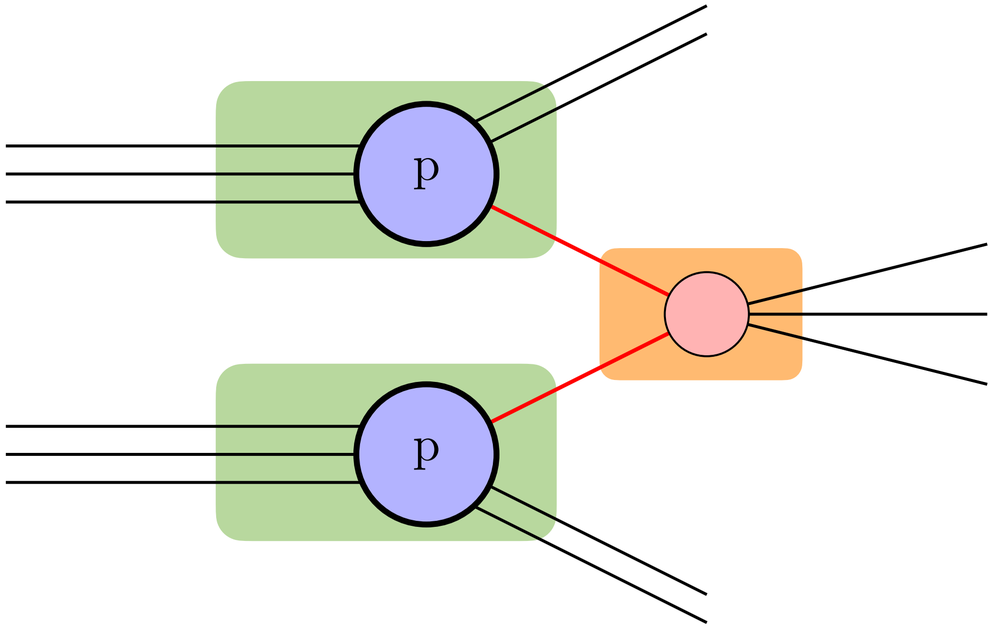
$$\int_0^1 xu(x)dx$$ → Fraction of the total momentum of the proton carried up by quarks.



NNPDF, 2109.02653

Factorisation and hadronic observables

Consider a proton-proton collision:



$$\sigma = \sum_{i,j} \int_0^1 dx_1 \int_0^1 dx_2 f_i(x_1) f_j(x_2) \hat{\sigma}_{ij}(x_1, x_2)$$

$x_{1,2}$: fraction of the hadron's momentum that is carried by the hadron

$\hat{\sigma}_{ij}$: partonic cross section

$f_i(x)$: PDF of parton of type i

- Radiative corrections introduce IR singularities that have to be factorised: $f(x) \to f(x, Q^2)$

Scale dependance of the PDFs

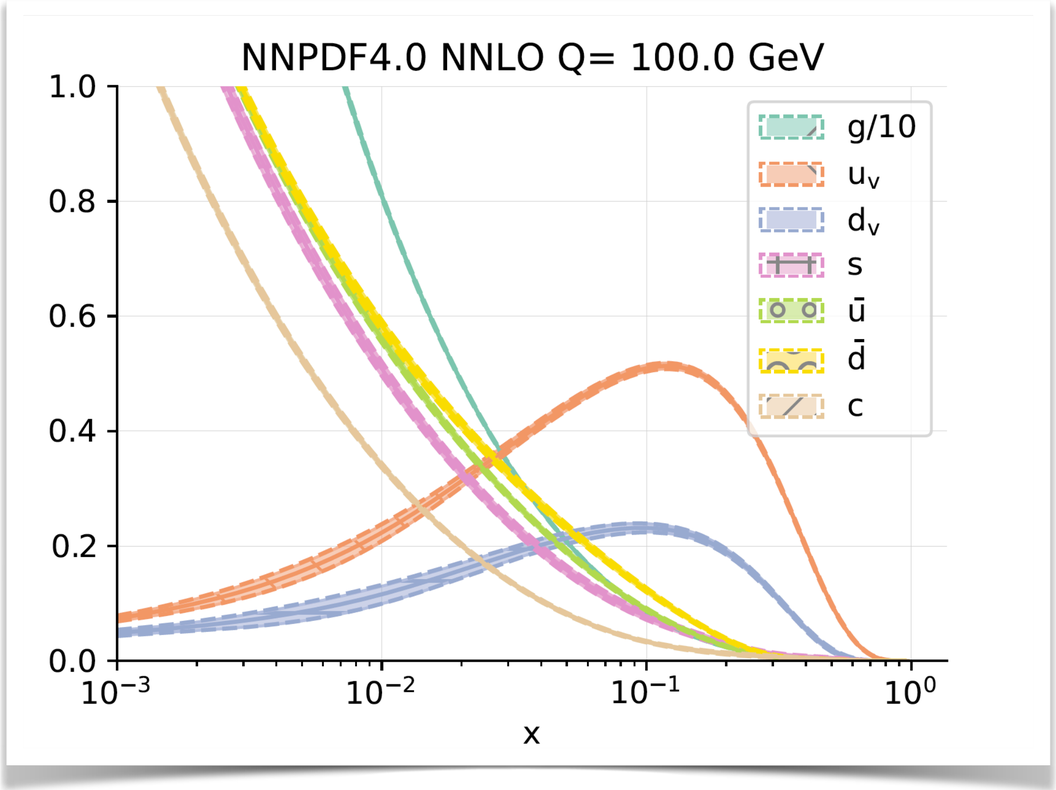
- Radiative corrections introduce IR singularities in the calculation of observables.
- PDFs have to be renormalised and acquire a dependence on one extra parameter: a factorisation scale Q^2.
- This factorisation scale separates short-distance from long-distance effects.
- Therefore, we promote:

$$f(x) \to f(x, Q^2)$$

PDF determination

PDFs **cannot** be calculated from first principles in perturbation theory, they have to be extracted from *data*.

$$f(x, Q^2)$$

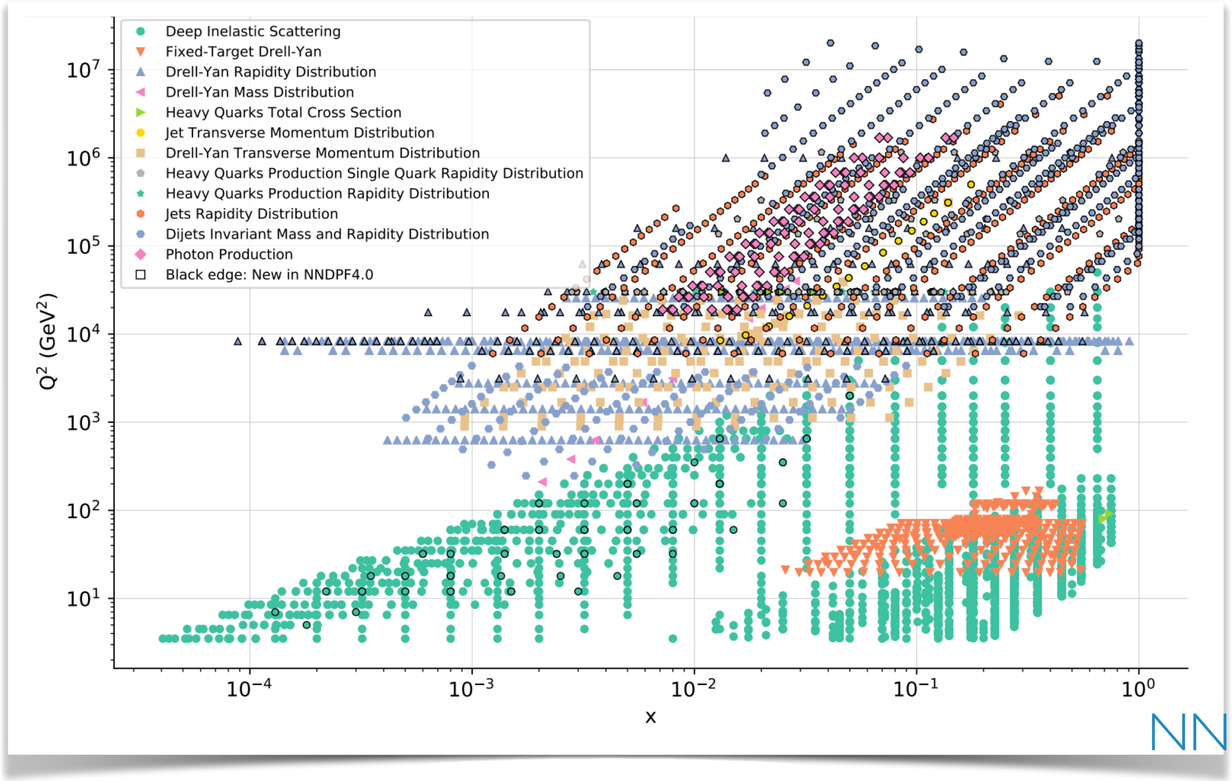


Some recent PDF fits include:

- NNPDF4.0, Ball et al., 2109.02653
- MSHT20, McGowan et al., 2012.04684
- CT18, Hou et al., 1912.10053

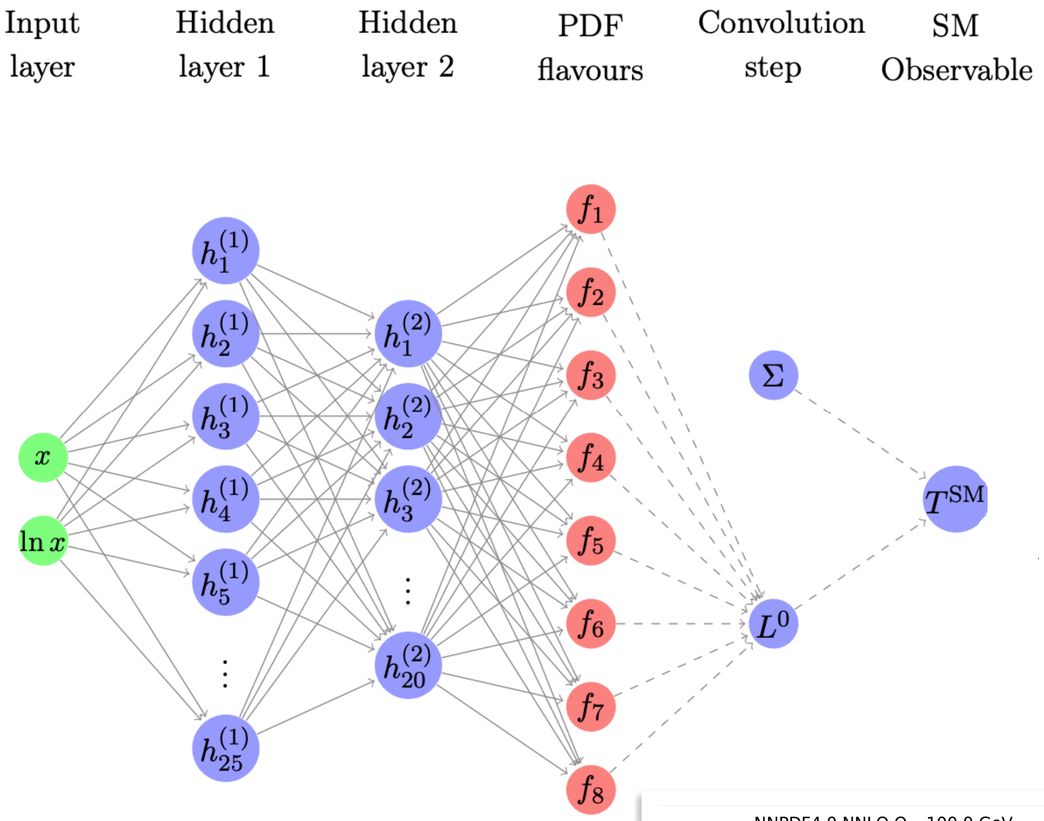
NNPDF4.0

- Thousands of collider measurements go into PDF fits.
- The kinematic coverage spans several orders of magnitude.



NNPDF, 2109.02653

NNPDF4.0



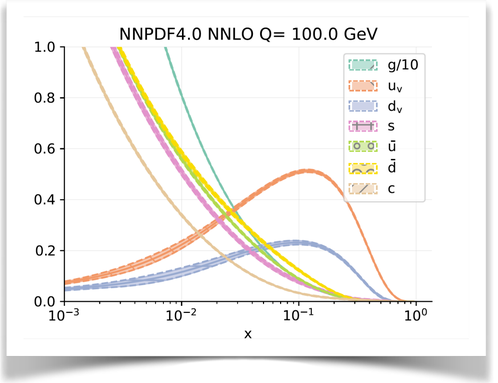
- PDFs are parametrised by neural networks (NNs):

$$f(x, Q_0^2) = \mathrm{NN}(x)$$

- The optimal weights of the NN are found by minimising a loss function:

$$\chi^2(\theta) = \frac{1}{N_{\rm dat}}(\mathbf{D} - \mathbf{T}(\theta))^T(\mathbf{cov})^{-1}(\mathbf{D} - \mathbf{T}(\theta)) \quad \theta = \theta_{opt}$$

- Uncertainty is propagated via de Monte Carlo replica method. 2109.02653, 2404.10056



PDFs through SR

The scaling in *x* of the PDFs is dictated by *Regge theory*:

$$\lim_{x\to 1} f(x,Q^2) = 0 \qquad\qquad \lim_{x\to 0} f(x,Q^2) \propto x^{\alpha}$$

We can integrate these requirements in the parametrisations we discussed:

$$f(x,Q^2) = x^{\alpha}(1-x)^{\beta}\mathrm{NN}(x,Q^2)$$

$$f(x,Q^2) = x^{\alpha}(1-x)^{\beta} f_{\mathrm{red}}(x,Q^2)$$

Could SR be of help?

$$f(x,Q^2) = x^{\alpha}(1-x)^{\beta}\mathrm{SR}(x,Q^2)$$

Corrections from Regge theory?

NN/analytical comparison?

Simpler functional forms with flexibility?

What about *uncertainty*?

NNPDF4.0

Input layer | Hidden layer 1 | Hidden layer 2 | PDF flavours | Convolution step | SM Observable

- PDFs are par
networks (N

$f(x,$

- The optimal
minimising a

$\chi^2(\theta) = \frac{1}{N_{\rm da}}$

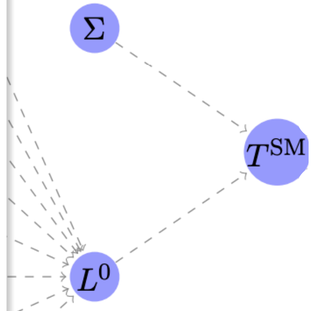
- Uncertainty is propagated via de Monte Carlo replica method. 2109.02653, 2404.10056

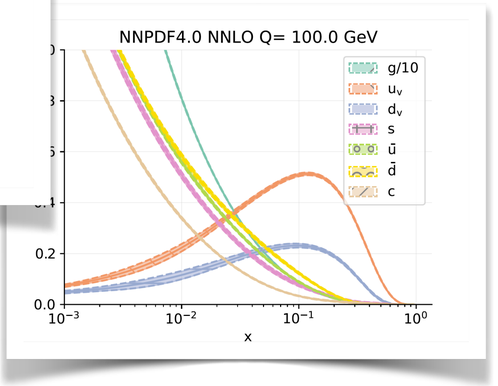
Other parametrisations are possible. For example, in the case of MSHT20, PDFs are parametrised with *fixed functional forms* of the type:

$$xf(x,Q_0^2) = A(1-x)^\eta x^\delta \left(1+\sum_{i=1}^{n} a_i T_i^{\rm Ch}(y(x))\right)$$

$T_i^{\rm Ch}$: Chebyshev polynomials

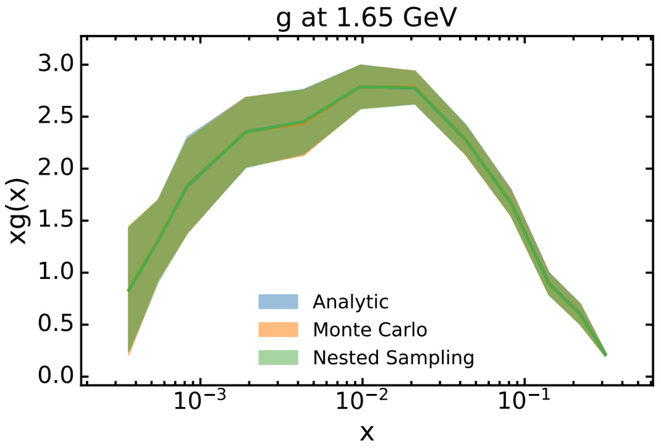
where

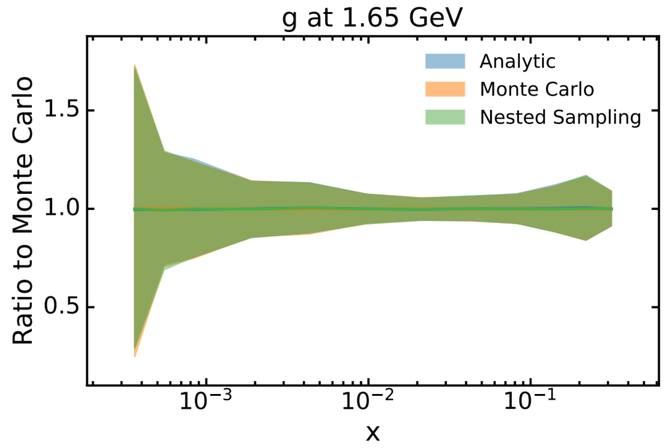




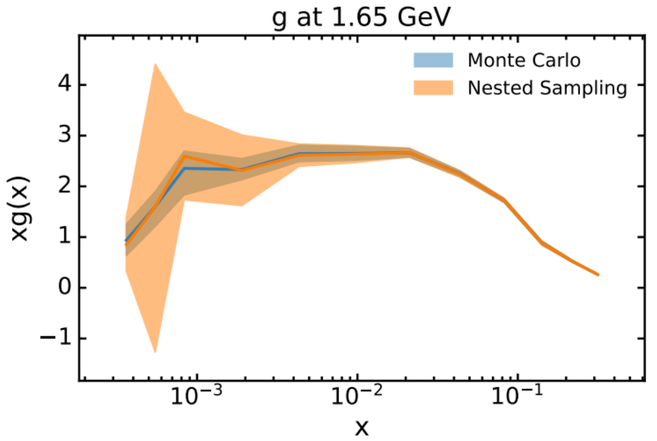
MC Replica vs Bayesian analysis [2404.10056]

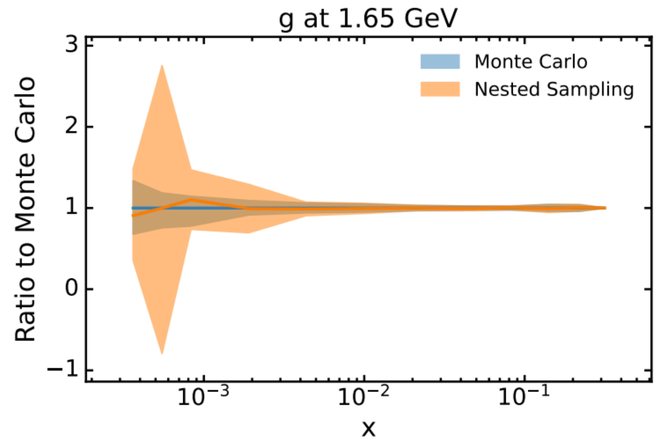
PDF: DIS case





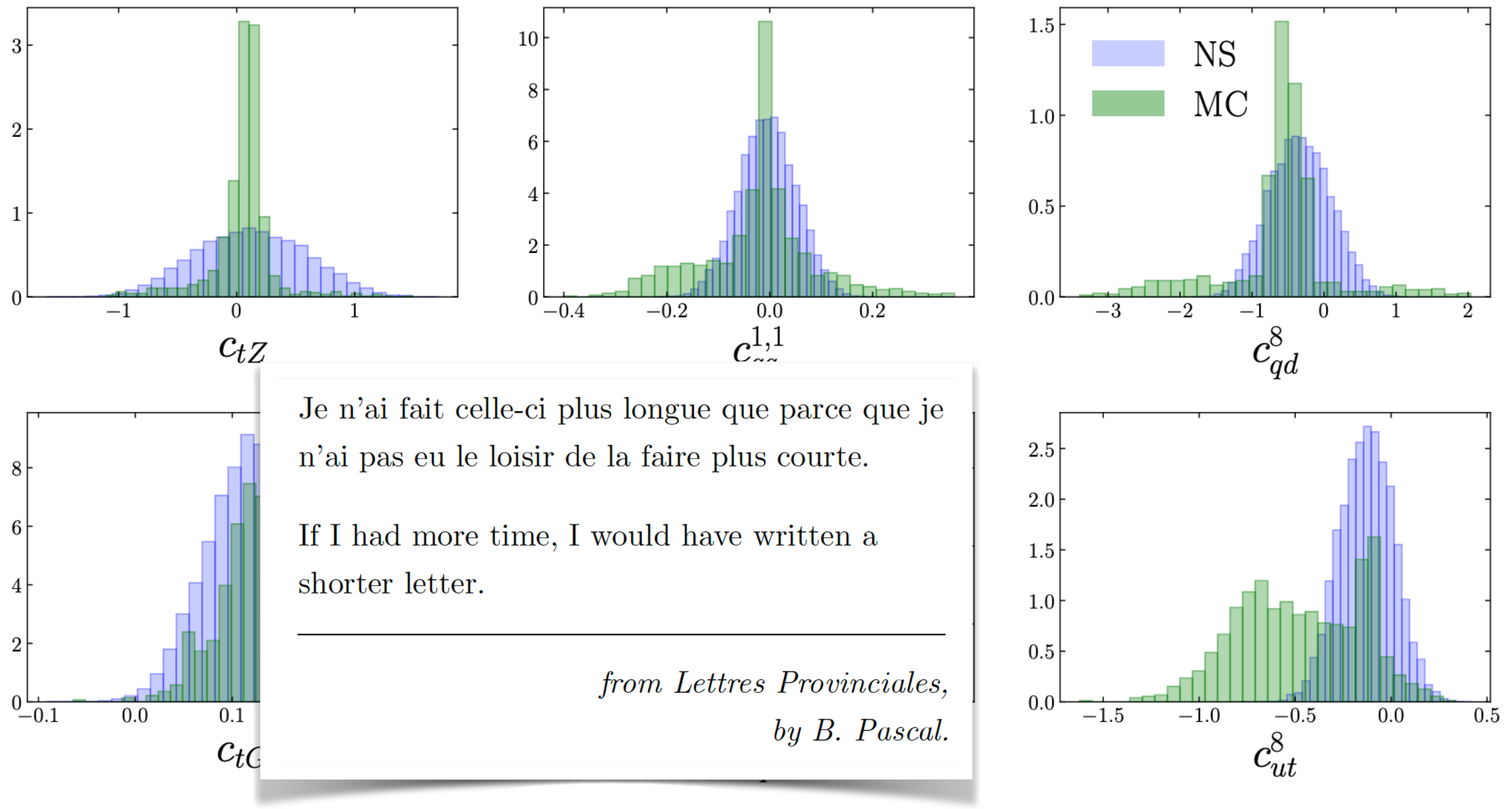
PDF: Hadronic case





MC Replica vs Bayesian analysis

Wilson coefficients



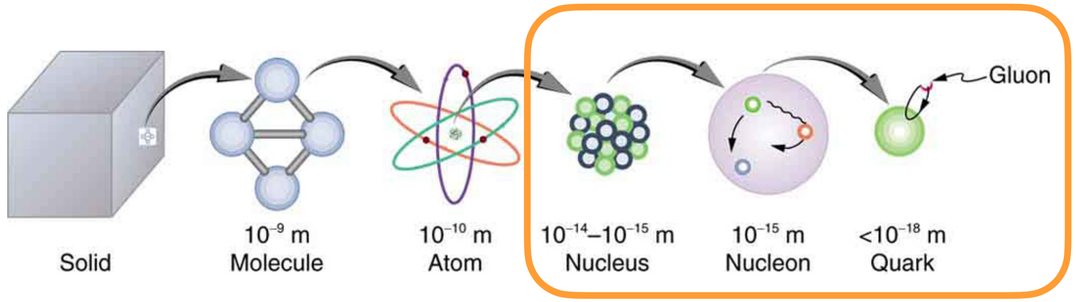
Je n'ai fait celle-ci plus longue que parce que je n'ai pas eu le loisir de la faire plus courte.

If I had more time, I would have written a shorter letter.

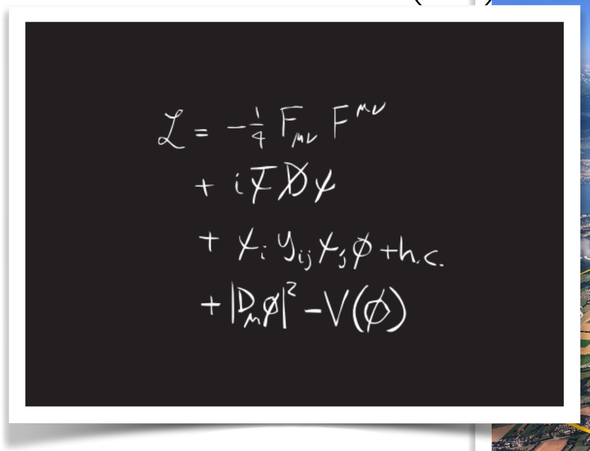
from Lettres Provinciales, by B. Pascal.

High energy physics

High energy physics: the description of fundamental particles and their interactions.



Standard Model (SM)



Colliders (e.g. LHC, Tevatron, HERA, and many others)