THEORY AND METHODOLOGY:
NEW DEVELOPMENTS

The NNPDF collaboration & N3PDF team

Amsterdam-Barcelona-Cambridge-Edinburgh-INFN-Milan-NIKHEF

PDF4LHC meeting

CERN, March 22, 2021

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 740006

THE PATH TO 1%

THE NNPDF COLLABORATION & N3PDF TEAM

AMSTERDAM-BARCELONA-CAMBRIDGE-EDINBURGH-INFN-MILAN-NIKHEF

PDF4LHC MEETING

CERN, MARCH 22, 2021

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 740006

SUMMARY

THEORY DEVELOPMENTS

- ELECTROWEAK CORRECTIONS **CHRISTOPHER SCHWAN** (MILAN)
- NUCLEAR AND DEUTERON UNCERTAINTIES **ROSALYN PEARSON** (EDINBURGH)

PDF PROPERTIES AND THEIR IMPLEMENTATION

- PDF POSITIVITY (TH) **FELIX HEKHORN** (MILAN)
- POSITIVITY (PH), INTEGRABILITY, ARCHITECTURE & BASIS **TOMMASO GIANI** (NIKHEF)

PDF DETERMINATION METHODOLOGY

- HYPEROPTIMIZATION AND K-FOLDING **JUAN CRUZ-MARTINEZ** (MILAN)
- METHODOLOGY CORRELATIONS **ROY STEGEMAN** (MILAN)

PDF VALIDATION

- CLOSURE TESTS **MICHAEL WILSON** (EDINBURGH)
- FUTURE TESTS **JUAN CRUZ-MARTINEZ** (MILAN)

DATASET SELECTION

- WEIGHTED PDFS **ZAHARI KASSABOV** (CAMBRIDGE)

PDF DELIVERY

- PDF REPLICA COMPRESSION **TANJONA RABEMANANJARA** (MILAN)

THEORY DEVELOPMENTS

NLO EW corrections for PDF fits

Ingredients?

- ✓ At least QED corrections in DGLAP
- ✓ Non-zero photon (lepton, . . .) PDF
- (✗) NLO EW (+ NLO QCD+EW) corrections for all PDF processes

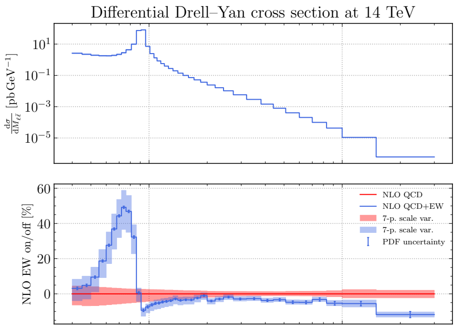
Motivation?

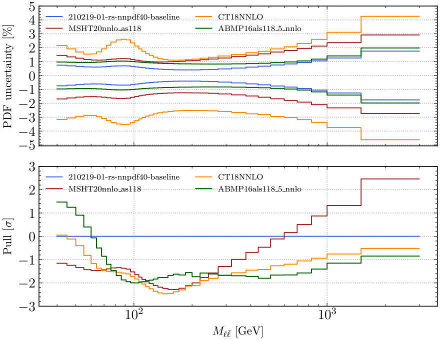
- Below 1 % NLO EW will matter!
- Cuts can be relaxed: DY large-mass region, large Z p_{T}, . . .
- User demand

Problems to be solved

- ✓ Need corrections in the form of interpolation grids: PineAPPL, interfacing with Madgraph5_aMC@NLO, see [S. Carrazza, E.R. Nocera, C.S., M. Zaro]
- ✓ Careful selection of data: no subtraction of FSR, no photon-initiated subtraction, proper observable definition see my LHCEWWG 2020 talk
- ✗ Write/verify runcards and run them (WIP)
- ✗ Implement changed data (WIP)
- → Run fit

Example: Drell–Yan @ 14 TeV





- binning from CMS DY @ 13 TeV: arXiv:1812.10529
- FSR distort the Z peak, weak corrections in the large-mass region

- PDF uncertainties for NNPDF4.0, CT18NNLO, MSHT20, and ABMP16

Deuteron and nuclear uncertainties

We use an **uncertainty** rather than a correcting the central value.

Theory covariance formalism previously developed in NNPDF.

Theory covariance matrix [Ball, Nocera, Pearson: Eur.Phys.J.C 79 (2019) 3, 282 & Eur.Phys.J.C 81 (2021) 1, 37]

$$S_{ij} = \frac{1}{N_{rep}} \sum_{k}^{N_{rep}} \Delta_i^{(k)} \Delta_j^{(k)} \tag{1}$$

$$\Delta_i^{(k)} = T_i^N[f_N^{(k)}] - T_i^N[f_p] \tag{2}$$

Deuteron: NNLO deuteron PDFs fitted in NNPDF methodology
Heavy nuclear: NLO heavy nuclear PDFs from nNNPDF2.0 [Abdul Khalek et al.: JHEP 09 (2020) 183]

Table: χ^2 per deuteron/nuclear dataset

Fit	**Total**	BCDMS d	SLAC d	NMC p/d	E866/NuSea p/d	E605 Cu	NuTeV Fe	CHORUS Pb
NNPDF4.0	**1.174**	1.015	0.4972	0.8194	0.3971	0.4907	0.4602	0.9372
No nuc unc	**1.265**	1.313	0.8217	0.8167	0.8195	1.154	0.4569	1.165

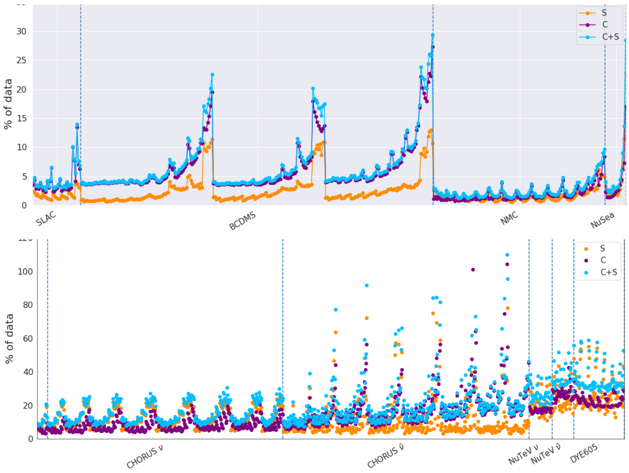
Per-point uncertainties

Deuteron (top) and heavy nuclear (bottom)

C: experimental uncertainties

S: theory uncertainties (nuclear)

C+S: total



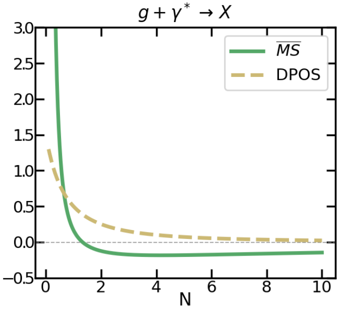
PDF PROPERTIES AND THEIR IMPLEMENTATION

Positivity - Theory

A. Candido, S. Forte, F. Hekhorn [JHEP 11 (2020) 129]
Can $\overline{\mathrm{MS}}$ parton distributions be negative?

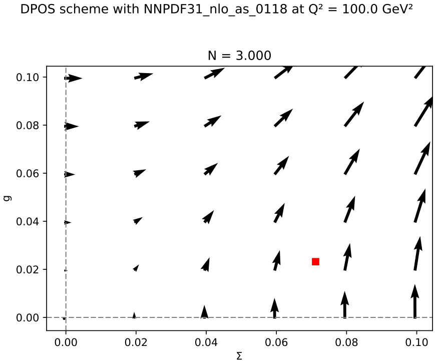
No!

DIS: $F = \sum_j c_j \otimes f_j$
LO: Structure Function = PDF ✓
gluon@NLO:



- $c_g^{(1),bare}(z, Q^2, \epsilon) > 0$ ✓
- $c_g^{(1),\overline{\mathrm{MS}}}(z) < 0$ for $z \to 1$
- $c_g^{(1),\mathrm{DPOS}}(z) > 0$ ✓
- $c_g^{(1),\mathrm{DIS}}(z) = 0$ ✓

Positivity - Theory



DPOS → $\overline{\mathrm{MS}}$: for $z < 1$ use perturbativity, for $z \to 1$ use *exact* transformation ⇒ $\overline{\mathrm{MS}} > 0$

To obtain a physical xs/PDF is positivity **sufficient?** no!

or even **necessary?** no!

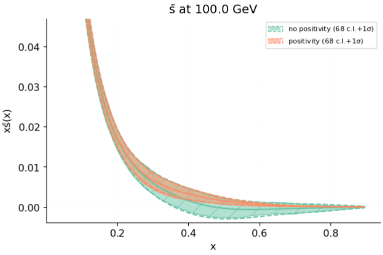
⇒ but we can prove they *are* positive and so it adds a cut in PDF space!

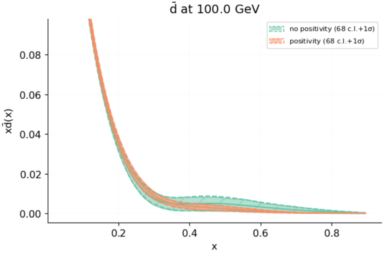
Positivity - Implementation

Quarks, anti-quarks and gluon $\overline{MS}$ PDFs q_k have to be positive: we add a term in the χ^2 penalizing negative distributions

$$\chi^2_{tot} = \chi^2_{exp} + \sum_k \chi^2_{k,\mathrm{pos}} \,,$$

$$\chi^2_{k,pos} = \Lambda_k \sum_i \Theta\left(-q_k\left(x_i, Q^2\right)\right) \,, \quad \text{with} \quad \Theta\left(t\right) = \begin{cases} t & \text{if} \quad t > 0 \\ 0 & \text{if} \quad t < 0 \end{cases} \,.$$





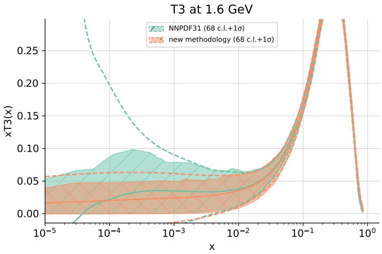
Integrability

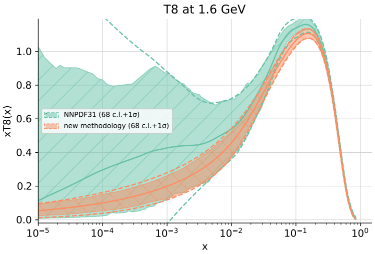
In order to satisfy valence and Gottfried sum rules the distributions $q_k = V, V_3, V_8, T_3, T_8$ have to be integrable at small-x

$$\lim_{x\to 0} xq_k\left(x, Q_0^2\right) = 0\,.$$

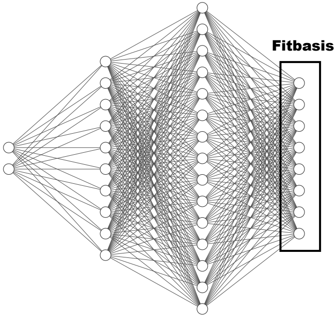
Similarly to what done for positivity, we add to the total χ^2 a penalty of the form

$$\chi^2_{k,integ} = \Lambda_k \sum_i \left[x_i\, q_k\left(x_i, Q^2\right)\right]^2\,.$$





Fitbasis



Flavour basis:
$g, u, \bar{u}, d, \bar{d}, s, \bar{s}, c$

Evolution basis:
$g, \Sigma, V, V_3, V_8, T_3, T_8, T_{15}$

- independently on the basis choice the same physical constraints have to be satisfied: positivity and integrability
- NNPDF4.0 will be hyper-optimized in the evolution basis
- the final results should not depend on the details of the methodology → fitbasis independence studies

PDF DETERMINATION METHODOLOGY

Key differences with respect to the 3.1 methodology

NNPDF 3.1 code	NNPDF 4.0 code
→ **Genetic Algorithm optimizer**	→ **Gradient Descent optimization**
→ One network per flavour	→ One network for all flavours
→ Physical constraints imposed independently of optimization	→ Physical constraints integrated in the optimization
→ Preprocessing fixed per each of the replicas	→ Preprocessing can be fitted within replicas
→ C++ monolithic codebase	→ Python object oriented codebase
→ In-house Machine Learning optimization framework	→ Freedom to use external libraries (default: TensorFlow)
→ Fitting times of up to various days	→ Results available in less than an hour
↓	↓
Fit parameters manually chosen (manual optimization of hyperparameters)	**Fit parameters chosen automatically (hyperparameter scan)**

Beyond the PDF fit: fitting the methodology

The main objective of NNPDF is to minimize choices that can bias the PDF:

- ✗ Functional form ⟶ Neural Networks
- ✗ However: NN are defined by set of parameters!

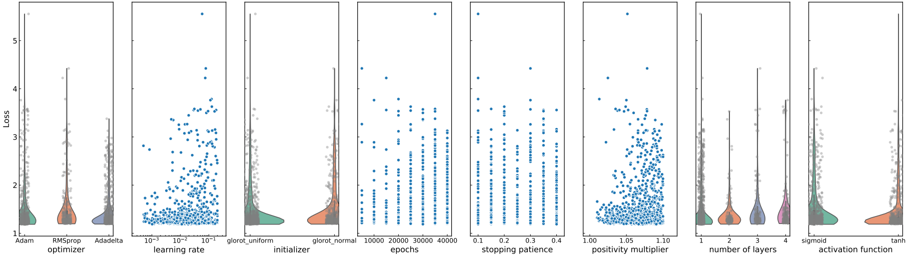
Humans are good at recognising patterns but selecting the best set of parameters is a slow process and systematic success is not guaranteed

To overcome this selection problem we implement a hyperparameter scan: let the computer decide automatically

- ✓ Scan over thousands of hyperparameter combinations
- ✓ Define a reward function to grade the model
- ✓ Check the generalization power of the model

Hyperparameter scan

Each blue dot corresponds to a fit of a different set of hyperparameters:



Thousands of fits for the hyperoptimization algorithm to choose:

- ✓ Optimizer
- ✓ Initializer
- ✓ Stopping Patience
- ✓ Number of Layers
- ✓ Learning Rate
- ✓ Epochs
- ✓ Positivity Multiplier
- ✓ Activation Function

Hyperoptimization: reward and generalization

If we use as hyperoptimization target the χ^2 of the fitted data, we risk finding the hyperparameter set that better overfits.

We avoid this problem by adopting ***k*-folding**:

- Divide the data into k sets.
- Leave one set out and fit the $k-1$ sets left.
- Optimize the average χ^2 of the k non-fitted sets.

Example of function to hyperoptimize:

$$\text{Loss}(\mathit{optimizer\_name},\ \mathit{depth\_of\_network}) = \frac{1}{k}\sum_{k}^{i}\frac{\chi_i^2}{N_i}$$

Where we are computing the χ^2 for the data that did not enter the fit. This ensures that the methodology can accommodate well even data that has never been seen by the fit.

Self-correlation of PDF sets

Is the data-induced correlation between PDF sets based on the same NNPDF methodology and underlying data 100%?

We calculate the data-induced correlation using PDF pairs fitted to the same data replicas.

No, the data-induced correlation is not 100%. This is a result of uncorrelated functional uncertainties.

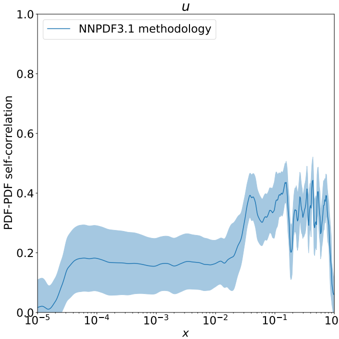


Fig: PDF-PDF self-correlation between two PDF sets based on the same NNPDF methodology and data, but different (random) initialization.

Self-correlation of PDF sets

Is the data-induced correlation between PDF sets based on the same NNPDF methodology and underlying data 100%?

We calculate the data-induced correlation using PDF pairs fitted to the same data replicas.

No, the data-induced correlation is not 100%. This is a result of uncorrelated functional uncertainties.

If the data-induced correlation is higher, this means the functional uncertainty is smaller if compared to the data uncertainty.

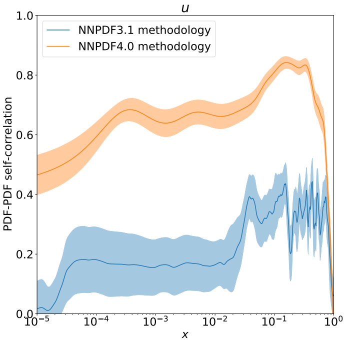


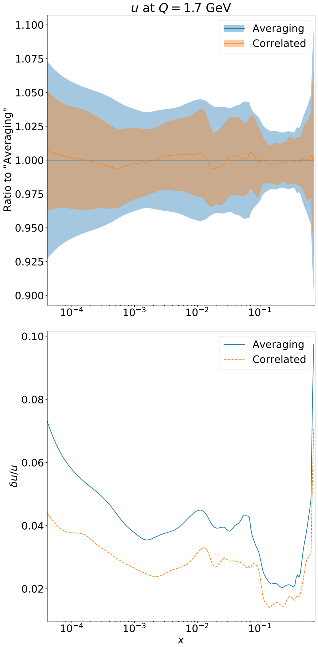
Fig: PDF-PDF self-correlation between two PDF sets based on the same NNPDF methodology and data, but different (random) initialization.

Combination of PDF sets

At present the PDF4LHC combination method is a simple averaging.

Can we achieve a more precise results by including underlying data-induced correlations?

No, this can lead to arbitrarily small uncertainties.



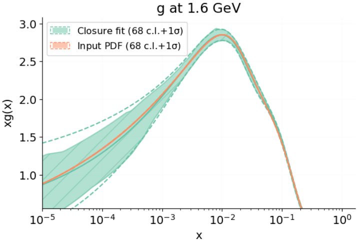
PDF VALIDATION

Closure Tests

Fit replicas to pseudodata in usual way

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{f} + \boldsymbol{\eta} + \boldsymbol{\epsilon} \\ &= \boldsymbol{z} + \boldsymbol{\epsilon},\end{aligned} \tag{3}$$

where $\boldsymbol{\eta} \sim \mathcal{N}(0, C)$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, C)$ are sampled independently.
Use predictions from an input PDF as proxy for $\boldsymbol{f}$.



Example closure fit and input PDF.

Allows testing of methodology, if the input assumptions hold.

For example:

Bias: difference between central prediction and true observable

Variance: uncertainty of replica predictions

Bias is a stochastic variable. If PDF uncertainty is faithful then

$$\mathbf{E}_{\eta}[\text{bias}] = \text{variance} \tag{4}$$

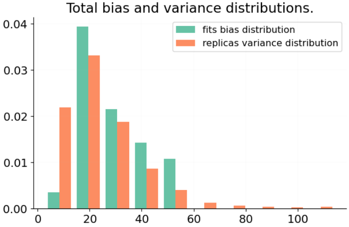
High demand on resources - made feasible with next generation fitting code.

Preliminary results

Compare first moments:

	$\sqrt{\mathbf{E}_\eta[\mathrm{bias}]/\mathbf{E}_\eta[\mathrm{variance}]}$
Total	1.11 ± 0.5

Alternatively look at the respective distributions

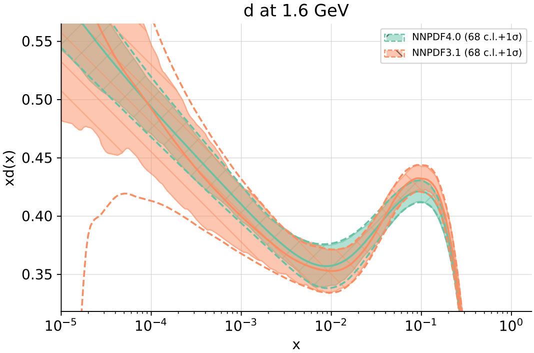


Bias distribution sampled with 25 fits, 40 replicas each.

How can we future-proof the methodology?

Do we trust our errorbands?

The smaller error bands in the NNPDF4.0 fits are driven both by the increased amount of data and the improved methodology. But there are still kin. regions not covered by data!



Ideally: design an experiment for the regions not covered by fitted-data!

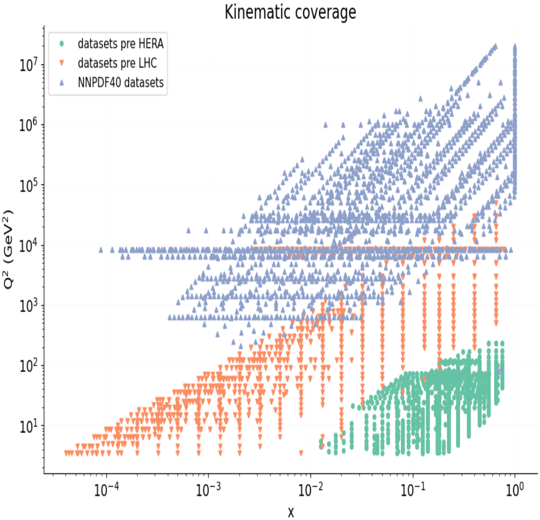
Problem: we want the results before 2050...

Fig: Other valid and certified future-testing methods

Solution: chronologically ordered subsets of data to test unseen regions, we named this “future tests“.

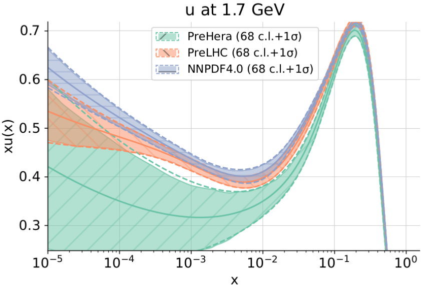
Future tests

for more information see arxiv:2103.08606



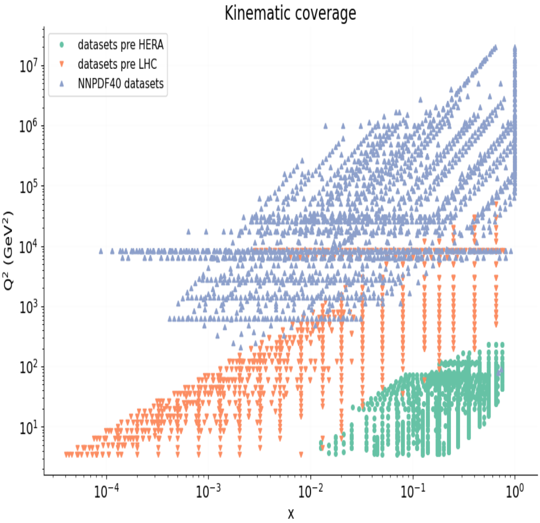
χ^2/N (only exp. covmat)

(dataset)	NNPDF4.0	pre-LHC	pre-Hera
pre-HERA	1.09	1.01	0.90
pre-LHC	1.21	1.20	**23.1**
NNPDF4.0	1.29	**3.30**	**23.1**



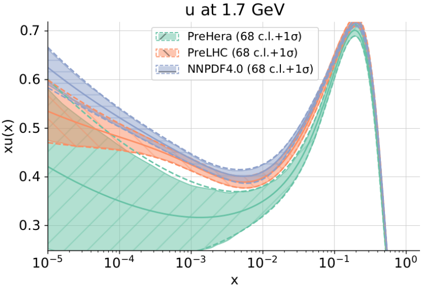
Future tests

for more information see arxiv:2103.08606



χ^2/N (exp. and PDF covmat)

(dataset)	NNPDF4.0	pre-LHC	pre-Hera
pre-HERA			0.86
pre-LHC		1.17	**1.22**
NNPDF4.0	1.12	**1.30**	**1.38**



DATASET SELECTION

Why can't we fit a dataset?

Three possibilities:

1. Problems with the experiment
2. Problems with the theory (including methodology)
3. Tension with other dataset (i.e. 1. or 2. but for other data)

Objective: Investigate the origin of issue
Tool: Constrained fits; Try to constrain the fit to agree with the dataset under investigation and see what breaks in the process.

Implementation

Idea from *Why α_s Cannot be Determined from Hadronic Processes without Simultaneously Determining the Parton Distributions* [Forte, Z.K., **arxiv:2001.04986**]

Instead of optimizing for the total χ^2, give special attention to the agreement of the dataset under investigation by giving more weight to its error χ^2_p.

$$\chi^2 + w\chi^2_p$$

where w is *big*.

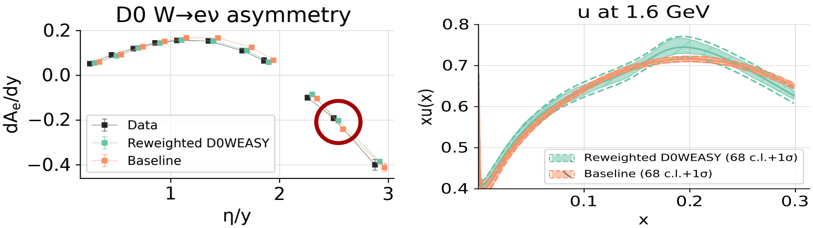
Consequences, compared to optimizing for χ^2:

- Total χ^2 will go up, because we are not optimizing for it any longer.
  - Observe which datasets get worse and how much: Assess **3**.
- Dataset error, χ^2_p will go down.
  - Observe if we can reasonably get a good agreement or the dataset is not self consistent. Conclude **1** or **2**, but likely **1** as our methodology is *very* flexible

Example: DO electron Asymmetry

We cannot fit the D0 electron assymetry dataset. Set $w = 411$.

χ^2/ndat	Baseline	Reweighted D0EASY
D0 e ASY	5.3	1.7
D0 μ ASY	2.0	5.4
Total	1.17	1.29



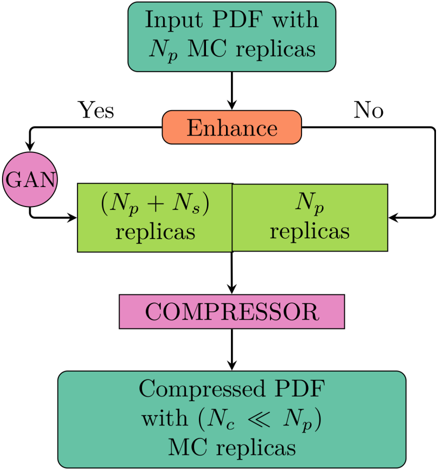
- Can lift the downward prediction but only at the cost of:
  - Introducing unnatural shapes
  - Increasing error in other datasets particularly D0 μ asymmetry.
- The large weight fit still obtains poor fit quality for D0EASY:
  - **Dataset not self consistent**.

PDF DELIVERY

A new methodology with GANs

S. Carrazza, J. Cruz-Martinez, T. Rabemananjara

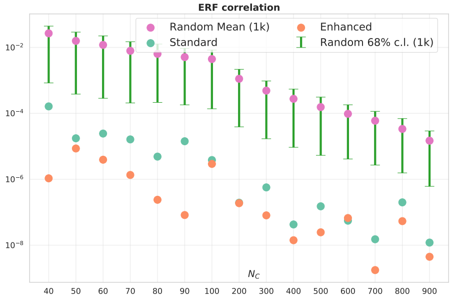
Goal: Provide a smaller set of MC replicas that best represents the Probability Distribution of a given PDF set with large samples.

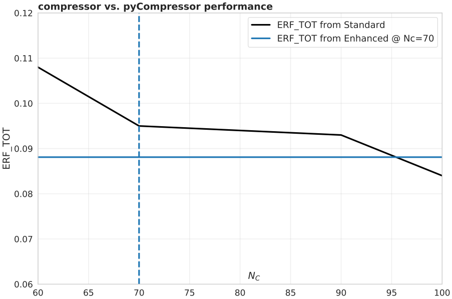

Input PDF with N_p MC replicas
Enhance
Yes
No
GAN
$(N_p + N_s)$ replicas
N_p replicas
COMPRESSOR
Compressed PDF with $(N_c \ll N_p)$ MC replicas

Compression results

Standard vs. GAN-Enhanced Compressor:

SETUP: $(N_p = 1000 + N_s = 2000) \longrightarrow N_c$





$$N_c(\text{GAN-Enhanced}) = 70 \sim N_c(\text{Standard}) = 95$$

OUTLOOK

- PROGRESS ⇒ ALL ASPECTS OF PDF DETERMINATION;

  FROM THEORY TO FINAL DELIVERY

- SIZABLE IMPACT ON PDF DETERMINATION:
  - SUBSTANTIAL SPEED-UP, BY MORE THAN FACTOR 10
  - INCREASED ACCURACY
  - SUBSTANTIAL INCREAS IN PRECISION

    ⇒ SMALLER PDF UNCERTAINTIES

- BIG IMPACT ON NNPDF4.0 ⇒ NEXT GENERATION PDF SET

  ⇒ SEE **E. NOCERA**'s talk

EXTRAS

EW corrections in PDF fits

$$\frac{\mathrm{d}\sigma_{ab}}{\mathrm{d}\mathcal{O}}(x_1, x_2, Q^2, \mathcal{O}) = \sum_{m,n} \alpha_\mathrm{s}^m(Q^2)\alpha^n \frac{\mathrm{d}\sigma_{ab}^{(m,n)}}{\mathrm{d}\mathcal{O}}(x_1, x_2, Q^2, \mathcal{O})$$

→ only lowest order in α included in PDF fits

Known:

- LO QED (+N^{2}LO QCD) PDFs are available (NNPDF, MMHT):
  - non-zero photon PDF: LUXQED [A. Manohar, P. Nason, G. P. Salam, G. Zanderighi], [A. Manohar, et al.]
  - lepton PDFs: LUXlep [L. Buonocore, P. Nason, F. Tramontano, G. Zanderighi]
  - QED effects in the DGLAP equation
- (some) QED effects subtracted in data

Unknown:

- EW corrections are not (systematically) included in PDF fits; only NNLO QCD corrections

→ Inclusion of fully differential NLO EW corrections (no K factors) for all PDF processes

Why include NLO EW corrections?

- NNLO QCD+NLO EW more accurately than plain NNLO QCD
- Do we need NLO EW corrections in PDF fits/Is LO QED enough?

→ Probably not now, but certainly in the future!

- Already now: NNPDF cuts off a few observables because of large EW corrections (e.g. $M_{\mathrm{e}\bar{\mathrm{e}}} \leq 210\,\mathrm{GeV}$ for the shown ATLAS measurement)

Programme: inclusion of NLO EW

- allows inclusion of observables with large EW corrections:
  DY: large $M_{\ell\bar{\ell}}$, Z boson: large p_{T}, ...
- more accurate description of observables: effects on PDFs?

→ systematic calculation of all corrections (including QCD–EW), for all processes
"the best PDFs demand the best predictions" as interpolation grids

→ demands a more consistent data treatment, study double-counting issues

Interpolation grids (I)

For PDF fitting we need PDF independent predictions. Use Lagrange interpolation,

$$f_a(x_1, Q^2) f_b(x_2, Q^2) \approx \sum_{i,j,k} f_a(x_i, Q_k^2) f_b(x_j, Q_k^2) L_i(x_1) L_j(x_2) L_k(Q^2),$$

with Lagrange polynomials L_i over the 3D grid $\left\{ (x_i, x_j, Q_k^2) \right\}_{i,j,k}$. Insert into master formula:

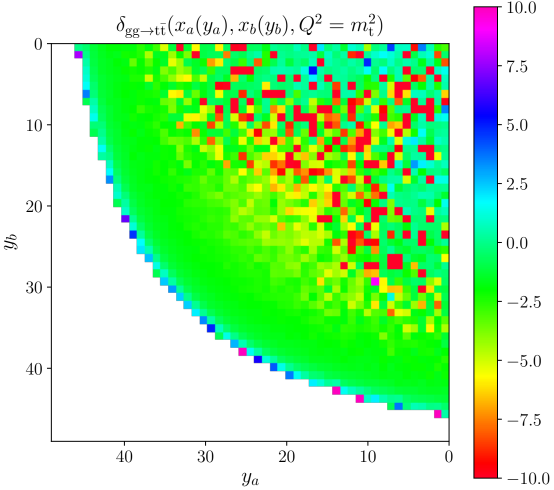
$$\begin{aligned} \frac{\mathrm{d}\sigma}{\mathrm{d}\mathcal{O}} &= \sum_{a,b} \int_0^1 \mathrm{d}x_1 \int_0^1 \mathrm{d}x_2 \int_{Q_{\text{min}}^2}^{Q_{\text{max}}^2} \mathrm{d}Q^2 \, f_a(x_1, Q^2) f_b(x_2, Q^2) \, \frac{\mathrm{d}\sigma_{ab}}{\mathrm{d}\mathcal{O}}(x_1, x_2, Q^2, \mathcal{O}) \\ &= \sum_{a,b} \sum_{i,j,k} \sum_{m,n} f_a(x_i, Q_k^2) f_b(x_j, Q_k^2) \alpha_{\mathrm{s}}^m(Q^2) \alpha^n \, \frac{\mathrm{d}\Sigma_{abijkmn}}{\mathrm{d}\mathcal{O}} \end{aligned}$$

where

$$\frac{\mathrm{d}\Sigma_{abijkmn}}{\mathrm{d}\mathcal{O}} = \int_0^1 \mathrm{d}x_1 \int_0^1 \mathrm{d}x_2 \int_{Q_{\text{min}}^2}^{Q_{\text{max}}^2} \mathrm{d}Q^2 \, L_i(x_1) L_j(x_2) L_k(Q^2) \, \frac{\mathrm{d}\sigma_{ab}^{(i,k)}}{\mathrm{d}\mathcal{O}}(x_1, x_2, Q^2, \mathcal{O})$$

→ generate $\frac{\mathrm{d}\Sigma_{abijkmn}}{\mathrm{d}\mathcal{O}}$ once, perform PDF convolutions very quickly off-line

Example: $\Sigma_{\mathrm{gg}ij021}/\Sigma_{\mathrm{gg}ij020}$, $\mathcal{O}(\alpha_\mathrm{s}^2\alpha)/\mathcal{O}(\alpha_\mathrm{s}^2)$ for $\mathrm{gg} \to \mathrm{t\bar{t}}$ @ 8 TeV



- no interpolation in y_a, y_b, or Q^2
- correction for ixs roughly $-0.5\,\%$
- $y_{a/b}(x) = -\ln x_{a/b} + 5(1 - x_{a/b})$, $y(1) = 0$
- lower left corner → production threshold
- at threshold: Coulomb singularity
- $y_a \leftrightarrow y_b$ symmetry: initial-state symmetry of $\mathrm{gg} \to \mathrm{t\bar{t}}$
- negative correction for larger x_a, x_b

Interpolation grids (II)

- Interpolation grids are an old idea:
  - APPLGRID [T. Carli et al.]
  - FASTNLO [T. Kluge, K. Rabbertz, M. Wobisch]
- data generation with
  - AMCFAST [V. Bertone, R. Frederix, S. Frixione, J. Rojo, M. Sutton] (MG5_AMC@NLO V2+APPLGRID) or
  - MCGRID [L.D. Debbio, N.P. Hartland, S. Schuhmann] (SHERPA+APPLGRID/FASTNLO)
  - dedicated MCs: MCFM, NLOJET++
- NNPDF uses APPLGRID
- None of the above support EW corrections
- APPLGRID is slow and difficult to use; FASTNLO has complicated interface

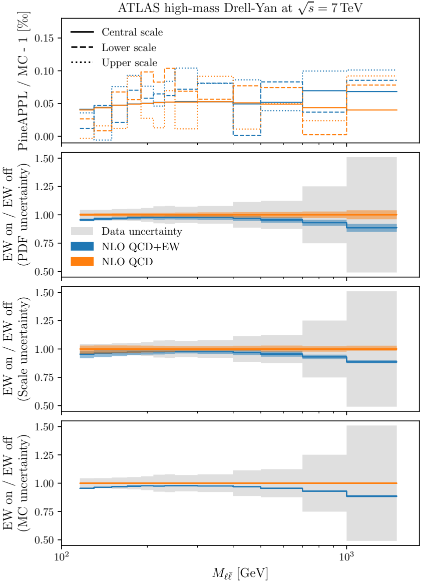
→ We wrote PINEAPPL (PINEAPPL Is Not an Extension of APPLGRID)

- supports arbitrary fixed-order calculations
- easily supports distributions with more than 1000 bins
- interfacing with
  - MG5_AMC@NLO V3.0.4 complete, will be released soon
  - SHERPA+MCGRID in progress
  - your custom MC (should be) easily possible!

PineAPPL

- [S. Carazza, E.R. Nocera, C. Schwan, M. Zaro]
- interpolation error typically sub-per mille (see right)
- off-line PDF uncertainty calculation in a few seconds
- command-line program to quickly produce predictions
- simply add: `set pineappl True` to `mg5_aMC` runcard
- `C`, `Python`, `Rust` interfaces available
- size of each order, partonic channels, observable
- support for 1D, 2D, 3D, . . . , nD distributions

`https://n3pdf.github.io/pineappl`



Process list for NNPDF4.0

Calculate theoretical predictions for the entire NNPDF4.0 dataset:

1. DIS: NMC, SLAC, BCDMS, CHORUS, NUTEV, HERA
2. Fixed-target Drell–Yan experiments
3. Collider experiments (CDF, DØ, ATLAS, and CMS)
   - Drell–Yan Z and $\mathrm{W}^{\pm}$
   - dijets ATLAS 7 TeV and CMS 7 TeV and 8 TeV (NEW)
   - single-inclusive jet production ATLAS 8 TeV (NEW); peculiar observable,

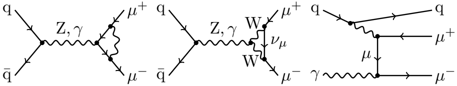
     see [M. Cacciari, S. Forte, D. Napoletano, G. Soyez, G. Stagnitto]
   - Top-pair production
   - Z transverse momentum
   - $\mathrm{W}^{\pm} + \mathrm{c}$ production (at NLO only)
   - W + jets (NEW)
   - single top t-channel production (NEW)
   - diphoton production (NEW)

→ for the time being only collider experiments; ($N_{\text{dat}} \approx 1200$)

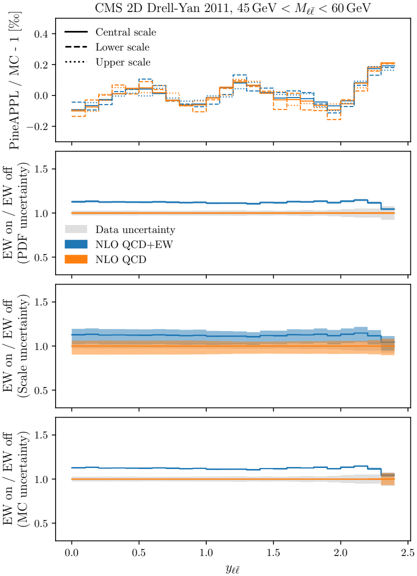
→ write runcards, validate

→ explore double-counting issues, problems with data

Drell–Yan: NNPDF4.0 dataset

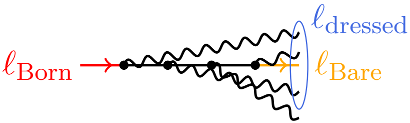


NNPDF ID	$\sqrt{s}$	NNPDF ID	
CDFZRAP	1.96 TeV	CMSWEASY840PB	7 T
D0ZRAP	1.96 TeV	CMSWMASY47FB	7 T
D0WMASY	1.96 TeV	CMSDY2D11	7 T
ATLASWZRAP36PB	7 TeV	CMSWMU8TEV	8 T
ATLASZHIGHMASS49FB	7 TeV	LHCBZ940PB	7 T
ATLASLOMASSDY11EXT	7 TeV	LHCBZEE2FB	8 T
ATLASWZRAP11CC	7 TeV	LHCBWZMU7TEV	7 T
ATLASWZRAP11CF	7 TeV	LHCBWZMU8TEV	8 T
ATLAS_DY2D_8TEV	8 TeV	LHCB_Z_13TEV_DIMUON	13 T
ATLAS_WZ_TOT_13TEV	13 TeV	LHCB_Z_13TEV_DIELECTRON	13 T



- Observables: $y_{\ell\bar{\ell}} \approx \frac{1}{2} ln \frac{x_1}{x_2}$ and $M_{\ell\bar{\ell}} \approx \sqrt{x_1 x_2 s}$
- analysis: [CMS Collaboration]
- NLO EW / NLO QCD: 11 %

Double-counting problem: subtraction of FSR



- pre-FSR data/Born leptons: observables of leptons "before they radiate", calculated using photon-shower inversion (PHOTOS), from
- post-FSR data/dressed leptons: observables using leptons with photons recombined around $\Delta R_{f\gamma}$, typically $\Delta R_{f\gamma} = 0.1$

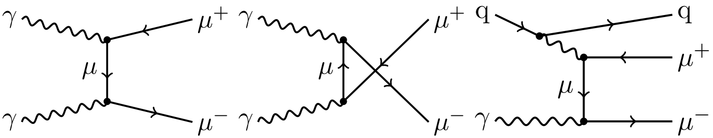
- pre-FSR data for comparisons with QCD-only theory predictions
- post-FSR data for comparisons with EW corrections (up to one photon emission)

- Some experiments—notably CMS—do not publish post-FSR data: double counting issue!
- dressing factors

$$C_{\text{dress}} = \frac{\mathrm{d}\sigma_{\text{post-FSR}}/\mathrm{d}\mathcal{O}}{\mathrm{d}\sigma_{\text{pre-FSR}}/\mathrm{d}\mathcal{O}}$$

  can be large, up to 20 % in invariant mass distributions
- Often C_{dress} (+uncertainty) and pre-FSR dataset given ⇒ need to change systematic uncertainties!

Subtraction of photon–photon contribution

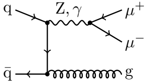


- For ATLAS and CMS it seems to be standard procedure to subtract double-photon induced contributions:

> *The photon-induced process, $\gamma\gamma \to \ell\bar{\ell}$, is simulated at LO using Pythia 8 and the MRST2004qed PDF set.*

- I am not sure why this is done
- This is a problem: proton contains photons, should be counted towards signal!
- Size of the LO contribution can become significant in large-invariant-mass bins (3 %) depending on the used PDF—up to twice as large for pre-LUXQED photon PDFs

Z transverse momentum



$$\mu = M_Z \text{ vs. } \mu = \sqrt{M_Z^2 + (p_T^{\ell\bar{\ell}})^2}$$

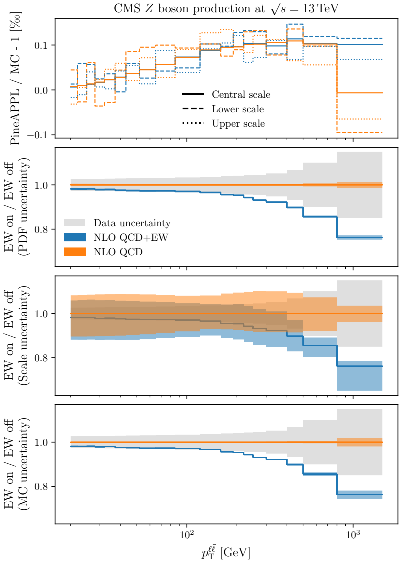
- FSR issues similar to DY
- no photon subtraction

static scale:

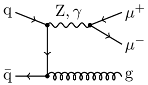
- accidental cancellation of NLO QCD correction → uncertainty band shrinks
- NLO EW are artificially enhanced because of normalisation

dynamic scale:

- scale variation is stabilised
- still significant EW corrections,



Z transverse momentum



$$\mu = M_{\mathrm{Z}} \text{ vs. } \mu = \sqrt{M_{\mathrm{Z}}^2 + (p_{\mathrm{T}}^{\ell\bar{\ell}})^2}$$

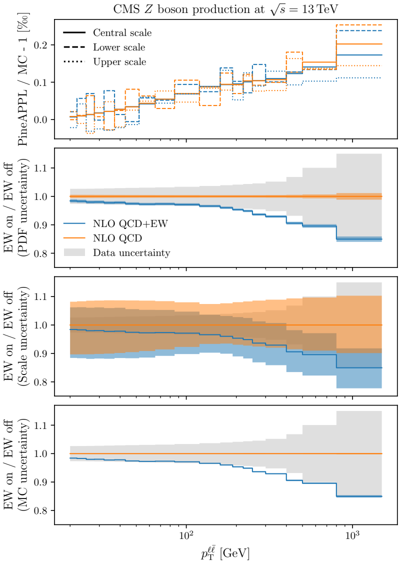
- FSR issues similar to DY
- no photon subtraction

static scale:

- accidental cancellation of NLO QCD correction → uncertainty band shrinks
- NLO EW are artificially enhanced because of normalisation

dynamic scale:

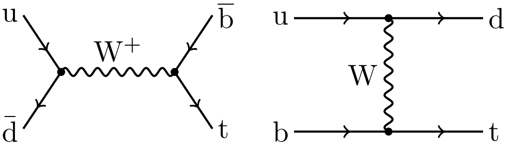
- scale variation is stabilised
- still significant EW corrections,



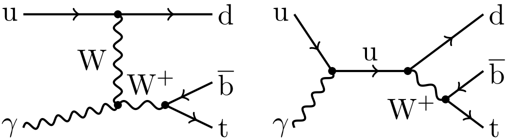
Single-top production

Not properly definable (!?) at NLO EW:

- Analyses, e.g. [ATLAS collaboration], treat *s*-channels as background
- single-production at LO:



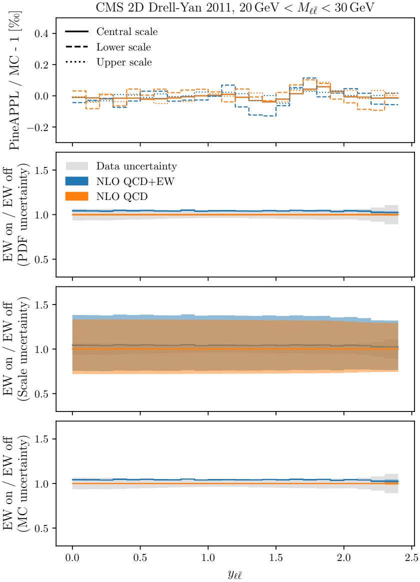
- but at NLO EW not (gauge-invariantly) separable:

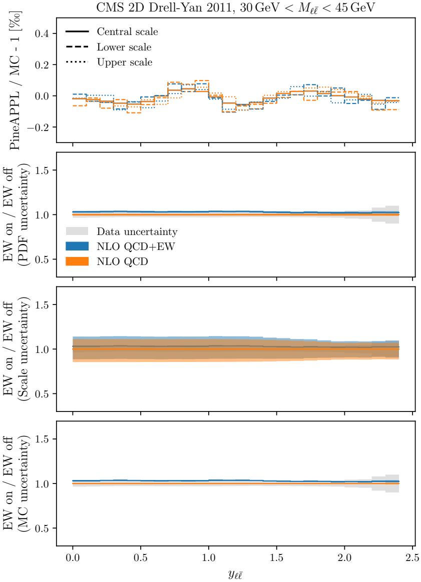


→ ignore these datasets?

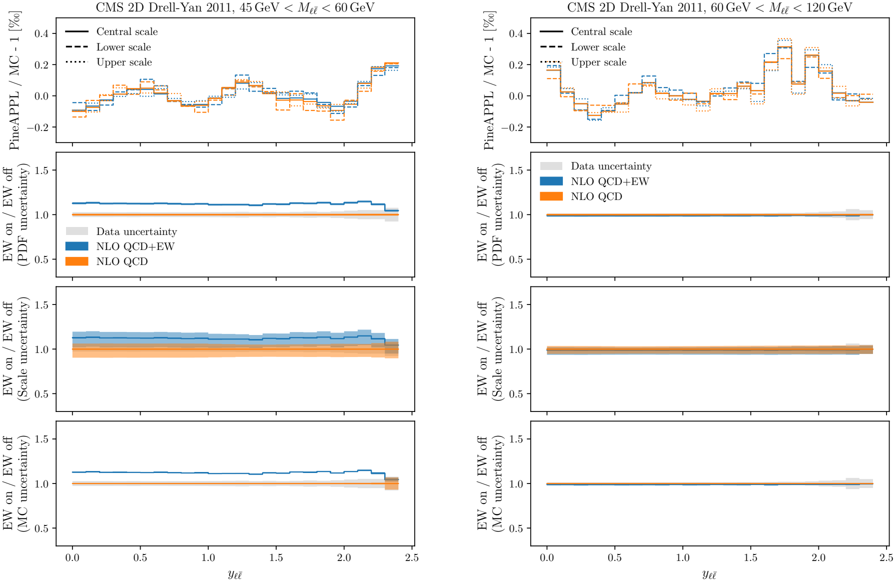
- probably not too important, but see [E.R. Nocera, M. Ubiali, C. Voisey]

CMS DY 2D (I)

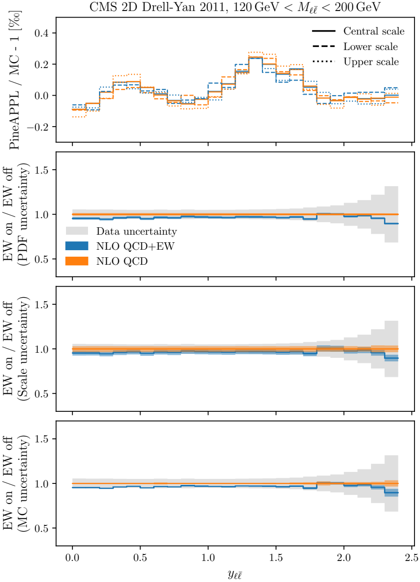

CMS 2D Drell-Yan 2011, 20 GeV < $M_{\ell\bar{\ell}}$ < 30 GeV
PineAPPL / MC - 1 [‰]
0.4
0.2
0.0
−0.2
Central scale
Lower scale
Upper scale
EW on / EW off (PDF uncertainty)
Data uncertainty
NLO QCD+EW
NLO QCD
1.5
1.0
0.5
EW on / EW off (Scale uncertainty)
EW on / EW off (MC uncertainty)
0.0
0.5
1.0
1.5
2.0
2.5
$y_{\ell\bar{\ell}}$

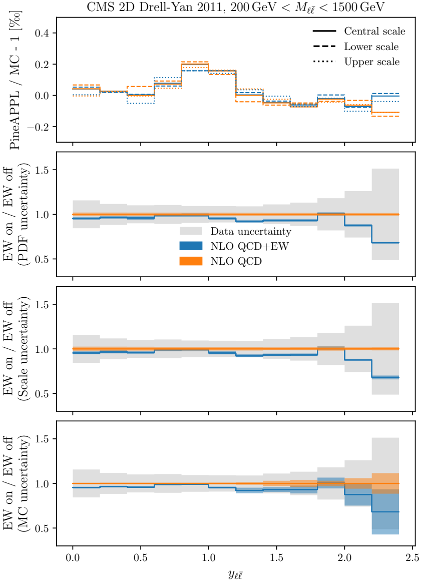

CMS 2D Drell-Yan 2011, 30 GeV < $M_{\ell\bar{\ell}}$ < 45 GeV
PineAPPL / MC - 1 [‰]
0.4
0.2
0.0
−0.2
Central scale
Lower scale
Upper scale
EW on / EW off (PDF uncertainty)
Data uncertainty
NLO QCD+EW
NLO QCD
1.5
1.0
0.5
EW on / EW off (Scale uncertainty)
EW on / EW off (MC uncertainty)
0.0
0.5
1.0
1.5
2.0
2.5
$y_{\ell\bar{\ell}}$

CMS DY 2D (II)


CMS 2D Drell-Yan 2011, 45 GeV < $M_{\ell\bar{\ell}}$ < 60 GeV
CMS 2D Drell-Yan 2011, 60 GeV < $M_{\ell\bar{\ell}}$ < 120 GeV
PineAPPL / MC - 1 [‰]
Central scale
Lower scale
Upper scale
0.4
0.2
0.0
−0.2
EW on / EW off (PDF uncertainty)
EW on / EW off (Scale uncertainty)
EW on / EW off (MC uncertainty)
Data uncertainty
NLO QCD+EW
NLO QCD
1.5
1.0
0.5
0.0
0.5
1.0
1.5
2.0
2.5
$y_{\ell\bar{\ell}}$

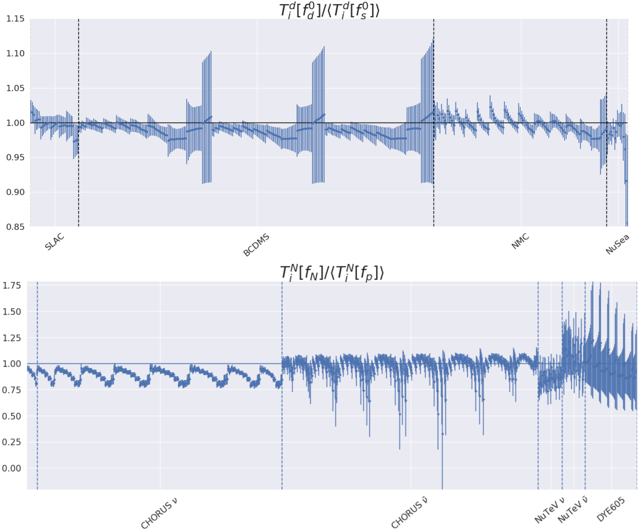
CMS DY 2D (III)


CMS 2D Drell-Yan 2011, 120 GeV < $M_{\ell\bar{\ell}}$ < 200 GeV
PineAPPL / MC - 1 [‰]
Central scale
Lower scale
Upper scale
EW on / EW off (PDF uncertainty)
Data uncertainty
NLO QCD+EW
NLO QCD
EW on / EW off (Scale uncertainty)
EW on / EW off (MC uncertainty)
$y_{\ell\bar{\ell}}$


CMS 2D Drell-Yan 2011, 200 GeV < $M_{\ell\bar{\ell}}$ < 1500 GeV
PineAPPL / MC - 1 [‰]
Central scale
Lower scale
Upper scale
EW on / EW off (PDF uncertainty)
Data uncertainty
NLO QCD+EW
NLO QCD
EW on / EW off (Scale uncertainty)
EW on / EW off (MC uncertainty)
$y_{\ell\bar{\ell}}$

Extra: Nuclear and deuteron observables

Deuteron (top) and heavy nuclear (bottom)



$T_i^d[f_d^0]/\langle T_i^d[f_s^0]\rangle$

$T_i^N[f_N]/\langle T_i^N[f_p]\rangle$

Extra: deuteron uncertainties

Deuteron data

Deuteron only F_2^d: SLAC, BCDMS $\rightarrow T_i^d[f_d]$
Mixed F_2^d/F_2^p & $\sigma_{pd}^{DY}/\sigma_{pp}^{DY}$: NMC, DYE866/NuSea $\rightarrow T_i^d[f_d, f_p]$

Standard is to use isoscalar PDFs in place of deuteron: $f_s \equiv \frac{1}{2}(f_p + f_d)$
Now

$$\Delta_i^{(k)} = \begin{cases} T_i^d[f_d^{(k)}] - T_i^d[f_s^{(0)}] & i \in \text{deuteron only} \\ T_i^d[f_d^{(k)}, f_p^{(0)}] - T_i^d[f_s^{(0)}, f_p^{(0)}] & i \in \text{mixed} \end{cases} \tag{5}$$

Extra: heavy nuclear uncertainties

Heavy nuclear data $T_i^N[f_N]$

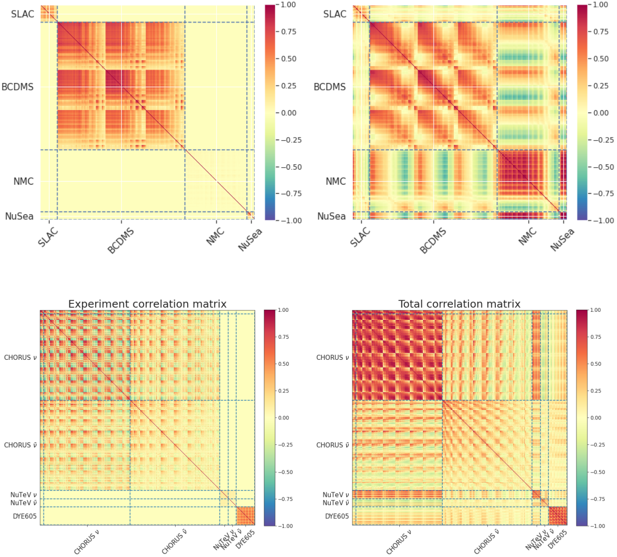
Cu: DYE605, $N = 64$
Fe: NuTeV (& EMC), $N = 56$
Pb: CHORUS, $N = 208$

$$\Delta_i^{(k)} = T_i^N[f_N^{(k)}] - T_i^N[f_p] \tag{6}$$

Where

$$\begin{aligned} T_i^N[f_N] &= \frac{1}{A}(ZT_i[f_{p/N} + (A-Z)T_i[f_{n/N}] \\ T_i^N[f_p] &= \frac{1}{A}(ZT_i[f_p] + (A-Z)T_i[f_n] \end{aligned} \tag{7}$$

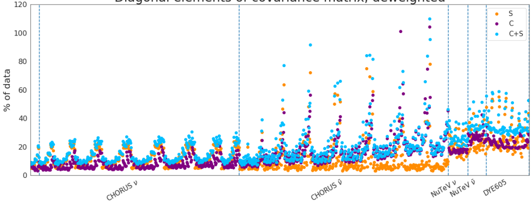
Extra: Correlation matrix



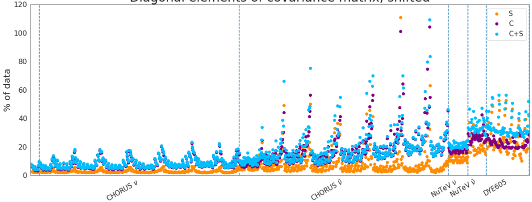
Extra: Nuclear correction

- Can also "shift" the observables $T_i^{(N/d)}[f_(p/s)^{(0)}] \to T_i^{(N/d)}[f_{(N/d)}^{(0)}]$
- Uncertainty is correspondingly reduced

Diagonal elements of covariance matrix, deweighted

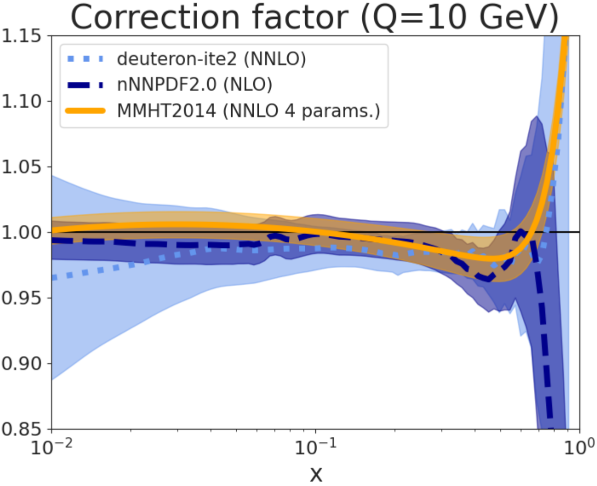


Diagonal elements of covariance matrix, shifted



Extra: Deuteron correction

Comparing deuteron correction to MMHT [Harland-Lang et al.: Eur. Phys. J. C 75(5), 204 (2015)]



Positivity - Theory - Backup

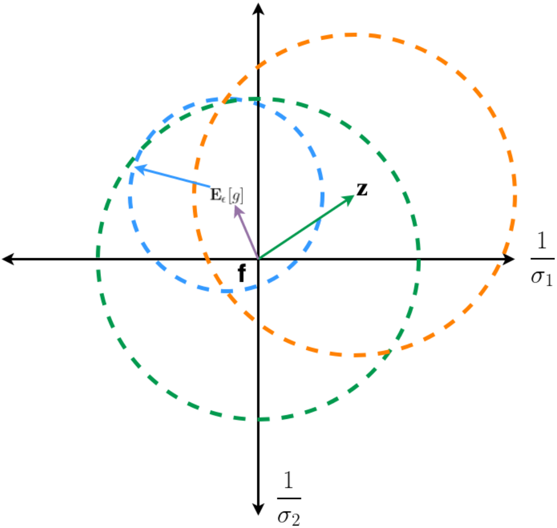
DIS: $F = \sum_j c_j \otimes f_j$
LO: Structure Function = PDF
gluon@NLO: $C_g^{(1),\text{bare}}(z, Q^2, \epsilon) = \frac{\Gamma(-\epsilon)\left(\frac{s}{4\pi\mu^2}\right)^{-\epsilon}[8P_{qg}(z) - 16T_R\epsilon(3-\epsilon(2-\epsilon))]}{16\pi(2-2\epsilon)\Gamma(3-2\epsilon)}$

- $C_g^{(1),\overline{\mathrm{MS}}}(z) = P_{qg}(z)\left(\ln\left(\frac{1-z}{z}\right) - 4\right) + 3T_R$ with $\mu^2_{\overline{\mathrm{MS}}} = Q^2$
- $C_g^{(1),\mathrm{DPOS}}(z) = 3\,[T_R - P_{qg}(z)]$ with $\mu^2_{\mathrm{POS}} = (k_T^{max})^2 = \frac{s}{4} = \frac{Q^2(1-z)}{4z}$
- $C_g^{(1),\mathrm{DIS}}(z) = 0 \Rightarrow$ Structure Function = PDF

quark@NLO: no problem due to *positive* logs (resummation)

Geometric Interpretation

Consider 2 data points on an axis in the basis which diagonalises C normalised by the square root of the eigenvalues:



Statistical estimators - more detail

Decompose the expectation value of the likelihood function, χ^2, by completing the square. Exposing some statistical indicators

$$\begin{aligned} \mathbf{E}_\epsilon[\chi^2(g;y)] &= \frac{1}{N_{\mathrm{data}}}\mathbf{E}_\epsilon[(\boldsymbol{g}-\boldsymbol{y})^T C^{-1}(\boldsymbol{g}-\boldsymbol{y})] \\ &= \mathrm{bias} + \mathrm{variance} + \mathrm{noise} - \mathrm{crossterm} \end{aligned} \tag{8}$$

focus on the first two terms:

$$\mathrm{bias} = \frac{1}{N_{\mathrm{data}}}(\boldsymbol{f}-\mathbf{E}_\epsilon[\boldsymbol{g}])^T C^{-1}(\boldsymbol{f}-\mathbf{E}_\epsilon[\boldsymbol{g}]) \tag{9}$$

$$\mathrm{variance} = \frac{1}{N_{\mathrm{data}}}\mathbf{E}_\epsilon\left[(\boldsymbol{g}-\mathbf{E}_\epsilon[\boldsymbol{g}])^T C^{-1}(\boldsymbol{g}-\mathbf{E}_\epsilon[\boldsymbol{g}])\right] \tag{10}$$

where $\mathbf{E}_\epsilon[\cdot]$ is the expectation across replicas.

- Faithful uncertainties if $(\boldsymbol{g}-\mathbf{E}_\epsilon[\boldsymbol{g}])$ and $(\boldsymbol{f}-\mathbf{E}_\epsilon[\boldsymbol{g}])$ have same distribution.
- Sample $(\boldsymbol{g}-\mathbf{E}_\epsilon[\boldsymbol{g}])$ through sampling ϵ - usual MC replica procedure
- Sample $(\boldsymbol{f}-\mathbf{E}_\epsilon[\boldsymbol{g}])$ distribution through sampling η - only possible in closure test!

Preliminary results

Breakdown of $\mathbf{E}_\eta[\text{bias}]/\mathbf{E}_\eta[\text{variance}]$ for out of sample data by experiment - fitted on NNPDF3.1 dataset and validated on additional datasets to be included in NNDPF4.0

	$\sqrt{\mathbf{E}_\eta[\text{bias}]/\mathbf{E}_\eta[\text{variance}]}$
ATLAS	1.17 ± 0.4
CMS	1.07 ± 0.5
LHCb	0.83 ± 0.6
Total	1.11 ± 0.5

Self-correlation of PDF sets

The PDF$_i$-PDF$_j$ correlation for a given flavour is defined as

$$corr_{i,j}(x) = \frac{\sum_{n=1}^{N_{rep}} (f_{i,n}(x) - f_{i,0}(x))(f_{j,n}(x) - f_{j,0}(x))}{\sqrt{\sum_{n=1}^{N_{rep}} (f_{i,n}(x) - f_{i,0}(x))^2} \sqrt{\sum_{n=1}^{N_{rep}} (f_{j,n}(x) - f_{j,0}(x))^2}}$$

where $f_{i,n}(x)$ is a PDF replica, and $n = 0$ corresponds to the central value of the PDF set.

Correlated combination of PDFs

When combining PDF sets f_i in a correlated way, for each momentum fraction x and flavour, the central value is calculated using

$$\langle f \rangle_{comb} = \sum_{i=1}^{N_{sets}} w_i f_{i,0}$$

and the variance using

$$V_{comb} = \sum_{i,j=1}^{N_{sets}} w_i \sigma_{ij} w_j$$

where σ is the covariance matrix, $f_{0,i}$ is the central value of PDF set i, and w_i are the weights

$$w_i = \frac{\sum_{j=1}^{N_{sets}} \left(\sigma^{-1}\right)_{ij}}{\sum_{k,l=1}^{N_{sets}} \left(\sigma^{-1}\right)_{kl}}$$

Choosing the weight

- Not a critical setting for this kind of study.
    - Typically we see PDFs bending and/or other datasets being poorly fitted.
- Rule of thumb: Have the dataset contributed roughly as much to the total error function as all the of the data (assuming χ^2/ndat=1)

$$w = \text{total ndat}/\text{dataset ndat}$$

e.g. DO e ASY has 11 points and we have 4524 in total, so we set a weight of 411.

Compression & GAN Methodologies

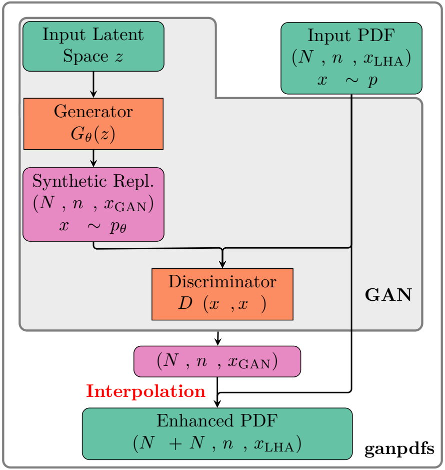
Error Function (ERF):

$$\mathrm{ERF}_k = \frac{1}{N_k} \sum_i \left(\frac{C^k(x_i) - P^k(x_i)}{P^k(x_i)} \right)^2$$

$$\mathrm{ERF}_{\mathsf{ToT}} = \frac{1}{N_{\mathsf{EST}}} \sum_k \mathrm{ERF}_k$$

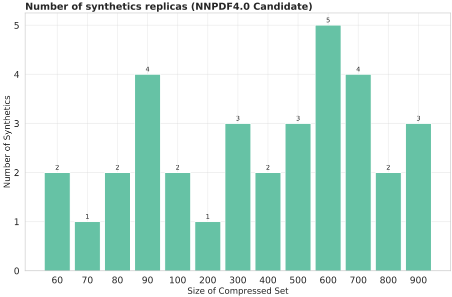
- P^k is the value of the estimator k for the PRIOR
- C^k is the value of the estimator k for the COMPRESSED (either drawn from the PRIOR or ENHANCED set)

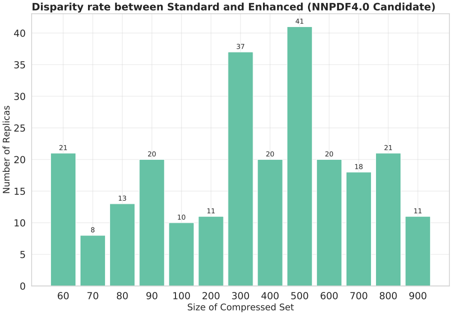
GANPDFs flowchart:


Input Latent Space z
Input PDF $(N\ ,n\ ,x_{\mathrm{LHA}})$ $x\ \sim p$
Generator $G_\theta(z)$
Synthetic Repl. $(N\ ,n\ ,x_{\mathrm{GAN}})$ $x\ \sim p_\theta$
Discriminator $D\ (x\ ,x\ )$
GAN
$(N\ ,n\ ,x_{\mathrm{GAN}})$
Interpolation
Enhanced PDF $(N\ +N\ ,n\ ,x_{\mathrm{LHA}})$
ganpdfs

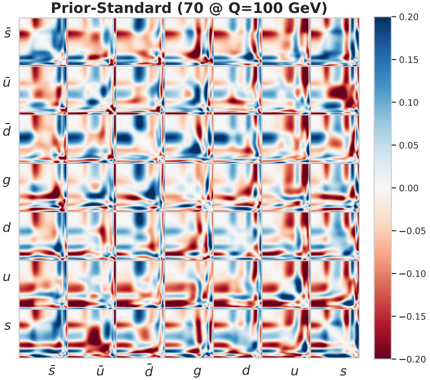
Final compressed sets

Number of Synthetic replicas in the compressed set:

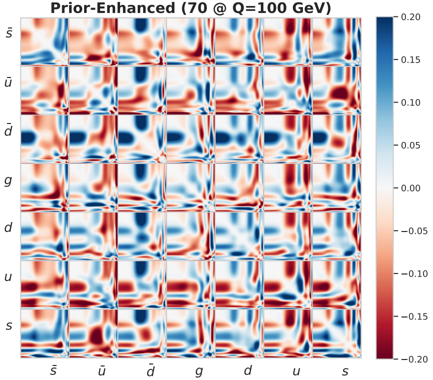




Comparing PDF correlation matrix

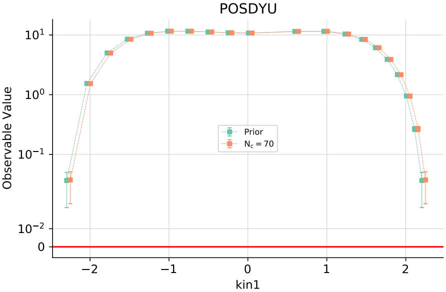


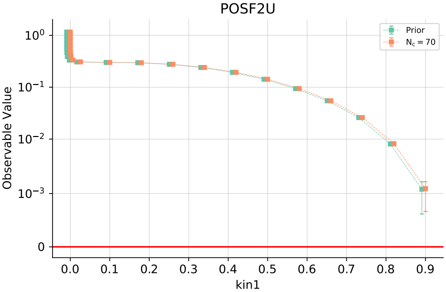
Prior-Standard (70 @ Q=100 GeV)



Prior-Enhanced (70 @ Q=100 GeV)

Positivity Constraints



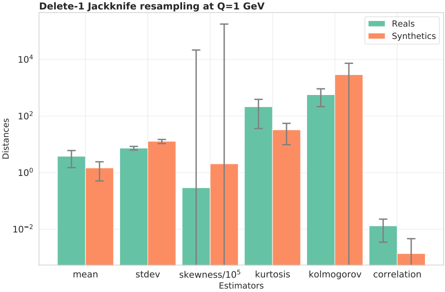


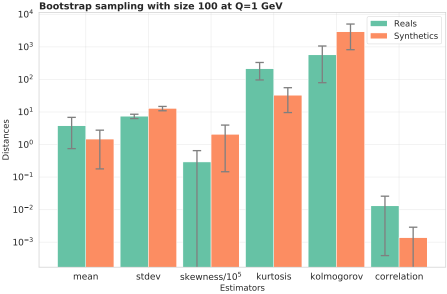
GANs for PDF fitting

Consider 3 different set of fits:

- 2 disjoint fits S_1 and S_2 with $N = 500$ replicas
- GAN fit S_3 with with $N = 500$ replicas determined from $N_0 = 100$

GREEN$\equiv$ ERF(S_1, S_2) and **ORANGE** $\equiv$ ERF(S_1, S_2) using different resampling methodologies.


Delete-1 Jackknife resampling at Q=1 GeV
Reals
Synthetics
Distances
10^4
10^2
10^0
10^{-2}
mean
stdev
skewness/10^5
kurtosis
kolmogorov
correlation
Estimators


Bootstrap sampling with size 100 at Q=1 GeV
Reals
Synthetics
Distances
10^4
10^3
10^2
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
mean
stdev
skewness/10^5
kurtosis
kolmogorov
correlation
Estimators