Using unfolded data in global QCD analyses

Sasha Zenaiev
(Hamburg University)

France-Berkeley Conference on Unfolding
11 Jun 2024

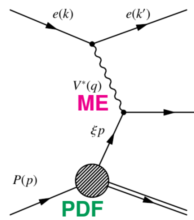
Overview

- In this talk I will discuss how we deal with experimental data in global QCD fits: what lessons we learnt, what issues may arise etc.
- Disclaimer: based on my limited experience with global QCD fits, mainly:
  - fits to *ep* HERA data
    H1 and ZEUS Coll., EPJ C75 (2015) 580 EPJ C78 (2018) 473
  - fits to HERA+LHCb+ALICE data
    PROSA Coll., EPJ C75 (2015) 396 JHEP 04 (2020) 118
  - data analysis in CMS and global fits to HERA+CMS data
    CMS Coll., EPJ C77 (2017) 459 EPJ C80 (2020) 658
  - global fits in the ABMP PDF framework
    Garzelli, Mazzitelli, Moch, Zenaiev JHEP 05 (2024) 321,
    to appear Alekhin, Garzelli, Moch, Zenaiev 24XX.YYYYY

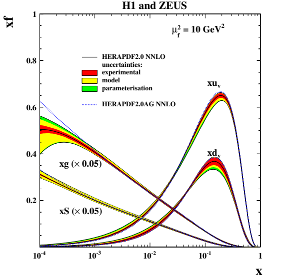
  $\rightarrow$ it is not exhaustive, other groups might have different experience
- Also I will touch some aspects of experimental data analyses
- All these fits were done using open source `xFitter` (former `HERAfitter`) program
  - define theory model with some free parameters
  - select experimental data (uncertainties+correlations are crucial here)
  - compare theory to data and extract best theory parameters

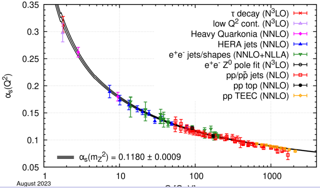
www.xfitter.org
Gitlab page

Global QCD anlyses: $\boxed{\text{PDFs},\alpha_S,\ldots \xleftarrow{\text{pQCD}} d\sigma/dO} \xleftarrow{\text{MC}}$ events



- Factorization theorem: $\boldsymbol{\sigma} = \textbf{PDF} \otimes \textbf{ME}$
- Parton distribution functions (PDFs) $\boldsymbol{f(x, \mu_f)}$ describe distribution of quarks and gluons in hadrons
- Matrix elements (ME) are calculated in perturbative QCD (pQCD) $\sigma = \Sigma_{i=0}^{n} \sigma_i \alpha_S^i$ requiring $\alpha_S(\mu_r) < 1$ ($\mu_r \gg \Lambda_{QCD}$)
- At low scales $\mu \sim 1$ GeV non-perturbative QCD effects are parametrised by PDFs **which are extracted using data**
  - typically shaped like $x^a(1-x)^b$ with a few tens of parameters (but there are different approaches e.g. NNPDF)
- At higher scales $\mu > 1$ GeV PDF evolution is predicted by pQCD
- Other unknown parameters: $\alpha_S(M_Z)$, masses of heavy quarks (fundamental free parameters of Standard Model)
  - can be fitted or fixed in global QCD analyses
- **Challenges:**
  - find suitable PDF parametrization
  - select PDF sensitive and consistent data sets
  - use appropriate statistical method (typically minimizing χ^2)
  - **most challenging are PDF uncertainties:** very much depend on all above





xFitter [https://xfitter.org] [https://gitlab.com/fitters/xfitter]

xFitter

Welcome to xFitter (former HERAFitter)

Proton parton distribution functions (PDFs) are essential for precision physics at the LHC and other hadron colliders. The determination of the PDFs is a complex endeavor involving several physics process. The main process is the lepton proton deep-inelastic scattering (DIS), with data collected by the HERA ep collider covering a large kinematic phase space needed to extract PDFs. Further processes (fixed target DIS, ppbar collisions etc.) provide additional constraining powers for flavour separation. In particular, the precise measurements obtained or to come from LHC will continue to improve the knowledge of the PDF.

The xFitter project is an open source QCD fit framework ready to extract PDFs and assess the impact of new data. The framework includes modules allowing for a various theoretical and methodological options, capable to fit a large number of relevant data sets from HERA, Tevatron and LHC. This framework is already used in many analyses at the LHC.

Downloads of xFitter software package

All the xFitter releases can be accessed HERE including **2.2.0 FutureFreeze** release
All the former (HERAFitter) releases can be accessed HERE.
Description: http://arxiv.org/abs/1410.4412

xFitter Meetings

- xFitter Workshop at CERN 2-5 May 2023
- User's Meetings: meetings to enhance communication between users and developers (open access)
- Developer's Meeting: technical weekly meetings to ensure communication among developers (restricted access)
- Steering Group's Meeting (restricted access)

xFitter representation

- Snowmass contrubution
- List of results
- List of collected talks

Developers Info (restricted to developers)

- Internal Developments

Organisation

- **Release coordinator/Librarian** (revision of the release candidates): Sasha Glazov, Oleksandr Zenaiev
- **DESY IT Contact:** Yves Kemp

Getting help

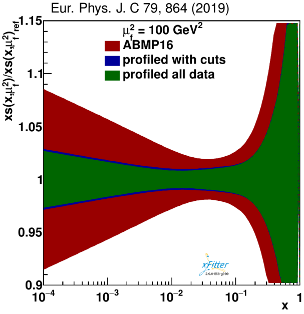
See our help forum https://groups.google.com/forum/#!forum/xfitter-users
In case of questions or problems, please post a message there (requires a google account) or send it via email xfitter-users@googlegroups.com (no account required)

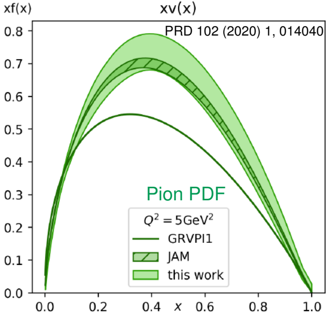
xFitter [xfitter.org] [gitlab.com/fitters/xfitter]

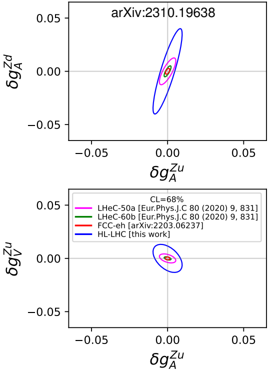
- **xFitter (HERAfitter before 2015) is a unique open-source QCD fit framework:**
  - extract PDFs and theory parameters
  - assess impact of new data
  - check consistency of experimental data
  - test different theoretical assumptions
  - . . . any exercise which involves data vs. theory
- It is widely used by LHC experiments and theorists ($>$ 100 publications)
- **Why is xFitter UNIQUE and so VERSATILE/FLEXIBLE/ADAPTABLE?**
  Because it is fully modular. E.g., hadron interactions are realized as
  - PDF parametrization at starting scale: it is enough to type your favourite formulas
  - PDF decomposition: valence, sea, gluon + automatic numerical integration for sum rules
  - PDF evolution: interfaced various codes (QCDNUM, OPENQCDRAD, APFEL, LHAPDF)
  - hard scattering (“reaction”): again, supports various options:
    - ⋆ various heavy-quark schemes for *ep* DIS
    - ⋆ some “simple” calculations, e.g. LO DY
    - ⋆ interfaced external packages, e.g. HATHOR, HVQMNR
    - ⋆ but main emphasis is put on interfaces to **fast intepolation tables**, such as **fastNLO, ApplGrid, PineAppl**: allows one to get recent higher-order calculations (e.g. MCFM, NNLOJET, MATRIX etc.) “for free”
  - χ^2 definition with various uncertainty treatment (additive, multiplicative etc.)
  - χ^2 minimization: MINUIT, CERES; error matrix by HESSE, MC replicas or custom methods for error matrix estimation (such as by Pumplin arXiv:hep-ph/0008191)
  - . . . and one can **change & mix & introduce new ingredients** freely!

Selected studies by the xFitter team

- "Probing the strange content of the proton with charm production in charged current at LHeC" EPJ C79, 864 (2019)
- "Parton Distribution Functions of the Charged Pion Within The xFitter Framework" Phys.Rev.D 102 (2020) 014040
- "Exploring SMEFT Couplings Using the Forward-Backward Asymmetry in Neutral Current Drell-Yan Production at the LHC" arXiv:2310.19638







χ^2 expression

$$\chi^2_{\text{exp}}(\boldsymbol{m}, \boldsymbol{b}) = \sum_{ij} \left(m_i - \sum_{\alpha} \Gamma^i_{\alpha} b_{\alpha} - \mu_i \right) C^{-1}_{\text{stat},\ ij} \left(m_i - \sum_{\alpha} \Gamma^i_{\alpha} b_{\alpha} - \mu_i \right) + \sum_{\alpha} b^2_{\alpha}$$

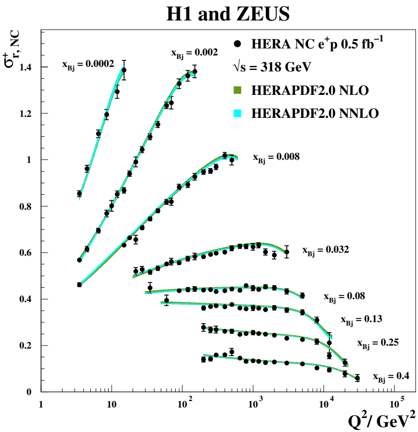
- m_i: data
- μ_i: theory
- $C_{\text{stat},\ ij}$: statistical covariance matrix
- b_α: nuisance parameters for correlated systematic uncertainties
- Γ_α: scaled correlated systematic uncertainties; might depend on m_i, μ_i:

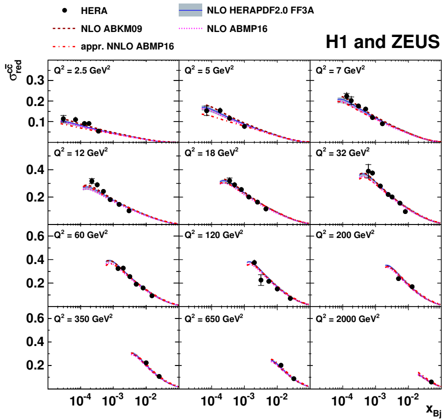
Treatment	Scaling rule (Γ^i_α)
Poisson	$\sqrt{m_i \mu_i}$
Multiplicative	m_i
Additive	μ_i

- Correlated uncertainties can be supplied as covariance matrix or source-by-source
- Also uncertainties can be included with offset method (external variations)

$\Longrightarrow$ Need to know what are uncorrelated and correlated uncertainties, and how they scale

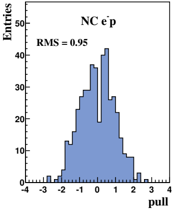
HERA DIS data

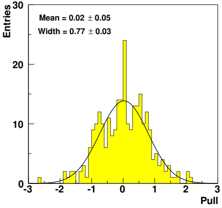




- HERA data on *ep* DIS scattering are a backbone of all global QCD analyses
  - direct constraints on valence and sea quark PDFs in a wide kinematic range
  - however only indirect sensitivity to gluon PDF and α_S
- HERA data on heavy quark (charm, bottom) and jet production in DIS:
  - direct constraints on gluon PDF, α_S, m_c, m_b

HERA DIS data: discussion

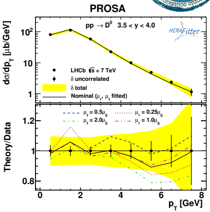


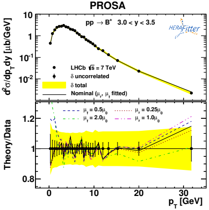


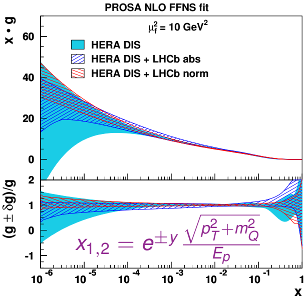
Q^2 GeV^2	$x_{\rm Bj}$	$\sigma^+_{r,\rm NC}$	$\delta_{\rm stat}$ %	$\delta_{\rm uncor}$ %	$\delta_{\rm cor}$ %	$\delta_{\rm rel}$ %	$\delta_{\gamma p}$ %	$\delta_{\rm had}$ %	δ_1 %	δ_2 %	δ_3 %	δ_4 %	$\delta_{\rm tot}$ %
70	0.922×10^{-3}	1.385	2.57	5.67	1.10	0.12	−0.16	−0.29	−0.04	−0.29	0.01	−0.80	6.39
70	0.100×10^{-2}	1.434	2.18	2.55	0.86	−0.08	−0.11	−0.20	−0.02	0.08	0.01	−0.93	3.60
70	0.110×10^{-2}	1.506	1.88	2.10	0.81	−0.08	−0.11	−0.17	−0.01	0.48	0.01	−0.97	3.14
70	0.124×10^{-2}	1.445	1.48	1.80	0.80	−0.15	−0.10	−0.16	−0.01	0.24	0.01	−0.95	2.66
70	0.130×10^{-2}	1.414	1.32	1.32	0.82	−0.16	−0.08	−0.12	0.00	0.45	0.00	−1.09	2.37
70	0.200×10^{-2}	1.246	1.15	1.36	0.78	−0.03	−0.09	−0.12	0.00	0.22	0.01	−0.96	2.19
70	0.250×10^{-2}	1.190	0.96	0.97	0.77	0.03	−0.09	−0.12	0.00	0.13	0.01	−0.91	1.82
70	0.320×10^{-2}	1.084	0.91	0.78	0.74	0.26	−0.10	0.00	0.01	−0.13	0.03	−0.52	1.53
70	0.500×10^{-2}	0.958	0.83	0.74	0.75	0.23	−0.09	−0.09	−0.01	−0.49	0.05	−0.43	1.51
70	0.800×10^{-2}	0.819	1.62	0.48	0.77	0.26	−0.05	−0.11	0.01	−0.06	0.07	−0.05	1.88
70	0.130×10^{-1}	0.716	1.89	0.51	0.86	0.46	−0.07	−0.62	0.01	−0.06	0.05	−0.04	2.27
70	0.200×10^{-1}	0.637	1.77	0.67	0.80	0.37	−0.11	−0.46	−0.01	−0.06	0.07	−0.05	2.14
70	0.320×10^{-1}	0.561	2.10	0.92	0.85	−0.02	−0.05	0.26	−0.04	−0.05	0.07	−0.08	2.46
70	0.500×10^{-1}	0.512	1.61	0.90	1.19	0.35	−0.16	−0.37	−0.12	−0.03	0.00	0.02	2.27
90	0.130×10^{-2}	1.479	0.66	1.49	0.87	0.06	−0.04	−0.13	0.00	−0.05	0.15	0.24	1.87
90	0.150×10^{-2}	1.418	1.12	1.03	1.33	1.02	−1.18	−0.21	0.01	−0.04	0.08	0.12	2.56
90	0.200×10^{-2}	1.328	0.79	0.95	0.76	0.53	−0.25	−0.01	−0.01	0.06	0.07	−0.10	1.57
90	0.320×10^{-2}	1.165	0.66	0.63	0.75	0.45	−0.10	−0.08	0.00	0.08	−0.01	−0.20	1.29
90	0.500×10^{-2}	1.023	0.71	0.58	0.74	0.46	−0.09	−0.06	0.00	0.04	−0.03	−0.13	1.27
90	0.800×10^{-2}	0.883	0.87	0.56	0.78	0.38	−0.07	−0.12	0.00	−0.07	0.06	−0.16	1.37
90	0.130×10^{-1}	0.754	0.91	0.69	0.80	0.33	−0.08	−0.22	−0.01	−0.05	−0.22	−0.09	1.47
90	0.200×10^{-1}	0.649	0.95	0.62	0.81	0.36	−0.09	−0.19	0.00	−0.04	0.19	−0.08	1.47

- HERA DIS data are **final combined** H1 and ZEUS data
  - essentially provided as a single data set (no overlap)
  - combinantion served as a data consistency test
- Very complete description of correlated uncertainties
- Bin-by-bin unfolding (very good resolution of kinematic variables Q^2,x_{Bj})
  - however, sometimes at phase space corners a coarse binning had to be used
- Data are reported at (Q^2,x_{Bj}) *values*
  - although experimental measurements were done in *intervals* of Q^2,x_{Bj}
  - these intervals were different in H1 and ZEUS measurements
  - interpolation procedure (*swimming*) was applied to provide data at (Q^2,x_{Bj}) values
  - → potential model dependence, however, corresponding uncertainties were estimated (also older fixed-target DIS data sets were provided at (Q^2,x_{Bj}) values)
  - recent ZEUS analysis "Study of proton parton distribution functions at high x using ZEUS data" PRD 101 (2020) 11, 112009 published event counts and response matrices, but it is not easy to use these data together with the combined H1+ZEUS data

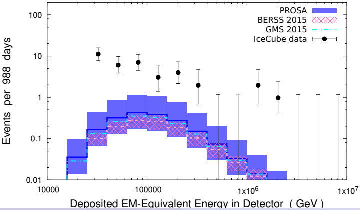
- LHCb measured:
  - charm $0 < p_T < 8$ GeV, $2 < y < 2.5$ NPB871 (2013) 1
  - beauty $0 < p_T < 40$ GeV, $2 < y < 2.5$ JHEP08 (2013) 117
- First QCD analysis of these data: PROSA Coll., EPJ C75 (2015) 396
- Improved gluon and sea-quark distributions up to $x \gtrsim 5 \times 10^{-6}$ (not covered by other experimental data)
  - used in next paper to predict IceCube background for very high energy cosmic ν PROSA Coll., JHEP05 (2017) 004
  - further update with ALICE and LHCb data JHEP04 (2020) 118







$$x_{1,2} = e^{\pm y} \frac{\sqrt{p_T^2 + m_Q^2}}{E_p}$$



Typical description of correlated systematic uncertainties

LHCb 5 TeV JHEP06 (2017) 147

Table 2: Fractional systematic uncertainties, in percent. Uncertainties that are computed bin-by-bin are expressed as ranges giving the minimum to maximum values. Ranges for the correlations between p_{T}-y bins and between modes are also given, expressed in percent.

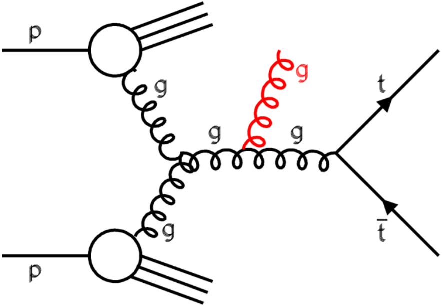
	Uncertainties (%)				Correlations (%)	
	D^0	D^+	D_s^+	D^{*+}	Bins	Decay modes
Luminosity	3.8				100	100
Tracking	3–5	5–7	4–7	5–7	90–100	90–100
Branching fractions	1.2	2.1	5.8	1.5	100	0–95
Simulation sample size	0–10	0–10	2–9	1–10	0	0
Simulation modelling	0.3	0.7	0.6	2	0	0
PID sample size	0–1	0–1	0–2	0–2	0–100	0–100
PID binning	0–30	0–10	0–20	0–20	0	0
Fit model shapes	0–3	0–3	0–3	0.0–1.0	0	0

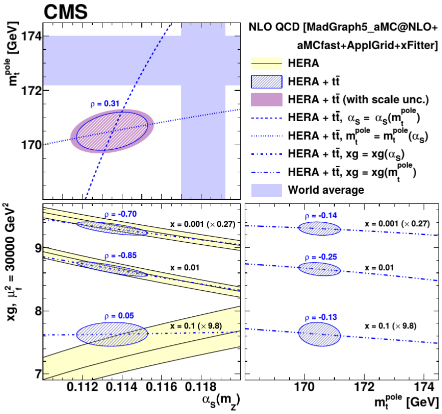
This information is not really sufficient:

- need to know contributions of different systematic uncertainties for each bin (not just ranges)
- need to know correlation betwen different *D* and energies
- *total* covariance matrices were provided for some LHCb data sets, but
  - some of them appeared to be not positive definite (issue of rounding?)
  - they still do not allow one to properly correlate different data sets

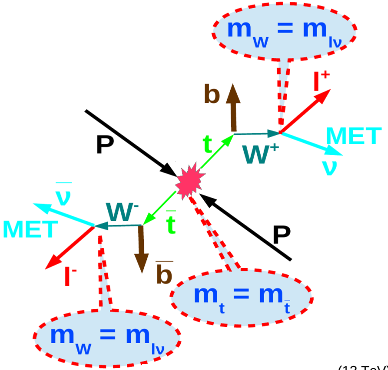
EPJ C80 (2020) 658 arXiv:2402.08486

- First measurement of triple-differential $t\bar{t}$ cross sections as function of $M(t\bar{t})$, $y(t\bar{t})$ and N_{jet}
  - $M(t\bar{t})$ constrains m_t: $M(t\bar{t}) > 2m_t$
  - $M(t\bar{t})$, $y(t\bar{t})$ constrain PDFs: $x_{1,2} = \frac{M(t\bar{t})}{\sqrt{s}} e^{\pm y(t\bar{t})}$
  - N_{jet} constrain α_s
- Undetected neutrinos affect detector resolution of $t\bar{t}$ kinematic observables: **unfolding is crucial**





CMS $t\bar{t}$ multi-differential x-sections: kinematic reconstruction



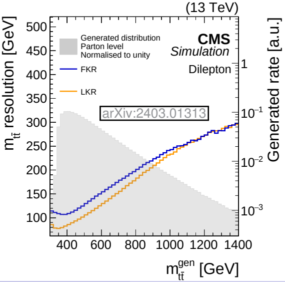
- Measured input: leptons, jets, MET
- Unknowns: $\bar{p}_\nu$, $\bar{p}_{\bar{\nu}}$ (6)
- Constraints:
  - m_t, $m_{\bar{t}}$ (2)
  - m_{W^+}, m_{W^-} (2)
  - $(\bar{p}_\nu + \bar{p}_{\bar{\nu}})_T = \mathrm{MET}$ (2)

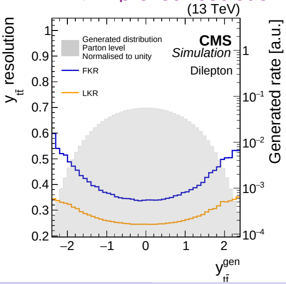
(1) Full kinematic reconstruction (FKR):

- reconstruct t, $\bar{t}$ using all constraints

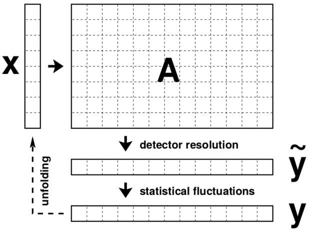
(2) Loose kinematic reconstruction (LKR):

- reconstruct $t\bar{t}$ (4 unknowns)
- m_t constraints not used
  $\rightarrow$ reliable for m_t extraction
  $\rightarrow$ improved resolution of $t\bar{t}$ kinematics





CMS $t\bar{t}$ multi-differential x-sections: unfolding

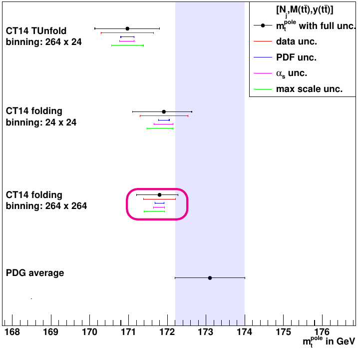


- **Unfolding problem:** $\mu = \boldsymbol{Ax}$
- $\boldsymbol{\mu}$: detector expectation (M_y bins), $\boldsymbol{V_{yy}}$ is its covariance
- $\boldsymbol{A}$: response matrix (taken from MC)
- $\boldsymbol{x}$: unknown truth, M_x bins ($M_x \leq M_y$)
- $\mathcal{L} = (\boldsymbol{y} - \boldsymbol{Ax})^T \boldsymbol{V_{yy}}^{-1} (\boldsymbol{y} - \boldsymbol{Ax}) + \tau^2 (\boldsymbol{x} - \boldsymbol{x_0})^T (\boldsymbol{L}^T \boldsymbol{L})(\boldsymbol{x} - \boldsymbol{x_0})$
- $\frac{\partial \mathcal{L}}{\partial x} = 0 \quad \Rightarrow \quad \boldsymbol{x} = \boldsymbol{x}(\boldsymbol{y}, \boldsymbol{V_{yy}}, \boldsymbol{x_0}),\ \boldsymbol{V_{xx}} = \boldsymbol{V_{xx}}(\boldsymbol{y}, \boldsymbol{V_{yy}}, \boldsymbol{x_0})$
- $\boldsymbol{x_0}$: bias vector for regularization (taken from MC)

- Solution obtained using regularised multidimensional unfolding with TUnfold
- Bin-to-bin correlations damped by biasing curvature to MC
- Regularised strength determined by minimising global correlation coefficient ('MinRhoAvg')
- Finer binning at detector level, limited only by Gaussian stat. $\rightarrow$ reduce "wide bins" problem
- Coarser binning at generator level, limited by resolution
- **Regularized vs unregularized unfolding**: moderate impact on QCD analysis, mainly due to systematic uncertainties affected by limited MC statistics

CMS $t\bar{t}$ multi-differential x-sections: forward folding

DESY 2018 summer school, L. Materne, bachelor thesis “Differential Top-Pair Production Cross Section with the CMS Detector - Optimization of Measurement Information”, Karlsruher Institut für Technologie (KIT), Bachelorarbeit, 2018 [ETP-Bachelor-KA/2018-11]



- Direct folding: convolute theory prediction with response matrix
- One can use finer binning (limited by Gaussian statistics only): 24 $\rightarrow$ 264 bins (in 3D)

$\rightarrow$ better sensitivity to theory parameters (m_t^{pole})

- Conceptually such measurement is simpler
- Extra step required to compare to theory predictions: multiply with the response matrix (trivial in `xFitter`)

Global ABMP16tt analysis

- Follows ABMP16 PDF fit [PRD 96 (2017) 014011]
- Input data:
  - DIS data (HERA and fixed-target experiments): backbone
  - Drell-Yan data (LHC and fixed-target experiments): improve flavour separation
  - $t\bar{t}$ and single t data (LHC and Tevatron): gluon, α_S, m_t
- Focusing on adding differential $t\bar{t}$ LHC data:
  - all measurements of total $\sigma(t\bar{t})$:
    - ★ 10 data points, including combined CMS+ATLAS cross section at 7 and 8 TeV
    - ★ no correlations between CMS and ATLAS is available
  - differential measurements $\frac{1}{\sigma(t\bar{t})}\frac{d\sigma(t\bar{t})}{dO}$:
    - ★ normalized cross sections (to avoid unknown correlation with total $\sigma(t\bar{t})$ and to reduce unknown correlations between different data sets)
    - ★ bin-by-bin correlations should be available (no Tevatron data)

$\sigma(t\bar{t})$

Experiment	decay channel	dataset	luminosity	$\sqrt{s}$
ATLAS & CMS	combined	2011	5 fb^{-1}	7 TeV
ATLAS & CMS	combined	2012	20 fb^{-1}	8 TeV
ATLAS	dileptonic, semileptonic	2011	257 pb^{-1}	5.02 TeV
CMS	dileptonic	2011	302 pb^{-1}	5.02 TeV
ATLAS	dileptonic	2015-2018	140 fb^{-1}	13 TeV
ATLAS	semileptonic	2015-2018	139 fb^{-1}	13 TeV
CMS	dileptonic	2016	35.9 fb^{-1}	13 TeV
CMS	semileptonic	2016-2018	137 fb^{-1}	13 TeV
ATLAS	dileptonic	2022	11.3 fb^{-1}	13.6 TeV
CMS	dileptonic, semileptonic	2022	1.21 fb^{-1}	13.6 TeV

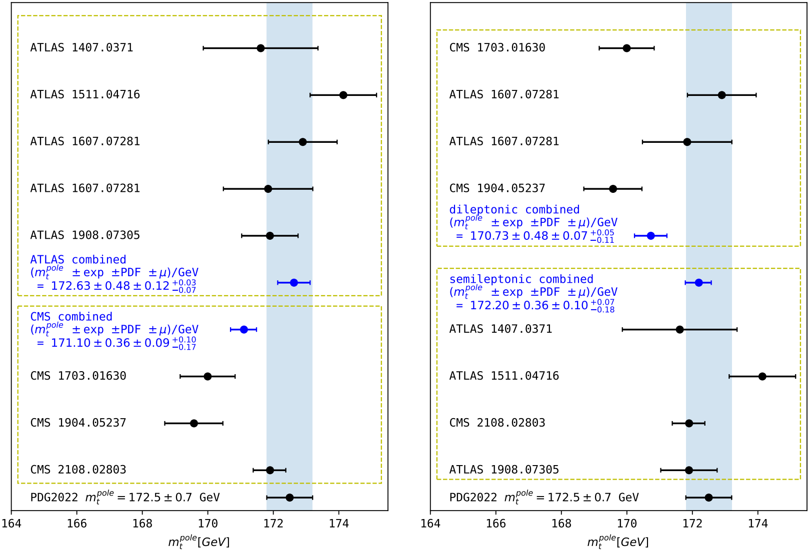
$\frac{1}{\sigma(t\bar{t})}\frac{d\sigma(t\bar{t})}{dO}$

Experiment	decay channel	dataset	luminosity	$\sqrt{s}$	observable(s)	n
CMS	semileptonic	2016–2018	137 fb^{-1}	13 TeV	$M(t\bar{t})$, $\|y(t\bar{t})\|$	34
CMS	dileptonic	2016	35.9 fb^{-1}	13 TeV	$M(t\bar{t})$, $\|y(t\bar{t})\|$	15
ATLAS	semileptonic	2015–2016	36 fb^{-1}	13 TeV	$M(t\bar{t})$, $\|y(t\bar{t})\|$	19
ATLAS	all-hadronic	2015–2016	36.1 fb^{-1}	13 TeV	$M(t\bar{t})$, $\|y(t\bar{t})\|$	10
CMS	dileptonic	2012	19.7 fb^{-1}	8 TeV	$M(t\bar{t})$, $\|y(t\bar{t})\|$	15
ATLAS	semileptonic	2012	20.3 fb^{-1}	8 TeV	$M(t\bar{t})$	6
ATLAS	dileptonic	2012	20.2 fb^{-1}	8 TeV	$M(t\bar{t})$	5
ATLAS	dileptonic	2011	4.6 fb^{-1}	7 TeV	$M(t\bar{t})$	4
ATLAS	semileptonic	2011	4.6 fb^{-1}	7 TeV	$M(t\bar{t})$	4

Discussion of $t\bar{t}$ data uncertainty treatment

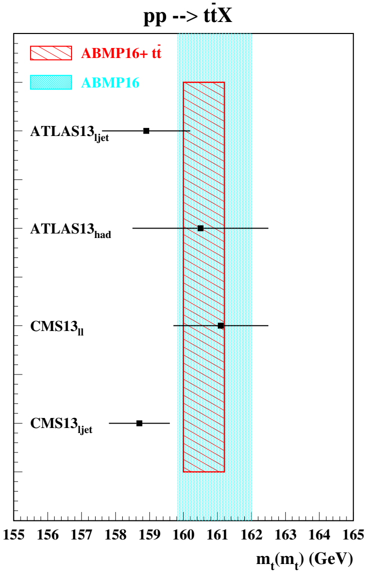
- only two data sets (CMS dilepton Run 1, CMS dilepton Run 2) report source by source systematic uncertainties
- all other data sets report covariance matrices: it is even not possible to separate systematic correlation from statistical (unfolding)
- for some data sets (ATLAS l+jet Run 2, ATLAS dilepton Run 1 (8 TeV)), covariance matrices are not singular as they should be for normalised x-sections (one degree of freedom is lost): issue of rounding?
- **we tried to minimize the impact of the lack of experimental systematic correlations by using the total x-section and normalised differential x-sections (many systematic uncertainties cancel for normalised x-sections), but:**
  - $\rightarrow$ please always report covariance matrix from unfolding and source by source systematic uncertainties
  - $\rightarrow$ please specify the sign of systematic variation (e.g. scale varied up or down)
  - $\rightarrow$ if some systematics is calculated as envelope of several variations, please report every variation
  - $\rightarrow$ for covariance matrix, please do some basic consistency checks (positive definiteness, singularity if normalised x-sections etc.)
  - $\rightarrow$ an effort from ATLAS and CMS on combining their differntial $t\bar{t}$ data will be useful

Extraction of m_t^{pole} JHEP 05 (2024) 321

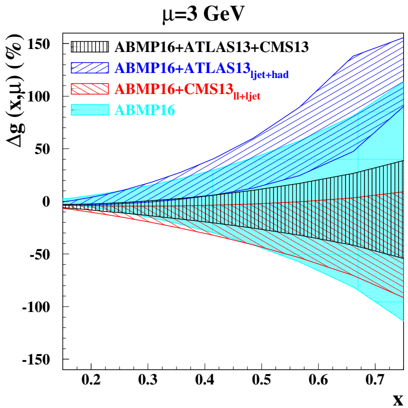


2.5σ tension ATLAS vs CMS, dilepton vs semileptonic: lack of info on correlated systematics?

Global ABMP16tt fit of PDFs, α_S, m_t (WIP) Alekhin, Garzelli, Moch, SZ 24XX.YYYYY



Data set	$m_t(m_t)$ (GeV)	χ^2/NDP
$ATLAS\,13_{ljet}$	158.9 ± 1.3	25.2/19
$ATLAS\,13_{had}$	160.5 ± 2.0	11.3/10
$CMS\,13_{ll}$	161.1 ± 1.4	13.9/15
$CMS\,13_{ljet}$	158.7 ± 0.9	37.4/34



some tension between ATLAS or CMS data remains

Summary: what data do we need for global QCD fits?

- Unfolded or forward folded data?
  - forward folding preserves maximum information
  - with unfolded data it is easier to conclude on (in)consitency of various data sets
- Uncorrelated and correlated (source-by-source for systematics) uncertainties
- Additive or multiplicative systematic uncertainties
  - multiplicative treatment is difficult to use properly in some cases (e.g. normalized data)
- Many data sets are used (a few tens): need info how their uncertainties are correlated
- Practical experience: it is very useful if quantitative data vs theory comparison is documented in the experimental publication, can be reproduced and used as starting point
  - e.g. χ^2 for data vs theory comparison (details of the theory calculation must be documented, or theory predictions should be provided explicitly)
  - IDEA: provide a demo routine which reads HEPDATA and computes χ^2 (could be even integrated in HEPDATA)?
- Try to provide unfolded data in a form such that one can restore its dependence (via response matrix and unfolding bias) on physical parameters
  - example: dependence of measured $\sigma(t\bar{t})$ on m_t
  - can one do more?.. (PDFs etc.)
- Dealing with inconsistent data sets:
  - adjust tolerance $\Delta\chi^2 > 1$?
  - downweight outliers?
  - use only consistent data keeping $\Delta\chi^2 = 1$?
- Probably, many of the differences between results of global fits can be traced back to the treatment of the data