Publishing unfolded measurements

2nd Pan-European Advanced School on Statistics in High Energy Physics

Carsten Burgard

March 29, 2022

Introduction

- Take a simple analysis: place a few cuts, and count events
- Measure signal strength μ in the **selected** phase space by fitting

$$N_{\text{obs}}^{\text{sel}} = \mu \cdot N_{\text{sim}}^{\text{sig,sel}} + N_{\text{sim}}^{\text{bkg,sel}}$$

- Publish the **total** cross-section

$$\sigma_{\text{obs}}^{\text{sig,tot}} = \mu \cdot \sigma_{\text{sim}}^{\text{sig,tot}}$$

Shortcomings & Caveats

- assume that μ on the selected phase space (sel) is the same as on the total phase space (tot)
- no uncertainty assigned on this extrapolation

“Improvement”

- Measure signal strength μ in the **selected** phase space by fitting

$$N_{\text{obs}}^{\text{sel}} = \mu \cdot N_{\text{sim}}^{\text{sig,sel}} + N_{\text{sim}}^{\text{bkg,sel}}$$

- Publish the **selected** cross-section

$$\sigma_{\text{obs}}^{\text{sig,sel}} = \mu \cdot \sigma_{\text{sim}}^{\text{sig,sel.}}$$

Shortcomings & Caveats

- need to reproduce selection at detector level in order to compare with any model
- not useful for anyone outside of **your** experimental collaboration

Improvement

- Measure signal strength μ in the **selected** phase space by fitting

$$N_{\text{obs}}^{\text{sel}} = \mu \cdot N_{\text{sim}}^{\text{sig,sel}} + N_{\text{sim}}^{\text{bkg,sel}}$$

- Define a **fiducial** phase space
  - similar to your selected phase space to reduce extrapolation
  - defined using **truth level** cuts to avoid relying on detector model
- Publish the **fiducial** cross-section

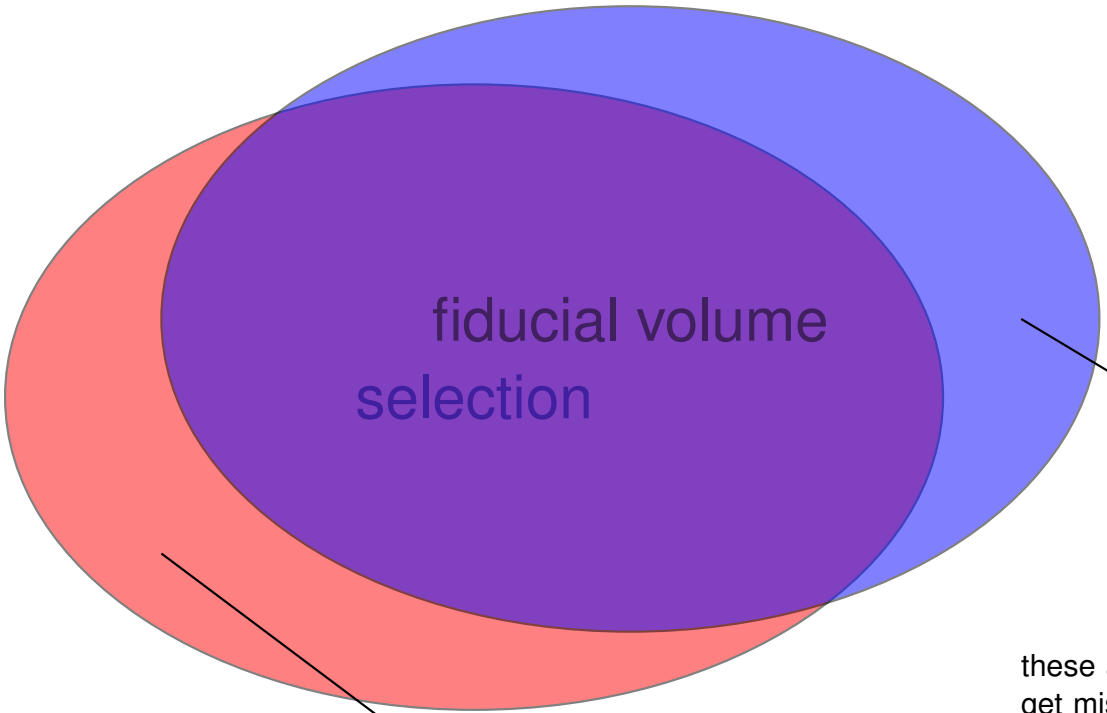
$$\sigma_{\text{obs}}^{\text{sig,fid}} = \mu \cdot \sigma_{\text{sim}}^{\text{sig,fid}}$$

Definitions

acceptance geometrical acceptance of your measurement:
$A = \sigma_{\text{sim}}^{\text{sig,fid}} / \sigma_{\text{sim}}^{\text{sig,tot}}$

efficiency accuracy with which your detector is able to measure your fiducial volume: $\epsilon = \sigma_{\text{sim}}^{\text{sig,sel}} / \sigma_{\text{sim}}^{\text{sig,fid}}$

More definitions



inefficiencies

these are events that are part of your fiducial volume, but get missed by your detector.

out-of-fiducial events (“fakes”)

these are events that you select with your cuts, but that are not part of your fiducial volume. They are “signal”, but behave like a background – for example, they pass one of your p_T cuts by virtue of a mismeasurement.

Going differential

- want to measure cross-section in several bins of an observable
- simple solution: conduct non-overlapping fiducial measurements for each bin separately
- this procedure has a name: **bin-by-bin unfolding** or “correction factor method”

$$N^{\text{sel}}_{\text{obs},i} = \mu_i \cdot N^{\text{sig,sel}}_{\text{sim},_i} + N^{\text{bkg,sel}}_{\text{sim},i}$$

$$c_i = \frac{\sigma^{\text{sig,fid}}_{\text{sim},_i}}{\sigma^{\text{sig,sel}}_{\text{sim},_i}} = \frac{N^{\text{sig,fid}}_{\text{sim},_i}}{N^{\text{sig,sel}}_{\text{sim},_i}}$$

$$\sigma^{\text{sig,fid}}_{\text{obs},i} = \frac{c_i}{\mathcal{L}} \cdot N^{\text{sig,sel}}_{\text{obs},i} = \frac{c_i}{\mathcal{L}} \cdot \mu N^{\text{sig,sel}}_{\text{sim},i}$$

- the correction factors c_i incorproate a generalized acceptance & efficiency correction

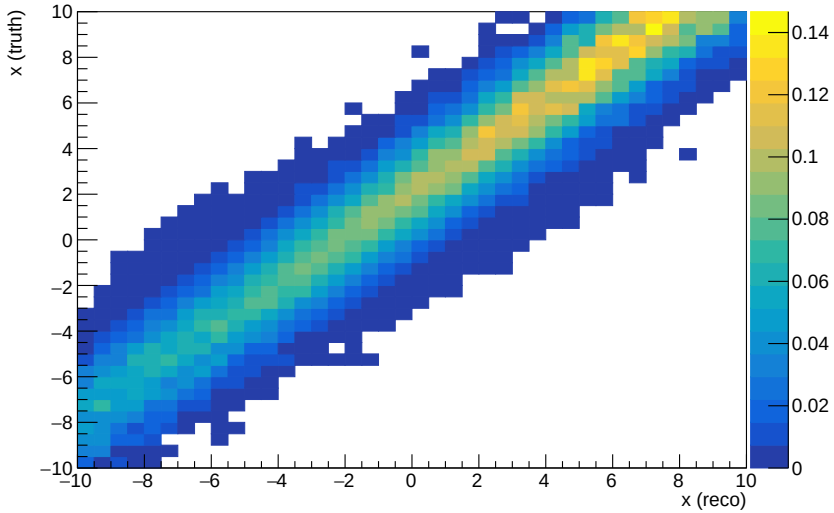
Formulating the problem

- segment truth-level phase-space into $n + 1$ regions
  - n bins in your fiducial signal region
  - one additional bin for everything outside the fiducial region
- segment detector-level phase-space into $m + 1$ regions
  - m signal regions ($m \geq n$)
  - one additional bin for everything outside the signal region
- find the number of signal events in each truth-level bin as

$$N_{\text{obs},i}^{\text{sig,fid}} = \sum_{j}^{m} \epsilon_{ij} N_{\text{obs},i}^{\text{sig,sel}}$$

- ϵ_{ij} is the "response function", the probability of an event from in signal region j to originate from truth bin i

The detector response



- using monte carlo
- construct migration matrix M_{ij}
- two-dimensional histogram mapping
  - the bins on truth-level
  - the bins on detector-level

- use underflow bin for out-of-fiducial events and misses
- normalize each row to the sum of events (including underflow)
- this “confusion matrix” $\widetilde{M}$ yields the probability of an event from truth bin i to end up in signal region j

$$N_{\text{sim},j}^{\text{sig,sel}} = \sum_{i}^{n} \widetilde{M}_{ij} \cdot N_{\text{sim},i}^{\text{sig,fid}}$$

Some linear algrebra

- recasting all of this in linear algrebra notation, we find

$$\widetilde{M}_{ij}\vec{N}^{\text{sig,fid}}_{\text{sim}} = \vec{N}^{\text{sig,sel}}_{\text{sim}}$$
$$\vec{N}^{\text{sig,fid}}_{\text{sim}} = \widetilde{M}^{-1}_{ij}\vec{N}^{\text{sig,sel}}_{\text{sim}}$$

- the inverse of the confusion matrix $\widetilde{M}^{-1}_{ij}$ is a MC estimate for the response function

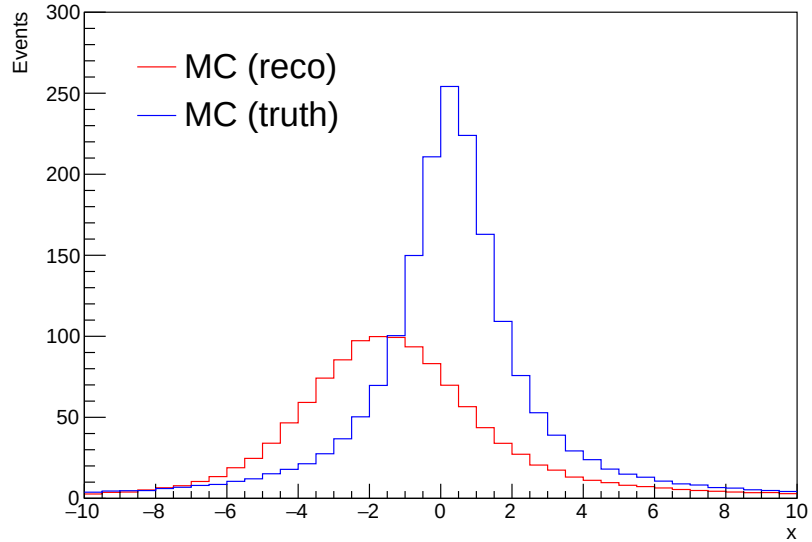
$$\vec{N}^{\text{sig,fid}}_{\text{obs}} = \epsilon_{ij}\vec{N}^{\text{sig,sel}}_{\text{obs}}$$

Caveats & further reading

- this method of unfolding is called “matrix inversion unfolding”
- straight-forward, unbiased and identical with the ML estimator
- some undesireable properties: unreasonably large uncertainties
- give rise to a whole zoo of methods of *regularization*

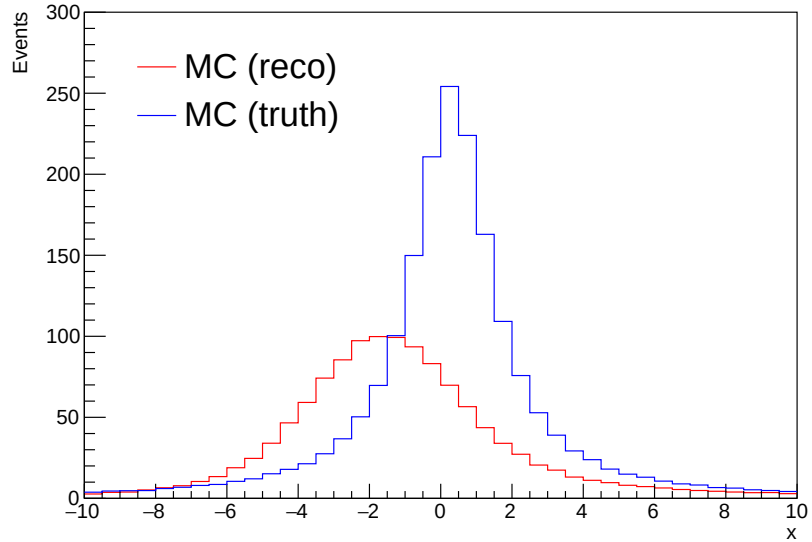
Unfolding in a nutshell

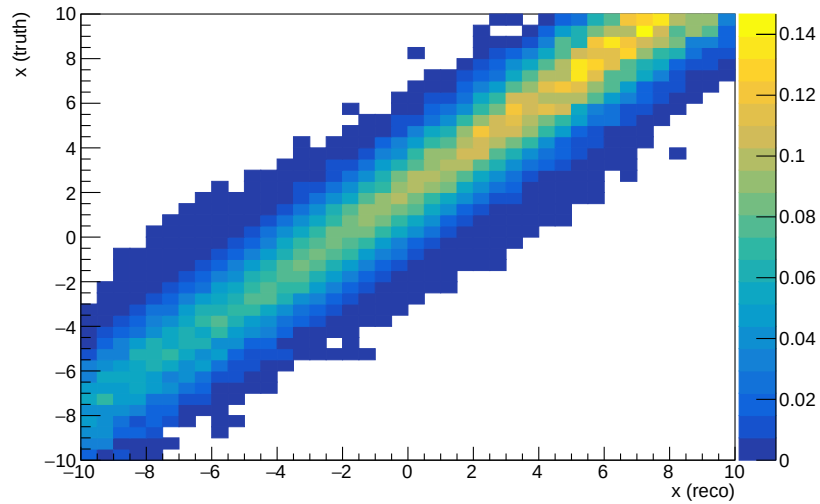
1. start with truth and reco distributions from MC



Unfolding in a nutshell

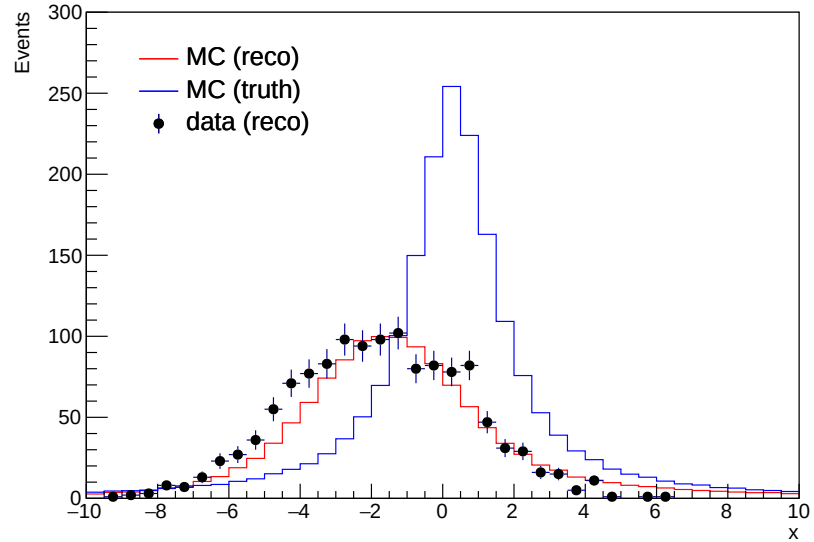
1. start with truth and reco distributions from MC
2. measure the migration or response matrix

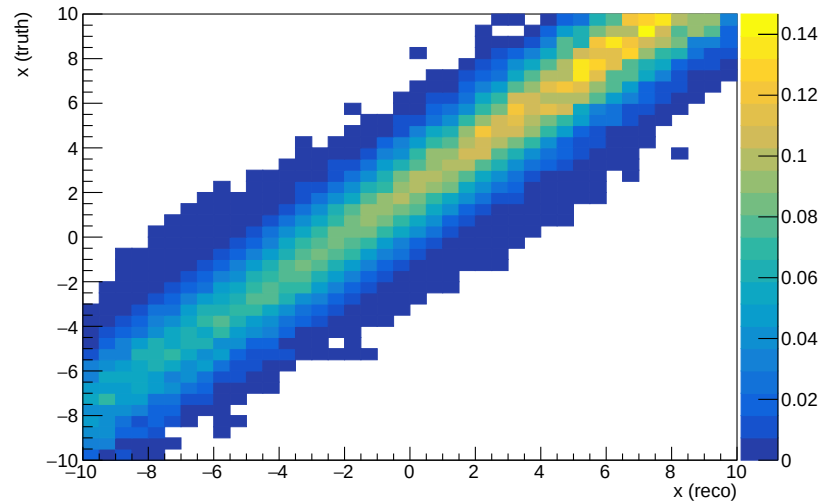




Unfolding in a nutshell

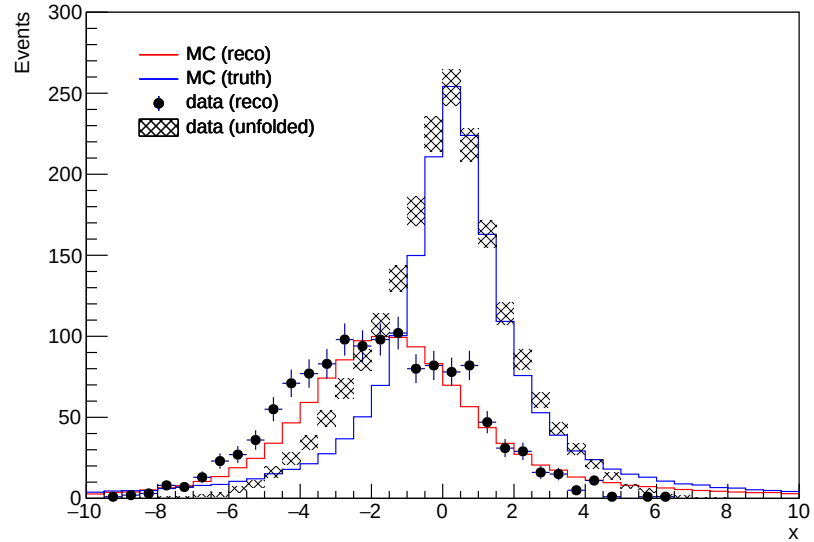
1. start with truth and reco distributions from MC
2. measure the migration or response matrix
3. measure data

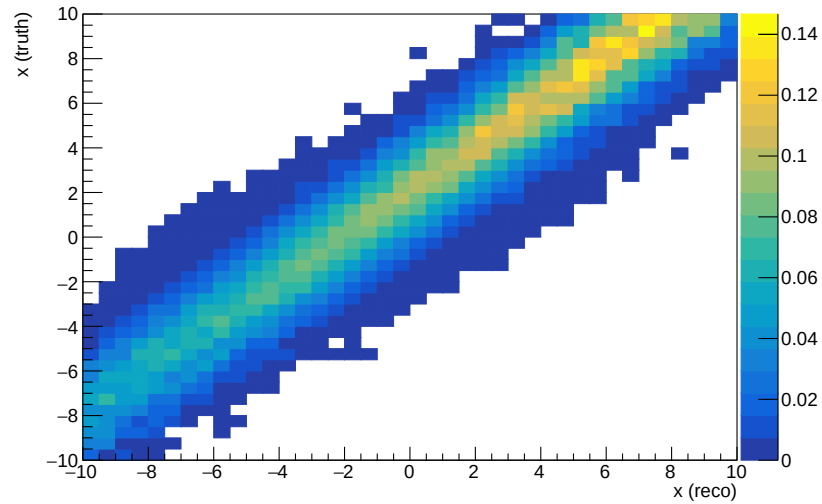




Unfolding in a nutshell

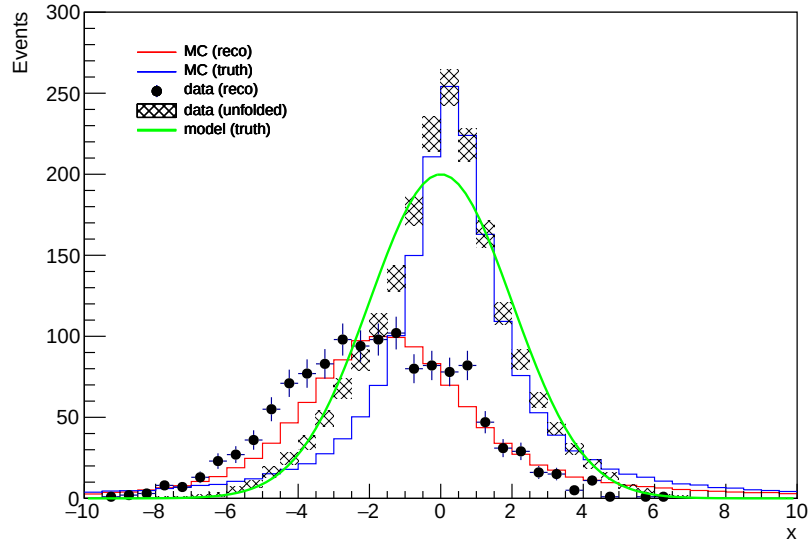
1. start with truth and reco distributions from MC
2. measure the migration or response matrix
3. measure data
4. unfold

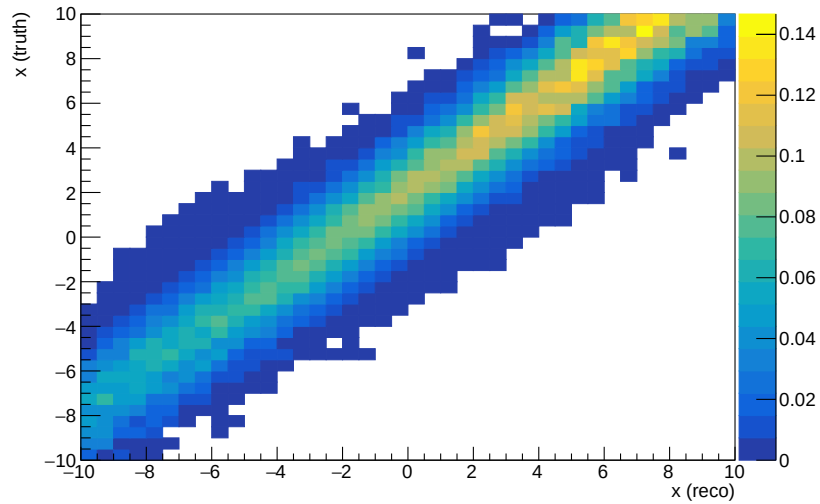




Unfolding in a nutshell

1. start with truth and reco distributions from MC
2. measure the migration or response matrix
3. measure data
4. unfold
5. compare to theory model (optional)





Likelihood-based unfolding

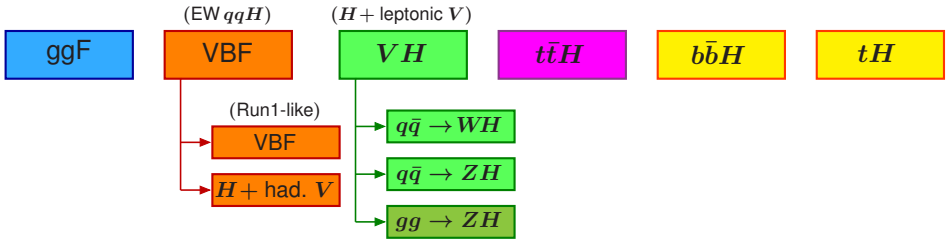
- most unfolding follows a two-step approach
  1. extract the signal in the selected phase space (typically using a fit)
  2. extrapolate form selected to fiducial phase space (unfold)
- two-step approach creates challenges for proper handling of systematic uncertainties
  - need to pay close attention to correctly treating sources of systematic uncertainty and their correlations
- possibility to perform signal extraction and unfolding simultaneously using a maximum-likelihood fit

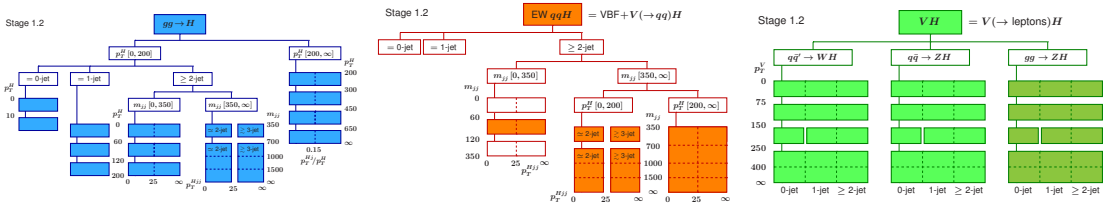
Receipe

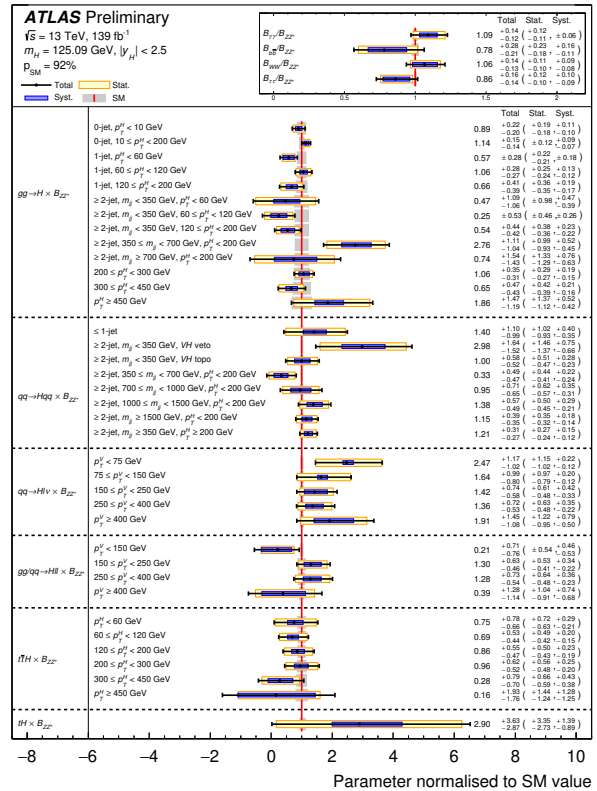
- slice signal prediction in truth bins, with separate scaling factors
- matrix of signal slices vs. signal regions: confusion matrix
- fit yields maximum likelihood estimators for unfolded predictions
- why are people not doing this?

Higgs physics: STXS

- **S**implified **T**emplate **X**-**S**ections
- signal sliced in production bins production mode, p_T^H, n_{jets}, . . .



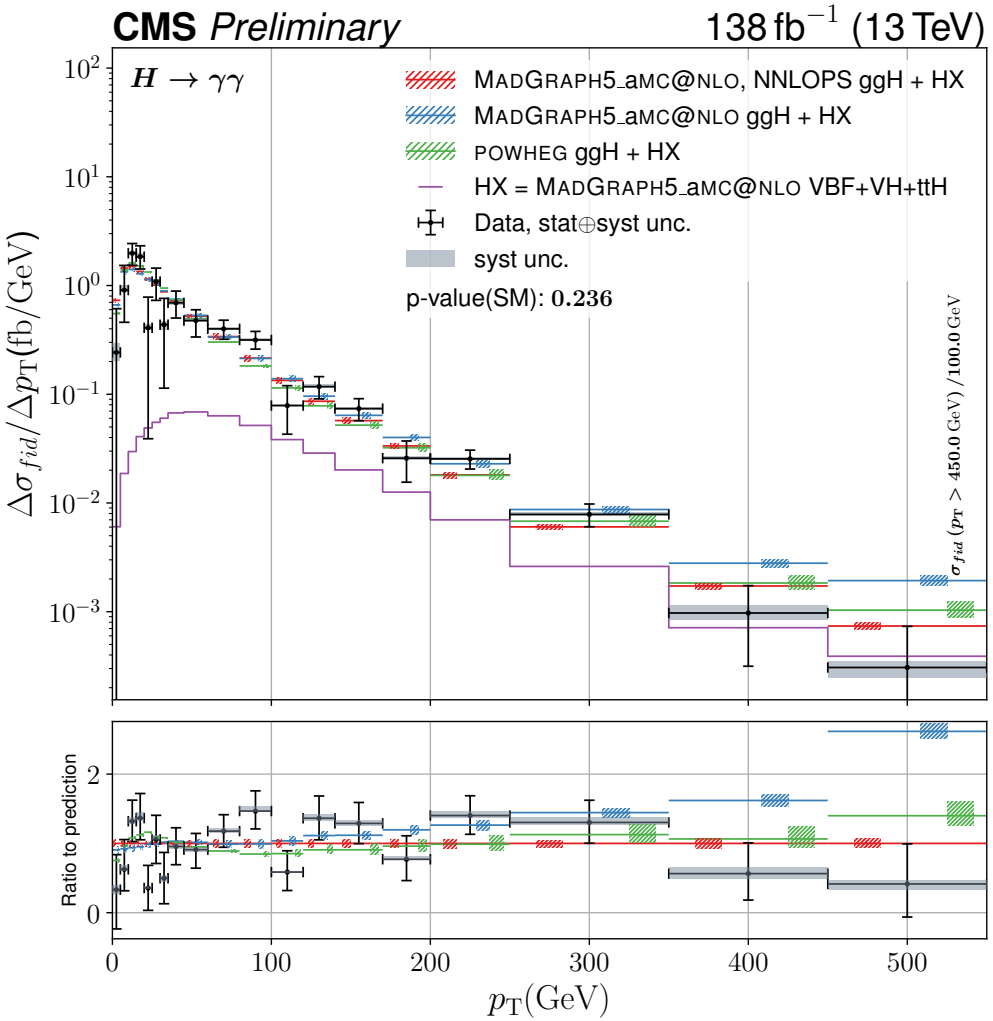




Caution!

- STXS is differential, but not fiducial
- . . . but latest publications started publishing acceptance

Using unfolded results

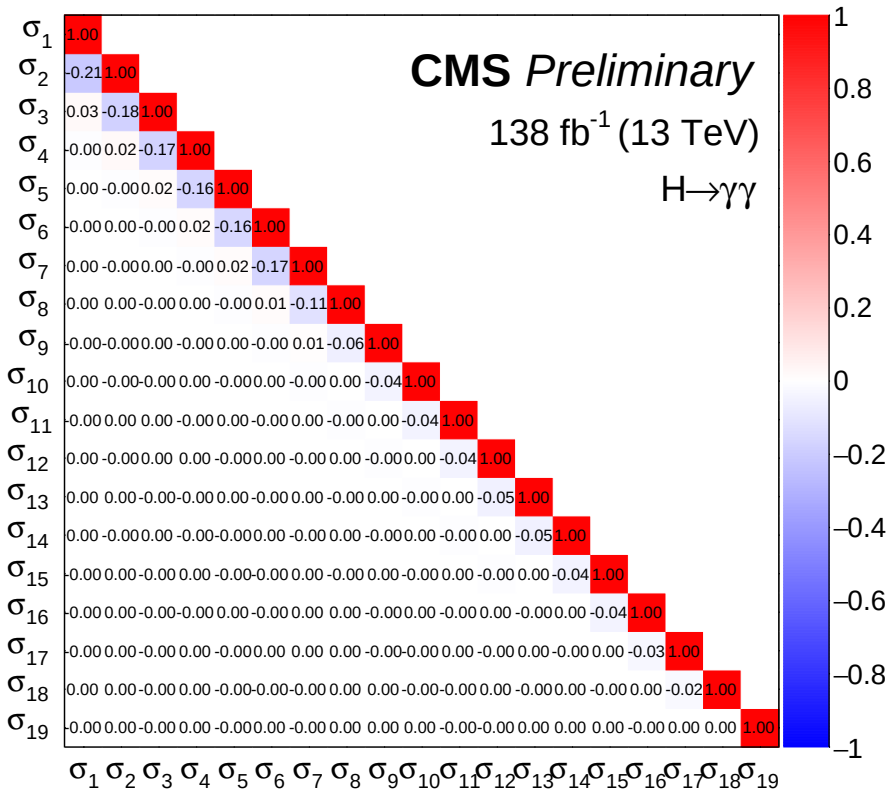


- the resulting unfolded distributions can be compared to theory predictions
- possible to do without access to detector simulation
- can compute *p*-value to quantify agreement

Using unfolded results

Caution!

- the unfolded data points are not independent
- need to know covariance matrix for quantitative studies
- **always publish the correlations of the unfolded data points**



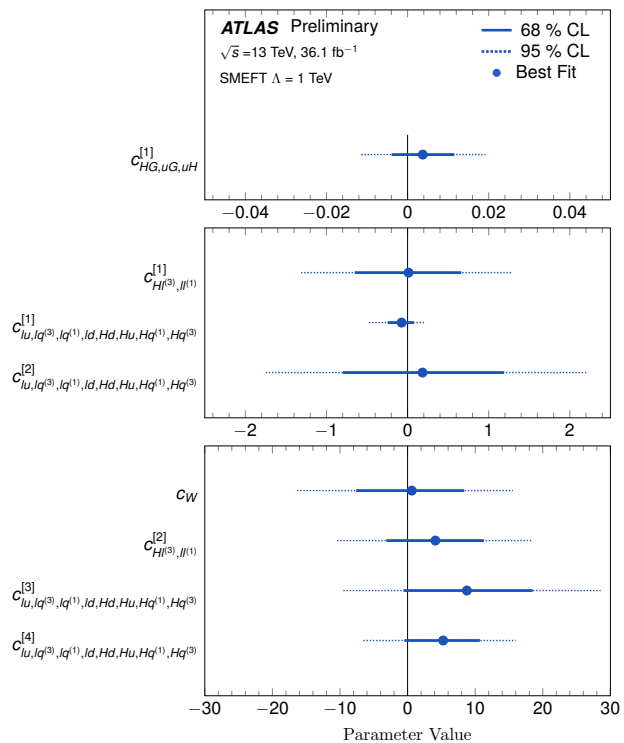
Using unfolded results

- Knowing the covariances C and the central values σ_{obs}, construct a Gaussian LH based of your favourite model on unfolded data

$$G(\boldsymbol{\Delta\sigma}) = \frac{1}{\sqrt{(2\pi)^{n_{\text{bins}}} \det C}} \exp\left(-\frac{1}{2}\boldsymbol{\Delta\sigma}^T C^{-1} \boldsymbol{\Delta\sigma}\right)$$

$$\Delta\sigma = \sigma^{\text{obs}} - \sigma^{\text{model}}$$

  - can plug-in arbitrary models
  - produce statistically sound, quantiative comparisons
  - re-interpretations inside and outside of collaborations



Summary

- Unfolding is a method to **add value** to your measurement
  - independent of your specific measurement apparatus
  - useful for theorists
  - re-interpretable at truth level
- Some care needed to do it correctly
  - carefully choose method of unfolding and/or regularization
  - publish covariance/correlation matrices for everything!

Further reading

- *Statistical Data Analysis* (Glen Cowan) – **Book**
- *Comparison of unfolding methods using RooFitUnfold* (Brenner, Verschuuren, Balasubramanian, Burgard, Croft, Verkerke) – **Paper**
- *Uncertainty quantification in unfolding elementary particle spectra at the Large Hadron Collider* (Mikael Kuusela) – **Thesis**