

## HEP Overview

---

Philippe Gras

June 11-13, 24

CEA/IRFU - Saclay

In this talk I will depict the unfolding motivations, describe the methods commonly used, and introduce new methods that have been emerging using machine learning techniques.

# Use of unfolding in HEP

- Unfolding is used for differential cross-section measurements, and more generally of observable distributions
  - Unfolding covers correction for all detector effects on the observed distributions: smearing, efficiency, misidentification, acceptance, and background.
- 
- Jets are the objects measured with the least resolution (with missing  $E_T$ ) and unfolding is especially relevant for jet-dependent observables.

# Cross section measurement in practice

To measure a differential cross section we typically define a histogram and count the number of event in each bin of the histogram.

$$\frac{d\sigma}{dX} \rightsquigarrow \frac{\delta_k \sigma}{\delta_k X} = \frac{(N_k^{\text{data}} - N_k^{\text{bkg}}) \cdot \mathcal{C}_k}{\delta_k X \cdot \mathcal{L}}$$

With  $\delta_k X$  the bins of a histogram,  $N_k^{\text{data}}$  the measured event yield in the bin,  $N_k^{\text{bkg}}$  the estimated background contribution,  $\delta_k \sigma$  the cross section integrated over the bin,  $\mathcal{L}$  the integrated luminosity,  $\mathcal{C}_k$  the correction for efficiency, bin-to-bin migration, and acceptance.

# Why unfolding (or not unfolding) measurement?

## Why unfolding a measurement ?

- Obtain a more fundamental result that does not depend on the apparatus.
- Ease comparison with results from other experiments.
- Ease comparison with other theoretical predictions: no need to simulate the detector response.

## Why not unfolding ?

- Unfolding is an ill-posed problem and regularization that it may require can bias the result.
- Unfolding can only reduce the information contents.

# Ill-posed problem

- Because of the detector finite resolution, we cannot infer  $d\sigma/dX$  from the measurement (= measurement of event yields) without regularity assumptions: cannot see variations below the resolution.
  - Needs to add a hypothesis on the regularity of the distribution to infer: Regularization
- Nevertheless, actually measuring  $\delta\sigma/\delta X$ 
  - reduced to a ill-conditioned problem, i.e. solutions with large variance, or well-conditioned.
  - If bin width,  $\delta X$ ,  $\approx$  resolution, then regularization is often not needed.
  - The case for many analyses.

# Unfolding classical approach

Extract from the Simulation the probability that an event in a bin  $i$  before the detector response (i.e. at generator level) ends up in the bin  $j$  after the detector response (i.e. at reconstruction level):

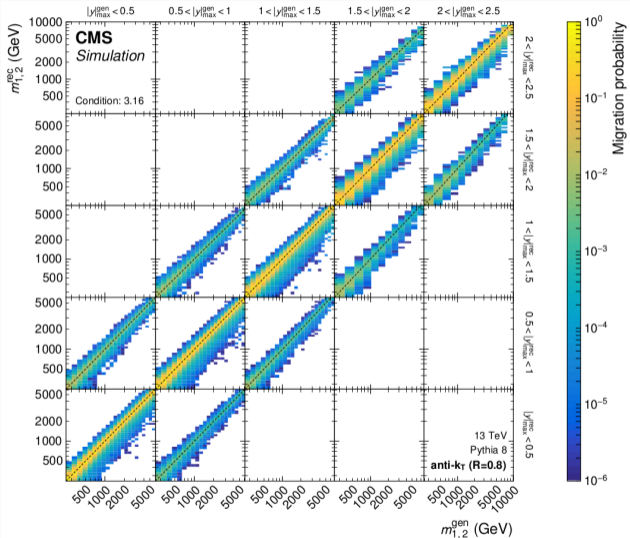
- Fill a 2D histogram of reco vs gen (*migration histogram*)
  - x-axis the generator level (gen) quantity (i.e. before the detector response).
  - y-axis: the reconstruction level (reco) quantity (i.e. after detector response simulation)
- Normalize the histogram such that the sum along the reco axis is equal to one (or to the efficiency) to obtain the probabilities.

This matrix, called **Response Matrix**, is then used to unfold the data. We need to solve,

$$\mathbf{N}_{\text{data}} = \mathbf{R} \mathbf{N}_{\text{unfold}} + \mathbf{N}_{\text{bkg}}$$

$\mathbf{N}$ : histograms, i.e. vectors of bin contents

# Response matrix



$$\mathcal{P}(E_i|C_j) \approx \frac{N_{i,j}}{\sum_j N_{i,j}}$$

$E_i$  event is in reco bin  $i$  (effect)

$C_j$  event is in gen bin  $j$  (cause)

**Tip:** to unfold a multidimensional distribution, map the bins to a 1-D axis.



# Boundaries

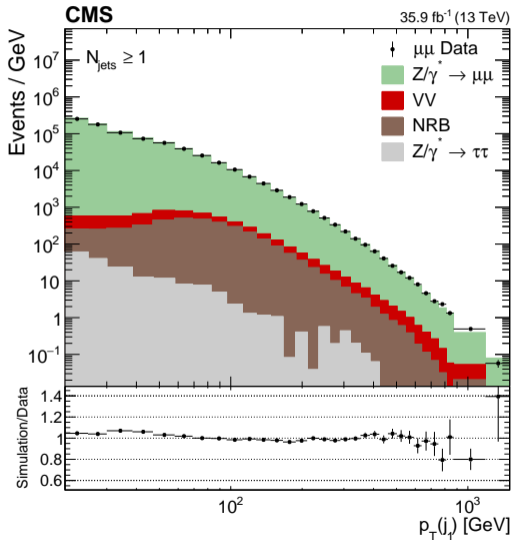
## Migration through boundaries

- Events migrating out of boundaries are treated as inefficiencies.
- Event migrating into the boundaries are treated as "fakes".

## Steep slope

Extra bins beyond boundaries with a steep slope are typically added to perform the unfolding and dropped from the final result.

E.g., in [PRD 108 \(2023\) 05204](#) two extra bins are added at low value to unfold the distribution on the right.



# Unfolding classical methods

## Three main methods

- Response matrix (pseudo-)inversion  $\equiv$  least-square method
- D'Agostini iterative method: converge to Maximum likelihood estimate (MLE)
- MLE

### Least-square

$$(R\mathbf{x} + \mathbf{b} - \mathbf{N}_{\text{data}})^T \Sigma^{-1} (R\mathbf{x} + \mathbf{b} - \mathbf{N}_{\text{data}})$$

*Gaussian approx.*

$\rightarrow$  *Linear algebra*

### MLE

$$-\sum_j \ln(\text{Poiss}(\mathbf{N}_{\text{data},j} | [R\mathbf{x}]_j + b_j))$$

*Unc. can be profiled*

$$\mathbf{x} \equiv \mathbf{N}_{\text{unf}}$$

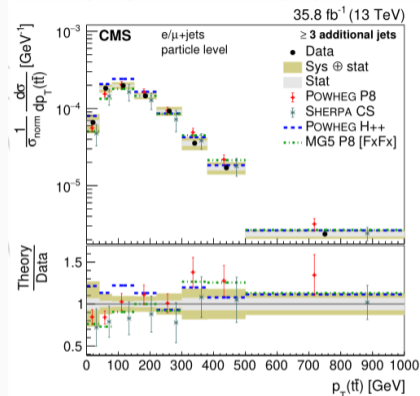
# Least-square method

- Invert the response matrix by minimizing,

$$\chi_{\text{unf}}^2 = (R\mathbf{x} + \mathbf{b} - N_{\text{data}})^T \Sigma^{-1} (R\mathbf{x} + \mathbf{b} - N_{\text{data}}) + \tau^2 \chi_{\text{reg}}^2$$

with  $\Sigma$  the data covariance matrix and  $\mathbf{x} \equiv N_{\text{unf}}$ .

- $\chi_{\text{reg}}^2 = (x - f * x_0)^T L^T L (x - f * x_0)$  used to favor regular solutions: Tikhonov regularization.
- matrix  $L$  used to select type of regularization: on the amplitude, the derivative or curvature.





CMS  $t\bar{t}$ ,  $1\ell$ +jets measurement,

[doi:10.1103/PhysRevD.97.112003](https://doi.org/10.1103/PhysRevD.97.112003)

Implemented by TUnfold from S. Schmitt, [JINST 7 \(2012\) T10003](#), included in ROOT.

# D'Agostini iterative method

[doi:10.1016/0168-9002\(95\)00274-X](https://doi.org/10.1016/0168-9002(95)00274-X) 

Also known as Lucy–Richardson deconvolution ([doi:10.1364/JOSA.62.000055](https://doi.org/10.1364/JOSA.62.000055) ,  
[doi:10.1086/111605](https://doi.org/10.1086/111605) )

An iterative method using the Bayes theorem ( $\rightarrow$  also called D'Agostini Bayes method)

$$\mathcal{P}(C_i|E_j) = \frac{\mathcal{P}(E_j|C_i) \mathcal{P}(C_i)}{\sum_{l=1}^{n_{\text{gen}}} \mathcal{P}(E_j|C_l) \mathcal{P}(C_l)} \quad (1)$$

$$\hat{N}_i^{\text{gen}} = \frac{1}{\epsilon_i} \sum_j \mathcal{P}(C_i|E_j) N_j^{\text{reco}} \quad (2)$$

$\mathcal{P}(E_j|C_i) \equiv R_{ji}$ : response matrix

$\epsilon_i = \sum_j \mathcal{P}(E_j|C_i)$ : reco efficiency.

# D'Agostini iterative method

$$\mathcal{P}(C_i|E_j) = \frac{\mathcal{P}(E_j|C_i) \mathcal{P}(C_i)}{\sum_{l=1}^{n_{\text{gen}}} \mathcal{P}(E_j|C_l) \mathcal{P}(C_l)} \quad (1)$$

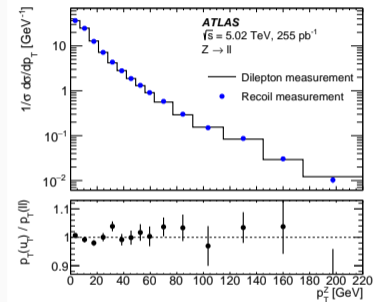
$$\hat{N}_i^{\text{gen}} = \frac{1}{\epsilon_i} \sum_j \mathcal{P}(C_i|E_j) N_j^{\text{reco}} \quad (2)$$

1. Start with some priors  $\mathcal{P}(C_i) = \mathcal{P}_0(C_i)$ : distribution from MC, flat prior, or some other choice;
2. Compute  $\hat{\mathcal{P}}(C_i|E_j)$  using eq. (1) with the priors;
3. Estimate  $\hat{N}^{\text{gen}}$  by injecting step-2  $\hat{\mathcal{P}}(C_i|E_j)$  in eq. (2);
4. Estimate new priors  $\mathcal{P}_1(C_i) = \hat{N}_i^{\text{gen}} / \sum_k \hat{N}_k^{\text{gen}}$  and repeat from step 2.

# D'Agostini iterative method

## Properties

- Converges to the MLE, although convergence can be slow in some cases.
- Runs fast.
- Regularization is obtained by stopping the iterations before convergence.
- $N_i^{\text{unf}}$  can be written as a linear combination of  $N_j^{\text{reco}}$ ,  $\mathbf{N}^{\text{unf}} = \mathbf{U} \cdot \mathbf{N}^{\text{reco}}$ .



ATLAS  $p_T(Z)$  measurement (uses D'Agostini unfolding). [arXiv:2404.06204](https://arxiv.org/abs/2404.06204)

# Maximum likelihood

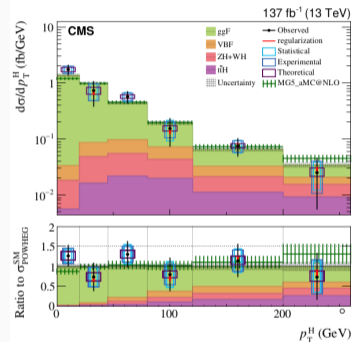
- Maximize a likelihood, e.g. using Minuit.
- If the measurement already uses a likelihood to extract reco-level event yields, use a single likelihood.
- See [doi:10.1007/JHEP03\(2021\)003](https://doi.org/10.1007/JHEP03(2021)003) measurement that uses this approach.

## Pros

- Simultaneous fit of signal, background and unfolding;
- Profiling of systematics;
- Poisson statistics.

## Cons

- Slow compared to the other methods that use linear algebra.
- Number of bin limit due to both computation time and fit stability: ok up to  $\mathcal{O}(100)$ . Use of ML fitting as in [doi:10.1103/PhysRevD.102.092012](https://doi.org/10.1103/PhysRevD.102.092012) may leverage this limitation.



CMS Higgs boson diff.

cross-section in  $WW(\rightarrow l\nu l\bar{\nu})$   
channel,

[doi:10.1007/JHEP03\(2021\)003](https://doi.org/10.1007/JHEP03(2021)003)

## Three regularization methods encountered in LHC data analyses

- Tikhonov regularization we saw before (Tikhonov, Soviet Math Dokl 4, 1035-1038). Can be used with both  $\chi^2$  and MLE methods.
- Early stopping in the D'Agostini iterative method
- SVD: smooth rejection of the smallest singular values  
([doi:10.1016/0168-9002\(95\)01478-0](https://doi.org/10.1016/0168-9002(95)01478-0))



# Choice of regularization strength

- To minimize bias it is important to make an objective selection of the regularization strength.
- Many methods on the market.
- Most used methods in LHC data analyses:
  - L-curve scan;
  - Minimization of global correlation;
  - Minimization of unfolding mean square error (MSE) using simulation;
  - Minimization of error on reunfolded data.

# Regularization strength choice: L-curve

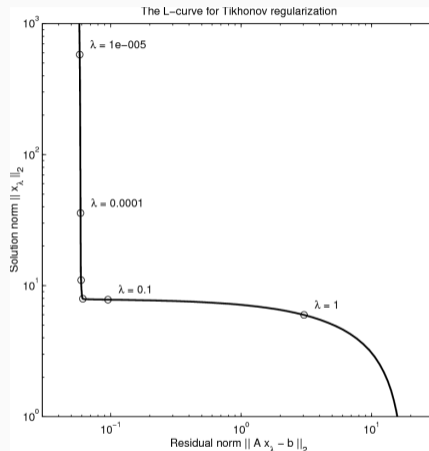
## L-curve

- Applies to minimization using Tikhonov regularization.
- Goal: find a compromise between fit residual minimization and solution regularity. [P. C. Hansen 2000, WIT Press](#)

## Method

- Draw the curve  $\log \chi_{\text{unf.}}^2$  vs  $\log \frac{\chi_{\text{reg.}, \tau}^2}{\tau^2}$ , with  $\tau$  as parameter.
- Select the  $\tau$  value of the point with the maximum curvature.

Specific to Tikhonov regularization. *Implemented in TUnfold*



from P. C. Hansen 2000

# Regularization strength choice: Global correlation

## Principle

Minimize the correlation between bins of the unfolded histogram.

## Implementation

- Scan the regularization strength values and select the value that minimizes the global correlation of the bins,  $\rho_i = \sqrt{1 - \frac{1}{(\Sigma_{ii} * \Sigma_{ii}^{-1})}}$
- Two options: use the mean or max of  $\rho_i$ .

*Implemented in TUnfold*

# Regularization strength choice: MSE

Method recommended by the [RooUnfold](#) manual

## Principle

- Minimize the error: difference between truth and estimation, including both bias and variance. Used the mean squared error (MSE), with average done over the bins.
- Use the simulation for which truth is known.

## Implementation

- Make replicas of the simulation reco histogram using a Poisson law for the bin content.
- Unfold each replica and compute the MSE with respect to the simulation gen histogram.
- Average MSE over the replicas.
- Select the regularization strength that minimizes the averaged MSE.
- Check that the error is small enough in every bin and uncertainty coverage is sufficient.

## Limitations

- Based on the simulation.
- If the shape of the truth distribution differs from the model used in the simulation, then the unfolding can behave differently.
- Especially, it seems important to use a flat prior for the D'Agostini iterative method and a flat bias for least square.
- Limitation can be alleviated by testing with different event generators, reweight the simulation to match with the observed reco histogram(s), or by distorting the distribution used for the test.

# Regularization strength choice: reunfold

used e.g. in [doi:10.1140/epjc/s10052-018-6373-0](https://doi.org/10.1140/epjc/s10052-018-6373-0)

Variation of MSE using data as template.

- Draw  $N$  replicas of the unfolded distribution using a Poisson law for the bin contents;
- “Fold” each replica by applying the response matrix and resample it.
- Unfold the  $N$  folded replicas and for each of them compute  $T = \sum_i (N_i^{\text{unf}} - N_i^{\text{gen}})^2 / N_i^{\text{gen}}$ .
- With a D’Agostini iterative unfolding,  $T$  will typically decrease with the number of iterations (approaching to the solution) and then increases (because of the fluctuations added by the unfolding). Select the minimum as working point.
- Check that the  $(N_i^{\text{unf}} - N_i^{\text{gen}})^2 / N_i^{\text{gen}}$  is small enough in every bin and uncertainty coverage is sufficient.

# Cross-checks: closure tests

Several cross-checks are usually performed to validate the unfolding.

## Closure test I

- Use MC to generate pseudo-data ( $\rightarrow$  gen- and reco- level distributions) and response matrix;
- Unfold the reco-level distribution
- Check that the unfolded distribution matches with the gen-level distribution

## Closure test II: sensitivity to the MC model

- Same as test I but using a different event generator for the MC sample used to extract the response matrix

## Bottom-line test

- The bottom line: unfolding should not enhance the measurement discrimination power between two models.
- The test:
  - Pick up a model for the true distribution  $\rightarrow \lambda_{\text{gen}}$ ;
  - Smear the model to obtain the reconstruction level distribution  $\rightarrow \lambda_{\text{reco}} = R\lambda_{\text{gen}}$
  - Compare the p-value of the  $\chi^2$ -tests of background-subtracted data vs  $\lambda_{\text{reco}}$  and of unfolded data vs  $\lambda_{\text{gen}}$ : the p-values must be similar and the one in the unfolded space should not be smaller than the one in the reco space.
- Beware the test is not valid in case of large regularization because the ndof for the unfold-space test is no more equal to the number of bins.  
<http://arxiv.org/pdf/1408.6500> provides a method to estimate ndof in such a case.



## coverage test

If the result is biased, then the uncertainty coverage will be too small.

$$\text{coverage} = \Phi\left(\frac{\text{bias}}{\sigma} + 1\right) - \Phi\left(\frac{\text{bias}}{\sigma} - 1\right)$$

with  $\Phi$ , the normal cumulative distribution function.

- Coverage can be checked using toy experiments.

- Reco-level statistical and systematics uncertainties to propagate to the unfolded measurement.
- Unfolding statistical uncertainties
- Unfolding model uncertainties: more in next slides

# Unfolding model uncertainties

## Limitations of the response matrix approach

- Sensitive to the modelling of the distribution within the bins;
- Dependency of event migration on other observables than the unfolded one(s) ignored  
→ e.g. unfolding of a  $p_T$  distribution sensitive to MC  $\eta$  distribution accuracy

## Unfolding model uncertainties

Because of this limitation the result depends on the accuracy of the event generator, and we should account for model uncertainties.

# Model uncertainties

Different methods used in LHC data analyses, based on computing alternative response matrices from:

- Gen. parameter variations (using weights produced by generators): energy scales (renormalization, factorization, parton showering), PDF,  $\alpha_S$  variations  
More variations can be included. E.g., for analyses with top quarks, colour reconnection, top mass, B-fragm.,  $h_{\text{damp}}$ .
- One (or more) alternative generator(s)  $\rightarrow$  used to derive an unc. or as cross check.
- Reweighted MC: variation based on the difference of data/MC reco distributions. E.g.,
  - *Measurements of differential cross sections for associated production of a W boson and jets in proton-proton collisions at  $\sqrt{s} = 8$  TeV, CMS collaboration, March 2017, [doi:10.1103/PhysRevD.95.052002](https://doi.org/10.1103/PhysRevD.95.052002)*
  - *A simultaneous unbinned differential cross section measurement of twenty-four Z+jets kinematic observables with the ATLAS detector, ATLAS collaboration, submitted to PRL, [arXiv:2405.20041](https://arxiv.org/abs/2405.20041)*

Note: in all methods, it is important to check at reco-level that the variations cover differences between data and simulation (by construction for the last one).

Machine learning opens a new avenue for unfolding our measurements

# A rich literature

- [OmniFold: A Method to Simultaneously Unfold All Observables](#)
- [Unfolding with Generative Adversarial Networks](#)
- [How to GAN away Detector Effects](#)
- [Machine learning approach to inverse problem and unfolding procedure](#)
- [Machine learning as an instrument for data unfolding](#)
- [Advanced event reweighting using multivariate analysis](#)
- [Unfolding by weighting Monte Carlo events](#)
- [Binning-Free Unfolding Based on Monte Carlo Migration](#)
- [Invertible Networks or Partons to Detector and Back Again](#)
- [Neural Empirical Bayes: Source Distribution Estimation and its Applications to Simulation-Based Inference](#)
- [Foundations of a Fast, Data-Driven, Machine-Learned Simulator](#)
- [Comparison of Machine Learning Approach to other Unfolding Methods](#)
- [Scaffolding Simulations with Deep Learning for High-dimensional Deconvolution](#)
- [Preserving New Physics while Simultaneously Unfolding All Observables](#)
- [Measurement of lepton-jet correlation in deep-inelastic scattering with the H1 detector using machine learning for unfolding](#)
- [Presenting Unbinned Differential Cross Section Results](#)
- [Feed-forward neural network unfolding](#)
- [Optimizing Observables with Machine Learning for Better Unfolding](#)
- [Unbinned profiled unfolding](#)

## Two approaches

- Iterative unfolding (Omnifold)
- Generative unfolding

*List from the  
HEPML Living  
Review*

## Principle

Exploit the following properties of binary classifiers: for two probability distributions of events, it approximates the likelihood ratio.

E.g. with a NN  $f(x)$  trained with a cross-entropy loss function,

$$\text{loss}(f(x)) = - \sum_{i \in \text{Cat.0}} \log f(x_i) - \sum_{i \in \text{Cat.1}} \log(1 - f(x_i))$$

we have<sup>1</sup>,

$$\frac{f(x)}{1 - f(x)} \approx \frac{p_0(x)}{p_1(x)}$$

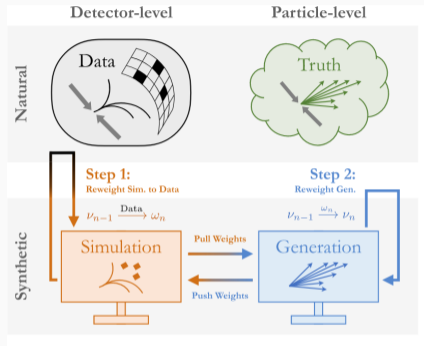
with  $p_i$  the probability to be in Category  $i$ .

<sup>1</sup>assuming the same number of events in both categories for the training

Generalizes the iterative D'Agostini method to unbinned unfolding of the full phase space

1. Train a classifier to distinguish if an event is from data or simulation  
 $\Rightarrow \mathcal{P}(\text{Data}|x_{\text{reco}})/\mathcal{P}(\text{Simu}|x_{\text{reco}})$
2. Reweight Simulated event with  $\mathcal{P}(\text{Data}|x_{\text{reco}})/\mathcal{P}(\text{Simu}|x_{\text{reco}})$
3. Train a second classifier to distinguish at gen. level if an event is from the original or the reweighted simulation  
 $\Rightarrow \mathcal{P}(\text{Reweighted}|x_{\text{gen}})/\mathcal{P}(\text{Original}|x_{\text{gen}})$
4. Reweight simulation with  $\mathcal{P}(\text{Reweighted}|x_{\text{gen}})/\mathcal{P}(\text{Original}|x_{\text{gen}})$
5. Repeat from 1

[PRL 124 182001 \(2020\)](#), [PRD 104 076027 \(2021\)](#)



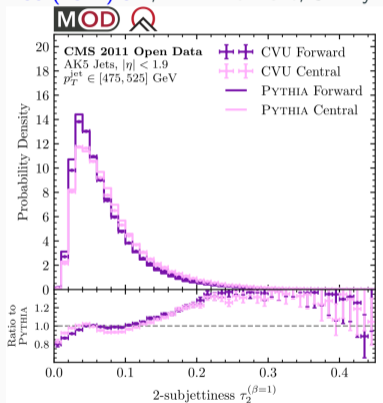
Can also be used on a limited number of observables: called Unifold for 1 observables and Multifold for more.



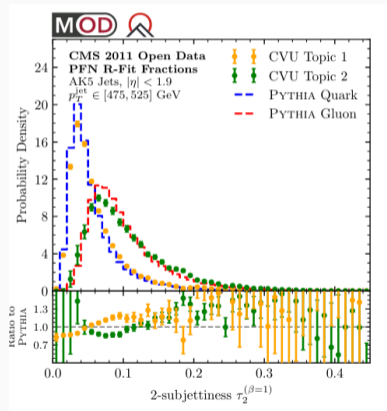
# Omnifold on LHC data 1/2

## Disentangling quarks and gluons in CMS open data

PRD 106 (2022) 9 [↗](#), P. T. Komiske, S. Kryhin, J. Thaler



Unfolded  $\tau_2$  distributions of the two jet categories compared to Pythia8.

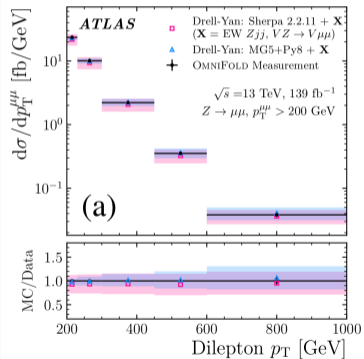


Distributions extracted for quark and gluons

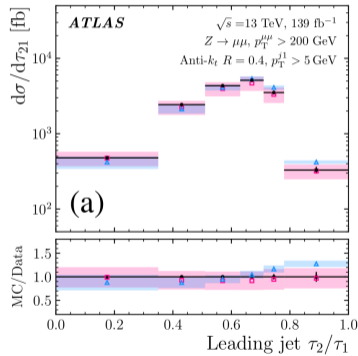
# Omnifold on LHC data 2/2

A simultaneous unbinned differential cross section measurement of twenty-four  $Z$ +jets kinematic observables with the ATLAS detector, ATLAS Collaboration, submitted to PRL, [arXiv:2405.20041](https://arxiv.org/abs/2405.20041).

Unbinned unfolding, although only binned distributions publicly released: 24 binned distributions.



Unfolded  $p_T(\mu\mu)$  distribution.



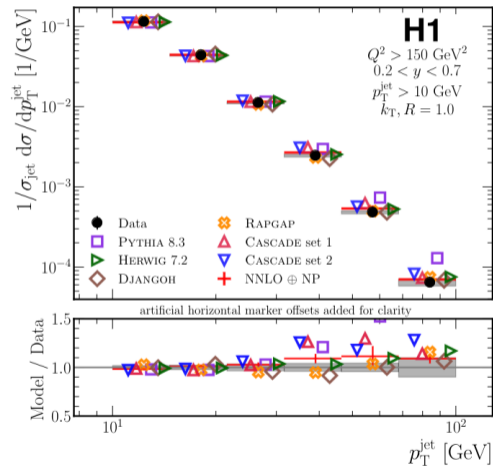
Distribution of the ratio of 2-subjettiness to 1-subjettiness

# Omnifold on Hera data

Measurement of lepton-jet correlation in deep-inelastic scattering with the H1 detector using machine learning for unfolding,

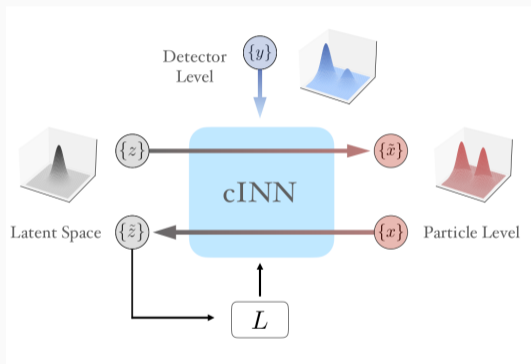
[doi:10.1103/PhysRevLett.128.132002](https://doi.org/10.1103/PhysRevLett.128.132002)

MultiFold of 8 observables,  $p_T^e$ ,  $p_z^e$ ,  $p_T^{\text{jet}}$ ,  $\eta_T^{\text{jet}}$ ,  $\varphi^{\text{jet}}$ ,  $q_T^{\text{jet}}/Q$ ,  $\Delta\varphi^{\text{jet}}$



# Generative unfolding

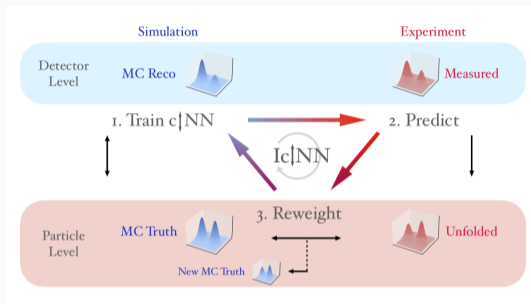
- Uses a conditional invertible neural network (cINN)
- Trained to generate a gen-level event on the condition of a reco-level event
- Apply to data to generate the unfolded distribution



[arXiv:2006.06685](#), [arXiv:1806.00433](#), [arXiv:1912.00477](#), [arXiv:2212.08674](#)

# Iterative generative unfolding

- Mitigate MC bias using iterations
- After the generative unfolding, use a classifier to learn ratio of unfolded to truth-level distribution and extract weights for the simulation
- Repeat the generative unfolding with the reweighted simulation



- Unfolding is widely used in HEP for differential cross-section measurements.
- Unfolding comes with a model uncertainty, difficult to estimate as often the case for systematic uncertainties.
- Machine learning allows handling of large numbers of dimensions opening a new avenue with unbinned full phase space unfolding.