Introduction to Unfolding
(Matrix- and Machine Learning-based methods)

Bogdan MALAESCU

LPNHE, CNRS

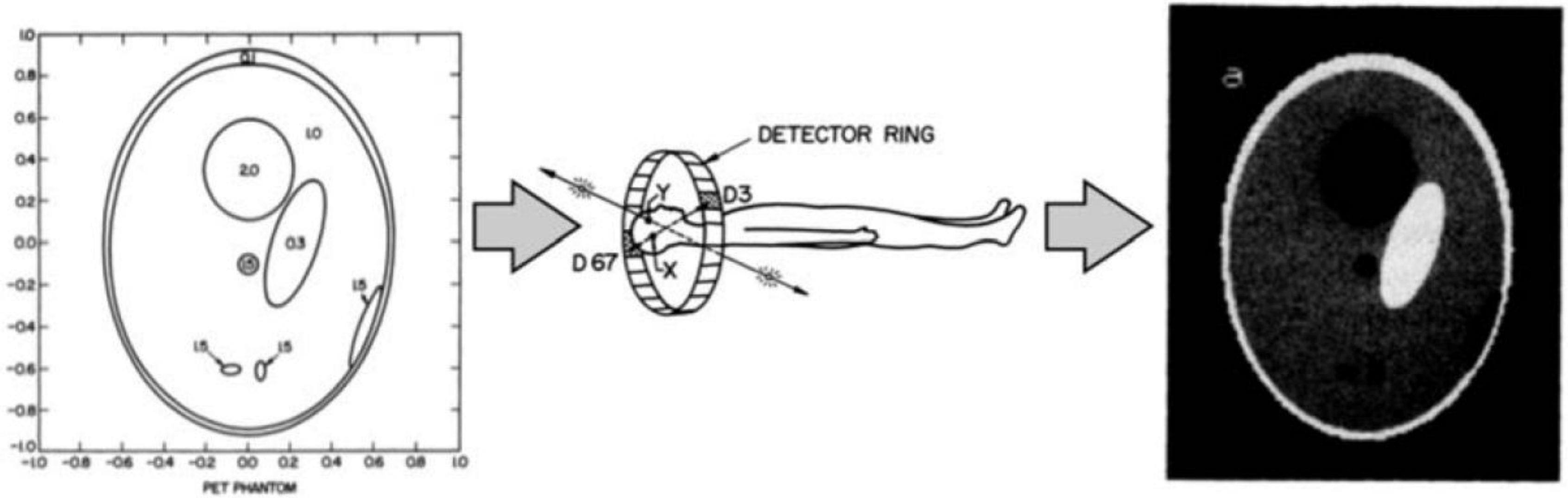
IML meeting
27/07/2023

Outlook

- Unfolding of detector effects: *why, where, how* ?
- Matrix-based methods, parameter setting, treatment of uncertainties
- How to publish unfolded results e.g. in HEPData
- Examples of physics studies using reconstructed-level / unfolded spectra
- Unfolding using ML-based methods:

  high potential and many new possibilities!

Examples of (un)folding problems: PET scan



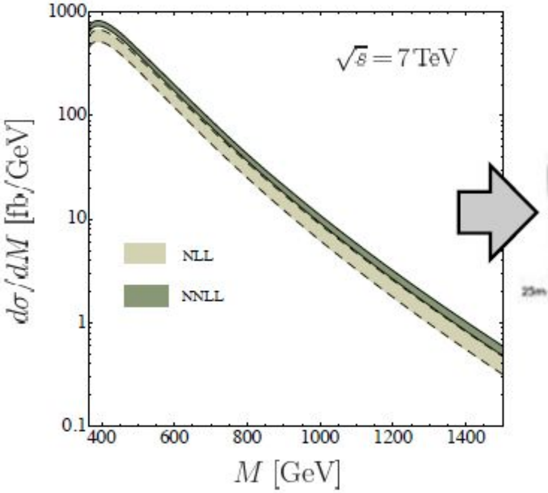
Positron Emission Tomography

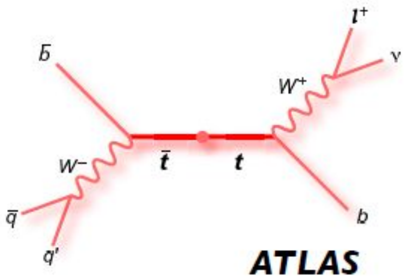
Y. Vardi et al.
http://www.jstor.org/stable/2288030

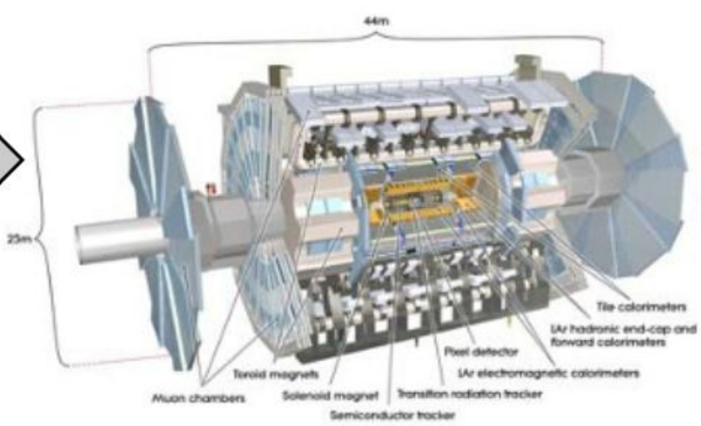
Examples of (un)folding problems: top pairs @ LHC

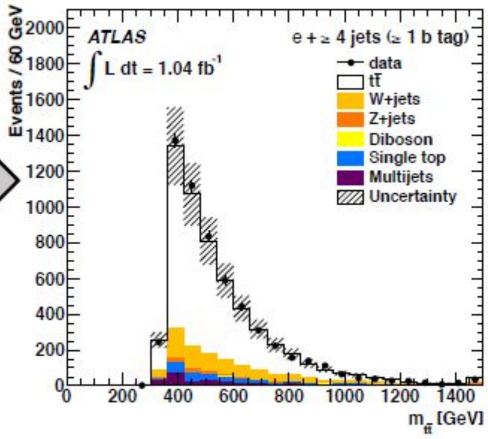
F. Spano

EPJ Web of Conferences 55, 03002 (2013)






44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

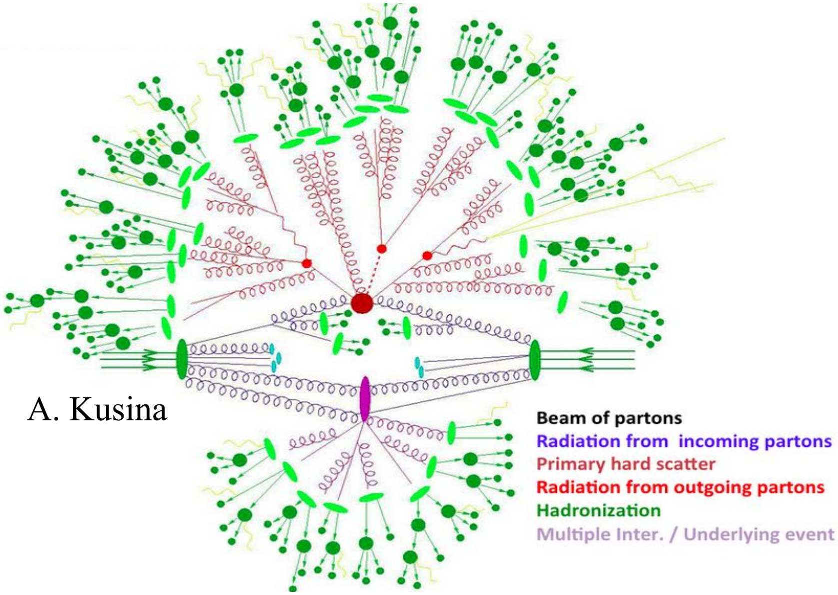


Folding

Unfolding

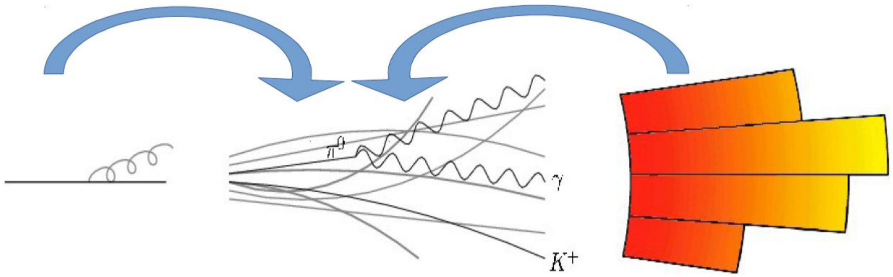
Environment and unfolding strategy for jet studies @ LHC

Typical proton-proton collision: a complex process in a difficult environment


Beam of partons
Radiation from incoming partons
Primary hard scatter
Radiation from outgoing partons
Hadronization
Multiple Inter. / Underlying event

A. Kusina

Pile-up



NP corrections

Hadronization & UE

Calibration+Unfolding

Jet energy response & resolution

parton level jet

particle level jet

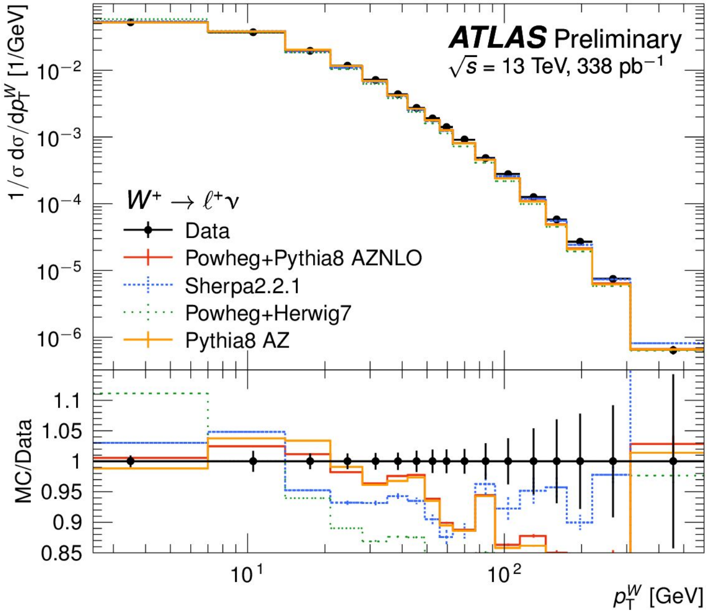
calorimeter (reconstructed) level jet

Data/theory comparison

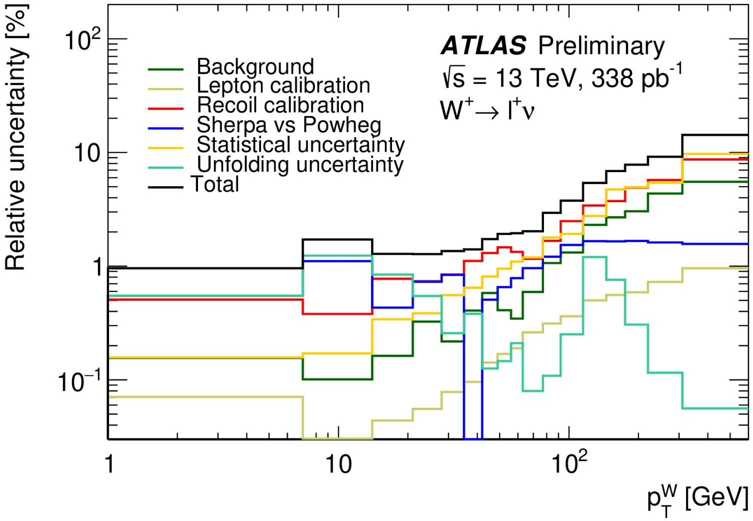
Goal: *publish data "corrected for detector effects" (on average, in the sense of an estimator), with minimal bias and minimal model dependence, with the full information needed for comparisons with theory predictions*

Unfolding problems: some challenging examples

→ $p_T(W)$: large resolution effects for MET reconstruction & need relatively fine binning in order to discriminate among theoretical predictions



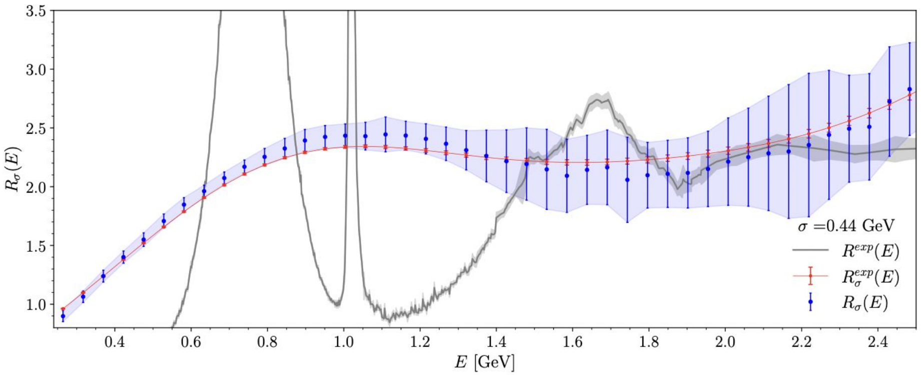
ATLAS-CONF-2023-028



→ Unfolding in a different context:

inverse Laplace transform to convert spacelike lattice QCD results into timelike quantities

arXiv:2212.08467

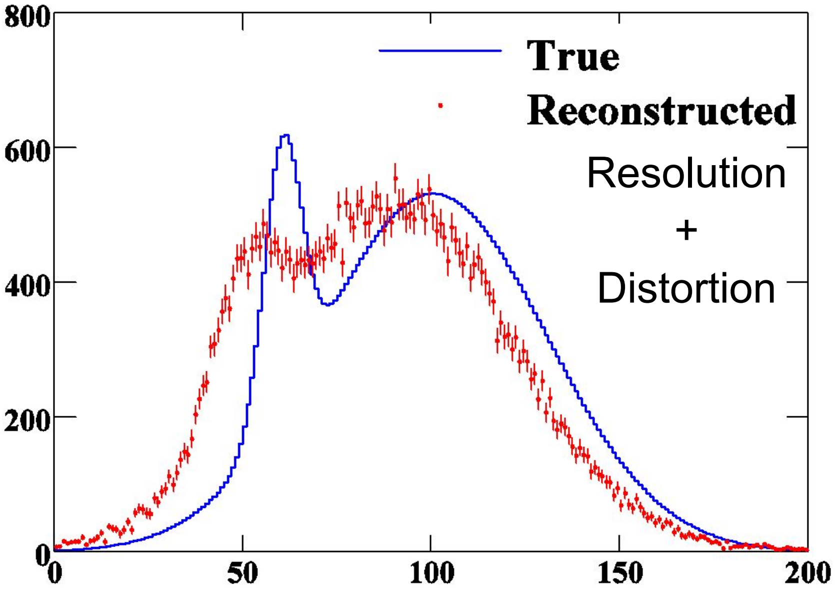


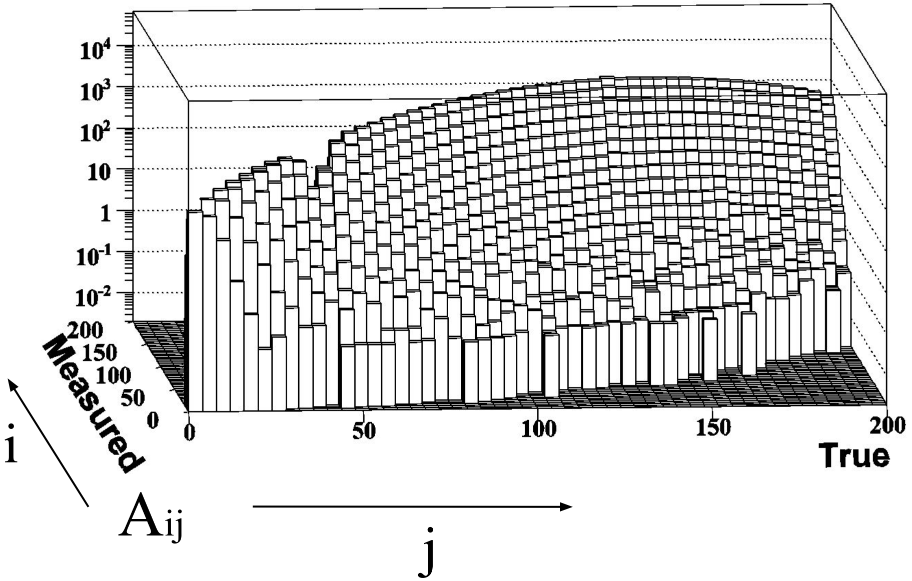
Why unfolding detector effects ?

- Enable interpretation “by eye” of images
- Direct comparison of measurements from different experiments
- Simplify phenomenological studies
- Data preservation and re-interpretation
- …

→ *Delicate exercise that has to be done with care*

Detector effects, folding and unfolding





- Folding: $f_{\mathrm{meas}}(y) = \int R(y|x)\, f_{\mathrm{true}}(x)\, dx$

$$P_{ij} = \frac{A_{ij}}{\sum_{k=1}^{n_d} A_{kj}} \quad ; d{=}P \cdot t$$

- Focus on unfolding of detector effects (acceptance correction factorized)
- Unfolding is generally not a simple numerical problem

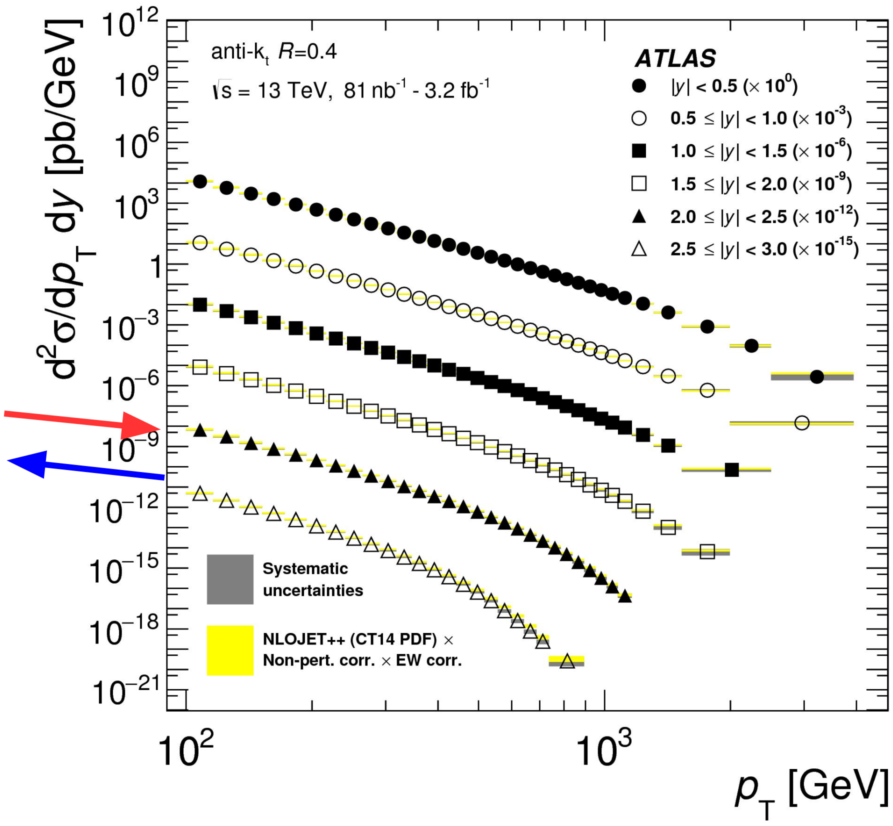
→ Regularization methods are often necessary

Choice of the phase-space

- Selection defining phase-space at “truth” level – as close as possible to the reconstructed-level selection: *minimize extrapolation to reduce model dependence*

- Include over-/under-flow bins when migrations to the region of interest are relevant

  → These extra bins are generally not published



Matrix-based unfolding methods + choice of method & settings

- *Maximum likelihood / matrix inversion*
- *SVD (+ Tikhonov regularization)*
- *Iterative Bayes-inspired regularized unfolding (IBU)*
- *Full Bayesian unfolding*
- *Iterative, dynamically stabilized (IDS) method*
- *Bin-by-bin correction* : d_i x $(T_i/R_i)^{MC}$→potentially large bias by relying on truth MC (used only when small bin-to-bin migrations & for statistics limited measurements e.g. Higgs differential Xsec; cross-check with matrix-based method)

- In general, recommended not to (dis)favor some particular method
- Recommended to evaluate the performance of *several methods & regularizations* and use the "optimal" one for the given unfolding study

→ Take into account: *systematic uncertainty related to the unfolding method (bias due to MC/data shape difference & regularization)*; impact on statistical uncertainties & correlations; constraints induced on binning choice

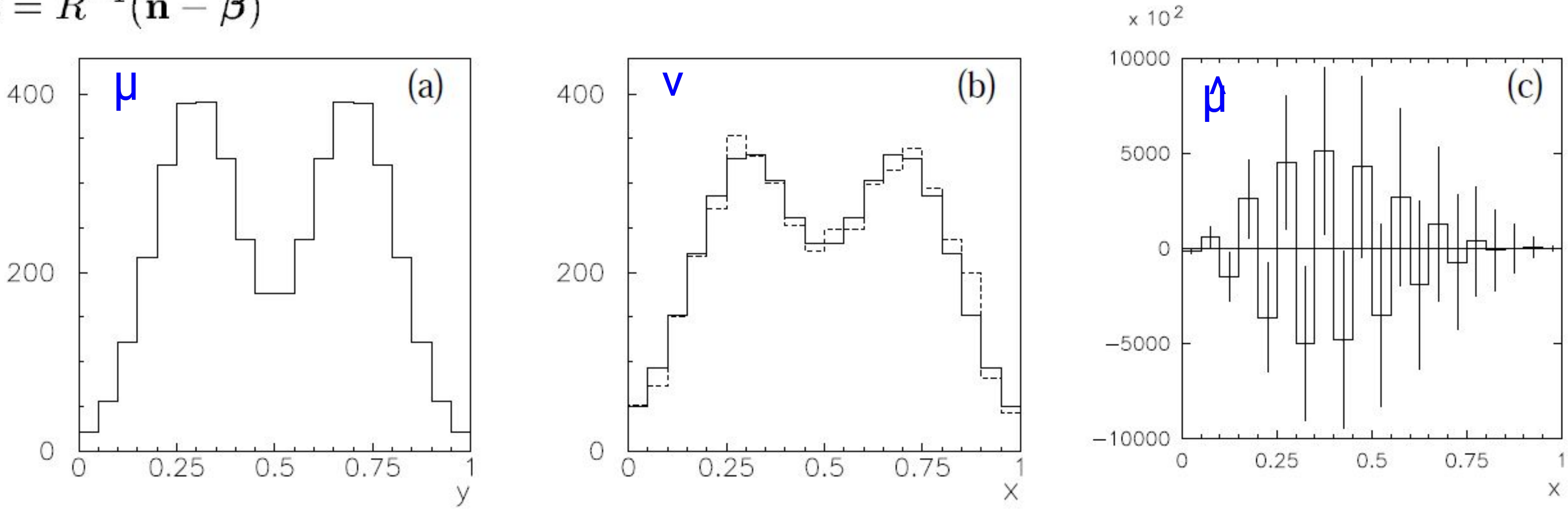
Matrix inversion

Folding of signal and background in data:

$$E[\mathbf{n}] = \boldsymbol{\nu} = R\boldsymbol{\mu} + \boldsymbol{\beta}$$

Unfolding based on matrix inversion:

$$\hat{\boldsymbol{\mu}} = R^{-1}(\mathbf{n} - \boldsymbol{\beta})$$



→ Inversion procedure unbiased, but induces large variances in unfolding result, as well as strong bin-to-bin (anti-)correlations: further use of the unfolded spectra require very precise determination of the covariance matrix (arXiv:2308.04221)

→ Unfolding is not a simple numerical problem → Regularization methods necessary

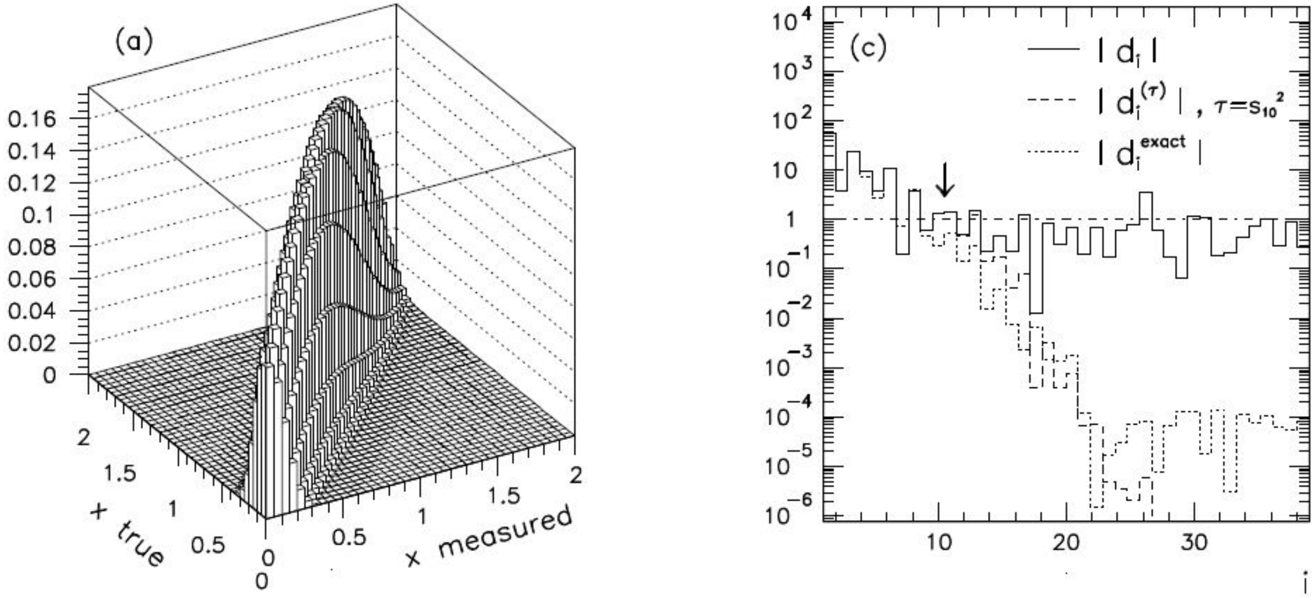
→ The binning itself provides a regularisation
Beware biases related to large binning (arXiv:2111.01091, ATLAS-CONF-2023-028, arXiv:1711.02692)

Singular Value Decomposition (SVD)

→ Inspired by the matrix inversion, but with regularization:

Suppress effect of small eigenvalues (~noise) + constraint on smoothness of the unfolded distribution → Regularization (may introduce bias)

$$S(\boldsymbol{\mu}) = -\sum_{i=1}^{M-2}[(\mu_{i+2}-\mu_{i+1})-(\mu_{i+1}-\mu_i)]^2$$



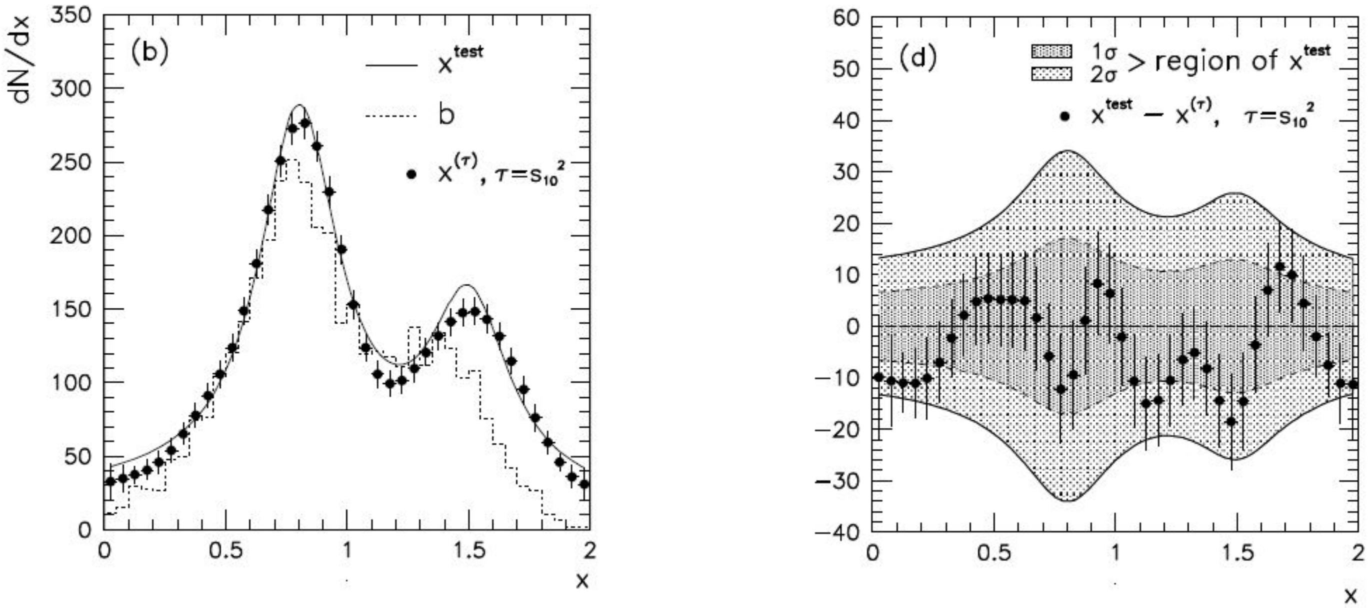
Nucl. Instr. Meth. A 372, 1996 (469)

Singular Value Decomposition (SVD)

→ Inspired by the matrix inversion, but with regularization:

Suppress effect of small eigenvalues (~noise) + constraint on smoothness of the unfolded distribution → Regularization (may introduce bias)

$$S(\boldsymbol{\mu}) = -\sum_{i=1}^{M-2}[(\mu_{i+2}-\mu_{i+1})-(\mu_{i+1}-\mu_i)]^2$$



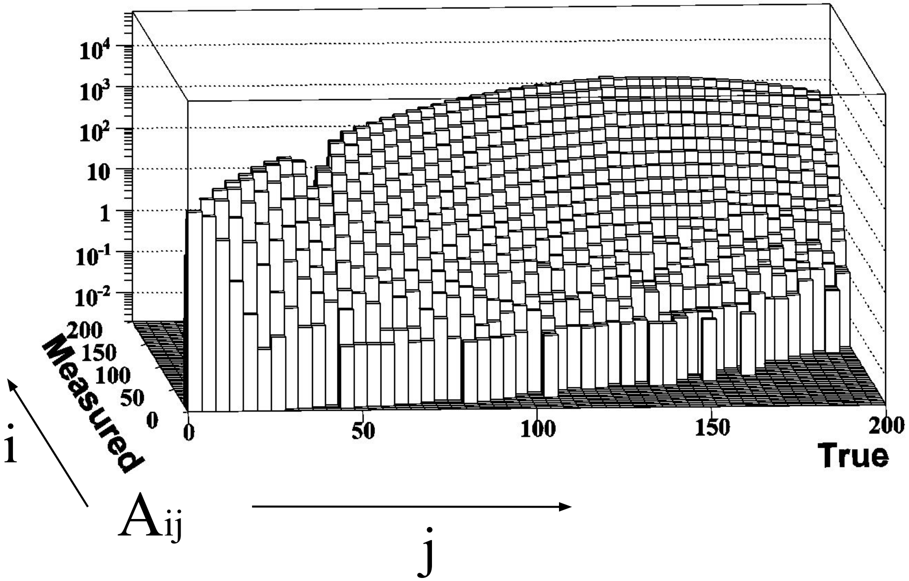
Nucl. Instr. Meth. A 372, 1996 (469)

An Iterative, Bayes-inspired Unfolding Method

$$P_{ij} = \frac{A_{ij}}{\sum_{k=1}^{nd} A_{kj}}$$

$$\tilde{P}_{ij} = \frac{A_{ij}}{\sum_{k=1}^{nu} A_{ik}} \; ; \boldsymbol{u} = \boldsymbol{\widetilde{P} \cdot d}$$

→ Note: $\tilde{P}_{ij}$ depends on the shape of the truth distribution in MC



- 1st unfolding, where the original transfer matrix is used

1)Transfer matrix improvement (hence of the unfolding probability matrix)

Reweight the truth MC distribution based on previous unfolding result.

2)Improved unfolding

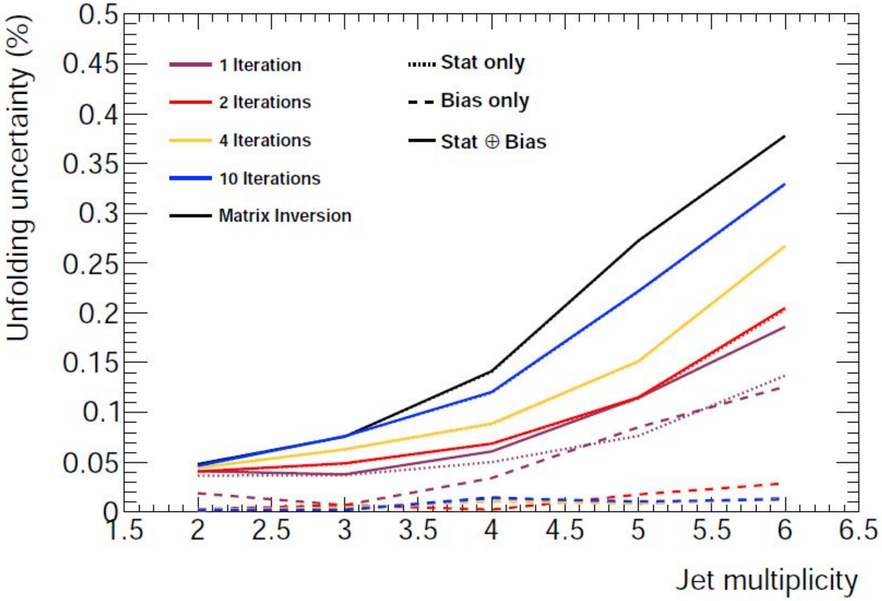
→ Choice on number of iterations = regularization (recommendations from previous slides apply)

→ Other methods exist, like e.g. dynamical local regularization in IDS (treatment of fluctuations in each bin, at each step of the procedure)

Iterative methods: choice of the number of iterations

- Number of iterations = regularization parameter: optimising variance / bias

Dustin Henry Urbaniec's PhD



- Compare data and the modified reconstructed MC: see how much information is left to be propagated from the data shape to the truth MC shape

→ bin-by-bin comparison or using a χ^2 (see e.g. arxiv:0907.3791, ATLAS-CONF-2023-028)

- Suggestion in IBU publication: compare results from consecutive steps (NIM A 362, 487 (1995))

→ risk of ~small changes between consecutive steps, while having a significant bias

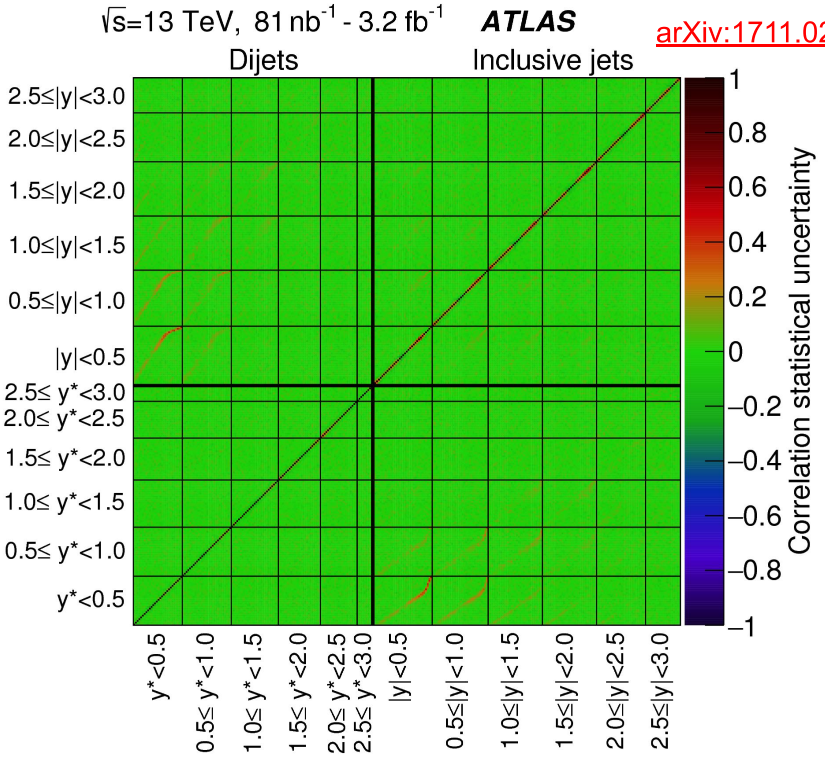
Statistical uncertainties

- Due to both data and MC
- Propagated using pseudo-experiments done separately/simultaneously for data and MC

→ Bootstrap method

- multiply event weights by random number: Poisson(1)
- seed given by event number
- allows to correlate measurements with overlapping samples

ATL-PHYS-PUB-2021-011
https://cds.cern.ch/record/2759945/
https://zenodo.org/record/5361038#.YTc7ni0Rpqs



arXiv:1711.02692

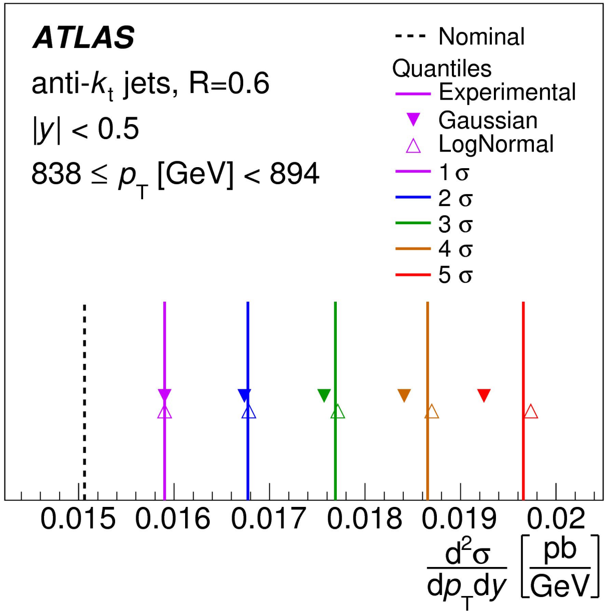
- Publish covariance matrix and/or a series of results based on each pseudo-experiment (i.e. Bootstrap replicas)
- Some unfolding methods provide estimates of the stat uncertainties

→ recommend cross-check with pseudo-experiments

Propagation of systematic uncertainties from inputs

- Modify input (pseudo-)data spectrum by ±1σ of the uncertainty, re-do unfolding and compare with nominal result

→ Can also use 1…5σ scans or pseudo-experiments



arXiv:1410.8857

→ Can shift reconstructed spectrum in transfer matrix instead of input spectrum: switched positive and negative variations

- For resolution uncertainties, perform smearing of the transfer matrix: smearing factor given by quadratic difference between resolution enhanced by 1σ and nominal resolution

Propagation of systematic uncertainties from inputs

- Bootstrap method to evaluate statistical uncertainties on the propagated systematics + rebinning/smoothing; (arXiv:1312.3524)

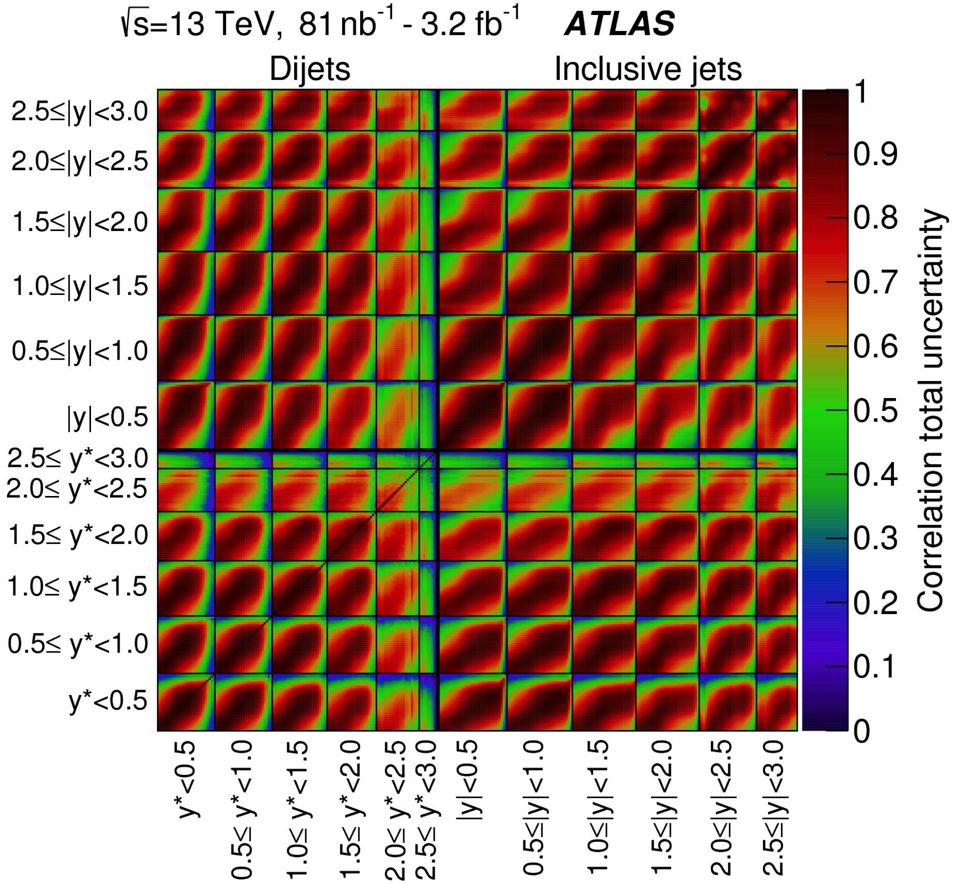
- Alternative propagation using pseudo-experiments (more difficult to probe e.g. 5σ effects)

- Alternative propagation option: include uncertainties as nuisance parameters in the definition of the response matrix + profile likelihood or Bayesian marginalization (often used for folding/template fits) (see e.g. arXiv:2304.03053)

Propagation of systematic uncertainties from inputs

- Split of systematics in sub-components (fully correlated in phase-space, independent between each-other) allows to evaluate correlations between different phase-space regions and between different measurements
- Information made available in HEPData tables (http://hepdata.cedar.ac.uk/)

$$Cov_{ij} = \sum_{k=1}^{N_{syst}} s_i^k \cdot s_j^k$$



arXiv:1711.02692

Tests of the unfolding

- “Technical closure test” → same MC for the transfer matrix and input distribution (pseudo-data) - expect perfect agreement between unfolding result and truth MC
- “Data-driven closure test” → allows to evaluate a systematic related to the unfolding method and the choice of regularization (see next slides)
- “Linearity test” → MC samples with various truth inputs; check linear dependence between unfolded and truth values of a quantity of interest
- “Pull test” → relevant only for unfolding methods providing an estimate of the statistical uncertainties (i.e. not from pseudo-experiments) - tests their reliability

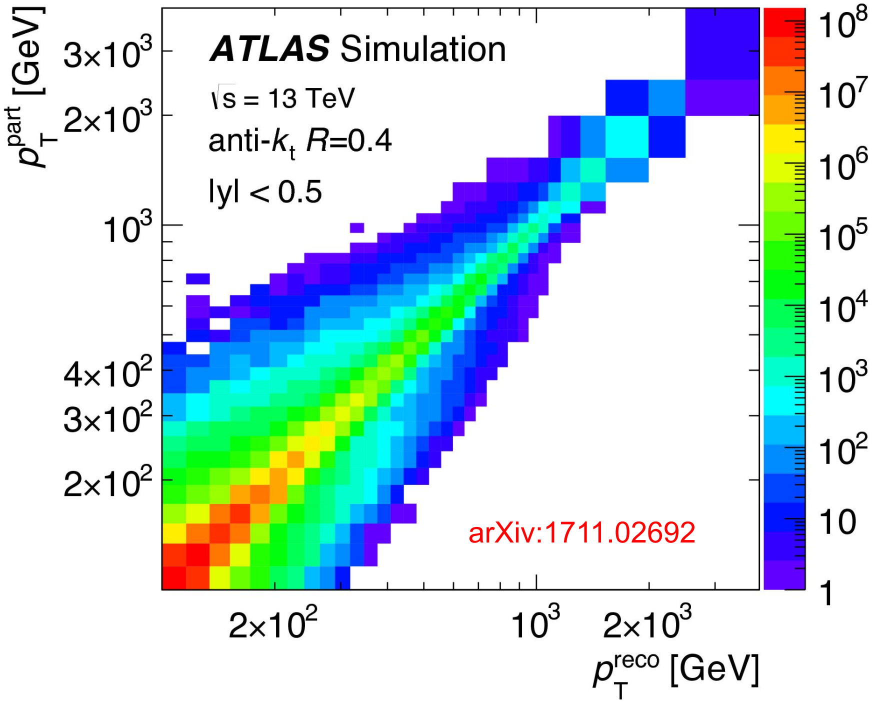
Data-driven closure test: motivation, procedure, example

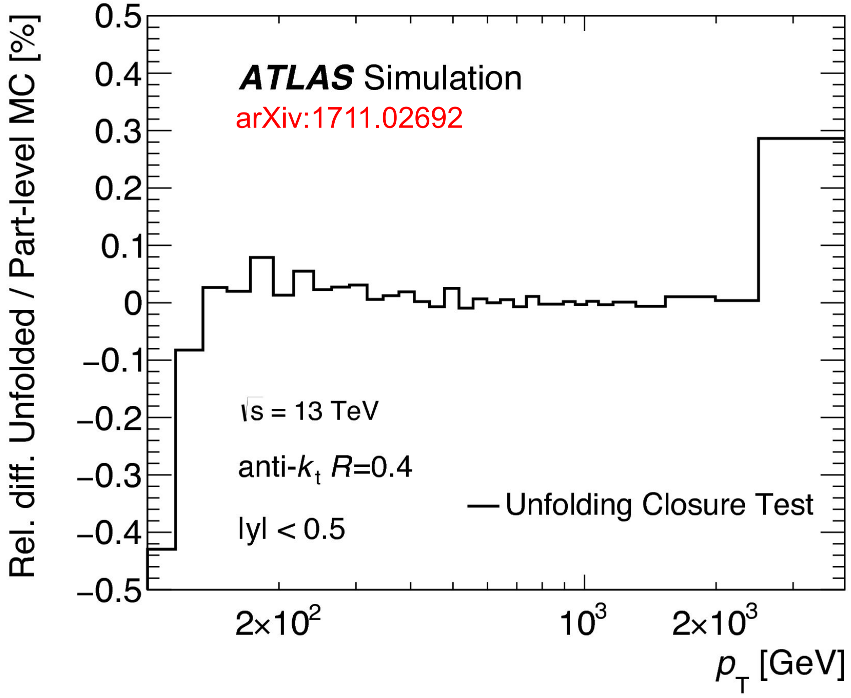
- In-situ (i.e. *realistic*) determination of the unfolding uncertainty related to the data/MC shape difference and to the regularization :

- reweight true MC by smooth function: improved data/recoMC agreement

- unfold the reweighted reconstructed MC

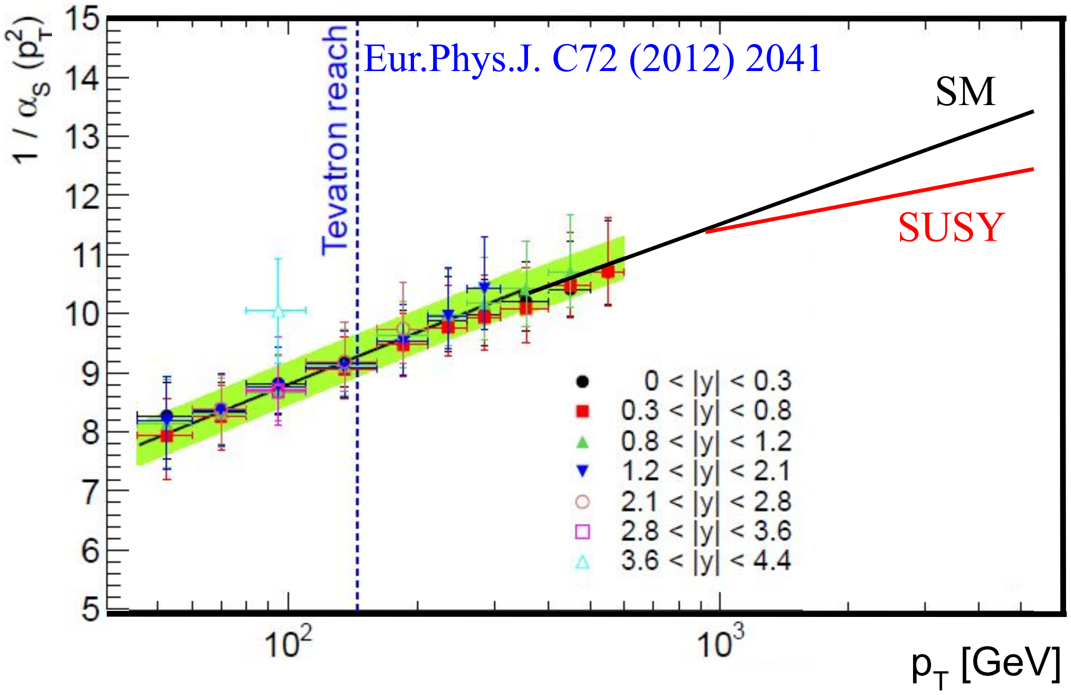
- compare with reweighted true MC

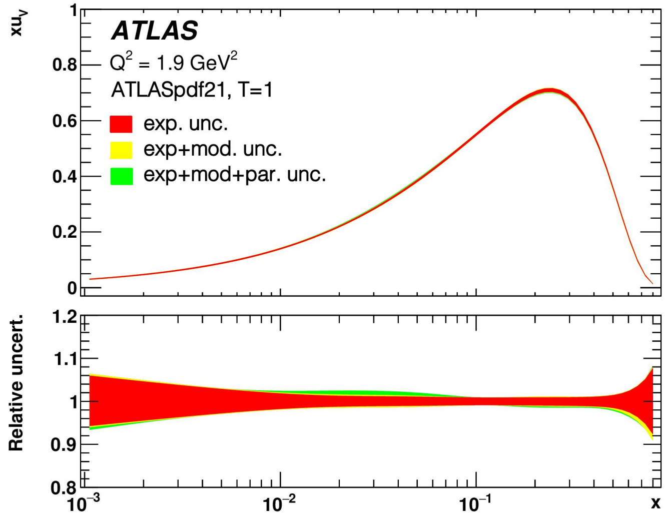


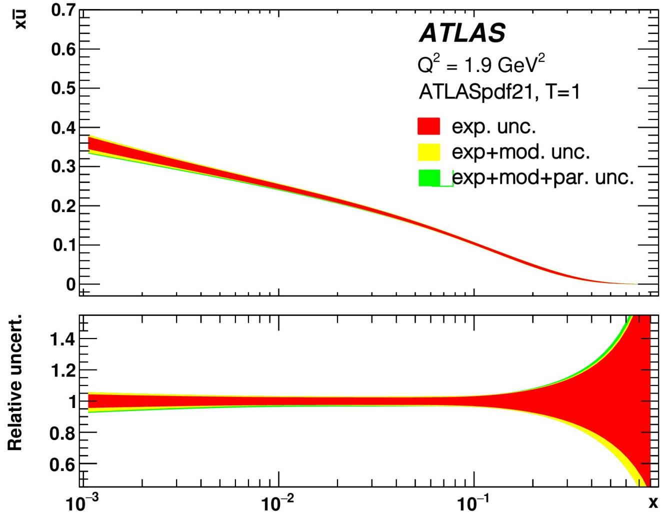


Method introduced in arXiv:0907.3791, used in arXiv:1112.6297 etc.

Extraction of Physics information from measurements



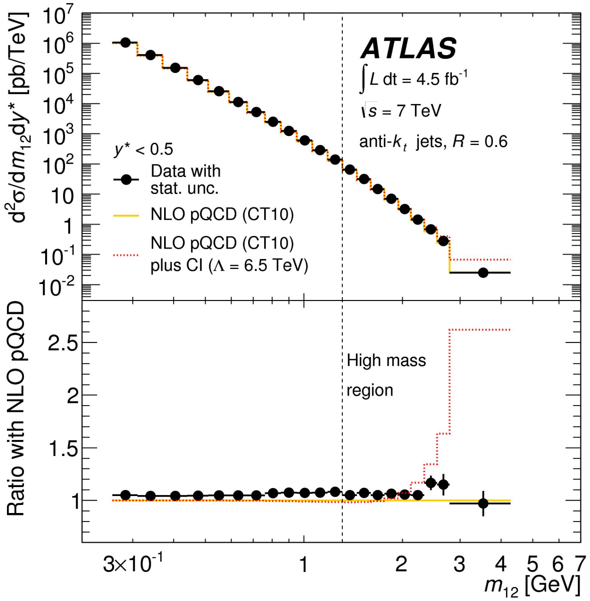


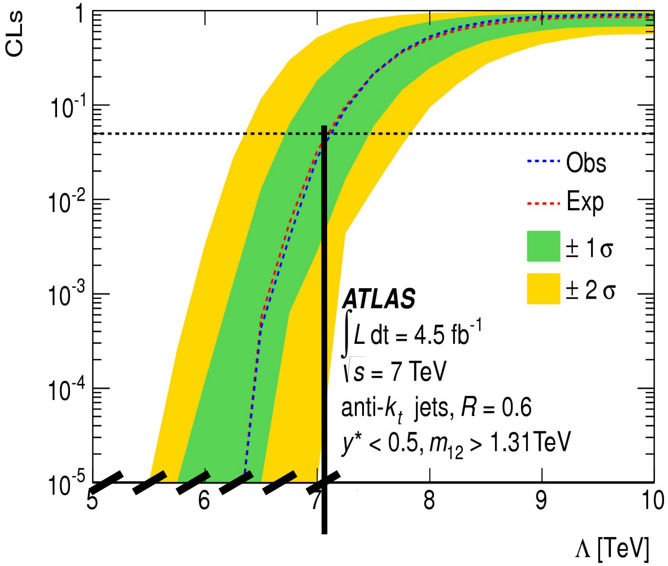


→ Involves using information on uncertainties and their correlations (between various measurement bins), keeping in mind that there are uncertainties impacting them too

Limits on New Physics using unfolded distributions

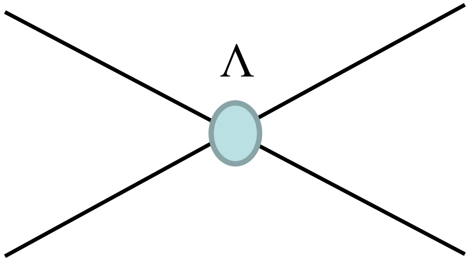
- Explore BSM physics directly at particle level





Contact Interaction Model (CI)

New force mediated by heavy particle



- Full frequentist analysis (CLs), with generalized χ^2 as test statistic

→ Accounts for correlations and asymmetries of uncertainties (stat. & syst.)

- Limits similar to the ones obtained by dedicated searches

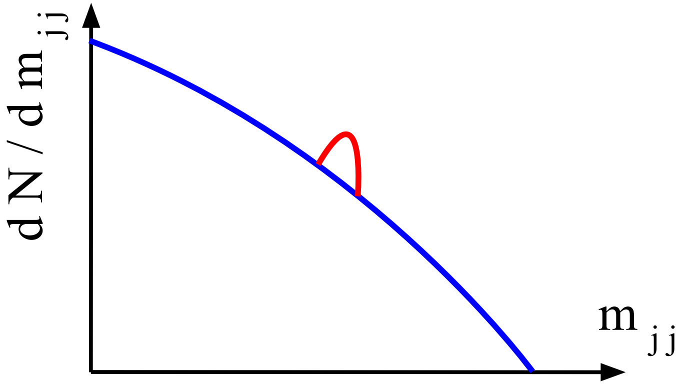
  (comparing reconstructed-level data with theory predictions folded with detector effects)

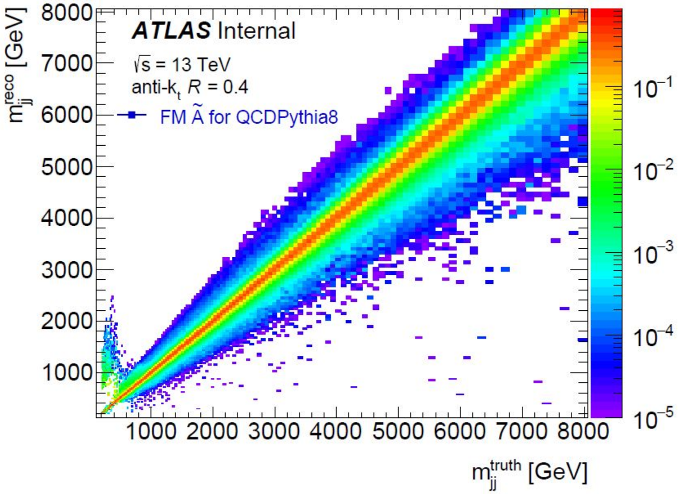
Generic Gaussian signals: folding-based method

→ Limits on generic Gaussian signals can be re-interpreted in terms of various signal models

→ Previously studied at reconstructed-level – hadron-level preferable

→ Folding method using MC-based transfer matrix allows to factorize physics & detector effects (publish limits more straightforward to use)

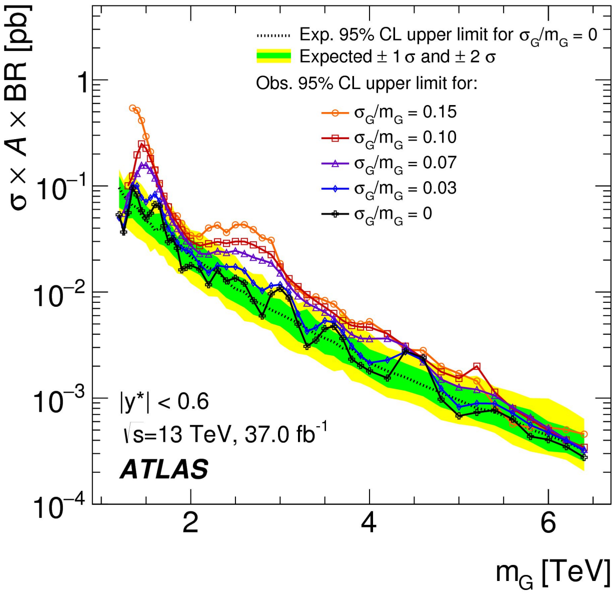




$f_y(M^{truth})$ = truth entries for a given model.

$F'_x(M^{reco})$ = the expected reco entries.

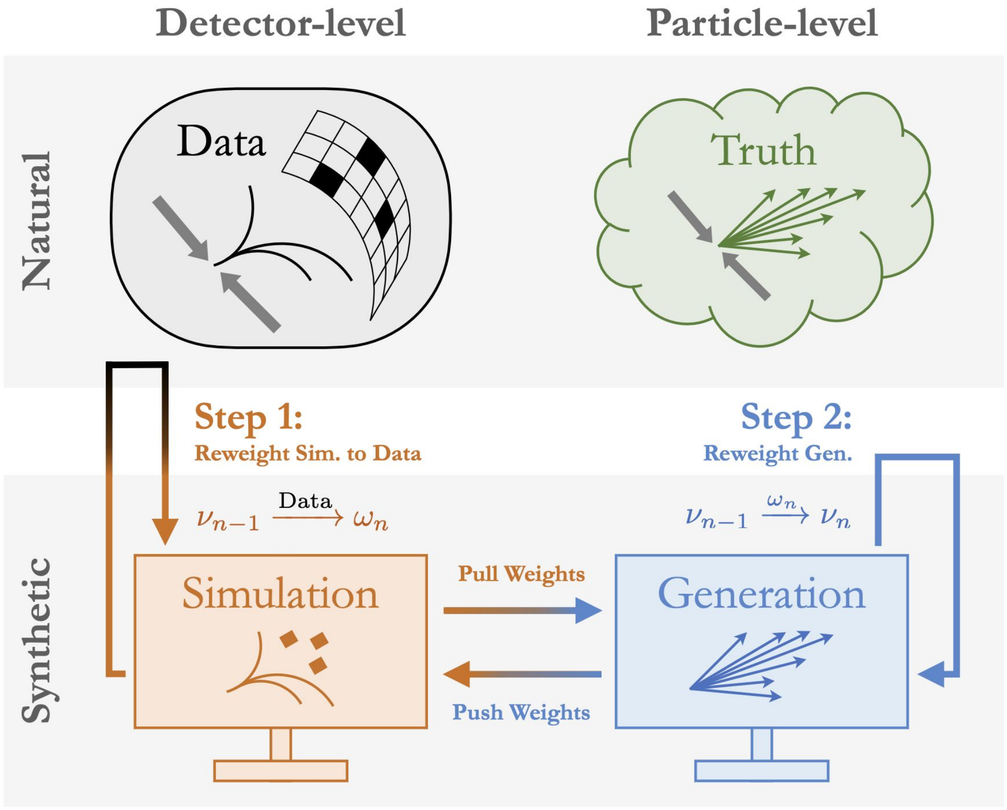
$$f_y(M^{truth}) \xrightarrow{\text{Folding}} F'_x(M^{reco}) = \sum_y f_y * \underbrace{E^T_y * A_{xy} / E^R_x}_{\tilde{A}}$$



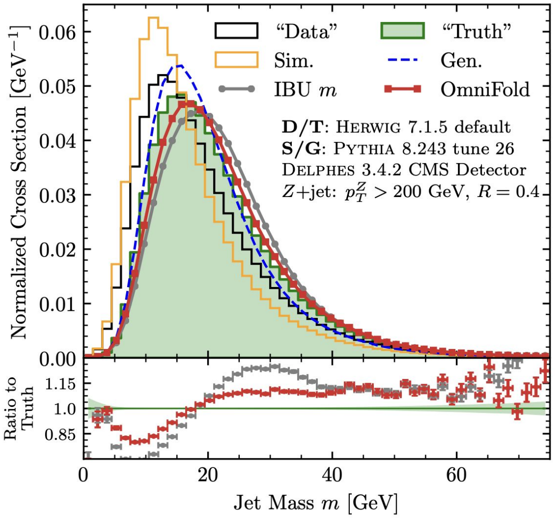
→ For resonance width ~ resolution: differences between folding result and reconstructed-level limits of up to 20% (different interpretation)

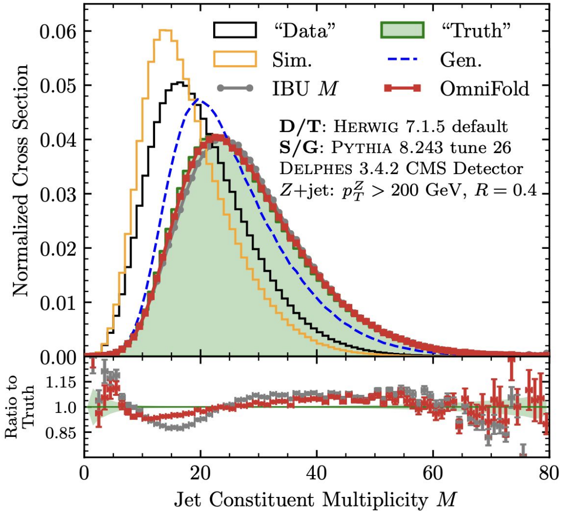
ML-based unfolding

- Binned (matrix-based) unfolding applicable up to 2-3 observables simultaneously (some of them being impacted by resolution effects more than others) : convert nD to 1D unfolding
- ML-based methods allow to enhance the dimensionality & obtain results event-by-event: enables computing secondary quantities arXiv:2109.13243
- OmniFold: iteratively improve (reweight) MC simulation; publish MC events & weights



arXiv:1911.09107; arXiv:2105.09923

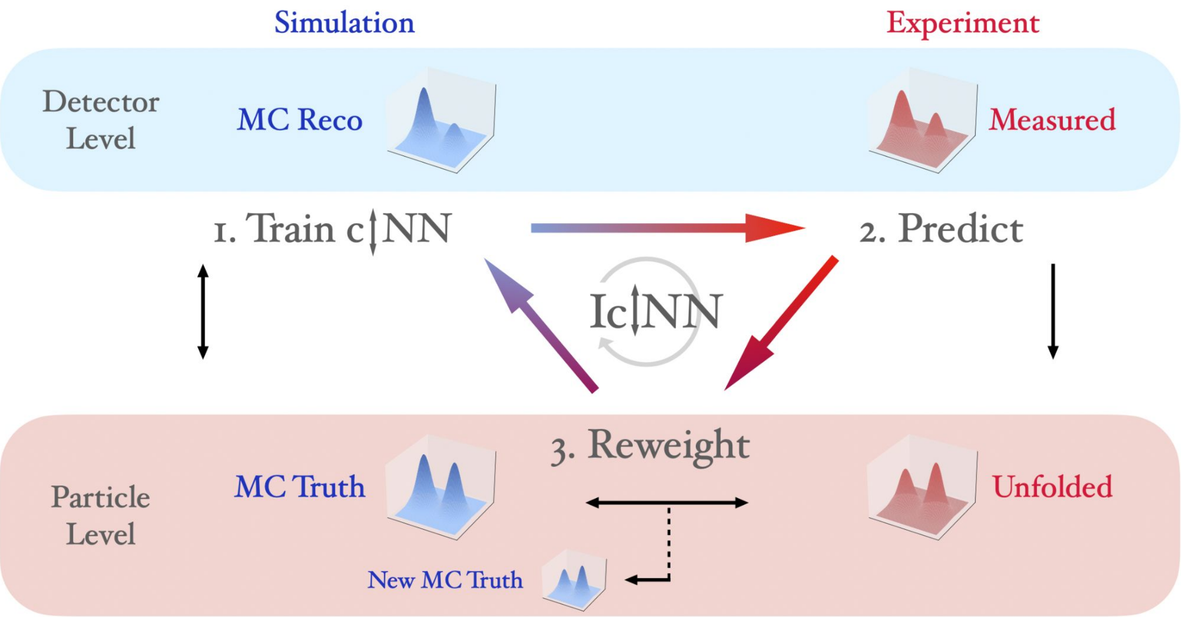


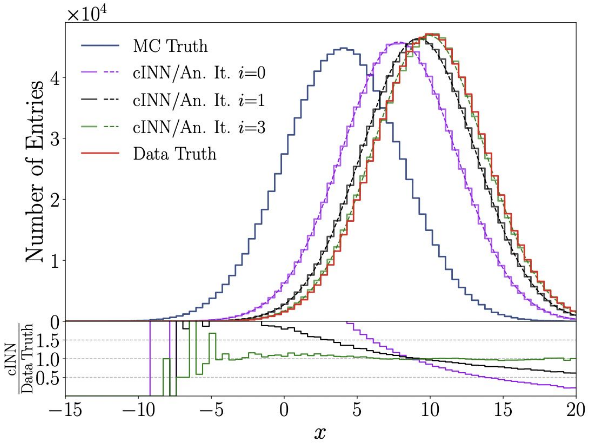


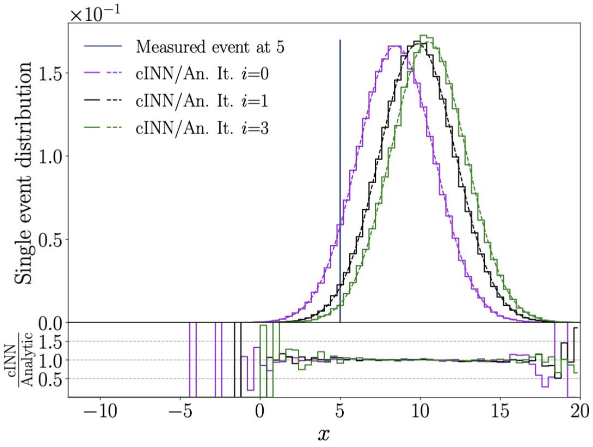
ML-based unfolding

- ML-based methods allow to enhance the dimensionality & obtain results event-by-event: enables computing secondary quantities arXiv:2109.13243

- IcINN: iteratively improve (reweight) MC simulation; publish unfolded distributions for each data event







arXiv:2212.08674

See talk by Mathias Backes

Instead of Conclusions

- Numerous topics on which we can have interesting discussions

Thank you !!!