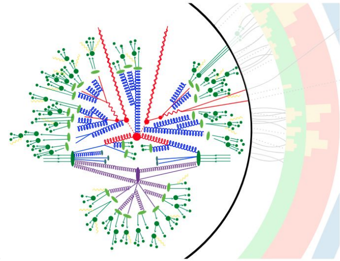
Unfolding Data with Machine Learning

vmikuni@lbl.gov

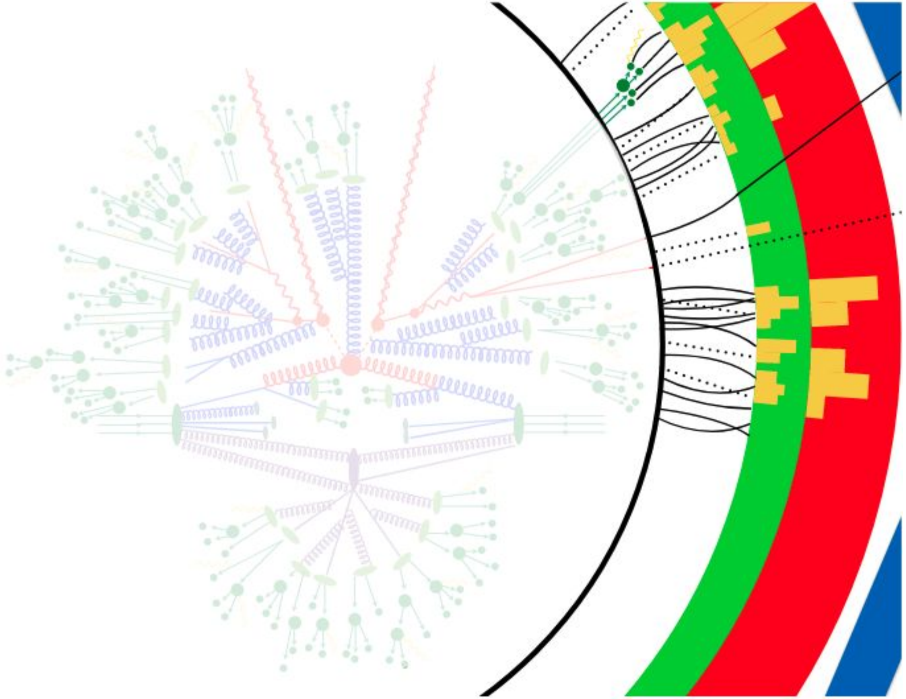
vinicius-mikuni

Vinicius Mikuni

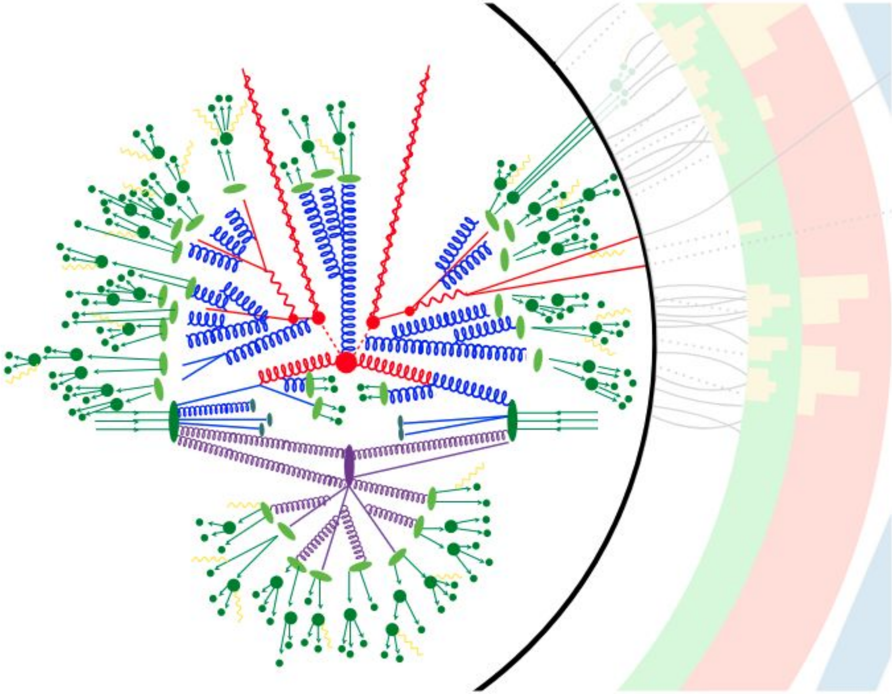
Unfolding



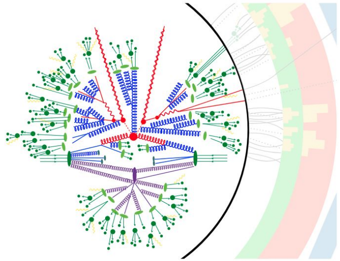
What we measure



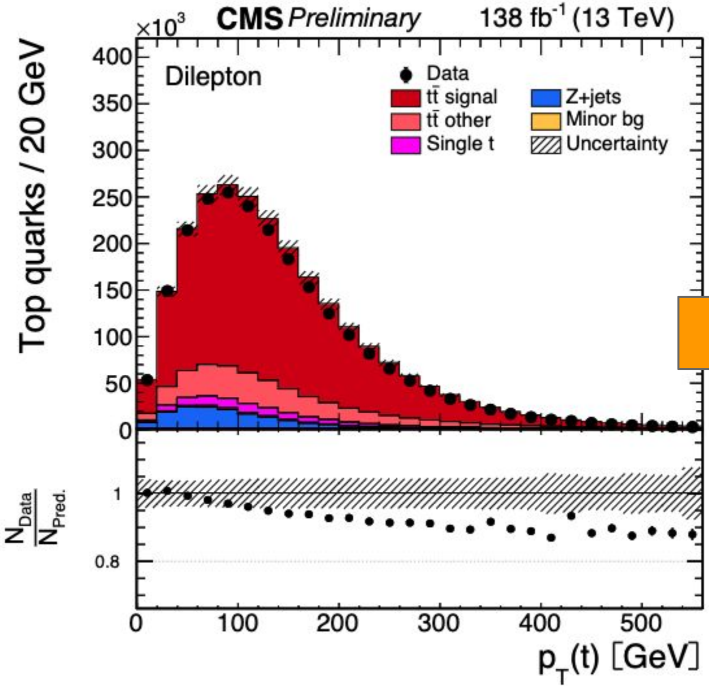
What we want

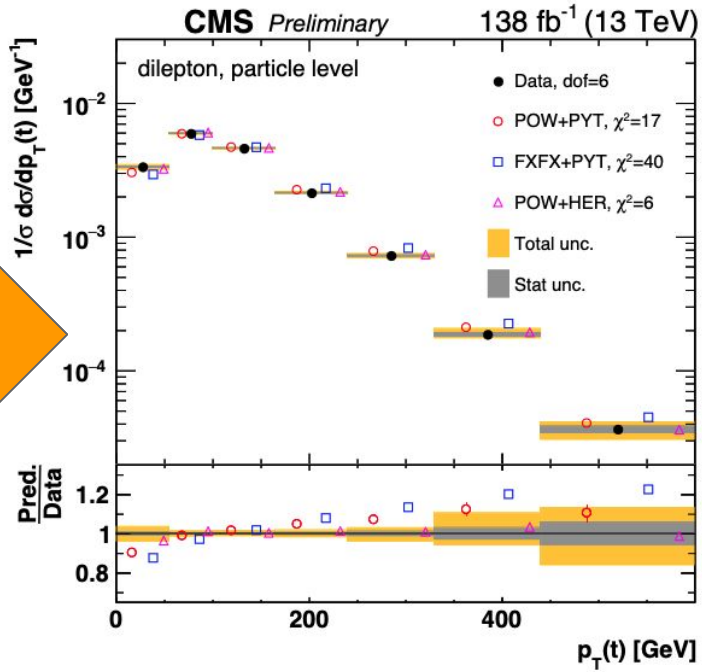


Unfolding



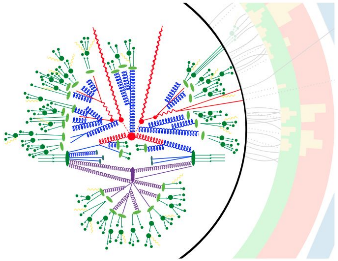
Source: CMS-PAS-TOP-20-006



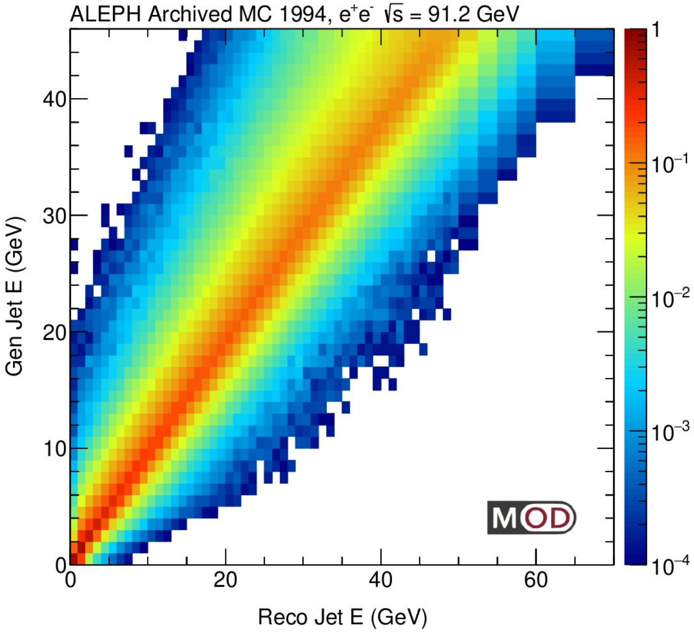


Traditional methods for **unfolding** use **histograms**

Unfolding



Source: MITHIG-MOD-21-001

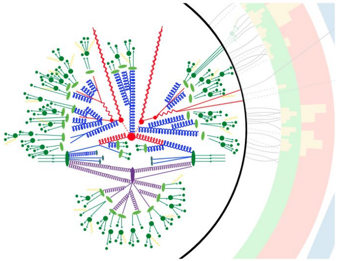


$$a_i = R_{ij} b_j$$

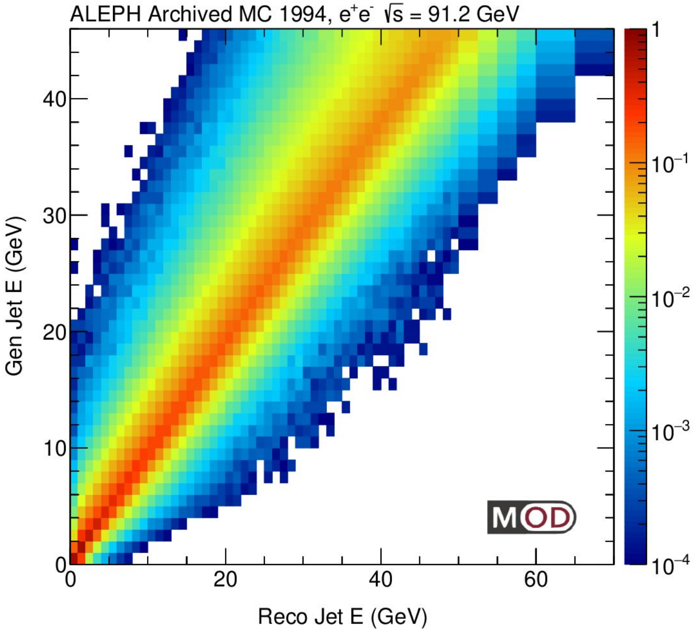
R_{ij} is the response matrix: **P(observed in bin i | true in bin j)**

Traditional methods for **unfolding** use **histograms**

Unfolding



Source: MITHIG-MOD-21-001

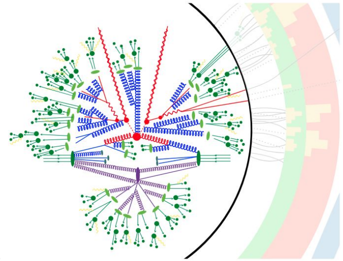


$$a_i = \mathbf{R_{ij}} b_j$$

$\mathbf{R_{ij}}$ is the response matrix: **P(observed in bin i | true in bin j)**

- Traditional unfolding is all about inverting the matrix R_{ij}

Traditional methods for **unfolding** use **histograms**

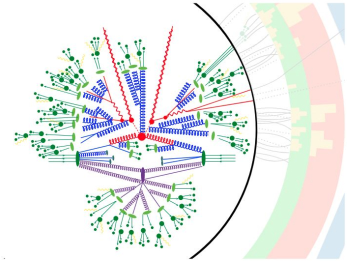


Unfolding

How to define the **optimal binning**?

- Choice depends on the **distribution** and **phase space**
- Need to compromise when **combining** results from **different experiments**

Unfolding



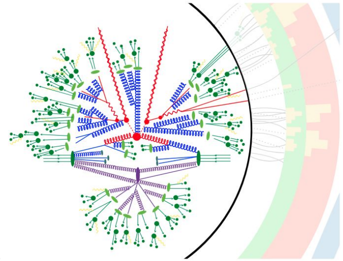
How to define the **optimal binning**?

- Choice depends on the **distribution** and **phase space**
- Need to compromise when **combining** results from **different experiments**

How to include **multiple distributions**?

- Histograms are hard to scale: **curse of dimensionality**
- Unfolding uncertainties can be reduced using **additional observables**

Unfolding



How to define the **optimal binning**?
- Choice depends on the **distribution** and **phase space**
- Need to compromise when **combining** results from **different experiments**

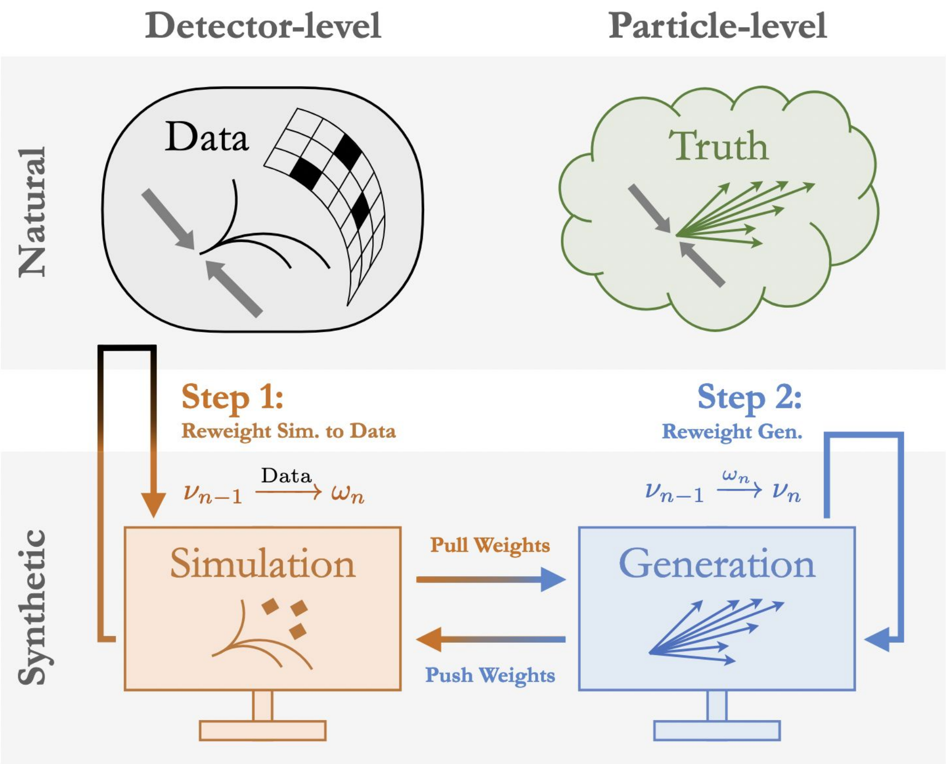
How to include **multiple distributions**?
- Histograms are hard to scale: **curse of dimensionality**
- Unfolding uncertainties can be reduced using **additional observables**

How to unfold distributions that are **not** defined for each event?
- Moments of distributions
- Energy Correlators

Going beyond histograms: Omnifold*

For unfolding using **invertible networks** see:

- SciPost Phys. 9 (2020) 074 e-Print: [2006.06685](2006.06685)



ML is used to define a method for unfolding that is unbinned and can use multiple distributions at a time

2 step iterative approach

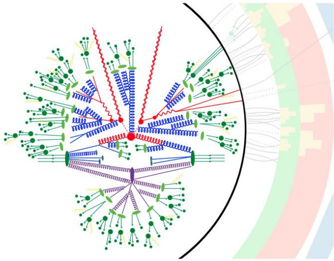
- Simulated events after detector interaction are reweighted to match the data
- Create a “new simulation” by transforming weights to a proper function of the generated events

Machine learning is used to approximate **2** likelihood functions:

- **reco MC to Data** reweighting
- **Previous** and **new Gen** reweighting

* Andreassen et al. PRL 124, 182001 (2020)

Omnifold



EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

Submitted to: Phys. Rev. Lett.

CERN-EP-2024-132
May 31, 2024

A simultaneous unbinned differential cross section measurement of twenty-four Z+jets kinematic observables with the ATLAS detector

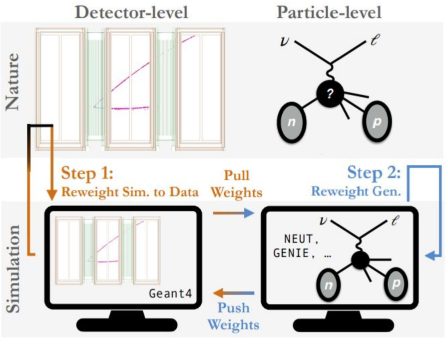
The ATLAS Collaboration

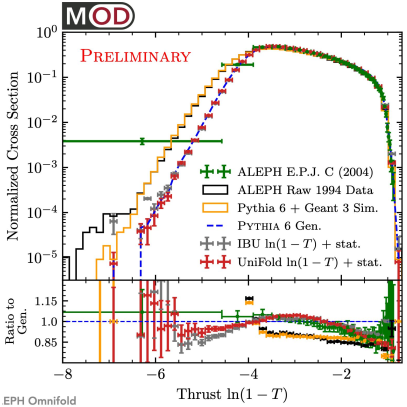
CMS Physics Analysis Summary

Contact: cms-pag-conveners-smp@cern.ch 2024/06/03

Measurement of event shapes in minimum bias events from pp collisions at 13 TeV

The CMS Collaboration



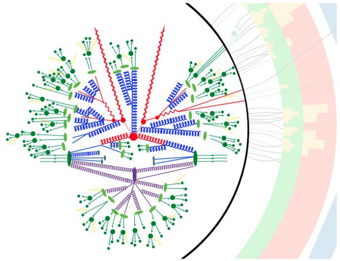


EPH Omnifold

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

V. Andreev,[23] M. Arratia,[35] A. Baghdasaryan,[46] A. Baty,[16] K. Begzsuren,[39] A. Belousov,[23,*] A. Bolz,[14] V. Boudry,[31] G. Brandt,[13] D. Britzger,[26] A. Buniatyan,[6] L. Bystritskaya,[22] A.J. Campbell,[14] K.B. Cantun Avila,[47] K. Cerny,[28] V. Chekelian,[26] Z. Chen,[37] J.G. Contreras,[47] L. Cunqueiro Mendez,[27] J. Cvach,[33] J.B. Dainton,[19] K. Daum,[45] A. Deshpande,[38] C. Diaconu,[21] G. Eckerlin,[14] S. Egli,[43] E. Elsen,[14] L. Favart,[4] A. Fedotov,[22] J. Feltesse,[12] M. Fleischer,[14] A. Fomenko,[23] C. Gal,[38] J. Gayler,[14] L. Goerlich,[17] N. Gogitidze,[23] M. Gouzevitch,[42] C. Grab,[49] T. Greenshaw,[19] G. Grindhammer,[26] D. Haidt,[14] R.C.W. Henderson,[18] J. Hessler,[26] J. Hladký,[33] D. Hoffmann,[21] R. Horisberger,[43] T. Hreus,[50] F. Huber,[15] P.M. Jacobs,[5] M. Jacquet,[29] T. Janssen,[4] A.W. Jung,[44] H. Jung,[14] M. Kapichine,[10] J. Katzy,[14] C. Kiesling,[26] M. Klein,[19] C. Kleinwort,[14] H.T. Klest,[38] R. Kogler,[14] P. Kostka,[19] J. Kretzschmar,[19] D. Krücker,[14] K. Krüger,[14] M.P.J. Landon,[20] W. Lange,[48] P. Laycock,[41] S.H. Lee,[3] S. Levonian,[14] W. Li,[16] J. Lin,[16] K. Lipka,[14] B. List,[14] J. List,[14] B. Lobodzinski,[26] E. Malinovski,[23] H.-U. Martyn,[1] S.J. Maxfield,[19] A. Mehta,[19] A.B. Meyer,[14] J. Meyer,[14] S. Mikocki,[17] M.M. Mondal,[38] A. Morozov,[10] K. Müller,[50] B. Nachman,[5] Th. Naumann,[48] P.R. Newman,[6] C. Niebuhr,[14] G. Nowak,[17] J.E. Olsson,[14] D. Ozerov,[43] S. Park,[38] C. Pascaud,[29] G.D. Patel,[19] E. Perez,[11] A. Petrukhin,[42] I. Picuric,[32] D. Pitzl,[14] R. Polifka,[34] S. Preins,[35] V. Radescu,[30] N. Raicevic,[32] T. Ravdandorj,[39] P. Reimer,[33] E. Rizvi,[20] P. Robmann,[50] R. Roosen,[4] A. Rostovtsev,[25] M. Rotaru,[7] D.P.C. Sankey,[8] M. Sauter,[15] E. Sauvan,[21,2] S. Schmitt,[14] B.A. Schmookler,[38] L. Schoeffel,[12] A. Schöning,[15] F. Sefkow,[14] S. Shushkevich,[24] Y. Soloviev,[23] P. Sopicki,[17] D. South,[14] V. Spaskov,[10] A. Specka,[31] M. Steder,[14] B. Stella,[36] U. Straumann,[50] C. Sun,[37] T. Sykora,[34] P.D. Thompson,[6] D. Traynor,[20] B. Tseepeldorj,[39,40] Z. Tu,[41] A. Valkárová,[34] C. Vallée,[21] P. Van Mechelen,[4] D. Wegener,[9] E. Wünsch,[14] J. Žáček,[34] J. Zhang,[37] Z. Zhang,[29] R. Žlebčík,[34] H. Zohrabyan,[46] and F. Zomer[29]
(The H1 Collaboration)

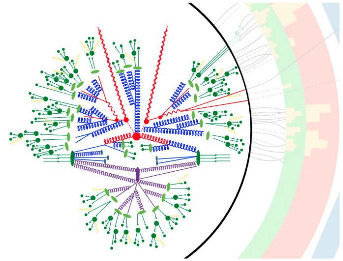
Increasing adoption by experimental collaborations: **ATLAS, CMS, H1, T2K, Aleph**



Omnifold

Reco level Data MC

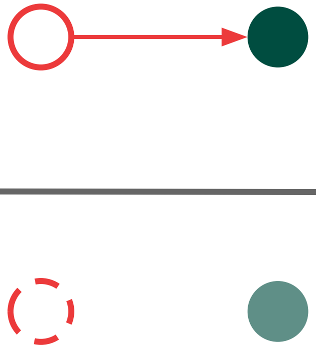
Generator level Data MC



Omnifold

Reco level ● Data ○ MC

Iteration 1



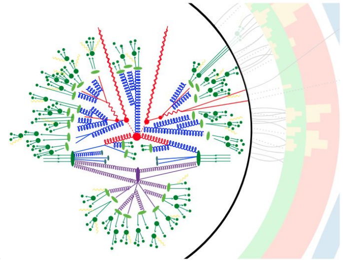
Generator level ● Data ◌ MC

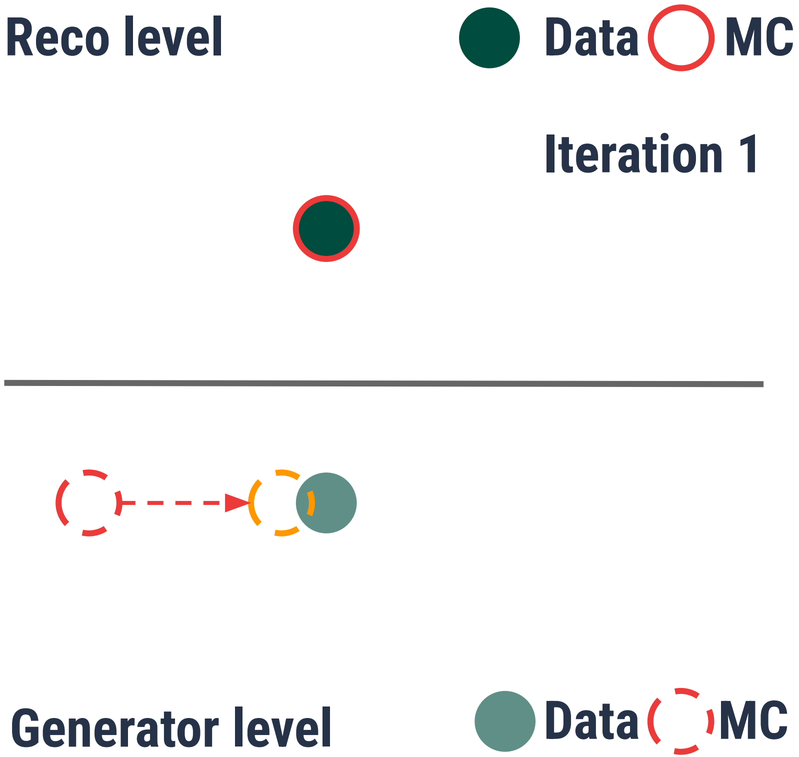
Step 1:

- Train a classifier to separate **data** from **MC** events
- Reweight **reco level MC** with weights:

$$W(\mathrm{reco}) = p_{\mathrm{Data}}(\mathrm{reco})/p_{\mathrm{MC}}(\mathrm{reco})$$

Omnifold



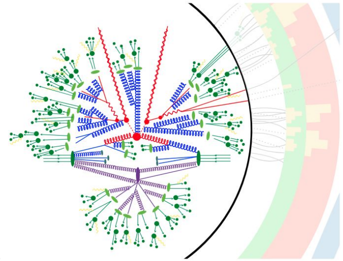


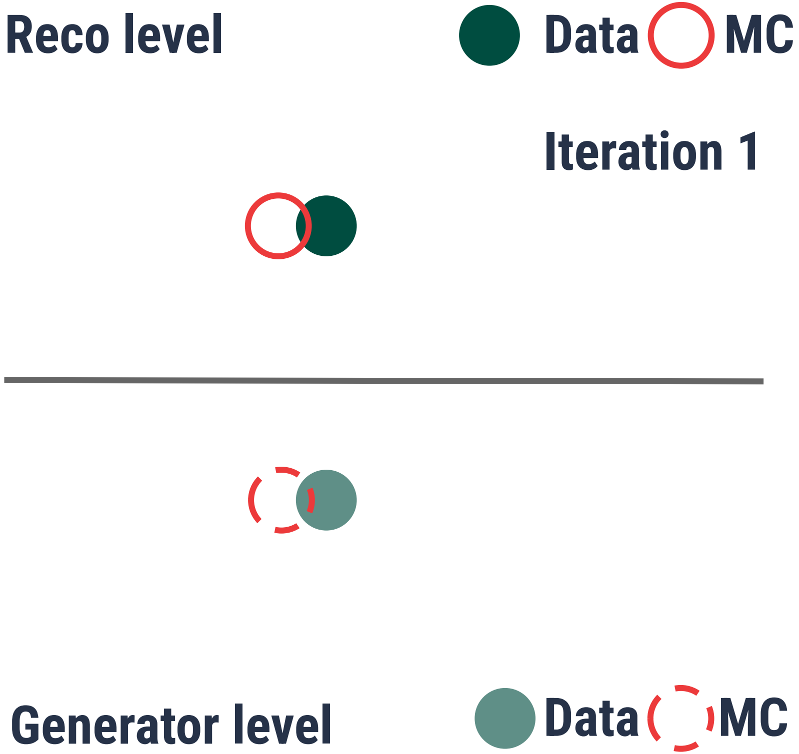
Step 2:

- **Pull weights** from **step 1** to generator level events
- Train a classifier to separate **initial MC at gen level** from **reweighted MC** events
- Define a **new simulation** with weights that are a **proper function of gen level kinematics**

$$W(\mathrm{gen}) = p_{\mathrm{weighted\ MC}}(\mathrm{gen})/p_{\mathrm{MC}}(\mathrm{gen})$$

Omnifold

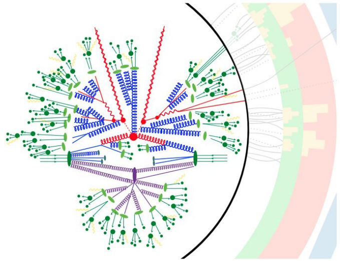


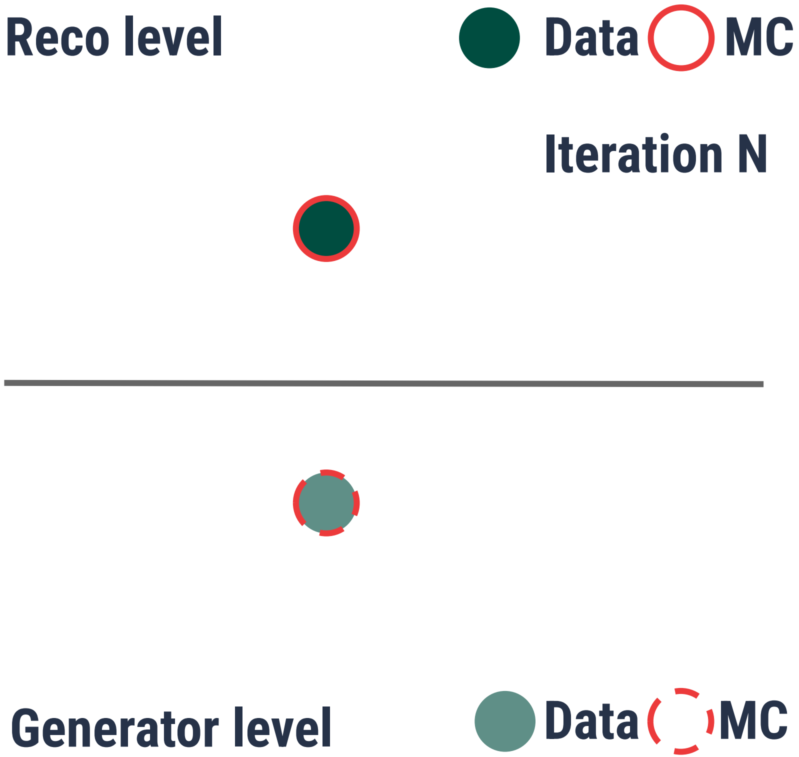


Start again from **step 1** using the **new simulation** after **pushing** the weights from **step 2**

- Guaranteed convergence to the maximum likelihood estimate of the generator-level distribution when number of iterations go to infinite
- In practice, less than 10 iterations are enough to achieve convergence

Omnifold





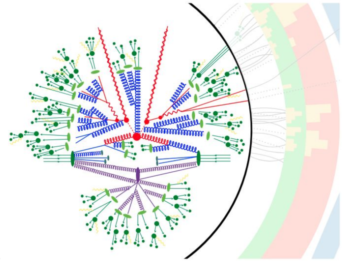
Start again from **step 1** using the **new simulation** after **pushing** the weights from **step 2**

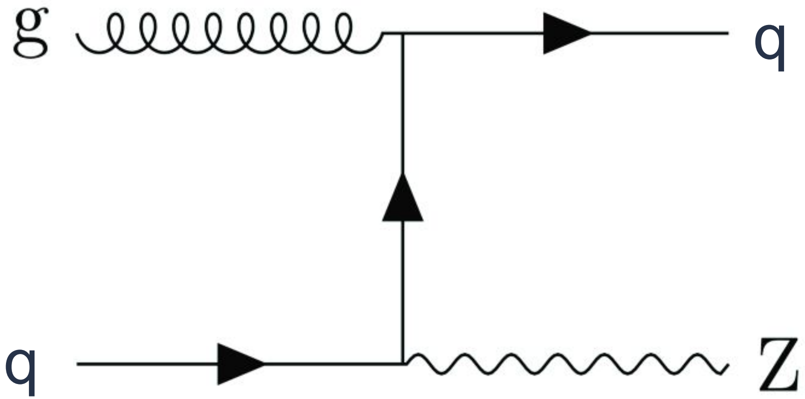
- **Guaranteed convergence** to the maximum likelihood estimate of the generator-level distribution when number of iterations goes to infinite
- In practice, **less than 10 iterations** are enough to achieve convergence

Part 2

Applications

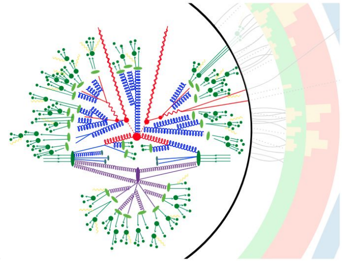
OmniFold dataset



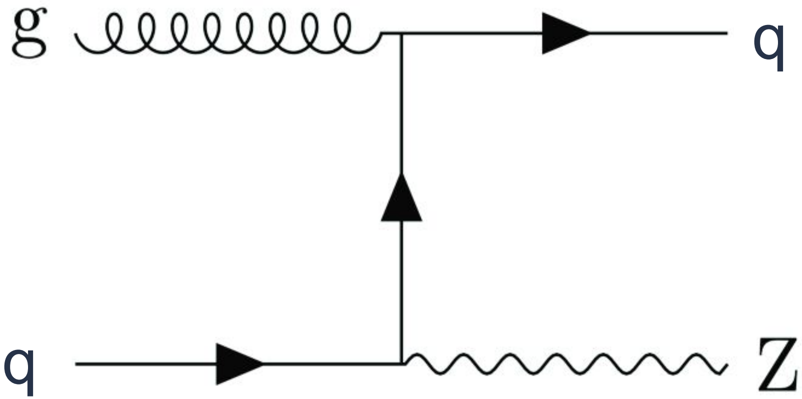


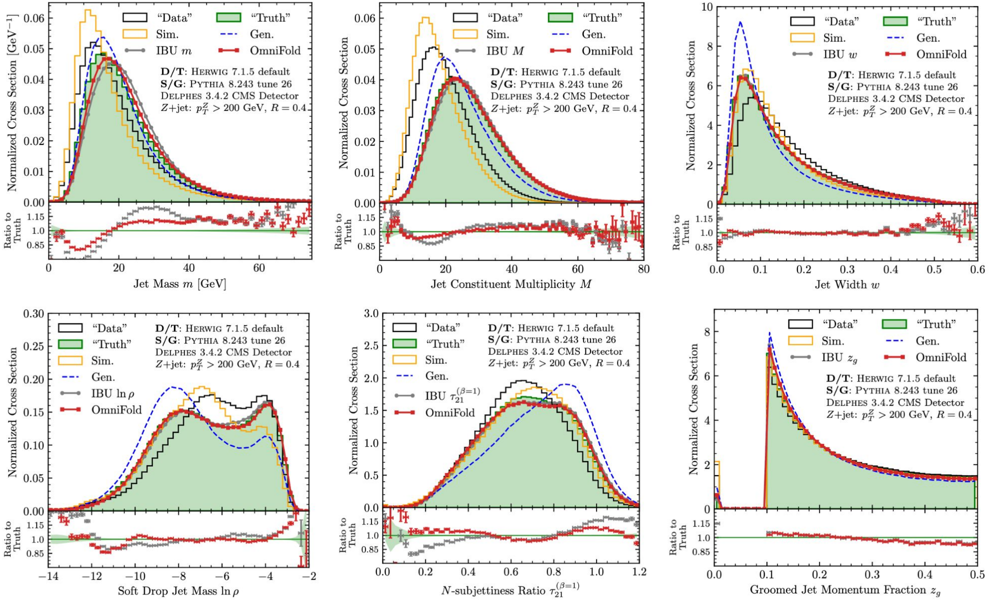
We are going to unfold **6 jet substructure observables** simultaneously using **OmniFold**

Only consider Z decaying to neutrinos:
mostly a single jet per event.



OmniFold dataset





Only consider Z decaying to neutrinos:
mostly a single jet per event.

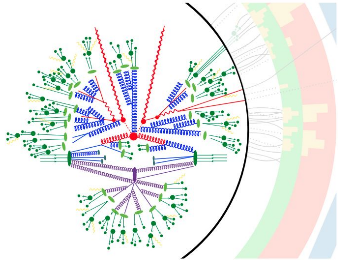
Phys. Rev. Lett. 124, 182001 (2020)

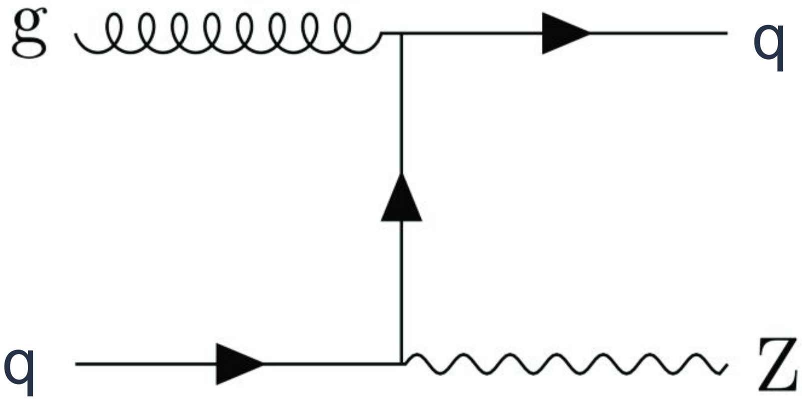
THANKS!

Any questions?

Backup

OmniFold dataset





Only consider Z decaying to neutrinos:
mostly a single jet per event.

- Observables:
  - Jet mass
  - Particle Multiplicity
  - $\tau_{21} = \tau_2 / \tau_1$ see Energ. Phys. 2012, 93 (2012).
  - Jet width (τ_1)
  - $\log \rho = 2 \log M_{SD}/p_T$
  - Momentum fraction z_g after using Soft Drop