

Unfolding Jet Substructure Observables with Machine Learning

Nicodemos Andreou

A Large Ion Collider Experiment (ALICE)

- One of the four major high energy experiment at the LHC, CERN
- Study Quark-Gluon Plasma (QGP)
- Lead – Lead (Pb-Pb) collisions
- Proton - Proton (pp) collisions

Why is QGP interesting?

- QGP produced during Pb-Pb collisions.
- State of the universe in the first instants after the Big Bang.
- Inner core of neutron stars?
- Other phenomena when QGP cools down.

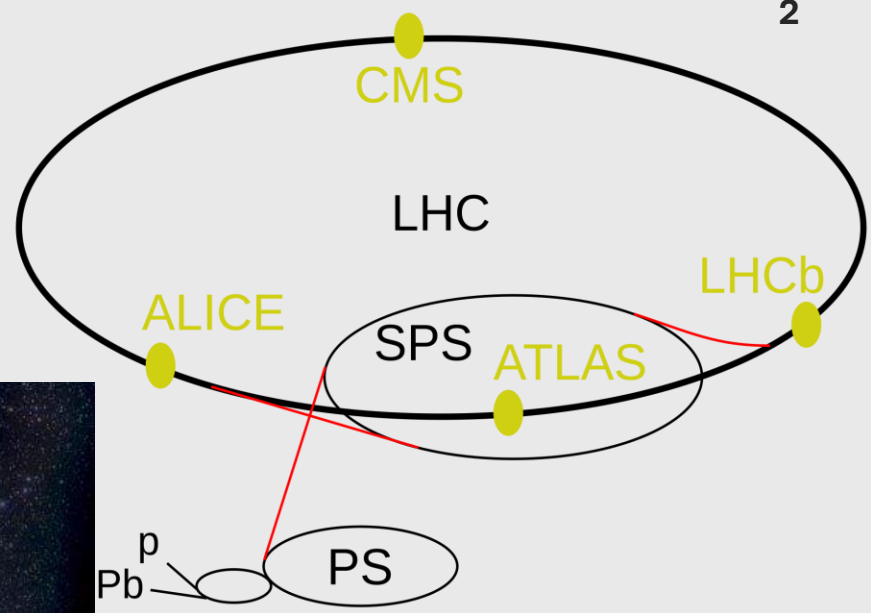


Figure 1: LHC ring diagram

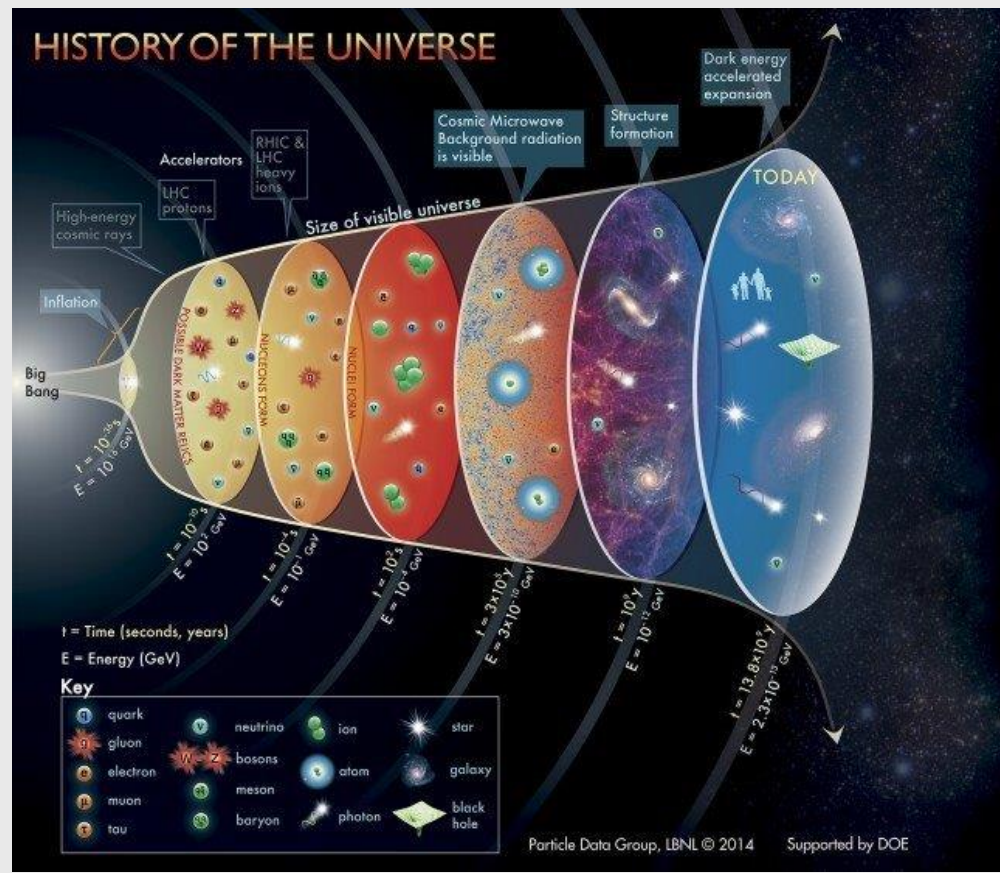


Figure 2: Evolution of the universe including LHC energies

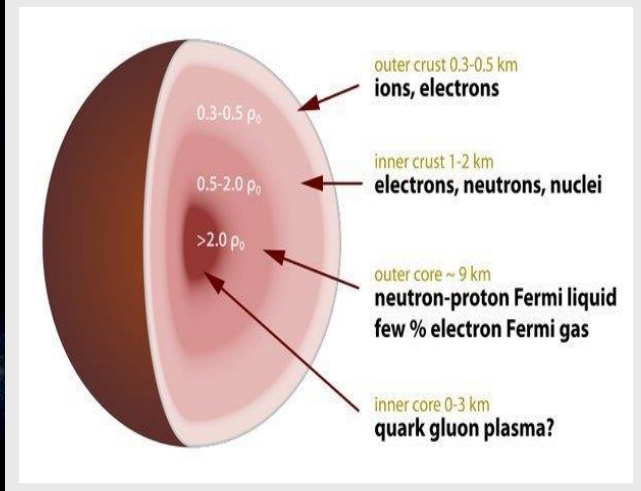


Figure 3: Neutron star

- Jets: Sprays of particles (pions (π), kaons (κ),...etc.) produced in high momentum transfers during high energy collisions (hard scattering).

- Jets have 10-100 times the typical energy of a particle
- Jets in pp = jets in vacuum due to the absence of dense medium.
- Jets in Pb-Pb collisions experience the entire evolution of QGP medium.
- Differences in similar jets from pp & Pb-Pb are due to QGP.
- Jets in Pb-Pb lose up to half of their energy due to the interaction with QGP.

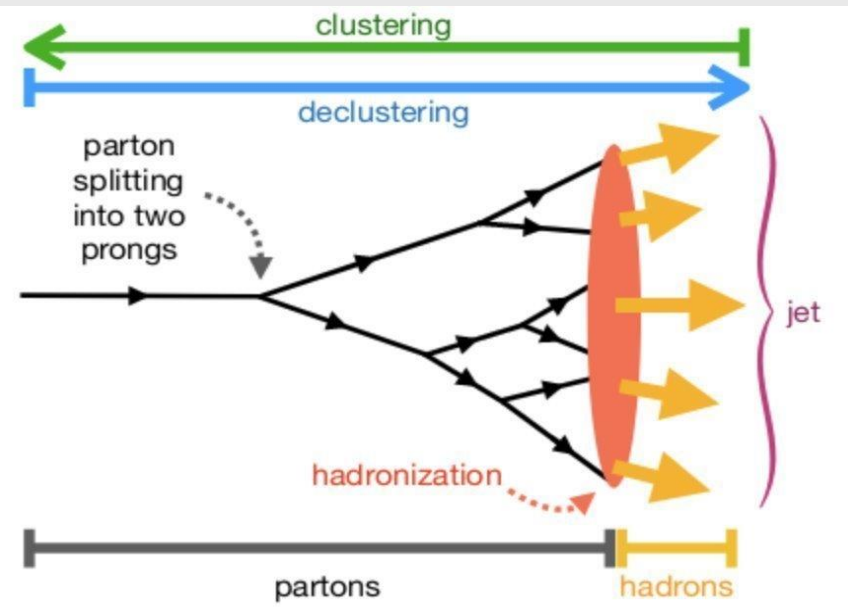
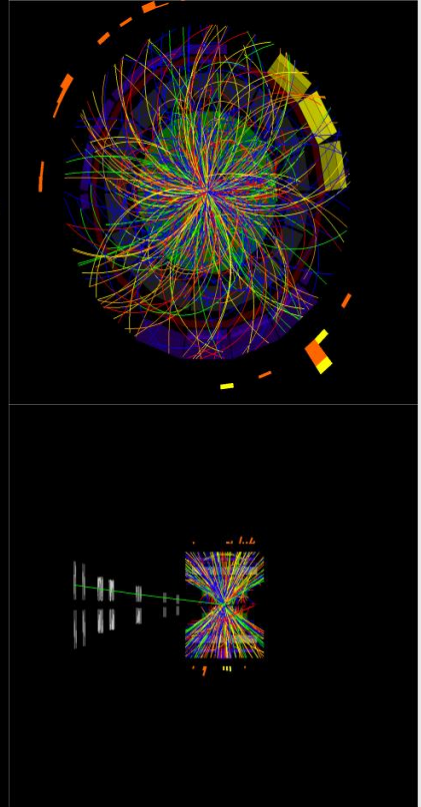
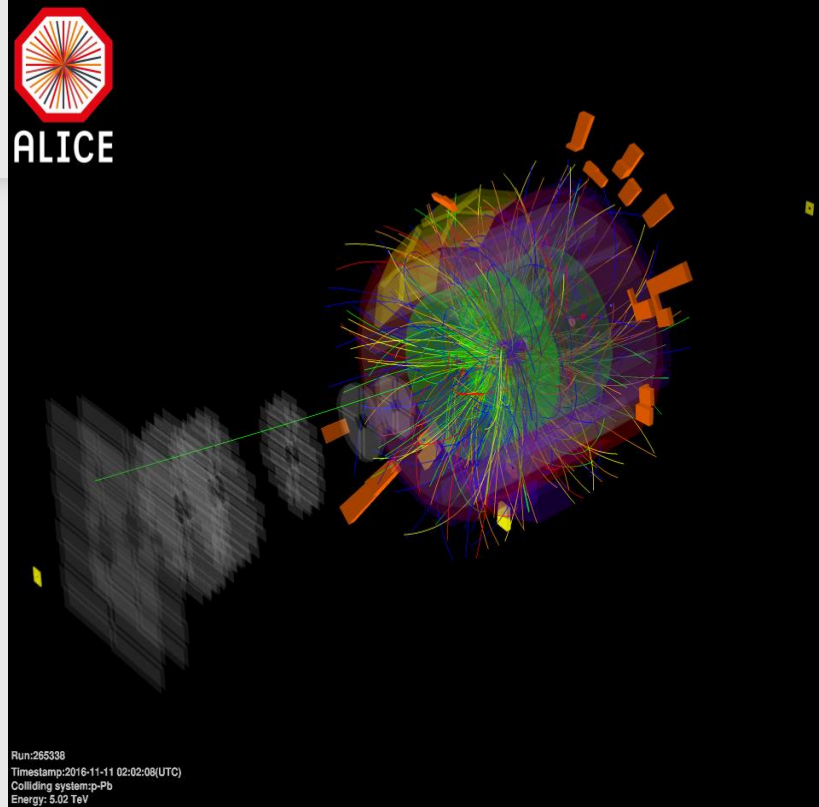
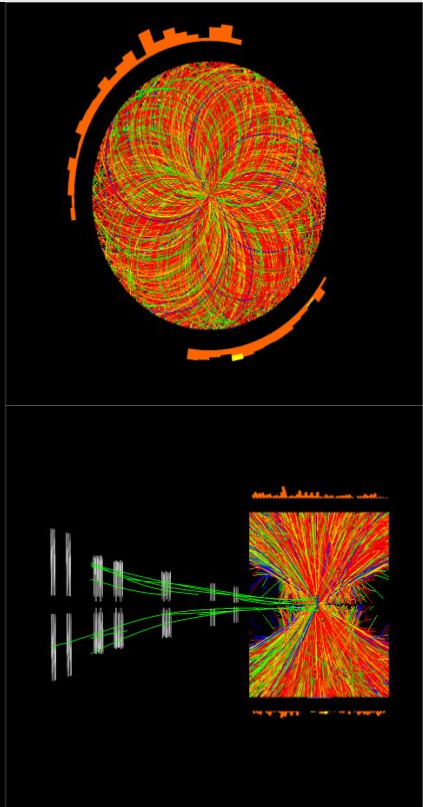
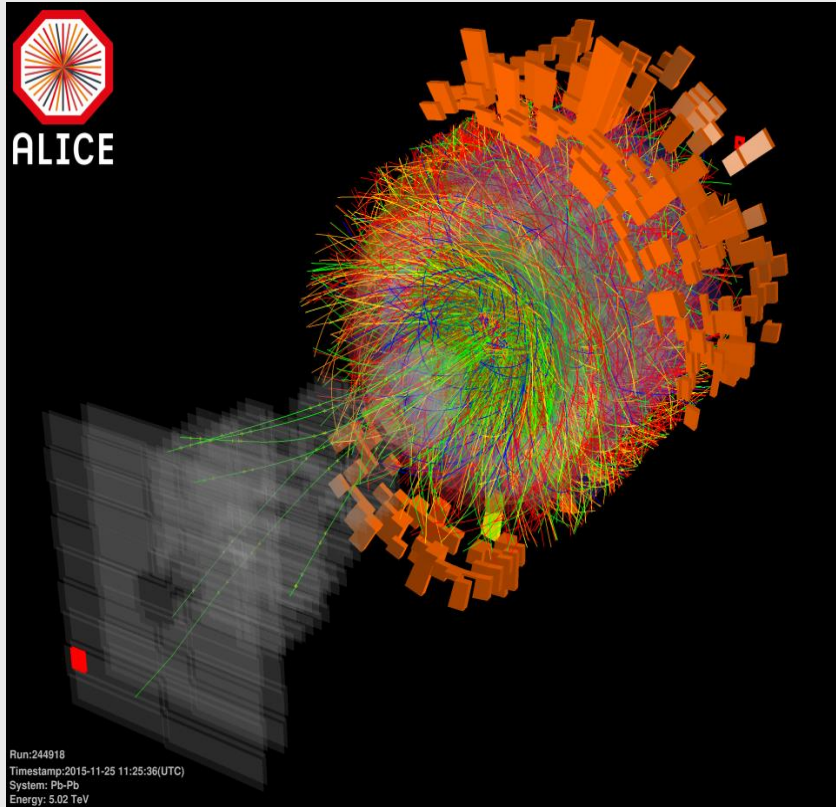


Figure 4: Jet diagram

Collision Background

Pb-Pb

pp



Observables

- p_T : Transverse Momentum
- k_T : Relative Transverse Momentum
- z_g : Groomed Momentum
- r_g : Groomed Radius
- nsd : Number of splittings
- η : Pseudorapidity

Groomed Substructure Observables:

- Grooming algorithms remove soft-wide angle radiation
- ↓
- Access the hard parton splittings
- ↓
- Isolate substructures well-controlled in pQCD
- ↓
- Constrain jet quenching effects

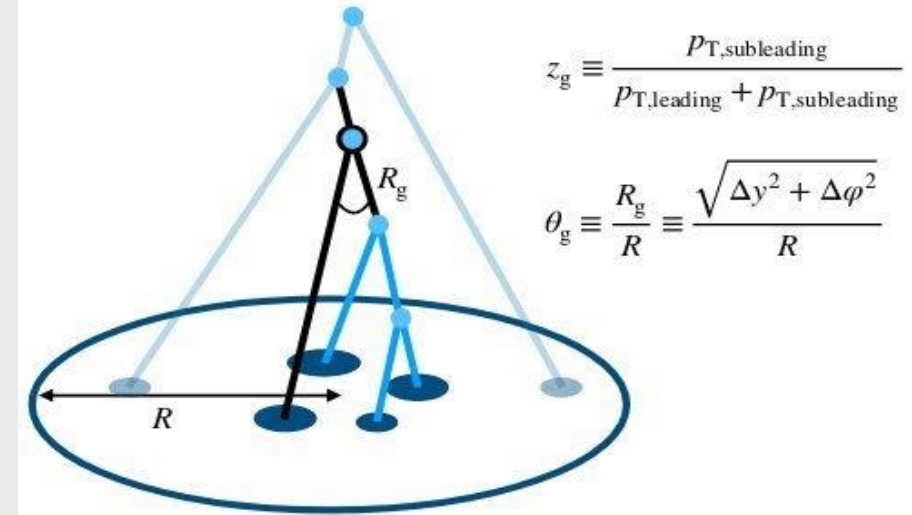
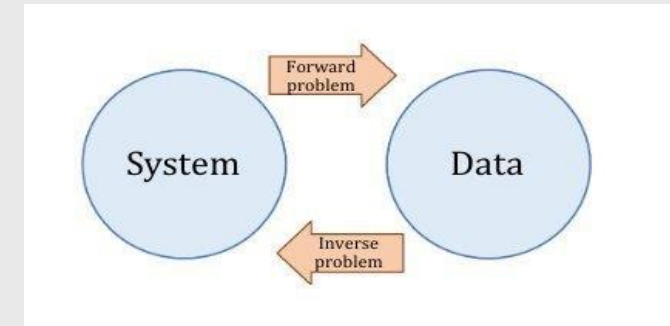


Figure 5: Angularly ordered jet constituents. Hardest splitting in black and groomed away splittings in light blue [1]

Unfolding – Inverse Problem

Measurement: $m = Rt$

- m : observed value
- t : true value
- R : instrumental response **R (matrix)**

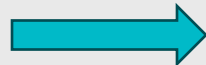


Sadri, M., Shariatipour, S. M., Hunt, A., & Ahmadinia, M. (2019). Effect of systematic and random flow measurement errors on history matching: a case study on oil and wet gas reservoirs. *Journal of Petroleum Exploration and Production Technology*, 9, 2853-2862.



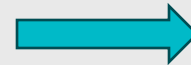
(i) True Data (t)

Measurement



(ii) Observed Data (m)

Unfolding



(iii) Recovered Data ($mR^{-1} \sim t$)

Figueiredo, M. A., & Nowak, R. D. (2003). An EM algorithm for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 12(8), 906-916.

NOTE:

- **ALL MEASUREMENTS/DETECTIONS INCLUDE INSTRUMENTAL EFFECTS**
- **TRUE DISTRIBUTIONS CAN ONLY BE ESTIMATED**

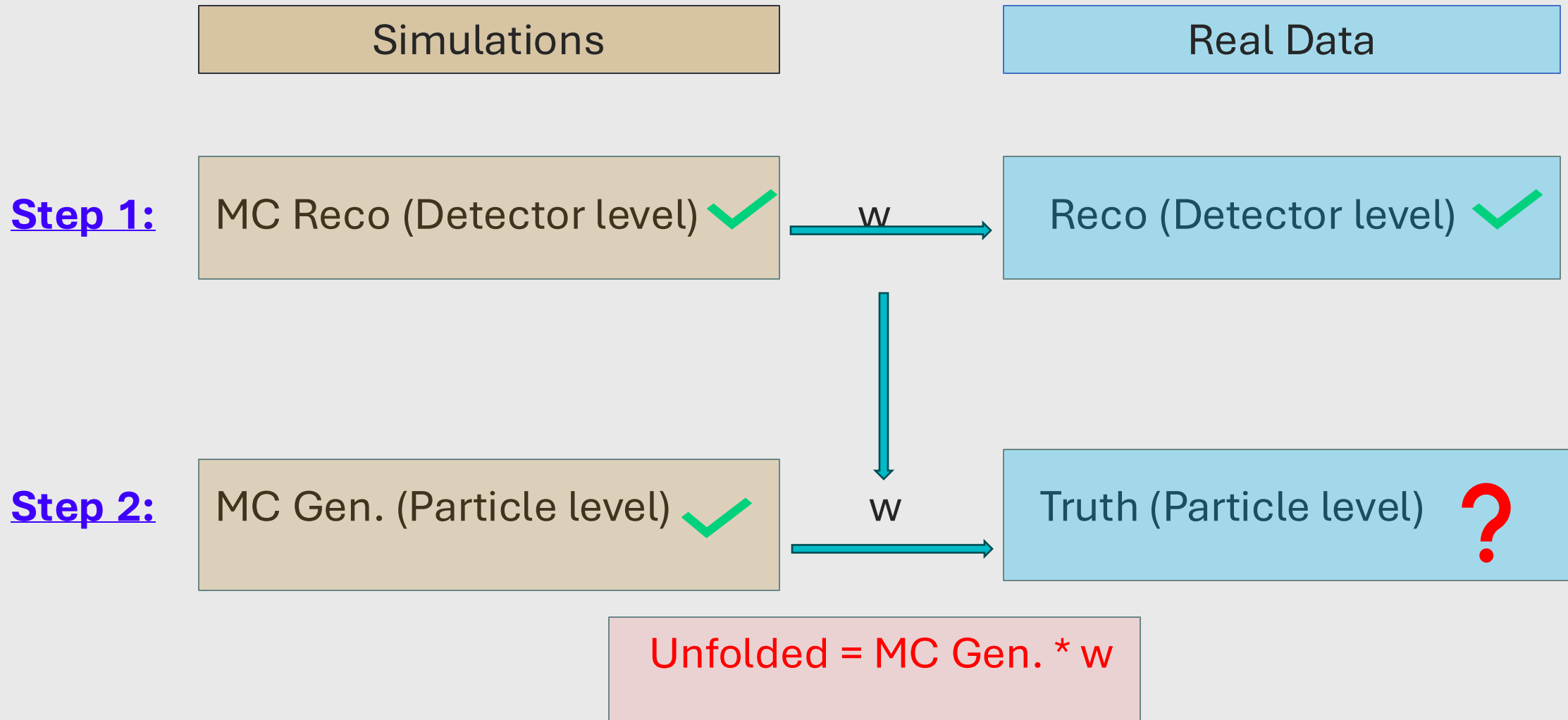
Omnifold (Multifold)

- Unbinned unfolding.
- Iterative reweighting event by event to match simulated data distributions to real data distributions.
- Neural Network (NN) classifiers to estimate likelihood ratios to update weights.
- Final weights converge via Maximum Likelihood Ratio.

Advantages:

- High-dimensional simultaneous unfolding
- More complete evaluation of detector response
- Explore multi- dimensional correlations among observables

Omnifold (Multifold)



Omnifold (Multifold) 4D (Simulated Data)

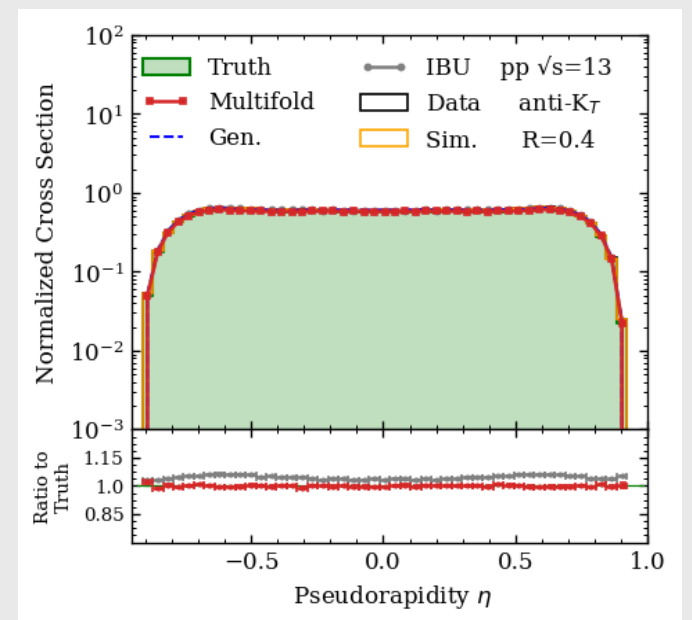
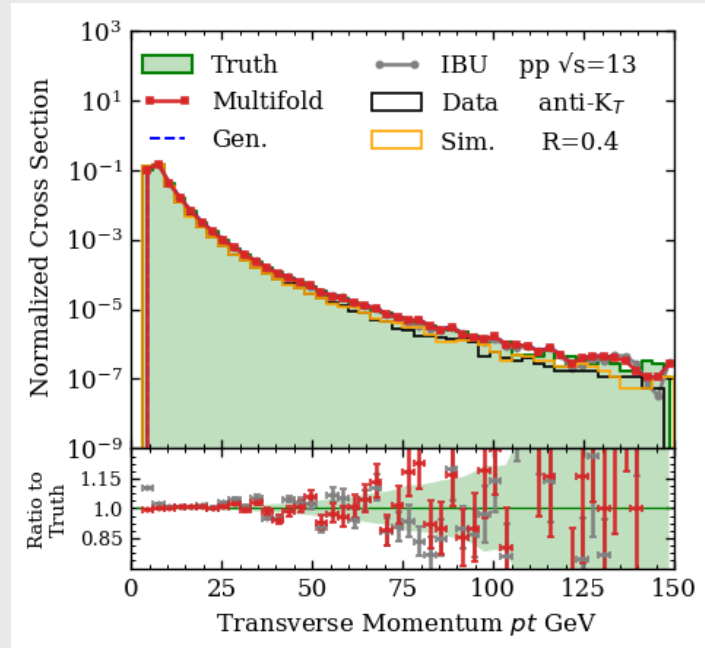
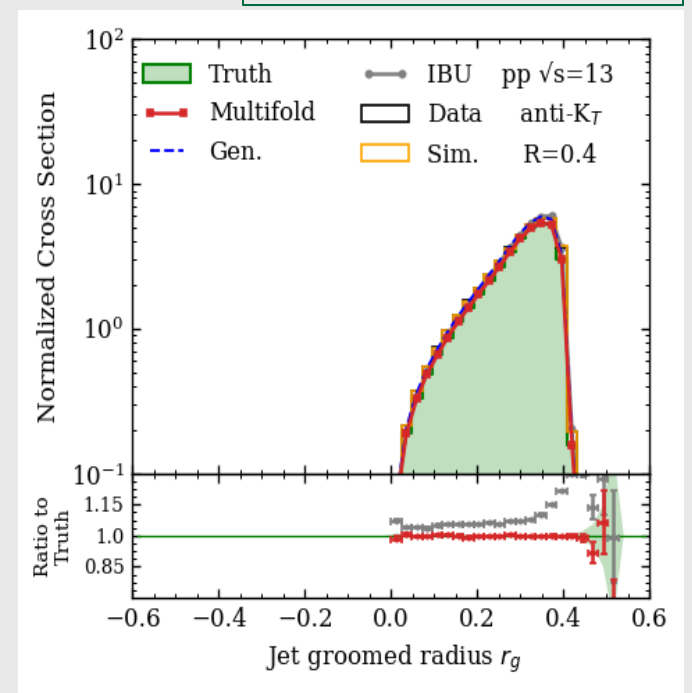
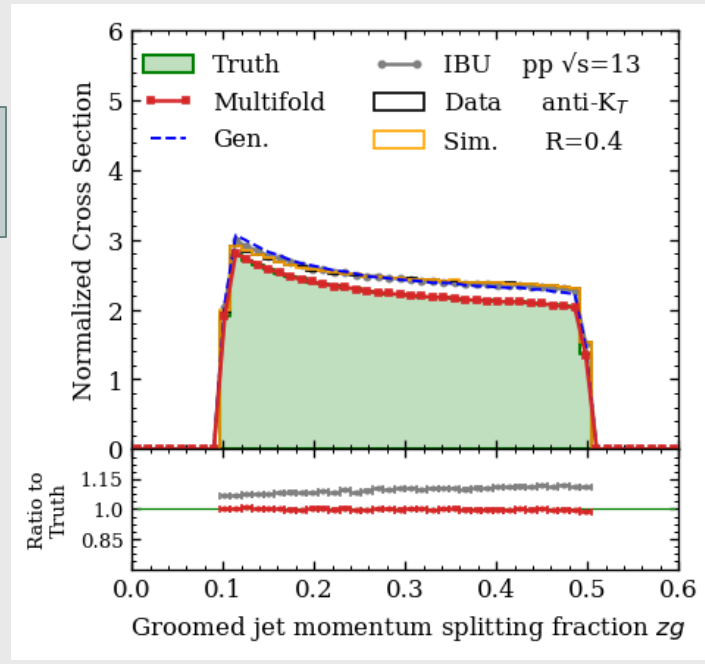
- Framework validation with MC (closure tests)
- Comparison with Iterative Bayesian Unfolding (IBU)

Data Format:
 AOD → UPROOT → NumPy Arrays

Datasets: MC PYTHIA Run 2 (2018)

IBU: Iterative Bayesian Unfolding
 Multifold: Omnifold in 4D
 Gen: MC Gen
 Truth: MC Gen (Truth)
 Sim: MC Reco
 Data: MC Reco (Reco)

Takeaway: Good performance over IBU in multidimensional spectrum compared to 'Truth'



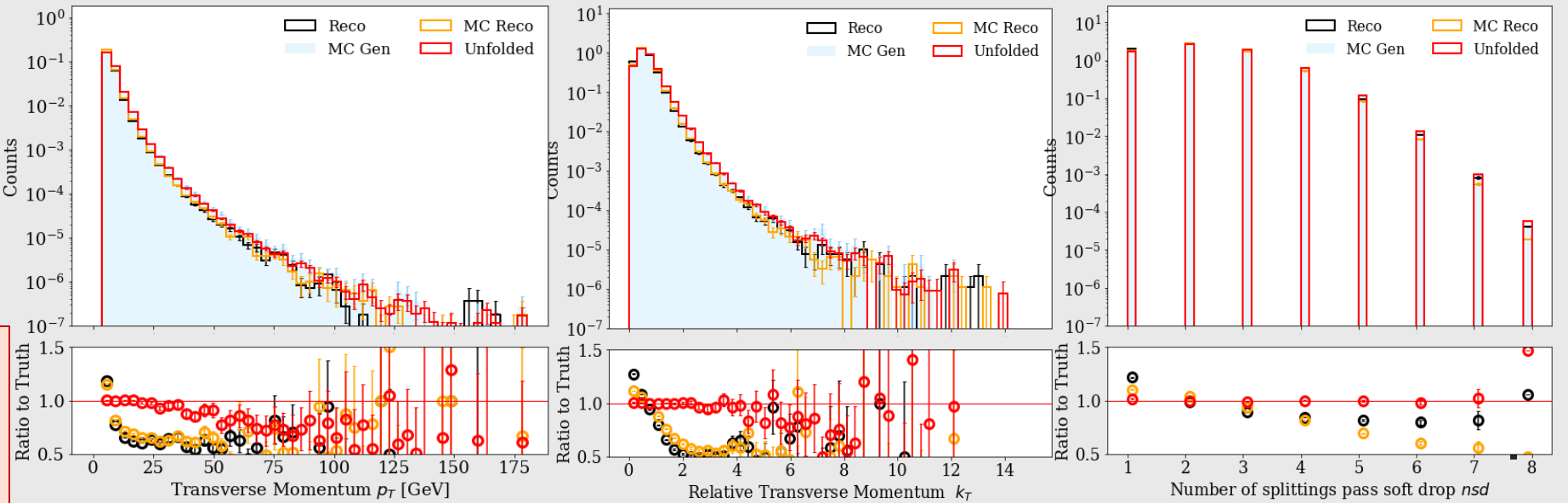
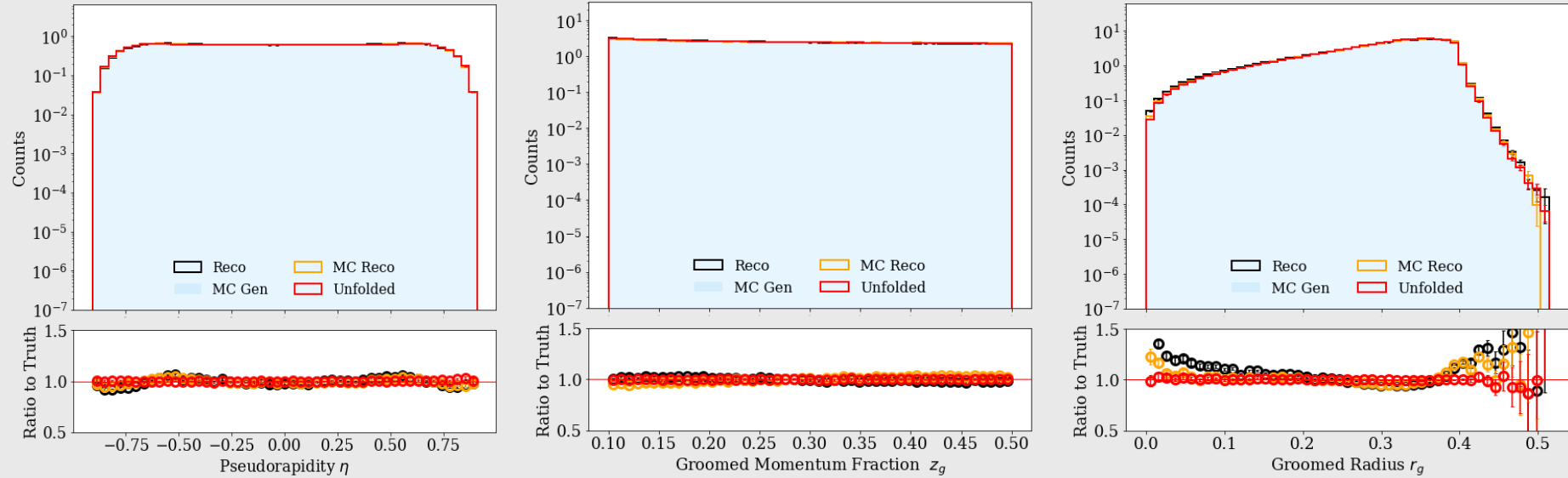
Omnifold 6D (Real Data)

Datasets:

- pp Run 2 (2018)
- MC pp (PYTHIA Run 2, 2018)
- $\sqrt{s} = 13$ TeV
- 3×10^6 jets

- Real data unfolding
- Model trained on MC Reco & Reco (Step 1)
- Weights from Step 1, applied to MC Gen (Step 2)

Work in progress! 10



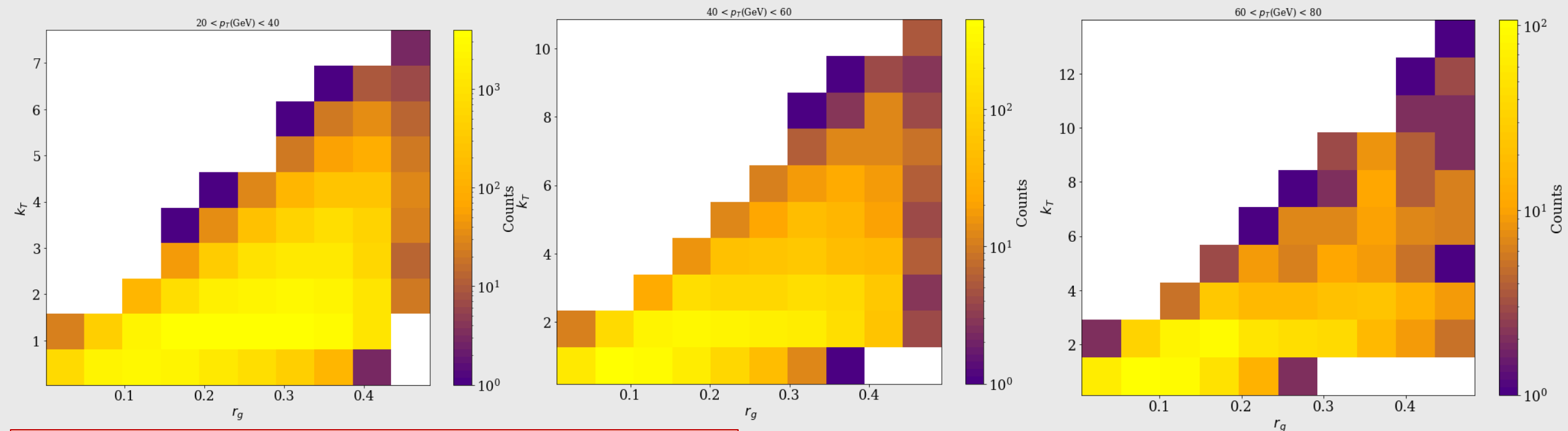
Takeaway:

- Good performance in 6D
- Nice agreement with 'MC Gen'
- Unstable in low-stats regions

Datasets:

- pp Run 2 (2018)
- MC pp (PYTHIA Run 2, 2018)
- $\sqrt{s} = 13$ TeV
- 3×10^6 jets

Work in progress!



Takeaway:

- Exploiting multidimensional correlations of unfolded distributions
- Any correlations among any observables can be studied

Statistical:

- Bootstrapping: weights from Poisson dist. with mean=1 applied to MC & Data (training data) ✓
- Ensembles: Averaging NN weights for more stability over random fluctuations of individual models ✓

Systematic:

- Different MC sources ✗
- Alternative weights $\sim 1\sigma$ from 'best model' ✓

Validation:

- Pseudo –data model with MC (pre-assigned 'truth' dataset) to test the model (closure-test) ✓
- Validation Test with 20% of data ✓

What's Next?

- Unfold entire Run 2 pp dataset
- Unfold Run 3 pp data

How about Pb – Pb?

- Complex background is a real challenge.
- Potentially if realistic background simulations could be produced.

Summary:

- Multidimensional unbinned unfolding with Omnifold
- Framework validation & comparison with IBU
- Run 2 pp simulated data
- Run 2 pp real data unfolded distributions

Conclusion:

- Omnifold was successfully used and validated using Run 2 data
- Powerful tool for unfolding multiple observables simultaneously and studying their correlations