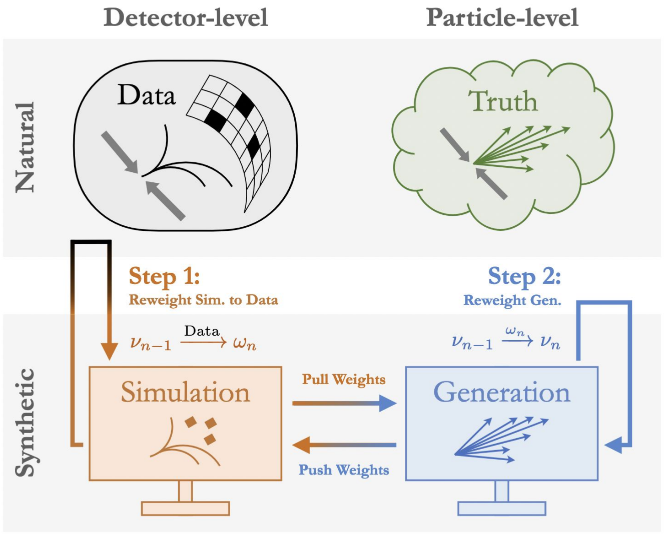
Multi-differential Jet Substructure Measurement in High Q^2 DIS Events with HERA-II Data

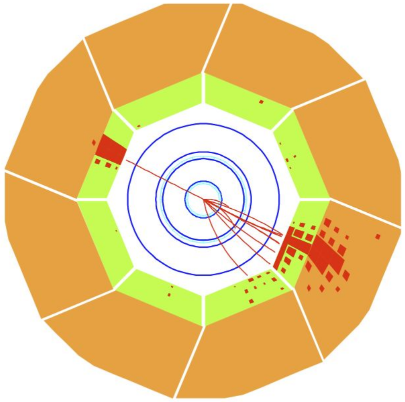
Vinicius Mikuni

Review

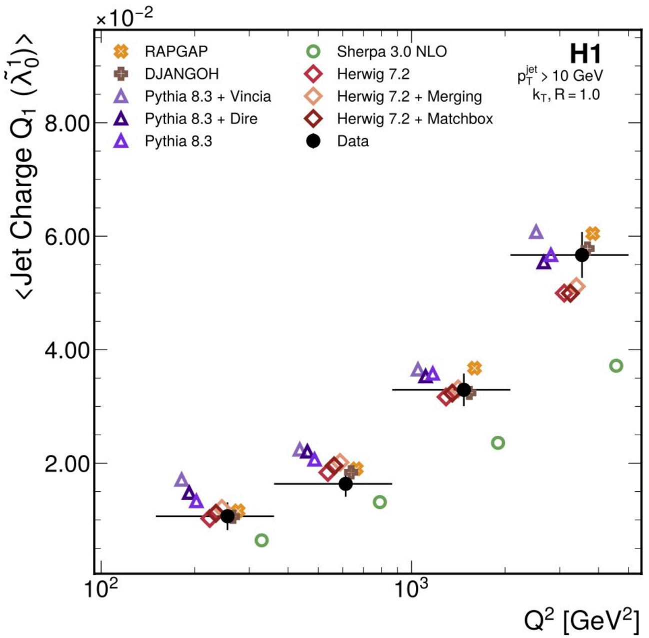
1: Unfolding methodology



2: Definition of measure observables

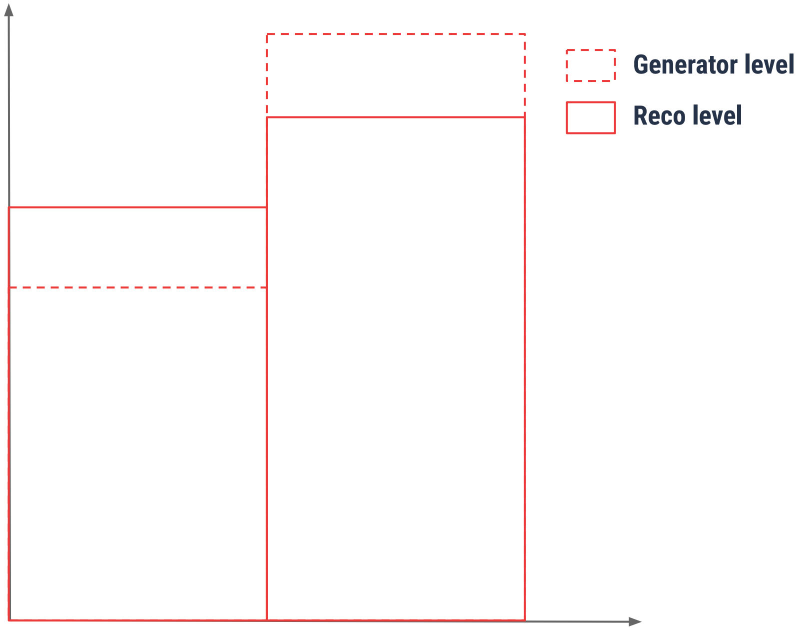


3: Multi-differential cross section results



Unfolding

- We only have access to observables at **reconstruction level**, i.e after detector effects
- When comparing different theories, we want to compare observables before detector interaction (**generator level**):
  - Don't require theorists to have expert detector knowledge to compare their predictions
  - Easier to maintain and incorporate new calibration routines for detector simulation
- What I'm **not** talking about today:
  - IBU/D'Agostini method
  - SVD
  - Matrix inversion
  - Other methods for unfolding using histograms



Unfolding

Traditional methods for unfolding are performed using **histograms**

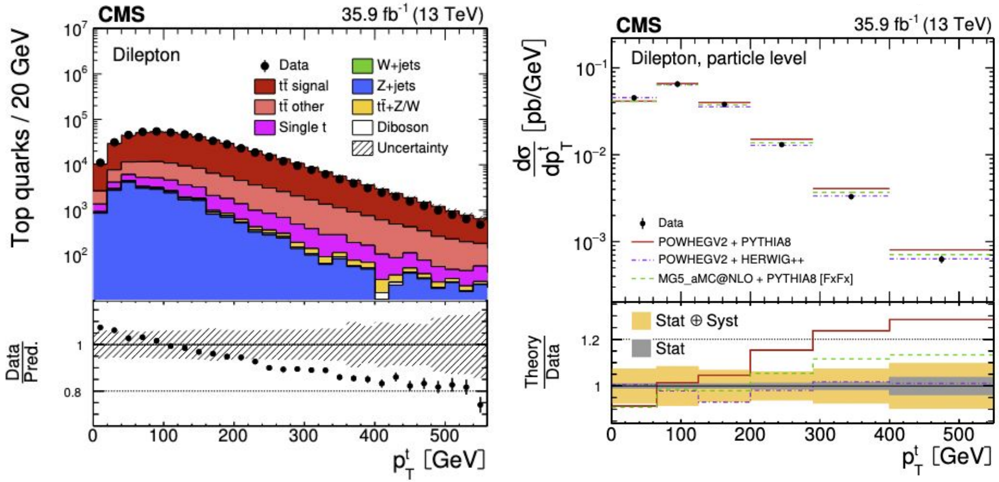
- Well understood statistical properties
- Clear convergence criteria

Limitations:

- Histograms need to be defined before unfolding.
  - If a different binning is required, the full unfolding routine needs to be redone
- Often able to address only 1 observable at a time
  - Multi-dimensional histograms are harder to deal with: **curse of dimensionality**

Reco level

Generator level

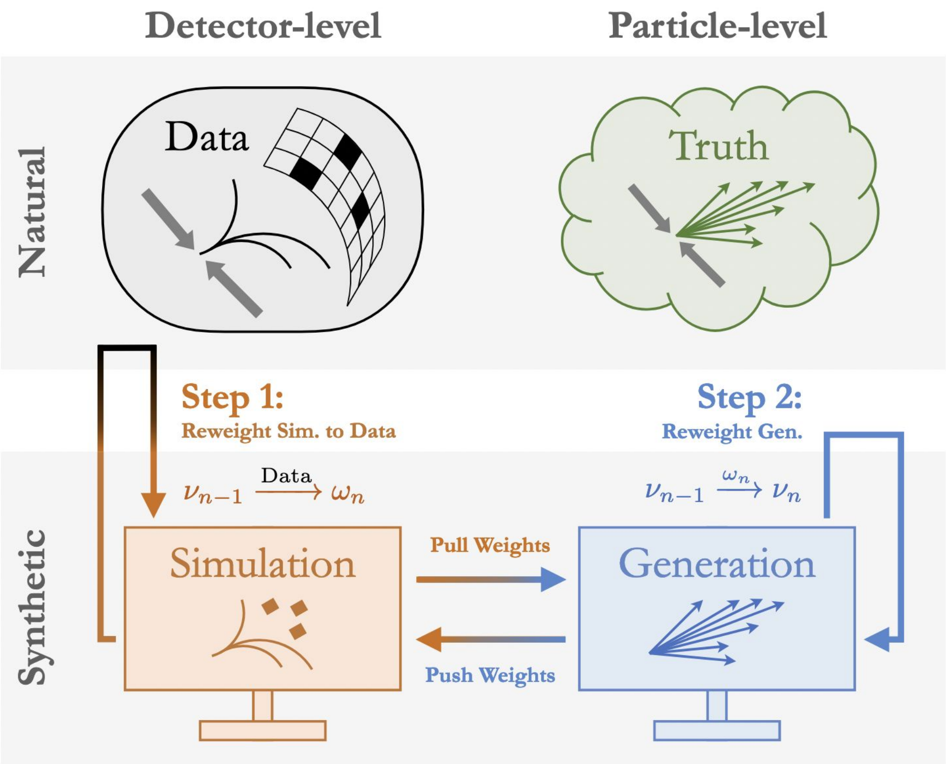


J. High Energ. Phys. **2019,** 149 (2019).

Omnifold*

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

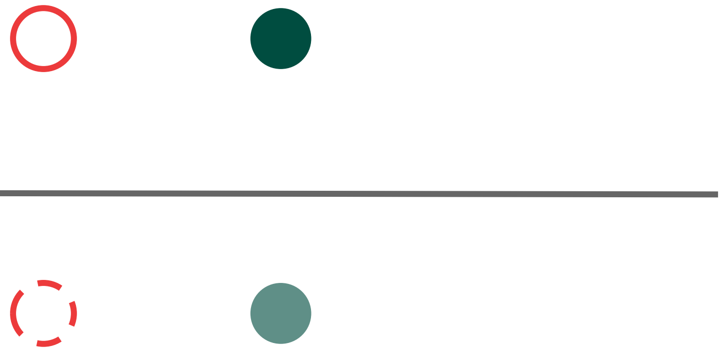
- **reco MC to Data** reweighting
- **Previous** and **new Gen** reweighting

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

Omnifold

Reco level

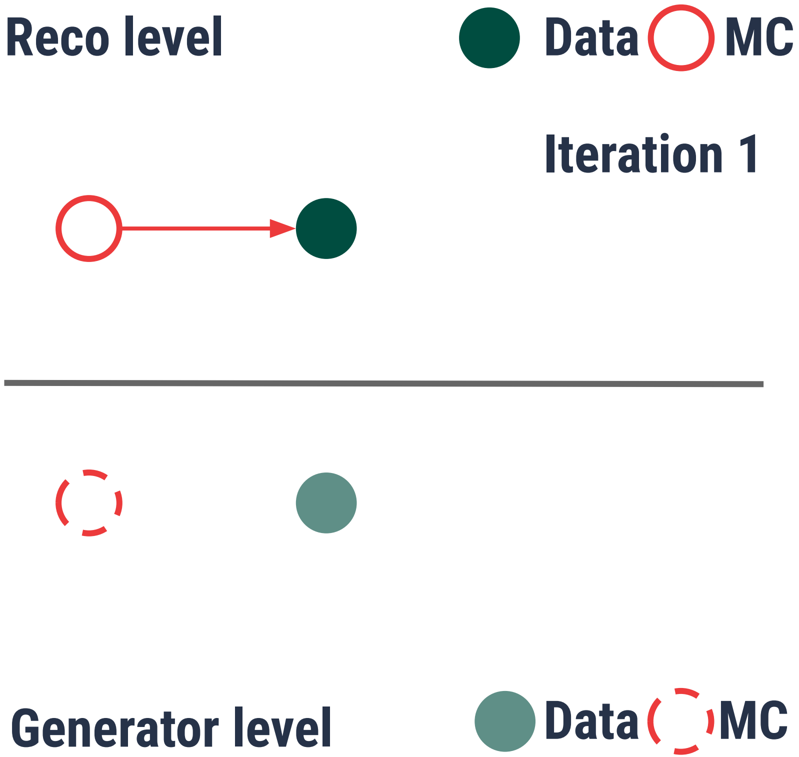
Data MC



Generator level

Data MC

Omnifold



Step 1:

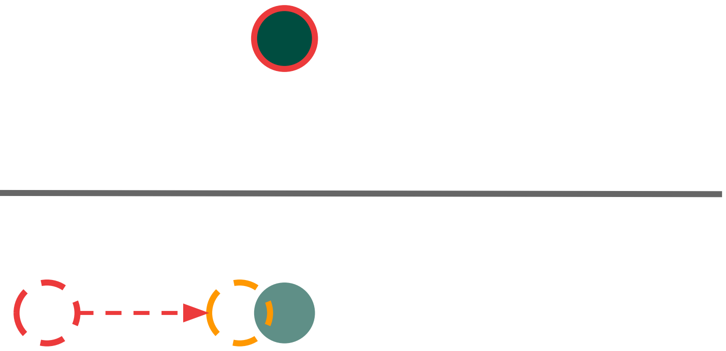
- Train a classifier to separate **data** from **MC** events
- Reweight **reco level MC** with weights:

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

Omnifold

Reco level ● Data ○ MC

Iteration 1



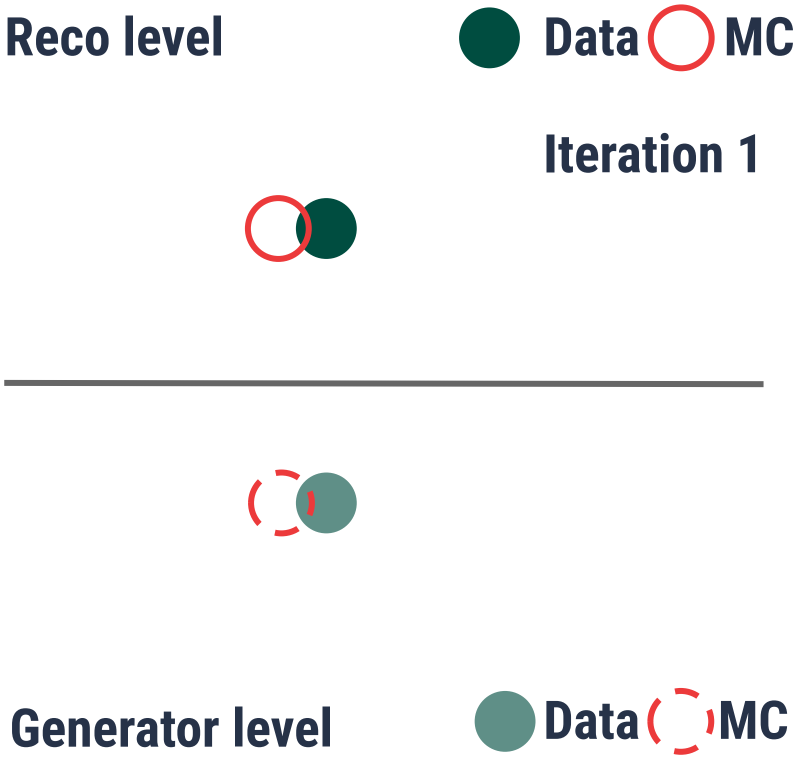
Generator level ● Data ◌ MC ◌ MC reweighted

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**

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

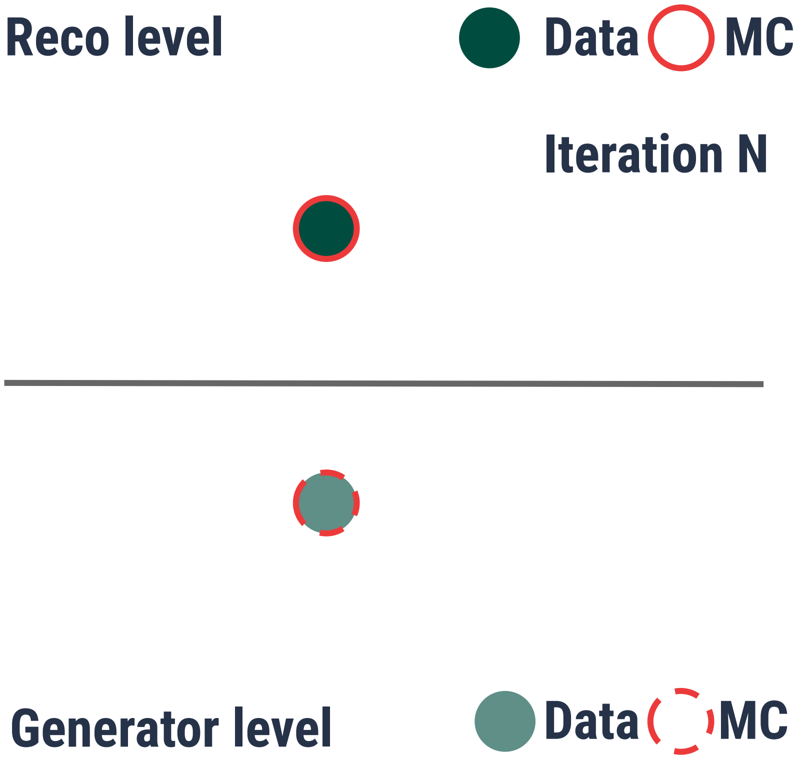
Omnifold



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

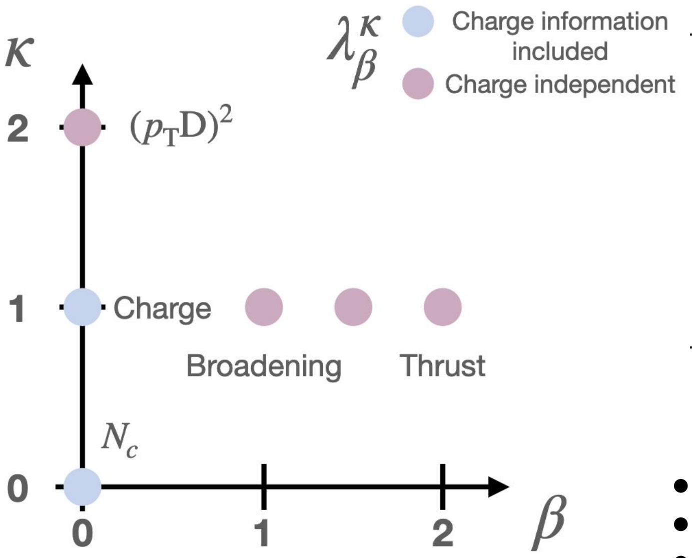
Part 2

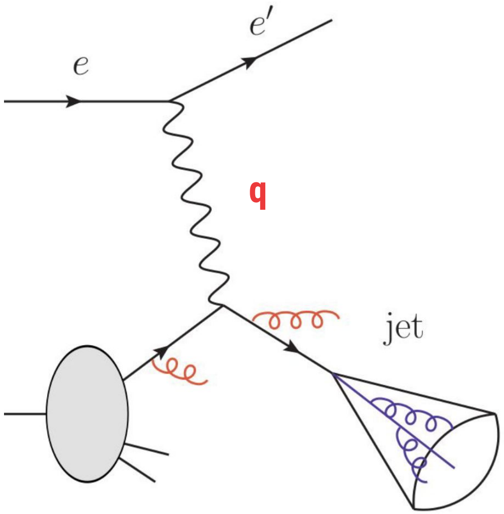
Physics case

Jet angularities

Use jet observables to study different properties of QCD physics:

- Infrared and collinear (IRC) safe λ^1_a, a = [0,0.5,1] and unsafe $\mathbf{p_TD}$ angularities
- Charge dependent observables: $\mathbf{Q_j}$ and $\mathbf{N_c}$
- Study the evolution of the observables with energy scale $\mathbf{Q^2 = -q^2}$



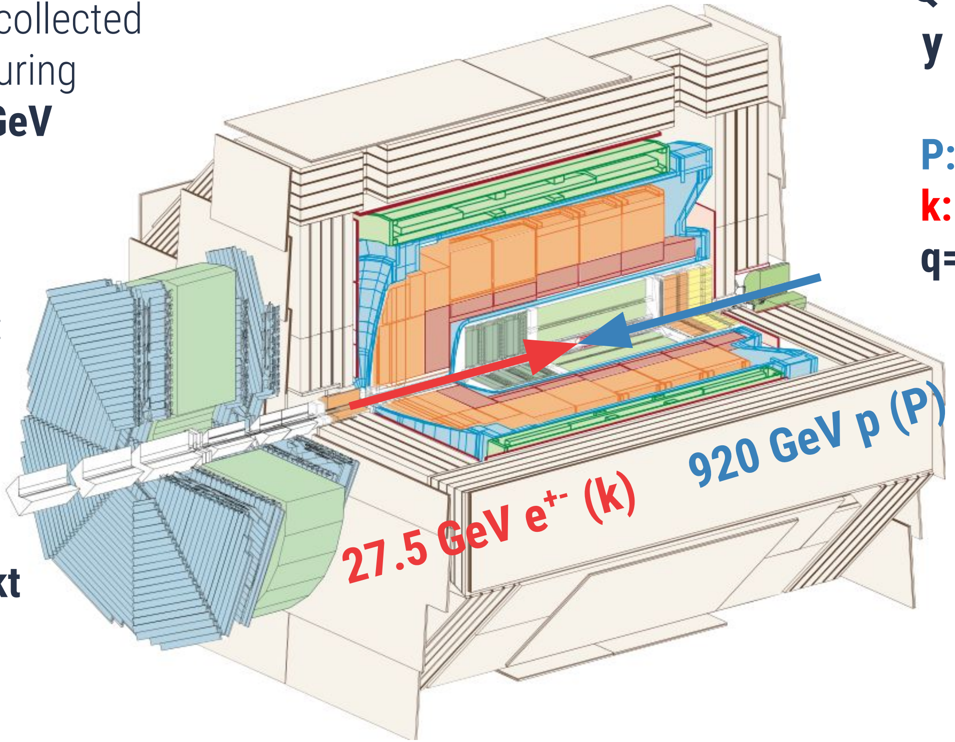


- z_i: longitudinal momentum fraction
- q_i: charge
- R_i distance from jet axis in (eta,phi)

$$\lambda^\kappa_\beta = \sum_{i\in\text{jet}} z_i^\kappa \left(\frac{R_i}{R_0}\right)^\beta \qquad \tilde{\lambda}^\kappa_0 = Q_\kappa = \sum_{i\in\text{jet}} q_i \times z_i^\kappa .$$

Experimental setup

Using **228 pb^{-1}** of data collected by the **H1 Experiment** during **2006** and **2007** at **318 GeV center-of-mass energy**



$Q^2 = - q^2$

$y = Pq / pk$

P: incoming proton 4-vector
k: incoming electron 4-vector
q=k-k' : 4-momentum transfer

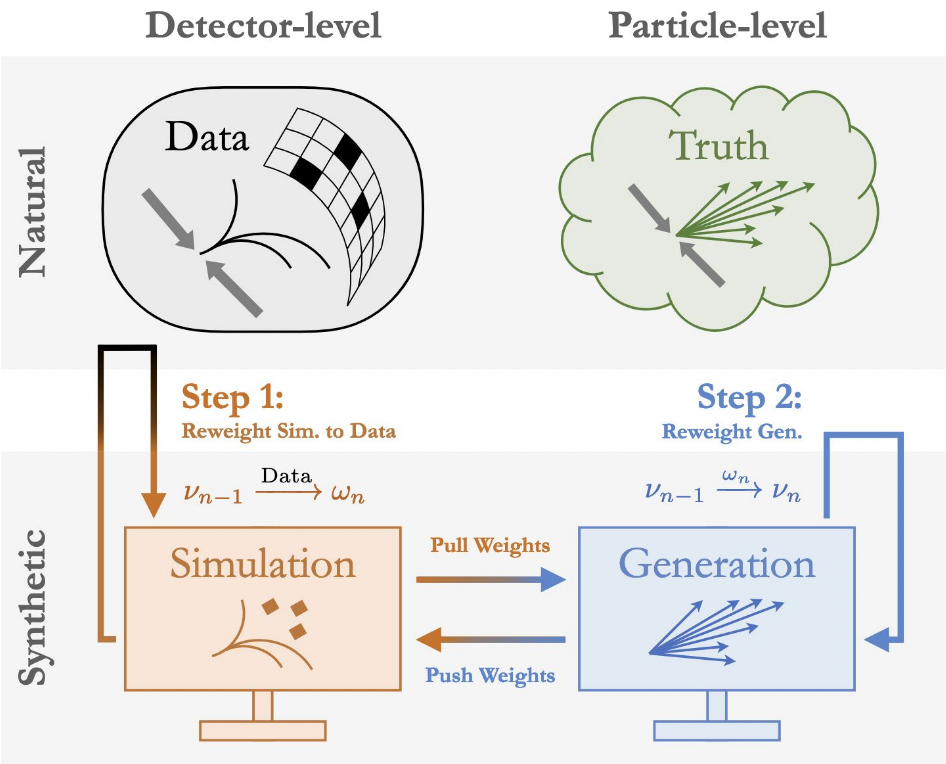
Phase space definition:

- $0.2 < y < 0.7$
- $Q^2 > 150$ GeV^2
- Jet $p_T > 10$ GeV
- $-1 < \eta_{lab} < 2.5$

Jets are clustered with **kt** algorithm with **R=1.0**

Reconstructed hadrons using combined detector information: **energy flow algorithm**

Omnifold*



2 step iterative approach

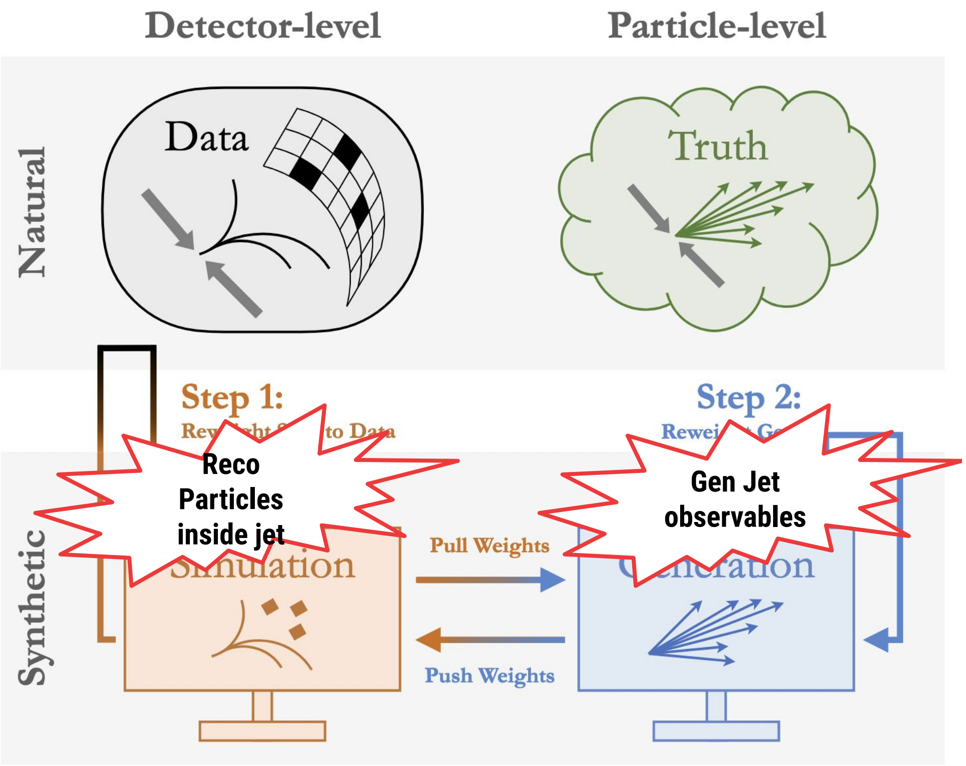
- Simulated events after detector interaction are reweighted to match the data
- Create a "new simulation" by transforming weights to a proper function of the generated events

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

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

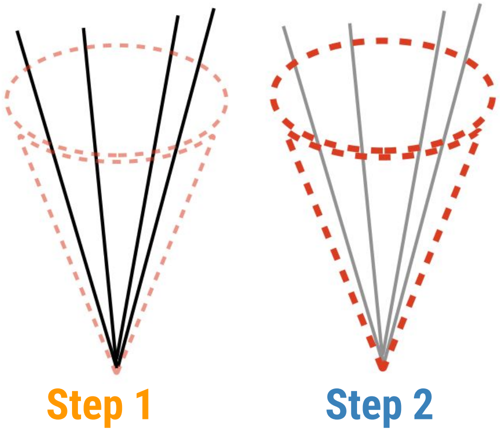
* Andreassen et al. PRL 124, 182001 (2020)

Omnifold



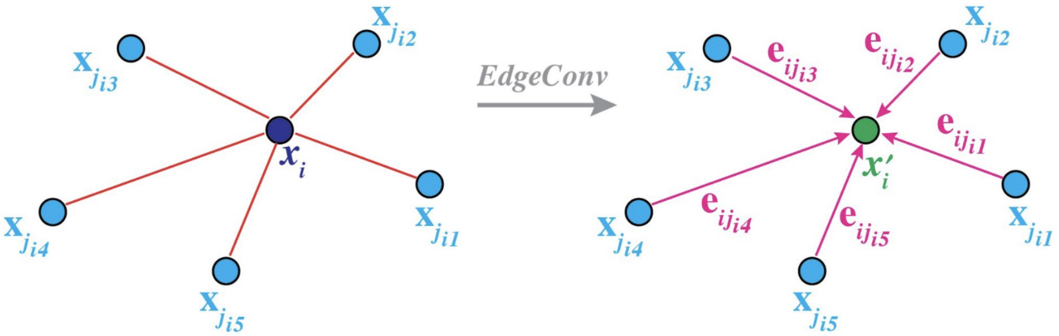
Different input levels for each step

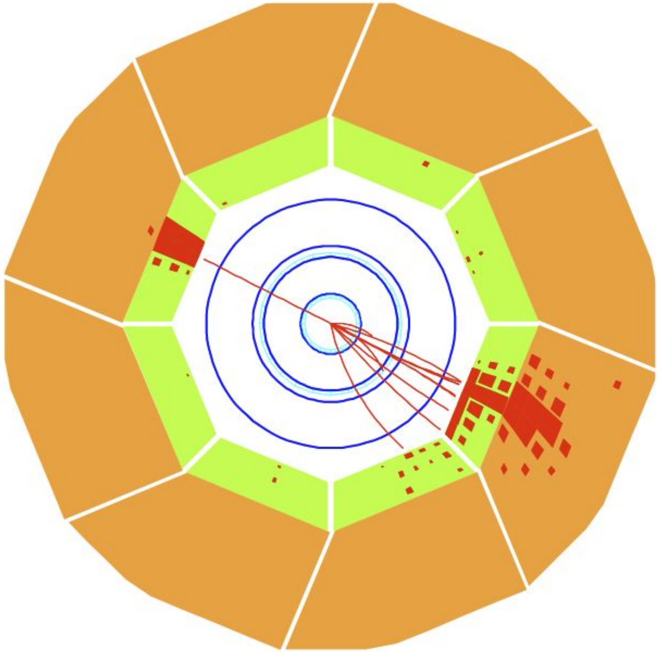
- **Step 1** particles are used as inputs
- **Step 2** uses the set of observables planned to unfold



Extracting particle information

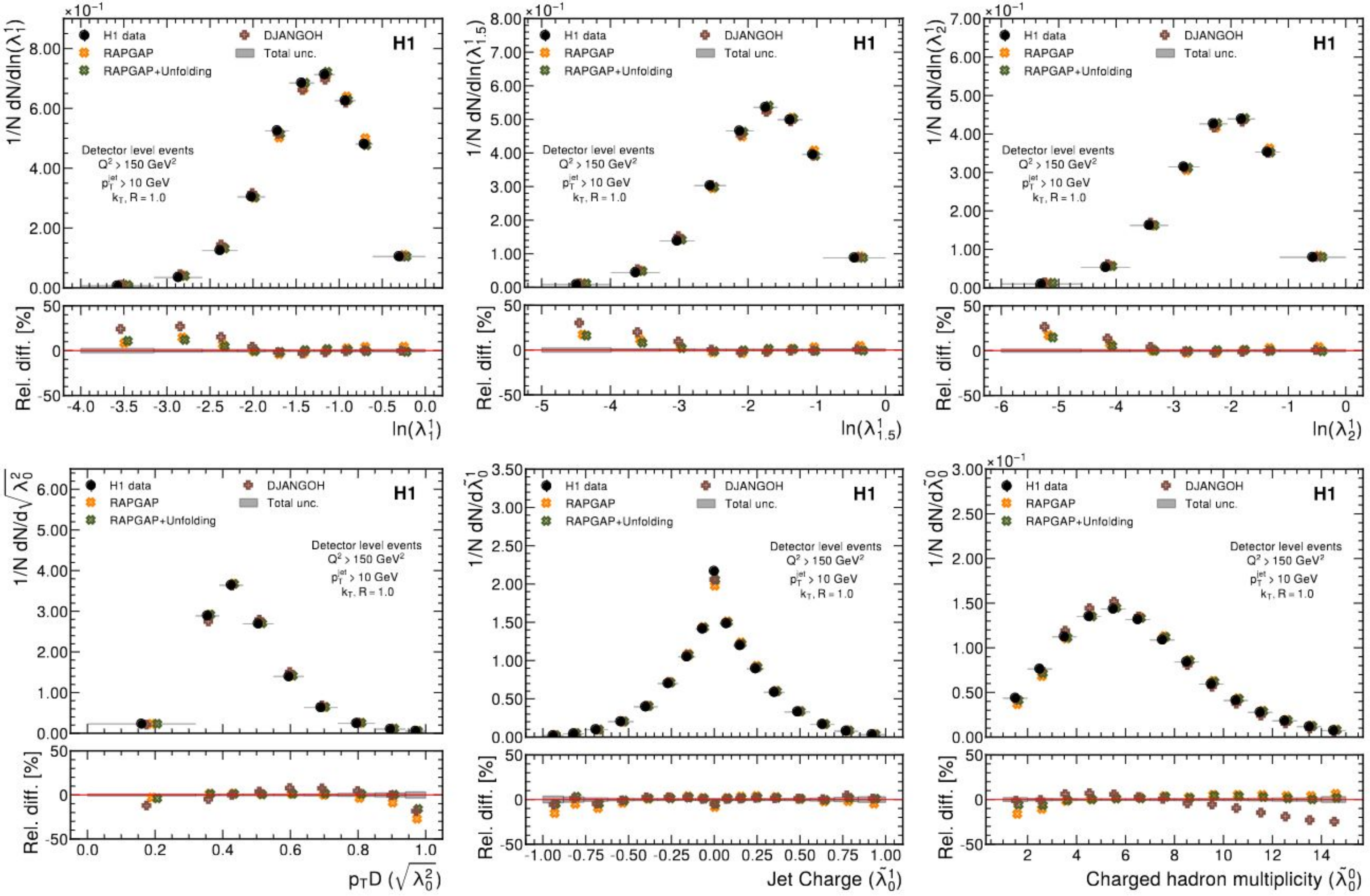
- Particle information is extracted using a **Point cloud transformer*** model
- Model takes **kinematic properties** of particles and use the distance between particles in η-φ to learn the relationship between particles
- Built in symmetries: **permutation invariance**
- Consider up to **30** particles per jet





* V. Mikuni and F. Canelli 2021 *Mach. Learn.: Sci. Technol.* **2** 035027

All distributions are **simultaneously** unfolded.



Outputs of the unfolding methodology are **weights** that are applied to the simulation

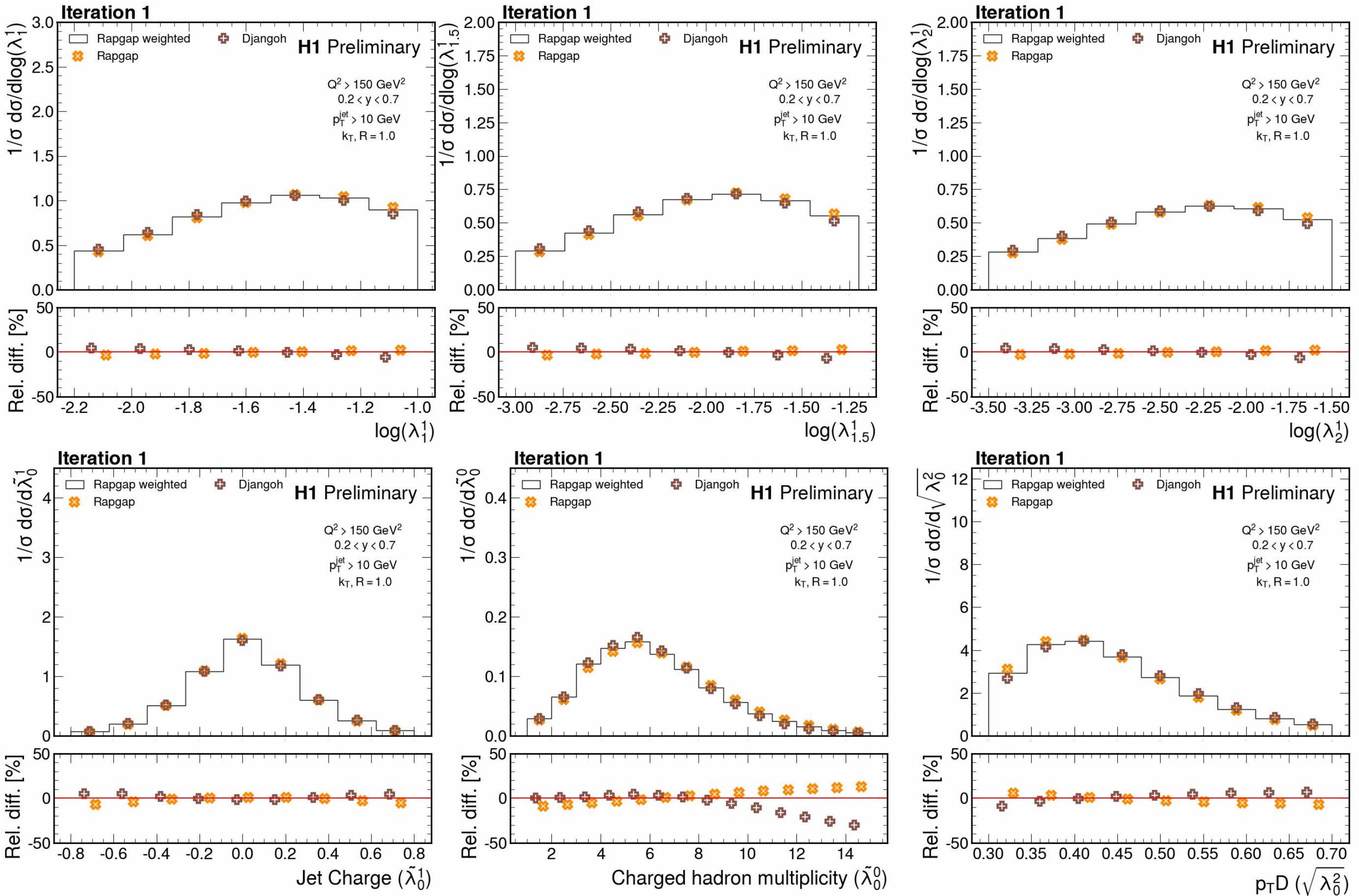
- **Green markers** represent the unfolded results **at reco level**
- Agreement with data **improves** compared to **initial Rapgap simulation**

Part 3

Unfolded results

Closure test

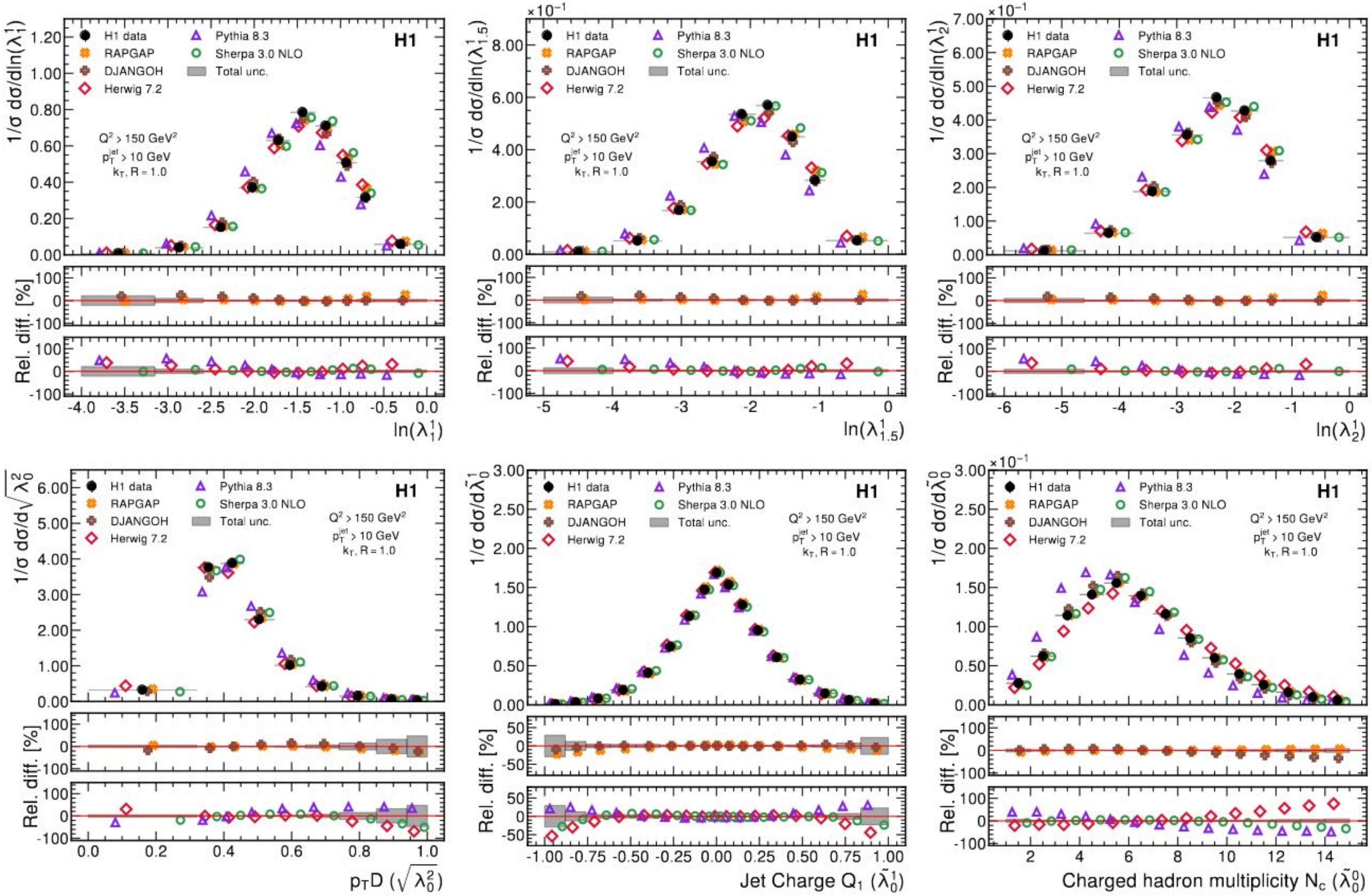
All distributions are unfolded **simultaneously without binning and without jet substructure information used at reco level!**



Verify the model **consistency**: start from the **Rapgap** simulation and unfold the response based on the **Djangoh** simulation

Total of **6 iterations** used to derive the main results

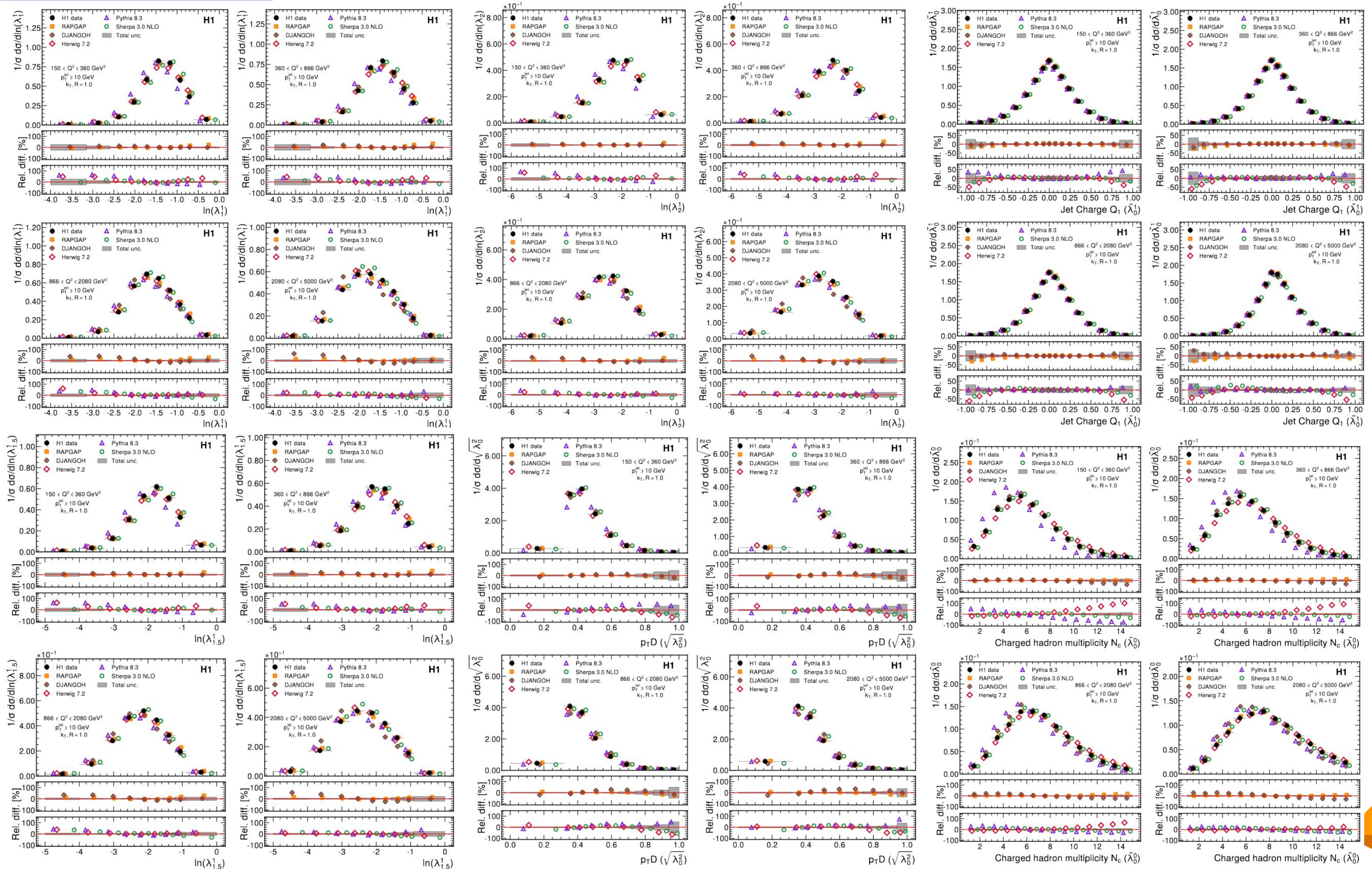
Inclusive



Dedicated DIS generators do a good job **everywhere**, especially **Rapgap**

Herwig, Pythia, and (yet **unreleased update** to) **Sherpa** do a decent job for most distributions

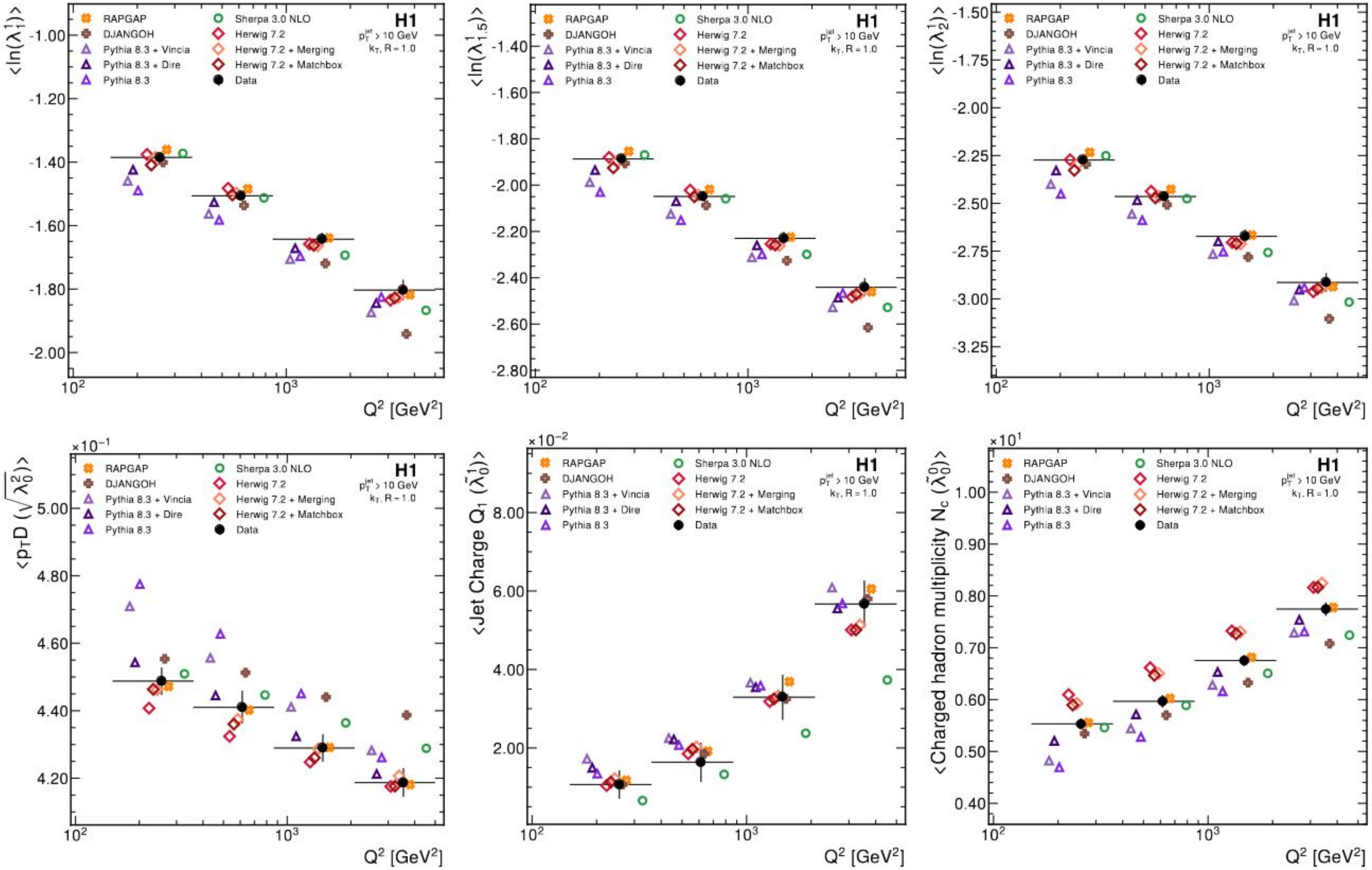
Multi-differential



Q^2 distribution is simultaneously unfolded, displaying the energy scale dependence of the observables, resulting in more than **30 unfolded distributions provided**

Multi-differential

Mean value of all distributions also unfolded for free

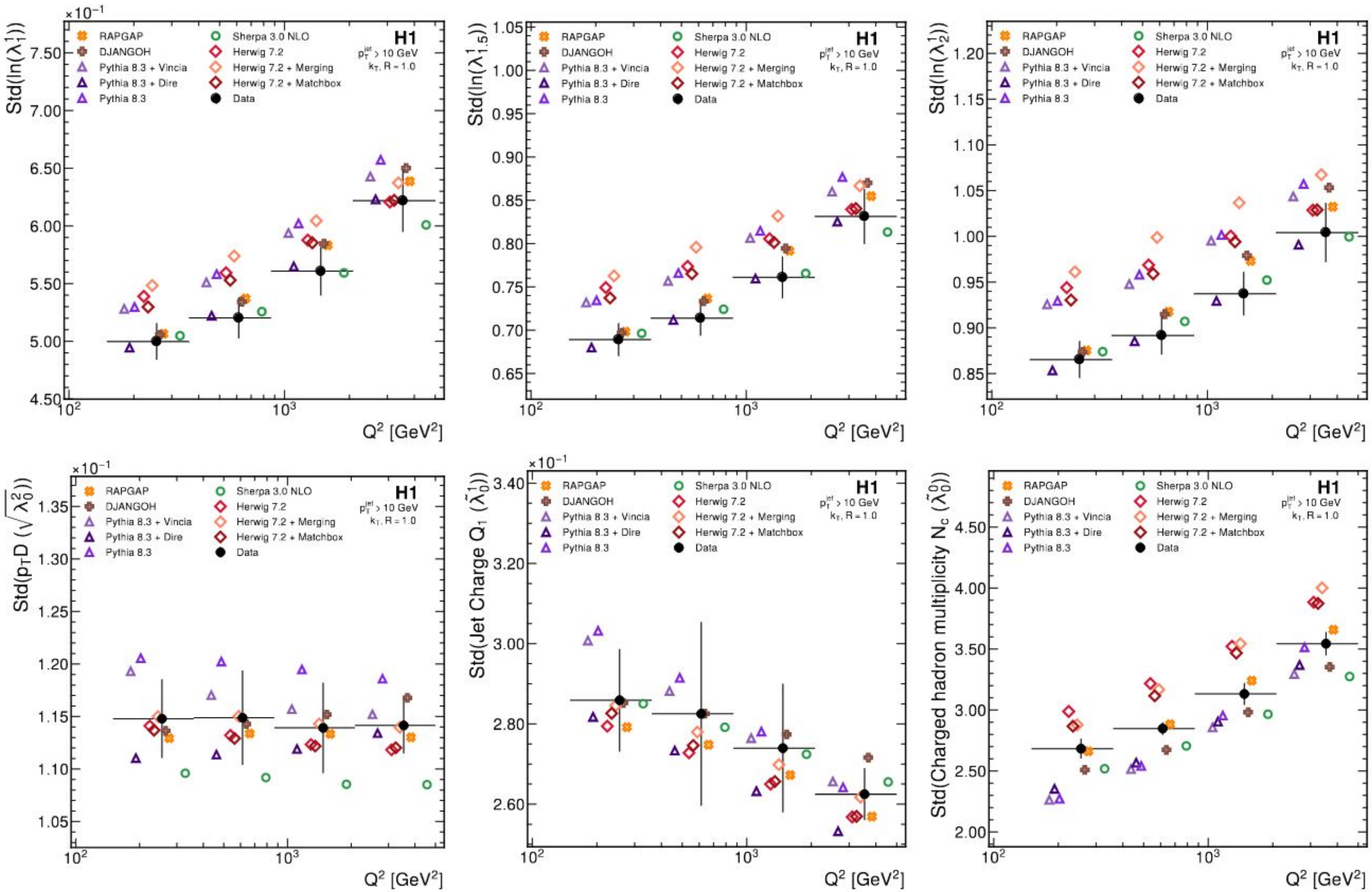


More quark-like behaviour at higher energies: mean jet charge becomes more positive

Agreement between general purpose generators **improve** at higher Q^2

Multi-differential

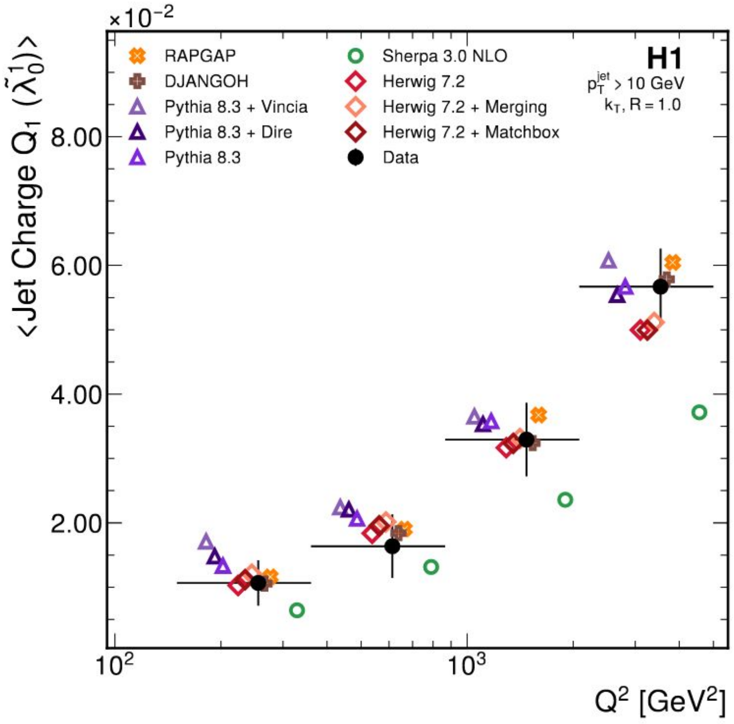
Standard deviation of all distributions also unfolded for free



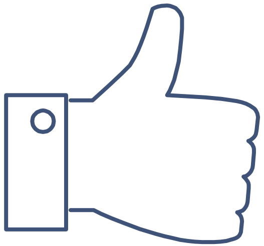
Worse general agreement between data and simulations

Conclusions

Conclusions and prospects



- Jet observables are an unique laboratory to study **QCD properties**
- **Energy scale** evolution for each jet observable measured in multiple **Q^2 intervals from 150 to 5000 GeV^2**
- Detector effects are corrected using the **Omnifold method** with particles as inputs using **graph neural networks**
  - Unbinned and simultaneous unfolding
- Unfolded the means and standard deviations without bin artifacts
- Good agreement for dedicated DIS generators, **worse** agreement for general purpose simulators
- Public results available at: **DESY-23-034**



THANKS!

Any questions?

Backup

Systematic uncertainties

Systematic uncertainties currently considered

- **HFS energy scale:** +- 1%
- **HFS azimuthal angle:** +- 20 mrad
- **Lepton energy:** +- 0.5% (mainly affects Q^2)
- **Lepton azimuthal angle:** +- 1 mrad (mainly affects Q^2)
- **Model uncertainty:** differences in unfolded results between Djangoh and Rapgap
- **Non-closure uncertainty:** Differences between the expected and obtained values of the closure test
- **QED uncertainty**: Use the variation of measured quantities when radiation is turned off in the simulation
- **Statistical uncertainty:** Standard deviation of 100 bootstrap samples with replacement

MC Generators

Lund string hadronization model and **CTEQ6L** PDF set

- **Djangoh:** Dipole model from Ariadne
- **Rapgap**: PS from leading log approximation

Pythia 8.3: default NNPDF3.1 PDF

- **Vincia**: p_T ordered antenna and NNPDF3.1 PDF
- **Dire**: dipole model, similar to Ariadne and MMHT14nlo68cl PDF

Herwig 7.2: Cluster hadronization and CT14 PDF set

Sherpa 3.0: Cluster hadronization pQCD at NLO accuracy for the 1 & 2 jet final states and LO for the 3 jet contribution.